The American Statistical Association's Comments on the NIH Draft Data Science Strategic Plan

Prepared under the guidance of its Committee on Funded Research

April 2, 2018

*** The appropriateness of the goals of the plan and of the strategies and implementation tactics proposed to achieve them**

The American Statistical Association applauds the NIH's efforts in developing a Strategic Plan for Data Science and specifically, seeking comments from the community on how best to take advantage of this important emerging field to serve NIH's mission. The NIH Big Data to Knowledge (BD2K) office provided an important and constructive impetus NIH's data science work and we're glad to see this important work continue.

We are concerned by what we consider an overall lack of acknowledgement of the important role of statistics—the science of learning from data, and of measuring, controlling, and communicating uncertainty—and statisticians and biostatisticians in the data science. The document properly acknowledges the immense advances in computing and other fields but scarcely mentions statistics.

With the vital role of statistics in data science broadly acknowledged,[1] we believe the eventual NIH data science strategic plan will be most successful with statistics fully integrated. Indeed, one description of data science is as a stack of various skills and expertise.[2] Statistics plays a central role in at least a few of the data science stack layers (e.g., modeling, machine learning, decision

[1] https://cra.org/data-science/; https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505347; https://www.nap.edu/read/24886/chapter/2; https://arxiv.org/pdf/1503.08776v2.pdf; http://magazine.amstat.org/blog/2015/10/01/asa-statement-on-the-role-of-statistics-in-data-science/.
[2] https://www.nitrd.gov/nitrdgroups/images/6/62/DataScienceTechnologyStack.pdf

support) and is essential in fundamental experimental design, whether this is based on "small" data or massive data. The more the NIH data science strategic plan acknowledges and integrates the various components of data science from the beginning—as BD2K did—the more successful it will be.

Just as there have been important advances in recent years in computing and other fields, there have been equally important advances in statistics, a scientific discipline that goes back to at least the 1800's. Throughout its long history, statistics has made innumerable and important contributions to biomedical science. As one indication of its impact, the 2014 *Nature* analysis of the 100 most cited research articles ever included nine articles by statisticians and two more that were heavily statistical.[3]

With the immense concern for the rigor and reproducibility of scientific research that has emerged in the last several years, the engagement of statisticians in biomedical research, and of more and more thorough statistical training for biomedical researchers, have each grown considerably. Indeed, in 2014 *Science* announced the creation of its Scientific Board of Reviewing Editors[4] to help verify the accuracy of results reported on its pages, acknowledging the critical role of competent and correct statistical argumentation in how scientists learn from data. *Nature* has also engaged statisticians in its efforts to ensure the publishing of more robust scientific research and has gone to the extent of contributing to statistical training with a regular "Matters of Significance" column in its *Nature Methods column,* subtitled, "Sound experimental design and analysis require improved statistical training."[5]

The NIH Center for Scientific Review (CSR) also acknowledges the important role of statisticians and statistics in its September 2016 publication[6] of Peer Review Notes reporting on a meeting of senior CSR staff and ASA members and staff. The lead paragraph states the key points included "how statisticians can (1) help improve rigor and reproducibility, (2) be key members of a research team, (3) identify common statistical issues, and (4) play valuable roles in peer review meetings."

The National Academy of Science, with leadership from its Committee on Applied and Theoretical Statistics, the Computer Science and Telecommunications Board, and the Board on Mathematical Sciences and Analytics, has also tacitly acknowledged the vital role of statistics in data science. See, for example the 2018 interim report, Envisioning the Data Science Discipline: The Undergraduate Perspective, or any of the highlights of the Roundtable on Data Science Postsecondary Education.

In addition to journals and funding agencies, the broader scientific community has also recognized the vital role of statistics in bettering science in the data science era. The 2016 "ASA Statement on p-Values" has been viewed 244,000 times as of April 2, 2018 and spawned widespread discussion of statistical evidence to bolster scientific research. A further indication of the growing recognition for the importance of statistics in the data science era is the reception of the June 2016 PLOS article, "Ten Simple Rules for Effective Statistical Practice." The article has

[3] http://community.amstat.org/blogs/steve-pierson/2014/11/10/statisticians-prominent-in-top-100-cited-article-list
[4] http://science.sciencemag.org/content/345/6192/9.summary
[5] https://www.nature.com/articles/nmeth.2638
[6] https://www.csr.nih.gov/CSRPRP/2016/09/statisticians-share-insights-for-applicants-and-reviewers/

216,000 views as of April 2, 2018, making it the sixth most viewed “Ten Simple Rules” piece in the series that goes back more than a decade. The ASA Committee on Funded Research published a document, “Statistical Issues Seen in Non-Statistics Proposals,” to “assist researchers of all disciplines in improving their scientific work.”

In short, statistics is a critical part of data science and NIH’s strategic plan on data science should recognize that stronger and more robust scientific results can be achieved through the integrative and vibrant engagement of statisticians and biostatisticians.

*** Opportunities for NIH to partner in achieving these goals**

Close partnership with organizations such as the National Science Foundation, the National Academies, CSR, and the ASA, as well as journals and publishers, are crucial for successful implementation of the goals - after revision to address the concerns listed above.

*** Additional concepts that should be included in the plan**

We refer the reader to the response in the first section where the need to incorporate statistical concepts is stated, including in training of biomedical scientists.

*** Performance measures and milestones that could be used to gauge the success of elements of the plan and inform course corrections**

/no response/

*** Any other topic the respondent feels is relevant for NIH to consider in developing this strategic plan**

/no response/