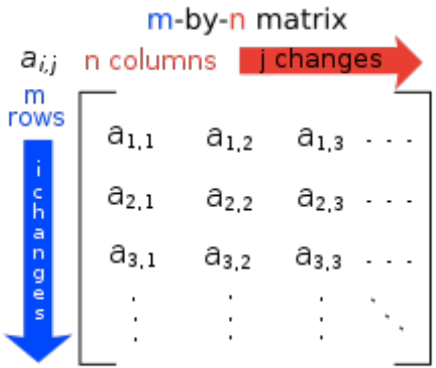
GENERAL ASSEMBLY

DATA SCIENCE

LECTURE 4: INTRODUCTION TO MACHINE LEARNING, CLASSIFICATION WITH K-NEAREST NEIGHBORS

FRANCESCO MOSCONI / ROB HALL / DAT-16

LAST TIME



I. LINEAR ALGEBRA REVIEW
II. DATA CLEANING
III. DATA VISUALIZATION

EXERCISES:
IV. NUMPY
V. PANDAS
VI. BOKEH

QUESTIONS?

QUESTIONS?

WHAT WAS THE MOST INTERESTING THING YOU LEARNT?

WHAT WAS THE HARDEST TO GRASP?

I. LINEAR ALGEBRA REVIEW
II. DATA CLEANING
III. DATA VISUALIZATION

EXERCISES:
IV. NUMPY
V. PANDAS
VI. BOKEH

QUESTIONS?

AGENDA

I. WHAT IS MACHINE LEARNING?
II. MACHINE LEARNING SOLUTIONS
III. CLASSIFICATION
IV. BUILDING EFFECTIVE CLASSIFIERS
V. K-NEAREST NEIGHBORS

EXERCISES:
VI. LAB: KNN CLASSIFICATION IN PYTHON

I. WHAT IS MACHINE LEARNING?

"A field of study that gives computers the ability to learn without being explicitly programmed." (1959)

Arthur Samuel, AI pioneer
Source: Stanford

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?

- **Machine learning** *is an area in computer science that studies and develops algorithms that can learn from data.*

- **Machine learning** *is a set of methods that can automatically detect patterns in data and use the discovered patterns to predict future data or perform other kinds of decision making*

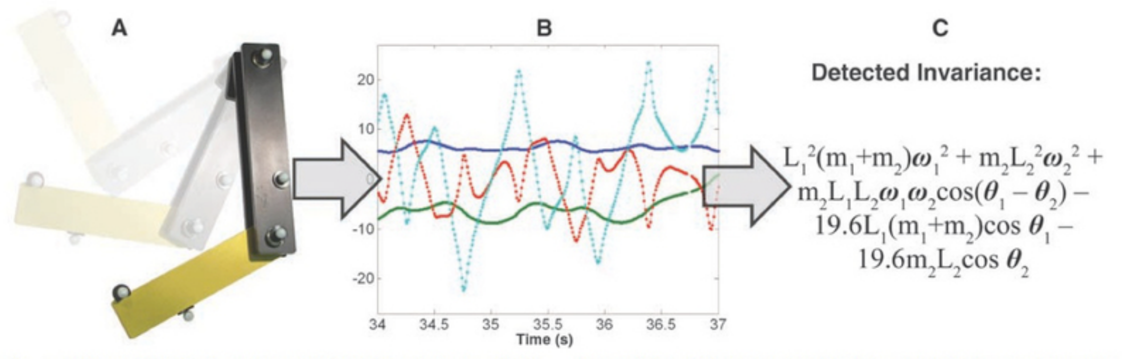
- *Statistical learning theory, Pattern recognition*

WHEN DO WE NEED MACHINE LEARNING?

Where we need it:

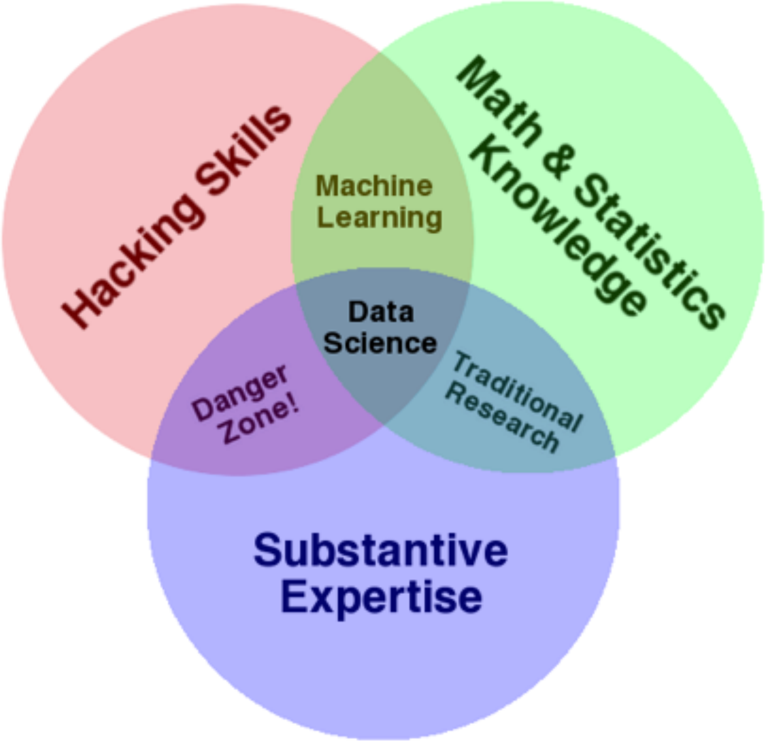
- *Some observable patterns exist*
- *There no explicitly known equations or dependencies (formulas)*
- *We have data on it*

Example: Newton's second law of motion, conservation of mechanical energy, pendulum motion



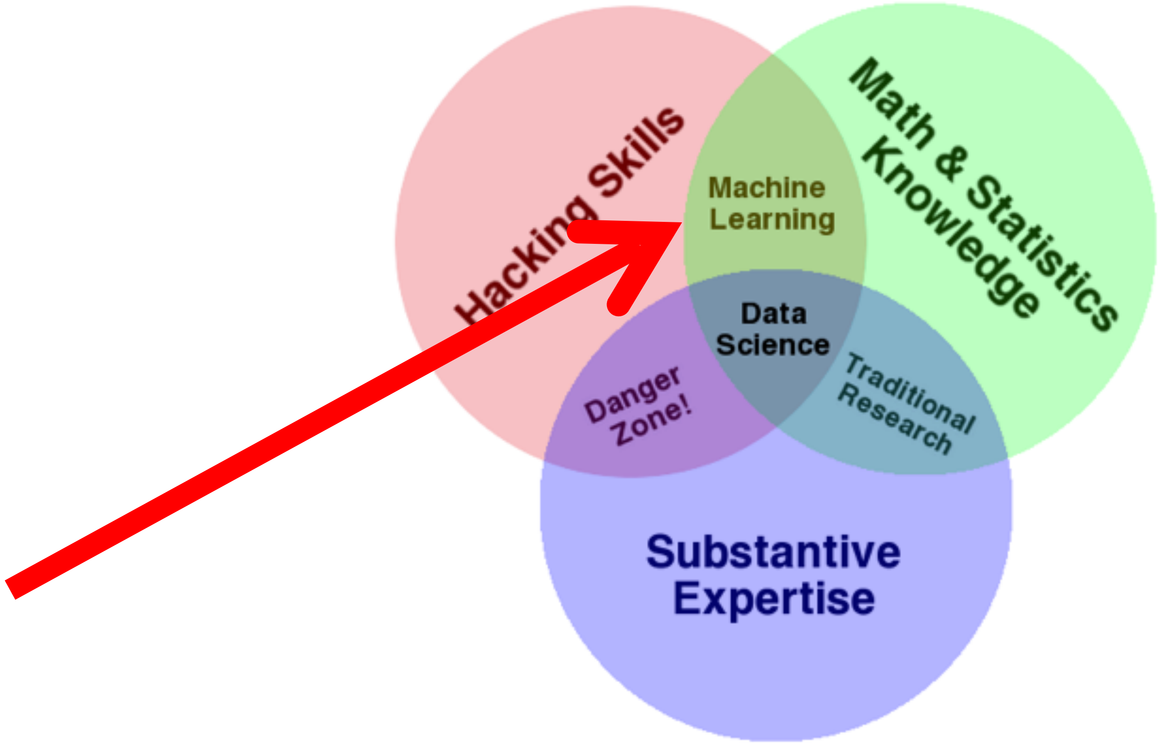
From "Distilling Free-Form Natural Laws from Experimental Data." M. Schmidt and H.Lipson. Science, 2009.

REMEMBER THIS?



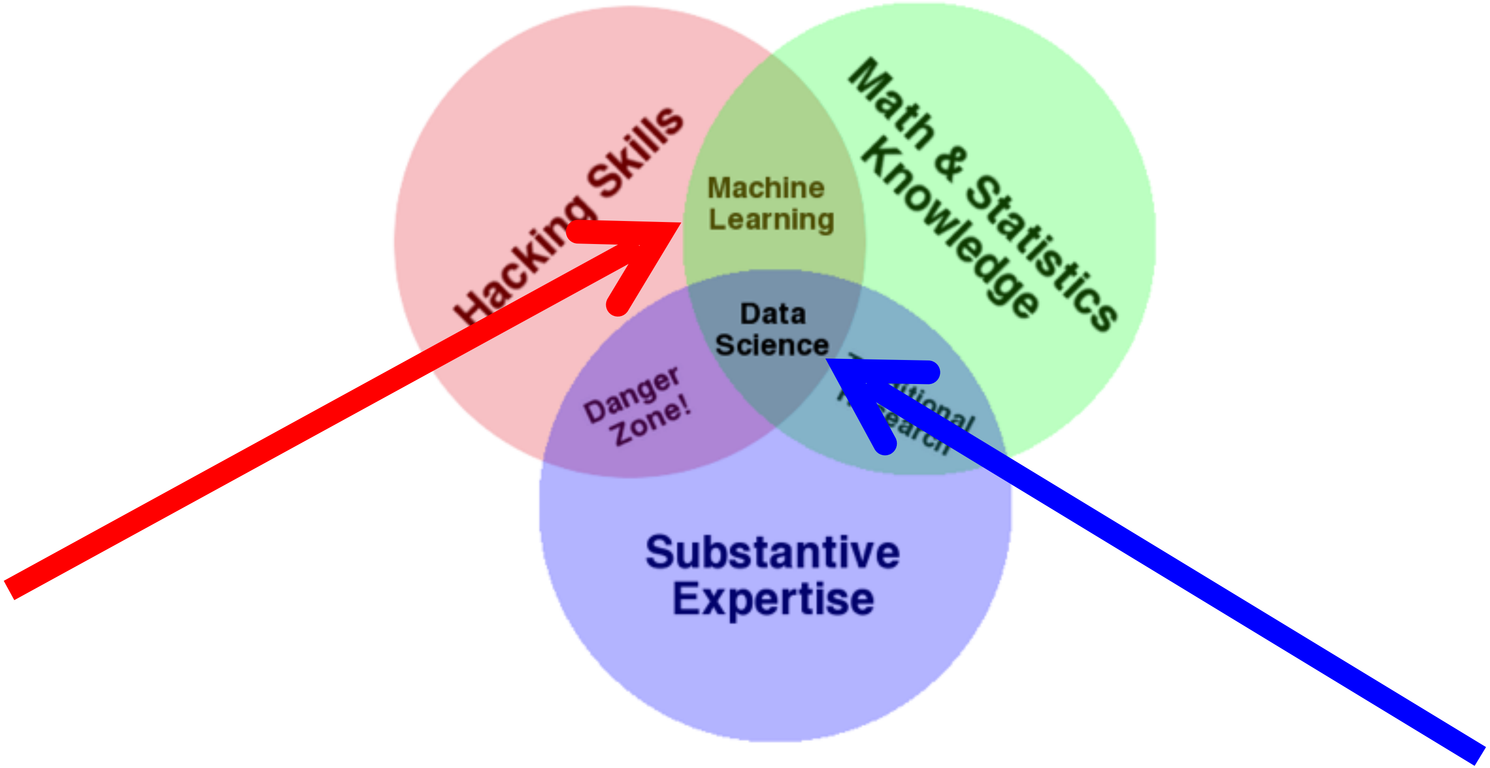
source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

