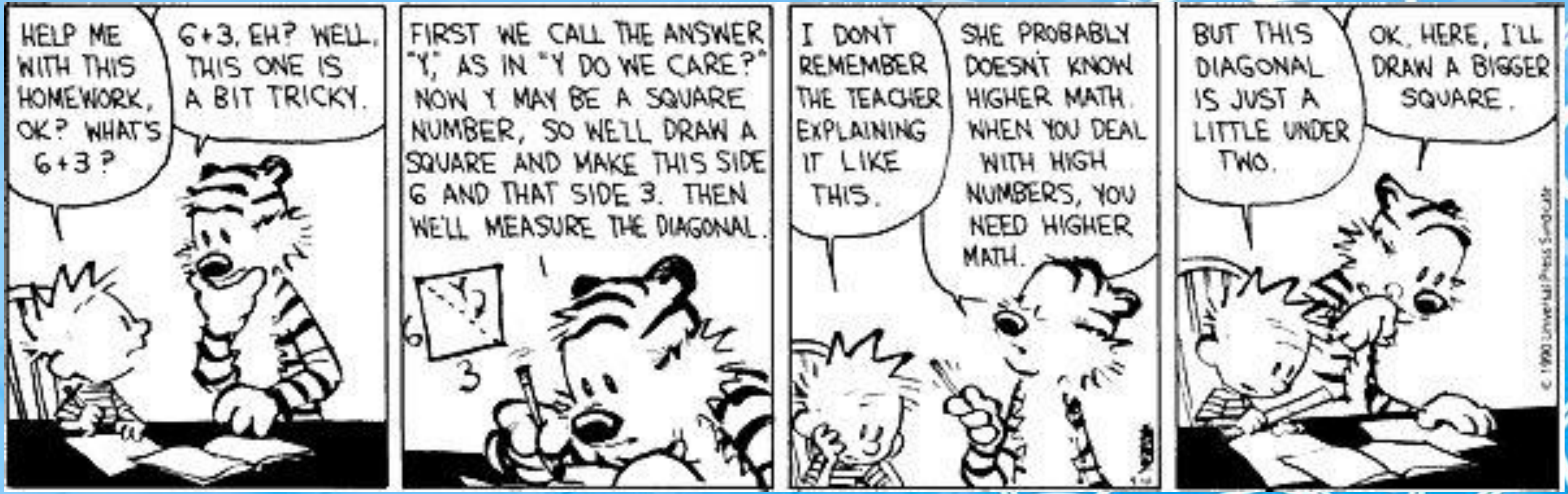


INFO 251: Applied Machine Learning

Welcome!

Good morning everyone!
The first lecture will start at 940

Outline

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

Quick Intros: Me

- Background
  - Undergrad: Computer Science, Physics
  - Grad: Machine Learning, Development Economics
  - Other: Microsoft Research, Internet startups
- Research Focus
  - Using novel data and methods to better understand the economic lives of the poor
  - See http://jblumenstock.com, http://didl.berkeley.edu

Quick Intros: Teaching team

- Emily Aiken (GSI)
  - Lead weekly lab/sections (Wednesdays 2-3pm)
  - Holds weekly office hours (Wednesdays 3-4pm)

- Lia Chin-Purcell and Uttam Ramesh (Graders)
  - Grade problem sets
  - Help manage Piazza
  - Hold weekly office hours (Fridays 10-11am)

