Teaching big data skills in the social sciences

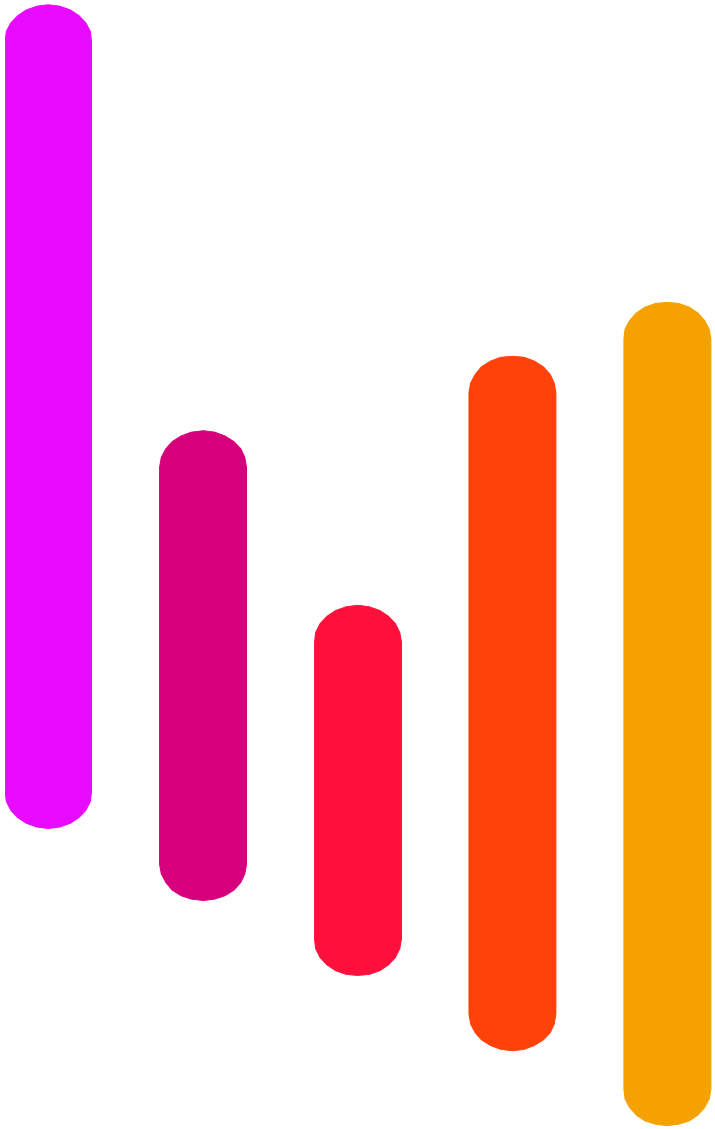
Dr Sarah King-Hele
UK Data Service
University of Manchester

IASSIST Conference 2017
Lawrence, Kansas
25 May 2017

UK Data Service

Overview

- UK Data Service big data skills training
  - Hadoop: Hive and Spark
  - Getting internet-based data
  - Basic IT skills
  - Appreciation of quality of big data
  - Appreciation of ethical issues
- Some reflections



UK Data Service

- a comprehensive resource funded

  by the ESRC in the UK

- a single point of access to a wide range of secondary social science data

- support, training and guidance

- Data types

  - survey microdata

  - UK census data

  - qualitative and mixed methods data

Site search FAQ Help Contact New users

UK Data Service

About us | Get data | Use data | Manage data | Deposit data | News and events

Register / Login

Explore the UK's largest collection of social, economic and population data resources.

