GENERAL ASSEMBLY

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

source: http://people.cs.umass.edu/~mcgregor/stocworkshop/langford.pdf

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

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:

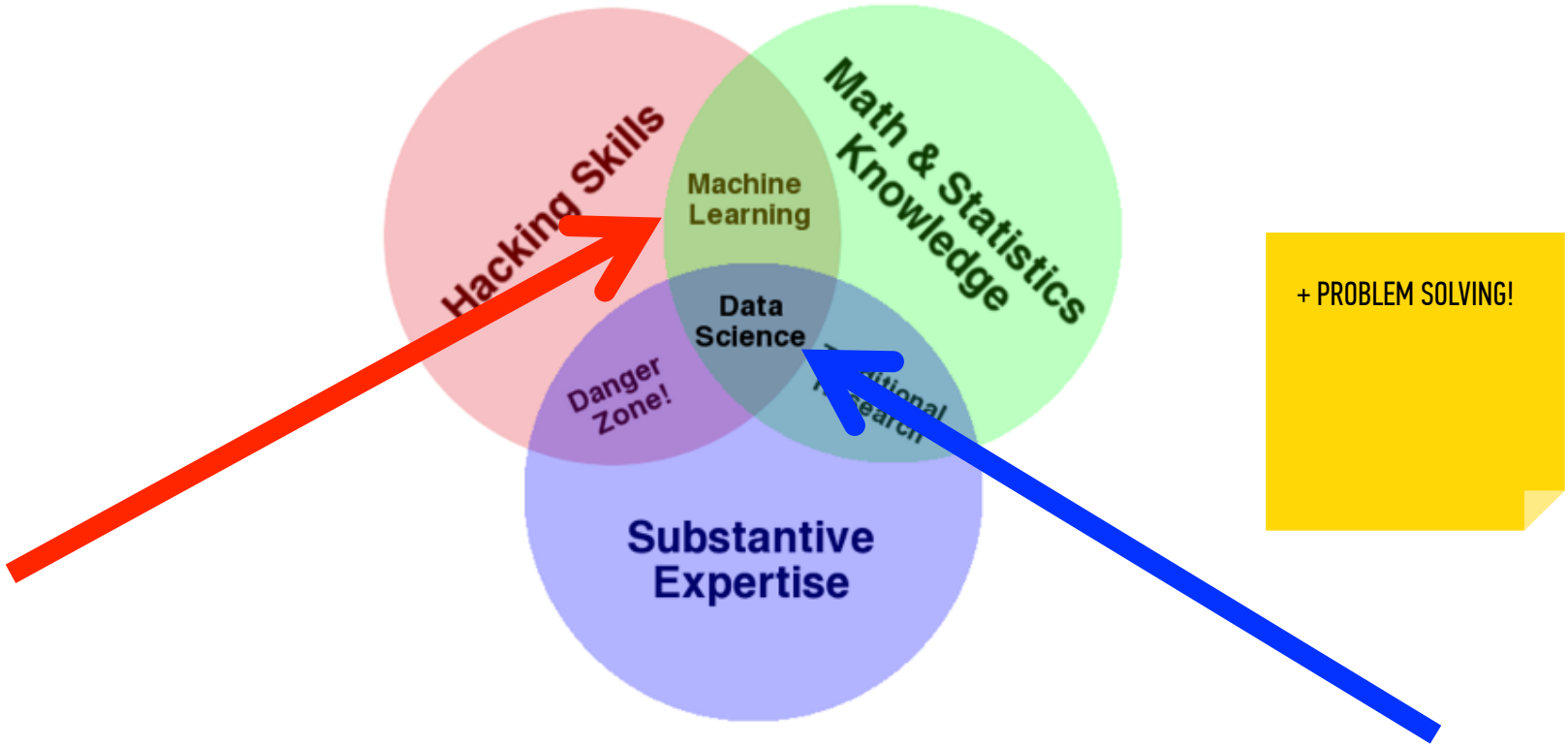
“Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can *learn from data*.”

