DOI:10.1145/3188721

Data science promises new insights, helping transform information into knowledge that can drive science and industry.

BY FRANCINE BERMAN, ROB RUTENBAR, BRENT HAILPERN, HENRIK CHRISTENSEN, SUSAN DAVIDSON, DEBORAH ESTRIN, MICHAEL FRANKLIN, MARGARET MARTONOSI, PADMA RAGHAVAN, VICTORIA STODDEN, AND ALEXANDER S. SZALAY

Realizing the Potential of Data Science

THE ABILITY TO manipulate and understand data is increasingly critical to discovery and innovation. As a result, we see the emergence of a new field—data science—that focuses on the processes and systems that enable us to extract knowledge or insight from data in various forms and translate it into action. In practice, data science has evolved as an interdisciplinary field that integrates approaches from such data-analysis fields as statistics, data mining, and predictive analytics and incorporates advances in scalable computing and data management. But as a discipline, data science is only in its infancy.

The challenge of developing data science in a way that achieves its full potential raises important questions for the research and education community: How can we evolve the field of data science so it supports the increasing role of data in all spheres? How do we train a workforce of professionals who can use data to its best advantage? What should we teach them? What can government agencies do to help maximize the potential of data science to drive discovery and address current and future needs for a workforce with data science expertise? Convened by the Computer and Information Science and Engineering (CISE) Directorate of

» key insights

- **Data science can help connect previously disparate disciplines, communities, and users to provide richer and deeper insights into current and future challenges.**
- **Data science encompasses a broad set of areas, including data-focused algorithmic innovation and machine learning; data mining and the use of data for discovery; collection, organization, stewardship and preservation of data; privacy challenges and policy associated with data; and pedagogy to support the education and training of data-savvy professionals.**
- **There is a growing gap between commercial and academic research practice for data systems that needs to be addressed.**

the U.S. National Science Foundation as a Working Group on the Emergence of Data Science (https://www.nsf.gov/dir/index.jsp?org=CISE), we present a perspective on these questions with a particular focus on the challenges and opportunities for R&D agencies to support and nurture the growth and impact of data science. For the full report on which this article is based, see Berman et al.[2]

The importance and opportunities inherent in data science are clear (see http://cra.org/data-science/). If the National Science Foundation, working with other agencies, foundations, and industry can help foster the evolution and development of data science and data scientists over the next decade, our research community will be better able to meet the potential of data science to drive new discovery and innovation and help transform the information age into the knowledge age. We hope this article serves as a basis for dialogue within the academic community, the industrial research community, and ACM and relevant ACM special interest groups (such as SIGKDD and SIGHPC).

The Data Life Cycle

Data never exists in a vacuum. Like a biological organism, data has a life cycle, from birth through an active life to "immortality" or some form of expiration. Also like a living and intelligent organism, it survives in an environment that provides physical support, social context, and existential meaning. The data life cycle is critical to understanding the opportunities and challenges of making the most of digital data; see the figure here for the essential components of the data life cycle.

As an example of the data life cycle, consider data representing experimental outputs of the Large Hadron Collider (LHC), an instrument of tremendous importance to the physics community and supported by researchers and nations worldwide. LHC experiments collide particles to test the predictions of various theories of particle physics and high-energy physics. In 2012, data on LHC experiments provided strong evidence for the Higgs Boson, supporting the veracity of the Standard Model of Physics. This scientific discovery was *Science Magazine*'s 2012 "Breakthrough of the Year"[3] and Nobel Prize for Physics in 2013.

