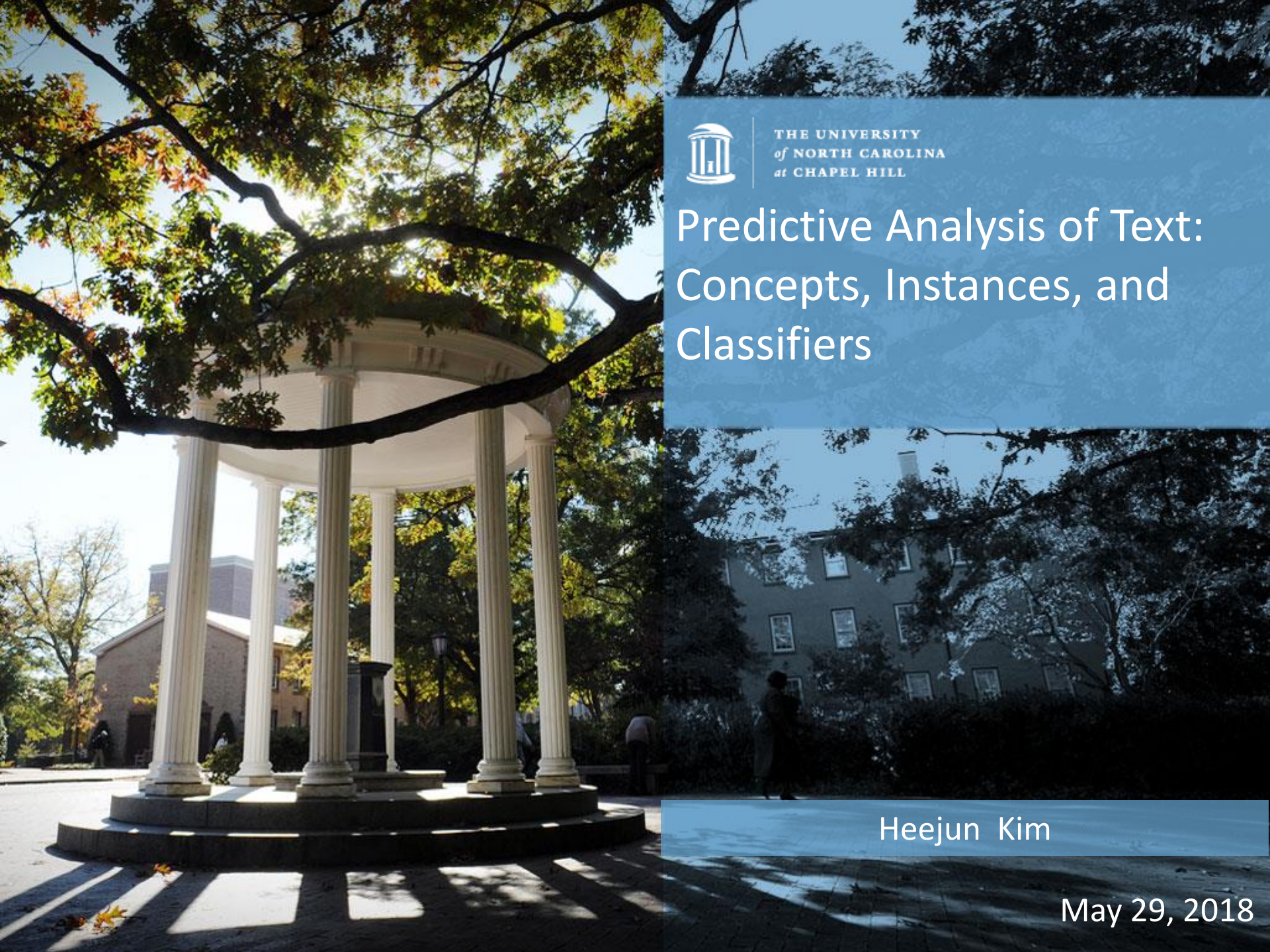


THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Predictive Analysis of Text: Concepts, Instances, and Classifiers

Heejun Kim

May 29, 2018

Predictive Analysis of Text

- **Objective:** developing computer programs that automatically predict a particular concept within a span of text

Predictive Analysis: Procedure

Performance Test

Test Data

Model

color	size	sides	equal sides	...	label
red	big	3	no	...	yes
green	big	3	yes	...	yes
blue	small	inf	yes	...	no
blue	small	4	yes	...	no
⋮	⋮	⋮	⋮	⋮	⋮
red	big	3	yes	...	yes

Training Data

Representation

color	size	sides	equal sides	...	label
red	big	3	no	...	yes
green	big	3	yes	...	yes
blue	small	inf	yes	...	no
blue	small	4	yes	...	no
⋮	⋮	⋮	⋮	⋮	⋮
red	big	3	yes	...	yes

Learning Algorithm

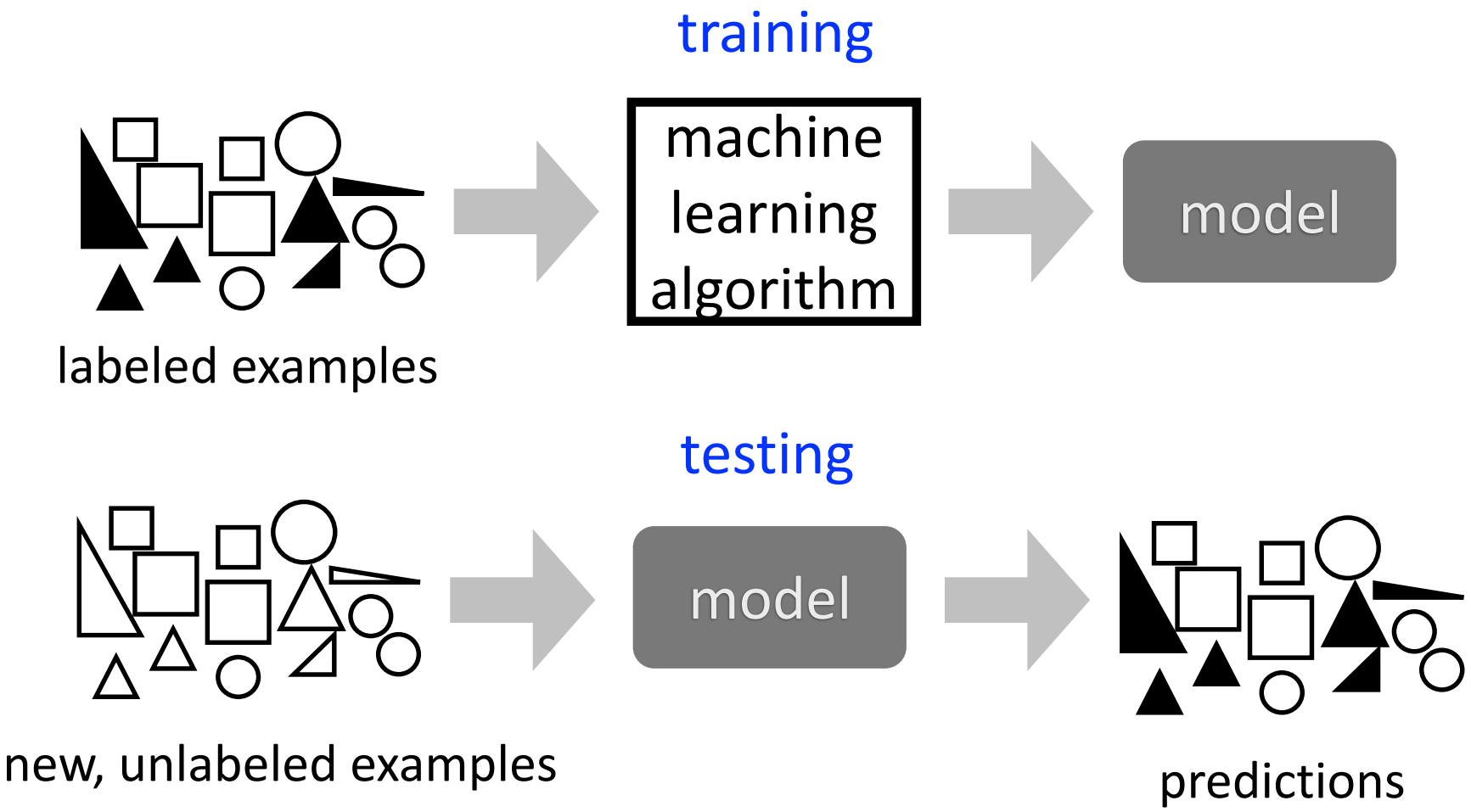
Predictive Analysis:
basic ingredients

- **Training data**: a set of examples of the labeled concept we want to automatically recognize
- **Representation**: a set of features that we believe are useful in recognizing the desired concept
- **Learning algorithm**: a computer program that uses the training data to learn a predictive model of the concept

Predictive Analysis:
basic ingredients

- **Model**: a function that describes a predictive relationship between feature values and the presence/absence of the concept
- **Test data**: a set of previously unseen examples used to estimate the model's effectiveness
- **Performance metrics**: a set of statistics used measure the predictive effectiveness of the model

Predictive Analysis:
training and testing



Predictive Analysis:
concept, instances, and features

features concept

instances

color	size	# slides	equal sides	...	label
red	big	3	no	...	yes
green	big	3	yes	...	yes
blue	small	inf	yes	...	no
blue	small	4	yes	...	no
⋮	⋮	⋮	⋮	⋮	⋮
red	big	3	yes	...	yes

Predictive Analysis:
Type of features

- Nominal: values that are distinct symbols (e.g., male and female). No ordering or distance.
- Numeric
  - Ordinal: ranked order of the categories (e.g., hot, mild, and cool). No distance.
  - Interval: ordered and measured in fixed and equal units (e.g., temperature and school year). 0 is arbitrary.
  - Ratio: measurement method inherently defines a zero point (e.g., distance). Ordered and measured in fixed and equal units.

