For each attribute we can have different level of measurement

A **VARIABLE** is a sort of **OPERATIONALIZED** version of the attribute

An **ATTRIBUTE** is just an intuitive general concept
The **VARIABLE** also define the set of all possible values that it can take

The **OPERATIONALIZATION** is the process of defining the values with a variable can take

## LEVEL OF MEASUREMENT

4 LEVELS.

- NOMINAL
- ORDINAL

Qualitative variables (Tall, short)

- INTERVAL
- RATIO

Quantitative variables (173 cm)

# QUALITATIVE ( CATEGORICAL )

- **NOMINAL**

When we can't establish a relation of order between variables we can say that the scale of measurement is nominal

- **ORDINAL**

derive from order.
is when we are observing variable that can be ordered.

When we can establish a relation of order between values we say that the scale of measurement is ordinal

$V_1 < V_2 \leq V_3$

ex: short < tall

The **order** is defined as a particular relation that has defined the property of

reflexivity ( $a \leq a$)
antisymmetry (if $a \leq b$ and $b \leq a$ then $a = b$)
transitivity (if $a \leq b$ and $b \leq c$ then $a \leq c$)

# QUANTITATIVE (NUMERICAL)

not all numerical variable are quantitative variable because not all number drive from a measurement

for example ID, credit card number are qualitative variables.

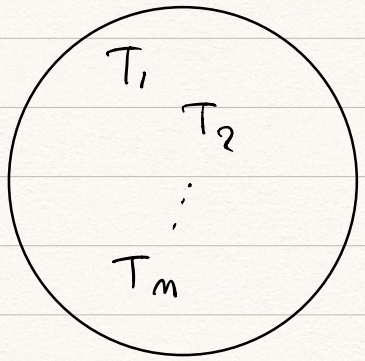
- **INTERVAL**

  the ratio does not make sense, but is important the differents between value

- **RATIO**

  The ratio between possible variables can make sense

# TIME SERIES ANALYSIS

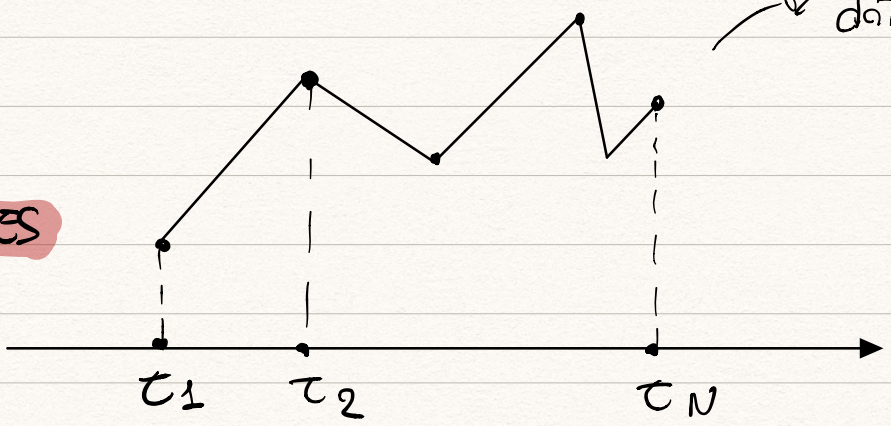



Instead of a set of items we have a set of **INSTANTS OF TIME**

For each istant of Time we have a set of observation (DATA POINT) which would be a The observation of some variable $(x_1, \ldots, x_K)$ at an istant

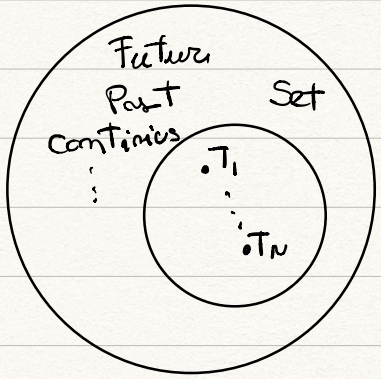
| time | $X_1$ | ... | $X_K$ |
|---|---|---|---|
| $t_1$ | $x_{11}$ | ... | $x_{1K}$ |
| ⋮ | ⋮ | | ⋮ |
| $t_N$ | $x_{N1}$ | ... | $x_{NK}$ |

**TIME SERIES**



A Time Series can be constituted by least one variable or multiple value.

