Harvard Data Science Review • Issue 3.1, Winter 2021

What Are the Values of Data, Data Science, or Data Scientists?

Xiao-Li Meng

Published on: Jan 29, 2021

DOI: <https://doi.org/10.1162/99608f92.ee717cf7>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

Many years ago, I was asked to provide statistical consultation for a grant project on mental health. The PIs (Principal Investigators) of the grant brought me in to help them to deal with many non-responses to their national surveys, for which they were unsure about the results provided by some research assistants. A mutual friend had bragged about my reputation as a ‘multiple imputer,’ i.e., someone who could magically fill in missing data multiple times. As much as I knew that the safest way to keep one’s status as a legend is to remain a legend, I couldn’t let my friend down. After looking into the problems and spending a considerable amount of time with the research assistants, we arrived at a set of results that were, indeed, quite different from the original ones. The PIs were very excited until they saw the new results. Their ensuing question—“Wait, Xiao-Li, why did we pay you to make our confidence interval wider?”—was not entirely sarcastic.

As much as I wish that this anecdote were only of humorous value, in fact, it highlights a thorny question about the value of statisticians, and now, more broadly, that of data scientists as well. The five *imaginaries* posited by philosopher Sabina Leonelli, in her most thought-provoking [“Data Science in Times of Pan\(dem\)ic,”](#) together with seven discussions from thought leaders with diverse perspectives, should compel all citizens of the data science community to ponder deeply a series of questions about values, including the question of what kinds of values we should prioritize, considering the various meanings of the term *value*. What are the values of data, data science, or data scientists? How might we best draw from these values under severe constraints, such as a pandemic? Should we trade non-negotiable, long-term values for situational, short-term ones during a crisis, or does that assume a false dichotomy to begin with? Should data scientists be value-driven, neutral, or some strategic mixture of both, if we take the neutral vs. value-driven distinction to be yet another false dichotomy?

Answers to these questions are complex, and some will be controversial. The challenge that I mentioned of dealing with survey non-responses is almost negligible compared to the profound challenges that the world currently faces due to the COVID-19 pandemic and its ramifications. Nevertheless, to illustrate the complexities entailed in answering questions pertaining to values, let’s return for a moment to the anecdote of my consulting experience, for the sake of providing a concrete context for contemplation. More specifically, let’s put ourselves in the shoes of the PIs who sought my help. Since grant funding is always tight, the justification for bringing a consultant on board has to be that it will result in some measurable increase in value that could not otherwise be obtained. In theory, the added value of having a statistical consultant would be better statistical results. Isn’t that obvious?

Not really. Whereas few would pay for deterioration (other than with spiteful intent), there is an elephant in the room, or rather in the ‘obvious’ statement above: *better* for whom, or for what? For our research publications? Our grantor? Our patients? Our profession?

A rhetorical answer might be: ‘As long as we do the right thing, it will be better for everyone.’ Unfortunately, this merely kicks the elephant into the context of yet another vexing question of values: what is the *right* thing? If by ‘better’ statistical results we mean results that are more likely to be published, then the right strategy

would seem to be to make the results as statistically significant as possible, with the apparently noble aim of ‘doing your best with the data.’ Yet, in practice, ‘doing your best with the data’ may give some analysts the perceived license to cherry-pick data, methods, results, or interpretations to inflate significance and exaggerate evidence. That no longer sounds all that noble or scientific, does it?

What Should Data Scientists Value?

Dealing with uncertainty, even in the most rigorous way, inevitably entails making some educated guesses. A guess is just a guess—as time goes by—and it always leaves room for exploration and exploitation. Indeed, uncertainty is a natural Petri dish for cherry-picking. We all have seen (and perhaps have also used) the tactic of reporting ‘as low as’ (quoting the left end of a confidence interval) or ‘as high as’ (quoting the right end of the same confidence interval), depending on which way the reporter intends to (mis)lead the readers. The reporter may have noble (e.g., to avoid unnecessary panic) or sinful (e.g., to escalate fear) reasons for doing so, and further cause to excuse oneself after the fact, since this practice does not involve choosing data, methods, or results, but rather, merely the best means of (mis)communication.

Yet things can go very wrong merely by selectively reporting. As Leonlli reminded us vividly, with respect to the current pandemic, “Leading pre-print repositories such as bioRxiv and medRxiv were inundated by manuscripts reporting on modeling results and related predictions, many of which had not yet been peer reviewed nor validated, and yet were quickly picked up as reliable findings by mainstream media looking for scientific evidence for specific political interventions.” The lethal pairing here is that of “quickly picked up” and “for specific political interventions,” which suggests a clear retrofitting to support a pre-chosen policy or conclusion. If one tortures data enough, the data will confess. Sadly, this statistical joke is no laughing matter here. When individuals or entities, be they political or not, ascribe value to data science as a means for advancing their already well-set agenda, it poses a conundrum for those of us who are capable of meeting their value propositions. Should we do it?

