Center for Science of Information (CSoI)
Science and Technology Center

Lead Institution: Purdue University
Partnering Institutions: Bryn Mawr College
Howard University
Massachusetts Institute of Technology
Princeton University
Stanford University
Texas A&M University
University of California, Berkeley
University of California, San Diego
University of Illinois, Urbana-Champaign

Cooperative Agreement Award: CCF-0939370
Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Section I:</td>
<td>General Information</td>
<td>1</td>
</tr>
<tr>
<td>Section II:</td>
<td>Research</td>
<td>26</td>
</tr>
<tr>
<td>Section III:</td>
<td>Education</td>
<td>108</td>
</tr>
<tr>
<td>Section IV:</td>
<td>Knowledge Transfer</td>
<td>149</td>
</tr>
<tr>
<td>Section V:</td>
<td>External Partnerships</td>
<td>174</td>
</tr>
<tr>
<td>Section VI:</td>
<td>Diversity</td>
<td>183</td>
</tr>
<tr>
<td>Section VII:</td>
<td>Management</td>
<td>200</td>
</tr>
<tr>
<td>Section VIII:</td>
<td>Center-Wide Outputs and Issues</td>
<td>206</td>
</tr>
<tr>
<td>Section IX:</td>
<td>Indirect/Other Impacts</td>
<td>269</td>
</tr>
<tr>
<td>Section X:</td>
<td>Budget</td>
<td>273</td>
</tr>
<tr>
<td>Appendix A</td>
<td>New Faculty Member Biography</td>
<td></td>
</tr>
<tr>
<td>Appendix B</td>
<td>Organization Chart</td>
<td></td>
</tr>
<tr>
<td>Appendix C</td>
<td>External Advisory Committee and Executive Committee Notes</td>
<td></td>
</tr>
<tr>
<td>Appendix D</td>
<td>Media Materials</td>
<td></td>
</tr>
</tbody>
</table>
I. General Information

The Center for Science of Information (CSol) was established on August 1, 2010 with Purdue University as the lead institution. The CSol is comprised of ten partner institutions, each with unique and complementary strengths in research, education, and outreach. The CSol is a research center sponsored by the National Science Foundation as a Science and Technology Center (STC Cooperative Agreement: CCF-0939370). The overarching vision of the CSol is to develop rigorous principles guiding the extraction, manipulation, and exchange of information, integrating elements of space, time, structure, semantics and context. These principles are motivated by, and validated on applications drawn from various scientific, engineering, and socio-economic domains. The research and development mission of the Center is complimented by an education and outreach plan focused on training the next generation of students in this rapidly emerging discipline, significantly enhancing the diversity of students and researchers, and exposing them to novel concepts at the intersection of the science of information and its applications.

1a. Provide the following general information:

Date submitted: November 1, 2013
Reporting period: November 1, 2012 – October 31, 2013
Name of the Center: Center for Science of Information
Name of the Center Director: Wojciech Szpankowski

Lead University: Purdue University

Contact information, if changed since last reporting period

Address: 250 N. University Street
Haas Hall, Suite 202
Purdue University
West Lafayette, IN 47907
Phone Number: (765) 494-2908

Fax Number: (765)494-1696

Email Address of Center Director: spa@cs.purdue.edu

Center URL: www.soilhub.org
Names of participating institutions, role, and (for each institution) name of contact person and other contact information, if changed since last reporting period

<table>
<thead>
<tr>
<th>Institution 1: Bryn Mawr College</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contact Person: Deepak Kumar</td>
</tr>
<tr>
<td>Address: 247 Park Science Building</td>
</tr>
<tr>
<td>Bryn Mawr College</td>
</tr>
<tr>
<td>Bryn Mawr, PA 19010</td>
</tr>
<tr>
<td>Phone Number: 610-526-7485</td>
</tr>
<tr>
<td>Fax Number: 610-526-5086</td>
</tr>
<tr>
<td>Email Address: <a href="mailto:dkumar@brynmawr.edu">dkumar@brynmawr.edu</a></td>
</tr>
<tr>
<td><strong>Role:</strong> Engage in Education and Research (Life Sciences and Knowledge Extraction Thrusts)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Institution 2: Howard University</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contact Person: Legand Burge</td>
</tr>
<tr>
<td>Address: L.K. Downing Hall</td>
</tr>
<tr>
<td>2300 Sixth Street, NW</td>
</tr>
<tr>
<td>Washington, DC 20059</td>
</tr>
<tr>
<td>Phone Number: 202-806-4852</td>
</tr>
<tr>
<td>Fax Number: 202-806-5859</td>
</tr>
<tr>
<td>Email Address: <a href="mailto:blegand@scs.howard.edu">blegand@scs.howard.edu</a></td>
</tr>
<tr>
<td><strong>Role:</strong> Engage in Education and General Research</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Institution 3: MIT (Massachusetts Institute of Technology)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contact Person: Peter Shor</td>
</tr>
<tr>
<td>Address: MIT, Room 2-369</td>
</tr>
<tr>
<td>77 Massachusetts Ave.</td>
</tr>
<tr>
<td>Cambridge, Massachusetts 02139</td>
</tr>
<tr>
<td>Phone Number: 617-253-4362</td>
</tr>
<tr>
<td>Fax Number: 617-253-4358</td>
</tr>
<tr>
<td>Email Address: <a href="mailto:shor@math.mit.edu">shor@math.mit.edu</a></td>
</tr>
<tr>
<td><strong>Role:</strong> Engage in Education and Research (Semantics and Knowledge Extraction Thrust)</td>
</tr>
</tbody>
</table>
### Institution 4: Princeton University

**Contact Person:** Sergio Verdu  
**Address:** B-308 Engineering Quad, EE  
Princeton University  
Princeton, NJ 08544  
**Phone Number:** 609-258-5315  
**Fax Number:** 609-258-3745  
**Email Address:** verdu@princeton.edu  
**Role:** Engage in Education and Research (Communication and Knowledge Extraction Thrusts)

### Institution 5: Stanford University

**Contact Person:** Andrea Goldsmith  
**Address:** Packard 371  
Dept. of Elec. Eng.  
Stanford, CA 94305  
**Phone Number:** 650-725-6932  
**Fax Number:** 650-723-9251  
**Email Address:** andrea@ee.stanford.edu  
**Role:** Engage in Education and Research (Communication and Life Science Thrusts)

### Institution 6: Texas A&M

**Contact Person:** PR Kumar  
**Address:** 3259 TAMU  
Room 331E, WERC  
College Station, TX 77843-3259  
**Phone Number:** 979-862-3376  
**Fax Number:** 979-845-2630  
**Email Address:** prk@tamu.edu  
**Role:** Engage in Education and Research (Communication and Knowledge Extraction Thrusts)
### Institution 7: University of California at San Diego

**Contact Person:** Shankar Subramaniam  
**Address:** 9500 Gilman Drive  
University of California at San Diego  
La Jolla, CA 92037  
**Phone Number:** 858-822-3228  
**Fax Number:** 858-822-3752  
**Email Address:** Shankar@sdsc.edu  
**Role:** Engage in Education and Research (Life Science Thrust)

### Institution 8: University of California, Berkeley

**Contact Person:** Bin Yu  
**Address:** University of California, Berkeley  
Berkeley, CA 94720  
**Phone Number:** 510-642-2021  
**Fax Number:**  
**Email Address:** binyu@stat.berkeley.edu  
**Role:** Engage in Education and Research (Communication and Knowledge Extraction Thrusts)

### Institution 9: University of Illinois Urbana-Champaign

**Contact Person:** Olgica Milenkovic  
**Address:** 126 Coordinated Science Lab, MC-228  
1308 W Main St  
Urbana, IL 61801  
**Phone Number:** 217-244-7358  
**Fax Number:**  
**Email Address:** milenkov@illinois.edu  
**Role:** Engage in Education and Research (Communication and Life Science Thrusts)
1b. Provide, in one page or less, brief biographical information for each new faculty member by institution. Attach as Appendix A.

See new faculty information in Appendix A

1c. Provide the name and contact information for the primary person to contact with any questions regarding this report.

Bob Brown

Managing Director

Purdue University
250 N University St, Suite 202
West Lafayette, IN 47907

Phone Number: (765) 494-2914
Fax Number: (765) 494-1696
Email Address: bobbrown@purdue.edu
2. **Context Statement**

The Context Statement should include a brief overview of the vision, goals, plans, and performance and management indicators for the Center. Any significant changes from the original plans for the Center should be described. This section also reports on progress toward meeting the goals set for the Center (described in detail in the remaining sections) and provides an overview of significant accomplishments during the reporting period. The Context Statement also should contain include a discussion of how the Center’s accomplishments in the past year fit within the overall Center accomplishments since the Center’s inception. In addition, the Context Statement should situate the work of the Center within the context of the disciplinary field(s) at large.

**Center for Science of Information Mission Statement**

*To advance science and technology through a new quantitative understanding of the representation, communication, and processing of information in biological, physical, social, and engineered systems.*

*The Center aspires to make advances in research, education, diversity, and knowledge transfer, particularly as they relate to the Science of Information, and more generally to the broad area of computing and communications.*

To realize this ambitious vision, the Center for the Science of Information (CSoI) sets forth a number of technical goals. These goals include models, methods, analyses, and software that cover aspects of information not addressed before. Building from a core of models and methods, the Center aims to apply these formalisms to diverse application areas. In particular, methods developed at the Center are validated on applications in *life sciences, communication, and knowledge management*, demonstrating novel insights in each of these diverse scientific domains.

**Research Overview and Goals**

Information is the basic commodity of our era—permeating every facet of our being. In 1948 Claude Shannon laid the foundations of modern data communication and storage and its ancillary trillion-dollar economic impact. Shannon introduced a general mathematical theory of the inherent information content in data and its reliable communication in the presence of noise. While Shannon’s Theory has profound impact, its application beyond storage (of sequences) and point-to-point communication (e.g., the Internet) poses fundamental challenges. The Center’s prime goal is to develop a new science of information that incorporates common features generally associated with data or information. It aims to develop a comprehensive science of information that includes the definition of core theoretical principles, development of meters and methods based on these principles, and applications to important problems in a diverse range of domains. Beyond these core technical objectives, the Center offers:

- A venue for multidisciplinary interactions and collaborations, while providing effective mechanisms for long-term, integrated scientific and technological research and education;
• A means for exploring effective ways to educate students and to train the next generation of researchers;
• Programs for broadening participation of underrepresented students and enhancing their training with deep and foundational problems in information sciences; and
• Mechanisms for timely transfer of advances in research to education and to technological advancement.

CSoI brings together accomplished researchers from diverse disciplines (computer science, information theory, life sciences, chemistry, physics, statistics, environmental sciences, economics, and social sciences) to develop a unique multidisciplinary perspective and to formulate solutions with significant broader impact.

**Integrated Research**

Our integrated research mission is to create a shared intellectual space, integral to the Center’s activities, providing a collaborative research environment that crosses disciplinary and institutional boundaries.

During the first three years of the CSoI, we have focused our efforts on two explicitly specified general research goals:

1. **Back off from Infinity**

   We propose to extend Claude Shannon’s findings to finite-sized data structures (graphs, sets, social networks, etc.); that is, to develop information theory of various data structures beyond first-order asymptotics, in order to predict the behavior of real systems with finite-length descriptions. We observe that many interesting information-theoretic phenomena appear in second-order terms. We accomplish these goals through nonasymptotic analysis valid for any length and development of full asymptotic expansions, e.g., using complex analytic tools applied to information-theory problems (known as analytic information theory). We have made significant progress toward this research goal, as highlighted below and discussed in depth in Part II of this report.

2. **Science of Information**

   In order for information theory to meet the new challenges of current applications in target domains—life sciences, communication, knowledge management, economics, and physics—we must understand aspects of structural, temporal, spatial, and semantic information in dynamic and cooperative networks. We also investigate representation-invariant information, using new information-theoretic metrics, to understand information in the submicroscopic world. For example, at the quantum scale, we study the relationship between computation and information, and finally, information extraction, when limited resources are available on the receiving side. Our ambitious long-term plan includes these challenges.

To provide a research environment that crosses disciplinary and institutional boundaries, we have organized several research workshops, starting with a kickoff workshop in Chicago (October 6–7, 2010), during which we formulated our first strategic plan. We followed with several opportunistic workshops at Allerton (September 2010 and 2011); Stanford (January 2011, May 2012); UCSD (February 2011 and 2012); Princeton (May 2011, March 2012, and September 2013); Purdue (September 2011 and July 2012); MIT (June 2012); and our 2013 Big Data Workshop.

Following last year’s NSF site-visit recommendations, we organized Grand Challenges Workshops to formulate our moderate and long-term research agenda. Following our External Advisory Board’s
recommendation, we created three Center-wide postdoctoral positions. For these positions, we have selected Tom Courtade to work jointly at Princeton and Stanford, Rui Ma for a joint position at UC San Diego and UC Berkeley, and Zhiying Wang to work at MIT, UIUC, and Stanford. These Center-wide postdocs are pursuing research goals that are closely aligned with the Center’s mission.

We made significant progress in our main research thrusts—namely, life sciences, communication, and knowledge management, as discussed in depth in Part II of this report. In this executive summary, we highlight some of our findings in understanding temporal, spatial, structural, learnable, and semantic information flow in dynamic networks with resource constraints and place them in their corresponding application context.

3. Structure

Structural information appears in diverse applications—from biology to social networks to material sciences. However, we do not have good metrics for information that is embodied in structure latent in “unconventional” data sets. For example, a graph can be represented as a binary matrix that further can be viewed as a binary sequence. However, such a sequence does not exhibit internal symmetries that are conveyed by the graph automorphism (such automorphisms make certain sequences or matrices “indistinguishable”). The main challenge in dealing with such structural data is to identify and describe regular structural relations. In fact, these “regular properties” constitute “useful (extractable) information,” understood in the spirit of Rissanen’s “learnable information.” Furthermore, such data structures often have two types of information—the information conveyed by the topological structure itself and the information conveyed by the data labels implanted in the structure. A significant outcome of our research would be an efficient way to jointly represent these two types of information. Currently, topological structure is relatively less understood. Therefore, we focused on topological structures.

Szpankowski and his collaborators (Choi 2012a, Choi 2012b) address this problem by studying information of unlabeled graphs. They first derived a relation between the entropy of a (labeled) graph and the structural entropy of (unlabeled) graphs. This gives, as a result, the design of the first provably asymptotically optimal compressor that achieves the structural entropy up to the first two leading terms. The next step is to extend these findings to the lossy case, but this will require us to define a distortion measure that preserves structural properties, which is a challenging task. Furthermore, extending these findings to other graph-generation models (e.g., the Albert–Barabási model) is nontrivial but crucial for applications.

In the last year, we concentrated on understanding the symmetry property of such power-law graphs as the Albert–Barabási model, uniform model, and others. Szpankowski and his student Manger made significant progress; however, this project is ongoing. Our aim is to achieve a final characterization by the fifth year of the grant. Once we reach this goal, we can then move toward a general theory of structures, as well as toward analogous results on related data structures, such as sets and trees and hypergraphs. In fact, Szpankowski’s group has a more ambitious goal—to develop a novel theory of graph compression with dependent labels (that is, to compress graph structure when labels are correlated). The main questions are: How much can one gain by compressing together the graph structure and dependent labels? and How can one design a provably optimal algorithm to achieve these optimal rates? This is a work in progress, and we expect to continue this investigation in the years ahead.

Center researchers Neville and Qi study graph structures and dependent substructures in social networks. The main premise of their work is that, in order to model decision making and behavior in
social-like networks, one must efficiently estimate joint distributions over possible network structures, and must accurately assess the significance of discovered patterns. In particular, Neville focused on developing statistical learning methods for a single network domain. Qi developed stochastic models for identifying latent structures in social and biological networks.

Qi observed that many complex systems can be studied from a network perspective; for instance, the study of biological networks has yielded new insights that would not have been obtained via individual gene studies. However, network analysis presents new challenges. First, many real-world networks are believed to be modular, but how can we learn network modules from data and determine the number of modules? Second, how do we link the modules to various properties or phenotypes of the whole system (e.g., cancer progression)? Third, networks (e.g., brain connectivity maps) evolve over time but modeling network dynamics remain elusive. To address these challenges Qi’s group (Zhe 2013) continues developing nonparametric Bayesian models on matrices or tensors. Subramaniam’s group also studied dynamic module detection (Narayanan 2011).

Neville’s group looks at social networks in which data comprise not only a set of entities (e.g., users), but also the observed relationships among them (e.g., friendships). Collective classification methods aim to exploit statistical dependencies that are naturally encoded by the observed links in the data. Previous models that were developed for collective classification have been shown to significantly improve predictive performance in many network settings. However, in some scenarios, the models exhibit poor performance that is even worse than conventional methods, which assume the entities are independent and ignore their relational links. To address this issue, Neville’s group developed a novel latent relational model based on copulas, which makes predictions collectively while ensuring identical marginals. Copula methods separate the specification of the models for the marginal distributions from the dependence structure that links these to form a joint distribution. This flexibility facilitates new approaches to modeling complex dependencies among entities in large heterogeneous networks.

Grama’s and Subramaniam’s groups, in a major undertaking, also studied a number of important problems associated with large sparse, noisy biological networks in order to discover topological similarities and motifs. Recent experimental approaches to high-throughput screening, combined with effective computational techniques, have resulted in large, high-quality databases of biochemical interactions. These databases hold the potential for fundamentally enhancing our understanding of cellular processes and for controlling them. Recent work on analyses of these databases has focused on computational approaches for aligning networks, identifying modules, extracting discriminating and descriptive components, and inferring networks. Network alignment, in general, poses significant computational challenges, because it is related to the subgraph isomorphism problem (known to be computationally expensive). For this reason, effective computational techniques focus on exploiting the structure of networks (and their constituent elements), on alternate formulations in terms of underlying optimization, and on using additional data for simplifying the alignment process. Mohammadi (2012) presents a comprehensive survey of these approaches, along with important algorithms for various formulations of the network-alignment problem.

Dynamic networks are very much on Lynch’s and her group’s research agenda. In her recent work with CoSI-sponsored student, Oshman, they introduce an abstract model for dynamic networks. In contrast to much of the literature on mobile and ad-hoc networks, their model makes fairly minimalistic assumptions; it allows the network topology to change arbitrarily from round to round, as long as the communication graph is connected in each round. They show that, even in this weak model, global computation is still possible, and any function of the nodes’ initial inputs can be computed efficiently.
Also, using tools from the field of epistemic logic, they analyze information flow in dynamic networks and study the time required to achieve various notions of coordination.

Realizing the importance of structural information, we discussed this issue in depth during our Grand Challenges Workshops in 2012 and 2013. We agreed to continue our effort on a number of important problems: (1) identification of statistically correlated graph modules (e.g., given a graph with edges weighted by node correlations and a generative model for graphs, identify the most statistically significant correlated networks); (2) significance measures for alignments (given an alignment of two graphs, evaluate the significance of the alignment); (3) models and methods for dynamic networks (generative models, conservation, and discriminants), (4) and network integration (generalizing conventional analysis in principled ways to multiscale heterogeneous networks).

Recently, during our 2013 workshop, we updated the grand challenges in the knowledge management thrust by explicitly searching for structural relationships in big data. We acknowledge that perhaps the greatest challenge in applying information-theoretic principles to a broader suite of problems—including biological systems, analytics for massive data sets, and social networks—is that of developing meaningful notions of “structural information” and establishing a set of corresponding fundamental results. Specifically with respect to the analysis of big data—the data size and complexity lead to the following challenge: To identify, encode, and test the underlying structure of “big” data, a trade-off exists between the accuracy of the data model and the amount of data needed to support and test the model.

4. Delay

The mathematical theory of information was originated from Shannon’s channel capacity, defined as the maximum rate that can be achieved over a channel with asymptotically small probability of error. Unfortunately, the Shannon capacity of a channel places no restrictions on complexity or delay in transmission or reception. Methods to properly characterize the complexity and the delay could potentially fill a large gap that would extend the Shannon capacity to dynamic networks with multipoint communication and often unpredictable delays. Furthermore, we emphasize that the increasing demands of wireless networks require delay guarantees. Applications include VoIP, video streaming, real-time surveillance, networked control, etc. A common characteristic of these applications is that they have a strict deadline associated with each packet, and the channel reliabilities of different clients can be different and even vary over time. These compelling reasons are why we need to understand the role of delay in the flow of information. We have made progress on several fronts in this endeavor, as exemplified in some of the research projects that follow.

In a series of foundational papers, Verdú, Polyanskiy, and Kostina extended the channel-coding theorem of Shannon to a finite-blocklength regime. Thanks to these results, we now understand that blocklength (closely related to delay) contributes a reciprocal of the square root of the blocklength to capacity. This is a nonasymptotic result (i.e., precise lower and upper bounds are presented), and it allows us to compute the degradation in capacity, even for small blocklengths.

In fact, Verdú’s group pioneered new finite-blocklength achievability bounds for multiuser information theory, based on a new fundamental nonasymptotic covering and packing lemmas. This machinery has succeeded in giving new achievability results for problems, such as Gelfand–Pinsker, Wyner–Ziv, Ahlswede–Korner, and inference and broadcast channels. Furthermore, Verdú’s group obtained results for channel inputs with associated costs (e.g., limited transmitted power). We showed that the maximum achievable coding rate under a cost constraint can be bounded in terms of the so called b-
tilted information that resembles the $d$-tilted information in lossy compression. Some publications resulting from this research are Kostina (2012, 2013), Polyanskiy (2013), and Verdú (2012).

In another line of research in a real-time coding system with look-ahead, Weissman (Asnani 2013) started investigating the impact of delay on expected distortion. The system we consider consists of a memoryless source; a memoryless channel; an encoder, which encodes the source symbols sequentially, with knowledge of future source symbols up to a fixed finite look-ahead, with or without feedback, of the past channel-output symbols; and a decoder, which sequentially constructs the source symbols using the channel output. The objective is to minimize the expected per-symbol distortion. Weissman provided one of the first results in this line of research. This bridges the gap between causal encoding (delay=0) and the infinite look-ahead case (delay=∞) where Shannon-theoretic arguments show that encoding–decoding separation is optimal.

Delay in networks and its relation to scheduling are discussed in P.R. Kumar’s recent work. He designed reliable scheduling policies with delay constraints for unreliable wireless networks. We focused attention on a formulation that provides a useful and tractable framework for modeling, analyzing, and designing real-time wireless communications. This framework is built on top of an analytical model that jointly considers three important challenges—a strict deadline for each packet, the timely throughput requirement specified by each client or application, and finally, the unreliable and heterogeneous nature of wireless transmissions. An important feature is that this model is suitable for characterizing the needs of a wide range of applications, and the model allows each application to specify its individual demand. Kumar presented a simple, yet extremely powerful model satisfying these constraints and designed an optimal scheduling policy.

More interestingly, Kumar and his group initiated fundamental research on security of wireless ad-hoc networks that are secure from the ground up, subject to certain axioms being validated. This is based on a min-max approach in which minimization is over all “bad behaviors” of the malicious nodes, and the maximization is over all protocols announced and faithfully followed by “good nodes.” The key ingredients to this approach are time and clocks. The ability to coordinate activity under a common time is a critical assumption to any scheduling that is part of any security design regime. In other words, in order to form a fully functional network that can efficiently and reliably carry information, it must be ensured that protocols operate at the same temporal and spatial scales. We believe this research will lead to a fundamentally new approach to security.

Temporal dynamic information is the key element in Ramkrishna’s group work on quantifying fluxes in a biological network. Their premise is to show that microbial metabolism is regulated by a survival strategy in controlling enzyme syntheses and activities that preferentially drive cellular reactions ensuring survival goals. Their theory predicts, and experiments confirm, discontinuous behavior as the growth rate increases, which never has been envisaged in traditional biological thinking.

In related research, Subramaniam’s group (Shirazi 2013) looked at time-varying causal inference from phosphoproteomic measurements in macrophage cells. In fact, they studied cellular signaling circuitry in eukaryotes by analyzing the regulation of protein phosphorylation and its impact on downstream mechanisms leading to a phenotype. A primary role of phosphorylation is to act as a switch to turn “on” or “off” a protein activity or a cellular pathway. Specifically, protein phosphorylation is the main step in the process of transducing molecular signals inside the cell. It is important to reconstruct context-specific signal-transduction mechanisms of the cell, because errors in transferring cellular information can alter the normal function and may lead to such diseases as diabetes, cancer, and autoimmunity. Recent advances in high-throughput "omics" technologies, combined with advanced
data integration, analysis, and interpretation, are facilitating a detailed and quantitative understanding of the underlying biochemical mechanisms. In this study, Subramaniam’s group has developed a novel framework for time-dependent reconstruction of signaling networks involved in activation of macrophage cells leading to an inflammatory response. Several signaling pathways have been identified in macrophage cells, but the time-varying causal relationship that can produce a dynamic directed graph of these molecules has not yet been explored. They used the Granger causality and apply a vector autoregressive model to the time–course data made available by the Alliance for Cellular Signaling (AfCS) in RAW 264.7 macrophage cells on phosphoproteins in single- and double-ligand experiments for 22 ligands. Through the reconstruction of the phosphoprotein network, Subramaniam’s group was able to estimate connectivity, Granger causality, and the dynamics of information flow. The time-varying framework results in a succession of the phosphoprotein network in three stages that represent the evolution of the biological subsystem as a time-dependent cascade.

In our Grand Challenges Workshops, we discussed further possible advances in this important area. We realize that recent work by Center members has made significant strides in extending Shannon theory to a finite blocklength regime. While understanding this model is of great importance, it does not fully capture the reality of many communication- and information-processing applications. Modern utilizations of these ideas demand the capability to process very large streams of data. Moreover, it is delay rather than blocklength that is the system-performance figure of merit. Characterization of the trade-offs between delay and other performance metrics—such as distortion, bit error rate, and complexity—is a challenge that Center members endeavor to fully analyze. This area promises to remain a focal point for the activity of several Center members in the years to come.

5. **Space**

In the modeling of cellular systems, Bialek’s group (Tkacik 2011, 2012) explored transmission of information in making spatial patterns. We showed that pattern formation in developing embryos is related to the information capacity of certain communication channels, which sets a limit on the complexity and reliability of the ultimate patterns formed. Bialek’s group is in the process of generalizing this result, and to turn it around, to show that the measurable “positional information” in patterns can be used to give a lower bound on the capacity of the key channels for communication of this information. If accomplished, this would allow us to show that the relatively macroscopic phenomenology of embryonic development can be used to reach conclusions about the minimal capacity of the microscopic cellular signaling and regulatory pathways.

Coleman’s group has recently developed (Kim 2013) a class of methods to succinctly capture spatiotemporal patterning of neural spiking in the cortex to reflect directed, causal functional network mechanisms. These methods are rigorous, computationally efficient, and scalable. The group has instantiated the theoretical results to guide recent neuroscientific findings that relate networks of functionally connected neurons to cortical-wave propagation within the primary motor cortex. It was established that this spatial layout accounts for the magnitude of task-relevant information contained in the sequential firing of functionally connected neurons. They further demonstrated cell-type specificity—excitatory neurons of thin spike width play a dominant role in this ensemble spike sequencing. Furthermore, Coleman’s team estimated (Quinn 2013a, 2013b) functionally connected networks of spiking neurons using a directed information analysis for point processes to demonstrate that a class of simultaneously recorded, single-motor cortical neurons in nonhuman primates spatially coordinates its spiking activity in a manner that mirrors the beta-wave axis. The circular distribution of
excitatory connection directions was bimodal and oriented close to the beta-wave axis in all three monkeys. Goldsmith’s group also used directed information to infer synapse connections between neurons.

6. Semantics

Shannon in his 1948 paper asserted: “Frequently, the messages have meaning, that is, they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem.” However, Sudan and his collaborators argue that the meaning of information does start to become relevant, whenever there is diversity across communicating parties and when parties themselves evolve over time. For example, when a computer attempts to communicate with a printer, it must talk the same language in the same format (e.g., the API of the “printer driver”). This led Sudan and his collaborators to consider communication in the setting where encoder and decoder do not agree a priori on the communication protocols, and thus the encoder and decoder do not understand each other. In several new findings, Sudan’s group proposed a mathematical theory of goal-oriented communication. These are among the first results that may lead to a new information theory of semantic communication.

Sudan in related work considered “learning PAC semantics”: namely, the design of efficient algorithms for reasoning by using a combination of explicit knowledge (given as logical formulas) and partial examples drawn from a (probabilistic) source. He observes that the use of random examples to draw broader (i.e., inductive) conclusions requires a weaker standard than the usual Tarskian semantics of logic, which is formally provided by Valiant’s PAC Semantics for logic. PAC Semantics provides a clean standard that captures the utility of the answers provided by such reasoning algorithms, which essentially amount to “rules of thumb” about a given domain.

Sudan also considers “communication amid uncertainty” to understand how information can be communicated effectively when the sender and the receiver are uncertain about each other. These two research goals—that of semantics and communication amid uncertainty—are related. The first considers a broader class of problems, while focusing in particular on not knowing each other’s prior beliefs exactly (but having approximate knowledge of these). The second considers the specific task of not “understanding” each other, but here it is harder to assume, or even define, approximate understanding, and the researchers hope to explore such questions.

Furthermore, Sudan and Tse recently initiated a new collaboration in this area, realizing that semantics is certainly important in communication. One aspect is the senders’ uncertainly about the background knowledge of the receiver, for instance, in human communication. Sudan and coauthors have analyzed a specific instance of this problem, in which the sender has to compress a message to send to the receiver, but the receiver has uncertainty about the distribution of the message. However, the fundamental limit of this problem turns out to be rather pessimistic. Moreover, as Tse has observed, the issue seems to be that there is no structure whatsoever assumed on the distribution. On the other hand, from basic information theory, it is known that if the message comes from a stationary ergodic distribution, then universal compressors like Lempel–Ziv can attain the entropy of the source without the receiver having any prior knowledge on the distribution. Nevertheless, stationary ergodic sources may be too restrictive a class of distribution to model human communication. In preliminary collaboration with Sudan, Tse and student Kamath, they seek a less restrictive but meaningful class of sources for which stronger statements can be made than assuming no structure in the source at all. One class may be the large-alphabet sources advocated by Orlitsky and recently discussed in Szpankowski (2012).
7. Cooperation and Dependency

In a major extension of the Shannon framework, Cover and his collaborators initiated a theory of cooperation and coordination in networks. A general understanding of the limits of dependence yields rate-distortion theory (data compression) as a special case and provides a general approach to distributed data compression and cooperation. It also elucidates such diverse processes as intercellular biological communication. The role of dependence is exemplified by the telephone system, wireless communication, the Internet, news services, the economies of large countries, and the internal workings of computer hardware. The efficacy of all of these systems depends on fast communication and consequent cooperative behavior. What limits on physical dependence are imposed by the speed of information? Are there energy constraints on computation? Some of these vast generalities can be addressed by developing a science of information for dependence and coordination. Fundamental for coordination is the ability for distributed agents to create joint probability distributions, which might then be used for strategic purposes. Anantharam and his student Kamath (Kamath 2102) asked: What joint probability distributions can be created over networks? This is a largely open problem. They brought tools from the area of hypercontractive inequalities into this area, with remarkable effect. For instance, they showed that the strong data-processing constant of a joint probability distribution is explicitly given in terms of the so-called hypercontractivity ribbon of the joint distribution, which is the nontrivial set of pairs for which hypercontractive inequalities hold. Other Center-affiliated faculty members with strong interest in this line of research are Polyanskiy, Raginsky, and Courtade. They have enjoyed and benefited from discussion in Center-sponsored workshops and meetings.

Dependency and rational expectation are critical ingredients in Sims’ work on modern dynamic economic theory. Much of modern dynamic economic theory formulates models by examining how continuously optimizing agents will interact in markets. This has been important in allowing consistent treatment of economic behavior, but the models postulate continuous optimization, implying very rapid response to policy changes and to market signals, whereas actual behavior is more sluggish. Approaches to address this (e.g., by postulating “adjustment costs”) have an ad-hoc flavor and are not grounded in direct microeconomic observations.

Sims observes that the existing “rational expectations” theories with continuous optimization imply infinite mutual information, in Shannon’s sense, between the stochastic process representing market signals and the stochastic process representing a person’s action. At least qualitatively, recognizing that this rate of information flow must be finite explains a broad array of observed facts about economic behavior that has, in the past, been explained with ad-hoc postulates of inertia or adjustment costs. Sims attempts to integrate a formal information-theoretic approach into dynamic economic theory. This seems to be a promising avenue for both explaining observations and improving the formulation of economic policy. In fact, thanks to feedback during the kick-off workshop in Chicago, Sims used, in his recent work, some novel techniques of rate distortion developed previously by the information-theory community. His efforts and those with his doctoral student Matejka resulted in his recent papers.

In the Grand Challenges Workshop, we agreed to continue this effort. In particular, we plan to concentrate on studying the impact of information flow on the dynamics of economic systems and the impact of agents with widely differing capabilities (information and computation) on the overall status of the systems. Raginsky, who joined the Center last year, collaborates with Sims on these issues. They also look at the value of information in economic context.
Finally, in systems biology, one of the focal points is the discovery of causal relationships among different components of biological systems. Gene regulatory networks, protein-protein interaction networks, chemical signaling, and metabolic networks all exhibit causal relationships between their agents that are crucial for proper functionality. Discovering such causal relationships through experiments may be a challenging task, due to the technical precision required from the experiments and due to the large number of interconnected and dynamically varying components of the system. It is therefore of great importance to develop an analytical framework for discovering causal connections between genes and for elucidating the gene interactome, based on limited experimental data. Analytically inferred interactions may consequently be used to guide the experimental design process, which would then help in further refining the modeling framework. One way to detect if a gene causally influences another gene is to observe the target gene’s expression levels and detect if changes in the expressions of the other genes affect changes in the expressions of the target. For this purpose, a number of authors suggested the use of Bayesian network modeling and other machine-learning tools, algebraic techniques, information-theoretic methods, and autoregressive Granger causality approaches. Coleman, Subramaniam, and Weissman use Granger causality to analyze biological neuron networks, as well as the stock exchange.

8. Learnable Information

A major goal in this area is to initiate “information theory of massive data.” In the last years, we made some progress in this direction—in particular, for information-theoretic modeling of recommender systems. Center-wide postdoc Courtade and Weissman, collaborating with Verdú, are using multiterminal source coding under logarithmic loss to present fundamental limits of recommendation systems, in order to estimate the need for and potential benefits of improved recommendation. More precisely, the recommendation can be viewed as investments and the actions of the customers as specific outcomes of the market. Indeed, by providing a limited number of recommendations (e.g., showing two different advertisements), the recommender is betting that the customer will follow the displayed recommendation. That is, if the customer clicks on a presented advertisement, the recommender wins (and collects its winnings); if the customer doesn’t click on a presented advertisement, the recommender gains nothing (and misses a chance at potential earnings). This simple market model is often called a horse race market, and it is well known that recommending according to a Kelly betting strategy will asymptotically outperform any other recommendation system (this is closely related to the concept of a log-optimal portfolio in more general markets). For example, Netflix invested over $1M in order to obtain a 10% improvement in their recommendation system. Should Netflix invest another $1M hoping for an additional 10% improvement? What if it was revealed that only a 1% improvement was theoretically attainable; would it still justify a $1M investment?

Roughly speaking, in a recent paper, Courtade and Weissman (Courtade 2013) provide a framework that characterizes the fundamental limits of recommendation systems in terms of attainable wealth, and thus partially answers the questions posed above. Specifically, a recommendation system can be broken into two parts: (1) a data-mining part that extracts relevant customer information from a database, and (2) a recommendation part that selects the best recommendations based on the extracted customer information. The former part can be viewed as a compression mechanism; that is, a large database is compressed to a representation that is useful to the recommendation part of the system. As with any rate distortion problem, there is a fundamental trade-off between the compression rate (i.e., the granularity of extracted information) and the wealth attainable by the recommendation system. Results presented give a precise characterization of this trade-off under
some mild and realistic assumptions on the statistics of the data. In fact, Courtade and Weissman were able to solve the Holy Grail of multiterminal source coding: lossy reconstruction of two correlated sources based on limited-rate descriptions of each of the sources separately (a problem referred to generically as ‘multiterminal source coding’). A complete solution to the problem was given in (Courtade 2013).

In a related work, Kulkarni (Shang 2013) has been investigating techniques for recommender systems. He and his colleagues have proposed a hybrid collaborative filtering model based on a Markovian random walk to address the data sparsity and cold-start problems in recommendation systems. More precisely, this work constructs a directed graph whose nodes consist of items and users, together with item content, user profile, and social network information. This approach incorporates the user’s ratings into edge settings in the graph model. The model provides personalized recommendations and predictions to individuals and groups.

In another direction, Center members continue making progress in statistical research of massive data. Yu’s research (Uhler 2013) continued to be both core statistics and interdisciplinary research. It concentrated on statistical machine-learning theory and solving data problems from such diverse fields as remote sensing, sensor networks, neuroscience, document summarization, robust speech recognition, computer systems, and image labeling with overlapping, multiple categories.

Furthermore, computational analysis has emerged as a powerful lens for examining massive data from various applications—for instance, analyzing high-throughput single-cell data for cancer stem-cell detection and processing ab initio computational data for new battery design. At the same time, massive data of growing complexity imposes new computational and statistical challenges—such as high dimensionality, complex interactions, and massive volume—which make many traditional data-analysis tools no longer applicable. To address these challenges, Center researchers Qi and Yu integrate concepts and techniques from Bayesian statistics, optimization, scientific computing, and parallel computing to develop principled and scalable computational tools to analyze complex massive data. These tools are being used to discover critical associations between genetic biomarkers and diseases (e.g., Alzheimer’s disease), capture biological, social, and atomic network interactions and their dynamics, detect rare populations of cancer stem cells from huge single-cell data sets, and help the virtual design of new materials for energy applications.

Neville’s recent work aims to formalize many of the “intuitions” behind past development and successful application of statistical relational learning (SRL) methods for single network domains (e.g., Facebook, the World Wide Web) in order to characterize algorithm performance analytically, and to use the understanding to drive the development of techniques that will be of greater use to data scientists. Neville’s group has formalized the notion of weak dependence for heterogeneous networks and has used this as a foundation to outline the conditions under which estimation methods are asymptotically consistent and normal, providing theoretical justification for current SRL approaches—which is, to the best of our knowledge, the first formal analysis for estimation in the context of single-network domains.

Atallah’s group is investigating privacy and security. The overall goal in this area is to design techniques that enable the extraction of useful knowledge from information that is both confidential and distributed among multiple parties who are reluctant to share it with each other. Atallah’s special focus within this area is the case where a client’s confidential information is stored at a remote untrusted site (e.g., in the cloud) and the client wishes the remote server to analyze and manipulate the information on behalf of the client, yet without revealing answers to the remote server about the...
stored information or the computation. In a recent publication (Blanton 2012), Atallah's group designed an algorithm that improves spectacularly the previous state-of-the-art (e.g., the algorithm runs in $O(m+n)$ instead of $O(mn)$).

Rivest's research efforts in this area have focused on the role of uncertainty in knowledge, particularly in two-party or multiparty contexts, where knowledge of the actions or beliefs of the other party (or parties) is relevant to one’s own actions or beliefs. The information one has about the actions or beliefs of the other party (or parties) may be naturally restricted, according to the situation at hand. In addition, one or more of the parties may be acting or communicating in a deceptive or malicious manner, so inference may be more difficult than in a cooperative setting. In particular, he has focused on the following two application areas: (1) the world of poorly kept secrets and (2) elections and voting.

In recent work Rivest and CSol-sponsored student Virza made a breakthrough in real-world implementation of zero-knowledge proofs for all of NP statements. Their system also tackles the problem of representation supporting verification of correct execution of arbitrary C programs. Zero-knowledge (ZK) proofs are crown jewels of modern cryptography: these proof systems reveal validity of the statement, but nothing more. For example, one could prove that a number is a composite without revealing its factorization. Such proofs are important building blocks of many cryptographic tools including secure multiparty computation, anonymous credentials, electronic voting, and others. It is known that ZK proofs of knowledge exist for all of NP, but known constructions are practical only when proving statements of special form that avoid generic NP reductions (e.g., pairing product equations).

Despite recent theoretical and practical progress, obtaining constructions that are both generic and efficient in practice remains a long-standing goal in cryptography. Research by Virza achieved exactly this. These results may have potential impact for advancing biology and communications. The communication ability to efficiently prove arbitrary NP statements has immediate applicability to outsourcing computation. That is, a computationally weak client can give his long-running computations to a third party, without worrying about its integrity as the results would be accompanied by an easy-to-verify mathematical proof. Biology-efficient zero-knowledge proofs have the potential to transform biomedical computations. For example, companies like 23andme maintain genome databases that could be very useful for advancing our understanding of health, but cannot be shared because of privacy concerns. They need to outsource computations to an untrustful but computationally powerful client.

In a novel line of research (Jacquet 2013), Szpankowski’s group provided a precise estimate for the joint-string complexity (Jacquet 2012b) to classify texts, such as Twitter texts. Finally, in collaboration with Amgen in (Apostol 2012) Szpankowski’s group also used information-theoretic tools for purity measurements in pharmaceutical data.

Finally, Kumar and Baryshnikov (Xie 2013) began working on the problem of data analysis of the large amount of real-time synchrophasor data generated by wide deployment of phasor measurement units in the power grid. To date, they have studied the underlying dimensionality of the data and established that it is low dimensional. Moreover, changes in data arising from events are detected very rapidly. This has made feasible the development of early detection of events, such as line tripping, unit tripping, and control input change. As a continuation of this effort, Kumar and his group are looking at nonlinear dimensionality reduction.
In our Grand Challenges Workshops, especially in the 2013 Big Data Workshop, we identified several open problems in this area that we plan to attack. For example, the first challenge is in developing formal methods for quantifying the trade-offs and deriving suitable metrics for compressed data analysis and efficient search in compressed data. These metrics must account for constraints of space and computation, overheads of distributed storage, and considerations of robustness and fault tolerance. Another challenge is associated with sparse high-dimensional data sets. As techniques for dealing with these data sets are developed, there is increasing recognition of the fact that the magnitude of this data poses significant challenges. Current approaches to this problem focus on either distributing the data sets (virtually all major Internet infrastructure—Google, Facebook, Twitter use this) or relying on out-of-core computations to deal with the data sets. We posit that compressed analysis would yield one to two orders of magnitude improvement in storage and network cost in distributed storage. The challenge is to support analyses and query tasks, such as dimensionality reduction on compressed data, clustering, projections, search, tensor decompositions, and visualization. Specifically with respect to the analysis of big data, we identify three main challenges:

1. **Discovering structure in data**: To identify, encode, and test the underlying structure of “big” data, a trade-off exists between the accuracy of the data model and the amount of data needed to support and test the model.

2. **Managing and querying data**: To collect, store, and query “big” data in data management systems, a trade-off exists between the efficiency (space) of the data structures and the efficiency (time) of the algorithms that query and access the data.

3. **Ranking findings and finding rankings in very wide data sets**: A very wide data set contains more features than records. By a finding, we mean here an association rule, a correlation, or a pattern. By a ranking, we mean a function that associates a measure of validity or interest with a finding. Such ranking function may return a \( p \)-value combined with some measure of interest of a finding.

**9. Resource Constraints**

Information processing within cells is limited by fundamental physical principles: individual molecular events are stochastic, which limits the precision and dynamic range of small signals. Bialek’s group (Tikhonov 2013, Tkacik 2013) has been working on deriving the structure of molecular networks that optimize information flow, subject to constraints on the total number of molecules being used. To their surprise, they found that information transmission is optimized by either self-activation or self-inhibition, depending on molecule concentration. This is the first step in more general work on networks with feedback. Coleman’s group at UCSD is also pursuing these concepts. Coleman is studying statistical causality in neural systems by using Granger principles, which allow us to assess how past values of, for instance, two signals affect the predictability of a single signal. Information-theoretic models are built to explore these questions.

Energy constraints in a wireless ad-hoc network may limit its capacity. Verdú’s and Goldsmith’s groups investigated the minimum energy per bit as a function of data length in Gaussian channels. Lynch’s group has studied (Sastry 2013) capabilities and limitations of distributed information processing, in settings exhibiting such complications as dynamic changes to the communication network, timing anomalies, and failures.

Pulkit Grover, a CSoI alumnus, recently initiated an interesting line of research by leading a group of students in Center-sponsored research on the interplay between computation and information. Grover asked whether information theory can be applied to information bottlenecks in the brain or in computational circuits. After all, a wire in a circuit, as well as an axon in the brain, represents a
communication channel. However, there is a crucial difference between these applications and the ones that Shannon was addressing. While computation at the transmitting and receiving ends is of secondary importance in Shannon’s long-distance communication problems—and indeed was neglected in Shannon’s seminal work—for understanding information bottlenecks in computation (e.g., an electrical or neuronal circuit), it would be self-contradictory to neglect computational costs. Computation at transmitting and receiving ends is precisely what Grover is trying to understand. His work (Grover 2013) shows that the traditional Shannon theory is insufficient and provides rather optimistic understanding of trade-offs between rate (e.g., speed of computing) and reliability. For instance, by ignoring computational costs, Shannon theory predicts that unboundedly low-error probabilities can be achieved with bounded transmit power. However, if one includes power consumed at the transmitter and receiver ends, then Grover’s results show that the total power grows unboundedly (and quantifies how fast this power must grow), as the error probability is driven to zero. Further, computational energy also diverges as one approaches Shannon capacity.


During the course of the grant, we made significant progress in understanding the behavior of lossy compression, compressed sensing, and control. In particular, Verdú’s group (Wu 2012) has accomplished an in-depth treatment of the fundamental limits of loss data compression in the nonasymptotic regime, including an exact expression for the performance of random lossy compressors leading to a tight achievability bound; a general converse bound using \( d \)-tilted information, a random variable that corresponds (in a sense that can be formalized) to the number of bits required to represent a given source outcomes within distortion \( d \). They designed several upper bounds. In particular, they suggested pragmatic lossy compression schemes and provided finite-blocklength bounds on the length of the short code required to achieve a certain rate-distortion point \((D + E, R(D) + \gamma)\). Furthermore, they used \( d \)-tight converse bound for joint source-channel coding setting.

Furthermore, Verdú’s group used Rényi’s information dimension to explain compressed sensing. The field of compressed sensing has seen an explosion of interest since the pioneering works of Candès, Donoho, Romberg, and Tao. Its central issue is how many real-valued measurements are sufficient to reconstruct a signal exploiting some prior knowledge about it? This area of research has a strong information-theoretic flavor; however, the formulations, methods, and solutions in the literature are rather different from those in Shannon theory. A fundamental disparity is that, in compressed sensing, signals are encoded by real numbers rather than bits. Moreover, the encoder or compressor is forced to be linear, the sources are typically not modeled probabilistically, and a worst-case (rather than average) approach to performance is adopted. Verdú’s group succeeded on the quest for an information-theoretic reformulation of the fundamental limits in compressed sensing. The result is the almost-lossless analog compression paradigm. Just as Shannon entropy is the fundamental limit of almost-lossless data compression, Verdú and his group have been able to show that information dimension, introduced by the legendary Hungarian mathematician Alfred Rényi in 1959, governs the number of measurements required to reproduce an analog source almost losslessly.

One of the most significant lines of exploration from the Goldsmith and Weissman group this past year pertains to communications stemming from an attempt to characterize the influence of controlled feedback on channel capacity (Mirghaderi 2014). They found a family of channels called POST (previous output is the state) channels where despite the memory, numerical evaluation has shown that feedback does not increase their capacity. Using convex optimization, those in the
Goldsmith and Weissman group were able to show that, if the output distribution induced by a code that achieves capacity of a channel with feedback can be achieved using a coding scheme without feedback, then the feedback does not increase capacity. In other words, the induced channel-output distribution is the important criterion for deciding whether or not feedback increases capacity. This result allowed us to establish the counterintuitive result that the nonfeedback capacity is in fact equal to the feedback capacity for a large class of POST channels.

Tse and his group (Tse 2013a, 2013b) initiated information theory of high-throughput (DNA) sequencing. DNA sequencing is the basic workhorse of modern-day biology and medicine. Shotgun sequencing is the dominant technique used—many randomly located short fragments called reads are extracted from the DNA sequence, and these reads are assembled to reconstruct the original sequence. Today multiple sequencing technologies exist, and many assembly algorithms are designed for them. A basic but yet open question is: Given a sequencing technology and the statistics of the DNA sequence, what is the minimum number of reads required for reliable reconstruction? This number provides a fundamental limit to the performance of any assembly algorithm, and thus provides an important benchmark. The long-term research goal is to answer this question in general.

Tse’s group formulated an analogy between the DNA sequencing and the classic communication channel problem. The efficiency of DNA sequencing was phrased in terms of channel capacity. In recent developments, Tse’s group is now trying to answer another fundamental question: in the presence of read errors, how much degradation in performance is a consequence of the specific assembly algorithm and how much is due to fundamental limits?

Finally, the Center sponsored a student-led research project directed by Courtade to understand which Boolean functions are most informative. The motivation behind this is quite simple—imagine a simple binary symmetric channel that flips bits. If a single bit is transmitted by a repetitive code, a decoder must make a decision between 0 and 1 by computing a Boolean function (e.g., majority) of the received output. The basic question Courtade’s group asked is: What is the trade-off between the energy of decoding (computing a Boolean function) at the decoder and the error probability? The question turns out to be very subtle. A partial answer is presented in their recent publication (Courtade 2013).

Recently, however, Bialek’s group (Tkacik 2013) made significant progress in understanding the flow of information to predict the behavior of biological systems. The problem of calculating information flow in moderately realistic models of transcriptional networks is challenging. Bialek’s group took the problem in steps, starting with a single transcription factor that “broadcasts” to many noninteracting target genes. Even this simple case is rich, illustrating situations in which the optimal network has substantial redundancy, a structure that most biologists would interpret as prima facie evidence against optimization. The next level of complexity allows for interactions in the network of target genes, but with no feedback loops. This already makes it possible for repressive interactions in the network to reduce redundancy, and the resulting patterns of expression begin to remind us of those in the fly embryo, though the class of networks we are considering still is too simple to be realistic.

This year Bialek’s group pushed the theory to include the simplest example of networks with feedback, in which a single gene is both responsive to inputs and self-regulating. Importantly, we show that optimal information transmission is never in the regime where self-interactions are strong enough to generate multistability, although there are parameter regimes in which the information transmission is maximally close to the critical point for the transition to multistability. An important aspect of these results thus far is the appreciation that certain “details” of the molecular mechanism
involved in connecting transcription-factor binding to DNA with the regulation of RNA polymerase activity can have a substantial effect on information transmission.

11. Quantum Information

It is now well accepted that the flow of information in macroscopic systems and submicroscopic systems may possess different characteristics. In fact, submicroscopic systems seem not to obey Shannon’s postulates of information. In the quantum world and on the level of living cells, traditional information often fails to accurately describe reality. On the other hand, some attempt to build quantum theory entirely on the flow of information. Peter Shor and Scott Aaronson are on the front line of these investigations. For example, in a recent paper, Aaronson developed the computational complexity of linear optics. In particular, Aaronson and his student proposed new photonic experiments that can help realize a fully universal quantum computer. In fact, this experiment is being explored currently at two labs—in London and Queensland.

In recent developments, Shor makes some progress on how best to do fault tolerance on a quantum computer; this problem was upgraded to one of our grand challenges. When quantum computation was first proposed, one objection to it was that it could not be made fault-tolerant. The standard techniques for making classical computation fault-tolerant involve either redundancy (duplicating information) or error-correcting codes. It was believed that these techniques would not work for quantum computation because of the quantum no-cloning theorem (Wootters and Zurek 1982). It appears, at first, that classical error-correcting codes require redundancy, so that quantum error-correcting codes would violate the no-cloning theorem. However, it was discovered that quantum error-correcting codes do exist (Shor 1995, Calderbank and Shor 1996). Furthermore, these error-correcting codes could be used to design fault-tolerant quantum computers.

The theoretical results on fault tolerance are usually given in threshold theorems; the exact details of these threshold theorems depend on the architecture chosen for the quantum computer. The threshold theorem says that if quantum gates have accuracy above some threshold, then any circuit for a quantum algorithm can be transformed into a quantum fault-tolerant circuit using a polylogarithmic factor more gates.

Education

During this past period, the Center’s education program has fostered continual engagement and learning by students, with increases in peer-to-peer and nonadvisor-faculty collaboration (36% of students this period, compared with 15% last year, on papers and presentations). This modest growth indicates that our education program continues to support a developing community where members are building the emerging field of the science of information. Students state they value their participation in the Center because it provides a venue for stimulating their own thinking about the science of information, provides productive networking opportunities with peers and leading scholars in the field, and fosters collaborations among students and faculty.

Outcomes for the Educational Program Efforts This Period

We formed four multidisciplinary student research teams around problems connected to the Center’s research thrusts. Seven universities and multiple departments are involved in the four teams, showing interdisciplinary interactions. See more at http://www.soihub.org/research-teams.php.
• The Center’s annual summer school involved organizing the IEEE IT Society North American School. There were 140 students and faculty members who participated, with 56 oral and poster presentations by the students. See more at http://www.soihub.org/summer-school.php.

• Including the Introduction to Science of Information course, faculty developed and taught 10 new courses and adapted new SoI knowledge into 8 existing courses for both undergrads and grad students, reaching more than 700 students.

• Monthly student brown-bag presentations (9 total) reached 161 students, postdocs, and faculty this year, while providing professional development for the advanced PhD students. See more at http://www.soihub.org/seminars.php?area=edu&type=22.

• An online learning hub has been developed to host our courses and module topics to make them widely available. We produced 18 modules, including 80 video tutorials. See more at http://learninghub.soihub.org.

• We developed an online (MOOC) version of the Introduction to Science of Information. Beta testing with a small cohort is taking place this winter. Courses will be offered for enrollment, starting in spring semester 2014, using the new Learning HUB.

• Two faculty workshops on teaching the science of information resulted in 52 attendees. Eight faculty from this group were selected to develop and teach 5 new SoI courses and 19 topic modules, all of which will populate the new Learning HUB.

• The total views of video tutorial and seminar content quadrupled when compared to last year, with 12,746 views in period 4 (as of Sept. 17, 2013), compared with 3,185 views in period 3.

• Sixty of our students and postdocs reported publishing 149 papers and 131 posters and oral presentations. Our students were recognized with 27 awards for these efforts.

• For the fourth straight period, we recorded an increase in the number of full-time student participants involved in Center work—138 undergrads, grads, and postdocs; 106 received financial support from the Center during this period.

Ninety-seven percent of students and postdocs polled say their participation in the Center has been valuable to them because the Center stimulates their own thinking about the science of information, provides ample networking opportunities with leading scholars in the field, and fosters collaborations among students and faculty. Two example testimonials are below.

“The Center has provided a good platform for looking at my research from a broader perspective. It has also provided easy access to faculty and students working in areas related to mine, and their expertise has been very helpful. And finally, I enjoy the opportunity to mentor undergraduates.”

“The CSoI seminars (including the weekly brown bags) are instrumental in providing novel perspectives and inputs. I believe that my experience with the CSoI will help me gain an edge in my future career.”

Highlighted Activities

Student Research Teams. Through a focused research workshop for students and postdocs last year and a subsequent follow-up workshop for students during period 4, the education program formed four multidisciplinary student research teams to work on projects that span all three primary thrusts of the Center. These teams gained experience in grant writing and reporting, creating interdisciplinary teams with members working at multiple institutions, and defining research questions in the emerging research areas of the science of information. The one-year working period for these teams has thus far resulted in 6 journal publications and 18 poster and oral presentations at conferences.

Summer School. The Center’s annual summer school is a signature program of the Center. This intensive short course helps orient students early in their PhD program in the science of information.
The school grew from 24 students representing 10 universities in 2011 to 43 students representing 12 universities in 2012 to 140 students, postdocs, and faculty attending in 2013. This year the Center organized and hosted the North American Information Theory Society School, offering a unique opportunity to broaden the impact of the Center. One hundred two external students and faculty joined 38 Center students and faculty at the summer school. Students deepened their understanding of how the science of information supports cutting-edge research in the life sciences, communication, and knowledge management. Students reported gaining insights into methods and areas of study that enhanced their understanding of the emerging field of the science of information and especially reported making stronger connections with their peers and faculty.

**SoI Learning HUB and Courses.** A robust, interactive online learning platform, called the SoI Learning HUB, is in the final stages of development this period. A massive open online course (MOOC) version of the introduction to science of information course, along with two short courses on probability theory and mathematical theory of communication, populate the Learning HUB currently. The beta version is being tested with a cohort of undergrads, with plans to offer the courses worldwide for spring semester 2014. Efforts are already underway to significantly broaden the offerings in the Learning HUB.

We held two focused trainings for faculty on the topic of teaching the science of information this period. These trainings have resulted in concrete plans for five faculty members to develop and teach new undergraduate and advanced courses at their respective universities, and eight faculty members at Center partners and external universities will develop topic modules that they will use in both physical courses and through the Learning HUB.

**Diversity**

Diversity efforts have continued to grow with the progressive development of the Center-wide CSoI Channels Scholars and Fellowship Programs and with the strengthening of partnerships with other minority STEM programs at partner institutions. The CSoI Undergraduate Channels Scholars program ([http://soihub.org/scholars](http://soihub.org/scholars)), which is a distributed research experience for undergraduates (DREU), targets underrepresented STEM students and provides a supportive cohort experience with a strong professional development component. The first cohort of six students successfully completed their term for the 2012–13 academic year—with two students continuing in the program for the 2013–14 year, three students going to graduate school, and one seeking a job after graduation. A female graduate student was also funded during the 2012–13 year as part of the Channels Program, and is continuing for the 2013–14 year.

Five undergraduate students (2 are females, and 4 of the 5 are from underrepresented groups) were selected as the 2013–14 CSoI Undergraduate Channels Scholars cohort. They hail from Purdue, UCSD, and Howard. A site REU proposal for a distributed summer program focusing on Science of Information was submitted to NSF in August 2013 in order to expand and broaden the CSoI Channels Scholars Program. Bryn Mawr will continue their own Fellowship Program with new female undergrad fellows.

A database in MS Access has been maintained to track our diversity efforts. Our present overall diversity within the Center is outlined in the table below. With the various Center efforts continually expanding and taking root, we anticipate seeing a gradual increase in the number of underrepresented individuals as the Center matures.
Other efforts related to increasing diversity have continued within the Center, such as:

- Growing efforts toward recruiting a diverse cohort of students for the CSol Summer School and Channels Scholars Program; for example, CSol exhibited at the 2012 Society for the Advancement of Chicanos and Native Americans in Science national conference and listed opportunities on the SACNAS website. CSol also sent a representative to the Grace Hopper Celebration and contacted the computer science department chairs at historically black colleges and universities.

- Diversifying the CSol Prestige Lecture series, which brings high-profile researchers to the Center for talks that engage Center members and the broader SoI community. For example, invited speakers have included a female engineering and computer science faculty member (Nancy Lynch, MIT), and an African American male MD/PhD (Emery Brown, Harvard/MIT).

- Strengthening and forging new relationships with other minority-serving STEM programs. Specifically, CSol has cultivated a strong collaborative relationship with the Center for Inclusion, Diversity, Excellence, and Advancement (IDEA Student Center) at the Jacobs School of Engineering at UCSD. This partnership has led to additional students being involved with CSol programs, and the IDEA Center helps to coordinate summer REU internship activities at UCSD. We hope to add new and leveraged partnerships such as this for other future diversity-related activities.

