STAT 432: Basics of Statistical Learning

Introduction

Ruoqing Zhu, Ph.D. <rqzhu@illinois.edu>

https://teazrq.github.io/stat432/

August 25, 2025

University of Illinois at Urbana-Champaign

Welcome to STAT 432

- Course Website
  - https://teazrq.github.io/stat432/
- Teaching Assistants
  - Sam Shi-Jun <wooimj2@illinois.edu>
- In-person office hour:
  - Mon 3:30 - 4:30pm at CAB 137 (Switch to Wed on Sep 15, Oct 13, Nov 10 and Dec 8)
  - Thur 2-3pm (Location TBA)
- Zoom office hour
  - Tue 6pm
  - Zoom: 85371847706, password: 432

My Research

- Machine learning in general
- Reinforcement learning, random forests, survival analysis, etc.
- Personalized medicine, decision making, computational challenges ...
- Applications to real world problems: sepsis, infectious diseases, nutrition and food, cancer, genetics ...
- more at sites.google.com/site/teazrq/

Overview

- Basic course information
  - Textbook
  - Course website
  - Homework
  - Exam
  - Project
- Topics and objectives
- ChatGPT, GitHub Copilot and other tools

Textbook

SMLR **Statistical Learning and Machine Learning with R**

by Zhu, R.

[online]

ISL **An Introduction to Statistical Learning: With Applications in R**

by James, G., Witten, D., Hastie, T. & Tibshirani, R.

[free PDF]

ESL **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**

by Hastie, T., Tibshirani, R. & Friedman, J.

[free PDF]

Course material goes beyond just a few textbooks!

Course Website

- Main website: https://teazrq.github.io/stat432/
  - post course material, homework, project and other info
- Canvas: https://canvas.illinois.edu/
  - Announcements
  - Discussion board
- Gradescope https://www.gradescope.com/courses/1082223
  - Submit HW and project
  - Entry code: **42RBX5**

Discussion Board

- Canvas discussion board as the primarily platform of communication
- For email communications, start with "Stat 432" in your email title.

Homework

- We have approximately 10 sets of homework (1 per week), depending on the course progression
- Assigned on Monday and due at Thursday (11:59PM) of the following week
- Late submission allowed: up to 3 days, 10% penalty per day
- The lowest score dropped
- Submit to gradescope (.pdf, with all code chunks visible)

Midterm Exam

- One in-class exam
- 20-30 multiple choices or short answer questions
- 10% of the total score
- Date: Nov 11th
- Examples on course website

Final Project

- Two options:
  - **[Option 1]**: Default project; Dataset and objectives provided
    - I will post a data analysis project with specific requirements
  - **[Option 2]**: Self-proposed project
    - Mandatory for graduate students
    - Complex data and goals. Cannot be a straightforward classification or regression problem.
    - Setup a meeting with me no later than Oct 24th to discuss project scope. Update project progression overtime.
    - In-class 15-min presentation
- Both options need to submit a 12-pages final report
- Maximum 3 members per team
- Previous projects can be found on the project page

Topics and Objectives

Data Analytic

Typical requirements for a data analytic:

- The model, algorithm, formulation
- Advantages/disadvantages: model flexibility, computational cost
- How to evaluate and tune parameters
- Domain knowledge, data properties, which model to use?
- Interpret results and communicate with others

