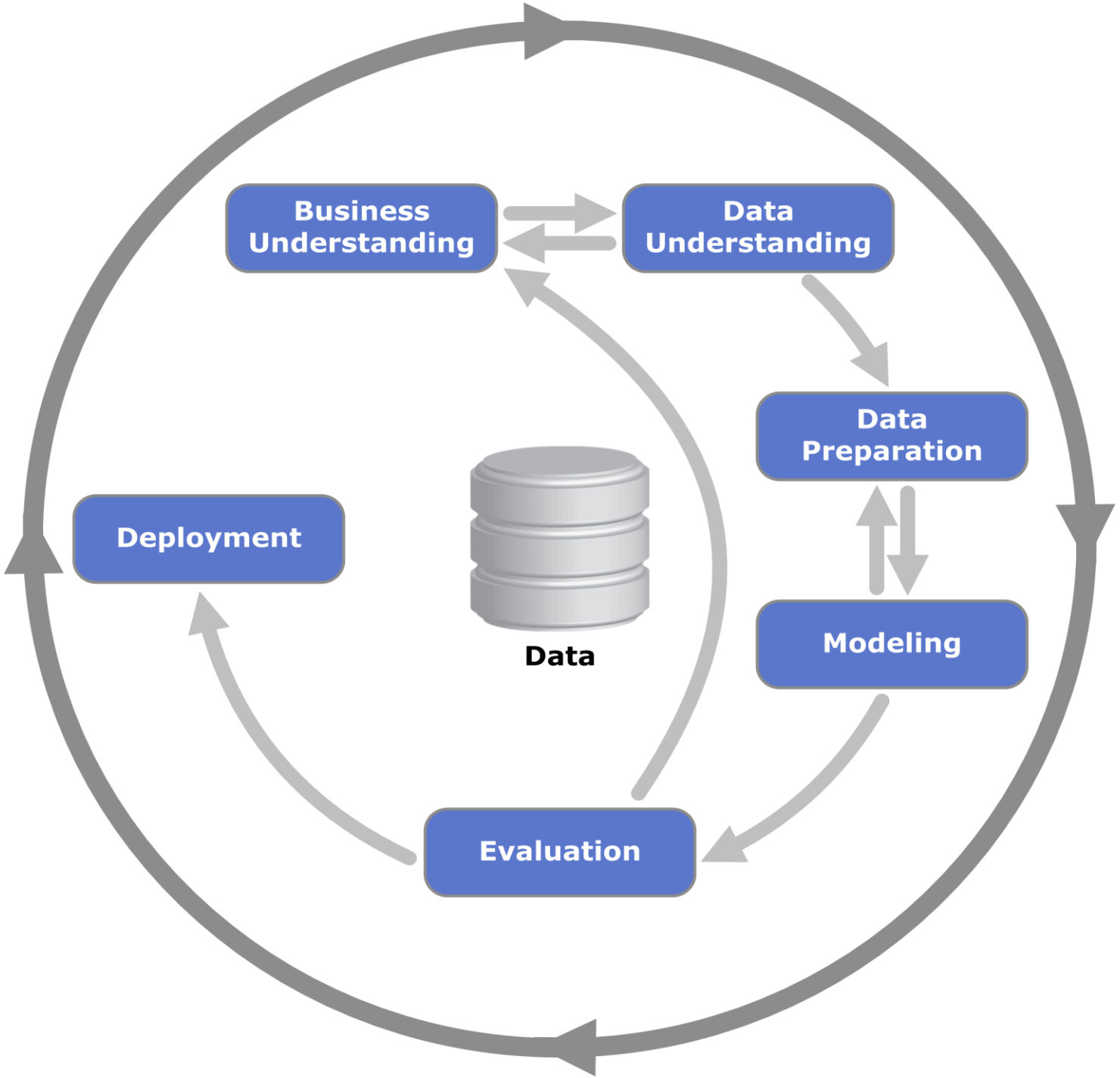
CRISP – DM
CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING

INTRODUCTION

- ***CRISP-DM*** stands for Cross Industry Standard Process for Data Mining. It is a data mining process model that describes commonly used approaches that expert data miners use to tackle problems. Polls conducted in 2002, 2004, and 2007 show that it is the leading methodology used by data miners.
- The only other data mining standard named in these polls was SEMMA.
- However, 3-4 times as many people reported using CRISP-DM. A review and critique of data mining process models in 2009 called the CRISP-DM the "de facto standard for developing data mining and knowledge discovery projects."
- Other reviews of CRISP-DM and data mining process models include Kurgan and Musilek's 2006 review, and Azevedo and Santos' 2008 comparison of CRISP-DM and SEMMA.

