CHAPTER 16

Next-Generation Data Scientists, Hubris, and Ethics

We want to wrap up by thinking about what just happened and what we hope for you going forward.

What Just Happened?

The two main goals of this book were to communicate what it's like to be a data scientist and to teach you how to do some of what a data scientist does.

We'd like to think we accomplished both of these goals.

The various contributors for the chapters brought multiple first-hand accounts of what it's like to be a data scientist, which addressed the first goal. As for the second goal, we are proud of the breadth of topics we've covered, even if we haven't been able to be as thorough as we'd like to be.

It's possible one could do better than we have, so the next sidebar gives you something to think about.

Thought Experiment: Teaching Data Science

How would you design a data science textbook?

It's not a well-defined body of knowledge, and there's no canonical corpus. It's popularized and celebrated in the press and media, but there's no "authority" to push back on erroneous or outrageous accounts. There's a lot of overlap with various other subjects as well; it could become redundant with a machine learning textbook, for example.

How does one measure the success and impact of a data science text? What's the counterfactual?

Can we set this up as a data science problem? Even better, can we use a causal modeling approach? This would require finding people who were more or less like you (dear reader) but didn't buy this book, and then use propensity score matching. Or we could run an experiment and randomly deny people access to the book (hard since Amazon is pretty open access), and you could get to it other ways. But that's neither here nor there, except it might be better than speculating whether the book made a difference to anyone.

In the industry, they say you can't learn data science in a university or from a book, that it has to be "on the job." But maybe that's wrong, and maybe this book has proved that. You tell us.

What Is Data Science (Again)?

We revisited this question again and again throughout the book. It was the theme of the book, the central question, and the mantra.

Data science could be defined simply as what data scientists do, as we did earlier when we talked about profiles of data scientists. In fact, before Rachel taught the data science course at Columbia, she wrote up a list of all the things data scientists do and didn't want to show it to anyone because it was overwhelming and disorganized. That list became the raw material of the profiles. But after speaking to different people after the course, she's found that they actually like looking at this list, so here it is:

- Exploratory data analysis
- Visualization (for exploratory data analysis and reporting)
- Dashboards and metrics
- Find business insights
- Data-driven decision making
- Data engineering/Big Data (Mapreduce, Hadoop, Hive, and Pig)
- Get the data themselves
- Build data pipelines (logs→mapreduce→dataset→join with other data→mapreduce→scrape some data→join)
- Build products instead of describing existing product usage
- Hack
- Patent writing
- Detective work
- Predict future behavior or performance
- Write up findings in reports, presentations, and journals
- Programming (proficiency in R, Python, C, Java, etc.)
- Conditional probability
- Optimization
- Algorithms, statistical models, and machine learning
- Tell and interpret stories
- Ask good questions
- Investigation
- Research
- Make inferences from data
- Build data products
- Find ways to do data processing, munging, and analysis at scale
- Sanity checking
- Data intuition
- Interact with domain experts (or be a domain expert)
- Design and analyze experiments
- Find correlation in data, and try to establish causality

But at this point, we'd like to go a bit further than that, to strive for something a bit more profound.

Let's define data science beyond a set of best practices used in tech companies. That's the definition we started with at the start of the book. But after going through this exploration, now consider data science to be beyond tech companies to include all other domains: neuroscience, health analytics, eDiscovery, computational social sciences, digital humanities, genomics, policy…to encompass the space of all problems that could possibly be solved with data *using* a set of best practices discussed in this book, some of which were initially established in tech companies. Data science happens both in industry and in academia, i.e., *where* or *what domain* data science happens in is not the issue—rather, defining it as a "problem space" with a corresponding "solution space" in algorithms and code and data is the key.

In that vein, start with this: data science is a set of best practices used in tech companies, working within a broad space of problems that could be solved with data, possibly even at times deserving the name science. Even so, it's sometimes nothing more than pure hype, which we need to guard against and avoid adding to.

What Are Next-Gen Data Scientists?