Handwritten Digit Recognition

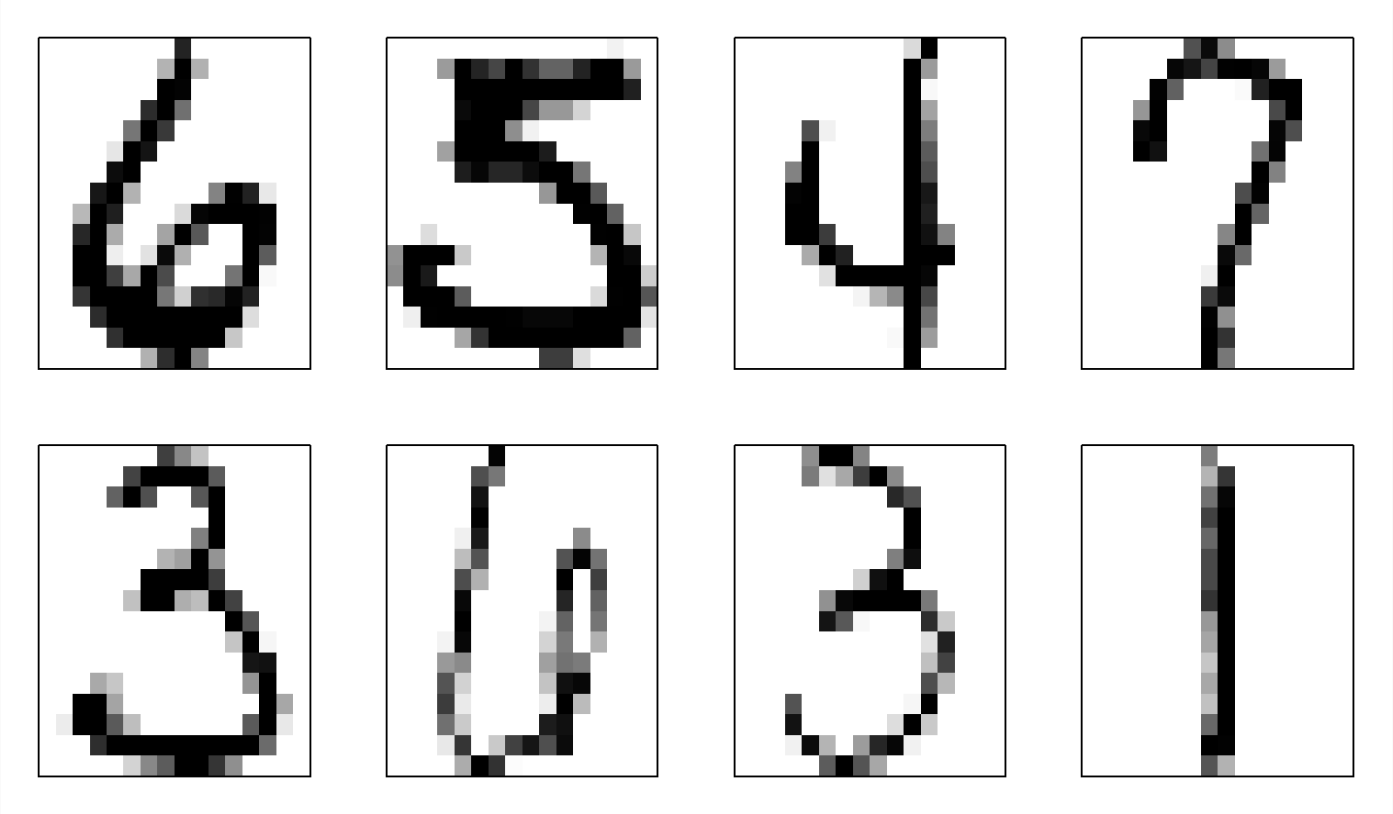


Figure 1: Hand written digit data from ElemStatLearn

- challenges: high-dimensionality, high correlation, non-linear

