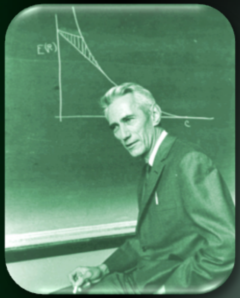
Center for Science of Information



NSF Center for Science of Information: Overview

Industrial Workshop, Chicago, 2013

Bryn Mawr
Howard
MIT
Princeton
Purdue
Stanford
Texas A&M
UC Berkeley
UC San Diego
UIUC

Outline

1. Science of Information

2. Center Mission

   STC Team & Staff

   Integrated Research

3. Grand Challenges

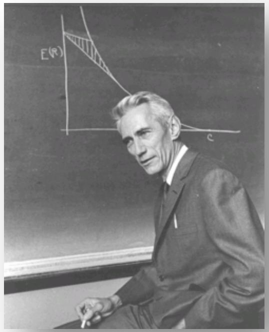
4. Research Results

Shannon Legacy

The Information Revolution started in 1948, with the publication of:

A Mathematical Theory of Communication.

The digital age began.



Claude Shannon:
Shannon information quantifies the extent to which a recipient of data can reduce its statistical uncertainty.
"semantic aspects of communication are irrelevant . . ."

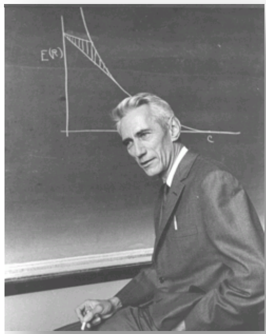
Fundamental Limits for Storage and Communication (Information Theory Field)

Applications Enabler/Driver:
CD, iPod, DVD, video games, Internet, Facebook, WiFi, mobile, Google, . .

Design Driver:
universal data compression, voiceband modems, CDMA, multiantenna, discrete denoising, space-time codes, cryptography, distortion theory approach to Big Data.

Shannon Legacy

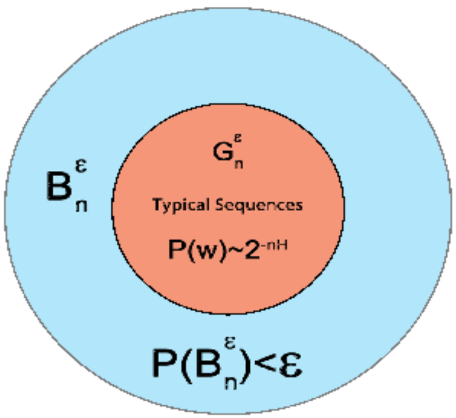
C. Shannon, A Mathematical Theory of Communication, *Bell System J.*, **27**, 379–423, 1948.



Theorem 1 & 3. (Shannon 1948; Lossless & Lossy Data Compression)

compression bit rate $\geq$ source entropy $H(X)$

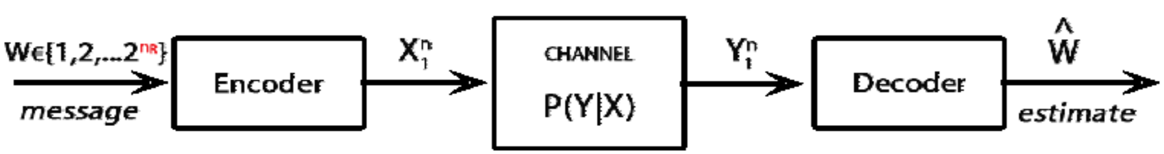
for distortion level D:
lossy bit rate $\geq$ rate distortion function $R(D)$



Theorem 2. (Shannon 1948; Channel Coding)

In Shannon's words:

It is possible to send information at the capacity through the channel with as small a frequency of errors as desired by proper (**long**) encoding. This statement is not true for any rate greater than the capacity.



Post-Shannon Challenges

1. **Back from infinity**: Extend Shannon findings to finite size data structures (i.e., sequences, graphs, sets, social networks, big data structures), that is, develop information theory of various data structures beyond first-order asymptotics.

2. **Science of Information**: Extend Information Theory to meet new challenges in

biology, massive data, communication, knowledge extraction, economics, …

In order to accomplish it we must understand new aspects of information in:

structure, time, space, and semantics,

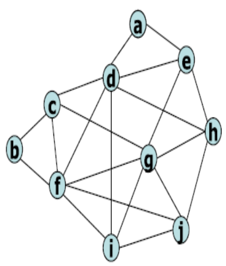
and

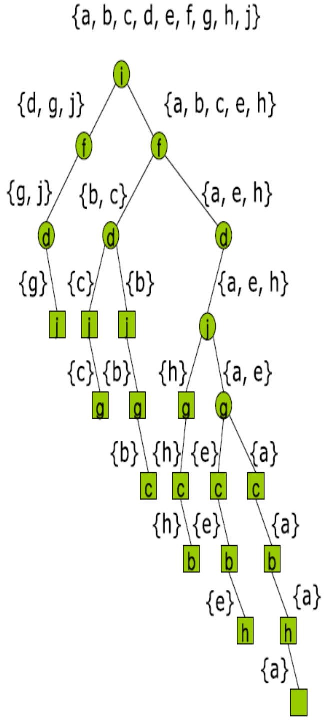
dynamic information, limited resources, complexity, representation-invariant information, and cooperation & dependency.

Post-Shannon Challenges

Structure:

Measures are needed for quantifying information embodied in structures (e.g., information in material structures, nanostructures, biomolecules, gene regulatory networks, protein networks, social networks, financial transactions).





B1 = 0100110100001110101

B2 = 1001011000000101

Szpankowski & Choi : Information contained in unlabeled graphs & universal graphical compression.

Grama & Subramaniam : quantifying role of noise and incomplete data, identifying conserved structures, finding orthologies in biological network reconstruction.

Neville: Outlining characteristics (e.g., weak dependence) sufficient for network models to be well-defined in the limit.

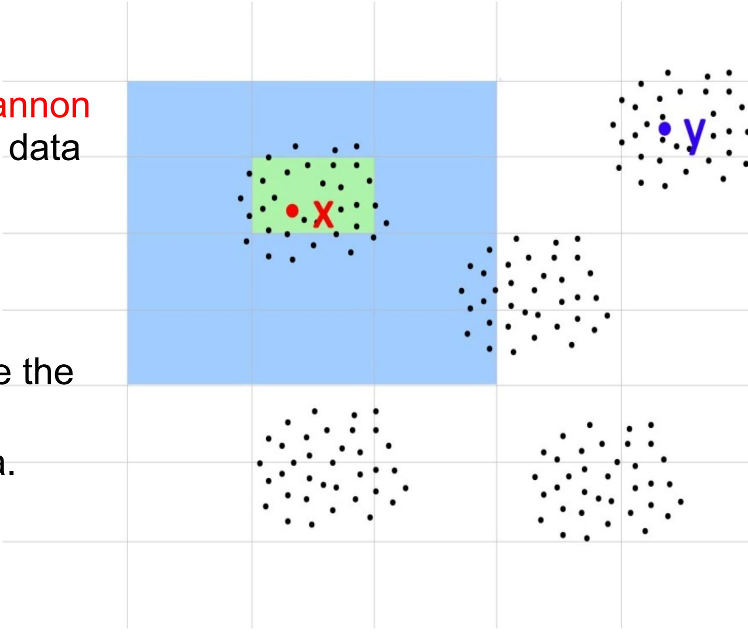
Yu & Qi: Finding distributions of latent structures in social networks.

Post-Shannon Challenges

Time:

Classical **Information Theory** is at its weakest in dealing with problems of delay (e.g., information arriving late may be useless of has less value).

Verdu , Polyanskiy, Kostina: major breakthrough in extending Shannon capacity theorem to finite blocklength information theory for lossy data compression.
Kumar: design reliable scheduling policies with delay constraints for wireless networks; new axiomatic approach to secure wireless networks.
Weissman : real time coding system with lookahead to investigate the impact of delay on expected distortion and bit rate.
Subramaniam: reconstruct networks from dynamic biological data.
Ramkrishna: quantify fluxes in biological networks by developing a causality based dynamic theory celluar metabolism taking into account a survival strategy in controlling enzyme syntheses.



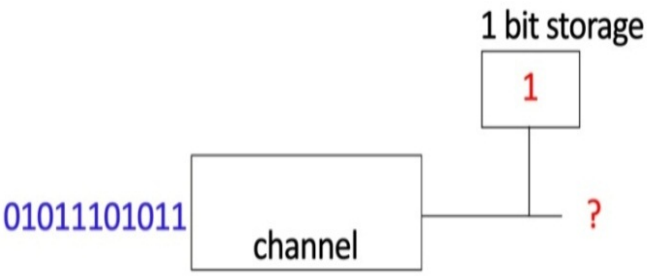
Space:

Bialek explores transmission of information in making spatial patterns in developing embryos – relation to information capacity of a communication channel.

