Data Mining

(Know the past to predict the future!)

P. Rullo
rullo@mat.unical.it

a.a. 2013-2014

Teaching Material

- Main reference books
  - T.M. Mitchell, Machine Learning, 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 and the general-to-specific ordering
- Elements of probability theory
- Entropy
- Decision Trees
- Rule learning
- Naïve Bayes classifiers
- Ensemble methods
- Text classification: Maximum Entropy, OlexGA
- Association rules
- Clustering - Kmeans

Outline

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

Motivations for DM

- **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!
  - Automated data collection tools, mature database technology and internet lead to tremendous amounts of data stored in databases, data warehouses and other information repositories created for operational purposes (querying)
  - Different data types: tuples, texts, images, temporal, spatial, etc.

Motivations for DM

- Such data contains valuable knowledge which is hidden and implicit - so, it is not represented in the database schema. SQL is not suitable to extract it
- How can we do to extract such a knowledge?
- Knowledge Discovery techniques from large amounts of data are needed

What is DM?

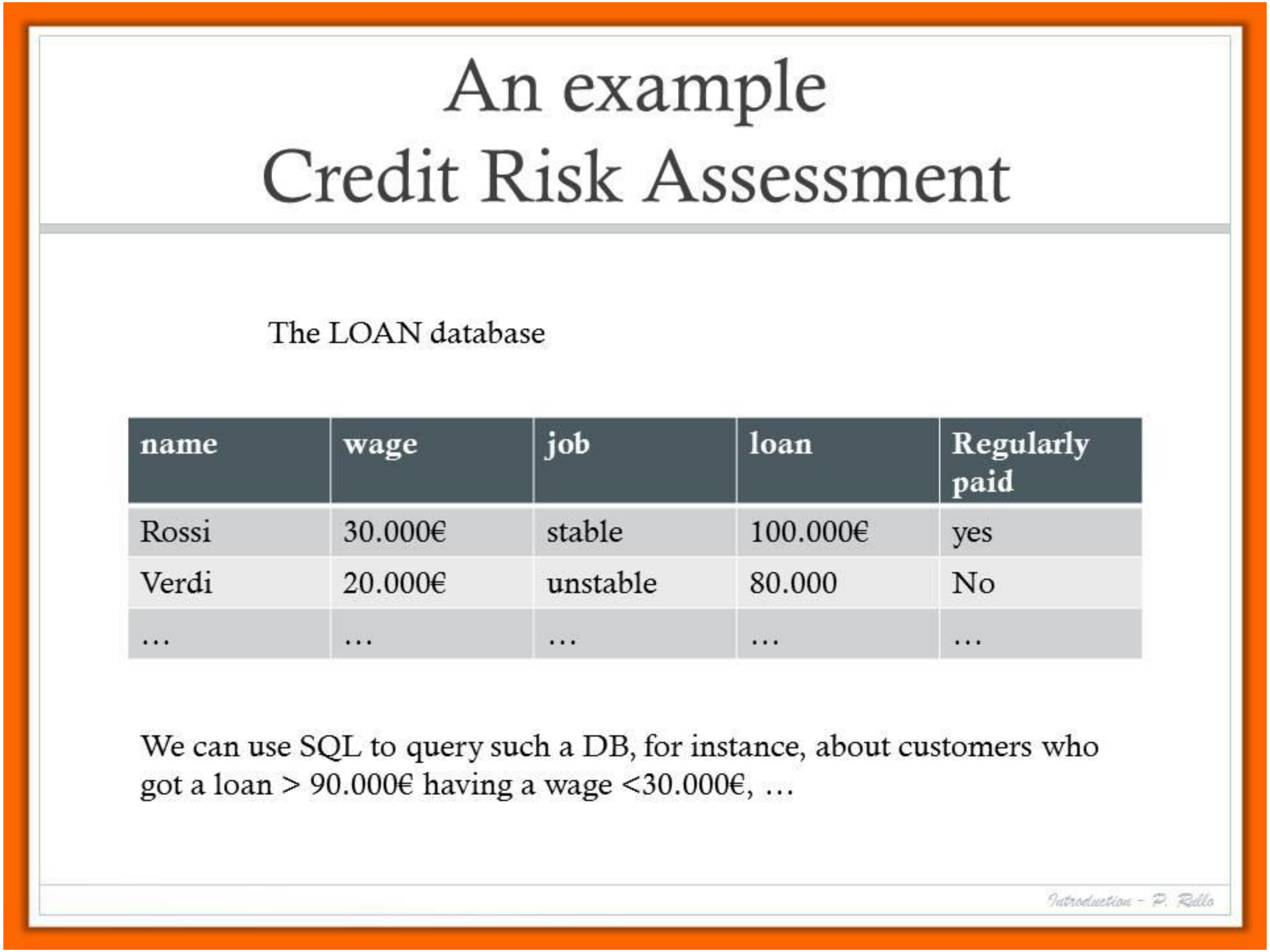
- A **miner** is a person who extracts coal, or other minerals from the earth
- Data miner is a person who extracts **knowledge** instead of minerals

What is DM?

- **Data mining (DM)**:
  - technology (algorithms, methods, tools) enabling the **extraction** of hidden and interesting knowledge (rules, regularities, patterns, constraints, …) from large amounts of data of different types and formats
- 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.,
  - Mr Rossi got a loan on 2002 of 100.000€
  - He regularly paid the mortgage in 10 years
  - Mr Rossi earns 30.000€/year, has a stable job, owns the apartment where he lives, is married, ….

