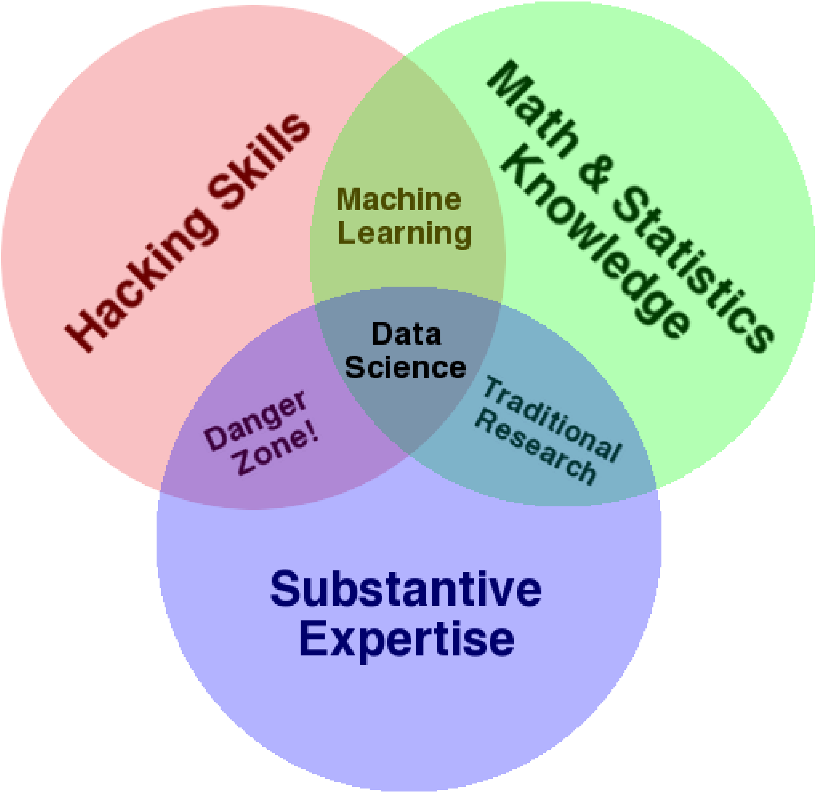
POLS3003 Data Science and Big Data Analytics

https://github.com/smikhaylov/POLS3003

Slava Mikhaylov
s.mikhaylov@ucl.ac.uk
Office: Room 2.01 29/30 Tavistock Square
Office hours: Wednesdays 11:00-13:00

Version: January 5, 2016



Overview and goals

Data Science and Big Data Analytics are exciting new areas that combine scientific inquiry, substantive expertise, programming, and statistical knowledge. One of the main challenges for businesses and policy makers when using big data is to find people with the appropriate skills. Data Science is no longer only the domain of computer scientists and engineers. 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.

Drew Conway's Data Science Venn diagram on the title page highlights the focus of this course. The 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. Students taking this module will be introduced to the most fundamental data analytic tools and techniques, and learn how to use specialised software to analyse real-world data and answer policy-relevant questions.

Learning outcomes

Upon successful completion of this module, students will:

- have a sound understanding of the field of data science and develop the ability to analyse real-world data using some of its main methods;
- become comfortable applying regression models for continuous and limited outcome variables;
- explore more complex models, such as the widely-used panel data models;
- develop familiarity with descriptive and predictive analytics, and their application to big data problems;
- explore methods of text analytics and automated data acquisition;
- have received a solid foundation for more advanced or more specialised study.

Prerequisites

This is an advanced course intended for students who've already had some training in quantitative methods for data analysis. An introduction to quantitative methods (statistics/econometrics) at any level would serve as a very useful foundation for this course, although no formal prerequisites are required. Familiarity with computer programming or database structures is a benefit, but not formally required.

If you are unsure whether your prior statistical training is sufficient to take this course please contact the course tutor for confirmation.

