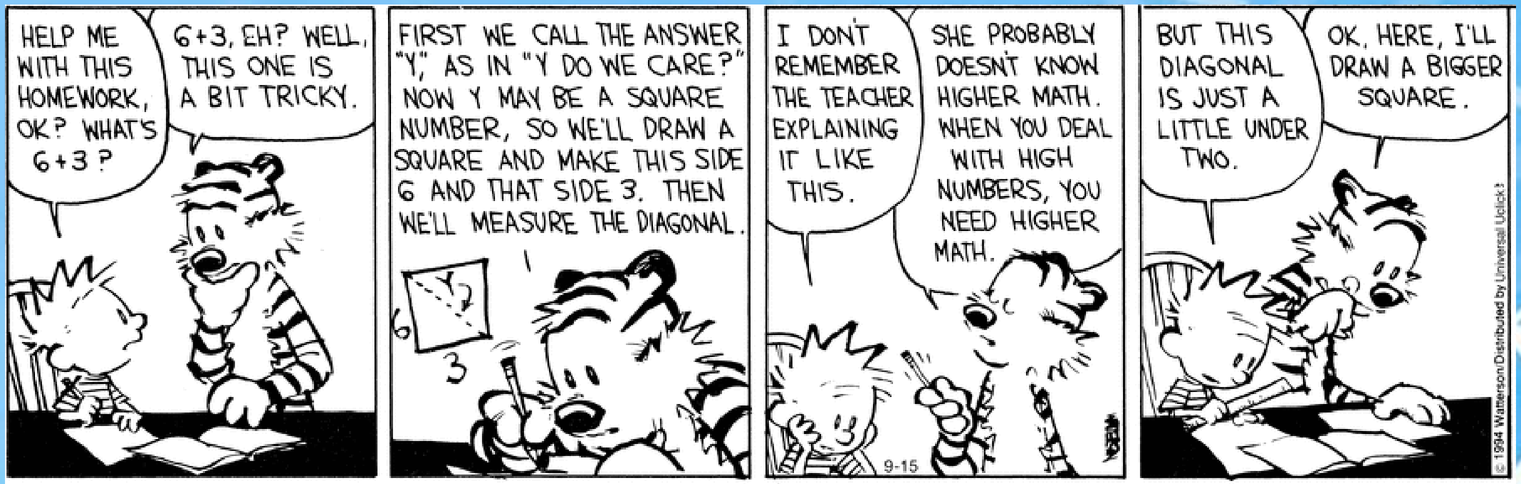


INFO 251: Applied Machine Learning

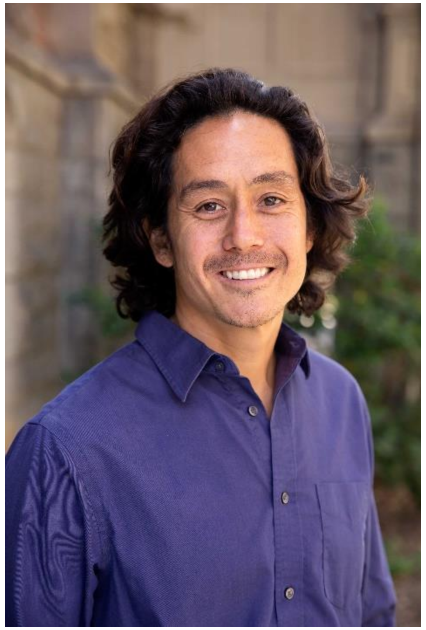
Welcome!

