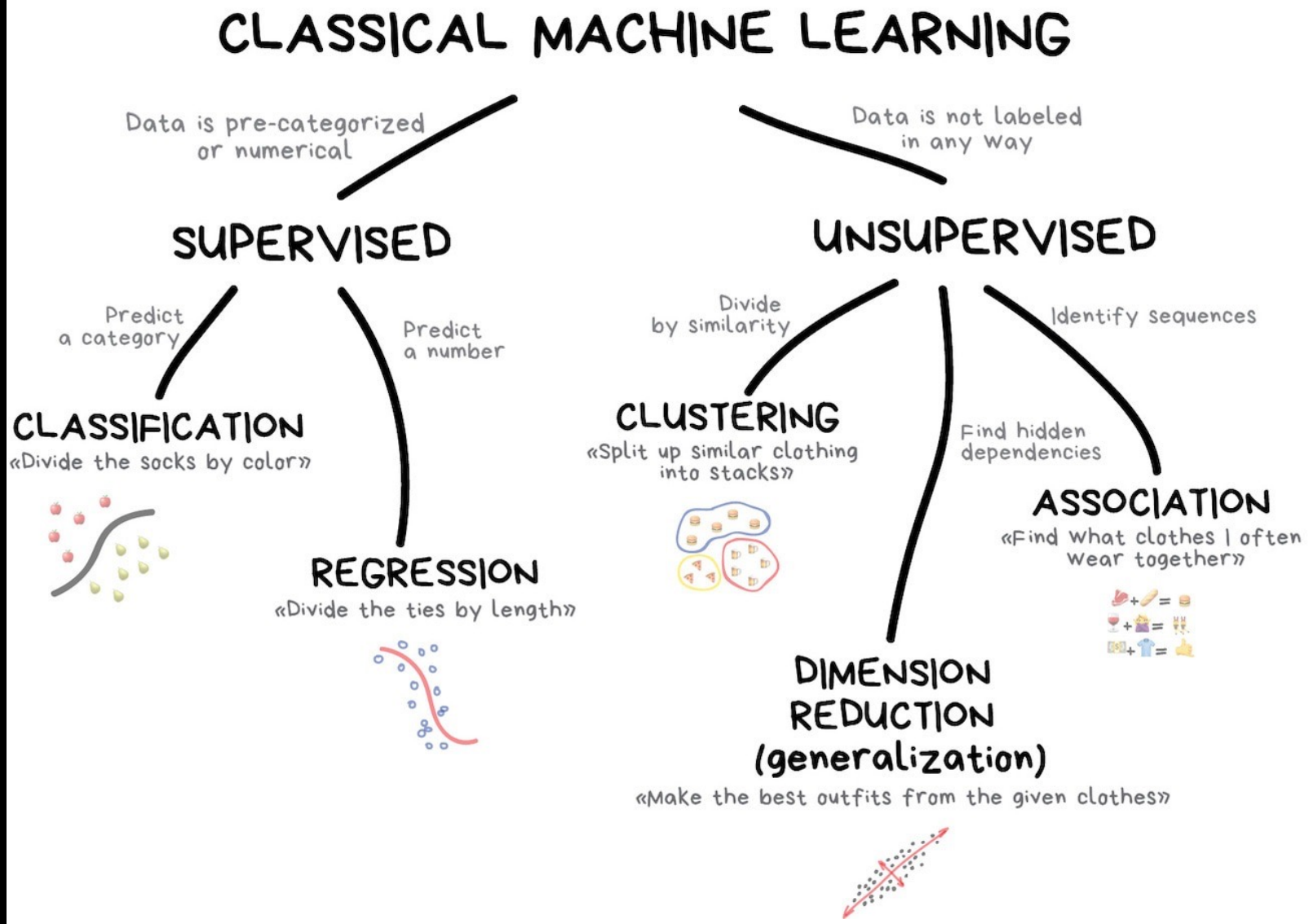
Machine Learning I

Carolina A. de Lima Salge
Assistant Professor
Terry College of Business
University of Georgia

Business Intelligence
Spring 2021

Terry College of Business
UNIVERSITY OF GEORGIA



https://vas3k.com/blog/machine_learning/

What Is Machine Learning (ML)?

No accepted definition but several are available:

- Field of study that gives computers the ability to learn without being explicitly programmed (Samuel 1959)
- **A computer program is said to learn** from experience E with respect to task T and some performance measure P, **if its performance** on T, as measured by P, **improves** with experience E (Mitchell 1998)

Quiz “Question”

A computer program is said to learn from experience E with respect to task T and some performance measure P, if its performance on T, as measured by P, improves with experience E (Mitchell 1998)

Suppose your email program watches which emails you do or do not mark as spam and based on that learns how to better filter spam. **What is the task T in this setting?**

1. Classifying emails as spam or not spam
2. Watching you label emails as spam or not spam
3. The number of emails correctly classified as spam/not spam
4. None of the above

Objective of ML

1. Draw causal insights
   > "**What is causing** our customers to cancel their subscription to our services?"*
2. Predict future events
   > "**Which customers** are likely to cancel their subscription next month?"*
3. Understand patterns in data
   > "**Are there groups of customers** who are similar and use our services in a similar way?"*

A Note on Inference/Prediction

Inference is concerned with understanding the drivers of a business outcome

- Models of inference are interpretable but less accurate

The prediction itself is the main goal

- Not easily interpretable (“black-box”) but more accurate

1. Draw causal insights
   > "**What is causing** our customers to cancel their subscription to our services?"*
2. Predict future events
   > "**Which customers** are likely to cancel their subscription next month?"*

Inference

Which of these affect the fraud probability the most?

	Transaction data A	Transaction data B	Transaction data C	Transaction data D
Transaction 1				
Transaction 2				
Transaction 3				
Transaction ...				
Transaction N				

Fraud probability

Prediction

	Transaction data A	Transaction data B	Transaction data C	Transaction data D
Transaction 1				
Transaction 2				
Transaction 3				
Transaction ...				
Transaction N				

Get the most accurate probability this is fraud

Fraud probability

Two Types of ML

Supervised

- We teach the computer how to learn something
  - E.g., Which transactions are fraudulent? (**prediction**)
  - E.g., What are the main drivers of fraud? (**inference**)
  - A specific purpose
  - Requires specifying a target

Unsupervised

- We let the computer learn by itself
  - E.g., Do our transactions naturally fall into different groups?
  - No specific purpose
  - No need to specify a target

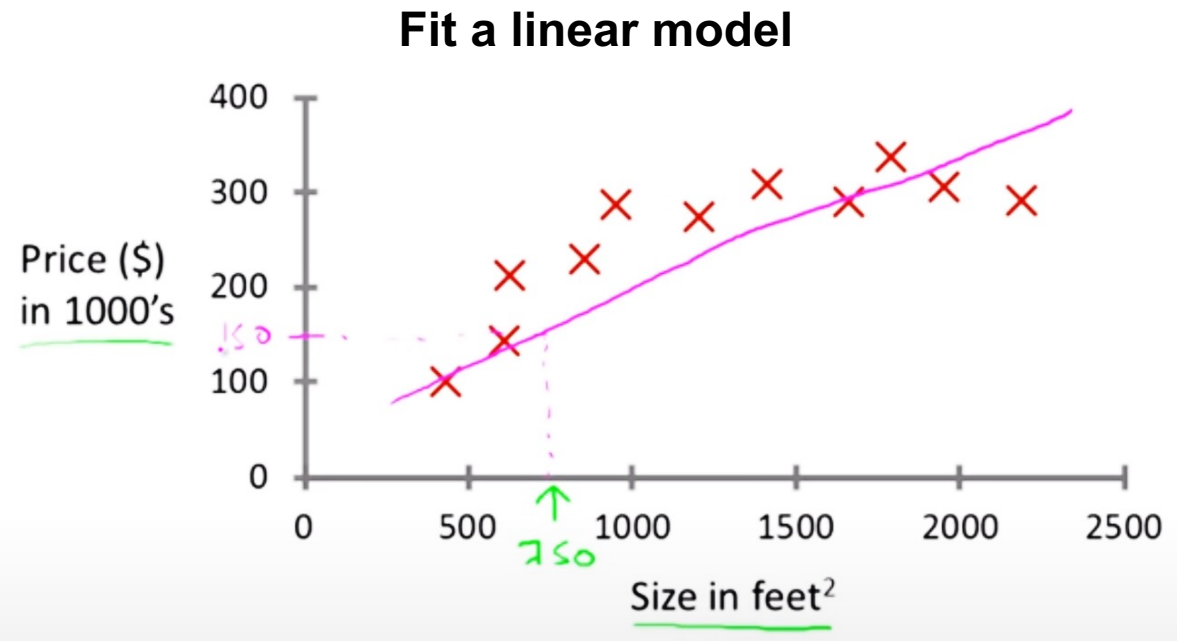
Others: Reinforcement learning, recommender systems

Focus on how to apply ML algorithms, not so much on theory or specifications (bunch of math!)

