**ME314: Introduction to Data Science and Machine Learning**

**Under Research Methods, Data Science, and Mathematics**

**The London School of Economics and Political Science**

Session Two 7/10-7/28

**Half Teaching / Half Tutoring**

May-Jul 6 hours each week x 6 weeks (36 hours)

7/10-7/28 90min each day x 13 weeks + 4 hours Exam Prep (24 hours)

Waive Preliminary Research and Prep (8 hours)

Payable to Bess Ho Weekly

**ME314 Course** https://www.lse.ac.uk/study-at-lse/summer-schools/summer-school/courses/research-methods/me314

**ME314 Course Outline** https://www.lse.ac.uk/ss-asset-library/course-outlines/2018/ME314-Course-Outline-2018.pdf

**Learning Languages**

R and SQL

**Learning Tools**

1. Terminal
2. GitHub
   1. Git Repository
   2. Git Command Line
   3. Git Desktop GUI
   4. GitHub Classroom
3. R Studio
4. SQLite
   1. SQLite Database
   2. Command Line Shell
   3. SQLDiff
   4. SQLite Analyzer
   5. SQLite Studio https://onecompiler.com/sqlite
5. DB Browser for SQLite https://sqlitebrowser.org/
6. Moodle (~Canvas)

**Git Lesson Plan**

**Excel Lesson Plan**

**R Lesson Plan**

**SQLite Lesson Plan**

SQLite Outline https://www.sqlitetutorial.net/

SQLite Online Editor   
https://extendsclass.com/sqlite-browser.html#

https://onecompiler.com/sqlite

https://www.mycompiler.io/view/ALjX79qiOUT

1. Simple Query [SELECT]
2. Sorting Rows [ORDER BY]
3. Filtering Data [DISTINCT] [WHERE] [LIMIT] [BETWEEN] [IN] [LIKE] [IS NULL]
4. Joining Tables (Inner Join, Left Join, Cross Join, Self Join, Full Outer Join)
5. Grouping Data [GROUP BY] [HAVING]
6. Set Operator [UNION] [EXCEPT] [INTERSECT]
7. Subquery (Subquery, Exist)
8. Querying Techniques [CASE]
9. Updating Data [INSERT] [UPDATE] [DELETE] [REPLACE]
10. Data Definition (Create, Alter, Rename, Drop, Vacuum)
11. Constraints (Primary Key, Not Null, Unique, Check, AutoIncrement)
12. Views (Create, Drop)
13. Indexes (Index, Index for Expressions)
14. Triggers
15. Full-Text Search
16. SQLite Tools (Commands, Show, Describe, Dump, Import CSV, Export CSV)

**R**

R Outline https://www.w3schools.com/r/r\_get\_started.asp  
R Online Editor   
https://www.jdoodle.com/execute-r-online/

https://www.mycompiler.io/new/r

**Primary Textbooks**

An Introduction to Statistical Learning: With applications in R by James et al. (2013) - Springer.

https://www.statlearning.com/  
R for Data Science by Garrett Grolemund and Hadley Wickham (2016) - O’Reilly Media.

https://r4ds.had.co.nz/index.html  
Practical Data Science with R by Zumel, N. and Mount, J. (2014) - Manning Publications.

**Additional Textbooks**  
Machine Learning with R by Lantz, B. (2013) - Packt Publishing.  
Mastering Machine Learning with R by Lesmeister, C. (2015) - Packt Publishing.  
Machine Learning for Hackers by Conway, D. and White, J. (2012) - O’Reilly Media.  
Mining of Massive Datasets by Leskovec, J., Rajaraman, A. and Ullman, J. (2011) - Cambridge  
University Press.  
Social Media Mining: An introduction by Zafarani, R., Abbasi, M. A. and Liu, H. (2014) = Cambridge University Press.

The Elements of Statistical Learning: Data mining, inference, and prediction by Hastie et al. (2009) - Springer.

