Harvard Data Science Review • Issue 3.2, Spring 2021

Data Science Goals and Pragmatics

Alfred Z. Spector^{1,2}

¹Research and Special Initiatives, Google Inc.,

²Two Sigma Investments, New York City, New York, United States of America

Published on: Jun 07, 2021

DOI: <https://doi.org/10.1162/99608f92.643b1bf8>

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

Berkeley’s data science curriculum effectively integrates many key topics in a pedagogically accessible and efficient manner. Adhikari et al.’s [“Interleaving Computational and Inferential Thinking: Data Science for Undergraduates at Berkeley,”](#) in this issue, describes it well (2021).

- The curriculum’s emphasis on teaching by example is important: The vignettes make data science more tangible to students and provide the context that is typically needed when solving most problems.
- The curriculum’s notion of connector classes is also critical, as other faculties teach material relevant to the application of data science. Many students will be better prepared for employment or further study when they have pursued the data science curriculum supplemented by in-depth education in a related field such as computer science, statistics, economics, medical informatics, or many others. The connectors facilitate this hybridization of data science, which I notate as $DS + X$ for many fields, X .

Focusing specifically on the meaning of their term “computational and inferential thinking,” there can also be no doubt that data science must fuse concepts such as probability, inferencing, modeling, visualization, the study of algorithms, and engineering. In this context, engineering refers both to the methodology of abstraction, encapsulation, and reuse and the pragmatics of applying computational, storage, and networking power to facilitate data capture, processing, and use. Data science must further unite traditional statistical modeling techniques built on conceptual models with newer empirically created, algorithmic models based on machine learning and similar approaches (Breiman, 2001).

Despite Berkeley’s approach having nailed these topics, I have two observations:

1. The article does not emphasize data science’s breadth of goals. For example, terms like ‘optimization’ or ‘objective function’ get short shrift, perhaps because the field of operations research is not called out as a contributor to data science.
2. The article does not enumerate many of the pragmatic, implementation-related topics (for example, computer security or abuse-resistance) that make the pursuit of data science applications gnarly.

I hypothesize more discussion of these topics will make Berkeley’s already thoughtful curriculum even more compelling—by providing opportunities to include important new material and to more fully introduce students to the breadth of the field’s challenges. Enumeration of the broader set of topics will also ensure that the evolving vignettes continue to illustrate the needed breadth of material.

Toward a Broader Set of Objectives

Adhikari et al. begin their article by noting the great value of creating an undergraduate curriculum based on the “grand conceptual achievements of a field, stripping away the inessentials and conveying the core ideas in a way that reveals their beauty, their universality, and their contemporary relevance.” While unquestionably a valuable basis for a curriculum, I think a curriculum must also take into account a field’s objectives; without a

comprehensive statement of these, a curriculum could be limited in both method and especially application. In their recent piece on education, Fayyad and Hamutcu also argue it is also important to state clearly the goals of data science education (Fayyad, 2021).

So, this leads to the question, how do Adhikari et al. define data science? While they do not include a definition, their article is entirely consistent with the definition that has been used throughout the history of DS8, Berkeley’s Introductory Course, as shown in Figure 1.

What is Data Science?

Learning about the world from data using computation

- **Exploration**
 - Identifying patterns in data
 - Uses visualizations
- **Inference**
 - Using data to draw reliable conclusions about the world
 - Uses statistics
- **Prediction**
 - Making informed guesses about unobserved data

Figure 1. Visual from Berkeley Data 8 defining data science (Sahai, 2021).

I endorse the use of the words ‘exploration’ and ‘inference,’ but feel the sole use of the word ‘prediction’ is constraining. By comparison, for our forthcoming book, Peter Norvig, Chris Wiggins, Jeannette Wing, and I begin our definition of data science by first evoking the concepts of exploration and inference under our term ‘insight,’ but we then supplement prediction with five other goals. Our list of these goals thus starts with “the **prediction** of a consequence”; but adds: “the **recommendation** of a useful action; a **clustering** that groups similar elements; a **classification** that labels a grouping; a **transformation** that converts data to a more useful form, or an **optimization** that will move a system to a better state” (Spector et al., 2021).

