Machine Learning CS229/STATS229

Instructors: Tengyu Ma and **Chris Ré**

The Teaching Team

Course Staff

Course Coordinator

John Cho

Head Course Assistant

Sauren Khosla

Course Assistants

Drew Kaul

Garrett Thomas

Hong Liu

Jeff Z. HaoChen

Kevin Li

Michihiro Yasunaga

Rui Deng

Shai Limonchik

Tengyu's Research: Machine Learning Tools/Theory

How do we

- train faster?
- pick the correct model (and hyperparameters)?
- regularize the models so that they can generalize with fewer samples to unseen scenarios?
- robustify the models?

Various settings:

- supervised learning
- unsupervised learning
- reinforcement learning

Tengyu Ma

My research threads Foundation Models, Systems+AI, Data-Centric viewpoint

- **Systems + ML/AI: 1st** thread of work since ~ 2010.
  - Scale up SGD, "*Hogwild*!" Scale impacts AI. Hardware, etc.
  - **Recent**: *FlashAttention* used in many foundation models, also MLPerf, Pytorch, etc. Tri Dao is amazing!
- **Data-Centric AI**: built systems (and companies) in the deep-learning era (sold companies, built products).
  - Snorkel 1st data-centric platform: data way to *"program ML/AI"* used in products from Google, Apple, etc.
- **Longer Sequence FMs.** Long sequence models via signal processing for foundation models (S4, H3, Hyena, etc.).
  - **Other**: theory to scale graphical models, joins, non-Euclidean geometry, distributed optimization, etc.

1. Administrivia

cs229.stanford.edu

(you may need to refresh to see the latest version)

2. Topics Covered in This Course

Pre-requisite

- Probability (CS109 or STAT 116)
  - distribution, random variable, expectation, conditional probability, variance, density
- Linear algebra (Math 104, Math 113, or CS205)
  - matrix multiplication
  - eigenvector
- Basic programming (in Python and NumPy)
- Will be reviewed in Friday sections (recorded)

This is a mathematically intense course. 😉
But that's why it's exciting and rewarding!

Honor Code

Do's

- form study groups (with arbitrary number of people); discuss and work on homework problems in groups
- write down the solutions independently
- write down the names of people with whom you've discussed the homework

Don'ts

- It is an honor code violation to copy, refer to, or look at written or code solutions from a previous year, including but not limited to: official solutions from a previous year, solutions posted online, solutions you or someone else may have written up in a previous year, and solutions for related problems.
- read the longer description on the course website

Course Project

- We encourage you to form a group of 1-3 people
  - same criterion for 1-3 people
- More information and previous course projects can be found on course website
- List of potential topics

  - Athletics & Sensing Devices
  - Audio & Music
  - Computer Vision
  - Finance & Commerce
  - General Machine Learning
  - Life Sciences
  - Natural Language
  - Physical Sciences
  - Theory
  - Reinforcement Learning

TA Lectures (Optional Attendance)

- Friday TA lectures
  - First few weeks, linear algebra, probability, NumPy review
  - Other weeks are about advanced topics

- Discussion sections
  - interactive sessions and in a small, traditional classroom settings
  - TAs will largely work through problems that are similar to or simpler than the homework, with the goal of making it easier to solve homework questions and midterm.

Other Information on Course Website

cs229.stanford.edu

- Ed:
  - All announcements and questions
    - For logistical questions, please take a look at course FAQ first
  - Finding study groups friends
  - If you enrolled in the class but do not have access to Ed, it should come within one or two days. If it has been more than that, send Sauren (sauren@stanford.edu) an email.
- Recorded videos on canvas
- Course calendar, syllabus, lecture notes, …
- Gradescope, late days policy **(no late days for final project)**, etc.
- Other FAQs in the logistics doc

1. Administrivia

cs229.stanford.edu

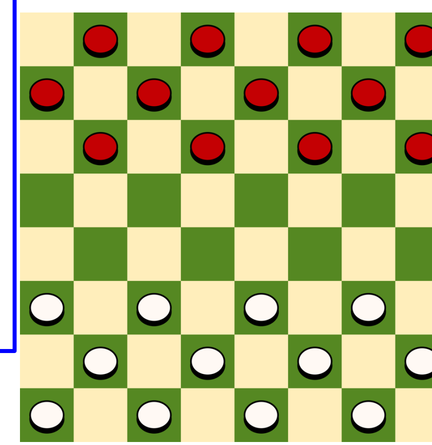
2. Topics Covered in This Course

Definition of Machine Learning

Arthur Samuel (1959): Machine Learning is the field of study that gives the computer the ability to learn without being explicitly programmed.

A. L. Samuel*

Some Studies in Machine Learning Using the Game of Checkers. II—Recent Progress



Photos from Wikipedia

Definition of Machine Learning

Tom Mitchell (1998): a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Experience (data): games played by the program (with itself)

