

DATA SCIENCE

MACHINE LEARNING / KNN

I. WHAT IS MACHINE LEARNING?

II. SUPERVISED LEARNING

III. UNSUPERVISED LEARNING

IV. SUMMARY

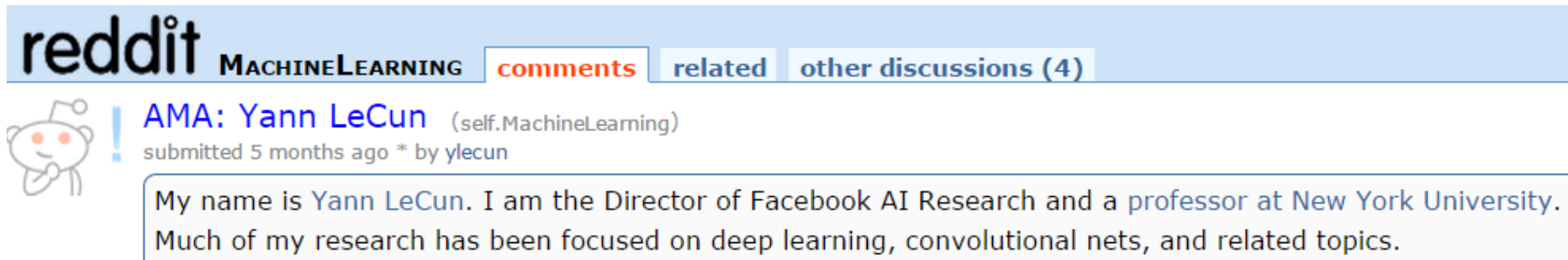
V. CLASSIFICATION WITH K-NEAREST NEIGHBORS

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



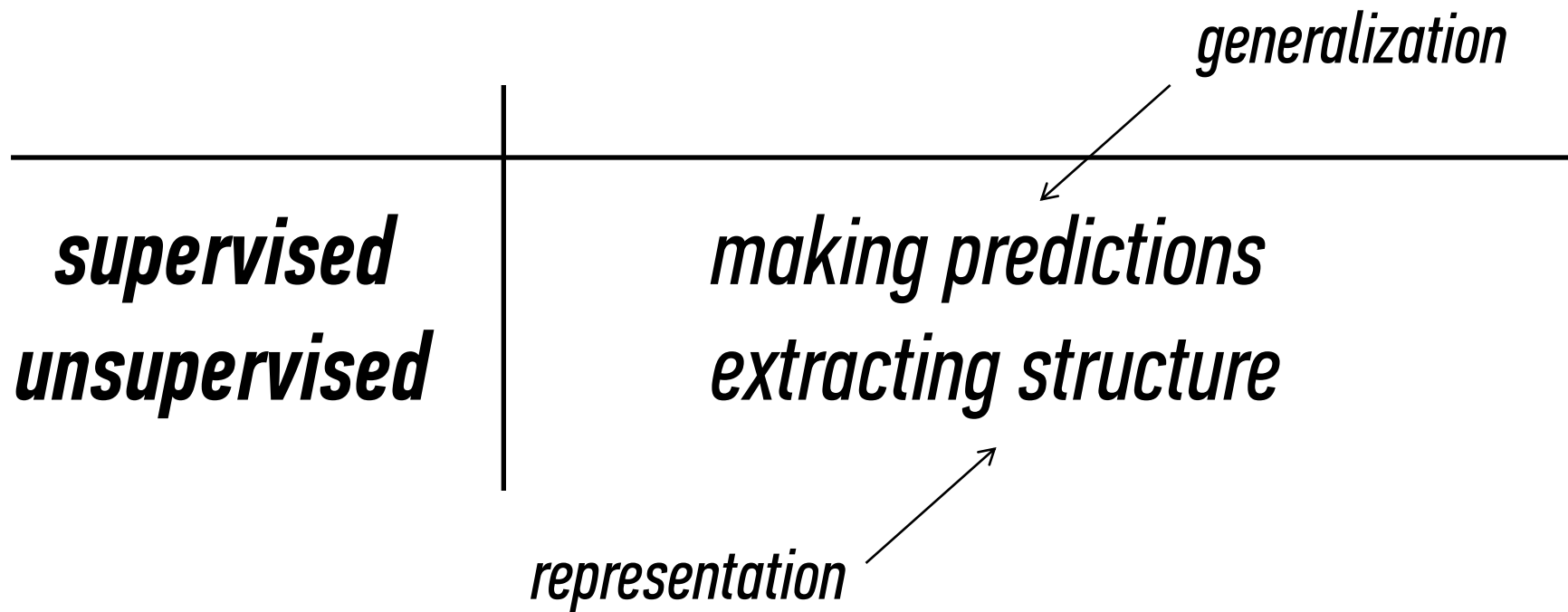
Seriously, I don't like the phrase "Big Data". I prefer "**Data Science**", which is the **automatic (or semi-automatic) extraction of knowledge from data**. That is here to stay, it's not a fad. The amount of data generated by our digital world is growing exponentially with high rate (at the same rate 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.

