Early-Career View on Data Science Challenges: Responsibility, Rigor, and Accessibility

Shuang Frost¹ Aleksandrina Goeva² William Seaton³ Sara Stoudt^{4,5}
Ana Trisovic⁶

¹Marshall School of Business, Lloyd Greif Center for Entrepreneurial Studies, University of Southern California, Los Angeles, California, United States of America,

²Macosko Lab, Stanley Center for Psychiatric Research, Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts, United States of America,

³Institute for Applied Computational Science, Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, United States of America,

⁴Berkeley Institute for Data Science, University of California Berkeley, Berkeley, California, United States of America,

⁵Statistical & Data Sciences Program, Smith College, Northampton, Massachusetts, United States of America,

⁶Institute for Quantitative Social Science, Harvard University, Cambridge, Massachusetts, United States of America

Published on: Sep 30, 2020

DOI: <https://doi.org/10.1162/99608f92.b211611a>

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

1. Introduction

As emerging voices in the data science community, we see a lot of promise in the future of the field and our roles in it. The articles by Jeannette Wing, Xuming He, and Xihong Lin present compelling challenges that our group of early career members—of graduate students and postgraduate researchers, rising leaders of industry, and recently appointed teaching faculty—believes gives us enough opportunity to last our whole careers. As new members, we are particularly focused on understanding how we should participate in the data science community and what our responsibilities are, to our colleagues and to broader society.

The fact that we are still asking what data science *is*, and debating whether it is even a discipline, can be seen as a mixed blessing. It signals that there might still be the opportunity for growth and plenty of research left to be done. It welcomes a broad range of participants, each with their own views. However, it also makes entry and advancement in the field difficult, with the desired skillset unstated and privileges awarded to those with an insider's perspective.

Without a clear definition provided to us, we have built our own understanding of the role of data science by focusing on what data scientists *do*, rather than trying to define who they *are* or *where* they come from. Data scientists contain multitudes; they juggle a variety of roles and skills throughout the work day. We identify some common questions that motivate our actions as we work with data throughout its life cycle, from collection and analysis to presentation and preservation:

- What data can I collect? What can I quantify?
- What can I learn from this data? Could the data be misleading?
- Which techniques and tools are most useful and accurate for this type of data?
- How can I communicate my approach and findings in a clear way?
- How can I ensure future reproducibility and reuse of my research?

Data scientists are engaged in the daily endeavor to represent the world around us in digital form and to understand what quantifiable metrics are needed to paint a fair and true picture of the people, objects, or natural phenomena under study.

Similar data science methodology is applied across different domains, showcasing the interdisciplinary utility of these methods and leaving us struggling to put data science in a box, or even a Venn Diagram ([Conway, 2010](#)). This commonality across domains allows researchers trained in one data-driven field to leverage and apply their skills into seemingly unconnected careers. The perhaps surprising pipeline of experimental physicists into computer science or quantitative trading careers is a good example.

Beyond a broadly useful skillset, the interdisciplinary nature of data science speaks to the commonalities in data itself: in its underlying forms of strings, integers, or Booleans, the shared ways we store and transfer it,

and how we as humans try to use data. As early practitioners, we are learning about the best ways to collect, store, clean, analyze, and disseminate data while at the same time realizing that a clean and ordered data set does not necessarily represent a data set of inherent value.

2. Useful Data

Wing introduces a notion of precious data, which she defines as data that is expensive to collect, artisanal, or containing a rare event and thus merits specialized techniques. This definition sparked a debate whether precious data necessarily had inherent value versus common data, or if there should be an alternate method of valuation. We discussed various frameworks for evaluating the value of data, including measuring data by quantity of scientific insights, market-driven monetary value, or the generated individual benefit.

In practice, we posit that precious data must be viewed through a lens of usefulness to guarantee we as data scientists are focused on data sets likely to produce something of value. Data that was expensive to collect may still lack documentation, be incomplete or noisy, or the cost of its reuse may be higher than the benefit we could obtain from it. Conversely, cheap or publicly available data can be very useful to a large number of researchers. The cost or rarity of data will not necessarily determine its usefulness for analysis. We acknowledge that trying to use a lens of usefulness still leaves unanswered challenges.