Performance measure: winning rate

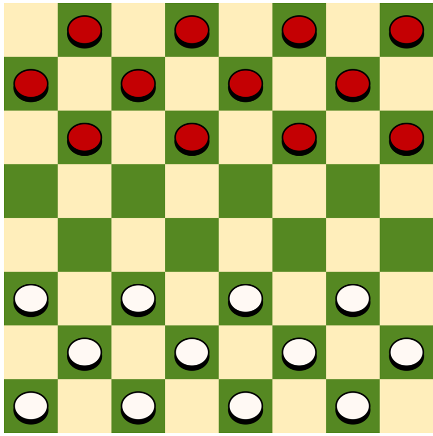


Image from Tom Mitchell's homepage

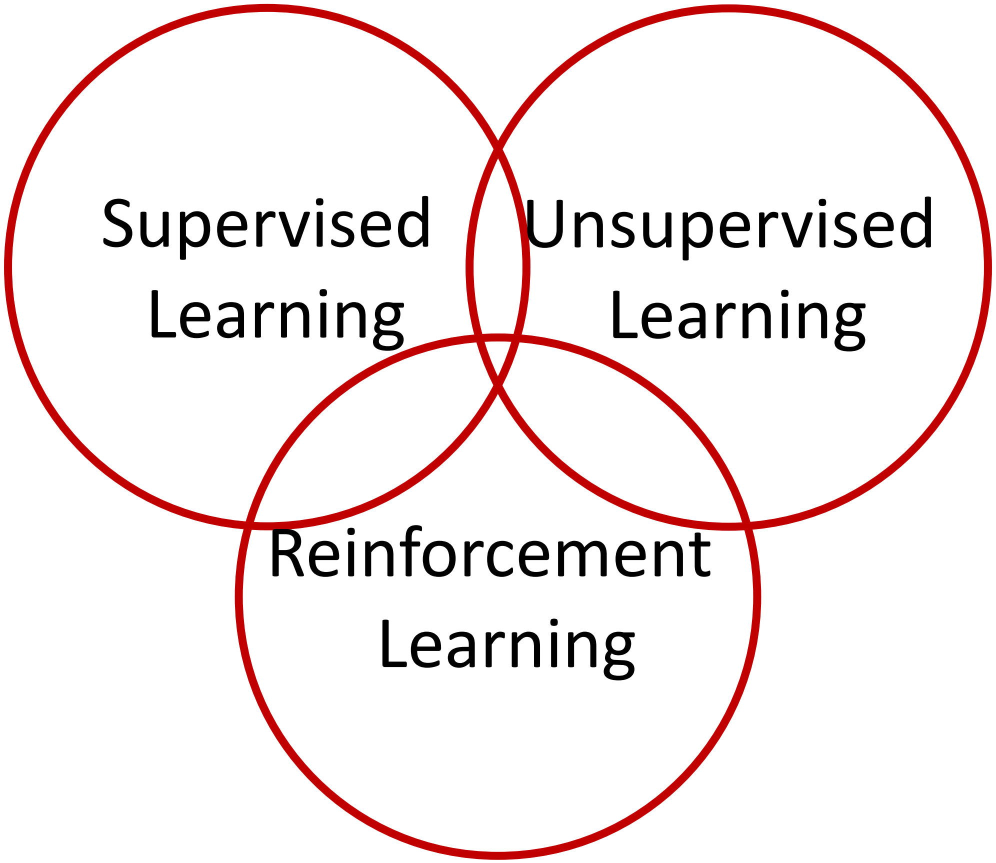
Taxonomy of Machine Learning
(A Simplistic View Based on Tasks)

Supervised Learning

Unsupervised Learning

Reinforcement Learning

Taxonomy of Machine Learning (A Simplistic View Based on Tasks)



can also be viewed as tools/methods

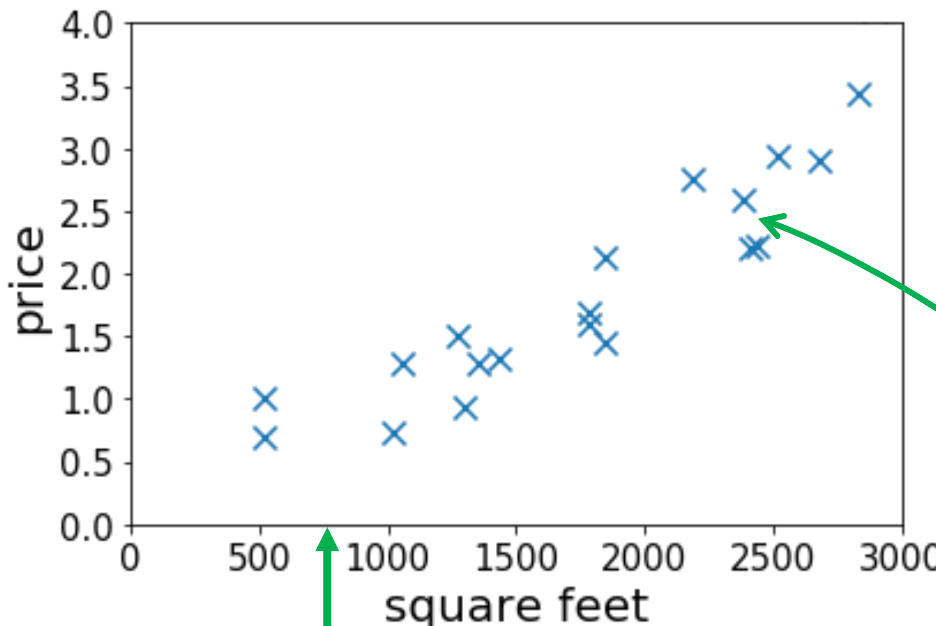
Supervised Learning

Housing Price Prediction

- Given: a dataset that contains n samples

$$\left(x^{(1)}, y^{(1)}\right), \ldots \left(x^{(n)}, y^{(n)}\right)$$

- Task: if a residence has x square feet, predict its price?



15th sample

$(x^{(15)}, y^{(15)})$

$x = 800$

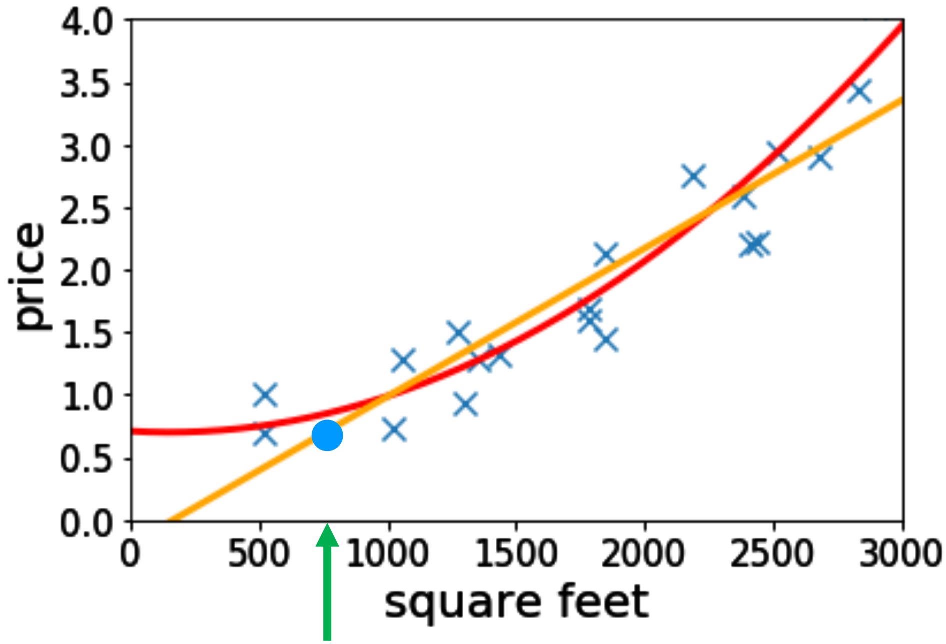
$y = ?$

Housing Price Prediction

- Given: a dataset that contains n samples

$$\left(x^{(1)}, y^{(1)}\right), \ldots \left(x^{(n)}, y^{(n)}\right)$$

- Task: if a residence has x square feet, predict its price?



