

# 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?

## 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?

---

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



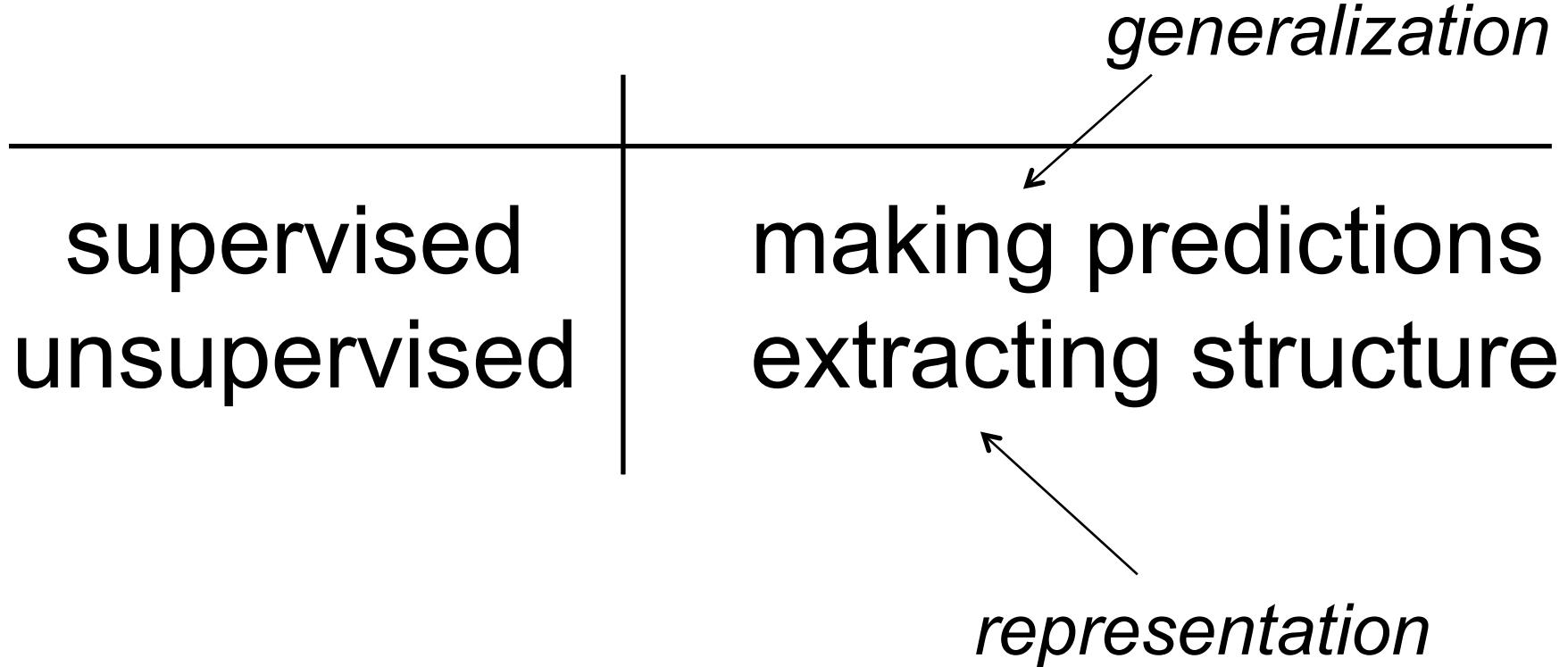
Face

# Test



Face?