The value of what can be produced from a data set, today and in the future, is very difficult to predict. In academia, data is considered valuable if it presents evidence for our understanding of the nature of the world and, thus, adds to our foundation of knowledge. An example of such data is a recording of a rare event like a gravitational wave (Abbott et al., 2016) or a decay of the Higgs boson (Aad et al., 2012; Chatrchyan et al., 2012), which may also be viewed as historical artifacts. In some cases, these data present a major technological and scientific breakthrough and lead to the rise of new disciplines, but these advancements cannot be guaranteed when designing the experiment.

Data sets can have different valuations depending on the researcher or analyst using the data. In industry, data sets are typically considered useful and valuable if they give insight into a company's users. The value system can be measured monetarily, since, for example, advertisers or service providers are eager to pay a fee to learn something potentially useful about customers. The fact that today's most-valued companies, such as Apple, Alphabet, Microsoft, Amazon, Facebook, and Alibaba (*Most Valuable Companies in the World - 2020*, n.d.), all own a rich inventory of user data, strengthens the assumption of intrinsic value and usefulness of (user) data. We observe that putting a price tag on such data is becoming prevalent, yet provoking discussion and raising ethical concerns around its further use.

3. Data Selves

Nearly all people-focused industries and disciplines today are struggling to define the quantifiable metrics needed to fairly represent a person, derive scientific insights from individual data, and implement applications

accordingly. At the same time, data scientists have to understand that people living in this digital era are rapidly growing their digital presence or data selves, the constellation of data points that represent who they are.

Individuals are no longer passive subjects whose data points are being collected by researchers, corporations, and governments. Although most people do not know exactly what data are being collected, how they are analyzed, and how institutions use insights from their data to get what they want, they are aware of the existence of their data selves. And they realize that their digital metrics are having an increasingly important impact on their everyday life. Some digitally savvy users, whether consciously or unconsciously, are starting to ‘train’ their apps, from personalized recommendation engines, to personalized pricing on digital platforms, to targeted advertising by political campaigns. For instance, one might carefully consider what videos to like, which ads to click on, or what hashtags to engage with in order to curate one’s self-representation that they want the machine to ‘see.’ In such ways, users perform to, try to control, even resist, the increasingly prevalent ‘dataveillance’ (Zuboff, 2019). Understanding people’s evolving relationship with their data selves, therefore, must also be part of data science.

As the first generation of data scientists who grew up interacting with our own data selves, we are familiar with being both subjects and researchers. We have conducted impromptu ‘controlled experiments,’ where a group of friends all hail a ride to and from matching locations, to understand our personalized pricing. We have learned to check flight prices while logged in and logged out of loyalty accounts, using different emails or even IP addresses, so as not to seem overeager to the price-setting algorithms. We are on early-career budgets after all. Many of us fell into data science when we dared to peek behind the algorithmic curtain. These experiences taught us that users have the curiosity and instincts to think like data scientists. With accessible general education, many citizens can develop data literacy that could help them keep track of all of these algorithms and be aware of their power over our lives.

4. Building Public Trust

As new data scientists, much of our careers so far have been filled with important revelations about bias within powerful models like those used for facial recognition, prison sentencing guidelines, or loan approvals. We have learned the wonders of things like targeted medicine, statistical space exploration, and big data predictions, while also learning about the risks in blindly trusting the machine. We have joined the data science community during a significant political era, when skepticism of experts and official reports is growing and too often justified. Continued media exposure reveals that the massive complexities of data science can be wielded as weapons for exploitation and manipulation. This all has been challenging for students of data scientists who were initially attracted to the promise of the field because of our optimistic belief that data could provide access to truth. The reality is that increasing complexity and opaqueness of data collection and processing make it difficult to make confident statements about truth. The fundamental question has sometimes shifted to no longer be ‘how can we gain the public’s trust regarding our work’ but ‘is our work even trustworthy?’

To answer ‘yes’ we need to collect data and apply analysis techniques with comprehensive social considerations in mind. Fundamental choices such as what data to collect and how to store that data may have major implications for data privacy and security. Data scientists must be responsible for the data they collect and use; they must consider the faces behind the data and prioritize the best interests of those represented in the data, sometimes over their own analytical interests.

