THE UNIVERSITY OF MEMPHIS

COMP7/8150:
Fundamentals of
Data Science

# Week 10: Machine Learning/ Supervised Learning

Max H. Garzon

# Standard Methodology

- **Problem Definition/goal**
  - Identify/specify goals of the data analysis
  - commit to specific deliverables
- **Data pre-processing**
  - Identify appropriate data
  - Acquire data (gather, lookup, understand)
- **Data processing**
  - Identify methods (gather, cleanse, store)
  - Carry out the analysis (patterns, trends, predictions?)
- **Data post-processing**
  - Visualize and present
  - Deploy and evaluate. Iterate, if necessary

# Learning Objectives

- To identify Machine Learning (ML) and define a framework for its methods
- To identify most common ML supervised algorithms
- To identify required steps for model fitting and model evaluation.
- To view case studies to gauge the degree of versatility and success of ML methods

# Example Application 1: Classify

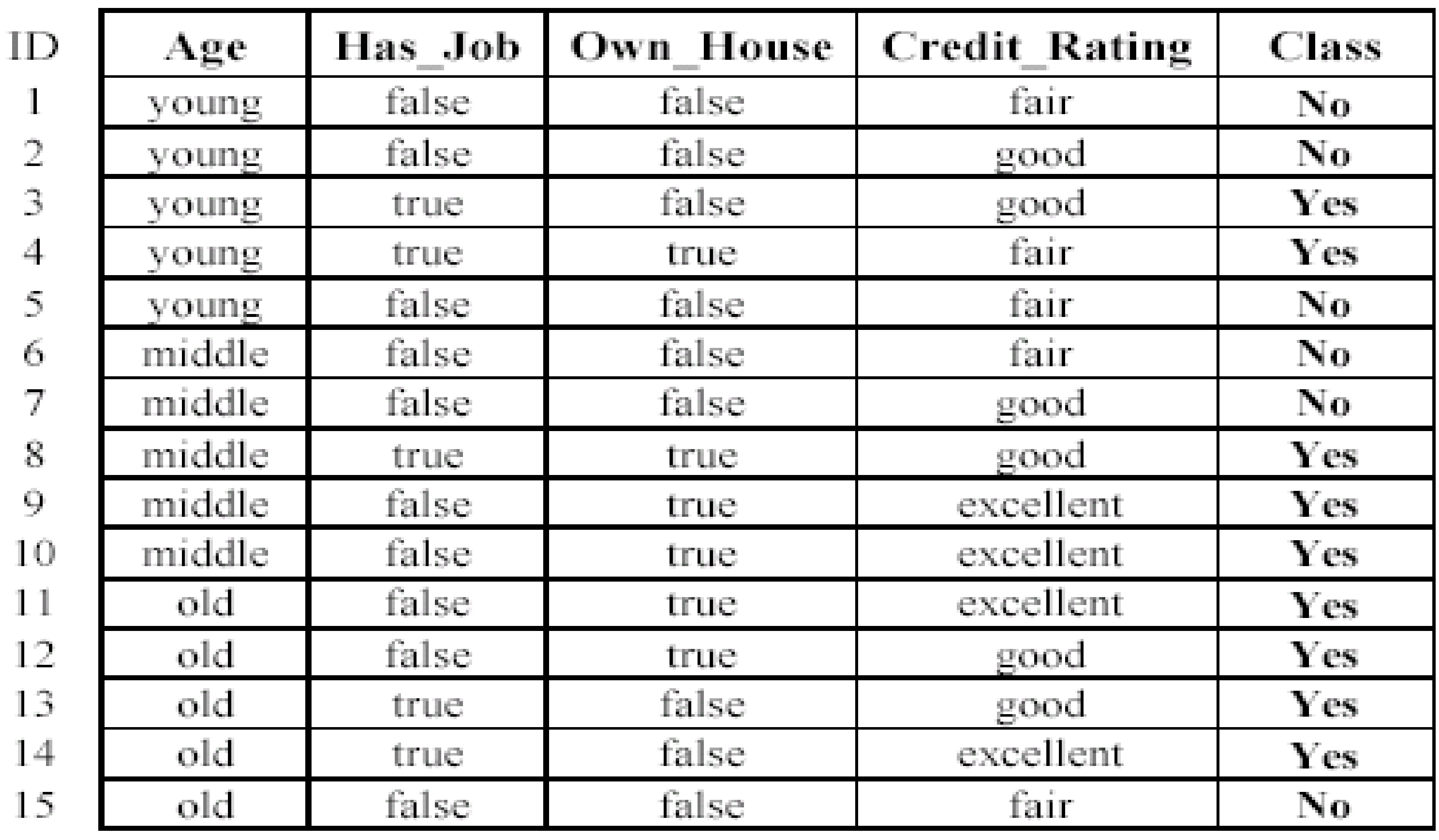
- An emergency room in a hospital measures 17 variables (e.g., blood pressure, age, etc.) of newly admitted patients.
- A decision is needed: Do they need put a new patient in an intensive-care unit (ICU)?
- Due to the high cost of ICU, those patients who may survive less than a month are given higher priority.
- Problem: to predict high-risk patients and discriminate them from low-risk patients.