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



II. SUPERVISED LEARNING

- Vector of “Predictors” X
 - Also known as features, independent variables, inputs, regressors, covariates, attributes
- “Response” Y
 - Also known as outcome, label, target, dependent variable
- Regression: Y is quantitative
 - e.g., price, blood pressure
- Classification: Y is qualitative (values in a finite, unordered set)
 - e.g., spam/ham, digit 0-9, cancer class of tissue sample
- Data is composed of “observations” (predictors and the associated response)
 - Also known as samples, examples, instances, records

150 observations
($n = 150$)

Fisher's *Iris* Data

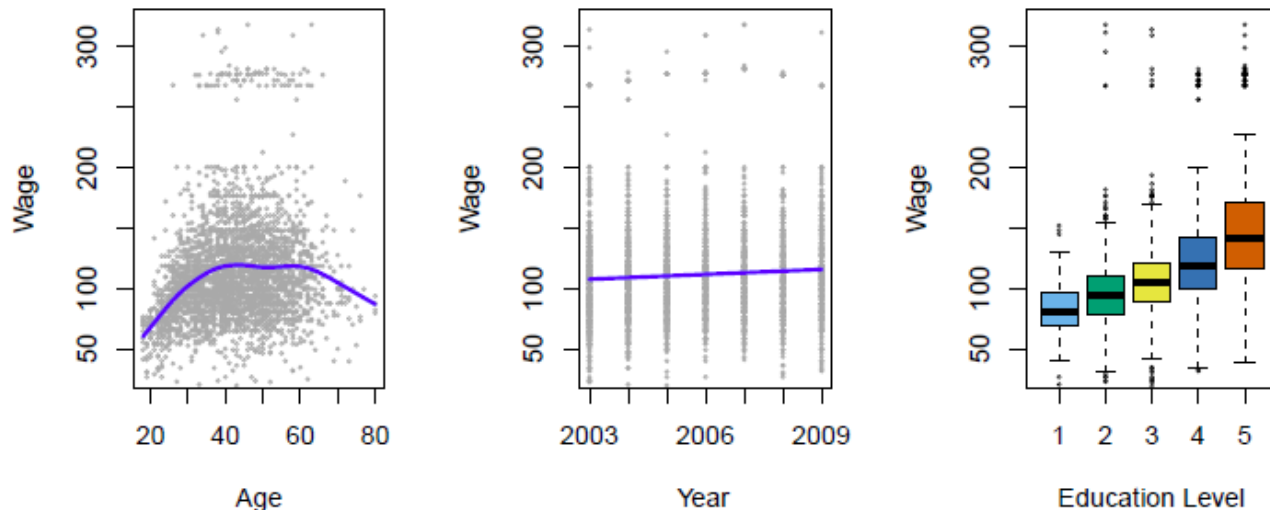
Sepal length ⬆	Sepal width ⬆	Petal length ⬆	Petal width ⬆	Species ⬆
5.1	3.5	1.4	0.2	<i>I. setosa</i>
4.9	3.0	1.4	0.2	<i>I. setosa</i>
4.7	3.2	1.3	0.2	<i>I. setosa</i>
4.6	3.1	1.5	0.2	<i>I. setosa</i>
5.0	3.6	1.4	0.2	<i>I. setosa</i>
5.4	3.9	1.7	0.4	<i>I. setosa</i>
4.6	3.4	1.4	0.3	<i>I. setosa</i>
5.0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

