Universidad
Rey Juan Carlos

Unit 1. Introduction

Artificial Intelligence and Learning

Contents

1.1 Artificial Intelligence and Machine Learning

1.2 Data collection, cleaning and pre-processing

1.3 Supervised and unsupervised learning

1.4 Selection and evaluation of machine learning models

1.5 Biomedical examples and applications

Contents

1.1 Artificial Intelligence and Machine Learning

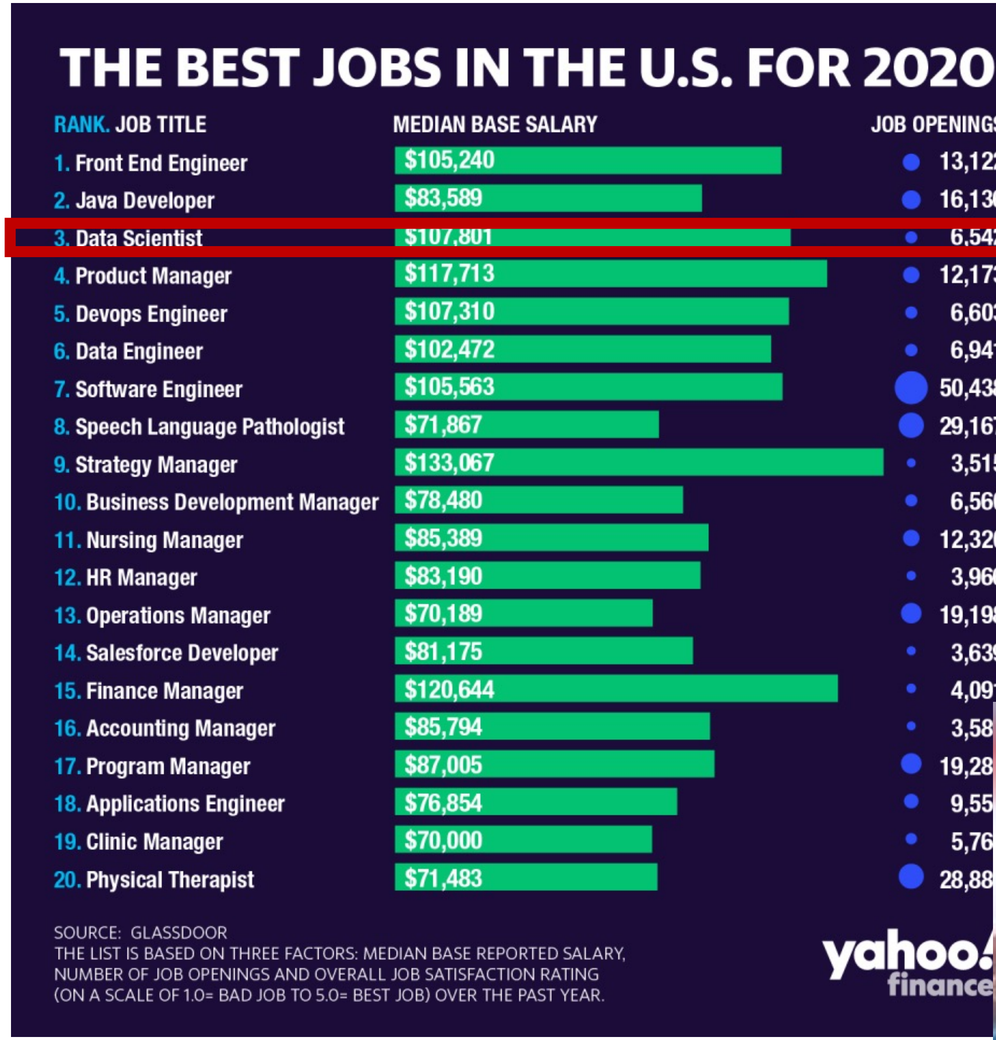
1.2 Data collection, cleaning and pre-processing

1.3 Supervised and unsupervised learning

1.4 Selection and evaluation of machine learning models

1.5 Biomedical examples and applications

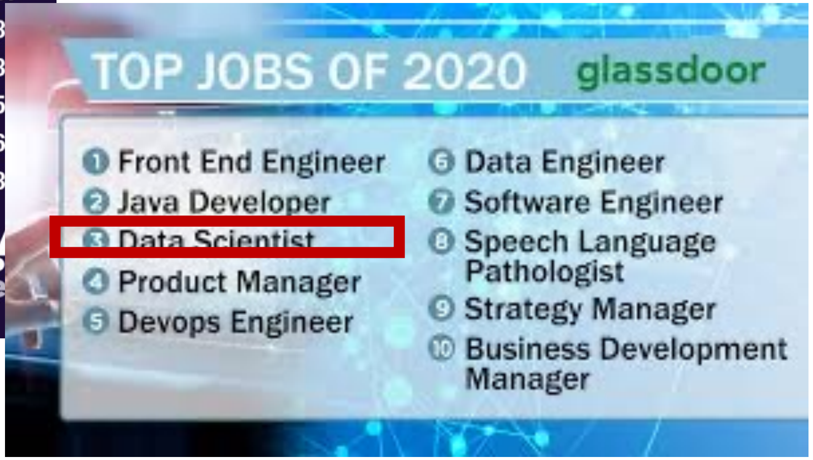
1.1 Artificial intelligence and machine learning



THE BEST JOBS IN THE U.S. FOR 2020

RANK. JOB TITLE	MEDIAN BASE SALARY	JOB OPENINGS
1. Front End Engineer	$105,240	13,122
2. Java Developer	$83,589	16,136
3. Data Scientist	$107,801	6,542
4. Product Manager	$117,713	12,173
5. Devops Engineer	$107,310	6,603
6. Data Engineer	$102,472	6,941
7. Software Engineer	$105,563	50,438
8. Speech Language Pathologist	$71,867	29,167
9. Strategy Manager	$133,067	3,515
10. Business Development Manager	$78,480	6,560
11. Nursing Manager	$85,389	12,320
12. HR Manager	$83,190	3,966
13. Operations Manager	$70,189	19,198
14. Salesforce Developer	$81,175	3,639
15. Finance Manager	$120,644	4,091
16. Accounting Manager	$85,794	3,58
17. Program Manager	$87,005	19,28
18. Applications Engineer	$76,854	9,55
19. Clinic Manager	$70,000	5,76
20. Physical Therapist	$71,483	28,88

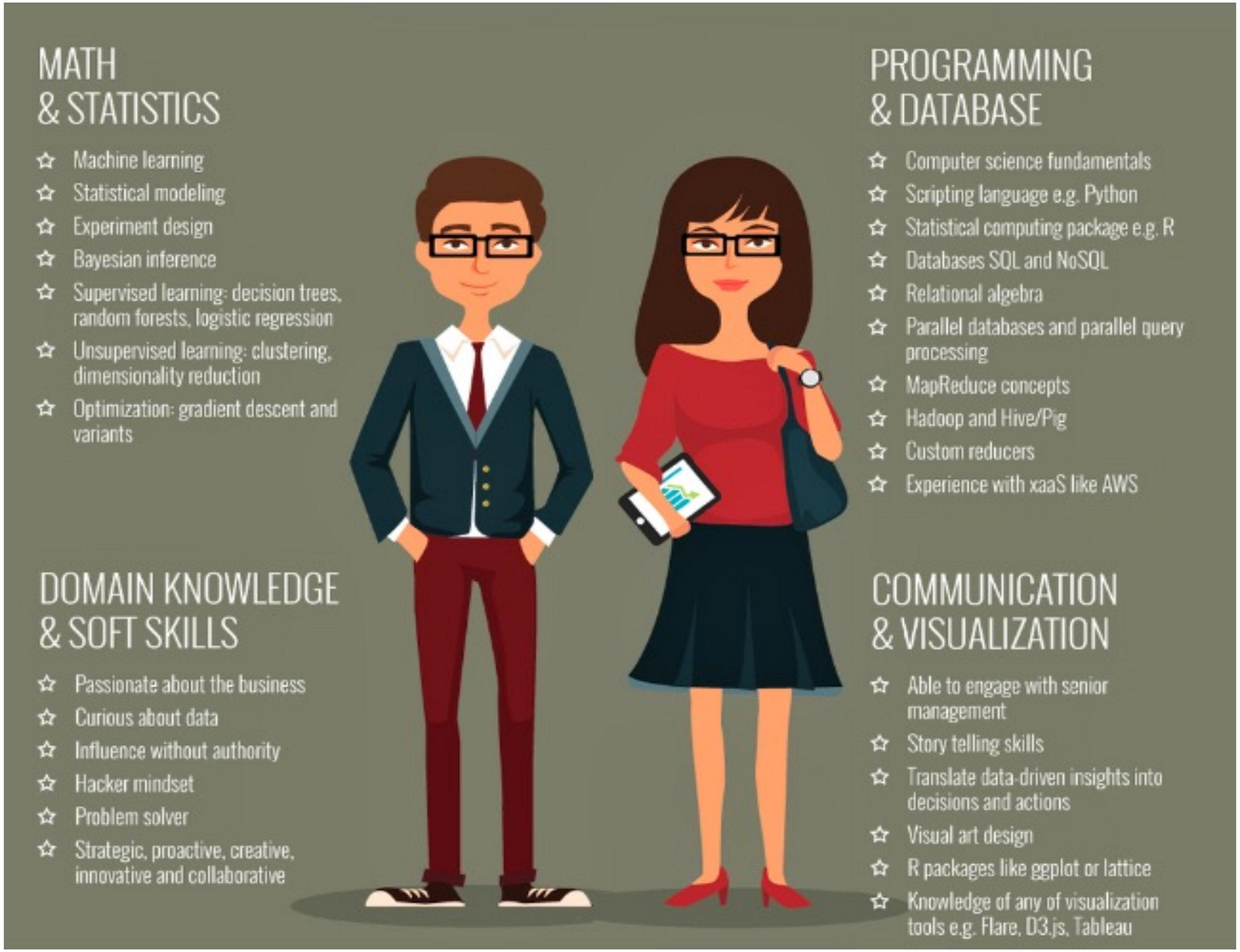
SOURCE: GLASSDOOR
THE LIST IS BASED ON THREE FACTORS: MEDIAN BASE REPORTED SALARY, NUMBER OF JOB OPENINGS AND OVERALL JOB SATISFACTION RATING (ON A SCALE OF 1.0= BAD JOB TO 5.0= BEST JOB) OVER THE PAST YEAR.



TOP JOBS OF 2020 glassdoor

1. Front End Engineer
2. Java Developer
3. Data Scientist
4. Product Manager
5. Devops Engineer
6. Data Engineer
7. Software Engineer
8. Speech Language Pathologist
9. Strategy Manager
10. Business Development Manager

1.1 Artificial intelligence and machine learning



MATH
& STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING
& DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE
& SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION
& VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

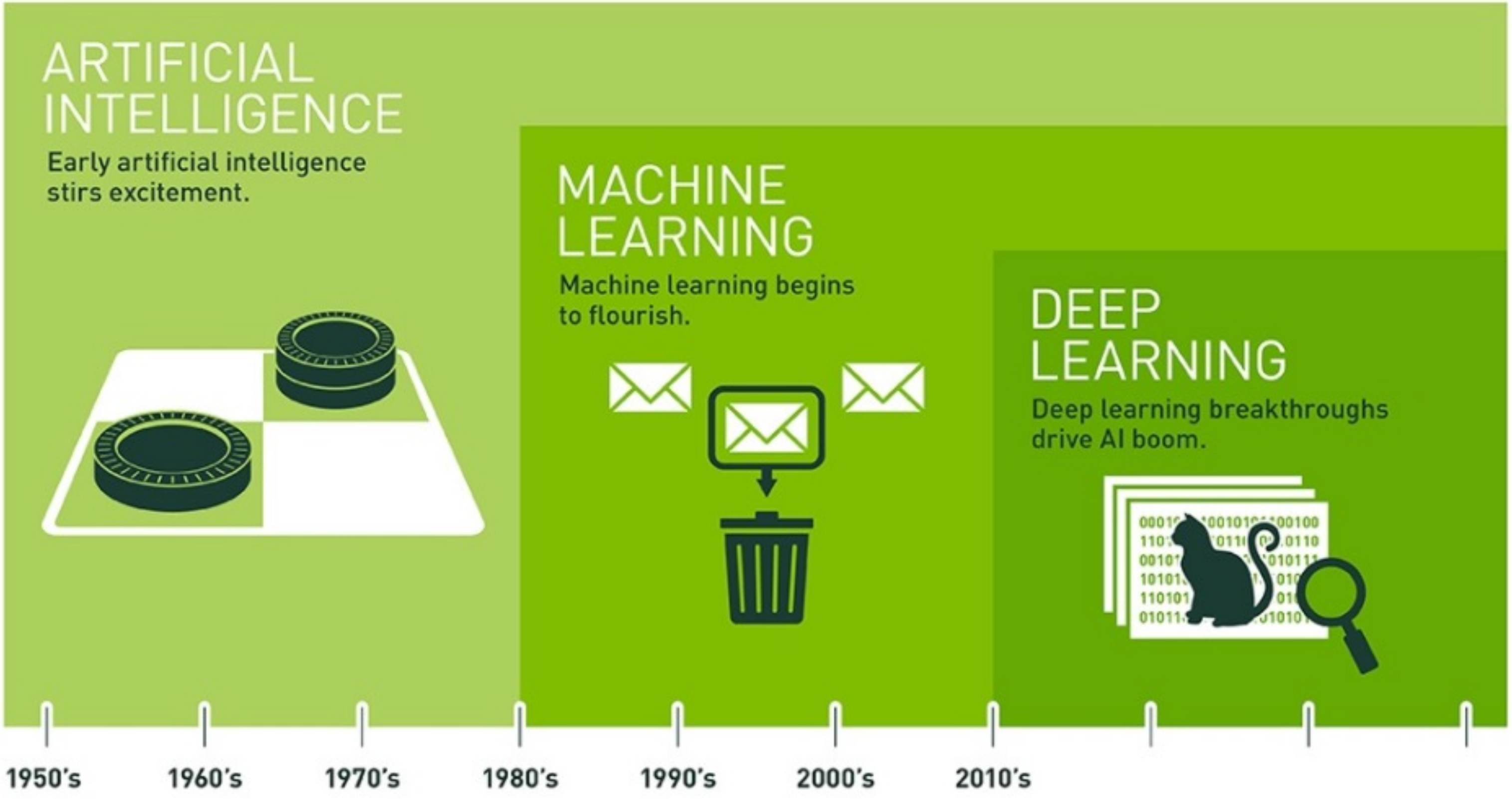
1.1 Artificial intelligence and machine learning

"**Artificial intelligence** is the part of computer science that is responsible for making machines **replicate the cognitive functions of the human mind**. Reasoning, **learning**, understanding, communicating..."

