CMPSC 448: Machine Learning

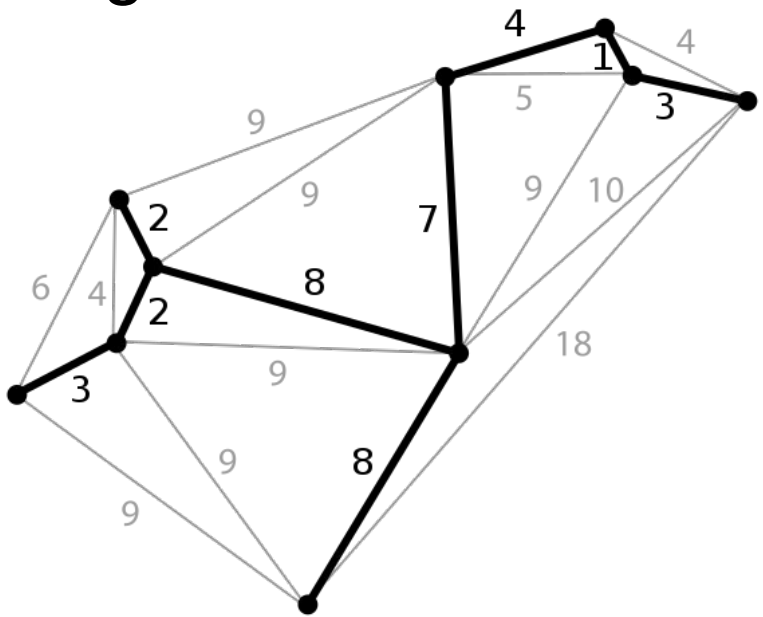
Lecture 1. Introduction & Logistics

Rui Zhang

Fall 2021

PennState

Minimum Spanning Tree



A classical problem in algorithm design: **Minimum Spanning Tree**

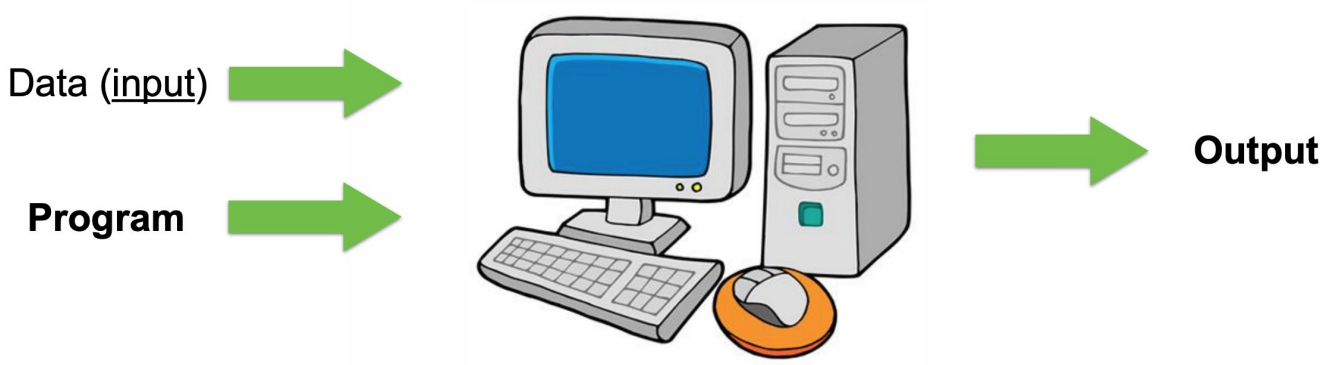
Input: A graph with cost for edges

Output: A spanning tree with minimum cost

Prim's algorithm, Kruskal's algorithm,...

Traditional Computer Science

- We are given a well-defined description of a problem to map a given input to its corresponding output.
- Input-output relations are well-defined.
- E.g., sorting, shortest path, spanning trees, scheduling, etc.
- The goal is to write a computer program that efficiently generates an output for a given input.



Algorithmic Problem Solving

Three important problem-solving tools

- **Data models**, the abstractions used to describe problems
  - graphs, trees, lists, etc

- **Data structures**, the programming-language constructs used to represent data models
  - stack, queue, binary trees, undirected graphs etc

- **Algorithms**, the techniques used to obtain solutions by manipulating data as represented by the abstractions of a data model, by data structures, or by other means
  - dynamic programming, greedy algorithms, divide & conquer, etc

Can I Eat This Mushroom?

I do not know what type it is?
I have never seen it before?
Is it edible or poisonous?

Data-Driven Problem Solving

Suppose we are given examples of edible and poisonous mushrooms:

- edible



- poisonous




Input/output relationship is difficult to specify.

Can we write a program to classify other (probably not seen) mushrooms?

Data-Driven Problem Solving

“**Machine Learning**” is a data-driven problem solving strategy!

- edible



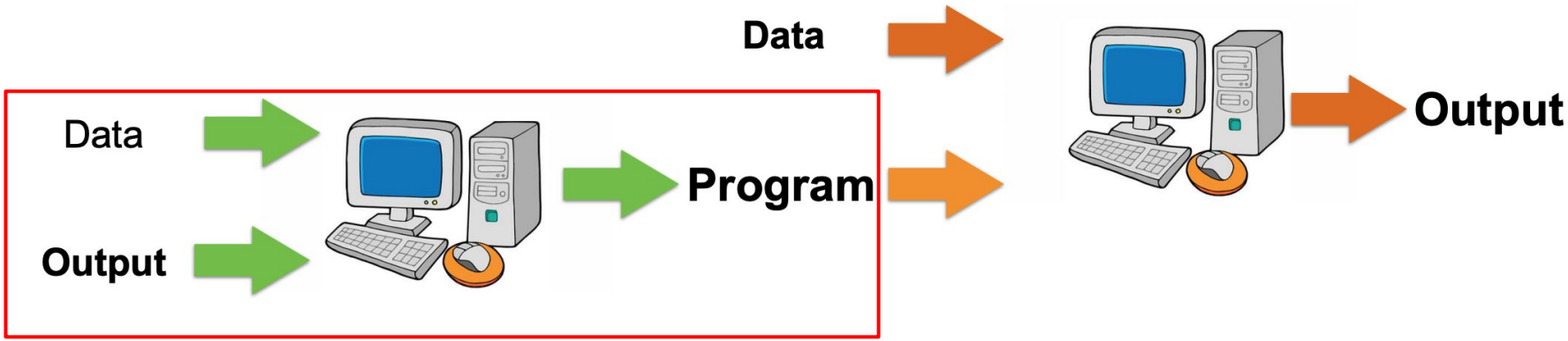
- poisonous




We are given a bunch of inputs and corresponding outputs!
Little knowledge on how the mapping works (sometimes even impossible to model the exact mapping)

Data-Driven Problem Solving

"**Machine Learning**" is a data-driven problem solving strategy!



Common theme is to solve a prediction problem:

Given an input $\boldsymbol{x}$,

Predict an "appropriate" output y (label, decision, action, etc).

