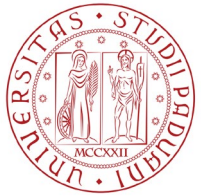


UNIVERSITÀ DEGLI STUDI DI PADOVA

DEPARTMENT OF INDUSTRIAL ENGINEERING

Machine Learning Lesson #1

Academic year 2025-2026

Prof. Pierantonio Facco

CAPE-Lab, Computer-Aided Process Engineering Laboratory

Email: pierantonio.facco@unipd.it

URL: https://research.dii.unipd.it/capelab/

Recap of the previous lesson

- We are living the **4th and 5th industrial revolution**
  - **large amount and variety** of **data** are available at high velocity in all the industrial sectors
  - in the phase of the development of new products and processes **few data** are usually available
- Data (even if limited in number) retain valuable **information** on the process/systems which generates them
- Turning data into information is the key to enhance the **knowledge** of the process/system we are studying
  - this requires appropriate **mathematical modelling** and **computing power**

- **Design of experiments** aids to create informative data to the purpose of discovery, improvement and optimization
- **Machine learning** helps to extract information, find correlations, identify patterns, classify, estimate, etc…

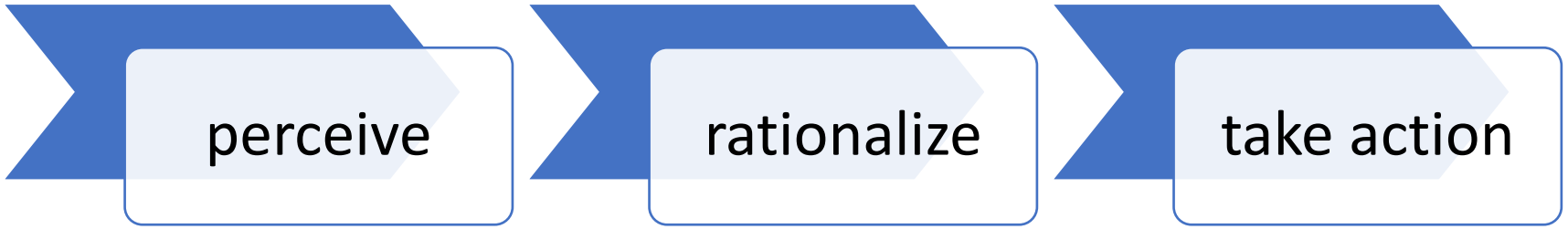
Today’s lesson

- Today’s lesson deals with:
  - why AI and machine learning in Industry 4.0 and 5.0?
    - definition and positioning
  - definition of machine learning
    - supervised learning
    - unsupervised learning
  - machine learning problems and tasks

Artificial Intelligence

Artificial intelligence

- **Artificial intelligence** (**AI**) is a set of technologies that, **mimicking human intelligence**, enable a **computational system** (such as, an information system, an automation system, or simply a computer) to perform a variety of **advanced functions and tasks** related to perception, learning, reasoning, problem-solving, and decision-making:
  - observe
  - read
  - understand
  - translate
  - analyze
  - solve problems
  - take (autonomous) decisions and action



Classification of the types of AI

- **Artificial intelligence** are classified depending on:
  - functionalities
  - capabilities
  - learning approaches
  - human interaction
  - creativity

Types of AI based on functionality

Reactive machines

- primitive AI
- produces outcomes from present data
- does not use context

Limited memory

- uses recent and past data to take decisions

Theory of mind

- understands emotions, sentiments and thoughts
- it is only concept

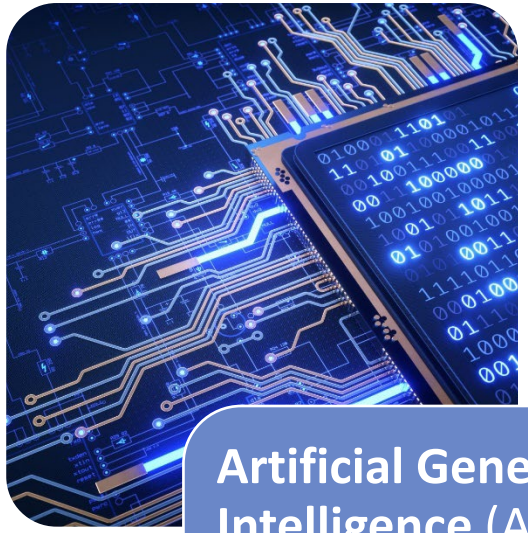
Self-awareness

- has its own feelings, needs, beliefs
- only hypothetical

Types of AI based on capability

Artificial Narrow Intelligence (ANI)

- solve single/relatively simple problems
- perform single tasks
- uses machine learning to process information
- now available for industry



Artificial General Intelligence (AGI)

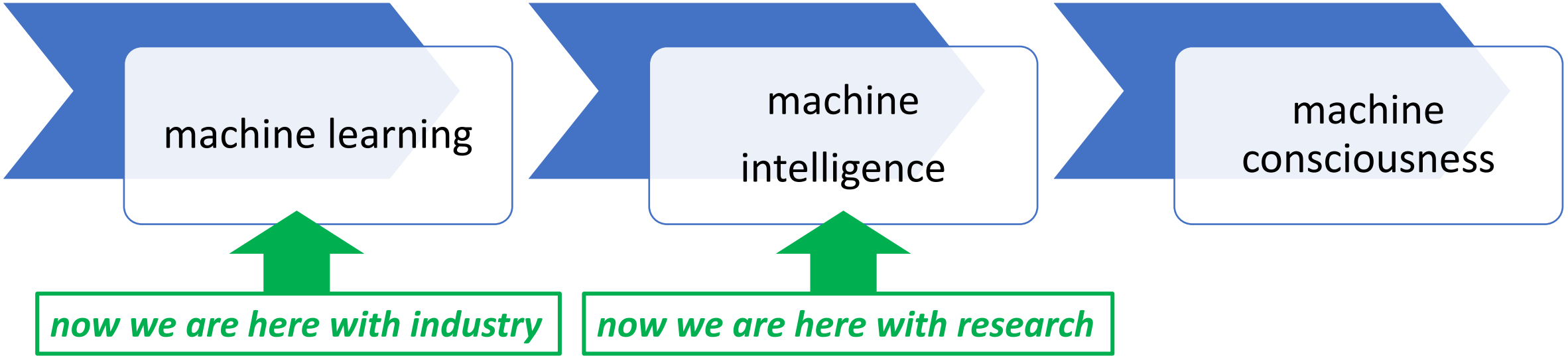
- human level of cognitive functions
- takes smart decisions
- comprises thousands of ANIs working together and communicating
- edge of research

Artificial Super Intelligence (ASI)

- surpasses human capabilities
- rational decision-making
- emotional relationships
- advance in realms that human being cannot even dream
- now still sci-fi for advanced robotics

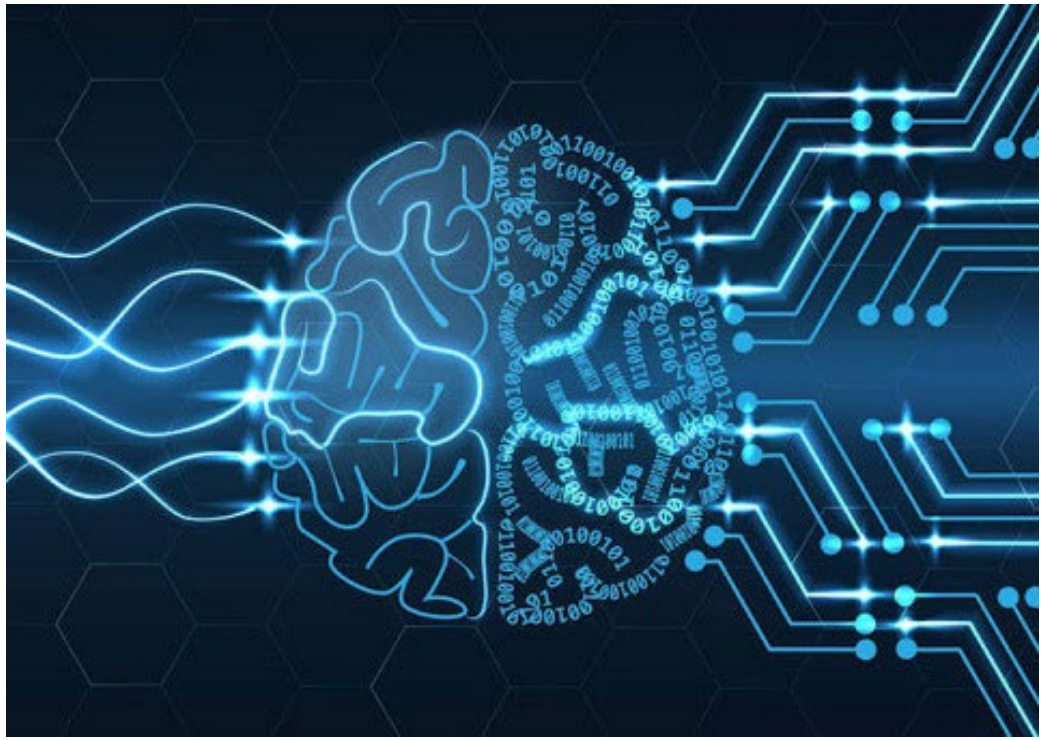
Where are we now?

- The path from ANI (Narrow Intelligence) to AGI (General) and ASI (Super) tracks the sequence of the passages between machine:
  - learning
  - intelligence
  - consciousness

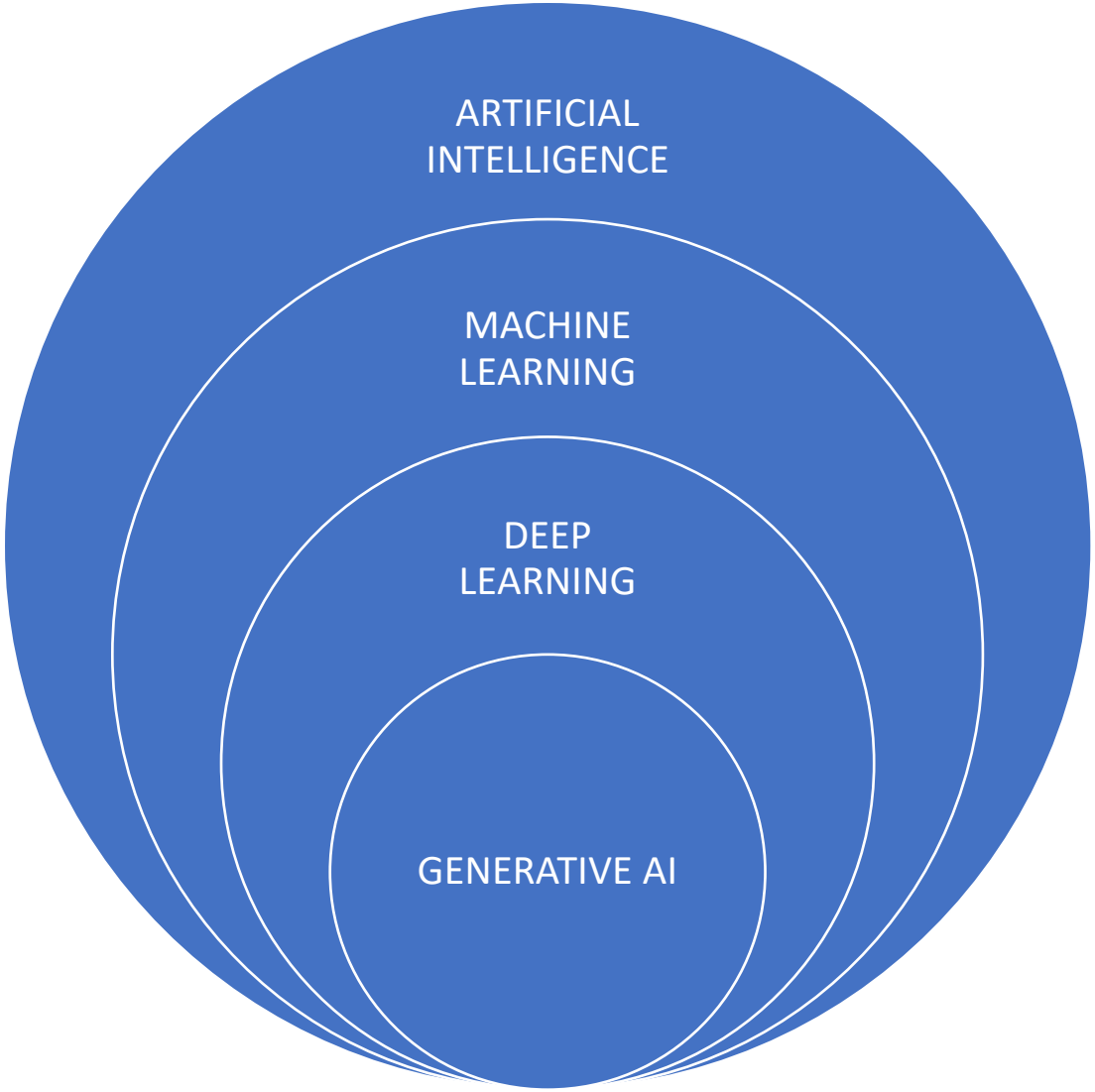


Machine learning: extracting value from data

- **Machine learning** is a branch of artificial intelligence that exploits algorithms (from mathematics, statistics and data science) that computer systems use to effectively **perform a specific task without using explicit instructions**, relying on patterns and inference instead
  - build a mathematical models of some "training data" in order to **take decisions without being explicitly programmed to perform the task**



