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/1XxQLORoTm2IkucQz6k3QdgNOARDrvU3zx_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.

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

Goal 1: NIH should be applauded for recognizing that historically, funding of data resources used funding approaches that were appropriate for research projects. We agree this must change. Also, we agree that tool development and data resources may often need distinct funding contexts/expectations.

Goal 1: We agree with the stated need to support the hardening and optimization (user-friendliness, computational efficiency, etc.) of innovative tools and algorithms.

Goal 1: The Data Commons has been driven from bottom-up development by the community. The Commons is in its early stages and should be allowed to function as is. The Commons is the appropriate testbed for many of the technological innovations and processes that NIH may ultimately which to explore at broader scales after sufficient development.

Goal 1: The implementation tactics proposed seem almost randomly selected, making it nearly impractical to criticize them. For example, one of the three bullet points indicates that new technologies should be adapted (somewhat meaningless, but not disagreeable), but then goes on to mention that GPUs should be used. That is weirdly specific, and not necessarily wrong, but why mention at this level of detail without some specific vision for the real biomedical informatics problems that are relevant here? This is like saying “calculus” should be used. Maybe; but such an out of context statement is ultimately a collection of buzzwords. At best, this is an indication that greater expertise is needed to reformulate this document.

Goal 1: Although this is a strategy document, it's implausible to imagine the linking of NIH datasets as described in objective 1-2 can be elucidated in a single thin paragraph. We have no idea how the "Biomedical Data Translator" will work but can only imagine it will need to function at least as well as the "Universal Translator" of Star Trek. Either enough detail for the strategy needs to be proposed here for the document to serve its purpose, or the space is better spent articulating the hard problems that need fixing.

Goal 2: The overall goal recognized (the need to avoid data siloing) is absolutely a correct (but difficult) target for NIH. It is impossible to believe that anything in the implementation or metrics will demonstrably achieve this. What appears to be one problem is likely several problems, each of which is worthy of study to generate actionable solutions. Usability is identified as target to optimize for and while this may be correct, no metrics proposed seem to measure this. While there may be real and important distinctions between databases and knowledgebases, we again suggest no convincing metrics have been proposed here. The statement that "Funding approaches used for databases and knowledgebases will be appropriate…" conveys no usable information.

Goal 2: For some reason, objective 2-2 contrasts the "small-scale" datasets produced by individual laboratories with the "high-value" datasets generated by NIH-funded consortia. Besides pointing out the needless condescension here, this kind of contrast actually belies the problem in designing data systems in which there are various classes of "citizenship." Treating the much larger quantity of data generated by the plurality of extramural investigators as somehow different my lead to policies which work for the NIH, but don't work for the community, leading to irreconcilable standards.

Goal 2: Implementation tactics proposed are bewildering. Under this subheading appears the bullet "Ensure privacy and security." While it is gratifying that the word privacy finally appears 11 pages into this 26-page document, this is not in any way an implementation tactic.

Goal 2: Evaluation metrics are horrific. Quantity of datasets and funded databases does nothing more than account for the fact that NIH spent money.

Goal 3: The strategy to leverage and support existing tool-sharing systems to encourage "marketplaces" for tools developers, and to separate funding of tool development and data generation is very important, and we support this direction of the NIH. This is a proven strategy to elevate high quality tools, for example, in the world of high-throughput genomics, one can consider the NHGRI funded Bioconductor Project as a decade-long successful use case in providing a unified interface for more than 1,000 "high-quality, open-source data management, analytics, and visualization tools".

Goal 3: In general, implementation tactics are plausible, but not evidence-based. Rather than propose that each step NIH takes to develop a tactic is supported by a body of research, the lowest-hanging fruit (and most productive solution) is to have the community develop an actual set of strategic targets, with clear metrics for evaluation.

Goal 3: The SPDS plan underestimates the pervasiveness and persistence of bad/outdated software and methods (See: https://www.the-scientist.com/?articles.view/articleNo/51260/title/Scientists-Continue-to-Use-Outdated-Methods/). It is completely unclear how separating evaluation and funding for tool development and dissemination from support for databases and knowledgebases (this sentence from the SPDS is itself unclear) will address this problem. This may help, but is our knowledge an unvetted hypothesis.