Supervised Learning

Most common machine learning problem (**also a regression problem**)

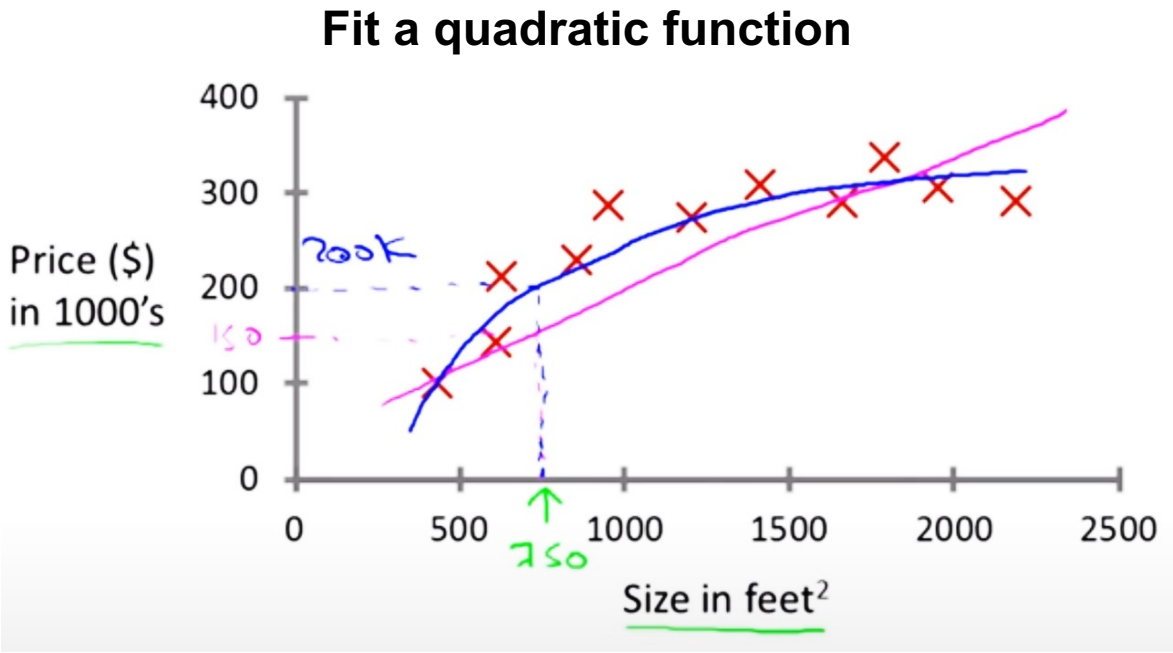
- Suppose you want to predict house prices, and you have some data about the price of a house (in thousands of $) over size (sqft)



Supervised Learning

Most common machine learning problem (**also a regression problem**)

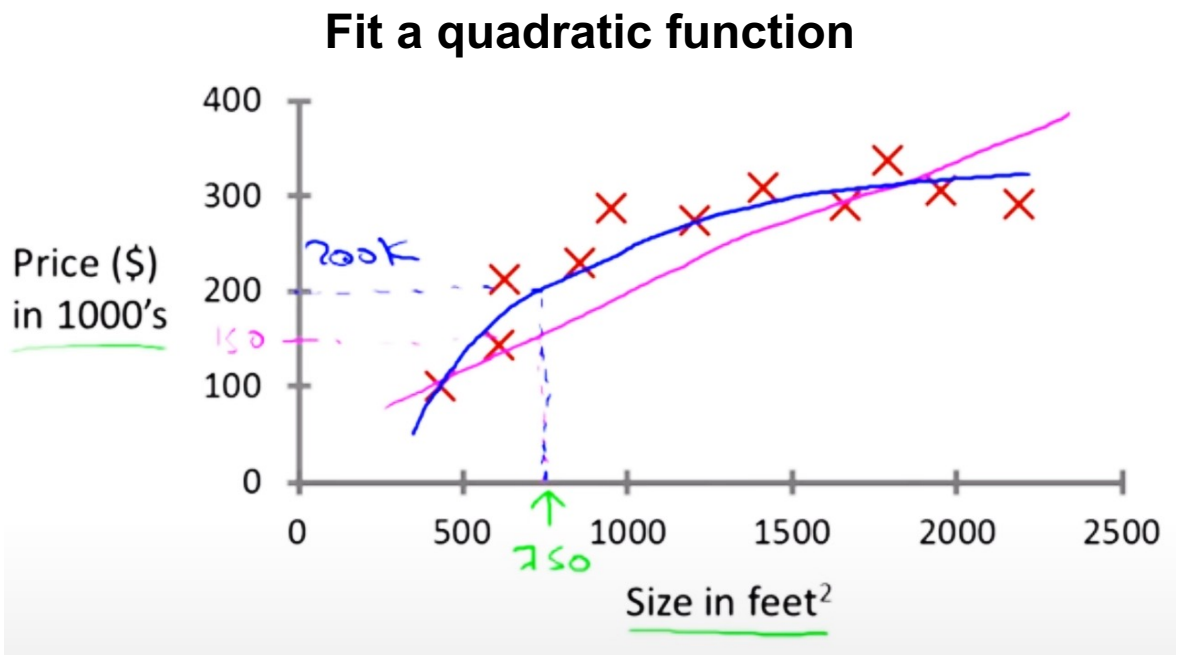
- Suppose you want to predict house prices, and you have some data about the price of a house (in thousands of $) over size (sqft)



Supervised Learning = "right answers" given

Most common machine learning problem (**also a regression problem**)

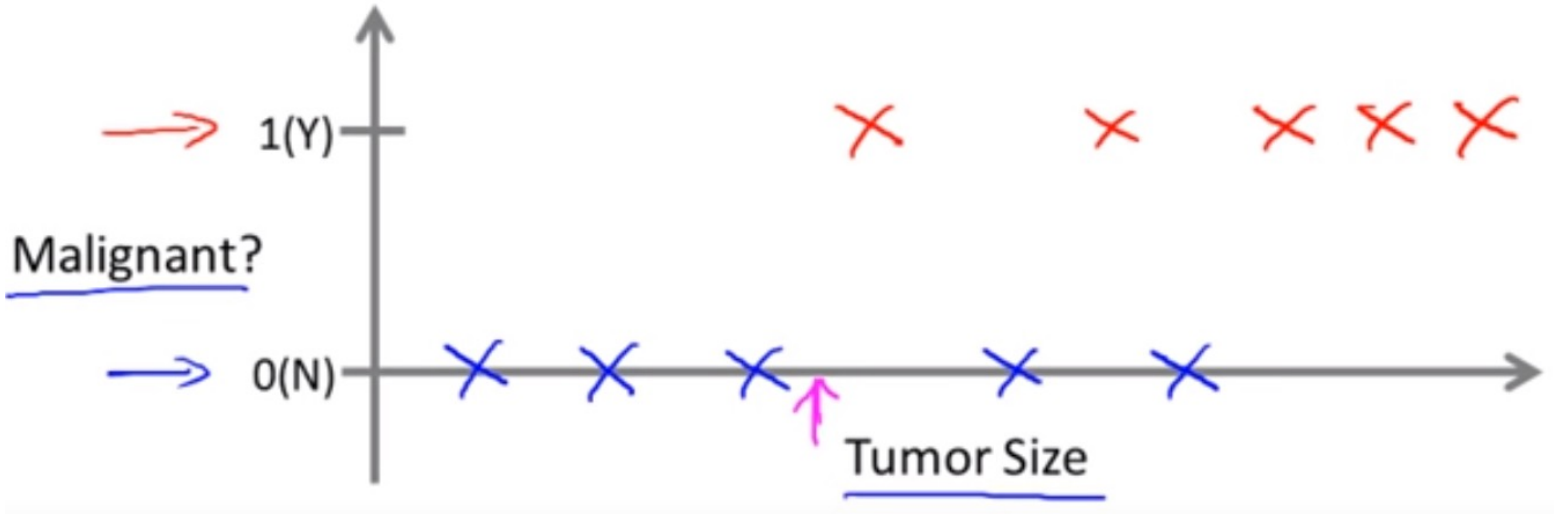
- Suppose you want to predict house prices, and you have some data about the price of a house (in thousands of $) over size (sqft)



