Machine Learning: Introduction

CS60050 (Sec 2)

July 24, 2024

Slides from Matt Gormley (CMU), Dr. Yaser S. Abu-Mostafa (CalTech, USA)

Somak Aditya
Assistant Professor
Department of CSE, IIT KGP

Sudeshna Sarkar
Professor
Department of CSE, IIT KGP
Centre of Excellence in AI, IIT KGP

Course Website

- https://machine-learning-2024-cs60050.netlify.app/

- Course Timings
  - Wed 11-12 PM, Thu 12-1 PM, Fri 8-9 AM
  - Section 2 (Dr. Somak Aditya, NC442), Section 1 (Dr. Sudeshna Sarkar, NC243)

- Office: CS 305 (Somak Aditya), CS 201 (Sudeshna Sarkar)

- Teaching Assistants
  - Manoranjan Behera, Ayan Maity, Utsav Bandyopadhyay Moulik, Saurabh Roy,Gadde Gayathri, Pritam Sarkar, Jadi Ajay, Mishra Saurabh Rajesh Mridulata

Books and Materials

- Tom Mitchell; "Machine Learning"; First Edition, McGraw Hill, 1997.
- Richard O. Duda, Peter E. Hart, David G. Stork; "Pattern Classification"; Second Edition, John Wiley & Sons, November 2000.
- Christopher Bishop; "Pattern Recognition and Machine Learning"; First Edition, Springer-Verlag New York, 2006.
- Ethem Alpaydin; "Introduction to Machine Learning"; Third Edition, The MIT Press, September 2014.
- Trevor Hastie, Robert Tibshirani, Jerome Friedman; "The Elements of Statistical Learning"; Second Edition, Springer, 2001.

Course Evaluation Plan (Tentative)

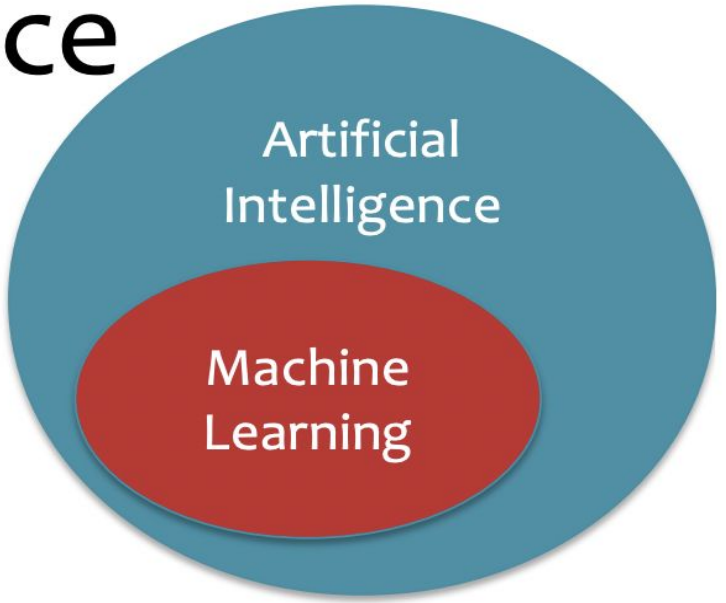
- Mid Term - 25%
- Final Exam - 40%
- Assignments (2) - 20%
- Class-Tests (Two) - 15%

Attendance Policy: Attendance record will be maintained and uploaded on the website.

- Drastic changes will be noted. Deregistering in the worst case.

What is Machine Learning?

Artificial Intelligence



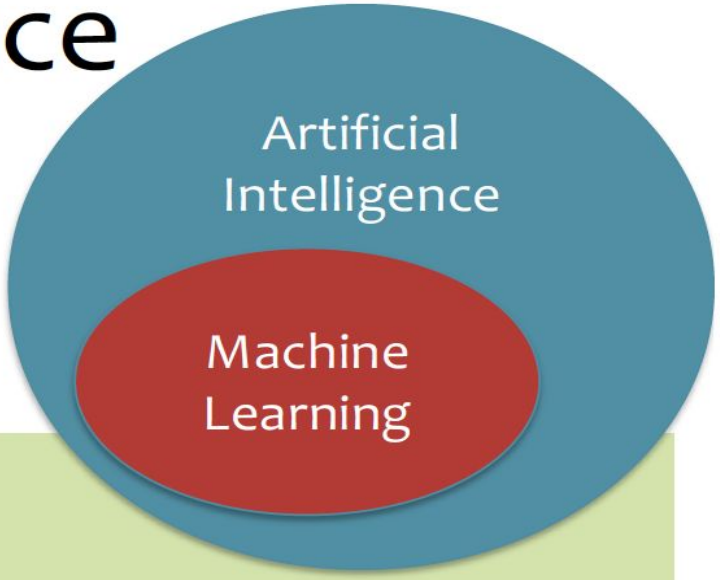
The basic goal of AI is to develop intelligent machines.

This consists of many sub-goals:

- Perception
- Reasoning
- Control / Motion / Manipulation
- Planning
- Communication
- Creativity
- Learning

Slides from Dr. Matt Gormley (CMU)

Artificial Intelligence



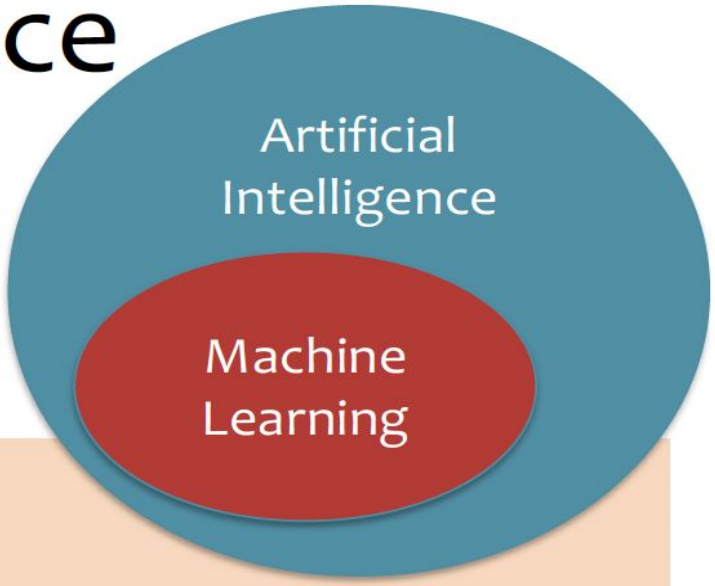
The basic goal of AI is to develop intelligent machines.

This consists of many sub-goals:

- Perception
- Reasoning
- Control / Motion / Manipulation
- Planning
- Communication
- Creativity
- Learning

Slides from Dr. Matt Gormley (CMU)

Artificial Intelligence



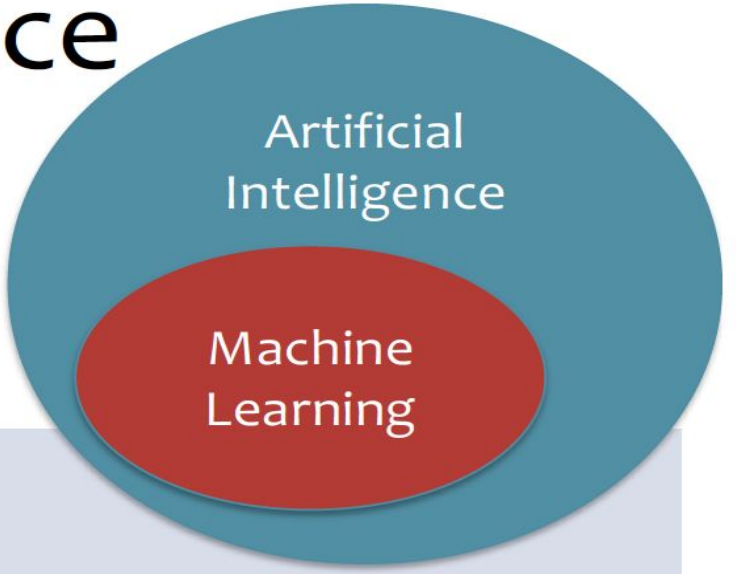
The basic goal of AI is to develop intelligent machines.

This consists of many sub-goals:

- Perception
- Reasoning
- Control / Motion / Manipulation
- Planning
- Communication
- Creativity
- Learning

Slides from Dr. Matt Gormley (CMU)

Artificial Intelligence



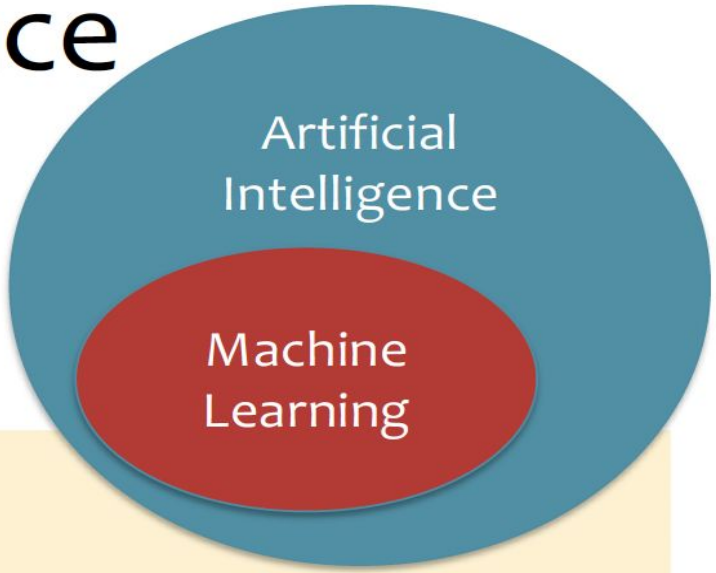
The basic goal of AI is to develop intelligent machines.

This consists of many sub-goals:

- Perception
- Reasoning
- Control / Motion / Manipulation
- Planning
- Communication
- Creativity
- Learning

Slides from Dr. Matt Gormley (CMU)

Artificial Intelligence



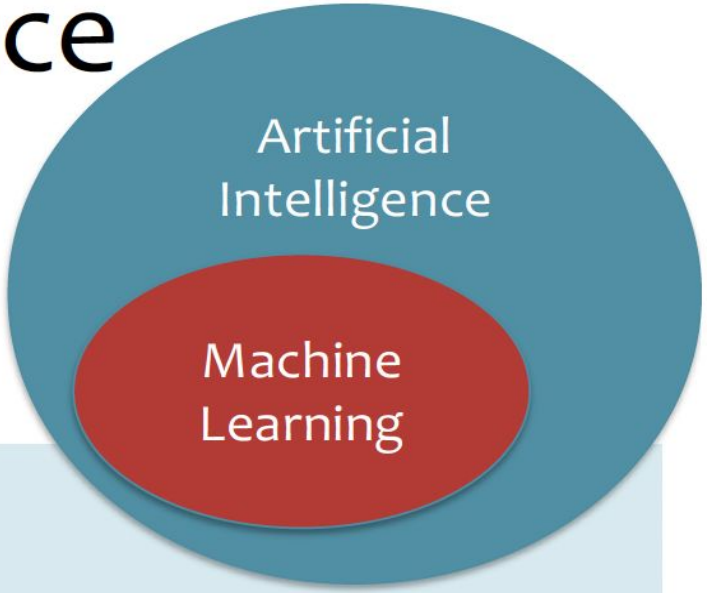
The basic goal of AI is to develop intelligent machines.

