2016 SDSC Summer Institute
Machine Learning

SDSC SAN DIEGO SUPERCOMPUTER CENTER

UC San Diego

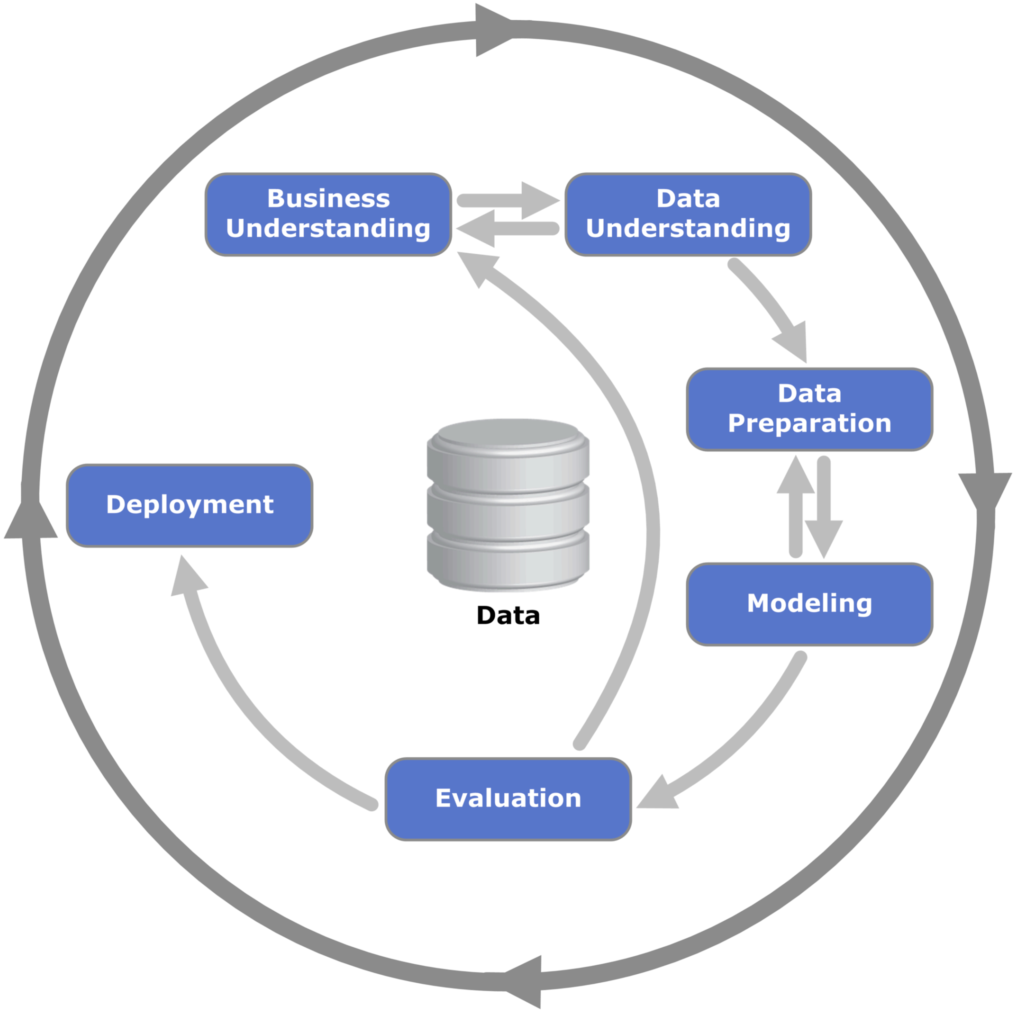
Machine Learning Process

Mai H. Nguyen

CRISP-DM

- **CRoss Industry Standard Process for Data Mining**
  - Process model describing steps in data mining process
- **Phases**
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment

CRISP-DM Diagram



Source: : https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

Phase 1: Business Understanding

- **Define problem or opportunity**
  - What is the problem of interest? Why is it interesting?
- **Assess situation**
  - Resources
  - Requirements, assumptions, and constraints
  - Risks and contingencies; costs and benefits
- **Formulate goals and objectives**
  - Goals and objectives
  - Success criteria
- **Create project plan**
  - Steps to achieve goals

Phase 2: Data Understanding

- **Data Acquisition**
  - Collect available data related to problem.
  - Consider all sources: flat files, databases, sensors, websites, etc.
  - Integrate data from multiple sources

- **Exploratory Data Analysis**
  - Preliminary exploration of data
  - To become familiar with data

Exploratory Data Analysis

Source:*http://www.greenbookblog.org/2013/08/04/50-new-tools-democratizing-data-analysis-visualization/*

- **Goal:**
  - Exploratory data analysis -> data understanding -> informed analysis
  - Also referred to as ‘data profiling’.
- **Techniques:**
  - Summary statistics
    - Mean, frequency, mode, range, variance, standard deviation, etc.
  - Visualization
    - Histograms, scatter plots, line graphs, etc.
  - Look for:
    - Correlations, general trends, outliers, etc.