Having these goals is important:

1. They establish the breadth of data science, setting forth such important topics as optimization of search and advertising, recommendation systems of all forms, related image matching and labeling, certain scientific applications of machine learning, automatic speech recognition and machine translation of language, financial portfolio selection, route finding, manufacturing optimization, and countless more.

2. These additional goals introduce complex challenges beyond those of prediction, both in technique and in precise objective. There are entire conferences devoted to them.
3. While Adhikari et al. suggest that their curriculum conjoins the ‘builder’ and ‘collaborator’ spirit, these additional goals add spirits such as ‘optimizer,’ ‘curator,’ and ‘transformer.’
4. Finally, many of these topics are excellent outlets for students’ post-graduation research and or other employment.

To check my analysis of Adhikari et al.’s focus, their first vignette on jury representation motivated me to create two columns of word frequencies, one from the authors’ article and one from the current draft of our manuscript, which builds on our broader definition of data science. I found only *five* contextually relevant references to the roots *optim** (e.g., optimization), *objective*, *recommend**, *transform**, *cluster**, and *classif** in comparison to a length-adjusted total of 70 in our present manuscript, thus illustrating the difference in focus.

Turning to the specifics, the argument for more emphasis on optimization arises from three directions:

First, operations research (OR) is a highly related predecessor discipline, for as Hillier and Lieberman write in their classic book, *Introduction to Operations Research*, the OR process, “begins by carefully observing and formulating the problem, including gathering all relevant data,” and “that OR frequently attempts to find a best solution (referred to as an optimal solution) for the problem under consideration” (Hillier, 2001). While traditionally, OR worked in a batch mode where models were built, calibrated with collected data, analyzed, optimized, and the resultant blueprint distributed for use, most data today are continually collected from the real world, fed into a model, and the model outputs used to do continual optimization of a system. Just as statistics is embracing algorithmic models, OR is hybridizing with machine learning aiming at increasingly adaptive and flexible approaches to optimization.

Second, *optimization* is an important use case of data science, bringing with it many types of algorithms as well as the enormous complexity of determining proper objective functions that balance competing interests. While objective functions (perhaps, by other names) are undoubtedly analyzed in the vignettes of Data 104—Human Contexts and Ethics of Data, I hypothesize it would make sense to call out the term ‘objective function’ as a critical concept for the curriculum as a whole, because deciding on objectives is one of the greatest challenges in deploying many data science applications. Notably, determining objective functions cannot just be placed under the concerns of ethics, for frequently their determination is mostly a complex commercial decision.

Third, many machine learning methods require the use of large-scale optimization in the training process—so optimization is involved not only in decision-making based on learned relationships, but also, in many cases, on learning those relationships from the available data.

Of the other terms not contextually referenced in Adhikari et al.’s article, *recommendation* seems particularly important to highlight. Though recommendation is related to prediction (a term the authors do emphasize),

recommendation is also related to optimization, and its contextual complexity provides unique problems and intellectual depth. Further, recommendation systems are perhaps the most common applications of data science, whether recommendations are for products, entertainment, news and social feeds, search results, or advertising. While recommendations are important for dealing with a barrage of information and generating the profit that funds the web, they are also problematic due to their ability to overly influence people and, more generally, the complexity of setting their objectives. For example, it is complex to balance the needs of consumers, publishers, advertisers, and exchanges in advertising systems. Thus, first class consideration would seem important to education in data science.

Turning to *transformation*, the combination of new deep learning models and large-scale training data has resulted in enormous progress on grand challenge problems, such as those of transforming speech to text or one human language to another. Beyond these highly visible results, data science is widely used to transform data into useful signals, in finance, medical diagnostics, security, and more.