This consists of many sub-goals:

- Perception
- Reasoning
- Control / Motion / Manipulation
- Planning
- Communication
- Creativity
- Learning

Slides from Dr. Matt Gormley (CMU)

Artificial Intelligence



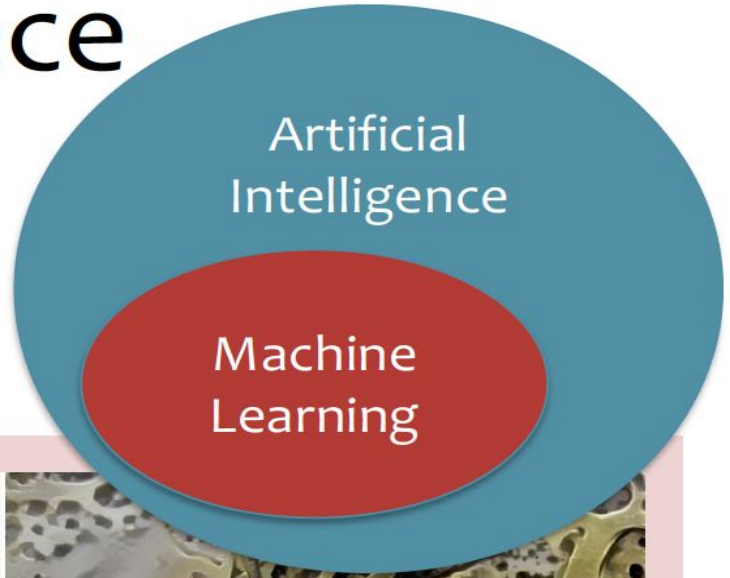
The basic goal of AI is to develop intelligent machines.

This consists of many sub-goals:

- Perception
- Reasoning
- Control / Motion / Manipulation
- Planning
- Communication
- Creativity
- Learning

Slides from Dr. Matt Gormley (CMU)

Artificial Intelligence



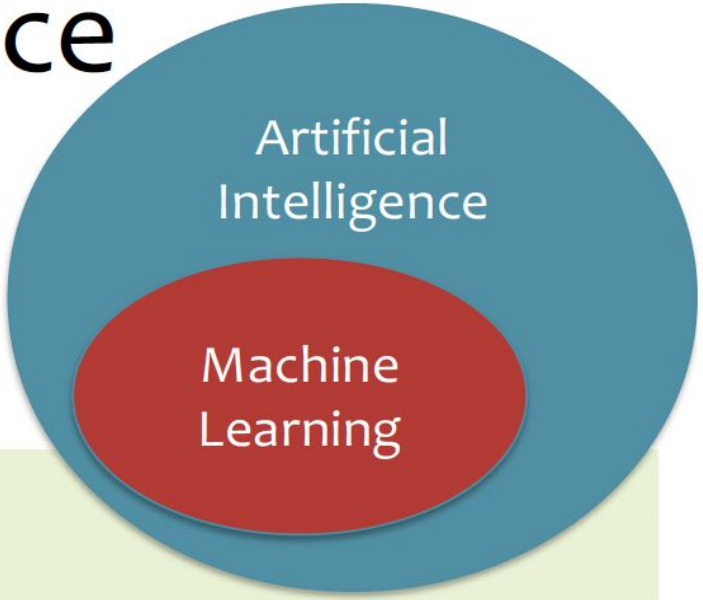
The basic goal of AI is to develop intelligent machines.

This consists of many sub-goals:

- Perception
- Reasoning
- Control / Motion / Manipulation
- Planning
- Communication
- Creativity
- Learning

Slides from Dr. Matt Gormley (CMU)

Artificial Intelligence



The basic goal of AI is to develop intelligent machines.

This consists of many sub-goals:

- Perception
- Reasoning
- Control / Motion / Manipulation
- Planning
- Communication
- Creativity
- Learning

Slides from Dr. Matt Gormley (CMU)

What is Machine Learning?

The goal of this course is to provide you with a toolbox:

Slides from Dr. Matt Gormley (CMU)

What is ML ?

Speech Recognition

1. Learning to recognize spoken words

THEN	NOW
"…the SPHINX system (e.g. Lee 1989) learns speaker-specific strategies for recognizing the primitive sounds (phonemes) and words from the observed speech signal…neural network methods…hidden Markov models…" (Mitchell, 1997)	

Figure from https://botpenguin.com/alexa-vs-siri-vs-google-assistant/

Robotics

2. Learning to drive an autonomous vehicle

THEN	NOW
"…the ALVINN system (Pomerleau 1989) has used its learned strategies to drive unassisted at 70 miles per hour for 90 miles on public highways among other cars…" (Mitchell, 1997)	 waymo.com

Games / Reasoning

3. Learning to beat the masters at board games

THEN	NOW
"…the world's top computer program for backgammon, TD-GAMMON (Tesauro, 1992, 1995), learned its strategy by playing over one million practice games against itself…" (Mitchell, 1997)	

Computer Vision

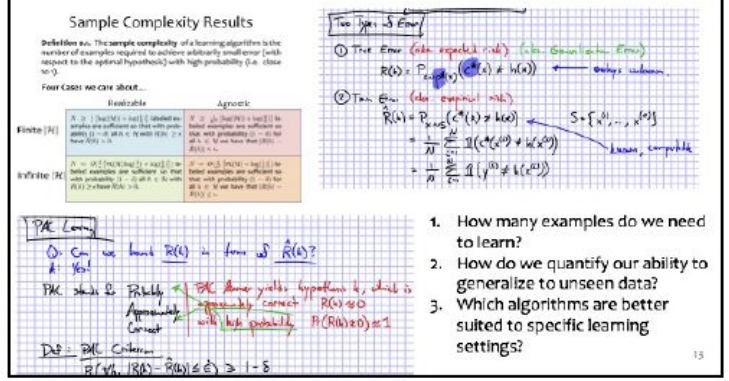
4. Learning to recognize images

THEN	NOW
"…The recognizer is a convolution network that can be spatially replicated. From the network output, a hidden Markov model produces word scores. The entire system is globally trained to minimize word-level errors…." (LeCun et al., 1995)	

Figure from https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Learning Theory

- 5. In what cases and how well can we learn?



1. How many examples do we need to learn?
2. How do we quantify our ability to generalize to unseen data?
3. Which algorithms are better suited to specific learning settings?

Slides from Dr. Matt Gormley (CMU)

More than Stats/Optimization

What ethical responsibilities do we have as machine learning experts?

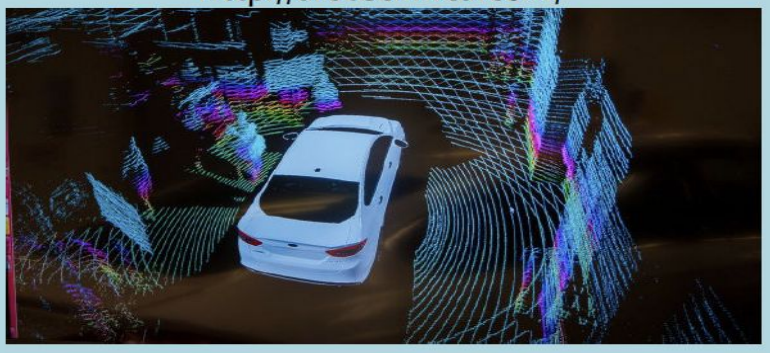
Question: What are the possible societal impacts of machine learning for each case below?

Answer:

1) Search results for news are optimized for ad revenue.

http://bing.com/

http://arstechnica.com/



2) An autonomous vehicle is permitted to drive unassisted on the road.

3) A doctor is prompted by an intelligent system with a plausible diagnosis for her patient.

https://flic.kr/p/HNUUzV

Slides from Dr. Matt Gormley (CMU)

Slides from Dr. Matt Gormley (CMU)

Societal Impacts of ML

The Washington Post
Democracy Dies in Darkness
Subscribe

A 72-year-old congressman goes back to school, pursuing a degree in AI

By Meagan Flynn

December 28, 2022 at 6:00 a.m. EST

Rep. Don Beyer (D-Va.) is pursuing a master's degree in machine learning at George Mason University with hopes of one day applying his AI knowledge to his legislative work. (Craig Hudson for The Washington Post)

Normally Don Beyer doesn't bring his multivariable calculus textbook to work, but his final exam was coming up that weekend.

"And I'm running out of time," he said, plopping the textbook and a scribbled notebook filled with esoteric-looking calculations on a coffee table in his office, "because I have all these—"

His phone was ringing. "I'll be there," Beyer told a colleague wondering when he would be returning to the House floor for votes.

It seemed study time would have to wait.

That's been the story of the year for Beyer (D-Va.), who has been moonlighting as a student at George Mason University in pursuit of a master's degree in machine learning while balancing his duties as a congressman. Beyer — a science wonk, economist and former car salesman — has been taking one class per semester in a slow but steady march toward the degree, with hopes of one day applying his artificial-intelligence knowledge to his legislative work as the technology evolves further.

Figure from https://www.washingtonpost.com/dc-md-va/2022/12/28/beyer-student-artificial-intelligence-degree/

ML Big Picture

Learning Paradigms:

What data is available and when? What form of prediction?

- supervised learning
- unsupervised learning
- semi-supervised learning
- reinforcement learning
- active learning
- imitation learning
- domain adaptation
- online learning
- density estimation
- recommender systems
- feature learning
- manifold learning
- dimensionality reduction
- ensemble learning
- distant supervision
- hyperparameter optimization

Problem Formulation:

What is the structure of our output prediction?

boolean	Binary Classification
categorical	Multiclass Classification
ordinal	Ordinal Classification
real	Regression
ordering	Ranking
multiple discrete	Structured Prediction
multiple continuous	(e.g. dynamical systems)
both discrete & cont.	(e.g. mixed graphical models)

Application Areas

Key challenges?

NLP, Speech, Computer Vision, Robotics, Medicine, Search

Theoretical Foundations:

What principles guide learning?

- ❑ probabilistic
- ❑ information theoretic
- ❑ evolutionary search
- ❑ ML as optimization

Facets of Building ML Systems:

How to build systems that are robust, efficient, adaptive, effective?

1. *Data prep*
2. *Model selection*
3. *Training (optimization / search)*
4. *Hyperparameter tuning on validation data*
5. *(Blind) Assessment on test data*

Big Ideas in ML:

Which are the ideas driving development of the field?