AI and types of learning



Deep learning

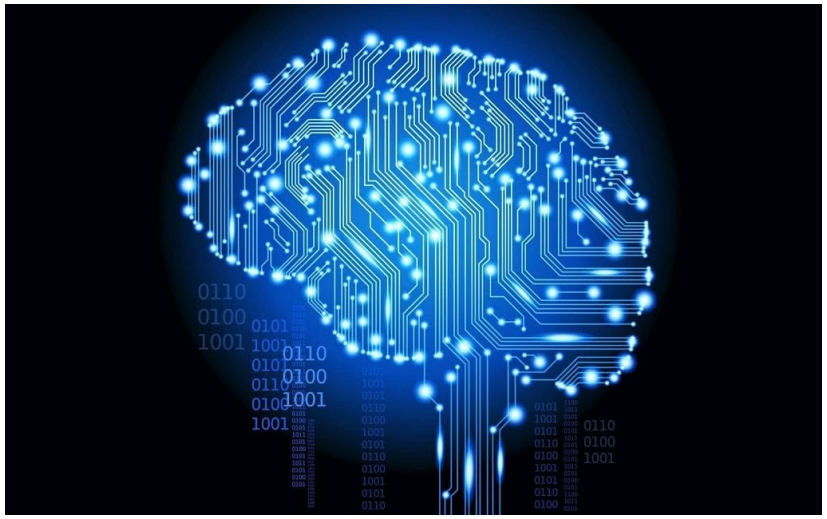
- uses artificial neural networks for predictive purposes

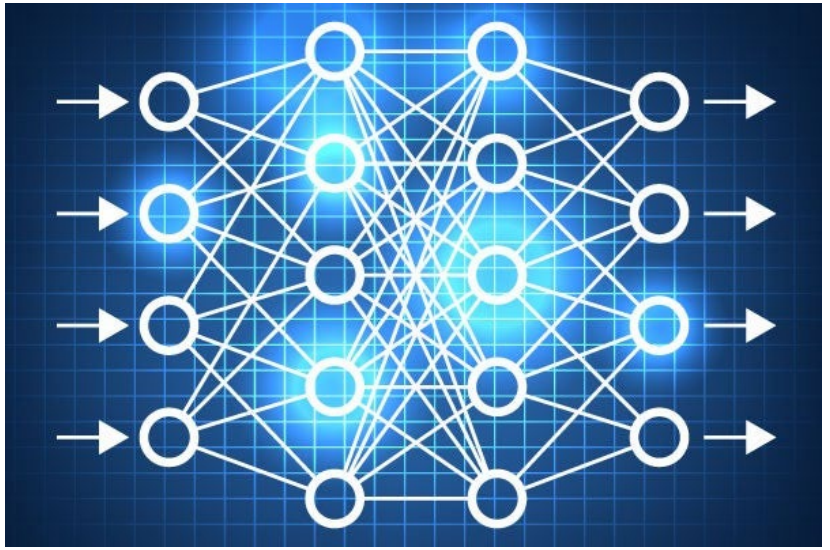
Generative AI uses generative models to produce text, images, videos, or other types of data:

- learns the patterns from training data
- use the learnt patterns to produce new data, which often comes in the form of natural language

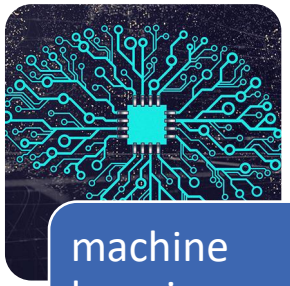
Deep Learning through Artificial Neural Networks

- **Artificial neural networks** (ANNs) are inspired to the biological structure of the neural network in the animal/human brain
  - made of a set of **neurons** (nodes of the network)
    - strongly interconnected
    - organized in layers
    - treat information in a strongly non-linear manner by activation functions
  - neurons are interconnected to exchange information between nodes, simulating **synapsis**
    - the strength of the signal is determined by weights
  - different architectures are available and different ways of data treatment are commonly used





Tasks performed by AI



machine learning

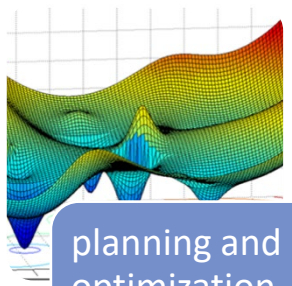
- data analytics
- deep learning

language processing

- extraction
- classification
- translation
- answering
- texting

expert systems

- mimic experts
- solve complex problems

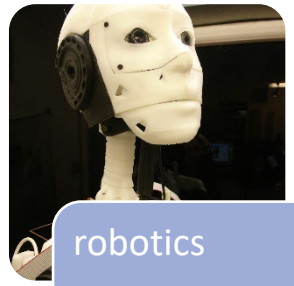


planning and optimization

- process optimization
- production scheduling

vision

- image recognition
- artificial vision systems



robotics

- perform human tasks

Industry 4.0 and digital transformation

Industry 4.0 in the process industry

- **Industry 4.0** integrates novel productive technologies through digitalization to improve business, productivity and quality of the production
- Classical technologies of digitization:
  - sensors and process analytical technologies
  - modelling tools
  - simulation software
  - advanced process control
  - soft sensors, etc.
- Tasks:
  - process understanding & troubleshooting
  - debottlenecking
  - monitoring
  - quality prediction
  - predictive maintenance
  - robotic-assisted manufacturing
  - process, product and technology transfer and scale-up

Data cahos

- Data volumes are increasing very much: almost 7x in few years
  - **global datasphere** is growing from **33 Zettabytes in 2018 to 221 Zettabytes by 2026**
  - ~50% of the world's stored data will reside in public cloud environments

- Consequences:
  - data growth and investment
  - Information Technology (IT) structural assessment
  - digital transformation competency
  - data value competency, etc.

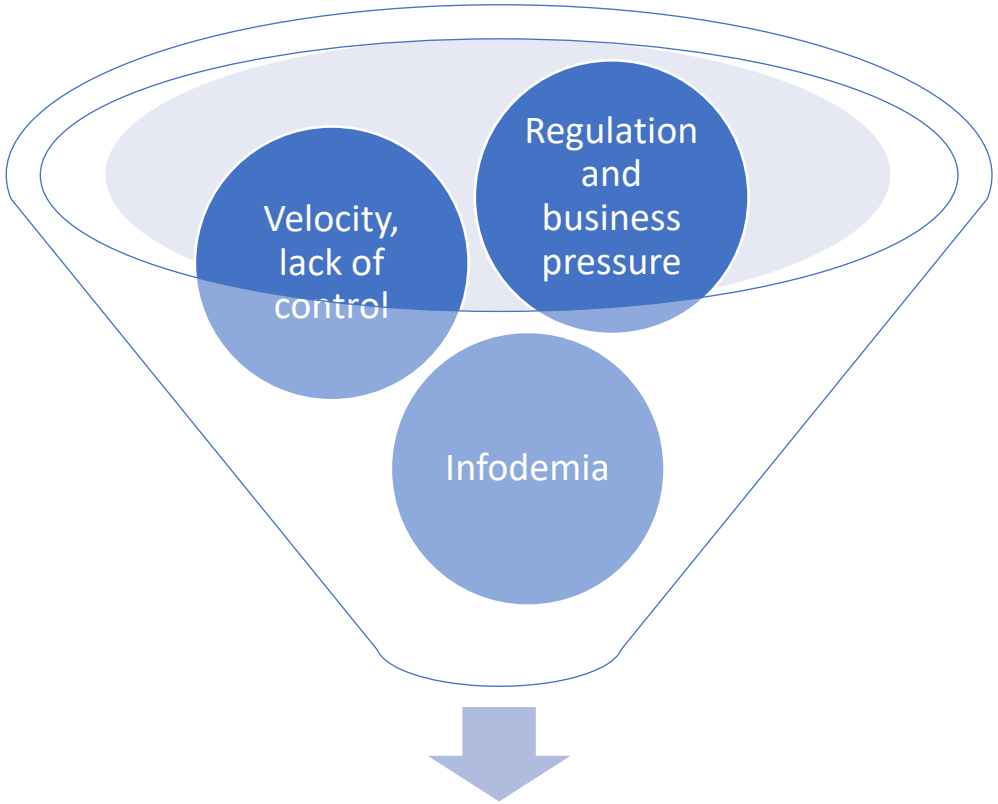
Impact of poor data quality

- **15 million dollars** average business losses due to poor data quality
  - loss of revenue
  - reduced efficiency
  - slow or flawed decision making
  - compliance risks
  - missed opportunities
  - reputation damage

Efficiency crisis

- Time spent on **data management and analysis**:
  - 19% analyzing
  - **81% managing**
    - 20% searching
    - 37% preparing
      - >11% data sorting, shaping, preparation for presentation
      - then cleaning, visualization, etc…
    - 24% protecting

Complex alchemy

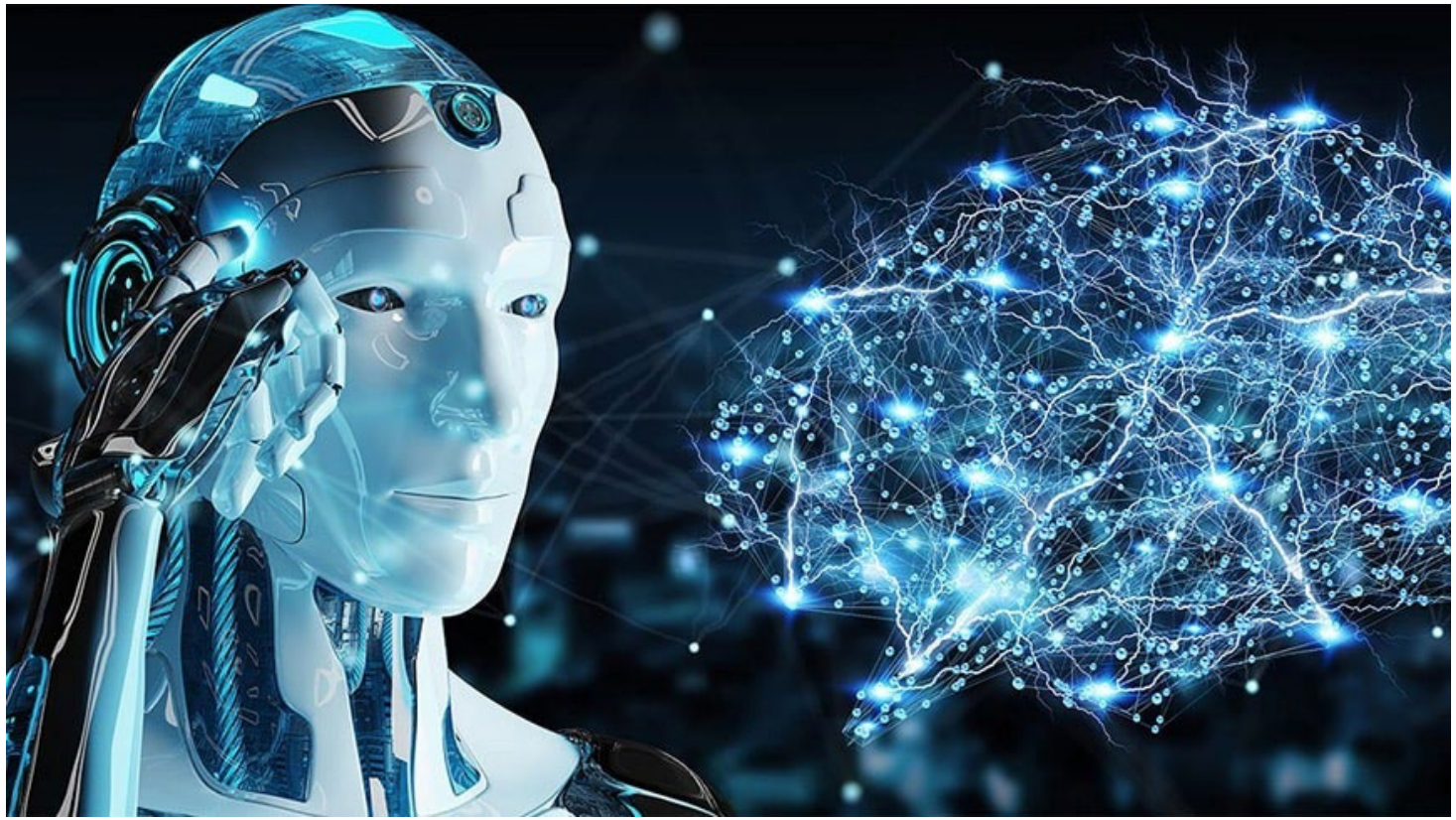


Needs

- Systems that **deal with data** should be:
  - trusted and reliable
  - fast
  - translated by trained personnel to the business
  - adaptability to:
    - volatility and changes
    - uncertainty
    - lack of stability
    - complexity
    - ambiguity (business, trends, competitors)

