The Science of Information: Big Data Analytics and Machine Learning

Shan Suthaharan
University of North Carolina at Greensboro

UNCG Course Code CSC495/CSC693

The development and delivery of the course is funded by the Center for the Science of Information, Purdue University through a sub-award approved by the National Science Foundation, and partially funded by UNCG.
From: [http://its.uncg.edu/telelearning/](http://its.uncg.edu/telelearning/)
Thanks to: Lane Ridenhour, Telelearning Center, UNCG

UNCG: 14 Undergraduate students and 10 Graduate students

UNCC and WCU: Expected to join
Why do we need this course?
Which one is Big?

Small

Big

Photo: by Praveen Suthaharan at the San Diego Zoo – August 2014

Photo: Samantha Henneke on flickr
Creative Commons License
What is Big Data?

The following link states: “African elephants are actually terrified of ants.”

http://www.dailymail.co.uk/sciencetech/article-1308415/Elephants-NOT-afraid-mice-terrified-ants.html
What is Big Data?

Today, there are an estimated 450,000 - 700,000 African elephants and between 35,000 - 40,000 wild Asian elephants.

- http://www.defenders.org/elephant/basic-facts

Scientists estimate that there are one quadrillion (1,000,000,000,000,000) ants living on the earth at any given time.


That is about 1351351351.3513513513513513513514 many ants per an elephant.

That is about 1.35 billion ants per an elephant.
What is Big Data?

Where is ant George?

Photo: Axel Rouvin on flickr - Creative Commons License
Intrusion Dataset
Topics

• Introduction
  – Conceptualization
  – Summarization

• Understanding of Data and Big Data
  – Data sets selection and analytics
  – Scalability and report writing

• Understanding of Computing Environment
  – Hadoop and MapReduce
  – Programming and Scikit-Learn
Topics

• Understanding of Machine Learning
  – Training, Validation and Testing
  – Support Vector Machine
  – Decision Trees and Random Forest
  – Deep Learning

• Scaling Up Machine Learning
  – PCA and Feature Hashing
  – SGD and Big Data Models
Study Planner
FLaSKU
Flexible Learning and Sequential Knowledge Update
Study Guide
Study Materials