In order to perform tasks in a manner that is considered "intelligent and automatic", the machine must have the ability to learn:

Machine learning (ML) allows the machine to learn the relationships between data.

1.1 Artificial intelligence and machine learning

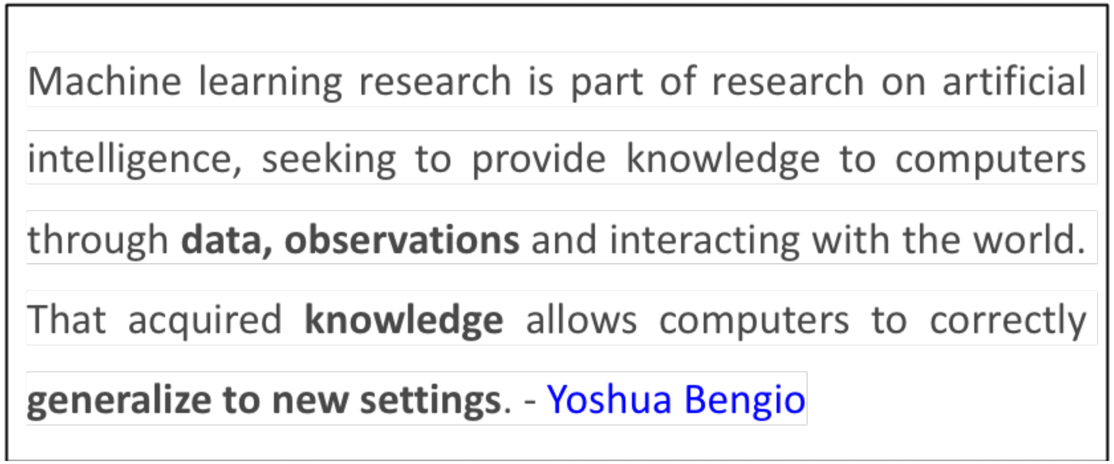

ARTIFICIAL INTELLIGENCE
Early artificial intelligence stirs excitement.
MACHINE LEARNING
Machine learning begins to flourish.
DEEP LEARNING
Deep learning breakthroughs drive AI boom.
1950's
1960's
1970's
1980's
1990's
2000's
2010's

1.1 Artificial intelligence and machine learning

Machine Learning

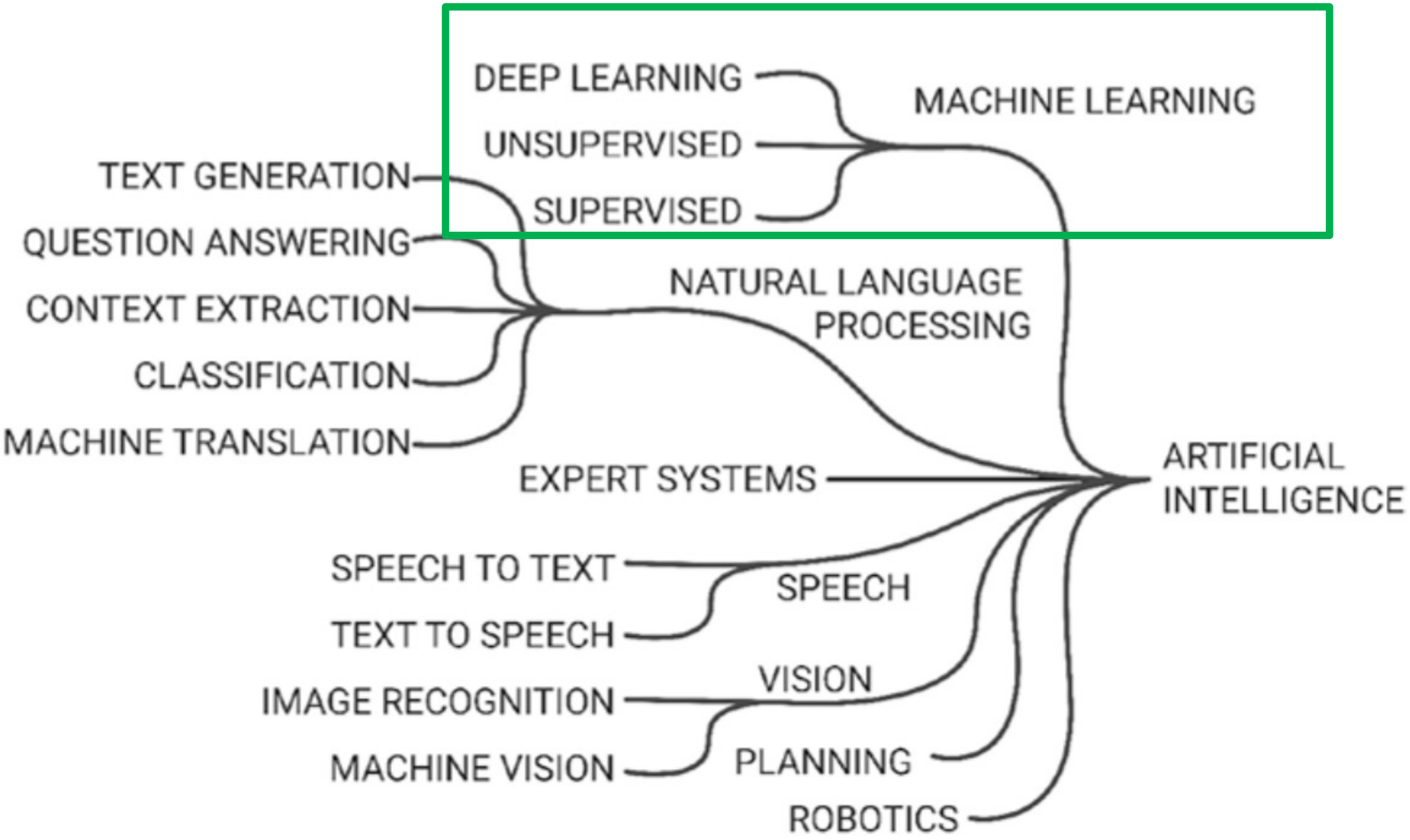
"Machine Learning at its most basic is the practice of using algorithms to parse **data, learn from it**, and then make a determination or **prediction** about something in the world." – Nvidia

"Machine learning is the science of getting computers to act without being explicitly programmed." – Stanford



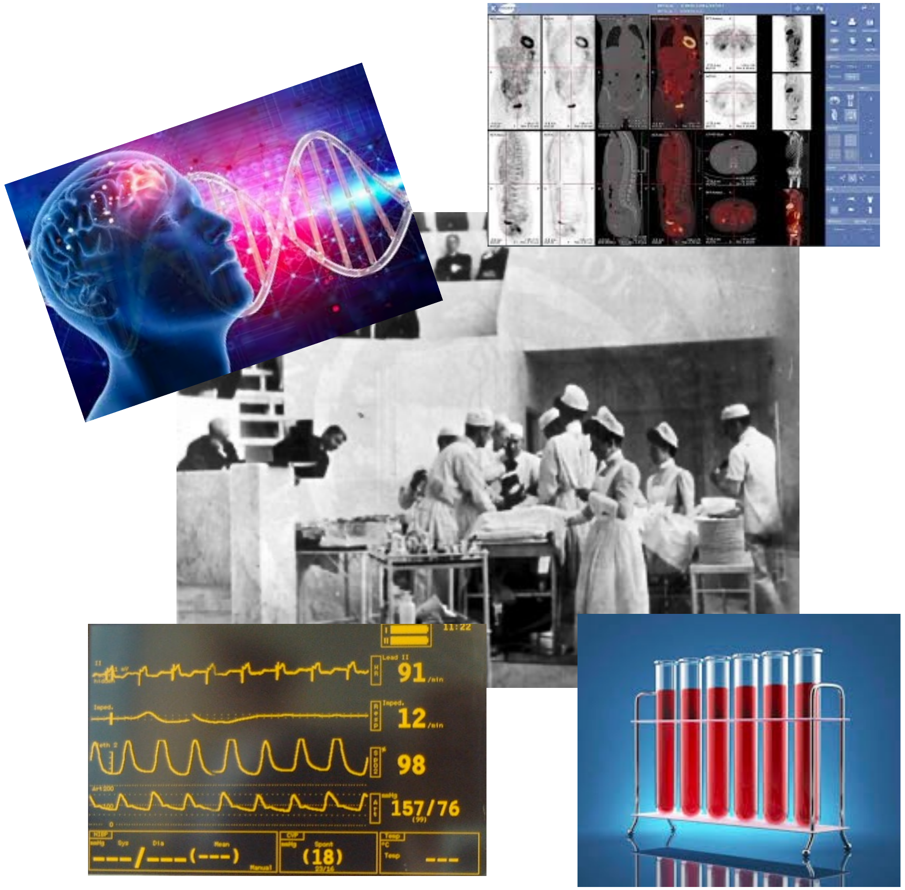
Machine learning research is part of research on artificial intelligence, seeking to provide knowledge to computers through **data, observations** and interacting with the world. That acquired **knowledge** allows computers to correctly **generalize to new settings**. - Yoshua Bengio

1.1 Artificial intelligence and machine learning



1.1 Artificial intelligence and machine learning

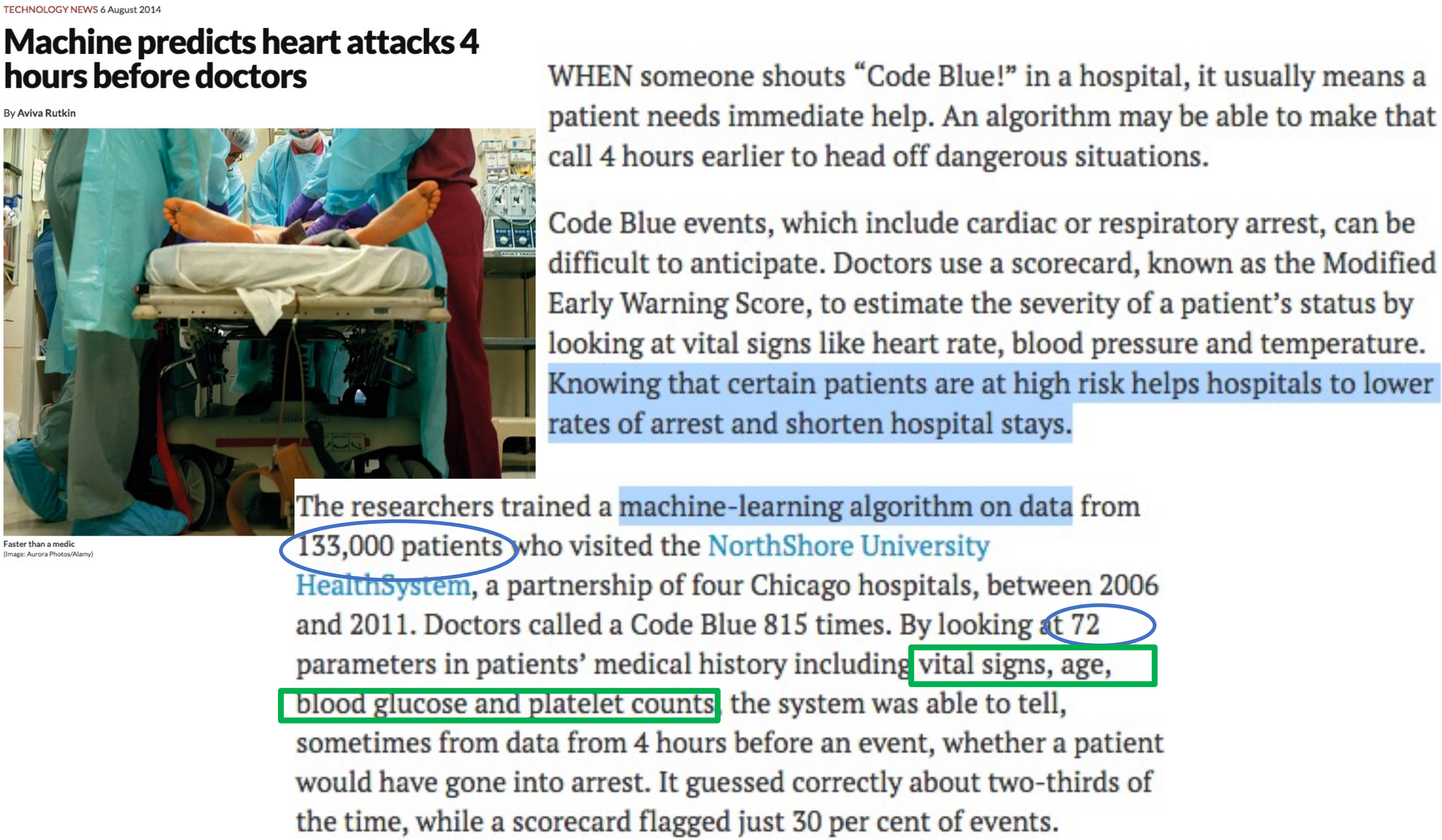
1.1 Artificial intelligence and machine learning



A large amount of data (of a complex nature) is currently generated

1.1 Artificial intelligence and machine learning

https://www.newscientist.com/article/mg22329814-400-machine-predicts-heart-attacks-4-hours-before-doctors/



TECHNOLOGY NEWS 6 August 2014

Machine predicts heart attacks 4 hours before doctors

By Aviva Rutkin

Faster than a medic
(Image: Aurora Photos/Alamy)

WHEN someone shouts "Code Blue!" in a hospital, it usually means a patient needs immediate help. An algorithm may be able to make that call 4 hours earlier to head off dangerous situations.