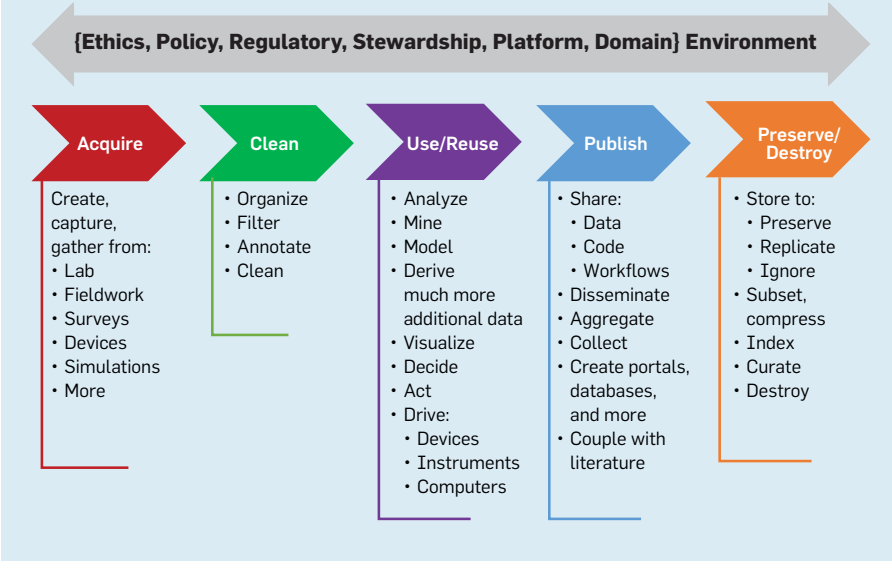
The life cycle of LHC data is fascinating. At "birth," data represents the results of collisions within an instrument carried out in a 17-mile tunnel on the France-Switzerland border. Most of the data generated is technically "uninteresting" and disposed of, but a tremendous amount of "interesting" data remains to be analyzed and preserved. Estimates are that by 2040, there will be from 10 exabytes to 100 exabytes (billion trillion bytes) of "interesting" data produced by the LHC. Retained LHC data is annotated, prepared for preservation, and archived at more than a dozen physical sites. It is published and disseminated to the community for analysis and use at more than 100 other research sites. Critical attention to stewardship, use, and dissemination of LHC data throughout its life cycle has played a key role in enabling the scientific breakthroughs that have come from the experiments.

In addition to development of data stewardship, dissemination, and use protocols, the LHC data ecosystem also provides an economic model that sustainably supports the data and its infrastructure. It is the combination of this greater ecosystem, community agreements about how the data is organized, and political and economic support that allow LHC data to meet its potential to transform our knowledge of physics and enable scientists to make the most of the tremendous investment being made in the LHC's physical instruments and facilities.

The data life cycle diagram outlined in the figure and the LHC example suggest a seamless set of actions and transformations on data, but in many scientific communities and disciplines today these steps are isolated. Domain scientists focus on generating and using data. Computer scientists often focus on platform and performance issues, including mining, organizing, modeling, and visualizing, as well as the mechanisms for eliciting meaning from the data through machine learning and other approaches. The physical processes of acquisition and instrument control are often the focus of engineering, or data as "dirty signals" or as control inputs for other equipment. Statisticians may focus on the mathematics of models for risk and inference. Information scientists and library scientists may focus on stewardship and preservation of data and the "back-end" of the pipeline, following acquisition, decisions, and action in the realm of publishing, archiving, and curation.

There is a significant opportunity for bridging gaps in development of effective life cycles for valuable data within and among the computer science, information science, domain, and physical science and engineering communities, for a start. There is also an opportunity for bridging gaps among machine learning, data analytics, and related disciplines (such as statistics

The data life cycle and surrounding data ecosystem from the *Realizing the Potential of Data Science Report*.[2]



and operations research). Here we focus on some opportunities.

National Data Science Research

Almost every stage of the data lifecycle, as outlined in the figure, provides deep research opportunities. Moreover, an overarching area of opportunity for a national data science agenda is to bridge the gaps in the life cycle, building stronger connections among the computer science, information science, statistics, domain, and physical science and engineering communities, as outlined earlier. That is, a business-as-usual research agenda is likely to strengthen individual technologies behind discrete steps in the data life cycle but unlikely to nurture broader breakthroughs or paradigm shifts that cut across existing disciplinary silos. It is an essential and defining attribute of data ("big" and otherwise) that it can connect previously disparate disciplines, communities, and users to provide richer and deeper insight into current and future challenges.