‘Wait, Xiao-Li, what’s the conundrum here if I support their agenda?’ you may ask, and add, ‘I’d only see difficulties if I’m going to be paid to do things that I don’t believe in or find morally objectionable.’ Correct, but in the latter case at least you would know that the right choice is to walk away. But if compelling factors, such as financial necessity, cause you to take the job, how (and how much) would or should your moral objections influence your professional judgment as to data science-related choices? Would you deliberately conceal or down-weight reliable evidence, which can be done seamlessly in skillful hands, that happen to support an agenda with which you disagree? Or would you provide the client with whatever service you are being paid to do, and suppress your personal value judgment?

Conversely, and more challengingly, say that you like their agenda and want to support it. But you also know—if you are a competent data scientist with basic professional ethics—that seeking data or methods solely for the purpose of supporting pre-chosen conclusions completely devalues data science as a trustworthy evidence

builder and erodes public trust in its credibility, regardless of the nobility of the conclusions or the lack thereof. This is a tougher conundrum because it is now *you*, not just them, who want the evidence to support the predetermined conclusions. You therefore will be negotiating between your moral and social principles and your conduct as a data scientist. Which one should be given higher value?

Whose Value Is It Anyway?

In the case of my own consulting experience, I valued the principles and integrity of my profession more than anything else, because I was brought in as a *statistical* consultant. The confidence intervals ended up wider because the original ones failed to take into account the uncertainty about the mechanism that had led to the non-responses. I do not blame the research assistants for not having been able to fix this problem on their own. It could not have been solved by any off-the-shelf software, and it was not simple enough to have had one definitive correct answer, since the mechanism could never have been fully understood. It therefore required practical wisdom derived from experience, a deep understanding of the substantive questions, and judicious judgment to provide substantively and statistically more principled results. I took into account as much uncertainty as could not be assumed away on the basis of evidence or sound judgment. The end result, predictably, was that the resulting confidence intervals were wider than before, implying that some of the treatment effects that the PIs had been hoping to establish were less clear.

While I didn't provide the value that the PIs had initially expected—namely, to strengthen the statistical significance of their results—they did appreciate the fact that the more cautious analysis spared them the possible professional embarrassment of announcing findings that could later have been found irreproducible (and this possibility turned out to be not hypothetical in one instance during my consultation period), which is of long-term professional value. When we go beyond personal and professional value propositions, however, the questions and answers become even more complex. For example, what's the value added by the more cautious analysis to the patients potentially affected by the aforementioned study?

Whereas it was my professional judgment that the original results were too optimistic, they still had a decent chance of tending in the right direction; indeed my more cautious analysis did not conclude they were wrong, but merely that there were more uncertainties about the findings. Delaying an effective treatment can unnecessarily prolong suffering for its potential beneficiaries. On the other hand, administering an ineffective treatment because of false data analysis would also inflict suffering, both physically and financially. The trouble is that, when facing uncertainties, especially inherently challenging ones (e.g., unobserved confounding factors in causal studies), there is no harmless strategy: the gain from one consideration is made at the expense of another. For example, minimizing false positive rates will increase the false negative rates, and vice versa.

Difficult trade-offs therefore need to be made, and this is where things can be deadly controversial—pun intended—when lives and livelihoods are involved, especially on a massive scale. As Leonelli asks, “What are the priorities underpinning alternative construals of ‘life with covid’? ... Whose advice should be followed,

whose interests should be most closely protected, which losses are acceptable and which are not?” Such questions clearly cannot (and should not) be answered by data science or data scientists alone, but the data science community has both the ability and responsibility to establish scientific and persuasive evidence to help to reach sustainable compromises that are critical for maintaining a healthy human ecosystem.

The Persuasive Value of Data Science

Persuasion is not about winning arguments. ‘Winning’ implies that there is a losing side, and few like to be a loser, especially when politics are involved. Persuasion is mostly about convincing others that our ideas or plans are better than theirs for *their* welfare, and hence, about resolving conflicts via elimination rather than suppression, which is rarely sustainable. Data science can persuade via the careful establishment of evidence from fair-minded and high-quality data collection, processing and analysis, and by the honest interpretation and communication of findings. In practice, these desiderata are rarely all achievable in a single study—the lack of high quality data is a common pitfall, for example (more below). General trust in data science processes (in contrast to data science products) then becomes critical in sustaining the persuasive power of data science. Because desires and aspirations often deviate substantially from reality, our results are not free from defects or flaws. But people can still trust us as long as they trust that the process is being guided by a commitment to exercising our best professional judgment, making a concerted effort to combat our own ideological biases and prejudices, and not engaging in cherry-picking or other unethical activities.

