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 back 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 these 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.

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.

RPG Review

If you cannot access the hyperlinks below,

visit http://grants.nih.gov/grants/peer/critiques/rpg_D.htm.

Application #: NIH Draft NIH Strategic Plan for Data Science

Principal Investigator(s): NIH

Overall Impact

Reviewers will provide an overall impact score to reflect their assessment of the likelihood for the project to exert a sustained, powerful influence on the research field(s) involved, in consideration of the following five scored review criteria, and additional review criteria. An application does not need to be strong in all categories to be judged likely to have major scientific impact.

Overall Impact **9**

Write a paragraph summarizing the factors that informed your Overall Impact score.

This strategic plan has very low likelihood to exert any substantive influence on research fields – including Data Science. Data Science is a dynamic and heterogeneous field that goes far beyond simply assembling, storing, and managing data, and to exert influence on this (or any other) field, this Strategic Plan should have been based on a needs analysis of any of those fields. “NIH supports the generation and analysis of substantial quantities of biomedical research data”; but NIH also supports the generation of substantial quantities of data that are not actually supportive of research aims. The Draft Strategic Plan for Data Science conflates “all data” with “research data”, resulting in a Plan that, even if fully realized, cannot purposefully support the NIH mission of seeking or applying knowledge “to enhance health, lengthen life, and reduce illness and disability.” What this Plan mainly supports is data management; to “allow NIH to adopt more cost-effective ways to capture, access, sustain, and reuse high-value biomedical data resources in the future”. However, two previous 10-year programs to fund infrastructure (National Centers for Biomedical Computing, NCBC; Big Data to Knowledge, BD2K) have cost NIH and taxpayers millions, and yet while

they clearly should have laid the groundwork for at least some part of the vision laid out in this Strategic Plan, neither their failures nor their successes are featured or reflected on at all. Since neither of those two programs had any lasting influence at all on NIH's current thinking, it is unlikely that this Plan could be expected to have either sustained or powerful influence on the NIH, much less on research. Moreover, the foundational weaknesses in the evaluations contemplated throughout the Plan highlight an unwillingness – or inability - to formalize evaluation – which may be the reason why prior programs never had any substantive successes OR failures to report. It is obvious that, because the NIH awards grants to investigators, if this Plan were implemented, there would be *financial* influence on anyone who chose to compete for these funds. However, the "power" of that influence would be questionable – and only lasting as long as the funding does. There is literally no way for this Plan to have any effect or impact on Data Science; since it was developed with no apparent input from practicing data scientists, they will ignore this Plan and worse, most probably continue to develop the discipline with no reference to this Plan at all. Then at the end of this initiative, anyone who pursued "Data Science" following the NIH Plan will not be competent, or even proficient, in actual Data Science. There does not appear to have been any consideration of how any biomedical science could be affected in any way by the execution of this Plan; therefore weakening any likelihood of this Plan having any real influence on the biomedical research fields that should be of greatest interest to NIH.

Scored Review Criteria

Reviewers will consider each of the five review criteria below in the determination of scientific and technical merit, and give a separate score for each.

1. Significance **9**

Strengths

None noted.

Weaknesses

Does the project address an important problem or a critical barrier to progress in the field?

No barrier to progress in any field is meaningfully defined or characterized. This Plan appears to have been motivated by the survey results published by “CrowdFlower” reporting “How a Data Scientist Spends Their Day (p. 6) – but while this may in fact describe how * data scientists * spend their day, it does not describe how NIH funded scientists or biomedical scientists who are not employed as data scientists interact with data. People who are employed as data scientists may actually be specifically hired to do data wrangling to free up biomedical scientists for the rest of the scientific process. The more appropriate interpretation of this survey result (that 80% of time is spent managing/cleaning data) is that biomedical researchers may need to hire someone to do this, not that biomedical scientists need NIH to manage data or further fund data oriented initiatives.

