Harvard Data Science Review • Issue 2.2, Spring 2020

Data, Data, Everywhere . . .

Nicole Lazar¹

¹Department of Statistics, Franklin College of Arts and Sciences, University of Georgia, Athens, Georgia, United States of America

Published on: May 27, 2020

DOI: <https://doi.org/10.1162/99608f92.a6e7a24e>

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

Column Editor's Note: *This inaugural column explores the big scene of data science for pre-college students and their parents and teachers, as well as for any novice who wishes to have a peek into the scene. It explains some 'buzz words,' discusses the growing role of data science in our daily lives, and presents examples as food for thought. The aim is to inspire young and new minds to consider data science as a necessary intellectual pursuit for an educated citizen in the digital age, even if one does not necessarily aspire to become a data scientist.*

Keywords: computer science, data literacy, data science, pre-college education, statistics, uncertainty

It is a common saying that we live now in a data-driven age. But what does this actually mean? There are many buzz words to describe this new reality: *data science*, *big data*, *machine learning*, even *artificial intelligence*—which is not exactly new but has gained new life in recent years. It's true that there are more data at our fingertips than ever before, with smartphones (and other so-called 'smart' devices, everything from watches to refrigerators) collecting information about our location and activities; web browsers that store (and pass on to others) our search history; and social media that track our interactions—and this is just the tip of the iceberg!

Indeed, it is an interesting time to be a data scientist or statistician (and some may ask: what is the difference between these two? More on this later.). In addition to the personal data that most of us generate on a daily basis, advances in medicine, psychology, biology, and the social sciences have all added to the steady stream of opportunity and challenge for people who are 'data literate' (another buzz word) and 'uncertainty literate.' We are truly awash in data in a way that would not have been imaginable not that long ago; along with that comes an increased need to make decisions under uncertainty, something that is a part of everyday life whether we like it or not. Even within the last five years, the data landscape has shifted significantly. Such rapid change is unprecedented in the history of statistics and data. For most of the past hundred years, the span of time in which many statistical methods were developed, data were small and relatively structured. That is no longer the case. Data are more varied and vaster than ever before, requiring the development of new ways of analysis—hence some of the buzz words that are bandied about in the popular press and scientific literature alike.

This is both exciting and overwhelming, especially if you are new to data science as a field, or have heard about it but are not really sure what it is. What, then, do people mean when they talk about 'data science'? For a simple starting point, you can think of data science as a big term for the fields that work with data, in an effort to learn what they can tell us about a phenomenon of interest and how they can guide us through its inherent uncertainty. This may include statistics—a central player in data science—as well as computer science, applied math, and some areas of engineering, depending on whom you ask.

In this column, and future columns in the *Minding the Future* series, the aim is to introduce the big picture of data science, particularly to younger readers, parents, and teachers, and to demystify what can, at times, seem

like a confusing array of terminology and concepts. Along the way, we hope to help you develop into more data literate and uncertainty literate consumers of information, a key part of being an educated citizen of the modern world.

Let's Talk About Data . . .

An important first step is to understand the different types of data available, their meanings and uses, and also to be aware of how data can be abused and misused. 'Data for good' is another term that has gained popularity in recent years. A group of students at the University of Georgia, where I teach, started an organization dedicated to using their data analysis skills for the betterment of society, completely unaware that this is a bigger, worldwide movement. The data for good movement has several aims: to increase the ethical use of data in the public sphere; to raise awareness of how data are used to make decisions that may be biased against certain groups of people (as in the popular book *Weapons of Math Destruction*); and to point out the ethical implications of big corporations having access to all of our personal information, much of which we freely, though often without realizing, give up to them. After all, who carefully reads the terms of agreement before signing on to a new type of social media, online game, or convenient website?

Governments also routinely collect data on their citizens. For the most part, this is done to assess the needs of the populace and hence to direct services provided by the government to the places where they are needed most: schools, medical care, or infrastructure such as roads and bridges. Yet there is also the concern that, in the wrong hands, such seemingly innocuous data could be used in ways that might cause potentially disproportionate harm to certain groups. An example that received a lot of publicity in the United States this past year dealt with the addition of a citizenship question to the 2020 census. Census data, collected every ten years in the United States, are used to allocate resources as well as seats in the House of Representatives. The concern was that inclusion of a citizenship question would deter some people—specifically non-citizens in this instance—from responding to the census, resulting in undercounting and misallocation of resources. The census is supposed to count all residents, not just citizens, so a severe undercount of that group would have been a real problem.

