Introduction: what is Data Mining?
Course Introduction to Machine Learning - AIQDSC18

Diane Lingrand and Frédéric Precioso
diane.lingrand@univ-cotedazur.fr
https://lms.univ-cotedazur.fr/course/view.php?id=22204

UNIVERSITÉ
CÔTE D'AZUR
Master Data Science M1

2020 - 2021

From statistics...

- A few hundred individuals
- Few variables
- Strong assumptions about the statistical laws followed
- Emphasis on the calculation
- Random sample
- Current challenges : Ultra-High dimensions (few samples, thousands of variables)

..to Data mining

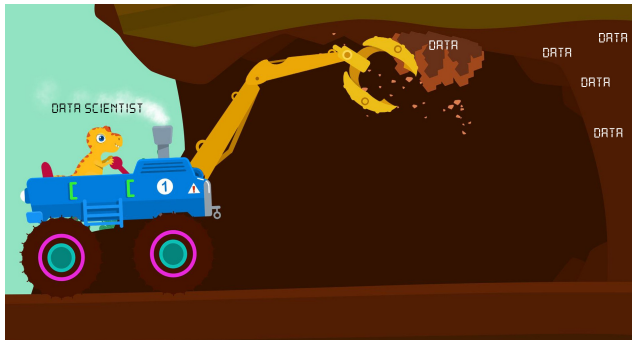
- Millions of individuals
- Hundreds to thousands of variables
- Data without prior study
- Need for quick calculations
- Not from a random sample

UNIVERSITÉ CÔTE D'AZUR

Data Mining

The data exploration or data mining is the analysis of large amounts of data in order to discover patterns and significant rules using automatic or semi-automatic methods based on statistical analysis and machine learning. (Berry et Linoff, 1997)

Data mining



- Data Mining $\neq$ Extraction of Data from the underground Data
- Data Mining $\neq$ Data Analysis
  - data mining extracts previously unknown informations from data
  - data analysis tests models and/or hypothesis on data
- Data Mining or Knowledge Discovery
  - clustering
  - anomaly detection
  - classification
  - regression
  - compact representation of data
  - dependency modeling

- **Data** form the core of the basic processes in most companies.
- **Data archiving** creates the memory of the company.
- **Data exploitation** i.e. 'data mining' creates the intelligence of the company.

The exploitation of data has become an industrial reality

- Data mining techniques exist for decades.
- The use of these techniques in the industry, however, is much more recent because :
  - data are produced electronically,
  - data are archived,
  - the necessary computing power is affordable,
  - the context is ultra-competitive,
  - many algorithms for the exploitation of data have emerged.
- In large volumes of data :
  - “Extraction of original information previously unknown, potentially useful”
  - “discovery of new correlations, trends and models”
  - “process of decision support in seeking interpretation of data models”

Fields of application of the Data mining (1)

- Business : retail, mail order
- Bank, insurance, determination of the consumer
  - Profile analysis
  - Customer segmentation
  - Analysis of the shopping basket
  - Sales prediction
  - Customer retention strategies
  - Risky customer identification
  - Fraud detection

Fields of application of the Data mining(2)

- Human resources management
  - Career plan forecast
  - Recruitment support
- Scientific activities
  - Medical diagnosis, Public health : ex, study of the genome
  - Chemical, biological and pharmaceutical analysis
  - Exploitation of astronomical data
  - Retrieval information in large volumes of multimedia data

- OLAP (Online Analytical Processing)
  - requires the user to formulate a specific question which is the subject of an adhoc query to provide a factual result : "How many shoes size 42 I sold these last three months ?"
  - Only counts but do not forecasts. The result is used to validate a hypothesis or provides information for an assessment.
  - focuses usually on current facts, uses predefined aggregate data, the establishment of factual results by adhoc queries finalized in the form of reports.

- The Data Mining
  - From a set of data, exploration techniques are applied to find relationships, often complex, and unknown models that have a meaning : “How many summer shoes of size 42, I should order for next season ?“
  - generally focuses on trends, estimates, discoveries or predictions, requires detailed data, implement statistical techniques, algorithms, and establishes models (express or implied, complete or partial).

UNIVERSITÉ CÔTE D'AZUR

Definition by Bill Inmon

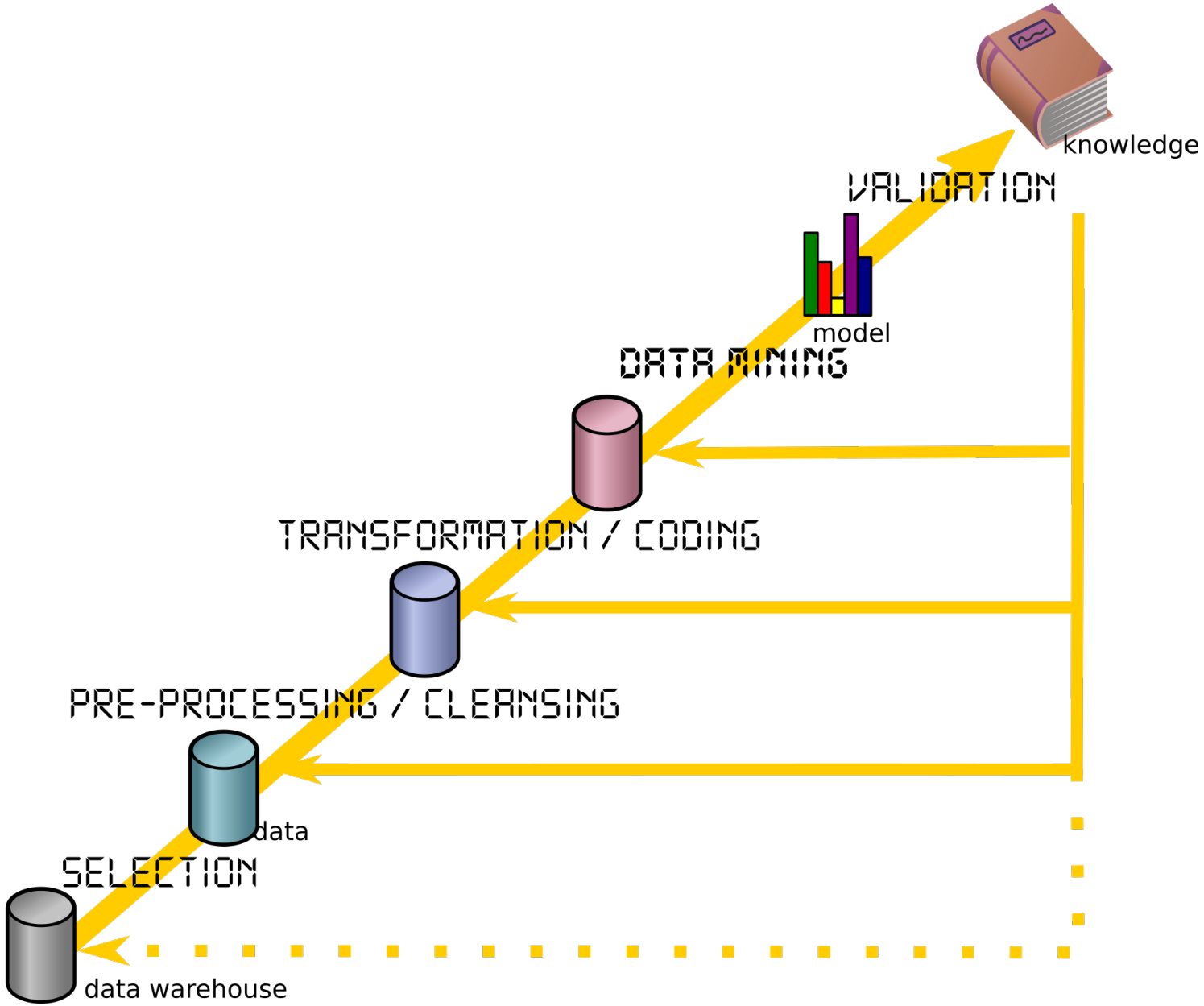
A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process.

- database in which are stored, after cleansing and homogenization, information from different systems of the company.
- The warehouse facilitates data mining but data mining can be done also on data extracted for the occasion.

Data mining as part of a process of knowledge discovery

1. Set down the problem
2. Data selection
3. Data cleansing
4. Data coding, actions/transformations on variables
5. Model, knowledge, or information retrieval : **Data mining**
6. Validation and interpretation of the result, with possible return on the previous steps
7. Knowledge Integration

Data mining as part of a process of knowledge discovery



1. Set down the problem

- It is understanding the scope, the already existing knowledge and the goals of the end user.
- It is defining what type of problem to process ? We know the classes, we want to identify the factors of assignment, or we want to create the classes differentiating factors.
- If one highlights many client groups, in a marketing survey can we review the process of marketing for each group ?

1. Set down the problem : an example

- A Bookstore sells 5 kinds of magazines : sport, car, house, music, and comics.
- Questions are :
  - 1. How many people have taken a subscription for a sport magazine this year ?
  - 2. Did we sold more subscriptions for sport magazines this year than last year ?
  - 3. Are comics magazine buyers also sports fans ?
  - 4. What are the main characteristics of the readers of car magazines ?
  - 5. Can we predict the loss of customers and provide measures to reduce them ?

