Data Science: A Happy Marriage of Quantitative and Qualitative Thinking?

Xiao-Li Meng^{1,2}

¹Department of Statistics, Faculty of Arts and Sciences, Harvard University, Cambridge, Massachusetts, United States of America,

²Harvard Data Science Review, Harvard Data Science Initiative, Harvard University, Cambridge, Massachusetts, United States of America

Published on: Jul 30, 2021

DOI: <https://doi.org/10.1162/99608f92.cee621a9>

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

‘Interesting rhetorical question, Xiao-Li! But isn’t it a meaningless one unless you can precisely define the terms *quantitative* and *qualitative*?’ If this was your first thought upon reading the title of this piece, then you just placed yourself on the quantitative-based thinking side of the quantitative-qualitative spectrum.

‘Well, Xiao-Li, you just did the same—only a quantitative thinker would try to arrange an orderly quantitative-qualitative spectrum!’ Very true. I was trained to be a quantitative thinker, and I surmise that’s the case for most people who currently feel comfortable calling themselves data scientists. But as the data science ecosystem evolves and expands, our comfort zones need to expand accordingly. Should then *qualitative thinking* be a part of our comfort zones as well? Here, the term ‘thinking’ is used with the same breadth and depth as it is in *computational thinking* and *inferential thinking*, as promoted in the [discussion article by Ani Adhikari, John DeNero, and Michael I. Jordan](#) in the last issue of *HDSR*; in particular, it includes methods and approaches that are guided by such thinking.

To be clear, qualitative thinking has always played an essential role in data science, as demonstrated by a series of articles in *HDSR* since its inaugural issue (e.g., [Borgman](#); [Leonelli](#); [Luciano and Cows](#); [Crane](#); [Pasquetto, Borgman, and Wofford](#); and [Gregory, Groth, Scharnhorst, and Wyatt](#)). Yet, broadly speaking, qualitative thinking currently does not receive nearly as much attention as quantitative thinking in data science research or education. The leading article in this issue, [“Why the Data Revolution Needs Qualitative Thinking”](#) by Anissa Tanweer, Emily Kalah Gade, P.M. Krafft, and Sarah K. Dreier, therefore, is timely. Drawing on qualitative social sciences, it introduces concepts that most of us do not use routinely to frame our thinking, research, or teaching, such as *interpretivism*, *abductive reasoning*, and *reflexivity*. Yet the processes described by these terms are rather relevant for data scientists, and many of us actually have engaged in such processes to various degrees at one time or another.

As a matter of fact, I must confess that the first time I understood what these terms describe, the thought that ‘Not a big deal—just some fancy terms describing what I have been practicing all these years ...’ came to my mind. But the better (more qualitatively sensible?) part of my brain came to the sense that *that* thought is not much different from a statistical novice’s reaction—upon learning that the maximum likelihood estimate (MLE) for a normal mean is simply the sample average—‘Oh what’s all the fuss about this MLE thing? I have been taking averages all my life!’

Machine Learning: An Algorithm, a Model, or a Function?

To remind ourselves of the importance and benefit of paying close attention to these ‘fancy terms’ in the social sciences and humanities, please allow me to share a recent experience about appreciating concepts and terms from different disciplines, especially when they seem to overlap with the ones for which we can claim expertise. A national committee I served on was discussing how to communicate to the scientific communities and the public about the impact of mathematical sciences upon scientific and societal advancement. The topic

of machine learning came up and the committee needed a pertinent technical concept to highlight the mathematical contributions. Terms such as ‘algorithm’ or ‘model’ came naturally to my mind, depending on whether I want to offend (or please) my statistics colleagues or computer science colleagues. To my surprise, the proposed term was ‘function.’

Function? Really? How many laypeople or even scientists have any grasp of the mathematical definition of the term? As my curiosity arose, so did my voice. But as the discussion went deeper, I started to recognize my disciplinary blind spots. The committee’s charge was to highlight the impact of mathematical sciences, and the most significant impacts are those that would not be possible without mathematical sciences. Many can (and do) build and implement machine learning algorithms without much mathematical rigor or understanding. But to establish their analytical properties and understand their general behaviors, especially the deep-learning ones, we need to dive deeply into the rich theory of iterated functions and recursive functions. And that is not something that can be advanced by most people, or by trial and error.