$x = 800$

$y = ?$

- Lecture 2&3: fitting linear/quadratic functions to the dataset

More Features

- Suppose we also know the lot size
- Task: find a function that maps

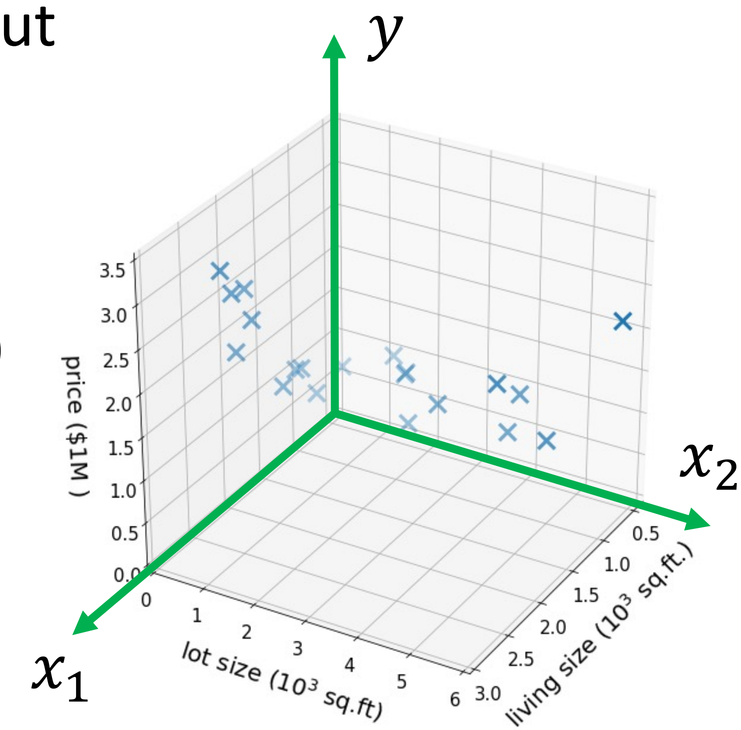
$$\underbrace{\text{(size, lot size)}}_{\text{features/input}} \rightarrow \underbrace{\text{price}}_{\text{label/output}}$$

$$x \in \mathbb{R}^2 \qquad\qquad y \in \mathbb{R}$$

- Dataset: $\left(x^{(1)}, y^{(1)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)$

where $x^{(i)} = \left(x_1^{(i)}, x_2^{(i)}\right)$

- "Supervision" refers to $y^{(1)}, \ldots, y^{(n)}$



High-dimensional Features

- $x \in \mathbb{R}^d$ for large d
- E.g.,

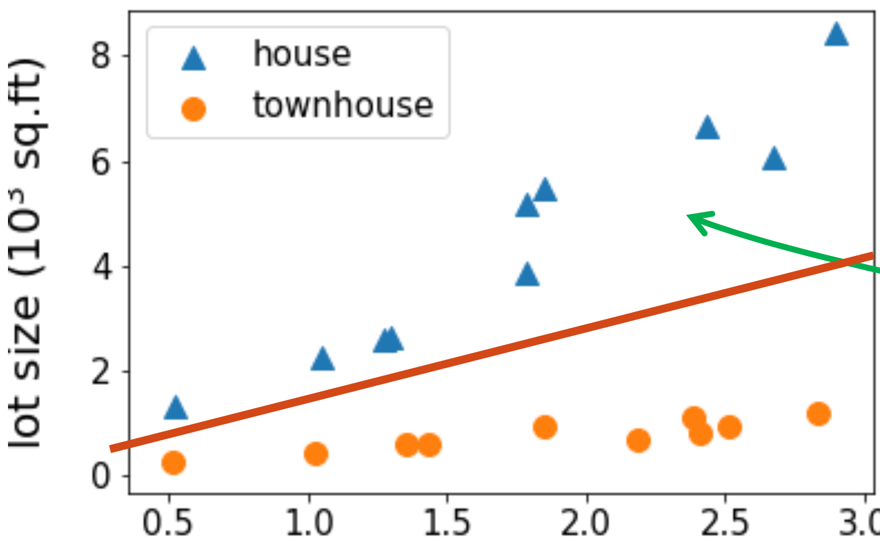
$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ \vdots \\ \vdots \\ x_d \end{bmatrix} \begin{array}{l} \text{--- living size} \\ \text{--- lot size} \\ \text{--- \# floors} \\ \text{--- condition} \\ \text{--- zip code} \\ \quad\vdots \\ \\ \end{array} \longrightarrow \quad y \text{ --- price}$$

- Lecture 6-7: infinite dimensional features
- Lecture 10: select features based on the data

Regression vs Classification

- regression: if $y \in \mathbb{R}$ is a continuous variable
  - e.g., price prediction
- classification: the label is a discrete variable
  - e.g., the task of predicting the types of residence

(size, lot size) $\rightarrow$ house or townhouse?



y = house or townhouse?

Lecture 3&4: classification

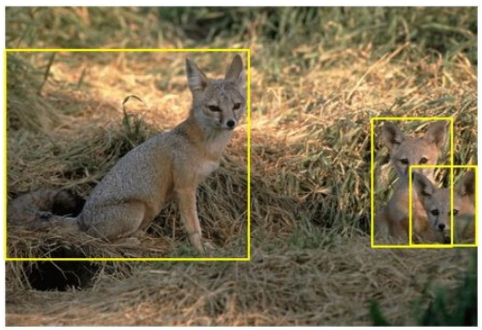
Supervised Learning in Computer Vision

- Image Classification
  - $x =$ raw pixels of the image, $y =$ the main object

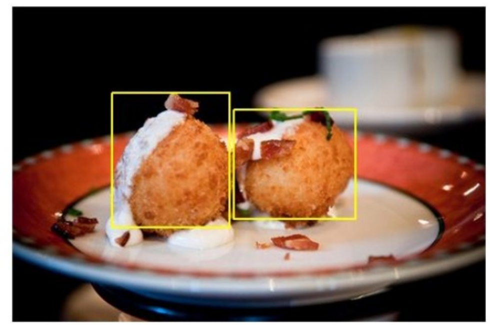
ImageNet Large Scale Visual Recognition Challenge. Russakovsky et al.'2015