Gene Expression and Classification in Vaccine Study

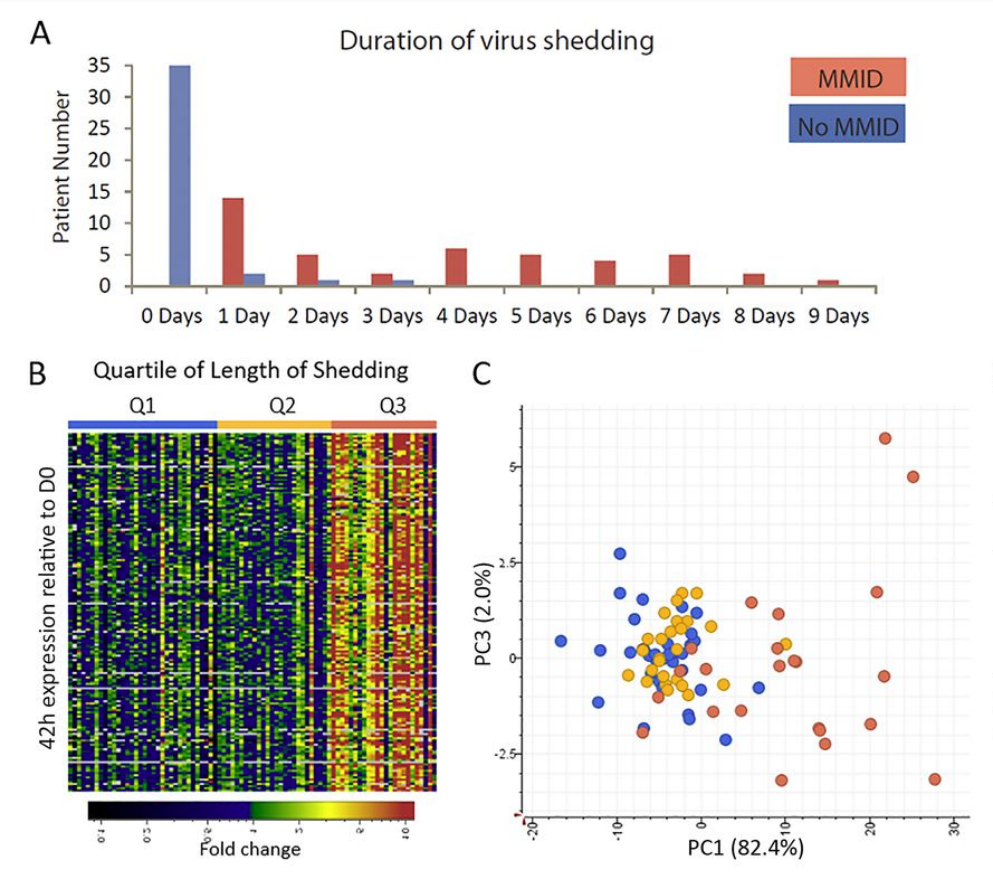


Figure 2: Gene expression changes after influenza infection, Walters et al., 2019

