



Open Universiteit

Online workshop Data Analytics

Lyana Curier, Daniele Di Mitri, Martine Hermans



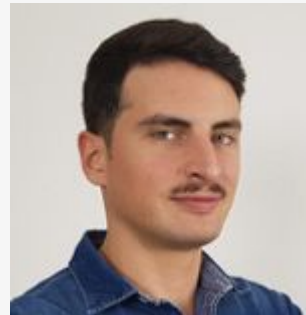
CAROU

**Center for
Actionable
Research of the
Open
University**

Who we are?



Lyana Curier
lyana.curier@gmail.com



Daniele DI MITRI
daniele.dimitri@ou.nl



Martine HERMANS
martine.hermans@ou.nl



Agenda





Introduction

data rich, data driven world.

You've probably heard of kilobytes, megabytes, gigabytes, or even terabytes

By 2025, it's estimated that 463 exabytes of data will be created each day globally – that's the equivalent of 212,765,957 DVDs per day!





The science of analyzing raw data in order to draw conclusions.



Techniques and processes are automated into mechanical processes and algorithms to prepare raw data for human consumption.

descriptive
diagnostics
predictive
prescriptive

DATA ANALYTICS



Companies large and small, in traditional and non-traditional sectors, are using data to optimize their performance.



BANKING AND SECURITIES

Helps in reducing Fraudulent Transactions



COMMUNICATIONS & MEDIA

For simultaneous real time reports of several Platforms



MANUFACTURING

To enhance Supply Chain Management



TRANSPORTATION

For better managing traffic plan and logistics



HEALTHCARE

To collect public health report and identify global spread of various viruses.



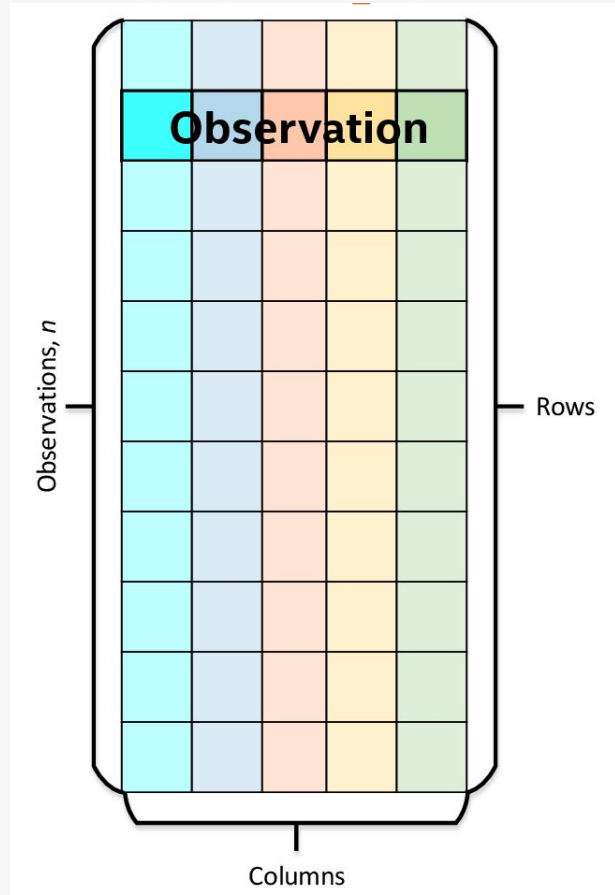
SPORTS

To monitor the performance of individual players and teams by analysis



A **dataset** is a collection of data that relate to a particular subject.

tabular data



sepal		petal		class
length	width	length	width	
6.3	2.3	4.4	1.3	versicolor
6.2	3.4	5.4	2.3	virginica
5.2	3.4	1.4	0.2	setosa
6.9	3.1	5.4	2.1	virginica
5.7	4.4	1.5	0.4	setosa
5.4	3.7	1.5	0.2	setosa
5	3.3	1.4	0.2	setosa
6.4	2.8	5.6	2.1	virginica
6	3	4.8	1.8	virginica
5.5	2.5	4	1.3	versicolor

Iris Data set

THE CONCEPT OF DATA SET



A **dataset** is a collection of data that relate to a particular subject.

tabular data

collection of
images

Observations, n											sepal		petal		class
	length	width													
0	1	2	3	4	5	6	7	8	9	4.4	1.3	versicolor			
0	1	2	3	4	5	6	7	8	9	5.4	2.3	virginica			
0	1	2	3	4	5	6	7	8	9	1.4	0.2	setosa			
0	1	2	3	4	5	6	7	8	9	5.4	2.1	virginica			
0	1	2	3	4	5	6	7	8	9	1.5	0.4	setosa			
0	1	2	3	4	5	6	7	8	9	1.5	0.2	setosa			
0	1	2	3	4	5	6	7	8	9	1.4	0.2	setosa			
0	1	2	3	4	5	6	7	8	9	5.6	2.1	virginica			
0	1	2	3	4	5	6	7	8	9	4.8	1.8	virginica			
0	1	2	3	4	5	6	7	8	9	4	1.3	versicolor			

