Big Data Analytics
Department of Political Science
Fall 2018

Instructors:
張佑宗 Yutzung@ntu.edu.tw
李宣緯 waynelee1217@gmail.com & hwwaynelee@ntu.edu.tw
Office Hour: By appointments at Office 642
Course Time: Thursday 15:30-17:20
Place: Classroom 501

Teaching assistants:
黃譯民 r06322020@ntu.edu.tw
項柏翰 r06322019@ntu.edu.tw
陳家洋 marcus124144414@gmail.com
Office Hour: Tuesday 15:30-16:30 and Friday 13:30-14:30 (TBA)

Course Description:
Data science is a field with goals overlapping with many disciplines, in particular, mathematics, statistics, algorithms, engineering, or optimization theory. It also has wide applications to a number of scientific areas such as natural sciences, social sciences, life sciences, business, or medicine. Data science has become an integral part of many research projects and started affecting social science reaches. The promise of the "big data" revolution is that in these data are the answers to fundamental questions of businesses, governments, and social sciences such as political science and sociology. Most importantly, these quantitative techniques provide ``better predictions'' across different systems. Many of the most astonishing results come from computational fields, which have little experience with the difficulty of social scientific inquiry. As social scientists, we have an extensive experience and observations of our own research fields and we can utilize the advance of these new computational methods to our studies.

The course objective is to study the theory and practice of constructing algorithms that learn from data. This is an applied graduate level course for social scientists. Students will learn practical ways to build machine learning solutions for their own researches. While some mathematical/statistical details are needed, we will have an overview of the quantitative tools we need and emphasize the methods with their conceptual underpinnings rather than their theoretical properties. Specifically, the course will cover: k-nearest neighbors methods, the naive Bayes method, decision trees, random forests, boosting, k-means clustering and nearest neighbors, kernels, scaling, and ensemble learning. We will also discuss topics related to best practices, including overfitting/underfitting of data, error rates, cross-validation, and the use of bootstrapping methods to develop uncertainty estimates.

By the end of this course, students should be able to: (1) Understand the fundamental concepts and applications of data science. (2) Learn the advantages and shortcomings of widely used machine learning algorithms. (3) Uncover patterns and structure embedded in data with machine learning methods. (4) Test and improve model specification and predictions. (5) Apply their learning to a social science research project. As a result, we hope that this course will appeal not just to mathematicians/statisticians but also to researchers in a wide variety of social science research fields.

Main References:
There are two required books for the course:

1. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2009. Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani.

2. *Applied Predictive Modeling*. Springer, 2013. Max Kuhn and Kjell Johnson.

Further Readings:
Here are some recommended readings. Students are not required to read all of these books prior to class.

1. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009. Trevor Hastie, Robert Tibshirani, and Jerome Friedman.

2. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2013. Larry Wasserman.

3. *Python Machine Learning*. PACKT Publishing, 2015. Sebastian Raschka.

Prerequisites:
One-year of calculus, basic linear algebra, basic probability theory, applied statistics, proficiency in Python/R/MATLAB or permission of the instructors.

Grading Policy:

Quizzes	10%
Assignments	30%
Midterm	30%
Final Exam	30%

Assignments:
There will be 5-6 problem sets during the semester, with 3-5 questions apiece, drawn mostly from the two textbooks. The datasets we will be using, but not limited to, are mainly fields of social sciences and business. You are encouraged to discuss with your classmates about the problems, but you must write and turn in your own answers. To be blunt, rote copying of an answer from your classmates or other sources is a waste of your time and the grader's time.

Statistical Software:
R is a programming language and free software environment for statistical computing and graphics that is supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. In this class we will primarily use the open source statistical software R. Go to http://cran.r-project.org/ to download R for free. Data Analysis Assignments will be submitted in R Markdown (http://rmarkdown.rstudio.com/). If you are already well versed in other statistical/computational softwares, feel free to use them, but you'll be on your own.

Class Policy:

1. An important component of this course is active engagement with the material in classes. Regular attendance is essential and expected.
2. Quizzes are closed book, closed notes.
3. No makeup quizzes will be given.
4. No foods in class.

Academic Honesty:
Lack of knowledge of the academic honesty policy is not a reasonable explanation for a violation.

