Intelligent System for Industry, Supply Chain and Environment

Indice

1 Introduzione **3**
1.1 Major AI advancements [overview] 4
1.1.1 Major AI advancements in the last 2 years 4

2 Legislation, Artificial and human learning, Gestalt, applications and opinions about AI **5**
2.1 Laws about AI in Europe 5
2.1.1 Four risk levels 6

3 Intelligent transportations and Vehicle to everything protocol Examples of IoT, IoT security Artificial Intelligence of Things (AIoT), HW/SW environments for IoT/AIoT **7**

4 Data Gathering, Data Preprocessing, Data Harmonization for intelligent system learning **8**
4.1 Value of Data Analytic 8
4.2 Data gathering - Step 1 9
4.2.1 IoT started to generate data... 9
4.2.2 Data heterogeneity and synchronization 10
4.2.3 Data Synchronization 10
4.3 Data Preparation - Step 2 11
4.3.1 Data wrangling 12
4.3.2 Missing Data 13
4.3.3 Structured and unstructured Data 14

5 Managing a small dataset in Python Degrees of freedom/parameters Data Leakage **15**
5.1 Degrees of freedom /parameters of the models - Step 1 and 3 . . 15
5.1.1 How much data? Degree of freedom/parameters 16
5.1.2 Similitude 1 16
5.1.3 Similitude 2 16
5.1.4 Number of Parameters of NNs 17
5.1.5 a 1D linear model 17
5.1.6 a 2D linear model 17

5.1.7 a 3D linear model 18
5.1.8 DoF in general 18
5.1.9 In brief... 18
5.2 Data leakage - Step 2 19
5.2.1 Data Leakage can happen 21
5.2.2 Time series: a special case for Regression and Classification 21
5.2.3 Checking the presence of Data Leakage 21

Capitolo 1

Introduzione

What is an intelligent system (IS)

a computer-based system that aims to replicate human cognitive abilities such as learning, perception, reasoning, and decision-making.

By utilizing Machine Learning (ML), and other related technologies, these systems are capable of processing and analyzing data to perform tasks that typically require human intelligence, make predictions, or provide insights.
Some examples of intelligent systems are:

- virtual assistant like Siri and Alexa
- autonomous vehicles
- image recognition
- fraud detection
- *and many others . . .*

Diffusion and relevance

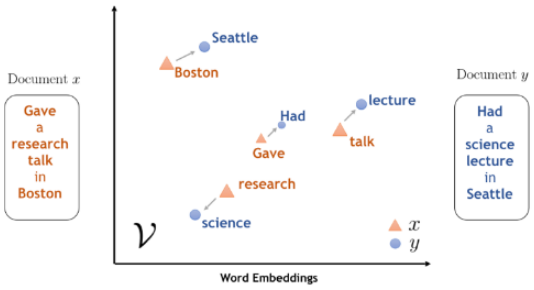
In particular, ML allows us to process data at unprecedented scales:

- see patterns
- detect problems earlier
- allocate resources more effeciently

1.1 Major AI advancements [overview]

Some recent improvements

- so-called **word embeddings** that are used as input to a Neural Network; it is a set of techiniques in NLP where words are mapped to **vectors of real numbers**



- solving **analogy puzzles**

  Paris is to France as Tokyo is to . . . ?

  `Japan`

1.1.1 Major AI advancements in the last 2 years

AI has made significant progress across various domains, revolutionizing industries and starting to create profits. . .

Capitolo 2

Legislation, Artificial and human learning, Gestalt, applications and opinions about AI

2.1 Laws about AI in Europe

The AI Act is a European law on artificial intelligence (AI), the first comprehensive law on AI by a major regulator anywhere. The AIA was published in the Official Journal of the EU on 12 July 2024 and entered into force on 1 August 2024 .

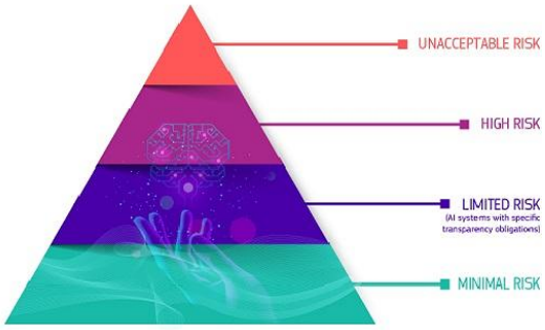
Why do we nee rules on AI?

