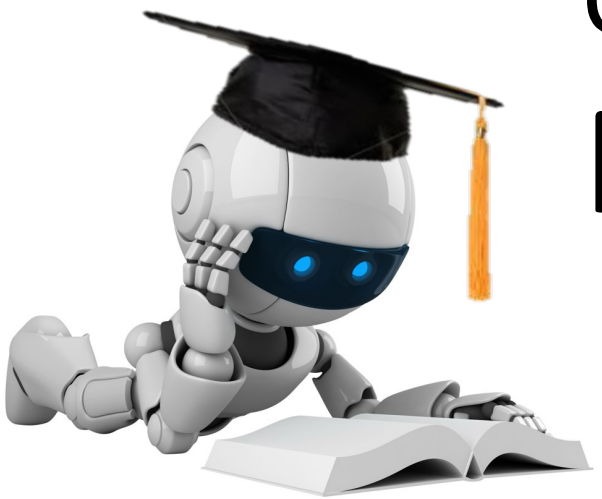


CMSI 5350/EECE 5998

Machine Learning

Dr. Mandy Korpusik

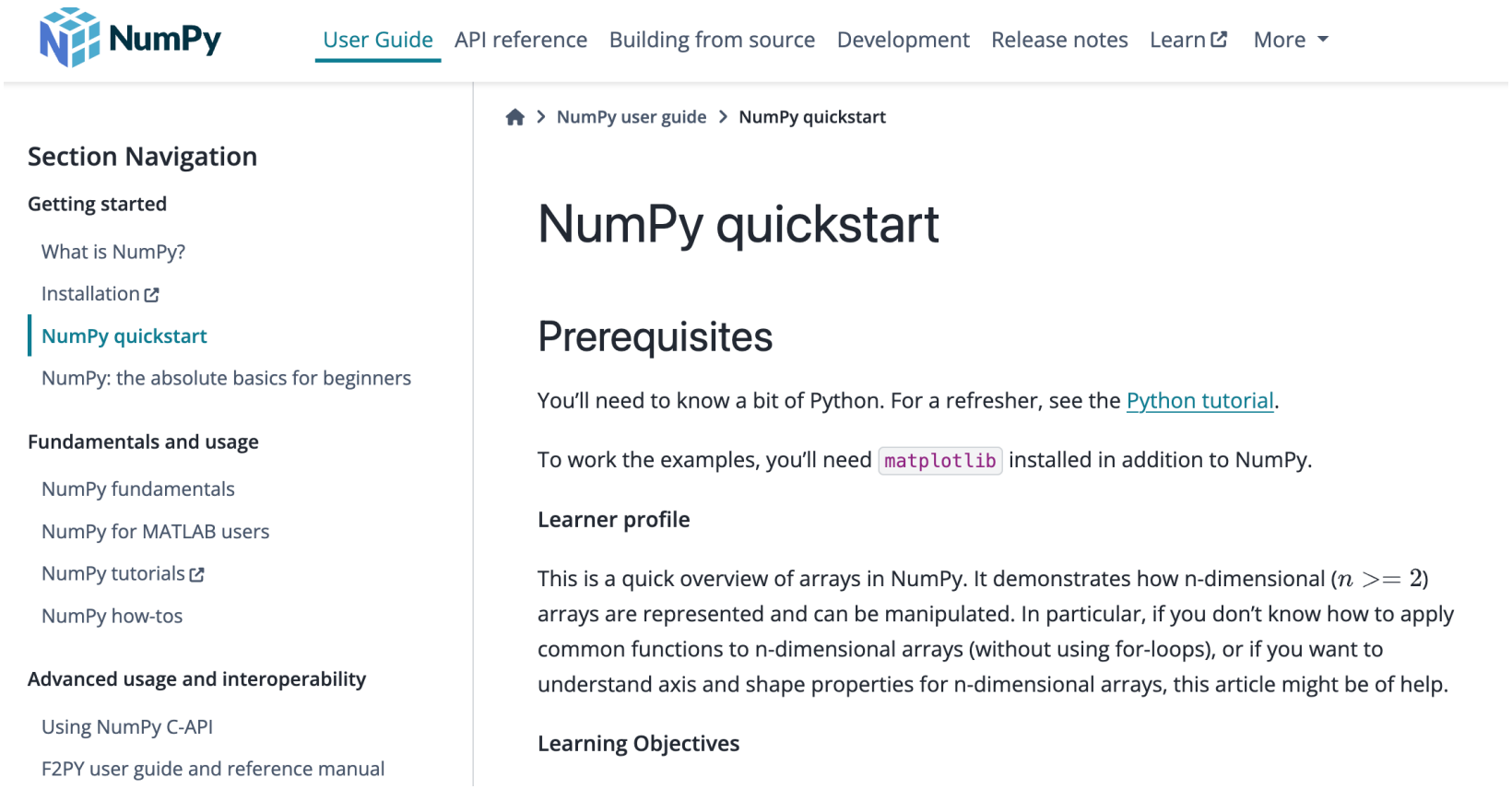
The instructor gratefully acknowledges Alex Wong (Yale), Jessica Wu (HMC), and the many others who made their course materials freely available online.

From last time… NumPy

```
import numpy as np

X = np.zeros((5, 4, 3))
print(X.shape)

X[:, :, 0] = 1 # Reminder: What did this do?
X[..., 1:3] = [2, 3] # And this?
X = np.reshape(X, (-1, 5)) # What's the shape?


def global_mean(x):
    return np.sum(x) / np.prod(x.shape)

global_mean(X) # What do we expect for the mean?
np.mean(X) # Let's check it
```

Resources for NumPy

https://numpy.org/doc/stable/user/quickstart



Scikit-learn library

```
import sklearn
import sklearn.datasets as skdata


housing = skdata.fetch_california_housing()
x = housing.data
y = housing.target
feat_names = housing.feature_names
```

Visualizing data

Let's plot our data using the matplotlib library!

```
from matplotlib import pyplot as plt
```

Let's start with how income varies with price.

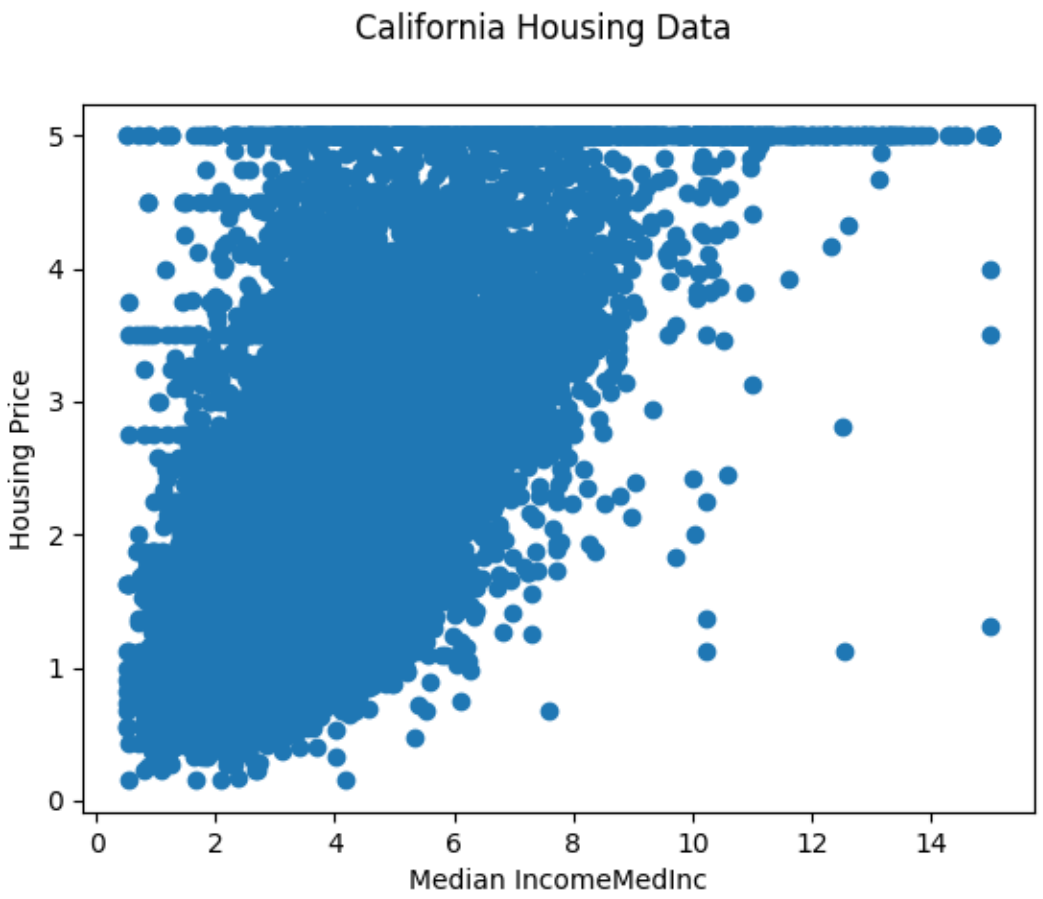
```
income = x[:, 0] # MedInc is the 1st dimension

fig = plt.figure()

ax = fig.add_subplot(1, 1, 1) # (row, col, idx)
ax.scatter(income, y)

plt.show(block=True)
```

Slide credit: Alex Wong

Visualizing data

Let's add labels to the axes:

```
fig = plt.figure()

ax = fig.add_subp
ax.scatter(income

fig.suptitle('Cal
ax.set_ylabel('Hou
ax.set_xlabel('Mec

plt.show(block=Tru
```

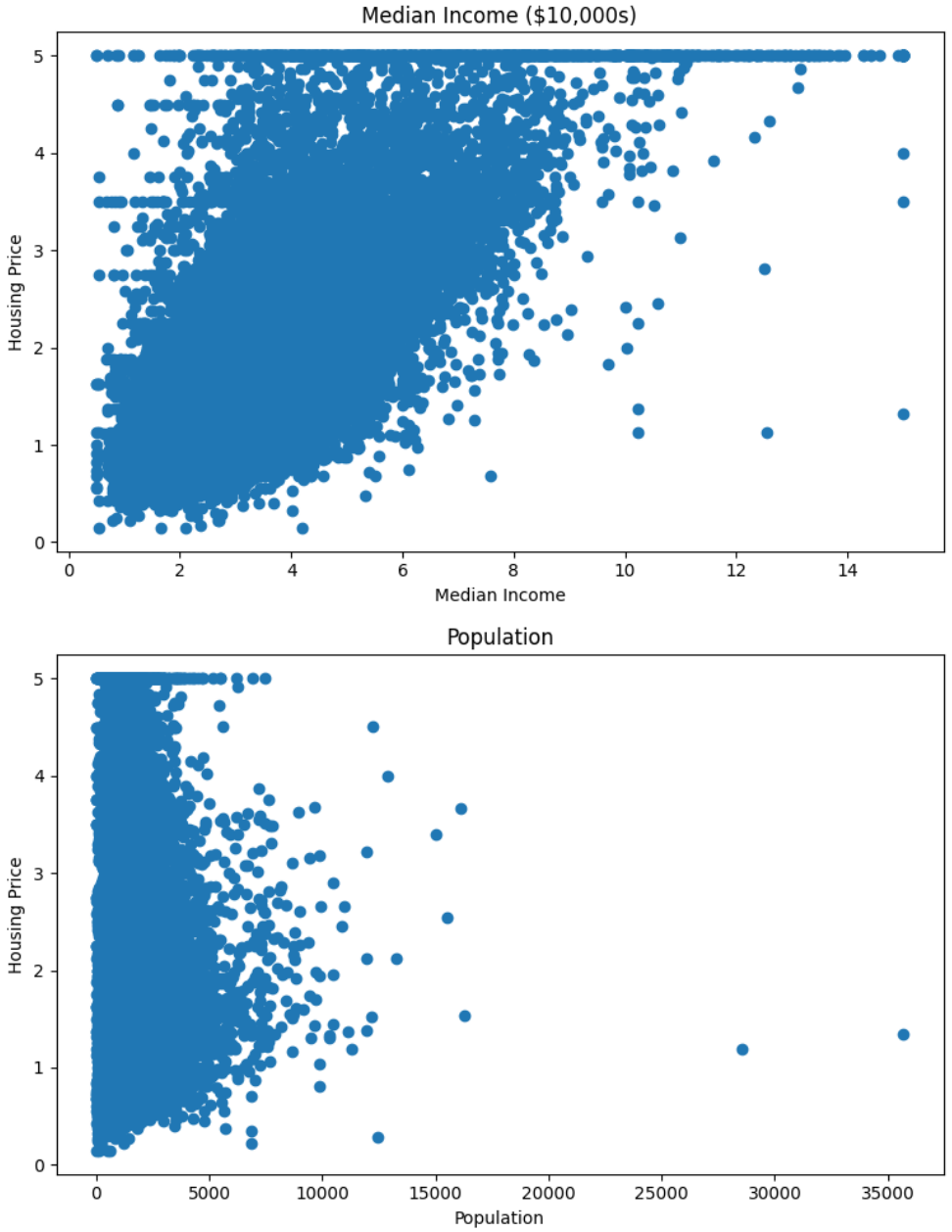


Slide credit: Alex Wong

Visualizing data

Let's see population vs

```
fig = plt.figure()
fig.suptitle('California

ax1 = fig.add_subplot(2,
ax1.set_title('Median Inc
ax1.set_ylabel('Housing P
ax1.set_xlabel('Median In
ax1.scatter(income, y)

population = x[:, 4]
ax2 = fig.add_subplot(2,
ax2.set_title('Population
ax2.set_ylabel('Housing P
ax2.set_xlabel(feat_names
ax2.scatter(population, y

plt.show(block=True)
```