“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

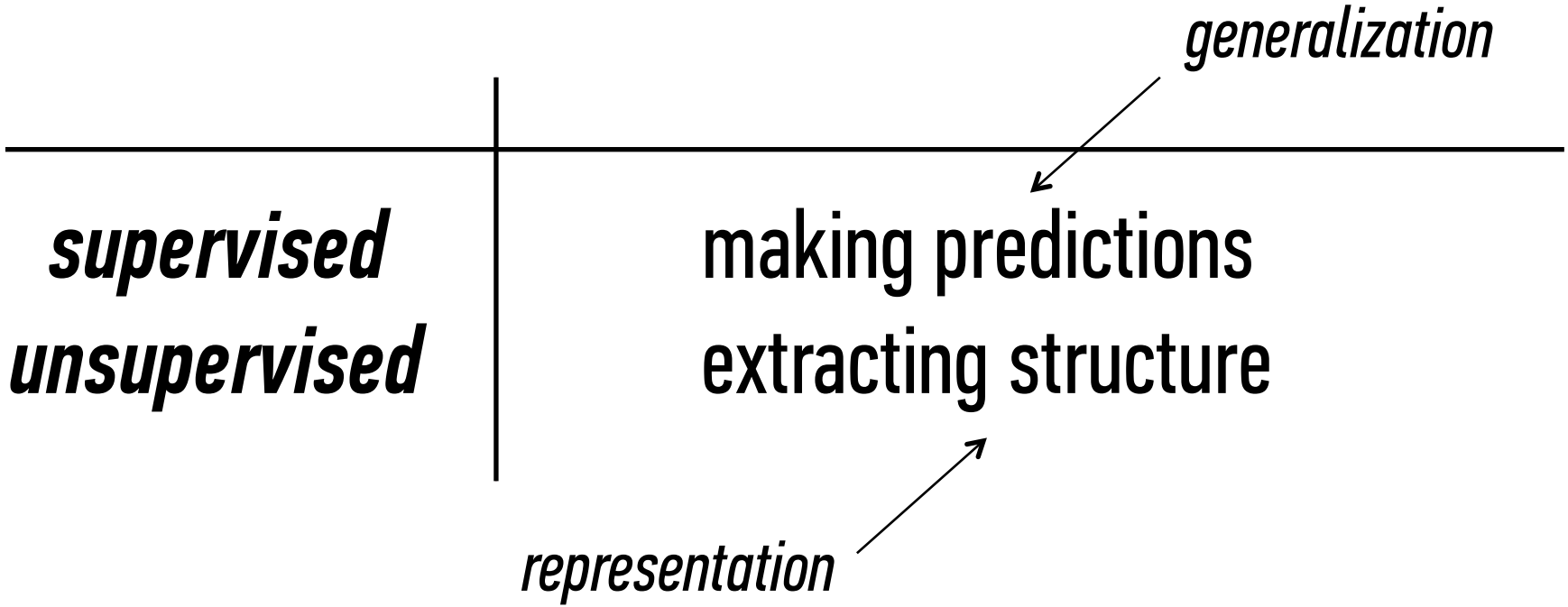
YOU WANT TO GO HERE

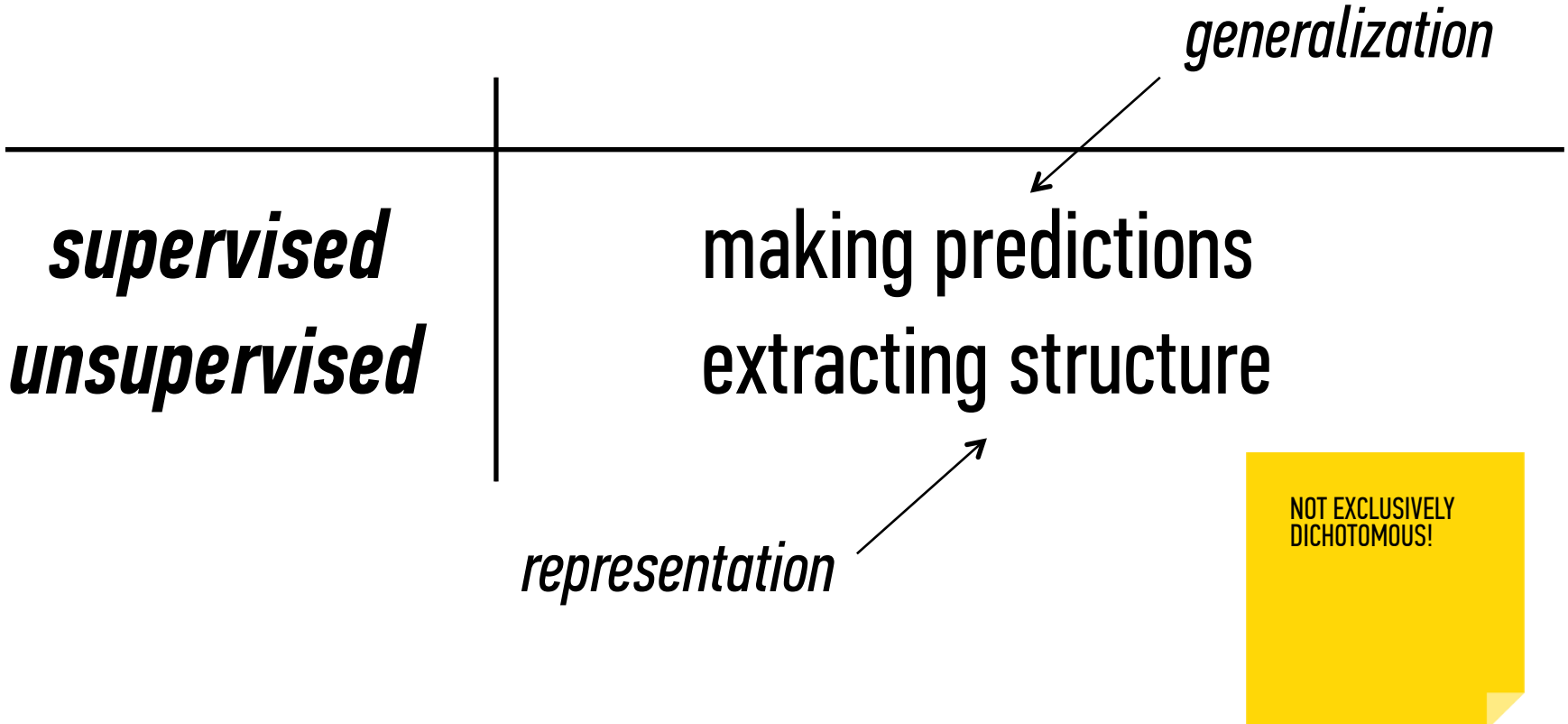


source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

II. MACHINE LEARNING PROBLEMS

supervised	making predictions
unsupervised	extracting structure


generalization
supervised
unsupervised
making predictions
