

# DATA SCIENCE

## INTRO TO MACHINE LEARNING & KNN

# **QUESTIONS?**

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

**WHAT WAS THE HARDEST TO GRASP?**

- Understand the goal of machine learning
- Be able to articulate the differences between supervised and unsupervised learning
- Understand classification problems
- Implement a classifier using the KNN algorithm with Python and scikit-learn

**I. WHAT IS MACHINE LEARNING?**

**II. SUPERVISED LEARNING**

**III. UNSUPERVISED LEARNING**

**IV. SUMMARY**

**V. CLASSIFICATION WITH K-NEAREST NEIGHBORS**

# **I. WHAT IS MACHINE LEARNING?**

---

## WHY MACHINE LEARNING?

---

Applications of machine learning are everywhere, with some very successful companies built around them:

- Ranking web search results
- Predicting CTR on paid search results
- Credit card fraud detection
- Spam filtering
- Handwriting recognition / OCR
- Speech recognition
- Object detection / recognition
- Recommending movies / songs / products
- Self-driving cars
- Predicting vehicle arrival times
- Network intrusion detection
- Social network analysis
- Predicting the function of proteins
- Customer churn prediction
- House price prediction
- “Smart” thermostats
- Lending decisions
- High frequency trading
- Weather prediction

---

## WHY MACHINE LEARNING?

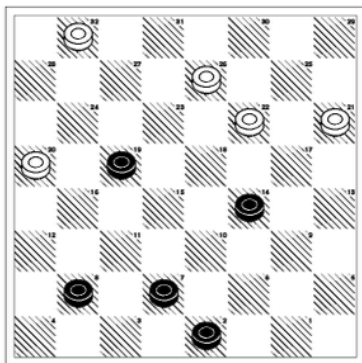
---



## THE ORIGINS OF MACHINE LEARNING

---

### *Shall we play a game?*



In 1958, computer scientist Arthur Samuel wrote a computer program to play checkers...

Initially, each board position was assigned a score reflecting its likelihood of leading to a win.

This worked, but was complicated.  
Also, performance wasn't great.

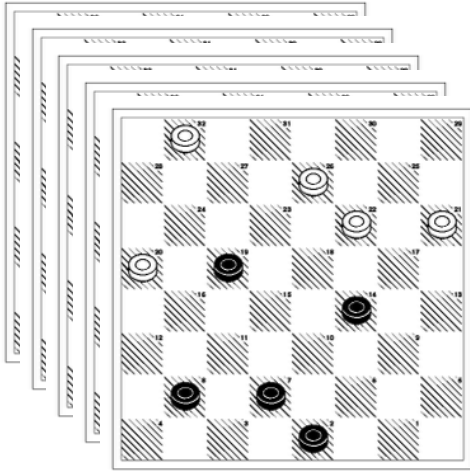


---

## THE ORIGINS OF MACHINE LEARNING

---

### *Shall we play a game?*



Then, he had an idea for improving performance:

Have the program play thousands of games against itself and use the results to improve the positional scoring.

He had written a program that was able to improve its performance through experience!

## WHAT IS MACHINE LEARNING?

---

*“A field of study that gives computers the ability to learn without being explicitly programmed.”*

*- Arthur Samuel (1959)*



## WELL-POSED LEARNING PROBLEM

---

*“A computer is said to **learn** from experience  $E$  with respect to some task  $T$  and some performance measure  $P$ , if its performance on  $T$ , as measured by  $P$ , improves with experience  $E$ .”*


*- Tom Mitchell (1997)*



source: <http://www.cs.cmu.edu/~tom/>

reddit MACHINELEARNING

comments related other discussions (4)

 ! AMA: Yann LeCun (self.MachineLearning)  
submitted 5 months ago \* by ylecun

My name is Yann LeCun. I am the Director of Facebook AI Research and a professor at New York University. Much of my research has been focused on deep learning, convolutional nets, and related topics.

Seriously, I don't like the phrase "Big Data". I prefer "**Data Science**", which is the **automatic (or semi-automatic) extraction of knowledge from data**. That is here to stay, it's not a fad. The amount of data generated by our digital world is growing exponentially with high rate (at the same rate our hard-drives and communication networks are increasing their capacity). But the amount of human brain power in the world is not increasing nearly as fast. This means that now or in the near future **most of the knowledge in the world will be extracted by machine and reside in machines**. It's inevitable. An entire industry is building itself around this, and a new academic discipline is emerging.

One definition: “Machine learning is the semi-automatic extraction of knowledge from data.”

- **Knowledge from data:** Starts with a question that might be answerable using data
- **Automatic extraction:** A computer provides the insight
- **Semi-automatic:** Requires many smart decisions by a human

Machine Learning is a class of algorithms which are data-driven. Unlike classical algorithms, it is the data that defines a “good” answer.

Example:

A **Non-Machine Learning** algorithm might “define” a face as having a roundish structure, two eyes, hair, nose, etc. The algorithm then looks for these “hard-coded” features in test cases.

A Machine Learning algorithm might only be given several pictures of faces and non-faces that are labeled as such. From the examples (called training set) it would “figure out” its own definition of a face.