- *inductive bias*
- *generalization / overfitting*
- *bias-variance decomposition*
- *generative vs. discriminative*
- *deep nets, graphical models*
- *PAC learning*
- *distant rewards*

Slides from Dr. Matt Gormley (CMU)

Syllabus (Tentaive)

- Foundations
  - Probability, MLE, MAP, Optimization
- Regression
  - Linear Regression
- Classifiers
  - Logistic Regression, Decision Tree, SVM
- Bayesian Learning
  - Bayesian Networks, Undirected Graphical Models, Learning and Inference
- Important Concepts
  - Bias, Variance
  - Kernels, Regularization and Overfitting
  - Experimental Design
- Unsupervised Learning
  - K-means, PCA, EM / GMMs
- Neural Networks
  - Feedforward Neural Nets, Basic architectures, Backpropagation, CNNs, LSTMs
- Learning Theory
  - Statistical Estimation
  - PAC Learning
- Other Learning Paradigms
  - Matrix Factorization,
  - Reinforcement Learning,

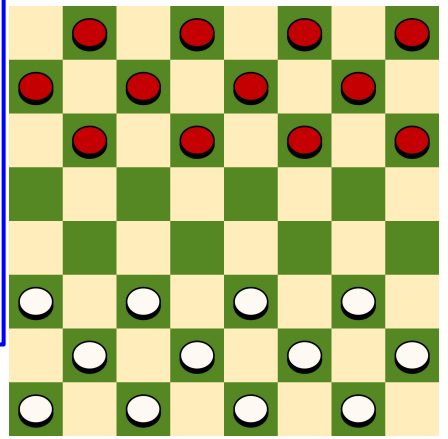
Defining Learning Problems

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

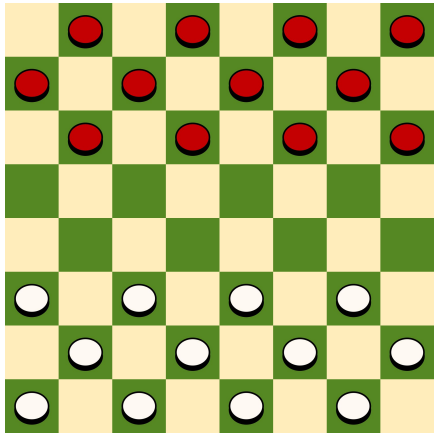


Image from Tom Mitchell's homepage

Well-Posed Learning Problems

Three components *<T,P,E>:*

- *1. Task, T*
- *2. Performance measure, P*
- *3. Experience, E*

Definition of learning*:*

A computer program **learns** if its performance at task T, as measured by P, improves with experience E.

Slides from Dr. Matt Gormley (CMU)

Example Learning Problem(s)

Learning to respond to voice commands (Siri)

- *1. Task, T:* 
- *2. Performance measure, P:* 
- *3. Experience, E:* 

Slides from Dr. Matt Gormley (CMU)

Capturing the Knowledge of Experts

Solution #1: Expert Systems

- Over 20 years ago, we had rule-based systems:
  1. Put a bunch of linguists in a room
  2. Have them think about the structure of their native language and write down the rules they devise

Give me directions to Starbucks
`If: "give me directions to X"` `Then: directions(here, nearest(X))`

How do I get to Starbucks?
`If: "how do i get to X"` `Then: directions(here, nearest(X))`

Where is the nearest Starbucks?
`If: "where is the nearest X"` `Then: directions(here, nearest(X))`

32

Slides from Dr. Matt Gormley (CMU)

Capturing the Knowledge of Experts

Solution #2: Annotate Data and Learn

- Experts:
  - **Very good at** answering questions about specific cases
  - **Not very good at** telling **HOW** they do it
- 1990s: So why not just have them tell you what they do on **SPECIFIC CASES** and then let **MACHINE LEARNING** tell you how to come to the same decisions that they did

Slides from Dr. Matt Gormley (CMU)

Capturing the Knowledge of Experts

Solution #2: Annotate Data and Learn

1. Collect raw sentences $\{x^{(1)}, \ldots, x^{(n)}\}$
2. Experts annotate their meaning $\{y^{(1)}, \ldots, y^{(n)}\}$

$x^{(1)}$: How do I get to Starbucks?	$x^{(3)}$: Send a text to John that I'll be late
$y^{(1)}$: `directions(here, nearest(Starbucks))`	$y^{(3)}$: `txtmsg(John, I'll be late)`
$x^{(2)}$: Show me the closest Starbucks	$x^{(4)}$: Set an alarm for seven in the morning
$y^{(2)}$: `map(nearest(Starbucks))`	$y^{(4)}$: `setalarm(7:00AM)`

Slides from Dr. Matt Gormley (CMU)

Example Learning Problems

Learning to respond to voice commands (Siri)

1. Task, T: predicting action from speech

2. Performance measure, P: percent of correct actions taken in user pilot study

3. Experience, E: examples of (speech, action) pairs

Slides from Dr. Matt Gormley (CMU)

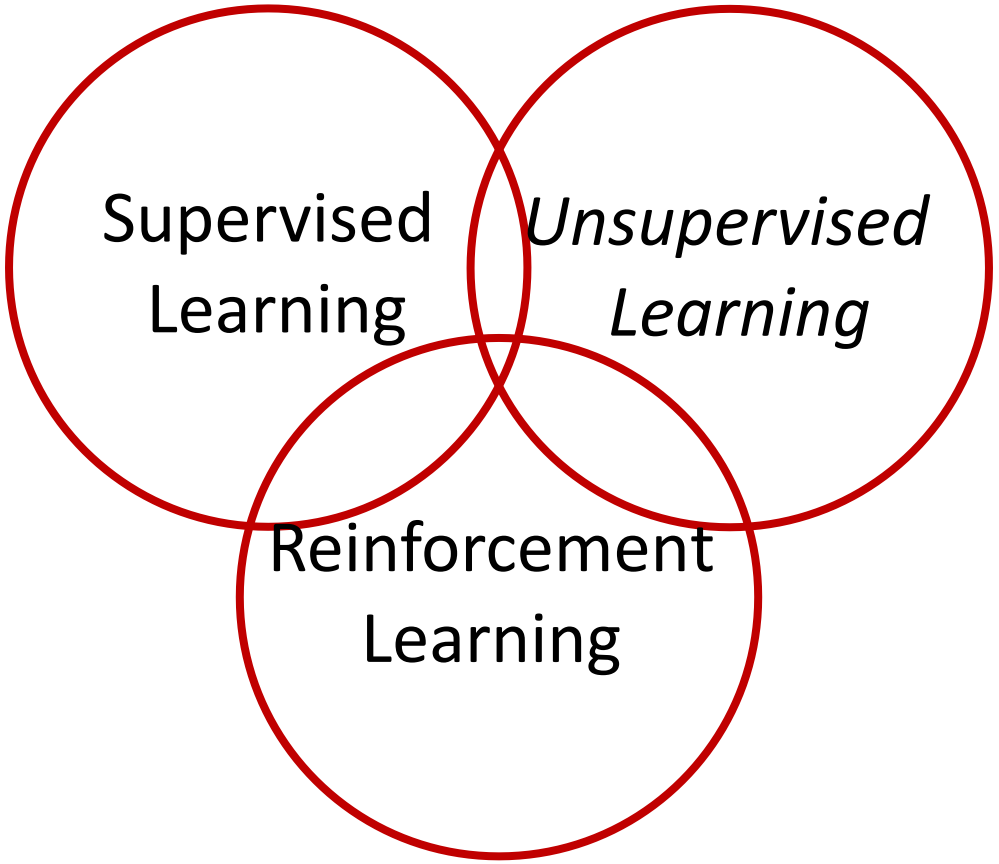
Example: Predicting how a viewer will rate a movie

10% improvement = 1 million dollar prize

- *The essence of machine learning:*
  - *A pattern exists.*
  - *We cannot pin it down mathematically.*
  - *We have data on it.*

Slides from Dr. Yaser S. Abu-Mostafa (CalTech, USA)

Taxonomy of Machine Learning

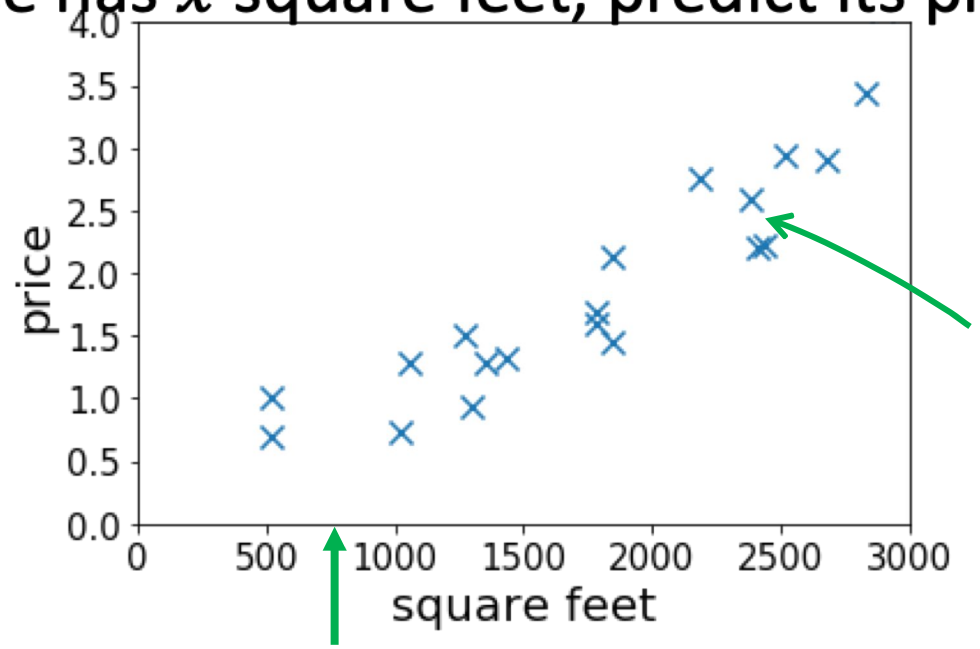


can also be viewed as tools/methods

Supervised Learning

Housing Price Prediction

- Given: a dataset that contains n samples
  - $(x^{(1)}, y^{(1)}), \ldots (x^{(n)}, y^{(n)})$
- Task: if a residence has x square feet, predict its price?



