CRA
Computing Research Association

Computing Research and the Emerging Field of Data Science

By CRA's Committee on Data Science: Lise Getoor (Chair), David Culler, Eric de Sturler, David Ebert, Mike Franklin, and H. V. Jagadish on behalf of the CRA Board

---

Our ability to collect, manipulate, analyze, and act on vast amounts of data is having a profound impact on all aspects of society. This transformation has led to the emergence of *data science* as a new discipline[1]. The explosive growth of interest in this area has been driven by research in social, natural, and physical sciences with access to data at an unprecedented scale and variety, by industry assembling huge amounts of operational and behavioral information to create new services and sources of revenue, and by government, social services and non-profits leveraging data for social good. This emerging discipline relies on a novel mix of mathematical and statistical modeling, computational thinking and methods, data representation and management, and domain expertise. While computing fields already provide many principles, tools and techniques to support data science applications and use cases, the computer science community also has the opportunity to contribute to the new research needed to further drive the development of the field. In addition, the community has the obligation to engage in developing guidelines for the responsible use of data science.

Data science starts with a strong set of foundations adapted from several fields including statistics, mathematics, social science, natural sciences, and computer science. Already, virtually all aspects of traditional computer science research have played a role in the development of data science. And looking forward, **data science will drive fundamentally new computing research**.

1. From a data management perspective, **data science requires a much deeper understanding and representation of** how **data** is acquired, stored and accessed. Data lineage, data quality, quality assurance, data integration, storage, privacy, and security all need to be rethought. The traditional approach of acquisition, followed by storage, and processing often does not work for high rate or sensitive data.
2. From a computational point of view, **very large data volumes, very high data rates, and very large numbers of users, demand new systems and new algorithms.** New system architectures that can accommodate the heterogeneity and irregular structure in data access and communication are needed. From an algorithmic perspective, there is a need for sublinear algorithms, online algorithms that support real-time data streams, and probabilistic and stochastic approaches to accommodate both scale and noise in the data.

---

[1] We use the term data science in its broadest sense, including data collection, data engineering, data analytics and data architecture.

3. **Furthermore, many classic statistical assumptions and machine learning techniques do not fit current data science needs**. Often derived from natural sources, data is increasingly likely to be biased, incomplete and highly heterogeneous. Systematic errors arising in automated data collection and semantic inconsistencies that result from stitching data together from multiple sources across longer time horizons present profound modeling challenges and opportunities for the development of new statistical methods and machine learning algorithms. Even in the small data setting, new techniques that can cope with heterogeneity and biased sampling are needed. While predictive modeling is important, many data science problems involve decision making, and the ability to reason about alternate courses of action is needed. In addition, understanding the curse of dimensionality, overfitting, and causality in these complex settings is critical.
4. The challenges in scale and heterogeneity also fundamentally change **how users interact with data and models,** how the data is visualized, what algorithms are needed to support understanding and interpretation of the results of data science models, how decisions are made, and how user feedback is acquired and incorporated. Human computer interaction and visual analytics will need to be more tightly integrated with data science models and algorithms. New use cases for natural language processing, speech, computer vision and other human-machine communication modes will emerge.
5. Because data science systems are often embedded in operational systems with changing demands and distributions, **supporting the entire data science lifecycle is important.** Ensuring the robustness of all aspects of the pipeline is important. New software engineering and computer programming best practices will need to be developed. Additionally, data artifacts will often persist beyond their initially planned usage, so longer-term curation and management must also be addressed.

The above research topics, and many others, will require foundational research into systems, computation and machine intelligence.

Furthermore, like colleagues in many other fields, **computing researchers are increasingly becoming users of data science**, as many subareas of computer science, including computer architecture, networking, software engineering, vision, robotics, education and user modeling, are becoming increasingly data-driven. Good empirical methodology is needed to ensure value and reproducibility, including proper data curation, rigorous system modeling, measurement and analysis, and sound methods for the presentation and interpretation of results. It is becoming increasingly important to train all computing researchers in basic data science skills.

Looking more broadly, data science provides new opportunities for creative collaborations between industry, academia and government for pure and applied research. In addition to sponsorship of research, industry partners can provide valuable insights to realistic problems, access to data, capacity to test theories at scale and in the wild, and complementary ways of seeking solutions. Academia, in turn, can provide innovative solutions and software, novel algorithms, and principled analyses of alternative approaches. Academia will also educate a cadre of well-trained data scientists to meet industry needs and help industrial partners explore cutting edge

research. These partnerships will also help inform the data science policy issues related to bias, data privacy, intellectual property, appropriate use, and regulatory issues. Open data initiatives and the open source software movement are particularly well suited for data science and can help smooth the path to commercialization and impact. In short, industry, academic and government data science collaborations will help drive new models for working together.

Finally, while data science offers many new opportunities for improving scientific inquiry and decision-making through increased utilization of data, these uses also offer new challenges. Both the context in which data is generated, and the application(s) for which it will be used are immensely important for accurate, fair and ethical data science. These data science efforts will require collaboration among subfields of computer science and between computer science and other disciplines. Both intradisciplinary as well as interdisciplinary skills need to be taught in order to help support this. As data generation and collection become ubiquitous, concepts of data ownership are evolving as well, and many legal and policy issues will need to be rethought in that context. **In order to understand how to use and share data ethically and responsibly, computer scientists will need to engage with domain scientists, policy makers, and ethicists** to understand the risks and assumptions being made. For example, understanding the social science behind the data science is important when answering questions about individuals and society (e.g., in education, economic policy, and policing). Important concerns include privacy, fairness, and transparency. In order to engage and contribute as productively as possible in emerging policy discussions around data science, computing researchers will need to develop new methods that are able to incorporate ethics, fairness and responsibility.

In summary, the computing research community has a unique opportunity to help define and shape the emerging field of data science. Together with statisticians, mathematicians, social scientists, data analysts, domain scientists and subject matter experts, computer scientists can develop the new theoretical foundations, algorithmic principles and systems upon which the foundations of data science will be built. The Computing Research Association is committed to supporting computing professionals and others in developing ethical and responsible data science research.