Supervised Learning in Computer Vision

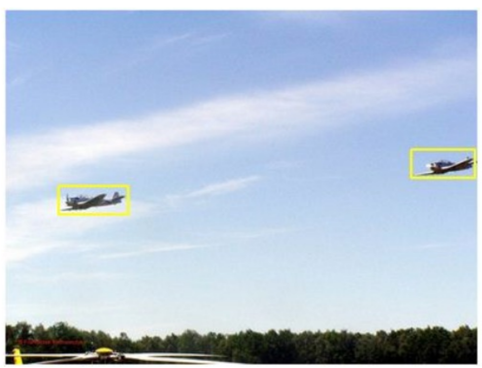
- Object localization and detection
  - $x =$ raw pixels of the image, $y =$ the bounding boxes



kit fox



croquette



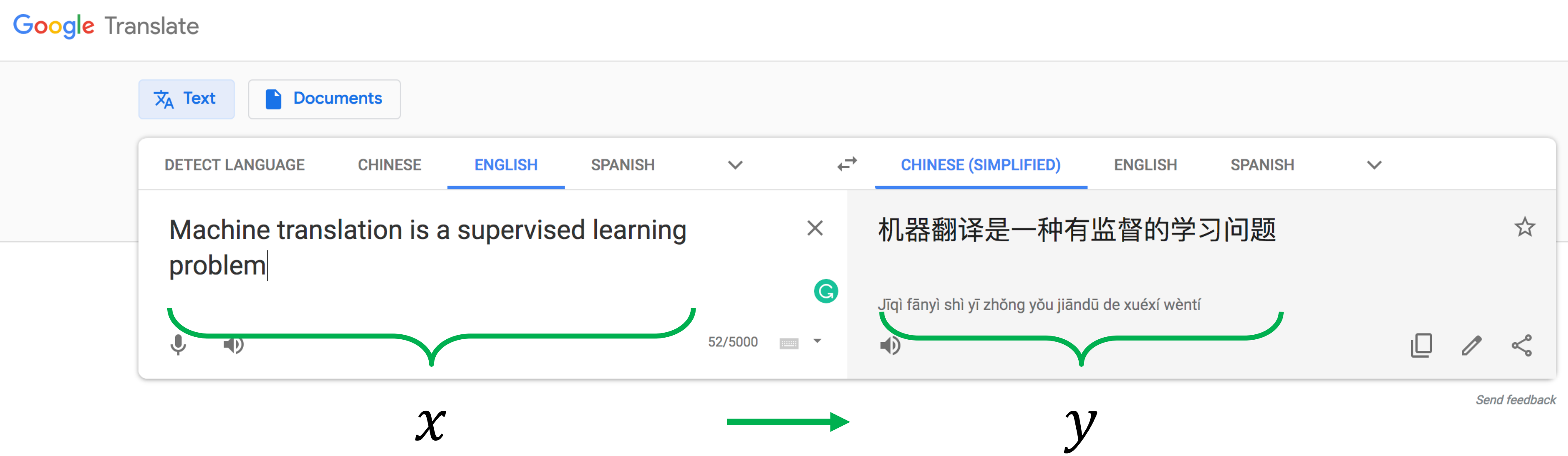
airplane

frog

ImageNet Large Scale Visual Recognition Challenge. Russakovsky et al.'2015

Supervised Learning in Natural Language Processing

- Machine translation



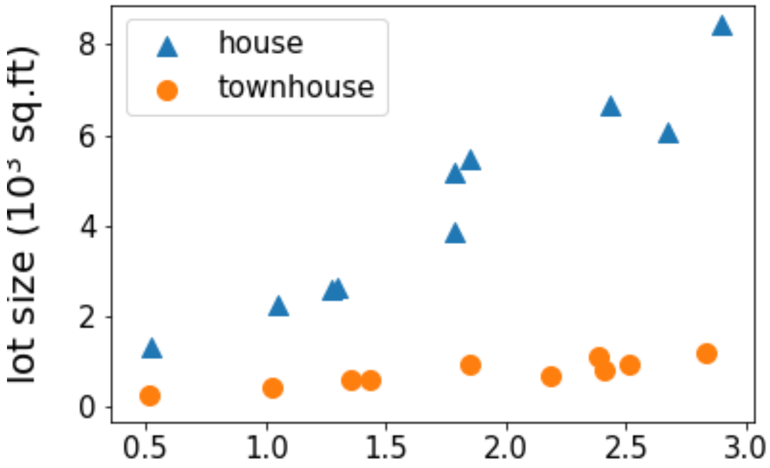
- Note: this course only covers the basic and fundamental techniques of supervised learning
- CS224N and CS231N would be more suitable if you are interested in the particular applications

Unsupervised Learning

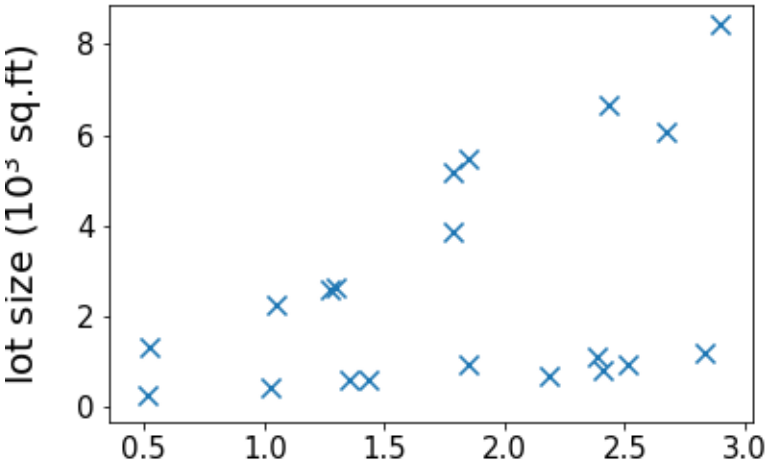
Unsupervised Learning

- Dataset contains no labels: $x^{(1)}, \dots x^{(n)}$
- Goal (vaguely-posed): to find interesting structures in the data

