Department of Mathematics
University of Calabria

Data Warehouse and Data Mining

Module II – Data Mining

Introduction to Data Mining

Ph.D. Ettore Ritacco

About the Course

- Instructor:

  Prof. Pasquale Rullo, rullo@mat.unical.it
  Department of Mathematics
  of the
  University of Calabria - UNICAL
  Office hours: by appointment

- Assistant:

  Ph.D. Ettore Ritacco, ritacco@icar.cnr.it
  High Performance Computing and Networking Institute - ICAR
  of the
  Italian National Research Council - CNR
  Office hours: by appointment (@ Cube 41C, 1st floor)

About the final exam

- It consists of two parts:
  - **Project**
    - It should be realized in group of 3-4 students.
    - The theme will be assigned during the course.
    - The results will be presented during a poster session.
  - **Oral Proof**

Teaching Material

- Main Text
  - Data mining: Concepts and Techniques, by J. Han and M. Kamber, Morgan Kaufmann Publishers.
- References
  - Machine Learning, by Tom M. Mitchell, McGraw-Hill.
  - Data Mining: Practical Machine Learning Tools and Techniques, by I. Witten, E. Frank and M. Hal, Morgan Kaufmann Publishers.
- Any useful material and information will be published on the course webpage :

  https://www.mat.unical.it/informatica/DataWarehouseEDataMiningModulo2

Technology

- Basically we're going to use Rialto and Weka.
  - Rialto: commercial product
    - http://www.exeura.eu/products/rialto (coming soon)
  - Weka: open source product (GNU General Public License - http://www.gnu.org/licenses/gpl.html)
    - http://www.cs.waikato.ac.nz/ml/weka/ (downloadable content)

- You can use any other DM software (even combinations):
  - Rapid Miner
  - Knime
  - R
  - Excel
  - ….

Outline

- The **WWWWH** questions:
  - **W**hat is data mining?
  - **W**hy mining data?
  - **W**hen and **W**here data mining?
  - **H**ow to do data mining?

Department of Mathematics
University of Calabria

Outline

- What

Department of Mathematics
University of Calabria

What is data mining?
(for dummies)

- The Duck Test

Department of Mathematics
University of Calabria

What is data mining?
(for dummies)

- The Duck Test

- If it
  - looks like a duck
  - swims like a duck
  - and quacks like a duck

What is data mining?
(for dummies)

- The Duck Test
- If it
  - looks like a duck
  - swims like a duck
  - and quacks like a duck
- Then?

Department of Mathematics
University of Calabria

What is data mining? (for dummies)

- It's a Duck! (probably...)

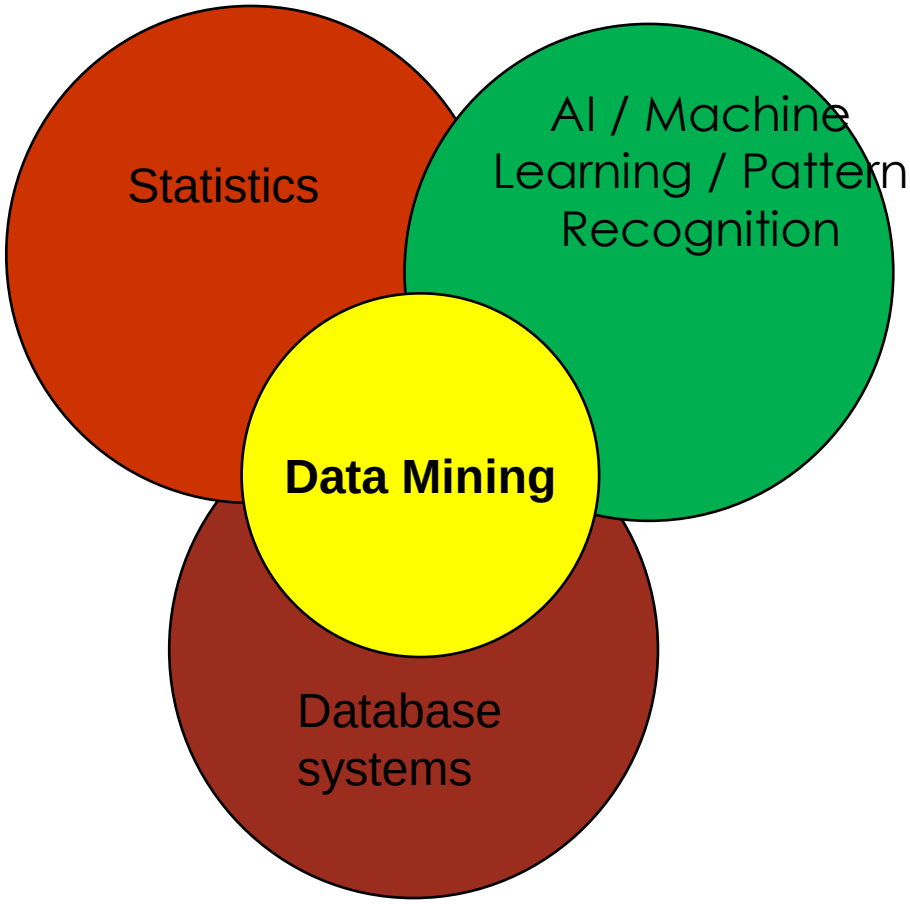
What is data mining?

- The process of **automatically** discover knowledge, models and patterns in **large** data repositories
  - Knowledge is the understanding of the phenomena that produce the observed data
  - Models are mathematical and logic sets of functions.
  - Patterns are discernible regularities within the data, whose elements and/or features repeat in a predictable manner

- Wish list:
  - Novelty
  - Generality
  - Utility
  - Understandability

What is data mining?

- Draws ideas from a lot of science fields…



What is data mining?

- … for instance:
  - Knowledge Discovery
  - Pattern Recognition
  - Artificial Intelligence
  - Machine Learning
  - Statistics
  - Graph Theory
  - Business Process Management
  - Data Management
  - Information Theory
  - …

What is data mining?

- Major challenges:
  - Scalability
  - Dimensionality
  - Complex and Heterogeneous Data
  - Data Quality
  - Data Ownership and Distribution
  - Privacy Preservation
  - Streaming Data

What is data mining?
(Student's perspective)

What is data mining?
(Student's perspective)

- Easy task:
  - Given the input x, some parameters θ and a function *f*, find:
    - $y = f(x, \theta)$

What is data mining?
(Student's perspective)

- Medium task:
  - Given the input x, the output *y* and a function *f*, find:
    - *θ such that y = f(x, θ)*

Department of Mathematics
University of Calabria

What is data mining?
(Student's perspective)

- (Quite) Medium task:
  - Given the output *y*, some parameters *θ* and a (invertible) function *f*, find:
    - $x = f^{-1}(y, \theta)$

What is data mining?
(Student's perspective)

- Hard task (**data mining – prediction**):
  - Given the output *y* and the input x, find:
    - *f, θ such that y = f(x, θ)*

What is data mining?
(Student's perspective)

- (Very) Hard task (**data mining – description**):
  - Given the input x, find:
    - *f, θ such that f(x, θ) governs x (f(x, θ) can explain and generate x)*

What is data mining?
(And what's not)

Yes	No
Find names that are more prevalent in certain locations	SQL query
Group together similar documents returned by search engine according to their context	Search documents through a search engine
Search for the most important numbers within a phone directory	Look up phone number in phone directory

Department of Mathematics
University of Calabria

Outline

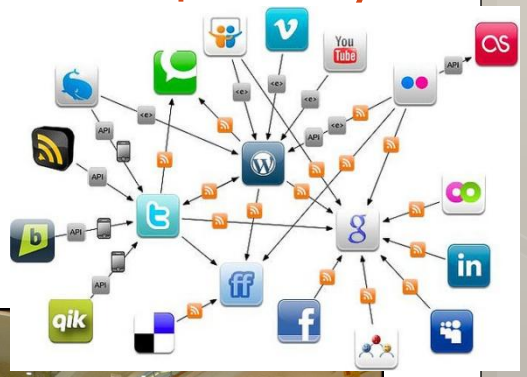
- Why

Why Mining Data?
(Scientific Viewpoint)

- Data collected and stored at enormous speeds (GB-TB-PB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in hypothesis definition

Why Mining Data?
(Commercial Viewpoint)

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - Purchases at department/grocery stores
  - Bank/credit card transactions
- Computers have become cheaper and more powerful
- Competitive pressure is strong
  - Providing better and customized services



