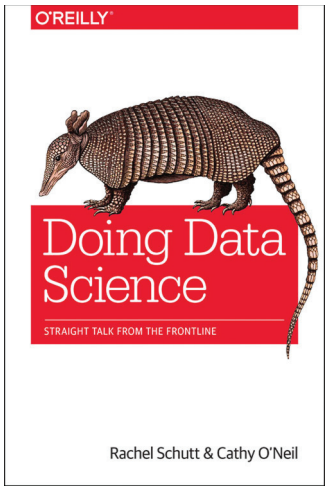


Doing Data Science

Reviewed by Brian Hayes

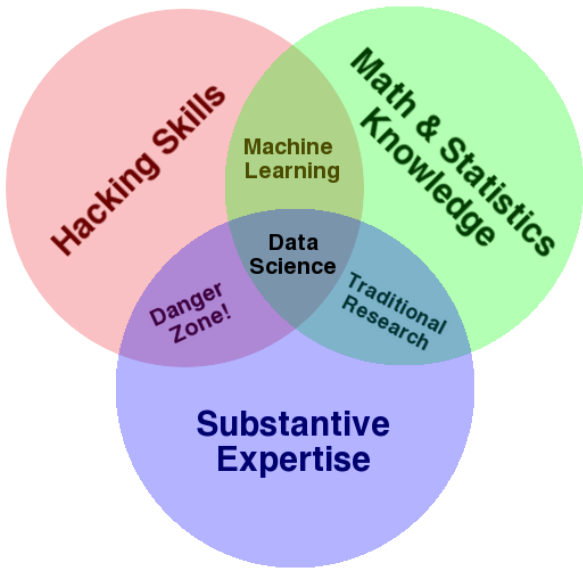
Doing Data Science
Rachel Schutt and Cathy O'Neil
O'Reilly Media, 2014
US$39.99, 375 pages
ISBN 978-1-449-35865-5

"Data Scientist: The Sexiest Job of the 21st Century"—that was the title of a 2012 article in *Harvard Business Review*. Many of us, I suspect, have never met a data scientist, and perhaps never heard of one. Although there's mild controversy about the provenance of the term, it seems the first business cards bearing that job title were printed in 2008. By 2011, Michael Rappa of North Carolina State University counted 394 individuals identifying themselves as data scientists. He came up with this number by doing a little data science of his own: He searched the LinkedIn social network, counting professional profiles with "data scientist" as part of a present or previous job title. In May of 2014 I repeated that experiment and found the population of data scientists on LinkedIn had grown to 4,696.

So what is this sexy new science of data? In *Doing Data Science* Rachel Schutt and Cathy O'Neil take up this question at the start of the first chapter, and it remains open for discussion in the final chapter. Here is one proposed definition:

> [A] data scientist is someone who knows how to extract meaning from and interpret data, which requires …tools and methods from statistics and machine learning, as well as being human.

Another attempt at a definition takes the form of a Venn diagram (created by Drew Conway), suggesting that data science lies at the three-way intersection of mathematical statistics, computing, and expertise in some particular subject domain.



(Why is the intersection of hacking skills and substantive expertise labeled a danger zone? Because those without grounding in mathematics and statistics risk producing results they don't understand.)

Brian Hayes is senior writer for American Scientist *magazine and writes the blog Bit Player,* `http://bit-player.org`. *His email address is* `brian@bit-player.org`.

DOI: http://dx.doi.org/10.1090/noti1167

A third definition is attributed to Josh Wills, Director of Data Science at Cloudera:

> Data scientist (noun): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

If none of those definitions gives you a clear sense of just what it is that data scientists do, maybe a few examples will prove more illuminating:

Recommendation engines. When you buy a book from an online merchant, the website presents a list of other items that might tempt you. Where do those suggestions come from? If you were shopping at a neighborhood bookshop (supposing that your neighborhood still has such quaint institutions), recommendations might come from a well-read clerk, relying on personal knowledge of both customers and literature. But such individualized services are not feasible for an online retailer with millions of customers and millions of items for sale. The solution is a "recommendation engine," which Schutt and O'Neil call "the quintessential data-science product." The main source of data to fuel the engine is the huge bipartite graph linking customers with the products they have bought. When you order a copy of *Doing Data Science*, the engine can consult the graph to find other customers who bought the same book (or browsed in it, or reviewed it), then look for other titles that also interested those people.

