NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

TRAINING STUDENTS TO EXTRACT VALUE FROM BIG DATA

A National Research Council Workshop
Sponsored by the National Science Foundation

COMMITTEE ON APPLIED AND THEORETICAL STATISTICS

APRIL 11-12, 2014

500 5TH STREET NW
WASHINGTON, DC

Workshop Objectives

This workshop on training undergraduate and graduate students to extract value from big data was designed to enable participants to share experience and perspectives on the following topics:

- What current knowledge and skills are needed by big data users in industry, government, and academia?
- What will students need to know to be successful using big data in the future (5-10 years out)?
- How could curriculum and training evolve to better prepare student for big data at the undergraduate and graduate levels?
- What options exist for providing the necessary interdisciplinary training within typical academic structures?
- What computational and data resources do colleges and universities need in order to provide useful training? What are some options for assembling that infrastructure?

Electronic Material

- All available workshop slides can be found here: http://sites.nationalacademies.org/deps/bmsa/deps_087192
- All videos of the workshop can be found here: http://vimeo.com/album/2861203

Day 1: Friday, April 11th

8:30 – 8:35 am	**Welcome**

Constantine Gatsonis, Brown University and CATS chair (slides available)
Video Available: http://vimeo.com/album/2861203/video/94383181

John Lafferty, University of Chicago, CATS member, and workshop planning co-chair
Video Available: http://vimeo.com/album/2861203/video/94383186

8:35 – 8:45 am	**Opening Remarks**

Suzanne Iacono, Deputy Assistant Director of the Directorate for Computer and Information Science and Engineering, National Science Foundation
Video Available: http://vimeo.com/album/2861203/video/94383183

8:45 – 10:00 am	**The Need for Training: Experiences and Case Studies**

Session chairs: Raghu Ramakrishnan, Microsoft Corporation, and John Lafferty, University of Chicago

Big Data - What is it? Why is it important? How should we train for it?
Raghu Ramakrishnan, Microsoft Corporation, CATS member, and workshop planning co-chair (slides available)
Video Available: http://vimeo.com/album/2861203/video/94383188

Training Students to do Good with Data
Rayid Ghani, University of Chicago (slides available)
Video Available: http://vimeo.com/album/2861203/video/94383189

The Need for Training in Big Data: Experiences and Case Studies
Guy Lebanon, Amazon Corporation (slides available)
Video Available: http://vimeo.com/album/2861203/video/94384884

10:10 – 10:15 am	**Break**

10:15 am– 12:45 pm	**Principles for Working with Big Data**

Session chair: Brian Caffo, Johns Hopkins University

Teaching about MapReduce
Jeffrey Ullman, Stanford University (slides available)
Video Available: http://vimeo.com/album/2861203/video/94384885

Big Data Machine Learning; Principles for Industry
Alexander Gray, Skytree Corporation (slides available)
Video Available: http://vimeo.com/album/2861203/video/94384887

Principles for Data Science Process
Duncan Temple Lang, University of California, Davis (slides available)
Video Available: http://vimeo.com/album/2861203/video/94384889

Principles for Working with Big Data
Juliana Freire, New York University (slides available)
Video Available: http://vimeo.com/album/2861203/video/94384893

12:45 – 1:45 pm	**Lunch**

1:45 – 4:15 pm	**Courses, Curricula, and Interdisciplinary Programs**

Session chair: Jim Frew, University of California, Santa Barbara

Experience with a First MOOC on Data Science
Bill Howe, University of Washington (slides available)
Video Available: http://vimeo.com/album/2861203/video/95192194

Data Science and Analytics Curriculum Development Rensselear (and the Tetherless World Constellation)
Peter Fox, Rensselear Polytechnic Institution (slides available)
Video Available: http://vimeo.com/album/2861203/video/94389369

Computational Training and Data Literacy for Domain Scientists
Joshua Bloom, University of California, Santa Barbara (slides available)
Video Available: http://vimeo.com/album/2861203/video/94389370

4:15 – 4:30 pm	**Break**

4:30 – 5:30 pm	**Q&A / Discussion**

Panelists
Joshua Bloom, University of California, Santa Barbara
Peter Fox, Rensselear Polytechnic Institution
Jim Frew, University of California, Santa Barbara
Bill Howe, University of Washington
Video Available: http://vimeo.com/album/2861203/video/95196879

5:30 pm	**Adjourn**

Day 2: Saturday, April 12th

8:30 – 11:00 am	**Shared Resources**

Session chair: Deepak Agarwal, LinkedIn

Can Knowledge Bases Help Accelerate Science?
Christopher Ré, Stanford University (slides available)
Video Available: http://vimeo.com/album/2861203/video/95196880

Divide and Recombine for Large Complex Data
Bill Cleveland, Perdue University (slides available)
Video Available: http://vimeo.com/album/2861203/video/95196882

Yahoo's Webscope Data Sharing Program
Ron Brachman, Yahoo (slides available)
Video Available: http://vimeo.com/album/2861203/video/95196883

Resource Sharing
Mark Ryland, Amazon (slides available)
Video Available: http://vimeo.com/album/2861203/video/95196884

11:00 – 11:15 am	**Break**

11:15 am – 1:00 pm	**Panel Discussion: Workshop Lessons**

Session chair: Rob Kass, Carnegie Mellon University and CATS member

Panelists
Deepak Agarwal, LinkedIn
Jim Frew, University of California, Santa Barbara
John Lafferty, University of Chicago, CATS member, and workshop planning co-chair
Claudia Perlich, New York University
Raghu Ramakrishnan, Microsoft Corporation, CATS member, and workshop planning co-chair
Video Available: http://vimeo.com/album/2861203/video/95199231

1:00 pm	**Adjourn**

Planning Committee

Co-Chairs

John Lafferty, University of Chicago and CATS member
Raghu Ramakrishnan, Microsoft Corporation and CATS member

Members

Deepak Agarwal, LinkedIn
Corinna Cortes, Google
Jeff Dozier, University of California, Santa Barbara
Robert Kass, Carnegie Mellon University and CATS member
Anna Gilbert, University of Michigan
Rafael Irizarri, Harvard University
Patrick Hanrahan, Stanford University
Prabhakar Raghavan, Google
Nathaniel Schenker, Centers for Disease Control and Prevention
Ion Stoica, University of California, Berkeley

Staff officer:
Neal Glassman
nglassman@nas.edu
202.334.1682