Data Mining
(Know the past to predict the future!)

MSc. Bui Quoc Khanh

khanhbq@hanu.edu.vn

Teaching Material

- Main reference books
  - T.M. Mitchell, MachineLearning, McGraw-Hill, 1997
  - J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann
  - P.N. Tan, M. Steinbach, V. Kumar: Introduction to Data Mining, Addison Wesley, 2006

Outline of the course

- Introduction to DM
- Concept learning
- Elements of probability theory
- Entropy
- Decision Trees
- Rule learning
- Naïve Bayes classifiers
- K-NN classifiers
- Text classification
- Ensemble methods
- Association rules
- Clustering –K-means

Outline of the introduction

- Motivations for DM
- Induction vs Deduction
- DM and the KDD Process
- Typical DM application

Motivations for DM -The Big Data Explosion

WE LIVE IN THE AGE OF DATA! EVERY PURCHASE WE MAKE IS DUTIFULLY RECORDED. EVERY MONEY TRANSACTION IS CAREFULLY REGISTERED. EVERY WEB CLICK ENDS UP IN A WEB CLICK ARCHIVE. WE HAVE DATA AVAILABLE LIKE NEVER BEFORE!

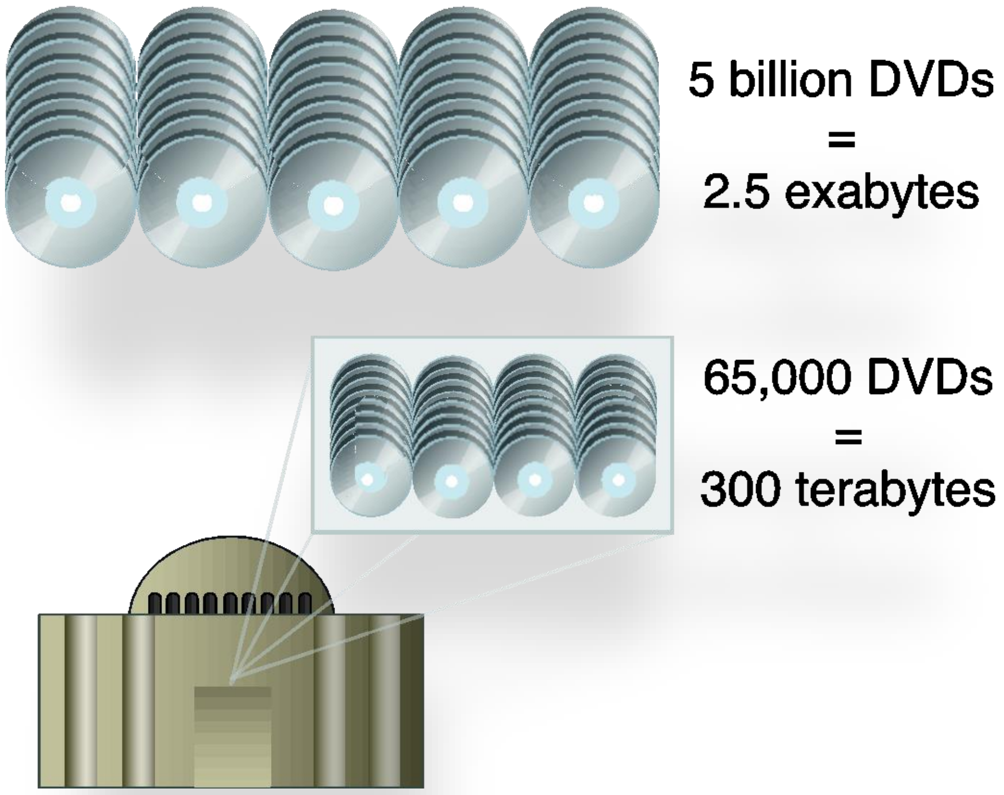
EVERY MINUTE, WE SEND 204 MILLION EMAILS, GENERATE 1.8 MILLION FACEBOOK LIKES, SEND 278 THOUSAND TWEETS, AND UPLOAD 200 THOUSAND PHOTOS TO FACEBOOK

DIFFERENT DATA TYPES: TUPLES, TEXTS, IMAGES, TEMPORAL, SPATIAL, ETC.

BIG DATA -THE 3 VS: VOLUME, VARIETY AND VELOCITY

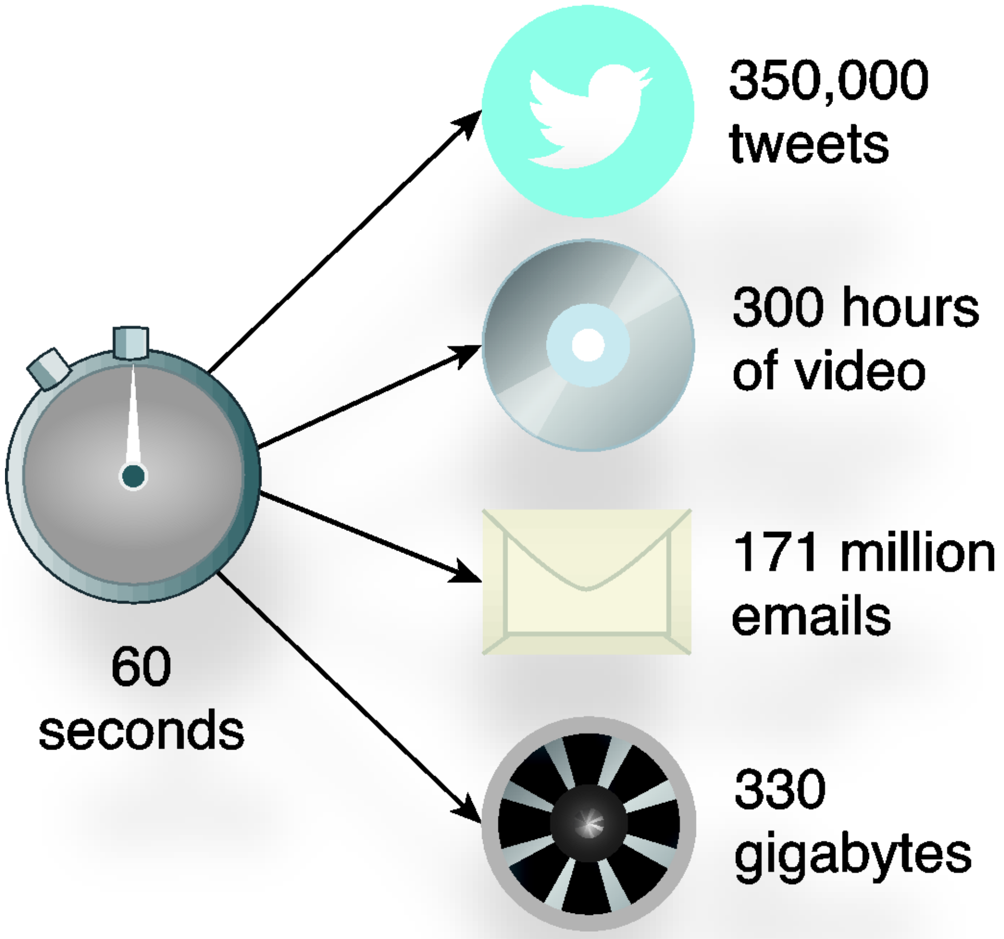
3 Vs of Big Data - Volume

- Organizations and users world-wide create over 2.5 EBs of data a day. As a point of comparison, the Library of Congress currently holds more than 300 TBs of data.



