# Machine Learning, Text Analysis, and Economics

**Econ 1680**
**Spring 2022**
**Brown University**
**Department of Economics**

## Course Information

Lecture Time: Mondays, 3:00pm-5:30pm
Lecture Location: Friedman Hall 101
Review Session: Thursdays, 9:00-10:00am
Review Location: 67 George St, room 104
Course Website: canvas.brown.edu
Github Classroom Website: https://classroom.github.com/classrooms/

## Instructor Information

**Professor Amy Handlan**
Email: amy_handlan@brown.edu
Website: sites.google.com/view/amy-handlan
Office: 70 Waterman St, Room 202
Office Hours by Appointment: Wednesdays, 1:30pm-3:30pm, Sign-up link [click here], in-person or over Zoom

**Teaching Assistant: Haoyu Sheng**
Email: haoyu_sheng@brown.edu
Open Office Hours: Tuesdays, 3:00pm-4:00pm, over Zoom: brown.zoom.us/my/haoyusheng
Office Hours by Appointment: Tuesdays, 4:00pm-5:00pm, Sign-up link [click here], over Zoom

## Course Overview

Economists need advanced methods to study data that is complex, high-dimensional, and unstructured. The goal of this course is to highlight the key challenges of working with such data in economics and what machine learning and text analysis methods can be used to address them. Private sector companies, government agencies, and economics graduate programs are all looking for individuals with an understanding of these advanced methods and experience in applying them to real-world problems. This course is designed to help to meet that demand.

This course will meet once a week. Lectures will introduce students to new material and include discussions of current economics research using machine learning and text analysis methods. They will also include applied exercises demonstrating methods covered in class and will focus on developing writing assignments and providing peer feedback.

## Course Objectives

By the end of this course, students will be familiar with a variety of machine learning and text analysis methods and their applications in economics. Students will be able to describe the methods covered in class and assess the appropriate use of these methods given different objectives and data conditions. Students will document their applications of multiple machine learning and text analysis methods in the writing assignments for the course. This class will help students build a foundation in these advanced methods and a project portfolio to help them pursue careers at the intersection of economics and data science.

## Course Prerequisites

Prerequisites are Econ 1110 or 1130; Econ 1629 or 1630; and an introductory computer science course (CS 040, 111, 112, 150, 170, or 190). While a university course specifically in Python is not required, students are expected to be either proficient in Python or willing to dedicate time to learning Python in parallel with the course. Knowledge of econometrics and programming is assumed.

## Course Materials

Readings for the course will include research papers and the books below. The main books we will be using are in bold. The list of research papers will be updated on the course website. Students will also need access to a computer with the internet and Python software.

**Textbooks**

Bengfort, Bilbro, and Ojeda. "Applied Text Analysis with Python." First Edition. 2018. ISBN: 1491963042.

Bird, Klein, and Loper. "Natural Language Processing with Python." First Edition. 2009. ISBN: 0596516495. The updated textbook is available at: http://www.nltk.org/book/

Goodfellow, Bengio, and Courville. "Deep Learning." 2016. ISBN: 9780262035613.
The textbook is available at: https://www.deeplearningbook.org/

**Grimmer, Roberts, and Stewart. "Text as Data: A New Framework for Machine Learning and the Social Sciences." 2022. ISBN: 0691207550. (This book is only available as an ebook right now, the physical copy will be available in March)**

**Hastie, Tibshirani, and Friedman. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction." 2nd Edition. 2017. ISBN: 0387848576. The textbook is available at: http://statweb.stanford.edu/~tibs/ElemStatLearn/**

Nielsen, Michael A. “Neural Networks and Deep Learning”, Determination Press, 2015. The textbook is available at: http://neuralnetworksanddeeplearning.com/index.html

**Papers**
Athey, Susan and Guido Imbens. 2019. “Machine Learning Methods Economists Should Know About.” arXiv:1903.10075. URL: https://arxiv.org/abs/1903.10075

