GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LECTURE 5: MACHINE LEARNING

I. WHAT IS MACHINE LEARNING?
II. MACHINE LEARNING PROBLEMS
III. CLASSIFICATION PROBLEMS
IV. KNN CLASSIFICATION

I. WHAT IS MACHINE LEARNING?

WHAT IS MACHINE LEARNING?

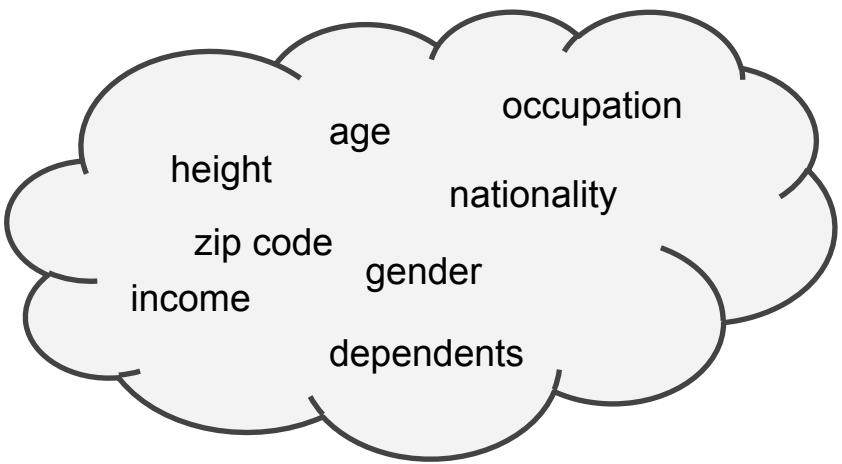
from Wikipedia:

"Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can *learn from data*."

source: http://en.wikipedia.org/wiki/Machine_learning

from Wikipedia:

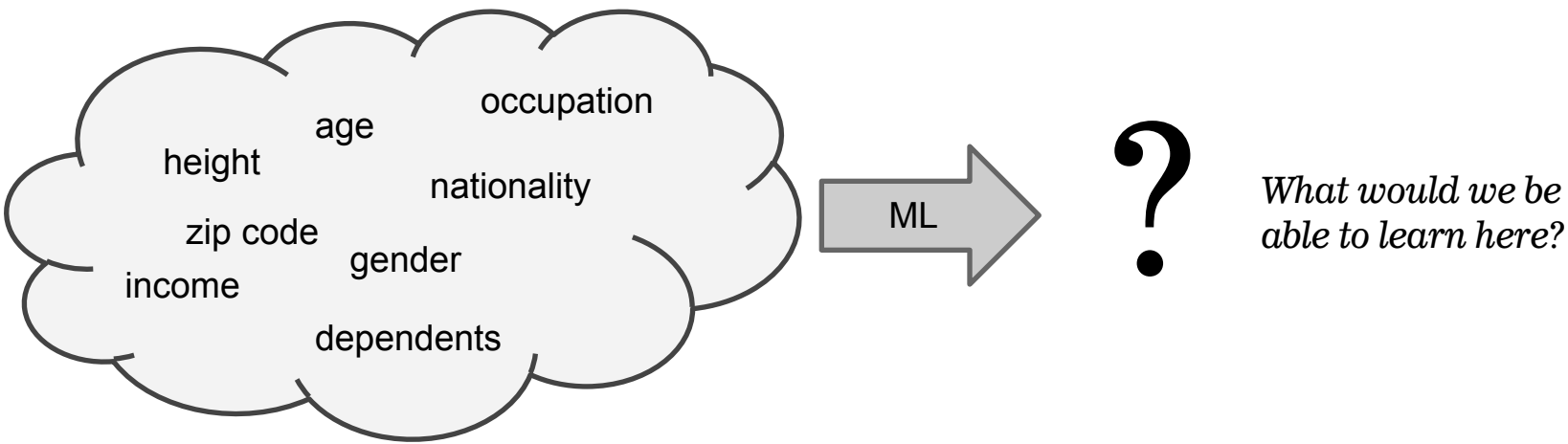
"Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can *learn from data*."



source: http://en.wikipedia.org/wiki/Machine_learning

from Wikipedia:

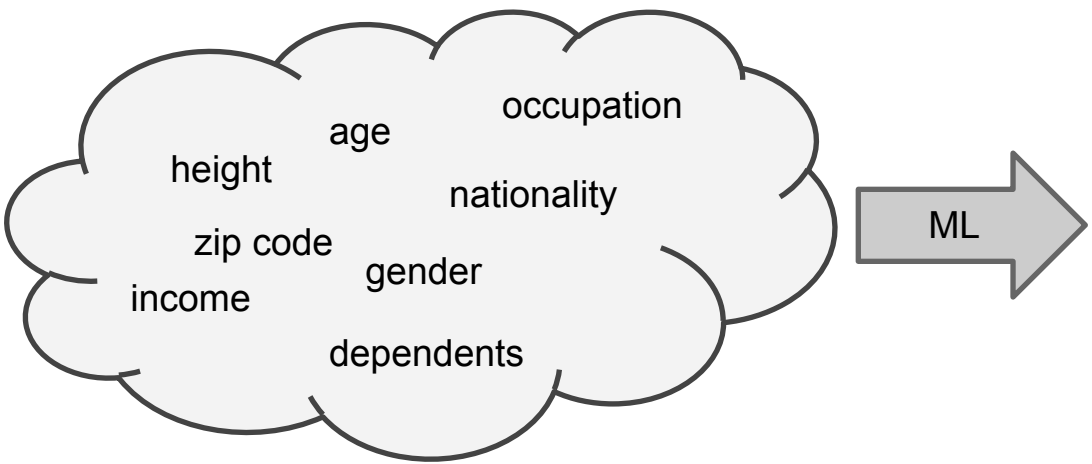
"Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can *learn from data*."



source: http://en.wikipedia.org/wiki/Machine_learning

from Wikipedia:

"Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can *learn from data*."



Algorithm has learned to:

- predict income of person
- cluster customers in segments
- predict if person has children

source: http://en.wikipedia.org/wiki/Machine_learning

WHAT IS MACHINE LEARNING?

from Wikipedia:

"Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can *learn from data*."

"The core of machine learning deals with *representation* and *generalization*…"

source: http://en.wikipedia.org/wiki/Machine_learning

WHAT IS MACHINE LEARNING?

from Wikipedia:

"Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can *learn from data*."

"The core of machine learning deals with *representation* and *generalization*…"

- *representation* – extracting structure from data

source: http://en.wikipedia.org/wiki/Machine_learning

WHAT IS MACHINE LEARNING?

from Wikipedia:

"Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can *learn from data*."

"The core of machine learning deals with *representation* and *generalization*…"

- *representation* – extracting structure from data
- *generalization* – making predictions from data

source: http://en.wikipedia.org/wiki/Machine_learning

WHAT IS MACHINE LEARNING?

from Coursera:

"Machine learning is the science of getting computers to act without being explicitly programmed."

source: https://www.coursera.org/learn/machine-learning/home/info

WHAT IS MACHINE LEARNING?

from Coursera:

“Machine learning is the science of getting computers to act without being explicitly programmed.”

Instead of programming a computer how to perform a task…

source: https://www.coursera.org/learn/machine-learning/home/info

WHAT IS MACHINE LEARNING?

from Coursera:

"Machine learning is the science of getting computers to act without being explicitly programmed."

Instead of programming a computer how to perform a task…

…we show the computer a history of how others performed the task, and tell him to figure it out from there.

source: https://www.coursera.org/learn/machine-learning/home/info

III. MACHINE LEARNING PROBLEMS

supervised	
unsupervised	

supervised	*labeled examples*
unsupervised	*no labeled examples*

supervised	*making predictions*
unsupervised	*discovering patterns*

generalization

supervised	*making predictions*
unsupervised	*discovering patterns*

representation

	continuous	*categorical*

	continuous	***categorical***
	quantitative	*qualitative*
	age, salary, height, etc.	city, yes/no, vote, etc.

	continuous	***categorical***
supervised	*regression*	*classification*
unsupervised	*dimension reduction*	*clustering*

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

Salary prediction

	continuous	***categorical***
	Salary prediction	Vote prediction
supervised	*regression*	*classification*
unsupervised	*dimension reduction*	*clustering*

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

Salary prediction

Vote prediction

customer segmentation

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

Salary prediction

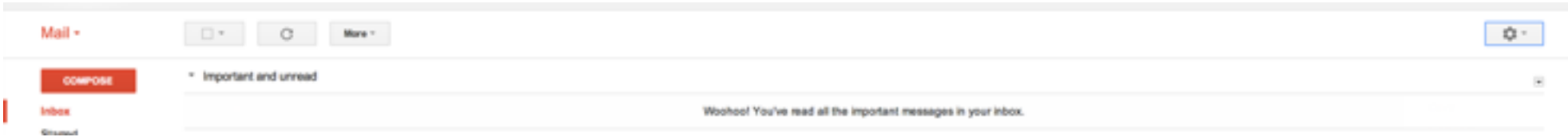
Vote prediction

face recognition

customer segmentation

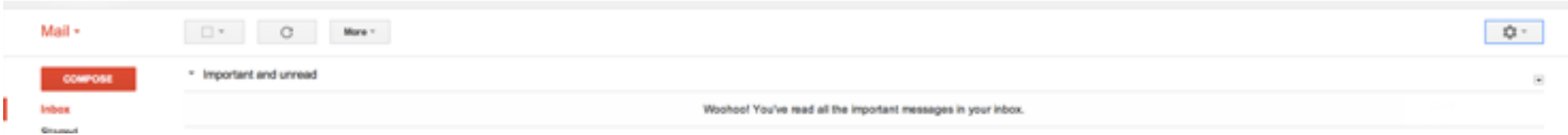
What type of problem is this?

Priority Inbox



What type of problem is this?

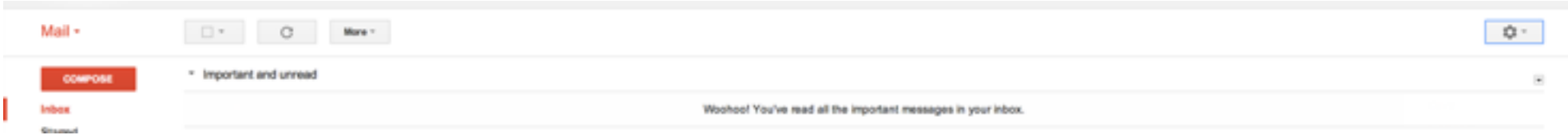
Priority Inbox



Probably either.

What type of problem is this?

Priority Inbox

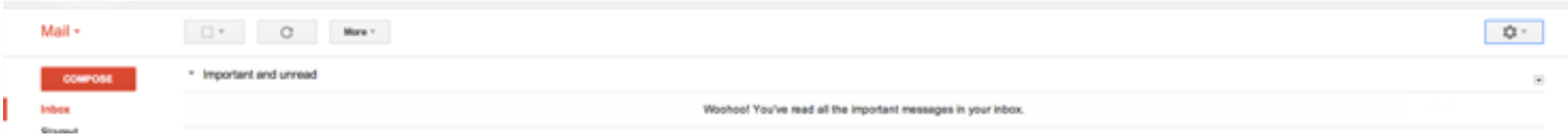


Supervised

- Predict which mails users are most likely to star

What type of problem is this?