Leonelli’s punchline that “Emergency data science can be fast, but should never be rushed” is a point that cannot be over-stressed if we data scientists want to gain general trust, and retain our value as a force of persuasion in an increasingly politicized and divided world. Rushed data science deprives us of the necessary thought process and due diligence to ensure that we do not accidentally inflict long-term suffering to some or even to all in the name of dealing with crises. The complex issue of protecting individual privacy and autonomy versus controlling the pandemic through societal surveillance presents just such a concern, as well-articulated in Leonelli’s article. The technological power to monitor individual movements 24/7 is already here. But power almost always finds a way to be abused and so, to inflict abuse. Worse, in the absence of structural checks and balances, power can and often does corrupt people who have the best intentions to do good, but nevertheless fail to guard themselves from misusing it.

Indeed, [Jill Lepore and Francine Berman’s conversation](#) on the rise and demise of Simulmatics, which most of us had never heard of until Lepore’s book (*If Then: How the Simulmatics Corporation Invented the Future*), provides a fascinating account of some very early history of the good, the bad, and the ugly of the power of digital data, from a precursor of Amazon’s predictive model to a predecessor of Cambridge Analytica. The conversation also reminds us of the importance of learning from historical lessons to guard against abusive power before it is out of control. As Lepore puts it succinctly, “no, you don’t kick the can down the road, you don’t just invent Facebook and see if it destroys democracy,” because if you do, it will.

The Value of Data and Data Quality

Of course, to build persuasive evidence, we need data, especially high quality data. A number of articles in this issue of *HDSR* directly address the issue of data quality in a variety of settings. The column by John Thompson, a former Director of United States Census Bureau, emphasizes that [“Understanding the Quality of the 2020 Census is Essential.”](#) The 2020 US Census “faced an unparalleled set of challenges beyond the control of the dedicated professional staff at the Census Bureau” for reasons most of us can list, which makes the issue of data quality ever more pressing. Thompson’s article centers on the recommendations made by the Census Quality Indicators Task Force of American Statistical Association, which was charged to produce a set of “scientifically-sound, publicly-available statistical indicators by which the quality, accuracy, and coverage of the 2020 Census can be assessed.” Such assessments have wide-ranging consequences, considering the critical roles that census data play in our lives and in research.

Along related lines, [the historical account of the creation and development of exit polls in US elections, as told by Murray Edelman](#), a pioneer of the exit poll, regales readers with tales of the efforts made by earlier generations to improve the quality of data in the domain of election forecast. These efforts also provide a great example of how to gain public trust, and hence, collect more reliable data, via innovation. Exit polls were invented in late 1960s initially to deal with the issue of multiple time zones within a state (e.g., Kentucky). They also turned out to induce trust in anonymity since “you don’t know the person’s phone number, you don’t know their address, you don’t know anything about them,” and, as such, people were and are generally willing to provide honest answers, even for rather private information at a time when there were more concerns about disclosing it.

Another fascinating historical account of improving data quality, or rather account of reconstructing historical data, is provided by Duo Chan, a newly-minted Ph.D. in earth and planetary sciences, who documents his heroic effort in [“Combining Statistical, Physical, and Historical Evidence to Improve Historical Sea-Surface Temperature Records.”](#) It is a heroic effort because the reconstruction involves “more than 100 million ship-based observations taken by over 500,000 ships from more than 150 countries”; these data went back to 1850, but only after 1970 some data were collected with the intention to study global warming. I’d challenge the most imaginative minds among the readers to contemplate how many and what kinds of data quality problems one would encounter with respect to such a project. Just to tease—it helps if you can read historical Japanese documents, as it might just provide a clue to why the Atlantic appeared to warm twice as fast as the North Pacific in the early 20th century and why sea-surface temperature seemed to warm up suddenly during World War II.