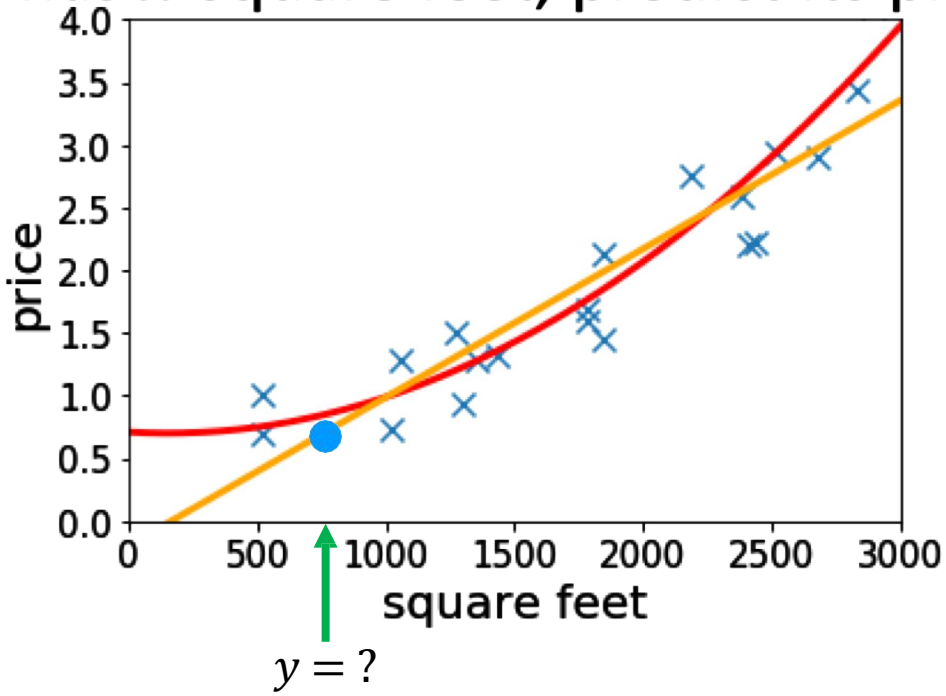
15th sample
$(x^{(0)}, y^{(0)})$

$x = 800$

$y = ?$

Housing Price Prediction

- Given: a dataset that contains n samples
  - $(x^{(1)}, y^{(1)}), \ldots (x^{(n)}, y^{(n)})$
- Task: if a residence has x square feet, predict its price?



- Fitting linear/quadratic functions to the dataset

More Features

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

(size, lot size) → price

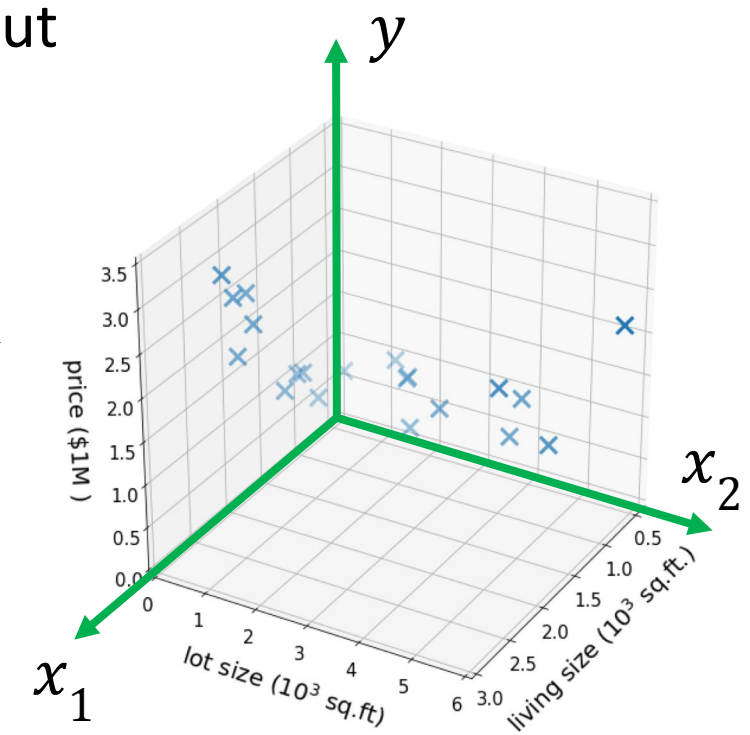
features/input
$x \in \mathbb{R}^2$

label/output
$y \in \mathbb{R}$

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

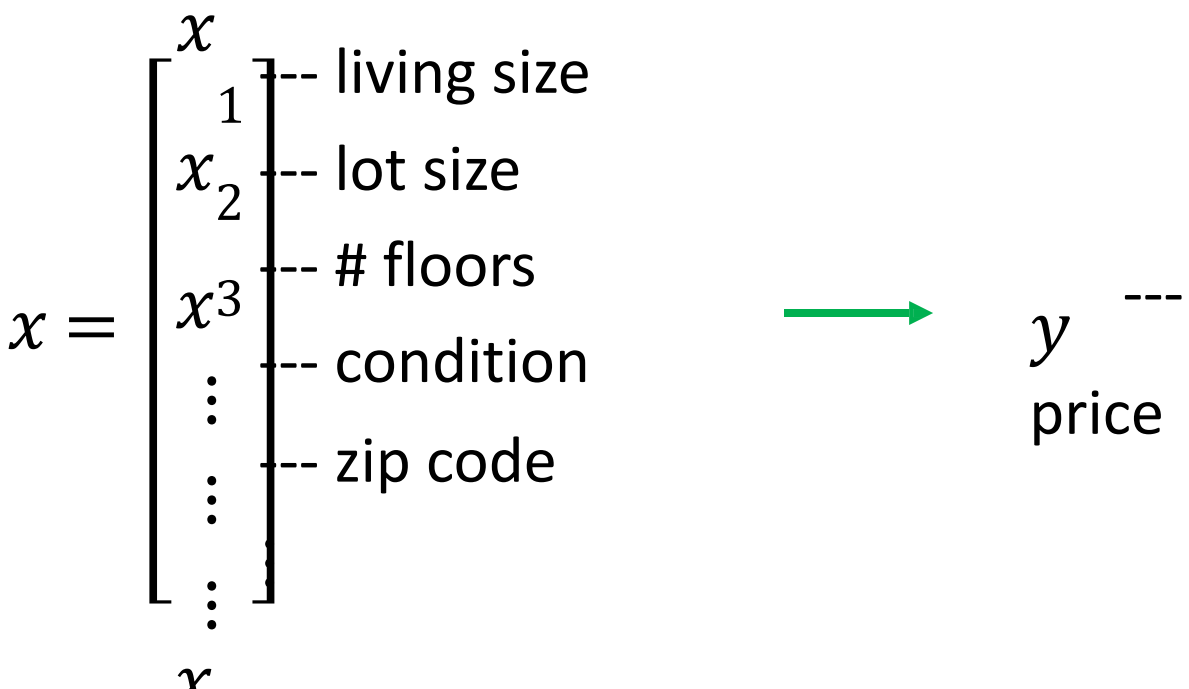
where $x^{(i)} = (x_1^{(i)}, x_2^{(i)})$

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



High-dimensional Features

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

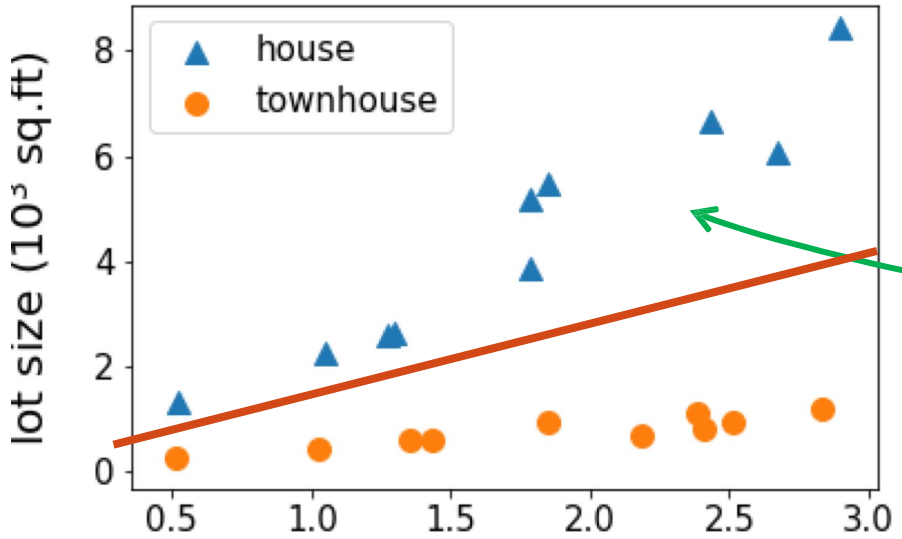


- SVM Kernels: infinite dimensional features
- Feature Selection: select features based on the data

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?

Components of Learning

- *Metaphor: Credit approval*
- *Applicant information:*

Age	23 years
Gender	male
Annual salary	$30000
Years in residence	1 year
Years in job	1 year
Current debt	$15000
…	…

- Approve Credit?

Slides from Dr. Yaser S. Abu-Mostafa (CalTech, USA)

Components of Learning

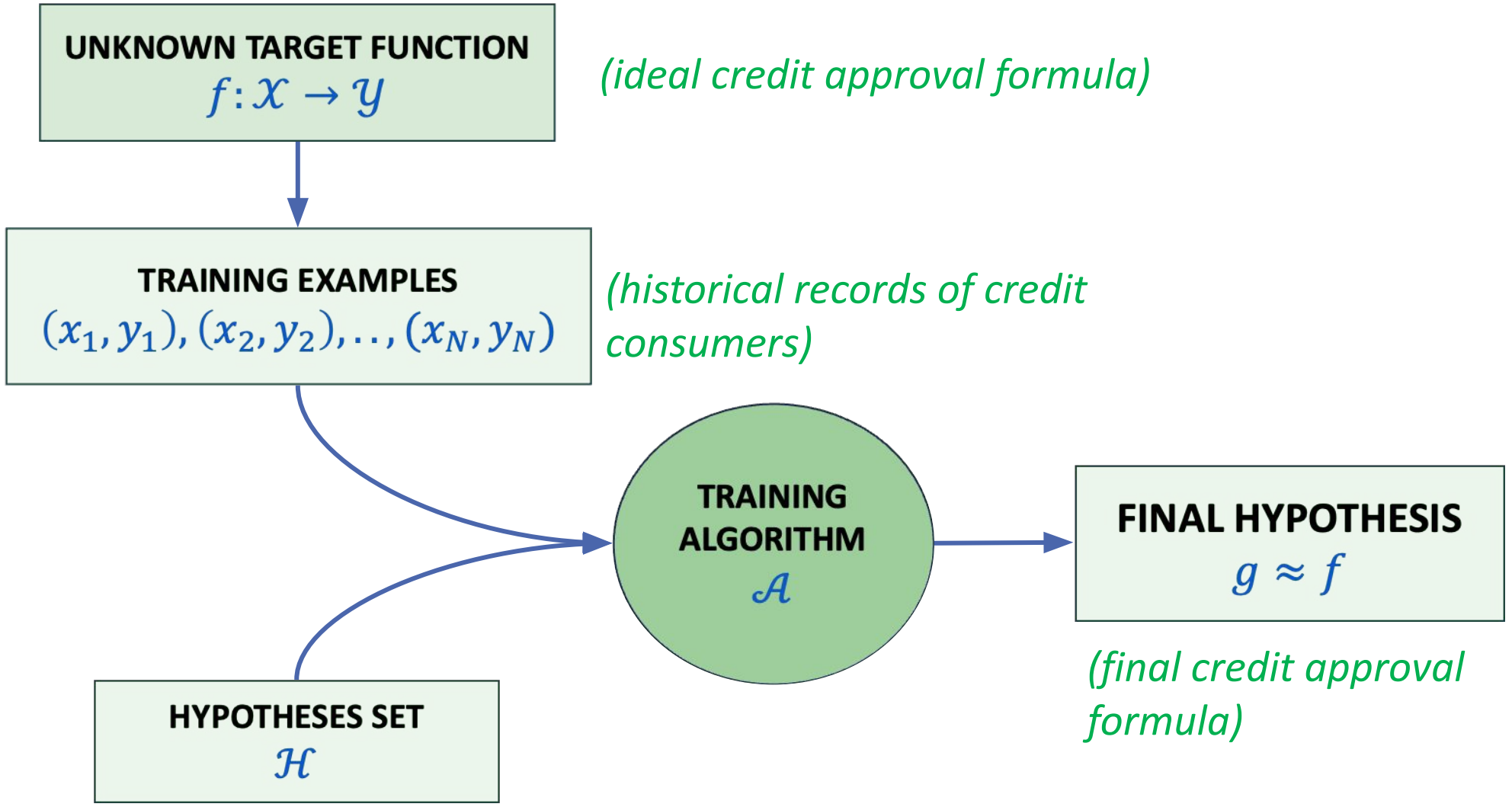
Formalization

- Input x (customer application)
- Output y (good/bad customer?)
- Target function $f: \mathcal{X} \rightarrow \mathcal{Y}$ (ideal credit approval formula)
- Data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$ (historical records)

