

# **INTRO to DATA SCIENCE**

## **MACHINE LEARNING / KNN**

## What's big data?

The practical viewpoint:

- ①  $O(n^2)$  algorithm feasible: small data
- ② Fits on one machine: medium data
- ③ Doesn't fit on one machine: big data

**I. WHAT IS MACHINE LEARNING?**

**II. MACHINE LEARNING PROBLEMS**

**III. CLASSIFICATION WITH K NEAREST NEIGHBORS**

# **I. 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](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*.”

“The core of machine learning deals with *representation* and *generalization*...”

source: [http://en.wikipedia.org/wiki/Machine\\_learning](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*.”

“The core of machine learning deals with *representation* and *generalization*...”

- *representation* – extracting structure from data

from Wikipedia:

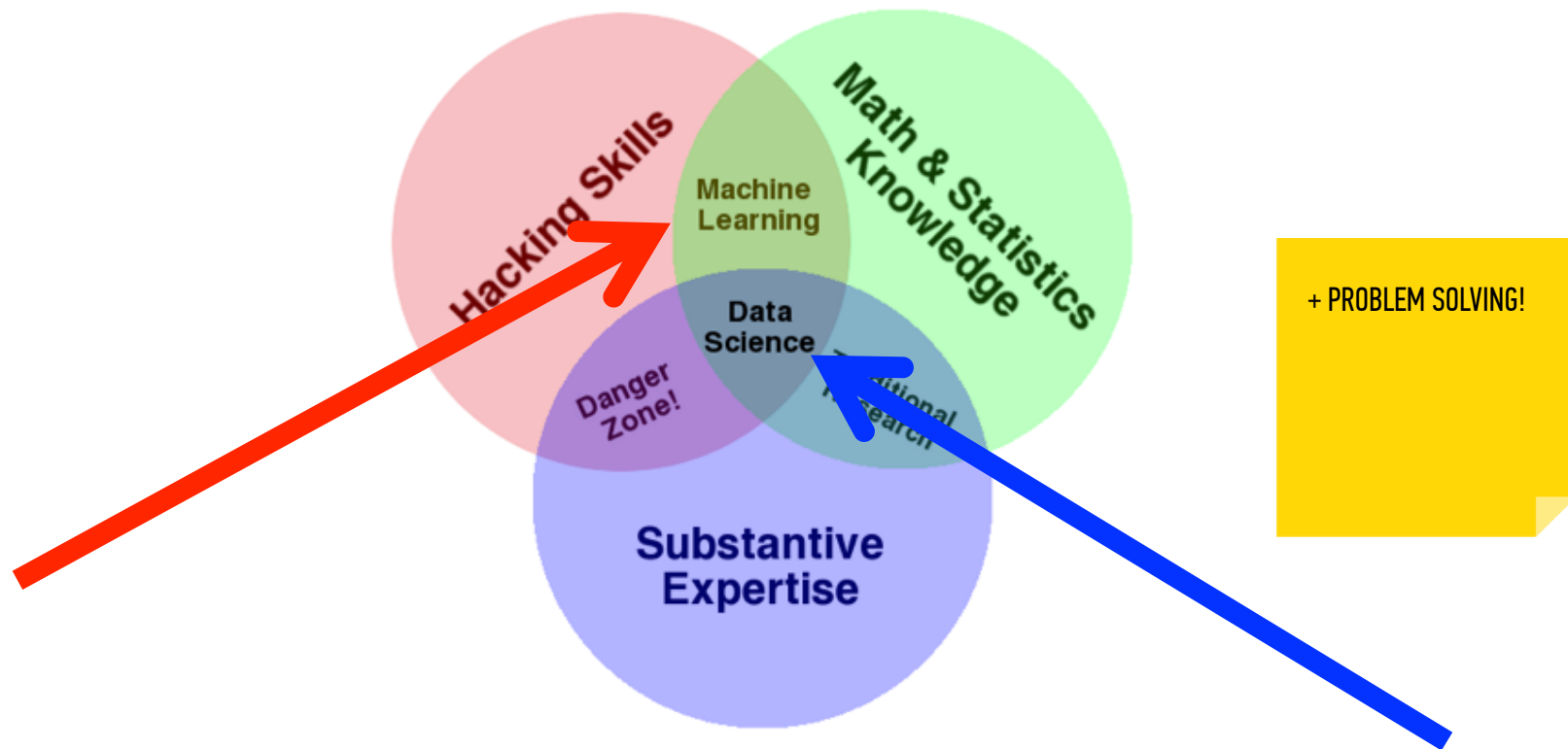
“Machine learning, a branch of artificial intelligence, is about the construction and study of systems that *can learn from data*.”