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

Ideally the generation of data scientists-in-training are seeking to do more than become technically proficient and land a comfy salary in a nice city—although those things would be nice. We'd like to encourage the next-gen data scientists to become problem solvers and question askers, to think deeply about appropriate design and process, and to use data responsibly and make the world better, not worse. Let's explore those concepts in more detail in the next sections.

Being Problem Solvers

First, let's discuss the technical skills. Next-gen data scientists should strive to have a variety of hard skills including coding, statistics, machine learning, visualization, communication, and math. Also, a solid foundation in writing code, and coding practices such as paired

programming, code reviews, debugging, and version control are incredibly valuable.

It's never too late to emphasize exploratory data analysis as we described in Chapter 2 and conduct feature selection as Will Cukierski emphasized. Brian Dalessandro emphasized the infinite models a data scientist has to choose from—constructed by making choices about which classifier, features, loss function, optimization method, and evaluation metric to use. Huffaker discussed the construction of features or metrics: transforming the variables with logs, constructing binary variables (e.g., the user did this action five times), and aggregating and counting. As a result of perceived triviality, all this stuff is often overlooked, when it's a critical part of data science. It's what Dalessandro called the "Art of Data Science."

Another caution: many people go straight from a dataset to applying a fancy algorithm. But there's a huge space of important stuff in between. It's easy to run a piece of code that predicts or classifies, and to declare victory when the algorithm converges. That's not the hard part. The hard part is doing it well and making sure the results are correct and interpretable.

What Would a Next-Gen Data Scientist Do?

Next-gen data scientists don't try to impress with complicated algorithms and models that don't work. They spend a lot more time trying to get data into shape than anyone cares to admit—maybe up to 90% of their time. Finally, they don't find religion in tools, methods, or academic departments. They are versatile and interdisciplinary.

Cultivating Soft Skills

Tons of people can implement k-nearest neighbors, and many do it badly. In fact, almost everyone starts out doing it badly. What matters isn't where you start out, it's where you go from there. It's important that one cultivates good habits and that one remains open to continuous learning.

Some habits of mind that we believe might help solve problems[1] are persistence, thinking about thinking, thinking flexibly, striving for accuracy, and listening with empathy.

Let's frame this somewhat differently: in education in traditional settings, we focus on answers. But what we probably *should* focus on, or at least emphasize more strongly, is how students behave *when they don't know the answer*. We need to have qualities that help us find the answer.

Speaking of this issue, have you ever wondered why people *don't* say "I don't know" when they don't know something? This is partly explained through an unconscious bias called the Dunning-Kruger effect.

Basically, people who are bad at something have no idea that they are bad at it and overestimate their confidence. People who are super good at something underestimate their mastery of it. Actual competence may weaken self-confidence. Keep this in mind and try not to over- or underestimate your abilities—give yourself reality checks by making sure you can code what you speak and by chatting with other data scientists about approaches.

Thought Experiment Revisited: Teaching Data Science

How would you design a data science class around *habits of mind* rather than technical skills? How would you quantify it? How would you evaluate it? What would students be able to write on their resumes?

Being Question Askers

People tend to overfit their models. It's human nature to want your baby to be awesome, and you could be working on it for months, so yes, your feelings can become pretty maternal (or paternal).

It's also human nature to underestimate the bad news and blame other people for bad news, because from the parent's perspective, nothing one's own baby has done or is capable of is bad, unless someone else

1. Taken from the book *Learning and Leading with Habits of Mind*, edited by Arthur L. Costa and Bena Kallick (ACSD).

somehow made them do it. How do we work against this human tendency?

Ideally, we'd like data scientists to merit the word "scientist," so they act as someone who tests hypotheses and welcomes challenges and alternative theories. That means: shooting holes in our own ideas, accepting challenges, and devising tests as scientists rather than defending our models using rhetoric or politics. If someone thinks they can do better, then let them try, and agree on an evaluation method beforehand. Try to make things objective.