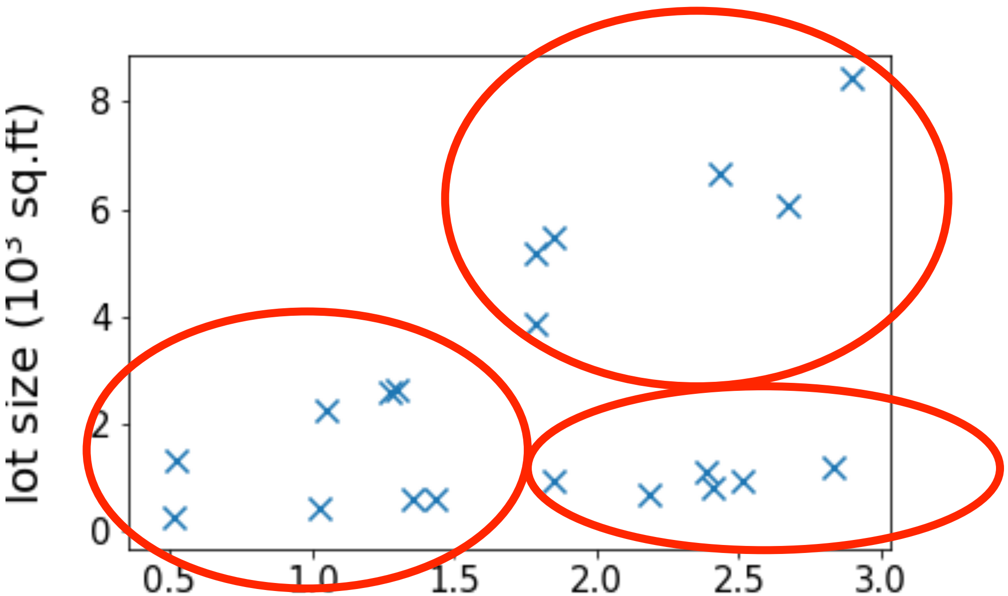
supervised



unsupervised

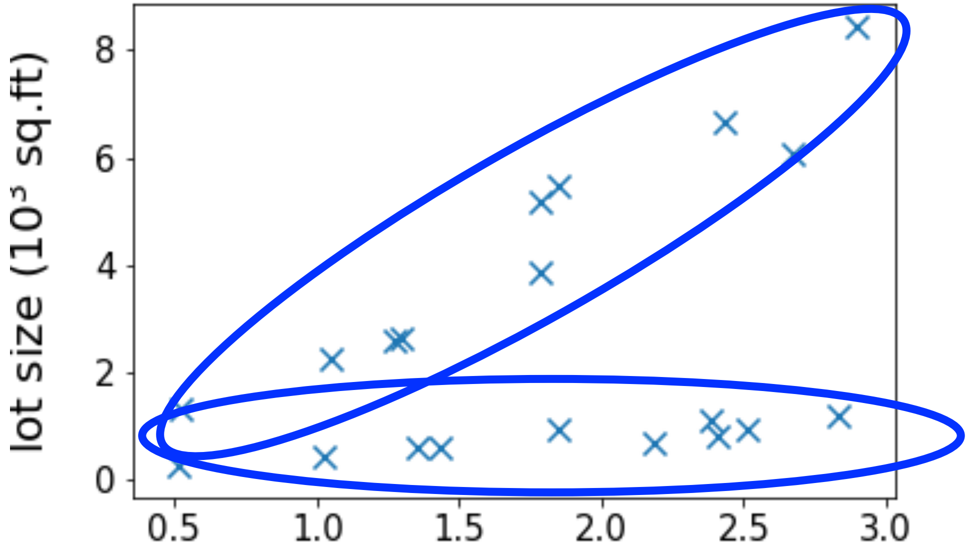


Clustering

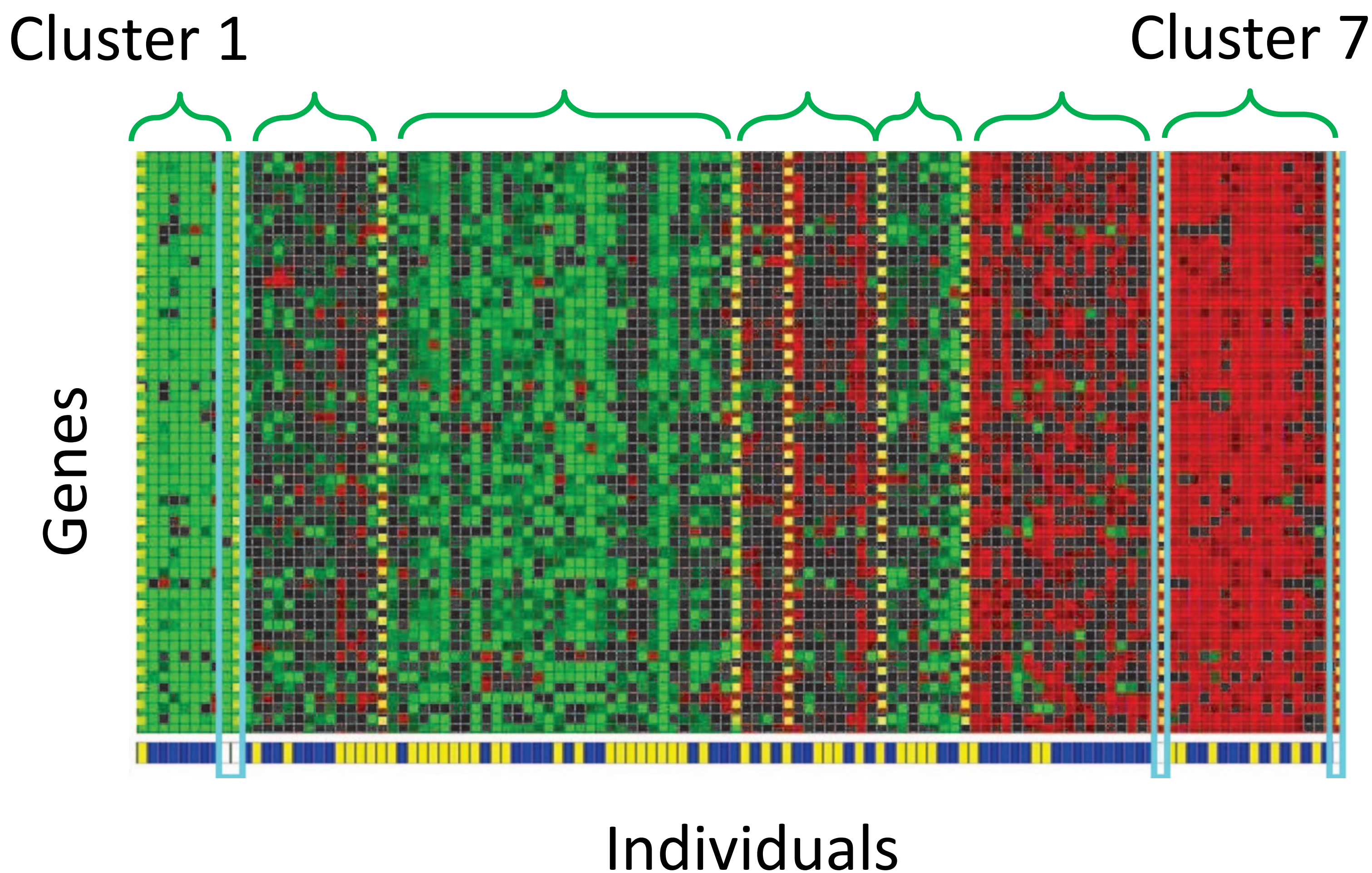


Clustering

- Lecture 12&13: k-mean clustering, mixture of Gaussians

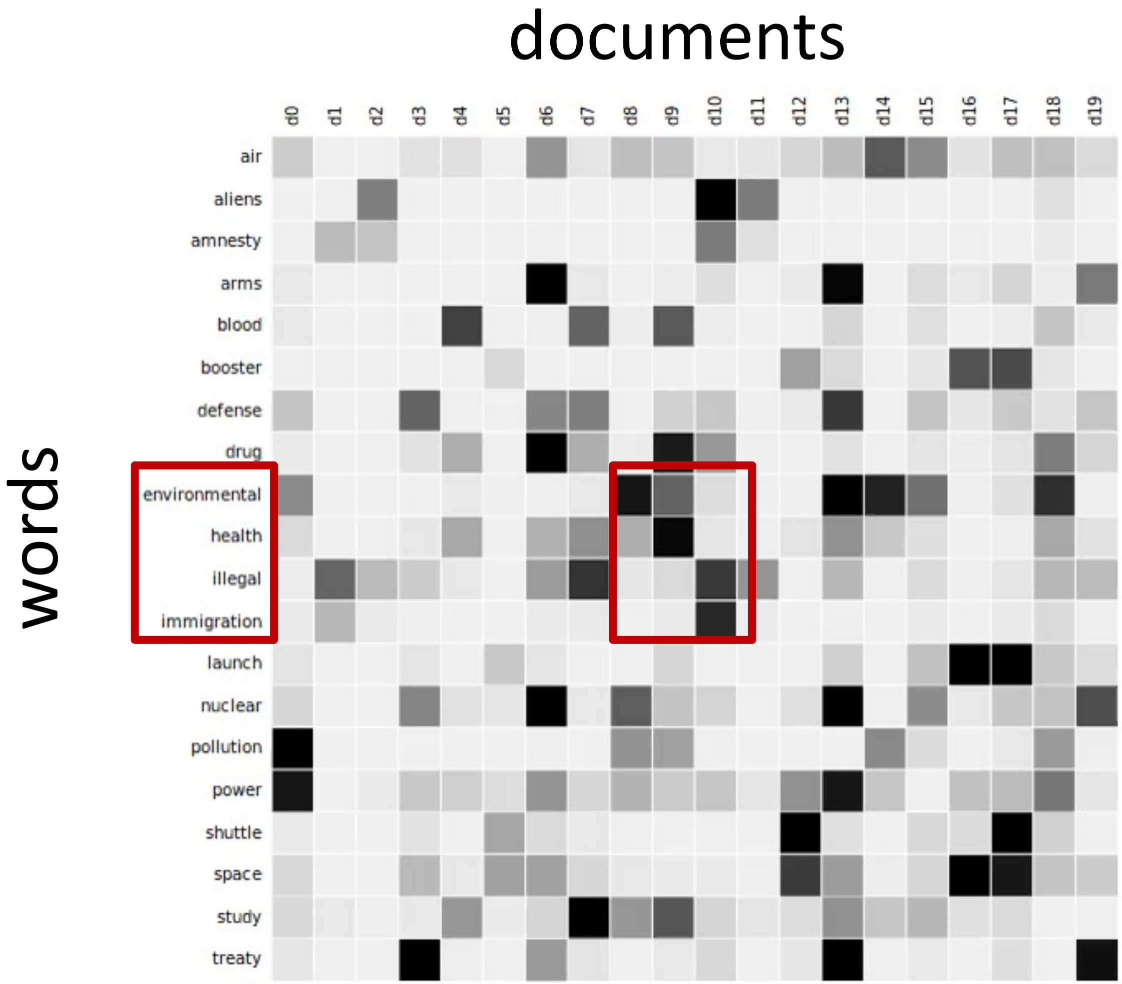


Clustering Genes



Identifying Regulatory Mechanisms using Individual Variation Reveals Key Role for Chromatin Modification. [Su-In Lee, Dana Pe'er, Aimee M. Dudley, George M. Church and Daphne Koller. ’06]