extracting structure
representation


generalization
supervised
unsupervised
making predictions
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	*...*

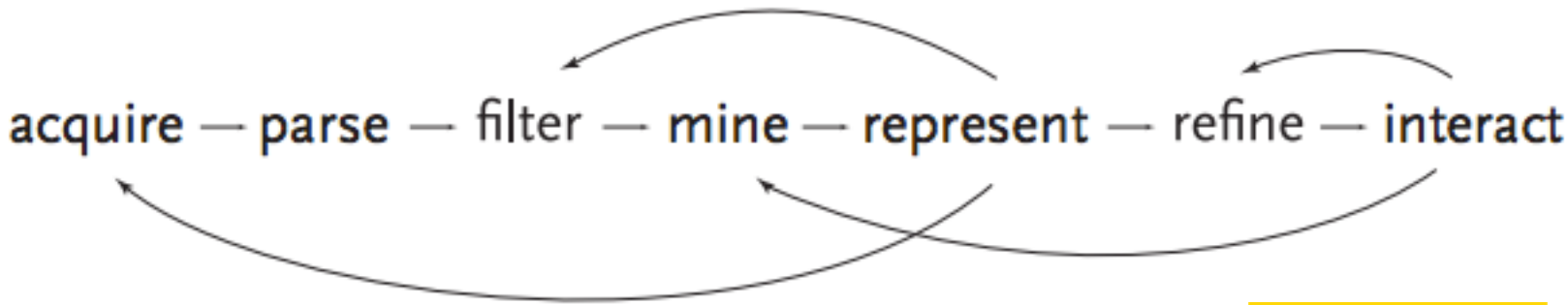
supervised *unsupervised*	*test out your predictions* *...*

ALSO

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

QUESTION

WHAT DO YOU DO WITH YOUR RESULTS?

