Response From The Data Science Institute, Columbia University

Director, Kathleen McKeown, *Rothschild Prof. of Computer Science, kathy@cs.columbia.edu*
Executive Committee Members
Garud Iyengar, *Prof. and Chair of Industrial Engineering and Operations Research*
Paul Sajda, *Prof. of Biomedical Engineering, Radiology (Physics) and Electrical Engineering*

The Data Science Institute at Columbia University is a broad interdisciplinary Institute with five Centers in application areas of data science and one Center focusing on the foundations of data science. Our mission is the development of technology that can exploit the massive amounts of data available today to help solve society's most challenging problems. We bring together researchers on interdisciplinary teams to enable the use of data science in disciplines from journalism to science. We have launched a certification in data science for working professionals, an MS in data science and are working on a PhD program in data science.

Amidst the arrival of massive, streaming, and heterogeneous datasets, we argue that funders must retain a focus on the science of big data. In order to maintain breakthroughs in the impact derived from big datasets, it is critical that fundamental progress continue on representation, algorithms, and statistical models surrounding the use of data. Without these, big data loses its potential. Cross agency efforts can accelerate this progress as the same methods may be developed for radically different big data applications. For example, a major innovation in novel probabilistic models, approximate inference algorithms, or theoretical guarantees can make a huge impact across many domains. By requiring innovation in data science techniques as well as in the application, cross agency efforts can accelerate research and the potential in the application itself. This approach pairs data scientists with experts in disciplines on all projects.

In the specific research challenges we present below, researchers from different disciplines must be able to work together to develop solutions. Strategies for bridging disciplinary boundaries and building knowledge drawn from diverse teams must be developed for both educational and research practice.

Specific Research Challenges

Health: In health analytics the lack of shared, linked, meaningful and interoperable datasets is a key obstacle to progress. Privacy of personal health information is critical and approaches that enable privacy preserving summaries over both clinical and genetic data are needed. Given that data is streaming over time, is heterogeneous, and includes unstructured text, images and numerics, new techniques are needed that can align and anchor multiple data types automatically to identify time-resolved salient events of medical/clinical significance. New research is also required to leverage the potential of wearable sensors as a medical/health data source and its integration with hospital and physician medical care records. This area is likely to result in a "data-driven revolution" into the way we diagnosis and treat many of our societies' most chronic and costly diseases, including obesity, diabetes and heart disease. The technical challenges are so broad, and the impact to the average American is so significant, that a multi-agency effort is warranted (NIH, NSF, NIST, DoD).

Disaster, risk and resilience: Several agencies are separately focused on different activities to improve disaster response, resilience, and associated risk management (DHS, DARPA, NIH, NIST). Similarly, disparate research communities are working on various aspects of risk management of disasters: analysis of social media for crisis informatics; simulation and computational modeling; developing resilient infrastructure; and understanding of social behavior in the context of epidemics and other public health crises. Coordination among agencies and research groups would lead to a significant impact. Further advantages are to be had by encouraging the development of methods that are able to jointly exploit different existing data sets: FEMA datasets, HAZUS models, hand-built taxonomies for specific communities, NOAA weather data, economic models; and new datasets, such as databases of global supply chains with precise geographic locations of manufacturing and distribution facilities so that risks associated with natural disasters could be addressed. There is a need to make all available data publically accessible in a format that would enable cross-data set analysis.

Natural and applied science: Research on the study of the ocean and our forests offers tremendous potential for the understanding of climate change. Using heterogeneous data collected by NOAA, NASA and NODC may help to predict the impact of the melting Arctic ice on the North Atlantic. Research that connects data science with material genomics, astronomy and physics is also key. Materials performance, for example, is the bottleneck in critical technologies such as sustainable energy (e.g., solar energy production). Big Data approaches are well suited to this but there have been virtually no approaches along these lines to date. Bringing scientists together with data scientists who have expertise in optimization has promise to yield computationally efficient solutions. DOE, NASA, and NSF can support such efforts.

Social science: Data-driven research is also needed to support policy decisions that require interaction between government and academia. New York City DOH's research on how the lunch program affects child obesity is a good example of the use of city data for public health. Estimating populations at risk for infectious diseases such as HIV/AIDS, Hepatitis C and estimating the effects of risk behavior combined with social structure and effective campaign and communication strategies to increase awareness of programs such as flu vaccines. Research challenges in all of these areas include: study design, data collection, processing and analysis, and policy analysis through simulation studies based on findings from field data analysis to support the decision making process and they cross agencies such as NIH and NSF.

Urban Data Science and Decision Support: Given the explosion in urban sensors and data sources, advances in data science tools and urban simulation platforms, and domain expertise across the engineering, natural and social sciences, we can now accelerate knowledge in the field of multi-scale urban dynamics and provide decision support tools for charting sustainable city futures. Funding needs to focus on scientific and technical advances within, and at the intersection of, urban sensing (from a wide variety of novel sources including advanced infrastructure monitoring networks, social networks, mobile data and transactions), computational urban simulation and modeling, and optimization and decision support tools.

Mechanisms for Funding Education

New mechanisms for funding are needed to support the educational enterprise around big data. A funding focus on big data as it relates to an entire ecosystem, as opposed to a specific research project, would enable major advances. For training, PhD fellowships that place a cohort of students who excel in the foundations of data science together with cohorts of students are interested in exploiting the big datasets in their discipline. This would enable the education of many more students than is possible with investigator focused research projects. The creation of national laboratories, each focusing on one or more disciplines in combination with data science, would also accelerate research. Finally, given the need to immerse oneself in both technology and data from different disciplines, longer periods of time before becoming a faculty member or joining industry is needed. Time to learn the vocabulary of other disciplines and the importance of different data is essential. Thus, we also advocate for the establishment of data science postdoctoral fellowships.