Goal 3: Although the SPDS does not make any strategy clear, the goal of supporting tools and workflows (objective 3-1) is a good one. We further agree that partnership is exactly the way that this needs to be pursued.

Goal 3: The metrics proposed for this sophisticated set of objectives are catastrophic.There is no way that the objectives stated for Goal 3 can be effectively measured or set a useful standard for success.

Goal 4: Feldon et.al (PNAS, 2017; https://doi.org/10.1073/pnas.1705783114) concludes that despite $28 million in investment by NSF and NIH in training (including workshops/boot camps relevant to biomedical data science), much of this training is “not associated with observable benefits related to skill development, scholarly productivity, or socialization into the academic community.” Clearly, if NIH intends to do better it needs to completely reconceive how it approaches training in data science.

Goal 4: Several reasonable priorities are identified, but only demonstrably ineffective and/or inappropriate approaches and evaluation metrics are proposed to achieve/evaluate these goals.

Goal 4: The inadequacy of the strategies suggested is epitomized by the proposed evaluation metric: "the quantity of new data science-related training programs for NIH staff and participation in these programs." This is as unconvincing as suggesting a research proposal will be measured in the number of experiments performed. In fact, the only metric proposed for evaluation of training is the number of training opportunities created. Such an arbitrary and crude metric would get any research proposal returned from study section without discussion. Number of training opportunities cannot be a plausible metric. Despite their being no shortage of training opportunities from MOOCs to workshops, there is a persistent, apparent, and urgent training gap. This inadequate metric is a clear red flag that training guided by the proposed plan will accomplish very little.

Goal 4: It is clear that training is under-prioritized by NIH. In the largest survey on unmet needs for life science investigators, NSF investigators report in Barone et.al (PLOS Comp. Bio, 2017; https://doi.org/10.1371/journal.pcbi.1005755) that their most unmet computational needs are not software, infrastructure or compute. Instead it is the need for training; specifically training in integration of multiple data types, data and metadata management, and scaling analyses to HPC and cloud.

Goal 4: Strategy failed to properly understand the role of training in biomedical data science and the need to define and measure what constitutes effective training. There is no mention made of a serious commitment to evidence-based teaching that is needed to design effective short-format courses and workshops. While the educational pipeline from at least the undergraduate level and beyond needs serious improvement to address biomedical data science, short-format training and workshops will play an important role. These workshops must not be constructed ad hoc. Typically, training is delivered by staff/faculty with a high level of bioinformatics/data science domain expertise, but little to no guidance in andragogy, cognitive science, or evaluation.

Goal 5: Here, and once in Goal 3 are the only mentions in the document of a community-driven activity (which actually needs to be brought to the entire

SPDS). The FAIR Data Ecosystem is a laudable goal, but the idea that NIH should "Strive to ensure that all data in NIH-supported data resources are FAIR" is still a goal without a plan. More than technological advances or implementation, this is a training activity that requires community awareness, understanding, input, and buy-in on FAIR principles. The implementation tactics are plausible, but ultimately without appropriate evaluation are too vague to establish success. Establishing open-source licenses or even promoting their use won't in itself FAIR. This section misses out on how the hard question of ELSI/privacy and biomedical data which NIH currently has not updated to accommodate the vision of what biomedical data science might achieve.

Goal 5: Evaluation is inappropriate/unrevealing count metrics that will not indicate whether FAIR principles are realized or not.

Opportunities for NIH to partner in achieving these goals

Goal 1: NSF has been exploring centralized computing models through XSEDE, and open-science clouds (CyVerse Atmosphere, XSEDE-Jetstream) for many years. These groups would be natural partners in addition to commercial cloud providers. The NSF resources will not match the capacity of commercial cloud but have optimized for the science use-cases and user profiles relevant to biomedical research.

Goal 2: ASAPbio, Force 11, Open Science Framework, Zonodo, FigShare, BiorXiv, and many other community-driven organization are exploring data lifecycle issues and metrics that are relevant to this discussion. The entire SPDS needs to be completely reconceived to include representation from individuals within these organizations who have scholarly reputations in data management and life science publication/communication.

Goal 3: There are a variety of groups the NIH can partner with. The number of potential individual investigators is too numerous to list, but these individuals should be relatively easy to identify by means of their scholarly contributions (carefully avoiding journal publications as a primary metric). Reaching out and partnering with groups such as the Open Bioinformatics Foundation and societies like ISMB would be an ideal way for NIH to foster deep community involvement.

