GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LECTURE 3: MACHINE LEARNING

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

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

from Wikipedia:

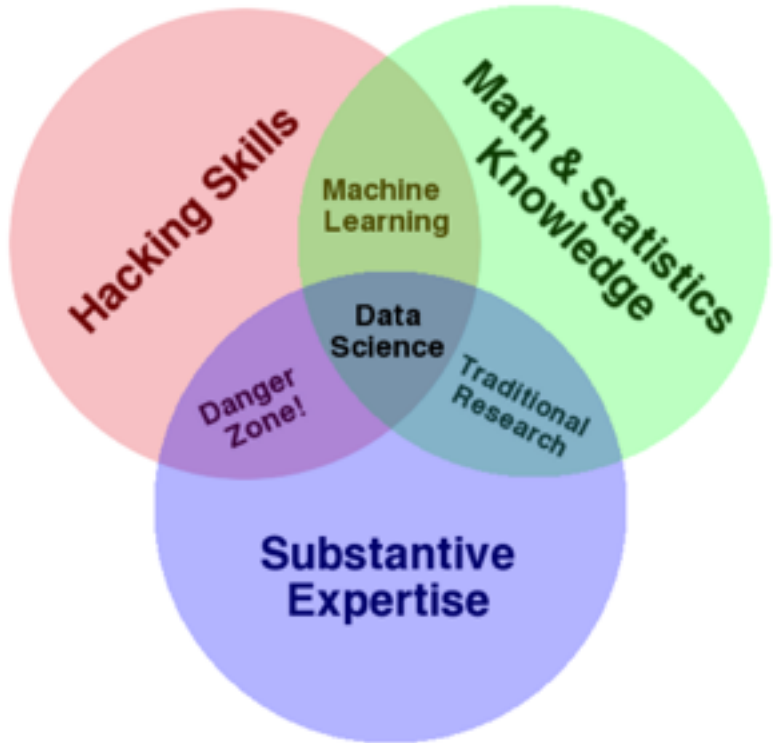
“Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can learn from data.”

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

- representation – extracting structure from data
- generalization – making predictions from data

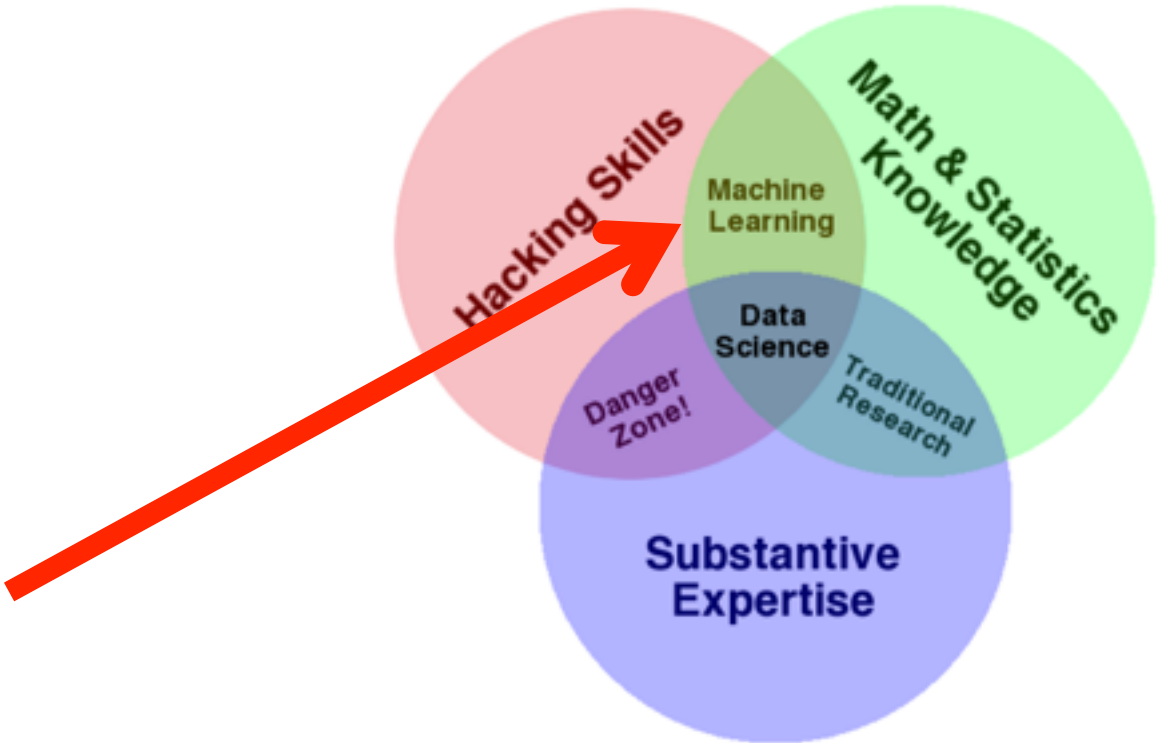
source: http://en.wikipedia.org/wiki/Machine_learning

REMEMBER THIS?



