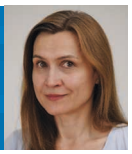
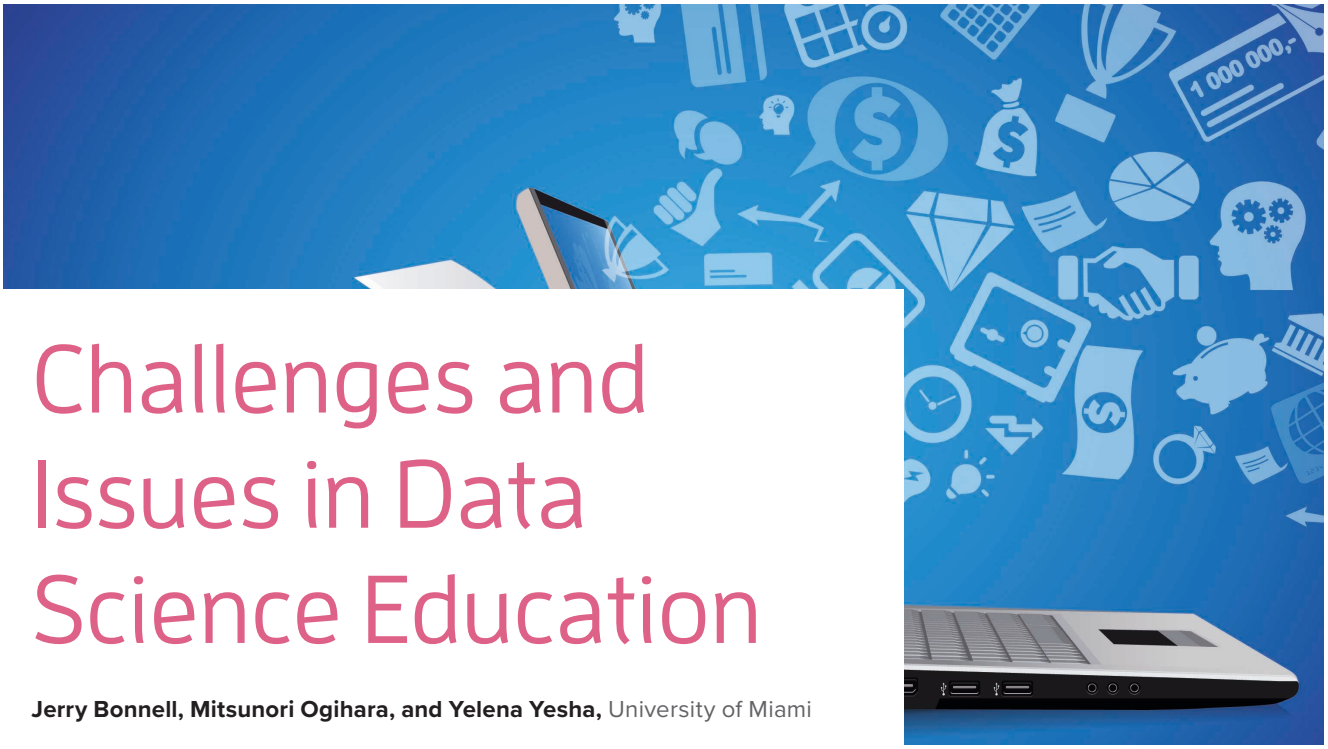
EDITOR IRENA BOJANOVA
NIST; irena.bojanova@computer.org





Challenges and Issues in Data Science Education

Jerry Bonnell, Mitsunori Ogihara, and Yelena Yesha, University of Miami

Data science is one of the fastest growing fields because of its ability to address analytic needs. However, the supply of data scientists has remained limited. Academic institutions must adapt to meet demand through programs to train future data scientists.

Data science is a burgeoning discipline that aims to develop methodologies and tools for analyzing large data sets to uncover insights that further research goals and facilitate decision making. Since its beginnings in the 1990s, the discipline has had origins in data mining and in the 2000s, with advances in computing, influences from big data analytics. Indeed, data, information collection, and the knowledge discovery pipeline describe what daily work in the field looks like. However, data science is also an interdisciplinary effort that does not exist independently or in isolation; instead, it should rightly be defined by its fruitful exchange of multiple knowledge areas: mathematics, statistics, computer science, communication, and relevant application domains.

C.F. Jeff Wu, in his 1985 paper, first used the term *data science*.[1] Later, in his 1997 inauguration speech as the H.C. Carver Chair at the University of Michigan, Wu suggested that statistics should be equivalent to "data science" and emphasized the need for large-scale computing and interdisciplinary training. With the recent and ongoing proliferation of data across all facets of life—generated by social media, smart devices, streaming services, real-time systems, and more—the desire to involve computation to effectively analyze data sets at scale to reduce uncertainty in decision making distinguishes data science today from its previous incarnations.

WHAT SKILL SETS ARE NECESSARY TO BE A SUCCESSFUL DATA SCIENTIST?