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



supervised  
unsupervised

*generalization*  
making predictions  
extracting structure

**Supervised: labeled data**

**Unsupervised: unlabeled data**

*representation*

# TYPES OF MACHINE LEARNING PROBLEMS

9

supervised  
unsupervised

*generalization*  
making predictions  
extracting structure

**Supervised: labeled data**

**Unsupervised: unlabeled data**

**Previous example  
was supervised!**

*representation*

# **II. SUPERVISED LEARNING**

- Vector(list) of “Predictors” X
  - Also known as features, independent variables, inputs, regressors, covariates, attributes
- “Response” y
  - Also known as outcome, label, target, dependent variable
- If y is continuous: **Regression**
  - e.g., price, blood pressure
- If y is categorical (values in a finite, unordered set): **Classification**
  - 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

## EXAMPLE #1: PREDICTING NEONATAL INFECTION

**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

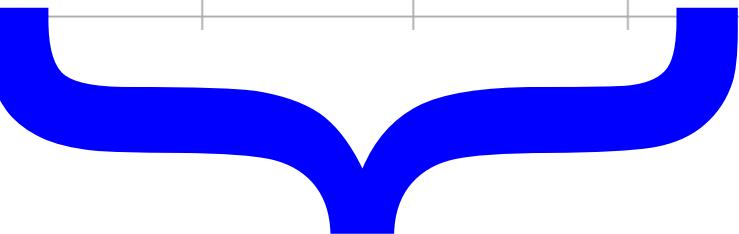


150  
observations  
 $(n = 150)$



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 ( $p = 4$ )



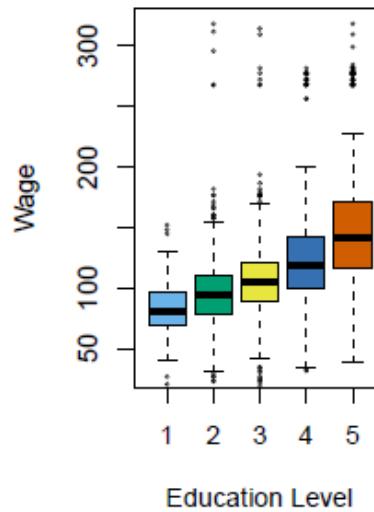
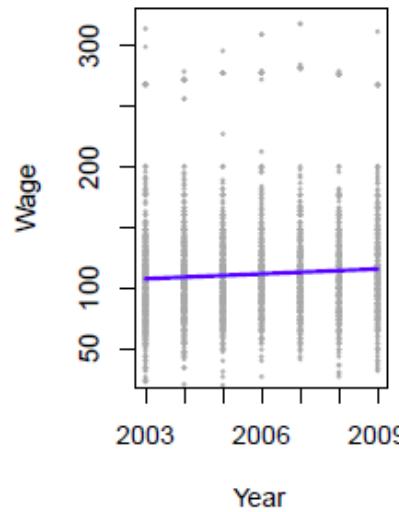
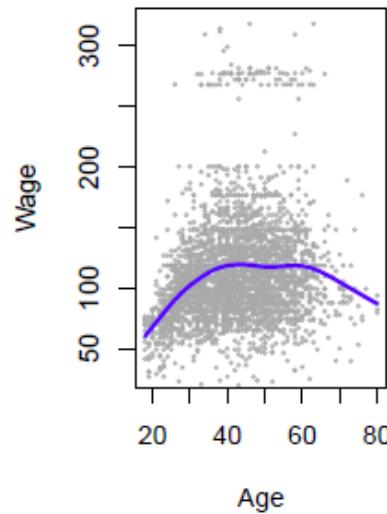
response

- Supervised Learning uses known/labeled “**training** cases” in order to:
  - Accurately predict unseen **test** cases
  - Understand which predictors affect the response, and how
  - Assess the quality of our predictions

