# Analisi dei dati ed estrazione di conoscenza

## Mastering Data Mining

Fosca Giannotti

Pisa KDD Lab, ISTI-CNR & Univ. Pisa

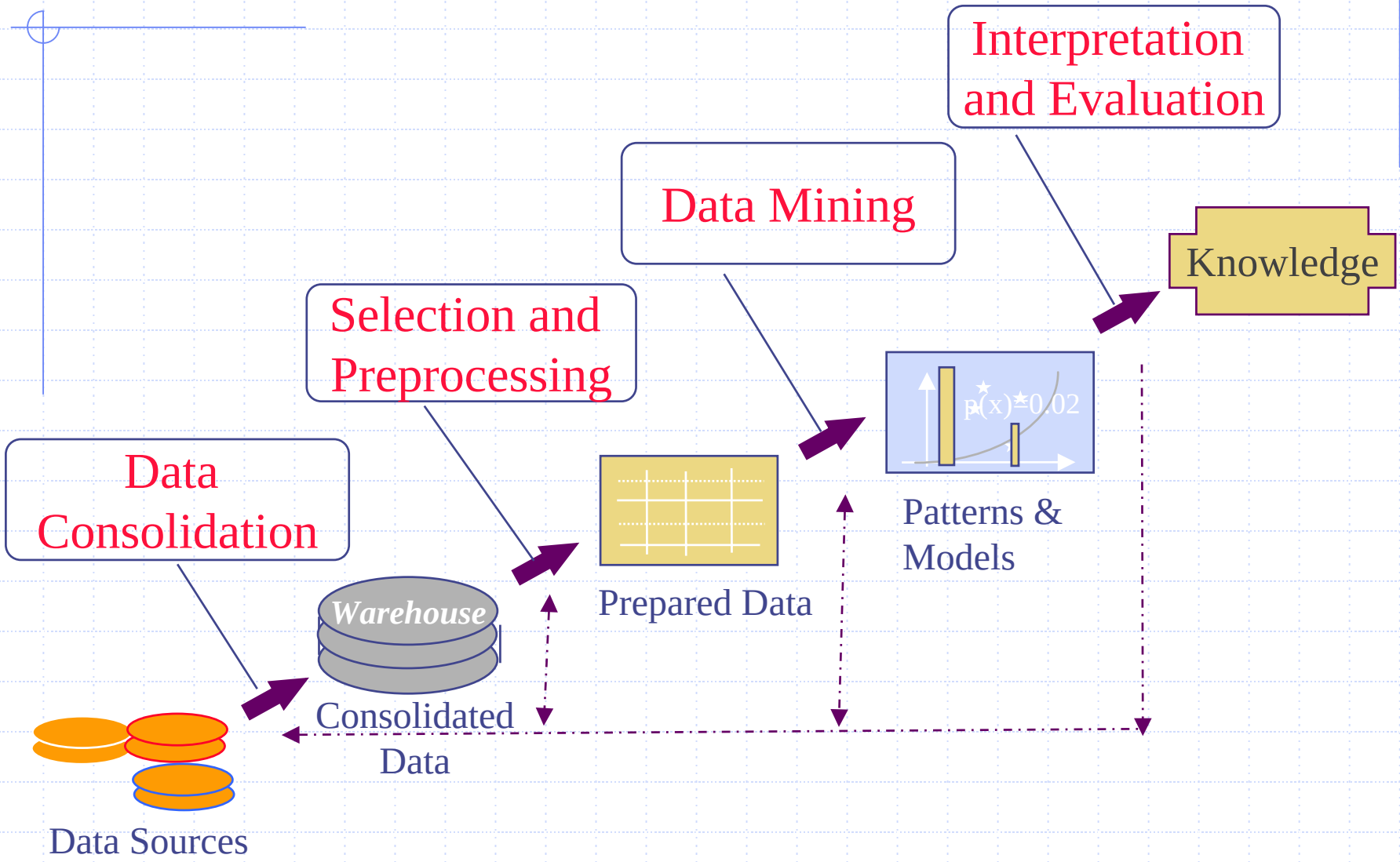
http://www-kdd.isti.cnr.it/

DIPARTIMENTO DI INFORMATICA - Università di Pisa

anno accademico 2005/2006

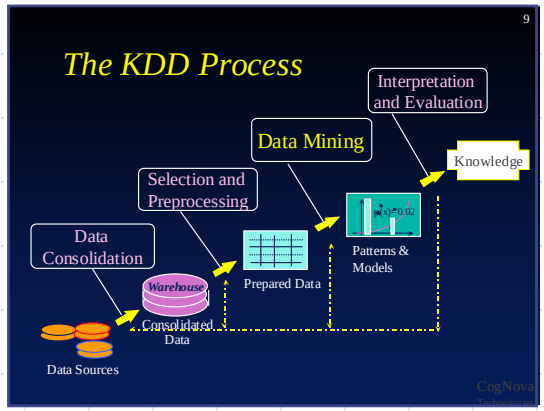
# Mastering Data Mining

# The KDD process



# The *virtuous cycle*

Problem




Knowledge

Identify Problem or Opportunity

Act on Knowledge

Strategy

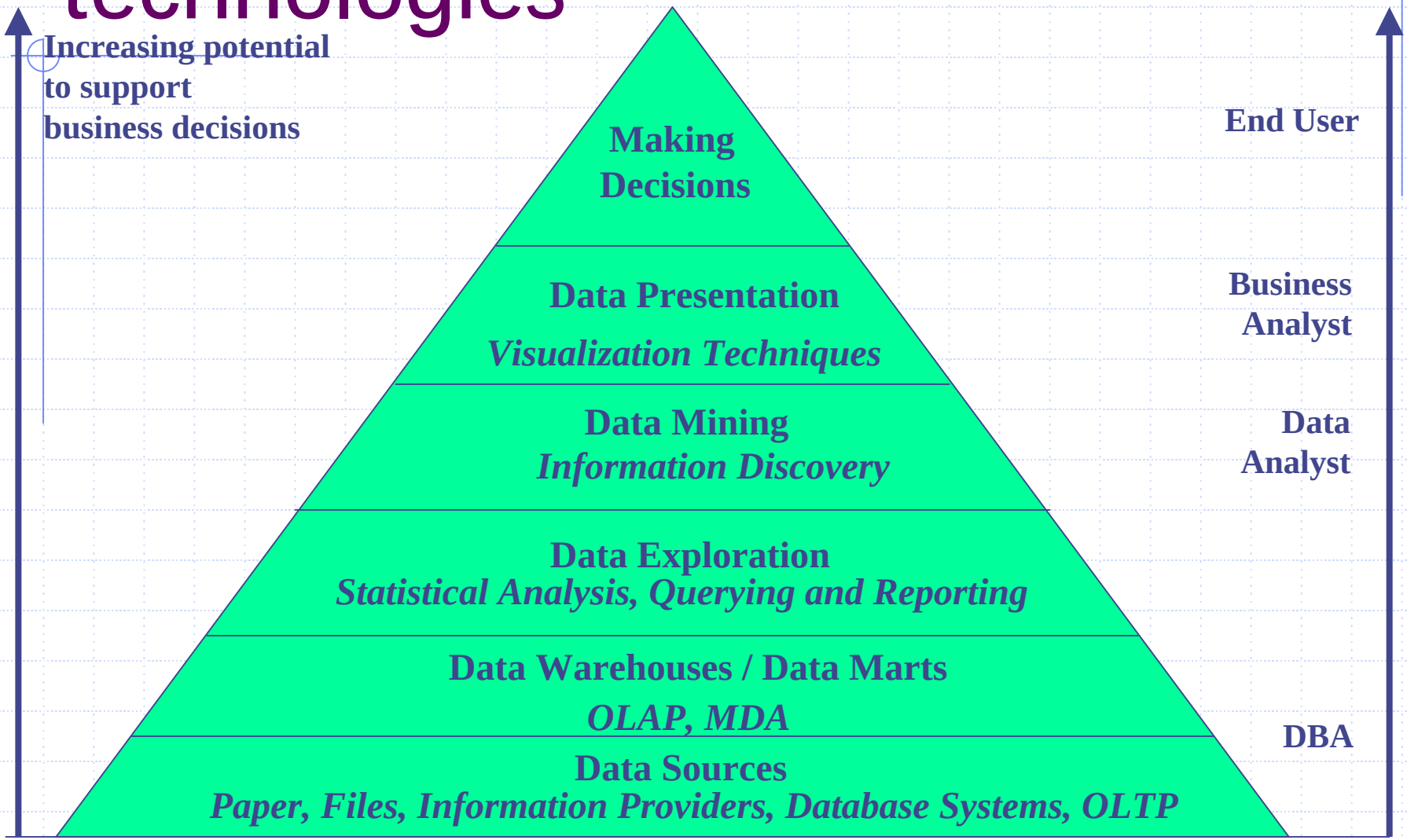
Measure effect of Action

Results