We can order istances of time. So the first thing that diferentiate the TIME SERIES SETUP from other setup is that the statistical units have defined a VARIATION OF ORDER



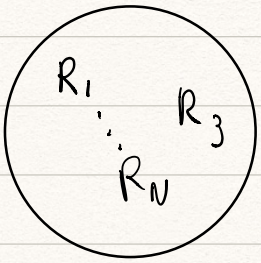

We can consider our set as a Sample

↳ INFERENTIAL nature intrisically

It is important as a financial instrument

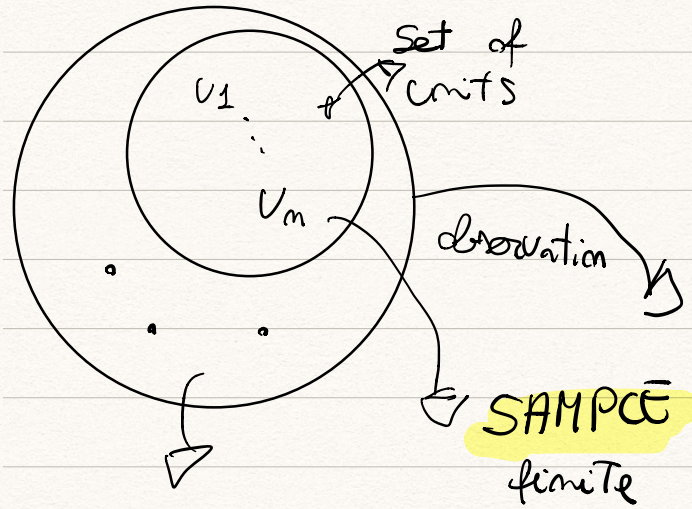
# SPATIAL DATA ANALYSIS

In Spatial Data analysis we have Region of space as statistical units




the regions of space haven't a natural order, but we can ab assumption To introduce some order.

We could order for longitude, latitude etc...