# REGRESSION EXAMPLE

15

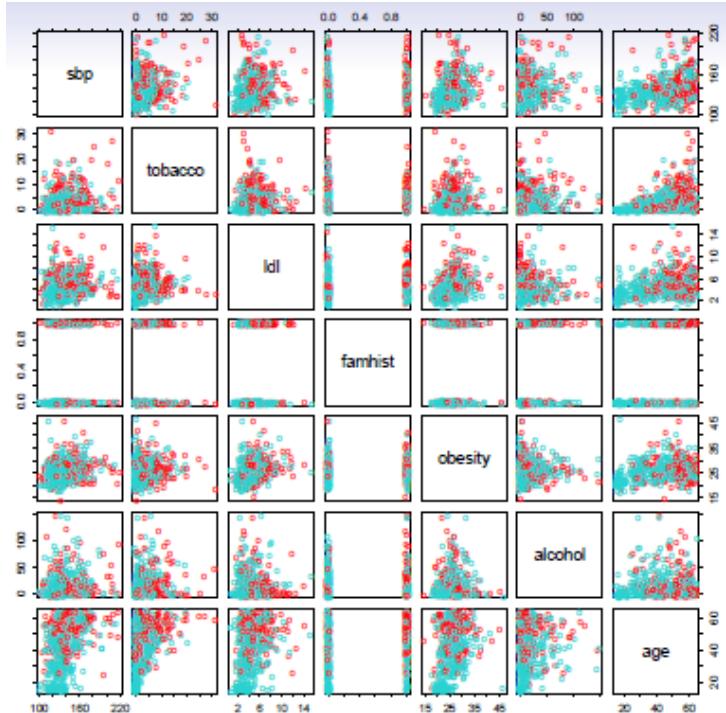
- Establish the relationship between salary and demographic variables in population survey data



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

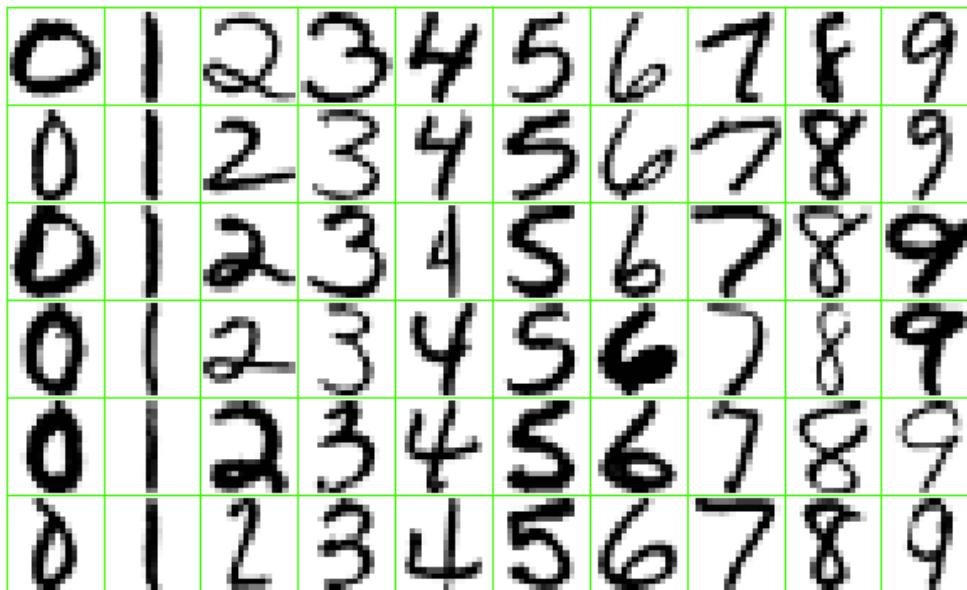
## CLASSIFICATION EXAMPLE

- 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



# CLASSIFICATION EXAMPLE: KAGGLE

kaggle

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).

