What is Data Science? Does It Matter?

Matthew Brett

Handout

Go to https://github.com/matthew-brett/bham-turing

Look for `bham_turing_th1_handout.pdf` in the `townhall-1` directory.

Thesis

- The UK is well behind the leaders in the US in its development of data science;
- To develop, we need to have a clear understanding of what data science means;
- If we are not to fall further behind, we need to invest our energy in reflection on the nature of data science, why it is important, and the changes it will bring to teaching and research.

About me

- https://www.turing.ac.uk/people/researchers/matthew-brett
- Analysis of functional brain images, since 1996;
- At UC Berkeley from 2003-2005, 2008-2017.
- Read more at What is data science?

A definition of data science ...

> 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)

. . . that would be a disastrous error

> 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'. . . .
>
> Choosing in this way is likely to miss out on the really important intellectual event of the next fifty years. – Donoho (2015)

Lesser and greater data science

- **Lesser data science**: superset of the fields of statistics and machine learning + technology for big data.
- ** Greater data science**: a radical change in approach to data analysis, founded on code, that emphasizes visualization, exploration and versatility.

Tukey's approach

> 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.

"The future of data analysis" (Tukey 1962)

Why it matters

We have to decide how to respond:

- Lesser data science:
  - Research: hire more machine learning experts
  - Teaching: expand advanced level courses in machine learning; add vocational courses for "data science" jobs.
- Greater data science:
  - Research: hire domain experts who have dealt with difficult data using computation. Emphasize collaboration across disciplines. Arrange university structures to support fluid collaboration.
  - Teaching: early undergraduate course in data analysis; transform teaching of data analysis across the curriculum.

Is Donoho right? Or can we live with Lesser?

- The origins of data science: Lesser or Greater?
- Developments in data science: Lesser or Greater?

History of "data science"


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

> ... 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?"

Who are these people?

> 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?"

Versatile

> ... on any given day, a team member could author a multistage processing pipeline in Python, design a hypothesis test, perform a

> regression analysis over data samples with R, design and implement an algorithm for some data-intensive product or service in Hadoop, or communicate the results of our analyses to other members of the organization.

Jeff Hammerbacher quoted in (Loukides 2010)

US developments in data science: teaching

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

National Academies of Sciences and Medicine (2018)

US developments in data science: research

- Careers
- Education and Training
- Tools and Software
- Reproducibility and Open Science
- Physical and Intellectual Space
- Data Science Studies

Moore-Sloan Data Science Environments: Themes. http://msdse.org/themes. ($37.8M).

Data science at UC Berkeley

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

Berkeley teaching programme

- Massive (~1500 student) course Foundations of data science. 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 using teaching methods from the foundation course.
- The greatest change in undergraduate teaching in a generation.

Recruiting for data science research

- Berkeley Division of data science.
- New recruitment half-time division of data science, half in their home department (of astronomy, psychology etc).

Conclusion

- "Lesser data science" was an early reaction from the universities to a much larger cultural change in data analysis.
- This change, maybe described as "Greater data science" is going to transform teaching and research.
- Is we want to lead in the UK, and internationally, we need to choose our vision.

The end

See: https://github.com/matthew-brett/bham-turing `townhall-1` for slides, handouts, source.

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.

Loukides, Mike. 2010. "What Is Data Science?" https://www.oreilly.com/ideas/what-is-data-science.

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.

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