Good morning everyone!
The first lecture will start at 940.
Please sign the attendance sheet

Outline

- **Quick Intros**
- Course objectives
- Course content & schedule
- Course logistics

Quick Intros: Me



- Instructor: Please call me “Professor” or “Josh”
  - Office hours: Tuesdays, 930am-1030pm
- Background
  - Undergrad: Computer Science, Physics
  - Grad: Machine Learning, Development Economics
  - Other: Microsoft Research, Internet startups
- Research Focus
  - AI and international development
  - See http://jblumenstock.com

Quick Intros: Teaching team

- Suraj Nair (GSI)
  - Office hours: Wednesdays 1030-12pm, South Hall 107

- Satej Soman (GSI)
  - Office hours: Thursdays 1245-145pm, South Hall 107

Today’s objective

- To help you understand if you should take INFO251
- To answer general questions
- To answer specific enrollment questions ***after*** lecture

- (there won’t be much substance today)

Outline

- Quick Intros
- **Course objectives**
- Course content & schedule
- Course logistics

Learning Objectives

- This course is designed to help you learn how to:
  1. Effectively design, execute, and critique experimental and non-experimental methods from machine learning, statistics, and econometrics.
  2. Understand the principles, advantages, and disadvantages of different algorithms for supervised and unsupervised machine learning.
  3. Implement canonical algorithms on structured and unstructured data, and evaluate the performance of these algorithms on a variety of real-world datasets.

Not Learning Objectives

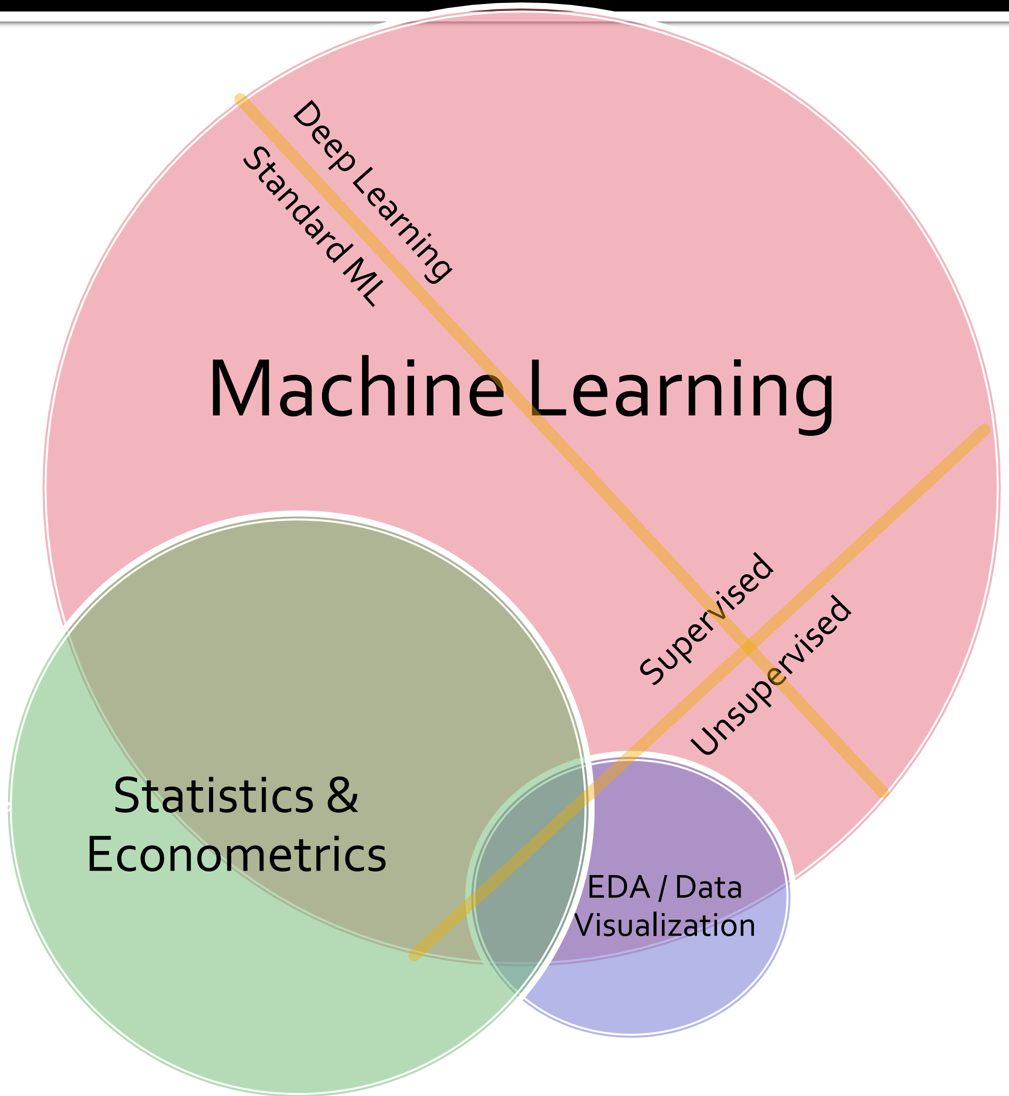
- This course will not:
  1. Teach you how to code in Python – you're expected to know this already
  2. Rely on "off the shelf" machine learning packages – you'll be coding everything from scratch
  3. Focus on proving theorems or deriving new estimators – take CS289 or CS281 or CS288 for that
  4. Spend much time dealing with working at scale (i.e., this is not a class on "big data")
  5. Go super-deep into any specific topic; this is a "survey" course