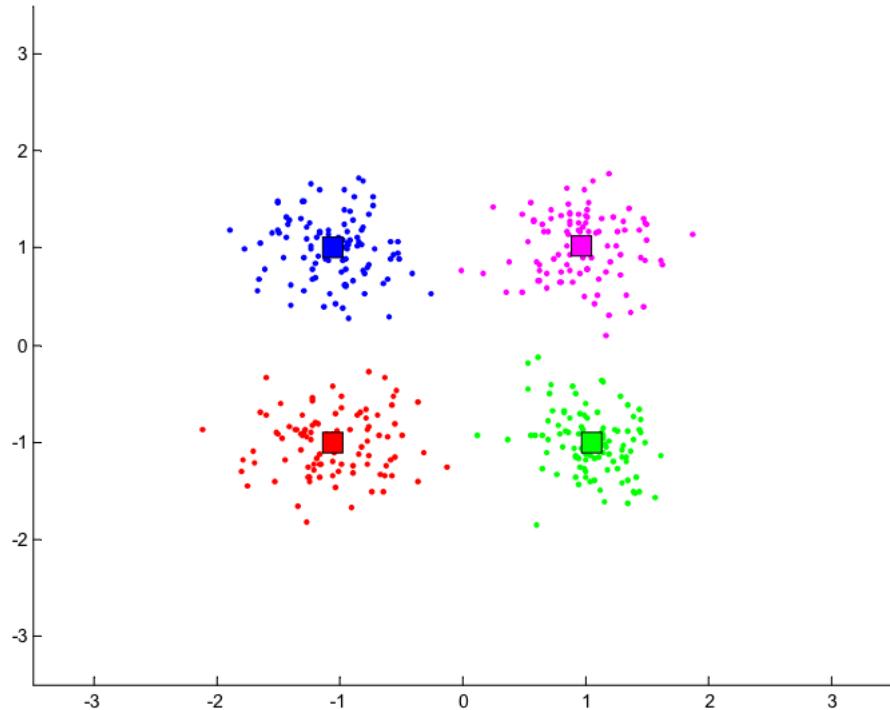
# **III. UNSUPERVISED LEARNING**

- No response variable  $y$ , just a set of predictors  $X$
- Objective is more open:
  - Find groups of observations that behave similarly
  - Find predictors that behave similarly
  - Find combinations of features that explain 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” (i.e., it hasn’t been labeled with a response)
- Sometimes useful as a preprocessing step for supervised learning
- Common techniques: clustering, principal components analysis