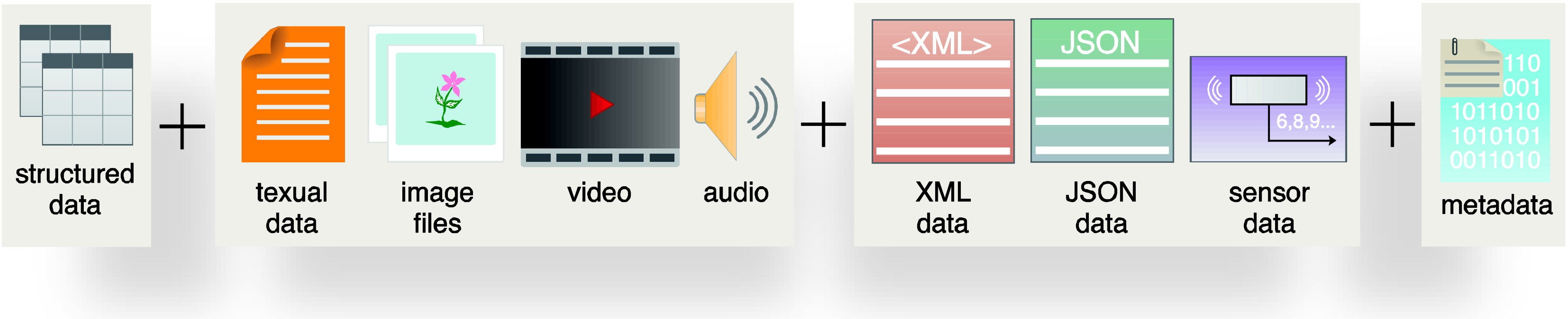
3 Vs of Big Data - Velocity

- Examples of high-velocity Big Data datasets produced every minute include tweets, video, emails and GBs generated from a jet engine.



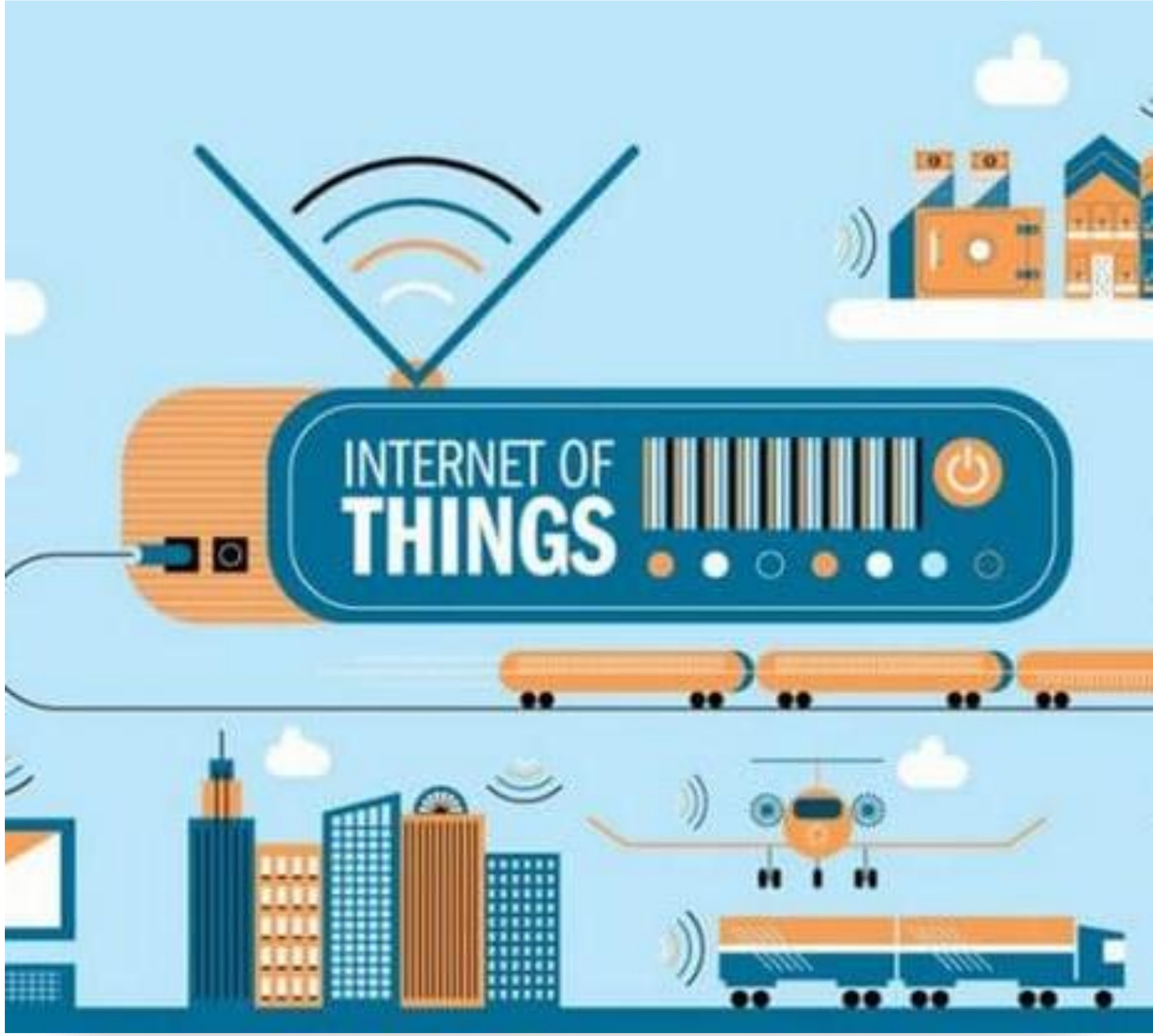
3 Vs of Big Data – Variety

- Examples of high-variety Big Data datasets include structured, textual, image, video, audio, XML, JSON, sensor data and metadata



Motivations for DM - The Internet of Things

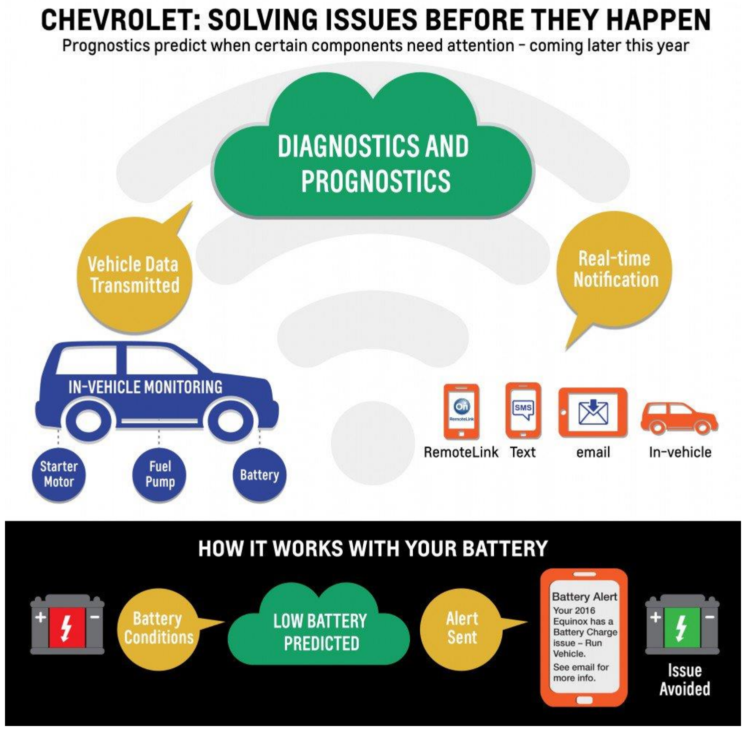
- The Internet of Things (IoT):
  - The Internet of Things (IoT) is the network of physical objects—devices, vehicles, sensors —that enables these objects to collect and exchange data
  - Train wagons are equipped with thousands of sensors generating data at a frequency of ms
  - The amount of devices that connect to the internet will rise from about 13 billions today to 50 billions by 2020