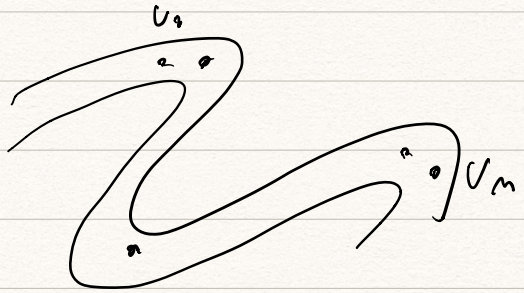


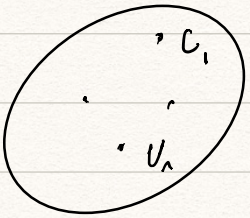

DATA SET

| DATA POINT (OBSERVATION) |
|---|

In Real world we consider a STREAM OF DATA of set of units of observation

DATA STREAM

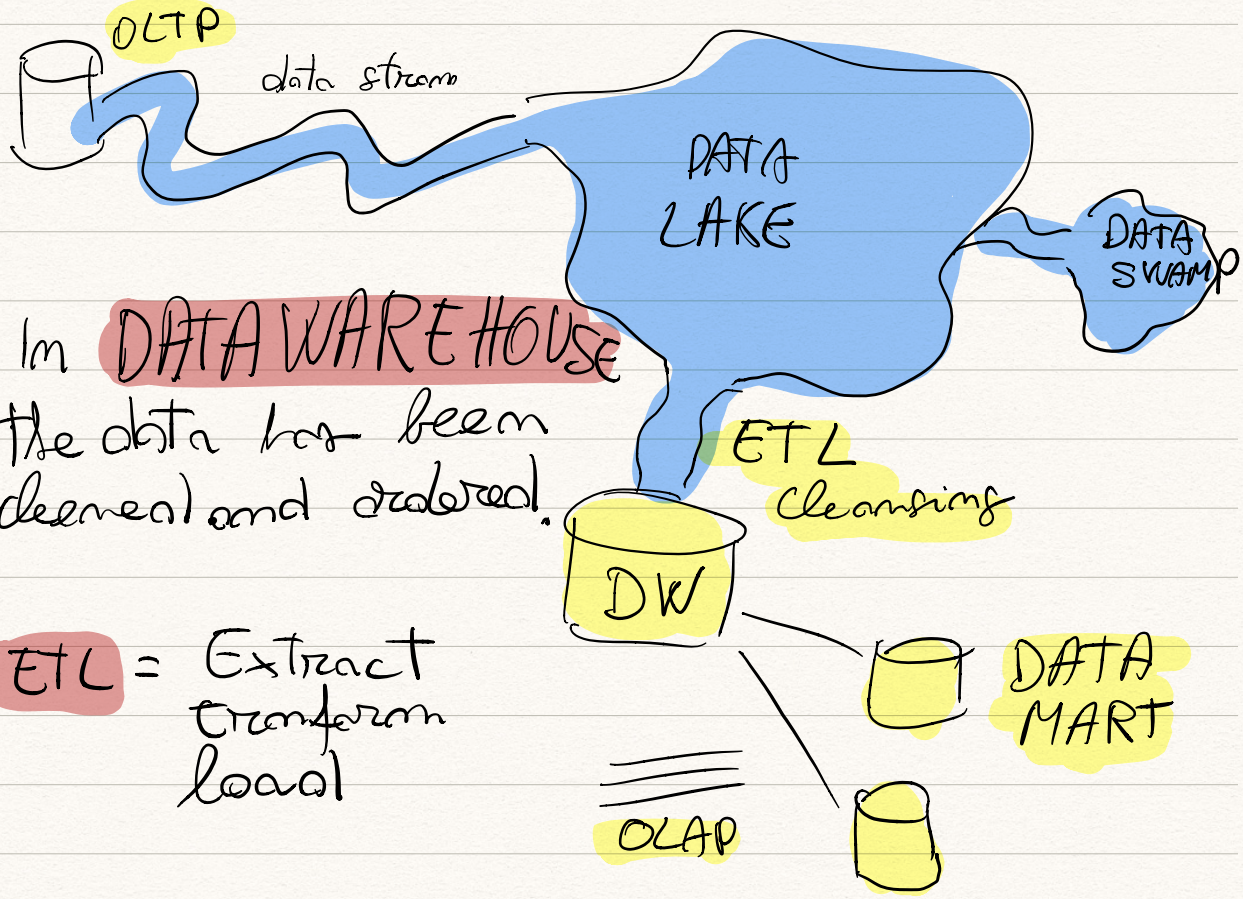



Set of units of observation

even if I have already collected the units I always Think as a stream establishing an order.

The data stream are stored in DATA LAKE or Data WAREHOUSE




In **DATA WAREHOUSE** the data has been cleaned and ordered.

**ETL** = Extract transform load

**DATA MART** are special data warehouse specialized

**OLAP** = Online Analysis Processing are operations of reporting, analysis and mining done to create DATA MART

**DATA SWAMP** = Same data more confusionary and less usefull.

OLTP = Online Transactional Processing
Operative type of system where can collect data