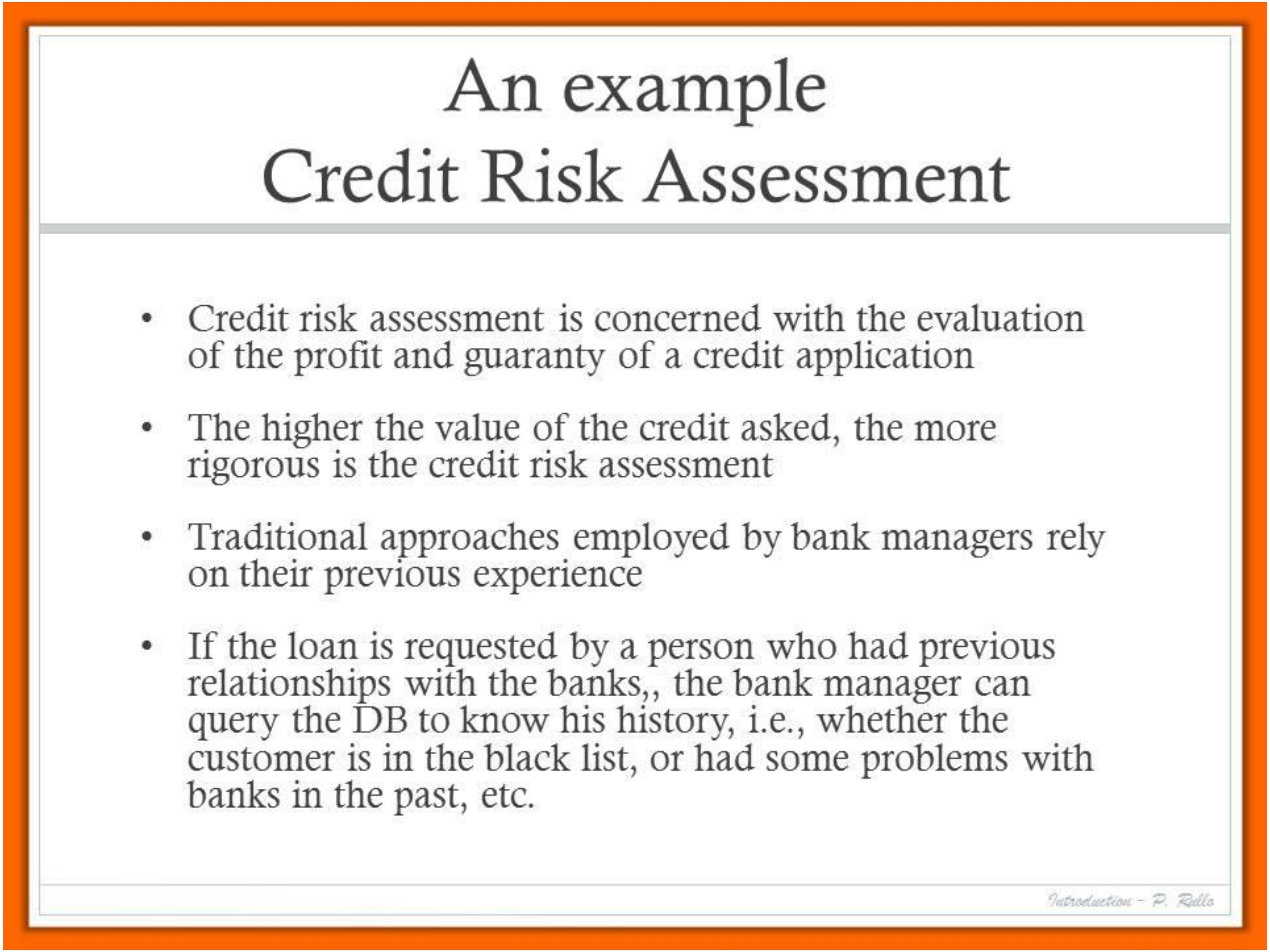


An example
Credit Risk Assessment

The LOAN database

name	wage	job	loan	Regularly paid
Rossi	30.000€	stable	100.000€	yes
Verdi	20.000€	unstable	80.000	No
…	…	…	…	…

We can use SQL to query such a DB, for instance, about customers who got a loan > 90.000€ having a wage <30.000€, …



An example
Credit Risk Assessment

- Credit risk assessment is concerned with the evaluation of the profit and guaranty of a credit application
- The higher the value of the credit asked, the more rigorous is the credit risk assessment
- Traditional approaches employed by bank managers rely on their previous experience
- If the loan is requested by a person who had previous relationships with the banks,, the bank manager can query the DB to know his history, i.e., whether the customer is in the black list, or had some problems with banks in the past, etc.

