THE UNIVERSITY OF CHICAGO
HARRIS SCHOOL OF PUBLIC POLICY

# PPHA 30545: Machine Learning for Public Policy

**Dr. Christopher Clapp**
**Dr. Jeff Levy**

**Syllabus, Winter 2025**

| | |
|---|---|
| Class Meetings (Clapp): | |
| Section 01 - MW 10:30am-11:50am | Keller 0007 |
| Section 02 - MW 01:30pm-02:50pm | Keller 0021 |
| Section 03 - TTh 09:30am-10:50am | Keller 0021 |
| Section 04 - TTh 11:00am-12:20pm | Keller 0021 |
| Class Meetings (Levy): | |
| Section 05 - TTh 02:00pm-03:20pm | Keller 0021 |
| Section 06 - TTh 03:30pm-04:50pm | Keller 0021 |
| Lab Sessions: | |
| Lab 01 (R) - F 01:30pm-02:50pm | Keller 2112 |
| Lab 02 (Python) - F 09:00am-10:20am | Keller 2112 |
| Lab 01 (Python) - F 01:30pm-02:50pm | Keller 1022 |
| Professors: | |
| Chris Clapp (he/him) | Keller 3039 |
| Jeff Levy (he/him) | Keller 3101 |

TAs: Ari Ainsfeld (Head TA; he/him), Steve Kim (Head TA; he/him), Alison Spencer (she/her), Andre Oviedo Mendoza (he/him), Arkadeep Bandyopadhyay (he/him), Michael O'Neil (he/him), Praveen Chandar Devarajan (he/him), Shuping Wu (he/him), Qilin Zhou (she/her)

## Course Description

It's an exciting time to study machine learning and data science more generally! We live in a digital era where many of our decisions and actions are tracked. Information is being produced and recorded at a stifling pace. While this may not seem novel to those who were born and have grown up in the Information Age, the amount of data available to researchers and policymakers is much more than what existed even a decade ago. Coupled with cheap computing power and expanded data storage, recent developments across statistics, computer science, and data-driven social sciences allow us to use all this data in a myriad of interesting ways. But what questions will we seek to answer with this newly available *big data* and these newly developed *machine learning* tools?

While these tools are already being used extensively in marketing, finance, and business, their application to public policy is in its infancy (despite the techniques being the same across disciplines). Early examples of questions with policy implications include: can we predict unavailable data we take for granted in the developed world from available information in a developing world context? Is it possible to improve the accuracy of judges' bail decisions that hinge on whether the accused will commit additional crimes? Or can we inform doctors about the trade-offs inherent in prescribing potentially addictive opioids to patients for short-term pain relief by predicting who is likely to develop an addiction in the long run?

In order to ask and inform questions like these, this class will introduce you to ways to detect patterns in data, then use what you have learned to predict important outcomes or describe the salient relationships among inputs. While this requires an understanding of how and why these tools work, we will emphasize the intuition and application of these techniques over their theoretical underpinnings. We will do so by exploring nascent, policy-relevant applications of these methods, but, ultimately, the full impact of how these machine learning techniques inform and influence policy has yet to be determined. That's up to you!

## Learning Objectives: "What's My Incentive for Taking This Course?"

Specifically, the purpose of the course is to introduce you to a wide array of the fundamental methods in modern machine learning. Each week, we will learn about and discuss a different set of techniques and their applications to public policy during lecture sections. During lab sessions, you will gain experience with those techniques by coding their implementation in Python or R.

Along the way you can expect to:

- Apply machine learning techniques to carry out policy-relevant analyses.
- Understand how the machine learning approach, which focuses on prediction, differs from the approach to fundamental statistical and/or causal inference you learned in Harris' core statistics classes.
- Gain an appreciation of why the bias-variance trade-off makes prediction inherently difficult.
- Recognize the different ways "long" and "wide" big data allow us to improve our predictions.
- Work with (often messy) real-world data.
- Continue developing your coding skills in Python or R as you learn new tools.
- Visualize, interpret, and convey your findings to audiences of different levels of technical sophistication.