# Training set



Face

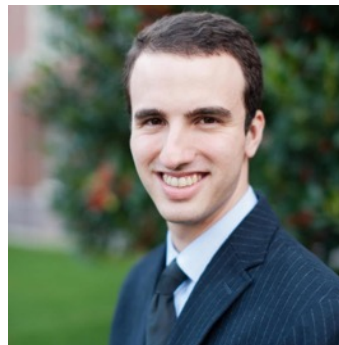


Not Face



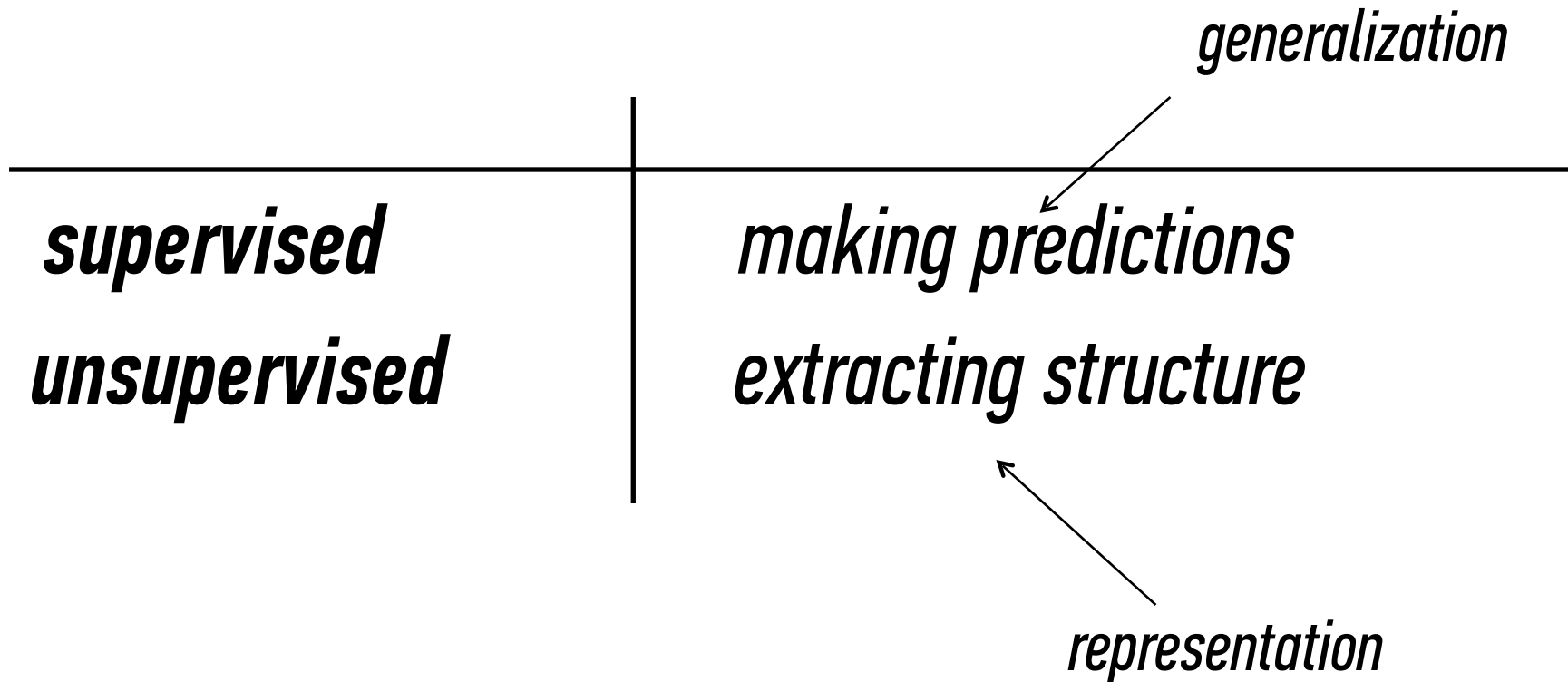
Face

# Test

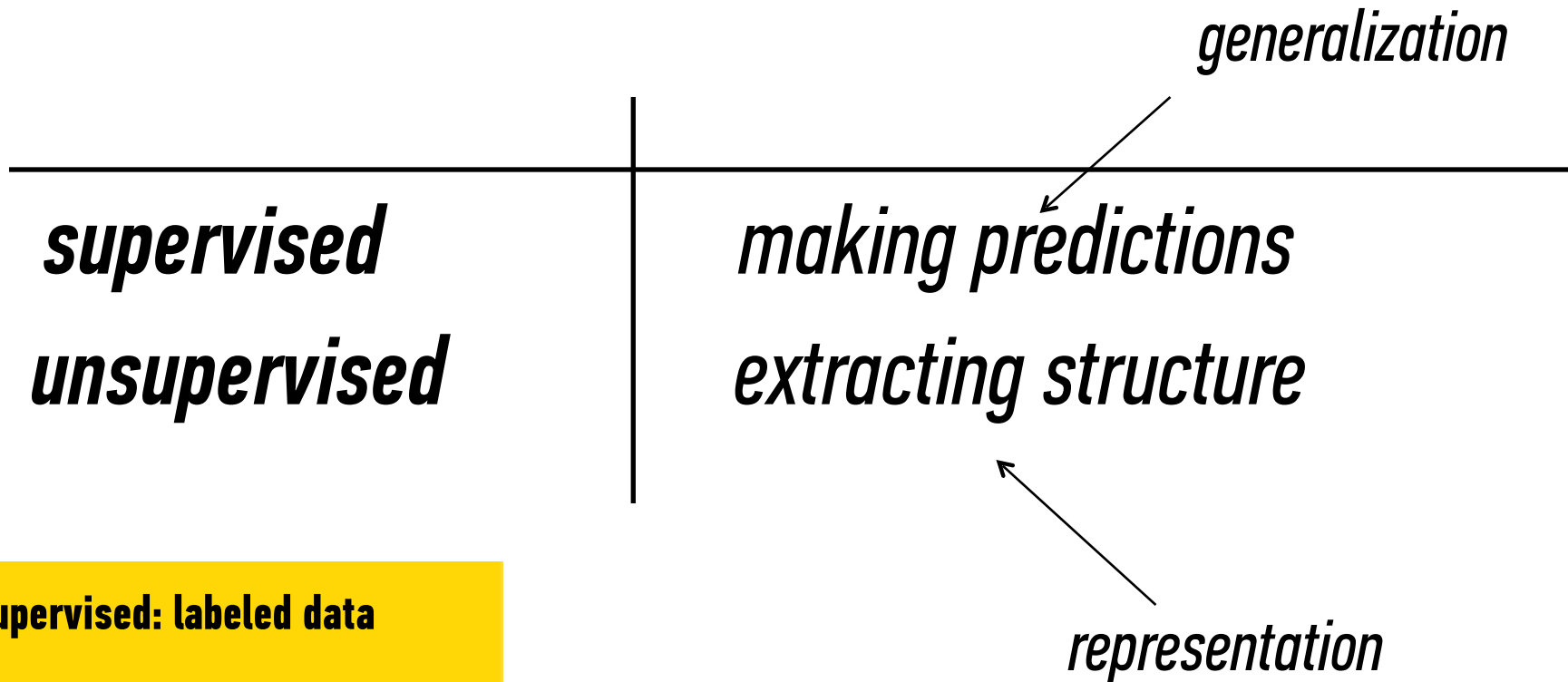


Face?

