CRISP-DM
(required for cw,
useful for any project…)

Based on Intro to Data Mining:

CRISP-DM

Prof Chris Clifton, Purdue Univ

Thanks also to Laura Squier, SPSS for some of the material

Center for Education and Research
in Information Assurance and Security

Data Mining Process

- Cross-Industry Standard Process for Data Mining (CRISP-DM) – a Methodology, not for Software Engineering, but data-analysis work
- European Community funded effort to develop framework for data mining and text mining tasks
- Goals:
  - Encourage interoperable tools across entire data mining process, by defining subtasks
  - Take the mystery/high-priced expertise out of simple data mining tasks – anyone can do it! (even students)

Why Should There be a Standard Process?

The data mining process must be reliable and repeatable by people with little data mining background.

- Framework for recording experience
  - Allows projects to be replicated, "real science"
- Aid to project planning and management
- "Comfort factor" for new adopters
  - Demonstrates maturity of Data Mining
  - Reduces dependency on "stars"

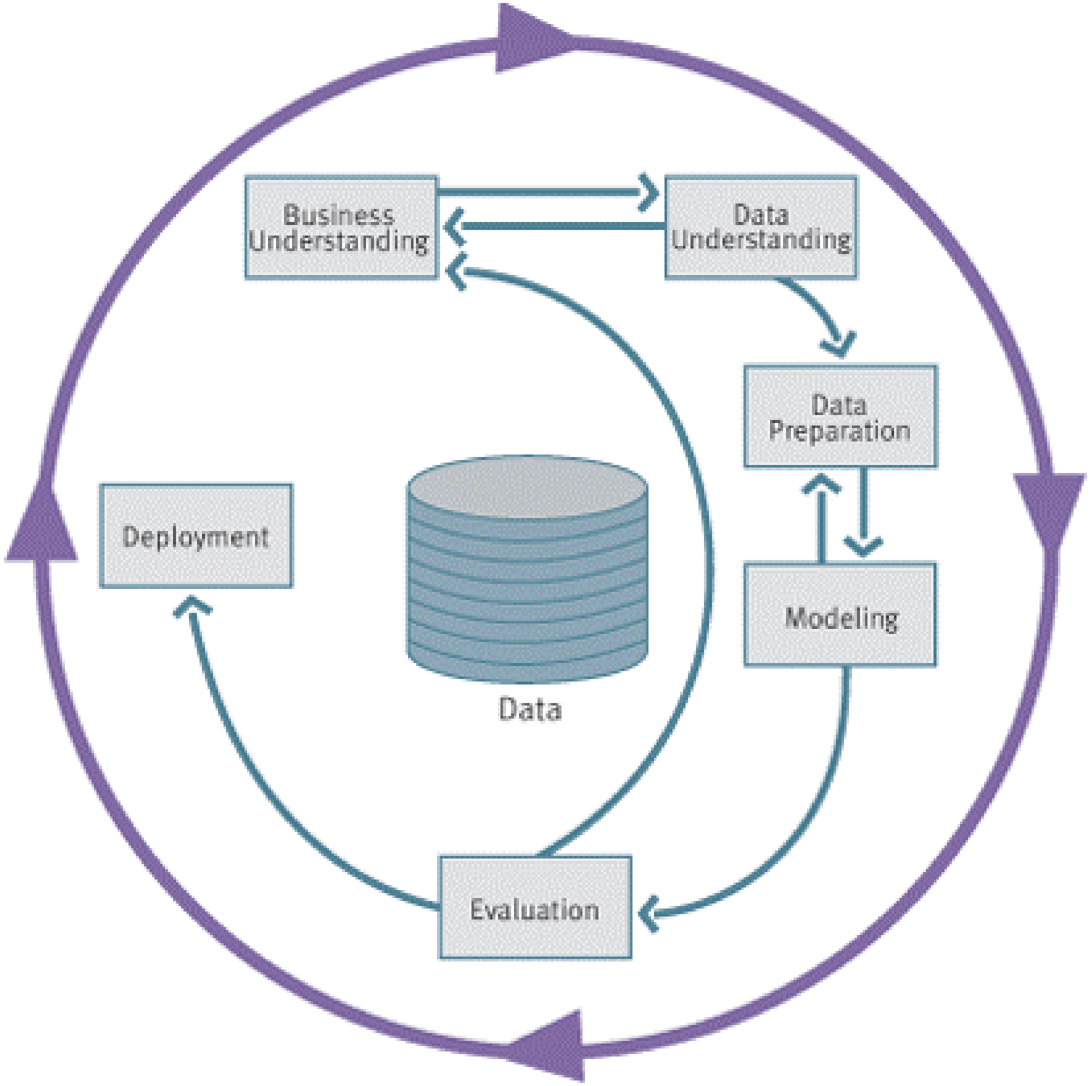
Process Standardization

- CRoss Industry Standard Process for Data Mining
- Initiative launched Sept.1996
- http://www.crisp-dm.org/
- SPSS/ISL, NCR, Daimler-Benz, OHRA