# .. Example App 2: Classify

- A bank receives thousands of applications from potential clients. Each application contains information about an applicant, such as
  - annual salary
  - age
  - outstanding debts
  - credit rating
  - marital status
- Problem: How to decide whether an application should be approved, i.e., to classify applications into two categories, approved/decline. Obviously, Co wants to minimize the risk of defaults.

# .. App2: Data for loan application

Approved or not

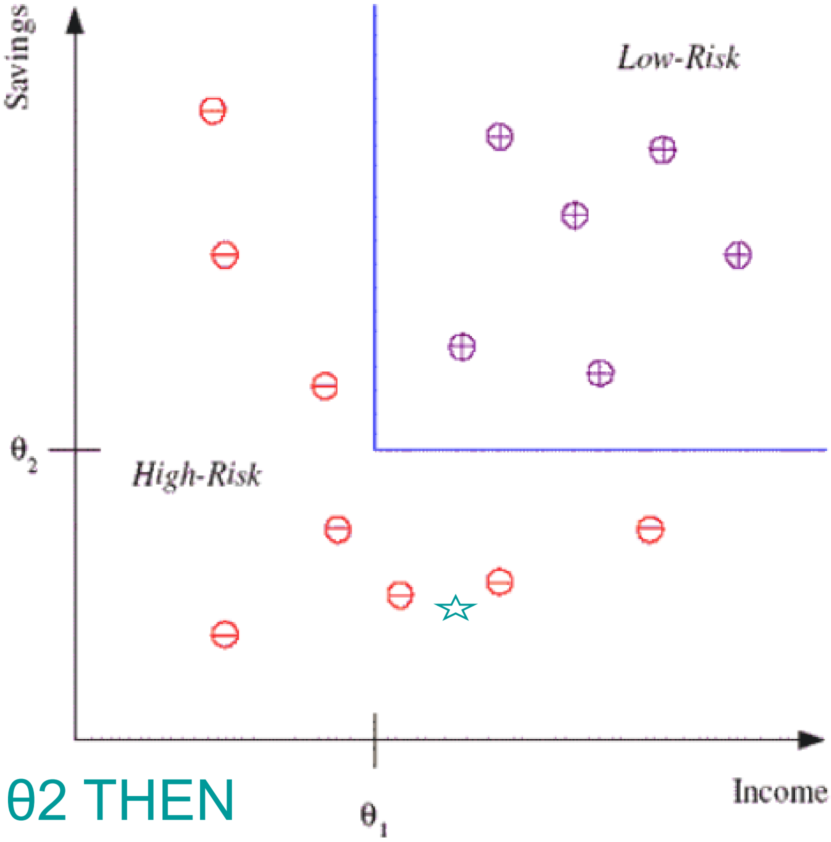


| ID | Age | Has_Job | Own_House | Credit_Rating | Class |
|---|---|---|---|---|---|
| 1 | young | false | false | fair | No |
| 2 | young | false | false | good | No |
| 3 | young | true | false | good | Yes |
| 4 | young | true | true | fair | Yes |
| 5 | young | false | false | fair | No |
| 6 | middle | false | false | fair | No |
| 7 | middle | false | false | good | No |
| 8 | middle | true | true | good | Yes |
| 9 | middle | false | true | excellent | Yes |
| 10 | middle | false | true | excellent | Yes |
| 11 | old | false | true | excellent | Yes |
| 12 | old | false | true | good | Yes |
| 13 | old | true | false | good | Yes |
| 14 | old | true | false | excellent | Yes |
| 15 | old | false | false | fair | No |

