Harvard Data Science Review • Issue 2.3, Summer 2020

Ten Research Challenge Areas in Data Science

Jeannette M. Wing

Published on: Sep 30, 2020

DOI: 10.1162/99608f92.c6577b1f

ABSTRACT

To drive progress in the field of data science, we propose 10 challenge areas for the research community to pursue. Since data science is broad, with methods drawing from computer science, statistics, and other disciplines, and with applications appearing in all sectors, these challenge areas speak to the breadth of issues spanning science, technology, and society. We preface our enumeration with meta-questions about whether data science is a discipline. We then describe each of the 10 challenge areas. The goal of this article is to start a discussion on what could constitute a basis for a research agenda in data science, while recognizing that the field of data science is still evolving.

Keywords: artificial intelligence, causal reasoning, computing systems, data life cycle, deep learning, ethics, machine learning, privacy, trustworthiness

Although data science builds on knowledge from computer science, engineering, mathematics, statistics, and other disciplines, data science is a unique field with many mysteries to unlock: fundamental scientific questions and pressing problems of societal importance.

In this article we enumerate 10 areas of research in which to make progress to advance the field of data science. Our goal is to start a discussion on what could constitute a basis for a research agenda in data science, while recognizing that the field of data science is still evolving.

Before we plunge into this enumeration, we preface our discussion by raising, but not answering, a meta-question: Is data science a discipline? Answering this meta-question is still under lively debate, including within the pages of this journal. Herein, we suggest additional meta-questions to help frame the debate.

Is Data Science a Discipline?

Data science is a field of study: one can get a degree in data science, get a job as a data scientist, and get funded to do data science research. But is data science a discipline, that is, a branch of knowledge? If not yet, will it evolve to be one, distinct from other disciplines? Here are a few meta-questions on whether data science is a discipline.

- *Are there driving deep question(s) in data science? If so, what are they?* Each scientific discipline (usually) has one or more 'deep' questions that drive its research agenda: What is the origin of the universe (astrophysics)? What is the origin of life (biology)? What is computable (computer science)? Does data science inherit its deep questions from all its constituent disciplines or does it have its own unique ones?

- *What is the role of the domain in the field of data science?* Many academics have argued (Wing et al., 2018) that data science is unique in that it is not just about methods, but also about the use of those methods in the context of a domain—the domain of the data being collected and analyzed; the domain in which, from this data, a question is to be answered. Is the inclusion of a domain inherent in defining the field of data science? Other methods-based disciplines, such as computer science, mathematics, and statistics, are used in the context of other domains, and are correspondingly inspired by problems from these domains. Can one study data science, as we do in computer science, mathematics, and statistics, without studying it in the context of a domain? Is the (more integral?) way a domain is included in the study of data science unique to data science?
- *What makes data science data science?* Is there a problem unique to data science that one can convincingly argue would not be addressed or asked by any of its constituent disciplines, for example, computer science or statistics? When should a set of methods, analyses, or results be considered data science, and not just methods, analyses, or results in computer science or statistics (or mathematics, etc.)? Or should all methods, analyses, and results in all these disciplines be considered part of data science?

Data science as a field of study is still too new to have definitive answers to all these meta-questions. Their answers will likely evolve over time, as the field matures and as members of the contributing established disciplines share scholarship and perspectives from their respective disciplines. We encourage the data science community to ponder and debate these meta-questions, as we make progress on more concrete scientific and societal challenges raised by the preponderance of data, data science methods, and applications of data science.

Ten Research Areas

So, let’s ask an easier question, one that also underlies any field of study: What are the research challenge areas that drive the study of data science? Here is a list of 10. They are not in any priority order, and some of them are related to each other. They are phrased as challenge areas, not challenge questions; each area suggests many questions. They are not necessarily the ‘top 10’ but they are a good 10 to start the community discussing what a broad research agenda for data science might look like. Given our discussion above, they unsurprisingly overlap with challenges found in computer science, statistics (Berger et al., 2019), social sciences, and so on. Given the author’s background, they are posed from the perspective of a computer scientist. The list begins, roughly speaking, with challenges relevant to science, then to technology, and then to society.