- Funding from European commission
- Over 200 members of the CRISP-DM SIG worldwide
  - DM Vendors - SPSS, NCR, IBM, SAS, SGI, Data Distilleries, Syllogic, Magnify, ..
  - System Suppliers / consultants - Cap Gemini, ICL Retail, Deloitte & Touche, …
  - End Users - BT, ABB, Lloyds Bank, AirTouch, Experian, ...
  - Linkedin.com group discussion

CRISP-DM

- Non-proprietary
- Application/Industry neutral
- Tool neutral
- Focus on business issues and practical problems
  - As well as technical analysis
- Framework for guidance
- Experience base
  - Templates and case studies for guidance and analysis

CRISP-DM: Overview



CRISP-DM: Phases

- **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
- **Modeling**
  - Run the data analysis and data mining tools
- **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

Phases and Tasks/Reports

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

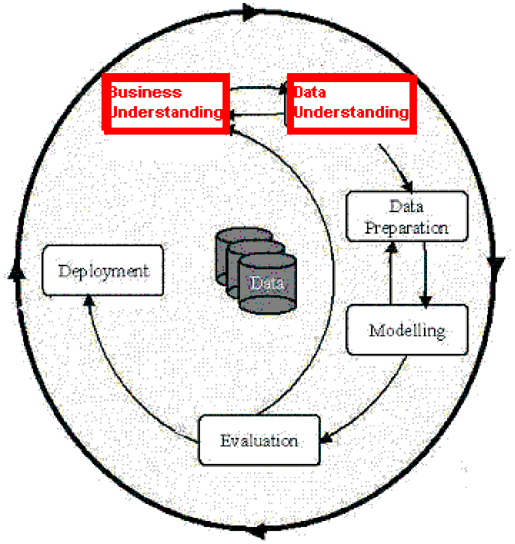
Plan Monitoring and Maintenance
Monitoring and Maintenance Plan

Produce Final Report
Final Report
Final Presentation

Review Project
Experience Documentation

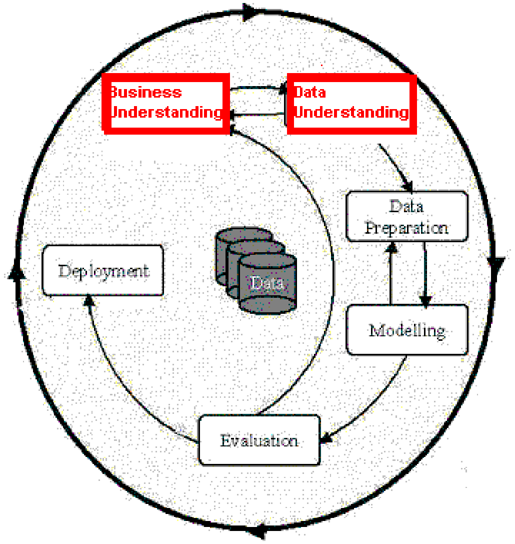
Phases in the DM Process (1)

- Business Understanding:
  - Statement of Business Objective
  - Statement of Data Mining objective
  - Statement of Success Criteria