4 predictors ($p = 4$)

response

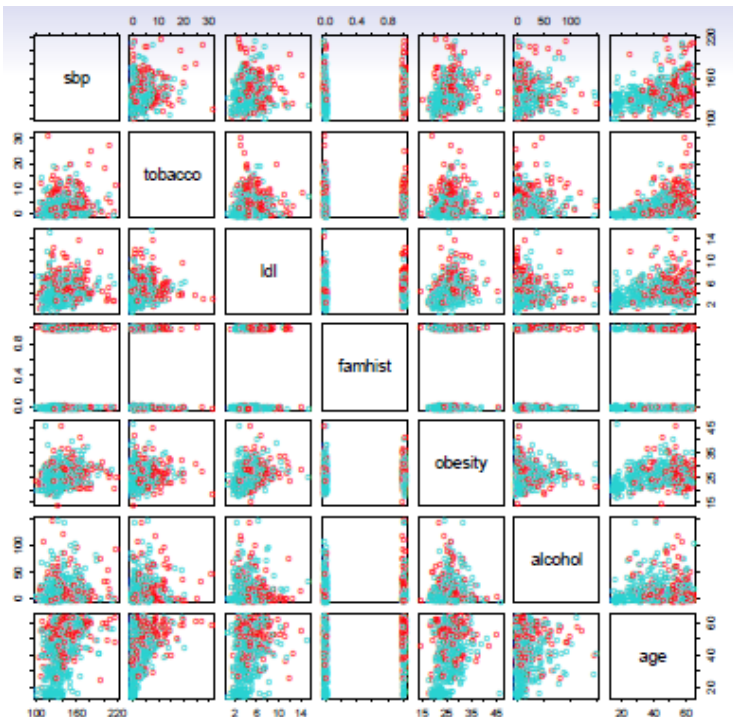
- Objectives of Supervised Learning:
 - Accurately predict unseen test cases
 - Understand which predictors affect the response, and how
 - Assess the quality of our predictions

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



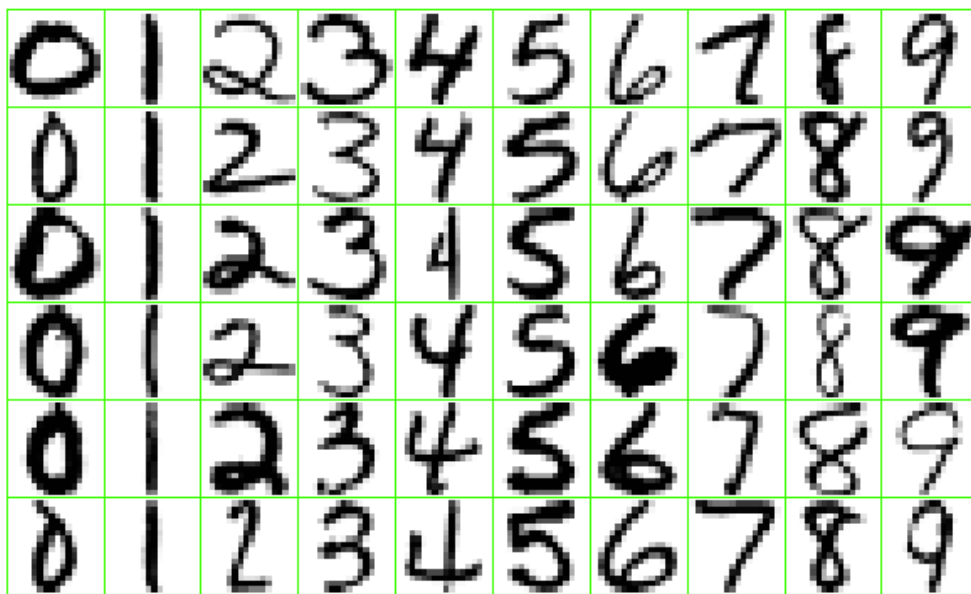
Income survey data for males from the central Atlantic region of the USA in 2009