The use of the expression “Plan for Data Science” suggests that executing this Plan may actually have some effects on Data Science as a discipline, but since there does not appear to have been any actual Data Scientists involved in crafting this Plan, that title is inappropriate. The first group to whom “workforce development” activities are directed are NIH employees (Objective 4-1). This is not a “critical barrier to progress in the field”, it is an NIH HR issue.

Is there a strong scientific premise for the project?

There is no scientific premise at all for the concept that NIH can contribute to Data Science by drafting a Plan with goals that are as weakly “evaluated” as the Plan describes. The idea that biomedical researchers need to harness the dynamic innovation that is always going on in Data Science makes sense, but this Plan does nothing to engage with actual data scientists, and it was clearly written without appeal to, or communication with, the work of Data Science or data scientists.

If the aims of the project are achieved, how will scientific knowledge, technical capability, and/or clinical practice be improved?

It is *possible* that both the NCBC and BK2K programs resulted in incremental gains in scientific knowledge and technical capabilities – for a very specific few in the biomedical research field, but it is *certain* that no evidence is included in this Plan suggesting any such improvements as a result of its implementation. Moreover, the lack of reflection on these two previous programs suggests that, like them, no real advances or improvements are likely.

How will successful completion of the aims change the concepts, methods, technologies, treatments, services, or preventative interventions that drive this field?

By offering grant programs and RFPs that target data scientific investigators – irrespective of their actual, documented ability to have any effect whatever on biomedical research or the NIH mission of improving health – this Plan has some potential to drive Data Science as a field as far off course in the support of biomedical research as is conceivable. There is no reflection at all on prior NIH initiatives, and no consultation with the actual Data Science community, in the crafting of this Plan; so anywhere this Plan drives the field of Data Science will reflect the will of those who know very little, if anything, about the domain. Moreover, "successful completion of the goals" set forth in this Plan is **tautologically** determined: all of the goal evaluations involve counting events that NIH must already do: numbers of funded initiatives, numbers of initiatives that meet NIH-specified criteria, and other countable events with no consideration of whether incrementing the count by 1 or 1,000 is desirable. The impact that this Plan can have on the field of Data Science is negative, if it is non-zero. The impact the Plan can have on biomedical research is similarly limited.

Are the scientific rationale and need for <this proposal> well supported by preliminary data, clinical and/or preclinical studies, or information in the literature or knowledge of biological mechanisms?

Neither rationale (scientific or otherwise) nor need for this Plan is articulated. As noted, no reference is made to NCBC or BD2K, or to either failures or successes in those projects. So although NIH does have the ability to reflect on whether these initiatives had any effect at all, much less the desired effects, this Plan contains no rationale, no preliminary data, and no information or knowledge about the actual field that is emerging as Data Science. There is also no real representation of the role Data Science (in reality or as conceptualized in this Plan) can actually make for biomedical research.

For <trials> focusing on clinical or public health endpoints, is this <clinical trial> necessary for testing the safety, efficacy or effectiveness of an intervention that could lead to a change in clinical practice, community behaviors or health care policy?

There is no evidence or even suggestion that this Plan is warranted, needed, or necessary. No hypotheses about the effectiveness or efficacy of the execution of this Plan, which are all testable, have been put forward. There appears to be no interest at all in the NIH for justifying this plan rationally or with evidence. No one would ever propose to influence health care policy or clinical practice using “strategic plans” with the same total lack of awareness of what needs to be done, and how to do it, that this Strategic Plan for Data Science represents.

For <trials> focusing on mechanistic, behavioral, physiological, biochemical, or other biomedical endpoints, is this <trial> needed to advance scientific understanding?