1. Scientific Understanding of Learning, Especially Deep Learning Algorithms.

As much as we admire the astonishing successes of deep learning, we still lack a scientific understanding of why deep learning works so well, though we are making headway (Arora et al., 2018; Balestriero & Baraniuk, 2018). We do not understand the mathematical properties of deep learning algorithms or of the models they produce. We do not know how to explain why a deep learning model produces one result and not another. We do not understand how robust or fragile models are to perturbations to input data distributions. We do not understand how to verify that deep learning will perform the intended task well on new input data. We do not know how to characterize or measure the uncertainty of a model's results. We do not know deep learning's fundamental computational limits (Thompson et al., 2020); at what point does more data and more compute not help? Deep learning is an example of where experimentation in a field is far ahead of any kind of complete theoretical understanding. And, it is not the only example in learning: random forests (Biau & Scornet, 2015) and high-dimensional sparse statistics (Johnstone & Titterington, 2009) enjoy widespread applicability on large-scale data, where gaps remain between their performance in practice and what theory can explain.

2. Causal Reasoning

Machine learning is a powerful tool to find patterns and to examine associations and correlations, particularly in large data sets. While the adoption of machine learning has opened many fruitful areas of research in economics, social science, public health, and medicine, these fields require methods that move beyond correlational analyses and can tackle causal questions. A rich and growing area of current study is revisiting causal inference in the presence of large amounts of data. Economists are devising new methods that incorporate the wealth of data now available into their mainstay causal reasoning techniques, for example, the use of instrumental variables; these new methods make causal inference estimation more efficient and flexible (Athey, 2016; Taddy, 2019). Data scientists are beginning to explore multiple causal inference, not just to overcome some of the strong assumptions of univariate causal inference, but because most real-world observations are due to multiple factors that interact with each other (Wang & Blei, 2019). Inspired by natural experiments used in economics and the social sciences, as more government agency and commercial data becomes publicly available, data scientists are using synthetic control for novel applications in public health, retail, and sports (Abadie et al., 2010; Amjad et al. 2019).

3. Precious Data

Data can be precious for one of three reasons: the data set is expensive to collect; the data set contains a rare event (low signal-to-noise ratio); or the data set is artisanal—small, task-specific, and/or targets a limited audience. A good example of expensive data comes from large, one-off, expensive scientific instruments, for example, the Large Synoptic Survey Telescope, the Large Hadron Collider, and the IceCube Neutrino Detector at the South Pole. A good example of rare event data is data from sensors on physical infrastructure, such as bridges and tunnels; sensors produce a lot of raw data, but the disastrous event they are used to predict is (thankfully) rare. Rare data can also be expensive to collect. A good example of artisanal data is the tens of millions of court judgments that China has released online to the public since 2014 (Liebman et al., 2017) or the two-plus-million U.S. government declassified documents collected by Columbia's History Lab (Connelly et al., 2019). For each of these different kinds of precious data, we need new data science methods and algorithms, taking into consideration the domain and the intended uses and users of the data.

4. Multiple, Heterogeneous Data Sources

For some problems, we can collect lots of data from different data sources to improve our models and to increase knowledge. For example, to predict the effectiveness of a specific cancer treatment for a human, we might build a model based on 2-D cell lines from mice, more expensive 3-D cell lines from mice, and the costly DNA sequence of the cancer cells extracted from the human. As another example, multiscale, spatiotemporal climate models simulate the interactions among multiple physical systems, each represented by disparate data sources drawn from sensing: the ocean, the atmosphere, the land, the biosphere, and humans. Many of these data sources might be precious data (see Challenge no. 3). State-of-the-art data science methods cannot as yet handle combining multiple, heterogeneous sources of data to build a single, accurate model. Bounding the uncertainty of a data model is exacerbated when built from multiple, possibly unrelated data sources. More pragmatically, standardization of data types and data formats could reduce undesired or unnecessary heterogeneity. Focused research in combining multiple sources of data will provide extraordinary impact.