Predictive Analysis:
training and testing

color	size	# slides	Equal sides	...	label
red	big	3	no	...	yes
green	big	3	yes	...	yes
blue	small	inf	yes	...	no
blue	small	4	yes	...	no
⋮	⋮	⋮	⋮	⋮	⋮
red	big	3	yes	...	yes

labeled examples

training

machine learning algorithm

model

color	size	# slides	Equal sides	...	label
red	big	3	no	...	?
Green	big	3	yes	...	?
blue	small	inf	yes	...	?
blue	small	4	yes	...	?
⋮	⋮	⋮	⋮	⋮	⋮
red	big	3	yes	...	?

new, unlabeled examples

testing

model

color	size	# slides	Equal sides	...	label
red	big	3	no	...	yes
green	big	3	yes	...	yes
blue	small	inf	yes	...	no
blue	small	4	yes	...	no
⋮	⋮	⋮	⋮	⋮	⋮
red	big	3	yes	...	yes

predictions

Predictive Analysis: questions

- Is a particular concept appropriate for predictive analysis?
- What should the unit of analysis be?
- How should I divide the data into training and test sets?
- What is a good feature representation for this task?
- What type of learning algorithm should I use?
- How should I evaluate my model's performance?

Predictive Analysis:
Concepts

- Learning algorithms can recognize some concepts better than others
- What are some properties of concepts that are easier to recognize?

Predictive Analysis:
Concepts

- Option 1: can a human recognize the concept?
- Option 2: can two or more humans recognize the concept independently and do they agree?
- Option 2 is better.
- In fact, models are sometimes evaluated as an independent assessor
- How does the model's performance compare to the performance of one assessor with respect to another?
  - One assessor produces the "ground truth" and the other produces the "predictions"

Predictive Analysis:

measures agreement: percent agreement

- **Percent agreement:** percentage of instances for which both assessors agree that the concept occurs or does not occur

$$\frac{(A + D)}{(A + B + C + D)}$$

	yes	no
yes	A	B
no	C	D

Predictive Analysis:

measures agreement: percent agreement

- Problem: percent agreement does not account for agreement due to random chance.
- How can we compute the expected agreement due to random chance?

Predictive Analysis:

measures agreement: percent agreement

- **Percent agreement:**

$$\frac{(80 + 10)}{(80 + 5 + 5 + 10)}$$

	yes	no
yes	80	5
no	5	10

- **Agreement due to random chance?**

Predictive Analysis:

measures agreement: percent agreement

- How can we compute the expected agreement due to random chance?
- **Kappa agreement:** percent agreement after correcting for the expected agreement due to chance (not covered in this course)
- For more details, refer to Wikipedia article or online video

Predictive Analysis: questions

- Is a particular concept appropriate for predictive analysis?
- What should the unit of analysis be?
- How should I divide the data into training and test sets?
- What is a good feature representation for this task?
- What type of learning algorithm should I use?
- How should I evaluate my model's performance?

Predictive Analysis:

turning data into training and test instances

- For many text-mining applications, turning the data into instances for training and testing is fairly straightforward
- Easy case: instances are self-contained, independent units of analysis
- topic categorization: instances = documents
- opinion mining: instances = product reviews
- bias detection: instances = political blog posts
- emotion detection: instances = support group posts

Topic Categorization:
predicting health-related documents

features

concept

instances

w_1	w_2	w_3	…	w_n	label
1	1	0	…	0	health
0	0	0	…	0	other
0	0	0	…	0	other
0	1	0	…	1	other
⋮	⋮	⋮	…	0	⋮
1	0	0	…	1	health

Opinion Mining

predicting positive/negative movie reviews

features

concept

instances

w_1	w_2	w_3	...	w_n	label
1	1	0	...	0	positive
0	0	0	...	0	negative
0	0	0	...	0	negative
0	1	0	...	1	negative
⋮	⋮	⋮	...	0	⋮
1	0	0	...	1	positive

Bias Detection

predicting liberal/conservative blog posts

features

concept

instances

w_1	w_2	w_3	…	w_n	label
1	1	0	…	0	liberal
0	0	0	…	0	conservative
0	0	0	…	0	conservative
0	1	0	…	1	conservative
⋮	⋮	⋮	…	0	⋮
1	0	0	…	1	liberal

Predictive Analysis: questions

- Is a particular concept appropriate for predictive analysis?
- What should the unit of analysis be?
- How should I divide the data into training and test sets?
- What is a good feature representation for this task?
- What type of learning algorithm should I use?
- How should I evaluate my model's performance?

Predictive Analysis:
training and test data

- We want our model to “learn” to recognize a concept
- So, what does it mean to learn?

Predictive Analysis:
training and test data

- The machine learning definition of learning:

> A machine *learns* with respect to a particular task T, performance metric P, and experience E, if the system improves its performance P at task T following experience E.
>
> -- Tom Mitchell

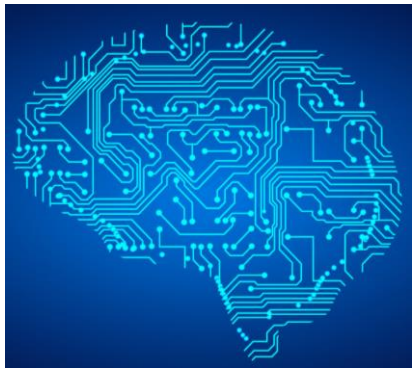
Predictive Analysis:
can we use the same data for testing?

Training Data

training

machine learning algorithm

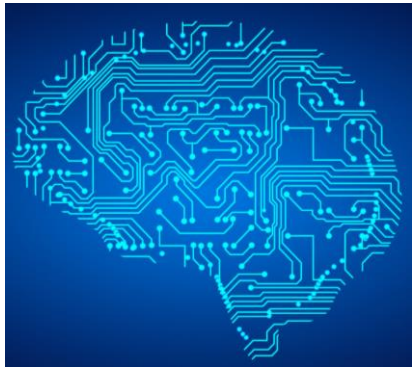
Spam Detection Model



Test Data

testing

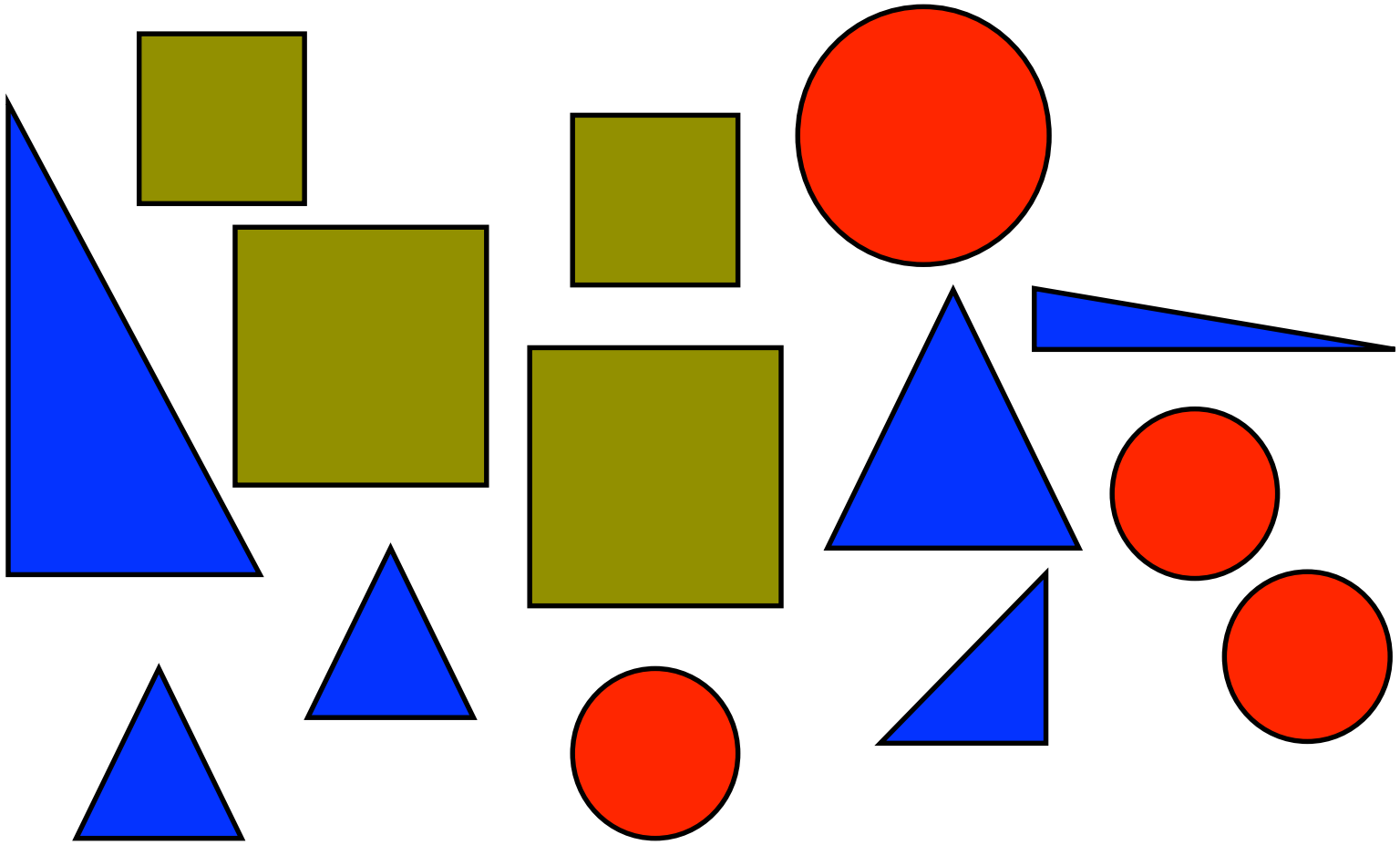
New Data



Predictive Analysis:
training and test data

- We want our model to improve its generalization performance!
- That is, its performance on previously unseen data!
- Generalize: to derive or induce a general conception or principle from particulars. -- Merriam-Webster
- In order to test generalization performance, the training and test data cannot be the same.
- Why?