Fraud detection. Credit-card transactions stream into a bank processing center at a rate of hundreds per second. Some small fraction of the transactions are fraudulent: The purchaser is presenting a stolen or counterfeited card, or perhaps the merchant is making an unauthorized charge to a customer's account. The data scientist's job is to identify these rogue transactions, using algorithms that have access to historical data for both the buyer and the seller. What features of individual transactions will most clearly discriminate between the illicit and the legitimate ones?

Social network analysis. A social network—such as the LinkedIn service mentioned above—can be represented as a mathematical graph: The people are vertices, and the connections between them are edges. Social graphs have a distinctive statistical structure. They are sparse graphs, with relatively few edges overall, and yet almost any two vertices are connected by a short path, traversing no more than a few edges. In other words, in these "small world" graphs, friend-of-a-friend links tie everyone together. Part of what makes the networks so cohesive is the presence of a few individuals with a very large number of contacts, and others who act as bridges between communities that would otherwise be isolated. Identifying these key individuals and the communities they influence is another job for a data scientist.

Masters of the Data Universe

What is it about tasks like these that accounts for the sex appeal of data science? Part of the thrill may be a simple matter of scale. The data scientist claims dominion over a planet-girdling empire of digital commerce and online life. For example, the largest social networks, such as Facebook, now have 10^9 nodes, approaching the size of the entire human population. The masters of this data universe, striding their realm with youthful swagger, are not always gentle as they sweep away the outworn ideas of earlier generations. One advocate of data-driven machine inference remarks: "Your decades of specialist knowledge are not only useless, they're actually unhelpful." In other words, get out of the way and let the algorithms do their work.

Most of the software tools and computational infrastructure built to deal with these huge data sets might properly be described as data engineering rather than data science. Yet there are issues of genuine scientific and mathematical interest underlying such activities. For example, the problem of extracting meaning from large, high-dimensional data arrays is not just a matter of data processing, to be solved by installing a bigger computer. Many of the inference and prediction procedures in data science rely on clustering algorithms, which partition data into subsets of values that are all near one another according to some metric. Those algorithms run up against an impossibility theorem for clustering, formulated by Jon Kleinberg of Cornell and reminiscent of the Arrow impossibility theorem for elections. Kleinberg lists three desirable criteria for a clustering function, which he calls scale invariance, richness, and consistency, and he shows that no algorithm can satisfy all three. Kleinberg's theorem is not mentioned in *Doing Data Science*, but other limitations of algorithms for clustering, classification, and ranking are discussed with some care. (Admittedly, such cautions may not dampen the boisterous enthusiasms of young people impatient to go out and change the world. Perhaps that's for the best.)

The two authors of *Doing Data Science* partake of the enthusiasms, but they also bring a measure of maturity and experience to the subject. Schutt is a mathematician and statistician, an adjunct professor at Columbia University; since the book was published she has become Vice President of Data Science at News Corp. O'Neil is a mathematician who left the academic world to work in finance, then turned away from that career as well, becoming active in the Occupy Wall Street movement; she is now a Data Science Consultant

at Johnson Research Labs in New York. She also writes a blog called `mathbabe.org`. In 2012 Schutt undertook to teach an introductory data science course at Columbia. O'Neil audited the course and reported on the experience in her `mathbabe` blog. The two authors then drew together the blog posts and other material to create the book.

Doing Data Science is not a tutorial or a textbook. It introduces lots of basic principles and techniques—probability distributions, linear regression, Bayes's theorem, various algorithms for machine learning—but none of these ideas are presented in great depth or detail. Most of the chapters are based on talks by guest lecturers, who chat about their tools, their tastes, and their habits of thought, then present one of their projects, perhaps illustrated with a few equations or snippets of code. It's like a television cooking show where every week a different celebrity chef comes to prepare a signature dish. Seeing the masters at work is entertaining and even inspiring, but when you go into the kitchen, you may realize you didn't learn quite enough to make that *vol-au-vent* on your own.

