LESSON 7

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




Outline

Data Gathering

Where the data is coming?

How much data is coming?

Data Preparation

Data control,

Unstructured data

Data harmonization

Data analysis

Data filtering

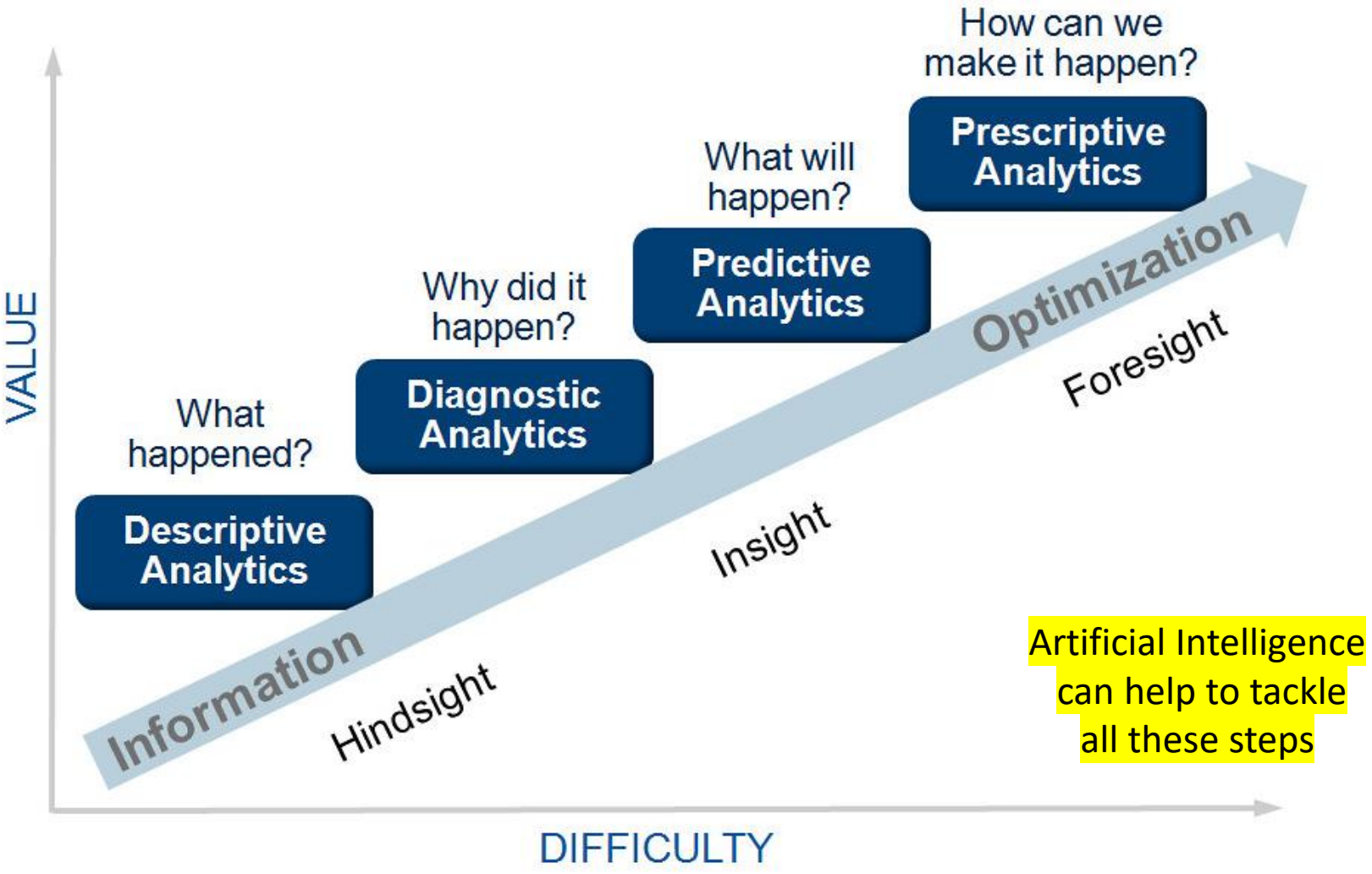
Main points

Open class discussion about data analysis

- What are your data analysis skills?
- How would you rate your data analysis skills?
- What are your previous experience?
- What are your favorite tools?
- Are you using data analysis tools for
  - Pre-processing?
  - Post-processing?
  - Results understanding?

Stop 1 minute the playback and take a note on a paper!

