medium.com /datadriveninvestor/data-science-project-management-methodologies-f6913c6b29eb

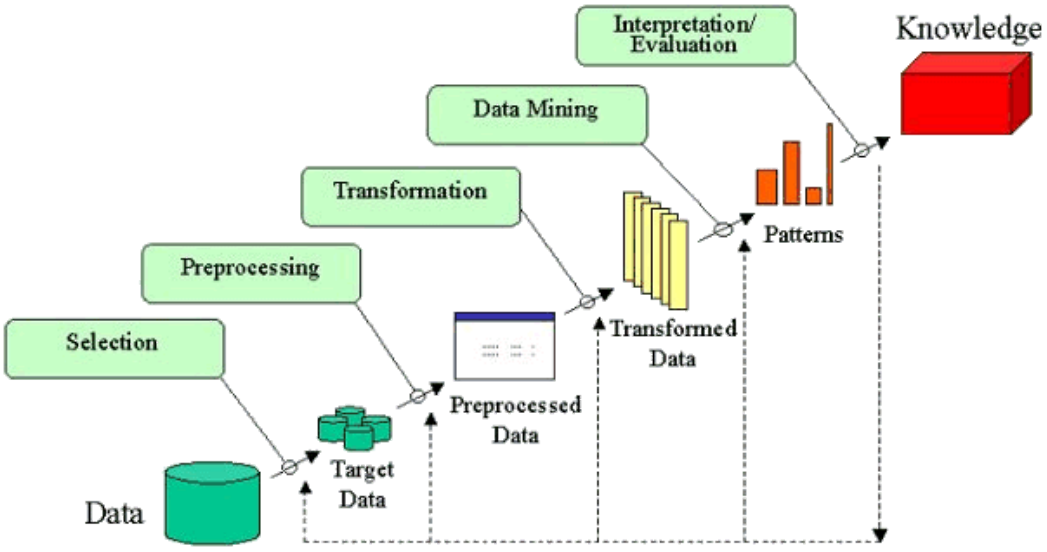
Data Science project management methodologies

Quantum

8-10 minutes

There are several data mining processes, that can be applied to modern Data Science projects. The most common of them are CRISP-DM, SEMMA, KDD. In this article, we are going to review and compare them.

Knowledge Discovery in Databases or KDD, for short, is a method of how specialists can extract patterns and/or required information from data. It consists of five stages — Selection, Preprocessing, Transformation, Data Mining, and Interpretation/Evaluation:



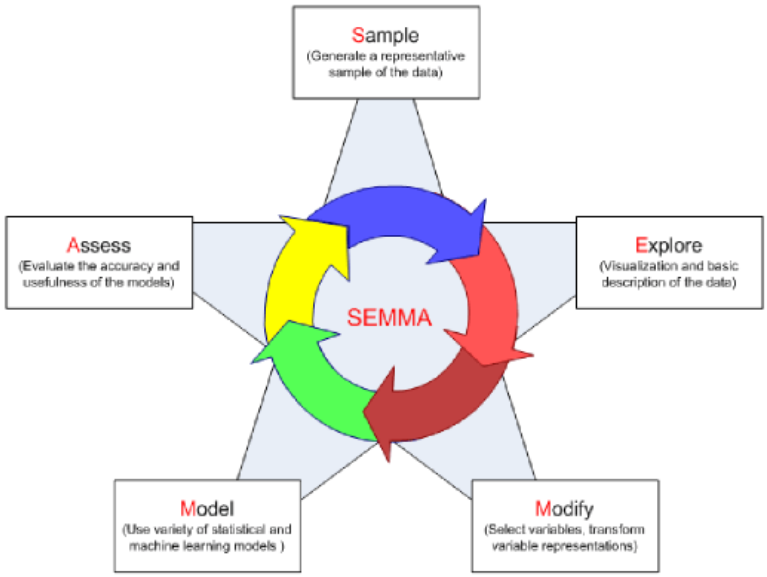
http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/kdd.gif

Let’s take a closer look at each stage. Every stage consists of a set of predetermined actions that are performed:

- **Selection** — creating a target data set, or focusing on a subset of variables or data samples that require further exploration;
- **Pre-processing** — target data pre-processing to obtain consistent data;
- **Transformation** — data transformation using dimensionality reduction or transformation methods;
- **Data Mining** — searching for patterns of interest in a particular representational form that depends on the Data Mining goal (e.g. prediction);
- **Interpretation/Evaluation** — interpretation and evaluation of the mined patterns.

After the cycle completes and all of the stages are completed, the specialist has the *Evaluation* information that indicates, if the *Knowledge* was indeed obtained. If not — the cycle repeats starting with any stage using updated targets until the goal is reached.

