Journal of Humanistic Mathematics

Volume 14 | Issue 1 January 2024

The Limits of Data Science

David E. Drew
Claremont Graduate University

The Limits of Data Science

Cover Page Footnote

I thank Brian Anderson, Megan Blanchard, Louis L. Bucciarelli, Dean Gerstein and anonymous peer reviewers for their comments on an earlier draft of this article.

The Limits of Data Science

David E. Drew

Claremont Graduate University, Claremont, California, USA
david.drew@cgu.edu

Synopsis

Data science can contribute valuable predictions in diverse fields. But I write to express some concerns and red flags. I suggest that data science is being oversold. This article contains three questions that I believe data science must address as this new discipline matures. *Is data science significantly different from statistics?* This is a question that has haunted the field since the term first was introduced. By creating algorithms based on current societal decision rules that may be biased, even bigoted, *does data science lock in and exacerbate inequality?* Scholars have identified a continuum from data to information to knowledge to wisdom, with the ultimate goal being wisdom. Data scientists seem to be acting as though data alone are enough. The big data systems are mathematically complex and have predictive power, but *can data science contribute significantly to our body of knowledge?*

Years ago, when I was in college, I had trouble choosing a major. Eventually I realized what the main problem was—there was no discipline that matched my interests. I wanted to work in the area where two Venn diagram circles overlapped: the social sciences and mathematics.

Today this would be called data science.

Back then I put together my own combination of courses, degrees, and work experiences to prepare myself for work in this hypothetical discipline. I took two years of physics and calculus, many statistics courses, and earned three degrees in sociology and social psychology. Between degrees, I worked four years as a computer programmer/software engineer and was head applications programmer at a leading research university. As I completed my Ph.D.,

I discovered an earlier classic essay by John Tukey [19], about applied data analysis as a valuable academic enterprise, and that helped to cement my professional identity.

I have since continued to enjoy teaching about, and conducting research using, quantitative research methods, particularly multivariate models.

All to say that I have been waiting for the development of data science as an exciting, productive emerging discipline. It often is described as a unique combination of mathematics, statistics, and computer science made possible by the ubiquity of massive data files and exponentially increasing computer power.

There have been some dramatic success stories in the world of data analytics, mostly in the business world. For example, a man complained bitterly to the manager of his local Target store that the company had erroneously sent advertisements and coupons about pregnancy products to his teenage daughter. Target responded quickly and appropriately, repeatedly apologizing for the error. Later, the man apologized to the Target manager. It turned out that his daughter was pregnant and the Target algorithms had known this before he knew it [3].

There also have been failures. For example, Google employed data analytic algorithms to predict where flu outbreaks might appear. Their model did not include any theory or empirical knowledge about the causes of flu; it just looked for patterns in the data. Google actually was able to make more accurate predictions than the Center for Disease Control one year. But, as often happens with big data, the solution was idiosyncratic. It was not replicated in other years, when the CDC clearly out-performed the Google predictions [9].

Data science and big data potentially can contribute valuable predictions in such diverse fields as medicine and management.

But I write now to express some concerns and red flags. I think data science is being oversold. I would like to present three challenges that I believe data science must address as this new discipline matures.

1. Is Data Science New?

Clearly neural networks and machine learning are so different from early pattern recognition algorithms that something new is happening. But much of what is presented as data science is a simple rebranding of statistics.

In the 1960s, Ford produced a modest sedan called the Falcon. It was a solid, conventional, unprepossessing car. Then Lee Iaccoca had a brilliant insight. Consumers could not afford to buy a sports car, but they wanted that image. He put a sports car-like shell over a Falcon engine and sold it at a price well below what sports cars cost, more like what a Falcon cost. The Mustang became a sensation. An early Mustang cost less than $2000, roughly equivalent to $16,000 today. The hit French movie *A Man and A Woman* was a love story that also was an ode to the Mustang. To this day it is a popular automobile and a classic example of successful rebranding.

For decades, when I have met people at a party, they ask what I do; I say I am a professor.

"What do you teach?" they ask. I answer "statistics". Then, inevitably, they start to back away and to look around me. They mumble something about how they had avoided statistics throughout college or they tell me how they almost did not graduate because of a stats course. I start to tell them that statistics can be mastered easily and can be fun, but it's too late. They are gone. I have lost them.

Recently I made a little change. Now when I am asked, I say that I teach data analytics. "Wow", they respond, "That is the hottest area. My nephew majored in data analytics and moved into a six-figure job after graduation. Tell me more about that."

Data analytics and data science have become the Ford Mustang of the academic world.