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

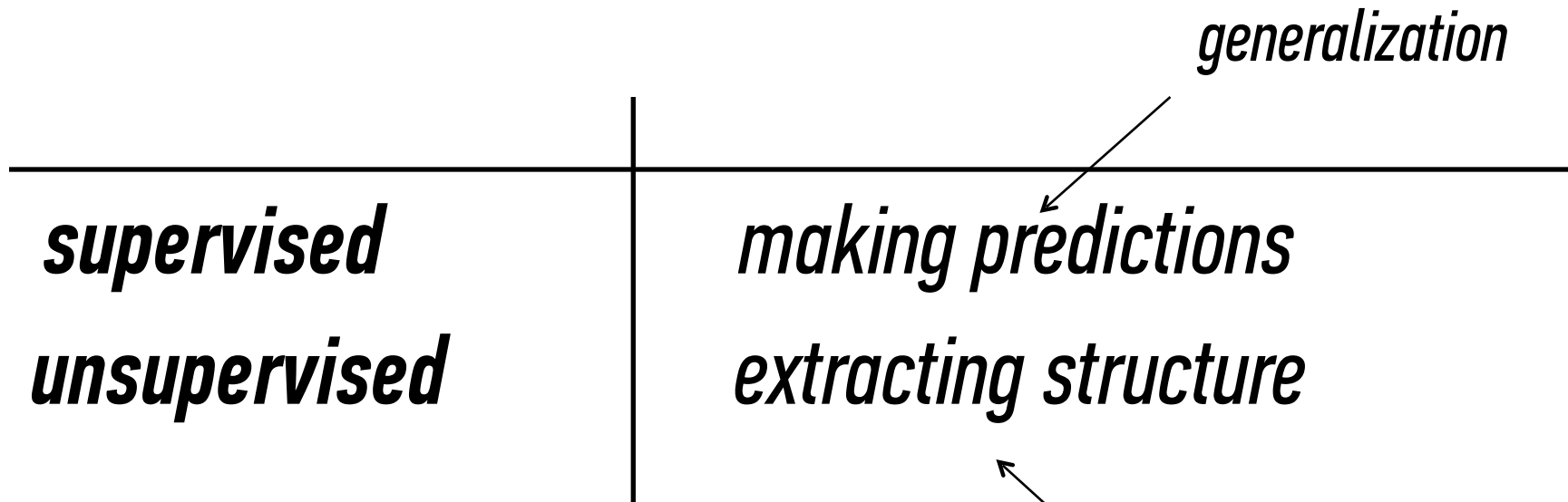






**Supervised: labeled data**

**Unsupervised: unlabeled data**



**Supervised: labeled data**

**Unsupervised: unlabeled data**

**Previous example  
was supervised!**

# **II. SUPERVISED LEARNING**

There are two main categories of machine learning: **supervised learning** and **unsupervised learning**.

**Supervised learning** (aka “predictive modeling”):

- Predict an outcome based on input data
- Example: predict whether an email is spam or ham
- Goal is “generalization”

## EXAMPLE #1: PREDICTING NEONATAL INFECTION

21

**Problem:** Children born prematurely are at high risk of developing infections, many of which are not detected until after the baby is sick

**Goal:** Detect subtle patterns in the data that predicts infection before it occurs



**Data:** 16 vital signs such as heart rate, respiration rate, blood pressure, etc...

**Impact:** Model is able to predict the onset of infection 24 hours before the traditional symptoms of infection appear

predictors

Sample response: Did the child develop an infection? True/False

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>

*150 observations*  
*( $n = 150$ )*

Feature matrix “X” has  $n$  rows and  $p$  columns

Response “y” is a vector with length  $n$

*4 features ( $p = 4$ )*

*response*

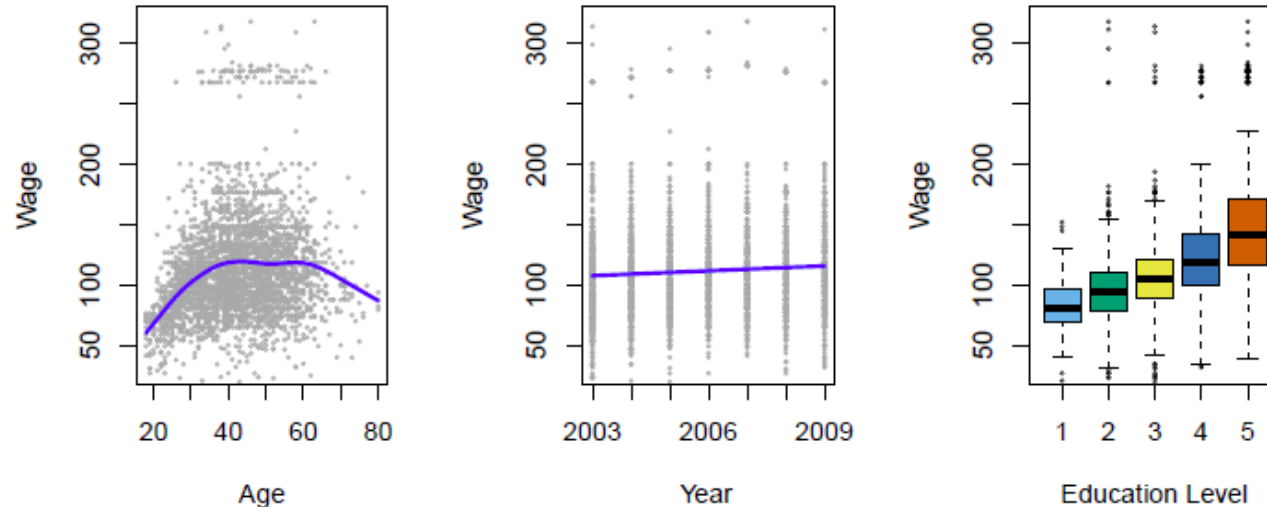
**Observations** are also known as: samples, examples, instances, records