SEMMA is the next process, that we will review. It has a similar structure to KDD, but as it does not focus as heavily on data-specific stages, it is easier to apply to general Data Science tasks. Also, it has strictly cyclic nature, unlike KDD. SEMMA is an acronym that stands for Sample, Explore, Modify, Model and Access:



https://slideplayer.com/slide/11825381/

Sample — a portion of a large data set is taken that is big enough to extract significant information and small enough to manipulate quickly.

Explore — data exploration can help in gaining understanding and ideas as well as refining the discovery process by searching for trends and anomalies.

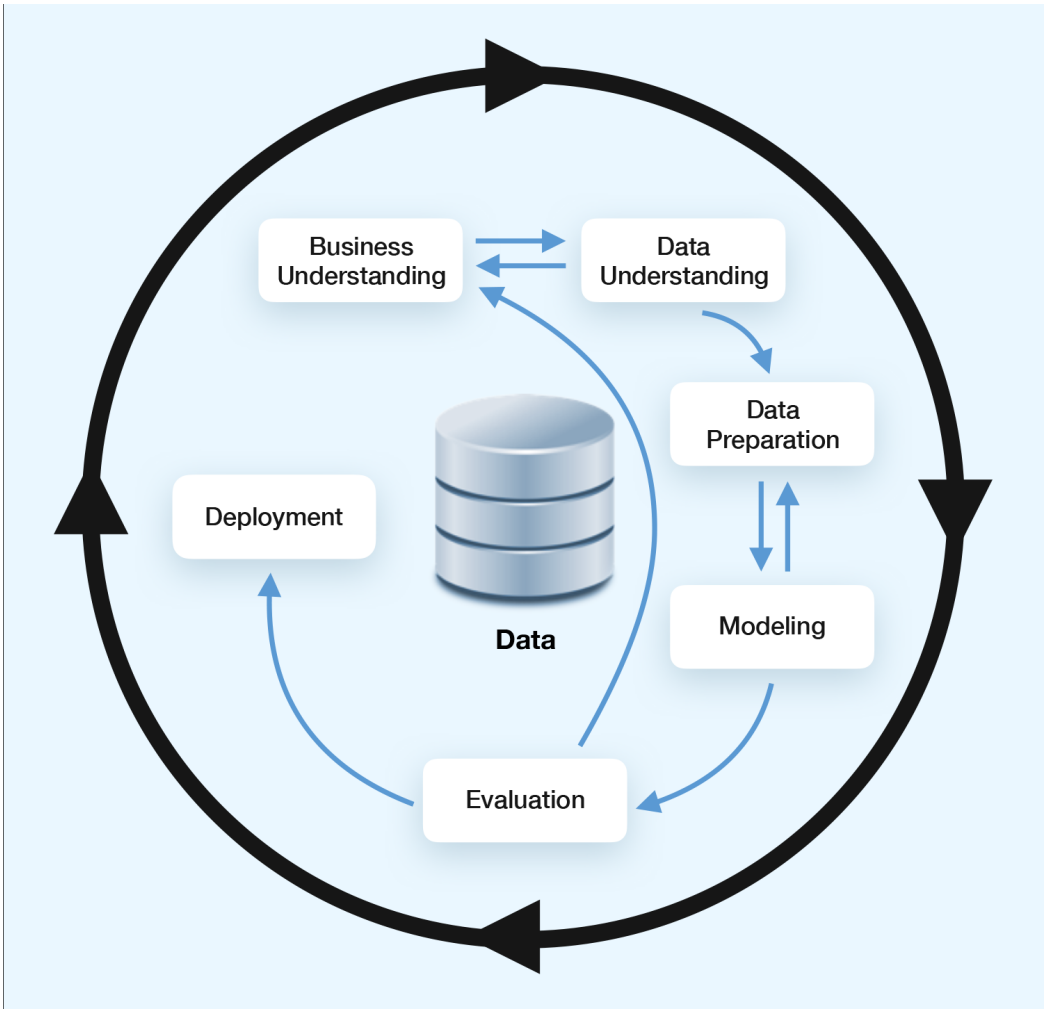
Modify — data modification stage focuses on creating, selecting and transformation of variables to focus model selection process. This stage may also look for outliers and reducing the number of variables.

Model — there are different modeling techniques present and each type of model has its strengths and is appropriate for a specific goal for data mining.

Access — this final stage focuses on the evaluation of the reliability and usefulness of findings and estimates the performance.

In the same way, as in KDD, SEMMA repeats itself until the set goal will be reached.

And last but not least, CRISP-DM. It stands for Cross-Industry Standard Process for Data Mining. This methodology was originally developed in IBM for Data Mining tasks, but our Data Science department finds it useful for almost all of the projects.



This model has the same cyclic nature as both KDD and SEMMA. The key difference in the structure is that the transitions between stages can be reversed. So if during the modeling stage the specialist found the data not sufficient to resolve the goal of the project, they can return to the data preparation stage and select different target variables, generate features, etc, without returning all the way to the start of the cycle.

