



# 10-301/10-601 Introduction to Machine Learning

Machine Learning Department  
School of Computer Science  
Carnegie Mellon University

# Course Overview

Matt Gormley  
Lecture 1  
Jan. 18, 2023

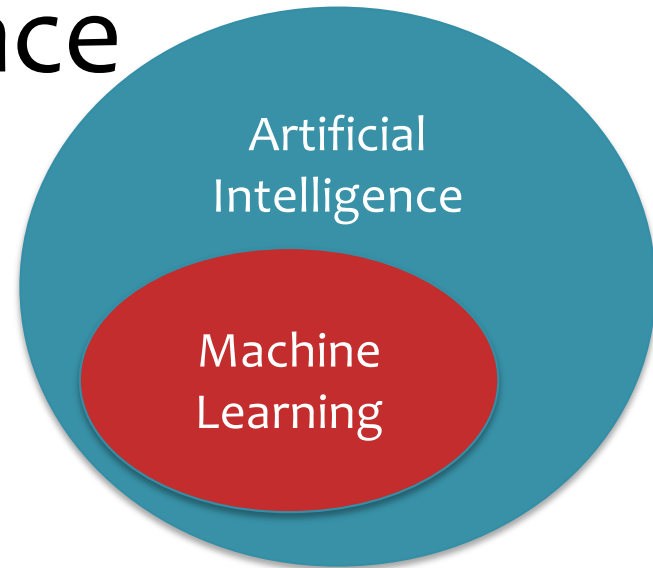
# **WHAT IS MACHINE LEARNING?**

# 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

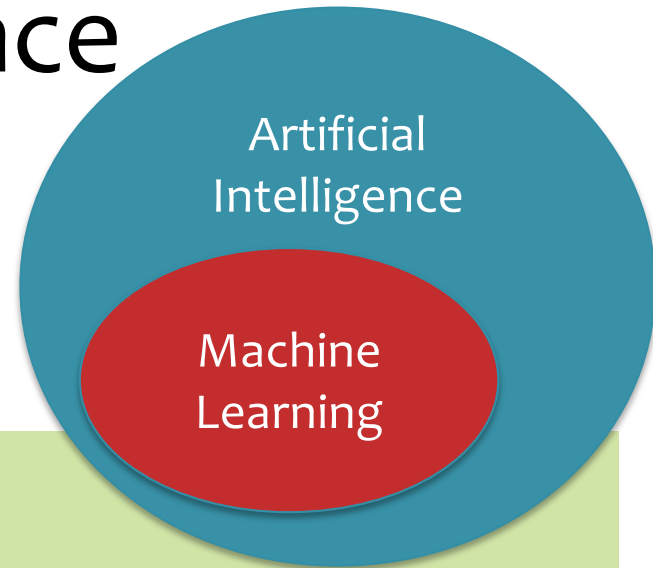


# Artificial Intelligence

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

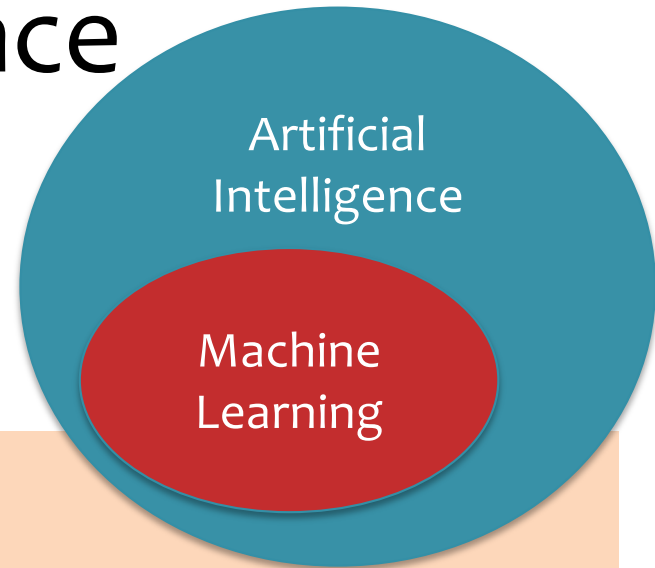


# Artificial Intelligence

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

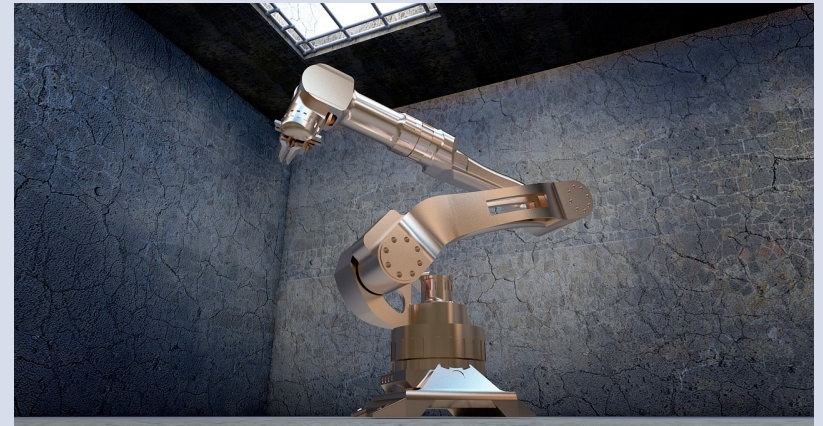
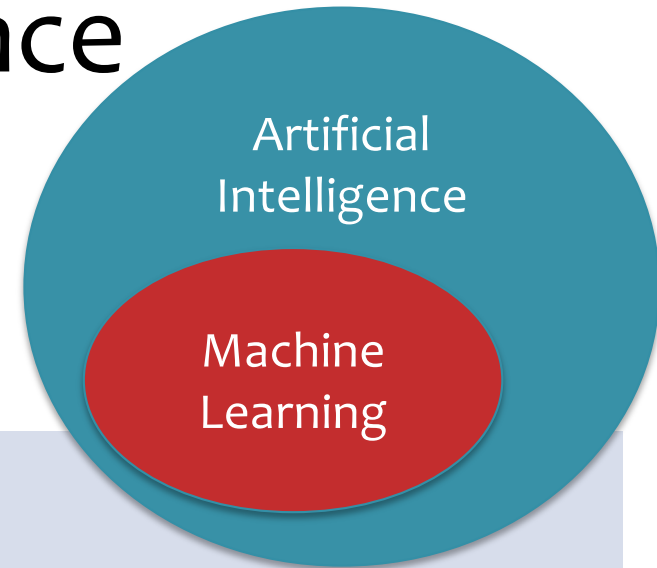


# Artificial Intelligence

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

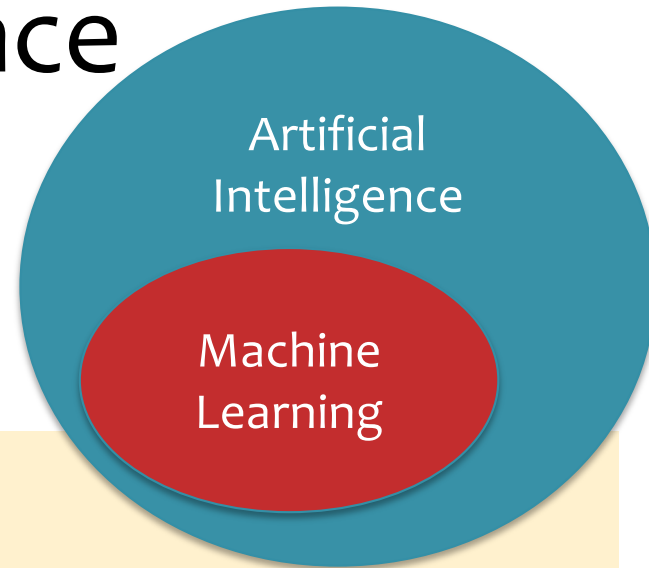


# Artificial Intelligence

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

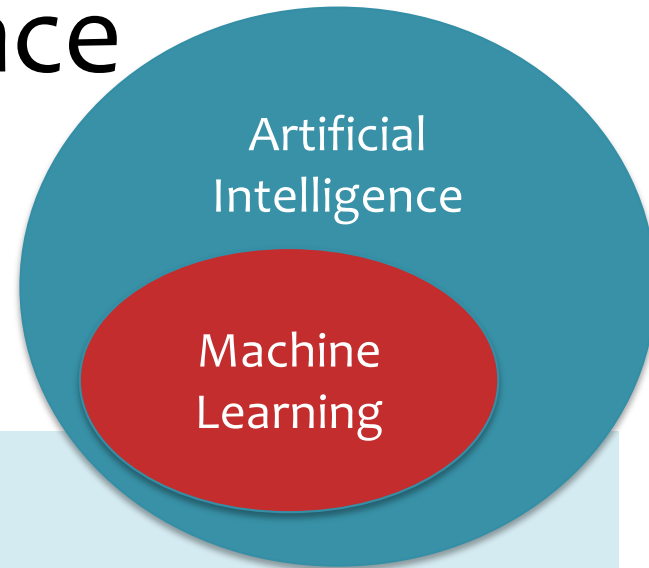


# Artificial Intelligence

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



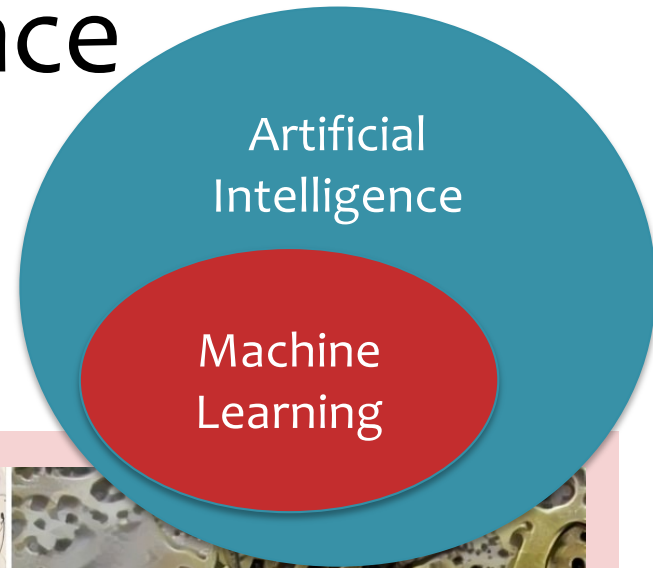


# Artificial Intelligence

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

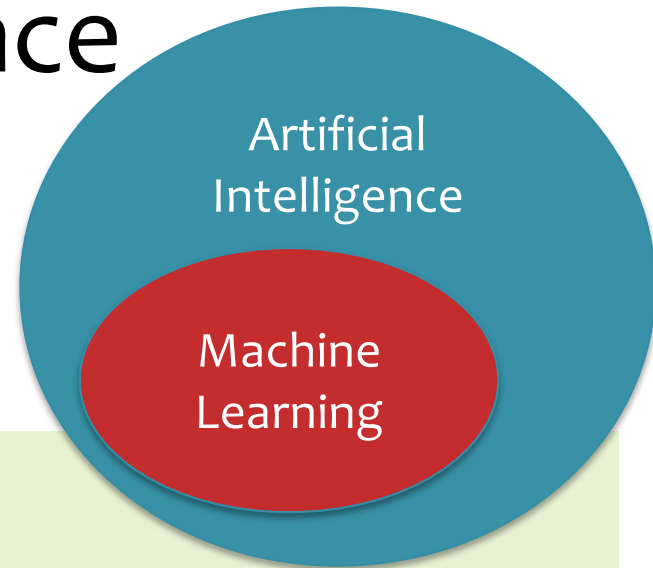


# Artificial Intelligence

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:

Machine Learning

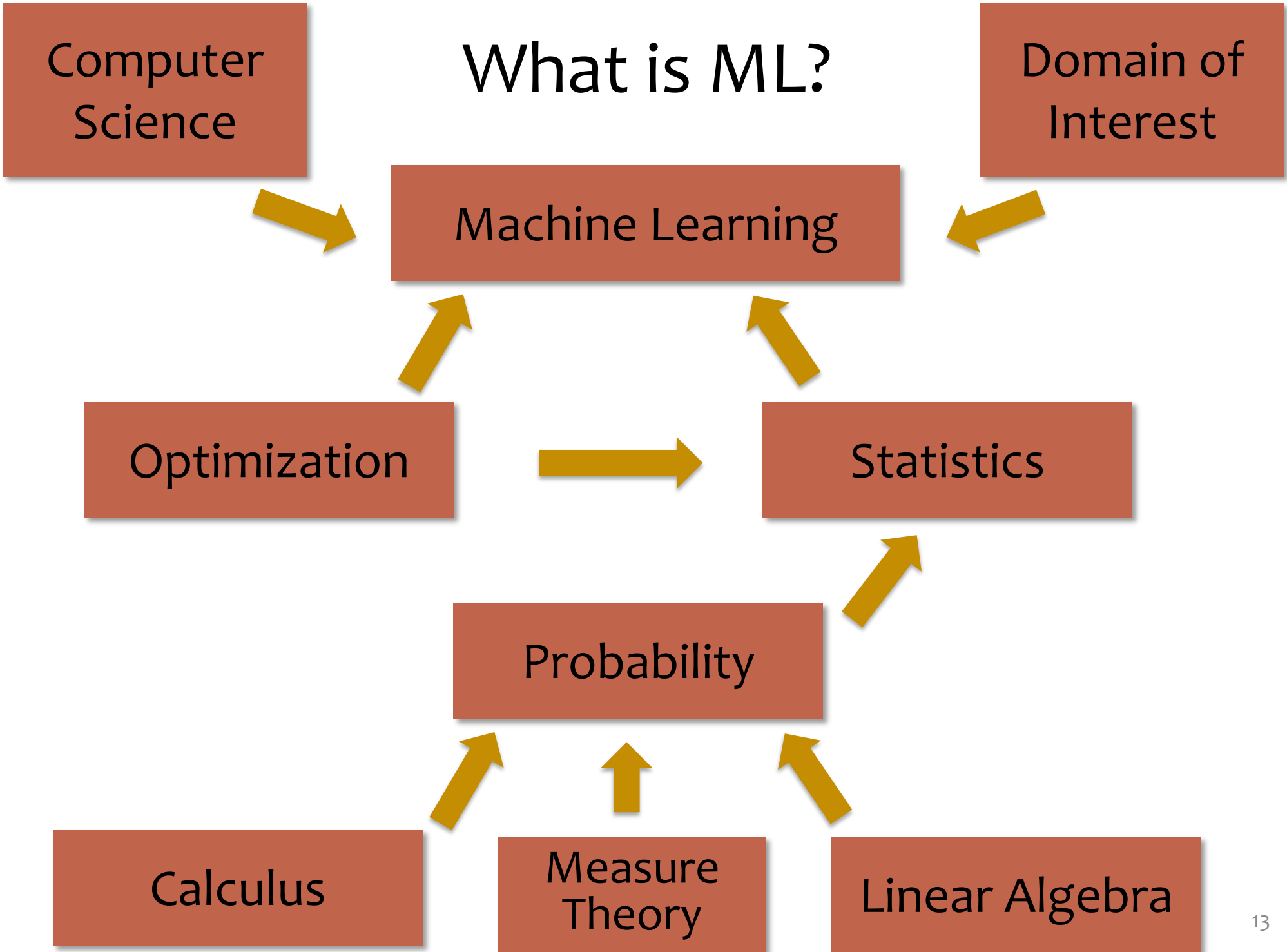
Statistics

Probability

Computer Science

Optimization





# What is ML?

## Speech Recognition

### 1. Learning to recognize spoken words

"...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..."



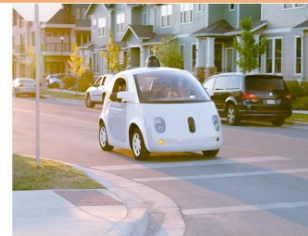
(Mitchell, 1997)

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

## Robotics

### 2. Learning to drive an autonomous vehicle

"...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..."



waymo.com

(Mitchell, 1997)



9

## Games / Reasoning

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

"...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..."



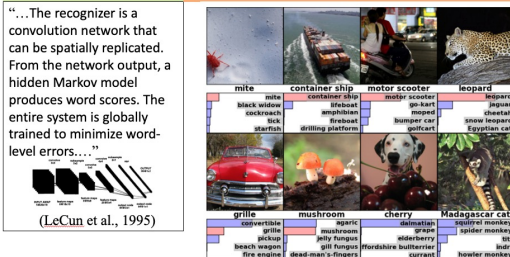
(Mitchell, 1997)

11

## 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/c399862d3bd6b76c8436e92468c45b-Paper.pdf>

## Learning Theory

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

#### Sample Complexity Results

Definition 6.4. The sample complexity of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

Four cases we care about...

Finite  $\mathcal{H}$

Infinite  $\mathcal{H}$

$$R(x) = \sum_{h \in \mathcal{H}} p_h(x) (h(x) - k(x))$$

$$R(x) = \sum_{h \in \mathcal{H}} p_h(x) (h(x) - k(x))$$

PK Lecture

Can we bound  $R(x)$  in terms of  $R(x)$ ?

A: Yes!

PK: What is probably? PK: Answer yields hypothesis  $h$ , which is probably correct.

Def: PK: Correct

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

13

# Speech Recognition

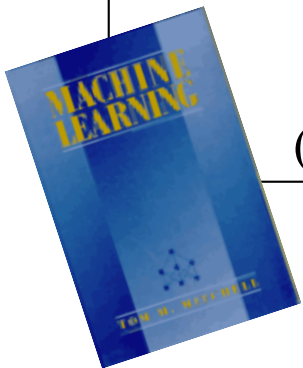
## 1. Learning to recognize spoken words

THEN

“...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)

NOW



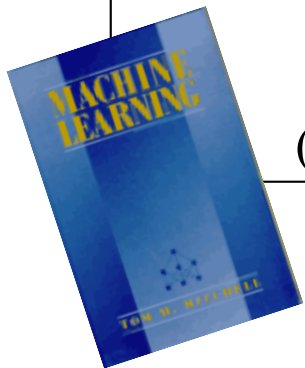
# Robotics

## 2. Learning to drive an autonomous vehicle

THEN

“...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)



NOW



waymo.com

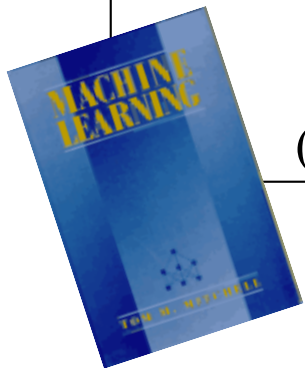
# Robotics

## 2. Learning to drive an autonomous vehicle

THEN

“...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)



NOW



aurora.tech



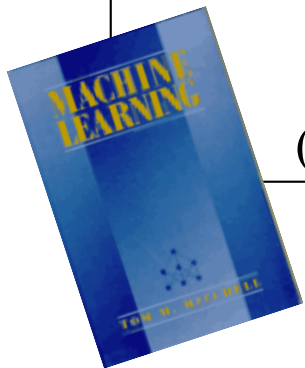
# Robotics

## 2. Learning to drive an autonomous vehicle

THEN

“...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)



NOW



[locomotion.ai](https://locomotion.ai)