- To avoid undesirable outcomes
- It is often not possible to find out why an AI system has made a decision or prediction and taken a particular action.
- It may become difficult to assess whether someone has been unfairly disadvantaged, such as in a hiring decision or in an application for a public benefit scheme

According to the European Union's Artificial Intelligence Act (AI Act), an AI system is defined as:

"a machine-based system designed to operate with varying levels of autonomy and that may exhibit, for explicit or implicit objectives , infers from the input it receives how to generate outputs such as predictions , content , recommendations, or decisionsthat can influence physical or virtual environments"

2.1.1 Four risk levels



- **Unacceptable risk:** All AI systems considered a clear threat to the safety, livelihoods and rights of people will be banned
- **High risk:**
  - critical infrastructures (e.g. transport), that could put the life and health of citizens at risk;
  - educational or vocational training, that may determine the access to education and professional course of someone's life (e.g. scoring of exams);
  - safety components of products (e.g. AI application in robotassisted surgery);
  - employment, management of workers and access to selfemployment (e.g. CV-sorting software for recruitment procedures);
  - essential private and public services (e.g. credit scoring denying citizens opportunity to obtain a loan);
  - law enforcement that may interfere with people's fundamental rights (e.g. evaluation of the reliability of evidence);
  - migration, asylum and border control management (e.g. verification of authenticity of travel documents);
- **Limited risk:** refers to AI systems with specific transparency obligations

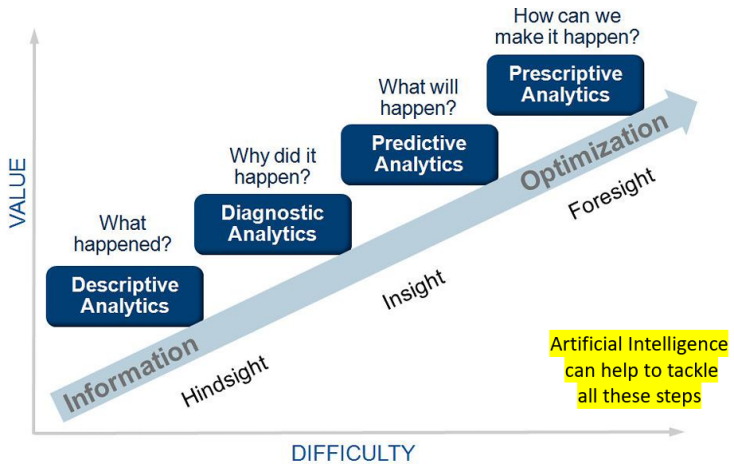
Capitolo 3

Intelligent transportations and Vehicle to everything protocol Examples of IoT, IoT security Artificial Intelligence of Things (AIoT), HW/SW environments for IoT/AIoT

Capitolo 4

Data Gathering, Data Preprocessing, Data Harmonization for intelligent system learning

4.1 Value of Data Analytic

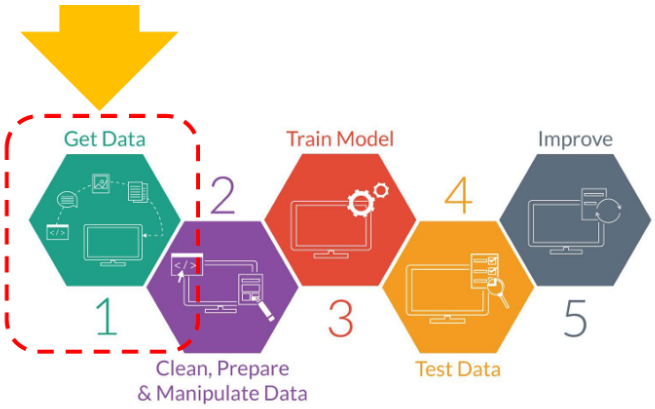


Four possibility increasing the value and the difficulty:

- Descriptive Analytics
- Diagnostic Analystics

- Predictive Analytics
- Prespective Analaytics

4.2 Data gathering - Step 1



We have so many powerful sources capable to generate data.
Data collection is the process of gathering and measuring information on targeted variables in an established system

4.2.1 IoT started to generate data...

The amount of data generated by connected internet of things (IoT) devices, forecast to grow to by 2025

- 41.6 billion connected devices
- 79.4 zettabytes (ZB) of data/year.

Data sources like phones, smart watches, ecc..

