Perspectives on Data Science Opportunities and Challenges

Charles P. Schmitt, Ph.D.
Director, NIEHS Office of Data Science

The field of data science has undergone rapid growth in recent years, due in part to the ongoing generation of large-scale data and in part to significant and broad interest and investments across industry, defense, and science sectors. These investments have led to a tremendous outpouring of new methods and technologies. Capitalizing on the growth in data science for Environmental Health research faces the challenges of adapting and translating emerging methods and technologies into day-to-day use by researchers and policy-makers as well as ensuring that the development of new methods is informed by and tailored towards the unique challenges of Environmental Health research.

NIH and NIEHS have developed closely-related strategic plans to foster the growth of data science and the increased use of data- and knowledge-driven approaches for biomedical research. Central to the NIH strategic plan is ensuring research data is findable, accessible, interoperable, and reusable (i.e., FAIR) and promoting the open sharing of well-managed research data and metadata. NIH investments are supporting the coordinated development of cyberinfrastructures for finding, accessing and working with large scale and sensitive research data in the cloud as well as supporting the development, curation, and operations of scientific database and knowledgebase as activities onto themselves. Other investments are targeting the importance of workforce development and seeking to increase engagement with data science experts in other domains and work sectors.

Advancement of data science within NIEHS has focused on a broad range of activities that align with NIH plans, including incorporation of data science into strategic planning, provisioning of cyberinfrastructures that support scientist led development and deployment of research databases and tools, piloting of cloud platforms, increased integration of data and metadata management systems with laboratory information management systems, targeted hiring, and increased training of scientific staff. In addition, NIEHS is targeting data science areas of marked importance for environmental health research, including promoting community-based development and application of environmental health and exposure ontologies, development of natural language processing methods to extract exposure data from research journals and published reports, development of training materials tailored to environmental health resources, development of statistical, machine learning, and artificial intelligence approaches that advance data interoperability across the diversity of environmental health data, and developing methods and tools that aid in bridging the gap from specific exposures to human relevant outcomes.