Latent Semantic Analysis (LSA)



- Lecture 14: principal component analysis (tools used in LSA)

Image credit: https://commons.wikimedia.org/wiki/File:Topic_detection_in_a_document-word_matrix.gif

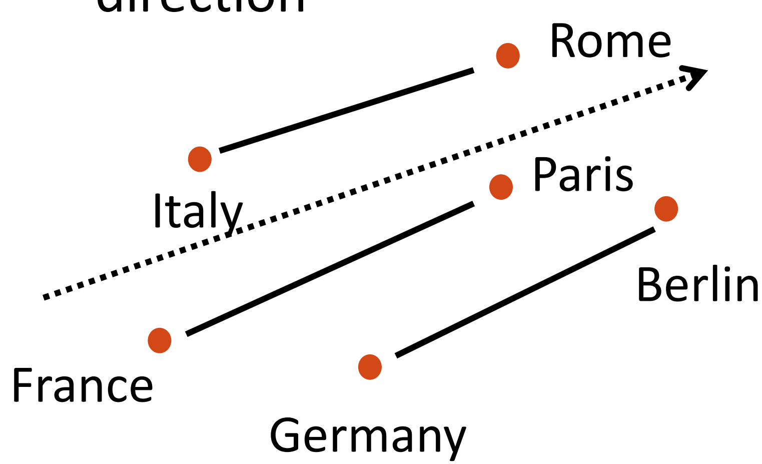
Word Embeddings

WIKIPEDIA
The Free ~~Encyclopedia~~
CORPUS

Unlabeled dataset

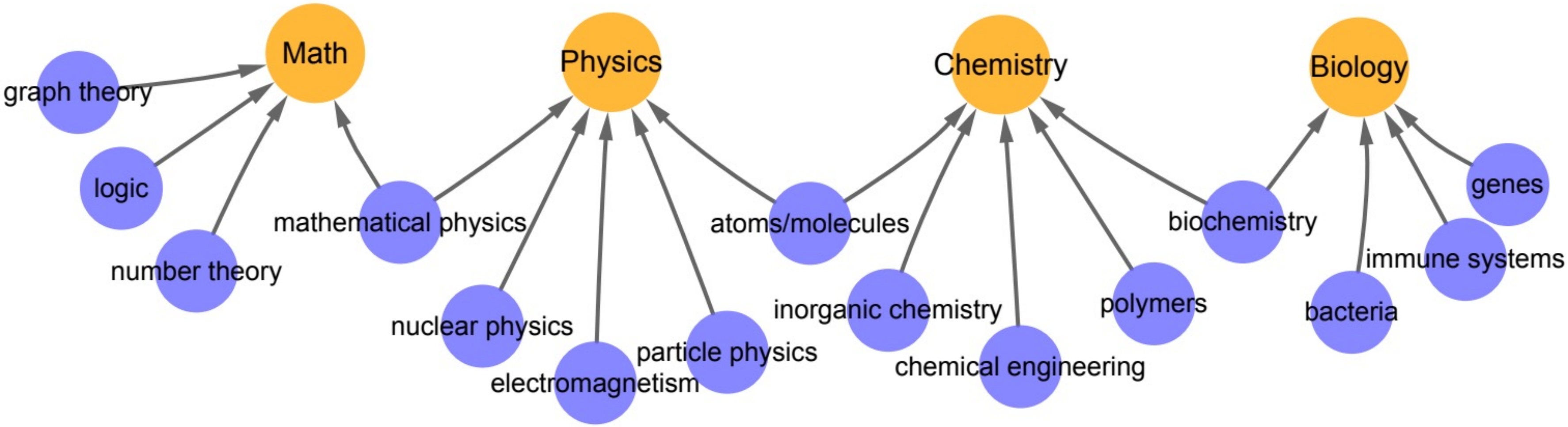
Represent words by vectors

- word $\xrightarrow{\text{encode}}$ vector
- relation $\xrightarrow{\text{encode}}$ direction