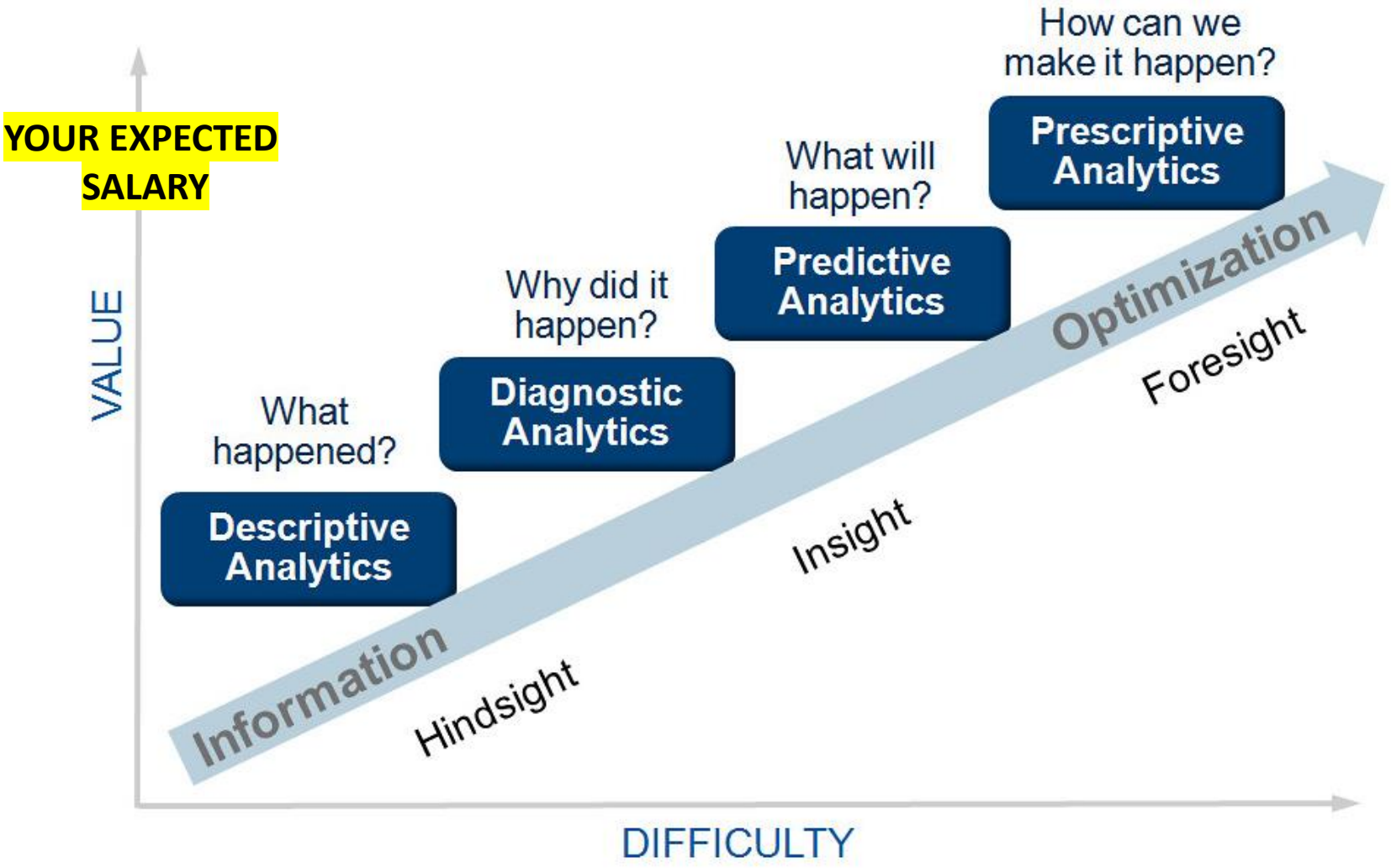
Value of Data Analytic



Artificial Intelligence can help to tackle all these steps

Gartner Group

Value of Data Analytic… For you



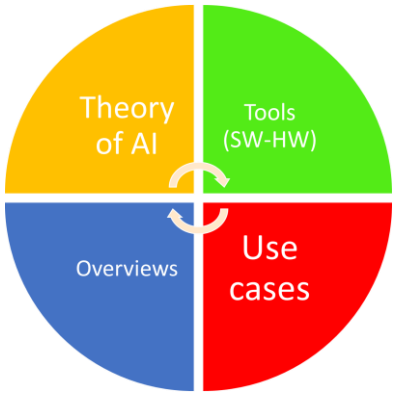
Gartner Group



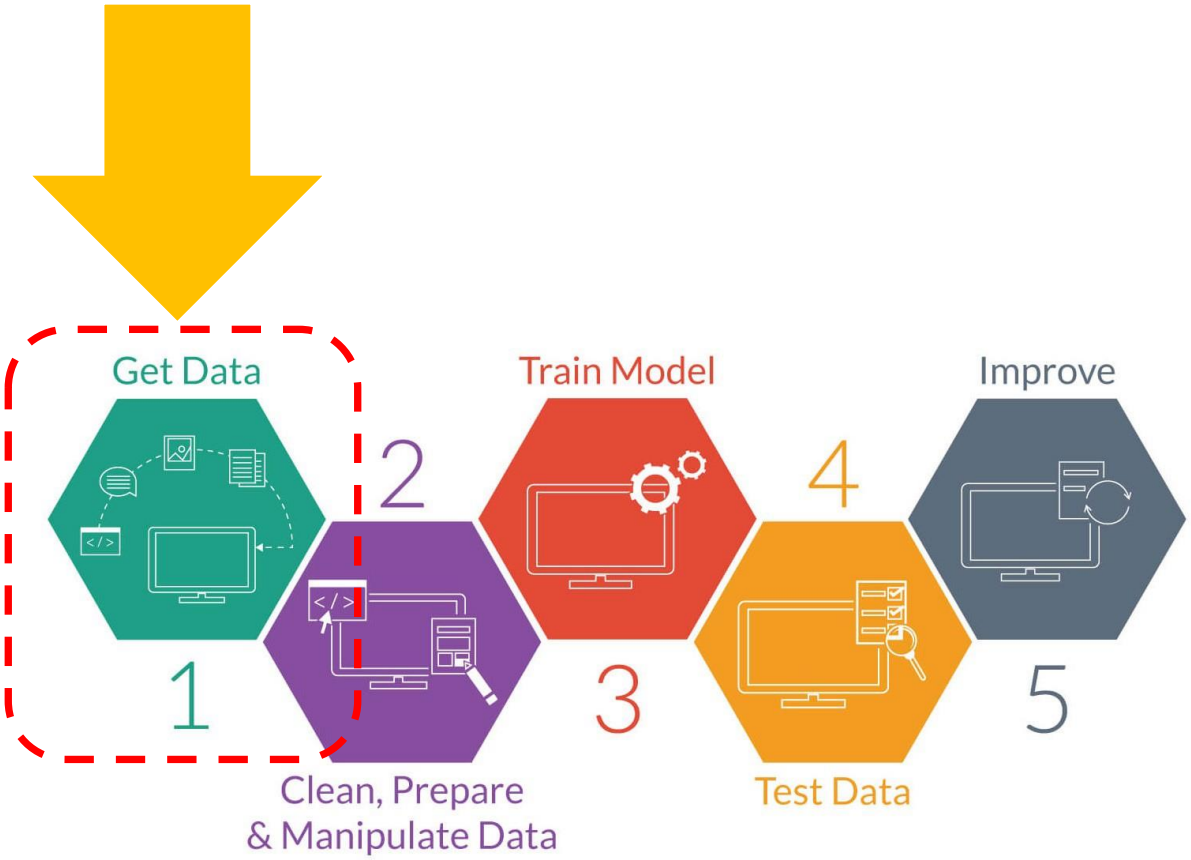

THEORY

Data gathering

Data collection and gathering: important points



Step 1 of the ML workflow

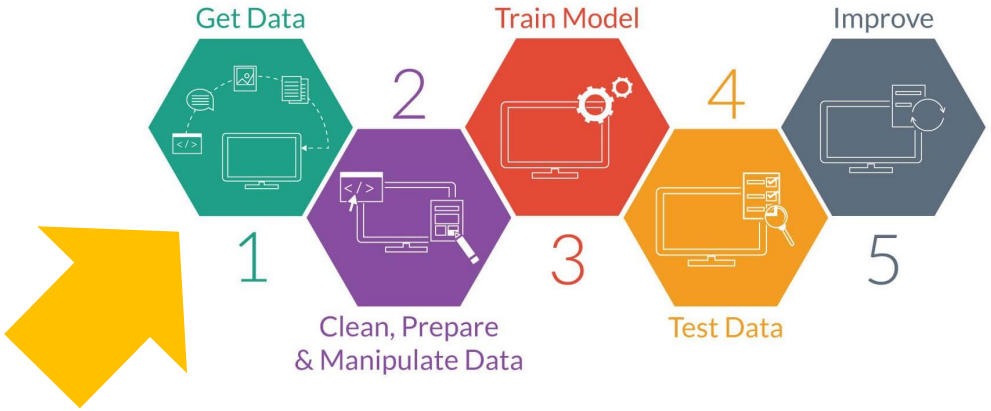


Step 1) in practice

- Classical DB query
- File processing
- Custom formats
- Stream of data from IoT and AIoT
- Online platform (AWS, Azure, Google, …)

Outside the focus of this course

We have so many powerful sources capable to generate data



Data collection

- Data collection is the process of gathering and measuring information on targeted variables in an established system
- Used in
  - Physical sciences
  - Industrial sector
  - Business
  - Humanities
    - Social sciences
    - In general, disciplines that study aspects of human society and culture
- Artificial intelligence is used in all this sectors / sciences
  - (even in highly unstructured data like just raw texts)

GIGO

NO «MAGICAL» NEURAL NETWORK WILL SAVE YOU FROM THIS!!!

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.
- Trending
  - Surveillance,
  - Smart home data,
  - Smart phones
    - Coordinates, health sensors, activities
  - Wearable
    - Smart watches, bands
  - Sensors,
  - …

Notes:

- Zettabytes = **ZB**
  → zetta indicates multiplication by the seventh power of 1000

Decimal		
Value	Metric	
1000	kB	kilobyte
1000^2	MB	megabyte
1000^3	GB	gigabyte
1000^4	TB	terabyte
1000^5	PB	petabyte
1000^6	EB	exabyte
1000^7	ZB	**zettabyte**
1000^8	YB	yottabyte

(*) International Data Corporation (IDC)

Iot started to generate data… (2)

- Video surveillance will drive a large share of the IoT data created
- 5G enable devices

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

GPS data into 48-64 bits minimum (GPS coordinates in ISO 6709 format)
https://developers.google.com/protocol-buffers/docs/encoding

Big Data….. Mining is needed

Public datasets (Annotated)

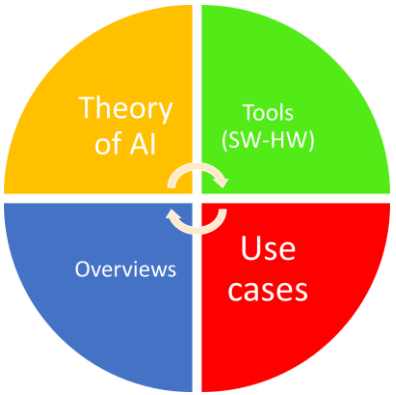
Primary source for training. Collections of Datasets!

- **UCI Machine Learning Repository**
- **Kaggle Datasets**
- **Amazon Datasets**
- **Google's Datasets Search Engine**
- **Microsoft Datasets**
- **Government Datasets**
  - EU Open Data Portal, US Gov Data, New Zealand's Government Dataset, Indian Government Dataset ,…
- **Computer Vision Datasets**
- **Lionbridge AI Datasets**
- **… the list is growing on a daily basis**

THEORY

Data heterogeneity and synchronization

Units, formats and exact timing



Data collection: data format heterogeneity

- Formats
- Name spaces
- Units
- Ranges
- Sizes
- …

Note: 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.

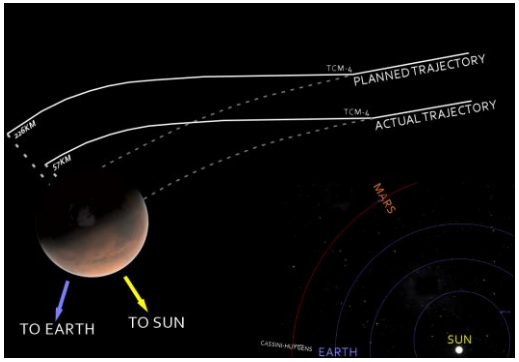
Data format heterogeneity (2)

Remember the Mars Climate Orbiter incident from 1999?

USE CASE

When NASA Lost a Spacecraft Due to a Metric Math Mistake
In September of 1999, after almost 10 months of travel to Mars, the Mars Climate Orbiter burned and broke into pieces. On a day when NASA engineers were expecting to celebrate, the ground reality turned out to be completely different, all because someone failed to use the right units, i.e., the metric units!
Mars Climate Orbiter was destroyed in the atmosphere…
…pound-force*seconds instead of the SI units of newton*seconds….

… The Mars Climate Orbiter, built at a total cost of $325 millions…
WRITTEN BY Ajay Harish

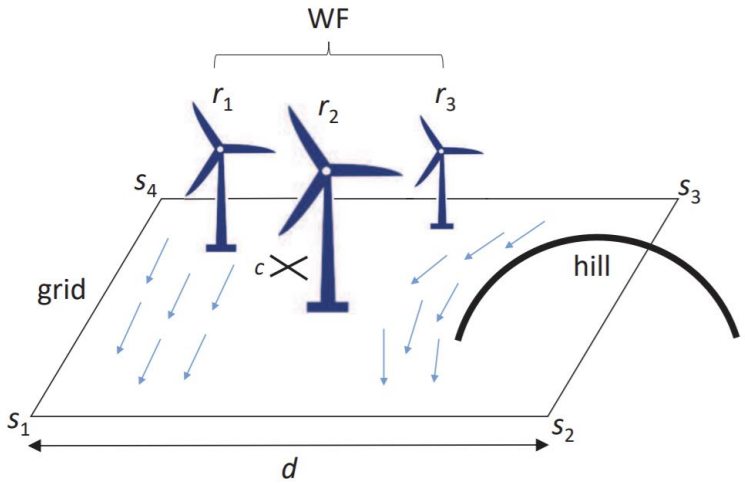


Use case: wind power prediction

Morning/afternoon breeze

Hills/occlusions

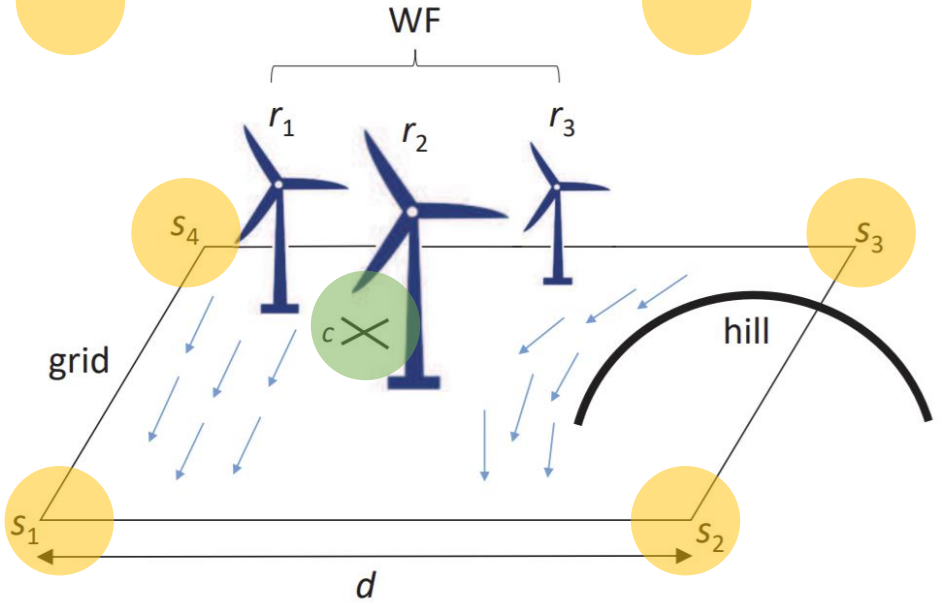
How to map the specificity of the specific site?



