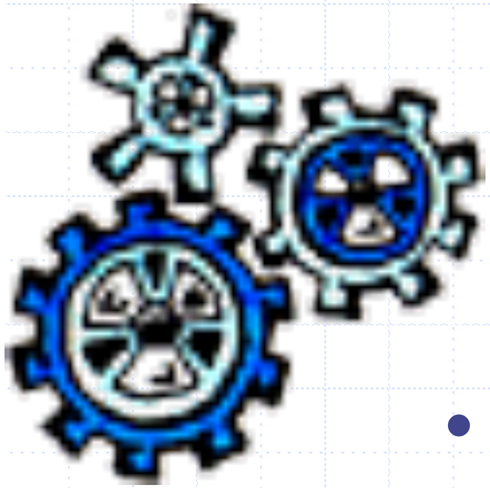
Data Mining2 – Advanced Aspects and Applications

Fosca Giannotti and Mirco Nanni

Pisa KDD Lab, ISTI-CNR & Univ. Pisa

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



- **DIPARTIMENTO DI INFORMATICA - Università di Pisa**

- **anno accademico 2013/2014**

Data Mining

- Monday 8:45-10:15 aula N1, Wednesday 8:45-10:15 aula L1
- Teachers:
  - Fosca Giannotti, ISTI-CNR, fosca.giannotti@isti.cnr.it
  - Mirco Nanni, ISTI-CNR, mirco.nanni@isti.cnr.it

- Question time: Wednesday 15-17, ISTI, Area Ricerca CNR, località San Cataldo, Pisa (send a request by e-mail)

Data Mining

- Reference
  - **Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to DATA MINING, Addison Wesley, ISBN 0-321-32136-7, 2006**
  - Barry Linoff Data Mining Techniques for Marketing Sales and Customer Support, John Wiles & Sons, 2002
- Slides available at: http://didawiki.cli.di.unipi.it
- Blog on privacy & DM
  - anna.monreale@isti.cnr.it
  - http://hd.media.mit.edu/wef_globalit.pdf

Data Mining- Theory

- KDD process: CRISP standard.
- Alternative Classification Methods,
- Sequential Pattern Mining,
- Mining of Time Series, spatial and spatio-temporal data
- Mining of graphs & Motifs
- Ethical issues of data mining - Data mining and personal data privacy

Data Mining – Applications

- Fraude Detection: Sogei1, DIVA (progetto 1)
- CRM: data set COOP, TargetMarketing, ChurnAnalysis: coop (progetto 2)
- E-health and Mining Official Data
- ICT digital traces: analysing GSM data: ORANGE, and WIND – visitor profile and .
- Mobility and Transportation: understanding human mobility. Progetto3

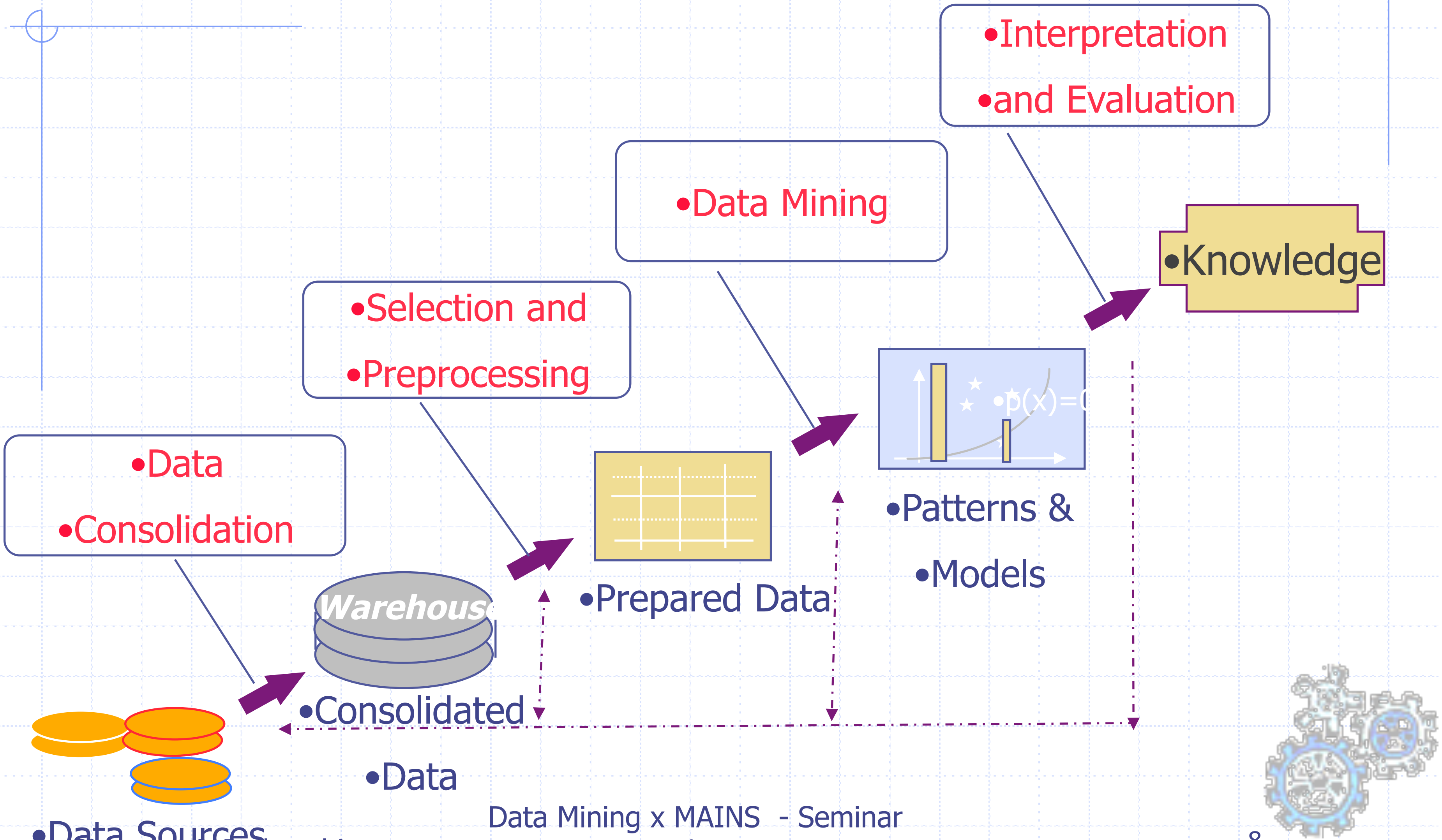
Evaluation

- Ongoing projects (on small datasets) with presentation to the class
- Team Project
  - Team of 2-3 person.
  - Unique grade.
  - Projects consist into the realization of some complete analytical processes on a given problem and a given dataset.
  - A final report followign the CRISP standard describing all steps: esploration, preparation and anaysis and final evaluation.
- Individual Project Discussion

Outline this lesson

- **KDD Process**
- CRM and Data Mining - AirMiles
- Health and Data Mining – Stulong
- CRISP model for developing Data Mining based services

The KDD process



The steps of the KDD process

- Learning the application domain:
  - relevant prior knowledge and goals of application
- Data consolidation: Creating a target data set
- Selection and Preprocessing
  - *Data cleaning* : (may take 60% of effort!)
  - *Data reduction and projection*:
    - find useful features, dimensionality/variable reduction, invariant representation.
- Choosing functions of data mining
  - summarization, classification, regression, association, clustering.
- Choosing the mining algorithm(s)
- Data mining: search for patterns of interest
- Interpretation and evaluation: analysis of results.
  - *visualization, transformation, removing redundant patterns*, …
- Use of discovered knowledge

The B.I. platform

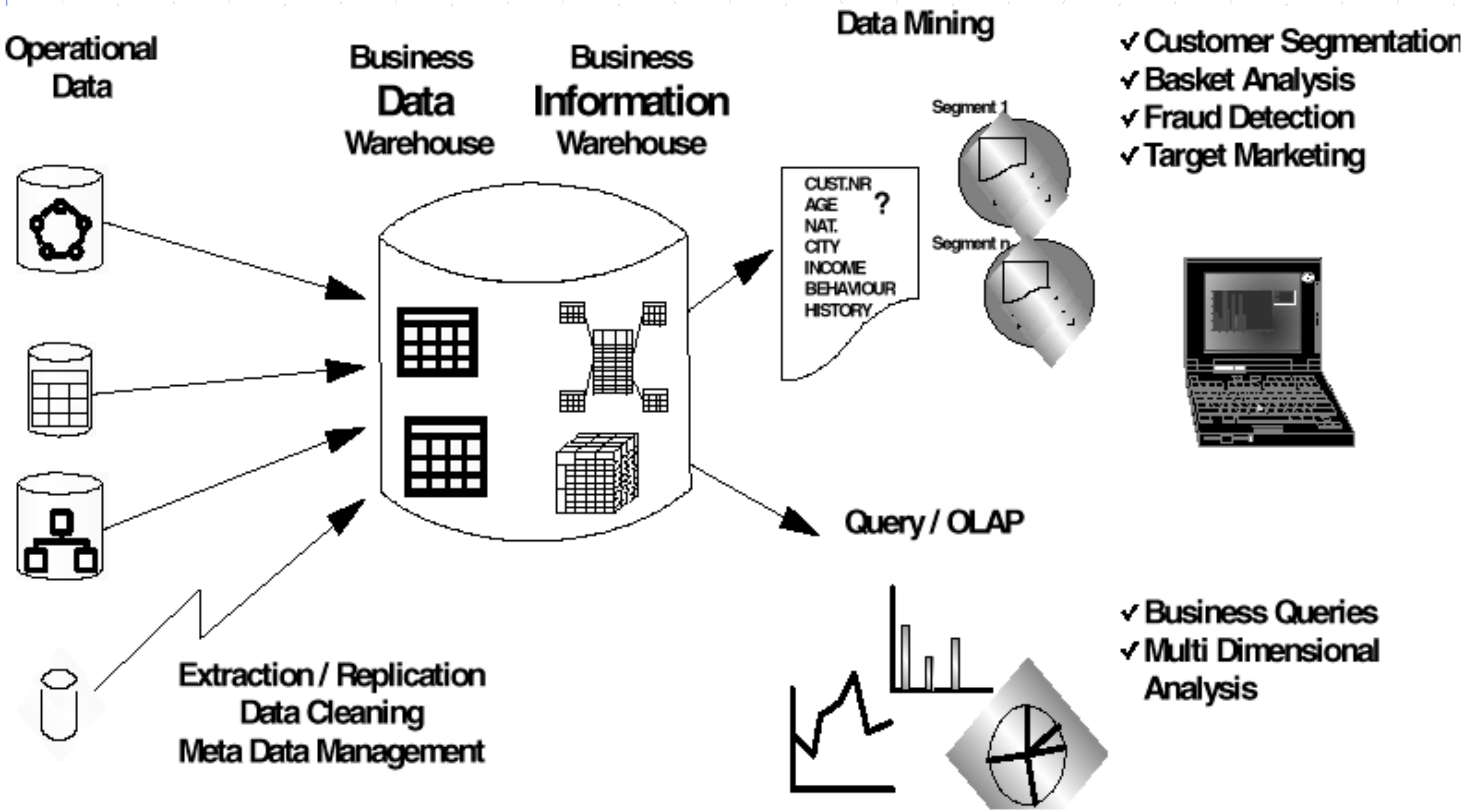


Figure per la B.I.

