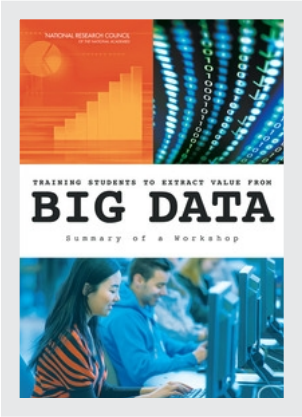


# Training Students to Extract Value from Big Data: Summary of a Workshop



## DETAILS

66 pages | 7 x 10 | PAPERBACK
ISBN 978-0-309-31437-4 | DOI 10.17226/18981

## AUTHORS

Committee on Applied and Theoretical Statistics; Board on Mathematical Sciences and Their Applications; Division on Engineering and Physical Sciences; National Research Council

# TRAINING STUDENTS TO EXTRACT VALUE FROM BIG DATA

Summary of a Workshop

Maureen Mellody, *Rapporteur*

Committee on Applied and Theoretical Statistics

Board on Mathematical Sciences and Their Applications

Division on Engineering and Physical Sciences

NATIONAL RESEARCH COUNCIL
*OF THE NATIONAL ACADEMIES*

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
**www.nap.edu**

**THE NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001**

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine.

This study was supported by Grant DMS-1332693 between the National Academy of Sciences and the National Science Foundation. Any opinions, findings, or conclusions expressed in this publication are those of the author and do not necessarily reflect the views of the organizations or agencies that provided support for the project.

International Standard Book Number-13: 978-0-309-31437-4
International Standard Book Number-10: 0-309-31437-2

This report is available in limited quantities from:

Board on Mathematical Sciences and Their Applications
500 Fifth Street NW
Washington, DC 20001
bmsa@nas.edu
http://www.nas.edu/bmsa

Additional copies of this workshop summary are available for sale from the National Academies Press, 500 Fifth Street, NW, Keck 360, Washington, DC 20001; (800) 624-6242 or (202) 334-3313; http://www.nap.edu/.



Printed in the United States of America

# THE NATIONAL ACADEMIES

*Advisers to the Nation on Science, Engineering, and Medicine*

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Ralph J. Cicerone is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. C. D. Mote, Jr., is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Victor J. Dzau is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Ralph J. Cicerone and Dr. C. D. Mote, Jr., are chair and vice chair, respectively, of the National Research Council.

**www.national-academies.org**

**PLANNING COMMITTEE ON TRAINING STUDENTS TO EXTRACT VALUE FROM BIG DATA: A WORKSHOP**

JOHN LAFFERTY, University of Chicago, *Co-Chair*
RAGHU RAMAKRISHNAN, Microsoft Corporation, *Co-Chair*
DEEPAK AGARWAL, LinkedIn Corporation
CORINNA CORTES, Google, Inc.
JEFF DOZIER, University of California, Santa Barbara
ANNA GILBERT, University of Michigan
PATRICK HANRAHAN, Stanford University
RAFAEL IRIZARRI, Harvard University
ROBERT KASS, Carnegie Mellon University
PRABHAKAR RAGHAVAN, Google, Inc.
NATHANIEL SCHENKER, Centers for Disease Control and Prevention
ION STOICA, University of California, Berkeley

***Staff***

NEAL GLASSMAN, Senior Program Officer
SCOTT T. WEIDMAN, Board Director
MICHELLE K. SCHWALBE, Program Officer
RODNEY N. HOWARD, Administrative Assistant

**COMMITTEE ON APPLIED AND THEORETICAL STATISTICS**

CONSTANTINE GATSONIS, Brown University, *Chair*
MONTSERRAT (MONTSE) FUENTES, North Carolina State University
ALFRED O. HERO III, University of Michigan
DAVID M. HIGDON, Los Alamos National Laboratory
IAIN JOHNSTONE, Stanford University
ROBERT KASS, Carnegie Mellon University
JOHN LAFFERTY, University of Chicago
XIHONG LIN, Harvard University
SHARON-LISE T. NORMAND, Harvard University
GIOVANNI PARMIGIANI, Harvard University
RAGHU RAMAKRISHNAN, Microsoft Corporation
ERNEST SEGLIE, Office of the Secretary of Defense (retired)
LANCE WALLER, Emory University
EUGENE WONG, University of California, Berkeley

***Staff***

MICHELLE K. SCHWALBE, Director
RODNEY N. HOWARD, Administrative Assistant

# Acknowledgment of Reviewers

This report has been reviewed in draft form by persons chosen for their diverse perspectives and technical expertise in accordance with procedures approved by the National Research Council's Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards of objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process. We thank the following individuals for their review of this report:

Michael Franklin, University of California, Berkeley,
Johannes Gehrke, Cornell University,
Claudia Perlich, Dstillery, and
Duncan Temple Lang, University of California, Davis.

Although the reviewers listed above have provided many constructive comments and suggestions, they were not asked to endorse the views presented at the workshop, nor did they see the final draft of the workshop summary before its release. The review of this workshop summary was overseen by Anthony Tyson, University of California, Davis. Appointed by the National Research Council, he was responsible for making certain that an independent examination of the summary was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this summary rests entirely with the author and the institution.

# Contents

# 1

# Introduction

Data sets—whether in science and engineering, economics, health care, public policy, or business—have been growing rapidly; the recent National Research Council (NRC) report *Frontiers in Massive Data Analysis* documented the rise of "big data," as systems are routinely returning terabytes, petabytes, or more of information (National Research Council, 2013). Big data has become pervasive because of the availability of high-throughput data collection technologies, such as information-sensing mobile devices, remote sensing, radiofrequency identification readers, Internet log records, and wireless sensor networks. Science, engineering, and business have rapidly transitioned from the longstanding state of striving to develop information from scant data to a situation in which the challenge is now that the amount of information exceeds a human's ability to examine, let alone absorb, it. Web companies—such as Yahoo, Google, and Amazon—commonly work with data sets that consist of billions of items, and they are likely to increase by an order of magnitude or more as the Internet of Things[1] matures. In other words, the size and scale of data, which can be overwhelming today, are only increasing. In addition, data sets are increasingly complex, and this potentially increases the problems associated with such concerns as missing information and other quality concerns, data heterogeneity, and differing data formats.

[1] The Internet of Things is the network of uniquely identifiable physical objects embedded throughout a network structure, such as home appliances that can communicate for purposes of adjusting their settings, ordering replacement parts, and so on.

Advances in technology have made it easier to assemble and access large amounts of data. Now, a key challenge is to develop the experts needed to draw reliable inferences from all that information. The nation's ability to make use of the data depends heavily on the availability of a workforce that is properly trained and ready to tackle these high-need areas. A report from McKinsey & Company (Manyika et al., 2011) has predicted shortfalls of 150,000 data analysts and 1.5 million managers who are knowledgeable about data and their relevance. It is becoming increasingly important to increase the pool of qualified scientists and engineers who can extract value from big data. Training students to be capable in exploiting big data requires experience with statistical analysis, machine learning, and computational infrastructure that permits the real problems associated with massive data to be revealed and, ultimately, addressed. The availability of repositories (of both data and software) and computational infrastructure will be necessary to train the next generation of data scientists. Analysis of big data requires cross-disciplinary skills, including the ability to make modeling decisions while balancing trade-offs between optimization and approximation, all while being attentive to useful metrics and system robustness. To develop those skills in students, it is important to identify *whom* to teach, that is, the educational background, experience, and characteristics of a prospective data science student; *what* to teach, that is, the technical and practical content that should be taught to the student; and *how* to teach, that is, the structure and organization of a data science program.

The topic of training students in big data is timely, as universities are already experimenting with courses and programs tailored to the needs of students who will work with big data. Eight university programs have been or will be launched in 2014 alone.[2] The workshop that is the subject of this report was designed to enable participants to learn and benefit from emerging insights while innovation in education is still ongoing.

## WORKSHOP OVERVIEW

On April 11-12, 2014, the standing Committee on Applied and Theoretical Statistics (CATS) convened a workshop to discuss how best to train students to use big data. CATS is organized under the auspices of the NRC Board on Mathematical Sciences and Their Applications.

To conduct the workshop, a planning committee was first established to refine the topics, identify speakers, and plan the agenda. The workshop was held at the Keck Center of the National Academies in Washington, D.C., and was sponsored by the National Science Foundation (NSF). About 70 persons—including speakers,

[2] See the Master's in Data Science website at http://www.mastersindatascience.org/ for more information, accessed June 5, 2014.

members of the parent committee and board, invited guests, and members of the public—participated in the 2-day workshop. The workshop was also webcast live, and at least 175 persons participated remotely.

A complete statement of task is shown in Box 1.1. The workshop explored the following topics:

- The need for training in big data.
- Curricula and coursework, including suggestions at different instructional levels and suggestions for a core curriculum.
- Examples of successful courses and curricula.
- Identification of the principles that should be delivered, including sharing of resources.

Although the title of the workshop was "Training Students to Extract Value from Big Data," the term *big data* is not precisely defined. CATS, which initiated the workshop, has tended to use the term *massive data* in the past, which implies data on a scale for which standard tools are not adequate. The terms *data analytics* and *data science* are also becoming common. They seem to be broader, with a focus on using data—maybe of unprecedented scale, but maybe not—in new ways to inform decision making. This workshop was not developed to explore any particular one of these definitions or to develop definitions. But one impetus for the workshop

**BOX 1.1**
**Statement of Task**

An ad hoc committee will plan and conduct a public workshop on the subject of training undergraduate and graduate students to extract value from big data. The committee will develop the agenda, select and invite speakers and discussants, and moderate the discussions. The presentations and discussions at the workshop will be 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?

was the current fragmented view of what is meant by analysis of big data, data analytics, or data science. New graduate programs are introduced regularly, and they have their own notions of what is meant by those terms and, most important, of what students need to know to be proficient in data-intensive work. What are the core subjects in data science? By illustration, this workshop began to answer that question. It is clear that training in big data, data science, or data analytics requires a multidisciplinary foundation that includes at least computer science, machine learning, statistics, and mathematics, and that curricula should be developed with the active participation of at least these disciplines. The chapters of this summary provide a variety of perspectives about those elements and about their integration into courses and curricula.

Although the workshop summarized in this report aimed to span the major topics that students need to learn if they are to work successfully with big data, not everything could be covered. For example, tools that might supplant MapReduce, such as Spark, are likely to be important, as are advances in Deep Learning. Means by which humans can interact with and absorb huge amounts of information—such as visualization tools, iterative analysis, and human-in-the-loop systems—are critical. And such basic skills as data wrangling, cleaning, and integration will continue to be necessary for anyone working in data science. Educators who design courses and curricula must consider a wide array of skill requirements.

The present report has been prepared by the workshop rapporteur as a factual summary of what occurred at the workshop. The planning committee's role was limited to planning and convening the workshop. The views contained in the report are those of individual workshop participants and do not necessarily represent the views of all workshop participants, the planning committee, or the NRC.

## NATIONAL EFFORTS IN BIG DATA

*Suzanne Iacono, National Science Foundation*

Suzanne Iacono, of NSF, set the stage for the workshop by speaking about national efforts in big data, current challenges, and NSF's motivations for sponsoring the workshop. She explained that the workshop was an outgrowth of the national big data research and development (R&D) initiative. The federal government is interested in big data for three reasons:

- To stimulate commerce and the economy.
- To accelerate the pace of discovery and enable new activities.
- To address pressing national challenges in education, health care, and public safety.

Big data is of interest to the government now because of the confluence of technical, economic, and policy interests, according to Iacono. Advances in technology have led to a reduction in storage costs, so it is easier to retain data today. On the policy side, data are now considered to be assets, and government is pushing agencies to open data sets to the public. In other words, there has been a democratization of data use and tools.