# Business Intelligence

- Business Intelligence is a global term for all the processes, techniques and tools that support business decision-making based on information technology.
- The approaches can range from a simple spreadsheet to a major competitive undertaking.
- Data mining is an important new component of business undertaking.

# Business intelligence technologies

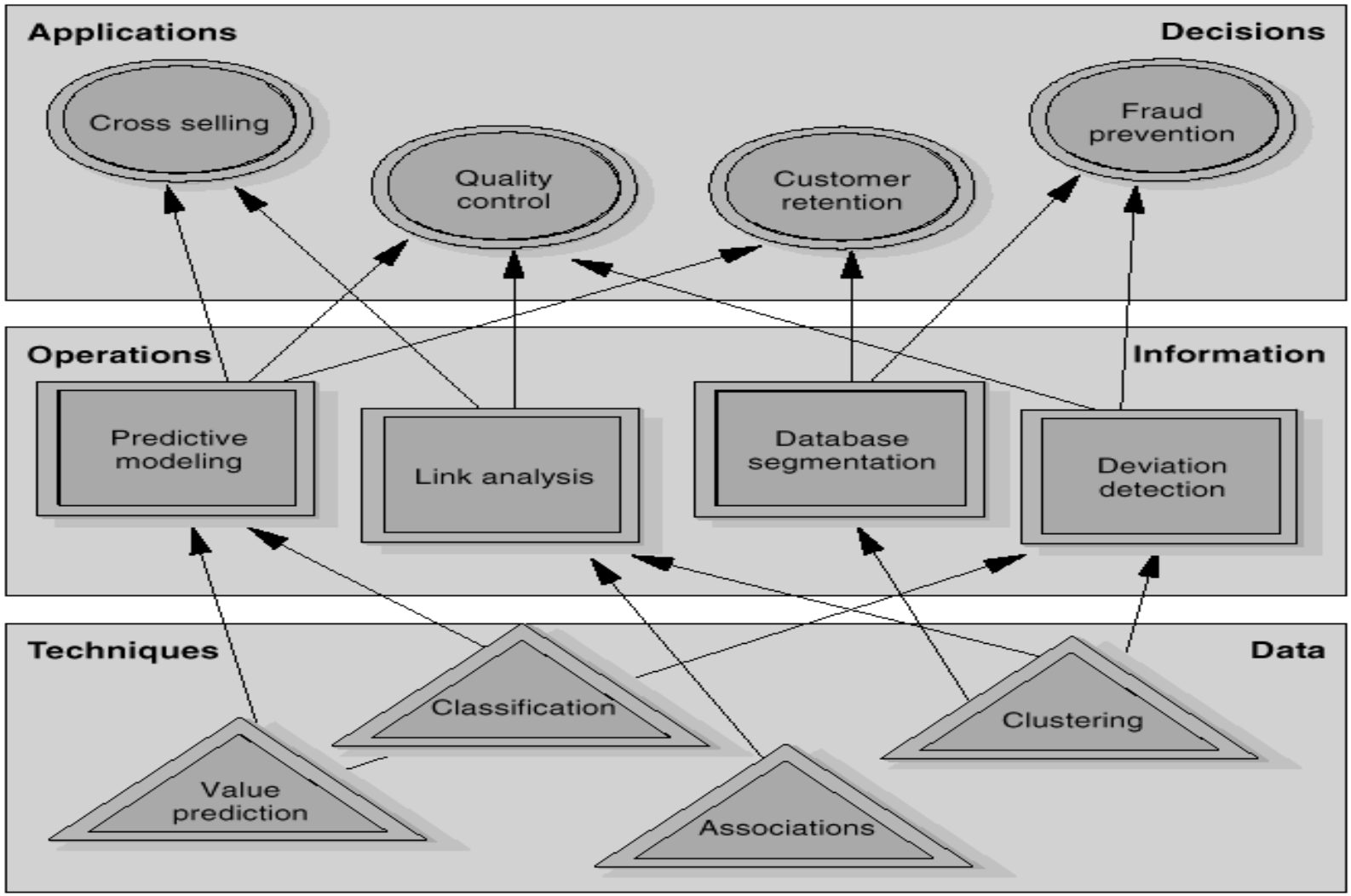


# Analogia: Piramide di Anthony

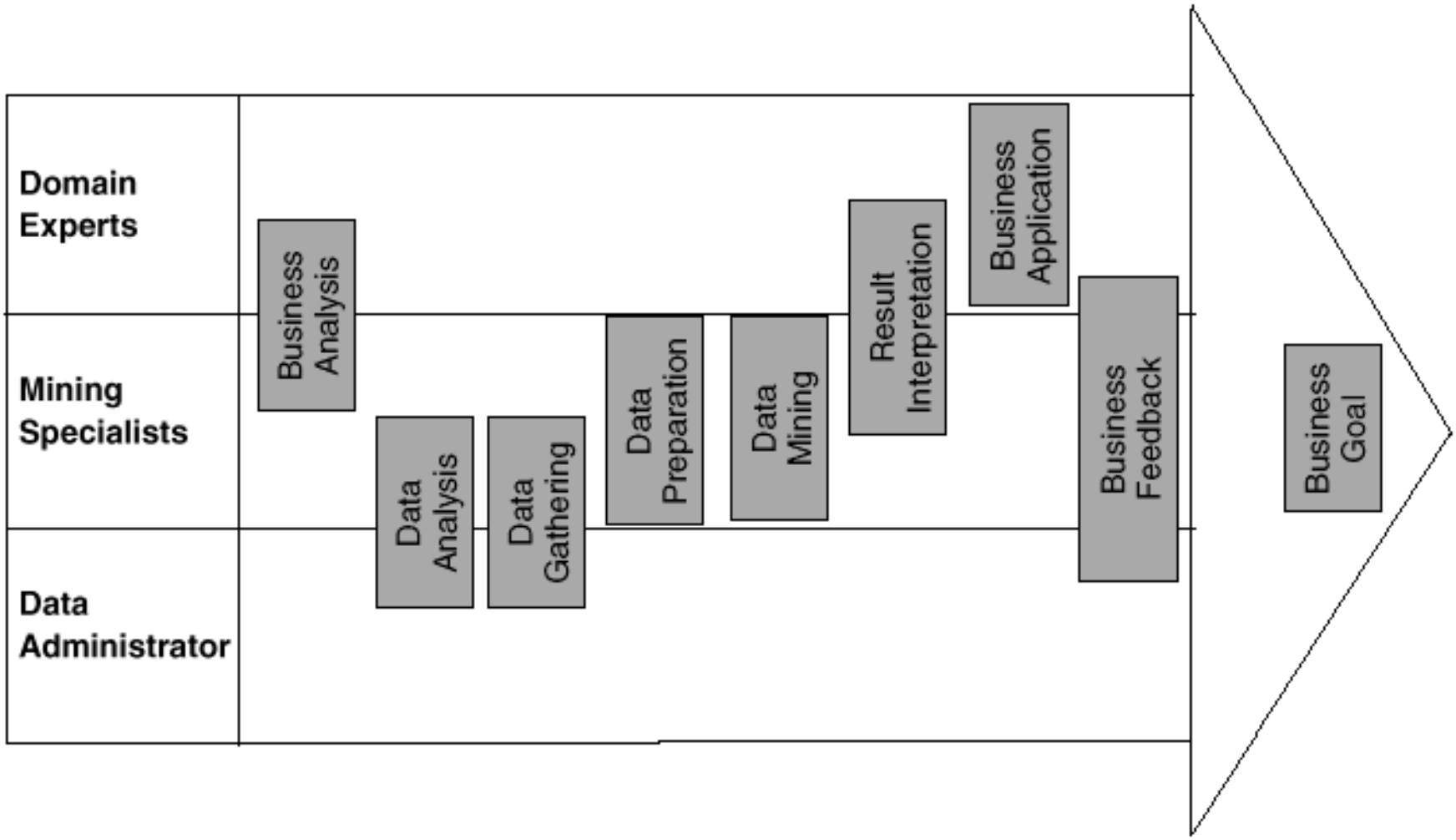
- classifica le attività svolte in un'organizzazione
- identifica il ruolo dei sistemi informatici a supporto di tali attività.