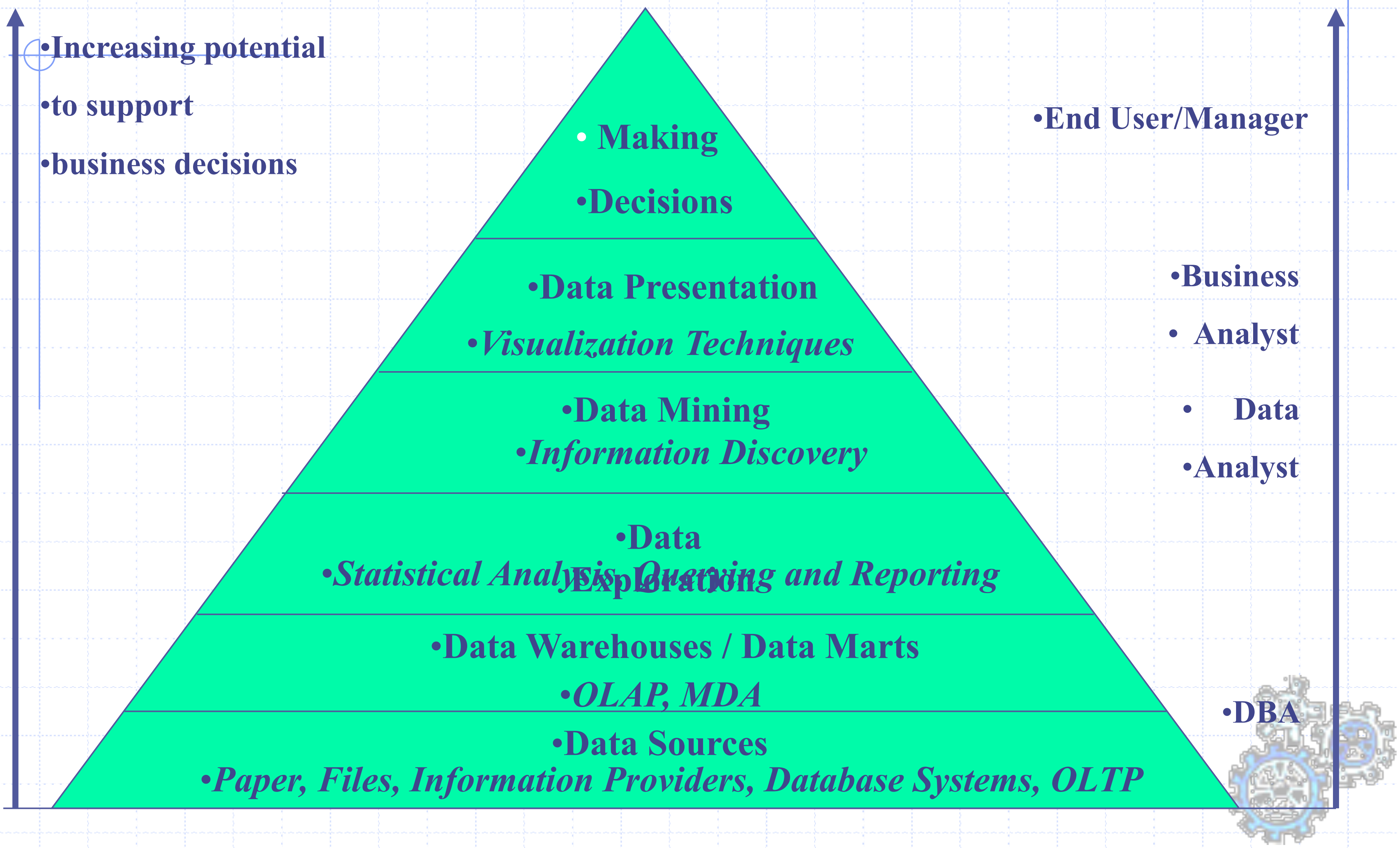
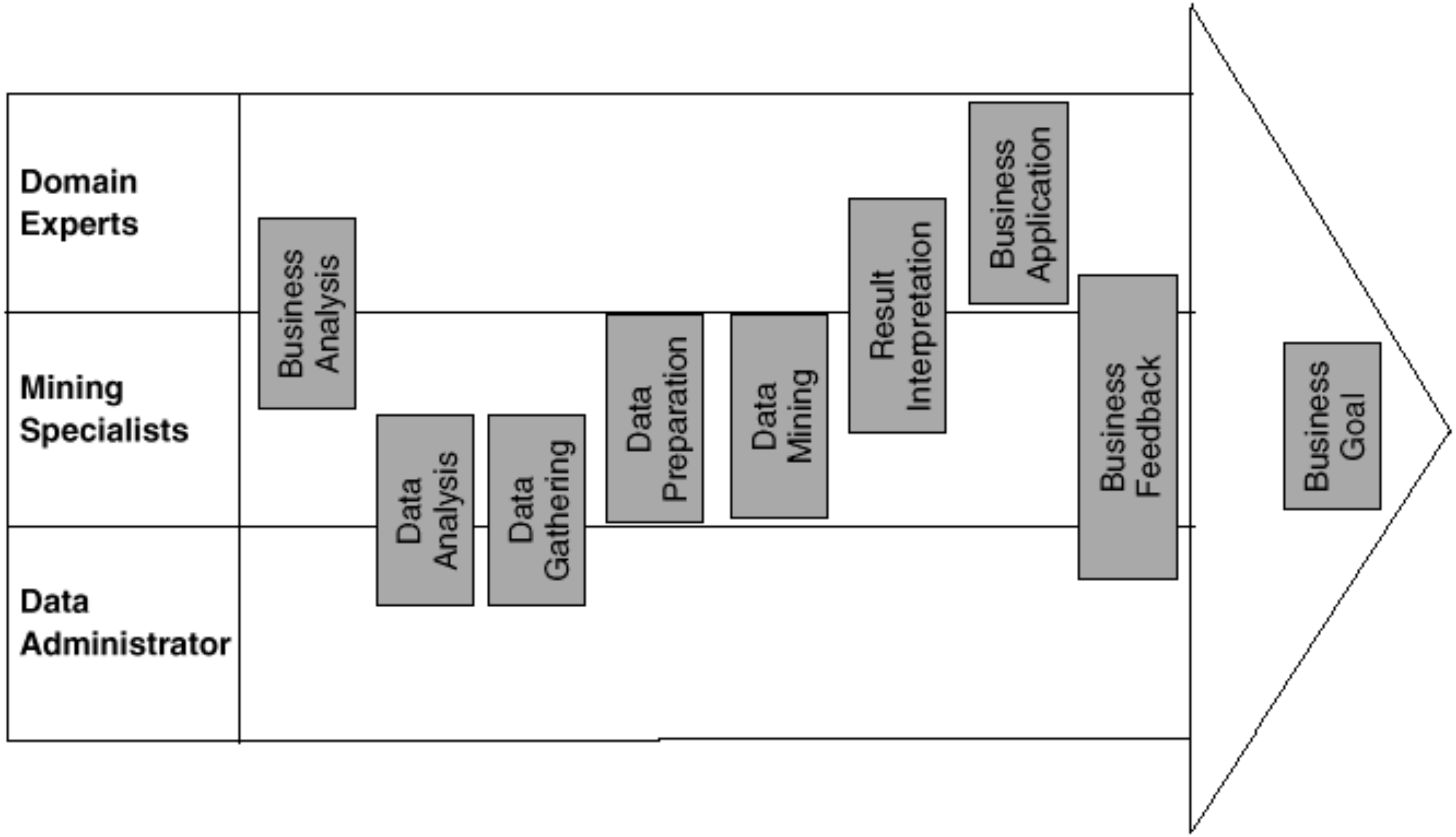


Figure nel processo di KDD



Intelligence/Value

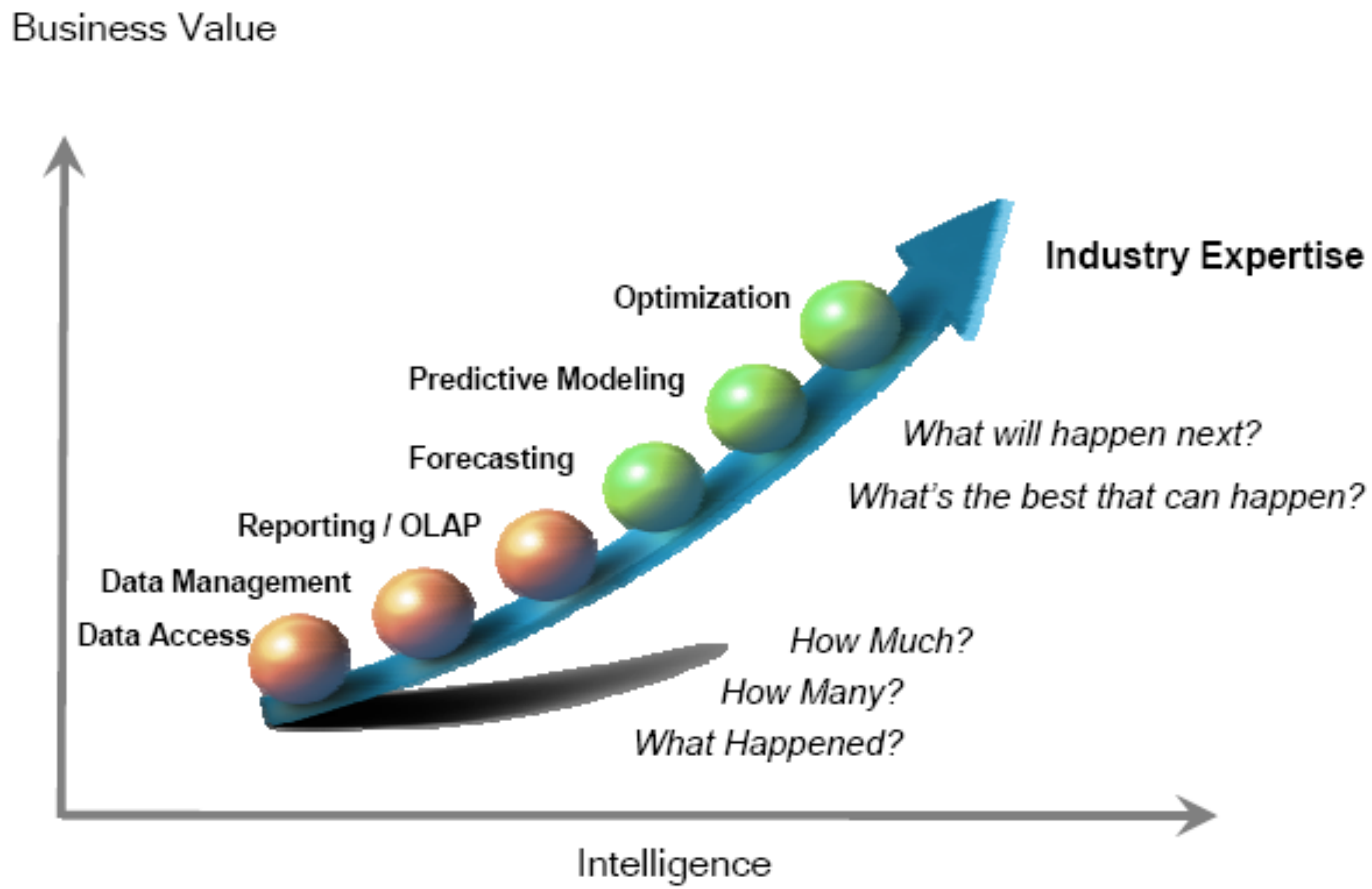


Figure 1: Business value increases exponentially with intelligence.

Outline this lesson

- KDD Process
- **CRM and Data Mining - AirMiles**
- Health and Data Mining – Stulong
- CRISP model for developing Data Mining based services

AIR MILES
a case-study on customer segmentation

From: G. Saarenvirta, “Mining customer data”, DB2 magazine on line, 1998
http://www.db2mag.com/98fsaar.html

Application: customer segmentation

- Given:
  - Large data base of customer data containing their properties and past buying records
- Goal:
  - Find groups of customers with similar behavior

Customer segmentation

- use customer-purchase transaction data to
  - track buying behavior
  - create strategic business initiatives.
- divide customers into segments based on "shareholder value" variables:
  - customer profitability,
  - measure of risk,
  - measure of the lifetime value of a customer,
  - retention probability.

Customer segments

- high-profit, high-value, and low-risk customer segment
  - typically 10% to 20% of customers who create 50% to 80% of a company's profits
  - strategic initiative for the segment is retention
- A low-profit, high-value, and low-risk customer segment may be also attractive
  - strategic initiative for the segment is to increase profitability
  - cross-selling (selling new products)
  - up-selling (selling more of what customers currently buy)

Behavioral vs. demographic segments

- Within behavioral segments, a business may create demographic subsegments.
- Customer demographic data are not typically used together with behavioral data to create segments.
- Demographic (sub)segmenting is used to select appropriate tactics (advertising, marketing channels, and campaigns) to satisfy the strategic behavioral segment initiatives.