Machine Learning

What is “Machine Learning”?

“Learning” (in nature): Using past experience to make future decisions or guide future actions

“Machine Learning” (an engineering paradigm): Using data and examples, instead of expert knowledge, to automatically create systems that perform complex tasks (e.g., make predictions or decisions)

Generalization

The ultimate goal of machine learning is **generalization**! Making future prediction on unseen examples as accurate as possible!

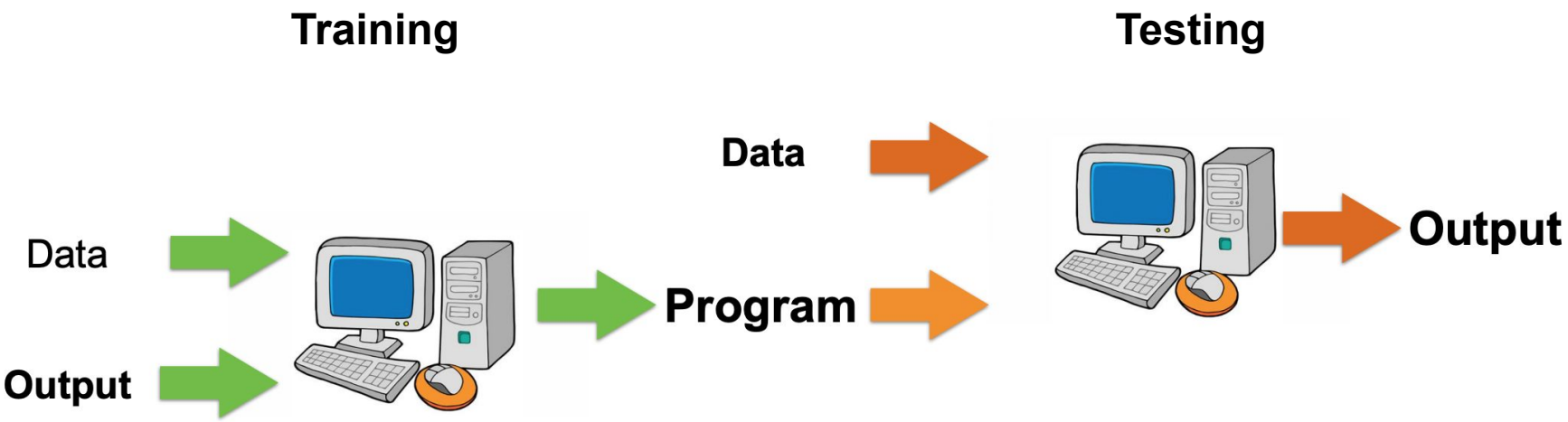


Image Recognition

“Machine Learning” is a data-driven problem solving strategy!

Data: a set of images
Output: the label of each image (dog or cat)

Question. Given a set of images and their labels (dog or cat), can you write a program that can label an input image as dog or cat for a photo you taken by your phone?

Credit Prediction

“Machine Learning” is a data-driven problem solving strategy!

- Using salary, debt, years in residence, etc., approve for credit or not!
- No magic credit approval formula.
- Banks have lots of data, whether or not they defaulted on their credit, and customer information such as salary, debt, etc.

Feature	Value
Age	32
Gender	Male
Salary	40,000
Debt	26,000
Years in job	2
Years at home	3

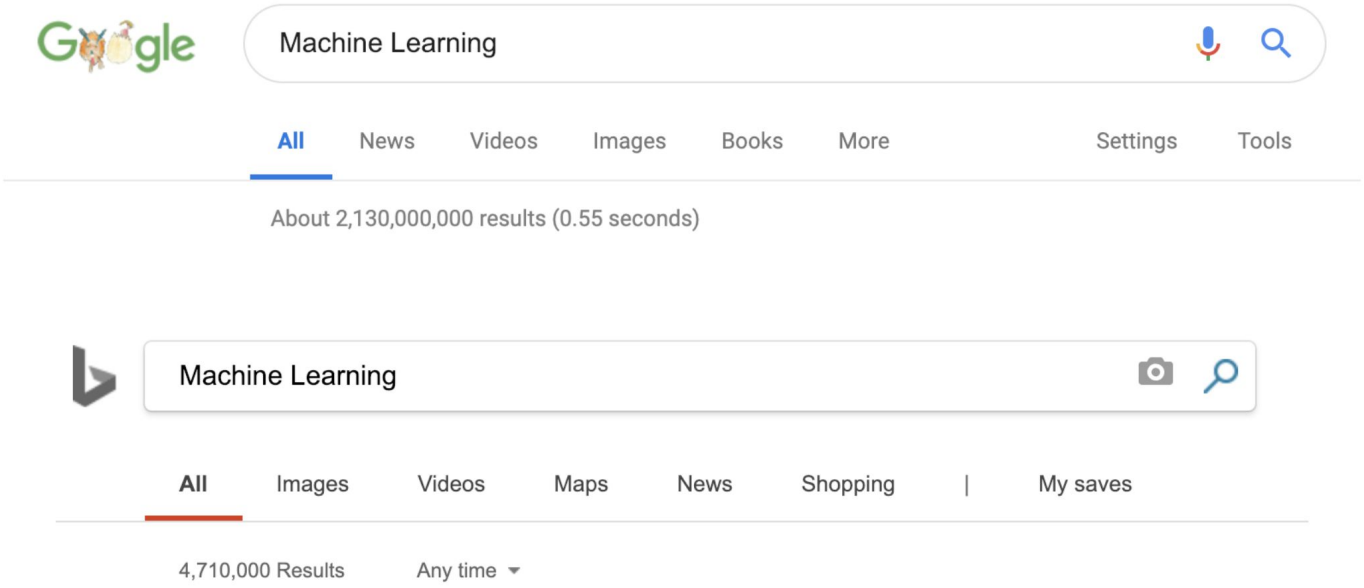
A pattern exists. We don’t know it. We have data to learn it.

Movie Recommendation

The Netflix challenge: improve our accuracy by 10% and win $1M

- No magic rating formula
- But, we have a lot of ratings and reviews issued by users!
- Can we learn from available data to learn the taste of future users and unseen movies for existing users?