The point of departure for this Plan is that data scientists should be spending less than 80% of their time wrangling data, without any consideration of the accuracy of that statement or its relevance for biomedical research. However, “do biomedical research” or indeed “do scientific research” is NOT one of the activities on the list of how data scientists spend their time, so spending less time wrangling data will not lead to “more time doing research”. Moreover, there is literally no need at all for this Plan to be executed in order to advance the understanding of Data Science; perhaps an RFP asking for informed, Data Science community-driven research into why data scientists spend so much time wrangling data (e.g., is that **because it is their job** OR is it **preventing them from doing their job**) would be informative -although reviews of applications responding to that RFP would require far more contextualization and articulated relationship to biomedical research than this Plan has. So far there appears to be no “advance of scientific understanding” that has accrued to the NCBC and BD2K initiatives. This Plan seems to strategize how to simplify the data scientist’s job – but has no real potential to affect the biomedical researcher’s job.

2. Investigator(s) **9**

Strengths

None noted.

Weaknesses

Are the PD/PIs, collaborators, and other researchers well suited to the project?

Given that this Plan is clearly lacking any input from actual, practicing data scientists, and the first group to whom “workforce development” activities are directed are NIH employees (Objective 4-1), there is clearly no one at NIH with the required qualifications for overseeing this Plan’s implementation, much less its drafting. The weaknesses in the alignment of the Plan goals and evaluations underscore the lack of qualifications of anyone drafting or reviewing this document at NIH to oversee its implementation or even its revision.

If Early Stage Investigators or those in the early stages of independent careers, do they have appropriate experience and training?

Data Science is a new domain. In spite of millions of dollars in funding, NIH has 20 years of experience (NCBC, BD2K) in this sort of infrastructure, but this proposal suggests it has still not moved any needle to promote or improve the discipline of Data Science. NIH has no evidence of the experience appropriate to undertake to guide or even influence Data Science. Meanwhile, statisticians, computer scientists, bioinformaticians, and business applications have all been actively defining Data Science and making tools and resources FAIR and accessible – mostly without interference or support from NIH.

If established, have they demonstrated an ongoing record of accomplishments that have advanced their field(s)?

With no data whatsoever on, and no reflections on the successes or failures of, the NCBC and BD2K initiatives, anyone at NIH who has been working in or around Data Science sufficiently long to be considered “established” does not appear to have advanced their field in any way. Similarly, the NIH has not demonstrably improved biomedical research or the abilities of biomedical researchers to do anything more than contribute yet more data to existing resources where such data are stored.

If the project is collaborative or multi-PD/PI, do the investigators have complementary and integrated expertise; are their leadership approach, governance and organizational structure appropriate for the project?

No. The team that authored this Draft Plan does not have the requisite expertise to lead a meaningful initiative that engages Data Science for biomedical research. The Plan represents a distinct lack of awareness of what Data Science as a field is and how it can be harnessed to support the NIH mission. This Plan could very well be used to find and fund investigators who similarly do not understand how Data Science can be used to strengthen biomedical research, simply by using this Plan to structure evaluations of the grant proposals that are submitted to whatever initiative this Plan may support. That would be a disappointing waste of money and resources. This Plan does not instill confidence that NIH has expertise or leadership sufficient to create and direct an initiative that meaningfully or effectively engages Data Science for biomedical research. Whomever commissioned or wrote this Plan needs to abandon it and start again. Ideally, a new Strategic Plan would focus on how Data Science can be leveraged for biomedical science, which would engage leaders and practitioners in the actual Data Science community who specifically support biomedical research. If the plan were drafted with independence, rather than with input from those who have had -and would rather maintain- NIH funding, it would be a promising start. The leadership that led to this Draft Plan needs to start over with a realistic and plausible Strategic Plan for engaging Data Science specifically in the support of biomedical research.

With regard to the proposed leadership for the project, do the PD/PI(s) and key personnel have the expertise, experience, and ability to organize, manage and implement the proposed clinical trial and meet milestones and timelines?