**Features** are also known as: predictors, independent variables, inputs, regressors, covariates, attributes

**Response** is also known as: outcome, label, target, dependent variable

**Regression problems** have a continuous response. **Classification problems** have a categorical response. The type of supervised learning problem has nothing to do with the features!

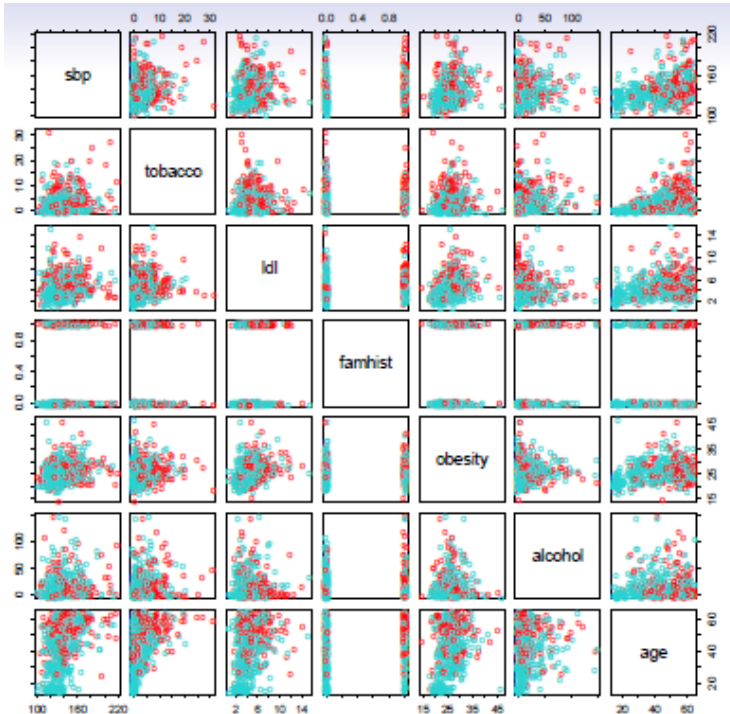
- Establish the relationship between salary and demographic variables in population survey data



Income survey data for males from the central Atlantic region of the USA in 2009

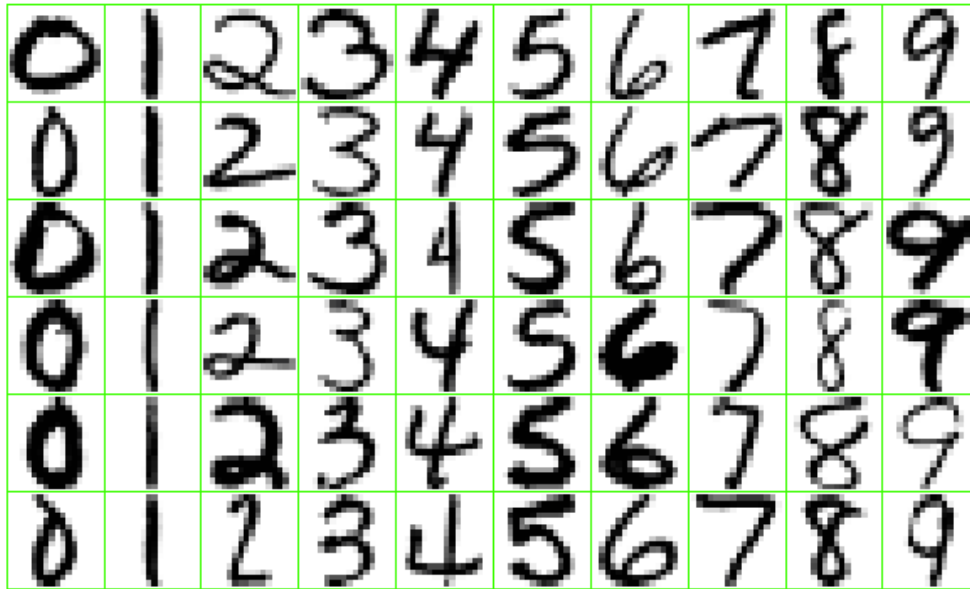


- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements



Case-control sample of men from South Africa  
 Red = heart disease  
 Blue = no heart disease

- Identify the numbers in a handwritten zip code





Customer Solutions

Competitions

Community ▾

Sinan Ozdemir

Logout



Knowledge • 1,029 teams

## Forest Cover Type Prediction

Fri 16 May 2014

Mon 11 May 2015 (3 months to go)

Dashboard

Home



Data



Make a submission



Information



Description

Evaluation

Rules

Competition Details » [Get the Data](#) » [Make a submission](#)

# Use cartographic variables to classify forest categories

[Random forests?](#) [Cover trees?](#) Not so fast, computer nerds. We're talking about the

The study area includes four wilderness areas located in the Roosevelt National Forest of northern Colorado. Each observation is a 30m x 30m patch. You are asked to predict an integer classification for the forest cover type. The seven types are:

- 1 - Spruce/Fir
- 2 - Lodgepole Pine
- 3 - Ponderosa Pine
- 4 - Cottonwood/Willow
- 5 - Aspen
- 6 - Douglas-fir
- 7 - Krummholz

