GENERAL ASSEMBLY

K NEAREST NEIGHBORS

Brian Chung

Whats cool in ML? http://www.sciencemag.org/content/350/6266/1332.full

Review / Exit Tickets
- Numpy

Project Overview
- Various topics
- Milestone 1

AGENDA

Goals for the session:

- Finish up Pandas
- First Learning model!

PANDAS OVERVIEW

Let’s see some more interesting tools using **pandas**

Pandas– A scientific computing library built for python on top of numpy, providing high performance data structures and operations on those data structures

Brings data aggregation and split-apply-combine features to Python, ala Excel++.

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?

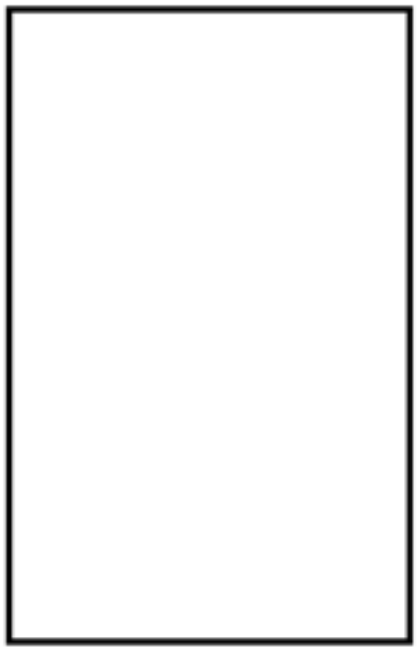
Q: What does "supervised" mean?

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

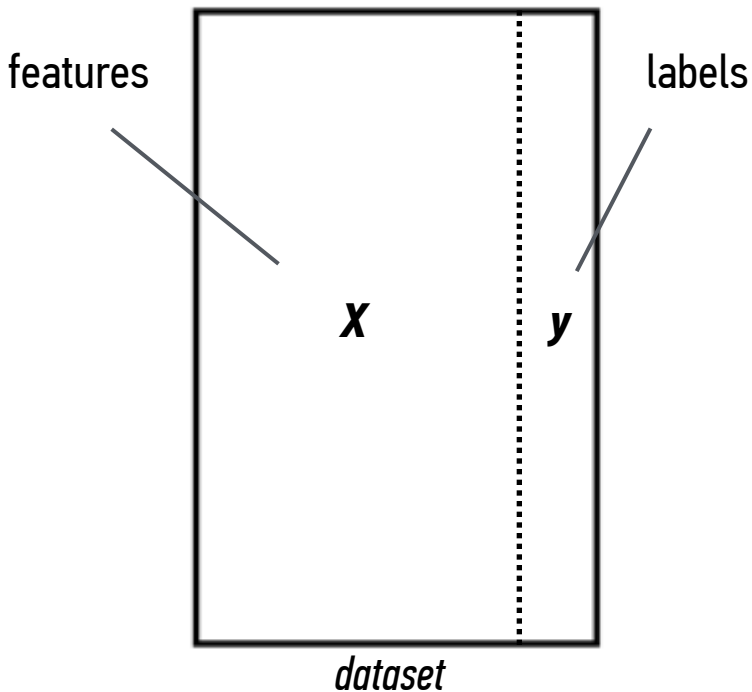
Q: How does a classification problem work?

Q: How does a classification problem work?

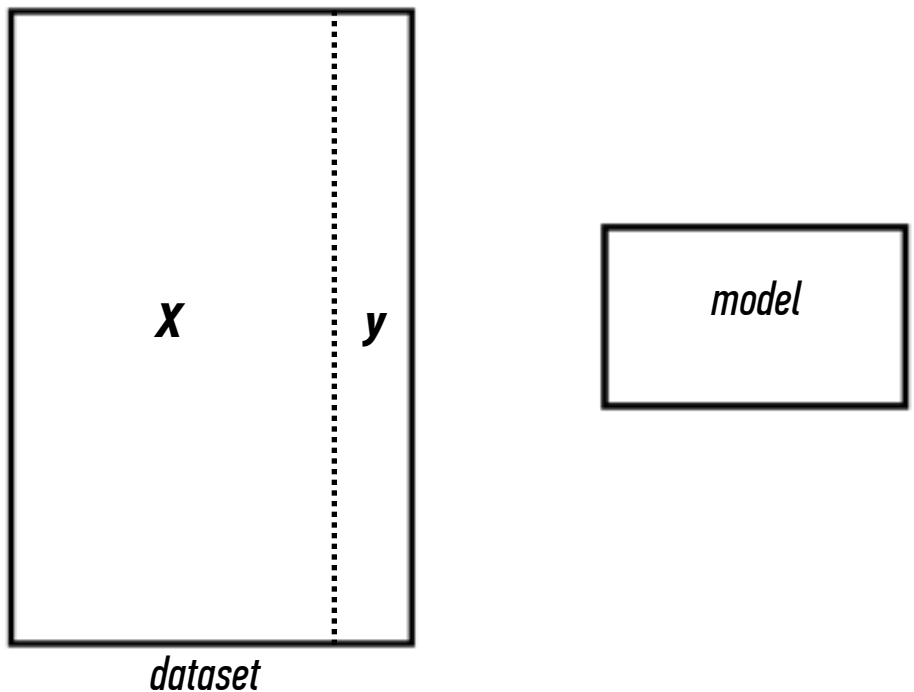


dataset

Q: How does a classification problem work?



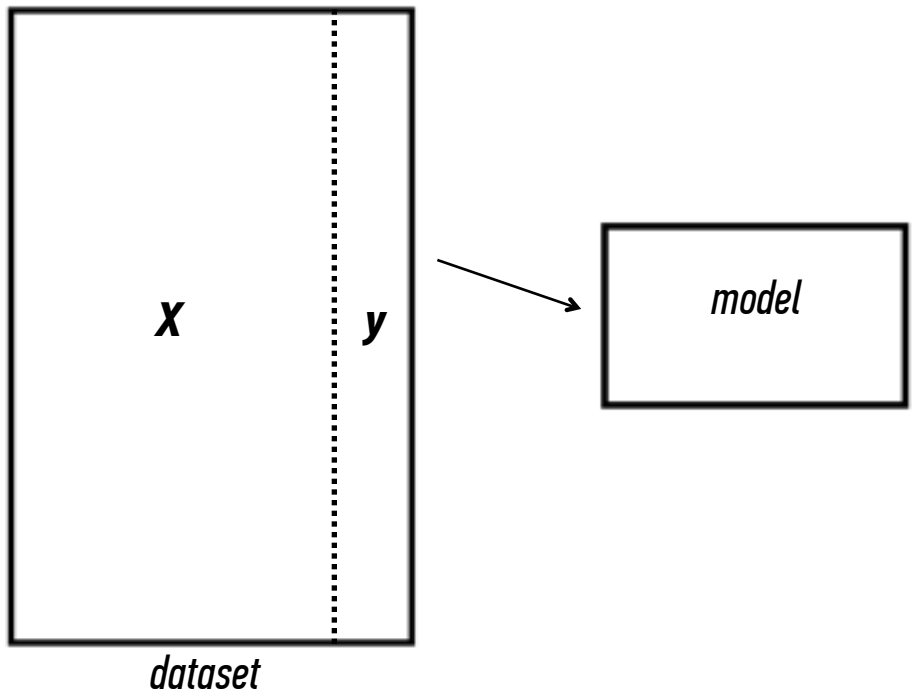
Q: How does a classification problem work?



Q: How does a classification problem work?

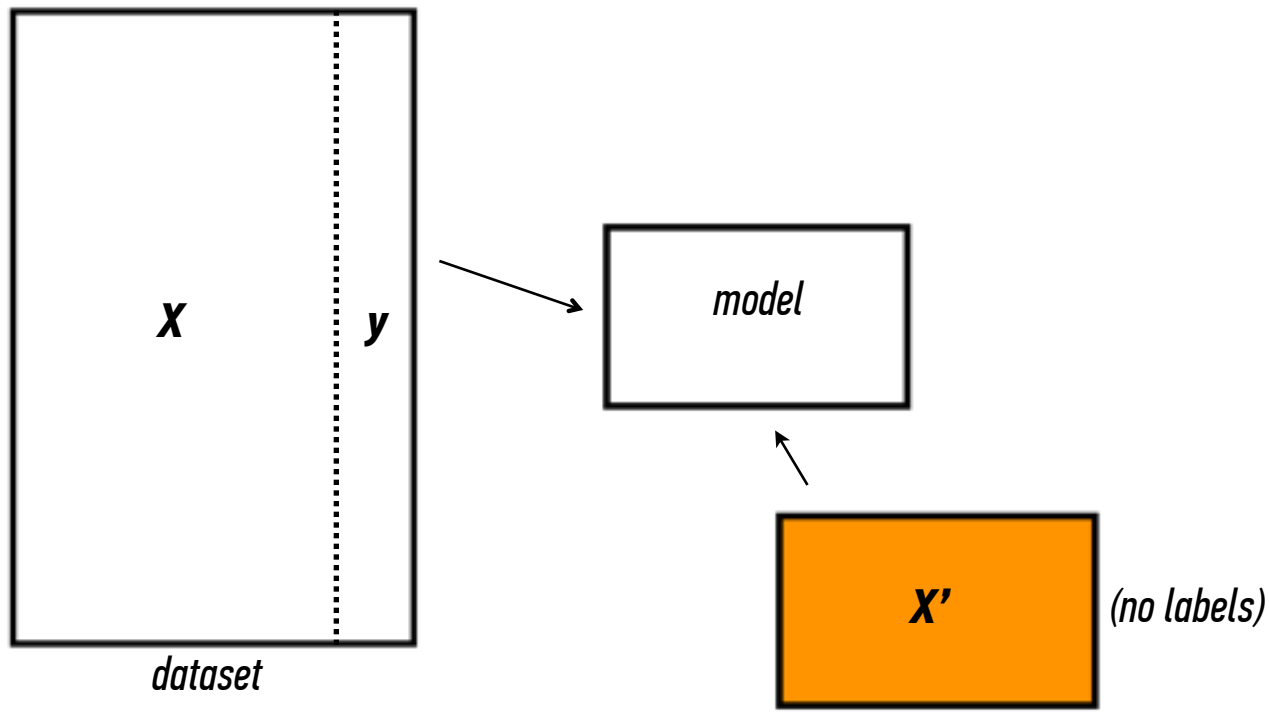
1) train model

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



Q: How does a classification problem work?