It is vital to encourage a broader and more holistic view of data as integrating research opportunities across the sciences, engineering, and range of application domains. One such opportunity is to invest in the full data life cycle and surrounding environment—as a central outcome itself, not as a side effect or intermediate step to another desirable outcome. In parallel with development of data science in depth as a core component of computer science, data science should also evolve in breadth to address the needs of domains outside computer science. Our community has a unique opportunity to advance data science, with respect to applying data-driven strategies to individual domain research and cross-domain research opportunities.

A second opportunity involves what might be called "embodied intelligence" scenarios that big data is enabling for the first time. Recent breakthroughs in a range of foundational artificial intelligence and "deep learning" technologies[1] have made it possible to create sophisticated software artifacts that "act intelligently." The key innovations are in mathematical-pattern-recognition techniques that take input from millions of training examples of correct responses to create software systems (soon likely hardware systems as well) able to better recognize images, decode human speech, discover critical patterns in legal and business documents, and more. As engineered artifacts, these artificial intelligence systems are embodied as complex mathematical formulae that are customized to purpose, or "trained," by a truly astounding volume of numerical parameters (such as 10 million for a decent image-classification system today).

These trained decision-oriented models are becoming core components in a range of novel software solutions to complex problems, creating cross-disciplinary challenges.[6] For example, what does it mean for such a component to be "correct" when it is perhaps only 70% accurate? What should the life cycle be for the data used to train and update these models? What are the policy implications (and designation of responsibility) for embodied intelligent agents trained on such data that behave with negative consequences (such as when blamed for an autonomous vehicle that crashes, or by a customer whose account is suspend-

Teaching Data Science: Many Flowers Blooming

To support research and workforce development for data science, we must determine how—and, interestingly, *where* in the institution—it should be taught. Much as the emergence of computer science in the 1960s created the first organizational units and degrees dedicated to computing in modern universities, the rise of data science is driving a range of interesting curricular experiments. For a sense of the rapidly evolving landscape, consider these five:

University of California. The University of California, Berkeley, Data Science Education Program[8] is part of its recently established Division of Data Sciences, at the same level as Berkeley's colleges and schools, integrating with them. The introductory class provides a foundation for students in all fields to engage with data and creates pathways to advanced-level work. The foundational course combines instruction in core computational and statistics concepts while enabling students to work with real data in a range of fields. It is designed to be accessible to undergraduates of any intended major without prior experience. A set of connector courses (mostly taken simultaneously with the foundational course) enable them to apply core skills from the foundational course to real-world issues that relate to their areas of interest. There are also advanced courses, including an upper-division integrative course called "Data 100 Principles and Techniques of Data Science."

University of Michigan. The University of Michigan Undergraduate Program in Data Science[11] is a new major offered as a joint program by the Electrical Engineering and Computer Science and Statistics Departments. The data science major is a rigorous program focusing on aspects of computer science, statistics, and mathematics relevant for analyzing and manipulating large datasets. It can be entered from either the College of Engineering or the College of Literature, Science, and Arts.

Columbia University. The Columbia University Data Science Institute's Master of Science in Data Science[4] offers a professional master's degree to students with any undergraduate degree that includes suitable quantitative prior coursework. It starts from a set of four foundational courses (that can be taken independently to yield a data science certificate), focusing on algorithms, probability/statistics, machine learning, and visualization.