# .. App2 solution: Naïve Classifier

**Algorithm**

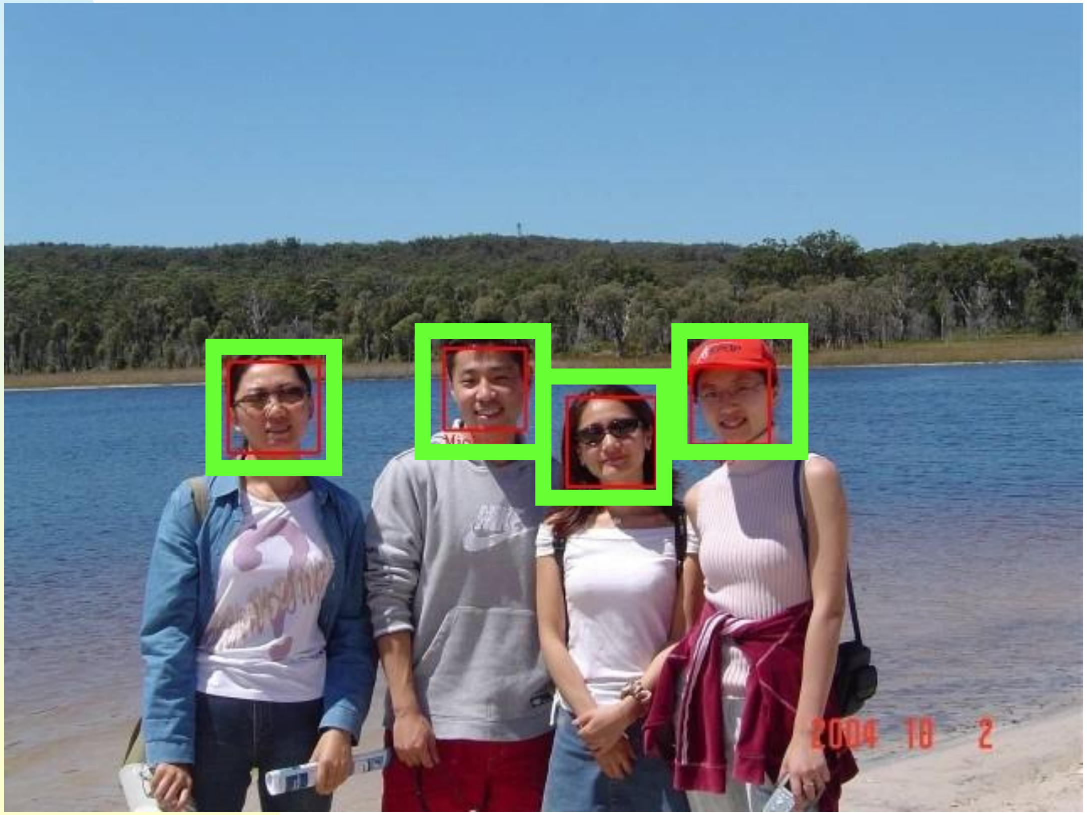
Discriminate low risk and high risk customers solely by their income and savings features



```
IF income > $\theta_1$ AND savings > $\theta_2$ THEN
        low-risk
ELSE
        high-risk
```

# .. App 3: Face Recognition

- Deciding whether a human face appears in a picture
- Deciding whether a picture contains me



# .. App 4: Signature Recognition

- Is that my signature?

  Structural similarities are difficult to quantify.

# Classification Problems

- Given a universe Ω and a partition Π (some disjoint groups/categories whose union exhaust Ω)
  CLASSIFICATION PROBLEM (Π)
  Instance: an element of x
  Question: which part/cat in Π does x belong to?
- A classifier is a solution to the classification problem, i.e., it places each input feature vector x into one of the parts/categories in Π.
- There are many types of classifiers:
  - Statistical (e.g. Gaussian)
  - Perceptrons / Support-Vector Machines (SVMs)
  - Feed-Forward Neural Networks (FNNs)

# Prediction Problems

- Given a function f: Ω → Y on a popualtion Ω
  PREDICTION PROBLEM(f)
  Instance: an element x of Ω
  Question: what is the value of f at x?
- A predictive model is a solution to the prediction problem, i.e., an algorithm that produces a (good approximation of) the value f(x) for every (or most) of the instances x in the population Ω.
- There are many types of predictive models:
  - Regression
  - Neural Networks (feed-forward, self-organizing)