Here are some examples from articles about "feature engineering" that were posted on the Medium website, a site which is popular with Silicon Valley techies. I do not mean to criticize these articles. They provide an excellent presentation of feature engineering. But this presentation reveals that feature engineering is essentially another term for what statisticians refer to as descriptive statistics.

The first article [16] notes that feature engineering in data science is the third step, of five, in model building. The first two are gathering data and cleaning data; the fourth and fifth are defining the model and "training, testing the model, and predicting the output".

It turns out that a 'feature' is a variable. The author gives an example of variable transformation; a feature or variable of date and time combined, e.g., 2018-02-15 18:10, is transformed into an hour of day feature, e.g., 18, and adds, "Here, creating the new feature 'Hour of_Day' is the feature engineering." "Crossed column feature engineering" is combining two variables into a third. Recoding an interval variable into a three category ordinal variable, e.g., recoding age into three categories, is "bucketized column feature engineering". In another instructional article [15] this process is called "binning". "We can further bin the hour of the day into four categories, namely, morning, afternoon, evening and night." "Fixed width binning" means that you use categories of the same size. "Normalization" reduces the effect of outliers and can be accomplished by creating either what we traditionally call z scores or percentiles.

One author [2], at the end of a clear exposition of feature engineering, adds, "I cannot find any books or book chapters on the topic. . . ..It is a hard topic to find papers on." But there are many books and papers about these topics. However, they use different words, the established vocabulary of statistics.

Essentially, the analysts who conduct machine learning use feature engineering because they usually don't collect their own data. They are given data, for example, date and time stamps of commercial sales, and need to convert that information into data that meet the assumptions of curve-fitting, for example, continuous interval variables instead of nominal variables. The problem is that the technical people in these emerging fields begin with the assumption that none of this has been done before. They wind up reinventing the wheel.

Universities are struggling to find the best strategies for incorporating big data, data analytics and data science into their organizational structures and curriculum.

In my opinion, Princeton University found the right balance for exploring new frontiers while recognizing the strengths of existing statistical knowledge. A task force consisting of distinguished professors from across the university

was charged with assessing new developments in data analytics and making recommendations about what those developments implied for possible new structures within that institution. They proposed, and the university has implemented, a new Center for Statistics and Machine Learning (CSML) [18]. This title and structure underscore the importance of statistics and focus on what is most new: machine learning. With respect to data science, the university policy is reflected in this statement: "The Data Science Core is an especially innovative aspect of the task force's recommendations: the Core includes scientific programmers, data scientists, data administrators, and software developers whose role is to support the computational work of CSML faculty and students." [17, page 4]

The University of California at Berkeley has taken a somewhat different approach. "... the University is creating a unique and powerful institutional structure, provisionally named the Division of Data Science and Information, that connects departments from the College of Engineering, the College of Letters and Science, and the School of Information. It incorporates the Berkeley Institute for Data Science (BIDS), and curates a new Data Science Commons that catalyzes formation of groups of faculty and students from across the University to open new research domains and develop new fields of study empowered by the data revolution." [10]

The differences between the strategies of these two leading universities are, in part, semantic. But the ambiguities, especially between what is statistics and what is data science, have been present ever since the concept 'data science' was championed by Chien-Fu Jeffrey Wu in 1997 in his inaugural lecture as the H.C. Carver Professor at the University of Michigan. The title of that landmark lecture was "Statistics=Data Science?". (Earlier, Peter Naur [12] had suggested that the term "data science" replace "computer science".)

These ambiguities have implications for those of us who teach statistics and data analytic methods. When teaching a standard statistics course, we should strive to trace for our students how a given technique translates in the world of big data and data science. And data scientists should become familiar with techniques that already are available in the world of statistics. The historical precursors of statistics also are the precursors of data science, e.g., ancient tally sticks and the analytical work of Florence Nightingale.

Reinventing the wheel takes time and effort that can better be spent analyzing your data.

2. Does Data Science Lock in and Exacerbate Inequality?

Data science is all about prediction. If you can predict a dependent variable or outcome, your model is successful. In this sense, it is similar to multivariate statistics. But experienced sociologists, observers, and policy-makers have become alarmed when joy over the success of prediction outweighs the realization that the status quo unfairly penalizes some people and groups, the realization that something needs to be changed.

Metropolitan police departments have developed algorithms to predict where crimes will be committed. But they do so based on arrest records. Yet criminologists know that someone committing a crime in a poor neighborhood is more likely to be arrested than someone committing a crime in an upper middle-class neighborhood. See, for example, [11]. This means that arrest records yield biased data.