https://hastie.su.domains/ElemStatLearn/

**Course Instruction**

https://lse-me314.github.io/instructions

**Course GitHub**

https://github.com/lse-me314/lse-me314.github.io

**Daily Exercise GitHub**

https://github.com/lse-me314

**Midterm - Week 5**

https://lse-me314.github.io/midterm-2018/

**Final Exam - Week 13**

https://lse-me314.github.io/finalexam/

**Additional Reading Assignments**

1. Lecture 1: Introduction to Data Science  
   R Software
   1. Patrick Burns, 2011. *The R Inferno*. Available [here](http://www.burns-stat.com/pages/Tutor/R_inferno.pdf).
   2. Lantz, Ch. 2.
2. Lecture 2: The Shape of Data  
   Statistical Datasets, RDB, DB Normalization, Data Restructure from "Wide" to "Long", Text Encoding, Date Formats, Sparse Matrix Format
   1. Reshaping data in Python: "[Reshaping and Pivot Tables](https://pandas.pydata.org/pandas-docs/stable/reshaping.html)".
   2. Robin Linderborg, "[Reshaping Data in Python](https://hackernoon.com/reshaping-data-in-python-fa27dda2ff77)", 20 Jan 2017.
3. Lecture 3: Working with Data  
   JSON, DB, SQL
   1. Lake, Peter. *Concise Guide to Databases: A Practical Introduction*. Springer, 2013. Chapters 4-5, Relational Databases and NoSQL databases.
   2. Nield, Thomas. *Getting Started with SQL: A hands-on approach for beginners*. O’Reilly, 2016. Entire text.
   3. [SQLite documentation](https://www.sqlite.org/docs.html).
   4. Bassett, L. 2015. [*Introduction to JavaScript Object Notation: A to-the-point Guide to JSON*](http://shop.oreilly.com/product/0636920041597.do). O'Reilly Media, Inc.
4. Lecture 4: Linear Regression  
   LR model, Supervised Learning
   1. Zumel and Mount, Chapter 7.1.
   2. Lantz, Chapter 6
5. Lecture 5: Classification  
   Lab 5 Midterm https://lse-me314.github.io/assignment05/  
   Logical Regression, Naive Bayes, Model Performance
   1. Lesmeister, Chapter 3.
   2. Zumel and Mount, Chapters 5, 6, 7.2.
   3. Lantz, Chapters 3-4, 10.
6. Lecture 6: Non-Linear Models and Tree-based methods  
   GAMs, Local Regression, Decision Trees, Random Forest, Bagging
   1. Lesmeister, Chapter 6.
   2. Zumel and Mount, Chapter 9.1-9.3.
   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.
7. Lecture 7: Resampling methods, model selection and regularization  
   Cross-validation, Bootstrap, Ridge, and Lasso
   1. Lesmeister, Chapter 4.
8. Lecture 8: Unsupervised Learning and Dimensional Reduction  
   Cluster Analysis, PCA
   1. Lesmeister, Chapter 5, 8-9.
   2. Zumel and Mount, Chapter 8.
   3. Lantz, Chapters 8-9
   4. Leskovec et al., Chapter 11.
9. Lecture 9: Text Analysis  
   Sentiment Analysis, Dictionary Methods
   1. Denny, M.J. and Spirling, A. (2018),[``Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It''](https://www.cambridge.org/core/journals/political-analysis/article/text-preprocessing-for-unsupervised-learning-why-it-matters-when-it-misleads-and-what-to-do-about-it/AA7D4DE0AA6AB208502515AE3EC6989E) *Political Analysis*
   2. Spirling, A. (2012), [``Democratization and Linguistic Complexity: The Effect of Franchise Extension on Parliamentary Discourse, 1832–1915.''](https://www.journals.uchicago.edu/doi/abs/10.1086/683612) *Journal of Politics*
   3. 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.
   4. Benoit, K., Munger, K., and Spirling, A. [``Measuring and Explaining Political Sophistication Through Textual Complexity.''](https://onlinelibrary.wiley.com/doi/full/10.1111/ajps.12423), *American Journal of Political Science* 63(2, April): 491–508. 10.1111/ajps.12423.
10. Lecture 10: Text Classification and Scaling  
    Naive Bayes Classifier, Wordscores, Wordfish
    1. [Statsoft, "Naive Bayes Classifier Introductory Overview."](http://www.statsoft.com/textbook/naive-bayes-classifier)
    2. An [online article by Paul Graham on classifying spam e-mail](http://www.paulgraham.com/spam.html).
    3. Bionicspirit.com, 9 Feb 2012, ["How to Build a Naive Bayes Classifier."](http://bionicspirit.com/blog/2012/02/09/howto-build-naive-bayes-classifier.html)
    4. Lowe, W. (2008). Understanding wordscores. Political Analysis, 16(4), 356-371.
    5. Benoit, Kenneth and Paul Nulty. 2013. “Classification Methods for Scaling Latent Political Traits.” Presented at the Annual Meeting of the Midwest Political Science Association, April 11–14, Chicago.
11. Lecture 11: Topic Modeling  
    Latent Dirichlet Allocation, Correlated Topic Model, Structural Topic Model
    1. 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.columbia.edu/~blei/papers/BleiLafferty2009.pdf).
    2. 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.
    3. Mimno, D. (April 2012). "Computational Historiography: Data Mining in a Century of Classics Journals." *Journal on Computing and Cultural Heritage*, 5 (1).
    4. Lesmeister Chapter 12.
12. Lecture 12: Data from the Web  
    Twitter API, Facebook API
    1. 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](https://github.com/lse-me314/lse-me314.github.io/blob/master/day11/Broniatowski.pdf)
    2. Barbera, Pablo., 2017. [``Less is more? How demographic sample weights can improve public opinion estimates based on Twitter data.''](http://pablobarbera.com/static/less-is-more.pdf) Working Paper
    3. Munger, Kevin., 2017. [``Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment''](https://link.springer.com/article/10.1007%2Fs11109-016-9373-5) Political Behaviour 39(3): 629-649
    4. Ginsberg et al., 2008. [``Detecting influenza epidemics using search engine query data''](https://www.nature.com/articles/nature07634) Nature 457: 1012–1014.
    5. Lazer et al., 2014. [``The Parable of Google Flu: Traps in Big Data Analysis''](http://science.sciencemag.org/content/343/6176/1203.full) Science 343: 1203-1205
    6. [Earthquake shakes Twitter users: real-time event detection by social sensors](http://www.ymatsuo.com/papers/www2010.pdf)
    7. <http://rcrastinate.blogspot.co.uk/2015/02/mapping-world-with-tweets-including-gif.html>
    8. <https://github.com/pablobarbera/streamR>
    9. Matthew Russell (2013). Mining the Social Web. O'Reilly Media. 2nd edition.
13. Lecture 13: Network Analysis
    1. Using Metadata to Find Paul Revere
    2. G. Erkan and D. Radev. 2004. “LexRank: Graph-based lexical centrality as salience in text summarization” Journal of Articial Intelligence Research 22, 457 - 479 Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. “Birds of a Feather:
    3. Homophily in Social Networks” Annual Review of Sociology 27, 415-44