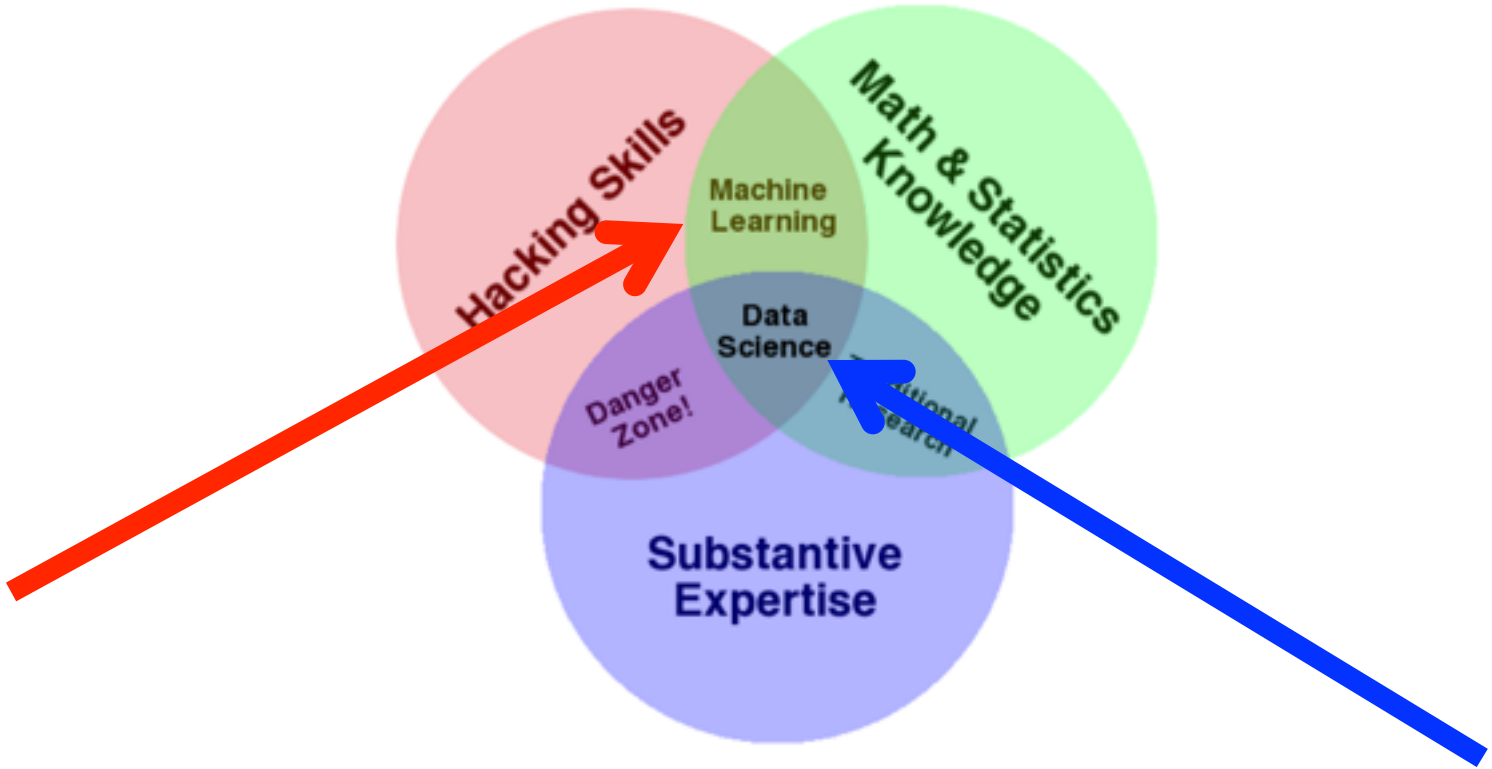
source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