WE ARE NOW HERE



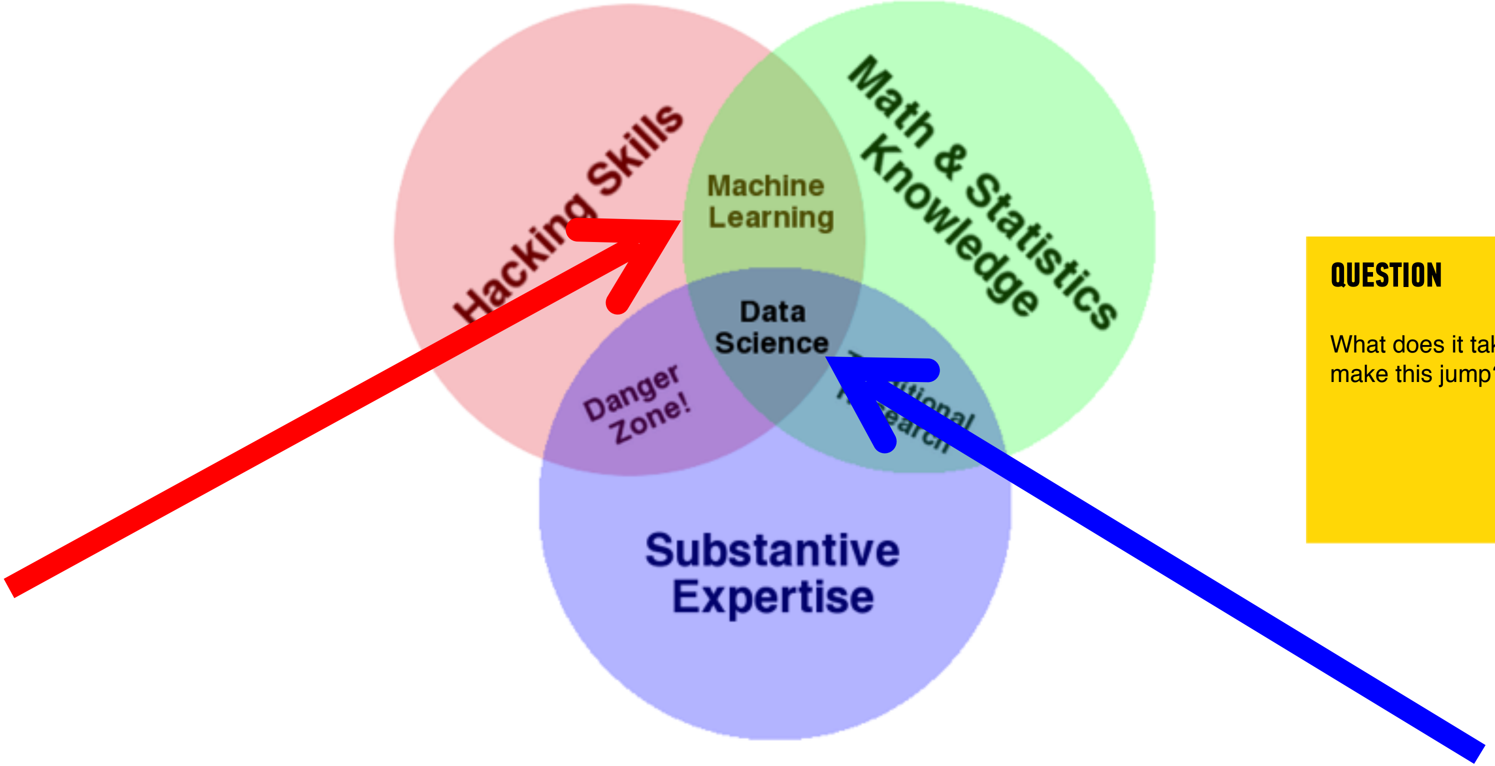
source: *http://www.dataists.com/2010/09/the-data-science-venn-diagram/*

WE WANT TO GO HERE



source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

WE WANT TO GO HERE

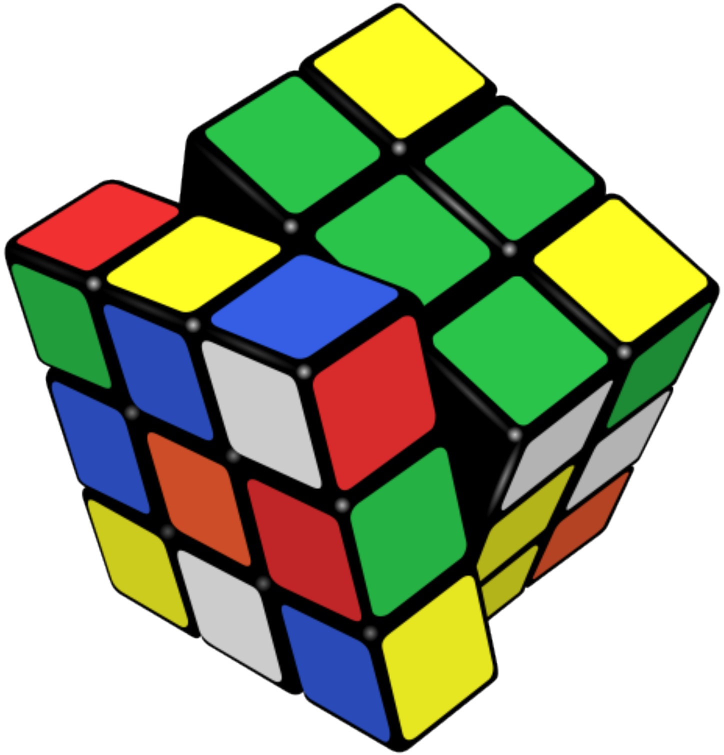


QUESTION

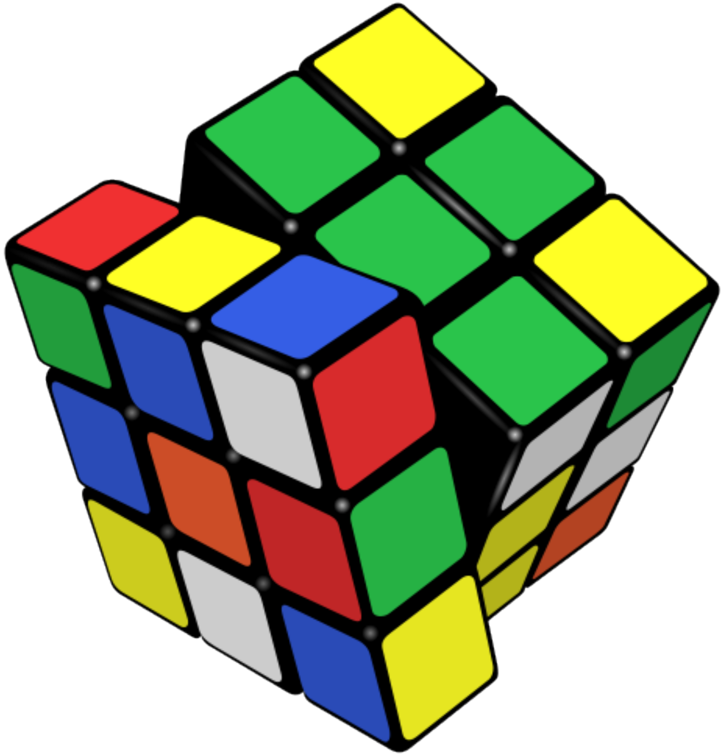
What does it take to make this jump?

source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

ANSWER: PROBLEM SOLVING!



ANSWER: PROBLEM SOLVING!



NOTE

Implementing solutions to ML problems is the focus of this course!

II. MACHINE LEARNING SOLUTIONS

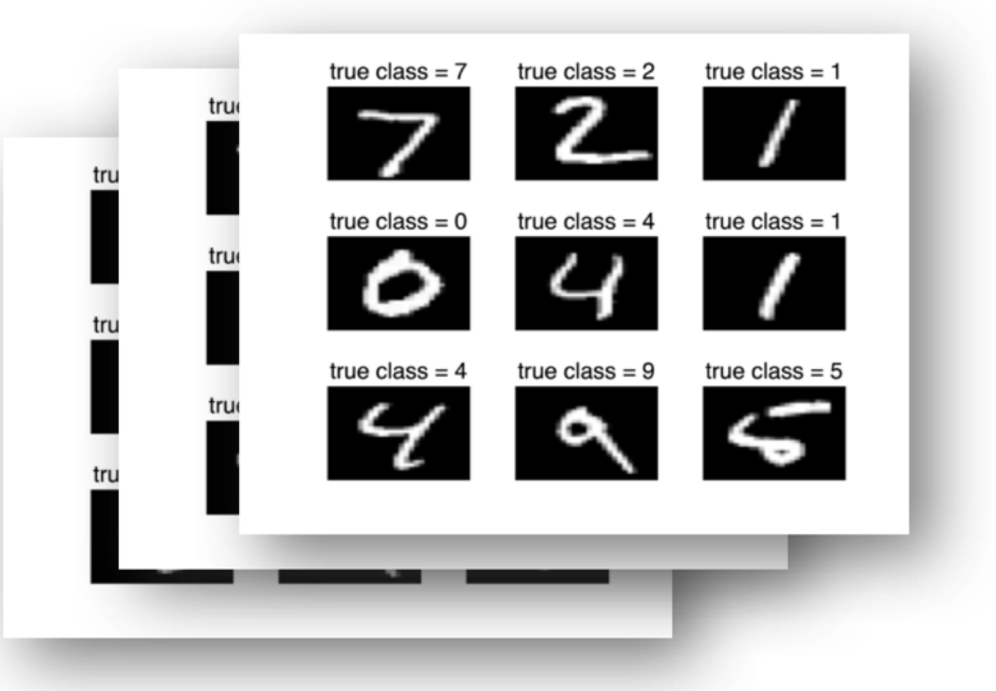
Learning is not about memorizing and being able to recall, it is about generalizing the conclusions to previously unseen examples

ML solutions can be described by the **type of question** for example:

Supervised learning: the goal is to learn mapping from given inputs x to outputs y, given a **labeled** set of input-output pairs

OCR


OCR


true class = 7
true class = 2
true class = 1
true class = 0
true class = 4
true class = 1
true class = 4
true class = 9
true class = 5

CREDIT SCORING

	Client 1	Client 2	Client 3
Age	23	30	19
Gender	M	F	M
Annual salary	$30,000	$45,000	$15,000
Years in residence	3 years	1 year	3 month
Years in job	1 year	1 year	1 month
Current debt	$5,000	$1,000	$10,000
Paid off credit	Yes	Yes	No

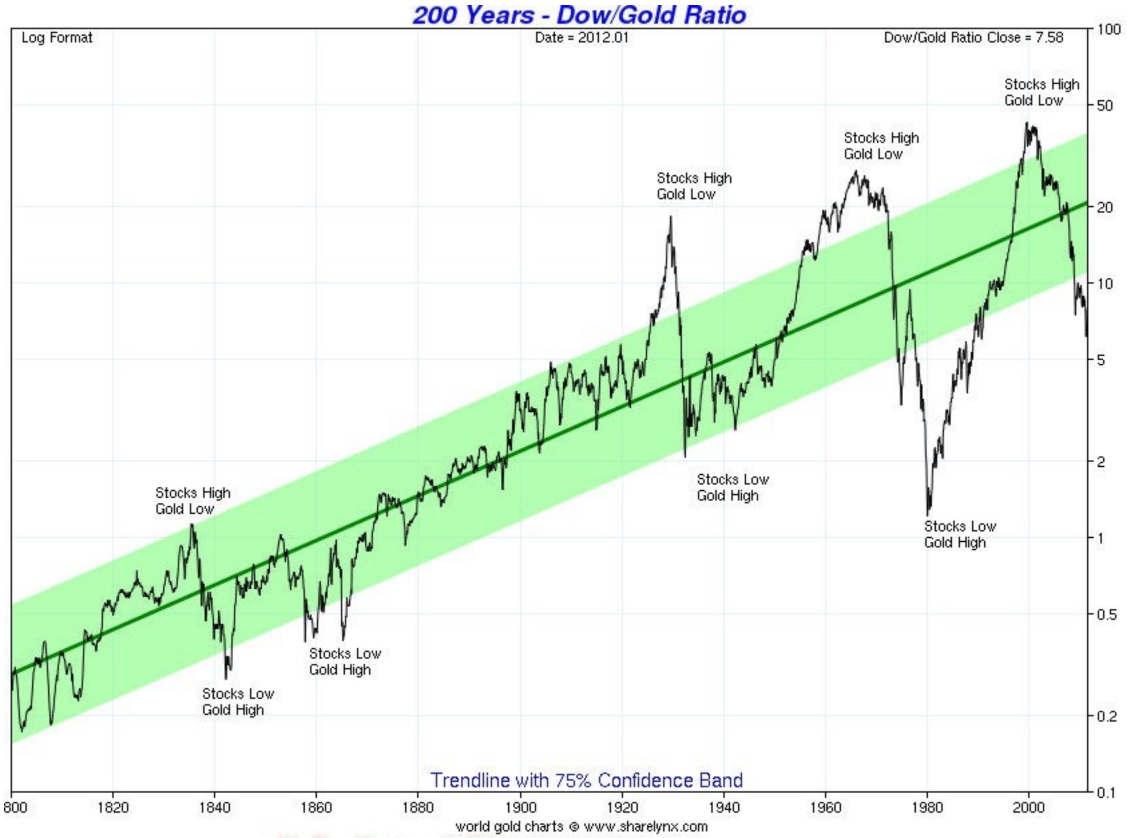
CREDIT SCORING

	Client 1	Client 2	Client 3
Age	23	30	19
Gender	M	F	M
Annual salary	$30,000	$45,000	$15,000
Years in residence	3 years	1 year	3 month
Years in job	1 year	1 year	1 month
Current debt	$5,000	$1,000	$10,000
Paid off credit	Yes	Yes	No