Code Blue events, which include cardiac or respiratory arrest, can be difficult to anticipate. Doctors use a scorecard, known as the Modified Early Warning Score, to estimate the severity of a patient's status by looking at vital signs like heart rate, blood pressure and temperature. Knowing that certain patients are at high risk helps hospitals to lower rates of arrest and shorten hospital stays.

The researchers trained a machine-learning algorithm on data from 133,000 patients who visited the NorthShore University HealthSystem, a partnership of four Chicago hospitals, between 2006 and 2011. Doctors called a Code Blue 815 times. By looking at 72 parameters in patients' medical history including vital signs, age, blood glucose and platelet counts, the system was able to tell, sometimes from data from 4 hours before an event, whether a patient would have gone into arrest. It guessed correctly about two-thirds of the time, while a scorecard flagged just 30 per cent of events.

1.1 Artificial intelligence and machine learning

Machine learning techniques allow you to tackle a problem by learning from experience.

The **experience** is represented through **observations** (also called cases or examples).

TECHNOLOGY NEWS 6 August 2014

Machine predicts heart attacks 4 hours before doctors

By Aviva Rutkin

Faster than a medic
(Image: Aurora Photos/Alamy)

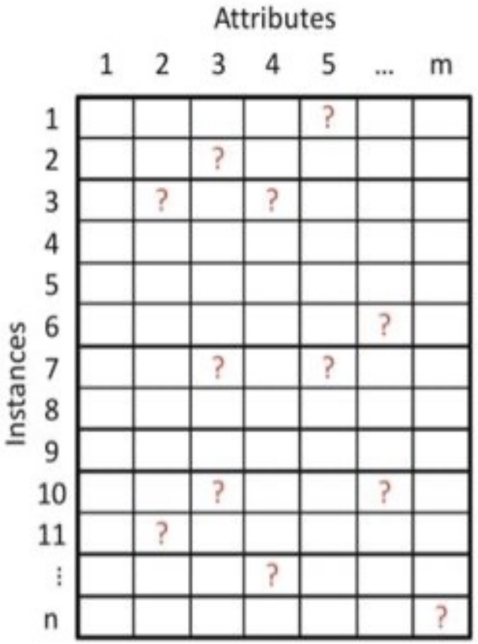
Input: The **examples/observations/samples** correspond to the 133,000 patients, each characterized by 72 **variables/attributes/features**.

Target/Output/Label: For each patient it is also known whether or not he suffered a heart attack at 4 hours.

Objective: to learn the **relationship** (model) between the 72 variables and whether or not the patient suffered a heart attack.

1.1 Artificial intelligence and machine learning

	Feature 1	Feature 2	...	Feature 72
Patient 1	$x_1^{(1)}$	$x_2^{(1)}$		$x_{72}^{(1)}$
Patient 2	$x_1^{(2)}$	$x_2^{(2)}$		$x_{72}^{(2)}$
Patient 3				
Patient 4				
…				
…				
…				
…				
…				
Patient 133000	$x_1^{(n)}$			$x_m^{(n)}$

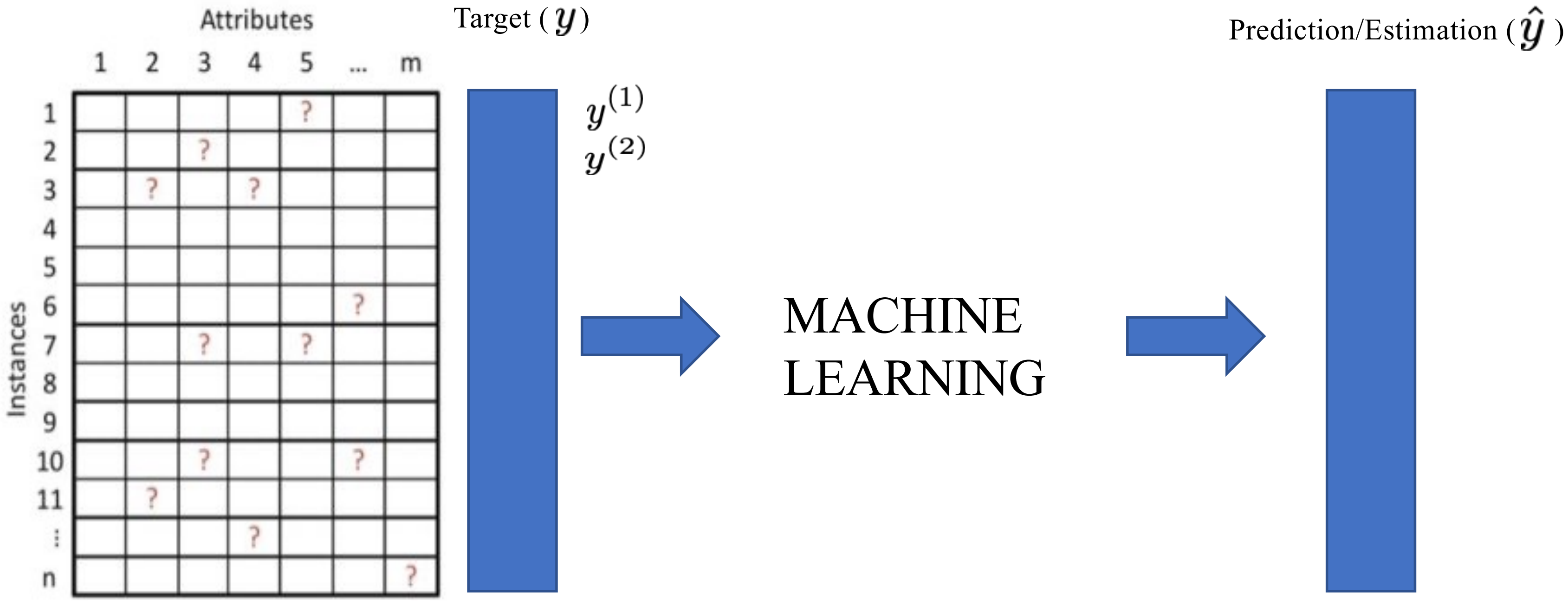


Attributes

Instances

	1	2	3	4	5	...	m
1					?		
2			?				
3		?		?			
4							
5							
6						?	
7			?		?		
8							
9							
10			?			?	
11		?					
⋮				?			
n							?

1.1 Artificial intelligence and machine learning



Attributes

Instances

	1	2	3	4	5	...	m
1					?		
2			?				
3		?		?			
4							
5							
6						?	
7			?		?		
8							
9							
10			?			?	
11		?					
⋮				?			
n							?

Target (y)

$y^{(1)}$

$y^{(2)}$

MACHINE LEARNING

Prediction/Estimation ($\hat{y}$)

1.1 Artificial intelligence and machine learning

The **models** built are based on learning/statistical tools by making use of the available cases to learn the underlying dynamics of the process to be modeled. For this reason, they are also called **data-driven models.**

The massive availability of data and the machine learning tools represent a real revolution in the Health Sector.

Are we ready for this revolution?

1.1 Artificial intelligence and machine learning

CHALLENGES!!!

- Scarce data
- Temporal data
- Irregular sampling
- High-dimensional data
- Sparse data
- Noisy labels
- Unbalanced database
- Feature selection
- …

Contents

1.1 Artificial Intelligence and Machine Learning

1.2 Data collection, cleaning and pre-processing

1.3 Supervised and unsupervised learning

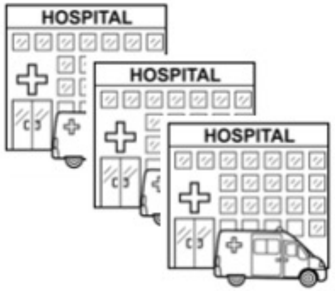
1.4 Selection and evaluation of machine learning models

1.5 Biomedical examples and applications

1.2 Data collection, cleaning and pre-processing

Data collection

- Voz
- Texto
- Imágenes



THE WORLD BANK

tripadvisor®

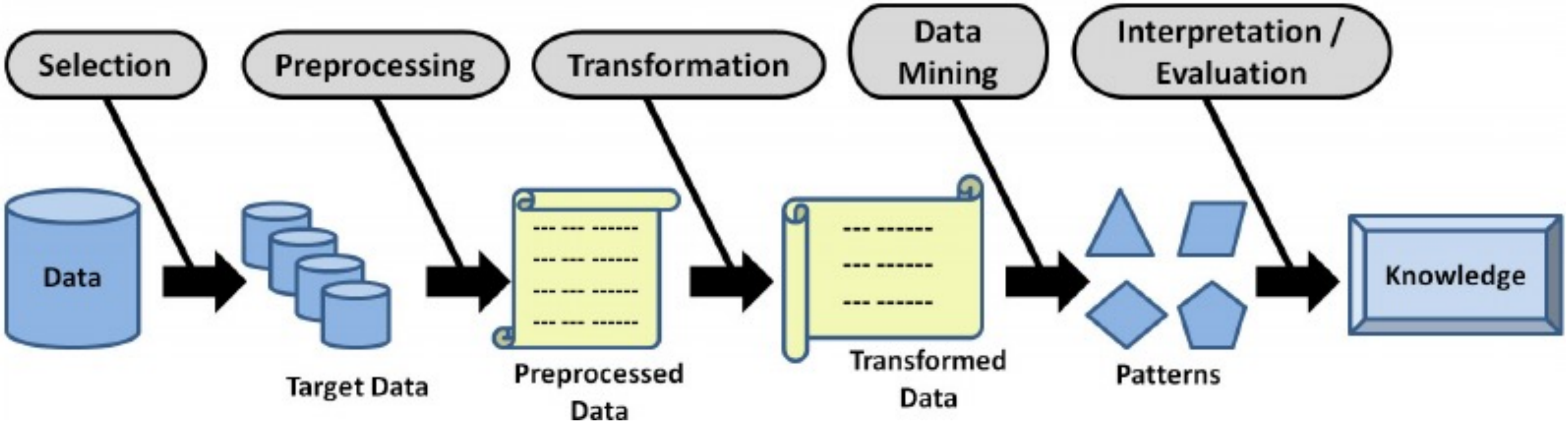
datos.gob.es
reutiliza la información pública

INE Instituto Nacional de Estadística

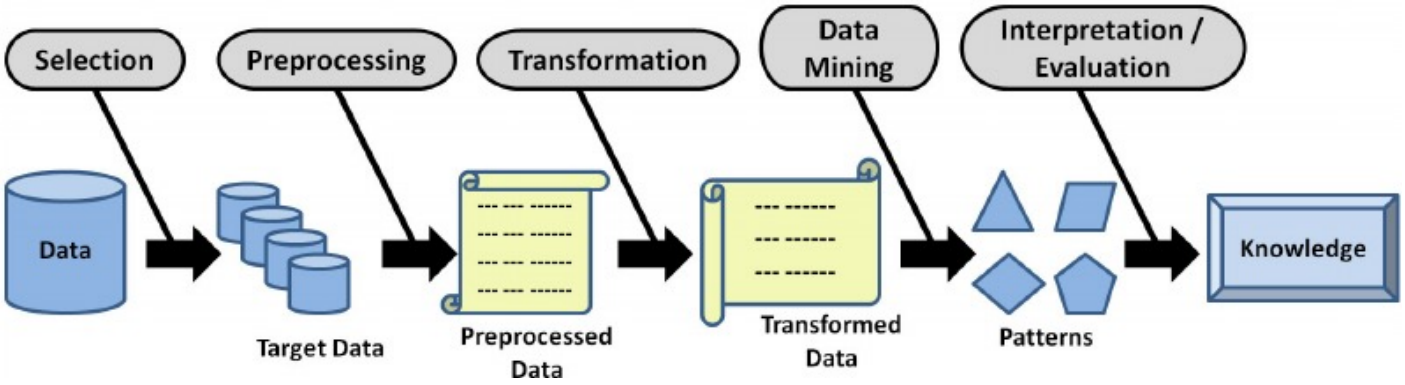
twitter

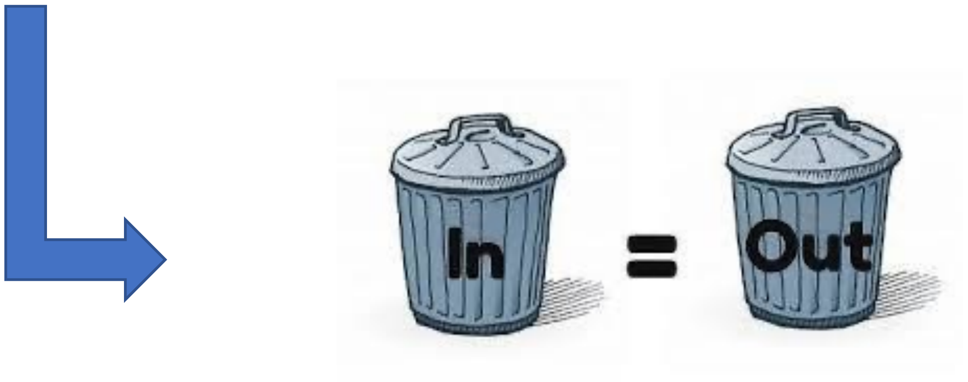
1.2 Data collection, cleaning and pre-processing

Knowledge discovery in databases



1.2 Data collection, cleaning and pre-processing





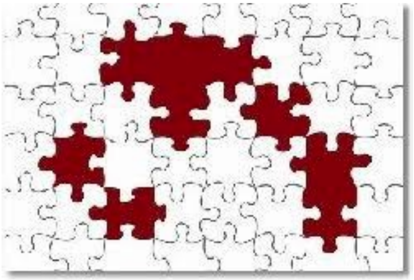
The success of the knowledge extraction process depends largely on the quality of the data used (GIGO).

