Departamento de Ciencia de la Computación (DC-UCSP)

Maestría en Ciencia de Datos

Statistical Learning

Daniel Alexis Gutierrez Pachas, PhD
dgutierrezp@ucsp.edu.pe

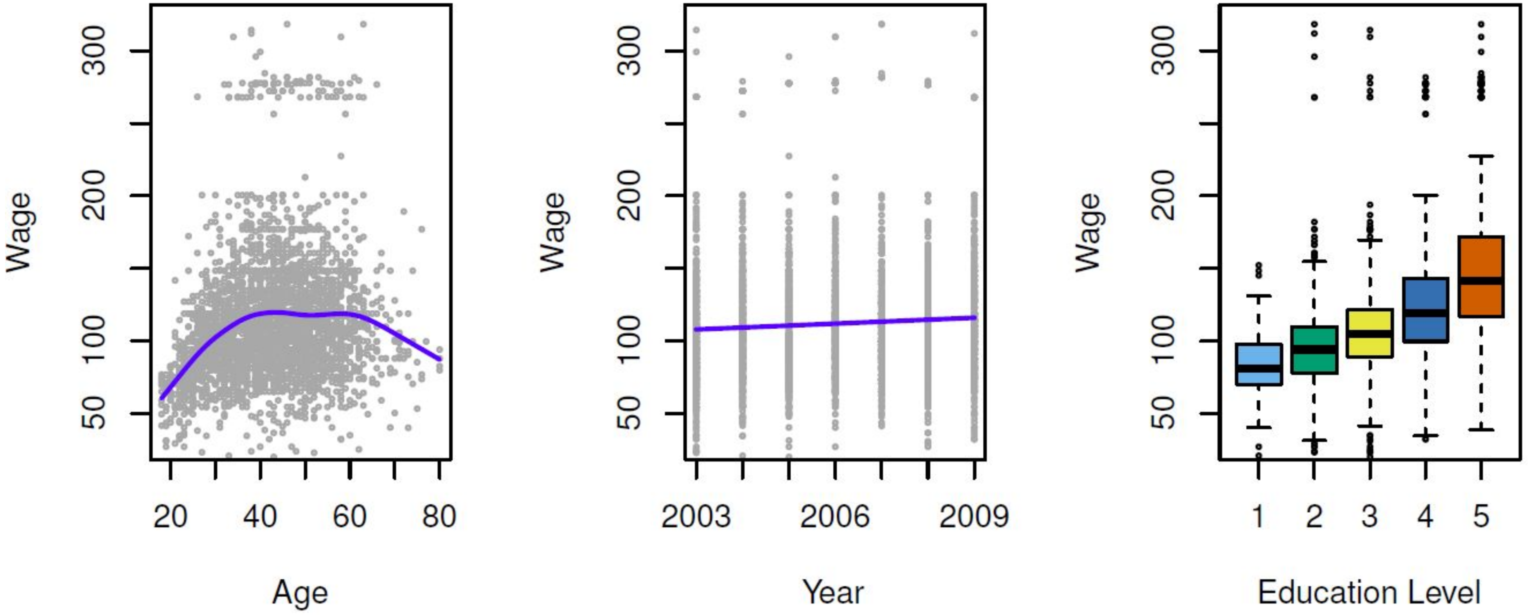
Universidad Católica San Pablo | Departamento de Ciencia de la Computación

Statistical learning refers to a vast set of tools for understanding data. These tools can be classified as **supervised** or **unsupervised**.