- challenges: longitudinal, small sample size, unsupervised

Model Evaluations for Predicting LTBI

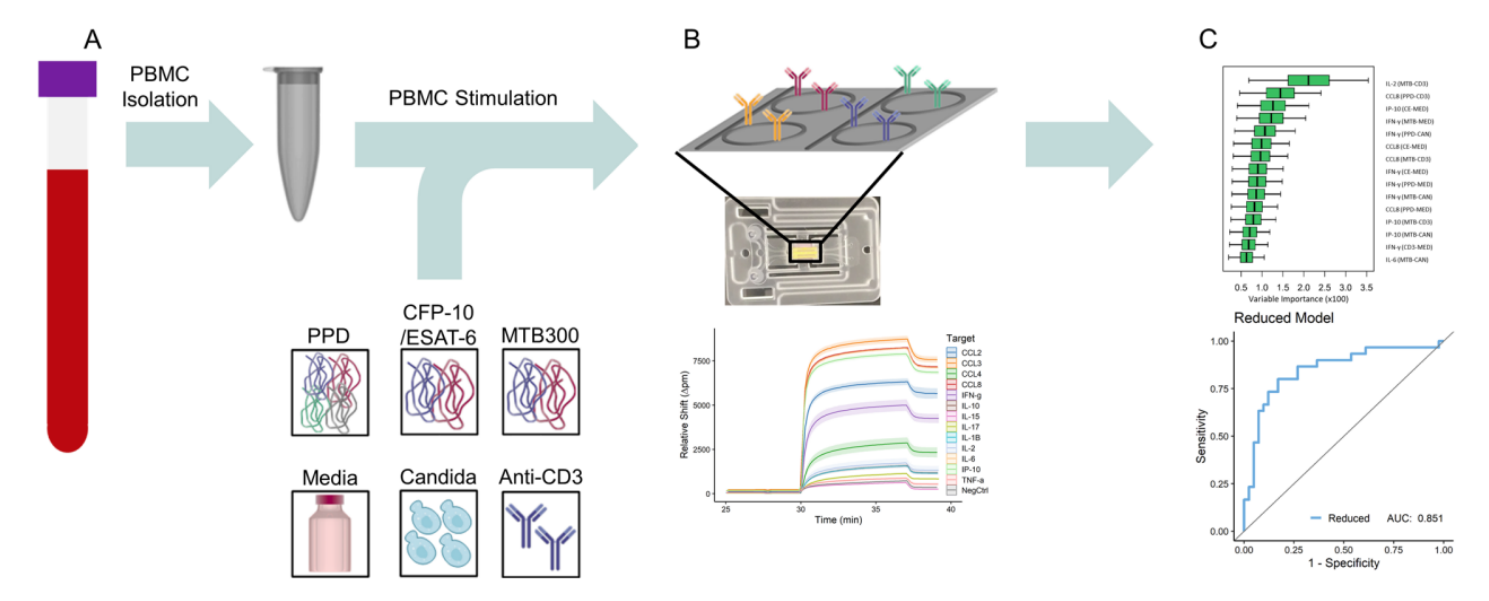


Figure 3: Identify latent tuberculosis infection with cytokine biomarkers, Robison et al., 2021

- challenges: low sample size, unbalanced group, ...