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

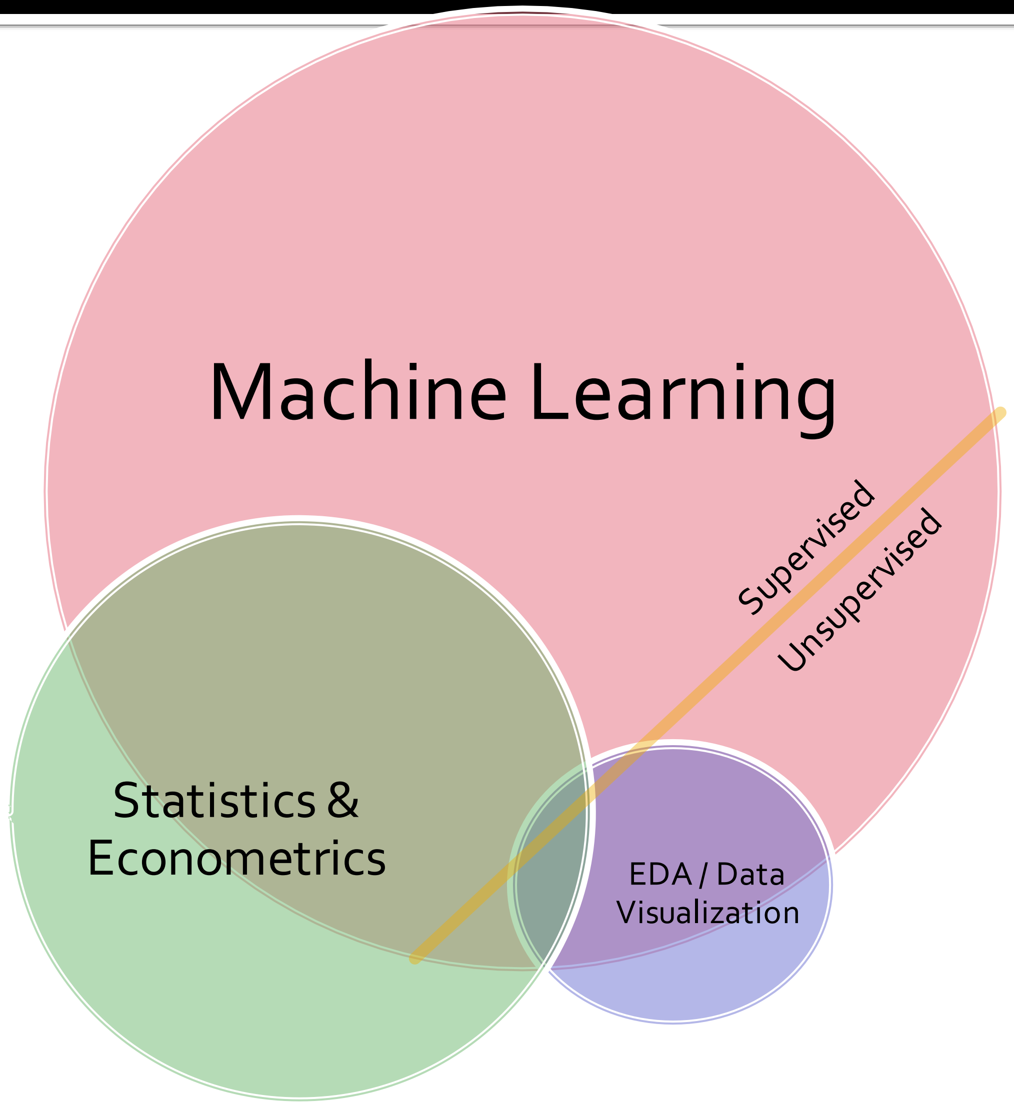
- This course will not:
  1. Teach you how to code in Python – you're expected to know this already
  2. Rely on "off the shelf" machine learning packages – you'll be coding everything from scratch
  3. Focus on proving theorems or deriving new estimators – take CS289 or CS281 for that
  4. Spend much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

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)
  - Neural networks, deep learning (2 weeks)
  - Fairness and bias (1 week)
  - ML Practicalities (1 week)
  - Unsupervised Learning (2 weeks)
- Special topics
  - Machine learning for causal inference

Some key concepts

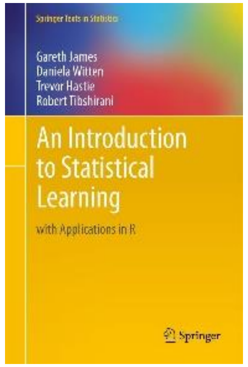
- Counterfactuals
- Double Difference Estimation
- Instrumental Variables
- Regression Discontinuity
- Evaluation and Optimization
- Cross-Validation
- Gradient descent
- Regularization
- Logistic regression
- Overfitting
- Model and feature selection
- Feature engineering
- Bootstrapping, Boosting and Bagging
- Naïve Bayes
- Fairness in ML
- The bias-variance tradeoff
- Perceptrons and MLPs
- Regression trees and forests
- Ensemble learning
- Gradient boosting
- Support vector machines
- Hyperplanes and linear separability
- Neural networks, back-propagation
- Convolutional Neural Networks
- Long Short-Term Memory Networks
- Auto-encoders
- Cluster analysis
- Principal component analysis
- ML for causal inference
- Collaborative filtering

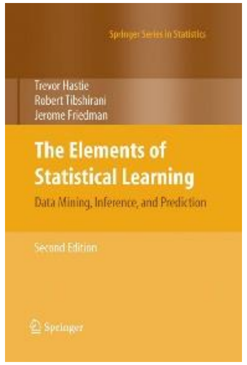
Is this the right course for you?

- 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 or ESL?
    - If a different class in ML, show me the syllabus/book
  - Could you write a stochastic gradient descent optimizer?
  - Could you write a back-prop algorithm yourself?





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 understand the OLS assumptions?
  - 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?
  - 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)

Is this the right course for you?