- Hypothesis g: $\mathcal{X} \rightarrow \mathcal{Y}$ (formula to be used)

Slides from Dr. Yaser S. Abu-Mostafa (CalTech, USA)

Components of Learning



Slides from Dr. Yaser S. Abu-Mostafa (CalTech, USA)

Solution Components

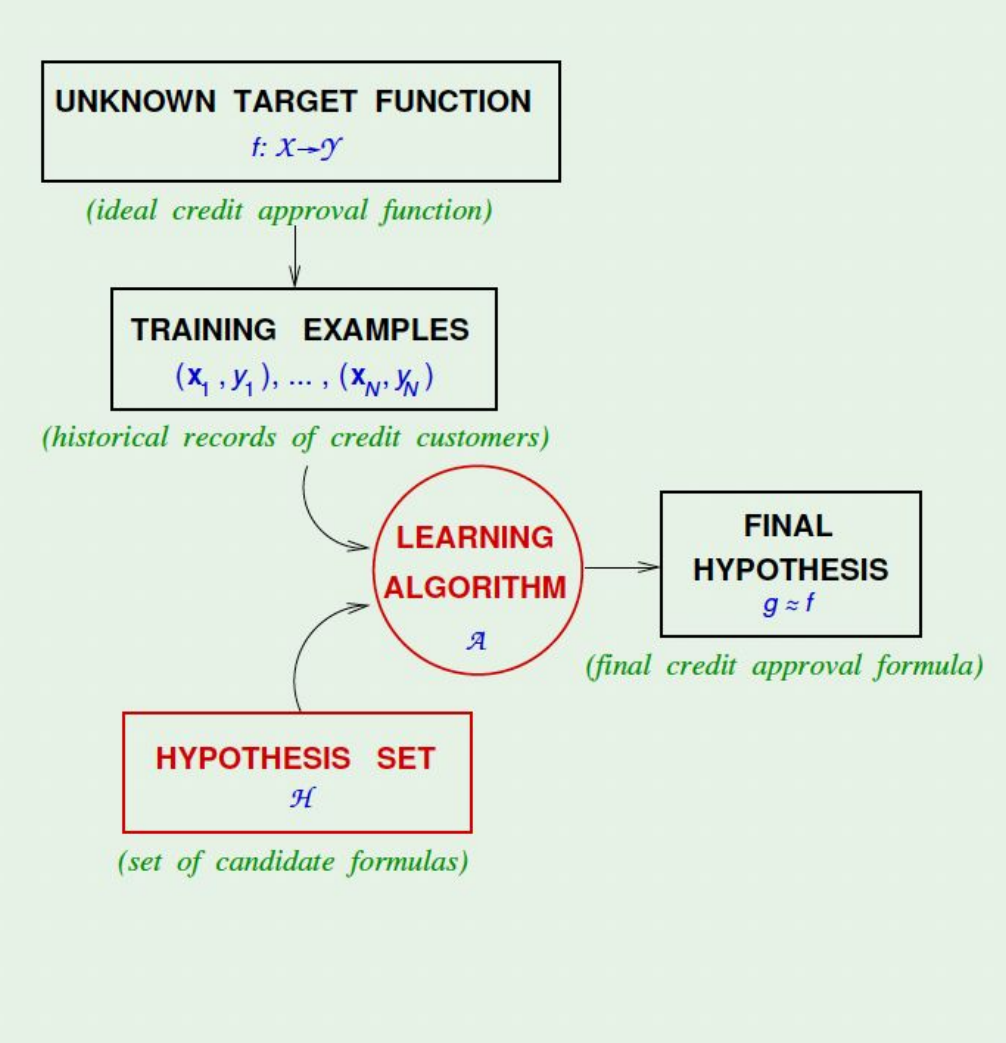
2 solution components of the learning problem

- The Hypothesis Set

  $\mathcal{H} = \{h\}, g \in \mathcal{H}$

- The Learning Algorithm

Together, they are referred to as the *learning model*.



Slides from Dr. Yaser S. Abu-Mostafa (CalTech, USA)

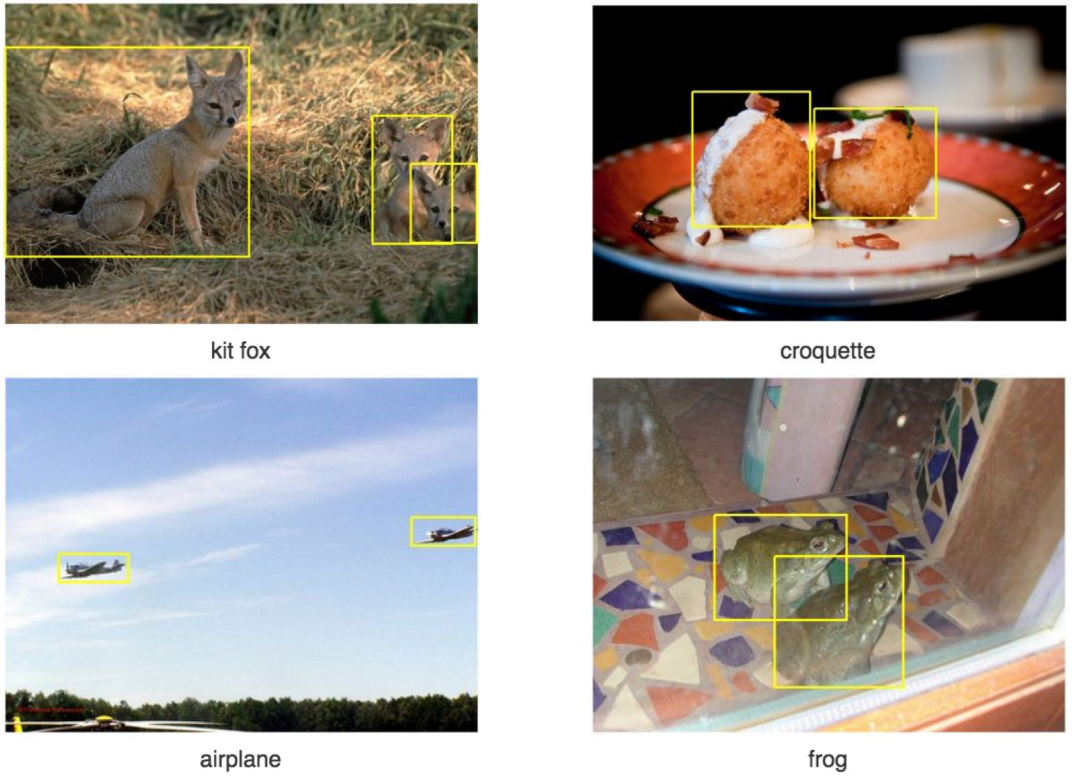
Supervised Learning in Computer Vision

- 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

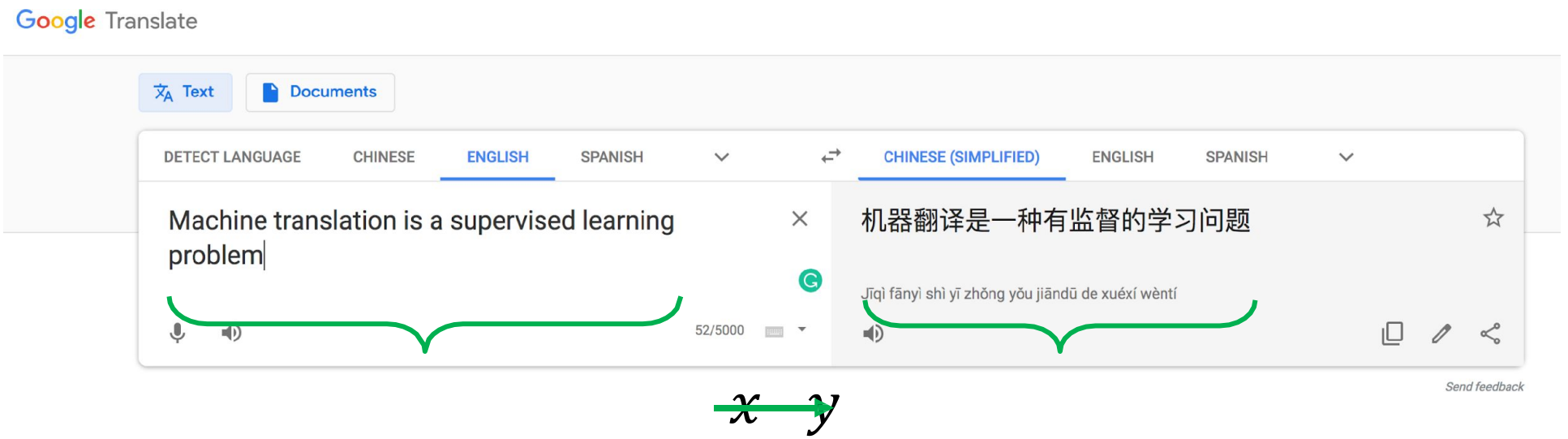
- Object localization and detection
  - x = raw pixels of the image, y = the bounding boxes



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

Supervised Learning in NLP

- Machine translation



- Note: this course only covers the basic and fundamental techniques of supervised learning

A simple hypothesis set – the 'perceptron'

For input $x = (x_1, \ldots, x_d)$ 'attributes of a customer'

Approve credit if $\Sigma_{i=1}^{d} w_i x_i >$ threshold,

Deny credit if $\Sigma_{i=1}^{d} w_i x_i <$ threshold,

Linear formula $h \in \mathcal{H}$ can be written as

$$h(x) = sign(\Sigma_{i=1}^{d} w_i x_i - threshold)$$

Slides from Dr. Yaser S. Abu-Mostafa (CalTech, USA)

