Better Data Science, Better Decisions

Isaac Newton Institute, Cambridge
14th November, 2018

David Johnson
Deputy Director, Knowledge Exchange
ONS Data Science Campus
@new_to_dave
Technology is easy
Culture is hard
“Although better use of [data] has the potential to transform the provision of economic statistics, ONS will need to build up its capability to handle such data.

This will take some time and will require not only recruitment of a cadre of data scientists but also active learning and experimentation.

That can be facilitated through collaboration with relevant partners – in academia, the private and public sectors, and internationally.”

Independent Review of Economic Statistics
Professor Sir Charles Bean, 2016, p.11
Building the Data Science Campus

- Funding approved in Mar 2016
- Start-up team of 2 in place in Jul 2016
- Temporary Campus open Aug 2016
- Research commenced in Sep 2016
- 1st Apprentices in Nov 2016
- 1st research output in Dec 2016
- Managing Director joined in Jan 2017
- Headcount reached 26 in Feb 2017
- Formal launch 27 March 2017
- Move to bespoke Campus in May 2017
- 55 FTE as of Nov 2018
- Headcount to reach 65 by March 2019
Short exploratory projects
• Improve our understanding of the UK’s economy, communities and people
• Up to 6 months in length, high tolerance of risk

Does it add value?
• Does it increase our understanding of the UK’s economy and society?
• What will we learn…
  - about a data source or technique?
  - building Data Science Campus capability?
  - building ONS and Government capability
• Can the learning be applied to other problems?

Working in partnership
• ONS and Government partners bring the challenges
• Academia and Industry partners provide technical and sectoral expertise
• Campus delivers project outputs and enables knowledge transfer
Why Data Science?

“Data scientists solve complex business problems using a combination of domain expertise, coding knowledge, machine learning and statistics skills on large and varied datasets.”

Government Data Science Partnership
“The 21st Century has brought new challenges in the analysis of data, and it is increasingly apparent that solutions to these are both statistical and computational. This has led to a great demand for people both in industry and in research who are able to draw upon the mathematics of both computation and probability to make sense of the large amounts of data that are collected in order to solve major problems.

Data science is an interdisciplinary response to this demand”

- University of Warwick
“The 21st Century has brought new challenges in the analysis of data, and it is increasingly apparent that solutions to these are both statistical and computational. This has led to a great demand for people both in industry and in research who are able to draw upon the mathematics of both computation and probability to make sense of the large amounts of data that are collected in order to solve major problems.

Data science is an interdisciplinary response to this demand”

- University of Warwick
“The 21st Century has brought new challenges in the analysis of data, and it is increasingly apparent that solutions to these are both statistical and computational. This has led to a great demand for people both in industry and in research who are able to draw upon the mathematics of both computation and probability to make sense of the large amounts of data that are collected in order to solve major problems.

Data science is an interdisciplinary response to this demand”

- University of Warwick
“The 21st Century has brought new challenges in the analysis of data, and it is increasingly apparent that solutions to these are both statistical and computational. This has led to a great demand for people both in industry and in research who are able to draw upon the mathematics of both computation and probability to make sense of the large amounts of data that are collected in order to solve major problems.

Data science is an interdisciplinary response to this demand”

- University of Warwick
Data Science in Government
Why Data Science in government?

• “Getting data right is the next phase of public service reform”

• Deliver more insight from the data we hold

• Drive more insight through use of new data sources

John Manzoni – Chief Executive of the UK Civil Service
“The public recognise] “the potential uses and relevance that the large volumes of data we now generate and collect have for public services.”

“Opportunities for data science were seen as greatest where a clear wider public benefit could be established and where the risk of taking no action was significant.”

“The most obvious opportunities for data science emerged where there was a clear remit for government, where the current status quo was seen as inadequate, where the outcome was appropriate, and where data science could complement other methods.”

*Public dialogue on the ethics of data science in government, (2016), Ipsos MORI, London p3-4*
Predicting Viral Outbreaks

Fig 1. Wordcloud of Norovirus Keywords

Fig 2. Norovirus Keywords in Tweets compared to reported incidents

2.8m
Cases of Norovirus per year in the UK

£120m
Estimated cost to the country in lost working hours due to Norovirus

£20k
Total cost of the project, including publicity campaign

Source: FSA
Mapping Tourist Travel Patterns

72.8m
Visits overseas by UK residents in 2017, an increase of 3% over 2016

39.2m
Visits by overseas residents to the UK in 2017, 4% more than in 2016

800k
International Passenger Survey interviews conducted per year at UK airports and ports

Travel Trends 2017, ONS, 20th July 2018
Early Indicators of GDP

Fig 1. UK GDP Growth Rate

-6%
Change in UK GDP between first quarter of 2008 and second quarter of 2009

5 years
Length of time from 2008 for the UK economy to return to pre-recession size

£12b
Estimated value for earlier identification of 2008 downturn

Early Indicators of GDP

Change in UK GDP between first quarter of 2008 and second quarter of 2009

-6%

Length of time from 2008 for the UK economy to return to pre-recession size

5 years

Estimated value for earlier identification of 2008 downturn

£12b

Ferry operators collect information on the contents of lorries and trade vehicles boarding their Ferries.