Priority Inbox



Supervised
- Predict which mails users are most likely to star

Unsupervised
- Group mails into groups and decide which group represents important mails

What type of problem is this?

Music Recommendation

What type of problem is this?

Music Recommendation

Probably either.

What type of problem is this?

Music Recommendation as Supervised Learning

Predict which songs a user will ‘thumbs-up’

What type of problem is this?

Music Recommendation as Unsupervised Learning

Cluster songs based on attributes and recommend songs in the same group

QUESTION

HOW DO YOU DETERMINE THE RIGHT APPROACH?

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

ANSWER

The right approach is determined by the desired solution **and the data available.**

QUESTION

HOW DO YOU REPRESENT YOUR DATA?

	continuous	***categorical***
	quantitative	*qualitative*

	continuous	*categorical*
color	*quantitative* *or*	*qualitative* *?*
ratings		

	continuous	*categorical*
color	*RGB-values*	*{red, blue}*
ratings	*1 – 10 rating*	*1-5 star rating*

QUESTION

HOW DO YOU MEASURE THE QUALITY?

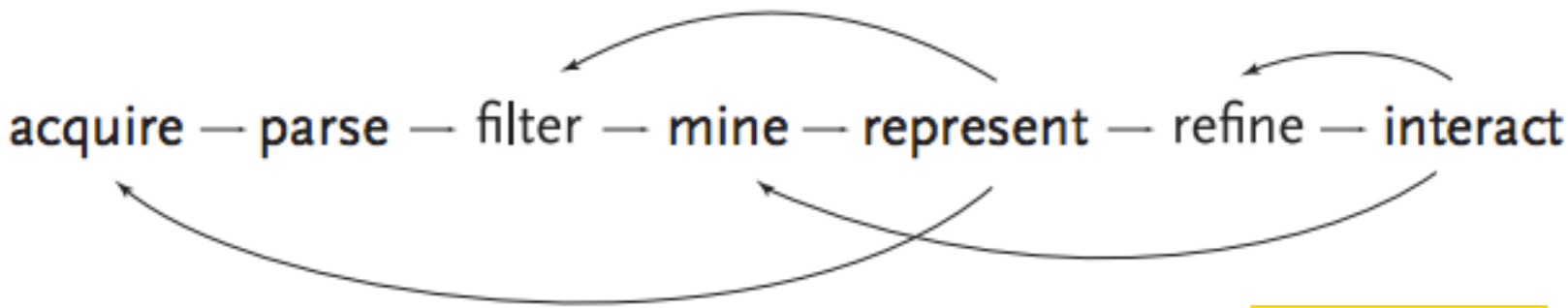
supervised	*making predictions*
unsupervised	*extracting structure*

supervised	*test out your predictions*

supervised	*test out your predictions*
unsupervised	*can't really (to be continued)*

QUESTION

WHAT DO YOU DO WITH YOUR RESULTS?



ANSWER

Interpret them and react accordingly.

source: http://benfry.com/phd/dissertation-110323c.pdf

IV. CLASSIFICATION PROBLEMS

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

Here's (part of) an example dataset:

Fisher's iris dataset (1936)

sepal_length	sepal_width	petal_length	petal_width	species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.3	3.3	6.0	2.5	virginica
5.8	2.7	5.1	1.9	virginica

Here's (part of) an example dataset:

Fisher's iris dataset (1936)

sepal_length	sepal_width	petal_length	petal_width	species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.3	3.3	6.0	2.5	virginica
5.8	2.7	5.1	1.9	virginica

independent variables
(also called *features*)

Here's (part of) an example dataset:

Fisher's iris dataset (1936)

sepal_length	sepal_width	petal_length	petal_width	species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.3	3.3	6.0	2.5	virginica
5.8	2.7	5.1	1.9	virginica

independent variables
(also called *features*)

class labels
(qualitative)

Here's (part of) an example dataset:

Fisher's iris dataset (1936)

sepal_length	sepal_width	petal_length	petal_width	species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.3	3.3	6.0	2.5	virginica
5.8	2.7	5.1	1.9	virginica

X = independent variables (also called *features*)

y = class labels (qualitative)

Q: What does "supervised" mean?

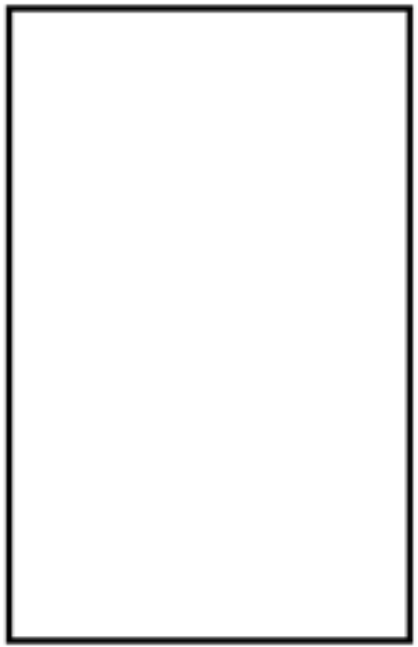
Q: What does "supervised" mean?

A: We know the labels.

sepal_length	sepal_width	petal_length	petal_width	species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.3	3.3	6.0	2.5	virginica
5.8	2.7	5.1	1.9	virginica

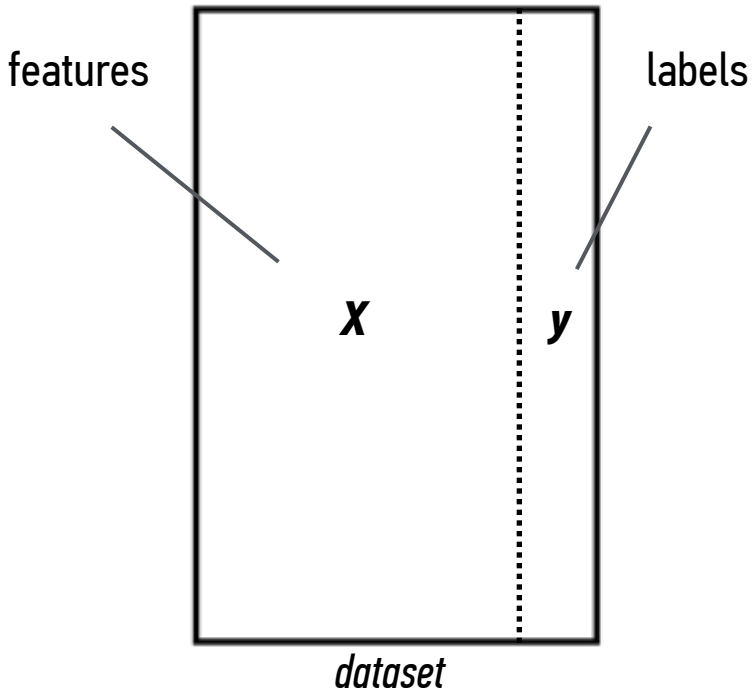
Q: How does a classification problem work?

Q: How does a classification problem work?

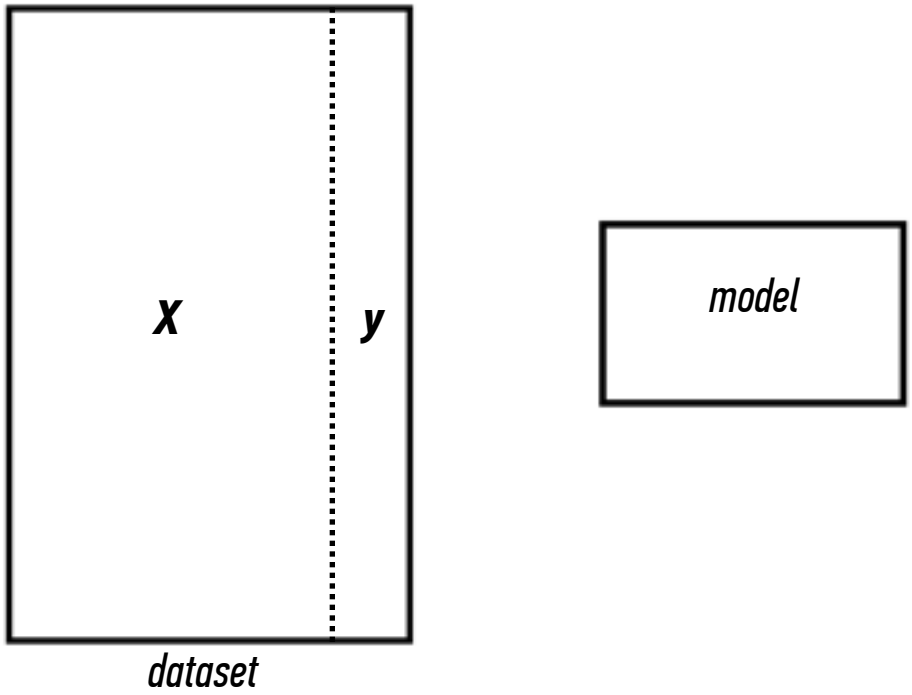


dataset

Q: How does a classification problem work?

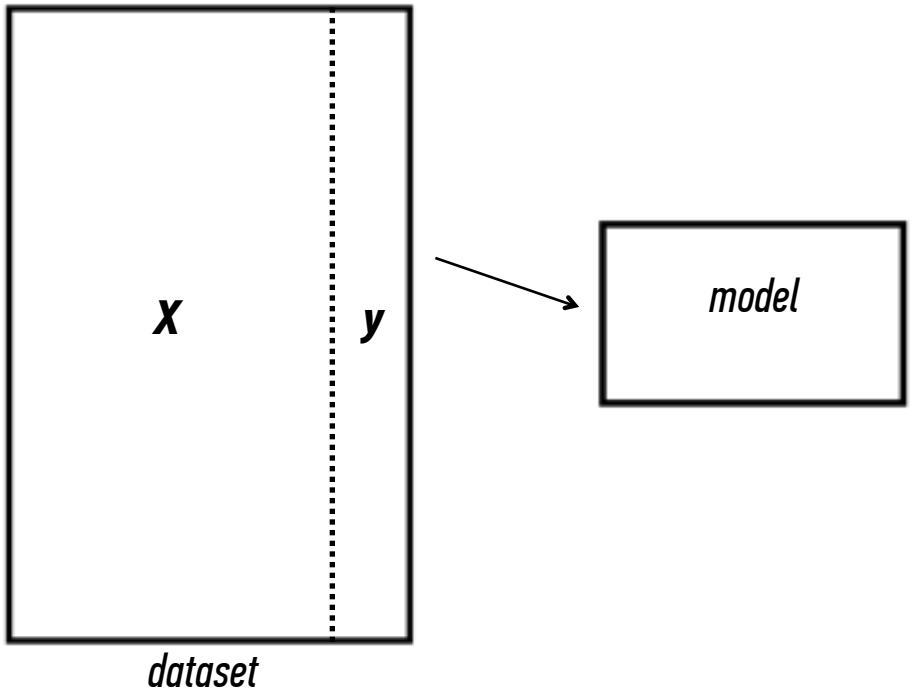


Q: How does a classification problem work?