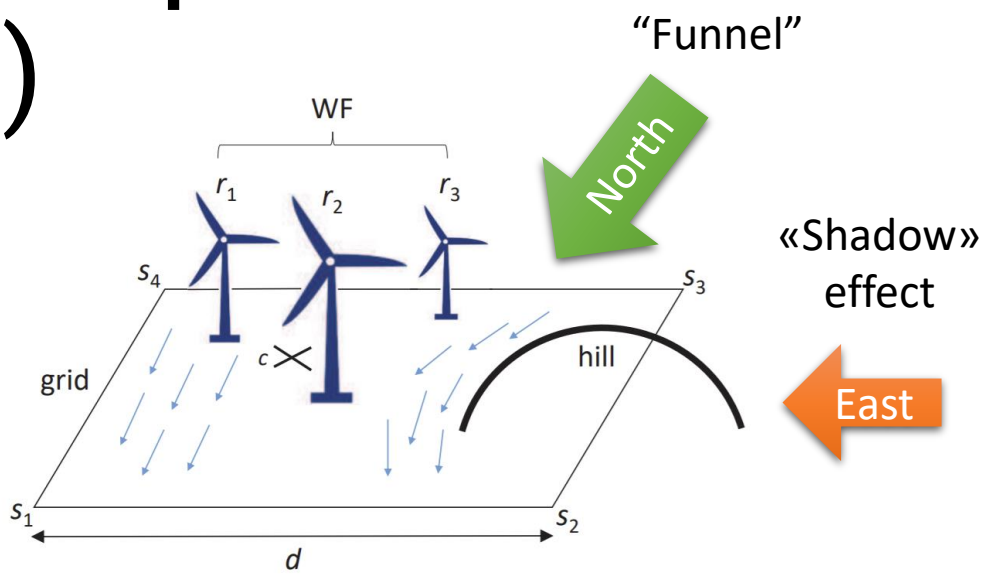
Use case: wind power prediction (2)

Weather data grid has a different altitude with respect to the plan

- The weather forecast data provided by different vendors and data collected by wind farm sensors could be expressed using different measurement units and reference systems
- Forecast data is different from site instrument station data (different locations)



Use case: wind power prediction (3)

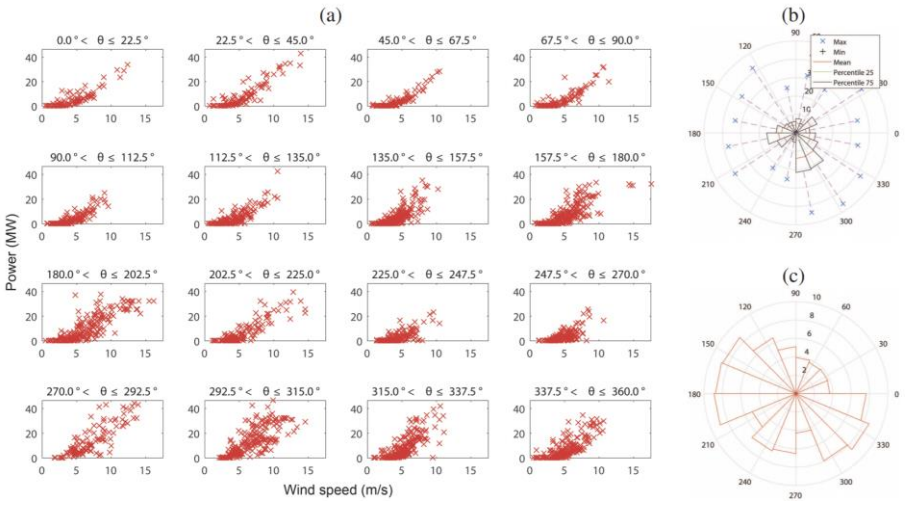


Many non idealities are present.

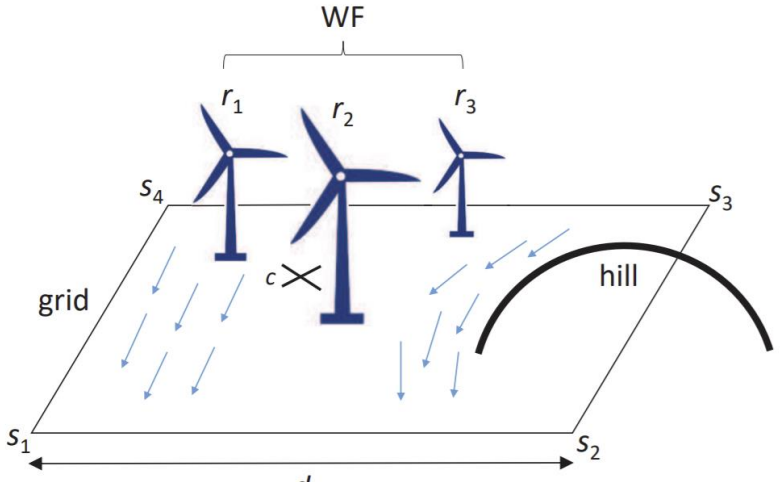
- Hills and occlusions can change of efficiency of the plant,
  Ex.: from East → very low power

- Wind turbines has minimum speed limit to start spinning (clipping)

- Same turbines have different speed/power curves due to the surroundings condition



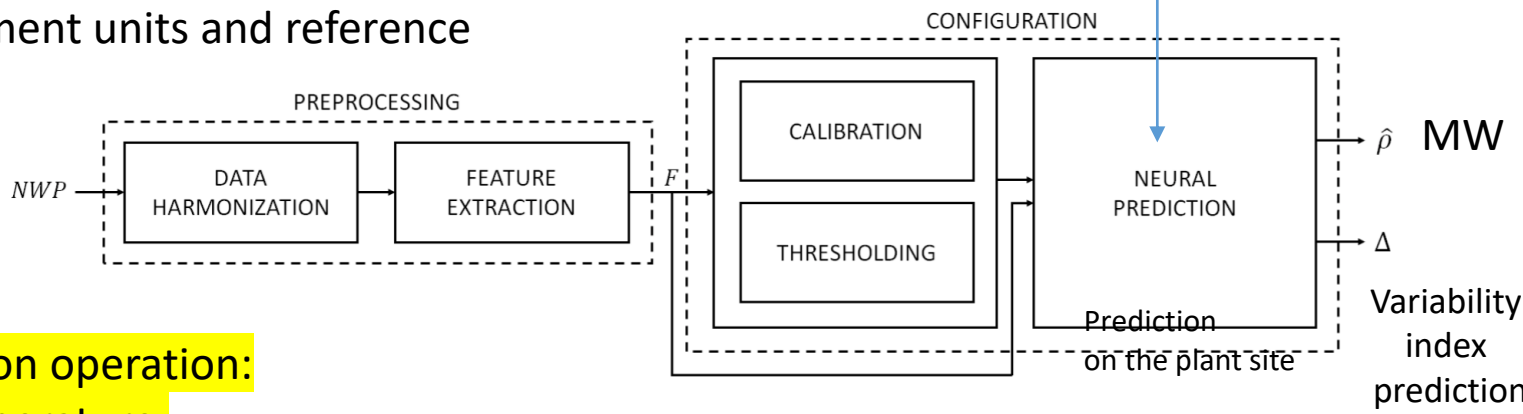
Use case: wind power prediction (3)



Goal: creation of 5y dataset from different sources of data

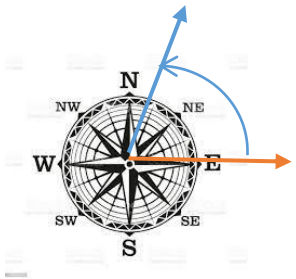
Problem: different weather forecast data format, measurement units and reference systems

Weather data on the grid and different altitude

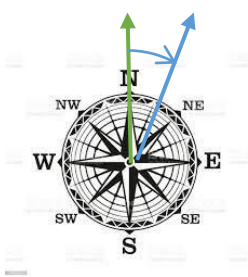


Data harmonization operation:

- °C for the temperature,
- m/s for the wind speed
- 0-360° angle northward directions (northward = 0°)
- hPa for atmospheric pressure,
- Scale from 0 to 1 for cloud coverage

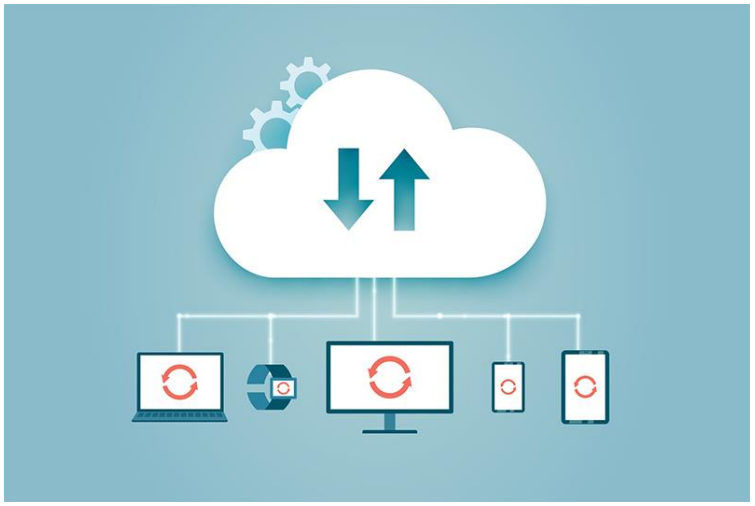


Wind angle def.1



Wind angle def.2

Data synchronization



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

In industrial and scientific applications this part is essential

Network Time Synchronization

Network Time Synchronization Devices

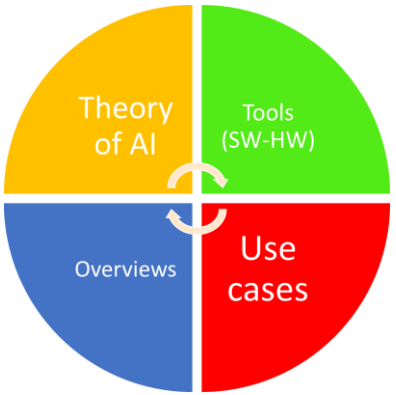
- The assumption that that computer clocks in servers, workstations and network devices are inherently accurate, is incorrect
  - Clocks are set by hand to within a minute or two of actual time and are rarely checked after that
  - Clocks are maintained by a battery-backed device that may drift as much as a second per day

  → It's impossible to have accurate time synchronization without a proper method
- **Solution 1**: Network Time Protocol (NTP) getting time from a public Internet Time Server (open in the firewall, UDP port 123,…)
- **Solution 2**: Dedicated network Time Server behind your firewall (devices synchronized to within 1/2 to 2 ms).

