CRISP - DM

Cross-Industry Standard Process for Data Mining

Data Mining Process

- Cross-Industry Standard Process for Data Mining (CRISP-DM)
- European Community funded effort to develop framework for data mining tasks
- Goals:
  - Encourage interoperable tools across entire data mining process
  - Take the mystery/high-priced expertise out of simple data mining tasks

Why Should There be a Standard Process?

- Framework for recording experience
  - Allows projects to be replicated
- Aid to project planning and management
- "Comfort factor" for new adopters
  - Demonstrates maturity of Data Mining
  - Reduces dependency on "stars"
- Encourage best practices and help to obtain better results

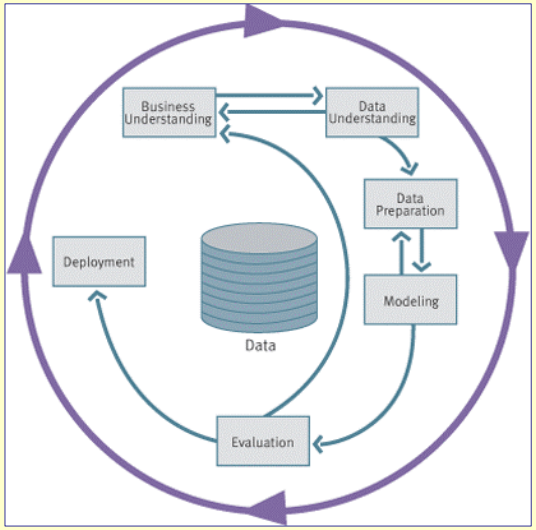
Process Standardization

- Initiative launched in late 1996 by three "veterans" of data mining market.
  - Daimler Chrysler (then Daimler-Benz), SPSS (then ISL) , NCR.
- Developed and refined through series of workshops (from 1997-1999)
- Over 300 organization contributed to the process model
- Published CRISP-DM 1.0 (1999)
- 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, ...

CRISP-DM

- Non-proprietary
- Application/Industry neutral
- Tool neutral
- Focus on business issues
  - As well as technical analysis
- Framework for guidance
- Experience base
  - Templates for Analysis

CRISP-DM: Overview



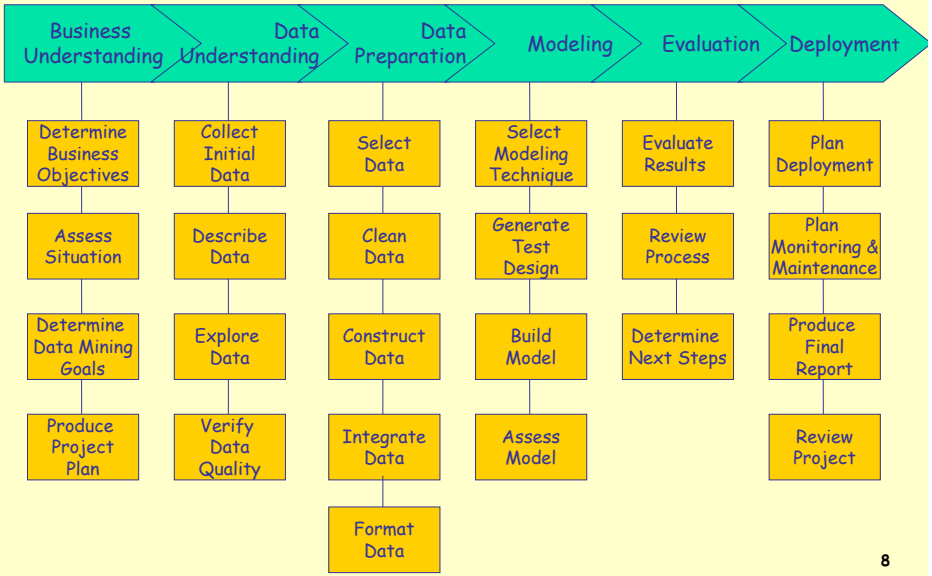
CRISP-DM is a comprehensive data mining methodology and process model that provides anyone—from novices to data mining experts—with a complete blueprint for conducting a data mining project.

CRISP-DM breaks down the life cycle of a data mining project into six phases.

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 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 continuous mining of the data

Phases and Tasks



Phase 1 - Business Understanding

- Statement of Business Objective
  - States goal in business terminology
- Statement of Data Mining objective
  - States objectives in technical terms
- Statement of Success Criteria

- Focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives
  - What the client really wants to accomplish?
  - Uncover important factors (constraints, competing objectives).