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

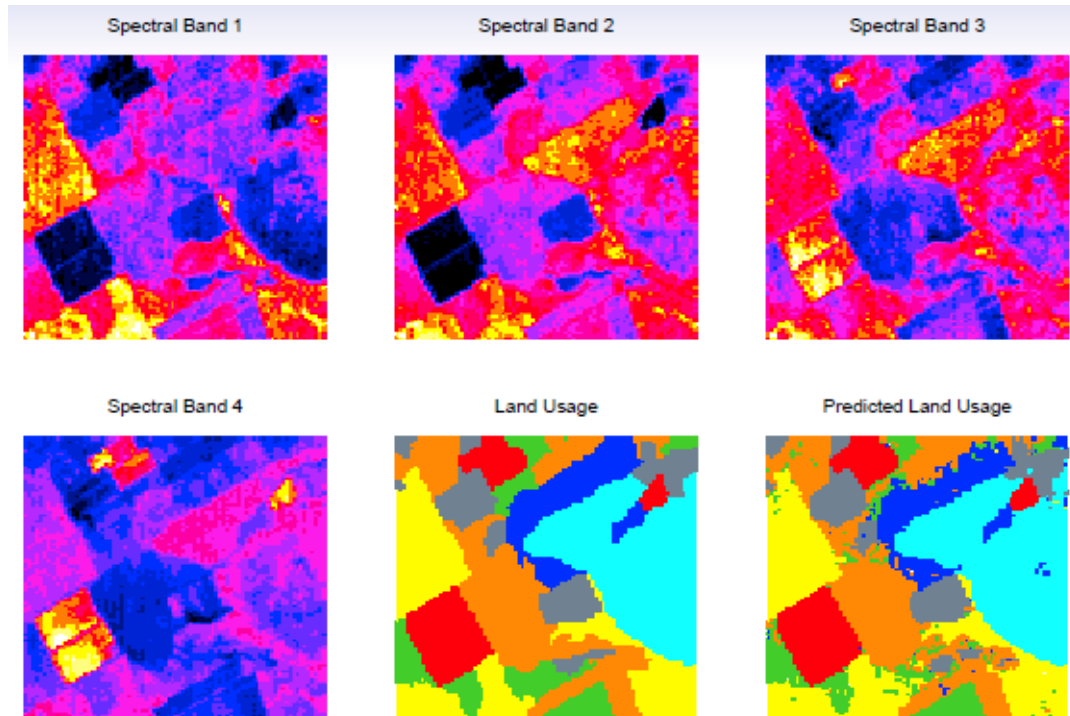


Case-control sample of men from South Africa
Red = heart disease
Blue = no heart disease