Iacono described a recent book (Mayer-Schönberger and Cukier, 2012) that outlined three basic shifts in today's data:

- There are more data than ever before.
- Data are messy, and there must be an increased acceptance of imperfection.
- Correlations can help in making decisions.

She then described the national Big Data Research and Development Initiative in more detail. A 2010 report from the President's Council of Advisors on Science and Technology argued that the federal government was not investing sufficiently in big data research and development and that investment in this field would produce large returns. A working group in big data, under the interagency Networking and Information Technology Research and Development (NITRD) program and managed by the Office of Science and Technology Policy, was charged with establishing a framework for agency activity. The result was that in 2012, $200 million was allocated for big data R&D throughout the NITRD agencies, including the Defense Advanced Research Projects Agency (DARPA), the Department of Energy (DOE) Office of Science, the National Institutes of Health (NIH), and NSF. Iacono showed the framework for moving forward with big data R&D, which included the following elements:

- *Foundational research.* Iacono stressed that this research is critical because data are increasing and becoming more heterogeneous.
- *Cyberinfrastructure.* New and adequate infrastructure is needed to manage and curate data and serve them to the larger research community.
- *New approaches to workforce and education.*
- *New collaborations and outreach.*

Iacono noted that policy envelops all four elements of the framework.

A 2013 White House memorandum directed executive branch agencies to develop plans to increase public access to the results of federally funded research, including access to publications and data, and plans are under way at the agency level to address this memorandum. Iacono noted that increased access to publications is not difficult, because existing publication-access methods in professional societies and some government agencies can be used as models. She also pointed

out that NIH's PubMed[3] program may be a useful model in that it shares research papers. However, she noted that access to data will be much more difficult than access to publications because each discipline and community will have its own implementation plan and will treat data privacy, storage duration, and access differently.

Iacono described foundational R&D in more detail. She explained that NSF and NIH awarded 45 projects in big data in 2012 and 2013. About half were related to data collection and management and one-fourth to health and bioinformatics. The remaining awards were spread among social networks, physical sciences and engineering, algorithms, and cyberinfrastructure. Seventeen agencies are involved in the Big Data Senior Steering Group, and each is implementing programs of its own related to big data. For example, DARPA has implemented three new programs—Big Mechanism, Memex, and Big Data Capstone; the National Institute of Standards and Technology maintains a Big Data Working Group; DOE has an Extreme Scale Science initiative; and NSF and NIH each has a broad portfolio related to big data. Iacono stressed that big data is a national issue and that there is substantial interest now in industry and academe, so she believes that government should consider multistakeholder partnerships.

Iacono discussed three challenges related to big data:

- *Technology.* She emphasized that technology alone cannot solve big data problems, and she cited several recent popular books that discuss the folly of technological solutionism (Mayer-Schönberger and Cukier, 2012; Mele, 2013; Reese, 2013; Schmidt and Cohen, 2013; Surdak, 2014; Webb, 2013).
- *Privacy.* Iacono pointed out that many of our behaviors—including shopping, searching, and social interactions—are now tracked, and she noted that a White House 90-day review to examine the privacy implications of big data was under way.[4] In general, Iacono noted the importance of regulating data use, as opposed to data collection; balancing interests; and promoting data sharing.
- *Education and workforce.* As noted above, the 2011 report from the McKinsey & Company predicted large shortfalls of big data experts. Iacono noted that the *Harvard Business Review* labeled data science as "the sexiest job of the 21st century" (Davenport and Patil, 2012). The *New York Times* has recently hired a chief data scientist. The bottom line, Iacono explained,

[3] See the National Center for Biotechnology Information's PubMed database at http://www.ncbi.nlm.nih.gov/pubmed (accessed May 25, 2014) for more information.

[4] That study has since been completed and can be found at Executive Office of the President, *Big Data: Seizing Opportunities, Preserving Values,* May 2014, http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_ report_may_1_2014.pdf.

is that the talent pool in data science must be expanded to meet current and future needs.

Iacono pointed out that there are traditional ways to educate students through school curricula but that there are also other ways to learn. Such companies as DataKind and Pivotal are matching data scientists with data problems in the nonprofit community. Universities, such as the University of Chicago, as discussed by Rayid Ghani (see Chapter 2), are also working to connect data scientists to problems of social good. Iacono concluded by stressing the many opportunities and challenges in big data that lie ahead.

## ORGANIZATION OF THIS REPORT

The remaining chapters of this report summarize the workshop presentations and discussions. To assist the reader, each chapter begins with a short list of important statements made by speakers during the workshop session. Chapter 2 outlines the need for training. Chapter 3 discusses some of the principles of working with big data. Chapter 4 focuses on courses and curricula needed to support the use of big data. Chapter 5 discusses shared resources, and Chapter 6 summarizes the group discussion of lessons learned from the workshop. Finally, Appendix A lists the workshop participants, Appendix B shows the workshop agenda, and Appendix C defines acronyms used in this report.

# 2

# The Need for Training: Experiences and Case Studies

**Important Points Made by Individual Speakers**

- Students often do not recognize that big data techniques can be used to solve problems that address societal good, such as those in education, health, and public policy; educational programs that foster relationships between data science and social problems have the potential to increase the number and types of students interested in data science. (Rayid Ghani)
- There may be a mismatch between some industry needs and related academic pursuits: current studies of recommendation systems, such as off-line score prediction, do not always correlate well with important industry metrics, such as sales and user engagement. (Guy Lebanon)
- Academia does not have sufficient access to practical data scenarios in industry. (Guy Lebanon)

Big data is becoming pervasive in industry, government, and academe. The disciplines that are affected are as diverse as meteorology, Internet commerce, genomics, complex physics simulations, health informatics, and biologic and environmental research. The second session of the workshop focused on specific examples and case studies of real-world needs in big data. The session was co-chaired by John Lafferty (University of Chicago) and Raghu Ramakrishnan (Microsoft Corporation), the co-chairs of the workshop's organizing committee. Presentations were made in Session 2 by Rayid Ghani (University of Chicago) and Guy Lebanon (Amazon Corporation).

## TRAINING STUDENTS TO DO GOOD WITH BIG DATA

*Rayid Ghani, University of Chicago*

Rayid Ghani explained that he has founded a summer program at the University of Chicago, known as the Data Science for Social Good Fellowship, to show students that they can apply their talents in data science to societal problems and in so doing affect many lives. He expressed his growing concern that the top technical students are disproportionately attracted to for-profit companies, such as Yahoo and Google, and posited that these students do not recognize that solutions to problems in education, health, and public policy also need data.

Ghani showed a promotional video for the University of Chicago summer program and described its applicant pool. Typically, half the applicants are computer science or machine learning students; one-fourth are students in social science, public policy, or economics; and one-fourth are students in statistics. Some 35 percent of the enrolled students are female (as Ghani pointed out, this is a larger proportion than is typical of a computer science graduate program). Many of the applicants are graduate students, and about 25 percent are undergraduate seniors. The program is competitive: in 2013, there were 550 applicants for 36 spots. Ghani hypothesized that the program would be appropriate for someone who had an affinity for mathematics and science but a core interest in helping others. Once in the program, students are matched with mentors, most of whom are computer scientists or economists with a strong background in industry.

He explained that the program is project-based, using real-world problems from government and nonprofit organizations. Each project includes an initial mapping of a societal problem to a technical problem and communication back to the agency or organization about what was learned. Ghani stressed that students need to have skills in communication and common sense in addition to technical expertise. The curriculum at the University of Chicago is built around tools, methods, and problem-solving skills. The program now consistently uses the Python language, and it also teaches database methods. Ghani emphasized the need to help students to learn new tools and techniques. He noted, for instance, that some of the students knew of regression only as a means of evaluating data whereas other tools may be more suitable for massive data.

Ghani described a sample project from the program. A school district in Arizona was experiencing undermatching—that is, students have the potential to go to college but do not, or students have the potential to go to a more competitive college than the one they ultimately select. The school district had collected several years of data. In a summer project, the University of Chicago program students built models to predict who would graduate from college, who would go to college,

and who was not likely to apply. In response to the data analysis, the school district has begun a targeted career-counseling program to begin intervention.

## THE NEED FOR TRAINING IN BIG DATA: EXPERIENCES AND CASE STUDIES

*Guy Lebanon, Amazon Corporation*

Guy Lebanon began by stating that extracting meaning from big data requires skills of three kinds: computing and software engineering; machine learning, statistics, and optimization; and product sense and careful experimentation. He stressed that it is difficult to find people who have expertise and skills in all three and that competition for such people is fierce.

Lebanon then provided a case study in recommendation systems. He pointed out that recommendation systems (recommending movies, products, music, advertisements, and friends) are important for industry. He described a well-known method of making recommendations known as matrix completion. In this method, an incomplete user rating matrix is completed to make predictions. The matrix completion method favors low-rank (simple) completions. The best model is found by using a nonlinear optimization procedure in a high-dimensional space. The concept is not complex, but Lebanon indicated that its implementation can be difficult. Implementation requires knowledge of the three kinds referred to earlier. Specifically, Lebanon noted the following challenges:

- *Computing and software engineering:* language skills (usually C++ or Java), data acquisition, data processing (including parallel and distributed computing), knowledge of software engineering practices (such as version control, code documentation, building tools, unit tests, and integration tests), efficiency, and communication among software services.
- *Machine learning:* nonlinear optimization and implementation (such as stochastic gradient descent), practical methods (such as momentum and step selection size), and common machine learning issues (such as overfitting).
- *Product sense:* an online evaluation process to measure business goals; model training; and decisions regarding history usage, product modification, and product omissions.

Lebanon described two problems that limit academic research in recommendation systems, both related to overlooking metrics that are important to industry. First, accuracy in academic, off-line score prediction does not correlate with

important industry metrics, such as sales and increased user engagement. Second, academe does not have sufficient access to practical data scenarios from industry. Lebanon posited that academe cannot drive innovation in recommendation systems; research in recommendation systems does not always translate well to the real world, and prediction accuracy is incorrectly assumed to be equivalent to business goals.

He then described a challenge run by Netflix. In the early 2000s, Netflix held a competition to develop an improved recommendation system. It provided a data set of ratings that had been anonymized and offered a $1 million prize to the top team. The competition created a boost in research, which saw a corresponding increase in research papers and overall interest. However, a group of researchers at the University of Texas, Austin, successfully deanonymized the Netflix data by joining them with other data. Netflix later withdrew the data set and is now facing a lawsuit. As a result of that experience, industry is increasingly wary about releasing any data for fear of inadvertently exposing private or proprietary data, but this makes it difficult for academe to conduct relevant and timely research.

Lebanon pointed out that the important result in a recommendation system is prediction of a user's reaction to a specific recommendation. For it to be successful, one needs to know the context in which the user acts—for instance, time and location information—but that context is not conveyed in an anonymized data set. In addition, methods that perform well on training and test data sets do not perform well in real environments when a user makes a single A/B comparison.[1] Lebanon proposed several new ideas to address those characteristics:

- Study the correlations between existing evaluation methods and increased user engagement in an A/B test.
- Develop new off-line evaluations to account for user context better.
- Develop efficient searches among the possibilities to maximize A/B test performance.

Few data sets are publicly available, according to Lebanon. Working with limited data, the research community may focus on minor improvements in incremental steps, not substantial improvements that are related to the additional contextual information that is available to the owners of the data, the companies. He pointed out that real-world information and context, such as user addresses and other profile information, could potentially be incorporated into a traditional recommendation system.

[1] In A/B testing, more formally known as two-sample hypothesis testing, two variants are presented to a user, and the user determines a winner.

Lebanon concluded with a brief discussion of implicit ratings. In the real world, one often has implicit, binary-rating data, such as whether a purchase or an impression was made. Evaluating that type of binary-rating data requires a different set of tools and models, and scaling up from standard data sets to industry data sets remains challenging.