Post-Shannon Challenges

Limited Resources:

In many scenarios, information is limited by available computational resources (e.g., cell phone, living cell).



Bialek works on structure of molecular networks that optimize information flow, subject to constraints on the total number of molecules being used.

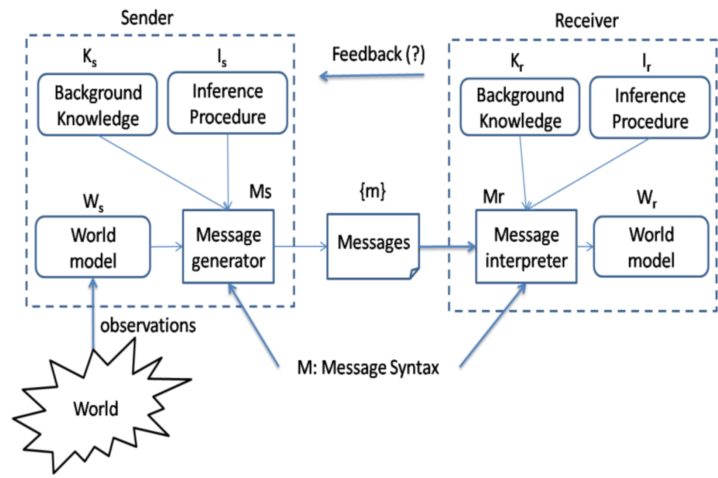
Verdu, Goldsmith investigates the minimum energy per bit as a function of data length in Gaussian channels.

Semantics:

Is there a way to account for the meaning or semantics of information?

Sudan argues that the meaning of information becomes relevant whenever there is diversity across communicating parties and when parties themselves evolve over time.

New collaboration between Sudan and Tse on human communication.



Post-Shannon Challenges

Representation-invariance:

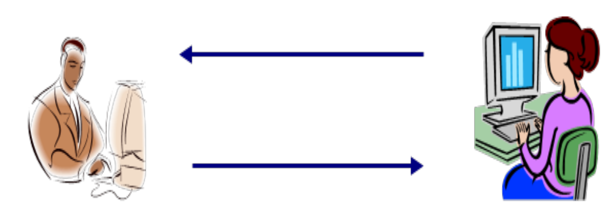
How to know whether two representations of the same information are information equivalent?

(Universal language a la Kieffer's weak compressibility?)

Learnable Information (BigData):

Data driven science focuses on extracting information from data. How much information can actually be extracted from a given data repository? How much knowledge is in Google's database? Information Theory of Big Data?

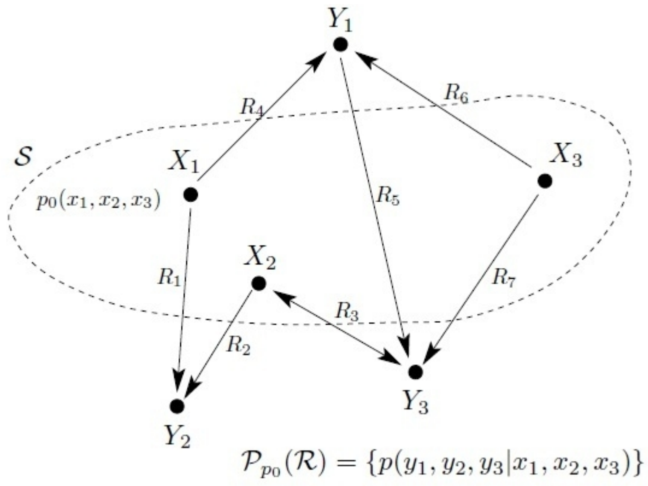
- Courtade, Weissman and Verdu developing information-theoretic approach to recommender systems based on novel use of multiterminal source coding under logarithmic loss.
- Atallah: how to enable the untrusted remote server to store and manipulate the client's confidential data.
- Grama, Subramaniam, Weissman & Kulkurani work on analysis of extreme-scale networks; such datasets are noisy and their analyses need fast probabilistic algorithms for effective search in compressed representations.



Owns data X but has no computing power to manipulate it, or room to store it locally

Computes outputs that depend on X without learning either X or the outputs

Post-Shannon Challenges



Cooperation & Dependency:

How does cooperation impact information (nodes should cooperate in their own self-interest)?

Cover initiated a theory of cooperation and coordination in networks, that is, they study the achievable joint distribution among network nodes, provided that the communication rates are given.

Dependency and rational expectation are critical ingredients in Sims' work on dynamic economic theory (value of information)

Coleman and Subramaniam study statistical causality in neural systems by using Granger principles.

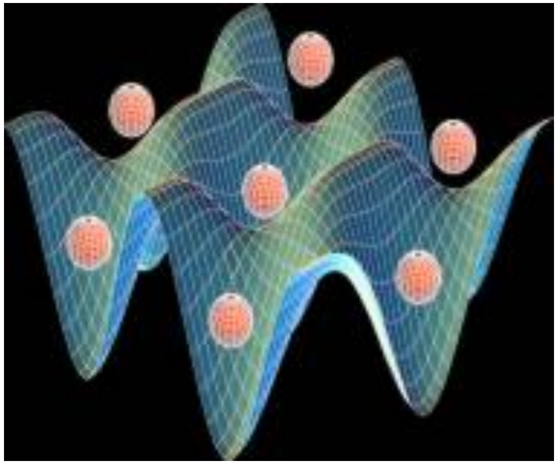
Quantum Information:

The flow of information in macroscopic systems and microscopic systems may posses different characteristics.

Aaronson developed computational complexity of linear systems.

Shor showed entangelment in particles with three spin states.

Quantum Fault Tolerant Computation.



Outline

1. **Science of Information**

2. **Center Mission**

   STC Team & Staff

   Integrated Research

3. **Grand Challenges**

4. Research Results

Mission and Center Goals

Advance science and technology through a new quantitative understanding of the representation, communication and processing of information in biological, physical, social and engineering systems.

Some Specific Center's Goals:

- define core theoretical principles governing transfer of information,
- develop metrics and methods for information,
- apply to problems in *physical* and *social sciences*, and *engineering*,
- offer a venue for multi-disciplinary long-term collaborations,

R
E
S
E
A
R
C
H

- explore effective ways to educate students,
- train the next generation of researchers,
- broaden participation of underrepresented groups,

Education
&
Diversity

- transfer advances in research to *education* and *industry*

Knowledge Transfer

STC Team

Bryn Mawr College: D. Kumar

Howard University: C. Liu, L. Burge

MIT: P. Shor (co-PI), M. Sudan (Microsoft)

Purdue University: (lead): W. Szpankowski (PI)

Princeton University: S. Verdu (co-PI)

Stanford University: A. Goldsmith (co-PI)

Texas A&M: P.R. Kumar

University of California, Berkeley: Bin Yu (co-PI)

University of California, San Diego: S. Subramaniam

UIUC: O. Milenkovic.

R. Aguilar, M. Atallah, C. Clifton, S. Datta, A. Grama, S. Jagannathan, A. Mathur, J. Neville, D. Ramkrishna, J. Rice, Z. Pizlo, L. Si, V. Rego, A. Qi, M. Ward, D. Blank, D. Xu, C. Liu, L. Burge, M. Garuba, S. Aaronson, N. Lynch, R. Rivest, Y. Polyanskiy, W. Bialek, S. Kulkarni, C. Sims, G. Bejerano, T. Cover, T. Weissman, V. Anantharam, J. Gallant, C. Kaufman, D. Tse,T.Coleman.

Wojciech Szpankowski, *Purdue*

Andrea Goldsmith, *Stanford*

Peter Shor, *MIT*

Sergio Verdú, Princeton

Bin Yu, *U.C. Berkeley*

STC Staff

Director – Wojciech Szpankowski

Managing Director – Bob Brown

Education Director – Brent Ladd

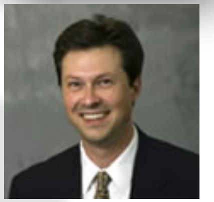
Diversity Director – Barbara Gibson

Multimedia Specialist – Mike Atwell

Administrative Asst. – Kiya Smith

Bob Brown
Managing Director

Brent Ladd
Education Director



Barbara Gibson,
Diversity Director

Mike Atwell
Multimedia Specialist

Kiya Smith
Administrative Asst.