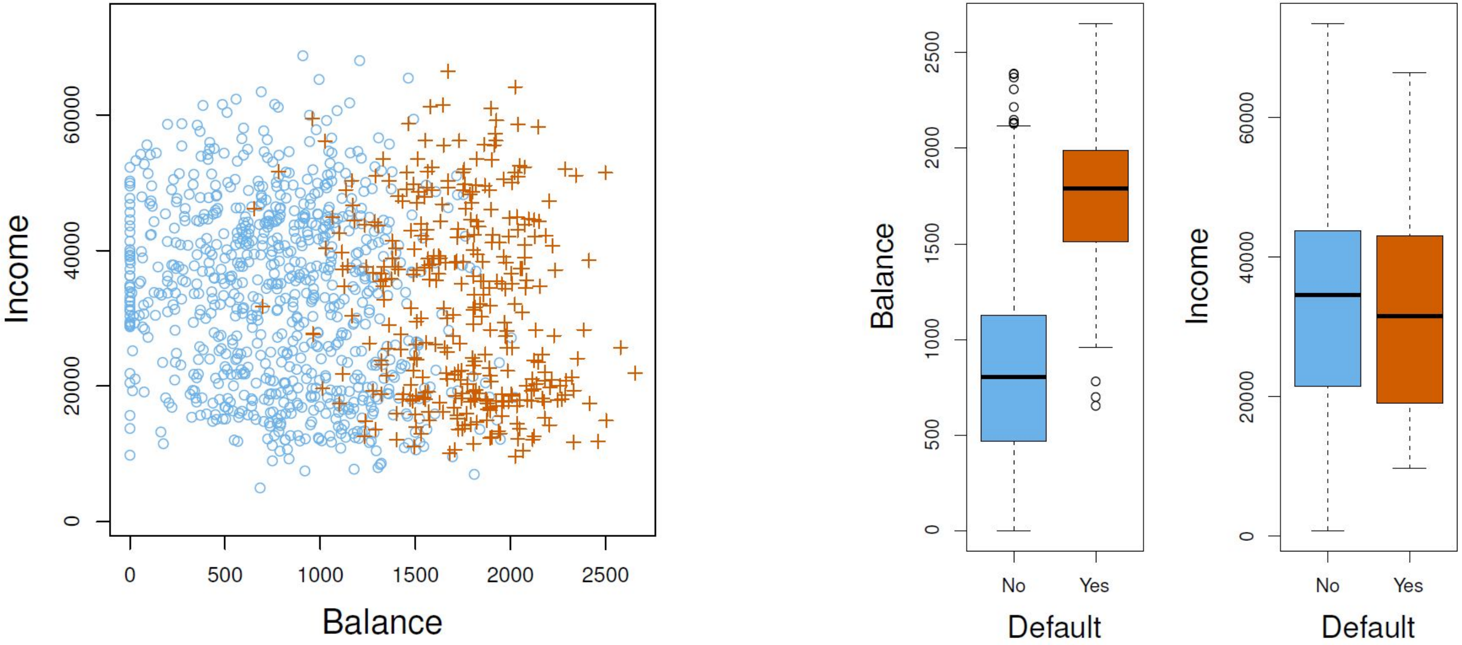
- **Supervised statistical learning** involves building a statistical model for predicting, or estimating, an output based on one or more inputs.

- **Unsupervised statistical learning**, there are inputs but no supervising output; nevertheless we can learn relationships and structure from such data.

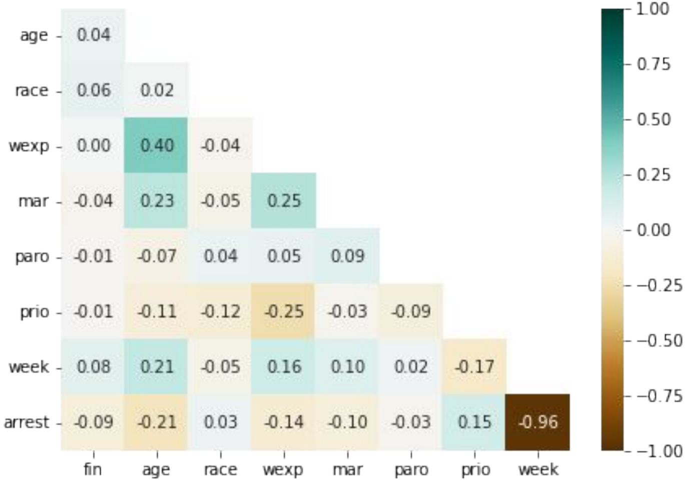
Example 1 (Wage Data): We examine a number of factors that relate to wages for a group of men from the Atlantic region of the United States. In particular, we wish to understand the association between an employee’s age and education, as well as the calendar year, on his wage.