1.2 Data collection, cleaning and pre-processing

Pre-procesing stage

The success of a machine/data mining/artificial intelligence learning process

does NOT only depend on having "a lot of data",



1.2 Data collection, cleaning and pre-processing

Challenges when working with real data

The data can:

- Not be in the same **database**
- Be stored in different **formats**
- Being **incomplete**: missing values, "missing data" (e.g., age = ?)
- Being contaminated by **noise**, may have **errors** (e.g., age = -10)
- **Inconsistent**: not corresponding to the attribute domain or contradictory with another, e.g., age = 50; year of birth = 1995

1.2 Data collection, cleaning and pre-processing

Techniques that allow to improve the quality of the data set.

Broadly speaking, the main tasks can be:

- Data integration
- Data cleaning
- Data transformation
- Data reduction

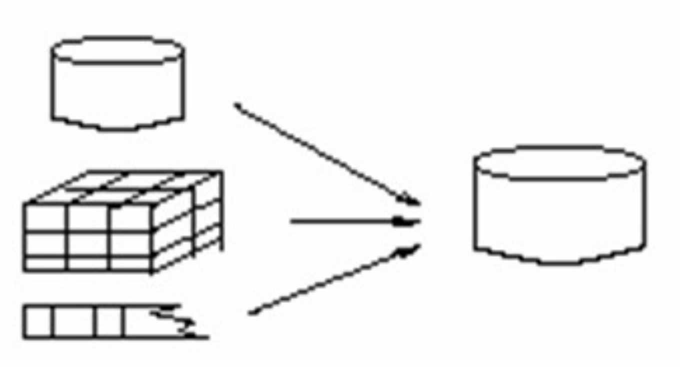
These tasks can be carried out in different ways, more than one can be done at the same time, ...

There is no established process.

1.2 Data collection, cleaning and pre-processing

Data integration

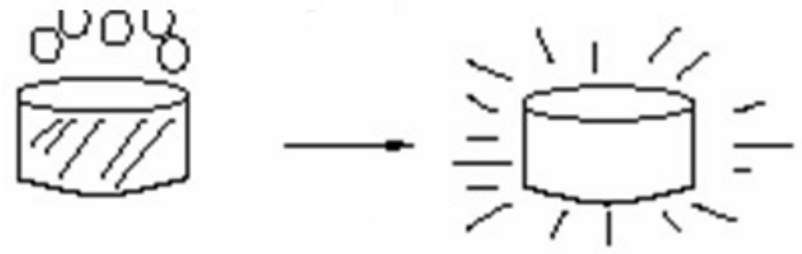
Integration of different databases, files, ... in a single database with a unique format



1.2 Data collection, cleaning and pre-processing

Data cleaning

- Complete incomplete records → **Missing values**
- Identify and eliminate erroneous data → **Outliers**
- Resolving inconsistencies
- Treating values with noise



1.2 Data collection, cleaning and pre-processing

Missing values

What to do with incomplete records? They are usually represented as: "?", "N/A", "0" or as a blank cell.

- **Nothing**

  - Some algorithms are robust to incomplete records (or records with missing data), most of them not.

- **Delete the record**

  - It is not effective when the percentage of lost values is very high.
  - Information available in other record attributes is lost.
  - It is not the best solution if the record is very significant.

- **Impute**

1.2 Data collection, cleaning and pre-processing

Missing values

- **Impute**
  - A global **constant,** for example, 0 value
  - A **statistic** dependent on the attribute, so that it preserves its average (continuos variables) or its mode (categorical variables)

It is important to note that the values used for imputation can affect the learning outcome

Sometimes a **new logical attribute** can be created to indicate whether the corresponding value of the original attribute was **incomplete or not**.

1.2 Data collection, cleaning and pre-processing

Outliers

- They do not match the general statistical behavior of the data. They can affect the quality of the data.

  They can be:

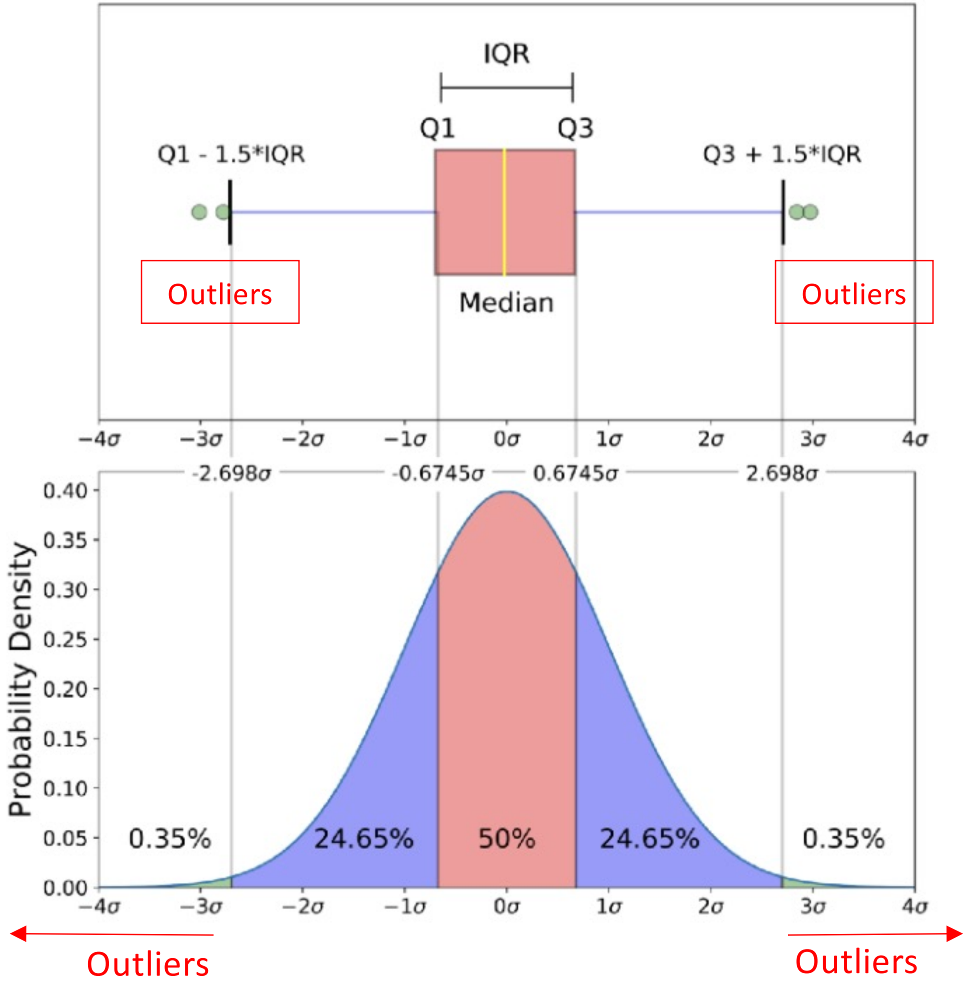
  - **correct values** that represent the reality
  - **wrong values**, e.g. because they are contaminated by noise ("age" = -200)

- There are statistical tools that suggest outliers. It is the user who must finally determine if they are wrong values or not

1.2 Data collection, cleaning and pre-processing

Outliers

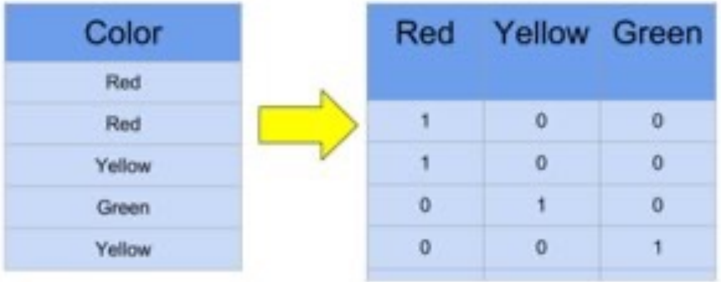
- **Boxplots**
- **Z-score**



1.2 Data collection, cleaning and pre-processing

Data reduction

Feature transformation – One-hot-encoding



Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

Normalization

① $x_{new} = \frac{x_{old}}{x_{max}}$

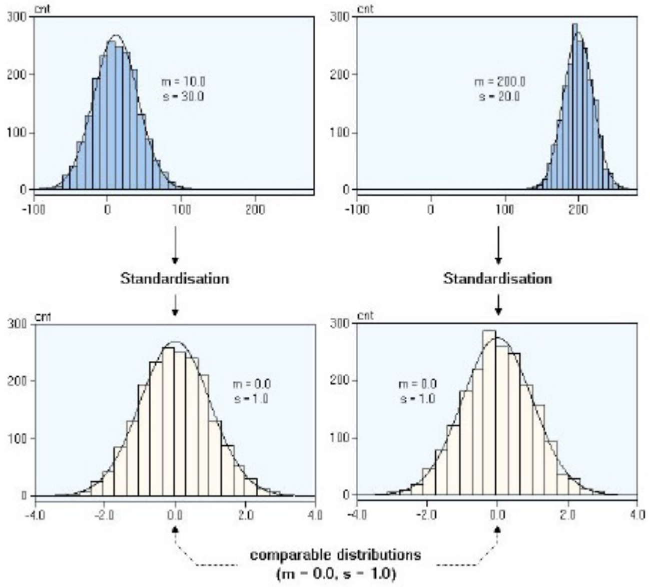
Simple Feature scaling

② $x_{new} = \frac{x_{old} - x_{min}}{x_{max} - x_{min}}$

Min-Max

③ $x_{new} = \frac{x_{old} - \mu}{\sigma}$

Z-score

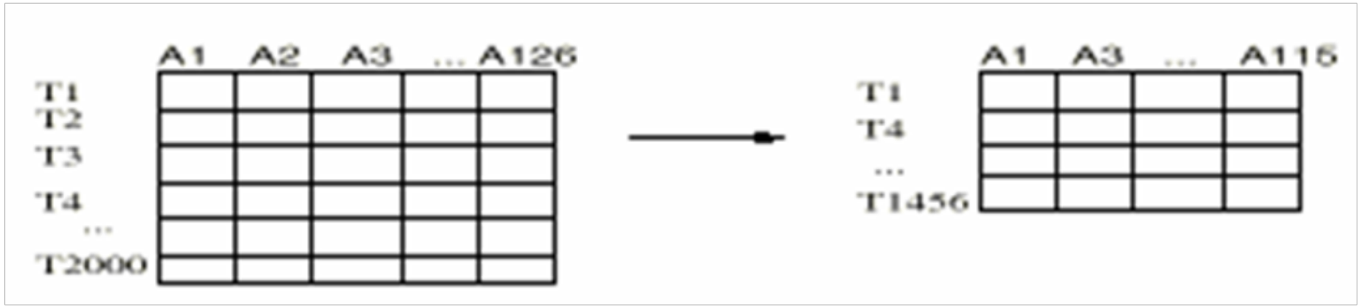


Z-score

1.2 Data collection, cleaning and pre-processing

Data Pre-processing

Data reduction



- Less but better quality

1.2 Data collection, cleaning and pre-processing

How can we know the quality of our data?

1.- We can make a descriptive analysis of:

- Calculate the number of different values
- Calculate statistics such as mean, average, fashion

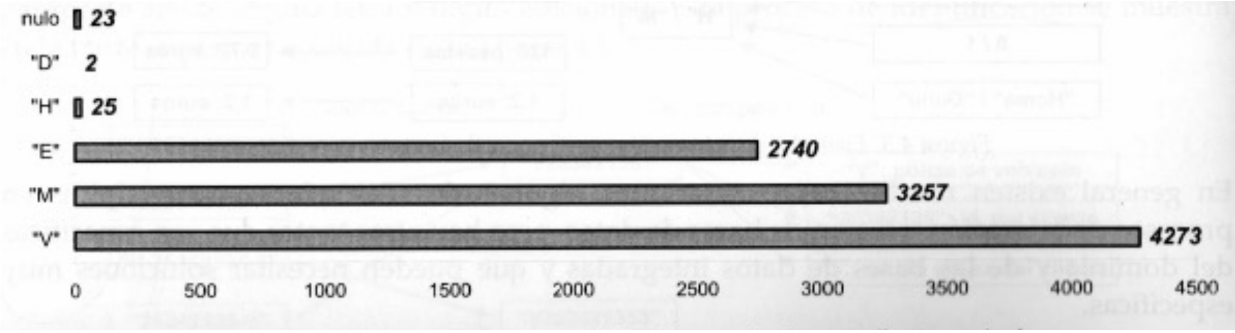
2.- We can use graphic tools to detect and solve conflicts in the data values.