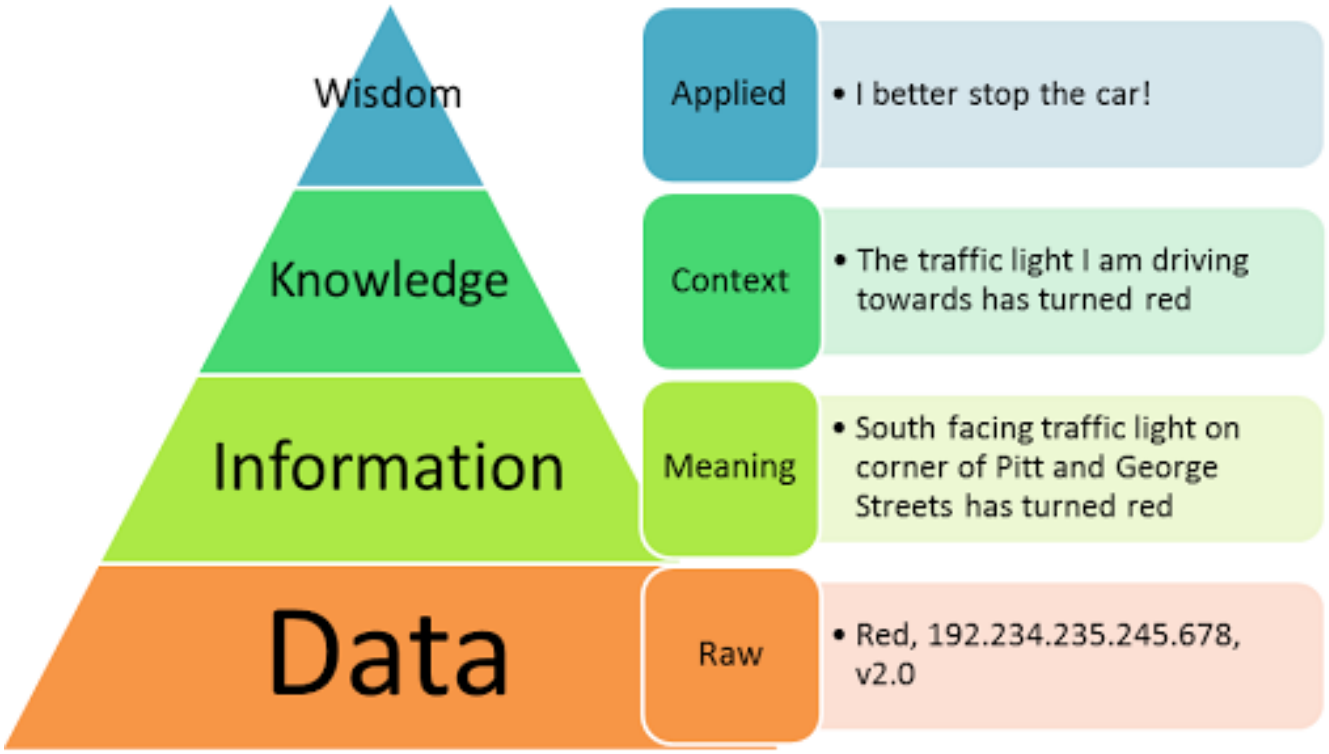
Department of Mathematics
University of Calabria

Outline

- When & Where

When and Where data mining?

- DIKW hierarchy



When and Where data mining?