We believe that what He and Lin call “fair and interpretable learning and decision making” and what Wing called “trustworthy AI” represent critical areas for data science research. If people think that our models are making generally fair and accurate decisions, then they may be more willing to share their data, to integrate algorithms into their daily life, and, thus, improve our collective decision making. We are glad to see both He and Lin as well as Wing in agreement, adding “Privacy” and “Statistical analysis of privatized data” to their lists. An increased emphasis on data privacy will hopefully increase the guarantee that for data participants, there will be no unexpected costs of participating in data research. While in the short-term, it is true that a data scientist could, to use Wing’s example, build a better predictive model of hospitals’ shared data, unfettered sharing of sensitive health records may serve to disincentivize patient participation over the medium-to-long term and increase restrictions, such that we will have less data over which to apply improved methods.

In this vein, we see collaborations between data scientists and policymakers as particularly necessary for taking on pressing societal challenges. For example, legislation affecting data collection made by the European Union’s General Data Protection Regulation or California’s Consumer Privacy Act are major steps forward in democratizing understanding about what data we give up and where it goes on a day-to-day basis and building trust between data collectors and data providers.

We make decisions at every step in the research process and to build trust it is especially important that these choices be transparently communicated and defended not only to our collaborators and stakeholders but also to the general public. The recent controversies of data manipulation have spurred public interest in understanding how individuals are represented and viewed through data, models, and algorithms (United States. Congress. Senate. Committee on Homeland Security and Governmental Affairs, 2019; Select Committee on Intelligence & United States Senate, 2019). By harnessing the rising data awareness and interest, we can further equip the public with knowledge and tools to make their own judgment whether certain data practices are trustworthy or not.

We have the opportunity to turn away from an expert-oriented view of trust-building, with which we simply tell the public that the experts have done a great job in protecting privacy and promoting welfare. Instead we can create transparent channels in which the public can express their opinions about how their data are handled and to build accountable institutions that protect the public’s data rights.

5. Transparency and Validation

Transparency is not only crucial for building trust between experts and the public, it is also essential for building trust between experts. Detailed documentation of our hypotheses, modeling choices, and, wherever possible, the explicit encoding of domain knowledge can streamline the process of evaluating one another's work.

In the era of big data, it has become tempting to proceed without invoking preestablished domain knowledge to form a hypothesis and just let the data speak. This hypothesis-free approach, well-established in the field of statistics by Tukey (1977), consists of the following basic steps—a lot of data is collected about a system, trying to capture as many dimensions as is feasible, modeling and inference targeted at finding structure and relationships are performed, and the results are scrutinized by data scientists and domain experts to generate insights. Any findings resulting from such an approach are a product of the data, the inference methods, and the way the results were interpreted. Thus, such findings carry uncertainty and may be biased. In line with He and Lin's call for further development of “Study design and statistical methods for reproducibility and replicability,” we add the need for further development of experimental design for the collection of ‘orthogonal’ data, or follow-up data to validate results originally found using a hypothesis-free approach.

Making fair comparisons across problem-solving approaches is crucial in both hypothesis-free and hypothesis-driven approaches; however, it is often far from trivial. Heterogeneity in data access and methodological preferences makes it challenging to synthesize competing methods in a standardized way, especially when ground truth is lacking. For example, complex models such as deep-learning architectures are enjoying widespread adoption, but how do we know they are necessary for every task (Rudin & Radin, 2019)? The use of simple baselines in comparison studies can both tell us if a complex model is really necessary, or if a simple model is just as good and should be preferred, and highlight when and why a more complex model is beneficial. However, the interpretation of a simple baseline model may not be straightforward. As Freedman & Freedman (1983) famously pointed out, one could get a statistically significant linear prediction model by using (a large number of) random features, while associations between the response and the randomly drawn features are scientifically useless and entirely spurious.