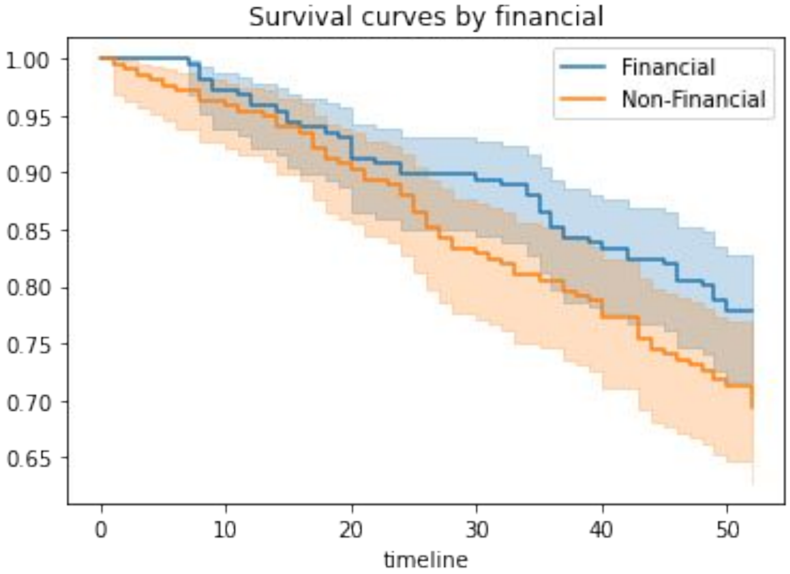


Example 2 (Default data): We are interested in predicting whether an individual will default on his or her credit card payment, on the basis of annual income and monthly credit card balance.

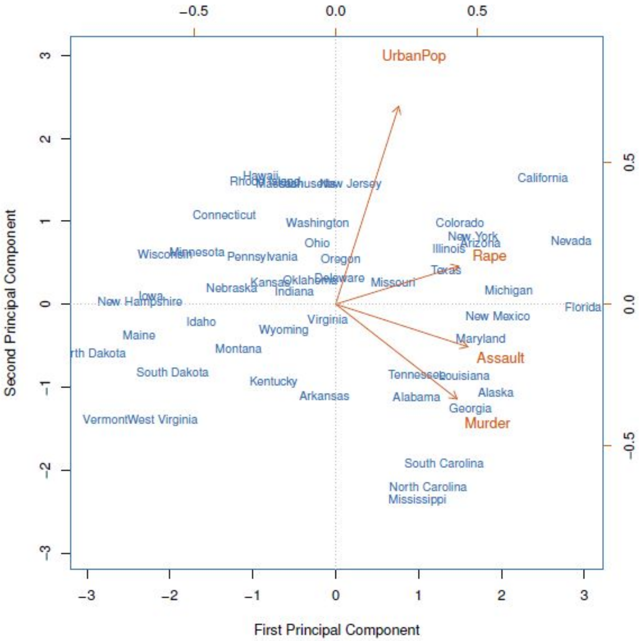


Example 3 (Recidivism data): We have information on 432 convicts who were released from Maryland state prisons in the 1970s and followed up for one year after release. Half the released convicts obtained financial support; half did not receive support. We will analyze the impact of this finance.