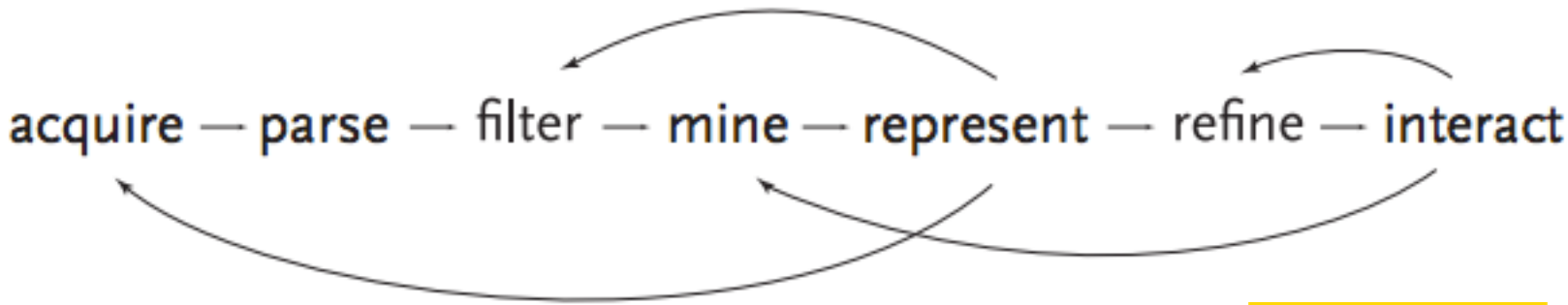


ANSWER

Interpret them and react accordingly - *application*.

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

THE DATA SCIENCE WORKFLOW



ANSWER

In
re

NOTE

This also relies on your problem solving skills!

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

III.
CLASSIFICATION WITH KNN

	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:

independent variables

Fisher's *Iris* Data

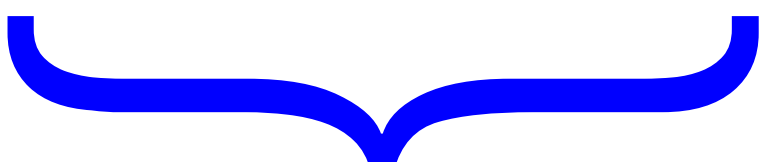
Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
4.6	3.4	1.4	0.3	*I. setosa*
5.0	3.4	1.5	0.2	*I. setosa*

Here's (part of) an example dataset:

independent variables

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*

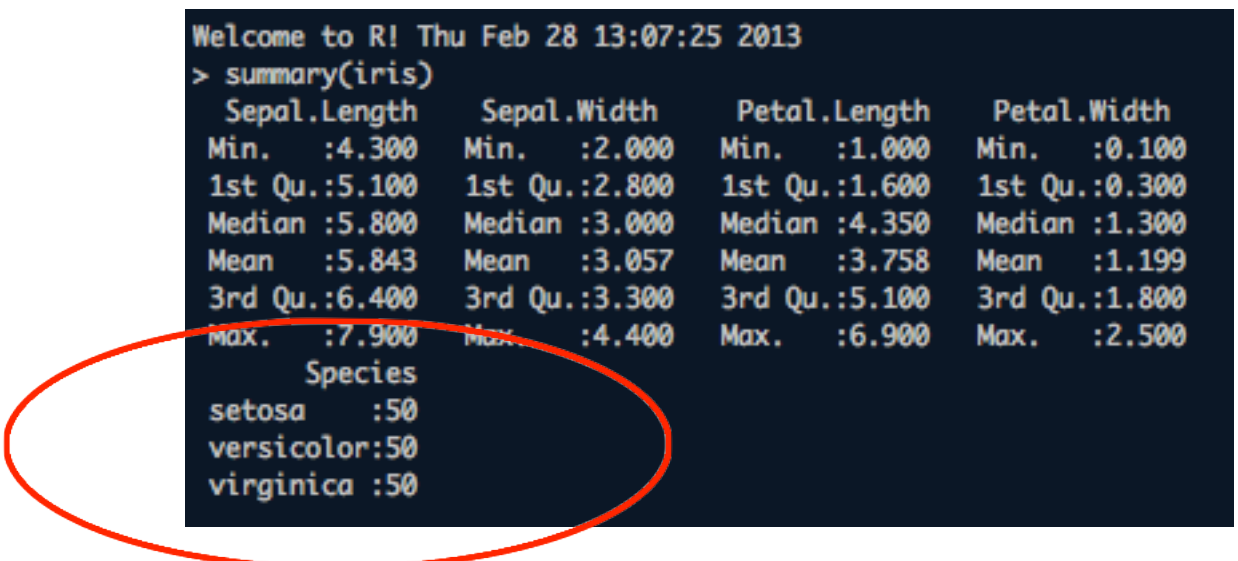


class labels *(categorical)*

Q: What does “supervised” mean?

Q: What does “supervised” mean?

A: We know the labels.



Q: How does a classification problem work?

Q: How does a classification problem work?

A: Data in, predicted labels out.

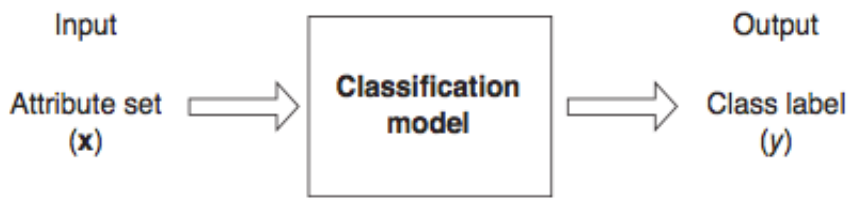
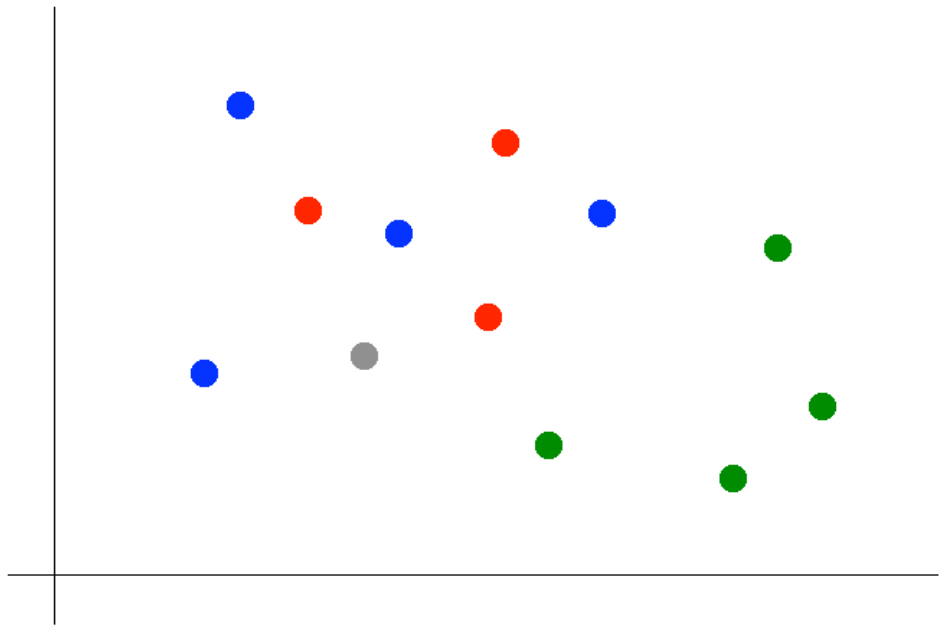


Figure 4.2. Classification as the task of mapping an input attribute set $\mathbf{x}$ into its class label y.