Separability

- The perceptron implements

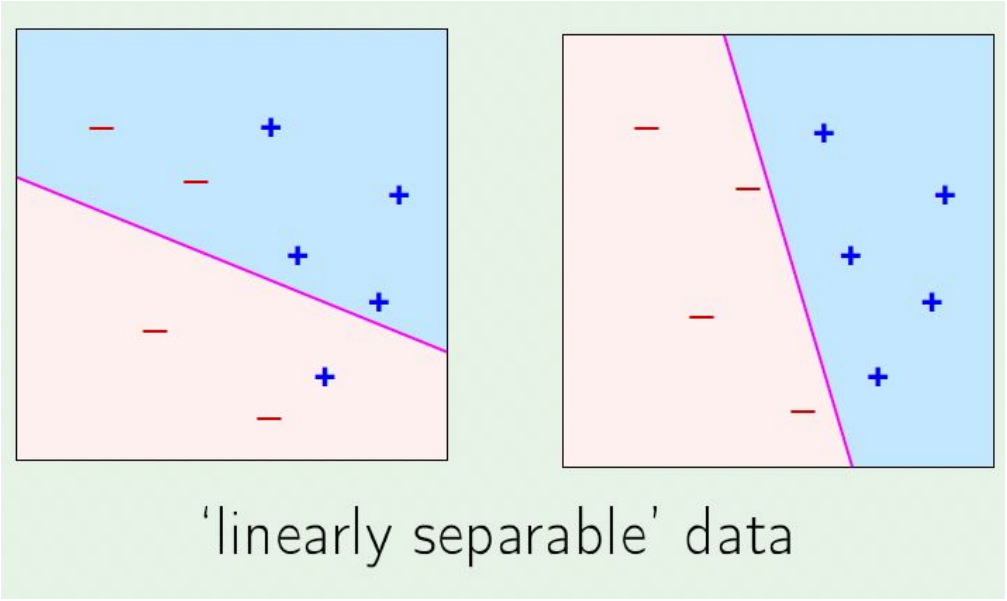
$$h(x) = sign(\Sigma_{i=1}^{d} w_i x_i + \quad w_0)$$

Introducing an artificial coordinate $x_0 = 1$

$$h(x) = sign(\Sigma_{i=0}^{d} w_i x_i)$$

In vector form, the perceptron implements

$$h(x) = sign(w^T x)$$



Slides from Dr. Yaser S. Abu-Mostafa (CalTech, USA)

A simple Learning Algo - PLA

The perceptron implements

$$h(\mathbf{x}) = \text{sign}(\mathbf{w}^{\mathsf{T}}\mathbf{x})$$

Given the training set:

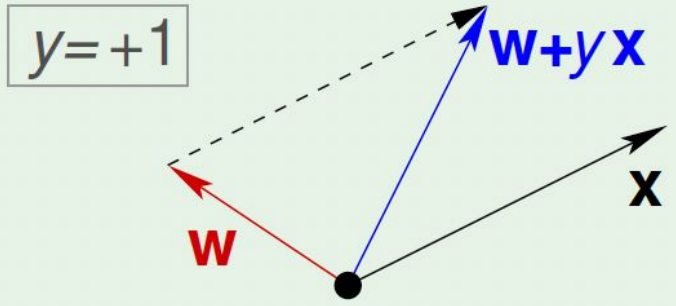
$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)$$

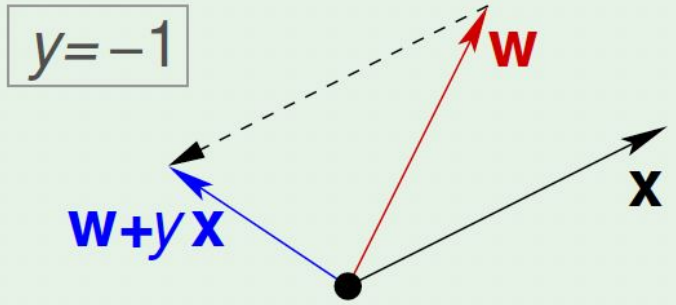
pick a misclassified point:

$$\text{sign}(\mathbf{w}^{\mathsf{T}}\mathbf{x}_n) \neq y_n$$

and update the weight vector:

$$\boxed{\mathbf{w} \leftarrow \mathbf{w} + y_n\mathbf{x}_n}$$





Slides from Dr. Yaser S. Abu-Mostafa (CalTech, USA)

Iterations of PLA

- One iteration of the PLA:

$$\mathbf{w} \leftarrow \mathbf{w} + y\mathbf{x}$$

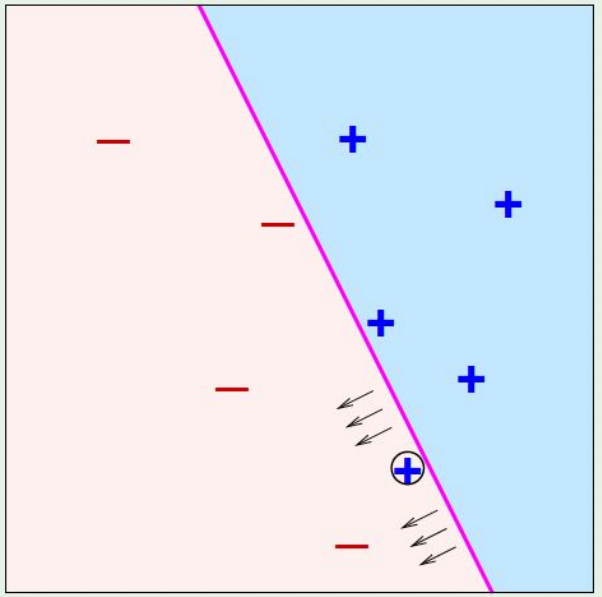
where $(\mathbf{x}, y)$ is a misclassified training point.

- At iteration $t = 1, 2, 3, \cdots$, pick a misclassified point from

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)$$

and run a PLA iteration on it.

- That's it!



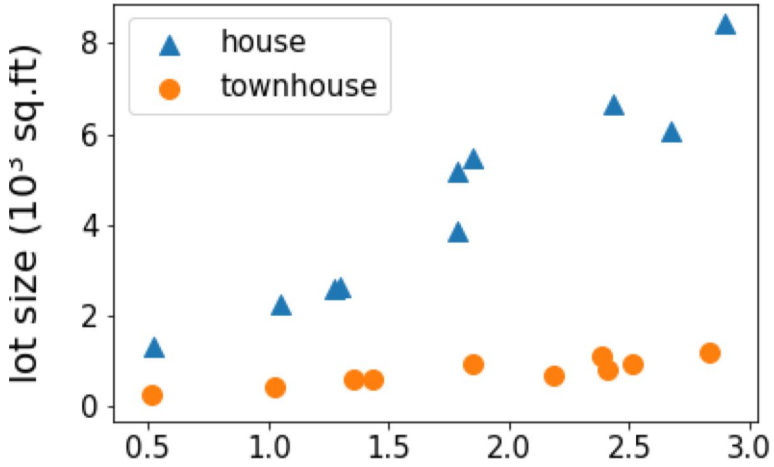
Slides from Dr. Yaser S. Abu-Mostafa (CalTech, USA)