# 3

# Principles for Working with Big Data

**Important Points Made by Individual Speakers**

- MapReduce is an important programming method designed for easy parallel programming on commodity hardware. (Jeffrey Ullman)
- There is an expertise gap between domain scientists and data scientists: domain scientists do not know what is possible technically, and data scientists do not understand the domain. (Juliana Freire)
- A data scientist should have expertise in databases, machine learning and statistics, and visualization; it is challenging, and perhaps unrealistic, to find people who have expertise in all three. (Juliana Freire and other discussion participants)
- Data preparation is an important, time-consuming, and often overlooked step in data analysis, and too few people are trained in it. (Juliana Freire)

Through better understanding of the tools and techniques used to address big data, one can better understand the relevant education and training needs. The third session of the workshop focused more specifically on how to work with big data. Presentations were made by Jeffrey Ullman (Stanford University), Alexander Gray (Skytree Corporation), Duncan Temple Lang (University of California, Davis), and Juliana Freire (New York University). The session was chaired by Brian Caffo (Johns Hopkins University).

## TEACHING ABOUT MAPREDUCE

*Jeffrey Ullman, Stanford University*

MapReduce (Dean and Ghemawat, 2004), explained Jeffrey Ullman, is a programming method designed for easy parallel programming on commodity hardware, and it eliminates the need for the user to implement the parallelism and to address recovery from failures. MapReduce uses a distributed file system that replicates chunks to protect against data loss, and it is architected so that hardware failures do not require that the job be restarted. Hadoop[1] is an open-source implementation of MapReduce, which is proprietary to Google.

MapReduce, Ullman said, consists of a map function and a reduce function. The map function converts a single element (such as a document, integer, or information record) into key-value pairs. The map tasks are executed in parallel; the code is sent to the data, and the task executes wherever chunks of input are. After the map function has been applied to all inputs, the key-value pairs are sorted by key. The reduce function takes a single key with its list of associated values and provides an output. Reduce tasks are also executed in parallel, and each key with its list of inputs is handled independently.

Ullman then described a data mining course being taught at Stanford University in which students are given access to Amazon Web Services, and many do choose to implement their algorithms by using Hadoop. The course uses real-world data from a variety of sources, including Twitter, Wikipedia, and other companies. Teams of three students propose projects, including the data set to use, the expected results, and how to evaluate their results. About a dozen teams are selected to participate in the course.

Ullman described a 2012 team project on drug interactions. The team used data from Stanford's medical school from which it extracted records for 3,000 drugs. It sought to identify drug interactions and examine each pair of drugs with a chi-squared test, a statistical test to evaluate the likelihood that differences in data arise by chance. The team was able to identify 40 of the 80 known drug combinations that lead to an increased risk of heart attack. More important, it identified two previously unknown pairs on which there was very strong evidence of interaction. Ullman explained that the team recognized that to make the problem more tractable, it needed to address it with fewer keys and longer lists of values, and it combined the drugs into groups, thereby reducing the number of comparisons and correspondingly reducing the amount of network-use time needed. Ullman stated that this example illustrated how communication time can often be the bottleneck in MapReduce algorithms.

[1] See Apache Software Foundation, "Apache Hadoop," http://hadoop.apache.org/ (accessed May 14, 2014), for more information.

Ullman then spoke more broadly about the theory of MapReduce models. Such models require three elements:

- *Reducer*[2] *size:* the maximum number of inputs that a given reducer can have, which leads to an upper bound on the length of the value list.
- *Replication rate:* the average number of key-value pairs generated by a mapper on one input. This measures communication cost per input; it is common for the replication rate to measure the length of time needed to run the algorithm.
- *Mapping schema:* a description of how outputs for a problem are related to inputs or an assignment of inputs to sets of reducers. No reducer is assigned more inputs than the reducer size; and for every output, there is some reducer that receives all the inputs associated with it.

Ullman showed that replication rate is inversely proportional to reducer size; this forces a trade-off between the two variables and provides a bound on replication rate as a function of reducer size. Ullman pointed out that the inverse relationship makes sense: when more work is done by a single reducer, less parallelism is needed, and the communication cost becomes smaller.

## BIG DATA MACHINE LEARNING—PRINCIPLES FOR INDUSTRY

*Alexander Gray, Skytree Corporation*

Alexander Gray began by briefly describing the first three phases of machine learning: artificial intelligence and pattern recognition (1950s-1970s), neural networks and data mining (1980s and 1990), and convergence of machine learning with statistics (middle 1990s to today). Gray considers that we are now seeing the beginning of a fourth phase, defined by big data with new scalable systems needed to support it.

Gray explained that almost every industry has big data and would be better served by understanding it. He noted a variety of situations in which machine learning is "mission-critical"; in general, this occurs when some extreme is needed, such as high volume, high speed, or extreme accuracy. Gray described a number of kinds of applications of big data, including science (the Search for Extra-Terrestrial Intelligence, the Sloan Digital Sky Survey, and the Large Hadron Collider), medicine (health-care cost reduction, predictive health,[3] and early detection), finance

[2] A reducer is a function that typically maps a larger set of values to a smaller set of values.

[3] The goal of "predictive health" is to predict the probability of future diseases to identify useful proactive lifestyle modifications and surveillance.

(improving derivative pricing, risk analysis, portfolio optimization, and algorithmic trading), and security (cybersecurity, crime prevention, and antiterrorism). In addition, Gray noted kinds of applications that he described as having lower stakes: recommendations, face tagging, dating matches, and online advertising. He posited that many companies would benefit from machine learning to compete and ultimately to survive.

Gray then asked how to maximize predictive accuracy and explained that overall prediction error decomposes into errors that result from the use of finite samples, the choice of model parameters (i.e., algorithmic accuracy), and the choice of models. He noted that one can increase computational speed by orders of magnitude by using smarter algorithms. In addition, speed is connected to accuracy in that speed allows the analyst more time to explore the parameter space. Gray then described weak and strong scaling, a high-performance computing concept that manages data either by using more machines (strong scaling) or by taking more time (weak scaling). With data sets that contain millions of items, parallelism can provide good scaling—for example, changing from one computer to five computers might lead to a 5-fold speed increase in calculation. Gray indicated that data sets that contain billions of items are not uncommon and said that his firm has worked with one client that had data sets that contained trillions of items. Gray noted that strong and weak scaling result in different errors.

In addressing algorithmic accuracy, Gray pointed out that stochastic methods are optimal but generally do not reach optimal results in a single iteration. That type of computation is useful for "quick and dirty" applications. In addressing model error, Gray emphasized the importance of understanding and using a variety of models, as the best model changes on the basis of the data set. He also indicated that the treatment of outliers can change the outcome of an analysis. And he pointed out the utility of visualizing data in a data-specific and domain-specific approach and indicated a need for improved exploratory data analysis and visualization tools. A workshop participant supported the use of visualization and emphasized the need to include the human in the loop; the user should be responsible for and involved in the visualization, not passive, and the visualization should enhance understanding of the data.

## PRINCIPLES FOR THE DATA SCIENCE PROCESS

*Duncan Temple Lang, University of California, Davis*

Duncan Temple Lang began by listing the core concepts of data science—items that will need to be taught: statistics and machine learning, computing and technologies, and domain knowledge of each problem. He stressed the

importance of interpretation and reasoning—not only methods—in addressing data. Students who work in data science will have to have a broad set of skills—including knowledge of randomness and uncertainty, statistical methods, programming, and technology—and practical experience in them. Students tend to have had few computing and statistics classes on entering graduate school in a domain science.

Temple Lang then described the data analysis pipeline, outlining the steps in one example of a data analysis and exploration process:

1. Ask a general question.
2. Refine the question, identify data, and understand data and metadata. Temple Lang noted that the data used are usually not collected for the specific question at hand, so the original experiment and data set should be understood.
3. Access data. This is unrelated to the science but does require computational skill.
4. Transform to data structures.
5. Perform exploratory data analyses to understand the data and determine whether the results will scale. This is a critical step; Temple Lang noted that 80 percent of a data scientist's time can be spent in cleaning and preparing the data.
6. Perform dimension reduction. Temple Lang stressed that it can be difficult or impossible to automate this step.
7. Perform modeling and estimation. Temple Lang noted that computer and machine learning scientists tend to focus more on predictive models than on modeling of physical behavior or characteristics.
8. Perform diagnostics. This helps to understand how well the model fits the data and identifies anomalies and aspects for further study. This step has similarities to exploratory data analysis.
9. Quantify uncertainty. Temple Lang indicated that quantifying uncertainty with statistical techniques is important for understanding and interpreting models and results.
10. Convey results.

Temple Lang stressed that the data analysis process is highly interactive and iterative and requires the presence of a human in the loop. The next step in data processing is often not clear until the results of the current step are clear, and often something unexpected is uncovered. He also emphasized the importance of abstract skills and concepts and said that people need to be exposed to authentic data analyses, not only to the methods used. Data scientists also need to have a statistical understand-

ing, and Temple Lang described the statistical concepts that should be taught to a student:

- Mapping the general question to a statistical framework.
- Understanding the scope of inference, sampling, biases, and limitations.
- Exploratory data analyses, including missing values, data quality, cleaning, matching, and fusing.
- Understanding randomness, variability, and uncertainty. Temple Lang noted that many students do not understand sampling variability.
- Conditional dependence and heterogeneity.
- Dimension reduction, variable selection, and sparsity.
- Spurious relationships and multiple testing.
- Parameter estimation versus "black box" prediction and classification.
- Diagnostics—residuals and comparing models.
- Quantifying the uncertainty of a model.
- Sampling structure and dependence for data reduction. Temple Lang noted that modeling of data becomes complicated when variables are not independent, identically distributed.
- Statistical accuracy versus computational complexity and efficiency.

Temple Lang then briefly discussed some of the practical aspects of computing, including the following:

- Accessing data.
- Manipulating raw data.
- Data structures and storage, including correlated data.
- Visualization at all stages (particularly in exploratory data analyses and conveying the results).
- Parallel computing, which can be challenging for a new student.
- Translating high-level descriptions to optimal programs.

During the discussion, Temple Lang proposed computing statistics on visualizations to examine data rigorously in a statistical and automated way. He explained that "scagnostics" (from *scatter plot diagnostics*) is a data analysis technique for graphically exploring the relationships among variables. A small set of statistical measures can characterize scatter plots, and exploratory data analysis can be conducted on the residuals.[4]

A workshop participant noted the difference between a data error and a data blunder. A blunder is a large, easily noticeable mistake. The participant gave the

[4] More information about scagnostics can be found in Wilkinson et al. (2005, 2006).

example of shipboard observations of cloud cover; blunders, in that case, occur when the location of the ship observation is given to be on land rather than at sea. Another blunder would be a case of a ship's changing location too quickly. The participant speculated that such blunders could be generalized to detect problematic observations, although the tools would need to be scalable to be applied to large data sets.

## PRINCIPLES FOR WORKING WITH BIG DATA

*Juliana Freire, New York University*

Juliana Freire began her presentation by discussing the tasks involved in addressing big data. She referred to a Computing Research Association (CRA) report[5] on the challenges posed by big data. CRA also documented the data analysis pipeline, which includes acquisition and recording; extraction, cleaning, and annotation; analysis and modeling; and interpretation. A simplified schematic of the pipeline is shown in Figure 3.1.

Freire posited that scaling for batch computation is not difficult—people have been working on this problem for several decades, and there is an infrastructure to support it. However, the human scalability is difficult; as the data size increases, it becomes more difficult for an analyst to explore the data space. The path from data to knowledge, she noted, is human-based and has many complicated elements.