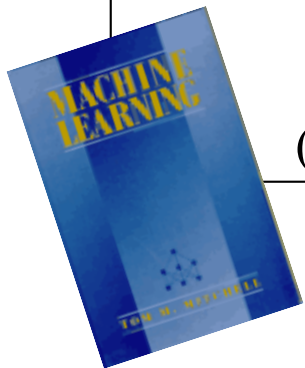
# Games / Reasoning

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

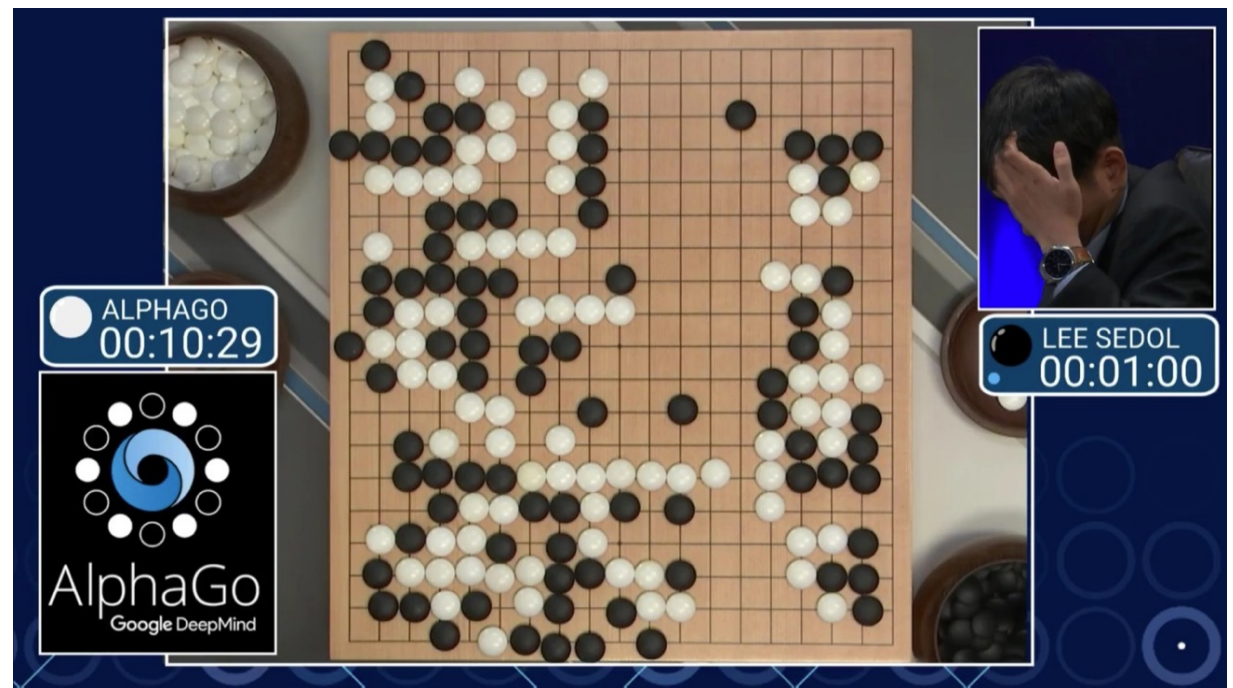
THEN

“...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)



NOW

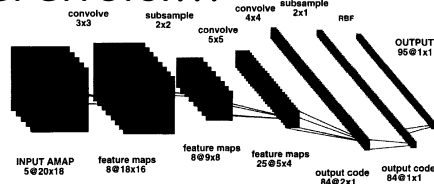


# Computer Vision

## 4. Learning to recognize images









THEN

“...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)

NOW

																							
<b>mite</b>	<b>container ship</b>	<b>motor scooter</b>	<b>leopard</b>																				
<table border="1"> <tbody> <tr><td>mite</td></tr> <tr><td>black widow</td></tr> <tr><td>cockroach</td></tr> <tr><td>tick</td></tr> <tr><td>starfish</td></tr> </tbody> </table>	mite	black widow	cockroach	tick	starfish	<table border="1"> <tbody> <tr><td>container ship</td></tr> <tr><td>lifeboat</td></tr> <tr><td>amphibian</td></tr> <tr><td>fireboat</td></tr> <tr><td>drilling platform</td></tr> </tbody> </table>	container ship	lifeboat	amphibian	fireboat	drilling platform	<table border="1"> <tbody> <tr><td>motor scooter</td></tr> <tr><td>go-kart</td></tr> <tr><td>moped</td></tr> <tr><td>bumper car</td></tr> <tr><td>golfcart</td></tr> </tbody> </table>	motor scooter	go-kart	moped	bumper car	golfcart	<table border="1"> <tbody> <tr><td>leopard</td></tr> <tr><td>jaguar</td></tr> <tr><td>cheetah</td></tr> <tr><td>snow leopard</td></tr> <tr><td>Egyptian cat</td></tr> </tbody> </table>	leopard	jaguar	cheetah	snow leopard	Egyptian cat
mite																							
black widow																							
cockroach																							
tick																							
starfish																							
container ship																							
lifeboat																							
amphibian																							
fireboat																							
drilling platform																							
motor scooter																							
go-kart																							
moped																							
bumper car																							
golfcart																							
leopard																							
jaguar																							
cheetah																							
snow leopard																							
Egyptian cat																							
																							
<b>grille</b>	<b>mushroom</b>	<b>cherry</b>	<b>Madagascar cat</b>																				
<table border="1"> <tbody> <tr><td>convertible</td></tr> <tr><td>grille</td></tr> <tr><td>pickup</td></tr> <tr><td>beach wagon</td></tr> <tr><td>fire engine</td></tr> </tbody> </table>	convertible	grille	pickup	beach wagon	fire engine	<table border="1"> <tbody> <tr><td>agaric</td></tr> <tr><td>mushroom</td></tr> <tr><td>jelly fungus</td></tr> <tr><td>gill fungus</td></tr> <tr><td>dead-man's-fingers</td></tr> </tbody> </table>	agaric	mushroom	jelly fungus	gill fungus	dead-man's-fingers	<table border="1"> <tbody> <tr><td>dalmatian</td></tr> <tr><td>grape</td></tr> <tr><td>elderberry</td></tr> <tr><td>ffordshire bullterrier</td></tr> <tr><td>currant</td></tr> </tbody> </table>	dalmatian	grape	elderberry	ffordshire bullterrier	currant	<table border="1"> <tbody> <tr><td>squirrel monkey</td></tr> <tr><td>spider monkey</td></tr> <tr><td>titi</td></tr> <tr><td>indri</td></tr> <tr><td>howler monkey</td></tr> </tbody> </table>	squirrel monkey	spider monkey	titi	indri	howler monkey
convertible																							
grille																							
pickup																							
beach wagon																							
fire engine																							
agaric																							
mushroom																							
jelly fungus																							
gill fungus																							
dead-man's-fingers																							
dalmatian																							
grape																							
elderberry																							
ffordshire bullterrier																							
currant																							
squirrel monkey																							
spider monkey																							
titi																							
indri																							
howler monkey																							

# Learning Theory

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

### Sample Complexity Results

**Definition 0.1.** The **sample complexity** of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

Four Cases we care about...

	Realizable	Agnostic
Finite $ \mathcal{H} $	$N \geq \frac{1}{\epsilon} [\log( \mathcal{H} ) + \log(\frac{1}{\delta})]$ labeled examples are sufficient so that with probability $(1 - \delta)$ all $h \in \mathcal{H}$ with $R(h) \geq \epsilon$ have $\hat{R}(h) > 0$ .	$N \geq \frac{1}{2\epsilon^2} [\log( \mathcal{H} ) + \log(\frac{2}{\delta})]$ labeled examples are sufficient so that with probability $(1 - \delta)$ for all $h \in \mathcal{H}$ we have that $ R(h) - \hat{R}(h)  < \epsilon$ .
Infinite $ \mathcal{H} $	$N = O(\frac{1}{\epsilon} [\text{VC}(\mathcal{H}) \log(\frac{1}{\epsilon}) + \log(\frac{1}{\delta})])$ labeled examples are sufficient so that with probability $(1 - \delta)$ all $h \in \mathcal{H}$ with $R(h) \geq \epsilon$ have $\hat{R}(h) > 0$ .	$N = O(\frac{1}{\epsilon^2} [\text{VC}(\mathcal{H}) + \log(\frac{1}{\delta})])$ labeled examples are sufficient so that with probability $(1 - \delta)$ for all $h \in \mathcal{H}$ we have that $ R(h) - \hat{R}(h)  \leq \epsilon$ .

### Two Types of Error

① True Error (aka. expected risk) (aka. Generalization Error)

$$R(h) = \mathbb{P}_{x \sim p^*(x)} (c^*(x) \neq h(x)) \quad \leftarrow \text{always unknown.}$$

② Train Error (aka. empirical risk)

$$\begin{aligned} \hat{R}(h) &= \mathbb{P}_{x \sim S} (c^*(x) \neq h(x)) \quad \leftarrow S = \{x^{(1)}, \dots, x^{(N)}\} \\ &= \frac{1}{N} \sum_{i=1}^N \mathbb{I}(c^*(x^{(i)}) \neq h(x^{(i)})) \quad \leftarrow \text{known, computable} \\ &= \frac{1}{N} \sum_{i=1}^N \mathbb{I}(y^{(i)} \neq h(x^{(i)})) \end{aligned}$$

### PAC Learning

Q: Can we bound  $R(h)$  in terms of  $\hat{R}(h)$ ?  
A: Yes!

PAC stands for Probably Approximately Correct

PAC learner yields hypothesis  $h$ , which is approximately correct  $R(h) \approx 0$   
with high probability  $\Pr(R(h) \approx 0) \approx 1$

Def: PAC Criterion

$$\Pr(\forall h, |R(h) - \hat{R}(h)| \leq \epsilon) \geq 1 - \delta$$

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?



# What is Machine Learning?

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

Machine Learning

Statistics

Probability

Computer Science

Optimization

To solve all the problems above and more



# Societal Impacts of ML

*What ethical responsibilities do we have as machine learning experts?*

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

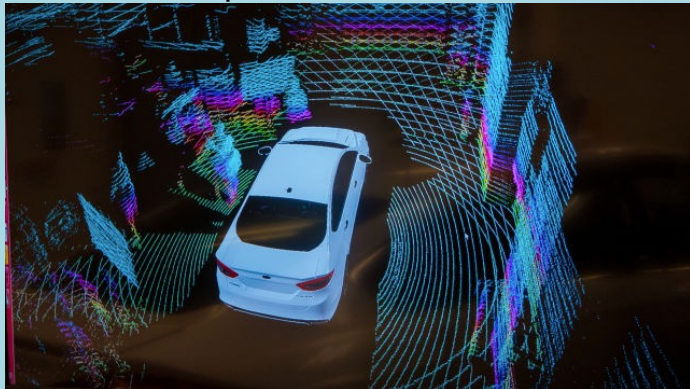
## Answer:

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/HNJUZV>

# 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.



# 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

# Topics

- Foundations
  - Probability
  - MLE, MAP
  - Optimization
- Classifiers
  - KNN
  - Naïve Bayes
  - Logistic Regression
  - Perceptron
  - SVM
- Regression
  - Linear Regression
- Important Concepts
  - Kernels
  - Regularization and Overfitting
  - Experimental Design
- Unsupervised Learning
  - K-means / Lloyd's method
  - PCA
  - EM / GMMs
- Neural Networks
  - Feedforward Neural Nets
  - Basic architectures
  - Backpropagation
  - CNNs, LSTMs
- Graphical Models
  - Bayesian Networks
  - HMMs
  - Learning and Inference
- Learning Theory
  - Statistical Estimation (covered right before midterm)
  - PAC Learning
- Other Learning Paradigms
  - Matrix Factorization
  - Reinforcement Learning
  - Information Theory

