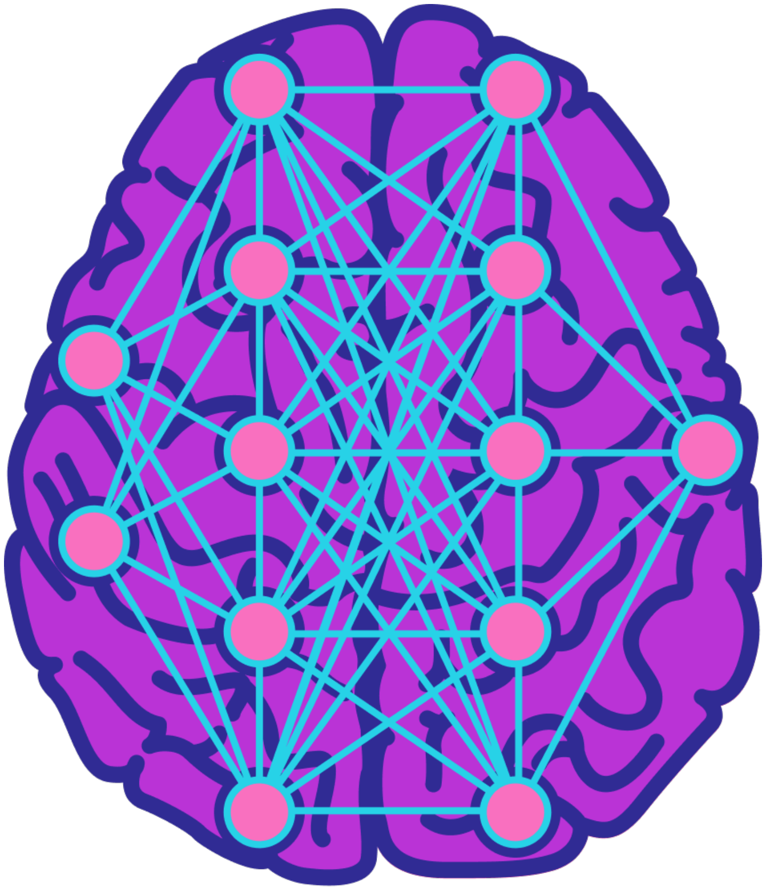
DS 4400

Machine Learning and Data Mining I
Fall 2020

Alina Oprea
Associate Professor
Khoury College of Computer Science
Northeastern University

September 10 2020

Welcome to DS 4400!