YOU ARE HERE



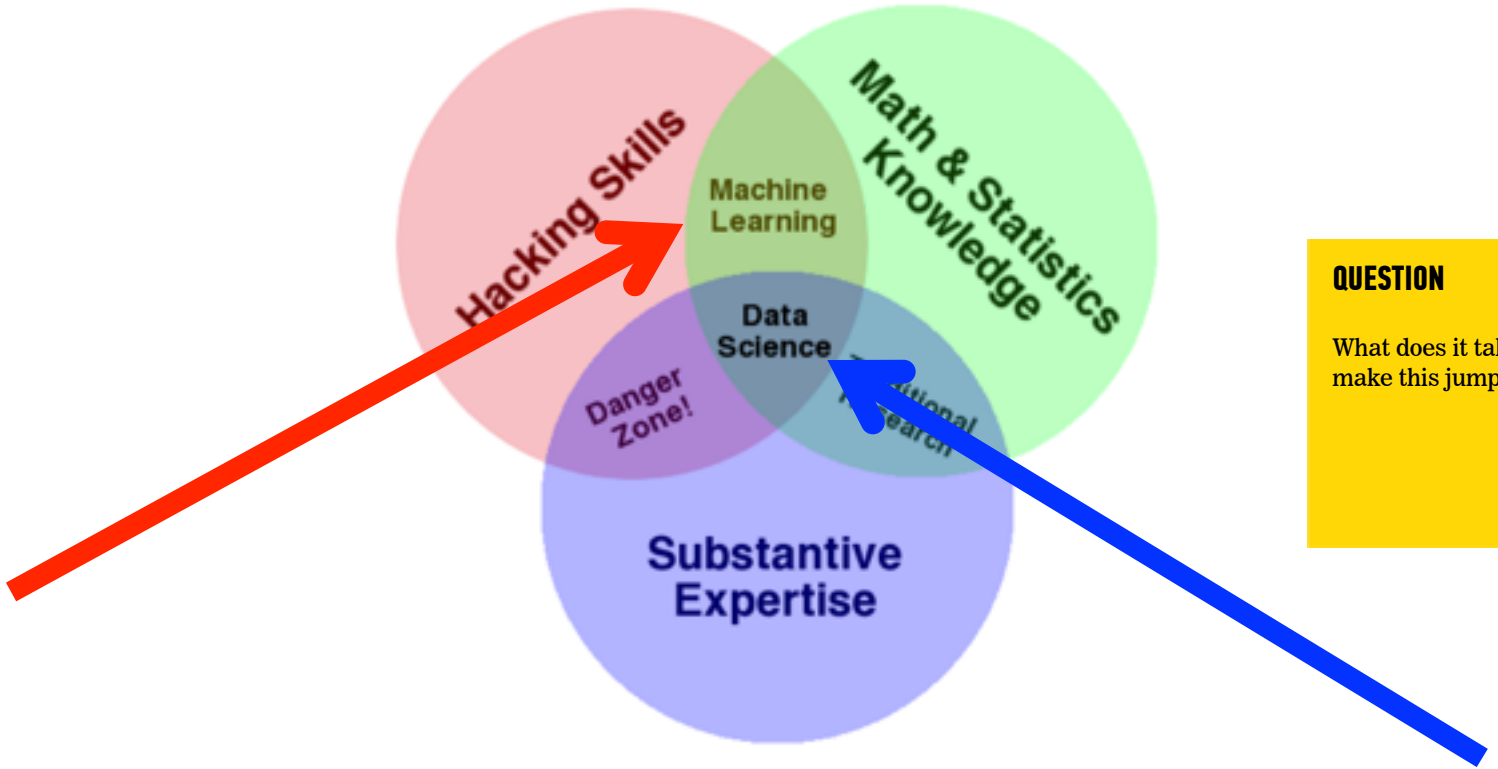
source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

