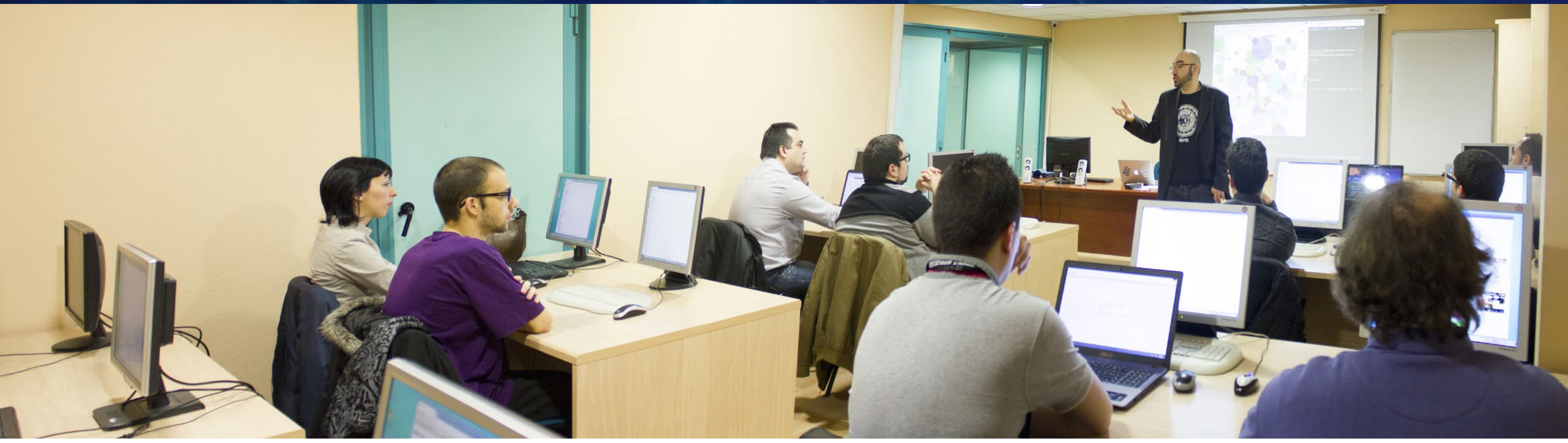
Data Service as a Platform

Upskilling social scientists in big data: 'Encounters' summer schools



The challenge

The past decade has seen a huge rise in studying social phenomena using data not initially collected for research and what we term 'big data' or new forms of data. Data gathered from the internet represent massive amounts of mineable information on the human condition, and data from digital sensors, financial transactions and administrative records are now viewed as important data commodities for the social scientist. However, despite the wide availability of training in data science and analytics, a significant gap still remains: how can the average social scientist make the transition into data science? The UK Data Service is well placed to help fill this gap by offering social scientists intensive training across the 'big data' lifecycle.

> Our aim was to support researchers in understanding and analysing large and complex datasets, focusing on using the power of popular statistical software like R in a big data environment.

The UK Data Service approach

The linked case study – Upskilling social scientists: introductory training – outlines the approaches we took to initiate our training in big data. Early engagement helped us assess user skills and needs, such as:

- moving from analysis of a social survey using SPSS to working with unfamiliar data formats
- calling huge datasets
- linking and mapping data
- using open source data tools

The big data summer schools we designed met a deliverable for the joint UK-South Africa Smarter Household Energy Data project, to run a free dedicated week-long course in South Africa as part of its knowledge exchange and capacity building agenda. The award built in sponsorship for researchers without funding to attend the course.

We pitched our summer school to the more skilled social science data analyst, and focused on teaching how to find, access and explore big data sources, keeping in mind the importance of questioning what lies behind the data and how to assess its provenance, trustworthiness, ethical entitlement and usability. A small team of UK expert advisors including Suzy Moat, David de Roure and Hugh Shanahan gave us an external view on the design of our course.

Encounters with Big Data: An Introduction to using Big Data in the Social Sciences was held in Cape Town in South Africa in February 2017 with 22 participants selected and vetted to meet the course requirements. They required experience using quantitative data in the social sciences, a good understanding of statistical methodology, and competence in writing commands in a statistical computing environment like Stata, R or SPSS. The course was repeated in the UK at the University of Essex in August 2017 as part of the Institute for Analytics and Data Science annual summer school, which attracted fees. There were 60 applicants, with places offered to 25 on a first come first serve basis.