Search data

About the UK Data Service

Guides and resources

- Dataset guides
- Topic guides
- Methods and software guides
- Guides to exploring online

See more

Video tutorials

See our growing range of training videos

See data from all over the world

Browse our data map

Data types

- Census data
- International macrodata
- Longitudinal studies
- Qualitative/mixed methods
- UK surveys

Professor Michaela Benzeval

The Director of Understanding Society discusses the impact of making data available through the UK Data Service

Featured data

Encouraging healthy lifestyles in young people

What is the link between

Latest data

- Small Business Survey, 2014
- Second Longitudinal Study of Young People in England: Wave 1, 2013: Safe Room Access
- Second Longitudinal Study of Young People in England: Wave 1, 2013: Secure Access

ukdataservice.ac.uk

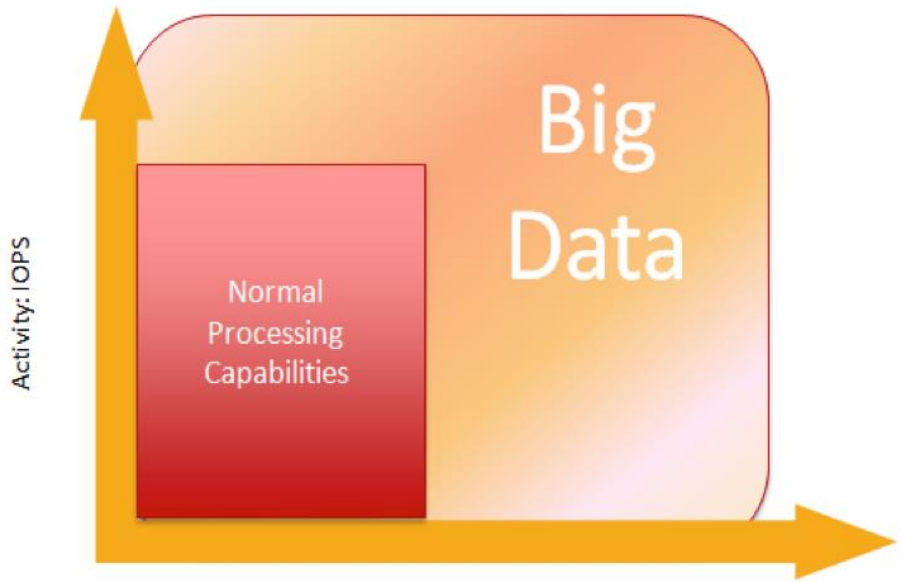
Background: UK Data Service big data projects

- Development projects: to prepare the UKDS for big data
  - Upgrade UKDS IT for storing and curating data and metadata
  - Provide big data training; upskilling for our user base
  - Links with others using big data, researchers and trainers e.g. big data centres

- Training team:
  - Manager - Sarah King-Hele (social survey research)
  - Trainer – Peter Smyth (IT and big data)
  - Other trainers: Chris Park, Libby Bishop, Louise Corti and Nathan Cunningham

What are ‘big data’?

Data sets that exceed the boundaries and sizes of normal processing capabilities, forcing you to take a non-traditional approach



New sources of data for social sciences research

due to technological change:

e.g.

- Social media data
- Smartmeter data – sensor data
- Data from businesses
- Administrative data