| | |
|---|---|
| **Attività strategiche** | -Scelta degli obiettivi aziendali<br>-Scelta delle risorse per il loro conseguimento<br>- Definizione delle politiche di comportamento aziendale |
| **Attività tattiche** | -Programmazione delle risorse disponibili<br>- Controllo sul conseguimento degli obiettivi programmati |
| **Attività operative** | -Conduzione a regime delle attività aziendali |

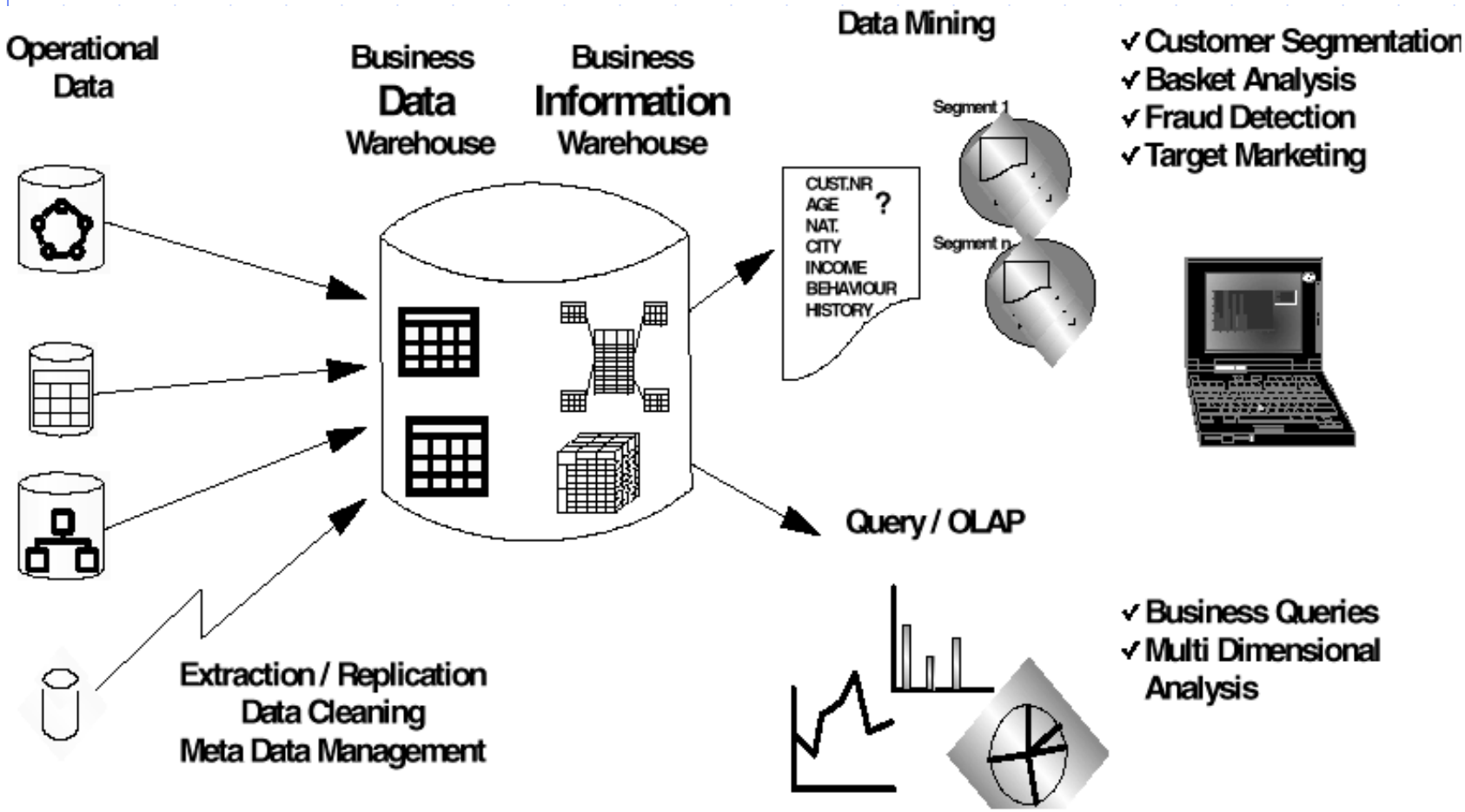
# Applications, operations,



# Roles in the KDD process

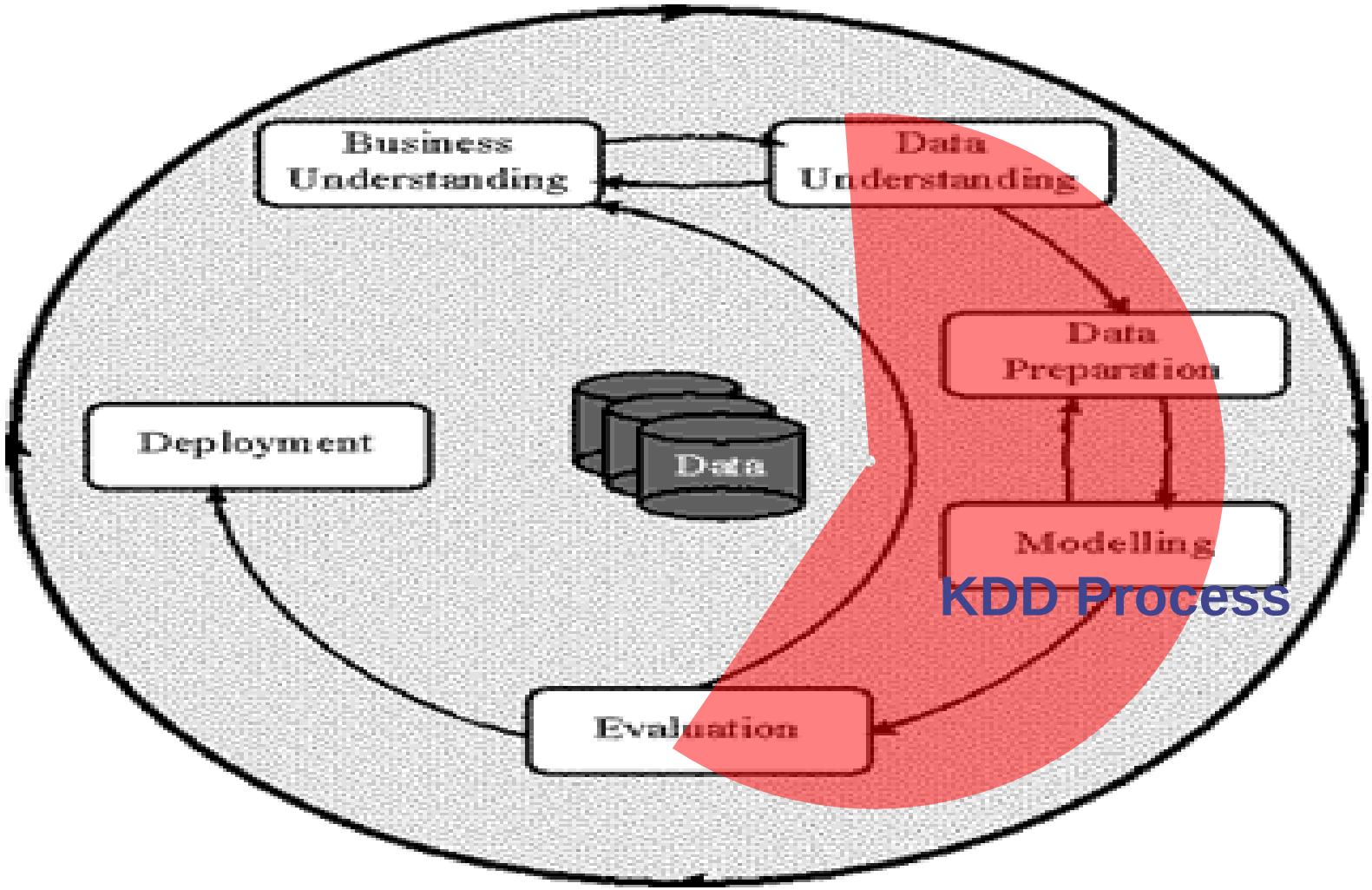


# A business intelligence environment



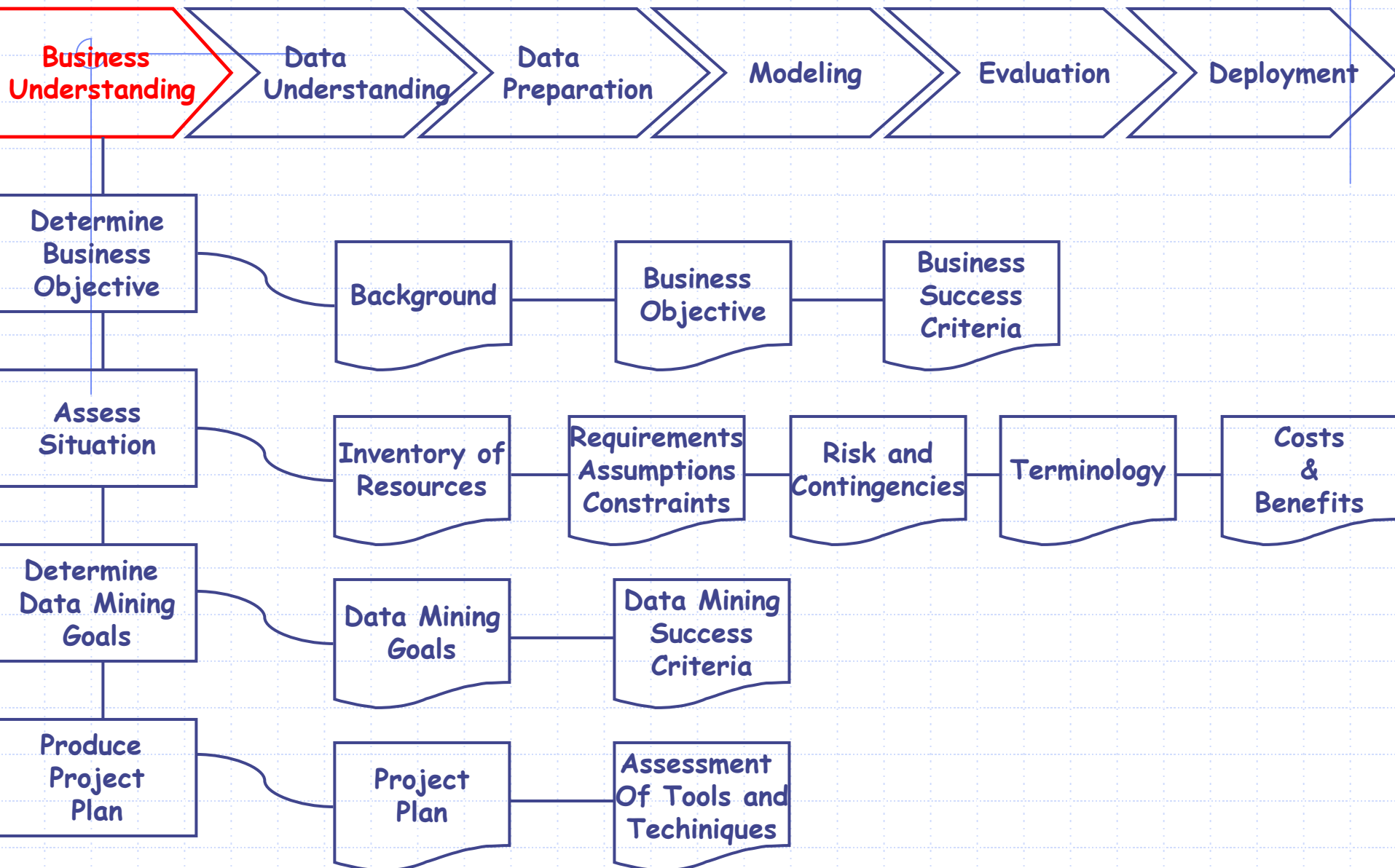
# How to develop a Data Mining Project?