Center Participant Awards

- **Nobel Prize** (Economics): Chris Sims
- **National Academies** (NAS/NAE) – Bialek, Cover, Datta, Lynch, Kumar, Ramkrishna, Rice, Rivest, Shor, Sims, Verdu.
- **Turing award** winner -- Rivest.
- **Shannon award** winners -- Cover and Verdu.
- **Nevanlinna Prize** (outstanding contributions in Mathematical Aspects of Information Sciences) -- Sudan and Shor.
- **Richard W. Hamming Medal** – Cover and Verdu.
- **Humboldt Research Award** -- Szpankowski.

Integrated Research

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


S. Subramaniam


A. Grama

Research Thrusts:

1. **Life Sciences**
2. **Communication**
3. **Knowledge Management** (BigData)

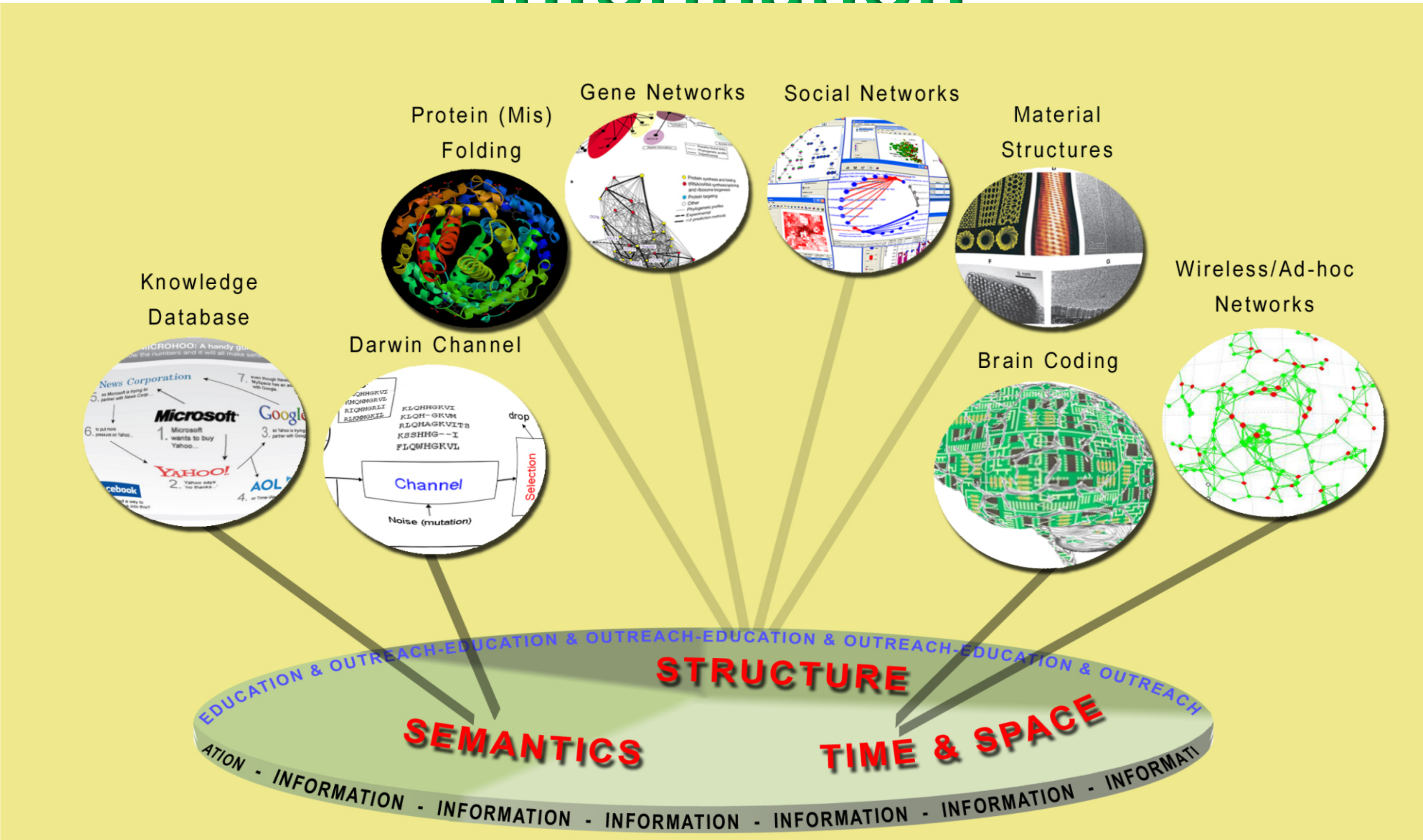

David Tse


T. Weissman


S. Kulkarni


M. Atallah

Application Thrusts vs Information



Outline

1. **Science of Information**

2. **Center Mission**

   STC Team & Staff

   Integrated Research

3. **Grand Challenges**

4. Research Results

Grand Challenges in Knowledge Extraction

Big data domain exhibits complexities that often violate (oversimplified?) assumptions used in IT:

- Large (peta and exa scale)
- Noisy (high rate of false positives and negatives)
- Multiscale (interaction at different levels of abstractions)
- Dynamic (temporal and spatial changes)
- Heterogeneous (high variability over space and time)
- Distributed (collected and stored at distributed locations)
- Elastic (flexibility to data model and clustering capabilities)
- Complex dependencies (structural, long term)
- High dimensional

Ad-hoc solutions do not work at scale!

We need new foundation and fundamental results

Grand Challenges in Knowledge Extraction

The greatest challenge in applying information-theoretic principles is to develop meaningful notions of "structural information" and establishing corresponding fundamental results.

Discovering Structure in Data: To identify, encode, and test the underlying statistically correlated structure of "big" data, there is a tradeoff between accuracy of the data model and the amount of data needed to 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 of the data structures and the efficiency (time) of the algorithms that query and access .

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.
Propose information-theoretic metrics that apply to sets of findings (context).

Grand Challenges in Communication

- Delay in dynamic networks
- Coordination over networks
  - Game theoretic solution concept for coordination over networks
  - Applications in control.
- Relearning information theory in non-asymptotic regimes
- Modeling energy, computational power, feedback, and delay
- Multiterminal source coding (with application to recommender systems)
- Back from Infinity (analytic and non-asymptotic information theory)
- Information and computation
  a. Quantifying fundamental limits of in-network computation, and the computing capacity of networks for different functions
  b. Complexity of distributed computation in wireless and wired networks
  c. Information theoretic study of aggregation for scalable query processing
- Value of information & Timeliness in dynamic economic systems with objective beyond reliable communication
- Quantum fault tolerant computation and quantum capacity of Gaussian channel

Grand Challenges in Life Sciences

- Sequence Analysis
  - Reconstructing sequences for emerging nano-pore sequencers.
  - Optimal error correction of NGS reads (suffix tree approach)
- Genome-wide associations
  - Rigorous approach to associations while accurately quantifying the prior in data
- Darwin Channel, and Repetitive Channel, and Protein-Protein Channel
- Semantic and syntactic integration of data from different datasets
- Network Reconstruction (model selection that fits sparse and temporal data)
- Construction of flux models guided measures of information theoretic complexity of models
- Quantification of query response quality in the presence of noise
- Rethinking interactions and signals from fMRI
- Develop in-vivo system to monitor 3D of animal brain (will allow to see the interaction between different cells in cortex – flow of information, structural relations)

Outline

1. Science of Information

2. Center Mission
   STC Team & Staff
   Integrated Research

3. Grand Challenges

4. Research Results (P. R. Kumar and A. Grama)

Research Accomplishments

- ✓ **Information Theoretic Approach to Recommender Systems and Data Compression for Similarity Queries**
  Courtade, Ingber, Weissman, and Verdu – Stanford & Princeton

- ✓ **Information Theory of DNA Sequencing**
  Tse, Motahari, Bresler, and Bresler – Berkeley (Purdue, UIUC)

- ✓ **Axiomatic Approach Clean Slate to Secure Wireless**
  Kumar, Ponniah, and Hu -- Texas A&M & UIUC (MIT)

- ✓ **Structural (Graphical) Compression**
  Szpankowski, Choi, Magner, and Grama, – Purdue (UIUC)

- ✓ **Pathways in Biological Regulatory Networks**
  Grama, Mohammadi, and Subramaniam – Purdue & UCSD (UIUC)

Outline

1. Science of Information

2. **Center Mission**

   Integrated Research

   Research Accomplishments

   Information Theoretic Approach to Recommender Systems and Data Compression for Similarity Queries

