GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

MACHINE LEARNING / KNN

I. WHAT IS MACHINE LEARNING?

II. MACHINE LEARNING PROBLEMS

III. CLASSIFICATION WITH K NEAREST NEIGHBORS

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

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:

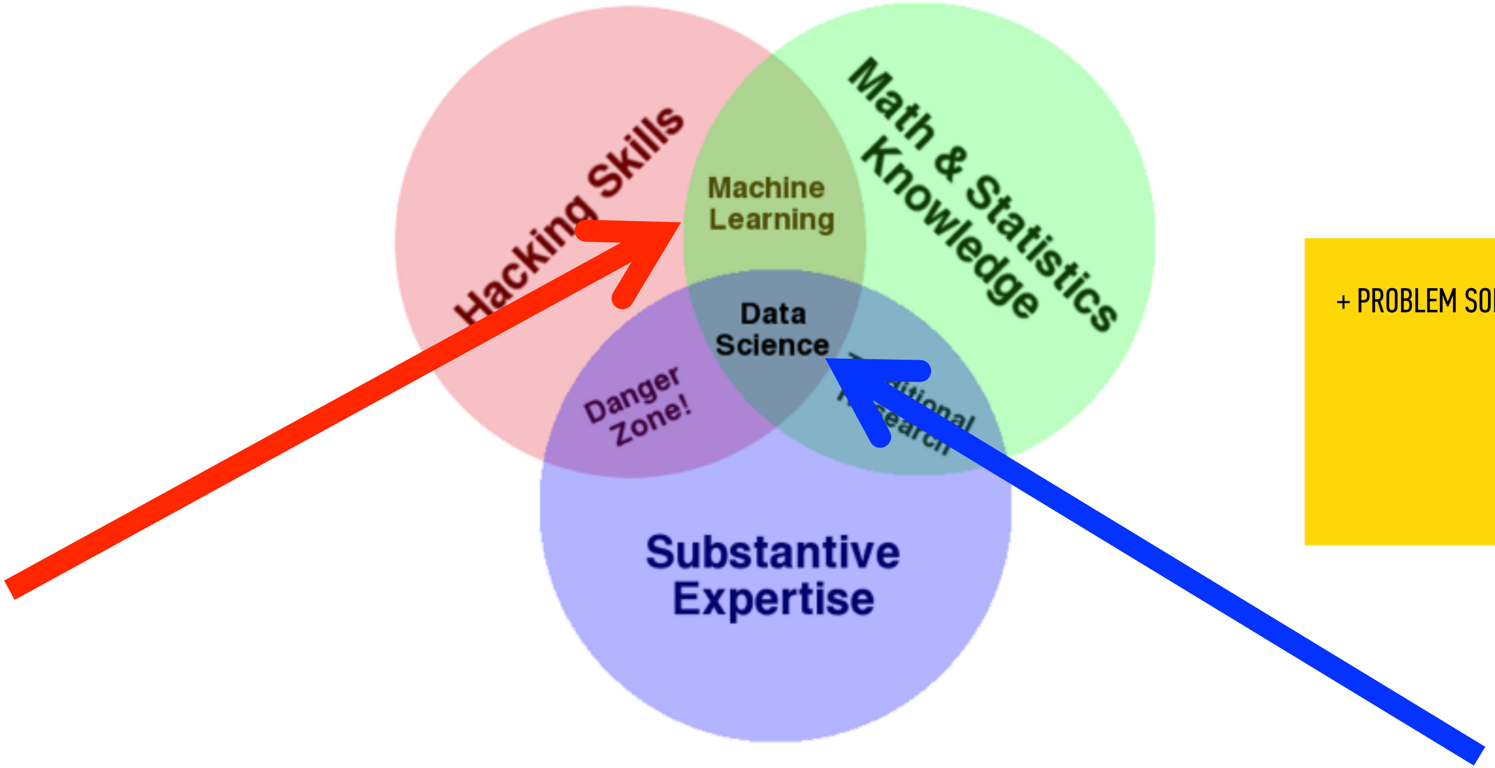
"Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can *learn from data*."

"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

YOU WANT TO GO HERE



+ PROBLEM SOLVING!