Search Engines

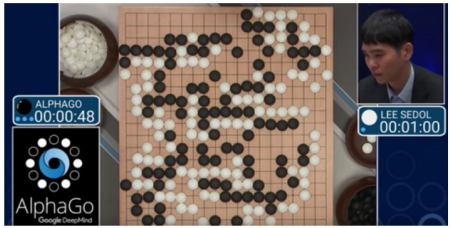


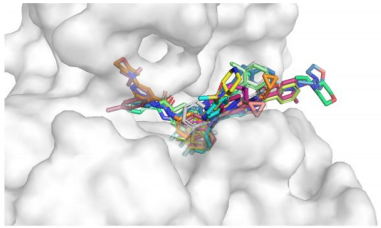
- How can you rank millions of documents to show top 10 to a user?
- We have millions of feedback from users who issued the same or similar query (search engine logs)

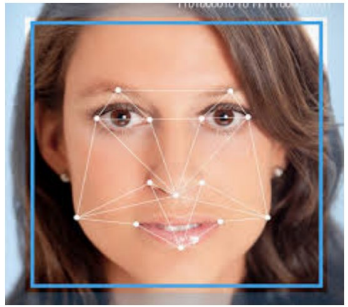
Healthcare

- Cancer detection
- Drug discovery

Machine Learning is Everywhere





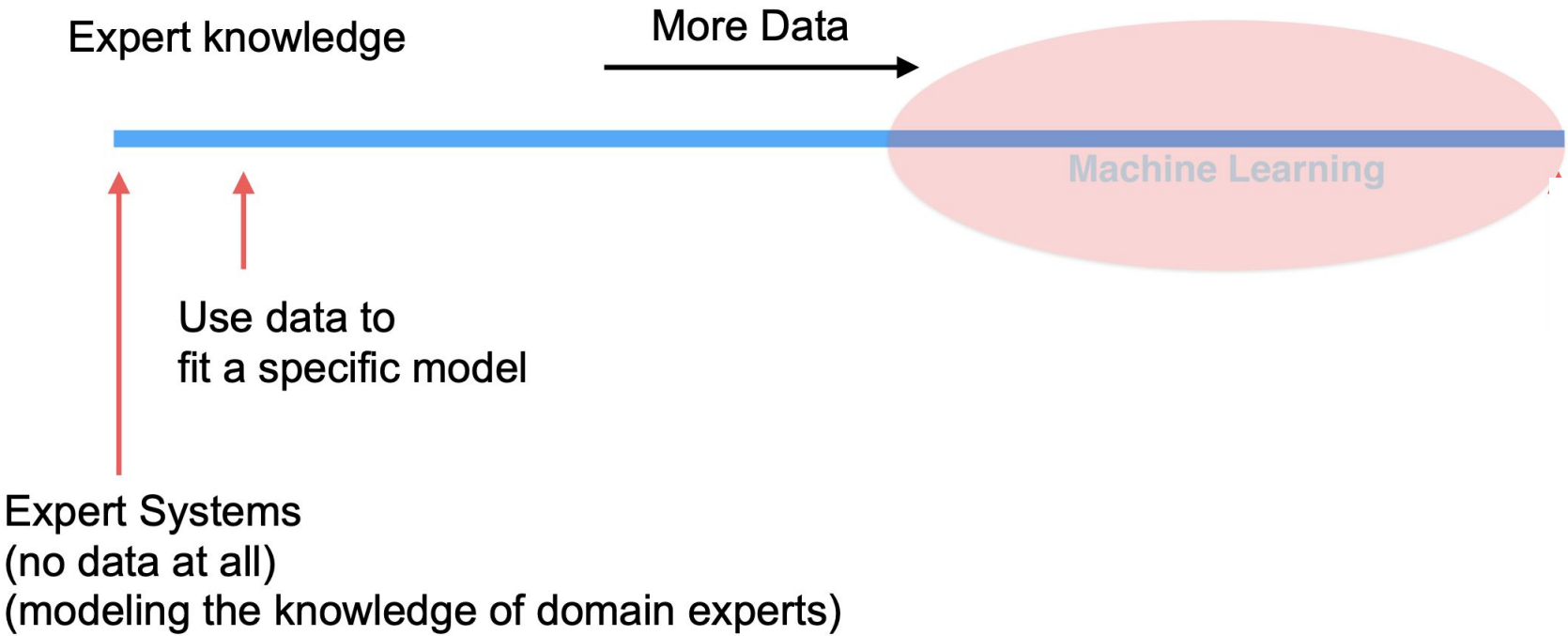


2σ
TWO SIGMA
VENTURES

Machine learning is starting to take over decision-making in many aspects of our life, including:

- Keeping us safe on our daily commute in self-driving cars
- Making accurate diagnoses based on our symptoms and medical history
- Pricing and trading complex securities
- Discovering new science, such as the genetic basis for various diseases.

More data, less expert knowledge



What types of ML are there?

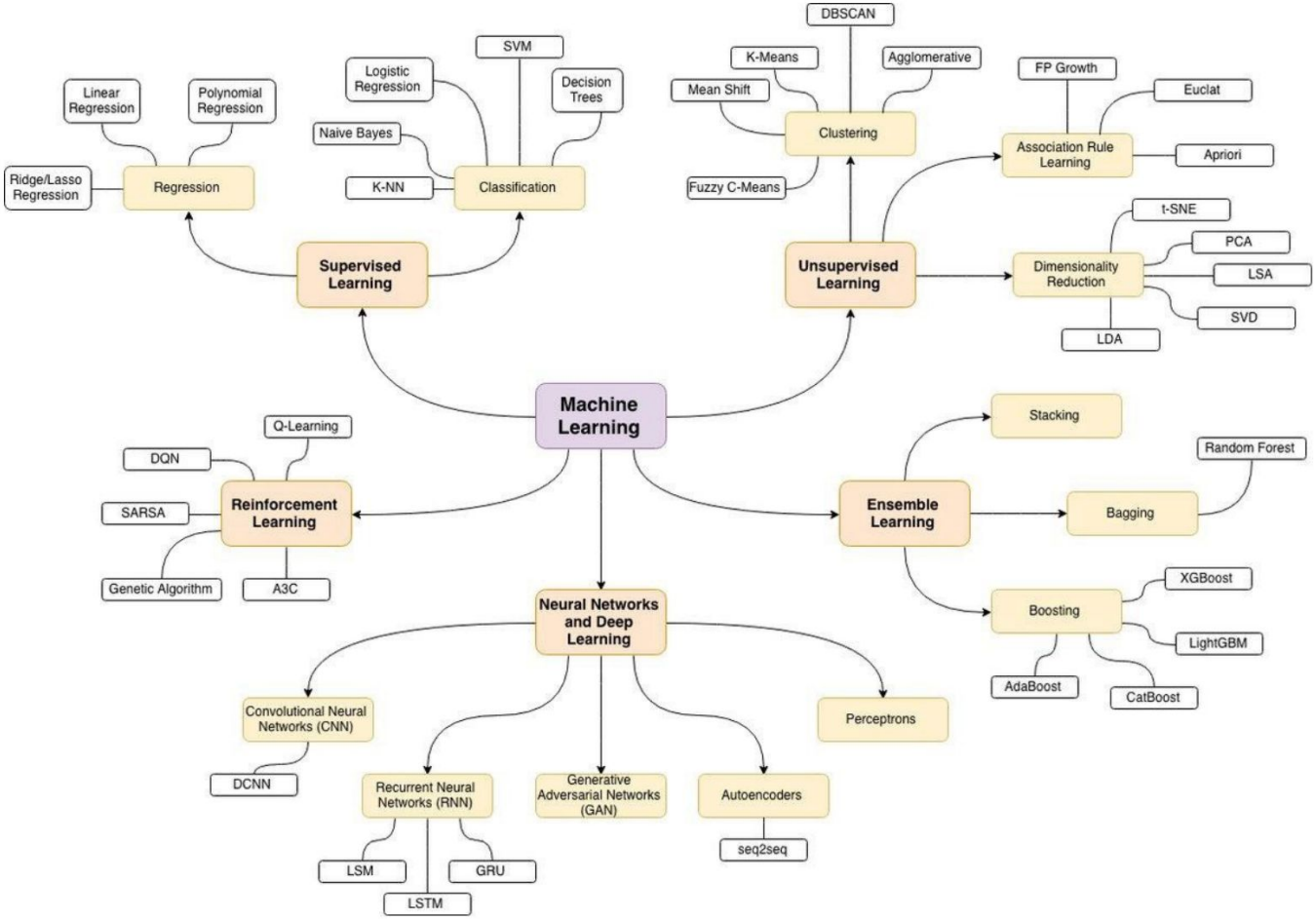
- The nature of data (statistical, adversarial, benign, etc)
- The availability of outputs
- The availability of data (streaming, batch, etc)
- Interaction with environment