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. “Text as Data.” *Journal of Economic Literature*, 57 (3): 535-74. URL: https://www.aeaweb.org/articles?id=10.1257/jel.20181020

Grimmer, Justin, and Brandon Stewart. 2013. “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts.” *Political Analysis,* pp. 1-13. URL: https://scholar.princeton.edu/sites/default/files/bstewart/files/tad2.pdf

Grimmer, Justin, Margaret Roberts, and Brandon Stewart. 2021. “Machine Learning for Social Science: An Agnostic Approach.” *Annual Review of Political Science.* 24: 395-419. URL: https://www.annualreviews.org/doi/pdf/10.1146/annurev-polisci-053119-015921

Additional papers will be added to the course website...

**Software**
We will be using Python as the programming language for the course. Python is open source and freely available online. The language has many machine learning and text analysis packages and strong online support documentation. While we will use Python in course notes, some tasks will be doable in other languages like R, Julia, MATLAB, and Stata.

## Course Grading

The course will be graded based on participation, homework, two writing-intensive projects, and a presentation.

**Grade Breakdown**
The final grade for the course will be weighted according to the following:

- 21% - Homework Assignments
- 35% - Writing Assignment 1
- 35% - Writing Assignment 2
- 7% - Presentation
- 2% - Class Participation

**Academic Honesty**
Submitting academic work that uses others' ideas, words, research, or images without proper attribution and documentation is a violation of Brown's academic code:
*"A student's name on any exercise (e.g., a theme, report, notebook, performance, computer program, course paper, quiz, or examination) is regarded as assurance that the exercise is the result of the student's own thoughts and study, stated in his or her own words, and produced without assistance, except as quotation marks, references, and footnotes acknowledge the use of printed sources or other outside help."* (Academic Code, p. 5)

**Regrade Policy**
Requests for reconsideration of grades are not encouraged and will be accepted only in writing, with a clear statement of what has been misgraded, within one week of receiving the graded assignment. Please submit your full assignment so grading on the entire assignment can be reconsidered.

## Course-Related Work Expectations

Over the spring semester, students will spend 2.5 hours per week in class (32 hours total). Required reading for the class meetings is expected to take up approximately 4 hours per week (56 hours). Students are expected to attend the review sessions with the teaching assistant which is 1 hour per week (14 hours total). Over the semester, the homework assignments will total approximately 30 hours and the projects will take approximately 50 hours.

## Homework Assignments

There will be three homework assignments for the course. These assignments are meant to complement material in lectures and provide a foundation for students to do well in their projects. Students will download the assignments from the course webpage. Student answers must be typed and formatted according to the course template and submitted to the course website. Remember to cite all your sources, including fellow classmates and others' code. Be sure to show all work for analytical problems. Short answer and essay questions should be answered with complete sentences and thorough explanations. Late homework will not be accepted. Assignment extensions must be requested in advance of assignment deadlines and will be granted according to university policy.

## Writing Assignments (Projects)

This class has two projects that will involve coding and a complimentary written analysis. Employers want to hire people who not only understand the advanced methods in this class, but can also explain those methods, their applications, and interpret the results.

There are two writing assignments (two projects) for this course:

1. Machine Learning Project
2. Text Analysis Project

Each project will total 35% of the final grade and that total is broken down in the following way:

- Topic Proposal (5%)
- Detailed Outline/Draft (10%)
- Final Paper (20%)

The scaffolding of the projects is meant to provide feedback throughout the projects' development.

**Topic Proposal**

Students must first enter their desired topic in the topics spreadsheet (link on the website), making sure no other student has chosen the same topic. Duplicate topics will not be accepted and are assigned on a first-come-first-serve basis. The guidelines for choosing a topic are that it must relate to economics and the methodology to study your topic must be among the methods covered in class. The objective here is to have students choose topics that they are passionate about and can help them develop skills for their specific careers. Once students have submitted their desired topic, they must then contact the teaching assistant to discuss the topic and methods to verify that it is suitable for the purposes of this course. These meetings will take place during a particular recitation section (see schedule below). You should move forward on your writing assignment only after the teaching assistant has approved your topic.

