Introductory big data training for social scientists

The challenge

It is only in the past decade that 'big data' from sources like social media, digital sensors, financial and administrative transactions have become available as data commodities for the social scientist. While data science courses have been around for some time, there are still relatively few introductory events aimed at social scientists. The panacea of big data has led to a focus on realising the power of analytics rather than on data quality. Yet researchers need to appreciate the structure, provenance and trustworthiness of data, and its ethical entitlements, as well as the computational methods and tools used to analyse the data. Data services like the UK Data Service are well placed to roll out training across the whole of the 'big data' life cycle.

> The advent of big data has seen a focus on the power of analytics rather than on data quality,

Our approach

In 2015, we began working on 'scaling up' approaches to data curation and user access for big data. We focused particularly on areas where researchers need larger or more complex data sources (for example, where they need to download survey datasets greater than 5GB) or where computationally intensive and iterative modelling is needed. As part of this, we worked with other ESRC big data investments to build capacity and anticipate users' needs.

Only a handful of very large research datasets were available through the UK Data Service at the time, so we chose to focus on teaching social scientists how to find, access, explore and assess the quality of new forms of data, without needing to be skilled data scientists.

Our 'What is' webinar series introduced some of the simplest elements of the Apache Hadoop system – the software we are using for processing and distributing big data. The webinars showed researchers what the interfaces look like and what they need to get started, and how to get and work with internet data, which both proved very popular. They also got hands-on experience through a series of workshops on Hive for querying tabular data, and a half-day course on 'Getting data from the internet'. In all of these courses, we emphasised the need to carefully interrogate data, assessing its provenance and ethical and legal issues, conducting exploratory analyses and plotting data. There were also sessions on data management, publishing and replication in research. We could see from these early webinars and courses that many attendees would benefit from more basic IT knowledge or skills courses, so we introduced further introductory webinars and workshops on topics such as 'Storing data in databases' and 'Introduction to programming using Python', always using examples from the social sciences. However, these short sessions could only offer a taster of their usefulness.

Over the final six months of the project, we ran two week-long intensive summer schools entitled 'Encounters with big data for the social sciences', which brought our existing courses together with new workshop materials around the software packages and programming languages we use.

We built links with other UK social science investments in big data, working with their training coordinators and inviting guest speakers to give webinars on research-related issues such as: 'Linking multimedia city data', 'Investigating demographic representation on Twitter' and 'Creating Top Metric Maps'.

UK Data Service

Overview of capacity building activities

Hadoop and related	• Webinar: What is Hadoop?* • Webinar: What is Hive? • Webinar: What is Spark? • 1 day workshop: Introduction to big data manipulation using Hive
Getting data from the internet	• Webinar: What are APIs? • Webinar: What is MongoDB? • Half day workshop: Getting internet-based data
Basic IT skills and visualisation	• Webinar: Storing data in databases* • Webinar: What is Open Database Connector (ODBC)? • Webinar: Putting data on maps* • 1 or 2 day workshop: Introduction to programming using Python
Methodological issues	• Webinar: Big data and social research ethics* • Half day workshop: Big data and ethics • Half day workshop: Understanding big data • Half day workshop: Assessing quality in big data • Half day workshop: Reproducibility of research using big data
Big data summer school "Encounters with big data in the social sciences"	• 5 day programme aimed at skilled statisticians for scaling up handing and analysis of larger web based data

*Most popular webinars

Insights from training

Overall, we found considerable interest in new data and data skills, with some webinars attracting over 80 attendees. The highest attendances were for the 'What is Hadoop?' and 'Practical ethics for big data research' webinars. The workshops were all free, and filled up very quickly, including the two week-long 'Encounters' summer schools, each with only 25 places available. All the courses attracted attendees with unusually diverse backgrounds from economics, criminology and demography in the social sciences to computer science, health informatics and physics.

The blend of webinars and workshops allowed us to test different levels of interaction with our audiences. Using a variety of teaching methods meant we could reach different groups and cover a range of areas: webinars for introductory topics and awareness raising; workshops for more practical topics and more concrete learning outcomes – always building in enough time to practice using the tools.

Our approach was backed up by a survey by SAGE Publications, 'Who Is Doing Computational Social Science? Trends in Big Data Research, which surveyed social scientists around the world about their engagement with research using big data and the challenges they faced in conducting 'computational social science'. Forty per cent said they wanted big data training, and most wanted introductory training on big data analytics or data science, many listing specific topics.

Future plans

Looking ahead, we plan to focus on:

- new data sources that we hold or link to
- in-house systems needed to access and work with these datasets
- tools and software packages needed to store, manipulate and analyse these data

In the immediate future, we plan to incorporate the open source package, R into our survey skills training, as it has a wide range of applications for data mining and machine learning – including the SparkR library used to analyse big datasets in Hadoop, or TwitteR to work with twitter data, both areas in which people are showing an interest.

For now, we will continue to teach the basic courses, using external trainers and partner organisations that also have a mandate for capacity building – such as the UK's National Centre for Research Methods. We would also like to work with government and industry as they scale up their big data capabilities. Training in these sectors will benefit from focusing on high quality data sources.

See our case studies on:

- Upskilling social scientists in big data: 'Encounters' summer schools
- Scaling up: digital data services for the social sciences

Authors:

Louise Corti and Sarah King-Hele, UK Data Service

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

NRF National Research Foundation

DataFirst

UK Data Service