Machine Learning and Data Mining I

Introduction

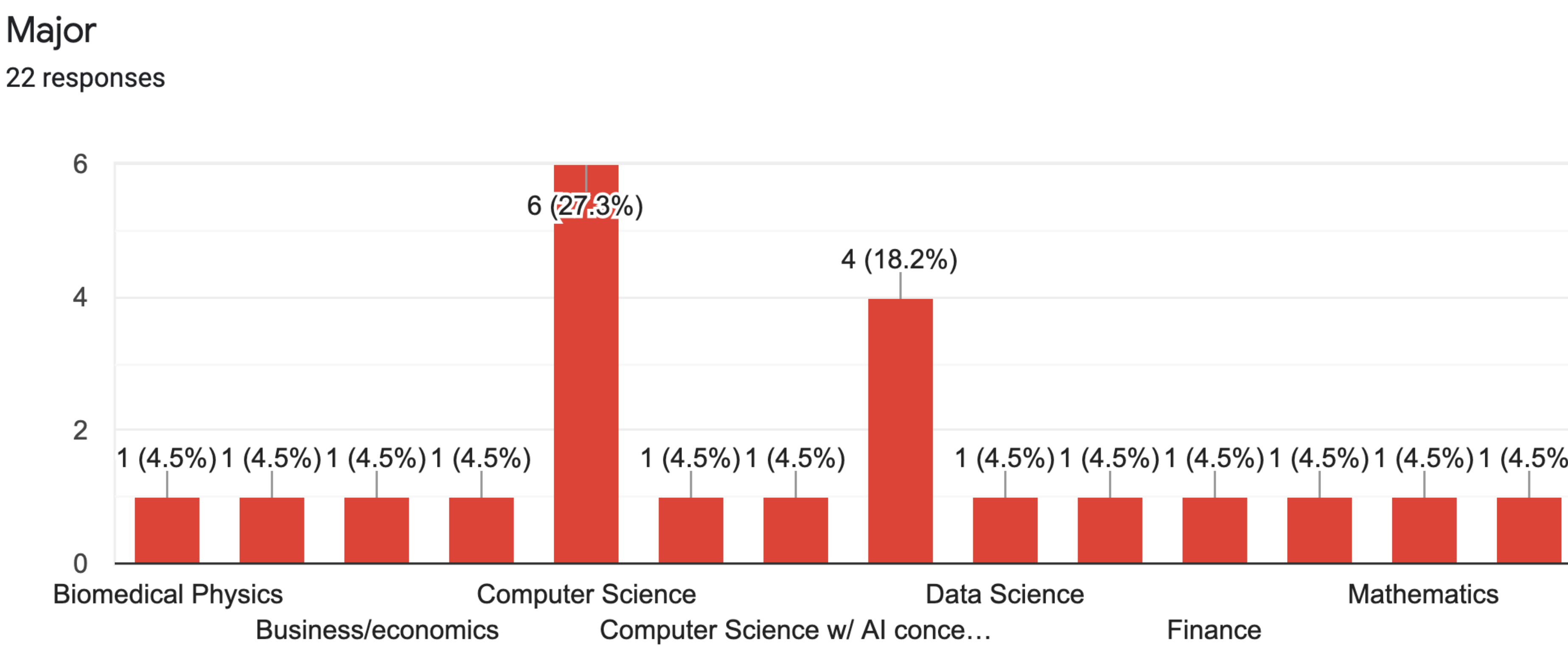
- Ph.D. at CMU
  - Research in storage security, cloud security, and cryptographic file systems
- RSA Laboratories
  - Cloud security, applied cryptography, game theory for security
  - ML/AI in security
- NEU Khoury College – since Fall 2016
  - NDS2 Lab part of the Cybersecurity and Privacy Institute
  - ML for security applications (attack detection, IoT, connected car security, collaborative defenses)
  - Adversarial ML (study the vulnerabilities of ML in face of attacks and design defenses)

TA Introduction

- Alex Wang
  - BS in CS, started 2016
  - Experience as data science co-op and took DS 4400 in Spring 2020
- Matthew Jagielski
  - PhD student in CS, started 2016
  - Research in adversarial ML, fairness, and differential privacy