Q: How does a classification problem work?

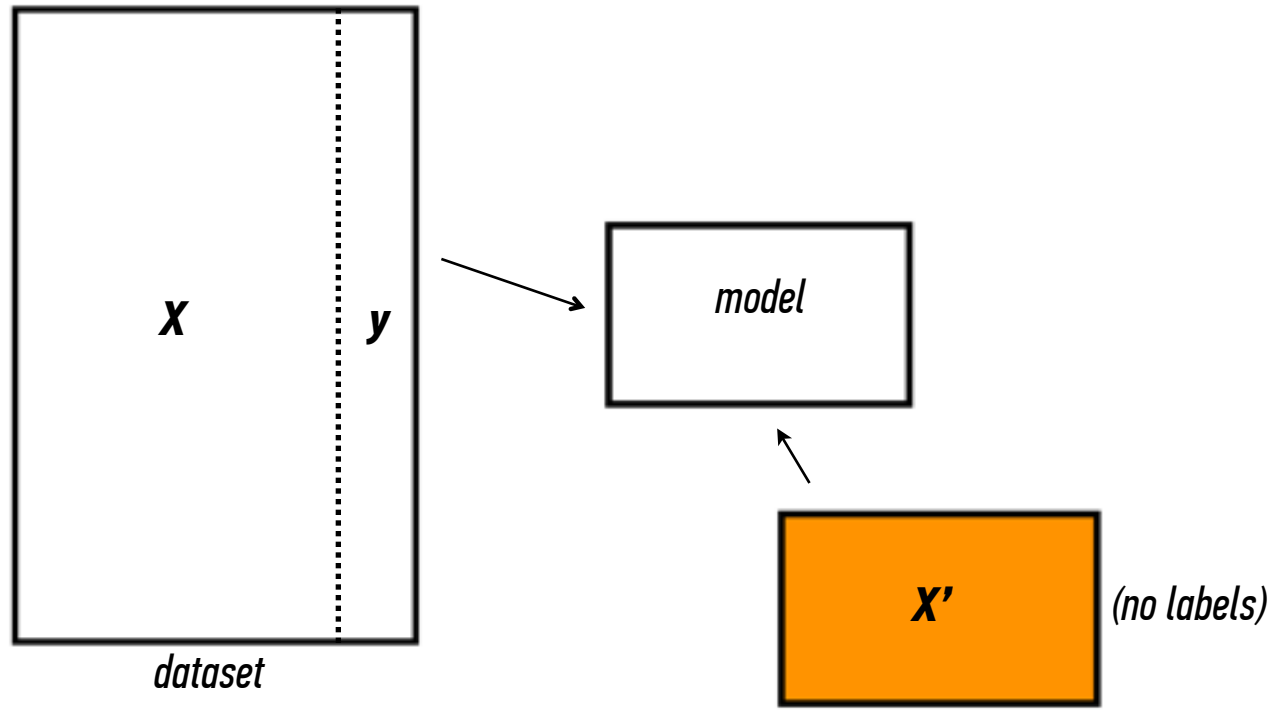
1) train model



*model 'learns' how **X** and **y** relate to each other*

Q: How does a classification problem work?

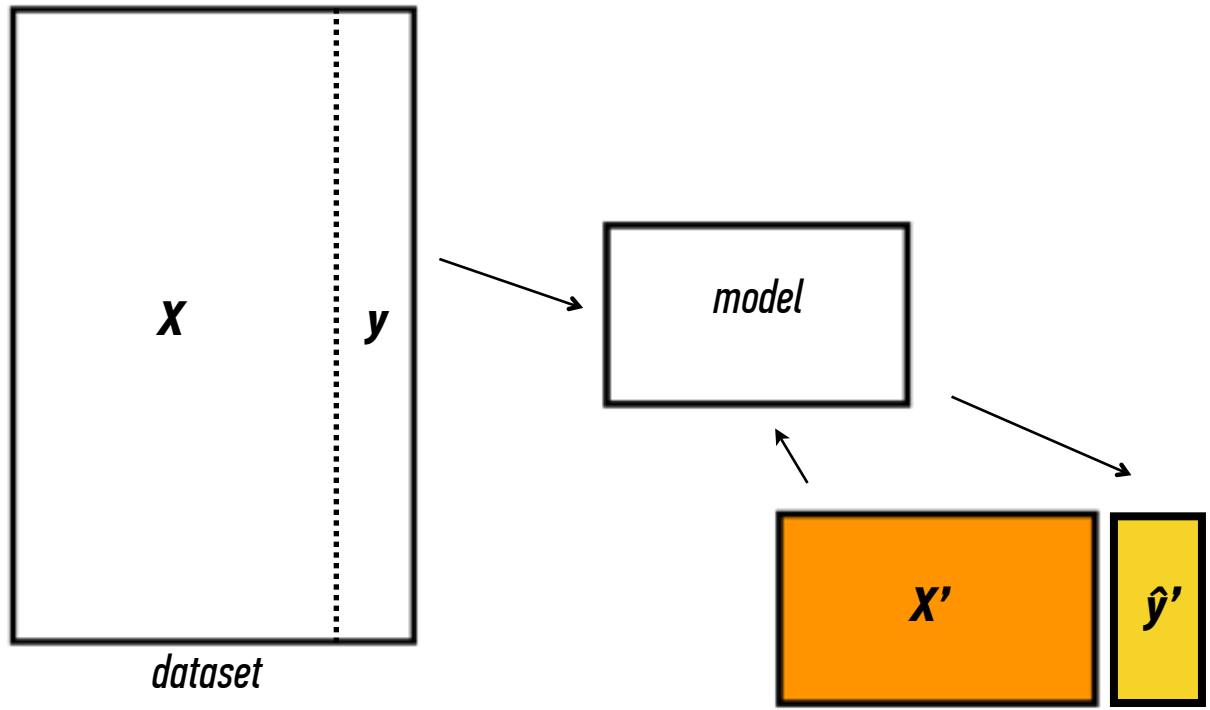
1) train model
2) make predictions



Q: How does a classification problem work?

1) train model
2) make predictions

model applies what it learned to new dataset ***X'***



Q: How does a classification problem work?

NOTE

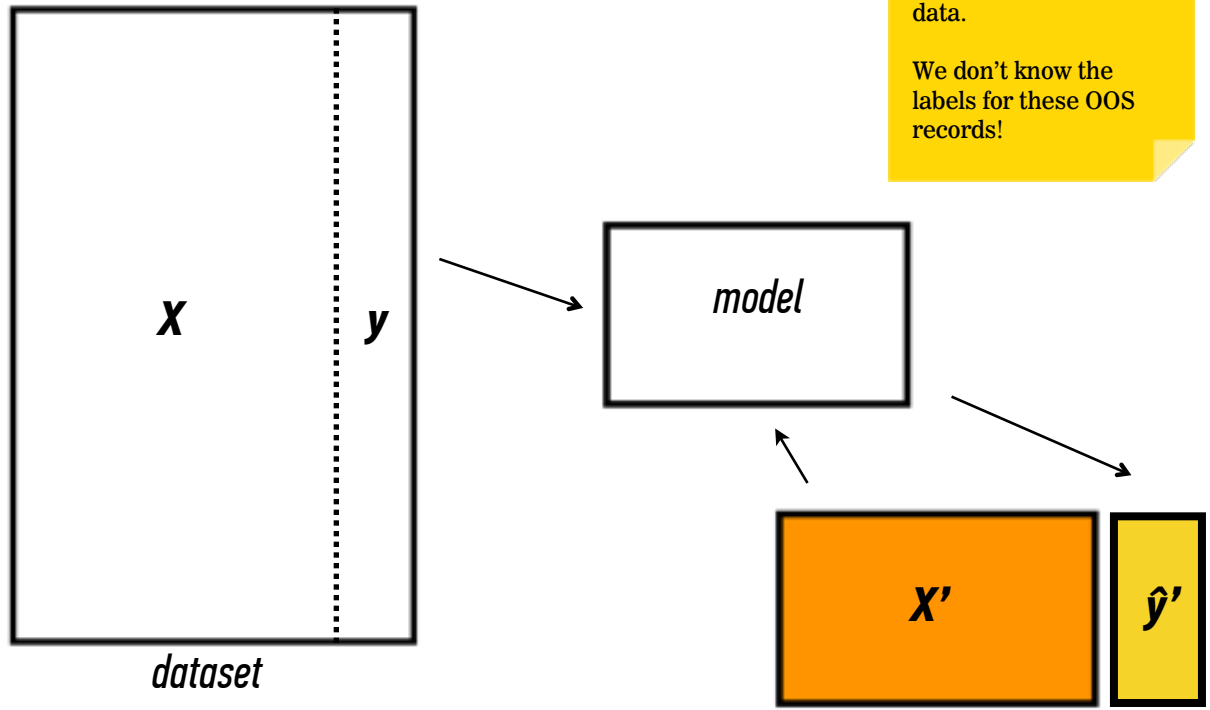
This new data is called *out of sample* data.

We don't know the labels for these OOS records!

1) train model

2) make predictions

model applies what it learned to new dataset ***X'***

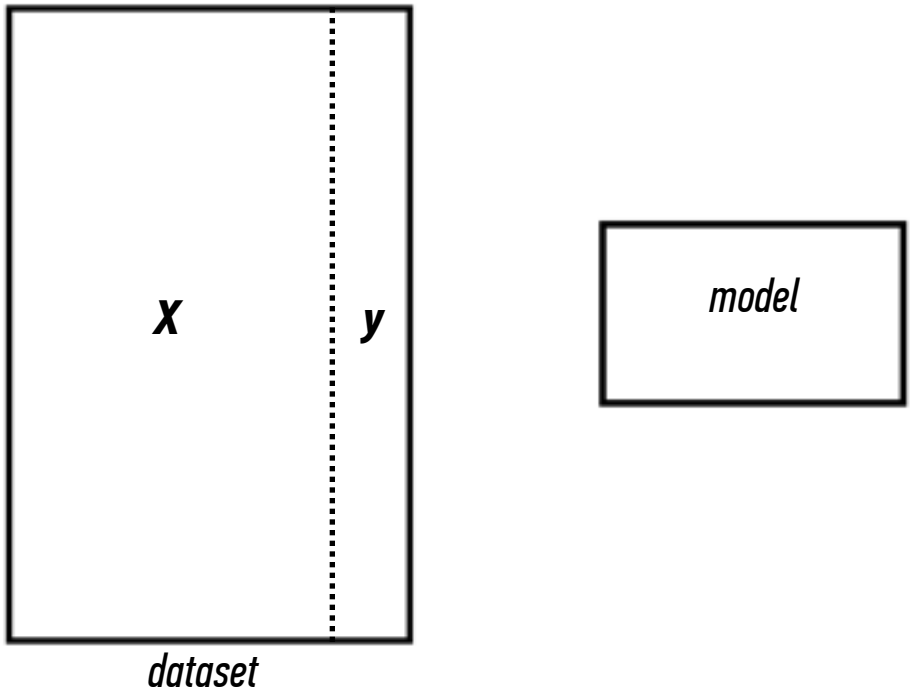


QUESTION

HOW DO YOU MEASURE THE QUALITY?