Supervised Learning = "right answers" given

Most common machine learning problem (**also a classification problem**)

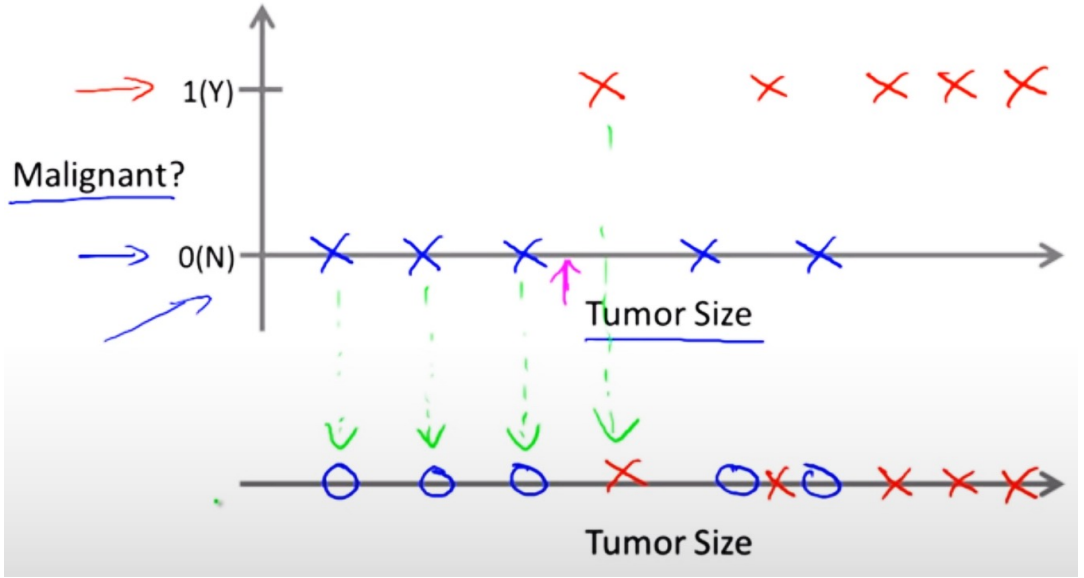
- Suppose you want to predict whether someone's breast cancer is malignant or benign, and you have some data about diagnosis (1s or 0s) over tumor size



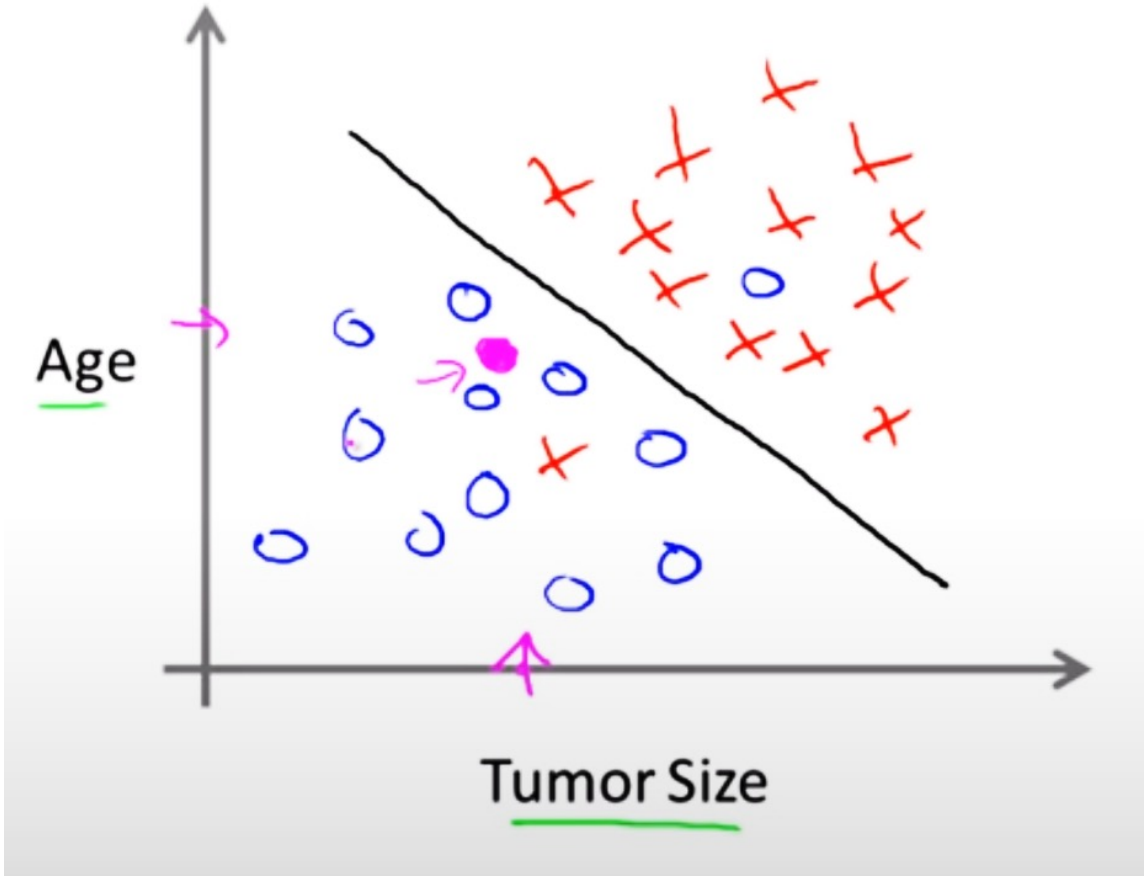
Supervised Learning = "right answers" given

Most common machine learning problem (**also a classification problem**)

- Suppose you want to predict whether someone's breast cancer is malignant or benign, and you have some data about diagnosis (1s or 0s) over tumor size



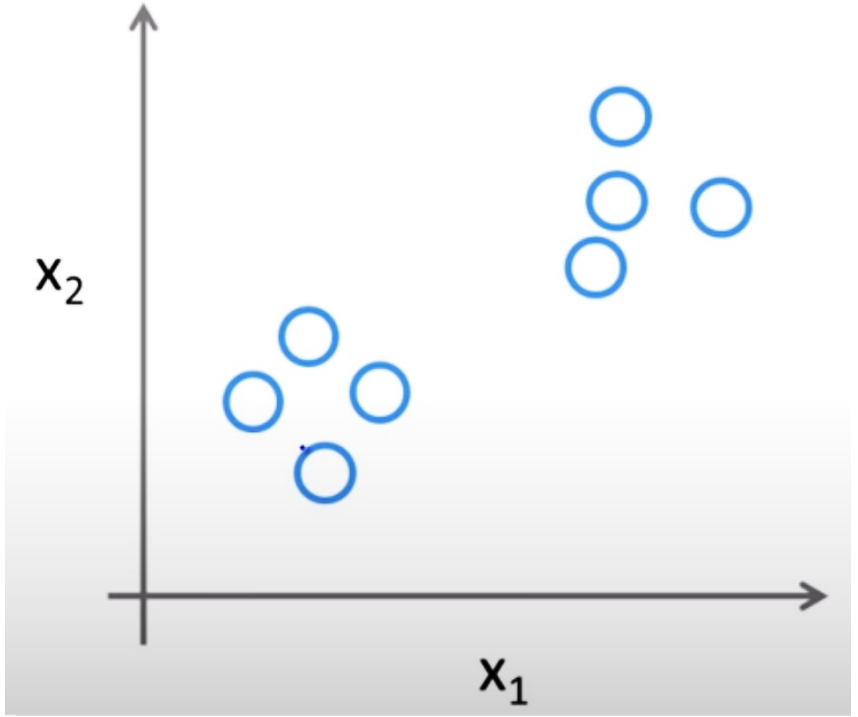
More than 1 attribute / independent variable