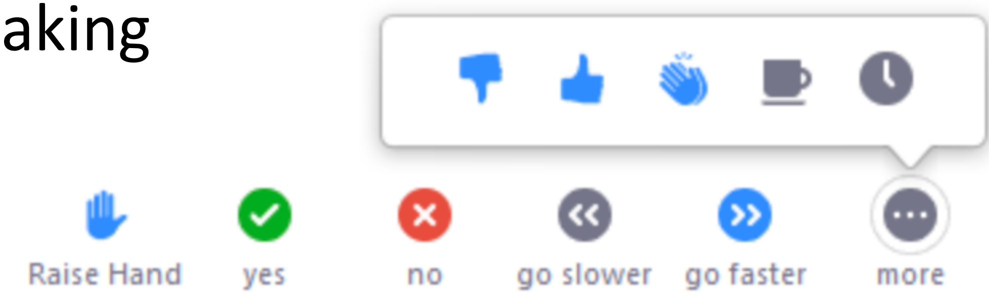
DS 4400 Class

- Enrollment of 36
- Diverse majors

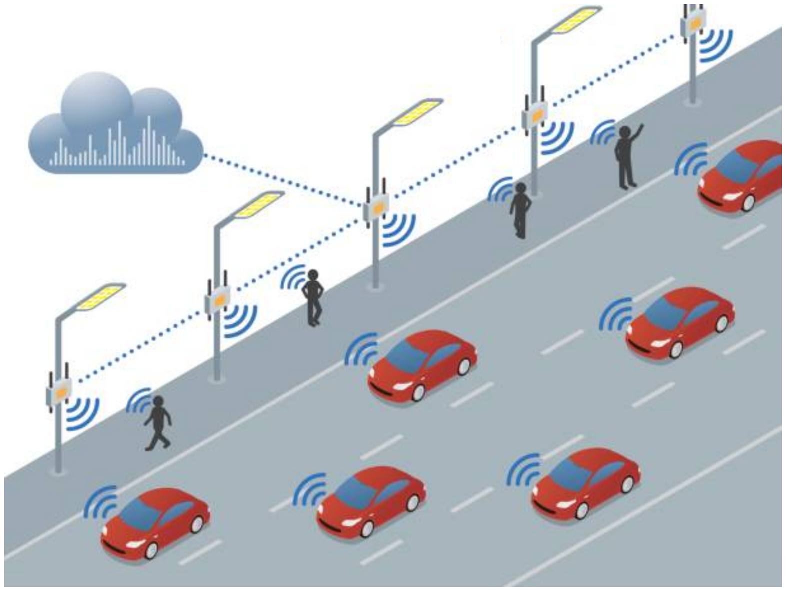


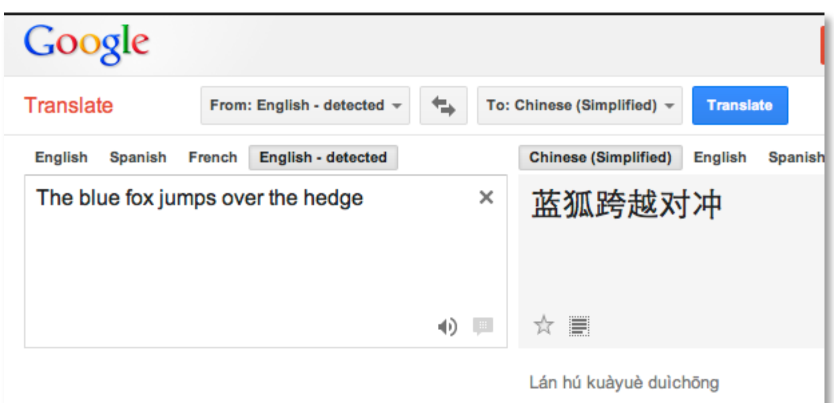
Online Classes

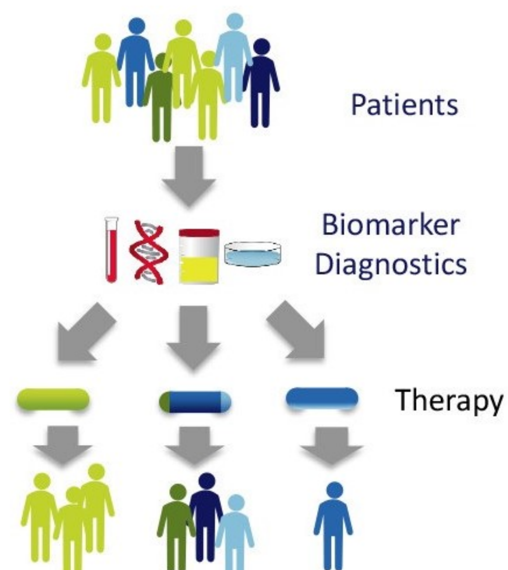
- Zoom conference call for class lectures
- Log in at northeastern.zoom.us
  - Upload a profile picture
  - Turn video on
  - Mute when not speaking
- Provide feedback
- To ask questions:
  - Raise hand
  - Use chat
- Discussion via breakout rooms
- Recording will be posted in Canvas



