

Introduction to Data Science and Big Data Analytics

ME314 2018

LSE Methods Summer Programme 2018

London School of Economics and Political Science

Kenneth Benoit (kbenoit@lse.ac.uk), Department of Methodology, LSE

Slava Jankin Mikhaylov (s.mikhaylov@essex.ac.uk), Institute for Analytics and Data Science, University of Essex

Jack Blumenau (J.Blumenau@ucl.ac.uk), Department of Political Science, UCL

This repository contains the class materials for the [Research Methods, Data Science, and Mathematics](#) course *ME314 Introduction to Data Science and Big Data Analytics* taught in July-August 2018 by Kenneth Benoit, Slava Mikhaylov, and Jack Blumenau.

Quick links to topics

Day	Date	Instructor	Topic
1	Mo 30 Jul	KB, SM	Overview and introduction to data science
2	Tu 31 Jul	KB	The Shape of Data
3	We 1 Aug	KB	Working with Databases
4	Th 2 Aug	KB	Linear Regression
5	Mo 6 Aug	SM	Classification
6	Tu 7 Aug	SM	Resampling methods, model selection and regularization
7	We 8 Aug	SM	Non-linear models and tree-based methods
8	Th 9 Aug	KB	Unsupervised learning and dimensional reduction
9	Mo 13 Aug	JB	Text analysis
10	Tu 14 Aug	JB	Topic modelling

Day	Date	Instructor	Topic
11	We 15 Aug	JB	Data from the Web
12	Th 16 Aug	KB	Network Analysis
13	Fr 17 Aug		Final Exam

Overview

Data Science and Big Data Analytics are exciting new areas that combine scientific inquiry, statistical knowledge, substantive expertise, and computer programming. One of the main challenges for businesses and policy makers when using big data is to find people with the appropriate skills. Good data science requires experts that combine substantive knowledge with data analytical skills, which makes it a prime area for social scientists with an interest in quantitative methods.

This course integrates prior training in quantitative methods (statistics) and coding with substantive expertise and introduces the fundamental concepts and techniques of Data Science and Big Data Analytics.

Typical students will be advanced undergraduate and postgraduate students from any field requiring the fundamentals of data science or working with typically large datasets and databases. Practitioners from industry, government, or research organisations with some basic training in quantitative analysis or computer programming are also welcome. Because this course surveys diverse techniques and methods, it makes an ideal foundation for more advanced or more specific training. Our applications are drawn from social, political, economic, legal, and business and marketing fields.

Objectives

This course aims to provide an introduction to the data science approach to the quantitative analysis of data using the methods of statistical learning, an approach blending classical statistical methods with recent advances in computational and machine learning. We will cover the main analytical methods from this field with hands-on applications using example datasets, so that students gain experience with and confidence in using the methods we cover. We also cover data preparation and processing, including working with structured databases, key-value formatted data (JSON), and unstructured textual data. At the end of this course students will have a sound understanding of the field of data science, the ability to analyse data using some of its main methods, and a solid foundation for more advanced or more specialised study.

The course will be delivered as a series of morning lectures, followed by lab sessions in the afternoon where students will apply the lessons in a series of instructor-guided exercises using

data provided as part of the exercises. The course will cover the following topics:

- an overview of data science and the challenge of working with big data using statistical methods
- how to integrate the insights from data analytics into knowledge generation and decision-making
- how to acquire data, both structured and unstructured, and to process it, store it, and convert it into a format suitable for analysis
- approaches to normalising data, using a database manager (SQLite), and working with SQL database queries
- the basics of statistical inference including probability and probability distributions, modelling, experimental design
- an overview of classification methods and related methods for assessing model fit and cross-validating predictive models
- supervised learning approaches, including linear and logistic regression, decision trees, and naïve Bayes
- unsupervised learning approaches, including clustering, association rules, and principal components analysis
- quantitative methods of text analysis, including mining social media and other online resources
- social network analysis, covering the basics of social graph data and analysing social networks
- data visualisation through a variety of graphs.

Prerequisites

Students should already be familiar with quantitative methods at an introductory level, up to linear regression analysis. Familiarity with computer programming or database structures is a benefit, but not formally required.

Preparing for the course

While we have also made a cloud server available to you for working with all of the assignments, if you want to use these locally or after the course, we recommend that you download and install [R](#) and [RStudio](#) on your computer.

Detailed instructions can also be found [here](#) for installing the tools you need and working with the lab materials.

If you are not already familiar with R, we strongly encourage you to attempt to become familiar before the start of the course. That way, you will spend much less time become familiar with the tools, and be able to focus more on the methods. The following links provide a basic introduction to R, which you can study at your own pace before the course begins.