https://www.youtube.com/watch?v=bQI5uDxrFfA

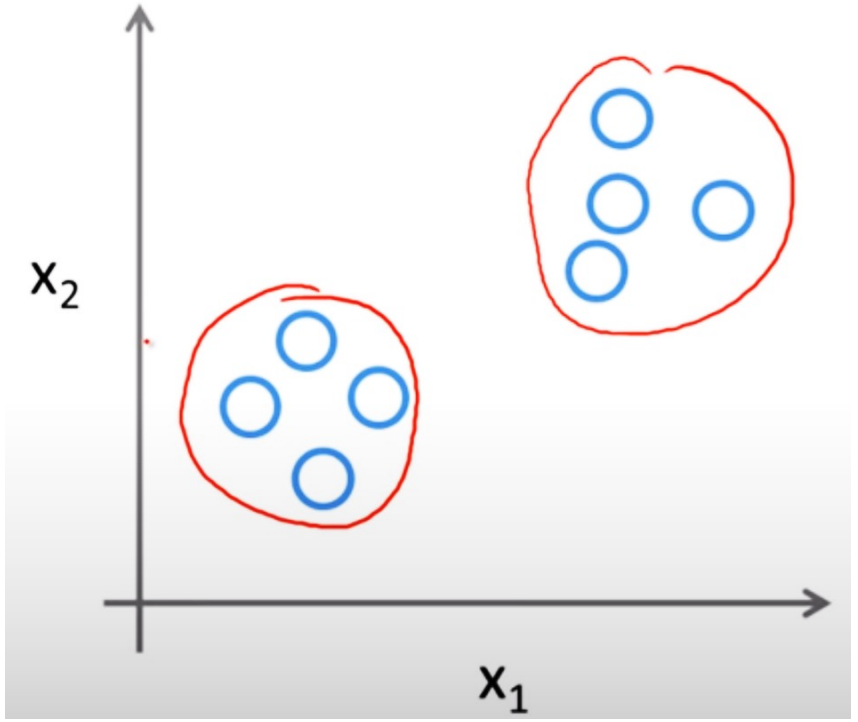
Unsupervised Learning = no target, just features

Here is a dataset, can you find some structure in the data?



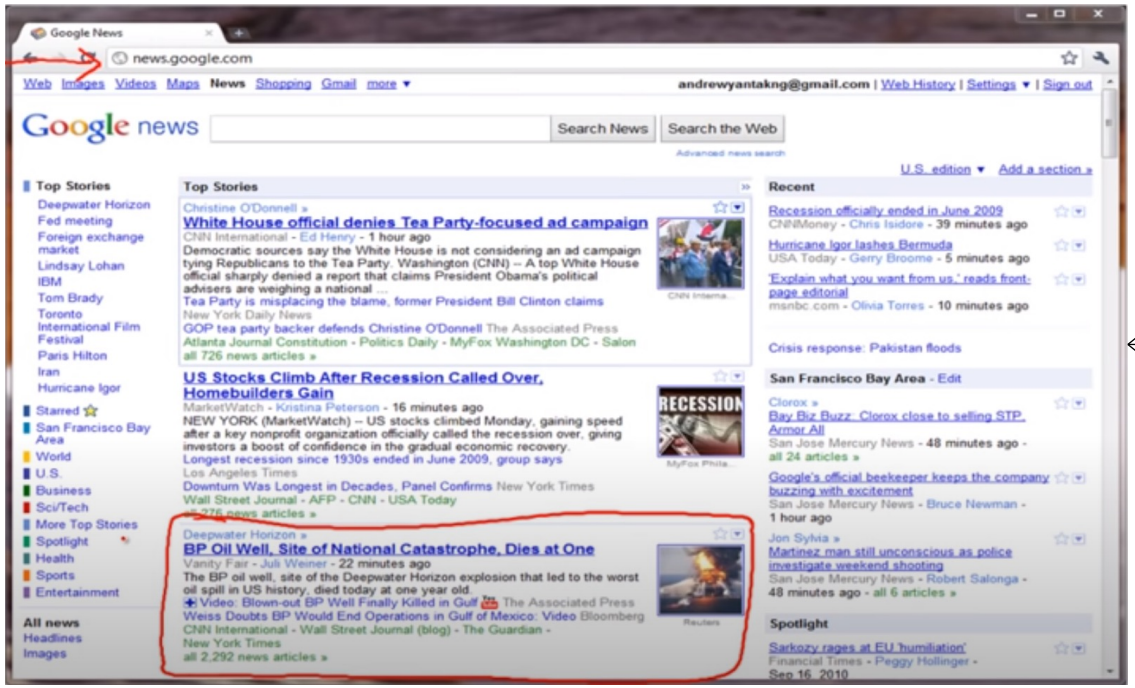
Unsupervised Learning = no target, just features

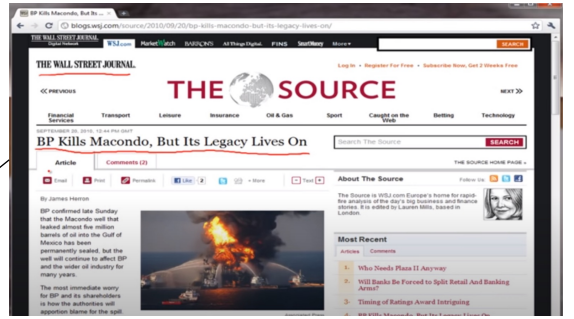
Here is a dataset, can you find some structure in the data?

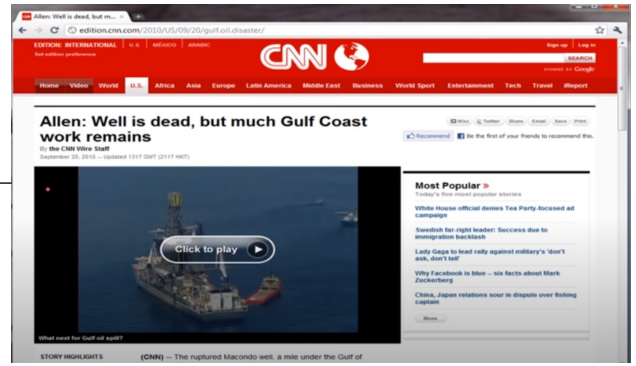


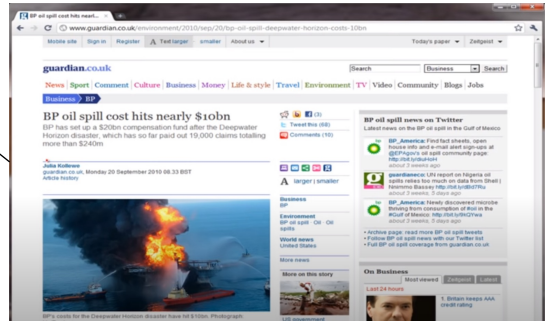
Unsupervised Learning = no target, just features