Machine Learning is Everywhere







Survey: Applications of ML

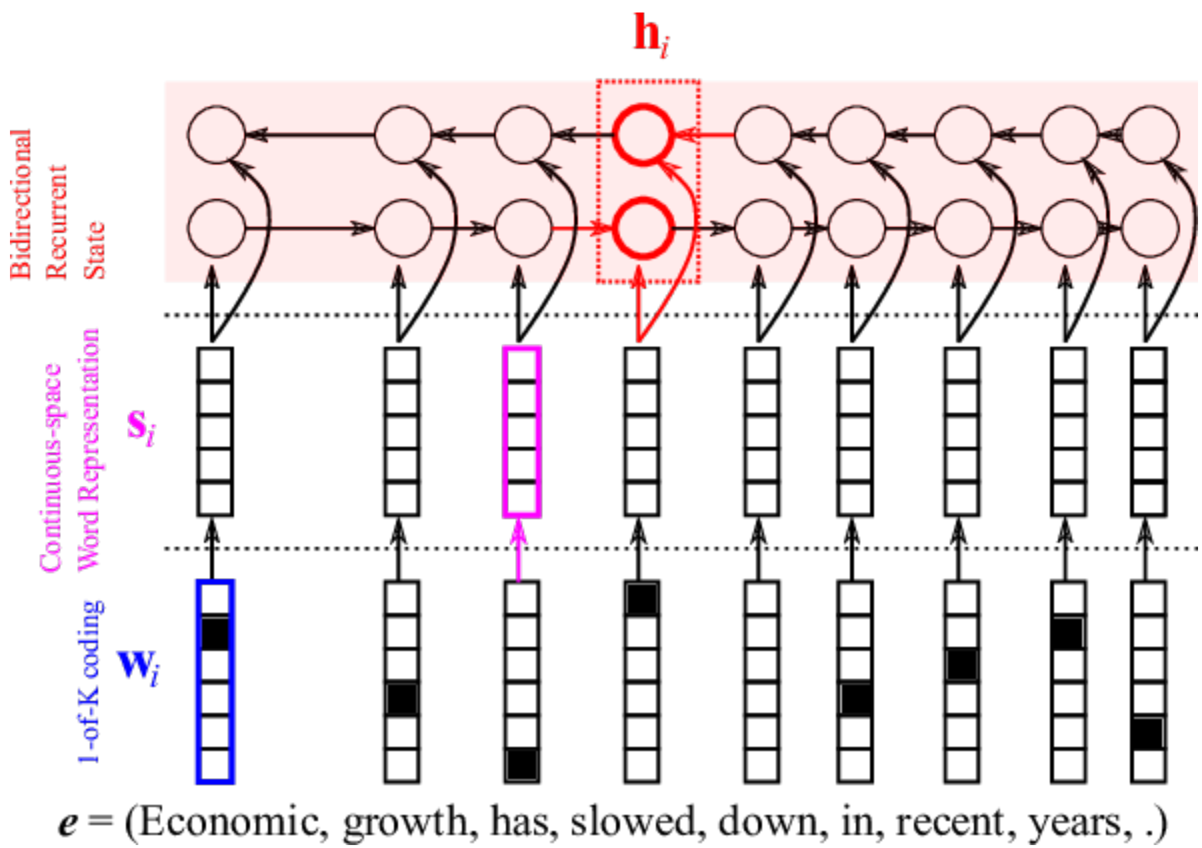
- Healthcare
- Vision
- NLP
- Speech recognition
- Self-driving cars
- Stock market analysis
- Recommendations
- Sentiment analysis
- Human behavior
- Quality of life
- Business
- Sports
- Bots / chatbots
- Science / engineering
- Bioinformatics
- Precision medicine
- Unsupervised learning
- Reinforcement learning

Class Breakout Intro

- Activity and discussion
  - Introduce to each other
  - Discuss most exciting ML applications
  - What are some of the concerns when using ML in the real world?