Technically, a machine learning procedure can be considered in terms of an *algorithm*, a *model*, or a (recursive) *function*. But these terms are not exchangeable, either for broad conceptualization or for in-depth inquiries. Each term signifies a distinctive line of inquiry with its own framing questions, investigative methods, theoretical insights, explorative intuitions, evaluative metrics, etc. They do not subsume each other, and indeed sometimes they pose competing inquiries (e.g., seeking predictive patterns verses identifying generative processes) under time and resource constraints. However, it is exactly their complementary and competing natures that make them indispensable for studies of, and in, machine learning, as it evolves into a viable scientific discipline.

My initial question, ‘Why would anyone even think about using the term function to describe machine learning?’ reflected my lack of appreciation of this big picture (notwithstanding the concern about effectively communicating technical terms to the public), and I am grateful to my mathematical colleagues for their enlightenment. Similarly, the concepts, methods, and examples from the qualitative social sciences, such as those documented in Tanweer et al., should enlighten those of us who have been too consumed by quantitative thinking. Even for those of us who have adopted some forms and habits of qualitative thinking into our data science inquiries, engaging them more systematically can further increase their usefulness and help us at least to avoid serious mistakes, such as those made about the inner workings of Wikipedia bots, a great lesson reviewed in Tanweer et al.

Enhance Sense and Sensibility of Data Science

Tanweer et al. categorizes concepts such as interpretivism, reflexivity, and abductive reasoning as *qualitative sensibilities*. And indeed, they are, injecting both *sense and sensibility* into quantitative data science, which is mostly about *patterns and processes*, such as seeking patterns for predictions and generative processes for understanding.

Specifically, in the data science context, *interpretivism* refers to an intentional effort to construct meaning in a given data study, not passively waiting for the data to speak, so to speak. Readers of [my first editorial](#) might recall a brief summary of philosopher [Sabina Leonelli's article](#), which emphasizes that there is no such thing as objective raw data. This point is extremely well reflected in the historical account by Kaneesha R. Johnson on “Two Regimes of Prison Data Collection,” who emphasizes that “The *content* of data that is gathered—what is being asked and recorded—and the initial intentions driving data gathering often dictates the parameters of our understanding.” (Emphasis is original.) The data science community therefore would be doing itself a great service, especially in the long-run, by putting as much effort into context-revealing deep probing as it has invested in pattern-seeking deep learning.

Setting appropriate contextual meaning is also a central focus of Tanweer et al.'s call for more systematic and reflective *abductive reasoning* in data science, in addition to inductive reasoning. They stated that “abductive reasoning—which is the process of using observations to generate a new explanation, grounded in prior assumptions about the world—is common in data science, but its application often is not systematized.” The need of contextual background then should be evident, as the same data can carry very different meanings and lead to different explanations in different contexts. Indeed, Tanweer et al provides a telling comparison between ‘labeling’ in machine learning and ‘coding’ in qualitative studies, illustrating that a labeling process without the careful qualitative-coding considerations of contextual meanings can be seriously misleading.

Reflexivity then is naturally another key qualitative sensitivity that the data science community should embrace. It refers to the process of examining how researchers' assumptions, experiences, and relationships influence their research, at both individual and collective levels. The quantitative data science community is broadly familiar with the concept and practice of checking assumptions, but it does so mostly with respect to technical and mathematical assumptions. We rarely conduct quantitative studies on the robustness of the findings relative to factors such as the ideological orientations, research experiences, collaborative capabilities of the research teams. Yet we all know these factors matter a great deal, at least in terms of our confidence in findings. Tanweer et al. provides a series of serious examples that are worthy of reflection for all of us. But here is a simple example. No matter how impressive a study on the benefit of wine drinking is, many of us would take the findings with a tall glass of water, if we know it is sponsored by *Chateau You-Fill-In*. Right there we know the researcher's affiliation and persuasion relationships matter.

