PA 470 - ML/AI for public sector

Week 1

Divij Sinha 2025-01-16

Introductions!

Course Goals

- Understand the fundamental concepts of machine learning.
- Grasp the current landscape of AI, including LLMs and their capabilities.
- Analyze the policy and ethical implications of AI in public administration.
- Utilize the R programming language for data analysis and ML model building.
- Apply R programming skills relevant to public sector data challenges.
- Develop simple machine learning models for real-world problems.
- Effectively use LLMs for relevant tasks.

Course Structure

- Each week's lecture will be split into consist of two ~75 minute sessions:

- Part 1: "Lab" sessions involving code-along exercises, statistical concepts, ML fundamentals, and AI theory.
- Break: A 10-minute break between sessions.
- Part 2: Discussions on real-world applications of AI/ML in the public sector, emphasizing practical implications and ethical considerations. There will be a 10-minute break between sessions.

Course Management

- GitHub
- https://github.com/uic-cuppa-pa470-spring2025/course-website

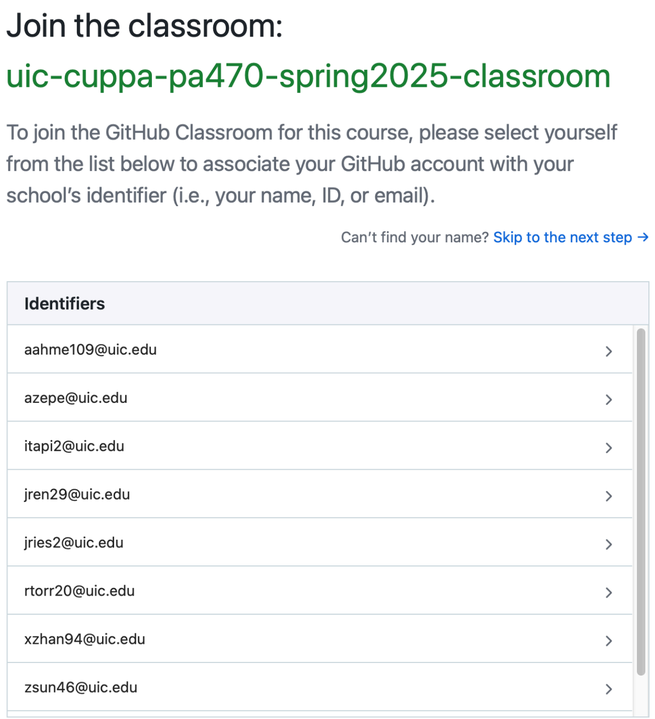
- Blackboard will have all relevant links + instructions to get you started

Grading

- Assignments - (30%)
  - These will consist of coding and writing assignments. Code will be graded on effort, thoughtfulness, and approach rather than perfect correctness. Writing will be evaluated based on argument strength, clarity, and understanding of the subject matter, not solely on language mechanics. Allowed use of tools like ChatGPT will be on a per question/ assignment basis.

Assignments

- To be submitted on GitHub (code AND writing)
- Assignment 1 invite link - https://classroom.github.com/a/lNo4S76r