Supervised Learning

Unsupervised Learning

Reinforcement Learning

What types of ML are there?



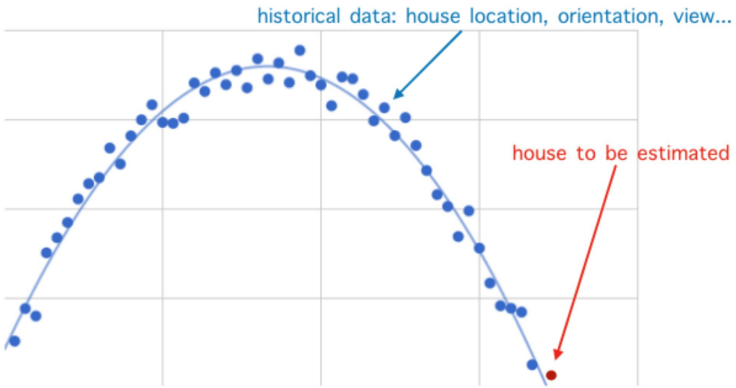
Supervised Learning

Given labeled examples, find the right prediction of an unlabeled example. (e.g. given annotated images learn to detect faces)

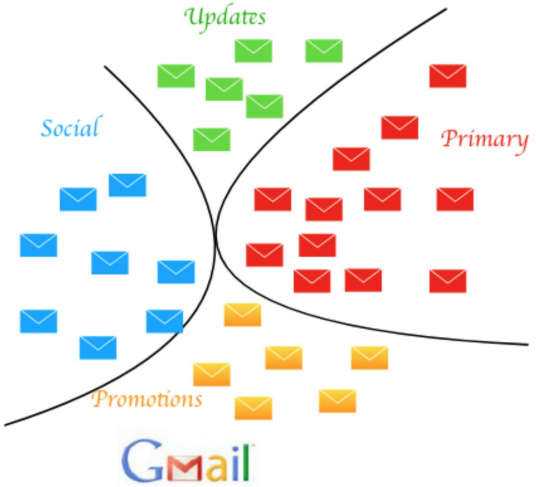
- Regression
- Binary classification (the label set is binary, e.g., spam detection)
- Multiclass classification (the label set is larger than two, digit recognition)
- Online learning (adversarial learning) (data are adversarial and come in a stream, e.g., stock prediction, weather prediction)
- Ranking (ranking instances, e.g., Search Engines)

Regression vs Classification

- Tracking a mortgage
- Houses data
- Banks: Estimate the loan based on location, orientation, view, etc
- Output values: continuous



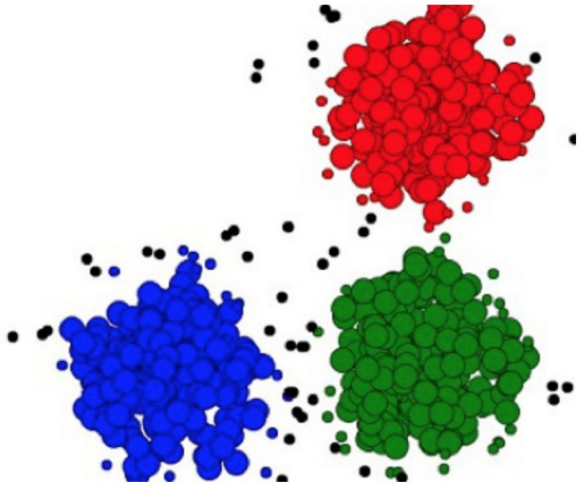
- Spam classification
- Incoming emails
- How to group email in categories
- Output values: discrete, categorical



Unsupervised Learning

Given data try to discover similar patterns, structure, sub-spaces (e.g. automatically cluster news articles by topic)

- No labeled are available
- But there is a pattern in data



Reinforcement Learning

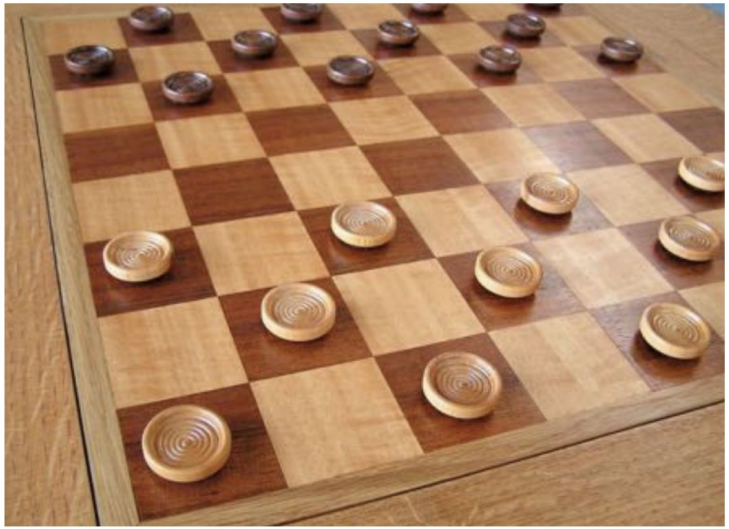
Try to learn from delayed feedback (e.g. robot learns to walk, fly, play chess)

- No training data at the beginning
- We have to interact with an unknown environment to learn

A Brief History of ML

1952: Game of Checkers

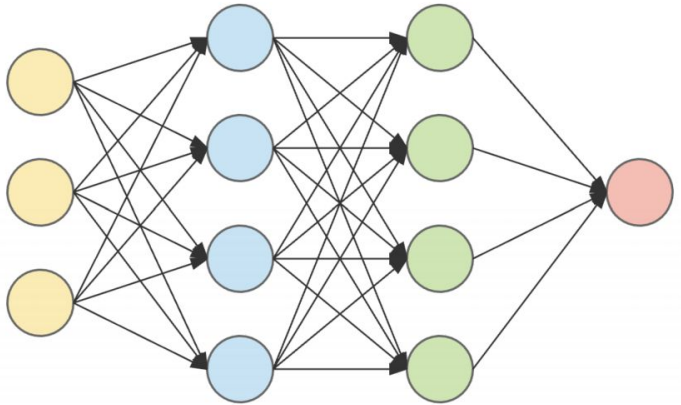
- Basically Shannon’s Minimax Algorithm (traditional AI)
- Included simple learning algorithm to improve board evaluation
- Player improved over time!!