Freire explained that the CRA data analysis pipeline tasks can be classified into two categories: data preparation (which includes acquisition and recording; extraction, cleaning, and annotation; and integration, aggregation, and representation) and data analysis (which includes modeling and interpretation). Data science includes statistics, machine learning, data mining, and visualization, but Freire noted that in many institutions it is synonymous with machine learning, and less emphasis is placed on the other elements. She pointed out that data visualization has been growing in importance and that there is a corresponding need for additional training in it. Freire emphasized that the data pipeline is complex and that what is shown in Figure 3.1 is an oversimplification; for instance, the pipeline is not linear. She also stressed the importance of research provenance: provenance of the exploration process should be captured for transparency, reproducibility, and knowledge reuse. She noted that provenance management is not often taught.

Freire acknowledged that people underestimate the effort required in preparing data. Few people have the expertise to prepare data, but there is a high demand for

[5] "Challenges and Opportunities with Big Data—A Community White Paper Developed by Leading Researchers Across the United States," http://www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf, accessed May 19, 2014.

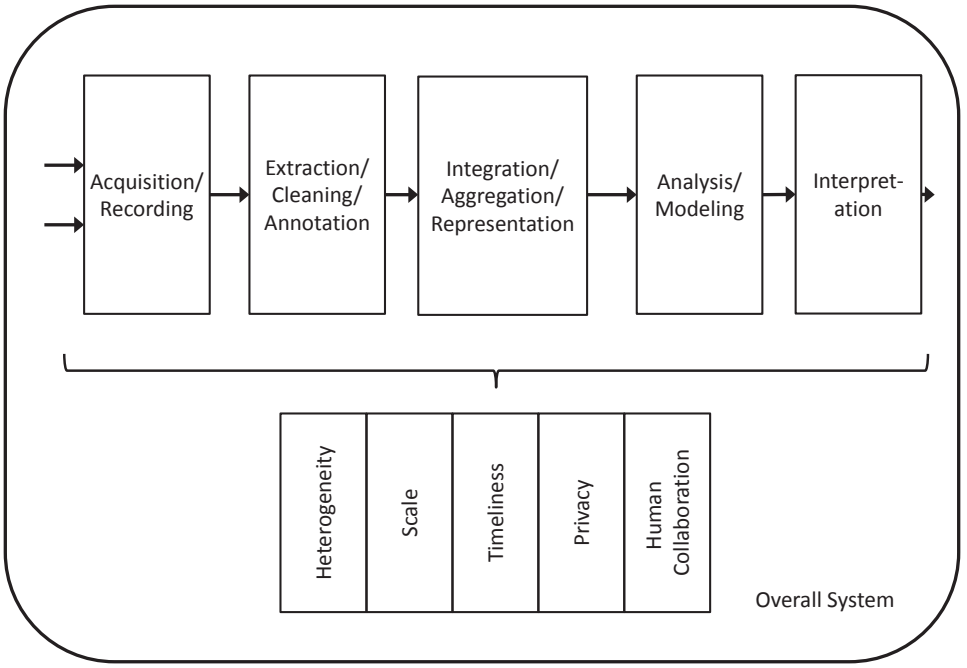


FIGURE 3.1 Simplified schematic of the big data analysis pipeline. Major steps in the analysis of big data are shown in the flow at top. Below it are big data needs that make these steps challenging. SOURCE: Computing Community Consortium, February 2012.

data preparation. In contrast, there are many experts to conduct the analysis, but relatively little time is needed for this step. She stated that data preparation takes a long time, is idiosyncratic, and can limit analyses. She also noted that new data sets continually provide new challenges in big data, and many needs are not met by existing infrastructure.

Freire then provided an example of recent work in applying data science principles to New York City taxis. The raw data set consisted of 500,000 trips per day taken for more than 3 years, which yielded 150 GB of data. The data were not enormous, but they were complex and had spatial and temporal attributes. The data show an unusual degree of regularity; one can easily see temporal changes related to weekends and holidays. The goal was to allow city officials to explore the data visually. The work involved developing a spatiotemporal index that was based on an out-of-core k-dimensional tree (Ferreira et al., 2013) and a new interactive map view.

Freire stated that domain scientists do not know what is possible to do with

their data, and technologists do not know the domain, so there is an expertise gap. Freire quoted Alex Szalay (Faris et al., 2011), who described the ideal scientist as "π-shaped," with deep knowledge in two fields and connections between them. Freire argued that although the data scientist is supposed to fill the expertise gap, in reality three people make up a "data scientist": a database expert, a machine learning and statistics expert, and a visualization expert. She said that computer science and data management research have partly failed in that it has not been able to create usable tools for end users. Freire stated that the complexity of data science problems is often underestimated.

Freire was asked by a workshop participant how to prepare students in software while teaching them their domain science. She suggested adding a new course for students who do not have a computer science background. She noted that there were several boot-camp-style programs for Ph.D. science students but that their overall effectiveness is not known.

Participants also discussed the requirements for a data analyst, a topic discussed by Temple Lang during his presentation. One person posited that the single expert in databases, machine learning and statistics, and visualization that Freire described should also be knowledgeable in systems and tools. The database expertise should include computational environments, not just databases. Another participant described the data analyst as a "jazz player" rather than a "symphony player"—in other words, a data analyst should improvise and make decisions rapidly, which cannot be done if the analyst does not know the subject matter well.

Some participants discussed tools. One person noted that commercial tools (such as Spotfire[6] and Tableau[7]) exist in a polished form and work in a variety of applications. Others responded, however, that students need training on these tools, and that a single tool does not usually solve complex data problems. A participant noted that students cannot afford a subscription to Tableau and argued that the existing tools should be open-source; however, open-source tools may not always be well curated.

[6] See TIBCO Software, Inc., "Spotfire," http://spotfire.tibco.com/, accessed June 9, 2014, for more information.

[7] See the Tableau Software website at http://www.tableausoftware.com/ (accessed June 9, 2014) for more information.

# 4

# Courses, Curricula, and Interdisciplinary Programs

**Important Points Made by Individual Speakers**

- A residual effect of training students to work with data is that the training will empower them with a toolkit that they can use in multiple domains. (Joshua Bloom)
- Boot camps and other short courses appear to be successful in teaching data computing techniques to domain scientists and in addressing a need in the science community; however, outstanding questions remain about how to integrate these types of classes into a traditional educational curriculum. (Joshua Bloom)
- Educators should be careful to teach data science methods and principles and avoid teaching specific technologies without teaching the underlying concepts and theories. (Peter Fox)
- Massive online open courses (MOOCs) are one avenue for teaching data science techniques to a large population; thus far, data science MOOC participants tend to be computer science professionals, not students. (William Howe)

By the end of 2014, more than 30 major universities will have programs in data science.[1] Existing and emerging programs offer many opportunities for lessons

[1] See the Master's in Data Science website at http://www.mastersindatascience.org/ (accessed June 5, 2014) for more information.

learned and potential course and content models for universities to follow. The fourth workshop session focused on specific coursework, curricula, and interdisciplinary programs for teaching big data concepts. The session was chaired by James Frew (University of California, Santa Barbara). Presentations were made in this session by Joshua Bloom (University of California, Berkeley), Peter Fox (Rensselaer Polytechnic Institute), and William Howe (University of Washington).

## COMPUTATIONAL TRAINING AND DATA LITERACY FOR DOMAIN SCIENTISTS

*Joshua Bloom, University of California, Berkeley*

Joshua Bloom noted that the purpose of graduate school is to prepare students for a career in the forefront of science. A residual effect of training students to work with data is that the training will empower the students with a toolkit that they can use even if they leave a particular domain. He pointed out that the modern data-driven science toolkit is vast and that students are being asked to develop skills in both the domain science and the toolkit.

Bloom then described upcoming data challenges in his own domain of astronomy. The Large Synoptic Survey Telescope is expected to begin operations in 2020, and it will observe 800 million astronomical sources every 3 days. A large computational framework is needed to support that amount of data, probably 20 TB per night. Other projects in radio astronomy have similar large-scale data production.

A goal in data science for time-domain astronomy in the presence of increasing data rates is to remove the human from the real-time data loop, explained Bloom—in other words, to develop a fully automated, state-of-the-art scientific stack to observe transient events. Often, the largest bottleneck is in dealing with raw data, but there are large-scale inference challenges further downstream.

Bloom pointed out that the University of California, Berkeley, has a long history of teaching parallel computing. The coursework is aimed at computer science, statistics, and mathematics students. Recently, "boot camps" that include two or three intensive training sessions have been initiated to teach students basic tools and common frameworks. The boot camp at Berkeley takes 3 full days. There are six to eight lectures per day, and hands-on programming sessions are interspersed. Bloom began teaching computing techniques to domain scientists, primarily physical-science students. His first boot camp consisted of several all-day hands-on classes with nightly homework. The student needed to know a programming language before taking the boot camp. In 2010, the first year, 85 students participated. By 2013, the boot camp had grown to more than 250 students. Bloom uses live streaming and archiving of course material, and all materials used are open-source. The course has been widely used and repeated; for instance, NASA Goddard Space Flight Center

used his materials to hold its own boot camp. Bloom noted, in response to a question, that instructors in his course walk around the room to assist students while they work. He posited that 90 percent of that interaction could be replaced with a well-organized chat among instructors and students; the course would probably take longer, and students would have to be self-directed.

Bloom explained that the boot camp is a prerequisite to Berkeley's graduate-level follow-on seminar course in Python computing for science. The graduate seminar was the largest graduate course ever taught in Berkeley's astronomy department; this indicated an unmet need for such a course at the graduate science level. Bloom said that the boot camps and seminars give rise to a set of education questions: Where do boot camps and seminars fit into a traditional domain-science curriculum? Are they too vocational or practical to be part of higher-education coursework? Who should teach them, and how should the instructors be credited? How can students become more (broadly) data literate before we teach them big data techniques? He emphasized that at the undergraduate level the community should be teaching "data literacy" before it teaches data proficiency. Some basic data-literacy ideas include the following:

- *Statistical inference.* Bloom noted that this is not necessarily big data; something as simple as fitting a straight line to data needs to be taught in depth.
- *Versioning and reproducibility.* Bloom noted that several federal agencies are likely to mandate a specific level of reproducibility in work that they fund.

Bloom suggested that there is a "novelty-squared" problem: what is novel in the domain science may not be novel in the data science methodology. He stressed the need to understand the forefront questions in various fields so that synergies can be found. For example, Berkeley has developed an ecosystem for domain and methodological scientists to talk and find ways to collaborate.

Bloom also noted that data science tends to be an inclusive environment that appeals to underrepresented groups. For instance, one-third of the students in the Python boot camps were women—a larger fraction than their representation in physical science graduate programs.

Bloom concluded by stating that domain science is increasingly dependent on methodologic competences. The role of higher education in training in data science is still to be determined. He stressed the need for data literacy before data proficiency and encouraged the creation of inclusive and collaborative environments to bridge domains and methodologies.

Bloom was asked what he seeks in a student. He responded that it depends on the project. He looked for evidence of prior research, even at the undergraduate

level, as well as experience in programming languages and concepts. However, he noted that a top-quality domain scientist would always be desirable regardless of computational skills.

A participant commented that as much as 80 percent of a researcher's time is spent in preparing the data. That is a large amount of time that could be spent on more fundamental understanding. Bloom responded that such companies and products as OpenRefine,[2] Data Wrangler,[3] and Trifacta[4] are working on data cleaning techniques. However, for any nontrivial question, it is difficult to systematize data preparation. He also suggested that a body of fundamental research should be accessible to practitioners. However, large-scale, human-generated data with interesting value do not typically flow to academe because of privacy and security concerns. He conjectured that the advent of the Internet of Things will allow greater data access, because those data will not be human data and therefore will have fewer privacy concerns.