Grading

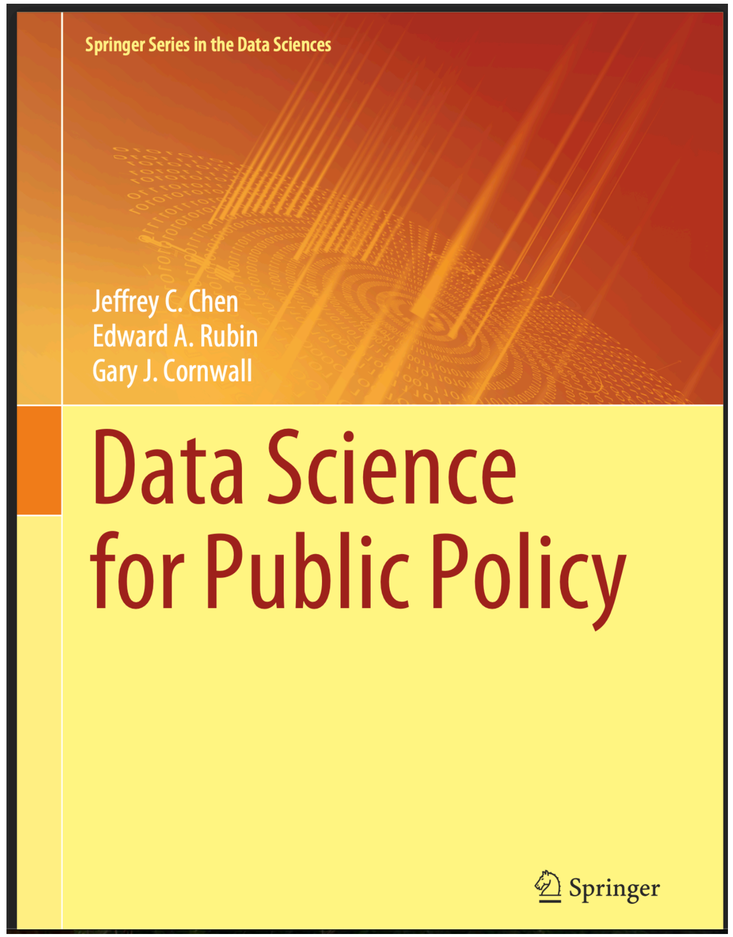
- Final Project - (45%)
  - This project will involve applying ML/AI techniques in R to a real-world problem relevant to the public sector. The project can be a proof of concept or a fully functional application. Grading will consider the ambition and effort invested, and we will discuss project scoping throughout the semester. A proposal is due in Week 5, with updates in Weeks 9 and 14. The final submission must also include a critical analysis of your project and its limitations.
  -

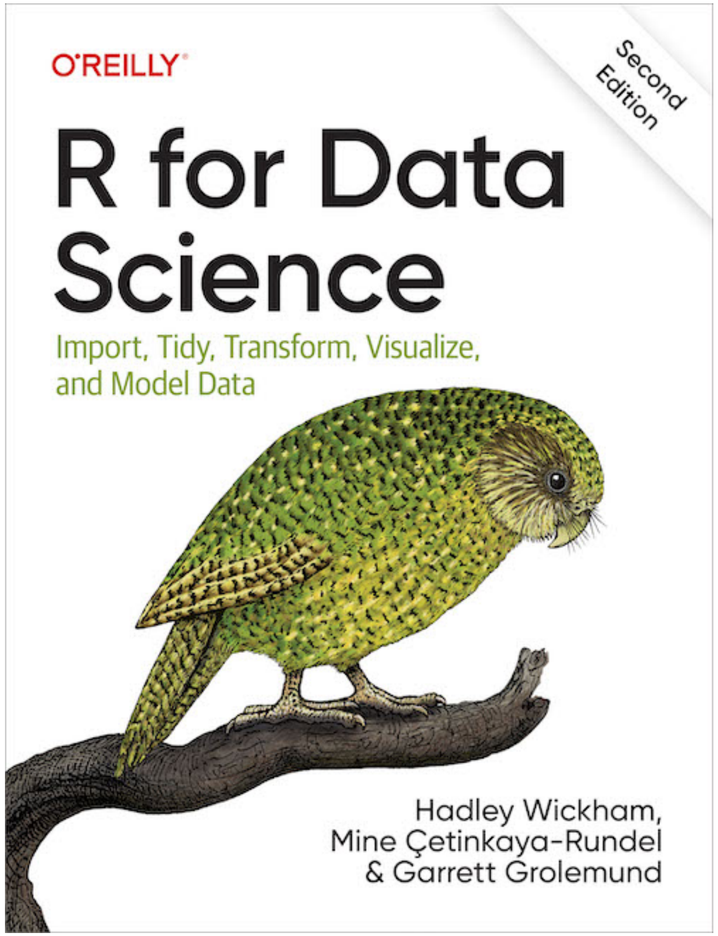
Grading

- Engagement - (25%)
  - Class is a dialogue.
  - Active participation in class discussions is essential!