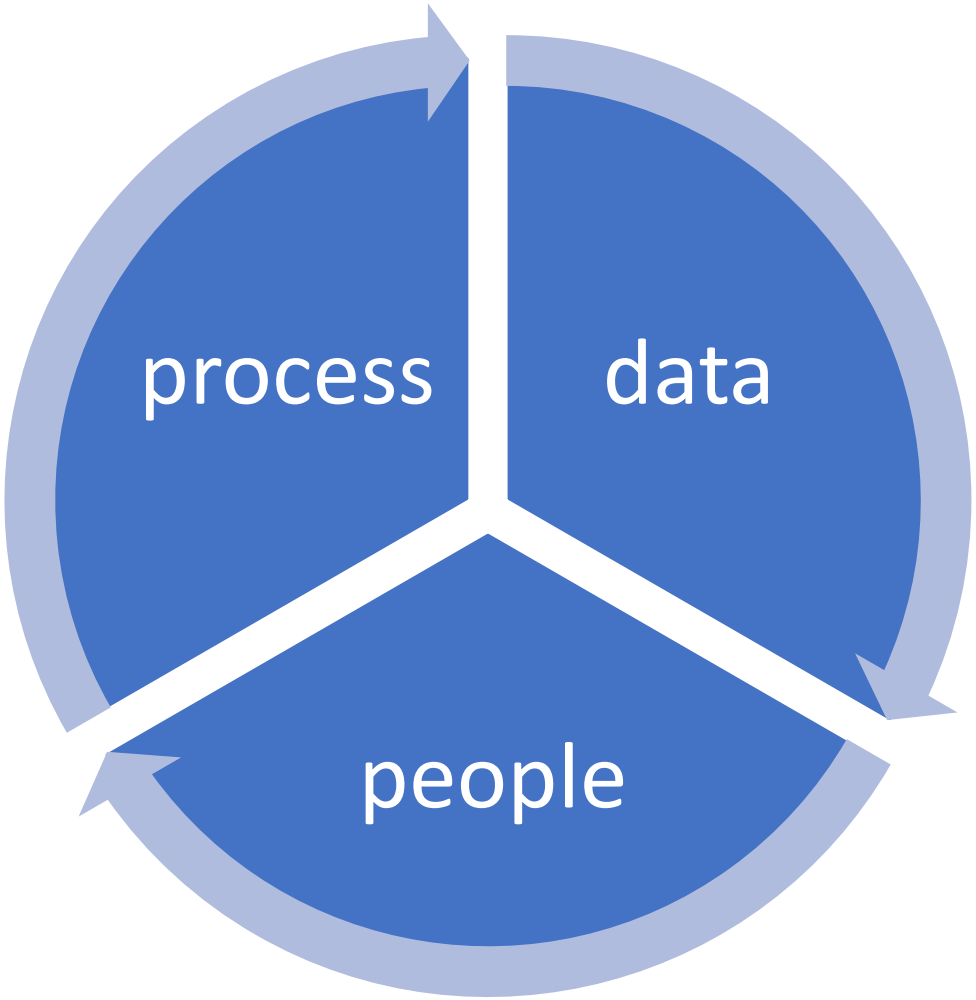
Challenges

- **Automation modernization**
- **Data analysis**
- Efficient infrastructures
- Endpoint devices
- Security
- Data integrity, protection and compliance
- Etc.

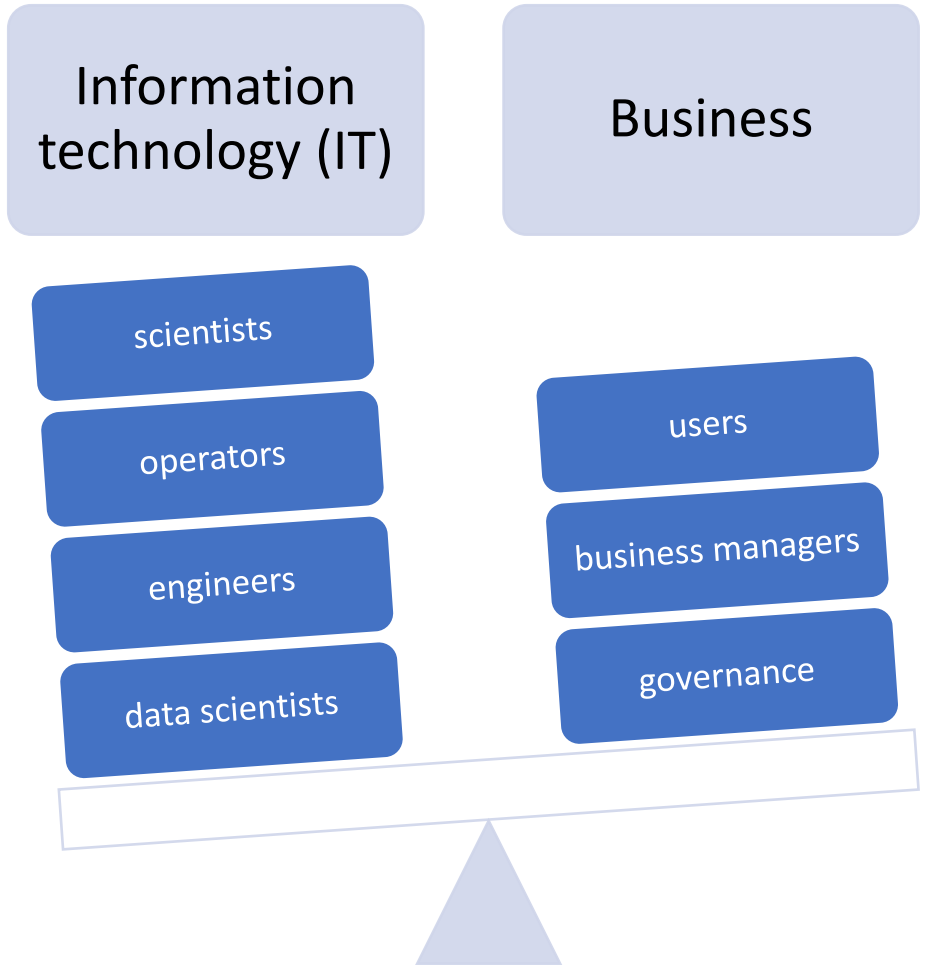


… so don’t think this is easy!

Key dimensions for the success

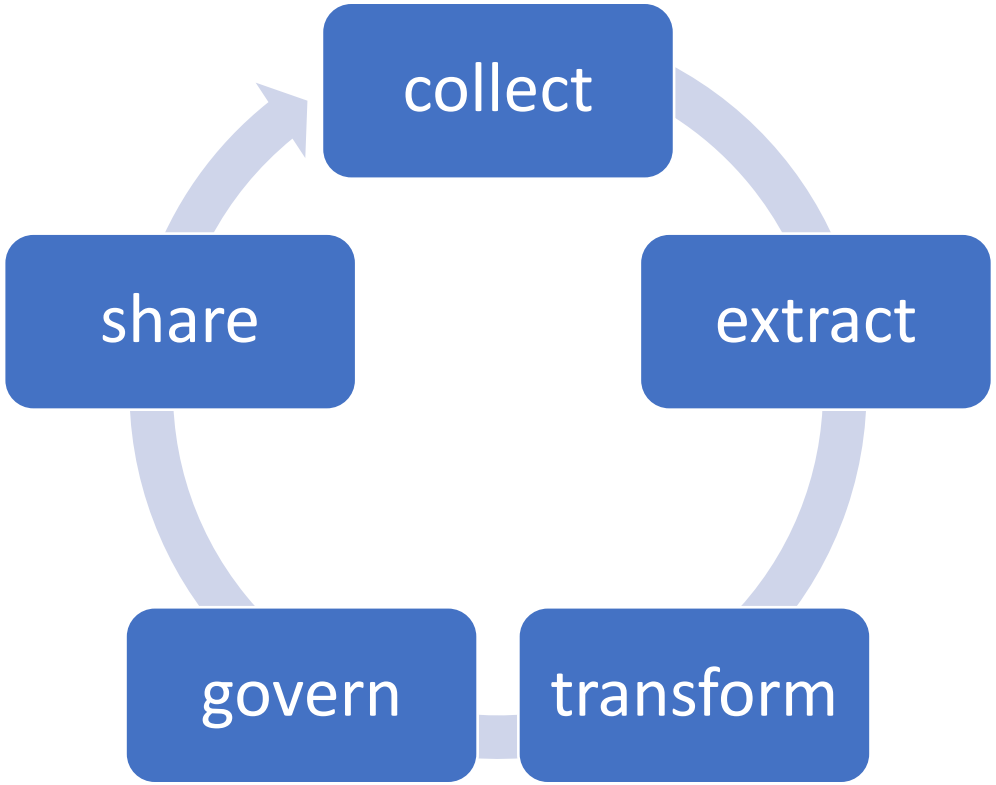


People


Information technology (IT)
Business
scientists
operators
engineers
data scientists
users
business managers
governance

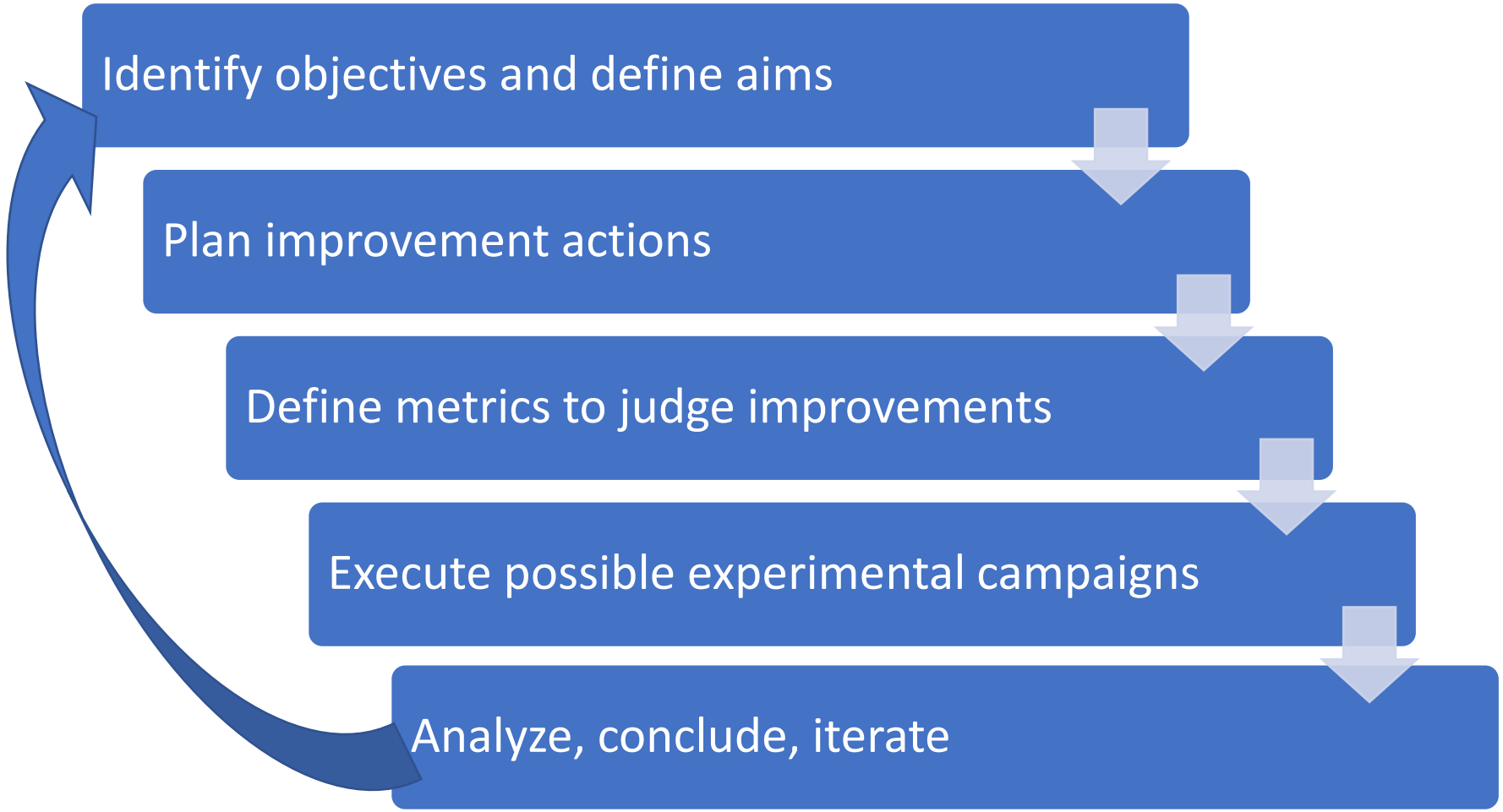
Team approach

- data analysis
  - IT developers
  - data analysts
  - **data translators**
- process
  - engineers
  - operators
  - scientists
- business
  - governance
  - management
  - costumers