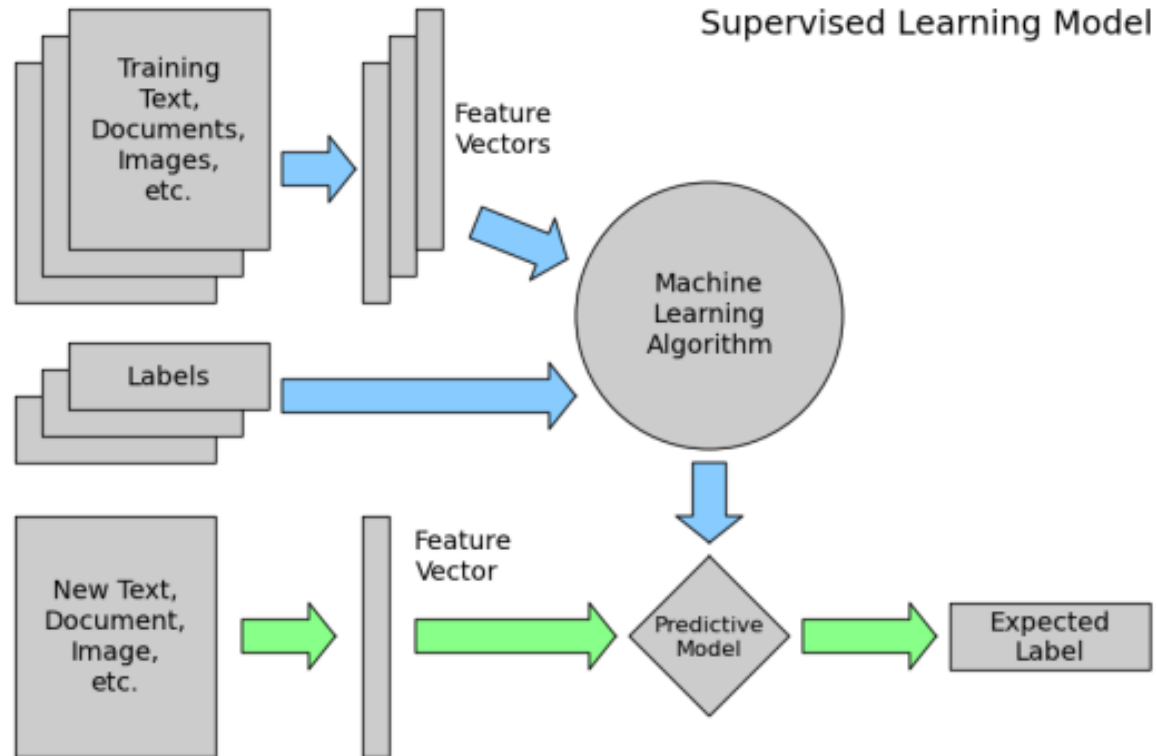
The training set (15120 observations) contains both features and the Cover\_Type. The test set contains only the features. You must predict the Cover\_Type for every row in the test set (565892 observations).

## How does supervised learning “work”?

1. Train a **machine learning model** using **labeled data**
  - “Labeled data” is data with a response variable
  - “Machine learning model” learns the relationship between the features and the response
2. Make predictions on **new data** for which the response is unknown

The primary goal of supervised learning is to build a model that “generalizes”:  
It accurately predicts the **future** rather than the **past**!

## How does supervised learning “work”?



## Supervised learning example: Dog detector

- Input data: Images from Google
  - Features: Numerical representations of the images
  - Response: Dog (yes or no), hand-labeled
1. Train a **machine learning model** using **labeled data**
    - Model learns the relationship between the image data and the “dog status”
  2. Make predictions on **new data** for which the response is unknown
    - Give it a new image, predicts the “dog status” automatically