An example
Credit Risk Assessment

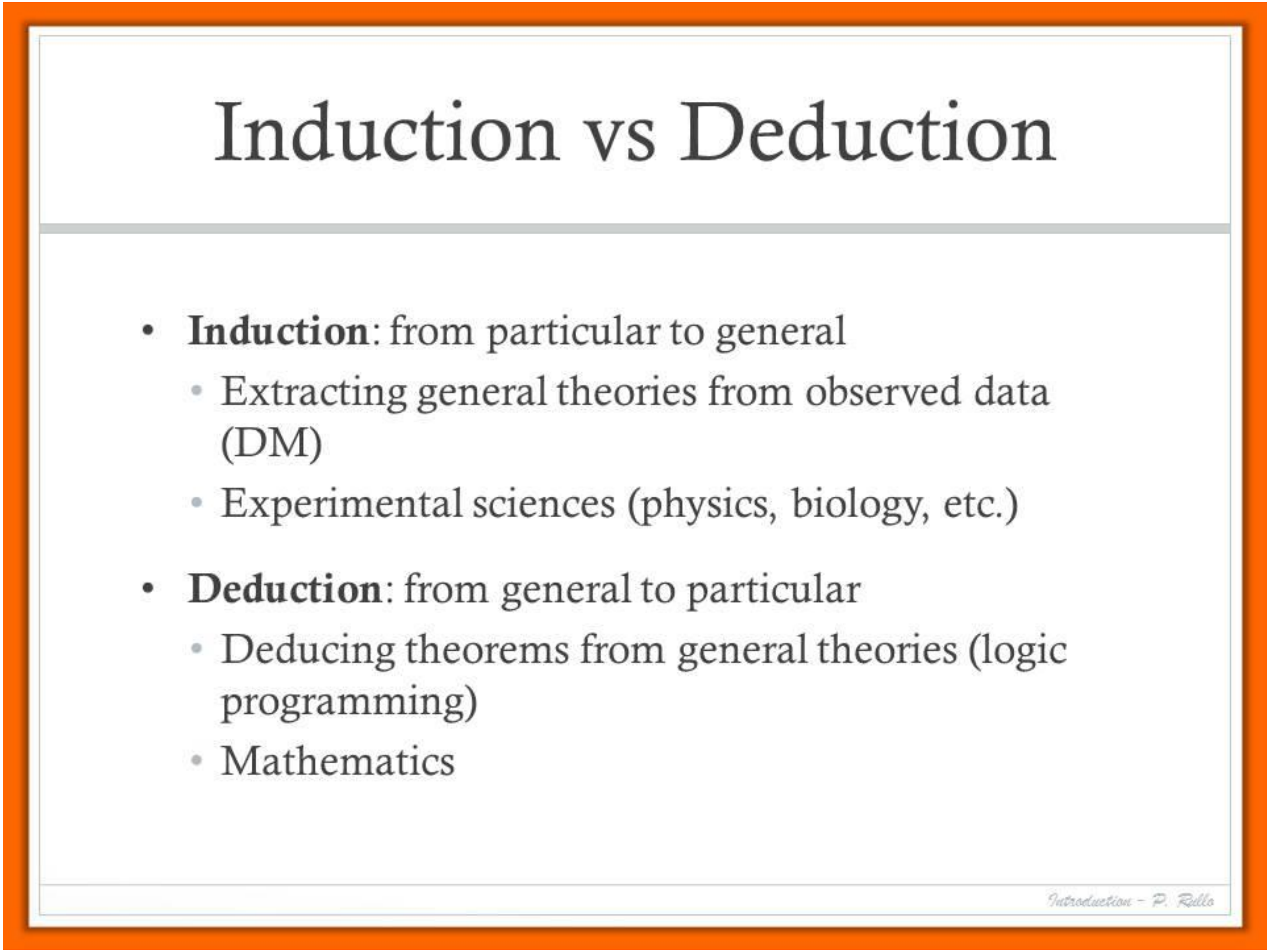
- But, what happens with new customers?
- What we actually need is a MODEL for concept "reliable customers", i.e., a description of the properties a customer should hold in order to be considered reliable
- What are the properties a customer should exhibit in order to be classified as "reliable"?

An example
Credit Risk Assessment

- DM comes to help:
  - a learning algorithm is trained over the LOAN database (training data) to learn the concept of "reliable customer"
  - e.g., a reliable customer is one who earns more than 30.000€/year, has a stable job, is not married ….
  - the learned concept (or model) is then used when e new customer asks for a loan to decide whether or not to grant it
- Predictive task: *Know the past to predict the future*!

