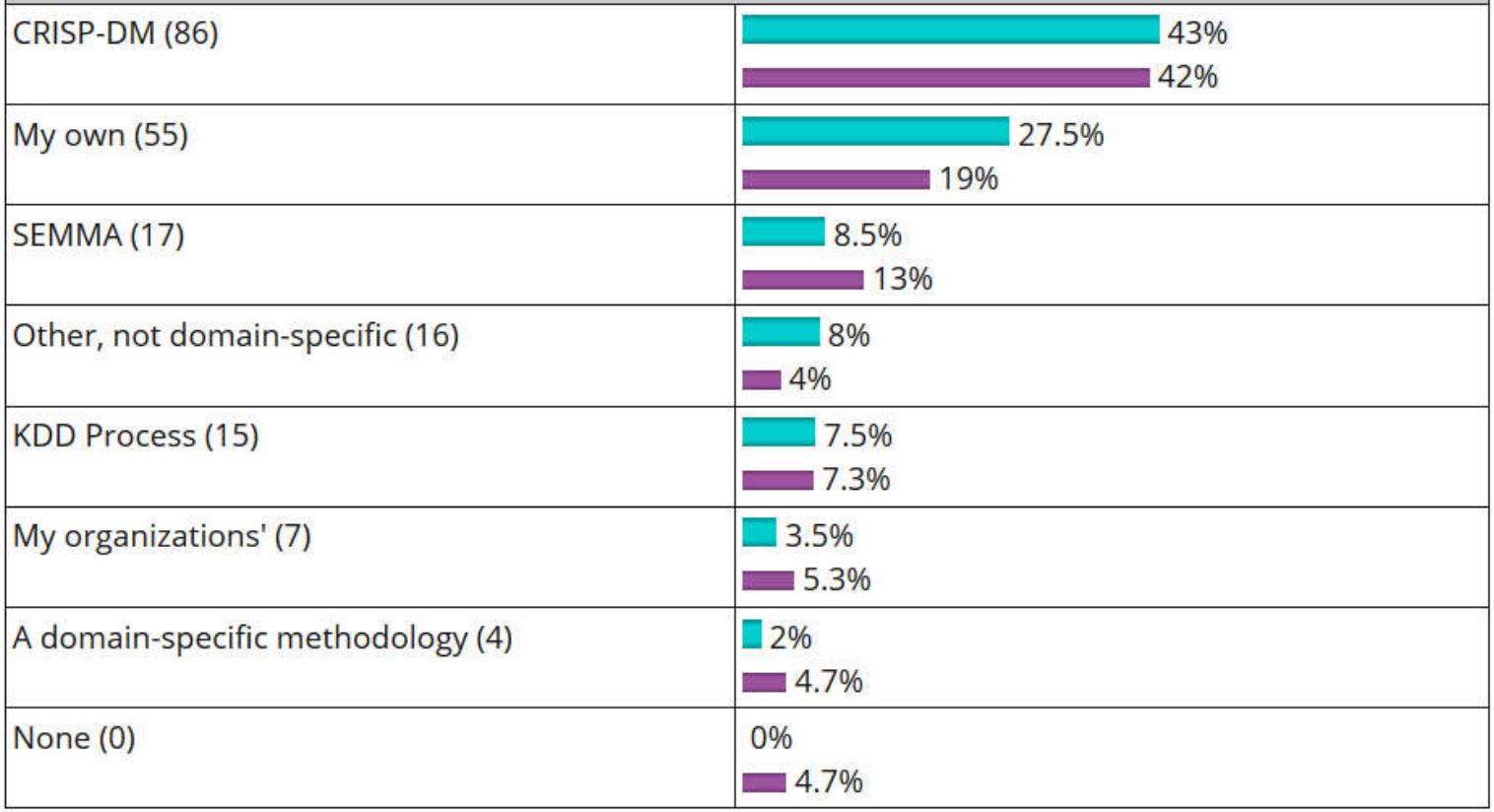
Main methodology for analytics, data mining, or data science projects