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

II. MACHINE LEARNING PROBLEMS

supervised	making predictions
unsupervised	extracting structure

generalization

supervised	making predictions
unsupervised	extracting structure

representation

generalization

supervised	making predictions
unsupervised	extracting structure

representation

NOT EXCLUSIVELY
DICHOTOMOUS!

	continuous	***categorical***
	quantitative	qualitative

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

	continuous	***categorical***
	quantitative	qualitative

NOTE

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

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

	continuous	*categorical*
supervised	regression	classification
unsupervised	dimension reduction	clustering

	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

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?

	continuous	*categorical*
supervised	regression	classification
unsupervised	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 YOU’RE
DOING WELL?

supervised	*making predictions*
unsupervised	*extracting structure*

supervised	*test out your predictions*

supervised	*test out your predictions*
unsupervised	*...*

supervised	*test out your predictions*
unsupervised	...

ALSO

There may be external sources of feedback, for example conversion rates in production systems.

TWO GENERAL THINGS WE WILL EMPHASIZE

Three decisions must be made when deciding on a machine learning method:

1. Model.
2. Method of fitting the model.
3. Validation method.

Three decisions must be made when deciding on a machine learning method. For HW1:

1. Model.

   Linear equation.

2. Method of fitting the model.

   Minimize the sum of squared residuals.

3. Validation method.

   Visually graph the fitted model.

- The **bias-variance tradeoff** is a way of intuitively comparing models.

- There is “**no free lunch**” – if a classifier performs well on some problems, it will not perform well on other problems.

- It conceptually explains why more complex models do not necessarily yield better results.

A model is bad because:

- Not accurate -- doesn’t match data well, or
- Not precise -- lots of variation in results, or
- Data has inherent irreducible error.

A model is bad because:

bias

- Not accurate -- doesn’t match data well, or

variance

- Not precise -- lots of variation in results, or

- Data has inherent irreducible error.

Bias: What is the average error of our predictor?

Variance: Given a different training set, how much would our predictor vary?

Bias: What is the average error of our predictor?

Variance: Given a different training set, how much would our predictor vary?

In general, if a model is:

complex -> **low bias** and **high variance**.

simple -> **high bias** and **low variance.**

Suppose our error measure is the expected value of the squared "error".

Suppose y is the target value and f^{hat} is the predicted y (a function of the features). Then, manipulation shows that:

$$\mathrm{E}\big[(y - \hat{f})^2\big] = \sigma^2 + \mathrm{Var}[\hat{f}] + \mathrm{Bias}[\hat{f}]^2$$

squared error *irreducible error*

III. CLASSIFICATION WITH KNN

	continuous	*categorical*
supervised	???	???
unsupervised	???	???

- Sounds scary, but we work with it every day!
- Here is some 5-dimensional data:

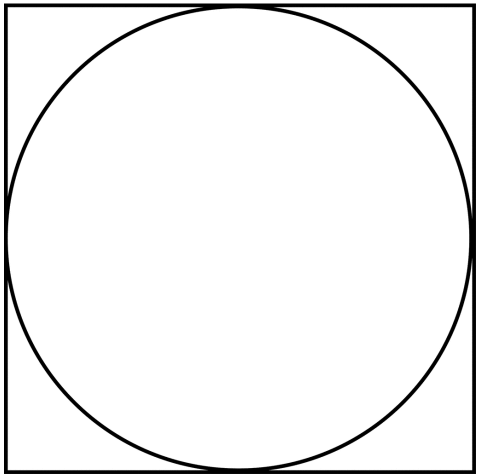
Fisher's *Iris* Data

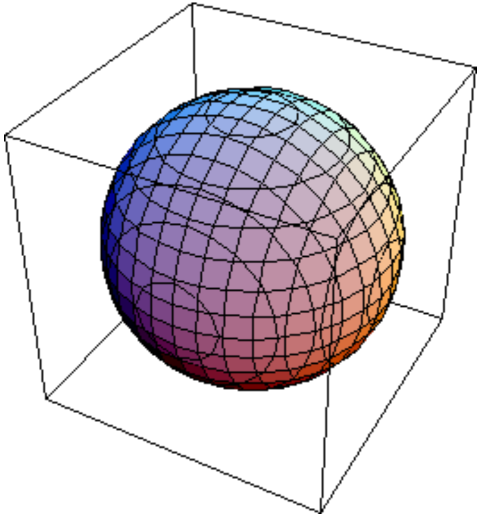
Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
4.6	3.4	1.4	0.3	*I. setosa*
5.0	3.4	1.5	0.2	*I. setosa*

- Sounds scary, but we work with it every day!
- Here is some 4-dimensional data with a target:

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*





$$x^2 < R^2$$

1-dimensional

$$x^2 + y^2 < R^2$$

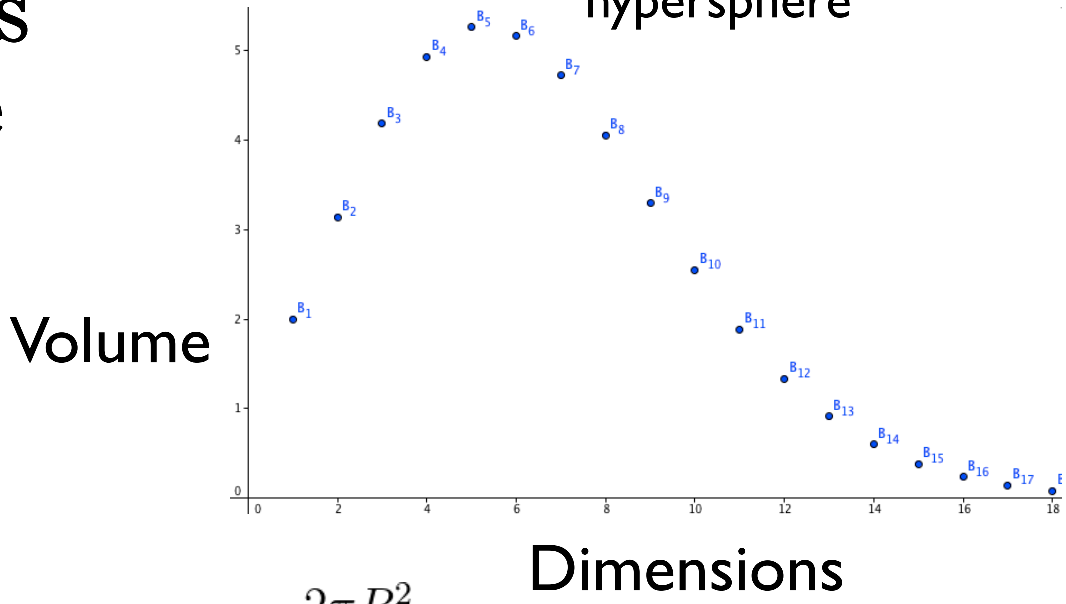
2-dimensional

$$x^2 + y^2 + z^2 < R^2$$

3-dimensional

- As the number of dimensions increases with fixed radius, the hypersphere volume approaches zero.

Volume of n-dimensional hypersphere



$$V_1(R) = 2R,\ V_2(R) = \pi R^2,\ \text{and}\ V_n(R) = \frac{2\pi R^2}{n} V_{n-2}(R),\ \text{for}\ n \geq 3.$$

Source: http://divisbyzero.com/2010/05/09/volumes-of-n-dimensional-balls/

	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 *(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.

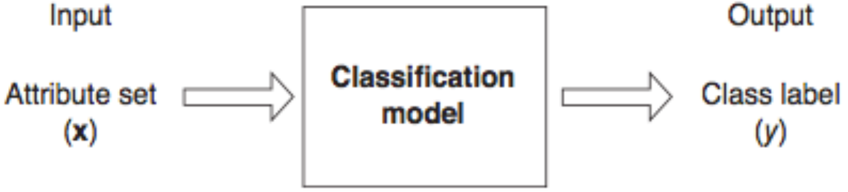
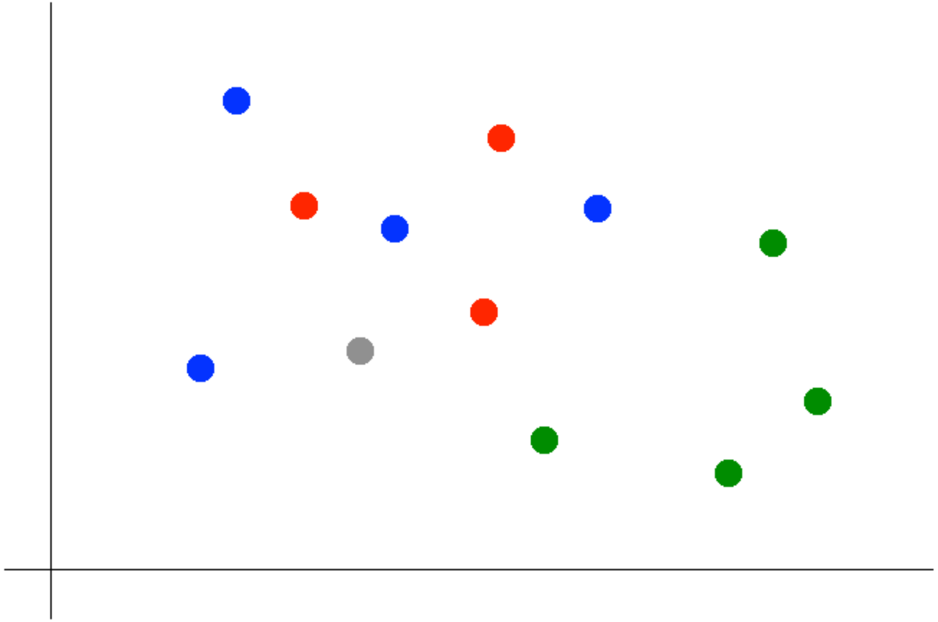