DM and Induction

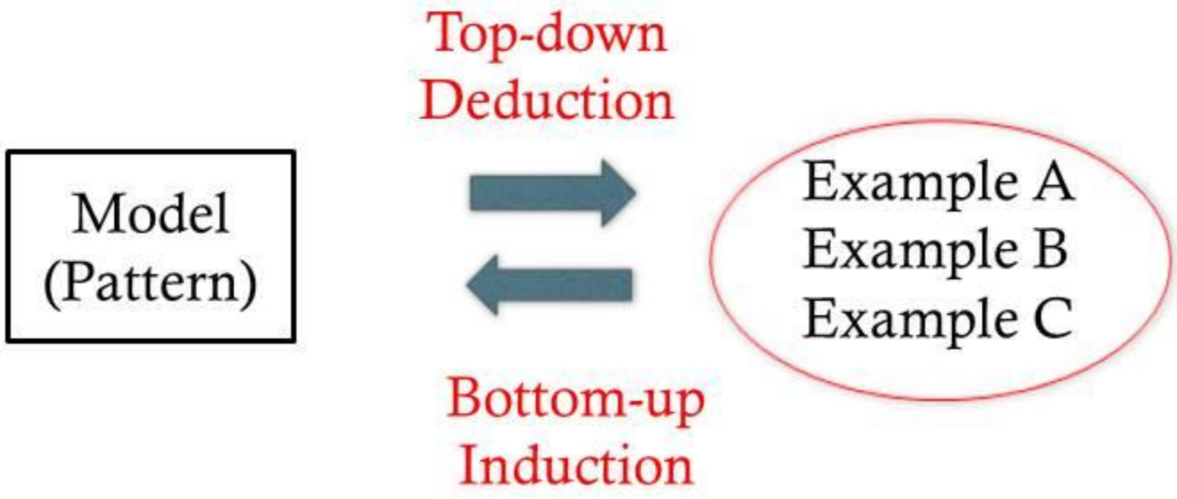
- DM is an inductive task, as it extracts general theories from observed data (empirical observations)
- Question: What is the model of the "reliable bank customer"?
- DM techniques allow to induce such a model by learning from a number of examples stored in the credit database (training examples)
- Purely inductive learning methods formulate general hypotheses by finding empirical regularities over the training examples.' (Tom M. Mitchell,1997,p334)



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



Top-down
Deduction

Model
(Pattern)

Example A
Example B
Example C

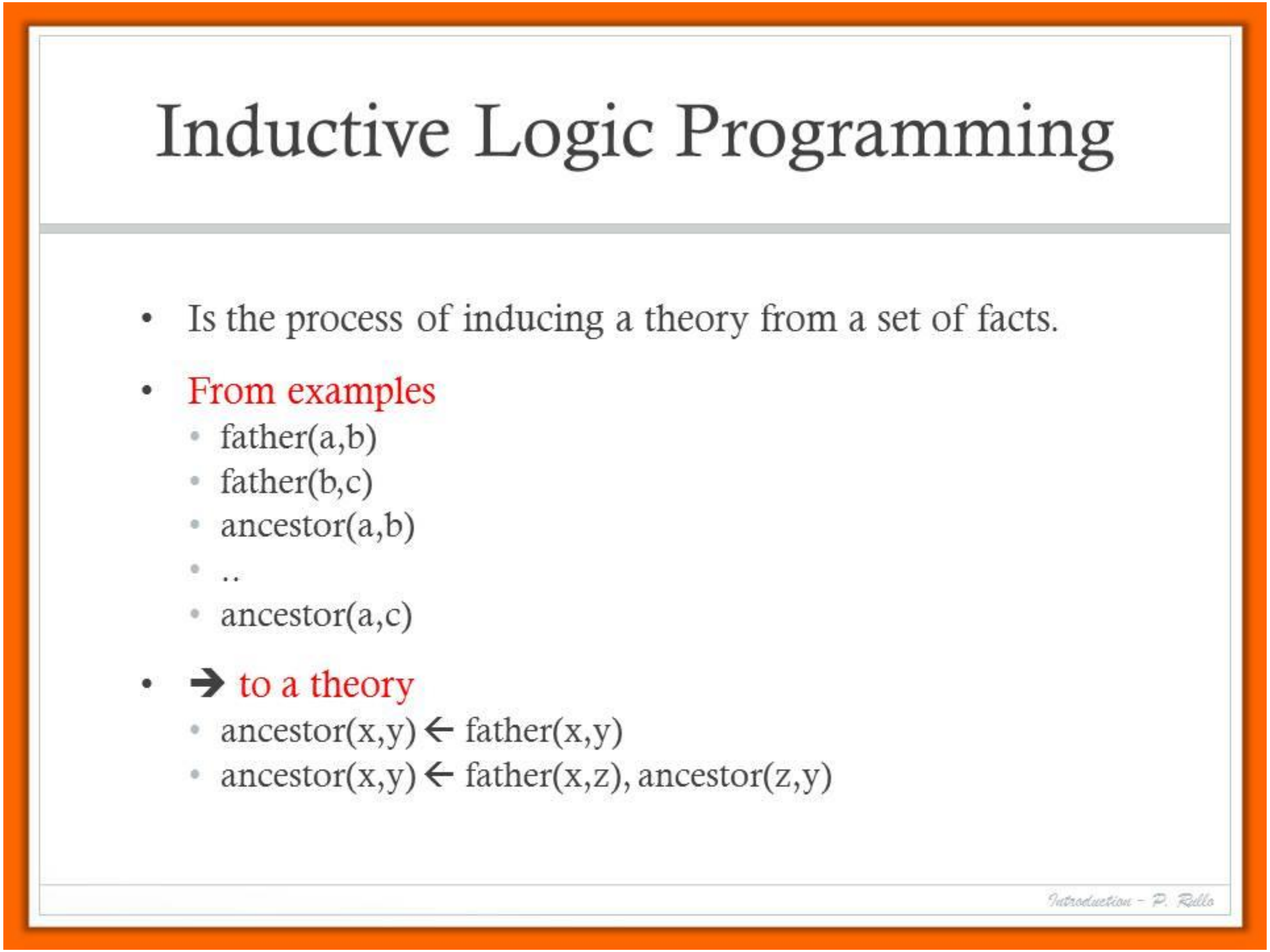
Bottom-up
Induction

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)