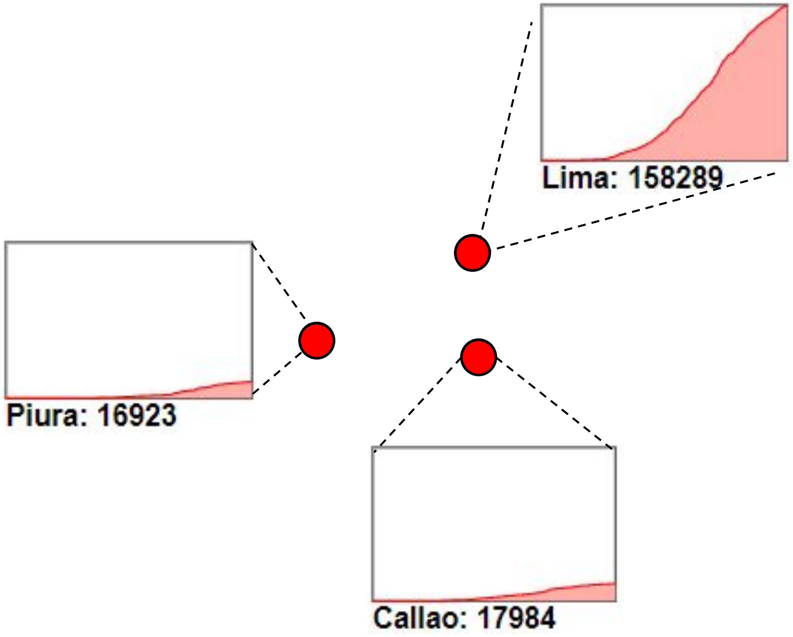


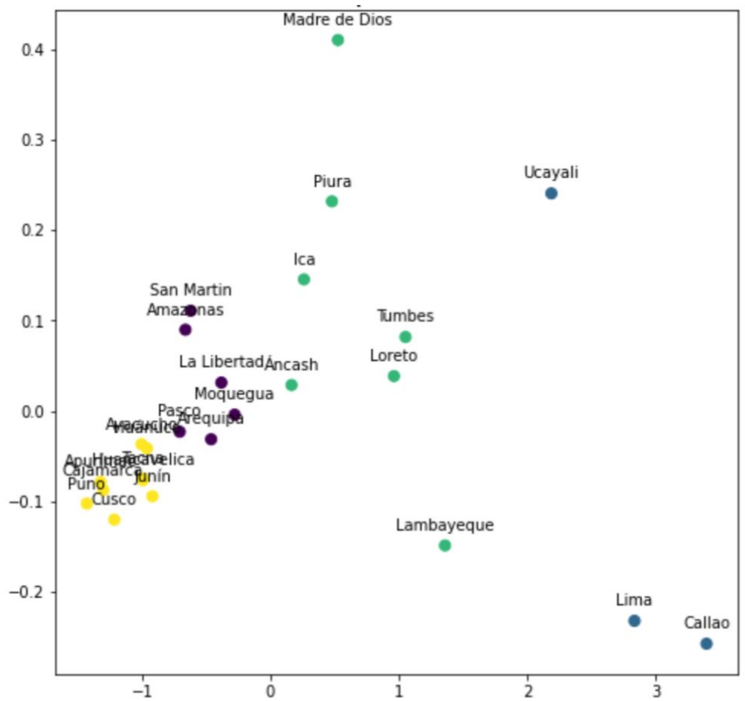


Example 4 (USArrests data): The data set contains the number of arrests per 100, 000 residents for each of three crimes: Assault, Murder, and Rape. We also record UrbanPop (the percent of the population in each state living in urban areas). We can examine differences between the states via the two principal component score vectors.