A single line description is recorded to detail the contents.

The data collection is not controlled enabling complete free text entries.

This significantly restricts the analysis that can be done.

Optimus is a pipeline that can group items from free-text lists by context that do not have accompanying classifications or codes.

The tool can generate labels for groups of items based on common syntax or, in some cases, synonyms. It can also handle inconsistencies in text records such as spelling mistakes, plurality and other syntactic variation.

35k
Lorry journeys in single month analysed during Phase 1

450k
Lorry journeys in 2017 to be analysed during Phase 2
Data + Technology = Data Transformation
## Data Sharing

### Why it's hard

<table>
<thead>
<tr>
<th>Technology</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bespoke, siloed IT systems</td>
<td>Data in hard-to-use formats like PDF</td>
</tr>
<tr>
<td>Legacy IT that makes data hard to extract</td>
<td>Data inconsistently entered</td>
</tr>
<tr>
<td>Outsourced IT providers charging for data access</td>
<td>Use of different standards</td>
</tr>
<tr>
<td>Lack of common platform for data sharing</td>
<td>Lack of common identifiers</td>
</tr>
<tr>
<td></td>
<td>Lack of open data</td>
</tr>
</tbody>
</table>

### How to make it easier

<table>
<thead>
<tr>
<th>Technology</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use tech conforming to common standards for interoperability</td>
<td>Record all data in machine-readable format</td>
</tr>
<tr>
<td>Insist all IT has open APIs</td>
<td>Enforce consistent data entry</td>
</tr>
<tr>
<td>Ensure contracts give full access to data</td>
<td>Use common standards</td>
</tr>
<tr>
<td>Invest in common platform for data sharing</td>
<td>Use unique IDs, e.g. UPRNs</td>
</tr>
<tr>
<td></td>
<td>Release non-personal data openly by default</td>
</tr>
</tbody>
</table>

### Legal

<table>
<thead>
<tr>
<th>Legal</th>
<th>Organisation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk averse leadership</td>
<td>Teams created to focus on their siloed remit</td>
</tr>
<tr>
<td>Staff unsure about data protection rules</td>
<td>Lack of dedicated time and resources for data collaboration</td>
</tr>
<tr>
<td>Most senior data professional in org is Data Protection Officer</td>
<td>Leaders lack understanding of role and means of using data</td>
</tr>
<tr>
<td>Lack of template data sharing agreements</td>
<td></td>
</tr>
</tbody>
</table>

### Organisation

<table>
<thead>
<tr>
<th>Legal</th>
<th>Organisation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train all staff in Privacy Impact Assessments</td>
<td>Establish Offices of Data Analytics</td>
</tr>
<tr>
<td>Appoint Chief Data Officer tasked with responsible data sharing</td>
<td>Free up time of in-house analysts to work on data science projects rather than KPI reporting</td>
</tr>
<tr>
<td>Make use of template Information Sharing Protocols</td>
<td>Leaders insist on using data to inform decisions</td>
</tr>
</tbody>
</table>

@EddieACopeland #IdeaOnAPage @nesta_uk | http://bit.ly/IdeaOnAPage
“Too often people with too little technical understanding are buying IT from suppliers who want to capture the buyer so they can’t ever go elsewhere. Suppliers’ interests are too often not aligned with the NHS’s interests, and the contracts badly managed. This supplier capture is common in IT, but it is not inevitable.

So we are seriously going to increase the in-house capacity to understand the technology, to procure the right things, to manage them better, to split up big contracts into smaller pieces, to ensure an agile, iterative approach focused on the question: what is the user need?

This takes a culture change. Out with the big service contracts. In with more agile in-house teams that can be smarter at contracting. I don’t want to see any more the automatic knee-jerk response to an IT problem of engaging big consultancies to tell us what the problem is, offer to sell us the solution, and then mark their own homework.

Instead we need to build in-house capacity to lead design and delivery, and this will also enable good providers to work more effectively with the NHS, deliver more successfully and operate with less risk.”

Secretary of State for Health and Social Care Matt Hancock
NHS Expo 2018, 6th September 2018
“Data Scientists identify complex business problems whilst leveraging data value. They work as part of a multidisciplinary team with Data Architects, Data Engineers, Analysts and others. They work with policy and operations teams to understand where data science can add value. They support strategic and operational decision making in order to create impact.