news.google.com clustering articles on the same topic together



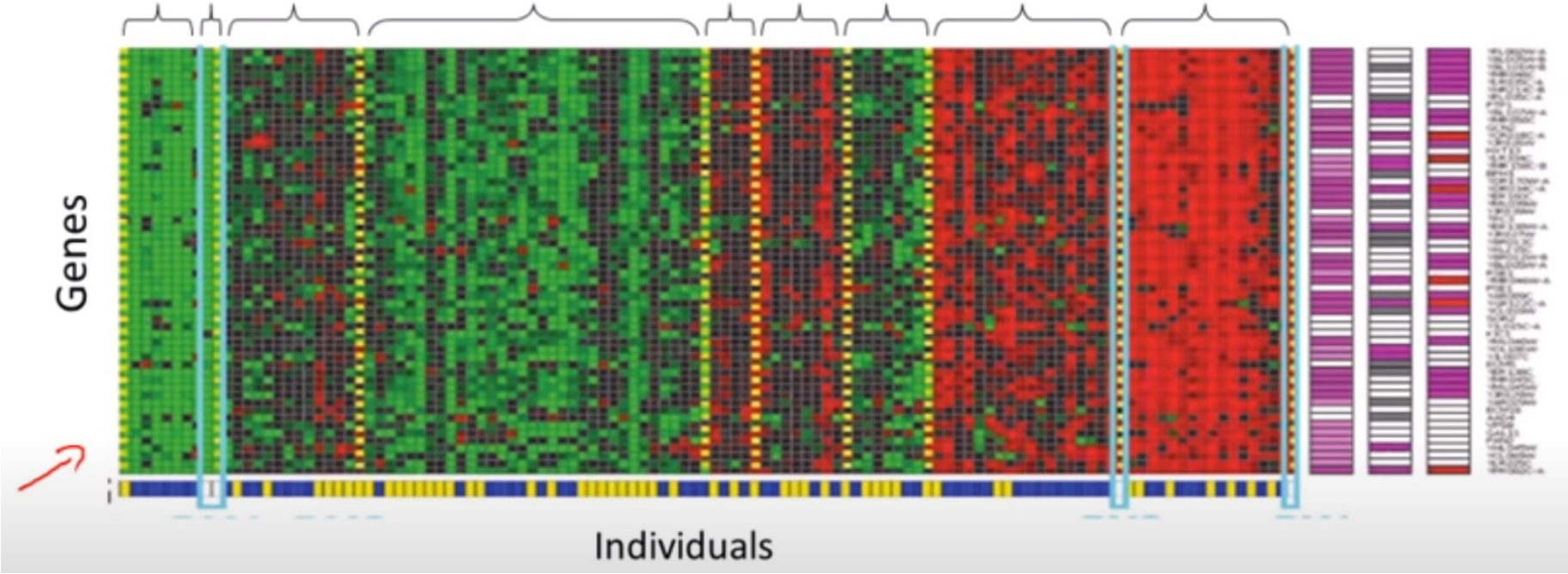






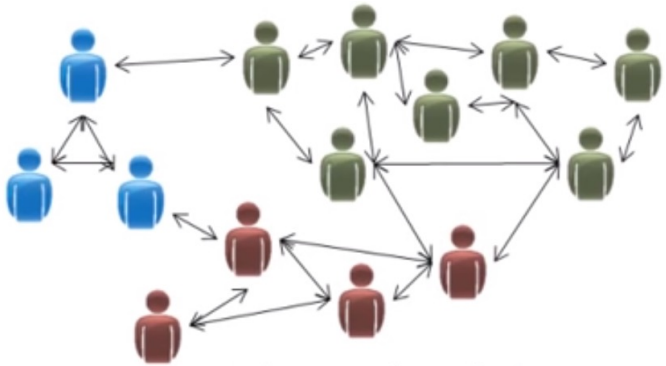
Unsupervised Learning = no target, just features

Understanding genomics – DNA data (automatically cluster different types of people based on their genes)



Unsupervised Learning = no target, just features

Organize computing clusters



Social network analysis

Market segmentation

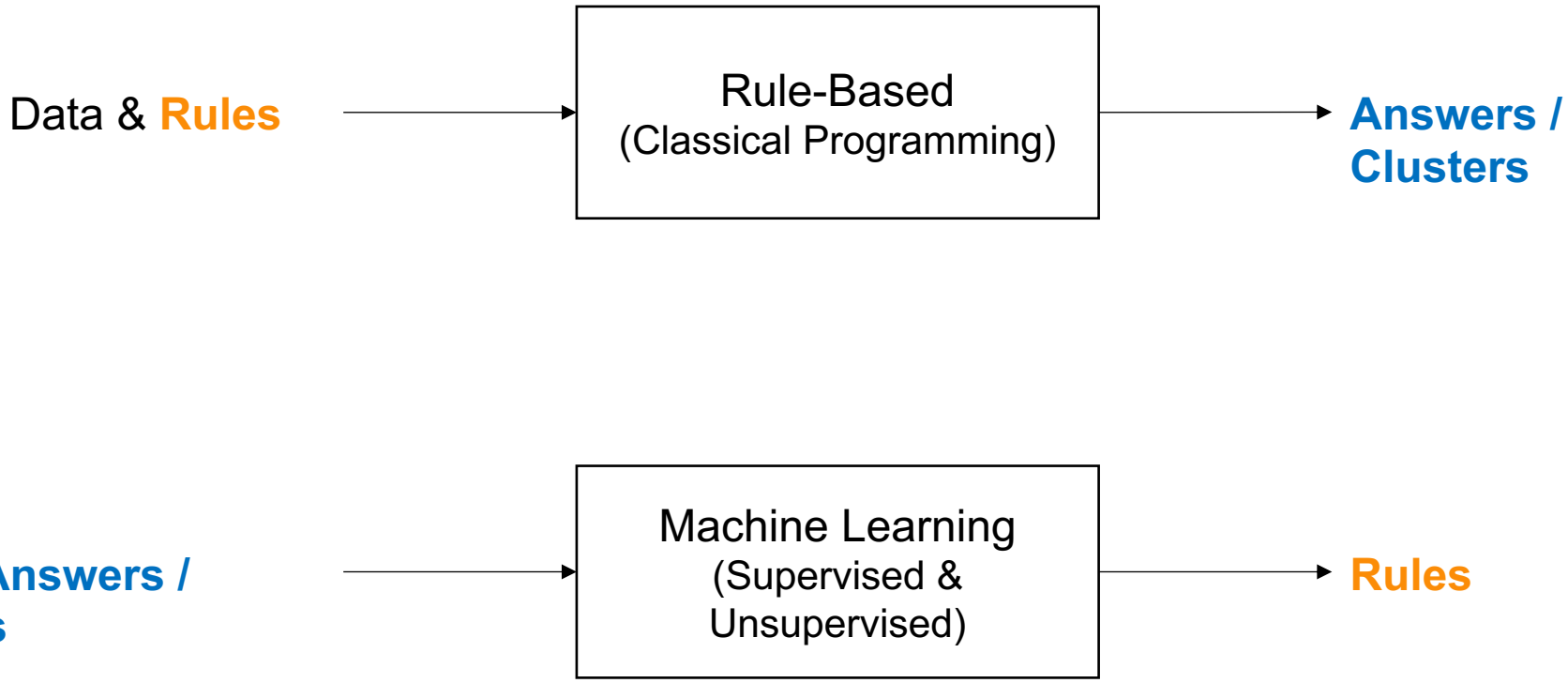
Astronomical data analysis

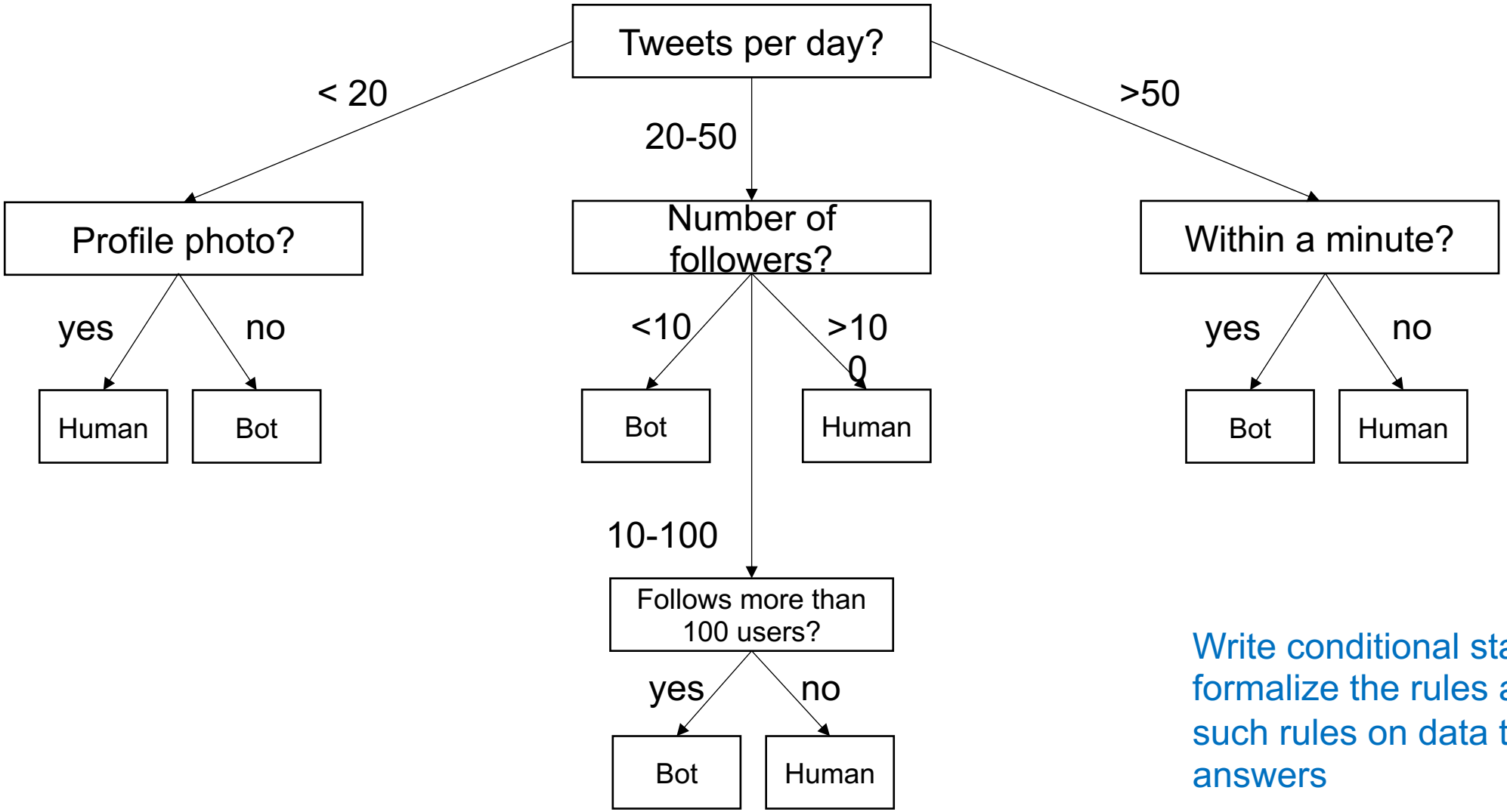
Quiz “Question”

Of the following examples, which would you address using an unsupervised learning algorithm? (Check all that apply.)