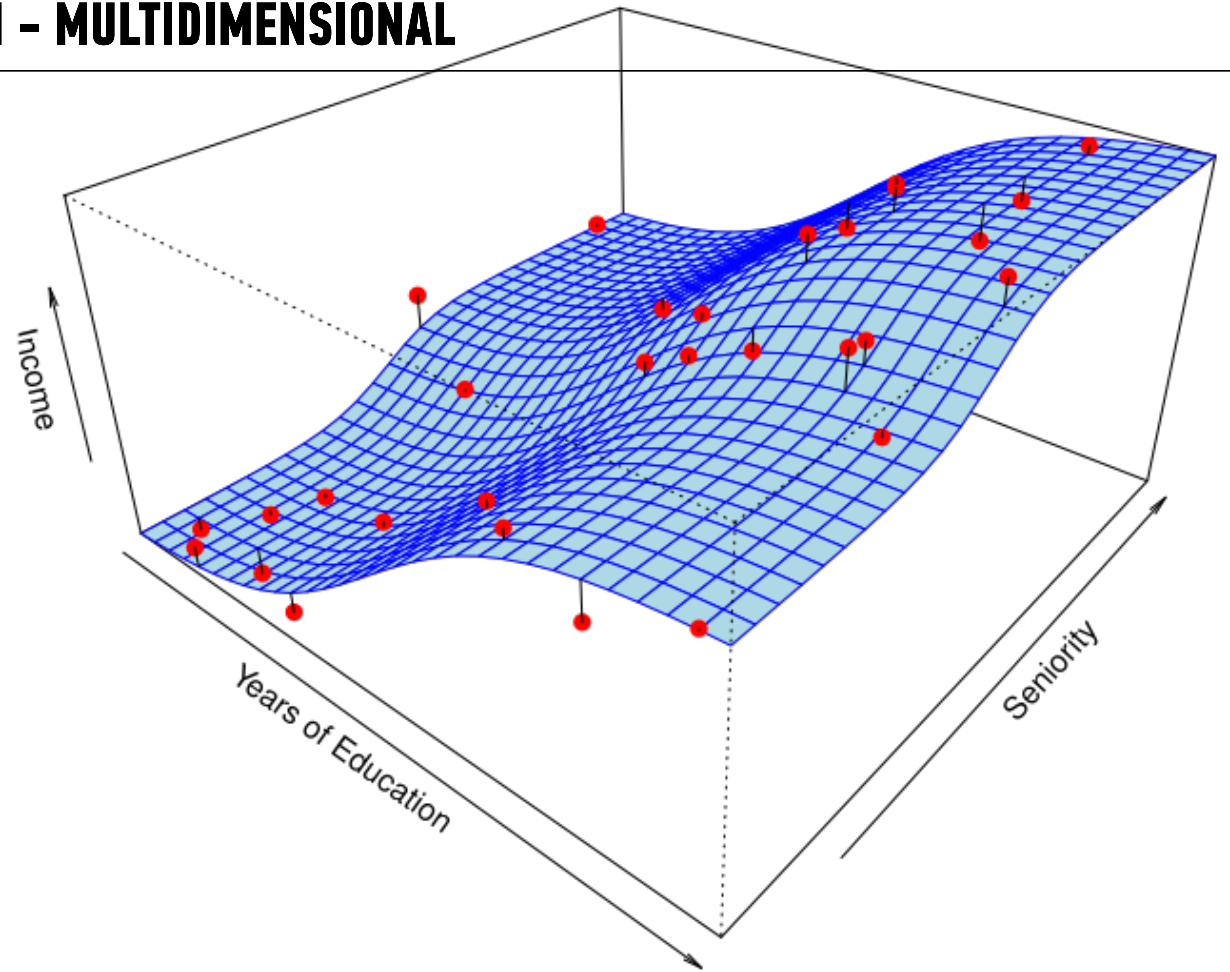
	Applicant
Age	25
Gender	M
Annual salary	$25,000
Years in residence	1 year
Years in job	2 year3
Current debt	$15,000
Credit decision/ score	???

EMAIL SPAM DETECTION

REGRESSION - STOCK PRICE PREDICTION



REGRESSION - MULTIDIMENSIONAL

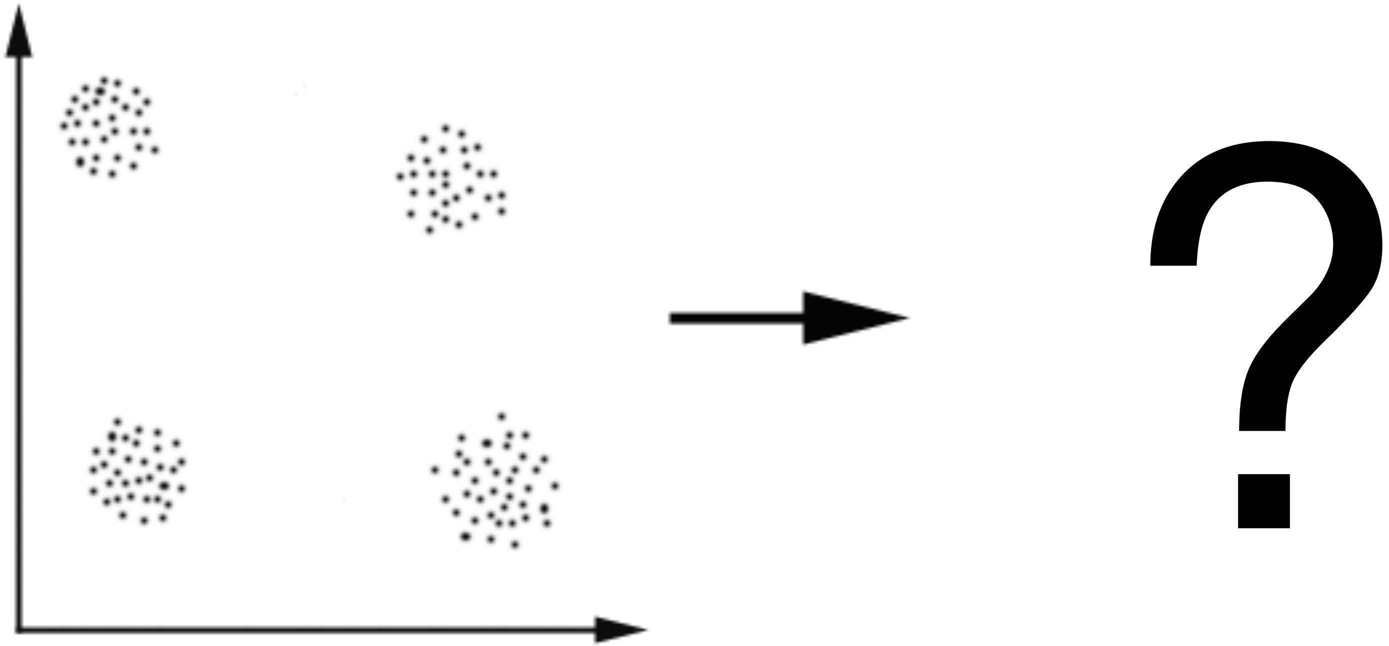


for example:

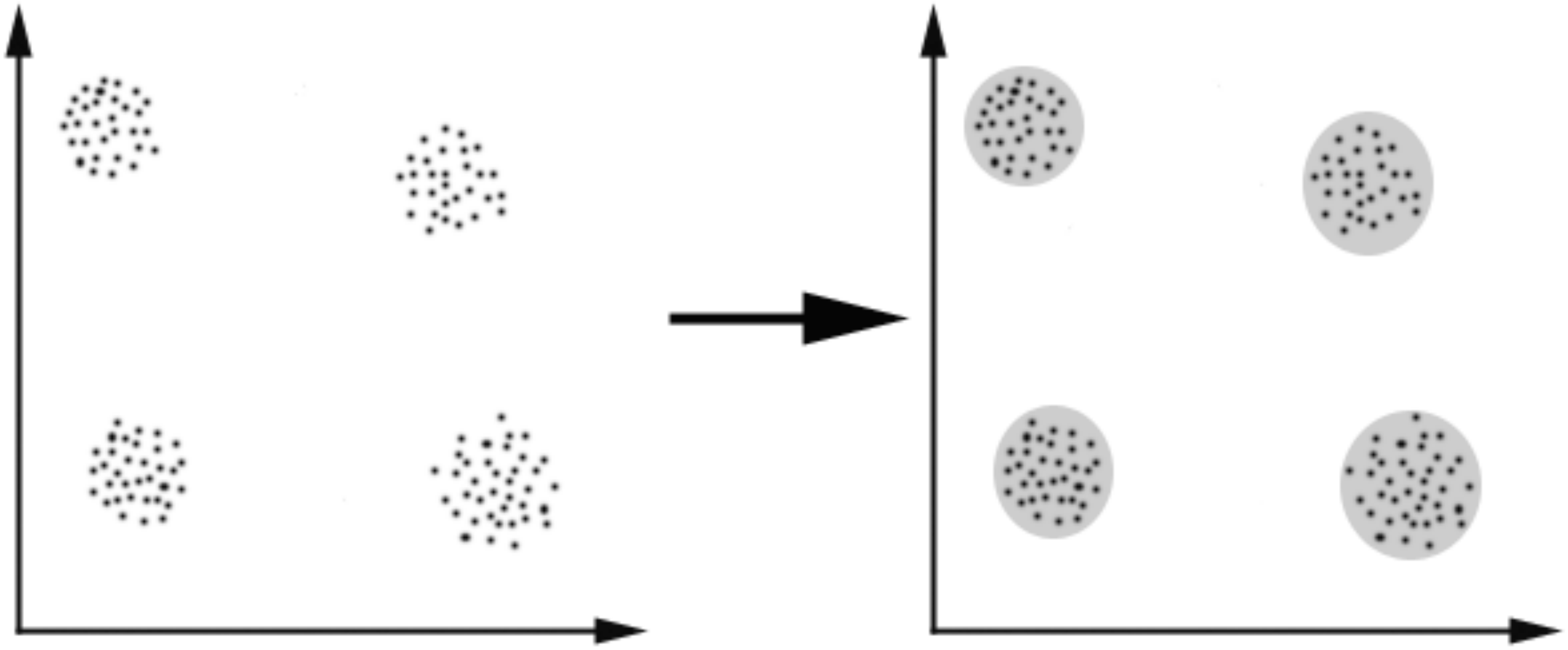
Unsupervised learning: the goal is to learn interesting patterns and structure in data given only inputs

no label information given at all

can you find structure in data given only inputs?



can you find structure in data given only inputs?



REMEMBER WHAT WE SAID BEFORE?

generalization

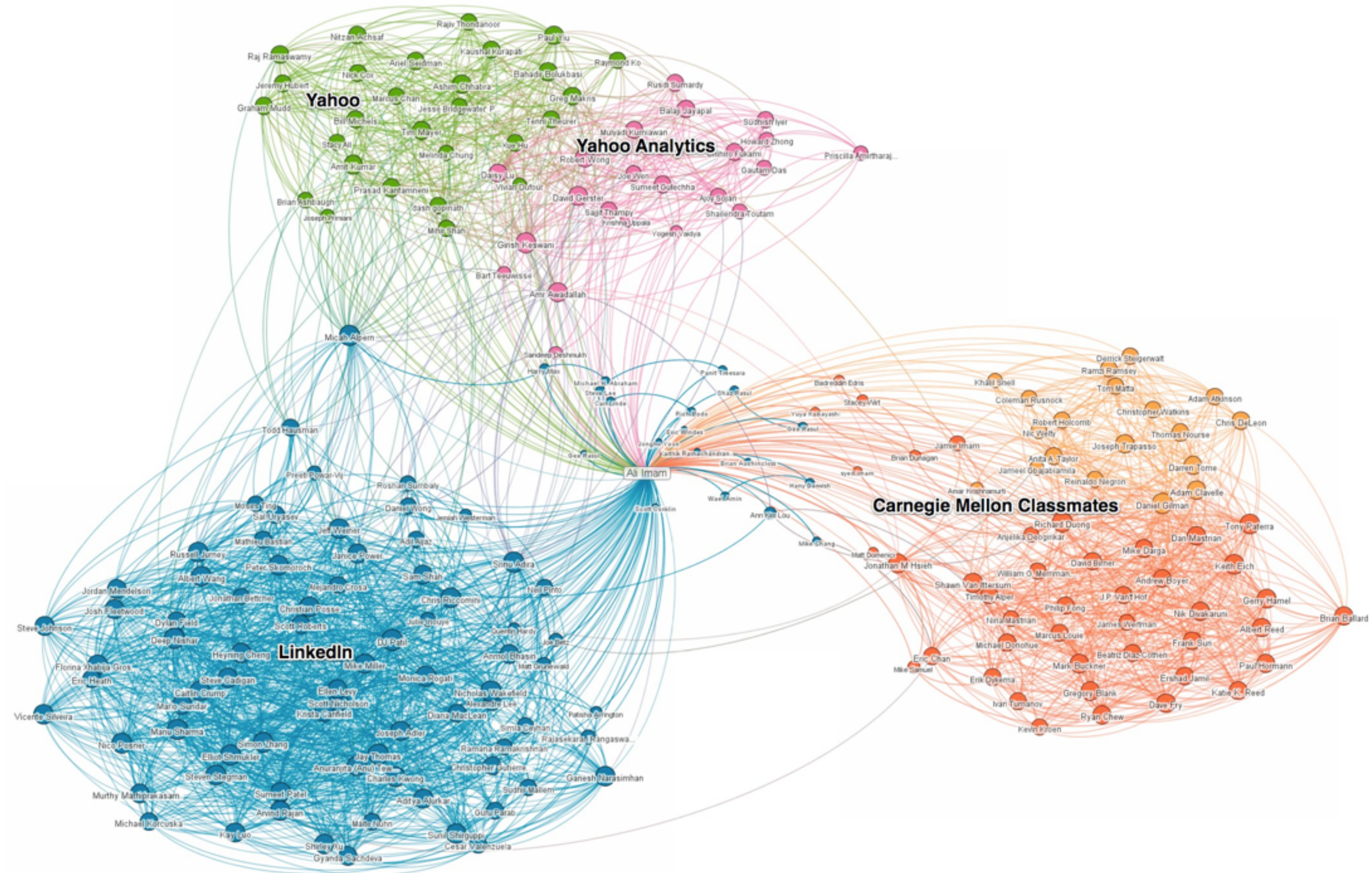
Supervised	*Making predictions*
Unsupervised	*Extracting structure*

representation

EXERCISE:

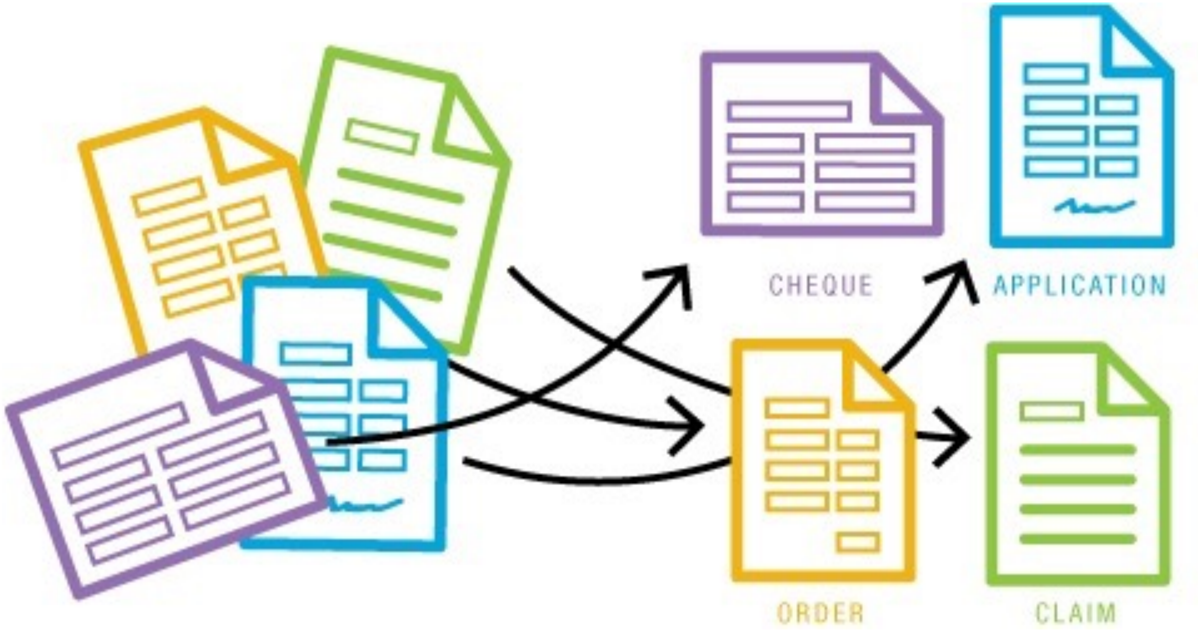
supervised or unsupervised?

COMMUNITY DETECTION IN SOCIAL NETWORKS



REGRESSION – HOUSE PRICE PREDICTION

DOCUMENT CLASSIFICATION



ML solutions can be described by the **type of data**

	Continuous	**Categorical**
	Quantitative	*Qualitative*

NOTE

The space where data live is called the *feature space.*

Each point in this space is called a *record*.

for example:

	Continuous	*Categorical*
	Height of children	*Eye colors*
	Weight of cars	*Courses at GA*
	Speed of the train	*Highest degree*
	Temperature	*Gender*
	Stock price	*Is email spam or not*

NOTE

The space where data live is called the *feature space.*

Each point in this space is called a *record*.

combined…

	Continuous	Categorical
Supervised	*regression*	*classification*
Unsupervised	*dimension reduction*	*clustering*

NOTE

We will implement solutions using *models* and *algorithms.*

Each will fall into one of these four buckets.

QUESTION

WHAT IS THE GOAL OF MACHINE LEARNING?

Supervised	*Making predictions*
Unsupervised	*Extracting structure*

ANSWER

The goal is determined by the type of problem.

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.

	Continuous	Categorical
Supervised	*regression*	*classification*
Unsupervised	*dimension reduction*	*clustering*

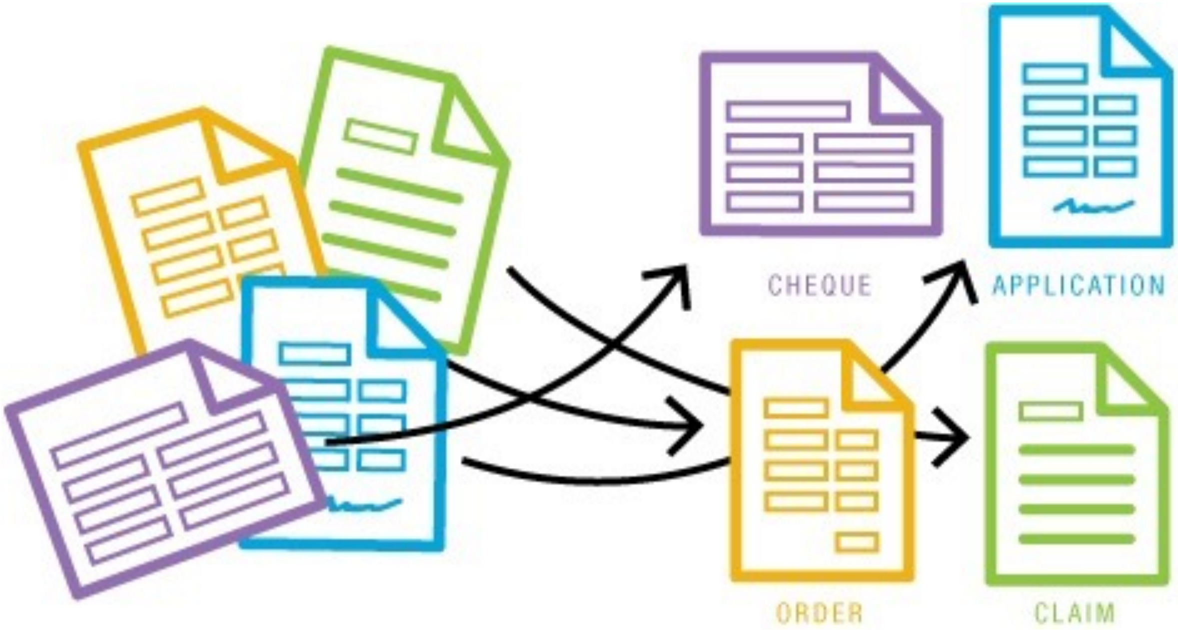
ANSWER

The
det
des

NOTE

All of this depends on your data!

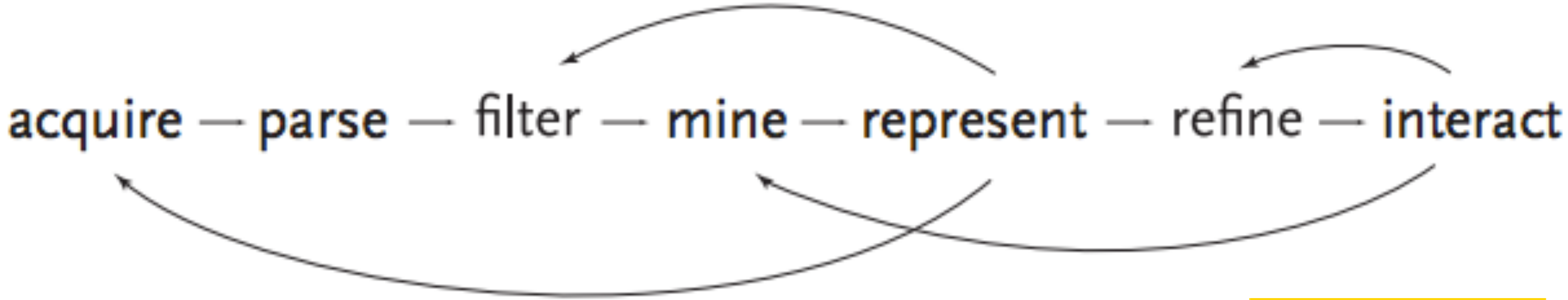
DO WE HAVE LABELS?



QUESTION

WHAT DO YOU DO WITH YOUR RESULTS?

THE DATA SCIENCE WORKFLOW

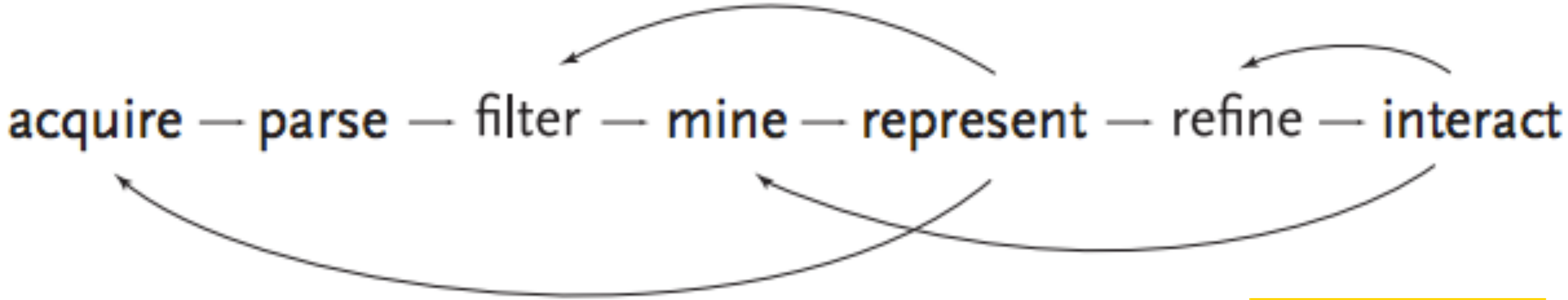


ANSWER

Interpret them and react accordingly.

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

THE DATA SCIENCE WORKFLOW



ANSWER

Int
rea

NOTE

This also relies on your problem solving skills!

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

III. CLASSIFICATION

	Continuous	***Categorical***
Supervised	*???*	*???*
Unsupervised	*???*	*???*

	Continuous	***Categorical***
Supervised	*regression*	*classification*
Unsupervised	*dimension reduction*	*clustering*

Here's (part of) an example dataset:

Fisher's *Iris* Data

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
4.6	3.4	1.4	0.3	*I. setosa*
5.0	3.4	1.5	0.2	*I. setosa*

Here's (part of) an example dataset:

Fisher's *Iris* Data

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
4.6	3.4	1.4	0.3	*I. setosa*
5.0	3.4	1.5	0.2	*I. setosa*

independent variables

Here's (part of) an example dataset:

Fisher's *Iris* Data

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
4.6	3.4	1.4	0.3	*I. setosa*
5.0	3.4	1.5	0.2	*I. setosa*

independent variables

class labels (qualitative)

Q: What does "supervised" mean?

Q: What does "supervised" mean?

A: We know the labels.

Fisher's *Iris* Data

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
4.6	3.4	1.4	0.3	*I. setosa*
5.0	3.4	1.5	0.2	*I. setosa*

class labels (qualitative)

Q: How does a classification problem work?

Q: How does a classification problem work?

A: Data in, predicted labels out.