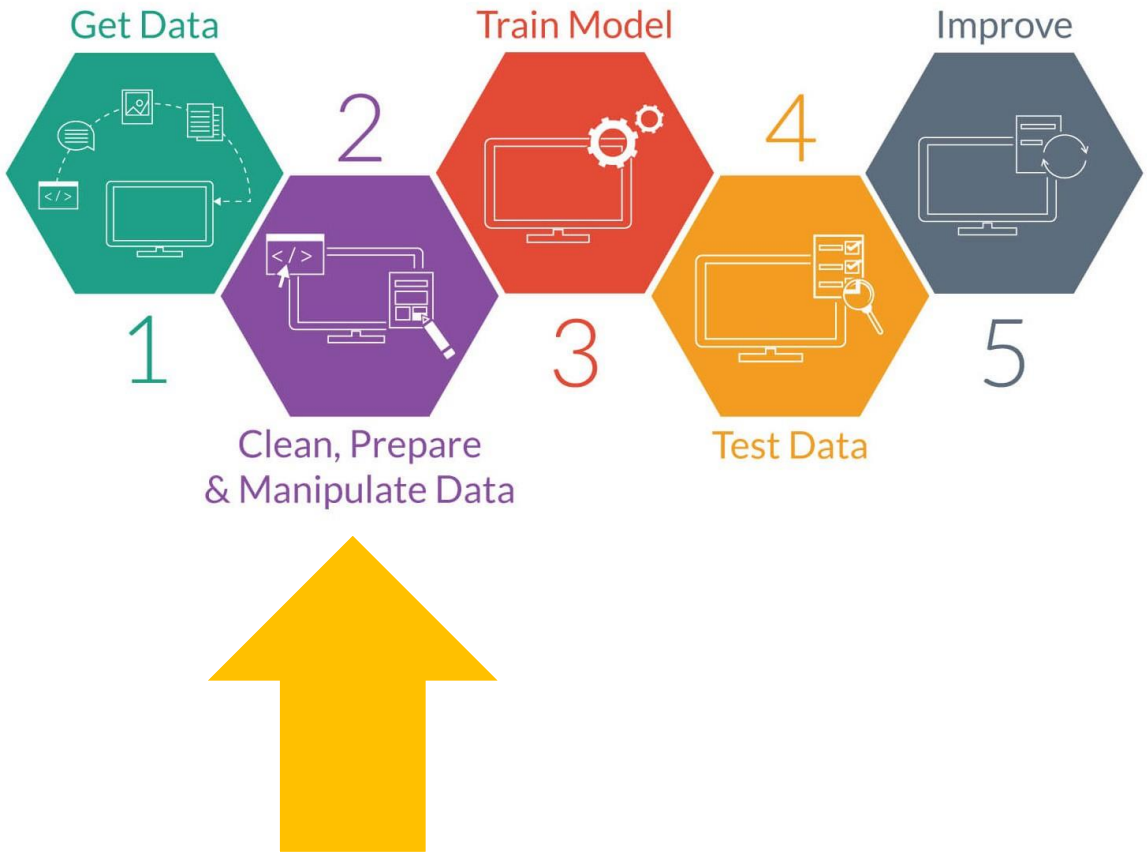
THEORY

Data Preparation

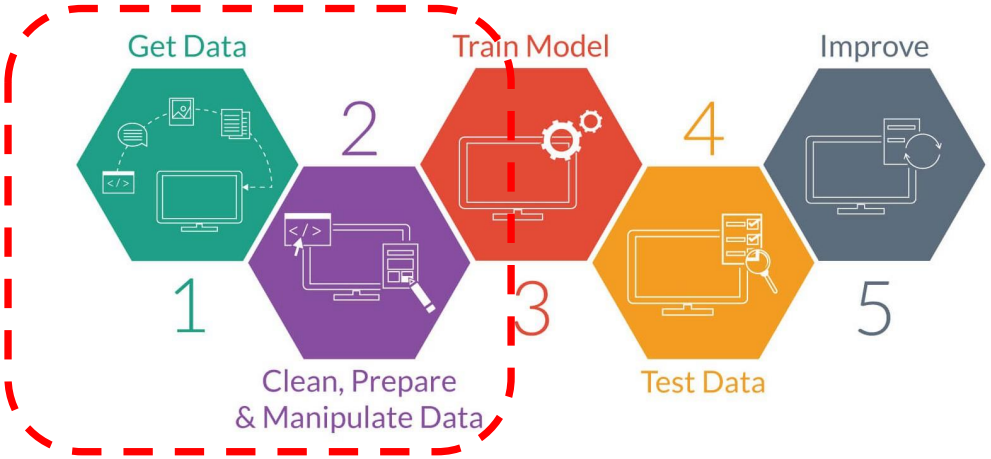
From noisy, messy, unstructured data
to the perfect dataset for your learning phase!



Step 2 of the ML workflow



«Entropy» → Structured data



5.1								
3.5								
1.4								
0.2								

Y = FUNC(X) that's it!!!

- For every dataset in machine learning or toolbox, is all about to create X and Y

- X

Sample, sample, ……

Features

5.1								
3.5								
1.4								
0.2								

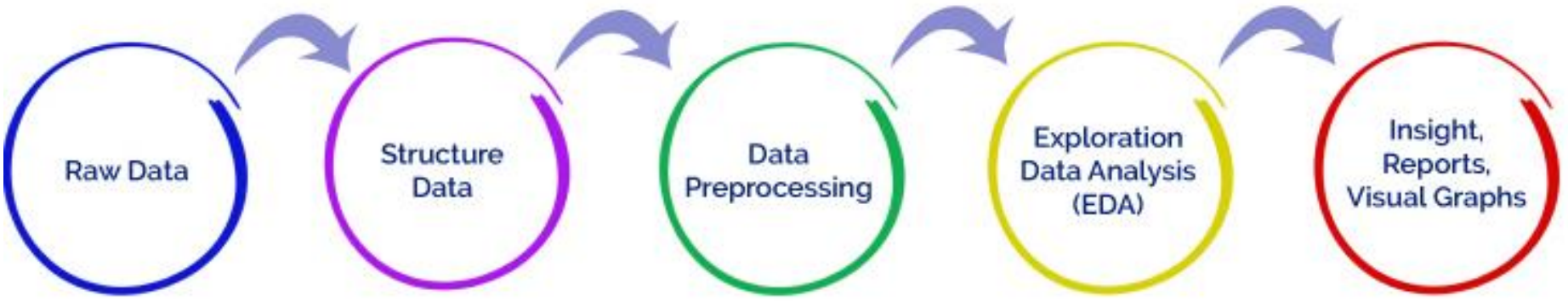
- Y

Class ID, or value to be learnt

1	**0**	**3**	**1**					

Data preparation

- A very important part of Data Science.
- It includes two concepts such as *Data Cleaning* and *Feature Engineering*.
- Two compulsory steps for achieving better accuracy and performance in the Machine Learning and Deep Learning projects.

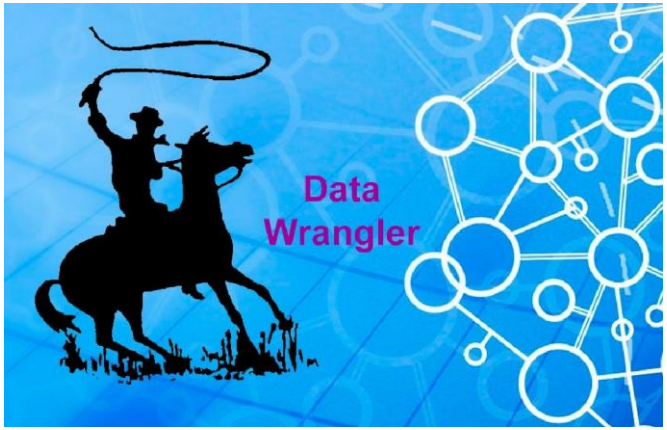


Data Preparation

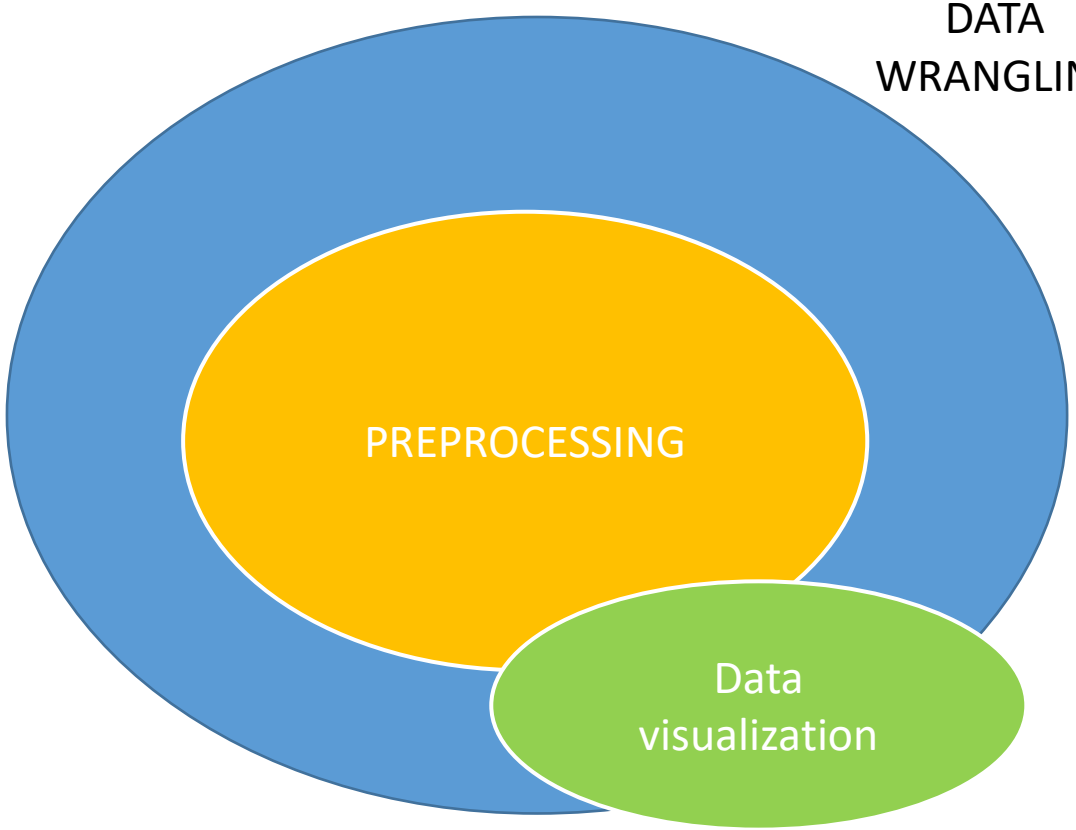
DATA PREPROCESSING

DATA WRANGLING

Data Preparation

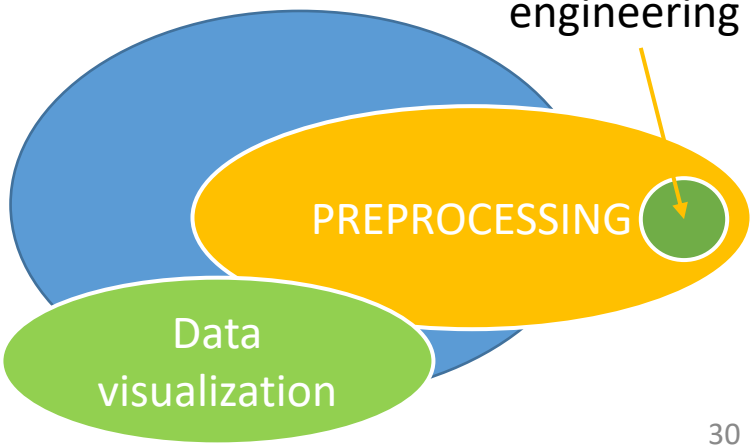


DATA
WRANGLING



DATA WRANGLING
re-def.