Slides by Pimprapai Thainiam

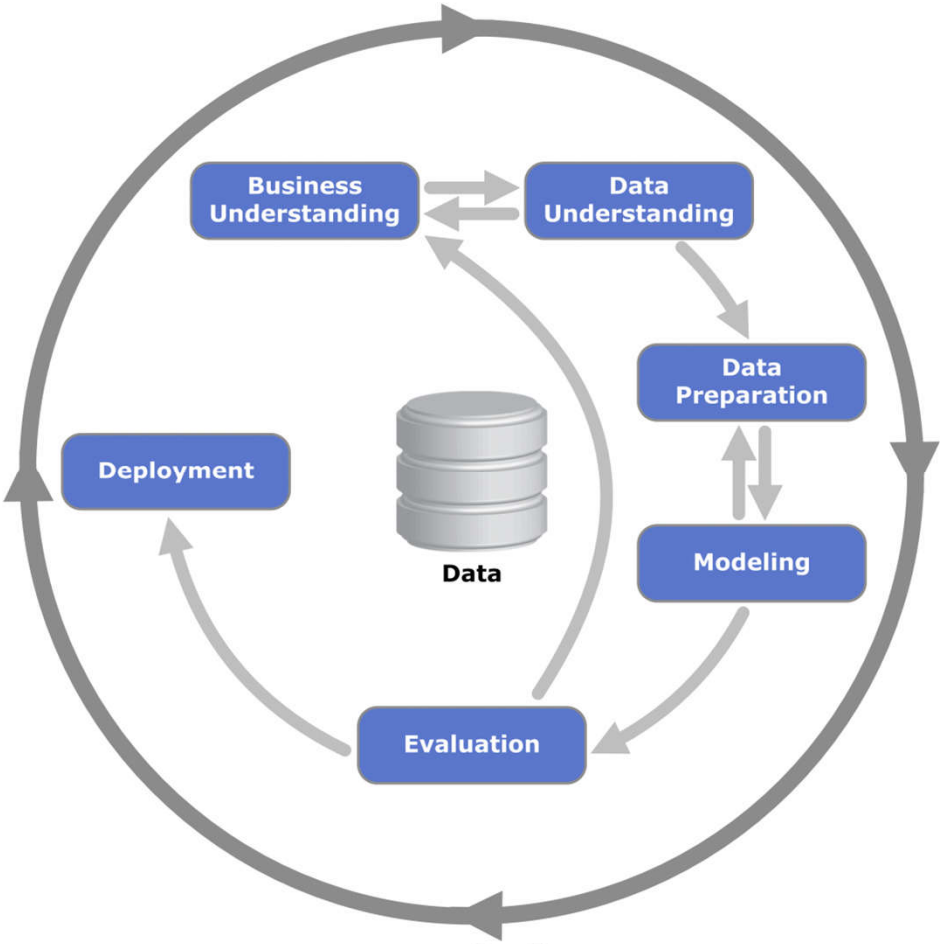
What main methodology are you using for your analytics, data mining, or data science projects ?
[200 votes total]
2014 poll 2007 poll



https://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science methodology.html

CRISP (CRoss-Industry Standard Process for Data Mining)

The virtuous loop of methodology phases



CRISP (CRoss-Industry Standard Process for Data Mining)

It consists on a cycle that comprises six stages

1. **Business understanding** - This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a DM problem definition and a preliminary plan designed to achieve the objectives.
2. **Data understanding** - The data understanding phase 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.
3. **Data preparation** - The data preparation phase covers all activities to construct the final dataset from the initial raw data.
4. **Modeling** - In this phase, various modeling techniques are selected and applied and their parameters are tuned to optimal values.

CRISP (CRoss-Industry Standard Process for Data Mining)

5. Evaluation - At this stage the model (or models) obtained are more thoroughly evaluated and the steps executed to construct the model are reviewed to be certain it properly achieves the business objectives.

6. Deployment - All the knowledge acquired to this point must be organized and presented to the "client" in a usable form. We must define, together with this client, a protocol to reliably deploy the data mining findings.