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

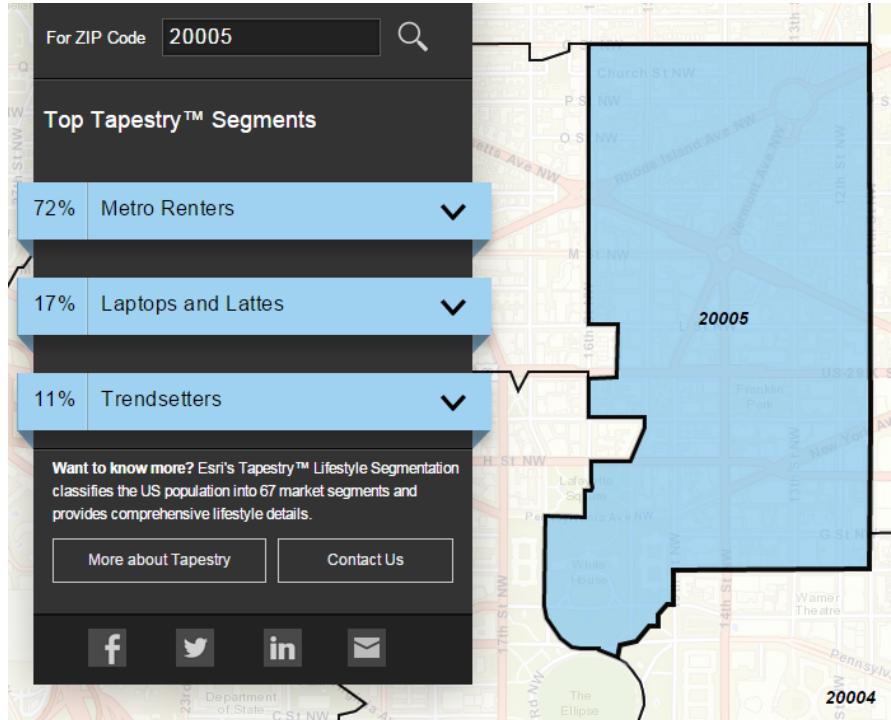
# CLUSTERING EXAMPLE

23



## CLUSTERING EXAMPLE

- 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/>

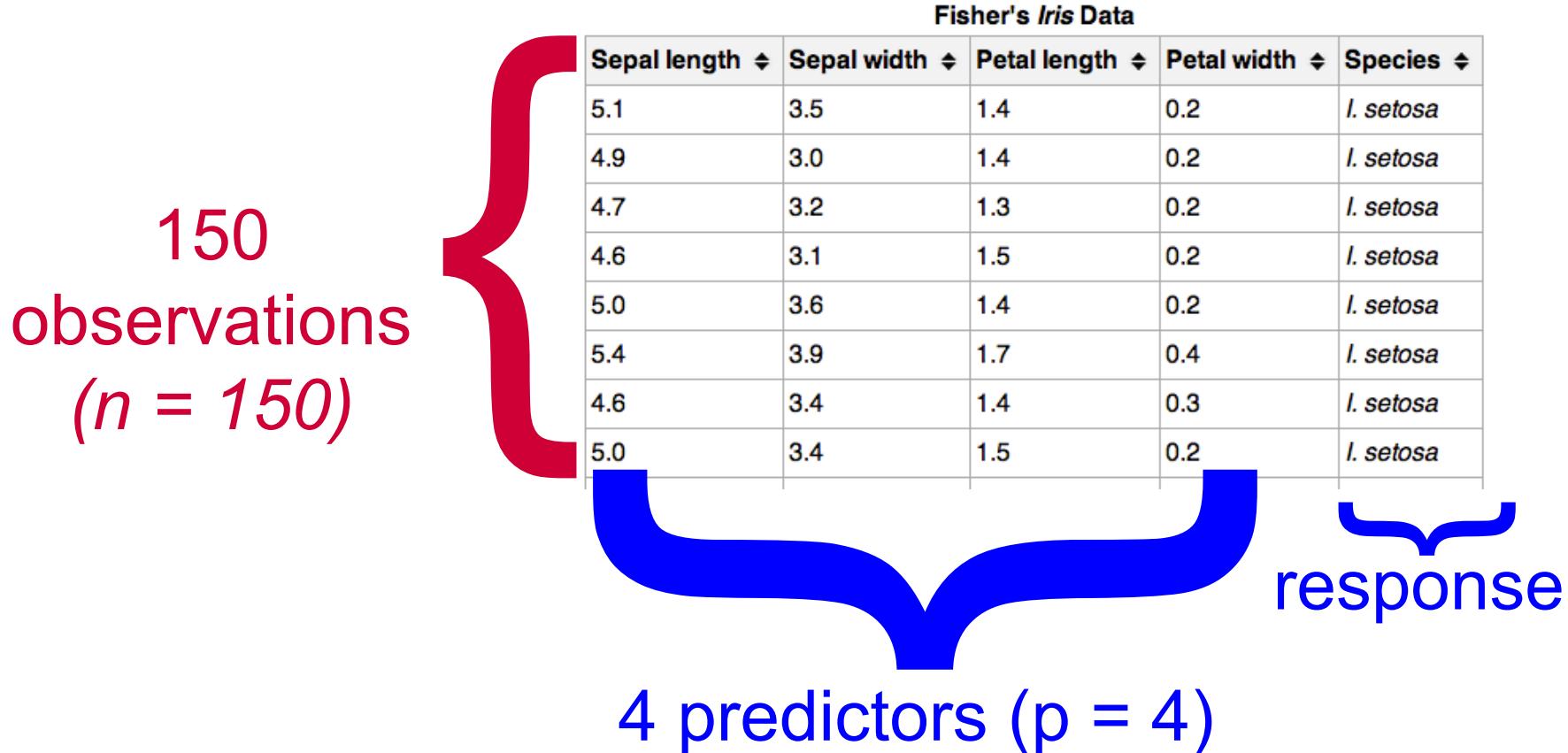
# **IV. SUMMARY**

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

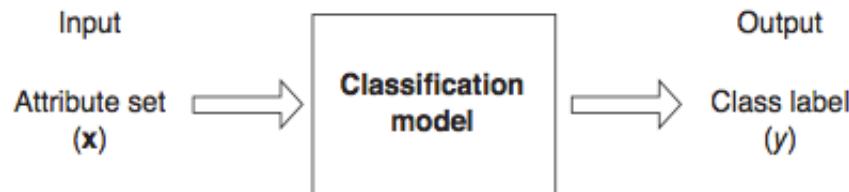
# V. CLASSIFICATION WITH K-NEAREST NEIGHBORS

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering



Q: How does a classification problem work?  
A: Data in, predicted labels out.