Motivations for DM -The Internet of Things

Wear able Tech

Smart Appliances

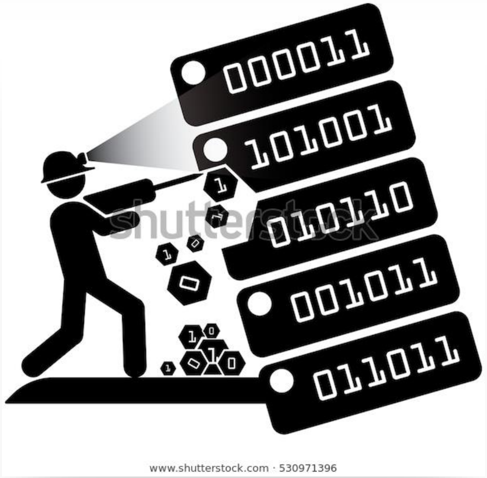

CHEVROLET: SOLVING ISSUES BEFORE THEY HAPPEN
Prognostics predict when certain components need attention – coming later this year
DIAGNOSTICS AND PROGNOSTICS
Vehicle Data Transmitted
Real-time Notification
IN-VEHICLE MONITORING
Starter Motor
Fuel Pump
Battery
RemoteLink
Text
email
In-vehicle
HOW IT WORKS WITH YOUR BATTERY
Battery Conditions
LOW BATTERY PREDICTED
Alert Sent
Battery Alert
Your 2016 Equinox has a Battery Charge issue – Run Vehicle.
See email for more info.
Issue Avoided

Motivations for DM

- Such large amount of data contains valuable knowledge which is hidden and implicit
- The critical challenge is today using this data to extract valuable information from it
- How can we do to extract such a knowledge?
- Data Mining (also called Knowledge Discovery) techniques from large amounts of data are needed

What is DM?

- A miner is a person who extracts coal, or other minerals from the earth
- Data miner is a person who extracts knowledge from data



www.shutterstock.com · 530971396

What is DM?

Data mining (DM):

technology (algorithms, methods, tools) enabling the extraction of hiddenand interestingknowledge(rules, regularities, patterns, constraints, …) from large amounts of data of different types and formats

technology aimed at finding correlations(or patterns) within data

DM relies on strong theoretical/mathematical foundations

Artificial Intelligence: Machine Learning & Logics

Statistics

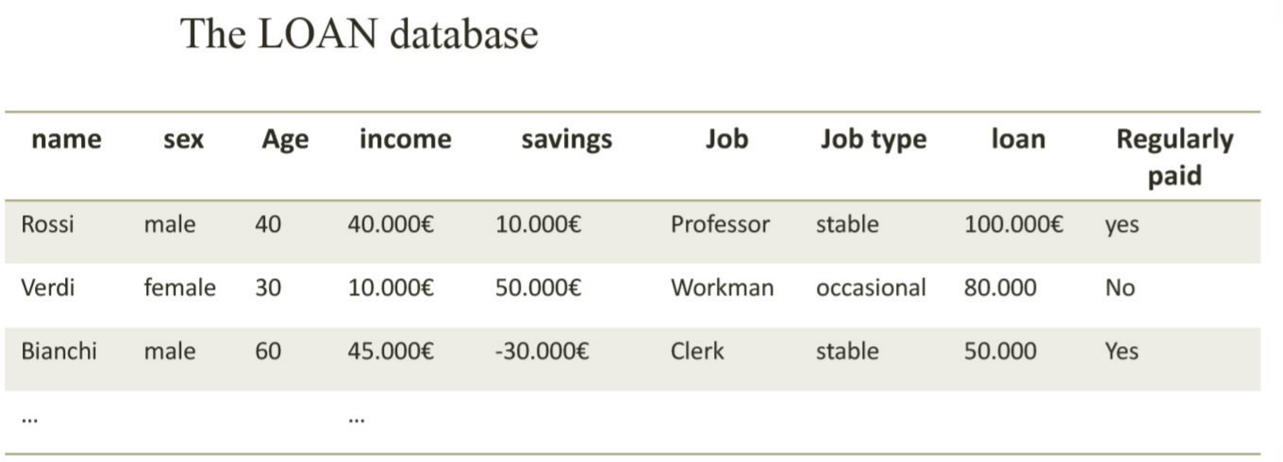
Database management systems

An example
Credit Risk Assessment

- Each bank owns a credit database storing all information about the past credit operations, e.g.,
  - MrRossi got a loan on 2002 of 100.000€
  - He paid the loan in 10 years
  - Payment was regular
  - MrRossi earns 30.000€/year, has a stable job, owns the apartment where he lives, is married, ….

An example
Credit Risk Assessment

- We can use SQL to query such a DB, for instance, about customers who paid regularly
- However, SQL can provide us with information about the past
- So what when a NEW customer asks for a loan?



The LOAN database

name	sex	Age	income	savings	Job	Job type	loan	Regularly paid
Rossi	male	40	40.000€	10.000€	Professor	stable	100.000€	yes
Verdi	female	30	10.000€	50.000€	Workman	occasional	80.000	No
Bianchi	male	60	45.000€	-30.000€	Clerk	stable	50.000	Yes
…			…					