YOU WANT TO GO HERE



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

YOU WANT TO GO HERE

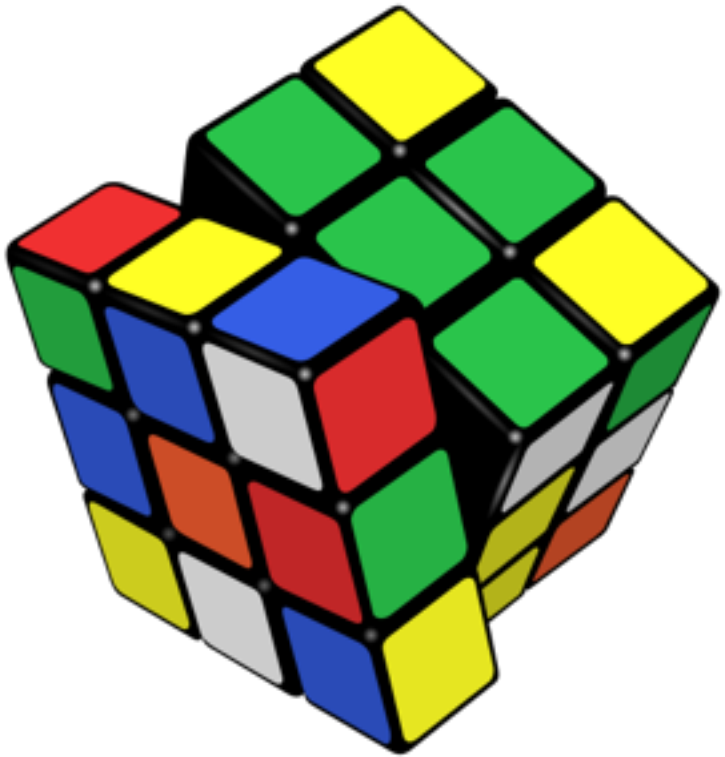


QUESTION

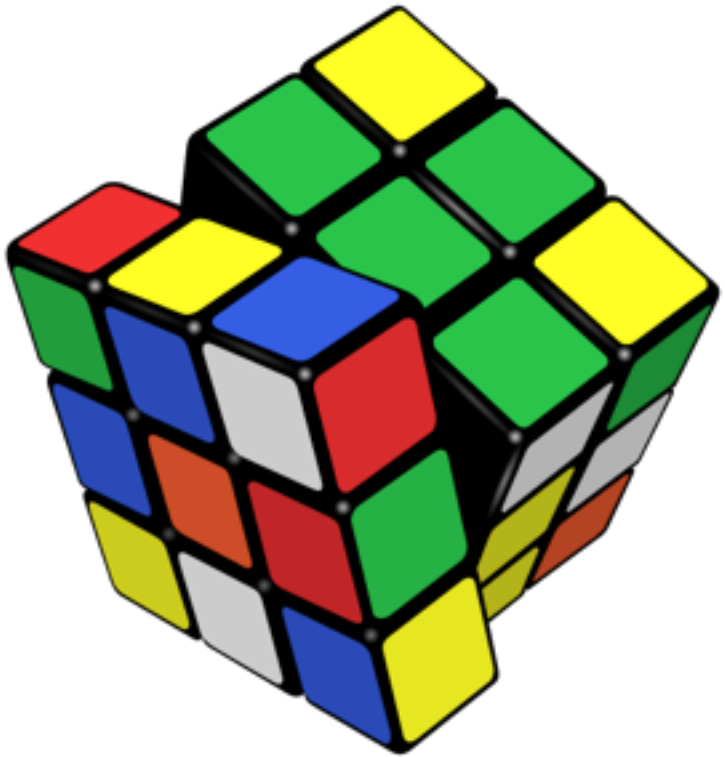
What does it take to make this jump?

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 PROBLEMS

Features: “Columns” of your data. Generally are independent variables that cover some numerical space (either as a continuous set or itemized matrix)

Features: “Columns” of your data. Generally are independent variables that cover some numerical space (either as a continuous set or itemized matrix)

Observations: “Rows” of your data. May not be unique, but each observation should represent one single representation of the feature space.

Features: “Columns” of your data. Generally are independent variables that cover some numerical space (either as a continuous set or itemized matrix)

Observations: “Rows” of your data. May not be unique, but each observation should represent one single representation of the feature space.

Algorithm: A “predetermined set of rules.” We will talk about a wide variety of algorithms throughout the next few weeks!

