

# MACHINE LEARNING: INTRODUCTION

## CS60050 (SEC 1)

August 2<sup>nd</sup>, 2023

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://sites.google.com/view/cs60050-fall-2023/home>
- Course Timings
  - Wed 11-12 PM, Thu 12-1 PM, Fri 8-9 AM
  - Section 1 (Dr. Somak Aditya, NC442), Section 2 (Dr. Sudeshna Sarkar)
- Office: CS 305 (Somak Aditya), CS 201 (Sudeshna Sarkar)
- Teaching Assistants
  - Sachin Vashistha, Paneliya Yashkumar Shaileshbhai, Ashwamegh Rathore, Biplab Roy, Aryan Singh, Tarun Mohanty, Shrinivas Khiste

# *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 (3-4) - 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.

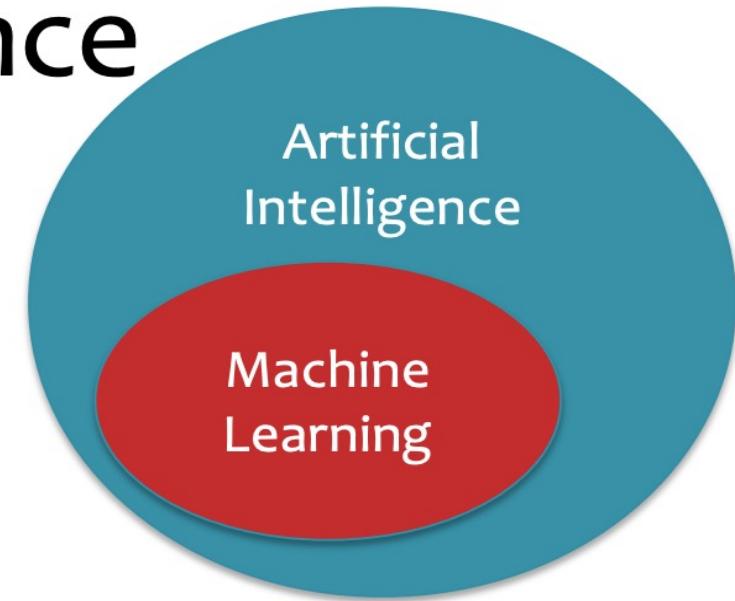