- Histograms
- Bar graphs

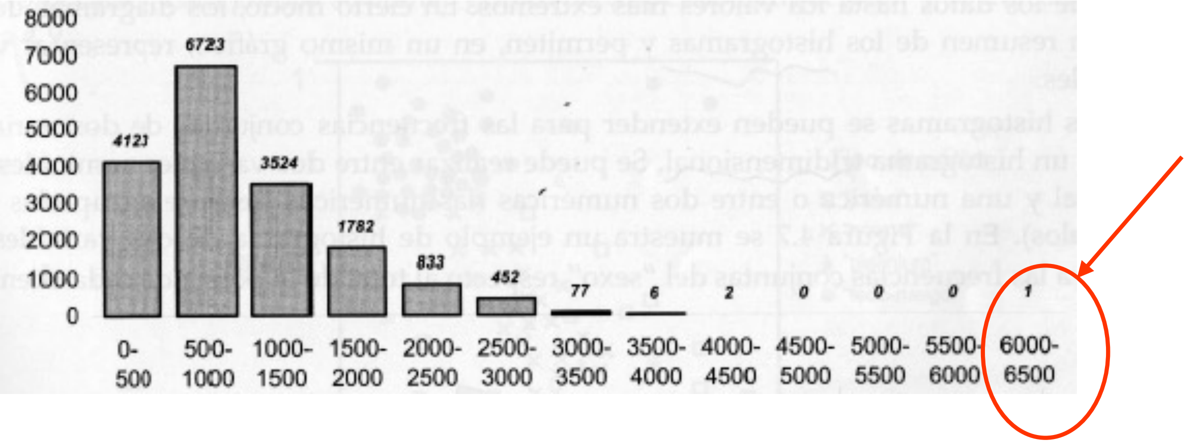
3.- Check the results obtained with the domain experts.

1.2 Data collection, cleaning and pre-processing

How can we know the quality of our data?∑



Is this wrong?



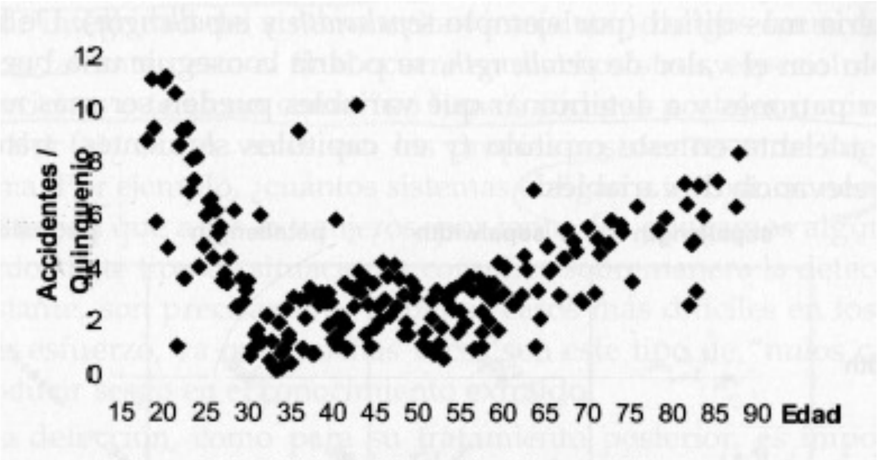
Without analyzing the data in detail it is not possible to determine whether it is an anomalous but correct value (e.g. the insurance of a very special car) or an erroneous data.

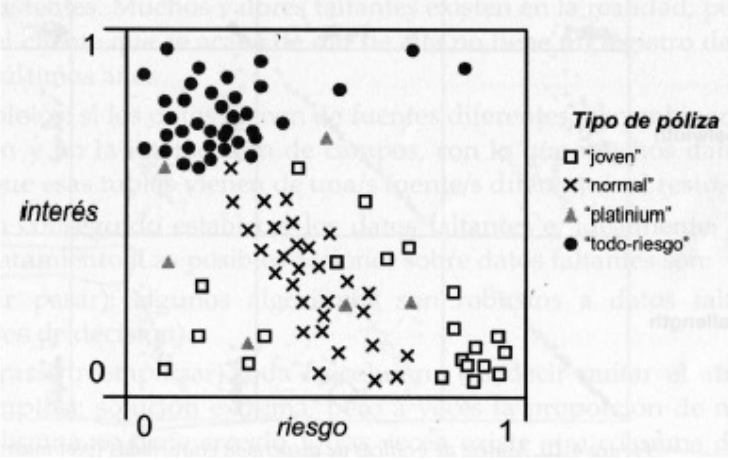
1.2 Data collection, cleaning and pre-processing

Scatter plots

The scatter plot is a diagram that represents in a Cartesian coordinate system the values of two attributes of the same individual.

It allows to study the type of relationship between two attributes.





The labelled scatter plot allows to show the relationship with a third nominal attribute.

Contents

1.1 Artificial Intelligence and Machine Learning

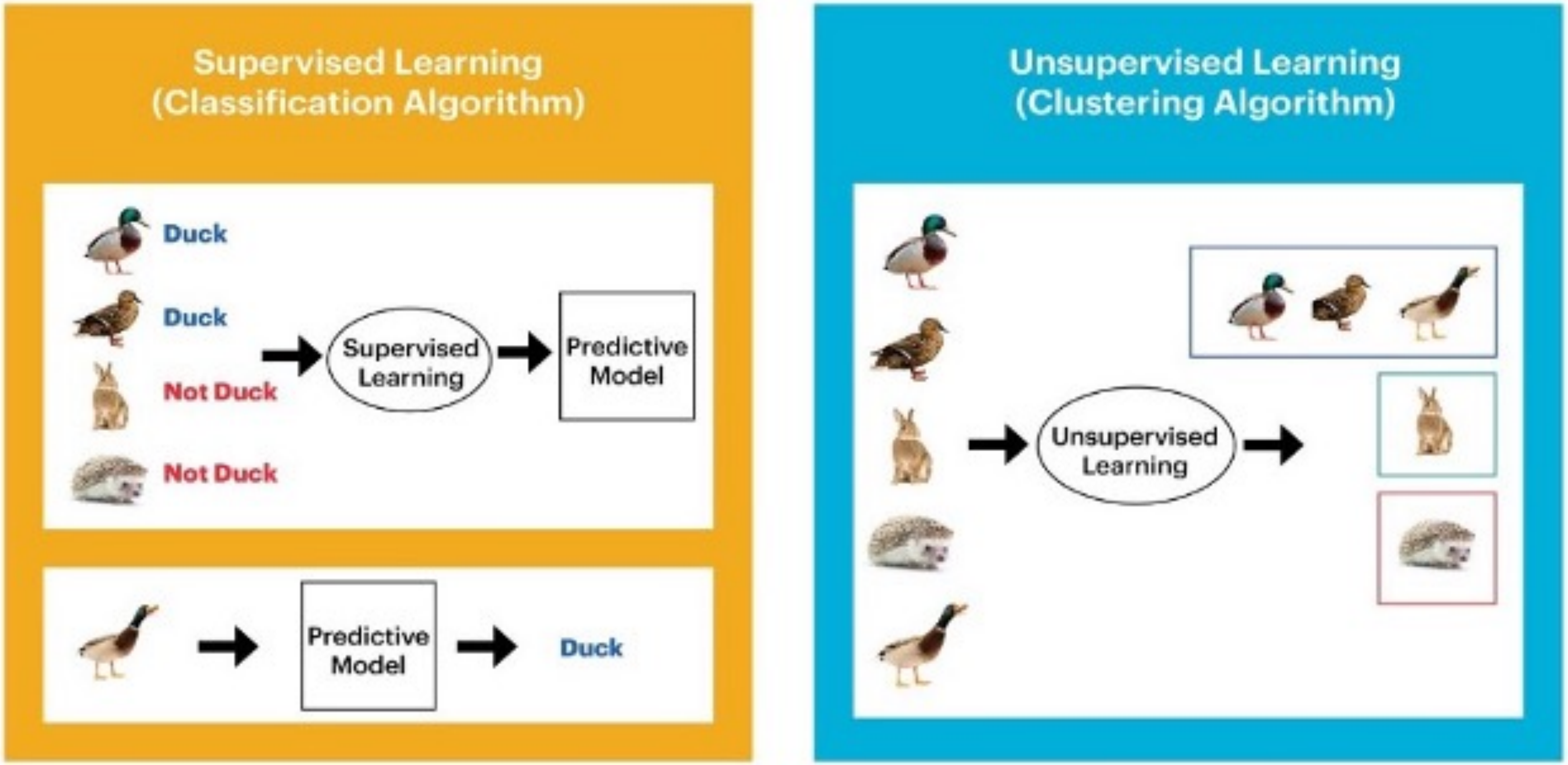
1.2 Data collection, cleaning and pre-processing

1.3 Supervised and unsupervised learning

1.4 Selection and evaluation of machine learning models

1.5 Biomedical examples and applications

1.3 Supervised and unsupervised learning



1.3 Supervised and unsupervised learning

- **Supervised learning**
  - Learn the relationship between feature vectors and the labels associated with each vector
  - There are two **types of supervised learning:**
    - **Classification:** The set of labels is numberable
    - **Regression:** The label set is not numeric
- **Unsupervised learning**
  - Learn the "structure" of the data by grouping examples into consistent groups with similar characteristics
  - It only considers the characteristics vectors

1.3 Supervised and unsupervised learning

What is the main difference between a supervised and unsupervised learning scheme?

Database:

* 133,000 patients
* 72 variables or characteristics/attribute

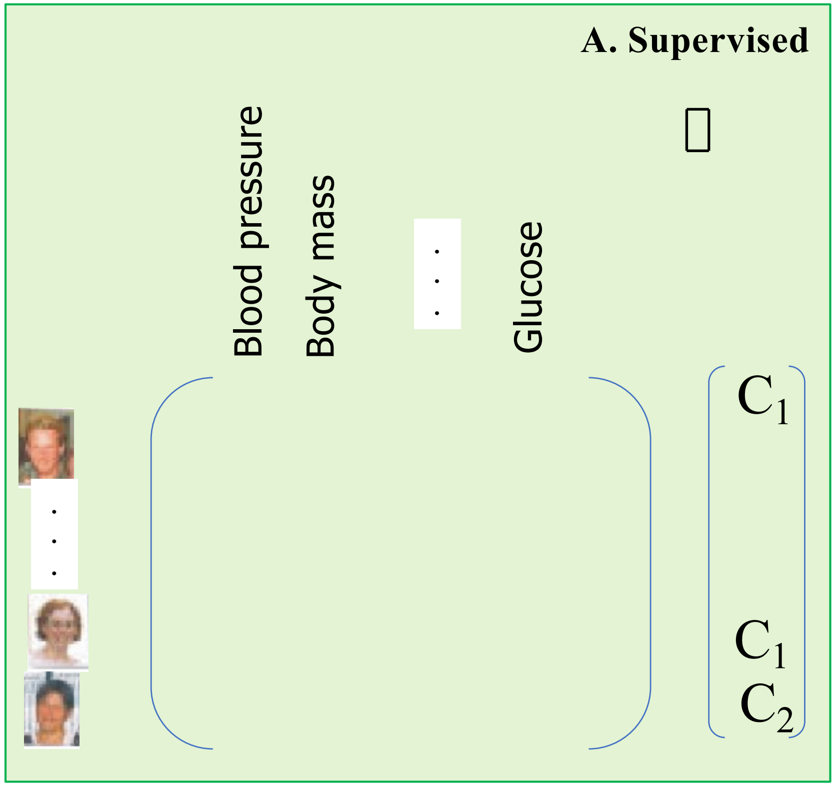
$$\begin{array}{r} \text{Blood pressure} \dashrightarrow \\ \text{Body mass} \dashrightarrow \\ \\ \text{Glucose} \dashrightarrow \end{array} \begin{bmatrix} x_1 \\ x_2 \\ . \\ x_{72} \end{bmatrix} \qquad \begin{array}{r} \text{Blood pressure} \dashrightarrow \\ \text{Body mass} \dashrightarrow \\ \\ \text{Glucose} \dashrightarrow \end{array} \begin{bmatrix} x_1 \\ x_2 \\ . \\ x_{72} \end{bmatrix}$$

Feature vector

1.3 Supervised and unsupervised learning

Database:

* 133,000 patients
* 72 variables or characteristics/attribute



A. Supervised

Blood pressure

Body mass

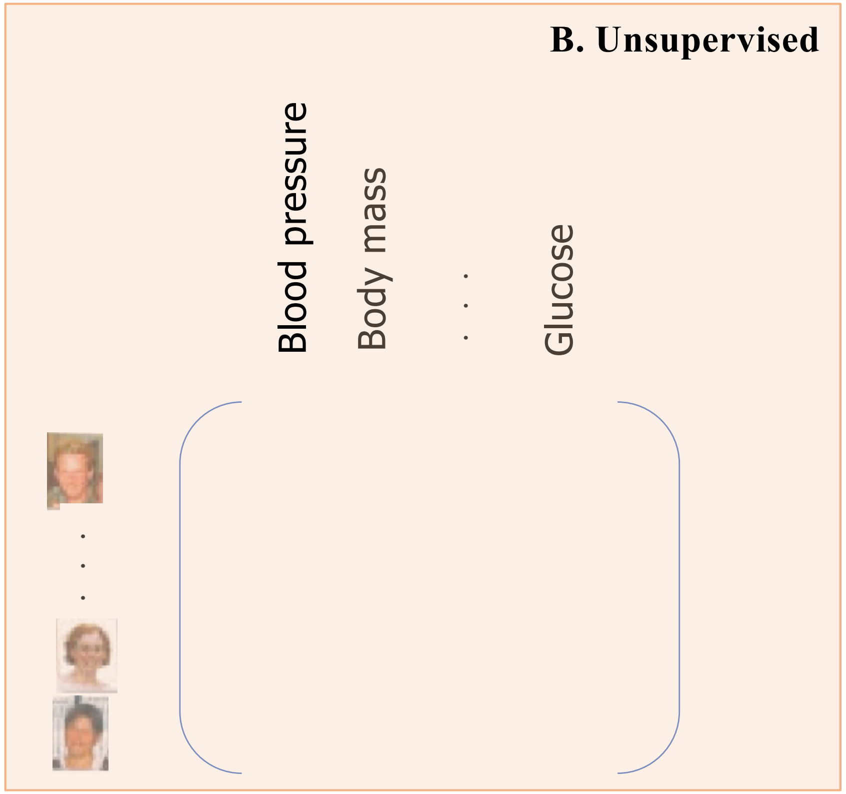
. . .