# PROCESS DATA STREAM

| | |
|---|---|
| Ex 1 | 30 |
| Ex 2 | 28 |
| Ex 3 | 20 |

1) Store the data
2) Data Processing

30
28
:
25

$$\frac{30 + 28 \dots + 25}{\text{count}}$$

OR

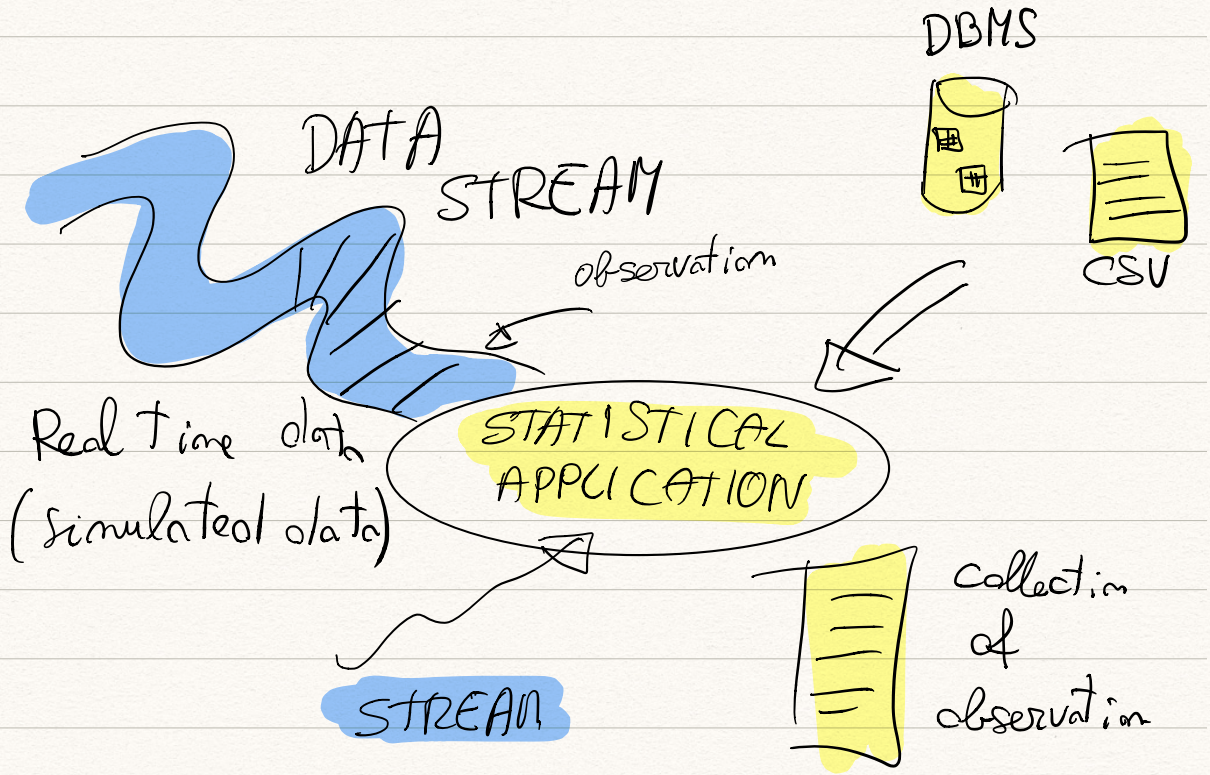
# STREAM PROCESSING

1) RECIEVE DATA
2) PROCESS AND UPDATE THE RESULT

$$Avg = \frac{\bar{x} + x_n \cdot n}{n+1}$$

$\bar{x}$ = latest data received

If we have an infinite stream of data we can't store all data and then process it, we need a streaming algorithm to do streaming processing




An application can work on datastream or an external collection of observation

- We can store all dataset in memory → BATCH PROCESSING (offline)
- We reserve a limit amount of data recieved from a stream → STREAM PROCESSING (online)

When we read only once the data we say that the algorithm is **ONE PASS** otherwise the algorithm is **MULTI PASS**

↓ single pass

This distinction is applicable to both Batch processing and stream processing

# FONDAMENTAL CONCEPTS

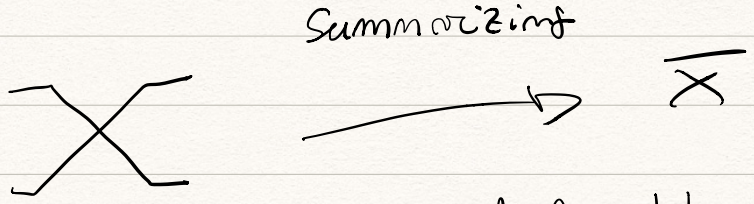
- **RANDOM OBJECT**
  generator of pseudo random number

- **TIMER**
  permit to generate event spaced by a determinated interval of time

**COLLECTION of object**

↳ collection of types

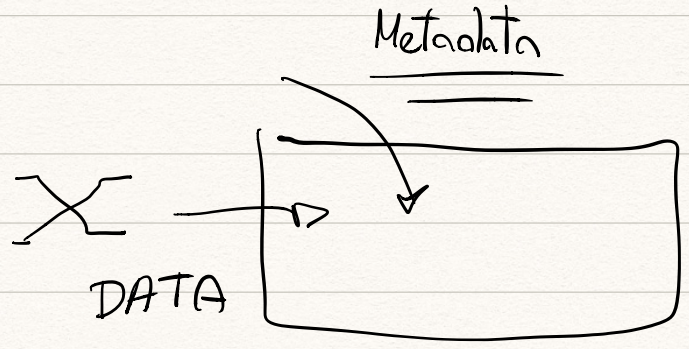
- Array
- List
- Dictionary
- Queue
- Stack
- Sorted list
- ⋮




the whole dataset into a unique "representative" value

for example in the case of an aritmetic mean we store the sum of the units as representative value.

METADATA is about the meaning of the ~~DATA~~ and they are also important.




Metadata will change the interpretation of the data, so is important to store it.

# PSEUDOCODE

IS a plain language description of how the algorithm is going to work

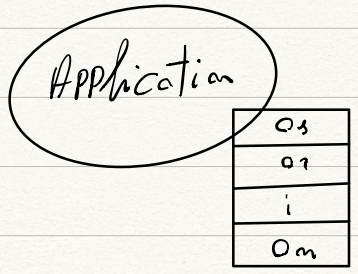
We need to define a **CLASS** that represent the statistical units and within the class we can define the properties that define the variable

↳ the instance of the class will be the **DATA POINT**

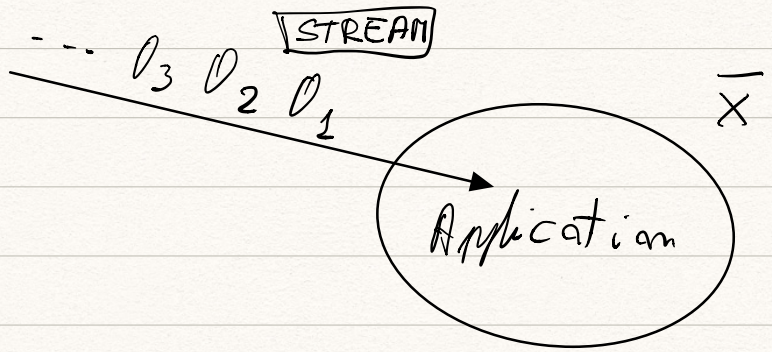
The **DATA** is just the content of the object

The **METADATA** is represented by the CLASS itself.

In the BATCH processing the application will load all the set of object



In STREAM Processing we feed the application with a continuous stream of object




$\bar{X}_i = \bar{X}_{i-1} +$ update Term

Online algorithm (On the fly)