“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](http://en.wikipedia.org/wiki/Machine_learning)

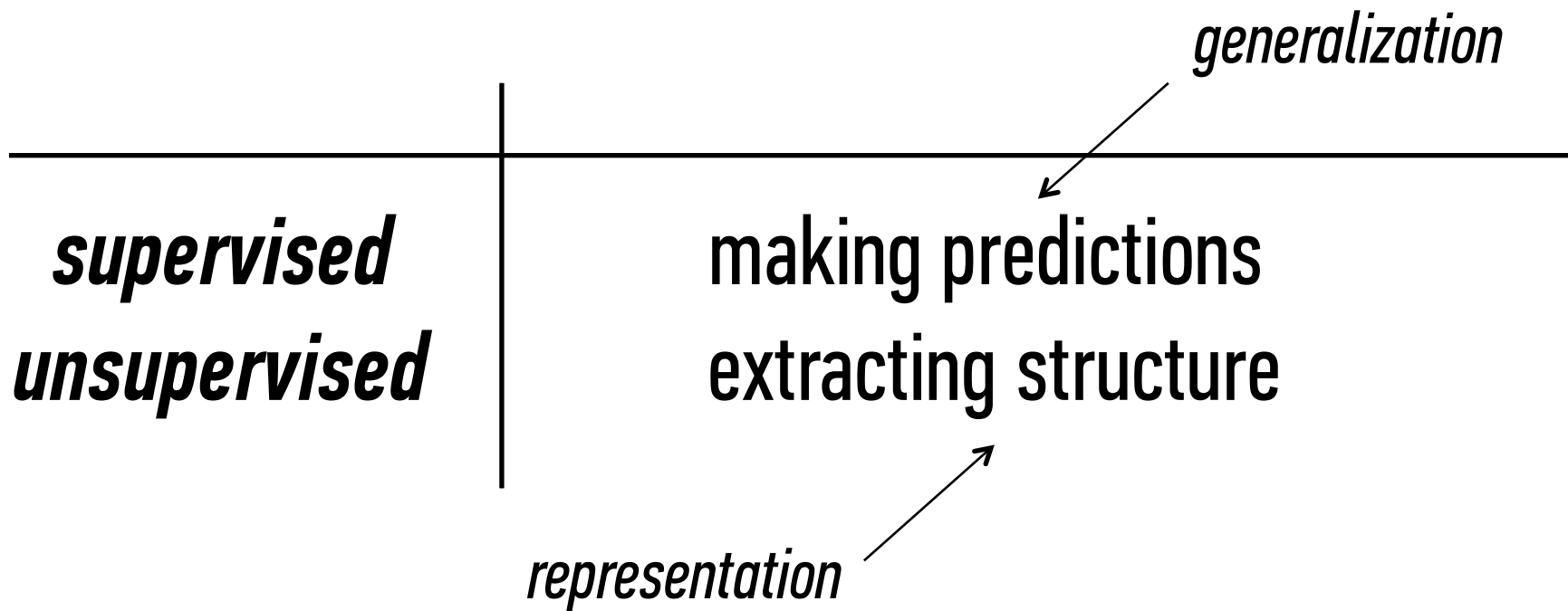


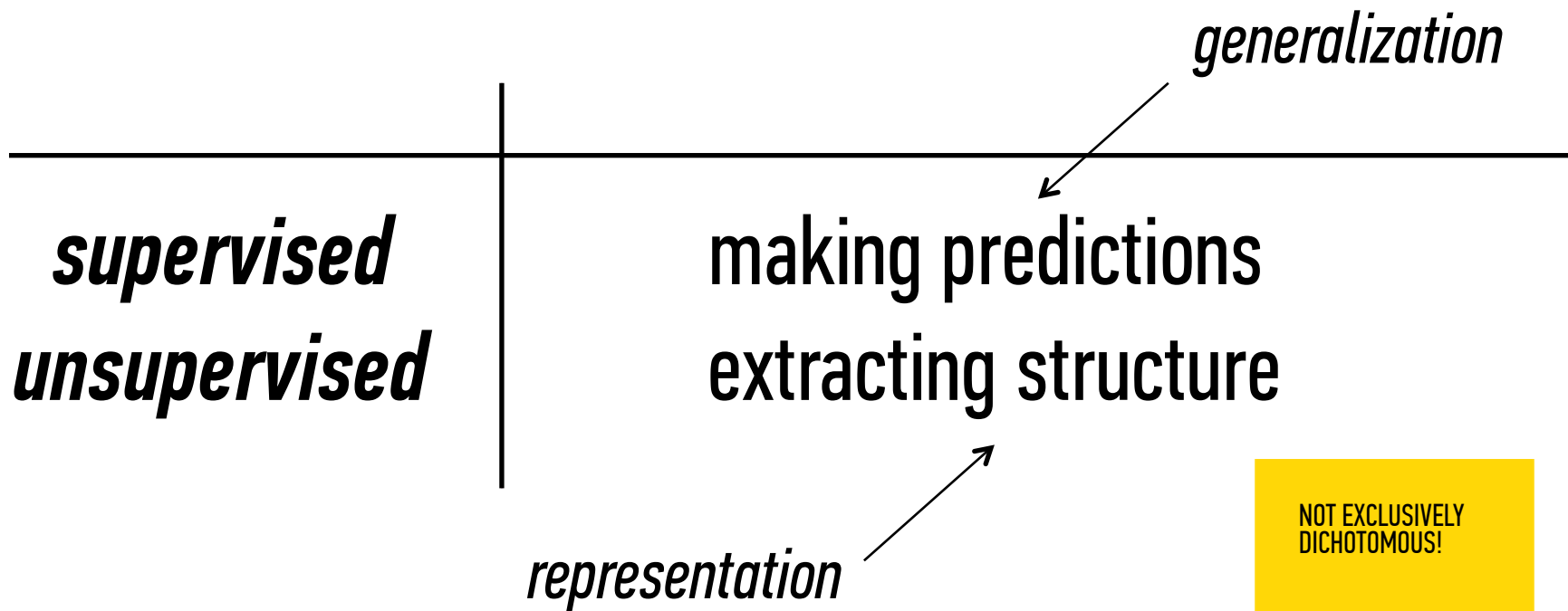


# **II. MACHINE LEARNING PROBLEMS**

---

<i><b>supervised</b></i>	making predictions
<i><b>unsupervised</b></i>	extracting structure





NOT EXCLUSIVELY  
DICHOTOMOUS!

	<b><i>continuous</i></b>	<b><i>categorical</i></b>
	<b>quantitative</b>	<b>qualitative</b>

*continuous*

*categorical*

quantitative

qualitative

## NOTE

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

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

	<i>continuous</i>	<i>categorical</i>
<i>supervised</i>	regression	classification
<i>unsupervised</i>	dimension reduction	clustering



	<i>continuous</i>	<i>categorical</i>
<i>supervised</i>	regression	classification
<i>unsupervised</i>	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***  
***unsupervised***

**making predictions**  
**extracting structure**

Academic goal: make good predictions by some metric.

Practical goal: provide insight and solve problems.

The goal is determined by the type of problem.

---

**QUESTION**

---

***HOW  
DO YOU  
DETERMINE  
THE RIGHT  
APPROACH?***

	<i>continuous</i>	<i>categorical</i>
<i>supervised</i>	regression	classification
<i>unsupervised</i>	dimension reduction	clustering

## ANSWER

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

What type of problem is this?