An example
Credit Risk Assessment

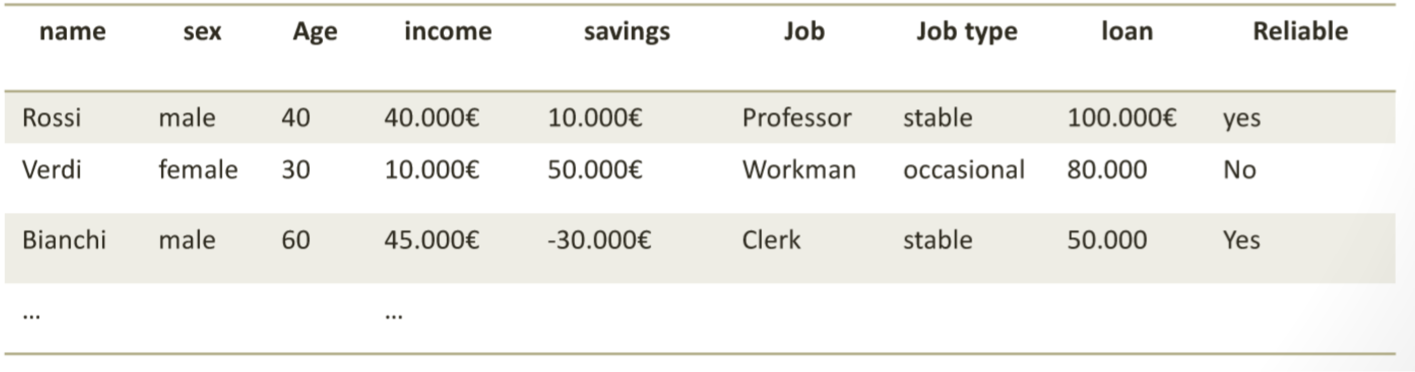
- To this end, a MODELfor concept“reliable customer” is needed, i.e., a description of the properties a customer should hold in order to be considered reliable (from the bank viewpoint)
    - What are the properties a customer should exhibit in order to be classified as “reliable”?
    - What is the correlationbetween reliability, from one side, and the other properties of customers?

An example
Credit Risk Assessment

- In principle, one may ask the bank manager to provide e model based on his personal experience, e.g.,
  - “a reliable customer as one who earns a high salary, has a stable job, has no litigation with the bank, etc.”
- Drawback: subjectivity, limited knowledge of the application domain, etc
- Another approach is that of learning from data (i.e., examples describing real past experience) -data does not lie! Let us learn from data!
- Top down (deductive) approach vs bottom up (inductive) approach
- Automatic learning

An example Credit Risk Assessment

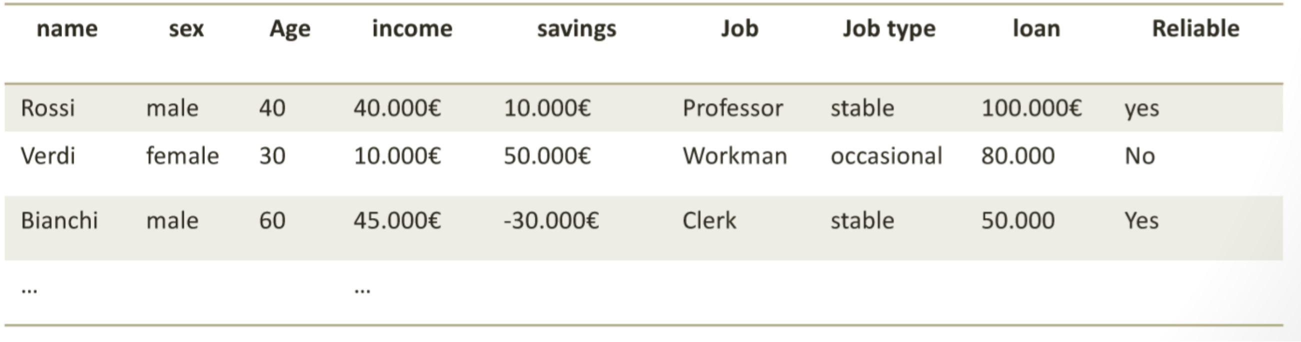
- We learn from the LOAN database –the training set
- A customer who paid back regularly the loan is seemed reliable
- So the tuples where Reliable = yes are the positive examples
- What is the Profile(or Model) of a reliable customer?



name	sex	Age	income	savings	Job	Job type	loan	Reliable
Rossi	male	40	40.000€	10.000€	Professor	stable	100.000€	yes
Verdi	female	30	10.000€	50.000€	Workman	occasional	80.000	No
Bianchi	male	60	45.000€	-30.000€	Clerk	stable	50.000	Yes
…			…					

An example Credit Risk Assessment

- Let us look at the properties shared by positive examples and not by negative ones
- Based on them, we may induce(or learn), say, the following model
  - Sex=male AND age≥40 AND Income≥40.000 AND job type=stable => Reliable



name	sex	Age	income	savings	Job	Job type	loan	Reliable
Rossi	male	40	40.000€	10.000€	Professor	stable	100.000€	yes
Verdi	female	30	10.000€	50.000€	Workman	occasional	80.000	No
Bianchi	male	60	45.000€	-30.000€	Clerk	stable	50.000	Yes
…			…					