# **CRISP-DM:** The life cicle of a data mining project



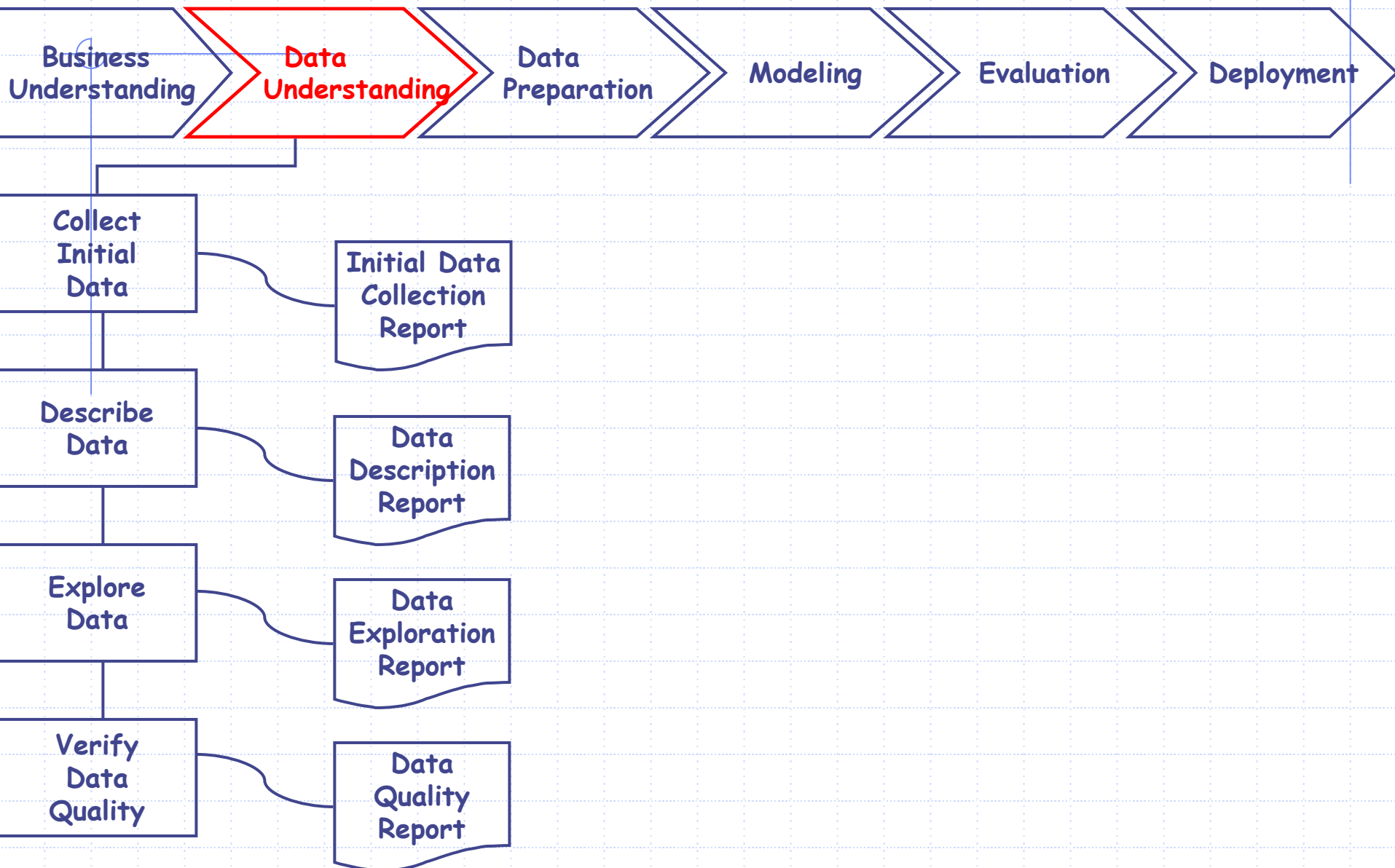
# Business understanding

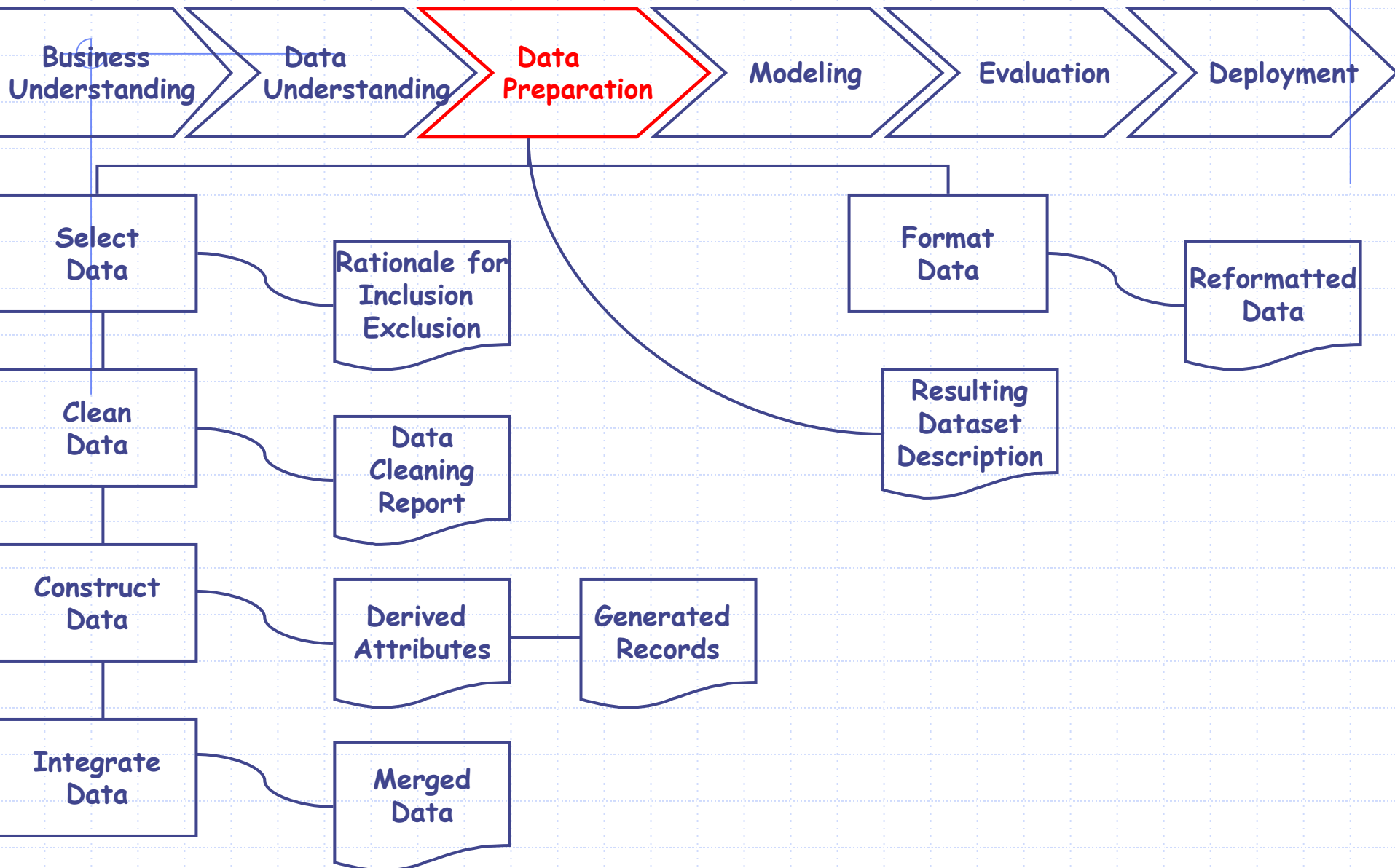
- Understanding the project objectives and requirements from a business perspective.
- then converting this knowledge into a data mining problem definition and a preliminary plan.
  - **Determine the Business Objectives**
  - **Determine Data requirements for Business Objectives**
  - **Translate Business questions into Data Mining Objective**


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Determine Business Objective
Background
Business Objective
Business Success Criteria
Assess Situation
Inventory of Resources
Requirements Assumptions Constraints
Risk and Contingencies
Terminology
Costs & Benefits
Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria
Produce Project Plan
Project Plan
Assessment Of Tools and Techiniques

