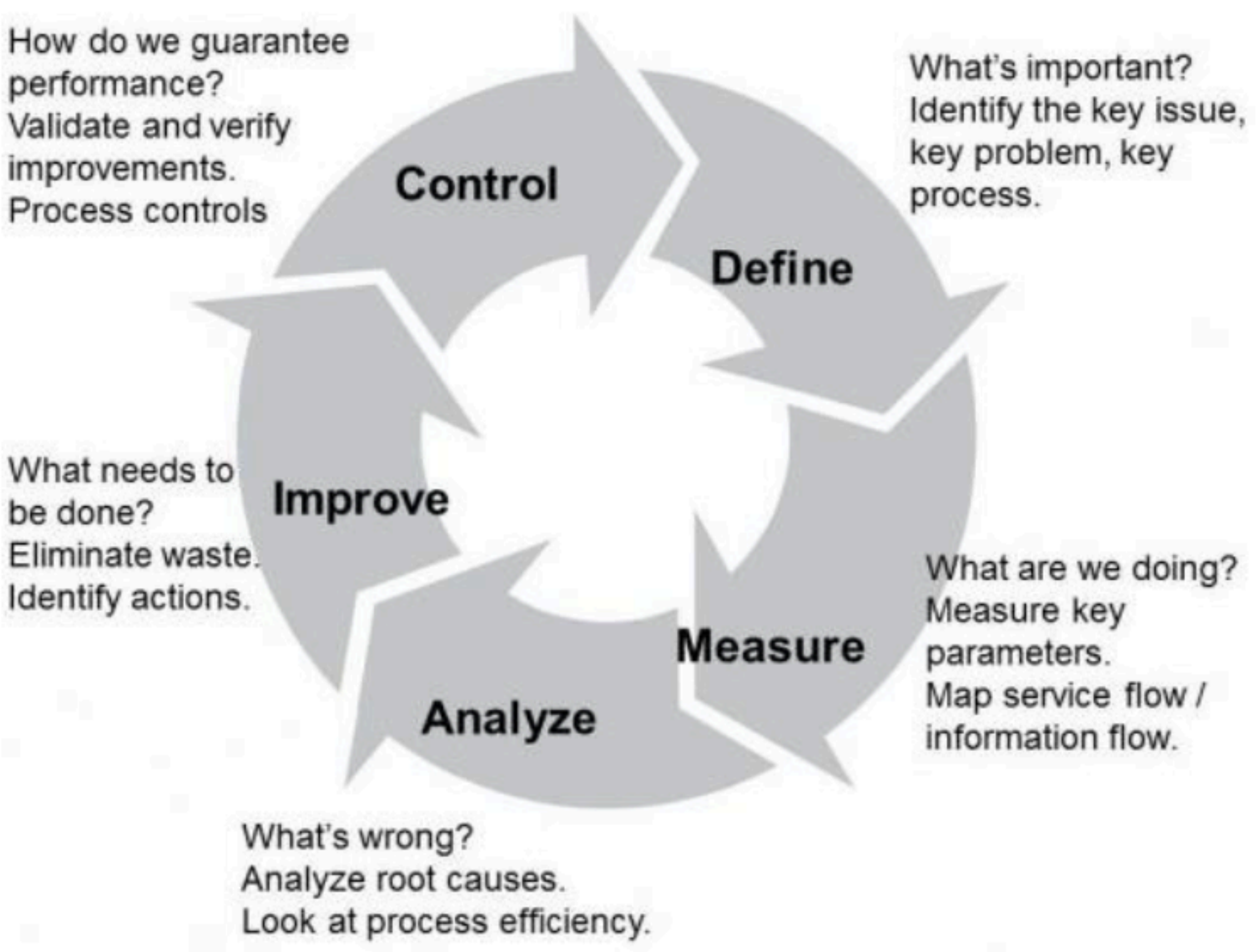
CS490DSC Data Science Capstone
CRISP-DM Methodology

Jean Honorio
Purdue University

Methodology

- What is a methodology?
  - A sequence of phases/steps when working on a project
- Why following a methodology?
  - To avoid obvious mistakes, e.g., misunderstanding the business/ user needs, misunderstanding the data, misunderstanding how we want the model to generalize
- Several methodologies
  - Six Sigma DMAIC (Define, Measure, Analyze, Improve, Control)
  - KDD (Knowledge Discovery in Databases)
  - SEMMA (Sample, Explore, Modify, Model, Assess)
  - TDSP (Team Data Science Process)
  - CRISP-DM (CRoss-Industry Standard Process for Data Mining)

Six Sigma DMAIC



Define

What's important?
Identify the key issue, key problem, key process.