Get used to going through a standard list of critical steps: Does it have to be this way? How can I measure this? What is the appropriate algorithm and why? How will I evaluate this? Do I really have the skills to do this? If not, how can I learn them? Who can I work with? Who can I ask? And possibly the most important: how will it impact the real world?

Second, get used to asking other people questions. When you approach a problem or a person posing a question, start with the assumption that you're smart, and don't assume the person you're talking to knows more or less than you do. You're not trying to prove anything—you're trying to find out the truth. Be curious like a child, not worried about appearing stupid. Ask for clarification around notation, terminology, or process: Where did this data come from? How will it be used? Why is this the right data to use? What data are we ignoring, and does it have more features? Who is going to do what? How will we work together?

Finally, there's a really important issue to keep in mind, namely the classical statistical concept of causation versus correlation. Don't make the mistake of confusing the two. Which is to say, err on the side of assuming that you're looking at correlation.

What Would a Next-Gen Data Scientist Do?

Next-gen data scientists remain skeptical—about models themselves, how they can fail, and the way they're used or can be misused. Next gen data scientists understand the implications and consequences of the models they're building. They think about the feedback loops and potential gaming of their models.

Being an Ethical Data Scientist

You all are not just nerds sitting in the corner. You have increasingly important ethical questions to consider while you work.

We now have tons of data on market and human behavior. As data scientists, we bring not just a set of machine learning tools, but also our humanity, to interpret and find meaning in data and make ethical, data-driven decisions.

Keep in mind that the data generated by user behavior becomes the building blocks of data products, which simultaneously are used *by* users and *influence* user behavior. We see this in recommendation systems, ranking algorithms, friend suggestions, etc., and we will see it increasingly across sectors including education, finance, retail, and health. Things can go wrong with such feedback loops: keep the financial meltdown in mind as a cautionary example.

Much is made about predicting the future (see Nate Silver), predicting the present (see Hal Varian), and exploring causal relationships from observed data (the past; see Sinan Aral).

The next logical concept then is: models and algorithms are not only capable of predicting the future, but also of causing the future. That's what we can look forward to, in the best of cases, and what we should fear in the worst.

As an introduction to how to approach these issues ethically, let's start with Emanuel Derman's Hippocratic Oath of Modeling, which was made for financial modeling but fits perfectly into this framework:

- I will remember that I didn't make the world and that it doesn't satisfy my equations.
- Though I will use models boldly to estimate value, I will not be overly impressed by mathematics.
- I will never sacrifice reality for elegance without explaining why I have done so. Nor will I give the people who use my model false comfort about its accuracy. Instead, I will make explicit its assumptions and oversights.
- I understand that my work may have enormous effects on society and the economy, many of them beyond my comprehension.

Something that this oath does not take into consideration, but which you might have to as a data scientist, is the politics of working in industry. Even if you are honestly skeptical of your model, there's always the chance that it will be used the wrong way in spite of your warnings. So the Hippocratic Oath of Modeling is, unfortunately, insufficient in reality (but it's a good start!).

What Would a Next-Gen Data Scientist Do?

Next-gen data scientists don't let money blind them to the point that their models are used for unethical purposes. They seek out opportunities to solve problems of social value and they try to consider the consequences of their models.

Finally, there are ways to do good: volunteer to work on a long-term project (more than a hackathon weekend) with DataKind.

There are also ways to be transparent: Victoria Stodden is working on RunMyCode, which is all about making research open source and replicable.

We want step aside for a moment and let someone else highlight how important we think ethics—and vanquishing hubris—are to data science. Professor Matthew Jones, from Columbia's History Department, attended the course. He is an expert in the history of science, and he wrote up some of his thoughts based on the course. We've included them here as some very chewy food for thought.

Data & Hubris

In the wake of the 2012 presidential election, data people, those they love, and especially those who idealize them, exploded in schadenfreude about the many errors of the traditional punditocracy. Computational statistics and data analysis had vanquished prognostication based on older forms of intuition, gut instinct, long-term journalistic experience, and the decadent web of Washington insiders. The apparent success of the Obama team and others using quantitative prediction revealed that a new age in political analysis has been cemented. Older forms of "expertise," now with scare quotes, were invited to take a long overdue retirement and to permit a new data-driven political analysis to emerge.