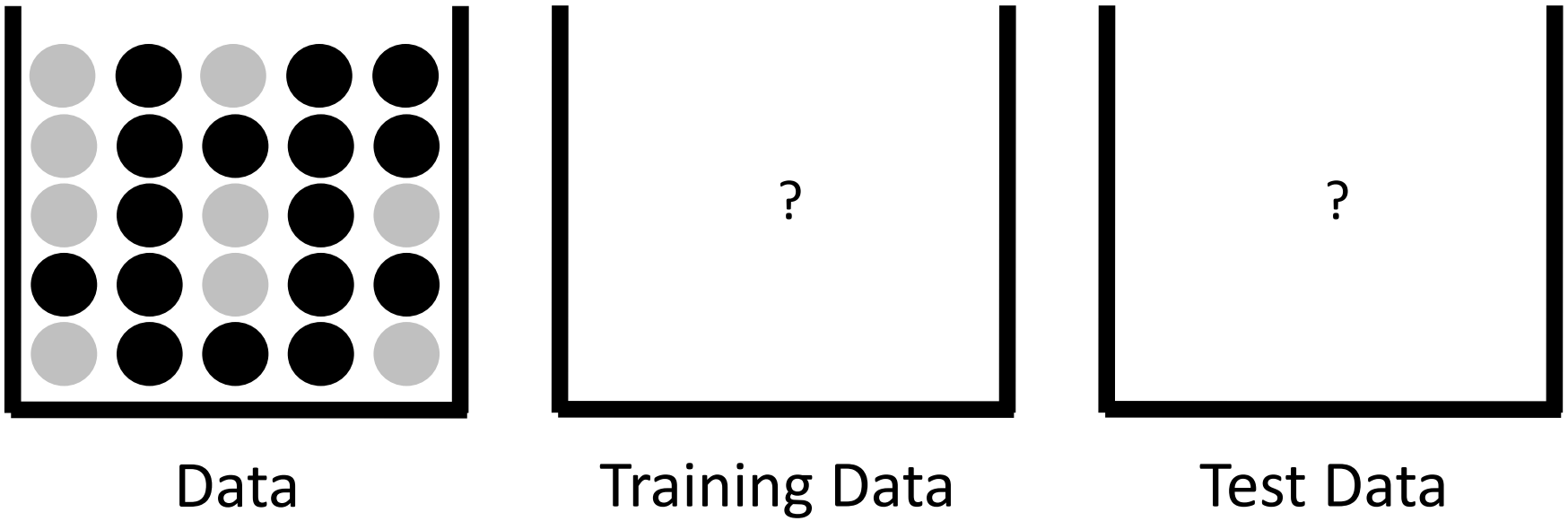
Training data + Representation:
what could possibly go wrong?



Predictive Analysis:
training and test data

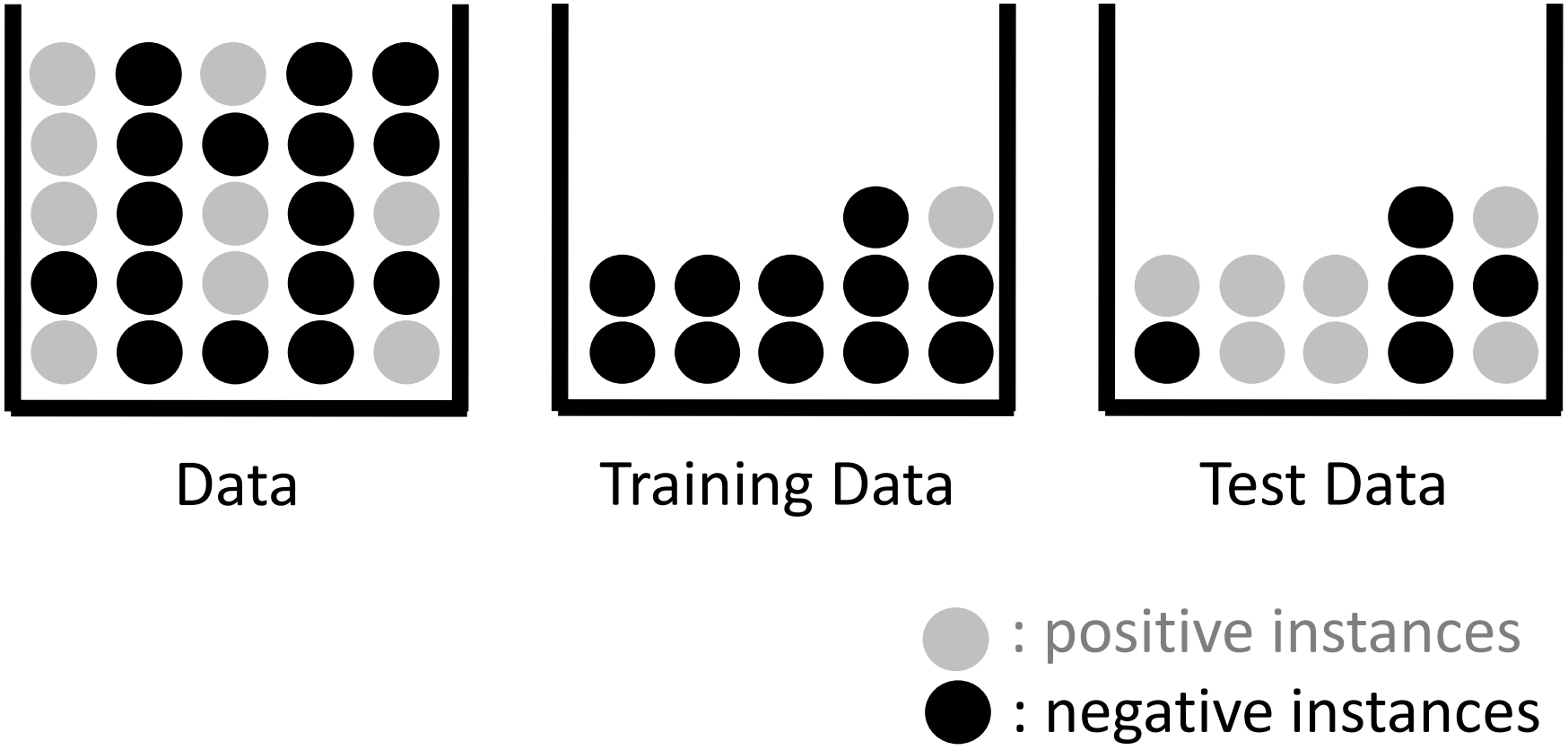
- While we don't want to test on training data, we want to have training and test set that are derived from the same "probability distribution".
- What does that mean?

Predictive Analysis:
training and test data

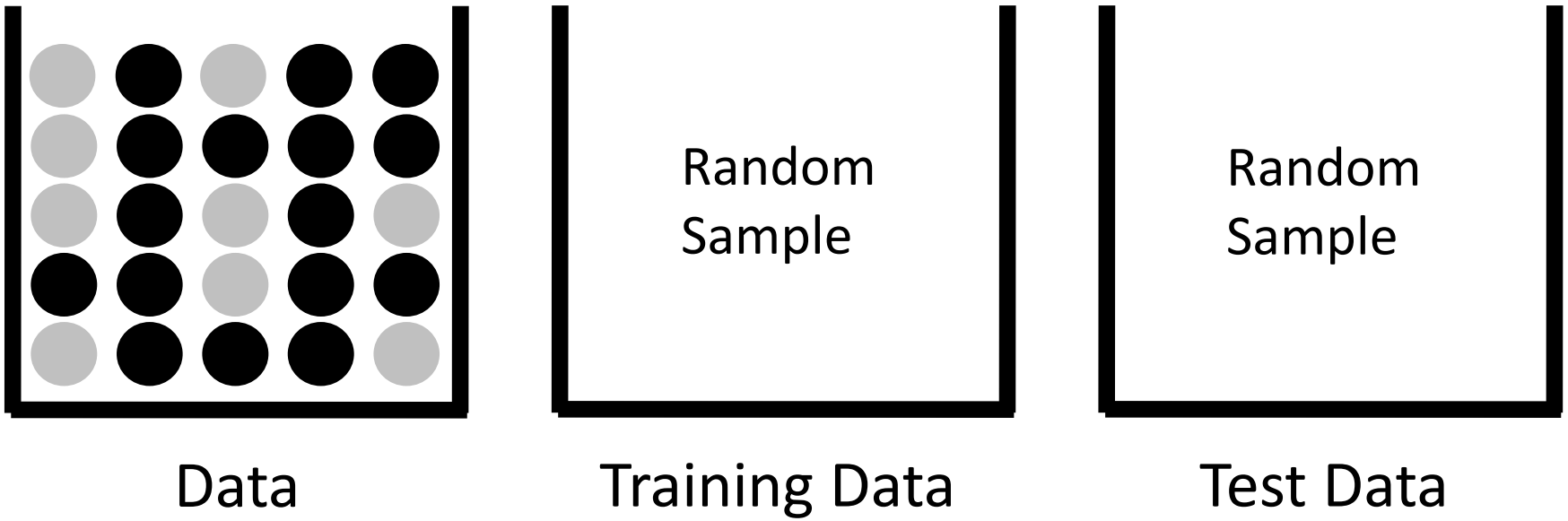


Predictive Analysis:
training and test data

- Is this a good partitioning? Why or why not?