s Data set

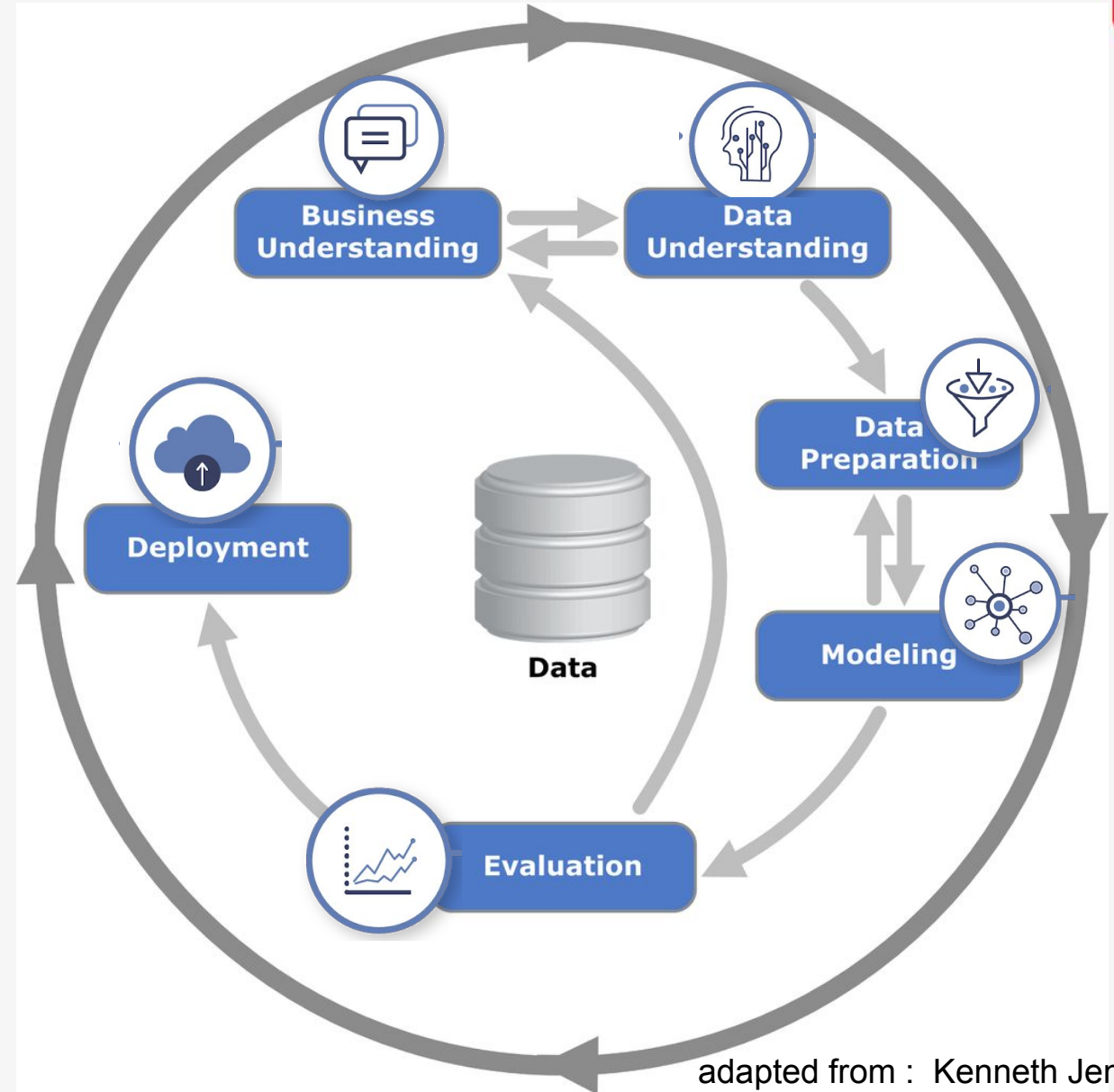
s Data set

THE CONCEPT OF DATA SET

CRISP DM[†]:

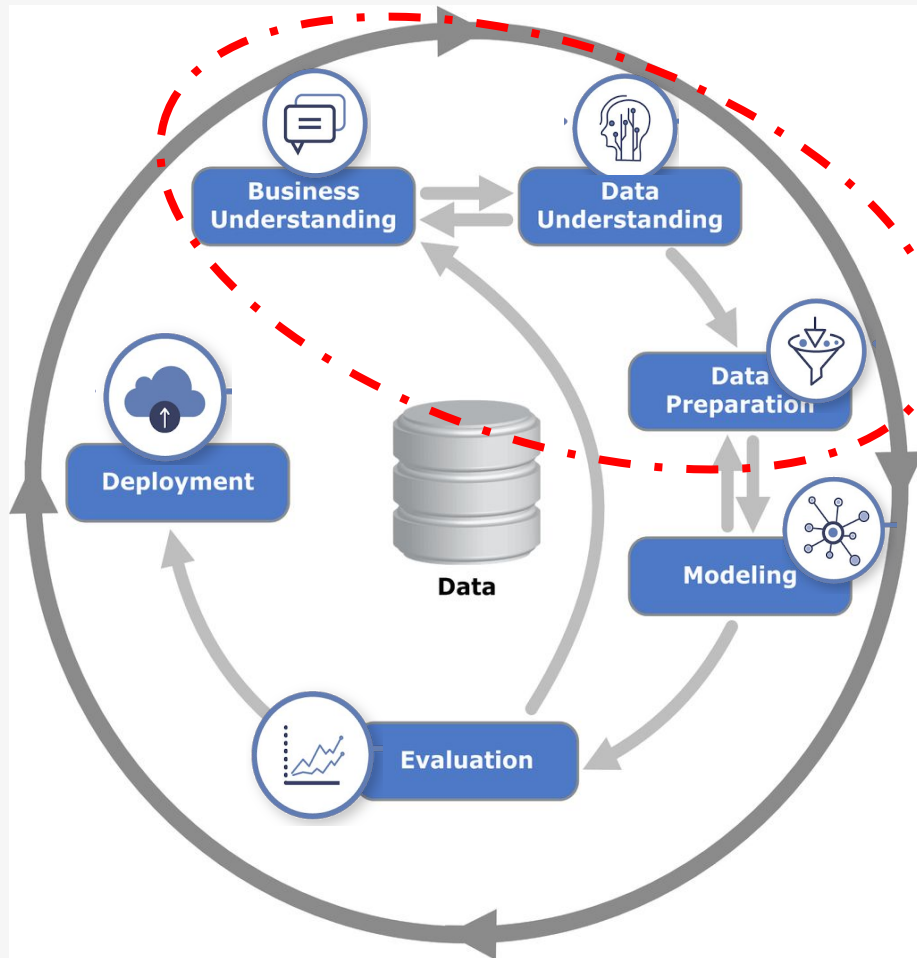
CRISP-DM breaks the process of data mining into six major steps.

sequence of the phases is not strict.



adapted from : Kenneth Jensen

CRISP DM: Phase I (1/2)



BUSINESS
UNDERSTANDING

thorough understanding of the business problem. Meet with stakeholders and domain/ subject experts to explicitly define “success criteria” for the project.



DATA
UNDERSTANDING

understand what data is collected in the business and for what purpose. Explore the data, present the data, assess quality and granularity . Formulate hypothesis on what can be found in the data.



