Text Data Mining:
Predictive and Exploratory Analysis of Text

Sandeep Avula
asandeep@unc.edu

August 22, 2018

Outline

Introductions

What is Text Data Mining?

Predictive Analysis of Text: The Big Picture

Exploratory Analysis of Text: The Big Picture

Applications

Introductions

- Hello, my name is _____.
- I'm in the _____ program.
- Prior experience with data?
- Any curious data problems you have thought of?
- I'm taking this course because I'd like to learn how to _____.

Other relevant details

- Course website: https://asandeepc.bitbucket.io/courses/inls613_fall2018/
- Channel of communication: MS Teams. Would you all be ok with that?
- Office hours: By appointment.
- Pronouns: He/Him.
- Assignment submissions: Sakai.

What is Text Data Mining?

- The science and practice of building and evaluating computer programs that automatically detect or discover interesting and useful things in collections of natural language text

Related Fields

- Machine Learning: developing computer programs that improve their performance with "experience"
- Data Mining: developing methods that discover patterns within large structured datasets
- Statistics: developing methods for the interpretation of data and experimental outcomes in reaching conclusions with a certain degree of confidence
- Data Science: Wait, what is this? (I think it is …) An interdisciplinary field which encompasses: ML, Stats, and some form of humanities.

Text Data Mining in this Course

- Predictive Analysis of Text
  - developing computer programs that automatically recognize or detect a particular concept within a span of text
- Exploratory Analysis of Text:
  - developing computer programs that automatically discover interesting and useful patterns or trends in text collections

Outline

Introductions

What is Text Data Mining?

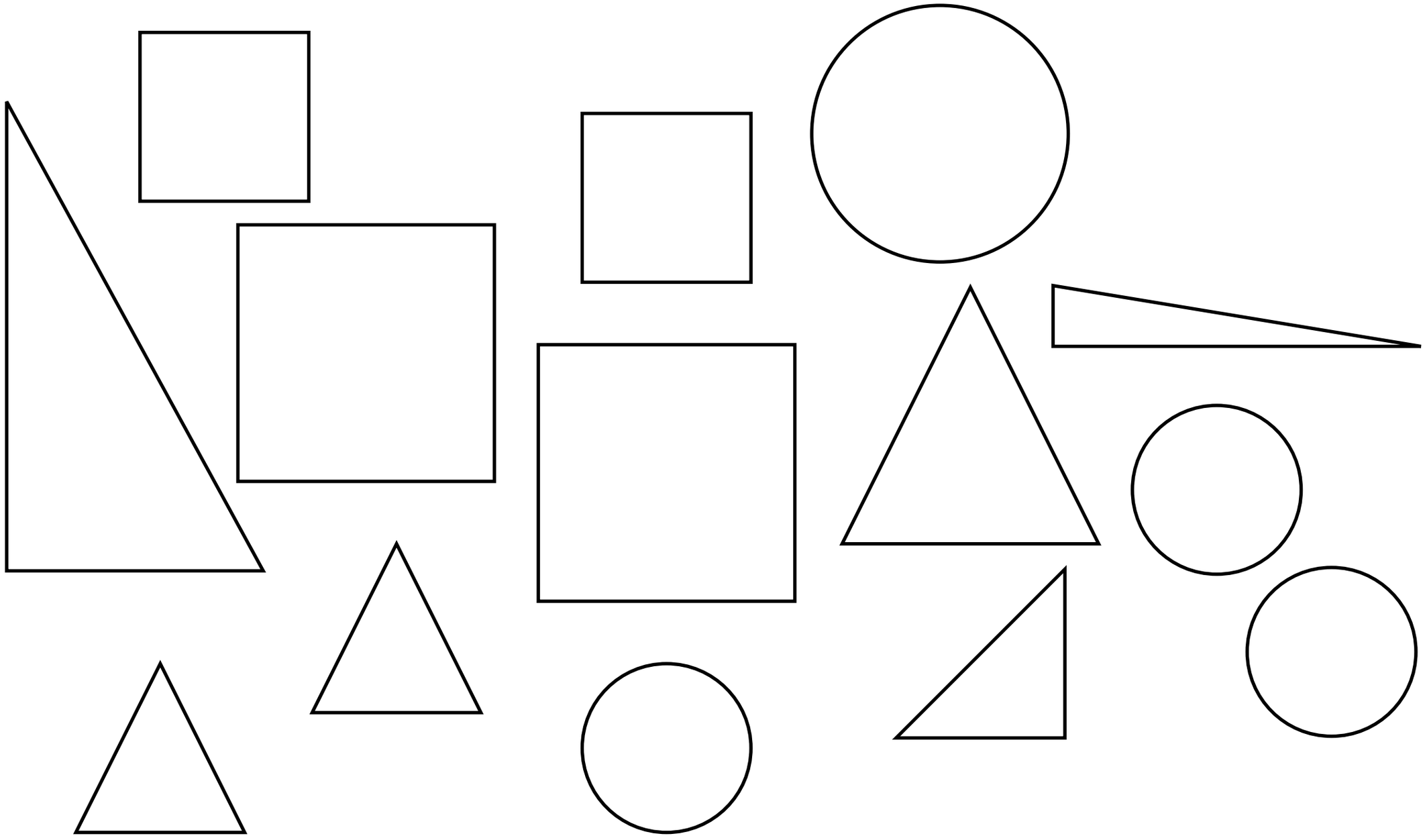
Predictive Analysis of Text: The Big Picture

Exploratory Analysis of Text: The Big Picture

Applications

Predictive Analysis

example: recognizing triangles



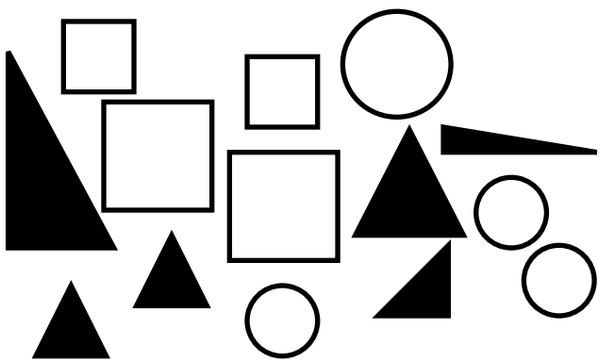
Predictive Analysis
example: recognizing triangles

- We could imagine writing a "triangle detector" by hand:
  - if shape has three sides, then shape = triangle.
  - otherwise, shape = other
- Alternatively, we could use supervised machine learning!

