ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

IT3160E
Introduction to Artificial Intelligence

Chapter Advanced topics – Machine learning

Lecturer:
Muriel VISANI

Department of Information Systems
School of Information and Communication Technology - HUST

Content of the course

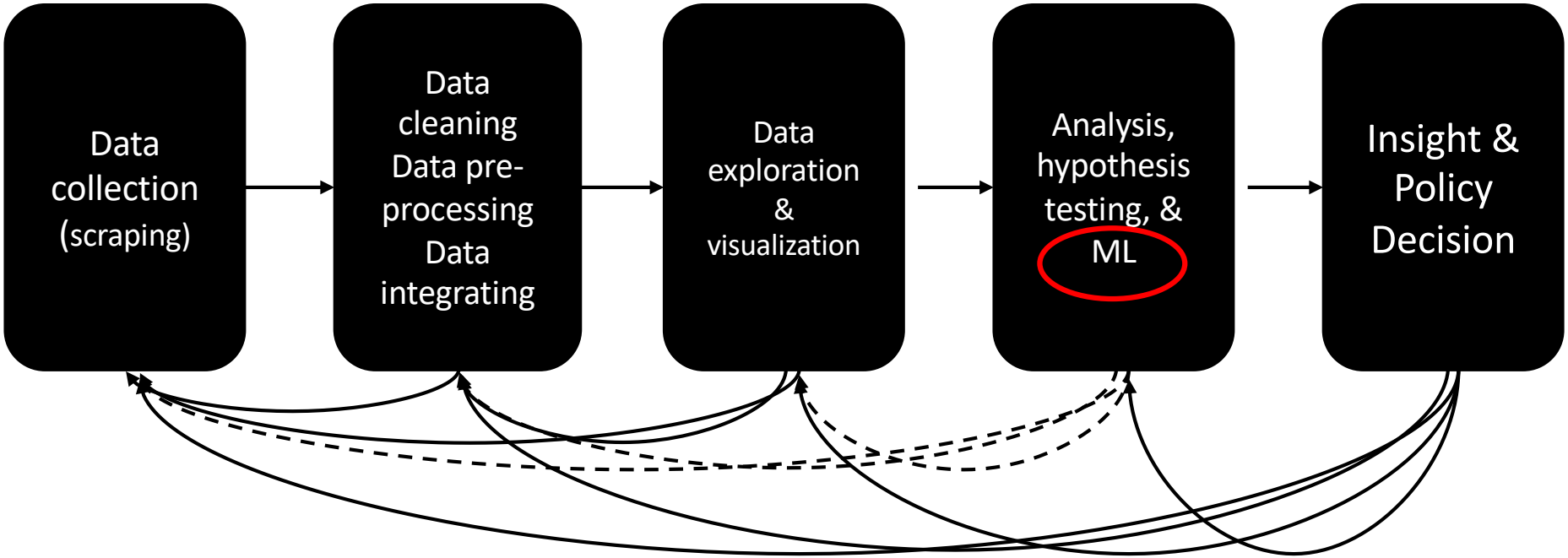
- Chapter 1: Introduction
- Chapter 2: Intelligent agents
- Chapter 3: Problem Solving
  - Search algorithms, adversarial search
  - Constraint Satisfaction Problems
- Chapter 4: Knowledge and Inference
  - Knowledge representation
  - Propositional logic and first-order logic
- Chapter 5: Uncertain knowledge and reasoning
- Chapter 6: Advanced topics
  - Machine learning
  - Computer Vision

Outline

- **Chapter 6: Machine Learning**
  - Introduction and definitions
    - Machine Learning Process
    - Main Goals of Machine Learning
  - Supervised *vs.* unsupervised learning
    - Supervised learning
    - Unsupervised learning
  - Supervised classification
    - Objective
    - A few words bout performance evaluation
    - Risk of over-fitting
    - Methods
    - Performance evaluation
  - Homework
  - Summary

Introduction and definitions

Machine Learning in the global framework of Data Science



What is Machine Learning?

- Machine learning is a sub-domain of Artificial Intelligence
- Sometimes related to cognitive sciences (*e.g.* neural networks)
- Based on **inference**
- Solves a wide variety of problems
  - Used in almost any application domains
    - Smart cars, diagnostic assistance for doctors, spam filters, targeted advertising, etc.

Different types of inference

- **Deductive** (Logical) inference:
  - From A and A $\rightarrow$ B, infer B
  - Deducing the consequences from the causes (premises)
  - Logical rules (see chapter 4)
- **Abductive** inference :
  - From B and A $\rightarrow$ B, infer A
  - Making hypothesis about the causes, from the consequences
  - Application to diagnostic systems

Different types of inference

- **Inductive** inference:
  - This is the type of inference mostly used in Machine Learning
  - Knowledge is extracted on the basis of specific examples

Valid only in probabilistic terms **(uncertainty)**
Watch out for hasty generalizations!

Learning for inductive inference

- Learning serves for **inductive** inference
- The aim is to extract a model from data samples (observations)
  - These observations are contained in a learning dataset

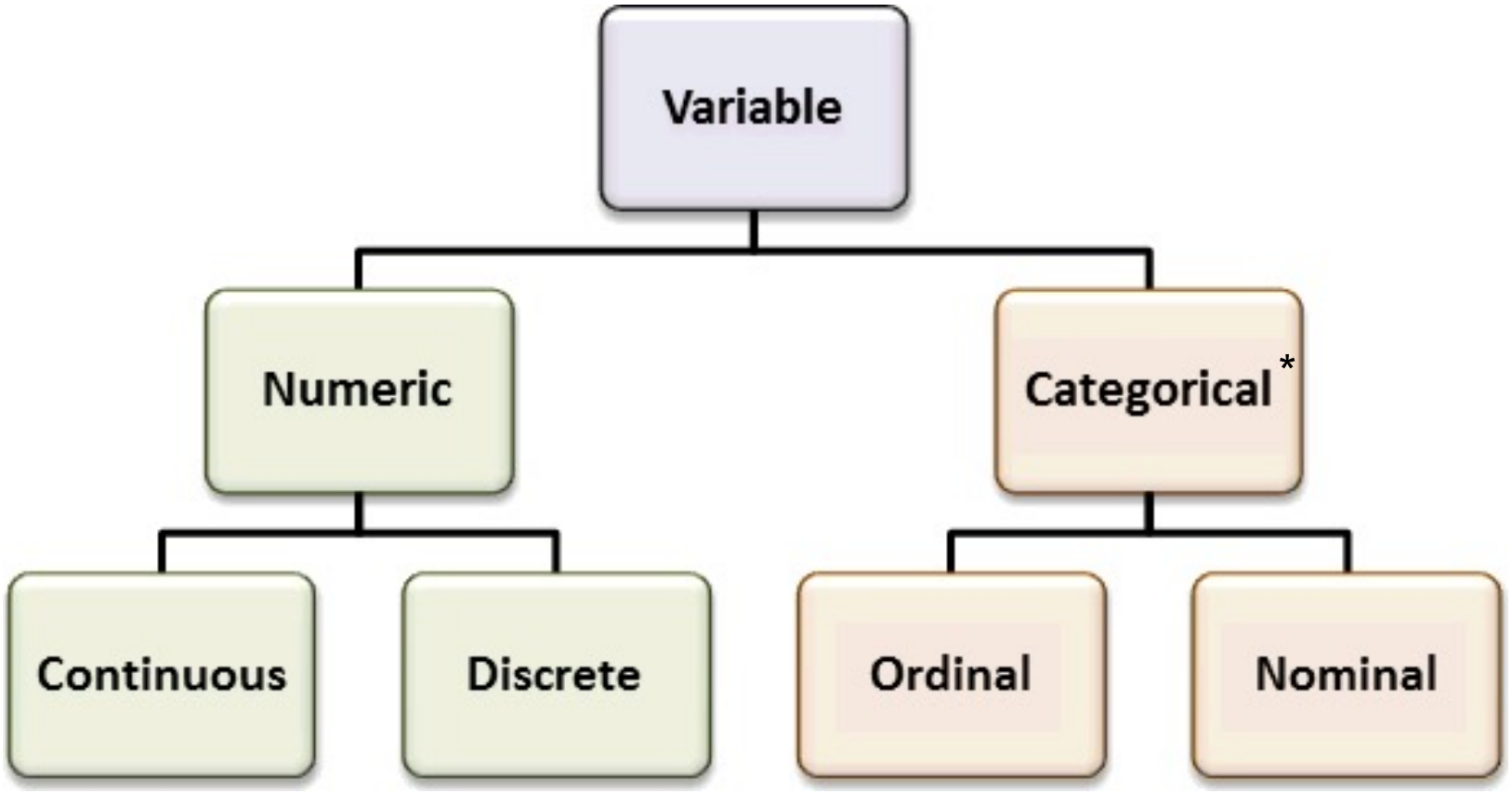
- A model is used to extract knowledge
  - The model can be improved with experience
    - *i.e.* with more observation in the learning dataset

Defs: observations and variables

- Data is an collection of observations (records, subjects)
  - Rows in the data table
- An attribute is a set of values describing some aspect across all observations, it is called a variable
  - Columns in the data table

HR Information		Contact	
Position	**Salary**	**Office**	**Extn.**
Accountant	$162,700	Tokyo	5407
Chief Executive Officer (CEO)	$1,200,000	London	5797
Junior Technical Author	$86,000	San Francisco	1562
Software Engineer	$132,000	London	2558