**Detailed Outline/Rough Draft**

The outline should cover the necessary components as listed on the course website. At this stage, your draft should allow you to answer the following: What are you doing? Why are you doing it? How are you doing it? The outline will have the following sections:

- Introduction
- Data Sources
- Methodology
- Results or Expected Results
- Conclusion
- Bibliography

Within each section, you should include:

- Lists of topics and concepts to discuss in each section
- Citations and descriptions for the methodology, data, topic of the project

There is no page minimum for the draft. Students are expected to sufficiently cover the above sections. The final paper will have a page maximum, but students can worry about condensing their writing later. The goal at this stage is to get the necessary components on paper. In class,

students will discuss each others' outlines, ask questions, and provide suggestions. In addition to peer feedback, students will submit their work for review by the professor and the teaching assistant.

**Rubric for writing assignments:**

| | Excellent | Good | Poor |
|---|---|---|---|
| Engagement (40 points) | Deep engagement with the problems. Attempts to conclusively answer deeper questions (puzzles) that arise. When problems cannot be solved with textbook approaches, generates new approaches to solve them. Comments on the desirability of these approaches. (35-40 points) | Good engagement with the problems. Attempts to solve deeper questions that arise, but does not go out of its way to do so. (20-34 points) | Does not engage with all problems, or misses the presence of deeper questions. (19 or below) |
| Understanding (40 points) | Writing displays a clear understanding of the problem, the methods, and associated limitations. (35-40 points) | Writing is unclear or does not demonstrate full understanding of the problem/methods/limitations. (20-34 points) | Writing is unclear or demonstrates substantial misconceptions. (19 or below) |
| Writing Basics (20 points) | Well written: clear and understandable. (18-20) | Writing is understandable. (13-17) | Does not follow proper structure, includes errors, or poorly written. (12 or below) |

**Final Paper**

The final paper must incorporate comments from the draft and possess the following characteristics:

- Content from the outline/draft should be flushed out to be in complete sentences and paragraphs.
- Given that students concisely discuss the above, the text of the paper should come to be about 1000 - 1500 words (this does not include graphs, tables, title page, or references).
- References are correctly cited throughout the paper and in the references page using Chicago style.

- Tables and graphs should all be at the end of the paper

The objective of these papers is to have no unnecessary text and to train students to succinctly write about an economic question, their data sources, the methodology needed to answer their question and to highlight the results. These papers are meant to be short but contain a lot of information. Mastering writing in this style is difficult. Nevertheless, it will greatly serve students in their postgraduate careers whether that be academia, industry, or policy.

## Presentation

At the end of the semester, students will choose one of their two projects to present to the class. Students will present their research question, the method they chose to address that question, their data sources, and summarize their findings. These presentations are meant to be short and efficient. The goal of the presentation is to help prepare students for similar types of presentations in their future interviews and careers where they need to share their work with supervisors, clients, professors, advisors, policymakers, or the public.

## Class Participation

Attendance is mandatory. Students are expected to participate in class meetings by:

- Attending class prepared for the day's lesson (up-to-date on assignments and readings)
- Actively listening and only using electronics for course activities (no cell phones)
- Commenting and questioning in class discussions on economics research and methods
- Providing peer feedback on writing assignments and student presentations

Students will be graded according to their engagement with the material on these dimensions.

## Guidance on Covid-19

This course will be taught in person from the start of the semester. If I must be absent during the semester, we may temporarily move the course to be remote, a substitute instructor may cover the class, or we may reschedule the class to a later date if necessary.

Students must follow university policy with regards to wearing masks, testing, and vaccination. Students should NOT attend class if they are feeling ill or have tested positive for Covid-19. Students who are sick and unable to attend class may request a Zoom link from the professor to live stream the course. Participation for these students will be evaluated by virtual attendance and engagement with online resources. If students miss a peer feedback and project development session, they are to email the professor/teaching assistant to be paired with someone virtually. Assignment extensions must be requested in advance of assignment deadlines and will be granted according to university policy. Deans in both the College and Student Support Services

may provide deans' notes on behalf of undergraduate students with specific personal or health circumstances that hinder their ability to complete academic work.