# Clustering Problems

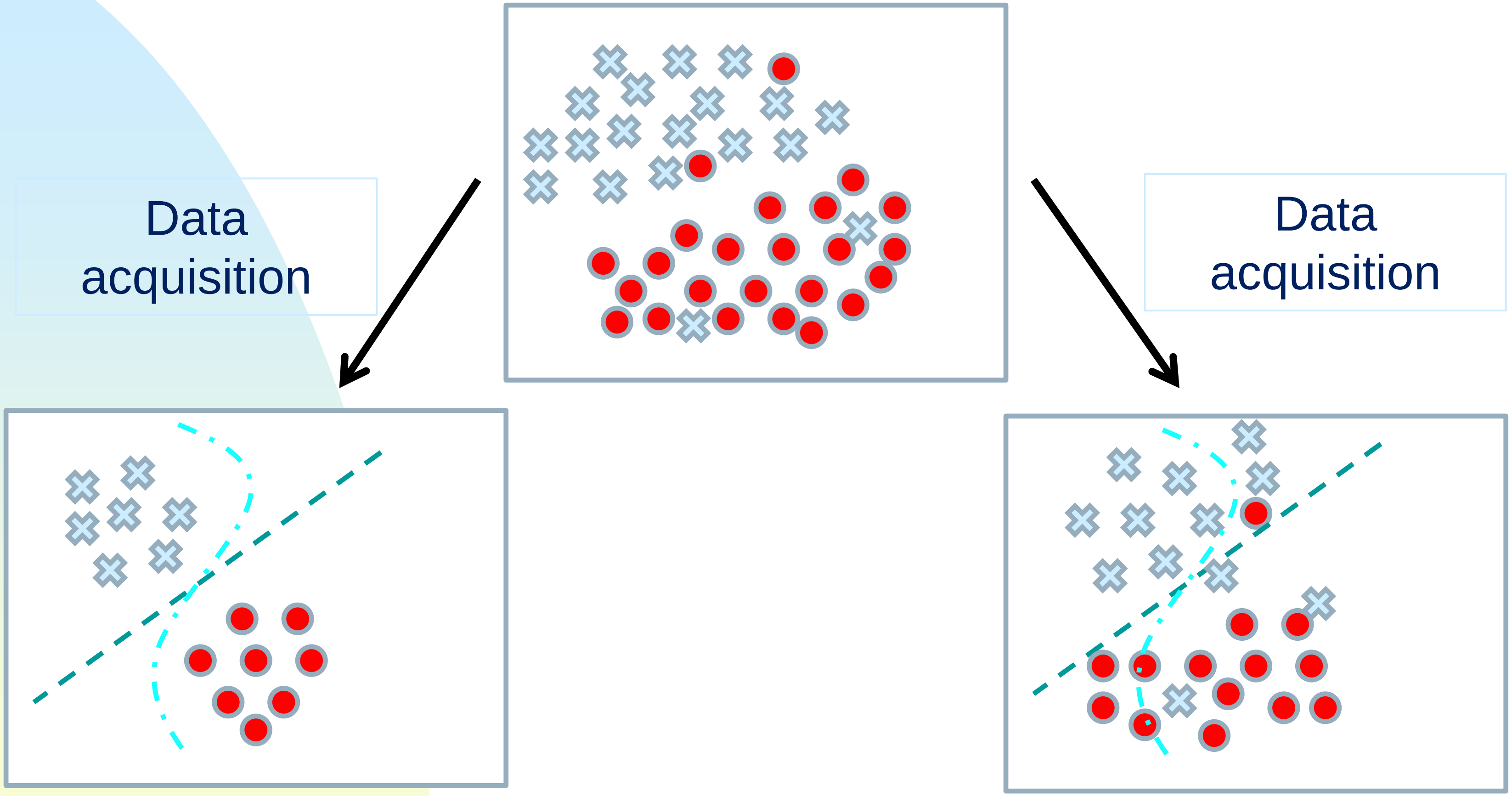
- CLUSTERING PROBLEM (Ω,*m*)
  Question: what is a partition Π of Ω where elements more similar according to measure *m* are put into the same part (cluster)
- A clustering algorithm is a solution to the clustering problem, i.e., an algorithm that produces the clusters in a partition Π of Ω.
- There are many types of clustering algorithms, primarily when m is a notion of distance:
  - Hierarchical
  - k-Means

# What is Machine learning (ML)?

- Given
  - a data set *D*,
  - a task *T*, and
  - a performance measure *M*,

  an algorithm/program P is said to **learn** from *D* to perform the task *T* if upon (repeated) execution, P’s performance on instances in D improves as measured by *M*.
- In other words, A helps a computer system perform *T* better as compared to a nonself-modifying program (Mitchell, 1997)