Predictive Analysis

example: recognizing triangles

training

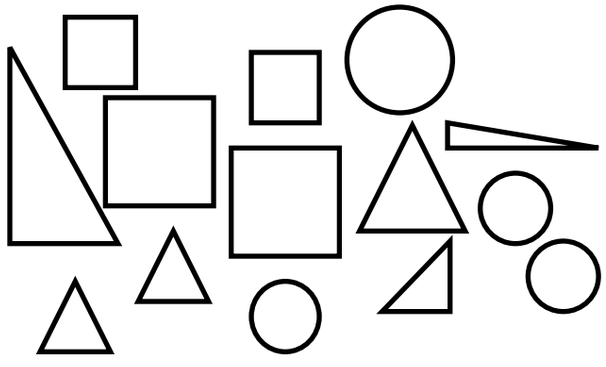


labeled examples

machine learning algorithm

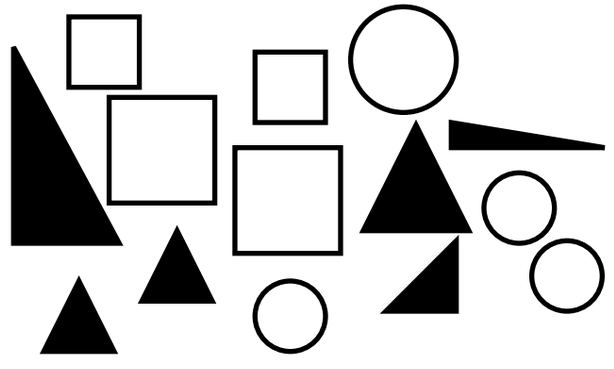
model

testing



new, unlabeled examples

model

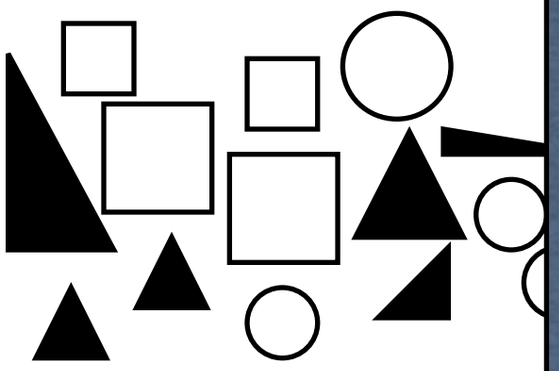


predictions

Predictive Analysis

example: recognizing triangles

training

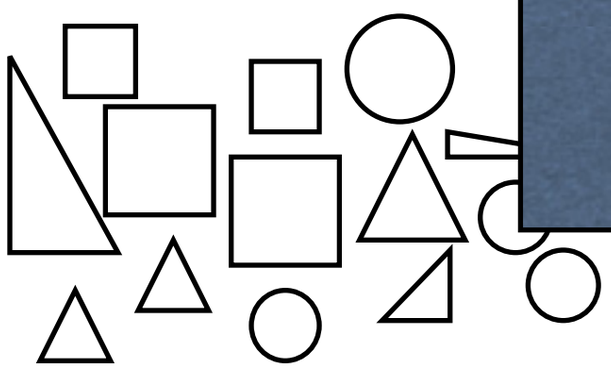


labeled exampl

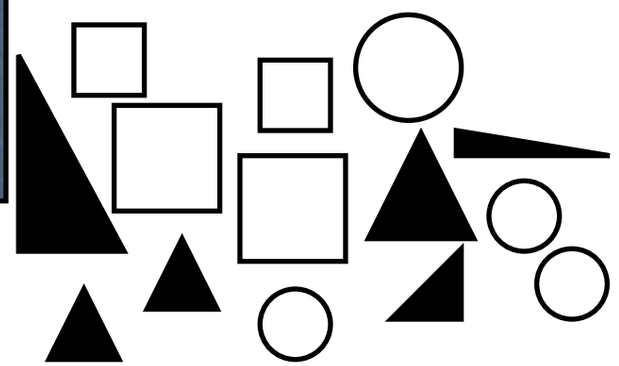
What is the part that is missing?

HINT: It's what most of this class will be about!

model



new, unlabeled examples



predictions

Predictive Analysis

representation: features

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

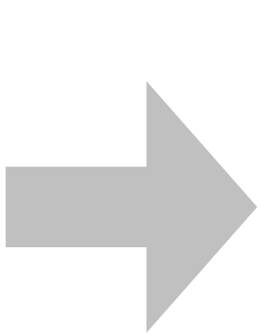
Predictive Analysis

example: recognizing triangles

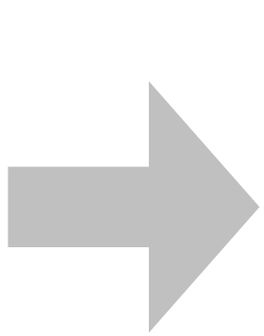
training

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

labeled examples



machine learning algorithm

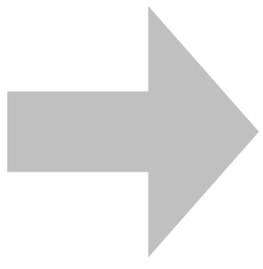


model

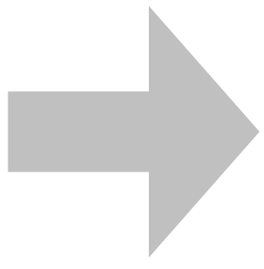
testing

color	size	sides	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



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

predictions

Predictive Analysis
basic ingredients

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

Predictive Analysis

basic ingredients

Highly influential!

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

Predictive Analysis
basic ingredients

4. Model: a (mathematical) function that describes a predictive relationship between the feature values and the presence/absence of the concept

5. Test data: a set of previously unseen examples used to estimate the model's effectiveness

6. Performance metrics: a set of statistics used measure the predictive effectiveness of the model

Predictive Analysis

basic ingredients: the focus in this course

1. Training data: a set of examples of the concept we want to automatically recognize
2. Representation: a set of features that we believe are useful in recognizing the desired concept
3. Learning algorithm: uses the training data to learn a predictive model of the "concept"

Predictive Analysis

basic ingredients: the focus in this course

4. Model: describes a predictive relationship between feature values and the presence/absence of the concept

5. Test data: a set of previously unseen examples used to estimate the model's effectiveness

6. Performance metrics: a set of statistics used measure the predictive effectiveness of the model

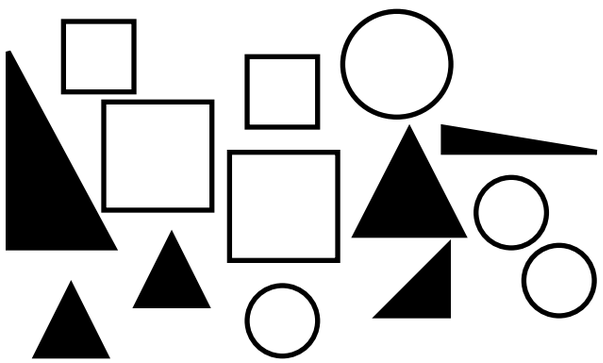
Predictive Analysis
applications

- Topic categorization
- Opinion mining
- Sentiment analysis
- Bias or viewpoint detection
- Discourse analysis (e.g., student retention)
- Forecasting and nowcasting
- Any other ideas?