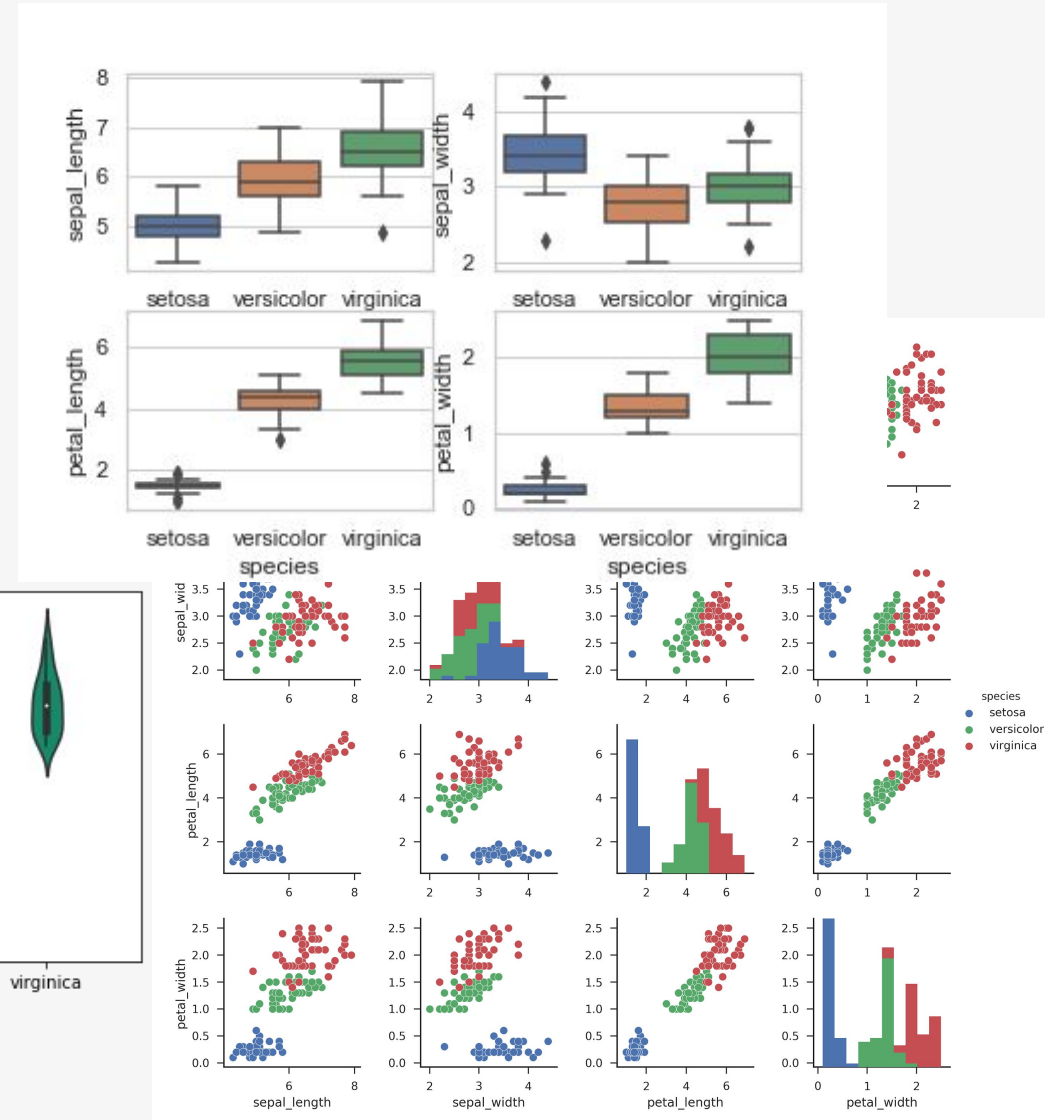
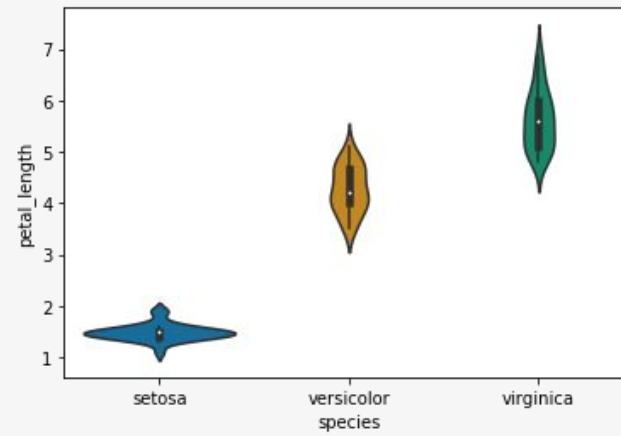
DATA
PREPARATION

given the hypothesis formulated in previous step access, transform, and condition available data into a format suitable for modeling and scoring i.e. for a machine learning prediction task

CRISP DM: Phase I (2/2)

Analyzing datasets to summarize their main characteristics
Visual Exploration

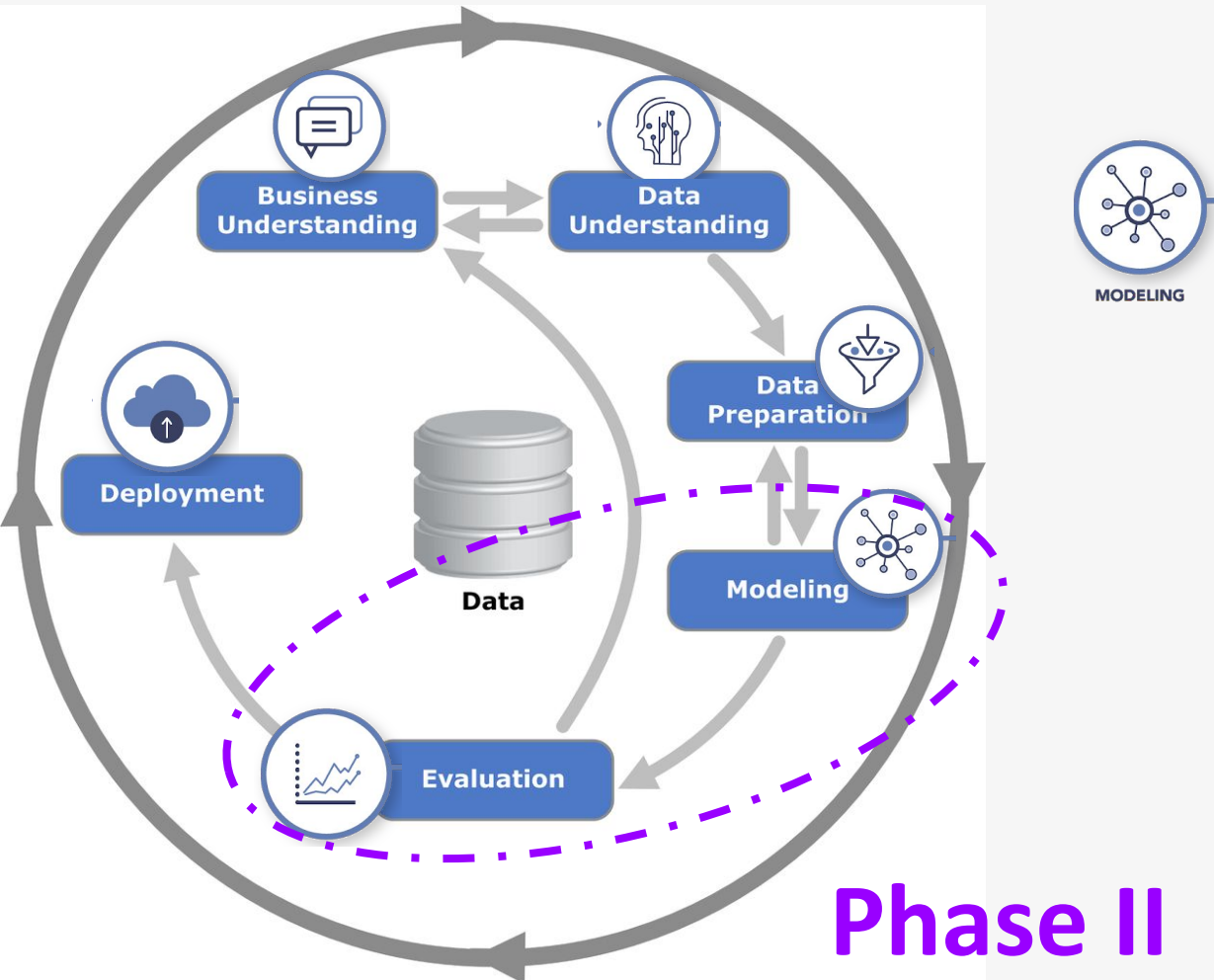
- ❑ uncover underlying structure;
- ❑ extract important variables/features;
- ❑ detect outliers and anomalies;
- ❑ test underlying assumptions;