Outline

- Quick Intros
- Course objectives
- **Course content & schedule**
- Course logistics

Course Content

- INFO251 Venn diagram:



Course Content

- Causal Inference
  - Experimental methods (1 week)
  - Non-experimental methods (1 week)
- Machine Learning
  - Design of Machine Learning Experiments, instance-based learning (1 week)
  - Linear Models and Gradient Descent (1+ week)
  - Non-linear models, ensembles (2 weeks)
  - Fairness and bias in ML (1 week)
  - Neural networks, deep learning (3+ weeks)
  - ML Practicalities (1+ week)
  - Unsupervised Learning (1 week)
- Special topics
  - Machine learning for causal inference

Some key concepts

- Counterfactuals
- Double Difference Estimation
- Instrumental Variables
- Regression Discontinuity
- Cross-Validation
- Gradient descent
- Regularization
- Logistic regression
- Overfitting
- Model and feature selection
- Feature engineering
- Bootstrapping, Boosting and Bagging
- Naïve Bayes
- Fairness in ML

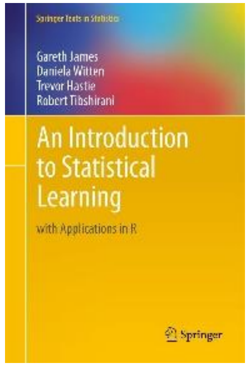
- Perceptrons and MLPs
- Regression trees and forests
- Ensemble learning
- Gradient boosting
- Support vector machines
- Neural networks, back-propagation
- Convolutional Neural Networks
- Long Short-Term Memory Networks
- RNN’s
- Attention
- Transformers
- LLMs
- Cluster analysis
- Principal component analysis
- ML for causal inference

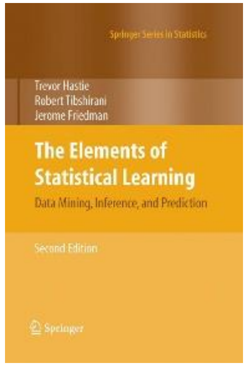
Is this the right course for you?

- Default response: “Yes”
  - After all, you’re here!
- Why might be the answer be “No”?
  - Not a good fit
    - Review learning objectives carefully!
  - Don’t have enough cycles to devote to class
    - This course has a significant workload
  - Underqualified / Overqualified (more on this…)

Is this the right course for you?

- Are you overqualified?
- You should answer “no” to most of the following:
  - Are you already comfortable with most of the “key concepts” on the last slide?
  - Have you taken a class that uses ISL, ESL, or similar?
    - If a different class in ML, show me the syllabus/book
  - Could you write a stochastic gradient descent optimizer?
  - Do you understand the math behind back-propagation?





Is this the right course for you?

- Are you underqualified?
- You should answer “yes” to all of the following:
  - Do you know how to interpret a regression table?
  - Do you know the differences between common probability distributions (normal, binomial, Bernoulli, etc.)?
  - Have you taken calculus?
  - Could you code a game of scrabble in Python (without CoPilot)?
  - Could you write a Python class that inherited methods and properties from other classes?

Is this the right course for you?