## DATA SCIENCE AND ANALYTICS CURRICULUM DEVELOPMENT AT RENSSELAER (AND THE TETHERLESS WORLD CONSTELLATION)

*Peter Fox, Rensselaer Polytechnic Institute*

Peter Fox began by describing the Tetherless World Constellation[5] at Rensselaer Polytechnic Institute (RPI). The research themes are divided loosely into three topics: future Web (including Web science, policy, and social issues), Xinfomatics (including data frameworks and data science), and semantic foundations (including knowledge provenance and ontology engineering environments). Fox indicated that his primary focus is Xinformatics. He deliberately did not define *X*, saying that it can mean any number of things.

Fox explained that to teach data science, one must "pull apart" the ecosystem in which data science lives. Data, information, and knowledge are all related in a data science ecosystem; there is no linear pathway from data to information to knowledge. He explained that he teaches or is involved in classes on data science, Xinformatics, geographic information systems for the sciences, semantic eScience, data analytics, and semantic technologies. The students in those classes have varied backgrounds. Last semester, his data science class had 63 students (most of them graduate students), and Xinformatics had about 35 students. Fox structures his

[2] See the OpenRefine website at http://openrefine.org/ (accessed June 9, 2014) for more information.

[3] See Stanford Visualization Group, "Data Wrangler alpha," http://vis.stanford.edu/wrangler/, accessed June 9, 2014, for more information.

[4] See the Trifacta website at http://www.trifacta.com/ (accessed June 9, 2014) for more information.

[5] See Rensselaer Polytechnic Institute (RPI), "Tetherless World Constellation," http://tw.rpi.edu, accessed May 22, 2014, for more information.

classes so that the first half of the semester focuses on individual work and gaining knowledge and skills. The second half focuses on team projects (with teams assigned by him) to demonstrate skills and share perspectives.

Fox explained that he teaches modern informatics and marries it with a method: the method that he teaches is iterative and is based on rapid prototyping applied to science problems. The framework for the iterative model is shown in Figure 4.1. Fox stressed that technology does not enter the method until well over halfway through the spiral; technology will change, so it is important to impart skills before adopting and leveraging technology.

Fox explained that a report was produced for NSF (Borgman et al., 2008) and that a diagram was developed that describes five types of mediation of increasing complexity (shown in Figure 4.2). The five generations of mediation were designed to apply to learning, but they hold true for science and research as well. Fox explained that, in contrast with most generational plots, all generations are present and active at once in both the learning and teaching environment and the science and research environment.

Fox explained that data analytics is a new course at RPI, and his desired prerequisites are not taught at the university; as a result, his class has no prerequisites. After teaching a variety of computer application languages simultaneously, Fox now

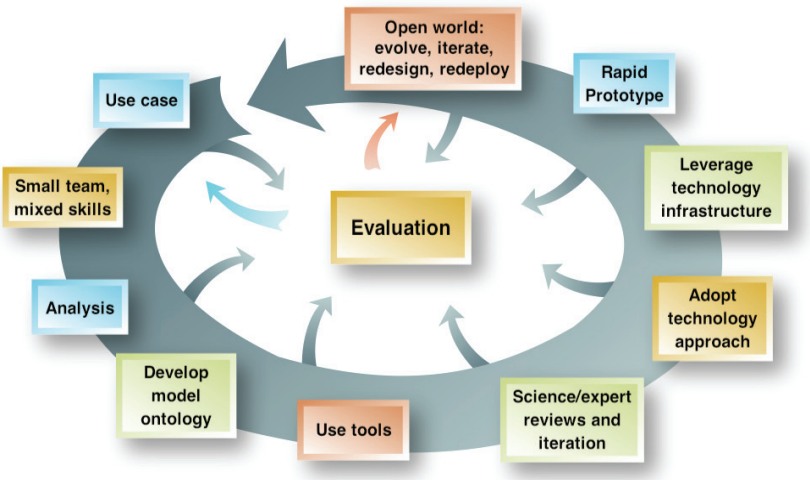


FIGURE 4.1 Framework for modern informatics. A technology approach does not enter into the development spiral until over halfway through the process. SOURCE: Fox and McGuinness (2008), http://tw.rpi.edu/media/latest/SemanticMethodologyPathwayPretty_v2.png.

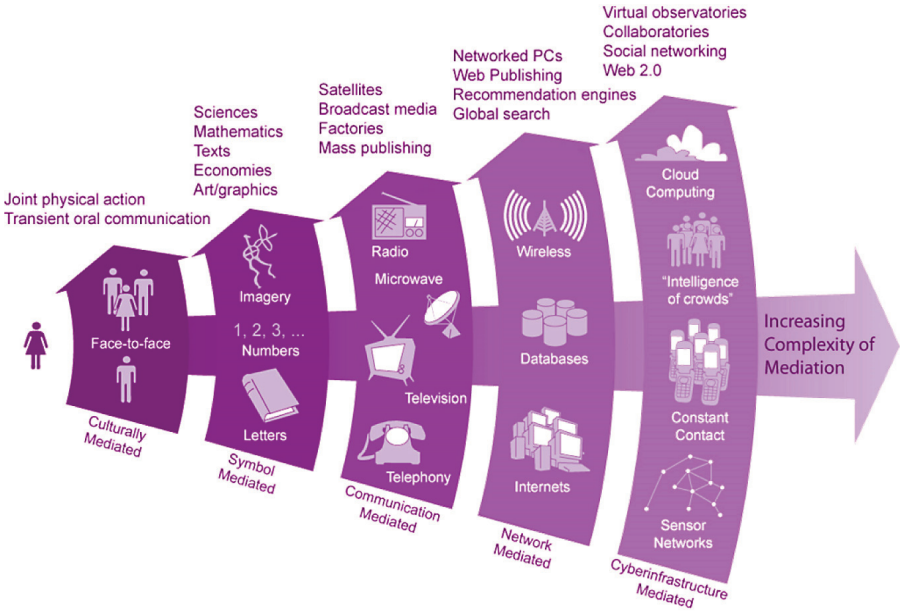


FIGURE 4.2 Generations of mediation, applied to the learning and teaching environment and the science and research environment. SOURCE: Illustration by Roy Pea and Jillian C. Wallis, from Borgman et al. (2008).

uses the R and RStudio[6] environment exclusively. (Students preferred the simplicity of learning a single language.) The data analytics course builds from data to processing to reporting to analytics, both predictive and prescriptive. Fox explained that in the ideal scenario, value would be added as one progresses from one step to the next. Part of the challenge is to teach students to understand the value added and to teach them to identify when value is *not* being added. He emphasized the importance of understanding the value and application of data analysis, not just learning the individual steps. In response to a later question, Fox clarified that students work with self-selected, application-specific examples.

Fox then described information technology and Web-science coursework at RPI. RPI has an interdisciplinary program that consists of a B.S. with 20 concentrations, an M.S. with 10 concentrations, and a multidisciplinary Ph.D. offering.[7] The program has four technical tracks—computer engineering, computer science,

[6] RStudio is an open source, professional user interface for R. See the RStudio website at http://www.rstudio.com/, accessed September 20, 2014.

[7] See the RPI Information Technology and Web Science website at http://itws.rpi.edu (May 22, 2014) for more information.

information systems, and Web science—with numerous concentrations in each track. Fox said that the M.S. was recently revised to include data analytics in the core curriculum and that the M.S. concentrations were updated. He noted in particular the addition of the Information Dominance concentration, which is designed to educate a select group of naval officers each year in skills needed to execute military cyberspace operations.

Fox talked about the Data Science Research Center[8] and the less formal Data Science Education Center at RPI. The centers are loosely organized; more than 45 faculty and staff are involved. RPI also maintains a data repository[9] for scientific data that result from on-campus research.

He listed some lessons learned after 5 years with these programs:

- Be interdisciplinary from the start; grow both technical and data skills simultaneously. Fox noted that teaching skills (such as how to manipulate data by using specific programming languages) can be difficult; skills need to be continually reinforced, and he cautioned that teaching skills may be perceived as training rather than education.
- Teach methods and principles, not technology.
- Make data science a skill in the same vein as laboratory skills.
- Collaboration is critical, especially in informatics.
- Teach foundations and theory.

Fox stated that access to data is progressing from provider-to-user to machine-to-user and finally to machine-to-machine; the burden of data access and usability shifts from the user to the provider. In the current research-funding paradigm, data are collected, data are analyzed by hand for several years, and the results are then published. Although that paradigm has served the research community well, Fox noted that it fails to reflect the change in responsibilities that is inherent in the new information era, in which the burden of access shifts from the user to the provider.

Fox concluded by positing that the terms *data science* and *metadata* will be obsolete in 10 years as researchers come to work with data as routinely as they use any other research tool.

Bloom noted that there was no mention of "big data" in Fox's presentation, only data. Fox stated that he does not distinguish big data from data. However, he acknowledged that, as a practical matter, size, heterogeneity, and structural representations will need to be parts of a student's course of study.

[8] See the RPI Data Science Research Center website at http://dsrc.rpi.edu/ (May 22, 2014) for more information.

[9] See RPI, "Rensselaer Data Services," http://data.rpi.edu, accessed May 22, 2014, for more information.

## EXPERIENCE WITH A FIRST MASSIVE ONLINE OPEN COURSE ON DATA SCIENCE

*William Howe, University of Washington*

William Howe stated that the University of Washington (UW) founded the eScience Institute in 2008 and that the institute is now engaged in a multi-institution partnership with the University of California, Berkeley, and New York University and is funded by the Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation to advance new data science techniques and technologies, foster collaboration, and create a cross-campus "data science environment." According to Howe, the strategy is to establish a "virtuous cycle" between the data science methodology researchers and the domain-science researchers in which innovation in one field will drive innovation in the other. The eScience Institute works to create and reinforce connections between the two sides, and six working groups act as bridges. One of the working groups is involved with education and training.

Howe explained that the education and training working group focuses on different ways to educate students and practitioners in data science. Through the working group, the UW eScience Institute has developed a data science certificate for working professionals, an interdisciplinary Ph.D. track in big data, new introductory courses, a planned data science master's degree, and a MOOC called "Introduction to Data Science." Howe focused in more detail on his experiences in developing and teaching the MOOC. Teaching a MOOC involves a large amount of work, he said, and the course is continuously developing. The goal of the MOOC is to organize a set of important topics spanning databases, statistics, machine learning, and visualization into a single introductory course. He provided some statistics on the data science MOOC:

- More than 110,000 students registered for the course. Howe noted that that is not a particularly relevant statistic, inasmuch as many people who register for MOOCs do not necessarily plan to participate.
- About 9,000 students completed the course assignments. Howe indicated that that is a typical level of attrition for a MOOC.
- About 7,000 students passed the course and earned the certificate.

He explained that the course had a discussion forum that worked well. Many comments were posted to it, and it was self-sustaining; Howe tried to answer questions posed there but found that questions were often answered first by other engaged students.

The syllabus that defined his 9-week MOOC consisted of the following elements:

- Background and scope of "data science."
- Data manipulation at scale.
- Analytics. Howe taught selected statistics concepts in 1 week and machine learning concepts in another week.
- Visualization.
- Graph and network analytics. Howe indicated this was a single, short module.

Howe explained that the selection of topics was motivated by a desire to develop the four dimensions of the course: tools versus abstractions (weighted toward abstractions), desktop versus cloud (weighted toward cloud), hackers versus analysts (balanced, although perhaps slightly in favor of hackers), and data structures and programming versus mathematics and statistics (weighted toward structures).

He conducted a demographic study of his MOOC participants and remarked that most of them were working professional software engineers, as has been reported for other MOOCs. He suggested that perhaps a MOOC could be used like a textbook, with instructors having their students watch some lectures and skip others, just as they do chapters of a book.

