Chapter 13

Knowledge Discovery Systems: Systems That Create Knowledge

Becerra-Fernandez, et al. -- Knowledge
Management 1/e -- © 2004 Prentice Hall

Chapter Objectives

- To explain how knowledge is discovered
- To describe knowledge discovery systems, including design considerations, and how they rely on mechanisms and technologies
- To explain data mining (DM) technologies
- To discuss the role of DM in customer relationship management

Knowledge Synthesis through Socialization

- To discover tacit knowledge
- Socialization enables the discovery of tacit knowledge through joint activities
  - between masters and apprentices
  - between researchers at an academic conference

Knowledge Discovery from Data – Data Mining

- Another name for Knowledge Discovery in Databases is data mining (DM).
- Data mining systems have made a significant contribution in scientific fields for years.
- The recent proliferation of e-commerce applications, providing reams of hard data ready for analysis, presents us with an excellent opportunity to make profitable use of data mining.

Data Mining Techniques Applications

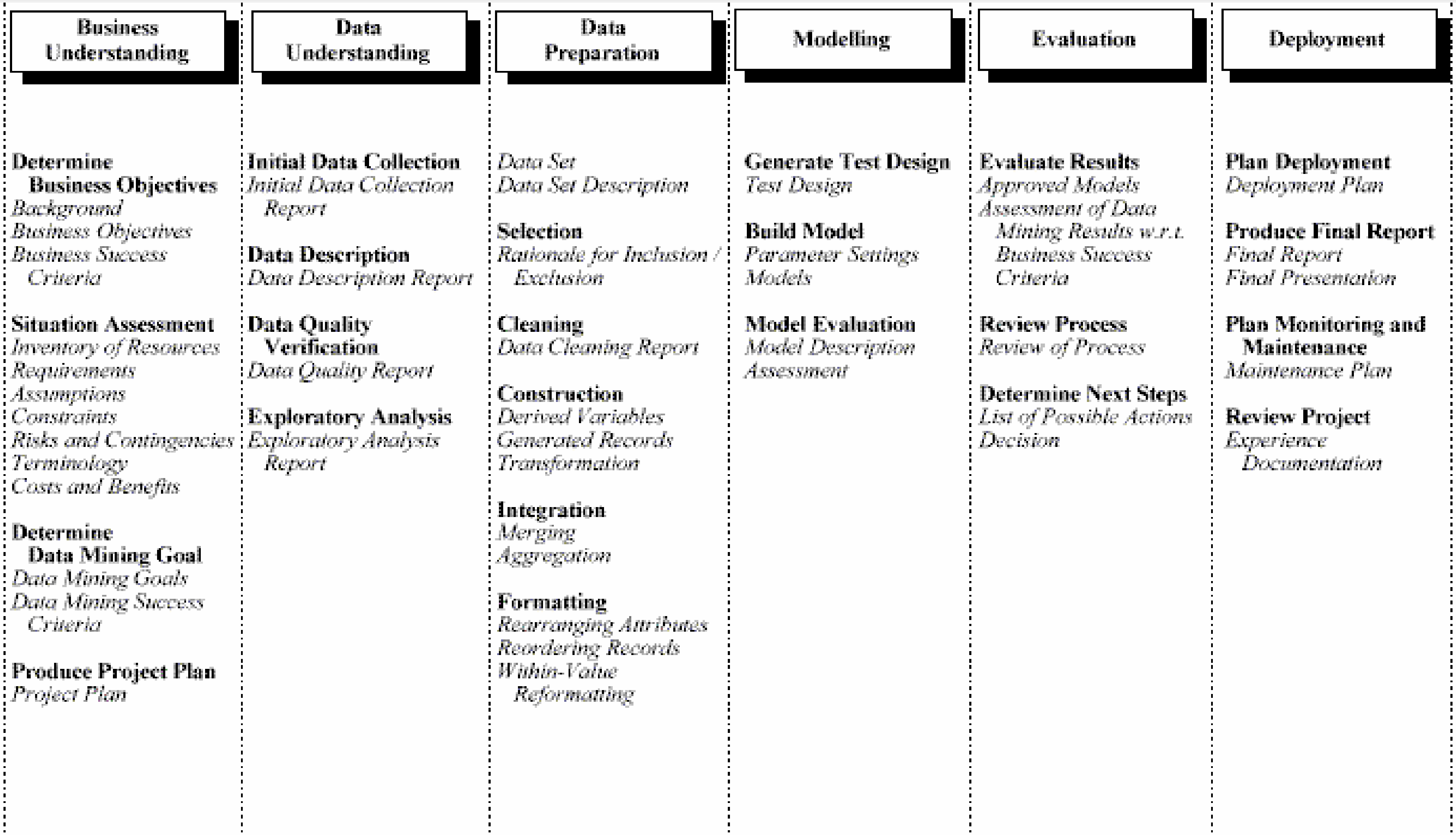
- **Marketing** – Predictive DM techniques, like artificial neural networks (ANN), have been used for ***target marketing*** including market segmentation.
- **Direct marketing** – customers are likely to respond to new products based on their previous consumer behavior.
- **Retail** – DM methods have likewise been used for ***sales forecasting****.*
- **Market basket analysis** – uncover which products are likely to be purchased together.