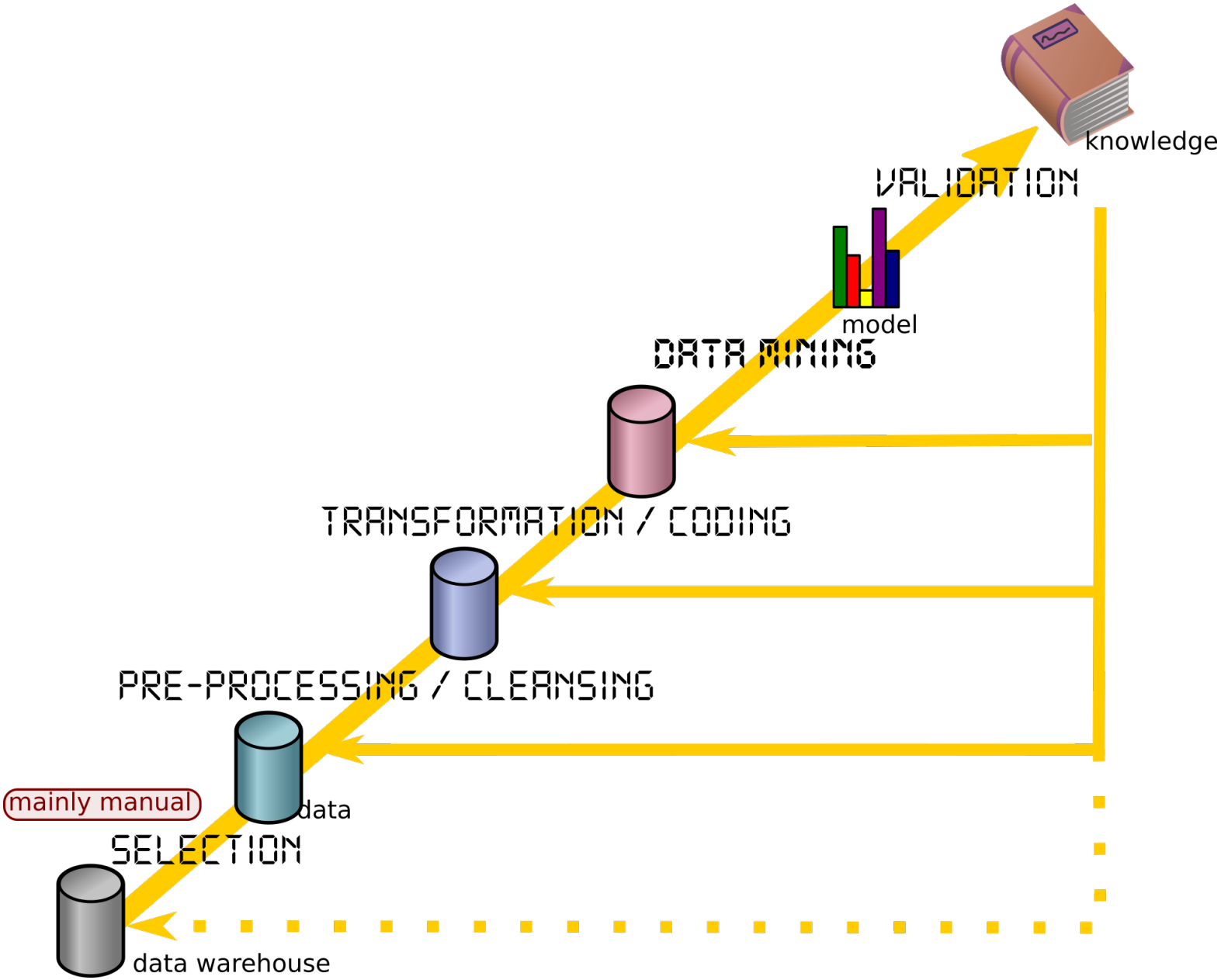
1. Set down the problem : an example

- A Bookstore sells 5 kinds of magazines : sport, car, house, music, and comics.
- Questions are :
  - 1. How many people have taken a subscription for a sport magazine this year ?
  - 2. Did we sold more subscriptions for sport magazines this year than last year ?
  - 3. Are comics magazine buyers also sports fans ?
  - 4. What are the main characteristics of the readers of car magazines ?
  - 5. Can we predict the loss of customers and provide measures to reduce them ?
- 1 and 2 are simple queries. In 2, there is the concept of time therefore data should get historical features. For 3, the answer might be worth considering the proba rule to be true. 3 can be generalized : we could search for frequent associations between buyers of magazine. 4 is more open, so as 5, it is really the field of data mining

2. Data selection

- Existing data or data to be built
  - data warehouse (Data Warehouse), store data, relational databases, temporal databases, Web...
- Sample or work on all data : depends on the available data, of the power machine, the desired reliability. Very often work on a sample is well suited to the data mining as it is an iterative process.

Data Mining Workflow



3. Data cleansing/preprocessing

- Duplicates, data entry errors, failures of sensors...
- Outliers : search peaks, the values outside a space determined by the mean and a number of standard deviations, visualization tools : histograms, scatterplots...
- Missing information : exclude incomplete records, replace missing data (average value, by default), keep the incomplete records if the search method can manage them.

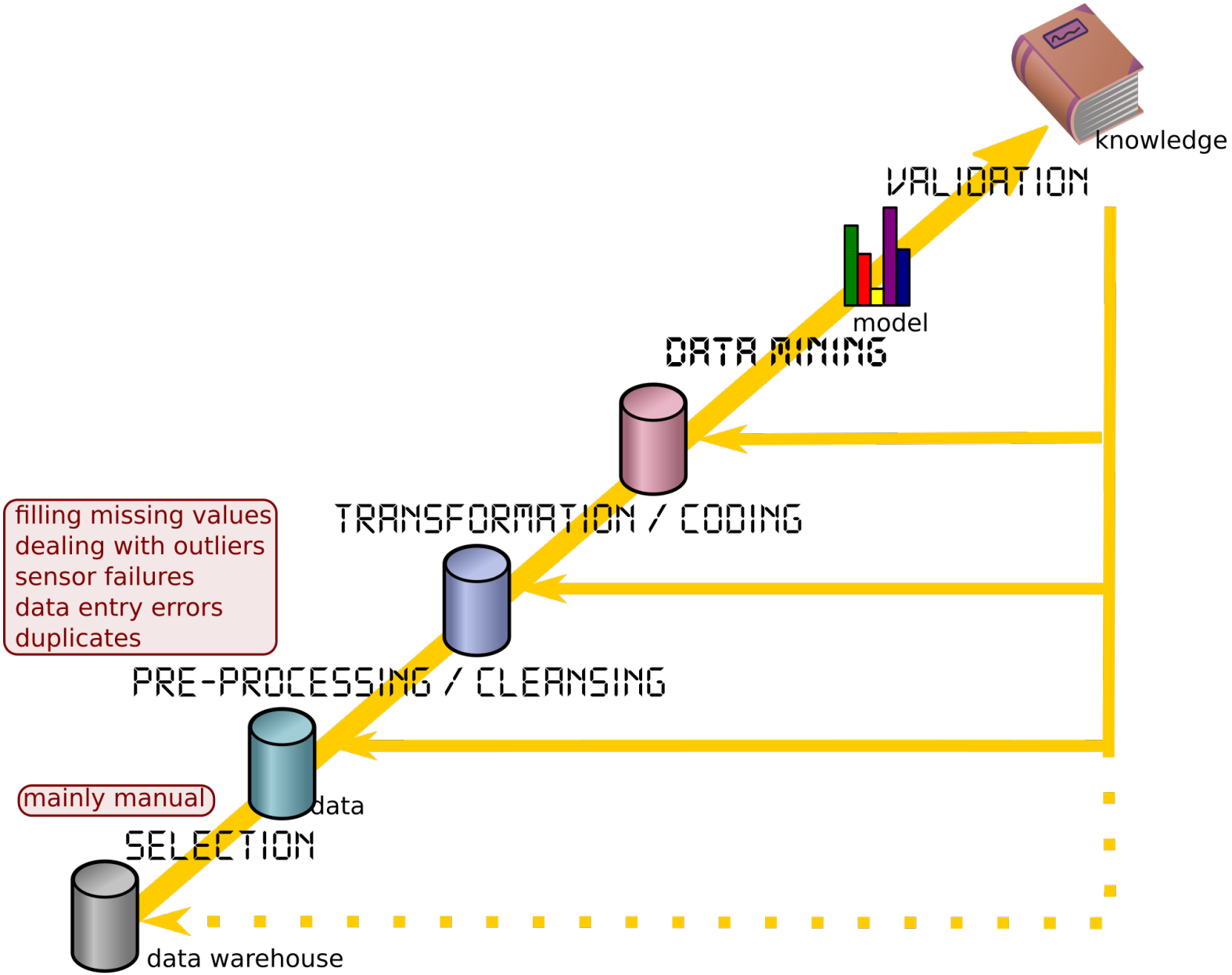
Missing values

- One can :
  - ignore the observation
  - use the average value (the worst !)
  - use the average value for samples of a same class
  - use the regression (more accurate but more complex)

Outliers

- A strategy should be defined to process outliers (data out of norm) or possibly to develop a model based on these values :
  - for example, if the objective is to predict attendance rates and revenues for sporting events, it must certainly eliminate the numbers of abnormal dating due to special events, strike, etc...
  - Instead in the case of fraud detection, it may be relevant to focus on certain outliers because they are perhaps the representation of fraudulent transactions.