- Providing word-of-mouth highlights about the Center’s various educational and research opportunities through the CSol Channels Scholars by participating in workshops and conferences that target underrepresented audiences, such as the Grace Hopper Women in Computing Conference and the SACNAS National Conference.

Finally, CSol faculty have been directly active in various diversity outreach efforts as well. A few examples include:

- N. Lynch (MIT) is a fairly frequent lecturer at various institutions and often meets with women students’ groups during her visits. For example, this year she gave a distinguished lecture at Carnegie-Mellon’s computer science department and spent several days meeting with CMU students, faculty, and postdocs. One of these meetings was with the women’s student group, and she answered many questions related to being a woman faculty member or a woman research leader in computer science.

- J. Neville (Purdue) is mentoring a female student from the Bronx High School of Science on her science-fair entry for the Intel Science Talent Search; and

- T. Coleman (UCSD) was a co-convener with W. Massey for the 2013 Conference for African-American Researchers in the Mathematical Sciences, which was held in July at UCSD.

**Broader Impact (Knowledge Transfer)**

The Center’s knowledge-transfer activities during the current period were wide-ranging and robust. These activities are all geared toward meeting the Center’s knowledge-transfer mission of developing
effective mechanisms for interactions between CSOl and external stakeholders, supporting the exchange of knowledge and data, and encouraging the application of new technologies. The Center achieves this predominantly through industry relations, international exchange, the broader scientific community, and educational initiatives such as development of the science of information course. These activities not only allow the Center to share its research with other entities, but also allow it to receive critical input and guidance relative to its overall agenda.

In 2013 CSOl took the advice of last year’s NSF site team and formed an industry advisory committee to advise and assist with industry engagement with additional input from the Center’s External Advisory Board. Now the Center has a clear concept for enlisting industrial participation and input. The Center included industry participation in its Big Data Workshop, and also held a full-day research workshop with representatives from Bell Labs that concentrated on science of information research and its grand challenges. Progress was made in garnering industrial support of Center-supported research and the addition of collaborative work between the Center and industry, as our faculty members have solid industry relationships and are routinely collaborating with many companies. Our students are in high demand from industry for both internships and permanent positions. Center faculty have had great success in leveraging their Center-supported research from companies, such as Google, IBM, Yahoo, Symantec, Microsoft, Intel, National Instruments, Huawei Technologies, NEC, Samsung, and Xerox, Support of these projects totals over $1 million in industrial support.

Center members continue to publish prodigiously in scientific journals and conferences about Center-sponsored research and challenges. They have led discussions about Center research and activities during numerous invited talks and research seminars. CSOl PhD students and postdocs again received much recognition, and several have joined prestigious universities as faculty members.

Our international partnerships formed in year 2 were active in this past year. Faculty exchanges took place with the Paris-based Laboratory for Information, Network and Communication Sciences (LINCS), as CSOl sent faculty to LINCS and hosted a LINCS member. We have formalized an agreement with the EAFIT University in Columbia, wherein that institution will assist us in further developing our science of information course for a broader audience, including translation of course content into Spanish. This initiative clearly has very positive implications for our long-term goal of impacting other participating communities throughout Latin America.

The full development of our Science of Information Community of Practice Learning Platform now offers itself as an excellent method of outreach and knowledge transfer as we ramp up collaborations among other higher education institutions to further develop teaching modules and teach the course. The CSol website continues to highlight Center activities and draws more and more users to our offerings. In fact, use of the Center’s online video tutorial and seminar content, which includes our graduate student research series and Prestige Lecture series quadrupled, going from 3,185 views in period 3 to over 13,000 in period 4—pointing to dramatic growth of our dissemination efforts.

Lastly, the Center has acted to ensure our presence on partner campuses: the 2012 summer school was held on Stanford’s campus; the 2014 summer school will be held at UC San Diego; our Bell Labs research and Grand Challenges Workshops were held at Princeton University; Bryn Mawr College has held multiple events sponsored by the Center; UC Berkeley and Stanford held a day-long workshop on Berkeley’s campus, which was entirely sponsored by the Center, called “New Directions in the Science of Information.” Similar events are planned at Texas A&M, MIT, and Howard in the coming spring.
II. Research

Goals, Metrics and Progress

1a. Overall Research Goals

Describe the Center's overall research goals/objectives. If the Center's overall research goals/objectives changed from the previous year, how did they change and why? [In section 2a below, please describe progress the Center has made toward reaching these goals/objectives.]

The Center's overall research goal is to advance science and technology through a new quantitative understanding of the representation, communication, and processing of information in biological, physical, social and engineered systems.

The decades since Shannon's original theory have seen orders of magnitude increase in our ability to collect and act on data. Information theory has enjoyed great success in establishing fundamental limits for problems related to relatively simple classes of random processes. Specifically, traditional analysis considers memoryless processes and, to a lesser extent, stationary ergodic sequences. The most celebrated of these successes include the channel and source coding theorems. However, many real-world domains exhibit complexities that violate the assumptions used to derive these fundamental results. In the new area of "big data," many domains consist of data with one or more of the following characteristics: very large scale (with possibly distributed storage), high dimensional, multi-scale patterns, dynamic, temporal, and/or geographic effects, heterogeneous structure, complex dependencies, and noisy records.

Perhaps the greatest challenge in applying information-theoretic principles to a broader suite of problems - including biological systems, analytics for massive datasets, and social networks - is that of developing meaningful notions of "structural, temporal, spatial, and semantic information" and establishing a set of corresponding fundamental results. Furthermore, nowadays the most common actions on datasets (e.g., at a large scale, such as Google's repositories) are search operations. Often such actions are performed on compressed (sketchy) data. The traditional goal of information theory, that of reliably reproducing data, must be replaced by reliably answering queries, often based on incomplete (e.g., compressed) information stored in distributed databases.

In view of these, traditional formalisms associated with information do not adequately address some key aspects of information. By incorporating these new elements of information into an integrated framework, the center fundamentally enhances the scope, scale, and effectiveness of applications on data.
**Structure**

Measures are needed for quantifying information embodied in geometric structures and networks (e.g., information in nanostructures, biochemical networks, social networks, networks of financial transactions). Often, these measures must account for associated context; and incorporate diverse (physical, social, economic) models of flow (dynamics).

**Time**

Timely delivery of partial information often carries a higher premium than delayed delivery of complete information (e.g., real-time control systems, decision processes in financial markets). The notion of timeliness, however, is closely related to the semantics (is the signal critical), system state (is the system under stress), and the sender/receiver.

**Space**

In interacting systems, spatial localization often impacts information transfer (e.g., interference rates can be significantly reduced in wireless networks by localizing communication, cells perform vital functions by localizing chemical reactions to certain parts of the cell).

**Information and Control**

In addition to delay-bandwidth tradeoffs, systems often allow modifications to underlying design patterns (e.g., network topology, power distribution, and routing). Information is exchanged in space and time for decision making. Thus, timeliness of information delivery along with reliability and complexity constitute basic objectives.

**Semantics**

In many scientific contexts, one is interested in the role of signals, in the absence of precise knowledge of their semantics, (e.g., DNA sequences, spike trains between neurons, whale songs). These signals are known to convey information, but little more than that can be assumed a priori. Is there a general way to account for the meaning of signals in a given context?

**Dynamic Information**

In a complex network, information is not just communicated, but also processed and generated (e.g., emergency response, stimuli). How can such considerations of dynamic sources be incorporated into an information-theoretic model?

**Learnable Information**

Data-driven science has received considerable recent research attention. How much information can actually be extracted from a given data repository? Is there a general theory that provides natural model classes for (more structured) data? What is the cost of learning the model, and how
does it compare to the cost of actually describing the data? Is there a scientific way to approach the problem of extracting relevant information?

**Resource Constraints**

Our ability to communicate and manipulate information is often limited by available resources (e.g., computing devices, bandwidth of signaling channels). How much information can be extracted and processed with limited resources? This question relates to complexity and information, where different representations of the same data may vary dramatically when complexity is taken into account.

Representation-invariant Information: How does one conclude whether two different data representations are information equivalent?

**Cooperation**

Often subsystems may be in conflict (e.g., the problem of Byzantine generals, denial of service or selfish attacks in computer systems) or in collusion (e.g., price fixing, insider trades). How does cooperation/collusion impact information?

As we discuss in the rest of this report, we have made significant progress towards incorporating these elements into models and methods and applied them to diverse applications with considerable success.

**Mission for Integrated Research**

To create a shared intellectual space, integral to the Center’s activities, providing a collaborative research environment that crosses disciplinary and institutional boundaries.

Creating a shared intellectual space that facilitates discussions, interactions, and problem formulation across disciplinary and institutional boundaries is essential to the Center’s success. Creation of this space requires multiple components, including informing PIs about the disciplinary and interdisciplinary expertise and interests of other Center PIs, creating and disseminating pedagogical resources such as tutorial documents and seminars; creating a virtual collaborative workspace via the Center’s website, and proactive planning and execution of small workshops and brainstorming sessions. These sessions are core facilitators for formulating a small number of high-impact research problems for interdisciplinary teams to coalesce around. We have two main mechanisms based on which Center workshops are organized: Regional and Opportunistic. We also have annual meetings to share results and progress and to create further opportunities for cross-disciplinary advancement.

**Regional Workshops**

The institutions associated with the Center coalesce in three main geographic regions: East Coast, Midwest, and West Coast. Each region, via the executive committee members and research thrust leaders at each institution, conducts one to two half or full-day workshops per year around specific
interdisciplinary topics. PIs from outside the region are also invited to attend. A part of the Center website is dedicated to these workshops, which can be used for both planning as well as reporting the outcomes, in particular any problem formulations that were discussed during the meeting. Additional input on the workshop outcome is solicited from all Center participants. Annual PI meetings also provide an opportunity for discussing these workshop outcomes. PIs are expected to participate in at least one of these regional workshops per year.

**Opportunistic Workshops**

Many Center PIs attend the same international meetings, which is an opportunity to carve out some time during these meetings for brainstorming sessions or workshops associated with the Center. If the meeting takes place close to one of the Center institutions, then PIs from that Institution are invited to attend. The agenda for these brainstorming sessions or workshops is set by the participants a few weeks in advance and shared with everyone via the Center website, where the workshop outcome and problem formulations that result are reported. The first of these opportunistic workshops was held in Sept. 2010 at UIUC, and the outcome was documented and also incorporated into the Strategic Plan. Workshops during this past three years were held at Stanford, UCSD, Illinois, MIT, Princeton, and Purdue, as well as at several international conferences. In 2013 we organized the Big Data Workshop in March 2013, and conducted two industrial meetings in Princeton and Chicago.

**Other Activities**

Other activities such as exchange visits among Center faculty, joint student supervision, and a systematic dissemination of cross-disciplinary solution techniques are being conducted. We believe that the latter, along with feedback, is at least as important as the tools and skill sets we may develop in the process of crossing barriers across disciplines. The dissemination process is facilitated by video records of presentations, a thesaurus and/or supplements (e.g., data sets) to foster model learning, a wiki space for questions and answers, and more standard publication and presentation methodologies. The process of building and disseminating such exemplars is meant to accelerate cross-disciplinary activity and thus enhance our integrative research through a learn-by-example approach.

We strongly believe in the Integrated Center being significantly more than the sum of its parts (researchers and research groups). To this end, we have made strong efforts and have achieved considerable success. Several new collaborations were established during the first year. To mention a few: Ramkrishna led a group of investigators from Purdue, Berkeley and UCSD to quantify fluxes in biological networks using information theoretic formalisms. During the UCSD workshop in February 2011, Bin Yu and Shankar Subramaniam initiated a joint research program between Berkeley and UCSD on statistical analysis of biological networks. To enhance these collaborations we created three center-wide postdocs: Tom Courtade works with Tsachy Weissman (Stanford) and Sergio Verdu (Princeton) while Rui Ma collaborates with Todd Colleman and Shankar Subramaniam (UCSD) and Venkat Anantharam (Berkeley). The third postdoc Z. Wang was added in 2013 and spends time
between Stanford, UIUC and MIT working with Tsachy Weissman, Olgica Milenkovic, and Nancy Lynch on compression of biological datasets. Furthermore, David Tse and his students initiated a collaboration with Madhu Sudan (MIT), while P.R. Kumar (Texas A&M) and Nancy Lynch (MIT) started a new collaboration on foundation of security.

The Center members also reach out outside the core Center. In 2013 we added three new collaborators: Y. Baryshnikov (UIUC), T. Imielinski (Rutgers) and M. Raginsky (UIUC). We plan to continue extending our collaboration and add new members to the Center in the research areas important to carry center’s mission.

1b. Performance and Management Metrics to Assess Progress

1b. Inform us of the performance and management indicators/metrics the Center has developed to assess progress in meeting its research goals/objectives, if changed from the previous reporting period.

The following are metrics for our integrative research efforts as developed for the strategic plan:

- Formulate a small number of high-impact research problems for interdisciplinary team to coalesce around in 2 years.
  - Grand Challenge workshops accomplished this.

- Identify two grand challenge problems in two years on a web bulletin board.
  - We identified several challenges during our Workshops in March 2012 and 2013.

- Conduct five investigator exchange visits for immersive activity.
  - Through our center post-doc program and opportunistic workshops we have an intensive exchange program. We added another center post-doc in 2013.

- Initiate five new collaborations through joint supervision, student exchange, joint publication, and presentations.
  - Student workshop that took place at Purdue and center post-docs are part of this initiative. We support grad student travel for students to continue these new collaborations.

- Develop two pedagogical resources (e.g. books, survey, papers, lecture series) at the interface of applications and theory in two years.
  - As mentioned above we initiated an “Introduction to Science of Information” course and have also developed online science of information educational modules.

- Develop proposals for sustained external funding within five years.
  - We submitted several proposals to NSF and AFOSR.
1c. Problems Encountered

Discuss any problems you have encountered in making progress toward the Center’s research goals/objectives during the reporting period as well as any problems anticipated in the next period. Include your plans for addressing these problems.

N/A
2a. Research Thrust Areas’ Accomplishments

Briefly describe the research thrust areas at the Center. Please provide basic information for each thrust area and details of significant accomplishments during the reporting period, including any research partnerships and their contributions to the Center (do not include publications, presentations, etc., that are reported in Section VIII, Center-wide Outputs and Issues). Include in the narrative a discussion of the goals, activities, and outcomes and/or impacts in the current reporting period, if changed from the previous reporting period. Be sure to discuss how the activities in the various research thrust areas enable the Center to meet its goals/objectives described above.

Life Sciences Thrust

We describe activities in various groups within the Center in the Life Sciences Thrust. These activities focus on aspects of space, time, interactions (networks), and control, as they relate to analytical formalisms in life sciences. These activities have engendered strong collaboration across groups and address fundamental problems in biological systems, scaling from individual molecules to entire ecosystems. Many high-profile publications and results are attributed to the Center.

We present, in order, results on analyses of biological sequences; high-dimensional vector data (gene expression, genome-wide associations); biological networks, including inference, functional characterization, phenotypic association, and network dynamics; neuronal sensing, modeling, and control, including spatiotemporal information and networks; and finally, structural information (cell structure, structure–phenotype correlation).

Information-Theoretic Approaches to Biological Sequences

Analysis of biological sequences has been a topic of active investigation for decades. Problems, such as matching, alignment, assembly, variant calling, motif finding, etc., have been well-studied, and efficient software tools have been developed. However, little is known about the fundamental complexity of these problems, the limits of algorithmic performance, and the efficiency of the device–software combination. These fundamental challenges drive both algorithm development, as well as ongoing improvements to experimental methodologies. Over the past year, Center researchers have made significant progress on these problems. We report on the major results here.

Optimal Shotgun Assembly for DNA Sequencing

A basic algorithmic problem in computational biology is the assembly of many hundreds of millions of short randomly sampled reads to reconstruct the original DNA sequences. The two key challenges in shotgun assembly are the long repeats in DNA sequences and errors in reads. What are the fundamental limits imposed on the performance of assembly algorithms due to long repeats and read errors? Which algorithms can approach the fundamental limits? Building on insights obtained
in previous years of the project, we have obtained a near-complete answer to these questions. Our current work aims to translate these theoretical results into improvements in state-of-the-art assemblers and the design of new assemblers, specifically for PacBio sequencers.

**Handling Long Repeats in DNA sequences**

As reported in Tse (2013a), we have developed a computational pipeline, which, given the empirical repeat statistics of any genome, indicates read lengths and coverage depth (number of reads) for which complete reconstruction of the genome is possible (with target success probability). Figure 1 shows such a plot for the repeat statistics of human chromosome 19.

![Figure 1](image)

*Figure 1.* The x-axis is the required read-length $L$ and the y-axis is the required number of reads as a ratio to the number of reads needed to cover the DNA sequence, a well-accepted baseline (Lander and Waterman) and a lower bound on the number of reads needed to reconstruct the sequence. The black curve is the lower bound on the read length and the coverage depth that holds for any assembly algorithm. The colored curves correspond to the performance of specific assembly algorithms.

There are two main uses of this framework:

1. It allows the comparison of different assembly algorithms in terms of the data requirements.
2. It provides insights for developing assembly algorithms, which we call MultiBridging.

MultiBridging can achieve performance close to the lower bound for empirical repeat statistics of a wide range of genomes.
Handling Read Errors

All sequencing technologies have a certain amount of errors in their reads, some more than others. For example, Illumina reads typically have about 1% substitution errors, while PacBio reads typically have 10% to 15% indel errors. What is the effect of errors on the read length and coverage depth requirements for assembly? In an earlier work (Tse 2012a), we have designed a modification of the greedy algorithm that can handle read errors. We have also evaluated the data requirements of this algorithm. One observation is that read errors always degrade the performance of this modified greedy algorithm in the sense that the minimum required read length for correct assembly also increases with the noise level. The question we pose: Is this degradation in performance fundamental? Or is it a consequence of the specific assembly algorithm used?

In Tse (2013b), we answered this question in the context of a simple i.i.d. genome model and substitution errors. We showed a surprising result—a threshold on the read noise level always exists, such that, below this threshold, read errors do not impact performance. The minimum read length and the minimum coverage depth required is the same asymptotically as in the case when there are no read errors. For a symmetric error channel, the threshold is 19%! That is, as long as the fraction of errors is less than 19%, there is no degradation in performance asymptotically in the limit of long genome and large number of reads. Moreover, we showed that a separation architecture is optimal when the noise level is less than the threshold; an error-correction phase and an assembly phase based on error-corrected phase is optimal.

De Novo RNA Sequencing

We generalize the information-theoretic formalism from DNA sequencing to RNA sequencing. High-throughput sequencing of RNA transcripts (called RNA-Seq) has emerged in the last few years as a powerful method that enables discovery of novel transcripts and alternatively spliced isoforms of genes, along with accurate estimates of gene-expression levels. In Tse (2013c), we study the fundamental limits of de novo transcriptome assembly using RNA shotgun sequencing. We propose a new polynomial-time algorithm for transcriptome assembly and derive sufficient conditions on the length of reads under which the algorithm will succeed. We then compare them with necessary conditions, which we derive, for reconstruction by any algorithm, and show that the proposed algorithm is near-optimal on a real data set. Along the way, we study the problem of deducing end-to-end network flows using link-level observations—and prove new results for this model.

We are currently working on converting various theoretical results we obtained on the DNA and RNA sequencing problems into computational tools that are useful for biologists. The specific tasks include the following:

Assembly Using PacBio Reads

Our theoretical result (Tse 2013b) suggests that performance of assemblers can be made very robust to read errors. Yet, read errors are a significant concern in state-of-the-art assemblers. We plan to use our theoretical insights to improve the capabilities of assemblers in dealing with read errors. In particular, we would like to focus on PacBio reads where the error rate is very high.
Empirical observations of performance of a state-of-the-art assembler for PacBio reads lead us to believe that, by adding a properly designed error-correction phase, genomes of much longer repeat lengths can be assembled (Tse 2013d).

**Estimation of Long Repeat Lengths**
Our result (Tse 2013a) says that the minimum read length required to assemble a genome depends on the tail end of the repeat statistics, i.e., the length of the longest repeats. However, the repeat statistics are not known a priori. This leads naturally to the question: How can one estimate the repeat statistics, especially the tail end, from the read data to determine whether a genome can be assembled with confidence? Tse’s group has started collaborating with Asif Khalak at Pacific Biosciences on this problem.

**De Novo RNA-Seq Assembly Tool**
Building on the theoretical results we obtained, we plan to build a scalable tool for de novo RNA–Seq assembly. Unlike DNA assembly, there are very few de novo assemblers for RNA-Seq data, and their performance is not very good. We expect that this is an opportunity for our theory to make significant impact.

**SNP and Structural Variant Detection**
We plan to apply our insights on repeats to the problem of detecting SNPs (single nucleotide polymorphisms). In particular, we would like to focus on detecting SNPs in repeat and paralogous gene regions. Repeat regions make read alignment difficult, and hence, SNP calling is difficult. We believe that, by combing RNA and DNA read data, we may find enough “diversity” against such repetitions.

**Compression and Storage of Genomic Sequences**
While storage capacity has increased significantly in the recent past, DNA-sequencing technology is developing and producing amounts of data, growing at rates that are surpassing that of storage capacity. The compression problem is important and timely, as the bottleneck is quickly moving from the acquisition to the storage and processing of the data. The problem is challenging at multiple levels—understanding the mechanisms underlying the generation of sequencing data, characterizing the fundamental limits on compression corresponding to such mechanisms, devising algorithms that approach these performance limits, and experimenting with them.

**Metagenome Read Compression**
We have been working on the development and application of new data-compression algorithms tailored to genomic data. An important use case is the problem of lossless compression of DNA of one individual in a population, given that the DNA of another individual is already available in the database (and thus available to the compressor and decompressor). This problem is becoming increasingly important, as we move into personalized genome-based medicine, where efficient downloading, copying, sharing, and processing of an individual genome are crucial. Weissman and collaborators have recently introduced a compressor that improves, significantly, on the current
state of the art. Further, unlike current techniques, our algorithm is guided not only by an understanding of the biological aspects of the data, but also the information-theoretic principles of universality and optimality.

When compressing reads, it is often necessary to simultaneously store “quality values” reported by most sequencers. Quality values (or q-factors) are also important when sequencing the genomes of a whole collection of species inhabiting a given ecosystem, as is done in metagenomics. We have developed a scheme for lossy compression of the quality values that results in significant compression gains relative to the lossless case, with little to no measurable compromise in performance of the downstream applications that use these quality values.

Milenkovic and collaborators have developed the first known approach to metagenome read compression, termed MCUIUC (metagenomic compression at the University of Illinois at Urbana–Champaign). The core of the proposed algorithm is in performing classification of reads based on unique organism identifiers, followed by reference-based alignment of reads for individually identified organisms and metagenomic assembly of unclassified reads. Once assembly and classification are completed, lossless reference-based compression is performed via a combination of difference coding, positional encoding via Golomb codes, and Lempel–Ziv compression. We evaluated our schemes on synthetic metagenomic samples containing randomly selected organisms, and are in the process of extending our studies to human-gut microbial screening.

**Gene Prioritization**

Gene prioritization is a class of methods aimed at discovering genes implicated in the onset and progression of a disease. As candidate genes are ranked based on similarity to known genes according to a set of criteria, the overall aggregation of these ranked data sets is a vital step of the prioritization procedure. Aggregation of different lists of ordered genes is accomplished either via classical-order statistical analysis or via combinatorial ordinal data fusion.

We have developed a new approach to combinatorial gene prioritization via linear programming optimization and used our newly developed weighted Kendall tau rank distance to assess similarities between rankings. The weighted Kendall tau distance allows for constructing aggregates that have higher accuracy at the top of the ranking, usually tested experimentally. Our method can also accommodate ties in rankings and handle negative outliers. We tested the performance of the prioritization method on a set of test genes pertaining to the Bardet–Biedl syndrome, schizophrenia, and human immunodeficiency virus (HIV). Also, for the purpose of disease gene discovery, we tested a number of genes not known to be implicated in glioblastoma, but suspected to be involved in neurofibromatosis, Turcot and Li–Fraumeni syndromes. Our specific goal is to identify unknown genetic pathways mutated in glioblastoma cells.

**Analytical Approaches to High-Dimensional Sparse Data**

One of the major challenges associated with different modalities of data in life sciences is its high dimensionality, sparsity, and need for statistical modeling. For instance, in correlation studies of
genomic variants (single nucleotide polymorphisms (SNPs, copy number variation, etc.) and phenotype (such higher-order as disease or such lower-order as microscopy), we may have a small number of samples (longitudinal patient data), a very large number of variants (millions of SNPs), and high noise (background variations). In such cases, analytical approaches are particularly difficult because of the lack of statistical power. The Center has made a number of important advances in solving these problems. We describe our major results in the past year.

**Sparsity Meets Structure and Nonlinearity**

For early diagnosis and effective treatment of Alzheimer’s disease (AD), it is of great importance to identify genetic variations and other biomarkers (e.g., imaging features extracted from MRI data) associated with AD. Researchers can use classical sparse learning methods, such as LASSO, elastic net, or sparse canonical correlation analysis, but these methods have several critical limitations. First, they do not model interactions between the variables, making the results difficult to interpret biologically. Second, they ignore invaluable structural information from prior knowledge (e.g., pathways). Third, they model linear relationships, despite the fact that any real complex system (e.g., a biological system) is highly nonlinear. To overcome these limitations, we have developed a series of novel Bayesian approaches.

**Selection of Correlated Variables**

Our key insight is to explore the correlation information between variables (i.e., biomarkers) embedded in the eigenstructure of the unlabeled data. We materialized this idea within a hybrid Bayesian model, called EigenNet, which constrains a sparse classifier in an eigensubspace of the data (Qi and Yan 2011). When the biomarkers are independent of each other, EigenNet is similar to LASSO; when the biomarkers are correlated, unlike LASSO and elastic net, EigenNet aligns the classifier with selected eigenvectors and chooses correlated biomarkers. We applied EigenNet to two applications of imaging genetics—selecting correlated SNPs that differentiate a healthy subject from an AD patient and predicting cognitive test scores based on interdependent brain-imaging features. For both tasks, EigenNet significantly outperformed LASSO and elastic net.

![Figure 2. Finding key associations between imaging markers and genetic variations helps to understand the genetic basis of Alzheimer’s disease.](image)

Biological pathways provide invaluable information for underlying biological mechanisms. Unfortunately, this information has not been utilized by most sparse learning methods. To address this issue, we have extended EigenNet to use regulatory and metabolic pathways extracted from certain pathway databases to guide the selection of genes associated with breast cancers. Our method has achieved significantly higher prediction accuracy compared to state-of-the-art sparse
learning methods. Collaborating with Eli Lilly researchers, we are applying this method to select biologically meaningful biomarkers for new drug targets.

**Multiview Learning**

We extend this line of work to multiview data analysis in which the task is to identify interactions between high-dimensional data sources, such as genetic variations and imaging features. Popular sparse canonical correlation analysis does not explicitly incorporate in their analysis known structural information (e.g., linkage disequilibrium structures among SNPs). To improve upon these methods, we incorporate the structural information as prior distributions and develop a supervised sparse Bayesian multiview model that uses disease-status information to guide the discovery of interactions (Zhe et al. 2013b). Collaborating with researchers from Eli Lilly and Indiana University Medical School, we use these models to elucidate the complex relationships among genetic variations, brain structures, and AD progression.

**Algorithms for Gene Expression**

Microarrays are used to determine gene expression (measurement of expressed genes) in tissue samples. For example, the genes expressed in a particular cancer-cell tissue can help predict the survival for that particular type of cancer. A microarray consists of chemical probes to which the chemicals in the genes attach, and we can use the pattern of attachment to discern the gene concentrations. However, many types of genes are in any given cell sample, some whose chemical composition is unknown, thereby complicating the mapping of microarray measurements to gene concentrations.

![Image](image.png)

**Figure 3.** Data analysis using hyperspectroscopy, a signal-processing technique, yields significantly better accuracy in gene expression from microarrays than existing techniques.

The process of mapping microarray data to expressed genes is shown in Figure 3. Goldsmith and collaborators have determined that this mapping problem is very similar to hyperspectroscopy in signal processing, and have proposed a method for gene expression based on this technique. In contrast to current methods, which rely on prior knowledge of the cell-type composition or signatures that are not available in most public data sets, our method identifies the cell-type composition, signatures, and proportions per sample without the need for information. The method
has been successfully tested on controlled and semicontrolled data a priori sets and performed as accurately as current methods that require additional information. As such, this method enables the analysis of cell-type specific gene expression, using existing large pools of publicly available microarray data sets.

We have also investigated the use of directed information to infer synapse connections between neurons. We propose a method that uses the delay profile of the causally conditioned entropy rate to find a measured delay range and its intersection with a connection-based delay range. Through simulations in Neuron, we show how this method outperforms connectivity inference based on directed information alone.

![Figure 4. Predicted delay range (shaded area) versus simulated delay range (dashed lines); when these ranges overlap, we infer a connection.](image)

**Information-Theoretic Methods for Biological Networks**

The Center has made significant contributions to analytical methods for modeling and analysis of biological networks. We have developed general methods that researchers can apply to broad classes of networks, as well as specific methods targeted to application contexts. A key element of our methods is deep reliance on statistical and analytical methods to quantify (statistical) significance of results. This is particularly challenging for network structures that do not draw from well-studied random models, as is the case in biological networks.

**Community Detection in Networks**

Community detection is a key problem of interest in network analysis. We present a novel framework for community detection that is motivated by a physical-system analogy. We model a network as a system of point masses and drive the process of community detection by leveraging the Newtonian interactions between the point masses. We designed our framework to be generic and extensible relative to the model parameters that are most suited for the problem domain. We illustrate the applicability of our approach by applying the Newtonian Community Detection algorithm on protein–protein interaction networks of *E. coli*, *C. elegans*, and *S. cerevisiae*. We also present a detailed analysis of the structural properties of the communities produced by our proposed algorithm, together with a biological interpretation that uses *E. coli* protein network as a
case study. We constructed a functional enrichment heat map with the Gene Ontology (GO) functional mapping, in addition to a pathway analysis for each community. Our analysis illustrates that the proposed algorithm elicits communities that are not only meaningful from a topological standpoint, but also possess biological relevance. We believe that our algorithm has the potential to serve as a key computational tool for driving therapeutic applications involving targeted drug development for personalized care delivery.

Duchenne muscular dystrophy (DMD) is an important pathology associated with the human skeletal muscle and has been studied extensively. Gene-expression measurements on skeletal muscle of patients afflicted with DMD provide the opportunity to understand the underlying mechanisms that lead to the pathology. Community-structure analysis is a useful computational technique for understanding and modeling genetic-interaction networks. We leverage this technique in combination with gene-expression measurements from normal and DMD-patient skeletal-muscle tissue to study the structure of genetic interactions in the context of DMD. We define a novel framework for transforming a raw data set of gene-expression measurements into an interaction network, and subsequently apply algorithms for community-structure analysis for the extraction of topological communities. We analyze the emergent communities from a biological standpoint in terms of their constituent biological pathways, and we present an interpretation that draws correlations between functional and structural organization of the genetic interactions.

We also compare these communities and associated functions in pathology against those in normal human skeletal muscle. In particular, we observe differential enhancements in the following pathways between pathological and normal cases: metabolic, focal adhesion, regulation of actin cytoskeleton and cell adhesion, and the implication of these mechanisms are supported by prior work. Furthermore, our study also includes a gene-level analysis to identify genes that are involved in the coupling between the pathways of interest. We believe that our results serve to highlight important distinguishing features in the structural–functional organization of constituent biological pathways, as it relates to normal and DMD cases, and provide the mechanistic basis for further biological investigations into specific pathways differently regulated between normal and DMD patients. Subramaniam and collaborators believe these findings have the potential to serve as fertile ground for therapeutic applications involving targeted drug development for DMD.

**Modularity in Networks**

Many recent studies have investigated modularity in biological networks and its role in functional and structural characterization of constituent biomolecules. A technique that has shown considerable promise in the domain of modularity detection is the Newman and Girvan (NG) algorithm, which relies on the number of shortest paths across pairs of vertices in the network traversing a given edge, referred to as the *betweenness* of that edge. The edge with the highest betweenness is iteratively eliminated from the network, with the betweenness of the remaining edges recalculated in every iteration. This generates a complete dendrogram, from which modules are extracted by applying a quality metric called *modularity*, which is denoted by $Q$. This exhaustive computation can be prohibitively expensive for large networks, such as protein–protein interaction
networks. We present a novel optimization to the modularity detection algorithm, in terms of an efficient termination criterion based on a target edge betweenness value, which may be terminated by using the process of iterative edge removal.

We validate the robustness of our approach by applying our algorithm on real-world protein–protein interaction networks of yeast, *C. elegans* and *Drosophila*, and demonstrate that our algorithm consistently has significant computational gains in terms of reduced run time, when compared to the NG algorithm. Furthermore, our algorithm produces modules comparable to those from the NG algorithm, qualitatively and quantitatively. We illustrate this using such comparison metrics as module distribution, module membership cardinality, modularity Q, and the Jaccard similarity coefficient.

**Connectivity, Causality, and Information Flow in Networks**

Researchers can study cellular signaling circuitry in eukaryotes by analyzing the regulation of protein phosphorylation and its impact on downstream mechanisms leading to a phenotype. A primary role of phosphorylation is to act as a switch to turn “on” or “off” a protein activity or a cellular pathway. Specifically, protein phosphorylation is the main step in the process of transducing molecular signals inside the cell. It is important to reconstruct context-specific, signal-transduction mechanisms of the cell, because errors in transferring cellular information can alter the normal function and may lead to such diseases as diabetes, cancer, and autoimmunity.

Recent advances in high-throughput “omics” technologies, combined with advanced data integration, analysis, and interpretation, are facilitating a detailed and quantitative understanding of the underlying biochemical mechanisms. We have developed a novel framework for time-dependent reconstruction of signaling networks involved in activation of macrophage cells leading to an inflammatory response. Several signaling pathways have been identified in macrophage cells, but the time-varying causal relationship that can produce a dynamic directed graph of these molecules has not been explored. We use the notion of Granger causality and apply a vector autoregressive model to the time–course data, made available by the Alliance for Cellular Signaling (AfCS), in RAW 264.7 macrophage cells on phosphoproteins in single- and double-ligand experiments for 22 ligands. Through the reconstruction of the phosphoprotein network, we were able to estimate connectivity, Granger causality, and the dynamics of information flow. We selected significant interactions through statistical hypothesis testing (t-test) of a linear model’s coefficients, and we used them to reconstruct the phosphoprotein-signaling network.

Our time-varying framework results in a succession of the phosphoprotein network in three stages that represent the evolution of the biological subsystem as a time-dependent cascade. The proposed approach was able to capture most of the known signaling interactions, such as protein kinase B (Akt) $\rightarrow$ glycogen synthase kinase 3α/β (GSKα/β), which appears only in the last two stages, and predicts such interactions as p38 $\rightarrow$ RSK that appears only in the first two stages (1 to 4 min. and 4 to 7 min., respectively), GSKα $\rightarrow$ moesin in the middle stage and GSKα $\rightarrow$ ezrin/radixin/moesin
in the third stage (7 to 10 min.). We also noted that the path p38 → GSKα/β only appears in the last stage.

**Pathway Inference**

The initiation of B-cell ligand recognition is a critical step for the generation of an immune response against foreign bodies. A wide variety of responses may be induced in B cells through the activation of different receptors. The regulatory mechanisms that are involved in B-cell response to antigenic stimulants are not very well understood. We sought to identify the biochemical pathways involved in the B-cell ligand recognition cascade and sets of ligands that trigger similar immunological responses. We used several comparative approaches to analyze the gene-coexpression networks generated from a set of microarray experiments spanning 33 different ligands. First, we compared the degree distributions of the generated networks. Second, we used a pairwise network alignment algorithm to align the networks based on the hubs in the networks. Third, we aligned the networks based on a set of KEGG (Kyoto encyclopedia of genes and genomes) pathways. We summarized our results by constructing a consensus hierarchy of pathways that are involved in B-cell ligand recognition. The resulting pathways that are shared across B-cell responses to different ligands were further validated through literature for their common physiological responses (e.g., both PGE and NPY trigger pathways that contribute to inflammation). Collectively, the results based on our comparative analyses of degree distributions, alignment of hubs, and alignment based on KEGG pathways showed a high degree of concordance and (1) provide a basis for molecular characterization of the immune-response states of B cells and (2) demonstrate the power of comparative approaches (e.g., gene-coexpression network-alignment algorithms) in elucidating biochemical pathways involved in complex signaling events in cells.

**Systems Modeling of Lipid Processing**

Lipids play an important role in physiology and pathophysiology of living systems. Until a few decades ago, the number of lipid molecules that were chemically characterized was a few hundred at most and were catalogued in monographs and compendia. The past two decades have witnessed two major advances in lipid biology. In the first, mass spectrometry has enabled the identification of thousands of lipid molecular species from cells and tissues, and this has pointed to the important need for developing a systematic ontology that can rationally name and catalog the molecules. Second, the ability to investigate the functional roles of lipid molecules through systematic phenotypic studies has led to the identification of lipids as extremely important players in physiology and pathophysiology of living species. In combination with proteins and nucleic acids, lipids are integrally involved in biochemical networks that lead to phenotypes, such as homeostasis, differentiation, and cell and tissue death. Any approach to systems characterization of living systems, of necessity, has to include lipids, along with other macromolecules and all complex cellular pathways involving lipid molecular species. Systems biology now extends in its scope to identify biosynthetic and metabolic lipid networks, cellular-signaling networks that explicitly include lipid molecules, and transcriptional and epigenetic networks where lipids play an integral role.
Several large-scale projects to characterize lipids and their functional roles have been initiated, as exemplified by the LIPID MAPS effort. The LIPID MAPS is an exemplar systems biomedical project of the National Institute of General Medical Sciences that measures cell-wide lipid changes in an attempt to reconstruct biochemical pathways associated with lipid processing and signaling. In this review, we have provided a comprehensive summary of extant developments in lipid bioinformatics and systems biology and discuss the outlook for the future integration of lipidomics into cellular and organismic biology studies of mammalian cells. Specifically, we present a detailed review on the development of classification, ontology, nomenclature, and structure representation tools for lipid molecules. Besides the lipidome, we also discuss lipid-related genes and proteins (lipid-genome and lipid-proteome databases). Challenges and approaches for integrated display of lipid-signaling pathways and lipidomic and transcriptomic data are presented. The systems biology aspect of this review deals with statistical analysis, correlations, and the integration of genomic and lipidomic data to study macrophages. We also review approaches and recent advances in quantitative kinetic modeling of lipid metabolism through integration of lipidomics and transcriptomics data.

**Data-Driven Reconstruction of Networks**

Data-driven reconstruction of biological networks is a crucial step toward making sense of large volumes of biological data. Although several methods have been developed recently to reconstruct biological networks, few systematic and comprehensive studies compare different methods in terms of their ability to handle incomplete data sets, high data dimensions, and noisy data. We use experimentally measured and synthetic data sets to compare three popular methods—principal component regression (PCR), linear matrix inequalities (LMI) and least absolute shrinkage and selection operator (LASSO)—in terms of root-mean-squared error (RMSE), average fractional error in the value of the coefficients, accuracy, sensitivity, specificity, and the geometric mean of sensitivity and specificity. This comparison enables us to establish criteria for selection of an appropriate approach for network reconstruction based on a priori properties of experimental data. For instance, although PCR is the fastest method, LASSO and LMI perform better in terms of accuracy, sensitivity, and specificity. Both PCR and LASSO are better than LMI in terms of fractional error in the values of the computed coefficients. Trade-offs such as these suggest that we need to take more than one aspect of each method into account when designing strategies for network reconstruction.

Motivated by these observations, Subramaniam and collaborators have developed a new method, called doubly penalized linear absolute shrinkage and selection operator (DP-LASSO), for reconstruction of dynamic biological networks. In this method, we have implemented principal component analysis as a supervisory-level filter to extract the most informative components of the network from the data set. In the lower-level reconstruction engine, we apply LASSO with extra weights on small parameters derived via partial least squares from the first-layer filter to maintain the principal components and nullify the remaining small coefficients. Simulation results show improvements in accuracy and sensitivity of the algorithm in reconstruction of networks.
Application of the DP-LASSO algorithm to experimental data sets for the cell-division cycle of fission yeast also shows the fidelity of the reconstructed network.

Studies of macrophage biology have been significantly advanced by the availability of such cell lines as RAW 264.7 cells. However, it is unclear how these cell lines differ from primary macrophages, such as thioglycolate-elicited peritoneal macrophages (TGEM). We used the inflammatory stimulus Kdo₂-Lipid A (KLA) to stimulate RAW 264.7 and TGEM cells. We concomitantly measured temporal changes of lipid and gene-expression levels and performed a systems-level analysis on the fold-change data. We presented a comprehensive comparison between the two cell types. Upon KLA treatment, both RAW 264.7 and TGEM cells show a strong inflammatory response. TGEM (primary) cells show a more rapid and intensely inflammatory response relative to RAW 264.7 cells. DNA levels (fold-change relative to control) are reduced in RAW264.7 cells, correlating with greater down-regulation of cell-cycle genes. The transcriptional response suggests that the cholesterol de novo synthesis increases considerably in RAW 264.7 cells, but 25-hyrdoxycholesterol increases considerably in TGEM cells. Overall, while RAW 264.7 cells behave similarly to TGEM in some ways and can be used as a good model for inflammation- and immune-function related kinetic studies, they behave differently from TGEM in other aspects of lipid metabolism and phenotypes used as models for various disorders, such as atherosclerosis.

Subramaniam and collaborators have also developed an information-theoretic approach to reconstruction of phosphoprotein-cytokine networks in RAW 264.7 macrophage cells. Cytokines are secreted upon activation of a wide range of regulatory signals transduced by the phosphoprotein network. Identifying these components can help identify regulatory modules responsible for the inflammatory responses. The information-theoretic approach is based on the estimation of mutual information of interactions by using kernel-density estimators. Mutual information provides a measure of statistical dependencies between interacting components. Using the topology of the network derived, we develop a data-driven parsimonious input–output model of the phosphoprotein-cytokine network. We demonstrated the applicability of our information-theoretic approach to reconstruction of biological networks. For the phosphoprotein-cytokine network, this approach not only captures most of the known signaling components involved in cytokine release but also predicts new signaling components involved in the release of cytokines. The results of this study are important for gaining a clear understanding of macrophage activation during the inflammation process.

**Reactions and Diffusion Processes in Networks**

Many biochemical processes at the subcellular level involve a small number of molecules. The local numbers of these molecules vary in space and time and exhibit random fluctuations that can only be captured with stochastic simulations. We have developed a novel stochastic operator-splitting algorithm to model such reaction–diffusion phenomena. The reaction and diffusion steps employ stochastic simulation algorithms and Brownian dynamics, respectively. Through theoretical analysis, we have developed an algorithm to identify if the system is reaction-controlled, diffusion-controlled
or is in an intermediate regime. The time-step size is chosen accordingly at each step of the simulation.

We have used three examples to demonstrate the accuracy and robustness of the proposed algorithm. The first example deals with diffusion of two chemical species undergoing an irreversible bimolecular reaction. It is used to validate our algorithm by comparing its results with the solution obtained from a corresponding deterministic partial differential equation models. In this example, we also compare the results from our method to those obtained using a Gillespie multiparticle (GMP) method. The second example, which models simplified RNA synthesis, is used to study the performance of our algorithm in reaction- and diffusion-controlled regimes and to investigate the effects of local inhomogeneity. The third example models reaction–diffusion of CheY molecules through the cytoplasm of *Escherichia coli* during chemotaxis. It is used to compare the algorithm’s performance against the GMP method. The Center researchers’ analysis demonstrates that the proposed algorithm enables accurate simulation of the kinetics of complex and spatially heterogeneous systems. It is also computationally more efficient than commonly used alternatives, such as the GMP method.

**Tissue-Specific Analysis and Identification of Silencing Mechanisms**

As increasing amount of interaction data becomes available, there is a realization that context plays a key role in shaping biological processes—networks are strongly shaped by the tissues within which they exist. To this end, we have initiated an effort at tissue-specific characterization of networks.

**Analysis of the Human Tissue-specific Networks**

The global human interactome consists of a static snapshot of potential physical interactions that can occur between pairs of proteins. However, it does not provide any information regarding the spatiotemporal characteristics of the actual protein bindings. In order to add a context to interpret each interaction, Bossi and Lehner (2009) overlaid the mRNA expression level of each transcript in different human tissues (Su et al. 2004) on top of the global human interactome, integrated from 21 PPI databases, and constructed a set of 79 tissue-specific networks, which forms the basis of our work and subsequent analysis. More information about this study is available at http://compbio.soihub.org/projects/ttsn.

The constructed tissue-specific networks have an average of approximately 37K interactions between 5,600 human proteins. Given the interaction density as the fraction of edges in the network over all possible protein pairs, Grama and collaborators find that these networks are denser on average than the global human interactome (0.0044 average density versus 0.0015). This may be explained by the fact that all of these networks share a common dense core of the universally expressed genes, corresponding to housekeeping proteins, but only retain a subset of loosely connected tissue-specific proteins. We analyze the connectivity of these networks in terms of the size and distribution of their connected components. The biggest connected component (CC) has an average size of about 5,500 across tissue-specific networks, whereas the total number of components is about 51 on average. Interestingly, only one CC has more than five vertices in each
network, which is known as the giant CC, and no singleton (isolated) vertex exists in any of the constructed networks.

These results may seem contradictory to the recent findings by Magger et al. (2012), where they suggest that node removal, the method used by Bossi and Lehner (2009), fragments the original network into many small connected components and significantly changes the connectivity of the underlying network. We argue, however, that the observed inconsistency is an artifact of the coverage difference in the global interactome used in these studies, combined with the high sensitivity of tissue-specific networks to missing edges, i.e., false negatives. The rationale behind this argument is that we know the tissue-specific proteins have lower degrees and are highly enriched with membrane proteins that are involved in the transporter and receptor activities within the cell (Bossi and Lehner 2009; Emig and Albrecht 2011; Emig et al. 2011). These proteins act as sources of information propagation within various signaling pathways. They commonly interact with secondary messengers to amplify the perceived signal, branch out, and eventually connect to the housekeeping proteins to modulate core-cellular functions. This induces a tree-like structure in the underlying graph, which makes it vulnerable to random deletion of top-level tree edges.

The observed sensitivity to missing edges, together with the known bias of experimental methods in identifying the interacting partners of membrane proteins (Helbig et al. 2010; Rabilloud 2009), support Grama’s hypothesis that the underlying connectivity of tissue-specific networks is more sensitive to false-negative than false-positive edges. The complete list of computed network properties and statistics for human tissue-specific networks is available for download from http://compbio.soihub.org/projects/ttsn.

**Identifying Housekeeping Proteins**

The term housekeeping (HK) was coined in 1965 (Watson et al. 1965) to describe genes that are universally expressed in the cell. It was later (Butte et al. 2001) linked to genes’ role in maintaining core-cellular functions (Dezso et al. 2008; Chang et al. 2011). These core functions include mitochondrial respiration, ribosome biogenesis, ubiquitin-dependent proteolysis, translation, RNA processing, intracellular transport, cell cycle, cytoskeleton remodeling, and energy metabolism (Dezso et al. 2008; Chang et al. 2011; Souiai et al. 2011). In terms of the connectivity, HK proteins constitute a dense core in the global human interactome (Dezso et al. 2008; Bossi and Lehner 2009; Lin et al. 2009; Souiai et al. 2011). Most of the tissue-specific proteins interact with this core to modulate tissue-specific functions, and most of, if not all, the housekeeping proteins have a significant number of tissue-specific interactions (Bossi and Lehner 2009).

Interestingly, tissue-specific interactions are themselves organized in a hierarchical structure based on their function. Interactions involving regulatory and developmental functions occupy more central positions, whereas interactions involving in organ physiological functions are usually in peripheral positions (Souiai et al. 2011). There are various ways of identifying housekeeping proteins using different high-throughput experiments (Chang et al. 2011; Dezso et al. 2008; Bossi and Lehner 2009). Here, we follow the definition proposed by Bossi and Lehner (2009), in which HK proteins are
identified based on the number of tissues in which they are active. Among 10,229 proteins in the human global interactome, 2,013 proteins are not active in any of the 79 tissues. Figure 5 shows the membership of 8,216 remaining proteins in different human tissues. Unlike previous studies suggesting a bimodal distribution for tissue specificity of the proteins (Yanai et al. 2005; Dezso et al. 2008), we observe a unimodal distribution with the peak occurring for HK proteins. We extract the list of proteins that are active in all 79 tissues, resulting in a total of 1,540 HK proteins, which are available for download at http://compbio.soihub.org/projects/ttsn.

Figure 5. Unsupervised hierarchical clustering of the normal human tissues based on the variation of their normalized mRNA expression profiles resembles an intuitive grouping of tissues based on their type. Expression profiles are normalized and colored using z-score normalization and clustered based on the Pearson correlation of normalized expression signatures.

The results discovered by Grama and collaborators recapture most of the previously known functions for HK proteins, such as terms related to oxidative phosphorylation, intracellular transport, and translation, but more interestingly, they also identify novel terms that have not been previously studied. The viral infectious cycle, antigen processing, and presentation of peptide antigen via MHC class I and mRNA splicing via spliceosome illustrate some of these new terms.

Alignment Consistency of the Housekeeping Proteins Across Different Human Tissues

Because the genes corresponding to the HK proteins are universally expressed across different human tissues, we expect that the computed network alignments should yield consistent results in terms of identifying the yeast partners of the HK proteins. In order to quantitatively assess this hypothesis, we first create an alignment consistency table, which summarizes the alignment results. For each HK protein, in each pair of aligned networks (tissue-versus-yeast), we find the corresponding yeast partner under the alignment. For unaligned HK proteins, corresponding to the proteins that are not conserved in yeast, we align them with a gap. Then, for each HK protein, we summarize the corresponding row of the table by identifying the most common yeast partners across different tissues and record the top-ranked and runner-up yeast partners, as well as the
percent of tissues in which they are observed. The final consistency table is available for download from http://compbio.soihub.org/projects/ttsn.

Preliminary analysis of the consistency table indicates that 366 HK proteins are not aligned to any yeast proteins in any of the aligned tissues (consisting ~ 23.77% of all HK proteins). For the remaining 1,174 HK proteins, we first look at their top-ranked yeast partner. If it is not a gap, we report it as their canonical yeast partner. Otherwise, if it is unaligned in a majority of tissues, we use the runner-up partner (second most common partner across different human tissues). Finally, we record the alignment score for each identified pair of proteins in yeast and humans. The final table of all partners and computed scores is available for download from http://compbio.soihub.org/projects/ttsn.

We can verify that 366 HK proteins are not aligned to any yeast protein by noting that only about 76.23% of HK proteins are aligned in at least one tissue. On the other extreme, 408 HK proteins are aligned to the same yeast protein in all 79 human tissues, constituting about 26.49% of all HK proteins. We call these proteins, which are highly conserved in yeast, the core HK proteins. We construct the alignment subgraph induced by these proteins and their corresponding yeast partners called CoreNet, as follows.

For each pair of HK proteins and pair of their yeast partners, we create a conserved edge if an edge exists in both the human and yeast interactome. Otherwise, if the edge only exists in one of the networks, we record a gap edge.

**Quantifying the Similarity of Human Tissue-specific Networks to the Yeast Interactome**

Given the network alignment between a human tissue-specific network and the yeast interactome, we are interested in assessing the significance of this alignment, which can be equivalently interpreted as the similarity or dissimilarity of the given tissue to yeast. We use simulation to assign an empirical p-value to each network alignment. For each tissue-specific network, we sample \( kR = 10,000 \) pseudo-random tissues, separately align them with yeast interactome, and compute various p-values and bounds. The comprehensive list of all computed statistics is available for download from http://compbio.soihub.org/projects/ttsn.

We extract the most and the least similar tissues to yeast by applying a suitable threshold to the p-value upper and lower bounds, respectively. Listed in Table 1 are a total of 23 of 79 tissues that show a significant similarity to the yeast interactome. Most blood cells show a very significant similarity to yeast. In fact, in 10,000 samples, there has not been even a single instance where either the alignment weight or the overlap of the random sample exceeds the original alignment. Similarly, cancer cell lines consistently show significant alignment p-values. On the other hand, 19 of 79 tissues show the least significant similarity to the yeast.
Table 2 shows the comprehensive list of dissimilar tissues. Ganglion tissues show the least similarity to the yeast. An interesting observation is that tissues and cell types that go on either extreme of the table, either very top or very bottom, usually have very high confidence values, i.e., both their topology and homology consistently align.
Transcriptional Similarity Across Different Human Tissues

The Center’s next goal is to identify whether or not the similarity of human tissues to yeast correlate with their pairwise similarity. In other words, if a pair of human tissues resembles a high similarity to each other, do they also have consistent alignment p-values with yeast? Conversely, given that a pair of human tissues is significantly similar to yeast, what can we infer about the pairwise similarity of these tissues? To address these questions, we perform a parallel experiment to our main analysis for identifying similar tissues to yeast, by constructing a tissue–tissue similarity network (TTSN) based on the transcriptional similarity of the human tissues. Unlike our previous experiments, where the main source of data was the combination of network topology with partial expression profile (restricted to the proteins in the human interactome and binarized), here we used a complementary data set of full transcriptional profile for all human ORFs (approximately 44K transcripts) but will not use the topology.

Figure 5 illustrates the hierarchical clustering of human tissues. We used one minus the Pearson correlation of the transcriptional signatures as the distance function between tissues, and the unweighted average distance via the UPGMA method for computing the distance between clusters. Columns and rows are permuted according to the clustering of genes and tissues, respectively. Coloring of each entry is based on the z-score of the expression value, computed across different tissues (i.e., column-wise normalization). Initial investigation of the clustergram indicates an inherent grouping among human tissues, which is highlighted in the figure. To analyze this underlying structure, we first compute the correlation matrix. Not surprisingly, almost all tissues exhibit a very high transcriptional correlation. This is largely due to the existence of universally expressed genes, corresponding to HK proteins, across these tissues. To remedy this, we subtract the baseline similarity from the correlation matrix by computing its linearized z-score. Positive entries in this matrix correspond to tissue pairs with higher than average transcriptional similarity. We construct the TTSN by retaining all the positive edges and leaving out all negative edges. The final constructed TTSN is available for download from http://compbio.soihub.org/projects/ttsn.

Finally, we apply a network-clustering algorithm, based on the affinity propagation method, to identify groupings of tissues in the TTSN. Figure 6 illustrates the top three identified clusters, each of which is labeled accordingly to represent the majority of the cell or tissue types in the cluster. Thickness of each edge represents the extent of transcriptional similarity between tissues. These clusters mainly represent brain tissues, blood cells, and cancer cell lines.
Figure 6. The TTSN is constructed on the basis of transcriptional similarities across different tissue, after removing baseline similarity due to the universal expression of housekeeping proteins. Only edges with positive z-score, i.e., similar tissues, are considered in this experiment. We applied a network-based, affinity propagation method to identify dense clusters in the TTSN. Each subfigure illustrates one of the identified clusters, labeled according to the majority of cell or tissue types in the identified cluster. Thickness and transparency of each edge represent the extent of transcriptional similarity between tissues.

Projecting Alignment p-values onto the TTSN

To better understand the correlation between the similarity of human tissues and their alignment p-value with yeast, we merge our previous results in a unified framework. We extract the subset of the most reliable edges in the TTSN by applying a stringent threshold on z-score over edges, which corresponds to a 95% confidence interval. Next, we mark each node, corresponding to a human tissue, by its alignment p-value with yeast. We log transform the mixed p-value score and then use min-max normalization to scale it between [-1, +1]. Finally, we assign it to the color spectrum of red-white-green to visually distinguish between similar and dissimilar tissues. The final network is illustrated in Figure 7. Here, green nodes are tissues with the most significant similarity to yeast, while red nodes show the least similar tissues. The thickness of edges illustrates the extent of similarity between human tissues.

Compared to clusters depicted in Figure 6, a significant overlap exists in functional groupings of the identified tissues. We observe that most of the nodes representing brain tissues, blood cells, and cancer cell lines are enriched in both networks. However, the high z-score threshold that is used here fragments the blood cell and cancer cell-line clusters into multiple small components with only the strongest edges retained.

Two additional groupings can be observed in this map—ganglion tissues and testis tissues. The former shows consistent dissimilarity to the yeast interactome. Interestingly, skin tissue and the atrioventricular node are also clustered with ganglion cells. On the other hand, testis tissues demonstrate higher variability in terms of their alignment p-value. Testis is a unique gland because it is involved in both endocrine secretion of testosterone, as well as exocrine creation of semen. Its structure consists of a variety of different types of tissues, working together to perform these
unique functions—some of them similar to yeast, while others are not. We argue that the variability of $p$-values is attributed to the tissue mixtures in different gross-cuts of this organ.

![Figure 7](image)

**Figure 7.** The TTSN thresholded at $1.96 \leq z$-score, colored with the $p$-value of alignment with the yeast interactome. Green nodes are tissues that have significant similarity to yeast, while red nodes are ones that do not. Intensity of colors is related to the significance value, while thickness of edges represent the extent of transcriptional similarity between tissues. Isolated nodes (tissues) are eliminated here.

**Inferring the Effective TOR-dependent Network in Yeast Underlying Aging**

Calorie restriction (CR) is one of the most conserved nongenetic interventions that extend health span in evolutionarily distant species, ranging from yeast to mammals. The target of rapamycin (TOR) has been shown to play a key role in mediating health-span extension in response to CR by integrating different signals that monitor nutrient availability and by orchestrating various components of cellular machinery in response. Both genetic and pharmacological interventions that inhibit the TOR pathway exhibit a similar phenotype, which is not further amplified by CR.

Grama and collaborators present the first comprehensive, computationally derived map of TOR downstream effectors with the objective of discovering key life-span mediators, their crosstalk, and high-level organization. We adopt a systematic approach for tracing information flow from the TOR
complex in the yeast interactome and use the approach to identify statistically significant signaling elements. By constructing a high-level functional map of TOR downstream effectors (Figure 8), we show that our methods are not only capable of recapturing previously known pathways, but also suggest potential targets for future studies.

**Figure 8.** Enriched Gene Ontology (GO) slim terms are identified by mHG *p*-value, computed for the ranked list of genes based on their information-flow scores. Each node represents a significant GO term, and edges represent the overlap between gene sets of GO terms. Terms in different branches of GO are color-coded with red, green, and blue. Color intensity of each node represents the significance of its *p*-value, while the node size illustrates the size of its gene set. Thickness of edges is related to the extent of overlap among geneses.

Information-flow scores provide an aggregate ranking of the relevance of proteins with respect to the TOR-signaling pathway. These rankings must be normalized for degree bias, appropriately interpreted, and mapped to associated roles in pathways. To gain a detailed mechanistic understanding of the roles of various proteins (and associated signaling elements), we present a novel statistical framework for integrating information-flow scores and for correlating them with the set of differentially expressed genes in response to rapamycin treatment. We also interpret these scores in the context of the yeast transcriptional regulatory network with the goal of deciphering most relevant transcription factors in mediating observed transcriptional response. Using our
algorithmic and statistical framework, we construct the effective response network (ERN) of the TOR pathway, consisting of 1,288 regulatory interactions between 17 key transcription factors and 181 targets. We hypothesize this network to play a key role in mediating the life-span extension in response to TOR inhibition (Figure 9).