Data scientists frequently engage and play a large role in multidisciplinary problems, working alongside domain experts. In the collaborative effort of developing an analysis, the domain expert's common sense and domain knowledge are routinely used in every step of the process. Decision making throughout the model-building process may occur informally, making it hard to document and track the logic of and, at the later stage, replicate the results. Another challenge of the multidisciplinary collaborative approach is that different experts working on the same problem could reach different conclusions if they are enforcing different beliefs about the data (Lindquist, 2020). These informal decisions obscure the path to the next step in the modeling process, and ultimately the final product. To keep the research process clear, it would be helpful to have

someone in the working group take on the role of documenting changes and decisions made throughout the analysis.

6. Myriad of Roles and Broad Access

The need for transparency brings us back to what a data scientist *does*. Ambiguity in the definition of data science is often manifested through unrealistically broad expectations of data scientists. It appears that a data scientist needs to have advanced knowledge of statistics, computer science, machine learning, data analytics, programming, and the cloud to succeed in what they do (Meng, 2019). Because the amount of things to learn is constantly overwhelming, many young researchers and data scientists across the board feel like imposters. This feeling can breed insecurity and make it hard to propose and develop the new ideas that our field needs.

We argue that creating new roles that view data science as a team activity would alleviate the pressure of data scientists to ‘do it all.’ New roles, like machine learning engineer or cloud data engineer, would allow those in the hiring position to be more specific and realistic about the skills they need for a particular job or to round out the skills of a data science team. For example, instead of searching for the elusive data science ‘unicorn,’ one could emphasize specific skills needed like ‘person with statistical background and comfort with AWS infrastructure.’ By being specific, more applicants with unique perspectives and applicable skills from a variety of domains may feel more prepared to apply to and to enter our field.

Academic data science can benefit from well-defined and realistic data science roles as well. As early career folks we still do not know what it takes to get tenure in data science. With traditional research papers being the currency of academia, it can be hard for those working on something like computing infrastructure or data management to be recognized for their work (Geiger et al., 2018). However, we are reliant on the increasing variety of dedicated data infrastructure at hand, from the creation of GPUs for improved deep-learning computation to the development of numerous database types to support time series, document, and unstructured data. The statistical challenges that come along with this increased variety in data and data structures are discussed by Wing (as “Multiple, heterogeneous data sources”) and He and Lin (under “Integrative analysis of different types and sources of data”) as are the computing infrastructure and data management (under “Computing systems for data-intensive applications” and “Cloud-based scalable and distributed statistical inference” respectively). Therefore, it is important for the academic data science system to provide credit toward career advancement for those doing the necessary work toward overcoming these challenges.

Because data science is not only interdisciplinary but also tied to our everyday lives, it is crucial to be able to promote data literacy and talk about our work in an approachable way. We have talked about ways to build public trust through transparency, but we also call for more opportunities to practice informal, yet clear, communication with broader audiences and increased support and recognition for outreach in our roles as data scientists.

In addition to improvements in data literacy and our communication with the public, we also note two systemic barriers to data science access. While Internet connectivity is increasing throughout the world, the portions without reliable access are still shut out. Technology and connectivity inequities are increasingly revealed as we face remote-learning challenges this coming school year. Many educational resources and publications exist in English only, limiting the ability of non-English speakers to join the discipline. Some of our own early career board members are unable to fully communicate their research in their native language due to the lack of translation of specific terms (*Physics in a Second Language*, n.d.). We hope that the data science community will accelerate the development of multilingual pedagogical materials.

7. Path Forward

Both articles conclude that to move the field forward we need to recruit emerging talent. Data might be universal, but the community infrastructure we operate within still results in inequality of opportunities. There are many identities who are underrepresented in data science jobs and leadership roles; we won't try to list them all. Instead, we emphasize that our group values people from all backgrounds and viewpoints and prioritizes inclusive spaces when seeking opportunities for ourselves and our peers. We think it is important to be intentionally and proactively inclusive to everyone, both personally and professionally, in order to improve our sense of community and strengthen our data science team. We urge everyone to take the steps needed to inform themselves, advocate for change, and make data science a place for everyone in order to best face the big challenges of our field.

Disclosure Statement

Shuang Frost, Aleksandrina Goeva, William Seaton, Sara Stoudt, and Ana Trisovic have no financial or non-financial disclosures to share for this article.

