Data Science: What is it and how to teach it?

Jeff Phillips, Ph.D. - Professor, University of Utah

Greg Jones, Ph.D., M.B.A - CEO, Revelat Analytics

Han-Wei Shen, Ph.D. - Professor, The Ohio State University

Chris Johnson, Ph.D. - Director, Scientific Computing and Imaging (SCI) Institute, University of Utah
Every two days we create as much data as we did from the beginning of mankind until 2003!
Big Data

Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...

Dan Ariely
Call for Papers
2015 IEEE International Conference on Big Data (IEEE Big Data 2015)
http://cci.drexel.edu/bigdata/bigdata2015/
Oct 29-Nov 1 2015, Santa Clara, CA, USA

• Big Data Science and Foundations
  a. Novel Theoretical Models for Big Data
  b. New Computational Models for Big Data
  c. Data and Information Quality for Big Data
  d. New Data Standards
Additional Topics

- Big Data Infrastructure
  a. Cloud/Grid/Stream Computing for Big Data
  b. High Performance/Parallel Computing Platforms for Big Data
  c. Autonomic Computing and Cyber-infrastructure, System Architectures, Design and Deployment
  d. Energy-efficient Computing for Big Data
  e. Programming Models and Environments for Cluster, Cloud, and Grid Computing to Support Big Data
  f. Software Techniques and Architectures in Cloud/Grid/Stream Computing
  g. Big Data Open Platforms
  h. New Programming Models for Big Data beyond Hadoop/MapReduce, STORM
  i. Software Systems to Support Big Data Computing

3. Big Data Management
   a. Search and Mining of variety of data including scientific and engineering, social, sensor/IoT/IoE, and multimedia data
   b. Algorithms and Systems for Big Data Search
   c. Distributed, and Peer-to-peer Search
   d. Big Data Search Architectures, Scalability and Efficiency
   e. Data Acquisition, Integration, Cleaning, and Best Practices
   f. Visualization Analytics for Big Data
   g. Computational Modeling and Data Integration
   h. Large-scale Recommendation Systems and Social Media Systems
   i. Cloud/Grid/Stream Data Mining- Big Velocity Data
   j. Link and Graph Mining
   k. Semantic-based Data Mining and Data Pre-processing
   l. Mobility and Big Data
   m. Multimedia and Multi-structured Data- Big Variety Data
4. Big Data Search and Mining
   a. Social Web Search and Mining
   b. Web Search
   c. Algorithms and Systems for Big Data Search
   d. Distributed, and Peer-to-peer Search
   e. Big Data Search Architectures, Scalability and Efficiency
   f. Data Acquisition, Integration, Cleaning, and Best Practices
   g. Visualization Analytics for Big Data
   h. Computational Modeling and Data Integration
   i. Large-scale Recommendation Systems and Social Media Systems
   j. Cloud/Grid/Stream Data Mining - Big Velocity Data
   k. Link and Graph Mining
   l. Semantic-based Data Mining and Data Pre-processing
   m. Mobility and Big Data
   n. Multimedia and Multi-structured Data - Big Variety Data

5. Big Data Security & Privacy
   a. Intrusion Detection for Gigabit Networks
   b. Anomaly and APT Detection in Very Large Scale Systems
   c. High Performance Cryptography
   d. Visualizing Large Scale Security Data
   e. Threat Detection using Big Data Analytics
   f. Privacy Threats of Big Data
   g. Privacy Preserving Big Data Collection/Analytics
   h. HCI Challenges for Big Data Security & Privacy
   i. User Studies for any of the above
   j. Sociological Aspects of Big Data Privacy

6. Big Data Applications
   a. Complex Big Data Applications in Science, Engineering, Medicine, Healthcare, Finance, Business, Law, Education, Transportation, Retailing, Telecommunication
   b. Big Data Analytics in Small Business Enterprises (SMEs),
   c. Big Data Analytics in Government, Public Sector and Society in General
   d. Real-life Case Studies of Value Creation through Big Data Analytics
   e. Big Data as a Service
   f. Big Data Industry Standards
   g. Experiences with Big Data Project Deployments
New Visual Analysis Techniques
Data Science Programs

- [Link](http://analytics.ncsu.edu/?page_id=4184)
- 19 MS programs in Data Analytics
- 8 MS programs in Data Science
- 28 MS programs in Business Analytics
- Several additional “tracks” or “concentration” programs
Report URLs


Computing at U.Utah View

www.cs.utah.edu/bigdata

**Fundamentals**
- Databases
- Algorithms
- Data Mining
- Machine Learning
- Visualization

**Advanced Topics**
- Distributed Systems
- Networking
- Randomized Algorithms
- Massive Data & Streaming
- Bayesian Modeling
- Natural Language Processing
- Information Retrieval
- Structured Prediction
- Visual Analytics

**Applications & Research**
- Scientific Computing
- Image Analysis
- Social Science
- Simulated Science
- Advertising Science
- Energy Engineering
Co-Authored National Reports
Computational Science: Ensuring America’s Competitiveness

President’s Information Technology Advisory Committee

June 2005

PITAC Report: www.nitrd.gov
Findings

- Traditional disciplinary boundaries within academia and Federal R&D agencies severely inhibit the development of effective research and education in computational science
- The paucity of incentives for longer-term multidisciplinary, multi-agency, or multi-sector efforts stifles structural innovation
Recommendations for Academia

• *Universities must significantly change their organizational structures* to promote and reward collaborative research

• *Universities must implement new multidisciplinary structures* to provide rigorous, multifaceted educational preparation for the growing ranks of computational scientists that the Nation will need to remain at the forefront of scientific discovery
Principal Finding

• Computational science is indispensable for solving complex problems in every sector, from traditional science and engineering domains to such key areas as national security, public health, and economic innovation.

• Advances in computing and connectivity and ability to capture and analyze huge amounts of data make it increasingly possible and practical to address these complex problems.

• *Universities and Federal government have not effectively recognized the strategic significance of computational science.*

• *These inadequacies compromise U.S. scientific leadership, economic competitiveness, and national security.*
Principal Recommendation

• Universities and Federal R&D agencies must *make coordinated, fundamental, and structural changes* that affirm the integral role of computational science
  – the most important problems are multidisciplinary, multi-agency, multi-sector, and collaborative

• The Federal government, in partnership with academia and industry, must also *create and execute a multi-decade roadmap* that directs coordinated advances in computational science and its applications in science and engineering
Recommendation of NSF Advisory Committee on Cyberinfrastructure (ACCI)

"The National Science Foundation should create a program in Computational and Data-Enabled Science and Engineering (CDS&E), based in and coordinated by the NSF Office of Cyberinfrastructure. The new program should be collaborative with relevant disciplinary programs in other NSF directorates and offices."

NSF can make a strong statement that will lead the Foundation, researchers it funds, and US universities and colleges generally, by recognizing Computational and Data-Enabled Science and Engineering as the distinct discipline it has clearly become.

Approved
Arden L. Bement, Jr.
Director
National Science Foundation

Date
05/27/2010
www.sci.utah.edu
crj@sci.utah.edu