- Identify the numbers in a handwritten zip code



- Classify the pixels in a LANDSAT image, by land usage



Satellite images of rural Australia

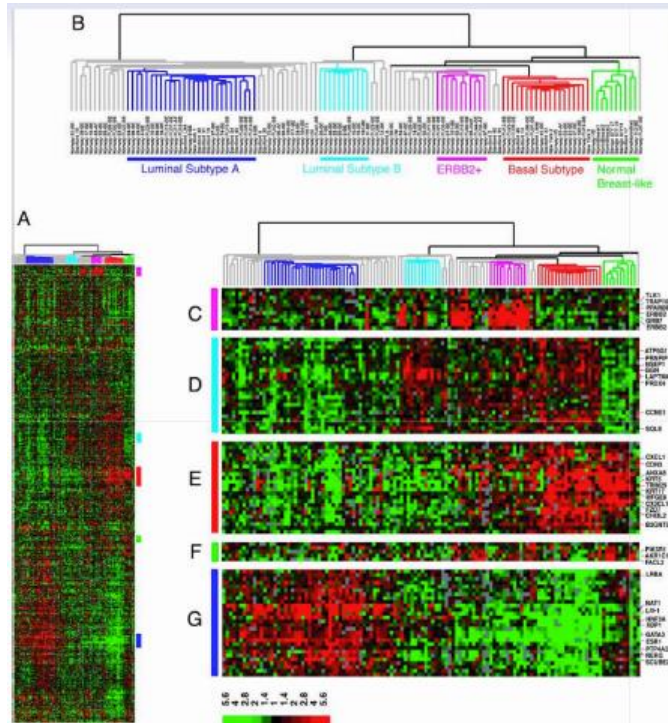
Land usage hand-labeled as one of six categories

Goal: Predict land usage using spectral bands at four frequencies

III. UNSUPERVISED LEARNING

- No response variable Y , just a set of predictors X
- Objective is more fuzzy:
 - Find groups of observations that behave similarly
 - Find predictors that behave similarly
 - Find linear combinations of features that explain most of the variation in the data
- Difficult to evaluate how well you are doing
- Data is easier to obtain for unsupervised learning since it can be “unlabeled”
- Sometimes useful as a preprocessing step for supervised learning
- Common techniques: clustering, principal components analysis

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



Gene expression data

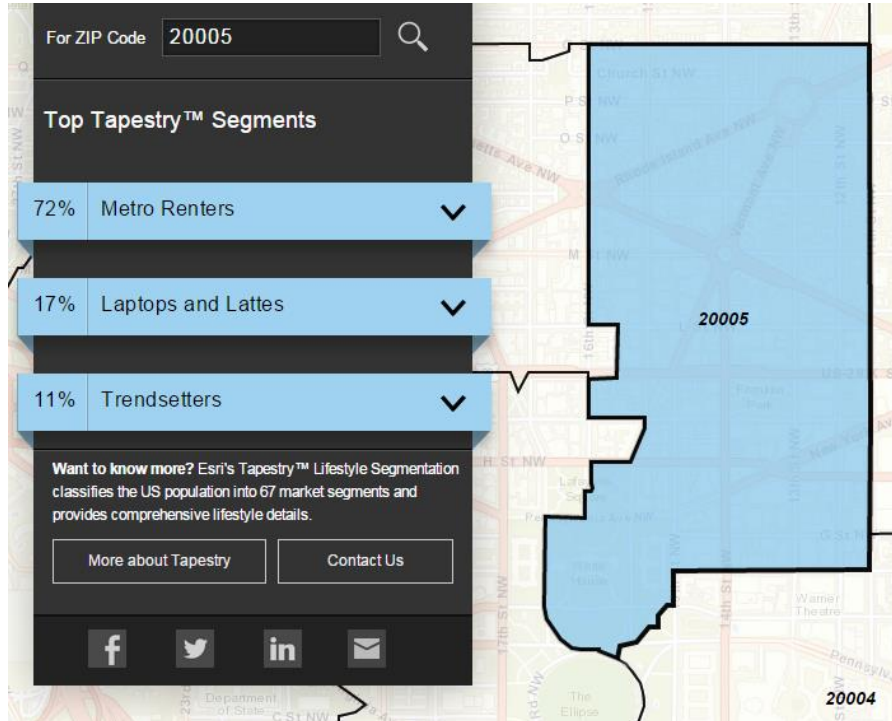
- Each row is a gene ($p=8000$)
- Each column is a woman with breast cancer ($n=88$)

Heatmap represents level of gene expression for each gene and each patient

Goal: Locate subcategories of breast cancer showing different gene expressions

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

- 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/>

IV. SUMMARY

	<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>

What type of problem is this?