Figure 4.2. Classification as the task of mapping an input attribute set $\mathbf{x}$ into its class label y.

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

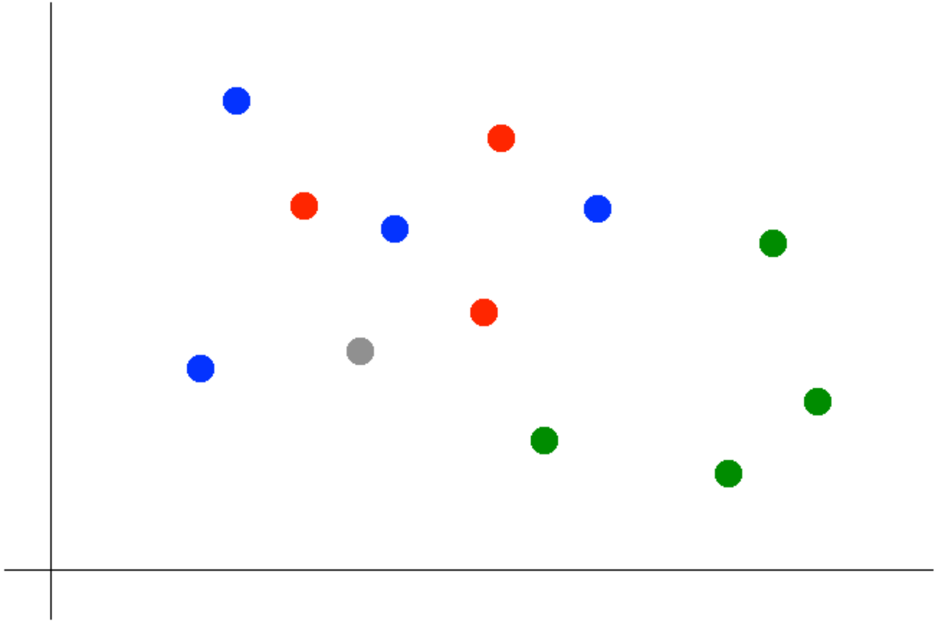
Suppose we want to predict the color of the grey dot.



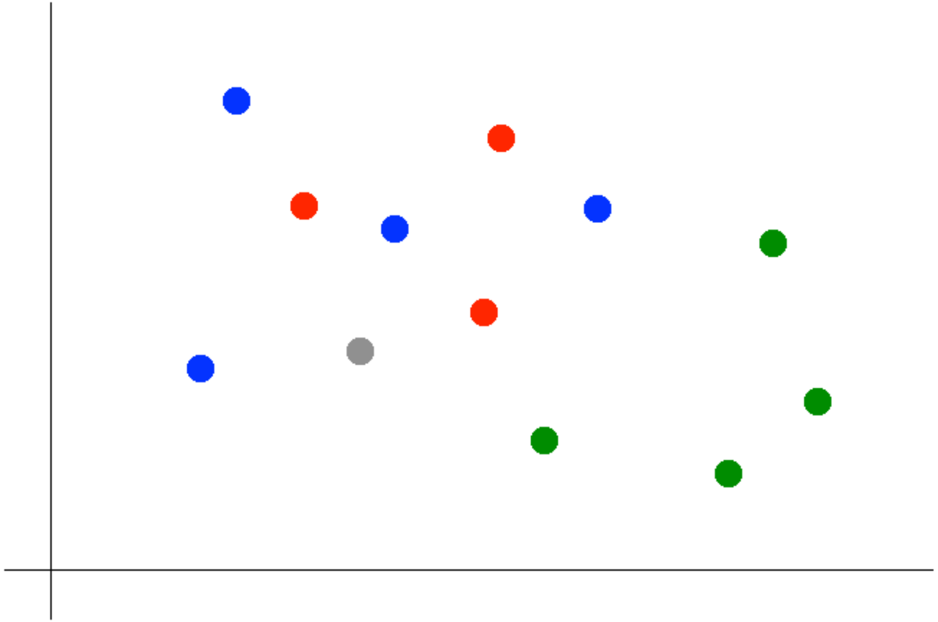
Suppose we want to predict the color of the grey dot.