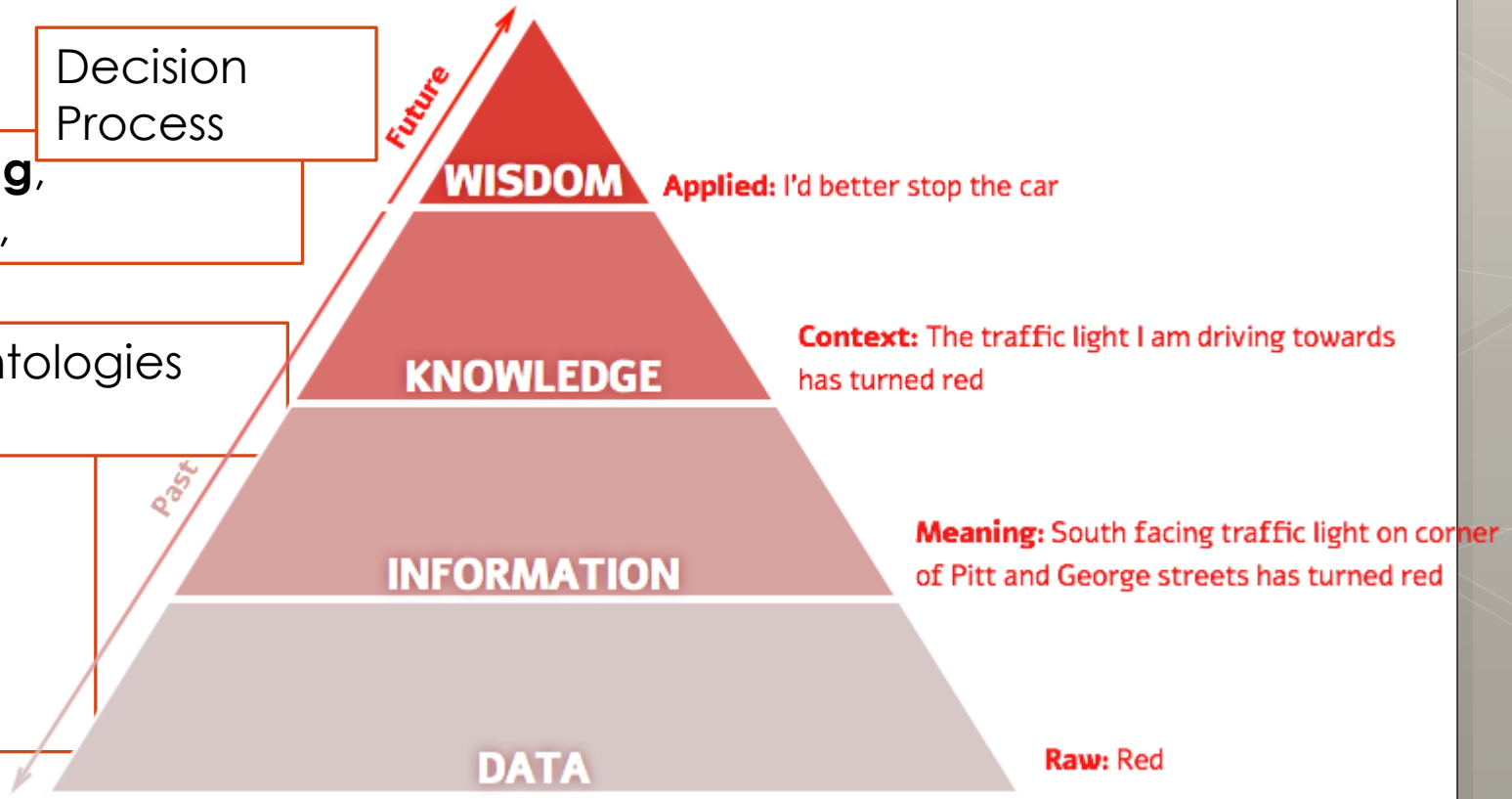
- DIKW hierarchy

Decision Process

Data Mining, Reasoning,

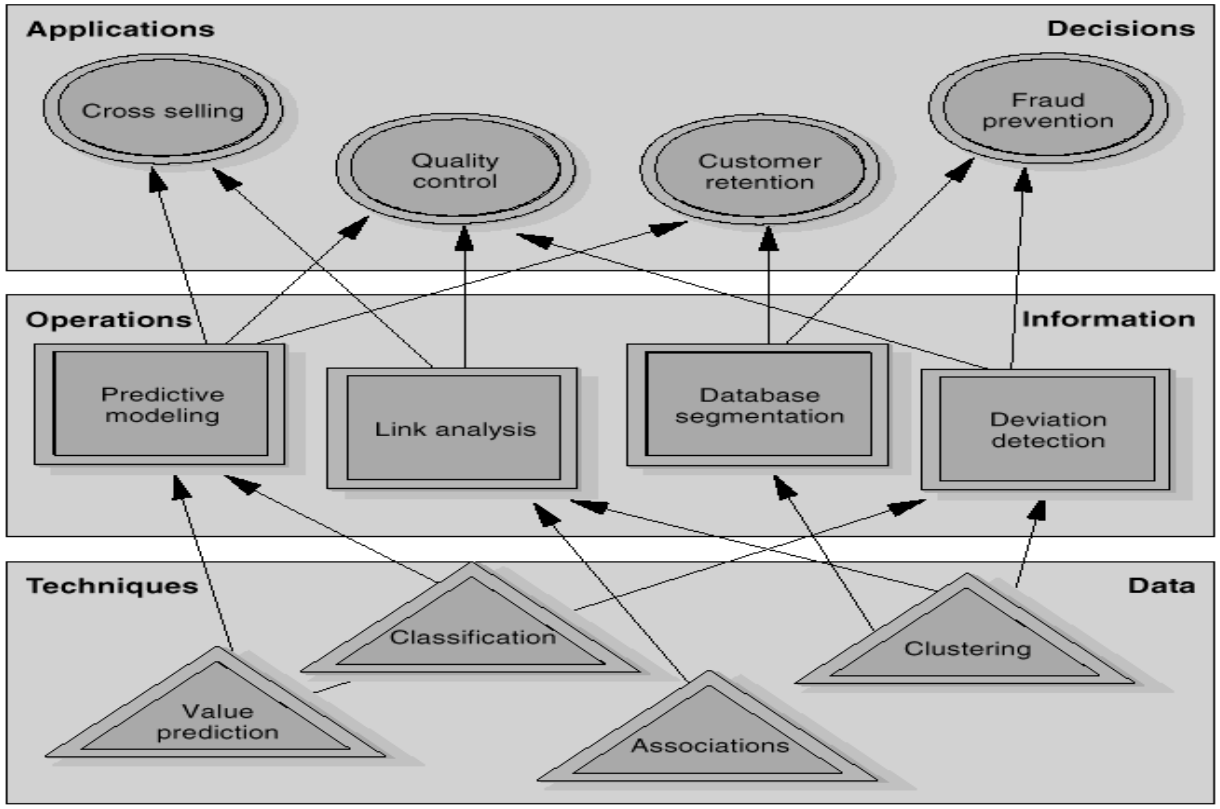
Ontologies

DBMS, Information Retrieval, Data Wharehouse



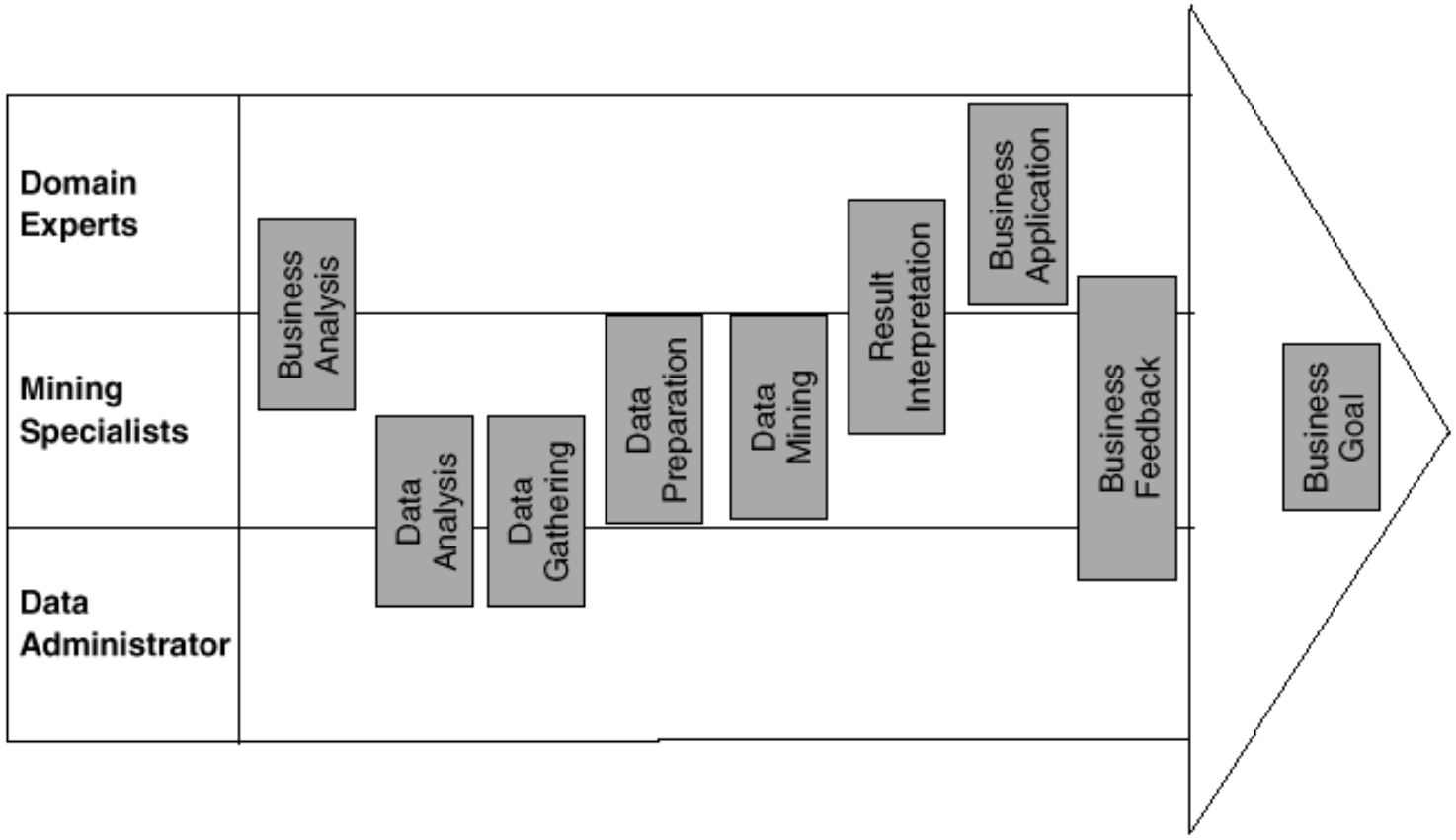
When and Where data mining?

- Applications, operations, techniques



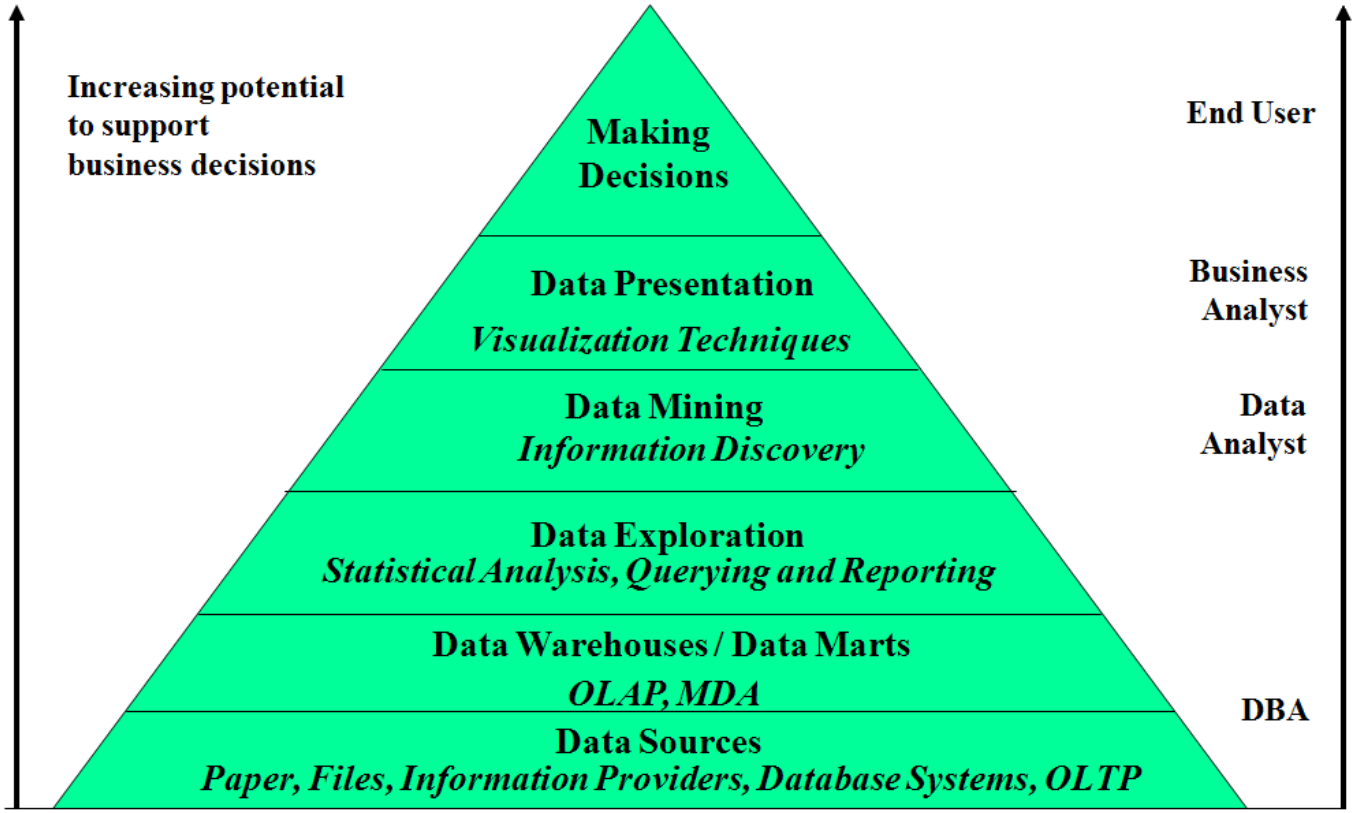
When and Where data mining?