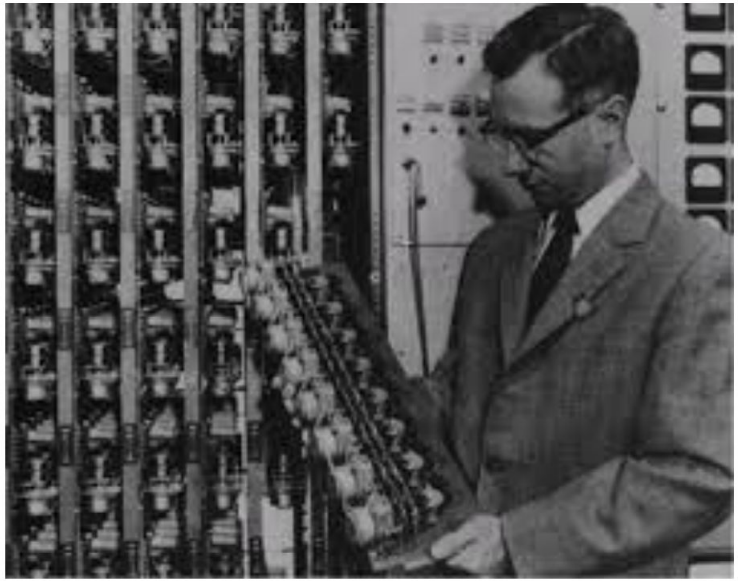


1957: Birth of Perceptron

A new electronic brain which hasn't been built, but which has been successfully simulated on the IBM 704 computer.

- Provable convergence properties
- Eventually leads to multilayer perceptron = Artificial Neural Networks = Deep Learning





Frank Rosenblatt

The invention of the perceptron generated a great deal of excitement and was widely covered in the media

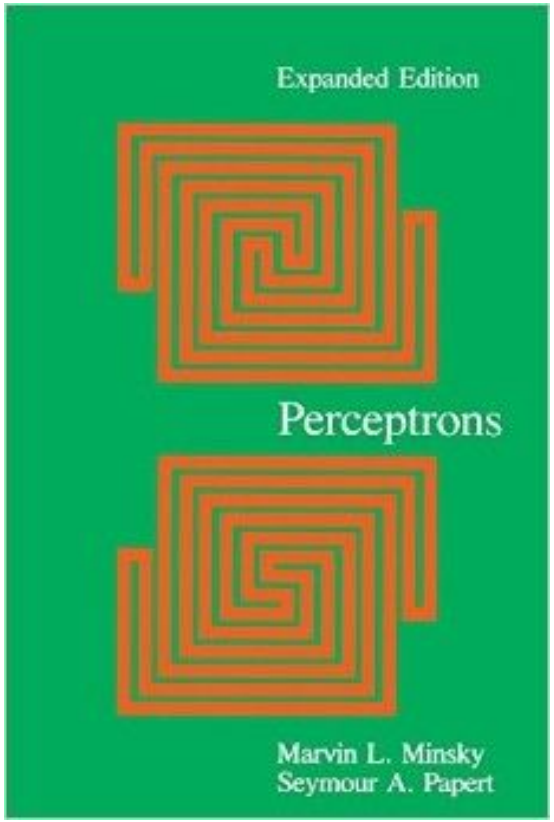
1969: AI Winter

Minsky & Papert “killed” AI

Mainstream research into perceptrons came to an abrupt end in 1969, when Marvin Minsky and Seymour Papert published the book *Perceptrons*, which was perceived as outlining the limits of what perceptrons could do (e.g., cannot compute XOR).

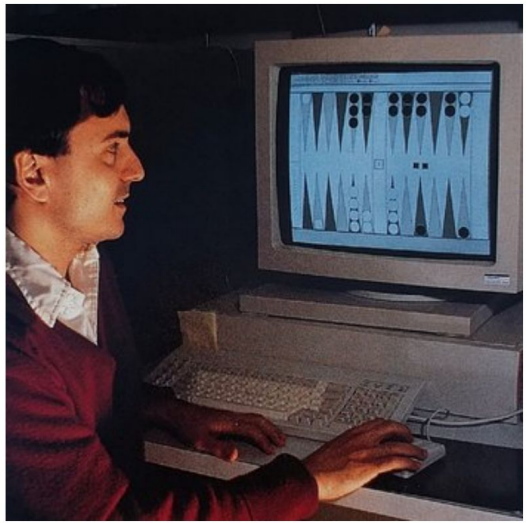
- Burst huge expectation bubble
- Funding for AI research collapsed for decades

The agencies which funded AI research (such as the British government, DARPA and NRC) became frustrated with the lack of progress and eventually cut off almost all funding for undirected research into AI (Wiki)



1994: TD-Gammon

- Gerry Tesauro (IBM) teaches a neural network to play Backgammon.
- The net plays 100K+ games against itself and beats world champion [Neurocomputation 1994]
- Algorithm teaches itself how to play so well using trained using temporal-difference(TD) learning !!!