Music Recommendation



What type of problem is this?

Music Recommendation

*It could be 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  
KNOW  
IF YOUR  
RESULTS ARE GOOD?***

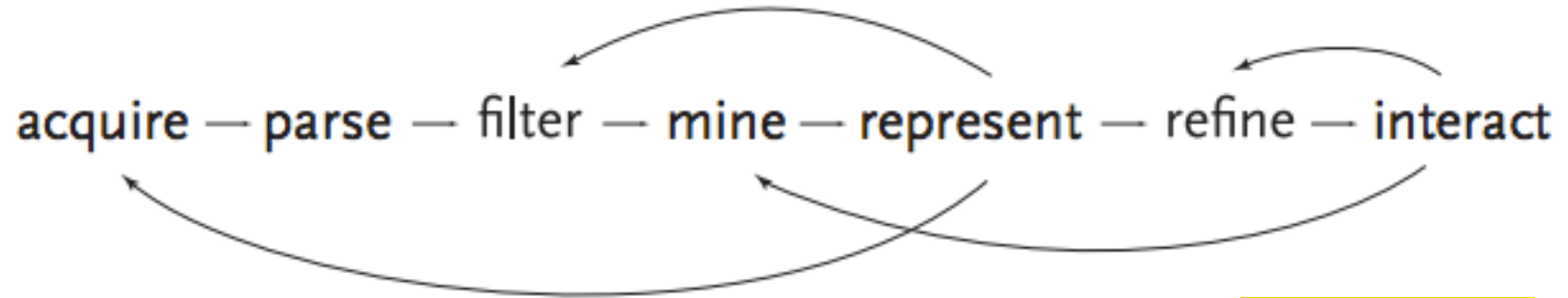
Evaluation metrics and procedures — about which more later.

---

## QUESTION

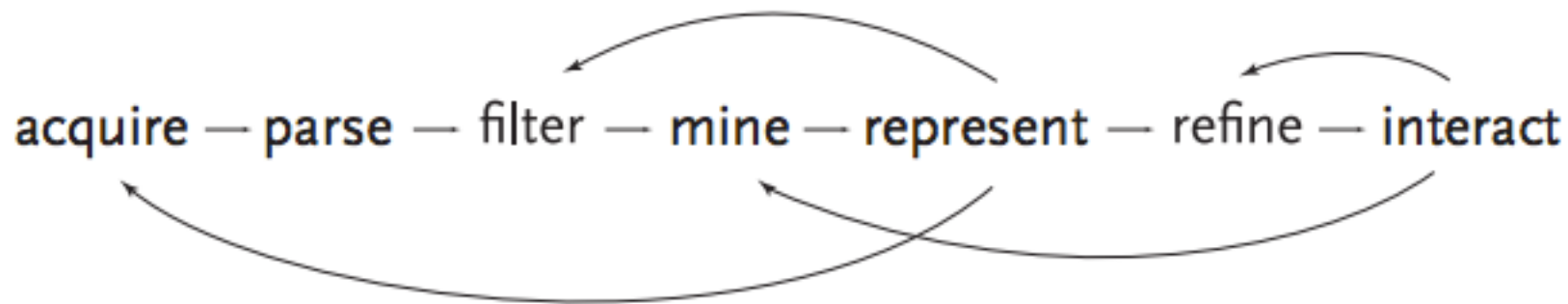
---

***WHAT  
DO YOU  
DO  
WITH YOUR  
RESULTS?***



## ANSWER

Interpret them and  
react accordingly -  
*application.*



ANSWER

NOTE

In:  
re

This also relies on your  
problem solving skills!

# **III.**

# **CLASSIFICATION WITH KNN**

	<i>continuous</i>	<i>categorical</i>
<i>supervised</i>	???	???
<i>unsupervised</i>	???	???