The Loyalty Group in Canada

- runs an AIR MILES Reward Program (AMRP) for a coalition of more than 125 companies in all industry sectors - finance, credit card, retail, grocery, gas, telecom.
- 60% of Canadian households enrolled
- AMRP is a frequent-shopper program:
  - the consumer collects bonuses that can then redeem for rewards (air travel, hotel accommodation, rental cars, theatre tickets, tickets for sporting events, …)

Data capture

- The coalition partners capture consumer transactions and transmit them to The Loyalty Group, which
- stores these transactions and uses the data for database marketing initiatives on behalf of the coalition partners.
- The Loyalty Group data warehouse currently contains
  - more than 6.3 million household records
  - 1 billion transaction records.

Before data mining

- The Loyalty Group has employed standard analytical techniques
  - Recency, Frequency, Monetary value (RFM) analysis
  - online analytic processing tools
  - linear statistical methods
- to analyze the success of the various marketing initiatives undertaken by the coalition and its partners.

Data mining project at AMRP

- Goal: create a customer segmentation using a data mining tool and compare the results to an existing segmentation developed using RFM analysis.
- data mining platform
  - DB2 Universal Database Enterprise parallelized over a five-node RS/6000 SP parallel system.
  - Intelligent Miner for Data (reason: has categorical clustering and product association algorithms which are not available in most other tools)

Data model

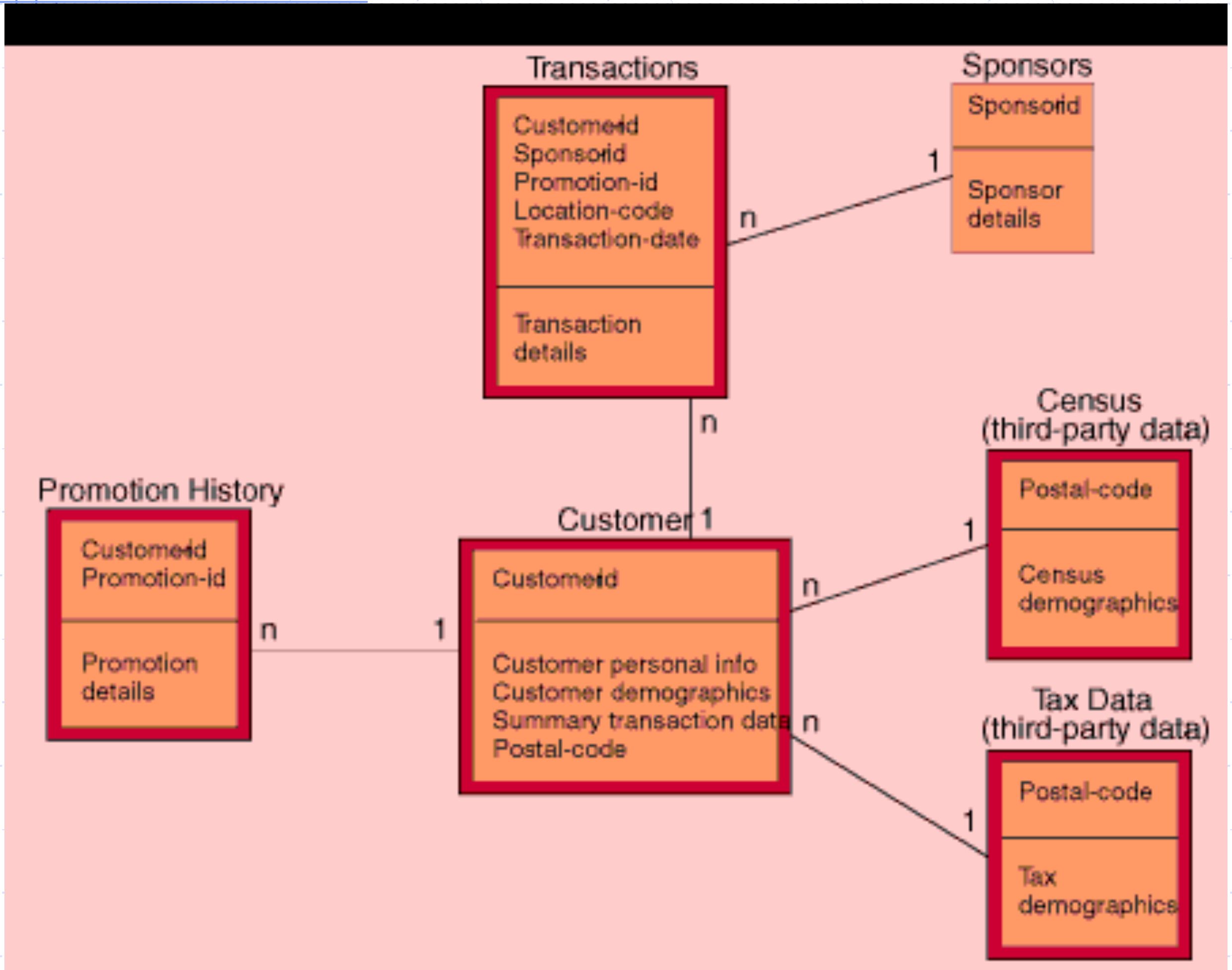


Figure 2. *AIR MILES case study data model.*

- ~ 50,000 customers and their associated transactions for a 12-month period.

Data preparation

- "shareholder value" indicators (variables)
  - revenue
  - customer tenure
  - number of sponsor companies shopped at over the customer tenure
  - number of sponsor companies shopped at over the last 12 months,
  - recency (in months) of the last transaction
- calculated by aggregating the transaction data and then adding them to each customer record

Data preparation (2)

- Dataset obtained by joining the transaction data to the customer file to create the input for clustering algorithms
- 84 variables =
  - 14 categories of sponsor companies ×
  - 3 variables per category ×
  - 2 quarters (first two quarters of 1997)

Data cleansing - missing values

- demographic data
  - is usually categorical
  - has a high % of missing values
  - the missing values can be set to either **unknown** or **unanswered** (if result of unanswered questions)
- if a large portion of the field is missing, it may be discarded.
- In the case study, missing numeric values set to 0

Data transformation

- Ratio variables.
  - E.g.: profitability = profit / tenure
- Time-derivative variables.
  - E.g.: profit 2nd quarter - profit 1st quarter
- Discretization using quantiles.
  - E.g., break points at 10, 25, 50, 75, and 90.
- Discretization using predefined ranges.
  - E.g., those used in census
- Log transforms.
  - E.g., for very skewed distributions