The analytic lifecycle in data science can be highly complex and broad, as it caters to specific domains, but one

Digital Object Identifier 10.1109/MC.2021.3128734
Date of current version: 14 February 2022

might extract the following, which may occur repeatedly, as the basic process. These steps do not always proceed linearly and, instead, are best viewed as an iterative and cyclical process:

1. *Identifying the domain-specific problem at hand*: What is the question that is being asked in the domain?
2. *Interpreting the problem as a mathematical, statistical, and computational problem*: What is an interpretation of the problem? Is it model development, optimization, simulation, classification, database development, or something else?
3. *Identifying data sets to use for solving the problem and collecting the data*: What data sets are necessary to address the question, and how large are they? What is their location—an on-site data server or a cloud server? Are there endpoints for accessing the data, and is the information available with open or closed access? What are the costs and rate limits associated with querying the data?
4. *Cleaning the data*: How do you preprocess and tidy the data, remove errors, handle missing values, and transform the information into a form that is suitable for downstream analyses?
5. *Establishing a solution*: What approaches are promising to solve the problem? Are there tools, models, algorithms, and codes for answering the question? Is it necessary to integrate them? Is it necessary to develop new methods?
6. *Conducting the analysis*: What does the application of the method reveal? Are the results satisfactory? How reliable are they? Do they suggest the need for further analysis at any stage of the lifecycle and, if so, how should these changes be coordinated and iterated?
7. *Visualizing and communicating insights*: What are possible interpretations of the results in the problem domain? What is the best way to visualize the outcomes and communicate the approach to others? How can the extracted knowledge best be communicated to facilitate informed and meaningful conclusions to be used in decision making?

The supply of data scientists, let alone those with the necessary domain knowledge, is limited and, consequently, the gap between supply and demand is widening.

The core skill set of data science must align well with data analysis processes. It includes the following:

- Basic information technology, for example, command line environments, remote login, and version control
- Programming language skills, for example, integrated development environments, software packaging (such as Anaconda), and coding in specific programming languages (such as R, Python, Julia, and MATLAB)
- Calculus and linear algebra, for instance, convergence, derivatives, integrals, vectors, and matrices
- Discrete mathematics, such as logic and inference, graphs, and combinatorics
- Data structures and algorithms, for instance, time complexity, search, and dynamic programming
- Statistical analysis, including hypothesis testing, statistical inferencing, bootstrapping, regression analysis, and experimental design
- Machine learning and deep learning, for instance, supervised and unsupervised learning, deep neural networks, generative models, and federated learning
- Databases and data warehousing, for example, Structured Query Language, online analytical processing, and distributed databases
- Simulation and optimization, including Markov chain Monte Carlo simulations, queuing theory, simulated annealing, gradient descent, and Newton's method
- Text, visual, and stream data, for example, the Fourier transform, wavelets, n-grams, word2vec, Bidirectional Encoder Representations From Transformers, object recognition, machine translation, and latent semantic analysis
- Data visualization such as categorical and numerical variables, the layered grammar of graphics, scatter plots, bar charts, violin plots, histograms, and geographical mapping
- Domain-specific knowledge, for instance, adjustment of gained knowledge to domains.

HOW LARGE IS THE SUPPLY–DEMAND GAP?

Applications of data science, and the desire to use data science, continue to rise. Unfortunately, the supply of data scientists, let alone those with the necessary domain knowledge, is limited and, consequently, the gap between supply and demand is widening.[2,3] The National Academy "Roundtable on Data Science Postsecondary Education" report[4] states that society has been generating information at an

unprecedented rate, prompting development of big data technologies and propelling demand for data skills.

HOW CAN THE PROBLEM BE SOLVED?

Despite the urgency for data science in research and industry, the domain continues to experience an absence of an effective, comprehensive educational plan that trains the existing workforce and present and prospective students for the field. Current efforts respond to the need in different forms: massive online open courses (MOOCs) (for example, Coursera), online courses (for instance, Linda.com), and degree/certificate programs (see the "Degree Programs in Data Science" section). These resources have been made widely available and are intended for learners with different backgrounds and time commitments, and yet the demand continues to exceed the supply of practitioners.

We identify three principal reasons for the deficiency in data science education. First, MOOCs experience high attrition rates despite allowing learners to progress using an individualized study plan; the time available for the workforce to participate in outside-work learning is also limited. Second, existing degree programs do not cater to all application areas, and time commitments for successful completion may be too demanding for people in the workforce. Third, a shortage of data science educators remains an obstacle to outreach and exposure to the field.

How can this gap be best addressed? For data science learning programs, there is no "one-size-fits-all" solution because students' educational backgrounds are diverse. We must consider using standard, formal courses in addition to alternate and more flexible modalities to cater to needs among different backgrounds: microcourses, tutorials, workshops, hackathons, bootcamps, and internships. One can combine these to build a variety of curricula. Microcourses, workshops, and tutorials are useful for workforce training; hackathons are valuable for advocacy, and internship opportunities benefit degree programs and can be integrated as part of the curriculum to enrich learning and provide real-world experience.

