**Machine Learning for Social Scientists**
**POLI 175, Spring 2020**
Tuesday, Thursday 8:00am - 9:20am Pacific Time
Location: `https://ucsd.zoom.us/j/8599340921`

**Professor**: Kirk Bansak
Contact: kbansak@ucsd.edu
Office Hours: Wednesday, 9:00am - 12:00pm (sign-up details below)

**TA**: Bertrand Wilden
Contact: bwilden@ucsd.edu
Office Hours: Tuesday, 12:00pm - 2:00pm (or by appointment)

# 1 Overview

Social scientists and policymakers increasingly use large quantities of data to make decisions and test theories. For example, political campaigns use surveys, marketing data, and previous voting history to optimally target get out the vote drives. Governments deploy predictive algorithms in an attempt to optimize public policy processes and decisions. And political scientists use massive new data sets to measure the extent of partisan polarization in Congress, the sources and consequences of media bias, and the prevalence of discrimination in the workplace. Each of these examples, and many others, make use of statistical and algorithmic tools that distill large quantities of raw data into useful quantities of interest.

## 1.1 Objectives

This course introduces techniques to collect, analyze, and utilize large collections of data for social science inferences. The ultimate goal of the course is to introduce students to modern machine learning techniques and provide the skills necessary to apply the methods widely. In achieving this ultimate goal, students will also:

1) Learn about core concepts in machine learning and statistics, developing skills that are transferable to other types of data and inference problems.

2) Develop their programming abilities in `R`.

3) Be introduced to substantive problems and participate in challenges applying the techniques from the course.

## 1.2 Prerequisites

Previous background in `R` is a vital, required skill for enrolling in this course. All homework assignments must be completed using `R`, and all programming examples and exercises covered during class will use `R`.

In addition, students should have taken POLI170A or POLI171 or POLI172 (or have equivalent coursework). If you have any questions regarding whether you're prepared for the class, please talk to the teaching staff.

## 1.3 Evaluation

Students will be evaluated across the following areas.

**Homework** 30% of final grade. Students will complete four homework assignments. The assignments are intended to expand upon the lecture material and to help students develop the actual skills that will be useful for applied work. Homework should be completed and submitted using `R Markdown`, a markup language for producing well-formatted HTML documents with embedded `R` code and outputs. `R Markdown` requires installation of the `knitr` package. We recommend using `RStudio`, an IDE for `R`, which is set up well for the creation of `R Markdown` documents.

More about `RStudio` can be found here:
`http://www.rstudio.com`

`R Markdown` can be found here:
`http://rmarkdown.rstudio.com`

Students are encouraged to work on homework assignments together, **but must write up their code and answers on their own!** Submitting any code or answers that are copied from another student is unacceptable and in violation of academic integrity. If you work together, you must indicate on your assignments who your co-workers were.

**Challenges** 30% of final grade. Students will complete two team-based machine learning challenges during the course. The challenges will allow students to apply the techniques learned in the course to real problems. The teaching staff will provide more details and specify the guidelines for each challenge upon assignment. The challenges will be the following:

1. Predicting recidivism. We will provide you with a curated data set of criminal defendants. You'll work in teams to build and train models to predict which defendants are likely to commit another crime. In the process of building and evaluating your predictive models, you will also explore the value, risks, and ethics of applying machine learning in the criminal justice setting.

2. Analyzing political text data. We will provide you with a data set of text pertaining to a particular political event or figure. Working in teams, you will employ methods you have learned in the class (supervised and/or unsupervised) to analyze the text data and identify key insights that would have been difficult or impossible to discover without computational methods.

**Midterm Exam** 20% of final grade. Students will complete a midterm exam, valued at 20% of the final grade.

**Final Exam** 20% of final grade. Students will complete a cumulative final exam, at the time and date delineated under UCSD's Spring Quarter exam schedule.

# 2 Logistics

## 2.1 Class Meetings

All class meetings will be conducted online at our scheduled class time (Tuesday, Thursday 8:00am - 9:20am Pacific Time) via the video-conferencing platform Zoom. It is recommended that everyone tries to connect through our Zoom link in advance of our first class meeting to ensure the software is working properly. Our Zoom link for all class meetings (unless otherwise noted) is the following:

In addition, all class meetings will be recorded and posted on our `canvas.ucsd.edu` page. However, students are highly encouraged to attend class at the scheduled meeting time in order to have the opportunity to participate and ask questions in real time. Individuals who do not want to have their surroundings visible are encouraged to use Zoom's virtual background feature, if feasible, or to participate without video. Please also be mindful of others who may not wish to be visible or recorded in the background.

## 2.2 Office Hours

I will hold office hours from 9:00am to 12:00pm Pacific Time on Wednesday via Zoom. Please make sure to sign up for office hours in advance using the Calendly link. If you would like to meet but have class during my office hours, please email me to arrange an alternative time.