Feature
engineering

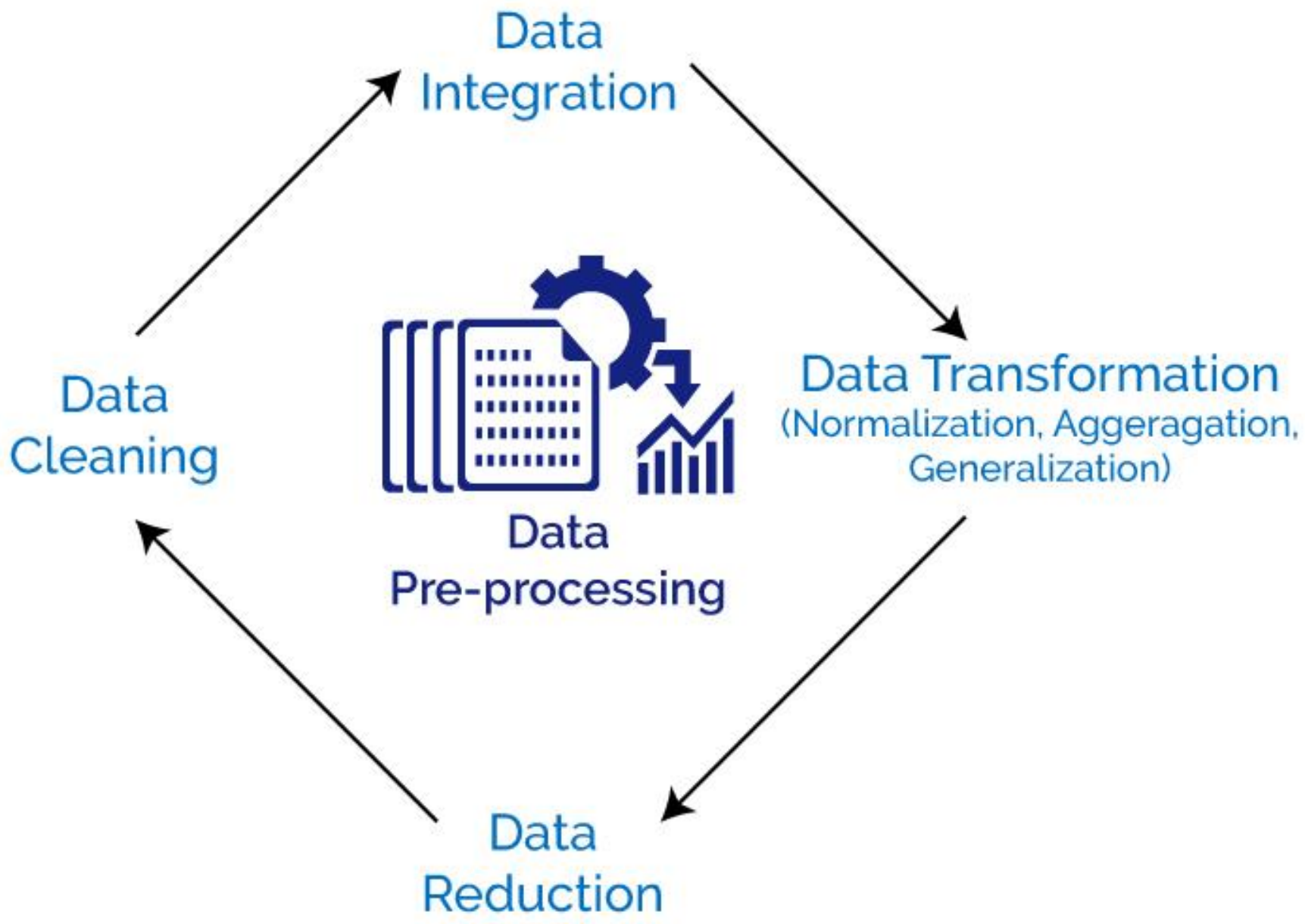


First things first!

"It's an absolute myth that you can send an algorithm over raw data and have insights pop up" (Jeffrey Heer)

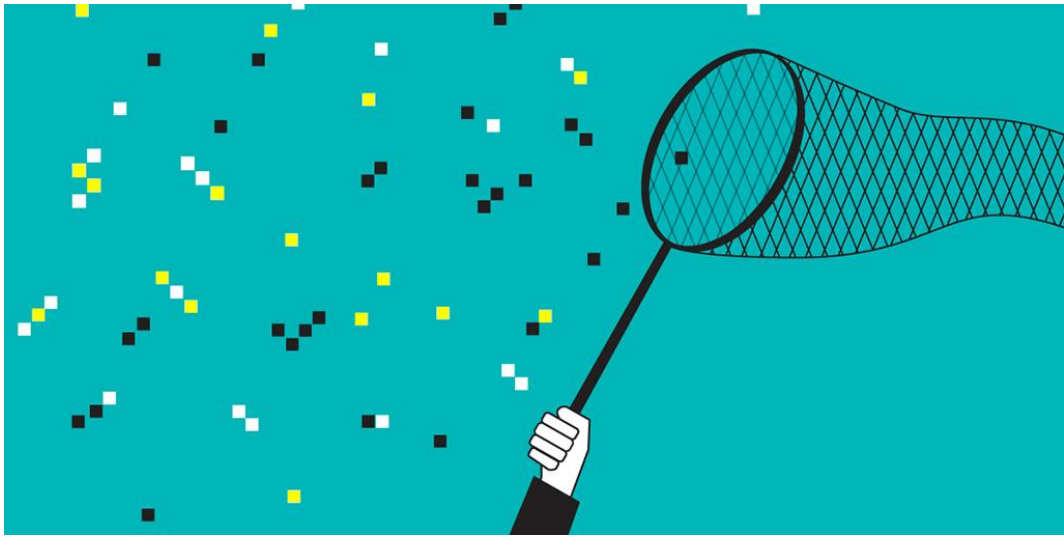
- The data wrangling problem is growing as different types of unstructured data or data in varying formats are pouring in from sensors, online and from traditional databases.
- All these data must be cleaned up and organized before data analytics/classifiers/regressors models can be applied.

Jeffrey Heer: expert of information visualization and interactive data analysis at the University of Washington


Data
Integration
Data Transformation
(Normalization, Aggeragation,
Generalization)
Data
Reduction
Data
Cleaning
Data
Pre-processing

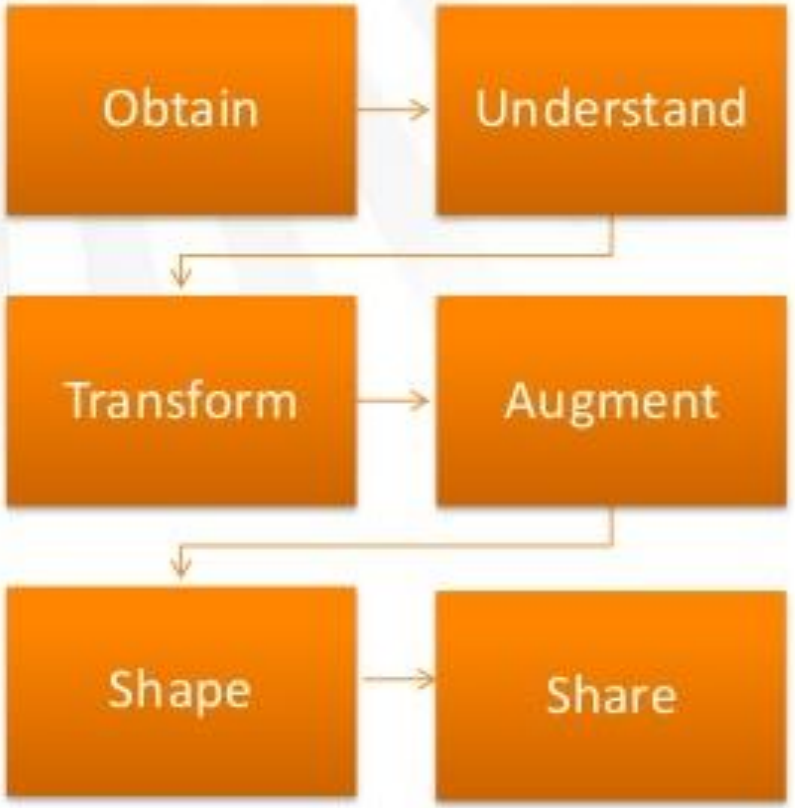



Data wrangling



Data wrangling steps

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



Data wrangling activities

- Data Preprocessing
- Data Preparation
- Data Cleansing
- Data Scrubbing
- Data Munging
- Data Transformation
- Data Fold, Spindle, Mutilate…

Tasks of Data Wrangling

Discovering

- Firstly, data should be understood thoroughly and examine which approach will best suit. For example: if have a weather data when we analyze the data it is observed that data is from one area and so primary focus is on determining patterns.

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 Preprocessing Tools

- **Matlab** (see next slides/lessons)
- **R**

R a framework that consists of various packages that can be used for Data Preprocessing like dplyr etc.

- **Weka**

Weka is a software that contains a collection of Machine Learning algorithms for Data Mining process. It consists of Data Preprocessing tools that are used before applying Machine Learning algorithms.

- **RapidMiner**

RapidMiner is an open-source **Predictive Analytics Platform** for Data Mining process. It provides the efficient tools for performing exact Data Preprocessing process.

- **Python (improving!!!!)**

Python is a programming language that provides various libraries that are used for Data Preprocessing.

Data pre-processing

- is a technique that is used to convert the raw data into a clean data set.
  - Whenever the data is gathered from different sources or it is collected in raw format is not (always) feasible for the analysis.
- Pre-processing includes
  - Data cleaning
  - Data integration
  - Data transformation
  - Data reduction

Data preprocessing and visualization for Data wrangling

- Data preprocessing and visualization are one of the first steps to understand the data you have to use
- To find…
  - Outliers
  - Errors
  - Missing values
  - Different scales
  - Distributions
  - Clusters in the dataset
  - Salient features

Data preparation

Feature engineering

This is an important list to be remembered

Why is Data Preprocessing is so important?

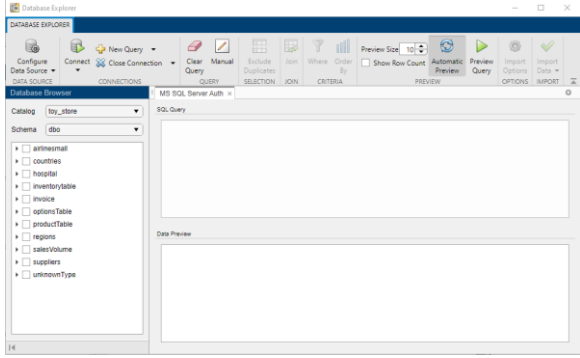
- Data Preprocessing is necessary because of the presence of unformatted real-world data.
  - **Inaccurate data (missing data)**

    Many reasons for missing data: data is not continuously collected, a mistake in data entry, technical problems with sensors/transmission/DoS/delayed ….
  - **The presence of noisy data (erroneous data and outliers)**

    Noisy data … could be a technological problem of device that gathers data, a human mistake during data entry, etc.
  - **Inconsistent data**

    Existence of duplication within data, human data entry, containing mistakes in codes or names, i.e., violation of data constraints and much more.

Tools: Matlab Database Explorer

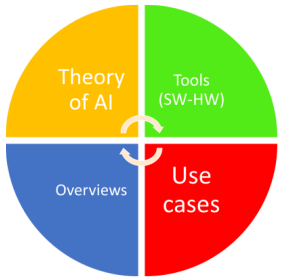


Quickly connect to a database, explore the database data, and import data into the MATLAB workspace in a visual way

Some interesting features:

- Create and configure ODBC and JDBC data sources.
- Establish multiple connections to the same or different databases.
- Select tables and columns of interest.
- Fine-tune selections using SQL query criteria.
- Preview selected data.
- Customize import options.
- Import selected data into the MATLAB workspace for analysis.
- Save generated SQL queries.
- Generate MATLAB code.