Visualizing data

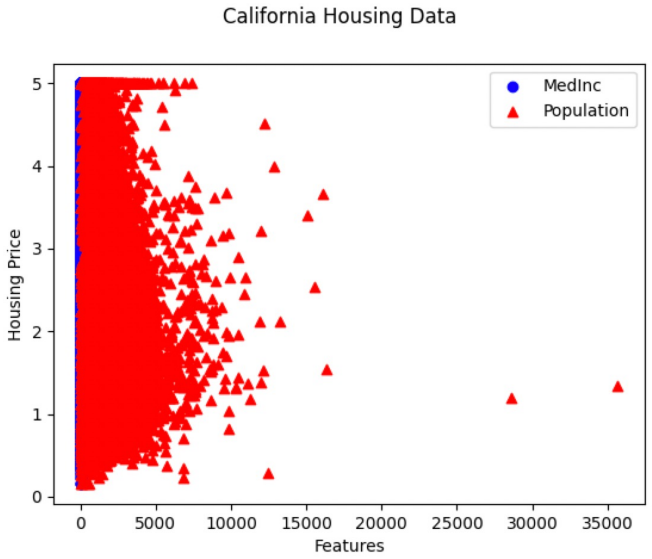
Let's try overlaying the two sets of features:

```
fig = plt.figure()
fig.suptitle('California Housing Data')

ax = fig.add_subplot(1, 1, 1)
ax.set_ylabel('Housing Price')
ax.set_xlabel('Features')

obs = (income, population)
ys = (y, y)
clr = ('blue', 'red')
marker = ('o', '^')
lbl = (feat_names[0], feat_names[4])

for o, yi, c, l, m in zip(obs, ys, clr, lbl, marker):
  ax.scatter(o, yi, c=c, label=l, marker=m)
  ax.legend(loc='upper right')

plt.show(block=True)
```

Slide credit: Alex Wong

Visualizing data

The scales are too different… ☹




```
print(np.min(income), np.max(income))
# MedInc: (0.4999, 15.0001)

print(np.min(population), np.max(population))
# Population: (3.0, 35682.0)
```

Q: How can we get them on the same scale?

Tuesday's learning objectives

Students will be able to:

- ✓ Explore datasets in scikit-learn (sklearn) library
- ✓ Plot data with the matplotlib library
- Implement min-max and standard normalization
- Visualize data in 3D

1) Min-max normalization

Normalize using min and max values:

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

```
def min_max_norm(x):
    return (x - np.min(x)) / (np.max(x) - np.min(x))

# Let's apply it to income and population!
income_minmax = min_max_norm(income)
populn_minmax = min_max_norm(population)

print(np.min(income_minmax),np.max(income_minmax))
print(np.min(populn_minmax),np.max(populn_minmax))
```

1) Min-max normalization

Now let's overlay the two feature sets again:

```
fig = plt.figure()
fig.suptitle('California Housing Data')

ax = fig.add_subplot(1, 1, 1)
ax.set_ylabel('Housing Price')
ax.set_xlabel('Min-Max Norm Features')

obs = (income_minmax, populn_minmax)
for o, yi, c, l, m in zip(obs, ys, clr, lbl, marker):
  ax.scatter(o, yi, c=c, label=l, marker=m)
  ax.legend(loc='upper right')

plt.show(block=True)
```

Slide credit: Alex Wong

1) Min-max normalization

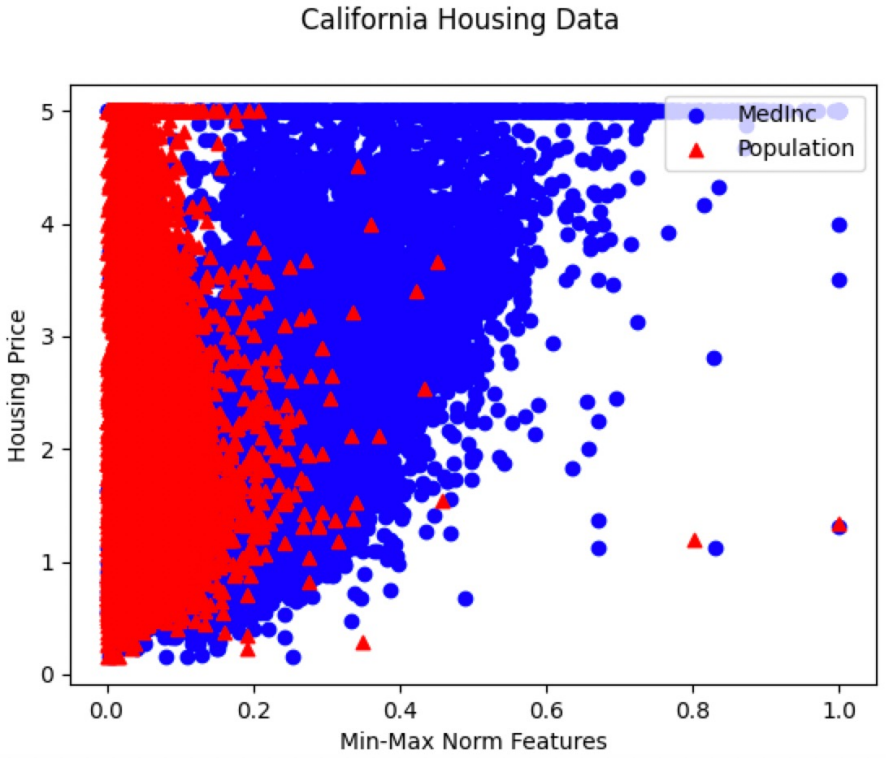
Now we can see trends in the data:

- Prices increase as median income increases
- Not a strong relationship between price and population

Q: Any issues with min-max?

Dividing population by max squishes observations btw. 0 and 0.3, with a few outliers…

Q: Is there a way to normalize data so it's better distributed?



Slide credit: Alex Wong

2) Standard normalization

Results in a mean of 0 and standard deviation of 1:

$$z = \frac{x - \mu}{\sigma}$$

```
def standard_norm(x):
  return (x - np.mean(x)) / (np.std(x))

# Let's apply it to income and population!
income_std = standard_norm(income)
populn_std = standard_norm(population)

print(np.min(income_std), np.max(income_std))
print(np.min(populn_std), np.max(populn_std))
```

Slide credit: Alex Wong

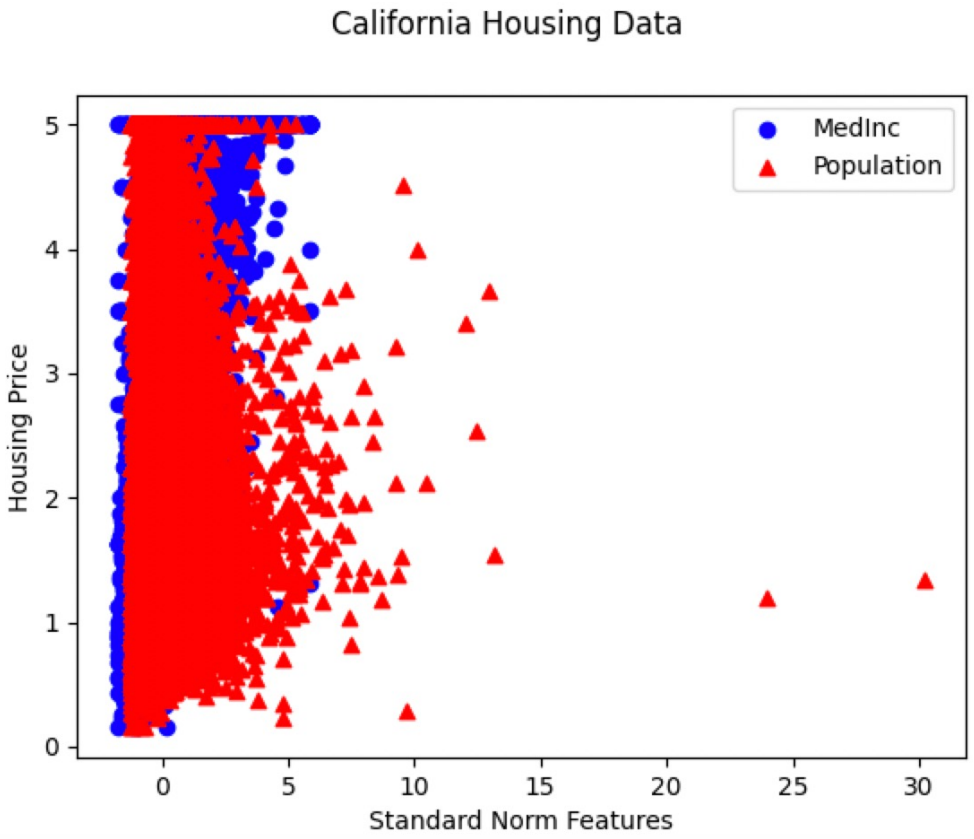
2) Exercise: Standard normalization

Overlay the two features again. What do you notice?

```
fig = plt.figure()
fig.suptitle('California Housing Data')

ax = fig.add_subplot(1, 1, 1)
ax.set_ylabel('Housing Price')
ax.set_xlabel('Standard Norm Features')

obs = (income_std, popuIn_std)
for o, yi, c, l, m in zip(obs, ys, clr, lbl, marker):
  ax.scatter(o, yi, c=c, label=l, marker=m)
  ax.legend(loc='upper right')

plt.show(block=True)
```

Slide credit: Alex Wong

2) Standard normalization



- Min-max has a bounded range, but this will squish the data.
- Standard norm is not bounded, but retains info about outliers.