They source, access, manipulate and engineer data processes with data that typically have characteristics of volume, velocity and/or variety. They build credible statistical models from the data and use best coding practices to generate reproducible work. They may draw on other technical and analytical standards from across government and industry.

They adhere to the Data Science Ethics Framework. They are open minded and demonstrate strong intellectual curiosity. They have an interdisciplinary focus, using techniques and knowledge from across the scientific spectrum.”

Data Scientist Role Description, Digital, Data and Technology Profession

Chin-Rui Tan, Head of Data Science at the Foreign and Commonwealth Office London, 2017
Growing Data Science Skills

Level 4 Apprenticeship in Data Analytics:
Two-year programme, 12 months at the Campus followed by 2 x 6 month rotations across ONS

Level 6 Apprenticeship in Data Science:
Three-year programme, approved for England in 2018, launching in England and Wales in 2019

Data Science Faculty:
In-house training unit delivering short courses in programming (R, Python) and fundamentals of Machine Learning, NLP etc

Accelerator:
12-week mentoring programme for Government analysts delivered with GDS, in parallel with in-house ONS Data Science Academy

Masters in Data Analytics for Government:
2 year part-time MSC in data science and statistics for government analysts, delivered by UCL, Oxford Brookes and University of Southampton. Expanding in 2020/21
Growing Data Science Skills

**University research programmes and PhD funding:**
Research partnerships with over a dozen universities and the Alan Turing Institute, which additionally hosts the *Data Science for Public Good* joint PhD programme

**Centres for Doctoral Training:**
13 UKRI and EPSRC bids supported for centres in AI, Data Science and Statistics, launching in 2019, with PhD co-supervision and Campus placements

**Graduate Student Placement Programme:**
Three-month MSc thesis and industry placements at the Campus

**Welsh Data Science Graduate Programme:**
2 year MSc programme with Welsh Contact Centre Forum that includes three eight month industry placements.

**STEM Outreach:**
Girls into STEM (Science, Technology, Engineering & Maths) programme for GCSE students; Nuffield Research Placements for A-level students
Partnerships

Working with Industry and Academia to enable data transformation

Joint Research Programmes range from specific projects with individual universities (WBS Data Science Lab) to the Turing-HSBC-ONS Economic Data Science Awards

Data Science for Public Good PhD Programme at the Alan Turing Institute launches in Sep with four co-funded PhD students, and a 6 month placement at DSC

Sunsetting PhD Co-funding Programmes in place at Cardiff and Lancaster

Graduate Student Placement Programme at the DSC to undertake MSC thesis work (Southampton and Cardiff Met), with PhD placements launching in Sep (UCL)

Centres for Doctoral Training will become a strategic focus from 2019/20. The DSC is currently supporting 13 UKRI and EPSRC bids in both AI and Statistics

Industry Collaboration Programmes focused on public good outputs with key partners (PWC, Barclaycard) include data sharing, secondments, apprentice development, knowledge sharing events etc

Skills and Standards collaborations with industry working groups led by Royal Statistical Society, the Royal Society, The Alan Turing Institute, and the ONS-led Level 6 Apprenticeship Trailblazer group

Government Data Science Partnership with the Government Digital Service and the Government Office for Science drives growth of data science across government through training (Accelerator), knowledge sharing events (Government Data Science Conference, Community of Interest) and communications (Slack)

Joint Training Programmes with GSS Learning Academy and GDS Academy

Direct Collaborations with individual government departments to provide training and undertake joint research

Data Enabled Change Accelerator Programme delivered with the Department for Digital, Culture, Media and Sport under John Manzoni’s Data Advisory Board provides DSC mentoring for strategic data science projects

Research Partnerships with international centres of data science excellence including the Centre for Big Data Statistics at Statistics Netherlands, and the United Nations Global Pulse Lab in Jakarta

Capacity Building Programmes that support the development of public sector data science and data analytics skills in less developed countries, with a flagship programme with National Institute of Statistics Rwanda

Learn: Act as a conduit for new data sets, technologies and methodologies to enter ONS
Grow: Support the growth of data science across the UK and internationally
Train: Develop the next generation of data scientists and analysts in Government
Recruit: Promote ONS and Government as first destination of choice for graduates
Technology is easy
Culture is hard
Technology is easy
Culture is hard

(but technology is also hard)
ONS Data Science Campus

Email: datasciencecampus@ons.gov.uk
Web: datasciencecampus.ons.gov.uk
GitHub: datasciencecampus.github.io
Twitter: @DataSciCampus

Image credits: ONS stock: (Dave Johnson); FT p3; TFL pp9.10; GDS pp11-13,22,23; FSA p14; WBS p15; Gerry Cranham / Fox Photos / Hulton Archive / Getty Images p8; Eddie Copeland/Nesta p20; drawnalism p28