Personalized Nutrition

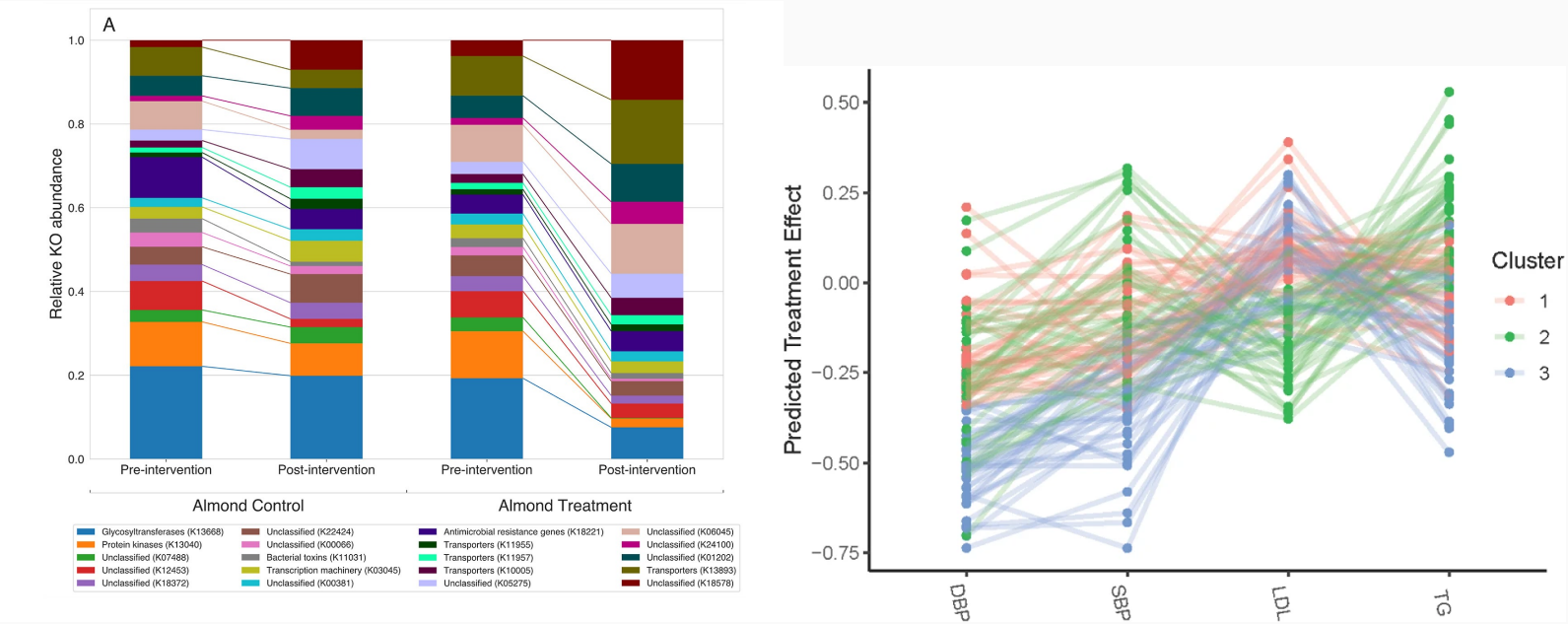


Figure 4: Decision making in nutrition study, Guo et al., 2023, Shinn et al., 2024

- challenges: Data integration

Clustering Analysis and Unsupervised Learning

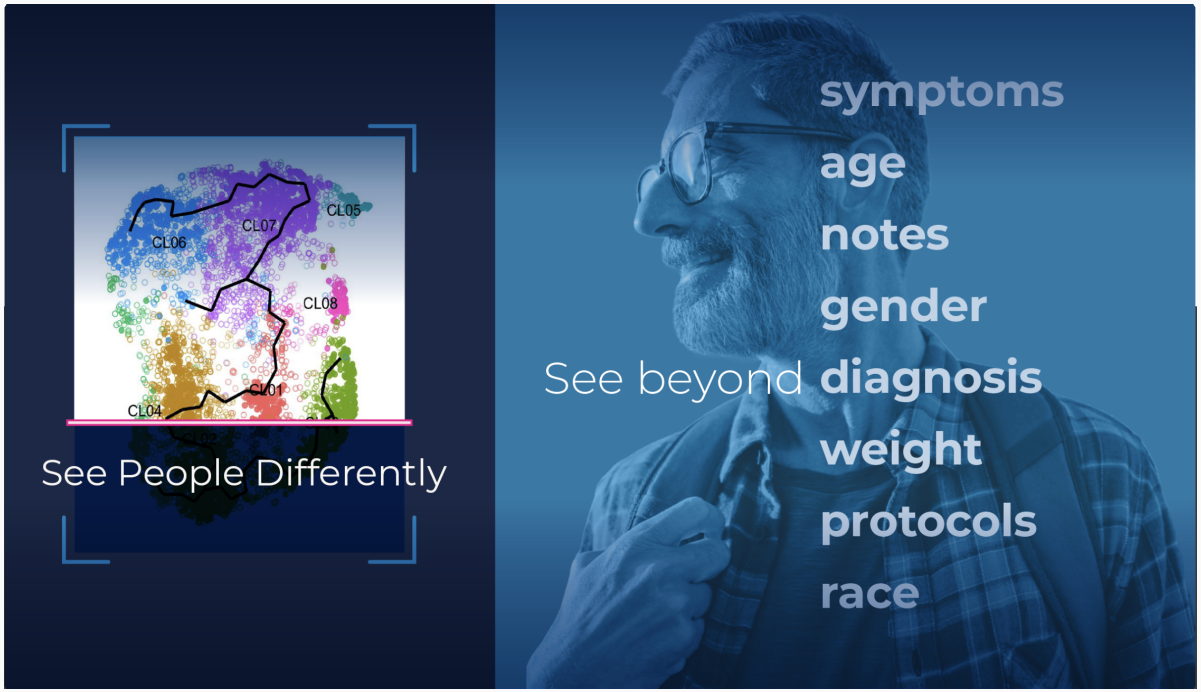


Figure 5: Clustering of patients with suspected sepsis, ImmunoScore™ by Prenosis Inc.