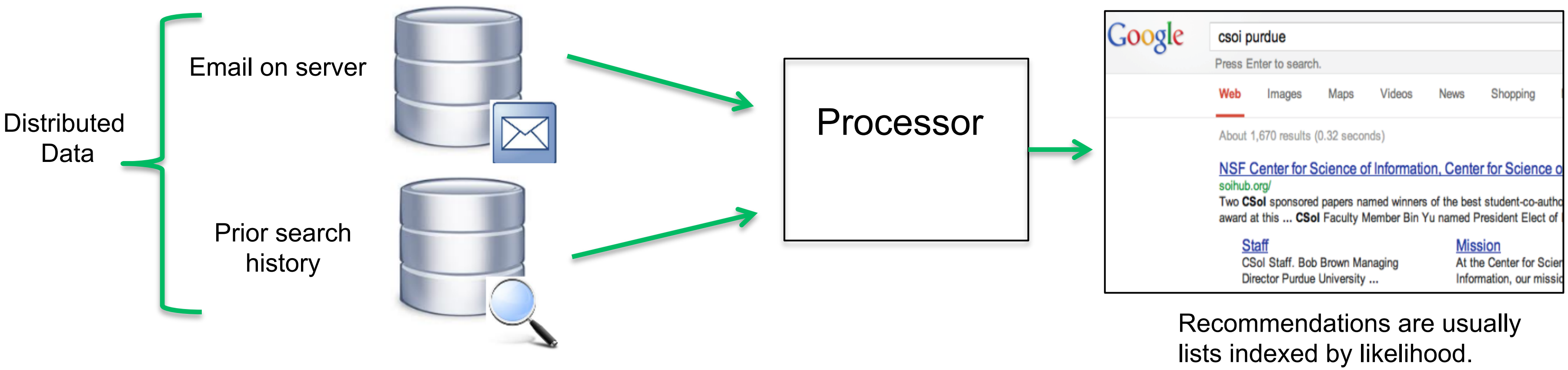
   Courtade, Ingber, Weissman, Verdu – Stanford & Princeton

3. Management
4. Budget

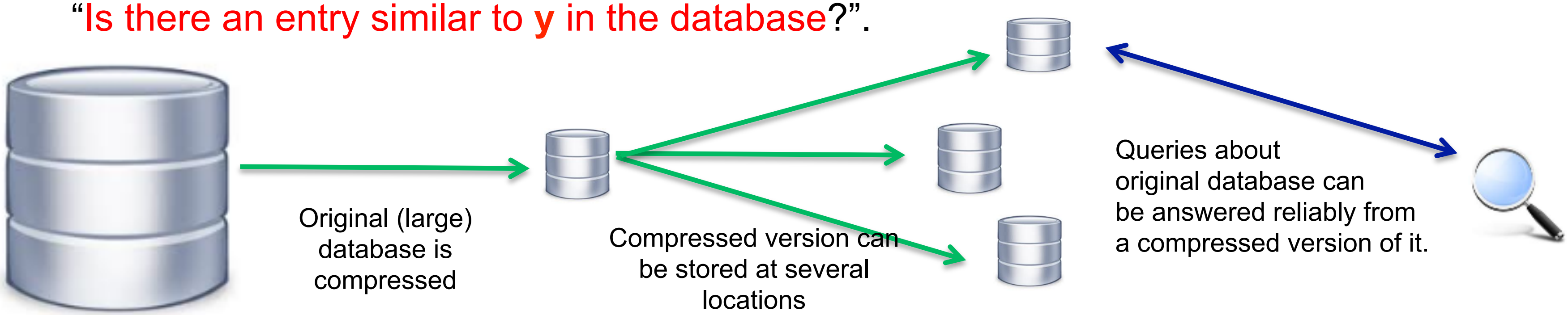
Modern Data Processing

Data is often processed for purposes other than reproduction of the original data: (new goal: *reliably answer queries rather than reproduce data*!)

- Recommendation systems make suggestions based on prior information:



- Databases may be compressed for the purpose of answering queries of the form: "Is there an entry similar to **y** in the database?".



Fundamental Limits

- **Fundamental Limits** of distributed inference

  (e.g., data mining viewed as inference problem: attempt of inferring desired content given a query and massive distributed dataset)

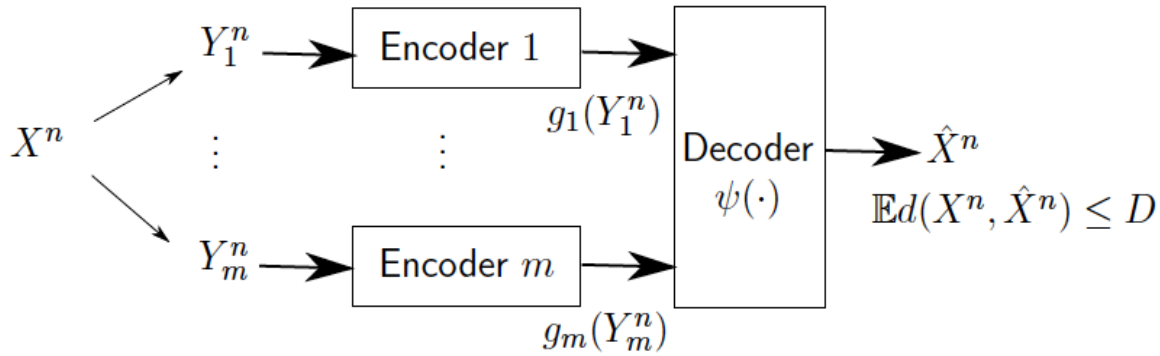
- **Fundamental Tradeoff**: what is the minimum description (compression) rate required to generate a quantifiably good set of beliefs and/or reliable answers

- **General Results**: *queries can be answered reliably if and only if the compression rate exceeds the identification rate*!

- Practical **algorithms** to achieve these limits

Information-Theoretic Formulation and Results

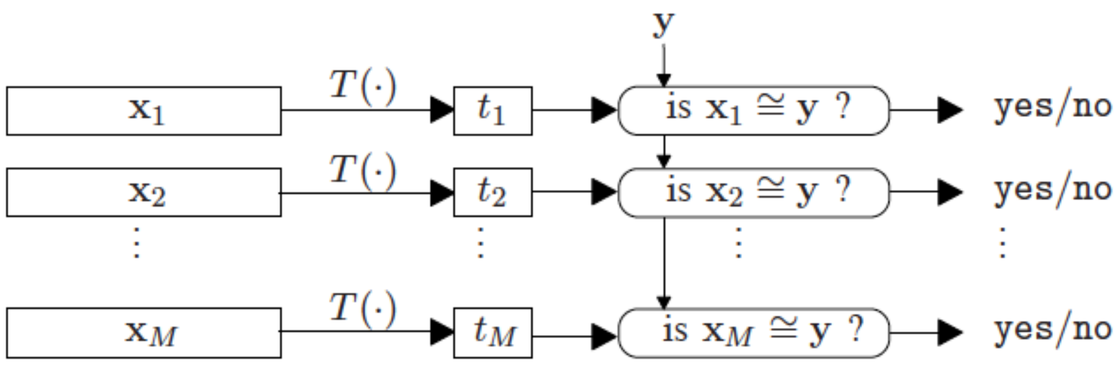
Distributed (multiterminal) source coding under logarithmic loss:



- Logarithmic loss is a natural penalty function when processor output is a list of outcomes indexed by likelihood.
- Rate constraints capture complexity constraints (e.g., limited number of queries to database, or compressed/ cached version of larger database).
- Multiterminal (distributed) problems can be solved under logarithmic loss!
- Such fundamental limits were unknown for the last 40 years(!) except in the case of jointly Gaussian sources under MSE constraints.

Information-Theoretic Formulation and Results

Quadratic similarity queries on compressed data



Goal: assuming there is no false negative minimize the false positive:

- Minimum rate of signature function *T*(.) has been characterized when **x** and **y** are Gaussian and similarity is measured by MSE.
- Query reliability exponent is explicitly characterized when **x** and **y** are Gaussian.
- Universal scheme posed which allows reliable queries, regardless of distribution on **x** and **y**.

