

# **DATA SCIENCE**

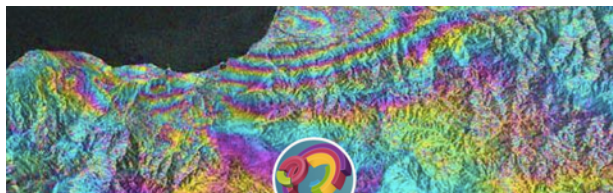
## **LECTURE 4: INTRODUCTION TO MACHINE LEARNING, CLASSIFICATION WITH K-NEAREST NEIGHBORS**

**FRANCESCO MOSCONI / ROB HALL / DAT-16**

## LAST TIME

- I. DATA SOURCES
- II. DATA FORMATS
- III. APIS
- IV. CLEANING DATA
- V. MISSING DATA

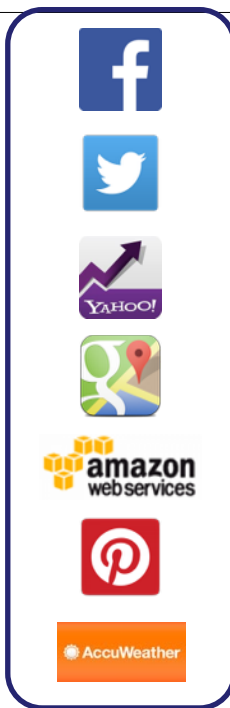
- EXERCISES:
- VI. KIMONO
- VII. PANDAS
- VIII. BOKEH



### WTF Visualizations

Visualizations that make no sense.  
For a discussion of what is wrong with a particular visualization, tweet at us [@WTFviz](#).  
Check out our friends [Thumbs Up Viz](#) and [accidental aRT](#), or [submit](#).

## Data



## Data ETL



---

**INTRO TO DATA SCIENCE**

---

**QUESTIONS?**

**WHAT WAS THE MOST INTERESTING THING YOU LEARNT?**

**WHAT WAS THE HARDEST TO GRASP?**

---

## **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**

- **UNDERSTAND THE GOAL OF ML**
- **UNDERSTAND THE DIFFERENCE BETWEEN SUPERVISED AND UNSUPERVISED LEARNING**
- **KNOW HOW TO ASSESS THE ACCURACY OF A MODEL**
- **BE ABLE TO PERFORM A SIMPLE CLASSIFICATION WITH KNN**

# **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



---

## WHAT IS MACHINE LEARNING?

---

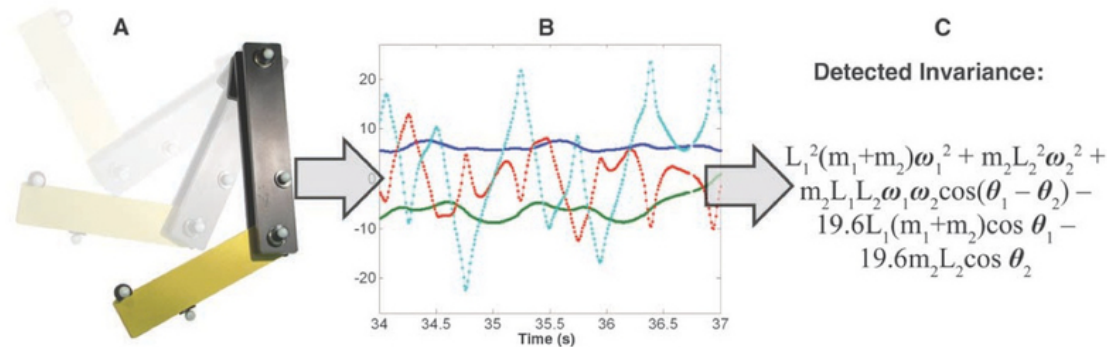
- **Machine learning** *is an area in computer science that studies and develops algorithms that can learn from data.*
- **Machine learning** *is a set of methods that can automatically detect patterns in data and use the discovered patterns to predict future data or perform other kinds of decision making*
- *Statistical learning theory, Pattern recognition*

# WHEN DO WE NEED MACHINE LEARNING?

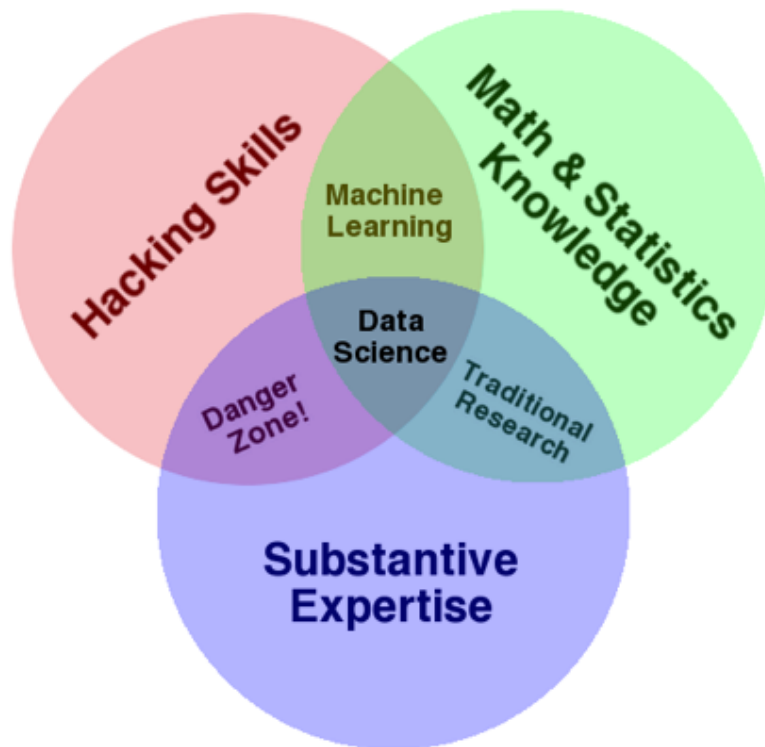
*Where we need it:*

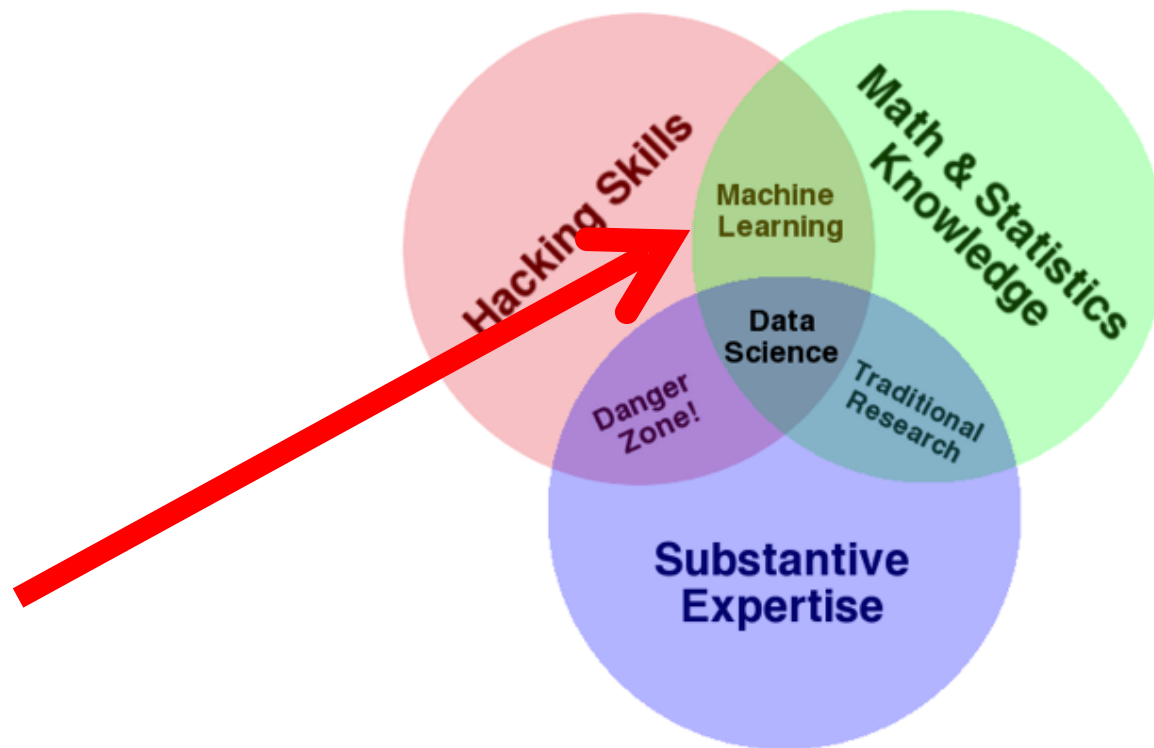
- *Some observable patterns exist*
- *There no explicitly known equations or dependencies (formulas)*
- *We have data on it*

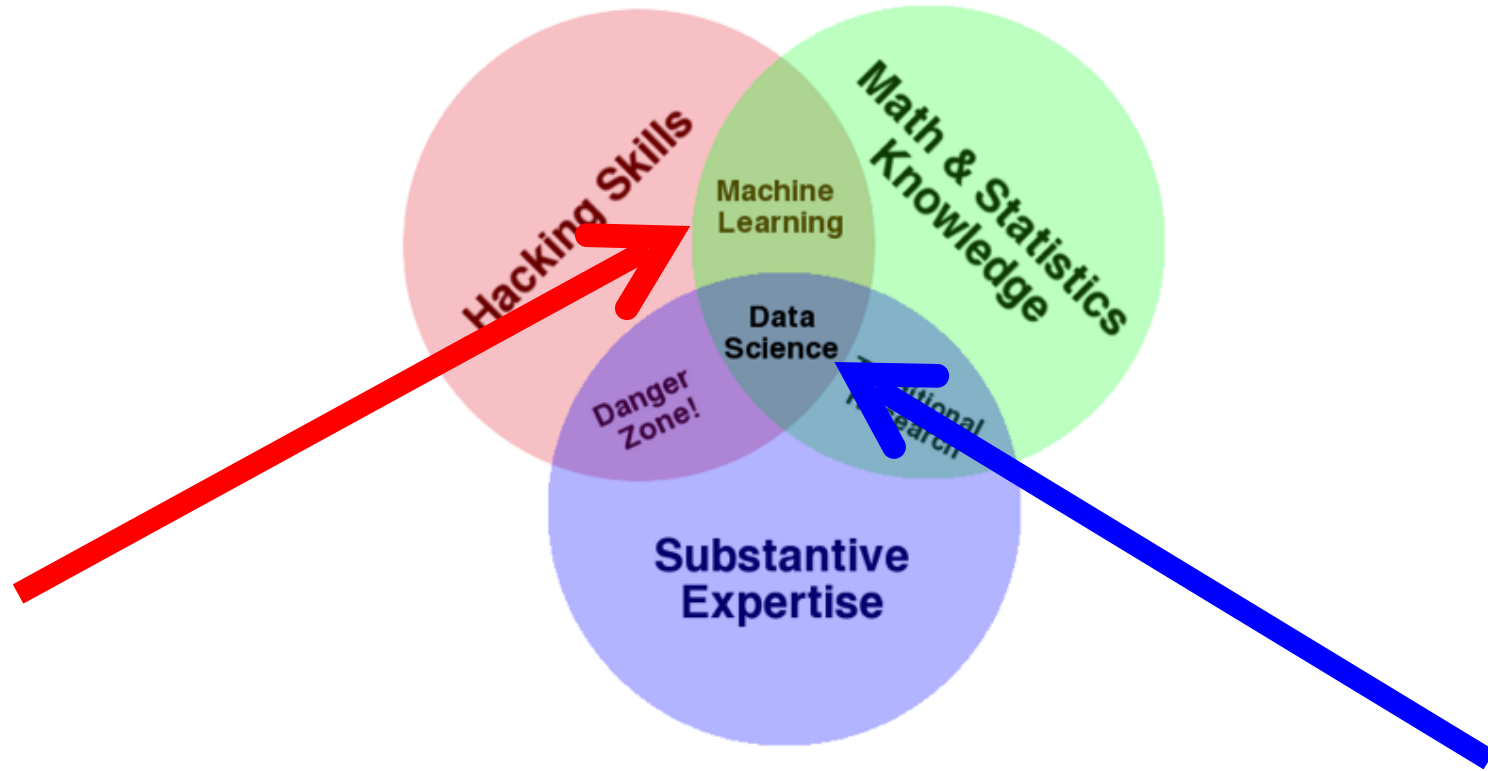
Example: Newton's second law of motion, conservation of mechanical energy, pendulum motion