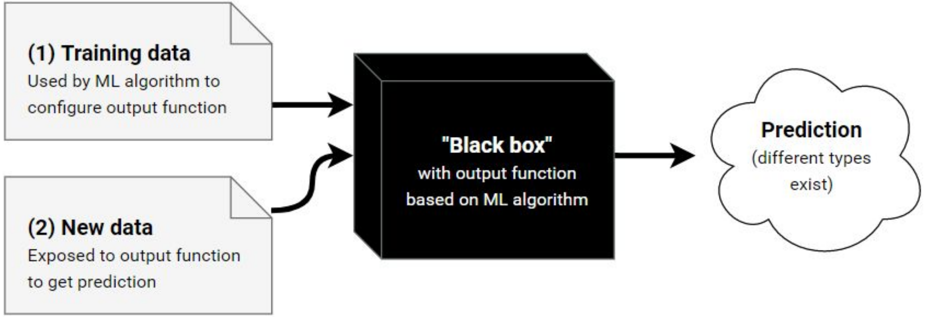
Example 5 (Covid19Peru data): Group the Peruvian departments according to the impact of Covid 19.





The Supervised Learning Problem

- Outcome measurement Y , Vector of p predictor measurements X (also called independent variables).
- In regression methods, Y takes quantitatives and continuous values.
- In the classification problem, Y takes values in a finite unordered set.
- In Survival analysis problem, Y=(T,E), where T and E are temporal and event variables.


(1) Training data
Used by ML algorithm to configure output function
(2) New data
Exposed to output function to get prediction
"Black box"
with output function based on ML algorithm
Prediction
(different types exist)

Examples 1, 2,and 3 are Supervised Learning Problems.

The Unsupervised Learning problem

- **No outcome variable**, just a set of predictors (features) measured on a set of samples.

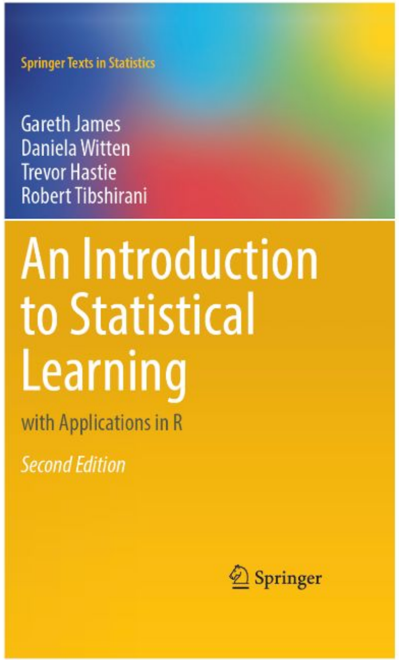
- Objective is **more fuzzy**:

    - Find groups of samples that behave similarly, our
    - Determine features that behave similarly (linear combinations).

- Different from supervised learning, but can be useful as a pre-processing step for supervised learning.

Examples 4, and 5 are Unsupervised Learning Problems.

References



Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

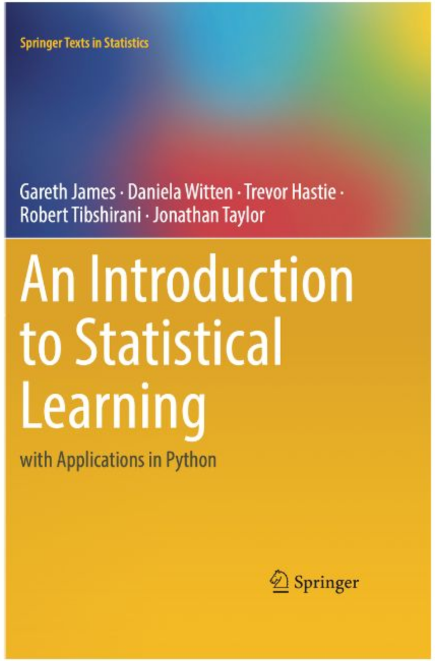
An Introduction to Statistical Learning

with Applications in R

Second Edition

Springer

Second Edition, 2022.