The overall course objective is for you to be able to use machine learning tools to inform better policy and make the world a better place, as well as to become an informed and critical consumer of policy recommendations based on machine learning techniques. Additionally, the course will allow you to market your newly gained machine learning knowledge and skills when applying for jobs.

## Prerequisites

In addition to assuming knowledge of the material in the "Harris Core" classes (Statistics for Data Analysis I & II), the prerequisites are:

- PPHA 30535/7 Data and Programming for Public Policy I and
- PPHA 30536/8 Data and Programming for Public Policy II.

This course is the third installment of the three-quarter core sequence of the Specialization in Data Analytics (https://harris.uchicago.edu/academics/design-your-path/specializations/specialization-data-analytics) at Harris. Students at Harris and from other parts of the University may enroll without having taken previous courses in the sequence *after students who have taken the prerequisites have had a chance to enroll*. However, it is necessary for MPP students to take the full sequence in order to meet the necessary requirements for the specialization.

**Considering the Course without the Prerequisites?** For anyone who has not taken the prerequisites and is considering taking this course, first, thanks for your interest in our class! This course introduces machine learning techniques, then has students practice and apply them via coding-based labs, problem sets, and mini-projects. Based on this overview, there are two things you should consider. First, all machine learning is based on statistics, so it is in your best interest to have completed the "Stats Core" before taking this class. Second, while the class doesn't directly follow the prerequisites (which teach general coding skills in Python/R), you will be responsible for knowledge of the material covered in those classes. It is in your best interest to take this class out of sequence only if you have sufficient experience coding in Python/R *and* are aware that they may be at a bit of a disadvantage relative to the majority of the students in the class who have taken the prerequisites.

If you are considering taking the class without having taken the prerequisites, we recommend looking over the syllabi for those classes and making sure that you're comfortable with the topics and techniques that are covered before making your decision on whether or not to enroll. It may also be helpful to take a look at the textbook, which is available online (for free; see the "Materials" section of this syllabus for details on how to access it).

## Evaluation

Your final grade in this course will be related to performance in several areas. The weight placed on each component will be as follows:

| | |
|---|---|
| Attendance | 05% |
| Problem Sets (4) | 40% |
| Mini-Projects (4) | 40% |
| Final Exam | 15% |
| Participation (Extra Credit) | 02% |

**Attendance** As per Harris policy (https://harris.uchicago.edu/student-life/dean-of-students-office/policies), regular attendance is required of all Harris students. It is also necessary (but not sufficient) to do well in the class. We will take attendance in each class. Please be sure to display your name tent in every class to facilitate the taking of attendance. Every absence beyond the first two will reduce your course grade by one percentage point, up to a maximum of five percentage points. Per the Harris policy, we will notify the Dean of Students about students who are chronically absent.

**Problem Sets and Mini-Projects** There are four problem sets and four mini-projects in this class. All assignments will be submitted on Canvas via the Gradescope option. You may submit assignments late for up to 24 hours after the due date with a four percentage point deduction per hour.[1] I will drop the lowest grade among these assignments when calculating your grade.

Problem sets will consist of more structured questions (primarily) from the textbook. They are designed to help students cement their understanding of the conceptual material covered in lecture and get practice both applying the tools we learn and with coding.

Mini-projects are designed to apply the machine learning concepts and tools covered in class to policy-relevant questions. As such, they are less structured, based on real-world data, and emphasize application to public policy over statistical concepts.

[1] These deductions are not fractional (e.g. turning an assignment in one second or 59 minutes and 59 seconds late will result in an four percentage point deduction).

You are welcome (and encouraged) to form study groups of no more than 2 students to work on the problem sets and mini-projects together. But you must write your own code and your own solutions. Please be sure to include the names of those in your group on your submission. Please also be sure to practice the good coding practices you learned in the Data and Programming classes and comment your code, cite any sources you consult, etc.[2]

**Final Exam** We will have an in-person final exam during finals week covering all material from the quarter. The final exam schedule will be published by Harris Registration in the first three weeks of the quarter. As per Harris policy, there will be no make-up final scheduled without an official accommodation from the Academic and Student Affairs team.