Data Mining Techniques Applications

- **Banking** – ***Trading*** and ***financial forecasting*** are used to determine derivative securities pricing, futures price forecasting, and stock performance.
- **Insurance** – DM techniques have been used for segmenting customer groups to determine ***premium pricing*** and predict ***claim*** frequencies.
- **Telecommunications** – Predictive DM techniques have been used to attempt to reduce churn, and to predict when customers will attrition to a competitor.
- **Operations management** – Neural network techniques have been used for planning and scheduling, project management, and quality control.

Designing the Knowledge Discovery System – CRISP DM

1. **Business Understanding** – To obtain the highest benefit from data mining, there must be a clear statement of the business objectives.
2. **Data Understanding** – Knowing the data well can permit the designer to tailor the algorithm or tools used for data mining to his/her specific problem.
3. **Data Preparation –** Data selection, variable construction and transformation, integration, and formatting
4. **Model building and validation** – Building an accurate model is a trial and error process. The process often requires the data mining specialist to iteratively try several options, until the best model emerges.
5. **Evaluation and interpretation** – Once the model is determined, the validation dataset is fed through the model.
6. **Deployment** – Involves implementing the ‘live’ model within an organization to aid the decision making process.



Business Understanding	Data Understanding	Data Preparation	Modelling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Situation Assessment** *Inventory of Resources* *Requirements* *Assumptions* *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 Data Collection** *Initial Data Collection Report* **Data Description** *Data Description Report* **Data Quality Verification** *Data Quality Report* **Exploratory Analysis** *Exploratory Analysis Report*	*Data Set* *Data Set Description* **Selection** *Rationale for Inclusion / Exclusion* **Cleaning** *Data Cleaning Report* **Construction** *Derived Variables* *Generated Records* *Transformation* **Integration** *Merging* *Aggregation* **Formatting** *Rearranging Attributes* *Reordering Records* *Within-Value Reformatting*	**Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* **Model Evaluation** *Model Description* *Assessment*	**Evaluate Results** *Approved Models* *Assessment of Data Mining Results w.r.t. Business Success Criteria* **Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Deployment** *Deployment Plan* **Produce Final Report** *Final Report* *Final Presentation* **Plan Monitoring and Maintenance** *Maintenance Plan* **Review Project** *Experience Documentation*

CRISP-DM Data Mining Process Methodology

1. Business Understanding process

a. **Determine Business objectives** – To obtain the highest benefit from data mining, there must be a clear statement of the business objectives .

b. **Situation Assessment** – The majority of the people in a marketing campaign who receive a target mail, do not purchase the product .

c. **Determine Data Mining Goal** – Identifying the most likely prospective buyers from the sample, and targeting the direct mail to those customers, could save the organization significant costs.

d. **Produce Project Plan** – This step also includes the specification of a project plan for the DM study .

2. Data Understanding process

a. **Data collection** – Defines the data sources for the study, including the use of external public data, and proprietary databases.

b. **Data description** – Describes the contents of each file or table. Some of the important items in this report are: number of fields (columns) and percent of records missing.

c. **Data quality and verification** – Define if any data can be eliminated because of irrelevance or lack of quality.

d. **Exploratory Analysis of the Data** – Use to develop a hypothesis of the problem to be studied, and to identify the fields that are likely to be the best predictors.