Business Understanding

Determining Business Objectives

1. Gather background information

- Compiling the business background
- Defining business objectives
- Business success criteria

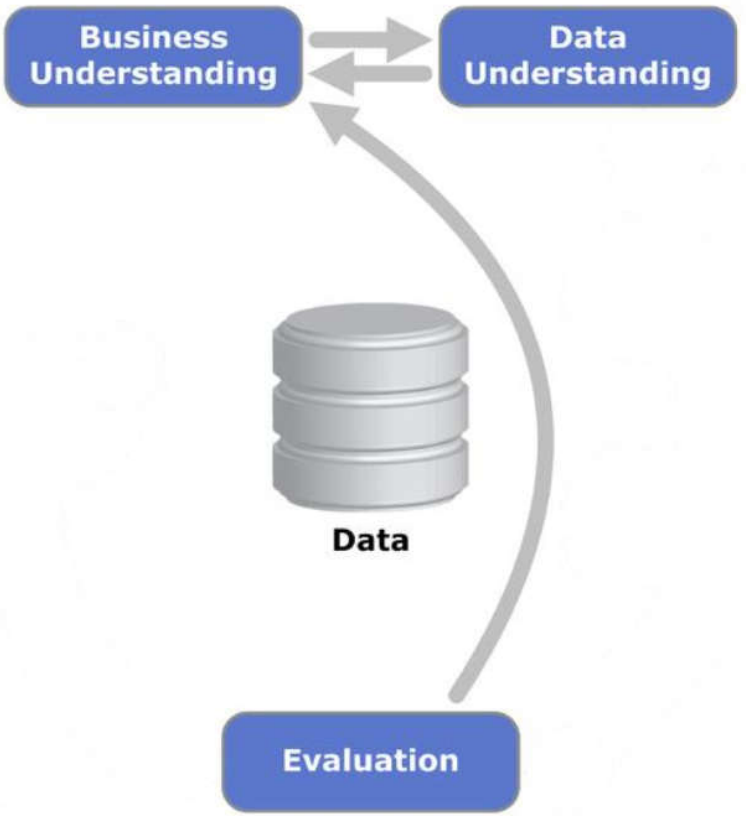
2. Assessing the situation

- Resource Inventory
- Requirements, Assumptions, and Constraints
- Risks and Contingencies
- Cost/Benefit Analysis

3. Determining data science goals

- Data science goals
- Data science success criteria

4. Producing a Project Plan



Ready for the Data Understanding?

From a business perspective:

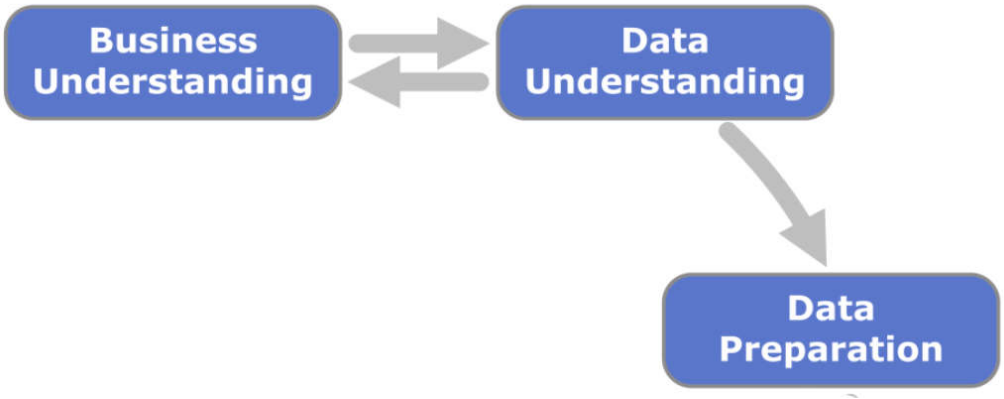
- What does your business hope to gain from this project?
- How will you define the successful completion of our efforts?
- Do you have the budget and resources needed to reach our goals?
- Do you have access to all the data needed for this project?
- Have you and your team discussed the risks and contingencies associated with this project?
- Do the results of your cost/benefit analysis make this project worthwhile?

From a data science perspective:

- How specifically can data mining help you meet your business goals?
- Do you have an idea about which data mining techniques might produce the best results?
- How will you know when your results are accurate or effective enough? (Have we set a measurement of data mining success?)
- How will the modeling results be deployed? Have you considered deployment in your project plan?
- Does the project plan include all phases of CRISP-DM?
- Are risks and dependencies called out in the plan?