**Class Participation** Class participation points will be based on your level of active, attentive, inquisitive participation during in-class discussions and/or on the discussion board. For in-class participation, note that regular class attendance is generally a necessary (but not sufficient) component of earning in-class participation points. Additionally, to earn credit, you must record each instance of your participation (e.g., when you ask a question, provide an answer, contribute to a class discussion, etc.) using the submission form linked on the main Canvas course page.[3] Please submit a separate entry each time you participate. You only need a brief description of your question/answer/etc. (enough to jog my memory) and you should record all participation within 24 hours after class ends. You do not need to record participation via the discussion board - just your in-class participation!

We will supplement in-class participation with the Ed Discussion discussion board on Canvas. Please use the discussion board to post questions, discuss the material covered in the lectures or on the assignments, and answer questions posed by your peers. As being a good colleague is both an important way to have social impact and is valued by employers, participation points can be earned by making posts that are helpful to your peers.[4] While this can take many forms, points will primarily be awarded for answering classmates' questions on the discussion board. In doing so, you may not explicitly share code, provide step-by-step solution algorithms (e.g., pseudo code), or direct solutions. You may clarify ambiguities in the assignments, discuss conceptual aspects of lectures or problems, show output and error messages, and provide general guidance on how to correct errors in understanding or code.[5] Additionally, you may post brief summaries of news articles that describe applications of machine learning techniques to public policy relevant issues.[6]

## Grades

This class requires a 60% or above to pass and is not curved. All passing letter grades will be determined based on the following intervals used in the Data Science Specialization sequence:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| A | $[95\% - 102\%]$ | A- | $[90\% - 95\%)$ | B+ | $[85\% - 90\%)$ | B | $[80\% - 85\%)$ | B- | $[60\% - 80\%)$ |

[2] The focus of the class is on applying machine learning techniques. So your focus in completing the assignments should be on developing and demonstrating your ability to apply those techniques. Part of both doing and demonstrating that requires using good coding style (in part because it makes it easier for the graders to see that you understand what you're doing). So while good coding style is secondary to applying the ML techniques, we may take points off if your code is hard to follow.

[3] You will have to be logged into your UChicago Google account to submit a response.

[4] Note that grades do not follow a curve in this class, so there is no penalty for helping others.

[5] For instance, a response to a peer that says, "to fix your error, the command should be '[...]'" is not permitted. Instead, saying, "I think you have a typo in the third argument of your command" is acceptable.

[6] Please note that in practice, the different means of class participation will be evaluated on an "either/or" basis. You are not required to participate in class via all possible modes of communication, although you are welcome to. There are multiple ways to participate because we want to give students as many opportunities to earn credit as possible, not because we want you to feel overwhelmed.

Pass/Fail (P/F), Withdrawal, and Incomplete grade requests will be handled in accordance with University and Harris policy. Students who wish to take the course pass/fail rather than for a letter grade must use the Harris P/F request form (https://harris.uchicago.edu/form/pass-fail) and must meet the Harris deadline, which is generally 9am on the Monday of the 5th week of courses. To earn a P grade, students taking the course P/F must: submit at least seven of the eight assignments and earn a grade that is overall equivalent to at least a B- letter grade.

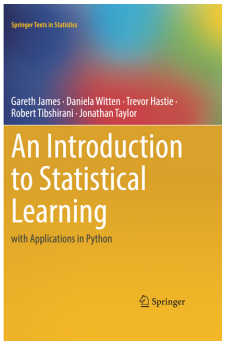
## Materials

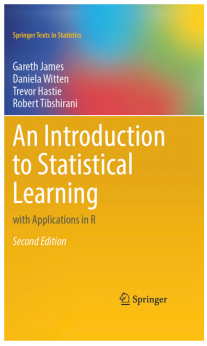
### Data Analysis and Statistical Software

You may use either Python or R software in this class, but you are required to pick one coding language and use that language for all assignments. To that end, students are required to register for both a lecture section (PPHA 30545) and a language-specific lab section (either PPHA 30547 in R or PPHA 30548 in Python). Python and R are free and open-source software that can be installed on all operating systems. The text editor or integrated development environment (IDE) you use for writing code for assignments is up to you.