Unfortunately, NIH chose to draft this document themselves and ask for input later, rather than asking actual data scientists and actual biomedical researchers to draft something on which THEY, the NIH, could comment. The Draft Strategic Plan for Data Science, as a document, concretely supports a conclusion that the expertise, experience, and abilities required for success are absolutely *lacking* at NIH. No meaningful milestones are included and those that are reflect NIH priorities (counting up the number of things NIH does/values) rather than priorities in the scientific communities around data or biomedical research.

Do they have appropriate expertise in study coordination, data management and statistics?

Clearly not – see Objective 4-1. This Plan also underscores the lack of expertise in the NIH staff to coordinate, manage, or analyze the results of this Plan's implementation. The type of training in data intensive methods that the NIH appears to favour has been documented not to work (PNAS 2017, http://www.pnas.org/content/pnas/114/37/9854.full.pdf). Thus, not only is there

inadequate experience at NIH to coordinate or even propose strategies relating to Data Science, this Plan does nothing to support the formal and meaningful preparation of new **or experienced** biomedical researchers to work effectively with large data sets or with data scientists. As with the BD2K initiative, a focus on "training new researchers" when experienced researchers who use a new type of methodology/technology DO NOT YET EXIST is absurd. This Plan continues that absurdity - acknowledging that Data Science is a new and dynamic field (albeit one that the authors of the Plan know little about) but ignoring the fact that experienced researchers - the best prepared biomedical investigators to incorporate data-intensive methods into ongoing and productive research programs - should probably be the first who are specifically trained and supported to engage with the new and dynamic field that is Data Science. There is highly *inappropriate* "expertise" in anything that promotes the belief that new biomedical researchers - those who are not yet very experienced investigators - can simply attend a course on Data Science and suddenly have a totally new perspective on how to do the science they're really just learning to do *at all*. Data Scientists may be able to train workshop participants in data-intensive methods, but because they're not biomedical scientists, they won't be very capable of training participants in **biomedical research** that uses data. This may in fact be why there are no real results from BD2K showing positive impact on biomedical research (and, see http://www.pnas.org/content/pnas/114/37/9854.full.pdf).

For a multicenter trial, is the organizational structure appropriate and does the application identify a core of potential center investigators and staffing for a coordinating center?

In this case, if NIH did identify such a core, it would most likely represent individuals intent on obtaining and/or maintaining funding, and not on improving Data Science or supporting the evaluable integration of Data Science into biomedical research. There is literally no evidence in this Plan that NIH values the evaluable integration of Data Science into biomedical research; the goals are not aligned with this overall objective and the evaluation metrics that are proposed are orthogonal to that objective. Contributors who are independent of NIH should be solicited to start over, to draft a Strategic Plan for the evaluable integration of Data Science into biomedical research. This plan cannot be salvaged to accomplish this.

3. Innovation **9**

Strengths

None noted. This is NCBC 3.0 or BD2K 2.0 – not innovative in terms of prior NIH work. Similarly, the goals that are articulated are all achievable, if not already achieved, by the global biomedical research community leveraging data intensive methods, tools, and resources.

Weaknesses

Does the application challenge and seek to shift current research or clinical practice paradigms by utilizing novel theoretical concepts, approaches or methodologies, instrumentation, or interventions?

The only paradigm this Plan will shift is to deflect real Data Science *away from novelty* to promote data wrangling and prioritize that ahead of science. No theoretical concepts, approaches, or methodologies relating to data or biomedical science are mentioned in this document (see Weaknesses in Investigators, above). Moreover, this Plan seems to encourage innovation in technology with no purpose except to make data wrangling and management easier for non-data scientists who then have no ability to use that data to improve health or their own research paradigms into biomedical problems.

Are the concepts, approaches or methodologies, instrumentation, or interventions novel to one field of research or novel in a broad sense?

Nothing in this Plan is either innovative or actually reflective of the realities of Data Science as a discipline. Moreover, the Plan is also not sensitive to, or reflective of, the requirements of biomedical researchers who are **actually scientists now** to utilize the data that the NIH plans to wrangle for them.