Data Understanding

1. **Collect initial data**
   - Existing data
   - Purchased data
   - Additional data
2. **Describe data**
   - Amount of data
   - Value types
   - Coding schemes
3. **Explore data**
4. **Verify data quality**
   - Missing data
   - Data errors
   - Coding inconsistencies



Ready for the Data Preparation?

- Are all data sources clearly identified and accessed? Are you aware of any problems or restrictions?
- Have you identified key attributes from the available data?
- Did these attributes help you to formulate hypotheses?
- Have you noted the size of all data sources?
- Are you able to use a subset of data where appropriate?
- Have you computed basic statistics for each attribute of interest? Did meaningful information emerge?
- Did you use exploratory graphics to gain further insight into key attributes? Did this insight reshape any of your hypotheses?
- What are the data quality issues for this project? Do you have a plan to address these issues?
- Are the data preparation steps clear? For instance, do you know which data sources to merge and which attributes to filter or select?

Data Preparation

1. Select right data

- Select training examples
- Select features

2. Clean data

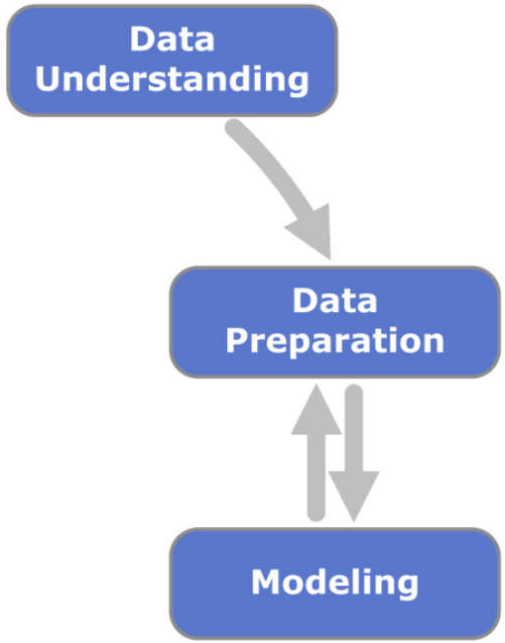
- Fill in missed data
- Correct data errors
- Make coding consistent

3. Extend data

- Extend training examples
- Extend features

4. Format data

- Put data in a format for training the model



Ready for the Modeling?

- Based upon your initial exploration and understanding, were you able to select relevant subsets of data?
- Have you cleaned the data effectively or removed unsalvageable items? Document any decisions in the final report.
- Are multiple data sets integrated properly? Were there any merging problems that should be documented?
- Have you researched the requirements of the modeling tools that you plan to use?
- Are there any formatting issues you can address before modeling? This includes both required formatting concerns as well as tasks that may reduce modeling time.

Modeling

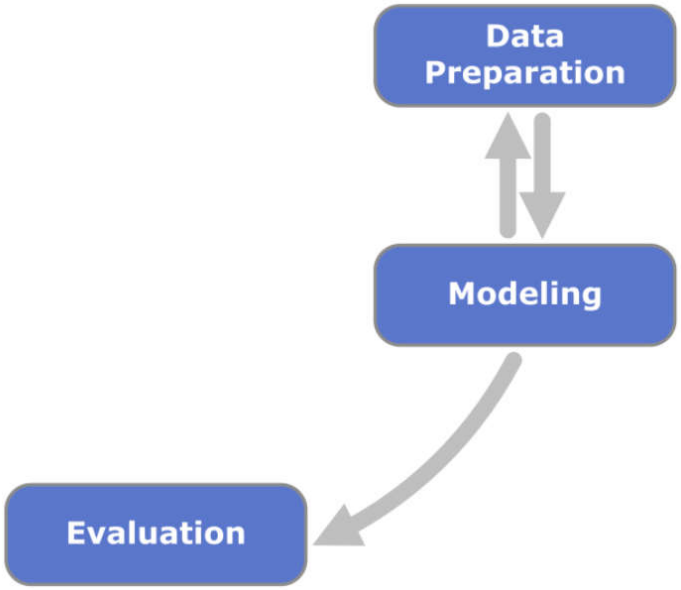
1. Select modeling techniques

- Select data types available for analysis
- Select an algorithm or a model
- Define modeling goals
- State specific modeling requirements

2. Build the model

- Set up hyperparameters
- Train the model
- Describe the result

3. Assess the model



Ready for the Evaluation?