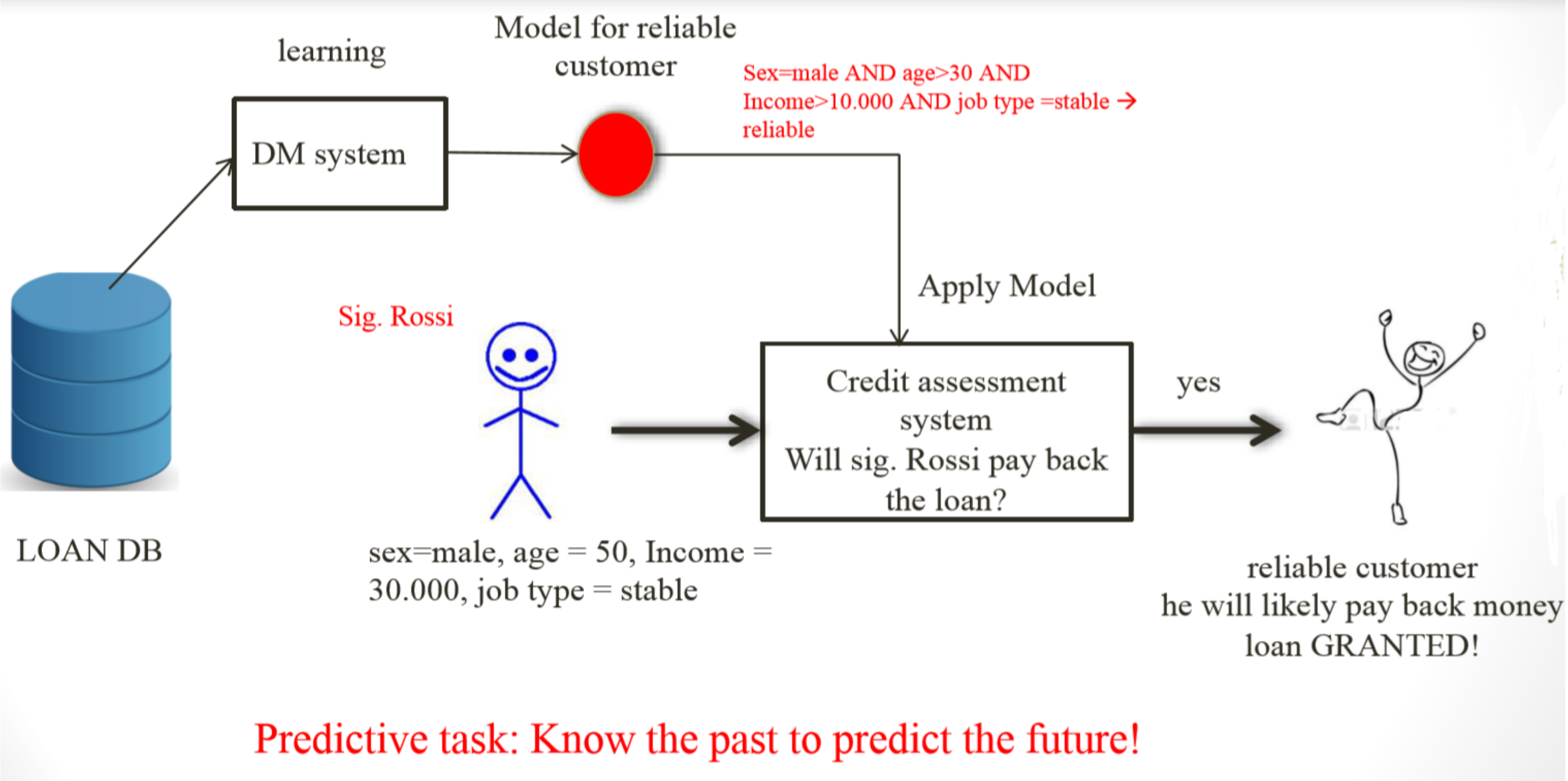
An example
Credit Risk Assessment

- The model
  - Sex=male AND age≥40 AND Income≥40.000 AND job type=stable => Reliable
- describes a relationship between the value of the attribute “reliable” (Yes or Not) and the values of the other attributes
- It enables to predict the value of the attribute “reliable” (Yes or Not) based on the values of the other attributes

An example
Credit Risk Assessment

- What do we do with the learned model?
- We can use it to predict whether a new bank customer asking for a loan will eventually pay back the money he got from the bank

An example
Credit Risk Assessment



learning

DM system

Model for reliable customer

Sex=male AND age>30 AND Income>10.000 AND job type =stable → reliable

Apply Model

LOAN DB

Sig. Rossi

sex=male, age = 50, Income = 30.000, job type = stable

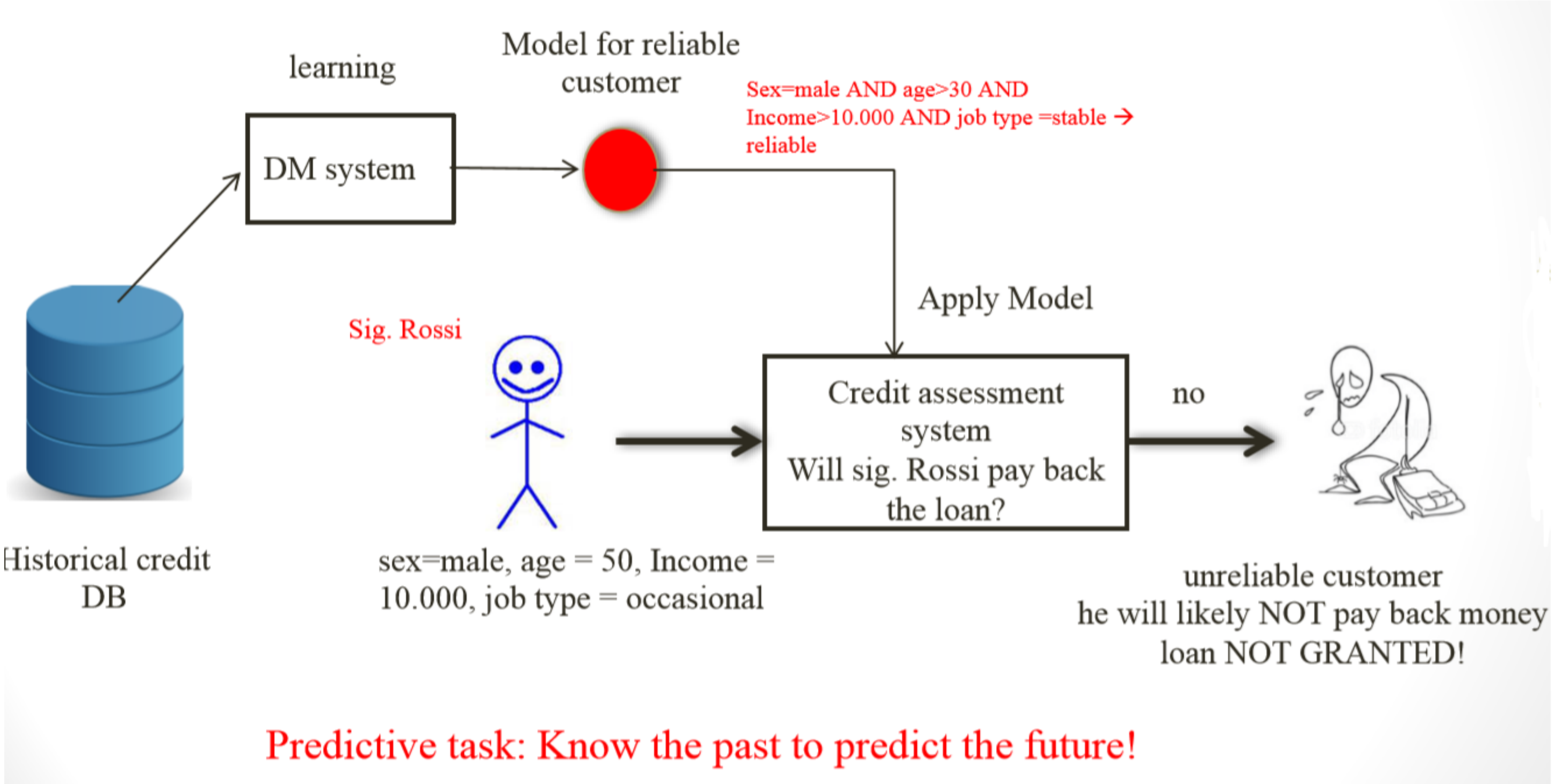
Credit assessment system
Will sig. Rossi pay back the loan?

yes

reliable customer
he will likely pay back money
loan GRANTED!

Predictive task: Know the past to predict the future!

An example
Credit Risk Assessment



learning

Model for reliable customer

Sex=male AND age>30 AND Income>10.000 AND job type =stable → reliable

DM system

Apply Model

Sig. Rossi

Credit assessment system
Will sig. Rossi pay back the loan?

no

Historical credit DB

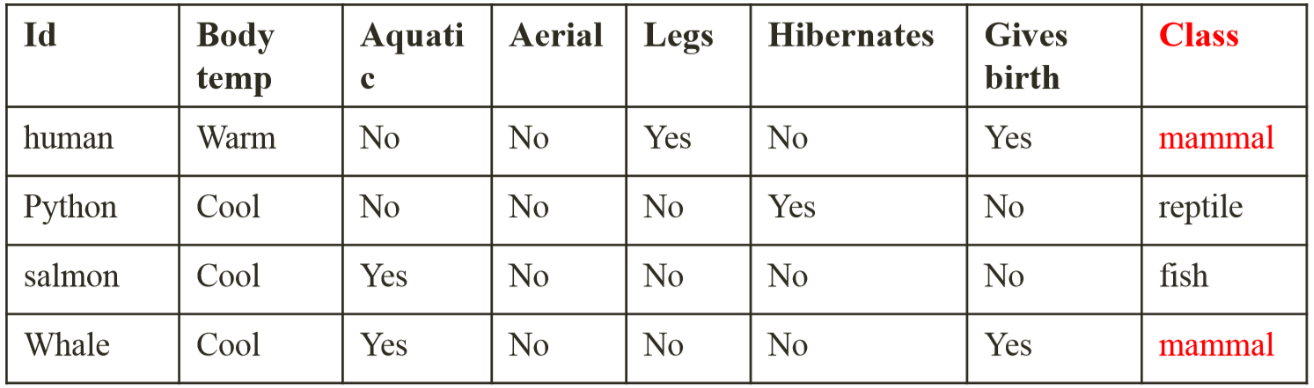
sex=male, age = 50, Income = 10.000, job type = occasional

unreliable customer
he will likely NOT pay back money
loan NOT GRANTED!