# Data understanding

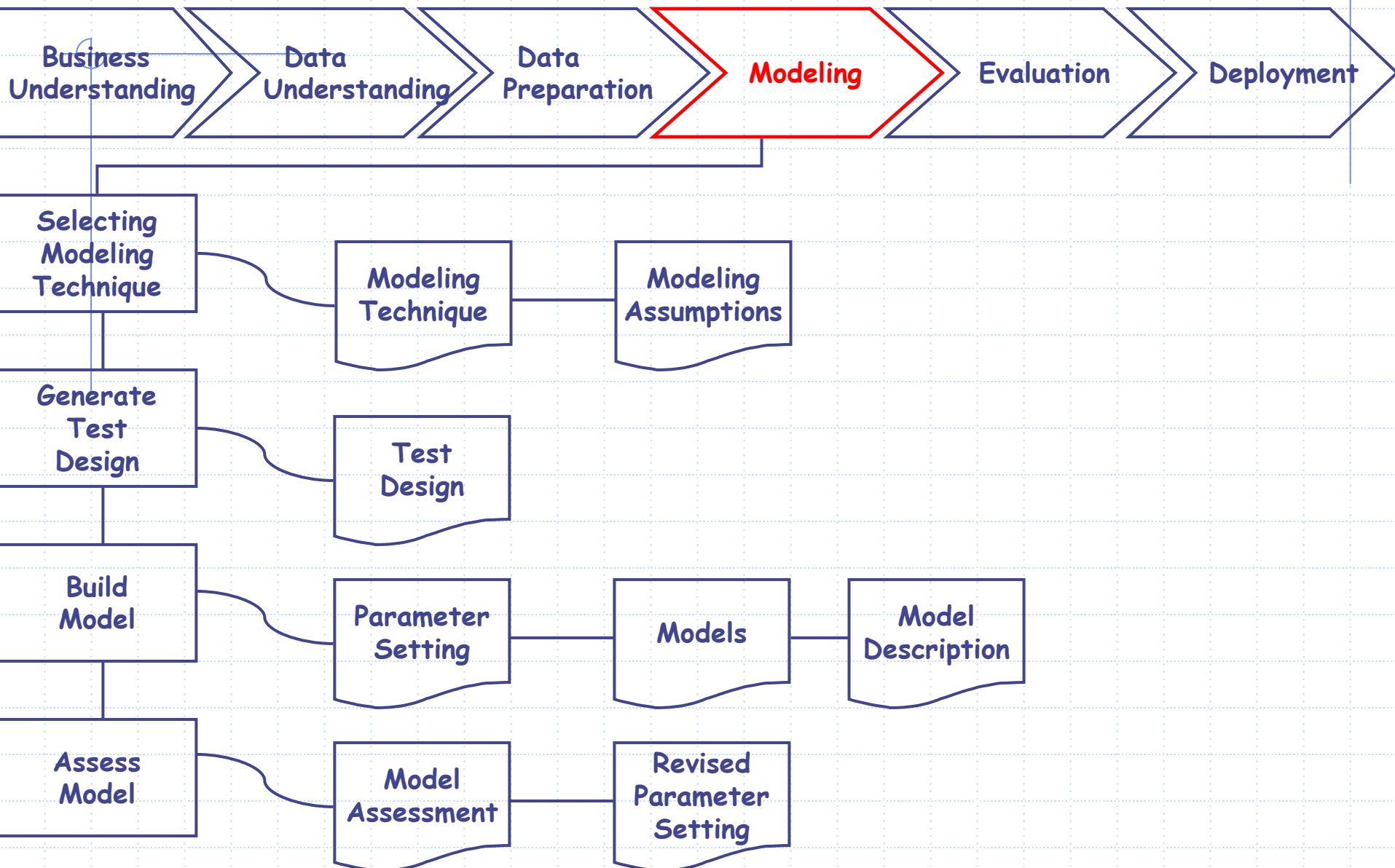
- **Data understanding:** characterize data available for modelling. Provide assessment and verification for data.


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Collect Initial Data
Initial Data Collection Report
Describe Data
Data Description Report
Explore Data
Data Exploration Report
Verify Data Quality
Data Quality Report


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
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
Resulting Dataset Description

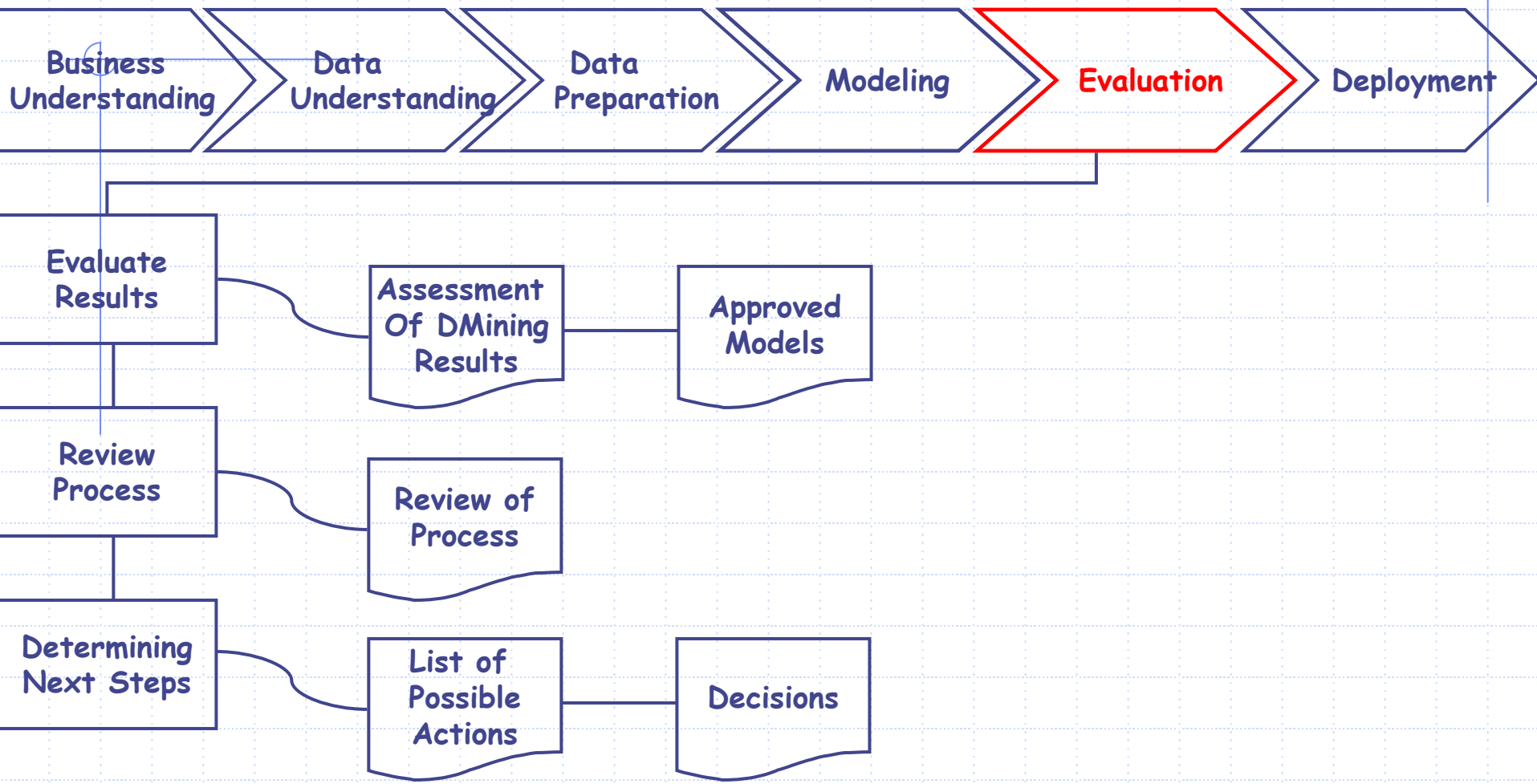
# Modeling:

- In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.
- Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data.
- Therefore, stepping back to the data preparation phase is often necessary.


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

# Evaluation

- At this stage in the project you have built a model (or models) that appears to have high quality from a data analysis perspective.
- Evaluate the model and review the steps executed to construct the model to be certain it properly achieves the business objectives.
- A key objective is to determine if there is some important business issue that has not been sufficiently considered.