- Roles in the KDD process



When and Where data mining?

- Data mining and business intelligence



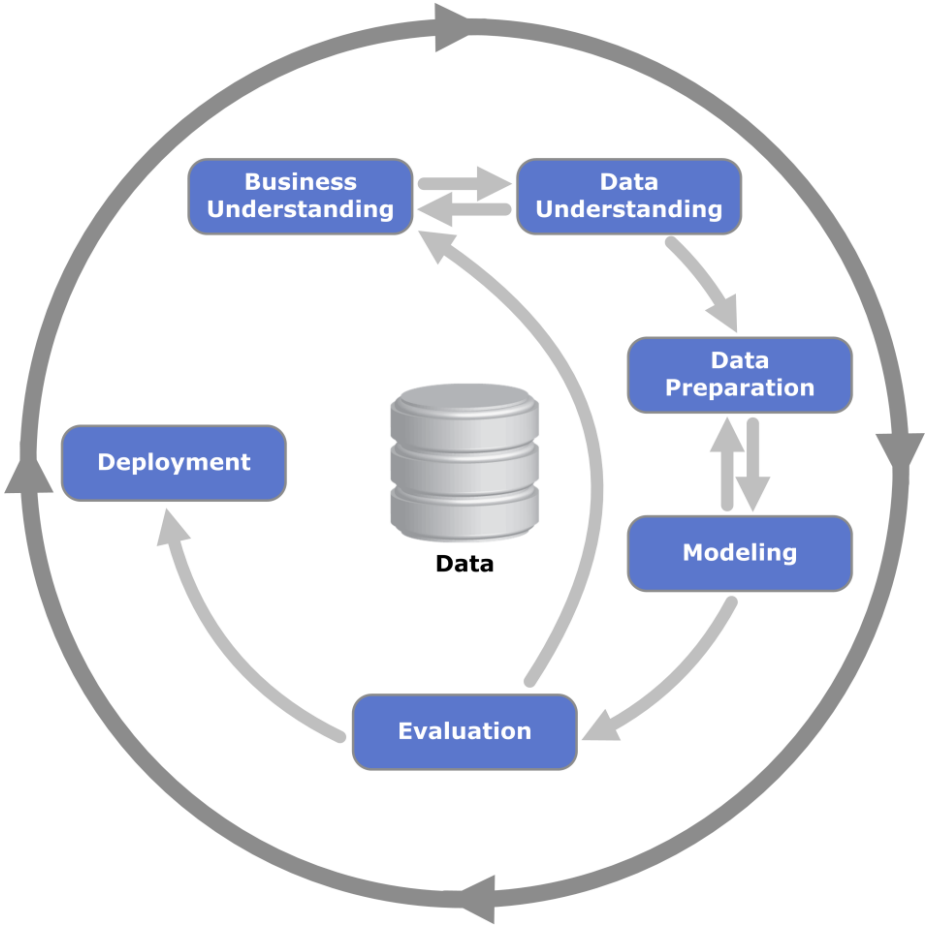
Department of Mathematics
University of Calabria

Outline

- How

How to do data mining?

- CRISP-DM Methodology
- (CRoss Industry Standard Process for Data Mining)



How to do data mining?

- CRISP-DM Phases (1 – 3):
  - Business Understanding
    - Understanding project objectives and requirements
    - Data mining problem definition
  - Data Understanding
    - Initial data collection and familiarization
    - Identify data quality issues
    - Initial, obvious results
  - Data Preparation
    - Record and attribute selection
    - Data cleansing

How to do data mining?

- CRISP-DM Phases (4 – 6):
  - Modeling
    - Run the data mining techniques
  - Evaluation
    - Determine if results meet business objectives
    - Identify business issues that should have been addressed earlier
  - Deployment
    - Put the resulting models into practice
    - Set up for repeated/continuous mining of the data

How to do data mining?

Business Understanding

Determine Business Objectives
Background
Business Objectives
Business Success Criteria

Situation Assessment
Inventory of Resources
Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits

Determine Data Mining Goal
Data Mining Goals
Data Mining Success Criteria

Produce Project Plan
Project Plan
Initial Asessment of Tools and Techniques

Data Understanding

Collect Initial Data
Initial Data Collection Report

Describe Data
Data Description Report

Explore Data
Data Exploration Report

Verify Data Quality
Data Quality Report

Data Preparation

Data Set
Data Set Description

Select Data
Rationale for Inclusion / Exclusion

Clean Data
Data Cleaning Report

Construct Data
Derived Attributes
Generated Records

Integrate Data
Merged Data

Format Data
Reformatted Data

Modeling

Select Modeling Technique
Modeling Technique
Modeling Assumptions

Generate Test Design
Test Design

Build Model
Parameter Settings
Models
Model Description

Assess Model
Model Assessment
Revised Parameter Settings

Evaluation

Evaluate Results
Assessment of Data Mining Results w.r.t. Business Success Criteria
Approved Models

Review Process
Review of Process

Determine Next Steps
List of Possible Actions
Decision

Deployment

Plan Deployment
Deployment Plan

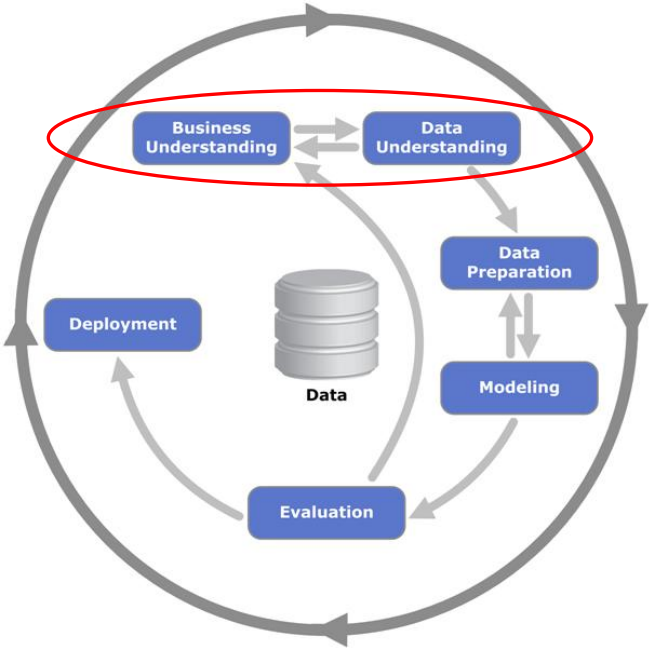
Plan Monitoring and Maintenance
Monitoring and Maintenance Plan