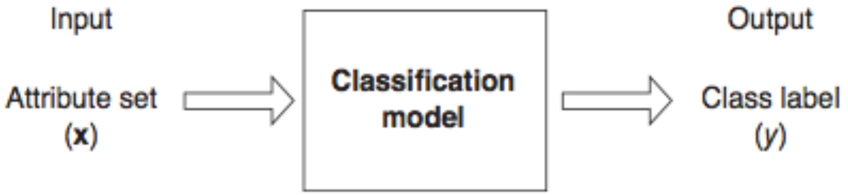
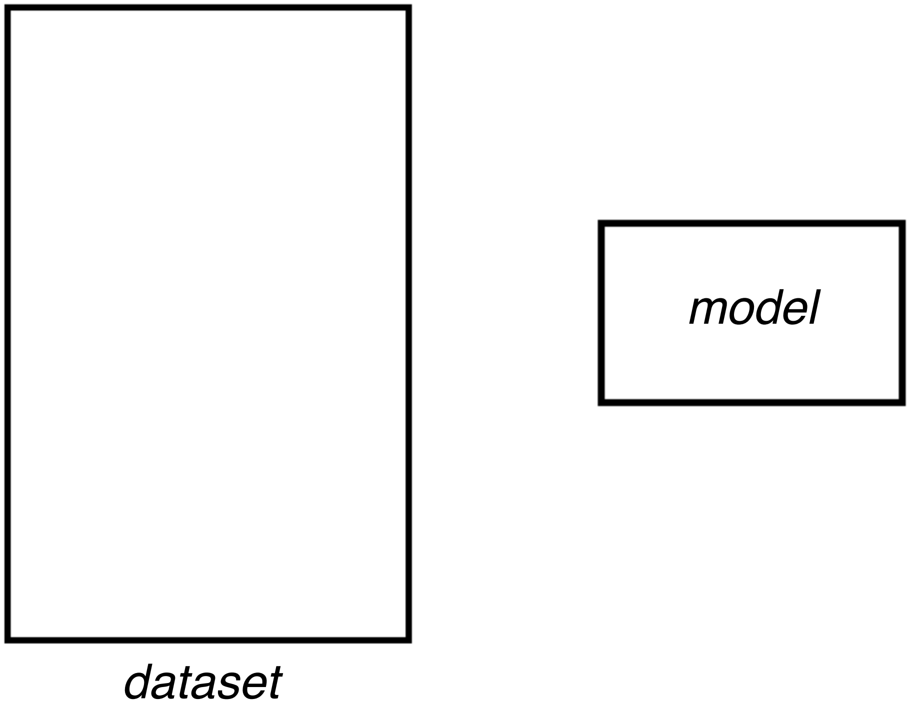


Figure 4.2. Classification as the task of mapping an input attribute set $\mathbf{x}$ into its class label y.

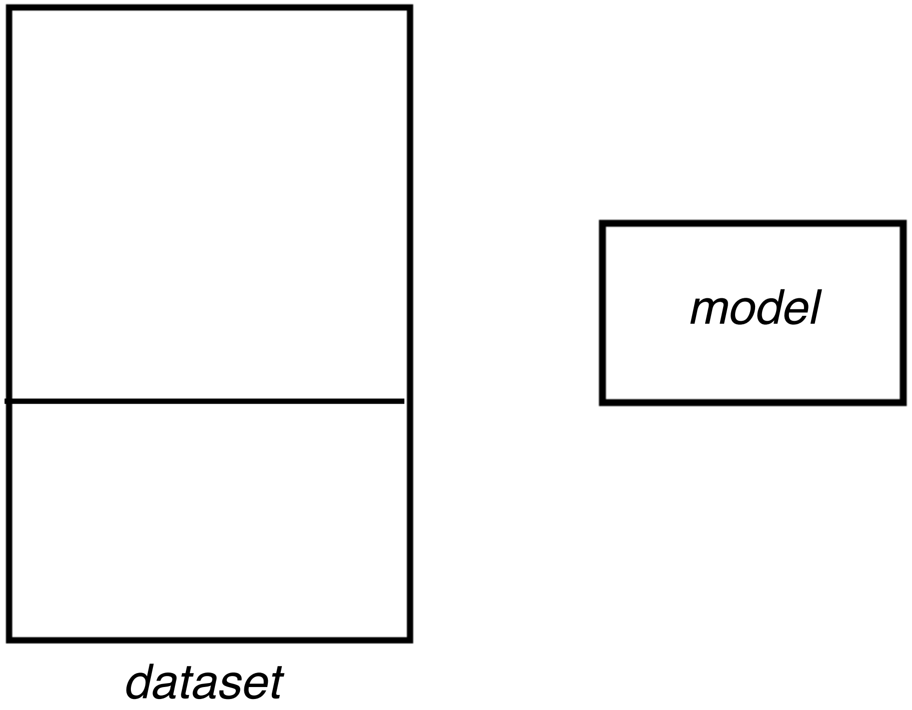
source: http://www-users.cs.umn.edu/~kumar/dmbook/ch4.pdf

Q: What steps does a classification problem require



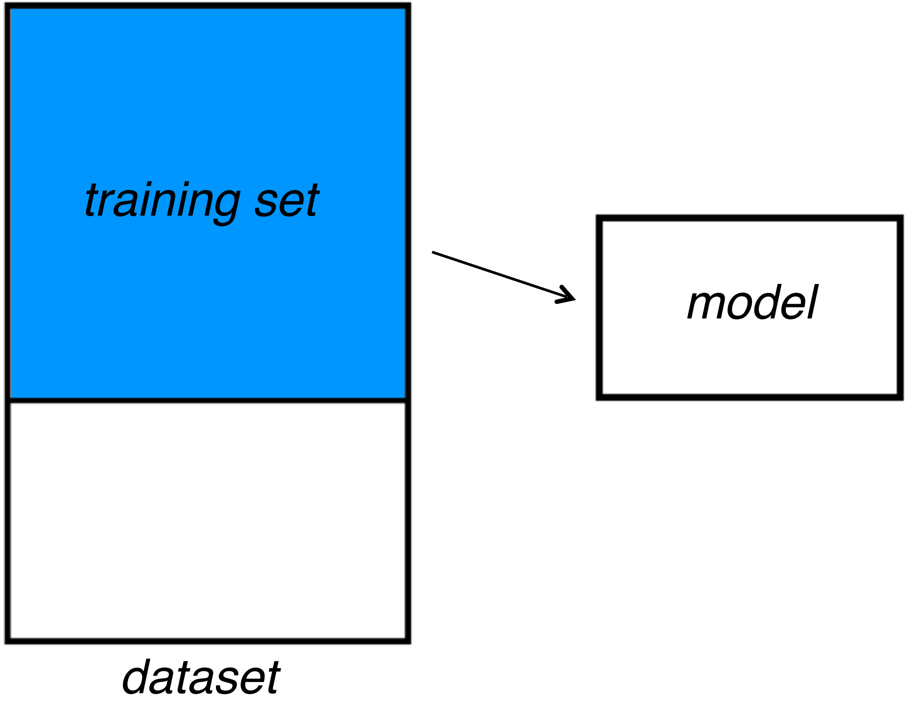
Q: What steps does a classification problem require

1) *split dataset*



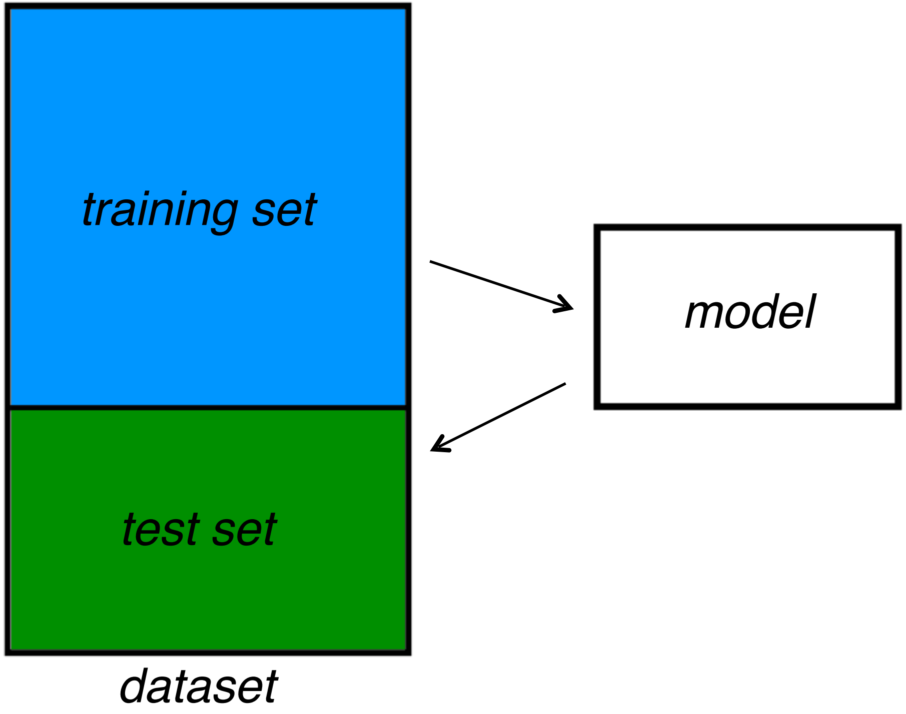
Q: What steps does a classification problem require

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



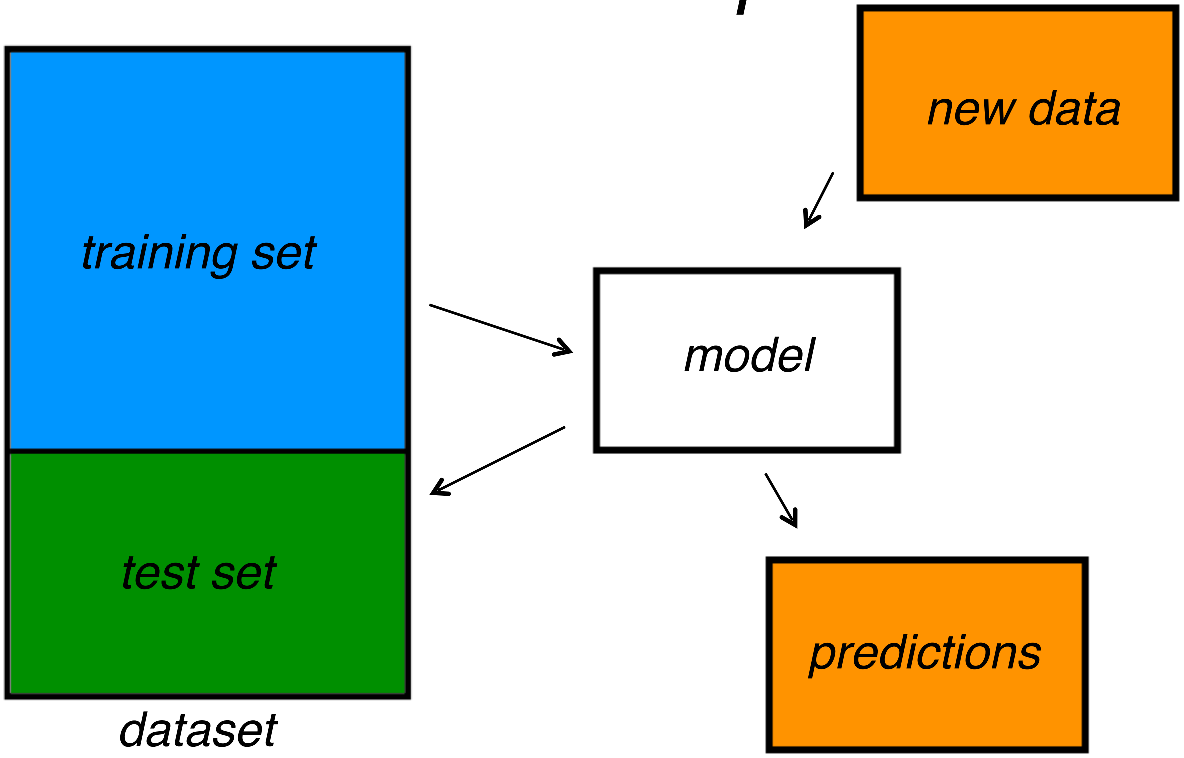
Q: What steps does a classification problem require

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



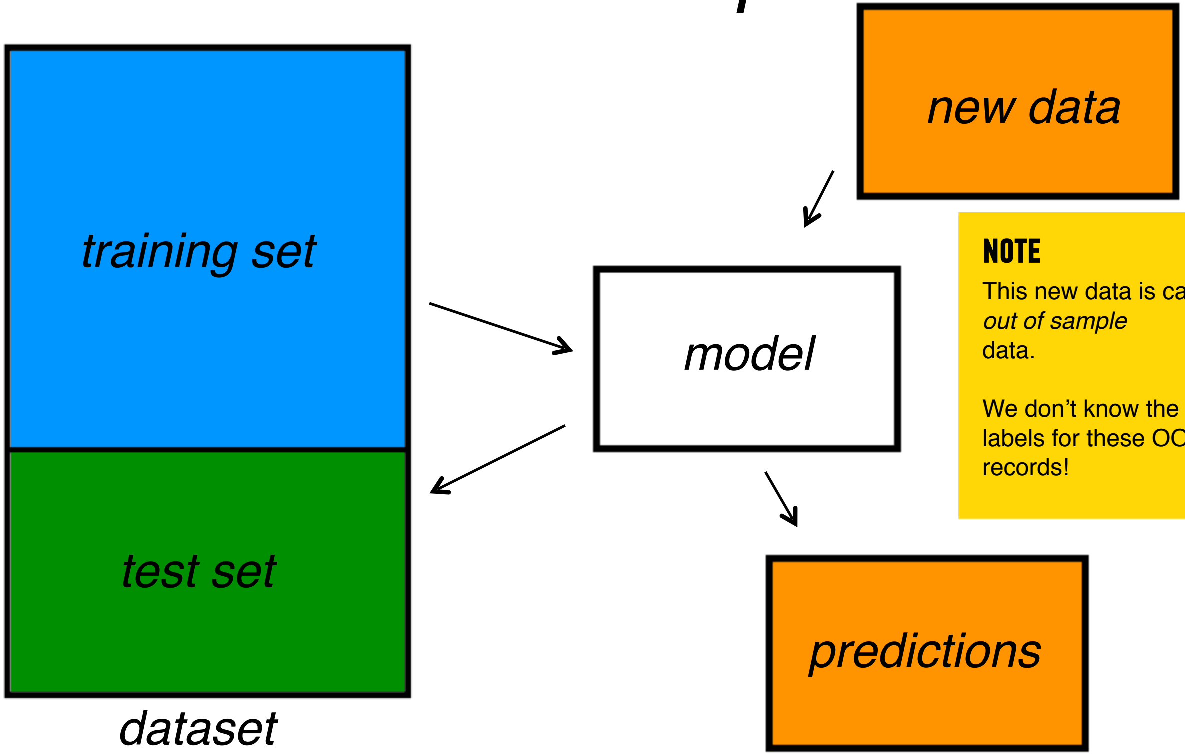
Q: What steps does a classification problem require