I'd venture to say that the data science community, which understands well the danger of cherry picking, should also engage itself in *meta-reflexivity*. That is, we should constantly ask ourselves that ‘Do I engage in reflexivity exercise only when I want a certain outcome, or don't like a certain outcome?’ More broadly, ‘Do I scrutinize those findings that appear to support my ideology or theory just as much as I do for those that don't?’ If the answer is no, then we owe to ourselves an explanation, derived from conscientious professional introspection, when we put on our hats as data scientists. Such meta-reflexivity is a difficult process, especially when we strongly believe in our noble causes. But that is exactly when it is most needed, because data science

ethics should compel our findings to be evidence-driven, not results- or passion-driven. As an example, [Rudin, Wang, and Coker](#) in their discussion article in issue 2.1 of *HDSR* (including the rejoinder) revealed problems in COMPAS' public documents and ProPublica's highly cited results (as cited by Tanweer et al.), both of which could be readily identified and avoided by a serious meta-reflexivity process by the respective researchers.

Quantitative and Qualitative: A Matter of Resolution?

Tanweer et al.'s call also provided me with a quantitative reflection on how to systematically integrate qualitative thinking into data science. I would suggest to my fellow data scientists an analogy between qualitative thinking and the investigative process that we undertake, when we have very little data or no data at all. When we do not have much data to fit models, reveal patterns, suggest relationships, etc., we tend to work and think much harder. We put much more effort into searching for theoretical insights and concepts, seeking possibly related data, and consulting substantive experts, all of which require much more human learning than machine learning. We will also (or should) be much more careful in making and checking assumptions, generating explanations, assessing uncertainties, providing caveats, communicating results, etc.

Some of us will find this process challenging, not so much because it requires more work, but because the process seems to be experience-driven, serendipity-laden, and guidance-free. That is, it seems messy and ad hoc. We cannot just use a machine learning algorithm or try to fit a likelihood function. Well, we can try, but we know a reputable journal would not accept our findings without asking us to do a lot more work beyond tuning the algorithm or maximizing the likelihood. For those of us who do not fear being labeled as (lazy) Bayesians, we do still have the Bayesian recipe to follow. But we know too well that having the recipe won't satisfy the hunger without the necessary ingredients. Yes, we can always cook up some prior to make a Bayesian omelet, but whether we can serve it with confidence would all depend on how much Bayesian Kool-Aid we have been drinking. Relying on the prior for a desperate rescue merely kicks the can into a wider west. There are no well-accepted and principled ways of constructing a sensible prior when there is not much information, since probability distributions are fundamentally unfit for capturing ignorance.

This is not a place to engage in a debate with my fellow Bayesians, but it is the place to emphasize that the principles and guidelines provided by the qualitative framework are also the ones for this messy and ad hoc process of ours. And those of us who are much more quantitative oriented should pay more respect and attention to qualitative researchers, who deal with such little-or-no-data situations routinely, and hence have developed an insight laden qualitative framework, parallel to the quantitative framework that is more familiar to us.

The decade-long painstaking work done by Christine Borgman and Morgan Wofford, [“From Data Processes to Data Products: Knowledge Infrastructures in Astronomy,”](#) is an extremely informative study both for its findings and for demonstrating qualitative thinking and approaches. As stated in its abstract, “Drawing upon a decade of interviews and ethnography, this article compares how three astronomy groups capture, process, and

archive data, and for whom.” The three groups are the Sloan Digital Sky Survey group, the Black Hole Group, and the Integrative Astronomy Group. A statistician would question immediately (and rightly) how representative this sample is, and how much one can learn from $n=3$? I will not want to take away my fellow statisticians’ satisfaction in answering these questions themselves, and hence invite them (and all readers) to dive into this tour-de-force of qualitative data science. My only preface for this invitation is to ask the readers to put yourself in the authors’ shoes, and contemplate: (1) How would I design and carry out such a study? (2) Would I be able to provide more informative and rigorous findings than those reported in this article? and (3) Would I be able to communicate my findings to quantitative oriented thinkers in more convincing ways?

Causal inference is another major area of intellectual inquiry for which contextual investigations and qualitative thinking are indispensable. The overview article by Francesca Dominici, Falco J. Bargagli-Stoffi, and Fabrizia Mealli, [“From Controlled to Undisciplined Data: Estimating Causal Effects in the Era of Data Science Using a Potential Outcome Framework,”](#) provides a timely overview of causal inference in the era of data science. They rightly emphasize in its abstract that “Although we advocate leveraging the use of big data and the application of machine learning (ML) algorithms for estimating causal effects, they are not a substitute of thoughtful study design.” Indeed, qualitative thinking is critical for study designs as well as for interpreting findings from a causal study, even for well controlled ones.