- challenges: unsupervised learning

Personalized Medicine for Managing Diabetes

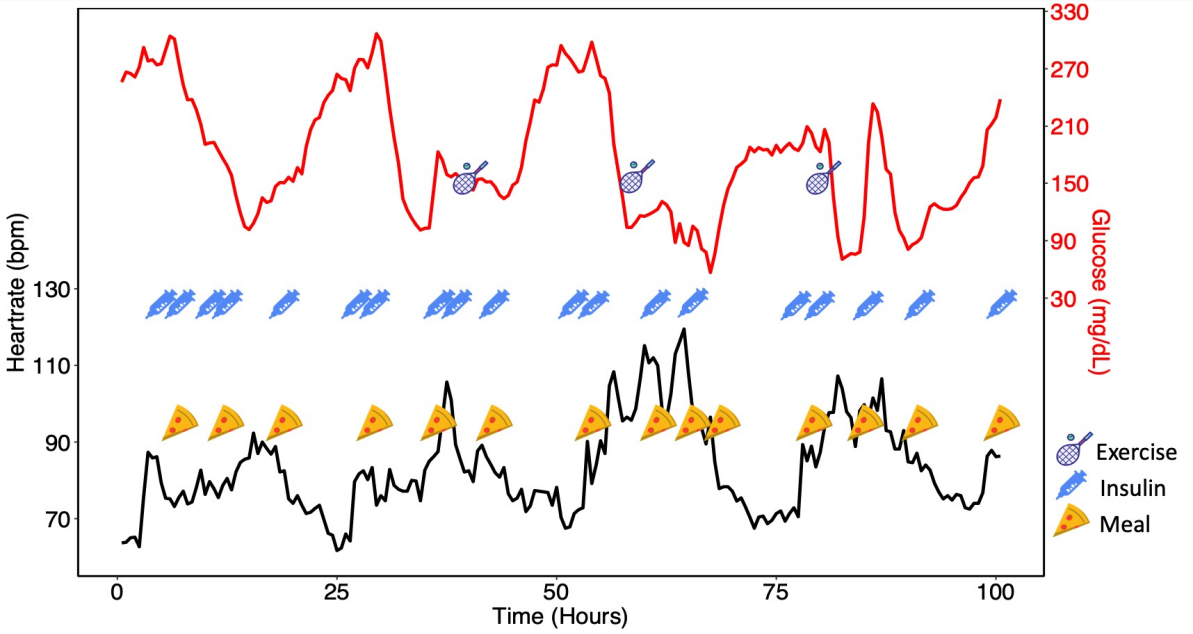
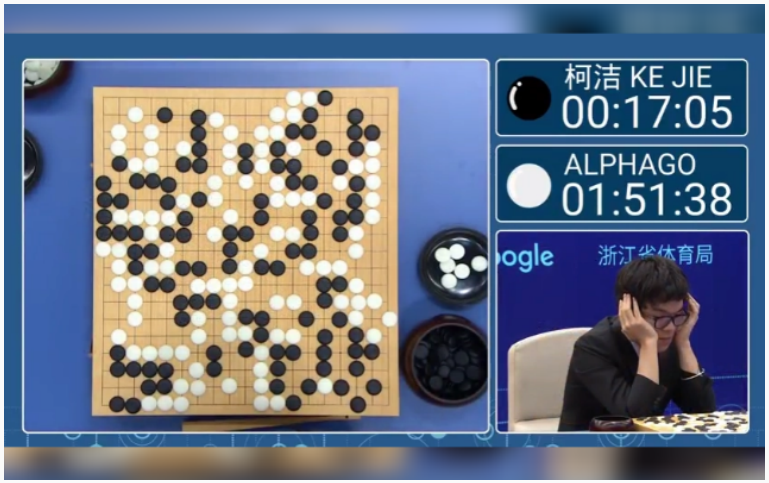