Example:
Basic GPS coordinates from smartphones @ITA

DataFromPositions/y = ItalianPopulation x CellPhoneRatio x 365 days x 24h/d x 60 min/d x 2coordinates/min = 65x10^6 x 0,83 x 365 x 24 x 60 x 2 x **8byte** = 697996800 byte ≈ 0,67GB

There are pubblic data centers, like Amazon's, Google's and Governative's ones

4.2.2 Data heterogeneity and synchronization

Heterogeneity in statistics means that your populations, samples or results are different. It is the opposite of homogeneity which means that the, population/data/results are the same. Qua fa un tot di esempi su come sia importante avere tutti i dati nello stesso formato (esempio di Marte e della NASA e della wind station)

4.2.3 Data Synchronization

The way a device adjusts its internal clock in order to align with the clocks of other devices in a network

Network Time Synchronization

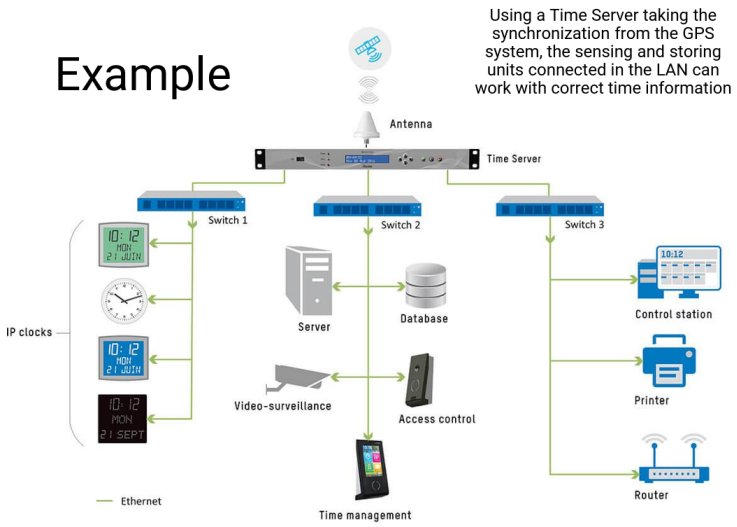
Computer clocks in servers, workstations and network devices are inherently not enough accurate Two problems:

- Clocks are set by hand to within a minute or two of actual time and are rarely checked after that
- Clocks are maintained by a battery-backed device that may drift as much as a second per day

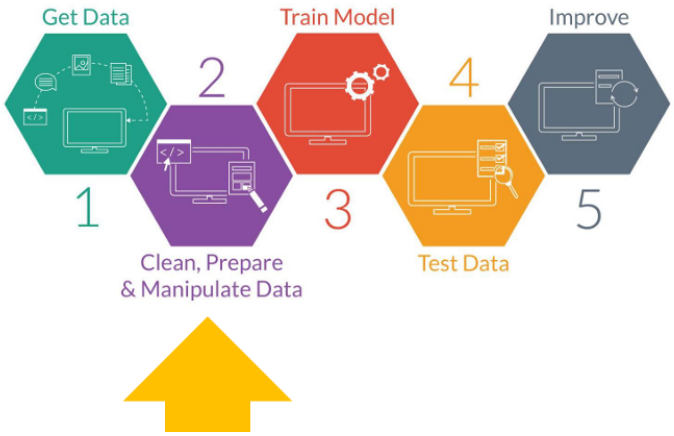
It's impossible to have accurate time synchronization without a proper method

Solutions

- **Network Time Protocol (NTP):** is a protocol for clock synchronization between computer systems over packetswitched, variable-latency data networks designed to mitigate local network latency
- **Time Server:** Dedicated network Time Server behind your firewall (devices synchronized to within 1/2 to 2 ms)