A teaching strategy that consists of both online and in-person components, he explained, has two possible approaches: offer the same course simultaneously online and in person or use the online component as a textbook and class time as an opportunity for practical application (juxtaposing the traditional roles of homework and classwork). There are examples of both teaching strategies, and it is unclear whether either will dominate. He also reiterated the importance of student-to-student learning that took place in his experience with a MOOC structure.

In the discussion period, a participant asked about the importance of understanding the foundations of programming and suggested that algorithms and data constructs should be taught to younger students, for example, in high school. (That last point generated some disagreement among participants. One suggested that even elementary school would be appropriate, but another was concerned that this might displace students from calculus and other critical engineering mathematics courses.) Howe replied that computer science enrollment is increasing at the undergraduate level by about 20 percent per year, and statistics departments are also seeing increased enrollment. Students understand that they need to understand the core concepts at the undergraduate level.

A workshop participant asked about other MOOC success stories in big data. Howe responded that he had only anecdotal evidence. Bloom concurred, noting that a graduate student who had participated in Berkeley's data science boot camp conducted large-scale parallel computing work that resulted in a seminal paper in his field (Petigura et al., 2014).

# 5

# Shared Resources

### Important Points Made by Individual Speakers

- Synthetic knowledge bases for domain sciences, such as the PaleoDeepDive system at Stanford University, can be developed by using the automatic extraction of data from scientific journals. (Christopher Ré)
- Divide and recombine methods are powerful tools for analysts who conduct deep examinations of data, and such systems can be used by analysts without the need for complex programming or a profound understanding of the tools used. (Bill Cleveland)
- Yahoo Webscope is a reference library of large, scientifically useful, publicly available data sets for researchers to use. (Ron Brachman)
- Amazon Web Services (AWS) hosts large data sets in a variety of models (public, requestor-paid, and private or community) to foster large-scale data sharing. AWS also provides data computation tools, training programs, and grants. (Mark Ryland)

The fifth session of the workshop was chaired by Deepak Agarwal (LinkedIn Corporation). The session had four speakers: Christopher Ré (Stanford University), Bill Cleveland (Purdue University), Ron Brachman (Yahoo Labs), and Mark Ryland (Amazon Corporation).

## CAN KNOWLEDGE BASES HELP ACCELERATE SCIENCE?

*Christopher Ré, Stanford University*

Christopher Ré focused on a single topic related to data science: knowledge bases. He first discussed Stanford University's experience with knowledge bases. Ré explained that in general scientific discoveries are published and lead to the spread of ideas. With the advent of electronic books, the scientific idea base is more accessible than ever before. However, he cautioned, people are still limited by their eyes and brain power. In other words, the entire science knowledge base is accessible but not necessarily readable.

Ré noted that today's science problems require macroscopic knowledge and large amounts of data. Examples include health, particularly population health; financial markets; climate; and biodiversity. Ré used the latter as a specific example: broadly speaking, biodiversity research involves assembling information about Earth in various disciplines to make estimates of species extinction. He explained that this is "manually constructed" data—a researcher must input the data by examining and collating information from individual studies. Manually constructed databases are time-consuming to produce; with today's data sources, the construction exceeds the time frame of the typical research grant. Ré posited that the use of sample-based data and their synthesis constitute the only way to address many important questions in some fields. A system that synthesizes sample-based data could "read" journal articles and automatically extract the relevant data from them. He stated that "reading" machines may be coming in the popular domain (from such Web companies as IBM, Google, Bing, and Amazon). The concept of these machines could be extended to work in a specific scientific domain. That would require high-quality reading—reading of a higher quality than is needed in a popular-domain application—in that mistakes can be more harmful in a scientific database.

Ré described a system that he has developed, PaleoDeepDive,[1] a collaborative effort with geologist Shanan Peters, also of Stanford University. The goal of PaleoDeepDive is to build a higher-coverage fossil record by extracting paleobiologic facts from research papers. The system considers every character, word, or fragment of speech from a research paper to be a variable and then conducts statistical inference on billions of variables defined from the research papers to develop relationships between biologic and geologic research. PaleoDeepDive has been in operation for about 6 months, and preliminary results of occurrence relations extracted by PaleoDeepDive show a precision of around 93 percent; Ré indicated that this is a very high-quality score.

[1] See the DeepDive website at http://deepdive.stanford.edu/ (accessed June 9, 2014) for more information.

Ré then stated the challenges for domain scientists related to synthetic knowledge bases:

- *Students are not trained to ask questions of synthetic data sets.* Ré noted that this may be changing; the University of Chicago, for instance, includes such training in its core curriculum. Stanford has an Earth-science class on how to use PaleoDeepDive.
- *Students lack skills in computer science and data management.* Ré indicated that this also may be changing; 90 percent of Stanford students now take at least one computer science class.
- *Some people are skeptical of human-generated synthetics.* Ré suggested that this is also changing as statistical methods take stronger hold.

Ré noted challenges for computer scientists related to synthetic knowledge bases:

- *Finding the right level of abstraction.* Ré posited that approaches to many interesting questions would benefit from the use of synthetic knowledge bases. However, PaleoDeepDive is not necessarily scalable or applicable to other disciplines.
- *Identifying features of interest.* Computer scientists, Ré noted, focus on algorithms rather than on features. However, synthetic knowledge bases are feature based and require a priori knowledge of what is sought from the data set.

A participant noted that noise, including misspelled words and words that have multiple meanings, is a standard problem for optical character recognition (OCR) systems. Ré acknowledged that OCR can be challenging and even state-of-the-art OCR systems make many errors. PaleoDeepDive uses statistical inference and has a package to improve OCR by federating open-source material together and using probabilistic inputs. Ré indicated that Stanford would be releasing tools to assist with OCR.

## DIVIDE AND RECOMBINE FOR LARGE, COMPLEX DATA

*Bill Cleveland, Purdue University*

Bill Cleveland explained the goals of divide and recombine for big data:

- Methods and environments that do not require reductions in dimensionality should be used for analyzing data at the finest level of granularity possible. The data analysis could include visualization.

- At the front end, analysts can use a language for data analysis to tailor the data and make the system efficient.
- At the back end, a distributed database is accessible and usable without the need for the analyst to engage in the details of computation.
- Within the computing environment, there is access to the many methods of machine learning and visualization.
- Software packages enable communication between the front and back ends.
- The system can be used continuously to analyze large, complex data sets, generate new ideas, and serve as a test bed.

Cleveland then described the divide and recombine method. He explained that a division method is used first to divide the data into subsets. The subsets are then treated with one of two categories of analytic methods:

- *Number-category methods.* Analytic methods are applied to each of the subsets with no communication among the computations. The output from this method is numeric or categoric.
- *Visualization.* The data are organized into images. The output from this method is plots. It is not feasible to examine all the plots, so the images are sampled. That can be done rigorously; sampling plans can be developed by computing variables with one value per subset.

Cleveland described several specific methods of division. In the first, conditioning-variable division, the researcher divides the data on the basis of subject matter regardless of the size of the subsets. That is a pragmatic approach that has been widely used in statistics, machine learning, and visualization. In a second type of division, replicate division, observations are exchangeable, and no conditioning variables are used. The division is done statistically rather than by subject matter. Cleveland stated that the statistical division and recombination methods have an immense effect on the accuracy of the divide and recombine result. The statistical accuracy is typically less than that with other direct methods. However, Cleveland noted that this is a small price to pay for the simplicity in computation; the statistical computation touches subsets no more than once. Cleveland clarified that the process is not MapReduce; statistical methods in divide and recombine reveal the best way to separate the data into subsets and put them back together.

Cleveland explained that the divide and recombine method uses R for the front end, which makes programming efficient. R saves the analyst time, although it is slower than other options. It has a large support and user community, and statistical packages are readily available. On the back end, Hadoop is used to enable parallel computing. The analyst specifies, in R, the code to do the division compu-

tation with a specified structure. Analytic methods are applied to each subset or each sample. The recombination method is also specified by the analyst. Cleveland explained that Hadoop schedules the microprocessors effectively. Computation is done by the mappers, each with an assigned core for each subset. The same is true for the reducers; reducers carry out the recombination. The scheduling possibilities are complex, Cleveland said. He also noted that this technique is quite different from the high-performance computing systems that are prevalent today. In a high-performance computing application, time is reserved for batch processing; this works well for simulations (in which the sequence of steps is known ahead of time and is independent of the data), but it is not well suited to sustained analyses of big data (in which the process is iterative and adaptive and depends on the data).

Cleveland described three divide and recombine software components between the front and back ends. They enable communication between R and Hadoop to make the programming easy and insulate the analyst from the details of Hadoop. They are all open-source.

- *R and Hadoop Integrated Programming Environment (RHIPE[2]).* This is an R package available on GitHub.[3] Cleveland noted that RHIPE can be too strenuous for some operating systems.
- *Datadr.*[4] Datadr is a simple interface for division, recombination, and other data operations, and it comes with a generic MapReduce interface.
- *Trelliscope.*[5] This is a trellis display visualization framework that manages layout and specifications; it extends the trellis display to large, complex data.

Cleveland explained that divide and recombine methods are best suited to analysts who are conducting deep data examinations. Because R is the front end, R users are the primary audience. Cleveland emphasized that the complexity of the data set is more critical to the computations than the overall size; however, size and complexity are often correlated.

In response to a question from the audience, Cleveland stated that training students in these methods, even students who are not very familiar with computer science and statistics, is not difficult. He said that the programming is not complex; however, analyzing the data can be complex, and that tends to be the biggest challenge.

---

[2] See Purdue University, Department of Statistics, "Divide and Recombine (D&R) with RHIPE," http://www.datadr.org/, accessed June 9, 2014, for more information.

[3] See GitHub, Inc., "R and Hadoop Integrated Programming Environment," http://github.com/saptarshiguha/RHIPE/, accessed June 9, 2014, for more information.

[4] See Tessera, "datadr: Divide and Recombine in R," http://hafen.github.io/datadr/, accessed June 9, 2014, for more information.

[5] See Tessera, "Trelliscope: Detailed Vis of Large Complex Data in R," http://hafen.github.io/trelliscope/, accessed June 9, 2014, for more information.

## YAHOO'S WEBSCOPE DATA SHARING PROGRAM

*Ron Brachman, Yahoo Labs*

Ron Brachman prefaced his presentation by reminding the audience of a 2006 incident in which AOL released a large data set with 20 million search queries for public access and research. Unfortunately, personally identifiable information was present in many of the searches, and this allowed the identification of individuals and their Web activity. Brachman said that in at least one case, an outside party was able to identify a specific individual by cross-referencing the search records with externally available data. AOL withdrew the data set, but the incident caused shockwaves throughout the Internet industry. Yahoo was interested in creating data sets for academics around the time of the AOL incident, and the AOL experience caused a slow start for Yahoo. Yahoo persisted, however, working on important measures to ensure privacy, and has developed the Webscope[6] data sharing program. Webscope is a reference library of interesting and scientifically useful data sets. It requires a license agreement to use the data; the agreement is not burdensome, but it includes terms whereby the data user agrees not to attempt to reverse-engineer the data to identify individuals.

Brachman said that Yahoo has just released its 50th Webscope data set. Data from Webscope have been downloaded more than 6,000 times. Webscope has a variety of data categories available, including the following:

- *Language and content.* These can be used to research information-retrieval and natural-language processing algorithms and include information from Yahoo Answers. (This category makes up 42 percent of the data in Webscope.)
- *Graph and social data.* These can be used to research matrix, graph, clustering, and machine learning algorithms and include information from Yahoo Instant Messenger (16 percent).
- *Ratings, recommendation, and classification data.* These can be used to research collaborative filtering, recommender systems, and machine learning algorithms and include information on music, movies, shopping, and Yelp (20 percent).
- *Advertising data.* These can be used to research behavior and incentives in auctions and markets (6 percent).
- *Competition data* (6 percent).