Process

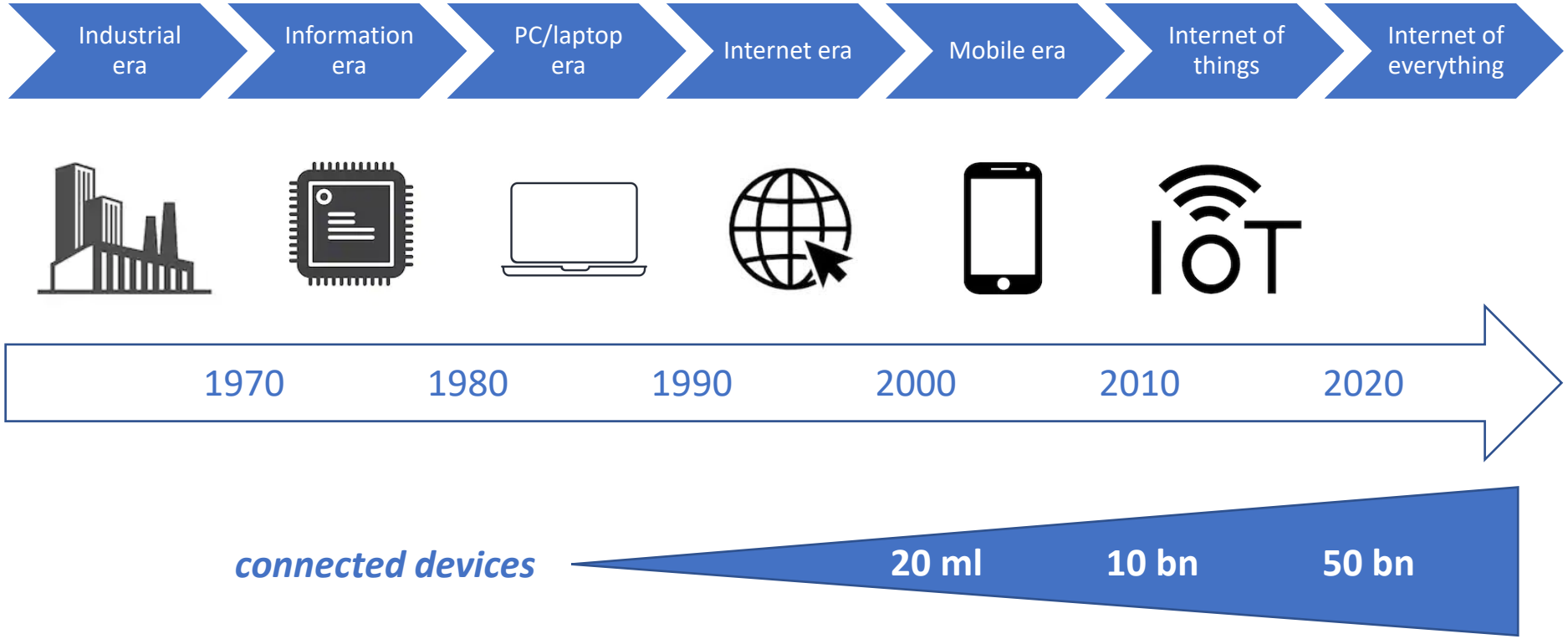
- Processes aligns people and data



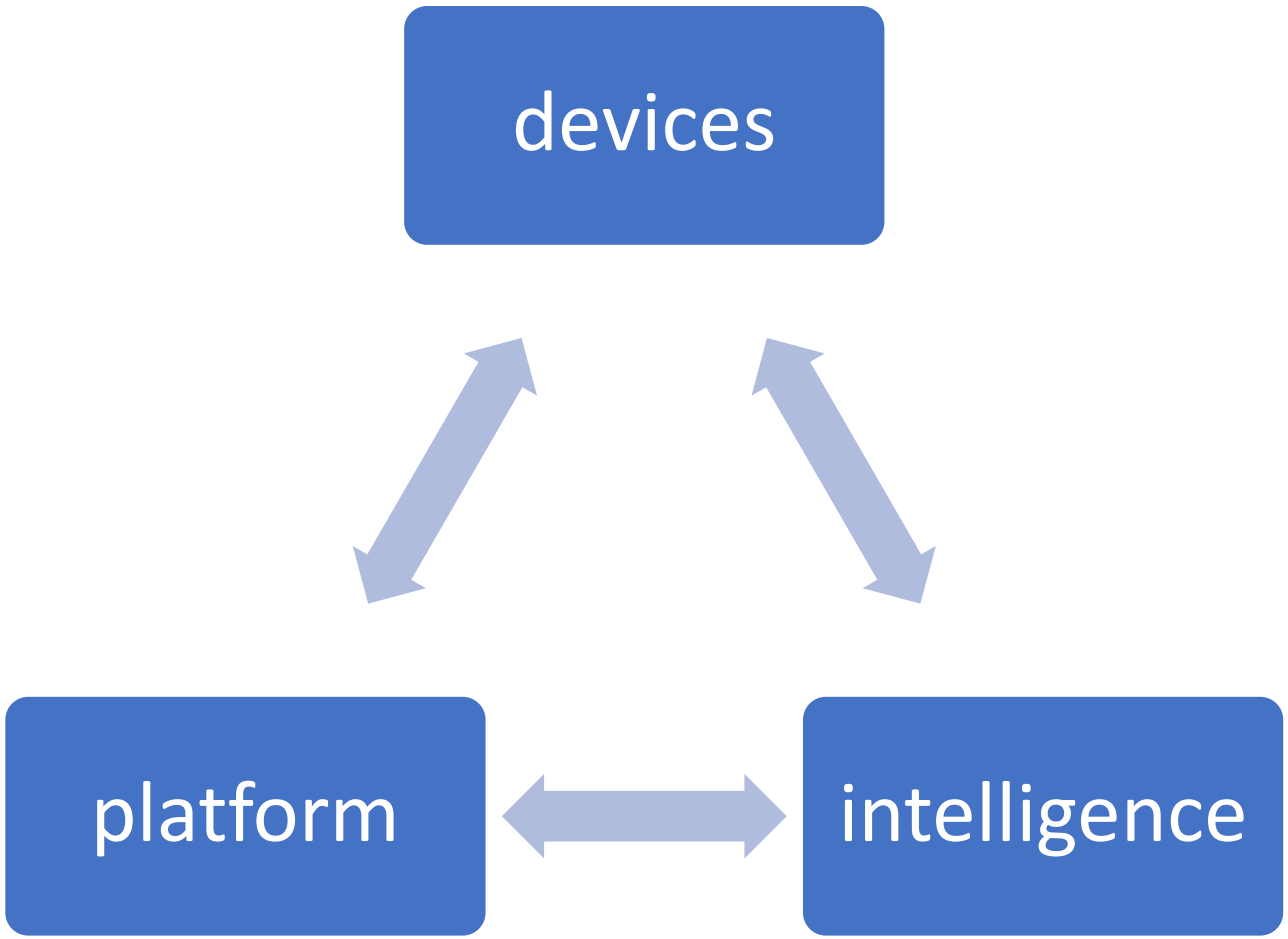
Internet of Everything

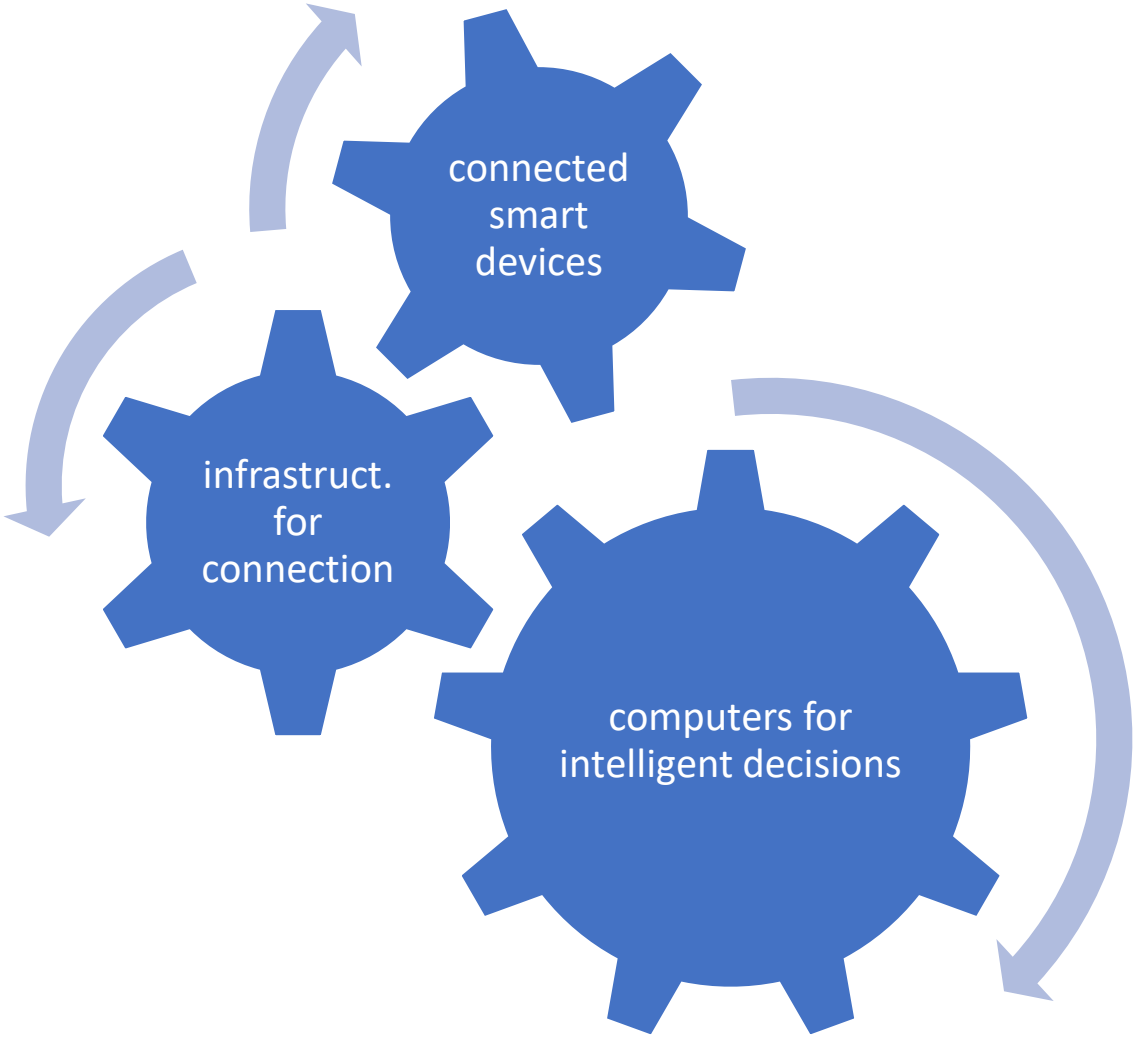
Internet of Things

- **Internet of Things** (IoT) is the giant network of interconnected devices (objects) which are able of making decisions without the human intervention



IoT principles






Welcome to the big
data era!