Figure 9. The Center computed the ERN for most relevant transcription factors, yielding a network of 1,288 transcription regulations between 17 transcription factors and 181 target genes. This is shown for illustration purposes only—the data are available for download and further analysis at http://compbio.soihub.org/projects/torc1.

Our approach, unlike experimental methods, is not limited to specific aspects of cellular response. Rather, it predicts transcriptional changes and post-translational modifications in response to the TOR inhibition. Figures 10 and 11 demonstrate fine-grained details revealed by our analyses of TOR-dependent transcription initiation and Gap 1 regulation. The constructed ERN greatly enhances understanding of the mechanisms underlying the aging process and helps in identifying new targets for further investigation of antiaging regimes. It also allows us to identify potential network biomarkers for diagnosis and prognosis of age-related pathologies.
**Figure 10.** Induced subgraph in the yeast interactome, constructed from the top-ranked genes in the information-flow analysis that are annotated with the transcription initiation Gene Ontology term. Different functional subunits are marked and color-coded appropriately.

**Figure 11.** The schematic diagram is based on literature evidence for the interactions. Each node in the signaling pathway is annotated with the rank of its information-flow score from TORC1. Ranking of nodes based on their information-flow scores coincides with our prior knowledge of the structure of this pathway.
**Spatiotemporal Modeling of Fluxes in Networks**

The Center’s work over the past year has focused on the development of a detailed causality-based, dynamic theory of cellular metabolism, based on cybernetic mechanisms, that postulates that regulatory processes underlying the activity of metabolic networks are inspired by the organism’s survival motives. We aim to glean detailed information from comprehensive metabolic data (e.g., gene-expression profiles, intracellular fluxes, etc.) with the help of information-theoretic measures for comparison with detailed predictions of the above cybernetic theory of metabolism. We also aim to generate a comprehensive framework for model comparison in systems biology using information-theoretic tools.

Organisms that currently exist in nature are a product of millions of years of evolution. They have prevailed over other competing species over time by the acquisition of elaborate control mechanisms, endowing them with regulatory capacity to survive in diverse environmental conditions. In developing a quantitative theory of metabolism, due regard must be accorded to account for such regulatory processes. Metabolic regulation is so intricate, diverse, and pervasive that its complete description would be beyond most mathematical frameworks. An alternative approach to a comprehensive accounting of regulation lies in the cybernetic idea that regulatory processes are inspired by the goal of survival.

We have developed a general cybernetic theory of metabolism based on the postulate that cells invest their constrained resources optimally on synthesizing enzymes and controlling their activities to selectively drive reactions that favor the organism’s survival in a dynamic environment. By further theorizing that survival is ensured by maximizing the carbon uptake rate of the organism, a detailed theory of metabolism has become available with unique attributes. First, it serves to “explain” bioinformatic data as the consequence of realizing the survival objective of the organism. Thus, whether observed gene-expression profiles as reflected, for example, in microarray data can be interpreted in terms of maximizing carbon-uptake rate becomes a highly organized effort in comparison with statistical approaches focusing purely on descriptive information with a listing of coexpressing genes. Second, it facilitates mathematical approaches to the control of bioprocesses so as to maximize the productivity bioproducts. Third, it becomes a mathematical tool for analyzing all metabolism-associated phenomena that include the onset and treatment of diseases, personalized medicine and so on.

Metabolic networks comprise multitudes of chemical reactions, each catalyzed by a specific enzyme that must result from the expression of a gene. As chemical changes must conserve mass, mass balances must be made of all metabolic species, intracellular and extracellular. As intracellular reactions occur at rates considerably higher than those occurring across cell boundaries, intracellular metabolites admit the pseudo-steady-state approximation yielding homogeneous algebraic equations. The more slowly varying extracellular variables satisfy differential equations. In our theory, the differential equations involve concentration variables and cybernetic variables—the latter representing regulatory effects derived by the use of optimal control theory. The cybernetic
variables dealing with enzyme synthesis appear in enzyme balances, whereas those concerned with the control of enzyme activity appear in the mass balance equations for cellular metabolites.

![Graph showing metabolic diversity](image)

**Figure 12.** Metabolic diversity when growing at a constant rate in a mix of carbon substrates.

Ramkrishna and Song (2012) have reviewed the development of cybernetic models from inception to current state of the art. Models addressing large networks use the concept of elementary modes (EM), which are pathway options for the organism. The uptake of substrate among the different EMs is distributed optimally to maximize the total carbon entering with the substrate, which is done by maintaining differential levels of enzymes catalyzing substrate uptake through the EMs. Kim et al. (2008) discussed this methodology, known as HCM (hybrid cybernetic models). An exciting consequence of this model was reported last year, which showed experimental validation of multiple steady-state behavior of fermentation products in bacterial metabolism in a continuous reactor fed with a mixture of carbon substrates, glucose, and pyruvate. Further exploration of this scenario has led to even higher metabolic diversity in bacterial cells when growing at a constant growth rate in a mixture of carbon substrates (Figure 12).

**Protein Structure Prediction**

Three-dimensional (3-D) protein structures determine the function of a protein. Classification of 3-D protein structures is therefore crucial to inferring protein functional information, as well as the evolution of interactions between proteins. Center researchers Liu, Burge, and Rwebangira designed different methods to predict the 3-D protein structures.

We used an artificial neural network (ANN) to classify the protein structures. The approach equally divides a 3-D protein structure into several parts and then extracts statistical features from each part. The different parts in spatial domain are viewed as the feature vectors that are input into the neural network. The experimental results show the method achieves 71% accuracy on protein-structure classification.
We developed a threshold-independent approach to deal with the gaps in the skeletons. The threshold-independent approach overcame the problem that the skeleton is sensitive to thresholds that may lead to gaps in the skeleton. The images were represented as graphs and a set of volume trees. The experimental tests with synthesized and authentic images show the percent of improvement is 117% and 40% for Gorgon and MapEM, respectively.

We proposed a graph process unit (GPU) based, parallel simulated annealing algorithm to improve the efficiency and accuracy of the protein-structure prediction. We introduced a compute unified device architecture (CUDA) and applied a statistical method to optimize the performance of the CUDA-based parallel algorithm. The experimental results show that the approach provides a feasible solution for protein-structure prediction.

We proposed a novel method for the rapid classification of 3-D structures using two-dimensional hidden Markov models (2DHMMs). We convert the tertiary structures of a protein into a 2-D inner-distance matrix and Ca to reduce the complexity and use the secondary structures and tertiary structures to represent super states and 1-DHMM states in the 2DHMM model. We then used the wavelets to reduce the noise in the 3-D structures. The results show that this method performs an effective and efficient classification of 3-D structures.

**Prediction of Protein–Protein Interactions**

Protein–protein interactions play a key role in the completion of cellular functions and usually correlate to each other in protein–protein interaction networks.

We built a cloud-computing–based computing pipeline to predict protein–protein docking sites in a massive protein data set. The pipeline combines several existing protein–protein docking site-prediction methods and takes advantage of distributed data-processing services. The results show that the method can highly improve the performance of protein–protein docking site prediction.

We proposed a new algorithm to study the interactions between proteins. We constructed a protein–protein interaction network from two query proteins. We then added all the neighbors of the two query proteins from an online protein database. We used an improved network-partition algorithm to split the network into subnetworks, and we then proposed a network-clusters–based scoring function to predict the interactions between proteins. The experimental results show that the scoring function predicts the protein–protein interactions accurately.

In order to predict protein–protein interactions over diverse data sets, we proposed a full Bayesian network (BN) model. First, we investigated the dihedral angle of atom C-alpha to describe flexible and rigid regions of protein structures and then design a domain–domain interaction (DDI) template library to predict the DDIs by the dihedral angle of atom C-alpha. We used two sequence encodings, the protein structures, the DDIs, and gene coexpression as the features for the BN model. We used a receiver-operating characteristic to evaluate the BN model's performance. Compared to the support vector-machine model, the sensitivity and specificity obtained from the BN model are increased 4% and 5%, respectively.
Neuronal Sensing, Modeling, and Control

Neurologic disorders and injuries are the leading cause of disability and the third leading cause of death in developed countries, constituting over 6% of the global burden of disease. In contrast to currently available multimodal ways to monitor the heart and lungs, we still have limited ways in which we can monitor the brain, preventing us from both understanding and managing neurological disorders in a timely manner. New advances in neurotechnologies, being accelerated with the BRAIN initiative, are leading to increasing ways to record ensemble neural processes across multiple modalities and brain regions. These new technologies will create opportunities for the Center to resolve physiological signals relevant to neurological functioning—hence, the diagnosis, treatment, and understanding of acute (e.g., traumatic brain injury) and chronic (e.g., epilepsy and Parkinson’s) neurological disorders. However, this “data deluge” will render our ability to make predictions, understand mechanisms, and suggest treatment to be limited, by the extent to which we can analyze and interpret the information in a succinct and predictive manner.

Figure 13.

Coleman and collaborators have recently developed a class of methods to succinctly capture spatiotemporal patterning of neural spiking in the cortex to reflect directed, causal or functional network mechanisms. These methods are rigorous, computationally efficient, and scalable. We have instantiated the theoretical results to guide recent neuroscientific findings that relate networks of functionally connected neurons to cortical wave propagation within the primary motor cortex. We then establish that this spatial layout accounts for the magnitude of task-relevant information contained in the sequential firing of functionally connected neurons. We further demonstrate cell-type specificity—excitatory neurons of thin spike width play a dominant role in this ensemble spike sequencing. We hypothesize that this spatiotemporal patterning may reflect the local somatotopic organization of motor cortex.
Figure 14. Patterns from three monkeys show: (a) Networks of significant directed connections on the array red (blue) arrows representing excitatory (inhibitory) connections. R, rostral. C, caudal. M, medial, and L, lateral denote anatomical orientation. (b) Circular distributions of directed excitatory (left) and inhibitory (right) connections for time window [-50 100] ms (purple) and window [150 300] ms (red). Inhibitory connections are in teal green and blue, respectively. Dotted black lines represent the beta-wave axis. (c) Same as in (a), but using neurons with firing rates of 1 spike/sec and 0.40 ms spike width.

Specifically, we estimated functionally connected networks of spiking neurons using a directed information analysis for point processes to demonstrate that a class of simultaneously recorded, single-motor cortical neurons in nonhuman primates spatially coordinates their spiking activity in a manner that mirrors the beta-wave axis. The circular distribution of excitatory connection directions was bimodal and oriented close to the beta-wave axis in all three monkeys.

By fitting the beta-wave directional distributions with a mixture of two von Mises functions with nonfixed dispersion parameters, we determined the mode and standard deviation of each wave mode and found that the modes of excitatory connections fell within 1.1 standard deviations of the beta-wave modes. In contrast, the circular distributions of inhibitory connection directions were also bimodal, but oriented almost orthogonal to the beta-wave axis. In contrast to the narrow spiking neurons, few, if any, statistically significant connections existed among wide spiking neurons in each time window.

We also demonstrated that sequential spiking activity of that class of neuron pairs contains task-relevant, target-direction information whose magnitude varies, according to the spatial orientation of the constituent neurons in a manner consistent with the beta-wave axis.
Network Models for Neurons

It is now possible to record simultaneously the activity of more than 100 neurons as they carry out the business of the brain. Center researchers have collaborated with our Princeton colleague MJ Berry II, whose laboratory has perfected these methods for recording from the vertebrate retina. These experiments involve monitoring not just a large number of neurons, but a large fraction of the neurons devoted to representing a small patch of the visual world. These neurons, the retinal ganglion cells, collect signals from a radius comparable to that of the recording electrode array, so that the cells we study are within an “interaction volume” of circuitry. Steady improvements in experimental methods since the start of our collaboration have made it possible to record up to 200 neurons, with 100 neurons being routine. These recordings are very stable, allowing the sampling of activity over an entire afternoon, as the retina views naturalistic movies, or other stimuli designed to probe specific features of neural coding and computation.

Bialek and collaborators’ initial work involved making maximum entropy models for groups of 10 to 15 neurons, models that captured the mean activity of individual cells and their pairwise correlations. We have now built such models for up to 160 neurons, matching the individual and pairwise behaviors, as well as the distribution of summed activity in the network (that is, the probability that \(K\) out of \(N\) neurons generate action potentials in the same small window of time). While models for small groups of cells can be tested just by comparing the predicted probabilities for every state with their observed frequencies, this is not possible in large networks—simply too many states. To test out models, then we have to look more subtly at the predicted statistical structure of the responses.

We have found that the most compelling test of our models is the prediction of how many states should have a particular probability—the distribution of the probabilities themselves. This is connected to the density of states vs. energy in statistical physics. The model assigns a probability to every state, and we can accumulate a distribution over the states that actually occur in the experiment, or we can compute this distribution of probabilities purely from the model. Astonishingly, theory and experiment agree that probabilities are so small that they correspond to states predicted to occur only once per month. Although we, of course, cannot “measure” the probability of such rare states individually, there are so many such states that they make up a macroscopic fraction of all states observed; the remarkable result is the number of rare states that we see in the data agrees so precisely with the theoretical prediction, with no adjustable parameters.

As the United States prepares to launch the BRAIN initiative, which will involve efforts to record the activity of ever-larger collections of neurons, the fact that we can build such accurate models for the distribution of activity in networks of more than 100 neurons seems very significant.

As in statistical physics itself, writing down the probability distribution for the states of a system is the beginning, not the end, of understanding. From statistical physics, we know that probabilistic models for systems with many degrees of freedom break into distinct “phases,” regions of
parameter space with qualitatively different behaviors that are separated by critical points, lines, and surfaces. In the context of neural networks, the different phases would correspond to varying kinds of ordered states—for example, the network recalls stored patterns of activity or generates stereotyped rhythms. The surprising result of our work is that the more accurate the models that we build and the larger the groups of neurons that we consider, the closer our model comes to sitting precisely on a critical surface. The idea that biological systems might poise themselves near criticality is an old one, but this is a clear example in which the data are driving us to this conclusion.

Bialek and collaborators have used the same maximum entropy approach to think about the emergence of order in flocks of birds, working in collaboration with Andrea Cavagna and Irene Giardina at the University of Rome (Sapienza). In a first try at the problem, we built models for the joint distribution of flight directions that were consistent with the mean correlation between directions of individual birds and their immediate neighbors. The result was parameter-free theory for the propagation of directional order through the flock, and this theory was in exceptionally precise agreement with data on the directional correlations vs. distance between birds, and even correctly predicts subtle higher-order correlations. As a second step, we considered maximum entropy models that match the similarity of vector velocities (not just direction, but also speed) among neighbors and the variance of individual speeds around the flock’s mean. This model again has no free parameters with which to fit the overall behavior of the flock—all parameters are set by measurements on neighbor correlations and velocity variance. Nonetheless, this model quantitatively reproduces the long-ranged correlations in both direction and speed fluctuations among birds in the flock.

The maximum entropy models of flocking that we generated have parameters that are determined entirely by the data. Also, they are equivalent to statistical mechanics problems, so we can ask where in the phase diagram the parameters of the real system are sitting. The answer is that the parameters are in the critical regime, the thin band of parameters (with width inversely related to the size of the system) surrounding the critical surface. We find it quite astonishing that both the flocks and the networks of neurons are poised near criticality.

The critical point is a place where many quantities are extremized. There is a growing appreciation that this extremization can be phrased in purely information-theoretic terms, generalizing the discussion of particular forms of order that has been more conventional in the physics literature. Although we can speculate, it is not yet clear which of the many special features of criticality are relevant for these biological systems, but we find the evidence for criticality to be increasingly compelling. Certainly, the maximum entropy construction is leading us to the conclusion that these systems are not at generic points in parameter space—and this by itself is an important conclusion. In the coming year, we will be pushing further on the idea of criticality, hoping to find the “smoking gun.”

As an aside, we have used the maximum entropy ideas to look at the patterns of voting on the Supreme Court. Although intended as a somewhat playful exercise for an undergraduate student with multidisciplinary interests, the resulting work provided a startlingly accurate and simple
account of the data. We take this as very concrete evidence that simple models, grounded in
statistical physics and information theory, can provide a very accurate description of collective
behavior, even in complex social and political contexts. In these models, the emergence of
consensus, for example, is not merely analogous to the ordering of more familiar physical systems, it
is mathematically identical. This work is already attracting some interest from the popular press.

**Optimization for Behavior Prediction**

The second major direction of our work concerns the use of optimization principles, grounded in
information theory, to predict the behavior of biological systems. In this direction, we continue to
make progress on a major effort to define the topology and parameters of genetic networks that
maximize information flow from the (input) concentrations of transcription factors to the (output)
expression levels of multiple interacting genes. In support of this theoretical effort, the Center has
an extensive collaboration with the laboratories of Thomas Gregor and Eric Wieschaus, using the
early events in the development of the fruit-fly embryo as a testing ground for our theories.

The problem of calculating information flow in moderately realistic models of transcriptional
networks is challenging. We took the problem in steps, starting with a single transcription factor
that “broadcasts” to many noninteracting target genes. Even this simple case is rich, illustrating
situations in which the optimal network has substantial redundancy, a structure that most biologists
would interpret as prima facie evidence against optimization. The next level of complexity allows for
interactions in the network of target genes, but with no feedback loops. This already makes it
possible for repressive interactions in the network to reduce redundancy, and the resulting patterns
of expression begin to remind us of those in the fly embryo, though the class of networks we are
considering still is too simple to be realistic.

This year we have pushed the theory to include the simplest example of networks with feedback, in
which a single gene is both responsive to inputs and self-regulating. Importantly, we could show
that optimal information transmission is never in the regime where self-interactions are strong
enough to generate multistability, although there are parameter regimes in which the information
transmission is maximally close to the critical point for the transition to multistability. We believe
that this analysis sets the stage for a more realistic calculation, in which we can search through the
full set of possible networks for those that maximize information flow. An important aspect of our
results thus far is the appreciation that certain “details” of the molecular mechanism involved in
connecting transcription factors binding to DNA with the regulation of RNA polymerase activity can
have a substantial effect on information transmission, and we are following this lead.

With our experimental collaborators (as described in part last year), we have shown how to actually
measure the amount of information that gene-expression levels carry about the position of cells in
the embryo. As we were doing this work, it became clear that really quite a lot of information is
being encoded—a handful of genes carry almost enough information to specify the identity of cells
uniquely along the main axis of the embryo. But to take these results seriously is to believe that the
measurements themselves are very accurate, and this started to strain credulity. The result was a
substantial effort by our experimental collaborators to establish the (surprisingly high!) precision of their measurements, a result that will have much wider implications as a whole community tries to address the dynamics of gene expression in more and more quantitative ways.

While this is not itself a result for which the Center can take credit, it is a convincing example of how theoretical work under the Center’s auspices can drive genuinely new experimental developments—even developments of experimental methods—that can have broad impact. In this case, we think it is fair to say that our collaboration was one of only a few groups in the world that could see why such precise measurements might be meaningful, and this is because of the theoretical picture that we articulate.

The measurement of positional information in the embryo revealed signatures of the optimization of information flow, the matching between the distribution of input signals and the noise characteristics of the channel through which these signals are being transmitted. In addition, more detailed analysis of the data raised subtle questions regarding information about absolute vs. relative positional information and the possibilities of error-correcting coding in the embryo. These issues are on the agenda for the coming year.

Genetic regulatory networks, including those in the fruit-fly embryo, can exhibit many qualitatively different behaviors, separated by critical surfaces. As noted above, there are conditions under which approaching the critical surface serves to optimize information flow. Although we need more theoretical work to be sure that this is the “solution” to the fly’s problem, we might be able to test the idea of criticality in the data more directly. At criticality, we should observe strong correlations in the fluctuations of different genes around their mean expression levels; a slowing of the dynamics along some, but not all directions, in the space of possible expression values; correlations of expression fluctuations over long distances in the embryo; and departures from a Gaussian distribution of these fluctuations. Analysis of recent experiments on the gap genes shows that all these signatures are observed, and that the different signatures are related in ways predicted by theory. While other explanations for these individual phenomena may exist, the confluence of evidence suggests that this genetic network is tuned to criticality.

Bialek’s initial interest in optimization principles came in the context of neural coding, where (many years ago) collaborations with experimental colleagues showed that sequences of action potentials generated by individual neurons convey information at rates that are a substantial fraction of the entropy rate, and that the mapping from sensory inputs to spike outputs adapts to the distribution of inputs in ways that serve to maximize this information transmission. In many ways, our work on information flow in the fly embryo is a direct translation of these ideas. An important caveat is that successful application of information-theoretic ideas to neural coding hinges on knowing which information is relevant—what is it that particular neurons are “trying” to encode? We would like, however, to have a more general theoretical approach, something that doesn’t rely on such detailed knowledge of particular neurons and circuits.
Almost all neural computations involve making predictions. Whether we are trying to catch prey, avoid predators, or simply move through a complex environment, the data we collect through our senses can guide our actions only to the extent that we can extract from these data information about the future state of the world. Although it is natural to focus on the prediction of rewards, prediction is a much broader problem, ranging from the seemingly simple extrapolation of the trajectories of moving objects to the learning of abstract rules that describe the unfolding pattern of events around us. An essential aspect of the problem in all these forms is that not all features of the past carry predictive power. Because costs are associated with representing and transmitting information, might sensory systems have developed coding strategies that are optimized, keeping only a limited number of bits of information about the past but ensuring that these bits are maximally informative about the future? Could we go further, and imagine successive stages of signal processing by the brain as attempts to predict future patterns of neural activity? Bialek has addressed these ideas in the context of the vertebrate retina in collaboration with the laboratory of MJ Berry II.

If the efficient representation of predictive information is an important principle, then we must learn how to measure the information that neural responses provide about the future. It is relatively easy to compute how much information is captured about particular aspects of the future (e.g., the position of an object 100 msec from now), but we would like not to be limited by our ability to guess the right feature. If we create in our experiments a world of sensory inputs generated from a known stochastic process, we can use this knowledge to generate many independent samples of a “past” that will lead to the same future, and we can do this for many different futures. In this way, we can measure the information that neural responses to the many independent pasts carry about their common future.

Using the “common future” method, we have found nearly every cell in the retina participates in a group of cells for which this predictive information is close to the physical limit set by the statistical structure of the inputs themselves. This contrasts with the well-known redundant coding of the immediate past in the retina. Groups of cells in the retina also carry information about the future state of their own activity, and we show that this information can be compressed further and encoded by downstream predictor neurons, which then exhibit feature selectivity that would support predictive computations, such as extracting epochs of constant motion. Efficient representation of predictive information is a new candidate principle that can be applied at each stage of neural computation and across all sensory modalities.
Analytic Approaches to Cell Structures

The Center's investigations are oriented toward developing new methods to apply the theory of information to the processing and interpretation of structural-microscopy data.

![Image of human and yeast cells under normal and deregulated epsin conditions]

**Figure 15.** Human and yeast cells (upper and lower panels, respectively) were imaged under normal and deregulated (increased amount of) epsin conditions. Epsin abnormalities lead to defects in cell proliferation, giving origin to cells that are concatenated, or unable to separate.

**Cell-division Phenotype Analysis**

Our research requires the analysis of large amounts of microscope images to identify factors that affect cell-division abnormalities, and we have been successful in identifying some of them (e.g., the endocytic protein epsin in Figure 14). To analyze these images, we used a quantitative approach where we counted the number of cells that were unsuccessful in completing cell division as a percentage of the total sample size. We also devised a method to estimate the severity of this cell-division defect based on the concept of circularity. Briefly, circularity is a measure of polarized extension, defined as follows.

\[
\text{circularity} = \frac{\text{area of a cell}}{\text{area of a circle of the same perimeter}}
\]

Preliminary analysis of images suggests that phenotype severity and circularity are inversely proportional. This robust quantitative approach will enable us to obtain more information from the images than the subjective visual inspection of data.
In this ANN diagram, circles represent the neurons in the network, and arrows, the connections between units. The input layer is made up of 5 clusters (one for each X position within the XXXYXXØ motif) containing 20 nodes each (representing the 20 possible residues—only 3 per cluster are shown), plus the Ø cluster with only 5 nodes (for F, M, I, L and V), yielding a total of 105 neurons. Neurons from the hidden layer are labeled h1 and h2, whereas the output neuron is marked o; both types of units rely on a logistic activation function, depicted as a sigmoidal output-input response. Final network output is denoted as V (interaction value). The weights associated to the input-hidden layer and hidden-output layer connections are indicated as $W_{ih}$ and $W_{ho}$, respectively.

**Intracellular Sorting of Proteins**

This effort pioneers the use of analytic methods to solve high-complexity problems in vesicle trafficking. A variety of transmembrane proteins are sorted to different intracellular compartments based on the presence of signals in their cytoplasmic domains. Among these sorting signals, the tyrosine-based motif YXXØ (Y is tyrosine, X any amino acid and Ø an amino acid bearing a bulky hydrophobic side chain) is one of the most widespread. It has been extensively demonstrated that the µ subunits of the clathrin-associated adaptor complexes constitute the receptors for those signals. Although there is a considerable degree of overlap, each µ subunit has distinct sequence preference for YXXØ signal recognition that depends on the nature of the X-residues. In addition, it is known that residues within the sequence of the signal can exert cooperative and inhibitory effects on each other, concerning the interaction with the µ subunit. Therefore, the inherent high complexity of the signal-recognition process makes it very difficult to predict a priori the specificity of a given Y-signal for the different µ subunits.

Because an extensive collection of experimental examples of interactions between YXXØ motifs and µ subunits is available, Aguilar and collaborators have developed and used a computational
approach based on the artificial neural network (ANN) paradigm to assess the problem of Y-signal specificity (Figure 15). After training, this ANN was able to predict the feasibility of specific YXXØ-µ interactions with accuracies > 90%. These ANN systems constitute powerful tools to evaluate or predict the individual contribution of each adaptor complex to the routing of a given transmembrane protein, therefore helping to understand the mechanics and dynamics of intracellular sorting of proteins.

Applications in Computer Systems Research

This year, as a Radcliffe Fellow, Lynch began a study of biologically inspired distributed algorithms. Such algorithms differ from other distributed algorithms in that they are generally more flexible and robust, and adaptable to changes during execution. They are also generally simpler (e.g., using less elaborate bookkeeping) and depend on simpler forms of communication (e.g., simple chemical signaling). The particular biological areas Lynch focused on were insect colonies, developing embryos, and the human brain. Our goals in doing this are both to understand how certain biological systems work and to use the insights to design better distributed algorithms.
2a. Research Thrust Areas’ Accomplishments

Briefly describe the research thrust areas at the Center. Please provide basic information for each thrust area and details of significant accomplishments during the reporting period, including any research partnerships and their contributions to the Center (do not include publications, presentations, etc., that are reported in Section VIII, Center-wide Outputs and Issues). Include in the narrative a discussion of the goals, activities, and outcomes and/or impacts in the current reporting period, if changed from the previous reporting period. Be sure to discuss how the activities in the various research thrust areas enable the Center to meet its goals/objectives described above.

Communication Thrust

We describe below our Center members’ activities over this past year within the Communication Thrust. These activities focus on the challenges we observe in communication systems. Productive collaborations between Center researchers examine fundamental problems in semantic communication, the coordination of networks (especially wireless), control applications, multiterminal source coding, finite-blocklength information theory, and information flow.

Hypercontractivity and Coordination

One of the Center’s challenge problems is to build a theory of coordination over networks. Toward the goal of coordination, distributed agents fundamentally need the ability to create joint probability distributions, which might then be used for strategic purposes. What joint probability distributions can we create over networks? The Center is tackling this largely open problem.

We have brought to bear tools from the area of hypercontractive inequalities into this area, with remarkable effect. For instance, we have shown that the strong data-processing constant of a joint probability distribution is explicitly given in terms of the so-called hypercontractivity ribbon of the joint distribution, which is the nontrivial set of pairs for which hypercontractive inequalities hold. A communication channel defines a linear map from functions on the output alphabet to those on the input alphabet, given by taking the conditional expectation. This maps \( L^q \) functions to \( L^q \) functions and is a contraction. However, generally speaking, more is true, in that this map is a contraction from \( L^q \) functions to \( L^p \) functions for a range of \( p > q \). This property is what is called hypercontractivity.

Our finding (Anantharam et al. 2013) contradicts a data-processing inequality result claimed by Erkip and Cover (1998), which has stood unchallenged in the literature for nearly two decades, and has been used in several works, which must now be revised.
Further, viewing the hypercontractivity ribbon as a characteristic of the underlying joint distribution allows us to give necessary conditions on which joint distributions can be realized, when coordinating over networks in some scenarios. The basic idea is that the ribbon cannot be made tighter than it already is when a joint distribution is created from another one.

Several Center-affiliated faculty, postdocs, and students have strong interests in this line of work. The Center enables meetings and discussions between those researchers. Those have been very conducive to the results obtained thus far, and will likely continue as we pursue this direction in the year to come.

During the coming year, we will also work to further develop the role of hypercontractive inequalities in the context of the communication thrust’s grand challenge of understanding coordination over networks. We are close to a complete characterization of the hypercontractivity ribbon in information-theoretic terms. Such a characterization would likely have several broad impacts and would naturally give rise to new investigations.

**Soft Decisions in Multiterminal Source**

Most multiterminal source coding scenarios have, to date, defied complete understanding of the achievable tradeoffs between rates and distortions. Perhaps the simplest case in point is lossy compression with a rate-constrained description of side information at the decoder: the ‘one-helper’ problem. The problem remains open, despite being but a slight seemingly innocuous extension of the unconstrained case, solved by Wyner and Ziv (1976) in a celebrated paper more than 35 years ago.

Recently, Courtade and Weissman (2012) solved the problem for the Holy Grail of multiterminal source coding: lossy reconstruction of two correlated sources based on limited-rate descriptions of each of the sources separately—a problem referred to generically as ‘multiterminal source coding’—for the case where distortion is measured by the logarithmic loss. Rather than a particular reconstruction symbol, the decoder is required to give a ‘soft’ reconstruction, which is a probability (belief) on the value of the original source symbol; the distortion is the logarithm of the reciprocal of the probability assigned to that symbol. While formally this is merely a special case comprised of one particular distortion measure and reconstruction alphabet (namely, a space of distributions), it does put a timely spotlight on the canonical idea of a ‘soft’ reconstruction in the context of lossy compression. Courtade and Weissman presented their complete solution to this problem at the 2012 International Symposium on Information Theory, and ISIT recognized it as the best student-authored paper award; the journal version will soon appear in the IEEE Transactions on Information Theory. We continue to investigate which fundamental limits are amenable to characterization (this is the exception, rather than the rule, in multiterminal Shannon theory). This project is yielding simple proofs (from first principles) of some celebrated results in the Gaussian case whose existing proofs are highly convoluted and confounding.
**Feedback Communication**

One of the Center’s most significant lines of exploration this past year stemmed from an attempt to characterize the influence of controlled feedback on channel capacity. We found a family of channels we refer to as POST (previous output is the state) in which, much to our surprise, despite the memory, numerical evaluation has shown that feedback does not increase its capacity. Using convex optimization, we showed that, if the output distribution induced by a code that achieves capacity of a channel with feedback can also be achieved feedback, then the feedback does not increase capacity. In other words, the induced channel-output distribution is the important criterion for deciding whether or not feedback increases capacity. This result allowed us to establish the counterintuitive result that the nonfeedback capacity is, in fact, equal to the feedback capacity for a large class of POST channels. We are still trying to develop intuition for this result. We are also still tackling the question of how to identify concrete schemes, in the absence of feedback that will achieve what our results indicate is achievable. We expect the insight gained thus far will be helpful in the investigation of a family of feedback communication settings we plan to explore in the following months.

**Structural Information**

Structural information appears in myriad applications, from biology to social networks to material sciences. However, we still do not have good metrics of information embodied in structure. Since the seminal paper of Shannon, we have become inundated with new data for which structural information plays crucial role. For instance, the Internet, social networks, biological networks, and medical records are all key examples that present grand challenges. These areas all can benefit from the extension of information theory to handle such “unconventional data.” Unconventional data often contains more sophisticated structural relations. For example, a graph can be represented by a binary matrix that further can be viewed as a binary sequence. However, such a sequence does not exhibit internal symmetries that are conveyed by the so-called graph automorphism (such automorphisms make certain sequences or matrices “indistinguishable”).

The main challenge in dealing with such structural data is to identify and describe these structural relations. In fact, these “regular properties” constitute “useful (extractable) information” understood in the spirit of Rissanen’s “learnable information.” Furthermore, such data structures often have two types of information: the information conveyed by the structure itself, and the information conveyed by the data labels implanted in the structure. A significant outcome would be an efficient way to jointly represent these two types. For now, structure itself is less understood, so we focused on graphical structures. Our recent attempt to address this problem was by studying information of unlabeled graphs. We first derived a relation between the entropy of a (labeled) graph and the structural entropy of (unlabeled) graphs. This gives, as a result, the design of the first provably asymptotically optimal compressor that achieves the structural entropy up to the first two leading terms. This work led to a new two-dimensional recurrence that was recently solved in the Center.
The next step is to extend these findings to the lossy case, but this will require the definition of a distortion measure that preserves structural properties, which is a challenging task. Abram Magner, a Center supported student, is working on this problem, along with extending existing results to other graph models such as preferential graph model generation. We already know that for such a model the underlying graph may be symmetric for some graph parameters. This directly impacts graph compression. Furthermore, another Center student Lan Vingh Troung is working on the capacity of a channel transmitting unlabeled graphs (structures) with some interesting and promising initial findings.

In a series of papers, we proposed novel tools for studying redundancy of memoryless and Markov sources, including Markov types. Specifically, we studied the min-max redundancy of universal coding for large alphabets over memoryless sources. In another recent work, we presented precise enumeration of Markov types, which turns out to be equivalent to the enumeration of Eulerian graphs and also the number of integer solutions of a system of equations characterizing the frequency matrix of the underlying Markov source.

This is also related to finding the underlying structure of Markov models. We are currently working on types for Markov fields that may provide deeper insights into Markov structure.

**Wireless Networks**

**Secure, Real-Time Communication**

The Center is investigating two of the major challenges associated with wireless networks: security and temporal guarantees in communication. Kumar and collaborators have developed a clean-slate approach to secure wireless communication. The traditional approach has been to build systems for performance and then provide patches as security vulnerabilities are identified. We have developed a comprehensive approach, which addresses the entire life cycle of wireless networks from initial formation to steady-state operation. The approach provides a guarantee of performance that is utility optimal in a precisely quantifiable game-theoretic way.

In our latest phase of study for wireless communication in real time, we have been studying a more fine-grained performance measure than throughput optimality, as well as fluctuations in packet deliveries. We also studied the issue in heavy traffic by employing a law of the iterated logarithm scaling and determining an upper bound that is touched infinitely often.

**Computing in Dynamic and Static Networks**

Today’s wireless networks tend to be centralized. They are organized around a fixed central backbone, such as a network of cellular towers or wireless access points. However, as mobile-computing devices continue to shrink in size and in cost, we are reaching the point where large-scale ad-hoc wireless networks, composed of swarms of cheap devices or sensors, are becoming feasible. In this work, Oshman (2012) and her Ph.D. advisor, Lynch, studied the theoretical computation power of such networks, and asked: What tasks are they capable of carrying out? How
long does solving particular tasks take? What is the effect of the unpredictable network topology on the network’s computation power.

In the first part of the work, we introduced an abstract model for dynamic networks. In contrast to much of the literature on mobile and ad-hoc networks, our model makes fairly minimalistic assumptions—it allows the network topology to change arbitrarily from round to round, as long as in each round, the communication graph is connected. We showed that, even in this weak model, global computation is still possible, and any function of the nodes’ initial inputs can be computed efficiently. Also, using tools from the field of epistemic logic, we analyzed information flow in dynamic networks and studied the time required to achieve various notions of coordination.

In the second part of our investigations, we restricted attention to static networks, which retain an important feature of wireless networks: they are potentially asymmetric. We showed that, in this setting, classical data-aggregation tasks become much harder, and we developed both upper and lower bounds on computing various classes of functions. Our main tool in this part of the work is communication complexity. We used existing lower bounds in two-player communication complexity, and also introduced a new problem, task allocation, and studied its communication complexity in the two-player and multiplayer settings.

Oshman is continuing her work on using computational complexity techniques to obtain lower bounds for fundamental problems in distributed computing.

In another aspect of distributed computing, Sastry and Lynch (Cornejo et al. 2012) have carried out an in-depth study of the power of failure-detection services to support solutions to fundamental problems of fault-tolerant distributed computing. Failure detectors are “oracles” that have been introduced to provide processes in asynchronous distributed systems with information about failures.

**Fundamental Limits on the Power of Wireless Networks**

The fundamental limits on the power of wireless networks have been a significant area for research among our researchers since the Center’s formation. We report our recent progress.

**Bounded-Contention Coding.** Censor and Lynch led an effort to develop a new bounded-contention coding (BCC) technique, which allows collisions in wireless networks with additive channels to be decoded, and whose costs depend on a bound on the contention among the transmitters. We showed how these channels can be used to implement reliable local broadcast in a particularly efficient way, with communication and time costs that depend on a bound on the amount of contention among the transmitters.

**Other Work on Coding-Based Communication.** More generally, we have been aiming to find ways in which coding techniques can improve the costs of solving higher-level global problems in graph-
based networks. For example, in Lynch, Khabbazian, Medárd, Kuhn, and Parandeh-Gheibi (2011), we reveal findings on global broadcast using algebraic coding methods.

We are also currently considering the costs of emulating atomic read/write shared memory in distributed message-passing systems. We are analyzing the costs of algorithms by Attiya, Bar-Noy, and Dolev and others, and developing new algorithms that use erasure coding to significantly reduce these costs. Our CAS (coded atomic storage) algorithm is superior to both previous algorithms in terms of communication costs. An optimization of CAS is also superior in terms of storage costs.

**Collision-Prone Networks.** The work on BCC is an outgrowth of another project that focuses specifically on reliable local broadcast. Lynch and collaborators are interested in algorithms that implement a reliable local broadcast layer over a variety of platforms, including platforms that simply lose messages when collisions occur. We have obtained many upper- and lower-bound results about the costs of local broadcast over such platforms, in particular, about algorithms that use probabilistic transmission. We want to compare the costs of these algorithms to those of coding-based implementations of reliable local broadcast.

We have also developed solutions to several nonlocal problems, including global broadcast, building maximal independent sets and dominating sets, and electing leaders over platforms with collisions and message losses. We intend to consider how solutions to these problems could benefit from use of coding methods.

**Communication in Wireless Networks: Network Coding vs. Routing**

The core task of any communication network is to transfer messages between different nodes. A fundamental challenge in designing network communication protocols is to achieve a high throughput. The classical method for disseminating messages over a network is routing. Messages are regarded as atomic tokens and routed from the source to the receivers; intermediate nodes only store and forward messages, without combining them. A newer and considerably more powerful approach is network coding, where intermediate nodes mix messages and send out coded versions of the received packets.

Network coding is known to often achieve the information-theoretically optimal throughput (typically the size of a minimum cut). However, this gain is usually accompanied by new complications and costs. Thus, determining the network-coding gain for different networking models is one of the important questions in the study of network coding. Network-coding gain is defined as the ratio of the throughput achievable via network coding versus the achievable throughput via routing. Determining the network-coding gain is of interest both in theory and in practice, and often has deep connections to combinatorial or graph theoretical problems.

Recently, interest in using network coding in wireless networks and in understating its gains over routing in these networks has increased. A prominent example is the work of Katti et al., where they
implemented a network-coding strategy for practical wireless networks and reported significant constant factor throughput gains (i.e., improvements over routing).

In four recent works, Ghaffari and Lynch present important theoretical results about the network-coding gain in wireless networks, and we also use network coding to achieve near-optimal algorithms for broadcast in wireless networks. These results are presented for two standard mathematical models of wireless networks: First, in the local broadcast model, above the MAC layer, each node can send one bounded-size packet per round to all of its neighbors. Second, in the radio-network model, below the MAC layer, each node is either silent or transmits a bounded-size packet, in each round. A node receives a packet if and only if it is listening and exactly one of its neighbors is transmitting. In particular, if two neighbors of a node transmit simultaneously, their transmissions collide (i.e., interfere) and are useless to the node; depending on the assumptions, the node may or may not be able to detect this collision.

Permutation Coding for Flash Memories

The Center has focused on the ideas of ordinal data representation and modulation coding in its study of novel coding schemes for flash memories. Permutations and multipermutations as information-representation formats have a long history, with early applications in communication theory dating back to the 1965 work of Slepian, who proposed using multipermutation codes for transmission in the presence of additive white Gaussian noise. More recently, Vinck proposed using permutation codes in the Hamming metric for combating impulse noise and permanent frequency noise in power grids. Permutation codes have received renewed interest in the past few years due to their promising application in storage systems, such as flash memories. The gist of the approach underlying such permutation coding is to store information via the relative order of charge levels of cells rather than their absolute charge values. This approach, termed rank modulation, alleviates the problems of cell over-injection, reduces the need for block erasures, and is more robust to errors caused by charge leakage.

Some of the Center’s researchers have pursued the study of a general “translocation error” model for ordinal data storage, based on the observation that increasing the number of charge levels to increase capacity may lead to charge drops that are significantly larger than the difference between adjacent charge levels. In addition, the proposed translocation error model adequately accounts for more general types of error, such as read-disturb and write-disturb errors. In this context, the distance measure of interest is the Ulam distance, related to the length of the longest common subsequence of two permutations and consequently, also related to the deletion-insertion, or edit, distance. The Ulam distance has also received independent interest from the bioinformatics and the computer science communities for the purpose of measuring the “sortedness” of (streaming) data. Milenkovic and collaborators are currently completing a line of work on devising codes for both permutation and multipermutation storage formats. Our design approach is based on new combinatorial constructions, termed “positional interleavers.”
**Empirical Distribution of Good Channel Codes**

Polyanskiy, Verdú, and collaborators study several properties of channel codes that approach the fundamental limits of a given memoryless channel with a nonvanishing probability of error. The output distribution induced by a capacity-achieving code is shown to be close, in a strong sense, to the capacity-achieving output distribution for discrete memoryless channels (DMC) and additive white Gaussian noise (AWGN). Relying on the concentration of measure (isoperimetry) property enjoyed by the latter, we show that regular (Lipschitz) functions of channel outputs can be estimated precisely and turn out to be essentially nonrandom and independent of the used code. We also found that the binary hypothesis testing between the output distribution of a good code and the capacity-achieving code cannot be performed with exponential reliability. Using related methods, we show that quadratic forms and sums of $q$-th powers, when evaluated at code words of good AWGN codes, approach the values obtained from a randomly generated Gaussian code word. The random process produced at the output of the channel is shown to satisfy the asymptotic equipartition property for DMC and AWGN.

**Fundamental Limits of Communication on Multi-Antenna Wireless Channels**

This line of the Center’s work investigates the maximal achievable rate for a given blocklength and error probability over block- and quasistatic multiple-input/multiple-output (MIMO) fading channels. Under mild conditions on the channel gains, it is shown that the channel dispersion is zero, regardless of whether the fading realizations are available at the transmitter or the receiver. The result follows from computationally and analytically tractable converse and achievability bounds. Through numerical evaluation, we verify that zero dispersion indeed entails fast convergence to outage capacity as the blocklength increases.

**Combinatorial Joint Source-Channel Codes**

Algorithms for lossy and lossless compression and error-correcting codes have been at the core of the digital revolution. In our work, we have focused on the particular set of applications in which both the lossy compression and noise resilience are required. Examples include storage of the high-resolution imagery on nonperfect flash memory and real-time video surveillance over jammed or noisy channels.

The state-of-the-art solution is “separation”: pair an off-the-shelf compressor with an off-the-shelf code; however, as was shown by another Center PI, for worst-case guarantees, the separated solution is far from being (even asymptotically) optimal. This realization is the main motivation for an omnidirectional line of investigation going on at the center of the combinatorial, geometric, algebraic, and information-theoretic aspects of the joint source-channel coding problem.

**Finite-Blocklength Information Theory**

Our Center researchers have pioneered new finite-blocklength achievability bounds for multiuser information theory based on new fundamental nonasymptotic covering and packing lemmas. A
radical departure from the conventional typical-sequence methods (as described, for example, in
the recent ElGamal-Kim textbook), this machinery has succeeded in giving new achievability results
in such multiterminal problems as Gelfand–Pinsker, Wyner–Ziv, Ahlswede–Korner, as well as
interference and broadcast channels.

By proving new tight finite-blocklength bounds for the best achievable lossy joint source-channel
code rate and by performing their refined analysis, we show that the situation is completely
different at finite blocklengths, where joint design can bring significant gains, both in terms of the
rate achieved and implementation complexity.

Verdú and collaborators have performed a finite blocklength analysis of the channel coding
problem, where channel inputs have costs associated to them, for example, due to limited transmit
power. We showed that the maximum achievable channel coding rate under a cost constraint can
be bounded in terms of the distribution of a random variable, termed b-tilted information, which
closely parallels the notion of d-tilted information in lossy data compression.

We considered the setup where the compressor observes a noise-corrupted version of a source
output (e.g., due to inherent measurement noise), and the goal is to minimize the distortion
between the true source output and its encoded version under a rate constraint. Asymptotically,
this problem is known to be equivalent to the traditional noiseless rate-distortion problem, but non-
asymptotically, a sizable gap exists between the two achievable distortions. The intuition behind this
result is that the noisy source-coding problem has inherently more randomness, which slows down
the rate of approach to the asymptotic fundamental limit.

We also have developed a new method based on Renyi’s information measures in order to obtain
finite blocklength bounds. We have generalized the d-tilted information converse bound to the joint
source-channel coding setting. Exploiting duality between list decoding and lossy compression, we
have shown another interesting converse result based on the best performance of binary hypothesis
testing. Our random-coding based achievability bound provides insights into the degree of
separation between source and channel coding required for optimal performance in the finite
blocklength regime. In particular, the conventional separate source-channel coding (in which the
source code is chosen without knowledge of the channel and the channel code is chosen without
knowledge of the source), which is known to achieve the asymptotic fundamental limit of joint
source-channel coding under certain conditions, is in general only suboptimal in the finite
blocklength regime.

For the transmission of a stationary memoryless source over a stationary memoryless channel, we
derive a Gaussian approximation of the maximal achievable coding rate. Interestingly, the dispersion
of joint source-channel coding decomposes into two terms: channel dispersion and source
dispersion. In those curious cases when the source and the channel are probabilistically matched, so
that symbol-by-symbol coding is optimal in terms of the average distortion achieved, we show that
it also achieves the dispersion of joint source-channel coding. In other words, not only such symbol-
by-symbol codes attain the minimum average distortion but also the variance of distortions at the decoder’s output is the minimum achievable among all codes operating at that average distortion. Even when there is a mismatch between the source and the channel, symbol-by-symbol transmission (though asymptotically suboptimal) might outperform separate source-channel coding and even outperform our random-coding joint source-channel coding achievability bound in the finite blocklength regime.

We have also studied the properties of channel codes that approach the fundamental limits of a given memoryless channel. The output distribution induced by a capacity-achieving code at a given error probability has been shown to be close in a strong sense to the capacity-achieving output distribution (for discrete memoryless channels and additive white Gaussian noise channels). We have also shown that the binary hypothesis testing between the output distribution of a good code and the capacity achieving one cannot be performed with exponential reliability. Using related methods, we have proved that quadratic forms and sums of \( q \)-th powers, when evaluated at code words of good AWGN codes, approach the values obtained from a randomly generated Gaussian code word. We have proved that the random process produced at the output of the channel satisfies the asymptotic equipartition property.

**Semantic Communication and Communication Under Uncertainty**

Sudan has been pursuing the following two main research goals within the context of the Center’s objectives:

- **Semantic Communication**: To know how the sender and receiver can effectively exchange information along with its intended meaning.
- **Communication amid Uncertainty**: To understand how information can be communicated effectively when the sender and receiver are uncertain about each other.

Many researchers seek the answers to these problems in today’s communication contexts. Communicating devices are becoming increasingly diverse. Twenty years ago, the only reason communicating devices might have shown diversity was that they had different manufacturers. Today almost every communicating device is highly programmable. Furthermore, users have the ability to control the exact set of programs that such a device contains. Hence, almost no pair of these devices is identical.

This diversity poses a serious challenge to the reliability of communication. While communicating devices are able to overcome “low levels” of uncertainty, e.g., those introduced by the channel of communication, they are not particularly robust to the “high levels” of uncertainty introduced by the diversity of the communication described above.

Sudan’s two research goals are related. The first considers a broader class of problems, while focusing in particular on not knowing each other’s prior beliefs exactly (but having approximate knowledge of these). The second considers the specific task of not “understanding” each other, but
here it is harder to assume, or even define, approximate understanding. Sudan’s research aims to resolve such questions.

**Research Ahead**

The Center’s future research in this area will proceed in various directions:

As most real-world sources of communication, such as text, images, video, and audio, are not memoryless, finite blocklength analysis of lossy compression for sources with memory has evident practical importance. The core results of our investigators’ thesis—namely, new tight achievability and converse bounds to the minimum achievable source-coding rate as a function of blocklength and tolerable distortion—allow for memory, analysis, and numerical computation of those bounds have been performed only in the most basic setting of lossy compression of a stationary memoryless source. It would be both analytically insightful and practically relevant to derive an analytical approximation to the minimum achievable finite blocklength-coding rate of Markov sources, similar in flavor to the Gaussian approximation we have obtained for memoryless sources.

In the related scenarios of lossless data compression of a Markov source and data transmission over a binary symmetric channel in which the crossover probability evolves as a binary symmetric Markov chain, such approximations have been derived by Polyanskiy and Verdú. In rate-distortion theory, however, such a pursuit meets unique challenges, as even for the simplest model of a source with memory—namely, a binary symmetric Markov source with bit error-rate distortion—the asymptotic fundamental limit is not known for all distortion allowances, let alone its finite blocklength refinements.

It would be enlightening to complement our finite blocklength results on excess distortion with those on average distortion. Most of the existing bounds in vector quantization are asymptotic. The most well-known bound is that, for fixed analog vector dimension, a signal-to-noise ratio achieved by fine quantization grows 6 dB for every increase of one bit; furthermore, uniform quantization is suboptimal by at most 1.53 dB, and scalar quantization of Gaussian sources suffers a penalty of 4.35 dB. Some progress in bounding the minimum average distortion achievable by a vector quantizer of a fixed dimension has recently been announced in July at the IEEE International Symposium on Information Theory.

Even more ground is left untouched in the realm of lossy joint source-channel coding under the average distortion constraint. Indeed, while the redundancy result of Zhang suggests that a similar beautiful expansion should hold in the joint source-channel coding setting, to date, not even a redundancy result of order $O(\log k/k)$ has been shown. Pilc’s achievability bound on a separated scheme leads to a redundancy result of order $O(\log k/k)$.

Systematic codes are those where each code word contains the uncoded information-source string plus a string of parity-check symbols. Shamai, Verdú, and Zamir (1997) characterized the
asymptotically achievable average distortion and found the necessary and sufficient conditions under which systematic transmission does not incur loss of optimality. It would be enlightening to quantify the penalty of systematic communication in the finite blocklength regime.

**Capacity of Sampled Analog Channels**

The capacity of communication channels has largely been studied in the digital domain, under the premise that sampling, if done above the Nyquist rate of the channel bandwidth, preserves information. However, hardware and power limitations often preclude sampling at this rate, especially for wideband channels. Moreover, the Nyquist rate is not always needed to preserve information; recent results in compressed sensing indicate that when signals exhibit sparse structure, they can be sampled below their Nyquist rate without losing information. These ideas give rise to several fundamental questions at the intersection of sampling theory and information theory: How is channel capacity affected by sampling below the channel’s Nyquist rate? What is the optimal sub-Nyquist rate sampling strategy to maximize capacity? For what channels is capacity preserved at sub-Nyquist sampling rates? We aim to explore these fundamental questions and provide some answers about fundamental trade-offs between sampling and capacity.

We recently derived the capacity of sampled analog channels for sampling following filter banks. Connections between sampling and MIMO Gaussian channels were also illuminated based on this analysis, along with optimal prefilters that maximize capacity. We also investigated a more general sampling strategy by adding modulation banks to filter-bank sampling, and showed a connection between this general sampling method and MIMO Gaussian channels. We considered the most general class of sub-Nyquist nonuniform sampling and showed that capacity was not increased by such techniques over uniform sampling followed by filter banks (Figure 1). These results were published in several *IEEE Transactions on Information Theory* papers this year.

![Figure 1: Capacity is monotonic in sampling rate for a sufficient number of filter bank branches.](image)

Recent work of Goldsmith and collaborators, who study channel capacity and sampling strategy, considers the capacity of subsampled analog channels when the sampler is designed independent of
the instantaneous channel realization. For this model, we investigated sampling methods that minimize the worst-case (min-max) capacity loss due to channel-independent (universal) sampling design, with equal power allocation employed over active subbands. We considered a class of periodic sub-Nyquist sampling, which subsumes as special cases sampling with modulation and filter banks. Our results demonstrated that under both Landau-rate and super-Landau sampling, the min-max capacity loss due to universal sampling design depends only on the band sparsity ratio and the undersampling factor with respect to the Nyquist rate, modulo a residual term that vanishes with signal-to-noise ratio. We also quantified the capacity loss under sampling with periodic modulation and low-pass filters, when the Fourier coefficients of the modulation waveforms are randomly generated (called random sampling). Our results indicated that random sampling achieves min-max capacity loss uniformly across all channel realizations, and is thus optimal in terms of the universal sampling-design sense (Figure 2). We presented these results at the 2013 IEEE International Symposium on Information Theory and have submitted them for publication.

Figure 2. Random sampling achieves min-max capacity loss uniformly across all channel realizations: (a) Min-max capacity loss per unit Hertz; (b) normalized min-max capacity loss per Hertz as a function of the sparsity ratio.

We have also recently extended these ideas to rate distortion, considering the rate-distortion performance bounds on Gaussian sources sampled below the Nyquist rate prior to compression. We have recently published initial results on this topic at the Allerton Conference on Communication, Control, and Computing.
Random Coding for Multiterminal Networks

We focus on the problem of obtaining a general achievable region based on random coding arguments for a channel with any number of transmitters and receivers. Through the use of graphical Markov models, we define a compact representation of any scheme combining coded time-sharing, rate-splitting, superposition coding, and binning. From this representation, we can readily obtain the corresponding achievable region. We call this model the chain graph representation of achievable rate regions (CGRAS). It has been successfully applied to several multiterminal models to obtain new achievable regions, including for the MIMO cognitive interference channel (Figure 3) and for cellular systems with relays.

Figure 3: Chain graph representation for MIMO Cognitive Interference Channels

Fundamental Performance Limits of Communication Systems with Feedback

Feedback is known to decrease the probability of error and complexity, and to increase the energy efficiency in communication systems. In this past year, Goldsmith with others in the Center have investigated the achievable error probability in communication over an AWGN discrete time memoryless channel with noiseless delay-less rate-limited feedback. For the case where the feedback rate $R_{FB}$ is lower than the data rate $R$ transmitted over the forward channel, we show that the decay of the probability of error is at most exponential in blocklength, and we obtain an upper bound for the increase in the error exponent due to feedback. Furthermore, we show that the use of feedback in this case results in an error exponent that is at least $R_{FB}$ higher than the error exponent in the absence of feedback, where $R_{FB}$ is the feedback rate of communication.

For the case where the feedback rate exceeds the forward rate, we propose a simple iterative scheme that achieves a probability of error that decays doubly exponentially with the code word blocklength $n$. More generally, for some positive integer $L$, we show that an $L$-th order exponential-error decay is achievable if $R_{FB} > (L - 1)R$. While the above results are proved under an average feedback rate constraint, we show that all the achievability results hold in a more restrictive case where the feedback constraint is expressed in terms of the per-channel-use feedback rate. Our results show that the error exponent as a function of the feedback rate has a strong discontinuity at
$R$, the forward rate of communication, where it jumps from a finite value to infinity. This discontinuity is shown in Figure 4 below.

**Figure 4.** The bounds on error exponents in terms of the feedback rate $R_{FB}$

In other recent work related to communication systems with feedback, we have studied real-time joint source-channel coding of a Markov source over a discrete memoryless channel with noiseless feedback. The encoder incurs a cost, which is minimized along with a real-time end-to-end distortion. We mapped the problem to a partially observable Markov decision problem and derived the corresponding optimality equations in the form of dynamic programming equations. As a consequence of the dynamic programming formulation, we established basic structural properties of the optimal encoding and decoding strategies. In addition, we showed the problem formulation and solution obtained for dynamic joint source-channel coding with noiseless feedback to encompass a much broader class of problems, including that of information acquisition and real time tracking. Earlier this year, we presented this work at the IEEE International Symposium on Information Theory.

**Fundamental Performance Limits of Uncoded Massive MIMO Systems**

In this work we consider a multiple-access channel with a large number of transmitters sending symbols from a constellation to the receiver of a multi-antenna base station. We investigate the joint decoding of the signals from all the users, using a low-complexity convex relaxation of the maximum likelihood decoder (constellation search). We show that, in a rich scattering environment,
and in the asymptotic limit of a large number of transmitters, reliable communication is possible even without employing coding at the transmitters. We also presented this work at the IEEE ISIT.

Figure 5. An uncoded massive MIMO system, which for a sufficiently large number of receive antennas, provides reliable communication, even in the absence of coding.

**Optimal Transport Theory**

The strong connections between optimal transport theory and information theory appear to be of basic importance in developing a theory of control over communication channels. The fundamental (by now classical) description of this connection is in the work of Jordan, Kinderlehrer, and Otto (1998), where it was determined that the heat equation is the gradient flow of the differential entropy in the metric on the space of densities given by the Wasserstein distance. Recently, analogs of this result for continuous time Markov chains have emerged. We are exploring these connections in the context of control over communication channels. One recent result was our discovery of an entropy power inequality (EPI) for sums of group-valued random variables over groups of size a size power of 2. The traditional EPI (for densities) is one of the key discoveries of Shannon and has played a central role in the resolution of some of the most interesting and important communication-theoretic problems, such as determining the capacity region of the MIMO Gaussian broadcast channel. Discovering a discrete analog of the EPI, as we have, is a significant achievement.

During the coming year, we will continue to work on further understanding the relations between optimal transport theory and information theory in the discrete context. The Mathematical Sciences Research Institute at Berkeley is running a semester-long program on optimal transport theory during Fall 2013, which will tap into and benefit this line of research.

**Multiterminal Source Coding**

Since Courtade and Weissman (2012) solved the problem of multiterminal compression under logarithmic loss, Center researchers have been intensively pursuing further development and
application of the tools and the point of view from our investigations to a much wider family of multiterminal source- and joint source-channel coding problems. This direction is exciting not only because it is giving rise to a very rich array of new distributed compression and processing problems—of the type increasingly encountered in the big-data era we are entering—for which the fundamental limits are amenable to characterization (the exception rather than the rule in multiterminal Shannon theory), but also because it is enhancing our understanding of key results in the current literature. For example, it allows us to re-derive simply—and from first principles—many of the celebrated results from the Gaussian case whose existing proofs are highly convoluted and confounding.

**Information Flow in Economic Systems**

The Center also has interest in economic systems because they share many common features with complex communication networks. Such systems consist of multiple entities (agents) with vastly heterogeneous capabilities for acquiring, storing, sharing, and processing information—and with differing degrees of authority for acting upon that information. Yet, for all the commonalities they share with communication networks, one crucial feature that distinguishes economic networks is the fact that agents have objectives that extend beyond simply reliable communication. In other words, in an economic system, information has value. One major challenge in economics is to formalize the notion of information value, particularly in dynamic settings involving multiple agents with different capabilities.