The King-of-the-Mountain Metric

Data science has its detractors. One of them is Cosma Shalizi, a statistician at Carnegie Mellon University. Schutt and O'Neil paraphrase his position as follows:

> Cosma basically argues that any statistics department worth its salt does all the stuff in the descriptions of data science that he sees, and therefore data science is just a rebranding and unwelcome takeover of statistics.

Shalizi may be right, but one could also argue that the problem with data science is that there are parts of statistics it has *not* yet assimilated. Some of the parts left out are really good parts.

Let us go back to the beginning of data science—or maybe it was before the beginning. In 2006 the movie rental company Netflix announced a contest: They would pay a million-dollar prize to anyone who could improve the accuracy of their recommendation engine by 10 percent or more. Some 20,000 teams registered for the competition. Contestants were given data showing how 500,000 viewers rated various subsets of 17,000 films; in all, there were about 100 million ratings in this training set. The challenge was to predict an additional three million ratings. A team from BellCore won a preliminary round of the contest. Their strategy was to apply a wide variety of algorithms to the training set—eventually they had 107 of them—then take a weighted sum of the predictions; the weights were tuned to maximize the score. In 2009 the BellCore team merged with two others, each of which added still more methods to the mix, and this consortium won the grand prize.

Elements of the Netflix contest seem to have become permanent fixtures of the data science scene. In particular, competitions remain a popular way of stimulating work on a problem and evaluating progress toward a solution. A company called Kaggle has made a business of conducting such contests. Moreover, many of the contest winners still favor a scattershot strategy, in which multiple algorithms are flung at the problem, with the final result being some weighted combination of their outputs. "Overfitting" is a constant hazard: When you work too hard at optimizing the weights, you may find you have tuned the model to mere noise in the training set, impairing performance on real-world data.

The continuing success of multi-algorithm mashups in open competition is undeniably an argument for their soundness. Nevertheless, I am disappointed to learn that we can't measure the performance of an optimization technique in a more meaningful way than to say that nobody has been able to beat it so far. This king-of-the-mountain metric tells us almost nothing about any fundamental bounds on accuracy or efficiency. You can know where you stand with respect to other contestants, but not how closely you might be approaching a true limit on predictive ability. And a program that combines outputs from more than 100 algorithms makes it hard to discern which techniques work best, or how to formulate more general procedures that can be applied to a wider variety of problems.

In statistics, by contrast, it's not the usual practice to choose a data model or estimator by holding a prize competition. There's a body of systematic knowledge that guides such decisions, generally leading to a single solution or a small set of alternatives, and quantifying the error and uncertainty in the results. Data science, as far as I can tell, has yet to develop its central limit theorem. Of course it is still very young.

So far, data science has evolved mainly outside the academic world, at Google and Facebook and a host of smaller startup companies. But if this new suite of ideas and techniques is to sustain itself, it will have to find a place in the university as well. A century ago, when statistics was emerging as a distinct academic discipline, there was some doubt about where it should make its intellectual and institutional home. The underlying ideas were clearly mathematical, but the new field also had strong affinities with social, political, and biological sciences. In the end, statistics did not become just another branch of mathematics, on the same level as number theory or combinatorics. Statisticians

stand at a slightly greater remove; to borrow a metaphor from politics, they are independents who caucus with the mathematicians.

Questions about intellectual and institutional affiliations arose again when computer science was born in the 1960s and 1970s. The outcome in that case was even greater autonomy, although the computing professions still have strong ties to mathematics on the one side and engineering on the other.

We may now be witnessing the birth of another new discipline. Will the community of data scientists be captured by either statistics or computer science? Or will it develop its own institutions—membership societies, journals, annual meetings, university departments?

The Students' View

Schutt and O'Neil allow the younger generation to have the last word in their book. In the final chapter the students in Schutt's course report their reactions to the curriculum and reflect on the careers they are about to launch. Interestingly, it is the students who most directly confront ethical issues and the broader role of data and data science in peoples' lives. They cite a comment from Jeff Hammerbacher, who was one of the two pioneers who first called themselves data scientists:

> The best minds of my generation are thinking about how to make people click ads…That sucks.

It's a sobering thought, and the students express determination to put their skills to better use. One naturally hopes that the brightest minds will be drawn to the deepest and most important questions. But sexy jobs also matter.