CRISP DM : Phase II (1/2)

A model is a representation of an object, system, or business process containing an optimal mix of core features relevant to the desired use cases, for example classification or prediction.



Definition of a Parsimonious model determine optimal factor settings to represent the data/observations. for example regression, classification, clustering, text mining, computer vision

“Learning a Model”

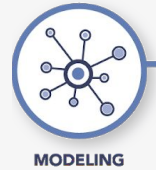
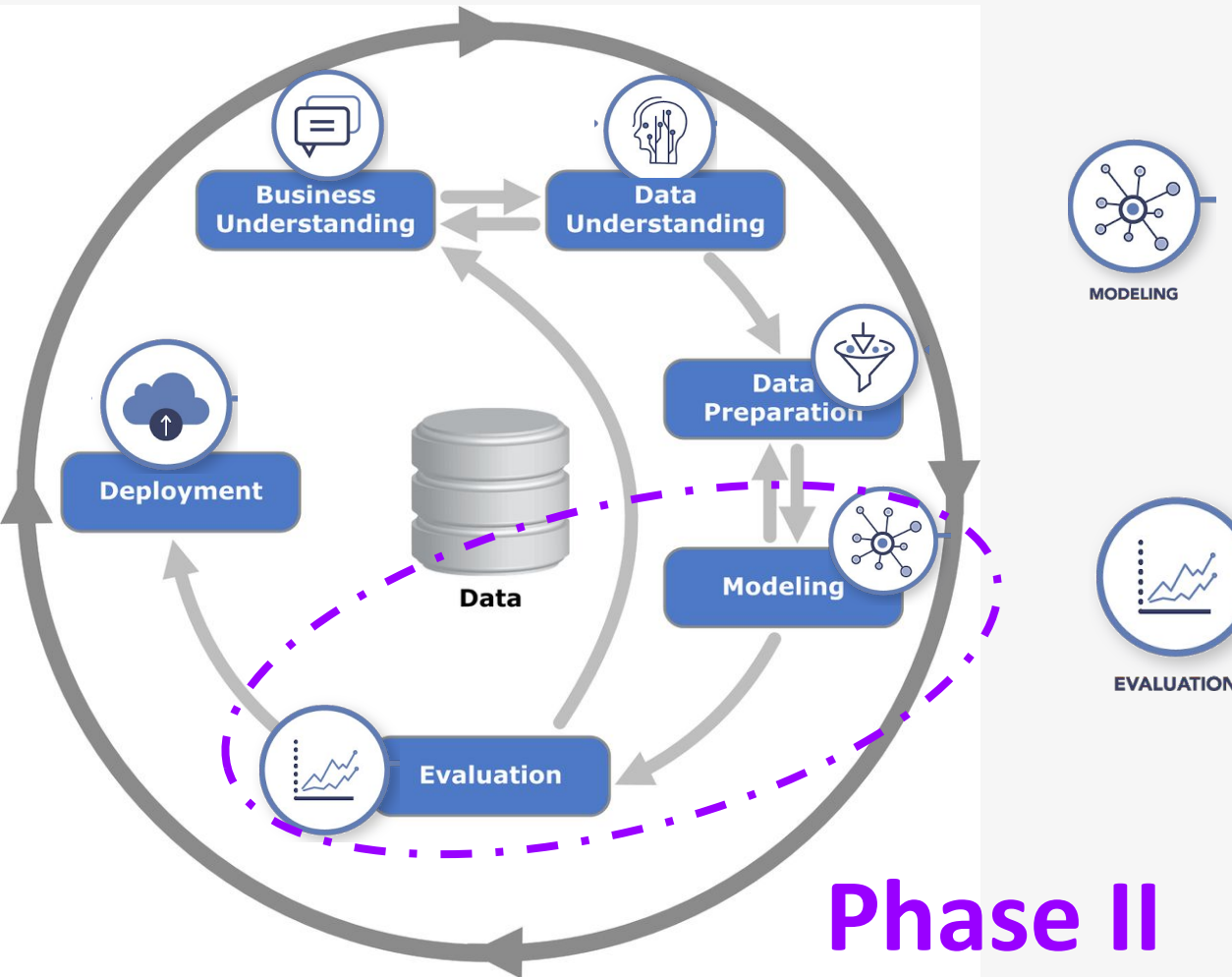
It is bottom up approach: maybe nothing or very little is known on the process generating the data.