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



Q: What steps does a classification problem require

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



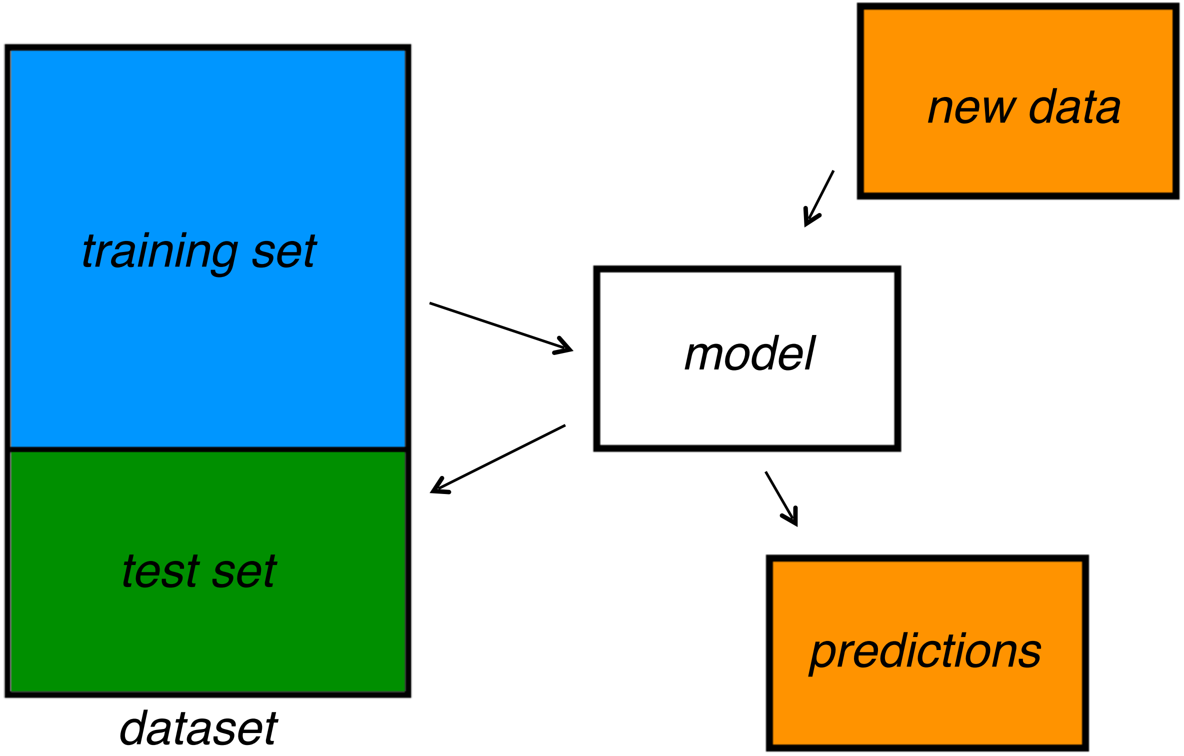
NOTE

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

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

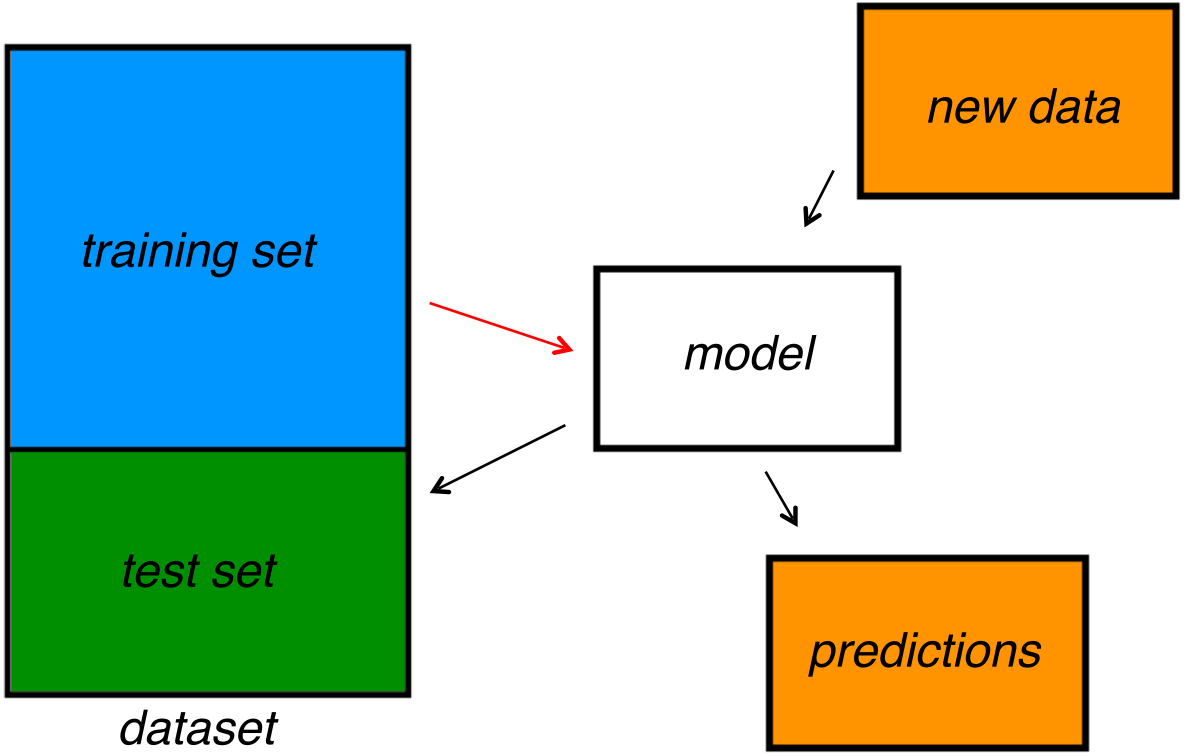
IV. BUILDING EFFECTIVE CLASSIFIERS

Q: What types of prediction error will we run into?



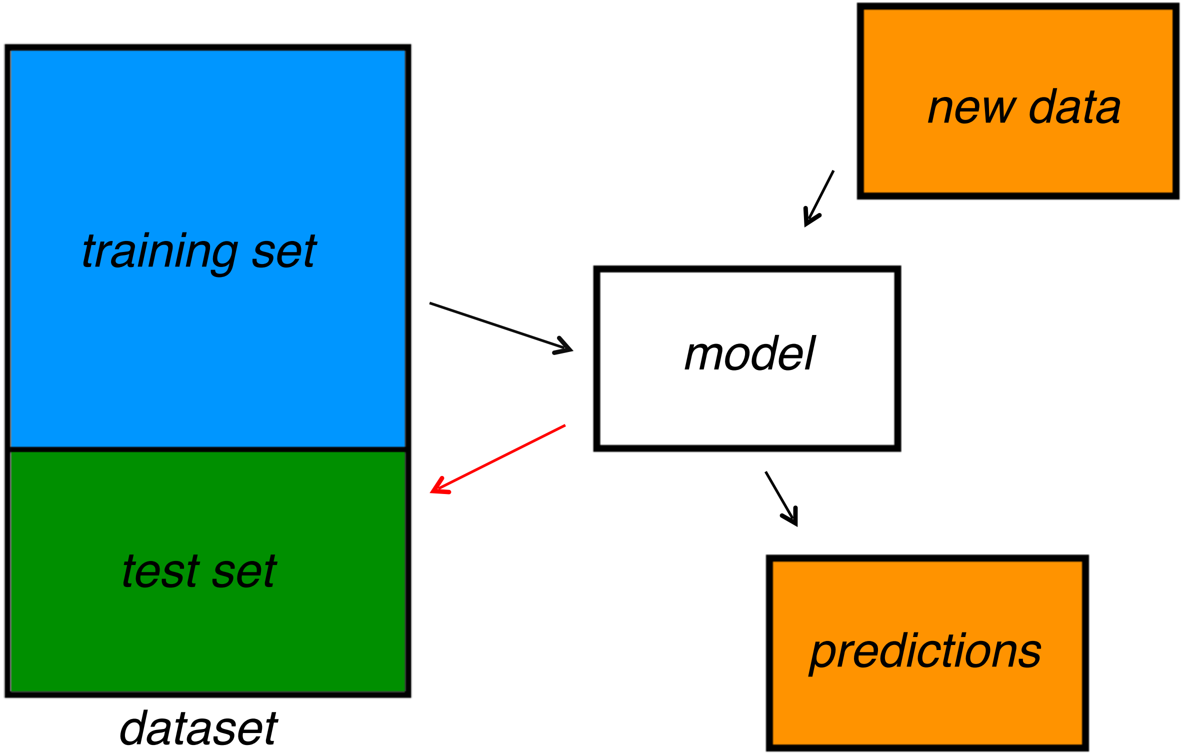
Q: What types of prediction error will we run into?

1) *training error*



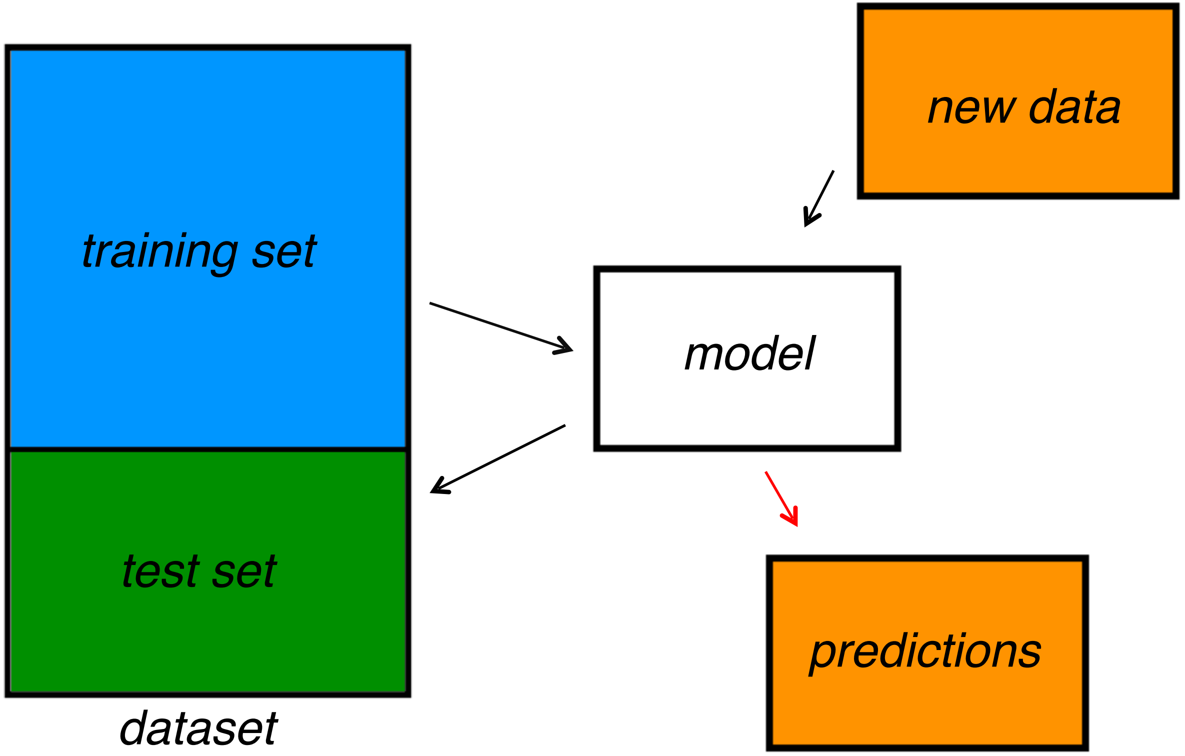
Q: What types of prediction error will we run into?