To date, there is no universally agreed-upon definition of information value, although several reasonable alternatives have been proposed since at least the late 1960s. However, if we abstract away the details of these various proposals, we can distill one basic underlying idea: The value of information has to do with the change of the information state of one or many agents, where, broadly speaking, an information state encapsulates all payoff-relevant knowledge available to the agent(s).

Sims and Raginsky have done some preliminary work on designing stationary finite-rate feedback control laws that operate in the presence of information constraints. The optimal control law turns out to be a solution to a rate-distortion problem, but the distortion function must incorporate a correction term that would account for the agent’s forecast of the future information state. These results apply to steady-state (ergodic) dynamics in a single-agent setting; we will build upon them to investigate what happens in networked settings. Moreover, we will explore the effect of transients on the information flow and on the dynamics of an economic system.

Another type of information constraint that arises in economic systems is not with the amount of information, but with its timeliness. The above discussion focused on instantaneous information flow, and the only limitation was the amount of information available to the agent(s) per time step. However, we must also understand the effect of delayed feedback on the dynamics of the information state. In many cases, it is possible to quantify the impact of delay in terms of
information-theoretic quantities, such as mutual information or directed information (the latter is more appropriate in situations involving feedback and causality). For example, we may consider the value of information as a function of the delay and optimize the observation channel subject to capacity and delay constraints. We hypothesize that, in many settings, a diminishing returns effect may exist: the net increase in information value is a monotonically decreasing function of the delay. When the delay is already small, decreasing it further may not yield appreciable information gain. The Center will investigate these and related issues within an information-theoretic framework.

For example, according to one definition proposed by Stratonovich in the Soviet Union and independently by Howard in the United States, the value of one bit of information acquired about a random variable of interest is the largest difference between expected utilities achievable with and without that additional one bit of information. Thus, we can speak of the “best” observation channel that can deliver one bit of information as the one that would provide the largest increase in expected utility relative to what one could achieve without any observation, based solely on prior knowledge.

Of course, now we have to specify (1) the utility function and (2) the nature of the information constraint. The choice of the utility function is conceptually similar to the choice of the distortion function in communication problems, and so we can assume that the utility function is a fixed exogenous quantity. On the other hand, the specification of the information constraint is a separate issue: We can require the observation channel to be implementable by a deterministic quantizer that partitions the state space into two disjoint regions, but then the problem of optimizing the choice of such a channel is fundamentally combinatorial. Alternatively, we may measure information in the sense of Shannon as the mutual information between the input and the output of the channel. If we adopt this mutual information criterion, we end up with the rational inattention framework proposed by Sims. In this case, the problem of optimizing the observation channel is exactly the Shannon rate-distortion problem, and the value of information is given by the Shannon distortion-rate function (DRF) with the negative utility playing the role of the distortion function. Another appealing feature of the rate-distortion viewpoint is as follows: If we are allowed to optimize not only the observation channel, but also the observation space, then it can be shown that the belief state induced by the channel that solves the rate-distortion problem is the best representation. This meshes well with the fundamental fact that the Bayesian posterior is the minimal sufficient statistic.

Of course, the Shannon DRF is only an asymptotic measure of performance, and it needs to be related to an operational criterion via an appropriate coding theorem. In a communication system, the relevant operational criteria pertain to the quality of signal reconstruction at the receiver, and the mutual information constraint is only an asymptotic abstraction of the channel’s reliability by way of the law of large numbers. However, in economics, it is not at all clear which operational interpretation one should attach to mutual information constraints faced by rationally inattentive agents. Finding an appropriate operational interpretation is a central challenge one faces when
applying information theory to economics. Useful applications include such questions as investment decisions (portfolio selection) in the presence of information constraints.

Another interesting and challenging problem related to information capacity constraints is a justification, from first principles, of the emergence of “coarse thinking” or quantization phenomena in situations involving strategically behaving agents. One vivid illustration is in the well-known work by Crawford and Sobel on strategic transmission of information: if one agent must communicate some signal to another agent who will in turn use that signal to select an action, if the action chosen by the second agent affects both agents’ utilities, and if the agents’ utility functions are different, then the optimal (equilibrium) strategy of the first agent is implemented by a quantizer. Moreover, the number of quantization bins increases as the disagreement between the two agents’ utility functions vanishes. Thus, strategic interaction may introduce hard information constraints, and this deserves further study within the rational inattention framework.

However, so far we have been discussing the simplest arrangement: static optimization by a single agent. What happens when we introduce dynamics into the picture? Indeed, any sufficiently complex economic system evolves in time: agents take actions based on information they obtain; these actions may, in turn, affect these and other agents’ information states; they may even affect the quality of any information that may become available in the future. Thus, the choice of an observation channel can be viewed as a control action that alters the information state. In general, the control law has a dual effect: it affects both the utility (or cost) at the current stage and the uncertainty about the information state at future stages. The presence of a mutual information constraint enhances this dual effect, since it prevents the agent from ever learning too much about the state. This, in turn, limits the agent’s future ability to optimize expected utility or cost. These issues need to be thoroughly understood both on a theoretical and on a practical level.
2a. **Center’s Research Thrust Areas**

*Briefly describe the research thrust areas at the Center. Please provide basic information for each thrust area and details of significant accomplishments during the reporting period, including any research partnerships and their contributions to the Center (do not include publications, presentations, etc., that are reported in Section VIII, Center-wide Outputs and Issues). Include in the narrative a discussion of the goals, activities, and outcomes and/or impacts in the current reporting period, if changed from the previous reporting period. Be sure to discuss how the activities in the various research thrust areas enable the Center to meet its goals/objectives described above.*

**Knowledge Management Thrust**

For a massive data set—whether it results from an Internet search, historical, financial, or genetic databases—information and noise acquire new meanings: What is noise for one query may well be useful information for another. The sheer amount of data contained in a database (e.g., the human genome or the World Wide Web) and the presence of conflicting data, may blur patterns and paralyze a user’s ability to perform any useful activity. The problems of discovering, extracting, and quantifying the amount of useful information are at the core of the grand scientific challenges of our time. How do we extract relevant, actionable information? Is there a scientific way to approach this problem? What is “useful information?” What is “learnable information?” How much semantic information can be extracted from massive data sets? These are some of the knowledge management thrust questions that the Center is trying to answer. The following reports on our efforts in this thrust.

**Proving the Correctness of Computations Performed**

One of the crown jewels of modern cryptography is the existence of zero-knowledge (ZK) proofs—proof systems that *reveal validity* of the statement, but nothing more. For example, one could prove that a number is a composite without revealing its factorization; or, with a stronger notion of ZK proof-of-knowledge, that one “knows” factors that he could reveal, if desired.

Such proofs are important building blocks of many cryptographic tools, including secure multiparty computation (Ben-Or et al. 1988), anonymous credentials (Belenkiy et al. 1988), electronic voting (Katz et al. 2001), and others.

ZK proofs of knowledge exist for any nondeterministic polynomial time (NP) problem, but known constructions are practical only when proving statements of special form that avoid generic NP reductions (e.g., pairing product equations [Groth 2006]). Despite recent theoretical and practical progress, obtaining constructions that are both generic and efficient in practice remains a long-standing goal in cryptography (Almeida et al. 2012).
An additional difficulty for studying proofs for arbitrary NP statements is the problem of representation. Proof systems are typically designed for inconvenient NP-complete languages, such as circuit satisfiability or coloring problems on expander graphs, while many of the statements we are interested in proving are best expressed via algorithms written in some high-level language.

Rivest, together with CSiL-supported students, has been working on a project that achieves the first real-world implementation of zero-knowledge proofs for all of NP. Our system also tackles the problem of representation, supporting verification of correct execution of arbitrary C programs. That is, the new proof system allows an untrusted third party to prove that it has returned the correct result to a computation of an arbitrary specified program on a database held by that third party.

While an important stepping stone toward full generality of zero knowledge, the proof system of Rivest’s CSiL team operates in preprocessing model. That is, before proving any statements, the verifier needs to execute—or delegate to a trusted generator—an expensive preprocessing step, namely, obtaining a pair of keys: a proving key and a verification key. This is a one-time step. Once obtained, the keys can later be used to prove and verify an arbitrary number of computations.

In other work, Rivest and collaborators have been looking at some theoretical problems that affect the proof system mentioned above:

- The expensive preprocessing step is necessary because the proving key is essentially an encoded version of one query to a proof string that the honest prover needs to prepare. If the prover knows exactly what is going to be queried, he can adjust his “proof” accordingly, to make the verifier accept an incorrect proof. This preprocessing step could be avoided if there was a way to obliviously sample from the distribution of key pairs.

- The run time for the circuit reduction underneath the proof system incurs a logarithmic overhead. Can this overhead be avoided—i.e., can we achieve linear-time conversion—if we are willing to introduce certain computational assumptions? While interesting in its own right, this question also has deep connections to other problems of complexity theory, in particular, to the existence of algebraic hash functions.

**Knowledge Extraction from Complex Higher-Order Functional Programs**

Jagannathan’s work has focused on techniques to automatically extract and verify useful and expressive safety properties from complex higher-order functional programs. Proving interesting and expressive safety properties of first-order imperative programs typically involves generating verification conditions that can be solved by a first-order decision procedure. In contrast, it is very difficult to find the necessary path constraints in functional programs, due to the presence of the higher-order functions.

This past year, Jagannathan’s team has explored solutions to this problem, which encodes higher-order features into pure first-order logic formulae, whose solutions can be derived by using a
lightweight, counterexample-guided refinement loop. This approach extracts initial verification conditions from dependent typing rules derived by a syntactic program scan. Specifications of higher-order functions are expressed via subtyping chains generated from these types, by treating such functions as uninterpreted first-order constructs.

This technique enables compositional verification inference of useful safety properties in higher-order programs. Moreover, this methodology (1) provides counterexamples that serve as witnesses of unsound assertions, (2) does not entail a complex translation or encoding of the original source program into a first-order representation, and, most importantly, (3) profitably employs the large body of existing work on verification of first-order imperative programs, resulting in a scalable and efficient analysis of higher-order functional programs.

**Knowledge Extraction from Online Consumer Data**

The Center finds multiple problems to solve in extracting data obtained online from millions of consumers. This past year, our researchers have studied ways to predict the number of user visits from advertising, to tag emotions from comments on online news sites, and to improve recommendation systems.

**Forecasting User Visits for Online Display Advertising**

Online advertising is one of the most profitable business models for Internet services. Display advertising is one of the major types of advertising. It comprised 38% of all advertising revenue in 2010 (IAB and PricewaterhouseCoopers 2011). In online graphical display advertising or online display advertising, advertisers buy (explicitly or implicitly) targeted user visits from publishers to promote their products. The products are displayed in graphical (e.g., image, video) advertisements (ads) on popular Web pages. An important problem in display advertising is how to forecast the count of user visits for a Web page (i.e., the number of impressions or ad views) during a particular period of time (e.g., day, hour, etc.). Overforecasting user visit volumes may lead to undesired ad-delivery outcomes, such as missing an advertiser's goal and advertiser attrition. Similarly, underforecasting the visit volumes may result in unsold user visits, often resulting in substantial revenue loss.

The team led by Si proposes to learn user-visit patterns hidden behind the online traffic of a large number of Web pages, by using a series of novel probabilistic latent class models. In particular, Si presents three such models that automatically identify different classes of Web pages and time stamps that share similar patterns of user visits, based on historical data. He then builds a regression model for each class of Web pages and time stamps, resulting in more accurate predictions.

The first latent class model identifies groups of Web pages with similar user-visit patterns. The second latent class model identifies groups of time stamps with similar user-visit patterns. Finally, the third model discovers groups of Web pages and time stamps that jointly share similar user-visit patterns. The proposed models are compared to traditional time-series regression models, as well
as to a large number of baselines that use past user-visit information as the forecast. An extensive set of experiments—along with detailed analysis on real-world proprietary data from Yahoo—shows the effectiveness of the proposed probabilistic latent class models and provides several insights about the performances of different models for forecasting user visits in online display advertising.

**Cross-Domain and Cross-Category Emotion Tagging for Comments of Online News**

With the explosion of social media over the past decade, more and more user-generated data are available on the Web for expressing users’ opinions and emotions. They exist in the form of user reviews of products, services, and businesses; posts on blogs and in social networks; and comments in forums and online news services. Among different types of social media, online news is an important type of information that attracts billions of users who read and actively respond by making comments. Users often express such subjective emotions in the comments as sadness, surprise, and anger. Such emotions can guide the understanding of preferences and perspectives of individual users. As a cumulative effect of the analysis of the users’ feedback and reactions, online publishers may be able to provide users with more personalized services.

To build a general approach to the classification of emotion-tagging comments in multiple domains, it is beneficial to adapt a classification model trained on some online news domains. Cross-domain emotion-classification methods are highly desirable to reduce the cost of manual labeling and the effort for training more models. There are two scenarios when predicting emotions for comments of cross-domain online news. In the first, the source domain and the target domain use the same set of emotion categories; in the second, they use different sets of categories.

Si’s research group proposes effective approaches for cross-domain and cross-category emotion tagging for comments on online news. When abundant labeled data in the source domain are available, a relatively small amount of labeled data in the target domain can be sufficient for training an accurate classifier in the target domain. Based on the relationship between the source and target domains, he has formulated a joint optimization problem to learn the classifier in the target domain by using data from both domains. This appears to be the first piece of research work on modeling cross-domain and cross-category emotion tagging for comments on online news.

Si conducted an extensive set of experimental studies on data from two popular online news services to evaluate the models. The proposed models significantly outperform several alternatives for the task of cross-domain emotion tagging for comments on online news. The models are effective whether the source domain and target domain share the same emotion categories or use different categories.

**Recommender Systems**

With the rise of e-commerce, recommendation systems have received significant attention. Early generations of recommendation systems have been commercialized and have achieved great success. Recommendation systems serve as an important component of various online retail services, e.g., such as Amazon and Netflix. Recommenders give customized suggestions about
purchases—such as books, movies, and other products or services—to users, based on their previous online purchases and preferences data.

With two CSol-supported students and other collaborators, Kulkarni has been investigating various improved techniques for recommender systems. They have proposed a hybrid collaborative filtering model based on a Markovian random walk. This method addresses the data sparsity and cold-start problems in recommendation systems that arise when users have no or little stored preference data.

More precisely, their approach constructs a directed graph, illustrated in Figure 1, whose nodes consist of items and users, together with item content, user profiles, and social network information. User ratings are incorporated into edge settings in the graph model. The model provides personalized recommendations and predictions to individuals and groups. The proposed algorithms are evaluated on MovieLens and Epinions data sets. Experimental results show that the proposed methods perform well compared with other graph-based methods, especially in the cold-start case.

**Model of a Personalized Recommendation System**

![Diagram](image)

**Figure 1.** This graphic model is used to provide recommendations to users. The nodes consist of items and users, together with item content, user profiles, and social network information.

In other work, they have considered other methods to address the case of sparse data. In particular, they propose an iterative collaborative filtering (CF) approach that updates the similarity and rating matrix. The improved CF incrementally selects reliable subsets of missing ratings based on an adaptive parameter and therefore produces a more credible prediction based on similarity. Experimental results on the MovieLens data set show that the new algorithm significantly
outperforms traditional CF, default voting, and singular value decomposition when the data is 1% sparse. The results also show that, in the dense data case, the new algorithm performs as well as state-of-the-art methods.

- **Fundamental Limits of Recommender Systems.**
  In a recent paper, Courtade and Weissman (2013) provide a framework to characterize the fundamental limits of recommendation systems. Specifically, a recommendation system can be broken into two parts: (1) a data-mining part, which extracts relevant customer information from a database, and (2) a recommendation part, which selects the best recommendations based on the extracted customer information. The data-mining part can be viewed as a compression mechanism, i.e., as a large database that has been compressed into a representation that is useful to the recommendation part of the system. As with any rate-distortion problem, a fundamental trade-off exists between the compression rate (i.e., the granularity of extracted information) and the wealth attainable by the recommendation system. Courtade and Weissman present a precise characterization of this trade-off under some mild and realistic assumptions on the statistics of the data. In fact, they solved the problem of distributed multiterminal source coding: lossy reconstruction of two correlated sources, based on separate, limited-rate descriptions of each of the sources. Courtade (2013) provides a complete solution to the problem.

- **Fundamental Limits for Efficient Query and Search.**
  Weissman has been pursuing the development and analysis of models to address fundamental challenges of efficient search in modern massive database systems. His goal is to obtain asymptotic fundamental limits that characterize the trade-offs between accuracy, compression rate, and search efficiency, and to develop practical algorithms that approach these ultimate benchmarks. A concrete problem in this domain is compression for efficient query and search; given a compressed representation, a key goal is to answer search queries about the compressed data. This is in stark contrast to traditional compression, where the data need be merely reconstructible (with or without distortion) from the compressed representation. The approach and models were motivated by discussions with CSoI’s Golan Yona about real problems encountered in biological data integration. The methodology is tailored to distributed database design, but it is also relevant to other compression scenarios involving search within the compressed domain. Efforts are now aimed at exploiting insight obtained from the theoretical results, with the goal of implementing practical schemes. Weissman’s initial results are very promising, and we expect that his findings will impact the design of systems for biological data compression and querying.

**Analysis of Data in the Power Grid**
Kumar has begun working on the problem of data analysis of the large amount of real-time synchrophasor data generated by wide deployment of phasor measurement units in the power grid. To date, Kumar’s group has studied the underlying dimensionality of the data and established it as low dimensional. Moreover, changes in data arising from events are detected very rapidly, which has made feasible the early detection of events, such as line tripping, unit tripping, and control input
change. Kumar’s group is now looking at nonlinear dimensionality reduction, and has begun working in this effort with Yuliy Baryshnikov at the University of Illinois.

**Bayesian Models for Scalable Inference**

Computational analysis has emerged as a powerful tool for examining massive data from various applications. Two key examples are (1) the analysis of high-throughput single-cell data for cancer stem-cell detection, and (2) processing *ab initio* computational data for new battery design. At the same time, massive data of growing complexity impose new computational and statistical challenges—such as high dimensionality, complex interactions, and massive volume—that require new, modern data-analysis methods.

To address these challenges, Qi is investigating the integration of concepts and techniques from Bayesian statistics, optimization, scientific computing, and parallel computing. He and his team are developing principled and scalable computational tools to analyze complex massive data. They are using these tools to discover critical associations between genetic biomarkers and diseases (e.g., Alzheimer’s disease); to capture and analyze biological, social, and atomic network interactions and their dynamics; to detect rare populations of cancer stem cells arising in huge single-cell data sets; and to implement the virtual design of new materials for energy applications.

**Some of Qi’s specific activities include the following:**

1. From network interactions to modules, to phenotypes, and to dynamics—nonparametric Bayes meets matrices and tensors. Many complex systems can be studied from a network perspective; for instance, the study of biological networks has yielded new insights that would not have been obtained via individual gene studies. However, network analysis presents new challenges. First, many real-world networks are believed to be modular, but how do we extract network modules from data and determine the number of modules? Second, how do we link the modules to various properties or phenotypes of the whole system (e.g., cancer progression)? Third, networks (e.g., brain connectivity maps) evolve over time but modeling network dynamics remains elusive. To address these challenges, Qi has developed nonparametric Bayesian models on matrices and tensors.

2. High-throughput analysis—Bayesian inference meets high-performance computing. Gigantic data volume necessitates the use of new parallel computing platforms. One example where Qi has combined Bayesian inference with parallel computing is the detection of cancer stem cells by parallel Bayesian learning on clusters. Qi has built a hierarchical Bayesian model, BayesFlow, and parallelized a coordinate descent-based training algorithm on a computer cluster with hundreds of nodes. Collaborating with Pyne at Harvard Medical School, Qi’s group has successfully identified a rare stem-cell population. The antibodies that define this population perfectly match the literature, validating the effectiveness of their approach. Furthermore, via multilevel parallelization of their code (e.g., MPI and multithreading), they have reduced the analysis time from a week to about ten minutes!
Another example involves extracting text topics by parallel inference on graphics processing units (GPUs). Besides using clusters, we have also designed a parallel Gibbs sampler and a parallel coordinate descent method (Yan et al. 2009) for training Latent Dirichlet Allocation models on GPUs. The new parallelization scheme effectively balances workloads and avoids memory access conflicts. Formerly, it took several days on a high-end workstation for a scheme to learn topics from ten years of the *New York Times*; this is now possible in two hours on an inexpensive GPU card. In contrast, a naïve parallelization on GPUs would have yielded no speedup, due to memory conflicts.

Besides the work in the above examples, Qi is also developing parallel Bayesian inference methods with researchers at Eli Lilly, to identify novel somatic splicing isoforms based on massive RNA-sequencing data (~20 terabytes).

**Quantized Consensus**

With a CSol-supported student and a collaborator, Kulkarni has analyzed a class of distributed quantized-consensus algorithms for arbitrary networks. In the initial setting, each node in the network has an integer value. Nodes exchange their current estimate of the mean value in the network, and then update their estimation by communicating with their neighbors in a limited-capacity channel in an asynchronous clock setting. Eventually, all nodes reach consensus with quantized precision.

The analysis starts with a special case of a distributed binary voting algorithm, then proceeds to the expected convergence time for the general quantized consensus algorithm proposed by Kashyap et al. The new approach uses the theory of electric networks, random walks, and couplings of Markov chains to derive an $O(N^3 \log N)$ upper bound for the expected convergence time on an arbitrary graph of size $N$, improving on the state-of-the-art bound of $O(N^4 \log N)$ for binary consensus and $O(N^5)$ for quantized-consensus algorithms. The result is not dependent on graph topology. They performed simulations on special graphs, such as star networks, line graphs, lollipop graphs, and Erdos-Renyi random graphs, to validate the analysis.

**Discovering Relevant Information by Data Reduction**

One of the goals of Szpankowski’s work is to discover relevant information by data reduction. The increasing use of proxies and of network address translation has made it hard to link Internet actions to their originating source. Sophisticated Web-bots deliberately use randomly varying sets of proxies to access target sites (so that ten simultaneous requests appear to come from ten different random sources). These Web-bots often ignore the site’s wishes to keep them out; they even try to masquerade as human users and to circumvent the bot-detection mechanisms that have been proposed and deployed. Proxies are used for many purposes, e.g., for collecting information about a competing corporation or a sensitive topic—or for simply foiling intrusion-detection and misuse-detection systems. As a result, event-logging mechanisms are unable to distinguish the true origin of an event. Even when the event-logging system knows that only some types of events may ultimately be of interest, it still typically records the other events too, in case they might be discovered to pertain to the events of interest.
In all these situations, whether the data originates as network traffic, in the audit trails of the events in a computer system, through actions by individuals, or as records of a corporation’s financial transactions monitored for internal compliance, many records of events tend to be massive. In haystacks of events lay buried valuable information, whose extraction would be easier if we could reduce the event record.

Consider this concrete scenario: suppose the FBI and its European counterpart collect data related to international money laundering. They store this data in an unstructured file that looks like an audit file. All sources of information are interleaved in a heap of mostly unactionable data. Suddenly, they intercept a lead about a new money transfer. Not much is known except that the FBI file has record of a similar situation. To find the stored file, they must dig out all relevant information from the huge FBI file, which contains mostly irrelevant information. As the first step, they must perform data reduction by de-interleaving all sources contributing to the FBI file, and extracting only those that contribute to the actionable information.

In Seroussi et al. (2012), Szpankowski and collaborators reported on our study of the problem of de-interleaving a set of finite-memory (Markov) processes over disjoint finite alphabets, which have been randomly interleaved by a finite-memory switch. The de-interleaver has access to a sample of the resulting interleaved process, but no knowledge of the number or structure of the component Markov processes or the switch. Szpankowski’s team studied conditions for uniqueness of the interleaved representation of a process, showing that certain switch configurations, as well as memoryless component processes, can cause ambiguities in the representation. We show that a de-interleaving scheme based on minimizing a penalized maximum-likelihood cost function is strongly consistent, in the sense of reconstructing (almost surely, as the observed sequence length tends to infinity) a set of component and switch Markov processes that are compatible with the original interleaved process. Furthermore, under certain conditions on the structure of the switch—including the special case of a memoryless switch—we show that the scheme recovers all possible interleaved representations of the original process. Our experimental results demonstrate that the proposed scheme performs well in practice, even for relatively short input samples.

**Rank Aggregation and Voting**

Milenkovic and collaborators focused on developing novel methods and techniques for ordinal data fusion, sometimes also referred to as rank aggregation. Rank aggregation is a classical problem frequently encountered in the social sciences, web search and Internet service studies, expert opinion analysis, and economics. Rank aggregation plays a special role in information retrieval based on different search models. Examples include cases where users initiate several queries for the information of interest to them, or in situations when one has to combine various sources of evidence or use different document surrogates. The problem can be succinctly described as follows: a set of voters or experts is presented with a set of candidates (objects, individuals, movies, etc.). Each voter’s task is to produce a ranking, that is, an arrangement of the candidates from the most preferred to the least preferred. The voters’ rankings are then passed to an aggregator. The aggregator outputs a single aggregate ranking to be used as a representative of all voters.
Milenkovic’s work addresses important shortcomings of classical distance measures—in particular, the Kendall tau distance—used in rank aggregation, by developing a new family of weighted distance functions. Weighted distances account for the fact that top positions in rankings (votes) are more relevant and more important to the voters; they also address candidates who share qualities that cause them to be viewed as similar by the voters. This work led to the derivation of the new distance measures via a general axiomatic approach, and also to the development of algorithmic solutions for approximating the weighted distance measures and the aggregates based on them with provable constant approximation guarantees. The approach was based on a nonuniform Markov chain method inspired by PageRank and weighted bipartite matching methods.

**Analysis of Network Data**

The Center reports on Neville’s work with collaborators this year on Markov networks for predicting large-scale network data and for improving classification accuracy through a new relational ensemble method.

**Copula Latent Markov Networks for Network Data**

The recent popularity of online social networks and social media has led to increased interest in learning predictive models for “linked data,” where the data comprise not only a set of entities (e.g., users), but also the observed relationships among them (e.g., friendships). Collective classification methods aim to exploit statistical dependencies that are naturally encoded by the observed links in the data. Previous models that were developed for collective classification have been shown to significantly improve predictive performance in many network settings. However, in some scenarios, the models exhibit poor performance that is even worse than conventional methods, which assume the entities are independent and ignore their relational links. This is due to the heterogeneous structure of the network and the availability of labeled attribute information on the nodes, which produces a complex interaction between the model components and the network, resulting in unequal marginals.

To address this issue, Neville and collaborators developed a novel latent relational model based on copulas, which makes predictions collectively while ensuring identical marginals. Copula methods separate the specification of the models for the marginal distributions from the dependence structure that links these to form a joint distribution. This flexibility facilitates new approaches to modeling complex dependencies among entities in large heterogeneous networks. While copulas have recently been used in machine learning for descriptive modeling, they have not been used for prediction in large-scale network data. This is among the very first approaches for the associated conditional inference problem.

Neville developed an approximate inference algorithm to apply the Copula Latent Markov Networks for prediction and applied them to a range of real-world relational classification tasks drawn from online purchasing, social network, social media, and biological domains. The proposed method consistently outperforms alternative collective classification methods, in some cases by a large margin. This demonstrates the validity of our intuitive assumption that network instances are
interdependent but nevertheless should be “identically distributed.” Copula-based methods that can constrain models to meet this assumption are more robust than other collective classification methods and thus result in significantly more accurate predictions.

**Analysis of Collective Classifier Ensembles to Improve Predictions in Networks**

Ensemble methods have been widely studied as a means of reducing classification error by combining multiple models for prediction. Much of the earlier work on ensemble techniques has focused on i.i.d. domains (where objects are independent and identically distributed), and models use exact inference techniques. While there has been some recent investigation of ensembles for relational domains, these previous works have a number of limitations in that: (1) no theoretical analysis exists to show the mechanism by which ensembles reduce model error in relational domains, (2) they focus on the reduction of only one type of error (due to either learning or inference), and (3) they make an assumption about the data sets having multiple relations.

In recent work, Neville and collaborators formulated a theoretical framework to compare the errors made by different relational ensembles and show the reason why some methods do better than others. Specifically, they used bias and variance decomposition for analysis, and extended it for the ensemble setting; as a result, they consider not just a single model, but also an ensemble of collective inference models. Specifically, we reason about two classes of ensemble models: (1) a relational ensemble model that runs the component classifiers independently for inference and aggregates the final predictions, and (2) an across-model approach, which runs the component models simultaneously for collective inference and aggregates intermediate predictions across the models during inference.

Based on their analysis, they found that aggregating the base models’ predictions at the end of inference reduces error due to variance in learning, while interleaving the models’ predictions during collective inference reduces the error due to inference. They also noted that the aggregation process reduces the variance in learning in proportion to the amount of true variance approximated by the learned models.

This led them to consider learning the ensemble in a way that can better approximate the true variance in relational data. Specifically, Neville’s group developed a relational resampling approach to capture the increased variance in relational data, allowing the ensembles to reduce more of the variance due to learning. They then combined that with their previously developed interleaved inference technique, which allows the ensembles to reduce more of the variance due to inference. In contrast to their ensemble method, previous approaches have focused on reducing errors due to variance in learning, or due to variance in inference, but not both simultaneously.

This work showed how their unique combination results in the largest improvement in classification accuracy, compared to previous ensemble methods. In addition, their new ensemble method is applicable to both single-graph and multigraph settings, thereby overcoming the limitation of existing relational ensembles.
Quantum Information Theory

Shor has been working on two grand challenges in quantum information theory. The first is the question of how best to do fault tolerance on a quantum computer. The second is to find the capacity of quantum Gaussian channels.

Quantum Fault-Tolerant Computation

When quantum computation was first proposed, one objection to it was that it could not be made fault-tolerant. The standard techniques for making classical computation fault-tolerant involve either redundancy (duplicating information) or error-correcting codes. It was believed that these techniques would not work for quantum computation because of the quantum no-cloning theorem (Wootters and Zurek 1982). It appears at first that classical error-correcting codes require redundancy, so that quantum error-correcting codes would violate the no-cloning theorem. However, it was discovered that quantum error-correcting codes do exist (Shor 1995; Steane 1996; Calderbank and Shor 1996). Furthermore, these error-correcting codes could be used to design fault-tolerant quantum computers (Preskill 1998). The theoretical results on fault tolerance are usually given in threshold theorems; the exact details of these threshold theorems depend on the architecture chosen for the quantum computer. The threshold theorem says that if quantum gates have accuracy above some threshold, then any circuit for a quantum algorithm can be transformed into a quantum fault-tolerant circuit, and the number of gates increases only by a polylogarithmic factor.

Despite these developments in the theory of threshold theorems, Shor and others are still quite far from achieving a practical fault-tolerant quantum computer. The accuracy required for the first threshold theorems was error less than $10^{-6}$. More recent results have improved this to the order of $10^{-2}$, but with an enormous overhead and using current techniques, fault tolerance with reasonable overhead is believed to require errors less than $10^{-4}$ or $10^{-5}$. These accuracies are still reasonably far from what experimenters can achieve. In experimental quantum computation, there currently is, and likely will continue to be, a tradeoff between the accuracy achievable and the number of qubits that can be built. To make experimental quantum computers useful as soon as possible, what we would like is to find fault-tolerance schemes that work with low overhead and require relatively low accuracy. Shor and others know several completely different methods for achieving fault tolerance, and it appears that they could discover more such methods. The currently known methods all have fairly high accuracy-overhead tradeoffs. The Center’s grand challenge is to discover a new method, or improve on one or more existing methods, to yield fault-tolerance schemes that can be operated at reasonable accuracy and with reasonable overhead.

Quantum Capacity of Gaussian Channels

For classical channels, Shannon’s theorem gives a formula for computing the capacity. This capacity is essentially the only capacity of a classical channel, and appears in many formulae describing the properties of a channel. Quantum channels have many capacities (Holevo and Giovannetti 2012); in particular, one is for the transmission of quantum information over the channel, and one is for the transmission of classical information over the channel. What is more frustrating is that while some
formulas are known for the capacity of quantum channels, we are still unable to calculate the capacity for many simple channels.

Possibly the most embarrassing case of this is the classical capacity of a quantum Gaussian channel, which are essentially those that can be implemented with linear quantum optics elements. We could find the classical capacity of an arbitrary quantum Gaussian channel if we could show that the minimum entropy output of a symmetric single-mode Gaussian channel with thermal noise was achieved on the input of a coherent state (Guha et al. 2008). This is a very simply described channel that is a good approximation to many real-life channels. It can be implemented by putting the input into one port of a beam splitter, and a thermal state of light into the other. While this seems to be a relatively straightforward problem, and there does not seem to be any better way of achieving a low-entropy output than a coherent state, it has eluded proof so far. It would follow from the proof of one of several entropy inequalities. (Koenig and Smith 2012; Guha et al. 2008)

**Privacy and Security in Knowledge Extraction**

The overall goal of Atallah’s work in this area is to design techniques that enable the extraction of useful knowledge from information that is both confidential and distributed among multiple parties who are reluctant to share it with each other. The special focus within this area is the case illustrated in Figure 2 below. The confidential information of a client (Bob) is stored at a remote untrusted site (e.g., in the cloud, see Alice) and the client wishes for the remote server to analyze and manipulate the information on the client’s behalf, yet without revealing anything to the remote server about the stored information or the results. See Figure 2 below.

**Problem of Extracting Data from Reluctant Users**

![Figure 2. Relationship between a client (Bob) who owns confidential information and an untrusted computational entity (Alice).](image)

The most significant result Atallah and his group achieved during the reporting period is the ESORICS paper (published Sept. 2012) because their improvements over the previous state-of-the-art are
spectacular. We found that the client does $O(m+n)$ work and communication, whereas the previous bounds were $O(mn)$; the protocol works in one round, whereas the previous methods took $O(mn)$; the server’s space usage is $O(m+n)$, whereas the previous was $O(mn)$; and the protocol uses only lightweight cryptography, whereas the previous used expensive homomorphic encryption and required oblivious transfer.

We also obtained a significant result in pattern matching. In approximate pattern matching, solutions that rely on algebraic approaches based on convolution naturally lend themselves to the use of existing protocols for securely and privately carrying out these algebraic computations. Several previous randomized algorithms made use of convolution to estimate the score vector of matches between a text string of length $N$ and a pattern string of length $M$, i.e., the vector obtained when the pattern is slid along the text and the number of matches is counted for each position. These algorithms ran in deterministic time $O(kN \log M)$, and found an unbiased estimator of the scores whose variance is $(M - c)(M - c)/k$ where $c$ is the actual score; here, $k$ is an adjustable parameter that provides a tradeoff between computation time and lower variance. The result is an algorithm that also runs in deterministic time $O(kN \log M)$, but achieves a lower variance of $\min(M/k, M - c)(M - c)/k$. For all score values $c$ that are less than $M - (M/k)$, the variance is essentially a factor of $k$ smaller than in previous work, and for $M - (M/k) < c \leq M$, it matches the previous work’s variance. As in the previous work, the estimator is unbiased, and no assumption is made about the probabilistic characteristics of the input or about the size of the alphabet. As desired, the solution extends to string matching with classes, class complements, “never match” symbols, “always match” symbols, the weighted case, and higher dimensions.

Atallah has also addressed privacy-preserving searching of image data, with emphasis on the most computationally expensive part: shape-based feature extraction. Much recent work has focused on the secure and private searching of data stored at a remote server, where an organization stores its data at untrusted remote cloud servers in an encrypted form, such that its own employees can query the encrypted data using weak devices (both computationally and storage-wise), or a weak client wants to outsource an expensive comparison-based searching task without revealing to the servers either the inputs or the computed outputs. The framework requires that the bulk of the computational burden of query processing be placed on the remote servers, without revealing to these servers anything about the data.

Most of the existing work in this area deals with nonimage data that is keyword-based. Atallah’s work deals with raw image data without any keyword annotations. Our team demonstrates that shape-based image-feature extraction—a particularly computationally intensive task—can be carried out within this framework. Two protocols have been proposed. The first is provably secure and is based on garbled circuit techniques. The second is practically secure and allows better response time. We have submitted our results in a research paper, and the results can be used in a number of practical situations. In one scenario, the client has images and wants to securely outsource feature extraction on them. In another scenario, the server has encrypted images, and the client wants a feature-extracted representation of those that are feature-rich.
Opportunities are rich to extract knowledge from high-dimensional, or big, data. Often, one is interested in causal relationships between variables under study. Causal modeling is the foundation for principled causal inference. However, traps are also paramount since opportunities to make mistakes are also rich with high-dimensional data. Robustness and stability considerations become pressingly important to increase reliability of results from data so they qualify as knowledge.

Recognizing the importance of causality, robustness and stability, Yu and collaborators worked on four thrusts of high-dimensional statistical theory: causal network modeling (Uhler et al. 2013; He et al. 2013), robust regression (Bean et al., 2013; Bean, Karoui, Bickel, and Yu 2013), and stability (Yu 2013) as a general concept or framework for reproducible statistical discovery, and Mairal (2013) is on unifying optimization methods. In addition, to ground themselves in real-world problems, they also worked on high-resolution aerosol retrieval (Moon et al. 2013) based on multi-angle satellite images and on using sparse modeling methods in bioinformatics data problems.

**Causal Inference**

In our big data age, causality questions abound. For example: *What genes or epigenetics factors cause cancer? Which web designs cause users to purchase more at a retail site?* Statistical-directed graphical models are the most common vehicles to reach causal conclusions from data.

However, many algorithms or methods for inferring causality rely heavily on the faithfulness assumption. The main justification for imposing this assumption is that the set of unfaithful distributions has Lebesgue measure zero, because it can be seen as a collection of hypersurfaces in a hypercube. Due to sampling error, the faithfulness condition alone is not sufficient for statistical estimation, and strong faithfulness has been proposed as assumed to achieve uniform or high-dimensional consistency. In contrast to the plain faithfulness assumption, the set of distributions that is not strong faithful has nonzero Lebesgue measure, and, in fact, can be surprisingly large, as Yu and collaborators (Uhler et al. 2013) have shown. They investigated the strong-faithfulness condition from a geometric and combinatorial point of view and gave upper and lower bounds on the Lebesgue measure of strong-faithful distributions for various classes of directed acyclic graphs. The results imply fundamental limitations for the PC algorithm and potentially also for other algorithms based on partial correlation testing in the Gaussian case.

As mentioned earlier, graphical models are popular statistical tools used to represent dependent or causal complex systems. Statistically equivalent causal or directed graphical models are said to belong to a Markov equivalent class. The Center has great interest in understanding the space of such classes. However, with currently known algorithms, sampling over such classes is only feasible for graphs with fewer than approximately 20 vertices. In He et al. (2013), Yu’s team designed reversible, irreducible Markov chains on the space of Markov-equivalence classes by proposing a perfect set of operators that determine the transitions of the Markov chain. The stationary distribution of a proposed Markov chain has a closed form and can be computed easily. Specifically, they construct a concrete perfect set of operators on sparse Markov-equivalence classes by introducing appropriate conditions on each possible operator. They provide algorithms and their
accelerated version to efficiently generate Markov chains and to explore properties of Markov-equivalence classes of sparse directed acyclic graphs (DAGS) with thousands of vertices. We found experimentally that, in most Markov-equivalence classes of sparse DAGs, (1) most edges are directed, (2) most undirected subgraphs are small, and (3) the number of these undirected subgraphs grows approximately linearly with the number of vertices. This is also related to recent work of Szpankowski on Markov types (Jacquet 2012).

Robust Statistics in High-Dimensional Regression

It is well known that in high-dimensional regression models, sample-to-sample variability is huge, even with Gaussian predictors and Gaussian errors, resulting in unstable statistical estimates and conclusions. How does this sample variability in the predictors interact with such heavy-tail errors as double-exponential distributions? Karoui et al. (2013) and Bean et al. (2013) begin a novel line of research to answer this question.

In Karoui et al. (2013), Yu and others study in regression M-estimates in the setting where $p$, the number of covariates, and $n$, the number of observations, are both large but $p >> n$. They find an exact stochastic representation for the distribution of M-estimator at fixed $p$ and $n$ under various assumptions on the objective function and our statistical model. A scalar random variable whose deterministic limit can be studied when $p/n = k > 0$ plays a central role in this representation. They discover a nonlinear system of two deterministic equations that characterizes the limiting value of the norm of the M-estimator. Interestingly, the system shows the limiting value depends on $k$ through proximal mappings, as well as various aspects of the statistical model underlying our study. Several surprising results emerge. In particular, we show that when $p/n$ is large enough, least squares become preferable to least-absolute deviations for double-exponential errors.

Bean et al. (2013) is set in the modern setting of high-dimensional statistics, the classic problem of optimizing the objective function in regression, using M-estimates when the error distribution is assumed to be known. Yu proposes an algorithm to compute this optimal objective function that takes into account the dimensionality of the problem. Though optimality is achieved under assumptions on the design matrix that will not always be satisfied, the analysis reveals generally interesting new families of dimension-dependent objective functions.

Stability as a General Concept or Framework for Reproducible Scientific Discovery

Reproducibility is imperative for any scientific discovery. More often than not, modern scientific findings rely on statistical analysis of high-dimensional data. At a minimum, reproducibility manifests itself in stability of statistical results relative to “reasonable” perturbations to data and to the model used. Jackknife, bootstrap, and cross-validation are based on perturbations to data, while robust statistics methods deal with perturbations to models.

In an overview article, Yu (2013) makes a case for the importance of stability in statistics by integrating results from other sources and papers. First, Yu motivates the necessity of stability for interpretable and reliable encoding models from brain fMRI signals. Second, Yu gives strong evidence in the literature to demonstrate the central role of stability in statistical inference, such as
sensitivity analysis and effect detection. Third, a smoothing parameter selector based on estimation stability (ES) for Lasso, ES-CV (cross-validation), is under investigation. ES-CV brings stability to bear on cross-validation. ES-CV is then utilized in the encoding models to reduce the number of predictors by 60% with almost no loss (1.3%) of prediction performance across more than 2,000 voxels. Last, Yu reviews a novel “stability” argument in Karoui et al. (2013) that is seen to drive new results that shed light on the intriguing interactions, alluded to earlier, between sample-to-sample variability and heavier tail error distribution (e.g., double-exponential) in high-dimensional regression models with $p$ predictors and $n$ independent samples.

**Iterative Optimization Methods**

In Mairal (2013), Yu and others study optimization methods that consist of iteratively minimizing surrogates of an objective function. By proposing several algorithmic variants and simple convergence analyses, we make two main contributions. First, we provide a unified viewpoint for several first-order optimization techniques, such as accelerated proximal gradient, block coordinate descent, or Frank-Wolfe algorithms. Second, we introduce a new incremental scheme that experimentally matches or outperforms state-of-the-art solvers for large-scale optimization problems typically arising in machine learning.

**High-Resolution Aerosol Retrieval and Isoform Identification**

In the attempt to tackle real-data problems, Yu and collaborators looked at satellite-retrieved aerosol optical depth (AOD), which can potentially provide a cost-effective way to expand the costly ground particulate air-pollution monitoring network. One of the current state-of-the-art AOD retrieval methods is NASA’s Multiangle Imaging SpectroRadiometer (MISR) operational algorithm, which has the spatial resolution of 17.6 km x 17.6 km. While the MISR’s operational aerosol products already lead to exciting research opportunities to study particle composition at regional scale, its spatial resolution is too coarse for analyzing urban areas, where the air pollution has stronger spatial variations and can severely impact public health and environment. Accordingly, Yu’s group recently developed a novel AOD retrieval algorithm with 4.4 km x 4.4 km resolution, based on hierarchical Bayesian modeling and Monte-Carlo Markov Chain (MCMC) inference method.

In Moon et al. (2013), in collaboration with environmental scientist Yang Liu at Emory University, Yu and collaborators carry out detailed quantitative and qualitative evaluations of the new algorithm, which they call HB-MCMC algorithm, using NASA’s recent DISCOVER-AQ mission data obtained in summer 2011. This data contain spatially dense ground measurements of AOD and other aerosol particle characteristics from the Baltimore, MD–Washington, DC region, a metropolitan area with complex terrain conditions and various aerosol emission types; hence, it serves as excellent validation data for Yu’s new algorithm. Our results show that the HB-MCMC algorithm has 41.1% more AOD retrieval coverage and improves RMSE with respect to the ground measured AOD by 26.8%, compared to the MISR’s operational algorithm. Furthermore, we demonstrate that HB-MCMC algorithm has additional capability of retrieving aerosol particle-size distributions, which is useful for identifying the aerosol types.
2b. **Indicators/Metrics**

2b. Describe how the Center is doing with respect to the indicators/metrics listed above. Include any data that have been collected on the indicators/metrics.

During the current year we have made progress in all aspects of the strategic plan and metrics/indicators, as discussed in item 1b.

**More specifically:**

We formulated several new collaborative research problems in life sciences and communication. In particular, we made significant progress in temporal and structural information as discussed in depth in part II.

In March 2012 and 2013 we held grand challenge workshops to settle on moderate and long term research objectives.

Several centers’ members visited partner institutions. For example, the Center Director visited MIT, Stanford, UIUC and Princeton. Thanks to adding three center-wide post-docs, new collaborations between Stanford, MIT, UIUC, and Princeton as well as UCSD and Berkeley were established. We continue to encourage center members to pursue collaborative research and engage students.

Robust collaborations throughout the Center have been established. For example, Ramkrishna from Purdue led a team of Purdue, Berkeley and UCSD faculty to quantify information flow in biological networks. Subramaniam from UCSD and from faculty from Berkeley initiated a collaboration following the 2012 UCSD workshop led by Subramaniam. Tsachy Weissman, Olgica Milenkovic and Nancy Lynch recently have started collaborating on biological dataset compression. Recently, Tse (Berkeley) initiated collaboration with Sudan (MIT), as well as Verdu (Princeton) who continues working with Polyanskiy (MIT). Kumar (Texas A&M) and Lynch (MIT) are establishing collaboration. Multiple cross institutional and interdisciplinary graduate student research groups have been formed and collaboration continues as a result.

We continue our Prestige Lecture Series on Science of Information, and initiated the first course on science of information taught at Purdue by Mark Ward. Most talks are now being transmitted to all members through the web via adobe-connect. The Science of Information course will now be taught at additional institutions including Carnegie Mellon University and MIT.

We have prepared a proposal for AFOSR that came as a result of our Industry Partners meeting in April. We obtained an NSF TUES grant, and applied for an REU site grant.
2c. **Research Plans**

2c. For annual reports, describe your research plans for the next reporting period with attention to any major upcoming changes in research direction or level of activity. Also, list plans for developing new research partnerships, if any, for the next reporting period.

The descriptions of the three research thrust areas contained in Sections 2.a. where each of our thrusts – Life Sciences, Communications, and Knowledge Management – are described including plans for research going forward. The plans for the next reporting period discussed during our Grand Challenges Workshop in Princeton are summarized in the following section.

**Life Sciences**

Utilize existing data sets for microphage cells, B-cells and cancer patients, initiate efforts on the following:

**Dynamical data**
- Model cellular biochemical pathway modeling, including hypergraph representations of the dynamics and conversion of the hypergraphs into models.

**Many-to-many networks**
- Explore methods for determining the minimal network for quantitative information on phenotypes and for assessing the sensitivity of network connections to phenotypes.

**Modularity in networks**
- Investigate and develop algorithms for identifying modularity in networks.

**Genome encoding and evolution**
- Work with genomic data, which is characterized as heterogeneous, noisy and missing data, to understand and reveal the underlying robust networks.

**Communication**

Continue with efforts on the following:

**Delay in information**
- Quantify the temporal value of information, develop information theory for finite block lengths, and identify tradeoffs between delay, distortion, and reliability in feedback systems.

**Information in networks**
- Develop an understanding of the communication between dynamic agents and between agents and the physical world; identify new classes of problems related to scaling laws and the microscopic-macroscopic transition.
Information and computation

- Quantify fundamental limits of in-network computation, and the computing capacity of networks for different functions; develop an understanding of the complexity of distributed computation in wireless and wired networks; study of aggregation for scalable query processing in distributed databases; define fundamental limits of interactive communication.

New measures and notions of information

- Develop new measures that quantify and represent soft information, semantics in information, and modern communication networks, such as the Internet and its overlaid networks (e.g. social networks).
- Collaboration with the Life Science Thrust

Knowledge Management

Perhaps the greatest challenge in applying information-theoretic principles to a broader suite of problems – including biological systems, analytics for massive datasets, and social networks – is that of developing meaningful notions of “structural information” and establishing a set of corresponding fundamental results. Specifically with respect to the analysis of big data—the data size and complexity leads to two main challenges with accuracy/efficiency tradeoffs:

Discovering Structure in Data

To identify, encode, and test the underlying structure of “big” data, there is a tradeoff between accuracy of the data model and the amount of data needed to support/test the model.

Managing and Querying Data

To collect, store, and query “big” data in data management systems, there is a tradeoff between the efficiency (space) of the data structures and the efficiency (time) of the algorithms that query and access the data.

Ranking findings and finding rankings in very wide data set

A very wide data set contains more features than records. By a finding we mean here an association rule, a correlation, or a pattern. By a ranking we mean a function which associates a measure of validity/interest with a finding. Such ranking function may return a p-value combined with some measure of interest of a finding.

We finalized a draft of our Grand Challenge document that is annually updated and reviewed. It serves to define the grand challenges that the Center will work on and used as a guide for center-wide projects that are fostered through these new collaborations.
III. Education

Vision, Mission, Goals & Key Outcomes

1a. Center’s Education Goals

 Describe the Center's overall education goals and/or objectives. If the Center’s overall education goals/objectives changed since the last reporting period, how did they change and why? [In section 2 below, please describe progress the Center has made toward reaching these goals or objectives.]

Vision

Our long term vision is to educate the next generation of scientists in various STEM disciplines with the core fundamentals, methods, algorithmic thinking, and understanding of applications in the Science of Information. Moreover, CSi will be a leader in increasing diversity among students, scientists and the STEM work force for disciplines related to the study of the science of information.

Mission & Goals

Our mission is to integrate cutting-edge, multidisciplinary research and education efforts across the Center to advance the training and diversity of the work force

In our efforts to achieve the Education vision and mission for CSi there are three overarching goals, each with supporting activities. The goals of the education team are as follows:

1. Educate students in the Science of Information.
2. Integrate research and education through collaboration.
3. Establish a community of practice around the emerging field of the Science of Information.
Figure 1. This graphic shows each goal with supporting activity categories. Metrics and measured outcomes for this period are detailed in Table 2, metrics section 2e.

**Key Outcomes for the Education Program this Period**

- Four multidisciplinary student research teams were formed around problems connected with the Center. Seven universities and as many departments are involved in the four teams, showing interdisciplinary interactions. [http://www.soihub.org/research-teams.php](http://www.soihub.org/research-teams.php)

- For our annual summer school we organized the IEEE IT Society North American School. There were 140 students and faculty that participated, with 56 oral and poster presentations by the students. [http://www.soihub.org/summer-school.php](http://www.soihub.org/summer-school.php)

- Including the Introduction to Science of Information course, faculty developed and taught 10 new courses, and adapted SoI content into 8 existing courses for both undergrads and grad students reaching more than 700 students enrolling.

- Student brown bag presentations (9 total) reached 161 students and postdocs this year. [http://www.soihub.org/seminars.php?area=edu&type=22](http://www.soihub.org/seminars.php?area=edu&type=22)

- An online Learning HUB has been developed to host our course and module topics to make them widely available to students and faculty. 18 new modules, including 80 video tutorials were produced. [http://learninghub.soihub.org](http://learninghub.soihub.org)

- An online (MOOC) version of Introduction to Science of Information has been developed. Beta testing with a small cohort is taking place this winter. It will be offered for enrollment starting in spring semester 2014 using the new Learning HUB.

- Two faculty workshops on teaching science of information resulted in 52 attendees. 8 faculty from this group were selected to develop and teach 5 new SoI courses and 19 topic modules, all of which will populate the new Learning HUB.
• Views of video tutorial and seminar content quadrupled with 12,746 views in period 4 (as of Sept. 17), compared with 3,185 views in period 3.

• Sixty of our students and postdocs reported publishing 149 papers, and 131 posters and oral presentations. Our students were recognized with 27 awards for these efforts.

• We increased the number of student participants involved in the Center (138 undergrad, grad, postdocs combined participating, with 106 received financial support during the period)

• 97% of students and postdocs polled say their participation in the Center has been valuable to them because the Center stimulates their own thinking about the science of information, provides ample networking opportunities with leading scholars in the field, and fosters collaborations amongst students and faculty. Two example testimonials are below (see additional student comments section 2e.) Student Testimonials of Center Activities).

“*The Center has provided a good platform for looking at my research from a broader perspective. It has also provided easy access to faculty and students working in areas related to mine, and their expertise has been very helpful. And finally, I enjoy the opportunity to mentor undergraduates.*”

“The CSol seminars (including the weekly brown bags) are instrumental in providing novel perspectives and inputs. I believe that my experience with the CSol will help me gain an edge in my future career.”

Community of Practice Framework

Learning occurs both individually and in multi-layered networks of mentoring. The education team focuses on conducting events to build a learning community that values individual scholarship, and meaningful collaboration. The Center community is a community of practice*, collectively learning about doing research within the emerging field of the Science of Information.

Students actively construct ideas about the science of information through their participation in the CSol Summer School, CSol research workshops, and within research labs of individual center faculty. Because students have been able to deepen their understanding of the science of information while connecting with others, four student teams responded with plans to embark on collaborative research this past period.
Graduate Student Expectations

The Center principally delivers value to students through diverse voluntary activities. Building a successful community of practice requires both public and private spaces for people to learn about the domain, develop skills associated with the practice, and participate regularly within the community. In 2013, the Center encouraged students to engage with the Center in a variety of ways. The Center hoped funded graduate students in their first or second year would complete the CSO1 summer school, and all students should attend at least two online research seminars, participate in at least one CSO1 research meeting, and present their research at least one conference.

1b. Education Performance and Management Indicators and Metrics

Inform us of the performance and management indicators (click for definition) the Center has developed to assess progress in meeting its education goals/objectives, if changed from the previous reporting period.

Broadly speaking, activities coordinated through the education program work to expand graduate students’ capacities to conduct research in the Science of Information. Some programs help students develop the critical knowledge base necessary to engage the Science of Information. Other activities, such as forming student research teams, provide platforms for advanced students to engage in interdisciplinary projects.

This year, the education program organized a number of opportunities for students to learn about the science of information, present their research to peers and faculty, and form project research teams. We have collected quantitative and qualitative data enabling the Center to gauge outcomes for each of the three overarching goals, and their supporting activities against the metrics of the education logic model (see Table 2 in the metrics section 2e.).

1c. Problems Encountered

Discuss any problems you may have encountered in making progress toward the Center’s education goals/objectives during the reporting period as well as any problems anticipated in the next period as appropriate. Include your plans for addressing these problems.

As in previous years, an ongoing challenge is learning about new students and postdocs added to the Center at our partner institutions, as we don’t often learn about these new students until annual faculty reports are sent in. Thus, they miss out on Center communications, events, and activities. One way we are attempting to learn about new students being supported on Center funds is through reports from the respective business offices. In addition, students are sometimes moved on and off of Center funds sporadically and this is difficult to track. Thus, a certain amount of follow-up
is necessary, yet time consuming. It is much appreciated when faculty communicate student and postdoc staffing changes to us as they occur.

Internal and External Educational Activities

2a. Internal Goals and Activities

Describe the Center's internal educational activities in the reporting period. Include in the narrative a discussion of how the various internal education activities enable the Center to meet its education goals/objectives described above.

The Center brings together a vibrant community of graduate students, post-docs, and undergraduate students. This year, the total number of active participants in this learning community grew slightly to 138, up from 129 in the last funding period. Student willingness to participate in Center-sponsored activities indicates that 1) these students intrinsically value CSol activities and 2) CSol reaches diverse populations of students by providing appropriate information and fostering a welcoming environment. These students choose to participate in the Center events to learn more about the science of information and to network with Center faculty and students. 32 students and postdocs not currently funded by the Center participated at the level of a funded graduate student. 36% of students reported collaborations on publications and presentations with other CSol members beyond their major professor. Table 1 shows an overview of participation and outputs by students in select activities.

<table>
<thead>
<tr>
<th>Period</th>
<th>Students &amp; Postdocs</th>
<th>Summer School Students</th>
<th>CSol internal research seminar attendees</th>
<th>Research Workshops</th>
<th>Research conf. presentations (oral + posters)</th>
<th>Publications</th>
<th>Degree of Collaborations Reported</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012-13 (period 4)</td>
<td>138 (+7%)</td>
<td>140 (+226%)</td>
<td>208 students/ 8480 (via the Web) (+166%)</td>
<td>99 (+153%)</td>
<td>131 (-32%)</td>
<td>149 (+&lt;1%)</td>
<td>36% of all students (+140%)</td>
</tr>
</tbody>
</table>

Table 1. Comparisons in number of students/postdocs attending select activities and scholarly outputs in the last reporting period over the previous period.

**Goal 1: Educate students in the Science of Information (Internal)**

The first goal of the education team is to educate students in the Science of Information. The Center coordinates diverse internal activities to expand students’ abilities to conduct independent research. Our signature program is our Annual Summer School. This year Purdue was the host location for a 4-day summer school with 140 post-docs, graduate students, and faculty. Due to the
nature of the much larger school this year with the partnership of the IEEE IT Society, the large majority of participants were external to the Center (38 Center, 102 External). Thus, this activity is reported in the External Activities section, 2c.)

The other major activities supporting the goal of educating students in the Science of Information pertains to developing science of information curriculum with new courses, new online topic modules, and infusing existing courses with the latest SoI knowledge and techniques. Due to the nature of reaching a much wider audience beyond the Center members, these curriculum and teaching efforts are described in the external activities section 2C.

**Goal 2: Integrate Research and Education through Collaboration (Internal)**

The second goal of the education team is to integrate research and education through collaboration. The Center seeks to involve students in the process of defining a research agenda around grand challenges in the Science of Information. Signature programs include Student Research Teams, and Virtual Brown Bag Student Seminars.

**Student Research Teams**

Following two focused research meetings and professional development in 2012, the education director formed and supported four interdisciplinary, multi-institutional student project teams. The teams officially started September 1, 2012. Each team presented overviews of their projects plans to their peers during the student research workshop on December 3. This allowed for a wider discussion and input from their peers as the teams were getting started. Specific research details and outcomes are reported throughout the Research section of this report. Research outcomes reports from each of the four student research teams are available at [http://soihub.org/research-teams.php](http://soihub.org/research-teams.php)

**Team 1: Understanding Information-Energy Interactions**

Post-doc PI/advisor: Pulkit Grover (EE, Stanford University & now Assistant Prof. at Carnegie Mellon, ECE)

Student PI: Karthik Ganesan (EECS, UC Berkeley, advisor: Jan Rabaey)

Student PI: George Alexandrov (EE. Stanford, advisor: Andrea Goldsmith)

This team brings together an information theorist with circuits and power transfer students. The team posits that understanding information-energy interactions is a tremendous opportunity for information scientists to expand the scope of traditional information theory. Problems they are researching include energy in communications and computation with error-correcting codes that attain minimum total communication power, and minimizing energy in computing the Fourier transform. In addition they are interested in wireless information and power/energy transfer looking at the concepts that build on ideas from Shannon and Tesla.
Team 2: Investigation of Metabolic Phenomena Using Information Theory

Student PI: Frank DeVilbiss (ChemEng, Purdue University, advisor: D. Ramkrishna)
Student PI: Pablo Robles-Granda (CS, Purdue University, advisor: Jennifer Neville)
Student PI: Mohan Gopaladesikan (STAT, Purdue University, advisor: Mark Daniel Ward)
Faculty Advisor: D. Ramkrishna, Chem Eng Purdue University
Faculty Advisor: Maxim Raginsky, ECE, University of Illinois Urbana-Champaign

An organism performs a range of metabolic reactions with the goal of survival, but the control goal(s) driving metabolism are not fully understood in this context. This team brings together a chemical engineer, an algorithms and random structures student, and a machine learning student with an information theory professor to address the team’s primary question of “does the sum of metabolic regulation converge upon the control goal of carbon uptake at maximum rates?” They will attempt to look at dynamic control goals that guide metabolic function as represented in complex sets of bioinformatics data. The team will first address how to best analyze gene expression data as a Big Data problem.