	<i>continuous</i>	<i>categorical</i>
<i>supervised</i>	regression	classification
<i>unsupervised</i>	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>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:

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

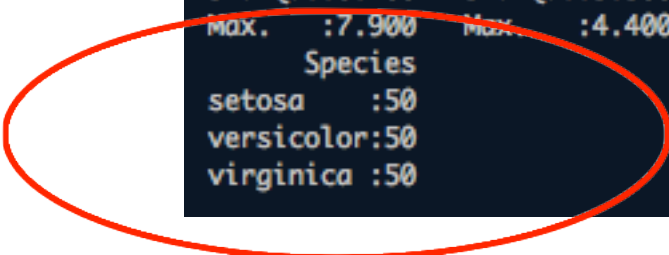
class  
labels  
(categorical)

**Q: What does “supervised” mean?**

Q: What does “supervised” mean?

A: We know the labels.

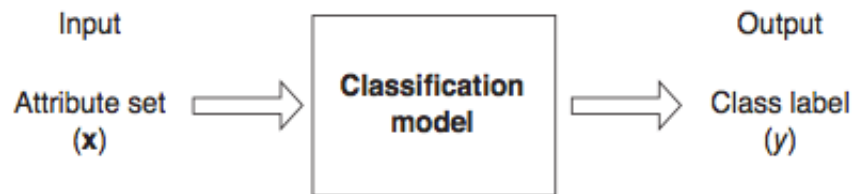
```
Welcome to R! Thu Feb 28 13:07:25 2013
> summary(iris)
  Sepal.Length Sepal.Width Petal.Length  Petal.Width
Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
Median :5.800   Median :3.000   Median :4.350   Median :1.300
Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
 Species
setosa   :50
versicolor:50
virginica :50
```



**Q: How does a classification problem work?**

Q: How does a classification problem work?

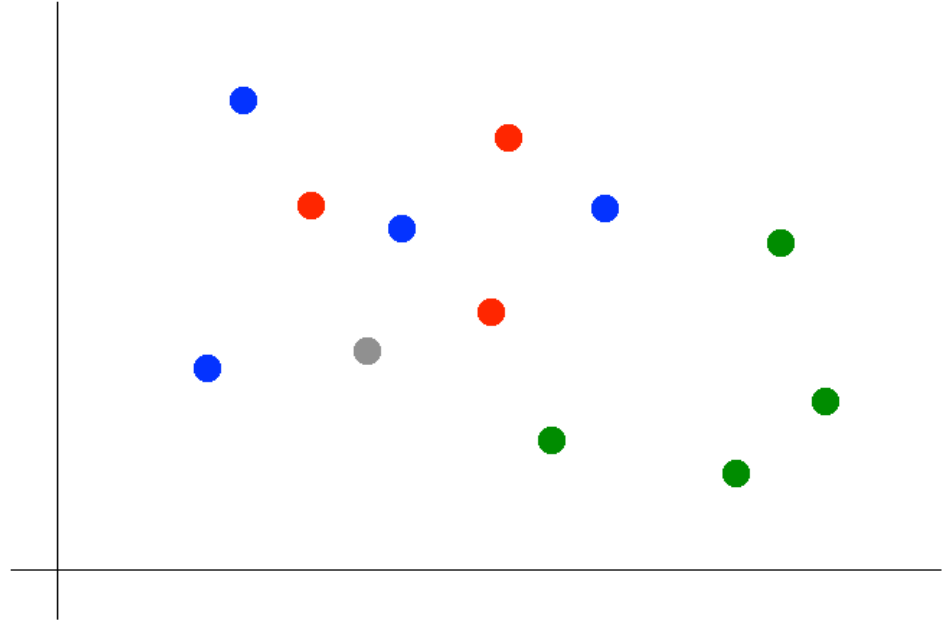
A: Data in, predicted labels out.