Class Topics and Readings:

Week 1
Overview of Data Science
Reading assignment: JWHT: Ch. 1; JK: Ch. 1
Optional Reading: *Finding a Place in Political Data Science.* Andrew Therriault, Democratic National Committee.

Week 2
Review Session 1
Reading assignment: JWHT: Ch. 2.3; JK: Appendix B
Optional Reading: *Google's R Style Guide.* Available at Google.

Week 3
Review Session 2
Reading assignment: JWHT: Ch. 2; JK: Ch. 2, 3, 4
Optional Reading: *We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together.* Justin Grimmer. Available at https://stanford.edu/~jgrimmer/bd_2.pdf

Week 4
Linear Regression 1
Reading assignment: JWHT: Ch. 3

Optional Reading: *The Cost of Racial Animus on a Black Presidential Candidate: Evidence Using Google Search Data.* Seth Stephens-Davidowitz. Journal of Public Economics. 118 : 26-40

Week 5
Linear Regression 2
Reading assignment: JK: Ch. 5, 6
Optional Reading: *Should I Use Fixed or Random Effects?* Tom S. Clark and Drew A. Linzer. Political Science and Research, 2015.

Week 6
Classification 1
Reading assignment: JWHT: Ch. 4
Optional Reading: *Tweet Sentiment: From Classification to Quantification.* Wei Gao and Fabrizio Sebastiani. Social Network Analysis and Mining, 2015.

Week 7
Classification 2
Reading assignment: JK: Ch. 11, 12
Optional Reading*: Automated Measurement of Mouse Social Behaviors Using Depth Sensing, Video Tracking and Machine Learning*. Weizhe Hong, Ann Kennedy, Xavier P. Burgos-Artizzu, Moriel Zelikowsky, Santiago G. Navonne, Pietro Perona, and David J. Anderson. Proceedings of the National Academy of Sciences of the United States of America, 2015.

Week 8
Resampling Methods
Reading assignment: JWHT: Ch. 5
Optional Reading: *Practical and Effective Approaches to Dealing with Clustered Data.* Justin Esarey and Andrew Menger. Political Science Research and Methods, 2018.

Week 9
Midterm Exam

Week 10
No class due to school anniversary

Week 11
Linear Model Selection
Reading assignment: JWHT: Ch. 6.1-6.3
Optional Reading: *Fraudulent Democracy? An Analysis of Argentinas Infamous Decade Using Supervised Machine Learning.* Francisco Cantu and Sebastian M. Saiegh. Political Analysis. 19: 409-433

Week 12
Regularization

Reading assignment: JWHT: Ch. 6.4-6.7
Optional Reading: *Enhancing Validity in Observational Settings When Replication is Not Possible.* Christopher J. Fariss and Zachary M. Jones. Political Science Research and Methods, 2017.

Week 13
Nonlinear Methods 1
Reading assignment: JWHT: Ch. 7
Optional Reading: *Bridging Analytical Approaches for Low-carbon Transitions.* Frank W. Geels, Frans Berkout, and Detlef P. van Vuuren. Nature Climate Change, 2016.

Week 14
Nonlinear Methods 2
Reading assignment: JK: Ch. 7
Optional Reading: *Studying User Income through Language, Behavior and Affect in Social Media.* Daniel Preoţiuc-Pietro, Svitlana Volkova , Vasileios Lampos , Yoram Bachrach, and Nikolaos Aletras. PLoS One, 2015.

Week 15
Tree Based Methods 1
Reading assignment: JWHT: Ch. 8
Optional Reading: *Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making.* Pete Burnap and Matthew L. Williams. Policy and Internet, 2015.

Week 16
Tree Based Methods 2
Reading assignment: JK: Ch. 8, 14
Optional Reading: *Improving Supreme Court Forecasting Using Boosted Decision Trees.* Aaron Russell Kaufman, Peter Kraft, and Maya Sen. Avalaible at http://www.aaronkauffman.com

Week 17
Unsupervised Learning
Reading assignment: JWHT: Ch. 10
Optional Reading: *Big Social Data Analytics in Journalism and Mass Communication.* Lei Guo, Chris J. Vargo, Zixuan Pan, Weicong Ding and Prakash Ishwar. Journalism and Mass Communication Quarterly, 2016.

Week 18
Final Exam