Springer Texts in Statistics

Gareth James · Daniela Witten · Trevor Hastie ·
Robert Tibshirani · Jonathan Taylor

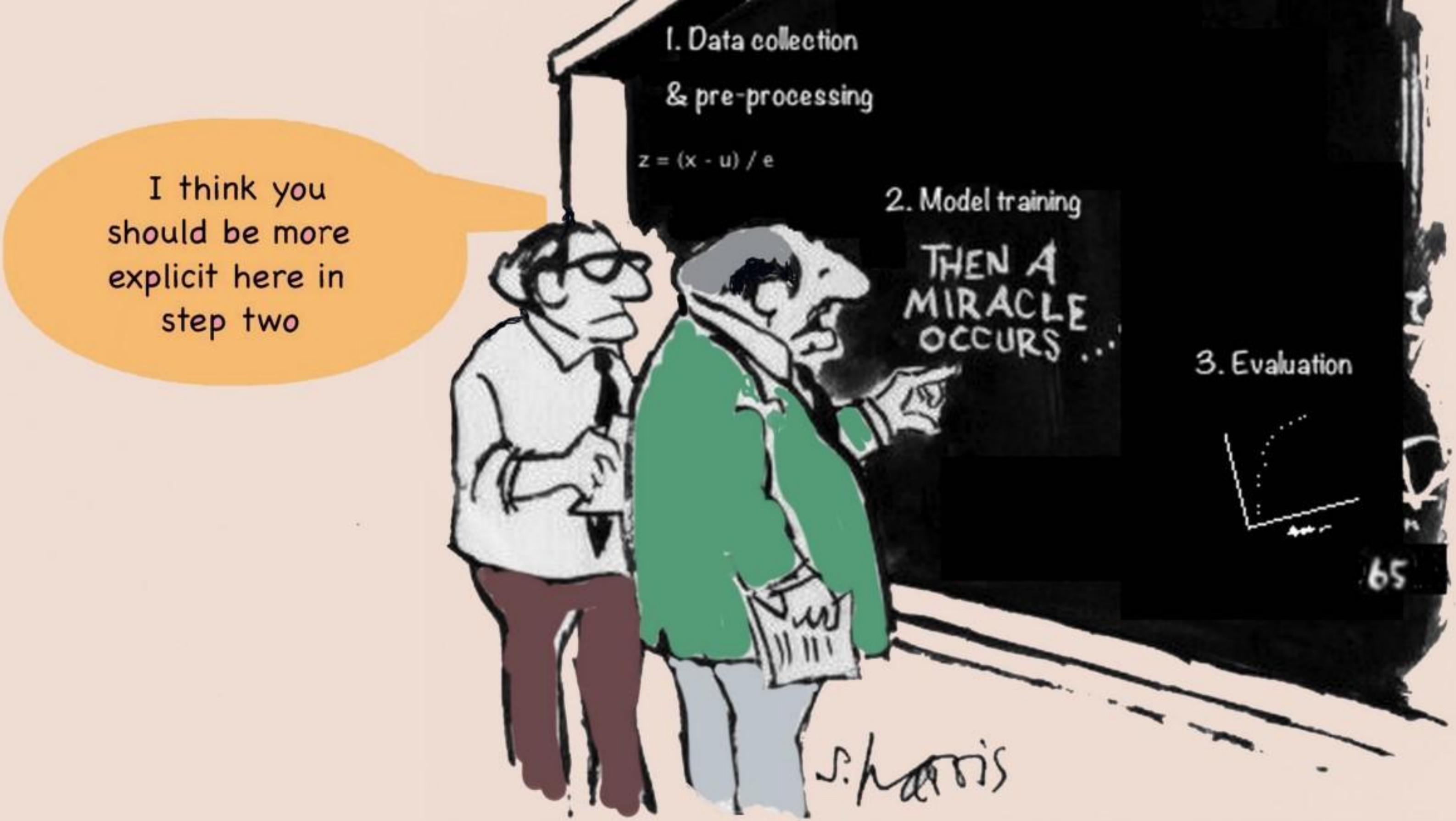
An Introduction to Statistical Learning

with Applications in Python

Springer

First Edition, 2023.

https://www.statlearning.com/



from *What's so Funny about Science?* by Sidney Harris (1977)