It could be :

- ❑ a statistical model
- ❑ a set of rules
- ❑ a set of vectors
- ❑ a “description”



CRISP DM :Phase II (2/2)



Definition of a Parsimonious model determine optimal factor settings to represent the data/observations. for example regression, classification, clustering, text mining, computer vision

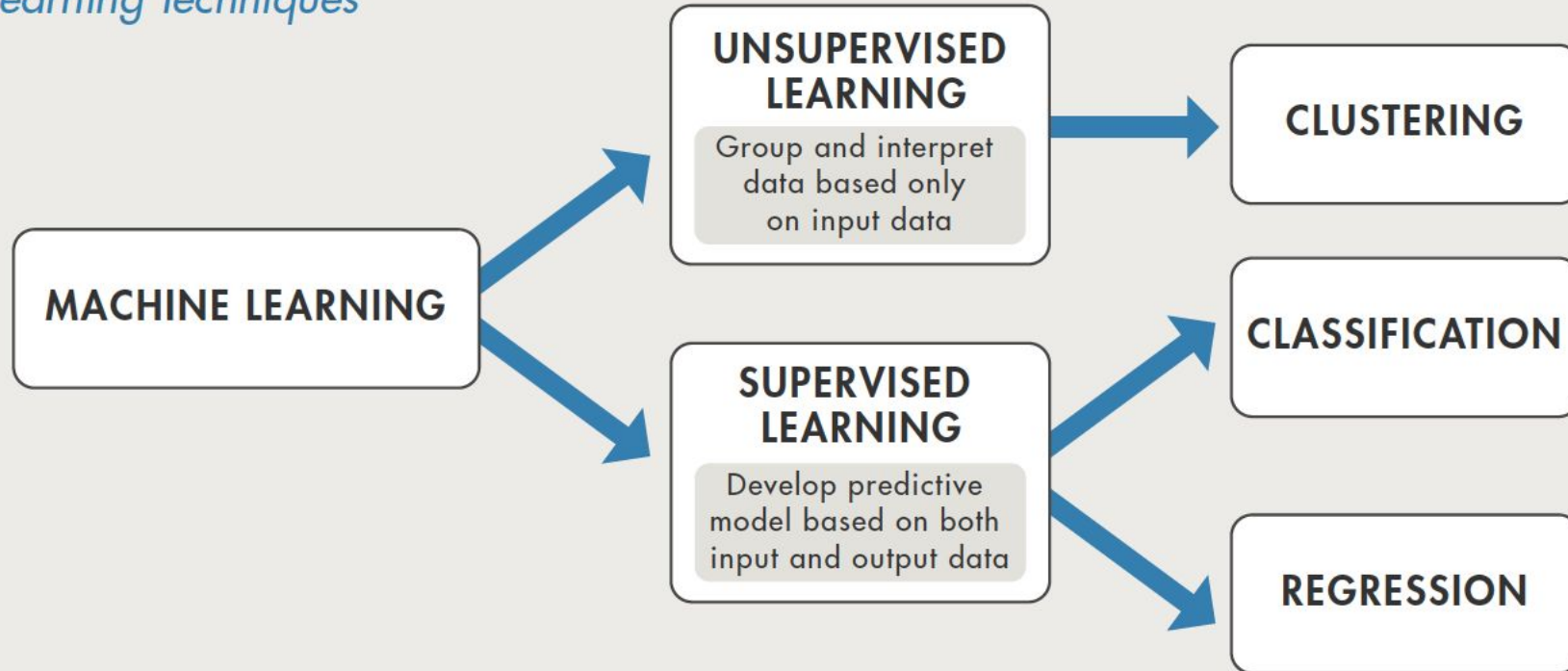


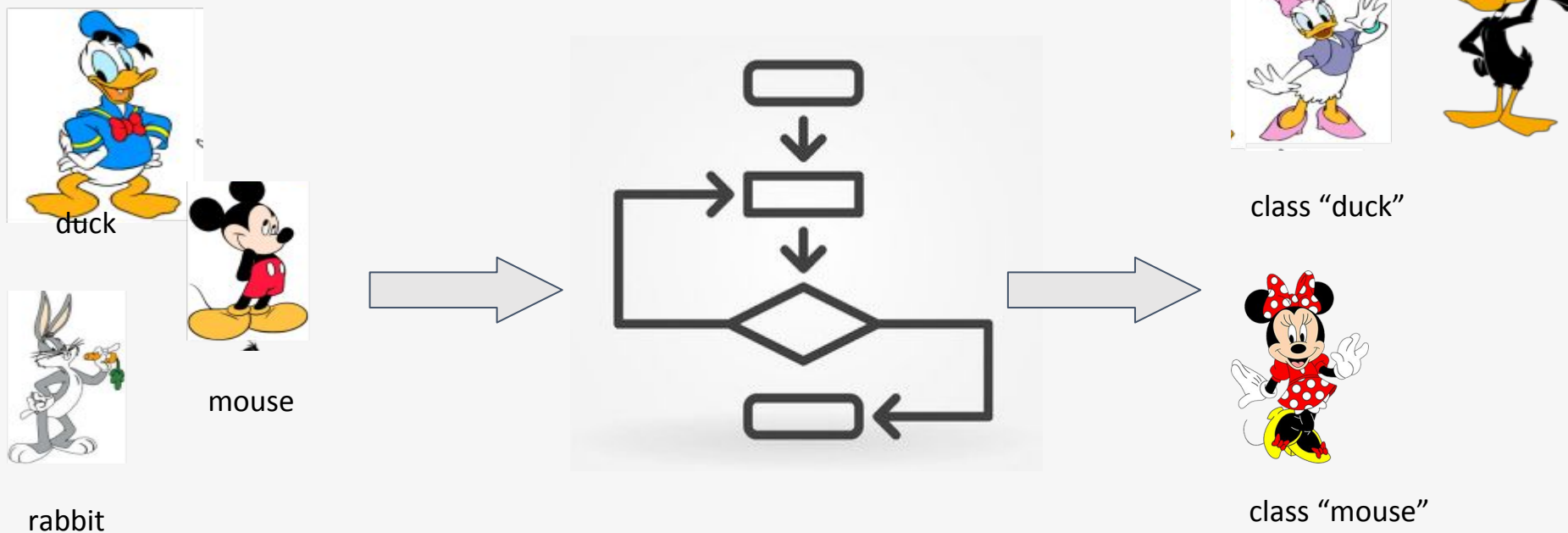
competing models are evaluated to determine which model (or model ensemble) best addresses the business objectives



CRISP-DM: Modelling

Machine Learning Techniques





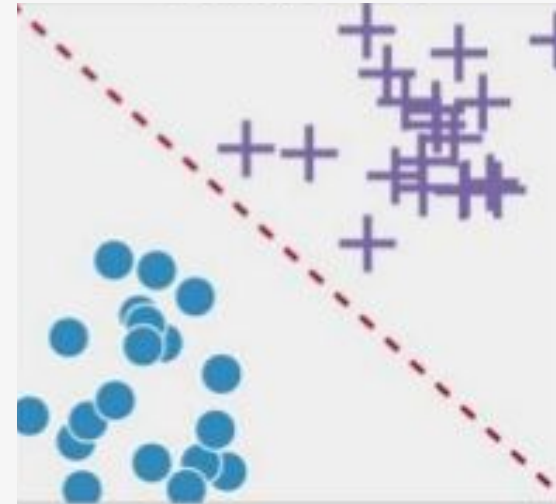
When you have a target

SUPERVISED LEARNING

Classification



attempts to predict, for each individual in a population, which of a (small) set of classes that individual belongs to



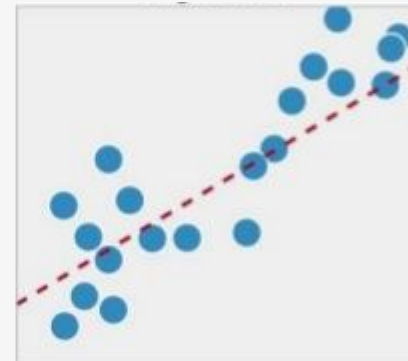
spam detection : binary classification spam and not spam.

many more example : credit approval, medical diagnosis, target marketing, land use, etc...