Predictive Analysis

example: recognizing triangles

training

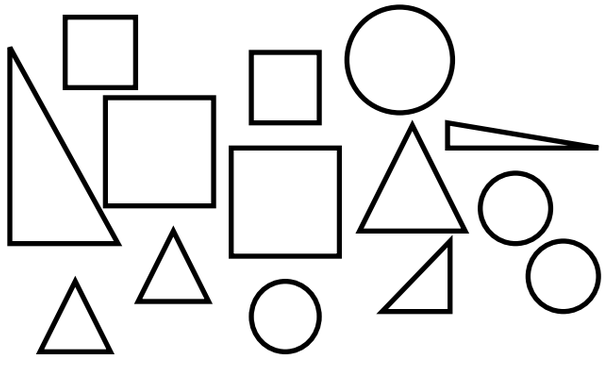


labeled examples

machine learning algorithm

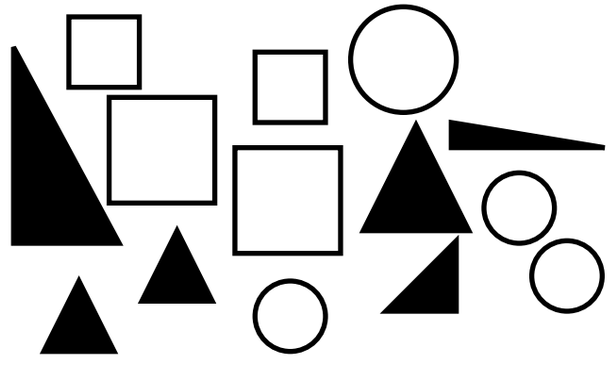
model

testing



new, unlabeled examples

model



predictions

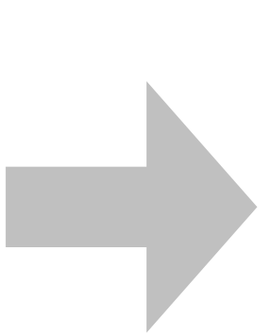
Predictive Analysis

example: recognizing triangles

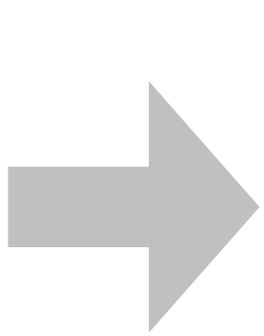
training

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

labeled examples



machine learning algorithm

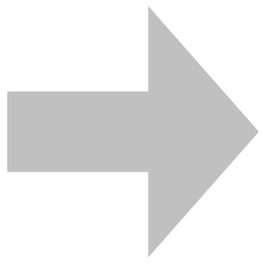


model

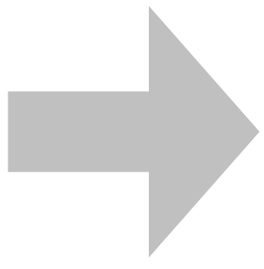
testing

color	size	sides	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



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

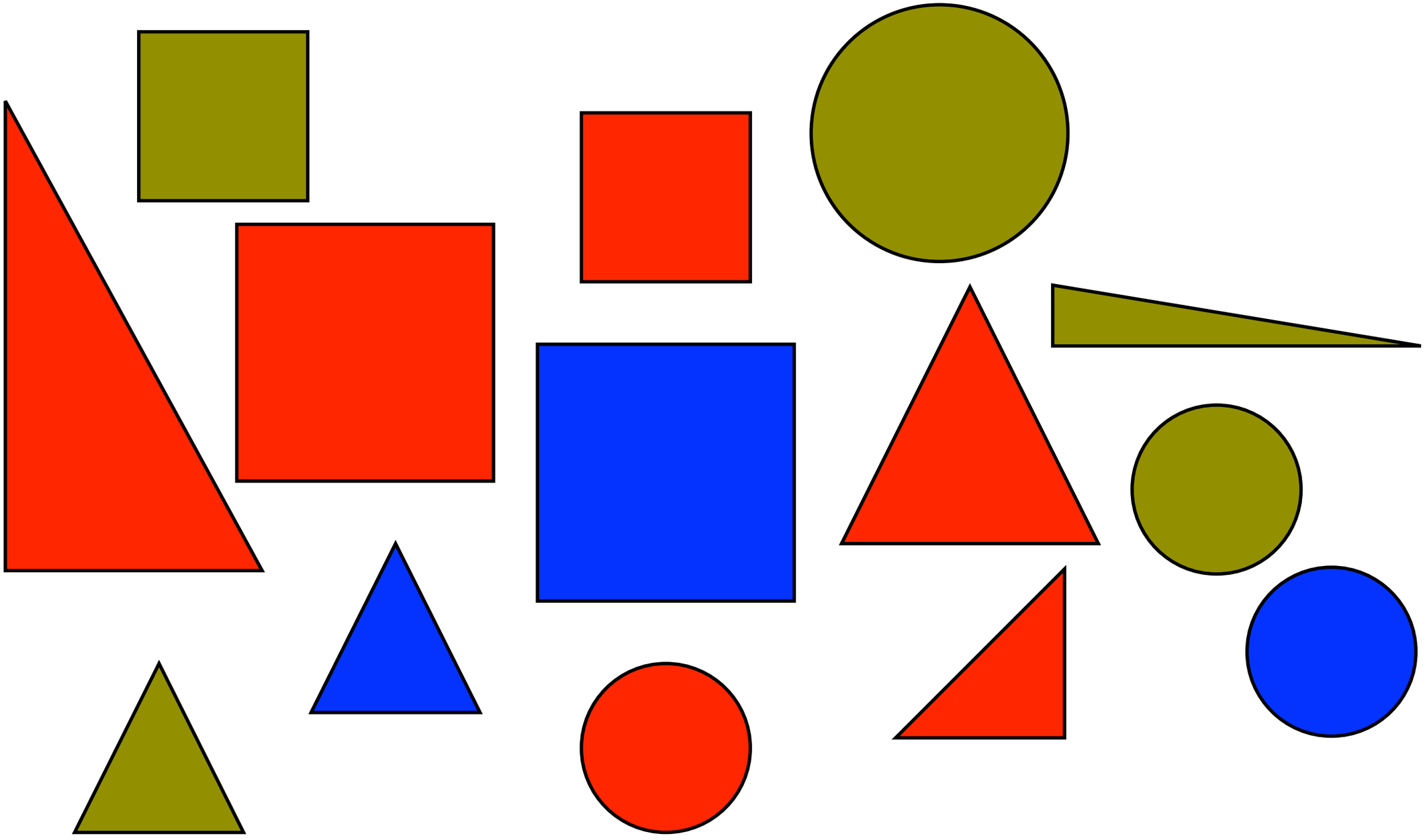
predictions

What Could Possibly Go Wrong?

1. Bad feature representation
2. Bad data + misleading correlations
3. Noisy labels for training and testing
4. Bad learning algorithm
5. Misleading evaluation metric

Training data + Representation

what could possibly go wrong?



Training data + Representation

what could possibly go wrong?

color	size	90 deg. angle	equal sides	...	label
red	big	yes	no	...	yes
green	big	no	yes	...	yes
blue	small	no	yes	...	no
blue	small	yes	yes	...	no
⋮	⋮	⋮	⋮	⋮	⋮
red	big	no	yes	...	yes