Produce Final Report
Final Report
Final Presentation

Review Project
Experience Documentation

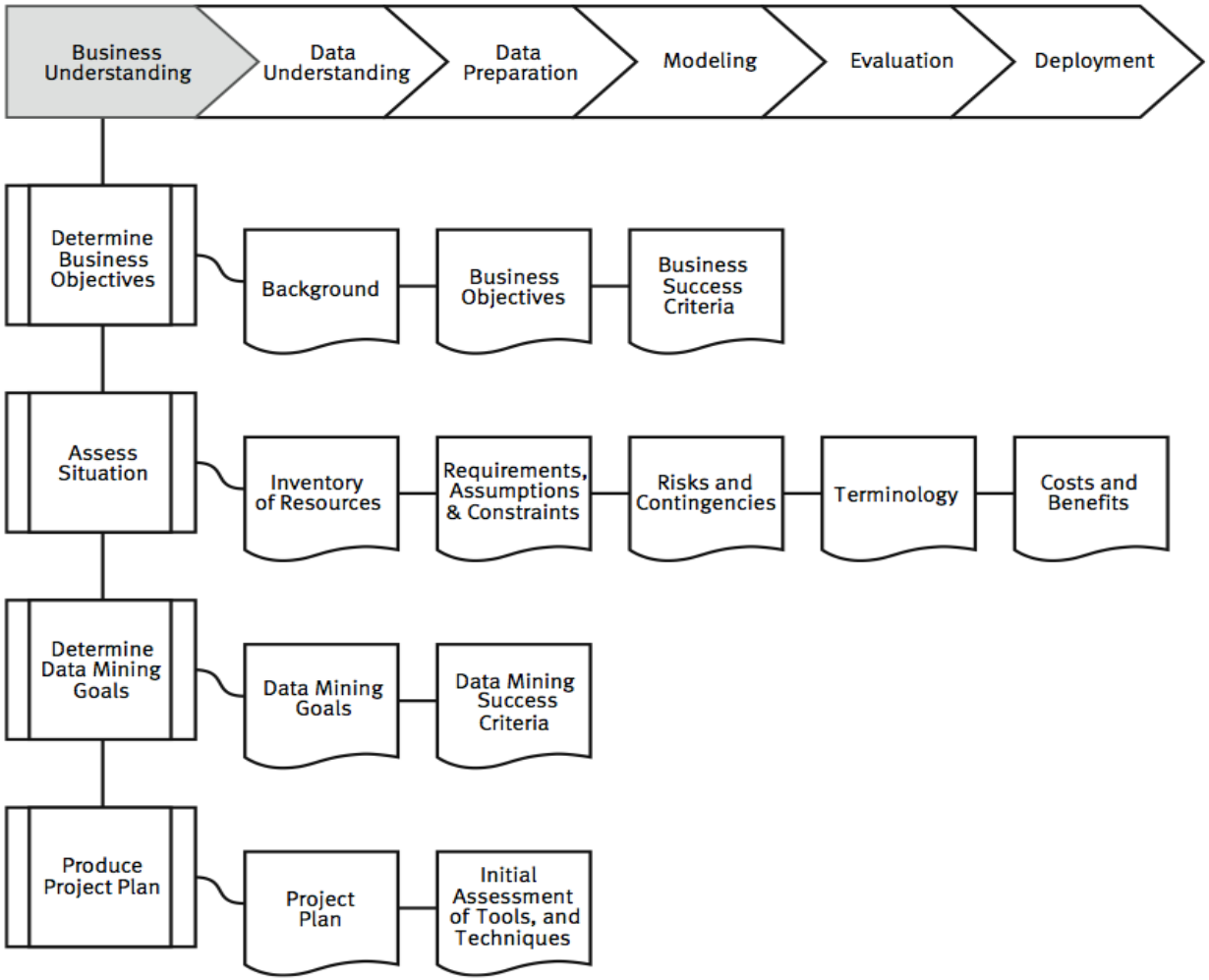
How to do data mining?

- Phases in the DM Process (1 & 2)
- Business Understanding:
  - Statement of Business Objective
  - Statement of Data Mining objective
  - Statement of Success Criteria
  - Project plan

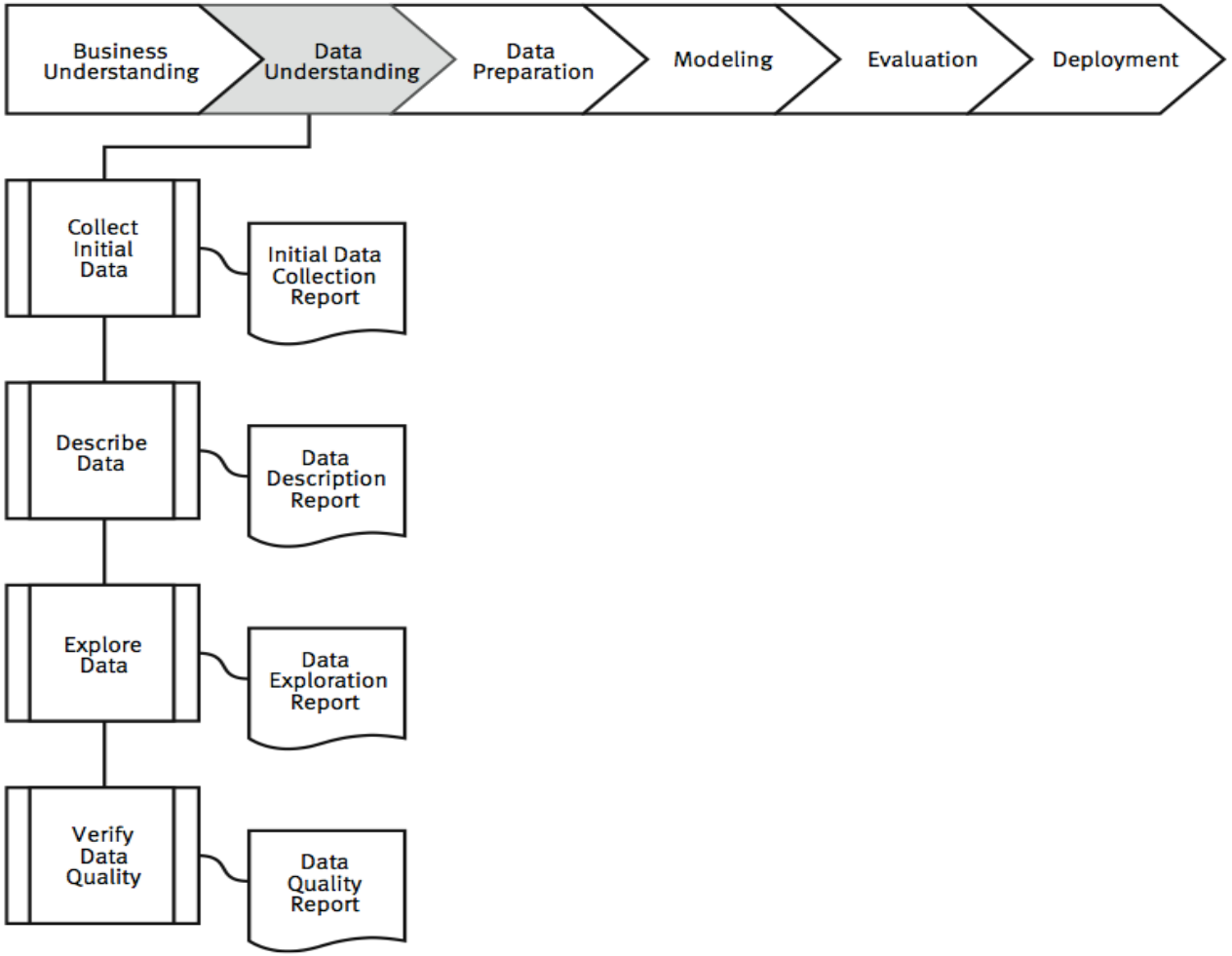


- Data Understanding
  - Explore the data and verify the quality
  - Find outliers