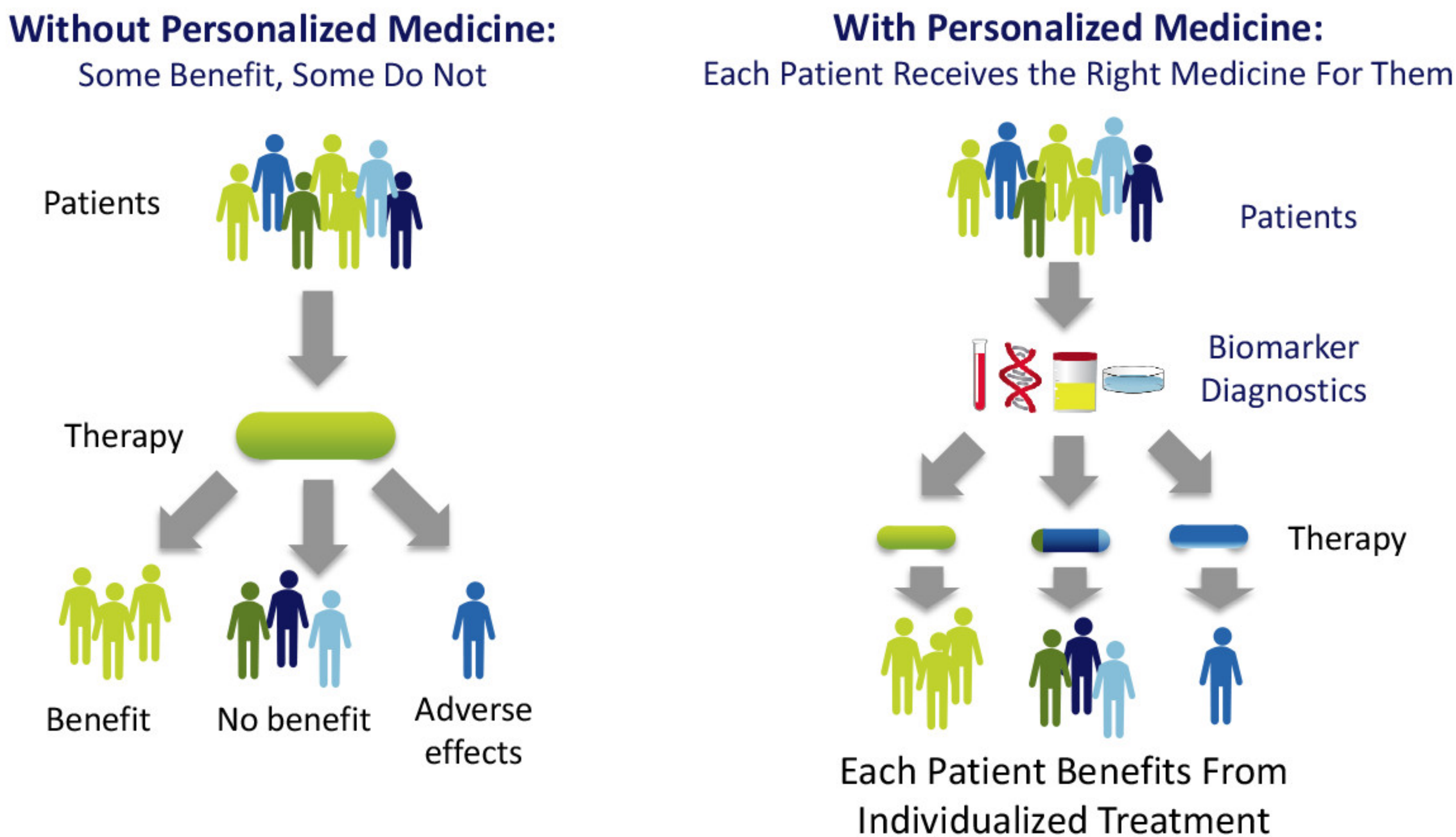
BREAKOUT
ROOMS

Natural Language Processing (NLP)