DEGREE PROGRAMS IN DATA SCIENCE

Perhaps the most critical component to the promotion of data science is the development and deployment of dedicated degree programs for university students. Dozens of U.S. institutions have begun this work and now offer B.S. degrees in data science, including Columbia University, Colorado State University, Iowa State University, Northeastern University, The Ohio State University, the University of Rochester, and Yale University. Also, M.S. degrees in data science have been made available at schools including Carnegie Mellon University; Georgia Tech; Northwestern University; Stanford University; Syracuse University; the University of California, Berkeley (UC Berkeley); the University of Miami; and the University of Washington. These programs aim to teach students fundamental knowledge and skills, give exposure to domain-specific applications, and, in many cases, provide opportunities through internships.

Perhaps the most critical component to the promotion of data science is the development and deployment of dedicated degree programs for university students.

BASIC DATA SCIENCE TRAINING FOR COLLEGE STUDENTS

As the importance of data science grows in society, a question arises: Have data science skills become so indispensable that every college student, regardless of major, must gain some basic mastery of the field? UC Berkeley, in addition to its data science degree programs, offers an interdisciplinary introductory course (Data 8) that enables students to gain a substantial amount of knowledge that can be adapted to any target domain of interest; the course has been highly influential and continues to have adaptations in the curricula of other universities, e.g., the University of Miami's Data Science for the World for freshmen. Data 8 and its adaptations have few or no prerequisites and introduce students to programming fundamentals and statistical inference. They also draw examples from real-life data so that they are relevant to students.

ADVOCACY

A prerequisite to growing a new generation of data scientists is to provide as much exposure as possible to the field and its practices. This, in turn, offers the best potential for attracting students to degree programs and, ultimately, careers in data science. Therefore, we envisage a thick pipeline that targets students at every step in their academic career, from high school to the completion of undergraduate and graduate programs. We would like to see many of these students who are interested in data science have exposure to the discipline through different formats and modalities. Advocacy campaigns should begin as early as possible to ensure that enough interest is generated: guest lectures to high-school students led by data scientists in industry and university professors of data science, career fairs and venues that promote jobs in data science, and hackathons that invite hands-on experience with data science work and opportunities for networking with practitioners.

TRAINING THE TRAINERS

To respond to the growing demand in data science, we need many educators. The basic skills in data science come from a combination of computer science, mathematics, and statistics.[5] To address the demand for data scientists, educators must use every available opportunity. Data science education should be made available at all levels and include technical and domain-specific training.

All the disciplines that utilize data science must incorporate training and encouragement for students to join the education force into their curricula.

Naturally, we entrust these disciplines to produce data science teachers, but the educators graduating from degree programs in these disciplines may be too few. Thus, we must take a more drastic approach. All the disciplines that utilize data science must incorporate training and encouragement for students to join the education force into their curricula. Education schools primarily produce teachers equipped with knowledge for leading general science classes and laboratory sessions. However, in teacher training, the use of computing—and especially how to teach computing—remains on the periphery of the curriculum. Because computing is a major component of data science, with perhaps the steepest learning curve, we suggest that a critical part of meeting the demand is to involve computation more aggressively in the curricula of education and scientific education disciplines so that the next generation of teachers becomes conversant with data and the use of computing tools when incorporating this knowledge into the classroom. Equally important is the training of data science educators.

REFERENCES

1. C. F. J. Wu, "Future directions of statistical research in China: A historical perspective," *Appl. Statist. Manage.*, vol. 1, pp. 1–7, 1986. doi: 10. 13860 /j . cnki . sl tj . 1986. 01. 001.
2. T. Davenport, "Beyond unicorns: Educating, classifying, and certifying business data scientists," *Harvard Data Sci. Rev.*, vol. 2, no. 2, pp. 1–3, 2020, doi: 10.1162/99608f92.55546b4a.
3. T. H. Davenport and T. D. Patil, "DataScientist: The sexiest job of the 21st century," *Harvard Business Review*, 2012. [Online]. Available: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century
4. The National Academies of Engineering, Science, and Medicine. *Roundtable on Data Science Postsecondary Education*. Washington, DC, USA: National Academies Press, 2020.
5. A. F. Wise, "Educating data scientists and data literate citizens for a new generation of data," *J. Learn. Sci.*, vol. 29, no. 1, pp. 165–181, 2020, doi: 10.1080/10508406.2019.1705678.

JERRY BONNELL is a senior Ph.D. student of computer science at the University of Miami, Coral Gables, Florida, 33146, USA. Contact him at j.bonnell@miami.edu

MITSUNORI OGIHARA is a professor of computer science, the director of the master's of science in data science, and the director of education and workforce development in the Institute for Data Science and Computing, University of Miami, Coral Gables, Florida, 33146, USA. Contact him at m.ogihara@miami.edu.

YELENA YESHA is a professor of computer science; the John S. and James L. Knight Foundation Endowed Chair of Data Science and Artificial Intelligence (AI); the director of the Institute for Data Science and Computing AI, and Machine Learning Program; and the innovation officer of the Institute for Data Science and Computing, University of Miami, Coral Gables, Florida, 33146, USA. Contact her at yxy806 @miami.edu.