5. Inferring From Noisy and/or Incomplete Data.

The real world is messy and we often do not have complete information about every data point. Yet, data scientists want to build models from such data to do prediction and inference. This long-standing problem in statistics comes to the fore as: (1) the volume of data, especially about people, that we can generate and collect grows unboundedly; (2) the means of generating and collecting data is not under our control, for example, data from mobile phone and web apps vary—by design—across different users and across different populations; and 3) many sectors, from finance to retail to transportation, embrace the desire to do real-time personalization. A great example of a novel formulation of this problem is the planned use of differential privacy for Census 2020 data (Abowd, 2018; Hawes, 2020), where noise is deliberately added to a query result, to maintain the privacy of individuals participating in the census. Handling 'deliberate' noise is particularly important for researchers working with small geographic areas such as census blocks, since the added noise can make the data uninformative at those levels of aggregation. How then can social scientists, who for decades have been drawing inferences from census data, make inferences on this ‘noisy’ data and how do they combine their past inferences with these new ones? Machine learning’s ability to better separate noise from signal can improve the efficiency and accuracy of those inferences.

6. Trustworthy AI

We have seen rapid deployment of systems using artificial intelligence and machine learning in critical domains such as autonomous vehicles, criminal justice, health care, hiring, housing, human resource management, law enforcement, and public safety, where decisions taken by AI agents directly impact human lives. Consequently, there is an increasing concern if these decisions can be trusted to be correct, fair, ethical (see Challenge no. 10), interpretable, private (see Challenge no. 9), reliable, robust, safe, and secure, especially under adversarial attacks. Many of these properties borrow from a long history of research on Trustworthy Computing (National Research Council, 1999), but AI raises the ante (Wing, 2020): reasoning about a machine learning model seems to be inseparable from reasoning about the available data used to build it and the unseen data on which it is to be used; and these models are inherently probabilistic. One approach to building trust is through providing explanations of the outcomes of a machine learned model (Adadi & Berrada, 2018; Chen et al., 2018; Murdoch et al., 2019; Turek, 2016). If we can interpret the outcome in a meaningful way, then the end user can better trust the model. Another approach is through formal methods, where one strives to prove once and for all a model satisfies a certain property. New trust properties yield new tradeoffs for machine learned models, for example, privacy versus accuracy; robustness versus efficiency; fairness versus robustness. There are multiple technical audiences for trustworthy models: model developers, model users (human and machine), and model customers; as well as more general audiences: consumers, policymakers, regulators, the media, and the public.

7. Computing Systems for Data-Intensive Applications

Traditional designs of computing systems have focused on computational speed and power: the more cycles, the faster the application can run. Today, the primary focus of applications, especially in the sciences (e.g., astronomy, biology, climate science, materials science), is data. Novel special-purpose processors, for example, GPUs, FPGAs, TPUs, are now commonly found in large data centers. Domain-specific accelerators, including those designed for deep learning, show orders of magnitude performance gains over general-purpose computers (Dally et al., 2020). Even with all these data and all this fast and flexible computational power, it can still take weeks to build accurate predictive models; however, applications, whether from science or industry, want *real-time* predictions. Distributing data, computing, and models helps with scale and reliability (and privacy), but then runs up against the fundamental limit of the speed of light and practical limits of network bandwidth and latency. Also, data-hungry and compute-hungry algorithms, for example, deep learning, are energy hogs (Strubell et al., 2019). Not only should we consider space and time, but energy consumption, in our performance metrics. In short, we need to rethink computer systems design from first principles, with data (not compute) the focus. New computing systems designs need to consider: heterogeneous processing, efficient layout of massive amounts of data for fast access, communication and network capability, energy efficiency, and the target domain, application, or even task.