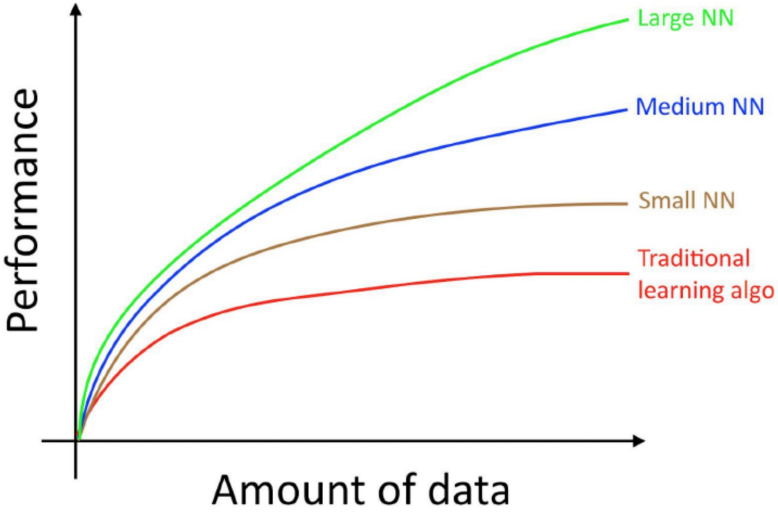
1997: IBM's Deep Blue

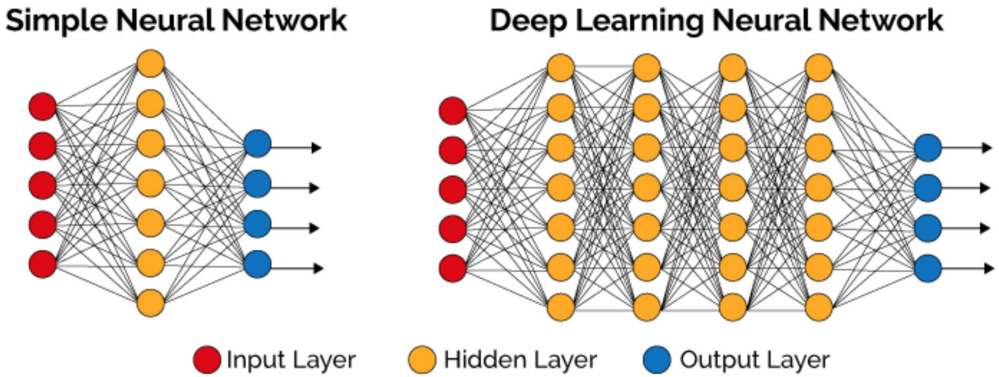
IBM’s Deep Blue wins against Kasparov in chess.

Crucial winning move is made due to Machine Learning (G. Tesauro)

2006: Deep Learning

Thanks to GPUs, huge amount of data, and some tweaks to traditional NNs with deep NNs was born!

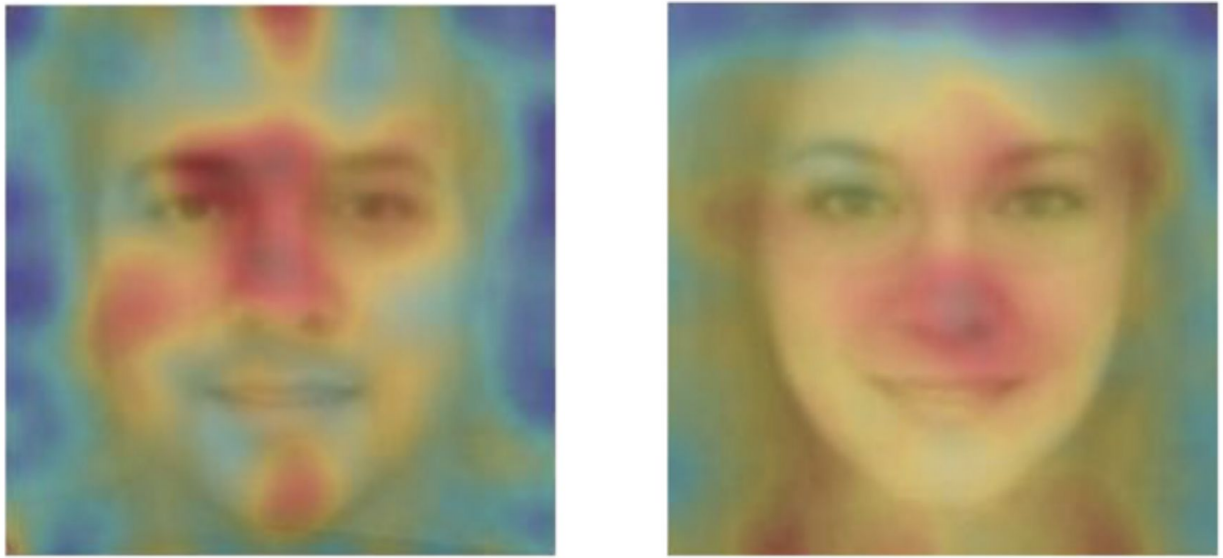




2011: IBM’s Watson

The Watson computer system competed on Jeopardy! against legendary champions Brad Rutter and Ken Jennings winning the first place prize of $1 million

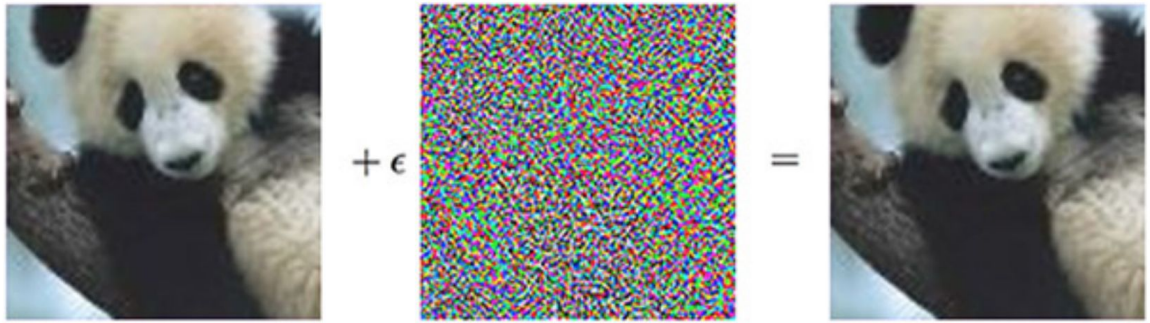
Privacy



Deep neural networks are more accurate than humans at detecting sexual orientation from facial images

[Wang and Kosinski 2017]

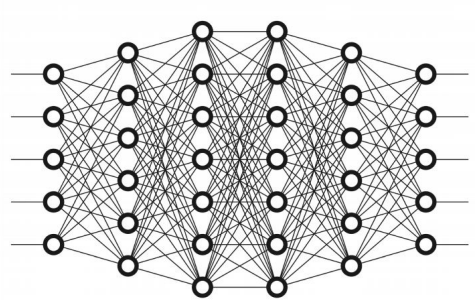
Security and Trustworthiness



Can we trust self-driving cars?

Interpretability & Human-in-Loop

Hey doctor, can you give feedback on my new cancer detection model?



Logistics

Course Information

Lecture

MWF, 9:05-9:55 am @ Leonhard Bldg 102

Instructor

Rui Zhang (https://ryanzhumich.github.io/)

Zoom Office Hour: https://psu.zoom.us/j/4954771701 Wednesday 5pm - 7pm

TA

Wenjian Miao and Ting-Yao Hsu

Office Hour: TBD

For most questions, please use Piazza.

Please only use Canvas email for personal issues.