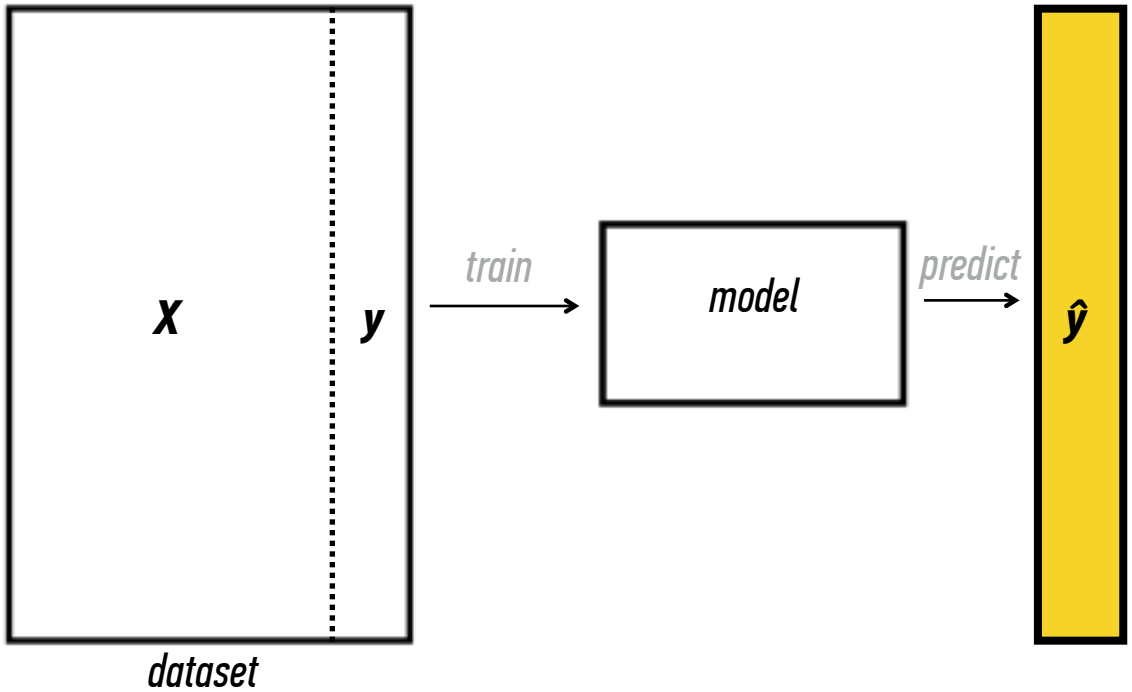
supervised	*test out your predictions*

Q: How do we test the model's predictions?



Q: How do we test the model's predictions?

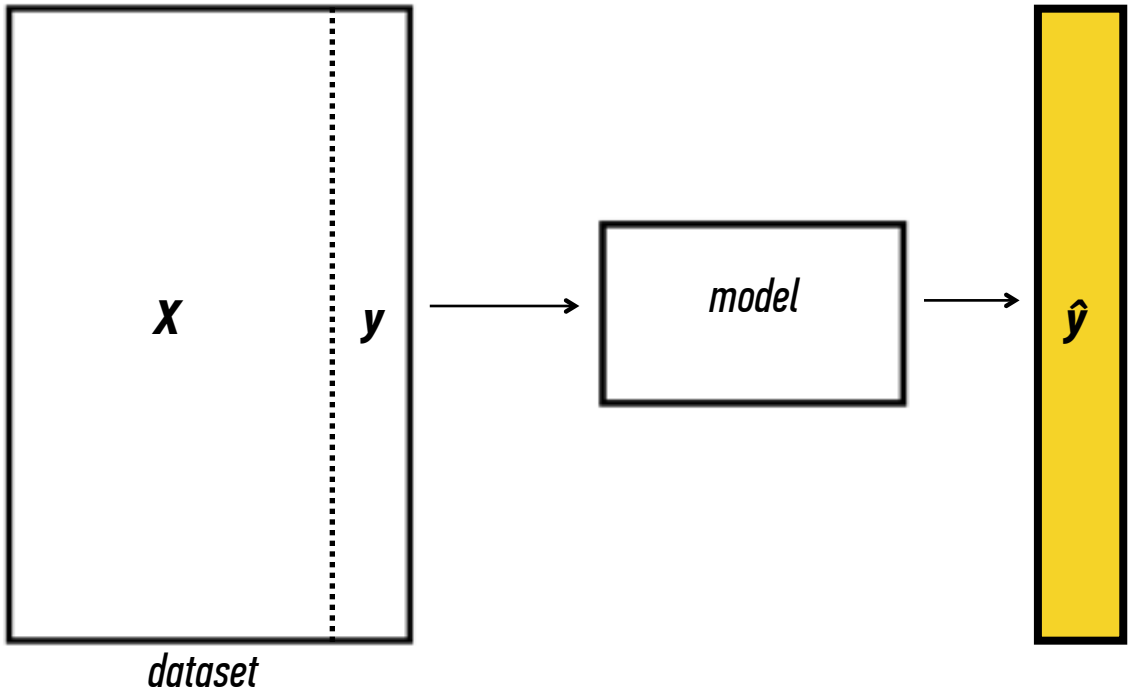
*We could apply the model on the given dataset **X** and test predictions **y***



Q: How do we test the model's predictions?

*We could apply the model on the given dataset **X** and test predictions **y***

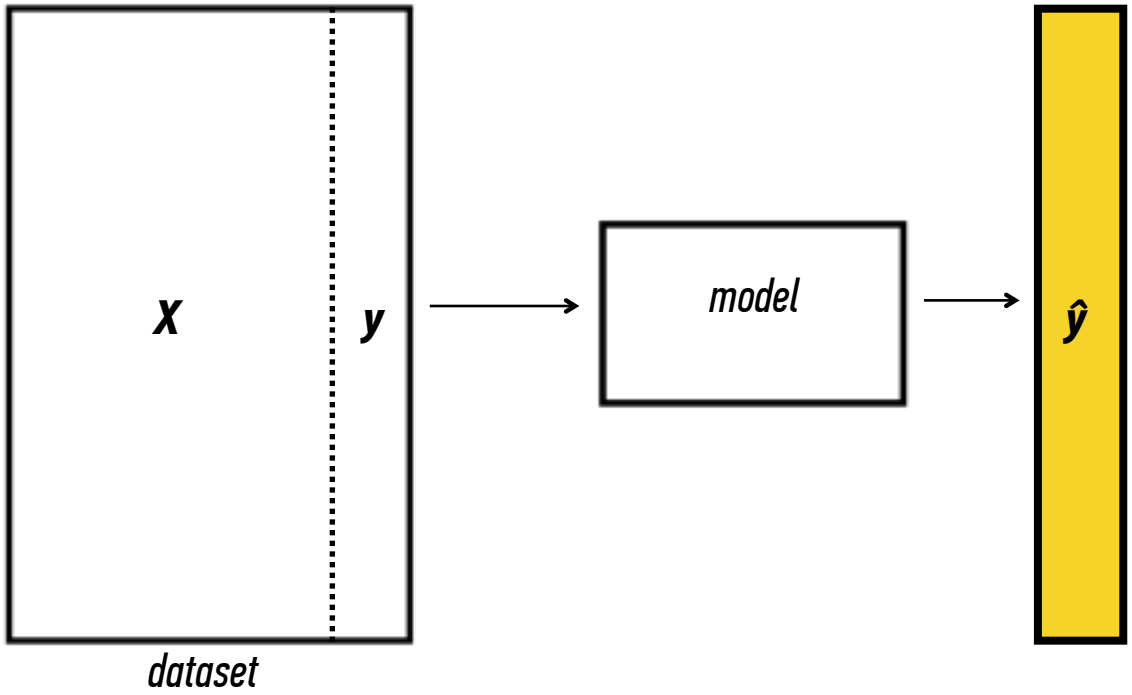
What could possibly go wrong here?



Q: How do we test the model's predictions?

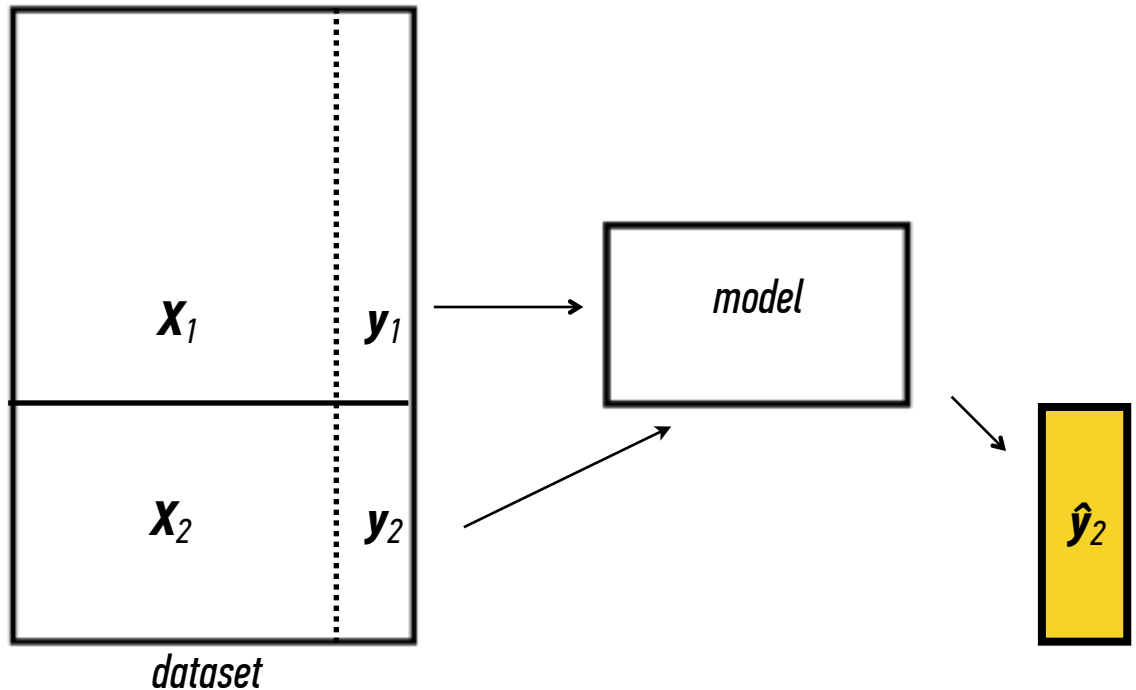
*We could apply the model on the given dataset **X** and test predictions **y***

Model could just have memorized all labels
(like a cheating student)



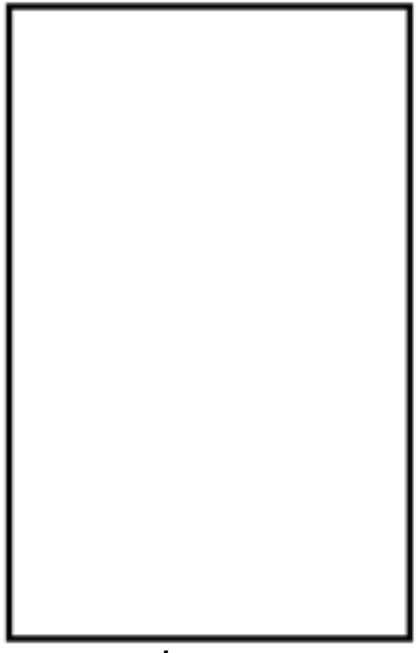
Q: How do we test the model's predictions?