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

Outline

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

Course Logistics

- Instructor: Please call me “Professor” or “Josh”
  - Office hours: Tuesdays, 11-12am
  - Feedback welcome, of all types
- GSI: Emily Aiken
  - Holds weekly office hours (Wednesdays 3-4pm)
- Readers: Lia Chin-Purcell and Uttam Ramesh (Graders)
  - Hold weekly office hours (Fridays 10-11am)

Random breakout session #1

- You will be randomly divided into groups of 3-4
  - This breakout room will only last a few minutes
  - It will be awkward, but that's okay
  - Introduce yourselves
    - Name, program, something you look forward to in 2022
    - Are you potentially looking for study partners?
    - Consider exchanging contact information!
  - This will just last a few minutes…

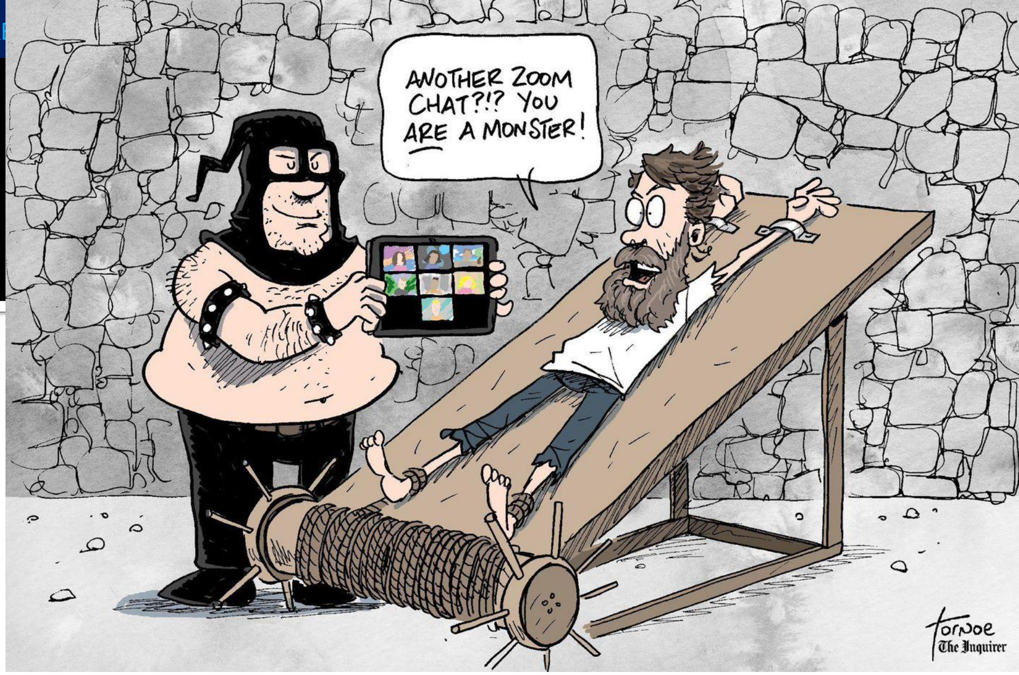
Course Logistics

- Lectures are conceptual
- Labs are practical
- The problem sets force you to implement
  1. Getting up to speed in Python (due Jan 26!)
  2. Causal inference
  3. Basics of machine learning, and a few algorithms
  4. Gradient Descent and Regularization
  5. Fairness and Bias
  6. Trees and Ensembles
  7. Neural Nets and Deep Learning
  - These will take time, and get harder
  - Interpretation is as important as "getting it right"

Course Logistics: Lectures

- For now, all lectures, discussions, and office hours are on Zoom
  - My hope is to eventually return to the classroom, but who knows…
  - If we do return to the classroom, I hope to be able to continue to record lectures, but recording quality may drop significantly
- Having everything remote is challenging
  - Harder to build a sense of community, form groups, etc.
  - Harder for me to “read the room”
  - Less time for impromptu discussion

Course Logistics: Zoom



- I anticipate there will be challenges
  - We're (still) all in this together
- Things I appreciate:
  - Please keep your video on, if at all possible!
  - Please keep zoom chat to a minimum. For now, please feel free to unmute yourself to ask a question
  - Please be respectful of others and the "classroom" environment
    - https://sites.ischool.berkeley.edu/remote-teaching/2020/03/09/for-students-online-classroom-expectations/
- Big picture: Feedback and constructive suggestions welcome
    - Feel free to communicate with me or the teaching team via email or Piazza

Course Logistics: Piazza

- Learn to love Piazza!
  - Especially during remote instruction, it's one way to connect with us and with each other
  - Access from bCourses or directly at piazza.com/berkeley/spring2022/info251
- Piazza also helps us be more efficient
  - Before emailing us a question, please consider posting it on Piazza

Course Logistics: Grades

- Problem Sets: 80%
- 2 Quizzes: 16%
- Participation and mini-assignments: 4%

- Note: I'm a stickler when it comes to late assignments – 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: Enrollment

- This course is oversubscribed by roughly 50%
  - To prioritize committed students, I do not permit auditors or S/U
  - Concurrent enrollment and other students rarely gain admission
- If you decide to drop, please do so officially and quickly!
- Will you get into this course?
  - Currently 40 on waitlist
  - Many people will drop: Last year, roughly 20 were eventually enrolled

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
  - Neural models
  - Unsupervised Learning
  - Practicalities, Fairness, Bias
- Special topics

Preparing for next class

- Note: no lab meeting tomorrow
- Take the online “Background Survey” on bCourses
- Read about impact evaluation and randomized experiments
- Get started on the first problem set