Data Science is a Process [DS4B Session 1c]

People often ask me how they can become a data scientist. To which my answers are usually 'why', 'what do you want to do with it' and 'let's talk about what it really is'. So let's talk about what it really is. There are many definitions of data science, e.g.:

- “A data scientist… excels at **analyzing data**, particularly large amounts of data, to help a business gain a competitive edge.”
- “The analysis of data using the **scientific method**”
- “A data scientist is an individual, organization or application that performs statistical analysis, data mining and retrieval processes on a large amount of data to **identify trends, figures and other relevant information**.”

We can spend hours debating which definition is ‘right’, or we could spend those hours looking at what data scientists do in practice, getting some tools and techniques under our belts and finding a definition that works for each one of us personally.

My own working definition of data science is “a process that helps people gain understanding through using data”. So let’s look at some of that process. The scientific method, mentioned above, is a process (O’Neill & Schutt, “Doing data science”):

- Ask a question
- Do background research
- Construct a hypothesis
- Test your hypothesis by doing an experiment
- Analyse your data and draw a conclusion
- Communicate your results

Most of science works like this: it’s all about creating explanations that fit our knowledge of the world, then testing those explanations with experiments. I like the cynicism embedded in this, the acknowledgement that everything is a working hypothesis that might turn out to be false, or false in different circumstances (see under Newton/Einstein), and those are all good things, but they don’t quite cover what data scientists do all day.

One of the process descriptions that data scientists use for themselves is the OSEMN process (Obtain-Scrub-Explore-Model-Interpret, pronounced ‘awesome’: your data is safe with data scientists, but not your acronyms…):

- Obtain datasets
- Clean, combine, transform data
- Explore the data
- Try models (classification, machine learning etc)

- Interpret and communicate your results

This is less about experiments, and more about the things that you need to do with data, but it loses what, to me, is the most important part: asking an interesting question. Data science isn't about data – it's about people, their problems and questions, and informing, persuading or entertaining them with your results (I'm with Sarah Cohen when she says "every good story starts with an idea, a question or an observation", and with anyone who says that a visualization isn't always the answer). So the process for these sessions, the thing we'll be working through slowly, is:

- Ask an interesting question
- Get the data
- Explore the data
- Model the data
- Communicate and visualize your results

with a healthy dose of cynicism, e.g. sanity-checking your results, in the context they're relevant to, which is especially important in a development data context, where data is hard to come by and may be erroneous, miss geographical areas or demographics and potentially be older than it looks.

Note that none of the quotes say "enormous amounts of data". We'll touch on big data in a later session (session 10), but most development data scientists work with small datasets, and that's nothing to be ashamed of: I'd rather have relevant, information-rich datasets than huge amounts of data that tells me almost nothing.

That process again

- **Ask an interesting question.** Write hypotheses that can be explored (Do people have more phones than toilets?, How is Ebola spreading? Is using wood fires sustainable in rural Tanzania? Can we feed 9 billion people?). Make them simple, actionable, and incremental (e.g. you can test different parts of the question separately).
- **Get the data.** There are many different data sources (e.g. datafiles, databases, APIs, text, maps, images, social media, people). Some of them are harder to get information out of than others, but they all contain data. Which means you'll often be extracting datasets from those sources (e.g. 80-page PDFs), and cleaning it. By cleaning, I mean getting the data into a shape that can be used by algorithms: dealing with file formats (pdfs!), badly-specified locations, human errors and differences between standards (e.g. "Tanzania" vs "Republic of Tanzania"). Although cleaning takes a lot of time, it's also time spent getting to know your datasets: what's in them, what's missing, what's strange, what potentially got lost in translation. Which leads us to:
- **Explore the data.** Once you've got the dataset in machine-readable form, you can start looking for more issues to deal with (e.g. these issues with different placename standards in

Tanzania) and, eventually, for potentially interesting patterns. Eyeballing (looking at) your data is usually a good place to start; often you have to take a subset of the whole dataset to do this, but it's usually worth it to get a better feel for what's contained in each dataset. Doing quick but ugly visualisations of your data is also a good thing to do.

- **Model the data.** Modelling is where we look for patterns and insights hidden in the data. It's where machine learning comes in. We'll look later at how to learn relationships between numbers, categories and graphs.
- **Communicate and visualize your results.** We want to get this effect: "I already knew that increased incarceration didn't lower crime, but I wasn't sure of the statistics. To see it on the graphs is really eye opening." (Pandey et al, The Persuasive Power of Data Visualisation), using whatever's appropriate (which might or might not be visualisations).

What we're aiming at is simple: "ask good questions, tell good stories" - if you can do this with data, you've won.

Data Scientists

Data scientists are the ~~mysterious beasts~~ people who do data science. The data science Venn diagram basically says that to be a data scientist, you need to know statistics, your business area and be able to code. But it's not quite like that. Although it's heresy to say this, many good data scientists don't code at all, and you can useful on a data science team without knowing everything, e.g. build insightful visualisations without using statistics, or specify a data science problem well enough for hardcore machine learning specialists to develop good algorithms for it.

That said, it does help a development data scientist to have expertise in development and statistics (these sessions were originally designed for people who had these skills: a statistics session is already in the pipeline…), and being familiar with data science skills and having the coding skills to get, clean and explore data will help you even if you never want to do anything more than create and be the 'client' for a data science problem specification.

At this point, you might be asking two questions:

- How do you become a data scientist?, and
- Should you become a data scientist?

You become a data scientist through learning and practice (that never stops: I'm still working on it myself). Yes, you need to learn a bunch of theory, but there's nothing like learning data science by doing it: you'll handle issues you didn't know existed, and learn many details about techniques by using them on non-sanitised (uncleaned) data. Good places to practice exploring and modelling data include:

- Kaggle - online datascience competitions

- Driven Data - social good datascience competitions
- Innocentive - some datascience challenges
- CrowdAnalytix - business datascience competitions
- TunedIt - scientific/industrial datascience challenges

Good places to practice asking good questions, getting data, communication and visualizing results include:

- Your own projects
- Data science for good groups (e.g. DataKind)

The answer to “should you become a data scientist” is “not necessarily”. There are lots of data science students desperate for good problems to work on, so you might want to become someone who can work **with** data scientists; which means learning how to specify data problems well. One of the places to see the work of these people who can specify problems (“problem owners”) is the competition sites listed above. The problem owner doesn’t have to do data modelling or machine learning themself, but they do need to be able to specify a problem well, find and clean data related to that problem so that competitors can access it easily (and all have the same starting dataset), and specify how the competition results will be marked (e.g. by accuracy on an unseen ‘test’ dataset). Go look at some of the problems listed on these sites, and think about how you would have done this yourself.

(session 1 slideset is here; cover image is from the Pump it Up challenge)