Team 3: Graph Inference based on Random Walks

Post-doc PI/advisor: Thomas Courtade (EE, Stanford/Princeton University)
Student PI: Victoria Kostina (EE, Princeton University, advisor: Sergio Verdu’)
Student PI: Suvidha Kancharla (CS, Purdue University, advisor: Jennifer Neville)
Faculty advisor: Jennifer Neville (CS & STAT, Purdue University)

This team brings together machine learning and information theory researchers. Estimating models of graph data from network samples is a task of fundamental importance for many problems today. For example, computing page rank for search engines, or peer influence in social networks both rely on an estimate of the underlying network model. However, in many cases the underlying graph is unknown, and must be estimated through sampling. By performing random walks several questions arise in terms of inferring the accuracy of the structure of the underlying graph, and what tradeoffs are there between compression rate of the data and the fidelity of the graph being reproduced?

Team 4: A Fresh Look at Boolean Functions

Post-doc PI/advisor: Thomas Courtade (EE, Stanford/Princeton University)
Post-doc PI/advisor: Pulkit Grover (EE, Stanford University & now Assistant Prof. at Carnegie Mellon University, ECE)
Student PI: Madars Virza (EE & CS, MIT, advisor: R. Rivest)
Student PI: Gowtham Kumar (EE, Stanford)

Nearly all information theoretic techniques rely on measure phenomena and non-asymptotic problem formulations are therefore much more difficult to deal with. New techniques and tools should be developed to address these problems. The primary goal of this team is bringing together researchers from the fields of information theory and theoretical computer science to take a fresh look at Boolean functions. Such problems have recently attracted computational biologists due to cell functions lending themselves to Boolean type representations.
Outcomes of Student Research Teams and Lessons Learned

The four teams were collectively very productive in the first year with a total of 18 presentations (posters and oral) at conferences, and 6 publications (published & in progress). Lists of each are below. In terms of collaboration lessons learned for these teams are that 1) frequent meetings, both face-to-face and virtual, are necessary for success, and 2) all teams alluded to the need for more direct involvement from faculty, including at the proposal stage of team formation. This will help in advisors plans for their students in terms of time management. Funds were available for each team to travel for meetings, and to present at conferences. However, in the case of two teams, timing was not always favorable for the members. The other two teams were able to take advantage of the funds and spent considerable time working together face-to-face.

Team participants mentioned personal growth in professional development this experience has given them, including the writing of their team project proposals, tackling interdisciplinary problems, learning to work in a team environment, along with reporting and presenting results.

Team Presentations (posters or oral)

Info-Energy Team

[1] ITA ’13: Talk about new fundamental limits on total (transmit + computation) energy consumed in communication obtained via consideration of informational-friction.


[3, 4, 5, 6, 7, 8] Seminars at CMU, IIT Bombay, Tata Institute of Fundamental Research (Bombay), LNMIIT (Jaipur, India), and presented two seminars at IIT Delhi, and National Brain Research Center (New Delhi, India).

Boolean Team


[10] G. Kumar and T. Courtade. “Which Boolean Functions are Most Informative?” Poster presentation at the 2013 Information Theory and Applications Workshop (ITA) at UCSD.


**Graph Inference Team**

[16] S. Kancharla, “Graph Inference Based on Random Walks.” Poster presentation at the December, 2012 NSF site visit at Purdue University.

**Metabolic-Info Team**


**Team Papers Published or in Progress**

**Info-Energy Team**


**Metabolic-Info Team**


**Virtual Brown Bag Research Discussion Series**

Throughout the year, nine student brown-bag discussions were convened where students presented high-level talks about their research that collectively spanned all three thrust areas of the Center. 161 students, faculty, and staff participated in the brown bags, for an average attendance of 17.

The virtual brown-bag seminars help students continue high-level conversations about their research. Using an Adobe Connect interface, presenters engage their audience in discussion.
throughout the presentation. Students and faculty join the seminar to learn about cutting edge research within the Science of Information. Titles included:

1. Rotem Oshman, M.I.T./U of Toronto - Data Aggregation in Wireless Networks (12 attendees).
2. Rui Ma, UC San Diego, Neuroscience Tutorial for Information Theorists (37 attendees).
6. Farzaneh Farhangmehr, UC San Diego, Information-theoretic Approaches for Biological Networks Reconstruction (21 attendees).
7. Madars Virza, M.I.T., Two problems in integrity verification (9 attendees).
8. Christine Task, Purdue University, Defining Differential Privacy for Social Network Analysis (7 attendees).
9. Grace (Xiao) Yang, Yale University, Large Alphabet Compression and Predictive Distributions through Poissonization and Tilting (19 attendees).

**Internal Seminars for Center Students**

David Tse (UC Berkeley) and Tsachy Weissman (Stanford) each gave advanced keynote tutorials to Center students during their annual student meeting to 33 students and postdocs. Titles included:

- David Tse - Optimal Whole Genome Shotgun Assembly: From Simple Models to Complex Data
- Tsachy Weissman - Fundamental Limits and Algorithms for Some Modern Data Processing Problems

**Internal Poster Sessions**

An internal research poster session was held for Center students and postdocs during the all-hands and NSF site visit meetings. Center students presented their research for peers, faculty, and NSF Review Board and staff (54 attendees). A second, larger effort was held in June with three days of poster sessions at the summer school. Those sessions are reported in the External activities section.

**Workshops and Collaboration**

Workshops described below, together with the student research teams, poster sessions, seminar series, and summer schools, have led to building a community in the Center where even students are now collaborating on their research together. This period, 36% of students reported presenting research in collaboration with other Center students and postdocs, up from 15% reported in period 3.
Several research focused meetings were held with student involvement. A primary meeting was organized around an All-Hands meeting last December. This was a follow-up meeting to the summer student research workshop.

Grad Student/Post Doc Research Workshop (33 student attendees).

CSol Big Data Workshop (20 student attendees).

CSol / Bell Labs Research Workshop – Princeton (3 student attendees).

Purdue Science of Information Research Discussion (19 attendees).

**Goal 3: Establish a community of practice around the Science of Information as an emerging field of study. (Internal)**

The third goal of the education team is to establish a community of practice around the Science of Information as an emerging field of study. This is a long term goal and the Center supports this goal with accrual of outcomes over time from Goals 1 and 2. In addition, a community of practice is being established through professional development activities, outreach to broader communities, and establishing the teaching of courses and modules. Much of this effort reaches a wider audience beyond the Center and is reported in the External activities section. In addition to the Professional Development activities (reported in section 2.b) the three primary internal efforts supporting this goal are:

**Mentoring**

Faculty members mentoring students more broadly throughout the Center, and this extends to students’ own peer-to-peer mentoring, as well as mentoring undergraduate students. 25% of our students report providing on-going mentoring to others. Representative comments about mentoring from faculty in the Center include:
“Although Ramki is my major advisor at Purdue, Maxim Raginsky at Illinois helped me greatly in developing a model comparison framework for systems biology. This has been mentioned in the publication in Current Opinion in Chemical Engineering and will be the subject of an upcoming publication.”

“I have received significant research mentoring from T. Weissman and S. Verdu. Also, both of these individuals have provided mentoring with respect to professional development over the past year.”

“Prof. A. Grama suggested directions for applying graph algorithms to biological datasets with the final target of extracting knowledge from the outcome, also advising for extending the applicability of the methods to a broader context, e.g. Web and social network datasets. Prof. W. Szpankowski emphasized the formally provable aspects of the approaches taken, suggesting the theoretical analysis of well-established graph models.”

“Prof. Todd Coleman and Prof. Venkat Anantharam provided mentoring on information theory and systems theory. The mentoring is very insightful, and provided important guidance leading me towards mature and independent thinking from a systems engineering perspective. With the mentoring from my CSOI center advisors, I’ve been able to apply principles in feedback information theory and dynamic systems theory to design my neuroscientific experiments. In this regard, Prof. Jack Gallant provided important guidance on the fundamental neuroscientific questions I should aim at answering. This unique combination of mentoring from multiple disciplines has greatly deepened my understanding and improved my capability to design and conduct research independently.”

“I had very useful discussions with Yury Polyanskiy, Madhu Sudan and Tsachy Weissman in the context of the problems I had been working on.”

“I have received excellent mentoring from Dr. R. Claudio Aguilar. I have also interacted with several CSOI faculty during summer schools and seminars and am grateful for their insights and suggestions. I received mentoring for Prof. Todd Coleman. He has been very generous in training me on topics I am interested in, which will benefit the current project, but also bolster my skill set and contribute to future careers goals. He has been incredibly approachable and supportive.”

Student-Postdoc Council

In addition to giving students voice in advancing cutting-edge research through research workshops, the Center also welcomes two-way communication from students through the Student Council. 12 students from seven institutions served on the Student Council.

The council has four primary objectives:

1. Stimulate peer-to-peer communication between students, and the Center staff. Members serve as a Center liaison for the other students at their institution.

2. Plan an all-hands student retreat/workshop annually for face-to-face student collaboration, poster session, and learning about student-conducted research within the Center.

3. Serve as framework for the students to cohesively have input into Center programs.
4. Decide and promote activities or opportunities the council wants to pursue that add value/will be beneficial to all students in the Center.

**Current Students and Alumni**

Evidence gathered over the first four periods indicates a growing community becoming established around science of information. Students have learned about problems, new concepts and techniques in summer schools, and workshops, and exchanged ideas through poster sessions and team discussions. These forums and interactions are leading to increased capacity of Center students to collaborate on multidisciplinary problems. These young scholars will form the backbone of the future science of information research community. The Center has now graduated a total of 84 members (undergrads, graduate, postdoc), with 37 in period 4 (32 in period 3, and 15 in period 2). Many alumni have found industry positions – some in the same company and location (i.e. Google, Facebook, IBM, Microsoft), however, many continue in academia. In some instances, they remain connected and involved in the Center. Two examples include Pulkit Grover, now an Assistant Professor at Carnegie Mellon University is developing in cooperation with the Center an advanced course and three topic modules. Tom Courtade has accepted a faculty position at UC Berkeley and will remain connected with the Center, and has expressed interest in teaching an SoI course in the future. Individual alumni can be viewed, along with links to their respective PhD theses (if available) at http://www.soihub.org/people.php?groupID=11. Alumni can join our csoi-friends listserv to stay abreast of center developments. Our alumni provide a valuable connection for current and future students.

**2b. Student Professional Development Activities**

*Summarize the participation of Center students in professional development activities in the reporting period. Include in the narrative a discussion of how the various professional development activities enable the Center to meet its goals/objectives and produce meaningful results.*

**Professional Development**

**Presentations, publications, & proposal writing:**

The student research workshop last summer involved 21 students who learned about the nuances and benefits of writing a proposal. Ten of these students undertook the proposal writing process as part of the student research team effort, and learned the basics of grant writing and reporting through this experience. Professional development also occurs through the process of students presenting their research at conferences, summer school, and brown bag center discussions (131 conference presentations, 9 brown bags, 56 oral – summer school, 56 – posters – summer school, 31 posters all-hands site visit, and through peer-reviewed publications (149 publications; listed in publication section). In addition to the training students receive from faculty and postdocs, they receive access to a number of professional development options through the diversity program. Two examples are constructing an effective CV, and understanding diversity in the hiring process.
Together, these professional development activities increase students’ capacity to conduct top level research, but also the understanding and experience to work across disciplines and in high stakes groups tackling grand challenge issues. This is at the core purpose of the Center.

The Center organized two workshops for faculty to learn about and help further develop the SoI teaching curriculum. These are reported in the External educational activities section.

**Internships**

The following students reported internships this period:

1. Naresh Rapolu, Summer Research Assistant, NEC Laboratories, Princeton, NJ.
2. Christine Task, Graduate research intern, Naval Research Labs Information Technology Division.
3. Mohan Gopaladesikan, Summer intern at MED Institute, West Lafayette, IN.
5. Tim LaFond, Intern at Lawrence Livermore National Labs Cyber Defenders.
6. Madars Virza, Internship with Prof. Eli Ben-Sasson at Technion-Israel Institute of Technology.
7. Rahul Singh, Bell Labs researcher in Alcatel Lucent Bell Labs, Murray Hill, NJ.
8. April Harry, Public Health Research Co-op, IBM Almaden Research Center, San Jose, CA.
9. Hongwei Li, Summer intern at Microsoft Research, Silicon Valley.
10. Himanshu Asnani, Engineering Intern, BNET DUIB IP SYST SYSTEM & TECHNOLOGYERICSSON.
11. Shang Shang, Research Scientist Intern, Amazon.
12. Idoia Ochoa-Alvarez, La Caixa Fellowship
13. Alexandros Manolakos, Summer Intern in Corporate R&D, Qualcomm, San Diego, CA.
15. Shumiao Wang, Research Intern, Microsoft Research.
16. Yuxin Chen, Research Intern, Microsoft Research.

**Teaching**

21 of our students reported teaching or serving as a teaching assistant for 30 undergraduate courses and one quarter of Center students report serving as a mentor for individual students. These activities have many benefits in terms of increasing their ability to explain difficult subject matter to both individuals and groups, as well as build their leadership and management skills.

1. Sudeep KamathEE20: Introduction to Signals and Systems
2. Jaewoo Lee Data mining (TA) 2013 Spring, Cryptography (TA)
3. Han-Hsuan Lin 8.06 Quantum Physics III, TA
4. Frank DeVilbiss ChE 456 - Process and Control
5. Philip Ritchey  
   CS 573 - Data Mining (TA)  
   CS 426 - Computer Security (TA)  
   CS 543 - Introduction to Simulation and Modeling of Computer Systems  
   (Guest Lecture)

6. Giorgos Kollias  
   CS525: Parallel Computing

7. Arpita Sen  
   Eukaryotic Cell Biology, teaching assistant.

8. Dustin Stansbury  
   Graduate Student Instructor for Computational Models of Cognition

9. Idoia Ochoa-Alvarez  
   TA for the information theory course and the Statistical Signal Processing

10. Albert No  
    TA of EE376A (Information Theory)  
    TA of EE378A (Statistical signal processing, part A)

11. Naresh Rapolu  
    (TA) for CS525 -- Parallel and Distributed Computing

12. Cagri Goken  
    TA for "Networks: Friends, Money, and Bytes"

13. Hongwei Li  
    Stat 241A Graphical models

14. Kavitha Mukund  
    Systems Biology and bioengineering 1- TA  
    Applied Bioinformatics-  
    TA - Introduction to networks and systems biology- Applied  
    Bioinformatics course- Guest lecture

15. Adam Bloniarz  
    TA - Stat 133, UC Berkeley  
    TA - IMA New Directions Short Course on Machine Learning, TA - Stat 215A, UC Berkeley

16. Marianne Catanho  
    TA for BENG 125 Modeling and Computation in Bioengineering

17. Alexandros Manolakos  
    TA in "Wireless Communication"

18. Gowtham Kumar  
    TA, EE476

19. Madars Virza  
    Teaching Assistant for MIT's 6.875 - Cryptography and Cryptanalysis

20. Victoria Kostina  
    Lecturer for Information Theory Forum seminar at Stanford

21. Michael Forbes  
    Lecturer for MIT 6.045 Automata, Computability, and Complexity

2c. **External Education Goals and Objectives**

*Describe the Center's external educational activities in the reporting period.*  
*Include in the narrative a discussion of how the various external educational activities enable the Center to meet its goals/objectives and*

The Center for the Science of Information is committed to establishing a community of practice around the science of information as an emerging field of study that extends far beyond the Center. The Center extends its reach through organizing lectures, special events, summer schools, along with new courses and topic models, and making this content available through the Learning HUB on soihub.org. Center faculty taught classes (18 reported) engaging hundreds (717 reported) of students with new knowledge found in the science of information. Additionally, students affiliated with the Center act as ambassadors for the science of information when they attend diverse conferences.
The first goal of the education program is to educate students in the Science of Information. The Center coordinates several external activities to introduce students to the Science of Information.

**Summer School**

The sixth annual North American School of Information Theory took place June 4-7, 2013 on the campus of Purdue University. Hosted by the Center for Science of Information (http://soihub.org), 140 total students, postdocs, faculty, and professional staff took part in the school this year. A concurrent workshop for faculty and postdocs on teaching science of information courses brought faculty from across the nation who also attended the lectures at the school. The school provides a venue where doctoral and postdoctoral students can meet to learn from distinguished professors in information theory, and form friendships and collaborations. This year the school introduced several interdisciplinary topics in the emerging field of science of information. Students the overall school mean rating of 4 out of 5.

The school format has courses by distinguished scientists followed by student poster presentations. The five featured speakers this year were Michelle Effros (Cal-Tech) who explained “Information Theory for Large Networks”, Scott Aaronson (M.I.T.) gave an enthusiastic tutorial on “Quantum Computing and Information”, P.R. Kumar (Texas A&M) and postdoctoral scholar Jonathan Ponniah co-presented “A Clean Slate Approach to Security of Wireless Networks”, Mehmet Koyuturk, gave a survey on “Complex Diseases and Information Theory”. Emina Soljanin (Bell Labs), gave the Padovani Lecture on “The Secret Lives of Codes: From Theory to Practice and Back”. Videos and slides of the lectures are viewable on the Science of Information Channel via http://soihub.org/summerschool, as well as on the IEEE IT Society 2013 School page http://www.itsoc.org/north-american-school-2013

Students gave fast-paced overviews of their research during the traditional “one-minute-madness” series of presentations that was enjoyed by all. Three poster sessions took place during the school. There was a broad scope of topics presented, and many students commented that the interdisciplinary nature of the posters presented provided much insight and helpful discussions. Many connections were made between the students and several mentioned possible future collaborations. Several students offered to lead open problem discussions in the evenings following dinner, and a professional development session on biases and diversity in hiring was also offered to the students one evening.

2013 sponsors included IEEE Information Theory Society, Center for Science of Information, Purdue Computer Science Department, and Vice President for Research, Princeton Electrical Engineering Department, UC Berkeley Departments of Electrical Engineering & Computer Sciences, Statistics, and ERSO, Bryn Mawr College Computer Science Department, and Texas A&M Electrical & Computer Engineering.

In addition to gaining experience presenting their research to their peers, students report a 3.7/4 on obtaining useful feedback to their research from talking with other students and faculty, and a 3.3/4
that they were able to start some level of professional connection with their peers for possible collaborations. Comments include:

“The poster sessions were a great opportunity to see the variety of information-theoretic research that is going on. They facilitated meaningful conversations. I hope to continue discussing my research with some of these students. The one-minute madness sessions were useful, too.”

“Sharing my research and getting to see what the other students’ research was certainly valuable. It is also good to speak to someone in the same stage of career and talk about practical aspects of the PhD life.”

“I think the summer school was a big success. I enjoyed the talks very much. The poster sessions were very interesting. I had several deep discussions with other researchers about my (and their) research which was very helpful for me.”

Curriculum Development and Teaching of SoI

Center faculty taught courses that incorporated the science of information and developed 10 new courses to educate students in the Science of Information. Additionally, center faculty report incorporating new science of information content into eight existing courses. This period, over 700 students received instruction from Center faculty in these Science of Information topics through these 18 courses. Specific SoI content for each of these courses is described below:

1. **EE226B, Faculty: Venkat Anantharam, UC Berkeley “Topics in Stochastic Processes”**
   Educational material with strong overlap with center thrusts was developed during this course, specifically notes on concentration inequalities a la Talagrand. About 8 -10 students and postdocs attended this course and benefited from it. Other material with less direct but still nontrivial relevance to the center which I covered in this course was: (1) a survey of heavy tailed and long range dependent stochastic process models and their applications to communications to control, and (2) a survey of mean field limit models and their applications. Each of the three units that I covered lasted for roughly two weeks (six lecture hours of course material).

2. **Measuring Uncertainty in Science of Information Faculty: Mark Daniel Ward, Purdue**
   First-year undergraduate Dean’s Seminar (essentially, an Honors course) on “Measuring Uncertainty” in fall 2013 at The George Washington University. I plan to teach it in the same style as the new Introduction to the Science of Information course that I offered in Purdue University’s University Honors Program in fall 2011.

3. **CS 380, Faculty: Deepak Kumar, Bryn Mawr “Recent Advances in Computer Science, Topic: Network Analysis”**
   This new course included content relating to Science of Information in the area of networks, their properties, and their analysis.

4. **ECE 55900, Faculty: Olgica Milenkovich, U of Illinois “Modern Topics in Coding Theory**
   Covered codes on graphs, polar codes, distributed storage, coding for modern storage systems, etc.

5. **Faculty: Sergio Verdu, Princeton**
   Tutorial on the results obtained under the sponsorship of this Grant at the IEEE Int.
Symposium on Information Theory and the attendance (over 150) broke all records for attendance at ISIT tutorials.

6. **BENG 100, Faculty: Todd Coleman, UC San Diego** “Introduction to Bioengineering Design”
   A general introduction to bioengineering design, including examples of engineering analysis and design applied to representative topics in biomechanics, bioinstrumentation, biomaterials, biotechnology, and related areas. A review of technological needs, design methodology, testing procedures, statistical analysis, governmental regulation, evaluation of costs and benefits, quality of life, and ethical issues.

7. **BENG 223, Faculty: Todd Coleman, UC San Diego** “Thermodynamics, Statistical Mechanics, Interfacial Phenomena in Living Systems”
   Thermodynamics, statistical mechanics, and interfacial phenomena that emphasize the chemical natures of living systems. Topics include intermolecular and surface forces, calculation of energetic processes, computation of electrical forces and fields, and principles of physics in multiscale engineering and design.

8. **Analytic Pattern Matching: DNA, Lempel-Ziv, and Trees, Faculty: Wojciech Szpankowski, Purdue** (while on sabbatical - Technology University of Gdansk, 2013.)

9. **Probability Theory Short Course – SoI Online Series Faculty: Mark Daniel Ward, Purdue**
   Probability theory tutorials were produced within the context of the Center. The online short course will aid students in preparing for more advanced topics in science of information, as well as assist students in getting maximum value from the introduction to science of information course.

10. **Mathematical Theory of Information Short Course - SoI Online Series, Faculty: Mark Daniel Ward, Purdue**
    Follows from the probability theory short course with more advanced tutorials for students to gain understanding of Shannon’s theorems. Provides foundational knowledge for students to prepare for more advanced topics in science of information.

**Existing Courses with New SoI Content Included**

11. **EE229A, Faculty: Venkat Anantharam, UC Berkeley** “Information Theory and Coding (Grad level course)”
    The treatment of some of the material was influenced by my center related activities. Specifically, the discussion of multiuser information theory included a discussion of issues of coordination over networks. We also discussed techniques for understanding the structure of error exponents of communication channels in the high SNR regime.

    Improved presentation of basic model of statistical pattern recognition and some of the techniques covered.

13. **CS 110, Faculty: Deepak Kumar, Bryn Mawr** “Introduction to Computing”
    This course was updated to reflect the importance of topic(s) related to Science of Information. Specifically, this course introduces the role of Big Data, techniques acquiring, storing, and subsequently analyzing the data followed by interactive visualization. This topic motivated a discussion of the move towards an information-centric approach to studying phenomena in various application areas like biology, economic, finance, etc.
14. CS 231, Faculty: Deepak Kumar, Bryn Mawr “Discrete Mathematics”
This course was updated to reflect the importance of topic(s) related to Science of Information. Specifically, this course incorporates fundamentals of probability theory and its applications in quantifying, storage, and transmission of information.

15. CS 325, Faculty: Deepak Kumar, Bryn Mawr “Computational Linguistics”
This course was updated to reflect the importance of topic(s) related to Science of Information. This course incorporated the statistical approaches to language processing that rely on the availability of large textual/language corpuses (like the Brown Corpus, Twitter feeds, etc.). The role of Big Data techniques applied to language processing including maximum entropy models for parsing and disambiguation.

Provides key foundations for our CSol research. A new unit this time will be on “failure detectors”, which is the main topic that Srikanth Sastry worked on for his CSol project. . .

17. Introduction to Science of Information, Faculty: Robert Rwebangria, Howard Version two of the course incorporates lessons learned from the previous year.

18. Introduction to Science of Information, Online Version, Faculty: Mark Daniel Ward, Purdue The online version has developed additional content to introduce the emerging field relevant to many undergraduate majors and practical problems. Additional content was developed to address life science topics. Multiple practice problems were developed to accompany each topic tutorial, along with weekly quizzes to gauge progress in understanding the content.

Courses to be Developed Next Period
As a result of faculty development through two workshops, eight faculty will be involved with teaching five new courses and developing 19 undergraduate and advanced topic modules in 2014.

19. Information of Cell Signaling (cross listed Biology/Computer Science)
Fall 2014, Undergraduate Upper Level, Villanova University
Faculty: Anil Bamezai, Lillian (Boots) Cassel, Vijay Gehlot, Karen Watanabe

20. Advanced Introduction to Science of Information Theory
Fall 2014, Graduate Level, Carnegie Mellon University, Faculty: Pulkit Grover

21. Introduction to Science of Information (with new topics on Error Correcting Codes, Statistics & Data Mining, and Networks & Complex Structures),

22. Fall 2014, Undergraduate Level, M.I.T., Faculty: Shan-Yuan Ho, Muriel Medard.

23. A General Education Bi-lingual Course on Science of Information
Fall 2014, Undergraduates in multiple disciplines, EAFIT University (Medellin – Colombia), Faculty: Juan Lalinde

24. Network Coding Short Course
Fall 2014, Graduate Students, Texas A&M University, Faculty: Alex Sprintson
Modules to be Developed Next Period

Physics of Information, spring 2014, Carnegie Mellon University, Faculty: Pulkit Grover

Understanding Information Flow in Circuits, spring 2014, Carnegie Mellon University, Faculty: Pulkit Grover

Goal-oriented communication from a multi-agent control viewpoint, spring 2014, Carnegie Mellon University, Faculty: Pulkit Grover

Network Coding Modules (Short Course), spring 2014, Graduate level, Texas A&M University
Faculty: Alex Sprintson

1) Introduction (two lectures);
2) Mathematical foundations, coding advantage, diversity coding (four lectures);
3) Polynomial and randomized algorithms for network code construction (four lectures);
4) Coding complexity (two lectures);
5) Network coding applications in network security and reliability (two lectures);
6) Wireless network coding (including the index coding problem) (four lectures);
7) Applications for network storage (two lectures);
8) Conclusions and future directions (two lectures).

Introduction to Information Theory for Biologists Modeling Proteins, spring 2014, Rose-Hulman Institute of Technology, Faculty: Yoshi Shibberu, Mark Brandt
Using Sequence Logos to Apply Information Theory in Biology, spring 2014, Rose-Hulman Institute of Technology, Faculty: Yoshi Shibberu, Mark Brandt

Evaluating Data Quality in Biological and Environmental Science, spring 2014, Oregon Health & Science University, Faculty: Karen Watanabe

Adapting Probability Theory and MTC Modules for Math Circles Program (Middle School Students), Spring 2014, Graduate level, Texas A&M University, Faculty: Alex Sprintson

Finding the Median, Fall 2013, M.I.T., Faculty: Shan-Yuan Ho

The Huffman Code, Fall 2013, M.I.T., Faculty: Shan-Yuan Ho

The BCH Code, Fall 2013, M.I.T., Faculty: Shan-Yuan Ho

Translation of Probability Theory & MTC Modules into Spanish, 2014, EAFIT University (Medellin – Colombia), Faculty: Juan Lalinde

**Goal 2: Integrate research and education through collaboration (External)**

The second goal of the education team is to integrate research and education through collaboration. The Center actively encourages funded graduate students to present their research at professional conferences. This year, the primary activity for reaching external audience of students was through our summer school organizing the North American Information Theory Society Annual Conference (reported above). 102 external students and faculty participated in the school this year. A series of poster sessions and networking sessions were held during the 2013 Center Summer School. Students gave 56 overview oral presentations followed by 56 posters during three separate sessions:

**Session 1**

<table>
<thead>
<tr>
<th>Author Name</th>
<th>Poster/Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abram Magner</td>
<td>Uniform Attachment Symmetry Results</td>
</tr>
<tr>
<td>Adam Williamson</td>
<td>An Improved Lower Bound on Rate for Variable-length Codes with Active Feedback</td>
</tr>
<tr>
<td>Alexandros Manolakos</td>
<td>Null Space Learning In MIMO Cooperative Cellular Networks Using Interference Level Feedback</td>
</tr>
<tr>
<td>Aman Chawla</td>
<td>On the reliability function of a neuronal channel when used in the presence of neuronal feedback</td>
</tr>
<tr>
<td>Amin Emad</td>
<td>CaSPIAN: A Causal Compressive Sensing Algorithm for Inference of Gene Interactions</td>
</tr>
<tr>
<td>Anderson Oliveira</td>
<td>Polynomial Representation of Codons by Galois Field Extensions and Labeling in the Study of the Genetic Code</td>
</tr>
<tr>
<td>Antonio Campello</td>
<td>Projections, Dissections and Bandwidth Expansion Mappings</td>
</tr>
</tbody>
</table>
Boulat Bash  
Low Probability of Detection Communication

Cátia Queiroz  
Maximal Quaternion Orders Derived from \{4g, 4g\} Tessellations

Cenk Sahin  
Channel Coding Over Finite Transport Blocks in Modern Wireless Systems

Cintya Benedito  
Hyperbolic Lattices Obtained from Arithmetic Fuchsian Groups via Hyperbolic Tessellations

Curt Schieler  
Rate-distortion Theory of Secrecy Systems

Daniel Bartolomeo  
Generating Dynamics by Varying an Entropic Framework

Diana Maamari  
Sum Capacity of K user Cognitive Interference Channel to within Constant Gap

Diego Mesa  
Convex Bayesian Inference Methods via Optimal Transport

Dmitry Krotov  
Morphogenesis at Criticality?

Ebrahim MolavianJazi  
On the Dispersion of Slow Rayleigh Fading Channels

Elyas Sabeti  
Atypical Information Theory for BIG DATA

Badih Ghazi  
LP Decoding of Spatially Coupled Codes

**Session 2**

Eva Song  
A Bit of Secrecy for Gaussian Source Compression

Fatemeh Arbabjolfaei  
Index Coding

Frank DeVilbiss  
Model Selection Problems in Systems Biology

Guoqiang Cai  
High Speed Train Safety Guarantee System

Hajar Mahdavi-Doost  
Sampling and Decoding Energy Tradeoff

He Zhu  
Compositional and Lightweight Dependent Type Inference for ML

Hendrik Vogt  
A Reflection Model for Secret-key Agreement

Idoia Ochoa  
A New Lossy Compression Algorithm for Quality Scores Based on Rate Distortion Theory

Jingbo Liu  
Robustness of Sparse Recovery via “F” minimization: A Topological Viewpoint

Jonathan Ligo  
A Controlled Sensing Approach to Graph Classification

Kai Tang  
A DNA Phosphatase Functions in Active DNA Demethylation in Arabidopsis

Kasra Vakilinia  
Non-Binary LDPC Code Design from Inter-Connected Cycles
Lele Wang  
Practical Coding Techniques for Network Communication Systems

Li Yan  
Adaptive Resource Allocation for Heterogeneous Traffic over OFDM Relay Networks

Li Yi  
Separation of the Collective and Intrinsic Particles Correlations in Relativistic Heavy Ion Collisions

Luzinete Faria  
MicroRNAs as Codewords of Cyclic Codes

Madars Virza  
SNARKs for C

Mario Enrique Duarte Gonzalez  
Linear Error Correcting Codes with Different input/output Abelian Alphabets for Modeling the Ribosome

Maryam Yammahi  
Approximate Searching of Very Large Data Files

Justin Richter  
Environmental data analysis

Session 3

Mikhail Tikhonov  
Genetic network complexity: weights matter more than topology

Minji Kim  
Gene Prioritization Via Rank Aggregation

Nikhil Gulati  
Bandit Learning for Heterogeneous Wireless Networks

Pablo Robles  
Incorporating Dependencies and Non-Stationarity into Kronecker Product Graph Models to Generate Realistic Networks

Qiong Wu  
Arsenomics: Transcriptomics of Arsenic Tolerance in P.vittata

Ramezan Paravi  
Markov Processes: Estimation in the Undersampled Regime

Ryan Muir  
Statistical Connection of Photon Counting and Photon Averaging

Sam Spencer  
Untitled

Sanket Satpathy  
Secure Cascade Channel Synthesis

Sinem Unal  
General Index Coding Problem with Three Decoders

Temitope Toriola  
Threats, Vulnerabilities, and Security Controls in Cloud Computing

Xiujie Huang  
Optimization of Achievable Information Rates and Optimal Detection for the Multilevel Flash Memory Channel

Yanzhu Ji  
Detecting transposable elements (TEs) in transcriptomes of three Heteromyid rodents

Yuguang Gao  
Achieving Capacity of Large-Alphabet Discrete Memoryless Channels

Zhiying Wang  
Codes for Network Switches
Yanying Chen  The Capacity Region of Three User Gaussian Inverse-compute-and-forward Channels
Eric Graves  Information Integrity

**Goal 3: Establish a community of practice around the Science of Information as an emerging field of study. (External)**

The third goal of the education team is to establish a community of practice around the Science of Information as an emerging field of study. The Center coordinates varied events to support community development such as faculty training workshops, meeting with members of industry, and public outreach events featuring the science of information. This year, the Center coordinated two faculty workshops on teaching SoI. The first workshop series targeted CS1 teachers, and the second workshop focused on disseminating the Introduction to the Science of Information course, and inviting faculty from diverse institutions to develop and adapt SoI content for undergraduate and graduate level students. The details of these workshops are in the tables below. We now describe each of these activities in detail.

**Professional Development**

Two faculty workshops on teaching a science of information course resulted in 52 attendees – the workshops were designed for attendees to learn more about the Center, its effort to develop a curriculum around SoI, and how they can get involved. Outcomes of these workshops resulted in faculty coordinating with the Center to develop five new courses and 19 new topic modules (*details in section 2c.*)

**Outreach for Broader Impact**

The Center was involved in a number of external seminars. Center faculty reported presenting 68 seminars with CSol content at other universities with students and others in the audience.

The Center sponsored a Computer Science and Math Theory series of lectures at Purdue for students and faculty (90 attendees).

**Other Outreach Activities**

The Center has planned several Science of Information Days. For the Bay Area, UC Berkeley and Stanford (hosted by Venkat Anantharam) planned a day for students and faculty to come together at Berkeley for posters, lunch, four guest speakers, and research discussions on November 4, 2013. A similar event at Purdue took place October 7, with an invited guest speaker, Grace Yang, Yale University, who spoke on Large Alphabet Compression and Predictive Distributions through Poissonization and Tilting.

The Center mailed packets to 300 university programs around the U.S. in science and engineering with our information bits posters, summer school information, and general invitation to view
content on our soihub.org website. This resulted in a large group attending our summer school, and an increase in views of our content such as seminar videos and tutorials.

The Center has had a presence at other events; Grace Hopper Celebration of Women in Computing, Minneapolis, MN October 2-5, 50 Years of Computer Science, Purdue University, April 5, 2013, and an introduction to the Center for new CS students at Purdue, August 15, 2013. Our information bits poster series has been popular at these events.

The Center maintains very active Facebook and Google + pages and communities. These serve as additional avenues for us to communicate to a broad audience our activities, opportunities, and announcements whenever our members are in the news.

2d. Research and Education Integration

*Describe and discuss the ways in which the Center integrated research and education in the reporting period, with examples as appropriate.*

The Center regards integrating research and education as a major education goal for the Center. Four main activities support this goal: 1) research posters and oral presentations, 2) research seminar series, 3) workshops and research meetings, 4) student research teams. These activities provide students with platforms to share their own research, take an active role in Center activities, and undertake cutting edge research in the Science of Information. The education team sees students as valuable and essential colleagues who will play a key role to help define the Science of Information. We will now discuss the programs integrating research and education in further detail.

**Research Posters and Oral Presentations**

Students most directly share their own research in poster sessions and oral presentations. In period 4, the Center sponsored two poster sessions. The first poster session occurred during the December 2012 All-Hands Meeting. 31 graduate students and postdocs presented a poster describing their research. The second poster session occurred during the Center’s Annual Summer School. A large group of students and postdocs attended the Summer School, where 56 students and postdocs presented a short oral presentation followed by their research posters. The Summer School provides a unique opportunity to integrate research and education because students new to the Center can see the work of their peers. The Center now has established mechanisms to collect student posters presented during the poster session. The posters presented at the Summer School are available on the soihub.org website, increasing student access to the posters. These were viewed 1,094 times online (nearly double that of period 3). Additionally, the Center encourages students to present their research at conferences. Students report presenting 131 posters and oral presentations at conferences. Research sponsored by the Center for the Science of Information has been presented at venues such as the ACM Symposium on Principles of Distributed Computing, the IEEE International Symposium on Information Theory, the Conference on Information Sciences and Systems, and the International Symposium on Chemical Reaction Engineering.
Seminar Series

The Center actively encourages students in advanced stages of their research to present their work in the Virtual Brown Bag seminars. These one-hour seminars permit high-level presentation of advanced research. To date, 23 PhD students and postdocs have presented their research projects during the Virtual Brown Bag seminars. Through seminars, students take on an active role of instructing their peers and receive questions and feedback.

Center graduate students have access to high-profile researchers through diverse Center seminar series. The Center has invited Luciano Floridi, Aurel Lazar, Leroy Hood, Emery Brown, Nancy Lynch, and Tomasz Imielinski to give Prestige lectures. The Prestige lectures help integrate research and education because students can see how researchers link core concepts within the Science of Information to pioneering research. Center faculty also provide students with further instruction in major research thrusts during the Center’s Annual Summer School. During the All-hands Student Research Meeting both David Tse and Tsachy Weissman gave keynote presentations about research in the Center. The Center welcomes students to attend a variety of guest lectures as they are offered throughout the year, and facilitates web-based viewing for all in the Center.

Monthly virtual brown bags with student presentation via the web.
**Workshops and Collaboration**

Workshops together with the student research teams, poster sessions, seminar series, and summer schools, we believe these have led to building a growing community in the Center where even students are now collaborating on their research together. This period 36% of students reported presenting or publishing research in collaboration with other Center students and postdocs. This is up from 15% reported in period 3.

Several research focused meetings were held with student involvement. A primary meeting was organized around an All-Hands meeting last December. This was a follow-up meeting to the summer student research workshop. Additional meetings included the Big Data Workshop in March, the Summer School, a Princeton-Bell Labs workshop, and two Science of Information Research Days; Bay Area at UC Berkeley (Nov. 4, 2013), and at Purdue (Oct. 7). The above workshops build upon the Grand Challenges workshop at Princeton University and a student research workshop in at Purdue in 2012 that focused on identifying grand challenges in the science of information. These workshops provided space for students and faculty to think critically about what constitutes the emerging field of the science of information.

All students have been invited to participate in research grand challenge discussions. Students will be involved in face-to-face discussions during the all-hands workshop in December, as well as form small discussion groups around specific topic areas.

**Student Research Teams**

This year, the Center has made considerable strides in supporting student research teams. Outcomes from the education programs 2-day student research workshop to foster interdisciplinary collaboration led to formation of four student research teams. This workshop drew upon the Center’s strengths of building community among students and launched the Center’s new project initiatives that help fund exploratory interdisciplinary collaborative research. 21 students and postdocs participated in this workshop, learned about the three thrust areas and reflected on how they could integrate their research on collaborative teams to address grand challenges. The Center provided support and facilitation to four collaborative teams to work on projects that span all three primary thrusts of the center. These teams gained experience in professional grant writing, creating interdisciplinary teams with members working at multiple institutions, and defining interdisciplinary research questions in the emerging research area of the science of information. Full team reports, overview presentations, and meeting details are available on our web:

http://www.soihub.org/research-teams.php

2e. **Education Indicators and Metrics**

*Describe how the Center is doing with respect to the indicators/metrics listed above. Include any data that have been collected on the indicators/metrics.*
A logic model based on the Center’s strategic plan for education has been used to evaluate activities against expected outcomes for the metrics below in Table 2. Following the metrics table is assessment findings narrative for two efforts – Introduction to Science of Information course findings regarding student learning and confidence, and 2) Faculty workshop on Teaching Science of Information.

**Education Metrics for Period 4**

<table>
<thead>
<tr>
<th>Goal/Programs</th>
<th>Expected Outcomes</th>
<th>Evaluation Metrics</th>
<th>Metrics Measured</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GOAL 1: Educate Students in the Science of Information</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annual Summer School</td>
<td>A majority of current Center students early in their course work attend</td>
<td>Number of Center students attending</td>
<td>38 (26 students, 12 faculty/staff)</td>
</tr>
<tr>
<td></td>
<td>Broader impact – outside students attend</td>
<td>Number of students outside the Center attending</td>
<td>102 (82 students, 20 faculty)</td>
</tr>
<tr>
<td></td>
<td>Students learn about the research agenda of the Center and important problems in the emerging field</td>
<td>Level of knowledge gained as self-reported by students</td>
<td>Rating of 3.25/4</td>
</tr>
<tr>
<td></td>
<td>Students learn from each other and share their own interests, ideas, and research</td>
<td>Level of networking success and peer exchange as reported by participants</td>
<td>Rating of 3.76/4</td>
</tr>
<tr>
<td></td>
<td>Students gain knowledge of specific skills, methods and tools for addressing SoI problems</td>
<td>Quality of presentations and understanding of skills, methods and tools taught as evaluated by students</td>
<td>Rating 3.97/5.</td>
</tr>
<tr>
<td></td>
<td>At least 15% of attendees represent minority and female students</td>
<td>Diversity of students</td>
<td>31% Female attendees</td>
</tr>
<tr>
<td>Courses &amp; Learning Modules</td>
<td>New courses, and new knowledge transfer into existing courses, support learning of SoI are developed to reach undergraduate and graduate students</td>
<td>1. Number of courses offered</td>
<td>18 (10 new, 8 existing)</td>
</tr>
<tr>
<td></td>
<td>2. Students enrolled</td>
<td>2. Students enrolled</td>
<td>717 (426 new, 291 existing)</td>
</tr>
<tr>
<td><strong>Goal/Programs</strong></td>
<td><strong>Expected Outcomes</strong></td>
<td><strong>Evaluation Metrics</strong></td>
<td><strong>Metrics Measured</strong></td>
</tr>
<tr>
<td>------------------</td>
<td>----------------------</td>
<td>------------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td></td>
<td>Learning modules on foundational and advanced topics are produced and offered online</td>
<td>Number of topic modules and video tutorials produced</td>
<td>18 new modules, 80 video tutorials</td>
</tr>
<tr>
<td></td>
<td>Students completing courses and modules demonstrate understanding of subject matter, and increased appreciation of the nature of information</td>
<td>Reflection essays, course projects, and survey results</td>
<td>See third party assessment of outcomes below this table.</td>
</tr>
<tr>
<td></td>
<td>Modules are developed and incorporated into a variety of courses by CSoI and other faculty.</td>
<td>Number of faculty developing or incorporating modules into new and existing courses</td>
<td>Planned: 8 faculty, 19 modules</td>
</tr>
</tbody>
</table>

**GOAL 2: Integrate Research and Education through Collaboration**

<table>
<thead>
<tr>
<th><strong>Student Research Teams</strong></th>
<th><strong>Expected Outcomes</strong></th>
<th><strong>Evaluation Metrics</strong></th>
<th><strong>Metrics Measured</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Students gain experience and skills in grant writing</td>
<td>Number of students submitting proposals</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Students gain experience working in interdisciplinary teams</td>
<td>Number of teams formed and funded</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>CSoI students start collaborating across institutions, sharing methods, ideas, and problems</td>
<td>Degree of interdisciplinary interaction (based on # departments or field areas involved)</td>
<td>7 universities, 7 depts.</td>
<td></td>
</tr>
<tr>
<td>New or improved methods and approaches are developed to address SOI problems</td>
<td>New or improved methods or solutions reported by teams</td>
<td>See <a href="http://www.soihub.org/research-teams.php">http://www.soihub.org/research-teams.php</a> for full reports</td>
<td></td>
</tr>
<tr>
<td>Student teams present results at poster sessions and conferences</td>
<td>Presentations of results (via posters or conference talks)</td>
<td>18 posters and presentations</td>
<td></td>
</tr>
<tr>
<td>Multi-author/multi-institution scholarly papers are published by team members in the wider community</td>
<td>Published papers</td>
<td>6 publications (each team has additional papers in progress)</td>
<td></td>
</tr>
<tr>
<td>Goal/Programs</td>
<td>Expected Outcomes</td>
<td>Evaluation Metrics</td>
<td>Metrics Measured</td>
</tr>
<tr>
<td>---------------</td>
<td>------------------</td>
<td>--------------------</td>
<td>------------------</td>
</tr>
<tr>
<td>Students become experienced working in teams across institutions and departments on broad research issues</td>
<td>capacity of students to work on interdisciplinary problems increases as measured by feedback and success of teams</td>
<td>Capacity has increased, as demonstrated by both presentations on the research and published papers, See full reports at <a href="http://www.soihub.org/research-teams.php">http://www.soihub.org/research-teams.php</a></td>
<td></td>
</tr>
<tr>
<td>Lessons learned regarding best practices on team building and facilitating student research groups across institutions</td>
<td>Lessons learned are incorporated and shared</td>
<td>Key lessons: 1) Teams require regular meetings and focused time for success, 2) Students require coordination and involvement with faculty to make it a priority</td>
<td></td>
</tr>
</tbody>
</table>

**Seminar Series - faculty and student**

| Students gain presentation skills through monthly virtual brown bag research presentations and discussions | Number student seminars Number attending | 9 monthly student seminars, 161 students, faculty attending |
| Students and faculty learn about leading edge of problems and solutions in SOI | Number of seminars and presentations given to students and faculty | 85 (17 internal, 68 external) |
| 2. Number attendees | Accurate counts unavailable for external. Known counts are given within in the report for each |

**Student Presentations**

<p>| Work in the Center is shared among broader audiences | Number of posters and presentations | 131 (62 posters, 69 oral presentations) |
| Students gain experience explaining their work to others and receive valuable critical feedback from peers and faculty | Student feedback of internal poster sessions indicate their value | 87 poster presentations, 56 oral. Students rated the value of these |</p>
<table>
<thead>
<tr>
<th>Goal/Programs</th>
<th>Expected Outcomes</th>
<th>Evaluation Metrics</th>
<th>Metrics Measured</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Quality and impact of posters and presentations increases</td>
<td>Number of awards/recognition for student research</td>
<td>27 student awards</td>
</tr>
<tr>
<td></td>
<td>Students publish their research in journals and conference proceedings</td>
<td>1. Number of Center students and postdocs publishing</td>
<td>60 students published</td>
</tr>
<tr>
<td></td>
<td>2. Number of publications</td>
<td></td>
<td>149 publications</td>
</tr>
<tr>
<td>Workshops &amp; Collaboration</td>
<td>As community builds, an increase in collaborations occurs</td>
<td>Degree of CSol collaboration</td>
<td>36% of students report joint publications and presentations with other CSol students. Compared to 15% in period 3</td>
</tr>
<tr>
<td></td>
<td>Collaboration is enhanced through student events and workshops</td>
<td>number of students participating in workshops</td>
<td>99 (19, Purdue research discussion, 31-Student research meeting 12/2012, 26 – center students at summer school met, 20 - Big Data workshop, 3 - Princeton-Bell Labs workshop).</td>
</tr>
</tbody>
</table>

**GOAL 3: Establish a community of practice around the Science of Information as an emerging field of study.**

<p>| Professional Development | Faculty trainings provide platform for development of new and adoption and use of existing SoI materials | 1. Number of faculty attending trainings | 52 (25 - Mellon 23/AALAC <em>What is Information?</em> workshop, 27 – Purdue Teaching) |</p>
<table>
<thead>
<tr>
<th>Goal/Programs</th>
<th>Expected Outcomes</th>
<th>Evaluation Metrics</th>
<th>Metrics Measured</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Students gain skills in conducting research in industry, presenting research to larger community, gain instruction experience in the classroom, and serve on a planning council with the Center</td>
<td>Number of new courses supporting SoI planned for development</td>
<td>SoI Workshop)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Number of new modules supporting SoI planned for development</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>Outreach for Broader Impact</td>
<td>Faculty mentoring other students in the Center</td>
<td>Teaching Assistant/courses</td>
<td>21 TA’s for 30 courses</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mentoring Others</td>
<td>25% of grad and postdocs report mentoring others. This is identical rate compared to period 3.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Internships</td>
<td>16 internships reported</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Student Postdoc Council</td>
<td>12 student reps</td>
</tr>
<tr>
<td></td>
<td>Faculty presentations to external student audiences</td>
<td>Evidence of broader mentoring by faculty</td>
<td>Students report mentoring occurring among a broad group of faculty in CSoI. See section 2A.3 for testimonials from students</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Number of presentations given</td>
<td>68 external seminars with CSoI content</td>
</tr>
<tr>
<td>Goal/Programs</td>
<td>Expected Outcomes</td>
<td>Evaluation Metrics</td>
<td>Metrics Measured</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------------</td>
<td>--------------------</td>
<td>------------------</td>
</tr>
<tr>
<td></td>
<td>Center web site (soihub.org) is used worldwide to access content</td>
<td>1. Total views and downloads via soihub.org</td>
<td>Period 4: 53,986 (Nov. 1, 2012- Sept. 17, 2013)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. views for seminar and brown bag videos</td>
<td>Period 4: 8,480. Period 3: 3,185 views</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3. views for summer school content and videos</td>
<td>Period 4: 5,568 (2,797 page views, and 2,771 video tutorials views) Plus additional 4,065 views via the IT Society school page</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4. views of SOI course content</td>
<td>learning system platform currently being developed and content integrated.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5. views of research posters</td>
<td>Period 4: 1,094</td>
</tr>
<tr>
<td></td>
<td>Materials for increasing awareness of the Center and the emerging field produced</td>
<td>Number of brochures, posters, and center materials distributed</td>
<td>Period 4: 1000 Center brochures, 1000 Information Bits series posters, table top displays distributed to partners, used at events</td>
</tr>
<tr>
<td></td>
<td>Materials reach prime candidates for student programs</td>
<td>Number of universities displaying materials targeting students</td>
<td>All partner universities (10) &amp; 300 programs across U.S. universities in 2013, and IEEE IT Society</td>
</tr>
<tr>
<td>Current Student &amp; Alumni</td>
<td>Students find the Center valuable to their development and future career</td>
<td>Percent of students reporting positive values of participation in the Center</td>
<td>97% of students report Center has provided a valuable community for their development. See comments in section 2E.5</td>
</tr>
</tbody>
</table>
### Goal/Programs

<table>
<thead>
<tr>
<th>Goal/Programs</th>
<th>Expected Outcomes</th>
<th>Evaluation Metrics</th>
<th>Metrics Measured</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of students and postdocs participating in the Center community grows</td>
<td>Number of students and postdocs participating</td>
<td>Period 4: 138 (17 Post docs, 41 Undergrads, 80 Graduate). See Table 1, section 2A for period comparisons.</td>
</tr>
<tr>
<td></td>
<td>Next generation of scientists working in science of information</td>
<td>Alumni of CSoI continuing in academia, industry, nonprofits</td>
<td>Total 84 (37 in Period 4, 32 in Period 3, 15 in Period 2) 17 PhD theses (publicly available) are linked in our Soihub resources page. One postdoc secured faculty position and will teach a new advanced course and build 3 advanced teaching SoI modules in cooperation with CSoI. Four Center postdocs participated in the Teaching SoI Workshop.</td>
</tr>
<tr>
<td></td>
<td>CSSoI graduates continue to collaborate and establish science of information field</td>
<td>This is difficult to measure at this early juncture. The Center will work toward how to best capture this knowledge through an alumni network.</td>
<td></td>
</tr>
</tbody>
</table>
### Education Goals, Expected Outcomes, Metrics, and Measured Outcomes

<table>
<thead>
<tr>
<th>Goal/Programs</th>
<th>Expected Outcomes</th>
<th>Evaluation Metrics</th>
<th>Metrics Measured</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Industry and Alumni Network</td>
<td>As the alumni become established in industry and academia, the Center will maintain contact through a listserv and future speaking invitations with current students. Alumni “LinkedIn” pages are linked with their alumni profiles.</td>
</tr>
</tbody>
</table>

**Table 2.** Education goals, expected outcomes, metrics, and measured outcomes for period 4.

#### Summary of Findings – Introduction to Science of Information

**Course Evaluation and Student Perceptions of Learning**

**NSF TUES Faculty Training Workshop Report**

This assessment, conducted by Dr. Dave Nelson (external evaluator), describes the combined analysis student responses to summative course evaluation questions related to learning and student confidence. 31 students completed pilot courses at Purdue University, Howard University, and Bryn Mawr College. A combined total of 24 students completed a voluntary assessment survey. Although the overall sample size for each institution remains small, the combined data set does allow for some initial conclusions.

First, student responses to reflective essay questions affirm increased understandings of the multidisciplinary nature of information, while highlighting differences between implementations of the course at different institutions. The single greatest similarity among all open-ended responses was an appreciation of the interdisciplinary nature of information. Representative comments include, “SoI is a very fundamental way of understanding the world,” “a way to understand what information means to our daily lives.” More than 50% of qualitative responses indicated a greater appreciation of the applicability of SoI to daily life.

Second, students who completed the course and survey data demonstrated confidence and understanding of four key skills and two key understandings. Six attitude statements rated highly (Means of 1.99 or better) among the students. These statements reflect student confidence in problem solving, persisting through initially difficult problems and unexpected challenges, and
scientific reasoning ability. Responses also demonstrate consistent understanding of the multidisciplinary nature and applicability of the Science of Information.

Similarly, students responded neutrally to the negatively coded item “Other students in the class seem to know more about the subject matter than I do,” received largely neutral responses, with a mean of 2.917 and a mode of 3. This may suggest that the classroom environments were sufficiently supportive to instill equality among student knowledge and skills while remaining sufficiently challenging to students’ existing perceptions.

Third, students indicated that they would likely pursue further study of the Science of Information. While no student indicated that s/he strongly agreed in such future pursuit, the mean response to the question was 2.750 and with a mode of 2 (Agree). Given that each of these courses were interdisciplinary in nature, introductory in approach and content, and theoretically challenging for students, the positive interest among students in further study illustrates the possible sustainability and scope of the course. Among students who indicated high likelihood of additional study in the subject, self-reported confidence in theoretical models correlated most strongly (.837 CC, P<0.01). Future iterations of the course may benefit from greater emphasis on theoretical modeling construction, and future assessment and evaluation might examine theoretical modeling aptitude in greater depth.

The declared academic major of the student does not appear to influence the student confidence in specific learning outcomes. Even when controlling for majors in a binary, with grouping computer science majors (9) vs others (15), a student’s declared major had no statistical correlative significance with declared competence in skills and understandings central to the course. However, the declared major of students at one particular institution correlates highly with perceived competence relative to peers in the class. At Bryn Mawr College, the four students in majors other than computer science indicated that they knew less than their peers in computer science majors (-.714 coefficient, p<0.01). Students in non-computer science majors indicated they were as likely to pursue additional studies in the Science of Information, were equally confident in their acquisition of key skills and understandings, but still believed they knew less than computer science students in the course. Perhaps the addition of online course modules and supplemental materials may provide additional content knowledge and confidence in foundational computer science understandings to students in multiple disciplines.

<table>
<thead>
<tr>
<th>Question</th>
<th>N</th>
<th>Mean</th>
<th>Standard Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>I am confident in my ability to persist until a solution is found</td>
<td>24</td>
<td>1.458</td>
<td>.5882</td>
</tr>
<tr>
<td>I am confident in my problem solving ability</td>
<td>24</td>
<td>1.625</td>
<td>.7697</td>
</tr>
<tr>
<td>I understand the multidisciplinary nature of the Science of Information</td>
<td>24</td>
<td>1.667</td>
<td>.5647</td>
</tr>
</tbody>
</table>
Participants at the June 5-6, 2013, workshop on teaching the Science of Information were solicited to complete two surveys; one administered two weeks before the event and one approximately a month following the workshop. Fifteen attendees completed both surveys, and their responses are analyzed here.

Among the participants who attended, evaluation reports judge the two-day workshop on teaching a Science of Information course to have been a success. From overall quality of the sessions (90% of the participants indicated the workshop was “excellent” or “very good”) to the individual likelihood of teaching a course in the future (85% of participants indicated they would teach or manage such a course), responses to the post-workshop survey demonstrate impact beyond the collegiality and discussion of teaching information science that occurred.

An analysis of the detailed responses indicates a significant increase in attendees’ willingness to incorporate information retrieval in future courses on information. In a survey conducted before the workshop, only 3 of 15 respondents indicated that they had incorporated Informational Retrieval or “Big Data” Analysis as topics in their courses. Following the workshop, 8 respondents indicated they planned to do so in the future. A smaller increase is present when attendees were asked about their desire to incorporate Shannon Theory into future course, with 9 of 15 respondents indicating they would do so, when only 7 had previously included the theory as part of the course.

Responses to questions concerning interest in particular subjects demonstrate that participants were now realizing the broad interdisciplinary aspects of Information Theory. Thirteen of 15 respondents indicated that attendance at the workshop had increased their interest in teaching computational and mathematical sciences a “great” or “good deal.” Given that the attendees come largely from computational and mathematical science disciplines, this result is hardly surprising. In fact, when asked before the workshop about their ability to incorporate such topics into courses, only 2 of the 15 respondents indicated that they were anything other than “Extremely” or “Very Prepared” to do so.

However, when asked before the workshop of their confidence levels in incorporating Biology and Physics into Science of Information courses, more participants indicated uncertainty about the
ability to integrate these units. Six of 15, and 5 of 15, respondents indicated that they were “Extremely” or “Very Prepared” to integrate Biology and Physics, respectively, into their courses. Following the workshop, 9 of 15 respondents indicated that they had gained a “Great” or “Good Deal” of interest in incorporating Biology into their Science of Information courses. Conversely, only 3 of 15 respondents expressed the same increased interest in Physics modules. It is possible that the presence of four faculty from biological science disciplines (and the corresponding absence of any physicists) at the workshop improved the confidence levels of the faculty in integrating Biology into their courses. While small, one can nonetheless conclude that the collegial nature of the workshops and sharing of pedagogical and content knowledge has broader the disciplinary scope of many faculty who will teach such courses in the future.

The Center for the Science of Information would be well-served with secondary inquiries to these fifteen attendees in future semesters. Asking these faculty members the degree to which they had incorporated elements from other disciplines and/or specific modules from the original Science of Information course would provide additional evidence of the broader impact of the Center’s educational work and the June 2013 workshop.