Regression



attempts to estimate or predict, for each individual in a population, the numerical value of some variable.



housing market : Predict sale prices of a house given the features of house.

A regression problem where input variables are ordered by time is called a time series forecasting problem

Examples: Classification vs. Regression

sepal		petal		class
length	width	length	width	
6.3	2.3	4.4	1.3	versicolor
6.2	3.4	5.4	2.3	virginica
5.2	3.4	1.4	0.2	setosa
6.9	3.1	5.4	2.1	virginica
5.7	4.4	1.5	0.4	setosa
5.4	3.7	1.5	0.2	setosa
5	3.3	1.4	0.2	setosa
6.4	2.8	5.6	2.1	virginica
6	3	4.8	1.8	virginica
5.5	2.5	4	1.3	versicolor

Classify the iris species (**virginica**, **setosa**, or **versicolor**) based on the **petal length**, **petal width**

Predict **petal length** based on the **iris species**, **petal width**, **sepal length**, **sepal width**.



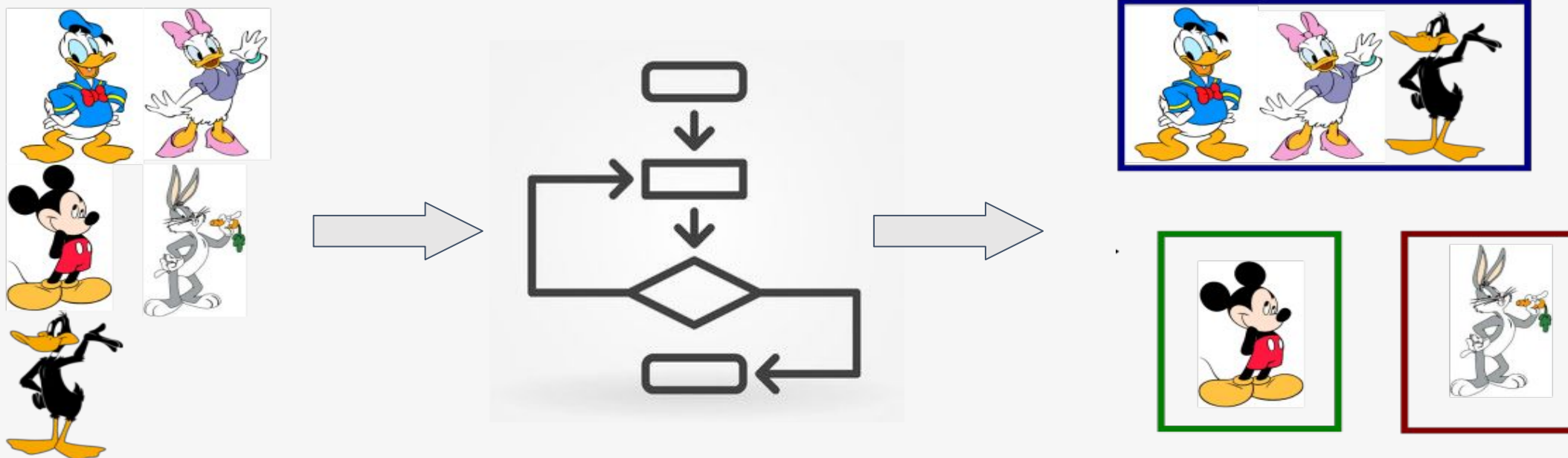
Classification

What will the weather be like tomorrow? Cold or Hot?



Regression

What will be the temperature tomorrow?

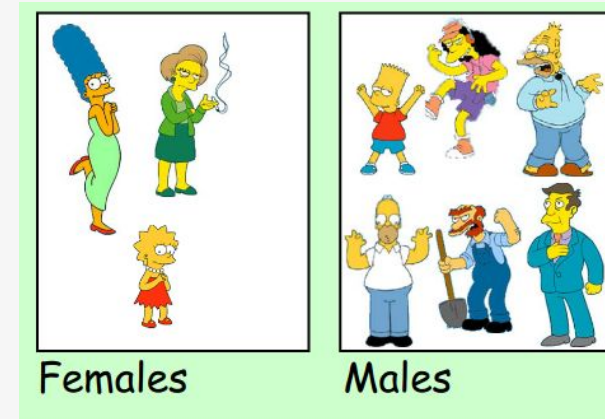
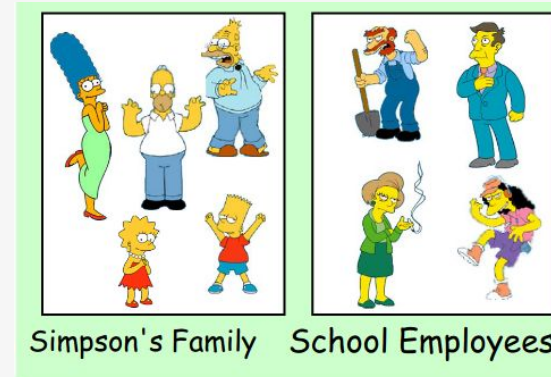
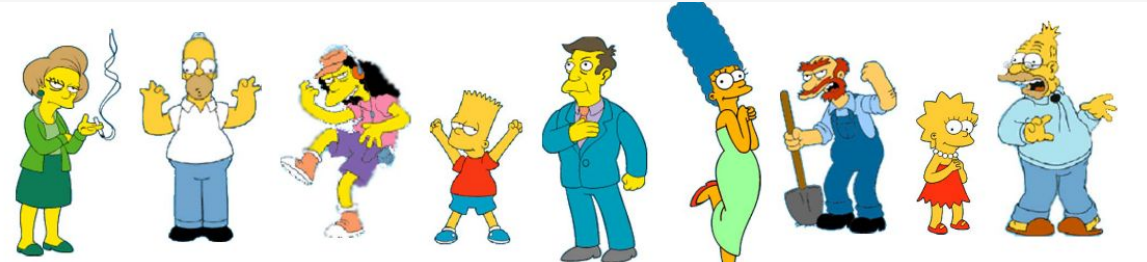


When you Don't have a target

UNSUPERVISED LEARNING

Clustering

Aims to find homogeneous subgroups such that objects in the same group (clusters) are more similar to each other than the others.