Types of variables



*The possible values of categorical variables are called modalities

Quiz

Introduction and definitions

Machine Learning Process

Machine learning (ML) process

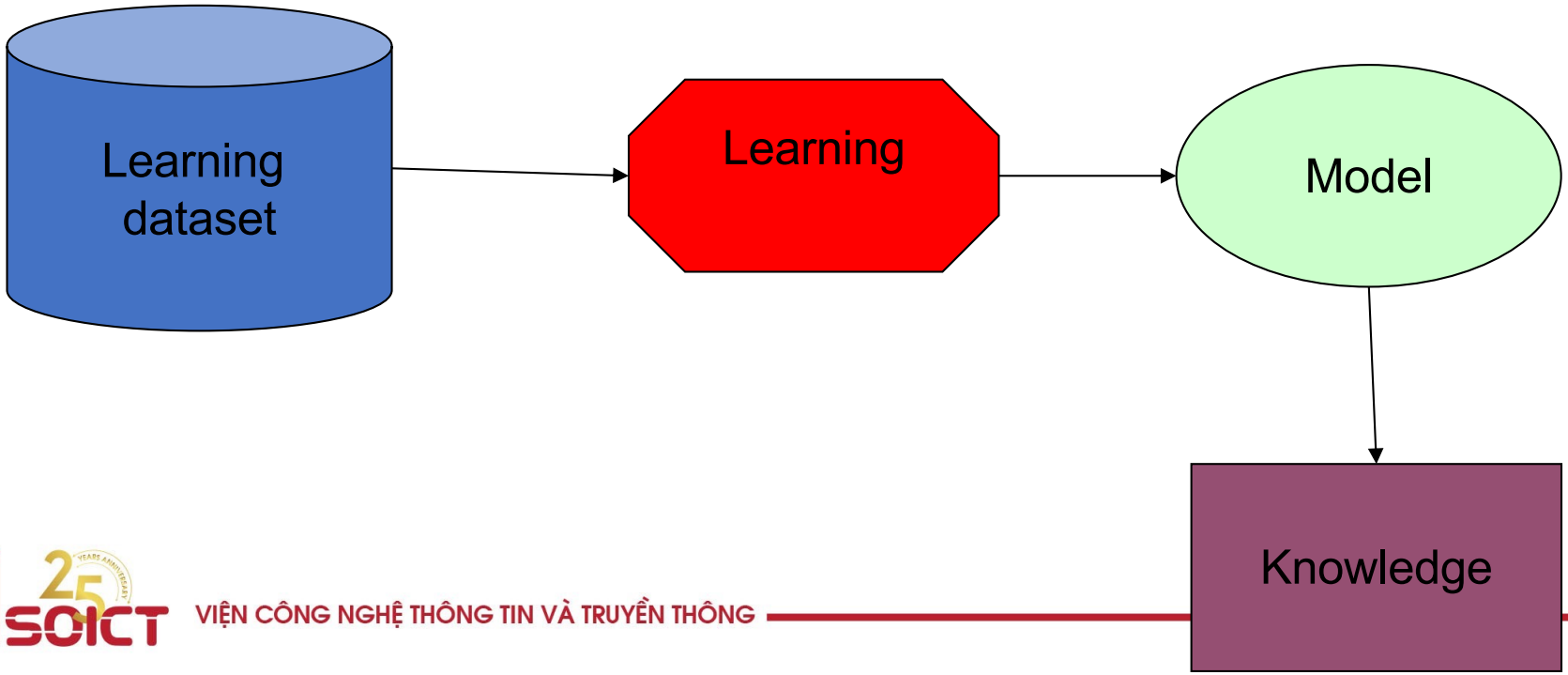
ML is a **two-step process**:
Goal: extracting a model to generate knowledge

Machine learning (ML) process

Step 1: Learning the model

ML is a **two-step process**:
Goal: extracting a model to generate knowledge

- on the learning dataset (model fitting)

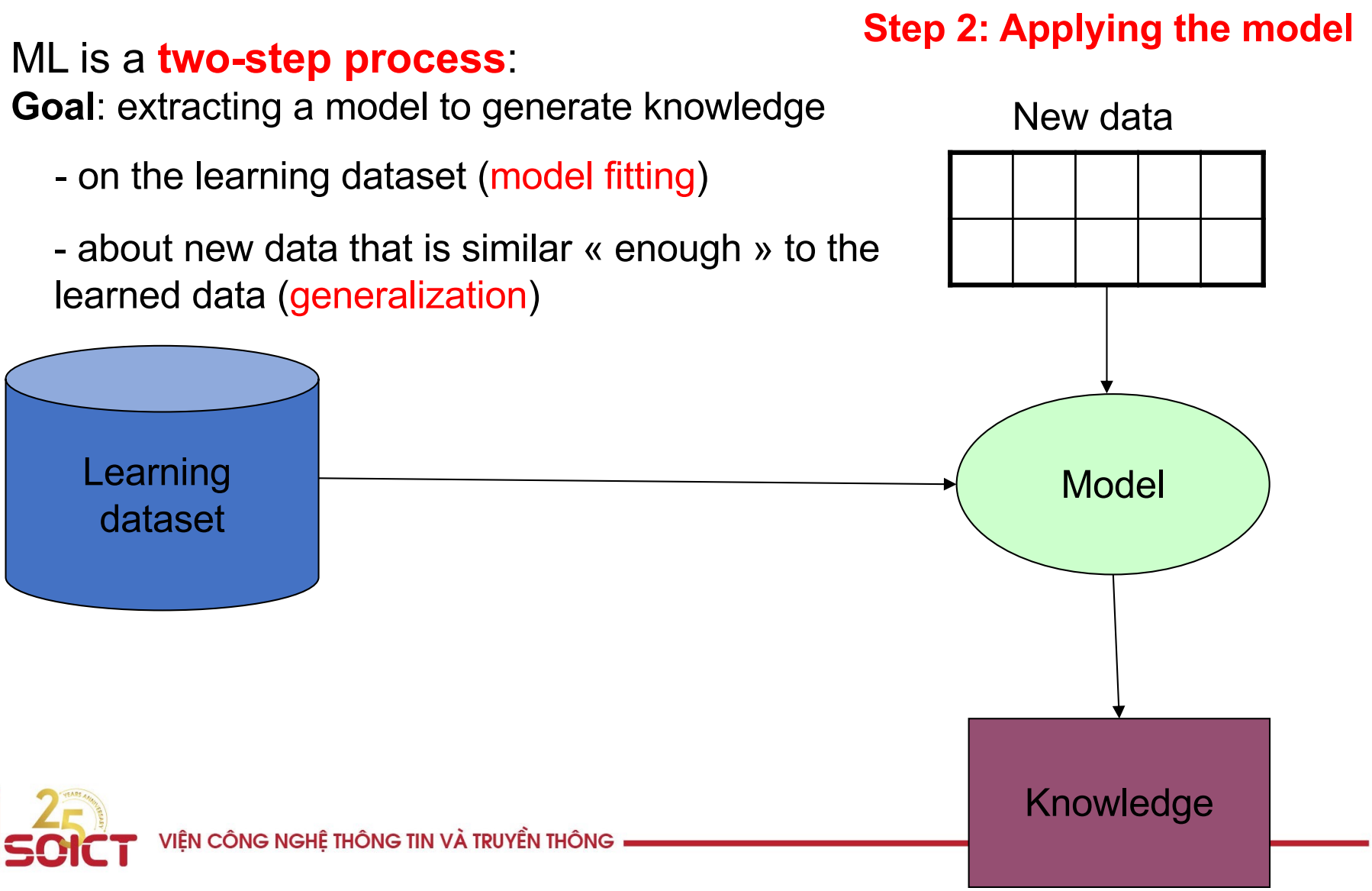


Machine learning (ML) process

ML is a **two-step process**:
Goal: extracting a model to generate knowledge

- on the learning dataset (model fitting)

- about new data that is similar « enough » to the learned data (generalization)



Introduction and definitions

Main goals of Machine Learning

Goals of ML

- The main objectives of Machine Learning are:
  - Description
  - Segmentation
  - Association
  - Prediction

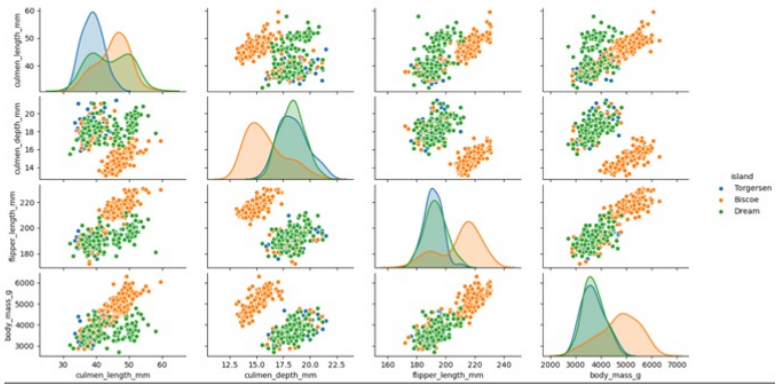
Main goals of Machine Learning
Data description

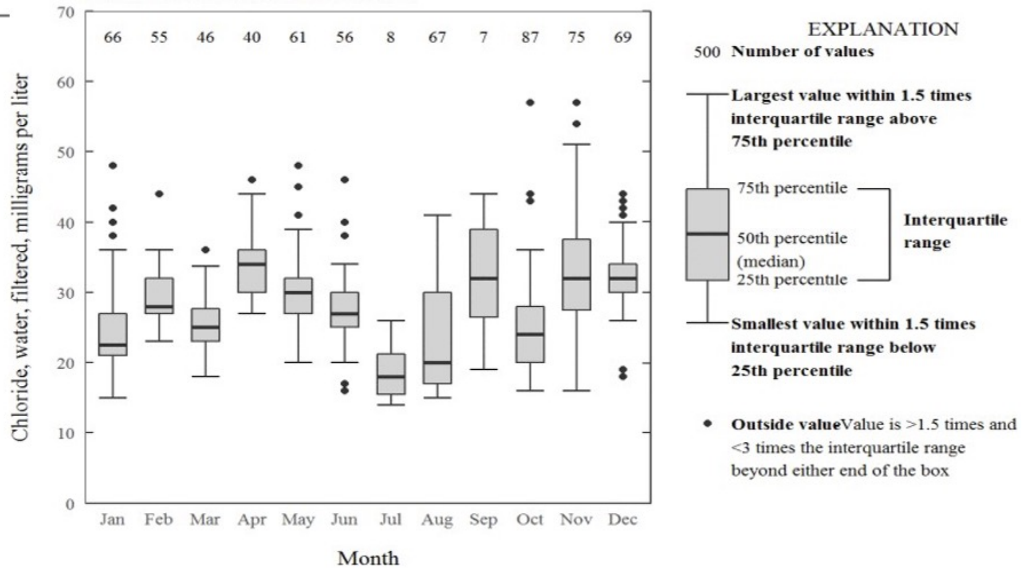
- Data description consists in summarizing the data in an “understable” way, either:
  - Through **exploratory data analysis**
    - Mostly descriptive statistics such as average, standard deviation, median, PCA…
  - Through **data visualization**

Main goals of Machine Learning
Data description through Exploratory Data Analysis

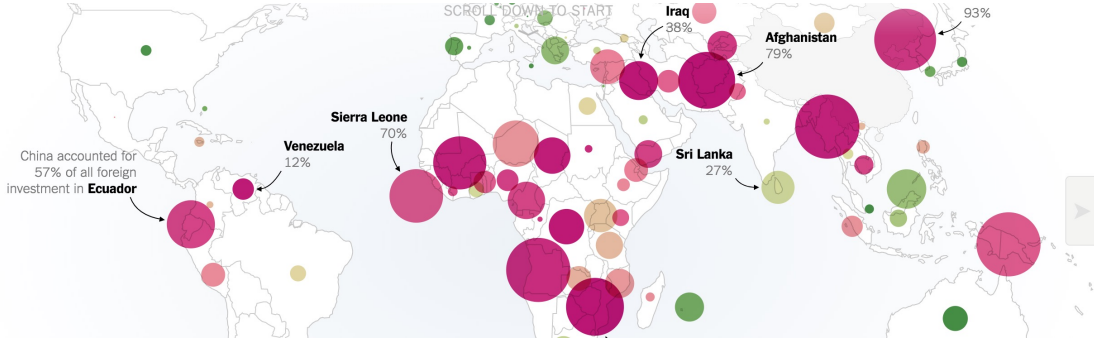
	INDIVIDUALS' AGE					
	Subadult		*Adult*		*Senior*	
Acoustic parameters	**median**	**range**	**median**	**range**	**median**	**range**
call duration [s]	10.3 (8.1, 12.8)	4.4–56.5	8.4(6.8, 10.6)	1.9–201.6	8.3 (7.2, 9.5)	1.7–29.8
number of elements	8 (6, 9)	4–50	9 (7, 12)	3–127	9 (8, 10)	1–33
element duration [s]	0.46 (0.20, 0.73)	0.01–1.73	0.27 (0.14, 0.58)	0.01–2.09	0.24 (0.13, 0.58)	0.02–1.73
interval duration [s]	0.88 (0.50, 1.24)	0.06–4.47	0.54 (0.32, 1.01)	0.04–4.34	0.46 (0.29, 0.97)	0.09–2.74
start F0 [Hz]	680 (640, 760)	530–1460	700 (620, 850)	390–1540	670 (600, 870)	10–1530
end F0 [Hz]	920 (820, 1040)	600–1700	950 (810, 1090)	480–1950	960 (820, 1100)	430–1570
max F0 [Hz]	930 (870, 1010)	640–1430	950 (840, 1040)	530–1480	930 (810, 1040)	500–1400
location of max F0 [s]	1040 (950, 1170)	710, 1900	1070 (930, 1210)	540–1950	1120 (960, 1210)	510–1680
number of elements	*N* = 2476		*N* = 5557		*N* = 2344	
number of calls	*N* = 259		*N* = 533		*N* = 257	