Glucose

C_1

C_1

C_2



B. Unsupervised

Blood pressure

Body mass

. . .

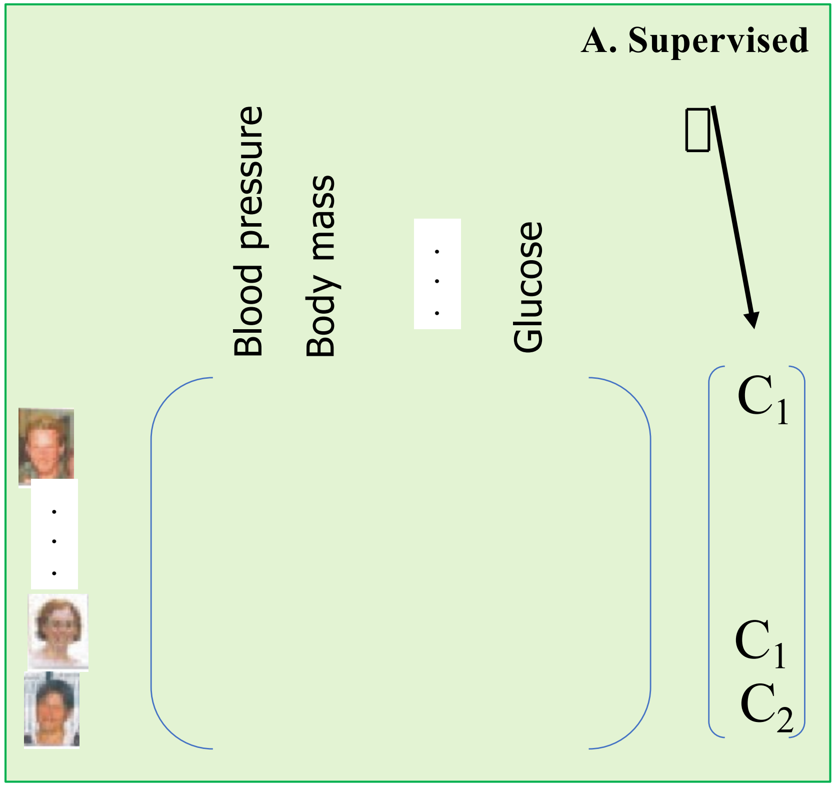
Glucose

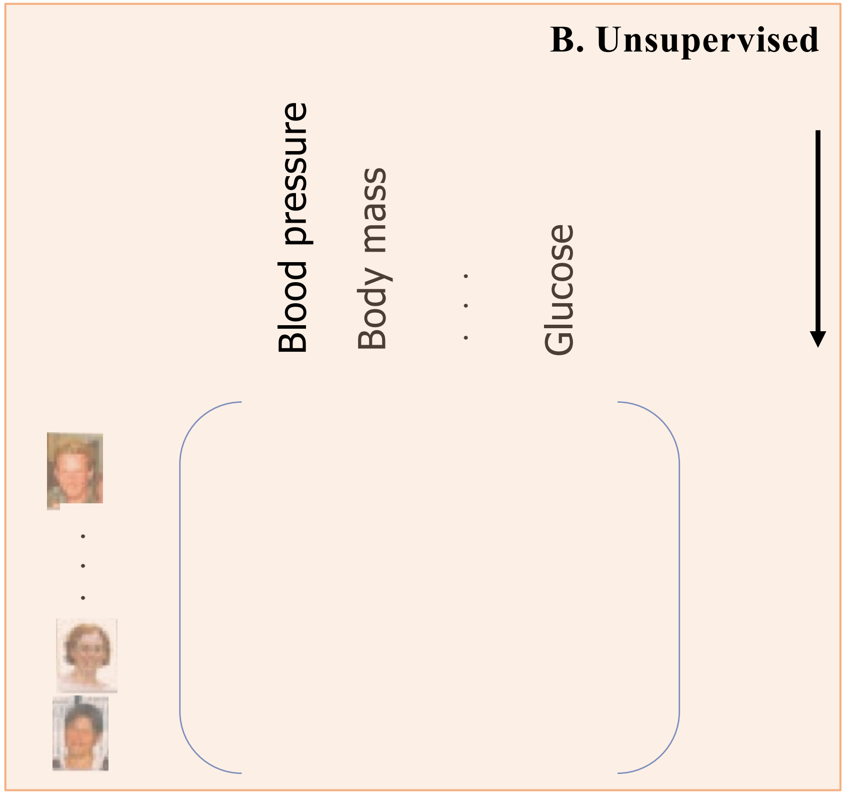
1.3 Supervised and unsupervised learning

Database:

* 133,000 patients
* 72 variables or characteristics/attribute

C_1 -> Attack C_2 -> Non-Attack





1.3 Supervised and unsupervised learning

Database:

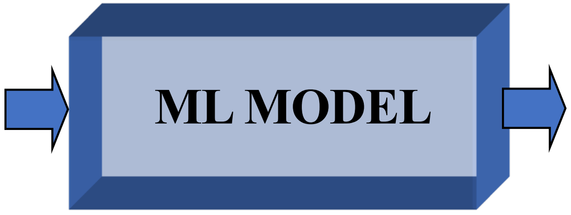
* 133,000 patients
* 72 variables or characteristics/attributes

$$\begin{array}{ll} \text{Blood pressure} \dashrightarrow & \\ \text{Body mass} \dashrightarrow & \\ & \\ \text{Glucose} \dashrightarrow & \end{array} \begin{bmatrix} x_1 \\ x_2 \\ . \\ x_{72} \end{bmatrix} \qquad \begin{array}{ll} \dashrightarrow & \\ \text{Body mass} \dashrightarrow & \\ & \\ \text{Glucose} \dashrightarrow & \end{array} \begin{bmatrix} x_1 \\ x_2 \\ . \\ x_{72} \end{bmatrix}$$

What is the relationship between the 72 variables and the risk of attack?

INPUT

$$\begin{array}{ll} \text{Blood pressure} \dashrightarrow & \\ \text{Body mass} \dashrightarrow & \\ & \\ \text{Glucose} \dashrightarrow & \end{array} \begin{bmatrix} x_1 \\ x_2 \\ . \\ x_{72} \end{bmatrix}$$



OUTPUT

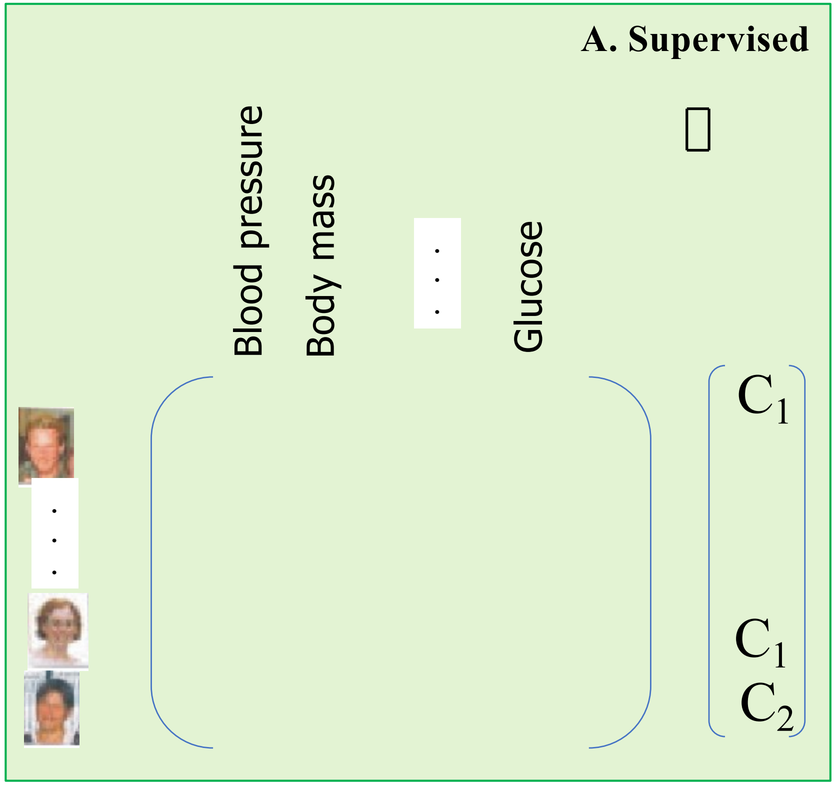
Attack--- "1"

Non-Attack--- "2"

1.3 Supervised and unsupervised learning

Supervised Learning

INPUT



A. Supervised

Blood pressure

Body mass

. . .

Glucose

. . .

C_1

C_1

C_2

Classification

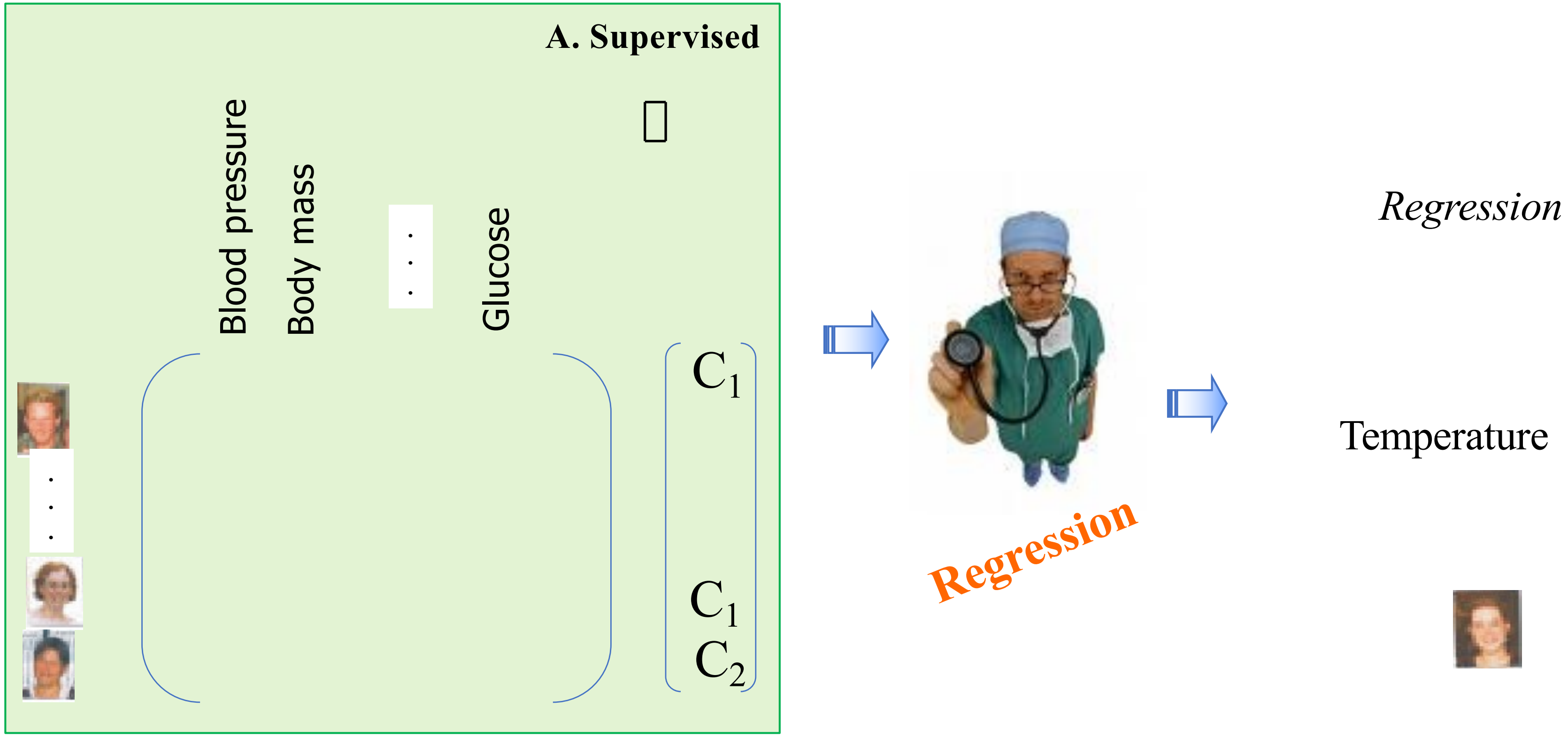
Binary classification

C_1 – Attack
C_2 – Non-attack

1.3 Supervised and unsupervised learning

Supervised Learning

INPUT



A. Supervised

Blood pressure

Body mass

. . .