Unsupervised Learning

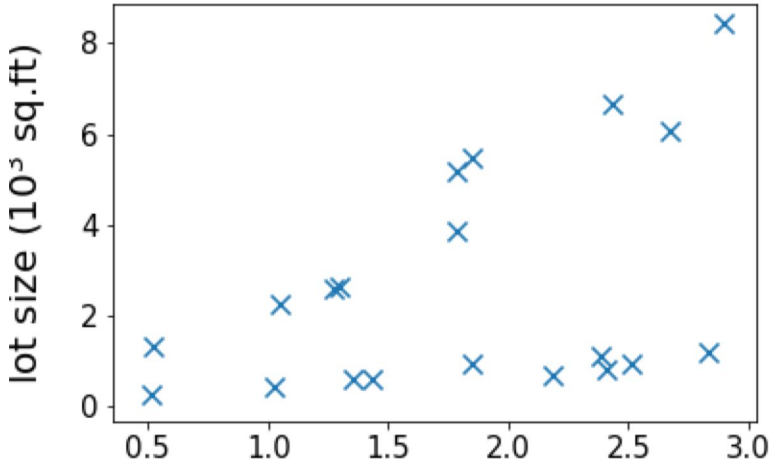
Unsupervised Learning

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

supervised

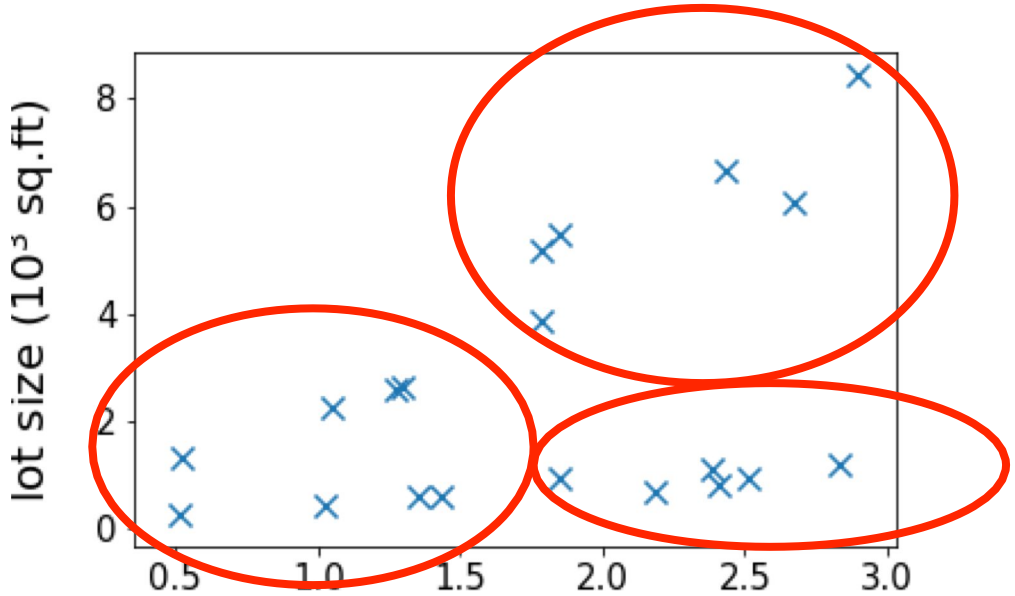


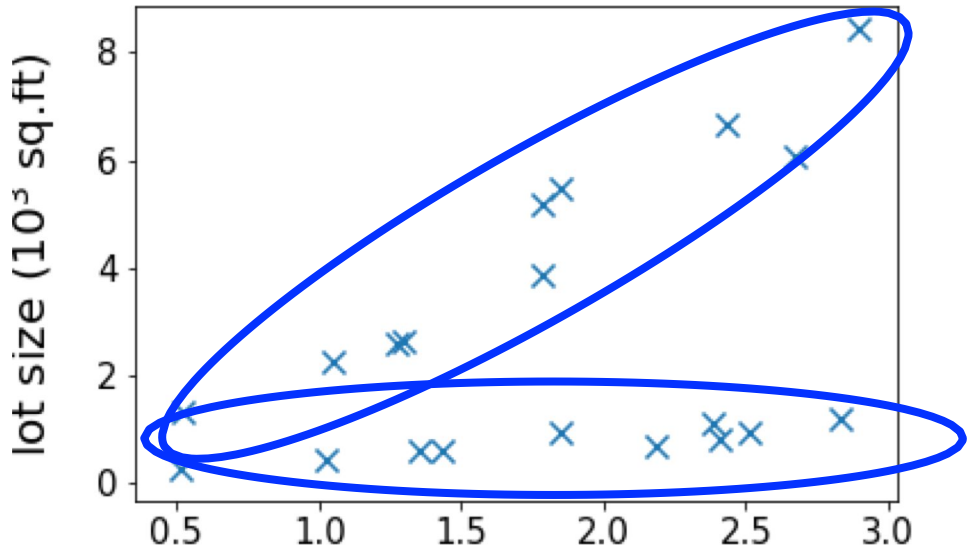
unsupervised



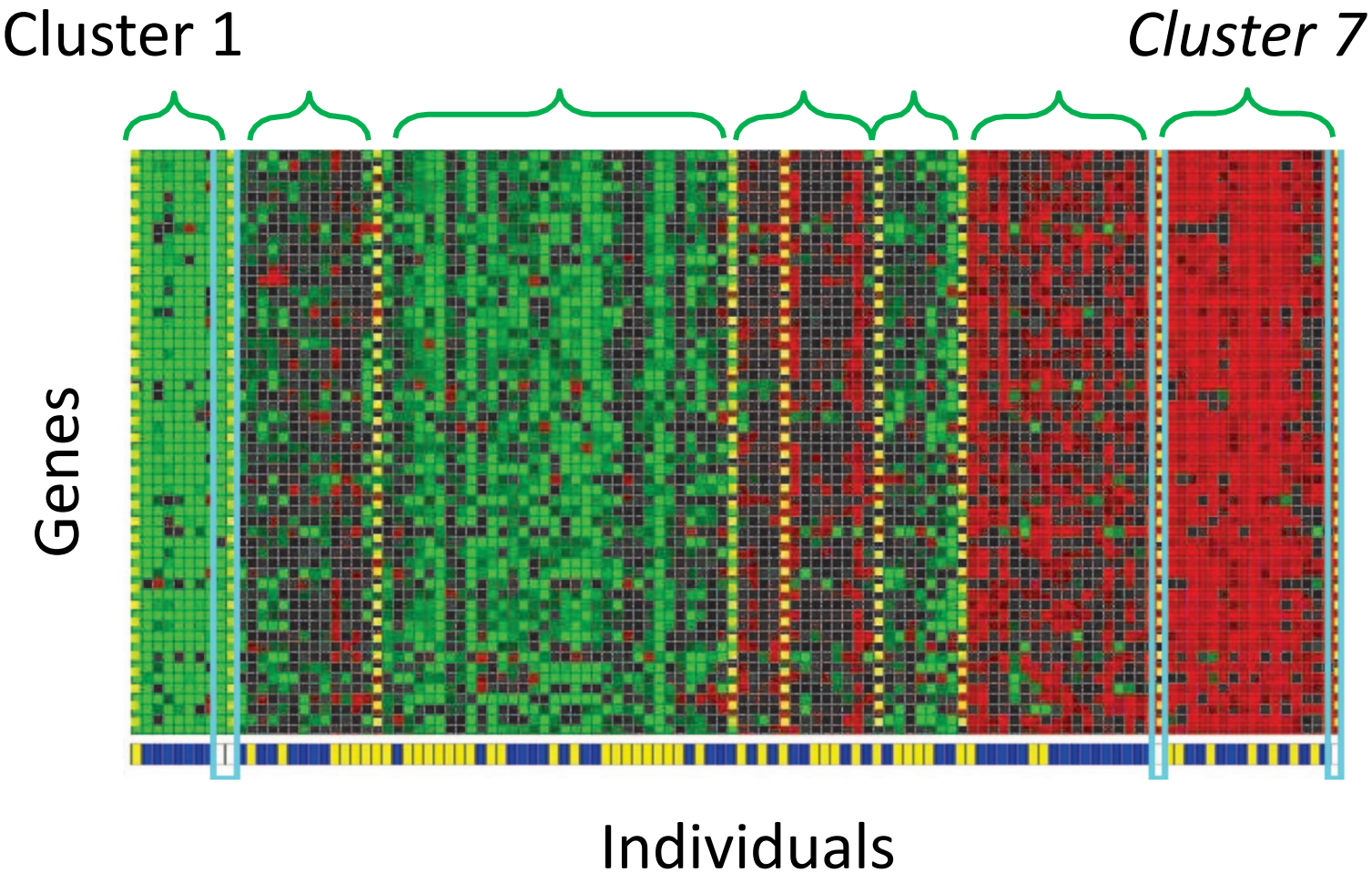
Clustering

- k-mean clustering, mixture of Gaussians



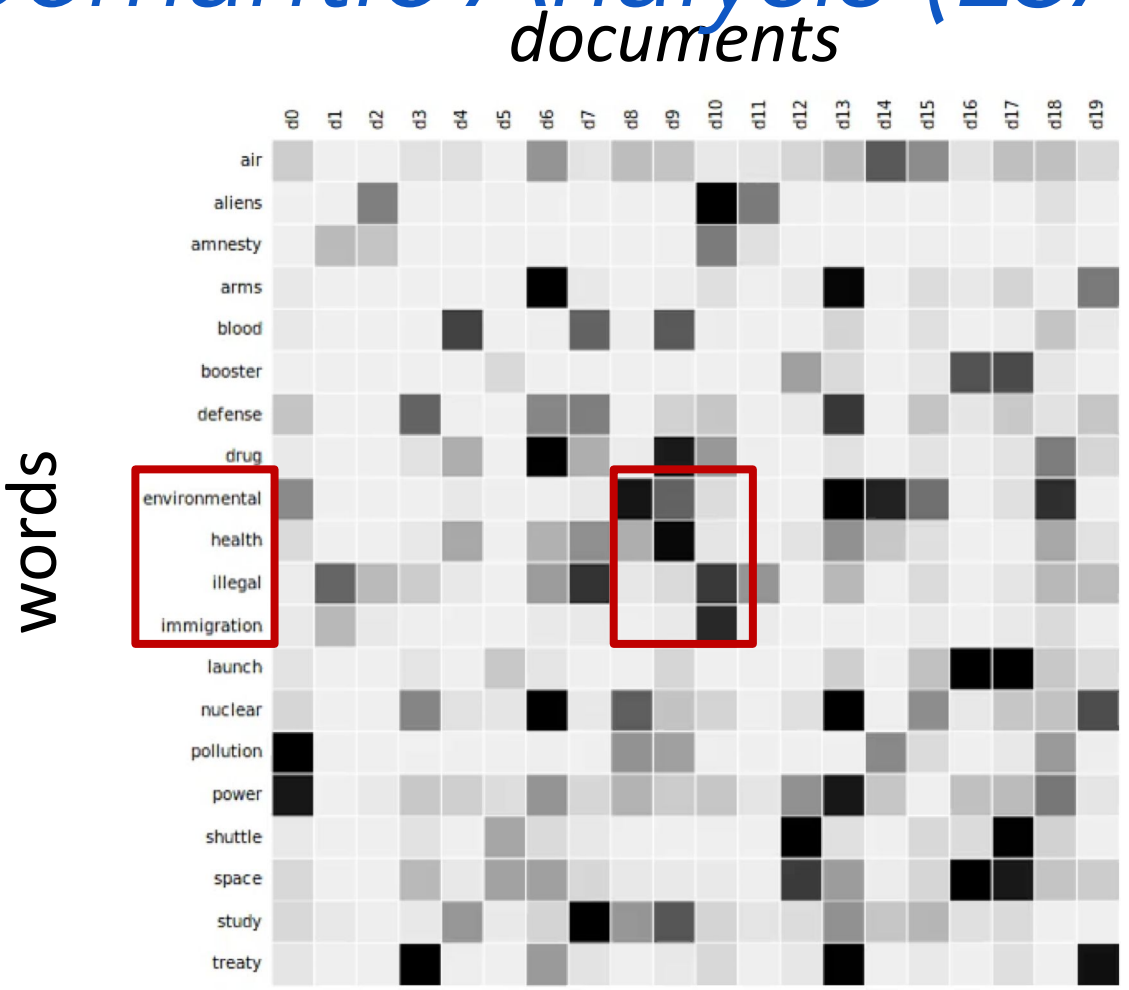


Clustering Genes



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)



- In syllabus: principal component analysis (tools used in LSA)

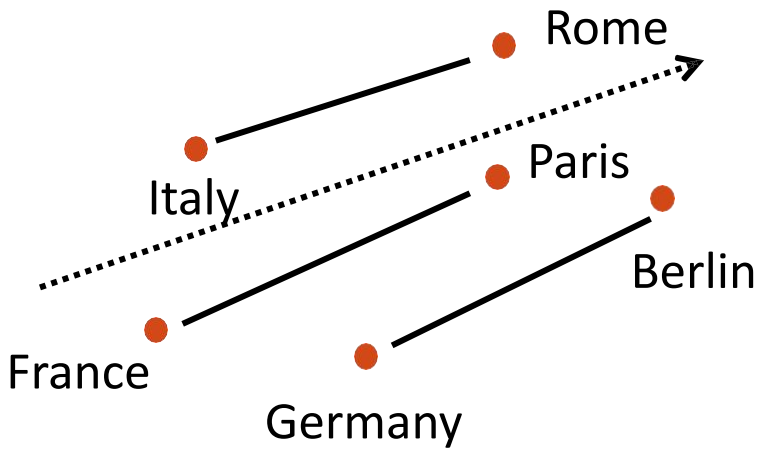
Image credit: https://commons.wikimedia.org/wiki/File:Topic_