Is a refinement, improvement, or new application of theoretical concepts, approaches or methodologies, instrumentation, or interventions proposed?

This Plan is a statement of vision, so no concrete refinements would be expected. What IS expected is that consideration of theoretical concepts, approaches, methodologies, instrumentations, or interventions from Data Science –and specifically, their

relevance to biomedical science – would have been proposed. Instead, there is only evidence that the drafters of the Plan do not understand either Data Science or its role in biomedical research.

Does the design/research plan include innovative elements, as appropriate, that enhance its sensitivity, potential for information or potential to advance scientific knowledge or clinical practice?

No. Even as a strategic vision, the Plan is irrelevant for Data Science as a field, for bioinformatics, and for biomedical research – mostly due to the lack of awareness so well captured in the document of what these three domains do and need relating to data, data intensive methods, and the harnessing of these to improve health and well being through biomedical research. Because NIH seems/seeks to reformulate how Data Science as a domain is perceived in the biomedical research community, the real potential is for a deviation of those funded by programs created based on this Plan from what the rest of the communities engaged with data/Data Science will be doing going forward. This Plan therefore has a worrying potential to wrongly direct scientists as it inappropriately diverts funding and attention from actual Data Scientists whose work would otherwise be fundable and useful in biomedical research.

4. Approach **9**

Strengths

None noted.

Weaknesses

Not only are no strengths noted in the Approach, and the Approach has inexplicable weaknesses that could actually result in negative momentum and ultimately, damage to Data Science and the reputation of biomedical researchers who actually need data intensive methods to do their work. The total lack of attention in the Plan to rigor and reproducibility as hallmarks of competent biomedical science is an unwelcome surprise. While data management and technology are engineered by those with training specific to practice in the domain to be rigorous and reproducible, these features of the data/data management/data wrangling systems do not translate simply to the science based on those systems.

Are the overall strategy, methodology, and analyses well-reasoned and appropriate to accomplish the specific aims of the project?

This Strategic Plan is actually relevant to NIH, and not to Data Science as the title suggests. The overall strategy may be acceptable for NIH human resources and training, but not for the domain of Data Science. The methods of evaluation are totally inappropriate for such a wide ranging and probably-costly initiative, and also inconsistent with formal program evaluation approaches that have been articulated by other segments of the federal government. Instead, there is no plan whatever for any real evaluation of the proposal. Every "evaluation plan" for a goal is basically a description of the items they plan to count. By contrast, the NSF has a program evaluation handbook (https://www.purdue.edu/research/docs/pdf/2010NSFuser-friendlyhandbookforprojectevaluation.pdf)
and the OPM has a "beginner's guide" to program evaluation (https://www.opm.gov/wiki/uploads/docs/Wiki/OPM/training/Program%20Evaluation%20Beginners%20Guide.pdf) that highlights questions **any reader of this Strategic Plan** would like to know the answers to at least half-way through the initative, namely:

- Does the program work? And how can it be improved?
- Is the program worthwhile?
- Are there alternatives that would be better?
- Are there unintended consequences?
- Are the program goals appropriate and useful?

NONE of these will be addressed by the “evaluation” plans in this Strategic Plan. It is worth noting that counting up the number of grants that are funded, or the number of times keywords articulated in this Plan are used in proposals chosen for funding because they are aligned with this Plan, cannot possibly be informative about whether “the program is working”. Those counts will always be greater than zero - i.e., will always increment - simply because NIH is doing its usual work of funding proposals that are responsive to NIH RFAs. It is also imperative to point out that those counts cannot ever be used to determine how a program can be improved.

The CDC also has a systematic approach to evaluation (from 1999):

1. Engage stakeholders

2. Describe the program.
3. Focus the evaluation. *****
4. Gather credible evidence. *****
5. Justify conclusions. *****
6. Ensure use and share lessons learned. *****

https://www.cdc.gov/eval/framework/index.htm

***** emphasis added to highlight the fact these are ignored in the Strategic Plan and do not appear to have been considered.

