Data science; what is it and why is it important

Matthew Brett

The origins of data science

Facebook, LinkedIn

> *. . . what data scientists do is make discoveries while swimming in data . . . At ease in the digital realm, they are able to bring structure to large quantities of formless data and make analysis possible. . . . Data scientists' most basic, universal skill is the ability to write code.*

Data Scientist: The Sexiest Job of the 21st Century, 2012.

Scientists who were good with data

> *Some of the best and brightest data scientists are PhDs in esoteric fields like ecology and systems biology. George Roumeliotis, the head of a data science team at Intuit in Silicon Valley, holds a doctorate in astrophysics.*

Data Scientist: The Sexiest Job of the 21st Century, 2012.

What went wrong?

> *The now-contemplated field of Data Science amounts to a superset of the fields of statistics and machine learning which adds some technology for 'scaling up' to 'big data'. This chosen superset ["lesser data science"] is motivated by commercial rather than intellectual developments. Choosing in this way is likely to miss out on the really important intellectual event of the next fifty years.*

(Donoho 2015)

Why did it go wrong?

- The wrong people were designing the courses.
- They missed the big picture.
- The wrong people because:
  - They weren't data scientists themselves
  - They were not training data-focussed researchers.

How did it go wrong?

- Learn to code, then data science.
- Don't learn to understand code, just use it like a recipe.
- Data science as specialization rather than foundation.

Greater data science

- ". . . embracing a reinvention of statistical education in the era of pervasive computation." (2015 Berkeley faculty report).
- Teaching statistics "assuming computers exist, rather than assuming they don't exist." (John Denero).
- "Express in code what we would otherwise express in equations." (John Denero).
- "**Visualize First** - represent the data graphically in order to motivate questions about inference and concepts of statistics" (Eric Van Dusen).

Berkeley course and beyond

- Data science as foundation for statistics, data analysis and understanding.
- No requirements for foundation course.
- Massive expansion throughout university
- Widely adopted, adapted and emulated

To the demo