Outline

1. Science of Information

2. **Center Mission**

   Integrated Research

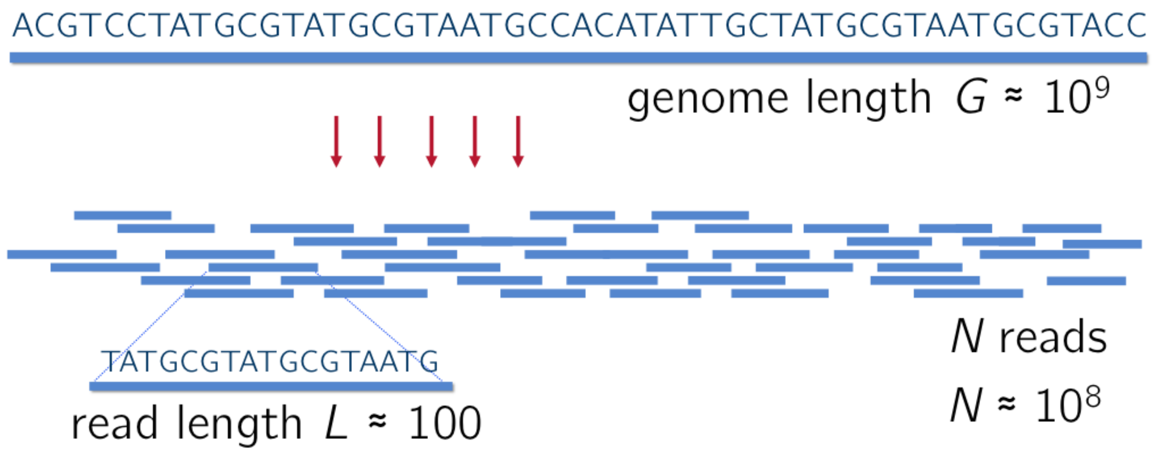
   Research Accomplishments

   Information Theory of DNA Sequencing

   (Tse, Motahari, Bresler and Bresler -- Berkeley)

3. Management
4. Budget

Information Theory of Shotgun Sequencing



Reads are assembled to reconstruct the original genome.

State-of-the-art: many sequencing technologies and many assembly algorithms, ad-hoc design tailored to specific technologies.

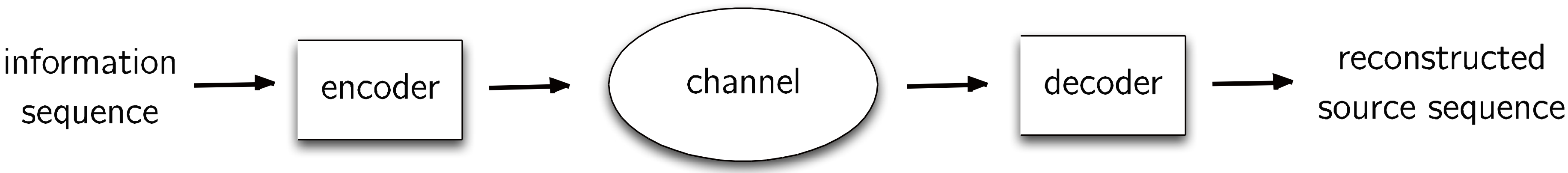
Our goal: a systematic unified design framework.

Central question: Given statistics of the genome, for what read length L and # of reads N is reliable reconstruction possible?

An optimal algorithm is one that achieves the **fundamental limit**.

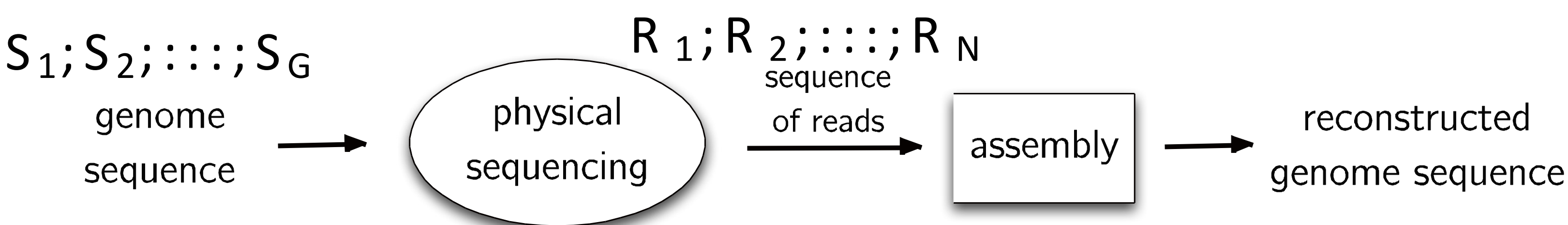
Communication and Sequencing: An Analogy

Communication:



$$\text{max. communication rate } R = \frac{C_{\text{channel}}}{H_{\text{source}}} \text{ source sym / sec.}$$

Sequencing:

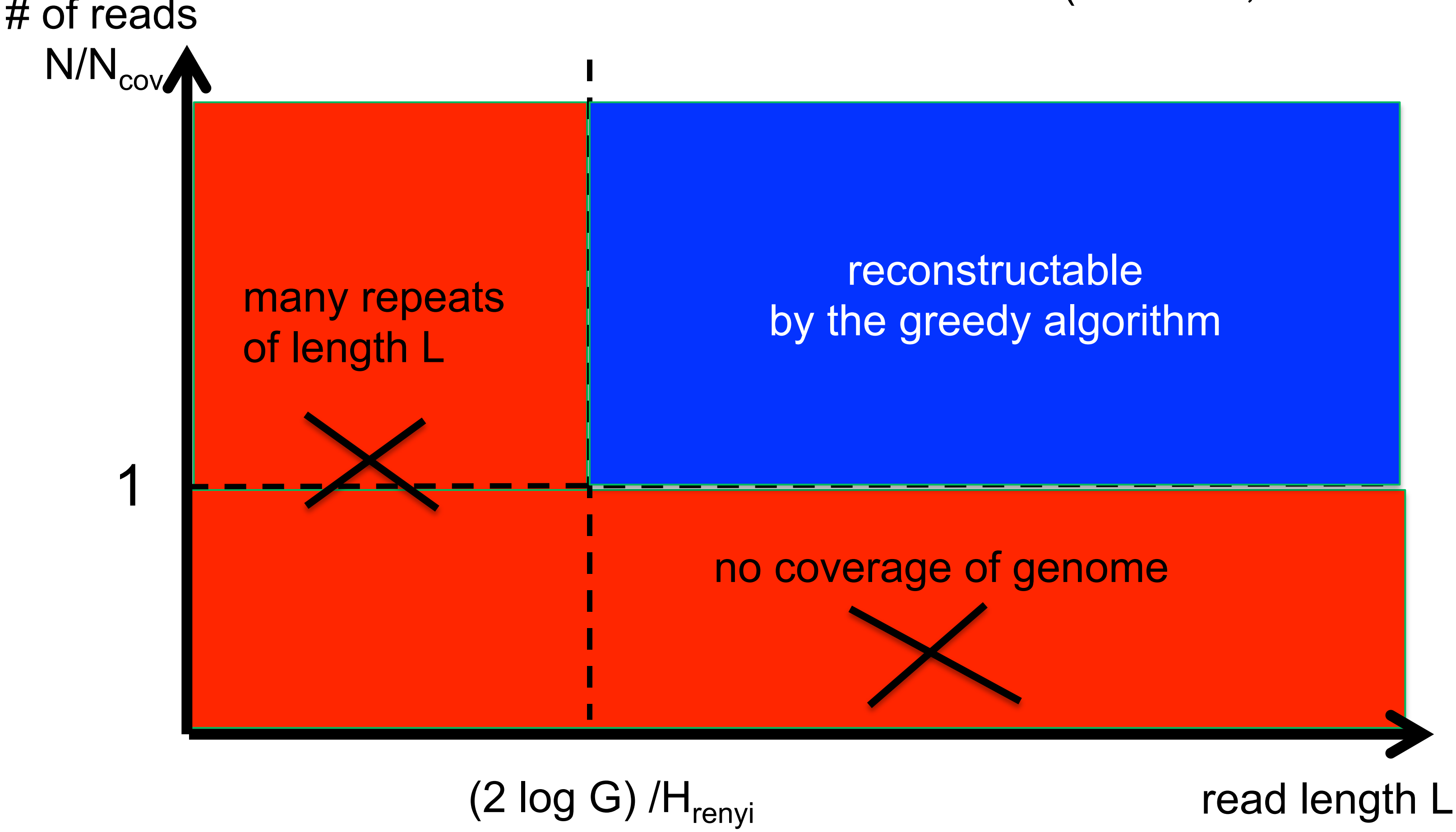


$$\text{max. sequencing rate } R = \frac{G}{N_{\min}} \text{ genome sym / read}$$

Computing N_{min} is equivalent to computing the max. sequencing rate.