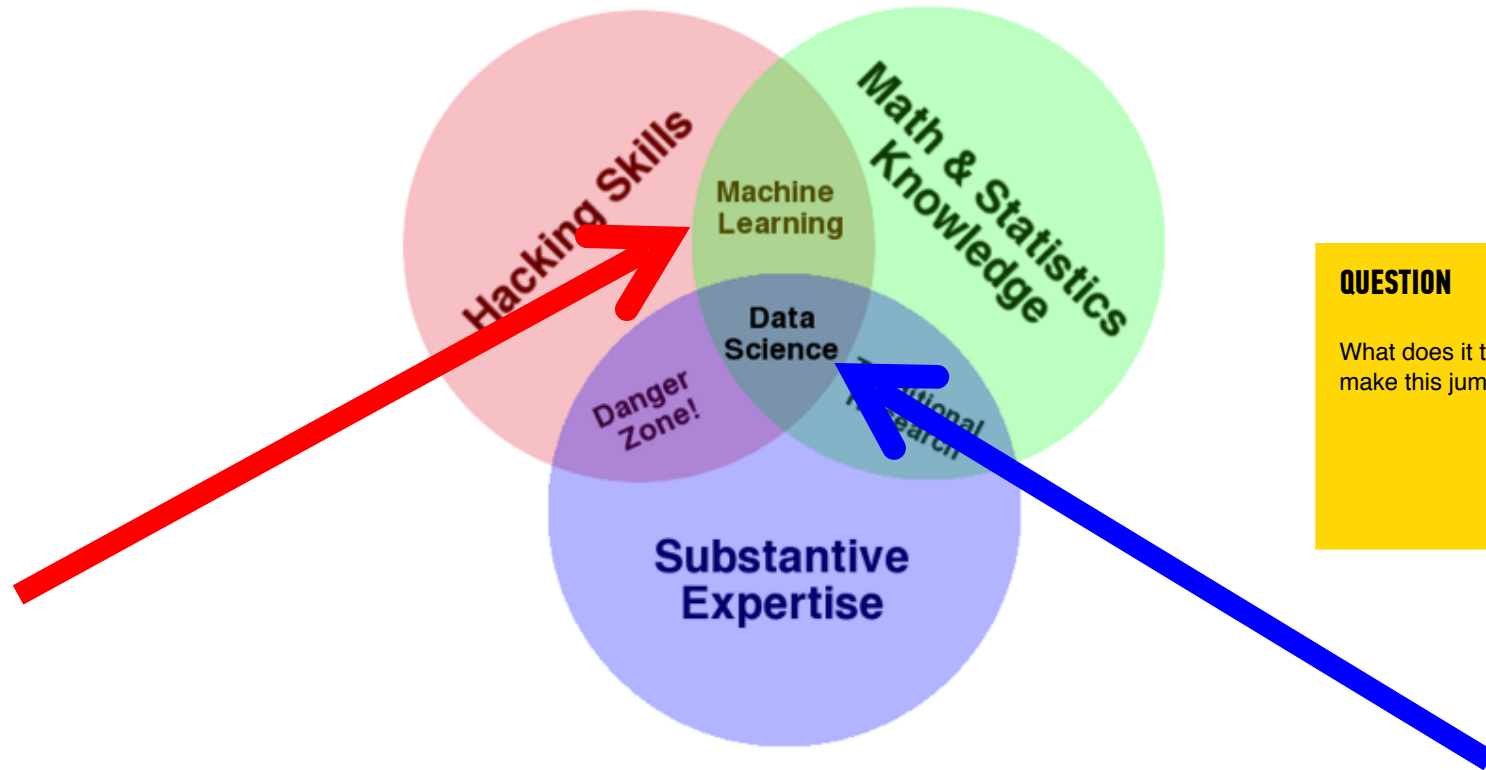


From "Distilling Free-Form Natural Laws from Experimental Data." M. Schmidt and H.Lipson. Science, 2009.



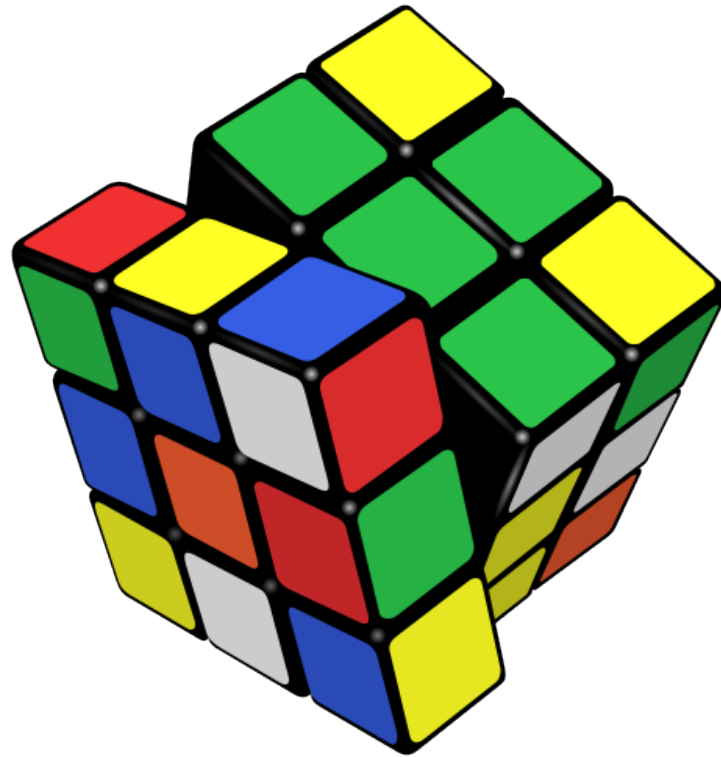


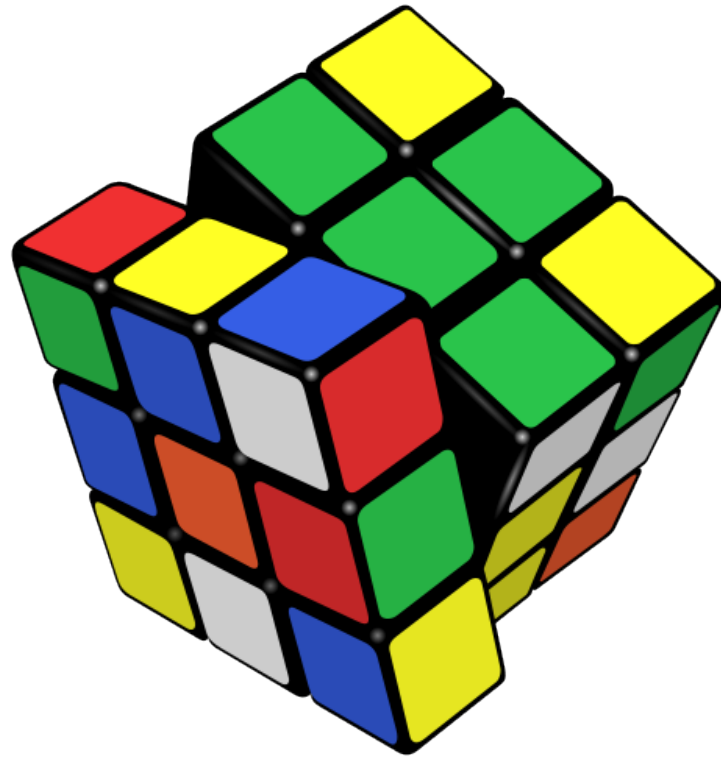




**QUESTION**

What does it take to make this jump?





**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**

---

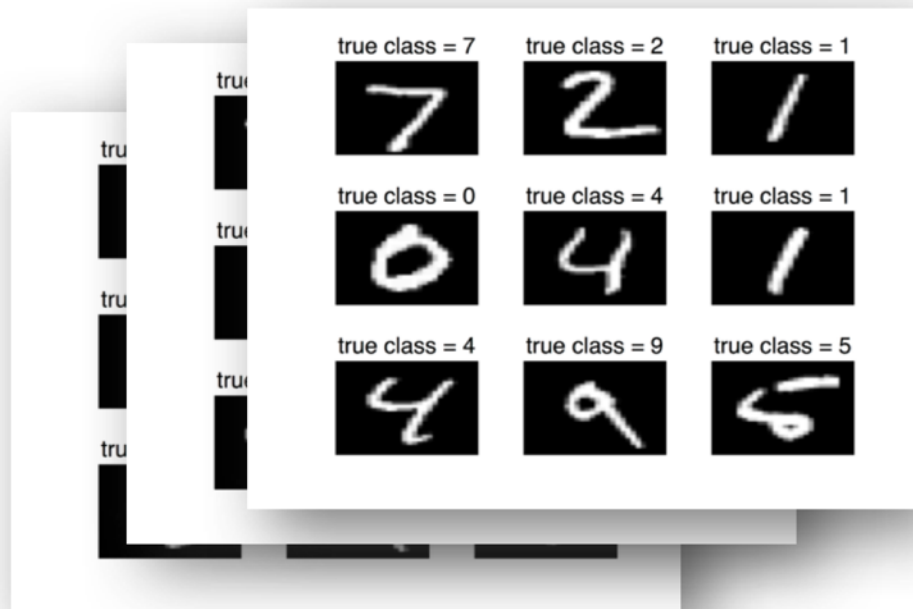
## TYPES OF LEARNING?

---

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



## CREDIT SCORING

**CLICK HERE  
TO APPLY TODAY!**



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

## CREDIT SCORING

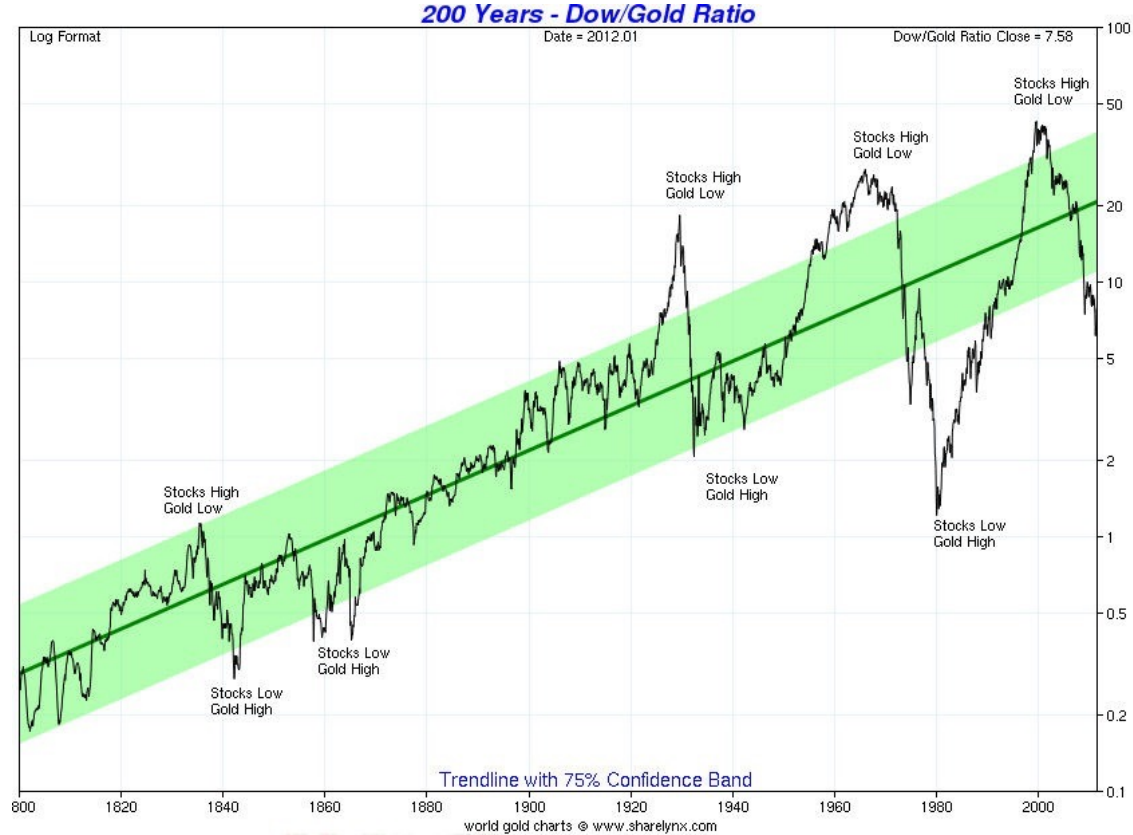
	<i>Client 1</i>	<i>Client 2</i>	<i>Client 3</i>
<i>Age</i>	23	30	19
<i>Gender</i>	M	F	M
<i>Annual salary</i>	\$30,000	\$45,000	\$15,000
<i>Years in residence</i>	3 years	1 year	3 month
<i>Years in job</i>	1 year	1 year	1 month
<i>Current debt</i>	\$5,000	\$1,000	\$10,000
<i>Paid off credit</i>	Yes	Yes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