- [An Introduction to R](#).
- Data Camp [R tutorials](#).
- Data Camp [R Markdown tutorials](#), first chapter.
- [R Codeschool](#).

We strongly recommend you spend some of July and August before the course working through the following materials:

- Garrett Golemund and Hadley Wickham (2016) *R for Data Science*, O'Reilly Media. Note: Online version is available from the authors' page [here](#).
- James et al. (2013) *An Introduction to Statistical Learning: With applications in R*, Springer, Chapters 1–2. Note: The book is available from the authors' page [here](#).

If you start preparing for the course (with the above materials) using your own laptop, it may be more convenient for you to continue using it during the summer school.

Important Specifics

Computer Software

Computer-based exercises will feature prominently in the course, especially in the lab sessions. The use of all software tools will be explained in the sessions, including how to download and install them. All of the class work will be done using R, using publicly available packages.

Main Texts

The primary texts are:

- James et al. (2013) *An Introduction to Statistical Learning: With applications in R*, Springer. Note: The book is available from the authors' page [here](#).
- Garrett Golemund and Hadley Wickham (2016) *R for Data Science*, O'Reilly Media. Note: Online version is available from the authors' page [here](#).
- Zumel, N. and Mount, J. (2014). *Practical Data Science with R*. Manning Publications.

The following are supplemental texts which you may also find useful:

- Lantz, B. (2013). *Machine Learning with R*. Packt Publishing.
- Lesmeister, C. (2015). *Mastering Machine Learning with R*. Packt Publishing.
- Conway, D. and White, J. (2012) *Machine Learning for Hackers*. O'Reilly Media.
- Leskovec, J., Rajaraman, A. and Ullman, J. (2011). *Mining of Massive Datasets*. Cambridge University Press.
- Zafarani, R., Abbasi, M. A. and Liu, H. (2014) *Social Media Mining: An introduction*. Cambridge University Press.
- Hastie et al. (2009) *The Elements of Statistical Learning: Data mining, inference, and prediction*. Springer. Note: The book is available from the authors' page [here](#).

Instructors

Kenneth Benoit is Professor of Quantitative Social Research Methods at the Department of Methodology, LSE. With a background in political science, his substantive work focuses on political party competition, political measurement issues, and electoral systems. His research and teaching is primarily in the field of social science statistical applications. His recent work concerns the quantitative analysis of text as data, for which he has developed a package for the R statistical software.

Slava Jankin Mikhaylov is Professor of Public Policy and Data Science at University of Essex, holding a joint appointment in Department of Government and School of Computer Science and Electronic Engineering Institute for Analytics and Data Science (IADS). He serves as a Chief Scientific Adviser to Essex County Council. His current work focuses on embedding artificial intelligence and data science in public service delivery.

Jack Blumenau is a Lecturer in Quantitative Methods at the UCL Department of Political Science. he was previously an ESRC “Future Research Leader” post-doc in the Methodology Department at the LSE. His research explores the effects of legislative leaders on the behaviour of parliamentarians in the European Parliament and the UK House of Commons.

Assessment

Daily lab exercises

These are not assessed, but will form the practical materials for each day’s labs. See these instructions for how to access and work with each day’s exercise.

See <https://lse-me314.github.io/instructions> for detailed instructions on obtaining and working with each day’s lab materials.

Mid-term project

The class assignment for Day 5 will count as the mid-term assignment, which will count for 25% of the grade. The [mid-term is available here](#).

Exam

Friday, 17 August. This will be a “take-home” exam. You are welcome to use a lab as a quiet room on Friday to complete the assignment, but you can also do this from home. Instructors will be available throughout the day by email for answering any procedural questions or clarifying anything about the exam that is unclear.

- **Instructions:** [The exam is here](#). Complete and submit the exam just as you would any lab assignment: by renaming the file, editing the R Markdown, knitting, and submitting

through Moodle your knitted HTML file.

- **Formatting:** Put your own textual answers in boldface (using `**boldface type**` in RMarkdown), so that we can easily identify them when reviewing your HTML file.
- **Deadline:** Monday 20 August 17:00 London time (GMT+1)

Detailed Course Schedule

1. Overview and introduction to data science

We will use this session to get to know the range of interests and experience students bring to the class, as well as to survey the machine learning approaches to be covered. We will also discuss and demonstrate the R software.