Before you take the course

You are strongly recommended before taking this course to complete the following:

- James, Witten, Hastie, and Tibshirani (2013) *Introduction to Statistical Learning with Applications in R*, Chapters 1-2.
- *An Introduction to R*, available from `http://cran.r-project.org/doc/manuals/R-intro.pdf`
- Downloading and installing RStudio, available from `http://www.rstudio.com`.
- A brief online introduction to R Markdown, which we will use for completing the exercises for the course, see `http://bit.ly/R_markdown`

What to expect

Reading

The handout lists the required readings for every week. This required reading should be completed prior to the lecture in a given week. Students are expected to read the material very carefully. You may even find it helpful to read it first, come to the lecture and then re-read it after the lecture.

Homework

This is a methodological course, developing skills in understanding and applying statistical methods. You can only learn statistics by doing statistics and, therefore, the homework for this course is extensive, including weekly homework assignments. The assignments consist of data analytic tasks that you will be asked to complete either on your own or in small groups. It is important to learn to work in small groups because that's how much of the private and public sector operates. You may also be asked to present the results of your group work at the subsequent seminar. This helps you develop public speaking skills and skills of presentation of the results of your analysis to a wider audience – something you'll have to do in your job after finishing the degree.

Homework and readings will keep you busy. But this is an intensive advanced level course where we cover a lot of ground in ten weeks. So if you take this course you shouldn't expect an easy ride. However, skills you learn in the course are transferrable well beyond your degree.

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 and RStudio as its integrated development environment (IDE): https://www.rstudio.com
- We are also using R Markdown as a format for all document (including final assessment) creation from R: http://rmarkdown.rstudio.com
- GitHub is used to distribute materials for the module: https://github.com/smikhaylov/POLS3003
- We will use Piazza to discuss any issues relating to the module: http://bit.ly/POLS3003-2016

Main Texts

The primary texts are:

- James et al. (2013) *An Introduction to Statistical Learning: With applications in R*. Springer. The book is available from the authors' page: http://www-bcf.usc.edu/~gareth/ISL/
- Hastie et al. (2009) *The Elements of Statistical Learning: Data mining, inference, and prediction*. Springer. The book is available from the authors' page: http://statweb.stanford.edu/~tibs/ElemStatLearn/

Additional recommended texts that we'll be using in the module:

- Kuhn, M. and K. Johnson (2013). *Applied Predictive Modeling*. Springer.
- Lesmeister, C. (2016). *Mastering Machine Learning with R*. Packt Publishing.
- Zumel, N. and Mount, J. (2014). *Practical Data Science with R*. Manning Publications.
- Leskovec, J., Rajaraman, A. and Ullman, J. (2014). *Mining of Massive Datasets*. 2nd edition. Cambridge University Press.

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

- Kromer, P. and R. Jurney (2016). *Big Data for Chimps*. O'Reilly.
- Lantz, B. (2013). *Machine Learning with R*. Packt Publishing.
- Conway, D. and White, J. (2012). *Machine Learning for Hackers*. O'Reilly Media.

Assessment

The course is assessed with a 3,000-word research paper in the form of a replication assignment (100% of the mark for the course). The goal of the research paper is to help you develop a deeper understanding of data analytic methods in your field of study and ability to apply them to a concrete empirical problem.

1. Your paper should address a substantive problem in your field of interest.
2. We will discuss the basic principles of replication and your essay assignment before the Reading Week, giving you an opportunity to work on the project during the Reading Week.
3. The research paper should replicate an academic article published in the last 5-8 years in a top ranking political science or economics journal. The definition of top ranking is any of the top 20 journals from the Google Scholar rankings:

   - Social Science (including any subcategories): http://bit.ly/Social_Science
   - Business, Economics & Management (including any subcategories): http://bit.ly/Business_Econ