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

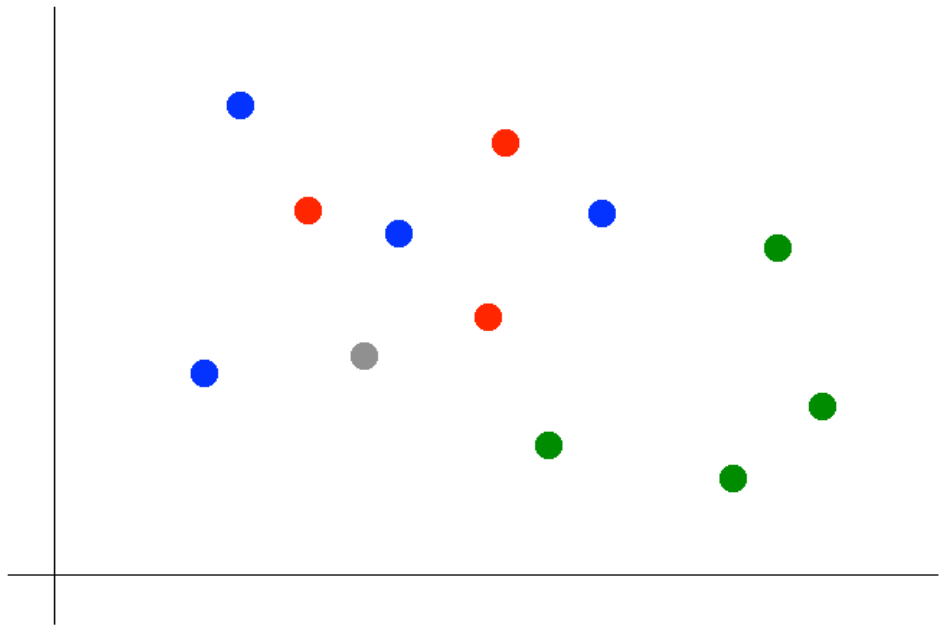
Suppose we want to predict the color of the grey dot.



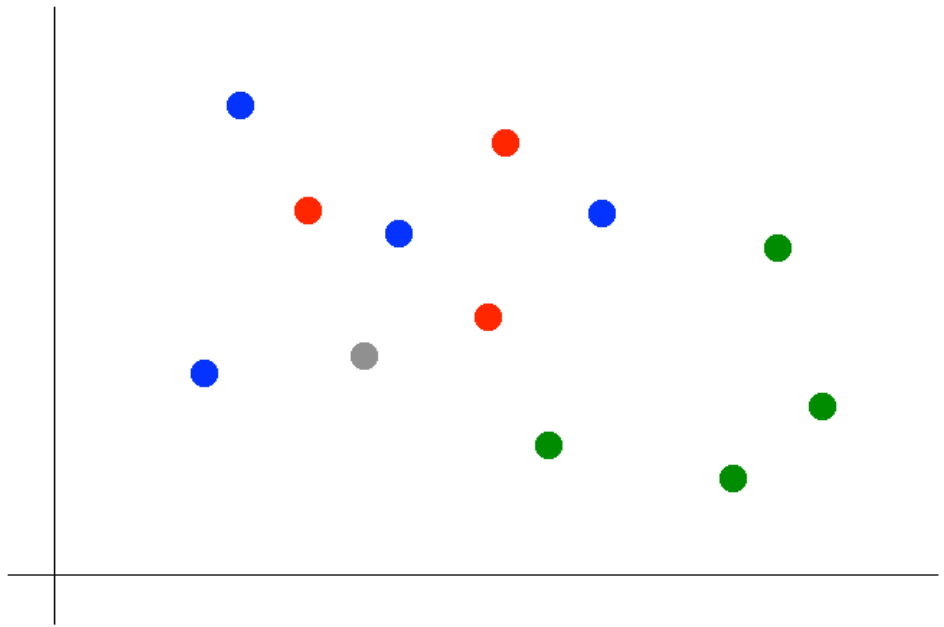
Suppose we want to predict the color of the grey dot.