The potential outcome framework explicitly defines the causal effect at the individual level, that is, the difference between *my* outcomes from being given a treatment and (say) a placebo. But what we can learn from the data, even without any issue of bias (e.g., due to confounding factors), is an estimate of the average causal effects in a population. This population average can still be very biased as an assessment of what the treatment would do to *me*. Translating the knowledge about the average causal effect to the individual causal effects requires attention to a host of qualitative considerations, such as the relevance of the population as proxy to individuals, as well as the questions of how we define/code *individuals*, how such coding affects the outcome, and how to assess the uncertainty in this translation when each individual is unique.

As it happens, a second article on causal inference in this issue, [“Individualized Decision Making under Partial Identification: Three Perspectives, Two Optimality Results, and One Paradox”](#) by Yifan Cui, studies these kinds of problems in making individualized causal assertions and decisions, but without the full information that would allow for the identification of confounding factors at individual levels. For example, unobserved educational preference is likely to be correlated with both education and wage, and hence, it would become a confounding variable for a study on the impact of education on earning. Cui’s article presents both rigorous theoretical and practical cases to show that, in situations where one does not have sufficient contextual understanding, one is likely to fall into Simpson-paradox type of traps. That is, one would be confused and hence make the wrong choice given the apparent contradictions between the evidence based on aggregated data and on disaggregated data, that is, data with different resolution levels.

Whether we conduct studies about specific individual groups or persons, or more broadly case studies, we are made aware of the need for qualitative thinking, from engaging in contemplation of different interpretations to reflecting more deeply on individual particularity, including our own preferences. When we have reliable data with sufficient replications, we tend to rely more on quantitative approaches to help reveal the patterns and relationships in the data. Here, by ‘sufficient replications’ I do not simply mean the data size, but rather the units of data objects that carry sufficiently similar information. For example, this issue of *HDSR* features a [survey study on COVID-19 contact tracing apps](#) conducted by Benjamin Levy and Matthew Stewart, where a total of 152 apps were reviewed. But this does not imply that we have a sample of 152 i.i.d. (independent and identically distributed) observations, to borrow some trite technical jargon from statistics and probability. The 152 apps came from 77 countries, and that fact alone should remind us of their potentially tremendous heterogeneity. Which of these differences are the signals that we seek to set them apart in terms of their compliance to existing ethical guidelines, which is the main goal of the study? And which of them make up the annoying idiosyncratic noise that we should remove because they mask the information we are seeking? How we code data—in the sense of Tanweer et al.’s ‘qualitative coding’—to categorize and quantify the 152 apps would directly impact how much information is there, and indeed, the problem of how best to define the information resolution.

In this sense, the main difference between quantitative and qualitative thinking is mostly a matter of resolution. The quantitative methods operate with low resolutions where replications and accumulations are more in abundance to make them applicable, whereas the qualitative ones can at least help us to make some sense out of high-resolution cases wherein the notion of ‘letting data speak’ simply is wishful thinking. Whereas quantitative methods tend to be too vulnerable for cases without enough replications to learn about patterns and processes, qualitative methods are more generally applicable regardless of the size or resolution of the data. This is because the very question of what to inquire into and how to inquire into a problem can never be handled by a purely quantitative process, however it is defined (which requires some qualitative thinking!).

Interweaving and Systemizing Quantitative and Qualitative Thinking

This is both a call, to echo Tanweer et al., and a reminder to ourselves that qualitative thinking is already in data science in various forms and to various degrees. The aim here is to clarify the roles and benefits of qualitative thinking, and to interweave it systematically with quantitative thinking in our data science related endeavors.