University of Illinois. The University of Illinois at Urbana-Champaign Master of Computer Science in Data Science degree[10] is offered as an online professional master's in computer science available on the Coursera massive open online course platform.[5] The degree seeks to create a global gateway into the discipline. The program builds expertise in four core areas of computer science—data visualization, machine learning, data mining, and cloud computing—and also offers courses in collaboration with the university's Statistics Department and School of Information Sciences. This collaboration specifically strives to cover the full data life cycle, including its mathematical, computational, and curation and stewardship components, in an integrated and comprehensive fashion.

University of Chicago. The University of Chicago Master of Science in Computational Analysis and Public Policy program[9] is offered jointly by the Department of Computer Science and the Harris School of Public Policy. As government decision making is increasingly data-driven, data use, data sharing, transparency, and accountability become increasingly important issues from both a public policy and a technological perspective. The program focuses on the intersection of policy and computer science. Students take courses across both areas, preparing them to make meaningful contributions to the design, implementation, and rigorous analysis of policies in the public sector.

ed inappropriately based on an automatic inference)? Software engineering, as a discipline, is challenged by such imprecision and with versioning and testing of the enormous data components—gigabyte-to-terabyte scale training data—for these systems. Existing notions of model verification/validation seem woefully insufficient. And the policy, stewardship, and curation questions go largely unasked and unanswered.

Note that the existence of predictive models is not unique to machine learning; for example, statistical models have been used in epidemiology, and physical models are common in weather prediction and nuclear simulations. The "training" aspect for data science may be novel in the context of the software engineering of solutions, in that the resulting models may lack the guarantees associated with statistical power and sample-size calculations.

Yet another opportunity is to address the growing gap between commercial and academic research practice for data systems at the edge of the state of the art. Much has been made of the increasing "reverse migration" of strong academic researchers into data-rich enterprises (such as Facebook, Google, and Microsoft). While this is likely good for the U.S. national economy in the near term, it is worrisome for the future of discovery-based open research, education, and training in the academic sector. In addition to the challenges of attracting sponsored research funding, another reason for the "brain drain" from the research community into the private sector may be declining infrastructure-support environments, including the sparsity of large datasets and adequate infrastructure in academia that support data science research at scale. When the best infrastructure environment for cutting-edge research is consistently in the private sector, the opportunity for innovation in the public sector deteriorates. Government support for strategic and committed public-private partnerships that build adequate and representative at-scale infrastructure in the academic community for researchers can unlock innovation in academic research and ultimately support the private sector through development of a more sophisticated, educated, better-trained workforce.

National Data Science Education and Training

Higher-education institutions across the U.S. recognize that data science is a critical skill for 21st-century research and a 21st-century workforce. In higher education, data science curricula have two audiences: new professionals in data science, and scientists and professionals who need data science skills to contribute to other fields. Data science curricula in higher education often focus on both, the same way curricula in computer science departments educate computer science students and provide training in computer skills to students from other disciplines to promote computer literacy.

It is important to note that, at present, there is no single model of which department, school, or cross-unit collaboration within higher-education institutions should have the responsibility for data science education and training. Data science programs are being sited in departments and schools of computer science, information science, statistics, and management. Many of the most successful, particularly at the undergraduate level, represent university-wide coalitions frequently sponsored by interdisciplinary institutes, rather than by a particular department or school. There is thus no common agreement as to where data science should "live" in the institution, though there is much interesting experimentation at this point (see the sidebar "Teaching Data Science" for several programmatic configurations). Note that when a university chooses to house "data science" in an existing department or college, it implicitly adopts the standards and culture of that existing organization. In contrast, when a university introduces "data science" as an interdisciplinary function, it confronts the heterogeneity of the new field up front but will likely deal with additional administrative overhead associated with a cross-organizational entity. We focus on trends in both data science education and training in the following paragraphs.