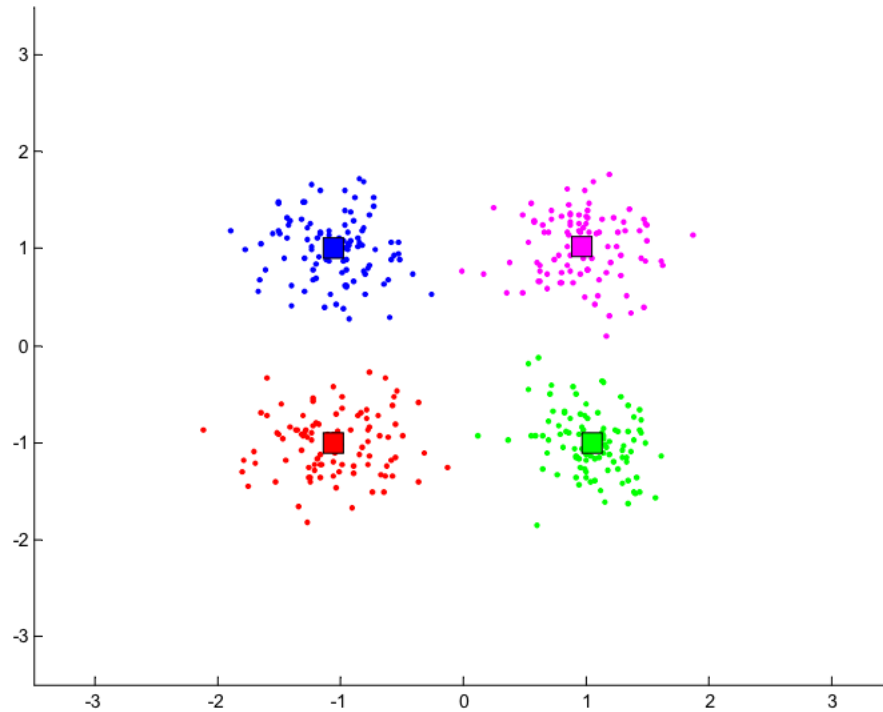
# **III. UNSUPERVISED LEARNING**



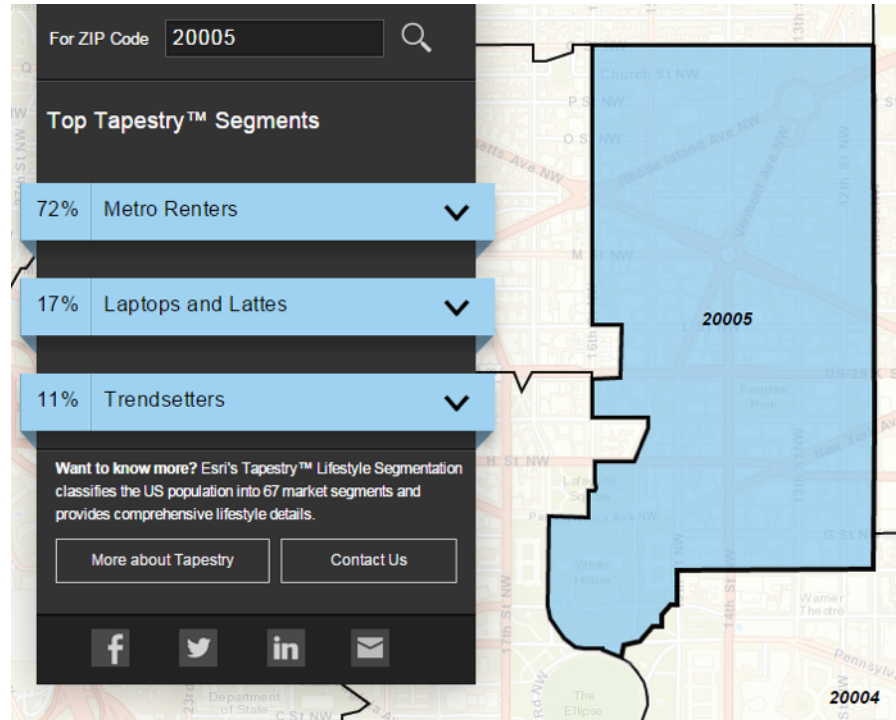
There are two main categories of machine learning: **supervised learning** and **unsupervised learning**.

**Unsupervised learning:**

- Extracting structure from data
- Example: segment grocery store shoppers into “clusters” that exhibit similar behaviors
- Goal is “representation”



- Classify US residential neighborhoods into 67 unique segments based on demographic and socioeconomic characteristics



#### Example of cluster: **Metro Renters:**

- Young, mobile, educated, or still in school
- Live alone or with a roommate
- Works long hours
- Buys groceries at Whole Foods and Trader Joe's
- Shops at Banana Republic, Nordstrom, and Gap
- Loves yoga, go skiing, and attend Pilates sessions.

Source: <http://www.esri.com/landing-pages/tapestry/>

Unsupervised learning has some clear differences from supervised learning.

With **unsupervised learning**:

- There is no clear objective
- There is no “right answer” (hard to tell how well you are doing)
- There is no response variable, just observations with features
- Labeled data is not required

## Unsupervised learning example: Image clustering

- Input data: Images from Google
- Features: Numerical representations of the images
- Response: There isn't one (no hand-labeling required!)

### I. Perform **unsupervised learning**

- Cluster the images based on “similarity”
- Might find a “dog cluster”, might not
- You're done!

Sometimes, unsupervised learning is used as a “preprocessing” step for supervised learning.

# **IV. SUMMARY**

---

## ML SOLUTIONS

---

	<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>

# EXERCISE:

supervised or unsupervised?