[6] See Yahoo! Labs, "Webscope," http://webscope.sandbox.yahoo.com, accessed May 20, 2014, for more information.

- *Computational-system data.* These can be used to analyze the behavior and performance of different types of computer systems architectures, such as distributed systems and networks and include data from the Yahoo Sherpa database system (6 percent).
- *Image data.* These can be used to analyze images and annotations and are useful for image-processing research (less than 4 percent).

Brachman explained that in many cases there is a simple click-through agreement for accessing the data, and they can be downloaded over the Internet. However, downloads are becoming impractical as database size increases. Yahoo had been asking for hard drives to be sent through the mail; now, however, it is hosting some of its databases on AWS.

In response to questions, Brachman explained that each data set is accompanied by a file explaining the content and its format. He also indicated that the data provided by Webscope are often older (around a year or two old), and this is one of the reasons that Yahoo is comfortable with its use for academic research purposes.

Brachman was asked whether any models fit between the two extremes of Webscope (with contracts and nondisclosure agreements) and open-source. He said that the two extremes are both successful models and that the middle ground between them should be explored. One option is to use a trusted third party to hold the data, as is the case with the University of Pennsylvania's Linguistic Data Consortium data.[7]

## RESOURCE SHARING

*Mark Ryland, Amazon Corporation*

Mark Ryland explained that resource sharing can mean two things: technology capabilities to allow sharing (such as cloud resources) and economic and cost sharing, that is, how to do things less expensively by sharing. AWS is a system that does both. AWS is a cloud computing platform that consists of remote computing storage and services. It holds a large array of data sets with three types of product. The first is public, freely available data sets. These data sets consist of freely available data of broad interest to the community and include Yahoo Webscope data, Common Crawl data gathered by the open-source community (240 TB), Earth-science satellite data (40 TB of data from NASA), 1000 Genomes data (350 TB of data from NIH), and many more. Ryland stated that before the genome data were publicly stored in the cloud, fewer than 20 researchers worked with those

[7] See University of Pennsylvania, Linguistic Data Consortium, "LDC Catalog," http://catalog.ldc.upenn.edu/, accessed May 14, 2014, for more information.

data sets. Now, more than 200 are working with the genome data because of the improved access.

A second type of AWS data product is requestor-paid data. This is a form of cost-sharing in which data access is charged to the user account but data storage is charged to the data owner's account. It is fairly popular but perhaps not as successful as AWS would like it to be, and AWS is looking to broaden the program.

The third type of AWS data product is community and private. AWS may not know what data are shared in this model. The data owner controls data access. Ryland explained that AWS provides identity control and authentication features, including Web Identity Federation. He also described a science-oriented data service (Globus[8]), which provides cloud-based services to conduct periodic or episodic data transfers. He explained that an ecosystem is developing around data sharing.

Sharing is also taking place in computation. Ryland noted that people are developing Amazon Machine Images with tools and data "prebaked" into them, and he provided several examples, including Neuroimaging Tools and Resources Clearinghouse and scientific Linux tools. Ryland indicated that there are many big data tools, some of which are commercial and some of which are open-source. Commercial tools can be cost-effective when accessed via AWS in that a user can access a desired tool via AWS and pay only for the time used. Ryland also pointed out that AWS is not limited to single compute nodes and includes cluster management, cloud formation, and cross-cloud capacities. AWS also uses spot pricing, which allows people to bid on excess capacity for computational resources. That gives users access to computing resources cheaply, but the resource is not reliable; if someone else bids more, then the capacity can be taken away and redistributed. Ryland cautioned that projects must be batch-oriented and assess their own progress. For instance, MapReduce is designed so that computational nodes can appear and disappear.

Ryland explained that AWS offers a number of other managed services and provides higher-level application program interfaces. These include Kinesis[9] (for massive-scale data streaming), Data Pipeline[10] (managed datacentric workflows), and RedShift[11] (data warehouse).

---

[8] See the Computation Institute, University of Chicago, and Argonne National Laboratory "Globus" website at https://www.globus.org/ (accessed May 14, 2014) for more information.

[9] See Amazon Web Services, "Amazon Kinesis," http://aws.amazon.com/kinesis/, accessed June 9, 2014, for more information.

[10] See Amazon Web Services, "AWS Data Pipeline," http://aws.amazon.com/datapipeline/, accessed June 9, 2014, for more information.

[11] See Amazon Web Services, "Amazon Redshift," http://aws.amazon.com/redshift/, accessed June 9, 2014, for more information.

AWS has a grants program to benefit students, teachers, and researchers, Ryland said. It is eager to participate in the data science community to build an education base and the resulting benefits. Ryland reported that AWS funds a high percentage of student grants, a fair number of teaching grants, but few research grants. Some of the research grants are high in value, however. In addition to grants, AWS provides spot pricing, volume discounting, and institutional cooperative pricing. In the latter, members of the cooperative can receive shared pricing; AWS intends to increase its cooperative pricing program.

Ryland explained that AWS provides education and training in the form of online training videos, papers, and hands-on, self-paced laboratories. AWS recently launched a fee-based training course. Ryland indicated that AWS is interested in working with the community to be more collaborative and to aggregate open-source materials and curricula. In response to a question, Ryland clarified that the instruction provided by Amazon is on how to use Amazon's tools (such as RedShift), and the instruction is product-oriented, although the concepts are somewhat general. He said that the instruction is not intended to be revenue-generating, and AWS would be happy to collaborate with the community on the most appropriate coursework.

A workshop participant posited that advanced tools, such as the AWS tools, enable students to use systems to do large-scale computation without fully understanding how it works. Ryland responded that this is a pattern in computer science: a new level of abstraction develops, and a compiled tool is developed. The cloud is an example of that. Ryland posited that precompiled tools should be able to cover 80 percent or more of the use cases although some researchers will need more profound access to the data.

# 6

# Workshop Lessons

Robert Kass (Carnegie Mellon University) led a final panel discussion session at the end of the workshop. Panelists included James Frew (University of California, Santa Barbara), Deepak Agarwal (LinkedIn Corporation), Claudia Perlich (Dstillery), Raghu Ramakrishnan (Microsoft Corporation), and John Lafferty (University of Chicago). Panelists and participants were invited to add their comments to the workshop; final comments tended to focus in four categories: types of students, organizational structures, course content, and lessons learned from other disciplines.

## WHOM TO TEACH: TYPES OF STUDENTS TO TARGET IN TEACHING BIG DATA

Robert Kass opened the discussion session by noting that the workshop had shown that there are many types of potential students and that each type would have different training challenges. One participant suggested that business managers need to understand the potential and realities of big data better to improve the quality of communication. Another pointed out that older students may be attracted to big data instruction to pick up missing skill sets. And another suggested pushing instruction into the high-school level. Several participants posited that the background of the student, more than the age or level, is the critical element. For instance, does the student have a background in computer science or statistics? Workshop participants frequently mentioned three main subjects related to big data: computation, statistics, and visualization. The student's background knowledge in each of the three will have the greatest effect on the student's learning.

## HOW TO TEACH: THE STRUCTURE OF TEACHING BIG DATA

Numerous participants discussed the types of educational offerings, including massive online open courses (MOOCs), certificate programs, degree-granting programs, boot camps, and individual courses. Participants noted that certificate programs would typically involve a relatively small investment in a student's time, unlike a degree-granting program. One participant proposed a structure consisting of an introductory data science course and three or four additional courses in the three domains (computation, statistics, and visualization). Someone noted that the University of California, Santa Barbara, has similar "emphasis" programs in information technology and technology management. These are sought after because students wish to demonstrate their breadth of understanding. In the case of data science, however, students may wish to use data science to further their domain science. As a result, the certificate model in data science may not be in high demand, inasmuch as students may see value in learning the skills of data science but not in receiving the official recognition of a certificate.

A participant reiterated Joshua Bloom's suggestion made during his presentation to separate data literacy from data fluency. Data fluency would require several years of dedicated study in computing, statistics, visualization, and machine learning. A student may find that difficult to accomplish while obtaining a domain-science degree. Data literacy, in contrast, may be beneficial to many science students and less difficult to obtain. A participant proposed an undergraduate-level introductory data science course focused on basic education and appreciation to promote data literacy.

Workshop participants discussed the importance of coordinating the teaching of data science across multiple disciplines in a university. For example, a participant pointed out that Carnegie Mellon University has multiple master's degree offerings (as many as nine) around the university that are related to data science. Each relevant discipline, such as computer science and statistics, offers a master's degree. The administrative structure is probably stovepiped, and it may be difficult to develop multidisciplinary projects. Another participant argued that an inherently interdisciplinary field of study is not well suited to a degree crafted within a single department and proposed initiating task forces across departments to develop a degree program jointly. And another proposed examining the Carnegie Mellon University data science master's degrees for common topics taught; those topics probably are the proper subset of what constitutes data science.

A workshop participant noted that most institutions do not have nine competing master's programs; instead, most are struggling to develop one. Without collective agreement in the community about the content of a data science program of study, he cautioned that there may be competing programs in each school instead of a single comprehensive program. The participant stressed the need to understand the core requirements of data science and how big data fits into data science.

Someone noted the importance of having building blocks—such as MOOCs, individual courses, and course sequences—to offer students who wish to focus on data science. Another participant pointed out that MOOCs and boot camps are opposites: MOOCs are large and virtual, whereas boot camps are intimate and hands-on. Both have value as nontraditional credentials.

Guy Lebanon stated that industry finds the end result of data science programs to be inconsistent because they are based in different departments that have different emphases. As a result, industry is uncertain about what a graduate might know. It may be useful to develop a consistent set of standards that can be used in many institutions.

Ramakrishnan stated that "off-the-shelf" courses in existing programs cannot be stitched together to make a data science curriculum. He suggested creating a wide array of possible prerequisites; otherwise, students will not be able to complete the course sequences that they need.

## WHAT TO TEACH: CONTENT IN TEACHING BIG DATA

The discussion began with a participant noting that it would be impossible to lay out specific topics for agreement. Instead, he proposed focusing on the desired outcomes of training students. Another participant agreed that the fields of study are well known (and typically include databases, statistics and machine learning, and visualization), but said that the specific key components of each field that are needed to form a curriculum are unknown.

Several participants noted the importance of team projects for teaching, especially the creation of teams of students who have different backgrounds (such as a domain scientist and a computer scientist). Team projects foster creativity and encourage new thinking about data problems. Several participants stressed the importance of using real-world data, complete with errors, missing data, and outliers. To some extent, data science is a craft more than a science, so training benefits from the incorporation of real-world projects.

A participant stated that an American Statistical Association committee had been formed to propose a data science program model for a statistical data science program; it would probably include optimization and algorithms, distributed systems, and programming. However, other participants pointed out that that initiative did not include computer science experts in its curriculum development and that that would alter the emphases.

One participant proposed including data security and data ethics in a data science curriculum.

Several participants discussed how teaching data science might differ from teaching big data. One noted that data science does not change its principles when data move into the big data regime, although the approach to each individual step

may differ slightly. Temple Lang said that with large data sets, it is easy to get mired in detail, and it becomes even more important to reason through how to solve a problem.

Ramakrishnan recommended including algorithms and analysis in computer science. He noted that although grounding instruction in a specific tool (such as R, SAS, or SQL) teaches practical skills, teaching a tool can compete with teaching of the underlying principles. He endorsed the idea of adding a project element to data science study.

## PARALLELS IN OTHER DISCIPLINES

Two examples in other domains that were discussed by participants could provide lessons learned to the data science community.