Participants – South Africa Encounters summer school	University of Cape Town, University of Witwatersrand, Stellenbosch University, and the University of Pretoria, government agencies including the South African National Space Agency and the Human Sciences Research Council of South Africa	Experienced researchers and lecturers using cross-sectional and longitudinal survey data in the fields of public health, transport, finances and satellite and statistics and methodology
Participants – UK Encounters summer school	UK home institutions and universities, public and private agencies from number of countries including Denmark, Spain, Italy, Malta, Canada, Korea, Mexico, Mongolia and the USA	Professors, research assistants, postgraduates, statisticians and data analysts spanning fields from economics, sociology, criminology and geography to marine ecology and genomics

Course content

The five-day course covered data extraction, exploration, basic analysis and visualisation of big data using a 'Sandbox' version of Hadoop, the system that underpins the UK Data Service Data Services as a Platform (DSaaP). Hadoop provides solutions that can deliver data at scale, with speed, and include Hive, Spark, and Zeppelin, which integrate seamlessly with popular data analysis environments like Python and R.

The first day offered an overview of issues in using big data for social science research from senior staff at the UK Data Service, covering:

- using the scientific method
- creating national statistics using big data
- legal and ethical challenges for big data

Over the next two days, our tutors delivered presentations and led group exercises focused on manipulating data using Hadoop-based and other analytic tools.

The more technical sessions began with an all-important introduction to Hadoop components and alternatives. Participants were shown how to create and query tables in Hive to examine the contents of the datasets and to 'slice' and 'dice' a dataset into smaller datasets, and also how to access data tables in Hive using Open Data Base Connector (ODBC), for use in local tools, for example, Excel and R. Ambari Views was shown as a user-friendly user interface for Hive, and Zeppelin notebooks were introduced as an open web-based notebook for carrying out interactive data analysis.

Using Apache Spark, a high performance, distributed computation engine designed for handling and analysing big data, students learned how to scale out small-scale analyses implemented in R with SparkR, and about distributed computing and how and when it could benefit research. The power of R's libraries for producing various analytic functions spatial visualizations were emphasised, with the group creating their own choropleth maps using Leaflet.

Day four covered tools and techniques for getting and converting external data called from APIs, and how to interpret the results in JSON format. The final taught session focused on meeting the transparency agenda in research, and what is involved in publishing replication data and code. Participants got to create their own Github account and repository where code could be published

In the final day and a half, participants moved onto formulating their own group projects to consolidate what they had learned over the past four days, making use of the techniques and tools demonstrated, and using open data from the internet. Tasks included accessing structured data from the web, importing, preparing and linking them for exploratory data analysis and mapping.

You can read blog posts on both workshops.

Technical infrastructure

When teaching big data skills, having access to powerful computers is essential. We used 20 high spec. laptops for our Hadoop summer schools which gave us full control over our training environment, including data and materials. While laptops are time-consuming to set up individually and to move around, most university-based infrastructure does not support training-oriented access to high performance computing or Hadoop-type systems. At present, there is still limited support out there for our students to access what they have learned once they finish a training course.

Hands-on exercises and group work

The importance of offering practical skills and allowing time to discuss and debate, as opposed to throwing theory at students cannot be over-emphasised. Time and again at the UK Data Service, we note the positive feedback gained from intensive training. We designed the sessions to follow the structure of:

- a higher level conceptual presentation
- an example and, where appropriate, a demonstration
- group work or hands-on exercises, with tutors present to help

In the first group work session, participants were asked to come up with a national statistic of their choice based on data from the internet, putting any legal or ethical barriers to one side. Groups devised some really creative ideas in the areas of public health, deprivation, crime rates, vehicle accidents, tourism, and financial policy and fraud detection.

The groups put what they had learned into practice using unfamiliar open source data sources and software packages. Data sources considered were from Google Trends, Twitter and Facebook, published crime statistics and police station reports, local weather station pollution data, mobile phone/cell tower records, flight statistics and movie databases. A range of R tools were used including ggplot2, Rtweet, gtrendsR and wordcloud(R).

There were prizes for the teams we felt had best tried to implement the skills taught.

- In Cape Town 'Team Synergy', a group of demographic and health data managers and economists undertook a spatial analysis of mortality in South Africa using Zeppelin, Hive, and SparkR.
- In the UK, two teams who had used dataframes, R tools and Leaflet won joint first prize. The first, 'Racists be damned' plotted by constituency the proportion of people who signed anti-immigration petitions in the UK. They extracted JSON data from the Parliamentary Petitions site's API, choosing petitions with a 'negative' sentiment. The second team, 'Auto Choice Model', aimed to build an app to help buy a car. Using multiple data sources, they constructed a plot of car engine size (economy and emissions) by brand, accident severity and accident location (LSOA).

The winning project teams

Feedback received

"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."

"Really enjoyed the R and Spark sessions. Felt like this is what I'll be most able to work on in my future research (economics). The exercises were very useful –because the content was accessible and also because of the way that the code was templated for us to follow along with Introductory lectures appreciated – was easy enough to follow. Chris's anecdotes very welcome (thanks!)."

"Very good course indeed. Good structure, material, pace, content and teaching."

See our case studies on:

- Upskilling social scientists in big data: introductory training
- Scaling up: digital data services for the social sciences

Authors:

Louise Corti, Chris Park, Sarah King-Hele and Peter Smythe, UK Data Service

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
NRF National Research Foundation
DataFirst
UK Data Service