# **DEFINING LEARNING PROBLEMS**

# Well-Posed Learning Problems

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

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

  $\rightarrow$  data

## Definition of learning:

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

# Example Learning Problems

## Learning to beat the masters at **chess**

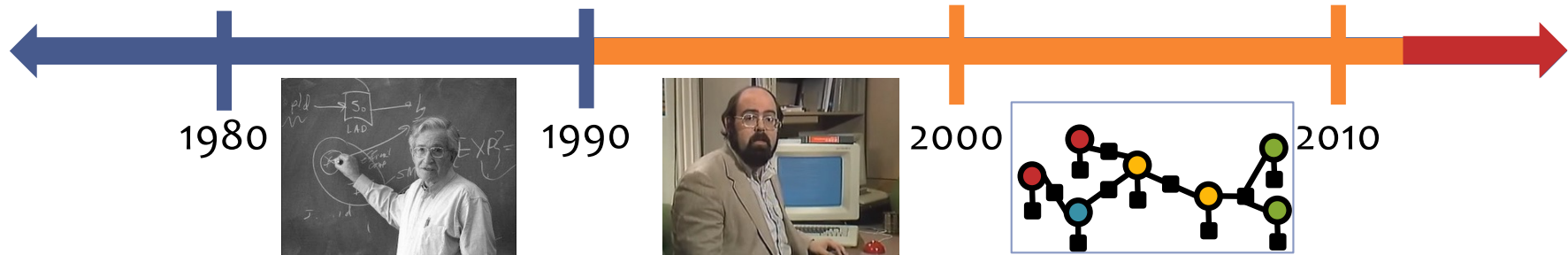
1. Task,  $T$ : *play chess*
2. Performance measure,  $P$ :
  - *ELO*
  - *win rate*
  - *moves per chess*
  - *speed of decisions*
  - *piece taken*
3. Experience,  $E$ :
  - *playing actual games*
    - ↳ *against itself or other algorithms*
    - ↳ *against masters*
  - *past games played by the masters*
  - *strategies from book*

# Example Learning Problems

## 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))`

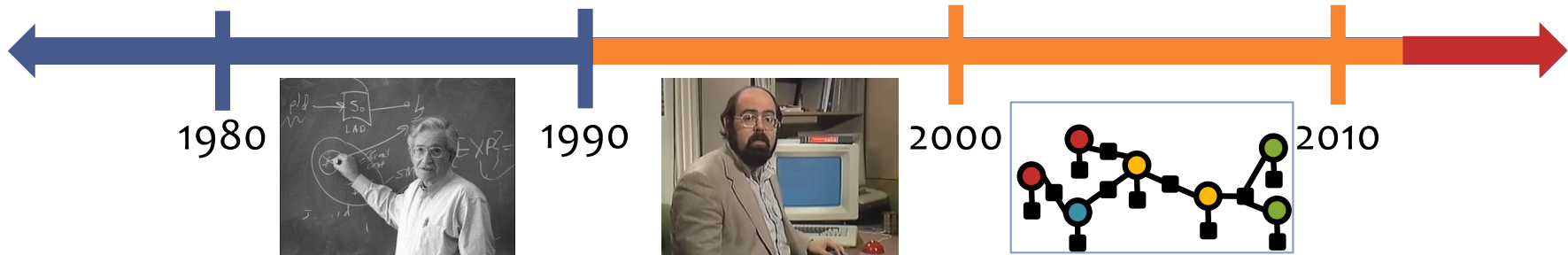
How do I get to Starbucks?

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 #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

I need directions to Starbucks

If: "I need directions to X"  
Then: directions(here, nearest(X))

Starbucks directions

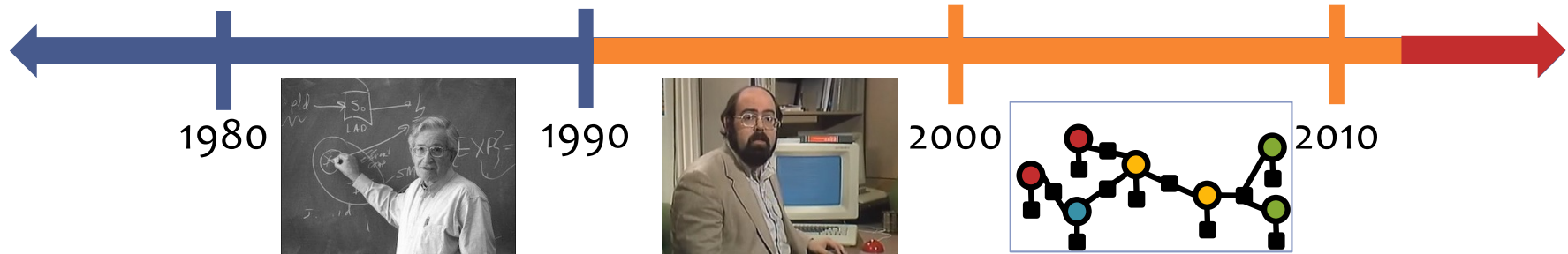
If: "X directions"  
Then: directions(here, nearest(X))

Is there a Starbucks nearby?

If: "Is there an X nearby"  
Then: directions(here, nearest(X))



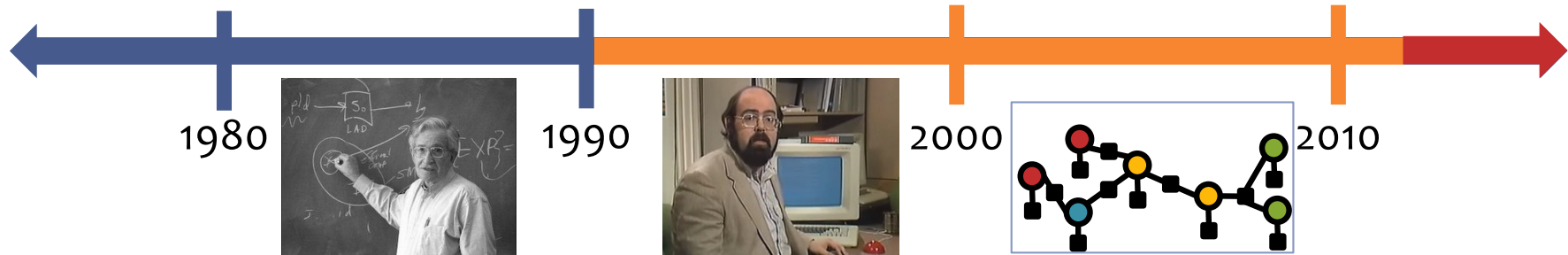
# Capturing the Knowledge of Experts



## 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**

# Problem Formulation

- Often, the same task can be formulated in more than one way:
- Ex: Loan applications
  - creditworthiness/score (regression)
  - probability of default (density estimation)
  - loan decision (classification)

## **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)

# Well-posed Learning Problems

## In-Class Exercise

1. Select a **task**,  $T$
2. Identify **performance measure**,  $P$
3. Identify **experience**,  $E$
4. Report ideas back to rest of class


## Example Tasks

- Identify objects in an image
- Translate from one human language to another
- Recognize speech
- Assess risk (e.g. in loan application)
- Make decisions (e.g. in loan application)
- Assess potential (e.g. in admission decisions)
- Categorize a complex situation (e.g. medical diagnosis)
- Predict outcome (e.g. medical prognosis, stock prices, inflation, temperature)
- Predict events (default on loans, quitting school, war)
- Plan ahead under perfect knowledge (chess)
- Plan ahead under partial knowledge (poker, bridge)

(without any math!)

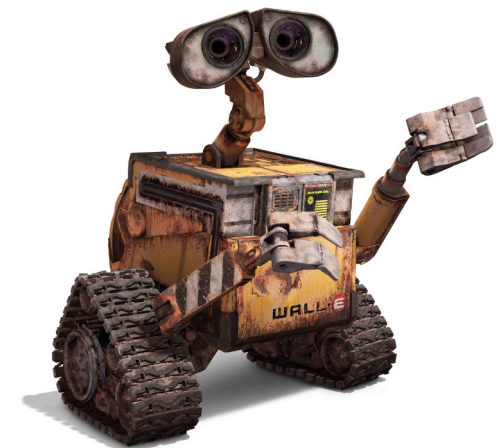
# **SUPERVISED LEARNING**

# Building a Trash Classifier

- Suppose the  ask CMU to build a robot for collecting trash along Pittsburgh's rivers
- You are tasked with building a classifier that detects whether an object is a piece of trash (+) or not a piece of trash (-)
- The robot can detect an object's *color*, *sound*, and *weight*
- You manually annotate the following dataset based on objects you find



trash?	color	sound	weight
+	green	crinkly	high
-	brown	crinkly	low
-	grey	none	high
+	clear	none	low
-	green	none	low



# WARNING!

Like many fields, Machine Learning is riddled with copious amounts of technical jargon!

For many terms we'll define in this class, you'll find four or five different terms in the literature that refer to the same thing.