Predictive task: Know the past to predict the future!

The Mammal example

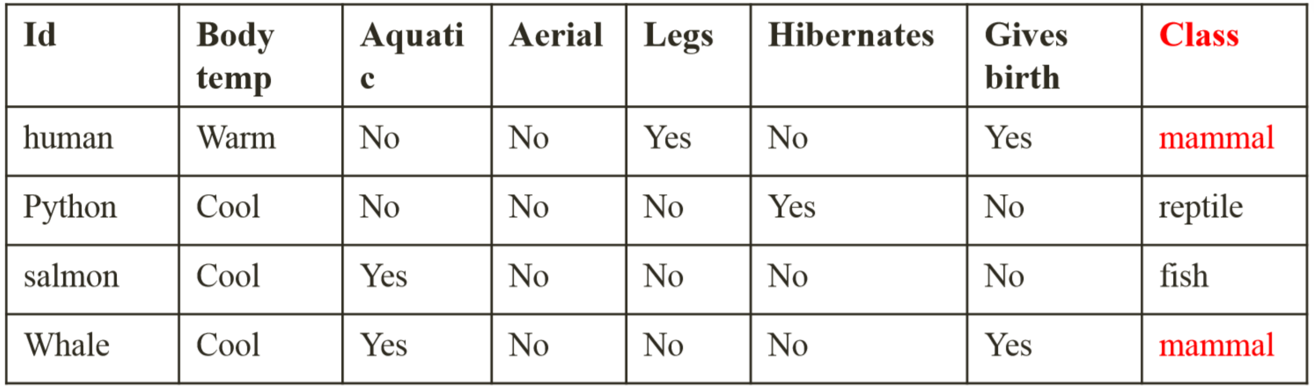
- We want learn the concept of mammal from our database (set of both positive and negative examples) –what are the properties characterizing mammals
- Mammal if aerial=no and hibernates =no and gives birth=yes
- According on our data, a mammal is a living being which is not aerial, does not hibernate and gives birth (no matter for the other attributes)
- This model is consistent with the training data



Id	Body temp	Aquatic	Aerial	Legs	Hibernates	Gives birth	Class
human	Warm	No	No	Yes	No	Yes	mammal
Python	Cool	No	No	No	Yes	No	reptile
salmon	Cool	Yes	No	No	No	No	fish
Whale	Cool	Yes	No	No	No	Yes	mammal

The Mammal example

- A turtle is <cool, yes, no, yes, no, no> -Is the turtle a mammal?
- Turtle does not belong to the training set –no SQL query
- Mammal if aerial=no and hibernates =no and gives birth=yes
- Since turtle does not satisfy the model, the answer is NO (according to our model of mammal)



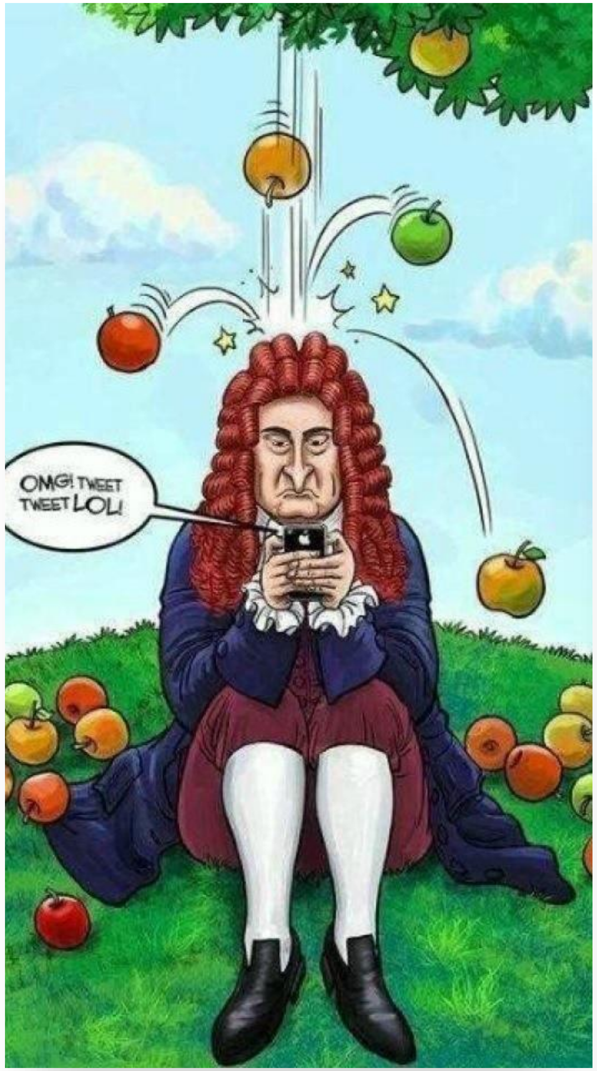
Id	Body temp	Aquatic	Aerial	Legs	Hibernates	Gives birth	Class
human	Warm	No	No	Yes	No	Yes	mammal
Python	Cool	No	No	No	Yes	No	reptile
salmon	Cool	Yes	No	No	No	No	fish
Whale	Cool	Yes	No	No	No	Yes	mammal

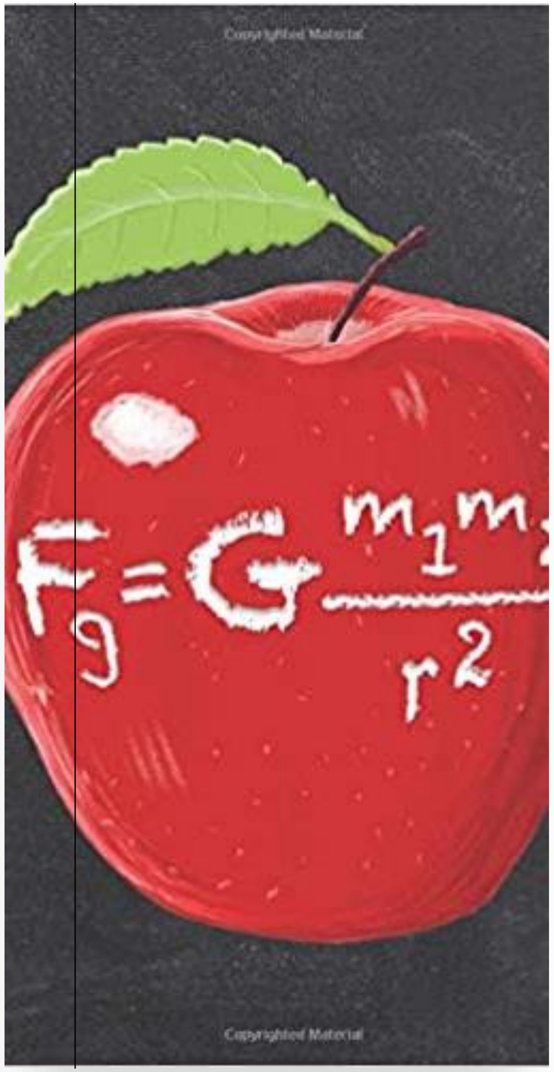
Learning Algorithms

- A learning algorithm is trained over the training data (e.g., LOAN database, MAMMAL database) to learn the concept (e.g., “reliable customer” or “mammal”)
- The learned concept (or model) is then used to classify new unknown instances
- The learned model is the knowledge we have extracted from the data describing past experience