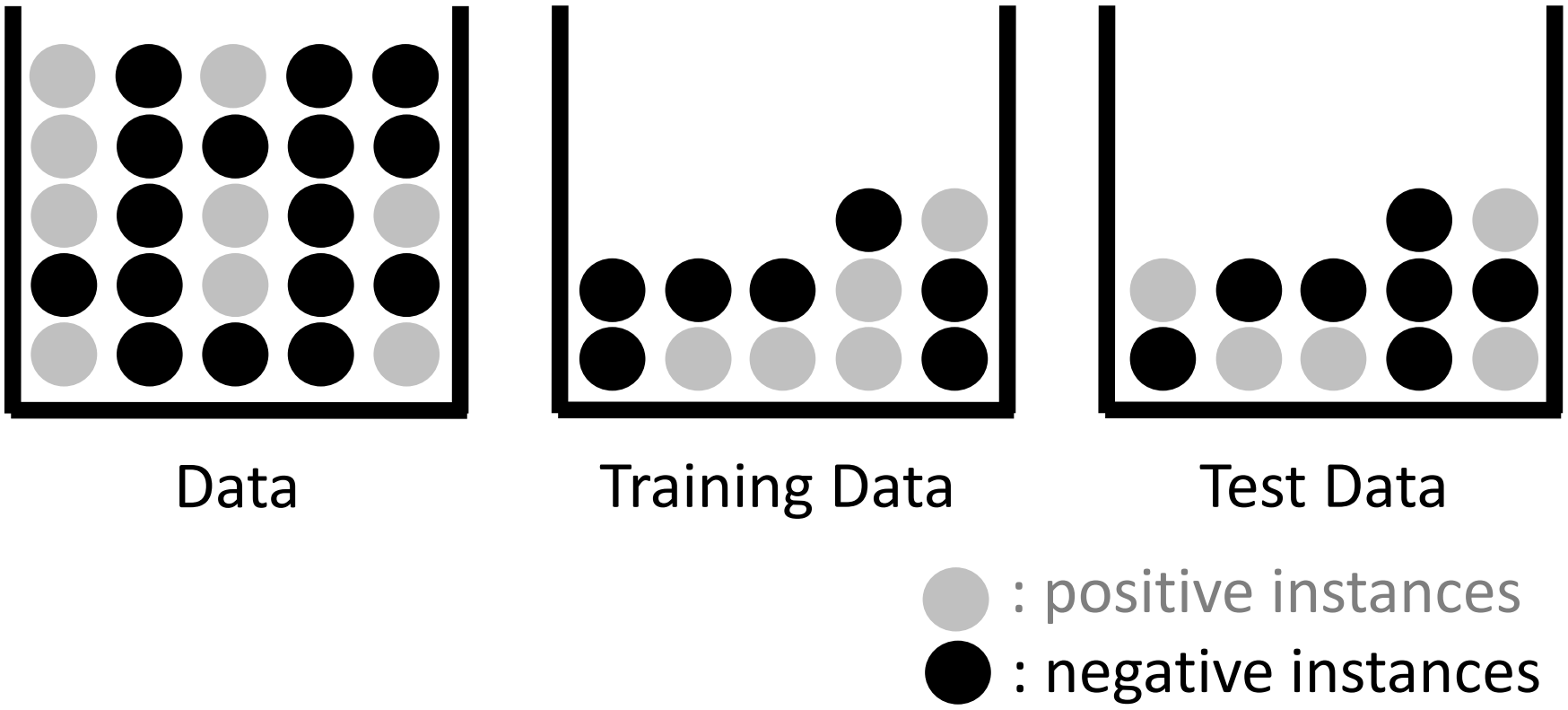


Predictive Analysis:
training and test data

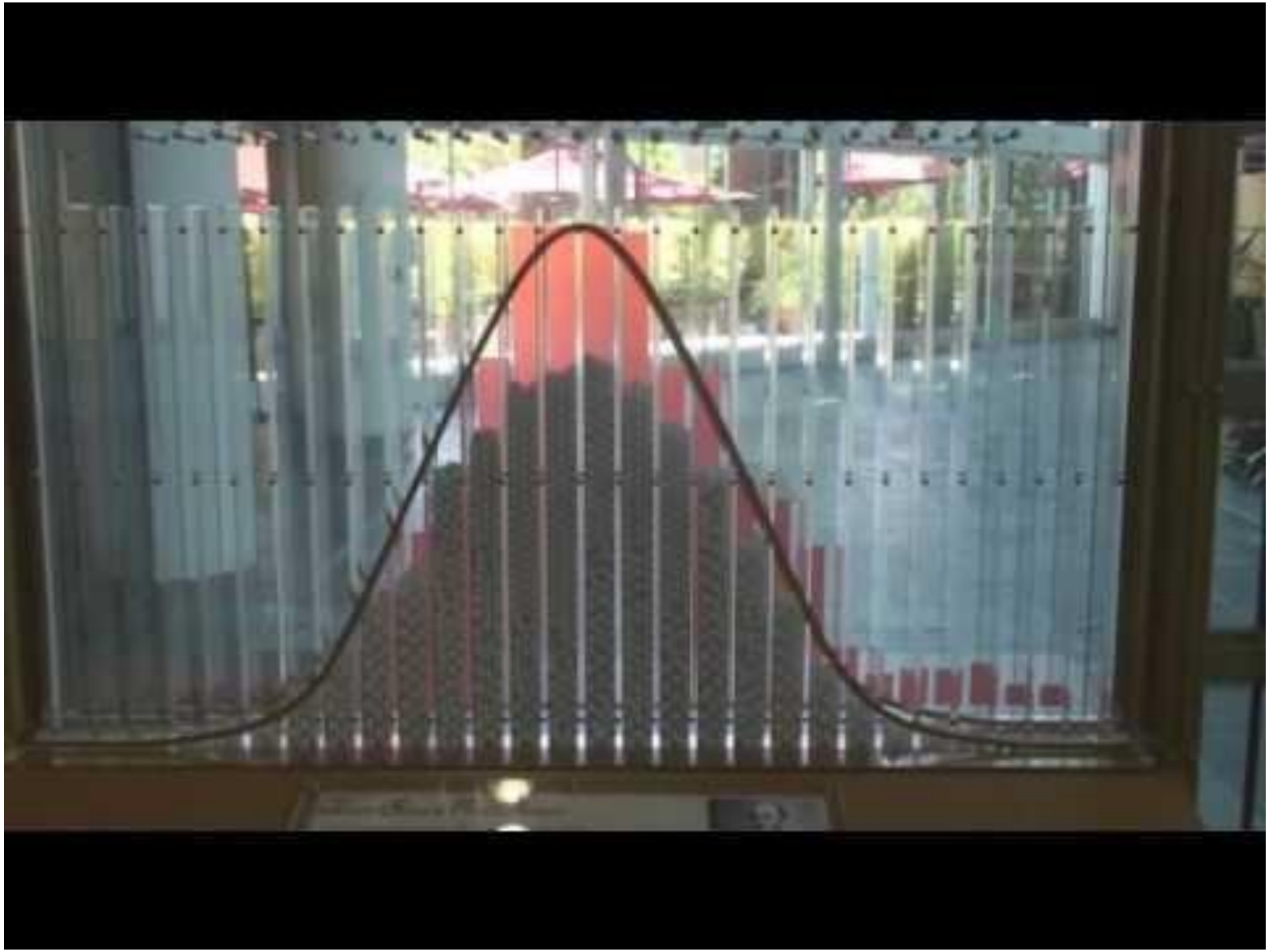


Predictive Analysis:
training and test data

- On average, random sampling should produce comparable data for training and testing