QUESTION:

What are the features?
What are the labels?

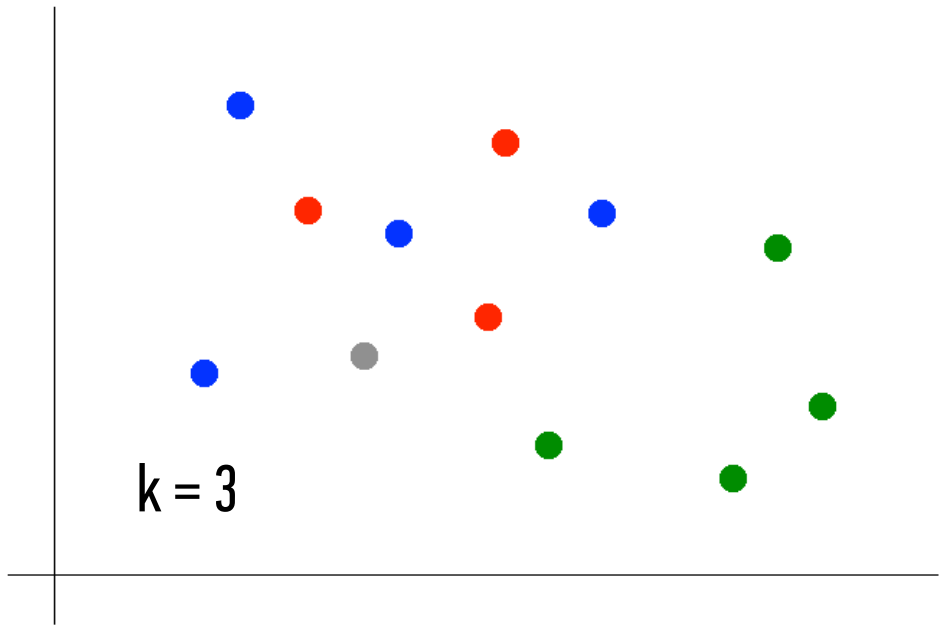


Suppose we want to predict the color of the grey dot.



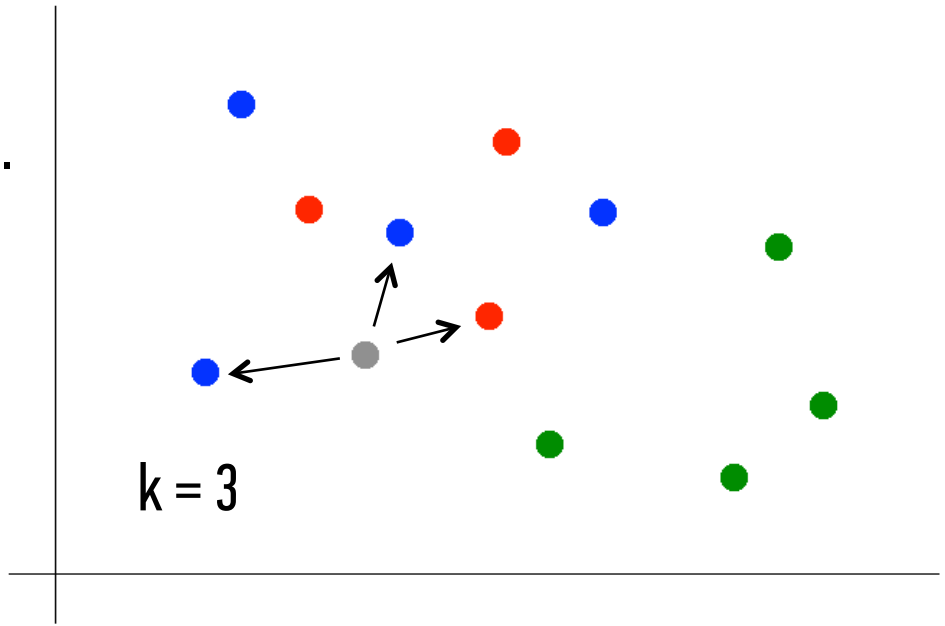
Suppose we want to predict the color of the grey dot.

1) Pick a value for k.



