

# **DATA SCIENCE**

## **MACHINE LEARNING AND KNN**

**I. WHAT IS MACHINE LEARNING?**

**II. SUPERVISED LEARNING**

**III. UNSUPERVISED LEARNING**

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

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


"A field of study that gives computers the ability to learn without being explicitly programmed." (1959)



Arthur Samuel, AI pioneer  
Source: Stanford

# WHAT IS MACHINE LEARNING?

5

 **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.

There are two main categories of machine learning:

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

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

## **Unsupervised learning**

- Extracting structure from data
- Example: create segments of voters
- Goal is “representation”

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

Feature matrix “X” has  
n rows and p columns

Response “y” is a  
vector with length n

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>

*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

Note: **Unsupervised learning** does not have a response, and does not require labeled data!



## **II. SUPERVISED LEARNING**

There are two categories of supervised learning:

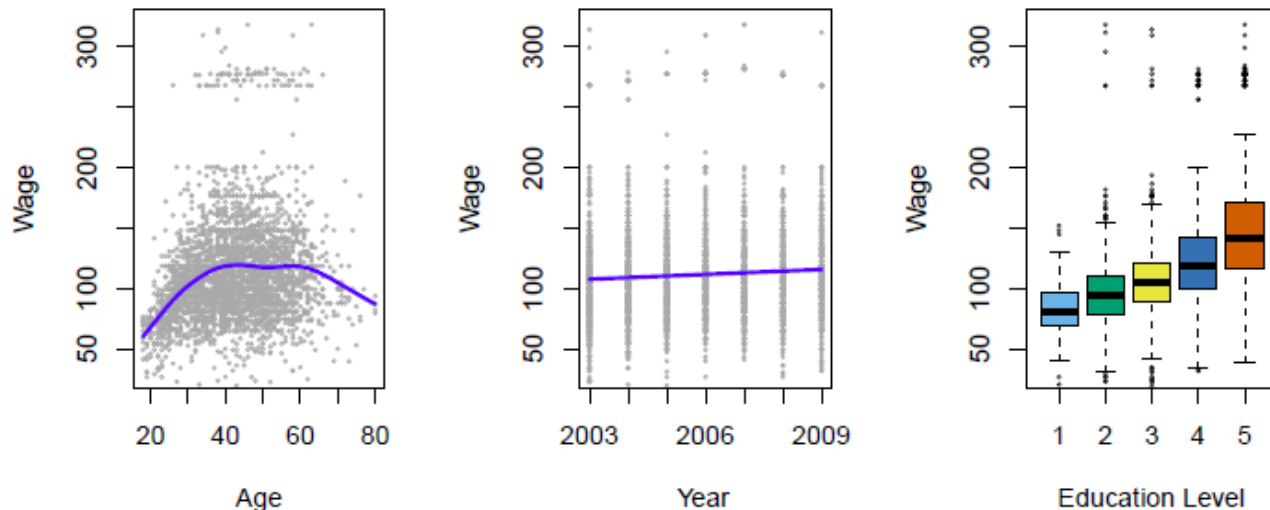
### **Regression**

- Response is continuous
- Examples: price, blood pressure

### **Classification**

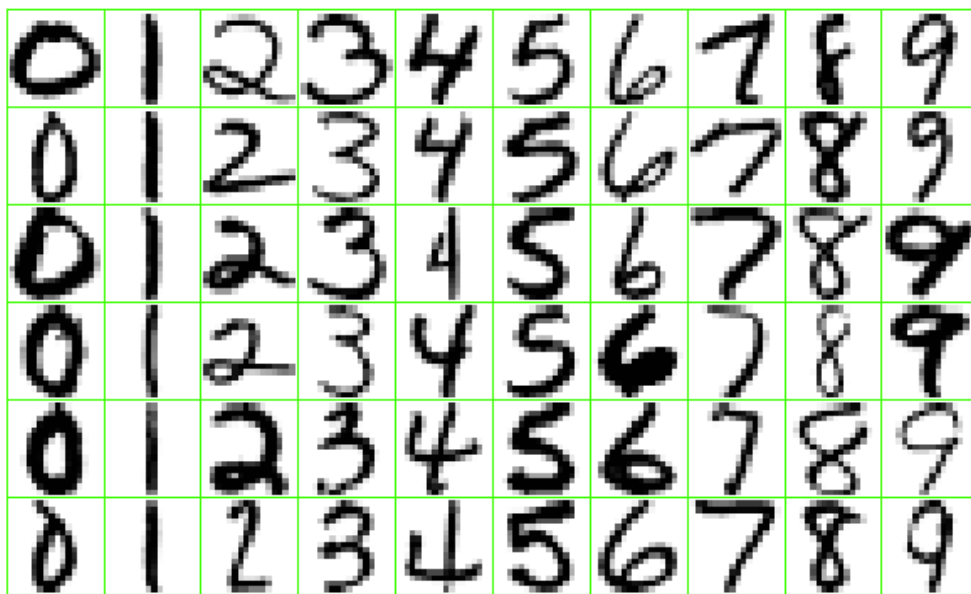
- Response is categorical (values in a finite, unordered set)
- Examples: spam/ham, digit 0-9, cancer class of tissue sample