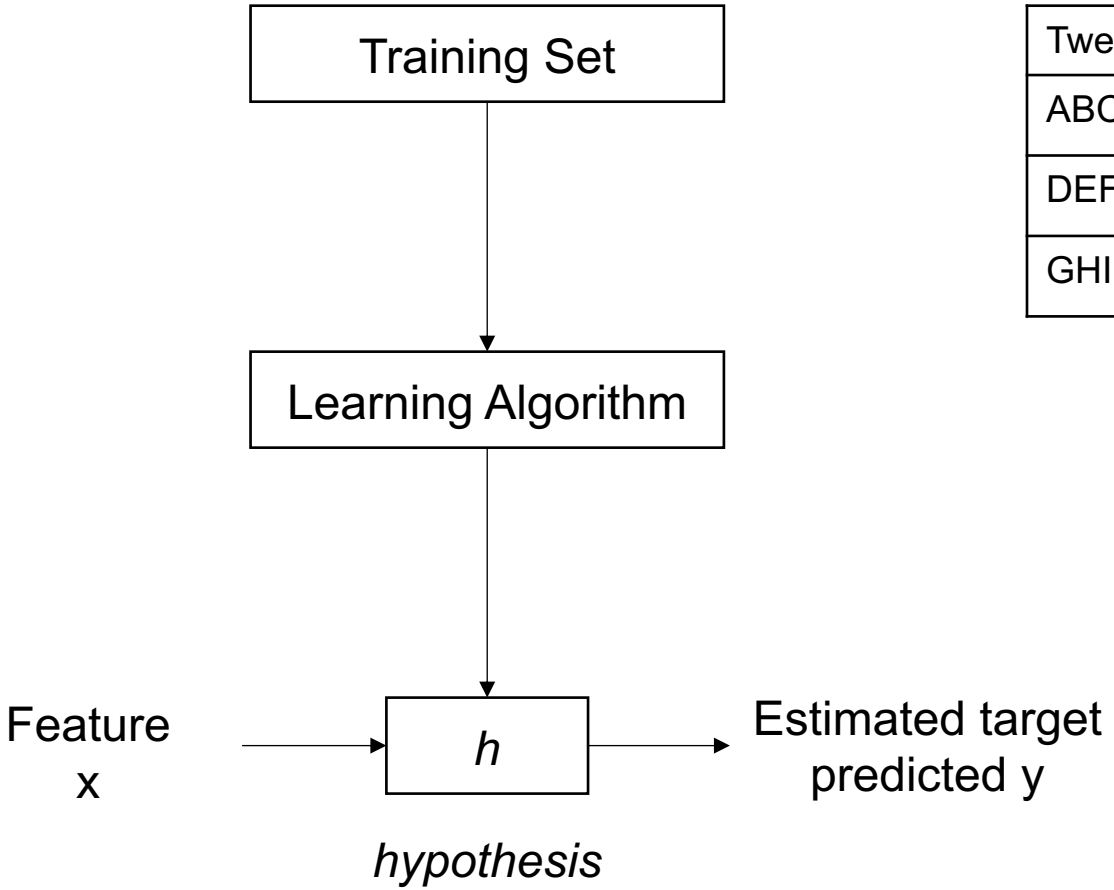
1. Given email labeled as spam/not spam, learn a spam filter
2. Given a set of new books found on the web, group them into set of books on the same topic
3. Given a database of customer data, automatically discover market segments and group customers into different market segments
4. Given a dataset of patients diagnosed as either having diabetes or not, learn to classify new patients as having diabetes or not

What Are the Differences?





Write conditional statements to formalize the rules and use such rules on data to get answers

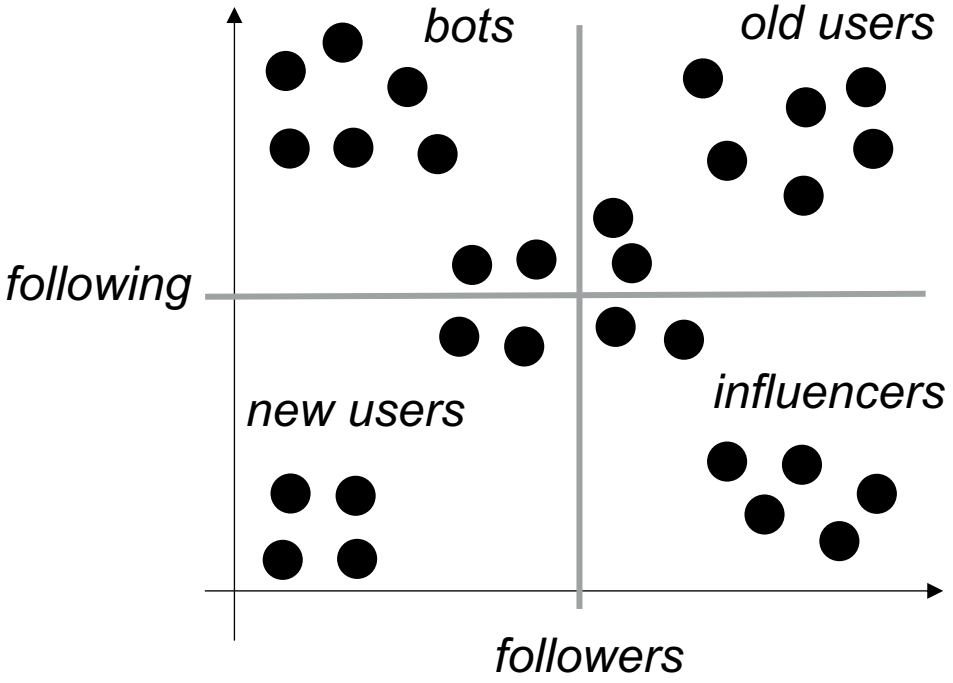


Tweets	Photo	Followers	Following	Timestamp	Human
ABC	1	10	100	11/2/20 5:55pm	1
DEF	0	5	5	11/2/20 5:55pm	0
GHI	0	100	1000	11/2/20 5:55pm	0

…

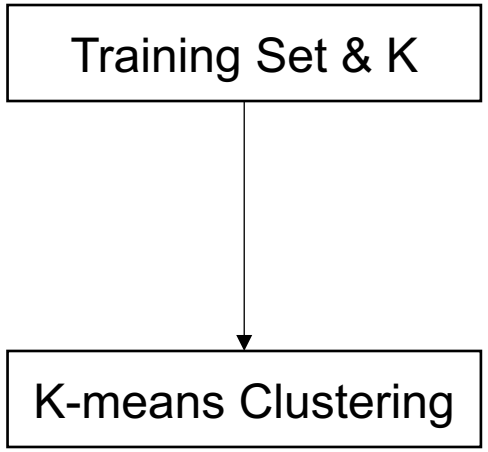
Regression
Decision Tree
Neural Networks
Support Vector Machines