The featured discussion article by Jessica Hullman and Andrew Gelman, [“Designing for Interactive Exploratory Data Analysis Requires Theories of Graphical Inference,”](#) is exactly an effort in this direction. Exploratory data analysis (EDA), by its very nature, employs a host of quantitative methods (e.g., conducting descriptive statistical investigations) and qualitative approaches (e.g., generating study hypothesis by contextualizing patterns suggested by EDA). As alluded to earlier, the qualitative components for EDA are

commonly perceived as being experience-driven and ascertainable only via trial and error, and largely they are in the current practice. Evidence for this observation may be found in the fact that almost every statistics department has multiple courses devoted to regression models or, more broadly, statistical modeling, but we rarely see a course focusing on EDA or even mentioning EDA. Few people have learned how to conduct EDA in a systematic way, let alone to teach it, and there is far less literature on the theory about EDA than that on statistical modeling.

At least for the most visible component of EDA, literally and figuratively, Hullman and Gelman argue that not only it is possible but necessary to have a structured theory to ensure that EDA is coherently connected to the later CDA (confirmatory data analysis) process, and hence to ensure a more reliable overall data analysis process. They suggest adopting the concepts and principles of Bayesian model checking for such a theoretical framework. I invite curious readers to dive into the six penetrating discussions (and the rejoinder) in addition to the article, so that you can sample the diverse opinions on Hullman and Gelman’s proposal and contemplate the benefit and challenges in integrating the relevant qualitative and quantitative processes.

Another fascinating article in this issue, the [“Self-Organizing Floor Plans”](#) by Silvio Carta, illustrates the power of integrating quantitative pattern-learning and qualitative individuality-seeking in auto-generating building plans that respect individual taste preference to various degrees. Few data scientists would be naïve enough to believe that architectural design is a purely quantitative process. Similarly, no respectable designer would claim that design (of any kind) is all about individual aesthetic expression with no patterns to seek or processes to follow. Evidently, quantitative learning and qualitative thinking must be interwoven if we want to ‘auto-generate’ designs with practical values.

I want to give a loud shout out to Carta’s article for its lucid demonstration of disciplinary *reflexivity* upfront (in its introduction), and its emphasis on contextualization throughout (e.g., “The notion of self organization has different meanings depending on its context.”). It warns explicitly about the media hype concerning the auto-generation of design plans through black boxes, entailing the displacement of human designers. I also particularly appreciate the emphasis on the hybrid approach, where “the contributions of machines and humans are deployed at different stages and to tackle different problems in a design process, playing on each specific strength.” Here, the contrast between machines and humans essentially parallels that between quantitative and qualitative thinking, providing a concrete demonstration as to why interweaving them is both a wise and necessary strategy to ensure high quality data science output.

Integrating Quantitative and Qualitative Skills Training

As documented in previous issues of and editorials in *HDSR*, themes and topics in data science always have impact on and implications for data science education. In that regard, I am very delighted that Deborah Nolan and Sara Stoudt contributed their pioneering work [“The Promise of Portfolios: Training Modern Data Scientists”](#) to *HDSR*, and I’d urge everyone who cares about the future of data science (who *doesn’t*?) to dive

into their proposal, and to offer constructive scrutiny and productive suggestions (comments can be made directly on *HDSR*'s site by hopping over any text you want to comment on, once you register for a free account on [PubPub.org](https://pubpub.org)).

Nolan and Stoudt are upfront on how their proposed portfolios approach goes significantly beyond documenting a student's accomplishment in quantitative frameworks ("...technical ones most often associated with data science: coding, statistical modeling, machine learning"), to include essential skills that "sometimes dismissed as 'soft' skills, such as communication, collaboration, and ethics." I fully agree that the term 'soft skills' does not capture the essential nature of these skills, but rather a sense of them being elusive, and hence hard to learn or train. Inspired by Tanweer et al.'s general framing and at the risk of being over-reaching, I think it might be useful to consider these as *qualitative skills*, which are as essential as quantitative skills and have similar organizing principles and pedagogical processes for training and learning (e.g., interweaving theoretical contemplation with practical exercises). The proposal by Nolan and Stoudt has much to be recommended for, because of its particular emphasis on interweaving the two kinds of skills, such as training 'coding as writing.' Again, I will invite readers to brainstorm and suggest innovations to enhance such integrated and interwoven training.