Predict salary using demographic data



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

Identify the numbers in a handwritten zip code



**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



# **III. UNSUPERVISED LEARNING**

**Supervised learning** has clear objectives:

- Accurately predict unseen test cases
- Understand which features affect the response, and how

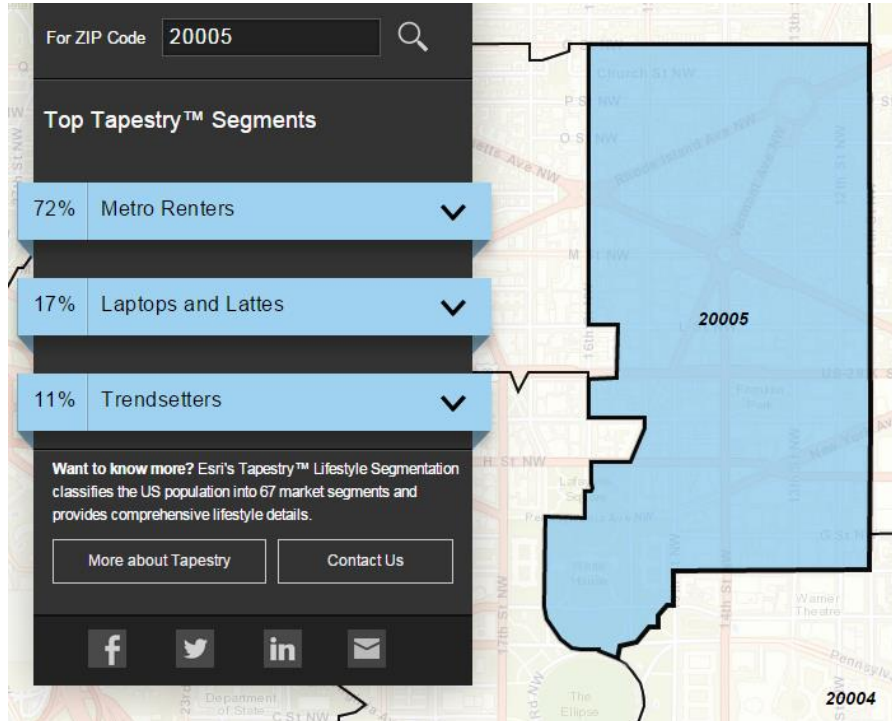
You can evaluate how well you are doing!

**Unsupervised learning** has fuzzy objectives:

- Find groups of observations that behave similarly
- Find features that behave similarly

It's difficult to evaluate how well you are doing!

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



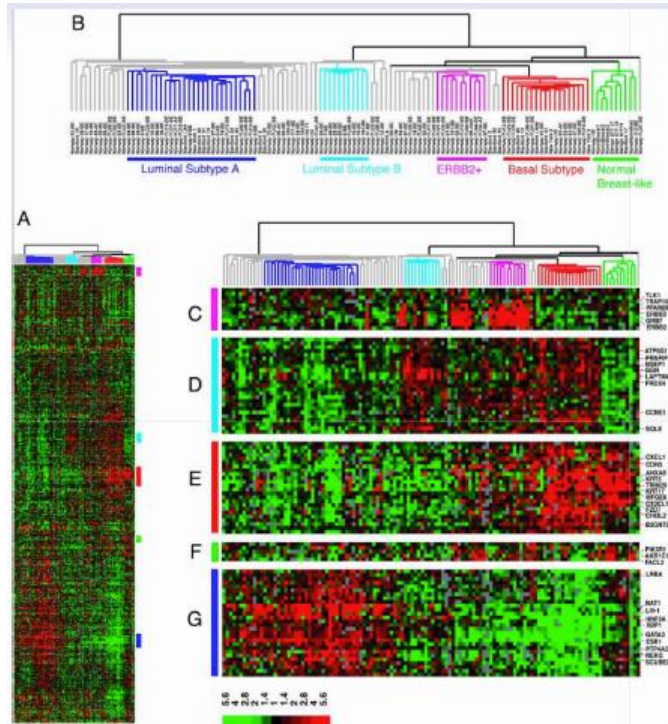
## Metro Renters:

Young, mobile, educated, or still in school, we live alone or with a roommate in rented apartments or condos in the center of the city. Long hours and hard work don't deter us; we're willing to take risks to get to the top of our professions... We buy groceries at Whole Foods and Trader Joe's and shop for clothes at Banana Republic, Nordstrom, and Gap. We practice yoga, go skiing, and attend Pilates sessions.