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



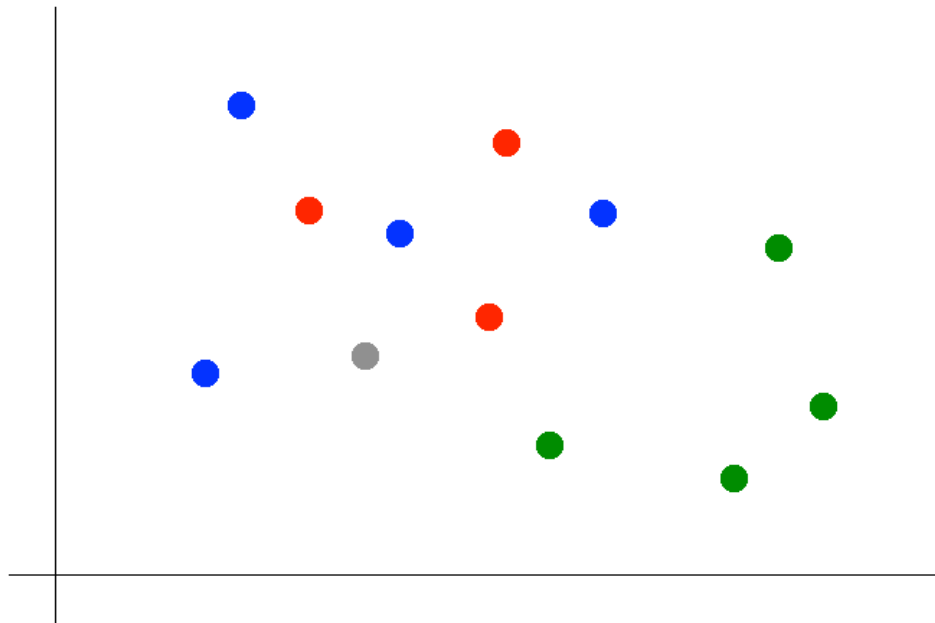
Suppose we want to predict the color of the grey dot.



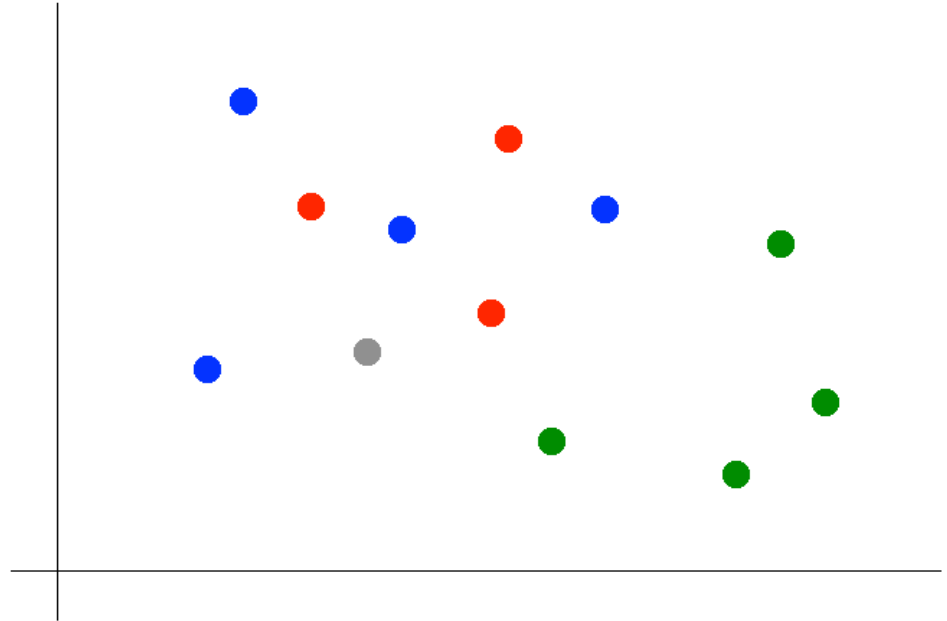
Suppose we want to predict the color of the grey dot.

QUESTION:

What are the features?  
What are the labels?

