



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

**Big Data Analytics Summer School, University of Essex**

**Organisers and tutors:** UK Data Service

**Course location:** Essex Business School (EBS) Room 1.1

**Lunch and coffee breaks:** EBS Foyer

## Monday 31<sup>st</sup> 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: Big Data – but we are still doing science - *Nathan Cunningham*

10:30 Coffee break

11:00 **Big data, social science and social surveys**

Presentation: Big Data – but we are still doing science (continued) - *Nathan*

*Cunningham*

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  
*Libby Bishop*

Case study: *James Allen-Roberson and Christian Kemp*

15:00 Coffee break

15:30 **Introduction to Hadoop and alternatives**

Presentation and exercise: Introduction to Hadoop: components and alternatives. *Peter Smyth*

17:00 Close

17:30-18:30 **Optional: Drinks reception at the UK Data Archive**

17.30-18.15 **Optional: IADS Summer School event:** Keynote speaker: Analysis and Forecasting of Complex Dynamical Systems with Neural Networks - *Hans-Georg Zimmerman, Senior Principal Research Scientist, Siemens Corporate Technology (EBS2.34 )*



## Tuesday 1<sup>st</sup> August: Manipulating data using Hive

8.30	Morning coffee
9:00	<b>Introduction to Hive and running Hive queries in Zeppelin notebooks</b> Presentation: Introduction to Hive and big data warehousing. <i>Peter Smyth</i> Demo and exercises: Introduction to Notebooks and Zeppelin. <i>Peter Smyth</i>
10:30	Coffee break
11:00	<b>Hive Exploratory data analysis using Ambari views</b> Demo and exercises: Exploratory data analysis with Hive. <i>Peter Smyth</i>
12:30	Lunch
13:30	<b>Creating and populating Hive tables using Zeppelin and Ambari Views</b> Demo and exercises: Advanced Hive. <i>Peter Smyth</i>
15:00	Coffee break
15:30	<b>Accessing Hive using Open Data Base Connector (ODBC)</b> Demo and exercises: ODBC in Excel and R. <i>Peter Smyth</i>
17:00	<b>Close</b>
19.30-22.20	<b>Optional: IADS Summer School event: Formal dinner at Wivenhoe House Hotel</b>

## Wednesday 2<sup>nd</sup> August: Manipulating and analysing data using Spark

8.30	Morning coffee
9:00	<b>Spark 1: Introduction</b> Presentation: Introduction to Computing with (Big) Data. <i>Chris Park</i> Demo and exercise: Introduction to R and Spark. <i>Chris Park</i>
10:30	Coffee break
11:00	<b>Spark 2: Exploring data with R and Spark – Part 1</b> Demo and exercise: Overview of data wrangling with R and Spark. <i>Chris Park</i>
12:30	Lunch
13:30	<b>Spark 3: Exploring data with R and Spark – Part 2</b> Demo and exercise: Basic data visualization and modelling with R and Spark. <i>Chris Park</i>
15:00	Coffee break
15:30	<b>Maps: Creating maps in R with leaflet</b> Demo and exercise: Creating interactive maps in R with Leaflet. <i>Peter Smyth</i>
17:00	Close
17.15-18.00	<b>Optional: IADS Summer School event: Keynote speaker: Beyond the Data and Digital transformation: Becoming a smart organisation – Antonio Alvarez/Javier Aparicio Gonzales, Head of Data Innovation, Isban UK/Santander UK (EBS2.34)</b>

**Thursday 3<sup>rd</sup> August: Tools and techniques for dealing with external data; Projects**

8.30	Morning coffee
9:00	<b>Tools and techniques for getting and converting data from external sources</b> Presentations and demos: APIs and JSON. <i>Peter Smyth</i>
10:30	Coffee break
11:00	<b>Review and getting your own Hadoop system</b> Review: End to end processing using the Hadoop stack. <i>Peter Smyth</i> Presentation: How to get your own Hadoop system. <i>Peter Smyth</i> Presentation: Introducing your group projects. <i>Louise Corti</i> <b>Formulate group projects</b>
12:30	Lunch
13:30	<b>Group projects</b>
15:00	Coffee break
15:30	<b>Group projects</b>
17:00	Close

**Friday 4<sup>th</sup> August: Transparency agenda and Github; Project presentations**

8.30	Morning coffee
9:00	<b>Meeting the transparency agenda in research</b> Presentation: Data replication and data publishing. <i>Louise Corti</i> Demo and exercise: Create your own Github account and repository. <i>Peter Smyth</i> <b>Continue group projects</b>
10:30	Coffee break
11:00	<b>Finalize group projects</b>
12:30	Lunch
13:30	<b>Group projects presentations and feedback</b> Group Presentation: Present group projects Discussion: Provide feedback on group projects
15:00	Coffee break
15:30	<b>Group projects presentations and feedback</b> Group Presentation: Present group projects Discussion: Provide feedback on group projects <b>Prize giving for best project!</b> <b>Summer school feedback and questions</b>
16:30	Close