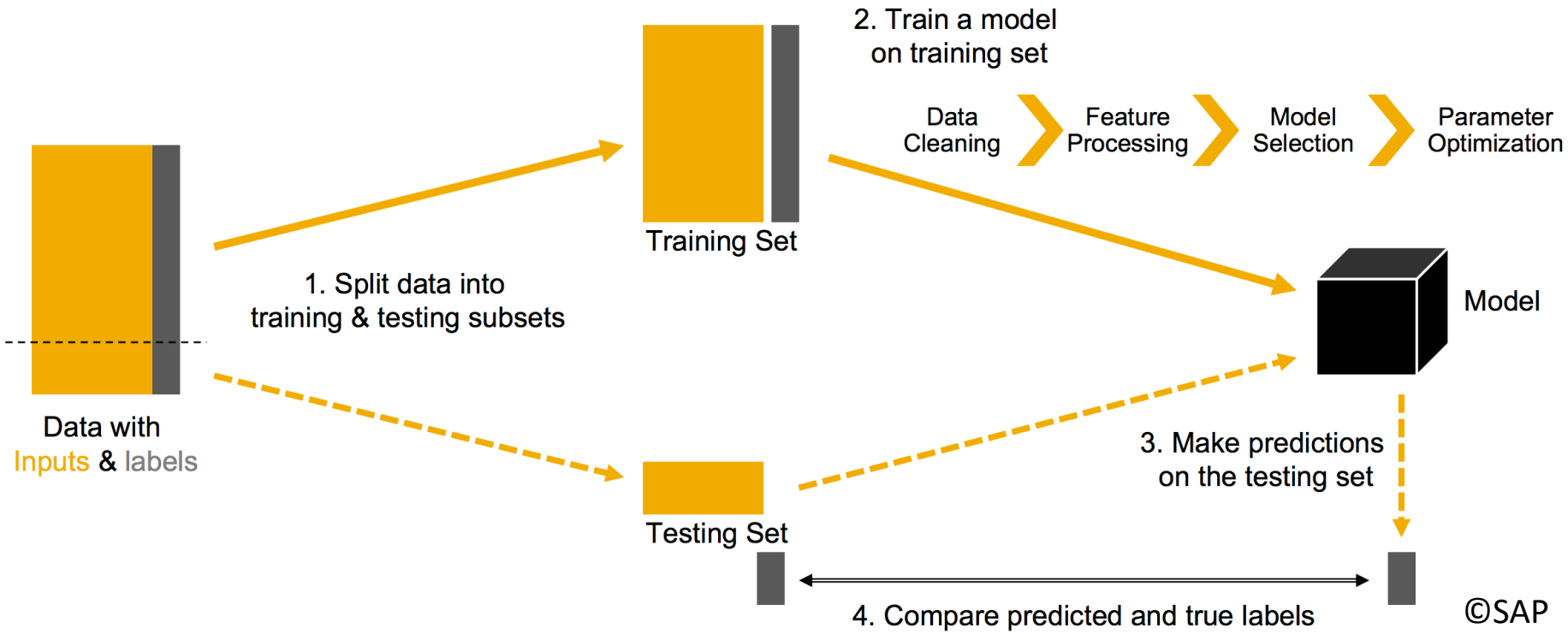


Statistical Estimation



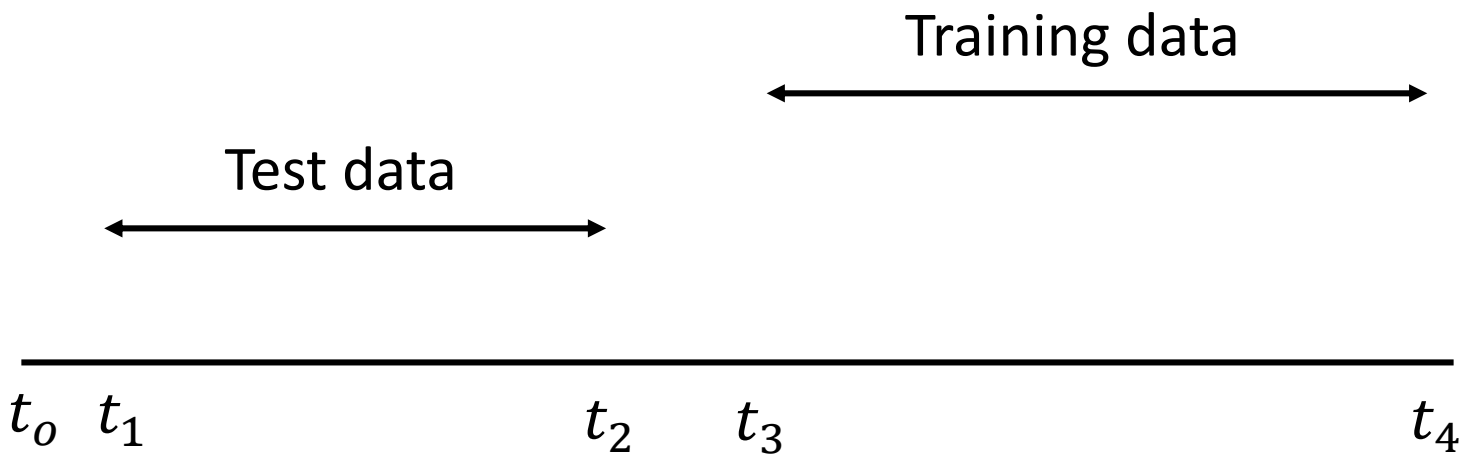
Link

Predictive Analysis:
training and test data



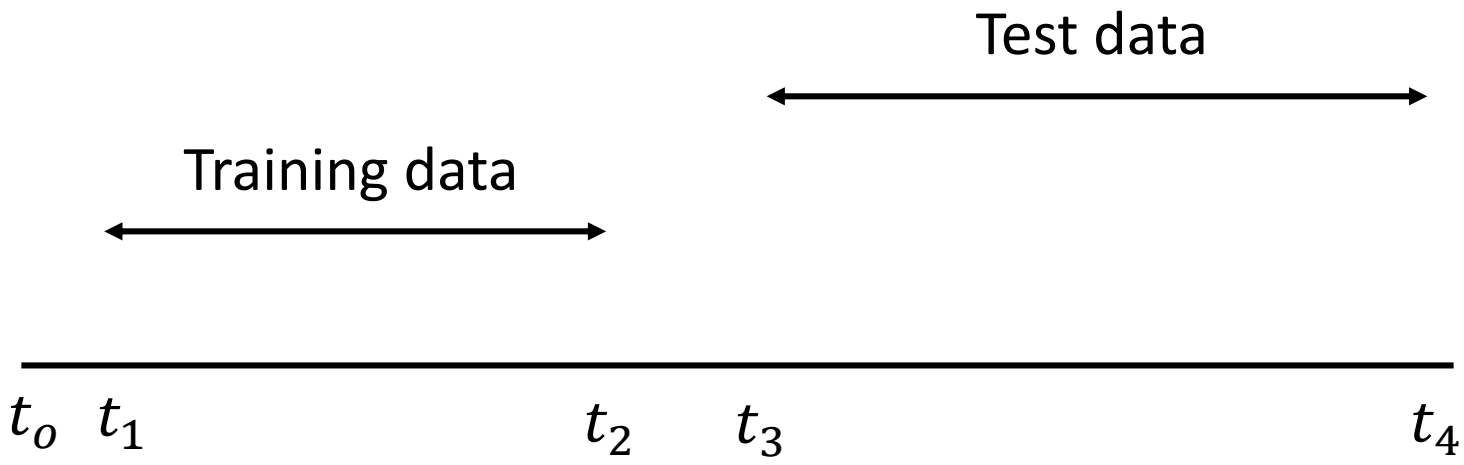
Predictive Analysis:
training and test data

- If you want to predict stock price by analyzing tweets, how the training and test data should be separated?



Predictive Analysis:
training and test data

- If you want to predict stock price by analyzing tweets, how the training and test data should be separated?



Predictive Analysis:
training and test data

- Models usually perform the best when the training and test set have:
  - a similar proportion of positive and negative examples
  - a similar co-occurrence of feature-values and each target class value

Predictive Analysis:
training and test data

- Caution: in some situations, partitioning the data randomly might inflate performance in an unrealistic way!
- How the data is split into training and test sets determines what we can claim about generalization performance
- The appropriate split between training and test sets is usually determined on a case-by-case basis

Predictive Analysis: discussion

- Spam detection: should the training and test sets contain email messages from the same sender, same recipient, and/or same timeframe?
- Topic segmentation: should the training and test sets contain potential boundaries from the same discourse?
- Opinion mining for movie reviews: should the training and test sets contain reviews for the same movie?
- Sentiment analysis: should the training and test sets contain blog posts from the same discussion thread?

Predictive Analysis: questions

- Is a particular concept appropriate for predictive analysis?
- What should the unit of analysis be?
- How should I divide the data into training and test sets?
- What is a good feature representation for this task?
- What type of learning algorithm should I use?
- How should I evaluate my model's performance?

Predictive Analysis:
three types of classifiers

- Linear classifiers
- Decision tree classifiers
- Instance-based classifiers

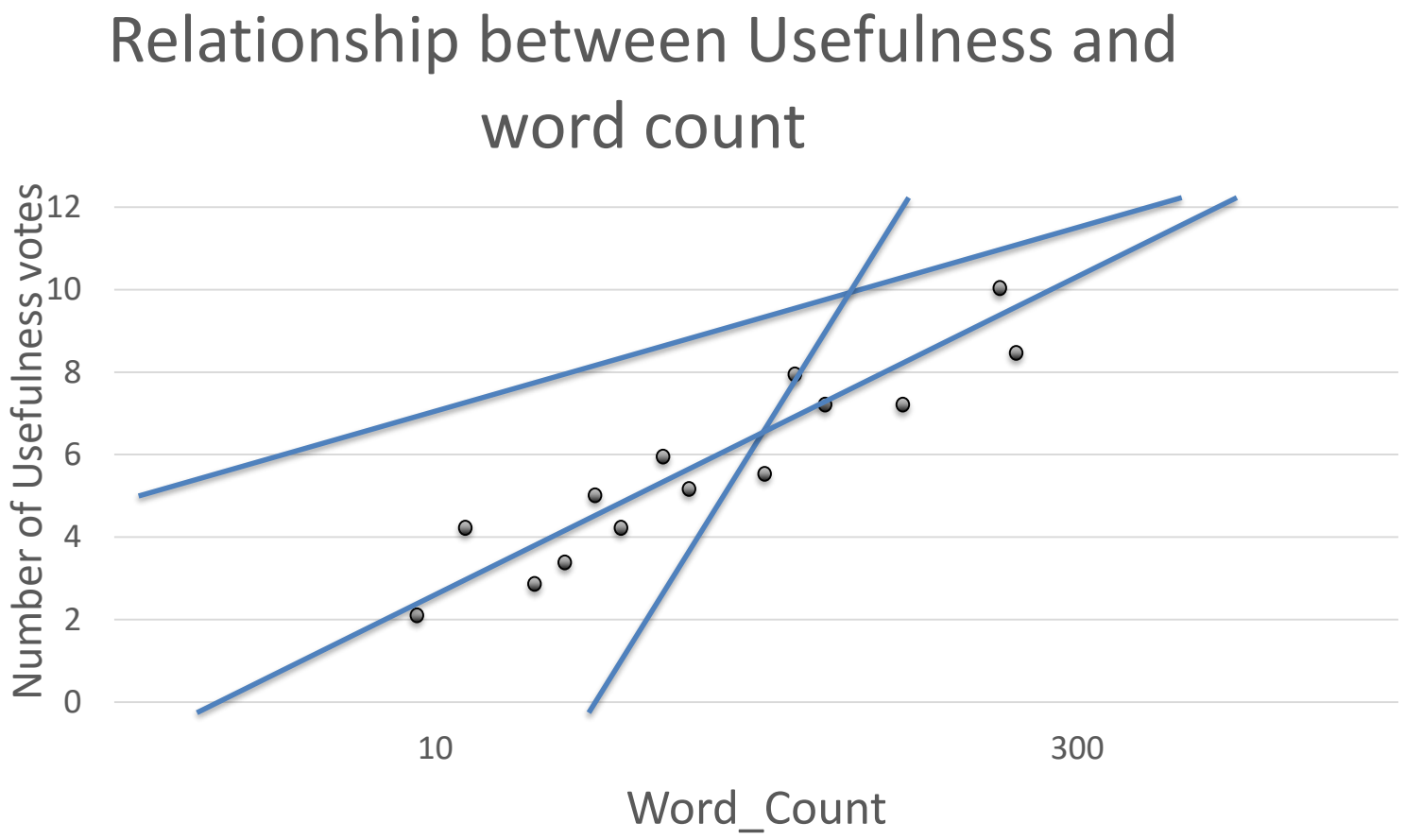
Predictive Analysis:
three types of classifiers

- All types of classifiers learn to make predictions based on the input feature values
- However, different types of classifiers combine the input feature values in different ways

Predictive Analysis:
three types of classifiers

$$y = \begin{cases} 1 & \text{if } w_0 + \sum_{j=1}^{n} w_j x_j > 0 \\ 0 & \text{otherwise} \end{cases}$$

Learning Algorithm + Model:
what could possibly go wrong?