Distribution of original data

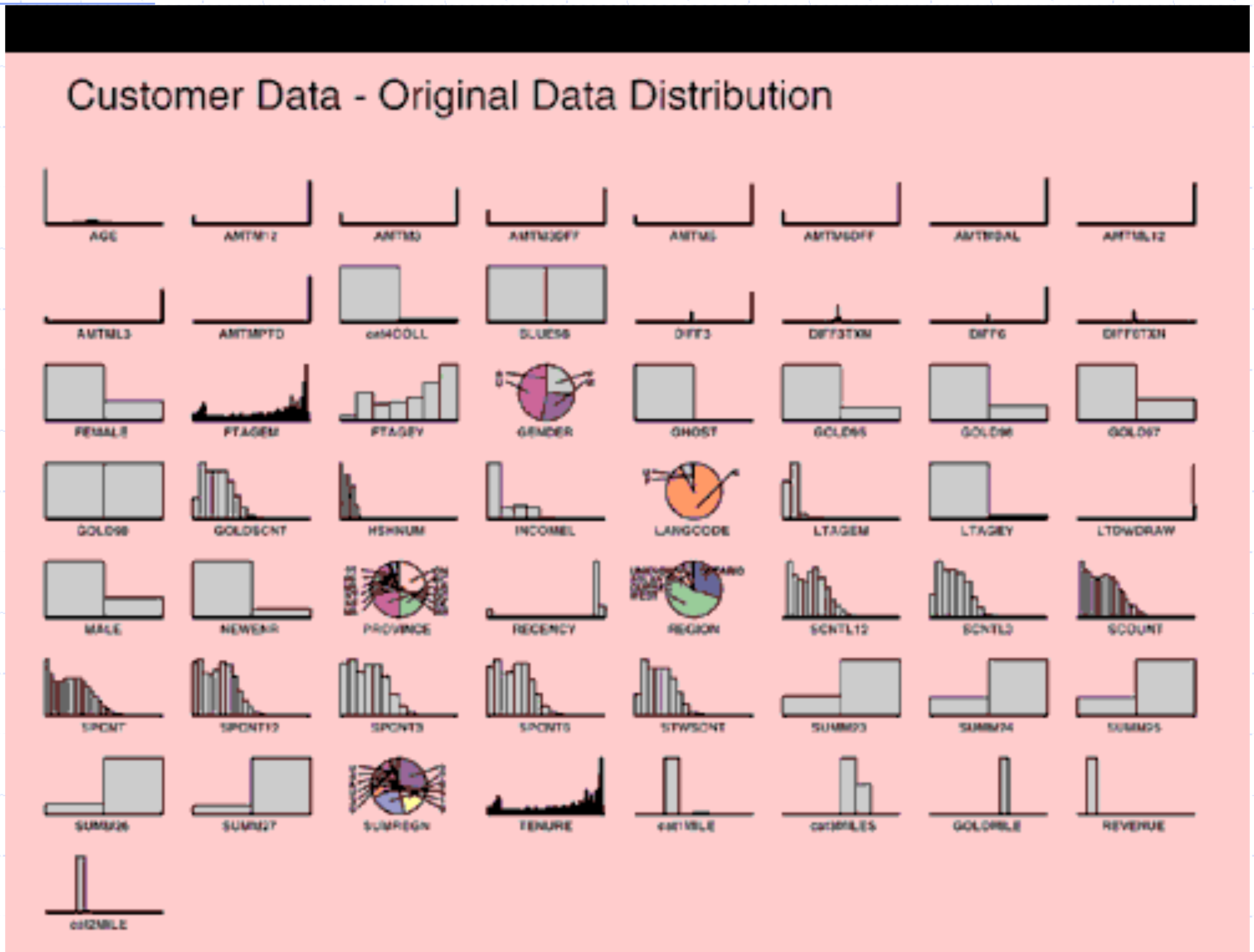


Figure3. *Original data.*

Distribution of discretized data

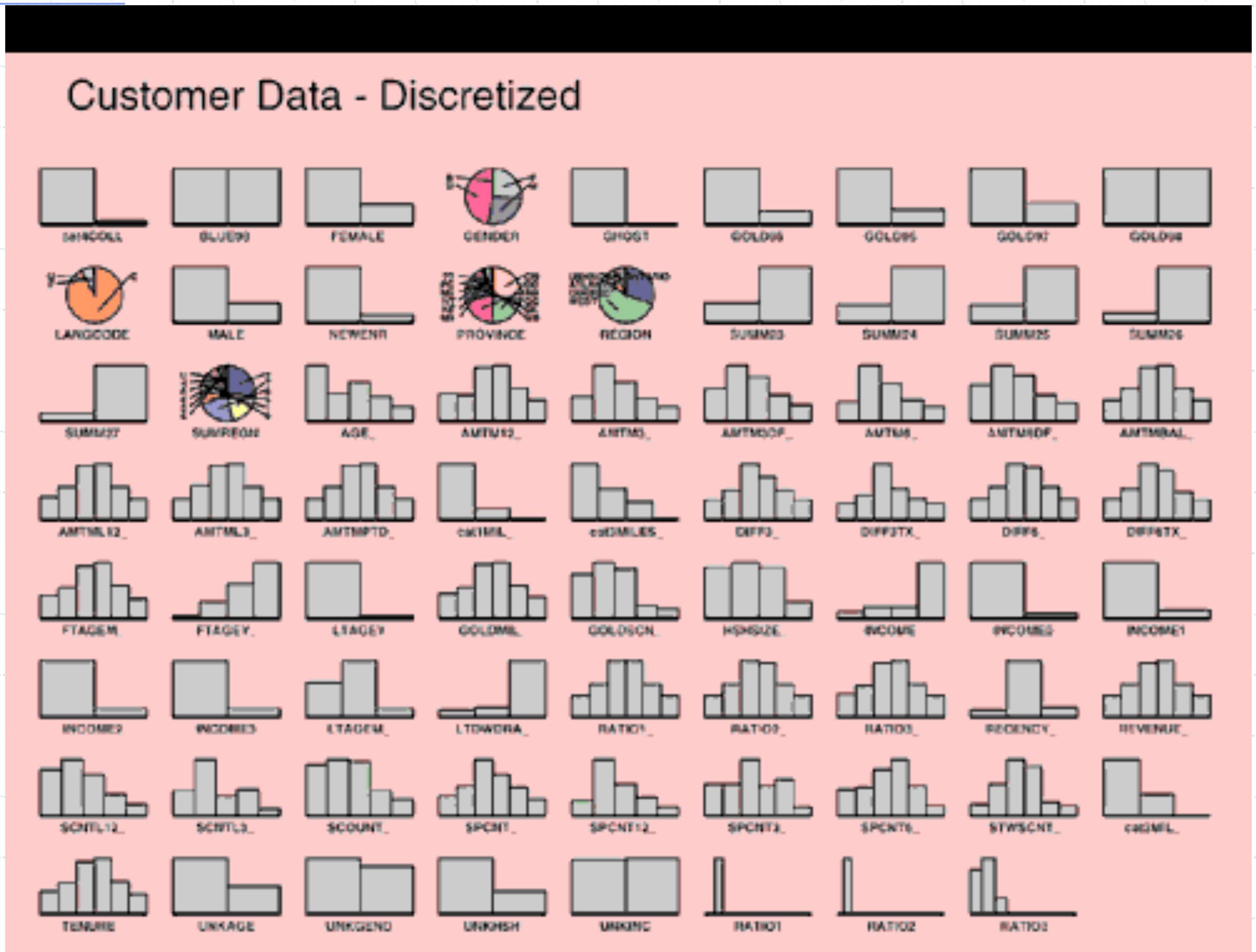


Figure 4. *Discretized data.*

Before/after discretization

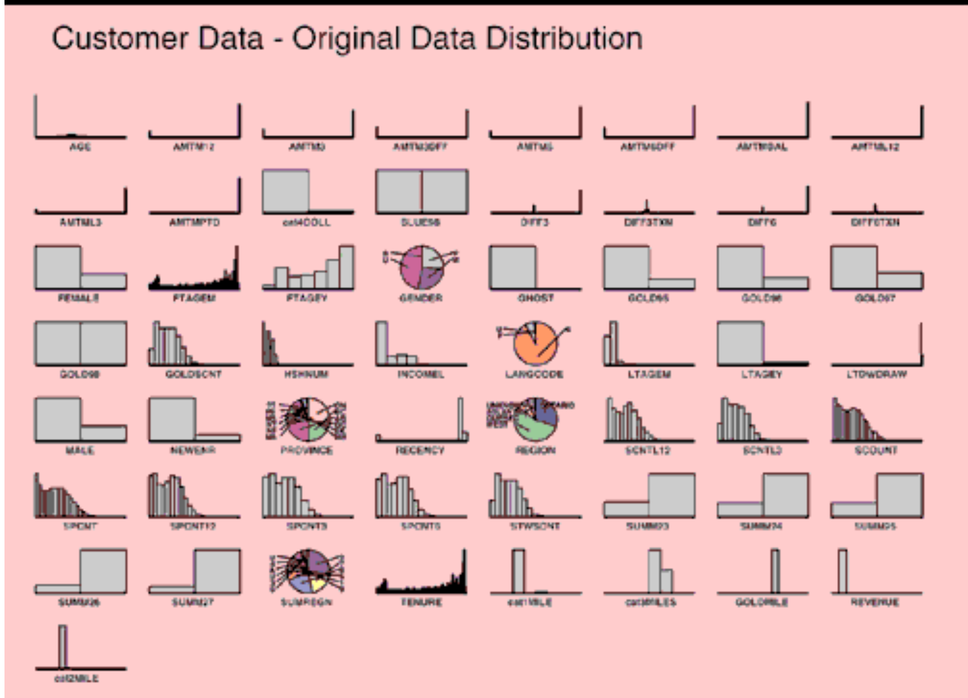


Figure3. *Original data.*

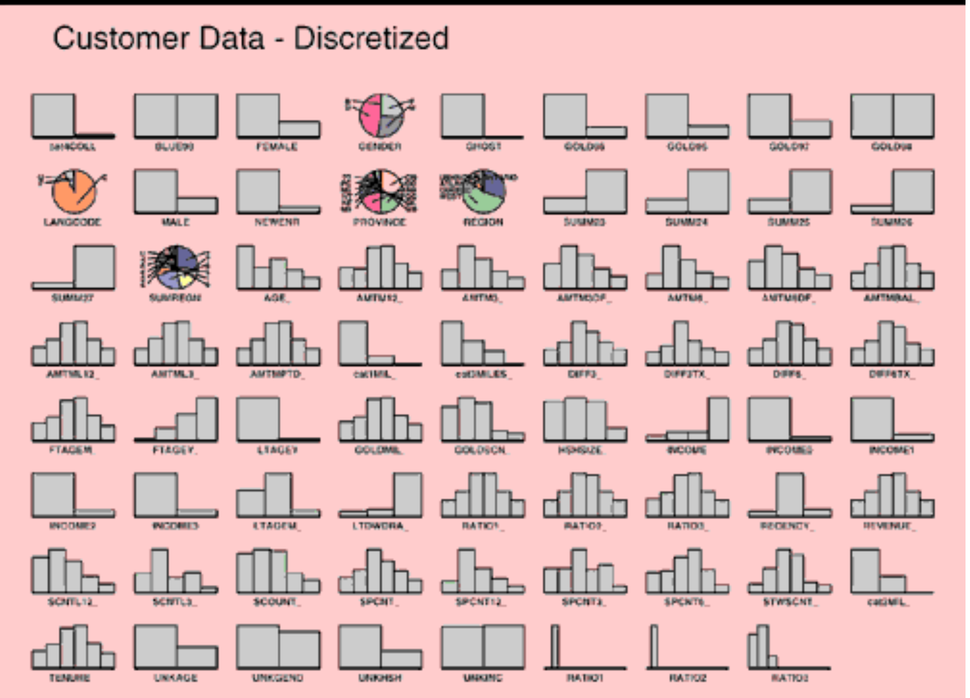


Figure 4. *Discretized data.*

Clustering/segmentation methodology

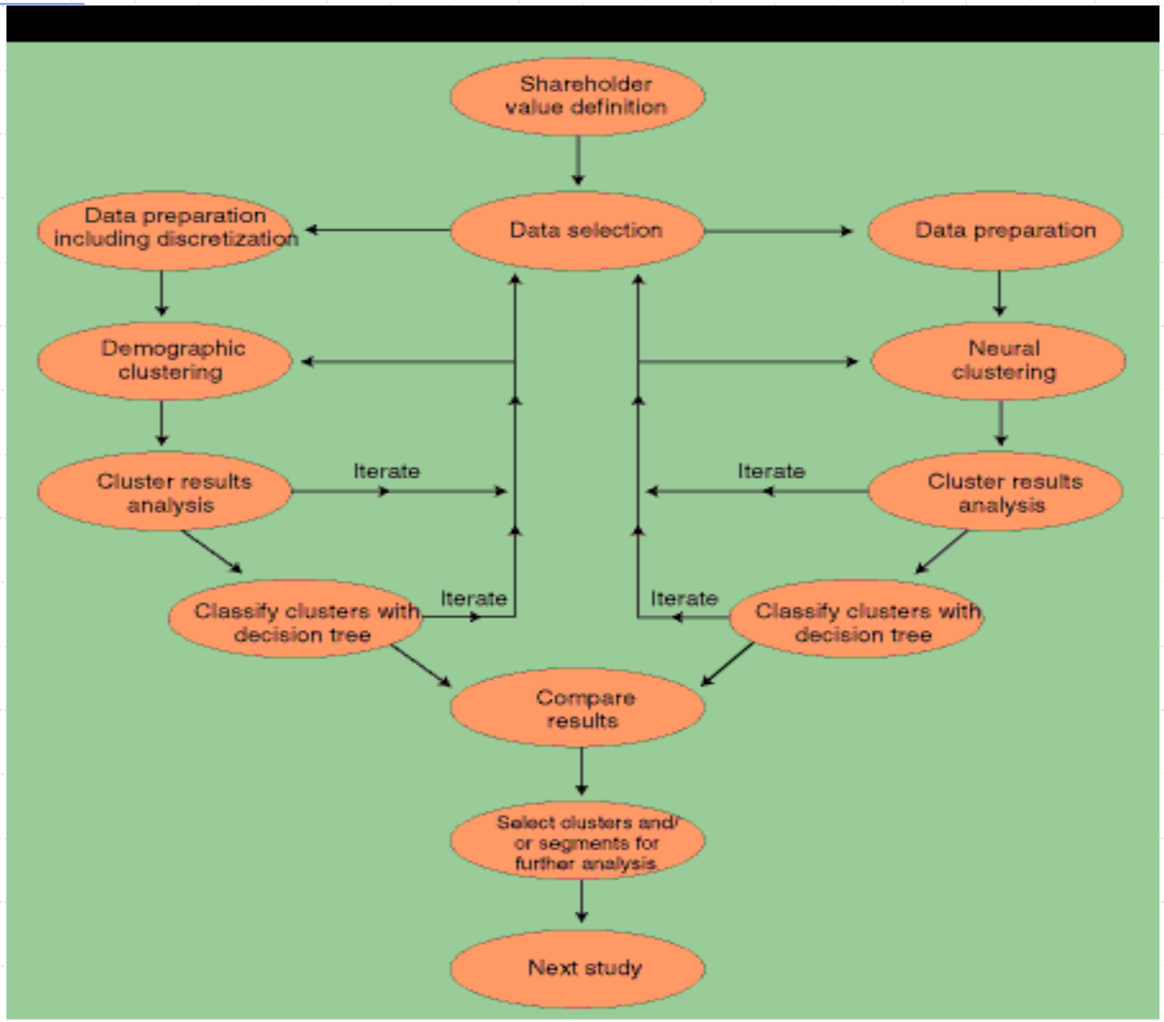


Figure 6. *Clustering workflow.*

IBM-IM demographic clustering

- Designed for categorical variables
- Similarity index:
  - increases with number of common values on same attribute
  - decreases with number of different values on same attribute
- # of clusters is not fixed a priori
  - only upper bound set

Input dataset

- dataset: all continuous variables discretized.
- input variables :
  - # of products purchased over customer's lifetime
  - # of products purchased in the last 12 months
  - Customer's revenue contribution over lifetime
  - Customer tenure in months
  - Ratio of revenue to tenure
  - Ratio of number of products to tenure
  - Region
  - Recency
  - Tenure (# of months since customer first enrolled in the program).

Input dataset

- Other discrete and categorical variables and some interesting continuous variables were input as supplementary variables:
- variables used to profile the clusters but not to define them.
- easier interpretation of clusters using data other than the input variables.