Model of ancestor

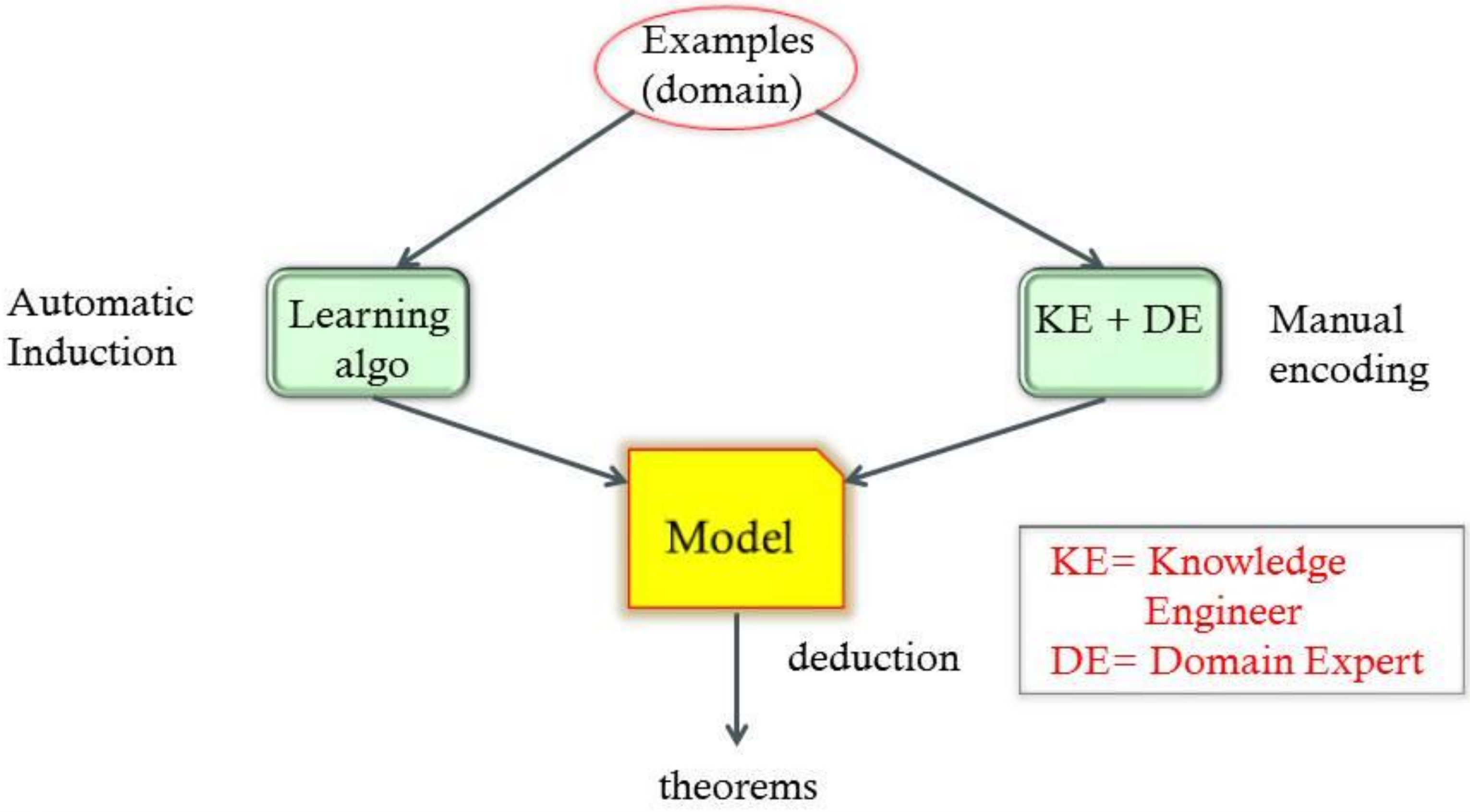
- to examples (theorems) by deduction :
  - ancestor(a,b), …, ancestor(a,c)



Inductive Logic Programming

- Is the process of inducing a theory from a set of facts.
- From examples
  - father(a,b)
  - father(b,c)
  - ancestor(a,b)
  - ..
  - ancestor(a,c)
- → to a theory
  - ancestor(x,y) ← father(x,y)
  - ancestor(x,y) ← father(x,z), ancestor(z,y)