1) *training error*
2) *generalization error*



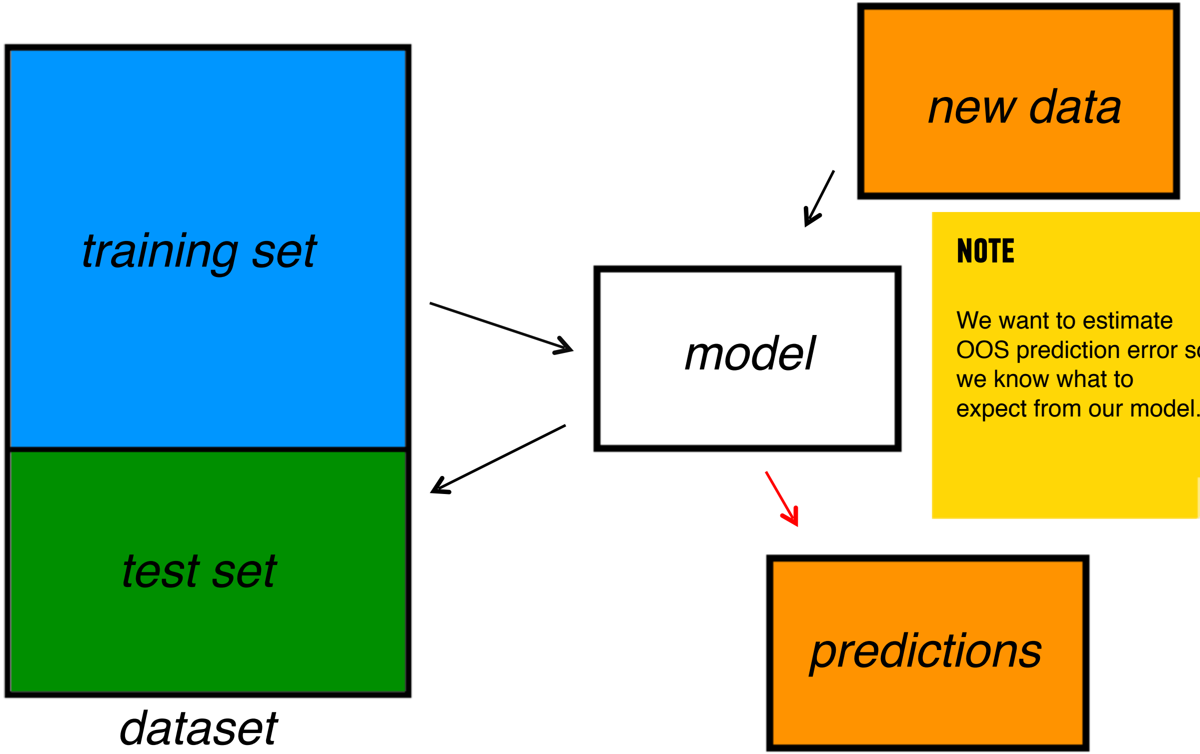
Q: What types of prediction error will we run into?

1) *training error*
2) *generalization error*
3) *OOS error*



Q: What types of prediction error will we run into?

1) *training error*
2) *generalization error*
3) *OOS error*



NOTE

We want to estimate OOS prediction error so we know what to expect from our model.

Q: Why should we use training & test sets?

Q: Why should we use training & test sets?

Thought experiment:
Suppose instead, we train our model using the entire dataset.

Q: Why should we use training & test sets?

Thought experiment:
Suppose instead, we train our model using the entire dataset.
Q: How low can we push the training error?

Q: Why should we use training & test sets?

Thought experiment:
Suppose instead, we train our model using the entire dataset.
Q: How low can we push the training error?

- *We can make the model arbitrarily complex (effectively "memorizing" the entire training set).*

Q: Why should we use training & test sets?

Thought experiment:
Suppose instead, we train our model using the entire dataset.
Q: How low can we push the training error?

- *We can make the model arbitrarily complex (effectively "memorizing" the entire training set).*

A: Down to zero!

Q: Why should we use training & test sets?

Thought experiment:
Suppose instead, we train our model using the entire dataset.
Q: How low can we push the training error?

- *We can make the model arbitrarily complex (effectively "memorizing" the entire training set).*

A: Down to zero!

NOTE

This phenomenon is called *overfitting*.

OVERFITTING

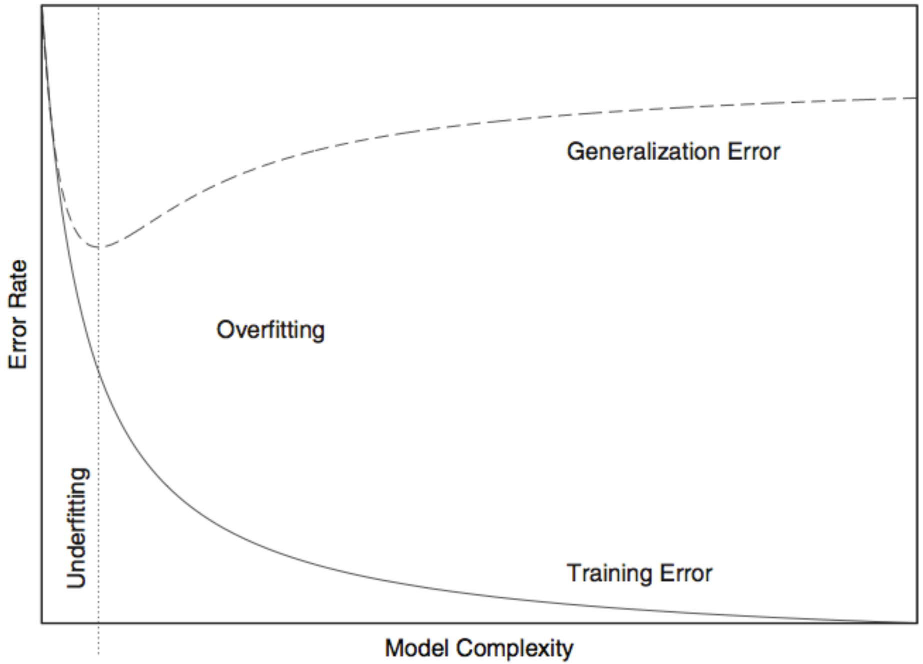
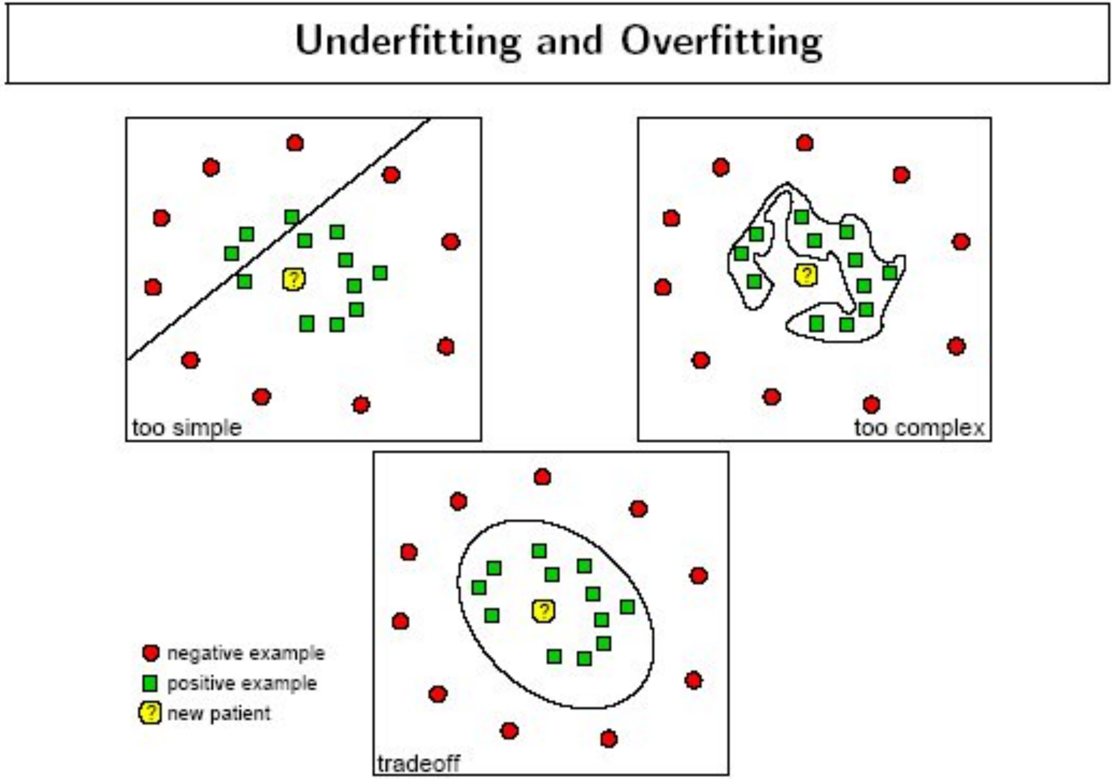


FIGURE 18-1. Overfitting: as a model becomes more complex, it becomes increasingly able to represent the training data. However, such a model is overfitted and will not generalize well to data that was not used during training.

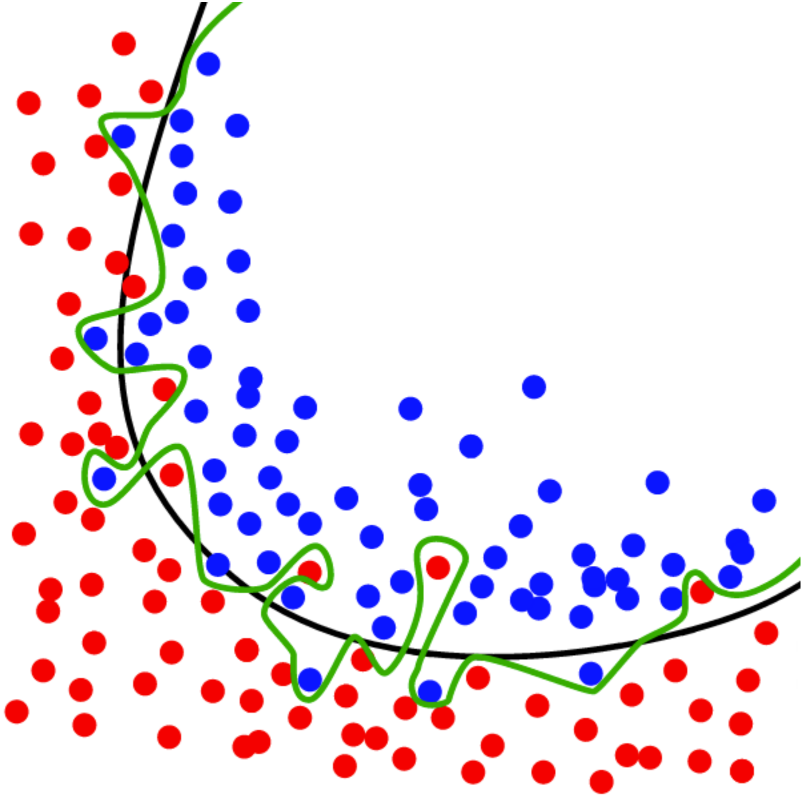
source: Data Analysis with Open Source Tools, by Philipp K. Janert. O'Reilly Media, 2011.

OVERFITTING - EXAMPLE



source: http://www.dtreg.com

OVERFITTING - EXAMPLE



source: http://www.dtreg.com

Q: Why should we use training & test sets?

Thought experiment:

Suppose instead, we train our model using the entire dataset.

Q: How low can we push the training error?

- *We can make the model arbitrarily complex (effectively "memorizing" the entire training set).*

A: Down to zero!

A: Training error is not a good estimate of OOS accuracy.

NOTE

This phenomenon is called *overfitting*.

Suppose we do the train/test split.

Suppose we do the train/test split.

Q: How well does generalization error predict OOS accuracy?

Suppose we do the train/test split.

Q: How well does generalization error predict OOS accuracy?

Thought experiment:

Suppose we had done a different train/test split.

Suppose we do the train/test split.

Q: How well does generalization error predict OOS accuracy?

Thought experiment:
Suppose we had done a different train/test split.
Q: Would the generalization error remain the same?

Suppose we do the train/test split.

Q: How well does generalization error predict OOS accuracy?

Thought experiment:
Suppose we had done a different train/test split.
Q: Would the generalization error remain the same?
A: Of course not!

Suppose we do the train/test split.

Q: How well does generalization error predict OOS accuracy?

Thought experiment:
Suppose we had done a different train/test split.
Q: Would the generalization error remain the same?
A: Of course not!

A: On its own, not very well.

Suppose we do the train/test split.

Q: How well does generalization error predict OOS accuracy?

Thought experiment:

Suppose we had done a different train/test split.

Q: Would the generalization error remain the same?