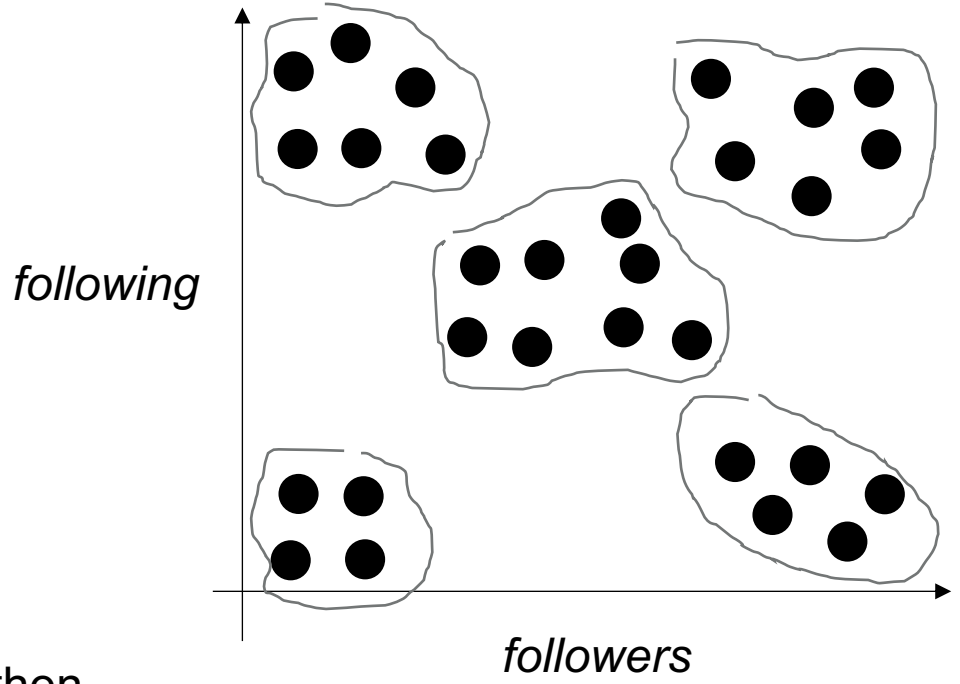
Use training set to learn the rules which can then be used to predict answers from testing set



No target, just features

Write conditional statements to formalize the rules and use such rules on data to derive clusters





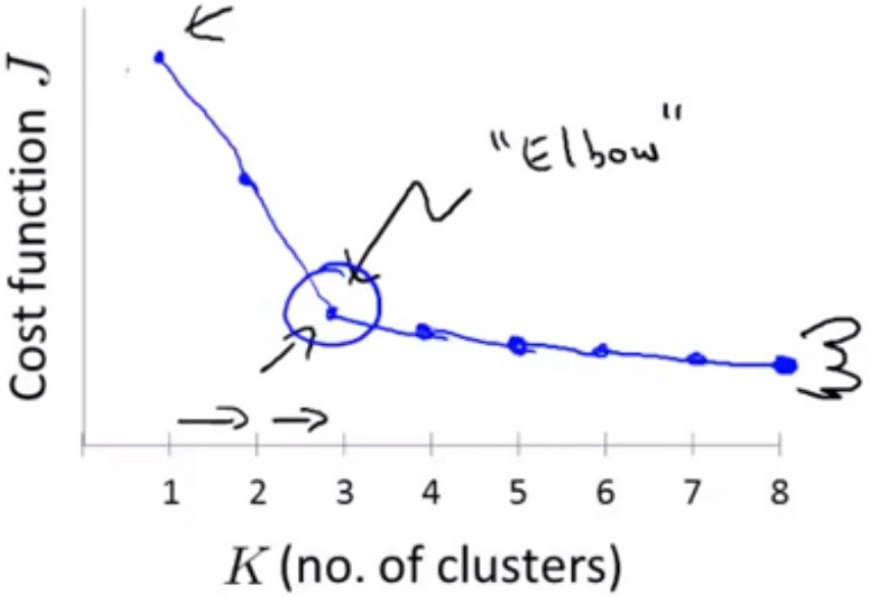
Randomly initialize K clusters centroids, then
For each value in the data, determine which is the closest *k* to it, then
Take average of values assigned to each cluster *k*

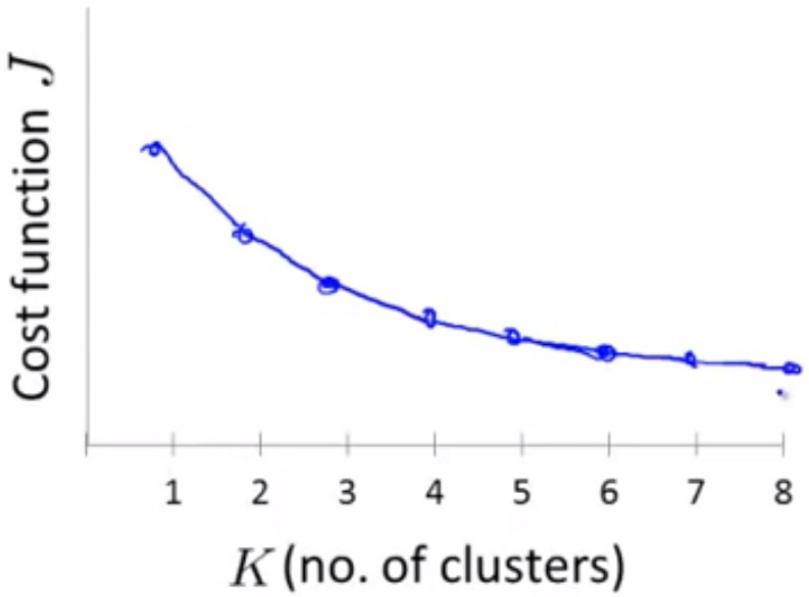
No target, just features

Use training set to learn the rules which can then be used to derive the clusters

Choosing K

Elbow method:





https://www.youtube.com/watch?v=lbR5br5yvrY&list=PLLssT5z_DsK-h9vYZkQkYNWcltqhIRJLN&index=80

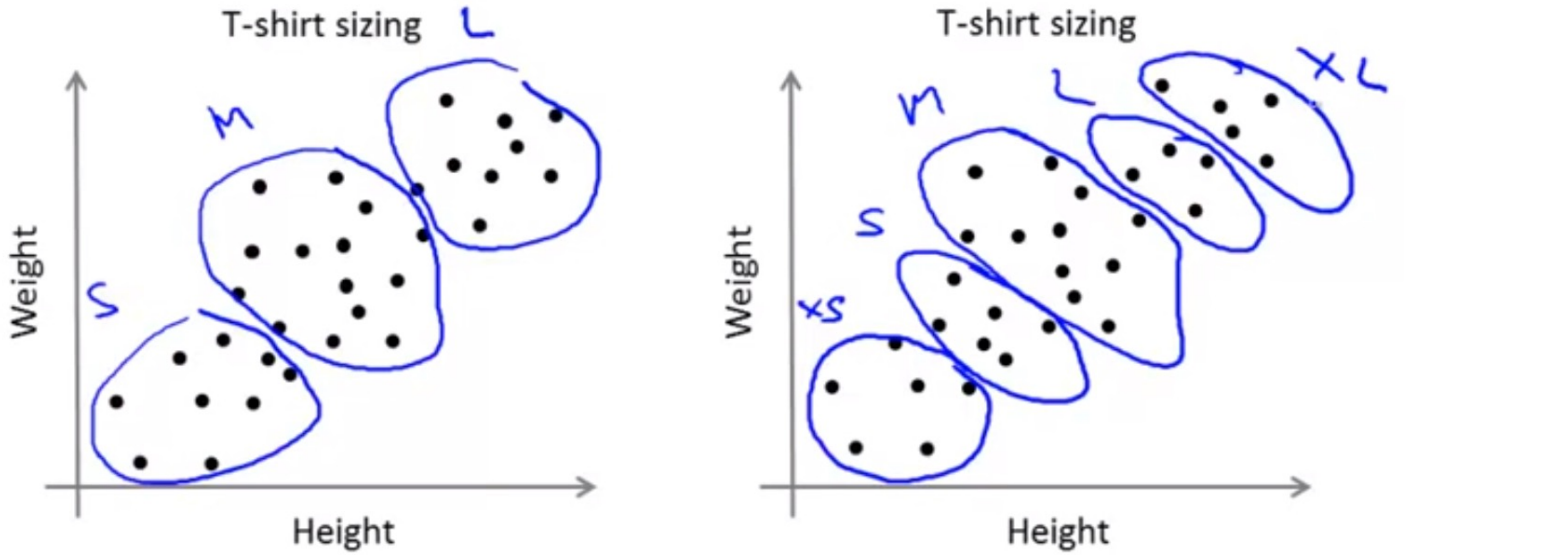
Choosing K

Sometimes, you're running K-means to get clusters to use for some later/downstream purpose. Evaluate K-means based on a metric for how well it performs for that later purpose.