Output of demographic clustering

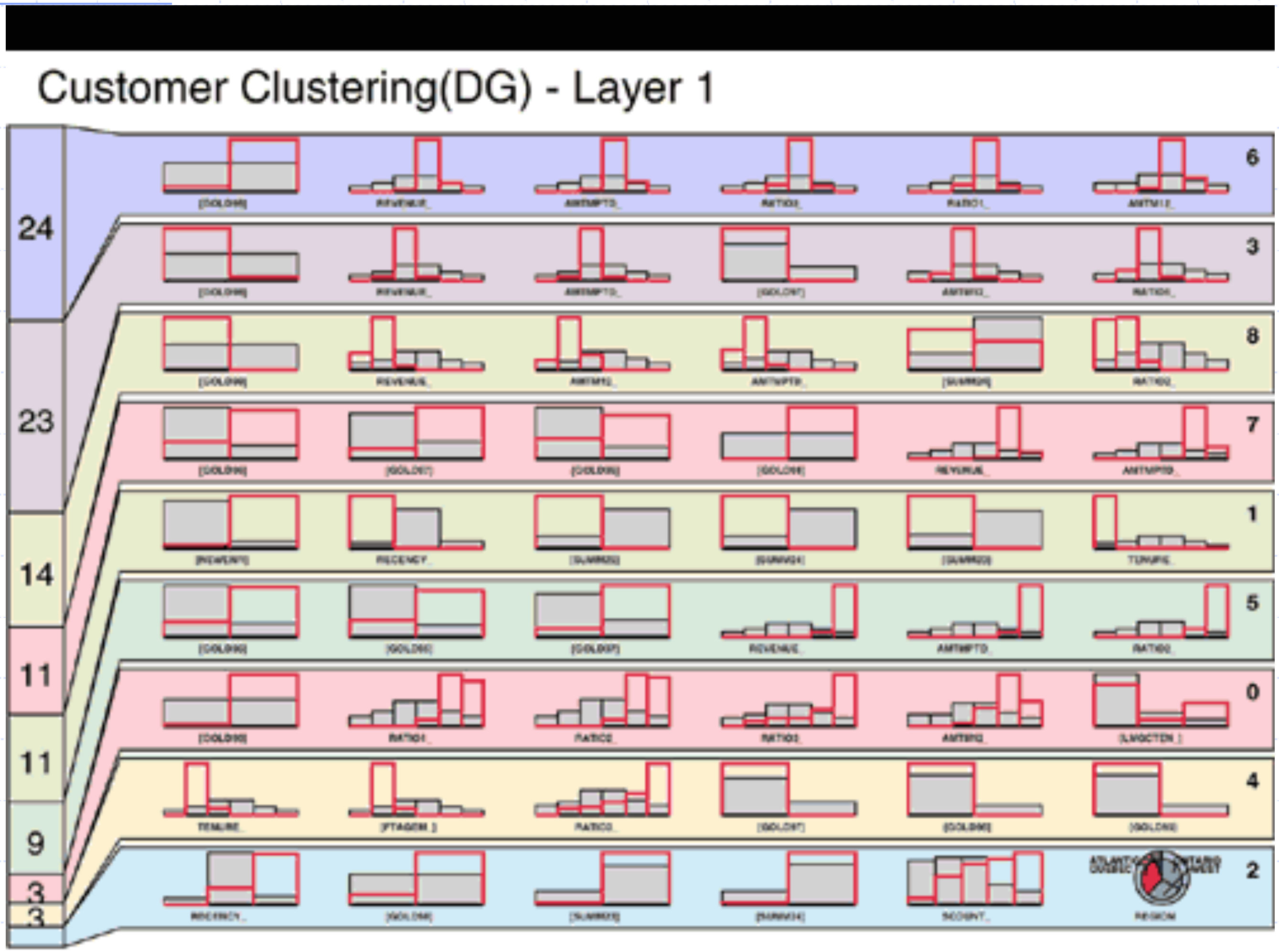


Figure 7. *Demographic clustering output.*

Visualization of clusters

- horizontal strip = a cluster
- clusters are ordered from top to bottom in order of size
- variables are ordered from left to right in order of importance to the cluster, based on a chi-square test between variable and cluster ID.
- other metrics include entropy, Condorcet criterion, and database order.

Visualization of clusters

- variables used to define clusters are without brackets, while the supplementary variables appear within brackets.
- numeric (integer), discrete numeric (small integer), binary, and continuous variables have their frequency distribution shown as a bar graph.

- red bars = distribution of the variable within the current cluster.
- gray solid bars = distribution of the variable in the whole universe.

Visualization of clusters

- Categorical variables are shown as pie charts.
- inner pie = distribution of the categories for the current cluster
- outer ring = distribution of the variable for the entire universe.

- The more different the cluster distribution is from the average, the more interesting or distinct the cluster.

Output of demographic clustering

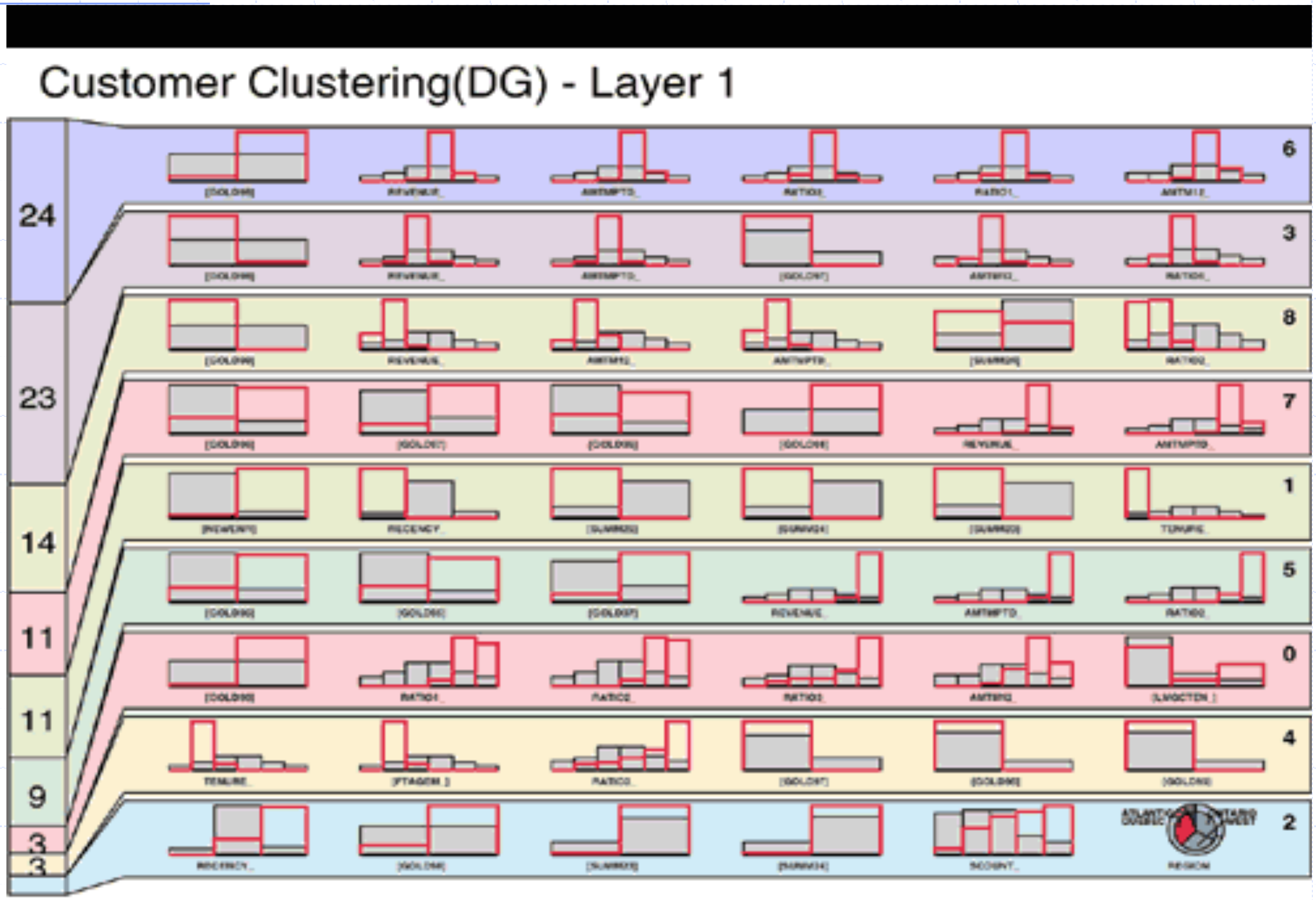


Figure 7. *Demographic clustering output.*

Qualitative characterization of clusters

- Gold98 is a binary variable that indicates the best customers in the database, created previously by the business using RFM analysis.
- The clustering model agrees very well with this existing definition: Most of the clusters seem to have almost all Gold or no Gold customers.
- Confirmed the current Gold segment!

Qualitative characterization of clusters

- clustering results
  - not only validate the existing concept of Gold customers,
  - they extend the idea of the Gold customers by creating clusters within the Gold98 customer category.
  - A platinum customer group
- Cluster 5
  - Almost all Gold98 customers,
  - revenue, bonus collected lifetime to date, revenue per month, and lifetime to date per month are all in the 50th to 75th percentile.

Qualitative characterization of clusters

- Cluster 3:
  - no Gold98 customers. Its customer revenue, bonus collected, revenue per month, are all in the 25th to 50th percentile.
- Cluster 5:
  - 9 %of the population.
  - revenue, bonus collected are all in the 75th percentile and above, skewed to almost all greater than the 90th percentile.
  - looks like a very profitable cluster

Detailed view of cluster 5

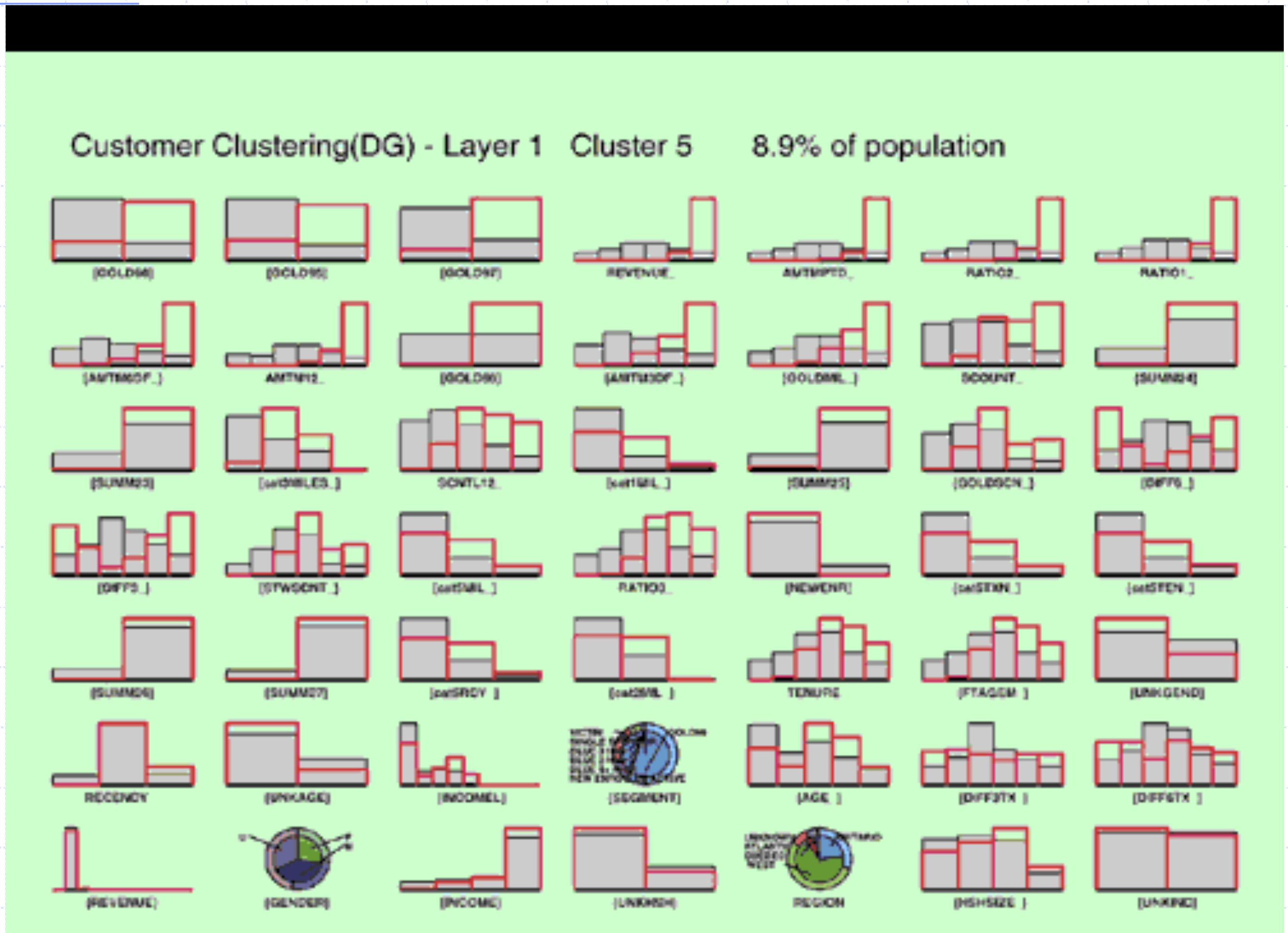


Figure 8. *Cluster 5 output.*

Profiling clusters

- Goal: assess the potential business value of each cluster quantitatively by profiling the aggregate values of the shareholder value variables by cluster.