Data Mining Workflow



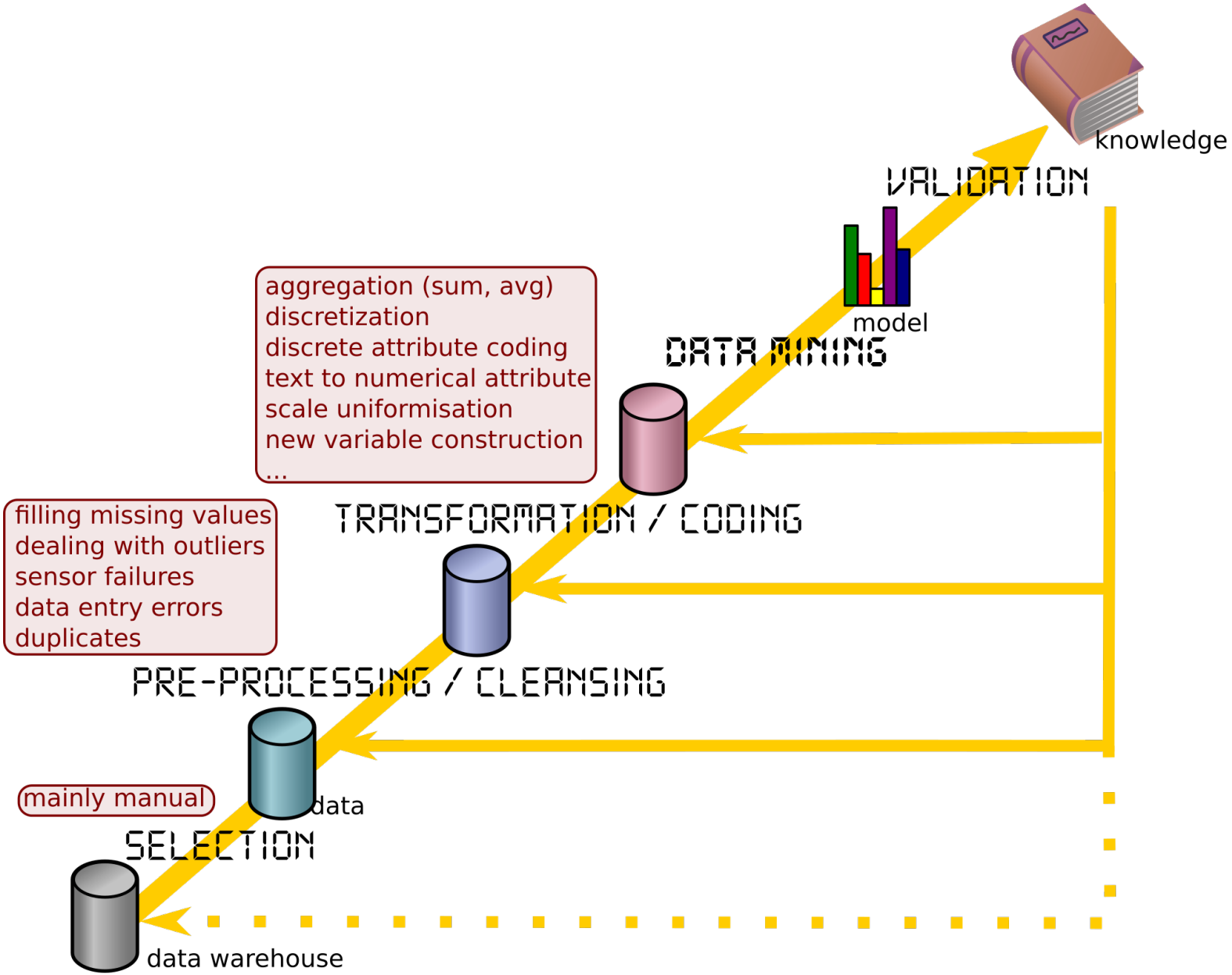
4. Data coding/transformation

- Aggregation (sum, average)
- Discretisation (reducing the number of values of a continuous variable by dividing the domain of values in intervals)
- Discrete attribute coding
- Scale uniformisation or standardisation
- New variables construction
- This main step of choosing good variables can be decisive for the success of the mining process.

Examples of actions on variables

- Transformation of variables :
  - Transformation of geographic data : City postal code into geographic data (longitude, latitude) to take into account the proximity of places in the reasoning (use geocoding in geomarketing)
  - Transformation of dates in durations : seniority of a client, duration between sending a catalogue and the 1st order
- Multi-variable transformations :
  - combine several variables in a new aggregated one, linear or non-linear multivariate combination : income and number of children combined into income/number of children
- Multimedia data :
  - images : better coding than a vector of pixels values
  - videos
  - sounds
  - text

Data Mining Workflow



- Algorithms of inspirations ...
  - Mathematics : statistics and Data analysis
  - Computational
    - Clustering
    - Decision tree
    - Association rules
    - Dynamic programming
    - Support Vector Machines (SVM)
  - Biological
    - Artificial neural networks (ANN)
    - Evolutionary algorithms

Supervised vs Unsupervised

- The classification, the logistic regression are supervised tasks
  - Predictive data mining (there is a dependent variable to predict or estimate denoted typically by Y).
- Clustering, searching for association rules are unsupervised tasks
  - Explanatory data mining (one looks more for explaining the relationships between the variables without having a dependent variable).

- Algorithms :
  - Unsupervised - Learning *a priori* in discovery mode
    - Clustering
    - Evolutionary algorithms
    - Association rules
  - Supervised - Learning *a posteriori* in Recognition/prediction mode
    - Artificial Neural Networks
    - Support Vector Machines
    - Decision tree
    - Dynamic programming

5. Data mining

- Classification
- Estimation
- Association search
- Clustering

- Assign an object to a class with respect to its features $A_1, ...A_n$
- Example
  - Determine if a message is a mail SPAM or not (2 classes)
  - Assign a web page in a theme category : Economy, Sports, Politics,... (multi-classes)
  - Diagnostic : risk of cerebro vascular accident or not (2 classes)

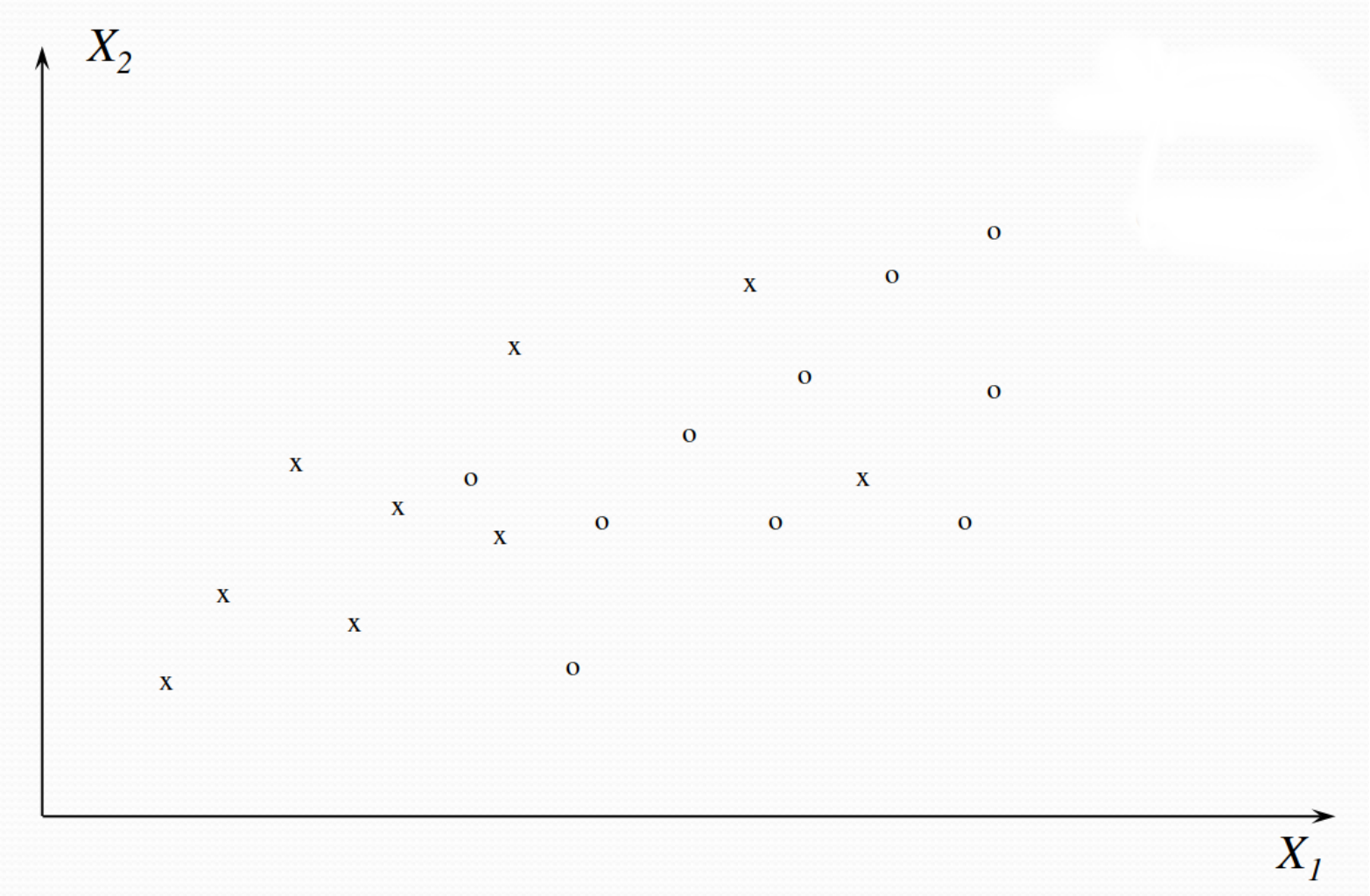
- If no knowledge *a priori* to define the class according to $A_1, ..., A_n$ then we study a set of examples for which we know $A_1, ..., A_n$, and the associated class and we build a model