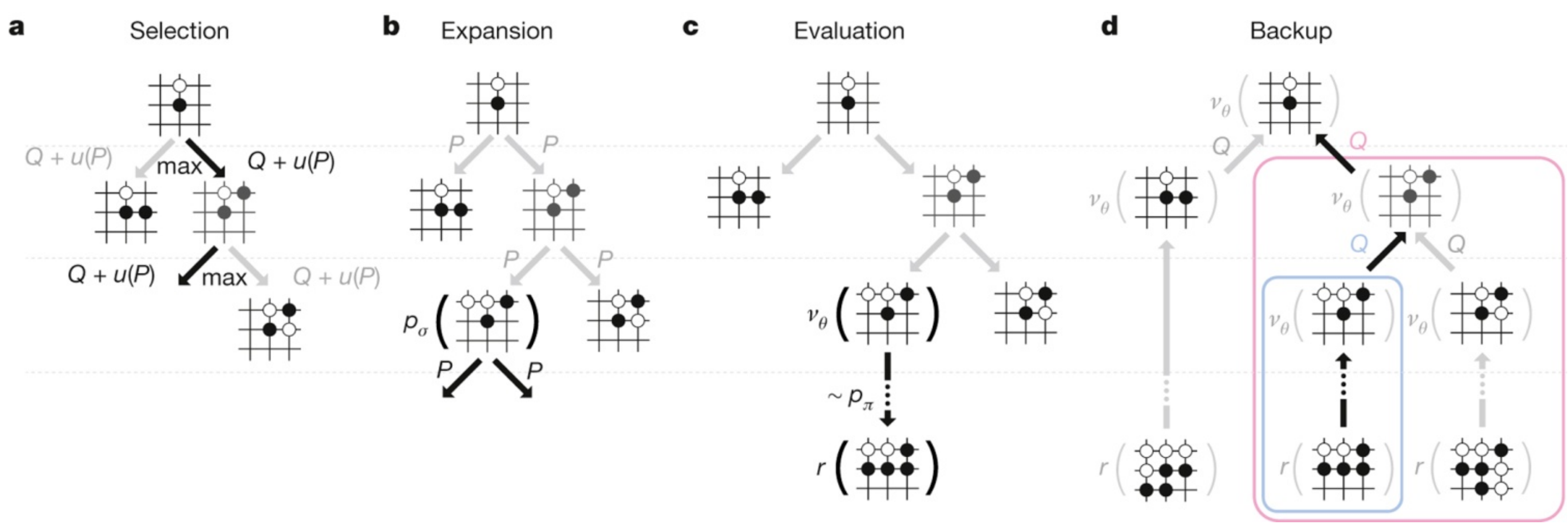
- Understand language semantics
- Real-time translation, speech recognition

Personalized medicine



- Treatment adjusted to individual patients
- Predictive models using a variety of features related to patient history and genetics

Playing games



Reinforcement learning

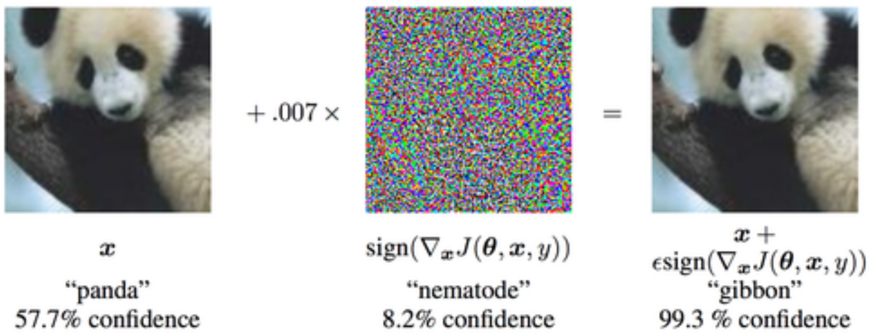
- AlphaGo
- Chess

Safety Concerns of AI

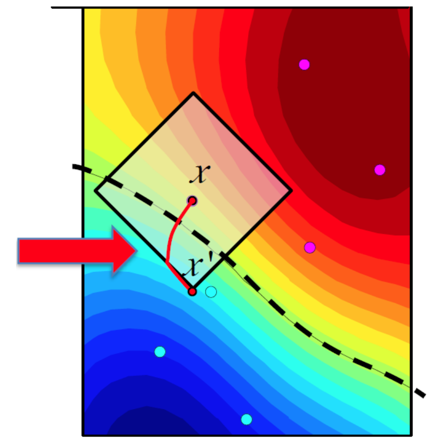
- Ethics and fairness of AI
  - Everyone is treated fairly
  - Robots will not perform harmful actions
  - Can the technology be used for nefarious purposes?
- Economic concerns
  - Might automate / displace some type of jobs in manufacturing, transportation, etc.
- Adversarial ML
  - ML can be manipulated
  - Small change in input results in different prediction