Slide credit: Alex Wong

Visualizing in 3D

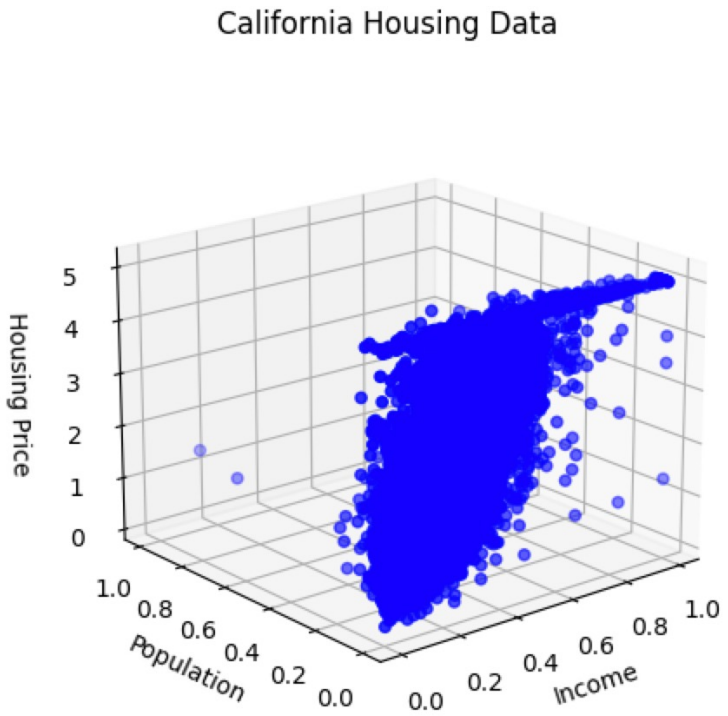
Let's visualize income and population in 3D:

```
from mpl_toolkits.mplot3d import Axes3D

fig = plt.figure()
fig.suptitle('California Housing Data')

ax = fig.add_subplot(1, 1, 1, projection='3d')
ax.set_zlabel('Housing Price')
ax.set_xlabel('Income')
ax.set_ylabel('Population')
ax.scatter(income_minmax, populn_minmax, y, c='blue',
marker='o')

plt.show(block=True)
```

Slide credit: Alex Wong

Visualizing in 3D

Can now see how the data is distributed together—how both features vary together with respect to housing price.



Slide credit: Alex Wong

Plotly

```
pip install plotly pandas
```

```
import plotly.express as px
import pandas as pd



data = pd.DataFrame(x, columns=feat_names)
print(data)

fig = px.scatter_geo(data, lat='Latitude',
    lon='Longitude', scope='usa', color='MedInc')

# Try assigning different features to color!
fig.show()
```

Exercise

Explore the UC Irvine Breast Cancer dataset:

```
import sklearn.datasets as skdata

breast_cancer_data = skdata.load_breast_cancer()

# Extract the data, feature names, and target
# Pick a couple features to plot against the target!
# Try normalizing the data
```

Target is binary (1 or 0) for benign or malignant.

Today's learning objectives

Students will be able to:

- Define Machine Learning.
- List three types of Machine Learning algorithms.
- Give examples of supervised learning tasks.
- Name an unsupervised learning technique.
- Explain how reinforcement learning works.
- Describe the learning task with math.

What is learning?

How do humans learn?

Observations → Learning → Action

How do machines learn?

Data (observations) → Learning → Decision (action)

Reminder: Machine Learning is…

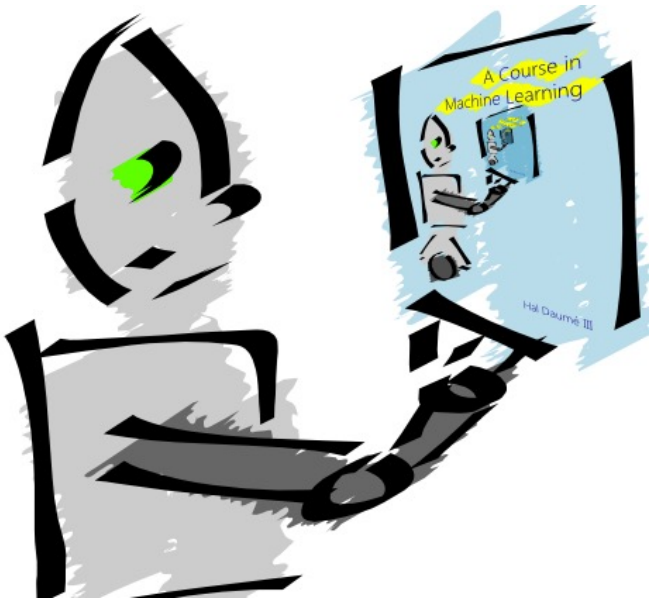
Machine learning, a branch of artificial intelligence, concerns the construction and study of systems that can learn from data.

Slide credit: David Kauchak

Machine Learning is...

Machine learning is about **predicting the future** based on the past.

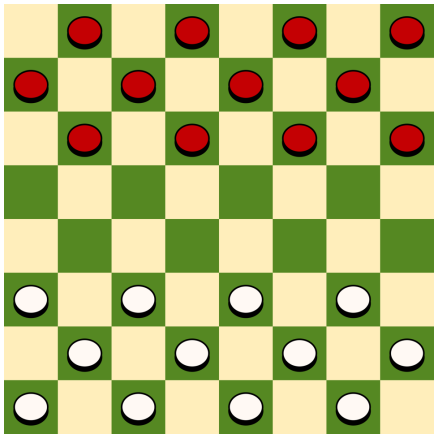
-- Hal Daume III



Slide credit: David Kauchak

Machine Learning is…

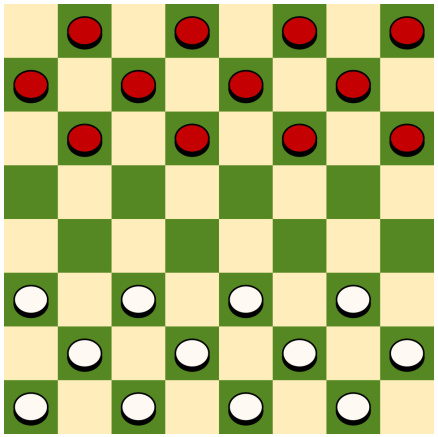
"the field of study that gives computers the ability to learn without being explicitly programmed." -Arthur Samuel (1959)



Photos from Wikipedia

Machine Learning is…

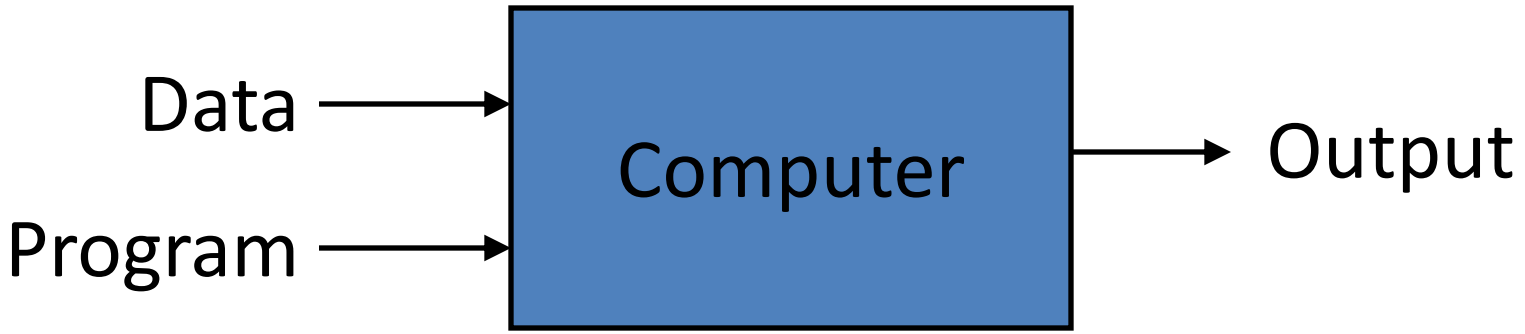
"a computer program is said to learn from experience E with respect to some class of tasks T and performance P, if its performance at tasks in T, as measured by P, improves with experience E." -Tom Mitchell (1998)



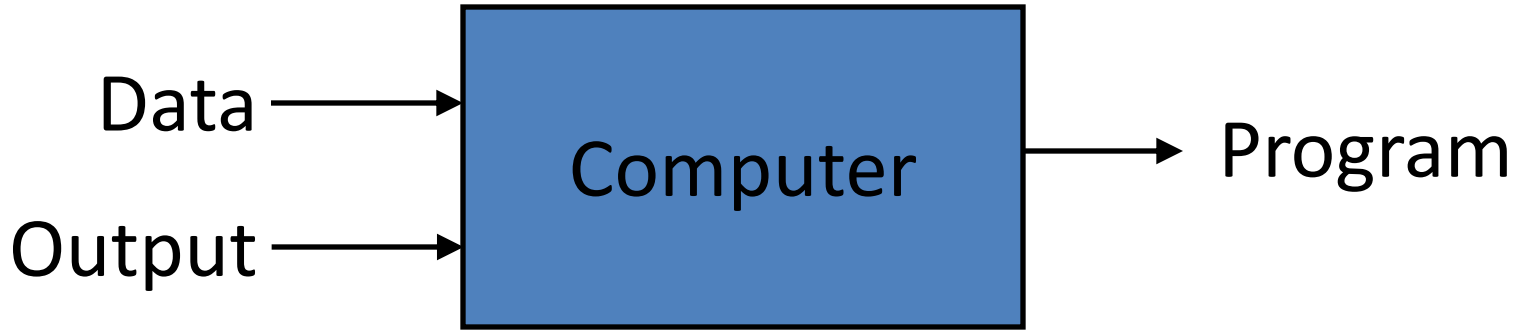
Based on slide by Eric Eaton

Image from Tom Mitchell's homepage

Traditional Programming



Machine Learning



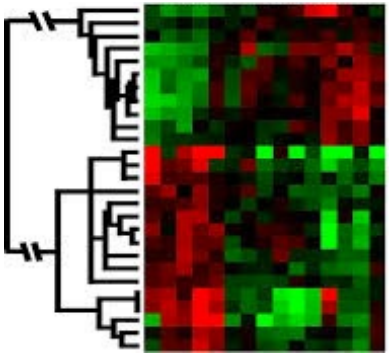
Slide credit: Pedro Domingos

When do we use Machine Learning?

ML is used when:

- Human expertise does not exist
- Humans cannot explain their expertise
- Models must be customized
- Models are based on huge amounts of data



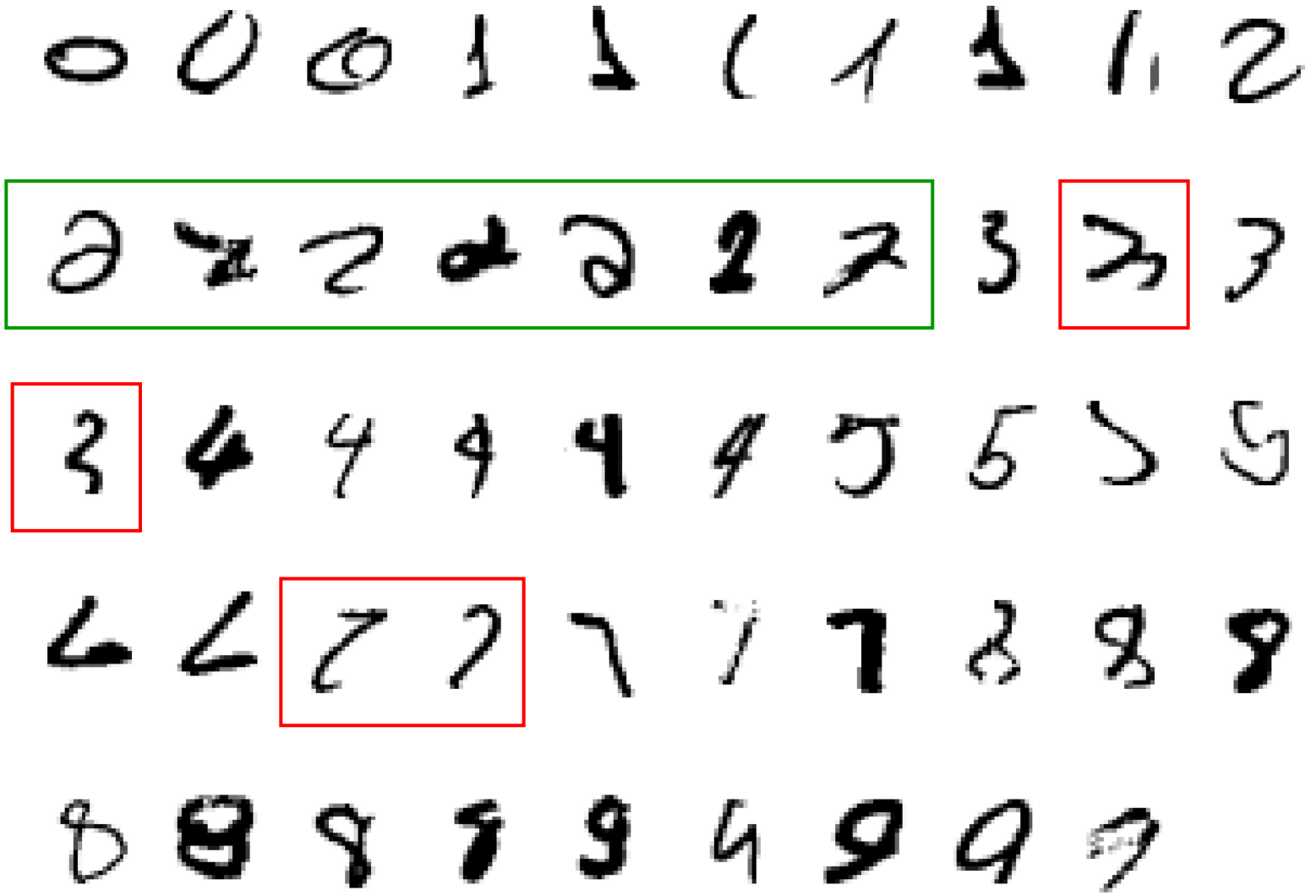



Learning is not always useful

Slide credit: Eric Eaton
Based on slide by E. Alpaydin

A classic example of a task that requires machine learning:

It is very hard to say what makes a 2



Slide credit: Geoffrey Hinton

Some examples of tasks that are best solved by using a learning algorithm

- Learning patterns
  - Predicting passenger survivability on the *Titanic*
  - Recognizing tweets as positive or negative
  - Clustering faces by identity
- State-of-the-art applications
  - Autonomous cars
  - Automatic speech recognition
  - Anomalies in credit card transactions
  - ChatGPT
- [Your favorite area here]

Slide credit: Jessica Wu

ML/AI jobs on the rise

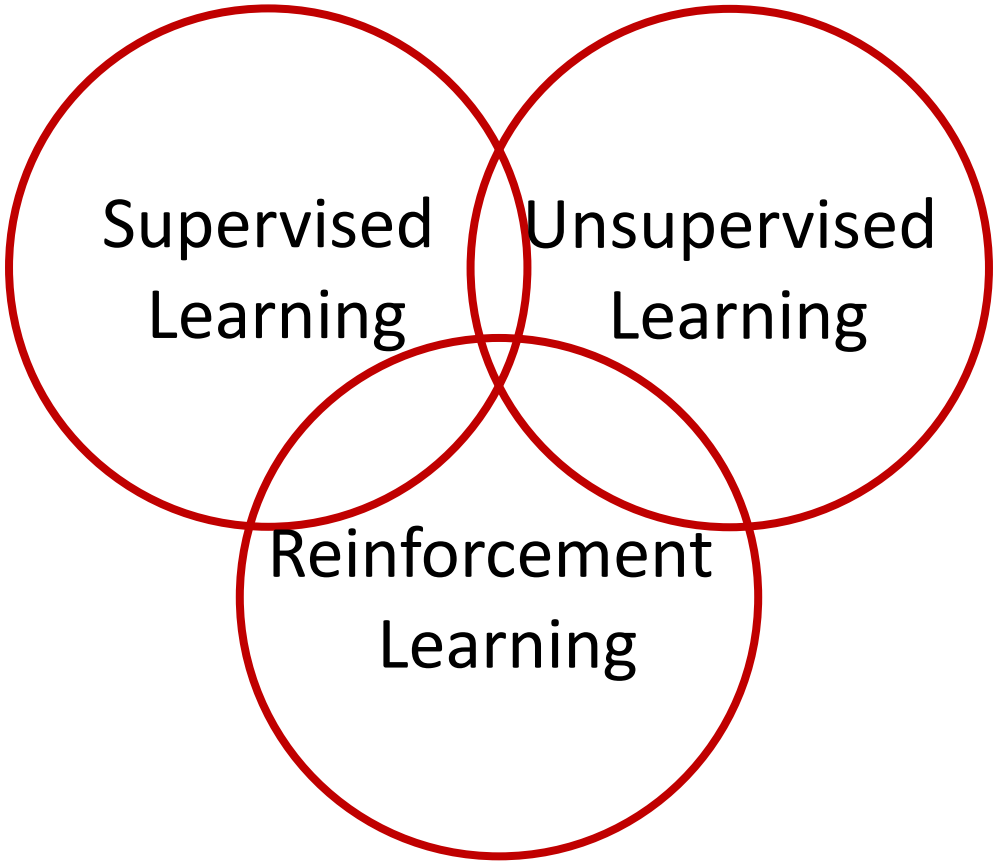
1. Chief Growth Officer
2. Government Program Analyst (**Data Analysis**)
3. Environment Health Safety Manager
4. Director of Revenue Operations
5. Sustainability Analyst (**Data Analysis**)
6. Advanced Practice Provider
7. VP of Diversity and Inclusion
8. **AI Consultant**
9. Recruiter
10. **AI Engineer**

Today's learning objectives

Students will be able to:

- ✓ Define Machine Learning.
- List three types of Machine Learning algorithms.
- Give examples of supervised learning tasks.
- Name an unsupervised learning technique.
- Explain how reinforcement learning works.
- Describe the learning task with math.

Types of Machine Learning

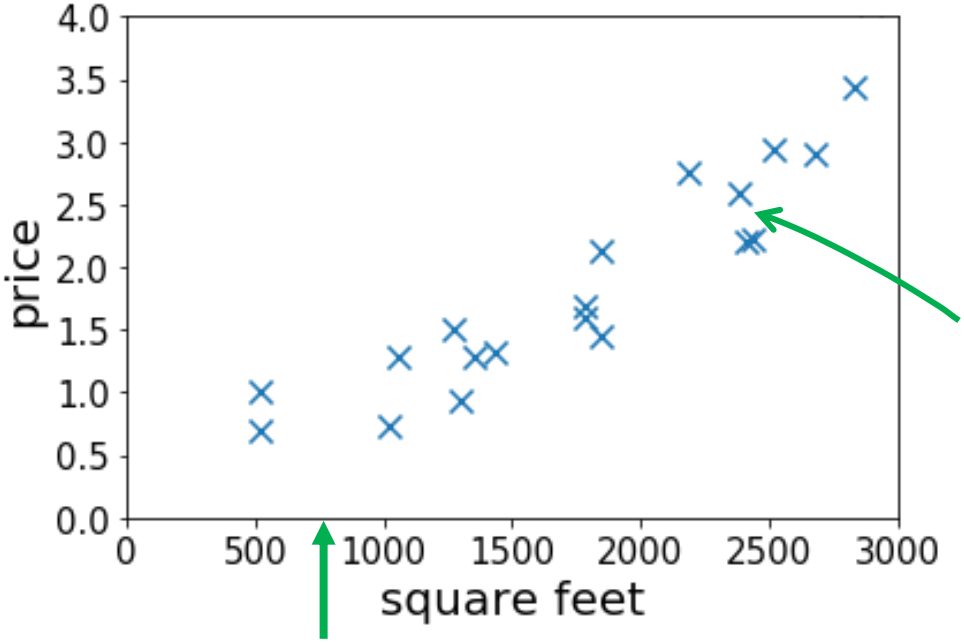


Slide credit: Moses Charikar, Tengyu Ma, Chris Re

Supervised learning: Housing price

Given: a dataset of n samples $\left(x^{(1)}, y^{(1)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)$

Task: if a residence has x sq ft, what is its price y?



Slide credit: Moses Charikar, Tengyu Ma, Chris Re

More features

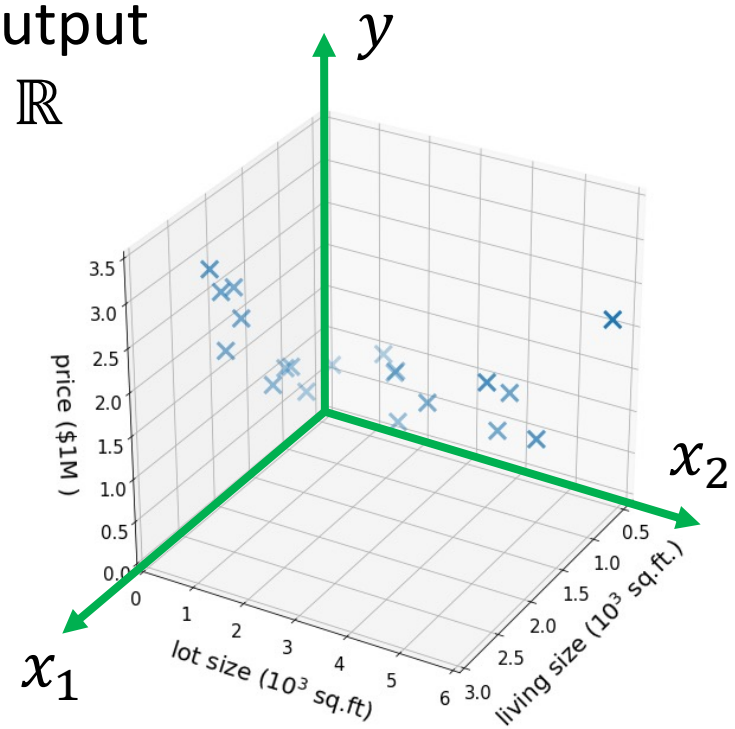
Suppose we also know the lot size

Task: find a function that maps

(size, lot size) → price

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

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



High-dimensional features

$x \in \mathbb{R}^d$ for large d

Ex)

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ \vdots \\ \vdots \\ x_d \end{bmatrix}$$

x_1 --- median income

x_2 --- house age

x_3 --- average rooms

$\vdots$

$\longrightarrow$ y --- price

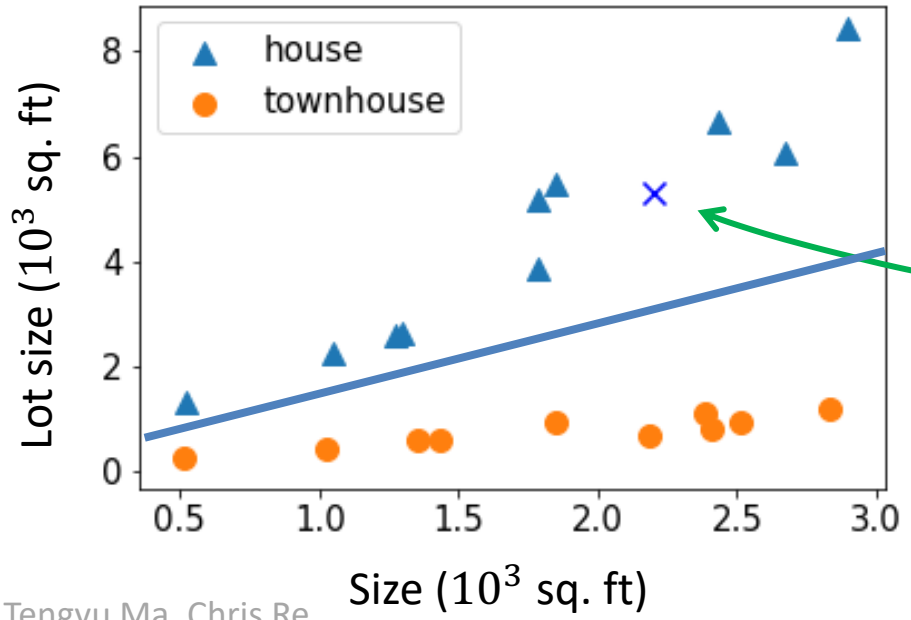
Regression vs. classification

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) → house or townhouse?