Induction vs Deduction

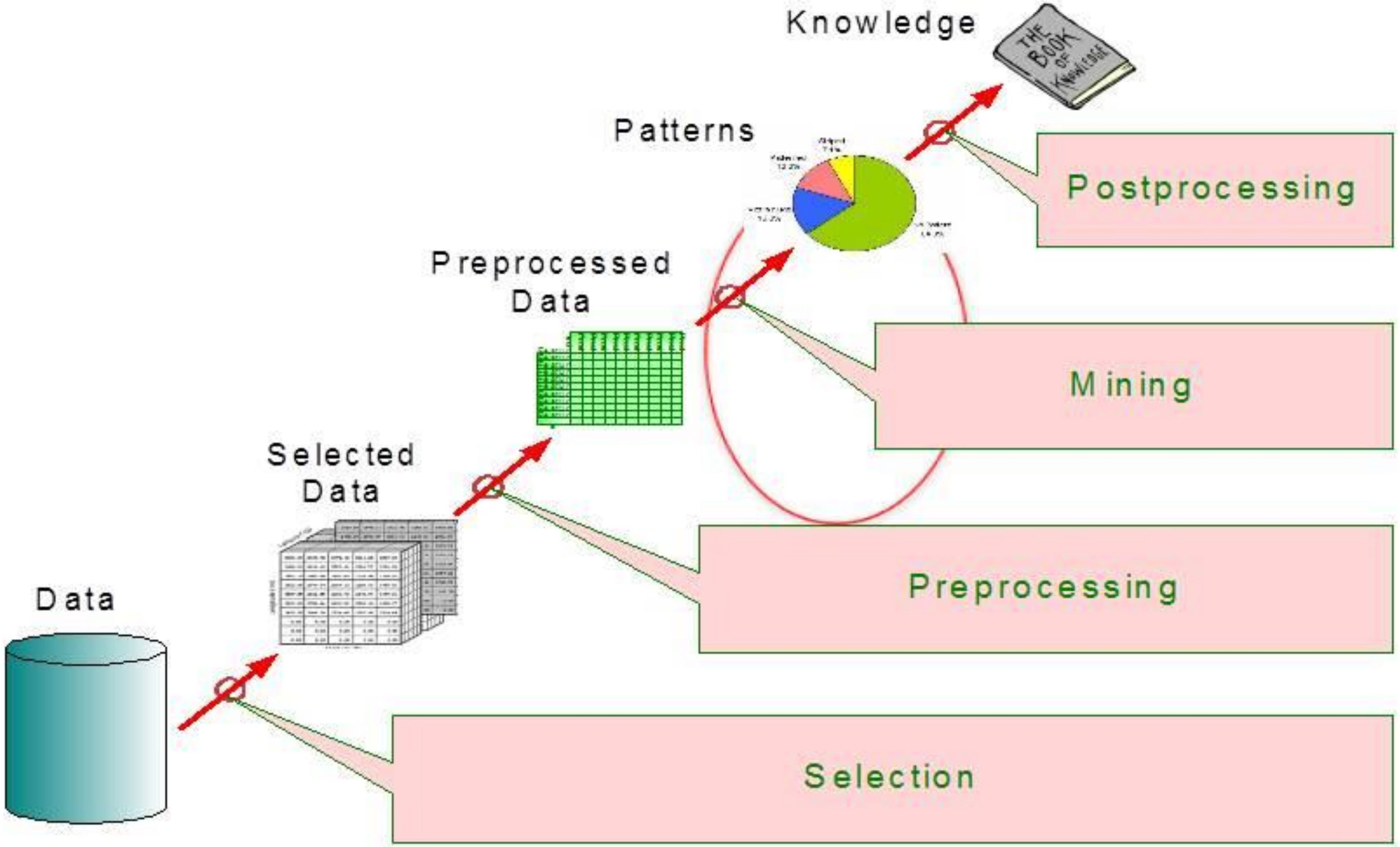


KE= Knowledge Engineer
DE= Domain Expert

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

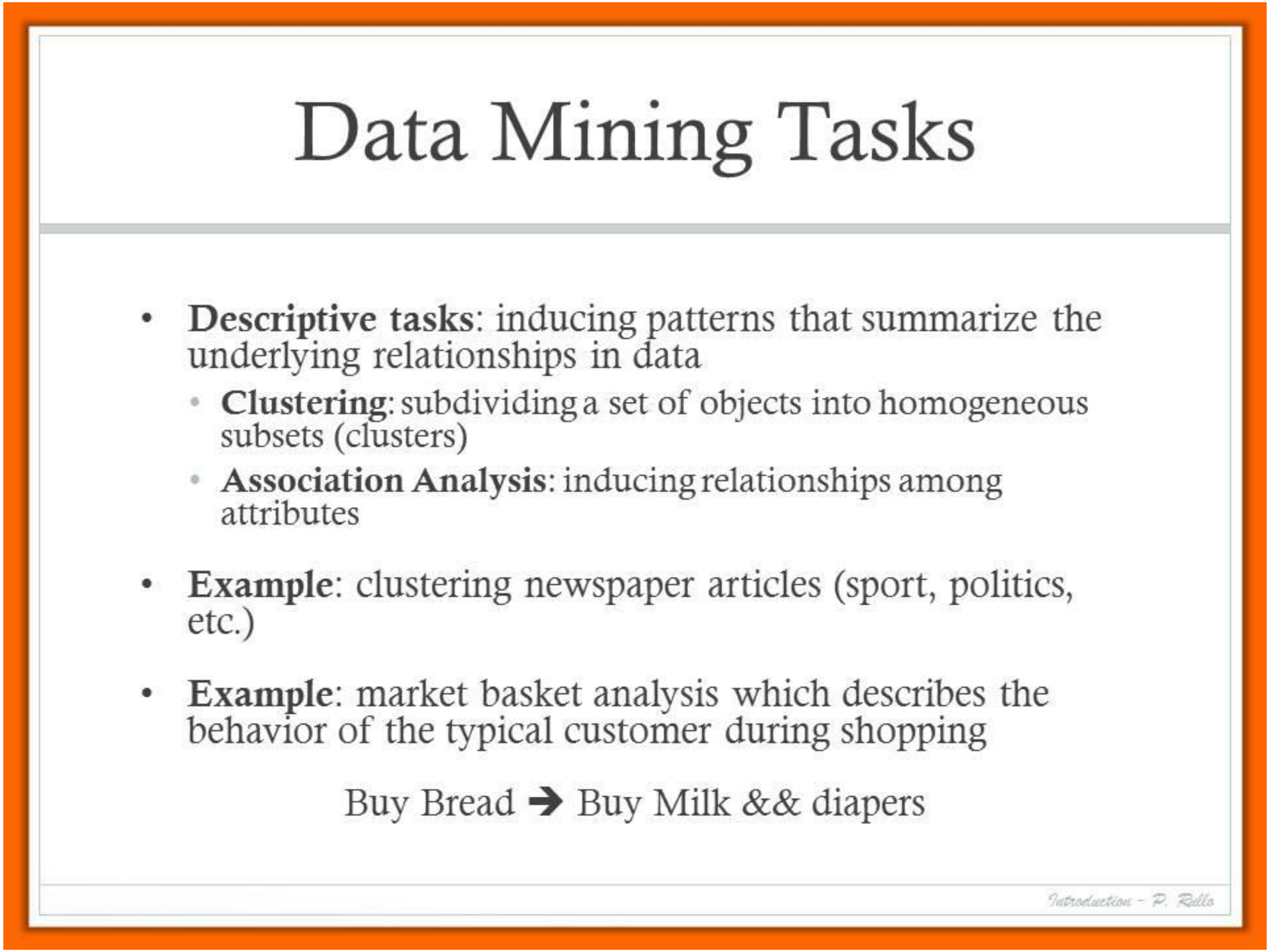


Pre-processing

- The data present in a database must be adequately prepared before data mining techniques can be applied to it. The main steps employed for data preparation are:
  - Preprocessing of the data to the format specified by the algorithms to be used
  - Reduction of the number of samples/instances
  - Reduction of the number of features/attributes