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

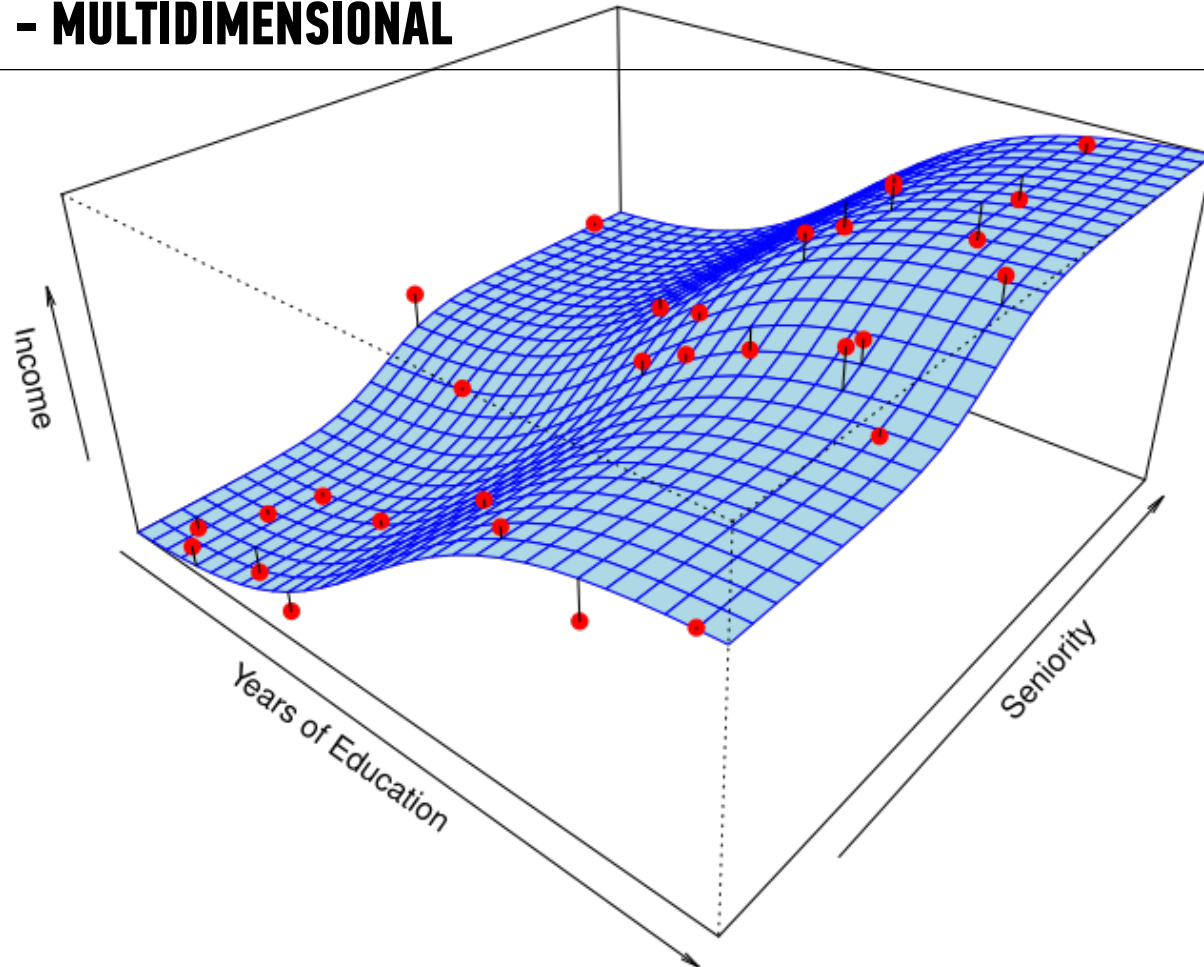




## REGRESSION - STOCK PRICE PREDICTION



## REGRESSION - MULTIDIMENSIONAL



---

## TYPES OF LEARNING?

---

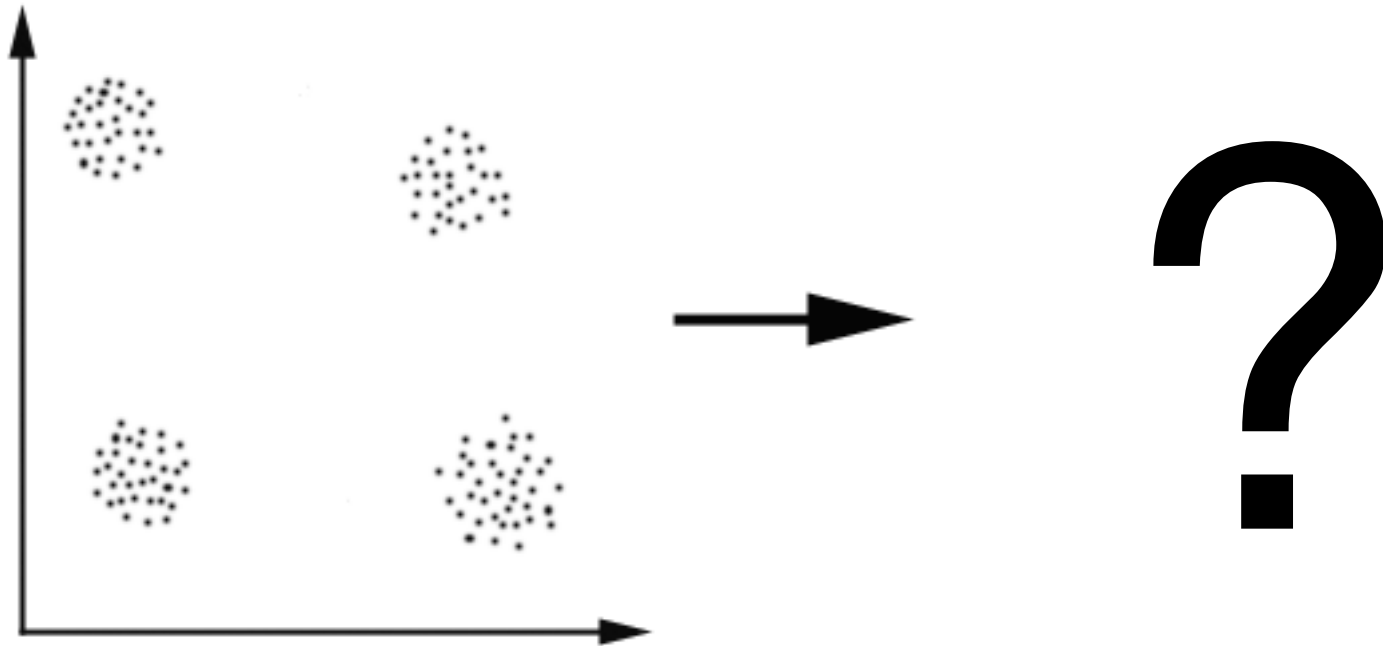
for example:

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

no label information given at all

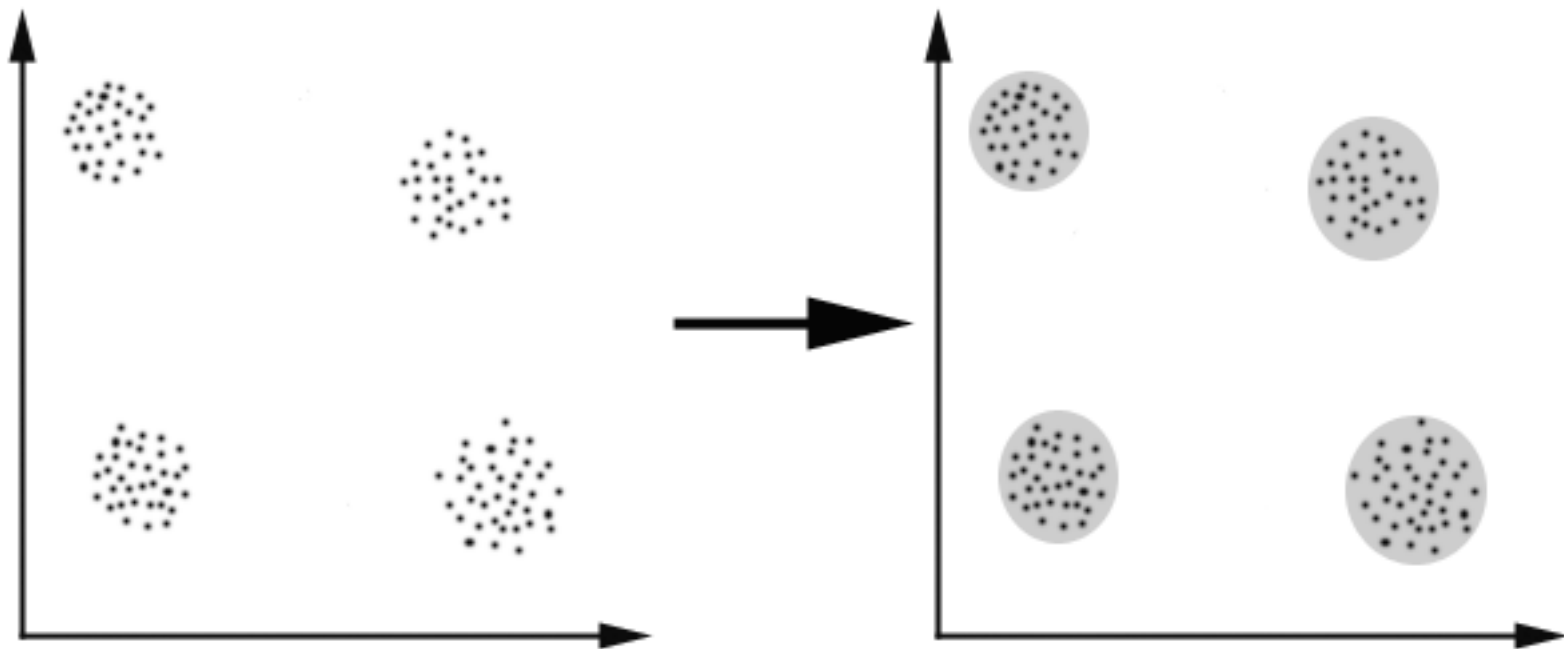
## CLUSTERING

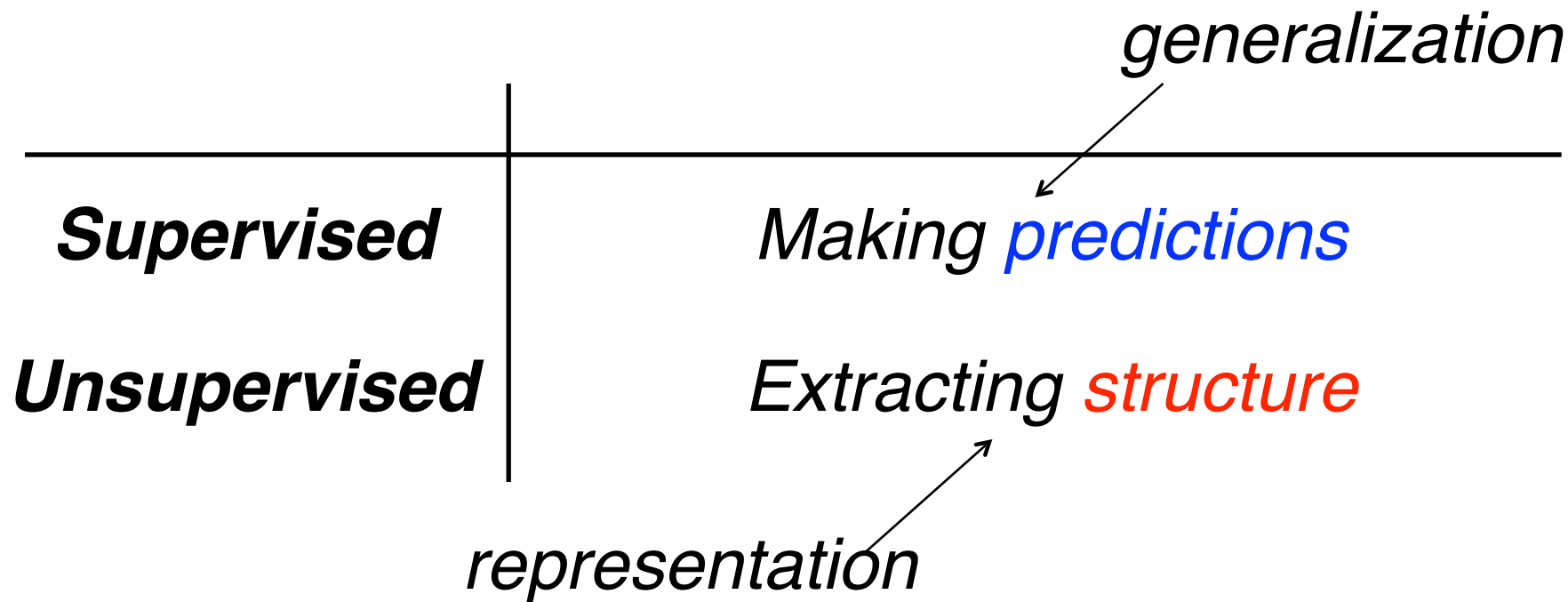
can you find structure in data given only inputs?



## CLUSTERING

can you find structure in data given only inputs?

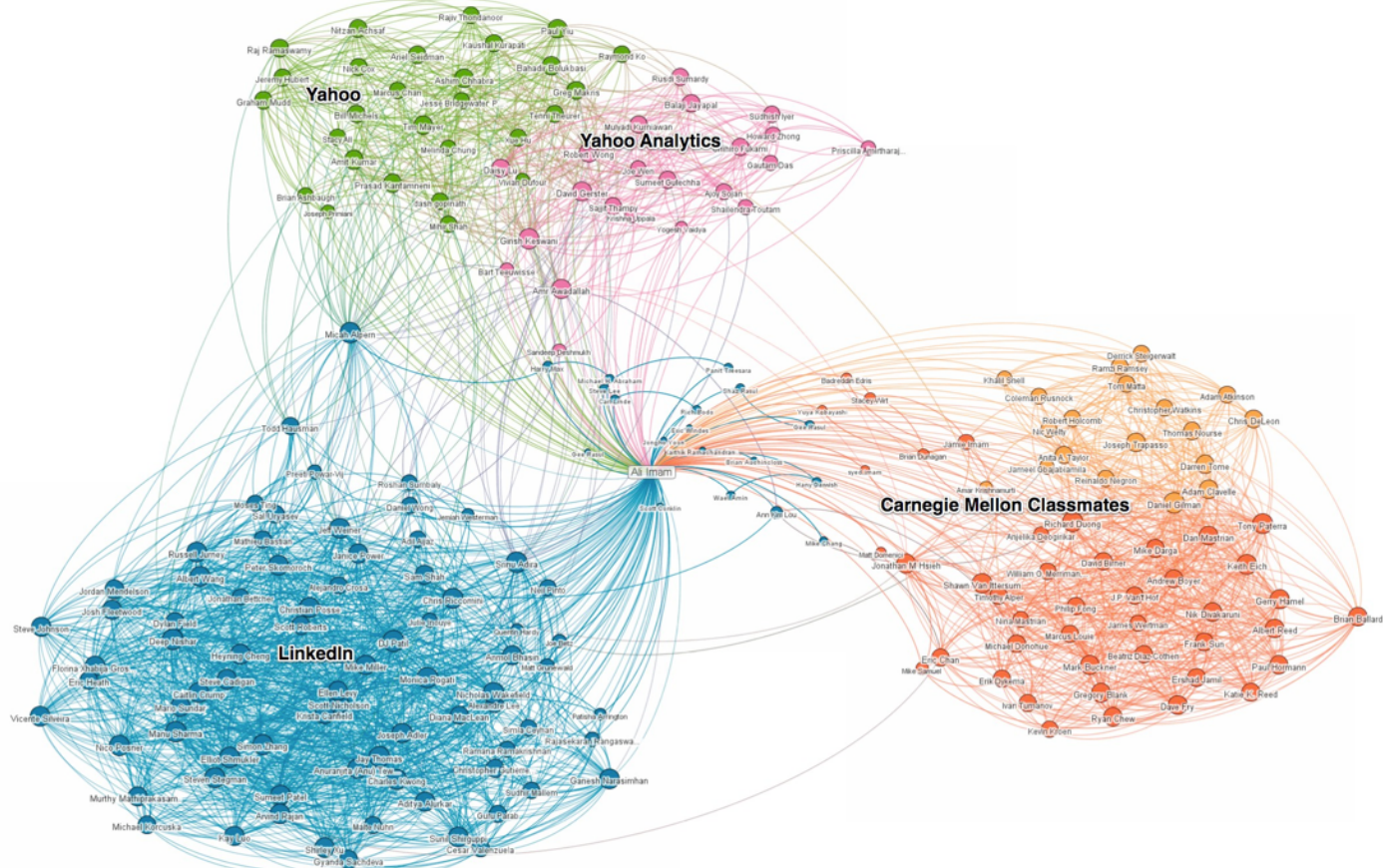




# EXERCISE:

supervised or unsupervised?

# COMMUNITY DETECTION IN SOCIAL NETWORKS





## REGRESSION - HOUSE PRICE PREDICTION



## DOCUMENT CLASSIFICATION



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

<i>Continuous</i>	<i>Categorical</i>
<i>Quantitative</i>	<i>Qualitative</i>

**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*

*Weight of cars*

*Speed of the train*

*Temperature*

*Stock price*

*Eye colors*

*Courses at GA*

*Highest degree*

*Gender*

*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*.

---

## TYPES OF DATA AND TYPE OF SOLUTION

---

combined...

	<i><b>Continuous</b></i>	<i><b>Categorical</b></i>
<i><b>Supervised</b></i>	<i>regression</i>	<i>classification</i>
<i><b>Unsupervised</b></i>	<i>dimension reduction</i>	<i>clustering</i>

**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?*



<i><b>Supervised</b></i>	<i>Making predictions</i>
<i><b>Unsupervised</b></i>	<i>Extracting structure</i>

**ANSWER**

The goal is determined  
by the type of problem.

---

QUESTION

---

*HOW  
DO YOU  
DETERMINE  
THE RIGHT  
APPROACH?*

	<i><b>Continuous</b></i>	<i><b>Categorical</b></i>
<i><b>Supervised</b></i>	<i>regression</i>	<i>classification</i>
<i><b>Unsupervised</b></i>	<i>dimension reduction</i>	<i>clustering</i>

**ANSWER**

The right approach is determined by the desired solution.

	<i><b>Continuous</b></i>	<i><b>Categorical</b></i>
<i><b>Supervised</b></i>	<i>regression</i>	<i>classification</i>
<i><b>Unsupervised</b></i>	<i>dimension reduction</i>	<i>clustering</i>

**ANSWER**

**NOTE**

The det  
des All of this depends on  
your data!

## DO WE HAVE LABELS?

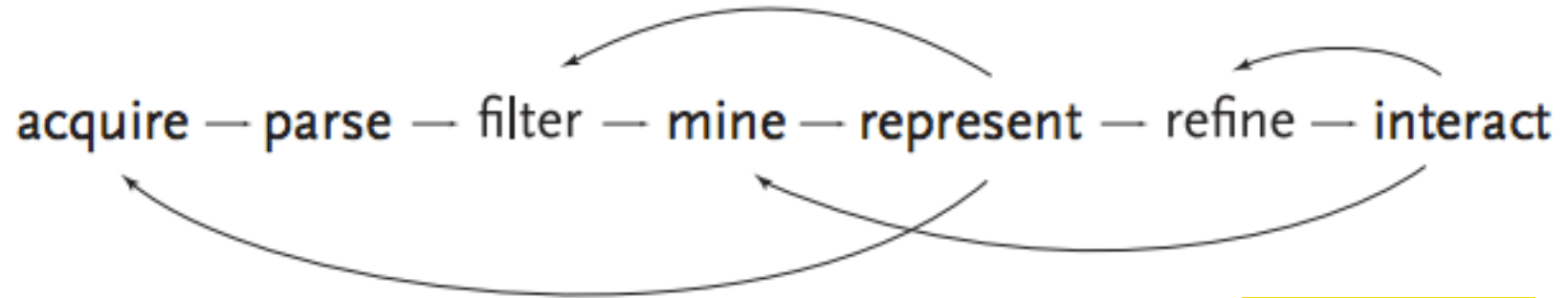


---

QUESTION

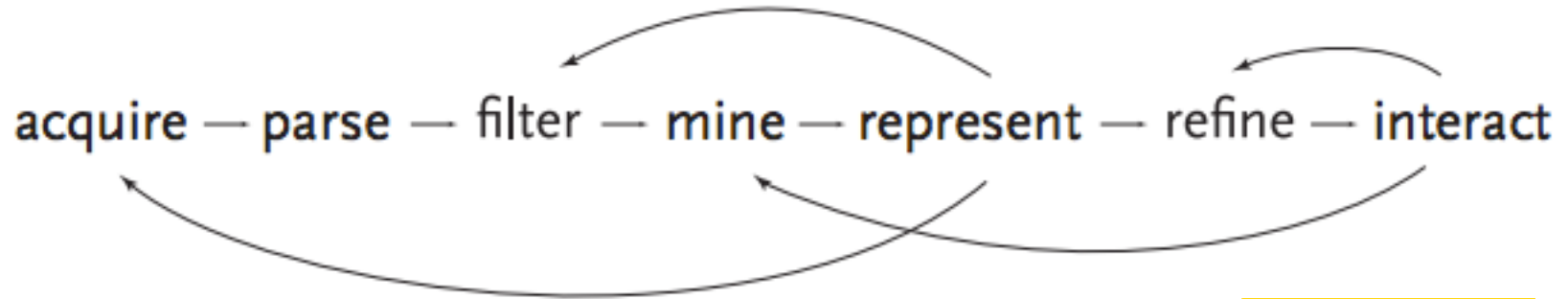
---

*WHAT  
DO YOU  
DO  
WITH YOUR  
RESULTS?*



### ANSWER

Interpret them and react accordingly.



### ANSWER

Int **NOTE**  
re:

This also relies on your  
problem solving skills!



# **III. CLASSIFICATION**

	<i><b>Continuous</b></i>	<i><b>Categorical</b></i>
<i><b>Supervised</b></i>	???	???
<i><b>Unsupervised</b></i>	???	???

