Big data skills in the social sciences - Interactive session notes

University of Manchester, 14 October 2016

Sarah King-Hele, CMI, University of Manchester

1. Question: What big data training would you like to see/what are the barriers to doing big data research around the following topics:

a) Data resources
b) Programming and software
c) Methods
d) Other?

The 30 participants put post-it notes on posters under each topic and a summary of their notes and subsequent discussion is given below.

2. Summary of notes and discussion

a) Data resources

Awareness of data for research and for teaching and examples of use cases

- There is a lack of awareness of what data is available – need a database of data? What are available? What are they good for? How can I use them? Overviews of what's available. Knowing what data is available
- As a teacher it would help to have recommendations of good datasets to use in training
- From a qualitative researcher: What was the most useful in today's session were the examples of projects that have been previously done with big data. I think more examples + more comprehensive overview of types of data available + how to get it is what would be most useful for me to get on board with big data.

Access and IP

- Access to data sources based on countries outside the UK
- Is there a way to access local authority data for the whole UK in one (or 2) place? Freedom of information request problem. If you get it, can you archive it?
- How to access resources, especially access to govt + local govt data relating to public services
- Negotiating access to data: BDN2 is addressing this in relation to local government and business data via the 3 data research centres (BLG at University of Essex, CDRC at University of Leeds and UCL, UBDC at University of Glasgow)
- IP

Big data infrastructure

- Research data secure space at University of Manchester is needed. A closed space where you could use secure data. There is a call out for ESRC 'pods' secure spaces at universities

which could be used for that would be limited in numbers. BLG at the University of Essex has such a space.
- Lack of infrastructure for big data research e.g. Hadoop systems, people to offer support. Haphazard in UK. Some areas of strength but not generally.

b) Software and programming

Programming

- Basic programming for research; Programming to get the data + also data links + matching
- Are there examples/ models of what a programming course would consist of at a UG social science level?
- Training in R. Introduction to programming skills for social science researchers
- Need to know more about the software available and what it is used for. Know more about programming – what courses are available?
- Knowledge analysis. Software skills need updating!

Data management

- Using Python for data management; perhaps data cleaning/prep/merging of secondary data (messy routinely recorded data maybe)
- Data storage / data management for big datasets
- Database management (e.g. SQL?) – aggregating /collating data from various formats into a unified format to be imported into mainstream statistical packages (for analysis).

Support for training, teaching and software

- Having a service to assist in a personal way to specific needs in software training
- IT support to set up teaching clusters so they are suitable for teaching with this kind of data… including software
- Free software? (or Univ to get)

Other

- NVivo training has been defined by the trainer but not from researcher's needs

c) Methods

Data accuracy: With so much data being generated, how much is being done around accuracy? Especially as it's dynamic and constantly changing. Example of how missing data in London Underground travel cards would lead to the conclusion that lots of people are disappearing in the underground every year if it were taken seriously.

Relating research questions and methods

Provide with training according to the moment – learning the method – applying method in your research – interpreting results

Appropriate statistical techniques e.g. dealing with missing data, problems with data mining

So much of the data analysis we learn is about inference – these new forms of data challenge that preoccupation… where you don't have probability sampling

Need to know more about what's out there in order to be clear about what methods might be most appropriate

Practical introduction to missing data imputation – what are the limitations and benefits? Model checking – steps of checking model assumptions after running parametric / non-parametric tests of association

Social Network Analysis SNA

Sampling issues when using social media data. Sample size!!

Analysis of social media data, including ethics

Prediction + simulation methods (in health in particular but other contexts could be informative as well); How simulations are used in practice.

Using big data in mixed methods research

Qualitative analysis of social media data (going beyond the numbers into details like analysing text, images etc)

d) Other training needs

Data visualisation

- GIS: Excel spreadsheet – can do GIS in R. But employers want expertise in proprietary software. ARCGIS etc. Free software is good sometimes and can be cutting edge.
- Out of the box graphical representation of results, tools for health researchers who are bored with bar-plots + scatterplots (also for graphing temporal processes)

Training in Metadata. XML Markup

Costs: the cost of training can be a barrier and disappointing when you are paying £13,000 annual fee for a BSc or MSc already

Ethics/legal

- Legal compliance – esp. when using your own data resources, e.g. data from Twitter, FB, website associated with public engagement activities
- Barriers – legal, different data standards, technical/integration, skills
- Ethical implications of using social media/ consumer data etc. Ethics in analytics – big questions about university reputation in using commercial data – what are implications for wider issues - say refusal of health insurance for unhealthy shoppers. If we can, does it mean we should?