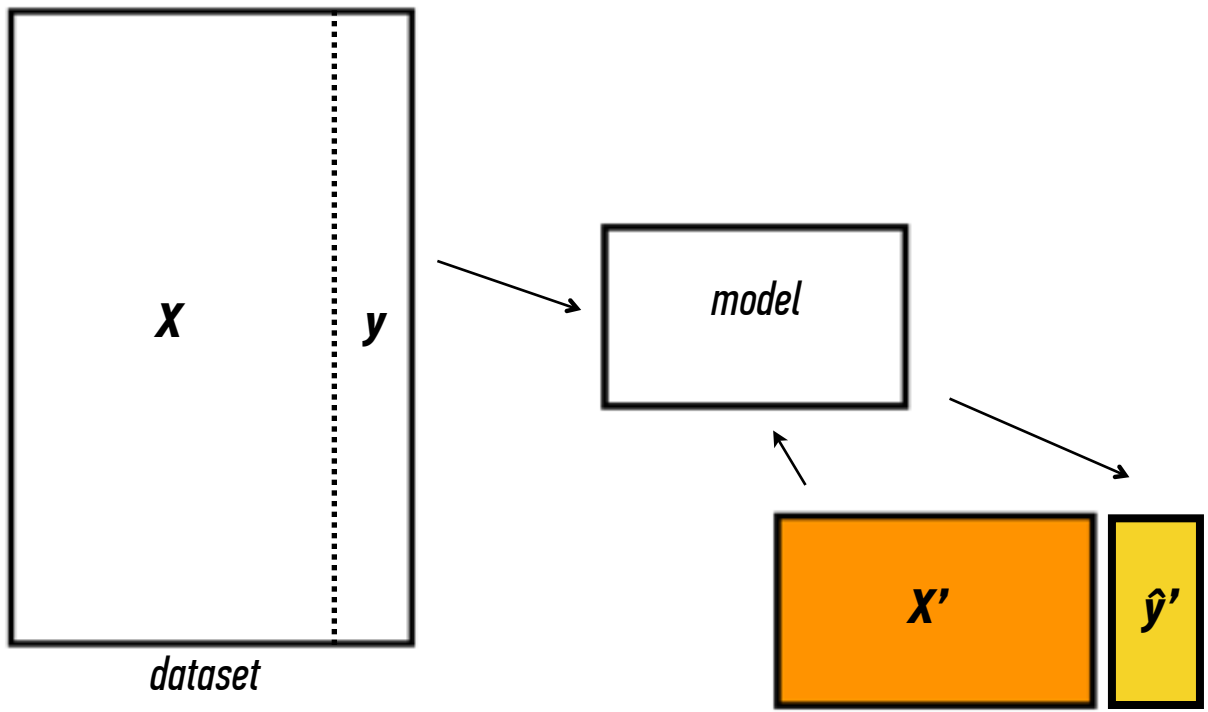
1) train model
2) make predictions



Q: How does a classification problem work?

1) train model
2) make predictions

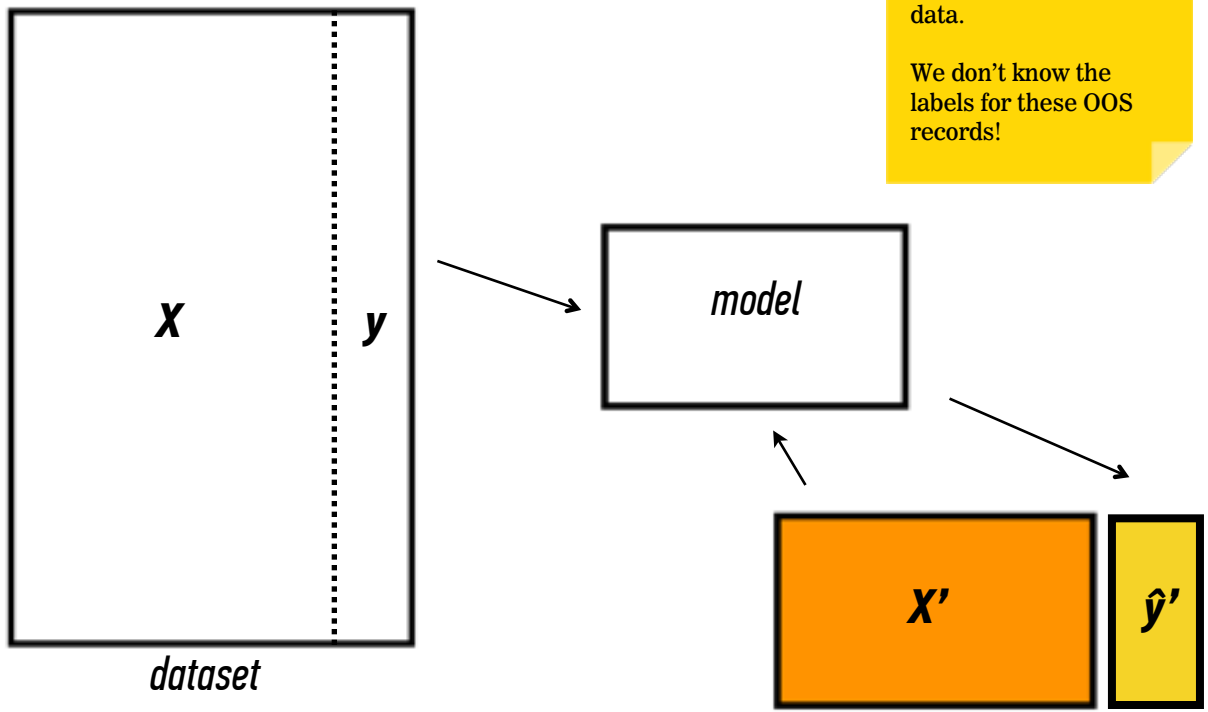
model applies what it learned to new dataset ***X'***



Q: How does a classification problem work?

1) train model
2) make predictions

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



NOTE

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

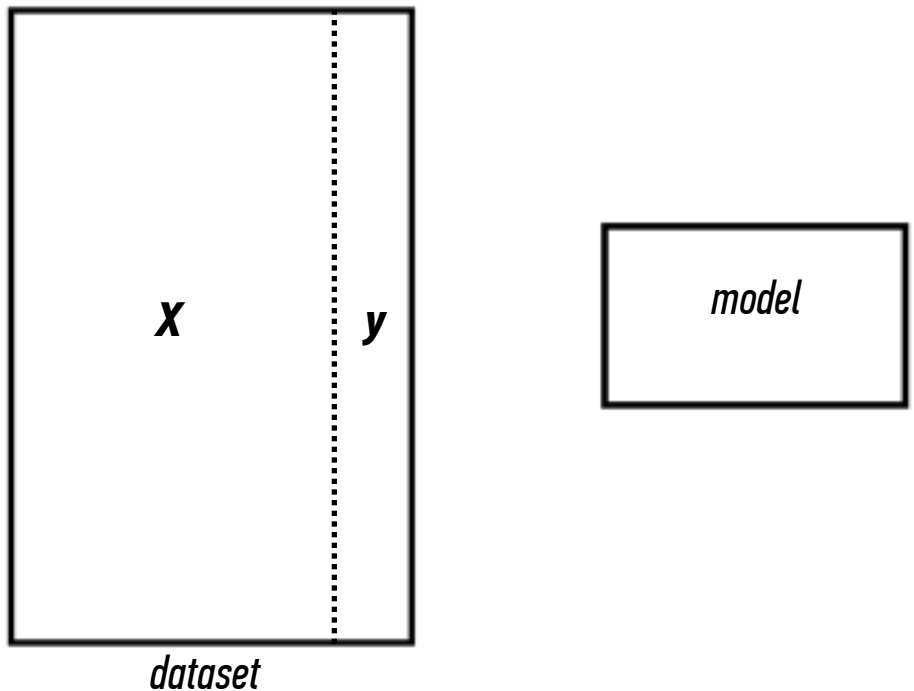
We don't know the labels for these OOS records!