Textbook(s)

- Only for review





Software

- You will have to run RStudio + GitHub for this course, nothing else should be necessary
- Windows machines - use WSL or (https://git-scm.com/downloads/win)
- If not, use the GitHub desktop app - slightly limited in functionality but should be fine

Questions?

Git/GitHub

What is git

- Git is a version control system
- Takes care of differences in files
- Important terms
  - `repo, branch`
  - `untracked, modified, committed etc.`

Basic git flow

- Create a repo - `git init`
- Check status - `git status`
- Add a file - `git add file.txt`
- Commit a file - `git commit -m “first commit”`
- Check history - `git log (--oneline)`
- Push code* - `git push`

What is gitHUB?

- Github is a provider that allows hosting of repos
- Allows collaboration

Common git flow

- **Clone** a repo - `git clone git@github.com:divij-sinha/my-first-repo.git`
- **Change** a file - `git add first.txt`
- Add a file - `git add second.txt`
- Commit a file - `git commit -m “second commit”`
- Push code* - `git push`

Resources

- https://www.atlassian.com/git
- https://git-scm.com/docs/gittutorial
- https://learngitbranching.js.org

R

But first, installations

- Install R - https://cran.rstudio.com
- Install RStudio - https://posit.co/download/rstudio-desktop/

Intro to R and RStudio

- RStudio interface
- Simple code and execution of code in console
  - Operators
  - Variables
  - Data Types
  - Vectors
  - Functions

Intro to R and RStudio

- Libraries
- File types
- Importing data
-

tidyverse

https://www.tidyverse.org

- A paradigm of manipulating data in R
- A set of packages sharing a design philosophy, grammar, and data structures
  - `tibble`
  - `dplyr`
  - `ggplot2`

tidymodels

https://www.tidymodels.org

- Statistical Modelling and ML using tidyverse rules

Statistics

- *INFERENCE* - Allows us to **understand** underlying processes that generate real world phenomenon
- *PREDICTION* - Allows us to **generate** new values that closely resemble reality (this is ML!)

ELIZA

```
  Welcome to
                EEEEEE  LL      IIII   ZZZZZZ   AAAAA
                EE      LL       II       ZZ   AA   AA
                EEEEE   LL       II     ZZZ    AAAAAAA
                EE      LL       II    ZZ      AA   AA
                EEEEEE  LLLLLL  IIII ZZZZZZ    AA   AA

  Eliza is a mock Rogerian psychotherapist.
  The original program was described by Joseph Weizenbaum in 1966.
  This implementation by Norbert Landsteiner 2005.


ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
```