- *Computational science.* A participant noted that computational science was an emerging field 25 years ago. Interdisciplinary academic programs seemed to serve the community best although that model did not fit every university. The participant discussed specifically how the University of Maryland structured its computational-science instruction, which consisted of core coursework and degrees managed through the domain departments. The core courses were co-listed in numerous departments. That model does not require new hiring of faculty or any major restructuring.
- *Environmental science.* Participants discussed an educational model used in environmental science. An interdisciplinary master's-level program was developed so that students could obtain a master's degree in a related science (such as geography, chemistry, or biology). The program involved core courses, research projects, team teaching, and creative use of the academic calendar to provide students with many avenues to an environmental-science degree.

# References

Borgman, C., H. Abelson, L. Dirks, R. Johnson, K.R. Koedinger, M.C. Linn, C.A. Lynch, D.G. Oblinger, R.D. Pea, K. Salen, M.S. Smith, and A. Szalay. 2008. *Fostering Learning in the Networked World: The Cyberlearning Opportunity and Challenge*. Report of the National Science Foundation Task Force on Cyberlearning. National Science Foundation, Washington, D.C.

Davenport, T.H., and D.J Patil. 2012. Data scientist: The sexiest job of the 21st century. *Harvard Business Review* 90(10):70-76.

Dean, J., and S. Ghemawat. 2004. MapReduce: Simplified data processing on large clusters. *Proceedings of the Sixth Symposium on Operating Systems Design and Implementation.* https://www.usenix.org/legacy/publications/library/proceedings/osdi04/tech/.

Faris, J., E. Kolker, A. Szalay, L. Bradlow, E. Deelman, W. Feng, J. Qiu, D. Russell, E. Stewart, and E. Kolker. 2011. Communication and data-intensive science in the beginning of the 21st century. *OMICS: A Journal of Integrative Biology* 15(4):213-215.

Fox, P., and D.L. McGuinness. 2008. "TWC Semantic Web Methodology." http://tw.rpi.edu/web/doc/TWC_SemanticWebMethodology.

Ferreira, N., J. Poco, H.T. Vo, J. Freire, and C.T. Silva. 2013. Visual exploration of big spatio-temporal urban data: A study of New York City taxi trips. *IEEE Transactions on Visualization and Computer Graphics* 19(12):2149-2158.

Manyika, J., M. Chu, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. Hung Byers. 2011. *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. McKinsey and Company, Washington, D.C.

Mayer-Schönberger, V., and K. Cukier. 2012. *Big Data: A Revolution That Transforms How We Work, Live, and Think.* Houghton Mifflin Harcourt, Boston, Mass.

Mele, N. 2013. *The End of Big: How the Internet Makes David the New Goliath*. St. Martin's Press, New York.

National Research Council. 2013. *Frontiers in Massive Data Analysis*. The National Academies Press, Washington, D.C.

Petigura, E.A., A.W. Howard, and G.W. Marcy. 2014. Prevalence of Earth-like planets orbiting Sun-like stars. *Proceedings of the National Academy of Sciences* 110(48):19273.

President's Council of Advisors on Science and Technology. 2010. *Federally Funded Research and Development in Networking and Information Technology*. Executive Office of the President, Washington, D.C.

Reese, B. 2013. *Infinite Progress: How the Internet and Technology Will End Ignorance, Disease, Poverty, Hunger, and War*. Greenleaf Book Group Press, Austin, Texas.

Schmidt, E., and J. Cohen. 2013. *The New Digital Age: Reshaping the Future of People, Nations and Business*. Knopf Doubleday, New York.

Surdak, C. 2014. *Data Crush: How the Information Tidal Wave Is Driving New Business Opportunities*. AMACOM Books, Saranac Lake, N.Y.

Webb, A. 2013. *Data, A Love Story: How I Gamed Online Dating to Meet My Match*. Dutton, New York.

Wilkinson, L., A. Anand, and R. Grossman. 2005. Graph-theoretic scagnostics. Pp. 157-164 in *IEEE Symposium on Information Visualization*. doi:10.1109/INFVIS.2005.1532142.

Wilkinson, L., A. Anand, and R. Grossman. 2006. High-dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions. *IEEE Transactions on Visualization and Computer Graphics* 12(6):1363-1372.

# Appendixes

# A

# Registered Workshop Participants

Agarwal, Deepak – LinkedIn Corporation
Albrecht, Jochen – Hunter College, City University of New York (CUNY)
Asabi, Faisal – Student / No affiliation known
Bailer, John – Miami University
Begg, Melissa – Columbia University
Bloom, Jane – International Catholic Migration Commission
Bloom, Joshua – University of California, Berkeley
Brachman, Ron – Yahoo Labs
Bradley, Shenae – National Research Council (NRC)
Bruce, Peter – Statistics, Inc.
Buechler, Steven – University of Notre Dame
Caffo, Brian – Johns Hopkins University
Christman, Zachary – Rowan University
Cleveland, Bill – Purdue University
Costello, Donald – University of Nebraska
Curry, James – National Science Foundation
Dell, Robert – Naval Postgraduate School
Dent, Gelonia –Medgar Evers College, CUNY
Desaraju, Kruthika – George Washington University
Dobbins, Janet – Statistics, Inc.
Donovan, Nancy – Government Accountability Office
Dozier, Jeff – University of California, Santa Barbara
Dreves, Harrison – NRC

Dutcher, Jennifer – University of California, Berkeley
Eisenberg, Jon – NRC
Eisner, Ken – Amazon Corporation
Fattah, Hind – Chipotle
Feng, Tingting – University of Virginia
Fox, Peter – Rensselaer Polytechnic Institute
Freire, Juliana – New York University
Freiser, Joel – John Jay College of Criminal Justice
Frew, James – University of California, Santa Barbara
Fricker, Ron – Naval Postgraduate School
Gatsonis, Constantine – Brown University
Ghani, Rayid – University of Chicago
Ghosh, Sujit – National Science Foundation
Glassman, Neal – NRC
Gray, Alexander – Skytree Corporation
Haque, Ubydul – Johns Hopkins University
Howard, Rodney – NRC
Howe, William – University of Washington
Hughes, Gary – Statistics, Inc.
Huo, Xiaoming – Georgia Tech, National Science Foundation
Iacono, Suzanne – National Science Foundation
Kafadar, Karen – Indiana University
Kass, Robert – Carnegie Mellon University
Khaloua, Asmaa – Macy
Kong, Jeongbae – Enanum, Inc.
Lafferty, John – University of Chicago
Lesser, Virginia – Oregon State University
Lebanon, Guy – Amazon Corporation
Levermore, David – University of Maryland
Liu, Shiyong – Southwestern University of Finance and Economics
Mandl, Kenneth – Harvard Medical School Boston Children's Hospital
Marcus, Stephen – National Institute of General Medical Sciences, National Institutes of Health (NIH)
Martinez, Waldyn – Miami University
Mellody, Maureen – NRC
Neerchal, Nagaraj – University of Maryland, Baltimore County
Orwig, Jessica – American Physical Society
Pack, Quinn – Mayo Clinic
Parmigiani, Giovanni – Dana Farber Cancer Institute
Pearl, Jennifer – National Science Foundation
Pearsall, Hamil – Temple University

Perlich, Claudia – Dstillery
Rai, Saatvika – University of Kansas
Ralston, Bruce – University of Tennessee
Ramakrishnan, Raghu – Microsoft Corporation
Ranakrishan, Raghunath – University of Texas, Austin
Ravichandran, Veerasamy – NIH
Ré, Christopher – Stanford University
Ryland, Mark – Amazon Corporation
Schwalbe, Michelle – NRC
Schou, Sue – Idaho State University
Shams, Khawaja – Amazon Corporation
Sharman, Raj –University at Buffalo, State University of New York (SUNY)
Shekhar, Shashi – University of Minnesota
Shipp, Stephanie – VA Bioinformatics Institute at Virginia Tech University
Shneiderman, Ben – University of Maryland
Spencer Huang, ChiangChing – University of Wisconsin, Milwaukee
Spengler, Sylvia – National Science Foundation
Srinivasarao, Geetha – Information Technology Specialist, Department of Health and Human Services
Szewczyk, Bill – National Security Agency
Tannouri, Ahlam – Morgan State University
Tannouri, Charles – Department of Homeland Security
Tannouri, Sam – Morgan State University
Temple Lang, Duncan – University of California, Davis
Torrens, Paul – University of Maryland, College Park
Ullman, Jeffrey – Stanford University
Vargas, Juan – Georgia Southern University
Wachowicz, Monica – University of New Brunswick, Fredericton
Wang, Rong – Illinois Institute of Technology
Wang, Youfa – University at Buffalo, SUNY
Wee, Brian – National Ecological Observatory Network (NEON), Inc.
Weese, Maria – MIA
Weidman, Scott – NRC
Weiner, Angelica – Amazon Corporation
Wynn, Sarah – NRC Christine Mirzayan Science and Technology Policy Graduate Fellow
Xiao, Ningchuan – Ohio State University
Xue, Hong – University at Buffalo, SUNY
Yang, Ruixin – George Mason University
Zhang, Guoping – Morgan State University
Zhao, Fen – National Science Foundation

# B

# Workshop Agenda

**APRIL 11, 2014**

8:30 a.m. **Opening Remarks**

Suzanne Iacono, Deputy Assistant Director, Directorate for Computer and Information Science and Engineering, National Science Foundation

8:40 **The Need for Training: Experiences and Case Studies**

| | |
|---|---|
| Co-Chairs: | Raghu Ramakrishnan, Microsoft Corporation |
| | John Lafferty, University of Chicago |
| Speakers: | Rayid Ghani, University of Chicago |
| | Guy Lebanon, Amazon Corporation |

10:15 **Principles for Working with Big Data**

| | |
|---|---|
| Chair: | Brian Caffo, Johns Hopkins University |
| Speakers: | Jeffrey Ullman, Stanford University |
| | Alexander Gray, Skytree Corporation |
| | Duncan Temple Lang, University of California, Davis |
| | Juliana Freire, New York University |

12:45 p.m. **Lunch**

1:45 **Courses, Curricula, and Interdisciplinary Programs**

| | |
|---|---|
| Chair: | James Frew, University of California, Santa Barbara |
| Speakers: | William Howe, University of Washington |
| | Peter Fox, Rensselaer Polytechnic Institute |
| | Joshua Bloom, University of California, Berkeley |

4:30 **Q&A/Discussion**

**APRIL 12, 2014**

8:30 a.m. **Shared Resources**

| | |
|---|---|
| Chair: | Deepak Agarwal, LinkedIn Corporation |
| Speakers: | Christopher Ré, Stanford University |
| | Bill Cleveland, Purdue University |
| | Ron Brachman, Yahoo Labs |
| | Mark Ryland, Amazon Corporation |

11:15 **Panel Discussion: Workshop Lessons**

| | |
|---|---|
| Chair: | Robert Kass, Carnegie Mellon University |
| Panel Members: | James Frew, University of California, Santa Barbara |
| | Deepak Agarwal, LinkedIn Corporation |
| | Claudia Perlich, Dstillery |
| | Raghu Ramakrishnan, Microsoft Corporation |
| | John Lafferty, University of Chicago |

1:00 p.m. **Workshop Adjourns**

# C

# Acronyms

| | |
|---|---|
| AOL | America OnLine |
| AWS | Amazon Web Services |
| BMSA | Board on Mathematical Sciences and Their Applications |
| CATS | Committee on Applied and Theoretical Statistics |
| CRA | Computing Research Association |
| DARPA | Defense Advanced Research Projects Agency |
| DOE | Department of Energy |
| MOOC | massive online open course |
| NASA | National Aeronautics and Space Administration |
| NIH | National Institutes of Health |
| NITRD | Networking and Information Technology Research and Development |
| NRC | National Research Council |
| NSF | National Science Foundation |
| OCR | optical character recognition |
| RHIPE | R and Hadoop Integrated Programming Environment |