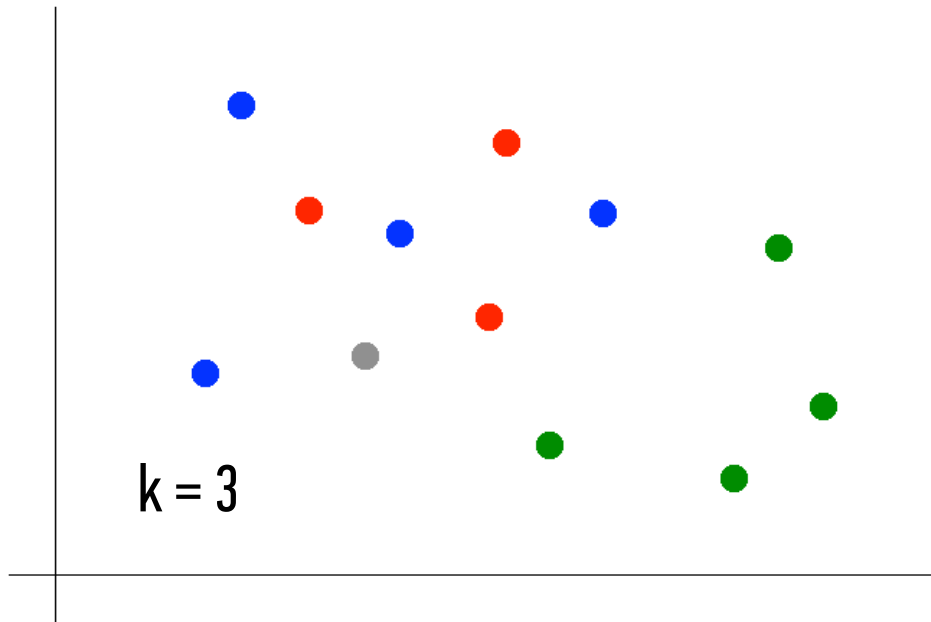


Suppose we want to predict the color of the grey dot.



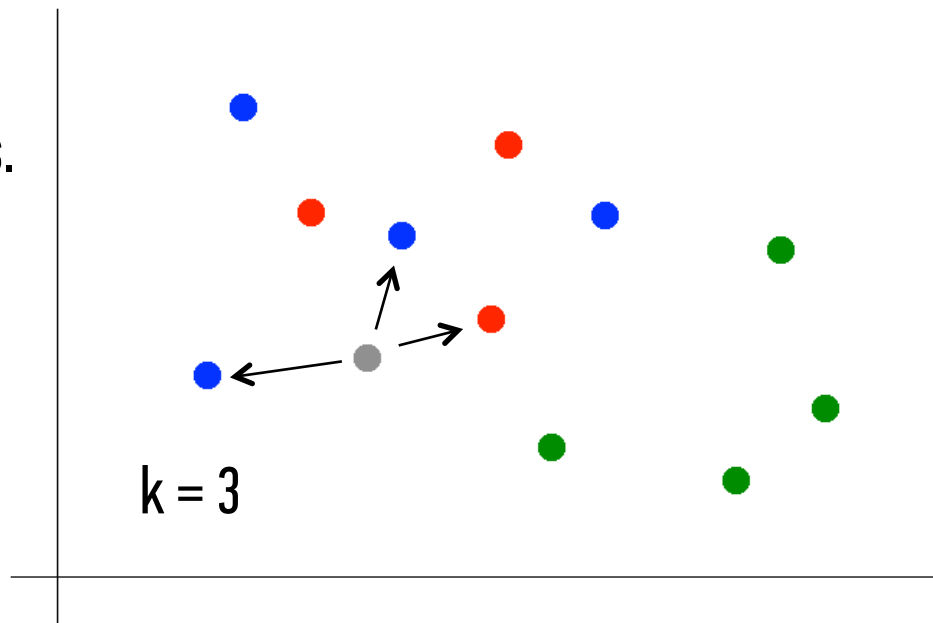
Suppose we want to predict the color of the grey dot.

1) Pick a value for  $k$ .