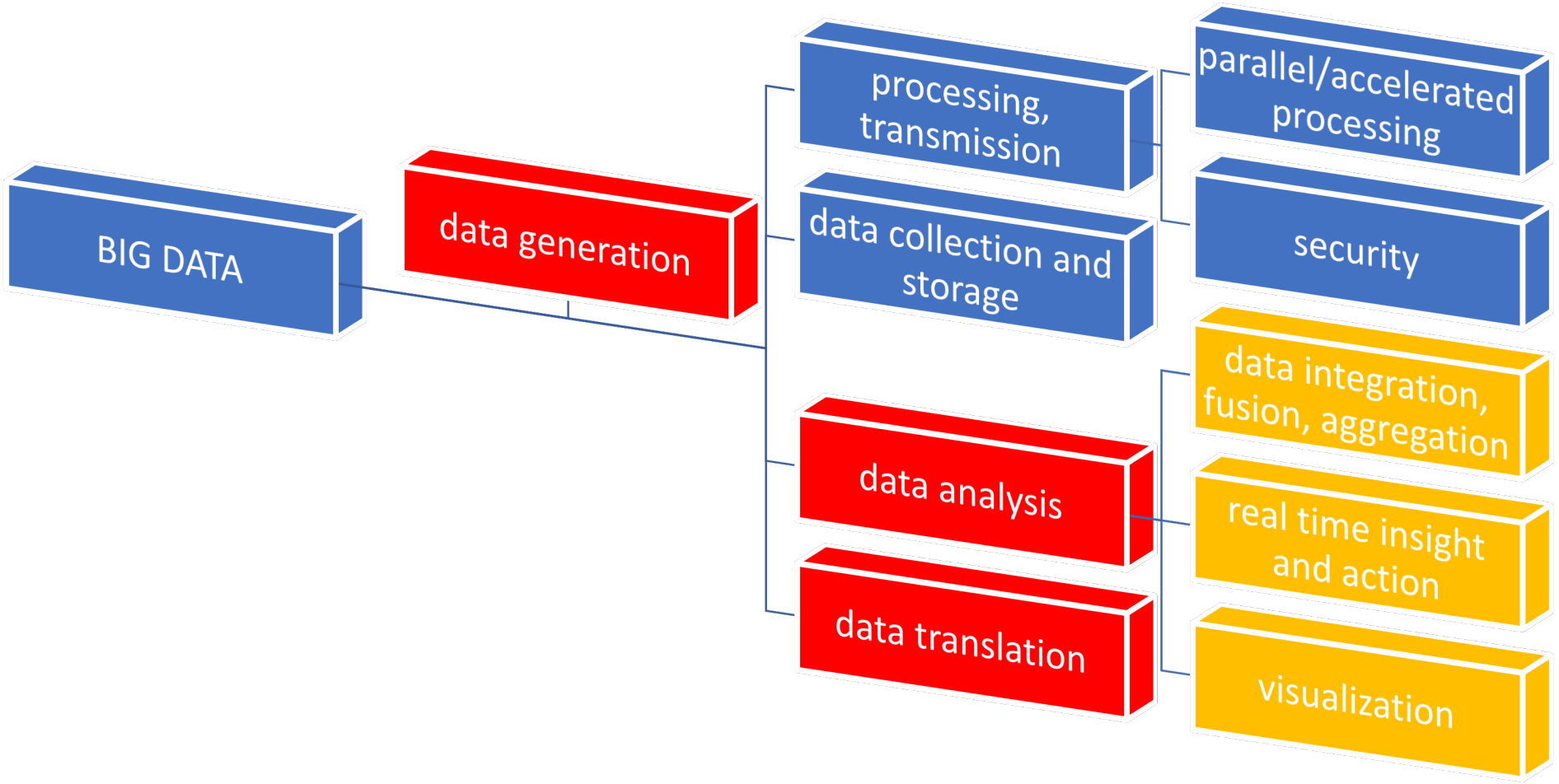
What really is big data?

- Someone said that:

"Big data is like sex for teenagers:
everyone talks about it,
nobody really knows how to do it,
but since everyone thinks everyone else is doing it,
everyone claim they are doing it!"

- Big data pose several challenges to engineers!

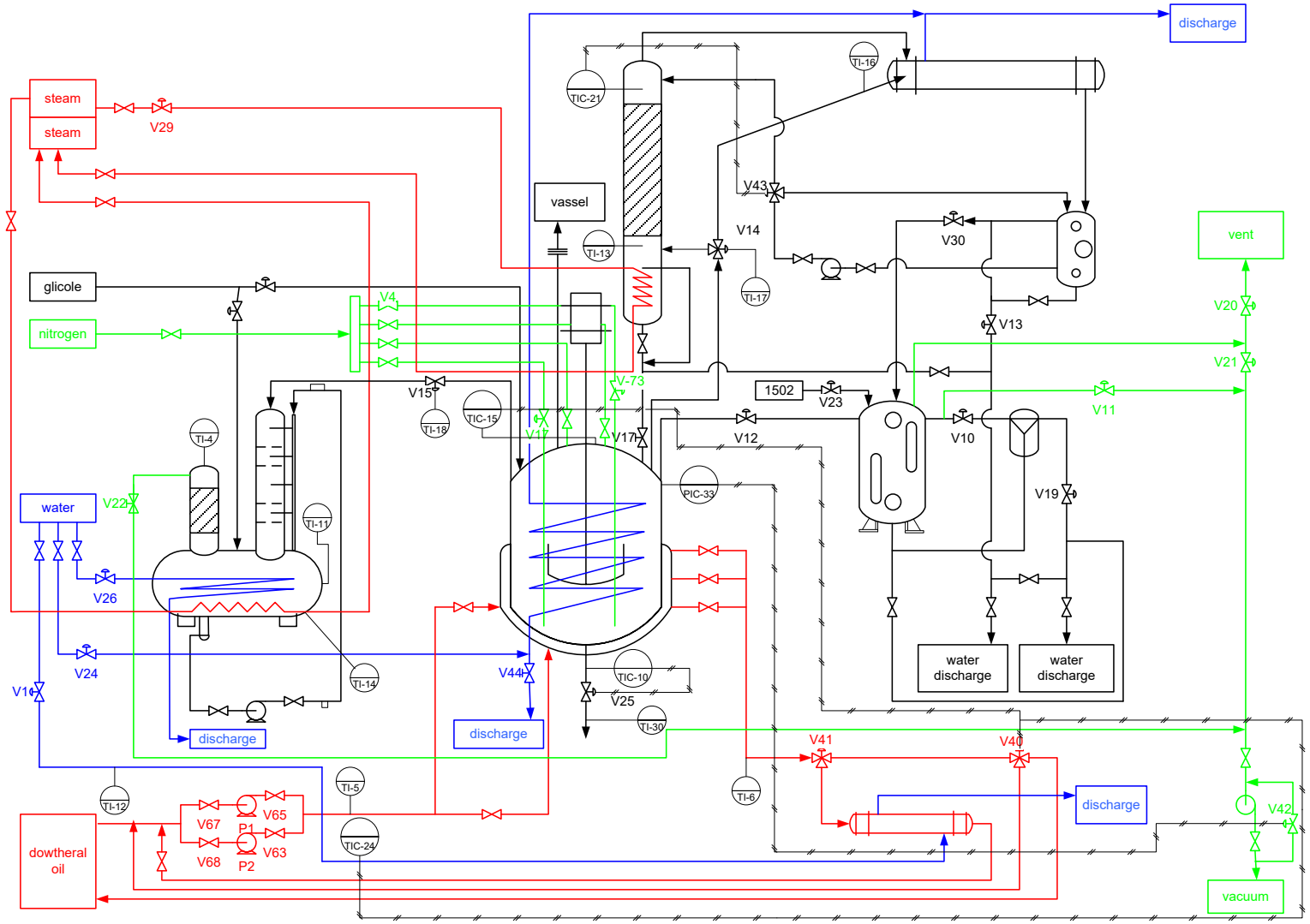
Big data challenges



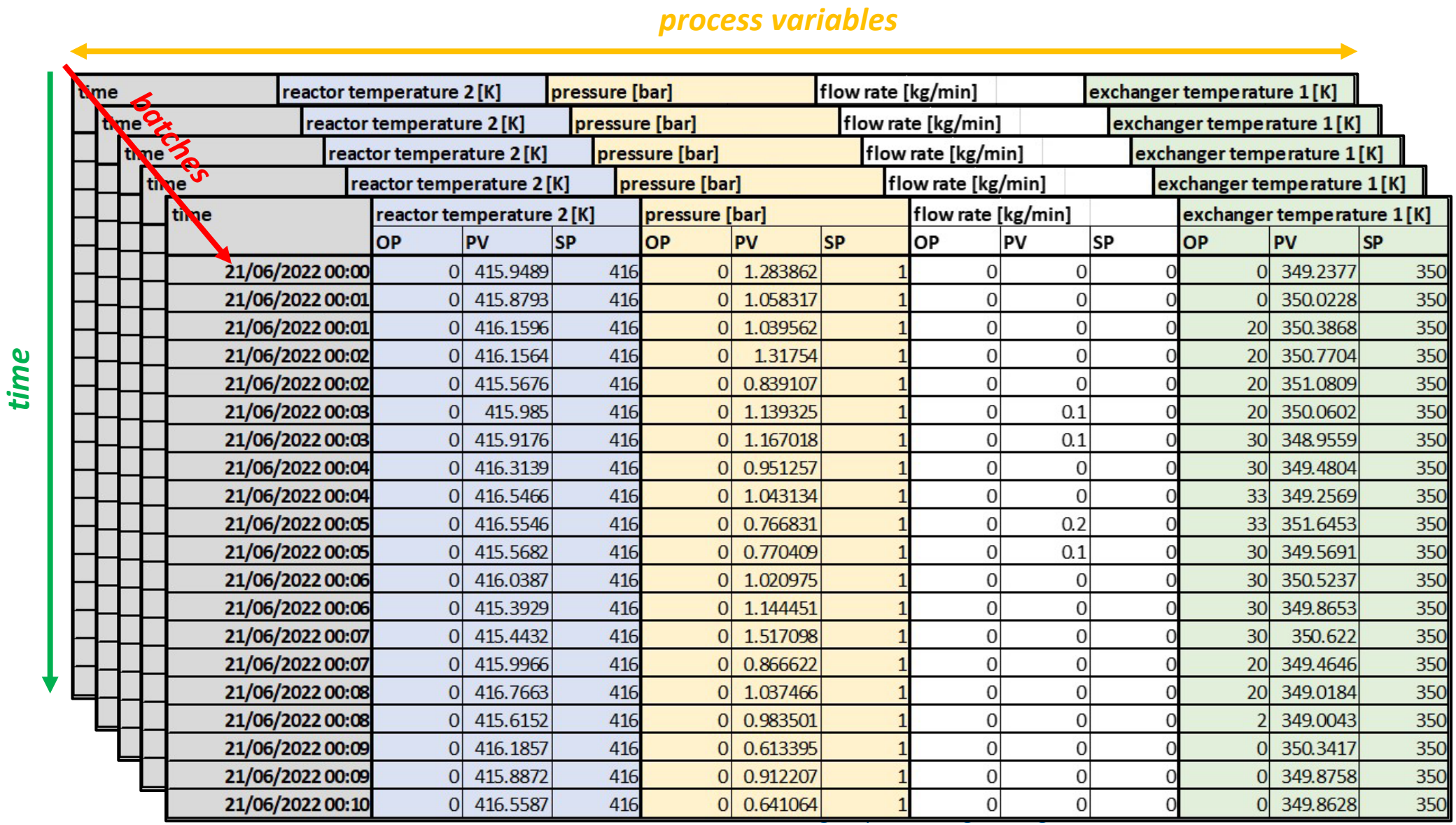
Pyramid discussion

- What are the data you think are involved in data analytics and machine learning?