Business Understanding



Data Understanding

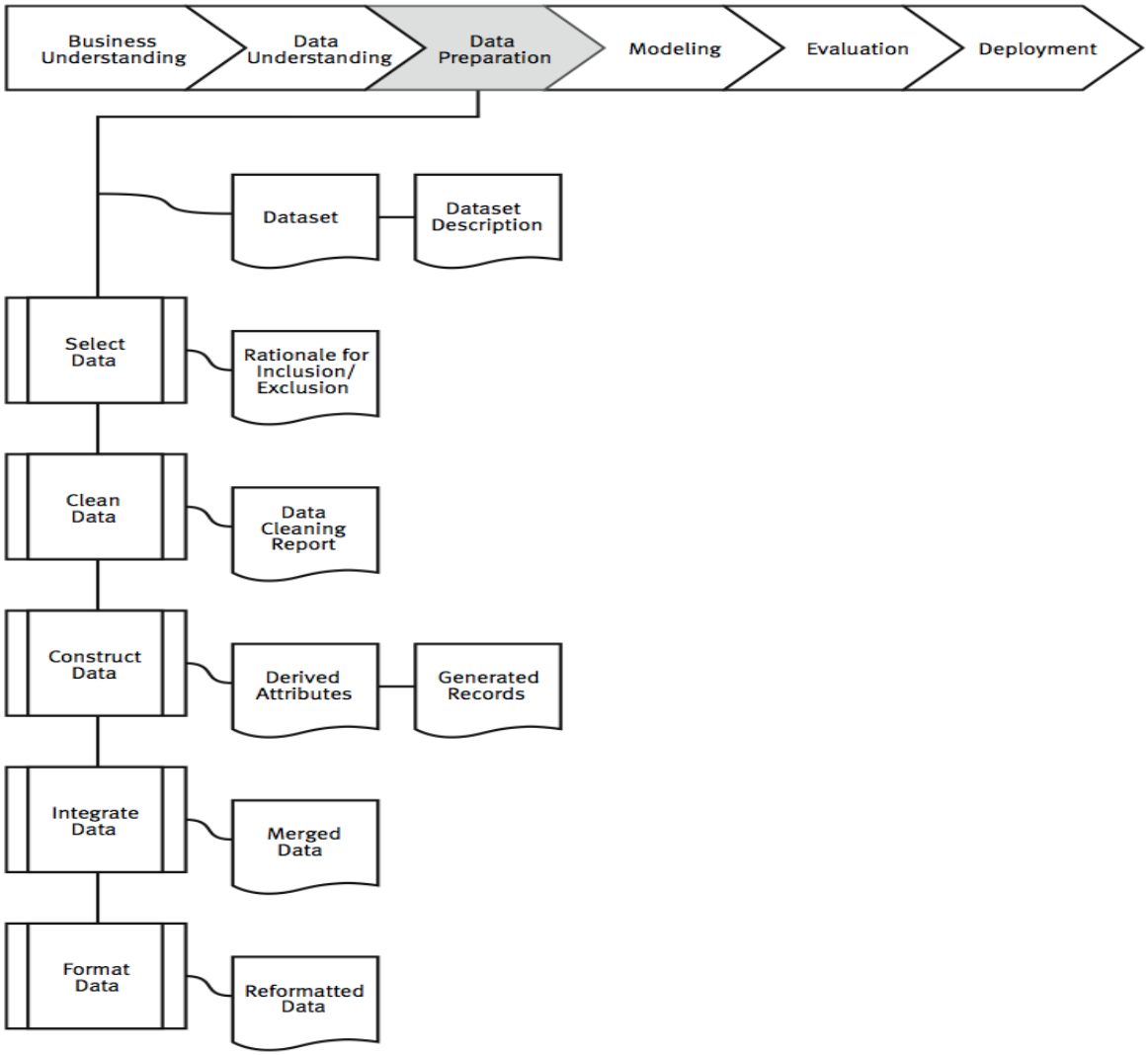


How to do data mining?

Data preparation:

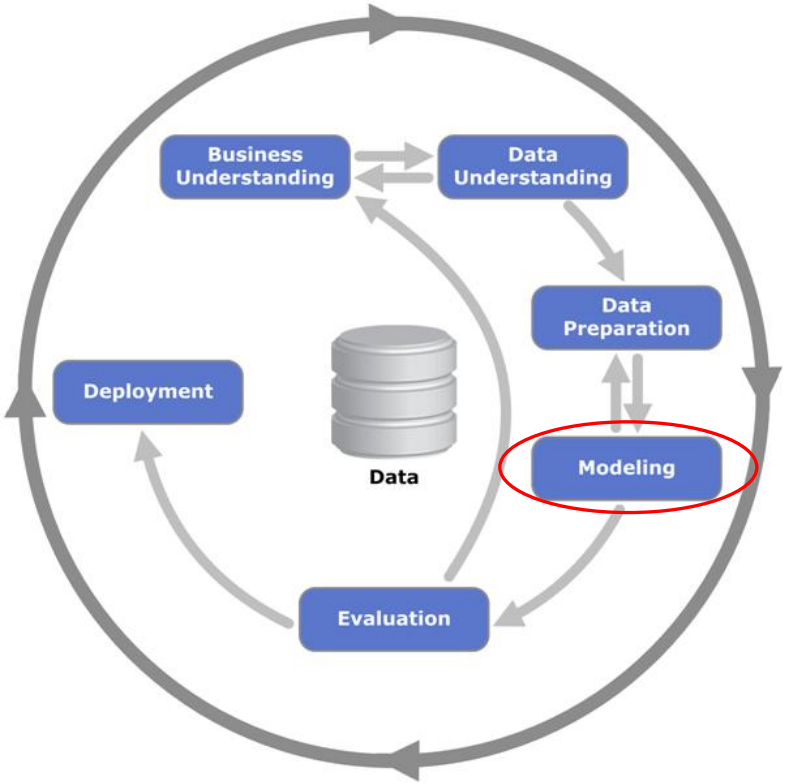
- Takes usually over 80% of the time
  - Collection
  - Assessment
  - Consolidation and Cleaning
    - table links, aggregation level, missing values, etc
  - Data selection
    - active role in ignoring non-contributory data?
    - outliers?
    - Use of samples
    - visualization tools
  - Transformations - create new variables

Data preparation

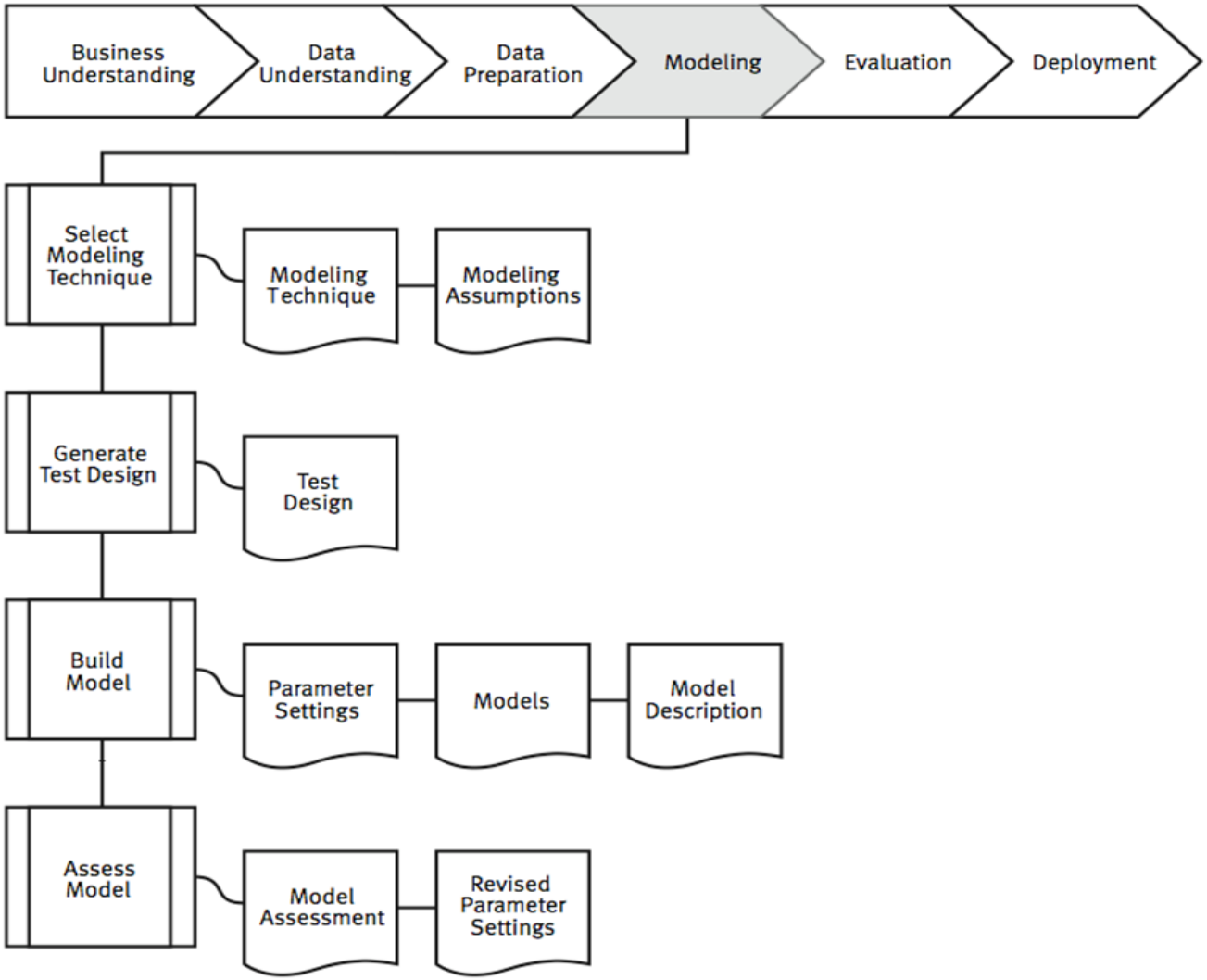


How to do data mining?

