Data Science = Coding + Simple Methods + Data Acumen

Patrick Perry
New York University

Envisioning the Data Science Discipline: The Undergraduate Perspective
The National Academies of Sciences, Engineering, and Medicine
December 12-13, 2016
Session #3: “What are the questions that should be asked to envision the future of data science for undergraduates?”

Why I Was Asked to Speak

- My job? (Statistics Professor)
- My education? (BS Math, MS EECS, PhD Statistics)
- My day-to-day activities? (Teaching, analyzing data, writing software, proving theorems, writing papers)
- My employer? (NYU Stern School of Business)

I’m not sure why I was asked to speak, some combination of the above.

Why You Should Listen to Me

I talked to some people in industry:

- Lukas Biewald, Chief Data Scientist, CrowdFlower
- Lee Dicker, Senior Research Scientist, Amazon
- Kyle Schmaus, Manager of Data Science, Stitch Fix
- Navin Sivanandam, Data Scientist, Likelihood
- Jon Sondag, Machine Learning Engineer, Spotify
- Anonymous, Software Engineer, Google
- Anonymous, Secretive Data-centric Organization

The reason you should listen to me, though, is that I talked to some people in industry about data science. This doesn't give me much credibility, but it does give me some.

Data Science Essentials

- Software engineering
- Linear algebra, optimization
- Basic statistics, linear and logistic regression
- Data Acumen (communication, perception, creativity)

The conclusion I've come to is simple: a data scientists needs a technical foundation in software engineering and basic methodology, and something else, which I'm calling "data acumen" — the ability to understand a problem and its broader context, come up with a creative data-driven solution, and communicate with non-technical others if necessary.

> “Having the same person devise and implement the method makes the whole process one thousand times more efficient.”
>
> –Lukas Biewald, Chief Data Scientist, CrowdFlower

How did I get to this conclusion? Let’s start with the technical foundation. There really is something effective about combining engineering skills with methodological skills that can distinguish data science from disciplines that focus on these skills individually.

Data Science Activities

organize
data and
computations

make decisions
using data

generate insights
from data

more
engineering

more
science

A data scientist moves along a spectrum of activities, some more focused on engineering, and others more focused on science. She doesn't need to be stellar at the engineering side or the science side, but she needs to be effective at both.

Coding: 80% or More

- Product infrastructure is typically not set up to support the data scientist. She may need to build tools to support herself.
- The data scientist is often tasked with "productionizing" analyses and predictions herself.
- A data scientist can spend over 80% of her job doing software engineering. (Source: NYTimes; This varies widely)

Regarding the engineering side, a data scientist can spend 80% or more of her time in this role. Some of this is "data cleaning"; other is setting up technical infrastructure to support herself. We should prepare data scientists for this part of the job by teaching them software engineering skills, possibly including data-specific software engineering ("tidy data", etc.).

> “If I had a MacGyver of data analysis and all he had was a t-test and regression, he would probably be able to do 99.9% of the analyses that we do that are actually useful.”
>
> –Jeff Hammerbacher, former head of Data team, Facebook, 2008

Regarding the methodology side, I’m going to take a somewhat controversial stance, and say that you do not need sophisticated methods to do good data science. The data scientists I spoke with overwhelmingly agreed that you can be effective with just basic statistics and regression. (I asked one data scientist what the most important method is, and he told me “arithmetic”.)

Methods: Simple

- You can do data science without using sophisticated machine learning
- Often training data is sparse, complex models are overkill.
- Latency and memory requirements constrain model complexity.
- Most important methods: linear and logistic regression.
- Need to understand the methods, so you need linear algebra, optimization, and basic statistics.

We hear a lot about “big data” solving all of our problems, but if you want, say, personalized recommendations, a recommendation for me for instance, then you need data from people who are similar to me. You probably won’t have much of this, so you won’t be able to fit a complex model. Even in situations where complex models (random forests, deep neural nets, etc.) do give better predictive performance, they are orders of magnitude slower than linear models, prohibitively slow for low-latency settings.

Data complexity and computational constraints push you towards linear models. If you understand linear models well (which requires knowledge of linear algebra, optimization, and basic statistics), then you can be an effective data scientist.

Uncertainty Quantification?

- Data scientists often deal with very subtle effects, they need tools for dealing with the uncertainty associated with these effects.
- Uncertainty quantification informs large-scale policy decisions (e.g. changing the deployed recommendation engine).
- Uncertainty quantification restrains “fishing expeditions”.
- However, uncertainty quantification is *not* used when making large-scale targeted predictions. (But could it be used?)