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



Classify a tissue sample into one of several cancer classes, based on gene expression data



- Each column is a woman with breast cancer (n=88)
- Each row is a gene (p=8000)
- Color represents level of gene expression

Goal: Locate subcategories of breast cancer showing different gene expressions

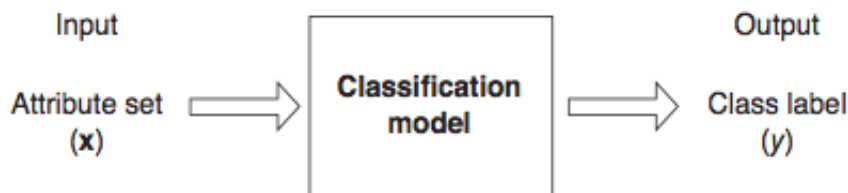
Technique: Hierarchical clustering applied to the columns, resulting in six sub-groups of patients

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

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

*Q: How does a classification problem work?*

*A: Features in, predicted response out.*

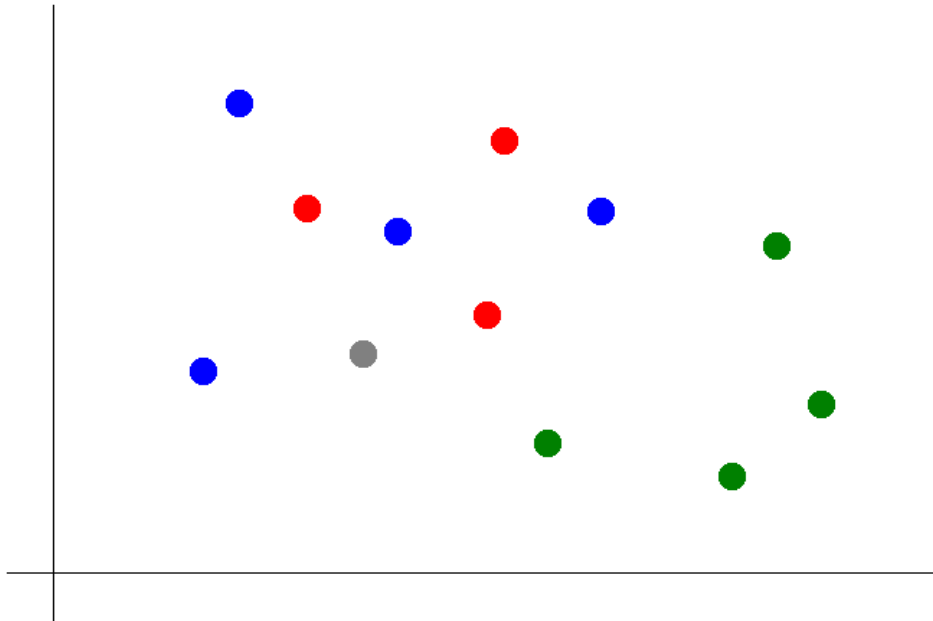


**Figure 4.2.** Classification as the task of mapping an input attribute set  $x$  into its class label  $y$ .

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

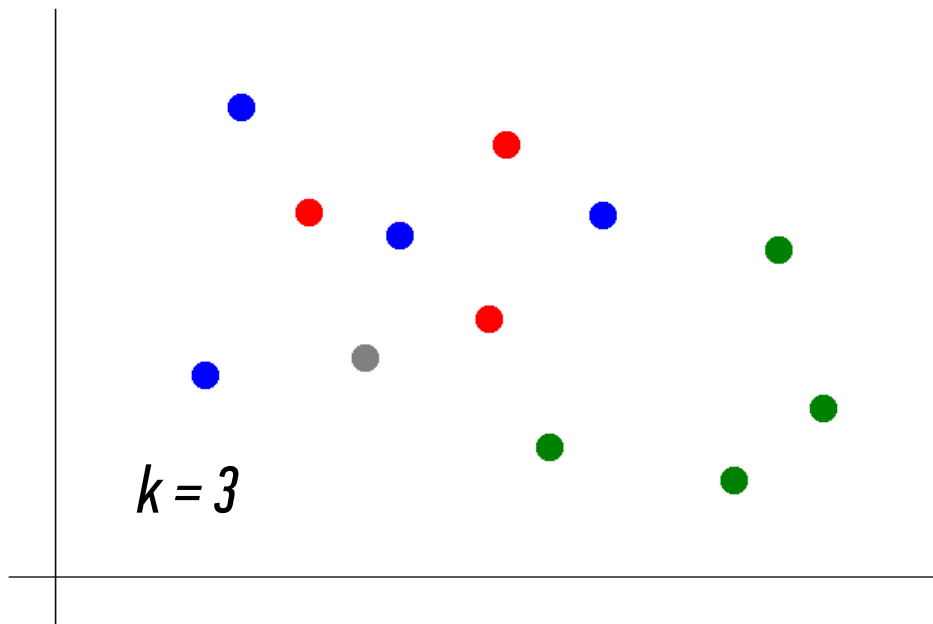
**QUESTION:**

What are the features?  
What is the response?