$$\text{Class} = f(A_1, ..., A_n)$$

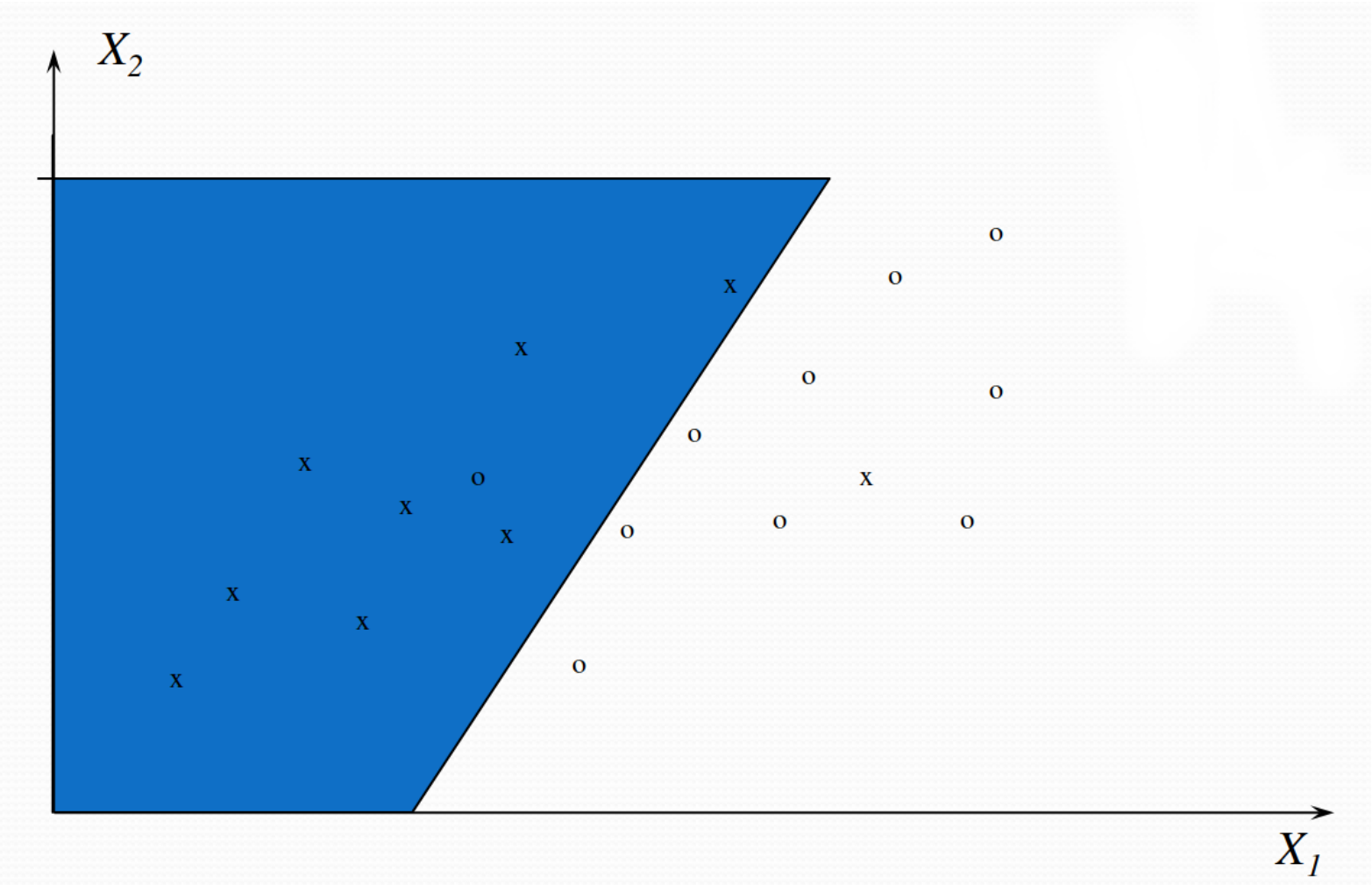
  - Discriminante analysis
  - Classification tree
  - Kernel machine

An example with 2 classes

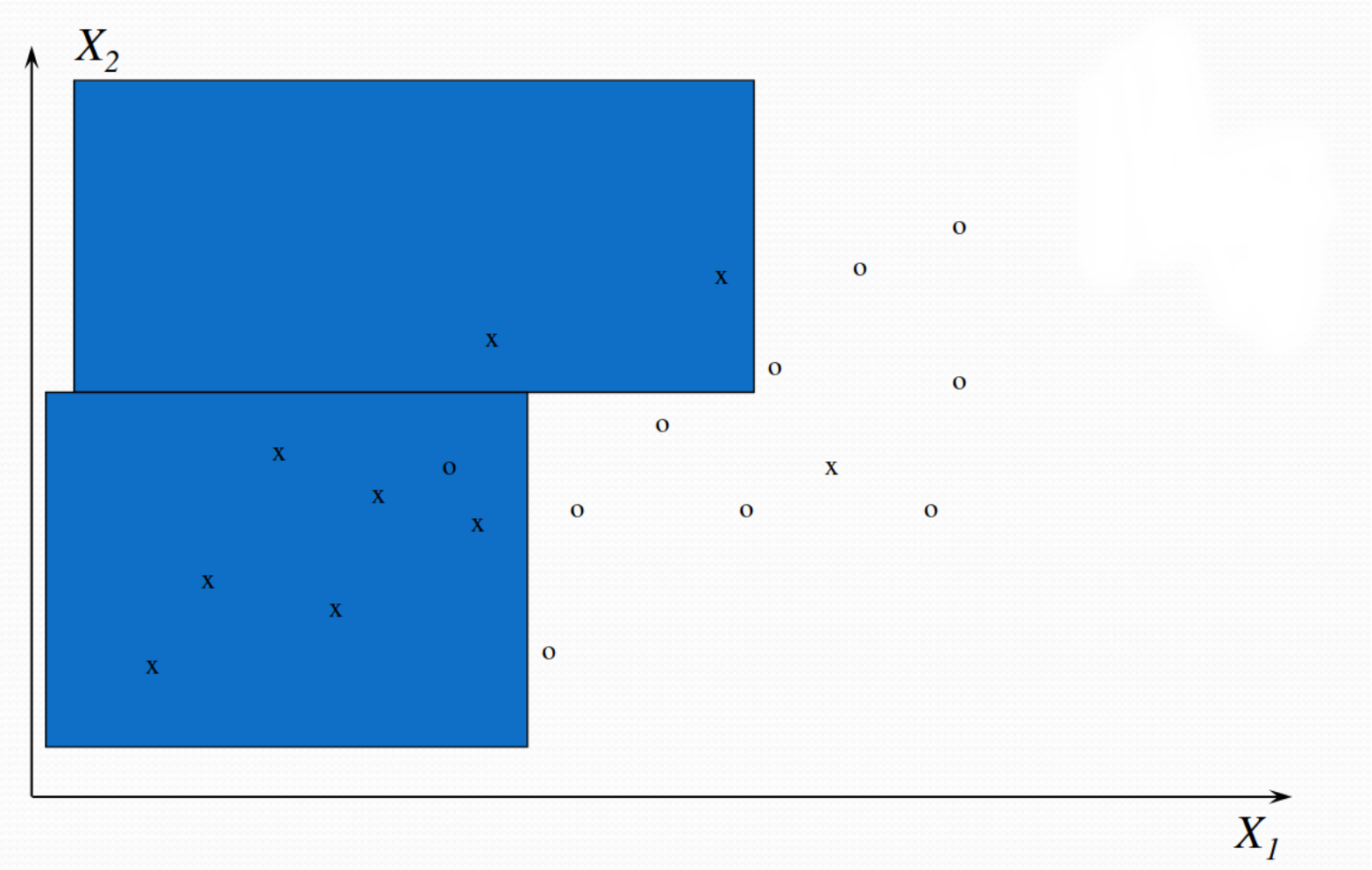
The classification is learning a function which allows to assign a new individual to a class or another.



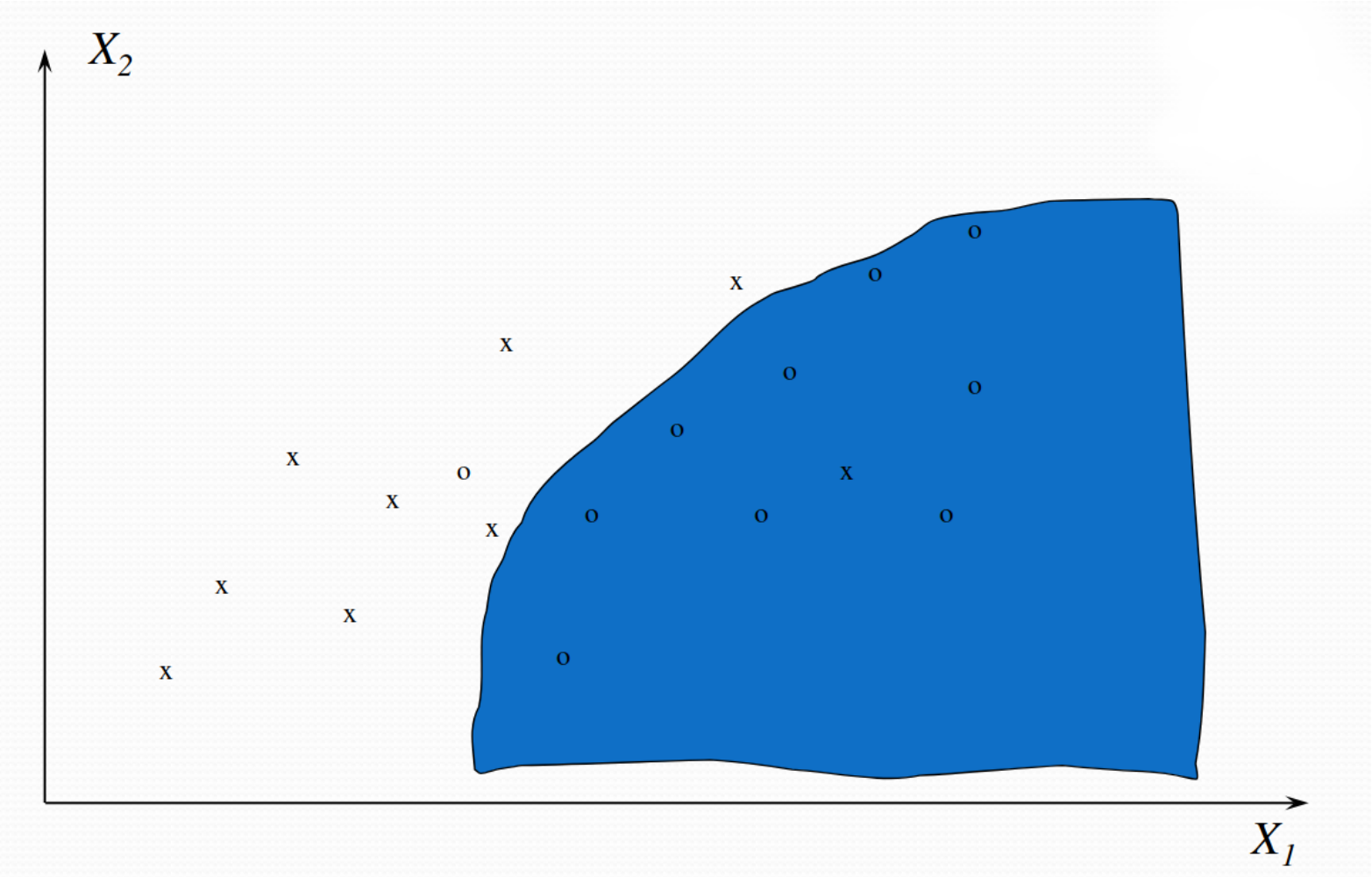
Classification with discriminante analysis