It's a compelling tale, with an easy and attractive bifurcation of old and new forms of knowledge. Yet good data scientists have been far more reflective about the dangers of throwing away existing domain knowledge and its experts entirely.

Origin stories add legitimacy to hierarchies of expertise. Data mining has long had a popular, albeit somewhat apocryphal, origin story: the surprising discovery, using an association algorithm, that men who buy diapers tend often to buy beer at the same time in drug stores. Traditional marketing people, with their quant folk psychologies and intuitions about business, were heretofore to be vanquished before what the press probably still called an "electronic brain." The story follows a classic template. Probability and statistics from their origins in the European Enlightenment have long challenged traditional forms of expertise: the pricing of insurance and annuities using data rather than reflection of character of the applicant entailed the diminution and disappearance of older experts. In the book that introduced the much beloved (or dreaded) epsilons and deltas into real analysis, the great mathematician Augustin-Louis Cauchy blamed statisticians for the French Revolution: "Let us cultivate the mathematical sciences with ardor, without wanting to extend them beyond their domain; and let us not imagine that one can attack history with formulas, nor give sanction to morality through theories of algebra or the integral calculus."

These narratives fit nicely into the celebration of disruption so central to Silicon Valley libertarianism, Schumpeterian capitalism, and certain variants of tech journalism. However powerful in extirpating rent-seeking forms of political analysis and other disciplines, the dichotomy mistakes utterly the real skills and knowledge that appear often to give the data sciences the traction they have. The preceding chapters—dedicated to the means for cultivating the diverse capacities of the data scientist—make mincemeat of any facile dichotomy of the data expert and the traditional expert. *Doing data science* has put a tempering of hubris, especially algorithmic hubris, at the center of its technical training.

Obama's data team explained that much of their success came from taking the dangers of hubris rather seriously, indeed, in building a technical system premised on avoiding the dangers of overestimation, from the choice and turning of algorithms to the redundancy of the backend and network systems: "I think the Republicans f**ked up in the hubris department," Harper Reed explained to Atlantic writer Alexis Madrigal. "I know we had the best technology team I've ever worked with, but we didn't know if it would work. I was incredibly

confident it would work. I was betting a lot on it. We had time. We had resources. We had done what we thought would work, and it still could have broken. Something could have happened."

Debate over the value of "domain knowledge" has long polarized the data community. Much of the promise of unsupervised learning, after all, is overcoming a crippling dependence on our wonted categories of social and scientific analysis, as seen in one of many celebrations of the Obama analytics team. Daniel Wagner, the 29-year-old chief analytics officer, said:

> The notion of a campaign looking for groups such as "soccer moms" or "waitress moms" to convert is outdated. Campaigns can now pinpoint individual swing voters. White suburban women? They're not all the same. The Latino community is very diverse with very different interests. What the data permits you to do is to figure out that diversity.

In productive tension with this escape from deadening classifications, however, the movement to revalidate domain expertise within statistics seems about as old as formalized data mining.

In a now infamous *Wall Street Journal* article, Peggy Noonan mocked the job ad for the Obama analytics department: "It read like politics as done by Martians." The campaign was simply insufficiently human, with its war room both "high-tech and bloodless." Unmentioned went that the contemporaneous Romney ads read similarly.

Data science rests on algorithms but does not reduce to those algorithms. The use of those algorithms rests fundamentally on what sociologists of science call "tacit knowledge"—practical knowledge not easily reducible to articulated rules, or perhaps impossible to reduce to rules at all. Using algorithms well is fundamentally a very human endeavor—something not particularly algorithmic.