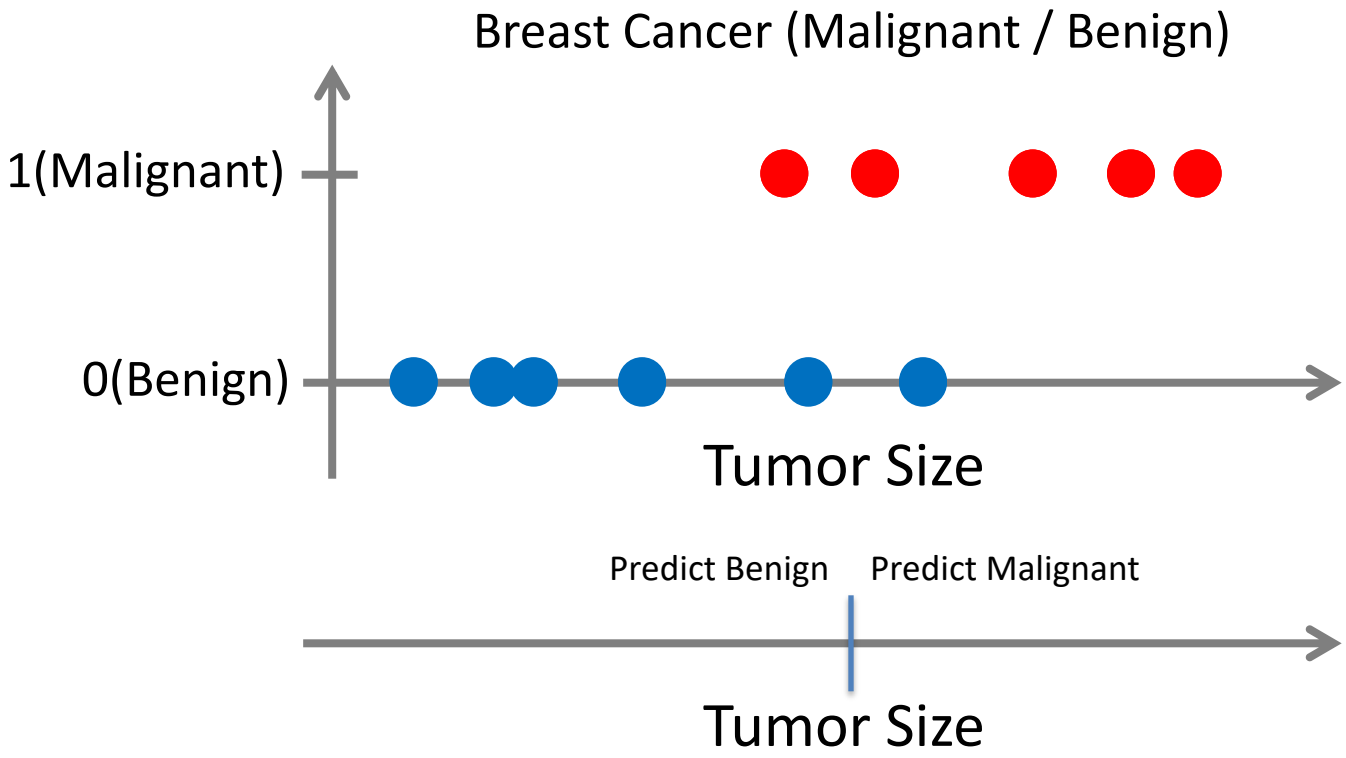
y = house or townhouse?

Slide credit: Moses Charikar, Tengyu Ma, Chris Re

Supervised Learning: Classification

Given: a dataset of n samples $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$

Task: given tumor size x, is it benign or malignant?

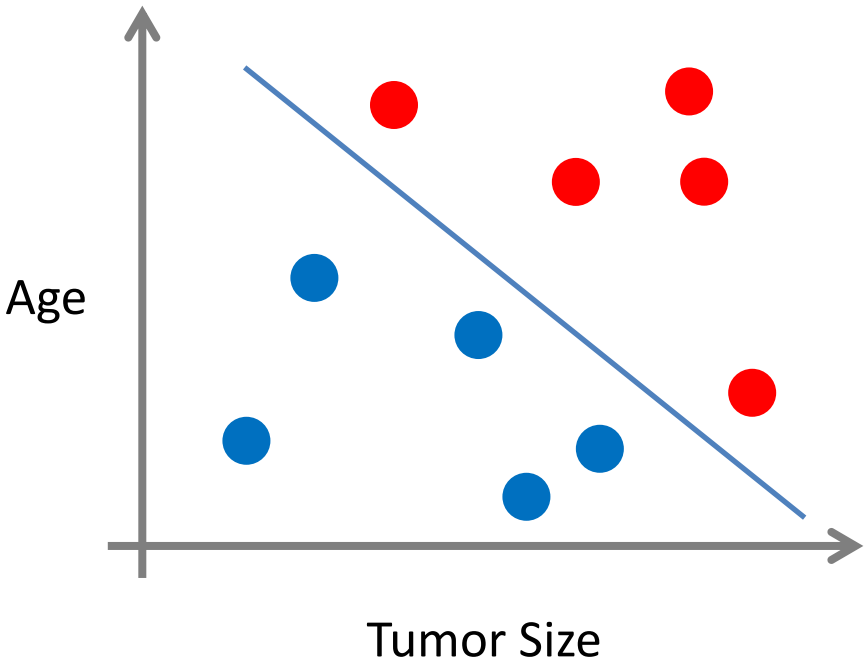


Based on slide by Eric Eaton
[example originally by Andrew Ng]

Supervised Learning

x can be multi-dimensional:

– each dimension corresponds to an attribute



- Clump Thickness
- Uniformity of Cell Size
- Uniformity of Cell Shape

…

Based on slide by Eric Eaton
[example originally by Andrew Ng]

Supervised learning: Vision

Image classification:

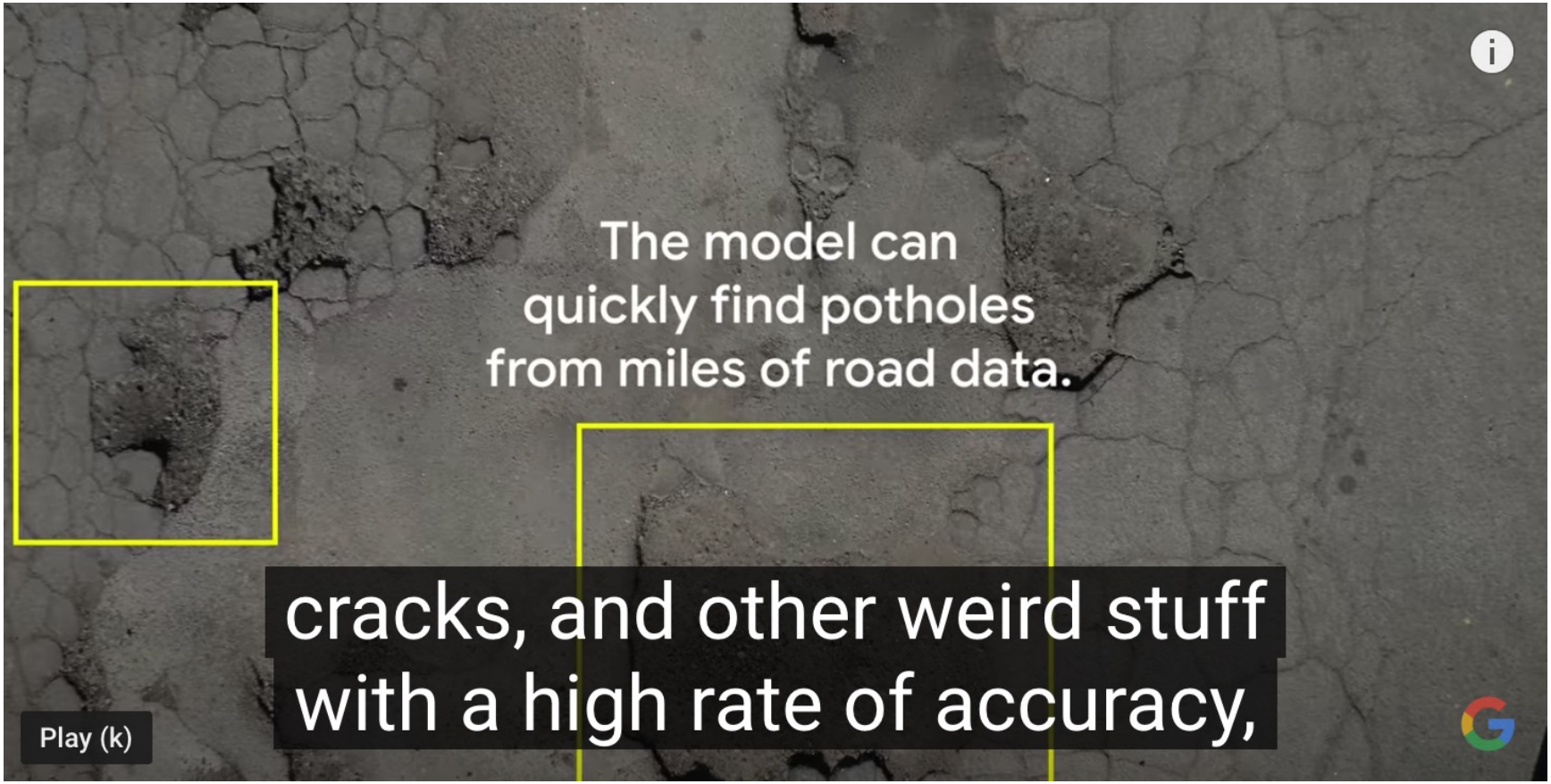
x = raw pixels of the image

y = the main object

Slide credit: Moses Charikar, Tengyu Ma, Chris Re

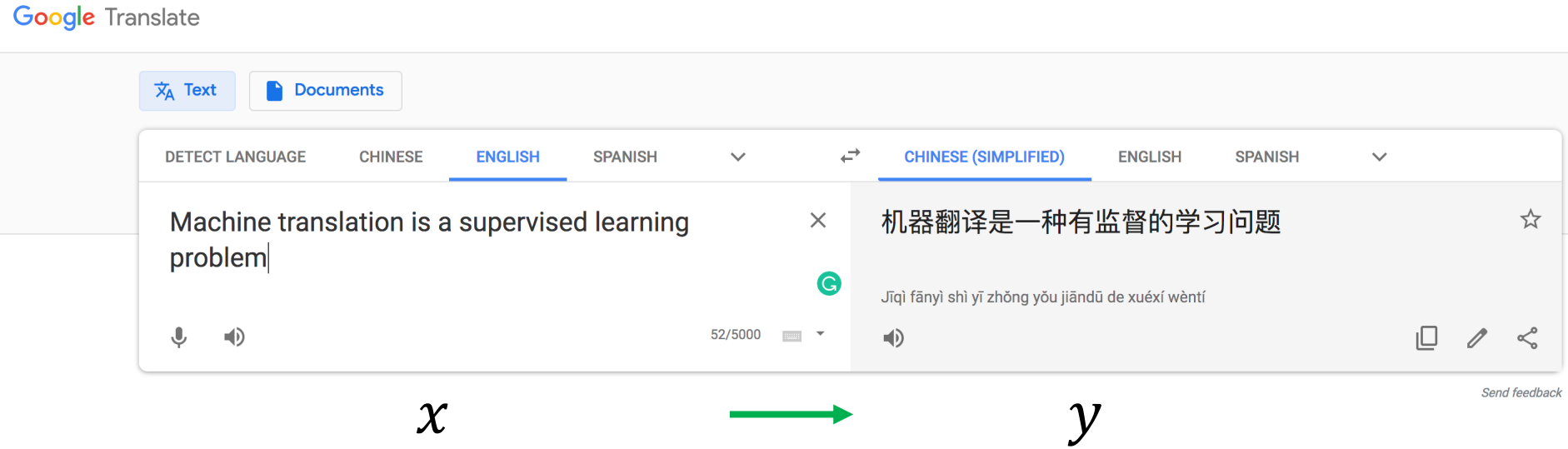
ImageNet Large Scale Visual Recognition Challenge. Russakovsky et al.'2015

Example: Pothole detection



Supervised learning: Natural Language Processing (NLP)

Machine translation



x → y

Note: this course only covers the basics of NLP—take 5370 next semester for more detail!

Slide credit: Moses Charikar, Tengyu Ma, Chris Re

Today's learning objectives

Students will be able to:

- ✓ Define Machine Learning.
- ✓ List three types of Machine Learning algorithms.
- ✓ Give examples of supervised learning tasks.
- Name an unsupervised learning technique.
- Explain how reinforcement learning works.
- Describe the learning task with math.