Example of a chemical plant layout



Online plant instrumentation data



time	reactor temperature 2 [K]			pressure [bar]			flow rate [kg/min]			exchanger temperature 1 [K]		
	OP	PV	SP	OP	PV	SP	OP	PV	SP	OP	PV	SP
21/06/2022 00:00	0	415.9489	416	0	1.283862	1	0	0	0	0	349.2377	350
21/06/2022 00:01	0	415.8793	416	0	1.058317	1	0	0	0	0	350.0228	350
21/06/2022 00:01	0	416.1596	416	0	1.039562	1	0	0	0	20	350.3868	350
21/06/2022 00:02	0	416.1564	416	0	1.31754	1	0	0	0	20	350.7704	350
21/06/2022 00:02	0	415.5676	416	0	0.839107	1	0	0	0	20	351.0809	350
21/06/2022 00:03	0	415.985	416	0	1.139325	1	0	0.1	0	20	350.0602	350
21/06/2022 00:03	0	415.9176	416	0	1.167018	1	0	0.1	0	30	348.9559	350
21/06/2022 00:04	0	416.3139	416	0	0.951257	1	0	0	0	30	349.4804	350
21/06/2022 00:04	0	416.5466	416	0	1.043134	1	0	0	0	33	349.2569	350
21/06/2022 00:05	0	416.5546	416	0	0.766831	1	0	0.2	0	33	351.6453	350
21/06/2022 00:05	0	415.5682	416	0	0.770409	1	0	0.1	0	30	349.5691	350
21/06/2022 00:06	0	416.0387	416	0	1.020975	1	0	0	0	30	350.5237	350
21/06/2022 00:06	0	415.3929	416	0	1.144451	1	0	0	0	30	349.8653	350
21/06/2022 00:07	0	415.4432	416	0	1.517098	1	0	0	0	30	350.622	350
21/06/2022 00:07	0	415.9966	416	0	0.866622	1	0	0	0	20	349.4646	350
21/06/2022 00:08	0	416.7663	416	0	1.037466	1	0	0	0	20	349.0184	350
21/06/2022 00:08	0	415.6152	416	0	0.983501	1	0	0	0	2	349.0043	350
21/06/2022 00:09	0	416.1857	416	0	0.613395	1	0	0	0	0	350.3417	350
21/06/2022 00:09	0	415.8872	416	0	0.912207	1	0	0	0	0	349.8758	350
21/06/2022 00:10	0	416.5587	416	0	0.641064	1	0	0	0	0	349.8628	350

Example of laboratory data

quality variables

time

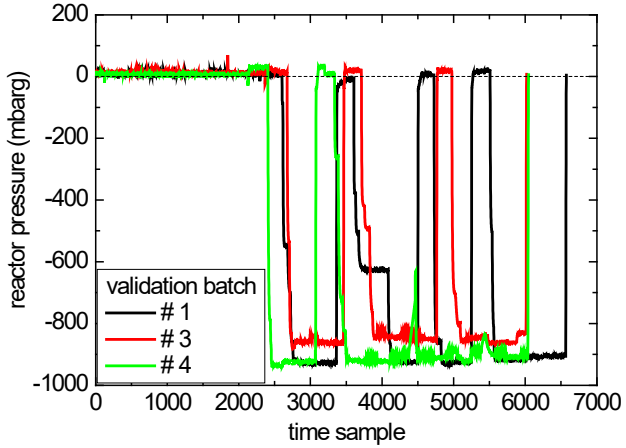
time	manual annotation	quality index 1	quality index 2
21/06/2022 00:00	start of new batch	13.68	31.36
21/06/2022 00:01		10.22	32.02
21/06/2022 00:01		12.21	29.73
21/06/2022 00:02		10.91	25.03
21/06/2022 00:02		12.61	23.38
21/06/2022 00:03		10.79	27.58
21/06/2022 00:03		12.97	27.33
21/06/2022 00:04		13.47	22.19
21/06/2022 00:04	correction	18.42	15.25
21/06/2022 00:05	after correction	11.61	20.62
21/06/2022 00:05		7.72	33.49
21/06/2022 00:06		10.32	26.88
21/06/2022 00:06		14.74	15.85
21/06/2022 00:07		9.85	31.81
21/06/2022 00:07		13.92	23.01
21/06/2022 00:08		12.24	25.65
21/06/2022 00:08		14.87	16.25
21/06/2022 00:09		8.07	37.22
21/06/2022 00:09		11.6	27.84
21/06/2022 00:10		9.58	29.93

Data types

- **Process data**
  - *hardware sensors*
  - *lab analysis*
  - *images*
  - *process analytical technologies*
    - *NIR spectra*
    - *etc…*

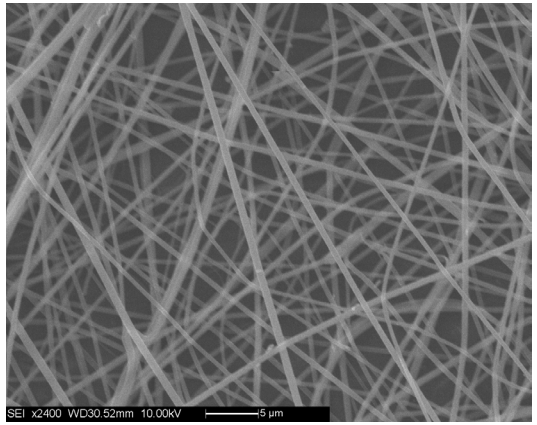
- **Data features**
  - *high quantity (giga/tera/zetta-byte)*
  - *correlation (also in space and time)*
  - *noise*
  - *missing data, etc…*

pressures, temperatures, etc…

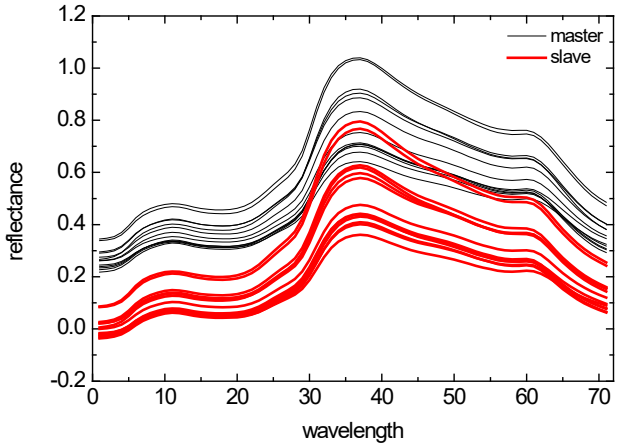


RGB images

SEM images

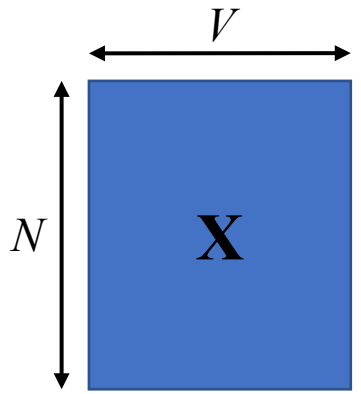


NIR spectra



Data tables

- Usually, **large data tables** are commonly available:
  - R&D
  - industrial data (large instrumentation of the manufacturing lines)
  - academia
  - every-day life



- Data tables are usually structured as matrices, such as $\mathbf{X}$:
  - N observations (units) on the rows
  - V variables on the columns

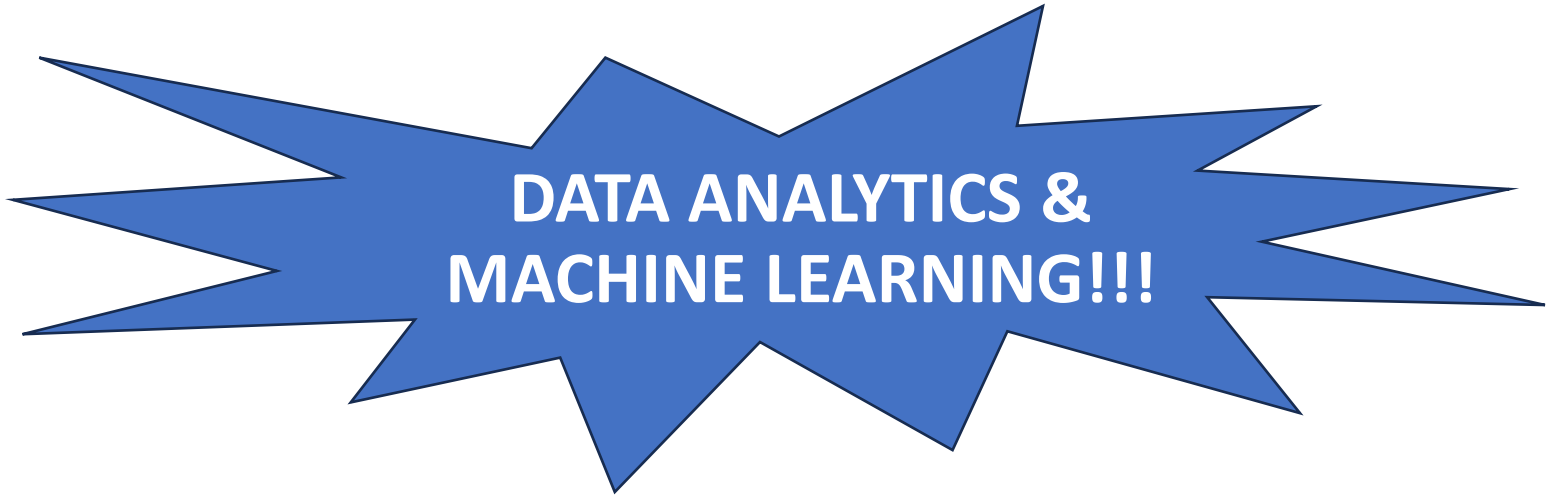
- However, large tables (large N and/or large V) does not mean having *appropriate and deep insight on the data*

- Data should be organized in a proper manner to be properly treated
  - data management is one of the keys to succeed in data analysis!

Extracting information from data

- It is difficult to extract meaningful information from **large amounts of raw data**
- Nowadays, observing few (statistical) indices on data is not sufficient

More sophisticated computer-based methodologies are needed!



Data analytics and machine learning

Supervised and unsupervised learning

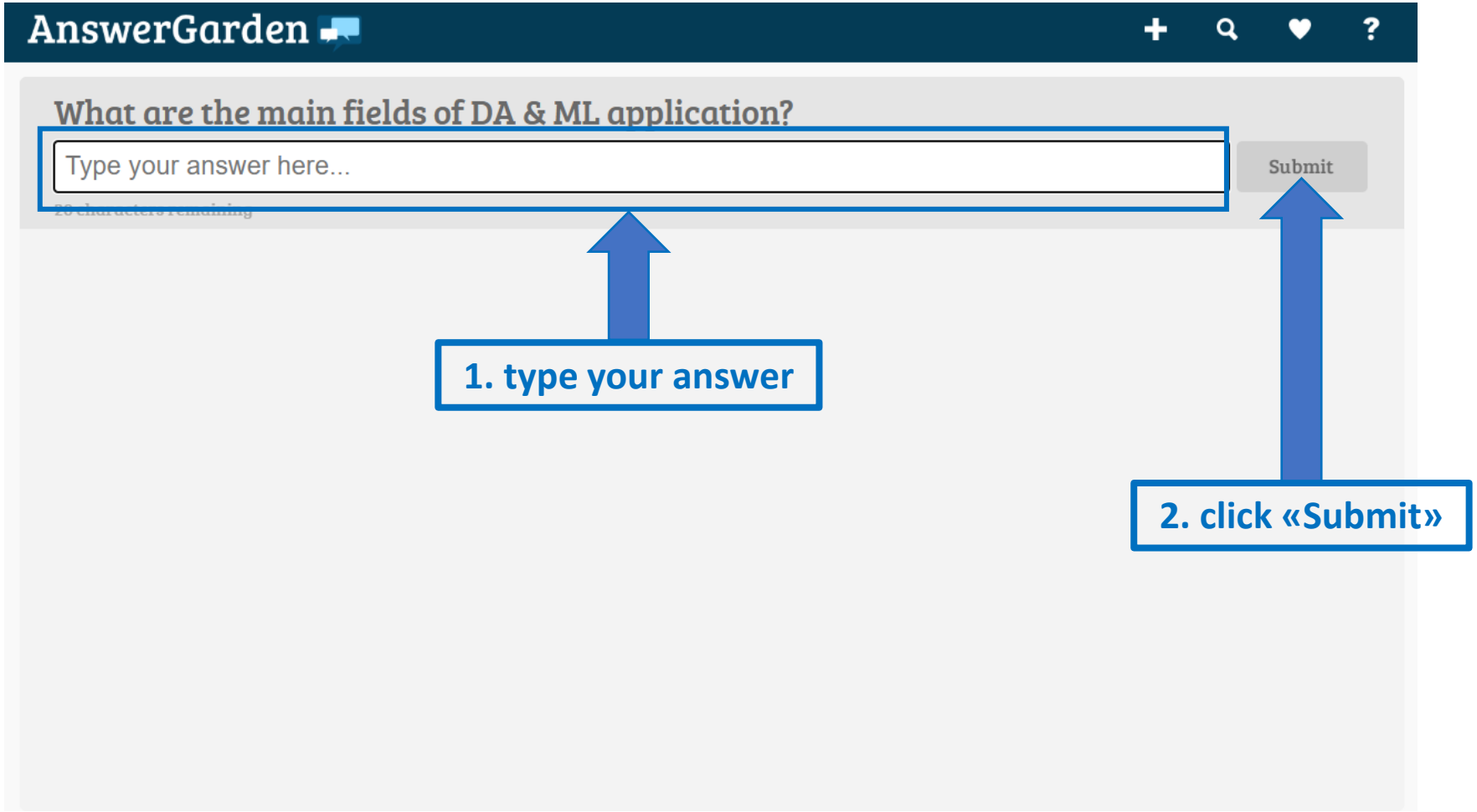
Data analytics

- **Data analytics** is the process of **examining data sets** in order to **draw meaningful conclusions** about the information they contain **with the aid of specialized systems and software**

What are the main fields of DA & ML application?