# Supervised Binary Classification

- Def: an **example** contains a **label** (aka. **class**) and **features** (aka. **point** or **attributes**)
- Def: a **labeled dataset** consists of rows, where each row is an example
- Def: an **unlabeled dataset** only has **features**

## Labeled Dataset:

	label	features		
index	trash?	color	sound	weight
1	-	brown	none	high
2	+	clear	crinkly	low
3	-	brown	none	low

## One example:

label	features		
trash?	color	sound	weight
-	brown	none	high

## Unlabeled Dataset:

	features		
index	color	sound	weight
1	brown	none	high
2	clear	crinkly	low
3	brown	none	low

# Supervised Binary Classification

- Def: an **example** contains a **label** (aka. **class**) and **features** (aka. **point** or **attributes**)
- Def: a **labeled dataset** consists of rows, where each row is an example
- Def: an **unlabeled dataset** only has **features**

## Labeled Dataset:

	label	features		
index	trash?	color	sound	weight
1	-	brown	none	high
2	+	clear	crinkly	low
3	-	brown	none	low

## One example:

label	features		
trash?	color	sound	weight
-	brown	none	high

## Unlabeled Dataset:

	features		
index	color	sound	weight
1	brown	none	high
2	clear	crinkly	low
3	brown	none	low

# Supervised Binary Classification

- Def: an **example** contains a **label** (aka. **class**) and **features** (aka. **point** or **attributes**)
- Def: a **labeled dataset** consists of rows, where each row is an example
- Def: an **unlabeled dataset** only has **features**
- Def: a **training dataset** is a labeled dataset used to **learn** a classifier
- Def: a **classifier** is a function that takes in features and predicts a label
- Def: a **test dataset** is a labeled dataset used to **evaluate** a classifier

Classifier  
features → label

## Training Dataset:

	label	features		
index	trash?	color	sound	weight
1	+	green	crinkly	high
2	-	brown	crinkly	low
3	-	grey	none	high
4	+	clear	none	low
5	-	green	none	low

## Test Dataset:

	label	features		
index	trash?	color	sound	weight
1	-	brown	none	high
2	+	clear	crinkly	low
3	-	brown	none	low

# Supervised Binary Classification

- *Def:* **predictions** are the output of a trained classifier
- *Def:* **error rate** is the proportion of examples on which we predicted the wrong label
- *Def:* a **classifier** is a function that takes in features and predicts a label
- *Def:* a **training dataset** is a labeled dataset used to **learn** a classifier
- *Def:* a **test dataset** is a labeled dataset used to **evaluate** a classifier

Classifier  
features → label

Test Predictions:	
	predictions
index	trash?
1	+
2	+
3	-

(Unlabeled) Test Dataset:			
	features		
index	color	sound	weight
1	brown	none	high
2	clear	crinkly	low
3	brown	none	low

# Supervised Binary Classification

- *Def:* **predictions** are the output of a trained classifier
- *Def:* **error rate** is the proportion of examples on which we predicted the wrong label
- *Def:* a **classifier** is a function that takes in features and predicts a label
- *Def:* a **training dataset** is a labeled dataset used to **learn** a classifier
- *Def:* a **test dataset** is a labeled dataset used to **evaluate** a classifier

## Test Predictions:

predictions

index	trash?	
1	+	✗
2	+	✓
3	-	✓

error rate = 1/3

## (Labeled) Test Dataset:

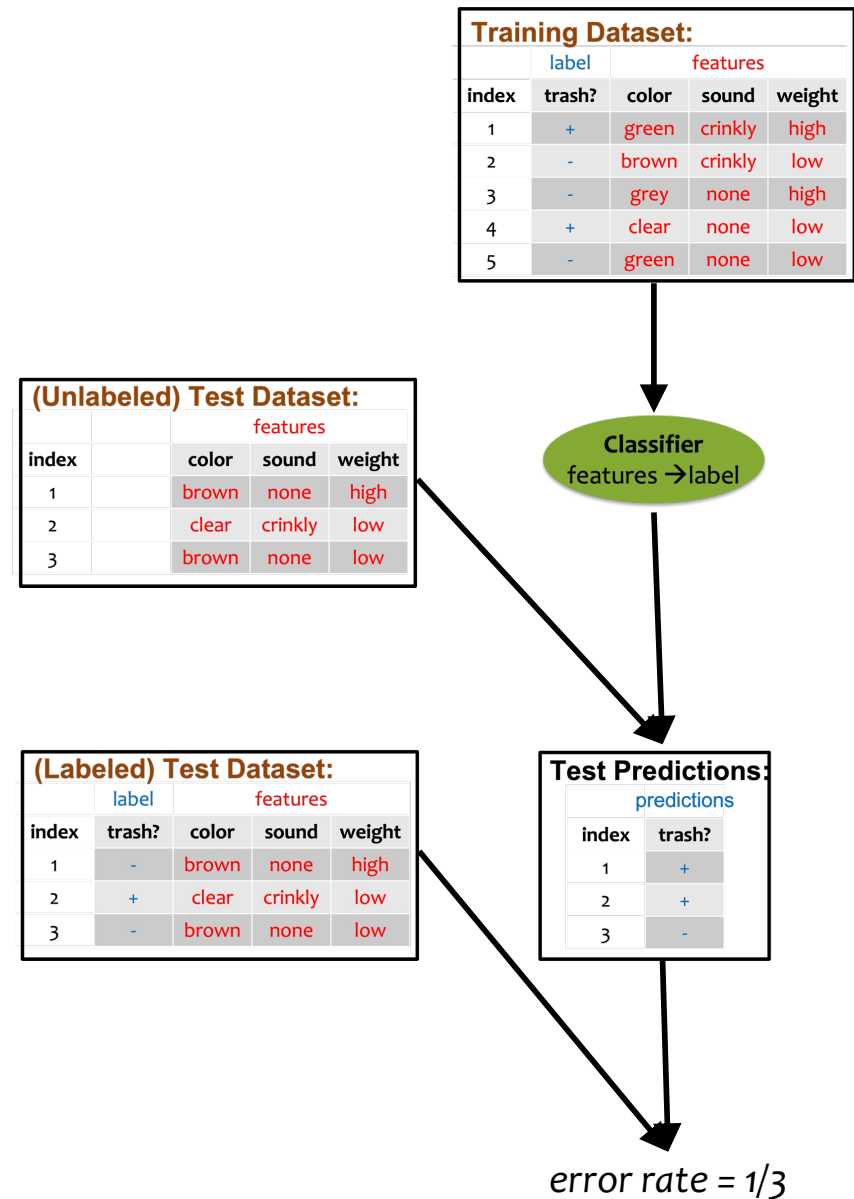
label

features

index	trash?	color	sound	weight
1	-	brown	none	high
2	+	clear	crinkly	low
3	-	brown	none	low

# Supervised Binary Classification

- Step 1: training
  - *Given*: labeled **training dataset**
  - *Goal*: learn a **classifier** from the training dataset
- Step 2: prediction
  - *Given*: unlabeled **test dataset**  
: learned classifier
  - *Goal*: **predict** a label for each instance
- Step 3: evaluation
  - *Given*: **predictions** from Step II  
: labeled **test dataset**
  - *Goal*: compute the **test error rate** (i.e. error rate on the test dataset)



# Supervised Binary Classification

- Step 1: training
  - Given: labeled **training dataset**
  - Goal: learn a **classifier** from the training dataset
- Step 2: prediction
  - Given: unlabeled **test dataset**  
: learned classifier
  - Goal: **predict** a label for each instance
- Step 3: evaluation
  - Given: **predictions** from Step II  
: labeled **test dataset**
  - Goal: compute the **test error rate** (i.e. error rate on the test dataset)

Training Dataset:				
	label	features		
index	trash?	color	sound	weight
1	+	green	crinkly	high
2	-	brown	crinkly	low
3	-	grey	none	high
4	+	clear	none	low
5	-	green	none	low

(Unlabeled) Test Dataset:				
		features		
index		color	sound	weight
1		brown	none	high
2		clear	crinkly	low
3		brown	none	low

Classifier  
features → label

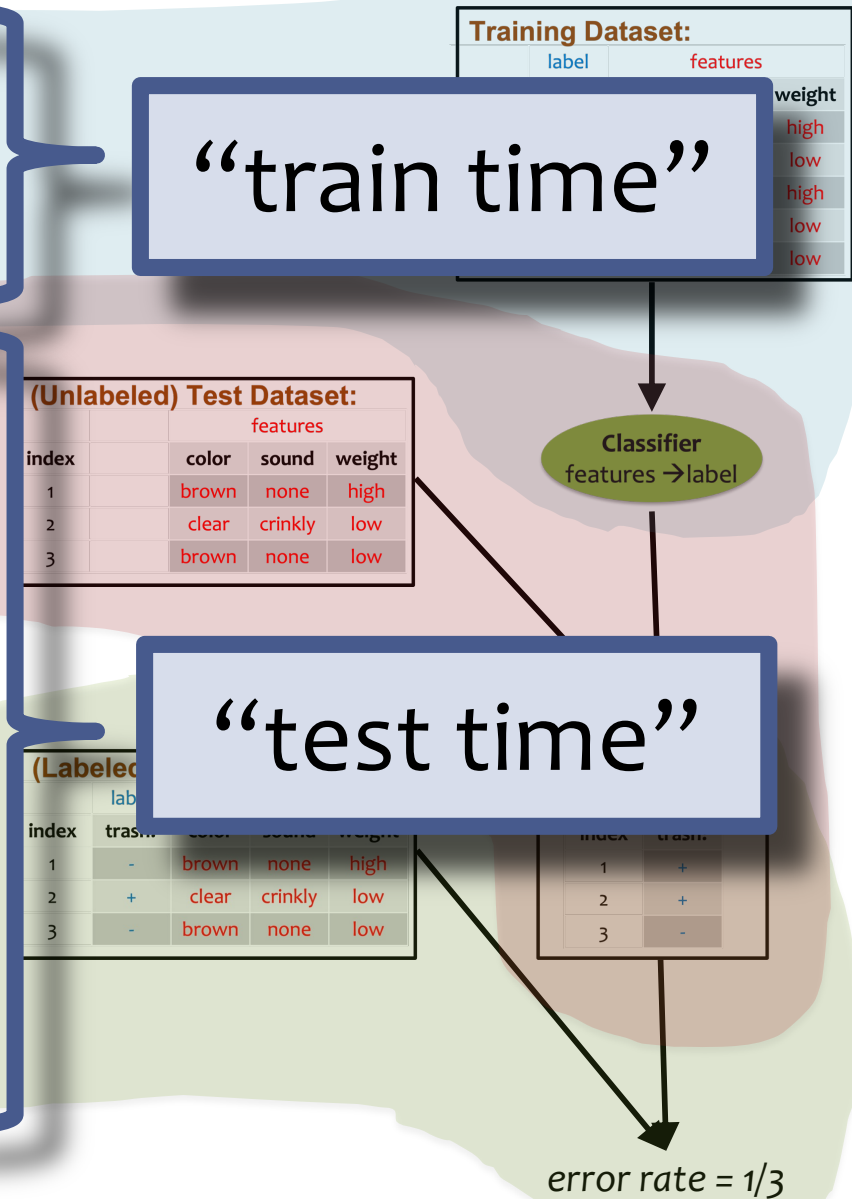
(Labeled) Test Dataset:				
	label	features		
index	trash?	color	sound	weight
1	-	brown	none	high
2	+	clear	crinkly	low
3	-	brown	none	low

Test Predictions:	
	predictions
index	trash?
1	+
2	+
3	-

error rate = 1/3

# Supervised Binary Classification

- Step 1: training
  - Given: labeled **training dataset**
  - Goal: learn a **classifier** from the training dataset
- Step 2: prediction
  - Given: unlabeled **test dataset**  
: learned classifier
  - Goal: **predict** a label for each instance
- Step 3: evaluation
  - Given: **predictions** from Phase II  
: labeled **test dataset**
  - Goal: compute the **test error rate** (i.e. error rate on the test dataset)





# Supervised Binary Classification

- Step 1: training
  - *Given*: labeled **training dataset**
  - *Goal*: learn a **classifier** from the training dataset
- Step 2: prediction
  - *Given*: unlabeled **test data**
  - *Given*: learned classifier
  - *Goal*: **predict** a label for each instance
- Step 3: evaluation
  - *Given*: **predictions** from classifier
  - *Given*: labeled **test data**
  - *Goal*: compute the **test error rate** (i.e. error rate on the test dataset)

	label	features		
index	trash?	color	sound	weight
1	+	green	crinkly	high
2	-	brown	crinkly	low
3	-	grey	none	high
4	+	clear	none	low
5	-	green	none	low

**Key question in  
Machine Learning:**

*How do we learn the  
classifier from data?*

# Random Classifier

The **random classifier** takes in the features and always predicts a random label.