Resources

- [Lecture Notes Part 1](#)
- [Lecture Notes Part 2](#)
- [Lab 1 materials](#)

Required reading

- James et al (2013), Chapters 1–2. Note: The book is available from the authors' page [here](#).
- *An Introduction to R*.
- Downloading and installing [RStudio](#) and [R](#) on your computer.
- Data Camp [R tutorials](#).
- Data Camp [R Markdown tutorials](#), first chapter.
- [R Codeschool](#).
- Garrett Golemund and Hadley Wickham (2016) *R for Data Science*, O'Reilly Media, Chapters 1-3. Note: Online version is available from the authors' page [here](#).

Recommended Reading

- Patrick Burns, 2011. *The R Inferno*. Available [here](#).
- Lantz, Ch. 2.

2. The Shape of Data

This week introduces the concept of data “beyond the spreadsheet”, the rectangular format most common in statistical datasets. It covers relational structures and the concept of database normalization. We will also cover ways to restructure data from “wide” to “long” format, within strictly rectangular data structures. Additional topics concerning text encoding, date formats, and sparse matrix formats are also covered.

Resources

- [Lecture Notes](#)
- [Lecture Notes as pdf](#)
- [Lab 2 materials](#)

Required reading

- Wickham, Hadley and Garrett Grolemund. 2017. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. Sebastopol, CA: O'Reilly. [Part II Wrangle, Tibbles, Data Import, Tidy Data](#) (Ch. 7-9 of the print edition).

Recommended Reading

- Reshaping data in Python: "[Reshaping and Pivot Tables](#)".
- Robin Linderborg, "[Reshaping Data in Python](#)", 20 Jan 2017.

3. Working with Databases

We will return to database normalization, and how to implement this using good practice in a relational database manager, SQLite. We will cover how to structure data, verify data types, set conditions for data integrity, and perform complex queries to extract data from the database. We will also cover authentication and how to connect to local and remote databases.

Resources

- [Lecture Notes](#)
- [Lecture Notes as pdf](#)
- [Lab 3 materials](#)

Required reading

- Lake, Peter. *Concise Guide to Databases: A Practical Introduction*. Springer, 2013. Chapters 4-5, Relational Databases and NoSQL databases.
- Nield, Thomas. *Getting Started with SQL: A hands-on approach for beginners*. O'Reilly, 2016. Entire text.

Recommended Reading

- [SQLite documentation](#).
- Bassett, L. 2015. *Introduction to JavaScript Object Notation: A to-the-point Guide to JSON*. O'Reilly Media, Inc.

4. Linear Regression

Linear regression model and supervised learning.

Resources

- [Lecture Notes](#)
- [Lab 4 materials](#)

Required Reading

- James et al., Chapter 3.

Recommended Reading

- Zumel and Mount, Chapter 7.1.
- Lantz, Chapter 6

5. Classification

Logistic regression, discriminant analysis, Naive Bayes, evaluating model performance.

Resources

- [Lecture Notes](#)
- [Lab 5 materials](#)

Required Reading

- James et al., Chapter 4.

Recommended Reading

- Lesmeister, Chapter 3.
- Zumel and Mount, Chapters 5, 6, 7.2.
- Lantz, Chapters 3-4, 10.

6. Resampling methods, model selection and regularization

Cross-validation, bootstrap, ridge and lasso.

Resources

- [Lecture Notes](#)
- [Lab 6 materials](#)
- [World Bank South Africa NDP](#)

Required Reading

- James et al., Chapter 5-6.

Recommended Reading

- Lesmeister, Chapter 4.

7. Non-linear models and tree-based methods

GAMs, local regression, decision trees, random forest, boosting.

Resources

- [Lecture Notes](#)
- [Lab 7 materials](#)
- [United Nations example](#)

Required Reading

- James et al., Chapter 7-8.

Recommended Reading

- Lesmeister, Chapter 6.
- Zume and Mount, Chapter 9.1-9.3.
- Muchlinksi, D., Siroky, D., Jingrui, H., Kocher, M., (2016) "Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data." *Political Analysis*, 24(1): 87-103.

8. Unsupervised learning and dimensional reduction

Cluster analysis, PCA, correspondence analysis, association rules.

Resources

- [Lecture Notes](#)
- [Lab 8 materials](#)

Required reading

- James et al., Chapter 10.

Recommended Reading

- Lesmeister, Chapter 5, 8-9.
- Zumel and Mount, Chapter 8.
- Lantz, Chapters 8-9
- Leskovec et al., Chapter 11.

9. Text analysis

Working with text in R, sentiment analysis, dictionary methods.