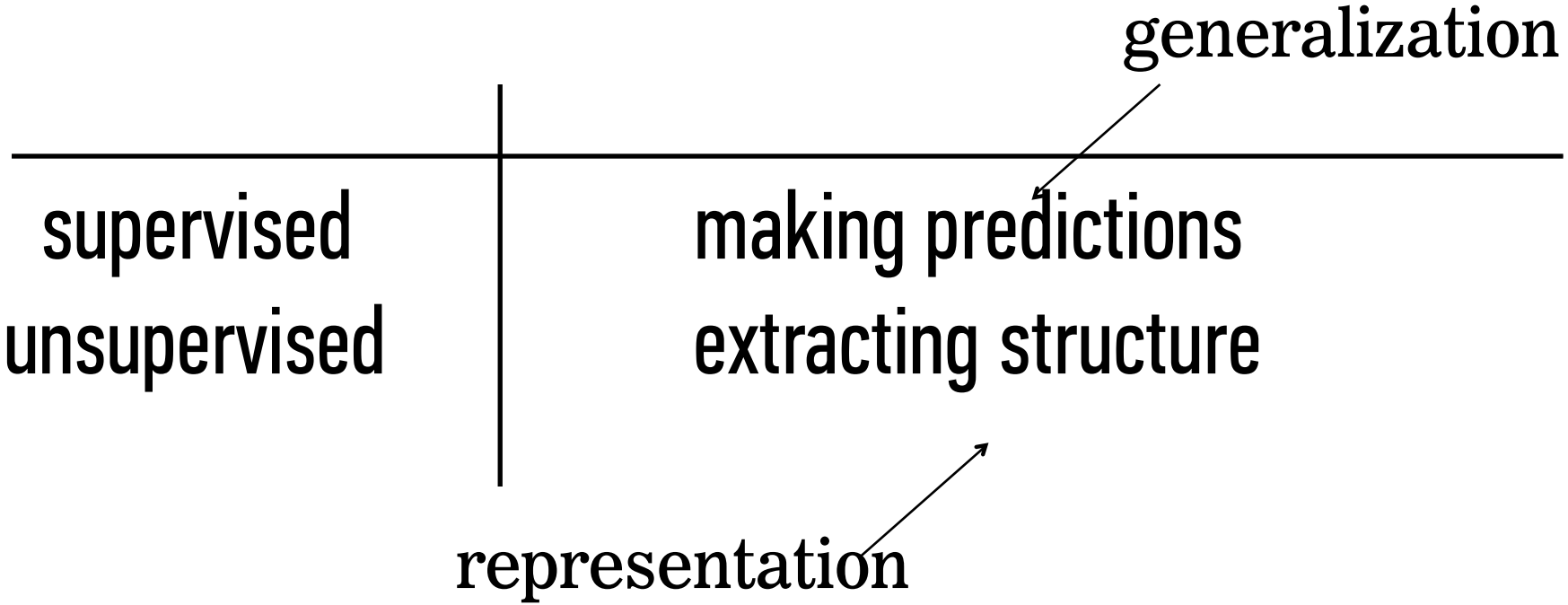
Features: “Columns” of your data. Generally are independent variables that cover some numerical space (either as a continuous set or itemized matrix)

Observations: “Rows” of your data. May not be unique, but each observation should represent one single representation of the feature space.

Algorithm: A “predetermined set of rules.” We will talk about a wide variety of algorithms throughout the next few weeks!

Model = Algorithm + Features + Observations

supervised	making predictions
unsupervised	extracting structure


generalization
supervised
unsupervised
making predictions
extracting structure
representation

	continuous	categorical
	quantitative	qualitative

	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

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

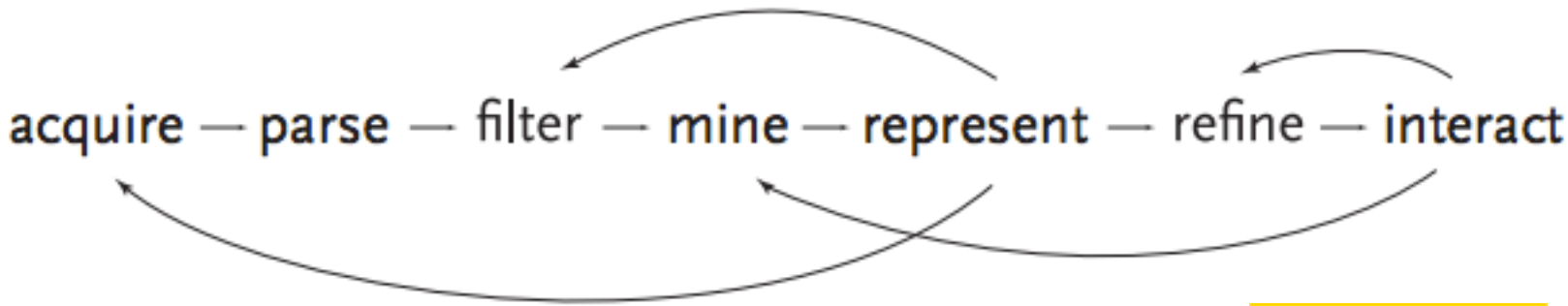
The
is d
des

NOTE

All of this depends on your data!