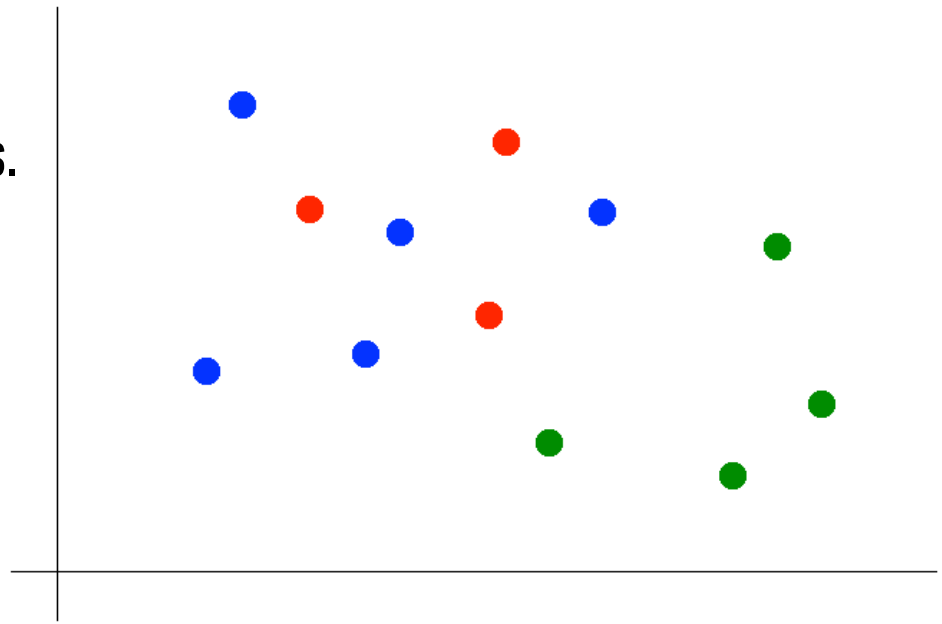
Suppose we want to predict the color of the grey dot.

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



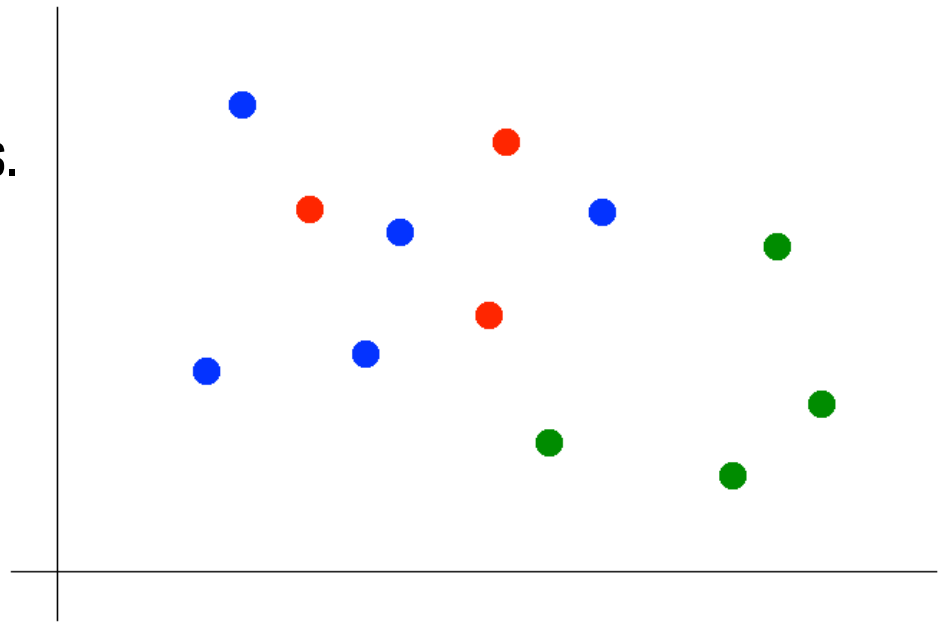
Suppose we want to predict the color of the grey dot.

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*.

INTRO TO DATA SCIENCE