*What is Machine Learning?*

# 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

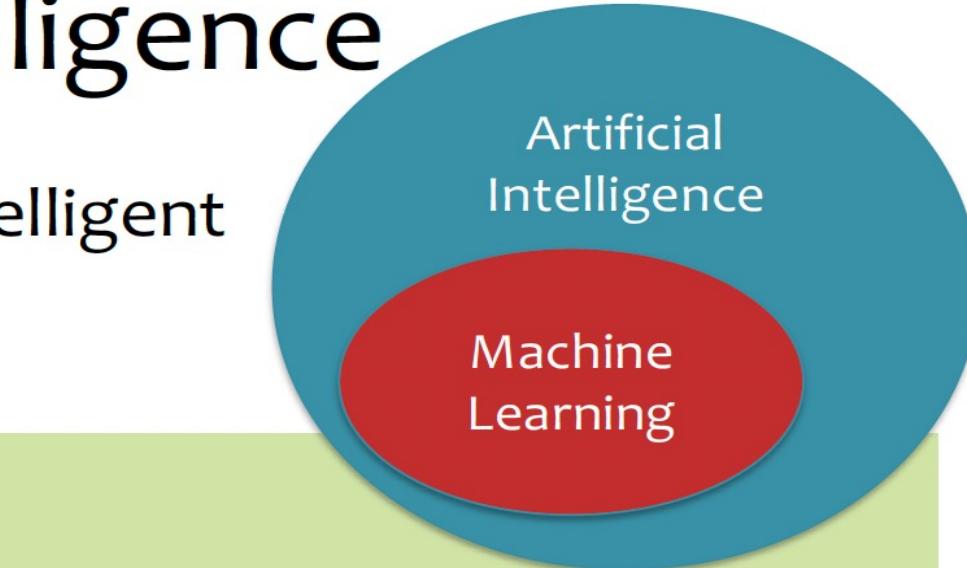


# Artificial Intelligence

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

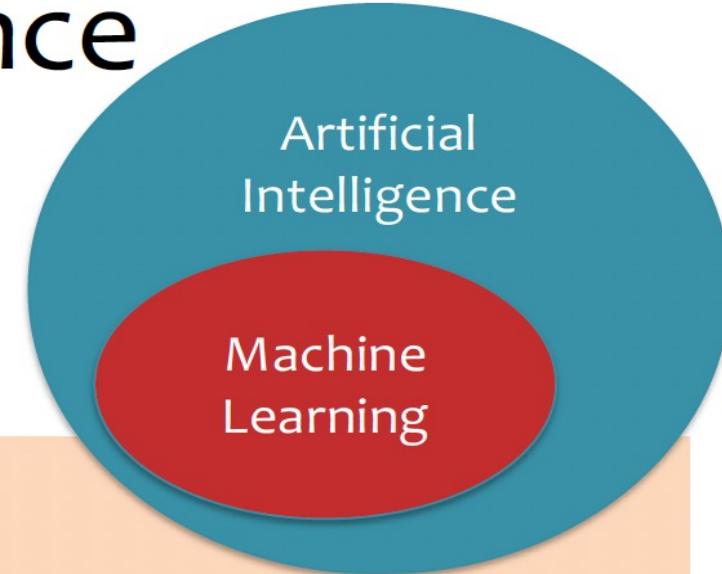


# Artificial Intelligence

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

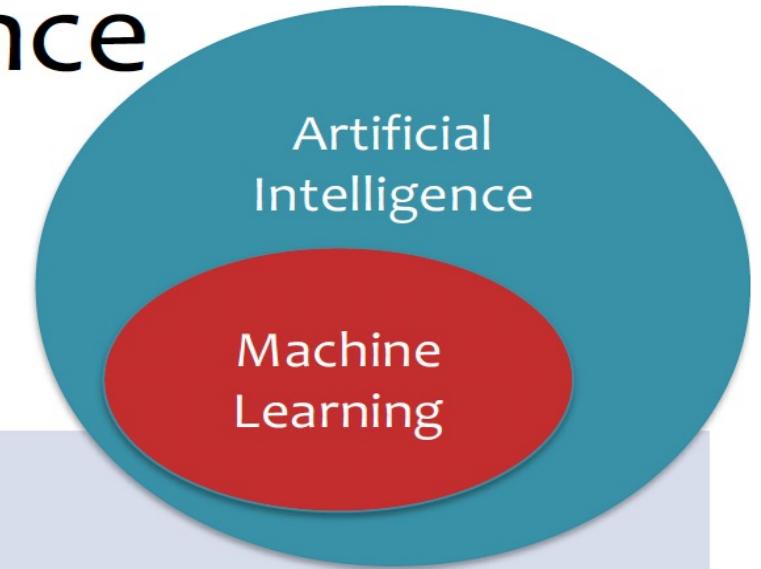


# Artificial Intelligence

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

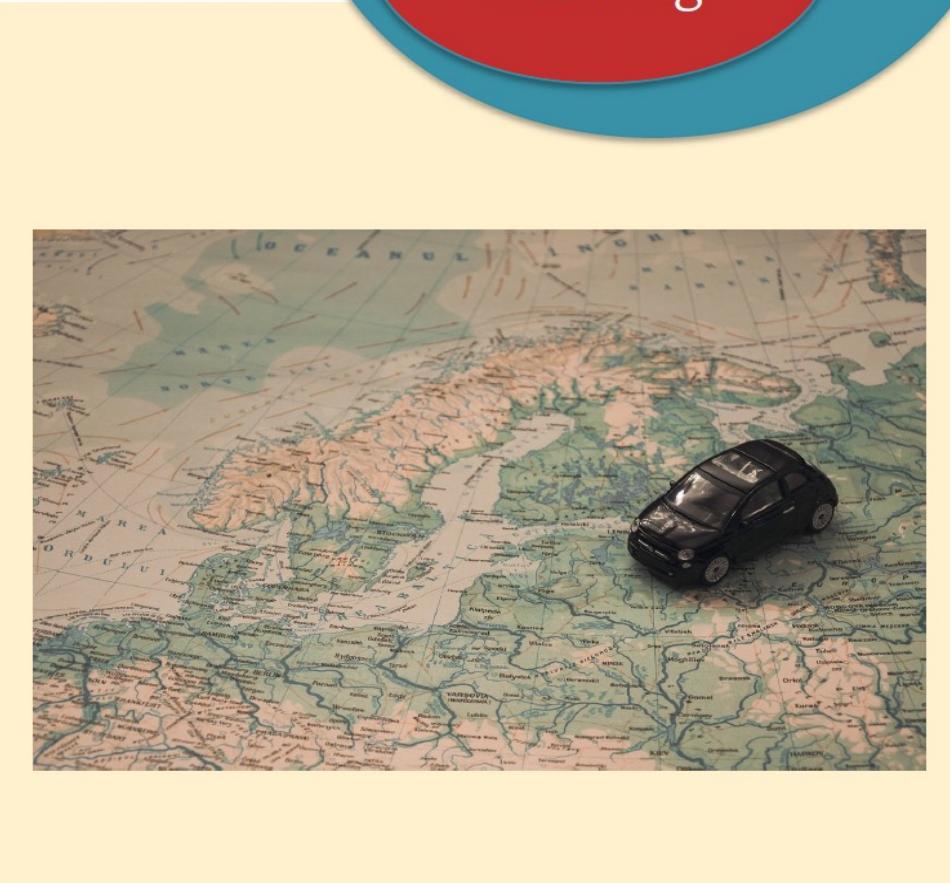
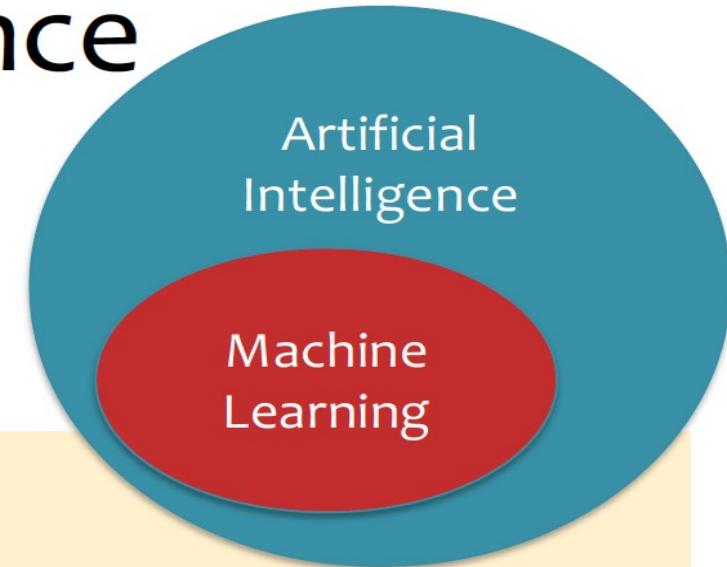