Glucose

C_1

C_1

C_2

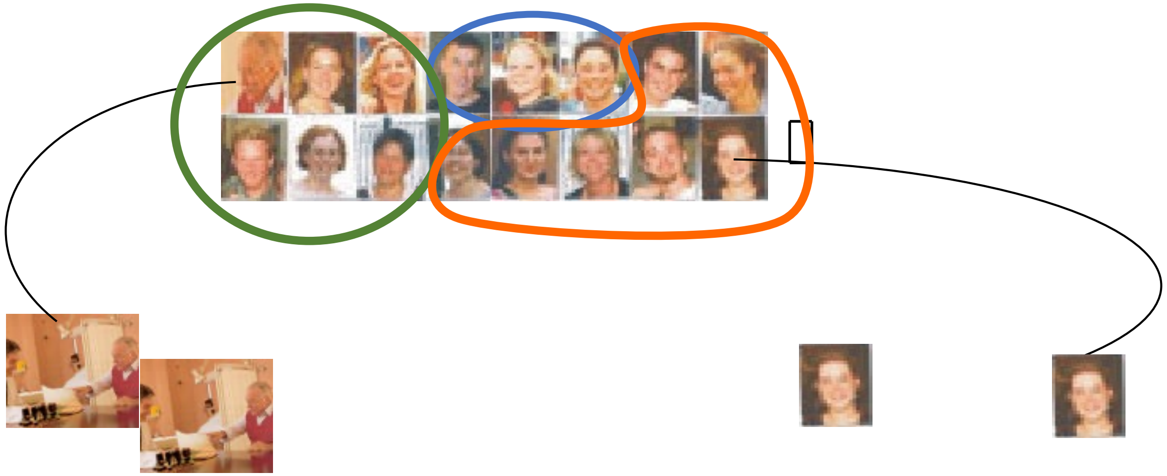
Regression

Temperature

Regression

1.3 Supervised and unsupervised learning

Unsupervised Learning



$$\begin{matrix}\text{Blood pressure} & \dashrightarrow \\ \text{Body mass} & \dashrightarrow \\ & \\ \text{Glucose} & \dashrightarrow\end{matrix}\begin{bmatrix}x_1\\x_2\\ .\\ x_{72}\end{bmatrix}$$

$$\begin{matrix}\text{Blood pressure} & \dashrightarrow \\ \text{Body mass} & \dashrightarrow \\ & \\ \text{Glucose} & \dashrightarrow\end{matrix}\begin{bmatrix}x_1\\x_2\\ .\\ x_{72}\end{bmatrix}$$

Contents

1.1 Artificial Intelligence and Machine Learning

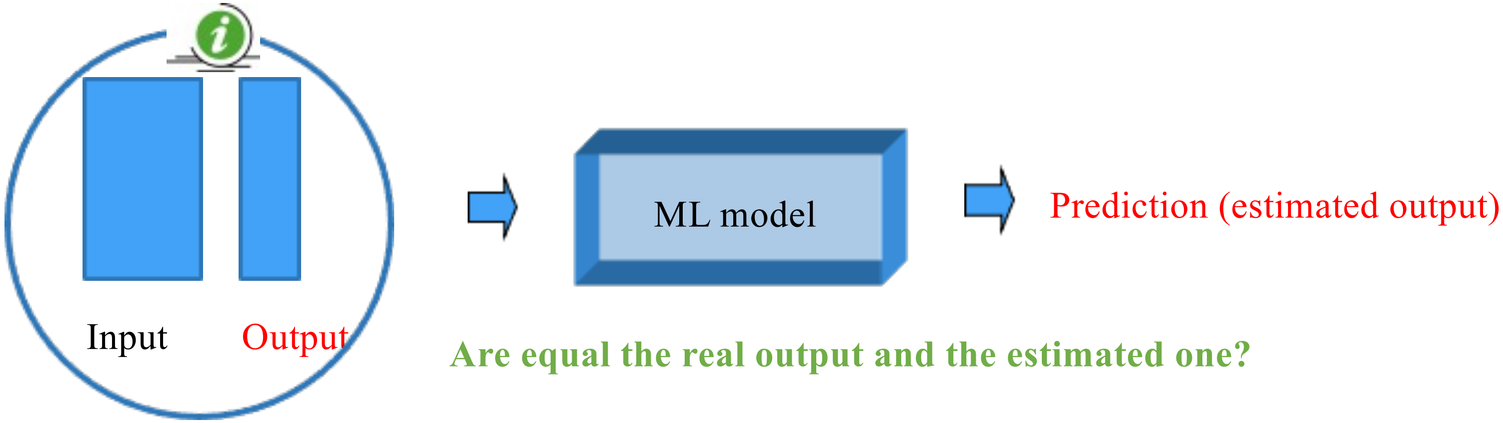
1.2 Data collection, cleaning and pre-processing

1.3 Supervised and unsupervised learning

1.4 Selection and evaluation of machine learning models

1.5 Biomedical examples and applications

1.4 Selection and evaluation of ML models



Input

Output

ML model

Prediction (estimated output)

Are equal the real output and the estimated one?

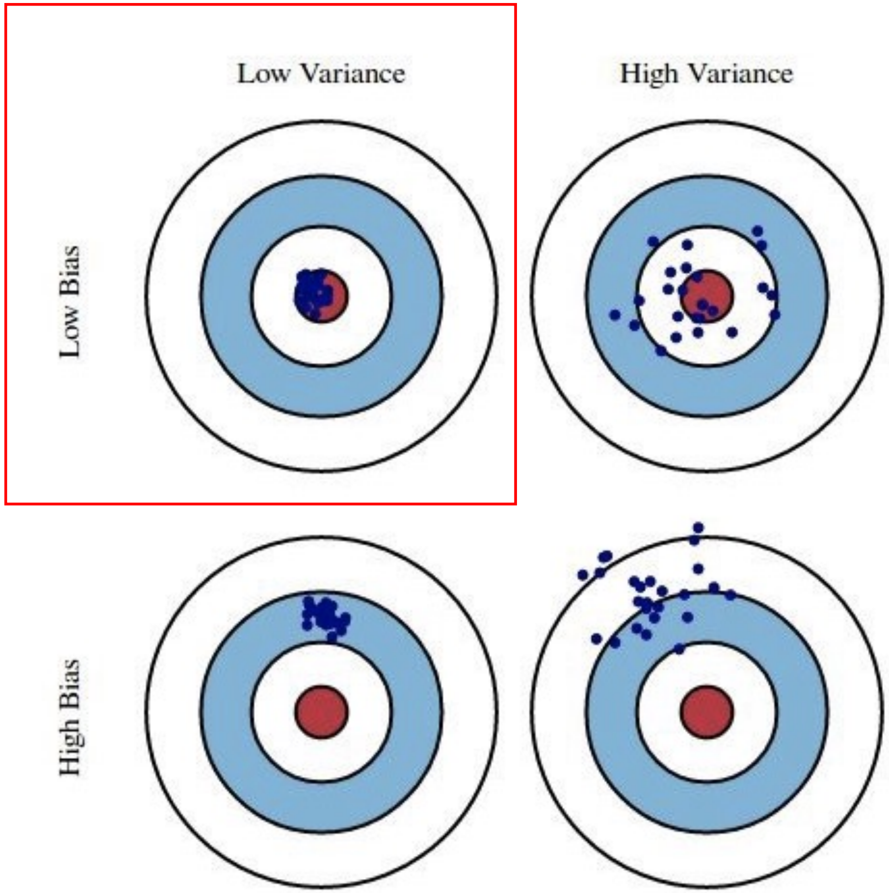
function(Input, *p1,p2,p3*, …..) = Output

p1,p2,p3, ..
p1,p2,p3, ..
p1,p2,p3, ..
p1,p2,p3, ..
p1,p2,p3, ..
p1,p2,p3, ..
p1,p2,p3, ..

In general, machine learning techniques use the examples (correct inputs and outputs) to iteratively search for the values of p1,p2,p3,... that best model the mathematical relationship between input and output.

1.4 Selection and evaluation of ML models

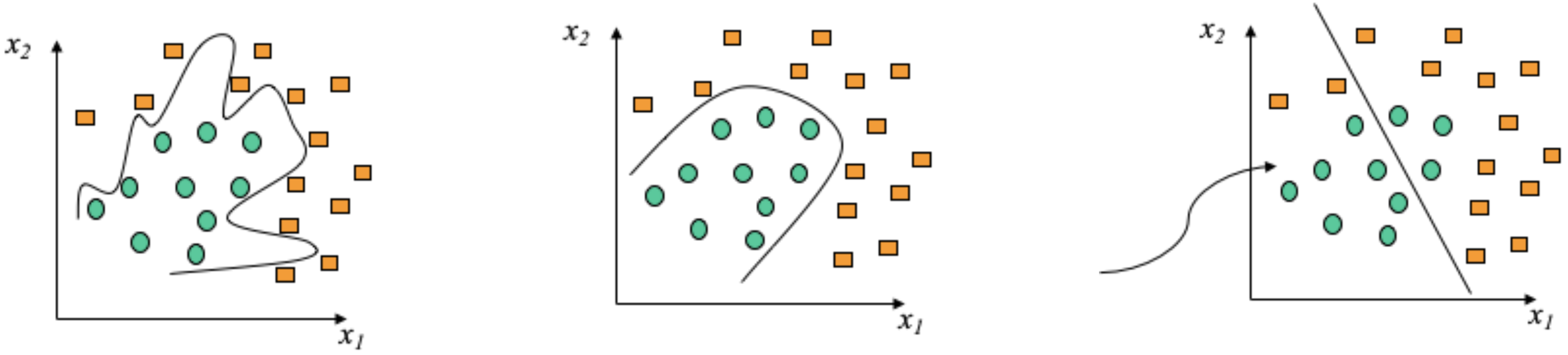
Trade-off between bias and variance



***High variance (overfitting)*.** The predictions/estimations are sensitive to sampling. When considering a different sample of data, the predictions will vary a lot

***High bias (underfitting)*.** The prediction/estimations based on the observed data are not close to the real value.

1.4 Selection and evaluation of ML models



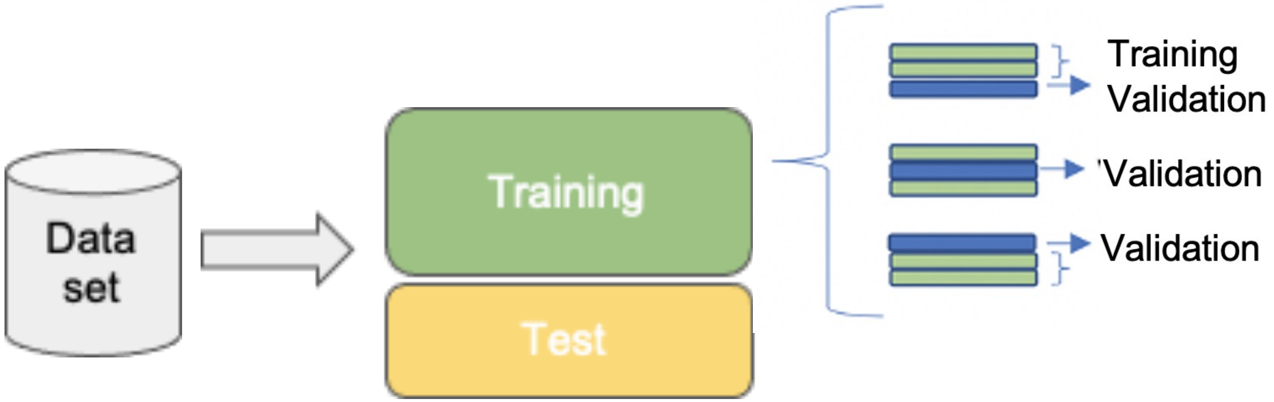
(Overfitting)

(Underfitting)

To achieve a model with good GENERALIZATION capacity, it is possible to evaluate the performance of the model with **cross validation techniques**

1.4 Selection and evaluation of ML models

How we select the best model?



K-Fold Cross Validation

Randomly shuffling the dataset and then splitting it into k subsets

Training
Test

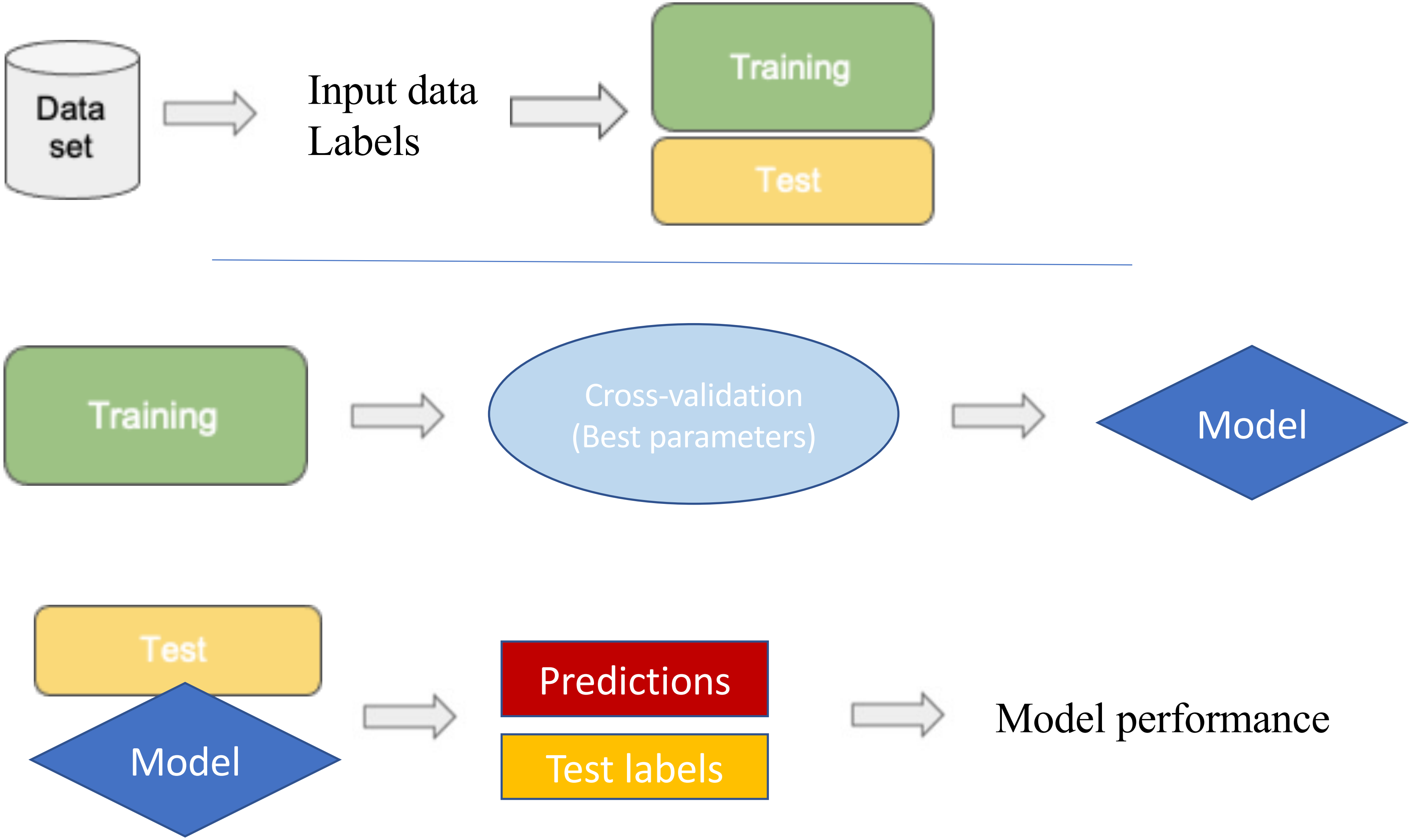
Desing the model

The best model is selected by choosing the one with the highest score in the validation/training set

Evaluate the designed model

1.4 Selection and evaluation of ML models

How we select the best model?