CRISP DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

SEMMA -DM

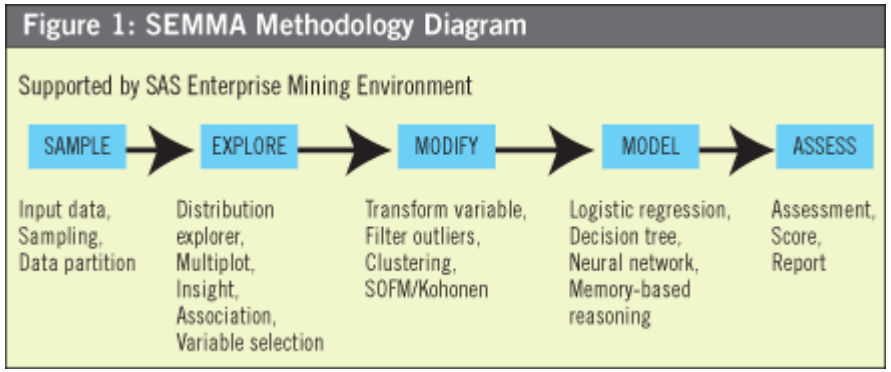


Figure 1: SEMMA Methodology Diagram
Supported by SAS Enterprise Mining Environment
SAMPLE
EXPLORE
MODIFY
MODEL
ASSESS
Input data, Sampling, Data partition
Distribution explorer, Multiplot, Insight, Association, Variable selection
Transform variable, Filter outliers, Clustering, SOFM/Kohonen
Logistic regression, Decision tree, Neural network, Memory-based reasoning
Assessment, Score, Report

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Assess Situation** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* **Determine Data Mining Goals** *Data Mining Goals* *Data Mining Success Criteria* **Produce Project Plan** *Project Plan* *Initial Assessment of Tools and Techniques*	**Collect Initial Data** *Initial Data Collection Report* **Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	**Select Data** *Rationale for Inclusion/ Exclusion* **Clean Data** *Data Cleaning Report* **Construct Data** *Derived Attributes* *Generated Records* **Integrate Data** *Merged Data* **Format Data** *Reformatted Data* *Dataset* *Dataset Description*	**Select Modeling Techniques** *Modeling Technique* *Modeling Assumptions* **Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Descriptions* **Assess Model** *Model Assessment* *Revised Parameter Settings*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models* **Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Deployment** *Deployment Plan* **Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation* **Review Project** *Experience Documentation*

CRISP – DM

1. **Pemahaman Bisnis(*Business Understanding*)**
   - Merupakan tahap awal yaitu pemahaman penelitian, penentuan tujuan dan rumusan masalah *data mining*.
2. **Pemahaman Data(*Data Understanding*)**
   - Dalam tahap ini dilakukan pengumpulan data, mengenali lebih lanjut data yang akan digunakan.
3. **Pengolahan Data(*Data Preparation*)**
   - Tahap ini adalah pekerjaan berat yang perlu dilaksanakan secara intensif. Memilih kasus atau variable yang ingin dianalisis, melakukan perubahan pada beberapa variable jika diperlukan sehingga data siap untuk dimodelkan.
4. **Pemodelan(*Modeling*)**
   - Memilih teknik pemodelan yang sesuai dan sesuaikan aturan model untuk hasil yang maksimal. Dapat kembali ke tahap pengolahan untuk menjadikan data ke dalam bentuk yang sesuai dengan model tertentu.
5. **Evaluasi (*Evaluation*)**
   - Mengevaluasi satu atau model yang digunakan dan menetapkan apakah terdapat model yang memenuhi tujuan pada tahap awal. Kemudian menentukan apakah ada permasalahan yang tidak dapat tertangani dengan baik serta mengambil keputusan hasil penelitian.
6. **Penyebaran (*Deployment*)**
   - Menggunakan model yang dihasilkan seperti pembuatan laporan atau penerapan proses *data mining* pada departemen lain.