Training data + Representation

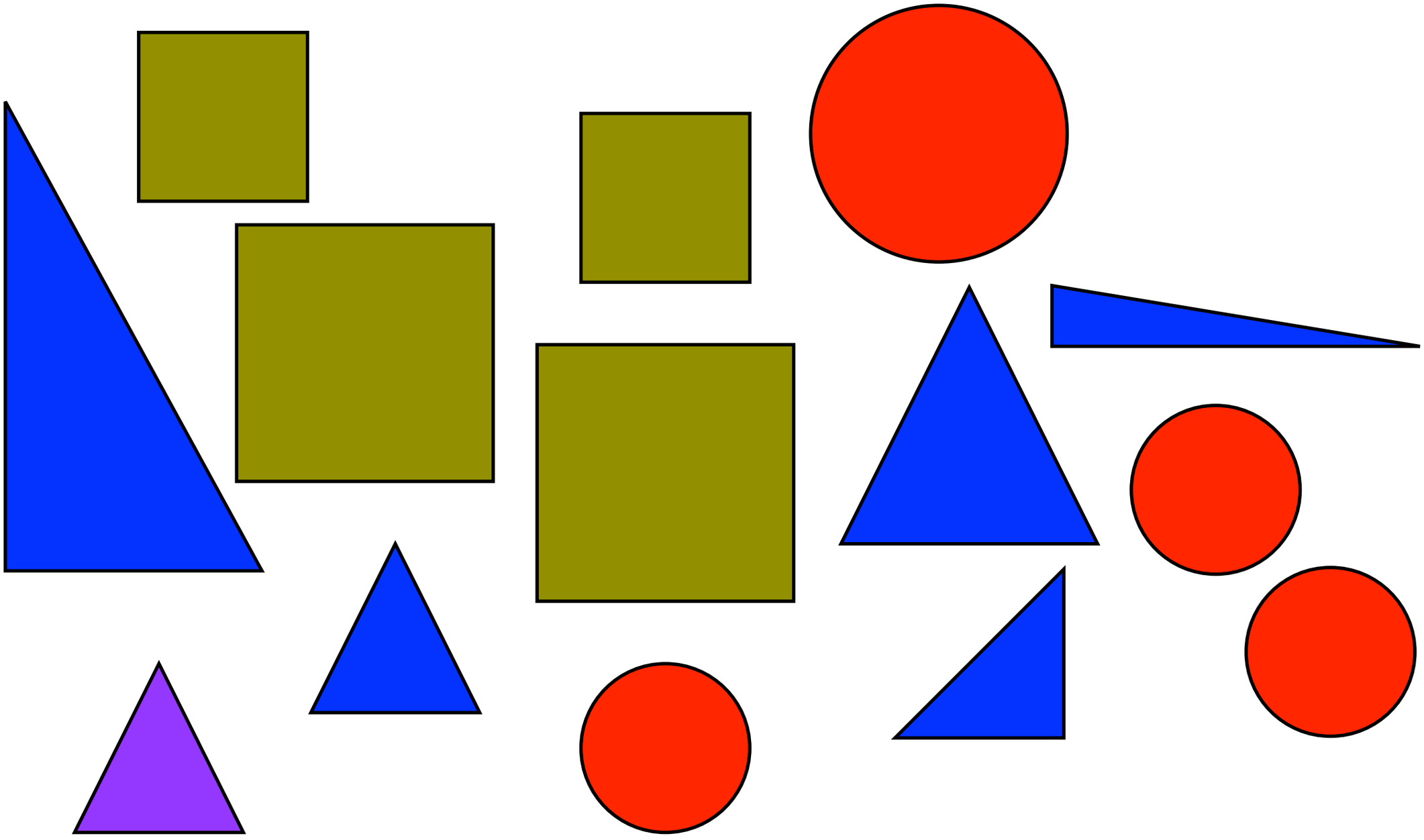
what could possibly go wrong?

color	size	90 deg. angle	equal sides	...	label
red	big	yes	no	...	yes
green	big	no	yes	...	yes
blue	small	no	yes	...	no
blue	small	yes	yes	...	no
⋮	⋮	⋮	⋮	⋮	⋮
red	big	no	yes	...	yes

1. bad feature representation!

Training data + Representation

what could possibly go wrong?



Training data + Representation

what could possibly go wrong?

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

Training data + Representation

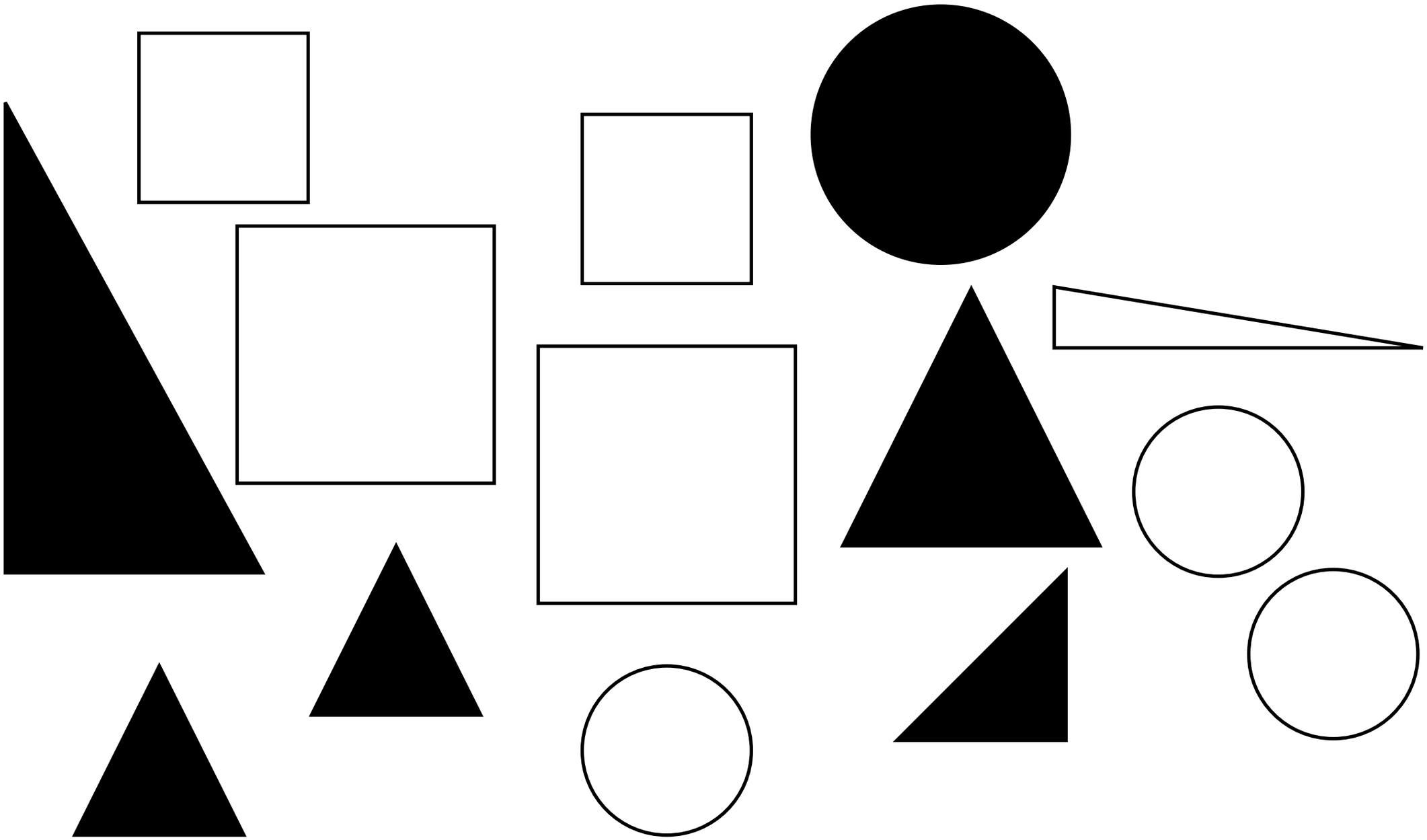
what could possibly go wrong?

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

2. bad data + misleading correlations

Training data + Representation

what could possibly go wrong?



Training data + Representation

what could possibly go wrong?

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

3. noisy training data!

Learning Algorithm + Model

what could possibly go wrong?

- Linear classifier

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

Learning Algorithm + Model

what could possibly go wrong?