# REGRESSION - HOUSE PRICE PREDICTION

---

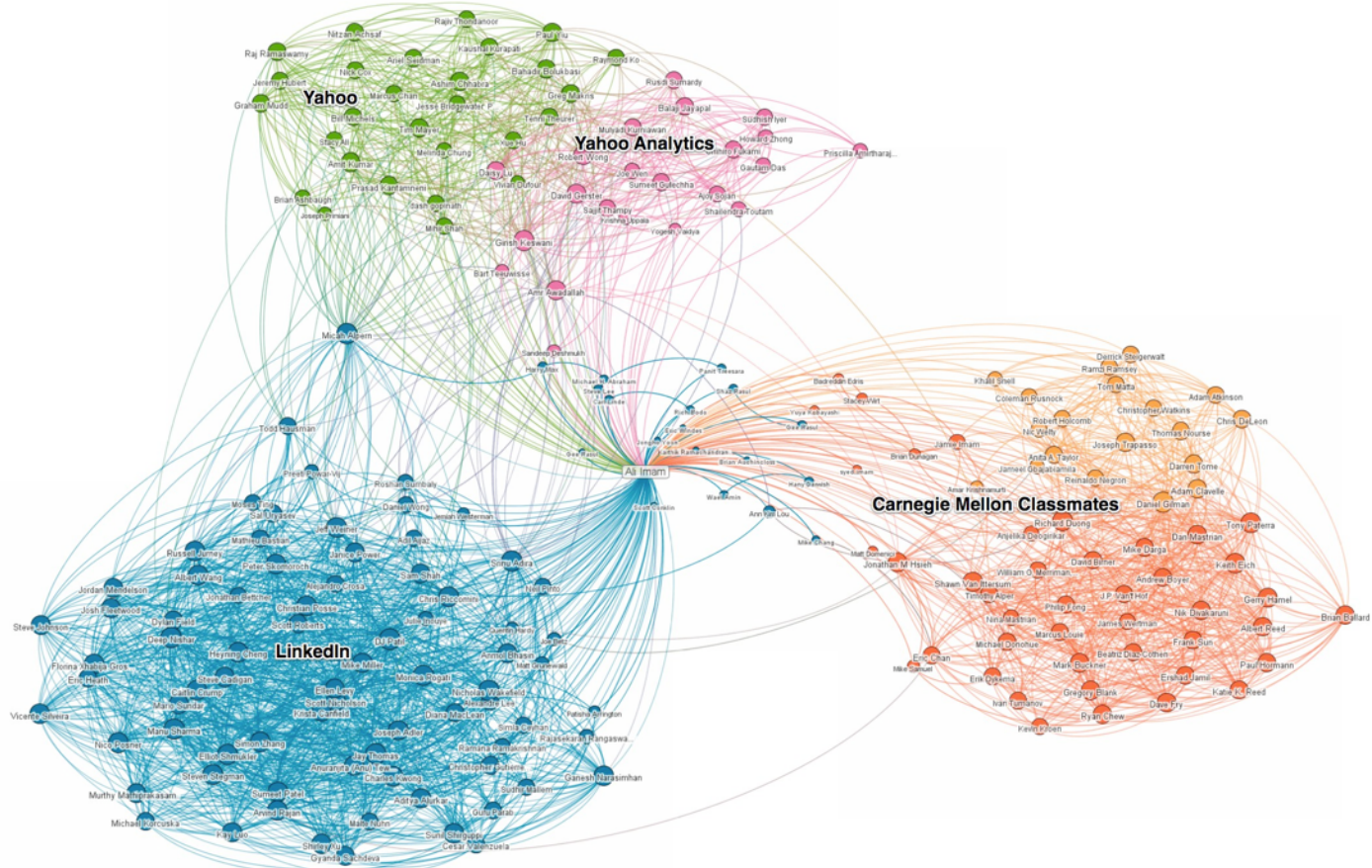


# DOCUMENT CLASSIFICATION

---



# COMMUNITY DETECTION IN SOCIAL NETWORKS



# **V. CLASSIFICATION WITH K-NEAREST NEIGHBORS**

---

## ML SOLUTIONS

---

	<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>

## ML SOLUTIONS

	<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>

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>

*150 observations*  
*( $n = 150$ )*

Feature matrix “X” has  $n$  rows and  $p$  columns

Response “y” is a vector with length  $n$

*4 features ( $p = 4$ )*

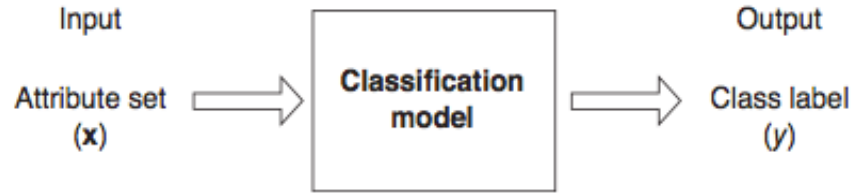
*response*

# CLASSIFICATION PROBLEMS

---

*Q: How does a classification problem work?*

*A: Data in, predicted labels out.*



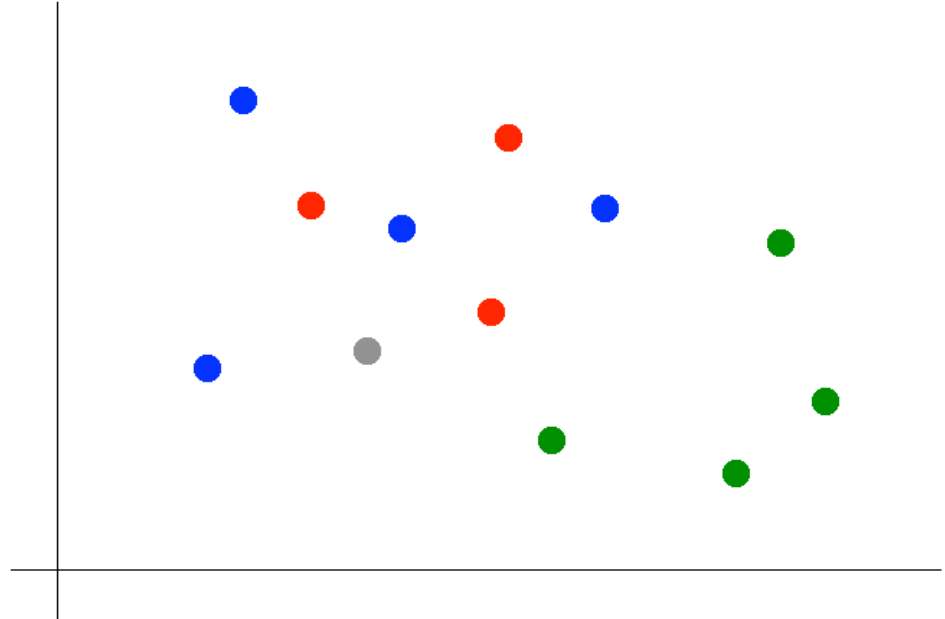
**Figure 4.2.** Classification as the task of mapping an input attribute set  $x$  into its class label  $y$ .



*Suppose we want to predict the color of the gray dot.*

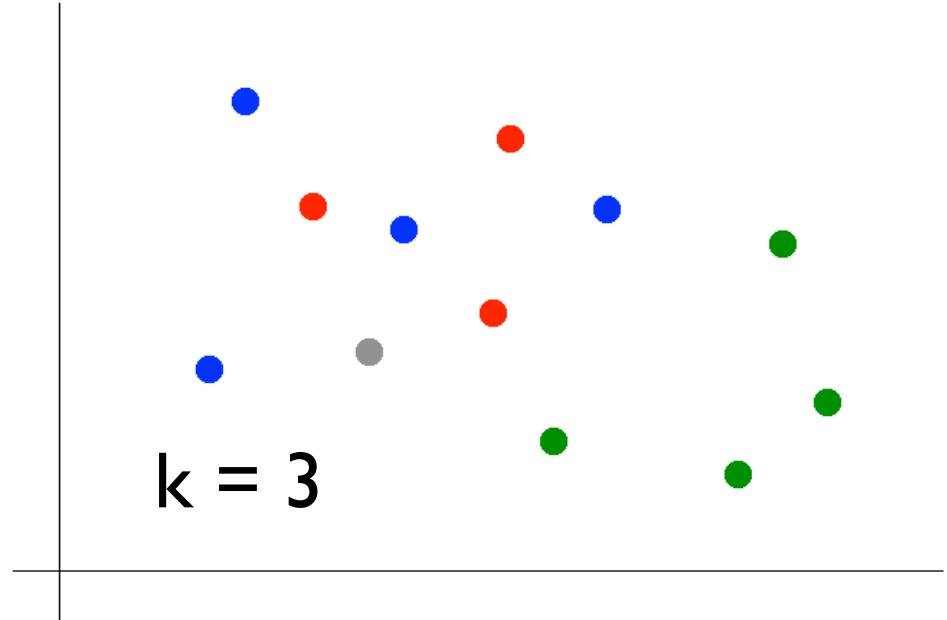
**QUESTION:**

What are the predictors?  
What is the response?