4.3 Data Preparation - Step 2

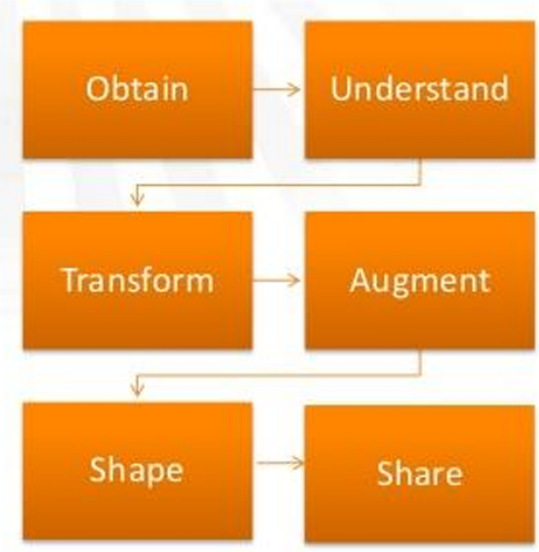


Data preparation includes two concepts such as Data Cleaning and Feature Engineering
The data wrangling problem is growing as different types of unstructured data or data in varying formats are pouring in from sensors, online and from traditional databases. All these data must be **cleaned up and organized** before data analytics/classifiers/regressors models can be applied.

4.3.1 Data wrangling

Data wrangling steps:

- Iterative process
- Understand
- Explore
- Transform
- Augment
- Visualize



Tasks of Data Wrangling:

- **Discovering:** Firstly, data should be understood thoroughly and examine which approach will best suit.
- **Structuring:** As the data is gathered from different sources, the data will be present in various shapes and sizes. Therefore, there is a need for structuring the data in proper format.
- **Cleaning:** Cleaning or removing of data should be performed that can degrade the performance of analysis.
- **Enrichment:** Extract new features or data from the given data set to optimize the performance of the applied model.

- ***Validating:*** This approach is used for improving the quality of data and consistency rules so that transformations that are applied to the data could be verified.

Data pre-processing: "is a technique that is used to convert the raw data into a clean data set" Pre-processing includes • Data cleaning • Data integration • Data transformation • Data reduction
Why is Data Preprocessing is so important? Three answers:

- Inaccurate data (missing data)
- The presence of noisy data/erroneous data/outliers
- Inconsistent data

4.3.2 Missing Data

What do we do when we have missing data?

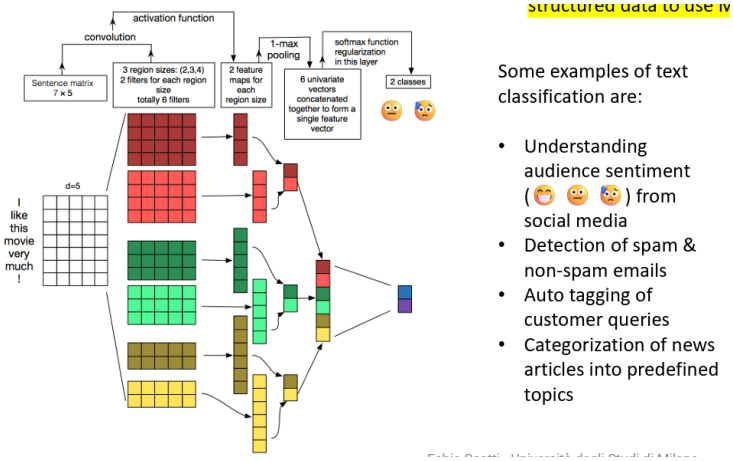
- **Ignoring the missing record:** is the simplest and efficient method for handling the missing data (not the best method when the number of missing values are immense or when the missing data problem and can solved (debugging/re-designredoing the experiment) and not just ignoring the problem causing the missing data.).
- **Filling the missing values manually:** one of the best-chosen methods, But there is one limitation that when there are large data set, and missing values are significant then, this approach is not efficient as it becomes a timeconsuming task.
- **Filling using computed values:** The missing values can also be occupied by computing mean, mode or median of the observed given values (ex: you can copy from the most similar column or generate values by using any ML or Deep Learning algorithm but it can generate bias within the data).

4.3.3 Structured and unstructured Data

Structured data usually resides in relational databases, This format is eminently searchable both with human generated queries and via algorithms using type of data and field names, such as alphabetical or numeric, currency or date. **Unstructured data** is essentially everything else. Unstructured data has internal structure but is not structured via pre-defined data models or schema (ex: sensor data, text files, emails, etc..).