# Artificial Intelligence

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

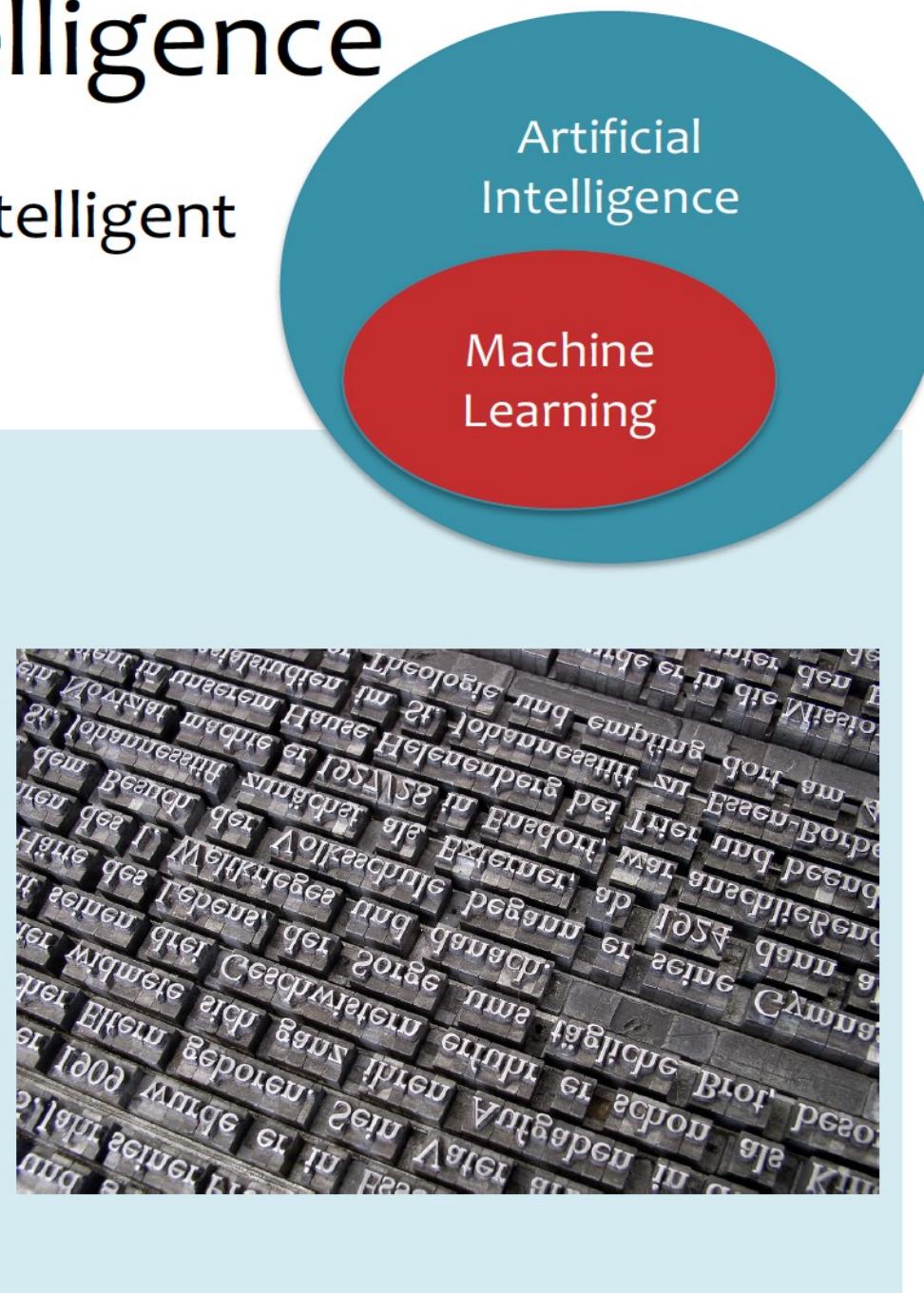


# Artificial Intelligence

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



# 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

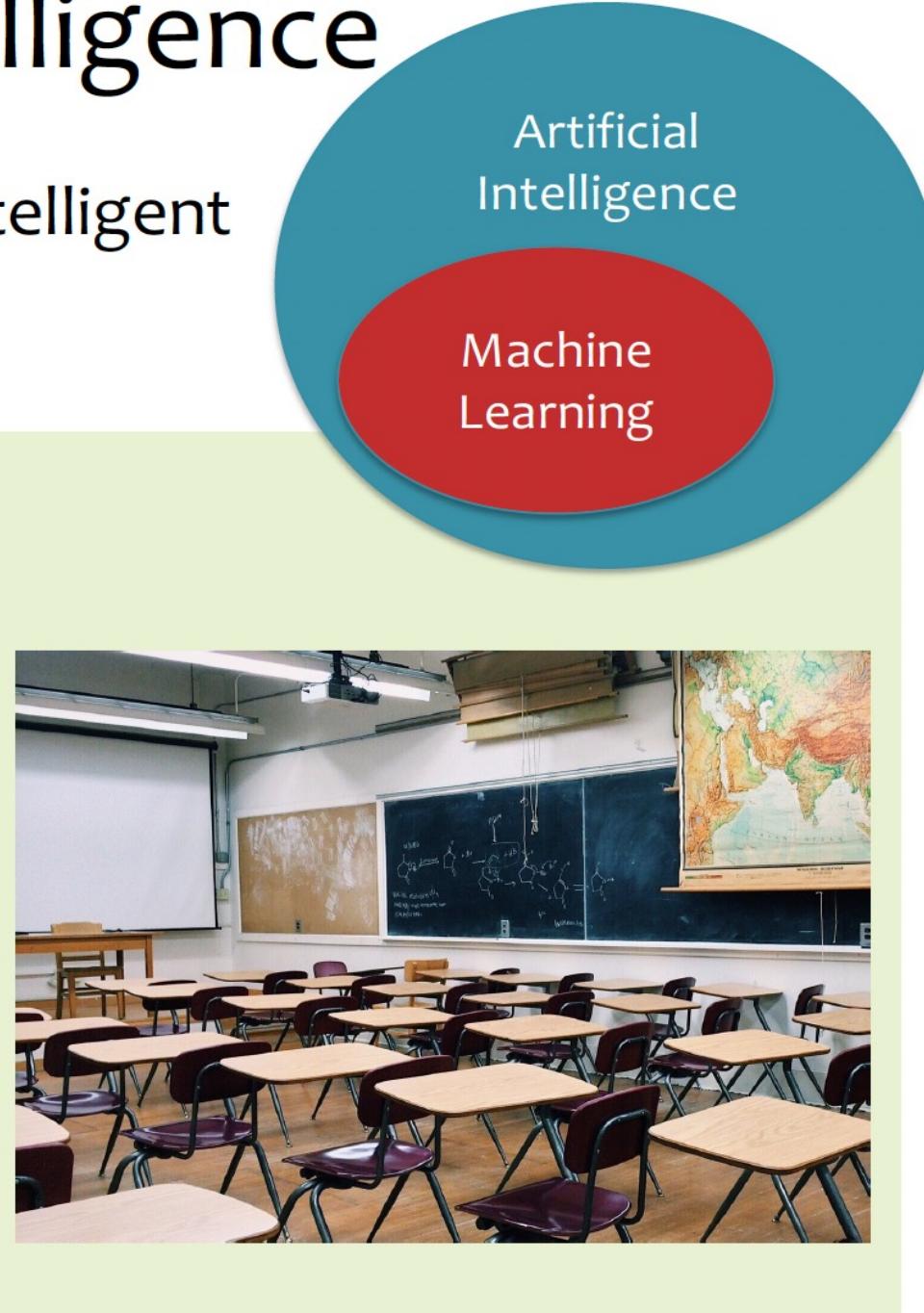


# Artificial Intelligence

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



# What is Machine Learning?

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



# What is ML?

## Speech Recognition

### 1. Learning to recognize spoken words

THEN	NOW
<p>“...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...”</p> <p>(Mitchell, 1997)</p>	

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

## Robotics

### 2. Learning to drive an autonomous vehicle

THEN	NOW
<p>“...the ALVINN system (Pomerleau 1989) has used its learned strategies to drive unsupervised at 70 miles per hour for 90 miles on public highways among other cars...”</p> <p>(Mitchell, 1997)</p>	 <p>waymo.com</p>

## Games / Reasoning

### 3. Learning to beat the masters at board games

THEN	NOW
<p>“...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...”</p> <p>(Mitchell, 1997)</p>	

## Computer Vision

### 4. Learning to recognize images

“...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/c399862d3b9d6b7c8436e924a68c45b-Paper.pdf>

## Learning Theory

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

#### Sample Complexity Results

Definition 1.1. The sample complexity of a learning algorithm is the number of examples required to achieve arbitrary small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

Four Cases we care about...

Realizable

Infinite  $\mathcal{H}$

Finite  $\mathcal{H}$

Infinite  $\mathcal{H}$

Unrealizable

Finite  $\mathcal{H}$

# 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/>



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