Educational curricula in data science have yet to "standardize" and appear today with many interesting course configurations. In general, data scientists are expected to be able to analyze large datasets using statistical techniques, so statistics and modeling are typically part of required coursework. Moreover, a comprehensive data science curriculum is more than machine learning and statistics, possibly including courses on programming, data stewardship, and ethics, in addition to other areas. Data scientists must be able to find meaning in unstructured data, so classes on programming, data mining, and machine learning are often part of the core. Data scientists must also be able to communicate their findings effectively, so courses on visualization may be offered, at least as an elective. In recognition of the challenges that arise from misuse of data and incorrect conclusions drawn from data, ethics is also becoming a part of responsible curricula for the field.

Other courses that appear either in the core or as an elective in various programs include research design, databases, algorithms, parallel computing, and cloud computing, all of which reflect skills an employer might expect from a data scientist. Many programs also require a capstone project that gives students experience in working through real-world problems in teams in a particular domain. Data science courses are also becoming a staple of quality online programs.

A strong data science curriculum requires faculty with appropriate expertise and engagement with the field. The pull of faculty with expertise in data science and related fields away from academia and toward industry creates a challenge for educational institutions in mounting such programs. It also presents a potential challenge to development of data science as a formal discipline.

To combat this trend, the Moore and Sloan Foundations in 2013 created a joint $38 million project, the Moore-Sloan Data Science Environments, to fund initiatives to create "data science environments,"[7] addressing challenges in academic careers, education and training, tools and software, reproducibility and open science, physical and intellectual space, and data science studies. This funding has been transformational, providing critical "worked examples" of data science programs useful for current and future efforts.

From the current diversity of curricula and programs, data science is going through an important and healthy period of experimentation. It is important

that we do not "standardize" data science too quickly, continuing to explore configurations of courses, areas, projects, faculty, and partnerships to gain critical experience in how to best educate new generations of data scientists.

In addition to "data science" programs and majors that serve to evolve data science as a discipline, data science skills are increasingly critical as training for other disciplines and professions as they become more and more data-enabled. Effective training will empower data-enabled professionals and domain scientists to utilize data effectively and operate within a broader data-driven environment, develop an appreciation of what data can tell us and what it cannot, acquire appropriate technical knowledge about how data should be handled, gain awareness that correlation in data does not necessarily imply causality, and begin to develop a sense of responsible methodologies and ethical principles in the use of data.

More specific training in the nuts and bolts of dealing with data is also critical for various data-driven professions. Training in programming and software engineering is useful for students who will be using data-driven simulations and models in their research. Training in version control and the subtleties of stewardship, including working with repositories for data and software, should be taught to computational researchers. And training in best practices for digital scholarship and reproducibility should be integrated into research-methodology curricula. The ethics of using (and misusing) data should be incorporated into all training programs to promote effective and responsible data use. Courses teaching these skills can be made available in a variety of venues, from university courses and modules to online courses to professional courses that could be developed by scientific societies and communities.

At present, there is no single model of which department, school, or cross-unit collaboration within higher-education institutions should have the responsibility for data science education and training.

Data Science Research and Education Infrastructure

Any innovative agenda in data science research and education will depend on a foundation of enabling data infrastructure and useful datasets. Research in data science needs access to sufficiently large and numerous datasets to illuminate and validate results. The datasets must be available for reproducible research and hosted by reliable infrastructure.

Lack of such infrastructure and datasets will inhibit success. Education and training in data science is most authentic in a setting where students can work on data that represents the datasets and environments they will see in the professional arena; that is, data that is both at-scale and embedded in a stewardship infrastructure that enables it to be a useful tool in analysis, modeling, and mining.