3. Data Preparation process

a. **Selection** – Requires the selection of the predictor variables and the sample set.

b. **Construction and transformation of variables** – Often, new variables must be constructed to build effective models.

c. **Data integration** – The dataset for the data mining study may reside on multiple databases, which would need to be consolidated into one database.

d. **Formatting** – Involves the reordering and reformatting of the data fields, as required by the DM model.

4. Model building and Validation process

a. **Generate Test Design** – Building an accurate model is a trial and error process. The data mining specialist iteratively try several options, until the best model emerges.

b. **Build Model** – Different algorithms could be tried with the same dataset. Results are compared to see which model yields the best results.

c. **Model Evaluation** – In constructing a model, a subset of the data is usually set-aside for validation purposes. The validation data set is used to calculate the accuracy of predictive qualities of the model.

5. Evaluation and Interpretation process

a. **Evaluate Results** – Once the model is determined, the predicted results are compared with the actual results in the validation dataset.

b. **Review Process** – Verify the accuracy of the process.

c. **Determine Next Steps** – List of possible actions decision.

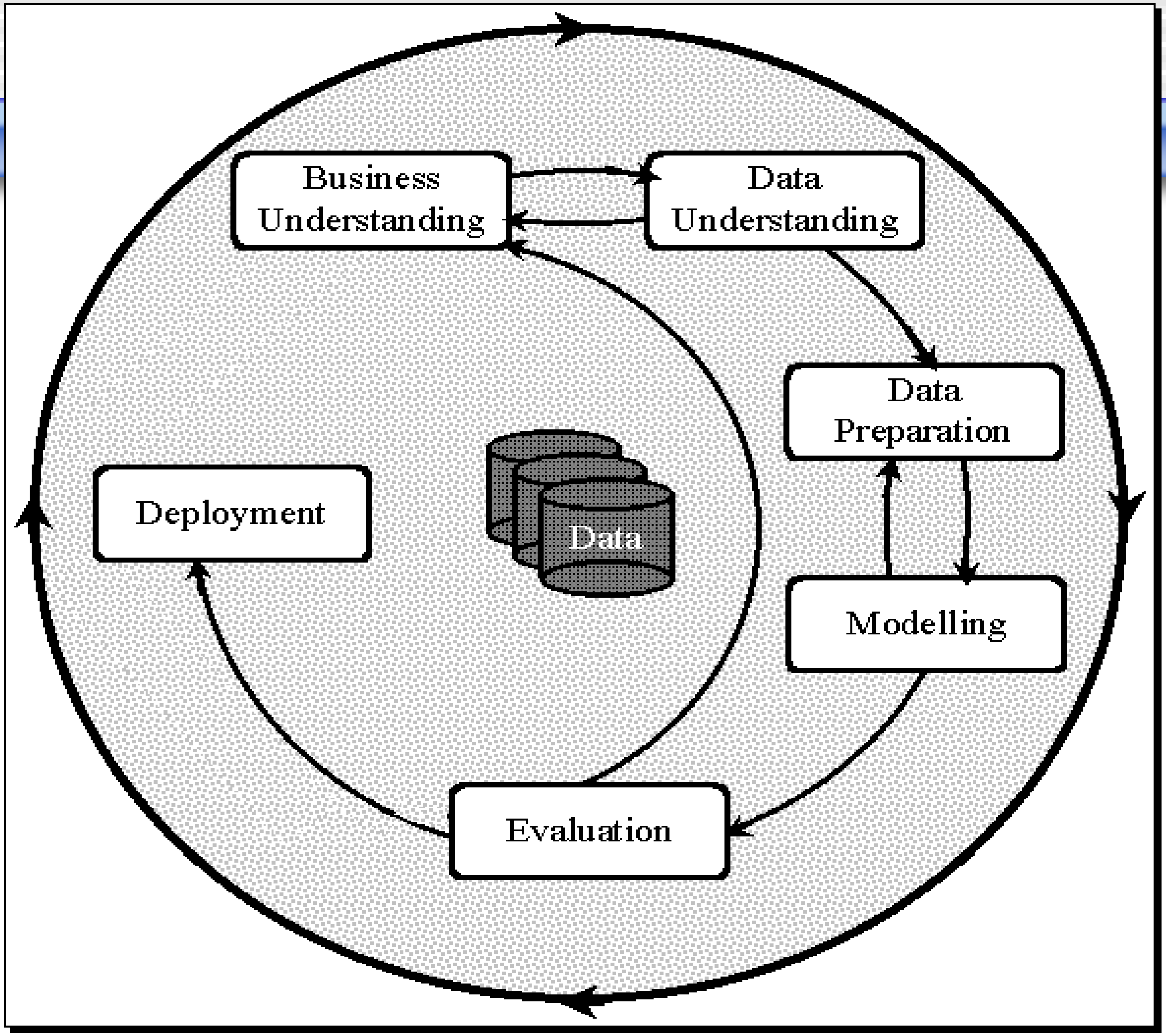
6. Deployment process

a. **Plan Deployment** – This step involves implementing the ‘live’ model within an organization to aid the decision making process..

b. **Produce Final Report** – Write a final report.

c. **Plan Monitoring and Maintenance** – Monitor how well the model predicts the outcomes, and the benefits that this brings to the organization.

d. **Review Project** – Experience, and documentation.



The Iterative Nature of the KDD process

Data Mining Techniques

1. **Predictive Techniques**
   - **Classification:** Data mining techniques in this category serve to classify the discrete outcome variable.