### Textbooks

There are two versions of the textbook, but you only need to use one of them. The machine learning content in them is the same. They differ only in the language used for the coding examples (Python or R). You can download a free PDF of both books from the author's website: https://www.statlearning.com/.




Either: *An Introduction to Statistical Learning with Applications in Python*, 1st Edition, by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor. (ISBN-10: 3031387465)

Or: *An Introduction to Statistical Learning with Applications in R*, 2nd Edition, by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. (ISBN-10: 1071614177)

## ~~Office~~ Student Hours

Our (and the TAs') office hours for this class are listed on Canvas. The intent of office hours is part of the "hidden curriculum" for some students, so we want to clarify our expectations about how students should view and approach office hours. Those hours are for you, so please make use of them! You do not need to make an appointment to come to office hours; just drop by (be it with questions about course material, to discuss ideas, or just to chat). We will be available during those times.

Please make your best effort to attend during the posted times, but if you have a conflict or want to talk with one of us one-on-one, you are welcome to make an appointment for another time. We are happy to meet with students outside of office hours.[7]

## Harris Tutoring Program

Harris offers free tutoring support for coding in Python, R, and Stata. See the Harris academic support programs (https://harris.uchicago.edu/student-life/dean-of-students-office/academic-support-programs) for more details. Any questions should be directed to your academic advisor or harrisdeanofstudents@uchicago.edu.

## Course Policies

### General

- Attendance is a requirement for Harris students and is part of how your grade will be determined. **That said, if you are experiencing COVID-19 symptoms or illness more generally, please do not attend class in person!** We will record classes on Zoom in order to make this easier.
  - If you need a more-permanent remote learning accommodation, please contact the Dean of Students, Kate Biddle (kbiddle@uchicago.edu). Per Harris policy, all such requests can only be approved centrally, not by individual instructors.
- We do not have a policy prohibiting electronic devices (e.g., laptops) in the classroom, but the use of screens in the classroom is discouraged (with the exceptions of those used as part of an accommodation and for hand-written note-taking on tablets laid flat on your desk). We reserve the right to modify this policy if we deem it necessary.
- The class webpage is available through the Canvas portal. We will use it to post announcements, assignments, and grades. Please check it regularly.
- Email, Canvas postings, and the discussion board are the official means of communication for out-of-class messaging. In other words, you are expected to check your UChicago email account and the Canvas site regularly.
- Email is inefficient. If you have a question about the class or the material, others probably do too! Questions and answers (knowledge) are public goods, so post your question to the discussion board, and feel free to answer questions your classmates ask. We will monitor and respond as well.
- If you have a question or concern about something you don't want to discuss publicly, feel free to email us. We will respond to email within 2 business days (Monday-Friday, 9:00am-5:00pm).

[7] We only ask that you do your best to attend the regularly scheduled office hours since we have many students and there are economies of scale in the production of knowledge.

- Any and all results of in-class and out-of-class assignments and examinations are data sources for research and may be used in published research. All such use will always be anonymous.

**Recording**

- We will record all lectures. Recordings are intended as a compliment to in-class instruction, not as a substitute for it.
- We will post recorded lectures only to Canvas in accordance with University and Family Educational Rights and Privacy Act (FERPA) guidelines.
  - The University has developed specific policies and procedures regarding the use of video/audio recordings (https://teachingremotely.uchicago.edu/recording-policy/).
  - FERPA is a federal statute that, broadly speaking, guarantees privacy over certain aspects of your educational records. You can view the details of the policy on the registrar's website (https://registrar.uchicago.edu/records/ferpa/).
- If you record a class, discussion section, office hours, or meeting without permission, or if you share any of the recorded videos without permission, you may be violating eavesdropping laws, copyright laws, or the FERPA statute. So do not post or share any such videos outside of Canvas. This also applies to any manipulated video.