QUESTION

HOW DO YOU MEASURE THE QUALITY?

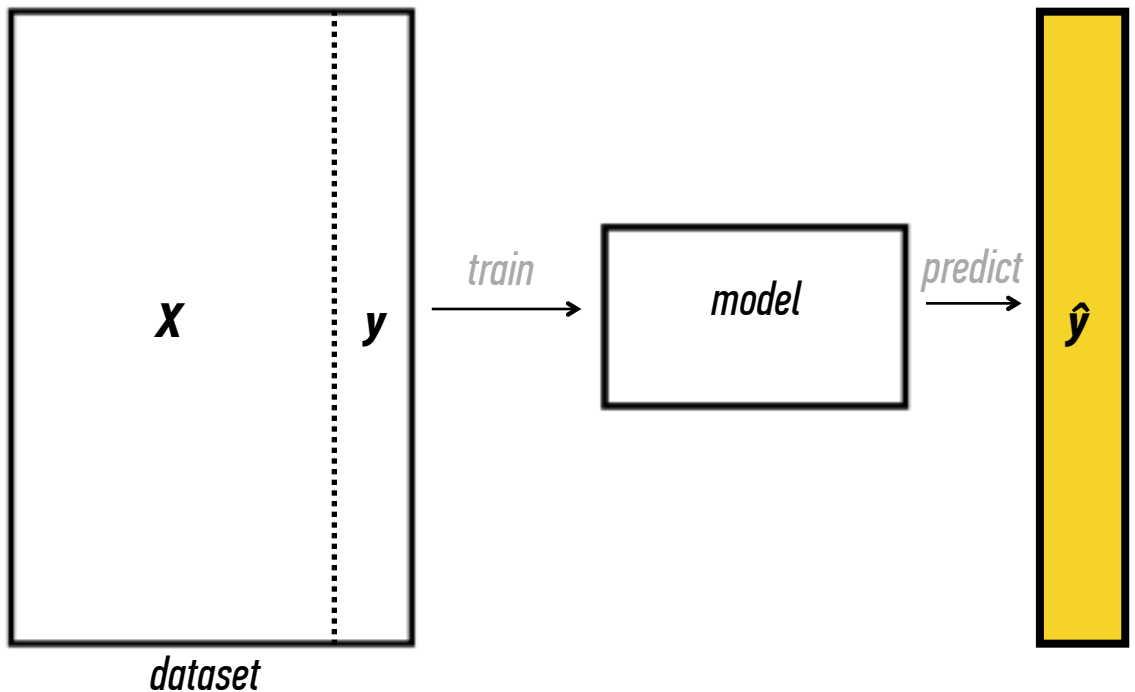
supervised	*test out your predictions*

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



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

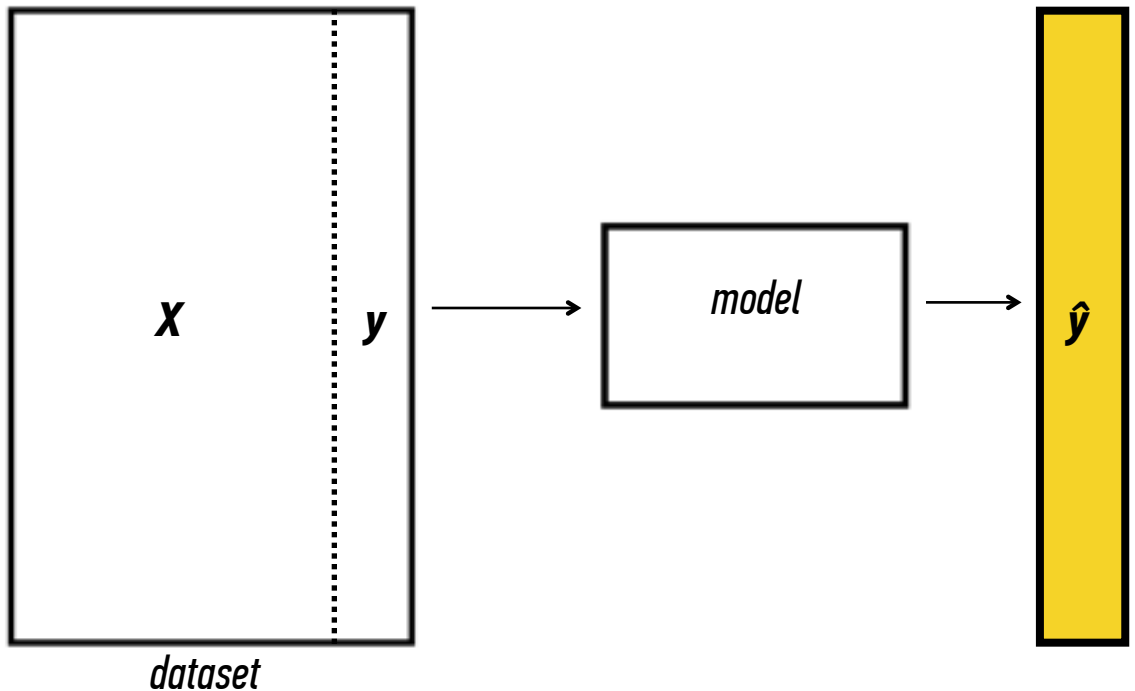
*We could apply the model on the given dataset **X** and test predictions **y***



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

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

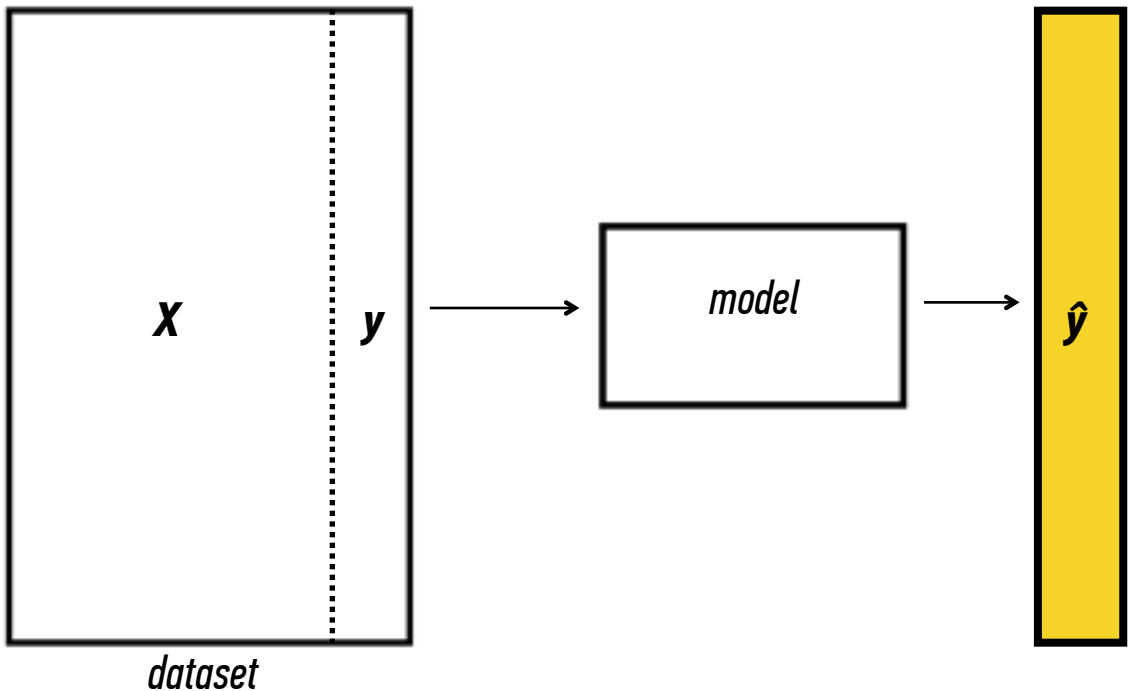
What could possibly go wrong here?