- Prerequisites
  - INFO206
    - Or an equivalent course in computer science
    - Data structures, OO-programming, algorithms, complexity
  - INFO271B
    - Or an equivalent course in statistical inference
    - Causal inference, hypothesis testing, regression
  - Python
    - (This is the last warning – check Lab 00 to make sure nothing there is unfamiliar or new)

Is this the right course for you?

- Other options on campus
  - DATA100/200
  - IEOR 265, IEOR242
  - CS189/289, CS281, CS288
  - CS282A
  - STAT254
- Sort of related
  - STAT215A / ECON 142
  - ECON241 / ARE213

Outline

- Quick Intros
- Course objectives
- Course content & schedule
- **Course logistics**

Course Logistics

- Lectures are conceptual
- Labs are practical (and required)
- The problem sets force you to implement
    1. Getting up to speed in Python (due Jan 28!)
    2. Causal inference
    3. Basics of machine learning, and a few algorithms
    4. Gradient Descent and Regularization
    5. Fairness and Bias
    6. Neural Nets, Trees and Ensembles
    7. LLMs
  - These will take time, and get harder
  - Interpretation is as important as "getting it right"

Course Logistics: Lectures

- All lectures, labs, and office hours are in person
  - Attendance is required for lecture and labs

Course Logistics: Ed

- Learn to love Ed!
  - Access from bCourses, or directly at
  - https://edstem.org/us/courses/70056/

- Ed helps us be more efficient
  - Before emailing us a question, please consider posting it on Ed
  - Logistics/administrative concerns can be communicated in a private post on Ed that is only visible to course staff

Course Logistics: Grades

- Problem Sets: 70%
  - We drop the worst problem set grade
- 2 Quizzes: 26%
  - These are **in-person**, closed book quizzes (no notes, ChatGPT, phones, etc.)
- Participation and mini-assignments: 4%

- Note: We have a strict late assignment policy – see syllabus for details
  - Moral of the story: don't turn in assignments late!
  - The real moral of the story: start your problem sets early!

Course Logistics: Collaboration

- Each student must submit independent work
  - You must type every character of your code with your own two hands
  - You must write all of your own responses and problem set interpretations
  - You may seek input from other students, but you should not share code
  - You may not reference material from past semesters
  - I take academic honesty very seriously – when in doubt, ask!
- Academic integrity and student conduct:
  - http://teaching.berkeley.edu/statements-course-policies

Course Logistics: ChatGPT

- Allowed: using ChatGPT (+ Claude, Copilot, etc.) to clarify concepts, debug code, or generate small code snippets for study or practice
- Not allowed: submitting large blocks of code or text generated by ChatGPT. Not okay to use ChatGPT to complete entire problems
- You must be able to explain all submitted code; you may be asked to demonstrate your understanding in follow-up assessments
- If ChatGPT is used, you students must include a comment specifying the tool and prompt/question used
  - I used ChatGPT with the prompt, 'How to implement a bubble sort in Python?'

Course Logistics: Enrollment

- This course is currently oversubscribed
  - To prioritize committed students, auditors and S/U are given last priority
  - Priority: I School > other grad > undergrad > CE/exchange > auditors
- If you decide to drop, please do so officially (and quickly)!
- Will you get into this course?
  - Last I checked, there were 20 on waitlist and 3 open seats
  - Many people will drop; be patient, and encourage your friends to drop early
  - *Make sure you do not have a conflicting class on your schedule; otherwise you cannot be added to the roster!*

Up Next: Experiments

- Causal Inference and Research Design
  - **Experimental methods**
  - Non-experiment methods
- Machine Learning
  - Design of Machine Learning Experiments
  - Linear Models and Gradient Descent
  - Non-linear models
  - Fairness and Bias in ML
  - Neural models
  - Deep Learning
  - Practicalities
  - Unsupervised Learning
- Special topics

Preparing for next class

- Note: First lab meeting (Lab 01) is tomorrow
  - Take a look at Lab 00 on bCourses to make sure you're comfortable with all the programming basics there
  - Read through Section 3.1 (inclusive) of the "Stats Refresher" on bCourses
- Take the online "Background Survey" on bCourses
- Read about impact evaluation and randomized experiments
- Get started on the first problem set!