Data Mining and Induction

- Induction: process of creating general theories from observed data (empirical observations)



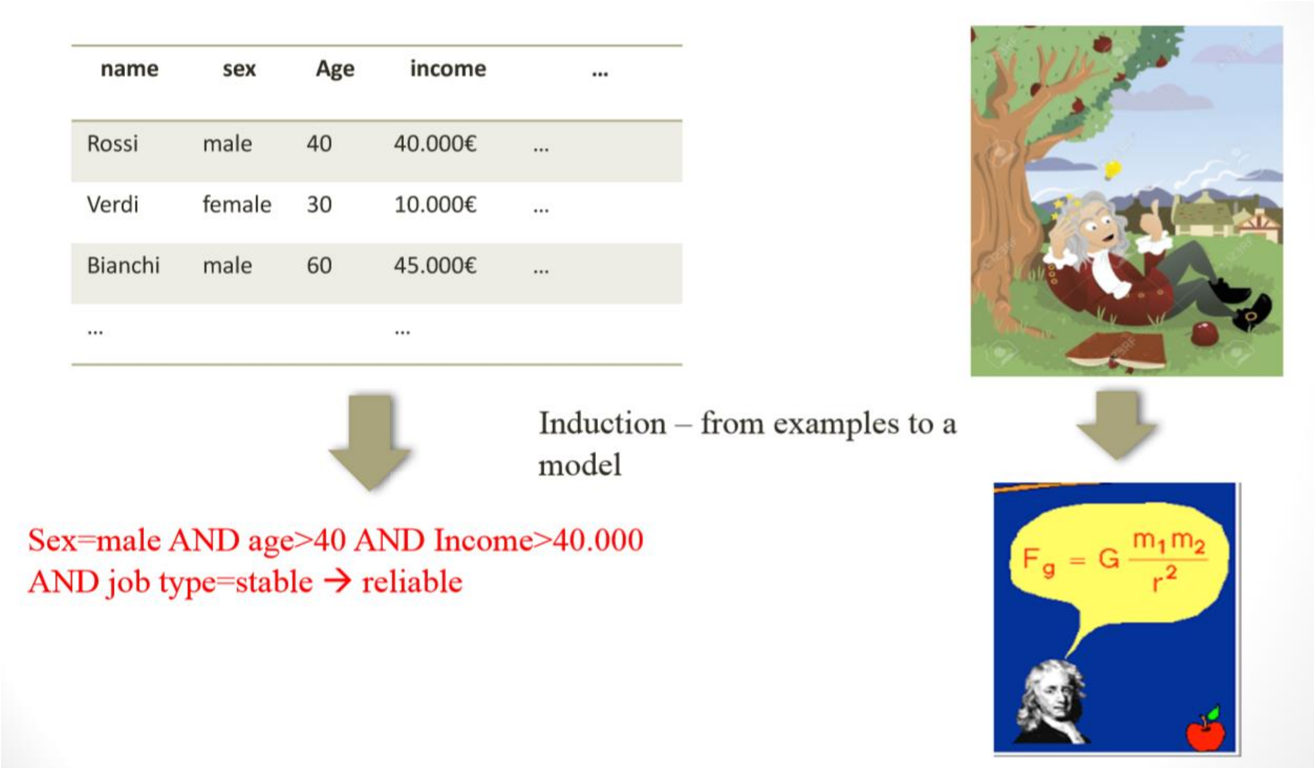


DM and Induction

- DM is an inductivetask, as it extracts general theories from observed data (empirical observations or examples), e.g., the model of the “reliable bank customer” from the LOAD database or that of “mammal” from the LIVING BEINGS database
- Purely inductive learning methods formulate general hypotheses by finding empirical regularities over the training examples.’ (Tom M. Mitchell,1997,p334)

Induction

- Inducing a model from empirical observations



name	sex	Age	income	...
Rossi	male	40	40.000€	...
Verdi	female	30	10.000€	...
Bianchi	male	60	45.000€	...
...			...	

Induction – from examples to a model

Sex=male AND age>40 AND Income>40.000 AND job type=stable → reliable

$$F_g = G\frac{m_1 m_2}{r^2}$$

Induction vs Deduction

Induction: from particular to general

Extracting general theories from observed data (DM)

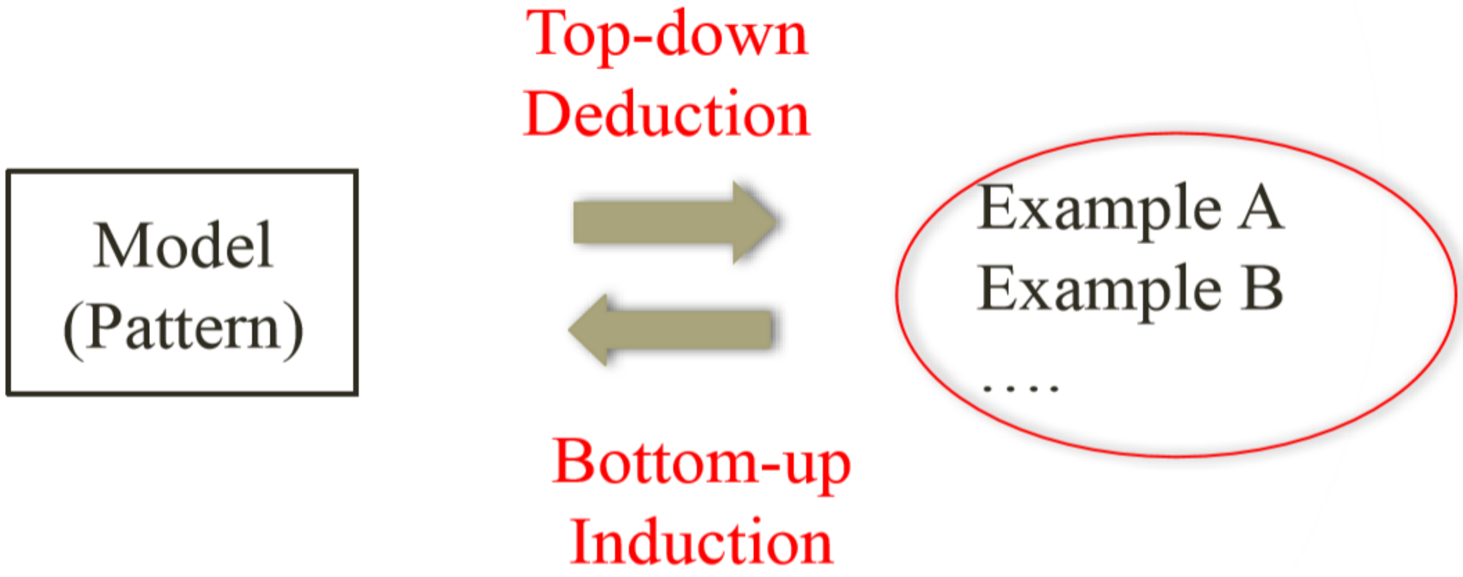
Experimental sciences (physics, biology, etc.)