- Model building
  - Selection of the modeling techniques
  - Parameter tuning
  - Model assessment (ranking)

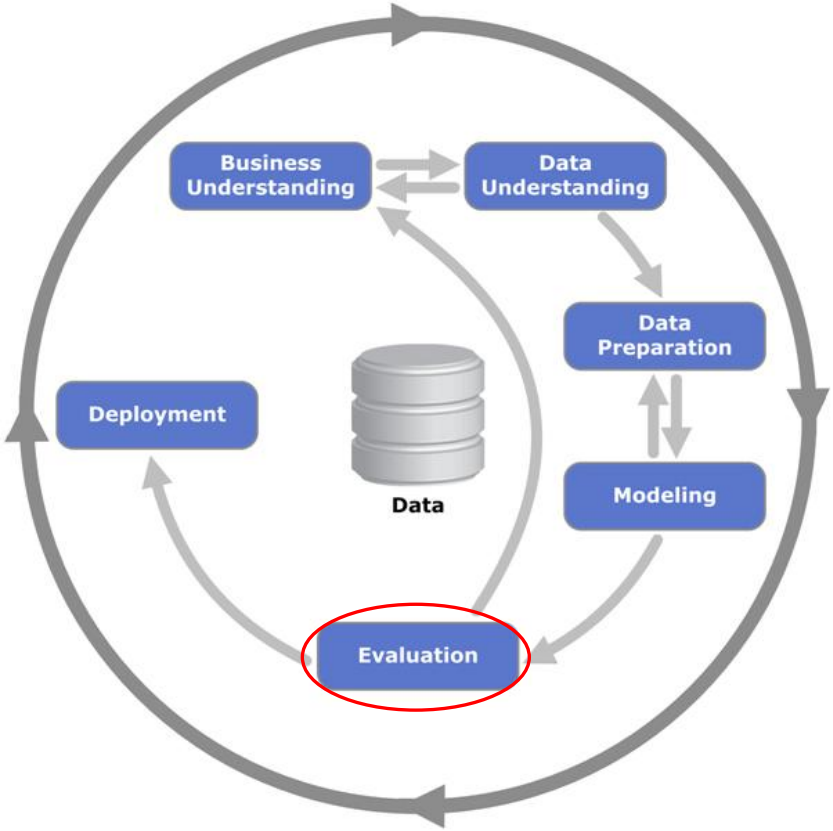


Modeling

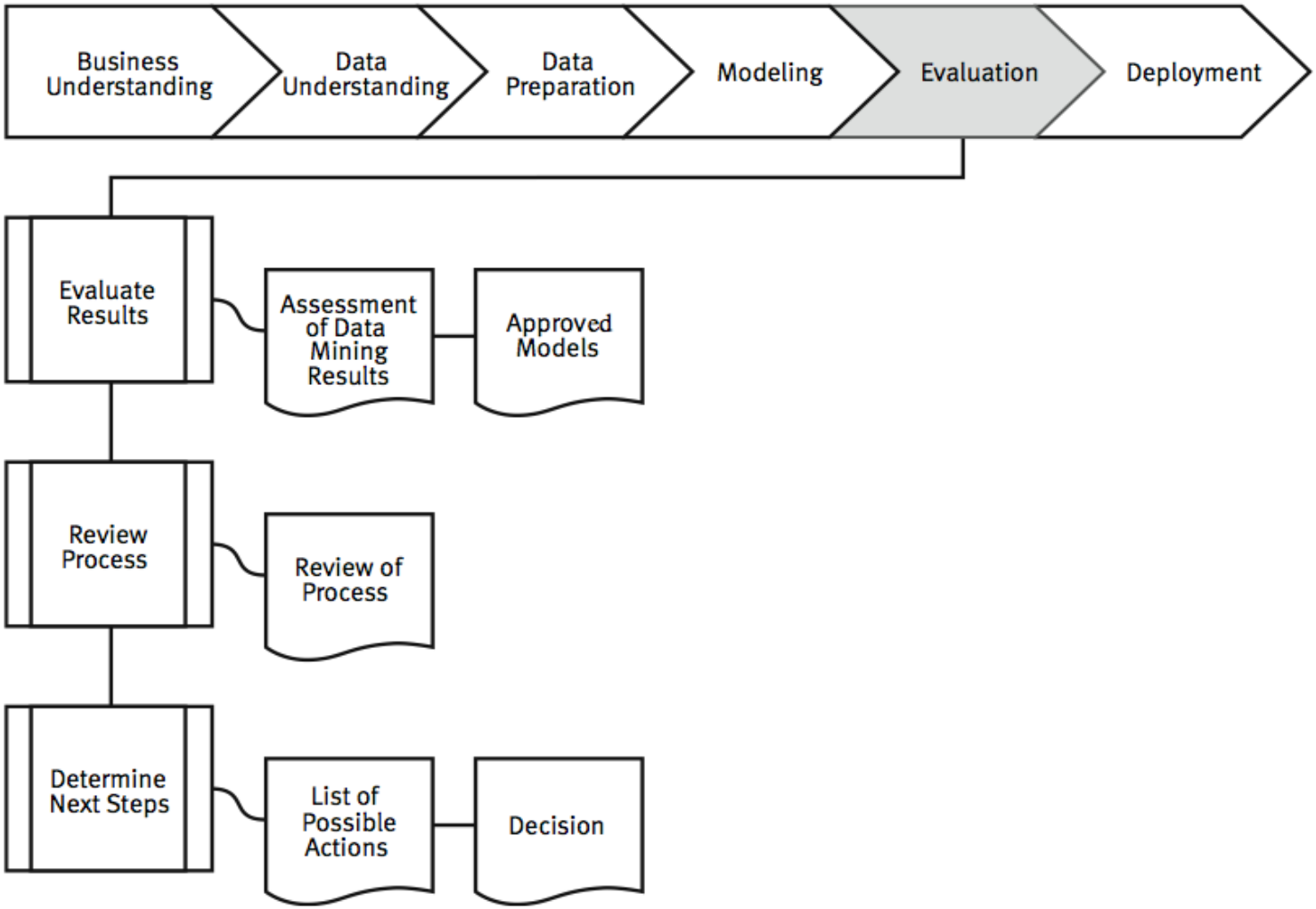

Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Select Modeling Technique
Modeling Technique
Modeling Assumptions
Generate Test Design
Test Design
Build Model
Parameter Settings
Models
Model Description
Assess Model
Model Assessment
Revised Parameter Settings

How to do data mining?

- Model Evaluation
  - Evaluation of model: how well it performed on test data
  - Methods and criteria depend on model type:
    - e.g., coincidence matrix with classification models, mean error rate with regression models
  - Interpretation of model: important or not, easy or hard depends on algorithm