https://answergarden.ch/5135125

Assess the internet, find answers and…



Different fields of application

- **Industrial sectors**:
  - basic chemicals
  - petrochemicals
  - polymer
  - plastics
  - fiber
  - resins, coatings, paints, adhesives
  - automobiles
  - (bio)pharmaceuticals
  - biotech
  - paper and pulp
  - food and beverages
  - mining
  - metals and materials
  - semiconductors
  - …
  - consumer science
  - marketing
  - telecommunication

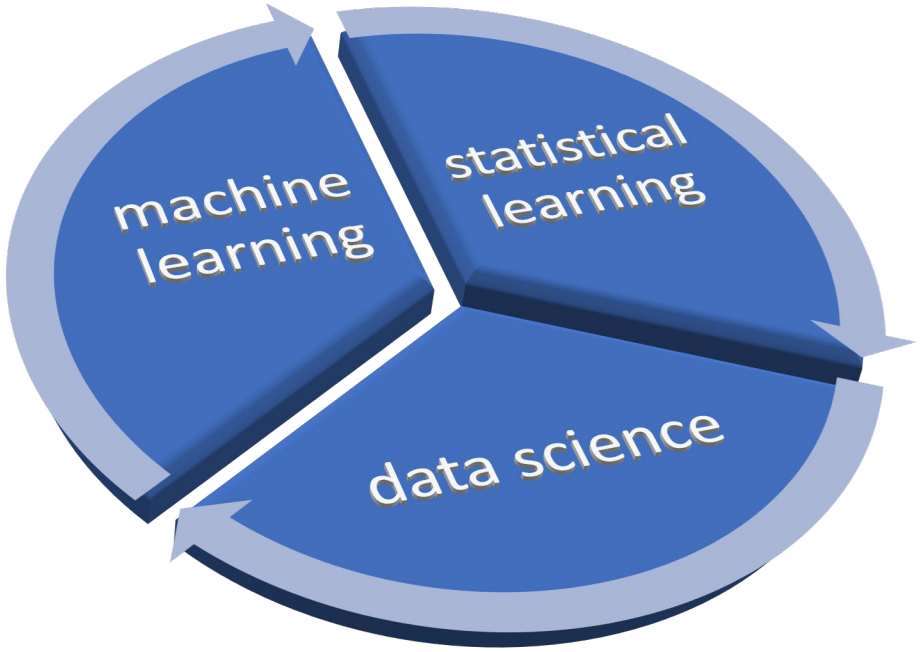
Areas:

- analytical technologies
- instrumentation calibration
- product formulation
- process development
- data mining
- process understanding
- process troubleshooting
- process monitoring
- quality control
- artificial vision systems
- security

Data analytics

- **Data analytics** is the process of **examining data sets** in order to **draw meaningful conclusions** about the information they contain **with the aid of specialized systems and software**
  - initiatives to:
    - enable organizations to make more-informed decisions
    - respond more quickly to emerging market trends and gain a competitive advantage over rivals
    - aid scientists and researchers to verify or disprove scientific models, theories and hypotheses
    - improve experimental campaigns
    - increase operational efficiency
    - help businesses to increase revenues
    - optimize marketing campaigns and customer service efforts
  - activities:
    - numerical data analysis and statistical analysis
    - understanding the content of information into data
    - dealing with a variety of data types: images, audios, videos, etc…
  - tools:
    - exploratory data analysis and data mining
    - pattern recognition
    - classification and clustering
    - predictive modelling
    - machine and deep learning
    - artificial intelligence

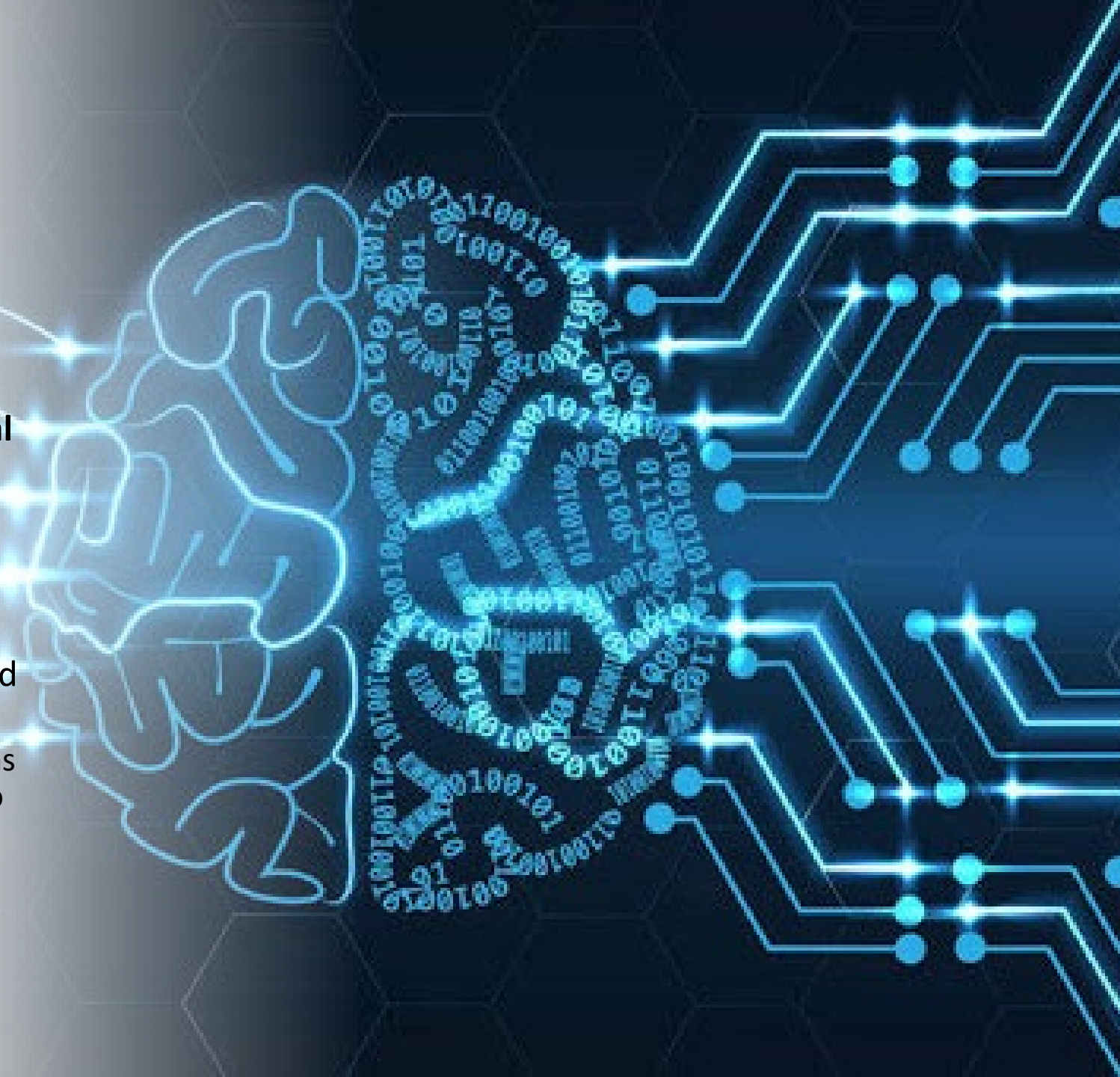
Data science



- **Data science** is the extraction of knowledge from data using:
  - mathematics
  - statistics
  - computer science, etc…
- **Statistical learning** is a branch of applied statistics with the main focus on statistical modelling and uncertainty assessment

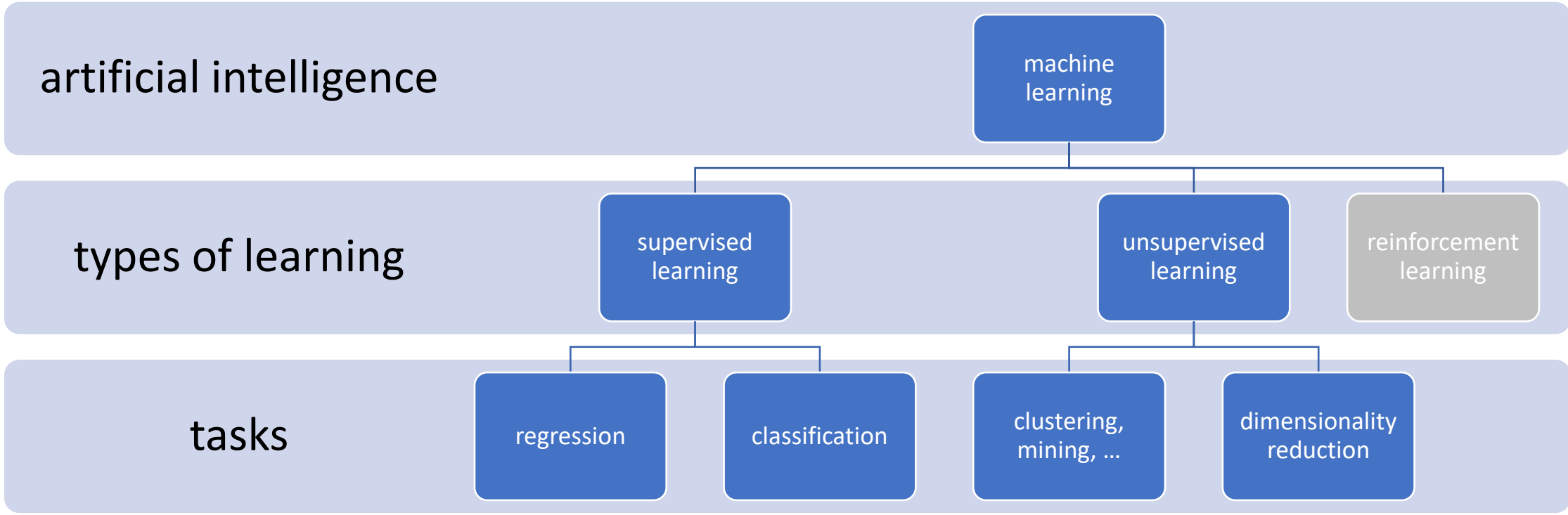
Machine learning: extracting value from data

- **Machine learning** is a branch of **artificial intelligence** that exploits algorithms (from mathematics, statistics and data science) that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead
  - build a mathematical models of some "training data" in order to take decisions without being explicitly programmed to perform the task



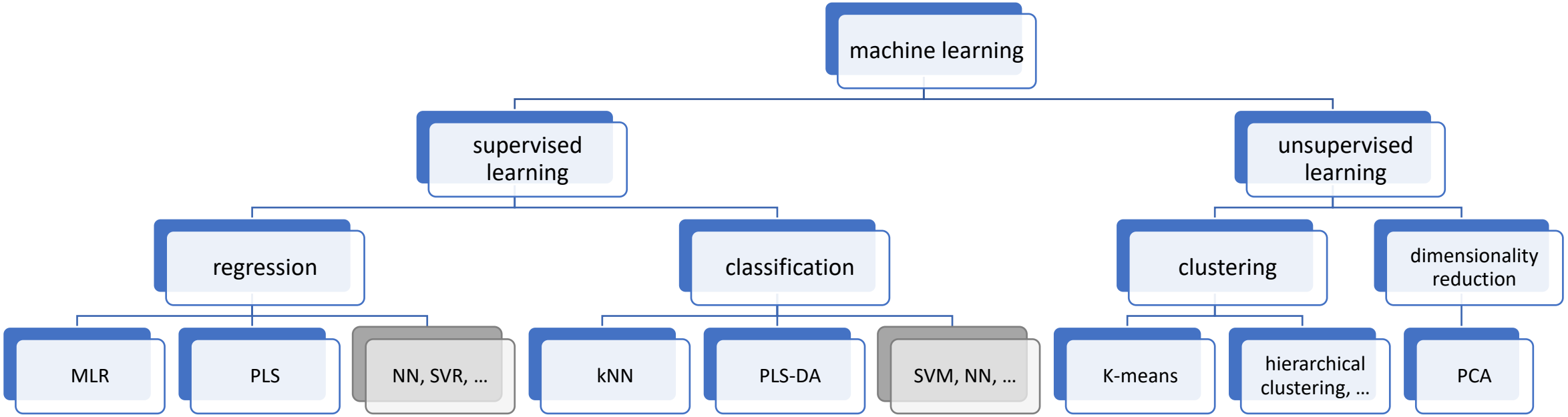
Machine learning

- Machine learning is based on supervised and unsupervised learning:
  - **unsupervised learning**: data grouping and interpretation based on **input data only**
  - **supervised learning**: predictive models based on **inputs-outputs relations**
  - **reinforcement learning**: an **agent** learns to **take decisions** by **performing actions** in a **dynamic environment** to maximize a **reward**



Machine learning techniques

- A lot of **modelling techniques** are commonly utilized in machine learning for both supervised and unsupervised learning
  - unsupervised learning methods are **descriptive**
    - finding natural grouping of observations in data
    - discovering interesting relations among variables
  - supervised learning methods are **predictive**
    - predicting a continuous/categorical attribute
    - learning a method to predict classes/variables from pre-labeled instances