THEORY

Missing Data

Dealing with data complexity

Missing data rapresentation

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

NaN

Member of a numeric data type that can be interpreted as a value that is undefined, unrepresentable, or missing (encoded with the IEEE 754 standard)

Missing data example

3 Hospitals are merging data of their patients over blood test screening exams
But the labs have different systems and instruments with a different number of features

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		Hospital 1				Hospital 2				Hospital 3				
2		PZ1	PZ2	PZ3	PZ4	PZ5	PZ6	PZ7	PZ8	PZ9	PZ10	PZ11	PZ12	PZ13
3	Feat. 1	0,744422	0,205193	0,766586	0,487325	0,775158	0,634885	0,765383	0,269233	0,301776	0,366692	0,83481	0,048661	0,068629
4	Feat. 2	0,341608	0,598078	0,431988	0,386859	0,311953	0,595674	0,914989	0,437615	0,264475	0,660899	0,230343	0,342688	0,181113
5	Feat. 3	0,406022	0,950997	0,045678	0,994392	0,848554	0,221909	0,034978	0,761679	0,763025	0,912281	0,165851	0,008296	0,016205
6	Feat. 4	0,456926	0,94847	0,094555	0,310251	0,503689	0,641912	0,380669	0,828563	0,446763	0,084337	0,057833	0,355379	0,738135
7	Feat. 5	0,380072	0,351067	0,658412	0,837005	0,019425	0,686687	0,191358	0,832462	0,833955	0,010785	0,608542	0,537297	0,845168
8	Feat. 6	0,662976	0,856301	0,696342	0,688191	0,552233	0,150594	0,517138	0,903131	0,717526	0,158101	0,57893	0,178603	0,291695
9	Feat. 7	0,25033	0,335838	0,395532	0,544964	0,198371	0,629432	0,843826	0,702729	0,962303	0,405384	0,965353	0,093214	0,439128
10	Feat. 8	0,456787	0,54951	0,217909	0,844122	0,527656	0,92943	0,18003	0,644716	0,408152	0,100464	0,372383	0,718214	0,901732
11	Feat. 9					0,929722	0,369789	0,170243	0,226735	0,575054	0,11046	0,513469	0,286501	0,696153
12	Feat. 10					0,116038	0,808077	0,579666	0,096338	0,588189	0,553363	0,745312	0,123346	0,028077
13	Feat. 11					0,793393	0,630495	0,406624	0,525547					
14	Feat. 12					0,253424	0,199476	0,47691	0,898241					

Data Harmonization is crucial here since the sources of data can be different!
Es: mg/dL → mmol/L → etc. etc.

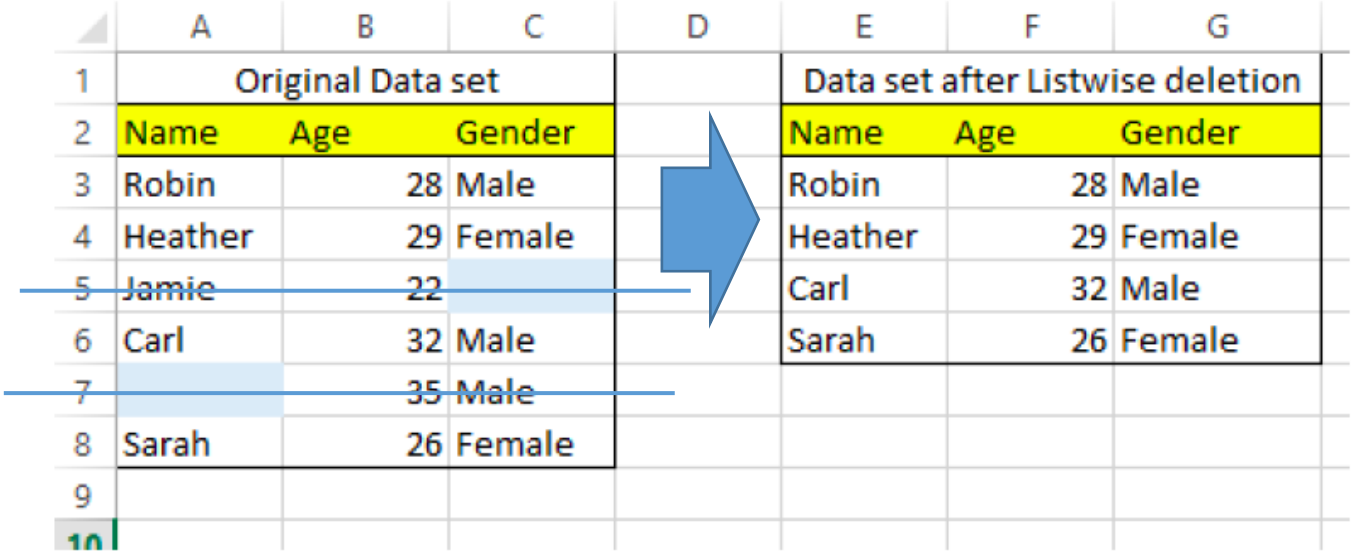
Remove incomplete features

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		Hospital 1				Hospital 2				Hospital 3				
2		PZ1	PZ2	PZ3	PZ4	PZ5	PZ6	PZ7	PZ8	PZ9	PZ10	PZ11	PZ12	PZ13
3	Feat. 1	0,744422	0,205193	0,766586	0,487325	0,775158	0,634885	0,765383	0,269233	0,301776	0,366692	0,83481	0,048661	0,068629
4	Feat. 2	0,341608	0,598078	0,431988	0,386859	0,311953	0,595674	0,914989	0,437615	0,264475	0,660899	0,230343	0,342688	0,181113
5	Feat. 3	0,406022	0,950997	0,045678	0,994392	0,848554	0,221909	0,034978	0,761679	0,763025	0,912281	0,165851	0,008296	0,016205
6	Feat. 4	0,456926	0,94847	0,094555	0,310251	0,503689	0,641912	0,380669	0,828563	0,446763	0,084337	0,057833	0,355379	0,738135
7	Feat. 5	0,380072	0,351067	0,658412	0,837005	0,019425	0,686687	0,191358	0,832462	0,833955	0,010785	0,608542	0,537297	0,845168
8	Feat. 6	0,662976	0,856301	0,696342	0,688191	0,552233	0,150594	0,517138	0,903131	0,717526	0,158101	0,57893	0,178603	0,291695
9	Feat. 7	0,25033	0,335838	0,395532	0,544964	0,198371	0,629432	0,843826	0,702729	0,962303	0,405384	0,965353	0,093214	0,439128
10	Feat. 8	0,456787	0,54951	0,217909	0,844122	0,527656	0,92943	0,18003	0,644716	0,408152	0,100464	0,372383	0,718214	0,901732
11	Feat. 9					0,929722	0,369789	0,170243	0,226735	0,575054	0,11046	0,513469	0,286501	0,696153
12	Feat. 10					0,116038	0,808077	0,579666	0,096338	0,588189	0,553363	0,745312	0,123346	0,028077
13	Feat. 11					0,793393	0,63[illegible]	0,40[illegible]	0,[illegible]					
14	Feat. 12					0,253424	0,19947[illegible]	0,47691	0,898241					

DELETE?

What if
Feature #12 is the feature you are looking for to detect the studied cancer?

Removing incomplete vectors



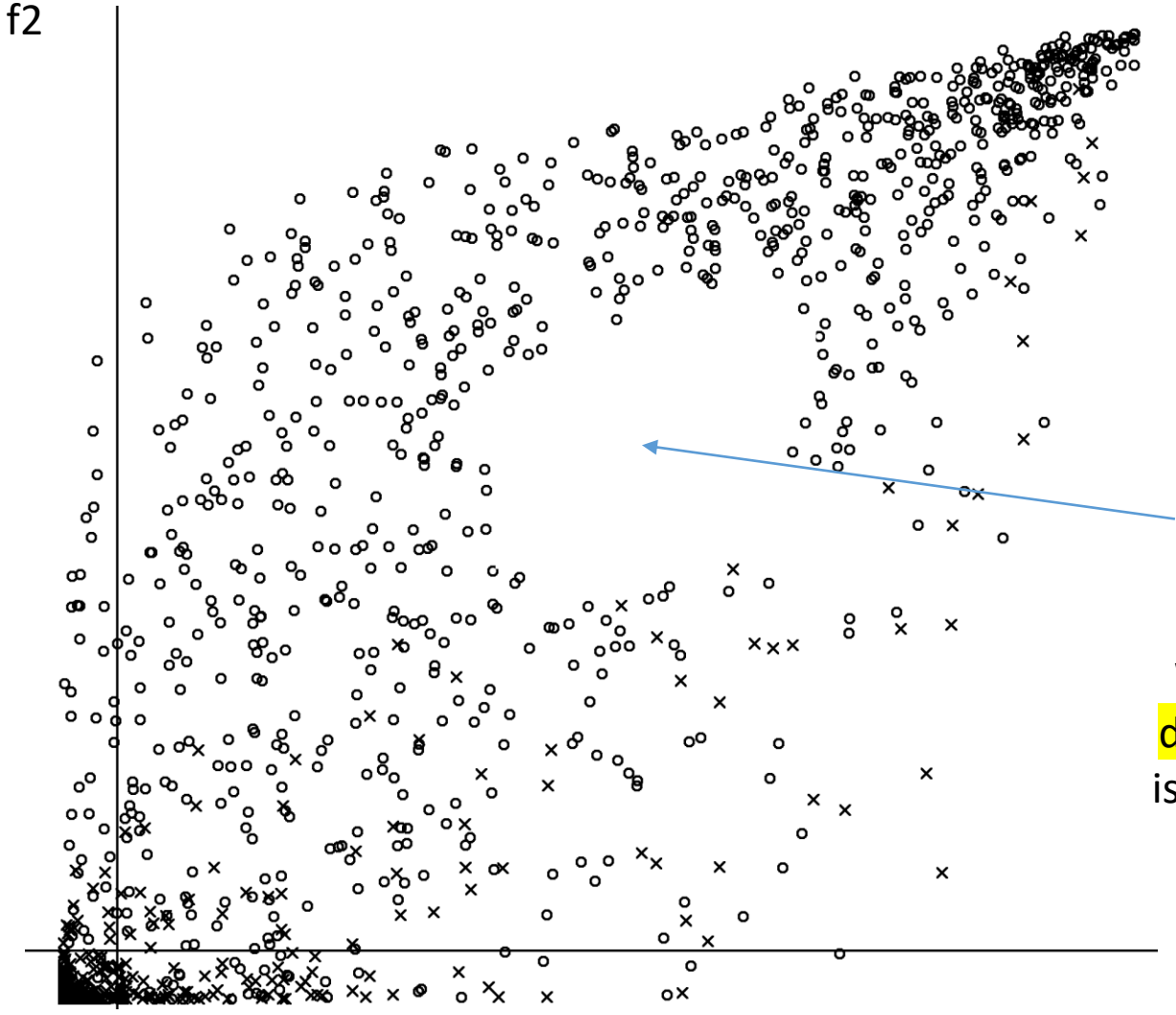
	A	B	C	D	E	F	G
1	Original Data set				Data set after Listwise deletion		
2	Name	Age	Gender		Name	Age	Gender
3	Robin	28	Male		Robin	28	Male
4	Heather	29	Female		Heather	29	Female
5	~~Jamie~~	~~22~~			Carl	32	Male
6	Carl	32	Male		Sarah	26	Female
7		~~35~~	~~Male~~				
8	Sarah	26	Female				
9							
10							

Bye-bye Jamie and Mr “35 Male” → **from 6 records to only (complete) 4**

Is this loss of information acceptable for your application?