Statisticians often emphasize “uncertainty quantification” as their most important contribution to data science, so it’s important that I touch on this. By far the most common way a data scientist uses uncertainty quantification is testing for very subtle effects, typically via a two-sample t-test (an “A/B test”). They need to be able to design and interpret the results of these experiments, and similar analyses for other large-scale policy decisions. Beyond that data scientists tell me that they use uncertainty quantification when performing “fishing expeditions,” finding significant effects when searching through thousands or more candidates, typically using tools like False Discovery Rates. So, uncertainty quantification is important. However, uncertainty quantification does *not* get used when making large-scale targeted predictions. Are data scientists leaving money on the table? Possibly, but they are able to function effectively without this. All this is to say that uncertainty quantification is important, but you do not need very sophisticated statistics to perform the kind of uncertainty quantification that data scientists need.

"Creativity in applying simple methodology
is more effective than
proficiency with complex methodology."

–Kyle Schmaus, Manager of Data Science, Stitch Fix

Simple methodology, and the ability to implement it, by themselves, are not sufficient to make a good data scientist. She needs to be able to *apply* the methodology creatively.

Technology +
Methodology + ????

Technology and methodology are not enough by themselves, there's something else that makes good data science. This "something else" is important, and it can offset deficiencies in the other two. Many descriptions of data science acknowledge this, using terms like "domain knowledge" or "substantive expertise" to describe the missing part, but those terms don't really capture the spirit of the missing skill set.

Data Acumen:

The ability to understand a problem and its broader context, come up with a creative data-driven solution, and communicate with non-technical others if necessary.

Related terms: communication skills, creativity, data-analytic thinking, domain knowledge, empathy, formulation expertise, insight, product sense, substantive expertise

I'm going to call this third skill set "Data Acumen."

Data Acumen: Important

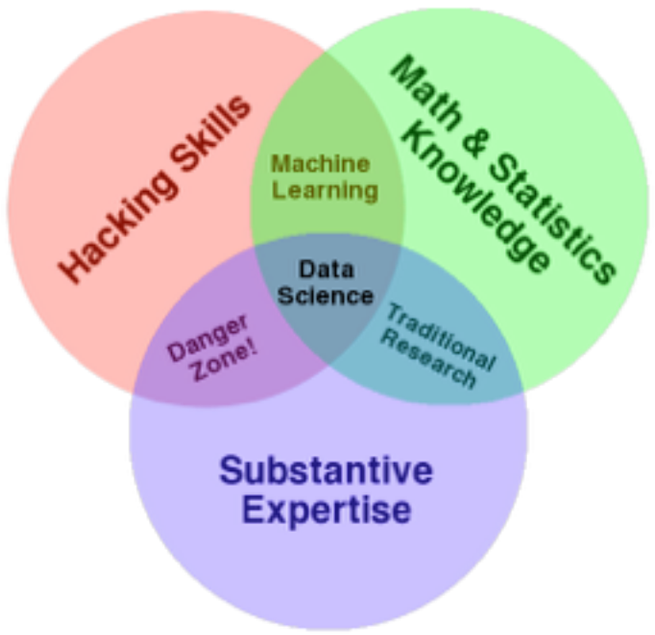
- Data science goals and objectives are often unclear or ill-defined.
- Data scientists must elicit priorities and “leads”, and they must validate their solutions.
- Communication with non-technical partners can deliver valuable intuition and domain knowledge.

Why is data acumen important? It’s not enough to have data analysis tools, and the ability to implement them. A data scientist needs to know how to apply her tools.

How to Teach Data Acumen?

- Partnerships with industry (projects, internships)
- Courses that showcase creative data analysis, including visualization
- Data-intensive courses taught by people in the natural and social sciences
- Focus on simple methods, complex data

The main the question we should be asking is “how can we teach data acumen?”. The cop-out is to say “we can’t” or “they’ll learn that part on the job.” Industry collaboration can help with this, but is not sufficient, certainly not if left for a “capstone” project at the end of the program. A culture of data acumen should pervade the program. My suggestion is to teach people data acumen by showing them examples of it, with courses that focus on creative data analysis, especially those taught by domain experts in the natural and social sciences. I don’t know if this is enough, and I am open to other ideas.



Drew Conway's *Data Science Venn Diagram (2013)*

This is misleading!

Data Science ≠
Machine Learning +
Substantive Expertise

Drew Conway's *Data Science Venn* Diagram has the right spirit, but it can easily be misinterpreted. It makes it look like to train a data scientist, you teach her Machine Learning, and then you teach her Substantive Expertise (or she learns it on the job). This is wrong. First, Machine Learning can be an elective, but it is not essential; second we should teach Data Acumen, not Substantive Expertise.

Data Science Essentials

- Technical foundation (software engineering, linear algebra, optimization, basic statistics)
- Simple models (linear and logistic regression)
- Data acumen (communication, perception, creativity)

    Key Challenge: How to teach data acumen?

The core of data science should focus on simple methods and proficiency in deploying them. The easy part is teaching the technology and the methodology. The challenge is teaching data acumen.