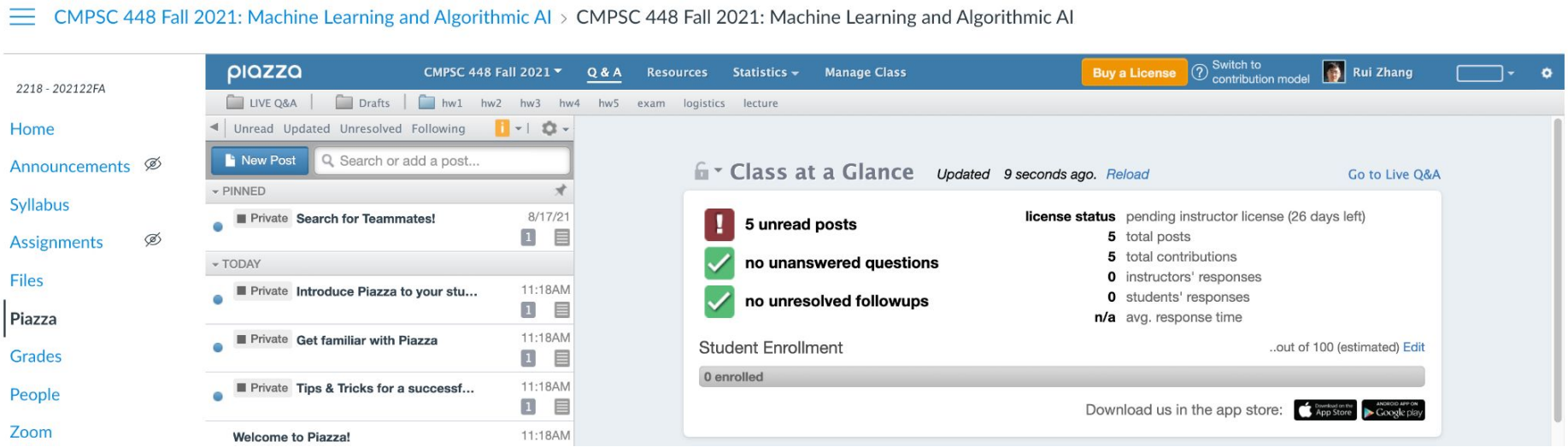
Don't use psu emails: It's hard for me to monitor class status using emails.

Canvas Page


CMPSC 448 Fall 2021: Machine Learning and Algorithmic AI > Syllabus
2218 - 202122FA
Home
Syllabus
Files
Piazza
Grades
People
Zoom
CMPSC 448 Fall 2021: Machine Learning and Algorithmic AI
Jump to Today
Course Syllabus
Course Schedule
Lecture
Time and Location: MWF, 9:05 am -9:55 am @ Leonhard Bldg 102
Mask
Penn State University requires everyone to wear a face mask in all university buildings, including classrooms, regardless of vaccination status. ALL STUDENTS MUST wear a mask appropriately (i.e., covering both your mouth and nose) while you are indoors on campus. This is to protect your health and safety as well as the health and safety of your classmates, instructor, and the university community. Anyone attending class without a mask will be asked to put one on or leave. Instructors may end class if anyone present refuses to appropriately wear a mask for the duration of class. Students who refuse to wear masks appropriately may face disciplinary action for Code of Conduct violations. If you feel you cannot wear a mask during class, please speak with your adviser immediately about your options for altering your schedule.
Contact
For most questions, please use Piazza. Please only use Canvas email for personal issues.

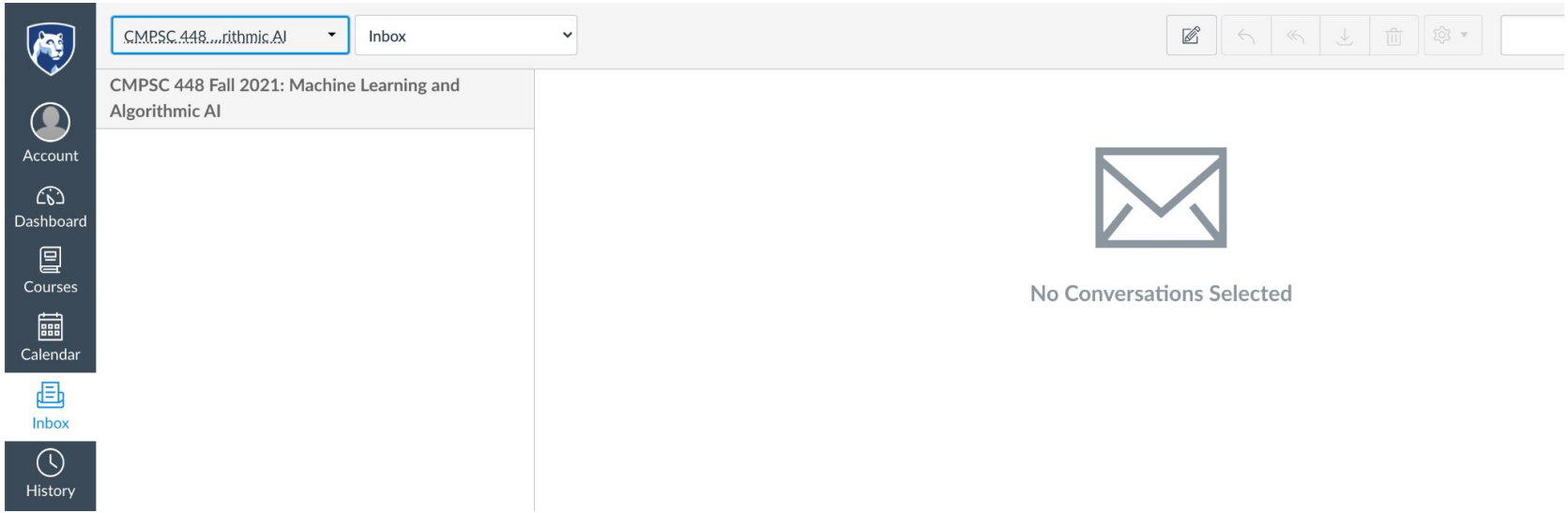
Piazza

For most questions, please use Piazza.



Canvas Email

Please only use Canvas email for personal issues.



Zoom


CMPSC 448 Fall 2021: Machine Learning and Algorithmic AI
CMPSC 448 Fall 2021: Machine Learning and Algorithmic AI
2218 - 202122FA
Home
Announcements
Syllabus
Assignments
Files
Piazza
Grades
People
Zoom
zoom
Your current Time Zone and Language are (GMT-04:00) Eastern Time (US and Canada), English
All My Zoom Meetings/Recordings
Schedule a New Meeting
Upcoming Meetings
Previous Meetings
Personal Meeting Room
Cloud Recordings
Get Training
Show my course meetings only
Start Time
Topic
Meeting ID
Mon, Aug 23 (Recurring)
9:00 AM
CMPSC 448 Fall 2021: Machine Learning and Algorithmic AI
932 3575 7587
Start
Delete

Textbooks

Not required.

- *Pattern Recognition and Machine Learning*. Christopher Bishop. available online
- *A Course in Machine Learning.* Hal Daumé III. available online
- *Machine Learning: A Probabilistic Perspective*. Kevin Murphy
- *Elements of Statistical Learning*. Hastie, Tibshirani, Friedman. available online
- *Reinforcement Learning: An Introduction*. Sutton and Barto. available online
- Dive into Deep Learning. Alex Smola et al. available online

You are strongly encouraged to read beyond slides.