Goal 4: We present comments here in the hopes that NIH will consider bold action in this area because the problem is solvable and the community of investigators with experience in training relevant to biomedical data science has been thinking deeply on the topic. The community is relatively small, well-connected, and should be extensively leveraged in developing robust scalable solutions. It should be easy to assemble the 10-20 most important practitioners and educators in biomedical data science, confident that they and their second order collaborators would constitute a reasonably sized working group that can bring in much needed solutions. Understandably, in fact by definition as an aspirational goal, NIH has identified the value in prioritizing data science as an area it needs to be at the forefront of but does not have the expertise to achieve alone. The community does.

Goal 4: Right now, the single best target for collaboration is the Software and Data Carpentry community (Greg Wilson: "Software Carpentry: Lessons Learned". F1000Research,2016, 3:62 (doi: 10.12688/f1000research.3-62.v2). There are many reasons why collaboration here will be tremendously important for NIH to succeed. First, the Carpentry community itself represents a global federation of researchers in the space of computation, data science, bioinformatics, and related fields with a strong interest in education. In short – this is a self-selected community of hundreds to thousands of researchers; the available expertise in biomedical data science is well-covered in this community. Additionally, there simply is no other community that has built a sustainable and scalable approach to building educational content relevant to biomedical data science with strong grounding in assessment and pedagogy. It would be a tremendous squandering of resources to not build on this foundation.

Goal 4: This is an area that especially calls for collaboration with the National Science Foundation. NSF already has several strong funded programs that are dedicated to understanding the problems of bioinformatics education and biomedical data science and developing solutions – the Network for Integrating Bioinformatics Education into Life Science (NIBLSE, https://qubeshub.org/groups/niblse) is just one of many. NIBLSE has for example, identified that the newest trained faculty with bioinformatics expertise are not bringing that training into the classroom, and the lack of training is the biggest barrier to bioinformatics education (https://www.biorxiv.org/content/early/2017/10/19/204420). The potential for synergy here is enormous for developing the k-16 pipeline. This alliance (especially leveraging NSF INCLUDES) could be a tremendous opportunity to do so in a way that enhances diversity. And while there are distinct career paths for biomedical vs. non-human life sciences, there is almost complete overlap in the study, preparation, and training for both student groups.

Additional concepts that should be included in the plan

Goal 1: Grafting “Data Science” onto NIH is essentially a massive retrofitting exercise. If we had to pick one area to think of, it is focusing on emerging techniques (Long-read sequencing, machine learning approaches, CIRSPR, etc.) and how NIH manages these data that could be a primary target for envisioning a data science-friendly ecosystem. The community of users is smaller, and fixing emerging challenges seems like a manageable focus for fomenting community consensus.

Goal 2: Conceptually, this goal needs to clearly differentiate technological obstacles from process obstacles. In reality, many of the needed technologies are either in place or will be generated from sectors outside of biomedicine. A few, perhaps, will be unique to the NIH use cases. More effort needs to be put into understanding the workflows and processes that investigators in a variety of contexts use to produce and consume data. This is an unaddressed research question in this document.

Goal 3: The proposal as currently mentioned does not mention (1) computational reproducibility, or (2) exploratory data analysis for data quality control. These two topics are critical for the high-level goal of "extracting understanding from large-scale or complex biomedical research data".

Goal 3: Computational reproducibility can be defined as the ability to produce identical results from identical data input or "raw data", and relies on biomedical researchers keeping track of metadata regarding the versions of tools that were used, the way in which tools were run, and the provenance and version of publicly available annotation files if these were used. This is very important for data science: if two groups observe discrepancies between their results, they absolutely must be able to identify the source, whether it be methodological or due to different versions of software or annotation data.

Goal 3: Exploratory data analysis (EDA) needs to be a key component of the data science plan, as this should be the first step of any data analysis involving complex biological data. EDA is often how a data scientist will identify data artifacts, technical biases, batch effects, outliers, unaccounted for or unexpected heterogeneity, need for data transformation, or other various data quality issues that will cause serious problems for downstream methods, whether they be statistical methods, machine learning, deep learning, artificial intelligence or otherwise. In particular, machine learning and statistical methods rely on the quality of the metadata and the ability to provide consistent terms and judgements on describing samples across all datasets from consortia and from individual labs. Downstream methods may either fail to detect the relevant signal (loosely categorized as "false negatives") or may produce many spurious results which are purely associations with technical aspects of the data ("false positives"). Furthermore, Basic EDA can uncover biological signal that may be missed, such as biologically relevant heterogeneity, e.g. subtypes of disease with signal present in molecular data.