	<i><b>Continuous</b></i>	<i><b>Categorical</b></i>
<i><b>Supervised</b></i>	<i>regression</i>	<i>classification</i>
<i><b>Unsupervised</b></i>	<i>dimension reduction</i>	<i>clustering</i>

*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>I. setosa</i>
4.9	3.0	1.4	0.2	<i>I. setosa</i>
4.7	3.2	1.3	0.2	<i>I. setosa</i>
4.6	3.1	1.5	0.2	<i>I. setosa</i>
5.0	3.6	1.4	0.2	<i>I. setosa</i>
5.4	3.9	1.7	0.4	<i>I. setosa</i>
4.6	3.4	1.4	0.3	<i>I. setosa</i>
5.0	3.4	1.5	0.2	<i>I. setosa</i>

*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>I. setosa</i>
4.9	3.0	1.4	0.2	<i>I. setosa</i>
4.7	3.2	1.3	0.2	<i>I. setosa</i>
4.6	3.1	1.5	0.2	<i>I. setosa</i>
5.0	3.6	1.4	0.2	<i>I. setosa</i>
5.4	3.9	1.7	0.4	<i>I. setosa</i>
4.6	3.4	1.4	0.3	<i>I. setosa</i>
5.0	3.4	1.5	0.2	<i>I. setosa</i>

*independent  
variables*



*Here's (part of) an example dataset:*

*independent variables*

*class labels (qualitative)*

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

*Q: What does “supervised” mean?*

*Q: What does “supervised” mean?*

*A: We know the labels.*

Fisher's Iris Data

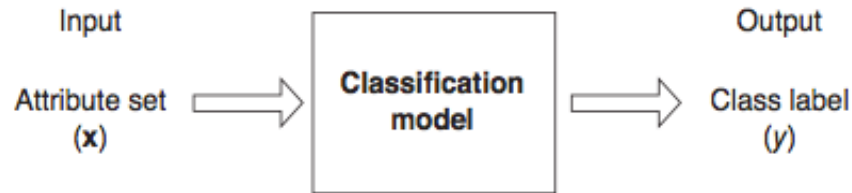
Sepal length ⇅	Sepal width ⇅	Petal length ⇅	Petal width ⇅	Species ⇅
5.1	3.5	1.4	0.2	<i>I. setosa</i>
4.9	3.0	1.4	0.2	<i>I. setosa</i>
4.7	3.2	1.3	0.2	<i>I. setosa</i>
4.6	3.1	1.5	0.2	<i>I. setosa</i>
5.0	3.6	1.4	0.2	<i>I. setosa</i>
5.4	3.9	1.7	0.4	<i>I. setosa</i>
4.6	3.4	1.4	0.3	<i>I. setosa</i>
5.0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

*class  
labels  
(qualitative)*



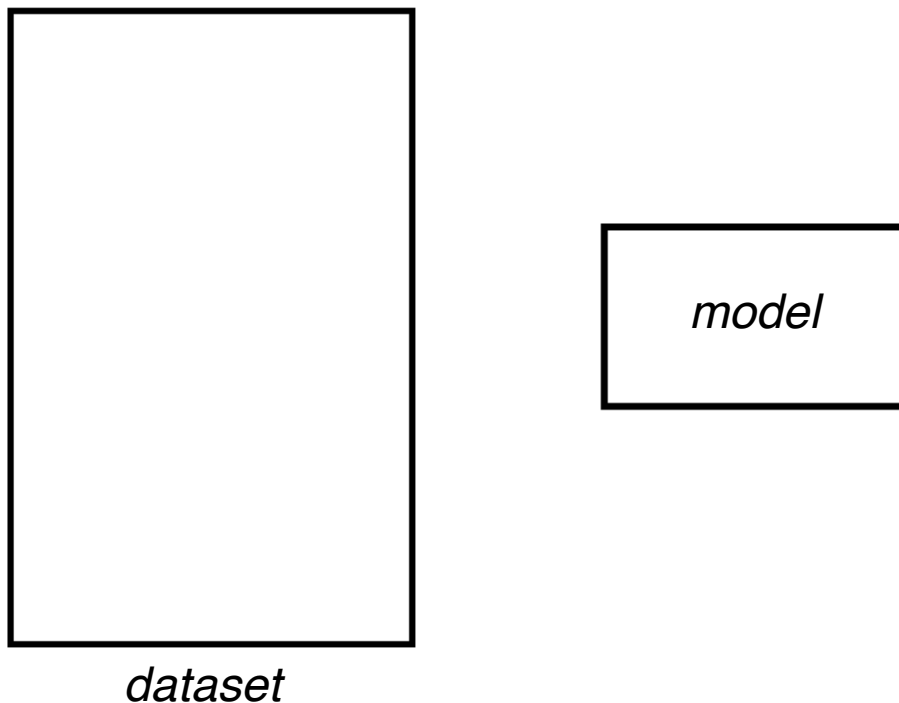
*Q: How does a classification problem work?*

*Q: How does a classification problem work?*  
*A: Data in, predicted labels out.*



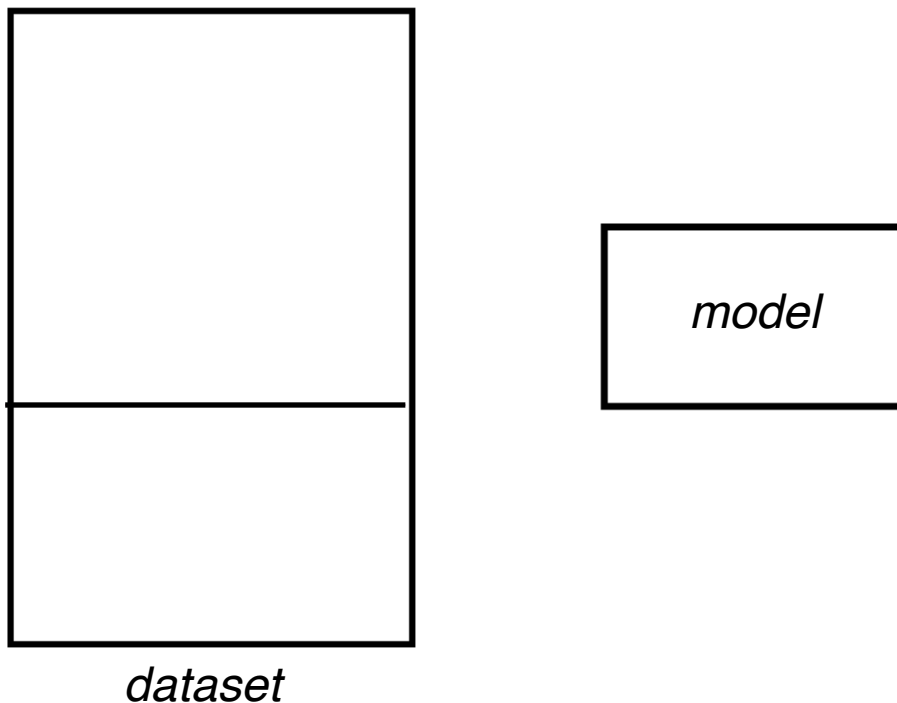
**Figure 4.2.** Classification as the task of mapping an input attribute set  $x$  into its class label  $y$ .

*Q: What steps does a classification problem require*



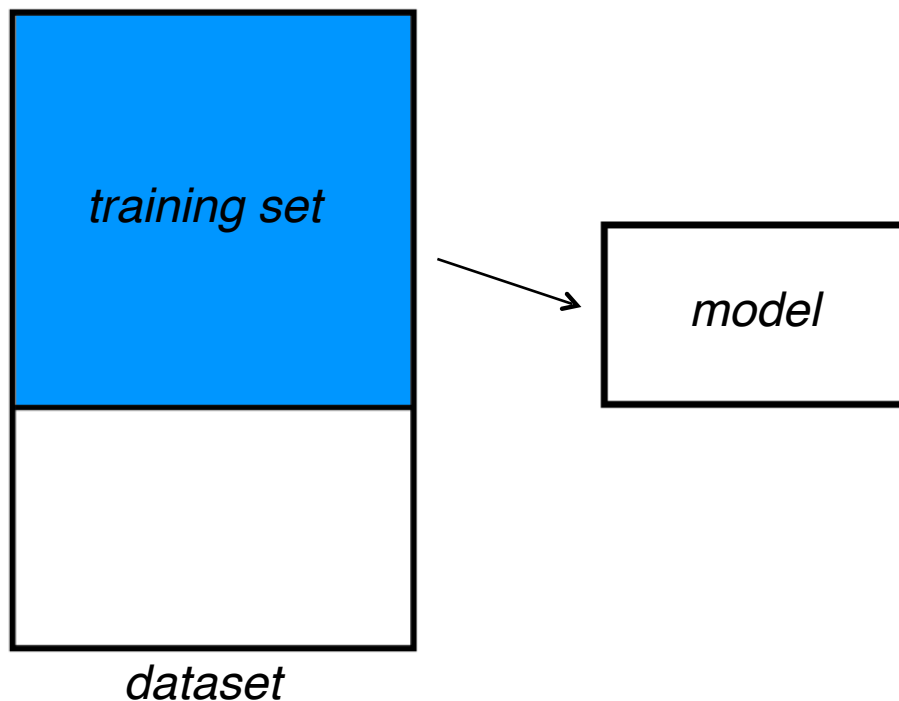
*Q: What steps does a classification problem require*

*1) split dataset*



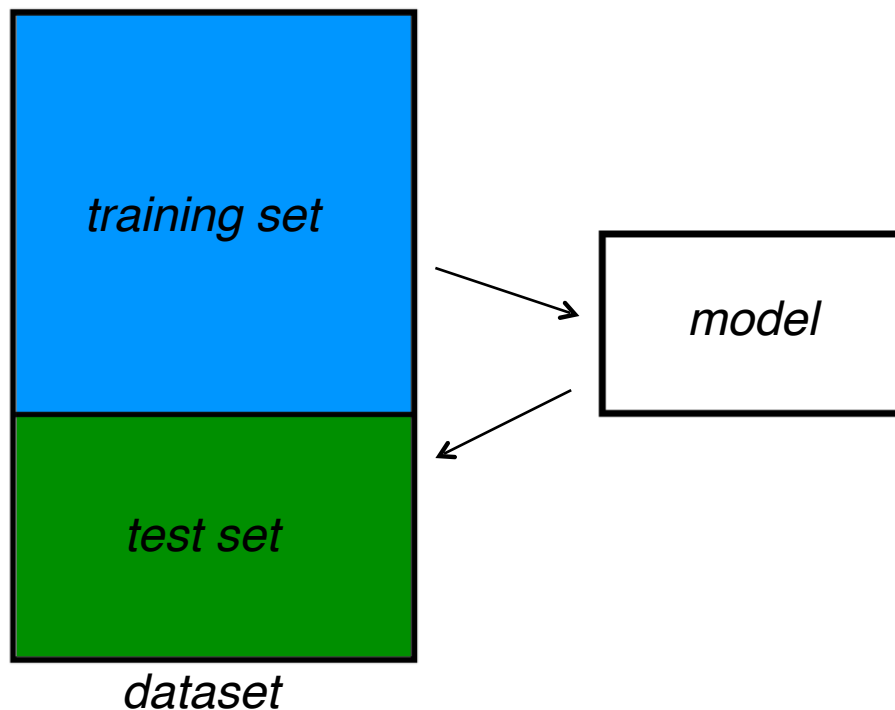
*Q: What steps does a classification problem require*

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



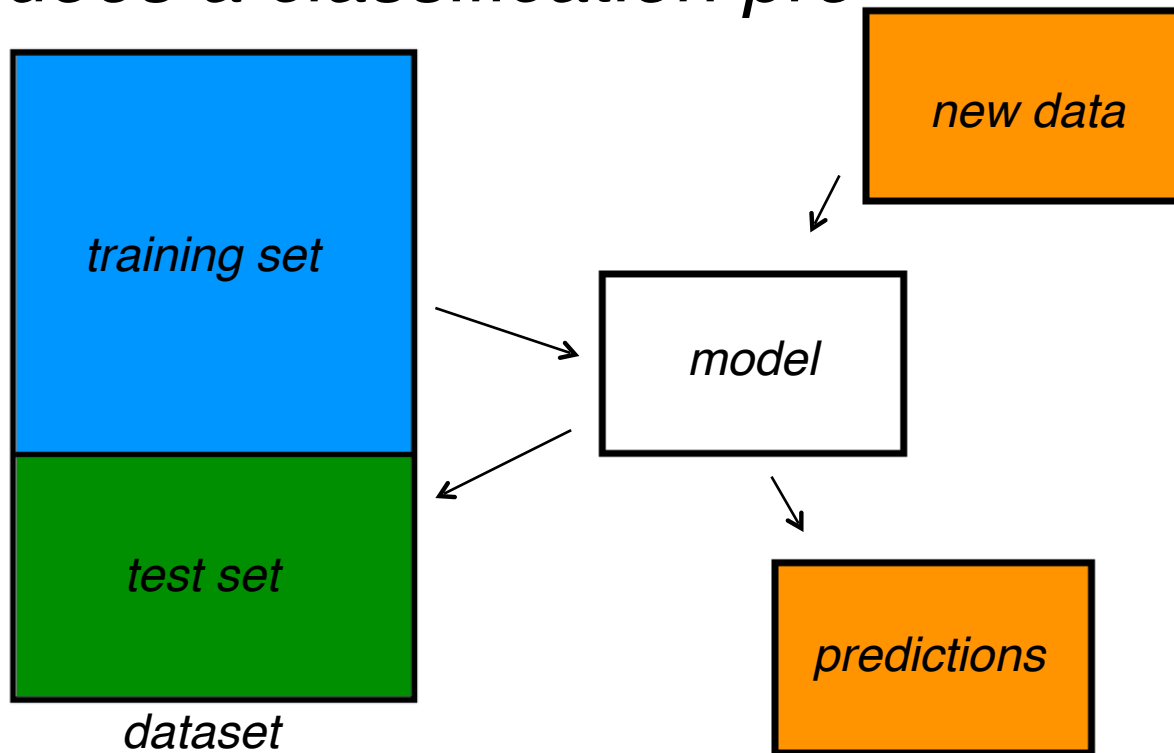
*Q: What steps does a classification problem require*