NONE of the features of the CDC or OPM evaluation processes have been considered in the drafting of this Plan or its evaluation elements.

The Obama white house issued a memorandum in 2012 describing the importance of "credible evidence of the effectiveness of the program" – and "describe how the agency plans to demonstrate or validate impact, or otherwise learn from the initiative, and how the agency plans to act on the new information"
https://obamawhitehouse.archives.gov/sites/default/files/omb/memoranda/2010/m10-32.pdf
and they emphasized **the importance of impact evaluations**.
https://obamawhitehouse.archives.gov/sites/default/files/docs/erp_2014_chapter_7.pdf

By contrast, this NIH Strategic Plan demonstrates a total lack of "cultural competence" with respect to either Data Science as a discipline or to the role of data in biomedical research. Together with a total failure to consider formal (or even useful) evaluations of this initiative, it suggests that anyone involved in the drafting of this Plan is totally unqualified to do this particular job).

Have the investigators presented strategies to ensure a robust and unbiased approach, as appropriate for the work proposed?

The opposite is actually true. None of the evaluations are robust and nothing in the document is sufficiently contextualized in actual Data Science to qualify as "unbiased" or robust.

Are potential problems, alternative strategies, and benchmarks for success presented?

Not only are potential problems and alternatives not considered, the NIH's own history with NCBC and BD2K are also not considered for drawing useful lessons. The benchmarks for success that are included are to simply count up the number of times the NIH actually does its job, funding projects where the stated Plan outcomes are featured, but never determining if any benefit to science, society, or health ever accrue (like NIH has done with NCBC and BD2K). The "benchmarks" are actually counts, with no indication of how high a number would be considered "good" or even "satisfactory"; and similarly, "success" in any of the evaluations appears to be a nonzero count of whatever non-impactful yet countable event is indicated. The weaknesses in the approach stem profoundly from a failure to contemplate plausible indicators of positive impact, but also from the total failure to consider formal evaluations that are easily done if planned from the outset. The total lack of any evaluation or even reflection on prior NIH data intensive initiatives eliminates any enthusiasm for any reader with the sole exception of the reader who has no intention of proposing impactful work in their own submissions to this initiative.

If the project is in the early stages of development, will the strategy establish feasibility and will particularly risky aspects be managed?

Absolutely no consideration whatsoever has been given to feasibility or managing risk. This is the 3rd consecutive 10 (or so) year initiative in data intensive programming from NIH, making the lack of this consideration totally unacceptable.

Have the investigators presented adequate plans to address relevant biological variables, such as sex, for studies in vertebrate animals or human subjects?

This is possibly the only NIH review criterion that isn't specifically relevant for the evaluation of this Strategic Plan.

If the project involves human subjects and/or NIH-defined clinical research, are the plans to address 1) the protection of human subjects from research risks, and 2) the inclusion (or exclusion) of individuals on the basis of sex/gender, race, and ethnicity, as well as the inclusion (exclusion) of children, justified in terms of the scientific goals and research strategy proposed?

The project undoubtedly involves data from human subjects and absolutely no consideration whatsoever is given to them. Moreover, although the 2009 NOT-OD-10-019 states, "**NIH requires that all trainees, fellows, participants, and scholars receiving support through any NIH training, career development award (individual or institutional), research education**

grant, and dissertation research grant must receive instruction in responsible conduct of research", according to the 2013 FAQ, this is actually NOT true*, not everyone* has to have RCR - according to #219 specifically,

"**Does the education requirement apply to awards that do not involve human subjects?**

No, but it is important for all investigators, even those working with tissues or specimens derived from human sources to understand when proposed research triggers regulatory and policy requirements.

Human subject as defined in 45 CFR part 46 means a living individual about whom an investigator (whether professional or student) conducting research obtains: (1) data through intervention or interaction with the individual, or (2) identifiable private information.