- Are you able to understand the results of the models?
- Do the model results make sense to you from a purely logical perspective? Are there apparent inconsistencies that need further exploration?
- From your initial glance, do the results seem to address your organization's business question?
- Have you used analysis nodes and lift or gains charts to compare and evaluate model accuracy?
- Have you explored more than one type of model and compared the results?
- Are the results of your model deployable?

Evaluation

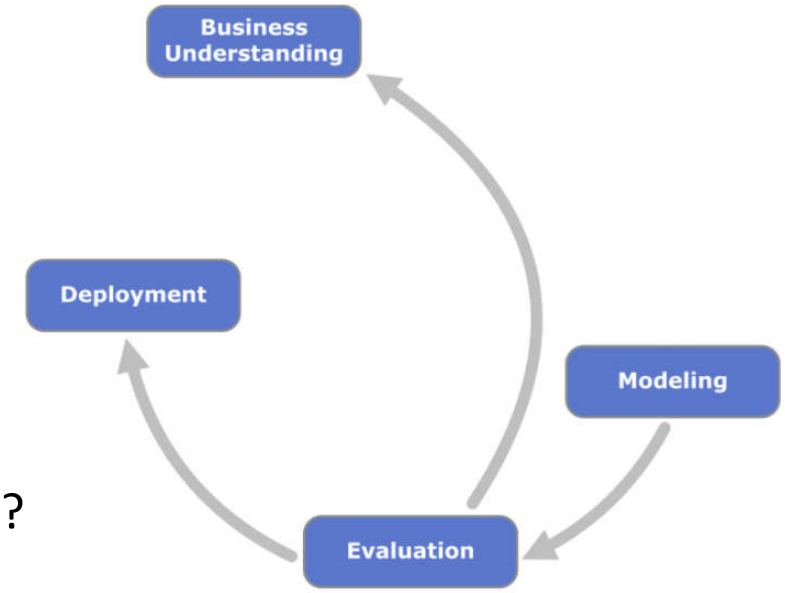
1. Evaluate the results

- Are results presented clearly?
- Are there any novel findings?
- Can models and findings be applicable to business goals?
- How well do the models and findings answer business goals?
- What additional questions the modeling results have risen?

2. Review the process

- Did the stage contribute to the value of the results?
- What went wrong and how it can be fixed?
- Are there alternative decisions which could have been executed?