Neural networks and unstructured data

It is not strictly compulsory to have structured data to use ML

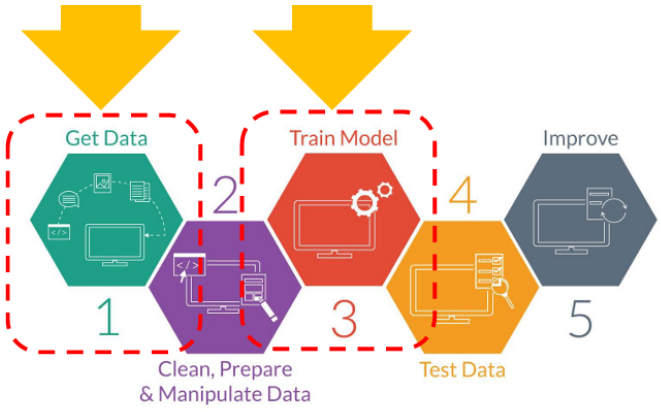


Capitolo 5

Managing a small dataset in Python Degrees of freedom/parameters Data Leakage

Lab in Pythone, c'è codice all'esame ma non che dobbiamo fare noi da zero(però possibili domande sul codice)

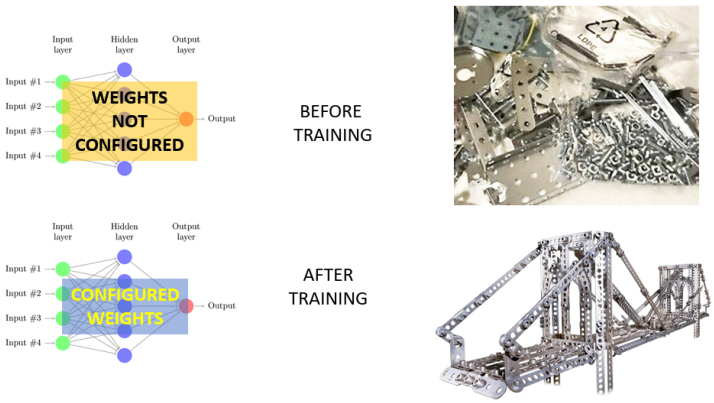
5.1 Degrees of freedom /parameters of the models - Step 1 and 3



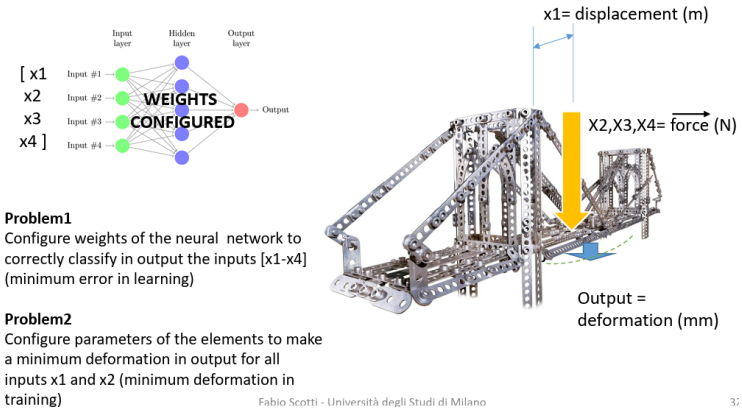
5.1.1 How much data? Degree of freedom/parameters

In physics, the degree of freedom of a mechanical system is the number of independent parameters that define its configuration.
Note: DoF is not exactly equivalent Par for complex systems, but they are strongly related.

5.1.2 Similitude 1



5.1.3 Similitude 2

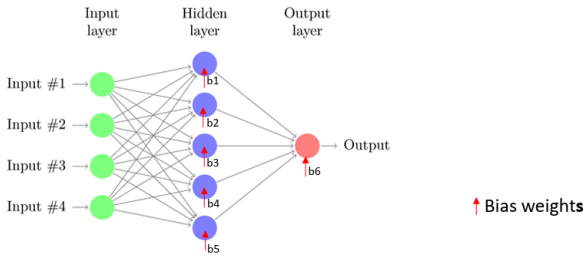


Problem1
Configure weights of the neural network to correctly classify in output the inputs [x1-x4] (minimum error in learning)

Problem2
Configure parameters of the elements to make a minimum deformation in output for all inputs x1 and x2 (minimum deformation in training)

Fabio Scotti - Università degli Studi di Milano 32

5.1.4 Number of Parameters of NNs



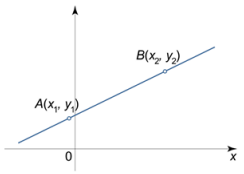
This simple neural network with 1 single hidden layer
has 4x5 (hidden weights) +5 (bias neuron) = **25** neuron weights (hidden layer),
and in the final neuron we have 5 connections +1 bias weight values
Total = 25 + 6 = 31 parameters to be fixed

How many input data [x1,x2,x3,x4] are need in the training to fix properly the weights?