Adversarial Attacks on ML

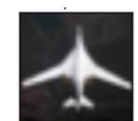
Evasion



Adversarial example



Poisoning



Data

Labels

Training Data

ML model

Wrong Prediction

Short History

- Legendre and Gauss – linear regression, 1805
  - Astronomy applications
- Probabilistic models
  - Bayes and Laplace - Bayes Theorem, 1812
  - Markov chains, 1913
- Fisher – linear discriminant analysis for classification, 1936
  - Logistic regression, 1940
- Widrow and Hoff ADELINE neural network, 1959
- Nelder, Wedderburn, generalized linear models, 1970
- “AI winter”, limitations of perceptron and linear models, 1970
- Breiman, Friedman, Olshen, Stone, decision trees (non-linear models), 1980
- Cortes and Vapnik , SVM with kernels, 1990
- IBM Deep Blue beats Kasparov at chess, 1996
- Geoffrey Hinton, Deep learning, back propagation, 2006
- C. Szedegy: Adversarial manipulation of image classification, 2013

DS-4400

- What is *machine learning*?
  - The science of teaching machines how to learn
  - Design predictive algorithms that learn from data
  - Replace humans in critical tasks
  - Subset of Artificial Intelligence (AI)
- Machine learning very successful in:
  - Machine translation
  - Precision medicine
  - Recommendation systems
  - Self-driving cars
- Why the hype?
  - Availability: data created/reproduced in 2010 reached 1,200 exabytes
  - Reduced cost of storage
  - Computational power (cloud, multi-core CPUs, GPUs)

DS-4400 Course objectives

- Become familiar with main machine learning tasks
  - Supervised learning vs unsupervised learning
  - Classification vs Regression
- Study most well-known algorithms
  - Regression (linear regression, spline regression)
  - Classification (SVM, decision trees, Naïve Bayes, ensembles, etc.)
  - Deep learning (different neural network architectures)
- Learn the theory and foundation behind ML algorithms and learn to apply them to real datasets
- Learn about security challenges of ML
  - Introduction to adversarial ML

https://www.ccs.neu.edu/home/alina/classes/Fall2020/

Class Outline

- Introduction – 1 week
  - Probability and linear algebra review
- Linear regression – 2 weeks
- Classification - 5 weeks
  - Linear classifiers: logistic regression, LDA,
  - Non-linear: kNN, decision trees, SVM, Naïve Bayes
  - Ensembles: random forest, boosting
  - Model selection, regularization, cross validation
- Neural networks and deep learning – 2 weeks
  - Back-propagation, gradient descent
  - NN architectures (feed-forward, convolutional, recurrent)
- Ethics of AI – 1 week
- Adversarial ML – 1 lecture
  - Security of ML at testing and training time

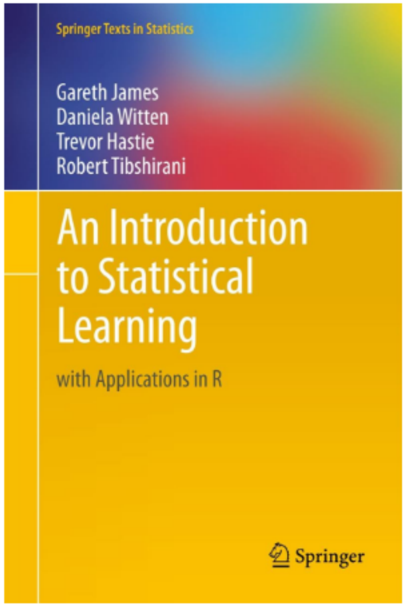
Textbook

An Introduction to Statistical Learning

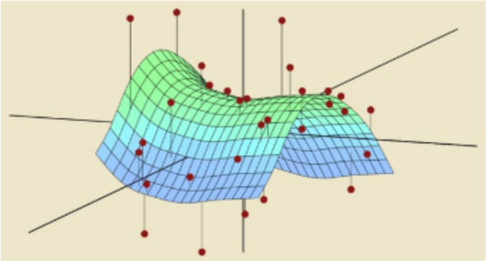
with Applications in R

Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani

- Home
- About this Book
- R Code for Labs
- Data Sets and Figures
- ISLR Package
- Get the Book
- Author Bios
- Errata