Phase 3: Data Preparation

- **Goal:**
  - Prepare data to make it suitable for modeling.
  - Also referred to as ‘data preprocessing’, ‘data munging’, ‘data wrangling’.

- **Activities:**
  - Identify and address quality issues
  - Select attributes to use
  - Create data for modeling

Data Quality

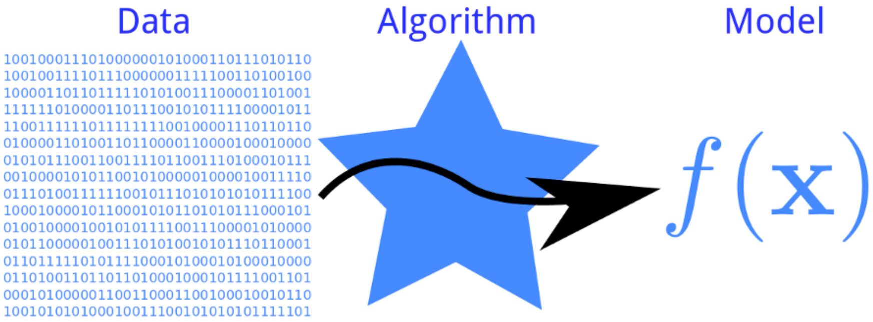
- **Data Quality Issues**
  - Missing Values
  - Duplicate Data
  - Inconsistent Data
  - Noise
  - Outliers

Source: *http://www.datasciencecentral.com/profiles/blogs/5-data-cleansing-tools*

- **Addressing data quality**
  - Also referred to as ‘data cleansing’ or ‘data cleaning’.
- **Important: Garbage in = Garbage out!**
  - Proper data preparation is crucial to machine learning process.

Phase 4: Modeling

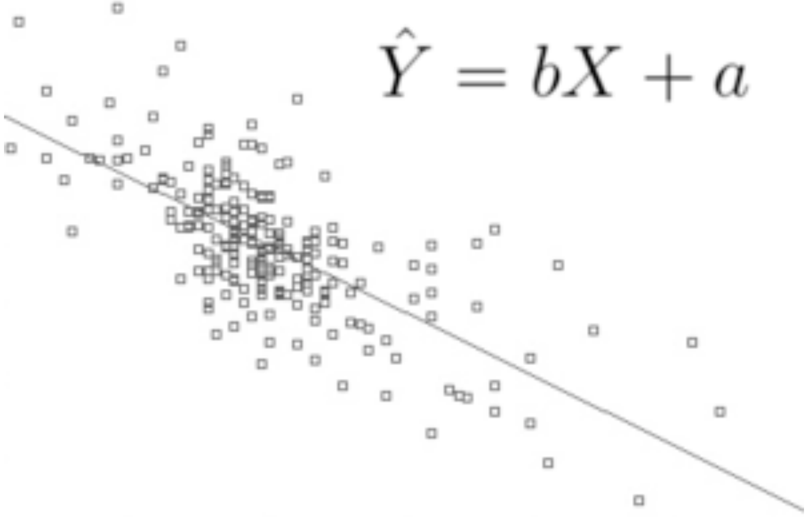
- **Determine type of problem**
  - Classification
  - Regression
  - Cluster analysis
  - Associative analysis
- **Select modeling technique(s) to use**
  - Decision tree
  - Linear regression
  - k-Means
  - etc.



Source: http://phdp.github.io/posts/2013-07-05-dtl.html

Building Model

- **Goal:**
  - Construct model that accurately predicts targets of training data as well as of new data.
  - This is called "generalization".
- **Process:**
  - Adjust model's parameters to minimize error using a learning algorithm.



Source: https://en.wikiversity.org/wiki/Linear_regression

Phase 5: Evaluation

- **Assess model performance.**
  - Determine metrics & methods to assess model results.
    - Accuracy measure
    - Confusion matrix
    - ROC chart
    - etc.
  - Evaluate model results w.r.t. success criteria.
    - Does model's performance meet success criteria?
    - Have all requirements been met?