Below you can find all 6 stages of CRISP-DM, depicted on the image above, their subprocesses, along with outputs for every subprocess.

Business Understanding

This stage is aimed toward getting a general understanding of the client's business. It is crucial in most cases to understand the application of the product to be developed. If it is skipped — you might end up with a large trained neural network, that has to be deployed to a mobile phone and work in realtime.

1. **Determine Business Objectives**
   - Background
   - Business Objectives
   - Business Success Criteria
2. **Assess the Situation**
   - Inventory of Resources
   - Requirements, Assumptions, and Constraints
   - Risks and Contingencies
   - Terminology
   - Costs and Benefits
3. **Determine Goals**
   - Data Mining Goals
   - Data Mining Success Criteria

4. **Produce Project Plan**
   - Project Plan
   - Initial Assessment of Tools and Techniques

Data Understanding

The second stage consists of collecting and exploring the input dataset. The set goal might be unsolvable using the input data, you might need to use public datasets, or even create a specific one for the set goal.

1. **Collect Initial Data**
   - Initial Data Collection Report
2. **Describe Data**
   - Data Description Report
3. **Explore Data**
   - Data Exploration Report
4. **Verify Data Quality**
   - Data Quality Report

Data Preparation

As we all know, bad input inevitably leads to bad output. Therefore no matter what you do in modeling — if you made major mistakes while preparing the data — you will end up returning to this stage and doing it over again.

1. **Select Data**
   - The rationale for Inclusion/Exclusion
2. **Clean Data**
   - Data Cleaning Report
3. **Construct Data**
   - Derived Attributes
   - Generated Records
4. **Integrate Data**
   - Merged Data
5. **Format Data**
   - Reformatted Data

- **Dataset**
- **Dataset Description**

Modeling

This stage is an execution of all of your findings from previous stages. You already know the input to the model, you can tell which models are compatible with the target platform. Now is the time to bring it all to life.

1. **Select Modeling Techniques**
   - Modeling Technique
   - Modeling Assumptions
2. **Generate Test Design**
   - Test Design
3. **Build Model**
   - Parameter Settings
   - Models
   - Model Descriptions
4. **Assess Model**
   - Model Assessment
   - Revised Parameter Settings

Evaluation

This stage is aimed at the evaluation of the obtained results. We need to check if the business goal was fulfilled and plan further steps of the project.

1. **Evaluate Results**
   - Assessment of Data Mining Results w.r.t. Business Success Criteria
   - Approved Models
2. **Review Process**
   - Review of Process
3. **Determine Next Steps**
   - List of Possible Actions
   - Decision

Deployment

If previous stages were successful and there was a decision made to deploy the model — this stage will be activated. You might need to put your model into an existing pipeline, create your own or deploy to cloud computing services.

1. **Plan Deployment**
   - Deployment Plan
2. **Plan Monitoring and Maintenance**
   - Monitoring and Maintenance Plan
3. **Produce Final Report**
   - Final Report
   - Final Presentation
4. **Review Project**
   - Experience Documentation

As you might have noticed, a lot of the outputs are reports. It might look like half of the project's time will be spent on filing those reports, but these outputs are simply recommendations. What you can do is create a report per stage to sum up the findings and bring the client up to speed.

KDD and SEMMA are almost identical in that every stage of KDD directly corresponds to a stage of SEMMA; the CRISP-DM process combines Selection-Preprocessing (KDD) or Sample-Explore (SEMMA) stages into Data Understanding stage. It also incorporates Business Understanding and Deployment stages.

KDD	SEMMA	CRISP-DM
---	---	Business Understanding
Selection	Sample	Data Understanding
Preprocessing	Explore	
Transformation	Modify	Data Preparation
Data Mining	Model	Modeling
Interpretation/Evaluation	Assess	Evaluation
---	---	Deployment

An important difference between CRISP-DM and two other methodologies is that transitions between stages in CRISP-DM can be reversed. This helps a lot when you work with real data — any misstep can be fixed without having to finish the whole cycle if you understand, that chosen target data will not lead to any knowledge.

In my opinion, if you only focus on the data and modeling, you can't see the forest for the trees. In order to step away from simply training models, and start providing solutions to your clients, I would recommend using the CRISP-DM process in your projects. It differs from two other reviewed processes for Data Mining in the addition of Business Understanding and Deployment stages that make the process seem less like *research* and more like *real-life product development*.

Here is where you can find more detailed information on the topic:

- KDD
- SEMMA

- CRISP-DM Guide [*ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.0/ru/ModelerCRISPDM.pdf*]
- CRISP-DM

Written by Nadiia Pyvovar
Proofread by Klym Yamkovyi, Mariana Vechirko