   Higher quality articles tend to appear in higher ranked, general field (rather than subfield) journals. In your replication exercise you should be learning from the best. The number of citations of an article on Google Scholar is another indicator of its importance to the discipline.
4. Your paper must use methods we have or will talk about in the course, or at about the same level of sophistication as the material we cover here.
5. You don't need to replicate every section in the original article. You can replicate only the part that you can justify as important (this may not be the part that the author considered important).
6. Do not choose an article unless you fully understand its argument, methods, and substance.
7. Your key task for the assignment is to improve and extend the analysis in one specific area. This should be reflected in your paper structure: replication of the original result should not take up more than a page or two of your paper, main emphasis should be on the substantive contribution and extension of the original results that you are making.
8. Improvements can include changing the way the original article dealt with missing data, selection bias, omitted variable bias, the model specification, the functional form, adding control variables or better measures, extending the time series and conducting out-of-sample tests, applying a better statistical model, etc.
9. By 10am in Week 10 upload a 2-page proposal outlining a clear replication and improvement plan to the corresponding folder on the Moodle page for the course. Your proposal should have a max 200-word abstract (just like any journal article that you read in this or any other modules) that contains a sentence on how replicating this article will help you with your dissertation project. In our Week 10 seminar we'll discuss your proposals.
10. Your paper should be proofread. Follow Harvard referencing style.
11. This assignment is a version of the replication paper exercise at Gary King's Gov2001 course at Harvard. For more details and guidance read Gary King "Publication, Publication," *PS: Political Science and Politics*, Vol. 39, No. 1 (January, 2006), 119-125. Updates are available here: http://gking.harvard.edu/papers/.

12. Follow Gary King's discussion in "Publication, Publication" for overall style, structure, and presentation of your paper.

13. Good examples of a replication exercise are (a) Gary King and Michael Laver "On Party Platforms, Mandates, and Government Spending" *American Political Science Review*; and (b) Bell, Mark and N. Miller "Questioning the Effect of Nuclear Weapons on Conflict" *Journal of Conflict Resolution*.

14. You should use R Markdown to write your replication project.

15. The final version of the paper is due on 25 April 2016. For the POLS3003 course, only an online submission is required. Please ensure you read the documents in the essay information folder on Moodle page for the course for full details on how to complete the submission. In addition you must also upload your complete replication package (R Markdown .Rmd file and full replication set) to the corresponding folder on Moodle.

The mark for your assessment consists of two parts: replication (50%) and extension (50%). The marks for replication and extension reflect how well you address points in the assessment description above.

Short Course Schedule

Below is a proximate schedule for the course. Some topics may need to be covered in more than one lecture. We will take as much time as needed on each topic, so we may not get to all the topics listed below.

Date	Topic	Details	Core Readings
Jan 12	Course overview and introduction to statistical learning	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 approaches to be covered. We will also discuss the concept of statistical learning	James et al. Chapters 1-2.
Jan 19	Linear regression	The linear regression model.	James et al. Chapter 3.
Jan 26	Classification	Logistic regression, LDA.	James et al. Chapter 4
Feb 2	Resampling Methods and Model Selection and Regularization	Cross-validation, bootstrap, shrinkage methods, ridge and lasso.	James et al. Chapters 5-6.
Feb 9	Non-linear Models and Tree-based Methods	GAMs, local regression, decision trees, random forests, boosting.	James et al. Chapters 7-8.
Feb 16	Reading week		
Feb 23	Support Vector Machines and Unsupervised Learning	SVM, principal components analysis, cluster analysis.	James et al Chapters 9-10.
Mar 1	Text analytics	Vector Space Model, tf-idf, Naive Bayes classifier, correspondence analysis, text clustering	Manning, Raghavan and Shutze Chapters 2, 6, 13-17.
Mar 8	Topic models	Latent Dirichlet Allocation, probabilistic topic models, dynamic topic models	Blei and Lafferty (2009)
Mar 15	Big Data	Hadoop, MapReduce, Pig	Kromer and Jurney Chapters 1-4
Mar 22	Hackathon		

Detailed Course Schedule

Tuesday, January 12: Course overview and introduction to statistical learning

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 approaches to be covered. We will also discuss the concept of statistical learning

Required Reading:

- James et al. Chapters 1-2.
- Hastie et al. Chapter 2.