Word Embeddings

Unlabeled dataset

Represent words by vectors

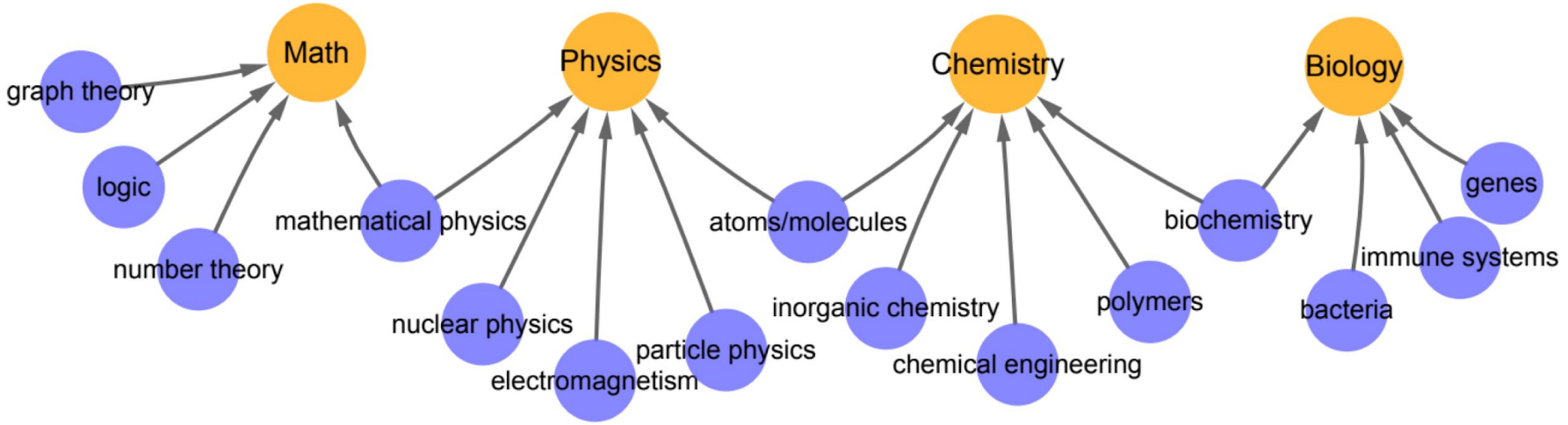
- word $\xrightarrow{enc\$\%e}$ vector
- relation $\xrightarrow{enc\$\%e}$ direction



Word2vec [Mikolov et al'13]

GloVe [Pennington et al'14]

Clustering Words with Similar Meanings (Hierarchically)

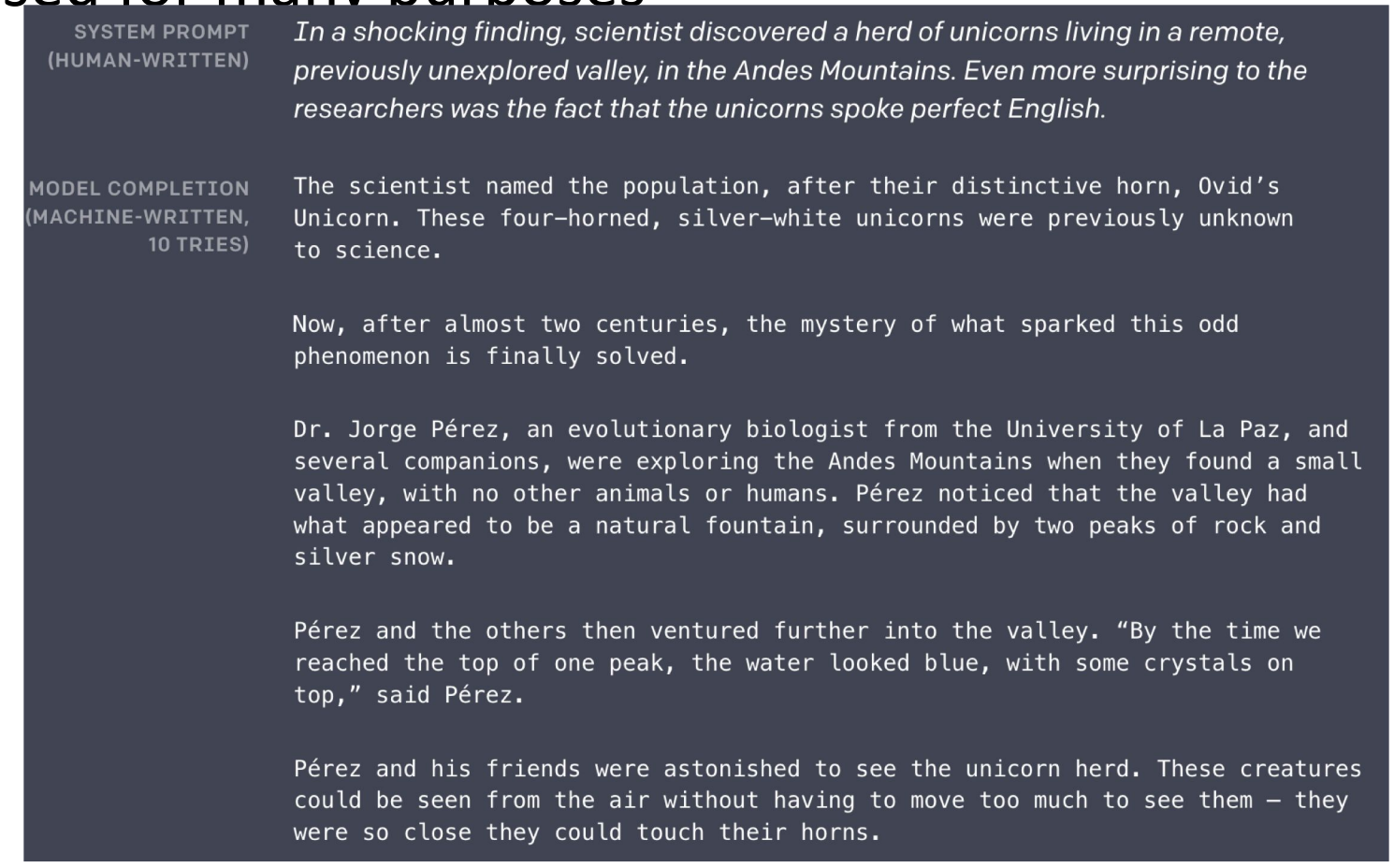


	logic deductive propositional semantics	graph subgraph bipartite vertex	boson massless particle higgs	polyester polypropylene resins epoxy	acids amino biosynthesis peptide
tag	*logic*	*graph theory*	*particle physics*	*polymer*	*biochemistry*

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

Large Language Models

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



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/

Types of Learning

"using a set of observations to uncover an underlying process"

Broad premise Many variations

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

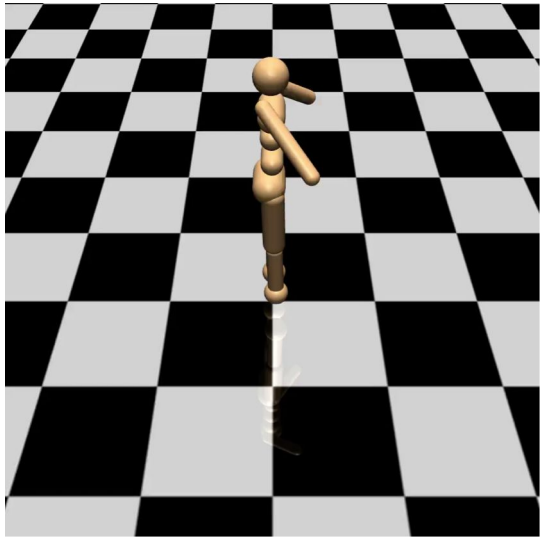
Slides from Dr. Yaser S. Abu-Mostafa (CalTech, USA)

Reinforcement Learning

Learning to make sequential decisions


AlphaGo
Lee Sedol
ALPHAGO

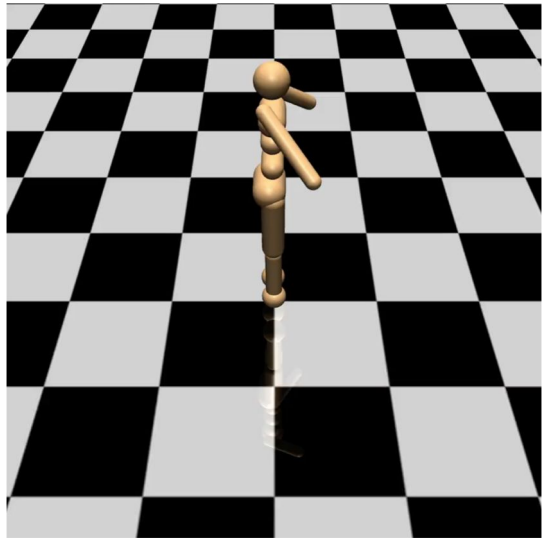
learning to walk to the right



Iteration 10

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

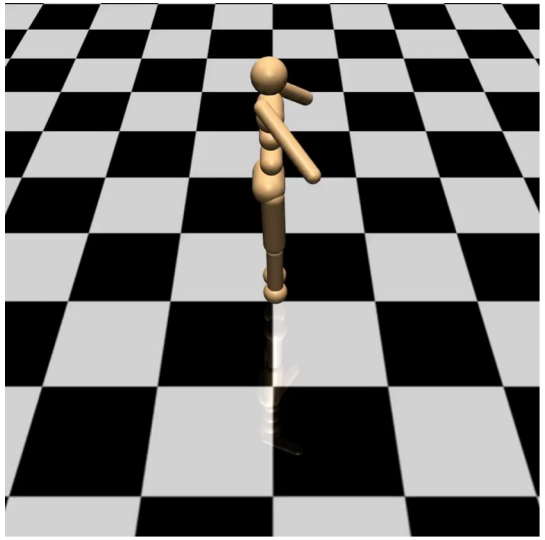
learning to walk to the right



Iteration 20

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

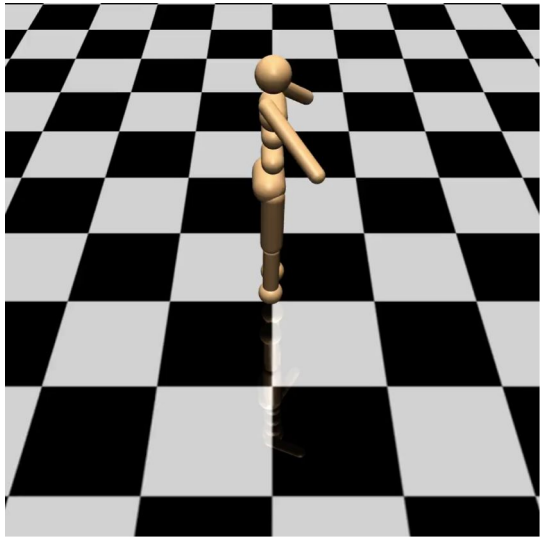
learning to walk to the right



Iteration 80

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

learning to walk to the right



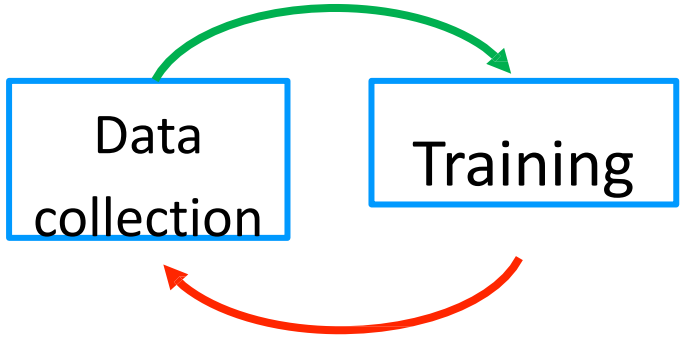
Iteration 210

[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