*Train model on a part of **X**, and test the results on the rest of the data*



Q: What steps does a classification problem require?

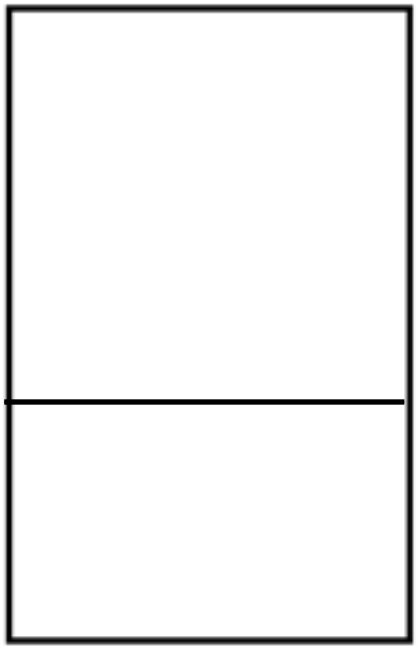
Q: What steps does a classification problem require?



dataset

Q: What steps does a classification problem require?

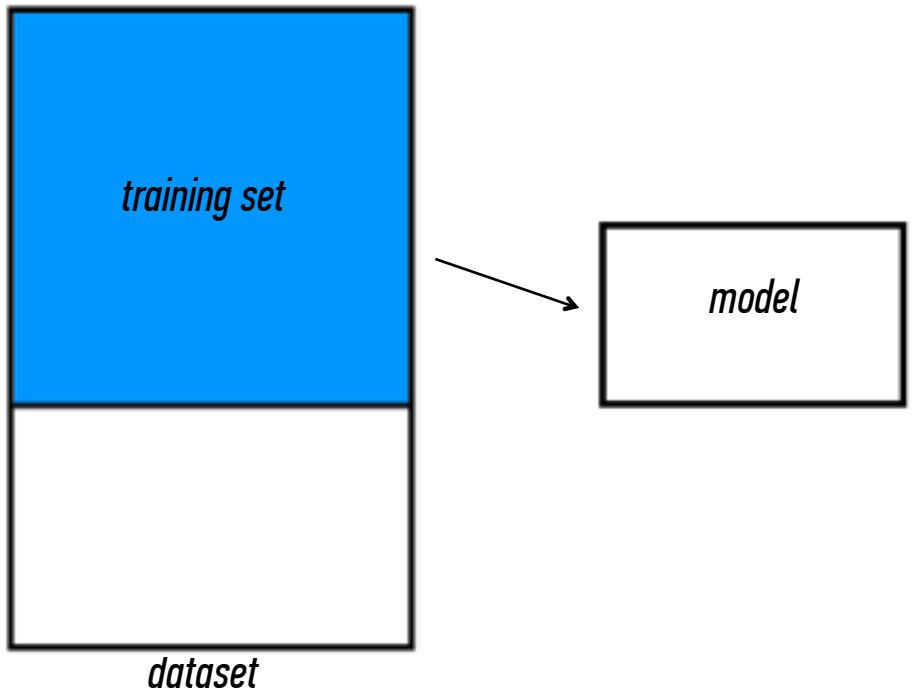
1) *split dataset*



dataset

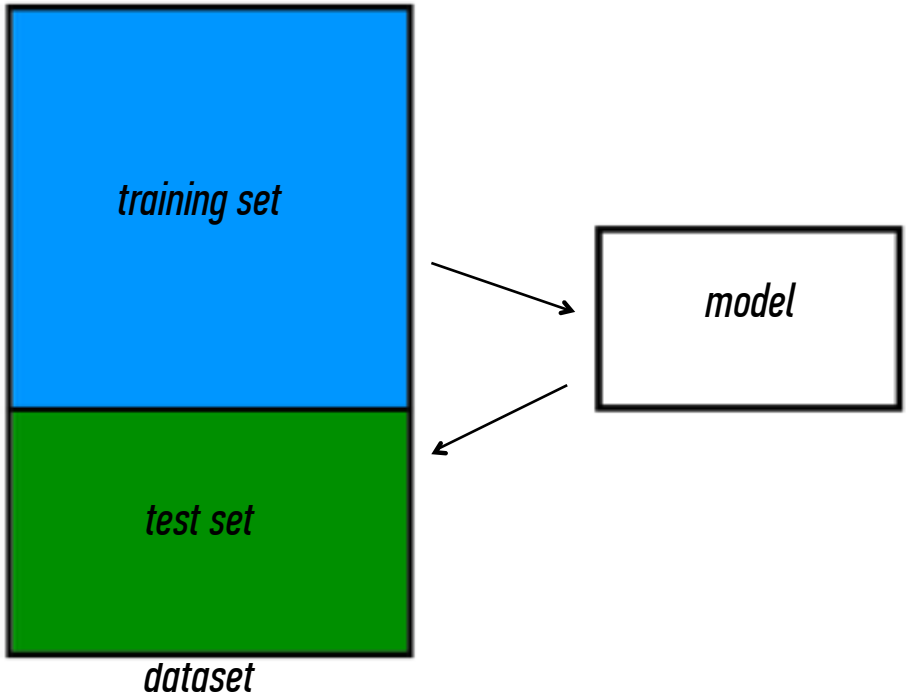
Q: What steps does a classification problem require?

1) *split dataset*
2) *train model*



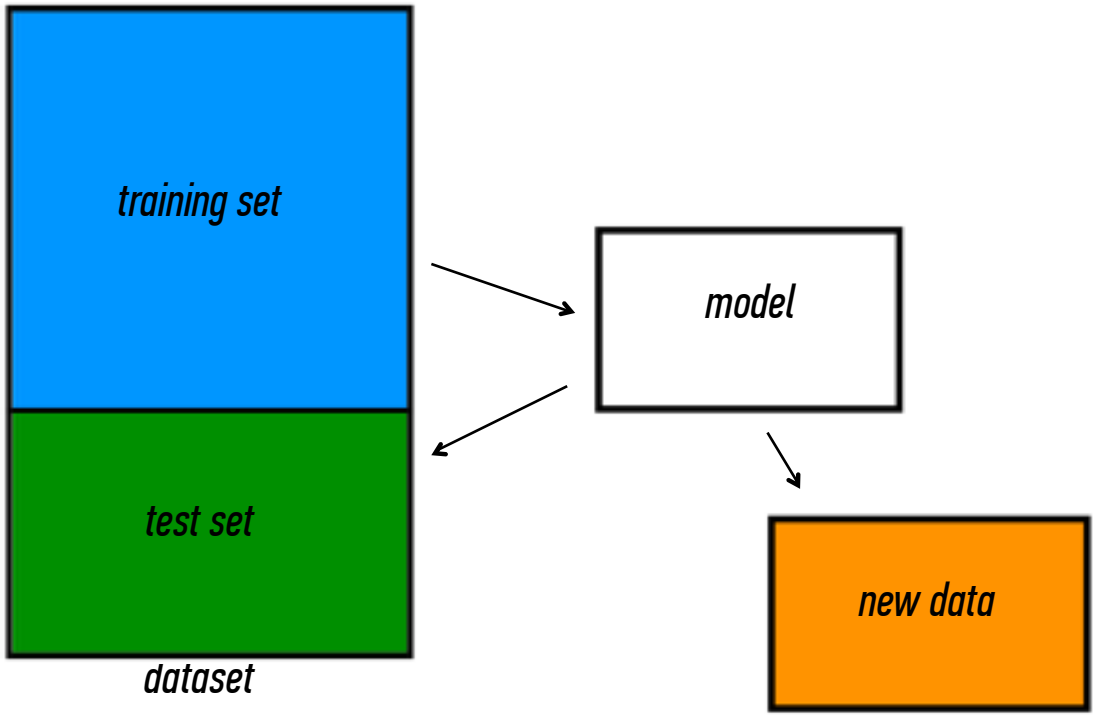
Q: What steps does a classification problem require?

1) *split dataset*
2) *train model*
3) *test model*



Q: What steps does a classification problem require?