CLUSTERID	REVENUE	CUSTOMERS	PRODUCT INDEX	LEVERAGE	TENURE
5	34.74%	8.82%	1.77	3.94	60.92
6	26.13%	23.47%	1.41	1.11	57.87
7	21.25%	10.71%	1.64	1.98	63.52
3	6.62%	23.32%	.73	.28	47.23
0	4.78%	3.43%	1.45	1.40	31.34
2	4.40%	2.51%	1.46	1.75	61.38
4	1.41%	2.96%	.99	.48	20.10
8	.45%	14.14%	.36	.03	30.01
1	.22%	10.64%	.00	.02	4.66

Table 1. *Profiling a cluster.*

Profiling clusters

- leverage = ratio of revenue to customer.
- cluster 5 is the most profitable cluster.
- as profitability increases, so does the average number of products purchased.
- product index = ratio of the average number of products purchased by the customers in the cluster divided by the average number of products purchased overall.
- customer profitability increases as tenure increases.

Business opportunities

- Best customers in clusters 2, 5, and 7. :
  - indication: retention
- clusters 2, 6, and 0
  - indication: cross-selling by contrasting with clusters 5 and 7.
  - Clusters 2, 6, and 0 have a product index close to those of clusters 5 and 7, which have the highest number of products purchased.
  - Try to convert customers from clusters 2, 6, and 0 to clusters 5 and 7. By comparing which products are bought we can find products that are candidates for cross-selling.

Business opportunities

- Clusters 3 and 4
  - indication: cross-selling to clusters 2, 6, and 0 •
- Cluster 1
  - indication: wait and see. It appears to be a group of new customers
- Cluster 8
  - indication: no waste of marketing dollars

Follow-up

- Reactions from The Loyalty Group
  - visualization of results allowed for meaningful and actionable analysis.
  - original segmentation methodology validated, but that refinements to the original segmentation could prove valuable.
  - decision to undertake further data mining projects, including
    - predictive models for direct mail targeting,
    - further work on segmentation using more detailed behavioral data,
    - opportunity identification using association algorithms within the segments discovered.

Outline this lesson

- KDD Process
- CRM and Data Mining - AirMiles
- **Health and Data Mining – Stulong**
- CRISP model for developing Data Mining based services

Atherosclerosis prevention study

**2nd Department of Medicine,
1st Faculty of Medicine of Charles University and Charles University Hospital,
U nemocnice 2, Prague 2
(head. Prof. M. Aschermann, MD, SDr, FESC)**

Atherosclerosis prevention study:

- The STULONG 1 data set is a real database that keeps information about the study of the development of atherosclerosis risk factors in a population of middle aged men.
- Used for Discovery Challenge at PKDD 00-02-03-04

Atherosclerosis prevention study:

- Study on 1400 middle-aged men at Czech hospitals
  - Measurements concern development of cardiovascular disease and other health data in a series of exams
- The aim of this analysis is to look for associations between medical characteristics of patients and death causes.
- Four tables
  - Entry and subsequent exams, questionnaire responses, deaths

The input data

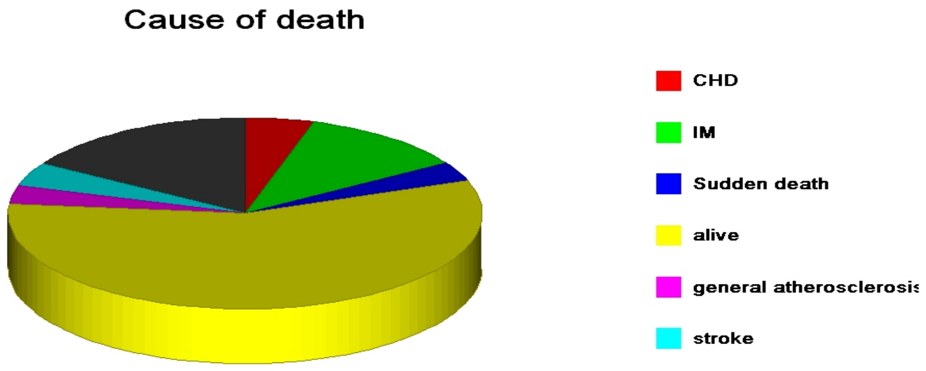
Data from Entry and Exams		
General characteristics	**Examinations**	**habits**
Marital status	**Chest pain**	**Alcohol**
Transport to a job	**Breathlesness**	**Liquors**
Physical activity in a job	**Cholesterol**	**Beer 10**
Activity after a job	**Urine**	**Beer 12**
Education	**Subscapular**	**Wine**
Responsibility	**Triceps**	**Smoking**
Age		**Former smoker**
Weight		**Duration of smoking**
Height		**Tea**
		Sugar
		Coffee

The input data

DEATH CAUSE	PATIENTS	%
myocardial infarction	80	20.6
coronary heart disease	33	8.5
stroke	30	7.7
other causes	79	20.3
sudden death	23	5.9
unknown	8	2.0
tumorous disease	114	29.3
general atherosclerosis	22	5.7
TOTAL	389	100.0

Data selection

- When joining “Entry” and “Death” tables we implicitely create a new attribute “Cause of death”, which is set to “alive” for subjects present in the “Entry” table but not in the “Death” table.
- We have only 389 subjects in death table.



The prepared data

Patient	General characteristics		Examinations		Habits		Cause of death
	Activity after work	**Education**	**Chest pain**	…	**Alcohol**	…..	
1	moderate activity	university	not present		no		Stroke
2	great activity		not ischaemic		occasionally		myocardial infarction
3	he mainly sits		other pains		regularly		tumorous disease
……	……..	……..	………..	..	…	……	alive
389	he mainly sits		other pains		regularly		tumorous disease

Descriptive Analysis/ Subgroup Discovery / Association Rules

Are there strong relations concerning death cause?

General characteristics (?) $\Longrightarrow$ Death cause (?)

Examinations (?) $\Longrightarrow$ Death cause (?)

Habits (?) $\Longrightarrow$ Death cause (?)

Combinations (?) $\Longrightarrow$ Death cause (?)

Example of extracted rules

- Education(university) & Height<176-180> $\Rightarrow$ Death cause (tumouros disease), *16 ; 0.62*
- It means that on tumorous disease have died 16, i.e. 62% of patients with university education and with height 176-180 cm.

Example of extracted rules

- Physical activity in work(he mainly sits) & Height<176-180> $\Rightarrow$ Death cause (tumouros disease), *24; 0.52*
- It means that on tumorous disease have died 24 i.e. 52% of patients that mainly sit in the work and whose height is 176-180 cm.

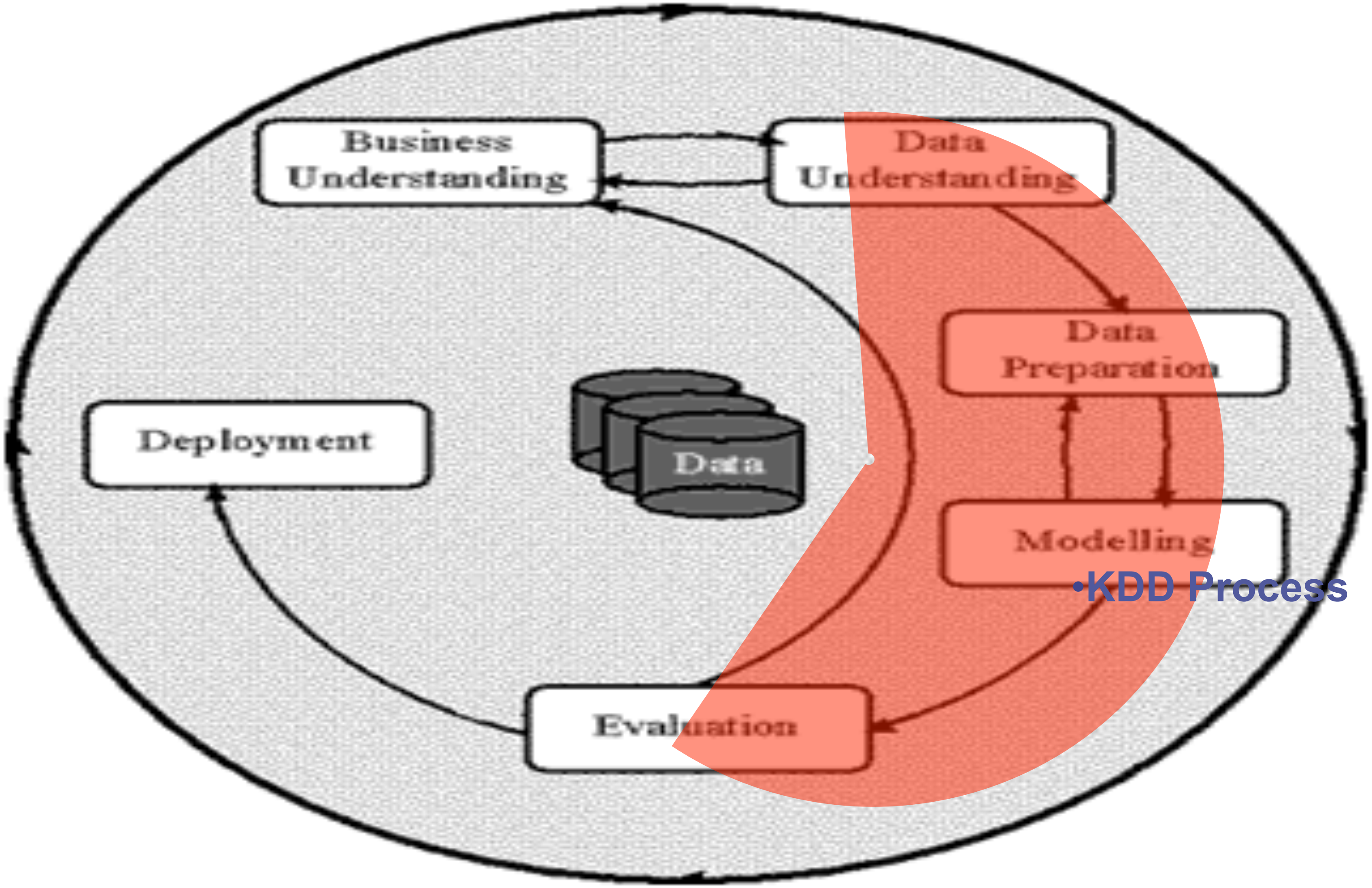
Example of extracted rules

- Education(university) & Height<176-180> ⇒Death cause (tumouros disease),
  16; 0.62; +1.1;
- the relative frequency of patients who died on tumorous disease among patients with university education and with height 176-180 cm is 110 per cent higher than the relative frequency of patients who died on tumorous disease among all the 389 observed patients