QUESTION

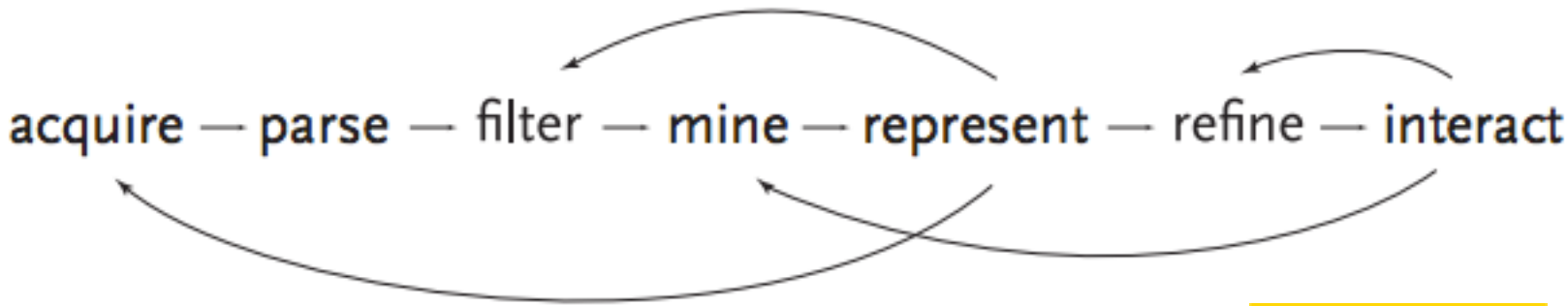
WHAT DO YOU DO WITH YOUR RESULTS?



ANSWER

Interpret them and react accordingly.

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



ANSWER

In re

NOTE

This also relies on your problem solving skills!

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

III. ‘CLASS’IFICATION ACTIVITY

One of the simplest machine learning algorithms is K-Nearest Neighbors (KNN for short)

One of the simplest machine learning algorithms is K-Nearest Neighbors (KNN for short)

KNN functions like this:

One of the simplest machine learning algorithms is K-Nearest Neighbors (KNN for short)

KNN functions like this:

1. Pick an observation to predict a classification

One of the simplest machine learning algorithms is K-Nearest Neighbors (KNN for short)

KNN functions like this:

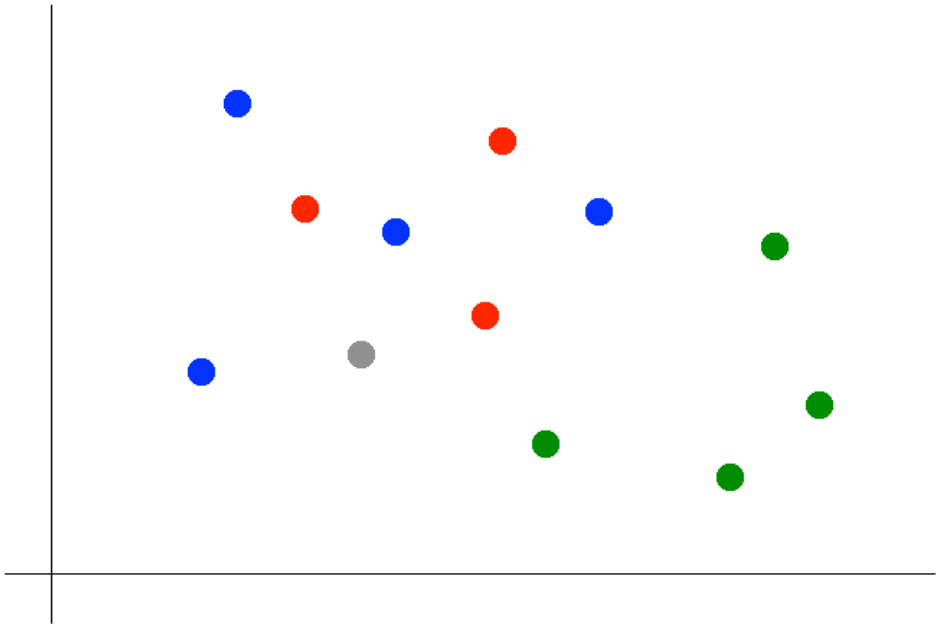
1. Pick an observation to predict a classification
2. Use a distance formula for neighbors k to find the nearest neighbors

One of the simplest machine learning algorithms is K-Nearest Neighbors (KNN for short)

KNN functions like this:

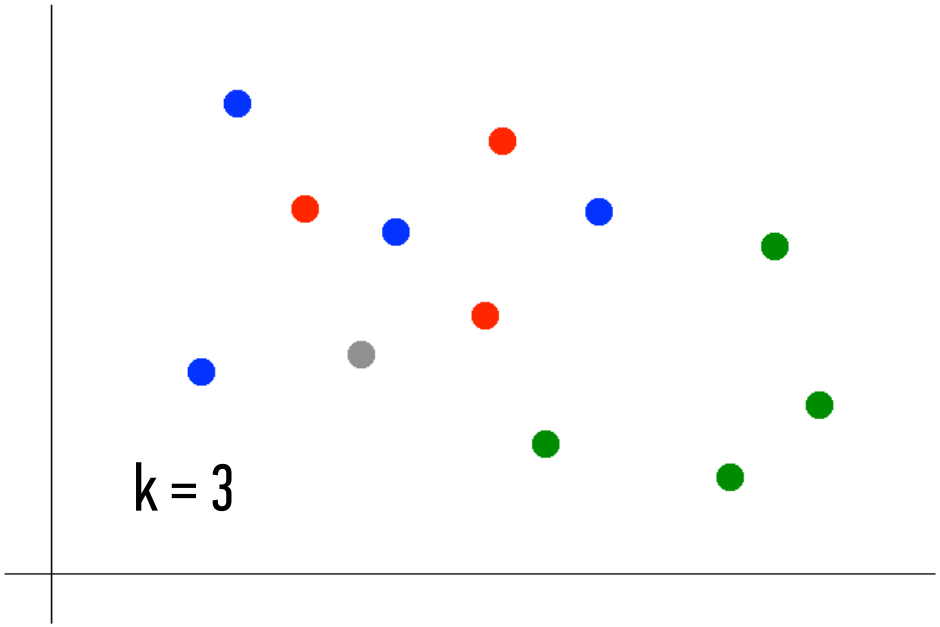
1. Pick an observation to predict a classification
2. Use a distance formula for neighbors k to find the nearest neighbors
3. Whichever neighbors have the highest representation of a class decide how to classify the unknown observation!

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



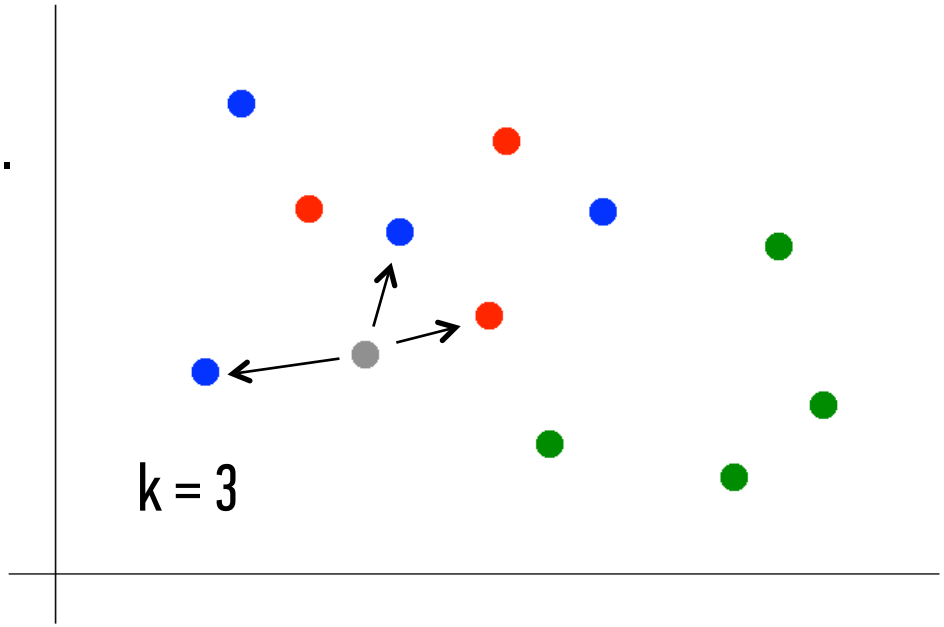
Suppose we want to predict the color of the grey dot.

1) Pick a value for k.