QUESTION:

What are the features?
What are the labels?

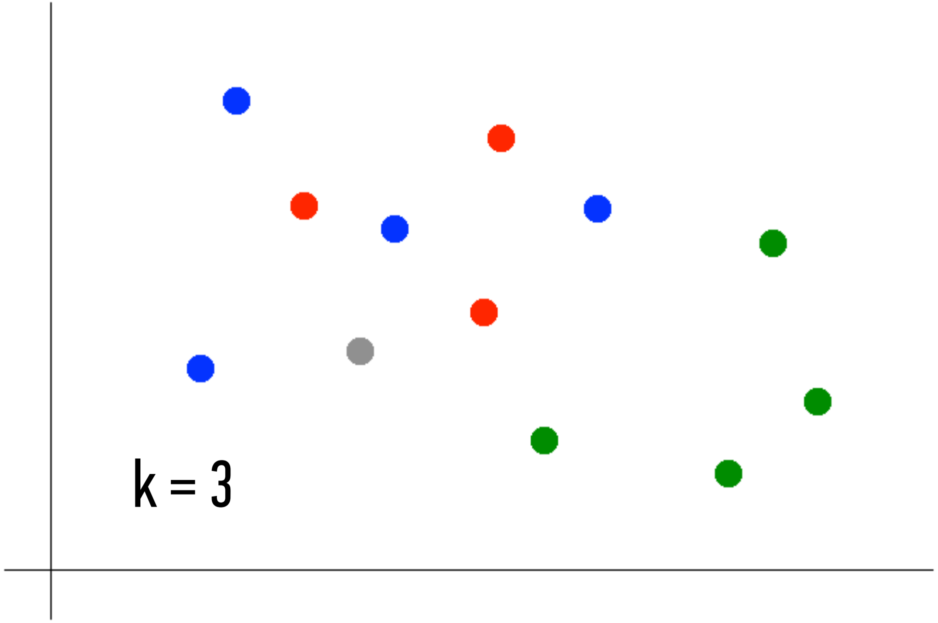


Suppose we want to predict the color of the grey dot.



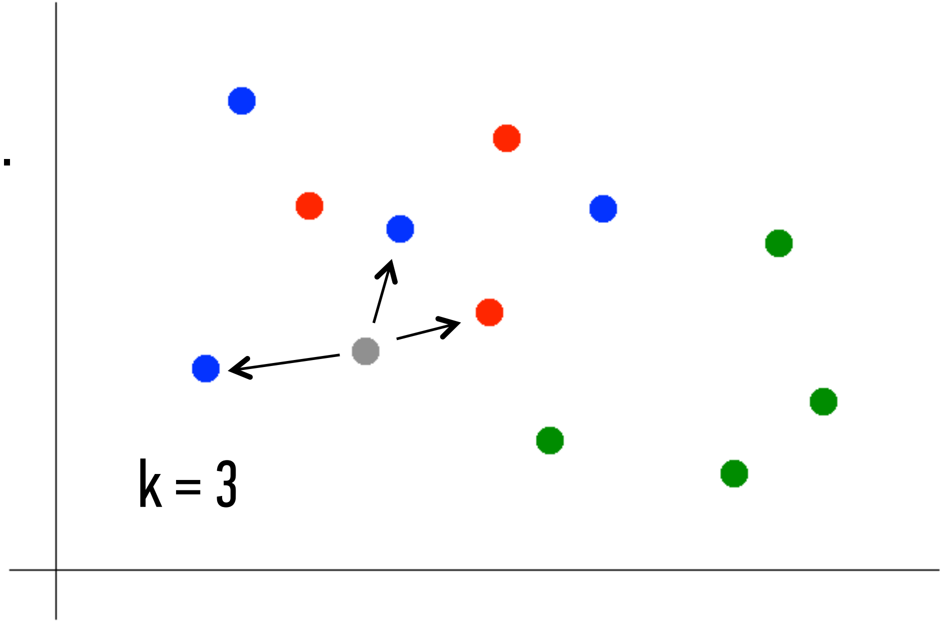
Suppose we want to predict the color of the grey dot.