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



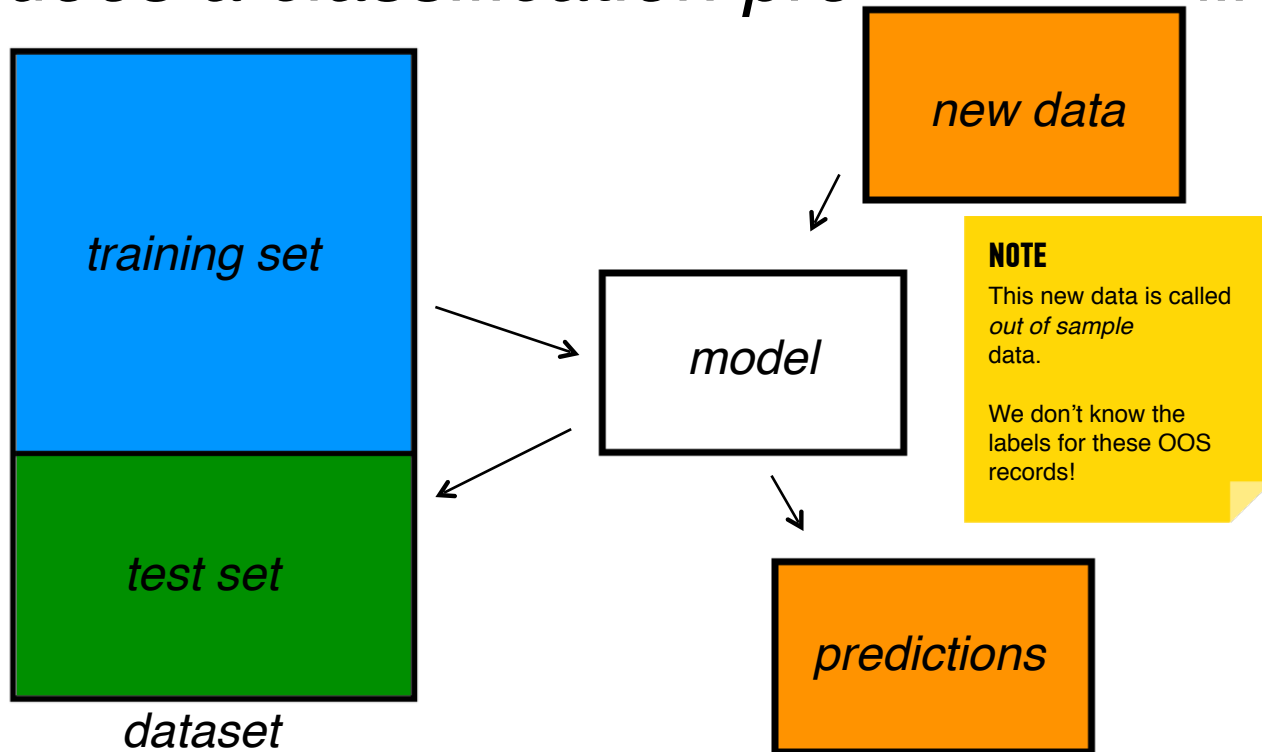
*Q: What steps does a classification problem require*

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



*Q: What steps does a classification problem require*

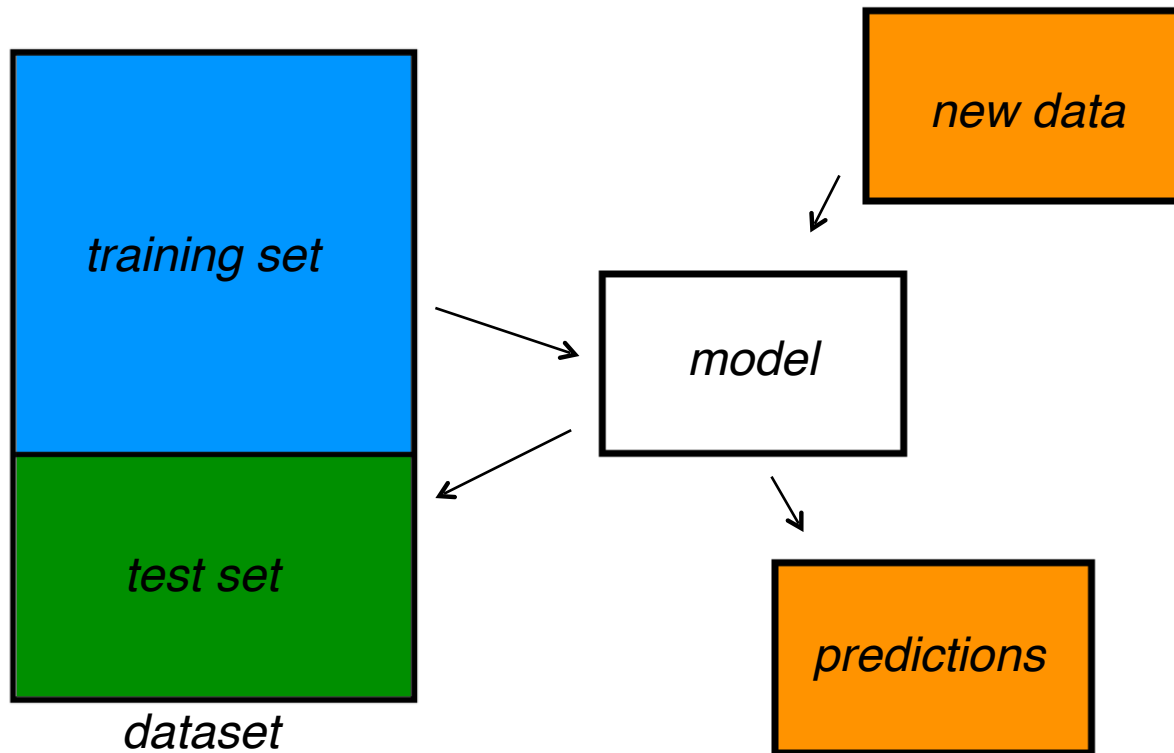
- 1) *split dataset*
- 2) *train model*
- 3) *test model*
- 4) *make predictions*





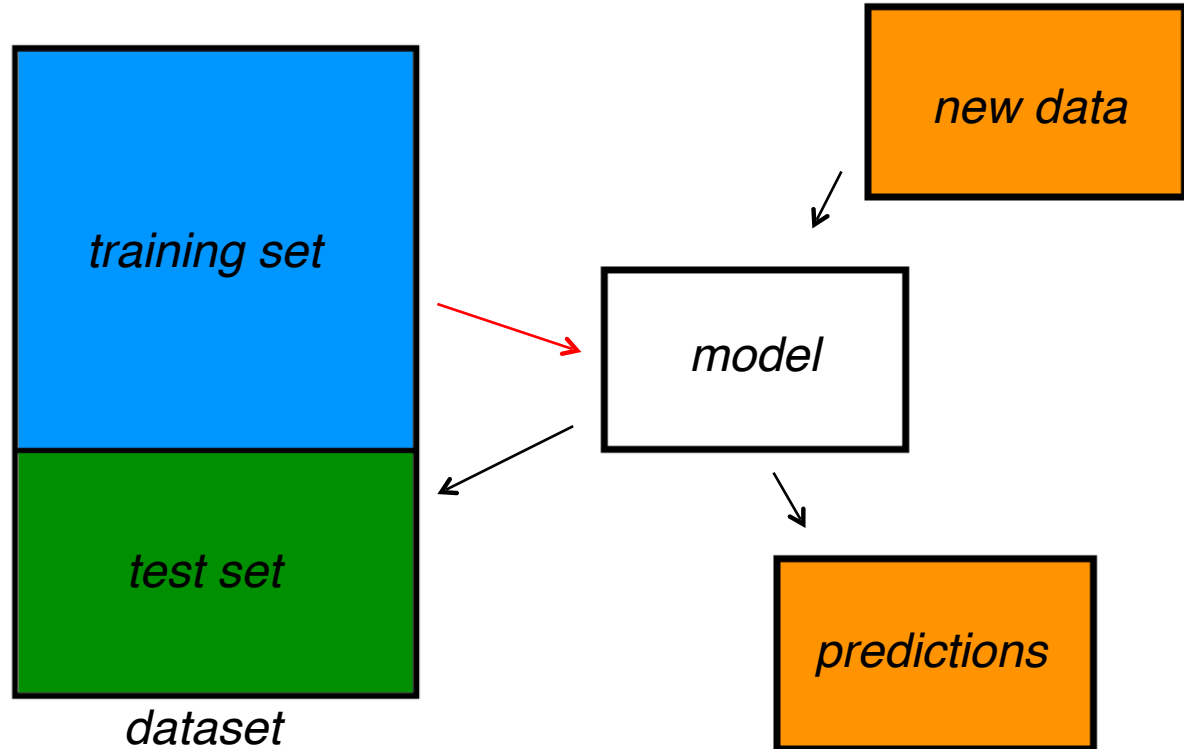
# BUILDING EFFECTIVE CLASSIFIERS

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



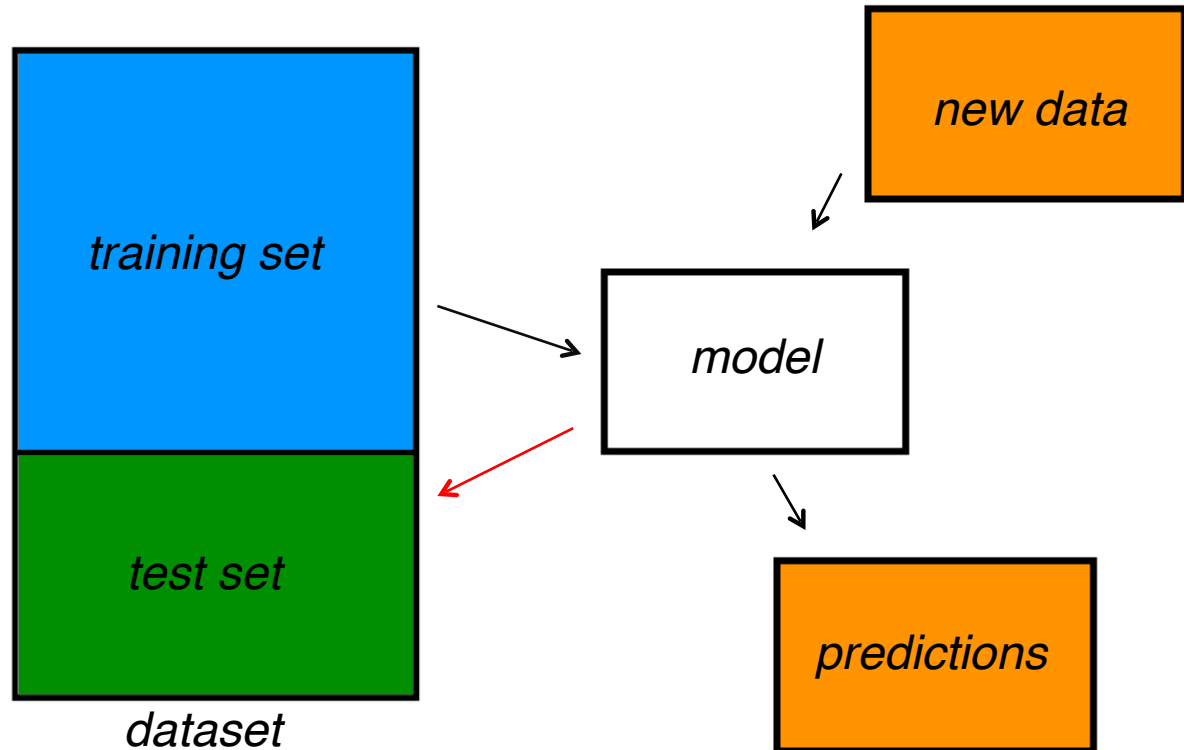
*Q: What types of prediction error will we run into?*