- Linear classifier

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

parameters learned by the model

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

Learning Algorithm + Model

what could possibly go wrong?

test instance

f_1	f_2	f_3
0.5	1	0.2

model parameters

w_0	w_1	w_2	w_3
2	-5	2	1

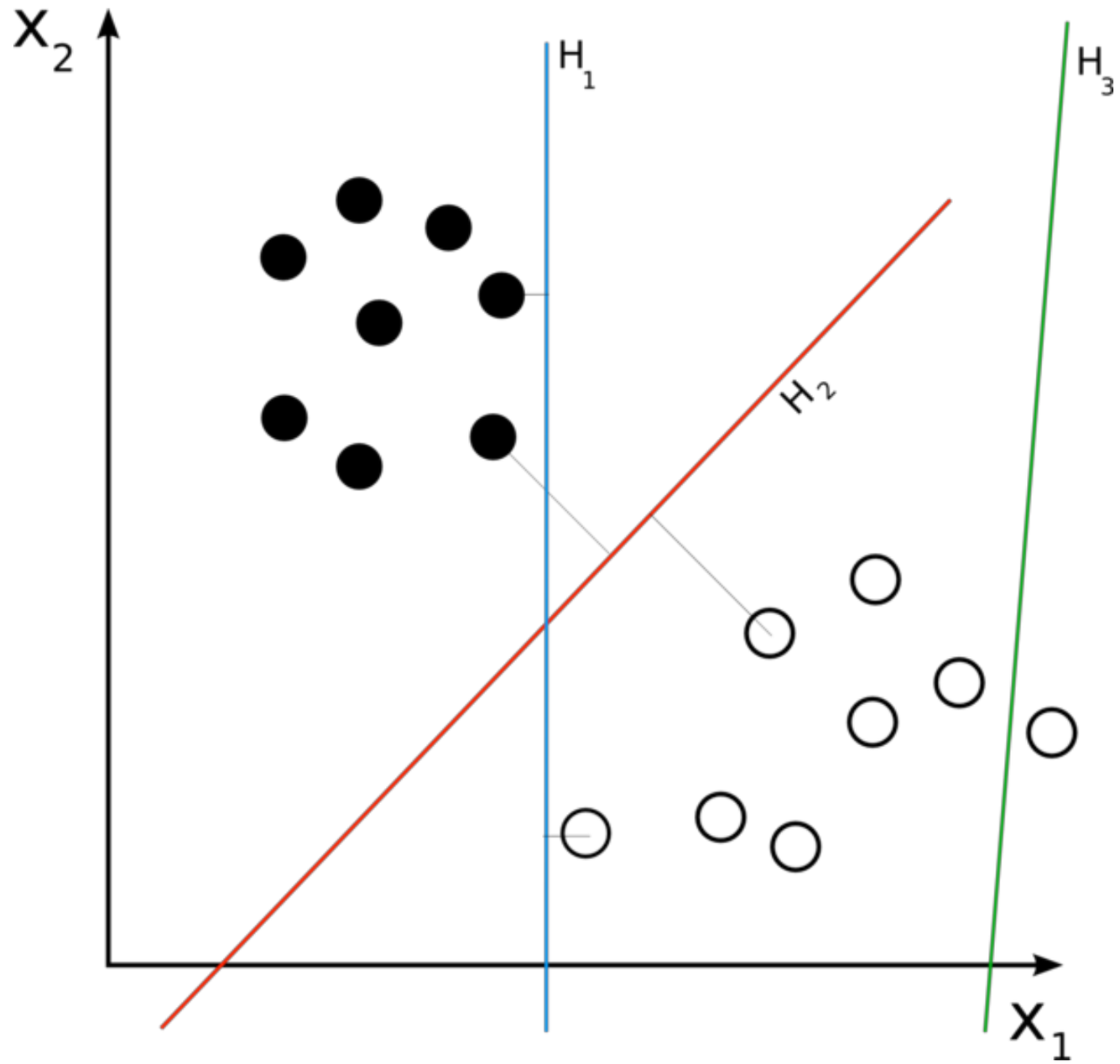
output = 2.0 + (0.50 x -5.0) + (1.0 x 2.0) + (0.2 x 1.0)

output = 1.7

output prediction = positive

Learning Algorithm + Model

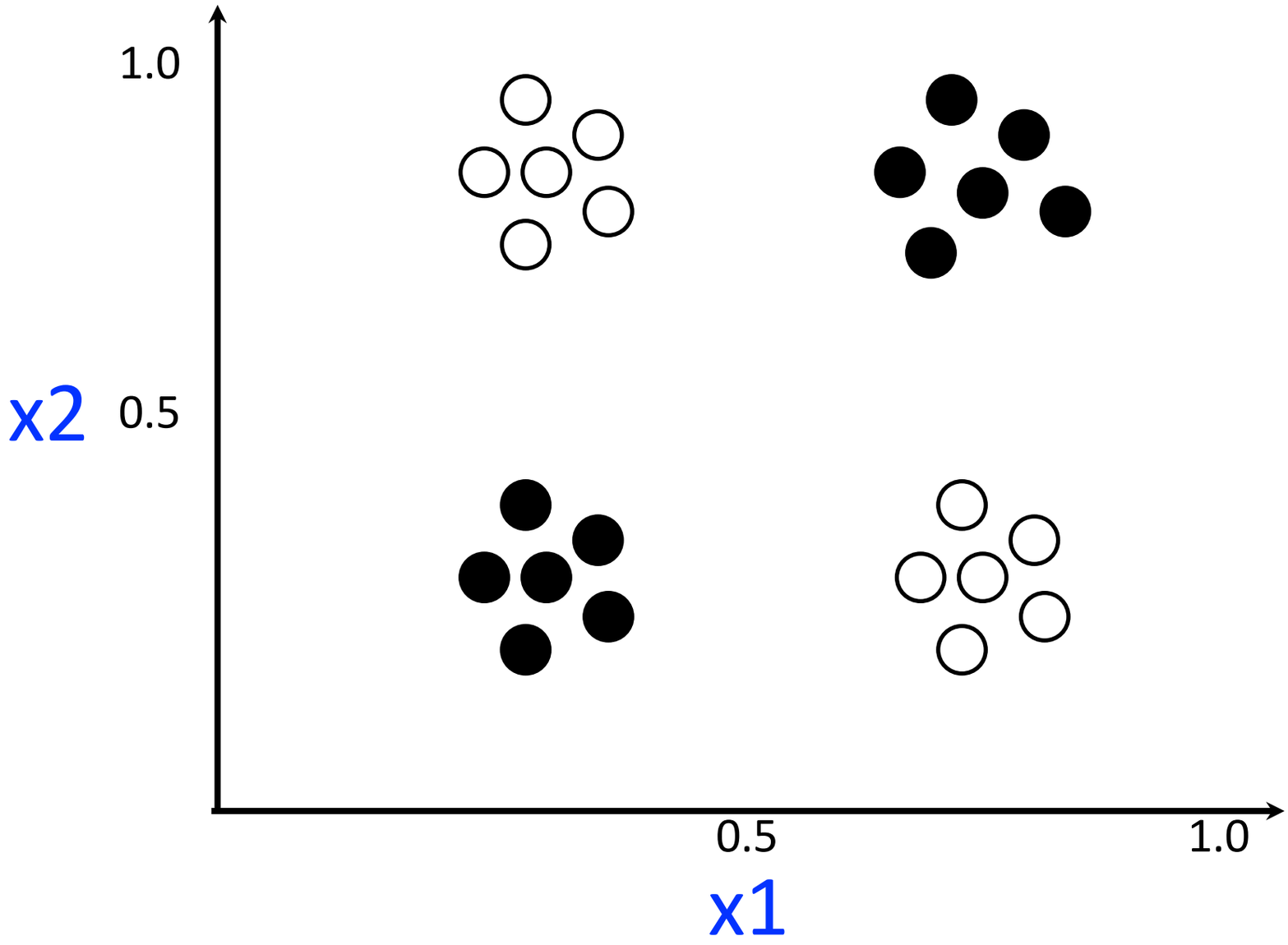
what could possibly go wrong?