Word2vec [Mikolov et al'13]
GloVe [Pennington et al'14]

Clustering Words with Similar Meanings (Hierarchically)



	logic deductive propositional semantics	graph subgraph bipartite vertex	boson massless particle higgs	polyester polypropylene resins epoxy	acids amino biosynthesis peptide
tag	*logic*	*graph theory*	*particle physics*	*polymer*	*biochemistry*

[Arora-Ge-Liang-M.-Risteski, TACL'17,18]

Large Language Models (Lecture 16)

- machine learning models for language learnt on large-scale language datasets
- can be used for many purposes

SYSTEM PROMPT (HUMAN-WRITTEN)	*In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.*
MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)	The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.
	Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.
	Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.
	Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.
	Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

Context →	Helsinki is the capital and largest city of Finland. It is in the region of Uusimaa, in southern Finland, on the shore of the Gulf of Finland. Helsinki has a population of , an urban population of , and a metropolitan population of over 1.4 million, making it the most populous municipality and urban area in Finland. Helsinki is some north of Tallinn, Estonia, east of Stockholm, Sweden, and west of Saint Petersburg, Russia. Helsinki has close historical connections with these three cities. The Helsinki metropolitan area includes the urban core of Helsinki, Espoo, Vantaa, Kauniainen, and surrounding commuter towns. It is the world's northernmost metro area of over one million people, and the city is the northernmost capital of an EU member state. The Helsinki metropolitan area is the third largest metropolitan area in the Nordic countries after Stockholm and Copenhagen, and the City of Helsinki is the third largest after Stockholm and Oslo. Helsinki is Finland's major political, educational, financial, cultural, and research center as well as one of northern Europe's major cities. Approximately 75% of foreign companies that operate in Finland have settled in the Helsinki region. The nearby municipality of Vantaa is the location of Helsinki Airport, with frequent service to various destinations in Europe and Asia. Q: what is the most populous municipality in Finland? A: Helsinki Q: how many people live there? A: 1.4 million in the metropolitan area Q: what percent of the foreign companies that operate in Finland are in Helsinki? A: 75% Q: what towns are a part of the metropolitan area? A:
Target Completion →	Helsinki, Espoo, Vantaa, Kauniainen, and surrounding commuter towns

Context →	Please unscramble the letters into a word, and write that word: taefed =
Target Completion →	defeat

Context →	L'analyse de la distribution de fréquence des stades larvaires d'I. verticalis dans une série d'étangs a également démontré que les larves mâles étaient à des stades plus avancés que les larves femelles. =
Target Completion →	Analysis of instar distributions of larval I. verticalis collected from a series of ponds also indicated that males were in more advanced instars than females.

Context →	Q: What is 95 times 45? A:
Target Completion →	4275

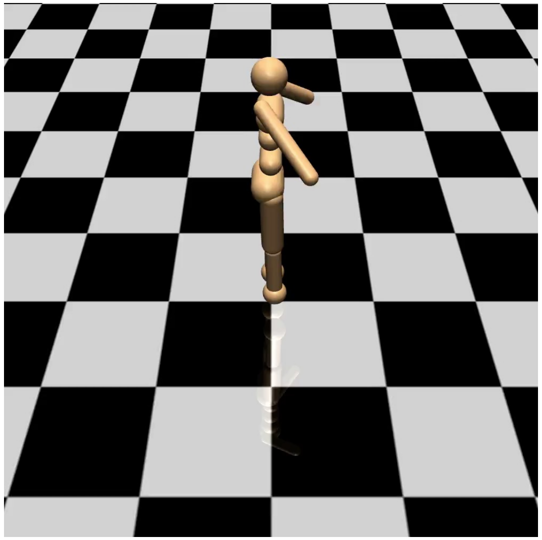
Language Models are Few-Shot Learners [Brown et al.'20]
https://openai.com/blog/better-language-models/

Reinforcement Learning

- Learning to make sequential decisions


AlphaGo
Lee Sedol
ALPHAGO

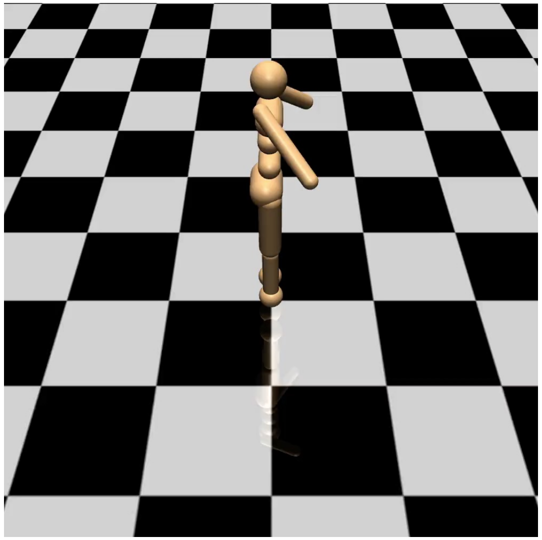
learning to walk to the right



Iteration 10

[Luo-Xu-Li-Tian-Darrell-M.'18]

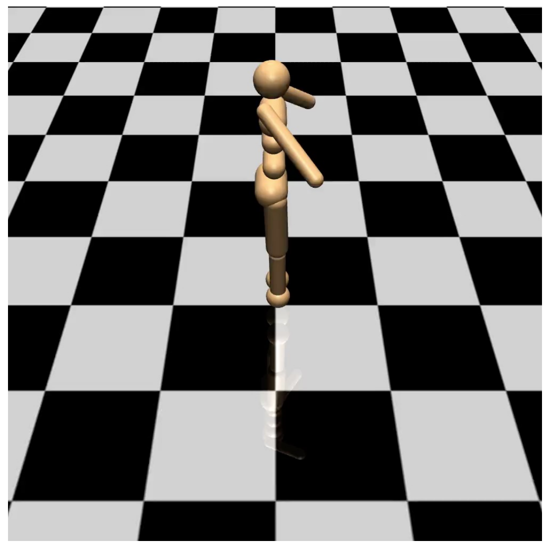
learning to walk to the right



Iteration 20

[Luo-Xu-Li-Tian-Darrell-M.'18]