*1) training error*



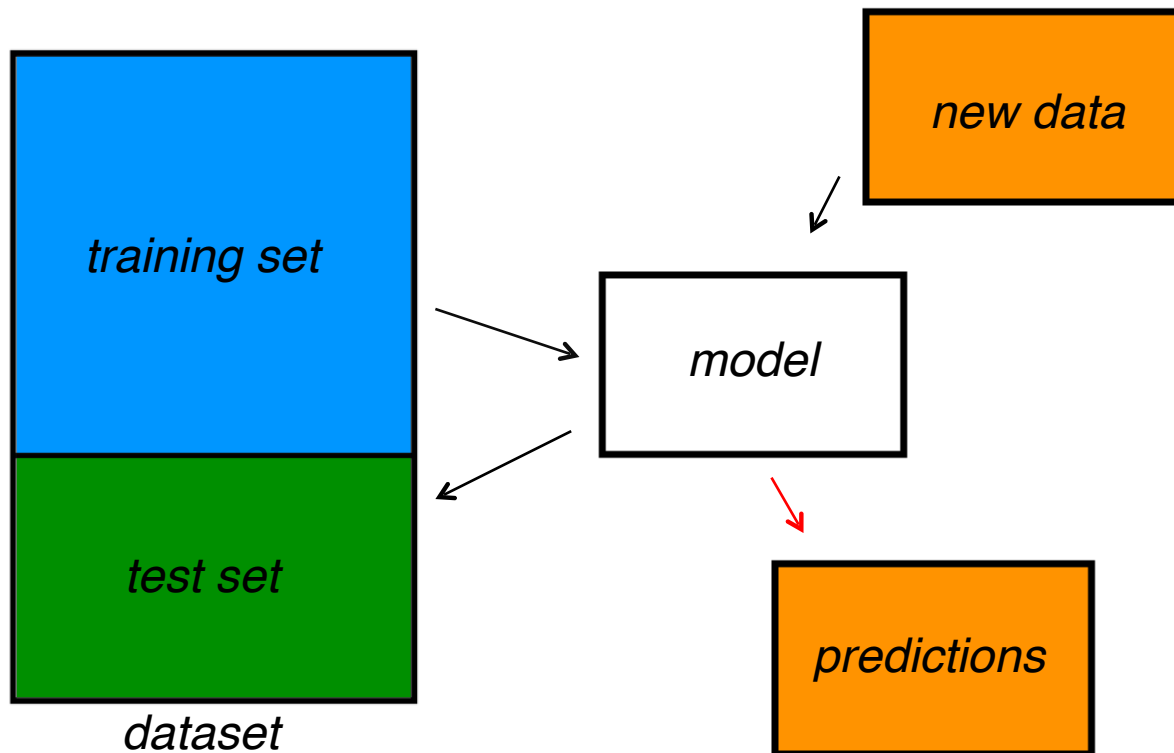
*Q: What types of prediction error will we run into?*

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



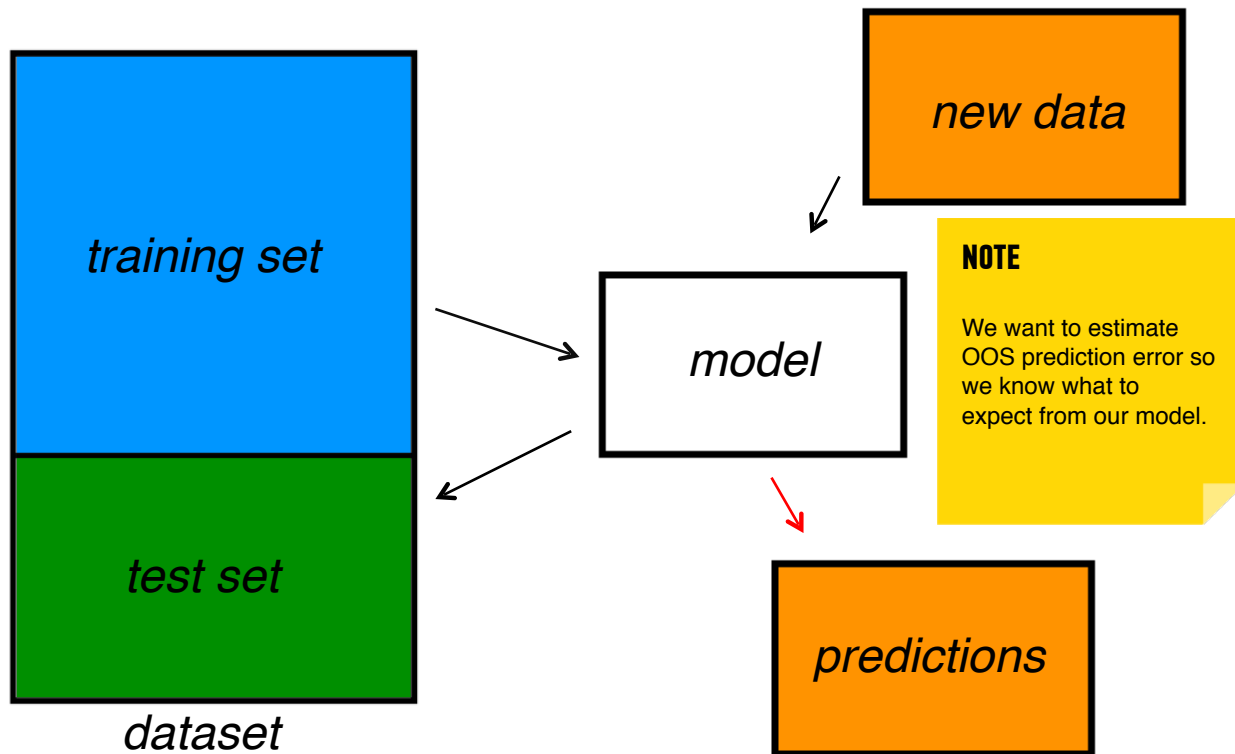
*Q: What types of prediction error will we run into?*

- 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*



*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.*

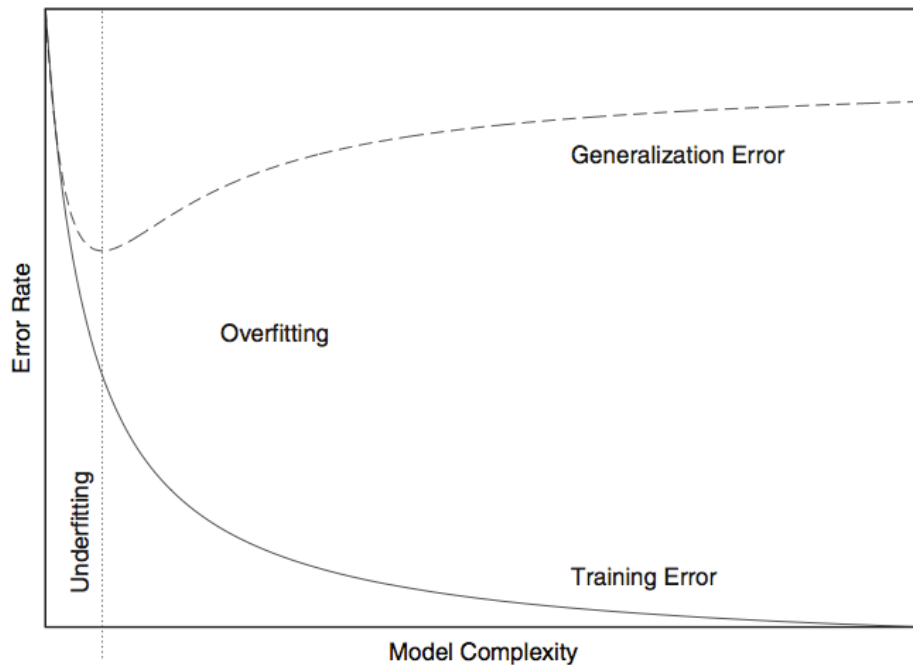
*Q: How low can we push the training error?*

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

*A: Down to zero!*

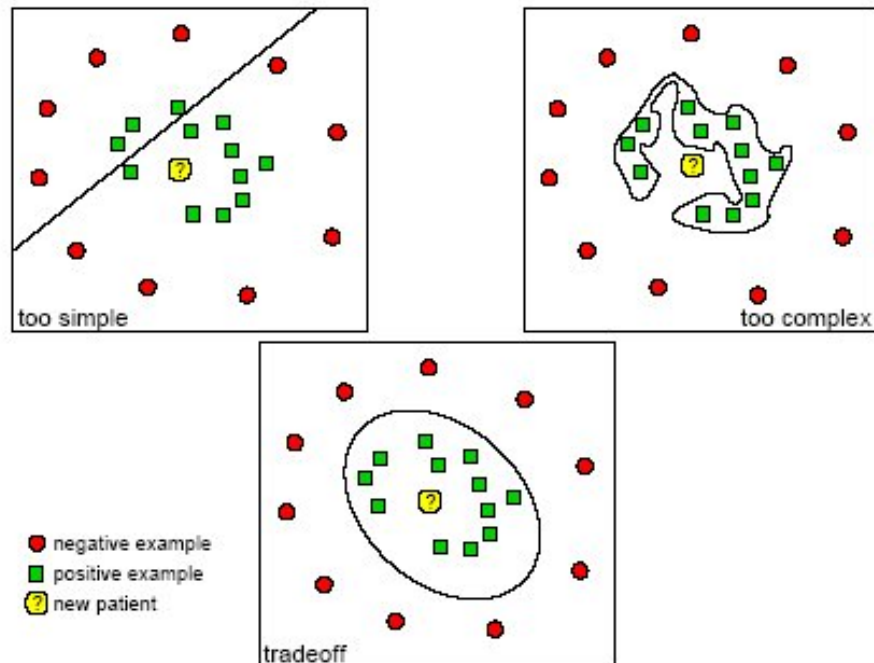
**NOTE**

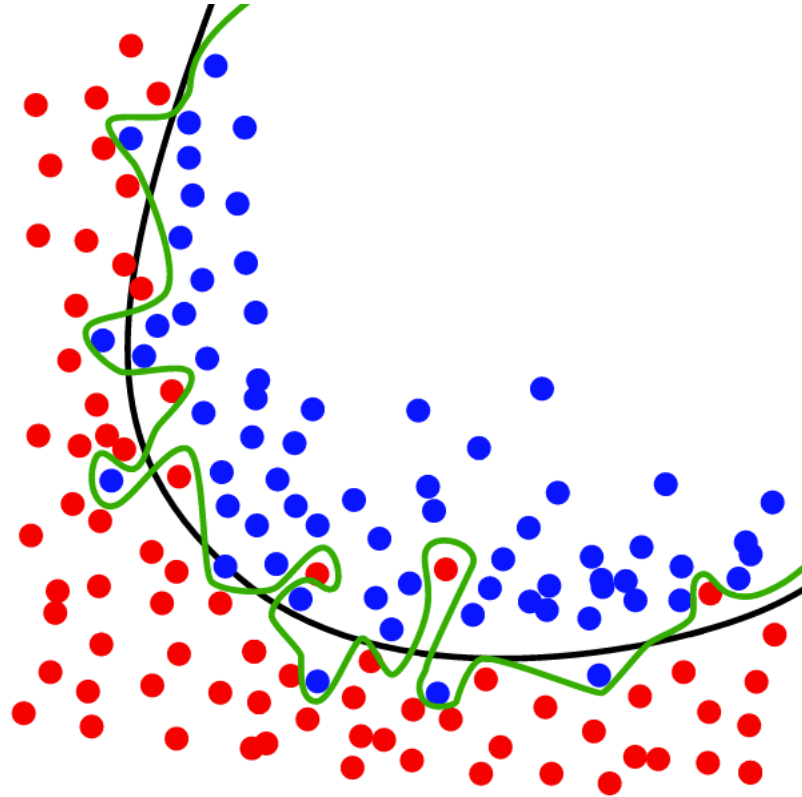
This phenomenon  
is called  
*overfitting*.