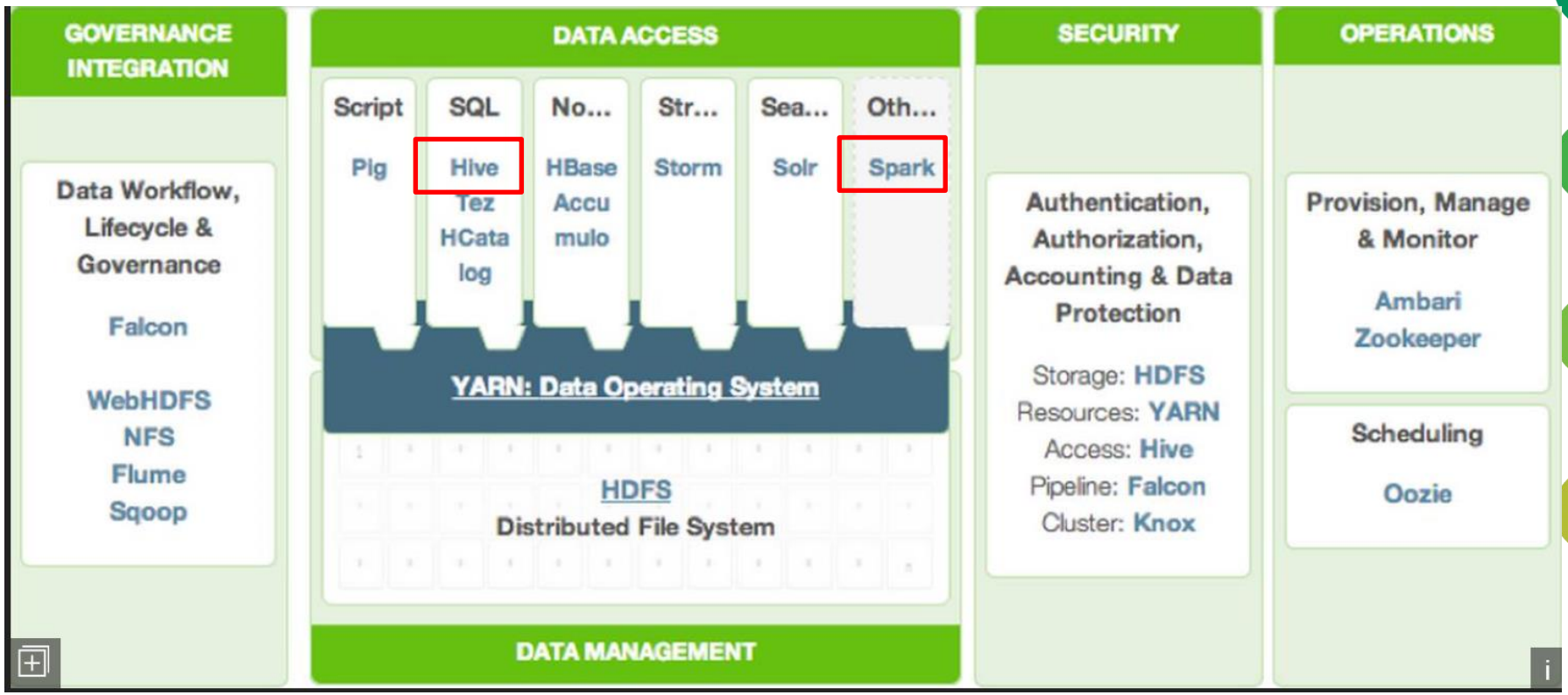
St Peter's Square, Rome

What big data skills should we teach?

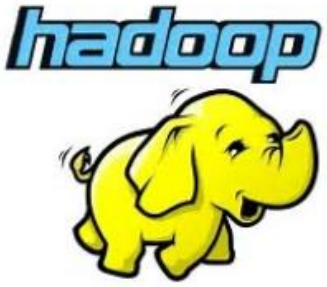
First thoughts:

- Hadoop ecosystem
- And obtaining data from the internet

Training in Hadoop tools: packages for storing and manipulating data



Hadoop and related training



Webinars: 45 minute introductory talks

- Basic understanding of
  - what Hadoop/Hive/Spark are and what they can be used for
  - what the interfaces look like
  - what users need to get started

- What is Hadoop?
- What is Hive?
- What is Spark?

Workshop: Introduction to big data manipulation using Hive

- How to use Hive in practice

Getting data from the internet training

Webinars: 45 minute introductory talks

- Basic understanding of what APIs are and how to use them

- What are APIs?
- What is MongoDB

More recently: Workshop: Getting internet-based data

What we learned:

- Very popular (though all courses were free)
  - Webinars: over 200 registered
- Some Hive workshop attendees had used SQL and others never used a database
- Gap in basic IT knowledge between average social sciences researchers and skills needed to use Hadoop and APIs

Basic IT skills training…

Webinars:

- Introduction to databases
- Introduction to ODBC

Workshops:

- Introduction to programming

Introduction to programming workshop

Aim: teach basics *principles* of programming (in any language)

- What is a computer program?
- Defining the problem in words and pictures: flow diagrams and pseudocode
- Python programming constructs
- Documentation
- Debugging
- Introduction to testing
- Re-using code and packages
- Writing small programs

Very popular but wide variety of expectations and knowledge

Other ‘big data’ training:

Practical ethics for big data research: An introduction (webinar)

October 2016

Libby Bishop (UK Data Service, University of Essex)

Assessing quality in big data (course)

Louise Corti (UK Data Service, University of Essex)

Summer school, Cape Town, Jan/Feb 2017

- 5 day programme aimed at skilled statisticians for scaling up handing and analysis of larger web based data
- Funded from grant between UK and South Africa Research Council
- Introduction to tools (via Hadoop Sandbox) and final group project
  - Big data in social sciences research, ethics, Hive, Spark

UK Data Service

Feedback…

- *”Great introduction to whole field, gentle ease in to more technical concepts”*
- *“Knowing what the other participants are working on is very helpful – a round robin type introduction saves a great deal of time during networking breaks as you can gravitate quickly towards those you want to follow up with. Really enjoyed the exercises – learning happens in action!”*
- *“The practicals on how to navigate Hive were interesting and mind opening”*
- *“I could easily follow the hdfs command lines and the effectiveness of Hadoop and think this software is going to be useful to my studies”*

Some reflections…

Huge interest in new packages and basic computing skills

- but

- attendees have diverse backgrounds
- diverse needs and interests

Should we teach single packages? Which?

- Multiple packages to do whole task?

Multiple modes of teaching works well

- Webinars – larger numbers, free, quick but limited
- Workshop – smaller numbers, in depth

More reflections…

Computing facilities can be problematic

- University clusters
- UKDS laptops for Hadoop training and summer schools

- UKDS platform for big data in development
  - teach how to use some instances using clear exemplars
  - where do we fit in with other training providers?

More reflections…

- Other trainers keen to collaborate

Cathie Marsh Institute for Social Research

Short courses

MANCHESTER 1824
The University of Manchester

NCRM National Centre for Research Methods

Methods@Manchester

DATA SCIENCE INSTITUTE
Gateway to The University of Manchester's data science activity

DATA CARPENTRY
Building communities teaching universal data literacy

- More collaboration needed across disciplines

Final comments

- Still early days for ‘big data’ in the social sciences

Issues:

- What data will be available and accessible to researchers?
- What packages and analysis methods will become standard?
- Will we need cross-disciplinary teams to enable richer scientific analysis?
- Elementary schools are teaching programming and children grow up with technology
  - what will their needs be in 10-15 years time?

Some lessons we learned…

- Great interest in learning about how to access, manipulate and analyse new forms of data
- Wide range of new data and skills: can't learn everything
- Linking data e.g. geo-enabled is a major use case
- Appreciation of the 'shape' of data from other disciplines
  - Knowledge exchange sessions disciplines, e.g biomedical and environmental science with social sciences
- Better basic IT skills are essential

Further information and acknowledgement

Encounters with big data: UK Data Service blog post

http://blog.ukdataservice.ac.uk/reflections-on-encounters-with-big-data-our-course-in-cape-town/

Webinar recordings:

https://www.ukdataservice.ac.uk/news-and-events/webinars

UK Data Service big data pages:

https://bigdata.ukdataservice.ac.uk/

Thanks to Nathan Cunningham for the use of his slide *What are big data?* and the images in *New sources of data for social sciences research*

Questions

Dr Sarah King-Hele

sarah.king-hele@manchester.ac.uk

ukdataservice.ac.uk/help/

Follow us on Twitter

https://twitter.com/UKDataService

or Facebook

https://www.facebook.com/UKDataService

Subscribe to the UK Data Service news list at

https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=UKDATASERVICE