1) *split dataset*
2) *train model*
3) *test model*
4) *make predictions*



	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

All supervised machine learning problems require using a training and test set

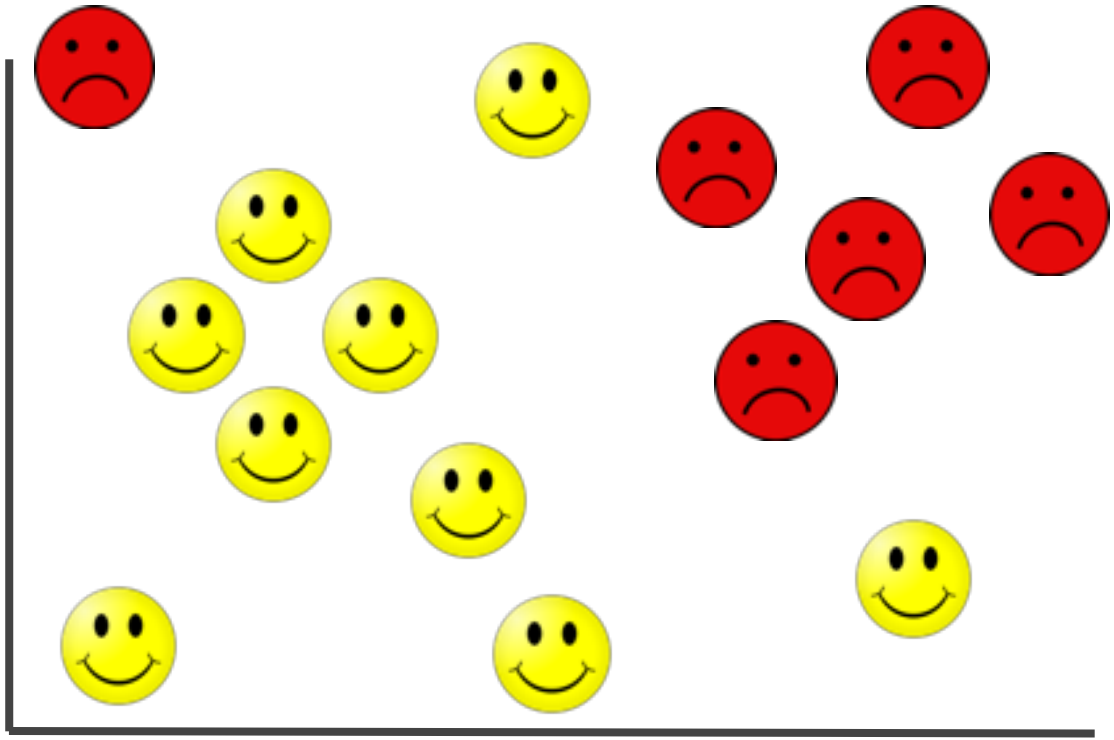
KNN CLASSIFICATION

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

kNN

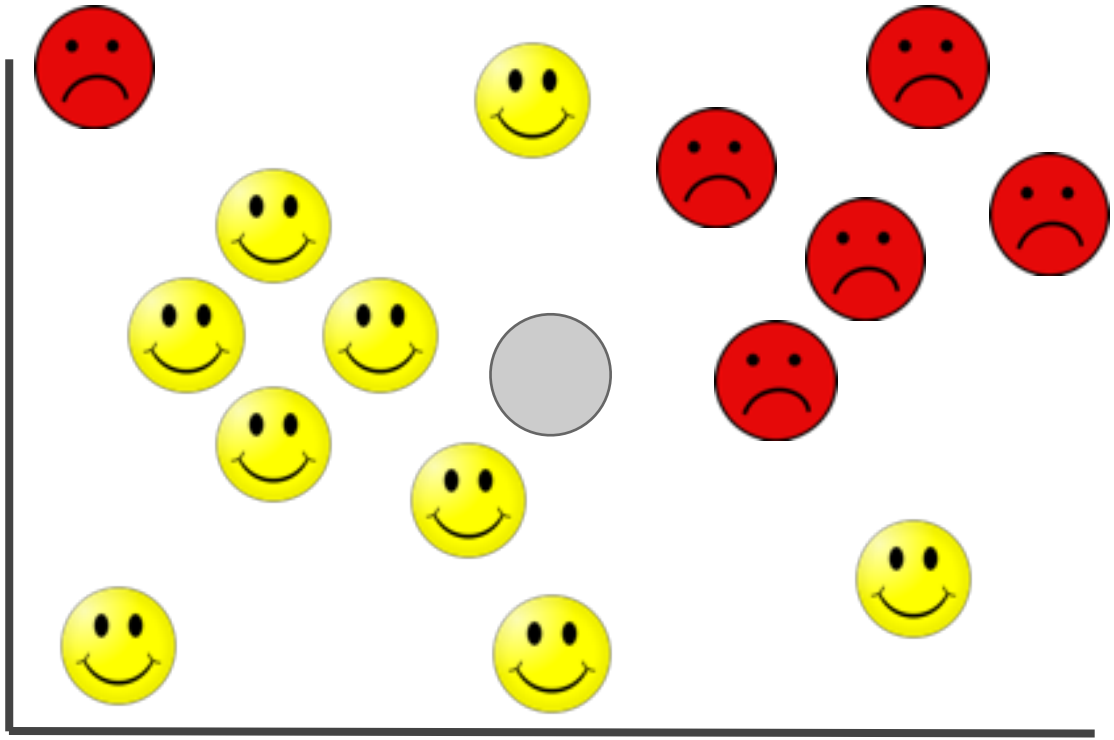
Supervised problem (labeled data)

Categorical data (happy vs. sad)



Want to predict:

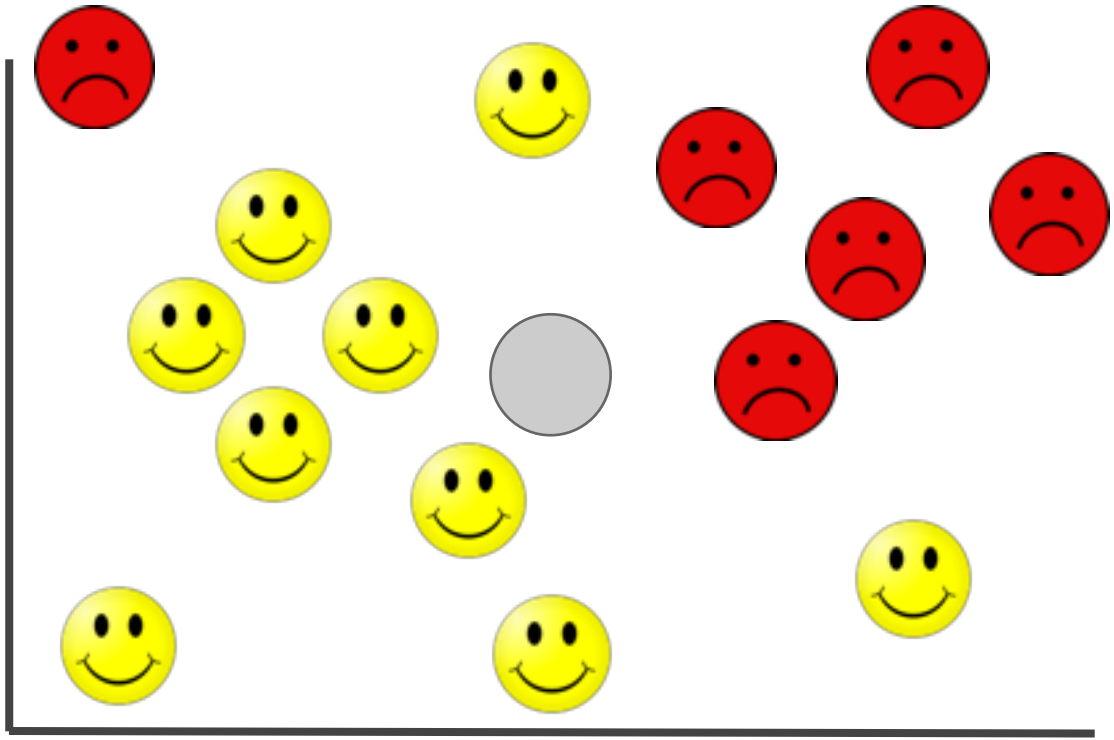
is the grey face happy?



Want to predict:

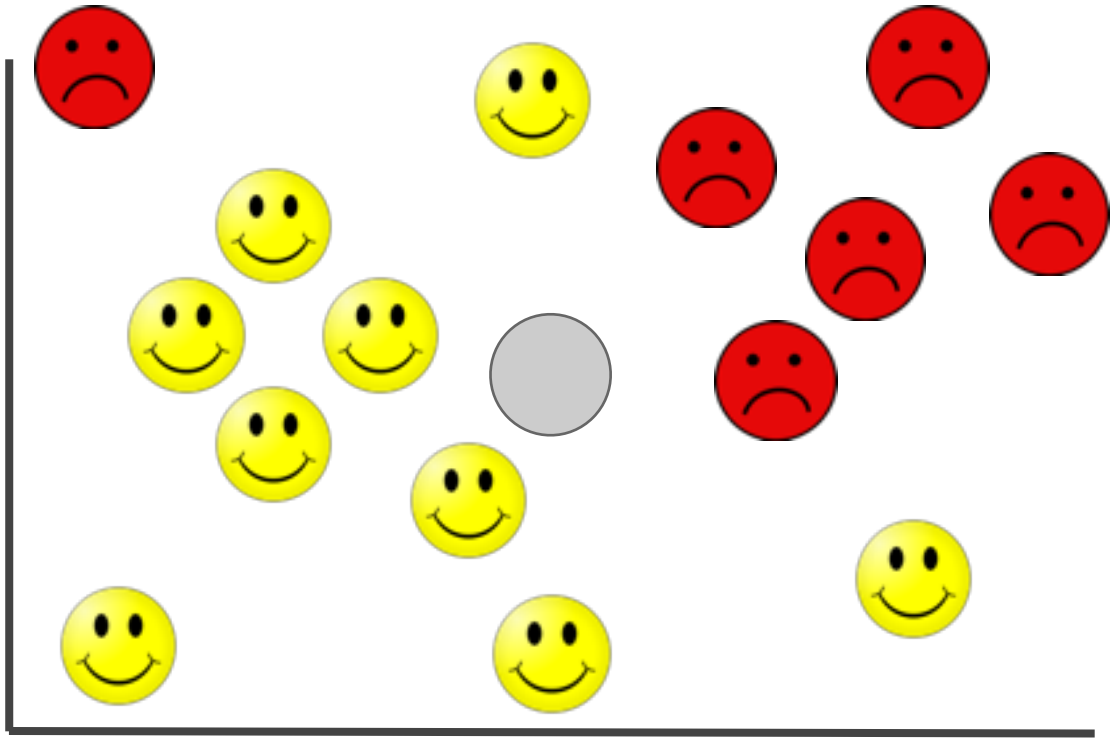
is the grey face happy?

*what do **you** think?*
why?