Outline this lesson

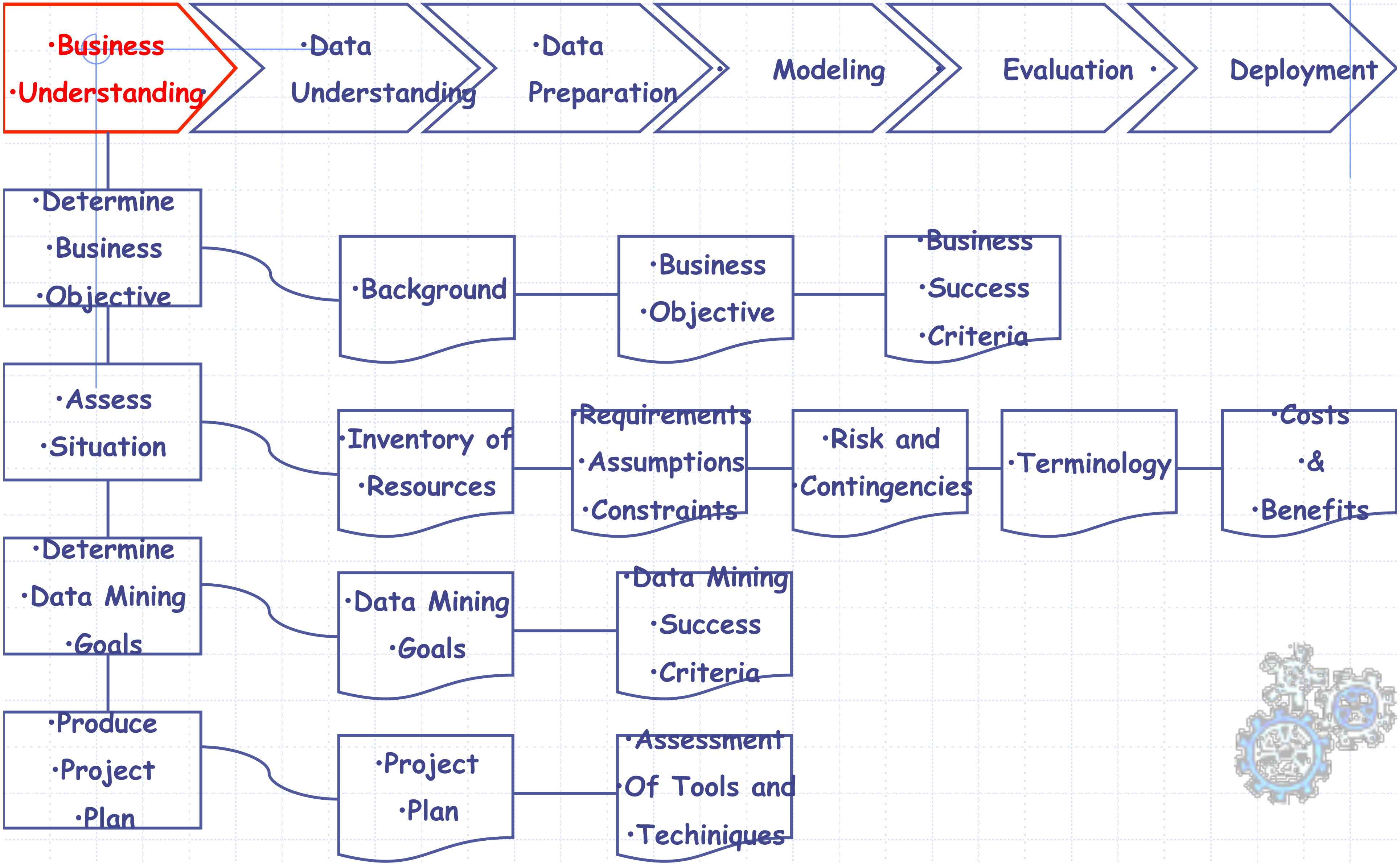
- KDD Process
- CRM and Data Mining - AirMiles
- Health and Data Mining – Stulong
- **CRISP model for developing Data Mining based services**

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



Business understanding

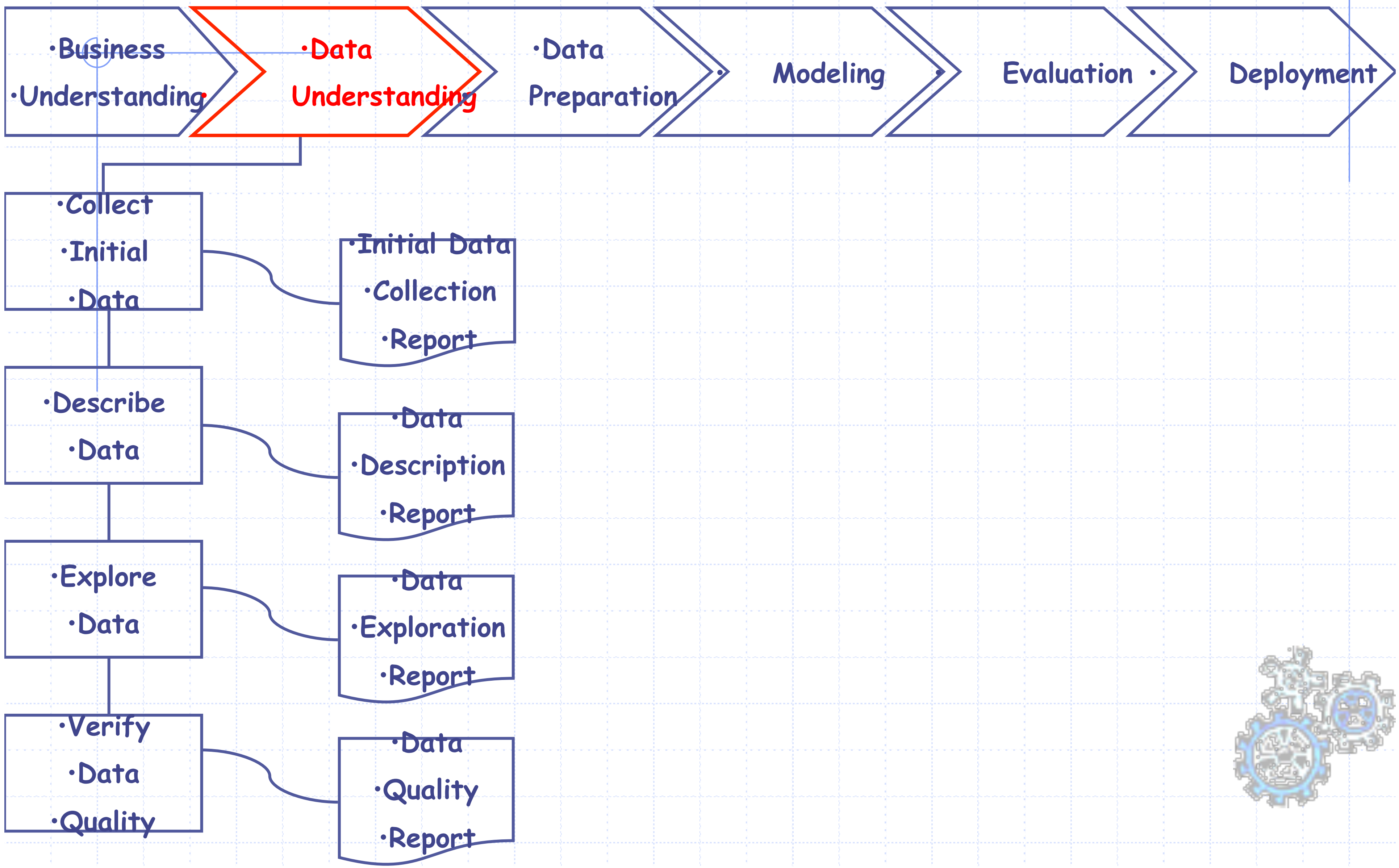
- Understanding the project objectives and requirements from a business perspective.
- then converting this knowledge into a data mining problem definition and a preliminary plan.
  - **Determine the Business Objectives**
  - **Determine Data requirements for Business Objectives**
  - **Translate Business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



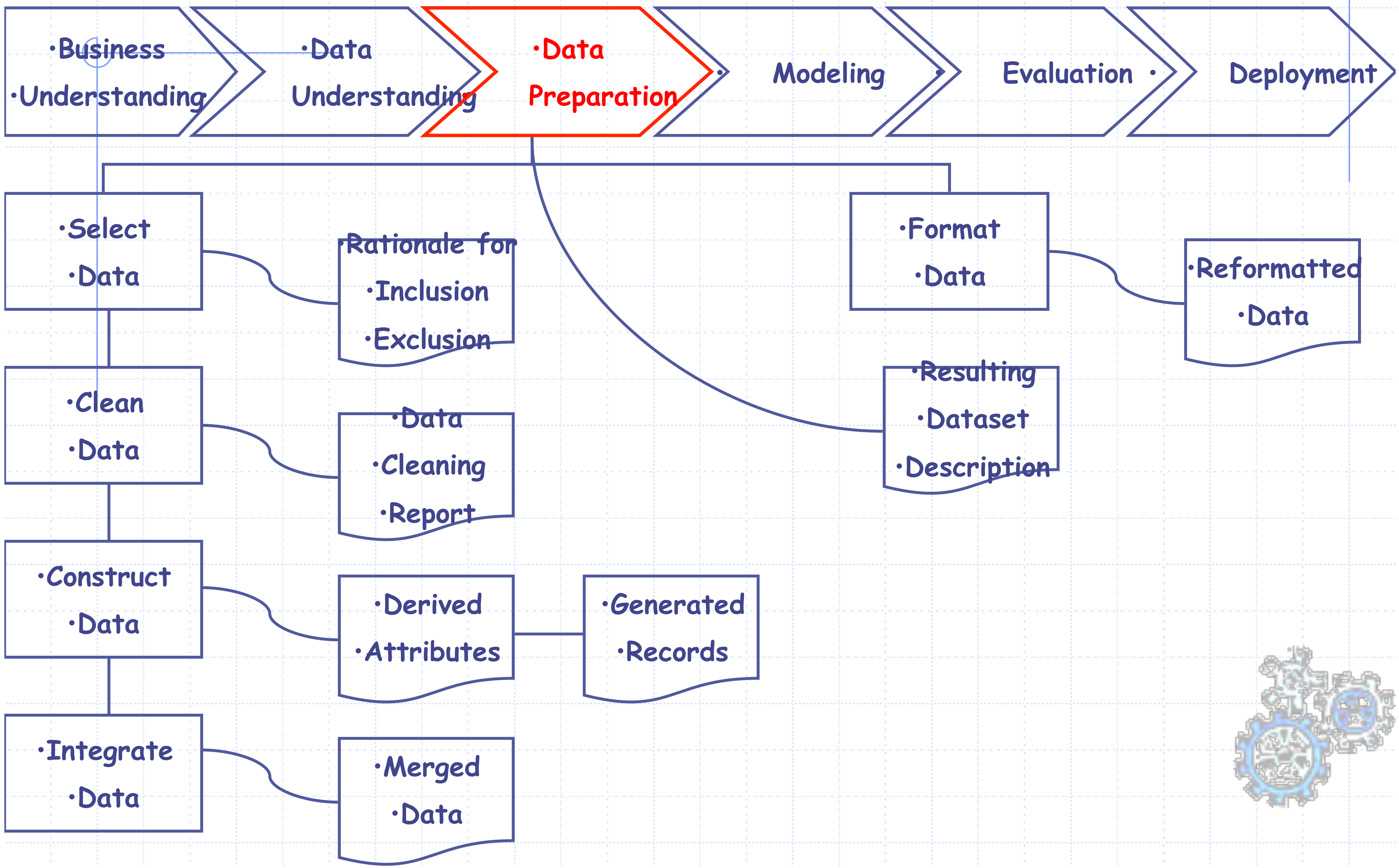
•Business
•Understanding
•Data
Understanding
•Data
Preparation
Modeling
Evaluation
Deployment
•Determine
•Business
•Objective
•Background
•Business
•Objective
•Business
•Success
•Criteria
•Assess
•Situation
•Inventory of
•Resources
•Requirements
•Assumptions
•Constraints
•Risk and
•Contingencies
•Terminology
•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.