... this is a terrible idea. It completely **ignores the training data!**

Classifier  
features → random!

## Training Dataset:

index	label	features		
	trash?	color	sound	weight
1	+	green	crinkly	high
2	-	brown	crinkly	low
3	-	grey	none	high
4	+	clear	none	low
5	-	green	none	low

## Test Predictions:

index	predictions
	trash?
1	-
2	-
3	+

error rate = 2/3

## Test Dataset:

index	label	features		
	trash?	color	sound	weight
1	-	brown	none	high
2	+	clear	crinkly	low
3	-	brown	none	low

# Random Classifier

The **random classifier** takes in the features and always predicts a random label.

... this is a terrible idea. It completely **ignores the training data!**

Classifier  
features → random!

## Training Dataset:

index	label	features		
	trash?	color	sound	weight
1	+	green	crinkly	high
2	-	brown	crinkly	low
3	-	grey	none	high
4	+	clear	none	low
5	-	green	none	low

## Test Predictions:

index	predictions
	trash?
1	+
2	+
3	-

error rate = 1/3

## Test Dataset:

index	label	features		
	trash?	color	sound	weight
1	-	brown	none	high
2	+	clear	crinkly	low
3	-	brown	none	low

# Random Classifier

The **random classifier** takes in the features and always predicts a random label.

... this is a terrible idea. It completely **ignores the training data!**

Classifier  
features → random!

## Training Dataset:

index	label	features		
	trash?	color	sound	weight
1	+	green	crinkly	high
2	-	brown	crinkly	low
3	-	grey	none	high
4	+	clear	none	low
5	-	green	none	low

## Test Predictions:

index	predictions
	trash?
1	+
2	-
3	+

error rate = 3/3

## Test Dataset:

index	label	features		
	trash?	color	sound	weight
1	-	brown	none	high
2	+	clear	crinkly	low
3	-	brown	none	low

# Majority Vote Classifier

The **majority vote classifier** takes in the features and always predicts the **most common label** in the training dataset.

... this is still a pretty bad idea. It completely **ignores the features!**

**Classifier**  
features → always predict “-”

## Training Dataset:

	label	features		
index	trash?	color	sound	weight
1	+	green	crinkly	high
2	-	brown	crinkly	low
3	-	grey	none	high
4	+	clear	none	low
5	-	green	none	low

## Test Predictions:

	predictions
index	trash?
1	-
2	-
3	-

error rate = 1/3

## Test Dataset:

	label	features		
index	trash?	color	sound	weight
1	-	brown	none	high
2	+	clear	crinkly	low
3	-	brown	none	low

# Majority Vote Classifier

The **majority vote classifier** takes in the features and always predicts the **most common label** in the training dataset.

... this is still a pretty bad idea. It completely **ignores the features!**

**Classifier**  
features → always predict “-”

The majority vote classifier even ignores the features if it's making predictions on the training dataset!

## Train Predictions:

predictions

index	trash?
1	-
2	-
3	-
4	-
5	-

error rate = 2/5

## Training Dataset:

label

features

index	trash?	color	sound	weight
1	+	green	crinkly	high
2	-	brown	crinkly	low
3	-	grey	none	high
4	+	clear	none	low
5	-	green	none	low

# Majority Vote Classifier

- Step 1: training
  - Given: labeled **training dataset**
  - Goal: learn a **classifier** from the training dataset
- Step 2: prediction
  - Given: unlabeled **test dataset**  
: learned classifier
  - Goal: **predict** a label for each instance
- Step 3: evaluation
  - Given: **predictions** from Step II  
: labeled **test dataset**
  - Goal: compute the **test error rate** (i.e. error rate on the test dataset)

Training Dataset:				
	label	features		
index	trash?	color	sound	weight
1	+	green	crinkly	high
2	-	brown	crinkly	low
3	-	grey	none	high
4	+	clear	none	low
5	-	green	none	low

(Unlabeled) Test Dataset:				
		features		
index		color	sound	weight
1		brown	none	high
2		clear	crinkly	low
3		brown	none	low

Classifier  
features → always predict “-”

(Labeled) Test Dataset:				
	label	features		
index	trash?	color	sound	weight
1	-	brown	none	high
2	+	clear	crinkly	low
3	-	brown	none	low

Test Predictions:	
	predictions
index	trash?
1	-
2	-
3	-

error rate = 1/3

# **SYLLABUS HIGHLIGHTS**



# Syllabus Highlights

The syllabus is located on the course webpage:

<http://www.cs.cmu.edu/~mgormley/courses/10601>

or

<http://mlcourse.org>

The **course policies** are **required** reading.

# Syllabus Highlights

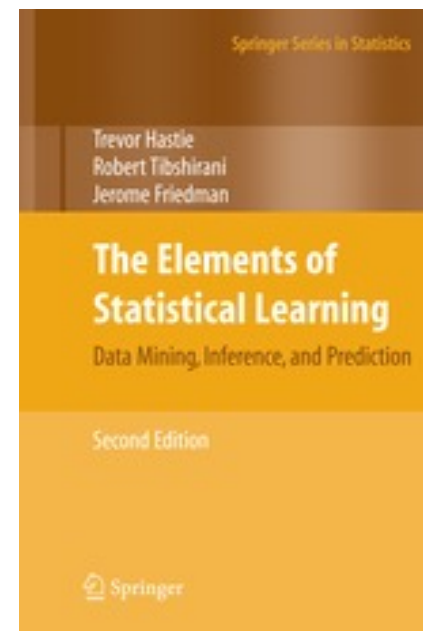
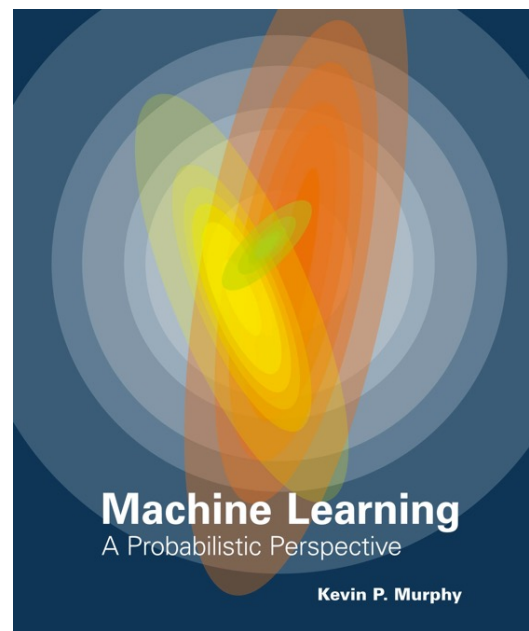
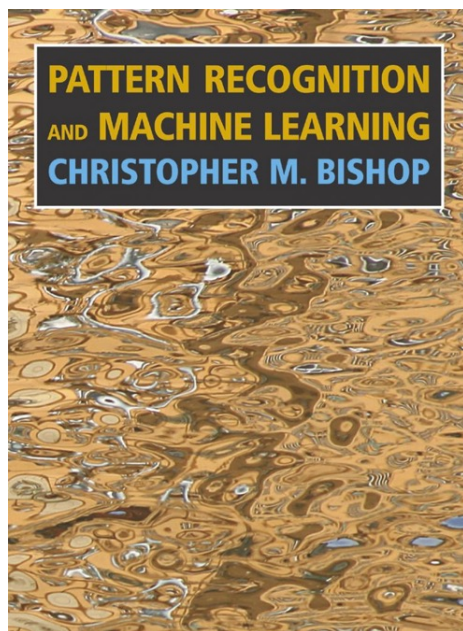
- **Grading:** 50% homework, 15% exam 1, 15% exam 2, 15% exam 3, 5% participation
- **Exam 1:** evening, Thu, Feb. 16
- **Exam 2:** evening, Thu, Mar. 30
- **Exam 3:** final exam week, date TBD by registrar
- **Homework:** 3 written and 6 written + programming (Python)
  - 6 grace days for homework assignments
  - Late submissions: 75% day 1, 50% day 2, 25% day 3
  - No submissions accepted after 3 days w/o extension; HW3, HW6, HW9 only 2 days
  - Extension requests: for emergencies, see syllabus
- **Recitations:** Fridays, same time/place as lecture (optional, interactive sessions)
- **Readings:** required, online PDFs, recommended for after lecture
- **Technologies:** Piazza (discussion), Gradescope (homework), Google Forms (polls)
- **Academic Integrity:**
  - Collaboration encouraged, but must be documented
  - Solutions must always be written independently
  - No re-use of found code / past assignments
  - Severe penalties (e.g. -100%)
- **Office Hours:** posted on Google Calendar on “Office Hours” page

# Lectures

- You should ask lots of questions
  - Interrupting (by raising a hand) to ask your question is strongly encouraged
  - Asking questions later (or in real time) on Piazza is also great
- When I ask a question...
  - I want you to answer
  - Even if you don't answer, think it through as though I'm about to call on you
- Interaction improves learning (both in-class and at my office hours)

# Textbooks

You are not *required* to read a textbook, but it will help immensely!