24

<https://flic.kr/p/HNUUJv>

Slides from Dr. Matt Gormley (CMU)

# Societal Impacts of ML

The Washington Post  
Democracy Dies in Darkness

## 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.

# 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

## Theoretical Foundations:

*What principles guide learning?*

- probabilistic
- information theoretic
- evolutionary search
- ML as 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)

## 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

## Application Areas

*Key challenges?*

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

# Syllabus (Tentative)

- *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*

# *Well-Posed Learning Problems*

***Three components***  $\langle T, P, E \rangle$ :

- 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.

# *Example Learning Problem(s)*

*Learning to respond to voice commands (Siri)*

- 1. *Task, T:*



- 2. *Performance measure, P:*



- 3. *Experience, E:*



# 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))

# *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

# Capturing the Knowledge of Experts



## Solution #2: Annotate Data and Learn

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

$x^{(1)}$ : How do I get to Starbucks?

$y^{(1)}$ : directions (here,  
nearest (Starbucks) )

$x^{(2)}$ : Show me the closest Starbucks

$y^{(2)}$ : map (nearest (Starbucks) )

$x^{(3)}$ : Send a text to John that I'll be late

$y^{(3)}$ : txtmsg (John, I'll be late)

$x^{(4)}$ : Set an alarm for seven in the morning

$y^{(4)}$ : setalarm (7 : 00AM)

# *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*

# *The Learning Problem - Outline*

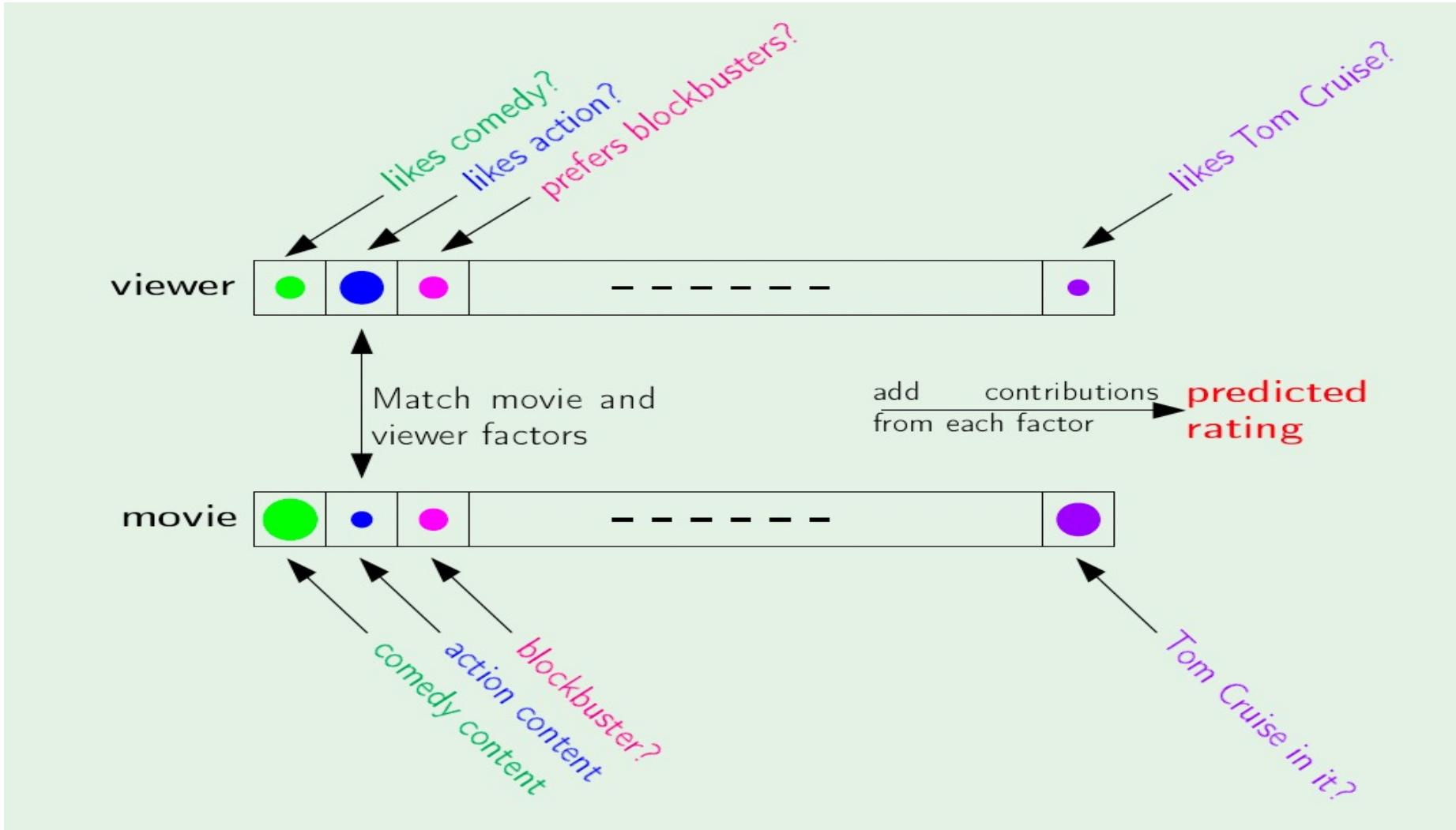
- *Example of machine learning*
- *Components of Learning*
- *A simple model*
- *Types of learning*
- *Puzzle*