## 2.3 Teaching Assistant

Bertrand Wilden (`bwilden@ucsd.edu`) will be the teaching assistant for this course. He will be holding office hours on Tuesday from 12:00pm to 2:00pm (or by appointment). Please

email Bertrand to schedule a meeting time.

## 2.4 Course Website

As our primary course website, we will use Piazza.

You can sign up on the Piazza course page directly from the above address. There are also free Piazza apps for mobile devices.

We will distribute course materials, including lecture slides and problem sets on our Piazza website. There is also a question-and-answer platform that is easy to use and designed to get you answers to questions quickly. It supports code formatting, embedding of images, and attaching of files.

If you have non-personal questions related to course material or logistics, we encourage you to post these questions on Piazza rather than emailing the course instructors. Using Piazza will allow students to see and learn from other students' questions. Course instructors will regularly check the board and answer questions posted, although everyone else is also encouraged to contribute to the discussion and help answer questions. Please be respectful and constructive in your participation on the forum.

In addition, we will also use our `canvas.ucsd.edu` page for a few select functions, including to record/access grades and post lecture recordings.

## 2.5 Required Readings

The required readings throughout the course are listed in the course outline below. Each reading is associated with a class meeting on a specific date. It is recommended that you complete each reading *in advance of* the associated class meeting.

As our primary reference, we will use the book listed below:

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*, Seventh Printing, 2013.

This book is referred to as *ISLR* in the course outline. The book can be accessed online and downloaded for free here:
`http://www-bcf.usc.edu/~gareth/ISL/index.html`

In addition, readings from other sources are also assigned.

### 2.6 Other Recommended Reference Books

You might also consider the following books as useful references.

Murphy, Kevin P. *Machine Learning: A Probabilistic Perspective*. A slightly more advanced text, but an excellent treatment of machine learning methods.

Bishop, Christopher M. *Pattern Recognition and Machine Learning*. A more computer science oriented treatment of machine learning, with more extensive treatment of the estimation techniques used for machine learning methods.

## 3 Course Outline

| | | | |
|---|---|---|---|
| Introduction | Week 1 | 03/31 | Introduction |
| | Week 1 | 04/02 | A Machine Learning Focus on Regression<br>*Read: ISLR pp. 15 - 39* |
| Unit 1: **Supervised Learning** | Week 2 | 04/07 | The Bootstrap and Linear Regression<br>*Read: ISLR pp. 59 - 92, 187 - 190* |
| | Week 2 | 04/09 | Linear Regression/Classification<br>*Read: ISLR pp. 127 - 130*<br>**HW 1 assigned** |
| | Week 3 | 04/14 | Classification and Logistic Regression<br>*Read: ISLR pp. 127 - 138* |
| | Week 3 | 04/16 | Training, Testing, Bias-Variance Tradeoff<br>*Read: ISLR pp. 33 - 36, 176 - 178*<br>**HW 1 due, HW 2 assigned** |
| | Week 4 | 04/21 | LASSO and Ridge Regression<br>*Read: ISLR pp. 214 - 225, 227 - 228* |
| | Week 4 | 04/23 | Cross-Validation<br>*Read: ISLR pp. 175 - 186*<br>**HW 2 due, HW 3 assigned** |
| | Week 5 | 04/28 | Classification and Regression Trees<br>*Read: ISLR pp. 303 - 316* |
| | Week 5 | 04/30 | Random Forests and Boosted Trees<br>*Read: ISLR pp. 316 - 324*<br>**HW 3 due, Challenge 1 assigned** |
| | Week 6 | 05/05 | Evaluating and Selecting Models<br>*Read: Drakos (2018) Part 1; Drakos (2018) Part 3* |
| | Week 6 | 05/07 | Machine Learning, Policy, and Ethics |

| | | | |
|---|---|---|---|
| | | | *Read: Kleinberg et al. (2016); Buchanan and Miller (2017)*<br>**Challenge 1 due** |
| | Week 7 | 05/12 | **Midterm** |
| Unit 2: **Unsupervised Learning** | Week 7 | 05/14 | Intro to Unsupervised Learning & Text as Data<br>*Read: ISLR pp. 373 - 374; Grimmer and Stewart (2013) pp. 1-7* |
| | Week 8 | 05/19 | Text Analysis and Dictionary Methods<br>*Read: Grimmer and Stewart (2013) pp. 7-14; Loughran and McDonald (2011)* |
| | Week 8 | 05/21 | Distance, Clustering, and Text Applications<br>*Read: ISLR pp. 385 - 401; Grimmer and Stewart (2013) pp. 14-17*<br>**HW 4 assigned** |
| | Week 9 | 05/26 | Topic Models for Text Analysis<br>*Read: Mohr and Bogdanov (2013); Grimmer and Stewart (2013) pp. 17-25* |
| | Week 9 | 05/28 | Principal Components Analysis<br>*Read: ISLR pp. 374 - 385*<br>**HW 4 due, Challenge 2 assigned** |
| | Week 10 | 06/02 | Other Topics in Dimensionality Reduction<br>*Read: TBD* |
| Conclusions | Week 10 | 06/04 | Review and Next Steps<br>**Challenge 2 due** |
| | | 06/09 | **Final Exam, 8:00am - 11:00am PT** |

