Encounters with big data: An introduction to using big data in the social sciences 2018

Big Data Analytics Summer School, University of Essex

Tutors: Louise Corti, Simon Parker, Myles Offord, Sharon Bolton, Cristina Magder, Darren Bell. UK Data Service

Course location: CSEE Lab 3, off Square 2

Lunch and coffee breaks: EBS Foyer

This 4-day course run by the UK Data Service introduces key concepts and discussions around using big data in the social sciences, and introduce attendees to approaches to and open source tools for exploring and analysing big data. It also looks at the challenges of replicability in science and covers best practices in being transparent about data and analysis when publishing results. The course, aimed at experienced researchers, statisticians, or data analysts, covers aspects of data evaluation (ethical, legal and practical), extraction, exploration, basic analysis and visualisation of data from the web using Spark R and R. Participants spend a full day on group projects applying what they have learned on real data. This course focuses on quantitative data and will not cover in any detail text, social media, audio etc.

Level: Introductory

Experience/knowledge required: Experience using quantitative research data in the social sciences. A good understanding of statistical methodology and concepts like standard error and standard deviation. Competence in writing commands in a statistical computing environment like Stata, R or SPSS.

Target audience: Aimed at experienced researchers, statisticians, or data managers

Monday 23 July: Introducing big data research	
8.30	Morning coffee
9:00	Introduction Presentation: Introduction to the summer school. Louise Corti Discussion: Participants' background and expectations. All Presentation: Technology for Big Data. Darren Bell
10:30	Coffee break
11:00	Big data, social science and social surveys Presentation: National statistics: Big data instead of social surveys? Louise Corti Exercise: National statistics experiment – discuss and investigate non-traditional data sources. Louise Corti
12:30	Lunch

13:30	Big data: ethics and disclosure risk Presentation: Ethics and rights in big data: risk, harm, governance, IPR, and 5 safes. Louise Corti Case study: Researching the Dark Web. Christian Kemp
15:00	Coffee break
15:30	Presentation: Introduction to Hadoop: components and alternatives. Darren Bell Presentation/Demo: How to set up your own Hadoop Sandbox on AWS. Darren Bell Demo: Introduction to documenting code: Jupyter Notebooks and R Markdown. Simon Parker
17:00	Close
18:00	Evening Reception: EBS Foyer

Tuesday 24 : Obtaining, assessing and exploring big data using Spark R and R	
8.30	Morning coffee
9:00	**Obtaining and managing big data. Simon Parker** Exercise: Keeping track of your work. Introducing Jupyter Notebooks and R markdown Demo and exercise: Introduction to R and Spark Demo and exercise: Overview of data wrangling with R and Spark, including linking and merging data sources
10:30	Coffee break
11:00	**Tools and techniques for getting and converting data from external sources. Myles Offord** Presentation, demo and exercise: Handling JSON formats Presentation, demo and exercise: Querying APIs Presentation, demo and exercise: SQL queries and using the Open Data Base Connector (ODBC)
12:30	Lunch
13:30	**Quality assessing big data** Presentation: Assessing and dealing with dirty data. Sharon Bolton Exercise : Assessing and dealing with dirty data. Simon Parker
15:00	Coffee break
15:30	Presentation: Assessing disclosure risk in data. Cristina Magder Exercise: Assessing disclosure risk in data. Cristina Magder
17:00	**Close**

Wednesday 25 July: Manipulating, exploring, analysing and publishing big data	
8.30	Morning coffee

9:00	**Exploring data with Spark and R. Simon Parker** Demo and exercise: Basic data visualization and modelling with R and Spark
10:30	Coffee break
11:00	**Maps: Creating maps in R with leaflet. Simon Parker** Demo and exercise: Creating interactive maps in R with Leaflet
12:30	Lunch
13:30	**Publishing and being transparent with big data** Presentation: Being transparent in science, publishing data and code. Louise Corti Demo and exercise: Create your own Github account and repository. Myles Offord
15:00	Coffee break
15:30	**Group projects** Introducing your group projects. Louise Corti Exercise: Brainstorm and formulate group projects
17:00	Close

Thursday 26: Projects	
8.30	Morning coffee
9.00	Group projects (supported by tutors and helpers)
10:30	Coffee break
11:00	Group projects
12:30	Lunch
13:30	Group projects
15:00	Comfort break/drinks and getting ready to present projects
15:15	Project presentations and prize ceremony
16:15	Close