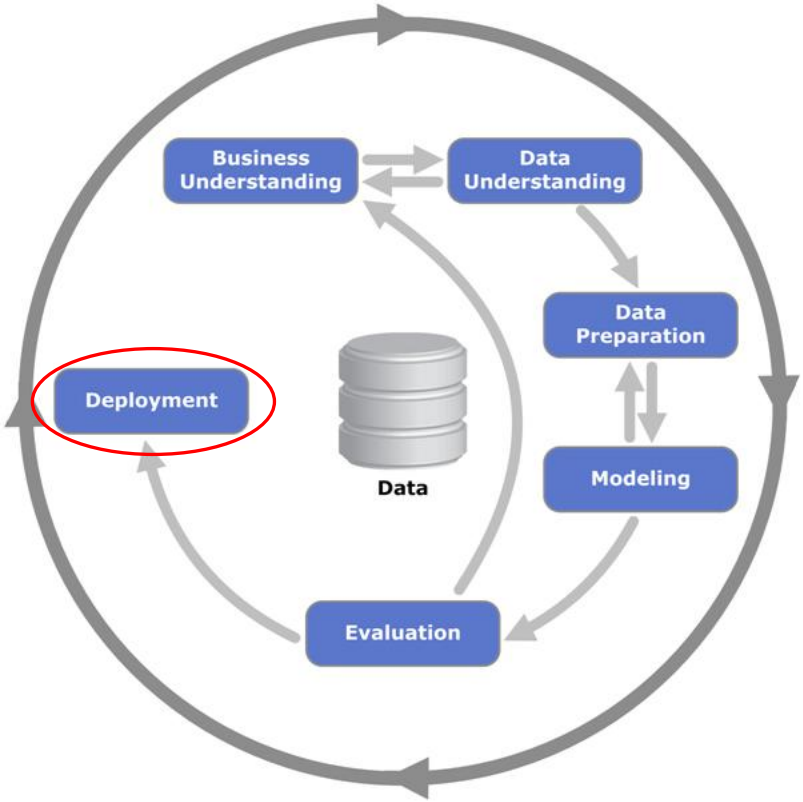


Evaluation

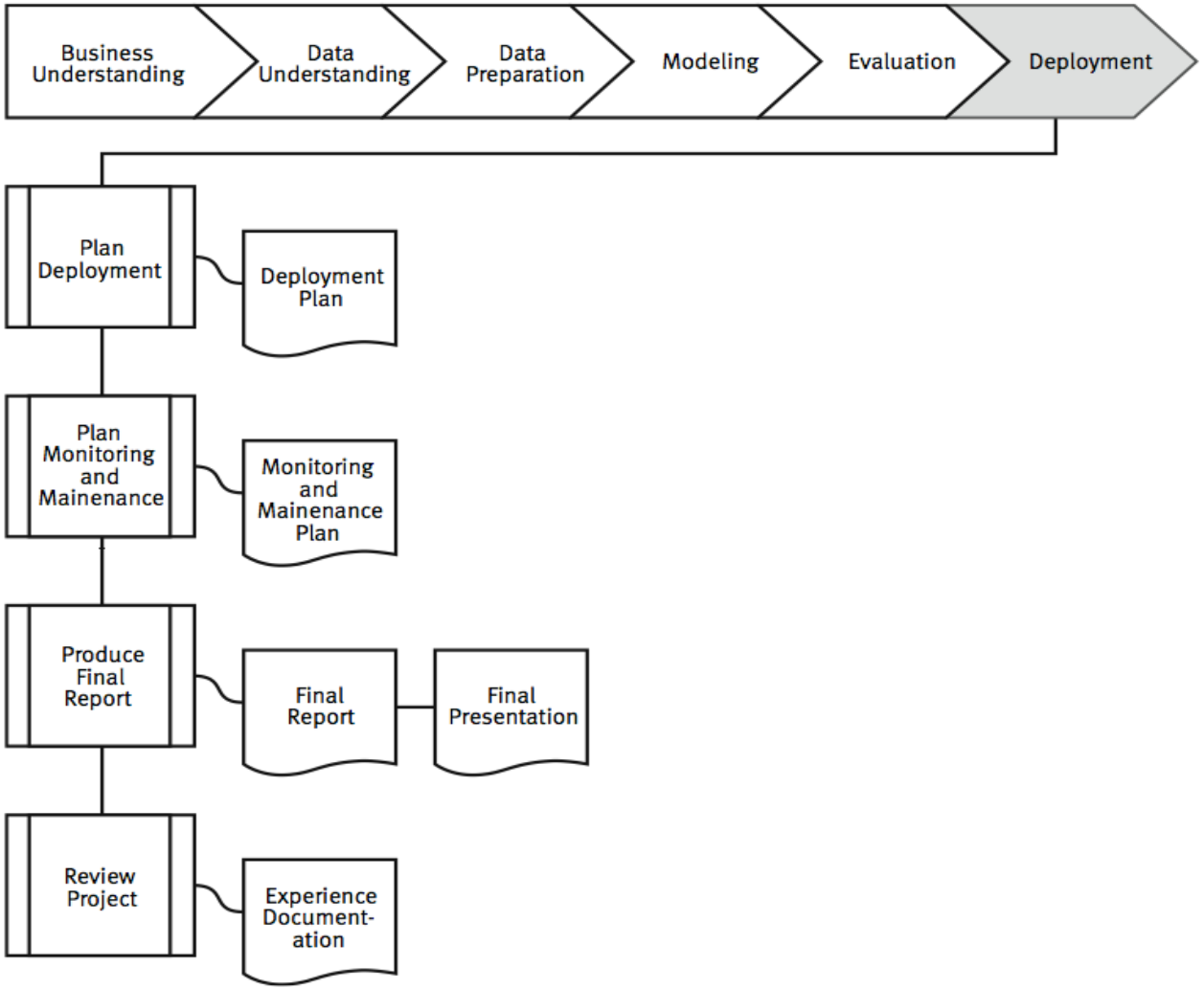

Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Evaluate Results
Assessment of Data Mining Results
Approved Models
Review Process
Review of Process
Determine Next Steps
List of Possible Actions
Decision

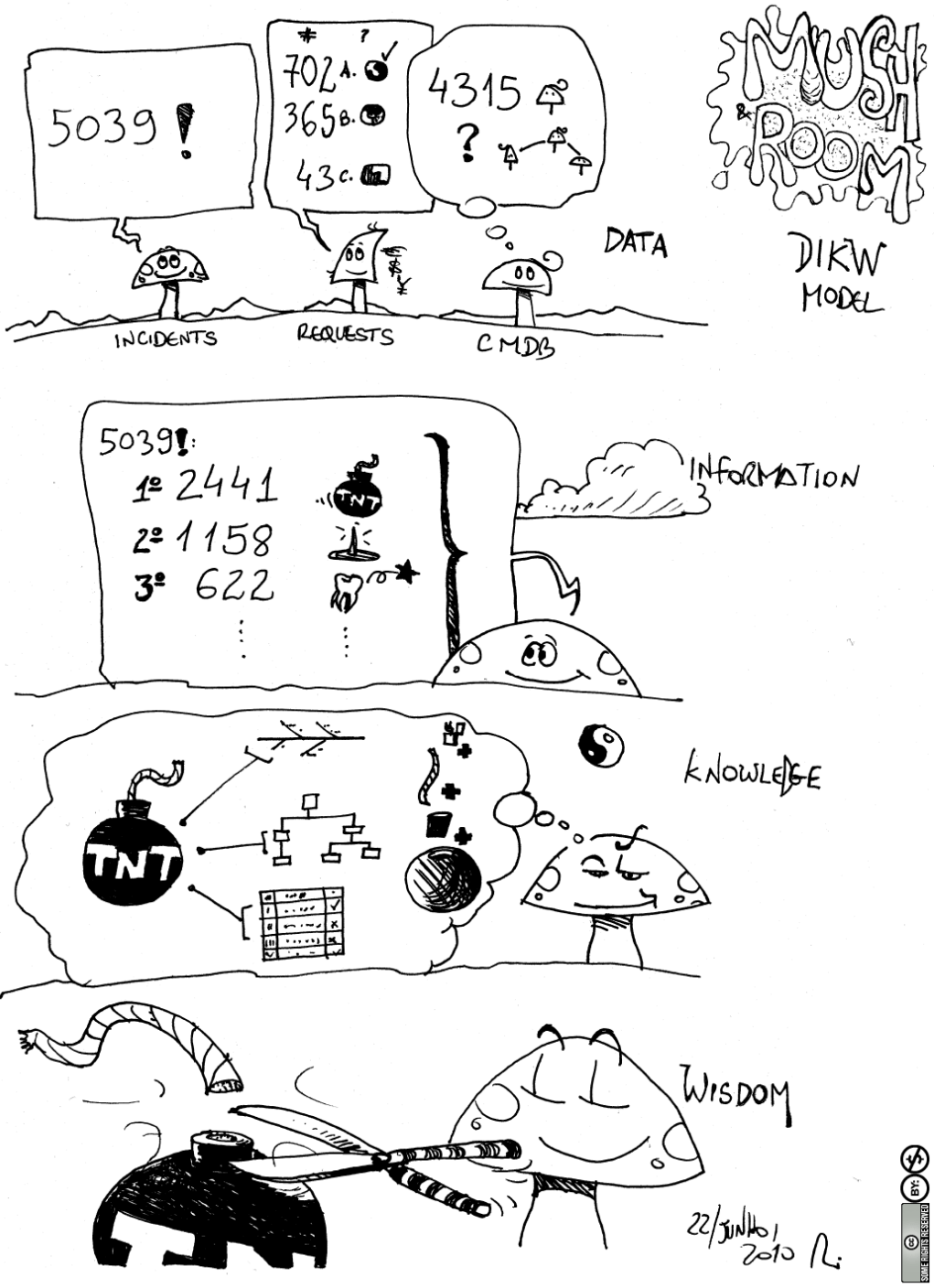
How to do data mining?

- Deployment
  - Determine how the results need to be utilized
  - Who needs to use them?
  - How often do they need to be used
- Deploy Data Mining results by:
  - Scoring a database
  - Exploiting results as business rules
  - Interactive scoring on-line



Deployment


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Plan Deployment
Deployment Plan
Plan Monitoring and Mainenance
Monitoring and Mainenance Plan
Produce Final Report
Final Report
Final Presentation
Review Project
Experience Document-ation



... but remember

Questions?