   - **Prediction or Estimation:** DM techniques in this category predict a continuous outcome (as opposed to classification techniques that predict discrete outcomes).
2. **Descriptive Techniques**
   - **Affinity or association:** Data mining techniques in this category serve to find items closely associated in the data set.
   - **Clustering:** DM techniques in this category aim to create clusters of input objects, rather than an outcome variable.

Web Data Mining - Types

1. **Web structure mining –** Examines how the Web documents are structured, and attempts to discover the model underlying the link structures of the Web.
   - ***Intra-page structure mining*** evaluates the arrangement of the various HTML or XML tags within a page
   - ***Inter-page structure*** refers to hyper-links connecting one page to another.
2. **Web usage mining *(Clickstream Analysis)* –** Involves the identification of patterns in user navigation through Web pages in a domain.
   - Processing, Pattern analysis, and Pattern discovery
3. **Web content mining –** Used to discover what a Web page is about and how to uncover new knowledge from it.

Data Mining and Customer Relationship Management

- CRM is the mechanisms and technologies used to manage the interactions between a company and its customers.
- The data mining prediction model is used to calculate a ***score***: a numeric value assigned to each record in the database to indicate the probability that the customer represented by that record will behave in a specific manner.

Barriers to the use of DM

- Two of the most significant barriers that prevented the earlier deployment of knowledge discovery in the business relate to:
  - Lack of data to support the analysis
  - Limited computing power to perform the mathematical calculations required by the DM algorithms.

Case Study

- **An application of Rule Induction to real estate appraisal systems**
  - In this case, we seek specific knowledge that we know can be found in the data in databases, but which can be difficult to extract.
  - Procedure to create the decision tree:
    - Data preparation and preprocessing
    - Tree construction
    - House pruning
    - Paired leaf analysis