# Where can I find...?

Date	Lecture	Readings	Announcements
Classification & Regression			
Mon, 1-Feb	Lecture 1 : Course Overview <a href="#">[Slides]</a>	<ul style="list-style-type: none"><li>• <a href="#">10601 Notation Crib Sheet</a>. Matt Gormley (2018).</li><li>• <a href="#">Command Line and File I/O Tutorial</a>. 10601 Course Staff (2020).</li><li>• <a href="#">10601 Learning Objectives</a>. Matt Gormley (2018).</li><li>• <a href="#">Visual Information Theory</a>. Christopher Olah (2015). blog.</li></ul>	
Wed, 3-Feb	Lecture 2 : Decision Trees, Overfitting <a href="#">[Slides]</a>	<ul style="list-style-type: none"><li>• <a href="#">Decision Trees</a>. Hal Daumé III (2017). CIML, Chapter 1.</li></ul>	HW1 out
Fri, 5-Feb	Recitation: HW1 <a href="#">[Handout]</a> <a href="#">[Solutions]</a>		
Mon, 8-Feb	Lecture 3 : Generalizing from examples - the Big Picture <a href="#">[Slides]</a> <a href="#">[Poll]</a>	<ul style="list-style-type: none"><li>• <a href="#">Limits of Learning</a>. Hal Daumé III (2017). CIML, Chapter 2.</li></ul>	
Wed, 10-Feb	Lecture 4 : k-Nearest Neighbors <a href="#">[Slides]</a> <a href="#">[Whiteboard]</a> <a href="#">[Poll]</a>	<ul style="list-style-type: none"><li>• <a href="#">Geometry and Nearest Neighbors</a>. Hal Daumé III (2017). CIML, Chapter 3.</li></ul>	HW1 due HW2 out
Fri, 12-Feb	Recitation: HW2 <a href="#">[Handout]</a> <a href="#">[Solutions]</a>		
Mon, 15-Feb	Lecture 5 : Model Selection <a href="#">[Slides]</a> <a href="#">[Whiteboard]</a> <a href="#">[Poll]</a>		
Wed, 17-Feb	Lecture 6 : Perceptron <a href="#">[Slides]</a> <a href="#">[Whiteboard]</a> <a href="#">[Poll]</a>	<ul style="list-style-type: none"><li>• <a href="#">The Perceptron</a>. Hal Daumé III (2017). CIML, Chapter 4.</li></ul>	HW1 solution session (Thursday)

# Where can I find...?

Home FAQ Syllabus People Schedule **Office Hours** Coursework Previous Links ▾

## Introduction to Machine Learning

10-301 + 10-601, S  
School of Comput  
Carnegie Mellon U

**10-301/601 Office Hours**

Today ◀ ▶ **Aug 29 – Sep 4, 2021** ▼ Print **Week** Month Agenda

	Sun 8/29	Mon 8/30	Tue 8/31	Wed 9/1	Thu 9/2	Fri 9/3	Sat 9/4
9am							
10am		<b>10:10 – 11:30</b> 10-301/601 Section A/C Mellon Institute		<b>10:10 – 11:30</b> 10-301/601 Section A/C Mellon Institute		<b>10:10 – 11:30</b> 10-301/601 Section A/C Mellon Institute	
11am		<b>11:30 - Matt and H</b>		<b>11:30 - Matt and H</b>			
12pm							
1pm							
2pm		<b>1:25p – 2:45p</b> 10-301/601 Section B/D CUC McConomy		<b>1:25p – 2:45p</b> 10-301/601 Section B/D CUC McConomy		<b>1:25p – 2:45p</b> 10-301/601 Section B/D CUC McConomy	
3pm		<b>2:45p - Matt and H</b>		<b>2:45p - Matt and H</b>			

# Where can I find...?

[Home](#)[FAQ](#)[Syllabus](#)[People](#)[Schedule](#)[Office Hours](#)[Coursework](#)[Previous](#)[Links](#)

## Introduction to Machine Learning

10-301 + 10-601, 9  
School of Computer Science  
Carnegie Mellon University

### Assignments

There will be 8 homework assignments during the semester in addition to the exams. The assignments will consist of both theoretical and programming assignments. Assignments will be released via a Piazza announcement explaining where to find the handout, starter code, LaTeX template, etc.

- Homework 1: Background Material (written / programming)  
[Handout](#)
- Homework 2: Decision Trees (written / programming)  
[Handout](#)
- Homework 3: KNN, Perceptron, and Linear Regression (written)  
[Handout](#)
- Mock Exam 1:  
[Handout and Solution](#)
- Homework 4: Logistic Regression (written / programming)  
[Handout](#)
- Homework 5: Neural Networks (written / programming)  
[Handout](#)
- Homework 6: Neural Networks and Reinforcement Learning (written / programming)  
[Handout](#)
- Homework 7: Graphical Models (written / programming)

# In-Class Polls

**Q:** How do these In-Class Polls work?

**A:** Don't worry about it for today. We won't use them until the second week of class, i.e. the third lecture.

Details are on the syllabus.



# PREREQUISITES

# Prerequisites

## What they are:

- Significant programming experience (15-122)
  - Written programs of 100s of lines of code
  - Comfortable learning a new language
- Probability and statistics (36-217, 36-225, etc.)
- Mathematical maturity: discrete mathematics (21-127, 15-151), linear algebra, and calculus

# Prerequisites

## What if you need additional review?

- Consider first taking 10-606/607: Mathematical/Computational Foundations for Machine Learning
- More details here:  
<https://www.cs.cmu.edu/~pvirtue/10606/>

## How to describe 606/607 to a friend

606/607 is...

a **formal** presentation of **mathematics** and **computer science**...

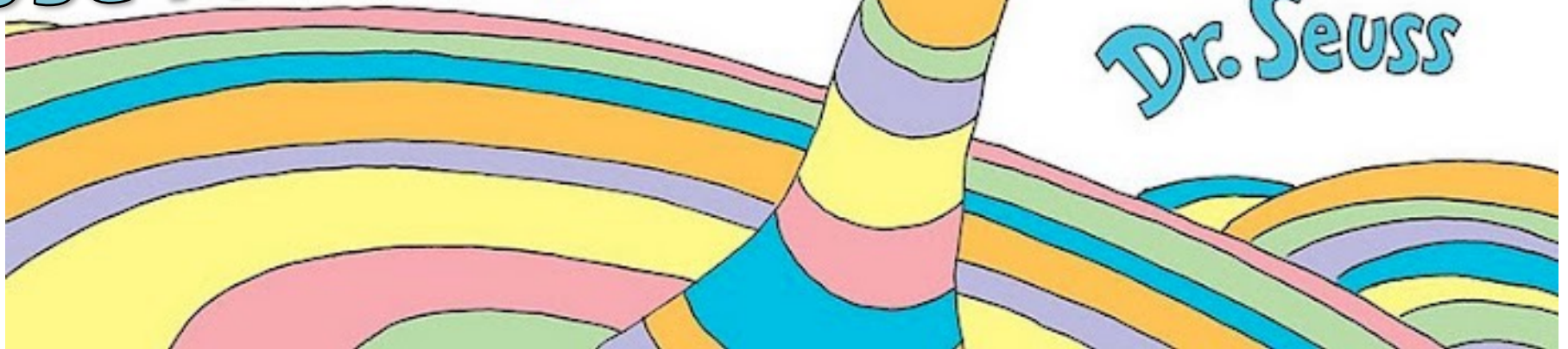
motivated by (carefully chosen) **real-world problems** that arise in **machine learning**...

where the **broader picture** of how those problems arise is treated **somewhat informally**.

Oh, the Places  
You'll  
Use Probability!



By  
Dr. Seuss



# Oh, the Places You'll Use Probability!

## Supervised Classification

- Naïve Bayes

$$p(y|x_1, x_2, \dots, x_n) = \frac{1}{Z} p(y) \prod_{i=1}^n p(x_i|y)$$

- Logistic regression

$$\begin{aligned} P(Y = y|X = x; \boldsymbol{\theta}) &= p(y|x; \boldsymbol{\theta}) \\ &= \frac{\exp(\boldsymbol{\theta}_y \cdot \mathbf{f}(x))}{\sum_{y'} \exp(\boldsymbol{\theta}_{y'} \cdot \mathbf{f}(x))} \end{aligned}$$

Note: This is just motivation – we'll cover these topics later!

# Oh, the Places You'll Use Probability!

## ML Theory

### (Example: Sample Complexity)

- Goal:  $h$  has small error over  $D$ .

$$\text{True error: } err_D(h) = \Pr_{x \sim D} (h(x) \neq c^*(x))$$

How often  $h(x) \neq c^*(x)$  over future instances drawn at random from  $D$

- But, can only measure:

$$\text{Training error: } err_S(h) = \frac{1}{m} \sum_i I(h(x_i) \neq c^*(x_i))$$