Choose k

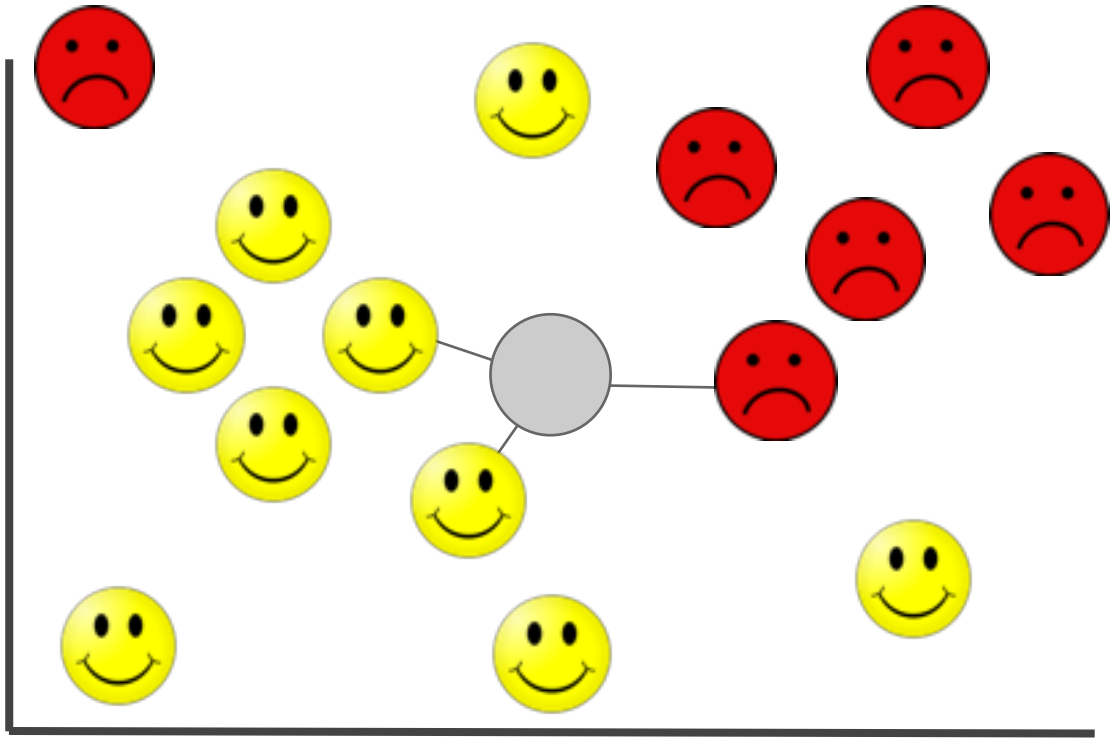
e.g., k = 3



Choose k

e.g., k = 3

Find k nearest neighbors

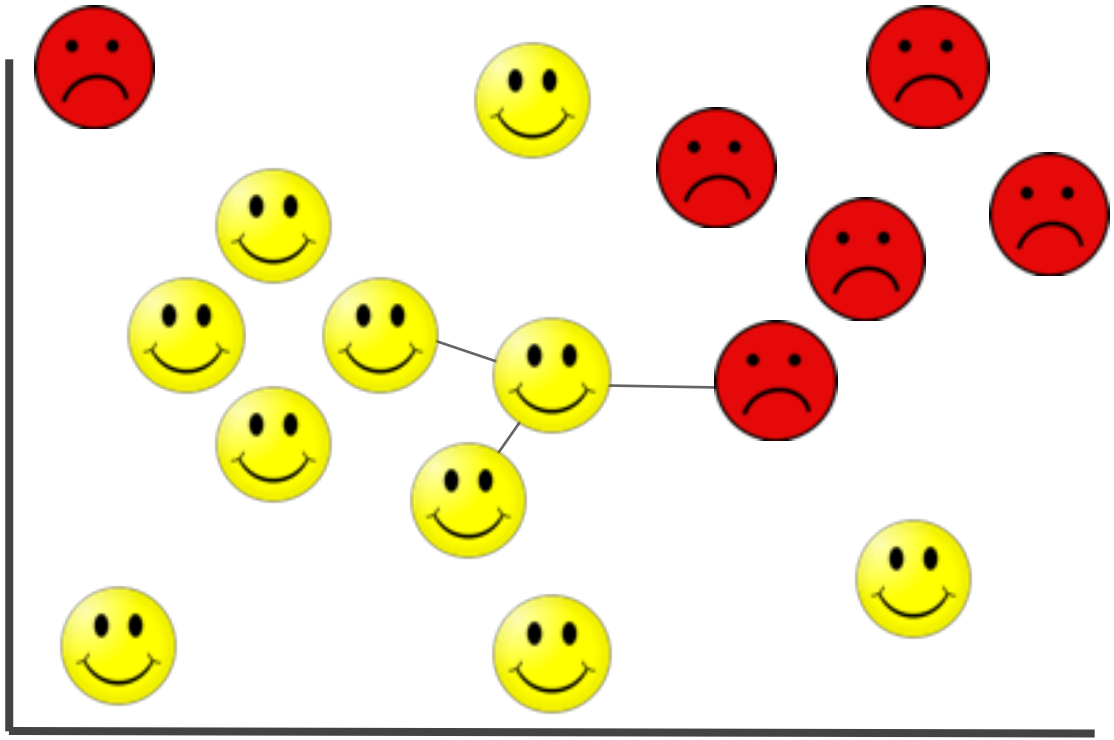


Choose k

e.g., k = 3

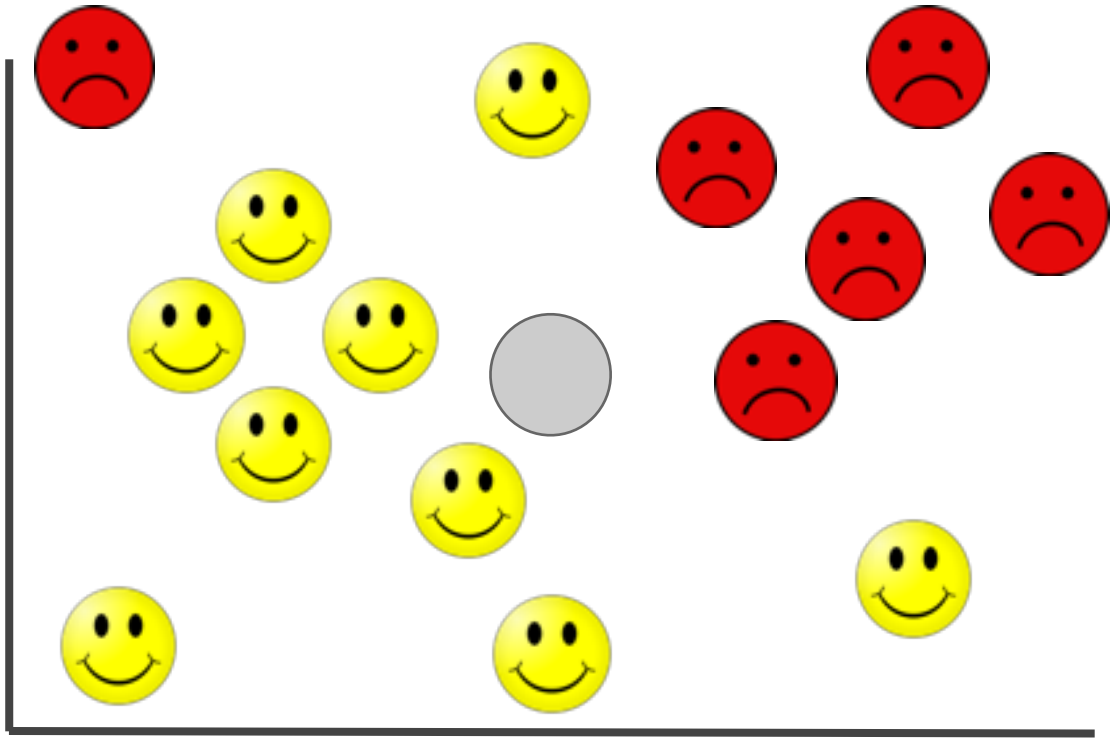
Find k nearest neighbors

Take majority vote



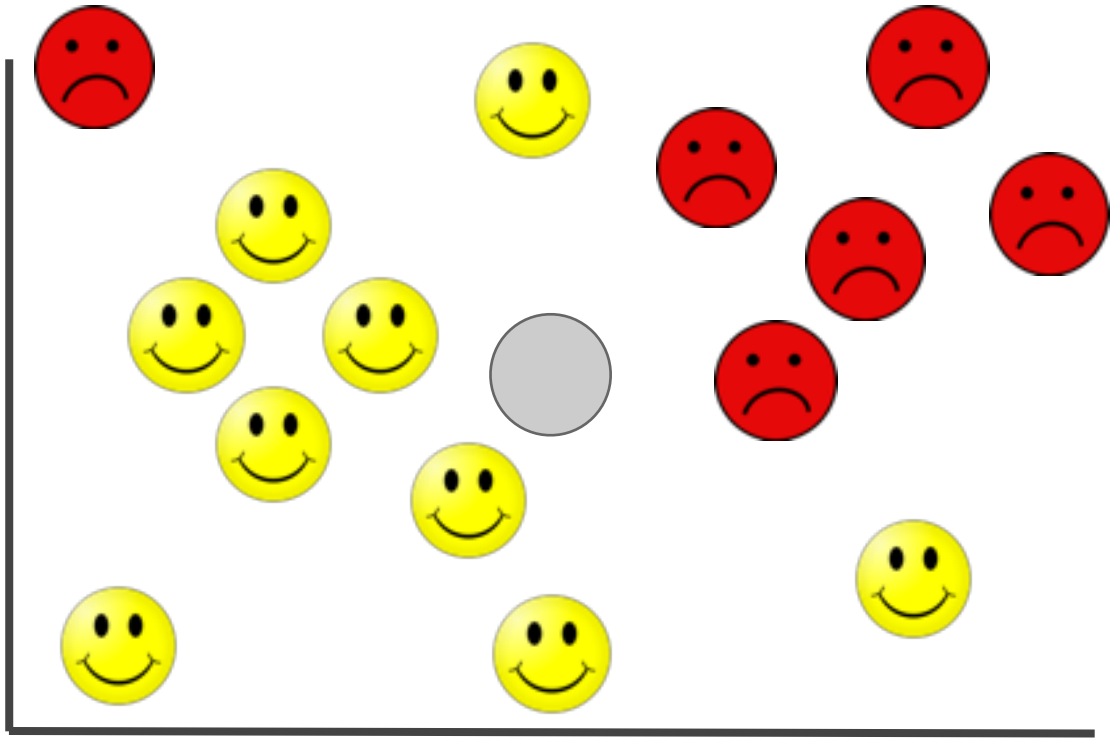
CAVEATS OF KNN

Q: What could possibly go wrong here?



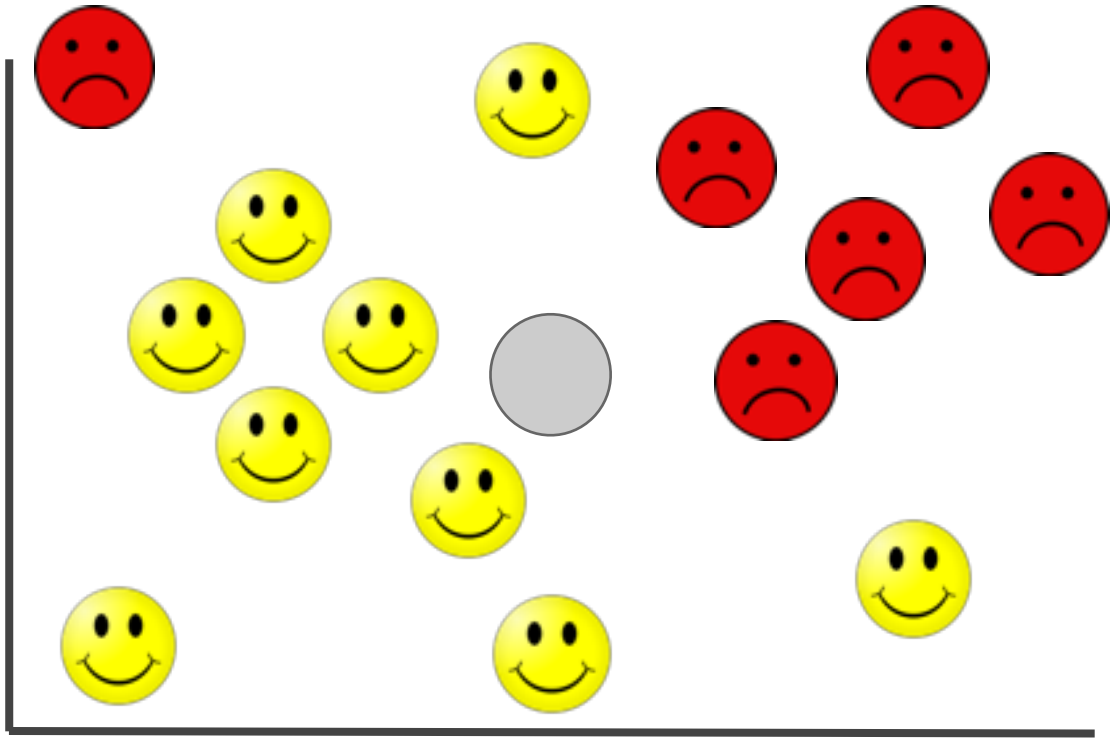
Q: What could possibly go wrong here?

What k?



Q: What could possibly go wrong here?