(source: http://en.wikipedia.org/wiki/File:Svm_separating_hyperplanes.png)

Learning Algorithm + Model

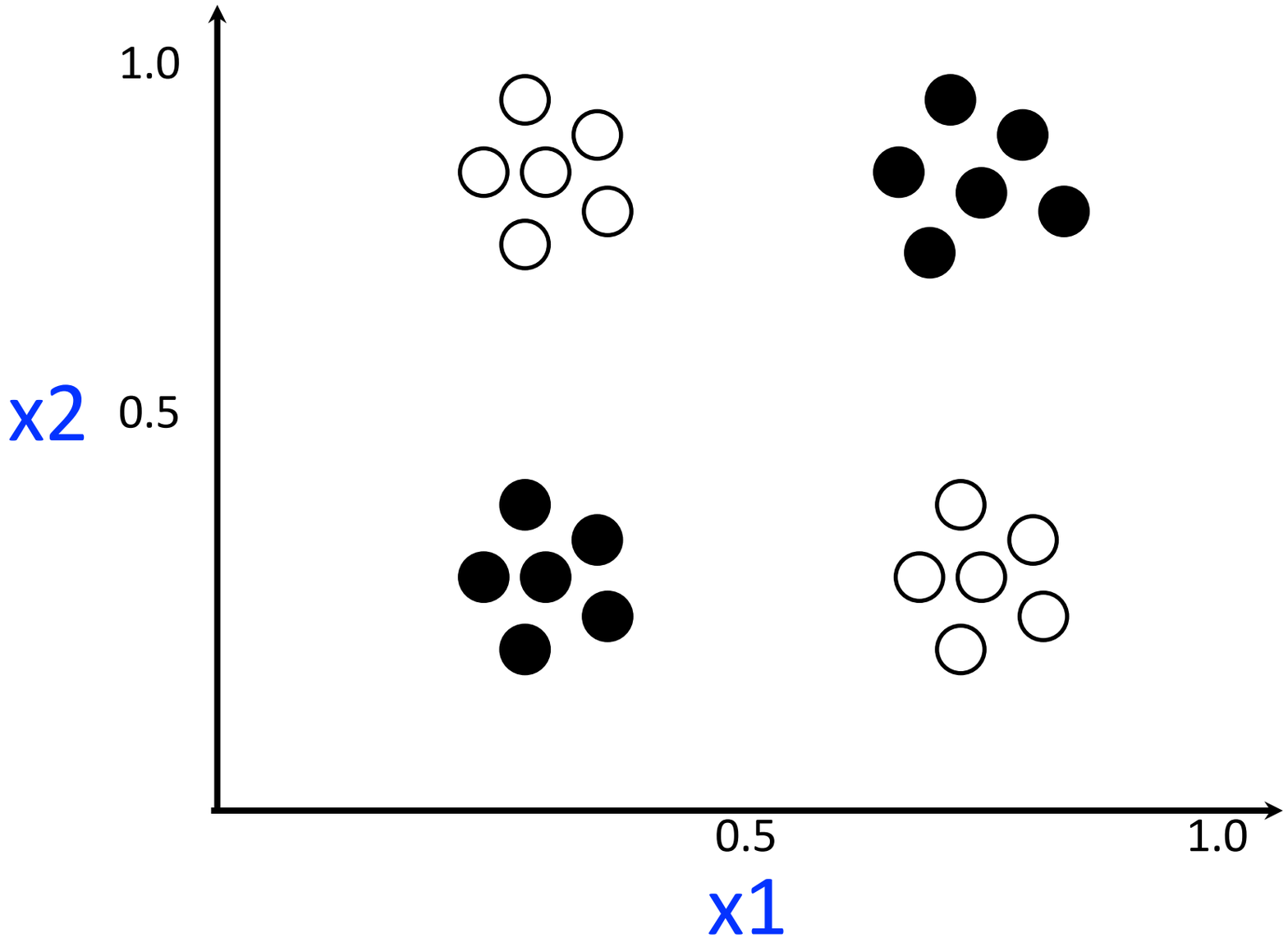
what could possibly go wrong?



- Would a linear classifier do well on positive (black) and negative (white) data that looks like this?

Learning Algorithm + Model

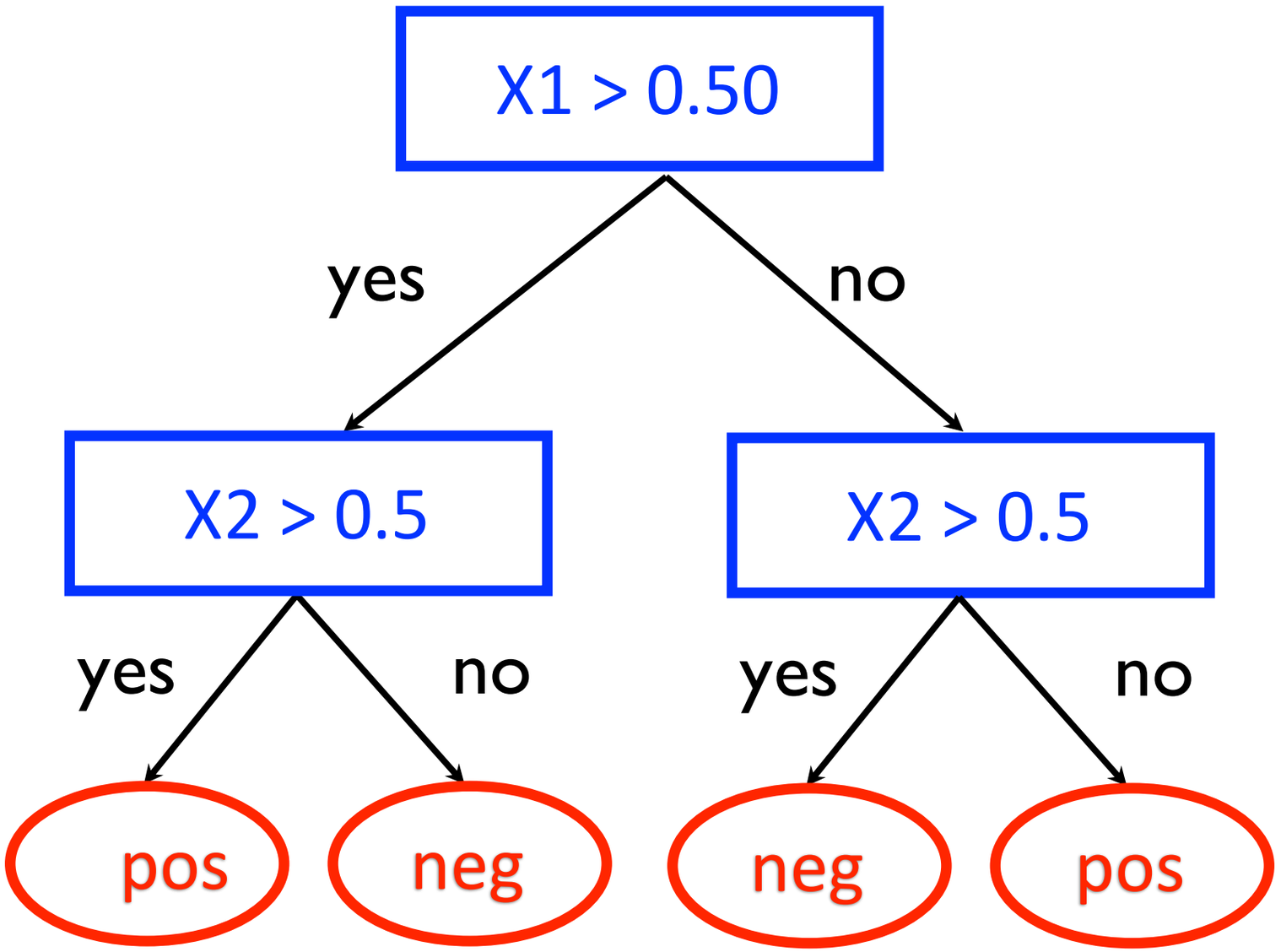
what could possibly go wrong?