Simple Model: I.I.D. Genome

(Motahari, Bresler & Tse 12)

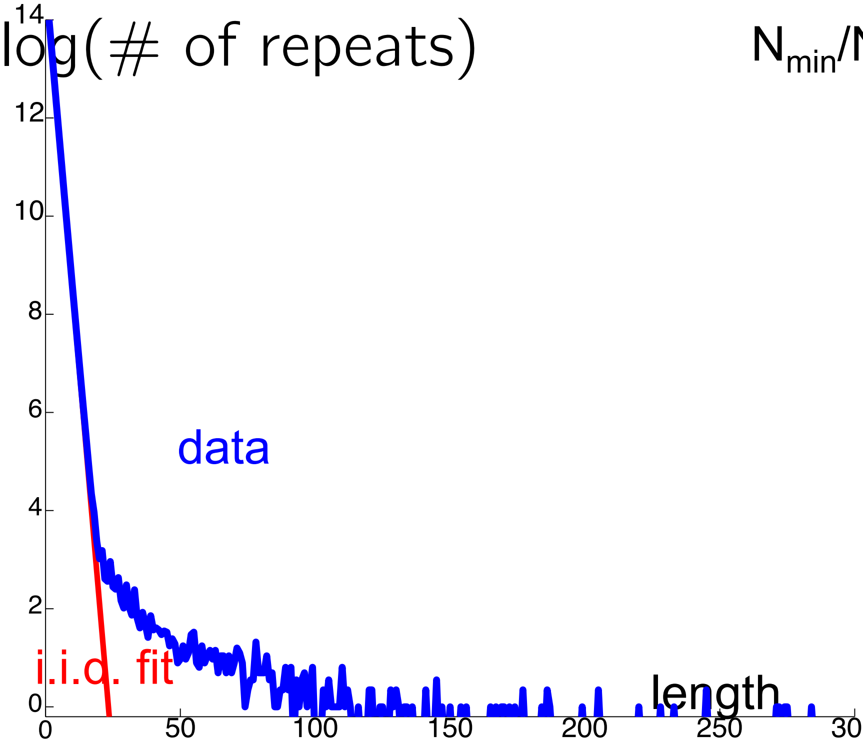


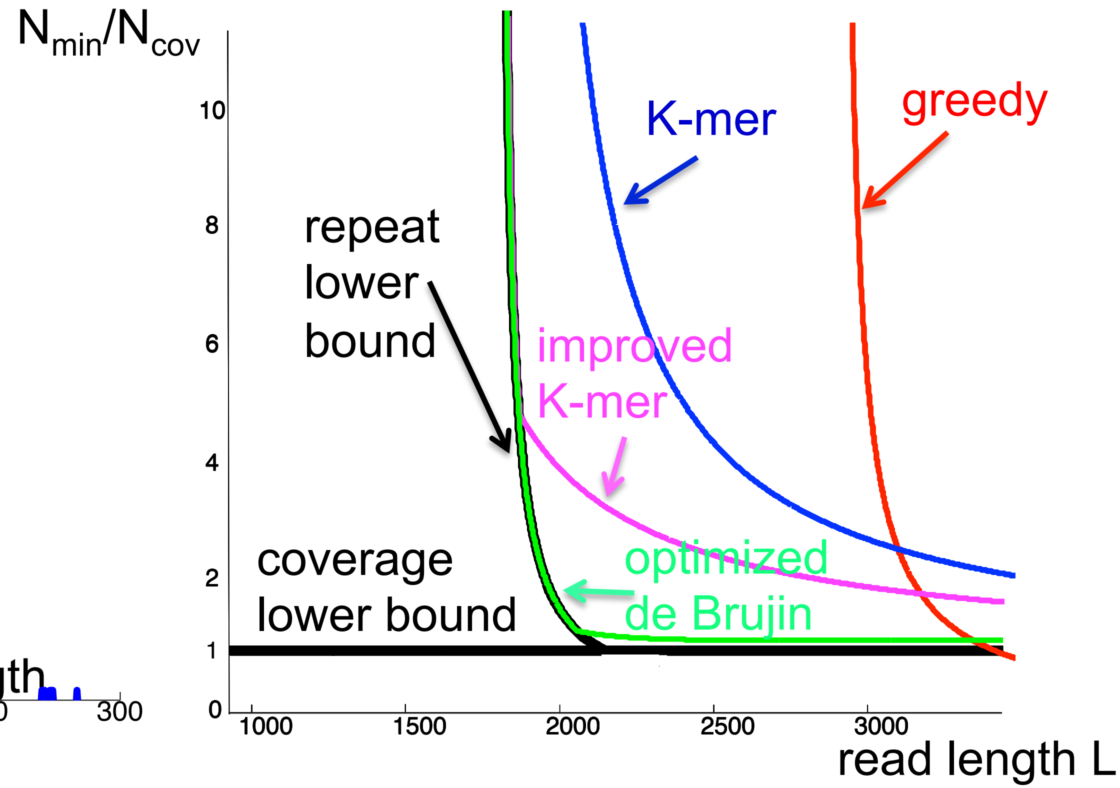
Real Genome Statistics

(Bresler, Bresler & Tse 12)

Example: Stapholococcus Aureus

upper and lower bounds based on repeat Statistics (99% reconstruction reliability)





A data-driven approach to finding near-optimal assembly algorithms.

Outline

1. Science of Information

2. **Center Mission**

   Integrated Research

   Research Accomplishments:

   Axiomatic Approach Clean Slate to Secure Wireless

   (Kumar, Ponniah and Hu -- Texas A&M & UIUC)

3. Management
4. Budget

Motivation

- Usual design approach, including in wireless networks, has been: build a system for good performance, often ignoring security concerns.
- Then
  - Some ATTACK is identified
  - A DEFENSE is developed for that attack
  - Then another ATTACK is identified
  - Another DEFENSE for that attack
- Result
  - A sequence of patches
  - An arms race
- Difficulty
  - We don't know what other attacks are possible
  - No guarantees of security

Wormhole attack
Sybil attack
Rushing attack

ARIADNE
TESLA

Goals

- Can we develop a principled and **holistic approach** to **security**?
  - Security addressed first, not an afterthought
  - Performance addressed second; optimized while preserving security
- Security objective
  - A clean slate approach to secure wireless networking
  - An axiomatic approach to security

    (e.g., imobile nodes, good nodes are half duplex, bad nodes: full duplex, full knowledge, clock synchronization, etc)
  - Defend against Byzantine behavior of malicious nodes
  - Provable security: Guaranteed if model assumptions satisfied
  - Complete suite of algorithms/protocols
  - An "existence theorem" and providing algorithms
- Also a performance guarantee: Max-Min Optimal
  - Max is over protocols, min is over all actions of malicious nodes

Security Performance: New Approach

Protocols yield network **Max-Min** optimal (for good nodes) with respect to utility

$$\underset{\text{Protocols}}{Max} \quad \underset{\text{All behaviors of bad nodes}}{Min} \quad U(x)$$

Ingredients of our approach:

Standard cryptographic primitives are assumed

Importantly: Clocks and synchronization

- *Time and synchronization* are essential ingredients
- Without *common* time, nodes cannot cooperate temporally, share resources
- Cooperative scheduling, etc., will be impossible

Fundamental Result:

Theorem:

There exists a protocol that enables to form a reliable network from start to fully functioning. Moreover, the policies of bad nodes can be reduced to one of two actions: cooperate or jam, which no protocol can prevent.

Outline

1. Science of Information

2. **Center Mission**

   Integrated Research

   Research Accomplishments:

   Structural (Graphical) Compression

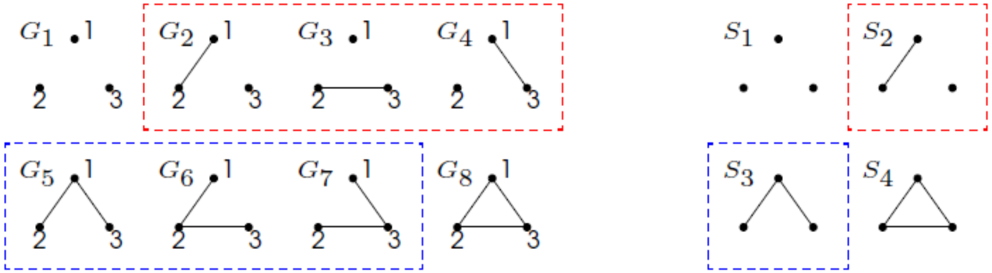
   (Szpankowski, Grama, Subramaniam – Purdue & UCSD)

3. Management
4. Budget

Structural Entropy on Graphs

Information Content of Unlabeled Graphs:

A structure model S of a graph G is defined for an unlabeled version.
Some labeled graphs have the same structure.



Graph Entropy vs Structural Entropy:

The probability of a structure S is: $P(S) = N(S) \cdot P(G)$

where $N(S)$ is the number of different labeled graphs having the same structure.

$$H_{\mathcal{G}} = \mathbf{E}[-\log P(G)] = -\sum_{G \in \mathcal{G}} P(G) \log P(G), \quad \text{graph entropy}$$

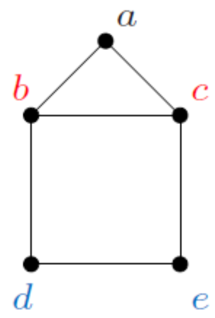
$$H_{\mathcal{S}} = \mathbf{E}[-\log P(S)] = -\sum_{S \in \mathcal{S}} P(S) \log P(S) \quad \text{structural entropy}$$

Relationship between H_G and H_S

Two labeled graphs G_1 and G_2 are called *isomorphic* if and only if there is a one-to-one mapping from $V(G_1)$ onto $V(G_2)$ which preserves the adjacency.