Download the book PDF
(corrected 7th printing)



Statistical Learning MOOC covering the entire ISL book offered by Trevor Hastie and Rob Tibshirani. Start anytime in self-paced mode.

This book provides an introduction to statistical learning methods. It is aimed for upper level undergraduate students

Specific chapters will be covered

Other resources

- Trevor Hastie, Rob Tibshirani, and Jerry Friedman, Elements of Statistical Learning, Second Edition, Springer, 2009.

- Christopher Bishop. Pattern Recognition and Machine Learning. Springer, 2006.

- A. Zhang, Z. Lipton, and A. Smola. Dive into Deep Learning

- Lecture notes by Andrew Ng from Stanford

Policies

- Instructors
  - Alina Oprea
  - TAs: Alex Wang, Matthew Jagielski
- Schedule
  - Tue 11:45am – 1:25pm, Thu 2:50-4:30pm EST
  - Zoom
  - Office hours:
    - Alina: Tue 4:00-5:30pm; Thu 4:30 – 5:30 pm (Zoom)
    - Matthew: Monday 3:00-4:00pm; Friday 9:00-10:00am (Zoom)
    - Alex: Wednesday: 5:00-7:00pm
    - Links on Canvas under "Syllabus"
- Online resources
  - Slides / recordings will be posted after each lecture
  - Use Piazza for questions
  - Canvas as course management system

Policies, cont.

- Your responsibilities
  - Please be on time, attend classes, and take notes
  - Participate in interactive discussion in class
  - Submit assignments/ programming projects on time
- Late days for assignments
  - 5 total late days, after that loose 20% for every late day
  - Assignments are due at 11:59pm on the specified date
  - We will use Gradescope for submitting assignments
  - No need to email for late days

Grading

- Assignments – 25%
  - 4-5 assignments and programming exercises based on studied material in class
- Final project – 35%
  - Select your own project based on public dataset
  - Submit short project proposal and milestone
  - Presentation at end of class (10 min) and written report
  - Team of 2 students
- Exam – 35%
  - One exam second half of November
  - Tentative date: November 19
- Class participation – 5%
  - Participate in class discussion/Zoom and on Piazza

Assignments

- Mostly programming exercises, occasionally some theory questions
- Language
  - Use R or Python
  - Jupyter notebooks recommended
- Submission
  - Submit PDF report
  - Includes all the results, as well as link to code and instructions to run it

Final project

- Goal: work on a larger data science project
  - Build your portfolio and increase your experience
- Requirements
  - Large dataset: at least 10,000 records (public source)
  - Not recommended to collect your own data
  - Pick application of interest, but instructor will also provide potential list of projects
  - Experiment with at least 3 ML models
  - Perform in-depth analysis (which features contribute mostly to prediction, which model performs best)
  - Teams of 2 students, will have a TA assigned
- Timeline
  - Proposal: mid class; milestone 3 weeks after (Instructors will provide early feedback)
  - Final presentation (10 mins) and report (5-6 pages)

Academic Integrity

- Homework is done individually!
- Final project is done in the team!
- Rules
  - Can discuss with colleagues or instructors
  - Can post and answer questions on Piazza
  - Code cannot be shared with colleagues
  - Cannot use code from the Internet
    - Use python or R packages, but not directly code for ML analysis written by someone else
- NO CHEATHING WILL BE TOLERATED!
- Any cheating will automatically result in grade F and report to the university administration
- http://www.northeastern.edu/osccr/academic-integrity-policy/