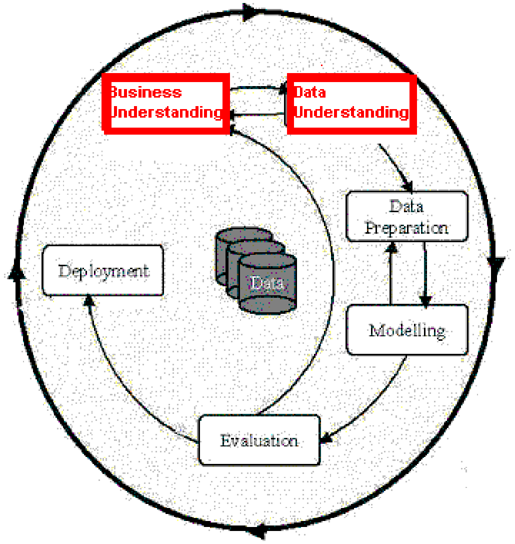
Phases in cw DM Process (1)

- Business Understanding:
  - Business Objective: learn Weka basics; learn to find texts on “subject of interest”
  - Data Mining objective: create and use arff files, explore classifiers and evaluate: measure “goodness”
  - Success Criteria: specific evidence: docx with answers and screenshots; set of attributes and values in an arff file, example classifier outputs, analysis of performance



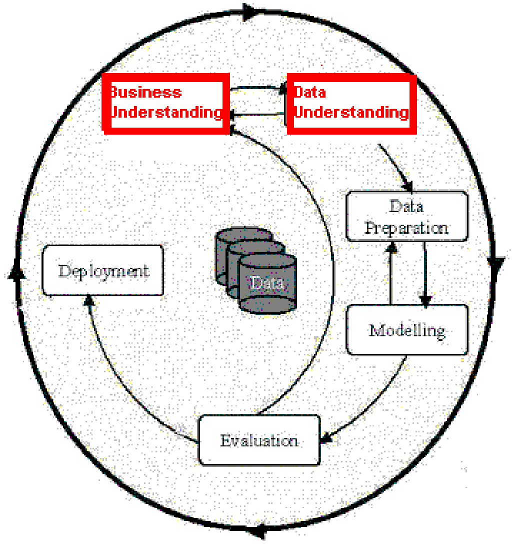
Phases in the DM Process (2)

- Data Understanding
  - Collect data
  - Describe data
  - Explore the data
  - Verify the quality and identify outliers



Phases in cw DM Process (2)

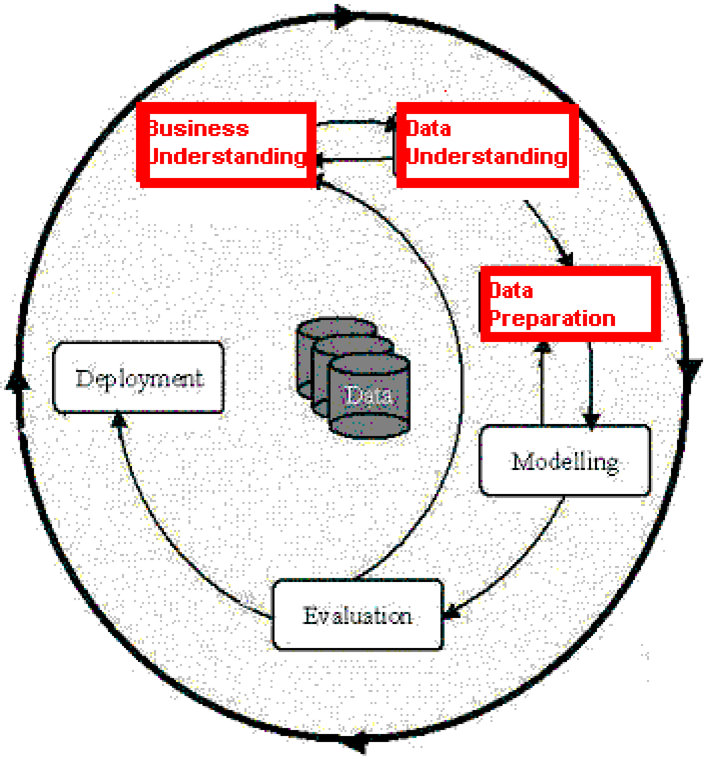
- Data Understanding
  - Data is provided; google ReutersCorn-train.arff
  - attribute: string, and CLASS of each news-story
  - Explore/verify that the features and values “seem relevant and sensible” – if not, how to transform?
  - identify outliers (values which don’t “belong”)