## Tentative Course Schedule

The official course schedule will be maintained on the course website. Readings are abbreviated by initials of author surnames.

| Week | Meeting Topic (Mondays) | Readings | Due (Fridays) |
|---|---|---|---|
| 1 | Intro and Econ MLTA Overview | GRS (2021); GS (2013); HTF 1; GBC 2-5; NLTK 0; BBO 1; Opt'l: AI (2019); GKT (2019) | |
| 2 | Regression/Classification: Linear, Logit, Polynomial Expansions, Additive Models, and Dimensionality | HTF 2-3,4.1,4.4, 5.1-5.2, 9.1; Løken, Mogstad, and Wiswall (2012) | HW1 |
| 3 | Dimension reduction and clustering | HTF 14,18 | Topic 1 |
| 4 | President's Day | - | |
| 5 | Neural Networks | HTF 11, GBC 6, 11 | HW2 |
| 6 | LDA and SVM + Project Development | HTF 12, 4.3,4.5 | Draft 1 |
| 7 | Text Analysis Intro-Word and Context Representations | NLTK, BBO 3-4 | Topic 2 |
| 8 | BOW: Sentiment Lists, TFIDF, Ngrams | BBO 5, 7 | Project 1 |
| 9 | Spring Break | - | |
| 10 | Topics, Clustering, and Similarity | BBO 6 | HW3 |
| 11 | *Virtual class: Project Development+ Guest Speaker (Dr. Marin-Aranega) | - | Draft 2 |
| 12 | Text Classification/Regression | BBO 12 | |
| 13 | Advanced Topics | TBD | Project 2 |
| 14 | Presentations (9min) | - | Slides Due |
| 15 | Presentations | - | |

## Acknowledgments

I thank Professor Daniel Björkegren for sharing his course materials. I also thank Haoyu Sheng for his assistance in developing course materials.

## Academic Support

**Accessibility and Accommodations Statement**
Brown University is committed to full inclusion of all students. Please inform me early in the term if you may require accommodations or modification of any of course procedures. You may speak with me after class, during office hours, or by appointment. If you need accommodations around online learning or in classroom accommodations, please be sure to reach out to Student Accessibility Services (SAS) for their assistance (seas@brown.edu, 401-863-9588). Undergraduates in need of short-term academic advice or support can contact an academic dean in the College by emailing college@brown.edu. Graduate students may contact one of the deans in the Graduate School by emailing graduate_school@brown.edu.

**Campus resources**
CAPS/SEAS/Academic Deans/Identity Centers/BWell/Chaplains Office/ Student Support Services Deans can be a helpful resource to discuss personal, family or health-related concerns, as well as a potential academic and personal plan.

**Writing Resources**
Brown's Writing Center provides support to students at every stage of the writing process. Associates at the Writing Center can help students, "with writer's block, audience awareness, argumentation, organization, grammar, research skills, the conventions of academic writing, English as a Second Language, and issues of clarity and style" (Writing Center Website).

Brown welcomes students from around the country and the world, and their unique perspectives enrich our learning community. To support students whose primary language is not English, an array of English support services are available on campus including language and culture workshops and individual appointments. For more information, contact english-support@brown.edu or (401) 863-5672.

**Course Expenses**
If your Brown undergraduate financial aid package includes the Book/Course Material Support Pilot Program (BCMS), concerns or questions about the cost of books and course materials for this or any other Brown course (including RISD courses via cross-registration) can be addressed to bcms@brown.edu. For all other concerns related to non-tuition course-related expenses, whether or not your Brown undergraduate financial aid package includes BCMS, please visit the

Academic Emergency Fund in E-GAP (within the umbrella of "E-Gap Funds" in UFunds) to determine options for financing these costs, while ensuring your privacy.