# *Example: Predicting how a viewer will rate a movie*

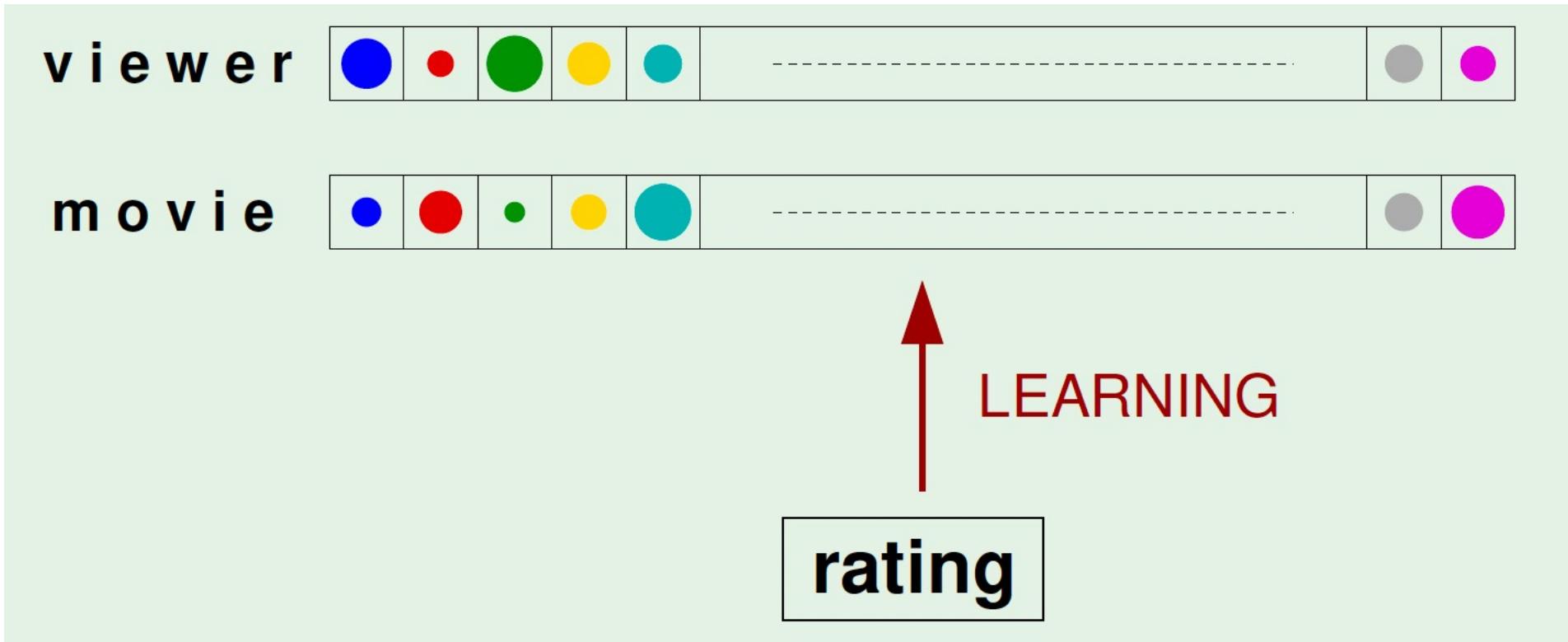
*10% improvement = 1 million dollar prize*

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

# Movie Rating – A Solution



# *The Learning approach*



# *Components of Learning*

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

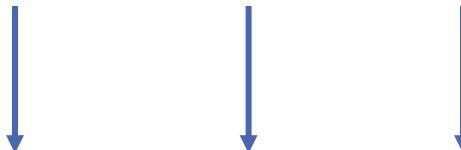
Age	23 years
Gender	male
Annual salary	\$30000
Years in residence	1 year
Years in job	1 year
Current debt	\$15000
...	...

- Approve Credit?