5.1.5 a 1D linear model

We have the weight (w1) and the bias (beta)

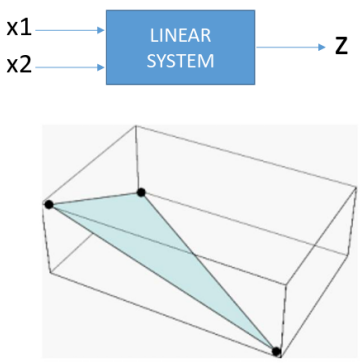
z = alpha x + beta
z = w1 x + b



DoF = #Par = 2
To completely describe the model
you need to fix 2 parameters:
w1, b.
→ You need 2 data points!

5.1.6 a 2D linear model

Now we have 3 parameters to fix



Vectorial form

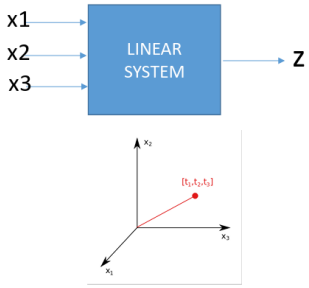
$$z = \mathbf{w} \cdot \mathbf{x} + b$$

$$z = [w_1\ w_2\ ] \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + b$$

DoF = #Par = 3
To completely describe the model you need to fix 3 parameters: w1, w2, b.
→ You need 3 data points!

5.1.7 a 3D linear model

Increasing the number of inputs I increas the number of parameters to fix



Vectorial form

$$z = \mathbf{w} \cdot \mathbf{x} + b$$

$$z = [w_1\ w_2\ w_3] \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + b$$

To completely describe the model you need to fix 4 parameters: w1, w2, w3, b.
→ You need 4 data points!

That is (almost) true also for non linear systems

By the way.... that is (also) the linear output of a neuron

Fabio Scotti - Università degli Studi di Milano 37

5.1.8 DoF in general

The degrees of freedom for a given problem are the number of independent problem variables which must be specified to uniquely determine a solution.

- degrees of freedom = variables - equations
- database are vectors so number of data = number of vectors in our database

5.1.9 In brief...

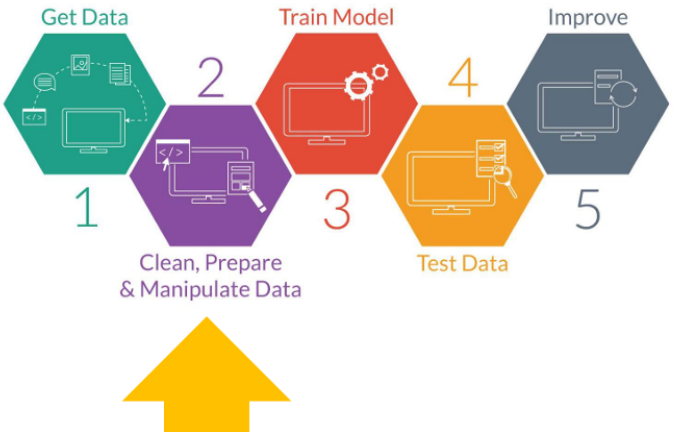
- The number of Par of the model (e.g., neural network) must be carfully tuned according to

- the size of the datasets (Number of vectors, Number of inputs)
- its complexity

≪Go deep≫ only if it is really necessary

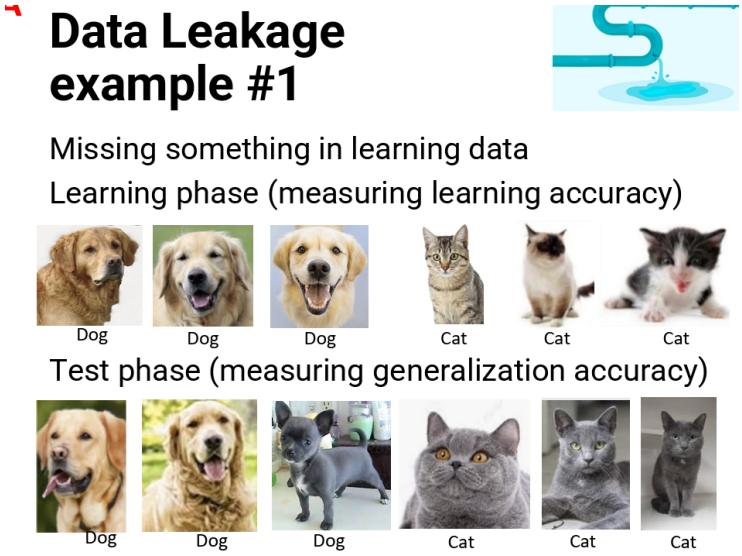
5.2 Data leakage - Step 2

Data Leakage is responsible for the cause of invalid Machine Learning/Deep Learning model due to the over optimization of the applied model.