Who owns your personal data: your shopping habits, your online activities, and your social network? These are data that somebody else collects about you, maybe through a website or a free app. Current laws do not fully address such questions, though some countries have started to consider issues of confidentiality and data collection. For example the European Union (EU) instituted the General Data Protection Regulation (GDPR) in May 2018. The regulation protects data privacy within the EU and the European Economic Area, giving individuals more control over their personal data and regulating the transfer of such data outside the EU and EEA. A similar law, the California Consumer Privacy Act, took effect in California on January 1, 2020. The law allows users to see what sorts of data companies collect on them, and gives them the ability to refuse companies the rights to use or sell those data. Also in the US, former Democratic presidential candidate Andrew Yang—noting that data are more valuable than oil—has stated that tech companies such as Facebook

and Amazon should pay people for the use of their data.¹ Who might take advantage of such an offer, and who might value their rights to privacy and to own their own data, more than any monetary gain? While people will weigh the options differently and hence reach different conclusions, it is not hard to imagine that some groups taken as a whole might be more willing to sell their data compared to other groups.

It is important for modern data analysts to think about questions of this type, and that is before one even gets to the data themselves. There is a myriad of interesting questions that can be explored using approaches from statistics, computer science, and the basic sciences: what we are broadly calling data science.

. . . And Collect Some Data Examples

A few examples show how wide the span is. Imaging technologies have given neuroscientists new insights into how the human brain functions. Techniques such as functional magnetic resonance imaging (fMRI), electroencephalography (EEG), magnetoencephalography (MEG), and others generate massive amounts of data on the brain in action. Such data are not only big (terabytes in size) but complicated in ways that pose challenges for the data analyst. The interesting brain images that you might see on the cover of a popular science magazine purporting to show differences between healthy individuals and those with some disease, are the output of sophisticated statistics or machine learning algorithms that take years to fully master.

Similar advances in genetics have led to a revolution in genomics data, whereby scientists can study in previously-impossible detail the very makeup of the human (or other) being. Services such as 23andMe, which give you a peek at your genetic background, are relatively recent arrivals on the scene; they also rely on data science to build their stories.

The field of digital humanities—which brings statistical analysis, machine learning techniques, and computational methods to the humanities—has a surprisingly long history. Already in 1963, well before the modern computing revolution, statisticians Fred Mosteller and David Wallace used statistical methods to analyze the 12 disputed Federalist Papers. It was long surmised that these 12 essays had each been written either by James Madison or Alexander Hamilton. Mosteller and Wallace used differences in word choice to attribute authorship of the papers to Madison. The discipline of digital humanities has particularly flourished and gained visibility in the last decade. The [Stanford Literary Lab](#), for instance, is a collective of researchers who use computational tools to study literature, focusing on topics as varied as fanfiction and the effects of translation on a text. Digital humanities labs and centers are now found in universities worldwide.

Businesses have also of course been quick to jump on the data bandwagon, as have sports teams, who use increasingly sophisticated data collection and analysis tools to improve their records (think the book and movie *Moneyball*). Many of these corporations have dedicated data science groups. The prevalence of such was brought home to me recently, when our department hosted a visitor from The Home Depot, a senior member of their data science team. For many of us, it was news that The Home Depot even had a data science team, although it probably shouldn't have been. His talk was enlightening in a number of regards.

First, he described the many ways that a chain that specializes in home repair, home projects, appliances, and supplies uses data science. Though his particular examples were from The Home Depot, most people these days have experience with online shopping, whether for patio furniture, clothes, school supplies, music, or games. The recommendations that pop up after a search, as well as the advertisements that you subsequently see when you go to other websites, are the results of algorithms developed and tested by data science teams.

Second, our guest discussed several algorithms that few of us in the audience had heard of, indicating that new developments continue to unfold at a rapid pace. As data scientists, there are always going to be new things for us to learn, since new types of data and new data analysis problems require different solutions to what worked in the past. The field has seen this before, with machine learning, for example, and we can anticipate that it will continue to be the case.