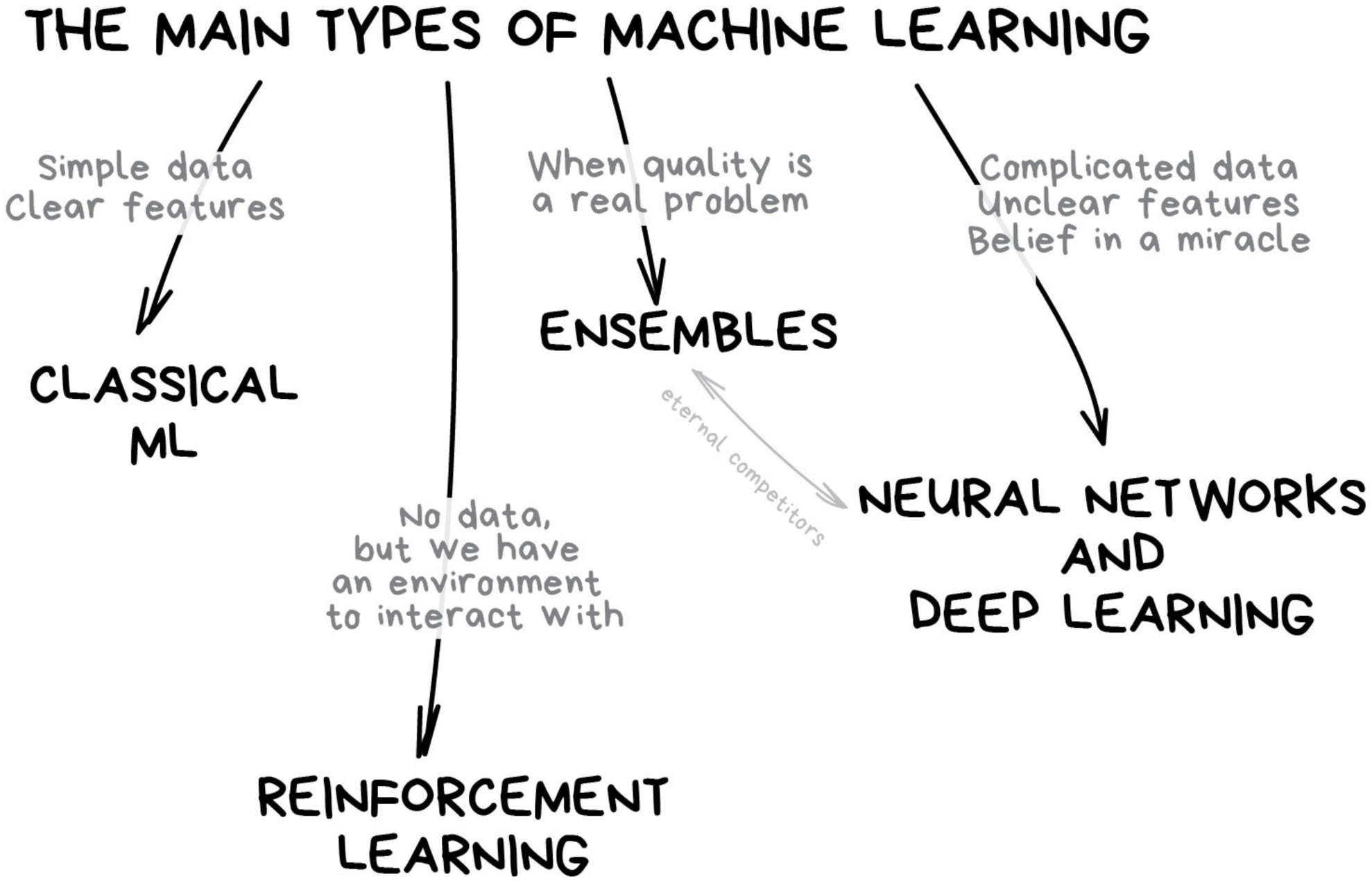
What is a model?

- Some formal mathematical creation that allows us to generate some random variable based on some set of simplified assumptions of reality
- More simply, given 0 or more inputs, a model is some sort of (potentially probabilistic) mathematical equations that generate one or more outputs
- What is the point of a model

Why do we care in public policy?

- What will be the inflation next week/month/year?
- Will better education lead to better wages? (Correlation? Causation?)
- Who will commit fraud on my welfare schemes?

- Inference vs Prediction?


THE MAIN TYPES OF MACHINE LEARNING
Simple data
Clear features
CLASSICAL ML
No data,
but we have
an environment
to interact with
REINFORCEMENT LEARNING
When quality is
a real problem
ENSEMBLES
eternal competitors
Complicated data
Unclear features
Belief in a miracle
NEURAL NETWORKS AND DEEP LEARNING

Linear Regression

- What is it?
- Assumptions?