8. Automating Front-End Stages of the Data Life Cycle

While the excitement in data science is due largely to the successes of machine learning, and more specifically deep learning, before we get to use machine learning algorithms, we need to prepare the data for analysis. The early stages in the data life cycle (Wing, 2019) are still labor intensive and tedious. Data scientists, drawing on both computational and statistical tools, need to devise automated methods that address data collection, data cleaning, and data wrangling, without losing other desired properties, for example, accuracy, precision, and robustness, of the end model. One example of emerging work in this area is the Data Analysis Baseline Library (Mueller, 2019), which provides a framework to simplify and automate data cleaning, visualization, model building, and model interpretation. The Snorkel project addresses the tedious task of data labeling (Ratner et al., 2018). Trifacta, a university spin-out company, addresses data wrangling (Trifacta, 2020). Complementing these needs, commercial services already support later stages in the data life cycle, in particular, automating construction of machine learning models, for example, Cloud AutoML (Google, 2020) and Azure Machine Learning (Microsoft, 2020).

9. Privacy

For many applications, the more data we have, the better the model we can build. One way to get more data is to share data, for example, multiple parties pool their individual data sets to build collectively a better model than any one party can build. However, in many cases, due to regulation or privacy concerns, we need to preserve the confidentiality of each party's data set. An example of this scenario is in building a model to predict whether someone has a disease or not. If multiple hospitals could share their patient records, we could build a better predictive model; but due to Health Insurance Portability and Accountability Act (HIPAA, 1996) privacy regulations, hospitals cannot share these records. We are only now exploring practical and scalable ways, using cryptographic and statistical methods, for multiple parties to share data, models, and/or model outcomes while preserving the privacy of each party's data set. Industry and government are already exploiting techniques and concepts, for example, secure multiparty computation, homomorphic encryption, zero-knowledge proofs, differential privacy, and secure enclaves, as elements of point solutions to point problems (Abowd, 2018; Ion et al., 2017; Kamara, 2014). We can also apply these methods to the simpler scenario where a single entity's data must be kept private prior to analysis.

10. Ethics

Data science raises new ethical issues. They can be framed along three axes: (1) the ethics of data: how data are generated, recorded, and shared; (2) the ethics of algorithms: how artificial intelligence, machine learning, and robots interpret data; and (3) the ethics of practices: devising responsible innovation and professional codes to guide this emerging science (Floridi & Taddeo, 2016) and to define institutional review board (IRB) criteria and processes specific for data (Wing et al., 2018). The ethical principles expressed in the Belmont Report (Belmont Report, 1979) and the Menlo Report (Dittrich & Kenneally, 2011) give us a starting point for identifying new ethical issues data science technology raises. The ethical principle of Respect for Persons suggests that people should always be informed when they are talking with a chatbot. The ethical principle of Beneficence requires a risk/benefit analysis on the decision a self-driving car makes on whom not to harm. The ethical principle of Justice requires us to ensure the fairness of risk assessment tools in the court system and automated decision systems used in hiring. These new ethical issues correspondingly raise new scientific challenges for the data science community, for example, how to detect and eliminate racial, gender, socioeconomic, or other biases in machine learning models.

Closing Remarks

As many universities and colleges are creating new data science schools, institutes, centers, and so on (Wing et al., 2018), it is worth reflecting on data science as a field. Will data science as an area of research and education evolve into being its own discipline or be a field that cuts across all other disciplines? One could argue that computer science, mathematics, and statistics share this commonality: they are each their own discipline, but they each can be applied to (almost) every other discipline.

What will data science be in 10 or 50 years? The answer to this question is in the hands of the next-generation researchers and educators. To advance and study data science will take a commitment to learn the vocabulary, methods, and tools from multiple, traditionally siloed disciplines. Integrating and applying this knowledge takes patience, but can be exhilarating. To today’s undergraduates, graduate students, postdoctoral fellows, and early-career faculty and researchers: Through the data science research problems you choose to tackle, you will shape this field!