Longer term outcomes from the workshop are that the Center’s education program has worked with eight of the faculty participants to develop plans for 5 new courses and 19 topic modules to be delivered in 2014.

**Online Platforms for Engaging Students**

In addition to supporting innovative research projects, the Center professionally records and produces high quality online content and makes it available through our soihub.org website ([http://soihub.org](http://soihub.org)). There were over 60,000 page views of our site this period. Next to the start page, the most frequently visited content on our soihub site, the summer school content online received 5,568 views this past period (2,797 page views, and 2,771 watched the video tutorials). Additionally, the IT School page on itsoc.org/north-american-school-2013 received 4,065 page views. The Center virtually published student posters presented during the annual Summer School, increasing visibility of research performed by students and postdocs. These posters were viewed online 1,094 times. The virtual brown bag student research series continues to increase knowledge of cutting edge research being conducted by our students and postdocs in the Center. Our students have used this web-based venue to present a high-level talk about their research to their peers, staff, and faculty. These talks are available online. Overall, CSoI organized brown bags, seminars, and summer school videos were viewed 12,746 times (through Sept. 17).

We are building a solid foundation for an online presence in the broader community through our soihub.org platform where more than quadruple the views for video content occurred this period compared with last period. When we launch the SoI Learning HUB in 2014 ([http://learninghub.soihub.org](http://learninghub.soihub.org)), we will reach an ever larger audience of students. In addition to this, a Spanish-English version of our Introduction to Science of Information online materials is planned for production in 2014.
**Student Testimonials of Center Activities**

To evaluate the Center’s efficacy of building a community of practice, we asked students whether their participation in the Center has been valuable, and to elucidate on their experiences. 97% of respondents reported that their experience had been valuable. Many students added further explanation of what they found valuable about the Center. A subset of testimonials are included below:

**Do you feel your participation in CSol has been valuable? - Please explain.**

“I participated in the CSol Summer School 2012 where I was given the opportunity to travel with, and get to know better, the post-doc who would ultimately be advising me, in addition to the plethora of interesting talks and workshops where I was introduced to what the science of information means and how it can be applied. Having a background in biology, this was especially interesting since it was new territory to me. Participation in the Center has offered many new opportunities that augment my biology degree.”

“Being part of the center for the Science of Information encouraged broad discussions with a broad spectrum of scientists, and this exchange of ideas helps find parallels in what appears at first to be a highly diverse spectrum of problems. Such exchange, I believe, is a powerful driving force, generating new approaches and ideas.

CSol is interdisciplinary and my participation made me realize the challenges and the opportunities when applying generic computational methods in diverse application domains.”

“CSol provides an interdisciplinary perspective on information which is difficult to find in most departments.”

“I can offer two concrete pieces of evidence which demonstrate how my participation in the Center has been valuable: 1) The previous year was very productive for me. I gave many talks and published several papers on a broad set of topics. This increased productivity and breadth is a direct result of my involvement with the center. Through my interactions with Center faculty and students (some collaborators, some not), I have gained a stronger sense of which problems are of particular/fundamental importance in information science, broadly defined. This is reflected in the broader set of projects that I am pursuing now compared to those I pursued prior to joining the center. 2) I joined CSol as a postdoc immediately after completing my PhD. My participation in the Center activities has led to interactions with an incredibly diverse set of faculty and students. This has played a large role in shaping my vision for research and education in information sciences. I feel that my success in obtaining a faculty position was greatly facilitated by my involvement with CSol.”

“I have been able to benefit from several gatherings and conferences hosted by the Center, especially the IT summer school which allowed me to discuss my research with other peers; in addition, I was able to learn from many researchers working on similar topics and discuss possible future collaborations.”

“I have learned much about information theory from the summer schools, and I have gained experience in presenting.”
“I very much enjoy, and have benefitted from, the lectures (especially the Prestige lectures). The center offers many great opportunities and resources, such as the CSIo summer school.”

“It has provided a good platform for looking at my research from a broader perspective. It has also provided easy access to faculty and students working in areas related to mine, and their expertise has been very helpful. And finally, I enjoy the opportunity to mentor undergraduates, as I’ve been able to at Howard.”

“My participation in CSol is a professionally rewarding experience. In addition to socializing with students from other institutions, it provides a great opportunity to hear presentations and to converse with other researchers. The opportunity of sharing research ideas early through poster sessions/ all-hands-meetings/ etc. and getting feedback from researchers is another advantage of participating in CSol events.”

“The CSol seminars (including the weekly brown bags) are instrumental in providing novel perspectives and inputs. I believe that my experience with the CSol will help me gain an edge in my future career.”

“Having a variety of different research perspectives has really helped to push my research forward.”

2f. Education Plans for Next Reporting Period

Describe your plans for internal and external educational activities for the next reporting period with attention to any major changes in direction or level of activity. Also, list plans for developing new educational partnerships, if any, for the next reporting period.

Annual Summer School

The Center will be conducting our fourth annual summer school at one of our partner locations, tentatively at UC San Diego during late June 2014. With a growing interest from students in bridging information theory, computer science, and life science topics, this will be an opportunity to develop the school theme around this topic for more in-depth learning and exchange.

Course & Module Development & Implementation

The education team will continue working to expand participation of undergraduate and graduate students by working with faculty to develop new content, courses, and modules. A coordinated effort with faculty for 2014 has five new courses planned, with 19 topic modules being developed. The Introduction to the Science of Information course has now been pilot tested at Purdue, Bryn Mawr College and Howard University. It is being taught this Fall 2013 at George Washington University, with plans to teach versions of this course at M.I.T. and EAFIT University in Colombia in 2014.

An online learning platform, SoI Learning HUB, will organize all SoI content available in the form of full and short courses, and individual learning modules for both faculty and students to access. A
beta test of the system will take place winter 2013 and a first course is planned for spring semester 2014 http://learninghub.soihub.org

**Seminar Series, Workshops, Collaborations**

We plan to continue the monthly virtual “brown bag” student research presentations where students and faculty throughout the Center and beyond can join real-time via the web. The Prestige Lecture series will continue in 2014 with three internationally known speakers.

The Center is working to form organized research discussion events at its partner institutions in the coming year. This includes a Midwest area at Purdue for period 4, and period 5 includes a Bay area event at UC Berkeley, Nov. 4, 2013, an East Coast event at M.I.T., with another at Texas A&M in spring of 2014.

A faculty workshop on teaching science of information is being planned for early 2014 to be hosted at in conjunction with the 2014 summer school at UCSD. This will be patterned similarly to the workshop that we offered at Purdue this past period, which resulted in a core of faculty interested in developing SoI curriculum.

**Student Research Teams**

We plan to continue supporting collaborative student teams. Two of the four current teams with active members will continue, while adding new members. We intend to develop another team around the economic questions of the Center research. All students will have additional opportunities to develop teams as we move forward. We will incorporate lessons learned from the first year of team formation to help ensure success.

**Alumni**

As students become established in academia and industry we plan to build an alumni network that benefits both alumni and current and future CsoI students. In addition to the network we plan to bring in at least one successful alumni speaker (or panel) at a future all-hands student event. As our postdocs and students achieve faculty positions we attempt to remain connected, and assist those interested in developing SoI courses and content.

**Educational Partnerships**

We will continue to build new partnerships with faculty for developing new courses and modules to reach an expanding audience of students.
IV. Knowledge Transfer

Goals, Objectives, Metrics, and Problems Encountered

1a. Knowledge Transfer Goals and Objectives

Describe the Center’s overall knowledge transfer goals and/or objectives. If the Center’s overall knowledge transfer goals/objectives changed since the last reporting period, how did they change and why? [In section 2 below, please describe progress the Center has made toward reaching these objectives.]

Knowledge Transfer Mission Statement

Develop effective mechanisms for interactions between the Center and external stakeholders to support the exchange of knowledge, data, and the application of new technologies.

To achieve this mission, the Center is continually working to engage with associated stakeholders. Knowledge outflow takes the form of disclosures to the public domain (including publications, presentations, software tools, and instructional material), licensing and patenting, and technical collaborations. Knowledge inflow is achieved through industrial roundtables, funded research projects, researchers-in-residence programs, and contributions to the Science of Information Hub. Through these mechanisms, the Center identifies stakeholders and then develops programs to engage stakeholders.

The Center’s overall goal is to develop a comprehensive set of activities associated with its knowledge transfer mission statement. The Center identifies its goals with respect to each group of external stakeholders. Specifically, the Center quantifies growth in terms of number of external partners at the three- and five-year marks. Similarly, the Center establishes qualitative and quantitative targets for the online portal, educational and research seminar development and dissemination, and other outreach efforts.

The Center also continually monitors and updates its models of engagement in response to comments from the NSF Review Board, as well as the External Advisory Board. Of particular note are comments relating to industry outreach (engaging with a small group of companies in a targeted fashion, as opposed to engaging superficially with a larger group – as suggested by the NSF Review Board of 2012), international collaborations (as suggested by the NSF Review Board of 2012 and the External Advisory Board), and cross-center initiatives (as suggested by the NSF Review Board of 2012 and External Advisory Board). We present
here results from a number of new initiatives, along with immediate outcomes of these efforts.

1b. Performance and Management Indicators and Metrics

Inform us of the performance and management indicators the Center has developed to assess progress in meeting its knowledge transfer goals/objectives.

The Center has identified mechanisms for evaluation of the knowledge transfer program. These mechanisms involve the Director of Knowledge Transfer, and periodically, the Center leadership.

Objectives and Metrics for this year:

- Increase the number of industry funded projects by 3 for 2014 and by 7 for 2015.
- Organize and convene the industrial advisory board meeting.
- Develop and make available a complete online course in science of information along with supporting educational modules.
- Develop, continually improve, and increase use of the Science of Information website each year.
- Institute and build upon the Center’s research and graduate student seminar series.
- Develop international collaborations and partnerships.
- Establish student internship opportunities/sponsored fellowships.
- Create a logic model for the Center’s website that will clearly define its primary purpose and functions in support of the Center and its knowledge transfer initiatives.

We have met or exceeded most of our goals for the current year, as we shall show in this report. We have also established aggressive goals for the next year, building on both the research contributions and established connections with stakeholders.

1c. Problems Encountered in Reaching Goals

Discuss any problems you have encountered in making progress toward the Center’s knowledge transfer goals/objectives during the reporting period as well as any problems anticipated in the next period as appropriate. Include your plans for addressing these problems.

During the first two years of existence, the Center has attempted several different approaches to engage industry and develop meaningful interactions. However, these attempts have been met with limited success. As the current economic climate in industry wanes, such engagement presents a serious challenge. Furthermore, balancing the
industrial engagement of individual researchers and project partners, with the overall engagement efforts of the Center have been challenging. While individual researchers and project partners have been extremely successful in engaging with industry (in terms of funded projects from industry) on topics of core interest to the Center, suitable appropriation of credit to the Center has been harder. After considerable thought and planning, the Centers role in knowledge transfer is now more formally and productively conceived. This strongly incorporates the suggestions of the NSF Review Board of 2012. The following steps have been taken:

- The Center recognizes its key role in publicizing technical contributions of the Center and ongoing research projects and interests within the Center to the broader industry.
- The Center engages with a small set of targeted partners in deep and meaningful ways to solve focused problems of broad impact.
- The Center convenes an Industrial Advisory Board, which meets annually and reviews Center activities and relevance to industry. In addition, the Center convenes a number of industry roundtables at various partner institutions. The first three of these events are at Chicago (hosted by Purdue, September 2013), Berkeley (hosted by UC-Berkeley, November 2013), and Cambridge (hosted by MIT, March 2014). The goal of these roundtables is to broadly advertise Center activities to industry.
- The Center is now actively engaged with multiple formal international collaborations. The goal of these collaborations ranges from research interactions to educational partnerships and exchange programs.

We report on all of these activities and their impact in the rest of this Section.

Knowledge Transfer Recipients

2a. Activities and Organizations

List organizations with which knowledge transfer occurs and the frequency and type of interactions. Describe the Center’s knowledge transfer activities in the current reporting period and discuss how they enable the Center to meet its knowledge transfer goals/objectives listed in 1a above.

Narrative: For each activity, briefly describe the activity, its goals/objectives, outputs and the outcomes or impacts in the current reporting period.

The Center’s knowledge transfer activities during the current period were wide-ranging and robust. These activities are all geared toward meeting Center knowledge transfer goals that
include industry relations, international exchange, and development of the Science of Information course. These activities not only allow the Center to share its research with other entities, but also allow it to receive critical input and guidance relative to its overall agenda.

The following meetings were organized by the Center over the past year. These meetings include participants from educational institutions, as well as industry. Almost all talks at these meetings are made available over the Center web site.

<table>
<thead>
<tr>
<th>Knowledge Transfer Activity – Big Data Workshop – March 18 – 20, 2013</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Led by</strong> – W. Szpankowski</td>
</tr>
<tr>
<td><strong>Participants</strong> – LinkedIn, AFOSR, Georgia Tech, Center Institutions</td>
</tr>
<tr>
<td><strong>Description</strong> – The Big Data Workshop was held with the objective of identifying Grand Challenges the Center faces in its Knowledge Management thrust. Outputs from the workshop have been incorporated into the Center’s research grand challenges document.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Knowledge Transfer Activity – Industry Advisory Meeting – September 2013</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Led by</strong> – A. Grama</td>
</tr>
<tr>
<td><strong>Participants</strong> – Technicolor, Bell Labs, John Deere, Motorola, HP, Intel</td>
</tr>
<tr>
<td><strong>Description</strong> – As a direct knowledge transfer goal of the Center – this meeting was held in order to get direct input and feedback concerning industry engagement in the Center. This will now become an annual event.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Knowledge Transfer Activity – Bell Labs Workshop, September 5, 2013, Princeton</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Led by</strong> – W. Szpankowski</td>
</tr>
<tr>
<td><strong>Participants</strong> – Bell Labs Researchers and CSoI Faculty</td>
</tr>
<tr>
<td><strong>Description</strong> – This activity directly relates to the Center’s objective of engaging industry, this</td>
</tr>
</tbody>
</table>
workshop was held with Bell Labs researchers from a variety of areas. CSiO faculty discussed the Center’s major research areas and attended presentations from Bell Labs and Center researchers on various research projects. Potential collaborations were discussed as well as mechanisms for formalizing opportunities for Center students.

<table>
<thead>
<tr>
<th>Knowledge Transfer Activity – CSiO Prestige Lecture Series – 4 lectures throughout year.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Led by</strong> – W. Szpankowski</td>
</tr>
<tr>
<td><strong>Participants</strong> – Emery Brown, Harvard; Nancy Lynch, MIT; Tomasz Imieliński, Rutgers</td>
</tr>
<tr>
<td><strong>Description</strong> – The Center hosted an annual lecture series that brings leading researchers to the Center to highlight and share their findings. These lectures are typically recorded and made available online for later viewing. The Center has recently started to schedule these lectures at partner institutions. For example, a prestige lecture is being held at UC Berkeley and another is scheduled for spring at MIT. In addition, visiting faculty throughout the Center often present their work, further helping the Center meet its goal of instituting and providing research seminars and exchanges.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Knowledge Transfer Activity – CSiO Student Brown Bag Research Seminars - Monthly</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Led by</strong> – B Ladd</td>
</tr>
<tr>
<td><strong>Participants</strong> – CSiO Grad students and Post Docs</td>
</tr>
<tr>
<td><strong>Description</strong> – The CSiO Brown Bag Research Presentations, as discussed in Section III Education, provides students an opportunity to discuss and present science of information topics.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Knowledge Transfer Activity – CSiO Annual Summer School – Annually</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Led by</strong> – W. Szpankowski, B. Ladd</td>
</tr>
<tr>
<td><strong>Participants</strong> – Students and faculty worldwide attended</td>
</tr>
<tr>
<td><strong>Description</strong> – This year our CSiO Summer School was held in conjunction with the IEEE</td>
</tr>
</tbody>
</table>
Information Theory Society’s Summer School. A success by all measures, it allowed the Center to continue its efforts to involve a wider range of students and faculty from outside the Center. Please see Section III Education for more information.

2b. Other Outcomes and Impacts of Knowledge Transfer Activities

Describe any other outcomes or impacts of knowledge transfer activities not listed above. Discuss, in particular, applications of Center research in industry, Federal Laboratories or elsewhere not discussed above.

Several additional knowledge transfer activities occurred during this last reporting period. These activities are described elsewhere in the report and are summarized briefly below. These included:

- Development of the Science of Information course and online teaching/tutorial modules.
- Center for Science of Information website.
- Invited talks and seminars.
- Industrial, governmental and international collaborations
- International collaborations with LINCS (Paris) and the Colombian government
- Visiting faculty exchanges

We elaborate on these outcomes in the following sections.

Development of the Science of Information Community of Practice Learning Platform

The full development of our Science of Information Community of Practice Learning Platform now offers itself as an excellent method of outreach and knowledge transfer as we ramp up collaborations among other higher education institutions to further develop teaching modules and actually teach the course. Please see Section III Education, for more details.

Center for Science of Information Website

The CSiO website continues to highlight Center activities and draws more and more users to our offerings. Please see web metrics below for the period Nov 1, 2012 – October 31, 2013.

17,254 visits (vs 16,499 the previous period, a 4.58% increase).

Of these, 10,129 were unique visitors (vs. 9,139 the previous period, a 10.83% increase).
57.47% were new visits – meaning visitors came to the site for the first time

In addition, it is worth mentioning that our Information Theory School page on the IT Society website received 4,065 views.

Visitors come from all over the world with 121 countries/territories represented. The top 10 countries represented by the number of visitors to the website are: (1) United States, (2) India, (3) Germany, (4) China, (5) Canada, (6) France, (7) United Kingdom, (8) Poland, (9) Iran, and (10) Colombia.

The Center’s video tutorial and seminar content quadrupled, going from 3,185 views in period 3 to over 13,000 in period 4. This indicator points to dramatic growth and dissemination of the Center’s online offerings.

**Invited Talks and Seminars 2013**

Center investigators presented a large number of seminars, worldwide. They had tremendous impact on the Center’s continuing goal of disseminating results of faculty research.

Below is a list of the 176 talks and seminars presented.

1. C. Aguilar. Presented at the Department of Biology, IUPUI, Indianapolis, IN, February 2011.


3. C. Aguilar. Presentation at the Institute for Biological and Physical Chemistry, University of Buenos Aires, January 2012.

4. V. Anantharam. “What risks lead to ruin?” Invited talk delivered at the Network Science Workshop, Chinese University of Hong Kong, Hong Kong, China, July 2012.


7. V. Anantharam. “Pointwise lossy source coding theorems for sources with memory.” Talk delivered at the MURI review meeting, Ohio State University, Columbus, Ohio, October 2012.


10. V. Anantharam. “What risks lead to ruin?” Distinguished lecture delivered at the Cal-IT Center, University of California at San Diego, November 2012.


20. E. Ben-Sasson. “Universal and affordable Computational Integrity, or, succinctly, from C to PCP,” Presented at the NYU theory seminar, April 2013.

22. E. Ben-Sasson. “Succinct Computational Integrity and Privacy Research (SCIPR),” Presented at the workshop on Computational Complexity, Banff International Research Station, Banff, CA, July 2013.


27. W. Bialek. “How much can we calculate?” Presented at the physics section meeting, National Academy of Sciences, April 2013.


47. J. Gallant. “Voxel-wise modeling and decoding” (VWMD).” Presented at the SAMSI Workshop on Neuroimaging Data Analysis, Research Triangle Park, NC, 2013.


52. J. Gallant. “Beyond localization: Systematic maps of visual and linguistic information across the human brain.” Presented at grand rounds, Departments of Psychiatry and Neurology, University of Texas Southwestern Medical Center, Dallas, TX, 2013.

53. J. Gallant. “Beyond localization: Systematic maps of visual and linguistic information across the human brain.” Presented at the Neurobiology and Anatomy Seminar Series, Department of Neurobiology and Anatomy, University of Texas Medical School, Houston, TX, 2013.


60. J. Gallant. “Reverse engineering the human brain.” Presented to the Department of Informatics, University of Amsterdam, Amsterdam, Netherlands, 2013.


68. A. Goldsmith. “Enabling the Wireless Cloud through Software-Defined Networking.” Presented a plenary talk at IEEE/CIC International Conference on Communications in China, Xi’an, China, August 2013.


78. V. Kostina. “Lossy joint source-channel coding in the finite blocklength regime.” Presented at the Technische Universitat Munchen, Germany, August 2012.


80. V. Kostina. “Non-asymptotic lossy compression.” Presented at the University of Illinois at Urbana-Champaign, Urbana-Champaign, IL, April 2013.


87. D. Kumar. “What is Information?” Talk to be given at The Quadrangle, an elder community in Haverford, PA, Fall 2013.

88. P. R. Kumar. Presented the keynote address at IEEE Wireless Communications and Networking Conference (WCNC 2013), Shanghai, China, April 2013.


90. P. R. Kumar. Presented the keynote talk at International Conference on Computing, Networking and Communications (ICNC), San Diego, CA, January 2013.

91. P. R. Kumar. Presented at the Center for Pervasive Communications and Computing Distinguished Seminar, University of California-Irvine, Irvine, CA, October 2012.

93. P. R. Kumar. Presented as the keynote speaker at the 7th International Conference on Wireless Algorithms, Systems, and Applications (WASA 2012), Yellow Mountains, China, August 2012.


97. C. Liu. “Predicting Protein-Protein Interactions Based on PPI Networks.” Presented at the 1st International Conference on Information and Intelligent Computing (ICIIC 2012), Chengdu, China, December 2012.


African Languages.” Presented at the 84th National Technical Association Conference (NTA 2012), Baltimore, MD, September 2012.


111. O. Milenkovic. “Novel Distance Measures for Vote Aggregation.” Presented at the University of Maryland, Baltimore, MD, October 2012.


118. A. Qi. “Bayesian approaches for correlated variable selection and online learning.” Presented at the Duke Workshop on Sensing and Analysis of High-Dimensional Data, Duke University, Durham, NC, July 2013.


120. A. Qi. “Scalable Bayesian learning for matrices and tensors.” Presented at the SAMSIFODAVA Workshop on Interactive Visualization and Analysis of Massive Data, December 2012.


122. A. Qi. “Constrained Sparse Bayesian Models for Selection of Imaging biomarkers.” Presented the keynote address at the International Molecular Imaging Summit, Beijing, China, August 2012.


129. A. Qi. “Scalable Bayesian learning for complex data.” Presented to the Computer Science Department, John Hopkins University, Baltimore, MD, March 2012.

130. A. Qi. “Scalable Bayesian learning for complex data” Presented at Beijing University, Beijing, China, June 2011.


140. M. Sudan. “Communication Amid Uncertainty.” Presented the EECS Distinguished Lecture at Northwestern University, Chicago, IL, October 2012.


146. W. Szpankowski. Presented as a Distinguished Colloquia Speaker, Concordia, Montreal, Canada, 2011.

151. W. Szpankowski. Presented a plenary talk (1/7) at the Polish Combinatorial Conference, Bedlewo, Poland, 2012.
159. S. Verdú “Non-asymptotics in Information Theory.” Presented to the Center for Control, Dynamical Systems and Computations, University of California-Santa Barbara, Santa Barbara, CA, February 15, 2013.
164. M. Ward. Presented seminar at the 4th biennial Canadian Discrete and Algorithmic Mathematics Conference (CanaDAM), St. John's campus of Memorial University of Newfoundland, June 2013.


170. T. Weissman. “Fundamental limits and algorithms for some modern data processing problems.” Presented keynote address at the Annual Science of Information Center Summit, Purdue University, West Lafayette, IN, December 2012.


175. B. Yu. Presented at the 11th International Workshop on Multiple Classifier Systems (MCS 2013), Nanjing University, China, May 2013.

176. B. Yu. Presented at the Fong Symposium at San Francisco State University, Mathematics Department, San Francisco, CA, April 2013.
Industrial, Governmental, and International Collaborations

Center-supported research has spawned and fostered many collaborations between Center members and industry. The following are examples of such projects and new collaborations.

Alan Qi and his team have been working with researchers from Eli Lilly and Company to model high-throughput RNA-sequencing data. They have developed a new Bayesian model, coupled with parallel computing, that, based on massive RNA-sequencing data, can model and detect isoforms linked to liver cancers. The developed tool is shared with Eli Lilly and Company. Qi has also worked with Yahoo on an integrated probabilistic model for reach- and frequency- modeling.

Ananth Grama is working with John Deere Corporation on projects related to handling larger sensor datasets. He is also working with Pfizer Corporation and organizing a proposal for a joint project between the Center and Pfizer Corporation.

P.R. Kumar is developing, with National Instruments, a MAC on FPGA that will allow programmability and experimentation at the MAC level. Hitherto experimentation has largely been confined to the network layer and above.

Bin Yu is discussing research discussions projects with Chris Dance from Xerox European Research Center. She also reports that her Center-supported “causal inference” work has been useful in consulting activities with Ebay, and her “stability” work is being used in a collaborative project with biology researchers at LBNL.

Andrea Goldsmith and her team frequently host visitors to their lab from NEC. Many of the Goldsmith lab’s ideas on software-defined wireless networks are being considered for future NEC network architectures. A CSoI-funded graduate student from the Goldsmith group interned at Broadcom this summer and brought some of their network optimization ideas to their chip design. Another student interned at Microsoft and used matrix completion and sparsity ideas for big data research there. Working with Huawei, Dr. Goldsmith advised on the state-of-the-art in reducing energy consumption of backbone cellular infrastructure. She has also collaborated with Ericsson on blind null space learning in MIMO cognitive radios to allow them to coexist with primary users without interfering with them.

Suresh Jaganathan is working with Microsoft Research on characterizing complexity/simplicity of proofs in program verification. This information will be used to understand
opportunities for proof reuse, optimization, and readability by precisely characterizing, in an information theoretic sense, simplicity measures in proof construction.

Jennifer Neville collaborates with Paul Bennett at Microsoft on active sampling methods for networks. They are developing principled methods to learn predictive models for a partially observed network, while actively querying for information about unseen parts. She has also worked with Lawrence Livermore National Laboratories (LLNL) on Anomaly Detection for Dynamic Networks – modeling node behavior in dynamic networks and their roles, and automatically detecting deviations from this behavior. She also interacts with Brian Gallagher at LLNL.

Doriswami Ramkrishna is working with Abbott Laboratories on control of crystal morphology towards better pharmaceutical products – exploring model-based control of super-saturation to manipulate crystal faces that increase bioavailability, and hence effectiveness of drugs. Dr. Ramkrishna is also in early discussions with Pfizer Corporation to work on a biological systems project dealing with personalized medicine.

David Tse closely collaborates with Asif Khalak of Pacific Biosciences (a sequencing company) on several research projects. In addition, he is initiating a collaborative project with Alex Copeland at the Joint Genome Institute (a DoE laboratory).

Luo Si jointly published his collaborative research with Yahoo researchers – “Forecasting User Visits for Online Display Advertising” (Journal of Information Retrieval). He has also partnered with Google on Expertise Search in Complex Networks – integrating heterogeneous information sources to identify experts and expertise in networks. In addition, Dr. Si designed and implemented the MapIN project, which visualizes expertise in Indiana Companies – a project with the Indiana Economic Development Corporation (IEDC).

Nancy Lynch continues to collaborate with Peter Musial at EMC and Srikanth Sastry at Google.

Sergio Verdu collaborates with Dr. A. Tulino at Alcatel-Lucent Bell Laboratories.

Wojtek Szpankowski published a paper with researchers from Amgen in Pharmaceutical Research, 2012. He has worked with HP on lossless compression for large alphabets to develop methodology to understand minimax redundancy for unbounded alphabets.

Tsachy Weissman is working with members of the Google Geo team on lossy compression for image compression. He also partners with the Samsung Israel Research Lab on real-time and limited delay compression and error correction.
Many of these interactions have also resulted in direct funding of projects from industry. These include:

Funding Institution: Microsoft Research  
Project Title: Active Sampling in Networks  
Amount: $25,000  
PI: J. Neville

Funding Institution: Lawrence Livermore National Laboratories  
Project Title: Modeling Graph Dynamics for Anomaly Detection  
Amount: $25,000  
PI: J. Neville

Funding Institution: Intel Corp.  
Project Title: Wireless Video Delivery  
Amount: $100k  
PI: P. R. Kumar.

Funding Institution: National instruments  
Project Title: Wireless Medium access control  
Amount: $15.8K  
PI: P. R. Kumar

Funding Institution: Huawei Technologies Do., Ltd.  
Project Title: Non-coherent communication schemes for multiuser massive MIMO systems  
Amount: $70,000  
PI: A. Goldsmith

Funding Institution: CableLabs  
Project Title: Cognitive Massive MIMO Wireless Networks  
Amount: $75,000  
PI: A. Goldsmith

Funding Institution: NEC China  
Project Title: Cloud-based architectures for future cellular systems  
Amount: $40,000 (processed through Stanford's Computer Forum)  
PI: A. Goldsmith

Funding Institution: Samsung  
Project Title: Heterogeneous networks with multiple radios.  
Amount: $20,000 (processed through Stanford's Computer Forum)  
PI: A. Goldsmith
Funding Institution: Xerox European Research Center  
Project Title: km-resolution aerosol retrieval based on remote sensing images  
Amount: $30K per year for three years  
PI: B. Yu

Funding Institution: Deutsche Telekom T-Labs  
Project Title: Distributed learning and collective decision-making  
Amount: $346,000  
PI: S. Kulkarni

Funding Institution: Google, Inc  
Project Title: Huds up Display to manage Healthcare for Users with Diabetes (HUD)^2.  
Amount: $25,000  
PI: T. Coleman.

Funding Institution: Yahoo!  
Project Title: An integrated Probabilistic Model for Reach and Frequency  
Amount: $15,000  
PI: L. Si

Funding Institution: Google  
Project Title: Expertise Search in Complex Information Network  
Amount: $60,000  
PI: L. Si

Funding Institution: Indiana Economic Develop Council (IEDC)  
Project Title: INDURE: Indiana Database of University Research Expertise (II) Map Indiana  
Amount: $150,000  
PI: L. Si

Funding Institution: Eli Lilly and Company  
Project Title: Isoform detection based on massive RNA-sequencing data for Hepatocellular carcinoma  
Amount: $160,000  
PI: A. Qi

Funding Institution: Microsoft Research  
Project Title: Network models for bioinformatics  
Amount: $100,000  
PI: A. Qi

Funding Institution: IBM  
Project Title: Big data for information security  
Amount: $100,000  
PI: A. Qi
2c. Progress in Meeting Metrics

Describe how the Center is doing with respect to the indicators/metrics listed above. Include any data that have been collected on the indicators/metrics.

The Center is meeting or exceeding most of the indicators/metrics listed above in 1b.

While progress was made in respect to industry collaboration and relationships, it has been necessary to reassess how the Center engages industrial companies. Excellent progress has been made in regards to industrial support of Center-supported research and the addition of collaborative work between the Center and industry. As mentioned above, with the recommendation of the 2012 NSF Site team, the Center formed an industry advisory committee to advise and assist with industry engagement and with the additional input from the Center’s External Advisory Committee – now have a clear model for enlisting industrial participation and input.

The Center made exceptional progress towards its educational goals, especially as it relates to the development of the SoI Course and Educational modules. This is evident by the increasing number of institutions outside of the Center that are teaching the Science of Information course.

Clear evidence exists that the Center’s online resources (publications, videos, educational modules/course materials) are making a significant impact.

The Center’s international partnerships are being formalized and moving forward, and the potential is high for additional collaboration in the coming year. Faculty members are extremely active internationally with their Center supported research initiatives.

The CSol website is well-designed and is an excellent resource for facilitating research across the Center by providing communication and research tools to all center participants.

CSol students are in high demand and have many opportunities for internships, job placement, and academic pursuits. In addition, the Center has provided significant opportunities for student-led interdisciplinary research teams to form and work together.
2d. Plans for Changes in Knowledge Transfer Activities, Research Direction, and New Partnerships

For annual reports: Describe your plans for knowledge transfer activities for the next reporting period with attention to any major changes in direction or level of activity. Include plans for new knowledge transfer partnerships, if any.

The following specific knowledge transfer activities are planned:

- Hold our second annual industry advisory meeting with a goal of increasing participants.
- Continue to engage companies/labs/institutions (through reciprocal visits, technical exchanges, targeted collaborations, and development efforts). A major emphasis will be placed on hosting research seminars and speakers from industry.
- Maintain support for students as they pursue internships, jobs, or academic positions. Additional support is made available for students to present relevant research at conferences, workshops, etc.
- In order to better communicate the Center’s efforts and successes, we will formalize a “Center Communications Plan” that emphasizes targeted messaging, utilizing various delivery methods, to relevant organizations and people throughout the world.
- Continue to develop the Science of Information Hub.
- As part the Sol Course that is being taught at institutions outside the Center, we will work to make all content available to a wide audience via the web.
- Coordinate an additional workshop that prepares faculty members to teach the Sol course
- Expand our working international relationships with LINCS (Paris) and Colombia. The Center will pursue potential research and knowledge transfer collaborations with representatives from a number of other countries (current talks include organizations in Argentina, China, and India).
V. External Partnerships

Goals, Objectives, Metrics, and Problems Encountered

1a. Partnership Goals and Objectives

Describe the Center's overall goals and/or objectives for developing external partnerships. If the Center's overall partnership goals/objectives have changed since the last reporting period, how did they change and why? [In section 2a below, please describe progress the Center has made toward reaching these goals/objectives.]

The Center's overall goals and objectives for developing external partnerships include developing partnerships (local, state, national, and international) with any entity interested in meaningful research, education, and diversity collaboration that can enhance the Center's capabilities and outreach—either through joint research, internships, exchange opportunities, or the like.

1b. Performance, Management Indicators and Metrics

Inform us of the performance and management indicators the Center has developed to assess progress in meeting its partnership goals/objectives.

- Industrial companies/research labs initiating research projects involving Center research and initiatives.
- Industrial companies engaged through introductory/exploratory meetings, presentations, industry affiliates program etc.
- Consultations made to industry.
- Involvement of industry partners in seminars, research exchanges, and opportunities for students.
- International collaborations and partnerships.
- Higher education partners beyond the Center.
- Educational and diversity partners.

1c. Problems Encountered in Reaching Goals

Discuss any problems you have encountered in making progress toward the Center's partnership goals/objectives during the reporting period as well as any problems anticipated in the next period. Please include your plans for addressing these problems.

None to report.
Knowledge Transfer Recipients

2a. External Partnership Activities

Describe and discuss the activities that are conducted as part of partnerships, which are not listed in another section of this report. Be sure to discuss how the Center’s partnership activities enable the Center to meet its partnership goals/objectives listed above.

### Partnership Activity – Research, Technology/Knowledge Transfer

*Led by – W. Szpankowski*

<table>
<thead>
<tr>
<th>Organizations Involved</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name of Organization</td>
</tr>
<tr>
<td>Gdansk University of Technology, Poland</td>
</tr>
</tbody>
</table>

**Outcomes or Impacts** – International activity involving research exchange and opportunities to promote the work of the Center. This collaboration will continue in future years.

### Partnership Activity – Teaching/Education, Technology/Knowledge Transfer

*Led by – A. Grama*

<table>
<thead>
<tr>
<th>Organizations Involved</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name of Organization</td>
</tr>
<tr>
<td>Colombian Government/EAFIT University – Juan Lalinde</td>
</tr>
</tbody>
</table>

**Outcomes or Impacts** – The Center has now entered into a formal agreement with EAFIT University to not only assist in developing and teaching our course, but also to translate our course and educational materials into Spanish. The potential impact is obviously enormous.

### Partnership Activity – Research, Student and Faculty Exchange, Technology/Knowledge Transfer

*Led by – W. Szpankowski, B. Brown*

<table>
<thead>
<tr>
<th>Organizations Involved</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name of Organization</td>
</tr>
<tr>
<td>Laboratory of Information, Network and Communication Sciences (LINCS)</td>
</tr>
</tbody>
</table>

**Outcomes or Impacts** – this collaboration has led to a strong international relationship which broadens presence in Europe and has led to faculty/student research exchanges.
**Partnership Activity – Diversity, Research, Education, Student Exchange**

*Led by – B. Gibson*

<table>
<thead>
<tr>
<th>Organizations Involved</th>
<th>Name of Organization</th>
<th>Shared Resources</th>
<th>Use of Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Inclusion, Diversity, Excellence, and Advancement (IDEA) Student Center at UCSD</td>
<td>None</td>
<td>Undergraduate Research Support</td>
</tr>
</tbody>
</table>

**Outcomes or Impacts –**  This has been an excellent opportunity to positively leverage our diversity efforts with another successful program at UC San Diego, thus giving opportunities to even more students.

**Partnership Activity – Diversity, Research, Education**

*Led by – B. Gibson*

<table>
<thead>
<tr>
<th>Organizations Involved</th>
<th>Name of Organization</th>
<th>Shared Resources</th>
<th>Use of Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Summer Research Opportunities Program (SROP) at Purdue</td>
<td>None</td>
<td>Undergraduate Research Support</td>
</tr>
</tbody>
</table>

**Outcomes or Impacts –**  Another example of a partnership with another group that has allowed us to better leverage our Center program and ultimately maximizing our impact.

**Partnership Activity – Research, Technology/Knowledge Transfer**

*Led by – A. Grama, W. Szpankowski*

<table>
<thead>
<tr>
<th>Organizations Involved</th>
<th>Name of Organization</th>
<th>Shared Resources</th>
<th>Use of Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bell Labs/Alcatel Lucent</td>
<td>Research, research presentations, website, student opportunities</td>
<td>Support for research seminar</td>
</tr>
</tbody>
</table>

**Outcomes or Impacts –**  An excellent opportunity to share knowledge and receive input on our agenda with a leading company in our field. We expect positive impacts for future student and faculty opportunities to further interact with Bell Labs.
### Partnership Activity – Research, Technology/Knowledge Transfer

**Led by – A. Grama, B. Brown**

<table>
<thead>
<tr>
<th>Name of Organization</th>
<th>Shared Resources</th>
<th>Use of Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Deere Corp.</td>
<td>Data, research, student research</td>
<td>Support for student research</td>
</tr>
</tbody>
</table>

*Outcomes or Impacts – This partnership has provided an opportunity to work directly with a company as it relates to its Big Data needs and strategies. It has provided undergraduate research opportunities and recently the Center was entrusted with the company's database as we move into more intense research.*

### Partnership Activity – Education (course/module development), Knowledge Transfer

**Led by – B. Ladd, M. Ward, D. Kumar**

<table>
<thead>
<tr>
<th>Name of Organization</th>
<th>Shared Resources</th>
<th>Use of Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rose Hulman University – Yoshi Shibberu</td>
<td>Web-based course materials, online modules</td>
<td>Course development support, travel/meeting support</td>
</tr>
</tbody>
</table>

*Outcomes or Impacts – See Villanova below.*

<table>
<thead>
<tr>
<th>Name of Organization</th>
<th>Shared Resources</th>
<th>Use of Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oregon Health and Science University – Karen Watanabe</td>
<td>Web-based course materials, online modules</td>
<td>Course development support, travel/meeting support</td>
</tr>
</tbody>
</table>

*Outcomes or Impacts – See Villanova below.*
### Partnership Activity – Education (course/module development), Knowledge Transfer

Led by – B. Ladd, M. Ward, D. Kumar

<table>
<thead>
<tr>
<th>Name of Organization</th>
<th>Shared Resources</th>
<th>Use of Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carnegie Melon University – Pulkit Grover</td>
<td>Web-based course materials, online modules</td>
<td>Course development support, travel/meeting support</td>
</tr>
</tbody>
</table>

**Outcomes or Impacts** – See Villanova below.

### Partnership Activity – Knowledge Transfer, Education, Workshop

Led by – M. Ward

<table>
<thead>
<tr>
<th>Name of Organization</th>
<th>Shared Resources</th>
<th>Use of Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Center for Coastal Margin Observation and Prediction (CMoP) STC</td>
<td>Data modeling workshop for faculty/students</td>
<td>Workshop development support, travel/meeting support</td>
</tr>
</tbody>
</table>

**Outcomes or Impacts** – This partnership was formed with one of our sister NSF Science and Technology Centers (CMoP). We are coming together to offer a data modeling workshop for faculty and students that can be offered to other STC’s as well. Because of the vast reach of the STC system the potential impact of this partnership is substantial.

### Partnership Activity – Curriculum Development, Education, Knowledge Transfer

Led by – D. Kumar, D. Xu

<table>
<thead>
<tr>
<th>Name of Organization</th>
<th>Shared Resources</th>
<th>Use of Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sidwell Friends High School</td>
<td>Creating methodologies for introductory computing courses</td>
<td>Workshop support, travel/meeting support</td>
</tr>
<tr>
<td>Martin High School</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Outcomes or Impacts** – Created in part with support from the Center, this is a K-12 outreach partnership that works to create methodologies for introductory computing courses and is leveraged with support from other grants.
### Partnership Activity – Education, Workshop, Knowledge Transfer

**Led by – D. Kumar, D. Xu**

<table>
<thead>
<tr>
<th>Organizations Involved</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Name of Organization</strong></td>
</tr>
<tr>
<td>Association for Advancement of Liberal Arts Colleges (AALAC)</td>
</tr>
</tbody>
</table>

*Outcomes or Impacts – Another effort as part of the Center’s commitment to offering professional development to faculty on teaching the Science of Information. This activity has started out very strong and we expect will continue to grow, offering itself to even greater impact in the future.*

### Partnership Activity – Education, Workshop, Knowledge Transfer

**Led by – D. Kumar, D. Xu**

<table>
<thead>
<tr>
<th>Organizations Involved</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Name of Organization</strong></td>
</tr>
<tr>
<td>Association of Computer Machinery – SIGCSE</td>
</tr>
</tbody>
</table>

*Outcomes or Impacts – Similar partnership as that above – offering faculty development and training.*

### Partnership Activity – Education, Workshop

**Led by – D. Kumar, D. Xu**

<table>
<thead>
<tr>
<th>Organizations Involved</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Name of Organization</strong></td>
</tr>
<tr>
<td>Computer Science Teachers Association</td>
</tr>
</tbody>
</table>

*Outcomes or Impacts – Another K-12 outreach partnership supported in part by the Center. This activity concentrates on training K-12 teachers to teach computing concepts.*
**Partnership Activity** – Education (course/module development), Knowledge Transfer

**Education (course/module development), Knowledge Transfer**

*Led by – B. Ladd, M. Ward, D. Kumar*

<table>
<thead>
<tr>
<th>Organizations Involved</th>
<th>Shared Resources</th>
<th>Use of Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Villanova University – Anil Bamezai</td>
<td>Web-based course materials, online modules</td>
<td>Course development support, travel/meeting support</td>
</tr>
</tbody>
</table>

**Outcomes or Impacts** – This partnership was formed as a means of further developing and teaching our Science of Information course and associated learning modules. It provides us with the opportunity to get valuable input from academic partners outside of the Center and to enlist a broader range of participants in this Center endeavor.

The Center continues to make progress toward its overall goals and objectives for developing external partnerships in all areas. During the last period, existing partnerships were enhanced, and new partnerships were created. The Center now has a full range of partnerships at the state, national, and international levels. These relationships have facilitated technology transfer, education/diversity activity, and overall Center outreach.

**International Partnerships**

As a result of the Center’s partnership with the Laboratory of Information, Networking and Communication (LINCS, Paris) and an established faculty research exchange program, several Center members, including Venkat Anantharam (UC Berkeley), P.R. Kumar (Texas A&M), and Wojtek Szpankowski (Purdue), visited LINCS during the reporting period. Phillipe Jacquet from LINCS twice visited CSoI headquarters at Purdue University and also participated in the CSoI/Bell Labs research workshop at Princeton University. In the coming year we will continue the faculty exchanges and add a student exchange of at least one student from each partner.

With the help of Dr. Juan Lalinde, the Center now has a formal agreement with Colombia that will broaden the Center’s educational development and outreach efforts while also expanding its faculty/student exchange opportunities. In addition, Dr. Lalinde has agreed to work with the Center to further develop the Science of Information course including translating the existing content into Spanish. Combined, this effort will dramatically increase the number of individuals the Center can potentially impact.

The Center is also in early discussions with faculty representatives from Argentina to create a partnership similar to the one with Colombia.

**Industry Partnerships**

As discussed in Section IV. Knowledge Transfer, CSoI has developed new strategic partnerships with Bell Labs and the John Deere Corporation. At the suggestion of last year’s site visit team, and to
continue to strengthen and develop industrial partnerships, the Center has established an industry advisory group – this is also discussed further in the Knowledge Transfer section IV.

**Education/Diversity Partnerships**

During this last reporting period, perhaps the most impactful partnership was with the IEEE IT Society. As reported in Section III. Education, CSol partnered with the IEEE IT Society to co-sponsor the North American IT Society summer school. It was a great success by all measures.

All other activity listed with the above partners is indicative of diversity/education sections of this report. It is also worth noting that the Center continues to grow its education and outreach efforts through multiple external higher education partnerships. Specifically, CSol partner Bryn Mawr College has established a number of initiatives in faculty/teacher development and course development for science of information related courses and workshops.

**2b. Outcomes and Impacts of External Partnership Activities**

*Describe any other outcomes or impacts of partnership activities not listed elsewhere.*

The Center has no additional outcomes or impacts to report.

**2c. Progress in Meeting Metrics**

*Describe how the Center is doing with respect to the indicators/metrics listed above. Include any data that have been collected on the indicators/metrics.*

The Center continues to make progress in its efforts to develop meaningful relationships with external partners, and now has multiple activities in each area marked as an indicator for success. Through our big data workshop, a stand-alone research workshop with Bell Labs, and an industry advisory board meeting in Chicago, we have actively worked to engage a wider representative field of industry. The goal of the industry advisory board is to help grow industry involvement and consultation in all aspects of our work. Center students are in high demand for internships and jobs after graduation, and there is solid industry support for a wide range of Center-related research.

*Center faculty are highly involved in international activities, and the Center now has several meaningful international collaborations that we expect to grow and lead to additional collaborations in future years.*

Our education and diversity partnerships are strong and growing. In addition, multiple new partnerships have been formed this year as the Center continues to develop its science of information course and other available educational modules.

In the coming year, the Center expects the number of CSol partnerships to continue to increase as additional industry representatives and higher education partners are engaged in teaching and developing the Science of Information course and educational modules.
2d. Plans for Changes in External Partnership Activities, Research Direction and New Partnerships

For annual reports: Describe your plans for partnership activities for the next reporting period with attention to any major changes in direction or level of activity.

The following planned activities will enhance external partnerships while paralleling Center knowledge transfer activities:

- Continue to secure suitable internship positions at relevant companies.
- Continue hosting researchers from industry at the Center and at related workshops or seminars.
- Seek input and involvement from industry relating to research opportunities and challenges whenever possible.
- Continue to solicit partnerships and develop mutual collaborations/exchanges. Meet with potential partner institutions and explain guidelines and benefits of partnership.
- Conduct reciprocal visits/exchanges with industry, including industry-led lectures and/or workshops throughout the Center.
- Enhance international partnerships and collaboration while seeking to add partners when appropriate.
IV. Diversity

Goals, Metrics and Problems Encountered

1a. Overall Diversity Goals/Objectives and CSoI Channels Program

Describe the Center's overall goals or objectives related to increasing diversity at the Center if there have been any changes in the Center’s overall goals/objectives and plans related to increasing diversity since the last reporting period, please discuss these changes and the reasons behind them. [In section 2a below, please describe progress the Center has made toward reaching these goals/objectives.]

Overall Diversity Goal

Broaden participation and success of underrepresented groups among students, faculty, and the STEM workforce in the Science of Information (SOI) disciplines.

Broadening participation within STEM is needed to enhance and facilitate training of the next generation of scientists within SOI fields of study and to increase workforce development in these fields. Therefore our long-term vision is to create a vibrant and diverse community of students and faculty engaged in the education, research, and scholarly activities of the center. We strive to be among the leaders in increasing diversity among students, scientists, and the STEM workforce in the areas relating to the science of information.

The goal of broadening participation within the disciplines of the Science of Information is integrated throughout all Center activities.

Objectives

The specific efforts that will target recruitment and retention of underrepresented groups in research and education initiatives within CSoI are:

- Develop and offer and create a Center-wide Research Experience for Undergraduates (REU) program that targets underrepresented students.
- Promote participation of non-Center post-docs and faculty from underrepresented groups in Center activities.
- Offer professional development meetings/activities to students, especially those that are underrepresented, to Center students and post-docs in order to make them more competitive when applying to graduate school and entering the workforce.
- Increase the overall diversity within the various Center classifications (student, faculty, post-doc) with respect to underrepresented groups within STEM, including females, minorities, first generation, and low-income students.
**CSol Channels Program**

Broadening participation within STEM disciplines is needed to enhance and facilitate training of the next generation of scientists within SOI fields of study and to increase workforce development in these fields.

Over the past year, CSol has continued to expand and enhance its Channels Program. The aim of the Channels Program is to diversify the Science of Information community by providing a broad range of opportunities for undergraduate and graduate students, post-docs, and faculty who are members of underrepresented groups through pathways that enable them to flourish in the field of Science of Information. These pathways or “channels” include undergraduate internships, summer research opportunities, graduate assistantships, postdoctoral mentoring travel awards, a professional development program, and collaborations between faculty at MSIs and Center scientists who are leaders in the field of Science of Information.

Recruitment of underrepresented individuals and groups into the Center’s activities related to education and research will be carried out by various efforts, including but not limited to:

- Outreach activities by various CSol students, faculty/researchers, and Diversity staff to underrepresented K-12 students interested in STEM. This could include having CSol undergrad and graduate students mentor K-12 students with science fair projects, or visiting classrooms to co-teach for a day.

- Distributing informational brochures. This literature is designed to be inclusive towards targeting minorities into CSol research and education activities.

- Facilitating collaborations between CSol researchers and faculty/students at MSI/HBCUs. Besides setting up possible research collaborations between faculty, ideally this could lead to recruiting of underrepresented graduate students to CSol partner institutions as well as new faculty within the Center; and

- Collaborations with underrepresented serving STEM groups (such as SACNAS, UCSD IDEA Center, ADVANCE, NGCP and others) and fellow leaders in the area of broadening participation in STEM disciplines. Establishing and maintaining such liaisons throughout the lifetime of the Center will lay a strong foundation for recruiting, mentoring, placement, and advancement of students, faculty, and researchers.

While the program will continue to evolve as needed, several initiatives of the Channels Program have been implemented and will continue to be advanced Y4 -Y5:

- Development and creation of a CSol Channels Program targeted at underrepresented undergraduate and graduate students doing research with the disciplines related to the Science of Information. Besides being active in CSol research, these students will be required to participation in a limited number of outreach/education efforts.
- A distributed academic year REU (CSol Undergraduate Channels Scholars) within the Center was implemented in Fall 2012 to target underrepresented STEM students and to provide a supportive cohort experience. This program will promote inclusion of a diverse population of undergraduates in Center research and educational initiatives. Currently 5 students are supported within the Center for the 2012-13 academic year.

- Similar to the CSol Channel Scholars Program, the CSol Graduate Channels Fellows Program provides a supportive environment for underrepresented graduate students pursuing research projects within the fields of Science of Information. These graduate students can also serve as peer mentors to CSol Channel Scholars. Currently the program supports one female graduate student.

- Development of Channels Faculty and Post-Doc Research Collaborations. Borrowing best practices from the National Posse Program, (http://www.possefoundation.org/) faculty and post-doc researchers from MSIs, or minority faculty or post-docs from other institutions, will visit a CSol Center facility or participate in a CSol function. This will expose and afford the participation of these faculty and post-docs to cutting edge research within Science of Information, and facilitate future collaborations after they return to their home institutions.

- Creation of a Professional Development Program that will provide CSol undergraduates, graduate students, and post-docs, especially those from underrepresented groups, with career, presentation, and enhanced research skills to increase their competitiveness in the workforce after they graduate. This will include, but are not limited to, developing and giving workshops on:
  - proposal writing,
  - resumes/CVs,
  - scientific presentations/posters,
  - submitting papers for publication,
  - applying to graduate school (undergrads only)
  - diversity within STEM, and
  - communicating science to the public.

Some of the topics will be fluid and be based on the needs and interests of the CSol students.

The CSol Channels Program will ensure that students and post-docs receive training targeted to broaden their areas of expertise in order to address the emerging interdisciplinary areas of research. We plan to collaborate with other STC’s on professional development materials and resources, as well as leverage partnerships with established programs such as the IDEA Center at UCSD that supports success of underrepresented students in engineering fields.
1b. Evaluation and Metrics, with Measurable Outcomes from Diversity Efforts

*Inform us of the performance and management indicators the Center has developed to assess progress in meeting its diversity goals/objectives.*

The goal of increasing diversity within CSol (and in turn SOI fields) will be integrated throughout all Center activities, with the Channels Program being the primary pathway. Below are noted specific outcomes and metrics that will be used to measure the success for targeting underrepresented groups for participation in research and education initiatives within CSol.

**Measurable Outcomes from Diversity Efforts Include**

- Broadening participation of underrepresented groups within all levels of research and education initiatives within CSol.
- Providing a collaborative and supportive cohort environment within Center activities to promote interest and success in SOI research with underrepresented students and post-docs.
- Providing underrepresented groups within CSol with cutting edge research experiences and professional development skills unique to SOI to make them more competitive within the emerging workforce.
- Creating informational brochures and other materials promoting CSol for its support of diversity within SOI.
- Broadening participation of underrepresented students through the attendance by Center staff, students, and faculty at appropriate conferences (such as SACNAS and Grace Hopper) for recruiting and presentations of CSol research projects.
- Creating a social network for collaboration/mentoring between CSol undergraduates and graduate students. Can also be used for long term tracking and recruitment of students.
- Offering formalized meetings/workshops related to professional development areas that will help make CSol students more competitive in the workforce or graduate school. Will be presented using multiple formats, including face-to-face and web (Google+ Hangouts).

**Evaluation and Metrics**

- Number of liaisons with STEM organizations and MSIs.
- Number of underrepresented students and post-docs supported by Center.
- Number of underrepresented students, post-docs, and faculty participating in Center activities.
- Number of outside fellowships, scholarships, internships awarded to underrepresented CSol students each year.
• Number of underrepresented undergraduate, graduate students, and post-docs attaining positions with academia, government, industry.

1c. Problems Encountered in Reaching Goals

 Discuss any problems you have encountered in making progress toward the Center’s diversity goals/objectives during the reporting period as well as any problems anticipated in the next period as appropriate. For annual reports include your plans for addressing these problems.

The Director of Diversity has transitioned to a new position at Purdue, and the Center is currently in a search for a replacement. This change of personnel may delay or impact certain programs in an as yet unknown manner.

We are continuing to strive to seek better ways to promote center-wide coordination of diversity activities, especially as related to the engagement of senior faculty. Currently, many of the current activities are spearheaded by the Purdue, Bryn Mawr, Howard, and UCSD campuses. Thus there is continuing need to engage the faculty at other institutions within the Center to directly participate in Center-wide diversity efforts. We will continue efforts to strategize and engage faculty to improve our Center-wide diversity endeavors.

Activities, Impacts, Progress, Plans

2a. Center Activities that Contribute to the Development of U.S. Human Resources in Science and Engineering

 Describe and discuss Center activities which contribute to the development of United States human resources in science and engineering at the postdoctoral, graduate, undergraduate, and pre-college levels. Please pay particular attention to those accomplishments and activities that aim to attract, increase, and retain the participation of US citizens, nationals, or lawfully admitted permanent resident aliens of the United States, women, underrepresented groups, and persons with disabilities. Include a discussion of any partnerships formed that allow the Center to meet its diversity goals/objectives.

Through the CSoI Channels Program, we are providing STEM pathways for underrepresented students, and are in the long-term directly addressing the nation’s needs for the development of human resources within STEM at a nationwide scale. The faculty and staff of CSoI acknowledge the need to broaden the participation of US citizens, nationals, and lawfully admitted permanent residents, women and underrepresented groups and persons with disabilities. Many of their activities to support these efforts are reported in the tables at the end of this section.
Furthermore, the existence of the Center is enabling a cross flow of students (undergraduate and graduate), faculty, and post-docs within all center institutions. For example, the students in the Center-wide Channels Scholars are engaged in several activities and events of the Center, and participate in SOI research projects, professional development, and other outreach programs into the local communities. Through the Channels Program, the CSoI has provided pathways for undergraduate students at Center institutions, as well as faculty from MSIs to participate in Center activities (such as the Summer School), created fruitful liaisons with Center researchers, and received career advice and mentoring.

2b. Outcomes and Impacts of Diversity Programs and Activities

*Discuss the impact of these programs or activities on enhancing diversity at the Center.*

At Bryn Mawr College, several undergraduate fellowships are awarded to a diverse group of undergraduate students who show interest in pursuing careers in the Science of Information. The Center-wide CSoI Undergraduate Channels Scholars Program has awarded this academic year REUs to six (6) underrepresented or female students at Purdue, Howard, and UCSD. These students participate in professional development and peer mentoring activities which helps to strengthen them both professionally and academically. Moreover, a female scientist was recently awarded one of the Center Postdoctoral Fellowships.

2c. Progress in Meeting Metrics by Individual Program or Activity

*Describe how the Center is doing with respect to the indicators/metrics listed above. Include any data that have been collected on the indicators/metrics.*

Specific details related to how CSoI faculty, staff and student activities are fulfilling the metrics as relative to diversity objectives follow.

<table>
<thead>
<tr>
<th>Activity</th>
<th>CSoI Channels Program</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>January 2012 – ongoing</td>
</tr>
<tr>
<td>Audience</td>
<td>undergraduates, graduate students, post-docs, and faculty</td>
</tr>
<tr>
<td>Led By</td>
<td>B. Gibson/D. Kumar/T. Coleman</td>
</tr>
<tr>
<td>Attendees</td>
<td>Pre-college Undergrad Graduate Post-doc Faculty Staff Other</td>
</tr>
<tr>
<td></td>
<td>27 2 2 2</td>
</tr>
</tbody>
</table>

**Brief Description:** The Channels Program of the CSoI provides multiple educational, research and mentoring opportunities for undergraduates, graduate students, post-docs, and faculty to advance in the field of Science of Information. A major goal of the Channels Program is to diversify the Science of Information community as well as the larger STEM communities by providing a broad range of opportunities for students and faculty who are members of underrepresented groups through pathways that enable them to flourish in this dynamic area of scientific endeavor. A grant
A proposal has been submitted to the NSF REU program to create a distributed, summer REU with a focus on SOI and underrepresented students. This program will support 6 students each year for three years in SOI disciplines.

In 2012-13, one female graduate student, one female post-doc and two underrepresented faculty were assisted though the Channels Program ranging from: sponsorship to attend the Summer School, recruiting, mentoring, and subsequent new collaborations with CSol researchers.

The CSol Undergrad Channels Scholars Program began in Fall 2012. Five students are part of the inaugural cohort: three females, two males; two are from underrepresented groups. A portion of one of the student’s travel to a research conference was covered by the program.

Three 2012-13 UG SOI Fellows were awarded at Bryn Mawr College.

### Activity

**Student Networking and Mentoring Lunches w/ Prestige Lecturers**

<table>
<thead>
<tr>
<th>Time</th>
<th>November 2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audience</td>
<td>Undergraduates/Graduate Students</td>
</tr>
<tr>
<td>Led By</td>
<td>B. Gibson</td>
</tr>
<tr>
<td>Attendees</td>
<td>Pre-college, Undergrad, Graduate, Post-doc, Faculty, Staff, Other</td>
</tr>
<tr>
<td></td>
<td>~ 30 people from various groups, mostly students</td>
</tr>
</tbody>
</table>

**Brief Description:** In conjunction with the Purdue Black Graduate Graduate Student Association (BGSA), CSol sponsored lunches with one of our Prestige Lecturer to promote networking and peer mentoring among underrepresented STEM students.