(c) Eamonn Keogh, eamonn@cs.ucr.edu

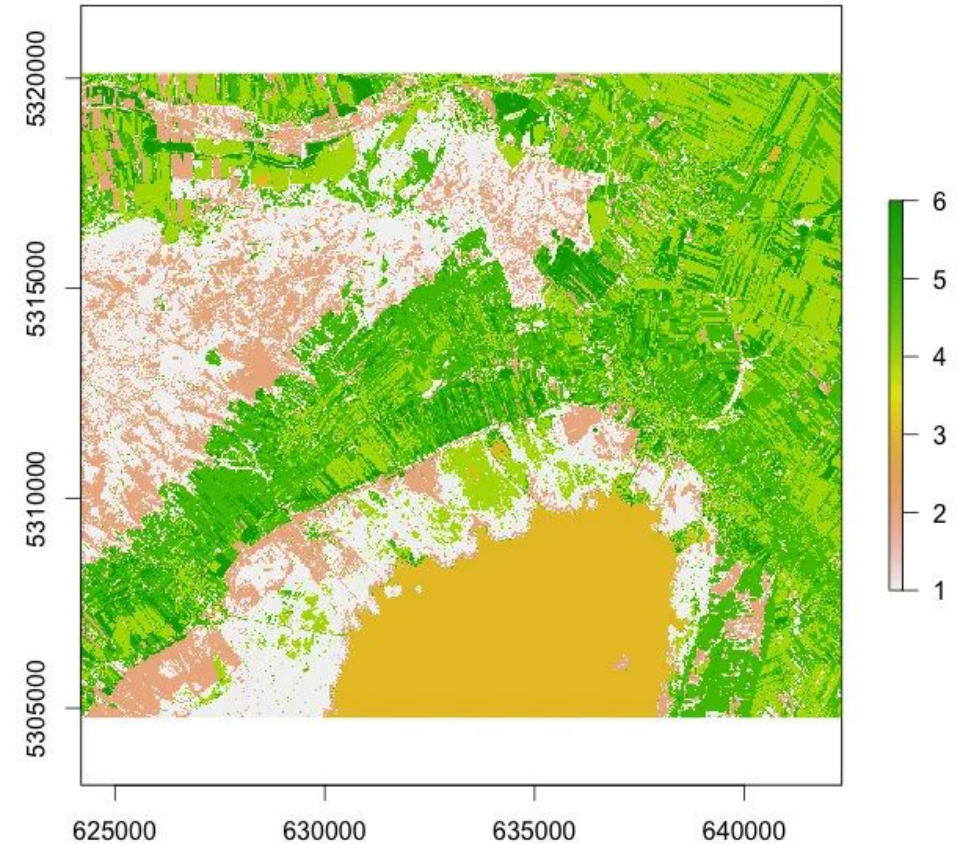
Examples: Clustering



true color composite (432)

a false color near infrared
composite (843)

Sentinel-2 image





Learning Associations

Rule based **machine learning** and data mining technique that finds important relations between variables or features in a data set.

Basket analysis:

❏ Given :

- ❏ database of transaction
- ❏ each transaction is a list of items purchased by a customer in a visit

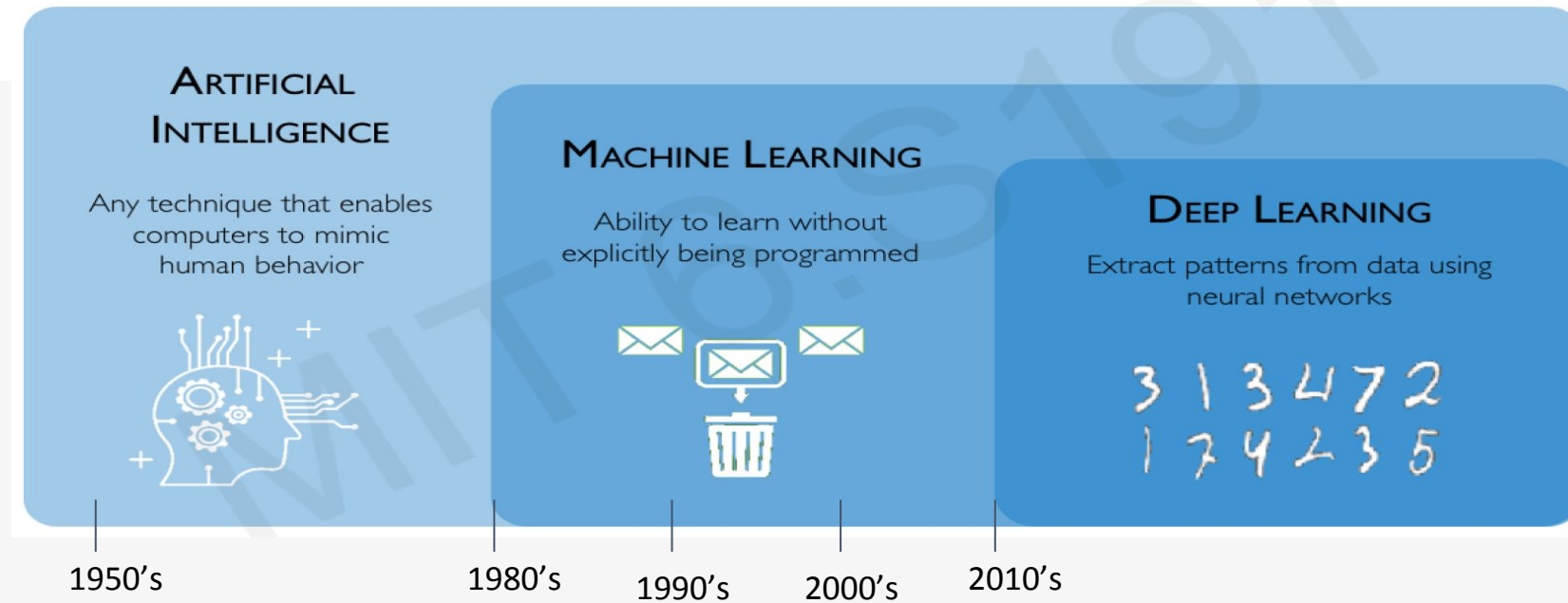
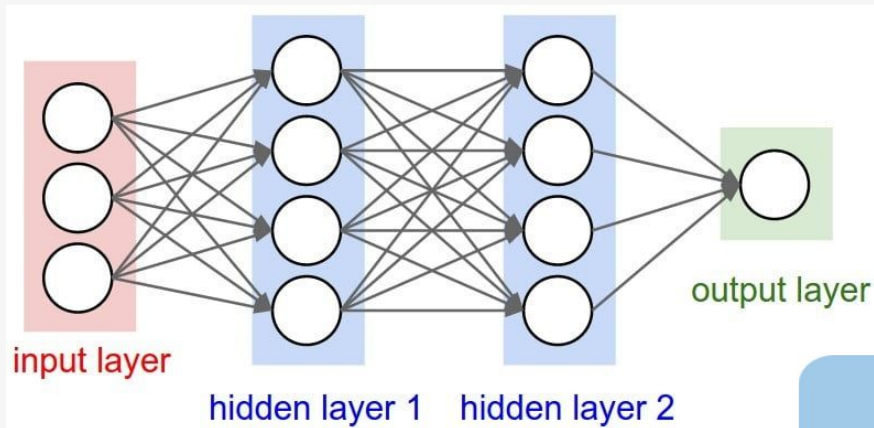
❏ Find