Perhaps, Adhikari et al. would argue that they have *classification and clustering* covered as they are merely excellent examples of inference and prediction. Perhaps this is true, but classification and clustering have become highly specialized domains, and they are incredibly important in solving problems in image recognition, recommendation, spam detection, and more. Arguably, they should therefore be considered as first-class foci of data science.

More aggressively, the authors might argue that the detail I propose is broadly unnecessary as all the goals arise from inference and prediction. I think their argument may be weakest where optimization is an essential ingredient, but in all cases, I suggest the enumeration of the additional explicit goals creates opportunities for fascinating and valuable course material (e.g., on optimization or image recognition techniques). It also better educates students as to data science's breadth of challenges.

Promote Pragmatics

While the essence of computing is covered under the authors' computational thinking umbrella, much of the effort in inventing, designing, and operating data science applications arise from the pragmatics of developing systems: Some of these topics are most certainly mentioned within Data 100: Principles and Techniques of Data Science; for example, the authors list issues of scale, efficiency, and data quality. However, there are topics that aren't mentioned, as confirmed by the word frequency analysis. Immense attention must be applied to building in *security* and *resilience* in the face of failure, concerted user *abuse*, and *adversarial attacks*. There are real gotchas in applying data science when models do not offer *explanation*, or when scientists desire scientific *reproducibility* but that data cannot be released. While the authors do mention *privacy*, it is an exceedingly complex topic in multiple dimensions (e.g., human factors, risk management, spookiness, notions of manipulation) going well beyond what seems to be a focus on differential privacy. *Regulatory and liability*-related issues are already a significant topic and will become more so.

While one could be tempted to state that data scientists should ‘leave the pragmatics to the engineers,’ the complex feedback loop between the art of the desired (e.g., developing the best, most flexible model) and the art of the feasible (e.g., reducing privacy, security, failure, regulatory, and other risks) renders this compartmentalization impractical. In most of the data science efforts in which I’ve been involved, addressing these pragmatics takes the bulk of the time.

The authors no doubt intend for many of these challenges to be illustrated through their vignettes, but as I argued previously, explicit enumeration would be useful. On the other hand, I fully understand that these topics cannot be addressed in depth given time limitations. Full coverage of these pragmatics will of necessity be left to courses (e.g., security, distributed systems, optimization, causal analysis) in other disciplines.

Summary

Lest readers feel this is in any way an indictment of Adhikari’s article or the Berkeley curriculum, they should set that thought aside. I believe their article and the curriculum it represents are an overwhelming benefit to the field and the students. This short essay merely argues in favor of the authors (1) explicitly addressing certain additional goals of data science and (2) paying more deliberate emphasis to the pragmatic challenges of making data science work. If there is anything particularly fundamental in this discussion, it might be that operations research, with its focus on optimization, should be considered a first-class progenitor of data science.

Acknowledgments

Thanks to many, including Benjamin Spector and Emily Spector, who have commented at short notice.

Disclosure Statement

Alfred Z. Spector has no financial or non-financial disclosures to share for this article.

References

- Adhikari, A., DeNero, J., & Jordan, M. I. (2021). Interleaving computational and inferential thinking in an undergraduate data science curriculum. *Harvard Data Science Review*, 3(2). <https://doi.org/10.1162/99608f92.cb0fa8d2>
- Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3), 199–231. <https://doi.org/10.1214/ss/1009213726>
- Fayyad, U., & Hamutcu, H. (2021). How can we train data scientists when we can’t agree on who they are? *Harvard Data Science Review*, 3(1). <https://doi.org/10.1162/99608f92.0136867f>

Hillier, F. S., & Lieberman, G. J. (2001). *Introduction to operations research* (7th ed.). McGraw-Hill Higher Education.

Sahai, S. (2021). *Data 8-S21-L01-2021-02-20: Introduction*. [Video]. YouTube. <https://youtube/nESjEnI20gw?t=444>

Spector, A., Norvig, P., Wiggins, C., & Wing, J. (2021). *A holistic view of data science*. [Manuscript in Preparation]. <https://bit.ly/3wJCagt>

©2021 Alfred Z. Spector. This article 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 article.