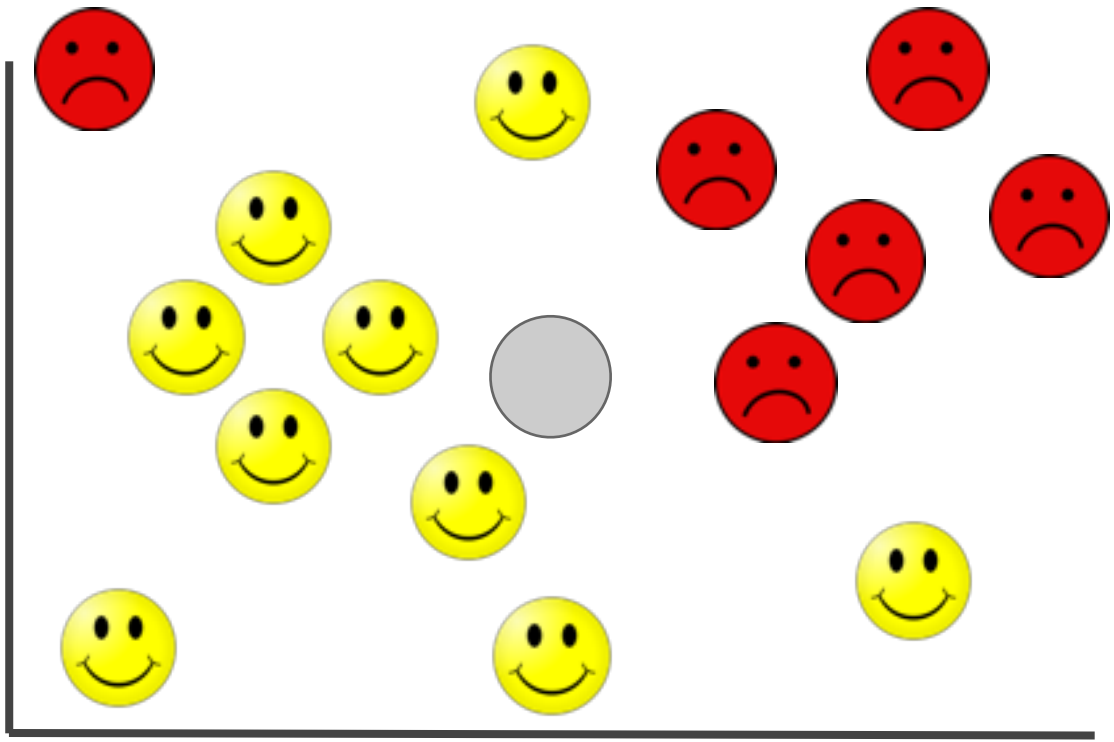
What k?
What if k = 1000?



Q: What could possibly go wrong here?

What k?
What if k = 1000?

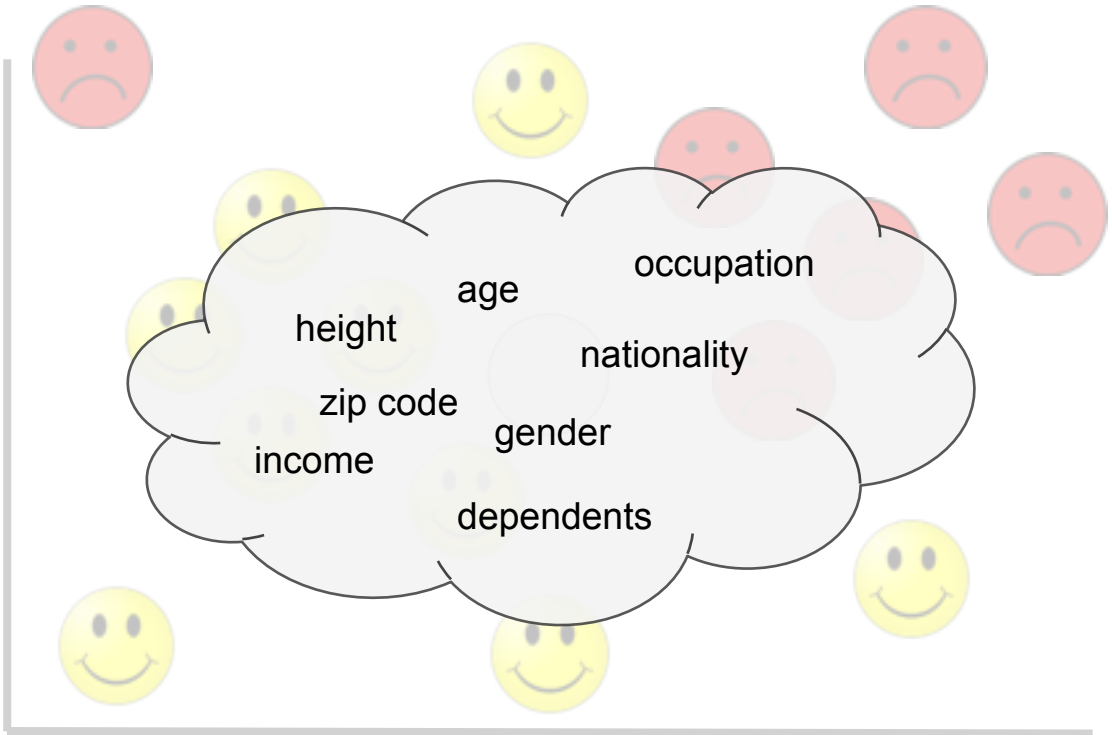
What is 'nearest'?



Q: What could possibly go wrong here?

What k?
What if k = 1000?

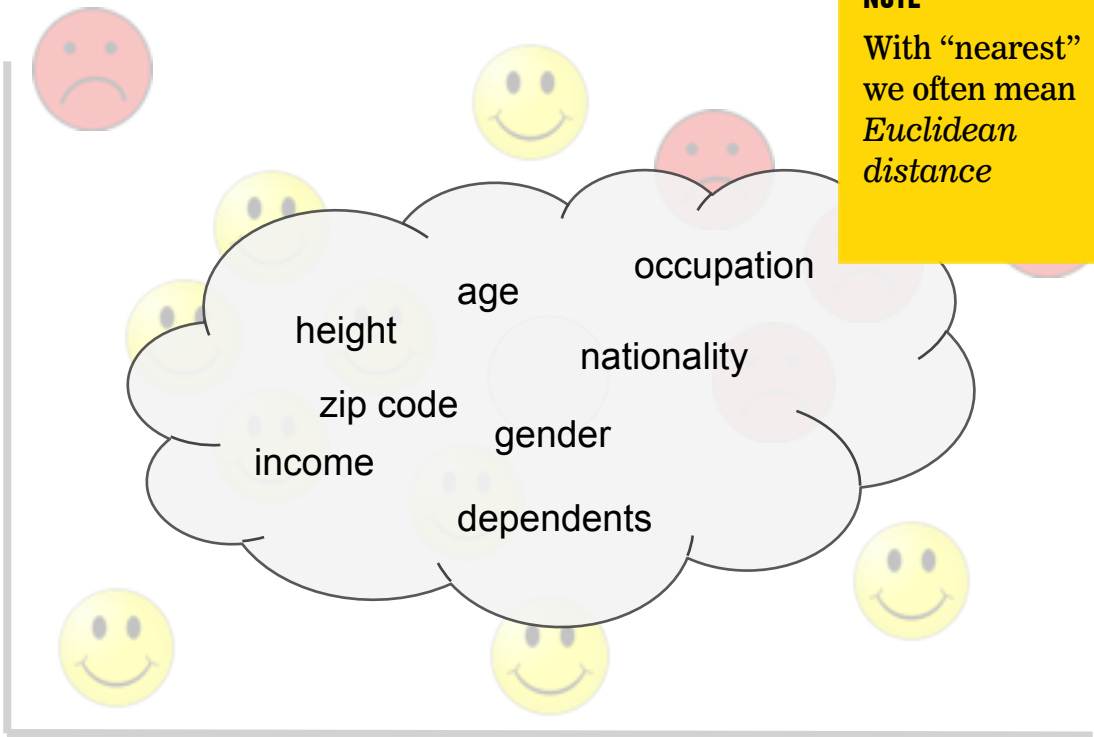
What is 'nearest'?



Q: What could possibly go wrong here?

What k?
What if k = 1000?

What is 'nearest'?



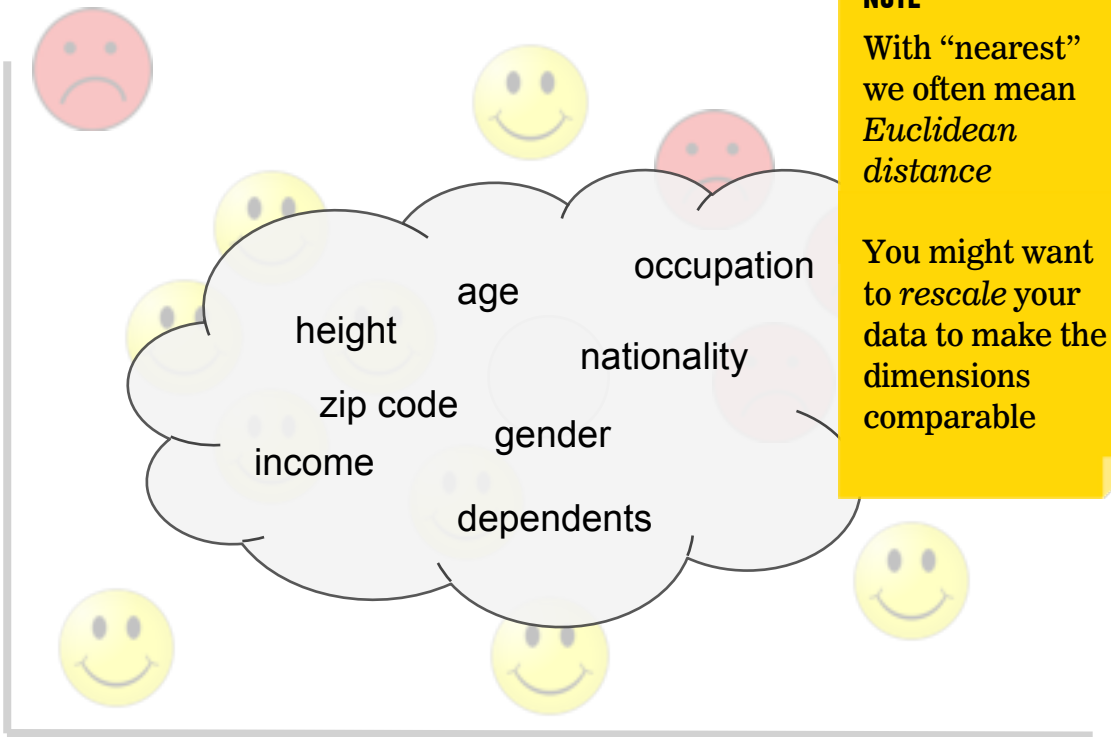
NOTE

With "nearest" we often mean *Euclidean distance*

Q: What could possibly go wrong here?

What k?
What if k = 1000?

What is 'nearest'?



NOTE

With "nearest" we often mean *Euclidean distance*

You might want to *rescale* your data to make the dimensions comparable

DISCUSSION