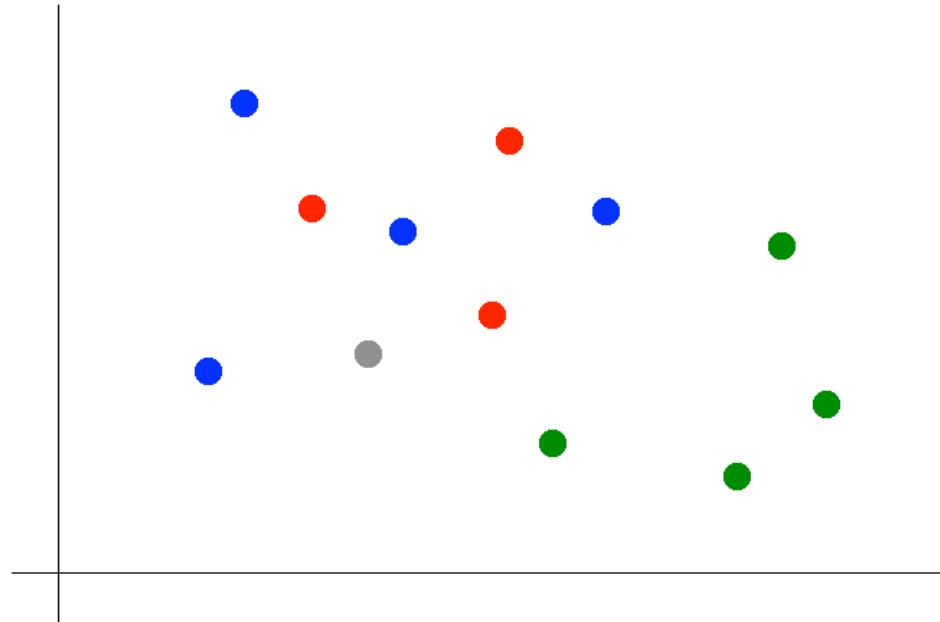


**Figure 4.2.** Classification as the task of mapping an input attribute set  $x$  into its class label  $y$ .

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

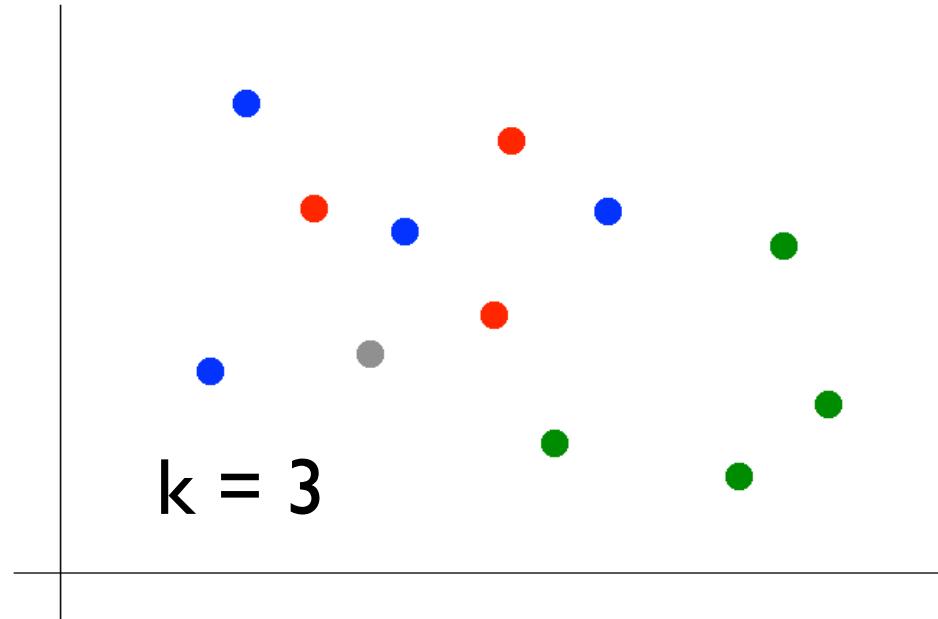
**QUESTION:**

What are the predictors?  
What is the response?



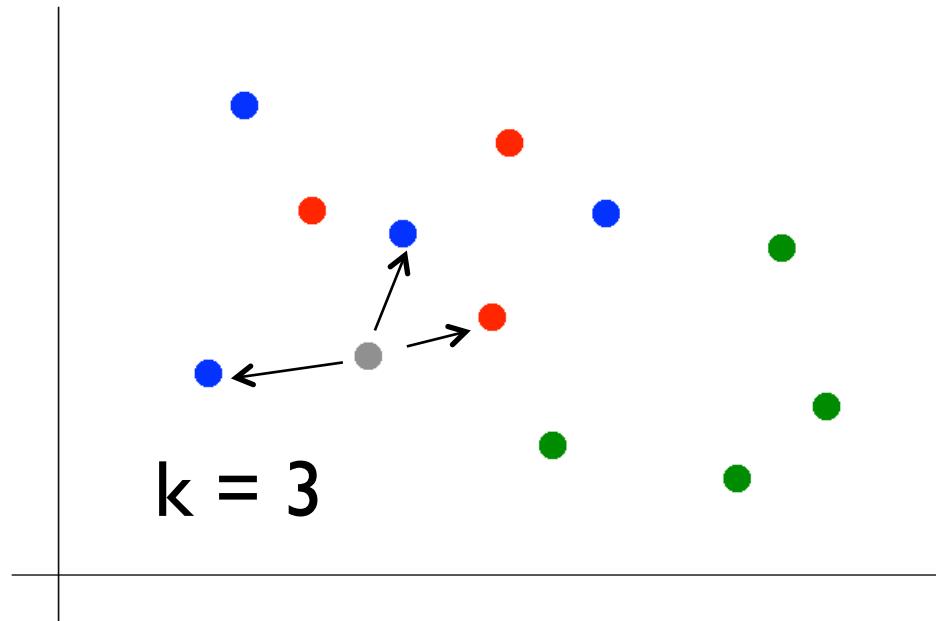
Suppose we want to predict the color of the gray dot.

- 1) Pick a value for k.