Unsupervised learning

- The goal of **unsupervised learning** (or “learning without a teacher”) is to directly infer, without the help of a supervisor, namely, without having a priori knowledge on the relation among observations
  - the dimension of $\mathbf{X}$ may be much higher than in supervised learning
  - the inputs $\mathbf{X}$ represents all the variables under consideration
- Pros and cons:
  - **good point**: it is not required to infer how the properties of $\mathbf{X}$ change, conditioned on the changing values of another set of variables $\mathbf{Y}$
  - **bad point**: with unsupervised learning there is **not a direct measure of success** that can be used to:
    - judge model adequacy
    - to compare the effectiveness of different methods over various situations/models
      - this situation led to heavy proliferation of methods

Supervised learning

- The problem is observing if a set of input variables (which are measured or preset) have some influence on one or more outputs:
  - the **inputs X** $[N \times V]$ are often called (the terms can be used interchangeably):
    - predictors
    - independent variables
    - factors (in Design of Experiments)
    - features (in the pattern recognition literature)
    - regressors
  - the **outputs Y** $[N \times M]$ are called:
    - responses
    - dependent variables

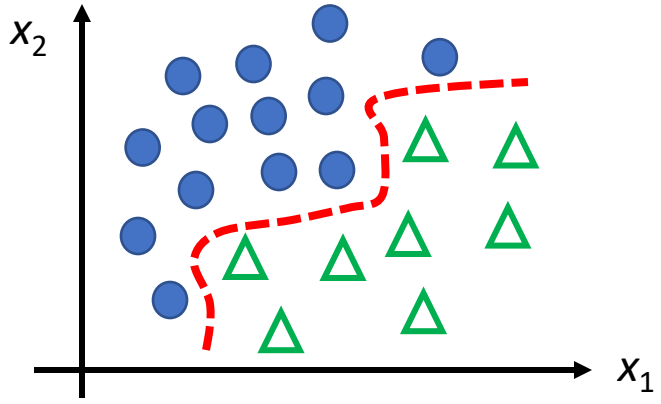
- The **goal** of **supervised learning** is to use the inputs to **predict/estimate** the values of the outputs

Regression and classification

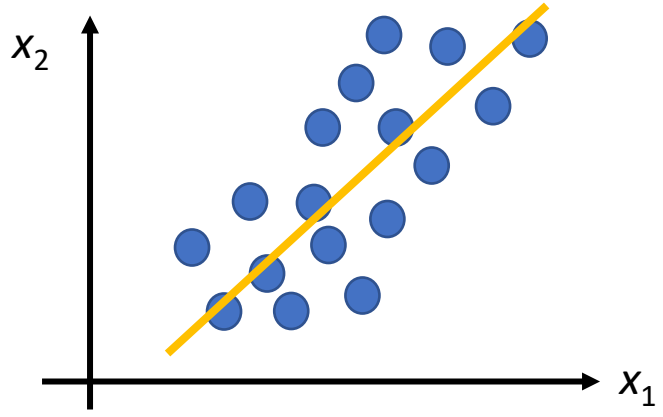
- Conventionally, different prediction/estimation tasks (which have a lot in common!) are determined by distinct output types:
  - **regression**: **quantitative outputs** prediction/estimation
  - **classification**: **qualitative outputs** prediction/estimation
    - both can be viewed as a task in function approximation

supervised learning

classification



regression



Applications of classification and regression

- For all the types of inputs and outputs pairs it makes sense to think of using the inputs to predict/estimate the output
  - regression
    - given some specific biological and chemical measurements (e.g.: viability, pH, dissolved oxygen) in the previous days of an experiments, the titer can be predicted
  - classification
    - given the atmospheric temperature, humidity, wind, etc., we want to forecast tomorrow's weather

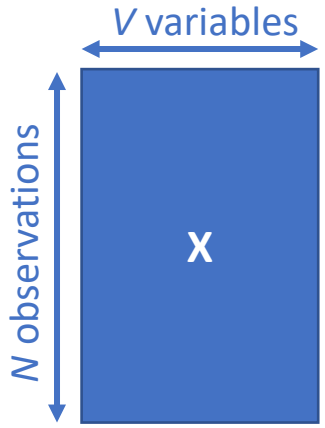
Machine learning methodologies

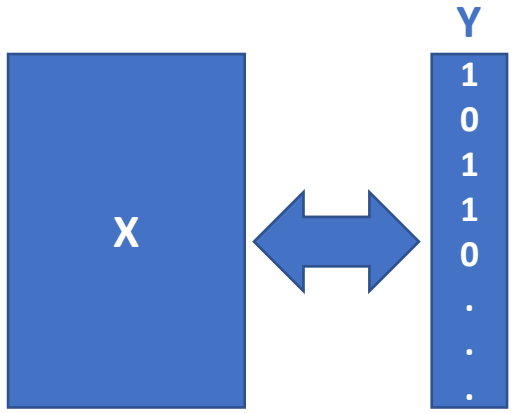
CLUSTERING vs. **CLASSIFICATION** vs. **REGRESSION**

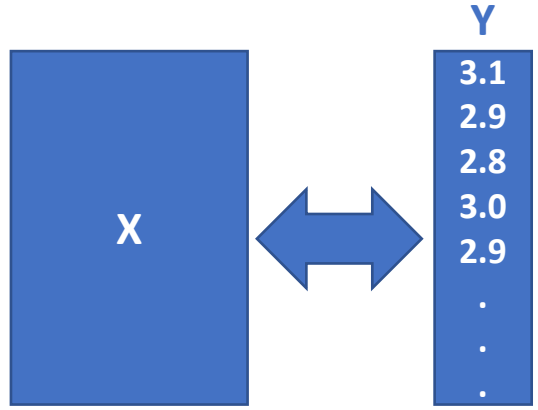
- only input variables **X**

- both inputs **X** and outputs **Y**
- outputs **Y** are categorical variables

- both inputs **X** and outputs **Y**
- outputs **Y** are continuous variables







Classification vs. clustering

- Clustering is discovering the inherent grouping in the data from input data only
  - somewhat similar to classification

criteria	classification	clustering
prior knowledge	yes	no
use	predict known classes for new samples	suggest grouping based on data patterns
algorithms	PLSDA, kNN, LDA, etc.	K-means, hierarchical clustering, etc.
needed data	labeled data: inputs and outputs on the available classes	unlabeled input data

Machine learning problem types

Basic machine-learning problem types

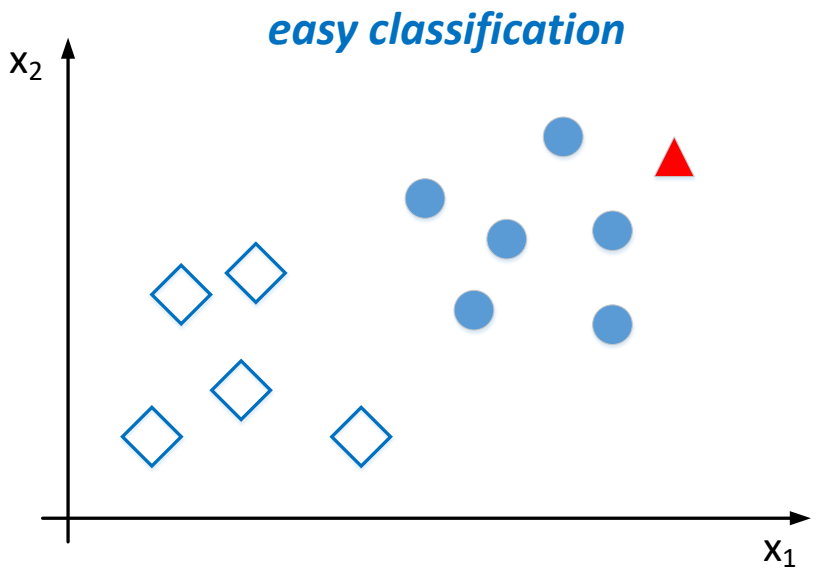
- Objective: exploiting **flexible and versatile tools** that, with few fundamental assumptions and (relatively simple) model structure, to face basic problems:
  1. **data exploration and mining**
     - identifying data pattern (pattern recognition)
     - similarity/dissimilarity among observations
  2. **discrimination** among group of observation
     - class attribution
  3. **estimation and prediction** (regression among two data tables $\mathbf{X}$ and $\mathbf{Y}$)

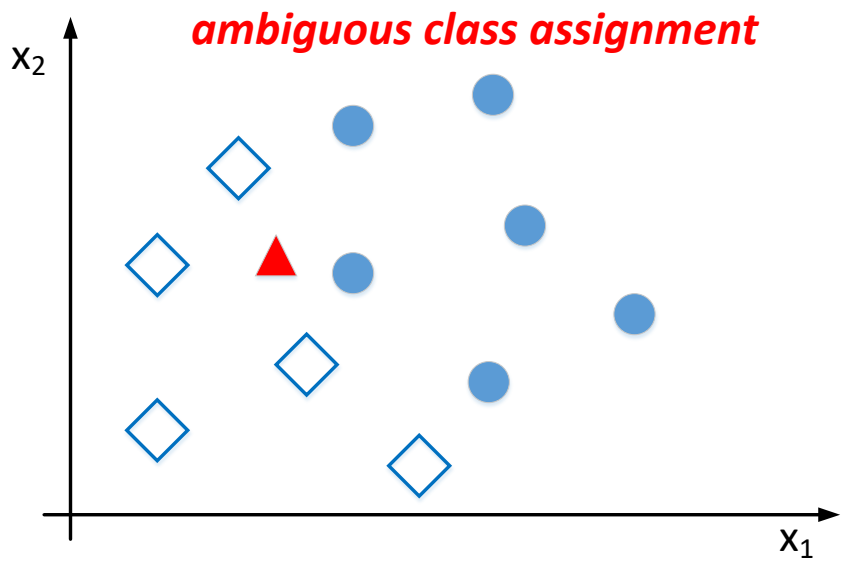
1. Data exploration and mining

- Data exploration and mining are used for general data **overview and summary**:
  - how the observation are related
  - detection of deviating observations
  - identification of different data classes (clusters)
  - understanding on the relationship between variables
  - assessing if some variables contribute in similar manner
  - etc.

2. Discrimination

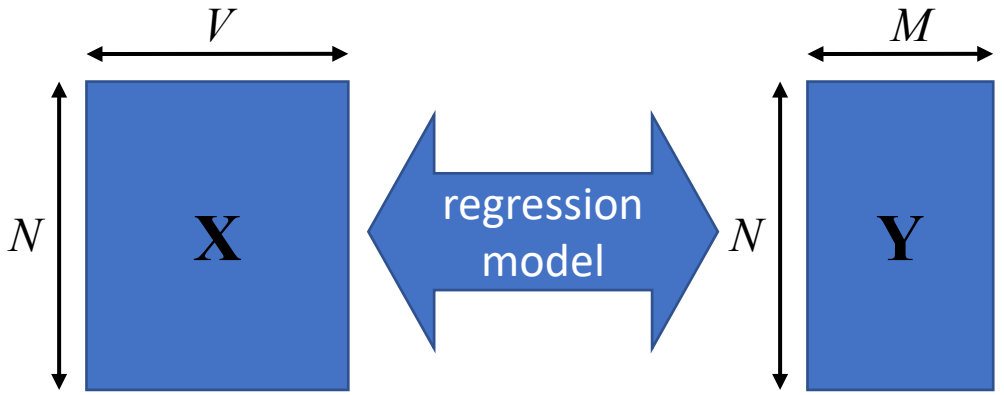
- Discrimination evaluates the **grouping among observations**:
  - once a model is built and the class model established it can be used to attribute the class to new observations
  - class assignment can be ambiguous
    - a probabilistic approach to class assignment can be utilized
  - it can be done in:
    - unsupervised manner through **clustering**
    - supervised manner through **classification**





3. Estimation and prediction

- Estimation and prediction are carried out by means of **regression models**:
  - accomplish reliable, fast and complex predictions/estimations of response variables
  - find out how predictors are quantitatively related to responses
  - give information on how factors can be used to adjust responses
- Two blocks of data are modelled
  - **predictors** (or factors): $\mathbf{X}$ $[N \times V]$
    - usually sampled frequently and at regular intervals
  - **responses** that are estimated: $\mathbf{Y}$ $[N \times M]$
    - often laborious, expensive and time-consuming measurements
    - available with low frequency



Warning on (data-based) modelling

- All the abovementioned modelling strategies are built on the available data
  - having *a lot of data* does not mean to have also *good data*
  - the effectiveness of these methodologies' performance strongly depends on the **quality of the available data**
- Who utilizes data-based models should be aware of the challenges which are related to the **quality of the available data**

… per sempre a fianco a me!