Goal 3: Computational reproducibilty and supporting EDA should be components of both NIH funded tool development, as well as the plan to "Enhance Workforce Development for Biomedical Data Science" (Goal 4).

Goal 4: There are a few basic concepts that must be included; and these are potentially at the right level for a strategic vision document:

- Training is the most unmet need of investigators. Investments that under-prioritize training will not realize the value of computational and data infrastructure developed.
- Biomedical data science education must not be solely delivered by domain experts/investigators training according to what they think is best. Instead, curriculum must be developed using evidence-based pedagogical principles.
- Collaboration is key. Training that is developed as the unitary creation of NIH will fail. Training must be developed by a community that can maintain and sustain learning content.
- Assessment is an integral part of training and cannot be ad hoc, it must generate evidence that learning has occurred, and be developed in a framework community of practice. This is hard – it is easy to count the number of CPU

cycles paid for on the cloud, or the size of a database.

- Data science must not about science, and not just data. It is easy to accumulate datasets, but not easy to develop training that is measurably effective – however it is definitely possible.
- Citizen Science is more than "individuals giving their brains for analyzing data using computer games". There are growing communities of patients and healthy individuals who are coming together to analyze biomedical data, either their own or using public data resources to perform science under their own lead (c.f. http://jme.bmj.com/content/early/2015/03/30/medethics-2015-102663 on participant-lead research). These community efforts are growing substantially and are bound to become important stakeholders in performing additional biomedical research. The needs of these communities should thus be targeted to.
- Diversity is a highly obtainable goal for biomedical data science education.

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

Goal 1: The proposed evaluation metrics are horrific. These metrics are more appropriate for cloud providers who capture the described metrics to develop their invoices. While NIH should look to drive down communal costs, Goal 1, like all of the goals in this document are – hard research problems – which require deep thought to understand what success is.

Goal 2: Without specific research questions, or the identification of relevant research that can be directly applied to the use cases NIH wants to advance, there are no additional milestones except a clearer definition of the goal.

Goal 4: This question is difficult to answer because there needs to be a defined and agreed-upon set of competencies for biomedical data science. From these will follow learning objectives and assessments for these objectives. At the next stage will be dissemination targets and measures of community use and buy in. A workshop could resolve and develop these over the course of a few months.

Any other topic the respondent feels is relevant for

NIH to consider in developing this strategic plan

Goal 1: The SPDS NIH correctly identifies that “The generation of most biomedical data is highly distributed and is accomplished mainly by individual scientists or relatively small groups of researchers.” This should be followed by the conclusion that any top-down approach must be matched by a correspondingly large-scale bottom-up approach. Individual investigators need the training and support to generate data in a way that fulfills the promise of FAIR principles. The Strategic Plan quotes the famous (and regrettable) statistic that 80% of Data Science is cleaning data, yet nothing proposed in this document will solve this. While NIH is in a position to pioneer (and appropriately fund) hard infrastructure (computation/storage, etc.), the greater attention must be paid to funding soft-infrastructure – training, documentation, and support that bring investigators into a community of practice. Note that Barone et.al (https://doi.org/10.1371/journal.pcbi.1005755) replicates the earlier findings of EDUCAUSE (https://net.educause.edu/ir/library/pdf/ers0605/rs/ers0605w.pdf); organizations planning for cyberinfrastructure development tend to underestimate and underfund the training needed to use infrastructure. Any infrastructure development must be matched by clear, measurable learning outcomes to ensure that investigators can actually make intended use of the investments.

Goal 3: There are so many potential emerging technologies and themes, understandably, this plan should not be a laundry list of things to try. This objective need to be reconceived to articulate how those solutions will be collected, pursued, and evaluated. No such vision is clearly present.

Goal 4: The work being done by NSF as well as the Data Science recommendations developed by the National Academies are highly relevant and need to be better integrated. There is also a unique connection in Data Science to industry. Industry partners will continue to lead advances in Data Science relevant to biomedicine.