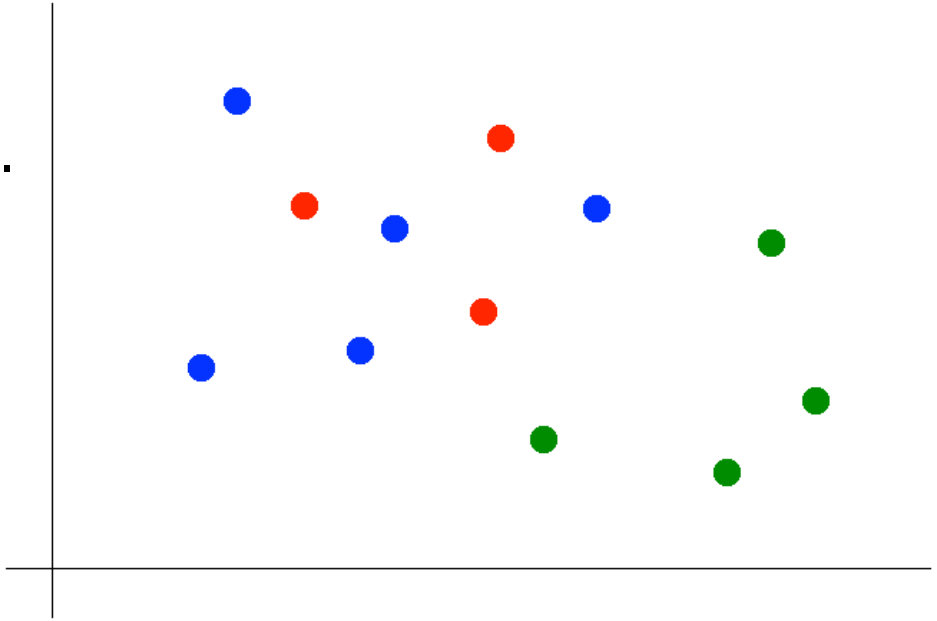
Suppose we want to predict the color of the grey dot.

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



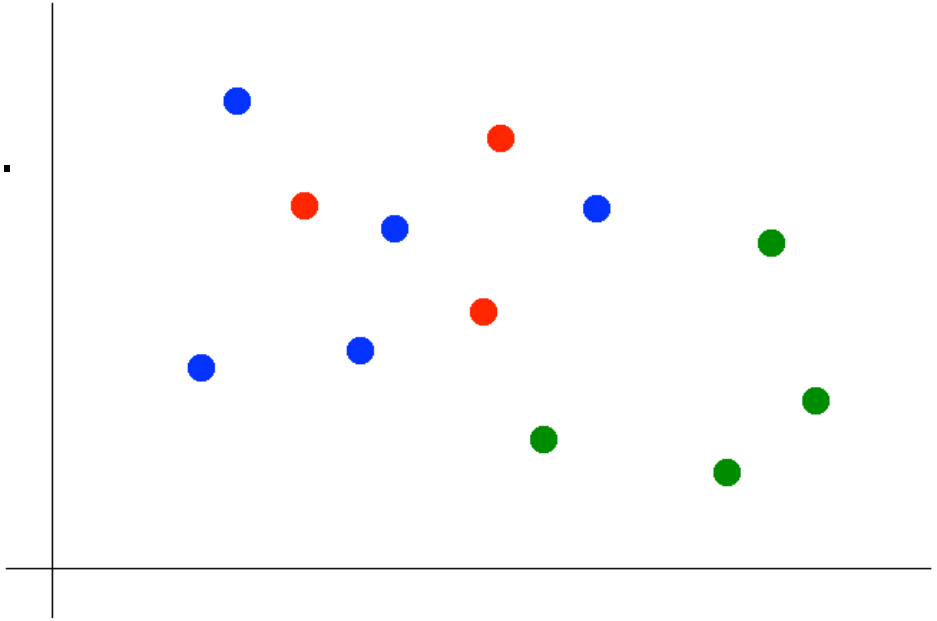
Suppose we want to predict the color of the grey dot.

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.*

EXERCISE – K NEAREST NEIGHBORS CLASSIFICATION IN R

KEY OBJECTIVES

- knn classification using train/test sets

R FUNCTIONS

- knn {class}

ASSIGNMENT – KNN WITH N-FOLD CROSS-VALIDATION

KEY OBJECTIVES

Extend the script we used in class to implement knn classification on the iris dataset using n-fold cross-validation.

(bonus: split code into functions)

for example:

```
knn.nfold <- function(n, … ) {
   # create n-fold partition of dataset
   # perform knn classification n times
   # n-fold generalization error = average over all iterations
}
```

As a class, consider our values as a feature space with 3 features:

1. On a scale of 1 to 10 (10 being most important), how important is it for you to work at a company where the product sells and brings in revenue?

As a class, consider our values as a feature space with 3 features:

1. On a scale of 1 to 10 (10 being most important), how important is it for you to work at a company where the product sells and brings in revenue? (continuous)
2. On a scale of 1 to 10 (10 being most important), how important is it for you to work at a company where the goal is to create an amazing product for everyone? (continuous)
3. Which region of the United States (or country) that you've spent the majority of your life.

Final thoughts/Reflection:

1. What worked well with that approach to classifying people?
2. What didn't work well with that approach to classifying people?
3. What other challenges could have came up with this approach?