Recommended Reading:

- Zumel and Mount Chapter 10.
- Kuhn and Johnson Chapter 1.

Tuesday, January 19: Linear regression

The linear regression model.

Required Reading:

- James et al. Chapter 3.
- Hastie et al. Chapter 3.2.

Recommended Reading:

- Lesmeister Chapter 2.
- Zumel and Mount Chapter 7.1.
- Kuhn and Johnson Chapters 5, 6.1-6.2.

Tuesday, January 26: Classification

Logistic regression, LDA.

Required Reading:

- James et al. Chapter 4.
- Hastie et al. Chapter 4.1-4.4.

Recommended Reading:

- Lesmeister Chapter 3.
- Zumel and Mount Chapter 7.2.
- Kuhn and Johnson Chapters 11, 12.1-12.3, 13.5-13.6.

Tuesday, February 2: Resampling methods and model selection and regularization

Cross-validation, bootstrap, shrinkage methods, ridge and lasso.

Required Reading:

- James et al. Chapter 5-6.
- Hastie et al. Chapter 3.3-3.5, 7.10-7.11.

Recommended Reading:

- Lesmeister Chapter 4.
- Kuhn and Johnson Chapters 4, 6.3-6.4, 12.4-12.5.

Tuesday, February 9: Non-linear models and tree-based methods

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

Required Reading:

- James et al. Chapter 7-8.
- Hastie et al. Chapter 9.1-9.4, 10.1.

Recommended Reading:

- Lesmeister Chapter 6.
- Zumel and Mount Chapter 9.1-9.3.
- Kuhn and Johnson Chapters 7.2, 8, 14.

Tuesday, February 23: Support Vector Machines and unsupervised learning

SVM, principal components analysis, correspondence analysis, cluster analysis.

Required Reading:

- James et al. Chapter 9-10.
- Hastie et al. Chapter 12.1-12.3, 14.3, 14.5.
- Leskovec et al. Chapter 11.

Recommended Reading:

- Lesmeister Chapter 5, 8-9.
- Zumel and Mount Chapters 8.1, 9.4.
- Kuhn and Johnson Chapters 7.3, 13.4.

Tuesday, March 1: Text analytics

Required Reading:

- Manning C., Raghavan P., and H. Shutze (2009). *An Introduction to Information Retrieval*. Chapters 2.2, 6.2-6.4, 13.1-13.4, 14-17. The book is available online from the authors' page: http://nlp.stanford.edu/IR-book/
- Jurafsky and Martin (2009) *Speech and Language Processing*, 2nd edition. Chapter 2.
- Laver, Michael, Kenneth Benoit and John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* 97: 311-331.
- Lowe, William. 2008. "Understanding Wordscores." Political Analysis 16(4): 356-371.
- Greenacre, M. (2007). *Correspondence Analysis in Practice*, 2nd edition. Appendix A & B.

Recommended 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*.
- Borcard, D., F. Gillet, P. Legendre (2011). *Numerical Ecology with R*. Springer.

Tuesday, March 8: Topic models

Required 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: http://www.cs.princeton.edu/~blei/papers/BleiLafferty2009.pdf.
- Lesmeister Chapter 12.

Recommended Reading:

- Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). "Latent dirichlet allocation." *Journal of Machine Learning Research* 3: 993-1022.
- 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).

Tuesday, March 15: Big data

Required Reading:

- Kromer, P. and Jurney, R. (2016). *Big Data for Chimps*. O'Reilly books. Chapters 1-4.
- Leskovec, Rajaraman and Ullman Chapter 2.

Recommended Reading:

- Lin, J. and Dyer, C. (2010). *Data-Intensive Text Processing with MapReduce*. Available from authors' page: https://lintool.github.io/MapReduceAlgorithms/MapReduce-book-final.pdf.

Tuesday, March 22: Hackathon

Required Reading:

- Zumel and Mount Chapters 1, 11.