Disclosure Statement

The author has nothing to disclose.

Acknowledgments

I would like to thank Cliff Stein, Gerad Torats-Espinosa, Max Topaz, and Richard Witten for their feedback on earlier renditions of this paper. Many thanks to all Columbia Data Science faculty who have helped me formulate and discuss these 10 (and other) challenges during our Fall 2019 DSI retreat. Final thanks to the anonymous reviewers whose comments helped me sharpen and enhance many of my points.

References

*Abadie, A., Diamond, A., & Hainmüller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's Tobacco Control Program. Journal of the American Statistical Association, **105**(490), 493-505.* https://doi.org/*10.1198/jasa.2009.ap08746*

Abowd, J. M. (2018). The U.S. Census Bureau adopts differential privacy. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2867. Association for Computing Machinery. https://doi.org/10.1145/3219819.3226070

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138-52160. https://doi.org/10.1109/ACCESS.2018.2870052

Amjad, M., Misra, V., Shah, D., & Shen, D. (2019). mRSC: Multi-dimensional Robust Synthetic Control. *Proceedings of the ACM on Measurement and Analysis of Computing Systems (Sigmetrics 2019)*, 3(2), 37:1-28 Association for Computing Machinery. http://dna-pubs.cs.columbia.edu/citation/paperfile/233/mRSC.pdf

Arora, S. Ge, R., Neyshabur, B., & Zhang, Y. (2018). Stronger generalization bounds for deep nets via a compression approach. *Proceedings of the 35th International Conference on Machine Learning*. *PMLR, 80*, 254–263. http://proceedings.mlr.press/v80/arora18b.html

Athey, S. (2016). *Susan Athey on how economists can use machine learning to improve policy*. Stanford Institute for Economic Policy Research. https://siepr.stanford.edu/news/susan-athey-how-economists-can-use-machine-learning-improve-policy

Balestriero, R., & Baraniuk, R. G. (2018). A spline theory of deep networks. *Proceedings of the 35th International Conference on Machine Learning*. *PMLR, 80,* 374–383. http://proceedings.mlr.press/v80/balestriero18b.html

Belmont Report. (1979). *The Belmont Report: Ethical principles and guidelines for the protection of human subjects of Research.* U.S. Department of Health, Education, and Welfare.

Berger, J., He, X., Madigan, C., Murphy, S., Yu, B., & Wellner, J. (2019). *Statistics at a crossroad: Who is for the challenge?* NSF workshop report. National Science Foundation. https://hub.ki/groups/statscrossroad

Biau, G., & Scornet, E. (2015). A random forest guided tour. *TEST*, ***25***, 197–227. https://doi.org/10.1007/s11749-016-0481-7

Chen, C., Lin, K., Rudin, C., Shaposhnik, Y., Wang, S., & Wang, T. (2018). An interpretable model with globally consistent explanations for credit risk. *NIPS 2018 Workshop on Challenges and Opportunities for AI in Financial Services: The Impact of Fairness, Explainability, Accuracy, and Privacy*. https://arxiv.org/abs/1811.12615

Connelly, M., Madigan, D., Jervis, R., Spirling, A., & Hicks, R. (2019). The History Lab. http://history-lab.org/

Dally, W. J., Turakhia, Y., & Han, S. (2020). Domain-specific accelerators. *Communications of the ACM*, *63*(7), 48–57. https://cacm.acm.org/magazines/2020/7/245701-domain-specific-hardware-accelerators/fulltext

Dittrich, D., & Kenneally, E. (2011). The Menlo Report: Ethical principles guiding information and communication technology research. U.S. Department of Homeland Security. http://www.caida.org/publications/papers/2012/menlo_report_ethical_principles/