Classification with decision tree



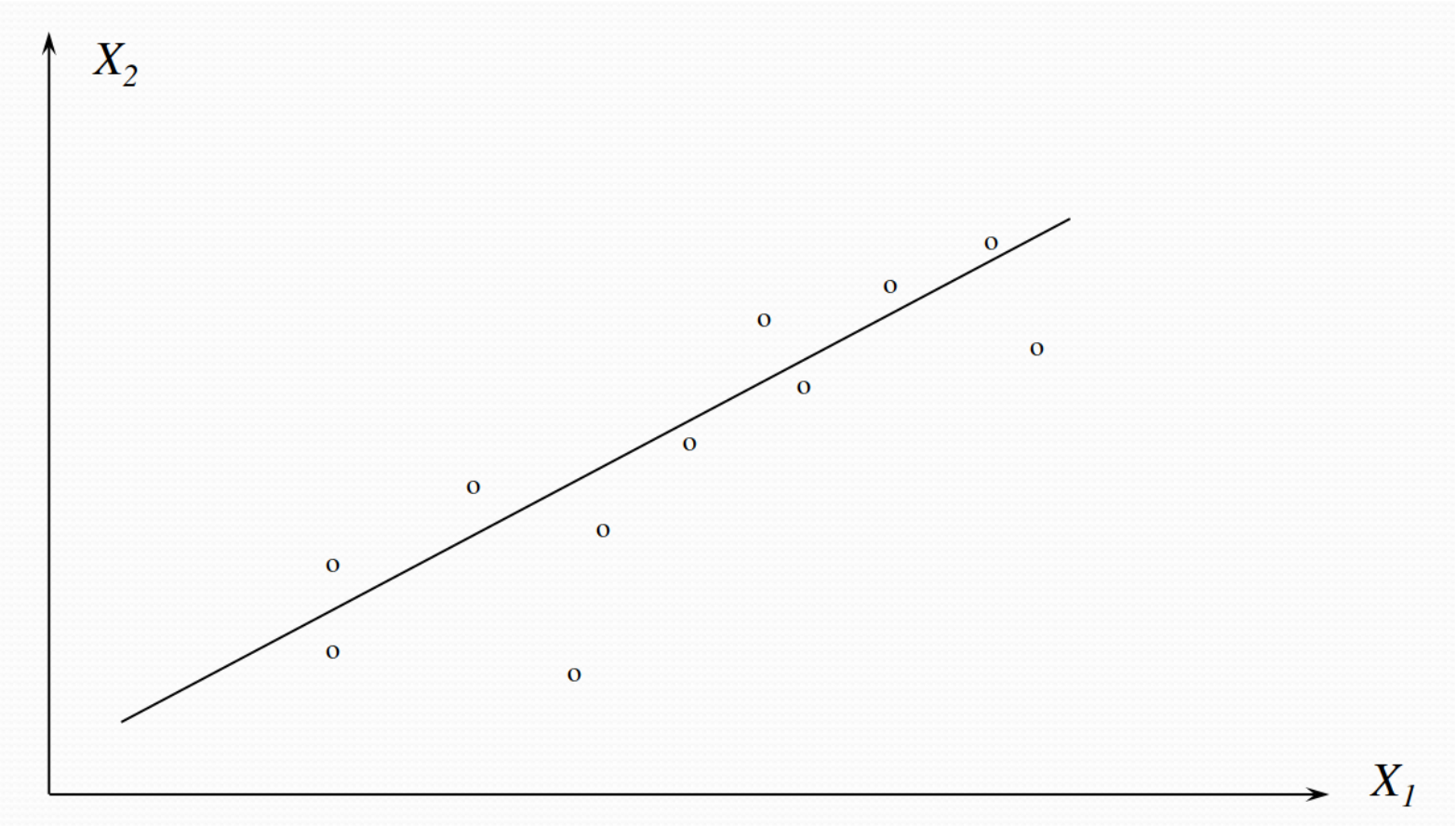
Classification with kernel machines

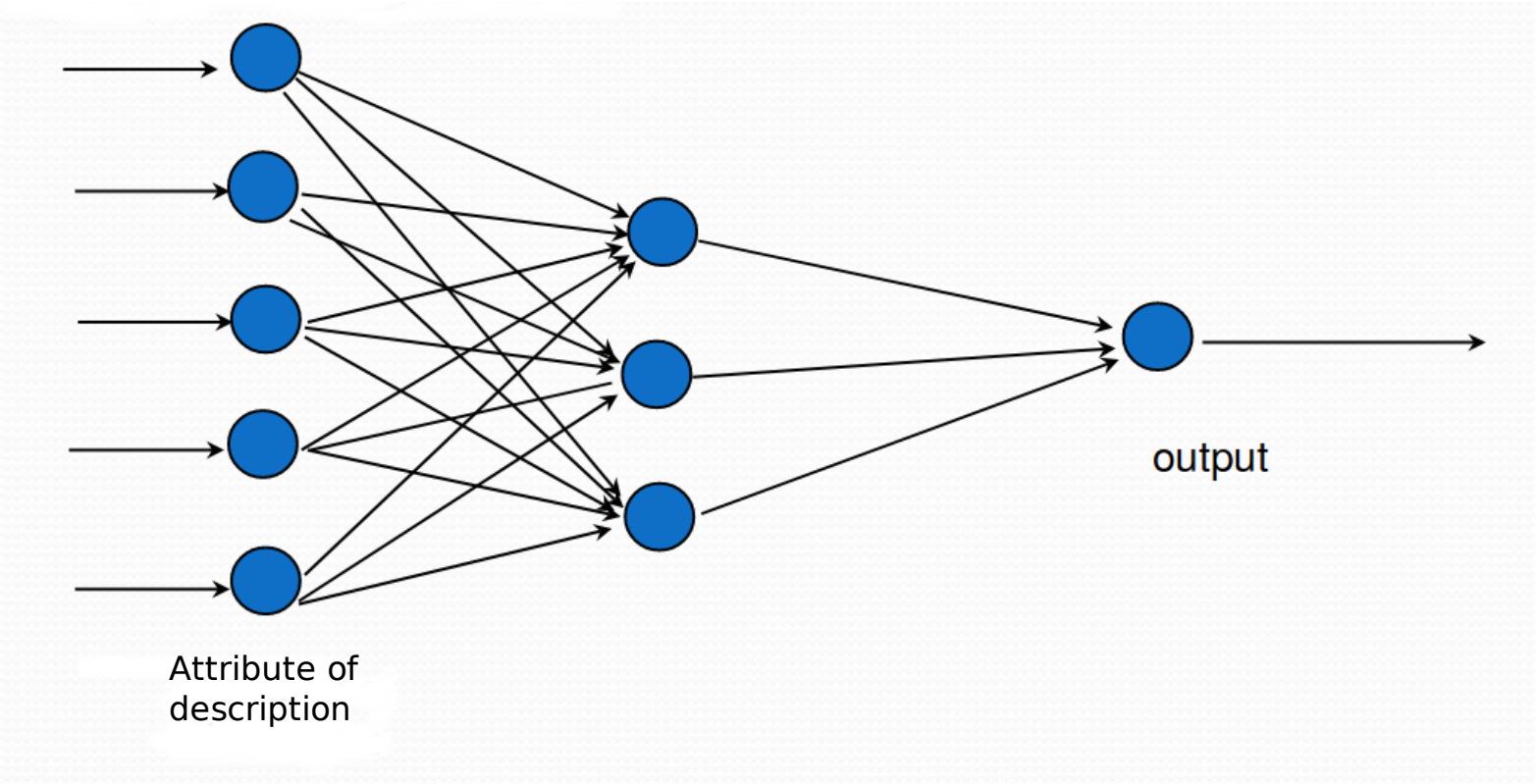


- Estimate (predict) the value of a variable with continuous values from the values of other attributes
  - Regression
  - Kernel machine : ANN, SVM

Simple linear regression

The regression explains the variations of a variable through a function of other variables : here X_2 is represented as a function of X_1. The result is poor because there is little or no correlation.





- The input layer corresponds to the initial data, the output layer to the result.
- It is a non-linear system.
- The learning will adjust the weight on the connexions but the architecture and the number of neurons on the hidden layer is arbitrary.