**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.*

## Underfitting and Overfitting





*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?*

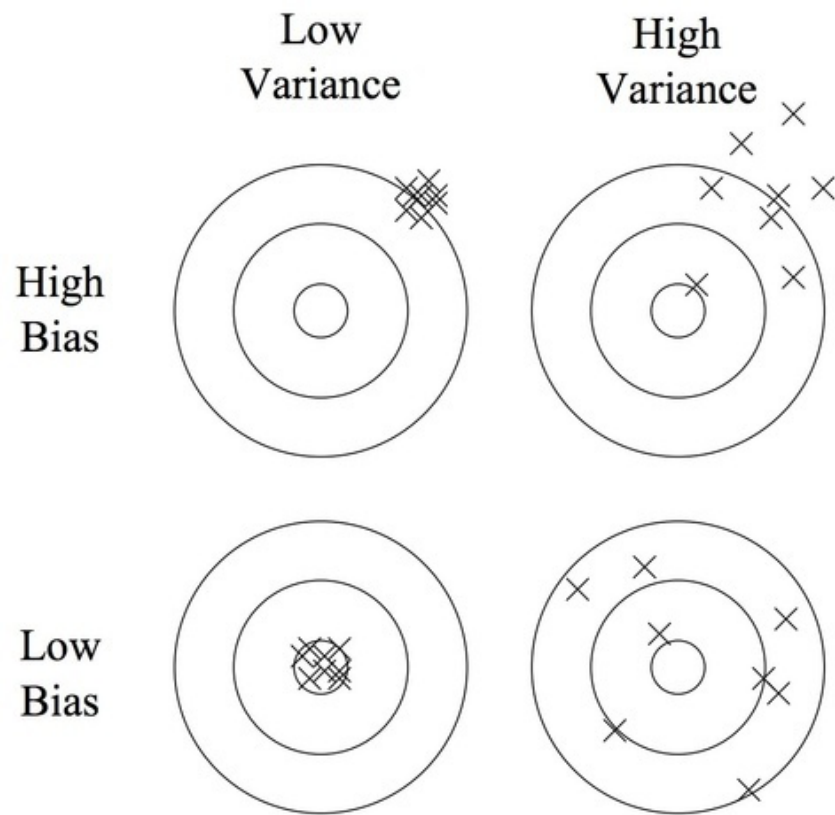
*A: Of course not!*

*A: On its own, not very well.*

**NOTE**

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

# BIAS-VARIANCE

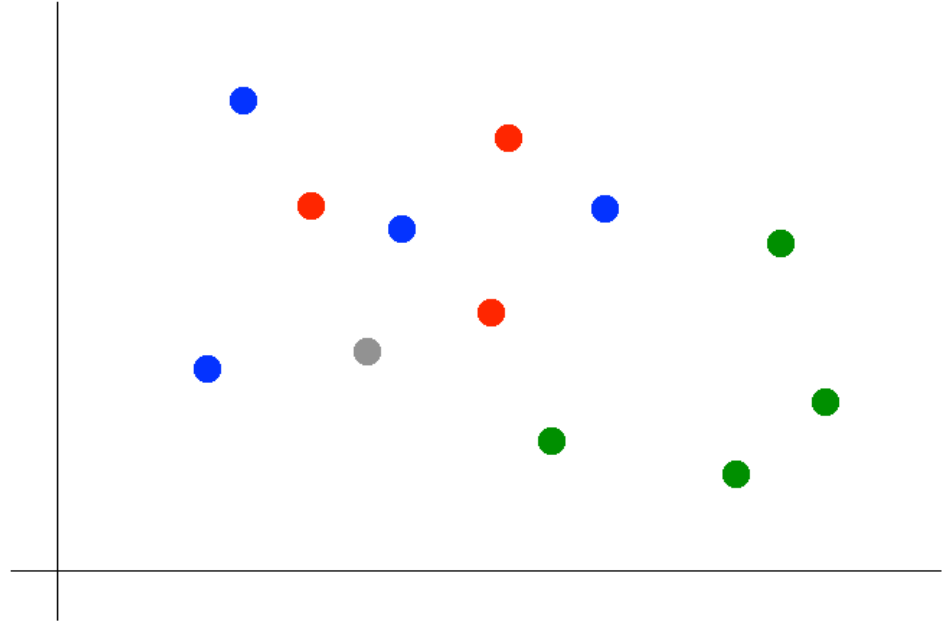




*We can do better than that.... as we will see in the next class....*

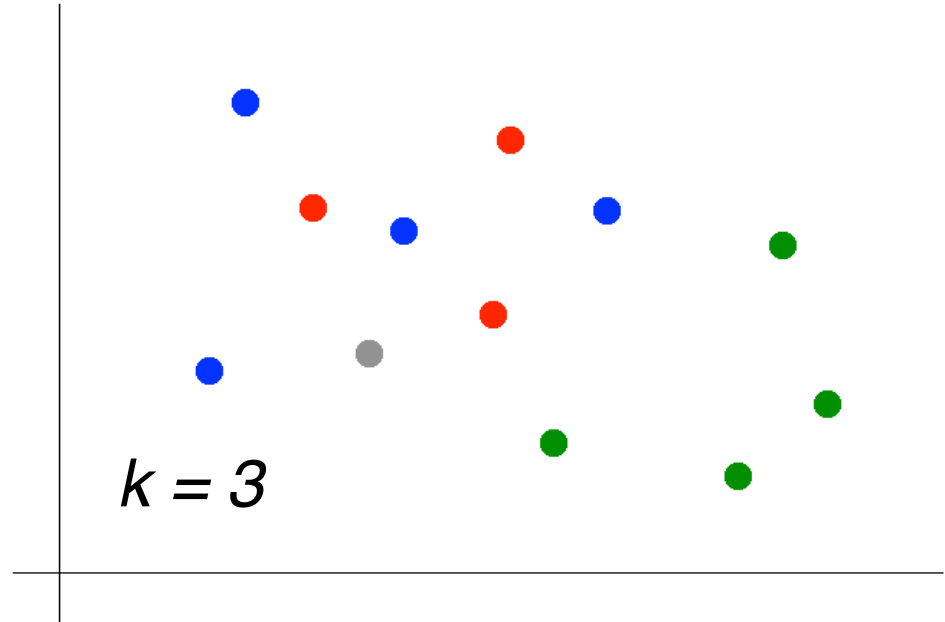
# K-NEAREST NEIGHBORS

*Suppose we want to predict the color of the grey dot*



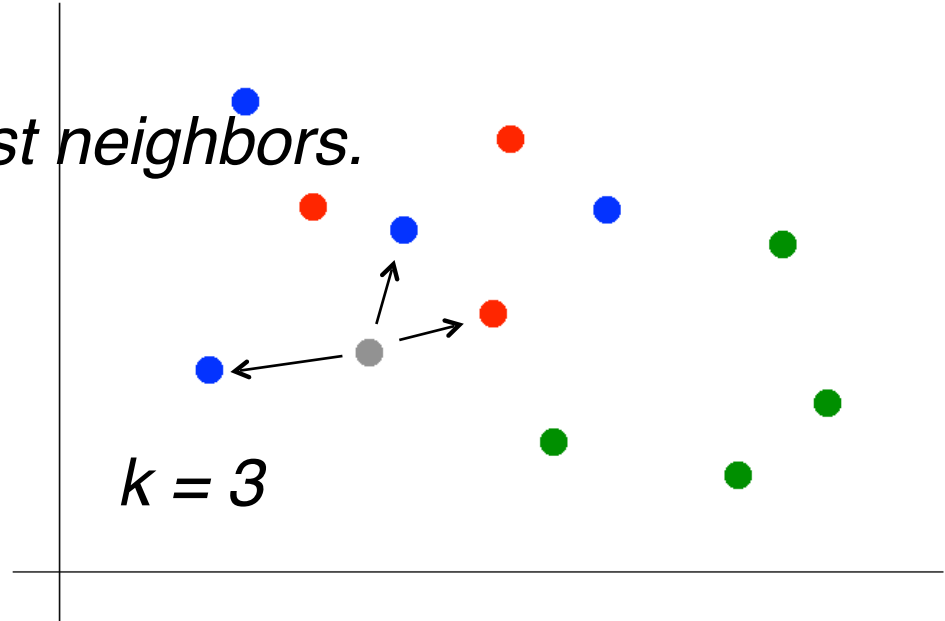
*Suppose we want to predict the color of the grey dot*

*1) Pick a value for  $k$ .*



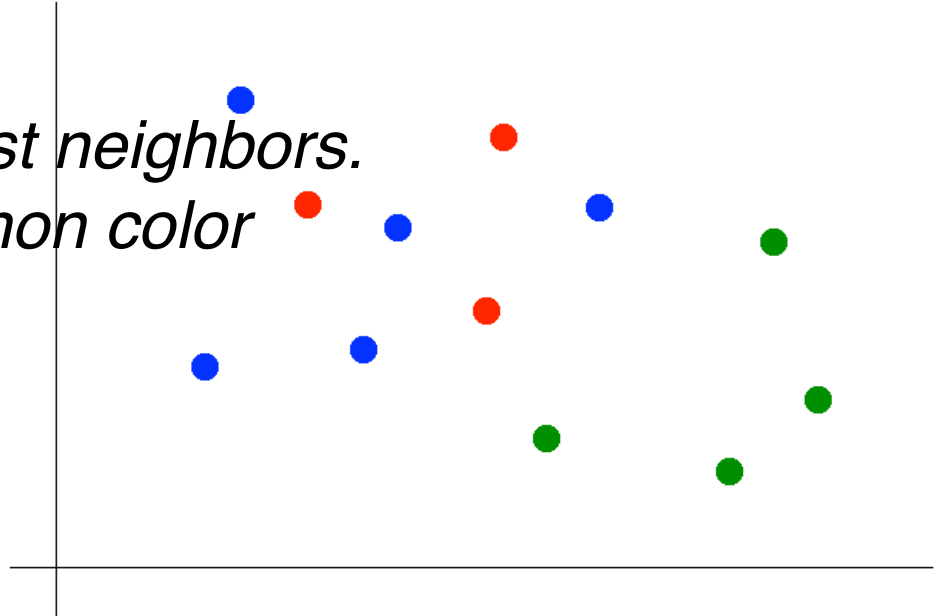
*Suppose we want to predict the color of the grey dot*

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



*Suppose we want to predict the color of the grey dot*

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



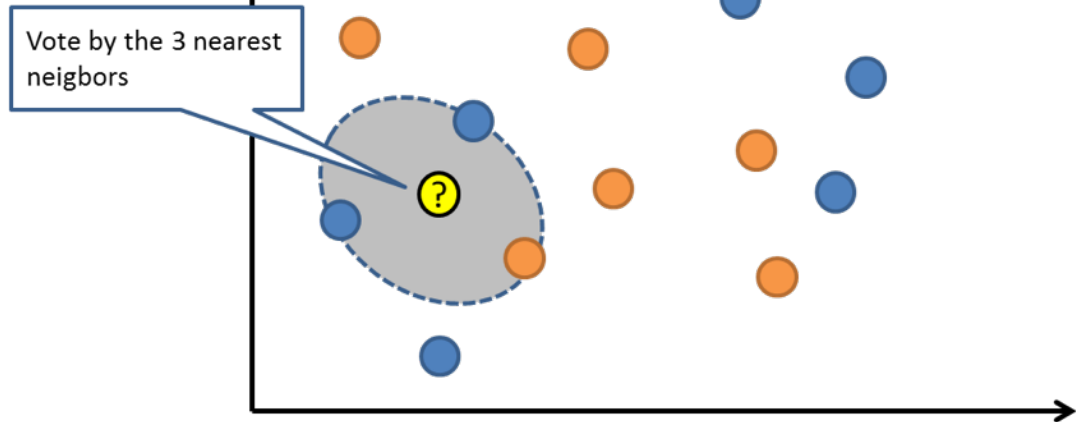
*Suppose we want to predict the color of the grey dot*

- 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**