Similarly, data analytics may make the same individual more likely to qualify for a loan if he or she lives in a more affluent area than if they live in a poor area or ghetto. The algorithms are excellent at predicting what will happen under the status quo and pay no attention to how the status quo should, and must, be changed.

There was a consciousness raising, and a social movement for justice, in this country following the brutal killing of George Floyd. More and more people are becoming aware that many embedded algorithms represent a codified form of systemic racism and injustice.

Cathy O'Neil [13] has written extensively about the dangers of algorithms in large data systems.

She notes, "They tend to punish the poor. . . ..The privileged, we'll see time and again, are processed more by people, the masses by machines." (page 8) She adds, "Automated systems stay stuck in time until engineers dive in to change them. If a Big Data college applications model had established itself in the early 1960s. we still wouldn't have many women going to college, because it would have been trained largely on successful men." (page 204) Data scientists, even when well-intentioned, create systems that tend to penalize the poor and vulnerable; they usually do not ask, "How can this system be made more equitable?"

3. Will Data Science Contribute Significantly to Our Knowledge Base?

Naomi Klein [7] has written about the branding phenomenon and how it changed the business world. Companies now market a lifestyle, rather than simply selling sneakers or coffee. New brands create excitement. But corporations then commit massive funds to marketing, to creating the new lifestyle associated with the brand. Simultaneously, in a corporate zero-sum game, they are committing fewer funds towards producing quality products. Some of this seems to be happening with data science.

Throughout history, human beings have tried to understand and predict the world around them. Science is an organized method for doing this. Traditionally science progresses when a scientist develops a theory and then tests it. Usually this takes the form of making predictions based on the theory and then seeing if the predictions hold up in empirical data [8, 14]. A key component of scientific reasoning is that the theory must be "falsifiable".

In data science the sole goal is to make successful predictions, even if there is no theory. Google programmers can make successful predictions about user behavior without a psychological or cognitive theory to guide them. In his article, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete", Chris Anderson [1], perhaps tongue in cheek, writes, "Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves."

Some might argue that this is simply a form of inductive science. But the goal is not to gather data and then use inferences to build a theory. The only goal is prediction. In addition, as Martin Fricke [6] argues, "Connectionism, neural nets, random forests and so on in machine learning are, at heart, a black box instrumentalist version of inductive algorithms... If anything, science needs more theories and less data." (pages 653, 655)

This limited focus on prediction is consistent with a disturbing move in American higher education away from traditional knowledge and basic research towards applied research. The fastest growing undergraduate majors all involve computer science or information systems. The needs of industry increasingly drive curriculum offerings in our colleges and universities.

So perhaps this concern (prediction only versus understanding) is a revisiting of the perennial debate about applied versus basic research. Scholars have identified a continuum from data to information to knowledge to wisdom, with the ultimate goal being wisdom. Data scientists seem to be arguing that data alone are enough.

In the middle of the 20th Century, polio was a deadly plague in the U.S. The public was told to avoid 'still water', for example in ponds, although no one knew why. A data science approach to all the available data might have been able to predict who would get polio (although I doubt that) and could have produced predictive models of which people would be helped most by iron lungs or by spinal fusion surgery (two widespread methods for treating polio victims). But only the path-breaking scientific research of Enders, Salk and Sabin, exploring issues that would not have been captured by a data science model, would lead to a vaccine and bring an end to this plague.

Or consider climate change. Faghmous and Kumar [4] argue that, "Despite the urgency, data science has had little impact on furthering our understanding of our planet in spite of the abundance of climate data. This is a stark contrast from other fields such as advertising or electronic commerce where big data has been a great success story... solely relying on traditional big data techniques results in dubious findings, and we instead propose a theory guided data science paradigm that uses scientific theory to constrain both the big data techniques as well as the results-interpretation process to extract accurate insight from large climate data." (page 155) They add, "... the way the data are represented (in data science) relies heavily on attribute-value representation without any notion of contextual information (in space or time)." (page 159)

This is perhaps the greatest challenge facing data science and big data. Consider the usual sources of big data. Many analysts are employed by corporations trying to predict who will buy their products. So they work with variables like, for example, information about the time, location, and amount of the sale, the demographic characteristics of the consumer, and what other products they buy elsewhere. Many analysts are employed by the giant tech firms, e.g., Google and Amazon. They worry about search engine optimization and how to predict which users will click on which ads. Social network firms, e.g., Facebook and Twitter, focus on patterns of interactions among their users. These analytical systems work with variables that meet the

assumptions of mathematical and statistical systems, e.g., they often work with interval variables with normal distributions, but those variables measure relatively superficial aspects of the human experience. They are not measuring nor considering variables that assess and improve our quality of life, the variables that drive human experience.

