What is data science?

Matthew Brett

Thesis

- Data science is a fundamental change in scientific practice.
- Fundamental changes need careful, committed, expensive planning.
- Life Sciences must lead.
- Success or failure will have dramatic effects on our future as a University.
- We are far behind our US counterparts.

A (bad) definition

> 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'.

(Donoho 2015)

Initial responses

- "Data science" vocational training.
- Competition between computer science and statistics departments.

Bad definition, bold warning

> 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 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)

Back to beginnings


DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

(Davenport and Patil 2012)

The data scientist in industry

> When Jeff Hammerbacher and I talked about our data science teams, we realized that as our organizations grew, we both had to figure out what to call the people on our teams. "Business analyst" seemed too limiting. "Data analyst" was a contender, but we felt that title might limit what people could do. After all, many of the people on our teams had deep engineering expertise. "Research scientist" was a reasonable job title used by companies like Sun, HP, Xerox, Yahoo, and IBM.

(Patil 2011)

The data scientist in industry

> ... 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.

(Davenport and Patil 2012) - "Who Are These People?"

The data scientist in industry

> 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.

(Davenport and Patil 2012) - “Who Are These People?”

The data scientist in industry

> Roumeliotis was clear with us that he doesn’t hire on the basis of statistical or analytical capabilities. He begins his search for data scientists by asking candidates if they can develop prototypes in a mainstream programming language ...

(Davenport and Patil 2012) - “Who Are These People?”

The origins of data science

> Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.

(Tukey 1962)

Data science as foundation

> ... academic institutions should encourage the development of a basic understanding of data science in all undergraduates.

National Academies of Sciences and Medicine (2018)

> I think it likely that in ten years’ time every undergraduate programme will have to include some teaching in data science.

Professor Sir Adrian Smith, Director of Turing Institute, October 2019.

What does that look like?

Data science in Berkeley

- February 2013: Supporting Data Science Workshop
- 2013: Berkeley Institute of Data Science
- 2015: Foundations of data science course
- 2018: National workshop on data science education
- 2018: Division of data science: announced

Berkeley teaching programme

- Massive (~1500 student) course Foundations of data science - "Data 8". No requirements in mathematics or programming. Running since 2015.
- Large (~900 student) intermediate course Principles and techniques of data science with further requirements in Python programming and linear algebra.
- 27 Connector courses: domain applications of teaching methods from the foundation course.
- "... embracing a reinvention of statistical education in the era of pervasive computation." Report by Data science education rapid reaction team
- The greatest change in undergraduate teaching in a generation.

Spreading across campus

Computing in the UK

- understand what algorithms are, how they are implemented as programs on digital devices, and that programs execute by following precise and unambiguous instructions;
- create and debug simple programs;
- use logical reasoning to predict the behaviour of simple programs
- use technology purposefully to create, organise, store, manipulate and retrieve digital content

Coding is not a specialist skill

- understand what algorithms are, how they are implemented as programs on digital devices, and that programs execute by following precise and unambiguous instructions;
- create and debug simple programs;
- use logical reasoning to predict the behaviour of simple programs
- use technology purposefully to create, organise, store, manipulate and retrieve digital content

National curriculum in computing: Key stage 1 (5-7 year olds).

Some aspects of data science

- Foundation in code
- Real, messy data
- Visualization before inference
- Reproducible practice
- Open-everything.

Dominated by open tools

- Python
- R

Why is this?

Our problem

- We have no statistics department.
- Our strategy is not clear.
- We are losing our LES data scientists at an astonishing rate.

Some questions

- What is data science? A speciality? Or a transformation in scientific practice?
- What will a successful university look like in 10 years time? Will it look like Berkeley? Is there another model?
- What are the costs of failure?
- What are the benefits of success?
- What changes do we need to make in:
  - Research culture?
  - Academic appointments?
  - Training?
  - Teaching?
- Who will do the work?

References

Davenport, Thomas H, and DJ Patil. 2012. "Data Scientist: The Sexiest Job of the 21st Century." *Harvard Business Review* 90 (10): 70–76. https:

//hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century.

Donoho, David. 2015. "50 Years of Data Science." In *Princeton NJ, Tukey Centennial Workshop*. http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf.

National Academies of Sciences, Engineering, and Medicine. 2018. *Data Science for Undergraduates: Opportunities and Options*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25104.

Patil, DJ. 2011. "Building Data Science Teams." "O'Reilly Media, Inc.".

Tukey, John W. 1962. "The Future of Data Analysis." *The Annals of Mathematical Statistics* 33 (1): 1–67. http://projecteuclid.org/euclid.aoms/1177704711.