NATIONAL SCIENCE FOUNDATION
4201 WILSON BOULEVARD
ARLINGTON, VIRGINIA 22230

NSF 17-049

Dear Colleague Letter: Request for Input on Federal Datasets with Potential to Advance Data Science

February 2, 2017

Dear Colleague:

Through this Dear Colleague Letter (DCL), the National Science Foundation's (NSF) Directorate for Computer and Information Science and Engineering (CISE) requests input on possible datasets held by federal departments, agencies, and offices that would be useful in furthering research in data science, including machine learning.

Over the past few years, Project Open Data (https://project-open-data.cio.gov/) has sought to identify and share best practices, examples, and software code to assist federal agencies with opening up access to data. Moreover, there have been efforts to scale up "open data" across various application sectors, including health, energy, climate, education and learning, finance, public safety, and global development, unlocking valuable data and improving decision making by making data resources more open and accessible to innovators and the public. NSF has established a national network of Big Data Regional Innovation Hubs and Spokes (BD Hubs and Spokes), comprising members from academia, industry, and government, with the goal of igniting new public-private partnerships across the Nation in big data research and development as well as training and education. Facilitating access to data is one of the objectives of the BD Hubs and Spokes. Collectively, these initiatives constitute an important first step in supporting the growing and interdisciplinary data science research community, which requires access to real-world datasets, e.g., as training data that can further data science, including machine learning capabilities, and enhance knowledge and decision making in various application sectors.

With this DCL, the CISE directorate is seeking input on the types of datasets that federal departments, agencies, and offices may possess and could make openly available for use in data science, including machine learning, research — and the potential associated broader impacts on science, engineering and society. In the longer term, planning grants may be made available in cases where well-defined efforts to publish specific government datasets in the open are described. Such projects may focus, for example, on various aspects of "data cleaning" (e.g., anonymization of data) that may be necessary to make the data openly accessible. Given the breadth of federal departments, agencies, and offices, the available datasets may be wide-ranging in scope and type.

Responses to this DCL should include the following information, for each distinct dataset:

- Submitter's name(s) and affiliation(s);
- Federal department/agency/office related to the dataset (if known);
- Information about the likely dataset, including:
  - What the dataset is about;
  - The type of data, e.g., structured/fielded data, unstructured text, image, video, etc.;
  - Approximate size of the data, in terms of the number of distinct "entries"/entities, as well as total bytes; and
  - How the data would be made available, e.g., via download from a server; physical media; etc.;
- Anticipated impact on data science, including machine learning;
- Potential scientific, engineering, and/or societal impacts that are anticipated by making the data available to data scientists and engineers; and
- Expected level of effort — in staff weeks/months and total dollar amount (note that the average level of effort is expected to be up to $50,000.

Submissions should be emailed to TrainingDataDCL@nsf.gov by March 31, 2017.

Inquiries about this DCL may be sent to Chaitan Baru, CISE Senior Advisor for Data Science, at cbaru@nsf.gov.

Sincerely,

Jim Kurose
Assistant Director, CISE