# .. What is Machine Learning?



Universal set

Data acquisition

Data acquisition

Training set
(observed)

Testing set
(holdout)

Practical use
X-validation set
(unobserved)

# .. What is Machine Learning?

- ML is about building artificial models / algorithms / software capable of
  - learning from past experiences (usually data/examples/cases, labeled or unlabeled )
  - Improving their performance as a result

- Some of these gadgets can be regarded as artificially intelligent programs
- Many of them follow closely methods observed to work in biological organisms
- Very different from traditional models (e.g statistical, where a prior analysis by experts is required)

# ML: Data and goal

- Data: A set of data records (also called *exemplars*, *instances*, or *cases*) described by
  - *A feature vector of n* attributes: ($x_1$, $x_2$, … $x_n$)
  - *a class*: Each example can (my not) be labeled with a pre-defined class.
- GOAL: produce a learning algorithm to classify the given data and exhibits good

  Generalization:

  it scales well if asked to predict classes of new instances (future, or test) that the algorithm has never seen before.

# ML: Un/Supervised Learning

- Supervised learning: classification is seen as supervised learning from exemplars.
  - Supervised: the algorithm is shown data (observations, measurements) including labels of the corresponding class, i.e., a “teacher” is available to tell the answer
  - Test data samples are classified into these classes too.
- Unsupervised learning (clustering)
  - Class labels of the data are unknown
  - Given a set of data, the task is to establish the existence of classes or clusters in the data

# ML: Supervised learning phases

a. Acquire Data:

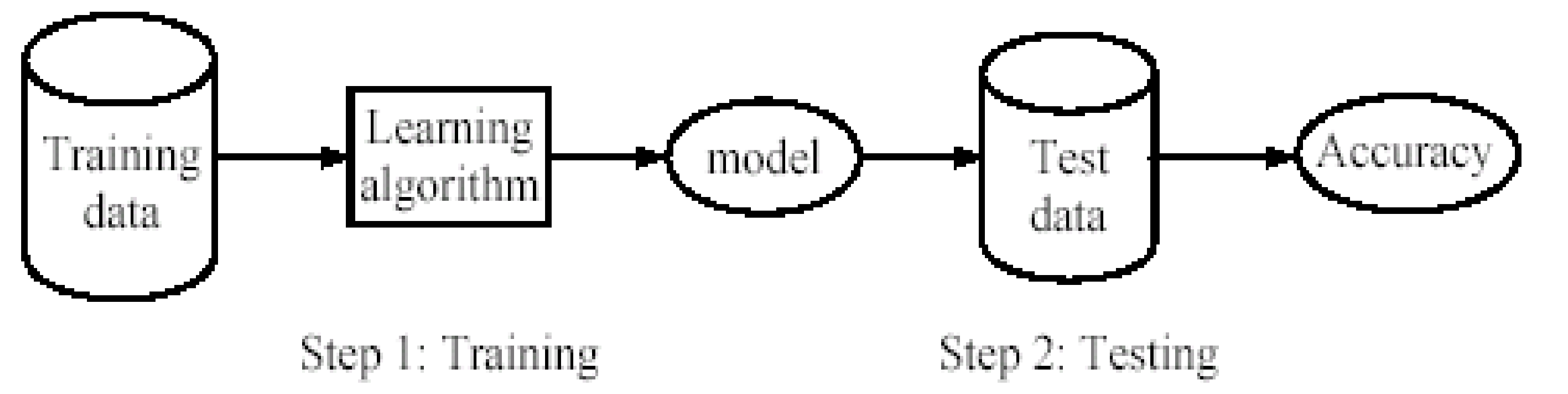
Training data: to show a model

Testing data: To test the model using test data unseen by the model to assess its accuracy

b. Learning Algorithm: to search for a good model

c. Validation: To use it to make sound predictions

$$Accuracy = \frac{\text{Number of correct classifications}}{\text{Total number of test cases}},$$



# Fundamental Requirement

- To eventually achieve good accuracy,
  the data corpus must be R$^3$S of all the population:
  - Reliable/Consistent
    real-world data has inherent consistency checks
    (and so must synthetic data to be useful, much harder)
  - Representative
    must include balanced points from all corners of the population
  - Relevant
    be about the problem, the whole problem and nothing but the problem
  - Sufficient
    contains sufficiently many points to be representative and statistically significant re: the population (else no generalization.)
- In practice, these requirements are rarely fully true

  Deviations will likely result in poorer solutions

# Questions?