Phases in the DM Process (3)

Data preparation:

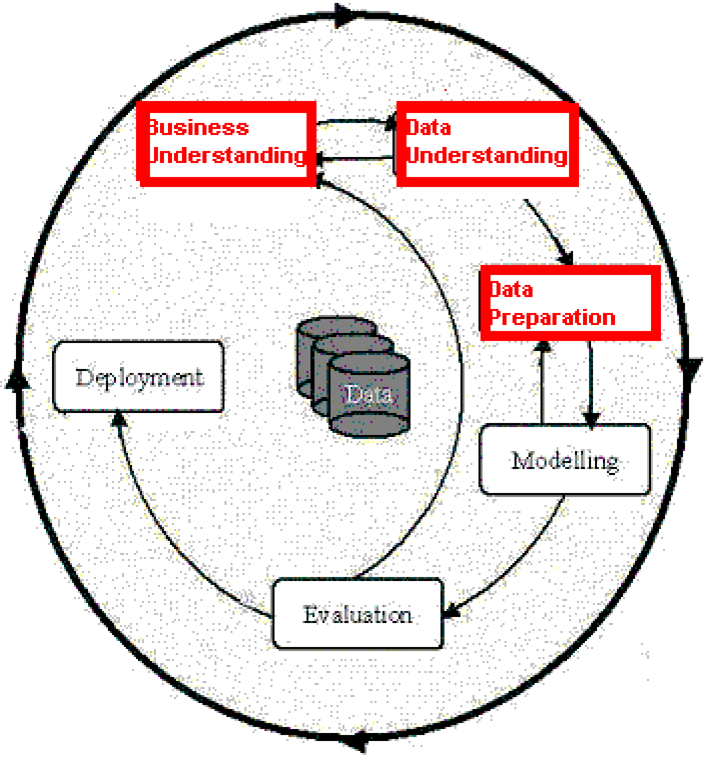
- Can take over 90% of the time
  - Consolidation and Cleaning
    - missing values
    - Remove “noisy” data, repetitions, etc
    - Remove outliers?
  - Feature selection
    - Select features
    - Use visualization tools
  - Transformations - create new variables, change formats



Phases in cw DM Process (3)

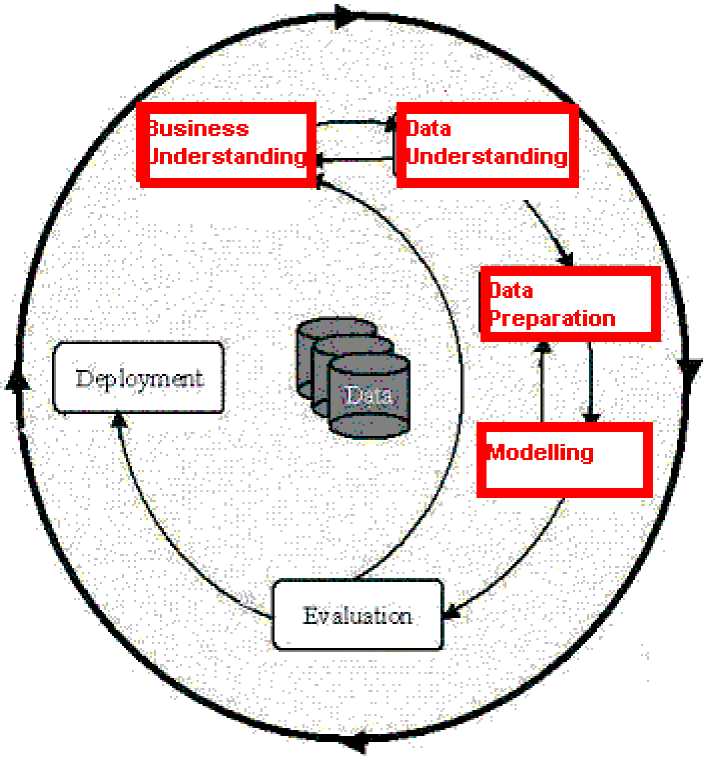
Data preparation:

- do NOT take up 90% of your time!
- **Clean Data**
- *Look for repetitions?*
- *Remove illegal values – eg text in number fields (you should not have this anyway!)*
- **Select Features**
- *Rationale for Inclusion / Exclusion: if it isn‘t relevant to classification – remove*
- **Transform Data**
- *(maybe) add attributes, eg StringToWordVector filter?*
- *Split into train and test parts?*



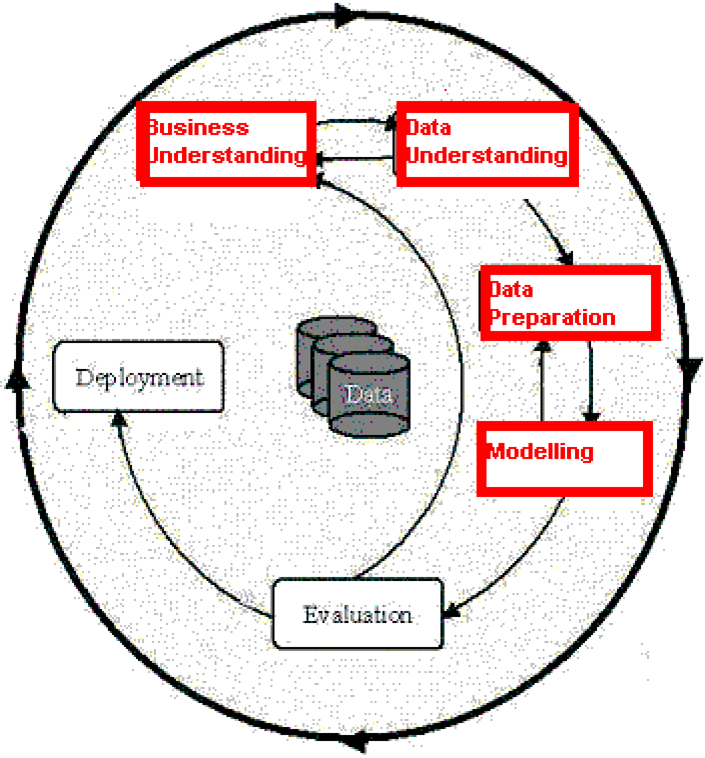
Phases in the DM Process(4)

- Model building
  - Selection of the modeling techniques is based upon the data mining objective
  - Modeling can be an iterative process; may model for description or prediction (or both)



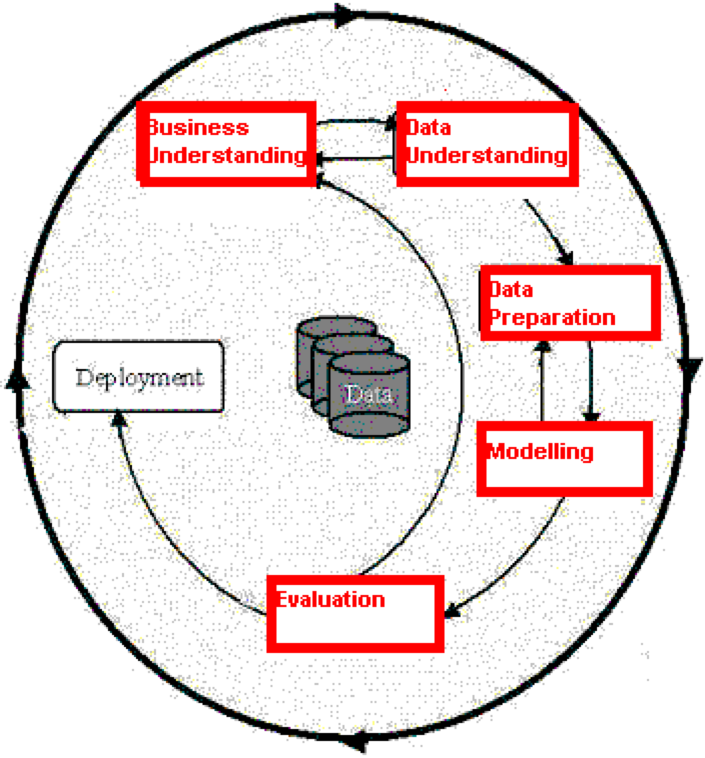
Phases in cw DM Process(4)