# *Components of Learning*

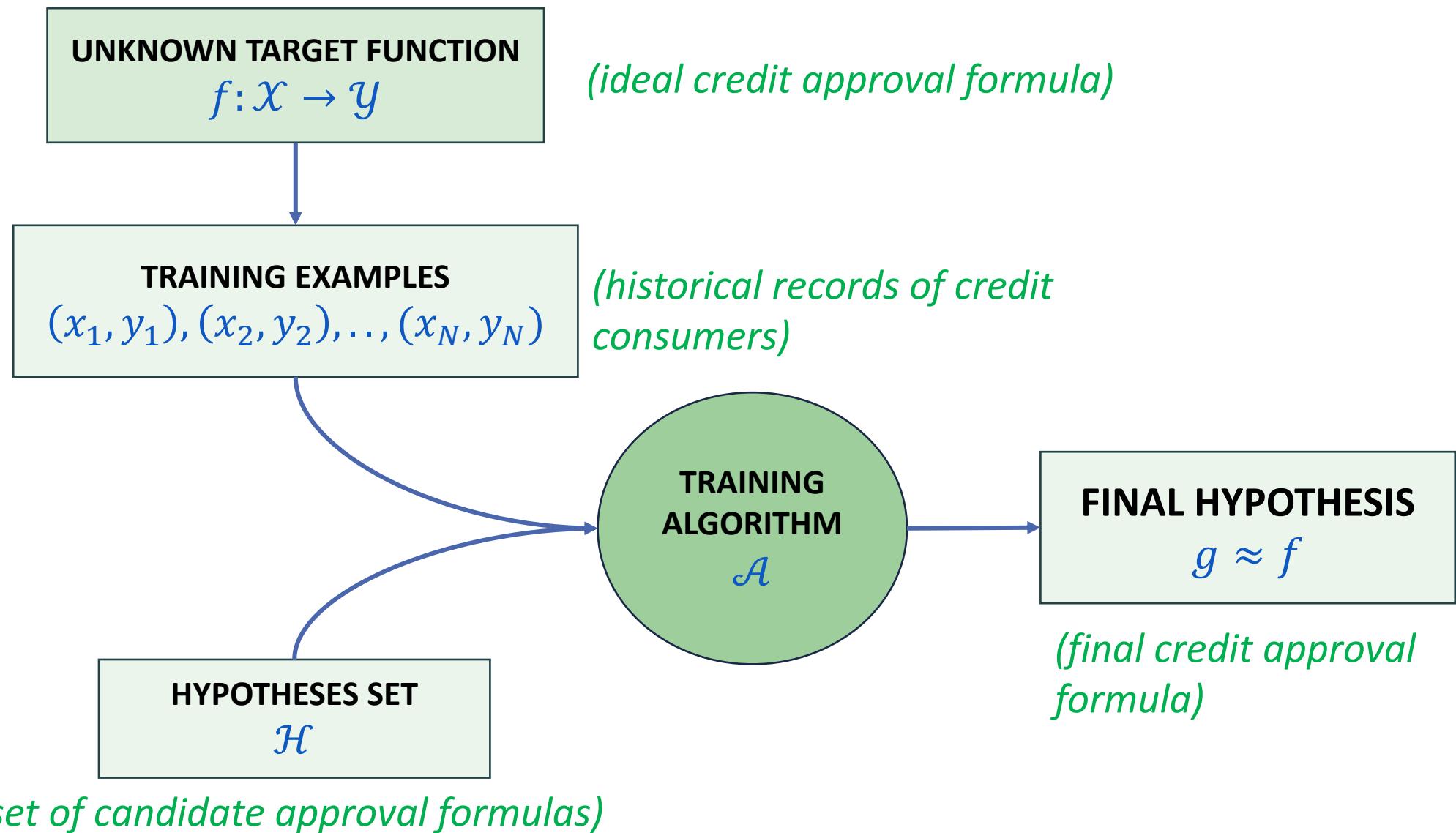
## Formalization

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



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

# Components of Learning

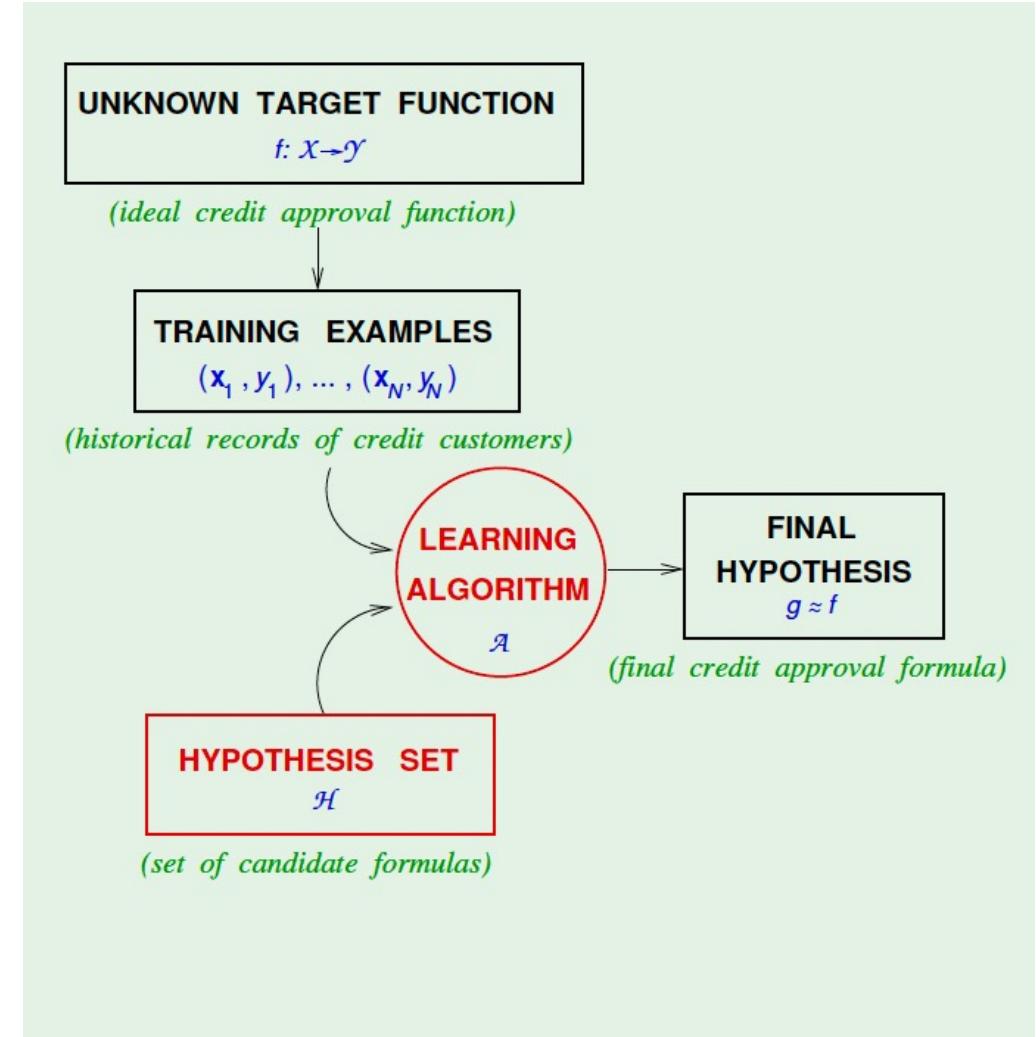


# Solution Components

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*.



# *A simple hypothesis set – the 'perceptron'*

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

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

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

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

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

# Separability

The perceptron implements

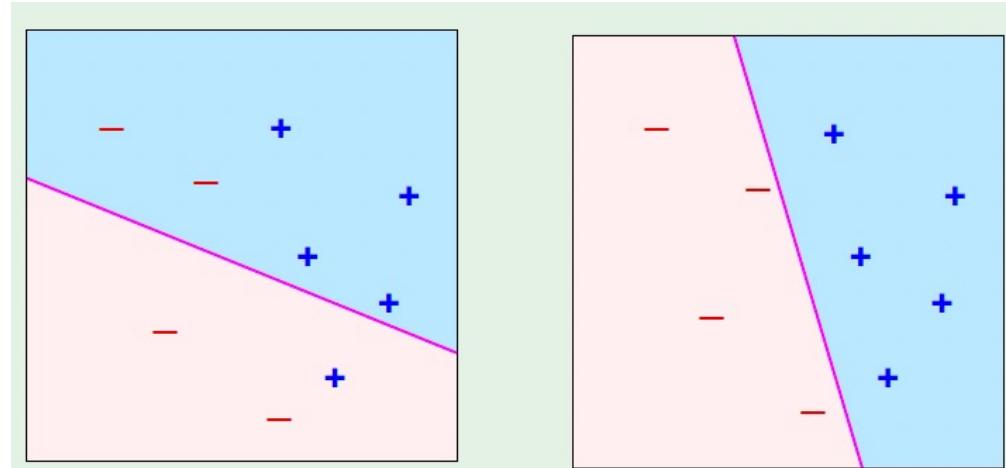
$$h(x) = \text{sign}(\sum_{i=1}^d w_i x_i + w_0)$$

Introducing an artificial coordinate  $x_0 = 1$

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

In vector form, the perceptron implements