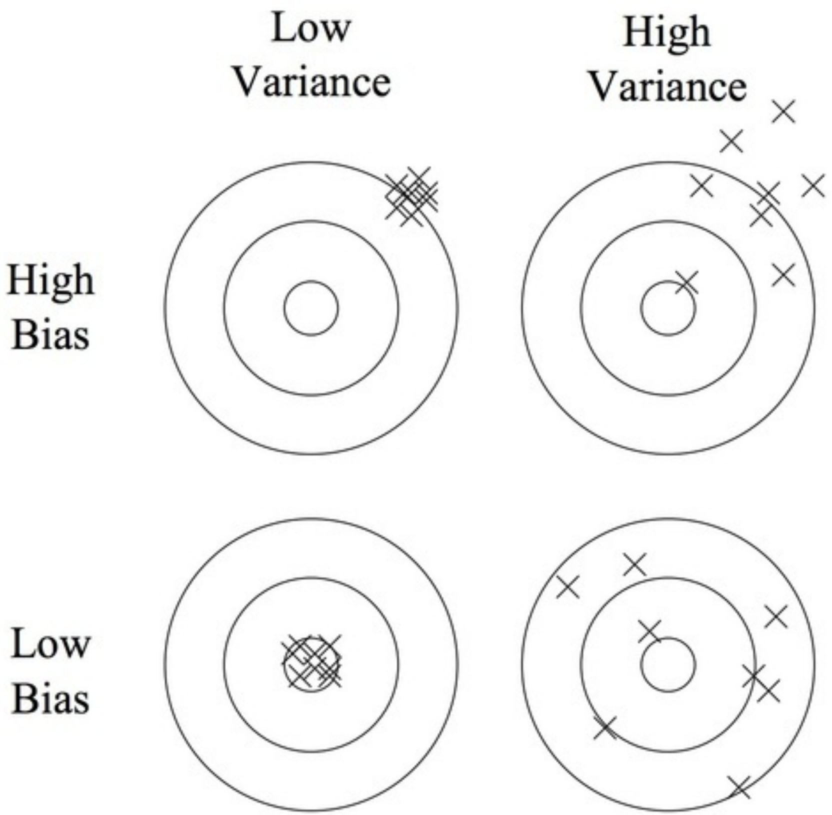
A: Of course not!

A: On its own, not very well.

NOTE

The generalization error gives a *high-variance estimate* of OOS accuracy.

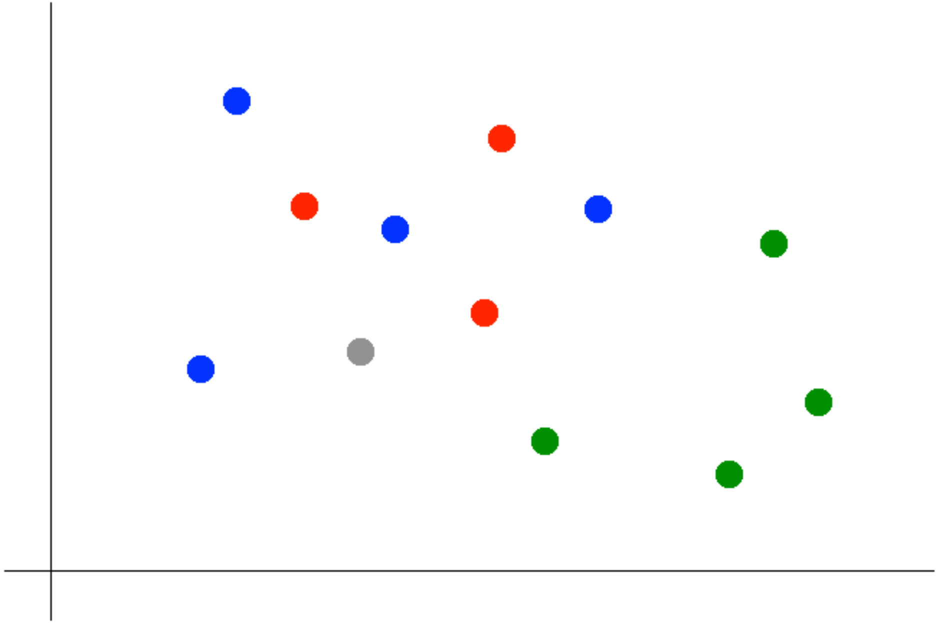
BIAS-VARIANCE



We can do better than that…. as we will see in the next class….

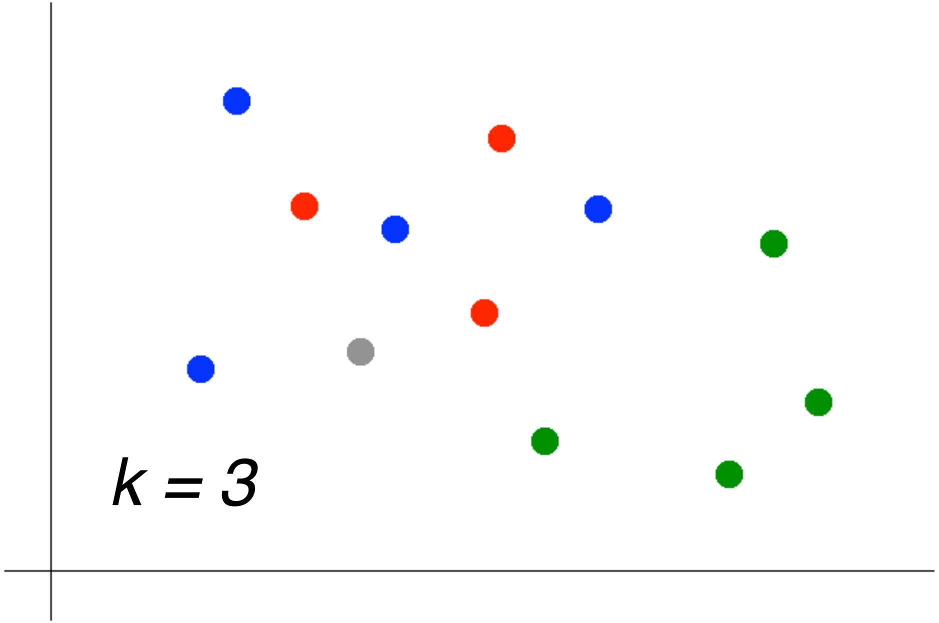
V. K-NEAREST NEIGHBORS

Suppose we want to predict the color of the grey do



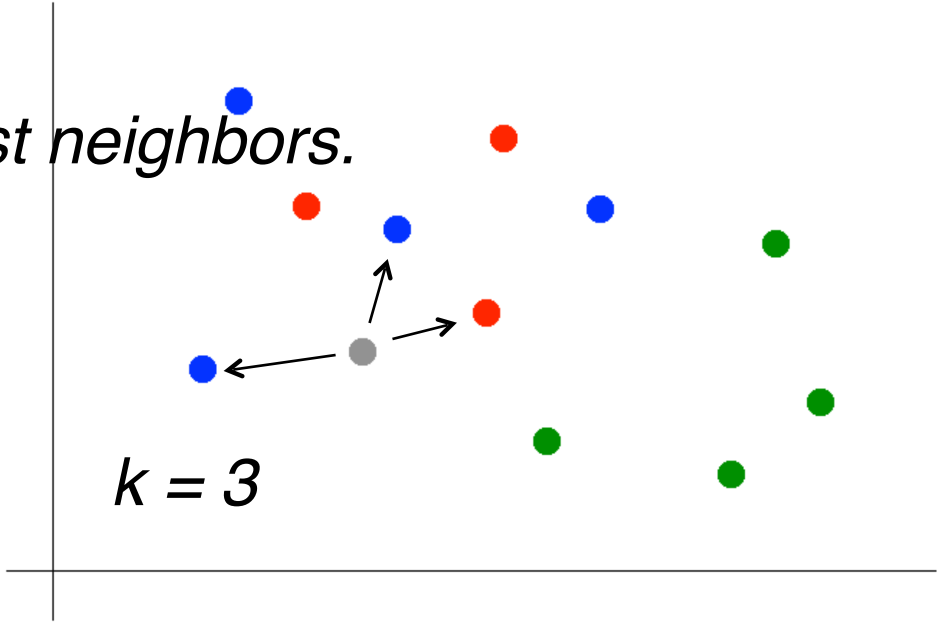
Suppose we want to predict the color of the grey do

1) *Pick a value for k.*



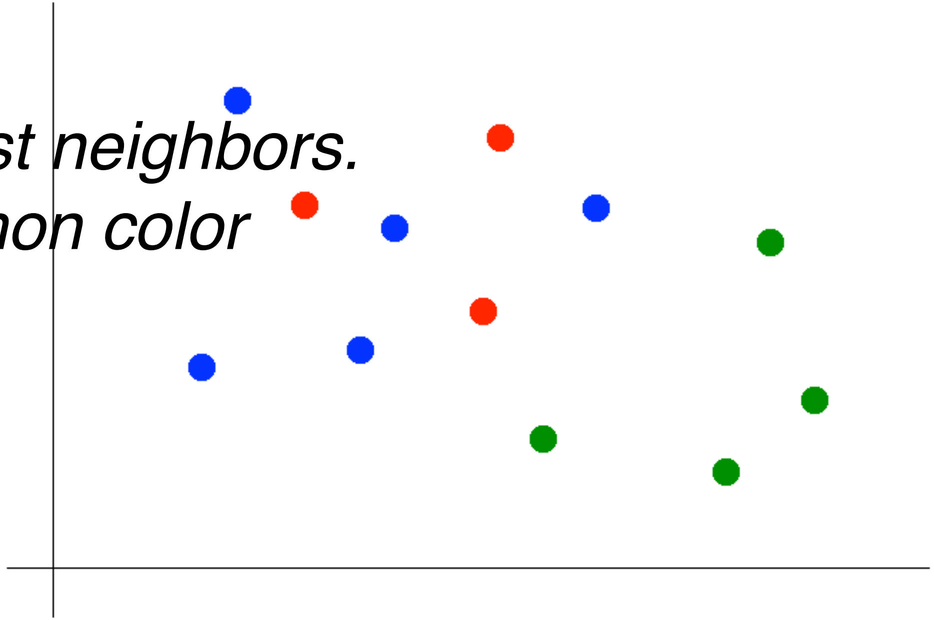
Suppose we want to predict the color of the grey do

1) *Pick a value for k.*
2) *Find colors of k nearest neighbors.*



Suppose we want to predict the color of the grey do

1) *Pick a value for k.*
2) *Find colors of k nearest neighbors.*
3) *Assign the most common color to the grey dot.*



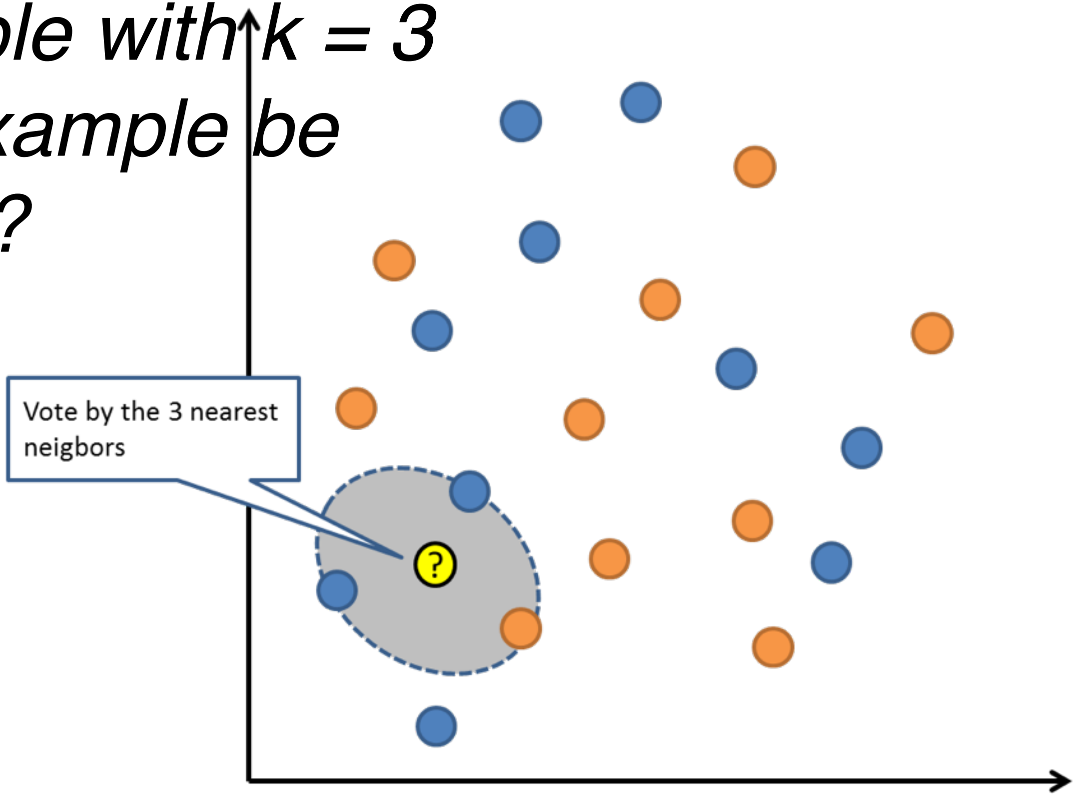
Suppose we want to predict the color of the grey do

1) *Pick a value for k.*
2) *Find colors of k nearest neighbors.*
3) *Assign the most common color to the grey dot.*

OPTIONAL NOTE

Our definition of "nearest" implicitly uses the *Euclidean distance function.*

Another example with k = 3
Will our new example be blue or orange?



LAB: KNN CLASSIFICATION