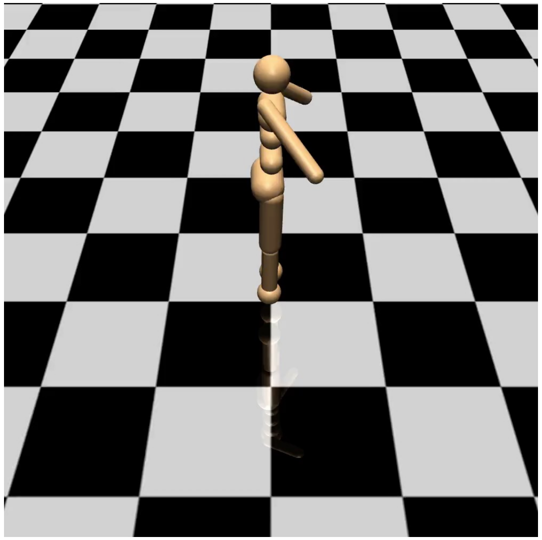
learning to walk to the right



Iteration 80

[Luo-Xu-Li-Tian-Darrell-M.'18]

learning to walk to the right



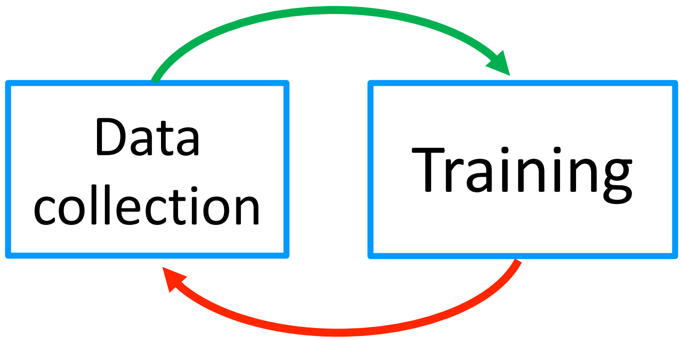
Iteration 210

[Luo-Xu-Li-Tian-Darrell-M.'18]

Reinforcement Learning

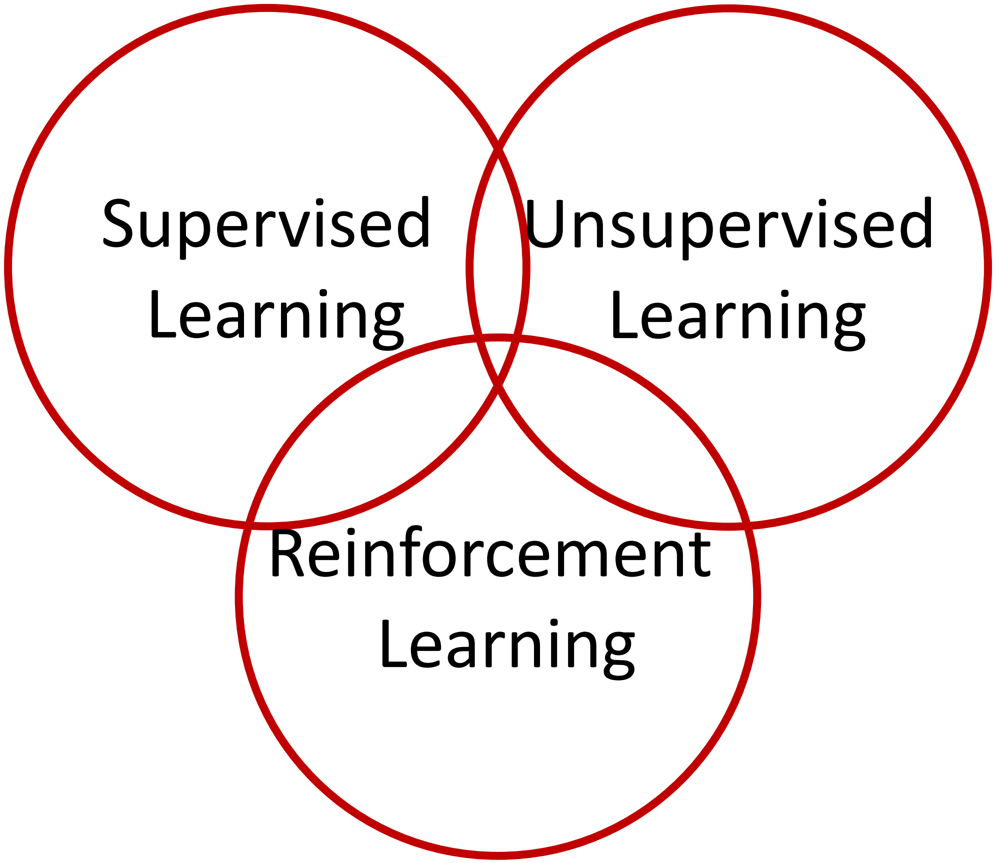
- The algorithm can collect data interactively

Try the strategy and collect feedbacks



Improve the strategy based on the feedbacks

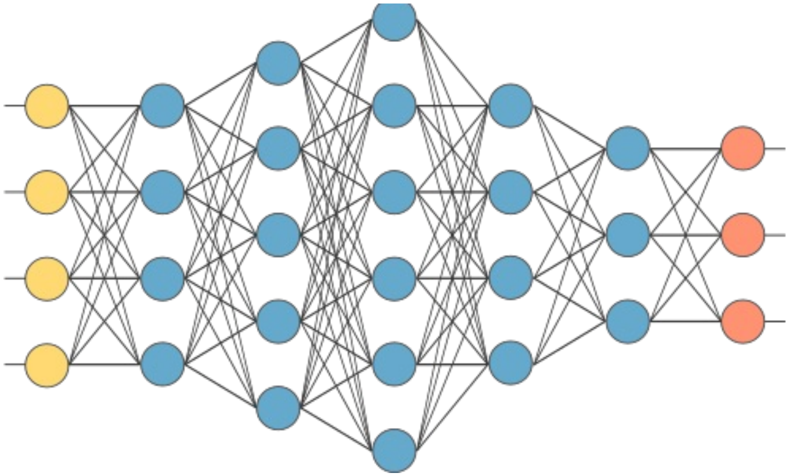
Taxonomy of Machine Learning (A Simplistic View Based on Tasks)



can also be viewed as tools/methods

Other Tools/Topics In This Course

- Deep learning basics



- Introduction to learning theory
  - Bias variance tradeoff
  - Feature selection
  - ML advice
- Broader aspects of ML
  - Robustness/fairness

Questions?

Thank you!