Phase 1 - Business Understanding

- **Determine business objectives**
  - Key persons and their roles? Is there a steering committee. Internal sponsor *(financial, domain expert).*
  - Business units impacted by the project *(sales, finance,...)* ? Business success criteria and who assesses it?
  - Users' needs and expectations.
  - Describe problem in general terms. Business questions, Expected benefits.
- **Assess situation**
  - Are they already using data mining.
  - Identify hardware and software available. Identify data sources and their types *(online, experts, written documentation).*
  - Identify knowledge sources and types *(online, experts, written documentation)*
  - Describe the relevant background.

Phase 1 - Business Understanding

- **Determine data mining goals**
  - Translate the business questions to data mining goals
    - *(e.g., a marketing campaign requires segmentation of customers in order to decide whom to approach in this campaign; the level/size of the segments should be specified).*
  - Specify data mining problem type
    - *(e.g., classification, description, prediction and clustering).*
  - Specify criteria for model assessment.
- **Produce project plan**
  - Define initial process plan; discuss its feasibility with involved personnel.
  - Put identified goals and selected techniques into a coherent procedure.
  - Estimate effort and resources needed; Identify critical steps.

Phase 2 - Data Understanding

- Acquire the data
- Explore the data (query & visualization)
- Verify the quality

- Starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

Phase 2 - Data Understanding

- **Collect data**
  - List the datasets acquired (locations, methods used to acquire, problems encountered and solutions achieved).
- **Describe data**
  - Check data volume ans examine its gross properties.
  - Accessibility and availability of attributes. Attribute types, range, correlations, the identities.
  - Understand the meaning of each attribute and attribute value in business terms.
  - For each attribute, compute basic statistics (e.g., distribution, average, max, min, standard deviation, variance, mode, skewness).

Phase 2 - Data Understanding

- **Explore data**
  - Analyze properties of interesting attributes in detail
    - *Distribution, relations between pairs or small numbers of attributes, properties of significant sub-populations, simple statistical analyses.*
- **Verify data quality**
  - Identify special values and catalogue their meaning.
  - Does it cover all the cases required? Does it contain errors and how common are they?
  - Identify missing attributes and blank fields. Meaning of missing data.
  - Do the meanings of attributes and contained values fit together?
  - Check spelling of values *(e.g., same value but sometime beginning with a lower case letter, sometimes with an upper case letter).*
  - Check for plausibility of values, e.g. all fields have the same or nearly the same values.

Phase 3 - Data Preparation

- Select and prepare data to be used
- Takes usually over 90% of the time

- Covers all activities to construct the final dataset from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record and attribute selection as well as transformation and cleaning of data for modeling tools.

Phase 3 - Data Preparation

- **Select data**
  - Reconsider data selection criteria.
  - Decide which dataset will be used.
  - Collect appropriate additional data (internal or external).
  - Consider use of sampling techniques.
  - Explain why certain data was included or excluded.
- **Clean data**
  - Correct, remove or ignore noise.
  - Decide how to deal with special values and their meaning *(99 for marital status).*
  - Aggregation level, missing values, etc.
  - Outliers?

Phase 3 - Data Preparation

- **Construct data**
  - Derived attributes.
  - Background knowledge .
  - How can missing attributes be constructed or imputed?
- **Integrate data**
  - Integrate sources and store result (new tables and records).
- **Format Data**
  - Rearranging attributes *(Some tools have requirements on the order of the attributes, e.g. first field being a unique identifier for each record or last field being the outcome field the model is to predict).*
  - Reordering records *(Perhaps the modeling tool requires that the records be sorted according to the value of the outcome attribute).*
  - Reformatted within-value *(These are purely syntactic changes made to satisfy the requirements of the specific modeling tool, remove illegal characters, uppercase lowercase).*

Phase 4 – Modeling

- Select the modeling technique
  - (based upon the data mining objective)
- Generate test design
  - Procedure to test model quality and validity
- Build model
  - Parameter settings
- Assess model (rank the models)

- 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.

Phase 4 – Modeling

- **Select modeling technique**
  - Select technique
  - Identify any built-in assumptions made by the technique about the data *(e.g. quality, format, distribution).*
  - Compare these assumptions with those in the Data Description Report and make sure that these assumptions hold.
  - Preparation Phase if necessary.
- **Generate test design**
  - Describe the intended plan for train, test and evaluate the models.
  - How to divide the dataset into training, test and validation sets.
  - Decide on necessary steps *(number of iterations, number of folds etc.).*
  - Prepare data required for test.

Phase 4 – Modeling

- **Build model**
  - Set initial parameters and document reasons for choosing those values.
  - Run the selected technique on the input dataset. Post-process data mining results *(eg. editing rules, display trees).*
  - Record parameter settings used to produce the model.
  - Describe the model, its special features, behavior and interpretation.
- **Assess model**
  - Evaluate result with respect to evaluation criteria. Rank results with respect to success and evaluation criteria and select best models.
  - Interpret results in business terms. Get comments by domain experts. Check plausibility of model.
  - Check model against given knowledge base *(discovered info. novel and useful?)*
  - Check result reliability. Analyze potentials for deployment of each result.