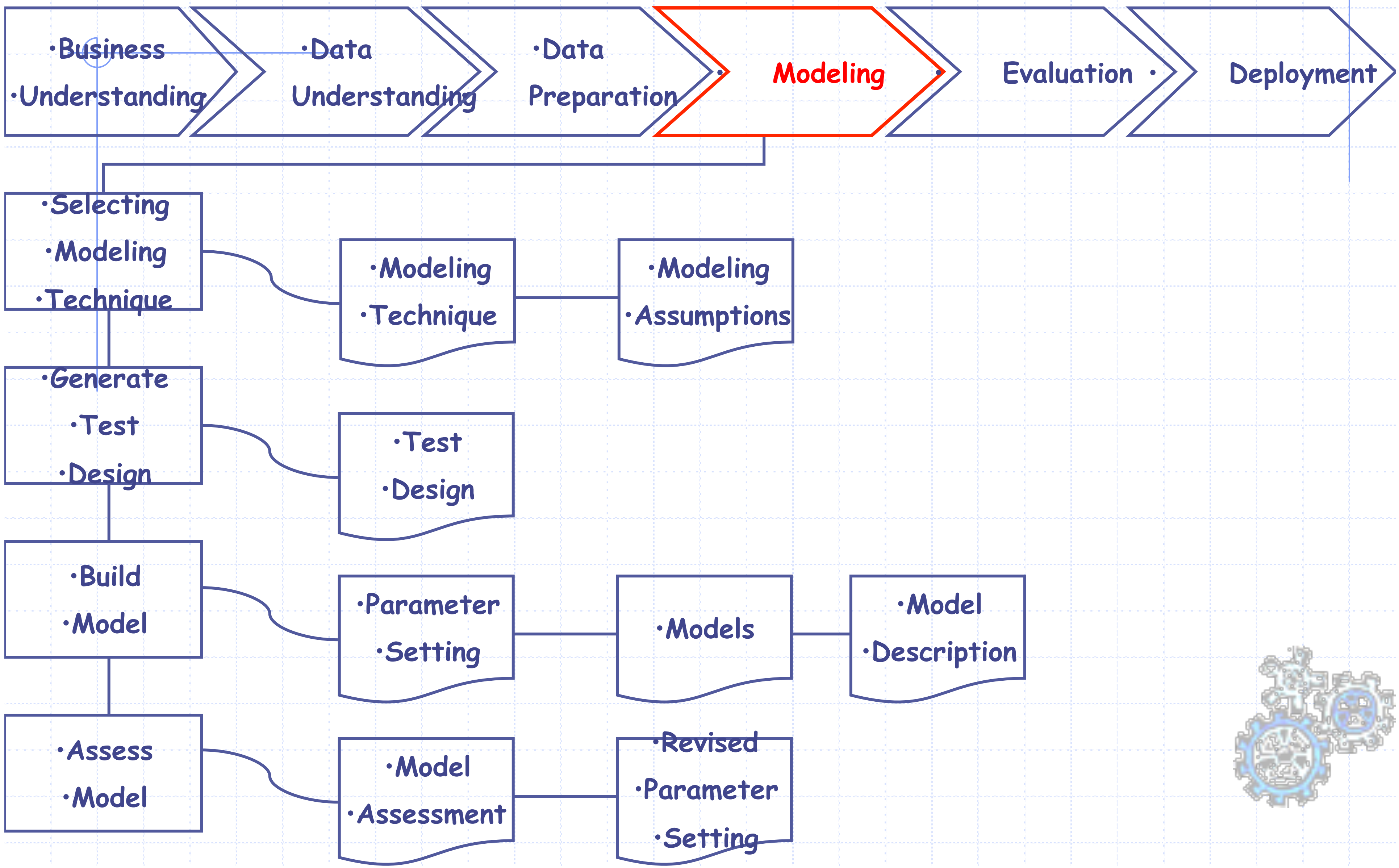
Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Collect Initial Data
Initial Data Collection Report
Describe Data
Data Description Report
Explore Data
Data Exploration Report
Verify Data Quality
Data Quality Report



•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:

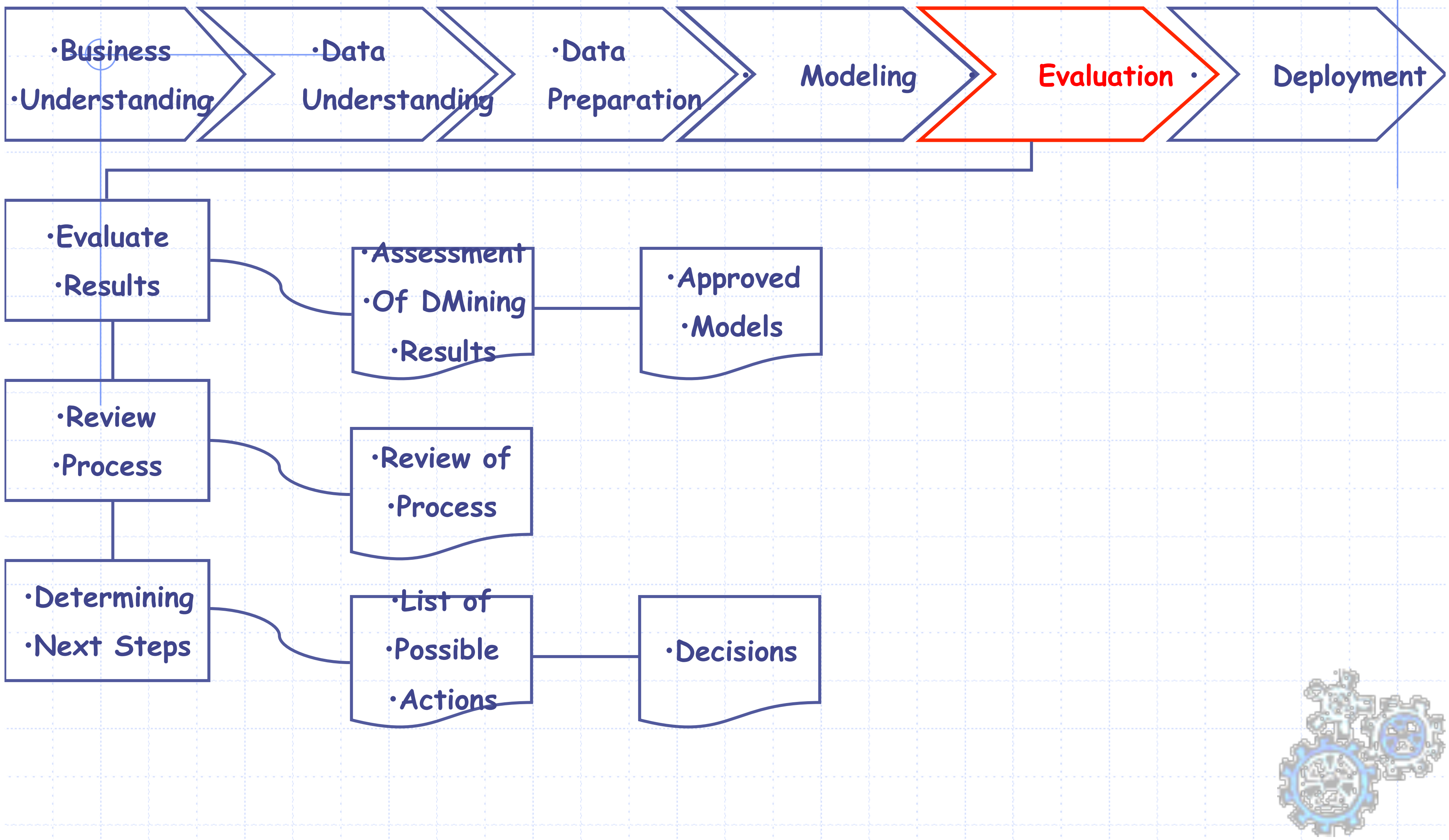
- In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.
- Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data.
- Therefore, stepping back to the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

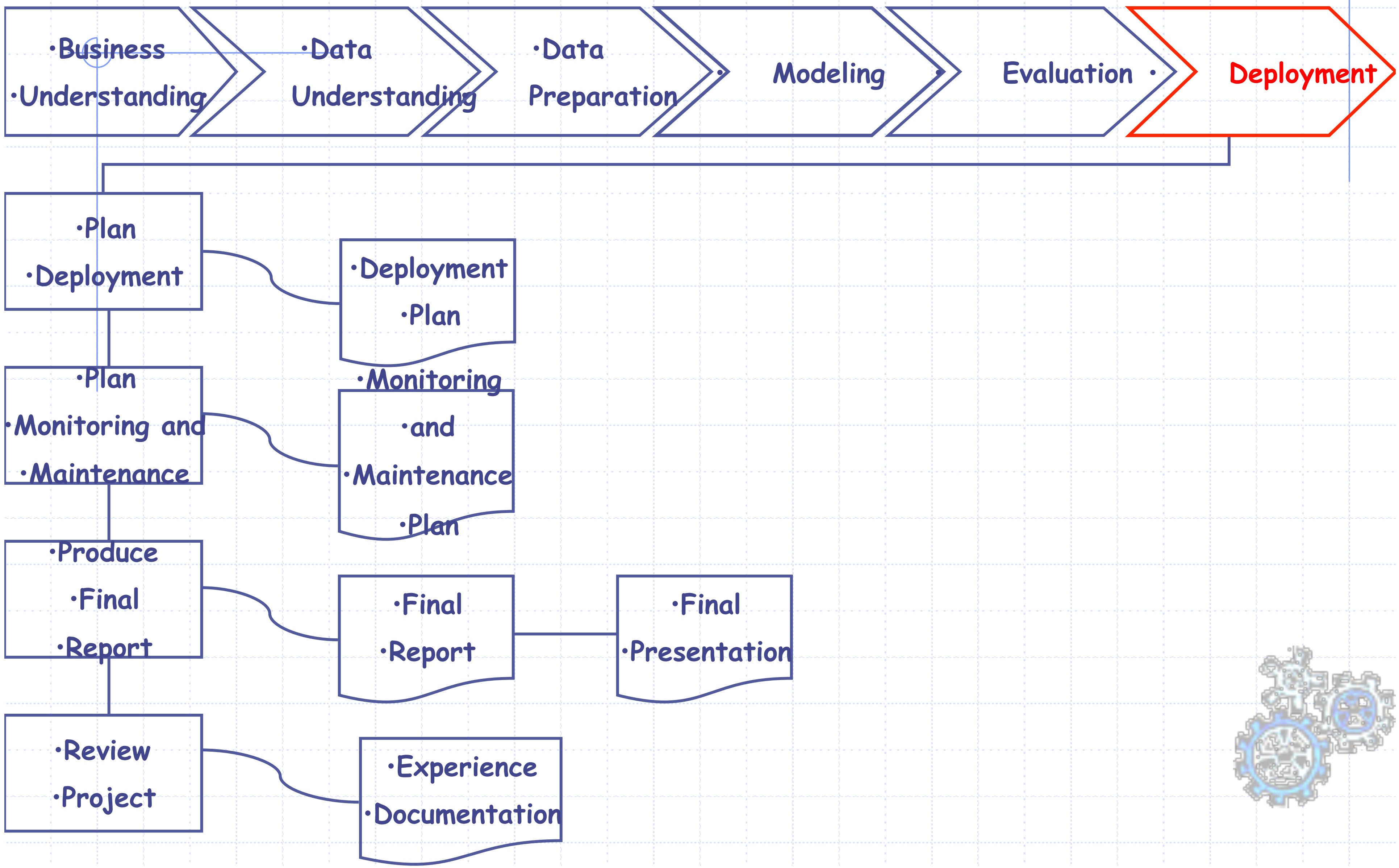
Deployment:

- 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

Pre-analisi	*Progettazione e pianificazione*	*Estrazione delle liste dei clienti*	*Erogazione attraverso i canali di contatto*	*Analisi dei risultati*
Segmentazione Profilatura	Definzione degli elementi della campagna ➡ costi ➡ offerta ➡ messaggi Segmentazione Disegno dell'albero della campagna Altre funzioni	Generazione della lista dei clienti target Pulizia della lista Rilascio della lista e della cadenza dei contatti ai canali	Outbound Inbound SMS E-Mail Web Direct Mail	Acquisizione dei dati sui contatti e sulle accettazioni Valutazione della efficacia e della effciineza Analisi delle efficacia per canale (clickstream) Altre analisi

•Mining Based Decision Support System: Adaptive Architecture