What should a national data science strategy look like?

Matthew Brett

Context

The demand for skills in data science and AI across all sectors is growing rapidly but this demand cannot be met. As a national institute The Turing has a key role to play in the national skills agenda and it is part of our mission to train new generations.

We are now looking for a highly motivated and inspirational leader to position the Institute within this landscape by developing and driving an ambitious skills strategy.

Some definitions

- Data science - ?
- Artificial intelligence.

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

Initial response

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)

This response is unlikely to be fruitful

> 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)

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)

Who are these people?

> ... 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?"

Life scientist - data scientist

> 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?"

What does this definition mean for education?

> ... 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 this look like in practice?

Berkeley data science initiatives:

- 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.

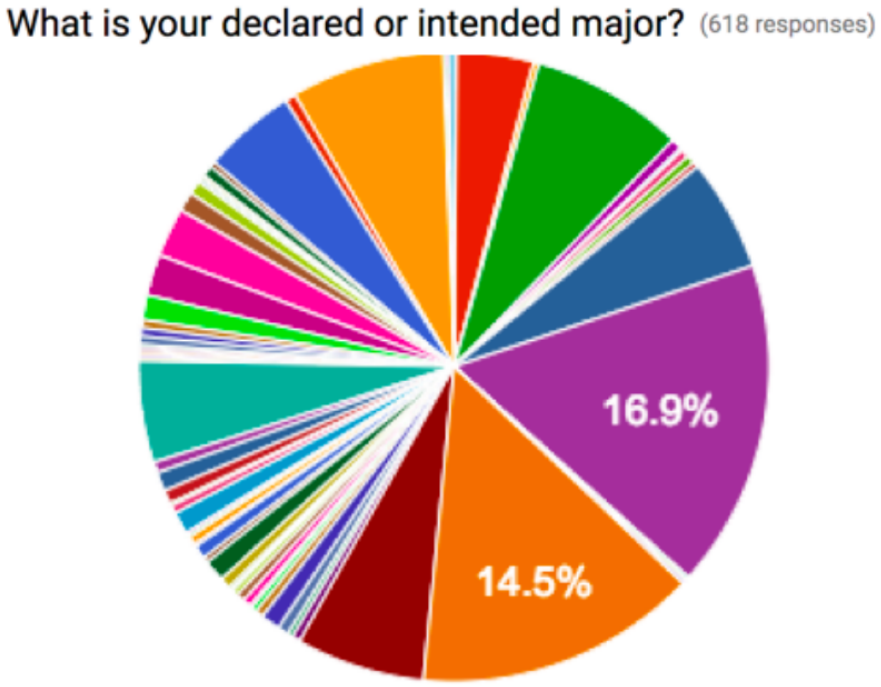
Principles of the introductory course

- Teaching statistics "assuming computers exist, rather than assuming they don't exist."
- "Express in code what we would otherwise express in equations."

John DeNero, 2018 Webinar

It's huge

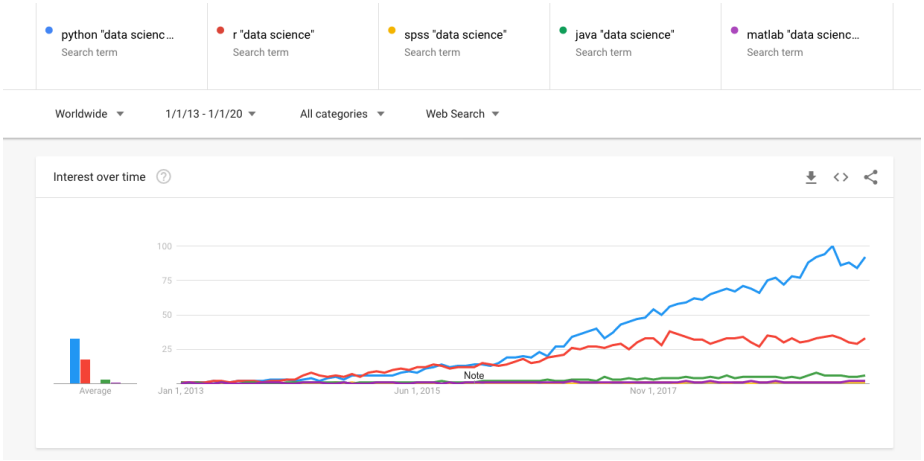
Students, Spring 2017



Wide range of majors, $> 14\%$ slices are economics, computer science – John DeNero, 2018 Webinar materials

Spreading across campus

On technology



Data science tools over time.

Future landscape in 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;

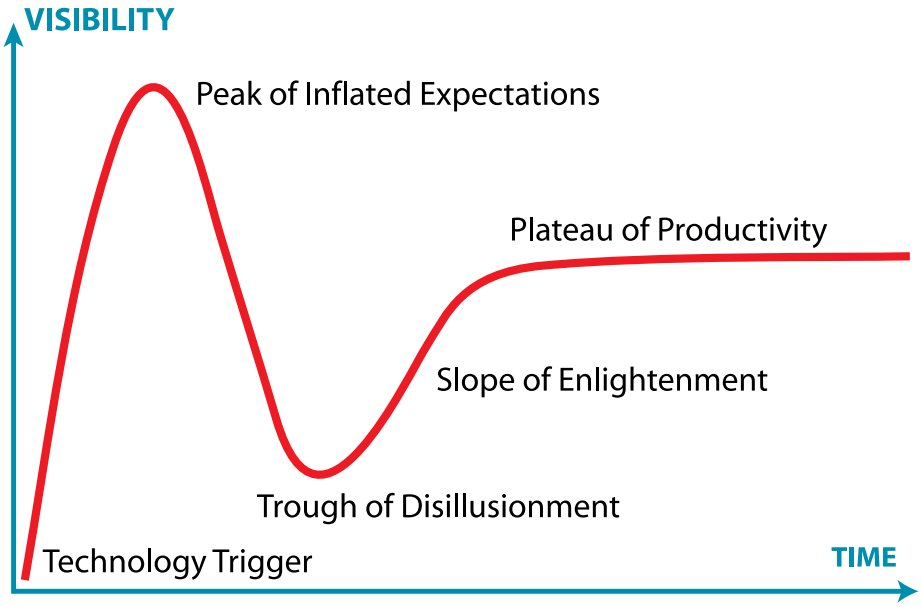
- use logical reasoning to predict the behaviour of simple programs
- use technology purposefully to create, organise, store, manipulate and retrieve digital content

National curriculum in computing from 2014: Key stage 1 (5-7 year olds).

Problems for a national strategy

- Little experience of undergraduate or post-graduate education
- Weak culture of open-source contribution.
- Little room for substantial cross-discipline courses within most degrees.
- Emphasis on grant money and REF metrics tends to discourage cross-discipline collaboration.

Data science and the hype cycle



Data science and the hype cycle. Graphic by Jeremykemp at English Wikipedia CC BY-SA 3.0

Options?

- Attempt to enforce a national curriculum of data science in universities?
  - Who says what that curriculum is?
- Build collaborative teaching programme.

  - The curriculum becomes the programme that universities want to use.

Collaborative teaching programme

- Online, editable textbooks:
  - https://www.inferentialthinking.com (but, license)
  - https://matthew-brett.github.io/cfd2019
- Python and R
- PDF / printed textbook: https://resampling-stats.github.io/resampling-with/
- Training resources for teachers
  - Workshops
  - Online
- Private question / homework bank for teachers.
- Rapid iteration with students.
- Your idea here.

The end

Materials at https://github.com/matthew-brett/ds-nat-teaching.

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.".