**Assignments**

- The goal of the assignments in this course is not just to demonstrate that you can write code and answer questions based on the output, but to help you develop an understanding of complex concepts and associated critical thinking skills. This only comes from grappling with the material (both alone and in discussion with peers). Artificial intelligence (AI) tools, including large language models (LLMs), can be a compliment or a substitute to this process. Examples of complimentary uses of AI to learning include getting help understanding a concept or an explanation of an error message. Examples of using these tools as a substitute for learning include asking Copilot to write the code for a question or ChatGPT provide an interpretation of your results.
  - While you are allowed to use AI tools in ways that are compliments to your learning, you are NOT allowed to use AI tools in this class in ways that are substitutes for learning.
  - To encourage the appropriate use of AI, any time you get help from an AI tool, you must provide in the initial query string you used and an explanation of how you used the AI tool's response as part of your answer.
  - If you are unclear if a use is appropriate or if something is an AI tool, please check with us.
- No assignments will be accepted after the 24 hour late period for any reason, valid or otherwise.[8] Not turning in an assignment, handing it in more than 24 hours late, or failing to turn it in before the link expires will result in a grade of zero.[9] We understand that students sometimes have legitimate reasons for

[8] Reasons include, but are not limited to: illnesses, athletic competitions, work trips, job fairs, job interviews, travel reservations, relative illnesses, relative funerals, out-of-town weddings, car accidents, car trouble, scooter trouble, tickets to see Billy Joel in concert, and emergency visits to the veterinarian with your dog.

[9] This is both because we post answer keys after the late period and because students from advantaged social backgrounds are more likely to make requests for extensions, so being responsive to these requests can increase educational inequality.

being unable to complete assignments on time or give their full effort, so your lowest assignment grade will be dropped. Dropping the lowest assignment grade is intended to cover ordinary illness and other emergencies. Only long-term issues of sufficient magnitude that warrant involving the Academic and Student Affairs team in the discussion can qualify for an exception to this policy.

## Academic Integrity[10]

We take the Harris Academic Honesty and Plagiarism Policies (https://harris.uchicago.edu/student-life/dean-of-students-office/policies) very seriously. All students suspected of academic dishonesty will be reported to the Harris Dean of Students for investigation and adjudication. The disciplinary process can result in sanctions up to and including suspension or expulsion from the University. In addition, if in our judgment, the preponderance of the evidence indicates that a student has committed an honor violation on an assignment, that student will receive an immediate grade of zero for that assignment and cannot earn a grade higher than a B- in the course, regardless of their performance on other assignments. This is regardless of the outcome of the disciplinary process. We trust every student in this course to fully comply with all of the provisions of UChicgo and Harris' integrity policies. Here are specific expectations:

- On assignments, it is expected that you will neither receive nor give aid, nor access any material other than items explicitly outlined in the instructions.
- For other assignments, you may (and should!) work with other students, but it is expected that you will collaborate on all parts of the assignment (as opposed to the "divide and conquer" method).
- During the entire quarter, it is expected that you will not access old problem sets, projects, answer keys, or any other class material at any time. This includes websites that post solutions under the guise of tutoring. (These sites both facilitate cheating and steal the intellectual property of the author.) This does not include the textbook authors' websites, Python/R documentation, or StackOverflow.
- During the entire semester and thereafter, it is expected that you will neither post any class material on the internet nor share any class materials with other students through any other means. Furthermore, if you become aware that this has occurred, you are obligated to let us know immediately.

### Americans With Disabilities Act

Students with disabilities needing an academic accommodation should contact UChicago's Student Disability Services (SDS). Please see their webpage for contact information (https://disabilities.uchicago.edu). If SDS determines a disability accommodation is appropriate, you should inform the Harris Dean of Students office by the end of the first week of class. The Harris Dean of Students office will work with the student and instructor to coordinate the implementation of the student's accommodations. Harris students are not required to submit their accommodations letter to the instructor, but please feel free to come talk to your instructor if you are comfortable doing so. We're happy to support your learning however we can.