A «topological» miss

The matrix of data is complete, but the feature space is not uniformly covered



x = Reject
o = Good

????

Without proper data visualization is hard to find this problem….

Data Preprocessing: missing data (1)

- **Ignoring the missing record –** It is the simplest and efficient method for handling the missing data.
- How your learning algorithm is managing them?!
- This method should not be performed at the time
  - when the number of missing values are immense
  - or when the missing data problem and can solved (debugging/re-design-redoing the experiment) and not just ignoring the problem causing the missing data.

	0
0	25
1	27
2	50
3	35
4	40
5	35
6	nan
7	48
8	50
9	37
10	21
11	nan
12	63

	0
0	35000
1	40000
2	54000
3	nan
4	60000
5	58000
6	52000
7	79000
8	83000
9	nan
10	24000
11	60000
12	70000

Data Preprocessing: missing data (2)

- **Filling the missing values manually**
  - This is 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 time-consuming task.

I'm sorry, I have some missing values from the sensors… Can we do that again?

For example, in medical trials
every single datum
is very important and expensive.

Excluding a vector
is not a good first strategy!

Data Preprocessing: missing data (3)

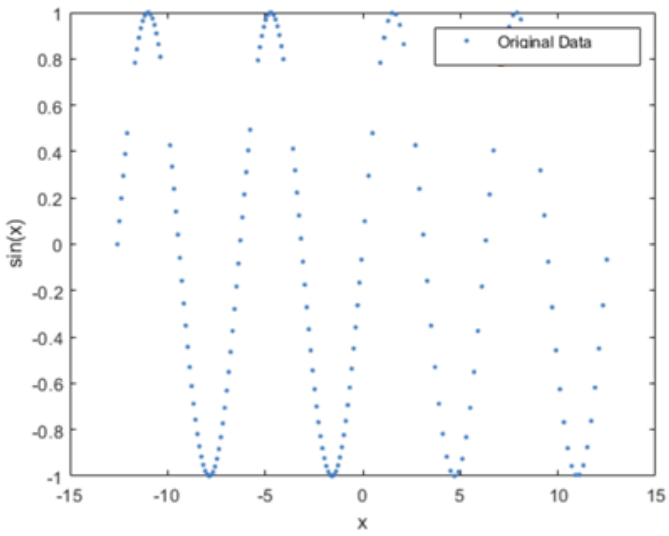
- **Filling using computed values**
- The missing values can also be occupied by computing mean, mode or median of the observed given values.
- You can also find the most similar column and use it to copy the missing values (ex. using kNN classifiers)
- Another method could be the predictive values that are computed by using any Machine Learning or Deep Learning algorithm.
  - Drawback: it can generate bias within the data.

A basic use case:

using a linear model

```
x = [-4*pi:0.1:0, 0.1:0.2:4*pi];
A = sin(x);
```

```
A(A < 0.75 & A > 0.5) = NaN;
```

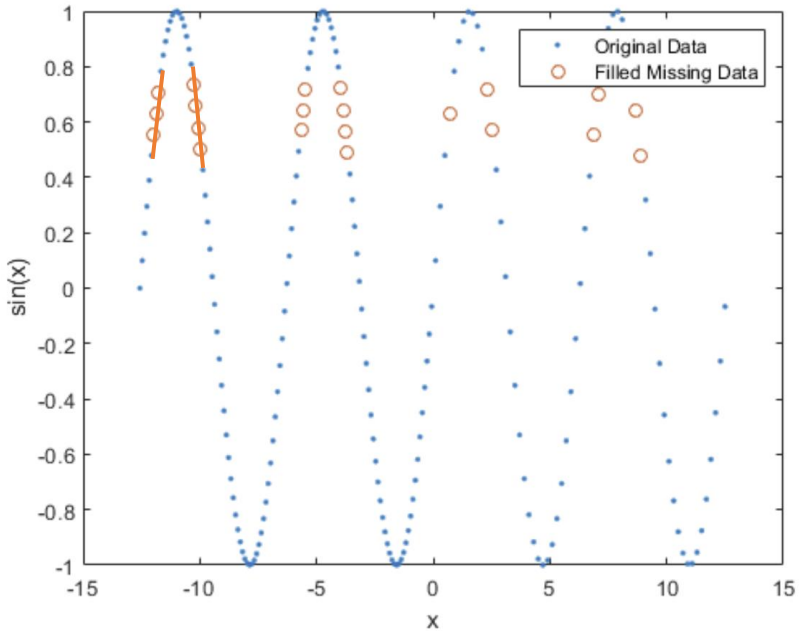


- F: Fill the missing data using linear interpolation
  TF: corresponds to the values of F that were filled

```
[F,TF] = fillmissing(A,'linear','SamplePoints',x);
```

Plot the original data and filled data.

```
plot(x,A,'.', x(TF),F(TF),'o')
xlabel('x');
ylabel('sin(x)')
legend('Original Data','Filled Missing Data')
```



The choice of the model to be used form interpolate the missing points will affect the learning of subsequent model

Missing data: Possibile solutions in brief

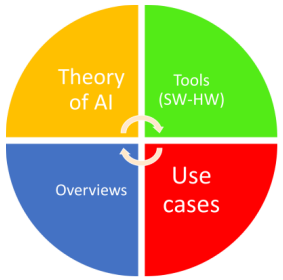
- STEP 0: Quantify the problem (Row? Cols?)
- Do Nothing
  - let the algorithm handle the missing data → ok, but what is it doing? Explainability and cross validation can be compromised. Some method will ignore, or crash, erratic, …
- Filling the missing values manually (if possible and ask the expert about data distribution and ranges)
- Statistic appr.: Mean, Mode, Median, Tabu-value (☹!) or constant
  - WARNING!! Do not make categorical features → real values
- K-Nearest Neighbor/s: find the most similar vector/vectors in input and copy the missing values

In brief…


1
Missing Data
→ Ignore
→ Fill Manually
→ Fill Computed Value
2
Noisy Data
→ Binning
→ Clustering
→ Machine Learning Algorithm
→ Remove Manually
3
Inconsistant Data
→ External Refrences
→ Knowledge Engineering Tools



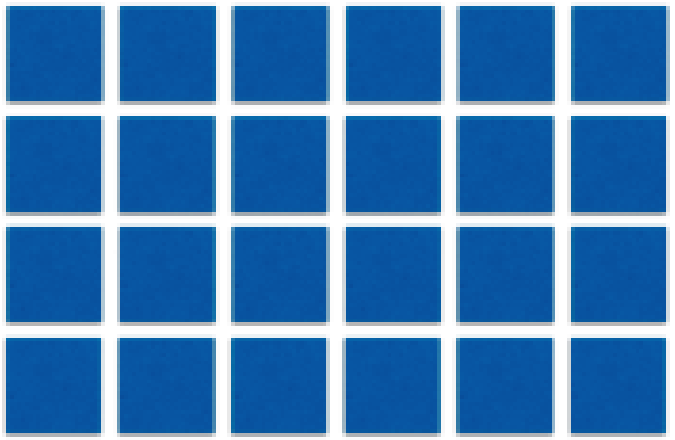

THEORY



Structured and unstructured Data

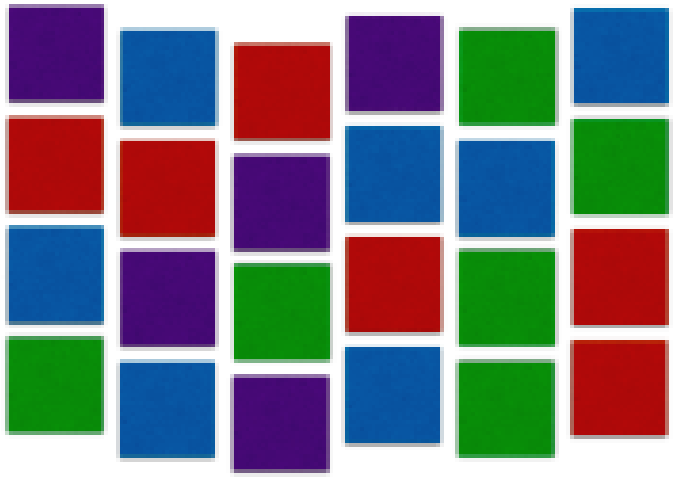
Dealing with data complexity

Structured Data



What you find in a DB
(typically)

Unstructured Data



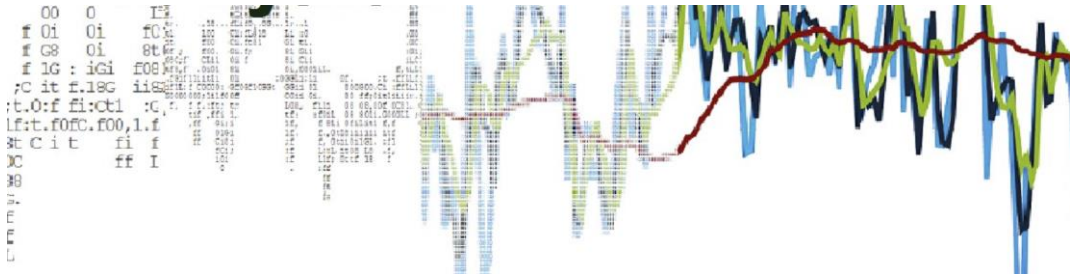
What you find in the 'wild'
(text, images, audio, video)

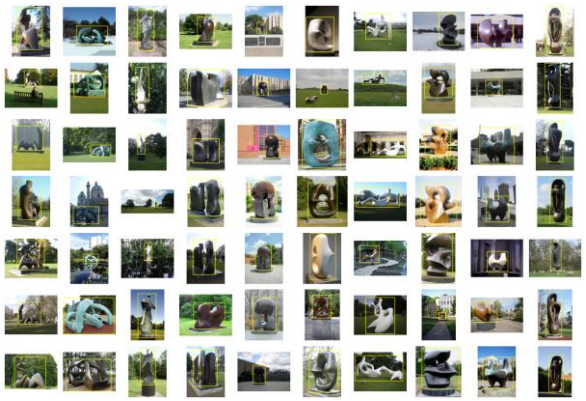
Structured data (Good!)