Case Study

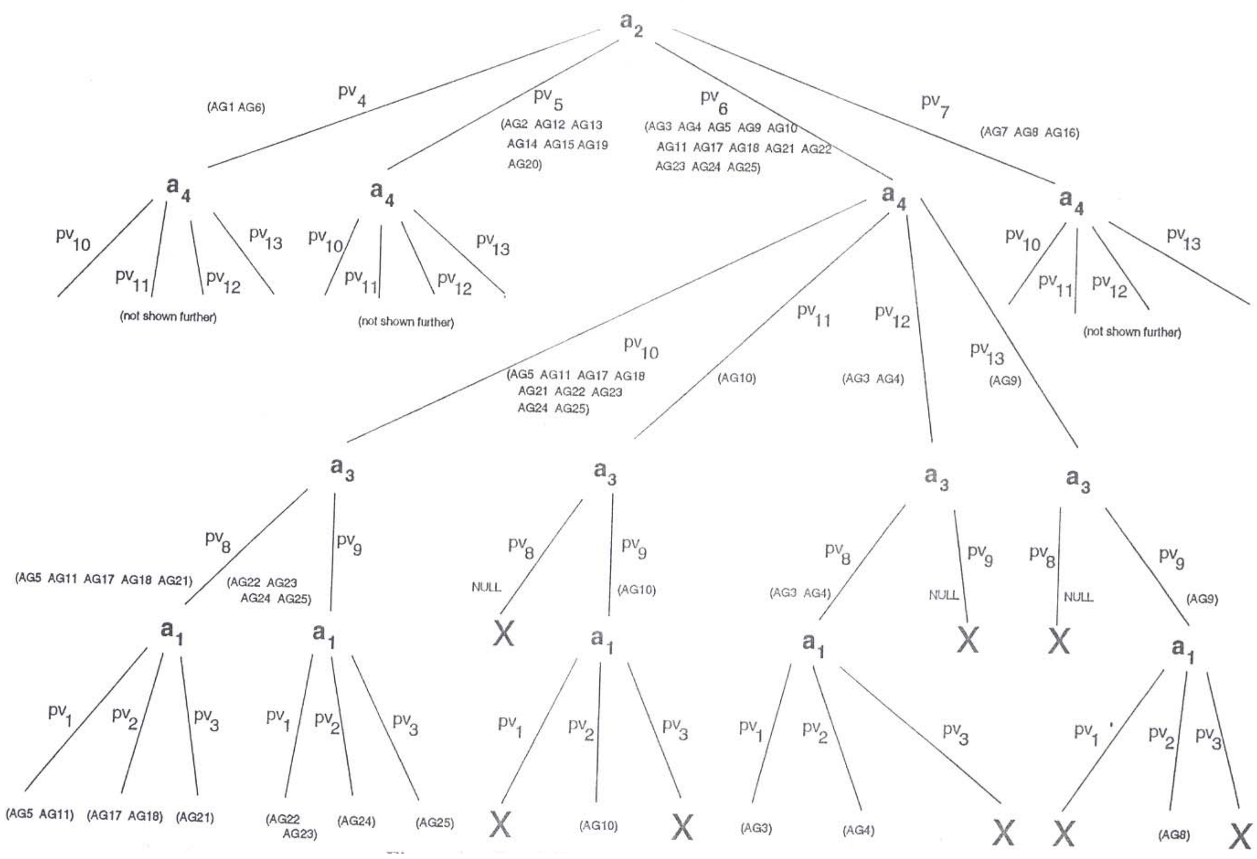
An application of Rule Induction to real estate appraisal systems

Attribute	Induction Results	Expert Estimate	Difference
Living Area	$15 - $31	$15 - $25	0 - 2.4%
Bedrooms	$4311 - $5212	$2500 - $3500	49 - 72%
Bathrooms	$3812 - $5718	$1500 - $2000	154 - 186%
Garage	$3010 - $4522	$3000 - $3500	0.3 - 29%
Pool	$7317 - $11697	$9000 - $12000	2.5 - 19%
Fireplace	$1500 - $4180	$1200 - $2000	25 - 109%
Year Built	1.2 - 1.7%	1.0 - 1.2%	20 - 42%

Summary of Induction Results

Case Study

An application of Rule Induction to real estate appraisal systems



Partial Decision Tree Results for Real Estate Appraisal

Case Study

- **An application of Web Content mining to Expertise Locator Systems**
    - NASA Expert Seeker Web Miner demo
    - A KM system that locates experts based on published documents requires:
        - *Automatic method for identifying employee names.*
        - *A method to associate employee names with skill keywords embedded in those documents.*

Conclusions

In this Chapter we:

- Described knowledge discovery systems, including design considerations, and how they rely on mechanisms and technologies
- Learned how knowledge is discovered:
  - Through through socialization with other knowledgeable persons
  - Trough DM by finding interesting patterns in observations, typically embodied in explicit data
- Explained data mining (DM) technologies
- Discussed the role of DM in customer relationship management

Chapter 13

Knowledge Discovery Systems: Systems That Create Knowledge

Becerra-Fernandez, et al. -- Knowledge
Management 1/e -- © 2004 Prentice Hall