Third, he emphasized the interdisciplinary nature of many data science teams. The people he works with come from a variety of backgrounds: statistics, computer science, and machine learning, to name a few. Data science teams may also include applied mathematicians, engineers, and domain experts. There is a lot of overlap among the different disciplines, and even to those of us working in the worlds of data, the distinctions between data science and statistics in particular are not always clear.

A popular description goes something like this: ‘A data scientist knows more computing than a typical statistician and more statistics than a typical computer scientist, but can communicate better than either of them.’ This somewhat tongue-in-cheek summary does, however, highlight important ideas for those who are interested in learning more. Mathematical skills are crucial, as is knowledge of computing (programming, database management, and the like). Communication refers to being able to explain what you did, be it in writing, orally, through the use of data visualizations, or other means. In addition, data scientists often have a serious, in-depth engagement with a data domain: neuroscience, genetics, politics, engineering, business, literature, or meteorology, to name a few.

What About Uncertainty?

Clearly, an understanding of data and how they are used is important for the modern citizen. Schools and universities hence are increasingly emphasizing the idea of data literacy for all their students. Equally important is to think about uncertainty, both in the data (which are almost always collected with some noise) and in how we use data to make decisions. This issue is playing out in front of all of us in real-time, as the world confronts the COVID-19 pandemic. Governments (and individuals) have had to make decisions about appropriate actions in the face of great uncertainty: how widespread is the virus? how easily does it spread? how lethal is it really? There are many unknowns in the current situation, in part because good data are hard to obtain in the midst of the crisis itself, and indeed they may never be fully available.

On a more prosaic note, we all need to make decisions with incomplete data sometimes.

As a student, you might be applying to universities early decision. How do you know which is your ‘best’ choice? What do you mean by ‘best’ for that matter? How do you decide which date to ask to the prom? How do school districts decide whether to close because of bad weather? In the Southern U.S. where snow is relatively rare and cities are unprepared to deal with it, even a hint of the possibility of snow may lead to school closures. It is easy to laugh when the snow fails to materialize and children are sitting at home for no reason, but think of the other side: schools do not close and a big snowstorm comes to a city with no plows. Sometimes the right decision (what one should have done) is clear in retrospect; other times, it will never be known. Grappling with these facts, understanding them, and not expecting or demanding solid answers in every circumstance is the sort of uncertainty literacy that is important in a world where (often incomplete, messy) data play a critical role.

Minding the Future

With all of these different types of data and data-related questions, it is no surprise that data scientists are in high professional demand. People who can analyze, interpret, and present data in meaningful, understandable ways will play an ever-increasing role in science, business, education, government agencies, and industries of all variety - the horizons are near-limitless. Likewise, those who can help society more generally work through the ethical implications and complications of our data-rich world should have an important voice in protecting the public from abuses and misuses of their personal information. The realms of data science and its constituents are vast, and you are invited to explore them. The goal of *Minding the Future* is to be a guide on this journey: to demystify and decode some of the buzz words; to inform about important and often confusing concepts; and to provide fun activities that you can try at home or in school to deepen your own data science knowledge. We welcome questions from readers, and especially suggestions for topics you would like to see covered in future columns.

Disclosure Statement

Nicole Lazar has no financial or non-financial disclosures to share for this article.

References

Lewis, M. (2003). *Moneyball: The art of winning an unfair game*, New York: W.W. Norton & Company.

Mosteller, F. and Wallace, D.L. (1963). Inference in an authorship problem. *Journal of the American Statistical Association*, 58(302), 275–309. <https://doi.org/10.1080/01621459.1963.10500849>

O'Neal, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York: Crown Books.

Sonnemaker, T. (2019, November 14). Andrew Yang wants you to make money off your data by making it your personal property. *Business Insider*. <https://www.businessinsider.com/andrew-yang-data-ownership-property-right-policy-2019-11>

Yang, A. (2020). *Data as a property right*. <https://2020.yang2020.com/policies/data-property-right/>

©2020 Nicole Lazar. 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.

Footnotes

1. Details about his thoughts on data as property can still be seen (as of March 2020) on the campaign's website. [↔](#)