- Structured data usually resides in relational databases (RDBMS).
  - Fields store length-delineated data phone numbers, Social Security numbers, or ZIP codes.
  - Even text strings of variable length like names are contained in records, making it a simple matter to search.
- Data may be human- or machine-generated as long as the data is created within an RDBMS structure.
- 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.

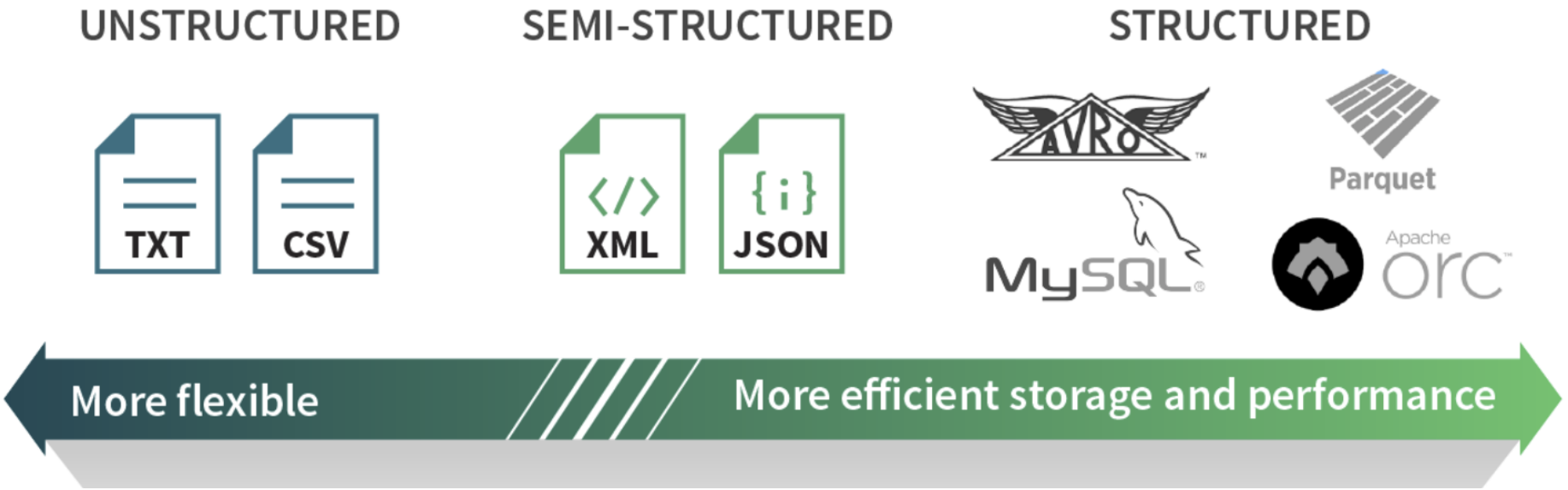
Unstructured Data (Bad…)

- Unstructured data is essentially everything else. Unstructured data has internal structure but is not structured via pre-defined data models or schema.
  - It may be textual or non-textual, and human- or machine-generated. It may also be stored within a non-relational database like NoSQL





Semi-structured



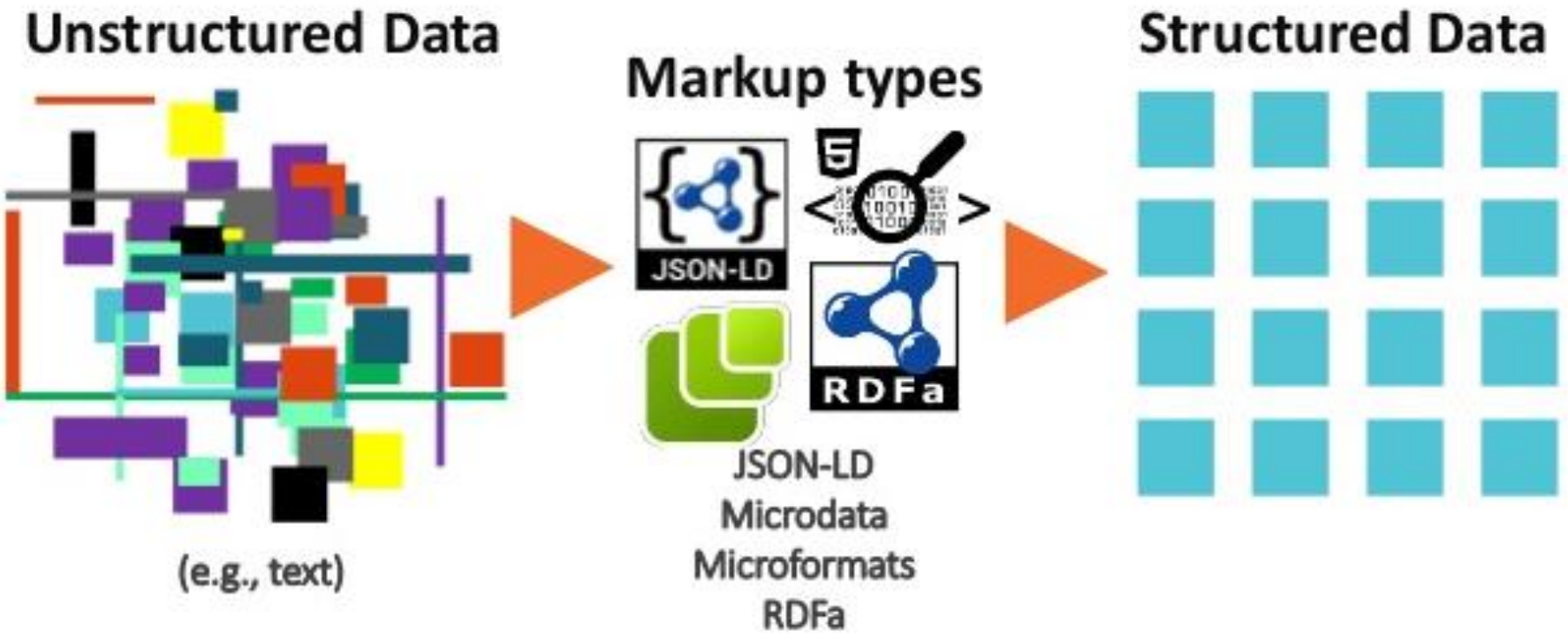
Typical human-generated unstructured data includes:

- **Text files:** Word processing, spreadsheets, presentations, email, logs.
- **Email:** Email has some internal structure thanks to its metadata, and we sometimes refer to it as semi-structured. However, its message field is unstructured and traditional analytics tools cannot parse it.
- **Social Media:** Data from Facebook, Twitter, LinkedIn.
- **Website:** YouTube, Instagram, photo sharing sites.
- **Mobile data:** Text messages, locations.
- **Communications:** Chat, IM, phone recordings, collaboration software.
- **Media:** MP3, digital photos, audio and video files.
- **Business applications:** MS Office documents, productivity applications.

Typical machine-generated unstructured data includes:

- **Satellite imagery:** Weather data, land forms, military movements.
- **Scientific data:** Oil and gas exploration, space exploration, seismic imagery, atmospheric data.
- **Digital surveillance:** Surveillance photos and video.
- **Sensor data:** Traffic, weather, oceanographic sensors.

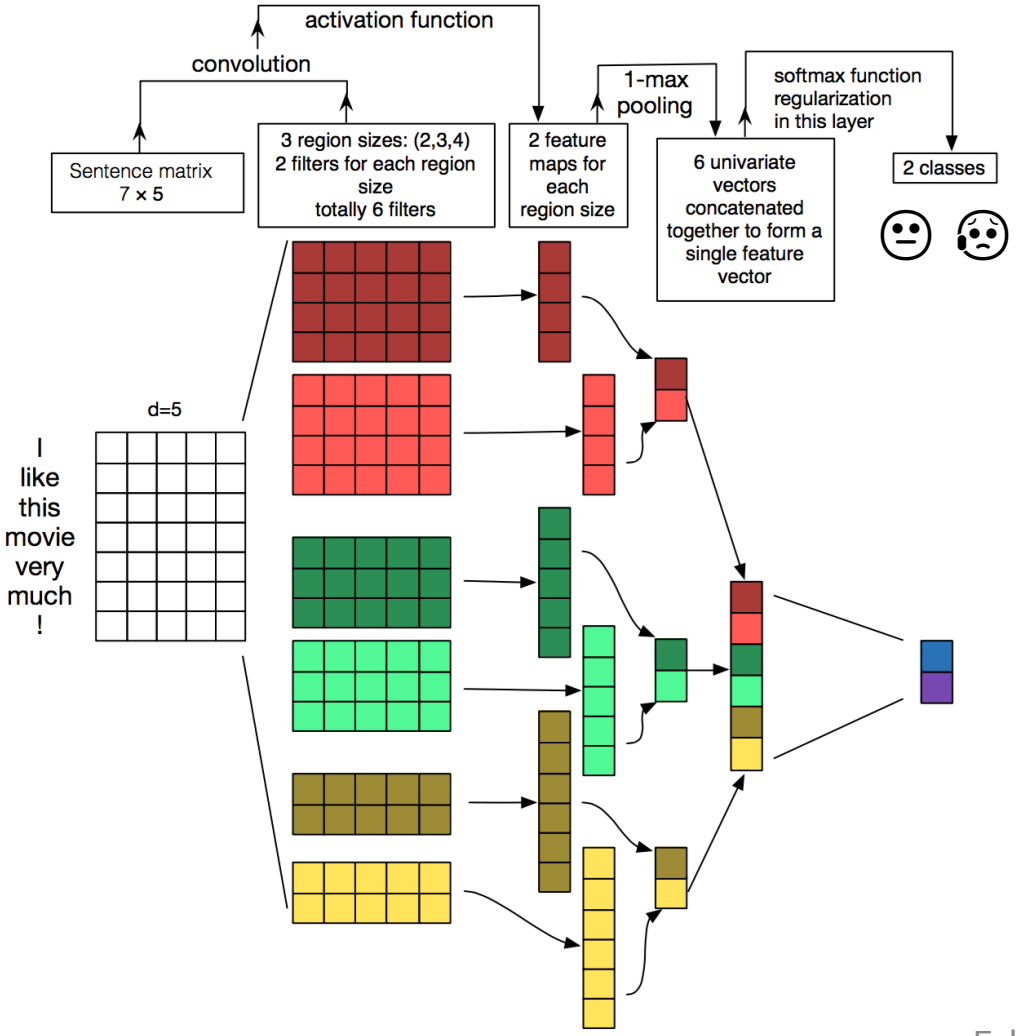
A web example of data tranformation



Resource Description Framework in Attributes (RDFa) is a W3C Recommendation that adds a set of attribute-level extensions to HTML, XHTML and various XML-based document types for embedding rich metadata within Web documents.

Neural networks and unstructured data

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



Some examples of text classification are:

- Understanding audience sentiment (😄 😐 😢) from social media
- Detection of spam & non-spam emails
- Auto tagging of customer queries
- Categorization of news articles into predefined topics

Main points

- Using datasets without previous activities is useless dangerous and it wastes your time
- Study, cleaning, harmonize…
  and comprehend basic relationship is mandatory
- Dealing with missing data
- Structured and unstructured data