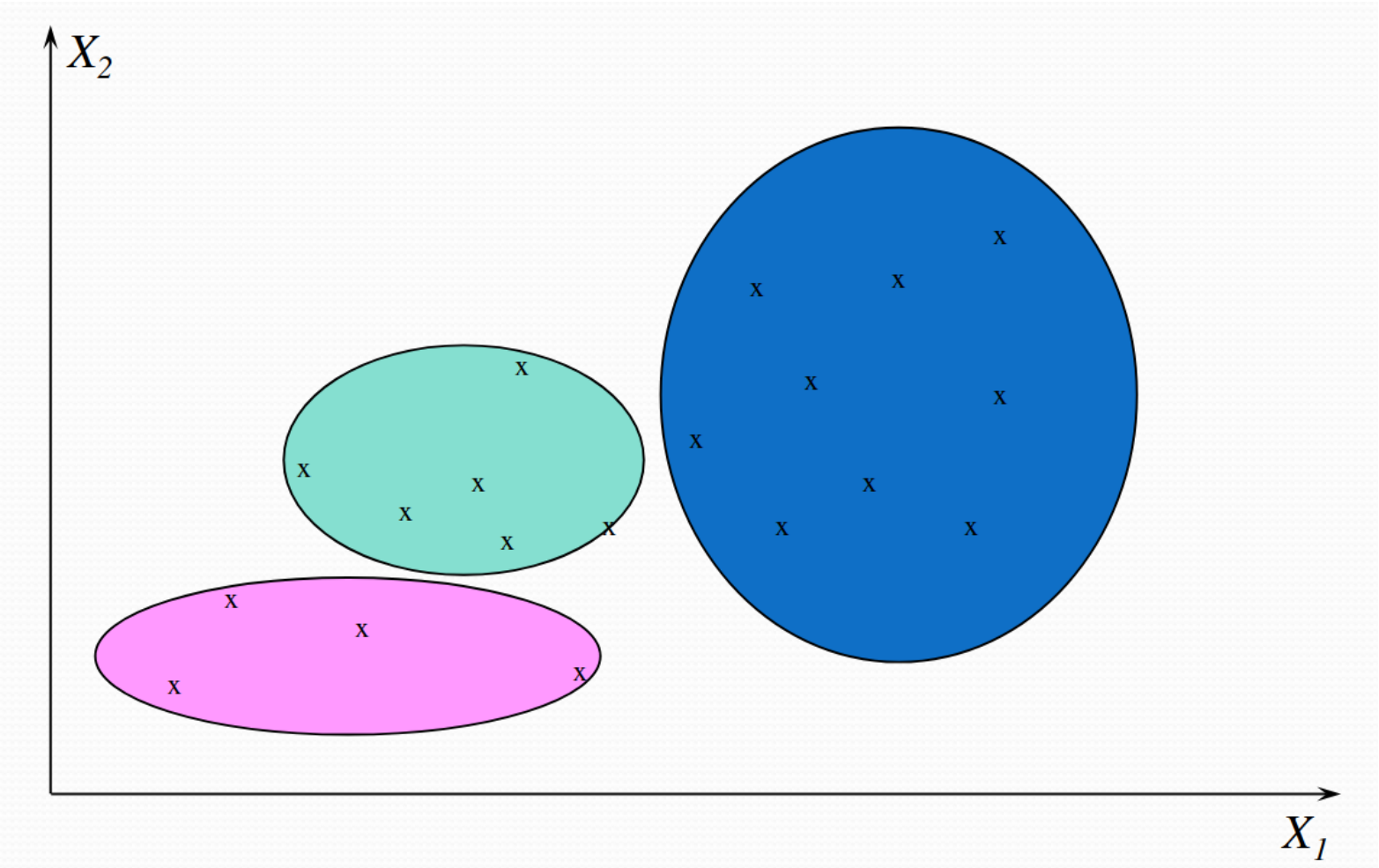
- Association rules : for instance analyse of shopping basket
  - ≪Satursday, consumers often buy together milk and eggs. ≫
  - Is there any causal relationships between buying the product P and another product P' ?
  - the beer and diapers legend. From `http://www.dssresources.com/newsletters/66.php` Blischok said "Yes, if we go back to the legend, we did discover that between 5 :00 and 7 :00 p.m. that consumers bought beer and diapers. This was an insight that the retailer had never seen before, and the fact that we discovered this affinity was not the real transformational event that occurred. What this showed Osco in this early pioneering effort was that it was possible to redesign the store based on consumer preferences at the center of all decisions. Their management team got it. They simply understood that they had the opportunity to change. Well, in reality they never did anything with beer and diapers relationships."

Data mining : Segmentation / partitioning (clustering)

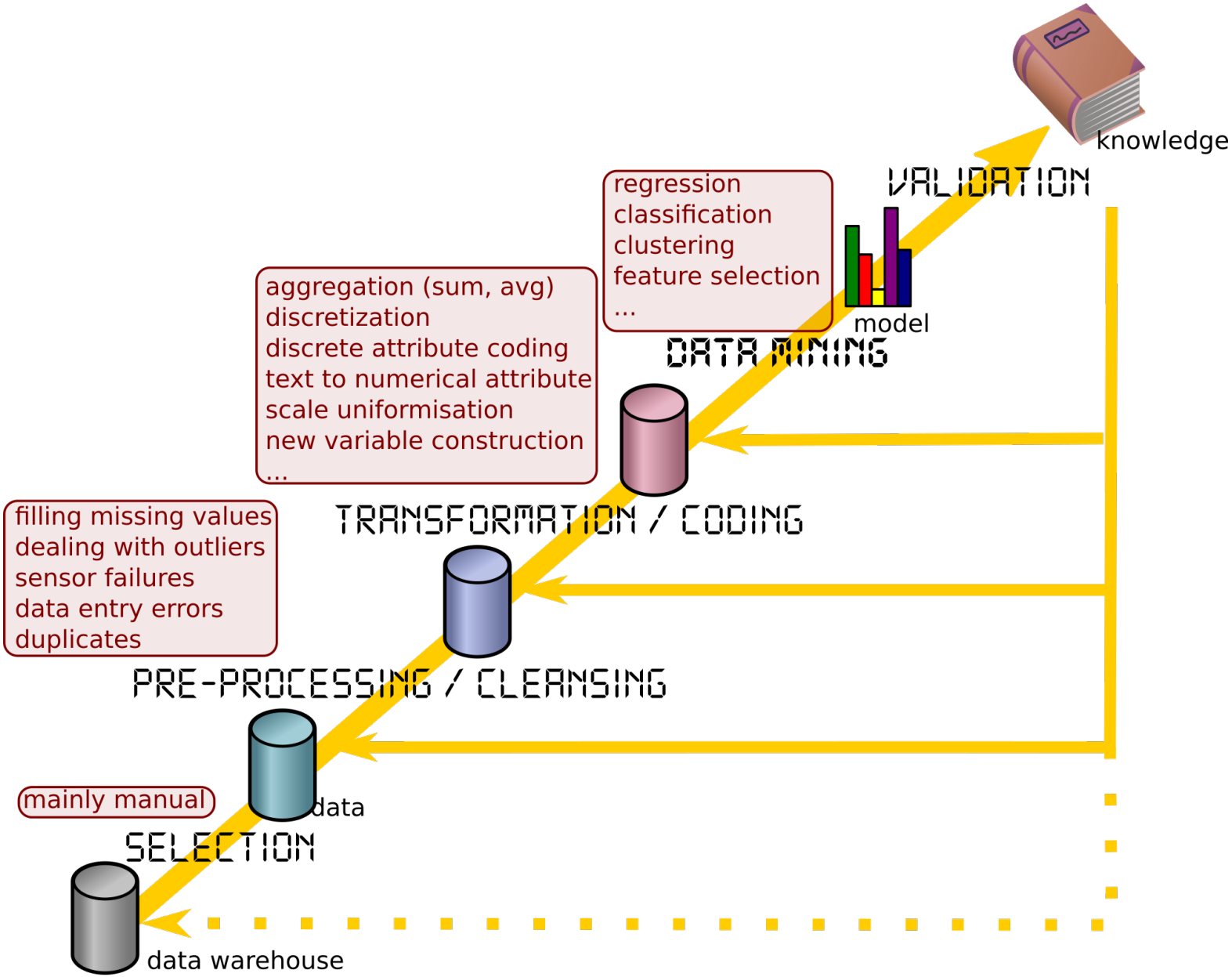
UNIVERSITÉ CÔTE D'AZUR

- Unsupervised learning : data are not classified, we isolate subgroups of records similar to each other
- Once the clusters are detected, we can apply modeling techniques on each cluster

No assignment to a known class at the start : it brings together individuals by their proximity into classes which may overlap.



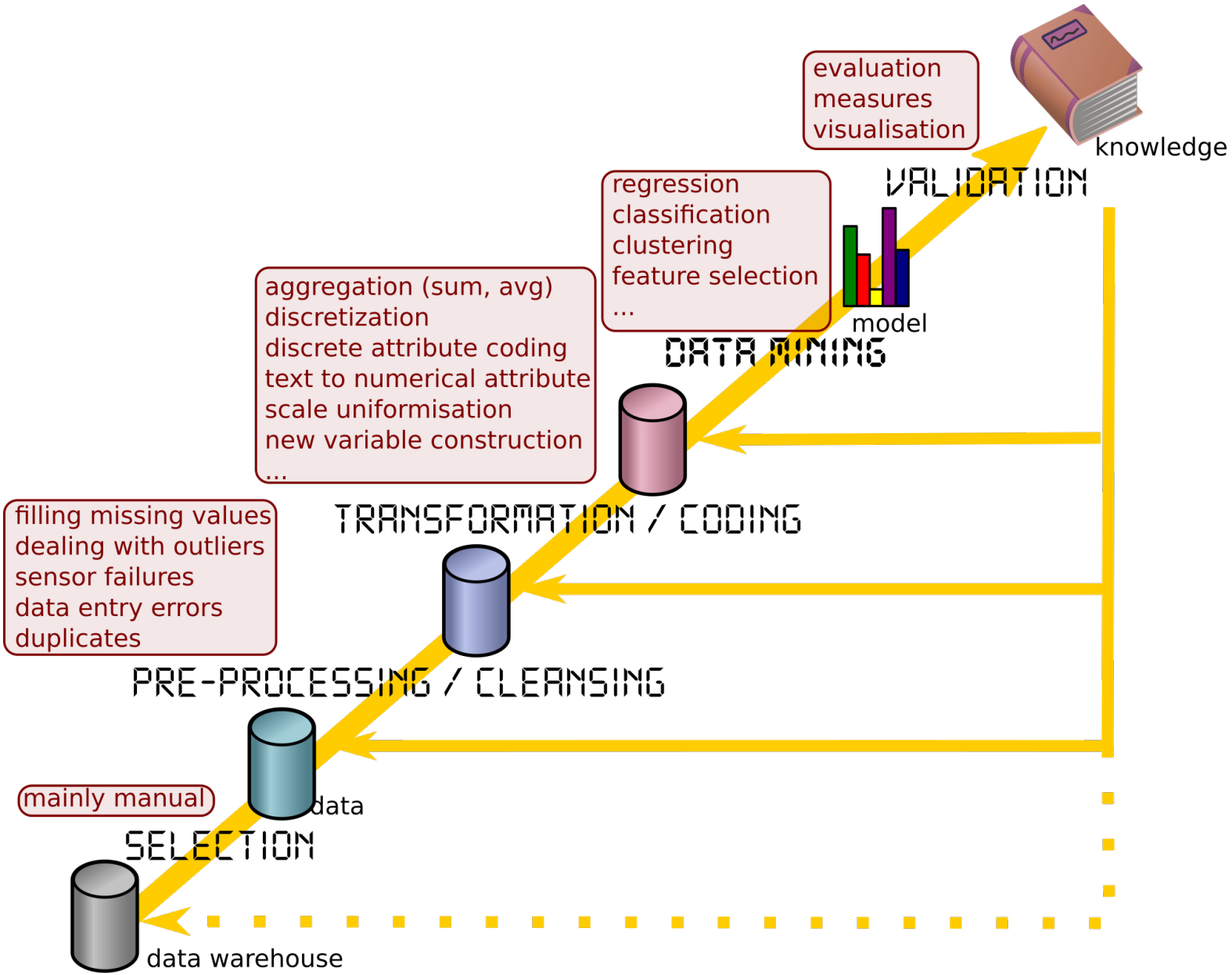
Data Mining Workflow



6. Validation in the supervised case

- Validation on test set
  - Data set is splitted into Training set and Test set
  - Construction of a model on the training set and test of the model on the test set for which the results are known
- Quantitative assessment (do not forget the confidence intervals)