In the best case, data infrastructure should support access to data for research and education that is equivalent to access to any other key utility; it must be "always on," it must be robust enough to support extensive use, and the quality must be good. In the world of data, this comes down to responsible stewardship, meaning there must be actors, plans, and both "social" and technical infrastructure to ensure the following:

Data is appropriately tracked, monitored, and identified. Who created, curated, and used the data? Can it be persistently identified? Are there adequate privacy and security controls?;

Data is well cared for. Who is committed to keeping it, in what formats, and for how long? Who is committed to funding data stewardship? And how will it be stored and migrated to next-generation media?;

Data is discoverable and useful. How is data made available and to whom? What services are needed to make good use of it? And what metadata and other information is needed to promote reproducibility?; and

Data stewardship is compliant with policy and good practice. Does stewardship comply with community standards and appropriate policy regarding reporting, intellectual property, and other concerns? Are the rights, licenses, and other properties that will determine appropriate use clear? And what data and metadata are to be kept, who owns it and its by-products, and who has access to it and its metadata or parts of it?

Since data will become the core for research and insight for a broad set of academic disciplines, access to it in a usable form on a reasonable time scale becomes the entry point for any effective research and education agenda.

Government R&D agencies (such as the National Science Foundation) have an opportunity to ensure the lack of adequate data infrastructure does not present a roadblock to innovative research and educational programs.

Developing and sustaining the infrastructure that ensures that research data is available to the public and accessible for reuse and reproducibility requires stable economic models. While there is much support for the development of tools, technologies, building blocks, and data-commons approaches, few U.S. federal programs directly address the resource challenges for data stewardship or provide help for libraries, domain repositories, and other stewardship environments to become self-sustaining and address the need for public access.

While the U.S. federal government cannot take on the entire responsibility for stewardship of sponsored research data and its infrastructure, neither should it shy away from providing seed or transition funding for institutions and organizations to develop sustainable stewardship options for the national community. We encourage the community, inside and outside of government, to support the development and piloting of sustainable data stewardship models for data-driven research and data science education through strategic programs, guidance, and cross-agency and public-private partnerships. Science-centric government agencies like the National Science Foundation should coordinate with peer agencies like the National Institutes of Health that focus on similar issues to leverage investments and provide economies of scope and scale.

Realizing the Potential

The research, education, and infrastructure discussions here focus on developing a foundation that can increase the pool of data scientists and data-literate professionals to meet the current and near-term challenges of data-driven efforts in all sectors, as well as the need to evolve data science as a discipline that can meet the challenges of future data-driven scenarios.

Data is everywhere, providing an increasingly important tool for a broad spectrum of endeavors. As systems grow "smarter" and take on more autonomous and decision-making capabilities, we will increasingly face data science technical challenges and the social challenges of governance, ethics, policy, and privacy. Addressing them will be critical to rendering data-driven systems useful, effective, and productive, rather than intrusive, limiting, and destructive. Such solutions will be particularly important in highly data-driven environments like the Internet of Things. Moreover, as fundamental computational platforms change in response to the looming end of Moore's Law scaling of semiconductors,[12] there will be tremendous opportunities to reimagine the entire hardware/software enterprise in the light of future data needs.

Conclusion

Our community must be prepared to deal with future scenarios by encouraging the initial research that lays the groundwork for innovative uses of data, well-functioning data-focused systems, useful policy and protections, and effective governance of data-driven environments. With both programmatic resources and a platform for community leadership, federal R&D agencies like the National Science Foundation play an important role in guiding the community toward innovation. Attention to deep efforts needed to expand the field and its impact, as well as broad efforts to help data science reach its potential for transforming 21st-century research, education, commerce, and life, are needed.

Acknowledgments

We would like to thank the National Science Foundation for convening this group and the institutions and organizations of the co-authors for their support for this work.