Data set

Input data
Labels

Training

Test

Training

Cross-validation
(Best parameters)

Model

Test

Model

Predictions

Test labels

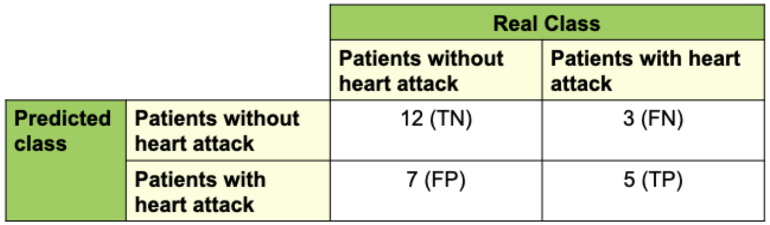
Model performance

1.4 Selection and evaluation of ML models

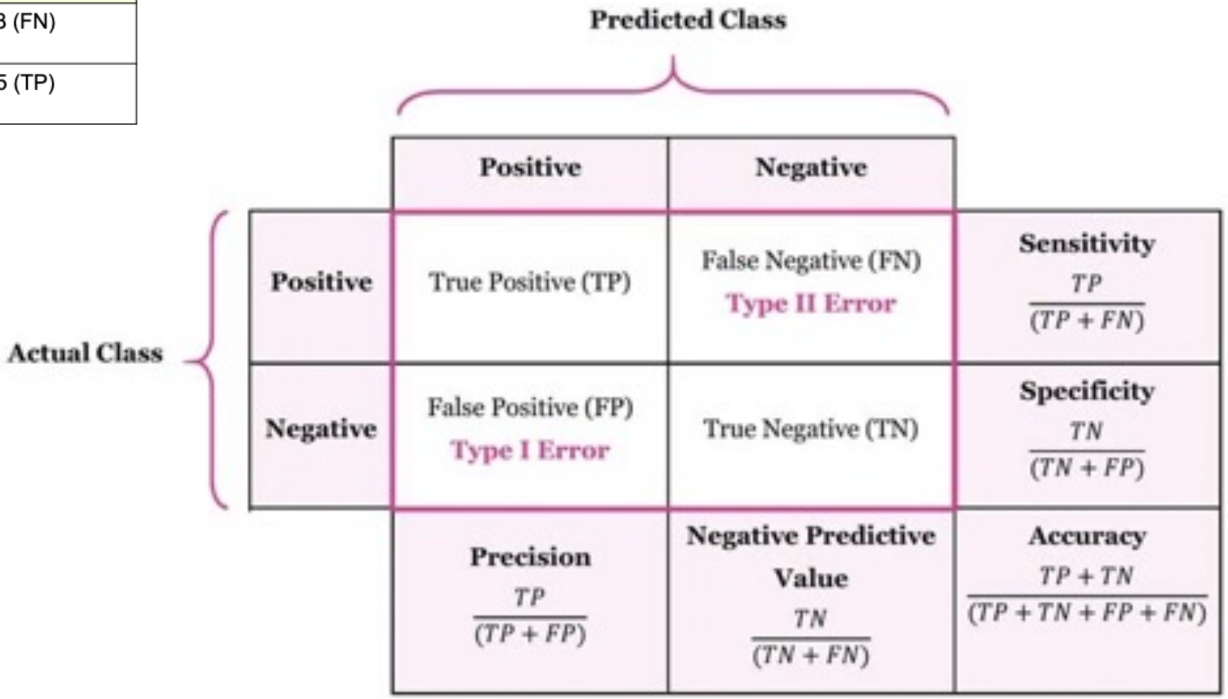
Model performance

Classification

Confusion matrix



		Real Class	
		Patients without heart attack	Patients with heart attack
Predicted class	Patients without heart attack	12 (TN)	3 (FN)
	Patients with heart attack	7 (FP)	5 (TP)



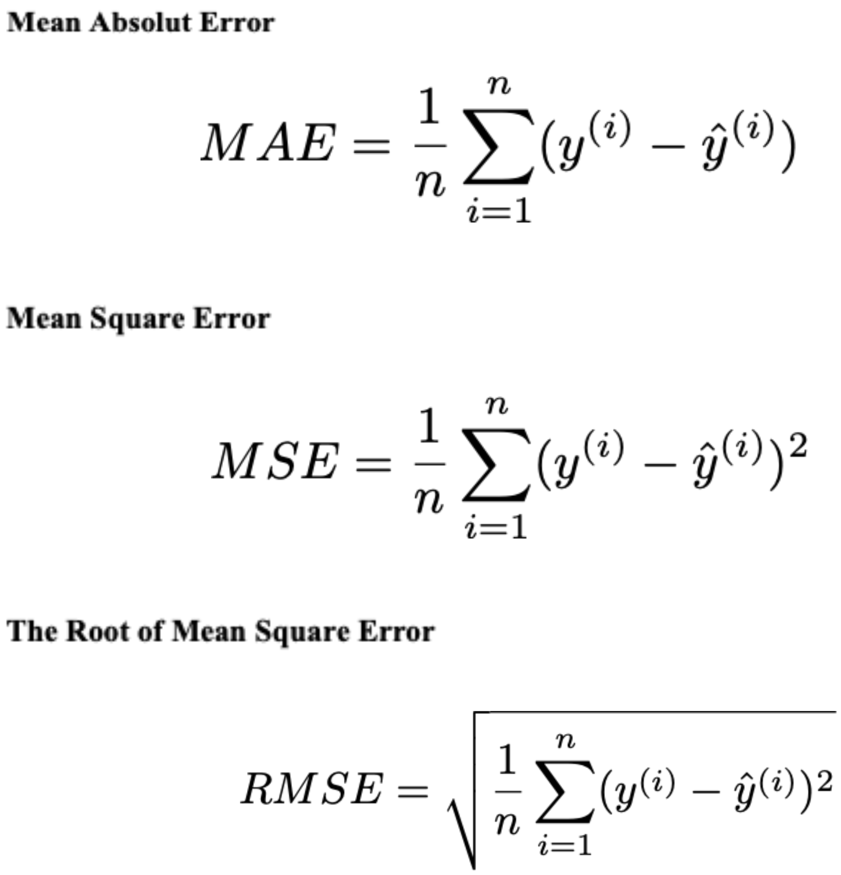
		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

ROC-AUC

1.4 Selection and evaluation of ML models

Model performance

Regression



Mean Absolut Error

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})$$

Mean Square Error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2$$

The Root of Mean Square Error

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2}$$

Contents

1.1 Artificial Intelligence and Machine Learning

1.2 Data collection, cleaning and pre-processing

1.3 Supervised and unsupervised learning

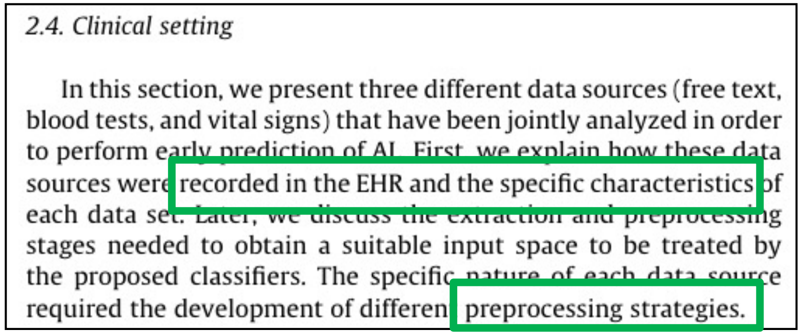
1.4 Selection and evaluation of machine learning models

1.5 Biomedical examples and applications

1.5 Biomedical examples and applications

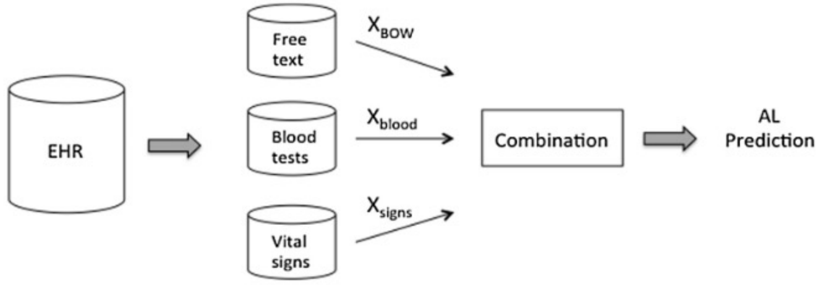
> J Biomed Inform. 2016 Jun;61:87-96. doi: 10.1016/j.jbi.2016.03.008. Epub 2016 Mar 12.

Predicting colorectal surgical complications using heterogeneous clinical data and kernel methods



2.4. Clinical setting

In this section, we present three different data sources (free text, blood tests, and vital signs) that have been jointly analyzed in order to perform early prediction of AL. First, we explain how these data sources were recorded in the EHR and the specific characteristics of each data set. Later, we discuss the extraction and preprocessing stages needed to obtain a suitable input space to be treated by the proposed classifiers. The specific nature of each data source required the development of different preprocessing strategies.



2.4.1. Free text. All documents related to both inpatient and outpatient visitsfrom 2004 to 2012 were extracted. The most frequent document types were nurses notes, journal notes, outpatient notes, radiologyreports, referrals, discharge letters, and admission notes.
Preprocesing. All words in the analyzed documents were transformed to lowercase, and all grammatical symbols, numbers, and stops words were removed. A **bag of word** (BOW) model was built based on the relative frequency of each word. Misspelledwords appeared relatively infrequently, thus, only words appearing at least 10 times were included in the BOW. This threshold actually reduced the dimensionality of BOW model from 65,328 to 13,188 words.

We considered it and the **mean zero and unit variance standardization**

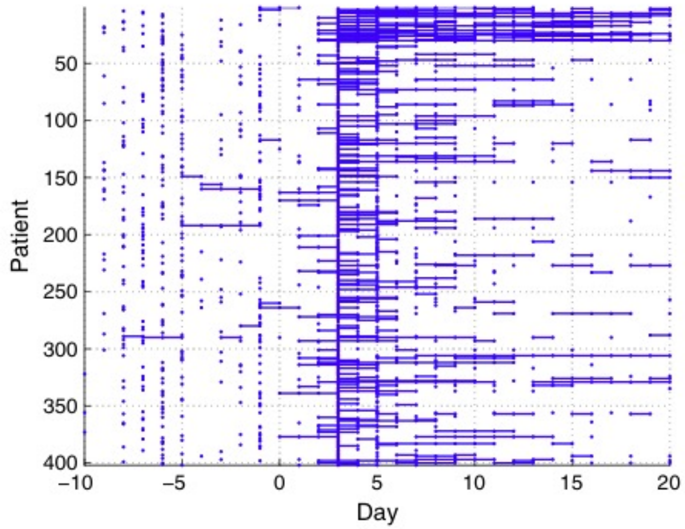
1.5 Biomedical examples and applications

> J Biomed Inform. 2016 Jun;61:87-96. doi: 10.1016/j.jbi.2016.03.008. Epub 2016 Mar 12.

Predicting colorectal surgical complications using heterogeneous clinical data and kernel methods

2.4.2. Blood tests. In this work, we analyzed structured data from nine different laboratory (blood) tests. These blood tests were recorded for a period of 10 days before the surgery and up to 20 days after the surgery. Note that the blood tests measurements are in general highly irregularly extracted in time. Hence, the observed data set is sparse over patients and time, which creates **challenges in the data processing.** From a data processing perspective, the data sparseness is equivalent to **missing data**, and the irregular sampling must be handled.

- When a relatively small number of samples are missing, skipping features or patients can be an option, but this was not the case in our problem.
- We followed an **imputation method** based on the nearest neighbour algorithm as in [36].



1.5 Biomedical examples and applications

> J Biomed Inform. 2016 Jun;61:87-96. doi: 10.1016/j.jbi.2016.03.008. Epub 2016 Mar 12.

Predicting colorectal surgical complications using heterogeneous clinical data and kernel methods

2.4.3. Vital signs. Three vital signs (temperature, blood pressure -high and low values-, and pulse) were extracted from different types of nurse'

Vital signs were normally recorded at least three times per day for each patient, for a period of 10 days before the colorectal surgery and up to 20 days after the surgery. However, these data were irregularly sampled by nature, thus, a causal imputation method based on the nearest neighbor algorithm was applied to obtain daily measures.

Temperature. The extraction process was restricted between 30.0 and 41.0 as normal values.
Blood pressure. The diastolic and systolic blood pressure of a patient was given as two integers separated by a /, for instance 120/80. The extraction process constrained it to be: (1) The first integer larger than the second integer; (2) The first integer larger than 60 and lower than 250; And (3) the second integer larger than 30 and lower than 200.
Pulse. The number of heart beats per minute was given as an integer. The extraction process restricted it to be between 41 and 250. Choosing 41 as the lower limit makes medical sense, though there might be rare cases of lower pulses than this. In these cases, the patient was probably anyway kept under tight control

References

- Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.