Evaluation Outcome

- **Determine next steps**
  - Go/No-go decision
  - Go:
    - Proceed to Model Deployment to apply model.
  - No-Go:
    - List of possible actions
      - Different modeling technique?
      - More data cleansing?
      - More data?

Source: *http://www.impactptac.com/?id=10*

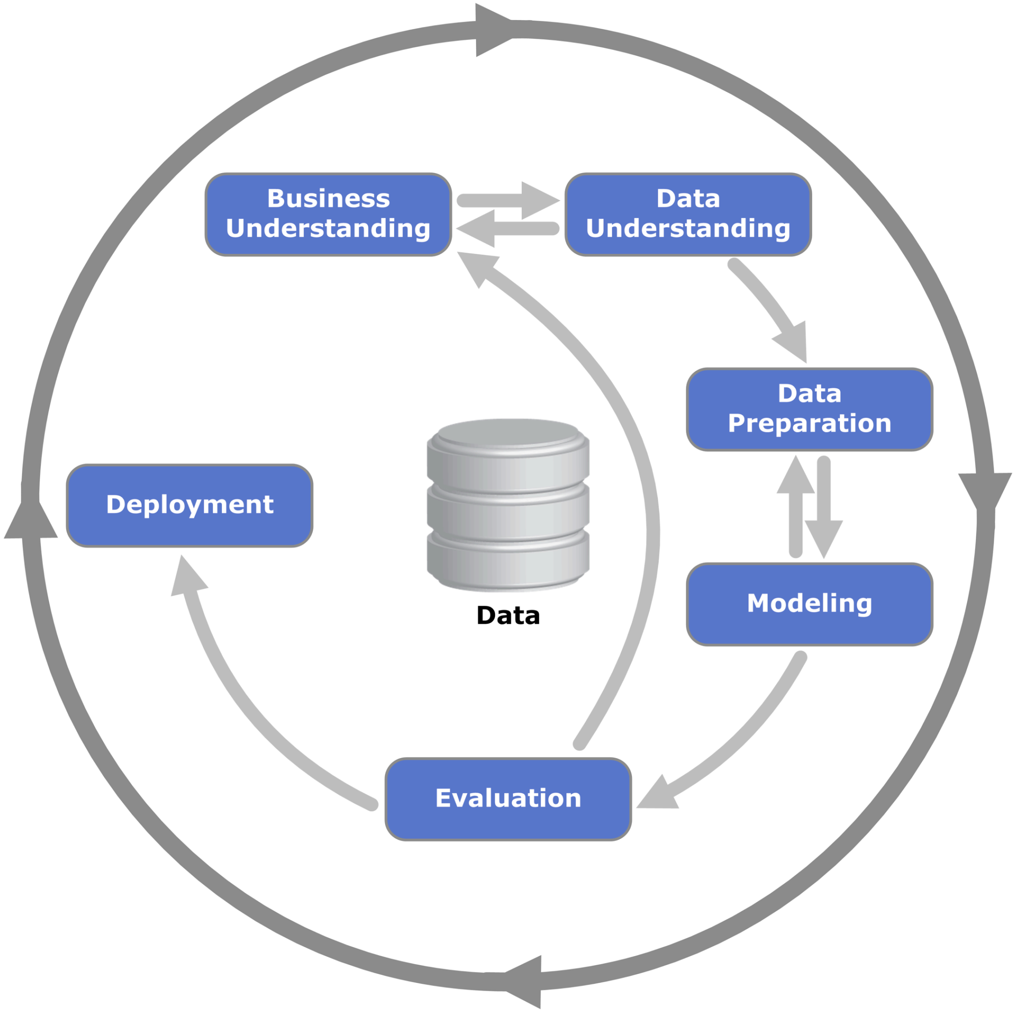
Phase 6: Deployment

- **Produce final report**
  - Summarize findings and recommend uses.
- **Deploy model**
  - Migrate model to production environment.
  - To integrate model into decision-making process.
- **Create plan for model monitoring & maintenance**
  - Monitoring model performance.
  - Plan for updating model.
- **Review and document project**

Model Deployment

- **Approaches**
  - Use data mining tool for scoring
  - Generate model in Java, C, …
  - Generate model in SQL for database use
  - Use cloud-based service (SaaS)
- **PMML**
  - Predictive Model Markup Language
  - Used to share & migrate model between applications and platforms
- **Also referred to as "operationalization".**

CRISP-DM: Iterative Process



DM Process – Key Points

- **CRISP-DM**
  - Process model that describes phases in data mining process
- **Phases**
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment

References

- SPSS. (2000). CRISP-DM 1.0. Retrieved from ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf

Questions?