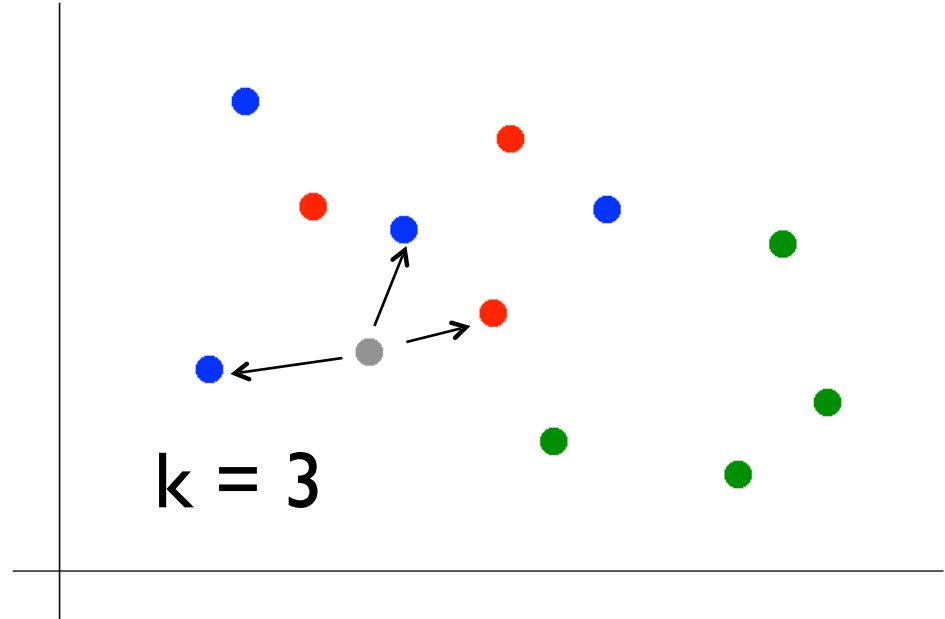
*Suppose we want to predict the color of the gray dot.*

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



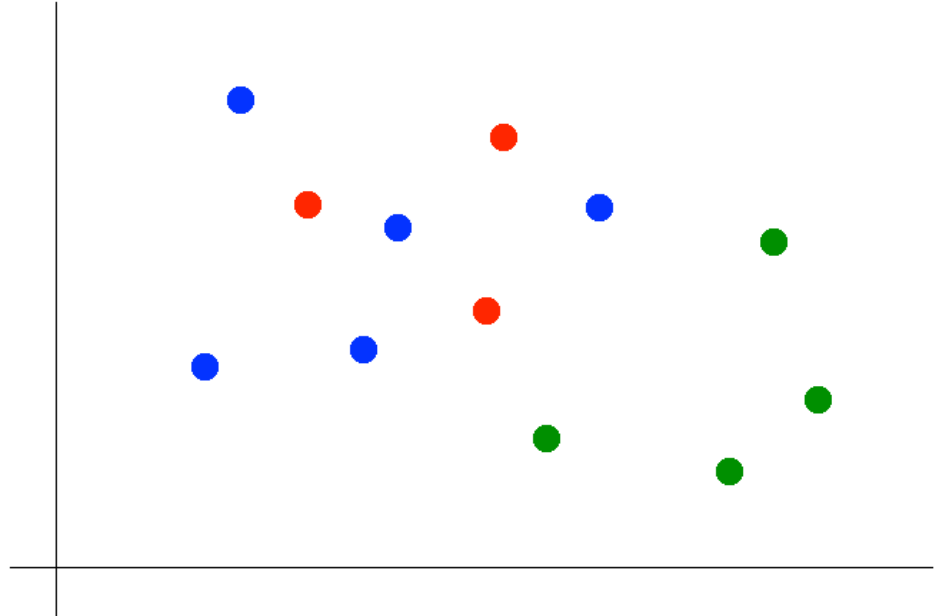
*Suppose we want to predict the color of the gray dot.*

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



*Suppose we want to predict the color of the gray dot.*

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

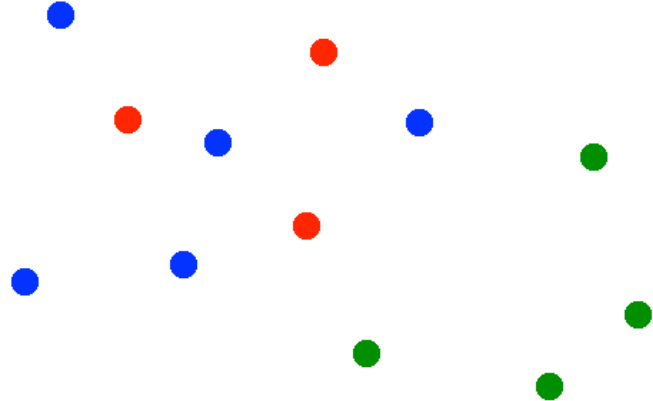


*Suppose we want to predict the color of the gray dot.*

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

**NOTE:**

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



### Advantages of KNN:

- Simple to understand and explain
- Model training phase is fast
- Non-parametric (does not presume a “form” of the “decision boundary”)

### Disadvantages of KNN:

- Prediction phase can be slow when  $n$  is large
- Sensitive to irrelevant features
- Very sensitive to feature scaling

---

**INTRO TO DATA SCIENCE**

---

**QUESTIONS?**