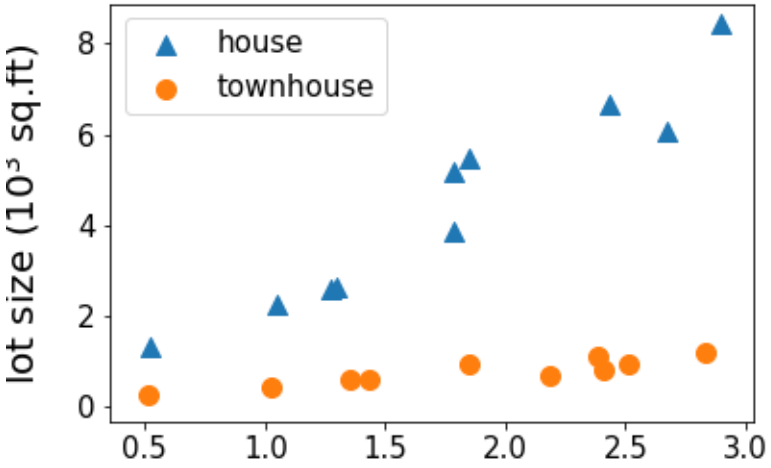
Unsupervised learning

Dataset contains no labels: $x^{(1)}, \ldots, x^{(n)}$

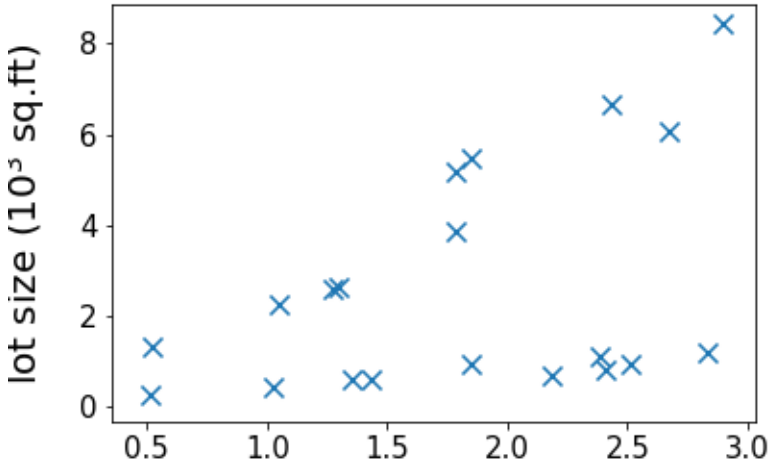
Goal: find hidden structures in the data

– e.g., clustering

supervised

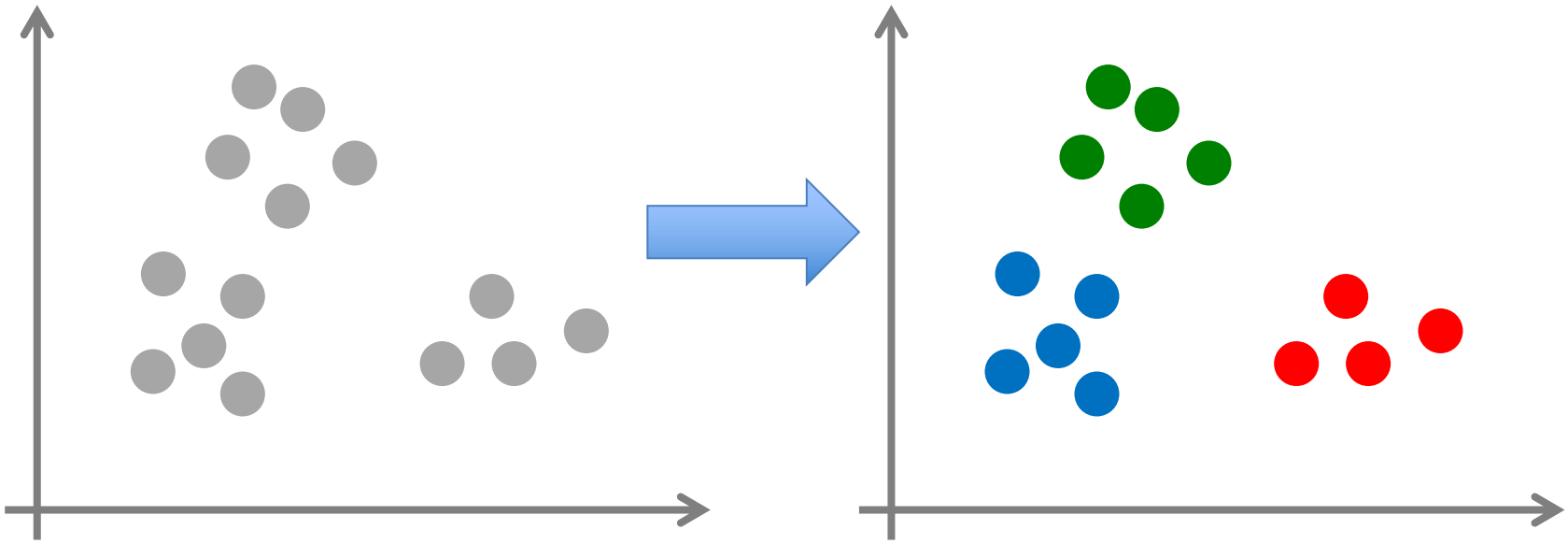


unsupervised



Slide credit: Moses Charikar, Tengyu Ma, Chris Re

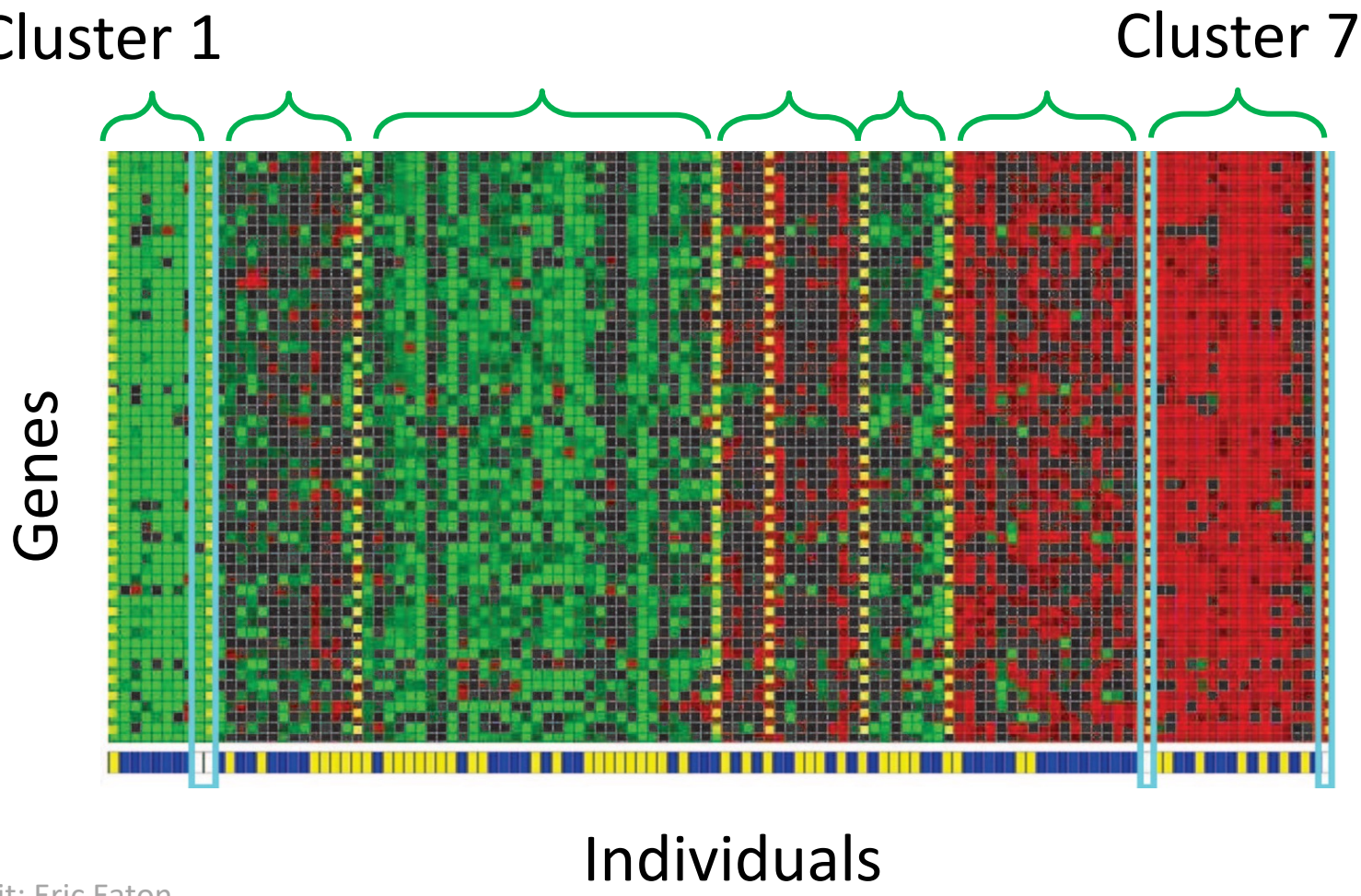
Clustering



Based on slide by Eric Eaton

Clustering genes

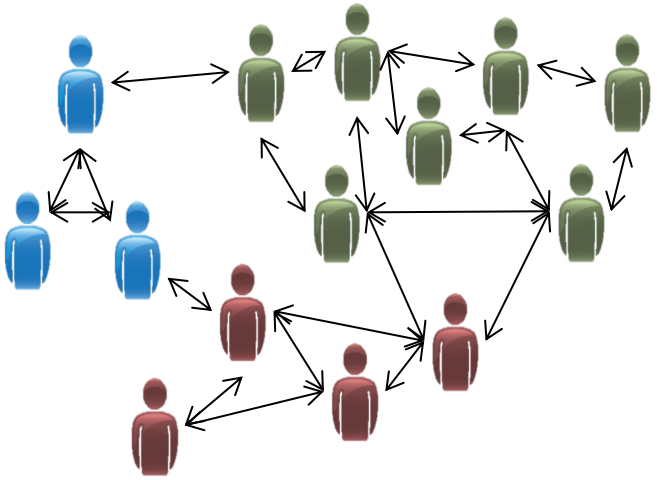
Genomics: group individuals by genetic similarity



Slide credit: Eric Eaton
[Source: Daphne Koller]

Clustering customer segments

Market segmentation



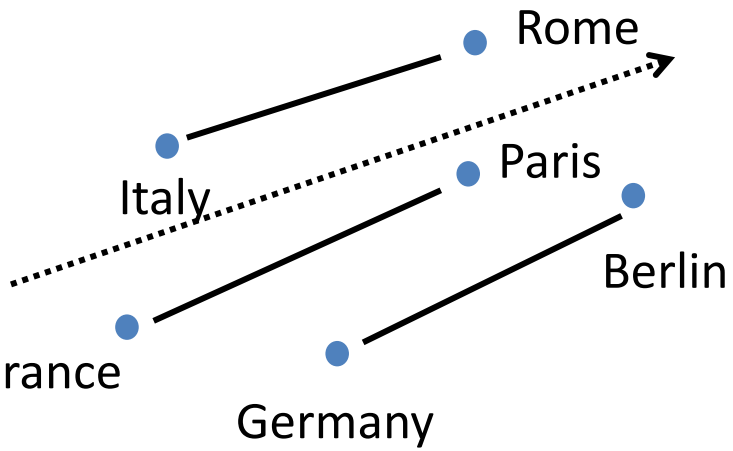
Social network analysis

Based on slide by Andrew Ng

Unsupervised NLP: Word embeddings

Represent words by vectors

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



Unlabeled dataset

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

Slide credit: Moses Charikar, Tengyu Ma, Chris Re

Today's learning objectives

Students will be able to:

- ✓ Define Machine Learning.
- ✓ List three types of Machine Learning algorithms.
- ✓ Give examples of supervised learning tasks.
- ✓ Name an unsupervised learning technique.
- Explain how reinforcement learning works.
- Describe the learning task with math.