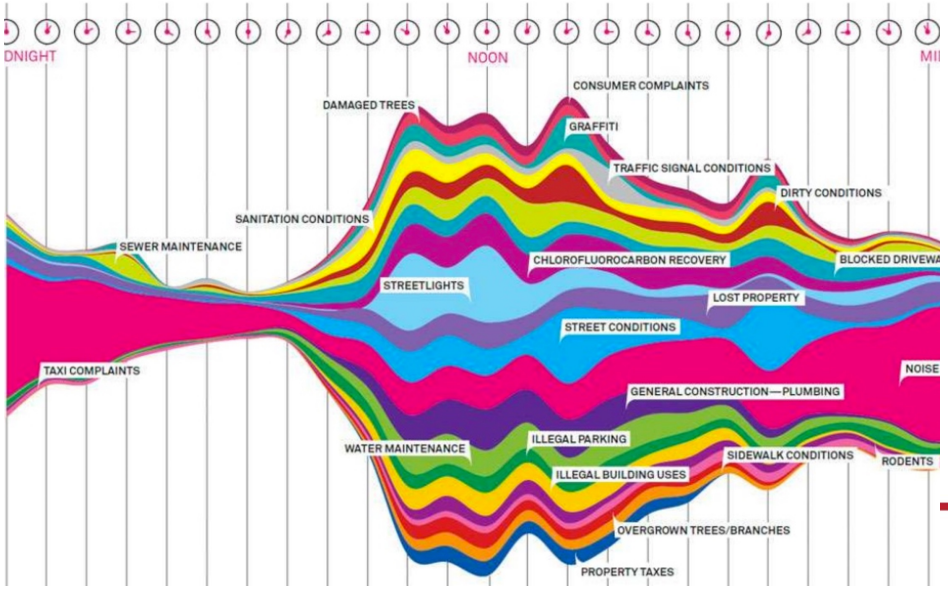
doi:10.1371/journal.pone.0082748.t004

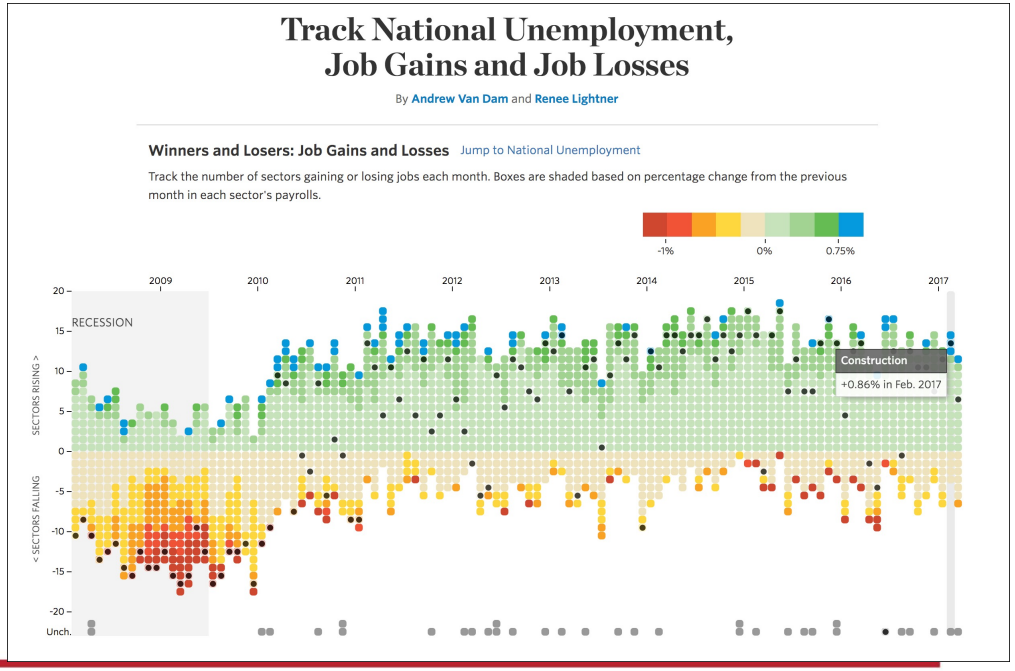




Main goals of Machine Learning
Data description through visualization





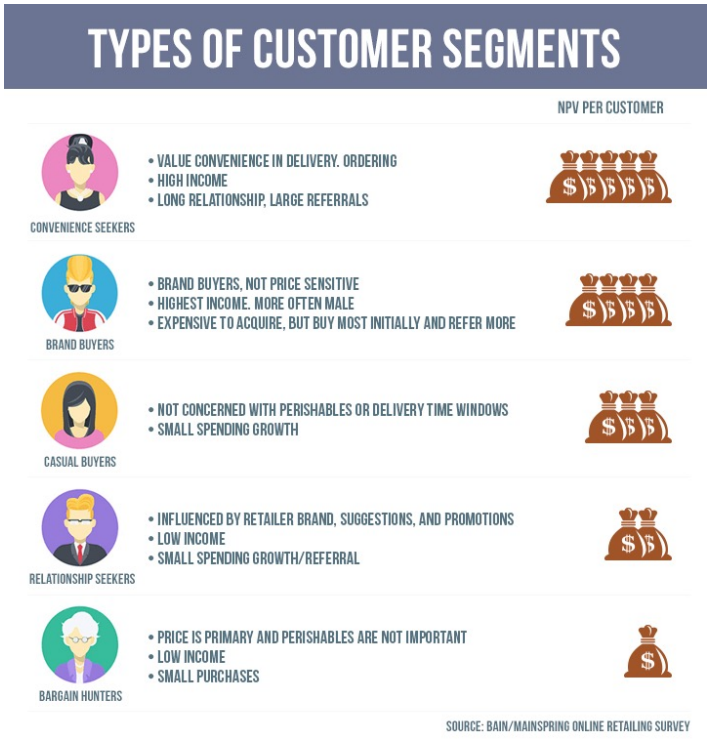


[http://graphics.wsj.com/job-market-tracker/]

Main goals of Machine Learning
Data segmentation

- Data segmentation consists in grouping the similar records into homogeneous groups (called **clusters**)
  - Records in a group have similar attribute values
  - Technically, the goal is to lean a “new” attribute (group#) from the record’s attributes
  - **Unsupervised** machine learning methods can be used

Main goals of Machine Learning
Data segmentation



	GEOGRAPHIC	DEMOGRAPHIC	PSYCHOGRAPHICS	BEHAVIOURAL	PERSONA	PREDICTIVE
SIMPLE	Where	Who	Why	What	Who, What, Why, Where	Who and When
WHAT IS IT?	Geographic segmentation divides customers into groups based on their location.	Demographic segmentation divides customers into groups based on census data.	Psychographic segmentation divides customers into groups based on personal interests and motivations.	Behavioural segmentation divides customers into what do - online/offline.	Persona segmentation divides customers into groups based on a blended data. as well as customer goals.	Predictive segmentation uses historical behavioral patterns to predict and influence future customer behaviors.
EXAMPLES	Countries Cities Urban, Suburban, Rural IP Addresses	Age Income Family/Single/Couple Gender Education	Interests Personality Lifestyle Social Status Activities, Interests, Opinions Attitudes	Benefits Sought Occasion Usage Rate Loyalty Buyer Readiness Actions taken e.g. online	Jobs to be done Pain/Gains Demographic data Psychographic data Behavioural data	Usupervised Learning Supervised Learning Reinforcement Learning
WHY USE IT	Dynamic Pricing Ease of use Country/Language differences Localized offers - stores	Easy to use Good for store profiling Ideal for life stages Good to supplement with other data	Uncovers motivations and reasons for product and brand purchases	Ideal for identifying patterns and triggers during buying process. Helps to tailor marketing to different stages.	Provides a rich profile of a customer segment. Proves a foundation to test hypothesis and testing to optimize results.	Uncovers hidden buying clusters of customers. Helps with customer discovery.

Main goals of Machine Learning
Data association

- Association consists in discovering association rules between records, according to pre-defined criteria
  - *E.g.* the items that are often bought during one single transaction
  - Technically, the goal is to lean a “new” information (association rules) from the record’s attributes
  - **Unsupervised** machine learning methods can be used