Deduction: from general to particular

Deducing theorems from general theories (logic programming)

Mathematics

Induction vs Deduction



Logic Programming and Deduction -an example

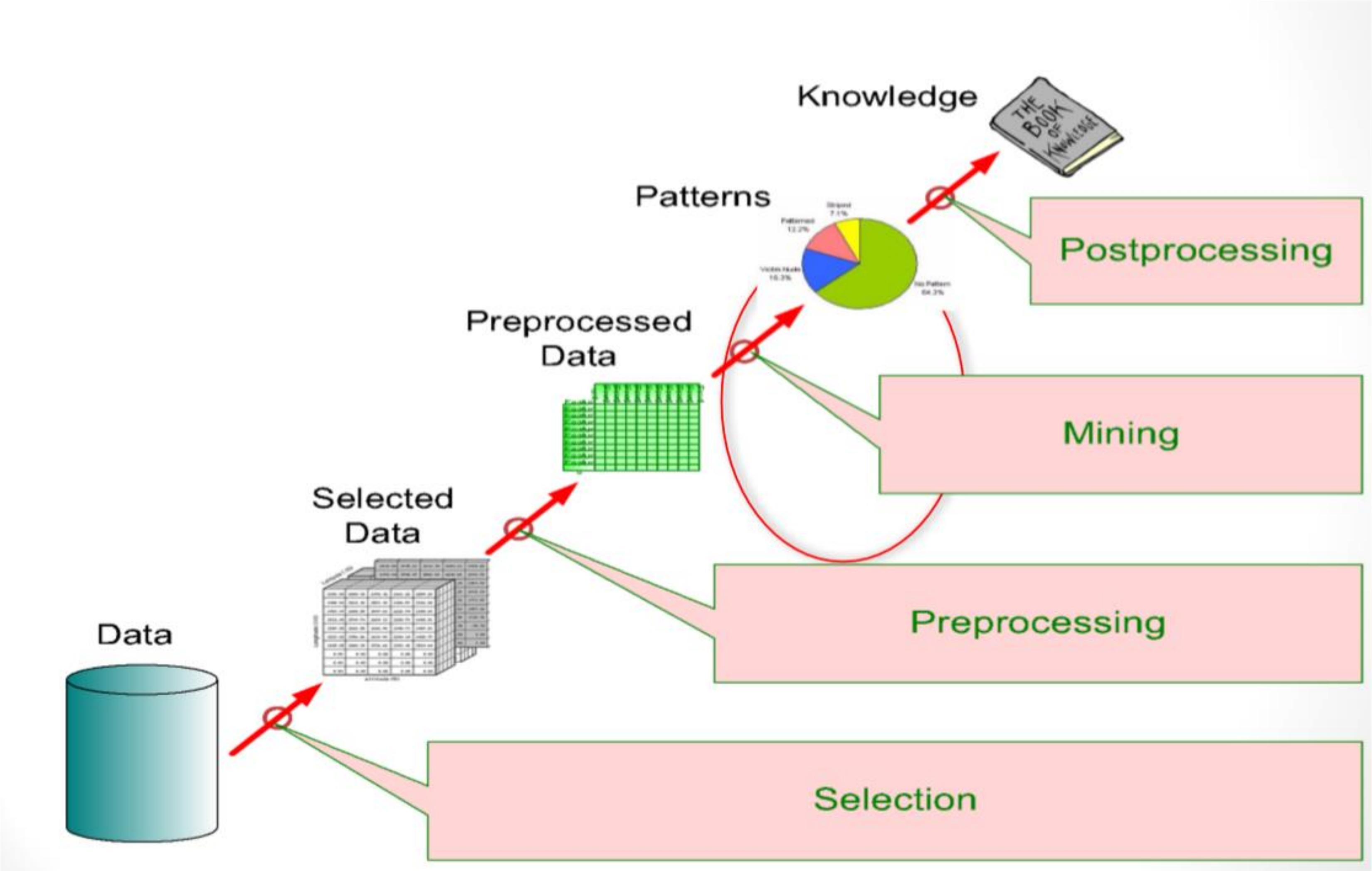
- From a model (theory)
  - father(a,b)
  - father(b,c)
  - ancestor(x,y) <- father(x,y)
  - ancestor(x,y) <- father(x,z), ancestor(z,y)
- to examples (theorems) by deduction :
  - ancestor(a,b), …, ancestor(a,c)

Model of ancestor

The KDD process

- The whole process of extraction useful knowledge from large databases is named knowledge discovery in databases(KDD)
- DM is a step of KDD

The KDD Process



Data Mining Tasks

- Predictive tasks: predicting the value of a particular attribute based on the values of the other attributes
  - Classification
  - Regression
- Example: predicting the value (true, false) of the attribute “reliable” based on the values of the other attributes describing a bank customer

Data Mining Tasks

- Descriptive tasks: inducing patterns that summarize the underlying relationships in data
    - Clustering: subdividing a set of objects into homogeneous subsets (clusters)
    - Association Analysis: inducing relationships among attributes
- Example: clustering newspaper articles (sport, politics, etc.)
- Example: market basket analysis which describes the behavior of the typical customer during shopping
- Buy Bread -> Buy Milk && diapers

Classes of Applications - Business

- ***User profiling***
  - Based on web-browsing behavior information (searches performed, pages accessed, goods bought, etc.) find clusters (groups) of “model” customers who share the same characteristics: interest, spending habits, etc. to suggest new products –for instance, suggest a new book based on the books recently purchased by similar users (recommendation systems)

Classes of Applications - Business

- ***Data mining in CRM***: rather than randomly contacting customers, DM allows to concentrate on customers that are predicted to have a high likelihood of responding to the offer (classification)
- ***Fraud Detection***: use historical data to build models for detecting and preventing fraudulent behaviors (e.g., tax evasion)
- ***Risk Analysis***: e.g., credit risk analysis (classification)

Classes of Applications - Social networks

- Every minute of every day, Facebook, Twitter, and other online communities generate enormous amounts of data like a customer likes Italian wines …
- An organization can exploit such data in planning future marketing initiatives. Social nets may function like a realtime CRM system, continually revealing new trends and opportunities.

Classes of Applications - Science

- **Genetic**s: find out the correlation between DNA structure and common diseases such as cancer
- **Astronomy**: JPL and the Palomar Observatory discovered 22 quasars with the help of data mining
- **Medicine**: find out patients at a high risk of breast cancer related to both genetics and life habits

Classes of Applications - Texts, images, spatial data

- Over 80% of human knowledge is represented in textual format
- Document classification and filtering –press review
- Sentiment Analysis or Opinion Mining -refers to the application of text analytics to identify and extract subjective information in texts
- Image clustering, classification, e.g., in radiology
- Spatial DM, e.g., public health services searching for explanations of disease clusters