- Model building
  - Data Mining objective is to explore and learn – so try several classifiers
  - “model” can be ZeroR rule, or J48 Decision Tree, or other classifiers
  - Try Data Visualization tools as well as Data Mining
  - For each, record accuracy and confusion matrix
  - Capture a few screenshots



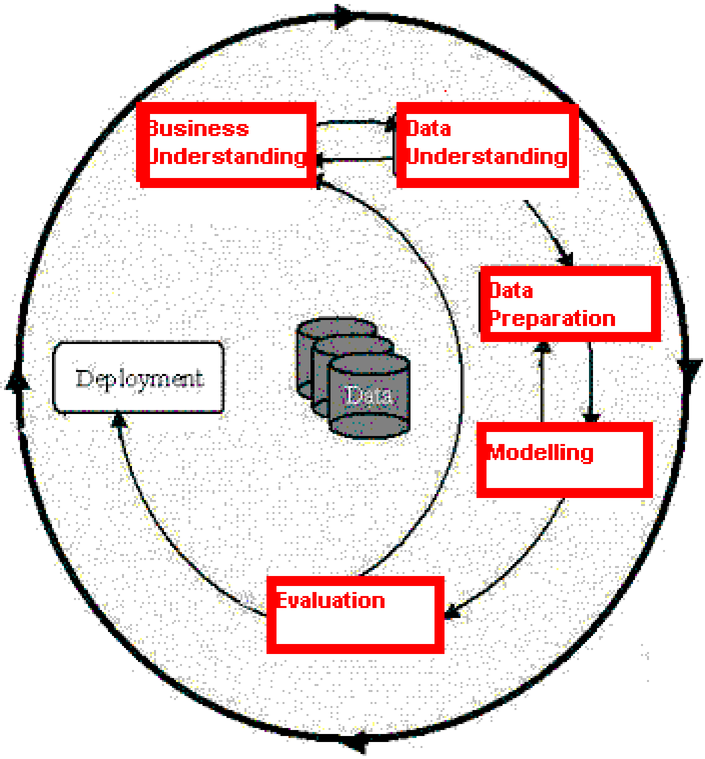
Phases in the DM Process(5)

- Model Evaluation
  - Evaluation of model: how well it performed, how well it met business needs
  - Methods and criteria depend on model type:
    - e.g., confusion matrix with classification models, ALSO meeting business goals: "understanding"
  - Interpretation of models: important or not, easy or hard depends on algorithm



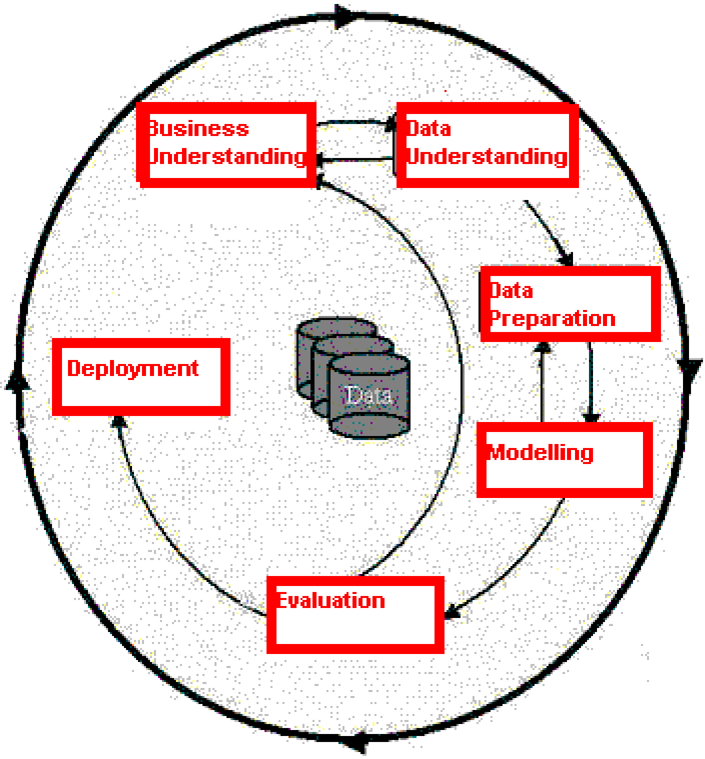
Phases in cw DM Process(5)