Main goals of Machine Learning
Data association

amazon

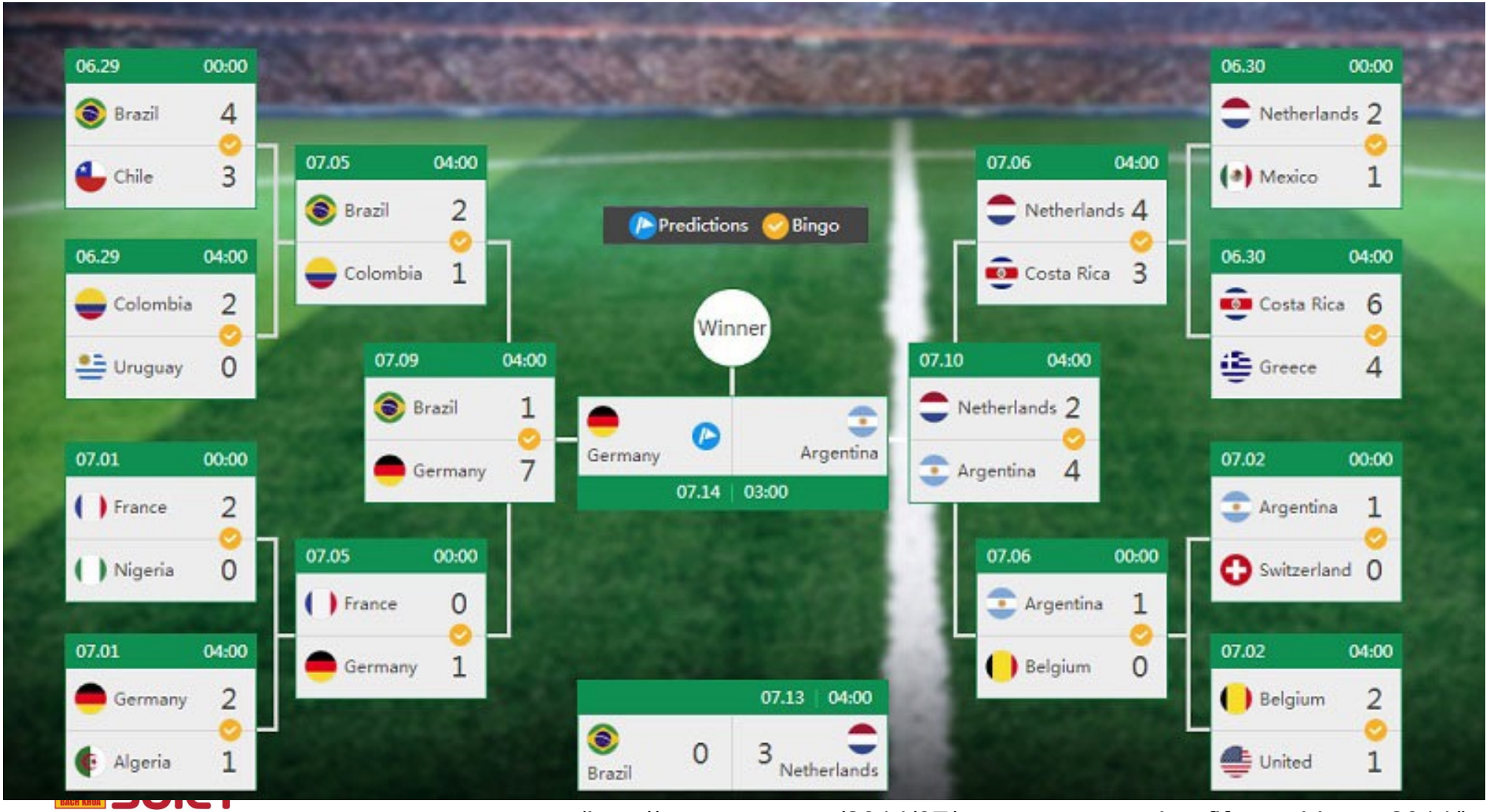
"The company reported a **29% sales increase** to $12.83 billion during its second fiscal quarter, up from $9.9 billion during the same time last year."
– Fortune, July 30, 2012

Main goals of Machine Learning
Data prediction

- Data prediction consists in either:
  - predicting (in the future) or estimating (in the present) the values of an attribute for a set or records
    - This attribute is known for other records
    - This knowledge is used to predict this attribute's values on our set of records
    - **Supervised** machine learning methods can be used

Accuracy ~93%.

Main goals of Machine Learning
Data prediction



Accuracy ~93%.

(http://yourstory.com/2014/07/germany-argentina-fifa-world-cup-2014/)

Goals of ML

- The main objectives of Machine Learning are:
  - Description
  - Segmentation
  - Association
  - Prediction
- Most of the time, the goal is to…
  - Infer the value of a new variable (response variable)…
  - …based on the observed values of explanatory variables…
  - …for the learning dataset and/or new data

Quizz

- A group of students wants to know if they can use people's height to predict their age
- They take a random sample of 50 people and record each individual's height and age
- **Question**: What is the learning dataset?
  - Solution:
- **Question**: What is the explanatory variable?
  - Solution:
- **Question**: What is the response variable?
  - Solution:

Supervised vs. unsupervised learning

Supervised *vs.* unsupervised: different goals

- Depending on the final objective, we might have to use a method based on supervised, unsupervised or semi-supervised learning
  - Supervised learning serves for **prediction**
  - Unsupervised learning might be used for **description, segmentation or association**
  - Semi-supervised learning can be used for segmentation or **prediction**

Machine learning

Supervised learning

Supervised vs. unsupervised learning

Supervised learning

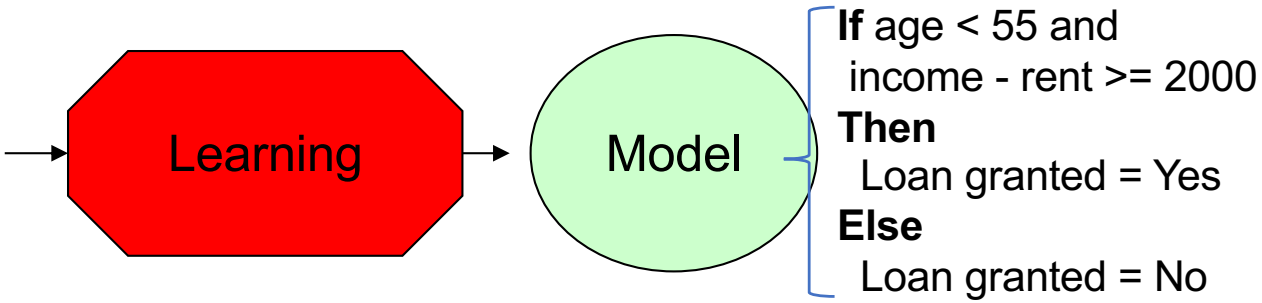
- **Supervised** learning methods
  - The learning dataset contains the values of the response variable(s)

Supervised learning

- **Supervised** learning methods
  - The learning dataset contains the values of the response variable(s)
  - Example

Step 1: Learning the model

Learning dataset			
Age	Rent	Income	Loan granted?
36	0	1299	Yes
55	240	2500	No
40	768	3000	Yes
39	0	2000	Yes
44	334	512	No
26	631	722	No

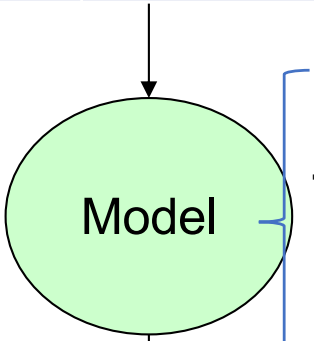


Supervised learning

- **Supervised** learning methods
  - The learning dataset contains the values of the response variable(s)
  - Example

Step 2: Applying the model

New example			
Age	Rent	Income	Loan granted?
30	260	1900	???



If age < 55 and
income - rent >= 2000
Then
Loan granted = Yes
Else
Loan granted = No

New example			
Age	Rent	Income	Loan granted?
30	260	1900	**No**

The model cannot always be expressed using rules
- it depends on the model
- *e.g.* decision trees give rules but neural networks don't

Supervised learning

- **Supervised learning** often serves for prediction
  - Main **tasks**: regression and classification

- **Regression** (linear regression, regression trees, neural networks...)
  - The response variable is **numeric**
  - The explanatory variables are the attributes observed

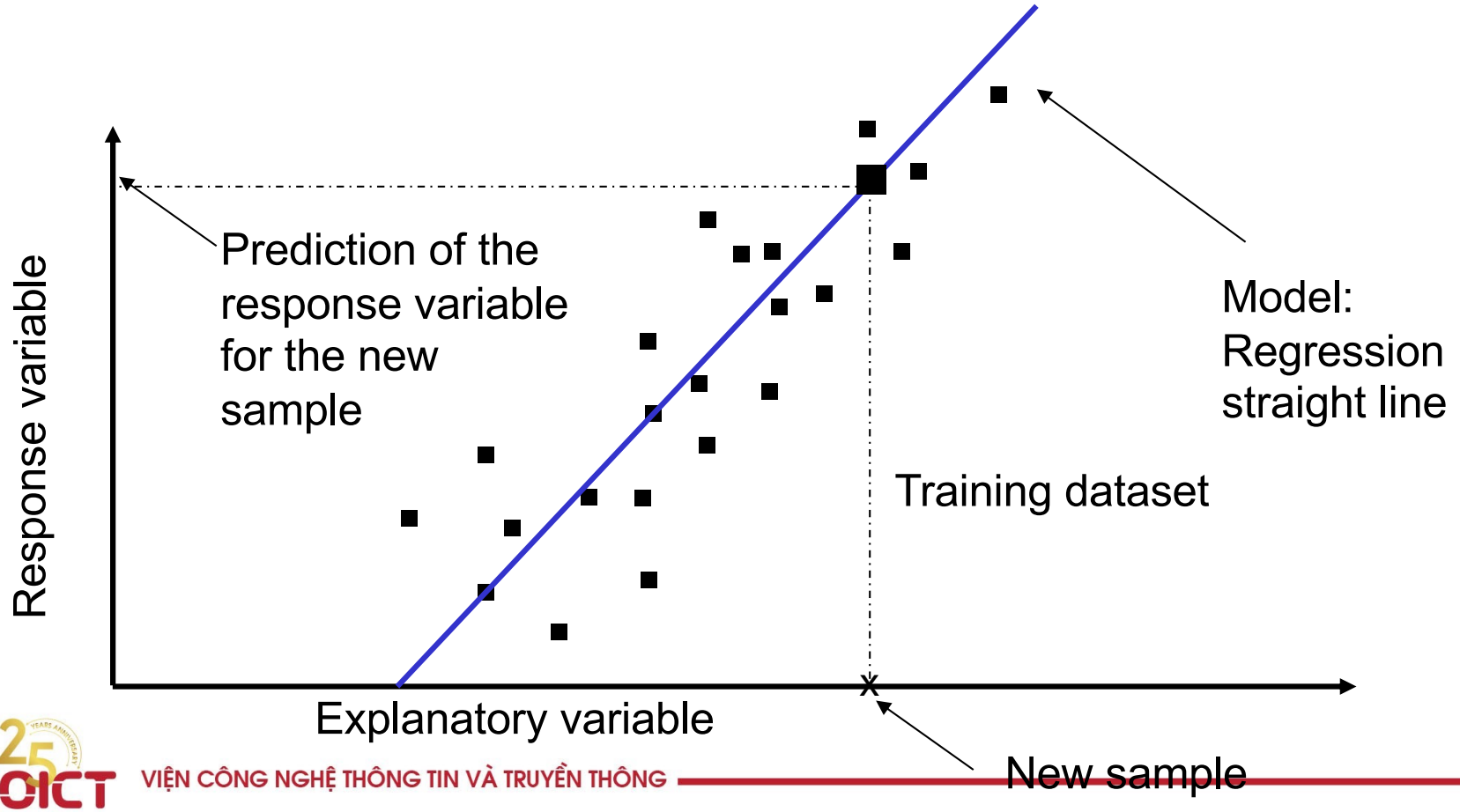
- **Classification** (Bayesian classifiers, classif trees, neural networks...)
  - Consists in assigning new data into pre-defined classes
  - The response variable is **categorical:** the class
  - The explanatory variables are the observed attributes

Quizz

- Let's consider the previous (bank) example
  - Is it a regression or a classification task?
    - Solution:
  - What is the response variable?
    - Solution:
  - What are the explanatory variables?
    - Solution:
  - What are the classes?
    - Solution:

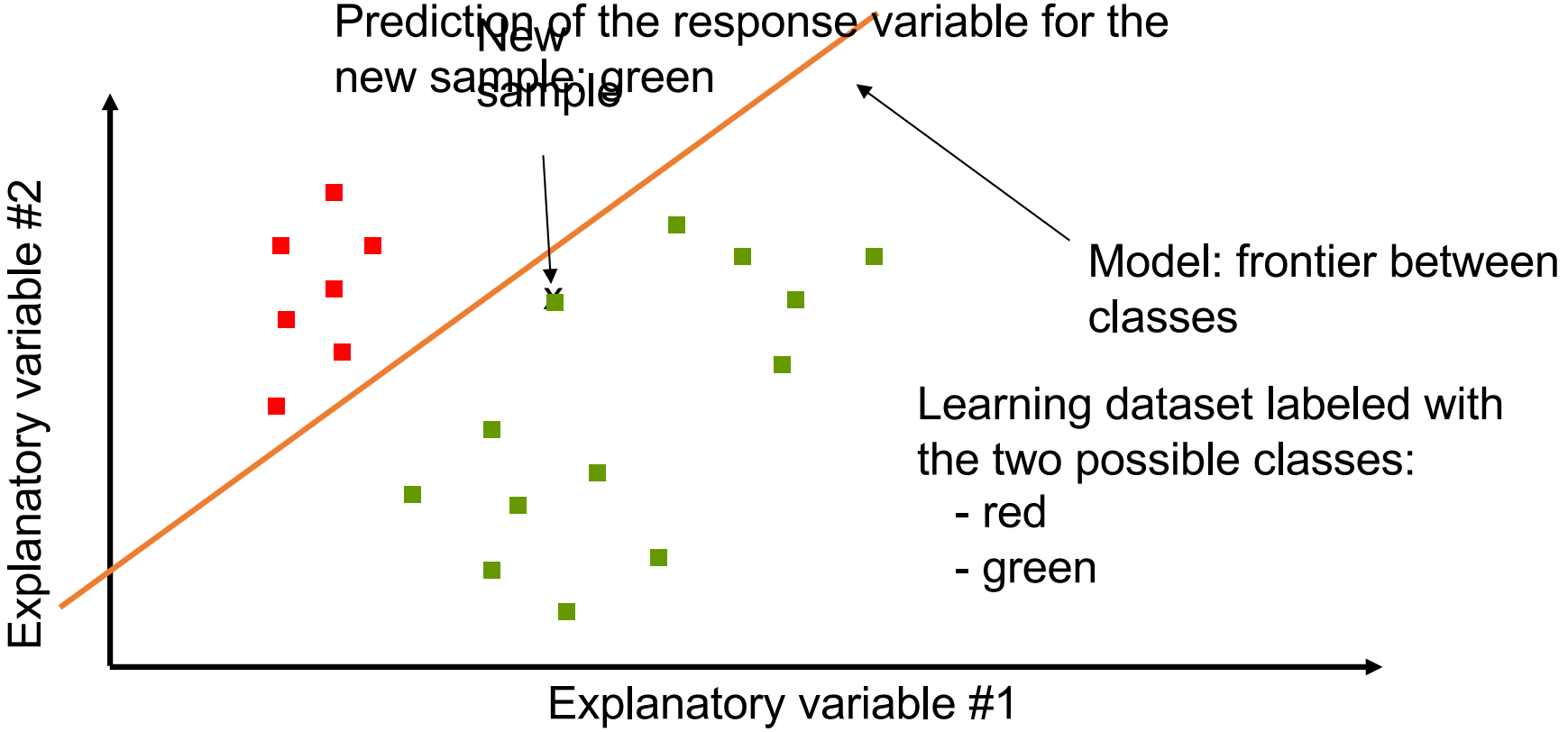
Supervised learning: regression

- Simple example: linear regression
  - With 1 explanatory variable and 1 response variable



Supervised learning: classification

- Simple example:
  - **2 explanatory variables** & **1 response variable** (with **2 classes**)



The model cannot always be expressed easily using frontiers
- it depends on the model (*e.g.* SVMs can, but not neural networks)

Machine learning

Unsupervised learning

Unsupervised learning

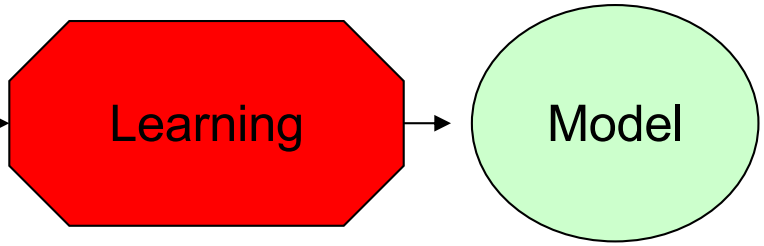
- **Unsupervised** learning methods
  - The learning dataset **does not contain** the values of the response variable(s)

Unsupervised learning

- **Unsupervised** learning methods
  - The learning dataset **does not contain** the values of the response variable(s)
  - Example

Step 1: Learning the model

Learning dataset		
Age	**Rent**	**Income**
36	0	1299
55	240	2500
40	768	3000
39	0	2000
44	334	512
26	631	722



	Customer Segments		
	Age	**Rent**	**Income**
Segment #1	[25 ; 45]	[250 ; 700]	<1000
Segment #2	[33 ; 60]	<400	>1200
Segment #3	[25 ; 55]	[500 ; 1000]	>2500

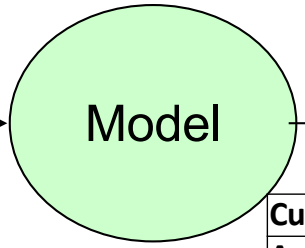
Unsupervised learning

- **Unsupervised** learning methods
  - The learning dataset **does not contain** the values of the response variable(s)
  - Example

Step 2: Applying the model

- to the training dataset…

Learning dataset		
Age	Rent	Income
36	0	1299
55	240	2500
40	768	3000
39	0	2000
44	334	512
26	631	722



	Customer Segments		
	Age	Rent	Income
Segment #1	[25 ; 45]	[250 ; 700]	<1000
Segment #2	[33 ; 60]	<400	>1200
Segment #3	[25 ; 55]	[500;1000]	>2500

Learning dataset			
Age	Rent	Income	Segment
36	0	1299	2
55	240	2500	2
40	768	3000	3
39	0	2000	2
44	334	512	1
26	631	722	1

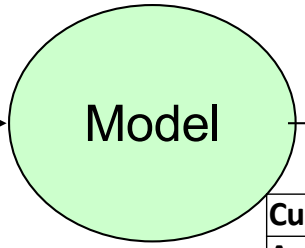
Unsupervised learning

- **Unsupervised** learning methods
  - The learning dataset **does not contain** the values of the response variable(s)
  - Example

Step 2: Applying the model

- to the training dataset… **and / or** to new data !

New sample			
Age	Rent	Income	Segment
30	260	900	???



New sample			
Age	Rent	Income	Segment
30	260	900	1

	Customer Segments		
	Age	Rent	Income
Segment #1	[25 ; 45]	[250 ; 700]	<1000
Segment #2	[33 ; 60]	<400	>1200
Segment #3	[25 ; 55]	[500;1000]	>2500

The model cannot always be expressed using data « slices »

- it depends on the model

Unsupervised learning: goals

- Unsupervised learning might be used for
  - Description
  - Association
  - Segmentation

Supervised classification

Objective

Focus of this chapter

- In the rest of this chapter, we will focus on **supervised** learning (not unsupervised learning)

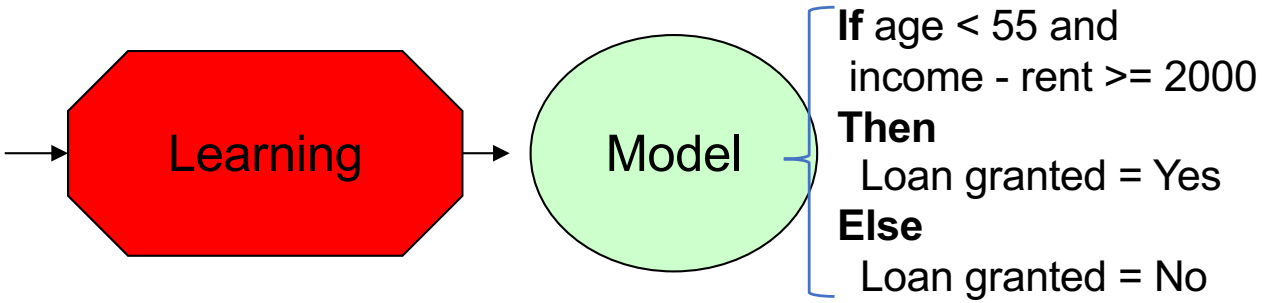
- More precisely, we will focus **on supervised classification** methods. Why?
  1. Because we don't have enough time to cover everything (this is only an intro course and there are thousands of methods)
  2. Because you will have a full ML course next semester
  3. Because supervised classification has most links with the previous chapters

Recall: supervised learning

- **Supervised** learning methods
  - The learning dataset contains the values of the response variable(s)
  - Example

Step 1: Learning the model

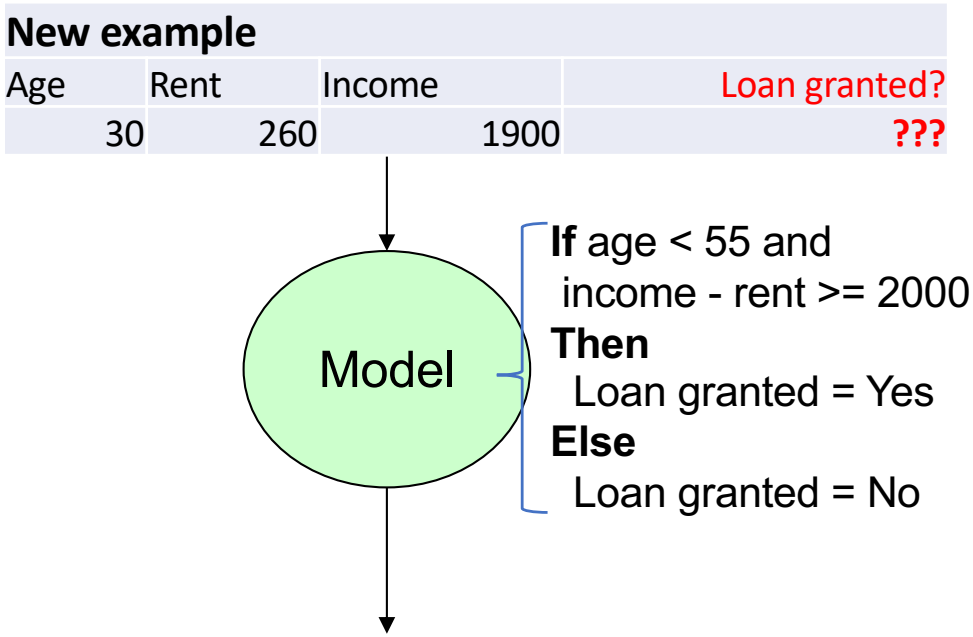
Learning dataset			
Age	Rent	Income	Loan granted?
36	0	1299	Yes
55	240	2500	No
40	768	3000	Yes
39	0	2000	Yes
44	334	512	No
26	631	722	No