K=3 S, M, L

K=5 XS, S, M, L, XL

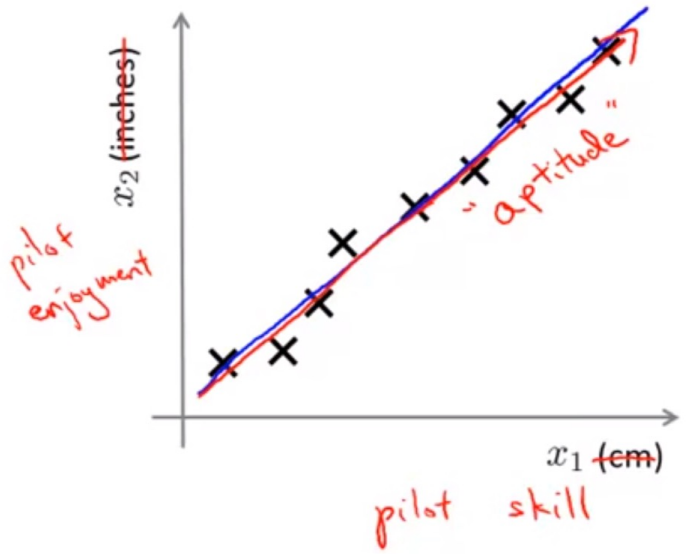
E.g.



Another Type of Unsupervised Learning

Dimensionality reduction

- Why? **Data compression** – allows you to speed up analysis and use up less computer memory



Reduce data from 2D to 1D

https://www.youtube.com/watch?v=Zbr5hyJNGCs&list=PLLssT5z_DsK-h9vYZkQkYNWcltqhIRJLN&index=81

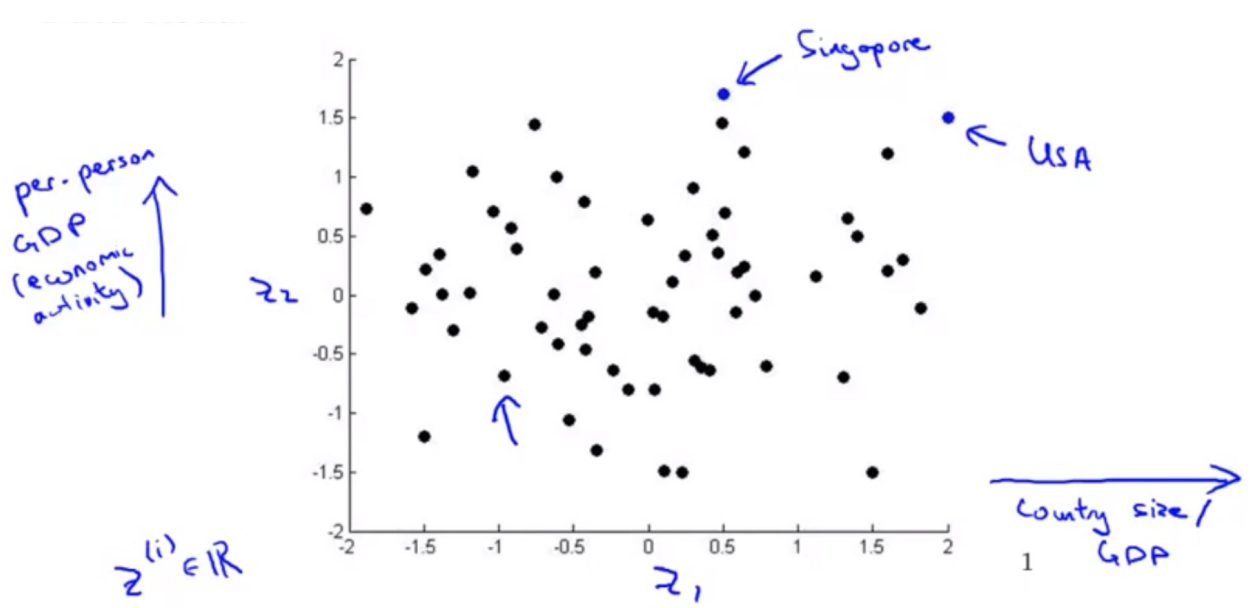
Another Type of Unsupervised Learning

Dimensionality reduction

- Why? **Data visualization** – easier to visualize a dataset with 5 dimensions (or variables) than one with 50 dimensions

Country	x_1 GDP (trillions of US$)	x_2 Per capita GDP (thousands of intl. $)	x_3 Human Develop-ment Index	x_4 Life expectancy	x_5 Poverty Index (Gini as percentage)	x_6 Mean household income (thousands of US$)	...
Canada	1.577	39.17	0.908	80.7	32.6	67.293	...
China	5.878	7.54	0.687	73	46.9	10.22	...
India	1.632	3.41	0.547	64.7	36.8	0.735	...
Russia	1.48	19.84	0.755	65.5	39.9	0.72	...
Singapore	0.223	56.69	0.866	80	42.5	67.1	...
USA	14.527	46.86	0.91	78.3	40.8	84.3	...
...	...	...	...	...	...	...	

Country	z_1	z_2
Canada	1.6	1.2
China	1.7	0.3
India	1.6	0.2
Russia	1.4	0.5
Singapore	0.5	1.7
USA	2	1.5
...	...	...

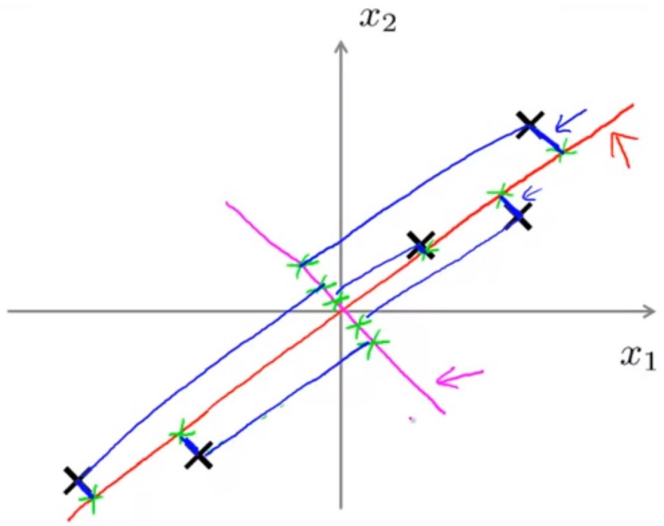


Another Type of Unsupervised Learning

How to reduce the dimensionality of the data?

- Principal Components Analysis (PCA)
  - In most cases, the goal is to reduce from n-dimension to k-dimension, meaning you will need to find k vectors onto which to project the data, as to minimize the projection error

From 2D to 1D



https://www.youtube.com/watch?v=T-B8muDvzu0&list=PLLssT5z_DsK-h9vYZkQkYNWcItqhlRJLN&index=83

Other Popular Types of Unsupervised Learning

Anomaly detection

- Which observations fall out of the discovered “regular pattern” and use it as an input in supervised learning (e.g., amount spent and fraud detection)

Similarity matching

- Recommending products or services to customers based on their similarity to other customers (e.g., Netflix movie recommendation)

Co-occurrence grouping

- Grouping things together based on their co-occurrences (e.g., Target placing products that are bought together next to each other on shelves)

https://www.youtube.com/watch?v=T-B8muDvzu0&list=PLLssT5z_DsK-h9vYZkQkYNWcItqhlRJLN&index=83

Common ML Tasks

Task	Supervised ML	Unsupervised ML
Classification	X	
Regression	X	
Clustering		X
Dimensionality Reduction		X
Anomaly Detection		X
Similarity Matching		X
Co-occurrence Grouping		X

A Final Note on ML for Business

When considering ML (supervised or unsupervised), think about:

- The **situation**
  - Did customers started to churn more recently?
  - Is fake news on the rise?
- The **opportunity**
  - Ability to reduce churn by X %, which could save USD Y in revenue
  - Can decrease the diffusion of fake news by X %, which could improve credibility and attract new users to the platform
- The **action(s)**
  - Identify and improve churn/fake news drivers
  - Detect customers/users at risk and introduce targeted campaigns

At Home Exercise

Review your writing from last class …

- … a well-defined question you have about it
  - **Is it a question of inference, prediction, or pattern understanding?**
- … what kind of data you will need, and where you can find such data
  - **Do you have a target variable? If yes, what is it? If not, should you find one?**
- … how you may need to wrangle the data
- … the ways you may explore and better understand the data
- … the model you may need to build, evaluate, and test
  - **Is it supervised (regression / classification) or unsupervised ML?**
- … the practical value of your model, and how to deploy it
  - **What are the opportunities and actions related to your question?**

Thank You!