Predictive Analysis

linear classifiers: perceptron algorithm

$$y = \begin{cases} 1 & \text{if } w_0 + \sum_{j=1}^{n} w_j x_j > 0 \\ 0 & \text{otherwise} \end{cases}$$

parameters learned by the model

predicted value (e.g., 1 = positive, 0 = negative)

Predictive Analysis

linear classifiers: perceptron algorithm

$$y = \begin{cases} 1 & \text{if } w_0 + \sum_{j=1}^{n} w_j x_j > 0 \\ 0 & \text{otherwise} \end{cases}$$

test instance

f_1	f_2	f_3
0.5	1	0.2

model weights

w_0	w_1	w_2	w_3
2	-5	2	1

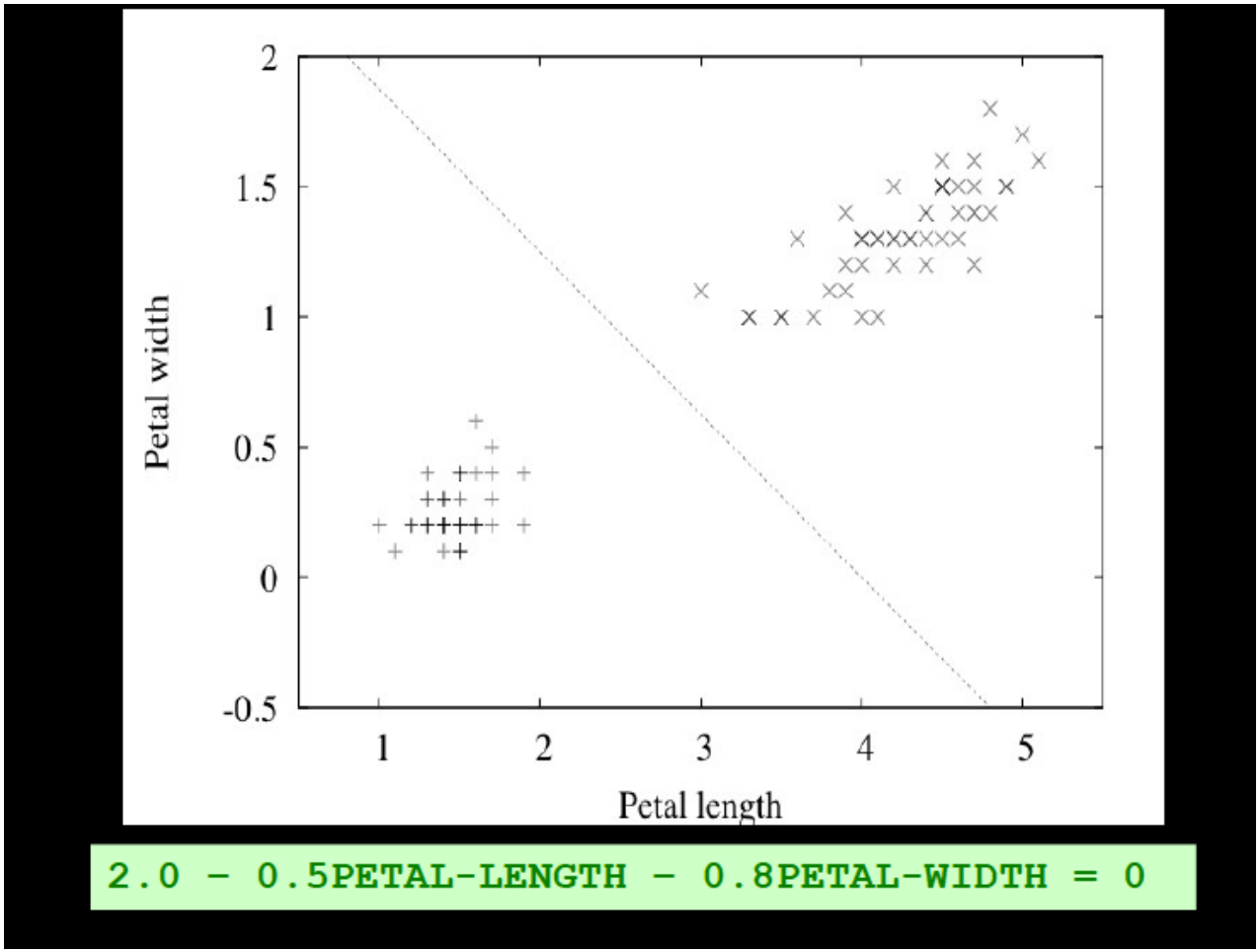
output = 2.0 + (0.50 x -5.0) + (1.0 x 2.0) + (0.2 x 1.0)

output = 1.7

output prediction = positive

Predictive Analysis

linear classifiers: perceptron algorithm



(two-feature example borrowed from Witten *et al.* textbook)

Predictive Analysis
linear classifiers: logistic regression

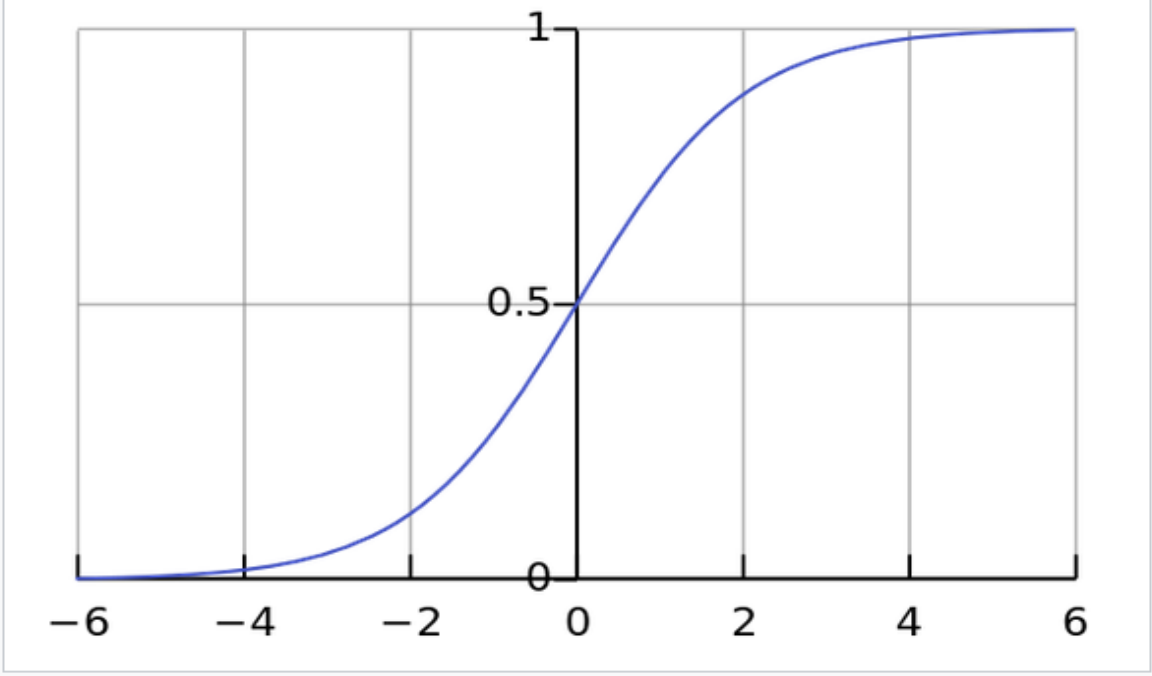


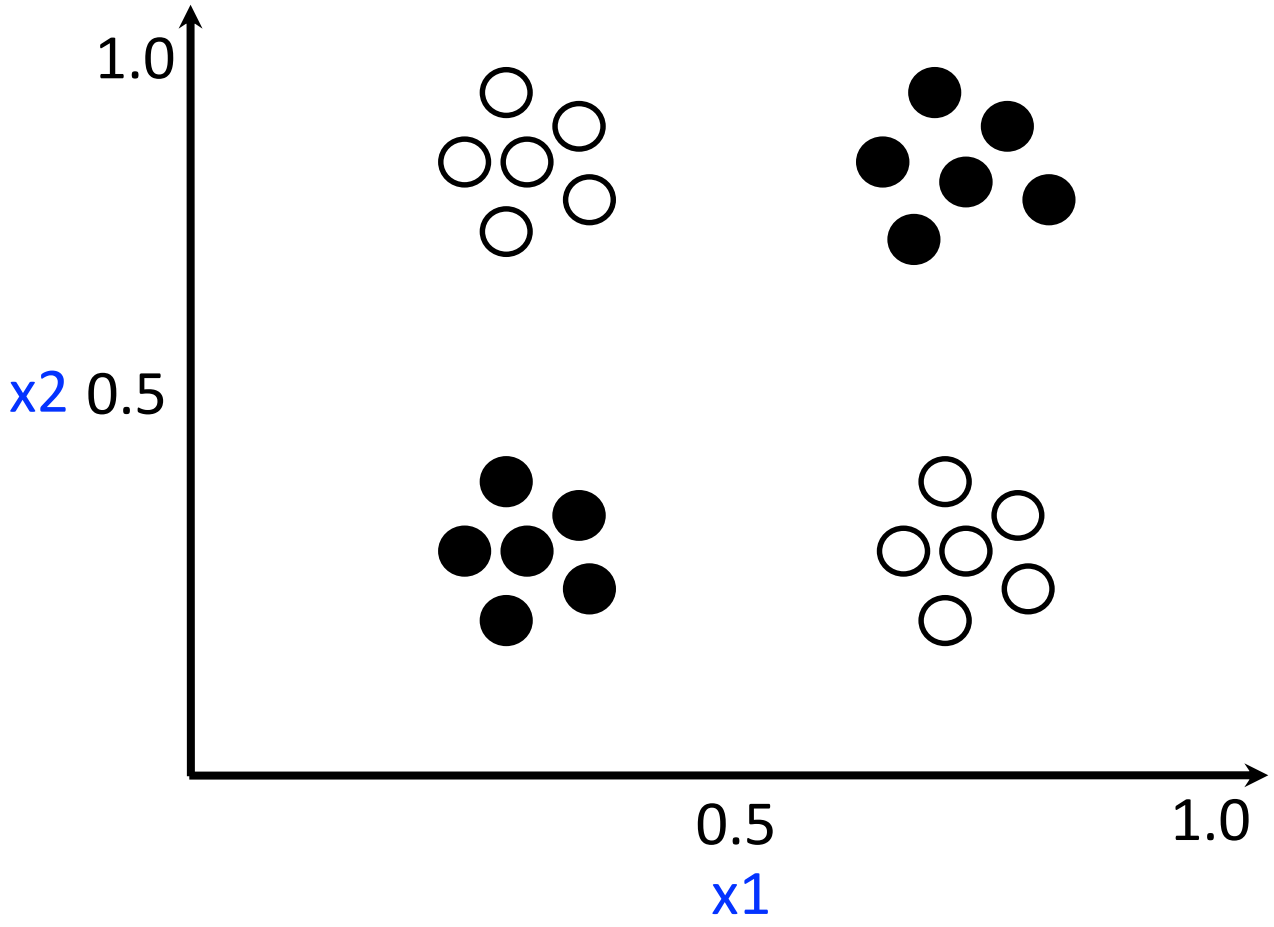
Figure 1. The standard logistic function $\sigma(t)$; note that $\sigma(t) \in (0,1)$ for all t.