  - Features construction, which is the combination of one or more attributes in order to transform irrelevant attributes to more significant attributes; and
  - Noise elimination and treatment of missing values.

Data Mining Tasks

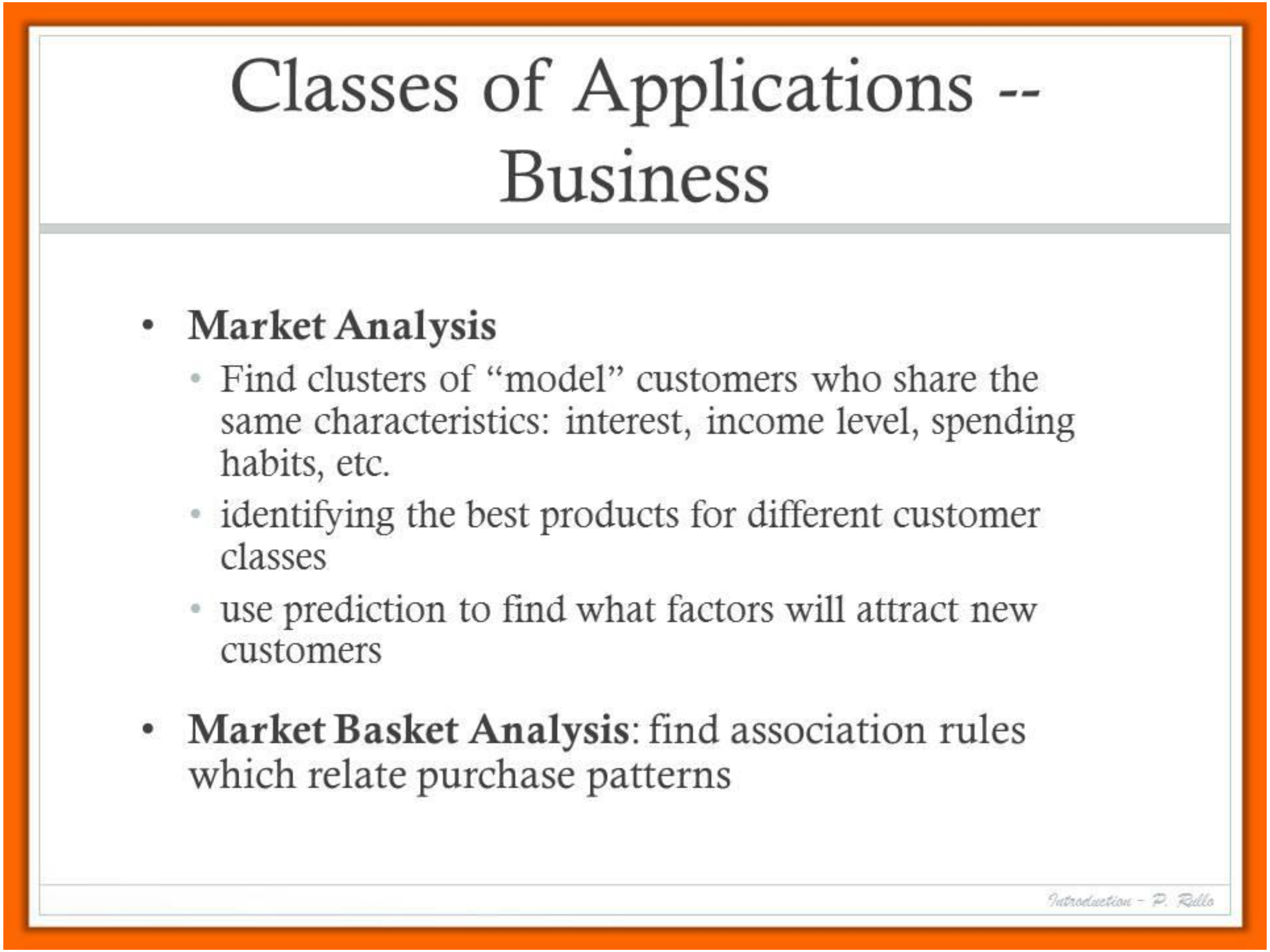
- **Predictive tasks**: predicting the value of a particular attribute based on the values of the other attributes
  - Classification
  - Regression
- **Example:** predicting the value (true, false) of the attribute "reliable" based on the values of the other attributes describing a bank customer