Music Recommendation

It could be either.

Supervised Learning:

Predict which songs a user will
'thumbs up'

Unsupervised Learning:

Cluster songs based on
attributes and recommend songs
in the same group



- Machine learning arose as a subfield of Artificial Intelligence
- Statistical learning arose as a subfield of Statistics
- There is much overlap:
 - Machine learning has a greater emphasis on large-scale applications and prediction accuracy
 - Statistical learning emphasizes models and their interpretability, and precision and uncertainty
- The distinction has become more and more blurred
- Machine learning has the upper-hand in marketing!

V. 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>

*150 observations
($n = 150$)*

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 predictors ($p = 4$)

response

Q: How does a classification problem work?

A: Data in, predicted labels out.

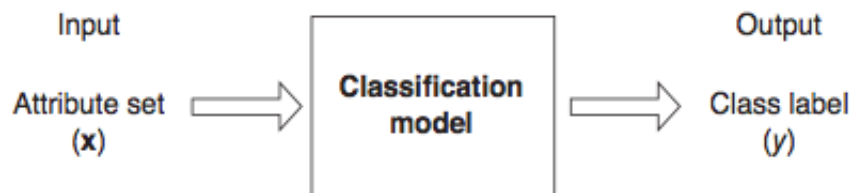
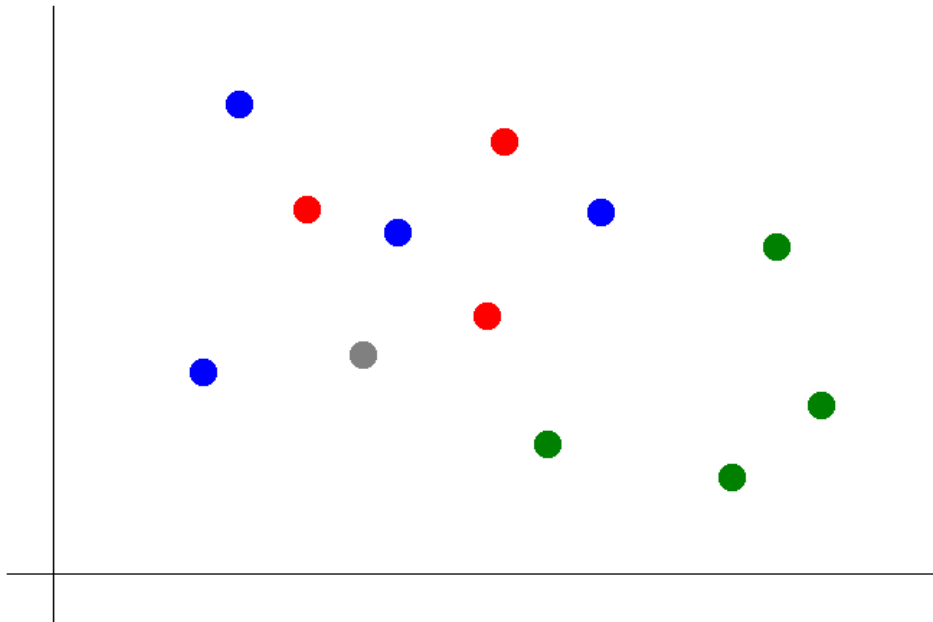


Figure 4.2. Classification as the task of mapping an input attribute set x into its class label y .