$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$

when $t = \beta_0 + \beta_1 x$

(source: https://en.wikipedia.org/wiki/Logistic_regression#/media/File:Logistic-curve.svg)

Predictive Analysis:
would a linear classifier work?

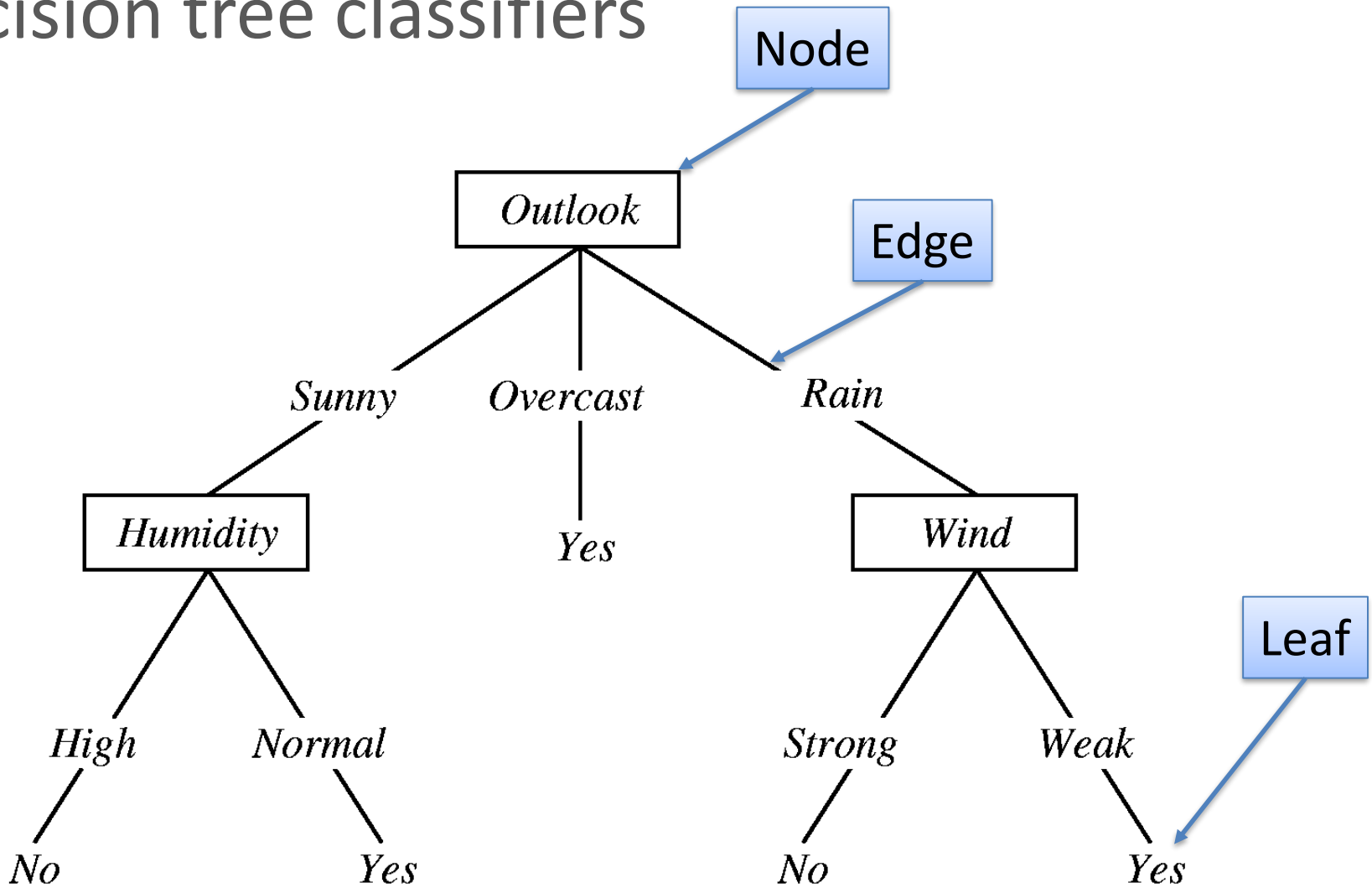


Predictive Analysis:
three types of classifiers

- Linear classifiers
- Decision tree classifiers
- Instance-based classifiers

Predictive Analysis

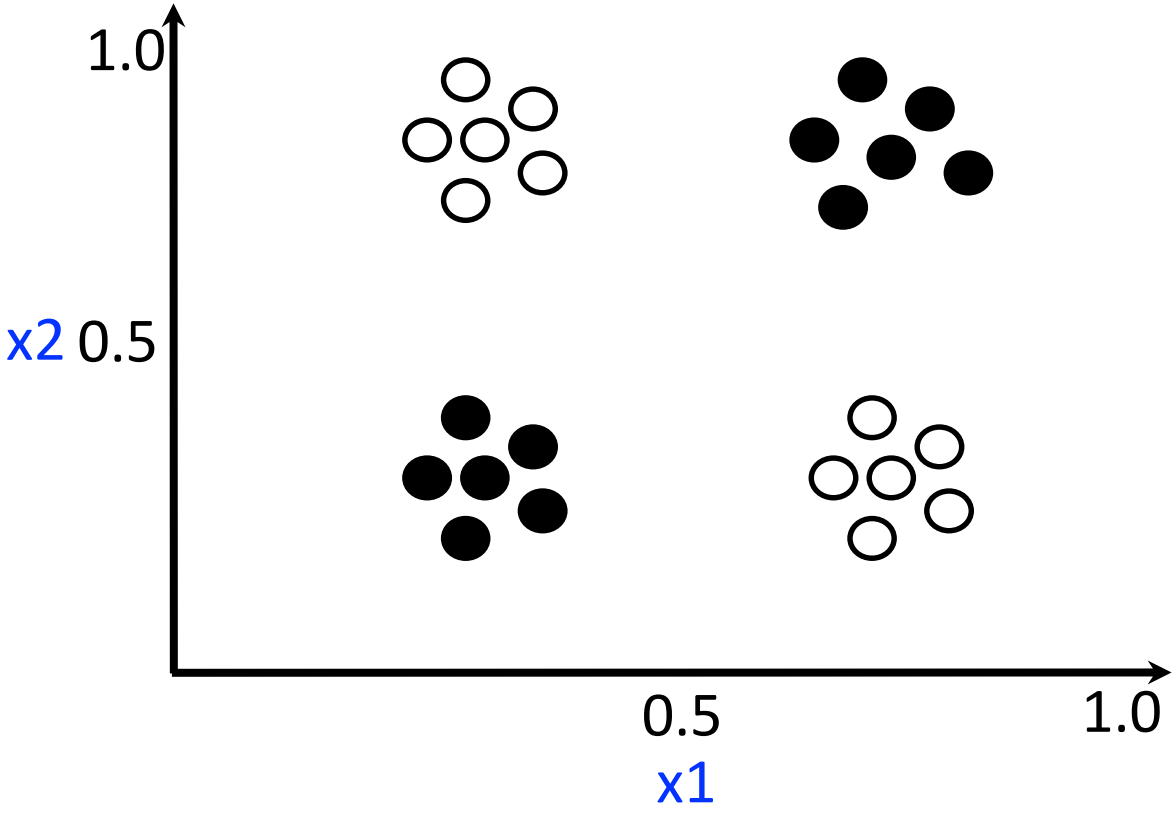
decision tree classifiers



Predictive Analysis
decision tree classifiers

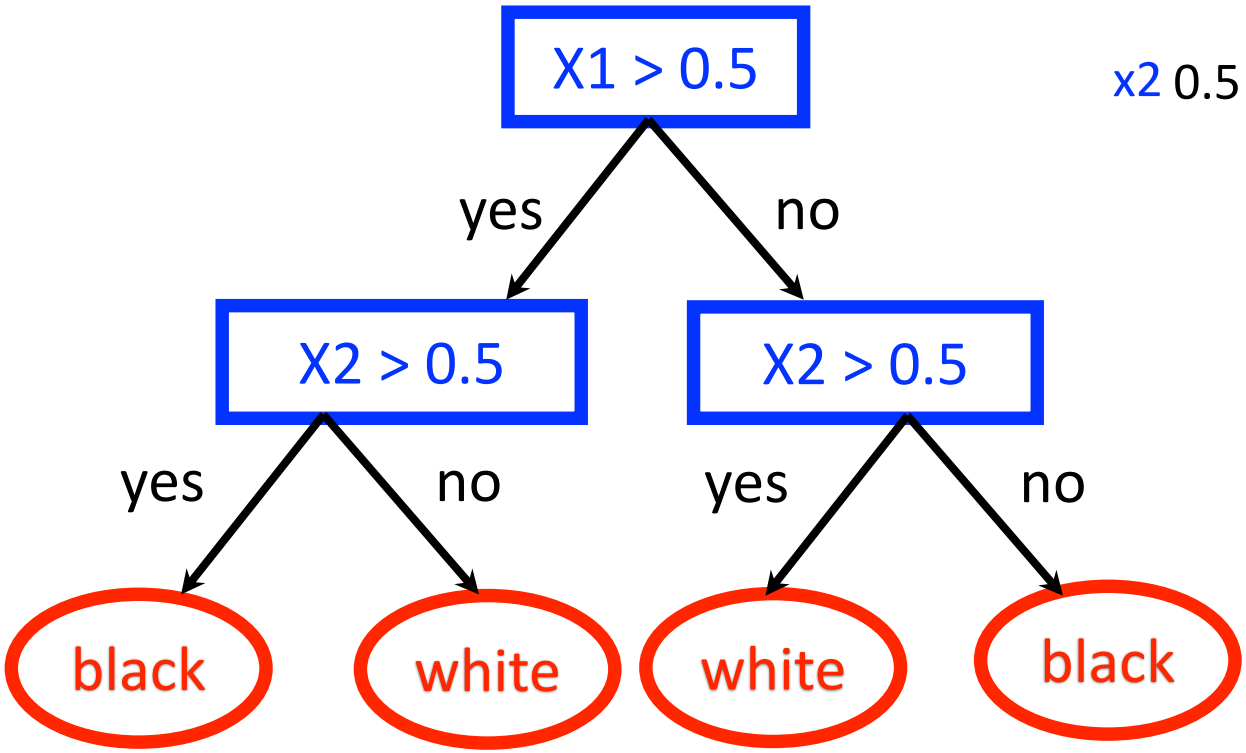
- Decision Tree
  - Special decision rules organized in form of tree data structure that help to understand the relationship among the attributes and class labels.
  - Attributes become nodes, edges are used to represent the values of these attributes, and predictions are made at each leaf.

