ME314 Introduction to Data Science and Big Data Analytics

LSE Methods Summer Programme 2018

Kenneth Benoit, Department of Methodology, LSE
Slava Mikhaylov, Institute for Analytics and Data Science, University of Essex
Jack Blumenau, 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.

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

Before the course you should:

- Download and install R *and* RStudio on your computer.

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 Grolemund 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 Grolemund 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.

Instructions for Submitting Homeworks

Each homework will be a single file in the RMarkdown format. The files linked below are *named very carefully*, to make it easy for us to identify your completed lab assignments.

1. Obtaining the starter files.

   Each day below will link the name of a starter file for you to download and work with. These are in the RMarkdown format. You should embed your answers, with code, into your version of the instruction files.

2. Renaming the starter files.

   For example, the first assignment file is named `ME314_assignment1_LASTNAME_FIRSTNAME.Rmd`, which you will need to rename by replacing the uppercase terms with your own last and first names, e.g. `ME314_assignment1_Bloggs_Joe.Rmd`.

3. From RStudio, you can create an HTML output file with your evaluated code, figure, and text answers by clicking the "Knit HTML" button in the top of the editor pane in RStudio. The resulting HTML file will be saved in your working directory.

4. You will need to upload the resulting HTML file – renamed! – to the course Moodle page, to the appropriate assignment folder.

We will walk you through this process in the Day 1 lab, so don't worry if it seems complicated the first time. This sort of careful workflow process and file management is part of learning practical data science too!

We do not mark the homework but we will walk through the solutions at the start of computer labs the following day.

Instructions for use of course materials

You have three options for downloading the course material found on this page:

1. You can download the materials by clicking on each link.
2. You can "clone" repository, using the buttons found to the right side of your browser window as you view this repository. This is the button labelled "Clone in Desktop". If you do not have a git client installed on your system, you will need to get one here and also to make sure that git is installed. This is preferred, since you can refresh your clone as new content gets pushed to the course repository. (And new material will get actively pushed to the course repository at least once per day as this course takes place.)
3. Statically, you can choose the button on the right marked "Download zip" which will download the entire repository as a zip file.

You can also subscribe to the repository if you have a GitHub account, which will send you updates each time new changes are pushed to the repository.

You can read more about using GitHub with RStudio in Hadley Wickham's *R Packages* book Chapter 13.

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 Mikhaylov is Professor of Public Policy and Data Science at Institute of Analytics and Data Science and Department of Government, University of Essex. He's a Chief Scientific Advisor to Essex County Council and a co-investigator in an ESRC Big Data infrastructure investment initiative – Consumer Data

Research Centre at UCL. His research and teaching is primarily in the field of computational social science and data science.

Dr. 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.

Course Schedule

Monday, 30 July: Overview and introduction to data science [SM,KB]

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
- Assignment 1 as R markdown
- Assignment 1 **solution** as R markdown

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 Grolemund 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.

Tuesday, 31 July: The Shape of Data [KB]

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
- Assignment 2 as R markdown
- Assignment 2 **solution** as R markdown
- Altaf’s resource on **dplyr** and the “tidyverse”
- Altaf’s resource on **ggplot2**

Required reading

- Wickham, Hadley and Garett 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.

___*Wednesday, 1 August: Working with Databases [KB*___]

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
- Assignment 3 as R markdown

- Assignment 3 **solution** as R markdown

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.

Thursday, 2 August: Linear Regression [KB]

Linear regression model and supervised learning.

Resources

- Lecture Notes
- Assignment 4 as R markdown
- Assignment 4 **solution** as R markdown

Required Reading

- James et al., Chapter 3.

Recommended Reading

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

Monday, 6 August: Classification [SM]

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

Resources

- Lecture Notes
- Assignment 4 as R markdown
- Assignment 4 **solution** as R markdown

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.

Tuesday, 7 August: Resampling methods, model selection and regularization [SM]

Cross-validation, bootstrap, ridge and lasso.

Resources

- Lecture Notes
- Assignment 6 as R markdown
- Assignment 6 **solution** as R markdown
- Dataverse replication example for PCR

Required Reading

- James et al., Chapter 5-6.

Recommended Reading

- Lesmeister, Chapter 4.

Wednesday, 8 August: Non-linear models and tree-based methods [SM]

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

Resources

- Lecture Notes
- RandomForest research example
- Assignment 7 as R markdown
- Assignment 7 **solution** as R markdown

Required Reading

- James et al., Chapter 7-8.

Recommended Reading

- Lesmeister, Chapter 6.
- Zumel 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.

Thursday, 9 August: Unsupervised learning and dimensional reduction [KB]

Cluster analysis, PCA, correspondence analysis, association rules.

Resources

- Lecture Notes
- Application example: smart meter analysis Notes
- Assignment 7 as R markdown
- Assignment 7 **solution** as R markdown

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.

Monday, 13 August: Text analysis [JB]

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

Resources

- Lecture Notes
- Assignment 9 as R markdown or html
    - sample zip file `UKimmigTexts.zip` of texts for building a corpus
- Assignment 9 **solution** as R markdown

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, Kathyrn Wagner Hill, and Benjamin Ginsberg.

Recommended Reading

- Spirling, A. (2012), "U.S. Treaty Making with American Indians: Institutional Change and Relative Power, 1784–1911." *American Journal of Political Science*, 56: 84–97.
- 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.

Tuesday, 14 August: Topic modelling [JB]

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

Resources

- Lecture Notes
- Assignment 10 as R markdown
- Assignment 10 **solution** as R markdown

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.

Wednesday, 15 August: Mining the Social Web [JB]

Working with the Twitter API, Facebook API, JSON data, and examples.

Resources

- Lecture Notes
- General examples from the lecture
- Streaming example code
- Rest Example code
- Assignment 11 as R markdown.

Required 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
- Twitter Authentication setup:
  - Official
  - Walkthrough

- Twitter API documentation:
  - Overview of REST API
  - Overview of streaming API

Recommended Reading

- 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/twitter/AnomalyDetection
- https://github.com/pablobarbera/streamR
- Zafarani et al., Chapters 1-4.
- Matthew Russell (2013). *Mining the Social Web.* O'Reilly Media. 2nd edition.

Thursday, 16 August: Network Analysis [JB]

Working with network data, network data types, centrality measures.

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

Assessment

Exam: Friday, 17 August, Time and Room TBC

- **Instructions:** 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)