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

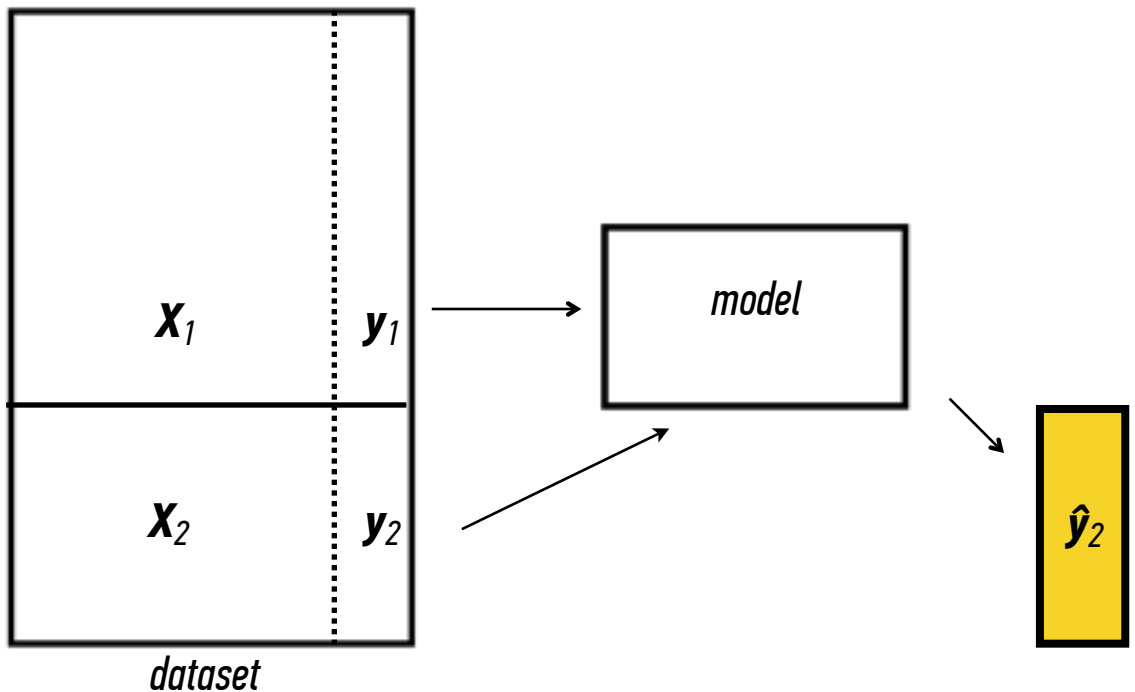
We could apply the model on the given dataset ***X*** *and test predictions* ***y***

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



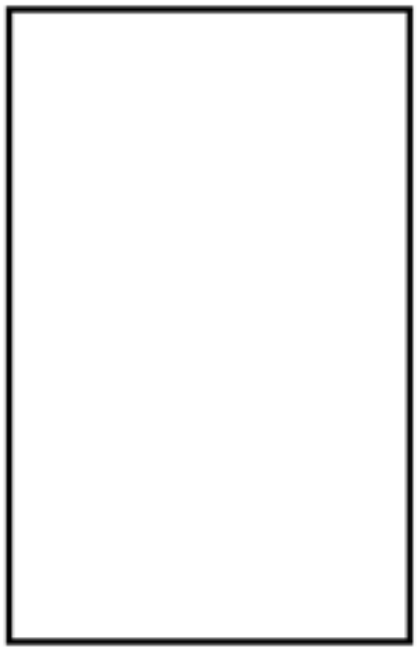
Q: How do we test the model's predictions?

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?

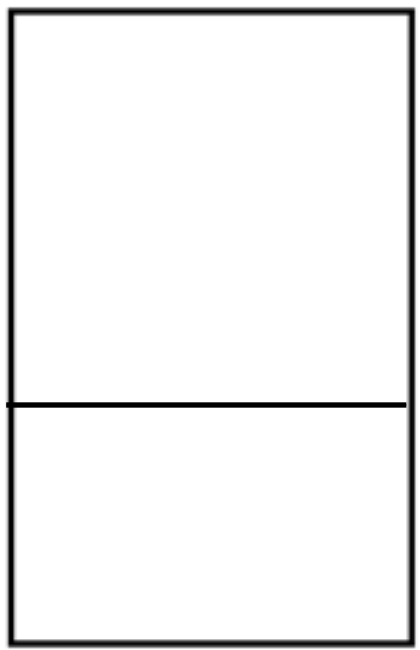
Q: What steps does a classification problem require?



dataset

Q: What steps does a classification problem require?

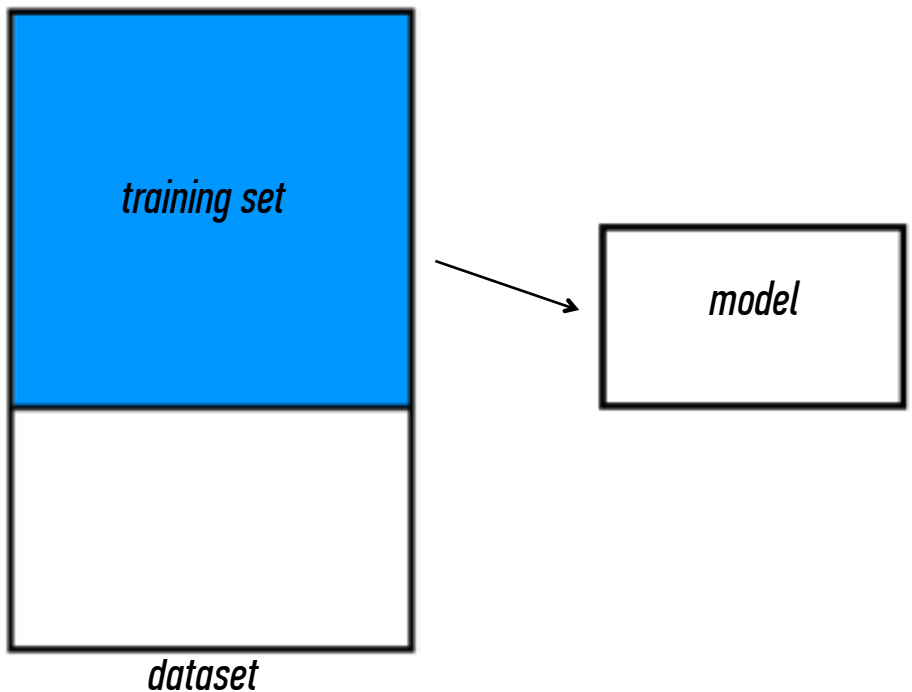
1) *split dataset*



dataset

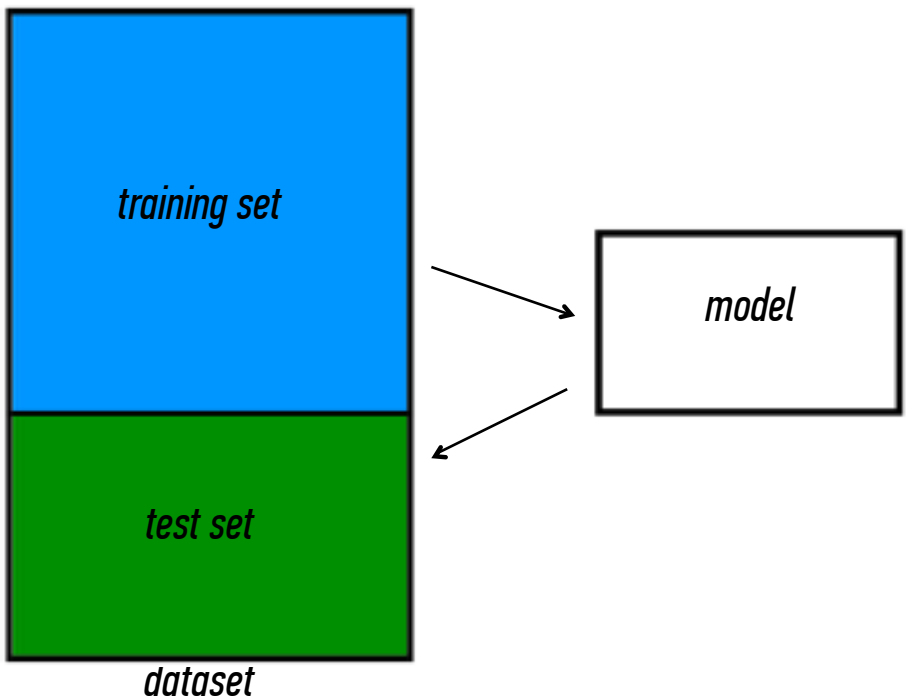
Q: What steps does a classification problem require?

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



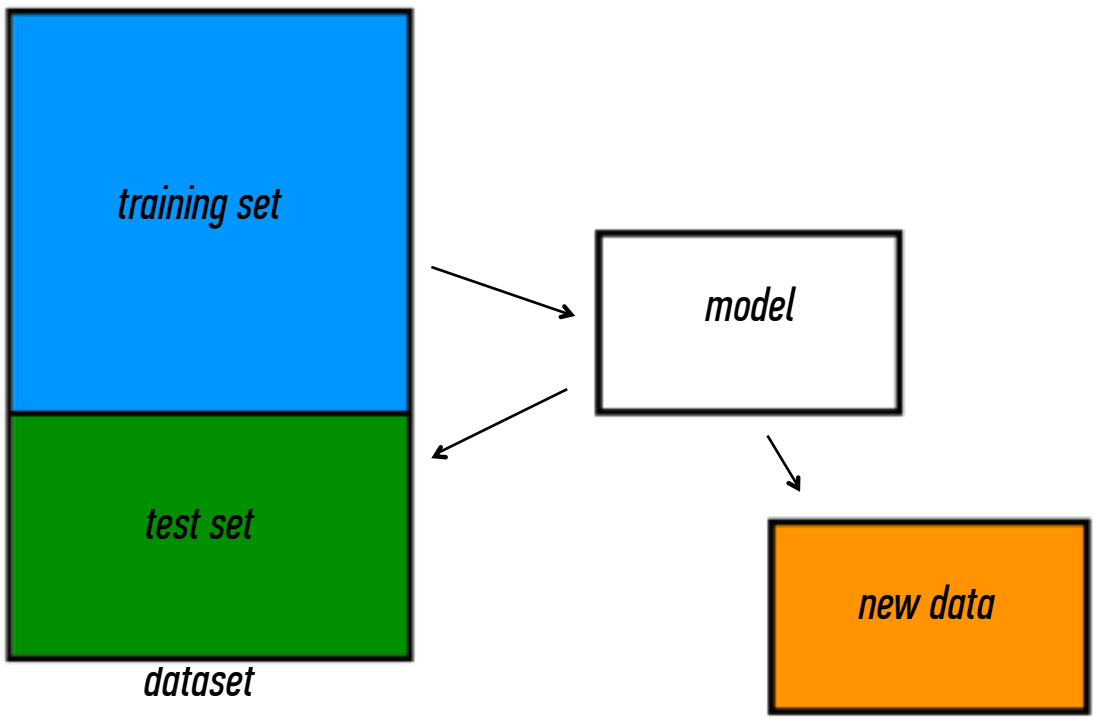
Q: What steps does a classification problem require?