References

1. Bengio, Y., LeCun, Y., and Hinton, G. Deep learning. *Nature 521* (May 28, 2015), 436–444.
2. Berman, F. (co-chair), Rutenbar, R. (co-chair), Christensen, H., Davidson, S., Estrin, D., Franklin, M., Hailpern, B. , Martonosi, M., Raghavan, P., Stodden, V., and Szalay, A. *Realizing the Potential of Data Science: Final Report from the National Science Foundation Computer and Information Science and Engineering Advisory Committee Data Science Working Group*. National Science Foundation Computer and Information Science and Engineering Advisory Committee Report, Dec. 2016; https://www.nsf.gov/cise/ac-data-science-report/CISEACDataScienceReport1.19.17.pdf
3. Cho, A. The discovery of the Higgs Boson. *Science 338*, 6114 (Dec. 21, 2012), 1524–1525.
4. Columbia University Data Science Institute. Master of Science in Data Science; http://datascience.columbia.edu/master-of-science-in-data-science
5. Coursera. Master of Computer Science in Data Science; https://www.coursera.org/university-programs/masters-in-computer-data-science
6. Dhar, V. When to trust robots with decisions, and when not to. *Harvard Business Review* (May 17, 2006); https://hbr.org/2016/05/when-to-trust-robots-with-decisions-and-when-not-to
7. Moore-Sloan Data Science Program; http://msdse.org/
8. University of California, Berkeley. Data Science Education Program; http://data.berkeley.edu/data-science-education-program
9. University of Chicago. Master of Science in Computational Analysis & Public Policy; https://capp.uchicago.edu/
10. University of Illinois, Urbana-Champaign, CS@ILLINOIS. Master of Computer Science in Data Science, Data Science Track; http://www.cs.uiuc.edu/academics/graduate/professional-mcs-program/mcs-data-science-track
11. University of Michigan. Undergraduate Program in Data Science; https://www.eecs.umich.edu/eecs/undergraduate/data-science/
12. Waldrop, M.M. The chips are down for Moore's Law. *Nature 530*, 7589 (Feb. 11, 2016), 144–146.

Francine Berman (bermanf@rpi.edu) is the Edward P. Hamilton Distinguished Professor in Computer Science at Rensselaer Polytechnic Institute, Troy, NY, USA, and Chair of the Research Data Alliance / U.S. She served as Co-Chair of the Data Science Working Group of the NSF CISE Advisory Committee.

Rob Rutenbar (rutenbar@pitt.edu) is a professor of computer science and electrical and computer engineering and Senior Vice Chancellor for Research at the University of Pittsburgh, Pittsburgh, PA, USA. He served as Co-Chair of the Data Science Working Group of the NSF CISE Advisory Committee.

Henrik Christensen (hichristensen@ucsd.edu) is a professor of computer science and Director of the Institute for Contextual Robotics at the University of California at San Diego, USA.

Susan Davidson (susan@cis.upenn.edu) is the Weiss Professor of Computer and Information Science at the University of Pennsylvania, Philadelphia, PA, USA.

Deborah Estrin (destrin@cs.cornell.edu) is Associate Dean and professor of computer science at Cornell Tech in New York City and a professor of public health at Weill Cornell Medical College, New York, USA.

Michael Franklin (mjfranklin@uchicago.edu) is the Liew Family Chairman of Computer Science and Senior Advisor to the Provost for Data and Computing at the University of Chicago, USA.

Brent Hailpern (bth@us.ibm.com) is a Distinguished Research Staff Member, Science Director of the IBM Cognitive Horizons Network, and Head of Computer Science for IBM Research, San Jose, CA, USA.

Margaret Martonosi (mrm@princeton.edu) is the Hugh Trumbull Adams '35 Professor of Computer Science at Princeton University, Princeton, NJ, USA.

Padma Raghavan (padma.raghavan@vanderbilt.edu) is a professor of computer science and computer engineering and Vice President of Research at Vanderbilt University, Nashville, TN, USA.

Victoria Stodden (vcs@illinois.edu) is an associate professor in the School of Information Sciences at the University of Illinois at Urbana-Champaign, USA.

Alex Szalay (szalay@jhu.edu) is Bloomberg Distinguished Professor in the Departments of Physics and Astronomy and Computer Science at the Johns Hopkins University, Baltimore, MD, USA.

Watch the authors discuss their work in this exclusive *Communications* video. https://cacm.acm.org/videos/realizing-the-potential-of-data-science