Graph Automorphism: For a graph G its automorphism is adjacency preserving permutation of vertices of G.

The collection $\mathrm{Aut}(G)$ of all automorphism of G is called *the automorphism group* of G.



Lemma 1. *If all isomorphic graphs have the same probability, then*

$$H_S = H_G - \log n! + \sum_{S \in \mathcal{S}} P(S) \log |\mathrm{Aut}(S)|,$$

where $\mathrm{Aut}(S)$ is the automorphism group of S.

Proof idea: Using the fact that

$$N(S) = \frac{n!}{|\mathrm{Aut}(S)|}.$$

Structural Entropy

Consider Erdös-Rényi graphs G(n,p).

Theorem 1 (Choi, W.S 2009). *For large n and all p satisfying $\frac{\ln n}{n} \ll p$ and $1-p \gg \frac{\ln n}{n}$ (i.e., the graph is connected w.h.p.),*

$$H_S = \binom{n}{2} h(p) - \log n! + O\left(\frac{\log n}{n^a}\right) = \binom{n}{2} h(p) - n \log n + n \log e + O(\log n), \quad a > 1$$

where $h(p) = -p \log p - (1-p) \log(1-p)$ *is the entropy rate.*

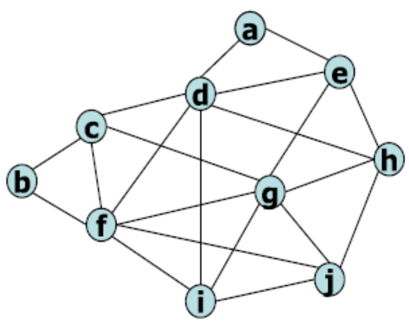
AEP *for structures:* $2^{-\binom{n}{2}(h(p)+\varepsilon)+\log n!} \leq P(S) \leq 2^{-\binom{n}{2}(h(p)-\varepsilon)+\log n!}.$

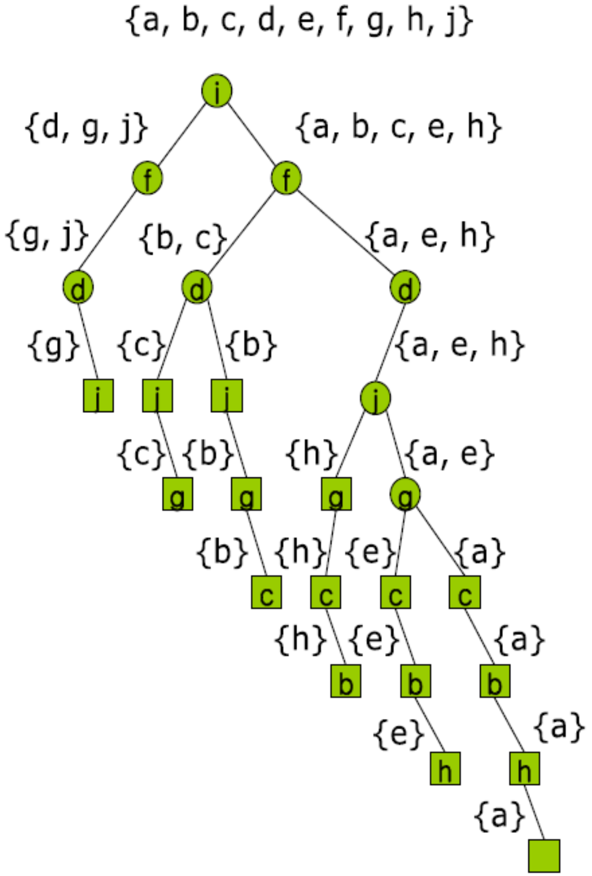
Proof idea:

1. $H_S = H_{\mathcal{G}} - \log n! + \sum_{S \in \mathcal{S}} P(S) \log |\mathrm{Aut}(S)|$.
2. $H_{\mathcal{G}} = \binom{n}{2} h(p)$
3. $\sum_{S \in \mathcal{S}} P(S) \log |\mathrm{Aut}(S)| = o(1)$ by asymmetry of $\mathcal{G}(n,p)$.

Structural Zip (SZIP) Algorithm

Compression Algorithm called Structural zip in short SZIP – Demo





B1 = 0100110100001110101

B2 = 100101100000101

Outline

1. Science of Information

2. **Center Mission**

   Integrated Research

   Research Accomplishments:

   Pathways in Biological Regulatory Networks

   (Grama, Mohammadi, Subramaniam – Purdue & UCSD)

3. Management
4. Budget

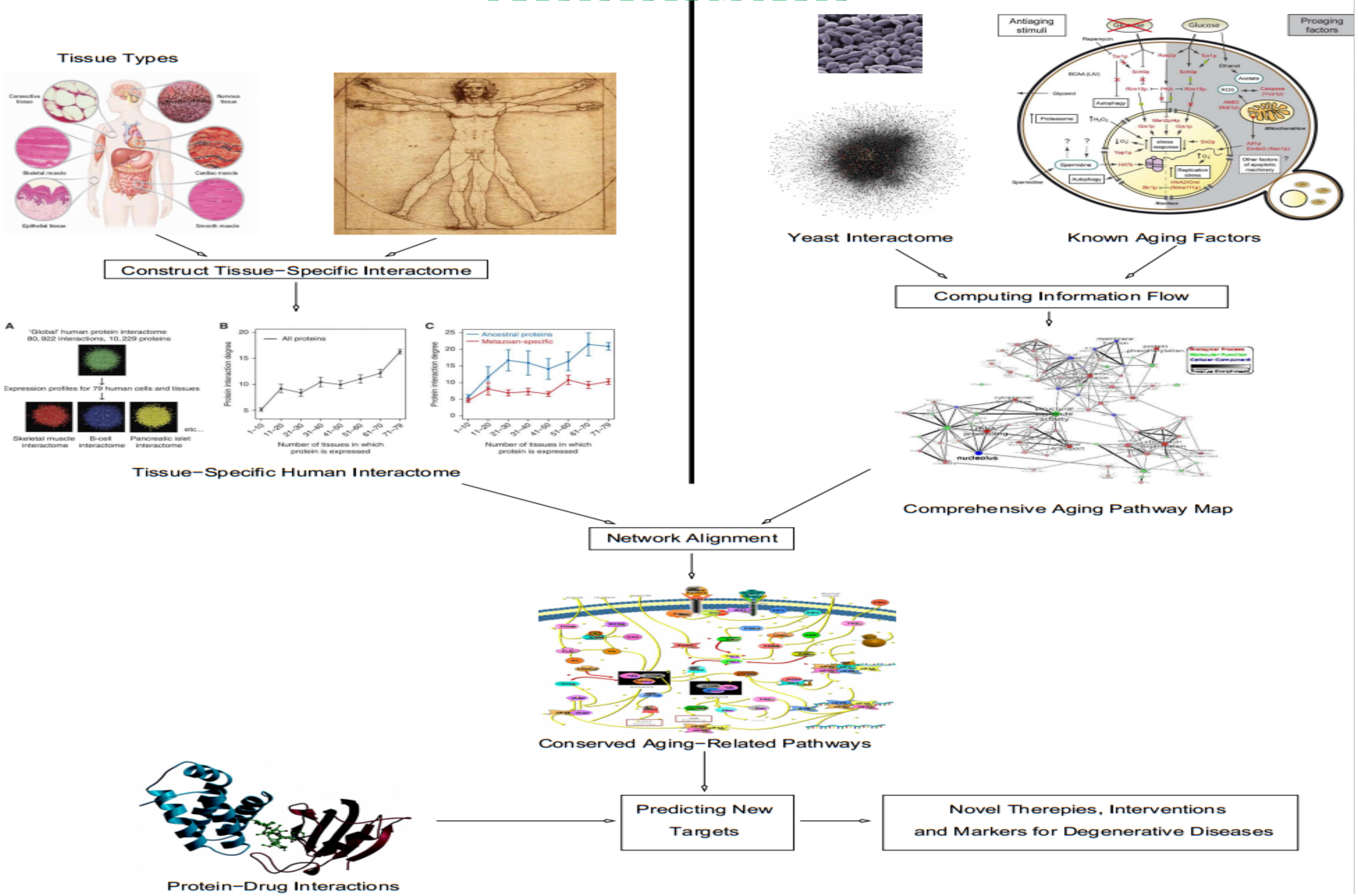
Decoding Network Footprints

- We have initiated an ambitions computational effort aimed at constructing the network footprint of degenerative diseases (Alzheimers, Parkinsons, Cancers).