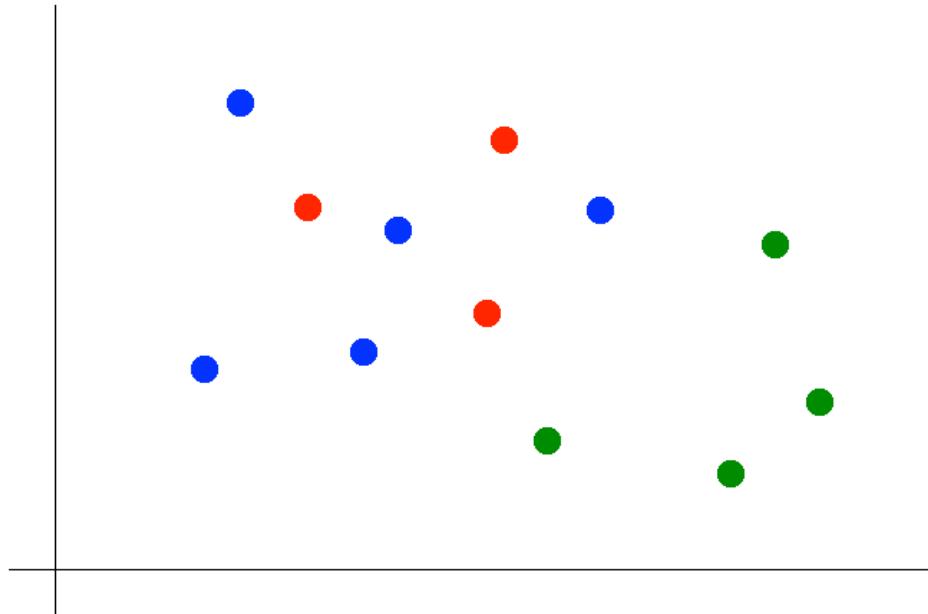
Suppose we want to predict the color of the gray dot.

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



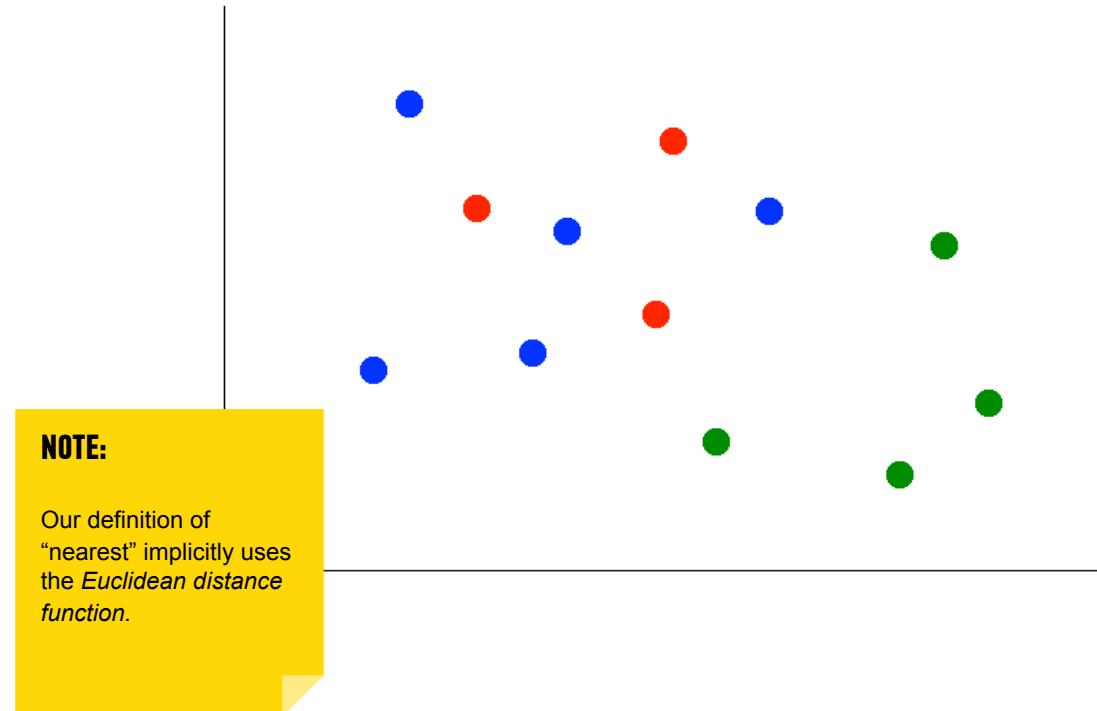
Suppose we want to predict the color of the gray dot.

- 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.



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

---

# DATA SCIENCE

---