4. Bad learning algorithm!

Learning Algorithm + Model

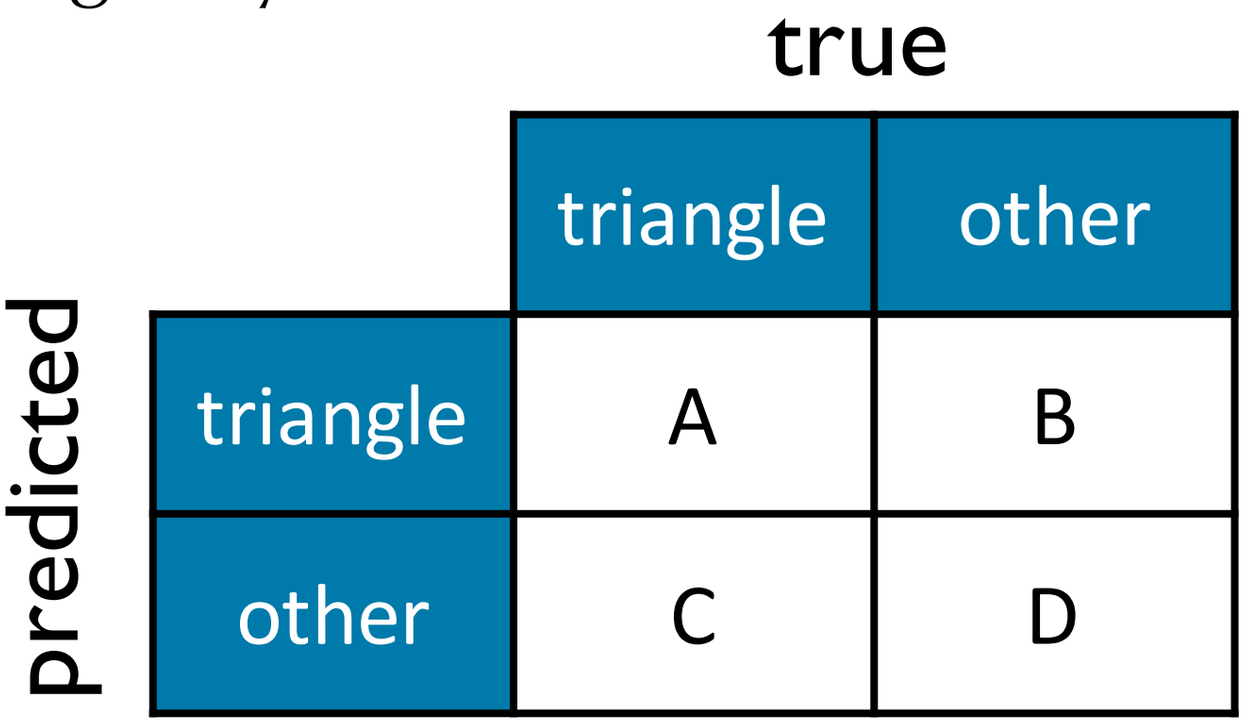
what could possibly go wrong?



Evaluation Metric

what could possibly go wrong?

- Most evaluation metrics can be understood using a contingency table



- What number(s) do we want to maximize?
- What number(s) do we want to minimize?

Evaluation Metric

what could possibly go wrong?

- True positives (A): number of triangles correctly predicted as triangles
- False positives (B): number of "other" incorrectly predicted as triangles
- False negatives (C): number of triangles incorrectly predicted as "other"
- True negatives (D): number of "other" correctly predicted as "other"

predicted \ true	triangle	other
triangle	A	B
other	C	D

Evaluation Metric

what could possibly go wrong?

- Accuracy: percentage of predictions that are correct (i.e., true positives and true negatives)

$$\frac{(? + ?)}{(? + ? + ? + ?)}$$

predicted \ true	triangle	other
triangle	A	B
other	C	D

Evaluation Metric

what could possibly go wrong?

- Accuracy: percentage of predictions that are correct (i.e., true positives and true negatives)

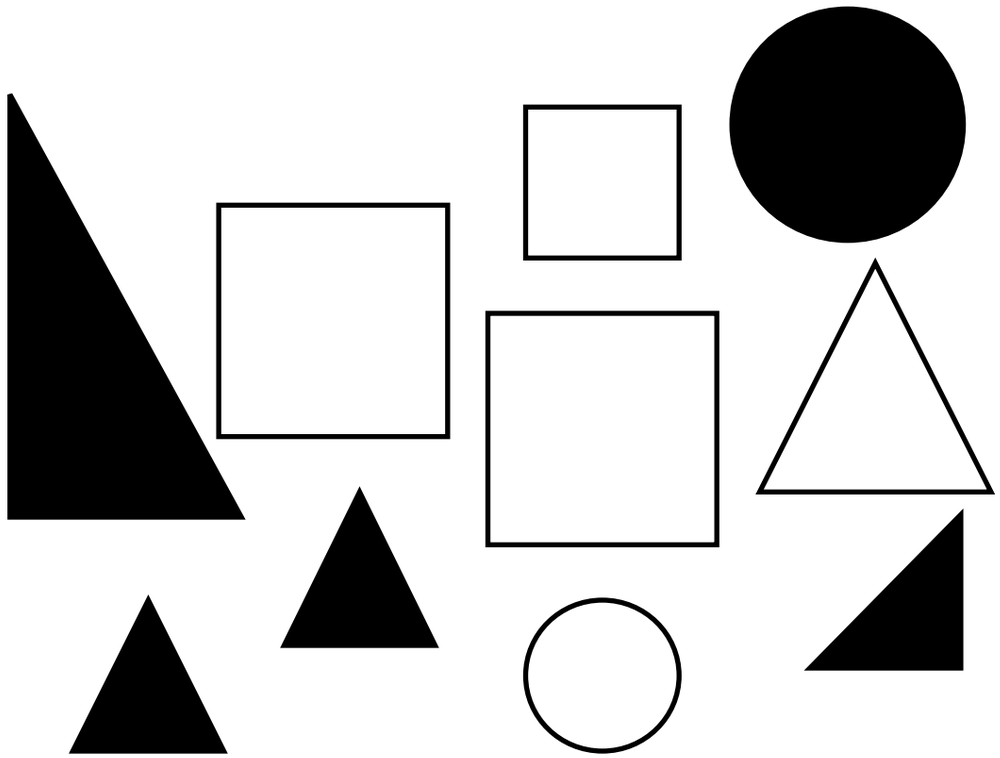
$$\frac{(A + D)}{(A + B + C + D)}$$

predicted \ true	triangle	other
triangle	A	B
other	C	D

Evaluation Metric

what could possibly go wrong?