Reinforcement learning

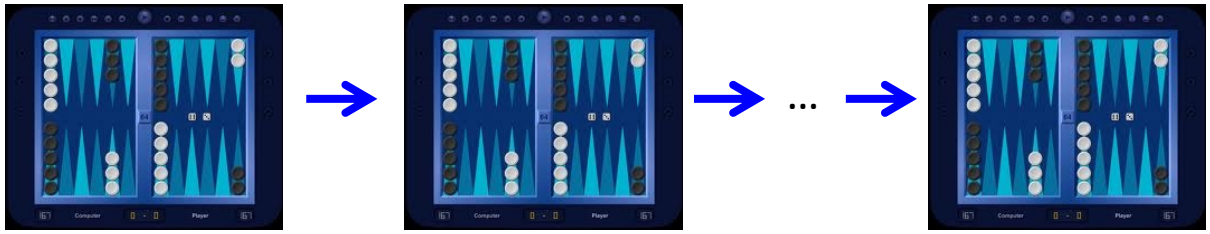
Collects data interactively

Goal: maximize the reward

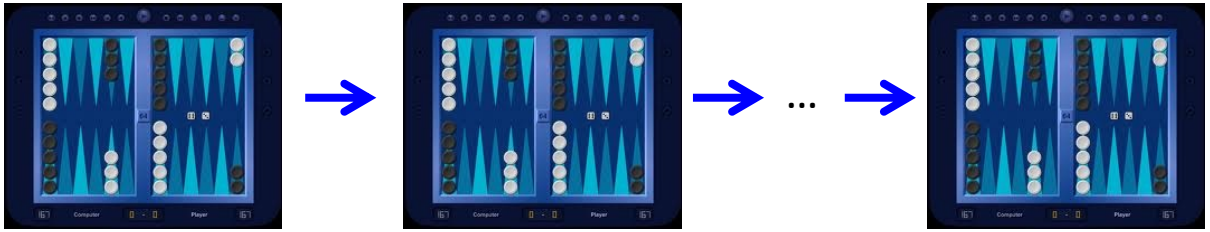
https://towardsdatascience.com/a-beginners-guide-to-q-learning-c3e2a30a653c

Reinforcement learning

Backgammon



WIN!



LOSE!

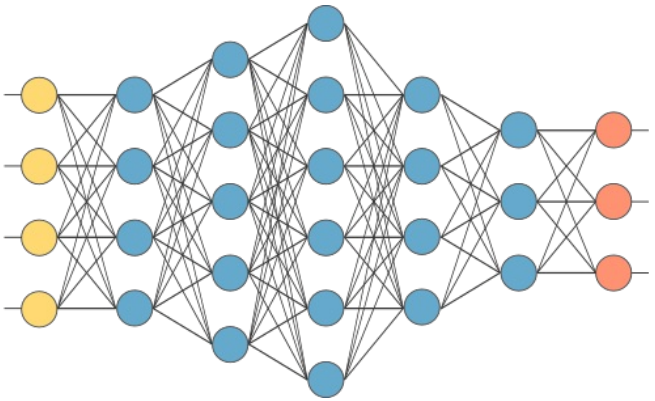
Given sequences of moves and whether or not the player won at the end, learn to make good moves

Slide credit: David Kauchak

Reinforcement learning

Brief Peek at Topics

- **Supervised learning**
  - Perceptron
  - Decision trees
  - K-nearest neighbors
  - Linear regression
  - Logistic regression
  - Support vector machines
  - Ensemble methods
- **Unsupervised learning**
  - Dimensionality reduction
  - Clustering
- **Reinforcement learning**
  - None: covered in-depth in AI (CMSI 3300)
- **Deep learning**



- **Bias and fairness in ML**

Today's learning objectives

Students will be able to:

- ✓ Define Machine Learning.
- ✓ List three types of Machine Learning algorithms.
- ✓ Give examples of supervised learning tasks.
- ✓ Name an unsupervised learning technique.
- ✓ Explain how reinforcement learning works.
- Describe the learning task with math.

Framing a Learning Problem

Learning Goals

- List stages of ML pipeline
- Name some key issues in ML

Defining the learning task

Improve on task T, with respect to
performance metric P, based on experience E

T: Playing checkers
P: Percentage of games won against an arbitrary opponent
E: Playing practice games against itself

T: Recognizing hand-written words
P: Percentage of words correctly classified
E: Database of human-labeled images of handwritten words

T: Categorize email messages as spam or legitimate
P: Percentage of email messages correctly classified
E: Database of emails, some with human-given labels

Q: Which type of ML algorithm does each task use?

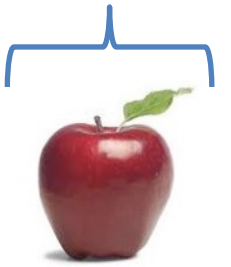
Slide credit: Ray Mooney

Representing examples

What is an example?

How is it represented?

examples



features

$feat_1$, $feat_2$, $feat_3$, $feat_4$,…

color, shape, leaf, weight, …

red, round, leaf, 3oz, …

green, round, no leaf, 4oz, …

yellow, curved, no leaf, 4oz, …

green, curved, no leaf, 5oz, …

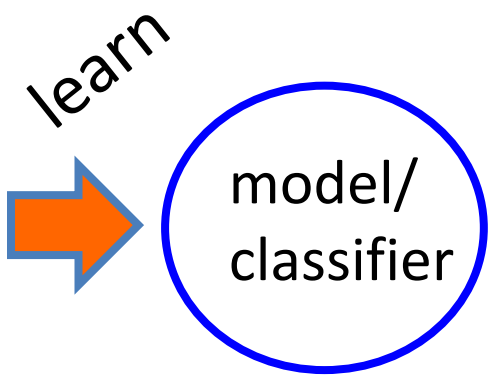
How our algorithms actually "view" the data

Features are the questions we can ask about the examples

Based on slide by David Kauchak

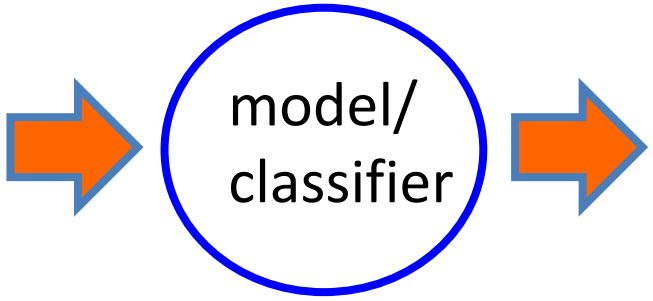
Learning system

examples' features	label
red, round, leaf, 3oz, …	apple
green, round, no leaf, 4oz, …	apple
yellow, curved, no leaf, 4oz, …	banana
green, curved, no leaf, 5oz, …	banana



During learning/training/induction, learn a model of what distinguishes apples and bananas *based on the features*

red, round, no leaf, 4oz, …



apple or banana?

apple Why?

The model can then classify a new example *based on the features*

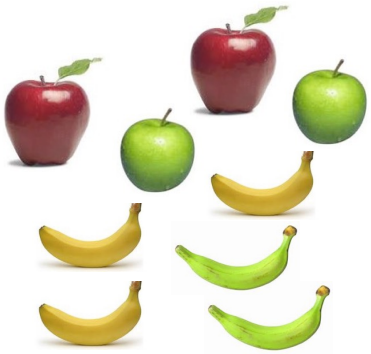
Based on slide by David Kauchak

Learning system

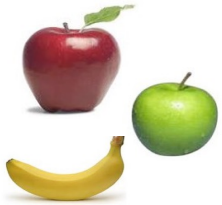
Learning is ***generalizing*** from training data.

What does this ***assume*** about train and test sets?

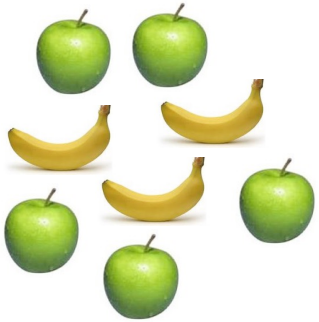
Training data



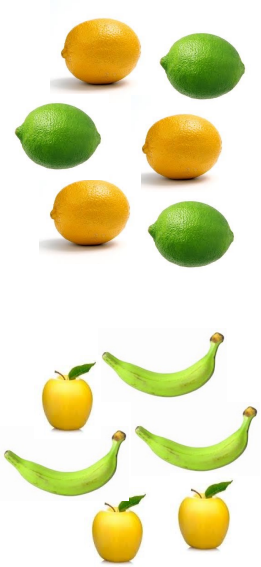
Test set



Training data



Test set



Not always the case, but we'll often assume it is!

Based on slide by David Kauchak

More technically…

We will use the ***probabilistic model*** of learning.

There is some (unknown) probability distribution over example/label pairs called the ***data generating distribution***.

Both the training data **and** the test set are generated based on this distribution.

- we call this i.i.d. (independent and *identically distributed*)

Based on slide by David Kauchak

ML as function approximation

Problem Setting:

- Set of possible instances X
- Set of possible labels Y
- Unknown target function $f : X \rightarrow Y$
- Set of function hypotheses $H = \{h \mid h : X \rightarrow Y\}$

Input: Training examples of unknown target function f

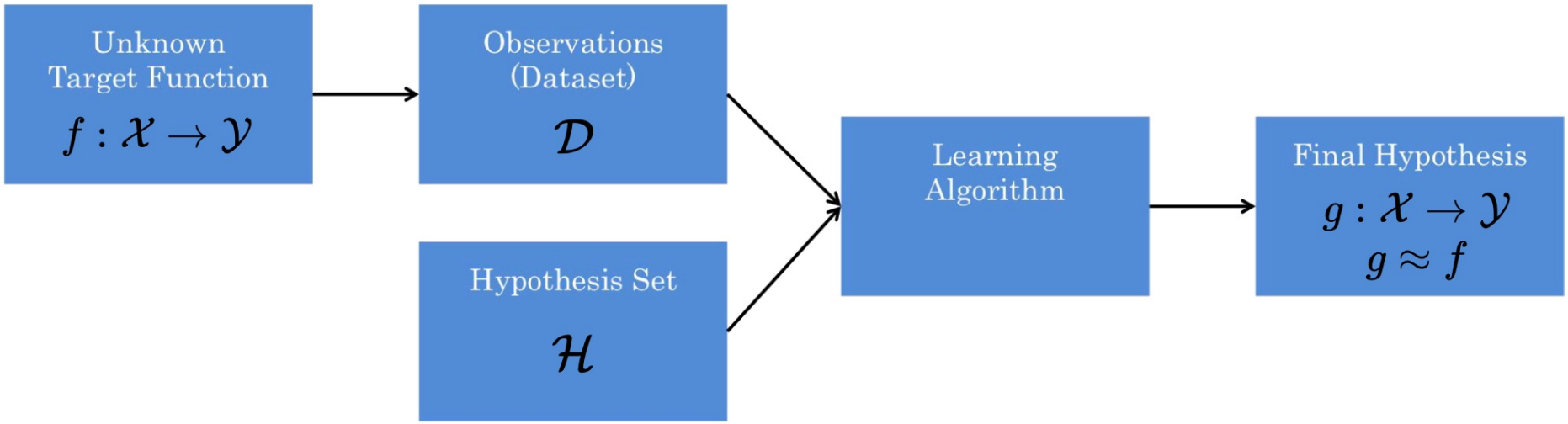
$$\left\{\left(x^{(i)}, y^{(i)}\right)\right\}_{i=1}^{n} = \left\{\left(x^{(1)}, y^{(1)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)\right\}$$

(superscript denotes example number)

Output: Hypothesis $h \rightarrow H$ that best approximates f

Slide credit: Eric Eaton
Based on slide by Tom Mitchell

The learning problem



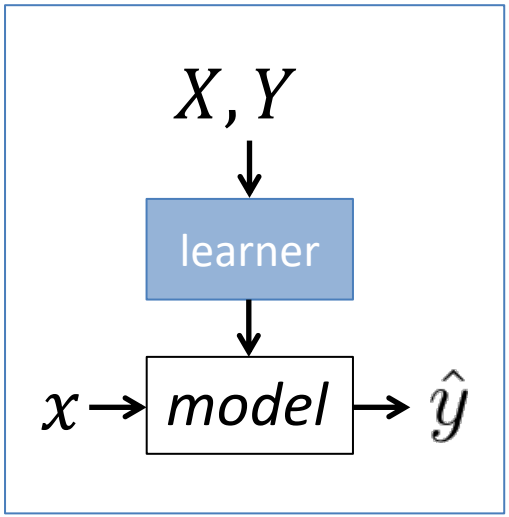
Slide credit: Alex Wong

Stages of Machine Learning

Given: labeled training data $X, Y = \left\{ (x^{(i)}, y^{(i)} \right\}_{i=1}^{n}$

