What is Biomedical Data Science and Do We Need an Annual Review of It?

We are pleased to bring you the first volume of the *Annual Review of Biomedical Data Science*. It spans a range of biological and medical research challenges that are data intensive and focused on the creation of novel methodologies to advance biomedical science discovery. The term "data science" describes expertise associated with taking (usually large) data sets and annotating, cleaning, organizing, storing, and analyzing them for the purposes of extracting knowledge. It merges the disciplines of statistics, computer science, and computational engineering. Many are irritated by the term—all of science depends ultimately on data, and many of the activities listed above sound like engineering (which is about solving problems) and not science (which is about discovery of new knowledge). If "data science" is not about science and the adjective "data" has no particular meaning, why does this term exist?

Indeed, the allied fields of informatics have existed for several decades in many forms—medical informatics, clinical informatics, health informatics, bioinformatics, and biomedical informatics—and variants all refer to the development of methods to analyze data, information, and knowledge within the space of biology and medicine. Practitioners of these fields are quick to point out that most if not all of data science falls within the purview of informatics. Informatics is a broad field that includes the social aspects of interacting with data, information, and knowledge; the challenges of human–computer interfaces; and the issues associated with introducing disruptive new computational interventions into systems (like hospitals and laboratories) with existing workflows. So why is the introduction of a new name for the field necessary?

The term "data science" has gained recognition, and the widespread comfort with it suggests it serves a useful purpose. Here we offer some observations on the diverse use of the moniker for many activities:

- Biology and medicine are not the only activities that have been revolutionized by the availability of data in volumes and velocities that were previously not seen. Many fields of endeavor (transportation, finance, media, entertainment, real estate, and others) have seen an increased ability to capture, digitize, and store data and a concomitant increase in business intelligence to improve processes, understanding, delivery, and efficiency. Often, the internet is the source of the data (from social media, finance, advertising, search engines, and others), and it has generated an acute need for experts in computer science, statistics, and engineering. These industries do not have a strong tradition of a specialized subdiscipline of informatics and they adopted the term "data science" because it captures the types of problems that they have and the types of workers they seek. These "data scientists" focus on solving

https://doi.org/10.1146/annurev-bd-01-041718-100001

the analytic challenges that emerge from the new data sources that support business decisions. Initially, many companies employed computer programmers in this role, but the folks they now seek have an overlapping but different skill set.

- In a similar way, biomedical researchers (biologists, physician-scientists, clinical trialists, and others) see an opportunity to transform the way they do their work using the data streams that are now available. Many of these scientists also paid no attention to the allied informatics disciplines, considering them relevant to only a few areas of inquiry. Big data sets, or data streams, are a now a big problem for these scientists, and they find the term "data science" useful in capturing the pressures on their research and delivery missions. These data streams can broadly be summarized in three bins: genomic data, sensor data, and health care data. The ability to cheaply and accurately sequence DNA has now been supplemented with similar abilities to measure the transcriptome and emerging capabilities for metabolomics, proteomics, and other large-scale measurements. These provide molecular data of unprecedented magnitude (and potential value) that require specialized capabilities to analyze. The ability to sense the environment generally, and individual patients specifically, has also created a stream of information about activity, heart rate, electrocardiogram (EKG) signals, electroencephalogram (EEG) signals, environmental exposures, and other continuous data that promise to redefine our understanding of normal physiology and the response to disease and therapy. Finally, the move toward universal electronic medical records, population health biobanks, and associated databases now provides information about the overall occurrence and co-occurrence of diseases (and healthy states) that can be used to understand the major trends in population health, including disease incidence, drug response, and device performance.
- Although most elements of data science have always been present in the informatics disciplines, there seems to be a particular skill set that is more pressingly relevant in current applications. Today's challenges include very large data sets that must be managed carefully because, for example, they do not fit in the working memory of a typical computer. In addition, there is a need for large-scale annotation and metadata that explain how the data were generated and what the sources of noise are. There is increased interest in applying machine learning to very large data sets. The high-profile success of neural network–based deep learning systems has created an active market for individuals with the knowledge required to build and deploy these systems. Thus, there seems to be a specific skill set for data science that is a subset of all of informatics and that addresses the pressing needs of those who need more data science. Interestingly, this workforce need has led to the creation of domain-independent data science training programs in which trainees learn the key skills for which there currently is a strong market. They become experts at managing data but may not have any specific knowledge in the area of application, depending on collaborators to provide domain knowledge to ensure that the questions they ask and answer are relevant and well formed.
- It seems clear that biomedical data science has special challenges associated with the complexity of its data and the complexity of the science questions. Although a data scientist with no biological or medical knowledge can probably make contributions, they would require careful supervision or collaboration because the nature of the data, the sources of noise, and the set of assumptions that are reasonable to make (and

conversely not reasonable!) require a level of sophistication that is nontrivial. Thus, we believe that the best data science work is likely to come from data scientists who have taken the time to dive more deeply into the domain and make good decisions on a minute-to-minute basis about how the data may best be sliced, diced, and analyzed.

These observations lead us to conclude that the terms "biomedical data science" and "biomedical data scientist" are reasonable and useful: They connote activities associated with the creation and application of methods to new and large sources of biological and medical data aimed at converting them into useful information and knowledge. They also connote technical activities that are data intensive and require special skills in managing the large, noisy, and complex data typical of biology and medicine. They may also imply the application of these technologies in domains where their collaborators previously have not needed data-intensive computational approaches. Fine—so why do we need an Annual Review of this field?

The Annual Reviews series of journals is devoted to creating an archival record, free from strict page limits, for summarizing progress and challenges in academic disciplines. The pressure on scientific journals to limit article length often consigns the best informatics and data science to the tiniest of fonts in the methods section—which are often made available only in an online supplement! Even journals devoted to methodology do not always accommodate a full review of related work, their strengths and weaknesses, and a full exposition of the design goals for new algorithms, their evaluation, and implementation details. We believe that there is a need for an annual volume that captures the most important contemporary data science challenges and provides scholarly reviews written to be useful to both specialists and interested scientists from the application disciplines. This has been the vision in creating this first volume, which comprises an eclectic set of reviews spanning basic molecular biology to clinical medicine and including important new technologies as they are applied to these application areas:

Deep learning
Text mining
Visualization
Sequence analysis
Protein–RNA interactions
Molecular interaction networks
Simulation of cells/tissues
Challenges for mass spectrometry
Deconvolution of gene expression signals
Clinical and genomic phenotyping
Cancer therapy resistance

We are excited to bring you this first volume and look forward to having the *Annual Review of Biomedical Data Science* become a regular source of deep information about the challenges and contributions of data science to the larger biological and medical research landscape.

Russ B. Altman
Michael Levitt

Co-Editors