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

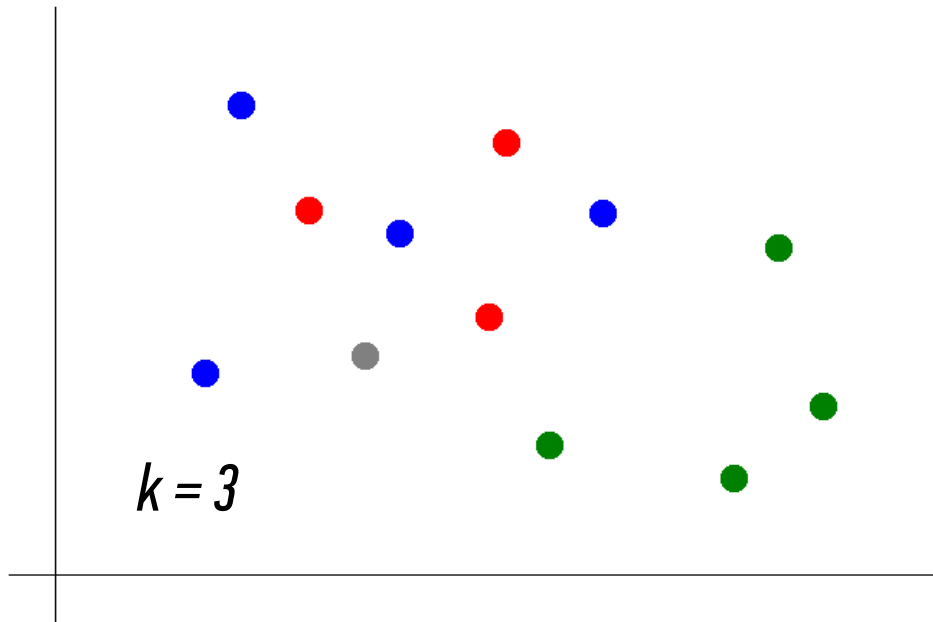
QUESTION:

What are the predictors?
What is the response?



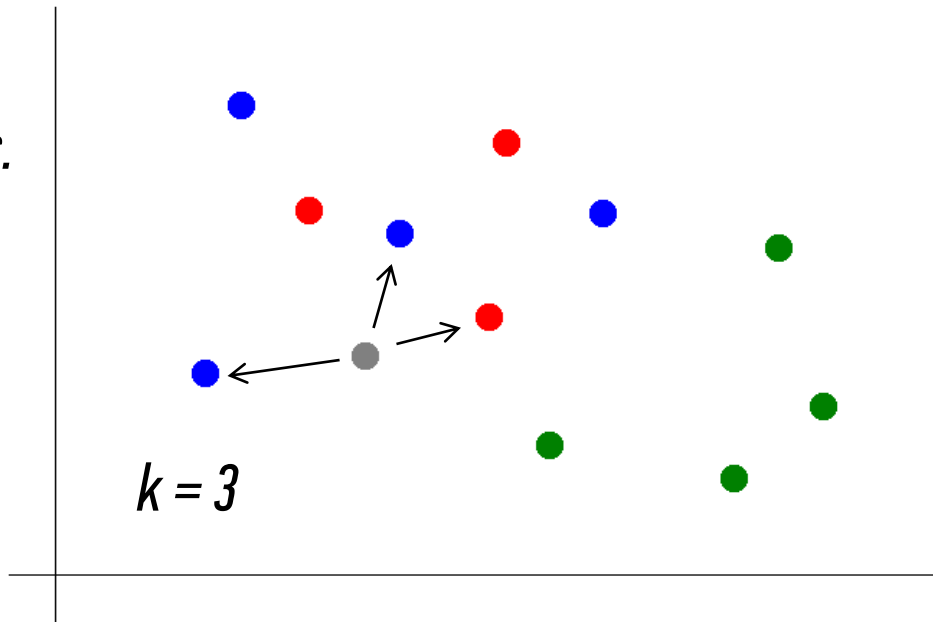
Suppose we want to predict the color of the gray dot.

1) Pick a value for k .