Resources

- [Lecture Notes](#)
- [Lab 09 materials](#)

Required reading

- Grimmer, J, and B M Stewart (2013), ``Text as Data: the Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis*.
- Benoit, Kenneth and Alexander Herzog. In press. ``Text Analysis: Estimating Policy Preferences From Written and Spoken Words.'" In *Analytics, Policy and Governance*, eds. Jennifer Bachner, Kathryn Wagner Hill, and Benjamin Ginsberg.

Recommended Reading

- Denny, M.J. and Spirling, A. (2018), ``Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It" *Political Analysis*
- Spirling, A. (2012), ``Democratization and Linguistic Complexity: The Effect of Franchise Extension on Parliamentary Discourse, 1832–1915." *Journal of Politics*
- Herzog, A. and K. Benoit (2015), ``The most unkindest cuts: Speaker selection and expressed government dissent during economic crisis." *Journal of Politics*, 77(4):1157–1175.
- Benoit, K., Munger, K., and Spirling, A. ``Measuring and Explaining Political Sophistication Through Textual Complexity.", *Working Paper*

10. Topic modelling

Latent Dirichlet Allocation, Correlated Topic Model, Structural Topic Model.

Resources

- [Lecture Notes](#)
- [Lab 10 materials](#)

Required reading

- David Blei (2012). "Probabilistic topic models." *Communications of the ACM*, 55(4): 77-84.
- Blei, David, Andrew Y. Ng, and Michael I. Jordan (2003). "Latent dirichlet allocation." *Journal of Machine Learning Research* 3: 993-1022.
- Blei, David (2014) "Build, Compute, Critique, Repeat: Data Analysis with Latent Variable Models." *Annual Review of Statistics and Its Application*, 1: 203-232.

Recommended Reading

- Blei, D. and J. Lafferty "Topic Models." In *Text Mining: Classification, clustering, and applications*, A. Srivastava and M. Sahami (eds.), pp 71-94, 2009. Chapter available [here](#).
- Blei, David M., and John D. Lafferty. "Dynamic topic models." In *Proceedings of the 23rd international conference on machine learning*, pp. 113-120. ACM, 2006.
- Mimno, D. (April 2012). "Computational Historiography: Data Mining in a Century of Classics Journals." *Journal on Computing and Cultural Heritage*, 5 (1).
- Lesmeister Chapter 12.

11. Data from the Web

The promises and pitfalls of social media data. The Twitter API. The Facebook API. Web scraping. Ethics.

Resources

- [Lecture Notes](#)
- [Lab 11 materials](#)
- [R script for twitterR lecture example](#)
- [R script for web scraping lecture example](#)

Recommended Reading:

- Broniatowski, David A, Michael J Paul, and Mark Dredze. 2013. "National and Local Influenza Surveillance Through Twitter: an Analysis of the 2012-2013 Influenza Epidemic" *PLoS ONE* 8(12): 83672-78. [PDF here](#)
- Barbera, Pablo., 2017. "Less is more? How demographic sample weights can improve public opinion estimates based on Twitter data." *Working Paper*
- Munger, Kevin., 2017. "Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment" *Political Behaviour* 39(3): 629-649

- Ginsberg et al., 2008. “Detecting influenza epidemics using search engine query data” *Nature* 457: 1012–1014.
- Lazer et al., 2014. “The Parable of Google Flu: Traps in Big Data Analysis” *Science* 343: 1203-1205
- [Earthquake shakes Twitter users: real-time event detection by social sensors](#)
- <http://rcrastinate.blogspot.co.uk/2015/02/mapping-world-with-tweets-including-gif.html>
- <https://github.com/pablobarbera/streamR>
- Matthew Russell (2013). *Mining the Social Web*. O'Reilly Media. 2nd edition.

To use the `twitterR` package in R:

- Twitter Authentication setup:
 - [Official](#)
 - [Walkthrough](#)
- Twitter API documentation:
 - [Overview of REST API](#)
 - [Overview of streaming API](#)

12. Network Analysis

More text analysis, including dictionaries and identifying keywords; and an introduction to network analysis. Will also include an exam review.

Resources

- [Lecture Notes](#)
- [Lab 12 materials](#)

Required reading

- Stephen Borgatti, Ajay Mehra, Daniel Brass, Giuseppe Labianca. 2009. “Network Analysis in the Social Sciences” *Science* 323, 892-895

Recommended Reading

- [Using Metadata to Find Paul Revere](#)
- G. Erkan and D. Radev. 2004. “LexRank: Graph-based lexical centrality as salience in text summarization” *Journal of Artificial Intelligence Research* 22, 457 - 479
- Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. “Birds of a Feather: Homophily in Social Networks” *Annual Review of Sociology* 27, 415-44

This page was generated by [GitHub Pages](#).