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

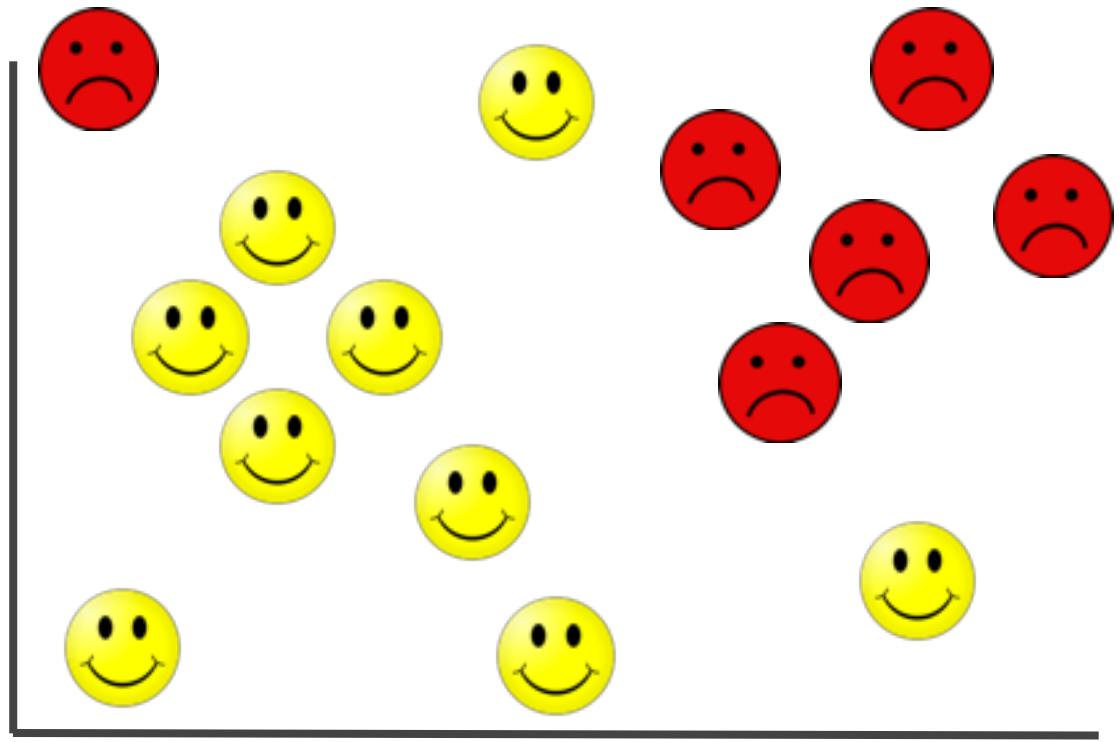
KNN CLASSIFICATION

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

kNN

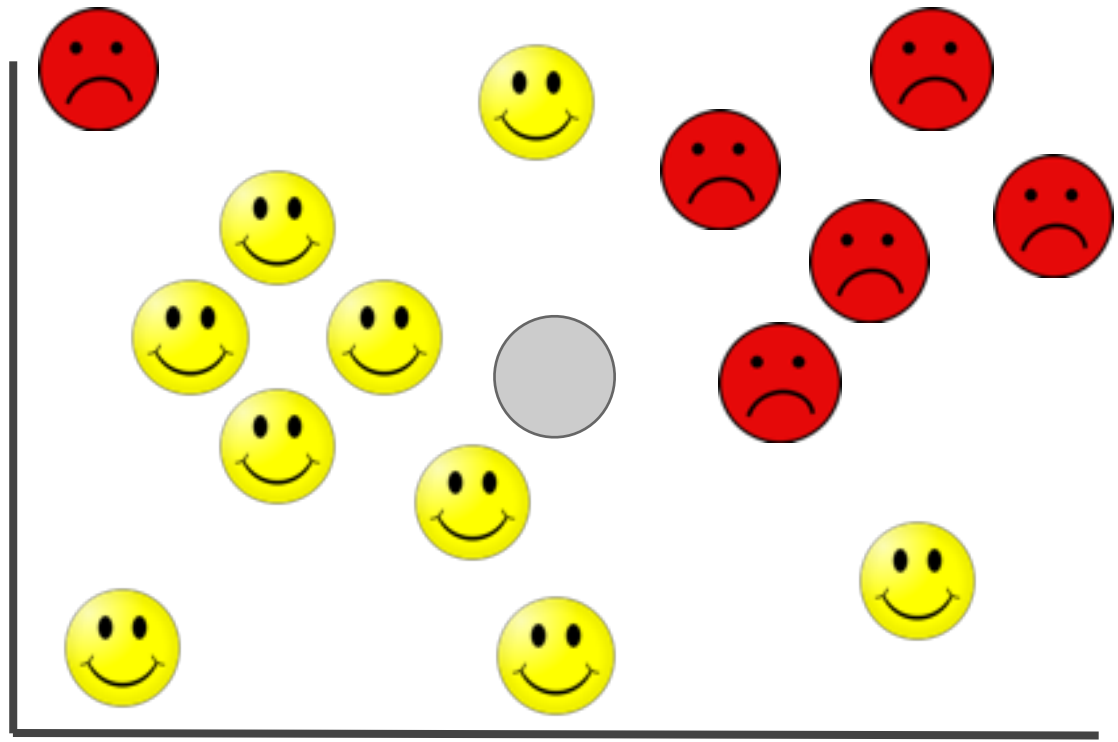
Supervised problem (labeled data)

Categorical data (happy vs. sad)



Want to predict:

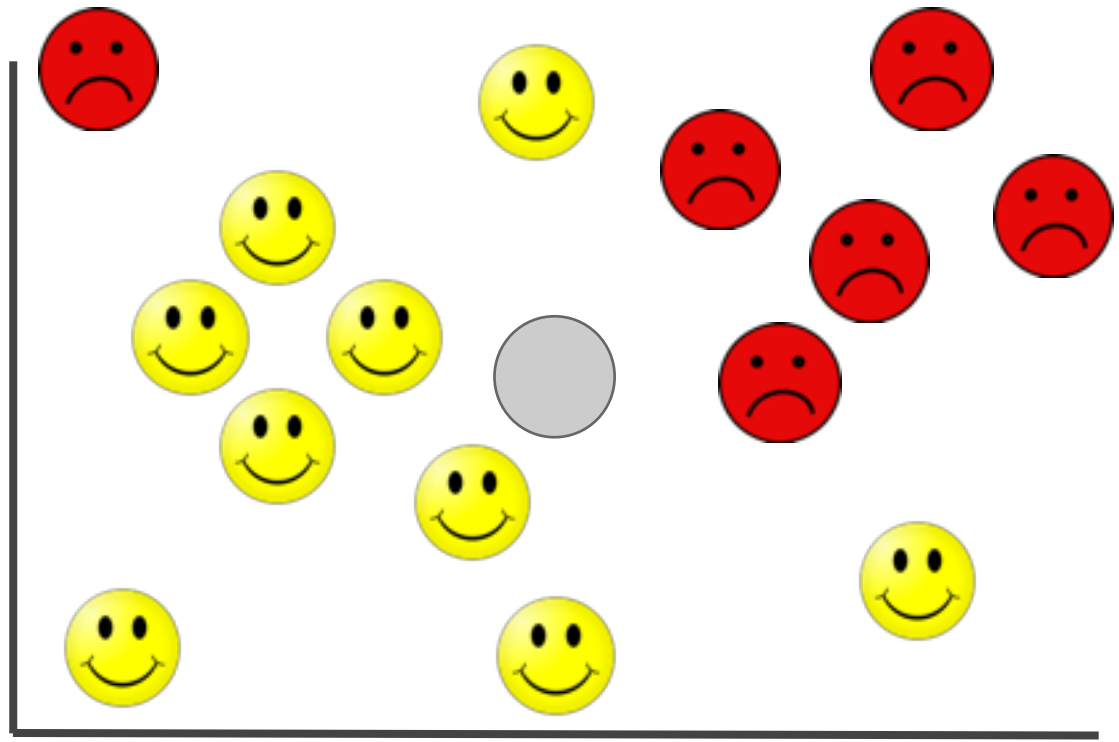
is the grey face happy?