The critical missing ingredient in most data science analyses is *meaning*. This has less to do with embedded algorithms, machine learning, neural networks and big data, and more to do with how you operationalize, or measure, abstract concepts—and, more to the point, whether you even consider them.

During World War II, Viktor Frankl, a psychiatrist, was imprisoned in a Nazi concentration camp [5]. He struggled to survive, and to understand the essential factor that prevented prisoners from giving up. "As this story is about my experiences as an ordinary prisoner, it is important that I mention, not without pride, that I was not employed as a psychiatrist in camp. . . .I was number 119,104 and most of the time I was digging and laying tracks for railway lines." (page 21) He concluded that "*Man's search for meaning is the primary motivation in his life* and not a 'secondary rationalization' of instinctual drives." (page 105, emphasis added)

As presently constituted, it does not appear that data science is contributing significantly to "man's primary motivation", the search for meaning.

4. In Conclusion

Data science will be strengthened if its proponents respond to these challenges:

- How can the education and training of data scientists be improved by coordinating terms and concepts with the existing body of knowledge in the field of statistics?
- How can we consider and integrate ethical values and social justice into algorithms, especially those that drive government and other social programs?
- How can we integrate and interpret current data science models with scientific theories and substantive knowledge?

References

[1] Chris Anderson (2008), "The end of theory: the data deluge makes the scientific method obsolete", *Wired Magazine,* June 27. http://www.uvm.edu/pdodds/files/papers/others/2008/anderson2008a.pdf

[2] Jason Brownlee (2014), "Discover feature engineering, how to engineer features and how to get good at it, *Data Preparation*, September 26, https://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/

[3] Charles Duhigg (2012), "How companies learn your secrets", *The New York Times Magazine,* February 16, https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?_r=1&ref=charlesduhigg

[4] James H. Faghmous and Vipin Kumar (2014), "A big data guide to understanding climate change: the case for theory-guided data science, *Big Data*, Volume 2, Issue 3, September 15.

[5] Viktor Frankl (1959) *Man's Search for Meaning*, Boston: Beacon Press.

[6] Martin Fricke (2015), "Big data and its epistemology", *Journal of the Association for Information Science and Technology,* Volume 66, Issue 4, pp 651-661.

[7] Naomi Klein (2000) *No Logo: Taking Aim at the Brand Bullies*. Canada: Random House, Picador.

[8] Thomas Kuhn (1962) *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

[9] David Lazer and Ryan Kennedy (2015) "What we can learn from the epic failure of flu trends", *Wired*, October 1, https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/

[10] Kara Manke (2018) "Berkeley inaugurates division of data science and information, connecting teaching and research from all corners of campus", *Berkeley News,* November 1. https://news.berkeley.edu/2018/11/01/berkeley-inaugurates-division-of-data-science-and-information-connecting-teaching-and-research-from-all-corners-of-campus/

[11] A Rafik Mohamed and Erik D. Fritsvold (2010) *Dorm Room Dealers: Drugs and the Privileges of Race and Class,* Boulder, CO: Lynne Rienner Publishers.

[12] Peter Naur (1974) *Concise Survey of Computer Methods.* Petrocelli Books.

[13] Cathy O'Neil (2016) *Weapons of Mass Destruction: How Big Data Increases Inequality and Threatens Democracy.* New York: Crown Publishing Group.

[14] Karl Popper (2002) *The Logic of Scientific Discovery.* Oxfordshire: Routledge.

[15] Karan Rajwanshi (2018) "Feature engineering: the key to predictive modeling", *Medium*, May 8, https://medium.com/@karanrajwanshi/feature-engineering-the-key-to-predictive-modeling-8f1935b3db4f

[16] Amit Shekhar (2018), "What is feature engineering for machine learning?" *MindOrks/Medium*, February 14, https://medium.com/mindorks/what-is-feature-engineering-for-machine-learning-d8ba3158d97a

[17] "Statistics and machine learning at princeton", a statement by President Christopher L. Eisgruber and Dean of the Faculty Deborah Prentice, Princeton University, August 22, 2016. https://strategicplan.princeton.edu/sites/strategicplan/files/response-to-statistics-and-machine-learning-report.pdf

[18] Task Force on Statistics and Machine Learning (2015), *The Future of Statistics and Machine Learning at Princeton University,* https://strategicplan.princeton.edu/sites/strategicplan/files/task-force-report-on-the-future-of-statistics-and-machine-learning.pdf

[19] John W. Tukey, (1962) "The future of data analysis", *Annals of Mathematical Statistics*, Vol. 33, No. 1, pp. 1-67.