Preamble

NIH's mission is "to seek fundamental knowledge about the nature and behavior of living systems and the application of that knowledge to enhance health, lengthen life, and reduce illness and disability." Data Science will play a prominent role in fulfilling that mission in the 21st Century. Unfortunately, and for several reasons, the current NIH Strategic Plan for Data Science (SPDS) will not further, and may even act against NIH's mission. On the surface, the SPDS has identified important goals. However, careful review of the current Plan reveals the impossibility of translating anything articulated into impactful, actionable, and meaningfully measurable implementations. If this document came from any organization other than NIH it might be ignored. Instead, the awesome weight of NIH recommendations - which demand an equally awesome obligation to the community for productive criticism - invites the plausible scenario that many investigators (within and outside of NIH) will be guided by an incomplete, inaccurate, and dangerously misguided document. While the tone of this document is necessarily blunt (far more so than if feedback was addressed to an individual investigator), it is with the belief that NIH does not simply want feedback that says what it wants to hear - the NIH mission is too important for criticisms to go unheard.

To conclude that the SPDS is fundamentally flawed, consider these representative metrics copied verbatim from the document:

- Goal 1: Support a Highly Efficient and Effective Biomedical Research Data Infrastructure
    - Sample evaluation metric: "quantity of cloud storage and computing used by NIH and by NIH-funded researchers"
- Goal 2: Promote Modernization of the Data-Resources Ecosystem
    - Sample evaluation metric: "quantity of databases and knowledgebases supported using resource-based funding mechanisms"
- Goal 3: Support the Development and Dissemination of Advanced Data Management, Analytics, and Visualization Tools
    - Sample evaluation metric: "quantity of new software tools developed"

- Goal 4: Enhance Workforce Development for Biomedical Data Science
    - Sample evaluation metric: “quantity of new data science-related training programs for NIH staff and participation in these programs”
- Goal 5: Enact Appropriate Policies to Promote Stewardship and Sustainability
    - Sample evaluation metric: “establishment and use of open-data licenses” (a count metric)

These metrics (which are wholly representative of evaluations proposed) would never be accepted by a scientific reviewer. If the SPDS were to be considered by one of the NIH’s scientific review panels, it would certainly be designated “Not Recommended for Further Consideration (NRFC)”. Simply put, there is no perceivable likelihood of this Strategic Plan to exert any positive influence on the fields of Data Science or data intensive biomedical research. The majority of metrics proposed are simply “counts” of the number of activities created or performed, without any meaningful and/or independent evaluation of their benefit. Most of these activities have to happen anyway -e.g., the “quantity of databases and knowledgebases supported” will increment every time the NIH gives a grant to someone proposing to collect data and create a database (following the NIH policy on data sharing, https://grants.nih.gov/policy/sharing.htm). There is no suggestion that the increments these metrics represent are meaningful - or that they do or might contribute to biomedical work to “enhance health, lengthen life, and reduce illness and disability.” These evaluation metrics are both tautological (i.e., if the NIH funds research that generates data, these metrics will increment and the Goal will appear to have been “achieved”) and vague - as long as at least one of each of these events occurs, the Goal could be argued to have been “achieved”). Thus, almost any proposal to collect a large quantity of data that is funded will be branded as “successful”. While these goals are important for the NIH to function as a data broker, they actually represent characteristics to strive for throughout NIH, rather than a Strategic Plan. This plan is surprisingly unreflective of prior NIH Data Science activities (NCBC, BD2K) or even praiseworthy and well-informed planning documents emerging in parallel (NLM Strategic Plan: https://www.nlm.nih.gov/pubs/plan/lrp17/NLM_StrategicReport2017_2027.pdf).

Data Science is not the progressive extension of clinical research applications, and a truly impactful NIH strategy for Data Science must by definition come from experts in Data Science - preferably, those with expertise in both Data Science and biomedical research. As stated in SPDS Goal 4, NIH as an organization does not have the expertise in Data Science to plan its strategic direction in this area. We recommend that if NIH is committed to a transformative strategy for Data Science (and it should be), the current SPDS should be discarded. A Strategic Plan with meaningful goals that promote the thoughtful integration of Data Science with biomedical research representing measurable (not "countable") impact, together with formal and independent evaluation, should be developed instead.

Such a Plan would require a sustainable vision that accurately reflects the discipline of Data Science and its potential role in biomedical research. Such a vision must be community-driven, perhaps through a call for nominations that would assemble national and international community members with evidence of expertise in the needed areas. Given the weaknesses in the current SPDS, this RFI has every chance of resulting in an inward-looking, self-fulfilling prophecy (e.g., such that any grant that is made to any data-generating proposal results in "success" according to these self-defined metrics). If the Data Science and data-intensive biomedical research community comments are ignored, and priority is given to existing investigators and internal stakeholders, the NIH will promote an inward-looking and essentially irrelevant program for integrating Data Science into biomedical research.

The Plan fails on multiple several technical merits, and the community members who have authored this document have assembled additional detailed remarks at this URL:

https://github.com/JasonJWilliamsNY/2018_nih_datascience_rfi

Additionally, as an exercise, the NIH review form has been filled out for this Plan to explore how a realistic NIH scientific reviewer (who did the exercise) might score such a proposal

https://docs.google.com/document/d/1XxQLORoTm2lkucQz6k3QdgNOARDrvU3zx_svT0PX_c0/edit#

These comments are provided to support a new effort at a realistic, plausible, and community-driven Strategic Plan for Data Intensive Biomedical Research. The community stands ready to assist NIH in this important work and we urge the organization to commit to making this a community effort. We thank NIH for the work done and going forward to develop the next stages of this plan.