## 3.1 Readings

- **ISLR**: Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Seventh Printing, 2013.
- **Buchanan and Miller (2017)**: Ben Buchanan and Taylor Miller, "Machine Learning for Policymakers: What It Is and Why It Matters," Belfer Center for Science and International Affairs, 2017.
- **Drakos (2018) Part 1**: Georgios Drakos, "How to Select the Right Evaluation Metric for Machine Learning Models: Part 1 Regression Metrics," Towards Data Science, *Medium*, August 26, 2018. Available at: https://towardsdatascience.com/how-to-select-the-right-evaluation-metric-for-machine-learning-models-part-1-regrression-metrics-3606e25beae0

- **Drakos (2018) Part 3**: Georgios Drakos, "How to Select the Right Evaluation Metric for Machine Learning Models: Part 3 Classification Metrics," Towards Data Science, *Medium*, September 12, 2018. Available at: https://towardsdatascience.com/how-to-select-the-right-evaluation-metric-for-machine-learning-models-part-3-classification-3eac420ec991

- **Grimmer and Stewart (2013)**: Justin Grimmer and Brandon M. Stewart, "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Political Analysis* Vol. 21, No. 3 (2013).

- **Kleinberg et al. (2016)**: Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, "A Guide to Solving Social Problems with Machine Learning," *Harvard Business Review*, December 8, 2016. Available at: https://hbr.org/2016/12/a-guide-to-solving-social-problems-with-machine-learning

- **Loughran and McDonald (2011)**: Tim Loughran and Bill McDonald, "When is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *The Journal of Finance* Vol. 66, No. 1 (2011).

- **Mohr and Bogdanov (2013)**: John W. Mohr and Petko Bogdanov, "Introduction–Topic Models: What They Are and Why They Matter," *Poetics* Vol. 41, No. 6 (2013).

# 4 Additional Information

## 4.1 Students with Disabilities

Students requesting accommodations for this course due to a disability must provide a current Authorization for Accommodation (AFA) letter issued by the Office for Students with Disabilities (`https://osd.ucsd.edu/`). Students are required to discuss accommodation arrangements with instructors and OSD liaisons in the department well in advance of any exams or assignments. The OSD Liaison for the Department of Political Science is Joanna Peralta; please connect with her via the Virtual Advising Center (`https://stark.ucsd.edu/students/vac/`) as soon as possible.

## 4.2 Academic Advising

Students who have questions pertaining to Political Science academic advising are asked to reach out the Department's Undergraduate Advisor, Natalie Ikker, who can be reached via the Virtual Advising Center (`https://stark.ucsd.edu/students/vac/`). Academic advising questions often include (but not limited to): add/drop deadlines, course enrollment policies, planning major and minor requirements, quarter-by-quarter plans, department petitions and paperwork, and referrals to campus and student support services.

## 4.3 UC San Diego Principles of Community

The University of California, San Diego is dedicated to learning, teaching, and serving society through education, research, and public service. Our international reputation for excellence is due in large part to the cooperative and entrepreneurial nature of the UC San Diego community. UC San Diego faculty, staff, and students are encouraged to be creative and are rewarded for individual as well as collaborative achievements.

To foster the best possible working and learning environment, UC San Diego strives to maintain a climate of fairness, cooperation, and professionalism. These principles of community are vital to the success of the University and the well being of its constituents. UC San Diego faculty, staff, and students are expected to practice these basic principles as individuals and in groups.

For the complete UC San Diego Principles of Community, see:
`https://ucsd.edu/about/principles.html`

## 4.4 Resources

| | |
|---|---|
| Library Help and Research Tools: | https://library.ucsd.edu/ask-us/triton-ed.html |
| Writing Hub: | https://commons.ucsd.edu/students/writing/index.html |
| Supplemental Instruction: | https://commons.ucsd.edu/academic-support/supplemental-instruction/si-students.html |
| Tutoring: | https://commons.ucsd.edu/academic-support/content-tutoring/index.html |
| Mental Health Services: | https://caps.ucsd.edu |
| Community Centers: | Learn about the different ways UC San Diego explores, supports, and celebrates the many cultures in our diverse community. https://students.ucsd.edu/student-life/diversity/index.html |
| Accessibility: | https://disabilities.ucsd.edu/ |
| Basic Needs: | Any student who has difficulty accessing sufficient food to eat every day, or who lacks a safe and stable place to live, and believes this may affect their performance in this course, is encouraged to contact: foodpantry@ucsd.edu and basicneeds@ucsd.edu |