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



'linearly separable' data

# A simple Learning Algo - PLA

The perceptron implements

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

Given the training set:

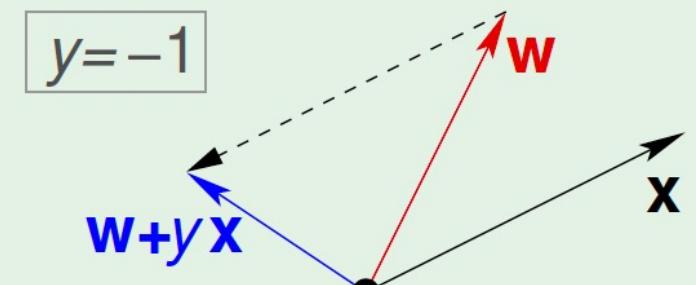
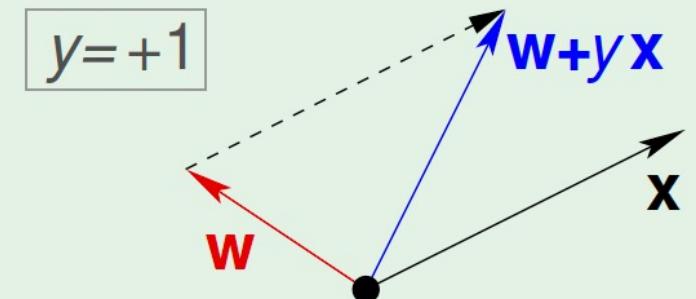
$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)$$

pick a **misclassified** point:

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

and update the weight vector:

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



# *Iterations of PLA*

- One iteration of the PLA:

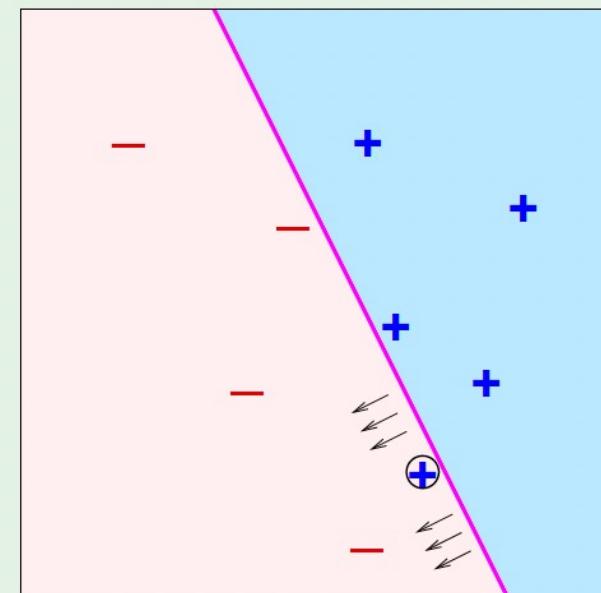
$$\mathbf{w} \leftarrow \mathbf{w} + y\mathbf{x}$$

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

- At iteration  $t = 1, 2, 3, \dots$ , pick a misclassified point from
$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)$$

and run a PLA iteration on it.

- That's it!