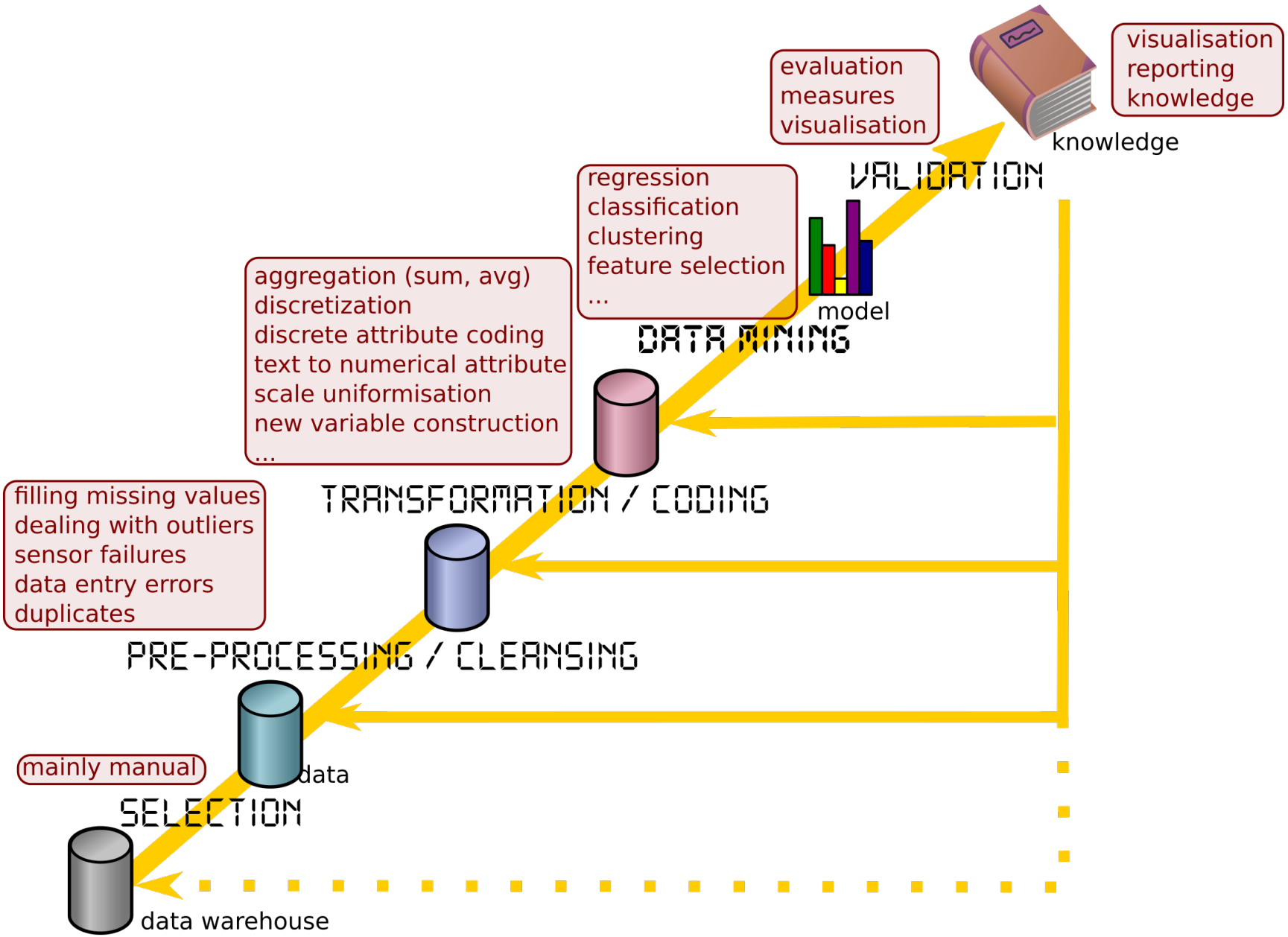
Data Mining Workflow



7. Integration of knowledge

- Decision-making with the extracted knowledge
- **Business experts are essential to give meaning to the information extracted !**

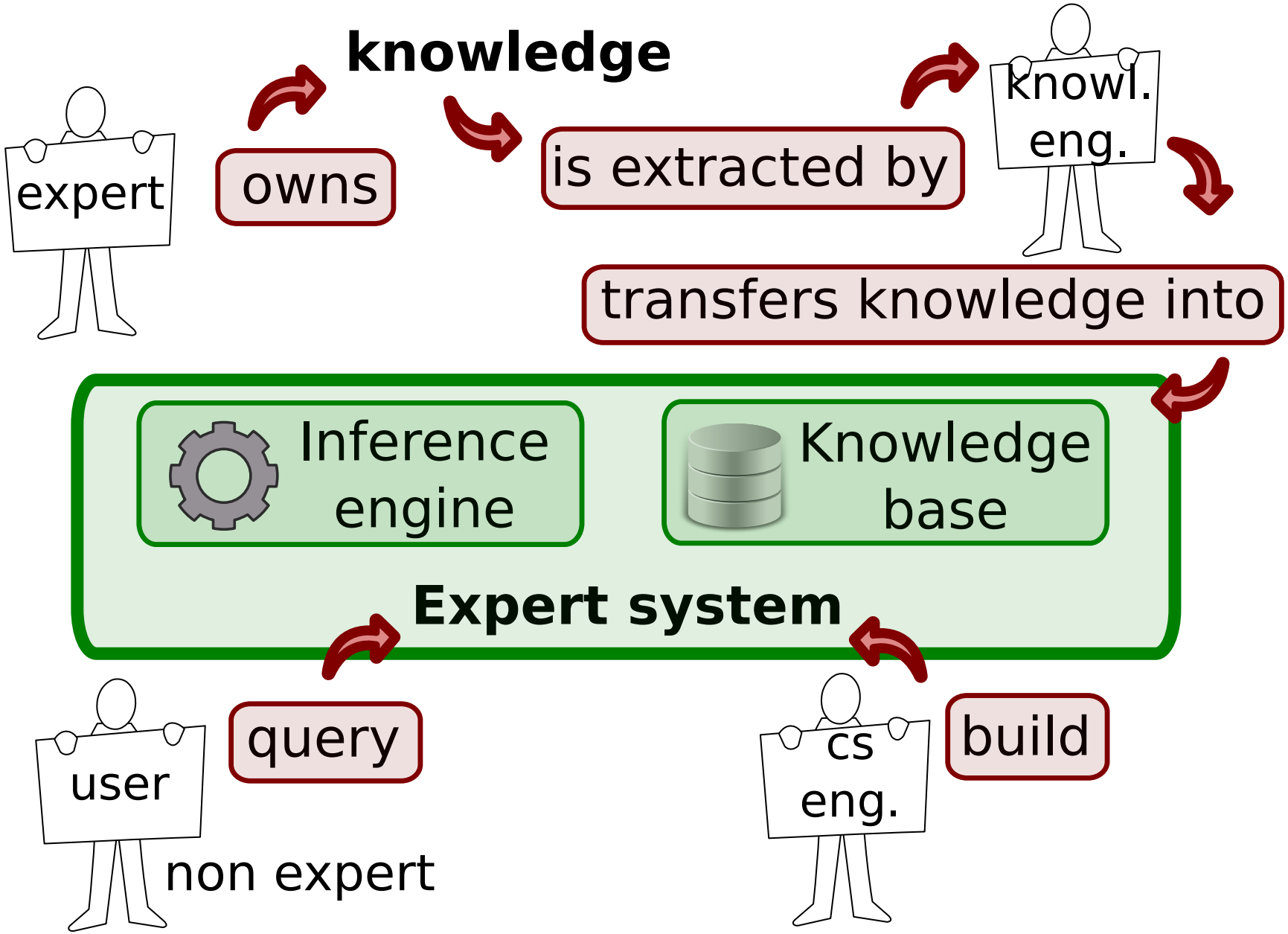
Data Mining Workflow



Five myths about Data Mining

1. Data mining could instantly prevent the future, like a crystalball.
2. Data mining is not yet viable for professional applications.
3. Data mining requires a separate and dedicated database.
4. It would be polytechnician to make of data mining (a Polytech'Nician will be just fine).
5. Data mining would be reserved for large companies that have a large volume of data client.