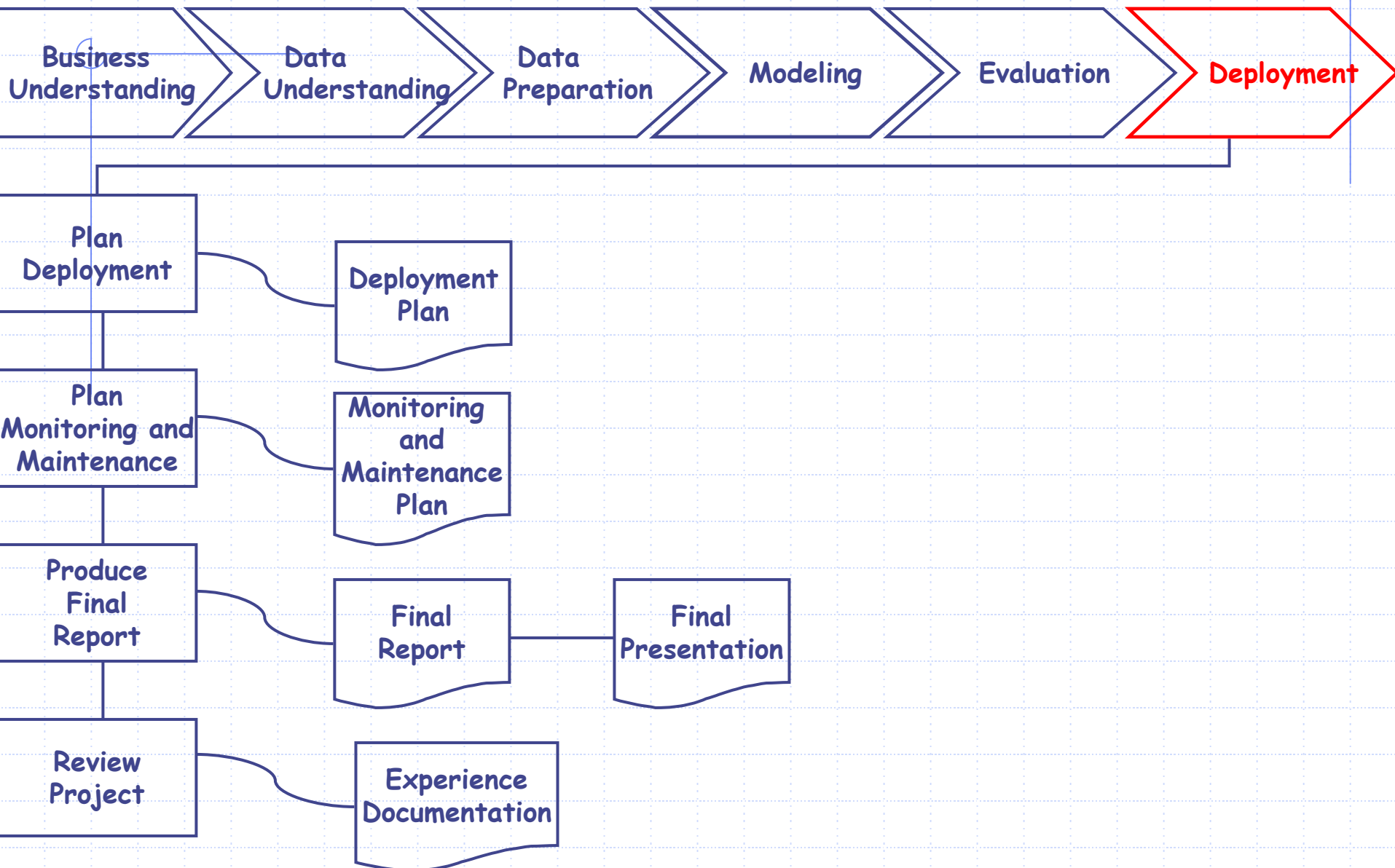

Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Evaluate Results
Assessment Of DMining Results
Approved Models
Review Process
Review of Process
Determining Next Steps
List of Possible Actions
Decisions

# Deployment:

- The knowledge gained will need to be organized and presented in a way that the customer can use it.
- It often involves applying “live” models within an organization’s decision making processes, for example in real-time personalization of Web pages or repeated scoring of marketing databases.

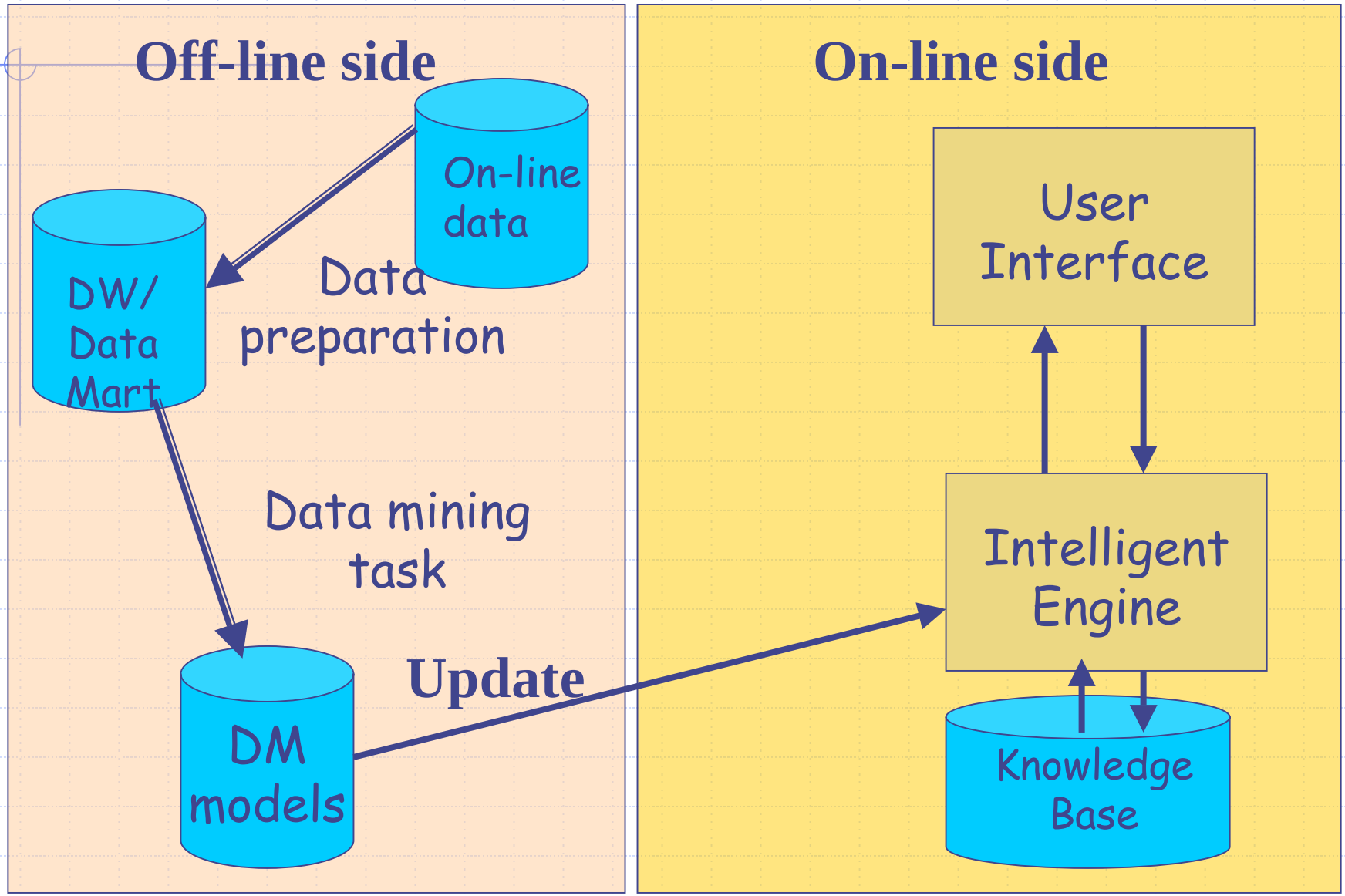
# Deployment:

- It can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.
- In many cases it is the customer, not the data analyst, who carries out the deployment steps.