How often  $h(x) \neq c^*(x)$  over training instances

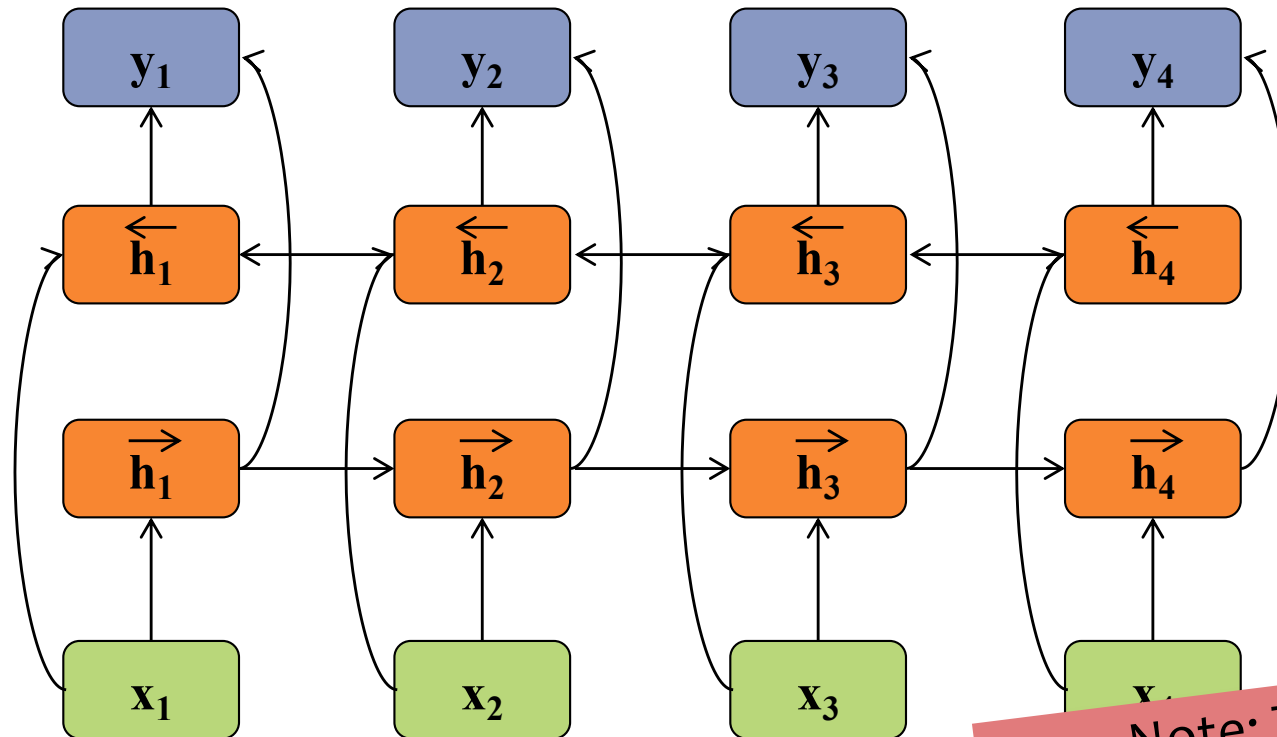
**Sample complexity: bound  $err_D(h)$  in terms of  $err_S(h)$**

Note: This is just motivation – we'll cover these topics later!

# Oh, the Places You'll Use Probability!

## Deep Learning

(Example: Deep Bi-directional RNN)

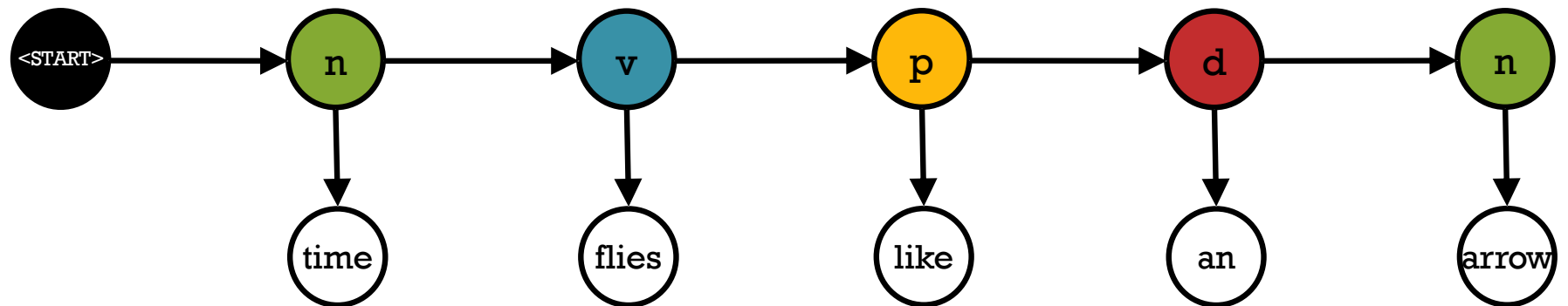


Note: This is just motivation – we'll cover these topics later!

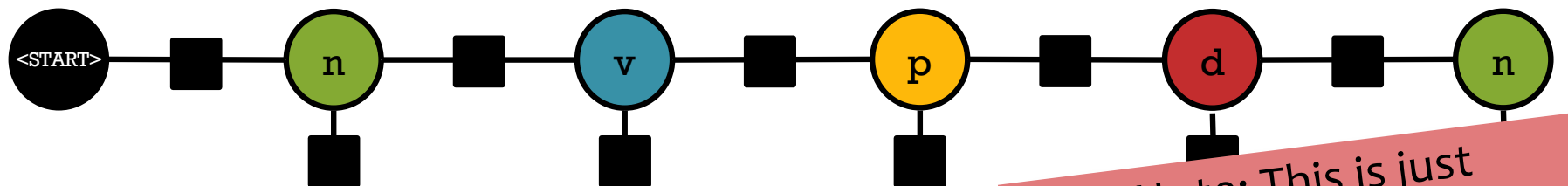
# Oh, the Places You'll Use Probability!

## Graphical Models

- Hidden Markov Model (HMM)



- Conditional Random Field (CRF)



Note: This is just motivation – we'll cover these topics later!



# Prerequisites

## **What if I'm not sure whether I meet them?**

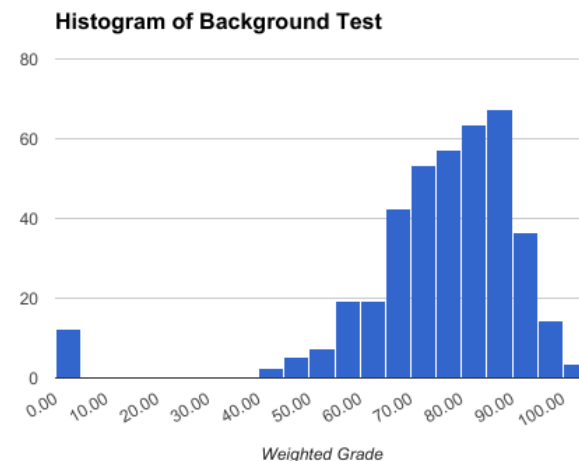
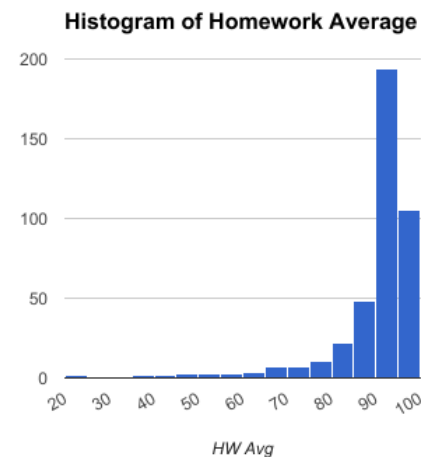
- Don't worry: we're not sure either
- However, we've designed a way to assess your background knowledge so that you know what to study!

# Syllabus Highlights

## Background Test

- **When:** Fri, Jan 20, in-class
- **Where:** this lecture hall
- **What:** prerequisite material (probability, statistics, linear algebra, calculus, geometry, computer science, programming)
- **Why:**
  - an assessment tool to show you what prereq topics to brush up on
  - to save you some time on HW1 if you already know it all
- **How:**
  - $\alpha$  = % of points on Background Test
  - $\beta$  = % of points on Background Exercises
  - Grade:  $\gamma = \alpha + (1-\alpha)\beta$

Your grade on HW1 will give you very little information about which topics to study. Hopefully, the Background Test does.



# Syllabus Highlights

## Background Test

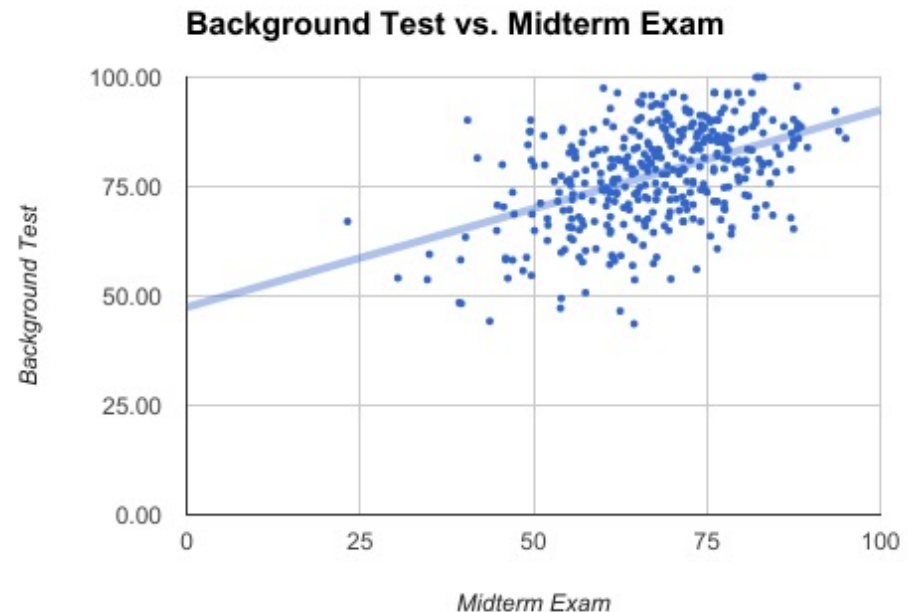
- **When:** Fri, Jan 20, in-class
- **Where:** this lecture hall
- **What:** prerequisite material (probability, statistics, linear algebra, calculus, geometry, computer science, programming)
- **Why:**
  - an assessment tool to show you what prereq topics to brush up on
  - to save you some time on HW1 if you already know it all
- **How:**
  - $\alpha$  = % of points on Background Test
  - $\beta$  = % of points on Background Exercises
  - Grade:  $\gamma = \alpha + (1-\alpha)\beta$

## Correlation between Homework Average and Midterm Exam:

- Pearson: 0.32 (weak - moderate)
- Spearman: 0.25 (weak)

## Correlation between Background Test and Midterm Exam:

- Pearson: 0.46 (moderate)
- Spearman: 0.43 (moderate)



# Reminders

- **Background Test**
  - Fri, Jan 20, in-class
- **Homework 1: Background**
  - Out: Fri, Jan 20
  - Due: Wed, Jan 25 at 11:59pm
  - Two parts:
    1. written part to Gradescope
    2. programming part to Gradescope

# Learning Objectives

*You should be able to...*

1. Formulate a well-posed learning problem for a real-world task by identifying the task, performance measure, and training experience
2. Describe common learning paradigms in terms of the type of data available, when it's available, the form of prediction, and the structure of the output prediction
3. Implement Decision Tree training and prediction (w/simple scoring function)
4. Explain the difference between memorization and generalization [CIML]
5. Identify examples of the ethical responsibilities of an ML expert

# Q&A

(my office hours are right now,  
just outside the lecture hall)