- ❏ all rules that correlate the presence of one set of items with that of another set of items

Example: $P(\text{chips} \mid \text{beer}) = 0.7$

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

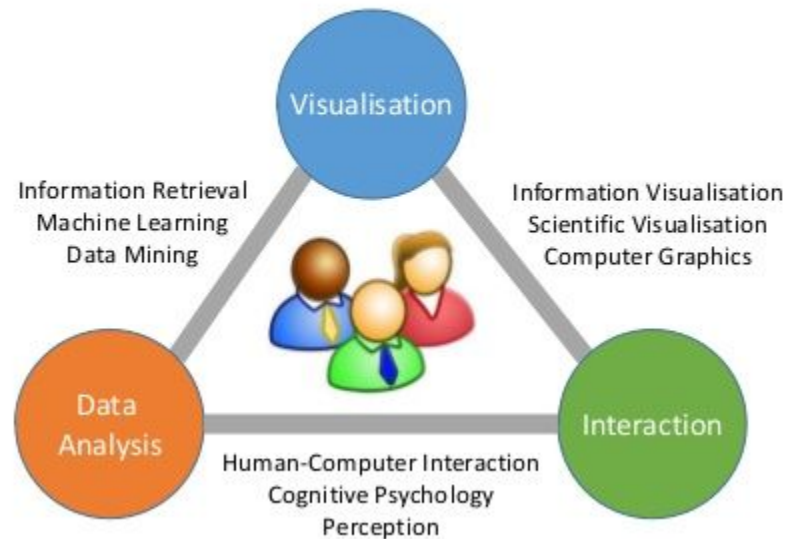
Neural Networks & Deep Learning



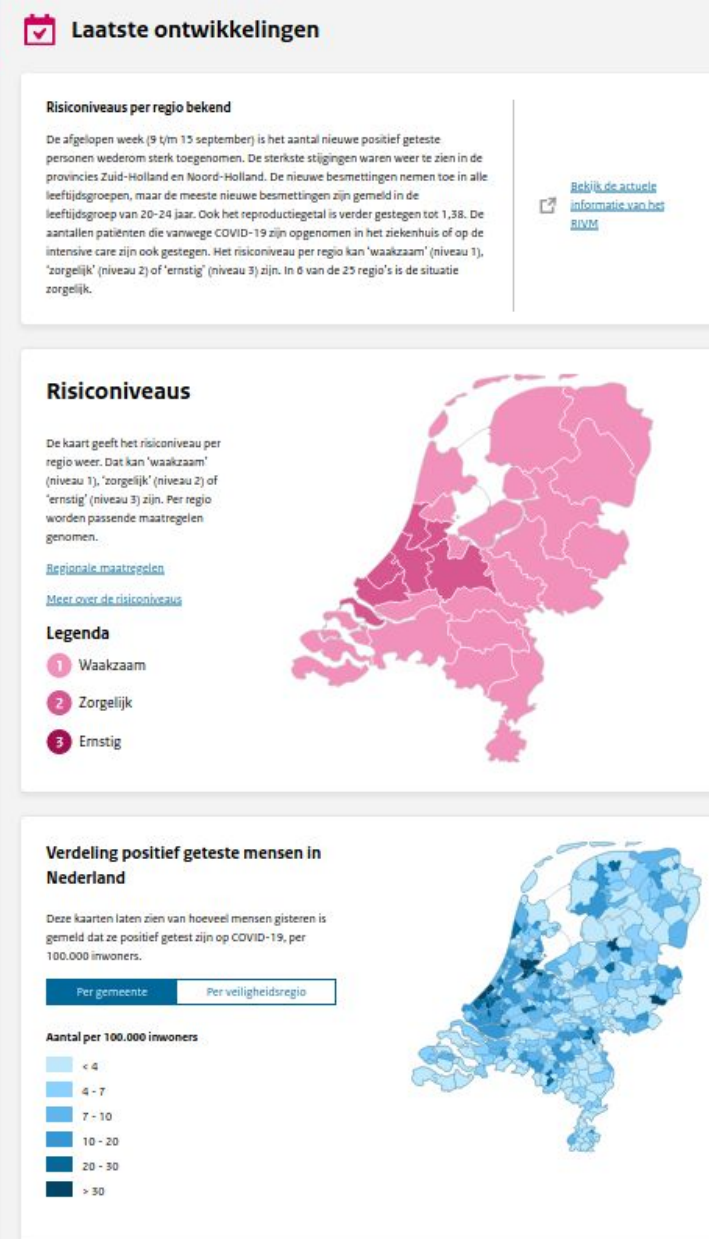
Visual Analytics

multidisciplinary field in which interactive visual interfaces are used to support analytical reasoning

Visual Analytics = Visualisation + (Data) Analytics



Visual Analytics



**BEDANKT
VOOR
UW
AANDACHT**

Open Universiteit



LEAD

Break-out groups session

Customer Churn

THE CHALLENGE

A telecommunication company because of the strenuous competitions given by the concurrence competitors has to switch from growth strategy to protective approach of their existing customers. Their standard 'carpet' marketing campaigns had been delivering poor ROI and so they asked your team for help.





Solution design: the *churn* use case

SOLUTION DESIGN Design your data-driven solution for 'churn' prediction.

1. How do you define 'churn'? Provide an operational definition (e.g. no purchases in the last N months)
2. What kind of input data you need? (demographic data, customer service calls, transactional data, etc.)
3. How many data samples are you considering?
4. What machine learning approach is suitable for this task: classification, regression, clustering?
5. How is the TLC company supposed to use the outcomes of your solution?
6. What would be the benefits for TLC company in using your solution in the longer term?



Live tutorial: Data analytics in Python