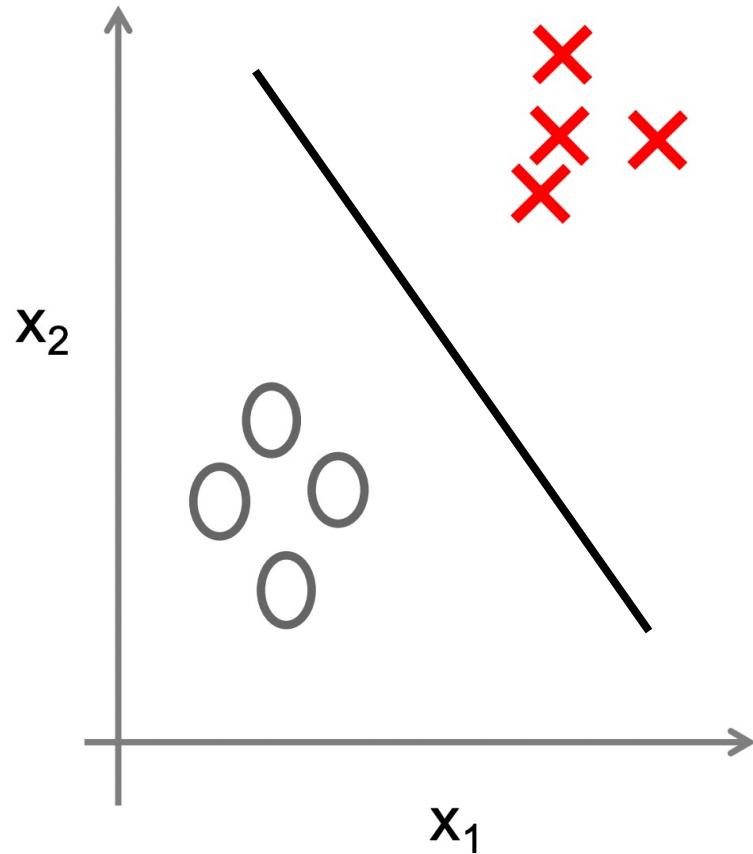
# *Types of Learning*

“using a set of observations to uncover an underlying process”

Broad premise → Many variations

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

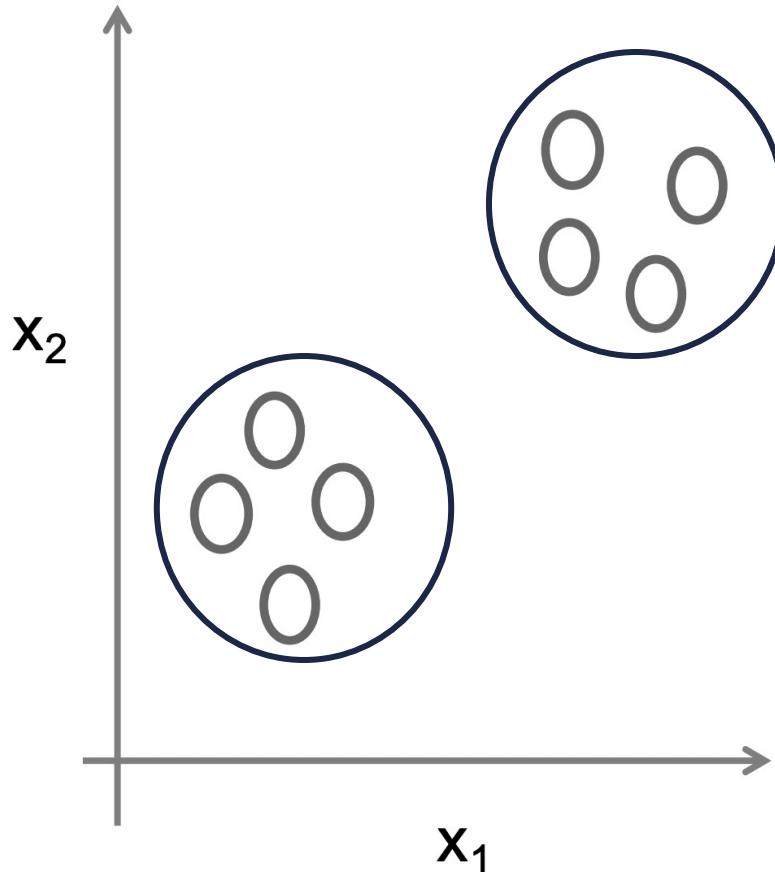
# Supervised Learning



Classification (one example)

- Input: (input-features, correct output)
  - <size, #rooms>, <cheap/costly>
- Output of learning algorithm
  - Function maps features to output
  - $F(<\text{size}, \#\text{rooms}>) = \text{cheap}/\text{costly}$

# *Unsupervised Learning*

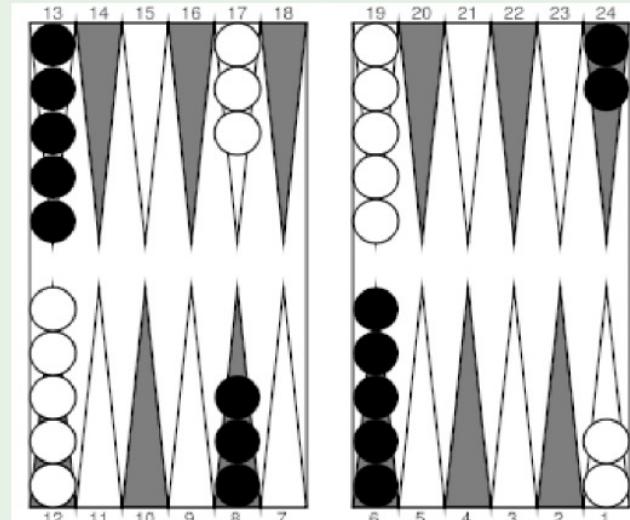


## Clustering

- Input: (input features, ?)
  - <size, #rooms>
- Output: Clusters
  - Cluster 1 probably cheap.  
Cluster 2 probably costly.

# Reinforcement Learning

Instead of (**input, correct output**),  
we get (**input, some output, grade for this output**)



The world champion was  
a neural network!