Whereas all these articles help us to better appreciate the fundamental importance and value of quality data, the column by Stephanie Kovalchik, [“Why Tennis Is Still Not Ready to Play Moneyball”](#) reminds us that the value of data, in its literal sense, can also be an obstacle to data science. As Kovalichik suggests, the fractured data landscape in tennis is a result of the structural fragmentation of its business enterprise, which features multiple

competing promoters, meaning that “decisions about data collection and sharing do not aim to improve performance but to improve business.” Consequently, “[d]espite the trove of data that tennis possesses, its analytics revolution has never come because it has not done more to remove the obstacles that stand between the data and the players and coaches who have the greatest interest in their analysis.” This negative example speaks vividly to the inevitability of *politics in data science*, not just politics of data science, a point that Leonelli further emphasizes [in her rejoinder](#) to the seven discussions.

The Value of Data Scientific Framing

Speaking of data science and politics, this issue of *HDSR* also features an article by Daryl DeFord, Moon Duchin, and Justin Solomon on [data science for politics, specifically the problem of redistricting](#). This is probably as political as it gets for a data science problem, because its main goal is to help courts, commissions, and the public to balance the political interests of different parties. The interest of each political party in redistricting is clear: to design and implement a redistricting map that will maximize the number of wins for *my* party. Without regulation, this explicit desire, given one-party control, can lead to highly unfair—and often ridiculous-looking—district maps, the results of so-called gerrymandering.

To prevent such extreme manipulations, various rules and restrictions apply to redistricting plans, such as population balance, contiguity, and ‘compactness,’ a demand for reasonable shapes. There are at least two major challenges in implementing these rules. The first and most fundamental one is that these rules are often qualitative, and some are left intentionally vague to accommodate the complexity of (political) life; indeed often a legislation can be adopted only because it is sufficiently pliant to gain broad support. The second is that even if we can precisely operationalize all the rules and restrictions, there will still be an astronomically large number of permissible redistricting maps from which to choose.

The authors’ aim is not to obtain one or few redistricting maps that satisfy the rules, but rather to provide a large and *representative* sample of all such maps. A court, for example, may find value in a representative sample as a benchmark to determine if there is strong evidence to cast doubt on the fairness of a proposed redistricting plan, because it is an outlier skewed to favor one party relative to the possibilities afforded by the rules.

So, how should we sample? When the rules can be quantified precisely, we can first define a class of permissible redistricting plans, and then design an algorithm to sample *uniformly* from this class, however vast. In such cases, the data science problem becomes purely computational, which is far from trivial, but nevertheless we have a well-defined theoretical target.

The problem becomes much more intricate when (some of) the rules are too flexible to have a precise mathematical formulation. For example, as the authors point out, “most states have a ‘compactness’ rule preferring regular district shapes, but few attempt a definition, and several of the attempted definitions are

unusable.” How can then we take into account such vague and qualitative rules in creating the benchmark samples?

This is where sound and creative data scientific framing comes in. The lack of a clear definition of ‘compactness’ rules out binary framing, that is, regarding each plan as either compact or not. But much of the data science enterprise is to deal with subjects that are not just black and white, but with many degrees or shades of ‘grayness.’ Indeed, DeFord, Duchin, and Solomon designed an iterative sampling algorithm that gives more compact plans higher probabilities to be sampled, and hence they can effectively model the intended vagueness of the rule by a probabilistic mechanism. A substantial practical advantage is that such ‘grayish’ strategy can be computationally far more efficient than the ‘black-white’ uniform sampling, and it even has substantial advantages over classical approaches inspired by statistical physics.

Of course there is no free lunch. Avoiding binary framing puts pressure on us to argue for the suitability of our alternative metrics, which leaves room for both exploration and exploitation by researchers (as well as their algorithms). And robust theory is also needed to understand the convergence behavior of the resulting iterative algorithms. All in all, this gives us a telling example of how creative data scientific re-framing can address substantive and computational problems simultaneously, opening up exciting new research directions.

The Value of Data Science Education

Much of the value of data science lies in the use of its theory and methods to provide concrete and actionable solutions to complex substantive problems. To do this well, however, takes much more than merely learning theory and methods in classrooms or from books. It requires ‘data acumen,’ a central emphasis of the [US National Academies’ report](#) on undergraduate data science education. Gaining data acumen requires one to be exposed to the complexity and scope of data science process in actual practical settings, just as there is no other way of learning to swim than actually diving into the water, and trying to stay afloat while moving. Consequently, a central value of data science education must lie in training many skillful swimmers in the even expanding and stormier ocean of data.