1) Pick a value for k.



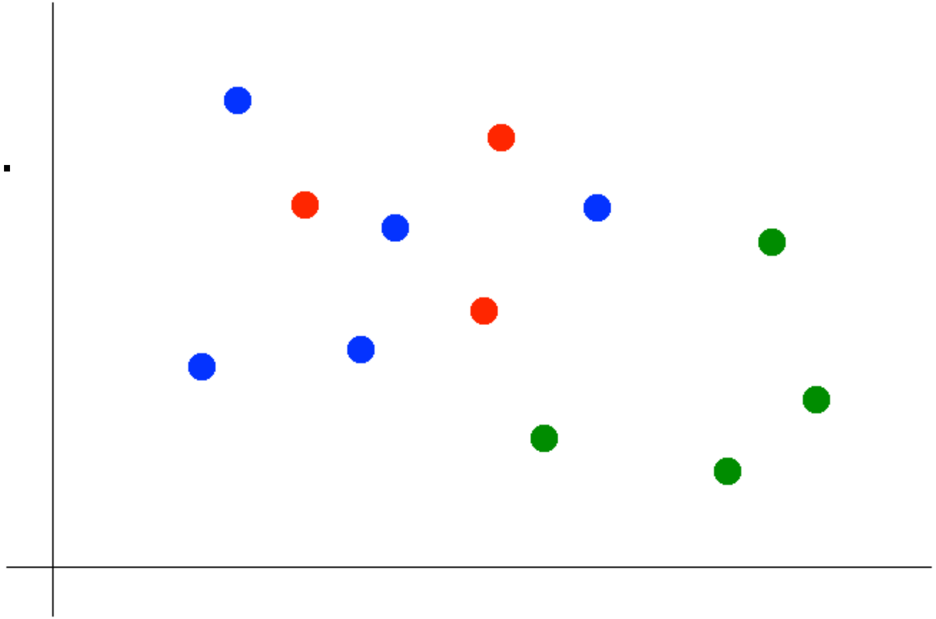
Suppose we want to predict the color of the grey dot.

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



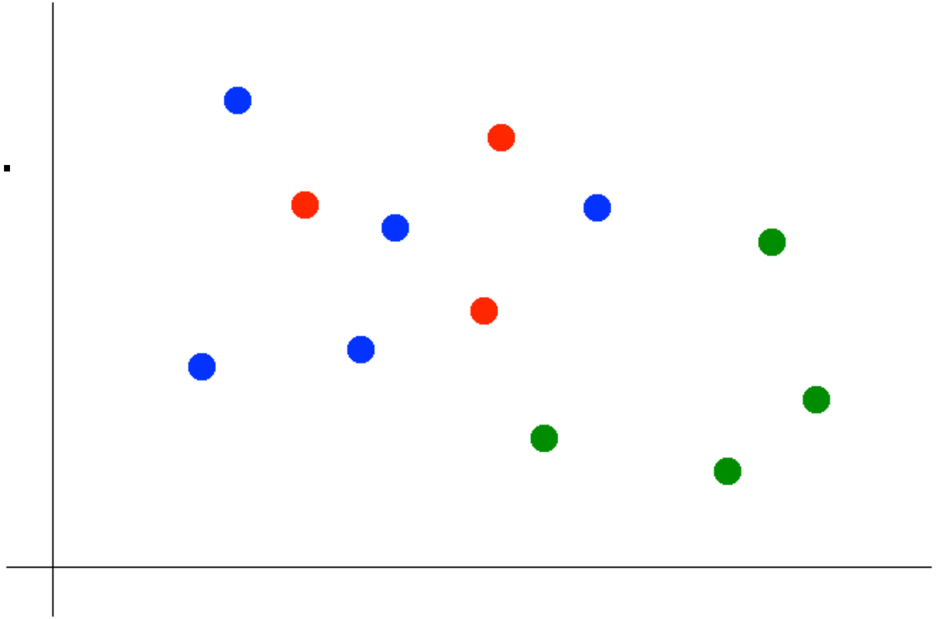
Suppose we want to predict the color of the grey dot.

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.*