- Accuracy: percentage of predictions that are correct (i.e., true positives and true negatives)

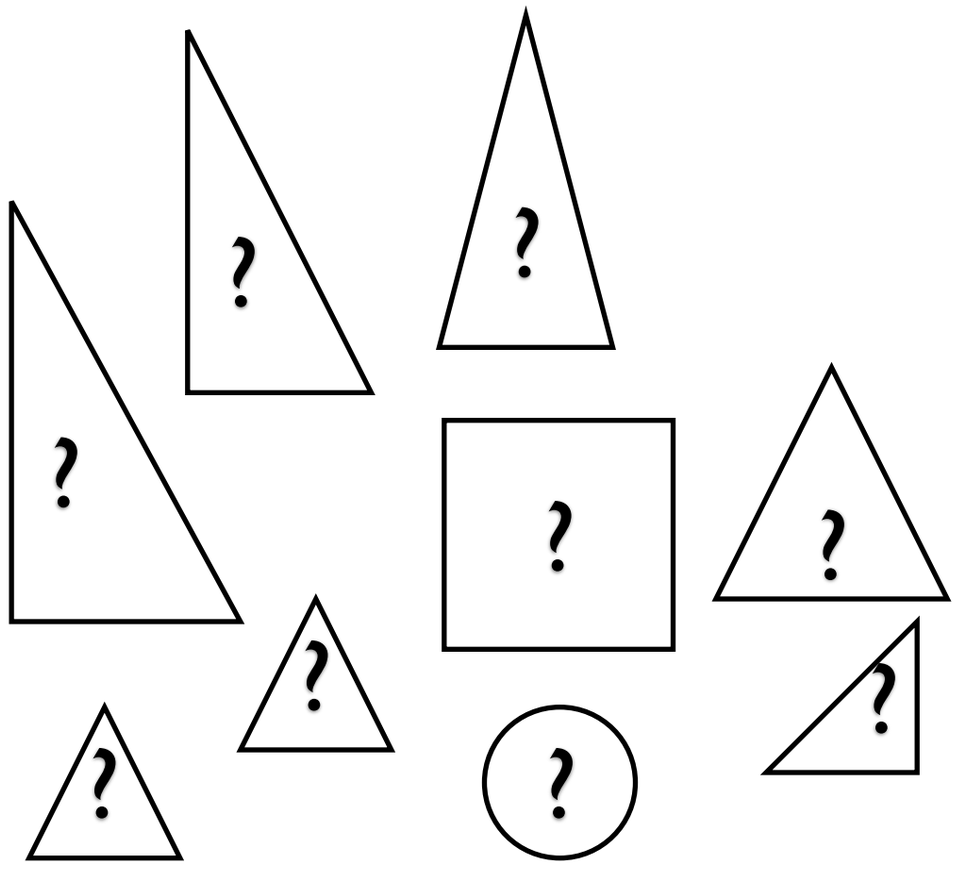


- What is the accuracy of this model?

Evaluation Metric

what could possibly go wrong?

- Interpreting the value of a metric on a particular data set requires some thinking ...

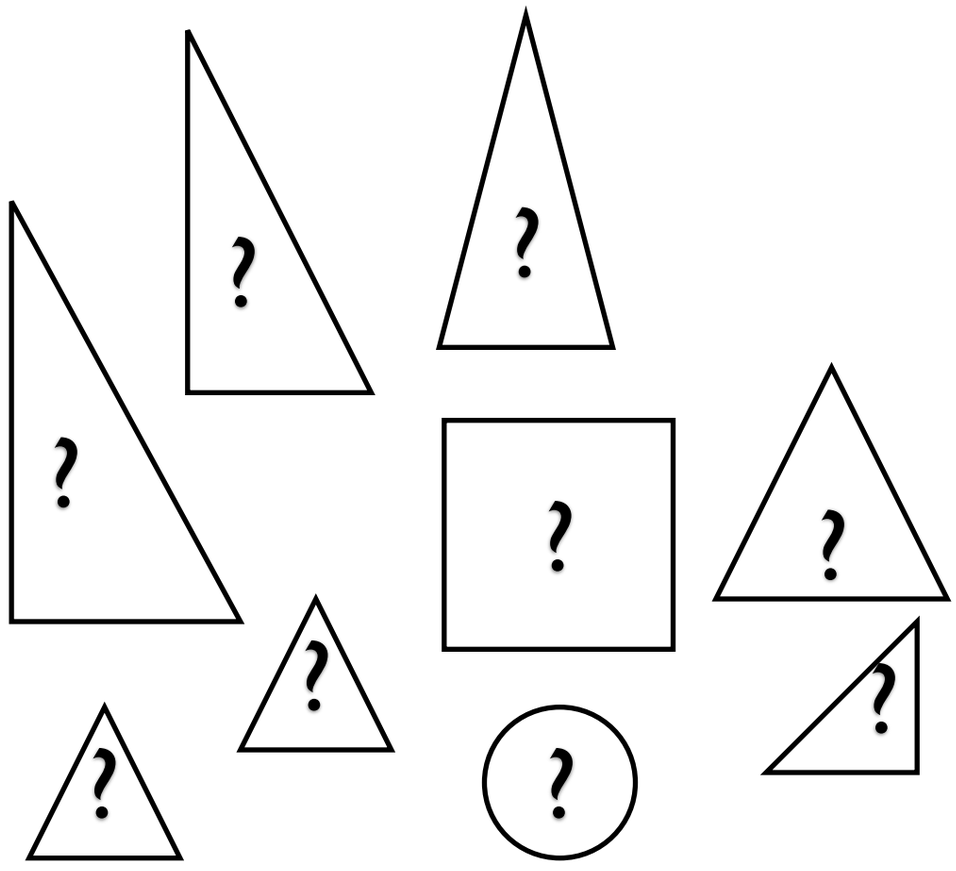


- On this dataset, what would be the expected accuracy of a model that does NO learning (degenerate baseline)?

Evaluation Metric

what could possibly go wrong?

- Interpreting the value of a metric on a particular data set requires some thinking ...



5. Misleading interpretation of a metric value!

What Could Possibly Go Wrong?

1. Bad feature representation
2. Bad data + misleading correlations
3. Noisy labels for training and testing
4. Bad learning algorithm
5. Misleading evaluation metric

Road Map

first half of the semester

- Predictive Analysis of Text
  - Supervised machine learning principles
  - Text representation
  - Feature selection
  - Basic machine learning algorithms
  - Tools for predictive analysis of text
  - Experimentation and evaluation
- Exploratory Analysis of Text
  - Clustering
  - Outlier detection (tentative)
  - Co-occurrence statistics

Road Map
second half of the semester

- Applications
  - Text classification
  - Opinion mining
  - Sentiment analysis
  - Bias detection
  - Information extraction
  - Relation learning
  - Text-based forecasting
  - Temporal Summarization
- Is there anything that you would like to learn more about?

Grading

- 30% homework
  - 10% each
- 20% midterm
- 40% term project
  - 5% proposal
  - 10% presentation
  - 25% paper
- 10% participation

Grading for Graduate Students

- H: 95-100%
- P: 80-94%
- L: 60-79%
- F: 0-59%

Grading for Undergraduate Students

- A+: 97-100%
- A: 94-96%
- A-: 90-93%
- B+: 87-89%
- B: 84-86%
- B-: 80-83%
- C+: 77-79%
- C: 74-76%
- C-: 70-73%
- D+: 67-69%
- D: 64-66%
- D-: 60-63%
- F: <= 59%

General Outline of Homework

- Given a dataset (i.e., a training and test set), run experiments where you try to predict the target class using different feature representations
- Do error analysis
- Report on what worked, what didn't, and why!
- Answer essay questions about the assignment
  - These will be associated with the course material

Homework vs. Midterm

- The homework will be more challenging than the mid-term. It should be, you have more time.

Course Tips

- Work hard
- Do the assigned readings
- Do other readings
- Be patient and have reasonable expectations
  - you're not supposed to understand everything we cover in class during class
- Seek help sooner rather than later
  - office hours: by appointment
  - questions via email or MS Teams
- Remember the golden rule: no pain, no gain

Questions?