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

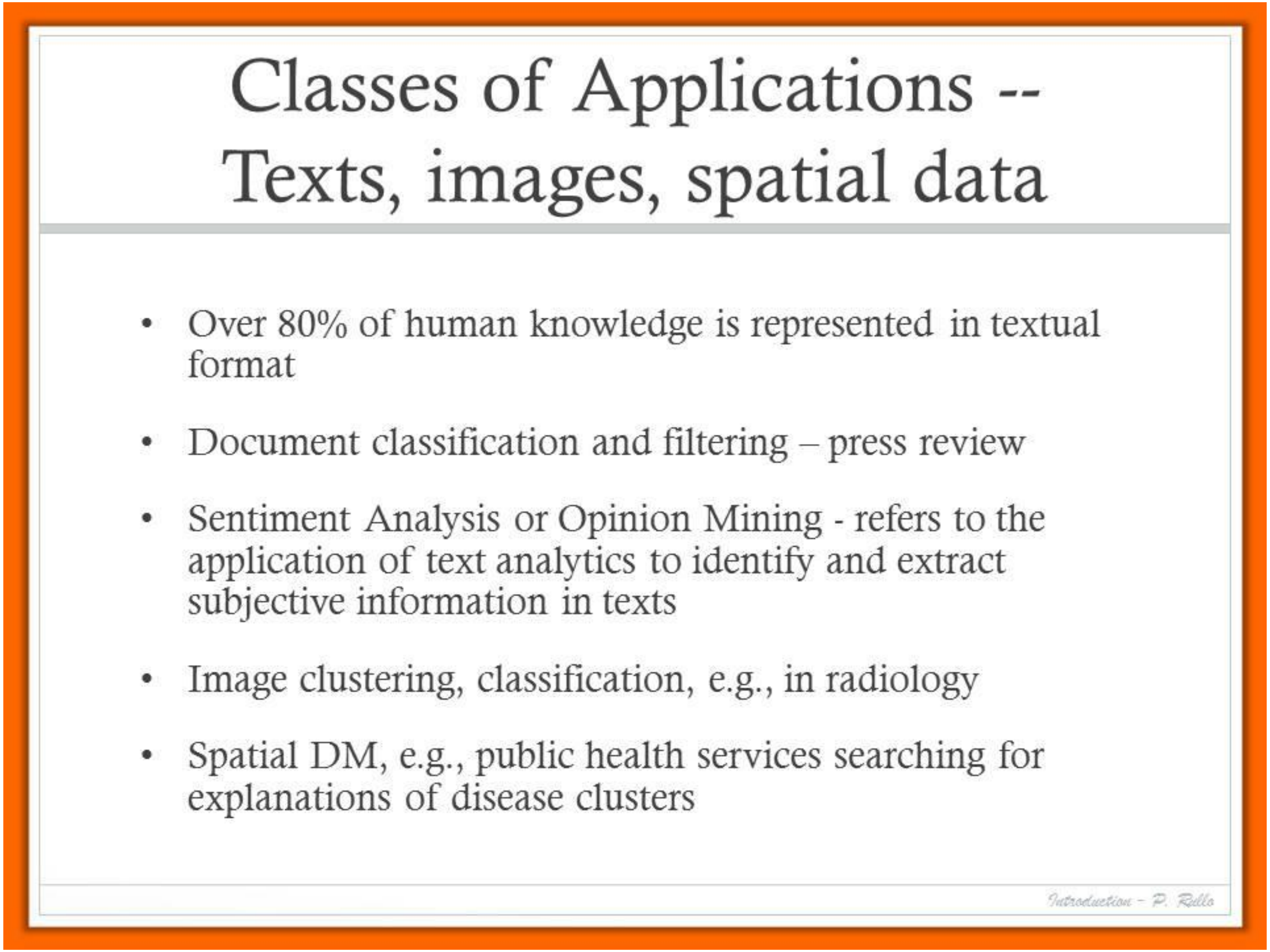
- **Market Analysis**
  - Find clusters of “model” customers who share the same characteristics: interest, income level, spending habits, etc.
  - identifying the best products for different customer classes
  - use prediction to find what factors will attract new customers
- **Market Basket Analysis**: find association rules which relate purchase patterns

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 -- Science

- **Genetics**: find out how the changes in an individual's DNA sequence affect the risk of developing 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

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 real-time CRM system, continually revealing new trends and opportunities.