### Mental Health Services

Students differ in how much they know about mental health services. Your use of UChicago's Student Health and Counseling Services (SHCS) is free, confidential, and not linked to your academic file. If you find yourself

[10] We apologize for the heavy handed tone of this section. It is intended to protect the many honest students who take the class and academic integrity as a whole.

suffering in silence, please do not hesitate to make use of the services provided by SHCS. Please see SHCS' mental health webpage for services and contact information (https://wellness.uchicago.edu/mental-health/). And if you are having serious mental, physical, or other problems, immediately contact the urgent medical care line at (773) 702-3625 (available 24 hours a day, 7 days a week).

### Diversity and Inclusion

UChicago is committed to diversity and rigorous inquiry that arises from multiple perspectives, and Harris encourages thought-provoking discourse that involves not only speaking freely about all issues but also listening carefully and respectfully to the views of others. We concur with this commitment and view the diversity that students bring to our class as a valuable resource and a benefit to learning. We expect to maintain a productive learning environment based on open communication, mutual respect, and non-discrimination. We strive to present materials in a way that is respectful of diverse student backgrounds. As there can always be a gap between intent and execution, suggestions for promoting a positive and open environment are welcomed. Please feel free to correct us on your preferred name and gender pronouns if necessary.

### Responsible Employees (Title IX)

All University of Chicago faculty and TAs are classified as "Responsible Employees." As such, they are required to report any discussions of sexual misconduct, dating violence, domestic violence or stalking to the Title IX Coordinator for the University. This includes the identities of the student making the complaint and alleged perpetrator. You will receive an email once a report is filed, but you are not obligated to meet with anyone or engage in the process. Alternatively, there are "Confidential Resource" employees at the University who do not have an obligation to share identifying information. For more information, including phone numbers, see the UChicago UMatter website (https://umatter.uchicago.edu/find-support/).

### Syllabus Change Policy

Except for changes that substantially affect implementation of the evaluation (grading) statement, this syllabus is a guide for the course and is subject to change with advance notice.

## Tentative Course Outline

The weekly coverage might change as it depends on the progress of the class. The "ISL" in the "Reading" column that follows is an abbreviation for the "An Introduction to Statistical Learning" textbook. "PS" is an abbreviation for "Problem Set," and "MP" is an abbreviation for "Mini-Project."

| Tentative Course Schedule | | | | |
|---|---|---|---|---|
| **Week** | **Day** | **Topic** | **Reading** | **Due** |
| 1 | M/Tu | Introduction | ISL Ch. 1 | |
| | W/Th | Statistical Learning | ISL Ch. 2 | |
| 2 | M/Tu | Linear Regression & Moving Beyond Linearity | ISL Ch. 3 & 7 | |
| | W/Th | | | |
| 3 | M/Tu | **MLK, Jr. Holiday! - No Class** | | |
| | W/Th | Classification | ISL Ch. 4 | PS 1 |
| 4 | M/Tu | | | |
| | W/Th | Resampling Methods | ISL Ch. 5 | MP 1 |
| 5 | M/Tu | | | |
| | W/Th | | | PS 2 |
| 6 | M/Tu | Linear Model Selection & Regularization | ISL Ch. 6 | |
| | W/Th | | | MP 2 |
| 7 | M/Tu | | | |
| | W/Th | Tree-Based Methods | ISL Ch. 8 | PS 3 |
| 8 | M/Tu | | | |
| | W/Th | Support Vector Machines | ISL Ch. 9 | MP 3 |
| 9 | M/Tu | | | |
| | W/Th | Deep Learning | ISL Ch. 10 | PS 4 |
| Exam Week | | | | MP 4 |

Please note that there is no class during week 3 on Monday, January 20th due to the due to the Martin Luther King, Jr. Day holiday. To keep coverage consistent across the sections, there will be no class on Tuesday, January 21st. So that the course will still cover a full nine weeks' worth of content, we will record an asynchronous lecture that will cover two "Classification" topics: the linear probability and logit models. Previous students have indicated that this material is partially a review.

Also during week 3, Friday lab section locations will be adjusted to accommodate Monday classes being held on Friday: the R lab will be held from 1:30-2:50pm in Keller 0021. There will be a Python lab from 9:00-10:20am in Keller 0001. The other Python lab will be held on Zoom from 1:30-2:50pm.