- Model Evaluation
  - Evaluation of model: have you found and quantified key features to classify the data?
  - Interpretation: don’t just present the results, try to explain possible reasons, e.g. news-story as a vector of words is a poor representation of its “meaning”



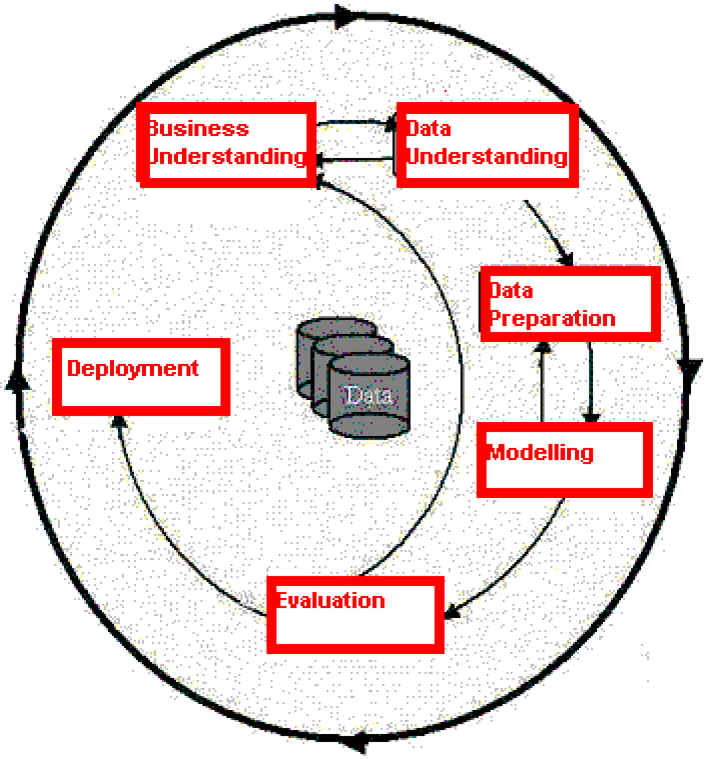
Phases in the DM Process (6)

- 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:
  - Producing report for users, with recommendations to improve their business
  - Deploy the results directly in the business



Phases in cw DM Process (6)

- Deployment
  - Write a report with answers to Ch17 questions
  - Deploy directly: use what you learned in your future studies - to pass the Exam, then in Project etc



Why CRISP-DM?

- The data mining process must be reliable and repeatable by people with little data mining skills (e.g. IT Consultants, students?...)

- CRISP-DM provides a uniform framework for
  - guidelines
  - experience documentation

- CRISP-DM is flexible to account for differences
  - Different business/agency problems
  - Different data

Why DM?: Concept Description

- Descriptive vs. predictive data mining
  - Descriptive mining: describes concepts or task-relevant data sets in concise, informative form: Decision Tree, Decision Rules, …
  - Predictive mining: Based on data and analysis, constructs models from the data-set, and predicts the trend and properties of unknown data: “model” need not be visualized, eg Neural Net, Ensemble
- Concept description:
  - Characterization: provides a concise and succinct summarization of the given collection of data

Data Mining v. Visualization

- Data Mining:
  - can handle complex data types of many attributes/features/dimensions
  - a more automated process
- OLAP Online Analytic Processing (Visualization):
  - restricted to a small number of dimensions and feature types (eg not so good for text)
  - user-controlled process

CRISP-DM: Summary

- **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
- **Modeling**
  - Run the data mining tools
- **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