No warning to young data padawans is as central as the many dangers of overfitting, the taking of noise for signal in a given training set; or, alternatively, learning too much from a training set to generalize properly. Avoiding overfitting requires a reflective use of algorithms. Algorithms are enabling tools requiring us to reflect more, not less. In 1997 Peter Huber explained, "The problem, as I see it, is not one of replacing human ingenuity by machine intelligence, but one of assisting human ingenuity by all conceivable tools of computer science and artificial intelligence, in particular aiding with the improvisation of search tools and with keeping track of the progress of an analysis." The word 'improvisation' is just right in pointing to mastery of tools, contextual reasoning, and the virtue of avoiding rote activity.

The hubris one might have when using an algorithm must be tempered through a profound familiarity with that algorithm and its particular instantiation.

Reflection upon the splendors and miseries of existing models figured prominently in the Obama campaign's job ad mocked by Noonan:

- Develop and build statistical/predictive/machine learning models to assist in field, digital media, paid media and fundraising operations
- Assess the performance of previous models and determine when these models should be updated
- Design and execute experiments to test the applicability and validity of these models in the field […]

The facile, automatic application of models is simply not at issue here: criticism and evaluation are. No Martian unfamiliar with territory could do this: existing data—of all kind—is simply too vital to pass up.

How to learn to improvise? In other words, what model would be best for educating the data scientist? Becoming capable of reflective improvisation with algorithms and large data demands the valorization of the muddling through, the wrangling, the scraping, the munging of poorly organized, incomplete, likely inconclusive data. The best fitting model for the training required is not narrow vocational education, but—of all things—the liberal arts in their original meaning.

For centuries, an art, such as mathematics or music, was called "liberal" just because it wasn't automatic, mechanical, purely repetitive, habitual. A liberal arts education is one for free people, in the sense of people mentally free to reflect upon their tools, actions, and customs, people not controlled by those tools, people free, then, to use or not to use their tools. The point holds for algorithms as much as literature—regurgitation need not apply in the creation of a data scientist worth the name. Neither should any determinism about technology. No data scientist need ever confuse the possibility of using a technology with the necessity of using it. In the sparse space of the infinite things one might do with data, only a few dense (and interesting) ethical pockets deserve our energy.

— Matthew Jones

Career Advice

We're not short on advice for aspiring next-gen data scientists, especially if you've gotten all the way to this part of the book.

After all, lots of people ask us whether they should become data scientists, so we're pretty used to it. We often start out the advice session with two questions of our own.

1. **What are you optimizing for?**

   To answer that question, you need to know what you value. For example, you probably value money, because you need some minimum to live at the standard of living you want to, and you might even want a lot. This definitely rules out working on lots of cool projects that would be cool to have in the world but which nobody wants to pay for (but don't forget to look for grants for projects like those!). Maybe you value time with loved ones and friends—in that case you will want to rule out working at a startup where everyone works twelve hours a day and sleeps under their desks. Yes, places like that still totally exist.

   You might care about some combination of doing good in the world, personal fulfillment, and intellectual fulfillment. Be sure to weigh your options with respect to these individually. They're definitely not the same things.

   What are your goals? What do you want achieve? Are you interested in becoming famous, respected, or somehow specifically acknowledged? Probably your personal sweet spot is some weighted function of all of the above. Do you have any idea what the weights are?

2. **What constraints are you under?**

   There might be external factors, outside of your control, like you might need to live in certain areas with your family. Consider also money and time constaints, whether you need to think about vacation or maternity/paternity leave policies. Also, how easy would it be to sell yourself? Don't be painted into a corner, but consider how to promote the positive aspects of yourself: your education, your strengths and weaknesses, and the things you can or cannot change about yourself.

There are many possible solutions that optimize what you value and take into account the constraints you're under. From our perspective,

it's more about personal fit than what's the "best job" on the market. Different people want and need different things from their careers.

On the one hand, remember that whatever you decide to do is not permanent, so don't feel too anxious about it. You can always do something else later—people change jobs all the time. On the other hand, life is short, so always try to be moving in the right direction—optimize for what you care about and don't get stagnant.

Finally, if you feel your way of thinking or perspective is somehow different than what those around you are thinking, then embrace and explore that; you might be onto something.