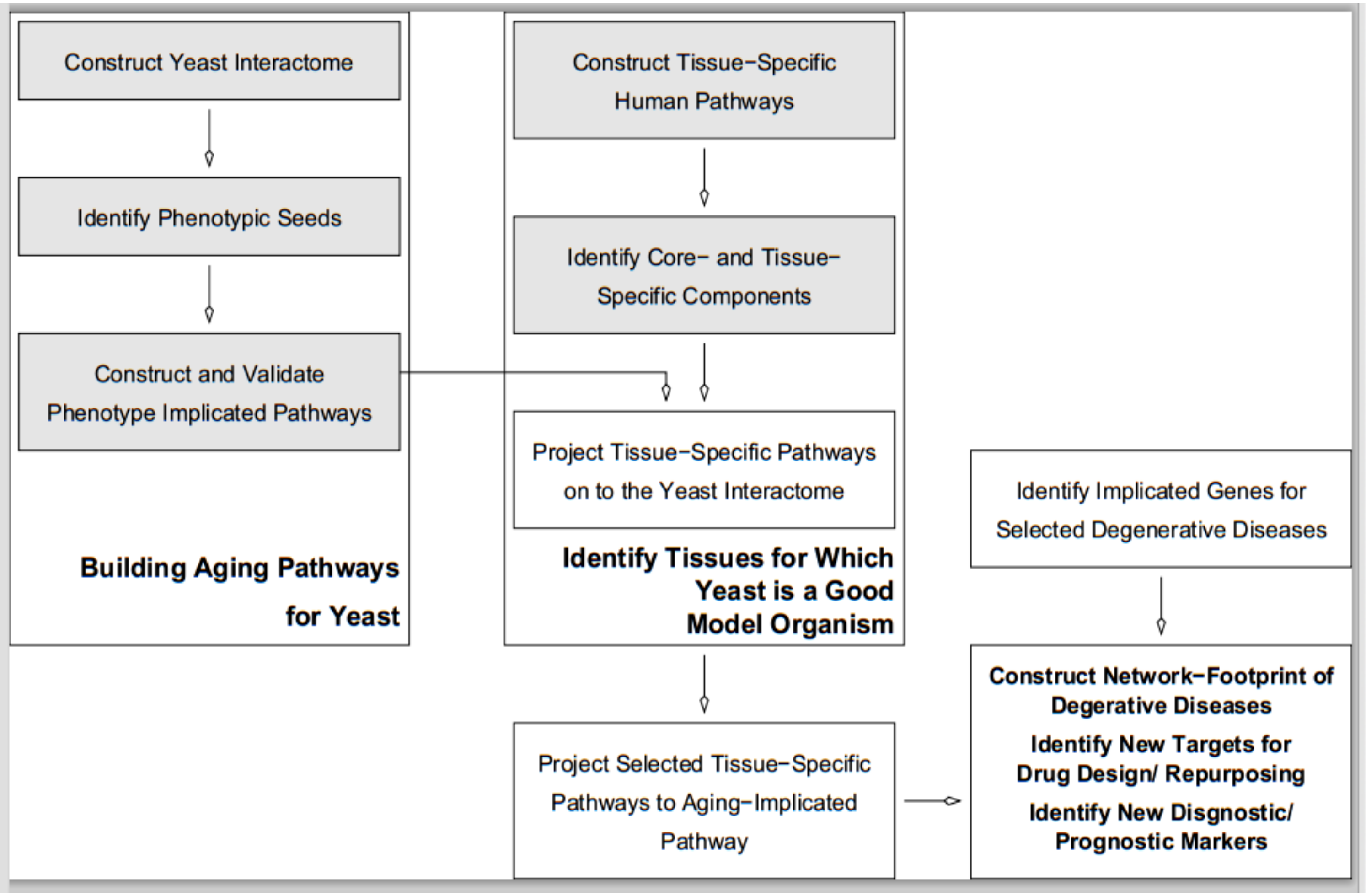
- Understanding pathways implicated in degenerative diseases holds the potential for novel interventions, drug design/ repurposing, and diagnostic/ prognostic markers.

- Using rigorous random-walk techniques, TOR complexes signaling are reconstructed – this identifies temporal and spatial aspects of cell growth!

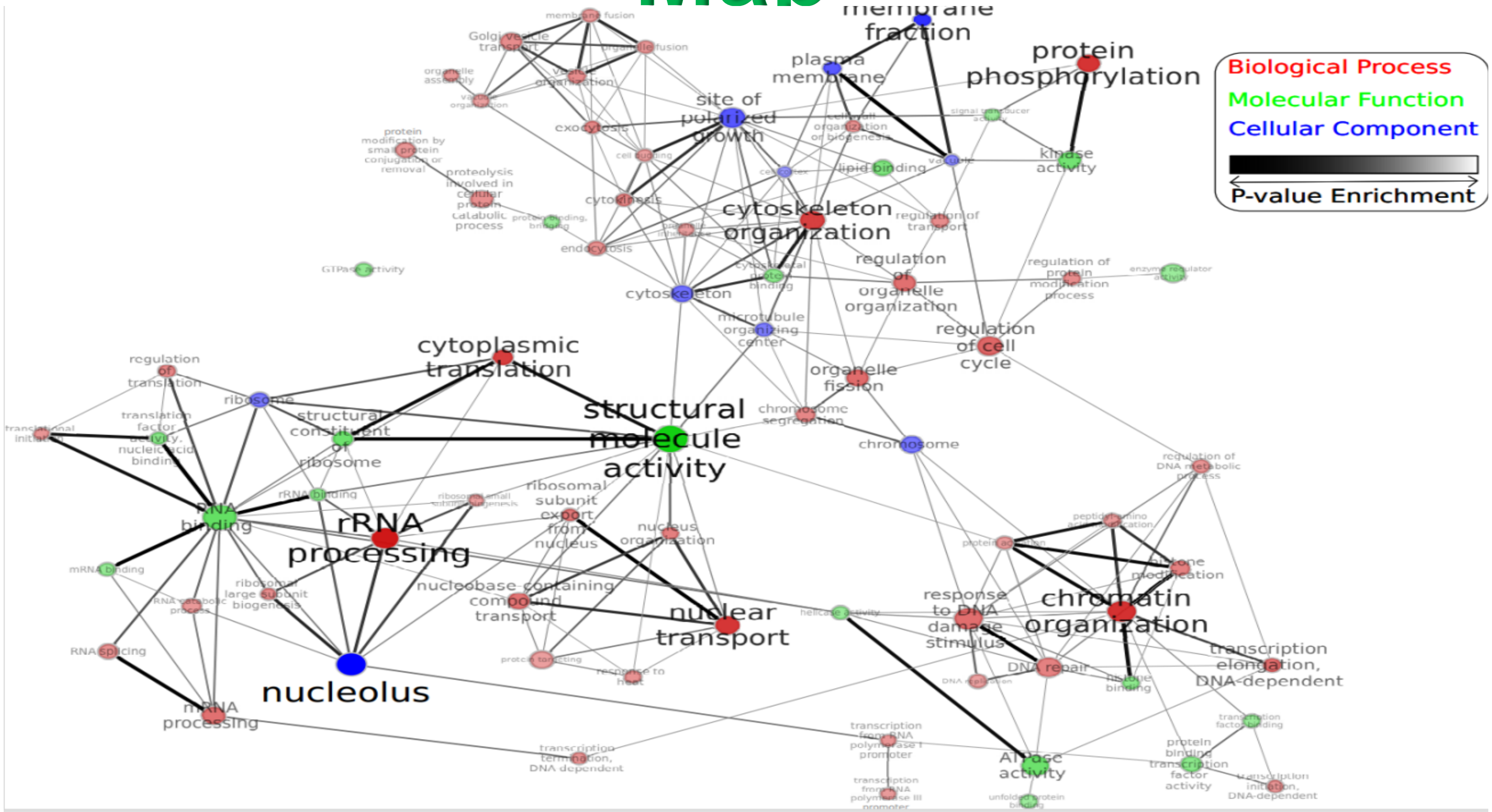
Network-Guided Characterization of Phenotype



Aging and Degenerative Diseases



Computational Derived Aging Map



This map includes various forms of interactions, including protein interactions, gene regulations, post translational modifications, etc. The underlying structural information extraction technique relies on rigorous quantification of information flow. It allows to understand temporal and spatial aspect of cell growth.

Really New Collaboration!

Prestige Lecture Series