Useful Resources

Math Resources

- Mathematics for Machine Learning: https://mml-book.github.io/
- Linear Algebra on Khan Academy: https://www.khanacademy.org/math/linear-algebra

Prerequisites

- Linear Algebra
- Multivariate Calculus
- Probability and Statistics
- Programming Skills

It is **strongly recommended** that you finish two courses (STAT 319 or STAT 415 and CMPSC 122 or prior programming experience) in these topics first before taking this course.

Course Topics

Background

- Calculus
- Linear algebra
- Probability, statistics, information theory
- Basic convex analysis and optimization (Gradient Descent, Stochastic GD)

Basics of Machine Learning

- The process of learning
- Training, evaluation, & generalization
- Model selection
- Regularization
- Bias-variance decomposition

Supervised Learning

- Regression (OLS, Ridge, Lasso)
- Perceptron
- Logistic Regression
- Nearest Neighbor
- Decision Trees
- Support Vector Machines
- Ensemble Methods (Bagging, Boosting)

Unsupervised Learning

- Clustering (k-means, k-means++, Mixture of Gaussians)
- Principal Component Analysis (PCA)
- Matrix Factorization

Reinforcement Learning

- Bandits
- Markov Decision Process
- Dynamic Programming
- SARSA and Q-learning

Schedule

	A	B	C	D	E	F
1	Week	Date	Day	Lecture	Topics	Homework
2	Part 1: The Basics of Machine Learning and Background					
3	1	8/23	Monday	Logistics & Introduction	What is machine learning: A brief history and applications	
4		8/25	Wednesday	The Processes of Learning	Data-driven problem solving, learning stages, feature engineering, training, model selection, generalization, optimization, and prediction	
5		8/27	Friday	Overview of key concepts	Overfitting, Underfitting, Regularization, The bias-variance decomposition	HW1 Out
6	2	8/30	Monday	Background	Linear algebra, vector and matrix calculus, Basics of convex analysis and optimization Exploratory Data Analysis (EDA) with pandas	
7		9/1	Wednesday			
8		9/3	Friday			
9	Part 2: Supervised Learning					
10	3	9/6	Monday	Labor Day - No Class		
11		9/8	Wednesday	Regression	The Ordinary Least Squares (OLS) and its solution, the geometric interpretation	
12		9/10	Friday		Regularization for regression: Geometric and algebraic motivation, Ridge Regression, Lasso, and Principle Component Regression (PCR), Gradient Descent (GD) and Stochastic Gradient Descent (SGD) for regression	HW2 Out
13	4	9/13	Monday			HW1 Due
14		9/15	Wednesday	Nearest Neighbors	Similarity measures, Decision boundaries, Model selection for NN, Metric learning, Efficient data structures for high-dimensional search (kd-tree, etc)	
15		9/17	Friday	Artificial Neural Networks	The binary classification, The Perceptron algorithm and its convergence, Multiple Layer Perceptron, A touch of Deep NN and its key challenges	
16	5	9/20	Monday	Logistic Regression	The predicting probabilities, The logit function, Logistic regression for classification and regularization, The exponential loss and loss minimization viewpoint, Gradient Descent (GD) for LR	
17		9/22	Wednesday			
18		9/24	Friday	Decision Trees	Entropy, Information gain, Decision Trees, Regularization	HW3 Out
19	6	9/27	Monday	Support Vector Machines	The margin intuition, distance of a point to a hyperplane, large-margin classification for separable data	HW2 Due
20		9/29	Wednesday		Hard-margin classification for non-separable data, The Hinge loss and loss minimization viewpoint of SVMs, SGD for largescale SVMs and a brief touch on kernel trick	
21		10/1	Friday			
22	7	10/4	Monday	Ensemble (Bagging and Boosting)	Ensemble learning, Bias-variance revisited, The Bootstrap aggregating The concept of Weak learners, The AdaBoost Algorithm, Introduction to Gradient Boosting and XGBoost algorithm	
23		10/6	Wednesday			
24		10/8	Friday			
25	8	10/11	Monday			HW3 Due
26		10/13	Wednesday	Midterm		

Schedule

25	8	10/11	Monday	Midterm		HW3 Due
26		10/13	Wednesday			
27		10/15	Friday			
28	Part 3: Unsupervised Learning					
29	9	10/18	Monday	Clustering	Introduction to unsupervised learning, k-means, k-means++	
30		10/20	Wednesday		Mixture of Gaussians and EM algorithm	
31		10/22	Friday			HW4 Out
32	10	10/25	Monday	Principle Component Analysis	Dimensionality reduction, minimum variance viewpoint, reconstitution error minimization viewpoint	
33		10/27	Wednesday			
34		10/29	Friday	Autoencoder	Power method and pitfalls of PCA, Autoencoder neural networks and connection to PCA	
35	11	11/1	Monday	Matrix Factorization	The Netflix problem and collaborative filtering	
36		11/3	Wednesday		Latent features and Matrix Factorization (MF), GD and SGD for solving MF, A brief introduction to low-rank models and Matrix Completion	
37		11/5	Friday			
38	Part 3: Reinforcement Learning					
39	12	11/8	Monday	Bandits	Reinforcement Learning, Multi Armed Bandits (regret, exploration versus exploitation, action valuation, greedy action selection, epsilon-greedy, Upper Confidence Bound (UCB), and Gradient Bandit algorithms), A/B Testing and its comparison to MAB	HW4 Due
40		11/10	Wednesday			
41		11/12	Friday	Markov Decision Processes	Finite Markov Decision Processes (MDP)	HW5 Out
42	13	11/15	Monday	Dynamic Programming	The key concepts: value function, state-action valuations, policies, and optimal policies, backup diagram, Bellman equation for an optimal policy	
43		11/17	Wednesday			
44		11/19	Friday		Dynamic Programming (DP) algorithm for solving Bellman recursive equations	
45	14			Thanksgiving - No Class		
46	15	11/29	Monday	Temporal Difference	The Temporal Difference algorithm, Sarsa and Q-learning algorithms	
47		12/1	Wednesday			
48		12/3	Friday			
49	Part 4: Advanced Topics					
50	16	12/6	Monday		Graphic Models, HMMs, Learning Theory, Multiclass SVM	HW5 DUE
51		12/8	Wednesday			
52		12/10	Friday			

Homeworks

5 bi-weekly homeworks, both on theory and programming.

Submit your write-up as a single PDF file and Python code on Canvas by 11:59 PM of the due date.

Not late homework accepted

- unless there is a compelling reason. Send me an Canvas email if you know you are going to be late.

Exams

There will be a midterm exam and a final exam.

- Midterm will cover the first two parts of the courses
- Final will be cumulative with more weight on after midterm material

For both exams, we use traditional in-person classroom format.

Grading

- 10%: Attendance and class participation
- 40%: Five Homeworks
- 20%: Midterm
- 30%: Final

Final letter grades will be curved and the cut scores will be determined after all grades are finalized.