References

- Aad, G., Abajyan, T., Abbott, B., Abdallah, J., Abdel Khalek, S., Abdelalim, A. A., Abdinov, O., Aben, R., Abi, B., Abolins, M., AbouZeid, O. S., Abramowicz, H., Abreu, H., Acharya, B. S., Adamczyk, L., Adams, D. L., Addy, T. N., Adelman, J., Adomeit, S., Adragna, P., ... Zwalinski, L. (2012). Observation of a new particle in the search for the standard model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B*, 716(1), 1–29. <https://doi.org/10.1016/j.physletb.2012.08.020>
- Abbott, B. P., Abbott, R., Abbott, T. D., Abernathy, M. R., Acernese, F., Ackley, K., Adams, C., Adams, T., Addesso, P., Adhikari, R. X., Adya, V. B., Affeldt, C., Agathos, M., Agatsuma, K., Aggarwal, N., Aguiar, O. D., Aiello, L., Ain, A., Ajith, P., Allen, B., ... LIGO Scientific collaboration and virgo collaboration. (2016). Observation of gravitational waves from a Binary Black Hole Merger. *Physical Review Letters*, 116(6), Article 061102. <https://doi.org/10.1103/PhysRevLett.116.061102>

Chatrchyan, S., Khachatryan, V., Sirunyan, A. M., Tumasyan, A., Adam, W., Aguilo, E., Bergauer, T., Dragicevic, M., Erö, J., Fabjan, C., Friedl, M., Frühwirth, R., Ghete, V. M., Hammer, J., Hoch, M., Hörmann, N., Hrubec, J., Jeitler, M., Kiesenhofer, W., Knünz, V. ... Wenman, D. (2012). Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B*, 716(1), 30–61.

<https://doi.org/10.1016/j.physletb.2012.08.021>

Conway, D. (2010, September 30). *The data science Venn diagram*—Drew Conway.

<http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>

Freedman, D. A., & Freedman, D. A. (1983). A note on screening regression equations. *The American Statistician*, 37(2), 152–155. <https://doi.org/10.1080/00031305.1983.10482729>

Geiger, R. S., Cabasse, C., Cullens, C. Y., Norén, L., Fiore-Gartland, B., Das, D., & Brady, H. (2018). Career paths and prospects in academic data science: Report of the Moore-Sloan data science environments survey. *SocArXiv*. <https://doi.org/10.31235/osf.io/xe823>

Lindquist, M. (2020). Neuroimaging results altered by varying analysis pipelines. *Nature*, 582(7810), 36–37.

<https://doi.org/10.1038/d41586-020-01282-z>

Meng, X.-L. (2019). Data science: An artificial ecosystem. *Harvard Data Science Review*, 1(1).

<https://doi.org/10.1162/99608f92.ba20f892>

Most valuable companies in the world - 2020. (n.d.). Retrieved August 23, 2020, from <https://fxssi.com/top-10-most-valuable-companies-in-the-world>

Physics in a second language. (n.d.). Retrieved August 23, 2020, from

<https://www.symmetrymagazine.org/article/physics-in-a-second-language>

Rudin, C., & Radin, J. (2019). Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harvard Data Science Review*, 1(2).

<https://doi.org/10.1162/99608f92.5a8a3a3d>

Select Committee on Intelligence, & United States Senate. (2019). *Report of the Select Committee on Intelligence United States Senate on Russian active measures campaigns and interference in the 2016 U. S. election: Volume 1: Russian efforts against election infrastructure with additional views*. Independently Published. https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume1.pdf

Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2). Addison-Wesley. http://theta.edu.pl/wp-content/uploads/2012/10/exploratorydataanalysis_tukey.pdf

United States. Congress. Senate. Committee on Homeland Security and Governmental Affairs. (2019). *Deepfake Report Act of 2019: Report of the committee on Homeland Security and Governmental Affairs, United States Senate, to accompany S. 2065 to require the Secretary of Homeland Security to publish an annual report on the use of deepfake technology, and for other purposes*. U.S. Government Publishing Office. <https://www.congress.gov/congressional-report/116th-congress/senate-report/93/1>

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.

©2020 Shuang Frost, Aleksandrina Goeva, William Seaton, Sara Stoudt, and Ana Trisovic. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](https://creativecommons.org/licenses/by/4.0/), except where otherwise indicated with respect to particular material included in the article.