Nolan and Stoudt also made repeated emphases on "strong communication skills," which include the ability to "write precise and accurate technical reports, translate findings to a broader audience in conversation or written form without sacrificing this precision and accuracy; and connect findings to a relevant context for a variety of audiences." I don't think I need to elaborate on any of these points to convince readers why they are important. As an educator, the number one request (or complaint) I have received from builders and leaders in any sector (business, governments, industries, NGO, etc) is 'please train your students with better communication skills.' But I do want to share one recent experience as a member of the ASA (American Statistical Association) President's Task Force on Statistical Significance and Replicability, as it provides a vivid example of the central role of communication in building consensus and getting messages across.

The committee's charge was to clarify ASA's position about the use and usefulness of p -value and significance level. Writing a committee report is never easy, especially on a controversial topic about which many people have strong opinions. Communication becomes essential, both internally (i.e., among the task force members to reach a consensus) and externally (e.g., to present a balanced message that will be appreciated by as broad an audience as possible). For those who have not been involved in such a task, what we went through might sound 'academic' in the pejorative sense of the word or even silly. Most of the meetings and email exchanges (and we have had many) were about wording, such as whether a term should be singular or plural. But these discussions were all substantive, because these choices were about communicating diverse opinions from professional communities. For example, the plural 'communities' here acknowledges that the issue of statistical significance does not concern only the statistical profession. Indeed, even within the statistical profession, there are a variety of communities (e.g., the Bayesian community) with very different takes on significance tests.

Such communication is challenging because too much balancing can lead to statements that appear to be non-informative or even like some political speeches with sound bites but no sound messages. I will leave the readers to be the judges, as the [task force's statement](#) is reprinted in this issue, so its message can reach as many members of the broad data science community as possible, given the prevalence of the use (and misuse) of statistical significance and the importance of the replicability of science ([see the special theme on this topic in HDSR](#)).

Last but not the least, this issue concludes with [a memorial to HDSR's Founding Co-editor of Data Science Education, Robert Lue](#). Since the memorial documented my fond memory of working with Rob and my profound sadness of losing him, I will not add more words to this already lengthy editorial, other than mentioning that Rob would fully support integrating qualitative and quantitative training in data science, because his vision was to reimagine the entire paradigm of liberal arts and science educations via the data science platform.

A Farewell and A Goodbye...

As it happens, completing this 10th editorial for *HDSR* also marks the beginning of a scheduled break for me. I am deeply grateful to the Co-Faculty Directors of the [Harvard Data Science Initiative](#), Francesca Dominici and David Parkes, who not only appointed me to this Editor-In-Chief (EIC) role initially, but also kindly agreed to serve as the interim Co-EICs for the 2021-2022 academic year, to enable me to take my regularly scheduled sabbatical. Indeed, they have already started their work since July 1, including putting more structure into *HDSR*'s editorial and review systems to ensure the long-term sustainability of *HDSR*, as it now enters a growth period. I invite readers and especially potential authors to listen to their welcome remarks and submission invitations.

I am of course also grateful to everyone who has been involved in launching and sustaining *HDSR* over the last three years, from the entire editorial and advisory boards to our publisher, MIT Press, and PubPub, and to all the authors, reviewers, and readers. The list is extremely long, but since this is not a farewell speech, I will single out only *HDSR*'s Editorial and Administrative Coordinator, Paige Sammartino. Paige has been with *HDSR* since its founding, and it suffices to signal how much work she has done by the fact that she provides editorial support to every single submission to *HDSR*, from communicating with the authors to coordinating on publications with the MIT Press. If you enjoy reading *HDSR*, please join me in thanking her for her indispensable contributions that have made all the enjoyable readings possible and wish her the very best as she leaves *HDSR* this summer to earn a teaching degree and become an educator herself (and I am sure she will be teaching both qualitative and quantitative thinking!).

And this brings me to my goodbye to you. As much as we all dislike uncertainties, excitement in life also comes from unlimited possibilities, vaguely formed at least in our imagination. My sabbatical three years ago changed my plan from serving as a second-term dean to launching *HDSR*, and I have yet to find out how this

sabbatical will unfold, including what new ideas it may generate for *HDSR*. Until then, please continue to support *HDSR* in ways you can, from writing and reviewing for *HDSR* to help promoting its content globally. I will be thanking you all remotely, whether I'll be conducting capture and recapture experiments, or participating in sampling and testing studies.

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.