Phase 5 – Evaluation

- Evaluation of model
  - More thoroughly evaluate model
  - Decide how to use results
- 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
- Thoroughly 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. At the end of this phase, a decision on the use of the data mining results should be reached

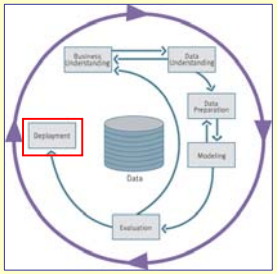
Phase 5 – Evaluation

- **Evaluate results**
  - Understand data mining result. Check impact for data mining goal.
  - Check result against knowledge base to see if it is novel and useful. Evaluate and assess result with respect to business success criteria
  - Rank results according to business success criteria. Check result impact on initial application goal.
  - Are there new business objectives? *(address later in project or new project?)*
  - State conclusions for future data mining projects.
- **Review of process**
  - Summarize the process review *(activities that missed or should be repeated).*
  - Overview data mining process. Is there any overlooked factor or task? *(did we correctly build the model? Did we only use attributes that we are allowed to use and that are available for future analyses?)*
  - Identify failures, misleading steps, possible alternative actions, unexpected paths
  - Review data mining results with respect to business success

Phase 5 – Evaluation

- **Determine next steps**
  - Analyze potential for deployment of each result. Estimate potential for improvement of current process.
  - Check remaining resources to determine if they allow additional process iterations (or whether additional resources can be made available).
  - Recommend alternative continuations. Refine process plan.
- **Decision**
  - According to the results and process review, it is decided how to proceed to the next stage (remaining resources and budget)
  - Rank the possible actions. Select one of the possible actions.
  - Document reasons for the choice.

Phase 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:
  - Scoring a database, utilizing results as business rules, interactive scoring on-line
  - The knowledge gained will need to be organized and presented in a way that the customer can use it. However, depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

Phase 6 – Deployment

- **Plan deployment**
  - How will the knowledge or information be propagated to users? How will the use of the result be monitored or its benefits measured?
  - How will the model or software result be deployed within the organization's systems? How will its use be monitored and its benefits measured (where applicable)?
  - Identify possible problems when deploying the data mining results
- **Plan monitoring and maintenance**
  - What could change in the environment? How will accuracy be monitored?
  - When should the data mining model not be used any more? What should happen if could no longer be used? (Update model, new data mining project)
  - Will the business objectives of the use of the model change over time?

Phase 6 – Deployment

- **Produce a final report**
  - Identify reports needed *(slide presentation, management summary, detailed findings, explanation of models, etc.).* How well initial data mining goals have been met.
  - Identify target groups for reports. Outline structure and contents of reports.
  - Select findings to be included in the reports. Write a report.
- **Review project**
  - Interview people involved in project. Interview end users. What could have been done better? Do they need additional support? Summarize feedback and write the experience documentation
  - Analyze the process (what went right or wrong, what was done well and what needs to be improved.).
  - Document the specific data mining process *(How can results and experience of applying the model be fed back into the process?).* Abstract from details to make the experience useful for future projects.

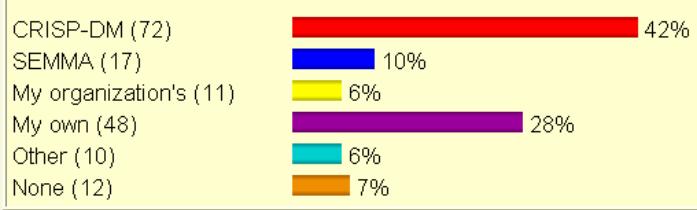
Why CRISP-DM?

- The data mining process must be reliable and repeatable by people with little data mining skills
- CRISP-DM provides a uniform framework for
  - guidelines
  - experience documentation
- CRISP-DM is flexible to account for differences
  - Different business/agency problems
  - Different data

KDnuggets : Polls : Data Mining Methodology (Apr 2004)

Poll

What main methodology are you using for data mining? [170 votes total]



SEMMA is not, however, a comprehensive project management template as CRISP-DM.

The acronym SEMMA -- sample, explore, modify, model, assess -- refers to the core process of conducting data mining.

References

- CRISP-DM 1.0 - Step-by-step data mining guide
  - Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) and Rüdiger Wirth (DaimlerChrysler)
  - http://www.crisp-dm.org/CRISPWP-0800.pdf
- The CRISP-DM Model: The New Blueprint for DataMining, Colin Shearer, JOURNAL of Data Warehousing, Volume 5, Number 4, pag. 13-22, 2000
- Introduction to Data Mining, Prof. Chris Clifton, http://www.cs.purdue.edu/homes/clifton/cs490d/Process.ppt
- CRISP – DM, Yi-Li, http://www.cs.ualberta.ca/~yli/CRISP-DM.ppt

Thank you !!!