Research using human specimens, tissues, or data that are unidentifiable may not be considered human subjects research. See: http://www.hhs.gov/ohrp/policy/cdebiol.pdf (PDF - 24 KB).

Investigators who conduct studies with human specimens, tissues, or data that are determined not to involve human subjects are not required to fulfill the education requirement."

(Emphasis added)

Study Design

Is the study design justified and appropriate to address primary and secondary outcome variable(s)/endpoints that will be clear, informative and relevant to the hypothesis being tested?

This vision statement may be justified for planning the “workforce development” of NIH staff, but it is wholly inappropriate for the domain of Data Science. It also does not promote real engagement by biomedical researchers with data or Data Science/scientists. It does not promote engagement with biomedical research by data scientists. The Plan does include clear “outcome variables”, but these count-based metrics do not support conclusions about relevance or impact of any activities that do - or do not - increment those counts.

Is the scientific rationale/premise of the study based on previously well-designed preclinical and/or clinical research?

The most important weakness here is that the previous two 10-year initiatives are not reflected on in any way in this Plan. Apparently, either nothing at all was learned from NCBC and BD2K or what was learned is being ignored in this current Plan.

Given <the methods used to assign participants and deliver interventions>, is the study design adequately powered to answer the research question(s), test the proposed hypothesis/hypotheses, and provide interpretable results?

This Strategic Plan is not devised in any way that can be considered “adequately powered”. The NIH-centric thinking behind the strategy, the absence of any formal outcomes that could be considered meaningful, and the lack of reflection by the authors of the strategy all point to tremendous weaknesses and a design that is completely inadequate. This Plan is not adequate to accomplish anything apart from supporting NIH Data Science workforce development, although given its own evaluation plans for such a goal, that will also not be accomplished in a meaningful way. Since no “results” of NCBC and BD2k have been interpreted or reported – ever, it is unreasonable to believe that this Strategic Plan would generate “interpretable results”. The example metrics given to evaluate achievement of stated goals are inherently uninterpretable: if the count is 1 (one event), that gives no information; if the count is 100 or 1000, the counts are equally uninterpretable. The circularity of these metrics is that if the NIH does fund any projects, then all of these counts will increment without actually providing any information about whether this Plan, its goals, or the NIH mission, are actually supported.

Is the trial appropriately designed to conduct the research efficiently?

No thought has been given to efficiency. Since the goals do not represent real issues – hard problems that are actually worth solving to improve biomedical research as well as health and well being – and there is no real plan for “evaluation” beyond counting up meaningless items, there is literally no sense in which any research done according to this Plan could be considered “efficient”.

Are potential ethical issues adequately addressed?

Absolutely not. See prior comment.

Is the process for obtaining informed consent or assent appropriate?

This Strategic Plan having been drafted – being actively considered by NIH for adoption apparently without ever having consulted with actual data scientists or biomedical researchers working with data/data intensive methods and resources, underlines that the process for obtaining assent from the communities that would arguably be most effective (and almost uniformly negatively) is NOT appropriate.

Is the eligible population available?

Oddly, all those who could have provided meaningful input to the process of drafting a real, useful, and evaluable Strategic Plan for Data Science in biomedical research were available – but for some reason, were not contacted by NIH.

Are the plans for recruitment outreach, enrollment, retention, handling dropouts, missed visits, and losses to follow-up appropriate to ensure robust data collection?

No. In fact, this Plan appears to have been designed specifically to ensure that data collection and storage are all supported while simultaneously ensuring that no biomedical researchers will ever really learn how to use those data. The Strategic Plan also includes nothing about encouraging NIH reviewers to prioritize, rather than penalize, grants that propose to analyze existing data.

Are the planned recruitment timelines feasible and is the plan to monitor accrual adequate?

Since no experienced or informed input was sought for this Strategic Plan, it is no surprise that whatever initiative this Plan ultimately seeds will not be feasible or adequate; moreover, the Plan specifies that whatever the result is, it will not be evaluated formally or meaningfully. Therefore, it is unlikely that anything resulting from this Plan will be feasible or adequate.