### Activity

**Undergraduate Mentoring**

<table>
<thead>
<tr>
<th>Location</th>
<th>Bryn Mawr College</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>Continuing</td>
</tr>
<tr>
<td>Led by</td>
<td>Deepak Kumar, Dianna Xu, Paul Ruvolo</td>
</tr>
<tr>
<td>Audience</td>
<td>Undergraduate female students</td>
</tr>
<tr>
<td>Participants</td>
<td>~20</td>
</tr>
</tbody>
</table>

**Brief Description:** Mentoring of undergraduate women for careers in computer science and related fields. Worked on extra-curricular projects, academic advising, professional development, research projects, etc.
| Activity | Invited Lecture: Modeling Creativity  
Tom De Smedt, St. Lucas University College of Art & Design, Belgium |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>Bryn Mawr College</td>
</tr>
<tr>
<td>Date</td>
<td>April 4, 2013</td>
</tr>
<tr>
<td>Led by</td>
<td>Deepak Kumar, Dianna Xu, Paul Ruvolo</td>
</tr>
<tr>
<td>Audience</td>
<td>Faculty, Undergraduate women, other students, and public</td>
</tr>
<tr>
<td>Participants</td>
<td>~60</td>
</tr>
</tbody>
</table>

**Brief Description:** This talk presented the thought-provoking question of whether machines can be creative, and explored computational models of creativity, through data mining, multilingual natural language processing, sentiment analysis, machine learning, graph analysis, and visualization.

<table>
<thead>
<tr>
<th>Activity</th>
<th>Postdoc Mentoring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>Bryn Mawr College</td>
</tr>
<tr>
<td>Date</td>
<td>Fall 2012-Spring 2013</td>
</tr>
<tr>
<td>Led by</td>
<td>Deepak Kumar, Dianna Xu</td>
</tr>
<tr>
<td>Audience</td>
<td>Graduating PhD student: Paul Ruvolo (UCSD)</td>
</tr>
<tr>
<td>Participants</td>
<td>1</td>
</tr>
</tbody>
</table>

**Brief Description:** Mentoring for careers in academia (especially liberal arts colleges). Activities included, collaborative research, co-teaching, student advising and mentoring, professional development, job interviewing, placement, etc.

<table>
<thead>
<tr>
<th>Activity</th>
<th>Recruitment of CSol Postdocs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>Bryn Mawr College</td>
</tr>
<tr>
<td>Date</td>
<td>Spring 2013</td>
</tr>
<tr>
<td>Led by</td>
<td>Barbara Gibson and several CSol members</td>
</tr>
<tr>
<td>Audience</td>
<td>Graduating PhD students</td>
</tr>
<tr>
<td>Participants</td>
<td></td>
</tr>
</tbody>
</table>

**Brief Description:** Recruitment of graduating PhDs for CSol Post-doc opportunities. Review and selection of finalists, etc.
### Science Demonstration at the Indiana Black Expo

**Location:** Indianapolis Convention Center  
**Date:** July 19, 2013  
**Led by:** D. Ramkrishna & Purdue College of Education  
**Audience:** Elementary Aged Students  
**Participants:** 40

**Brief Description:** For this activity, we designed a science demonstration to teach the principles of color and light in an engaging way. Children at the Indiana Black Expo had the opportunity to do hands on activities with color mixing and color chromatography.

### Academic workshop, Society for Advancement of Chicanos and Native Americans in Science (SACNAS)

**Location:** Purdue University  
**Date:** September 2012  
**Led by:** SACNAS Purdue Chapter  
**Audience:** Undergraduate and graduate students  
**Participants:** 30

**Brief Description:** M. Ward participated in this workshop. The goals of the workshop were: to connect students with potential mentors; discuss organizations and activities on campus where students can get support from faculty; broaden the students’ understanding of the range of research experiences on campus; discuss conference opportunities with a focus on diversity in academia; faculty share advice and experiences that will be helpful for students in their own college career.

### Panel on diversity

**Location:** ISIT, Istanbul, Turkey  
**Date:** July 10, 2013  
**Led by:** WITHITs (women in the information theory society)  
**Audience:** Women and minority graduate students, postdocs, and young faculty  
**Participants:** 20

**Brief Description:** A. Goldsmith was on a panel and open discussion about the challenges and opportunities facing women and minority young researchers in establishing their academic careers.
Activity: Conference for Africa-American Researchers in the Mathematical Sciences - CAARMS

<table>
<thead>
<tr>
<th>Location</th>
<th>La Jolla, CA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>July 24-27, 2013</td>
</tr>
<tr>
<td>Led by</td>
<td>Todd Coleman</td>
</tr>
<tr>
<td>Audience</td>
<td>African American Researchers in Mathematical Sciences</td>
</tr>
<tr>
<td>Participants</td>
<td>70</td>
</tr>
</tbody>
</table>

**Brief Description:** Addresses critical issues involving African-American researchers and graduate students in the mathematical sciences. Photos below:

*Left to right: Bioengineering professor Todd Coleman; Albert P. Pisano, incoming dean of the Jacobs School of Engineering; UC San Diego Chancellor Pradeep K. Khosla; Linda S. Greene, UC San Diego Vice Chancellor for Equity, Diversity and Inclusion; and Geert Schmid-Schönbein, bioengineering professor and chair, attend the Conference for African Americans in the Mathematical Sciences at UC San Diego*

*Chancellor Khosla speaking at CAARMS kickoff event*
<table>
<thead>
<tr>
<th>Activity</th>
<th>UCSD-Howard Pathways Initiative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>La Jolla, CA</td>
</tr>
<tr>
<td>Date</td>
<td>June-Aug, 2013</td>
</tr>
<tr>
<td>Led by</td>
<td>UC Office of the President/Todd Coleman</td>
</tr>
<tr>
<td>Audience</td>
<td>Howard University research students</td>
</tr>
<tr>
<td>Participants</td>
<td>20</td>
</tr>
</tbody>
</table>

**Brief Description:** The Partnership plays a key role in building strong and lasting connections between Howard University and UC San Diego research students.

Students who are part of this program and get admitted to UCSD get a full fellowship for graduate students through the UC Office of the President. In summer 2013, I advised Daril Brown, a rising senior at Howard in Mechanical Engineering on a projet to use Epidermal Electronics for unvoiced speech recognition.
### Activity: High School Student Summer Internships

<table>
<thead>
<tr>
<th>Location</th>
<th>Cyber-Physical Systems Lab of P. R. Kumar at Texas A&amp;M University</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>Summer 2013</td>
</tr>
<tr>
<td>Led by</td>
<td>P. R. Kumar</td>
</tr>
<tr>
<td>Audience</td>
<td>There were two students: (i) Shreya Shankar, female (ii) Akshans Verma (male)</td>
</tr>
<tr>
<td>Participants</td>
<td>2</td>
</tr>
</tbody>
</table>

**Brief Description:** Students carried out projects in the laboratory.

### Activity: Unconventional WSDM: Working Group on Gender Diversity

<table>
<thead>
<tr>
<th>Location</th>
<th>6th ACM International Conference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>2012</td>
</tr>
<tr>
<td>Led by</td>
<td>J. Neville</td>
</tr>
<tr>
<td>Audience</td>
<td>Students/Post-docs/Faculty</td>
</tr>
<tr>
<td>Participants</td>
<td></td>
</tr>
</tbody>
</table>

**Brief Description:** J. Neville organized a special Special workshop at the 6th ACM International Conference on Web Search and Data Mining: *Unconventional WSDM: Working Group on Gender Diversity*

### Activity: Statistics women lunch and discussion, UC Berkeley

<table>
<thead>
<tr>
<th>Location</th>
<th>UC Berkeley</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>Dec 2012 / May 2013</td>
</tr>
<tr>
<td>Led by</td>
<td>B. Yu</td>
</tr>
<tr>
<td>Audience</td>
<td>Women faculty, postdocs and graduate students in statistics and biostatistics</td>
</tr>
<tr>
<td>Participants</td>
<td>20-40</td>
</tr>
</tbody>
</table>

**Brief Description:** This is a regular activity where women faculty, postdocs and students interact socially and network and discuss work and family related issues.
<table>
<thead>
<tr>
<th>Activity</th>
<th>Recruitment Activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>Continuous</td>
</tr>
<tr>
<td>Audience</td>
<td>Undergraduates; Graduate Students; Post-docs</td>
</tr>
<tr>
<td>Led By</td>
<td>CSol Faculty and Staff</td>
</tr>
<tr>
<td>Attendees</td>
<td>Pre-college Undergrad Graduate Post-doc Faculty Staff Other (see below)</td>
</tr>
</tbody>
</table>

**Brief Description:** Faculty and staff invested themselves in a number of recruitment visits to underrepresented audiences, as well as helping to establish agreements with several STEM partners.

- C. Aguilar recruited one undergraduate student from his Biol42000 “Eukaryotic cell biology” course, Andrea Jaramillo, to his research team. In addition, he is actively participating in the NSF-backed Louis Stokes Alliance for Minority Participation (LSAMP) program in campus to host students from underrepresented groups to have a first-hand experience in research. From these experience, Margaret Ramos, McKeith Pearsons II and Andrea Jaramillo decided to join his group on permanent basis. He participated in the Purdue Summer Research Opportunity Program (SROP) hosted by the Purdue graduate school. As a result of his involvement in this program, he hosted undergraduate students Fernando J. Bonilla Valentin, Alberto Perez Medina and Reynat Jimenez from the University of Puerto Rico, Jeanette Vargas from Chaminade University and Shaleeka Cornelius from Howard University. In addition, he has been regularly hosting two Hispanic high-school students: Agustina de la Fuente and Rodrigo Lugo and another undergraduate volunteer, Brendan Knapp.

- M. Ward assisted Edray Goins and Alejandra Alvarado with their application to potentially host the 2015 Infinite Possibilities Conference at Purdue University.

- M. Rwebangira took 4 Howard undergraduates to visit Carnegie Mellon for a graduate student recruitment visit and learn about the admissions process. Two of them ended up applying for PhD programs.

- T. Weissman was involved in the hiring of a new woman faculty in his department: Ayfer Ozgur, and serves as her faculty mentor.

- T. Weissman serves as the primary PhD advisor to Idoia Ochoa (female) and postdoctoral host to Zhiying Wang (female).

- T. Coleman: Is the faculty mentor for the UCSD chapter of the National Society of Black Engineers and holds monthly dinners with them

- T. Coleman: Is a faculty member of the IDEA center at UCSD.

- B. Gibson, Purdue University: Shared an exhibit booth with Purdue Graduate School at 2012 SACNAS conference for recruiting of underrepresented students to Center.
### Activity | Minority Mentoring
--- | ---
Time | continuous
Audience | undergraduates, graduate students, post-docs, and faculty
Led By | B. Gibson/D. Kumar/T. Coleman
Attendees | Minority #students Undergrad #female students Graduate # female students Post-doc Faculty Staff Other (see below)

### Brief Description:

- A. Goldsmith mentored a female who was sponsored on a CSOI project as a summer undergraduate researcher.
- A. Goldsmith participated in a panel on diversity and work/life balance issues at ISIT 2013.
- A. Goldsmith mentors a young black female faculty member at Stanford, and another young female (non-minority), both assistant professors in EE.
- C. Aguilar is on the Faculty Committee for Diversity (College of Science, Purdue University)
  - Chair of the Committee for Diversity of the Dept. of Biological Sciences.
  - Purdue Latino Faculty Committee
  - American Society for Cell Biology, Minorities Affair Committee
- J. Neville is mentoring Nikita Rau from the Bronx High School of Science on her science fair entry for the Intel Science Talent Search
- J. Neville has several URM grad students that she actively mentors. She currently has 2 female grad students, 2 grad students from South America, and 2 US citizen grad students. She graduated two female PhD students in 2012.
- M. Ward was a mentor for Tyler Sandmann (undergraduate Statistics student) in Purdue’s HORIZONS Student Support Program (for first-generation college students).
- M. Ward mentored Mayfawny Bergmann (undergraduate, female, US Citizen) on a research project throughout summer 2013. Him and the student are almost ready to submit a journal paper for publication.
- C. Liu mentored two undergraduate students and a Master’s student from minority groups: Jean-Claude Tounkara: M.S. student; and Oluwaseun Ademuwagun, David Harris.
- L. Si was the advisor of Mariheida Cordova, who is a female minority Ph.D. student in Purdue Computer Science.
- N. Lynch often meets with women students’ groups when traveling for lectures. For example, this year she gave a Distinguished Lecture at Carnegie-Mellon’s Computer Science department and spent several days meeting with CMU students, faculty, and postdocs. One of these meetings was with the women’s student group where she answered many questions related to being a woman faculty member, or woman research leader, in computer science.
- N. Lynch, in conjunction with her research associate, Dr. Peter Musial, finished up their work
with MIT’s Primes program in summer of 2012. This involved supervising two advanced high school students in a project to design and analyze algorithms for maintaining atomic memory in fault-prone distributed systems

- R. Rivest’s Ph.D. student (Emily Shen) who has been supported by the Center has successfully completed her doctoral program, and now has a position at Lincoln Laboratories in Lexington, Massachusetts.

- T. Coleman actively helped two undergraduate students (one African-American, one Hispanic) who do research in my group at UCSD and have outstanding GPAs win a prestigious Howard Hughes Medical Institute EXROP program fellowship. This allows them to do research over the summer with any HHMI scientist of their choosing and it makes them eligible for a prestigious graduate school fellowship that fully funds them throughout their graduate career. (Ismael Munoz, Bioengineering, Hispanic; Cheryl Fonteh, Bioengineering, Hispanic).
Diversity Metrics

In addition to the above activities, metrics are also shown through the diversity numbers for various groups within the Center. These are presented below. (Since the records for prior years were incomplete, we do not have a baseline to compare with previous years).

- Race Distribution by Center Classification
- Ethnicity by Center Classification
- Gender by Center Classification
2d. Plans for Changes in Diversity Activities

For annual reports: Describe your plans for programs, activities, or partnerships to enhance diversity for the next reporting period with attention to any major changes in direction or level of activity. Be sure to discuss how the planned activities will enable the Center to meet its diversity goals/objectives.

The Center will continue efforts with other STEM programs committed to promoting diversity. This will help with our recruitment metrics, and our objective to broaden participation of underrepresented students and post-docs outside the Center in CSoI activities. We will also strengthen and establish partnerships with the educational and human resources programs of other STCs in order to build upon their best practices, and “cross pollinate” research and outreach activities related to students and post-docs where appropriate. Finally, we will attempt to expand our recruiting and mentoring efforts to engage more CSoI faculty in these efforts, especially at the advanced academic levels within CSoI, such as graduate students and post-docs. The best method to achieve this is still being investigated through literature review and discussions with other STEM professionals.
VII. Management

Organization Strategy, Rationale, Metrics, Progress and Problems

1a. Center’s Organizational Strategy and Rationale

Describe the Center’s organizational strategy and its underlying rationale, if changed since the last reporting period. To assist in your description, attach the organization chart of the Center during the reporting period as Appendix B (if changed from last period). If there have been any changes in the Center’s organization or management since the last reporting period, discuss these changes and the reasons behind them.

Integrating diverse, geographically separated teams of researchers into an effective common unit continues to be the Center’s overall management focus.

As described in previous reports and in order to pursue the Center’s goals, the Center adopted a management structure (see Appendix B) that leverages extensive existing resources to integrate project sites and teams, and to articulate a clear shared vision of information science as a key enabling discipline. The management plan’s goal is to put forth procedures aimed at achieving research, development, engagement, education, and outreach goals. The management plan remains consistent with the plan from the last reporting period.

At the suggestion of last year’s NSF site team, the Center added an oversight committee that is chaired by Richard Buckius, Purdue’s Vice President for Research, and includes various administrative representatives from throughout the Center. The purpose of this group is to provide executive oversight to the Center’s management team and Executive Committee. The committee will accomplish this through a review of the Center’s annual report along with an informational interview with the Center Director. This will be done once a year in between the completion of the annual report and the site visit. This year, the committee will meet in November 2013.

There has been only one management personnel change. At the very end of the reporting period, the Center’s Diversity Director, Barbara Gibson left the Center to accept a position in an academic department at Purdue University. The Center immediately started a search to replace the Diversity Director and is currently interviewing candidates for the position. It is expected that the position will be filled by the time of the Center’s December 2013 site visit. In the meantime, the former Diversity Director has agreed to assist the Center through the end of the calendar year as this transition is made.
1b. Performance and Management Indicators and Metrics

Inform us of the performance and management indicators the Center has developed to assess its progress in organizational and management goals/objectives.

The following are metrics for CSoi management as developed for the strategic plan:

- CSoi management (Director, Executive Committee) will hold formal activities at least once per year to allow all Center participants feedback in Center decision making.
- Executive Committee will conduct annual evaluations of each investigator’s contributions.
- The Center Director and/or Managing Director will continue visiting partner institutions and meeting with participants in person.
- The Center will clearly articulate expectations for each participating investigator.
- The Center’s website will be a key resource, making accomplishments visible and serving as a platform for current and future work.
- The Center will announce timely press releases to the general public highlighting research accomplishments.
- The Center will foster a reputation for excellence in research and innovation.

1c. Progress in Meeting Indicators and Metrics

Describe how the Center is doing with respect to the indicators/metrics listed above. Include any data that have been collected on the indicators/metrics.

Participant Feedback

The Executive Committee meets routinely over the course of the year to discuss Center business and evaluate its progress. At these meetings, representatives from each of the partnering institutions have the opportunity to ask questions and communicate concerns regarding their respective participants. (Meeting notes in Appendix C) In addition, an annual “all hands” meeting is conducted coinciding with our annual site visit, and presents another opportunity for faculty and student representatives from all institutions to engage.

Annual Evaluations

During the Executive Committee meetings, participant contributions were discussed and evaluated. Feedback was noted and implemented. Again this year, the process included the formation of a stand-alone evaluation committee, chaired by P.R. Kumar, and consisting of research thrust leaders and the managing director. The committee evaluated progress and participation of all PI’s and accordingly made recommendations to the Executive Committee. The Director with the advice of the Executive Committee implemented these decisions.
Visiting Partner Institutions

Each partner institution has now received a personal visit from either the Center Director or the Managing Director. Routine visits will continue to be a part of the management plan. In addition the Center has acted to assure our presence on partner campuses: the 2012 summer school was held on Stanford’s campus; the 2014 summer school will be held at UC San Diego; our Bell Labs research and grand challenges workshops were held at Princeton University; Bryn Mawr College has held multiple events sponsored by the Center; UC Berkeley and Stanford held a day-long workshop on Berkeley’s campus that was sponsored by the Center called “New Directions in the Science of Information.” Similar events are planned for Texas A&M, MIT, and Howard in the coming spring.

Articulating Expectations

After the Executive Committee’s evaluations, expectations of each investigator are reviewed and articulated. Clear expectations are presented in the form of a document that has been shared with all investigators. A Grand Challenge document has been drafted and will be distributed among all PI’s. This document informs PI’s about the Center’s research priorities.

CSol Website

The CSol website continues to serve the Center at an ever expanding level. Impressive programmatic details of the website are shared principally in Section III Education. The website experienced continued growth in the number of unique visitors, and previous efforts to build web-based research collaboration space and tools were utilized by researchers for multidisciplinary and cross-institutional research projects across the Center.

The site offers a custom database of Center participants, news, events and other resources that can be linked and queried in many ways. These are tagged with keywords allowing for the discovery of new collaboration opportunities.

CSol Connect/Video Offerings

The goal to “create a virtual collaborative workspace via the Center’s website” was realized last year by integrating Adobe Connect. This is accessible through http://soihub.org/connect. This popular and easy-to-use system has enhanced virtually every aspect of Center communication and collaboration efforts.

Using this system, the Center can stream events live and/or record them for later viewing and link to them to the CSol website. In the last year alone, various Center video offerings were accessed over 13,000 times as compared to just over 3,000 views the year before. Further growth is expected in the coming years.

CSol Connect was used for CSol staff meetings, executive committee meetings, seminars, student brown bag seminars and research workshops. This system will actually be used in the upcoming site
visit to meet with administrative representatives from throughout the Center. More detailed website statistics are available in Section IV Knowledge Transfer.

The Center also continues to grow its online social media presence by utilizing web-based sites such as Facebook, YouTube, LinkedIn, Google Plus, and Twitter. In addition, the Center is using a customized Wiki as a development site for the CSol learning hub.

**Press Releases**

The Center’s easy-to-use, news and events website interface allows easy distribution of press releases about the Center. Such news is shared with other relevant publications and posted to multiple social media sites.

**Foster a Reputation for Excellence in Research and Innovation**

The Center continues to build a reputation for excellence in research as indicated by the large numbers of publications and conference proceedings, as well as the numbers of investigators who have been awarded various distinctions within their fields.

1d. Problems Encountered in Reaching Goals

*Discuss any problems (e.g., technical, personnel, communication) you may have encountered in realizing the Center’s organizational strategy or management goals/objectives in the reporting period as well as any problems anticipated in the next period. Include your plans for addressing any problems.*

None to report.

**Management and Communication Systems**

2. Management and Communication Systems Developed to Fully Integrate the STC

*Describe and discuss the management and communications systems being used to develop a fully integrated STC as well as any problems encountered in achieving this integration, if changed from the previous reporting period.*

As the management team and administrative responsibilities are distributed widely across the Center, this system helps integrate the various parts of the STC. The Executive Committee membership is representative of the participating partners that make up the Center. All six research thrust leaders come from different institutions. Two faculty investigators serve as associate directors and directly advise the Center director on issues pertaining to education and diversity. A third faculty associate director advises on knowledge transfer.
The Center strives to maintain a consistent and open line of communication with all partners and members. Center staff continually updates the website with news and details about all upcoming events, and sends routine email announcements and reminders about important news and happenings (see Appendix D). Further, the Center maintains current promotional and informational materials to share with partners and other interested parties and that highlight the Center and its mission. This includes a monthly “e-brief” emailed as well as a quarterly newsletter sent to the entire constituency.

Center efforts remain consistent with the overall plan. The team continues to work to improve its efforts to fully integrate Center activities and communication.

**Internal and External Advisors**

3. **Listing of Internal and External Advisors or Advisory Bodies**

*Provide a list of names and affiliations of the Center’s internal and external advisors or advisory bodies in the reporting period. Attach summary minutes of advisory committee meetings as Appendix C.*

### CSol External Advisory Committee Members

<table>
<thead>
<tr>
<th>Name</th>
<th>Affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Jerry Gibson (chair)</td>
<td>University of California-Santa Barbara</td>
</tr>
<tr>
<td>2 Tony Ephremides</td>
<td>University of Maryland</td>
</tr>
<tr>
<td>3 Javier Garcia Frias</td>
<td>University of Delaware</td>
</tr>
<tr>
<td>4 Mark Guzdial</td>
<td>Georgia Institute of Technology</td>
</tr>
<tr>
<td>5 Michael Honig</td>
<td>Northwestern University</td>
</tr>
<tr>
<td>6 Ness Schroff</td>
<td>The Ohio State University</td>
</tr>
<tr>
<td>7 Gadiel Seroussi</td>
<td>Consultant- Information Theory Research and Applications</td>
</tr>
<tr>
<td>8 Emina Soljanin</td>
<td>Bell Laboratories</td>
</tr>
</tbody>
</table>

### CSol Executive Committee Members

<table>
<thead>
<tr>
<th>Name</th>
<th>Affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Venkat Anantharam</td>
<td>University of California-Berkeley</td>
</tr>
<tr>
<td>2 Ananth Grama</td>
<td>Purdue University</td>
</tr>
<tr>
<td>3 P.R. Kumar</td>
<td>University of Illinois at Urbana-Champaign</td>
</tr>
<tr>
<td>4 Andrea Goldsmith</td>
<td>Stanford University</td>
</tr>
<tr>
<td>No.</td>
<td>Name</td>
</tr>
<tr>
<td>-----</td>
<td>----------------------</td>
</tr>
<tr>
<td>5</td>
<td>Deepak Kumar</td>
</tr>
<tr>
<td>6</td>
<td>Shankar Subramaniam</td>
</tr>
<tr>
<td>7</td>
<td>Peter Shor</td>
</tr>
<tr>
<td>8</td>
<td>Sergio Verdu</td>
</tr>
</tbody>
</table>

**Ex Officio Members:** Wojtek Szpankowski (Center Director) and Bob Brown (Center Managing Director)

## Changes to Strategic Plan

4. **Describe Changes to Center’s Strategic Plan**

*Describe and discuss any changes to the Center’s strategic plan since its last submission.*

No major changes were made to the strategic plan. As would be expected, minor adjustments and additions have been made throughout.
VIII. CENTER-WIDE OUTPUTS AND ISSUES

Center Publications

1a. List of Center Publications

List all Center publications in the reporting period using a standard citation format. Please distinguish among the following publication types:
- peer reviewed journal articles
- submitted journal articles
- books and book chapters
- submitted books and book chapters
- technical reports
- submitted technical reports
- peer reviewed conference papers
- submitted conference papers

Peer Reviewed Publications


64. G. Kollias, E. Gallopoulos, and A. Grama. "Surfing the Network for Ranking by Multidamping." *IEEE Transactions on Knowledge and Data Engineering*, 2013, accepted.


**Articles Submitted for Publication**


1b. List of Conference Presentations

List all conference presentations in the reporting period using a standard citation format.


2012.


Workshop on Cloud Intelligence (Cloud-I, a VLDB 2012 Workshop), Istanbul, Turkey, August 2012.


H. Li, C. Liu, X. Liu, and W. Southerland. “Collaboration Visualization on Large Dataset for


Books and Book Chapters


N. Lynch, did publish a Second Edition of her book, The Theory of Timed I/O Automata. Timed I/O Automata serve as the foundation for all our work on distributed algorithms, so this could be considered related.


Books and Book Chapters Submitted for Publication


Technical Reports and Other Non-peer Reviewed Publications


**Technical Reports and Other Non-peer Reviewed Submitted for Publication**


### Other Outputs and Activities

#### 1c. Dissemination Activities

*Briefly describe any other dissemination activities not included elsewhere in the report.*

None other to report

#### 2. Awards and Honors

*List all awards and other honors with names of those honored and source in the reporting period. Please classify the award type indicating whether the award or honor is scientific, education-related, industry-related, a fellowship, or other.*

<table>
<thead>
<tr>
<th>Recipient</th>
<th>Reason for Award</th>
<th>Award Name and Sponsor</th>
<th>Date</th>
<th>Award Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pulkit Grover</td>
<td>Research</td>
<td>Leonard G. Abraham Award from IEEE Communication Society</td>
<td>2013</td>
<td>Research</td>
</tr>
<tr>
<td>Yury Polyanskiy</td>
<td>Research</td>
<td>NSF Career grant</td>
<td>2013</td>
<td>Research</td>
</tr>
<tr>
<td>Doraiswami Ramkrishna</td>
<td>Research</td>
<td>Sigma Xi Faculty Research Awards</td>
<td></td>
<td>Research</td>
</tr>
<tr>
<td>Andrea Goldsmith</td>
<td>Research/Industry</td>
<td>CTC Technical Achievement Award</td>
<td>2013</td>
<td>Research</td>
</tr>
<tr>
<td>Scott Aaronson</td>
<td>Research/Scientific</td>
<td>Alan T. Waterman Award</td>
<td>2012</td>
<td>Research</td>
</tr>
<tr>
<td>Claudio R. Aguilar</td>
<td>Research/Scientific</td>
<td>Lowe Syndrome Trust (Toward a greater understanding of the cellular biology underlying Lowe syndrome through integration with Cerebro-Renal diseases.)</td>
<td>2011-2013</td>
<td>Research</td>
</tr>
<tr>
<td>Andrew Drucker</td>
<td>Research</td>
<td>Best student paper “High Confidence Predictions under Adversarial Uncertainty” at ITCS</td>
<td>2012</td>
<td>Research</td>
</tr>
<tr>
<td>Andrew Drucker</td>
<td>Research</td>
<td>Ronald V. Book Award for best student paper at CCC 2012. “Limitations”</td>
<td>2012</td>
<td>Research</td>
</tr>
<tr>
<td>Name</td>
<td>Type</td>
<td>Award Description</td>
<td>Year</td>
<td>Category</td>
</tr>
<tr>
<td>---------------------</td>
<td>----------------</td>
<td>-----------------------------------------------------------------------------------</td>
<td>-------------</td>
<td>------------</td>
</tr>
<tr>
<td>Amin Emad</td>
<td>Research</td>
<td>Sundaram Seshu International Student Fellowship</td>
<td>2012</td>
<td>Research</td>
</tr>
<tr>
<td>Jennifer Neville</td>
<td>Research</td>
<td>NSF Career Award</td>
<td>2012-2016</td>
<td>Research</td>
</tr>
<tr>
<td>Olgica Milenkovic</td>
<td>Research/Scientific</td>
<td>Center for Advanced Studies Associate Award</td>
<td>2012</td>
<td></td>
</tr>
<tr>
<td>Sergio Verdu</td>
<td>Education</td>
<td>Tenth Viterbi Lecturer, Ming Hsieh Electrical Engineering Department, Viterbi School of Engineering, University of Southern California, Los Angeles</td>
<td>2012</td>
<td>Education</td>
</tr>
<tr>
<td>Tsachy Weissman</td>
<td>Research/Industry</td>
<td>Hewlett Packard Labs Innovation Research Award</td>
<td>2012-2015</td>
<td>Research</td>
</tr>
<tr>
<td>Tsachy Weissman</td>
<td>Research/Industry</td>
<td>Google Research Award</td>
<td>2011-2012</td>
<td>Research</td>
</tr>
<tr>
<td>Kartik Venkat</td>
<td>Best student paper award</td>
<td>ISIT Best student paper award “Pointwise Relations between Information and Estimation in Gaussian Noise”</td>
<td>2012</td>
<td>Research</td>
</tr>
<tr>
<td>Shang Shang</td>
<td>Best paper award</td>
<td>IEEE International Workshop. Hot Topics in Peer-to-Peer Computing and Online Social Networking</td>
<td>2012</td>
<td>Research</td>
</tr>
<tr>
<td>Dan Zhang</td>
<td>Research</td>
<td>IBM Ph.D. Fellowship award</td>
<td>2012</td>
<td>Research</td>
</tr>
<tr>
<td>Dan Zhang</td>
<td>Research</td>
<td>Facebook Ph.D. fellowship finalist</td>
<td>2012</td>
<td>Research</td>
</tr>
<tr>
<td>Arpita Sen</td>
<td>Research</td>
<td>American Society for Cell Biology Travel award</td>
<td>2012</td>
<td>Research</td>
</tr>
<tr>
<td>Ming Yang</td>
<td>Best Theory Job Market paper</td>
<td>Finance Theory Group Award “Optimality of Securitized Debt with Engogenous and Flexible”</td>
<td>2012</td>
<td>Research</td>
</tr>
<tr>
<td>Name</td>
<td>Domain</td>
<td>Award/Grant</td>
<td>Year</td>
<td>Type</td>
</tr>
<tr>
<td>-----------------------</td>
<td>-----------------</td>
<td>------------------------------------------------------</td>
<td>------</td>
<td>-------------</td>
</tr>
<tr>
<td>Pulkit Grover</td>
<td></td>
<td>Eli Jury Award UC Berkeley, EE, for dissertation work</td>
<td>2012</td>
<td>Research</td>
</tr>
<tr>
<td>Nancy Lynch</td>
<td></td>
<td>Career Athena Award</td>
<td>2012</td>
<td>Research</td>
</tr>
<tr>
<td>P. R. Kumar</td>
<td></td>
<td>Research ACM SIGMOBILE Outstanding Contribution Award</td>
<td>2012</td>
<td>Research</td>
</tr>
<tr>
<td>William Bialek</td>
<td></td>
<td>Career National Academy of Sciences</td>
<td>2012</td>
<td>Research/</td>
</tr>
<tr>
<td>Supriyo Datta</td>
<td></td>
<td>Career National Academy of Engineering</td>
<td>2012</td>
<td>Foundation</td>
</tr>
<tr>
<td>Sudeep Kamath</td>
<td>Research</td>
<td>Eliahu Jury Award, by EECS Department, UC Berkeley</td>
<td>2013</td>
<td>Research</td>
</tr>
<tr>
<td>Sean Kearney</td>
<td>Research</td>
<td>NSF Graduate Research Fellowship</td>
<td>2013</td>
<td>Fellowship</td>
</tr>
<tr>
<td>Alexandros Manolakos</td>
<td>Research</td>
<td>A.G. Leventis Foundation Scholarship, Greece</td>
<td>2013</td>
<td>Education</td>
</tr>
<tr>
<td>Naresh Rapolu</td>
<td>Research</td>
<td>Best student paper award at 4th IEEE International Conference on Cloud Computing (CLOUD) 2013</td>
<td>2013</td>
<td>Research</td>
</tr>
<tr>
<td>Arpita Sen</td>
<td>Research/Scholarship</td>
<td>Bilsland dissertation fellowship, Purdue University</td>
<td>2013</td>
<td>Fellowship</td>
</tr>
<tr>
<td>Mikhail Tikhonov</td>
<td>Scientific</td>
<td>Kusaka Memorial Prize in Physics, Princeton University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Venkat Kartik</td>
<td>Research</td>
<td>ISIT Student Best Paper Award at International Symposium on Information Theory, Istanbul</td>
<td>2013</td>
<td>Research</td>
</tr>
</tbody>
</table>
### Student Graduation Details and Placement

List any undergraduate, M.S. and Ph.D. students who graduated during the reporting period. Include their current placement. Include the number of years taken since entering graduate school to complete the Ph.D. List postdoctoral associates who left the STC during the reporting period, and include their current placement.

<table>
<thead>
<tr>
<th>Student Name</th>
<th>Degree(s)</th>
<th>Yrs</th>
<th>Placement</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oluwaseun Ademuwagun</td>
<td>BS</td>
<td></td>
<td>Software Engineer</td>
<td>Amazon</td>
</tr>
<tr>
<td>George Alexandrov</td>
<td>BS</td>
<td></td>
<td>Graduate Student</td>
<td>UC Berkeley</td>
</tr>
<tr>
<td>Yuval Benjamini</td>
<td>PhD</td>
<td></td>
<td>Stein Fellow</td>
<td>Statistics Dept., Stanford Univ.</td>
</tr>
<tr>
<td>J. Grayson Camp</td>
<td>Post-Doc</td>
<td></td>
<td>Postdoc</td>
<td>Stanford</td>
</tr>
<tr>
<td>Luis Campos</td>
<td>Masters</td>
<td></td>
<td>Research Statistician</td>
<td>UC Davis Health System</td>
</tr>
<tr>
<td>Marianne Catanho</td>
<td>Masters</td>
<td></td>
<td>Graduate Research Assistant</td>
<td>UC San Diego</td>
</tr>
<tr>
<td>Jenny Chen</td>
<td>BS</td>
<td></td>
<td>AmeiCorps VISTA</td>
<td></td>
</tr>
<tr>
<td>Tai-Jung Choi</td>
<td>PhD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caitlyn Clabaugh</td>
<td>BS</td>
<td></td>
<td></td>
<td>NetApp, Sunnyvale, CA</td>
</tr>
<tr>
<td>Wei Dai</td>
<td>Post-Doc</td>
<td></td>
<td>Lecturer</td>
<td>Imperial College London</td>
</tr>
<tr>
<td>Ronald Doku</td>
<td>BS</td>
<td></td>
<td>Software Engineer</td>
<td>EMC</td>
</tr>
<tr>
<td>Julien Dubuis</td>
<td>PhD</td>
<td></td>
<td></td>
<td>Boston Consulting Group</td>
</tr>
<tr>
<td>Alexandra Funk</td>
<td>BS</td>
<td></td>
<td>Software Developer</td>
<td>Vanguard</td>
</tr>
<tr>
<td>Karthik Ganesan</td>
<td>BS</td>
<td></td>
<td>Graduate Student</td>
<td>Stanford</td>
</tr>
<tr>
<td>David Harris</td>
<td>BS</td>
<td></td>
<td>Graduate Student</td>
<td>University of Maryland, Baltimore</td>
</tr>
<tr>
<td>Farzad Hassanzadeh</td>
<td>PhD</td>
<td></td>
<td>Postdoctoral Scholar</td>
<td>CalTech</td>
</tr>
<tr>
<td>Sudeep Kamath</td>
<td>PhD</td>
<td>5</td>
<td>Postdoctoral Scholar</td>
<td>UC San Diego</td>
</tr>
<tr>
<td>Victoria Kostina</td>
<td>PhD</td>
<td></td>
<td>Postdoctoral Scholar</td>
<td>Princeton University</td>
</tr>
<tr>
<td>Rachel Lee</td>
<td>BS</td>
<td></td>
<td>Senior Consultant</td>
<td>Dominion Digital</td>
</tr>
<tr>
<td>Reza Mirghaderi</td>
<td>PhD</td>
<td></td>
<td></td>
<td>Google</td>
</tr>
<tr>
<td>Marissa Mocenigo</td>
<td>BS</td>
<td></td>
<td>Software Engineer</td>
<td>Rethink Robots</td>
</tr>
<tr>
<td>Meagan Neal</td>
<td>BS</td>
<td></td>
<td></td>
<td>Vanguard Inc., Philadelphia, PA</td>
</tr>
<tr>
<td>Yair Noam</td>
<td>Post-Doc</td>
<td></td>
<td>Faculty</td>
<td>Ben Gurion University</td>
</tr>
</tbody>
</table>
### General Outputs of Knowledge Transfer

List, to the extent known, the general outputs of knowledge transfer activities since the last reporting period. Include:

<table>
<thead>
<tr>
<th>Patent Name and Inventors/Authors</th>
<th>Number</th>
<th>Application Date</th>
<th>Receipt Date (leave empty if pending)</th>
</tr>
</thead>
</table>
the number of participants (total number of people who utilize center facilities; not just persons directly supported by NSF). Please EXCLUDE affiliates (click for definition) this value should match the total number of participants listed in Section VIII, Item 5 (above)

8. Media Publicity

Describe any media publicity the Center received in the reporting period. Provide in Appendix D any appropriate media materials that can be used to disseminate information on Center accomplishments and activities to the public.

Please see CSoI related media materials in Appendix D

In Addition

Jack Gallant (UC Berkeley)

Gallant’s Neuron Paper (Huth et al., 2012) was covered in several dozen newspapers and science blogs around the world. The online brain browser that he made public at the time of publication received over one hundred thousand hits within the first few weeks after publication.

Nature Neuroscience paper (Cukur et al., 2013) was covered in several dozen newspapers and science blogs around the world.

Brain mapping: From the basics to science fiction. Quest, NPR, April 19, 2013


Zzzzs please. Big Picture Science, NPR. May 27, 2013

Marcus, G. The problem with the neuroscience backlash. The New Yorker. June 19, 2013

Gallant laboratory was featured on Through the Wormhole with Morgan Freeman (Season 4 Episode 6). July 2013.

Alan Qi (Purdue)

“Making Stem Cells More Available and Useful” in Drug Discovery & Development, August 2012 Issue
Bin Yu (UC Berkeley)
An interview published in the Bulletin of ICSA (International Chinese Statistical Association), July 2013 and it is entitled “A conversation with Professor Bin Yu”.

Andrea Goldsmith (Stanford)
Interviewed by the local Bay Area news regarding robustness of cellular phones to natural disasters

Nancy Lynch (MIT)
Profile in CACM after winning the Athena Award

Todd Coleman (U.C. San Diego)
TedX San Diego 2013, “Towards Inconspicuous and Instrumental Brain-Computer Interfaces,”
MSN News – “Temporary Tattoos that can measure brain signals wirelessly”
Time – “Finally, Tattoos That Let You Control Objects with Your Mind”
New York Observer- “These Temporary Tats Could Let You Move Objects with Your Mind”
Discovery News – “Electronic Telekinesis from Temporary Tattoo”
Smithsonian – “These Temporary Tattoos Could Fly Drones”
NBC News – “How Neuroscientists are hacking into Brain Waves to Open New Frontiers”
AAAS News: “Machines That Communicate With the Body”
Irish Times: “Brain Research Faces Ethical Issues”
## IX. INDIRECT/OTHER IMPACTS

### Other Outputs and Activities

1. **International Activities**

   Please describe any international activities in which the Center has engaged. If they are described elsewhere in the report, highlight them here without going into great detail.

   Additional information about international partnerships can be found in section V. External Partnerships.

<table>
<thead>
<tr>
<th>Faculty</th>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mikhail Atallah</td>
<td>Ongoing collaboration with colleagues in Qatar University and the University of Paris 5 that involves frequent mutual visits. This resulted in 1 published paper for each, during the reporting period. Prof. Qutaibah Malluhi, Qatar University.</td>
</tr>
</tbody>
</table>
What can we decode from the human brain? (2013). *BrainGain Breinfest*, Donders Institute, Nijmegen, Netherlands.  
Semantic spaces for perception and language, and their modulation by attention (2013). *Faculty of Social Science*, Donders Institute, Nijmegen, Netherlands.  
Mapping perceptual and cognitive representations across the human brain (2013). *Cognitive Science Center Amsterdam (CSCA) Lecture Series*, University of Amsterdam, Amsterdam, Netherlands.  
Reverse engineering the human brain (2013). *Department of Informatics*, University of Amsterdam, Amsterdam Netherlands. |
<table>
<thead>
<tr>
<th>Name</th>
<th>Activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andrea Goldsmith</td>
<td>Goldsmith spent 10 days in China giving plenary talks at two conferences and visiting the university at Funming. She also gave a plenary talk at VTC in Dresden with many European participants.</td>
</tr>
<tr>
<td>Ananth Grama</td>
<td>Initiated collaboration with the Indian Institute of Science.</td>
</tr>
<tr>
<td></td>
<td>Initiated collaboration between Center and the Government of Colombia.</td>
</tr>
<tr>
<td></td>
<td>Working on setting up a Parallel Computing course online (lectures recorded, final course being packaged).</td>
</tr>
<tr>
<td>Sanjeev Kulkarni</td>
<td>Mete Ozay, visiting Ph.D. from Middle East Technical University.</td>
</tr>
<tr>
<td></td>
<td>Research grant and collaboration with Deutsche Telekom T-Lab</td>
</tr>
<tr>
<td>Deepak Kumar</td>
<td>Deepak Kumar, Dianna Xu, and Paul Ruvolo attended the Data Visualization Workshop which included a master class where attendees learned how to acquire data from external streams and visualize it in a relevant way. The workshop, organized by Experimental media Research Group of the St. Lucas University College or Art &amp; Design (Antwerp, Belgium).</td>
</tr>
<tr>
<td></td>
<td>Visited Prof. Walter Daelemans Research Lab (Computational Linguistics &amp; Psycholinguistics – CLiPS) at University of Antwerp (Belgium). February 4-8, 2013. Established collaboration with these folks. Later, in April 2013 Tom De Smedt from St. Lucas was invited to lead a hands-on workshop at Bryn Mawr College (see above/below).</td>
</tr>
<tr>
<td>PR Kumar</td>
<td>Keynote Speaker, IEEE Wireless Communications and Networking Conference (WCNC 2013), Shanghai, China, April 7–10, 2013.</td>
</tr>
<tr>
<td></td>
<td>Keynote Speaker, 7th International Conference on Wireless Algorithms, Systems, and Applications (WASA 2012), Yellow Mountains, China, August 8–10, 2012.</td>
</tr>
<tr>
<td></td>
<td>Member of the Scientific Committee of Laboratory of Information, Networking and Communication Sciences, Paris, France</td>
</tr>
</tbody>
</table>
| Chunmei Liu | “Prediction of Protein-Protein Docking Sites Based on a Cloud-Computing Pipeline”, 2012 The 1st International Conference on Information and Intelligent Computing (ICIIC 2012), December 2012, Chengdu, China.  
| Nancy Lynch | We have an extensive list of international collaborators, as you can see from our publication list.  
Members of our group travel and attend many international conferences; these also appear in our publication list. |
| Doraiswami Ramkrishna | Manish Mehta (Undergraduate student, Indian Institute of Technology, Kanpur, India)  
Tuhina Ghanty (Undergraduate student, Indian Institute of Technology, Mumbai, India) |
| Ronald Rivest | Madars Virza has been working closely with the team of Professor Eli Ben-Sasson from the Technion (Haifa, Israel).  
Professor Eli Ben-Sasson, an expert in proof complexity from the Department of Computer Science at the Technion in Haifa, Israel, has been visiting our laboratory (CSAIL) during portions of the past academic year. One of the primary purposes of this visit was to work with MIT colleagues (specifically, Madars Virza) on aspects of proving the correctness of computations. (This research is described in more detail above.) |
| Luo Si | Luo Si: Conference PC Co-Chair, The 9th International Conference on Semantics, Knowledge & Grids, Beijing, China, 2013. |
| Wojtek Szpankowski | LINCS, Paris, France. Also hosted LINCS faculty member at CSol Technology University of Gdansk, Poland  
ETH, Zurich, Switzerland |
<table>
<thead>
<tr>
<th>Name</th>
<th>Information</th>
</tr>
</thead>
</table>
| Sergio Verdu       | Advisory Board, Foundations and Methodologies for Future Communication and Sensor Networks - COMONSENS, Centro de Estudios e Investigaciones Técnicas de Guipuzcoa, San Sebastiàn, Euzkadi, Spain  
Dr. Roy Timo, University of South Australia (visiting research scholar at Princeton)  
Javier Payan Somet, Universidad de Sevilla (visiting research scholar at Princeton)  
Shlomo Shamai, Technion, Israel (joint Grant)  
Ioannis Kontoyiannis, Athens University of Economics and Business, Greece (joint research) |
| Bin Yu             | With a guest appointment at Peking University, I have been advising a PhD student Hanzhong Liu at Peking University, who is now a visiting student at Berkeley.  
In brief, I visited and gave talks at Peking University, Nanjing University in China, ENSAE in Paris, Cambridge University and Oxford University and UCL in England, Univ of Padova in Italy and University of Oslo in Norway. I met and talked to students one-on-one at all the above universities except at Oxford, UCL and Padova.  
Joan Bruna (Aug, 2013), Postdoc from Couran Institute: one week research visit. Peter Buhlmann (Professor, Math Department ETH), seminar in Berkeley, Sept. 2013. |
| Venkat Anantharam  | Visited CSol Partner LINCS during Summer 2013 |

2. Other outputs, Impacts, or Influences

2. Please use this space to describe other outputs, impacts, or influences related to the Center’s progress and achievement during the current reporting period that may not have been captured in another section of the report. (optional)

No additional information to report.
X. Budget
Appendices

Appendix 1
New Faculty Member Biography

Appendix 2
Organizational Chart

Appendix 3
External Advisory Committee and Executive Committee Notes

Appendix 4
Media Materials
Yuriy Baryshnikov - UIUC

(a) Professional Preparation

MIT, Cambridge, MA, USA  Applied Math & Physics M.Sc., 1982

(b) Appointments

2011-present  University of Illinois at Urbana-Champaign, Professor of Mathematics and Electrical Engineering
2003-2011  Columbia University, Adjunct Professor in Electrical Engineering
2001-2010  Bell Labs, Alcatel-Lucent, Department Head (before 2006, Member of Technical Staff), Industrial Mathematics and Operations Research
2000-2001  University of Versailles, Professor in the Department of Mathematics
1998-2000  Technical University of Eindhoven, Researcher at EURANDOM
1995-1997  University of Hull, Kingston-upon-Hull, Lecturer in the School of Mathematics
1993-1995  Deutsche Forschungsgemeinschaft, Habilitation Fellow
1992-1993  University Osnabrück, Wissenschaftlicher Mitarbeiter, Department of Mathematics
1991-1992  University Osnabrück, Alexander-von-Humboldt Research Fellow at the Department of Mathematics
1986-1991  Institute for Control Sciences, Moscow, Research Fellow

(c) Recent publications


Maxim Raginsky - UIUC

(a) Professional Preparation

Northwestern University, Evanston, IL, USA  Electrical Engineering  Ph.D., 2002
Northwestern University, Evanston, IL, USA  Electrical Engineering  M.S., 2000
Northwestern University, Evanston, IL, USA  Electrical Engineering  B.S., 2000

(b) Appointments

2012-present  University of Illinois, Urbana-Champaign, Assistant Professor, Department of Electrical and Computer Engineering, Coordinated Science Laboratory
2010-2012  Duke University, Durham. Assistant Research Professor, Department of Electrical and Computer Engineering
2007-2010  Duke University, Durham. Research Scientist, Network and Imaging Science Laboratory, Department of Electrical and Computer Engineering
2004-2007  University of Illinois, Urbana-Champaign. Beckman Foundation Postdoctoral Fellow, Beckman Institute for advanced Science and Technology
2002-2004  Northwestern University, Center for Photonic Communication and Computing, Postdoctoral Researcher
2000-2002  Northwestern University, Graduate Research Assistant, Department of Electrical and Computer Engineering

(c) Recent publications


**d) Graduate and postdoctoral advisors**

PhD thesis advisors: P. Guan, X. Guang A. Xu, Northwestern University

MS thesis advisors: A. Kalieva, Northwestern University
Tomasz Imielinski - Rutgers

(a) Professional Preparation
Polish Academy of Science, Warsaw, Poland Computer Science Ph.D., 1982
Politechnika Gdanska Electrical Engineering M.E., 1977

(b) Appointments
2010-present Member of the Board of Directors, Connotate Technologies
1996-present Professor. Department of Computer Science, Rutgers University, New Brunswick, NJ
2000-present Founder, Connotate Technologies, New Brunswick, NJ
2010-2012 Vice President, Pronto, New York, NY (on leave from Rutgers)
2009-2010 Chief Scientist, Ask.com, Edison, NJ (on leave from Rutgers)
2004-2009 Vice President/Senior Vice President, Ask.com, Edison, NJ (on leave from Rutgers)
2000-2003 Director of the Division of Computer and Information Sciences, Rutgers University, New Brunswick, NJ
1996-2003 Chair, Department of Computer Science, Rutgers University, New Brunswick, NJ
1990-1991 Visiting Scientist, IBM Almaden Research Center, San Jose, CA
1987-1996 Associate Professor, Department of Computer Science, Rutgers University, New Brunswick, NJ
1983-1987 Assistant Professor, Department of Computer Science, Rutgers University, New Brunswick, NJ
1982-1983 Assistant Professor, Department of Computer Science, McGill University, Montreal, Canada

(c-i) Related publications
1. R. Agrawal, T. Imielinski, A. Swami, “Mining associations rules between sets of items in large databases”, ACM SIGMOD, May 1993, pp. 207-217 (winner of 2003 Test of Time SIGMOD award, one of the most cited cs papers with 13,000 citations)

(c-ii) Other publications


Appendix C

Provide a list of names and affiliations of the Center’s internal and external advisors or advisory bodies in the reporting period. Attach summary minutes of advisory committee meetings as Appendix C.

CSOl Executive Committee Members:  PR Kumar (Chair) (Texas A&M), Ananth Grama (Purdue), Deepak Kumar (Bryn Mawr), Andrea Goldsmith (Stanford), Venkat Anantharam (Berkeley), Shankar Subramaniam (UC San Diego), Peter Shor (MIT), Sergio Verdu (Princeton) (Ad hoc members: Wojtek Szpankowski, Bob Brown)

CSOl External Advisory Board Members:  Tony Ephremides, Javier Garcia-Frias, Jerry Gibson (Chair), Michael Honig, and Emina Soljanin

SOI STC Executive Committee Meeting by Conference Call
Wednesday, October 10, 2012

Attendees:
Wojtek Szpankowski
Bob Brown, Managing Director
PR Kumar, Chair of EC
Ananth Grama
Deepak Kumar
Andrea Goldsmith
Venkat Anantharam
Peter Shor
Sergio Verdu
Kiya Smith

Notes:

Managing Director Remarks
- Bob reminded everyone that we have a website to share updated items for the executive committee.
- We have a new international initiative in Paris with LINCS. We will be collaborating with research and faculty/student exchanges among other things in the future. There is an opportunity to host a couple of students if anyone is interested. This is a nice step toward more international collaboration.
• We would also like to promote additional lectures/seminars within Center institutions. The Center will offer travel grants to each institution interested in inviting speakers to their respective institutions. Reimbursements will be made through Purdue. Brent Ladd will be sending out a formal announcement soon.
• The NSF site visit is approaching soon. Students/post docs will be arriving a day early for a research workshop. Bob will be sending email later this week with preliminary agenda and details.
• Venkat suggested international collaborations with Hong Kong and India.

Chair’s Remarks
P.R. Kumar shared the following. The evaluation committee has reviewed reports and given feedback to Wojtek. There were three individuals who were discussed questions were raised regarding their lack of published papers and center participation. Also, questions were raised about whether some of their work properly aligned with the goals of the Center. The committee has made the recommendation not to provide additional funding for these individuals. They will remain a part of the Center and will be reassessed in the future.

Director’s Remarks
Wojtek led discussion with the committee about the evaluation committees report and then gave an update on where he thought we were as a Center, etc:

Evaluation Committee

We have 3 cases where individuals did not submit evaluations in the last year and as you’ve heard from the evaluation committee, additional concerns have been raised about each individual. Extended conversation commenced between members of the committee about this.

The decision was made to not allocate funding for next year for these three individuals. They will continue with the Center, which leaves a door open for future funding and additional collaboration.

This conversation is confidential and should remain with the Executive Committee.

State of the Center

We have three objectives for the Center, Research, Knowledge Transfer, and Education/Diversity.
Knowledge Transfer – we have lots of outlets for knowledge transfer and while many of our faculty are engaged with industry and are sharing Center sponsored/related research, we have not yet gotten direct support for the Center. We expect this to take some time and will concentrate on a broad outlook for knowledge transfer.

Education/Diversity – 93% of students agree that the Center is doing a great job for them. The Center has made a difference in student life and given them excellent opportunities to collaborate in a multi-discipline, cross-institutional setting. The student workshop had 23 participants and was very successful. Summer School 2012 was successful, and looking forward to 2013. Brown Bag seminars are given by students on a monthly basis and are available on in real-time on the web.

Barbara Gibson has done a great job of defining out diversity initiatives and has an excellent plan going forward. She has engaged our partners with her channels program and assures our presence at multiple conferences etc. throughout the country.

Mark Ward taught our newly developed Science of Information course at Purdue last fall, and it is currently being taught at Bryn Mawr and Howard. Feedback will be used to further develop the class and other institutions will begin teaching it in the future – we will expand the class to other institutions outside the Center.

Research – we need to continue to focus our efforts as a Center (particularly as it relates to grand challenge questions and other larger issues within the Science of Information – we have examples (several were mentioned) but we will continue to look for more. We are making progress in the right direction. Our portfolio is looking much better in Biology than last year. We will work to enhance and broaden our agenda in knowledge extraction (Big Data) in the coming year and are planning a workshop for that exact topic.

Participation of partners – Participation from partners has overall been good – we are looking for MIT and Berkeley to be at full capacity and increase their involvement in all Center related activity.
SOI STC Executive Committee Meeting by Conference Call
Thursday, December 20, 2012

Attendees:
Wojtek Szpankowski
Bob Brown, Managing Director
PR Kumar, Chair
Deepak Kumar
Andrea goldsmith
Bin Yu
Peter Shor
Shankar Subramaniam
Brent Ladd

Notes:

Minutes from the previous meeting were approved.

Managing Director Remarks

- Bob thanked everyone that was able to attend for the good outcomes of the NSF site visit. Things went very smoothly.
- The NSF team recommends creating an industry advisory board. Andrea suggests coordinating an industry workshop in the bay area to help make connections. Bob asked all for input. Bob suggested an audit of all connections in industry that we already have. Bob seconded Andrea’s suggestion of combining this with another activity, such as external advisory board. She also suggested that the key is holding the meetings where the companies are located. Bob acknowledged that we had looked at starting an affiliates program similar to other STC’s (TRUST Center at Berkeley for one) but the idea has not been popular. This problem is pervasive at most STC’s. Bob feels the key is getting the right person at the right company. He will look at what we have to date and work with Ananth and Andrea, as well as get assistance from Purdue’s VPR office to proceed. Bin mentioned that the Center can be the backbone of the new tools that others use to solve “the big data: problems.
- NSF review indicated that partner “administrators” were not involved in the visit and were not present – even electronically. Bob will poll partner PI’s for who should be involved. No expectation was made in earlier site visits. Question was asked about the appropriate level – Dept head? Dean? Provost? Dean or
Associate Dean at the college level might be appropriate. Also Dept Head. This is also tied to Center presence at partner locations – there should be better Center presence at each institution. Wojtek reiterated that everyone should send Bob names of admins at their locations.

- Workshops – ITA workshop in February, San Diego. Can Rui and Todd organize a CSoI workshop. Shankar indicated this could be done.
- Summer School – four speakers confirmed. We are happy with this. Still need a Padovani lecturer – Sanjeev Kulkarni has been nominated. June 4-7, 2013. Any other recommendations?

**Director’s Remarks**

- Wojtek announced that a Big Data workshop is to be planned for 2-3 days and will be held around March 18. It will consist of small working groups with one speaker in the morning followed by afternoon discussions and some 20 minute presentations by participants. He suggested certain people to attend – also machine learning and statistics interface is important. Need to make sure the people who attend can really help us with the vision the Center has for this, can speak with us in a way that is accessible, and can understand how the Center can be a contributor in this area. Shankar – NIH initiative is an important opportunity – an important connection. Kumar suggested giving Wojtek suggested names by the next Friday.

- NSF review response: are we still focused on Shannon 2.0? We had a good response, but should think about it. Do we need to update our vision for the Center? Need feedback. It was pointed out that they would like to see more PI’s at the review. It will always be during the first week of December – please mark your calendars. More visible activity is needed at each partner site. We will have a speaker that the Center will fund at each site, and engage in a day of discussion, talks, poster session, etc. Anyway, such activity should extend beyond Purdue. We will have more definitive links with each partner on our web.

- Wojtek stressed that partners need to conduct small regional meetings, bring students, have talks, work as a group.

- Research side of the report: common thread is needed for the projects. The Center post docs are helping with this, but how can we make this more
coherent? For instance, the team asked “how is quantum connected?” Peter explained and there are connections there.

- Does the Center have mechanisms for feedback from faculty and students? Newsletter at least 2x a year. Email news once a month on most relevant events coming up. How can we better communicate, especially with faculty? We are open to constant improvement. Please share all ideas.

- Other questions that were raised – who evaluates the Exec Committee? – presumably the site team and NSF. Will get other ideas from Purdue’s VPR. How do we better integrate the thrusts – more coherence needed in writing the research section of the report.

- Wojtek reiterated that all PI’s must reserve the first week of December and make sure to attend the site visit. It is NSF’s expectation that everyone attend and is taken seriously. As a last resort, connecting through the web etc. may be an option.

- Kumar stated that he and others are happy with the report.

**Post Docs**

Wojtek stated that NSF is very happy with our post doc program. We would like to have a 3rd Center post doc at this point, to which we have a few options. We can go the same route as before and conduct a full search. Or could select a group that was identified last time, but didn’t get a postdoc – meaning we could select that group for this 3rd post doc. Wojtek will think about this and discuss it at our next meeting.