Want to predict:

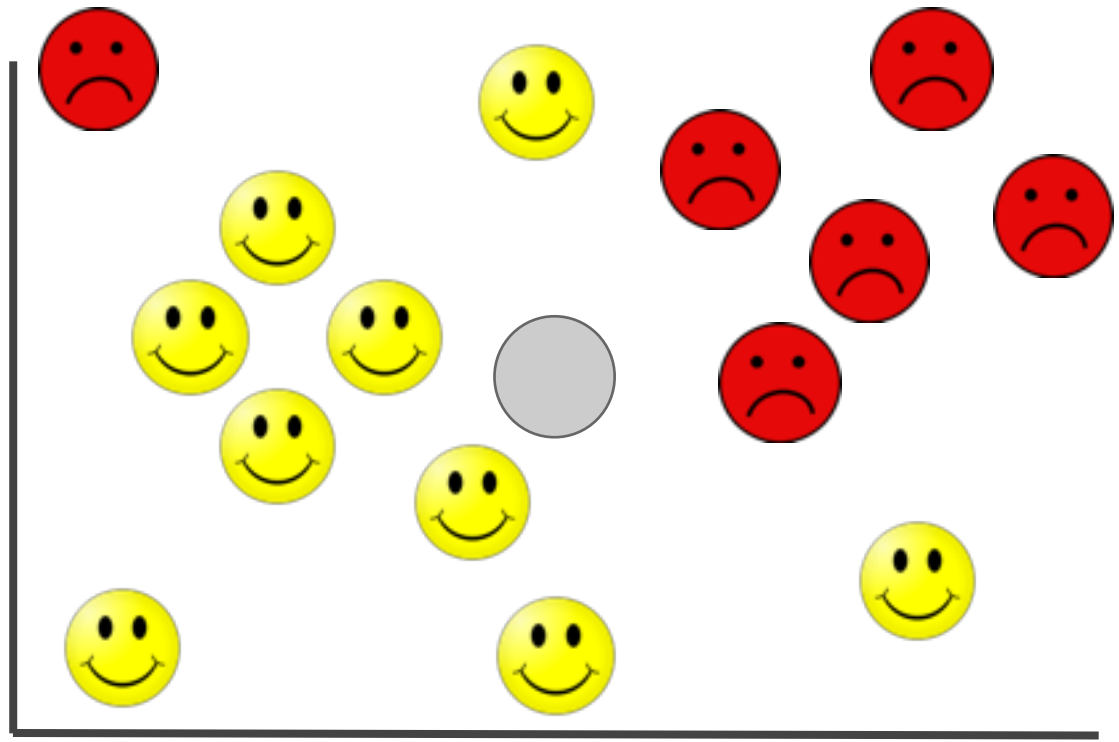
is the grey face happy?

