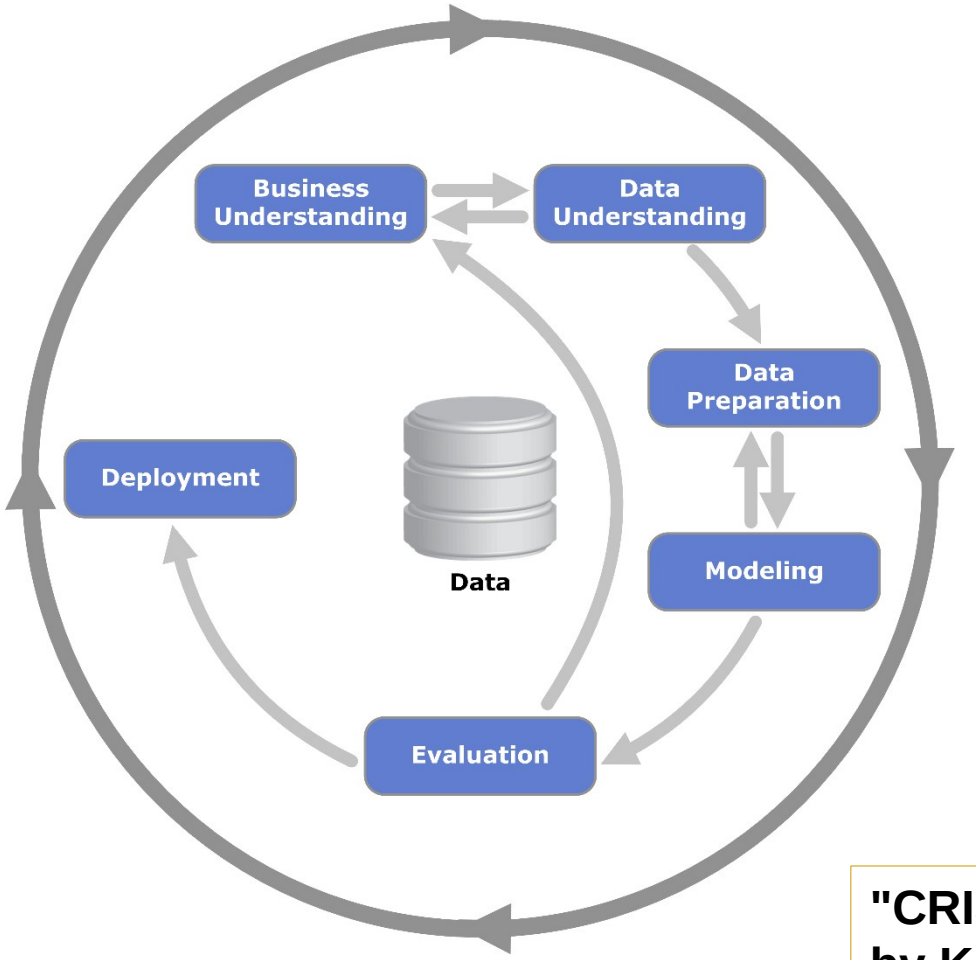
The Data Mining Process

CRISP-DM

- Cross Industry Standard Process for Data Mining.

CRISP-DM



"CRISP-DM Process Diagram" by Kenneth Jensen

Business Understanding

- What is the problem to be solved and in particular what is the usage scenario.
- How can the business problem be formulated as one or more data mining problems.

Data Understanding

- What data is available?
- What is the target variable?
- What are the features.
- Statistics and visualization of feature and target variable values
- Statistics and visualization of relationships between variables.
- Are there missing values?

Data Preparation

- Often data needs to be cleaned before it can be used.
- Deal with missing data.
- Deal with feature values with the wrong type.
- Be careful to avoid leakage. Leakage is where a feature is used in the model but in fact values for hat feature are not available when the model is to be used.
- For example, if the number of page views in a session was used to predict when the session will end!

Modelling

- Determining the model and any parameters.
  - Decision Tree, kNN, Logistic Regression etc.
  - Linear Regression Model (Line of Best Fit)

Evaluation

- Evaluate the model before deployment.
- Normally done by setting aside a test data set which is not used for training.
- Often another iteration is required, going back to the Business Analysis Phase.

Deployment

- The model itself is deployed.
- Sometimes deployment requires that the system need to be re-implemented for performance reasons (R/Python -> Java/Hadoop).
- [And the project is passed over from data scientists to data science engineers.]

CRISP is not a Software Dev. Cycle

- CRISP is based on exploration.
- Outcomes are far less predictable.
- Results may fundamentally change the understanding of the problem.
- More throw away prototypes are required.
- Engineering and deploying a solution too early can be a very costly mistake.