Next Executive Committee meeting will be held on February 8, 2013 at 3:00 EST.
SOI STC Executive Committee Meeting by Conference Call
Friday, February 8, 2013

Attendees:
Wojtek Szpankowski
Bob Brown, Managing Director
PR Kumar, Chair
Sergio Verdu
Ananth Grama
Andrea Goldsmith
Venkat Anantharam
Shankar Subramaniam
Kiya Smith

Notes:

P.R. Kumar brought the meeting to order. Notes from the previous meeting were approved by the committee.

Managing Director Remarks

- Summer School, which this year is jointly held with the IEEE IT Society, is scheduled for June 4-7. All speakers have been scheduled. We are asking each of our partner institutions to consider a small sponsorship for the School. Bob will contact each of you soon regarding this. Sergio will be attending the Board of Governors meeting next week and will present a short update about the summer school. Bob will have Brent Ladd send Sergio information to share. We are especially interested in any ideas about potential sponsorship.
- As our fiscal year ended January 31, 2013, STC funds have been extended at each institution until funds for period 4 arrive.
- CSoi staff will be working with each partner institution to plan events/activity on each campus and/or region. Sponsored speakers/lecturers, research poster sessions, and industry roundtables are all possibilities. This will be discussed further down in our agenda – especially as it relates to industry and knowledge transfer.

Director’s Remarks

- Wojtek shared comments about particular areas of the Center.
• During our last site visit the term “Shannon 2.0” was used relative to the progress we were making as a Center. Wojtek is concerned that this term has been misused and that we need to make it clear that we are setting goals and expectations through our grand challenges. Discussion followed – the term “Shannon 2.0” was apparently used by a reviewer during our initial site visit and shouldn’t be used – the term is too undefined and perhaps not appropriate.

• The Center is suffering some from miscommunication about evaluation results etc. Andrea suggested an “all hands” video conference/virtual meeting take place once a quarter. Bob added that Center staff members are committed to a once-a-month update email and a more formal e-newsletter once a quarter.

• Wojtek feels like Center members need to stress collaboration more between each other as we move forward. There are good things happening, but we need to assure all are involved.

• Wojtek suggested the following action items:
  o Executive Committee to send a letter to all clearly stating research goals and Center expectations.
  o Step up collaboration. Ask each member of the Executive Committee to establish a new collaboration within the coming year.
  o Plan an all-hands meeting with all PI’s in attendance – 1 day before site visit to meet and discuss research, collaboration, opportunities, etc.
  o Show more activity on campus. Per Bob’s comments above.

• Ananth stressed that these items need to get underway immediately. And asked if there was a way to give further incentives for collaboration.

• Sergio responded further to the question about “Shannon 2.0”. “We don’t want to be known or labeled as the Shannon 2.0 Center.”

• Wojtek suggested that a clear letter be written to all PI’s in the Center. He asked Kumar, Ananth and Sergio to write a first draft and send to members of the committee. Bob will send a copy of the letter that was sent in 2010 for reference.
Conversation ensued about the letter. The letter will be from the Executive Committee on where we are going and what is expected. Questions about consequences were asked, “What steps are we going to take to be more forceful”? “Showing up for one meeting is not acceptable.” Several members were impacted in the last evaluation, should we identify others at risk and send a targeted letter to those individuals? We can look at past reports to pinpoint individuals. Kumar, Ananth, and Sergio agreed to draft a letter.

- Center Post Doc – we will add another post doc to the Center. The same procedure as before will be used as before with a deadline of March 15.

- Wojtek brought to everyone’s attention that David Tse has been in touch with him about a possible collaboration between our Center and the Simons Institute on Theory of Computing. Simon Center is suggesting we propose an information theory program that would be one semester long and involve 3-4 key people from the Center. They would have to be amenable to spending one semester at Berkeley – probably spring 2015. Please think about possible suggestions for participants.

Kumar and Wojtek suggested that we schedule another meeting in the next two to three weeks as we were running out of time and still needed to discuss a couple of things – and it would be good as a follow up to some of these suggestions as well. We will send out a meeting survey early next week.
SOI STC Executive Committee Meeting by Conference Call
Thursday, March 7, 2013

Attendees:
Wojtek Szpankowski
Bob Brown, Managing Director
PR Kumar, Chair
Sergio Verdu
Ananth Grama
Tsachy Weissman
Venkat Anantharam
Deepak Kumar

Notes:

Notes from the previous meeting were approved.

Managing Director Remarks

- Summer School Update: which this year is jointly held with the IEEE IT Society, is scheduled for June 4-7. Posters will be sent out today – please post them. We have made solicitations for sponsorship. Pledges have been received from the Purdue VPR, Purdue CS Department, Bryn Mawr, TX A&M and Princeton.
- Requests for Period 3 financial reports will go out this week. Please share any questions you may have.
- Our monthly update email went out this week. Please let us know your news so that we can share in future updates.
- Please be reminded during this time of student and faculty recruitment that we as a Center value diversity and are happy to assist in your recruitment efforts of underrepresented students/faculty.

Director’s Remarks

- Wojtek shared comments about two specific issues.
  - The idea of engaging a senior research scientist in the Center that would play an active part in the research of the Center and assist with industry relations etc. was discussed. Wojtek has been made aware of someone that is
Discussion centered on expectations for having a research agenda while also gaining industry support. Tsachy mentioned the possibility of Stanford potentially providing office space. Wojtek feels like 50 percent of time should be for knowledge transfer – especially dealing with industry. Funding such a position is an issue. Comments were made about the need to have an open mechanism for such a search. Also that the job description needs to be more concrete – and that we would need to be realistic about support goals. Wojtek tabled the conversation for further consideration.

- Acknowledging the Center – pushing the Center vision is important when possible by all Center participants (papers, talks, conferences). Deepak spoke to how important it is for outreach and education to acknowledge the Center. Everyone should make a point to mention the Center when they can. Perhaps we can make something available that will assist others in this effort. Wojtek also talked about trying to get a piece about the Center in the Science Times or other publication – now is the right time.

**Letter from Executive Committee**

- Kumar led discussion about the letter that will be sent out to Center faculty. It will be sent out by Bob through email on behalf of the committee. It will not go to students.

**Third Post Doc**

- We have received just a couple of applications to date for our third Center post doc. We will extend the deadline to April 15. “April 15 or until filled”. Barbara Gibson is staffing this and will form a committee to review apps, etc.
Attendees:
Wojtek Szpankowski
Bob Brown, Managing Director
PR Kumar, Chair
Sergio Verdu
Ananth Grama
Shankar Subramaniam
Peter Shor
Venkat Anantharam
Andrea Goldsmith
Deepak Kumar
Kiya Smith

Notes:

Notes from the March 7 meeting were approved.

Managing Director Remarks

- NSF Funds have arrived and are being processed. They should clear to your universities in about 10 days. Any questions please contact Bob.
- A web address was included in the email sent to you on April 26 for the budget period 3-4 for our broader impact budget and expenses.
- Summer school attendance is expected at well over 100. Bob thanked the following for their funding support: Princeton, Bryn Mawr, Berkeley, TX A&M, Purdue VPR, Purdue CS, and the IT Society.
- At the last site visit it was suggested that we need administrative contacts present. Bob will be contacting each of you to identify the representative at your institution. Please respond to this request. The representatives will need to be available by conference/video call – if not in person. The representatives will need to be prepared.

Director’s Remarks

- From a conversation and email from John, he expects a similar site visit team as last year with perhaps several replacements. We are requesting that each
institutions have faculty participation and bring 2 to 3 students to speak about their accomplishments and participate in the all-hands meeting.

- Wojtek introduced and discussed new collaborations being formed in the Center including those between Shankar and Ramki, as well as Kumar and Lynch.
- Bob will send notice sometime in June to start the evaluation process.
- Research scientist, Marcelo Weinberger from HP, will bring in as Senior Scientist – supporting some travel and CSol related events.
- Yuli Baryshnikov and Max Raginsky from UIUC will be asked to join the Center to conduct research with current Center members.
- Third postdoc update – the applicants have been sorted and the committee has chosen 7. The faculty is in the process of selecting one. Deadline is May 3. Decision needs to be made by May 10 and announced by May 15.
- Wojtek briefly introduced a funding system from a Michigan group – to encourage different disciplines to collaborate – they take proposals for research projects and fund them (for this particular example is was around $20 k). This is a proactive way to support collaboration and should be discussed going forward.

**Other**

- Shankar suggested a collective paper from the Center – interdisciplinary efforts. Discussed the idea of exploring an interdisciplinary journal from many different disciplines.
- What are our action items from the Grand Challenges? How are we going to address each? We will discuss at our next meeting.
- Upcoming events. September – Princeton Bell Labs Workshop – October External Advisory Committee and Industry Advisory. December Site Visit.
SOI STC Executive Committee Meeting by Conference Call  
Monday June 10, 2013

Attendees:
Wojtek Szpankowski
Bob Brown, Managing Director
PR Kumar, Chair
Sergio Verdu
Ananth Grama
Peter Shor
Venkat Anantharam
Andrea Goldsmith
Kiya Smith

Notes:

Notes from the April meeting were approved.

Managing Director Remarks

- An email was sent to each committee asking them for administrative representation from their institution for the site visit. If you have not responded, please do.
- We are working on the activity report for faculty evaluations. This is the most important evaluation yet as we prepare for our renewal. Everyone will be asked to help in the completion.
- We are continuing to work with partner institutions to assure we are active on their campuses. Various events are being planned on campuses Berkeley (with Stanford), Princeton (Bell Labs), Howard/Bryn Mawr, MIT, TX A&M, San Diego (next summer school)
- 2013 summer school was held last week at Purdue (jointly with IEEE IT Society). It was attended by over 130 faculty and student participants. A faculty workshop was also held. The feedback has been extremely positive. A diversity workshop was one of the evenings and attended by many and poster sessions each afternoon created engaging conversation with the students. Thank you to everyone for a successful event.
Director’s Remarks

- Marcelo Weinberger had the title of Distinguished Scientist at Hewlett Packard and will keep that title with the Center – it was agreed to let him do so. Marcelo will not receive funding directly from the Center at this time. We will try to provide some admin support as well as travel and for CS/1 related events, etc.
- Discussion was held about how different rules are seen on various campuses relating to administrative expenses etc. Visiting grad students and post docs are treated differently at different campuses – for the Center, this needs to be more uniform. This should be discussed among administrators at the December meeting.
- David Tse at Berkeley is putting in a proposal for the Center to have a semester long research symposium at Berkeley sponsored by Simons Institute.
- Post Doc update – 30 applications were received. The committee chose 7. 2 received bids. Zhiying Wang was chosen to work with Olgica Milenkovic, Nancy Lynch, and Tsachy Weissman. Zhiying is finishing her Ph.D. from Cal Tech in July and will start immediately after that.
- Yuli Baryshnikov and Max Raginsky from UIUC will be asked to join the Center to conduct research with current Center members.

Other

Grand Challenges – Wojtek asked that Kumar call a meeting to set clear goals before the renewal. This will be discussed further at the December 3 All Hands Meeting. Shankar suggested at last meeting to write a vision paper for the Center. Ananth will take the lead and prepare an outline. It was agreed to have a special meeting to discuss.

Upcoming events:

September 5 – Princeton Bell Labs Workshop
September 30 – Industry Advisory Committee - Chicago
October 1 – External Advisory Committee Meeting – Chicago
November 4 – CS/1 event at Berkeley – with Stanford
November 22 – Undergraduate research symposium at Howard – with Bryn Mawr
December 3 – All Hands Meeting -Purdue
December 4 – Site Visit - Purdue
June 2013 – Summer School UC-San Diego (Date to be determined)
SOI STC Executive Committee Meeting by Conference Call
Friday, August 30, 2013

Attendees:
Wojtek Szpankowski
Bob Brown, Managing Director
PR Kumar, Chair of EC
Deepak Kumar
Shankar Subramaniam
Peter Shor
Sergio Verdu
Venkat Anantharam
Andrea Goldsmith
Kiya Smith

Notes:

Minutes from the June 10 meeting were approved.

Managing Director, Bob Brown remarks:

• The NSF Site visit will be Dec 4-5. Dec 3 is an all-hands meeting. We need as many people as possible to attend. We are still in need of administrative representation from each institution. Bob will be sending out another email. The representative should be a dean, dean of research, department head or chair. We are working on the best time to meet all time-zones.
• The NSF report is due 1 month prior to site visit. Brent is still in need of 25% more of the student reports to be turned in.
• Purdue’s Vice President for Research office will be assisting with editing the research content of the report. Deadlines are very important in order to give them the time needed.
• Budgeting process – Purdue Sponsored Programs needs information ahead of time.
• September 30-October 1 is the Industrial Meeting to be held in Chicago. We are in need of a few additional attendees. Bob will follow-up with Andrea for possible names to contact.
• October 1 is the External Advisory meeting in Chicago. Bob is asking for a few of you to call in during the meeting.

Wojtek Szpankowski remarks:

• Barbara Gibson has left the Center to take a new position at Purdue. She will be with us for 10% of her time until December.
• Wojtek, Bob, Brent and two graduate students attended the STC directors meeting last week in Portland, Oregon.
• Bob, Andrea and Ananth are working with industry. We have industry collaboration and support and need to report it accurately in our annual report.
• Wojtek discussed the perception surrounding the term “Shannon 2.0.” We have actually gotten the question – why are we not doing Shannon 2.0? We did not coin this phrase. It was used by one of the reviewers of our preproposal. We are doing exactly what our mission states. We need to show at our next site visit and in our report what our accomplishments are and what our mission is.
• Grand Challenges. We need to show continuation and need to continue to have new ideas. Members discussed their thoughts about our latest grand challenge draft. We need to rewrite the communication and life sciences sections. Shankar agreed to work on the life sciences challenges.
• Members spoke of their thoughts on the latest draft.
• It was announced at the STC Directors meeting that our current grant will close-out 6 months after end of July 2015.
• Evaluation Committee – Kumar, Ananth, Sanjeev, David, and Bob, will form a committee chaired by Kumar and will evaluate this year’s faculty reports. They will make recommendations on how to proceed for renewal.

Upcoming Events:
September 5 – Princeton Bell Labs Workshop
September 30 – Industry Advisory Committee – Chicago
October 1 – External Advisory Committee meeting – Chicago
November 4 – CSol event at Berkeley – with Stanford
November 22 – Undergraduate research symposium at Howard – with Bryn Mawr
December 3 – All Hands Meeting – Purdue
December 4-5 – Site Visit – Purdue
Spring 2013 – Csol event at Texas A&M
Spring 2013 – Csol event at MIT
June 2013 – Summer School UC-San Diego (date to be determined)

The next meeting will be scheduled for late September or early October.
December 3 – All Hands Meeting – Purdue
December 4-5 – Site Visit – Purdue
Spring 2013 – CsoI event at Texas A&M
Spring 2013 – CsoI event at MIT
June 2013 – Summer School UC-San Diego (date to be determined)

The next meeting will be scheduled for November 2013

National Science Foundation
Center for Science of Information
External Advisory Board Meeting
Sofitel Hotel, Chicago
October 1-2, 2013

Advisory Board Members in Attendance: Tony Ephremides, Javier Garcia-Frias, Jerry Gibson (Chair), Michael Honig, and Emina Soljanin

Meeting Summary: The Board met with the CSoI Leadership and staff on October 1, 2013, and heard presentations on an overview of the Center activities, initiatives, and accomplishments for the past year and the future plans for the Center for the coming year. The Board specifically heard presentations from Wojciech Szpankowski on an overview of center activities and some research project details followed by research presentations from Ananth Grama and P. R. Kumar. Brent Ladd gave an update on educational activities and Deepak Kumar discussed diversity efforts. Bob Brown presented some details on activities for knowledge transfer. An afternoon teleconference with Nancy Lynch (MIT), Sergio Verdu (Princeton), Bin Yu (UC-Berkeley), and Andrea Goldsmith (Stanford) also was very fruitful and led to considerable insight into the Center operations and revealed the enthusiasm of participating faculty for the Center’s activities. The discussions were highly interactive, with continual questions by the board and discussions among all present.

Discussions touched on the most recent CSoI Annual Report to NSF, last year’s NSF Site Visit Committee’s report and responses, progress in the past year, and plans for the future. The CSoI continues to evolve as the leadership, staff, and participating faculty gain experience with the structure, and while there are always challenges, the Center appears to be in a strong position for the future.

The Board prepared a draft set of Strengths, Weaknesses, Risks/Challenges, Opportunities, and Recommendations, and two board members, Emina Soljanin and Jerry Gibson, discussed the report with Wojciech Szpankowski, Bob Brown, and Brent
Ladd on the morning of October 2, 2013. The observations of the External Advisory Board are summarized below.

**Strengths:**

The Board feels that the progress of the center has been excellent this past year, with particular achievements in education, research, collaboration, and diversity. Below is a partial listing of the significant strengths of the Center and its operations.

- We continue to be impressed by the high quality of the research performed by the faculty, postdocs and students in the Center, as reflected by the quality and quantity of publications and in the presentations attended by the Advisory Board.

- We commend the Center for the development of mechanisms aimed at fostering collaboration across research areas and institutions. These mechanisms include collaborations among participant students through seminars and workshops, as well as Center-wide postdoctoral positions shared among several institutions.

- The Center has been very successful in the placement of alumni at top-tier institutions, either as faculty or postdocs. We are also encouraged by the resourceful and innovative ways developed by the Center to engage new faculty.

- We are impressed by the exceptional educational component present in the Center, including the development of new courses, the participation of students in seminars and workshops, and training activities for participating faculty.

- The Center is developing strong outreach activities, such as focused workshops and the organization of an annual Summer School, held in 2013 in conjunction with the Information Theory Society Summer School.

- The Center has developed and implemented a solid diversity plan, and it is carrying out specific efforts to foster diversity throughout the Center activities.

- We are encouraged by the success of individual faculty in attracting grants from industry in research areas related to the mission of the Center.

- We are very pleased by the high morale and optimism we observed among the faculty and students involved in the Center across different institutions.

**Weaknesses:**

The presentations and discussions during this External Advisory Board review indicate that the Center has achieved a significant level of maturity in its research activities, collaborations, educational efforts, and operations. While the Board feels that there are no significant gaps that need to be corrected, the Board feels that there are some areas where the Center could improve certain aspects and increase the Center’s effectiveness and impact. These areas are highlighted here as “weaknesses” that are easily
The visibility of the Center’s activities and opportunities within the community and beyond does not reflect the strengths of the Center’s accomplishments. The Center has an impressive team of researchers from the most revered institutions and has an extraordinarily ambitious research agenda. Research efforts on the Science of Information have led to research advances and collaborations that almost certainly would not have occurred without this Center. However, in spite of these achievements and the extraordinary efforts in education and outreach, the Center does not have the visibility commensurate with these successes.

Although the center has set aside 15-20% of the funds for budget reallocations, and this flexibility has led to an enormously successful and productive postdoctoral program, there is limited flexibility for reallocating funds across faculty and institutions to adapt to changing needs and opportunities. As the research interests and participation of the Center faculty change and as new research opportunities emerge, there is a need for further flexibility in the reallocation of funds on an annual basis by the Center Director and the Executive Committee.

Although the inherent connection between the vast agenda of the Center and the research projects is strong, this fact is not clear from the presentations at the Board meeting. In particular, the research plan was not directly reflected in the presentations of the vision, thrust areas, and the research projects described.

Many of the slides presented at the Advisory Board meeting contained an all-inclusive listing of topics that makes it difficult for reviewers to parse out the numerous and substantial contributions and efforts of the Center. As a result, this conceals the underlying merits and achievements of the Center. The presentations should include a hierarchy or more systematic organization of the contributions.

Given the size and scope of Center operations, including the large number of participating institutions, faculty, and students involved, the range of outreach and educational activities, and the needs for planning meetings and disseminating information concerning Center activities, the administrative staff appears to be stretched.

The Center has attempted several different approaches to engage industry and develop meaningful interactions. It appears that the Center has now developed an approach going forward for these interactions and this plan needs to be formalized, communicated, and implemented.
Risks and Challenges:

The Center, like all organizations of this kind, being in the early stages of its development, continues to be exposed to certain risks that represent possible threats for its future as well as some challenges that it needs to overcome. Despite its impressive progress to date, the Center needs to consider these risks and develop a strategy to avoid them and/or overcome them. The three major areas of risk outlined below are interrelated and can be addressed jointly.

1. The Advisory Board has clearly assessed that the Center is producing new knowledge of the highest caliber in the area of the Science of Information. However, its visibility within the community of Information Theory and, more importantly within the communities of the sister fields it serves (like Biology, Computer Science, Economics, etc) is insufficient. If this persists, there is a risk that the Center’s Legacy will be weakened and its impact diminished. A more aggressive strategy of making its accomplishments known is required. A more proactive stance is needed to overcome this challenge.

2. Although the main objective of the Center focuses on the development of the Science of Information, the lack of broad Industry engagement represents a risk of muting or diminishing the impact of its achievements. The Information Industry dominates the world stage today and affects Society and the world in a profound way. The efforts of the Center to engage this Industry have met with limited success. The economic climate in the industry today is perhaps not the strongest and, hence, such engagement does represent a serious challenge.

3. The Advisory Board considers the size of the administrative support of the Center rather limited. Some of the staff is seriously overloaded and, hence, their efforts in, for example, increasing the Center’s visibility and expanding the Center’s ties with Industry may not be as effective as necessary. This, again, represents a threat to the Center. For an effort of its size, the number of administrative personnel seems limited. And given the geographical distribution of the large number of partner Institutions, the concentration of the staff only at the site of the lead Institution increases its exposure to the risk of ineffective promulgation of its successes and reputation to the broader community.

Opportunities:

The impressive scientific results achieved by the center members, collaboration between members, as well as broad center diversity and outreach activities have created a number of opportunities for advancing the center’s agenda. Following are some opportunities that the Advisory Board members think can be immediately utilized.
• The Center continues to have unique opportunities to lay the groundwork for new research directions in the science of information, because of the exceptional qualifications and particular expertise of the Center members, their enthusiasm, and the extremely strong dedication and effectiveness of the supporting staff. The Center has the opportunity to make a strong mark on the information sciences and related areas.

• There are opportunities to promote and advertise the Center through other organizations that promote similar efforts (e.g., IEEE Information Theory Society, DIMACS) and their newsletters (e.g., the quarterly IEEE Information Theory Society Newsletter), because many members of the Center and its Advisory Board are affiliated with such organizations. Similarly, the use of MOOCs would expand the influence of the Center worldwide.

• There is an opportunity to form a followers’ database from the very many people who have already visited and interacted with the Center. Furthermore, such a database would be useful for promoting the Center in a more targeted way. There is an opportunity to add researchers and universities worldwide to be affiliated with the center.

• The Center has the opportunity to tap other resources that could be utilized to strengthen the Center activities. For example, specific strengths as well as individual member’s interests located at some of the partner institutions outside of the lead institution can be used for activities such as industry interactions or promoting diversity, in addition to research efforts.

Recommendations:

Based on the preceding discussion and observations, the board makes the following recommendations, which are intended to address the perceived weaknesses, and better exploit potential opportunities:

1. Additional effort should be devoted to presenting a more concise and coherent view of center goals along with connections to the research thrusts and more specific research efforts. The organization of the presentation(s), in addition to content, should be designed to clarify how the excellent achievements of the Center and ongoing activities are fulfilling the mission of the Center. The presentation of center activities should also highlight related interactions with industry, including new projects and grants in areas that overlap center activities.

2. The Center should consider hiring more staff to facilitate external relations and interactions with industry. For example, that could help to ensure that the Center’s activities are included in appropriate publications, and advertised at
appropriate venues (as described in the subsequent recommendation). The Center should also seek out and take advantage of potential support from the lead institution. However, it should be kept in mind that in general, additional staff does not need to be located at the base of the lead institution. Rather, the location should be selected according to the position. (For example, an industrial relations staff should have easy access to relevant companies.)

3. The Center should consider ways to increase its visibility. Some possibilities for this may include: (1) Reporting Center events and activities in society newsletters; (2) Creating a mailing list of affiliates, colleagues, and alumni to which announcements of activities are periodically sent; and (3) Distributing Center brochures at conferences related to the various center themes.

4. The Center should consider expanding the flexibility of its resource allocation mechanism so that it is better able to adapt and shape its agenda and seed new initiatives. That may even include the possibility of a zero-base for annual budgeting across projects and institutions.

5. The Center should continue to adapt its research vision in accordance with evolving trends. Including big data as part of its research agenda is a good example of this, and the Center may consider sponsoring additional focused workshops in other areas that may benefit from its strengths.

6. Some effort should be devoted to quantifying measures of success for the Center, establishing associated milestones, and describing the process through which the goals (in both research and education) are being achieved.

7. The Center should consider transforming its existing excellent on-line course materials to create a MOOC on Information Science.
Report on the Sixth Annual North American School of Information Theory

Brent T. Ladd, Director of Education, Center for Science of Information

The sixth annual North American School of Information Theory took place June 4-7, 2013 on the campus of Purdue University. Hosted by the Center for Science of Information (http://soihub.org), 140 total students, postdocs, faculty, and professional staff took part in the school this year. A concurrent workshop for faculty and postdocs on teaching science of information courses brought faculty from across the nation who also attended the lectures at the school. The school provides a venue where doctoral and postdoctoral students can meet to learn from distinguished professors in information theory, and form friendships and collaborations. This year the school introduced several interdisciplinary topics in the emerging field of science of information.

The school format has courses by distinguished scientists followed by student poster presentations. The five featured speakers this year were Michelle Effros (Cal-Tech) who explained "Information Theory for Large Networks", Scott Aaronson (M.I.T.) gave an enthusiastic tutorial on "Quantum Computing and Information", P.R. Kumar (Texas A&M) and postdoctoral scholar Jonathan Ponniah co-presented "A Clean Slate Approach to Security of Wireless Networks", Mehmet Koyuturk, gave a survey on "Complex Diseases and Information Theory". Eminol Soljanin (Bell Labs), gave the Padovani Lecture on "The Secret Lives of Codes: From Theory to Practice and Back". Videos and slides of the lectures are viewable on the Science of Information Channel via http://soihub.org/summerschool, as well as on the IEEE IT Society 2013 School page http://www.itsoc.org/north-american-school-2013

Students gave fast-paced overviews of their research during the traditional "one minute madness" series of presentations that was enjoyed by all. Three poster sessions took place during the school. There was a broad scope of topics presented, and many students commented that the interdisciplinary nature of the posters presented provided much insight and helpful discussions. Many connections were made between the students and several mentioned possible future collaborations. Several students offered to lead open problem discussions in the evenings following dinner, and a professional development session on biases and diversity in hiring was also offered to the students one evening.

September 2013

IEEE Information Theory Society Newsletter
We had the most perfect weather we could have asked for during the week of the school, with blue skies and temperatures in the mid 70's. Built into the schedule were lunches and dinners that allowed ample time for students to network and discuss issues in their respective fields. The social program included a sit-down roundtable dinner on the first evening of the school, with an outdoor BBQ cookout on Wednesday evening (with veggie sandwiches and frisbee games too!).

Many people helped make the 2013 school a success. Wojtek Sapankowski and Brent Ladd put the program together, Mike Atwell took care of the many web and print media duties, Kiya Smith and Erica Wilson assisted with organizing meals, lodging, lecture hall, and t-shirts, Bob Brown served as treasurer and organized key sponsorships, Barbara Gibson organized student diversity components including a session on recognizing biases in hiring. Sergio Verdu presented our school proposal and updates to the IT BoG, and the Center for Science of Information Executive Committee provided overall support. Erin Blakeslee, Mike Atwell, and Bob Brown all took great photographs - a full school album available at https://www.facebook.com/media/set/?set=a.477580742319450.1073741827.13209520201341&type=3. Robynne McCormick assisted with financial items, and Andy Thompson tirelessly ran video equipment and edited the final lectures. Deepak Kumar and Mark Ward facilitated a workshop for faculty interested in teaching science of information courses.

A big thank you to all of the sponsors that made the 2013 school possible. In particular, IEEE Information Theory Society, Center for Science of Information, Purdue Computer Science Department, and Vice President for Research, Princeton Electrical Engineering Department, UC Berkeley Departments of Electrical Engineering & Computer Sciences, Statistics, and ERSO, Bryn Mawr College Computer Science Department, and Texas A&M Electrical & Computer Engineering.

We close this brief report with feedback and representative quotes from students and faculty. In addition to gaining experience presenting their research to their peers, students report a 3.7/4 on obtaining useful feedback to their research from talking with other students and faculty, and a 3.4/4 that they were able to start some level of professional connection with their peers for possible collaborations.

"The poster sessions were a great opportunity to see the variety of information-theoretic research that is going on. They facilitated meaningful conversations. I hope to continue discussing my research with some of these students. The one-minute madness sessions were useful, too."

"Sharing my research and getting to see what the other students' research was certainly valuable. It is also good to speak to someone in the same stage of career and talk about practical aspects of the PhD life, questions that you would not feel so comfortable to ask to your advisor, in general."

"I think the summer school was a big success. I enjoyed the talks very much. The poster sessions were very interesting. I had several deep discussions with other researchers about my (and their) research which was very helpful for me."

"The best thing about the summer school for me is getting to know students from other universities, learn about their research and share ideas with them. And even talk about future collaborations!"

"I will look forward to attending this conference again, especially if the accommodations and food are as decent as they were at Purdue. Thank you."

"Great Summer School! Thank you for a really well organized and enriching event!"
Q&A

The Power of Distribution

Nancy Lynch talks about achieving consensus, developing algorithms, and mimicking biology in distributed systems.

Yes, my husband and I had a two-body problem, so we kind of moved around. But math departments were not hiring very much, so in 1977, I went to Georgia Tech as an associate professor of computer science. At Georgia Tech, I was surrounded by applied computer scientists, so I abandoned working in abstract complexity theory and started looking at computer systems. Distributed systems were just beginning to be important at that time, and there were other people at Georgia Tech who were interested in building them. I decided there must be some interesting mathematics to be developed, and began to work on developing a theory for distributed systems.

That work put you back in touch with Michael Fischer, with whom you had worked at MIT.

Yes, Mike and I started working together on this, going back and forth between Georgia Tech and the University of Washington, where he was at the time. We made a lot of progress on this new theory very quickly, and in 1981, on the strength of that work, I went to MIT on a sabbatical, got a tenure offer for the next year, and stayed. And we have been here ever since.

Your most famous result in distributed computing is the so-called FLIP impossibility proof of 1985, which proves that asynchronous systems cannot reach consensus in the presence of one or more failures. Can you talk about how you reached it?

We were studying different models of distributed computing, both synchronous and asynchronous. In synchronous models, computation occurs in lock-step rounds. In asynchronous models, there is no common notion of time, and processors can move at arbitrarily different speeds.

Researchers like Leslie Lamport, Danny Dolev, and Boy Shroyer were studying consensus...
problems in the presence of failures in synchronous models, in the form of Byzantine agreement. They were also studying fault-tolerant clock synchronization. From that problem, I defined an easier problem of "approximate agreement" on real values, where everybody starts from a real value and has to agree on some value that is in the range of all the other values. We studied that first in synchronous models, and then we saw how we could extend the result to asynchronous models. Putting it all together, it seemed pretty natural to consider the problem of exact agreement in asynchronous systems.

Another impetus was the then-current work on database transaction commits. This is a critical example of a practical problem of exact agreement on whether a transaction should commit or abort. It is important in practice for the solution to tolerate some failures, though not necessarily Byzantine failures—just simple stopping failures. And an asynchronous model would be appropriate, because you couldn’t realistically assume absolute bounds on the message delays.

How did your work proceed from there?

At first I thought that we might come up with an algorithm for the asynchronous case of this problem, like we had for approximate agreement. But our attempts failed, so we started trying to find an impossibility result. We went back and forth, working on both directions. We narrowed in on the solution relatively quickly—it didn’t take more than a few weeks. Formulating the ideas nicely, in terms of concepts like bidirectional, came a bit later.

When did you realize FLP’s significance?

I think I understood the practical significance for transactions relatively quickly, but we did not predict the impact it would have on later research. Theoreticians have developed many results that extend FLP to other problems, and many results that circumvent the limitation using such methods as randomization and failure detectors. Most interestingly, I think, is that FLP triggered the development of algorithms that established a clear separation of requirements for fault-tolerant consensus problems: safety, properties of agreement and validity, which are required to hold always, and termination properties, which are required to hold during stable periods. These algorithms are not only interesting theoretically, but provide interesting guidelines for development of practical fault-tolerant systems.

In the 1980s, you also began work on interactive or I/O, automata, which are used to model distributed algorithms. Mark Tuttle and I developed the I/O automata modeling framework for asynchronous distributed systems early on, in 1987. We had some asynchronous distributed algorithms and we wanted to prove that they worked, but we were doing a lot of work to define our models and found that we were repeating that work in different papers. So we stepped back and developed a rigorous math model for systems with interacting components.

Later, you extended the work to cover synchronous systems, as well. The I/O automata framework doesn’t deal with timing, so we defined another model, the Timed I/O Automata model, to cover synchronization. This is what we use as the foundation of our work on algorithms for mobile systems and wireless networks. My student Roberto Segala also worked with me to develop probabilistic versions, which are useful for describing randomized algorithms and security protocols.

So you have various frameworks that support the description of individual components in a system, and can then be used to produce a model for the entire system.

I don’t think the effort in developing these models is done yet. It would be nice to combine all the frameworks into one that includes discrete, continuous, timed, and probabilistic features, which is what’s needed to understand modern systems.

Let’s talk about some of your more recent work.

For the past 10 years or so, my group and I have been working on distributed algorithms for dynamic networks, in which the network changes over time because participating nodes can join, leave, fail, recover, and move, all while the algorithm is operating. We have designed algorithms that maintain consistent data, synchronize clocks, compute functions, and coordinate robots. We have also worked quite a bit recently on top-level wireless communication issues—managing contention among different senders in wireless networks.

Are there certain techniques, principles, or characteristics you have found helpful, or does every field network bring its own set of problems?

Some common techniques emerge. For example, we try to implement abstract layers, which are basically simpler models, over more complex models. You could have a Virtual Node layer that adds fixed nodes at known locations to a mobile wireless network and makes it easier to write higher-level algorithms. Or a Reliable Broadcast layer that makes it easier to send messages to content management in wireless networks, producing a more reliable substrate for writing higher-level algorithms. Various algorithmic techniques do also recur, such as quorum-based reliable data management, random methods for information dissemination, and back-off protocols for scheduling transmissions.

You have also begun working on biologically inspired distributed algorithms. Can you talk a bit about that work?

I’m really just starting on this, but the idea is that biological systems are a lot like distributed algorithms. Why? Because they consist of many components, interacting to accomplish a common task, and communicating mainly with nearby components. So I’m reading about, for example, self-organizing systems of insects and bacteria, systems of cells during development, and neural networks, and I’m trying to apply a distributed algorithms viewpoint. It’s too early to see what will emerge, but surely we can design models, state problems, describe systems at different levels of abstraction as distributed algorithms, analyze the algorithms, and maybe even prove lower bounds.

Laure Hoffmann is a technology writer based in Brooklyn, NY.

© 2013 ACM 0001-0782/13/03
2013 North American School of Information Theory
June 4-7, Purdue University
West Lafayette, IN, USA

Tutorials by: Scott Aaronson, MIT ◆ Michelle Effros, CalTech
Mehemet Koyuturk, Case Western Reserve ◆ P.R. Kumar, Texas A&M
Padovani Lecturer: Emina Soljanin, Bell Labs

Daily Research Poster Sessions
Registration is open at http://www.soihub.org/itschool

Host: Center for Science of Information NSF Science & Technology Center

soihub.org/itschool
Teaching a Science of Information Course
Faculty Workshop June 5-6, 2013
Purdue University, West Lafayette, IN, USA

Team building for collaborative Science of Information courses:
- Discuss ideas, approaches, and share course content
- Discover funding opportunities for teaching a SoI course

Early Registration: April 1, 2013
http://www.soihub.org/course-workshop

Funds are available for travel and lodging
Contact Dr. Mark D. Ward, Purdue University, mdw@purdue.edu

Center for Science of Information NSF Science & Technology Center
EAEO. Funded under grant agreement CCF-0939370
2012-13 Prestige Lecture Series on Science of Information

Distributed Algorithms for Wireless Networks
Nancy A. Lynch
NEC Professor of Software Science and Engineering
Dept. of Electrical Engineering and Computer Science
PI, Computer Science and Artificial Intelligence Laboratory
MIT

October 15th, 2012
2:30pm - 3:30pm
LWSN 1142
Purdue University

Data Mining - Association Rules: Twenty Years and Beyond
Tomasz Imieliński
Professor
Dept. of Computer Science
Rutgers University

October 22nd, 2012
2:30pm - 3:30pm
LWSN 1142
Purdue University

A Look at the Unconscious Brain Under General Anesthesia
Emery N. Brown
Professor of Computational Neuroscience and Health Sciences and Technology
MIT

Warren M. Zapol Professor of Anaesthesia
Massachusetts General Hospital
Harvard Medical School

November 5th, 2012
2:30pm - 3:30pm
LWSN 1142
Purdue University

Center for Science of Information
NSF Science & Technology Center
soihub.org
Distributed Computing Theory for Wireless Networks

Abstract:
Nearly all modern computer systems are distributed. Most are based on platforms that change dynamically, and many rely on wireless communication. These systems must deal with complications such as node mobility and message collisions. Consequently, such systems are hard to understand and hard to build.

As yet, there is no comprehensive theory to help us out. Such a theory should identify key problems and sub-problems, and should include algorithms, lower bounds, and ways of composing algorithms to build more complex algorithms. The theory should span from the basic communication model to high-level data-oriented and control-oriented applications.

In this talk, I will provide a high-level overview of my research group’s recent work on a theory for dynamic systems, especially wireless networks and mobile systems. I will start by describing our work on networks with reliable communication channels, including algorithms for maintaining atomic memory in dynamic networks, Virtual Node abstraction layers, algorithms for coordinating robot swarms, and algorithms for computing functions in dynamic networks. I will then describe our work for models with unreliable channels, in particular, channels that exhibit message collisions and resulting losses. I will finish with a discussion of some issues involving unpredictable message delivery range.

Bio:
Nancy Lynch is the NEC Professor of Software Science and Engineering in the Department of Electrical Engineering and Computer Science at MIT. She heads the Theory of Distributed Systems research group in MIT’s Computer Science and Artificial Intelligence Laboratory. She is also currently a Fellow at the Radcliffe Institute for Advanced Study.
A Look at the Unconscious Brain Under General Anesthesia

Abstract:
General anesthesia is a drug-induced, reversible condition comprised of five behavioral states: unconsciousness, amnesia (loss of memory), analgesia (loss of pain sensation), akinesia (immobility), and hemodynamic stability with control of the stress response. The mechanisms by which anesthetic drugs induce the state of general anesthesia are considered one of the biggest mysteries of modern medicine. We have been using three experimental paradigms to study general anesthesia-induced loss of consciousness in humans: combined fMRI/EEG recordings, high-density EEG recordings and intracranial recordings. By using a wide array of signal processing techniques, these studies are allowing us to establish precise neurophysiological, neuroanatomical, and behavioral correlates of unconsciousness under general anesthesia. Combined with our mathematical modeling work on how anesthetics act on neural circuits to produce the state of general anesthesia we are able to offer specific hypotheses as to how changes in level of activity in specific circuits lead to the unconscious state. We will discuss the relation between our findings and two other important altered states of arousal: sleep and coma. Our findings suggest that the state of general anesthesia is not as mysterious as currently believed.

Speaker Bio:

He is an anesthesiologist/statistician whose methodology research develops signal processing algorithms to characterize how the brain represents and transmits information. His experimental research uses a systems neuroscience approach to study how anesthetic drugs act in the brain to create the state of general anesthesia.

Dr. Brown is currently a member of the Burroughs-Wellcome Fund Board of Directors, NSF Mathematical and Physical Sciences Advisory Committee, and the Board of Directors of the International Anesthesia Research Society. Dr. Brown is a fellow of the IEEE, the American Statistical Association, the American Association for the Advancement of Science, the American Academy of Arts and Sciences, and a member of the Institute of Medicine.

Dr. Brown is the recipient of a 2007 NIH Director's Pioneer Award, the 2011 Sacks Award from the National Institute of Statistical Sciences, and a 2012 NIH Director's Transformative Research Award.
Data Aggregation in Wireless Networks

Abstract:
Today’s wireless networks tend to be centralized: they are organized around a fixed central backbone such as a network of cellular towers or wireless access points. However, as mobile computing devices continue to shrink in size and in cost, we are reaching the point where large-scale ad-hoc wireless networks, composed of swarms of cheap devices or sensors, are becoming feasible. My research studies the theoretical computation power of such networks, and asks what tasks are they capable of carrying out, how long does solving particular tasks take, and what is the effect of the unpredictable network topology on the network’s computation power.

I will describe one particular set of results, focusing on the hardness of aggregating data in wireless networks which suffer from asymmetric communication. I will show that asymmetry can be a major concern: simple tasks such as computing the size of the network, which require only O(D) rounds in symmetric networks of diameter D, now require \( \Omega(n) \) rounds in asymmetric networks, even when their diameter is 2; even “easier” tasks, such as computing the minimum input, require \( \Omega(\sqrt{n}) \) rounds. These and related lower bounds are shown by reduction from well-known problems in communication complexity, including Set Disjointness and Gap Hamming Distance. I will also describe a new communication complexity problem, called Task Allocation, give a lower bound on its 2-player communication complexity, and show how it can be applied to obtain an \( \Omega(n) \)-round lower bound on finding rooted spanning trees in networks of diameter 2.

Speaker Bio:
Rotem is currently a Postdoctoral Scholar at Toronto University. She received her PhD at MIT, advised by Prof. Nancy Lynch. Before coming to MIT, she completed a B.A. and an M.Sc. at the Technion, Israel Institute of Technology. Her M.Sc. was in formal verification of hardware, under the supervision of Prof. Orna Grumberg.
An Upper Bound on the Convergence Time for Quantized Consensus

Abstract:
We analyze a class of distributed quantized consensus algorithms for arbitrary networks. In the initial setting, each node in the network has an integer value. Nodes exchange their current estimate of the mean value in the network, and then update their estimation by communicating with their neighbors in a limited capacity channel in an asynchronous clock setting. Eventually, all nodes reach consensus with quantized precision. We start the analysis with a special case of a distributed binary voting algorithm, then proceed to the expected convergence time for the general quantized consensus algorithm proposed by Kashyap et al. We use the theory of electric networks, random walks, and couplings of Markov chains to derive an $O(N^3 \log N)$ upper bound for the expected convergence time on an arbitrary graph of size $N$, improving on the state of art bound of $O(N^4 \log N)$ for binary consensus and $O(N^5)$ for quantized consensus algorithms. Our result is not dependent on graph topology. Simulations on special graphs such as star networks, line graphs, lollipop graphs, and Erdős-Rényi random graphs are performed to validate the analysis.

This work has applications to load balancing, coordination of autonomous agents, estimation and detection, decision-making networks, peer-to-peer systems, etc.

Speaker Bio:
Shang is currently a graduate student from the Department of Electrical Engineering, Princeton University. She is co-advised by Prof. Sanjeev R. Kulkarni and Prof. Paul W. Cuff. Shang earned her M.A. degree in 2011 from the Department of Electrical Engineering, Princeton University, and her B.E. degree in 2009 from the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University. Shang’s research interests include recommendation systems, social networks and information aggregation.
Building Random Trees from Blocks

Abstract:
We study the stochastic behavior of parameters of a network using a tree model of growth, from building blocks that are themselves rooted trees. We give asymptotic results about the number of leaves, depth of nodes, total path length and height of such trees.

Joint work with:
Dr. Hosam Mahmoud, Department of Statistics, The George Washington University,

Dr. Mark Daniel Ward, Department of Statistics, Purdue University, West Lafayette IN.
Network Reconstruction of Dynamic Biological Systems: Doubly Penalized LASSO

Reconstruction of biological networks is a crucial step in extracting information from a large volume of experimental data. Various methods have been developed to reconstruct networks from data, each of which possesses its own strength and disadvantages. In this talk, I will first describe examples of network reconstruction techniques including classic optimization-based approaches (e.g., least-squares methods), dimensionality reduction methods (e.g., statistical significance tests combined with either principal components regressions (PCR), or partial least squares (PLS), Bayesian networks and hybrid methods (e.g., Linear Matrix Inequalities (LMI) and Least Absolute Shrinkage and Selection Operator (LASSO)). Next, I will introduce our most recent method, called Doubly Penalized Linear Absolute Shrinkage and Selection Operator (DPLASSO) for network reconstruction with the intent to combine the beneficial features of a regression and statistical significance testing-based method and a penalized optimization method. I will present results from applications to simulated data from synthetic random networks as well as from a biological system, namely the cell cycle in budding yeast.

http://www.soihub.org/connect
Information-theoretic Approaches for Biological Networks Reconstruction

High-throughput technologies for biological measurements generate vast amounts of quantitative data, which necessitates the development of advanced approaches to data analysis to help understand the underlying processes and networks. Reconstruction of biological networks from measured data of different components is a significant challenge in systems biology, whose goal is to develop network models capable of the quantitative mapping of inputs to responses resulting in a given phenotype. In this presentation, I use an information theoretic approach to develop a data-driven parsimonious input-output model of the phosphoprotein-cytokine network. Then I propose an information theoretic approach to reconstruct dynamic biological networks from time-course data. The results of this study are essential for understanding the functional and dynamical behaviors of biological networks.

http://www.soihub.org/connect

Supported by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370.
Two problems In Integrity Verification

With advent of cloud computing it is possible for weak clients to delegate their computational workloads to much more powerful off-site workers. In many scenarios, this asymmetry in computational power has been accompanied by asymmetry of trust, where clients have limited means to check correctness of results provided by the cloud service. In this presentation I will survey our work on two problems in verified computation: database query verification and verifying execution of arbitrary programs. We build on recent theoretical progress to arrive at practical solutions to both of them. In the former we achieve integrity, completeness and freshness verification for certain queries in dynamic outsourced databases, while in the latter we provide a practical non-interactive argument system for arbitrary NP languages.
Defining Differential Privacy for Social Network Analysis

Privacy of social network data is a growing concern that threatens to limit access to this valuable data source. Analysis of the graph structure of social networks can provide valuable information for revenue generation and social science research, but unfortunately, ensuring this analysis does not violate individual privacy is difficult. Simply anonymizing graphs or even releasing only aggregate results of analysis may not provide sufficient protection. Differential privacy is an alternative privacy model, popular in data-mining over tabular data, that uses noise to obscure individuals’ contributions to aggregate results and offers a very strong mathematical guarantee that individuals’ presence in the data-set is hidden. Analyses that were previously vulnerable to identification of individuals and extraction of private data may be safely released under differential-privacy guarantees. We review two existing standards for adapting differential privacy to network data and analyze the feasibility of several common social-network analysis techniques under these standards. Additionally, we propose out-link privacy and partition privacy, novel standards for differential privacy over network data, and introduce powerful private algorithms for common network analysis techniques that were unfeasible to privatize under previous differential privacy standards.
Large Alphabet Compression and Predictive Distributions through Poissonization and Tilting

Abstract—This talk introduces a convenient strategy for coding and predicting sequences of independent, identically distributed random variables generated from a large alphabet of size $m$. In particular, the size of the sample is allowed to be variable. The employment of a Poisson model and tilting method simplifies the implementation and analysis through independence. The resulting strategy is optimal within the class of distributions satisfying a moment condition, and is close to optimal for the class of all i.i.d distribution or strings of a given length. Moreover, the method can be used to code and predict sequences with a condition on the tail of the ordered counts. It can also be applied to distributions in an envelope class.
Mathematical Theory of Communication

Mathematical Theory of Communication (MTC):

Pioneered by Claude Shannon of Bell Laboratories in the late 1940’s, MTC is the foundational theory behind data encoding and transmission. Without it we would not have cell phones, iPods, databases on hard drives, or the internet we know today. Shannon developed the concepts of redundancy, and information entropy. Ultimately, MTC deals with transmission between a sender and receiver of symbols encoded in well-formed strings of signals.


Center for Science of Information
NSF Science & Technology Center
http://soihub.org

Our mission is to advance science and technology through a new quantitative understanding of the representation, communication and processing of information in biological, physical, social, and engineered systems.

informationbits is a program of the Center for Science of Information promoting appreciation and understanding of the role of information in science, technology, and society. We welcome your inquiries. Visit http://soihub.org
What is Information?

The word information comes from the Latin *informare* meaning “give form to”. The Oxford definition, *Knowledge communicated concerning some particular fact, subject or event; connects the word to both knowledge and communication.*

At the physical fundamental level is the science of information theory, which defines information as a quantitative measure of communicative exchanges irrespective of the context or meaning.

\[
I(X;Y) = H(Y) - H(Y|X)
\]

\[
H(X) = \sum_i p_i \log_2 \left( \frac{1}{p_i} \right)
\]

Whatever your own concepts are, there is no doubt that most of us are swimming in a flow of information that is increasing at an exponential rate.

How can we best capture useful knowledge from this flow and make better informed decisions? How might elements of space, time, semantics, structure, and context be assimilated into a scientific understanding of information?

At the Center for Science of Information, our mission is to attempt to answer these questions ... come join the discussion! [http://www.soihub.org](http://www.soihub.org)
Professors Sergio Verdu and Tsachy Weissman start an information theory forum

Sergio Verdu (Princeton; currently on leave at Stanford) and Tsachy Weissman (Stanford) have started an information theory forum at Stanford, which includes a weekly meeting for students and faculty to present tutorials on topics related to the science of information and a blog on information theory (the Princeton-Stanford blog): http://blogs.princeton.edu/blogs/

CSOI Howard undergraduates present at the National Technical Association annual conference

CSOI funded Howard University undergraduates, Chawasse Ademuyiwaag and David Harris, presented at the National Technical Association annual conference at Morgan State University in September 2012. David Harris presented on "Predicting Marked Code-switching in African Languages"; Chawasse Ademuyiwaag presented on "Empirical Evaluation of Strategy-Prefering of Majority Judgments". The National Technical Association Inc. was established in 1928 with the purpose of creating a medium for expressing minority professional concerns for the lack of minority participation in the mainstream architectural and engineering activities as well as modern technology in America.

Deepak Kumar contributes a 'personal perspective' via an article in IEEE Computer

Deepak Kumar (Bryn Mawr) was invited by the Guest Editors of the March 2013 issue of IEEE Computer's Special Issue on Gender Diversity in Computing to provide a personal perspective. The piece is titled, "Serving Women with a Purposeful Vision", and can be accessed here: http://www.computer.org/csdl/magpc/03/03/index.html

Nancy Lynch talks about achieving consensus, developing algorithms, and mimicking biology in distributed systems


Awards

Purdue graduate student, Apjita Sen, awarded National Science Foundation Fellowship


CSOI Channel scholars receive several awards

For more info: http://csolab.org/news-events.php?id=579

Student member of CSOI web team wins awards

CSOI Student Web Production Assistant Edgar Marta wins an award for 'Social Innovation' during the Purdue Coding Camp and was a finalist in the West Michigan Coding Competition held at Purdue.

Pulkit Grover Receives Leonard G. Abraham Award

Former CSO post-doc, Dr. Pulkit Grover, has earned the Leonard G. Abraham Award from the IEEE Communication Society. Presented annually, the award honors the best original paper published in any journal financially sponsored or co-sponsored by the society in the past three years.

Mark Ward and Co-i awarded NSF Grant

Mark Ward was the PI on a team of faculty who were recently awarded a NSF grant titled, "METP: Sophomore Transitions: Bridges into a Statistical Major and Big Data Research Experiences via Learning Communities". This Statistics Using Learning Communities (STAT) LLC project will address two transitions in the training of undergraduate statistics students: (1) The bridge from first-year general curriculum into sophomore year Statistics major courses; and (2) The bridge into a student's first research experience in data analysis, especially with big data. For more info: http://www.nsf.gov/awardsearch/showAward?AWD_ID=1246813

UPCOMING EVENTS

External Advisory Committee Meeting: October 1, Chicago, Selfel
51st Annual Allerton Conference on Communication, Control and Computing: October 2, Allerton Retreat Center, Monticello, IL
48 Hands Meeting: December 3, Purdue University
NSF Site Visit: December 4 - 5, Purdue University
Dr. Ramkrishna and others submit an invited manuscript to "Current Opinion in Chemical Engineering"

Hyoung-Seob Seng, Frank Demtröder and Dasarowami Ramkrishna submitted the following invited manuscript to Current Opinion in Chemical Engineering: "Modeling Metabolic Systems: The Need for Dynamics". This manuscript was the result of support from CSoi.

CSoi Welcomes
New Faculty

CSoi is happy to announce that three new faculty have been added to the Center.

Yuli Baryshnikov, Professor of Mathematics and Electrical Engineering at The University of Illinois

Tomasz Jerkowski, Professor, Computer Science, Rutgers University

Moein Razeghi, Assistant Professor of Electrical and Computer Engineering at The University of Illinois

Distinguished Scientist

As part of the Center's directive to continually enhance our research mission and to further our technology transfer initiatives, we are pleased to announce that Dr. Marcelo Weinberger is joining as a Distinguished Scientist.

CSoi Awards

Zhijing Wang was selected as this year's Center Postdoctoral Research Fellow. She received a B.S. degree from Tsinghua University in 2007, and an M.S. degree from CalTech in 2009, and will finish her Ph.D. in Electrical Engineering from CalTech in July, 2013. Her research interests include information theory, error-correcting codes and coding for storage devices. Zhijing will be working primarily with Dr. Olgica Milenkovic (UIC), Dr. Tsachi Weissman (Stanford) and Dr. Nancy Lynch (MIT).

CSoi Information

Videos of lectures with integrated slides are now posted for the 2013 Center summer school, along with institute student posters and school photos album: http://bit.ly/13csoi_summer


Like the Center for Science of Information on Facebook

https://www.facebook.com/pages/Center-for-Science-of-Information-CSoi/123899558032091?

and follow us on Twitter

https://twitter.com/ScienceofInfo

Always feel free to contact us

We are here to help. Contact our administrative assistant, Kiya Smith (kiya@purdue.edu) and/or Managing Director, Rob Brown (Rob.Brown@purdue.edu).
The 2013 School of Information Theory is organized by Center for Science of Information (http://www.soihub.org), a National Science Foundation Science and Technology Center, and is sponsored by the IEEE Information Theory Society. Hosted at Purdue University from Tuesday, June 4 to Friday, June 7, 2013, the school provides a venue where doctoral and postdoctoral students can meet to learn from distinguished professors in Information Theory, and form friendships and collaborations. Advanced undergraduates and M.S. students are also welcome. This year the school will introduce several interdisciplinary topics in the emerging field of science of information. Students will present their own research via a poster during the school. Although the focus is on Information Theory, interdisciplinary topics are welcome, e.g., topics related to mathematics, physics, biology, control, networking, etc. A separate, but concurrent, workshop for faculty and post-docs will be running during afternoons of June 5-6. The topic is teaching a science of information style course. Details on the workshop can be found at http://www.soihub.org/course-workshop

Announcements and Important Deadlines
Last updated: July 15, 2013 10:55 AM EDT
Video recordings of Invited talks
Open Problem Sessions [PDF]
Submitted One Minute Madness Slides & Posters [Google Spreadsheet]
One Minute Madness Presentation Order [PDF]
Activities/Walking Map (PDF) | [Google Map]
Announcements - May 21
Registration Closed: May 15, 2013
Faculty Workshop June 5-6, 2013

Team building for collaborative Science of Information courses:

- Discuss ideas, approaches, and share course content
- Discover funding opportunities for teaching a Sol course

The interdisciplinary topic is "Science of Information". This workshop is aligned with a $25 million NSF initiative: The Center for Science of Information.

The organizers will share materials from courses we have taught in Science of Information. Attending faculty will likewise share their own course experiences.

As a key outcome, faculty will be prepared to teach their own Science of Information courses spanning the computational and mathematical sciences, biology, engineering, physics, and beyond.

The workshop is especially well-suited for junior faculty and/or post-docs who are interested in discussing and learning about teaching courses in the emerging Science of Information field.

Funding for Travel, Lodging

A limited number of funds for travel and rooms are available for up to twenty participants on a competitive basis decided by the Workshop PIs (with up to $750 travel/lodging/meals reimbursement per person).
Our mission is to advance science and technology through a new quantitative understanding of the representation, communication, and processing of information in biological, physical, social, and engineered systems.
Encryption: Security Through Mathematics

Overview

Encryption is the transformation of data into a form unreadable by anyone without a secret decryption key. Its purpose is to ensure privacy by keeping the information hidden from anyone for whom it is not intended, even those who can see the encrypted data. For example, one may wish to encrypt files on a hard disk to prevent an intruder from reading them.

Start!

Instructor

Mark Daniel Ward
Associate Professor,
Dept. of Statistics,
Purdue University

Credentials

Those completing all Quizzes and receiving 80% or higher on Final Exam receive

Syllabus/Suggested Schedule

Week 1 - Ciphers
Week 2 - Cryptography Theory and Entropy
Week 3 - Public Key Cryptography
Week 4 - The RSA Algorithm Under the Hood Part 1
Week 5 - The RSA Algorithm Under the Hood Part 2
2013 North American School of Information Theory, June 4–7

Report
The 2013 School of Information Theory is organized by Center for Science of Information (http://www.scihub.org), a National Science Foundation Science and Technology Center, and is sponsored by the IEEE Information Theory Society. Hosted at Purdue University from Tuesday, June 4 to Friday, June 7, 2013, the school provides a venue where doctoral and postdoctoral students can meet to learn from distinguished professors in information theory, and form friendships and collaborations. This year the school will introduce several interdisciplinary topics in the emerging field of science of information. Students will present their own research via a poster during the school. Although the focus is on information theory, interdisciplinary topics are welcome, e.g., topics related to mathematics, physics, biology, control, networking, etc.

Announcements and Important Deadline:
Recorded video lectures, posters, photos and summary are available online at http://www.scihub.org/summer-school

Open Problem Sessions [PDF]
Submitted One Minute Madness Slides & Posters [Google Spreadsheet]
One Minute Madness Poster Presentation Schedule [PDF]
Activities/Walking Map [PDF] | [Google Map]
Inferring the Effective TOR-Dependent Network: A Computational Study in Yeast

Shahin Mohammadi, Shankar Subramaniam, and Ananth Grama

Background
Calorie restriction (CR) is one of the most conserved non-genetic interventions that extends healthspan in evolutionarily distant species, ranging from yeast to mammals. The target of rapamycin (TOR) has been shown to play a key role in mediating healthspan extension in response to CR by integrating different signals that monitor nutrient-availability and orchestrating various components of cellular machinery in response. Both genetic and pharmacological interventions that inhibit the TOR pathway exhibit a similar phenotype, which is not further amplified by CR.

Results
In this paper, we present the first comprehensive, computationally derived map of TOR downstream effectors, with the objective of discovering key lifespan mediators, their crosstalk, and high-level organization. We adopt a systematic approach for tracing information flow from the TOR complex and use it to identify relevant signaling elements. By constructing a high-level functional map of TOR downstream effectors, we show that our approach is not only capable of recapitulating previously known pathways, but also suggests potential targets for future studies.

Information flow scores provide an aggregate ranking of relevance of proteins with respect to the TOR signaling pathway. These rankings must be normalized for degree bias, appropriately interpreted, and mapped to associated roles in pathways. We propose a novel statistical framework for integrating information flow scores, the set of differentially expressed genes in response to rapamycin treatment, and the transcriptional regulation network. We use this framework to identify...