Floridi, L., & Taddeo, M. (2016). What is data ethics? *Philosophical Transactions of the Royal Society A*, *374*(2083), Article 20160360. https://doi.org/10.1098/rsta.2016.0360

Google. (2020). Cloud AutoML. https://cloud.google.com/automl/

Hawes, M. B. (2020). Implementing differential privacy: seven lessons from the 2020 United States Census. *Harvard Data Science Review, 2*(2). https://doi.org/10.1162/99608f92.353c6f99

HIPAA (1996), Health Insurance Portability and Accountability Act, US Congress, Pub.L. 104–191, 110 Stat. 1936, enacted August 21, 1996.

Ion, M., Kreuter, B., Nergiz, E., Patel, S., Saxena, S., Seth, K., Shananhan, D., & Yung, M. (2017). Private intersection-sum protocol with applications to attributing aggregate ad conversions. *Cryptology ePrint Archive,* Report 2017/738. https://eprint.iacr.org/2017/738

Johnstone, I. M., & Titterington, D. M. (2009). Statistical challenges of high-dimensional data. *Philosophical transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *367*(1906), 4237–4253. https://doi.org/10.1098/rsta.2009.0159

Kamara, S., Mohassel, P., Raykova, M., and Sadeghian, S. (2014). Scaling private set intersection to billion element sets. In N. Christin & R. Safavi-Naini (Eds.), *Financial cryptography and data security (pp. 195–215)*. Springer. https://doi.org/10.1007/978-3-662-45472-5_13

Liebman, B. L., Roberts, M., Stern, R. E., & Wang, A. (2017). Mass digitization of Chinese court decisions: How to use text as data in the field of Chinese law. *UC* San Diego School of Global Policy and Strategy, 21st Century China Center Research Paper No. 2017-01; Columbia Public Law Research Paper No. 14-551.
https://scholarship.law.columbia.edu/faculty_scholarship/2039

Microsoft. (2020). What is automated machine learning (AutoML)? https://docs.microsoft.com/en-us/azure/machine-learning/concept-automated-ml

Mueller, A. (2019). Data Analysis Baseline Library. GitHub. https://libraries.io/github/amueller/dabl

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(44), 22071–22080. https://doi.org/10.1073/pnas.1900654116

National Research Council. (1999). *Trust in cyberspace*. National Academies Press.
https://doi.org/10.17226/6161

Ratner, A., Bach, S., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. (2018). Snorkel: Rapid training data creation with weak supervision. *Proceedings of the 44th International Conference on Very Large Data Bases, 11(3), pp. 269-282.* http://www.vldb.org/pvldb/vol11/p269-ratner.pdf

Strubell E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 3645-3650.* https://www.aclweb.org/anthology/P19-1355.pdf

Taddy, M. (2019). *Business data science: Combining machine learning and economics to optimize, automate, and accelerate business decisions*. McGraw Hill.

Thompson, N. C., Greenewald, K., Lee, K., & Manso, G. F. (2020). *The computational limits of deep learning*. https://arxiv.org/abs/2007.05558

Trifacta. (2020). https://www.trifacta.com/

Turek, M. (2016). Defense Advanced Research Projects Agency, Explainable AI Program. https://www.darpa.mil/program/explainable-artificial-intelligence

Wang, Y., & Blei, D. M. (2019). The blessings of multiple causes. *Journal of the American Statistical Association*, *114*(528), 1574-1596, https://doi.org/10.1080/01621459.2019.1686987

Wing, J. M. (2019). The data life cycle. *Harvard Data Science Review*, *1*(1).

Wing, J. M. (2020). *Trustworthy AI*. https://arxiv.org/abs/2002.06276

Wing, J. M., Janeia, V. P., Kloefkorn, T., & Erickson, L. C. (2018). *Data Science Leadership Summit.* Workshop Report. National Science Foundation. https://dl.acm.org/citation.cfm?id=3293458