Recall: supervised learning

- **Supervised** learning methods
  - The learning dataset contains the values of the response variable(s)
  - Example

Step 2: Applying the model



The model cannot always be expressed using rules
- it depends on the model
- *e.g.* decision trees give rules but neural networks don't

Supervised classification

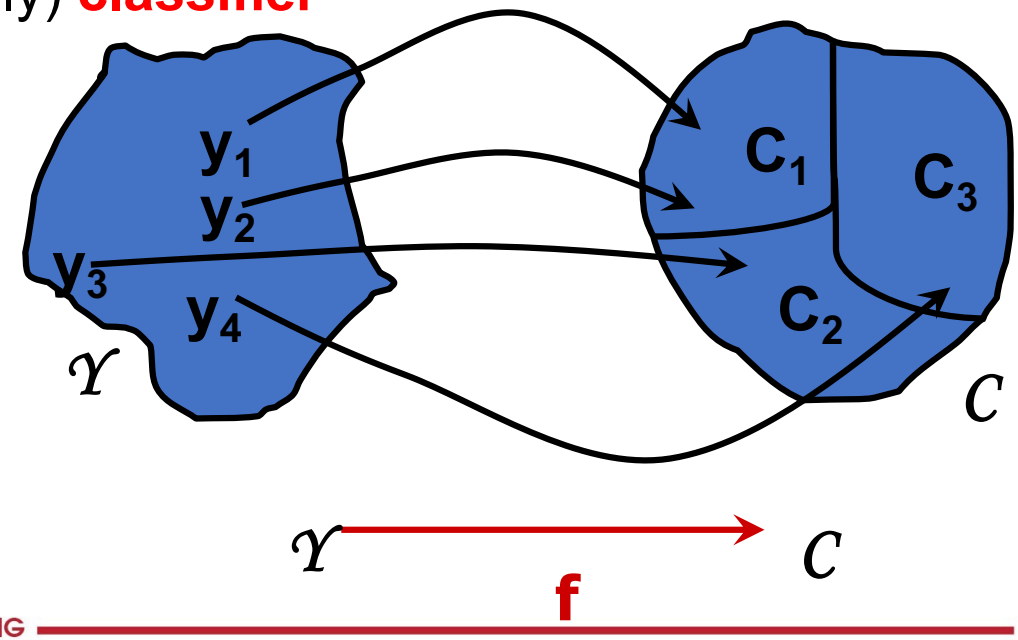
- Let $\boldsymbol{y}=(y^1,\ldots, y^p)^T$ be the vector of explanatory variables of an observation
- Let *c* be the class of the observation ***y*** (response variable)
- For each object in the Learning Dataset, we know the joint values of the *(**y**; c)* (by definition of supervised classification)
- For any new sample (to classify), we only know ***y***, and we want to infer its class *c*
- We’re aiming at learning a « **classifier** » *f* that allows:
  - to classify the training dataset as correctly as possible (fitting)
  - to best predict the class of new samples (generalization)

Supervised classification

- Therefore, we are looking for a function $f : \mathcal{Y} \to \mathcal{C}$, $Y \mapsto f(Y)$

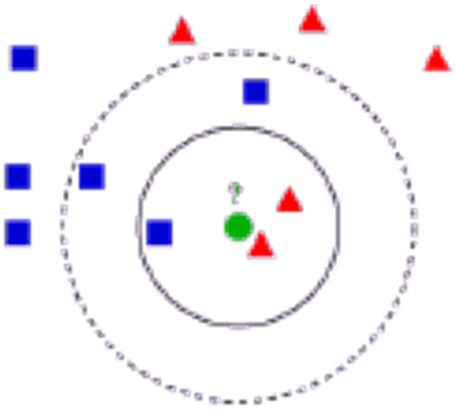
  - Such as *f* (***y***) = *c* with $\mathcal{C} = \{c_1, \ldots, c_k\}$ being the set of all possible classes (**modalities** of the response variable)

  - *f* is often called (wrongfully) **classifier**



Supervised classification: k-nn

- **Example** of a basic classifier: *k* nearest neighbours (*k*-nn)
  - The training dataset is labeled with the corresponding class
    - Blue square or red triangle
  - The new sample in green, $\boldsymbol{y}=(y^1,\ldots, y^p)^T$, with unknown class, is assigned to the majority class among its *k* nearest neighbours from the training dataset



k: classifier parameter
If k = 3, the new sample is labelled « triangle »
If k = 5, the new sample is labelled « square »

Supervised classification: k-nn

☺ Extremely simple classifier: does not even rely on learning a model

☺ Simple to implement

☹ Can be applied only on **numeric** explanatory variables (not to categorical variables)

☹ Sensitive to outliers

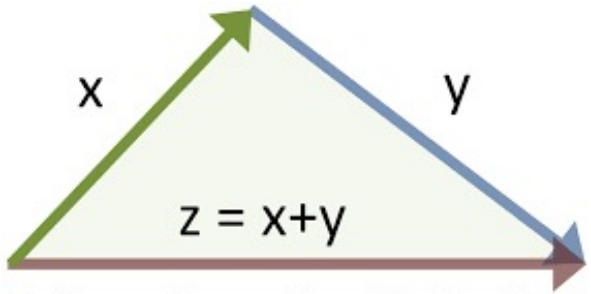
☹ **Very** dependent on its parameter *k*

☹ **Very** dependent on the distance used

Notion of similarity - dissimilarity

- Most classification methods require a similarity or dissimilarity measure
  - The similarity between x and y is high when x and y are close (in the representation space)
  - The dissimilarity between x and y is high when x and y are far (in the representation space)
  - A distance is a special case of a dissimilarity measure
    - Distances verify the triangle inequality, but dissimilarity measures do not necessarily verify it

Triangle inequality



$$\|z\| = \|x+y\| < \|x\|+\|y\|$$

Some useful distances / similarity measures

Minkowski distances:

$$D(x, y) = \left(\sum_{i=1}^{m} |x_i - y_i|^r \right)^{1/r}$$

- Euclidean distance

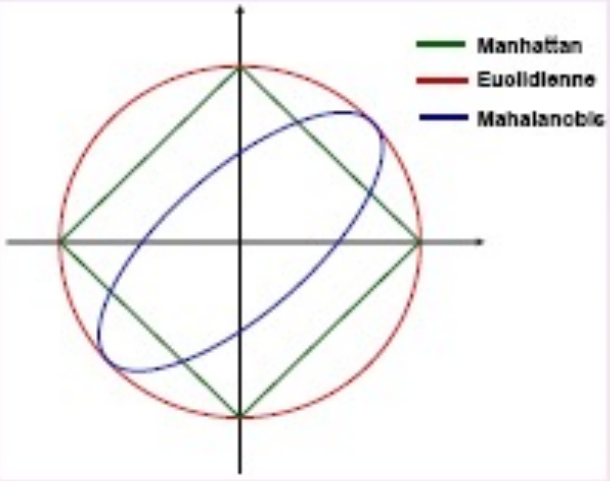
$$D(x, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$$

- Manhattan / city-block distance

$$D(x, y) = \sum_{i=1}^{m} |x_i - y_i|$$

Mahalanobis distance

$$D(x, y) = [\det V]^{1/m} (x - y)^T V^{-1} (x - y)$$ with *V*: covariance matrix



Unit circles:

circles with a radius of 1

Cosine similarity measure

$$s(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} = \frac{\sum x_i y_i}{\sqrt{\sum (x_i)^2} \times \sqrt{\sum (y_i)^2}}$$

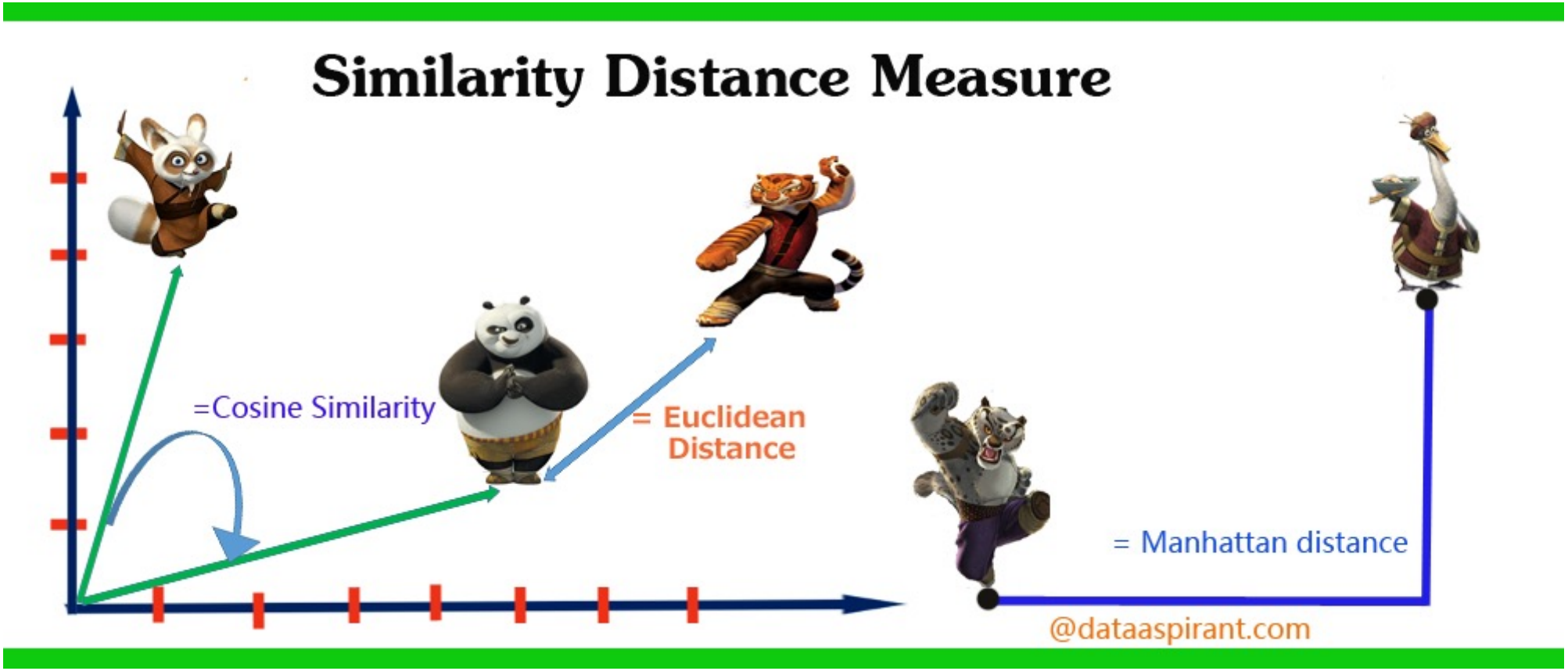
The computed similarity resides on the interval `[-1,1]`, where vectors with the same orientation have a similarity equal to `1`, orthogonal orientation a similarity equal to `0`, and opposite orientation a similarity equal to `-1`. The cosine distance seeks to express vector dissimilarity in positive space and does so by subtracting