3. Determine the next steps



Deployment

1. Planning for deployment

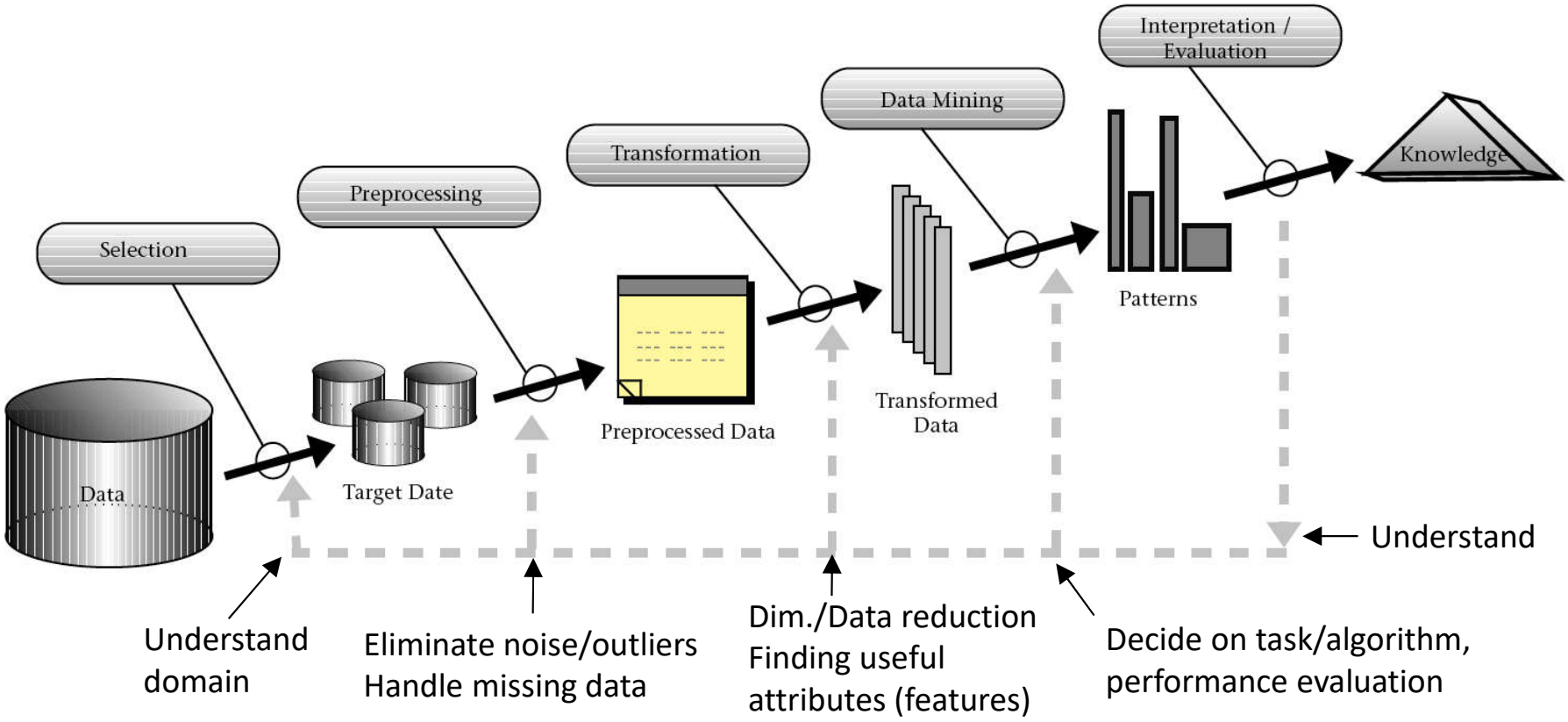
- Summarize models and findings
- For each model create a deployment plan
- Identify any deployment problems and plan for contingencies

2. Plan monitoring and maintenance

- Identify models and findings which require support
- How can the accuracy and validity be evaluated?
- How will you determine that a model has expired?
- What to do with the expired models?

3. Conduct a final project review

KDD (Knowledge Discovery in Databases)



From: Fayyad, Piatetsky-Shapiro, Smyth, "From Data Mining to Knowledge Discovery: An Overview"

KDD (Knowledge Discovery in Databases)

KDD consists of five stages as follows:

1. **Selection** - This stage consists on creating a target data set, or focusing on a subset of variables or data samples, on which discovery is to be performed
2. **Pre-processing** - This stage consists on the target data cleaning and pre processing in order to obtain consistent data
3. **Transformation** - This stage consists on the transformation of the data using dimensionality reduction or transformation methods
4. **Data Mining** - This stage consists on the searching for patterns of interest in a particular representational form, depending on the DM objective (usually, prediction)
5. **Interpretation/Evaluation** - This stage consists on the interpretation and evaluation of the mined patterns.

SEMMA (Sample, Explore, Modify, Model, Assess)

The SAS Institute considers a cycle with 5 stages for the process:

1. **Sample** - This stage consists on sampling the data by extracting a portion of a large data set big enough to contain the significant information, yet small enough to manipulate quickly.
2. **Explore** - This stage consists on the exploration of the data by searching for unanticipated trends and anomalies in order to gain understanding and ideas.
3. **Modify** - This stage consists on the modification of the data by creating, selecting, and transforming the variables to focus the model selection process.
4. **Model** - This stage consists on modeling the data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome.
5. **Assess** - This stage consists on assessing the data by evaluating the usefulness and reliability of the findings from the DM process and estimate how well it performs.

Summary of the correspondences between KDD, SEMMA and CRISP-DM

KDD	SEMMA	CRISP-DM
Pre KDD	----------	Business understanding
Selection	Sample	Data Understanding
Pre processing	Explore	
Transformation	Modify	Data preparation
Data mining	Model	Modeling
Interpretation/Evaluation	Assessment	Evaluation
Post KDD	----------	Deployment