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



Choose k

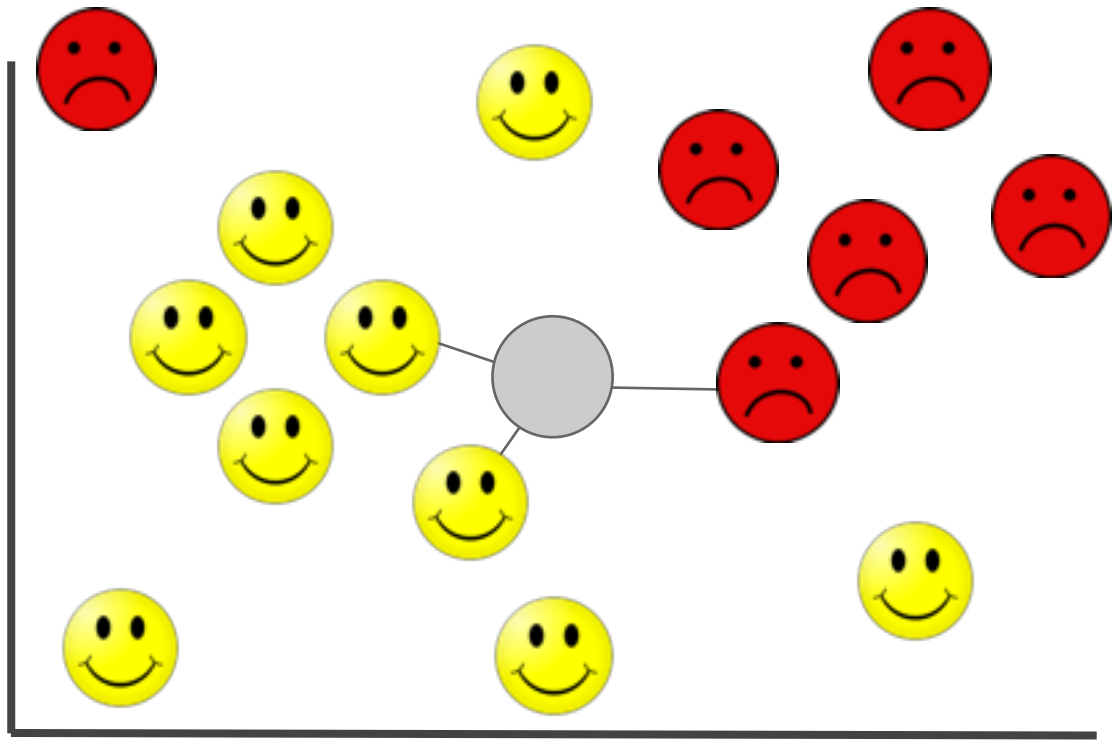
e.g., k = 3



Choose k

e.g., k = 3

Find k nearest neighbors

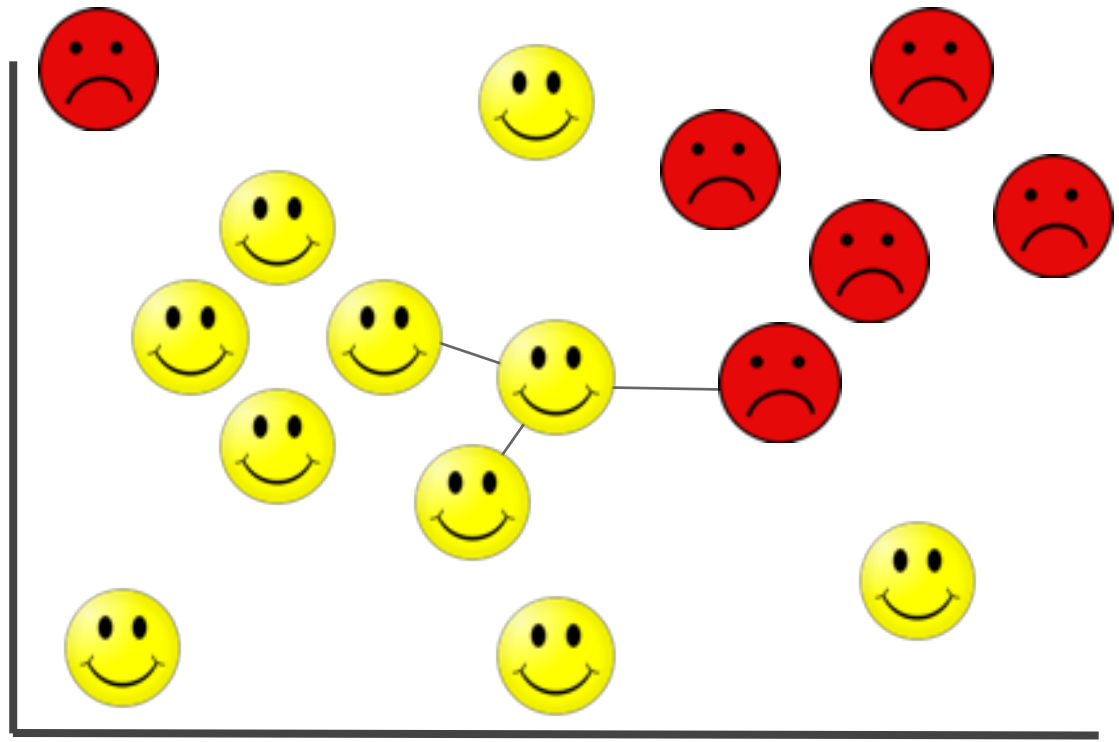


Choose k

e.g., k = 3

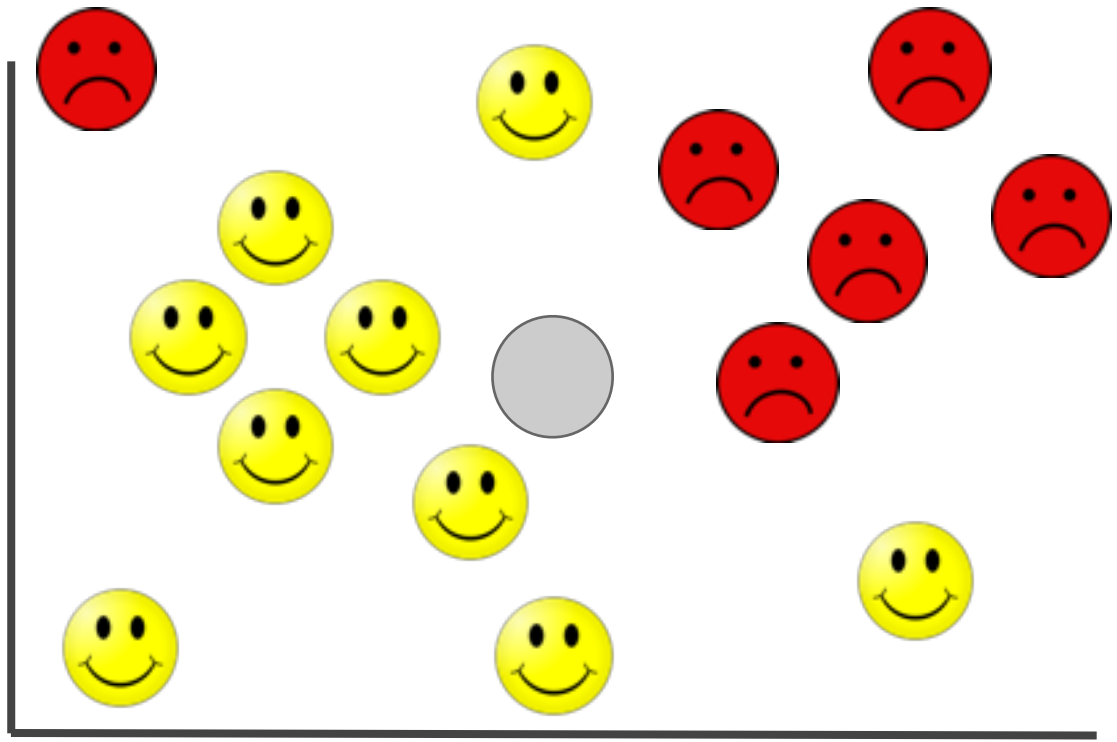
Find k nearest neighbors

Take majority vote



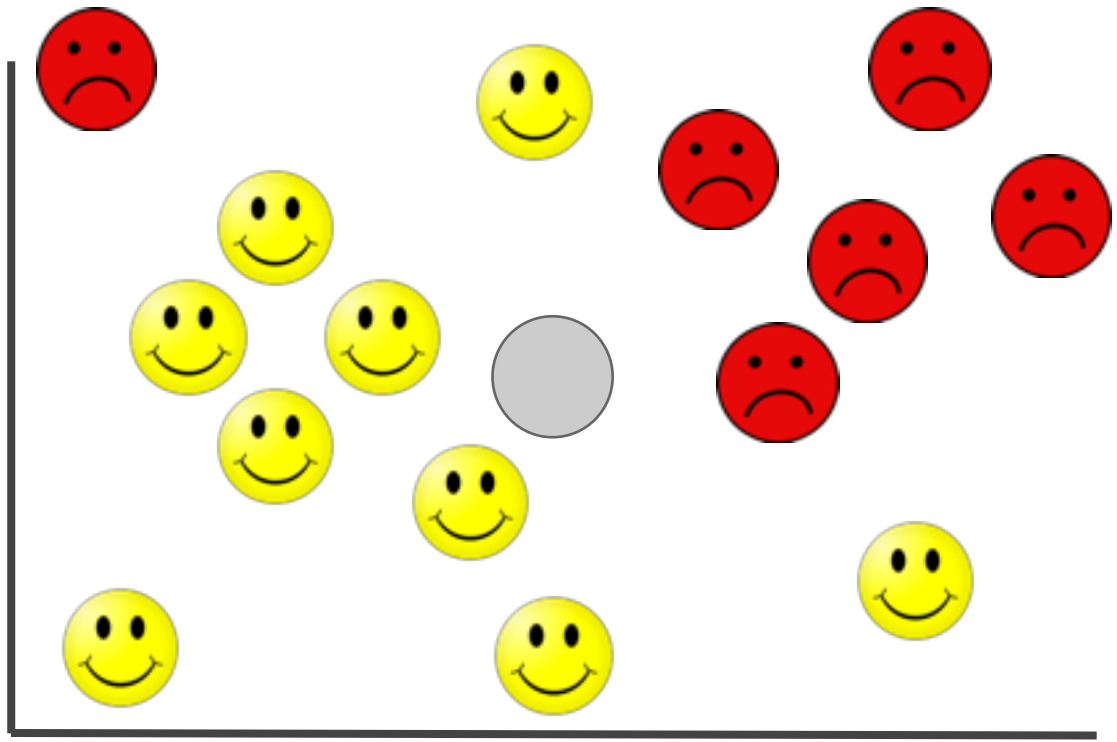
CAVEATS OF KNN

Q: What could possibly go wrong here?



Q: What could possibly go wrong here?

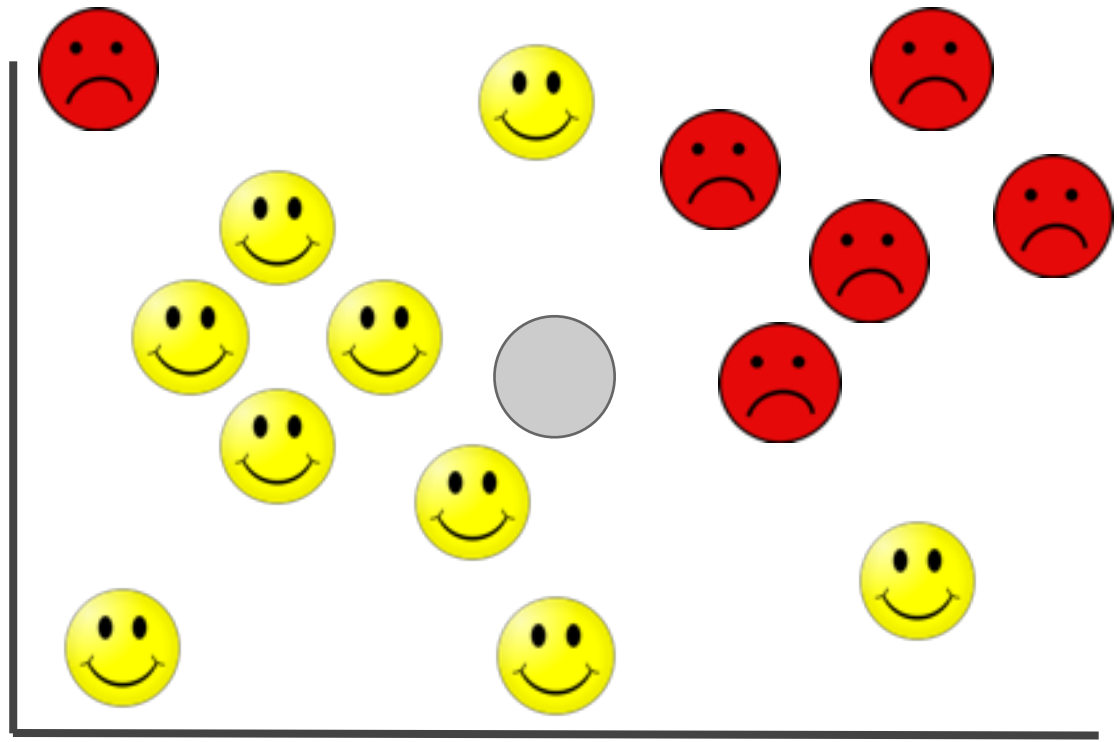
What k?



Q: What could possibly go wrong here?

What k?

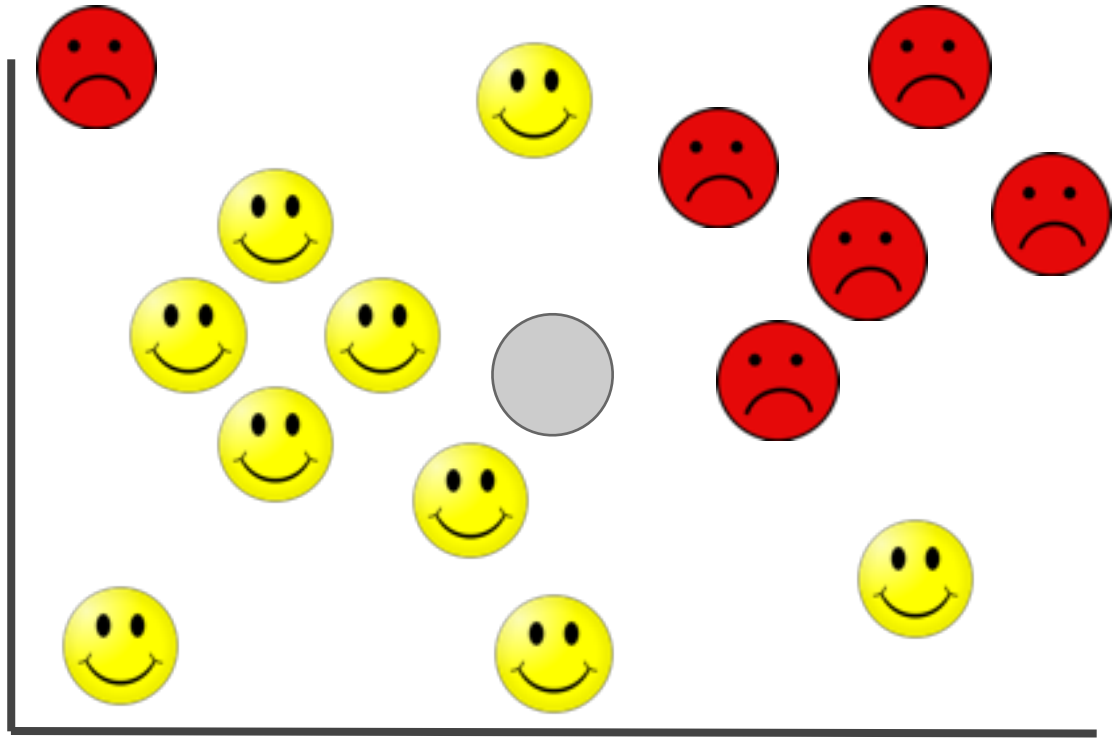
What if k = 1000?