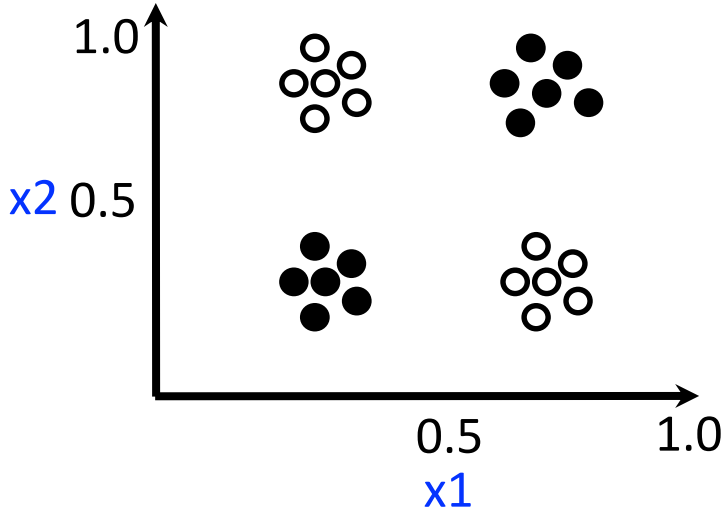
Predictive Analysis:
decision tree classifiers



- Draw a decision tree that would perform perfectly on this training data!

Predictive Analysis:
examples of decision tree classifiers

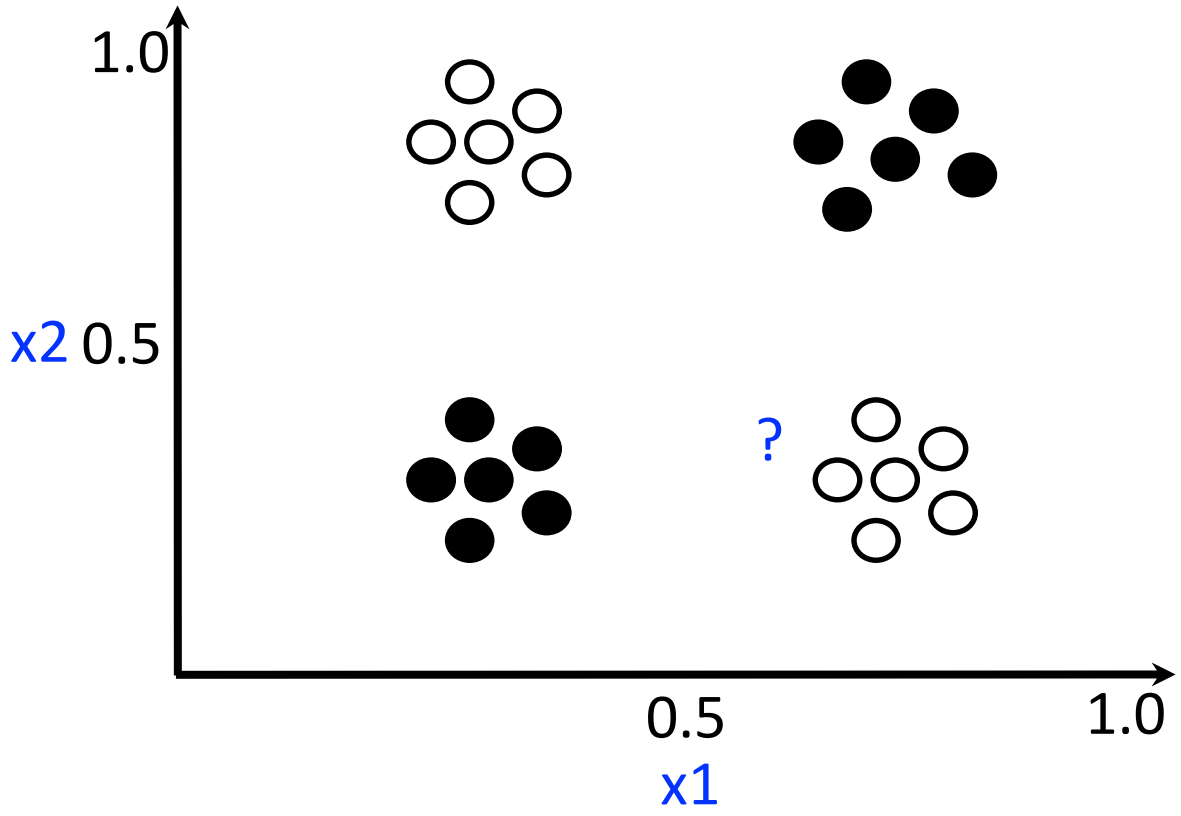




Predictive Analysis:
three types of classifiers

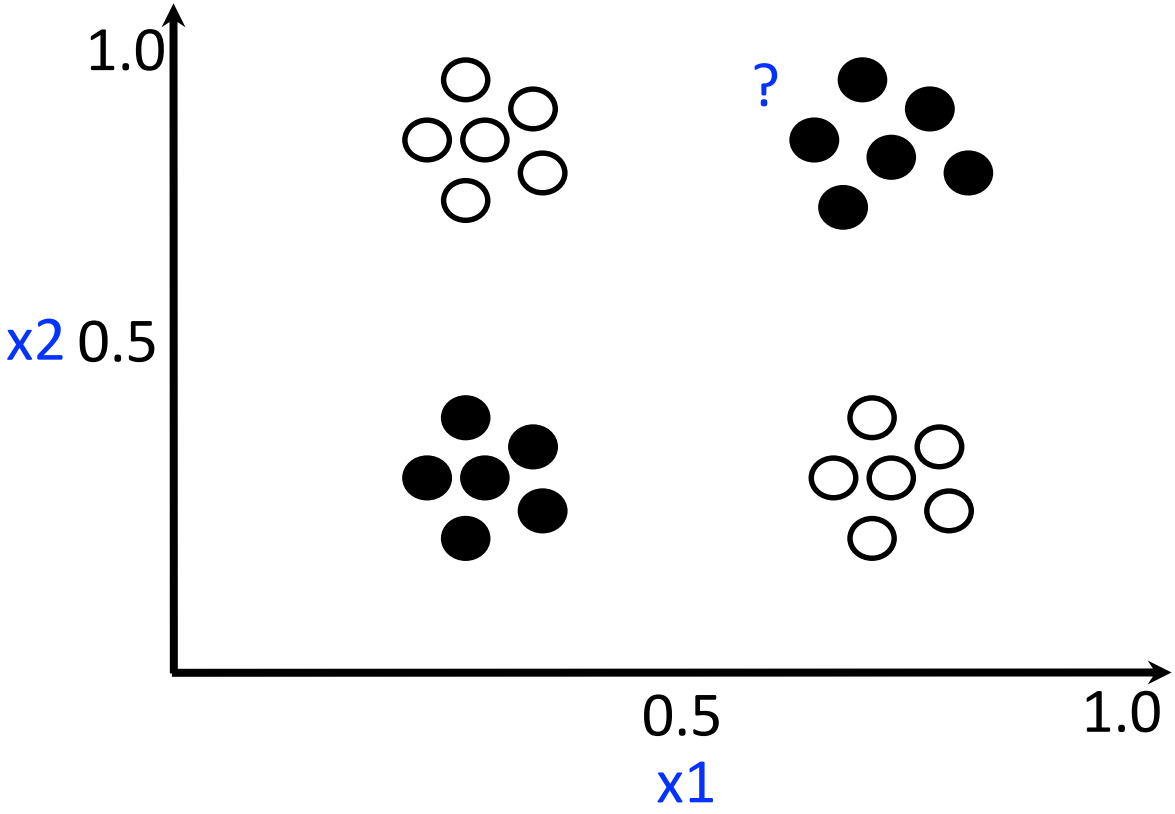
- Linear classifiers
- Decision tree classifiers
- Instance-based classifiers

Predictive Analysis: instance-based classifiers



- predict the class associated with the most similar training examples

Predictive Analysis: instance-based classifiers



- predict the class associated with the most similar training examples

Predictive Analysis: instance-based classifiers

- Assumption: instances with similar feature values should have a similar label
- Given a test instance, predict the label associated with its nearest neighbors
- There are many different similarity metrics for computing distance between training/test instances

Predictive Analysis: questions

- Is a particular concept appropriate for predictive analysis?
- What should the unit of analysis be?
- How should I divide the data into training and test sets?
- What is a good feature representation for this task?
- What type of learning algorithm should I use?
- How should I evaluate my model's performance?

Any Questions?

Text Representation

Next Class