Expert system



Data mining vs Expert Systems

- Deduction : basis of expert systems
  - logical schema allowing a theorem from axioms
  - the result is correct and valid, but the method requires the knowledge of rules (axioms)
- Induction : basis of data mining
  - method to draw conclusions from a series of facts
  - generalization "somewhat abusive"
  - confidence indicators allowing the weighting of the decision

Data mining vs Expert Systems

- A large International Bank wanted to control the costs generated by clients when they used ATMs of other banks :
  - What constitutes an excessive use by the clients of the ATMs of the contenders ?
  - Who are the clients that generate the excessive costs by the use of the ATMs of the contenders ?
  - What is the value these clients represent to our Bank ?
  - What should we pay attention when we use these results ?

Data mining vs Expert Systems

- 1. The first discovery made by the laboratory in Data Mining of Teradata (`http://www.teradata.com/`) was that 10% of the clients of the Bank generated 90% of the cost of competing ATM. This finding could perhaps have been made using traditional statistical techniques. However, thanks to data mining, we were able to highlight that 10% of clients who generate extra costs, 80% were low-value clients.
- 2. Several data mining techniques accounting for multiple variables allowed to understand the potential value of each of the low-value clients and quantify the concept of "excessive use".

Data mining vs Expert Systems

- 3. Data mining also allowed to respond to the next logical question that should be asked by the Bank : "Should we review service to these 80% of low-value clients ?
  "Analysis of behavioral patterns revealed that approximately 30% of these low-value clients were high-potential clients, namely : students. It certainly wasn't a startling discovery but it would have been difficult to achieve this without the techniques of analyzing multiple variables and detailed data.
- 4. The most interesting discovery was that a competitor targeted University campuses in order to install new ATMs. No other Bank developed actions on campus and it enjoyed almost exclusive presence.
- $\Rightarrow$ Through this experience in data mining, the Bank could meet its initial questions. But what is even more important, is that they were able to discover the strategy of a competitor.

Practical today

- python3 installations :
  - install numpy
  - install scikit-learn
  - install tensorflow and keras
- play with some data
  - load the Iris dataset
    - how many data ?
    - what dimension ?
    - how to visualise the data ?
  - load the MNIST dataset
  - load the FMNIST dataset

Determine species using Iris dataset

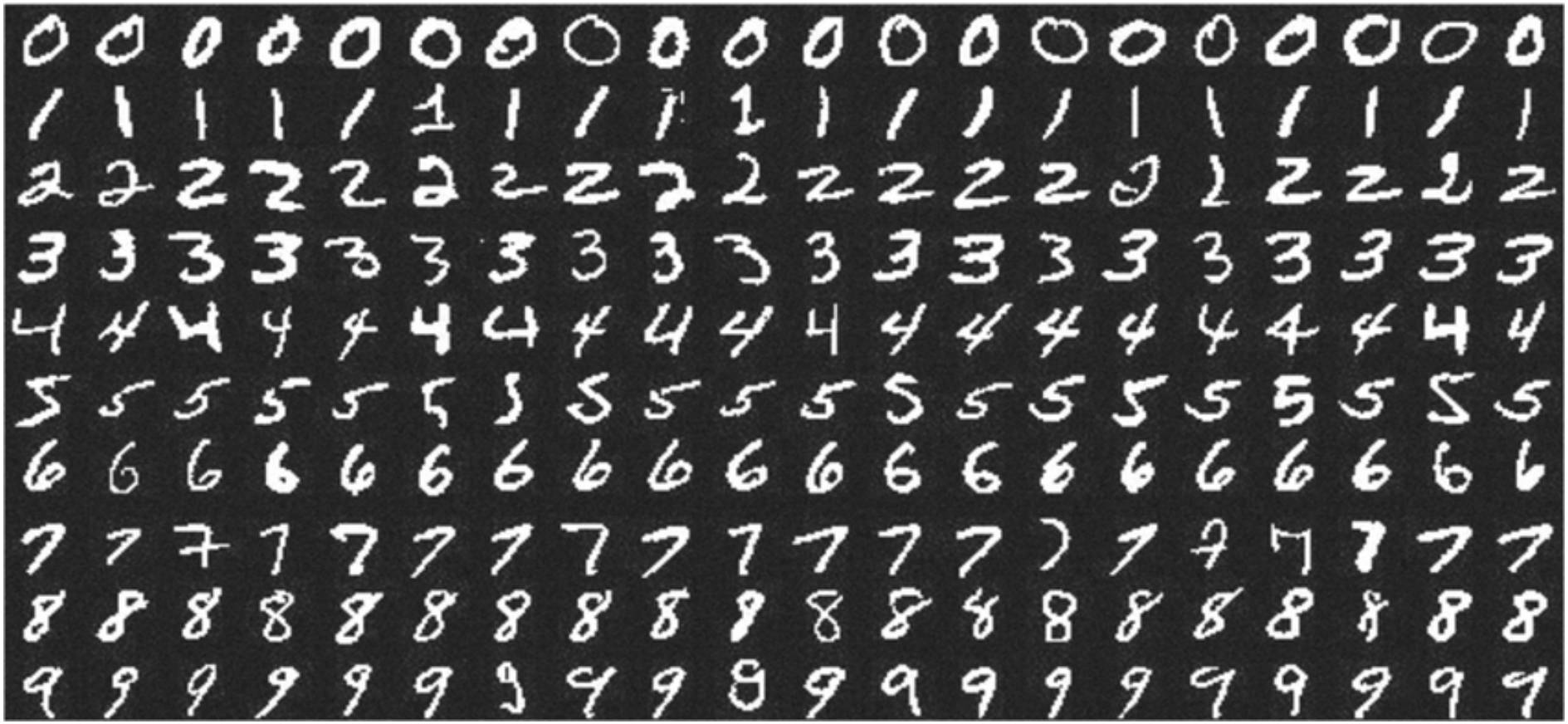


`https://archive.ics.uci.edu/ml/datasets/iris`

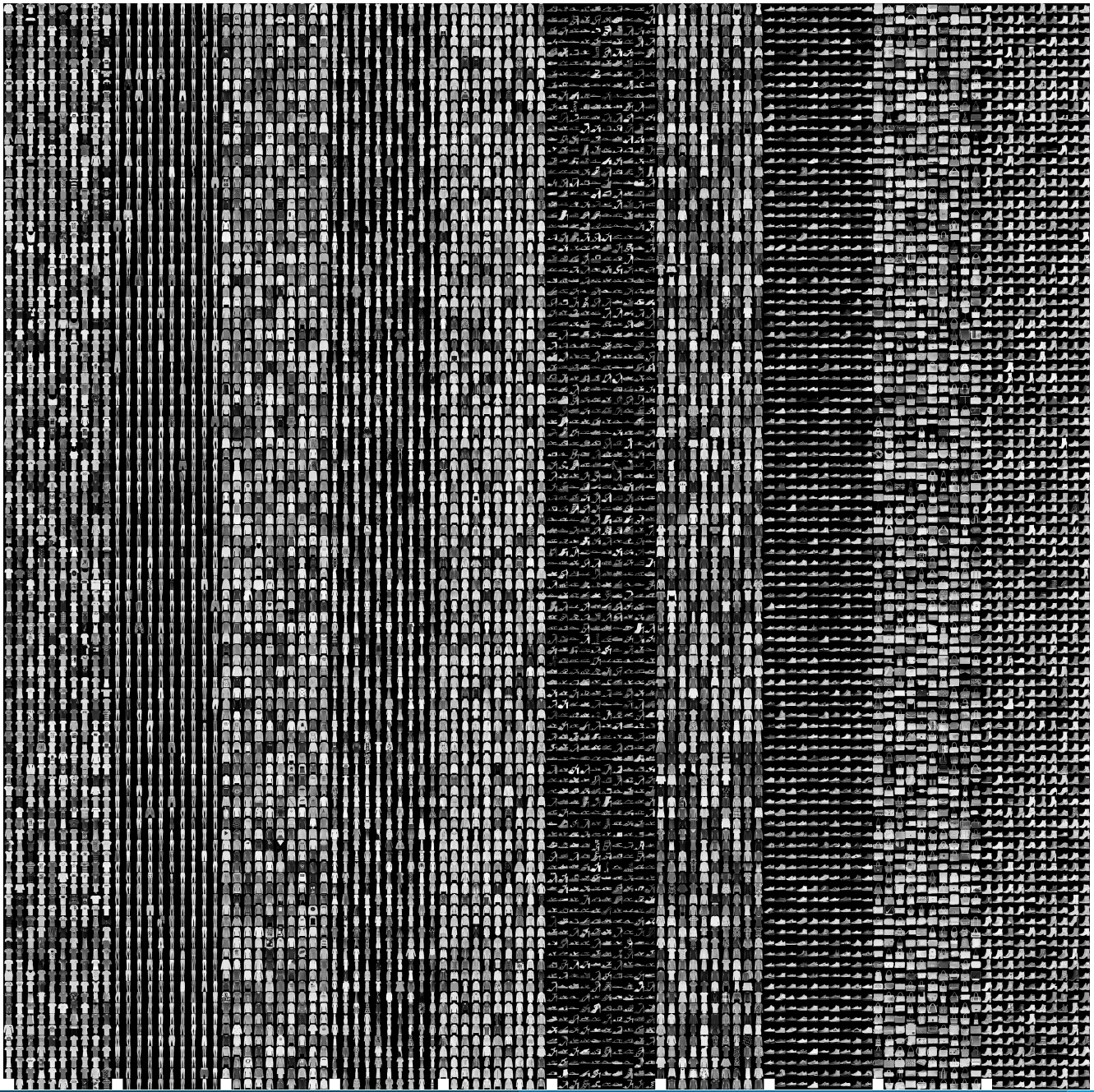
- 150 data
- 4d data :
  - sepal length and width in cm
  - petal length and width in cm
- classes :
  - Iris Setosa
  - Iris Versicolour
  - Iris Virginica

Classify handwritten digits using MNIST dataset

Dataset of 60000+10000 grayscale squared images (28x28).



Fashion MNIST



Dataset of 60000+10000 grayscale squared images (28x28).

Label	Description
0	T-shirt/top
1	Trouser
2	Pullover
3	Dress
4	Coat
5	Sandal
6	Shirt
7	Sneaker
8	Bag
9	Ankle boot