Are the plans to standardize, assure quality of, and monitor adherence to, the trial protocol and data collection or distribution guidelines appropriate?

There are no such plans, not because this Strategic Plan is a statement of NIH's vision, intended to be forward looking and describe the ideal state with respect to integrating Data Science with biomedical research (or even supporting this integration), but rather because there is insufficient expertise or experience to even conceptualize the field of Data Science in the future.

Does the application propose to use existing available resources, as applicable?

Not only are existing available resources not discussed – including evidence about the impact of prior initiatives, in terms of their strengths and weaknesses – but neither are prior experiences of NIH in data intensive initiatives mentioned or discussed. What *is* mentioned in Objective 4-1 is specifically that the NIH workforce *needs training in Data Science* - and is therefore NOT a resource that could be utilized. This Plan doesn't use the NIH's prior experience as a resource, and also doesn't mention institutional lack of experience as a limitation to be overcome. Neither is there any mention made of multiple initiatives and resources that are *already* freely available worldwide, and no mention is made of the lessons learned from any of those, either. This is a truly uninformed Plan, and that lack of informedness highlights the similar lack of plausibility of the NIH as an author of (or contributor to) a "Strategic Plan for Data Science", and undermines any enthusiasm for anything this particular Plan might eventually turn into.

Data Management and Statistical Analysis

Are planned analyses and statistical approach appropriate for the proposed study design and methods used to assign participants and deliver interventions?

The Plan is notable for the total lack of any mention of planned analyses; the specific disregard for appropriateness of statistics or statistical thinking/reasoning in how Data Science and biomedical research could interface removes all enthusiasm any reader might have.

Are the procedures for data management and quality control of data adequate at clinical site(s) or at center laboratories, as applicable?

The Strategic Plan, as noted earlier, is striking in its total lack of “quality control”. The evaluations that are proposed are stunningly inadequate, and the failure to use the NIH’s own prior experience -or apparently that of any informed participant in either Data Science or biomedical research requiring data scientific methodologies – suggests there *can be no quality control* in this Plan if it is executed.

Have the methods for standardization of procedures for data management to assess the effect of the intervention and quality control been addressed? Is there a plan to complete data analysis within the proposed period of the award?

Based on the total lack of reflection on prior efforts by the NIH, and the stark absence of input from informed data scientists or biomedical research requiring data scientific methodologies, the standardization of ANYTHING resulting from this Strategic Plan is truly worrisome. As noted earlier, among the worst features of the Plan is that if it is implemented, either those who are funded and follow this Strategic Plan will be/become the least knowledgeable Data Scientists – isolated from practicing communities by adhering to an absurd vision that is not grounded in reality; or they will gain absolutely nothing beyond having had a grant funded under this Plan. Neither of those is a desirable outcome; so this Plan should not be used or even “revised”. This Plan should be scrapped.

5. Environment **9**

Strengths

NIH has money to support data intensive initiatives, apparently.

Weaknesses

Overall, the NIH has demonstrated it has insufficient qualifications to strategize about Data Science. The document lacks any form of meaningful evaluation, and although counting up the number of times the NIH grants funding to applicants is apparently a key metric for NIH success, this has absolutely no relevance to the taxpayer or to the Data Science community.

In addition to excluding any meaningful evaluation of the impact of this Plan, this Plan also excludes any consideration of prior similarly data-intensive Plans and initiatives from the same environment (NIH). The failure to recognize what works and what does not work when NIH strategizes about data marks the NIH as a superlatively weak environment in which to propose any sort of strategy for Data Science.

This Strategic Plan is NIH centric and as such, essentially unrelated to the actual work required for effective and impactful integration of Data Science into biomedical research. The significant and profound weaknesses in this Plan underscore the marginality of the NIH environment for proposing, much less overseeing, a plan for that kind of integration.