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 students 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](http://sites.nationalacademies.org/deps/bmsa/deps_087192)
- All videos of the workshop can be found here: [http://vimeo.com/album/2861203](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)

John Lafferty, University of Chicago, CATS member, and workshop planning co-chair

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

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)

Training Students to do Good with Data
Rayid Ghani, University of Chicago (slides available)

The Need for Training in Big Data: Experiences and Case Studies
Guy Lebanon, Amazon Corporation (slides available)

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)
### Big Data Machine Learning; Principles for Industry
Alexander Gray, Skytree Corporation (slides available)

### Principles for Data Science Process
Duncan Temple Lang, University of California, Davis (slides available)

### Principles for Working with Big Data
Juliana Freire, New York University (slides available)

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>12:45 – 1:45 pm</td>
<td>Lunch</td>
</tr>
<tr>
<td>1:45 – 4:15 pm</td>
<td>Courses, Curricula, and Interdisciplinary Programs</td>
</tr>
<tr>
<td></td>
<td><strong>Session chair:</strong> Jim Frew, University of California, Santa Barbara</td>
</tr>
<tr>
<td></td>
<td><strong>Experience with a First MOOC on Data Science</strong></td>
</tr>
<tr>
<td></td>
<td>Bill Howe, University of Washington (slides available)</td>
</tr>
<tr>
<td></td>
<td><strong>Data Science and Analytics Curriculum Development Rensselaer (and the Tetherless World Constellation)</strong></td>
</tr>
<tr>
<td></td>
<td>Peter Fox, Rensselaer Polytechnic Institution (slides available)</td>
</tr>
<tr>
<td></td>
<td><strong>Computational Training and Data Literacy for Domain Scientists</strong></td>
</tr>
<tr>
<td></td>
<td>Joshua Bloom, University of California, Santa Barbara (slides available)</td>
</tr>
<tr>
<td>4:15 – 4:30 pm</td>
<td>Break</td>
</tr>
<tr>
<td>4:30 – 5:30 pm</td>
<td>Q&amp;A / Discussion</td>
</tr>
<tr>
<td></td>
<td><strong>Panelists</strong></td>
</tr>
<tr>
<td></td>
<td>Joshua Bloom, University of California, Santa Barbara</td>
</tr>
<tr>
<td></td>
<td>Peter Fox, Rensselaer Polytechnic Institution</td>
</tr>
<tr>
<td></td>
<td>Jim Frew, University of California, Santa Barbara</td>
</tr>
<tr>
<td></td>
<td>Bill Howe, University of Washington</td>
</tr>
<tr>
<td>5:30 pm</td>
<td>Adjourn</td>
</tr>
</tbody>
</table>
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)

**Divide and Recombine for Large Complex Data**
Bill Cleveland, Perdue University (slides available)

**Yahoo’s Webscope Data Sharing Program**
Ron Brachman, Yahoo (slides available)

**Resource Sharing**
Mark Ryland, Amazon (slides available)

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

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