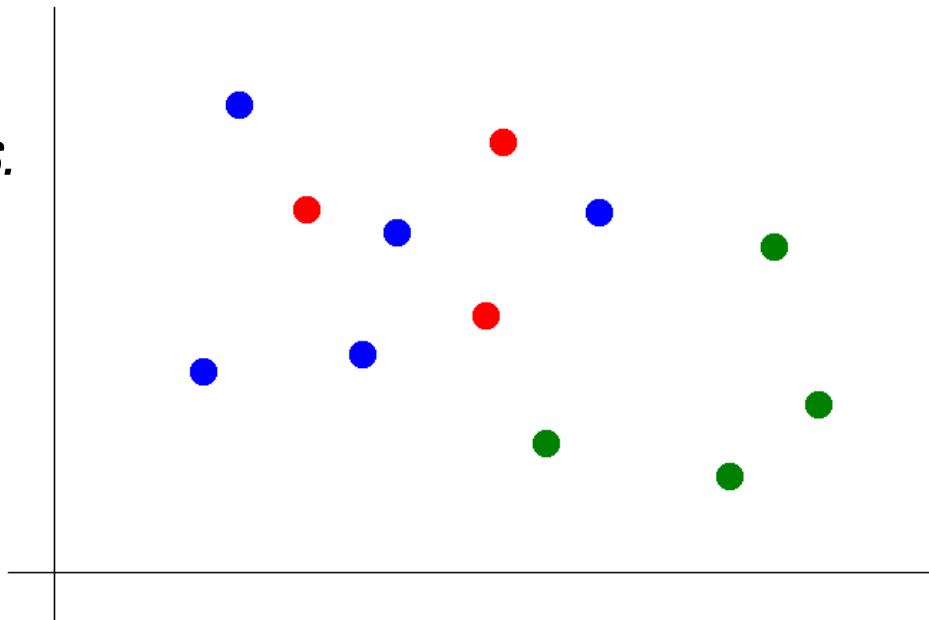
Suppose we want to predict the color of the gray dot.

- 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 grey dot.*

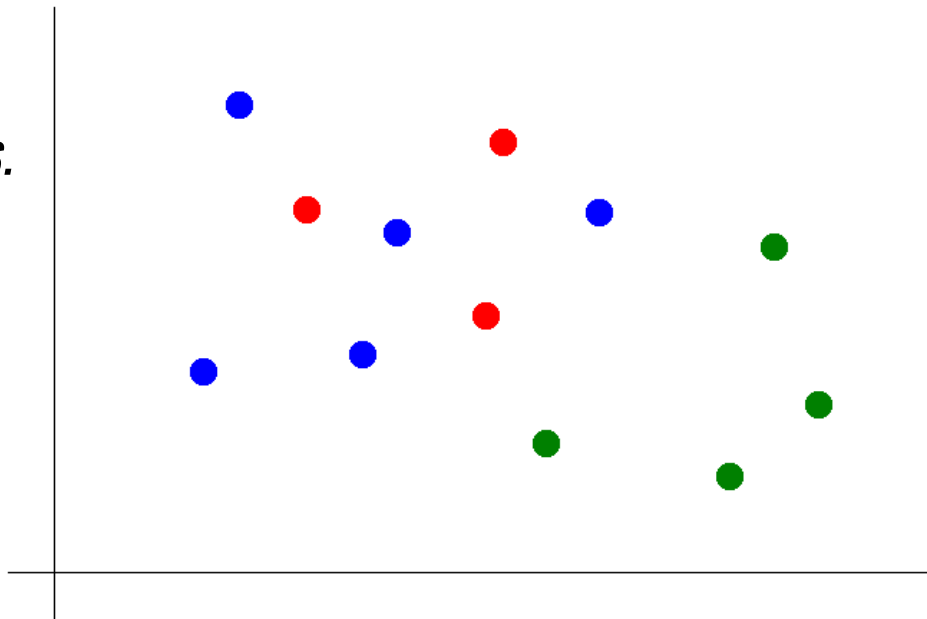


Suppose we want to predict the color of the gray dot.

- 1) Pick a value for k .*
- 2) Find colors of k nearest neighbors.*
- 3) Assign the most common color to the grey 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
- Parametric methods will be better than KNN if the selected form is close to the “true” form
- Parametric methods tend to be better than KNN for small n and large p
- Does not provide coefficients or p -values
- Sensitive to irrelevant features
- Sensitive to unscaled data

DATA SCIENCE