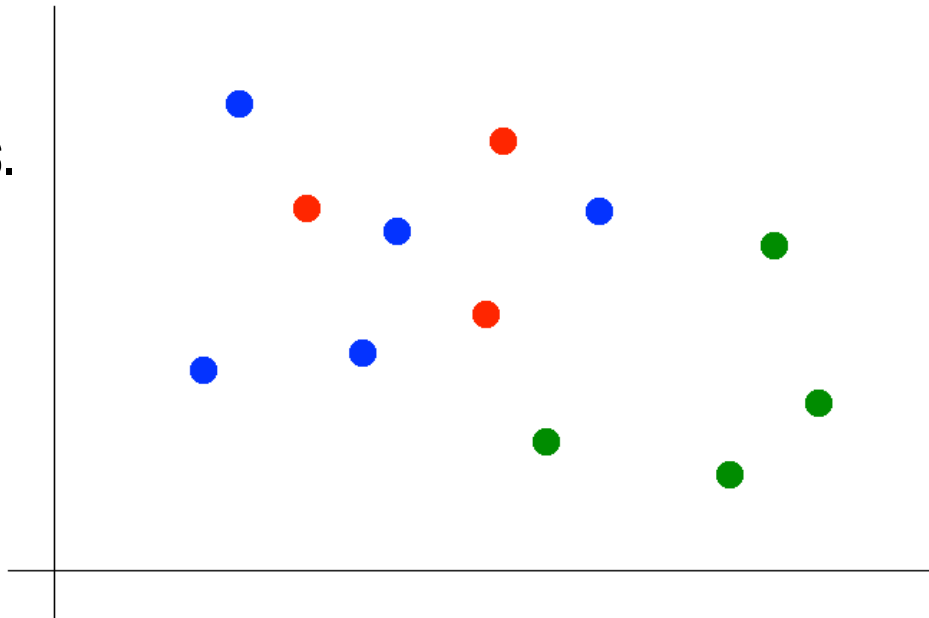
Suppose we want to predict the color of the grey dot.

- 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.

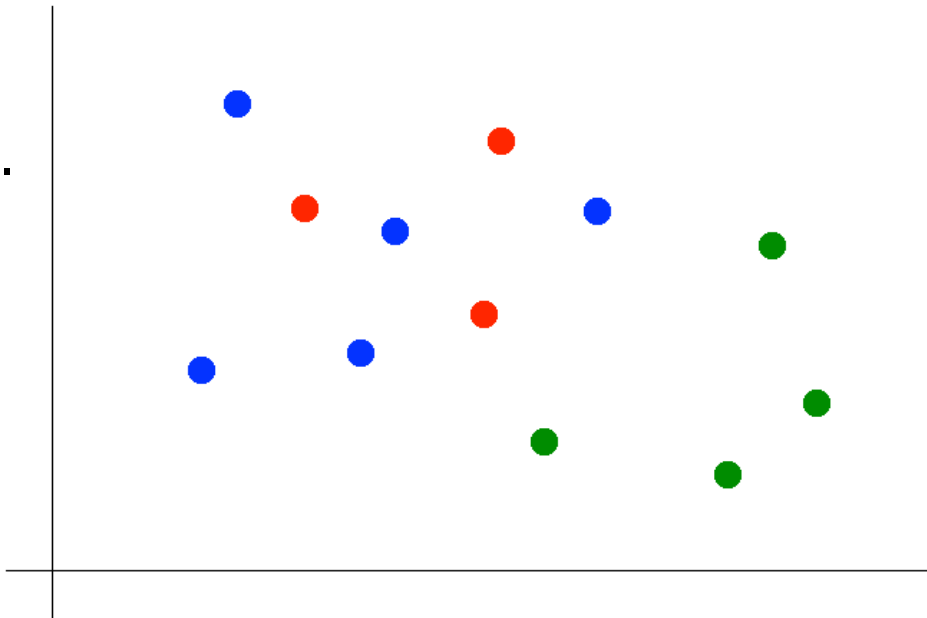


Suppose we want to predict the color of the grey dot.

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

**NOTE:**

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



---

# INTRO TO DATA SCIENCE

---