Q: What could possibly go wrong here?

What k?
What if k = 1000?

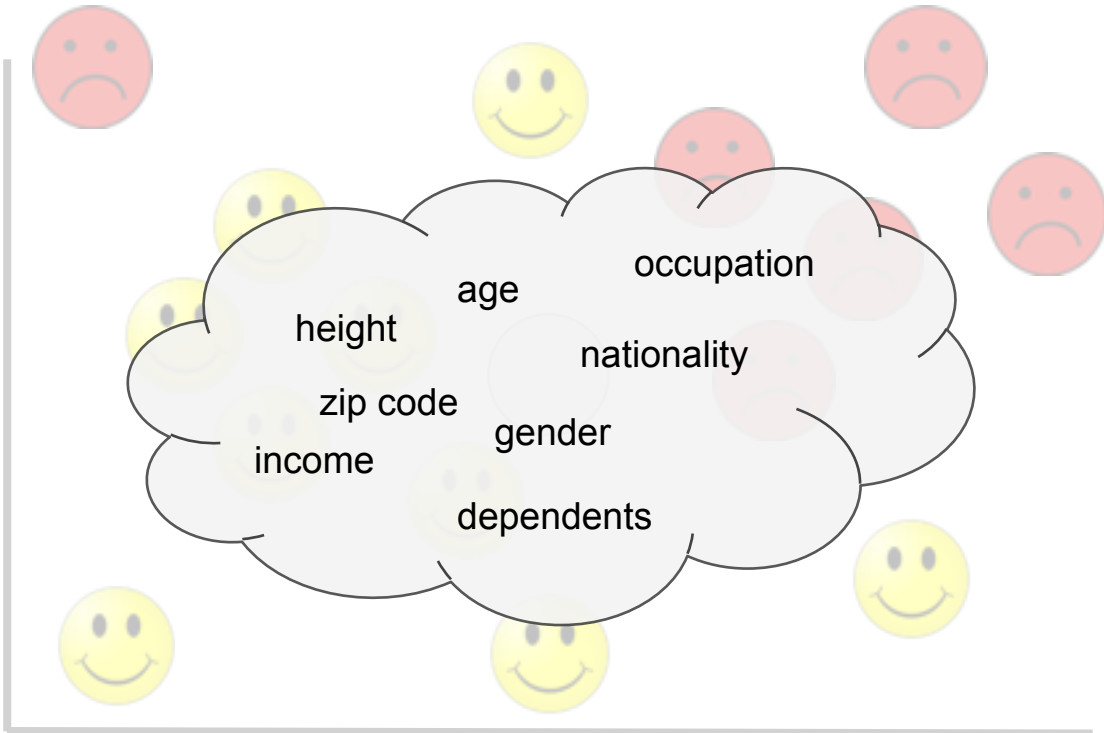
What is 'nearest'?



Q: What could possibly go wrong here?

What k?
What if k = 1000?

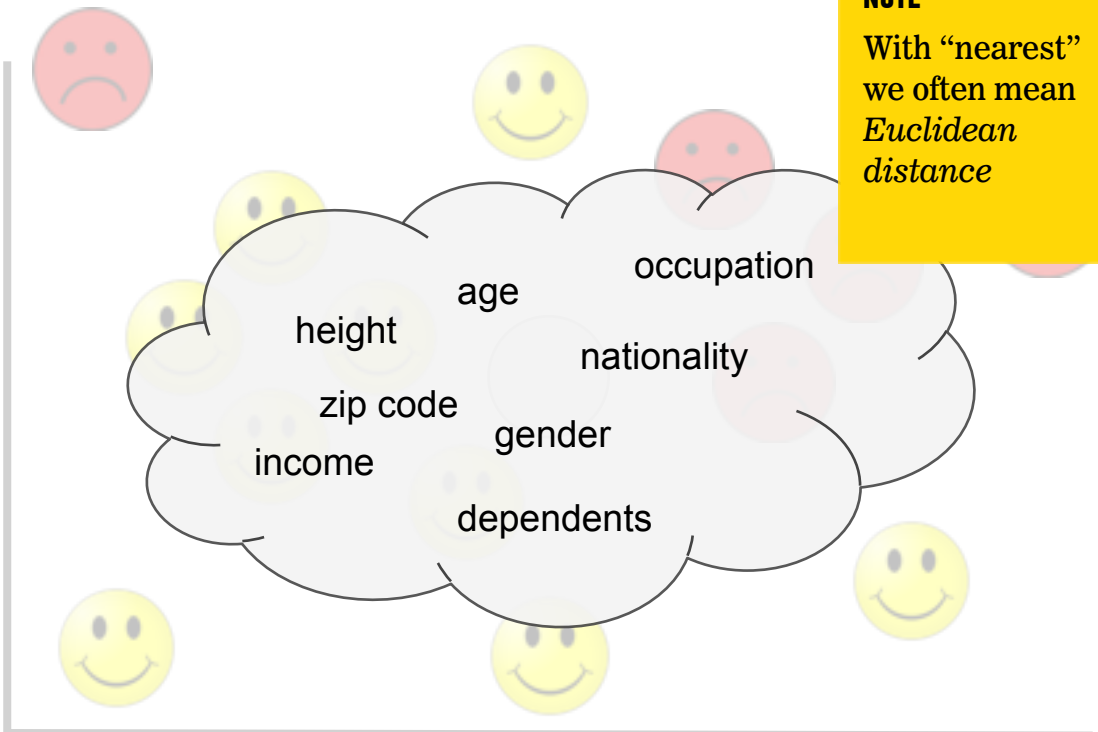
What is 'nearest'?



Q: What could possibly go wrong here?

What k?
What if k = 1000?

What is 'nearest'?



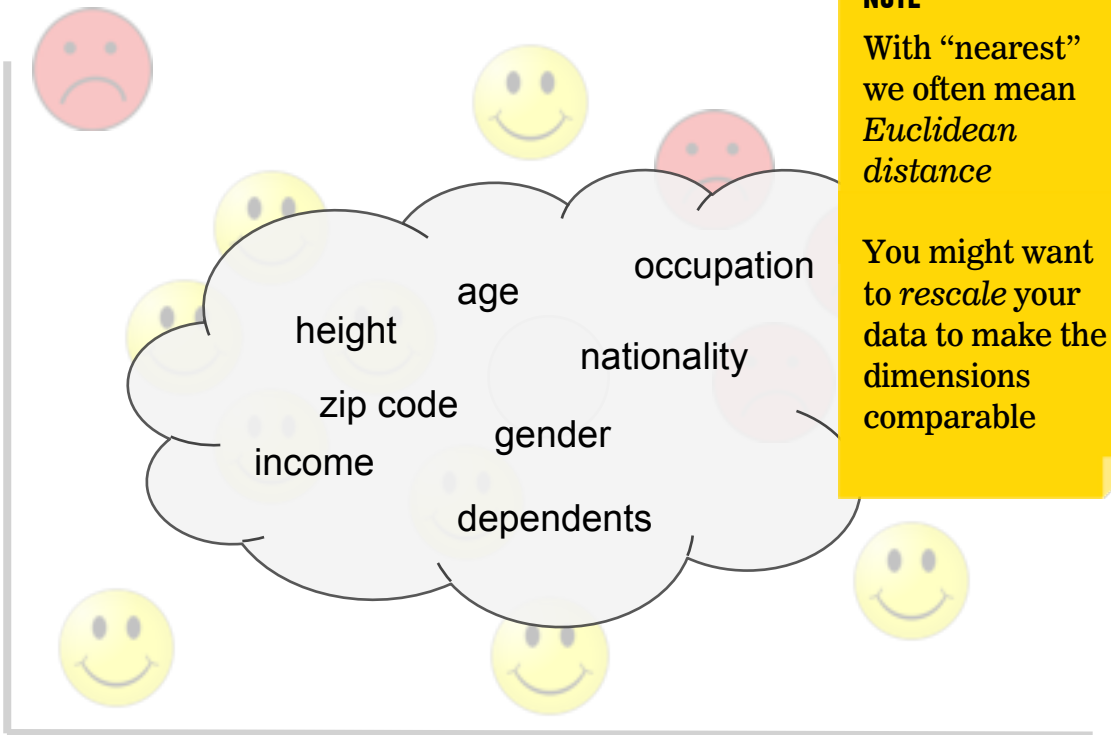
NOTE

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

Exit Tickets! DAT-1, Lesson 3, KNN

Homework 2 Due Dec 16 before class

I will be available after class to help with homeworks

Project Milestone 1 Due Jan 21 before class