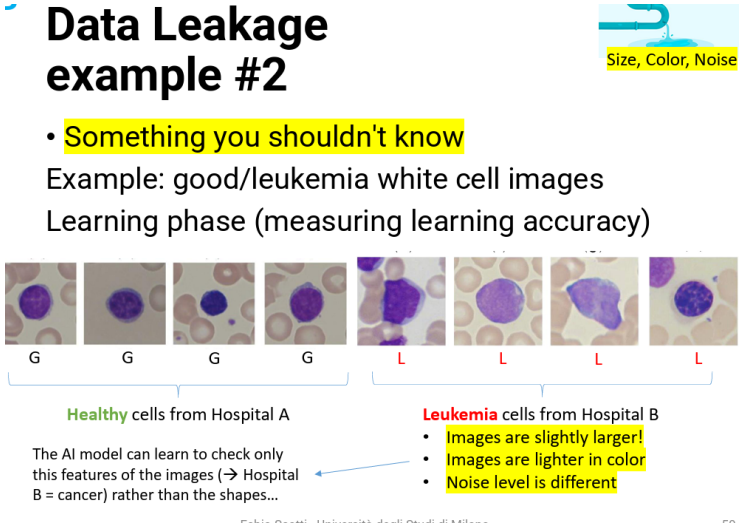


Two main topics:

- **Missing relevant features:** For example, when we want to use a particular feature for performing Predictive Analysis, but that specific feature is not present at the time of training of dataset (Example: you want to add to your dataset the concentration of OrmonX to predict CancerZ but OrmonX is not (almost) present in the trainig dataset.)



- **Adding something more..:**When information from outside the "expected" training information in the dataset is used to create the model. This additional information can allow the model to learn or know something that it otherwise would not know and in turn invalidate the estimated performance of the model being constructed. This additional learning of information by the applied model will disapprove the computed estimated generalization performance of the model, your performance estimation of the model once it will be deployed tends to be too much optimistic.



5.2.1 Data Leakage can happen

- The Leakage of data from test dataset to training dataset
- Leakage of future data into the past data
- Usage of data outside the scope of the applied algorithm

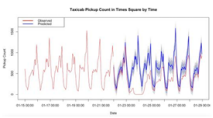
In brief, we have two primary sources of data leakage in Machine Learning algorithms:

- **Feature attributes (variables are saying too much. . .)**
- **Training data set (chunk of data used in the wrong phase)**

5.2.2 Time series: a special case for Regression and Classification

Time Series Prediction

Definition: Estimating **future** values based on **past data**.



Time Series Classification

Definition: Assigning **labels** based on **past data**.

Problem: in both cases, the model can be effective is the past data is consistent

- *training* (input to the learning method)
- *inference*(input to the model)

Data Leakage is observed in time-related complex datasets such as: dividing time series the dataset can be an error-prone problem

5.2.3 Checking the presence of Data Leakage

Data Leakage is observed in timerelated complex datasets such as:

- Storage of analog observations in the form of audios and images in separate files having a defined size and timestamp
- Implementation of sampling in a graphical problem is a complex task

The cropping problem

The **cropping problem** is a form of data leakage that affects multiple types of datasets and applications. It occurs when a model learns patterns from cropped or contextually biased data, leading to misleadingly high performance but poor generalization.
Is a form of data leakage where you cat to much or not enought information.

- **Images:** A model trained to classify objects may unintentionally rely on cropped edges or artifacts from image preprocessing rather than actual object features.
- **Audio:** A speech recognition model may learn background noise patterns from cropped samples instead of focusing on spoken words.
- **Structured Data:** In medical diagnosis, if training data is cropped to contain only extreme cases, the model might fail on intermediate cases.
- **Unstructured Data (Text):** A sentiment analysis model trained on social media posts might learn to rely on specific truncated phrases instead of full context.