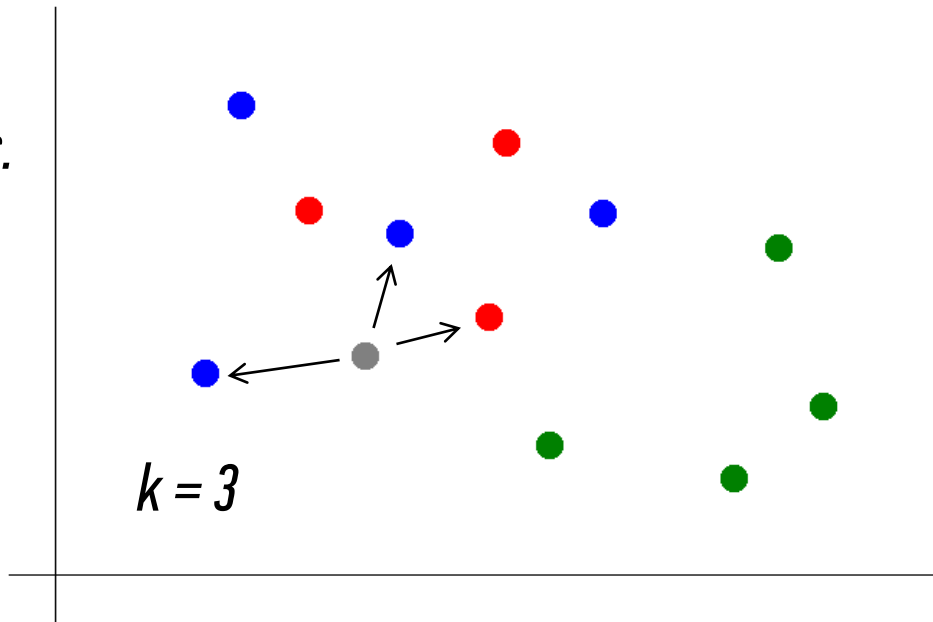
*Suppose we want to predict the color of the gray dot.*

*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.*

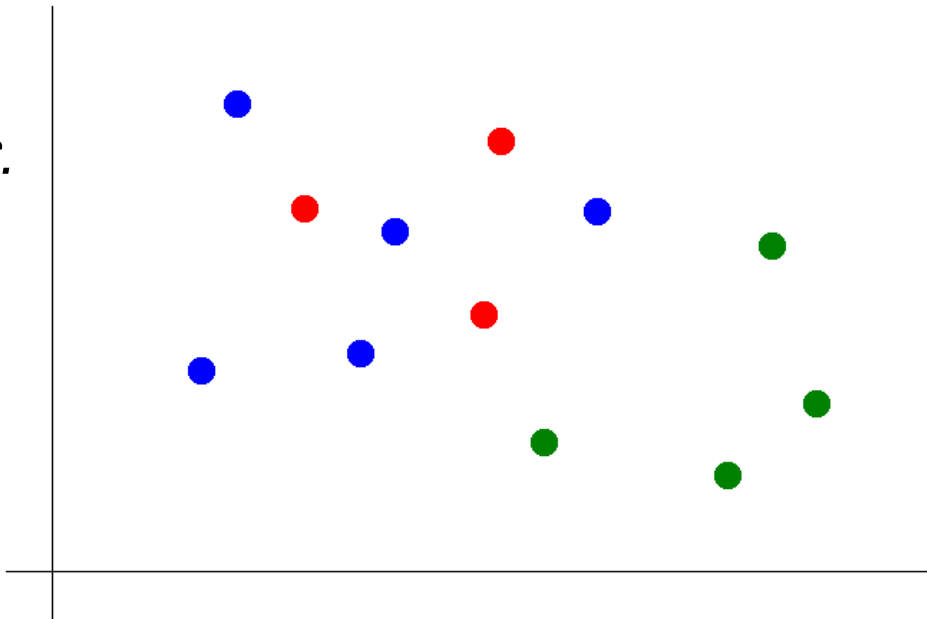


*Suppose we want to predict the color of the gray dot.*

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

**NOTE:**

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





Advantages of KNN:

- Simple to understand and explain
- Model training is fast
- Can be used for classification and regression!

Disadvantages of KNN:

- Prediction phase can be slow when  $n$  is large
- Sensitive to irrelevant features
- Accuracy is generally not competitive with the best supervised learning methods