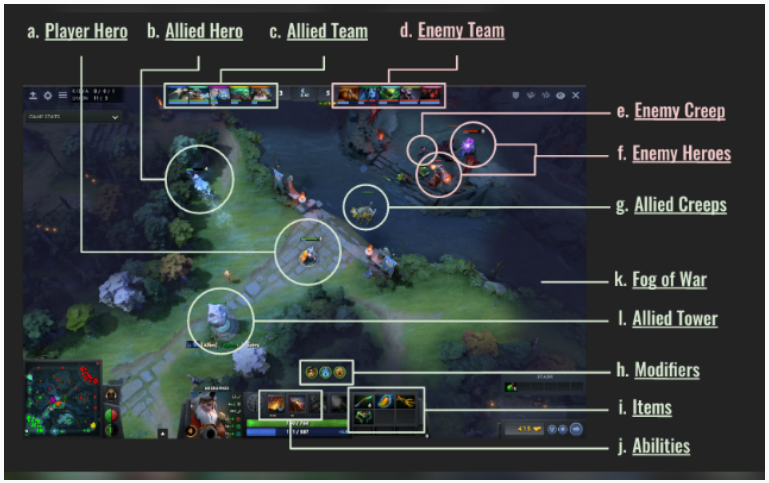
Figure 6: Dynamic treatment regime for diabetes, Zhou et al., 2021

- challenges: decision making,

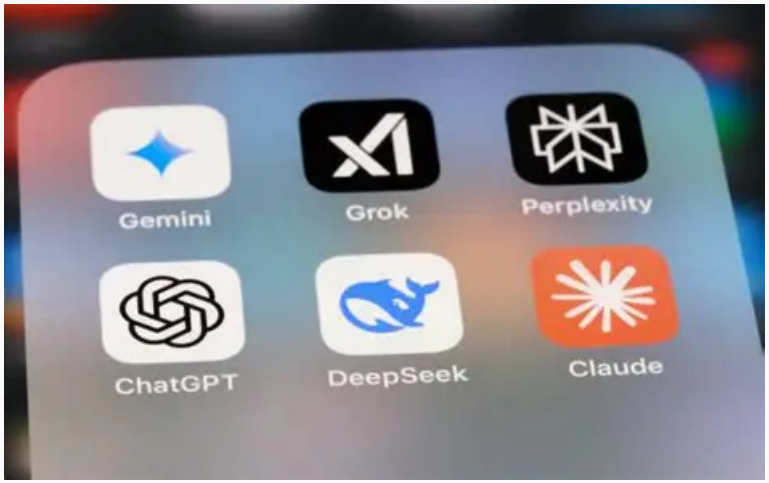
Reinforcement Learning



AlphaGo



Gaming: OpenAI Five



LLM (chain of thoughts)

Robotics

ChatGPT

- ChatGPT (any version), Copilot, Claude, Gemini, etc. Use them as much as you need.
- How to write effective prompts?
- Correctness — your responsibility
- Some other resources
  - Info from CITL

What's covered?

- Understand the models
  - Supervised / unsupervised, Regression / classification
  - Suitable for high-dimensional, high correlated data?
  - Principles of tuning parameters. Which one(s) should I tune?
  - Computational cost
  - How to evaluate models?
  - Interpretation

- Understand the data ...
  - Integrating domain knowledge
  - How the data were collected?

What will we learn?

- Fundamental statistical concepts
  - Bias-variance trade-off
  - Cross-validation
  - Resampling
  - Statistical simulation
- Practical skills
  - Using R for implementation
  - Data processing
  - Using AI tools to assist learning and programming

Expectations

- Intense weekly schedule!
- Several homework assignments are very challenging
- Pay attention to weekly objectives
- Practice and ask questions, bring laptops if you need

Prerequisites

- Probability: probability and random variables, distributions
- Statistics: estimators, likelihood, linear regressions
- Mathematics: linear algebra (matrix operations) and calculus
- Some prior knowledge of R
- Be able to use AI tools to learn new stuff

Next Step

- Install/update to the lasted version of R and R Studio (or VS Code)
  - Install packages: `devtools` and `tidyverse`
  - Update out-of-date packages, if any
- Read homework 1

Questions?