Train model:

model ← *learner*.train(X, Y)



Apply model to new data:

Given: new unlabeled instance x

$\hat{y}$ ← *model*.predict(x)

Example: Regression

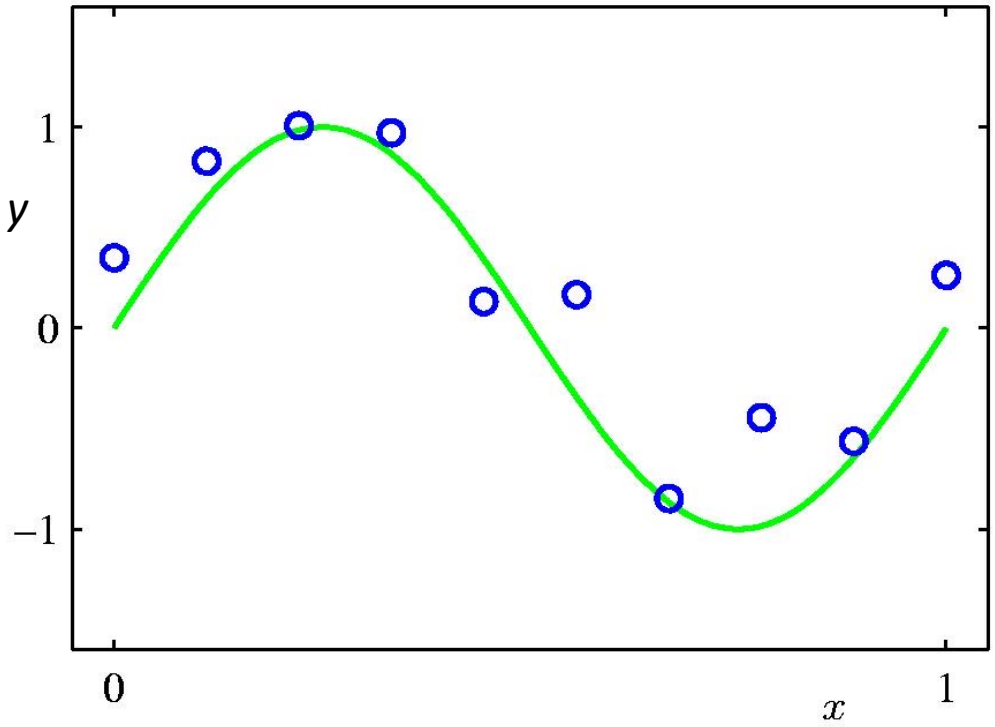
Consider regression:

- $f : X \to Y$
- $x, y \in \mathbb{R}$

Question 1: How should we pick the hypothesis space H?

Question 2: How do we find the best h in this space?

Dataset: 10 points generated from sine function with noise



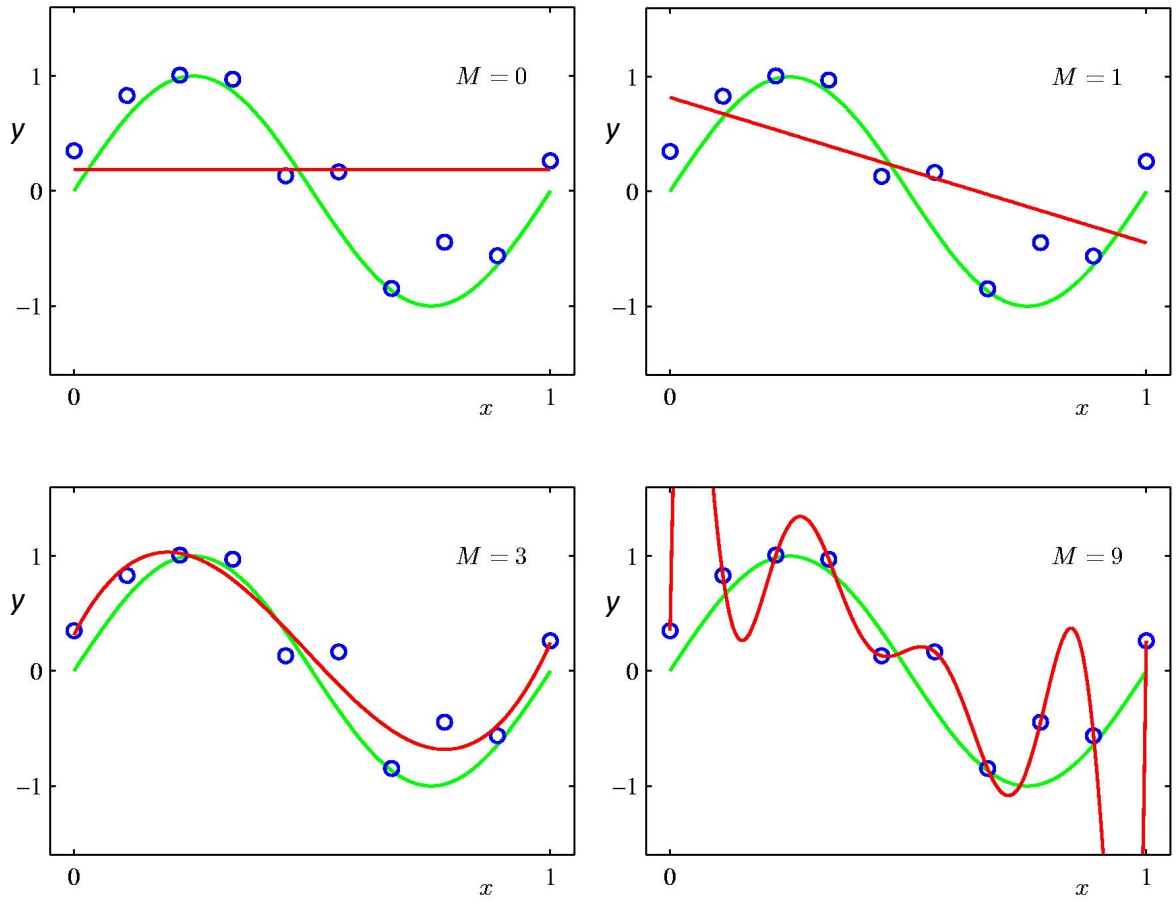
Based on slide by David Sontag
Images from Bishop [PRML]

Hypothesis space: Degree-M polynomials

Infinitely many hypotheses

$M < 9$ inconsistent with dataset, but $M \geq 9$ is consistent

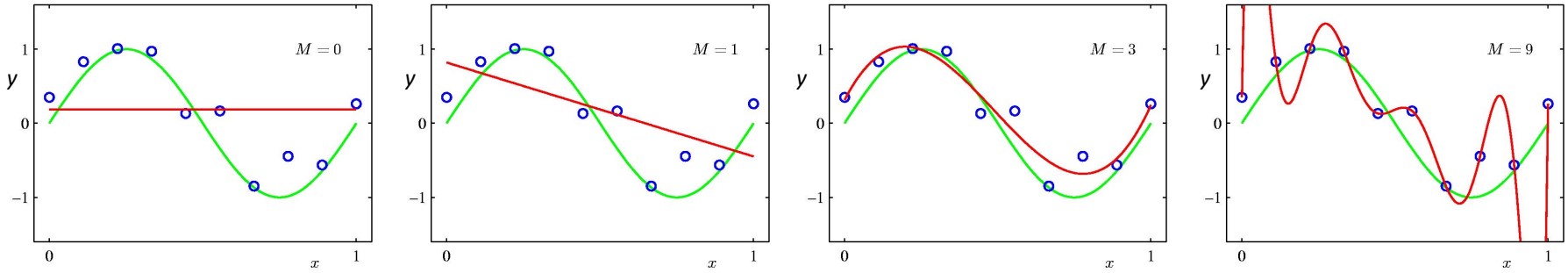
Which one is **best**?



$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Based on slide by David Sontag
Images from Bishop [PRML]

Hypothesis space: Degree-M polynomials



We measure error using a loss function $L(y, \hat{y})$.

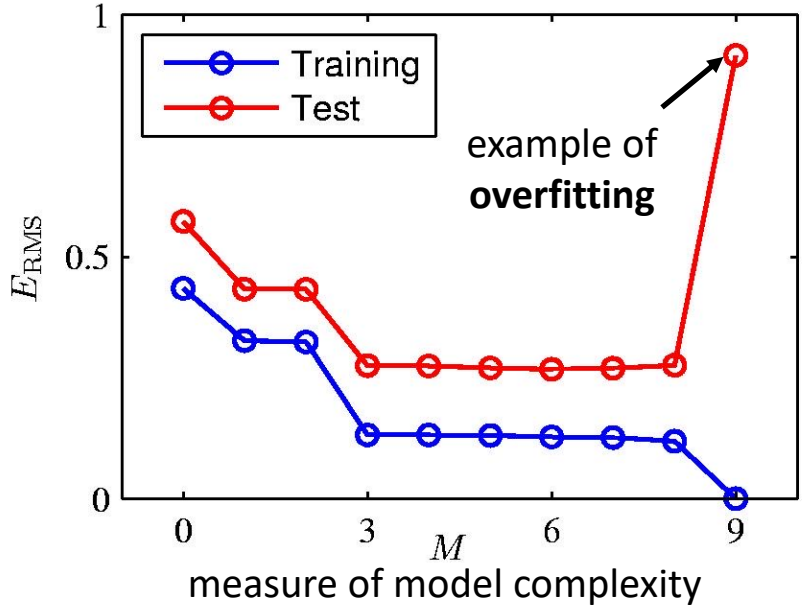
For regression, common choice is squared loss:

$$L\left(y^{(i)}, h(x^{(i)})\right) = \left(y^{(i)} - h(x^{(i)})\right)^2$$

Empirical loss of function *h* applied to training data:

$$\frac{1}{n}\sum_{i=1}^{n} L\left(y^{(i)}, h(x^{(i)})\right) = \frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - h(x^{(i)})\right)^2$$

Learning Curve



Based on slide by David Sontag
Images from Bishop [PRML]

Key issues in Machine Learning

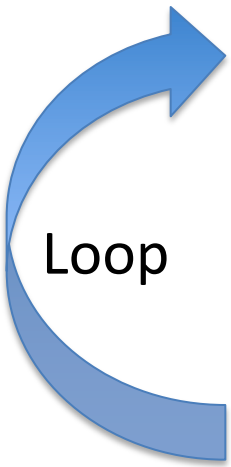
Representation: How do we choose a hypothesis space?

Optimization: How do we find the *best* hypothesis?

Evaluation: How can we gauge the accuracy of a hypothesis on unseen testing data?

Based on slides by Eric Eaton and by David Sontag

Machine Learning in practice



Understand domain, prior knowledge, and goals

Data integration, selection, cleaning, pre-processing, etc.

Learn models

Interpret results

Consolidate and deploy discovered knowledge

Slide credit: Eric Eaton
Based on a slide by Pedro Domingos

Today's learning objectives

Students will be able to:

- ✓ Define Machine Learning.
- ✓ List three types of Machine Learning algorithms.
- ✓ Give examples of supervised learning tasks.
- ✓ Name an unsupervised learning technique.
- ✓ Explain how reinforcement learning works.
- ✓ Describe the learning task with math.

Exercise

Pick 2 applications and frame them as a learning problem with input (features) X and predicted output y.

- Web search
- Finance
- Healthcare
- E-commerce
- Natural language processing
- Robotics
- Social networks
- [Your favorite area]

Based on slide by Pedro Domingos

Quote of the day

Phoebe Rusack, 2023, weareteachers.com

Socrative: Optional feedback

http://socrative.com
Go to student login (it's anonymous!)