Cosine distance

$$d(\mathbf{x}, \mathbf{y}) = 1 - s(\mathbf{x}, \mathbf{y})$$

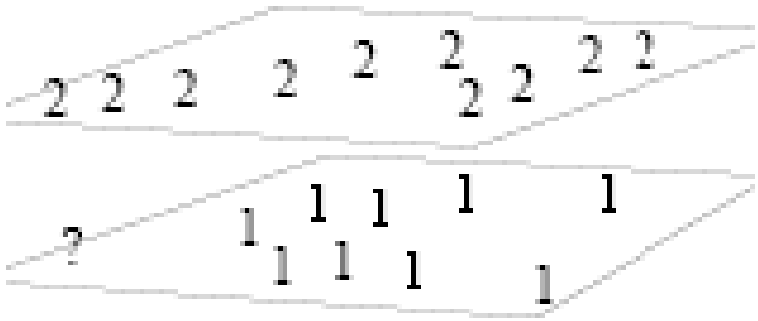
Some useful distances / similarity measures

- Summary (partial)
- Now, the main difficulty is to pick the distance / similarity that is best for our problem
    - Quite a tough question, but for later…



Quizz

- Let's consider the following example, in 3D, with 2 classes (1 & 2)
- The training data is labeled with its class (1 or 2)
- The new sample is marked as "?"
- Let us consider the 2 nearest neighbour classifier
- **Q1:** what is the predicted class for "?", if we use an Euclidean distance?
  - Solution:
- **Q2:** what is the predicted class for "?", if we use the Mahalanobis distance?
  - Solution:
- **Q3:** which distance should we choose, given the training data distribution?
  - Solution:



Supervised classification

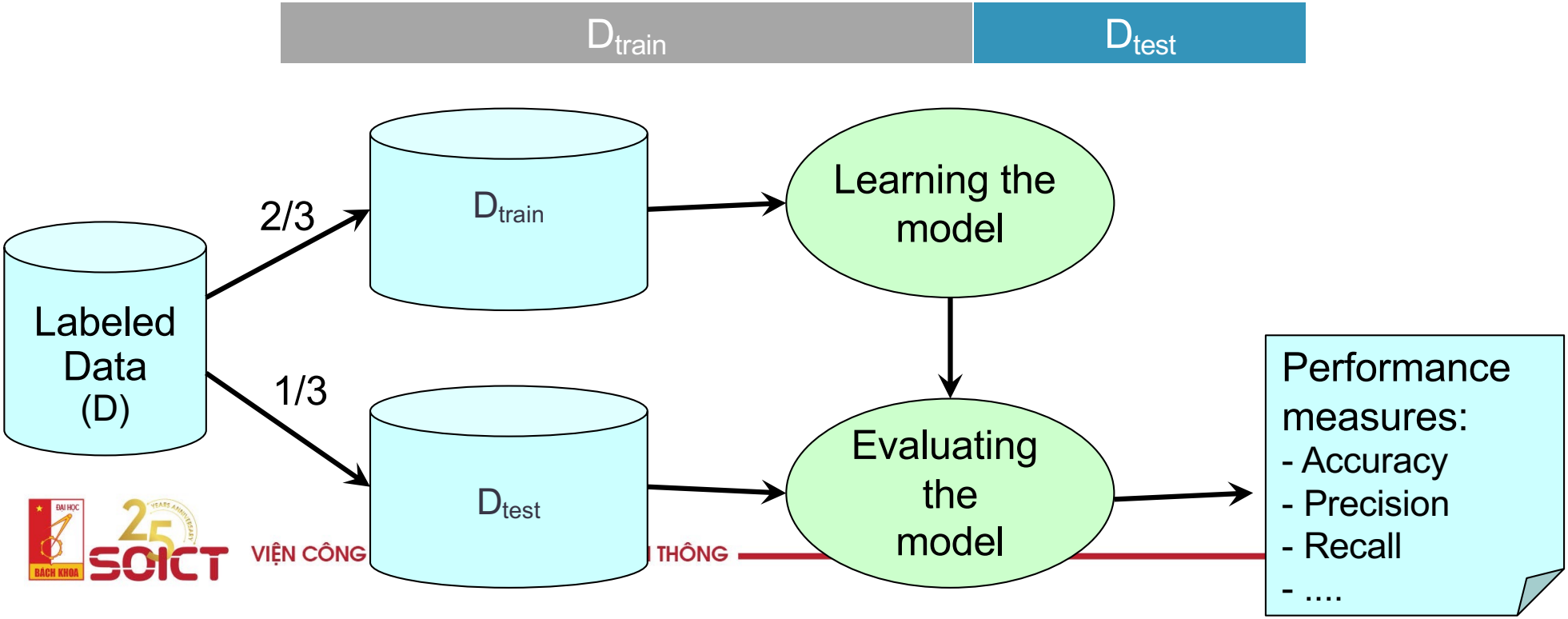
A few words about performance evaluation

Supervised classification – main issues

- In a real-world application, a classifier won't be able to properly classify all the observations
  - Real-life is not perfect!
- Hence the search for a **compromise** between **fitting** and **generalization**
- It is **easy** to assess the **fitting** performances of the classifier
  - by comparing the real class *c* with the class predicted *f(**y**)*...
  - ... for each example of the training dataset
- It is **more difficult** to assess the **generalization** capacity of the classifier
  - Indeed, on new samples, we know the explanatory variables ***y*** and we can infer the predicted class *f(**y**)…*
  - but we don't know the real class *c*!

Evaluation strategies: example of hold-out (random splitting)

- The observed, labeled dataset D is randomly split into 2 non-overlapping subsets:
  - D_{train}: used for training (learning dataset)
  - D_{test}: used to test performance (test dataset)



Evaluation measures: example of accuracy

- The most basic evaluation measure for supervised classification is accuracy:

  (# of well-classified records) / (total # of records)

  - Accuracy can be calculated:
    - On the training dataset (to evaluate the fitting performance)
    - On the test dataset (to evaluate the generalization performance)
  - One can define the classification error as 1-accuracy:
    - Classif. error= (# of wrongly-classified records) / (total # of records)

Evaluation measures: example of accuracy

- In any case, the model is trained (step 1) using the training dataset (data + labels), and evaluated (step 2) using the test dataset as follows:
  - Apply the model on the data in the test dataset
  - Collect the predicted labels
  - Compare it to the true labels of the data in the test dataset -> calculate the accuracy (and possibly other evaluation measures)

Supervised classification

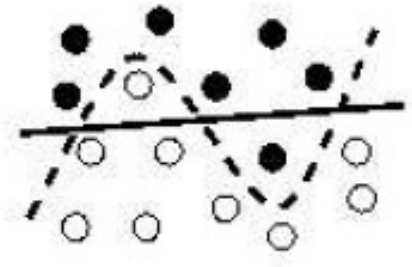
Risk of over-fitting

Supervised classification – over-fitting

- The issue of over-fitting occurs often when we don't have enough observations in the training dataset
  - If we try too hard to perfectly fit the classifier to the training data…
  - … then the classifier will be perfectly fitting the training data…
  - … but might not be good for new samples with the same distribution!!!
- The model has **learnt « by heart »** the learning dataset and is incapable of using its knowledge in a slightly different context
  - The learning dataset is just a sub-sample of the more general reality

Supervised classification – over-fitting

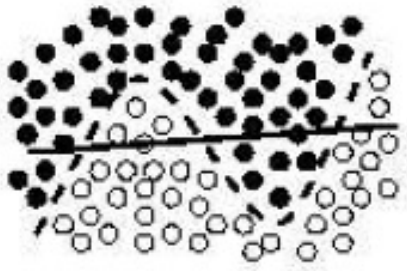
- Illustrative example:



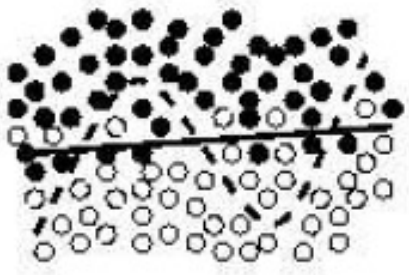
Training dataset:
Few observations
2 classes (white / black)

Problem: which is the best classifier (here frontier between classes)?
The dashed curve or the solid line?

- Well, it all depends on the underlying distribution of the data!



In this case, the dashed curve is better



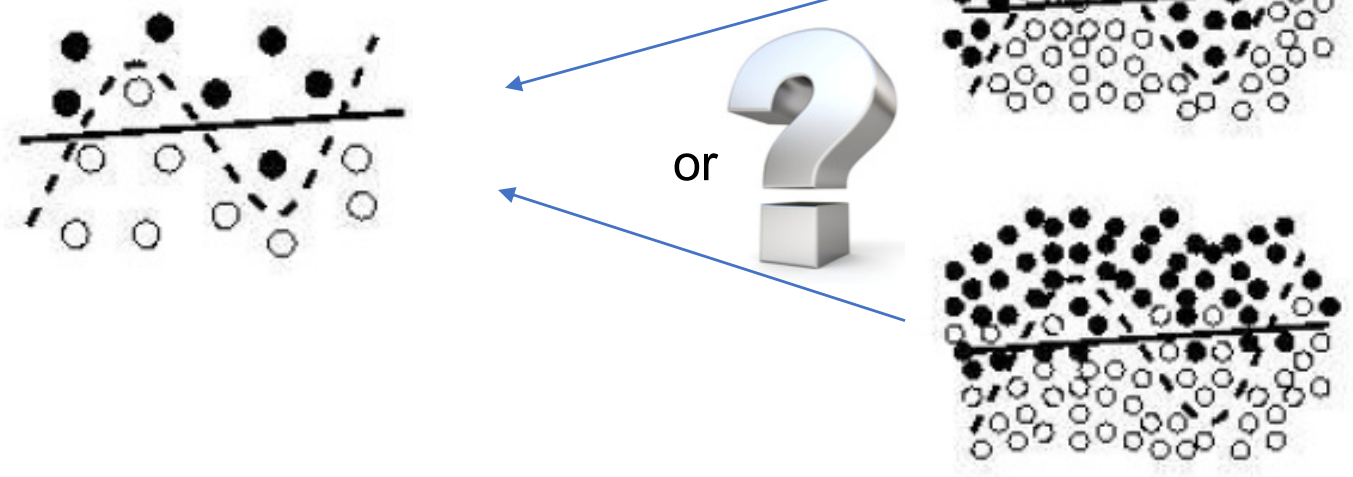
In this case, the solid line is better

Supervised classification – over-fitting

- Illustrative example:

Possible underlying distributions

Training dataset:
Few observations

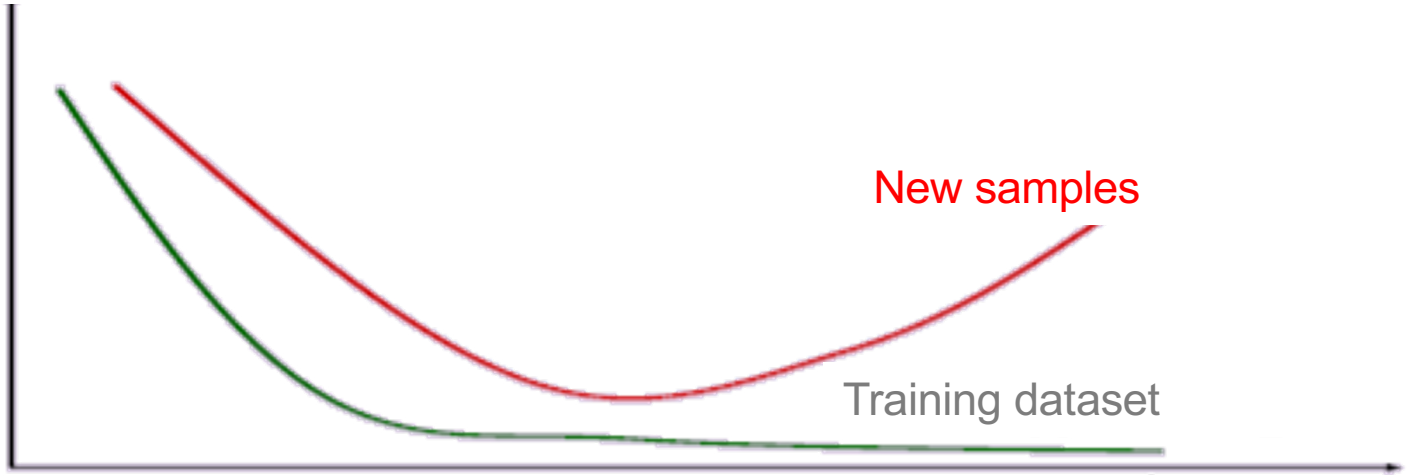


- Because we generally don't know the underlying distribution, it is sometimes better to choose a model that is simpler and does not fit perfectly the training data (here the solid line)
  - To avoid over-fitting (learning by heart) the training data

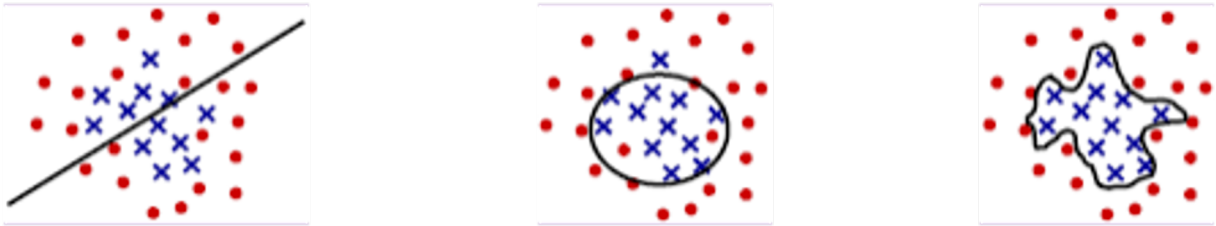
Supervised classification – over-fitting

- Another illustrative example:
  - Which classifier is the best?

Classification error rate



Training dataset with different classifiers



Supervised classification – over-fitting

- If we evaluate the performance of the classifier on the training dataset
  - Called apparent (or resubstitution) error rate
  - Not representative of the classifier's generalization capacity
    - Too optimistic
  - That's why we have to use a separate dataset for performance evaluation usually called **test dataset**)
    - *E.g.* a random subsample of the training dataset, that we **do not** use for learning the classifier
      - We only use it for assessing its generalization performances

Supervised classification

Methods

Supervised classification methods

- There are three main types of supervised classification methods:
  1. Probabilistic methods
     - Aim: to minimize the classification error rate based on probability estimates
     - For categorical or numeric attributes (assumptions on their distribution)
     - Examples: Maximum likelihood rule, Bayesian classifiers...
  2. Symbolic methods
     - Aim: to infer the most likely possible decision-making rules
     - Often, for categorical or « sliced » numeric attributes
     - Examples: decision trees, rule extraction...
  3. Statistical methods
     - Aim: to minimize an estimate of the classifier's generalization error rate
     - Often for quantitative attributes (assumptions about their distribution)
     - Examples: discriminant analysis, SVM, (deep) neural networks...

Supervised classification methods

- In this course, we will focus only on:
  - Probabilistic methods
    - Link with the last chapter about uncertain knowledge and reasoning
    - More specifically, we will focus on Bayesian classifiers
    - Recall: slide from Chapter 5:

Different possible frameworks for uncertain knowledge representation and reasoning

- Probabilistic Decision Theory
- Fuzzy logic
- Possibility theory
- Dempster–Shafer theory
- Info-gap decision theory
- …

Supervised classification

Probabilistic methods

Probabilistic methods

- Notations :
  - ***y***: vector of explanatory variables (description)
  - P[***y***]: probability of observing the description ***y***
  - P[*c*]: probability for an observation to belong to class *c*
  - P[*c*/***y***]: probability, for an observation with description ***y,*** to belong to class *c*:

$$P[c/y] = \frac{P[\ y\ \&\ c\ ]}{P[y]}$$

  - P[***y***/*c*]: probability, for an observation belonging to class *c,* to have description ***y***
  - Problem: we are looking for P[*c*/***y***] (this will define our classifier) but we cannot estimate it easily (unknown distribution of classes given the explanatory variables)
    - But, we can estimate P[***y***/*c*] on the training dataset!
  - So, we can use Bayes formula:

$$P[c/y] = \frac{P[y/c]P[c]}{P[y]}$$

Bayes classifier

- Bayes classifier:
  - To each observation ***y***, we assign the class *c* such that P[***y***/*c*]P[*c*] is maximum
  - Can be applied to either categorical or numeric variables, by making assumptions
    - *E.g.* the variable is Bernouilli or Gaussian
  - **Theorem**: the Bayes classifier is optimal on the training dataset
    - But, prone to overfitting…

Bayes classifier

- **Exercise**:
  - A (very clever) child wants to predict his Mom's answer when he'll ask her if he can play outside after school
  - He figured this depends a lot on if he finished his homework, or not
  - He collected data for 10 days, and made the assumption that « Homework » is a **Bernouilli variable**

Day	H: did I finish my homework before asking?	'c' : decision
1	1	Yes
2	1	Yes
3	1	Yes
4	1	Yes
5	1	Yes
6	0	Yes
7	0	No
8	0	No
9	1	No
10	0	No

Quizz

- Let's consider the previous example
  - What is the response variable?
    - Solution:
  - What are the explanatory variables?
    - Solution:
  - What are the classes?
    - Solution:

Bayes classifier

- **Exercise**:
  - Give the result of the Bayes classifier f_{Bayes} under the form $f_{Bayes}(H=0)=xx$ and $f_{Bayes}(H=1)=xx$, where xx must be replaced by « yes » or « no »
  - Give the apparent classification error = 1-accuracy (# of misclassified observations / # of observations) using this rule

Day	H: did I finish my homework before asking?	'c' : decision
1	1	Yes
2	1	Yes
3	1	Yes
4	1	Yes
5	1	Yes
6	0	Yes
7	0	No
8	0	No
9	1	No
10	0	No

Bayes classifier

- **Solution**:
  - $f_{Bayes}(D=0)$= ?? and $f_{Bayes}(D=1)$= ??
  - Apparent classification error =

Day	H: did I finish my homework before asking?	'c' : decision
1	1	Yes
2	1	Yes
3	1	Yes
4	1	Yes
5	1	Yes
6	0	Yes
7	0	No
8	0	No
9	1	No
10	0	No

Bayes classifier

- **Multi-variable Bayes classifier**
  - The child figures it does not only depends on his homework. So now, he wants to take into account other variables:
    - His Mom's mood (M), the weather (W), if he had a snack (S)

Day	H: did I finish my homework before asking?	M: is Mom in a good mood?	W: is it sunny (1) or rainy (0)?	S: Did I have a snack?	'c' : decision
1	1	1	1	1	Yes
2	1	0	1	1	Yes
3	1	0	1	0	Yes
4	1	0	1	0	Yes
5	1	1	1	0	Yes
6	0	1	1	1	Yes
7	0	0	0	0	No
8	0	1	1	0	No
9	1	1	0	1	No
10	0	0	1	1	No

Bayes classifier

- **Question:**
  - How many configurations are possible?
    - Solution:
  - Did we observe all of them in the learning set?
    - Solution:
  - What can we do then?
    - Is putting their estimated probabilities at 0 fair?
      - Solution:
    - **Solution: naive Bayes**

Naive Bayes classifier

- Naive Bayes Classifier
  - Assumption of **independence** of explanatory variables y_i **conditionally** to their class of belonging c:
  - For each y, we assign the class c s.t. P[y/c]P[c] is maximum, where

$$P[Y = y/c] = P[Y^1 = y^1/c]P[Y^2 = y^2/c] \dots P[Y^p = y^p/c]$$

  - Simple to implement, relatively good performance
  - Can solve non-linearly separable (so relatively complex) classification problems
  - Based on a generally **false** hypothesis
    - But often gives good results in practice
    - Serves as a benchmark for evaluating more elaborate methods

Homework

Homework

- By hand, calculate / give the Bayes rule for the couple of explanatory variables H and W
  - Assignment on Teams

Day	H: did I finish my homework before asking?	M: is Mom in a good mood?	W: is it sunny (1) or rainy (0)?	S: Did I have a snack?	'c' : decision
1	1	1	1	1	Yes
2	1	0	1	1	Yes
3	1	0	1	0	Yes
4	1	0	1	0	Yes
5	1	1	1	0	Yes
6	0	1	1	1	Yes
7	0	0	0	0	No
8	0	1	1	0	No
9	1	1	0	1	No
10	0	0	1	1	No

Summary

Summary of chapter 6

- Learning can serve for multiple final objectives
  - Description
  - Prediction/estimation
  - Association
  - Segmentation
- Learning can be **supervised** or not
  - Supervised learning can be used mostly for prediction / estimation
  - Unsupervised learning can be used for description, association, segmentation
- In this lecture, we focused on supervised learning
  - Supervised learning is used mainly for 2 tasks:
    - Regression (response variable is numeric)
    - Classification (response variable is categorical)
  - In this lecture, we focused mainly on **supervised classification**
    - Because it's the most linked to the previous chapters
    - Main risk associated with supervised classification: **over-fitting**

      -> Need to use separate datasets for training and evaluation

Summary of chapter 6

- Different **supervised** methods for classification:
  - Probabilistic methods
    - Pros: if D_{train} is very representative of the data, gives very good results
    - Cons: risk of over-fitting (bad generalization to new data)
  - Symbolic methods
    - Pros: very easily interpretable by a human
    - Cons: limited performance (especially with quantitative variables)
  - Statistical methods
    - Pros: excellent performance (especially on numeric data)
    - Cons: usually, lack of interpretability (black box models)
  - Many possible evaluation strategies and measures: depends mainly on the volume of data and your application
  - There is no « best » method overall: everything depends on your data (volume, distribution, etc.) and your final objective

Chapter 6

Questions

ĐẠI HỌC BÁCH KHOA

25 YEARS ANNIVERSARY SOICT

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Thank you for your attention!

soict.hust.edu.vn/ fb.com/groups/soict