I am therefore very pleased and proud that this issue of *HDSR* features a discussion article, [“Statistics Practicum: Placing ‘Practice’ at the Center of Data Science Education”](#) by Eric Kolaczyk, Haviland Wright, and Masanao Yajima from Boston University. While increasingly more statistics and data science programs include a practicum component, the vast majority institute it as a consulting or capstone course, providing “a final experience prior to leaving the program—with the implicit assumption being that the individual will ‘figure the rest out’ on the job.” The BU team argues that such a “final experience” training is insufficient to meet the ever-increasing demand for data science and data scientists from all sectors of our global community, and hence, they call for “a paradigm shift, placing a so-called Practicum course at the center of a data science program, intentionally organized as a hybrid between an educational classroom and an industry-like environment.” They back up their call with detailed documentation and a discussion of how BU implemented

such an idea in the past five years in its Master of Science program in Statistical Practice. Over 20 discussants—from current students to industrial leaders and to preeminent educators—provide wide-ranging perspectives, accounts of experiences, and suggestions on how to establish, sustain, and enhance practicum-centered data science education at all levels. For anyone who cares about present or future data science education, this featured discussion set is truly a must read.

The Value of Data Science Communication and Teamwork

Similarly essential is the column by Thomas Davenport and Katie Malone, [“Deployment as a Critical Business Data Science Discipline,”](#) which stresses the importance of “getting data science models deployed into production within organizations,” and how the success of deployment depends vitally on how early we integrate deployment planning into data science processes and training programs. As Davenport and Malone point out, “Starting with the algorithm first, and only at the end of the project thinking about how to insert it into the business process, is where many deployments fail.” This echoes the key point by the BU team, that when practical training is brought into play as the “final experience” of data science education, it comes too late.

Davenport and Malone also discuss how the success of deployment requires effective communication and mutual trust among data scientists, business managers, and other key stakeholders responsible for a given process or product. The necessity of effective communication was discussed in details in a [previous column](#) by Malone, and the importance of team work was the subject of another [insightful article](#) by Davenport. Together with yet another previous article in *HDSR*, [“How to Define and Execute Your Data and AI Strategy,”](#) by Ulla Krühse-Lehtonen and Dirk Hofmann, this collection of articles provides a clear picture of the vital value of communication skills and teamwork for fully realizing and enhancing the value of data science.

As another example of the effectiveness of teamwork, especially during the pandemic, [Power-Fletcher et. al. detailed their multi-disciplinary effort](#) to “better understand translational research within the context of pandemics, both historical and present day, by tracking publication trends in the immediate aftermath of virus outbreaks over the past several decades.” The scope of this study is such that it required substantive expertise in epidemiology, medicine, library sciences, and the humanities, because its aim is to identify and track language patterns in published literature. As the authors argued and demonstrated, both qualitative and quantitative expertise were essential for the successful implementation of an NLP (natural-language processing) based method for detecting patterns and trends, reflecting the anxieties and efforts to expedite and facilitate scientific discovery from the ‘bench’ to the ‘bedside’ to the wider population. This study also “demonstrates how the different sectors of biomedical research respond independently to public health emergencies and how translational research can facilitate greater information synthesis and exchange between disciplinary silos.”

Indeed, as I wrote in my [first editorial](#) for *HDSR*, data science has evolved into and as an ecosystem. Its full value therefore cannot be realized without ongoing and ever-expanding collaborative efforts from all its contributing disciplines and sectors. *HDSR*, with its aim to feature *everything data science and data science for everyone*, was launched exactly as a global forum for coordinating such efforts, by disseminating and exchanging relevant ideas, experiences, lessons, methods, results, and beyond. As it happens, the value of having such a synergistic forum has been just recognized by the Association of American Publishers, which has chosen *HDSR* as the winner of its [2021 PROSE award](#) for the best new journal in Science, Technology, & Medicine. I am extremely grateful to all the authors and reviewers for this issue and previous issues (we have published more than 150 pieces since the initial launch) for their tremendous contributions, and to the readers around the globe for your support and endorsement; our over a million page views came from 233 (out of a total of 236) countries and regions, with more than 50% outside of the US. Without all your contributions and support, *HDSR* could not possibly achieve such recognition and global presence in fewer than 20 months. So thank you all, on behalf of the entire *HDSR* team. You made *HDSR* valuable and, in return, we are more motivated than ever to provide you with excellent value for the precious time that you allocate to browsing and reading *HDSR*.

Until the next editorial, please help to spread the word about *HDSR* to the last three regions and countries: Kiribati, Vatican City, and Wallis and Futuna.

Disclosure Statement

Xiao-Li Meng has no financial or non-financial disclosures to share for this editorial.

©2021 Xiao-Li Meng. This editorial is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](#), except where otherwise indicated with respect to particular material included in the editorial.

Preview image by Riccardo Annandale