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Plan Deployment
Deployment Plan
Plan Monitoring and Maintenance
Monitoring and Maintenance Plan
Produce Final Report
Final Report
Final Presentation
Review Project
Experience Documentation

# *Es: Automatic* Target Marketing

# Mining Based Decision Support System: Adaptive Architecture



# How to bring Data Mining to bear on a company's business problem

# A photography metaphor

- Mastering data mining means learning how to get data to tell a true and useful story
- Similar to mastering the art of photography – *Mastering Data Mining*, Barry Linoff 2002

# Using an automatic Polaroid

- Purchasing Scores from outside vendors as for example from Nielsen,
- Aggregate information from Istat
- Purchasing demographic overlay and surveys

# Using a fully automated camera

- To purchase software that embodies DM expertise directed toward a particular application
- Vertical products
- Neural Net for Credit Card Fraud detection
- Churn Management
- Customer Relationship Management (Decisionhouse)

# Hiring a wedding photographer

- By hiring outside consultants to perform predictive modelling for you for special projects
- Valuable in early stages
- Failing when all models, data, and insights generated are in the end of outsiders.
- The problem is **How** to use outside expertise
- “A prophet of another land may have more success in persuading the management of a new approach”
- Pilot projects with DM Labs.

# Building your own dark-room and becoming a skilled photographer

- Developing in house expertise
- A long term goal
- People which understand both the data and the business will build better models.

# The frontier of Data Mining

# New data and new applications

- specificità della struttura dei dati da analizzare (sequenze, grafi, stream, testi, dati semistrutturati) tipiche in settori applicativi emergenti quali bioinformatica, biologia ed il mondo Web.
- Specificità dell'applicazione finale come la necessità di incapsulare le funzionalità di mining all'interno di processi automatici (Invisible Data Mining).

# Vertical DM and privacy

- Necessità di fornire all'utente possibilità di interazione ad alto livello in tutti i passi per personalizzare e validare il processo di estrazione di conoscenza rispetto ad una specifica conoscenza di dominio.
- Infine, un'altra problematica interessante proviene dalla necessità di garantire gli aspetti di privacy e sicurezza degli individui pur estraendo informazione aggregata e globale.

# Mining Data Streams:

- In many emerging applications data arrives and needs to be processed on a continuous basis, i.e., there is need for mining without the benefit of several passes over a static, persistent snapshot.

# Data Mining in Bioinformatics

- High-performance data mining tools will play a crucial role in the analysis of the ever-growing databases of bio-sequences/structures.

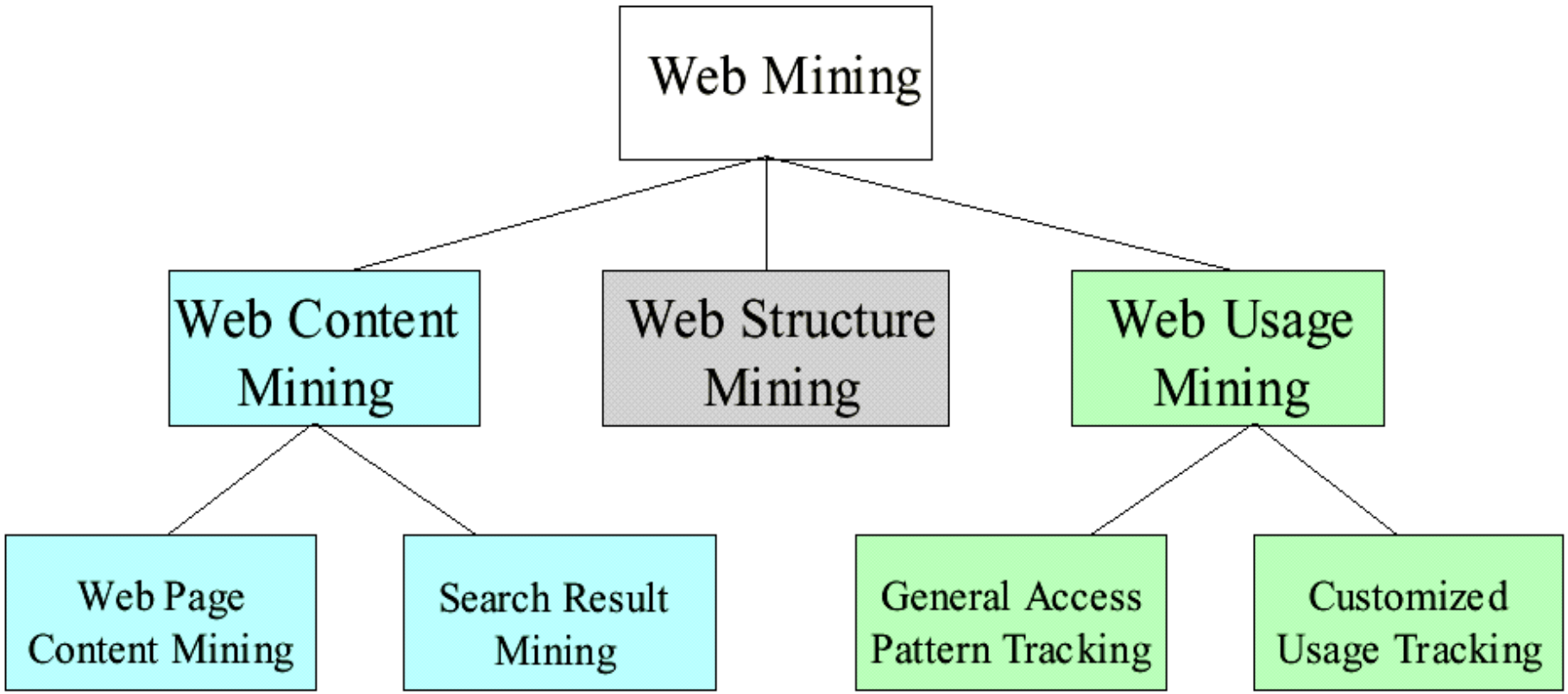
# Semi/Un-Structured Mining for the World Wide Web:

- The vast amounts of information with little or no structure on the web raise a host challenging mining problems such
- as web resource discovery and topic distillation; web structure/linkage mining;
- intelligent web searching and crawling; personalization of web content.

# Web Mining: A Fast Expanding Frontier in Data Mining

- Mine what Web search engine finds
- Automatic classification of Web documents
- Discovery of authoritative Web pages, Web structures and Web communities
- Meta-Web Warehousing: Web yellow page service
- Web usage mining

# Web Mining Taxonomy



# OLAP Mining: An Integration of Data Mining and Data Warehousing

- **Data mining systems, DBMS, Data warehouse systems coupling**
  - No coupling, loose-coupling, semi-tight-coupling, tight-coupling
- **On-line analytical mining data**
  - integration of mining and OLAP technologies
- **Interactive mining multi-level knowledge**
  - Necessity of mining knowledge and patterns at different levels of abstraction by drilling/rolling, pivoting, slicing/dicing, etc.
- **Integration of multiple mining functions**
  - Characterized classification, first clustering and then association