Measure

What are we doing?
Measure key parameters.
Map service flow / information flow.

Analyze

What's wrong?
Analyze root causes.
Look at process efficiency.

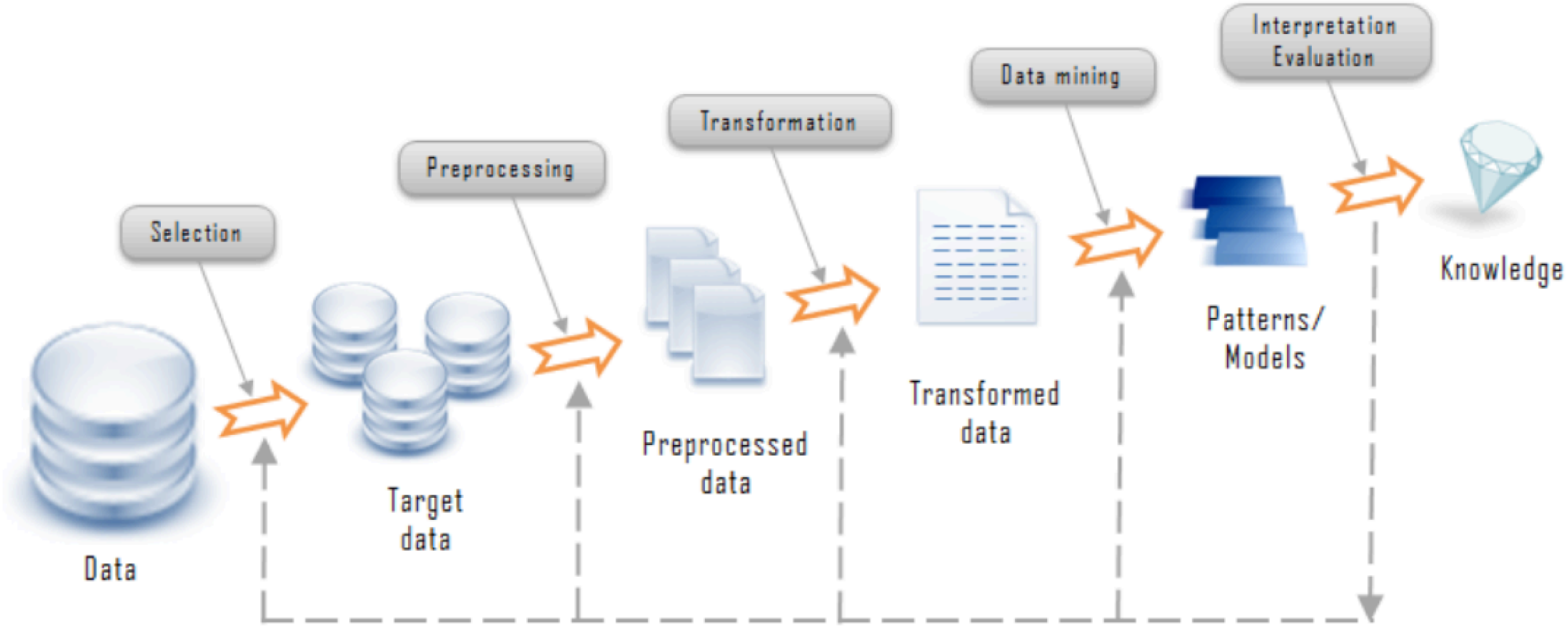
Improve

What needs to be done?
Eliminate waste.
Identify actions.

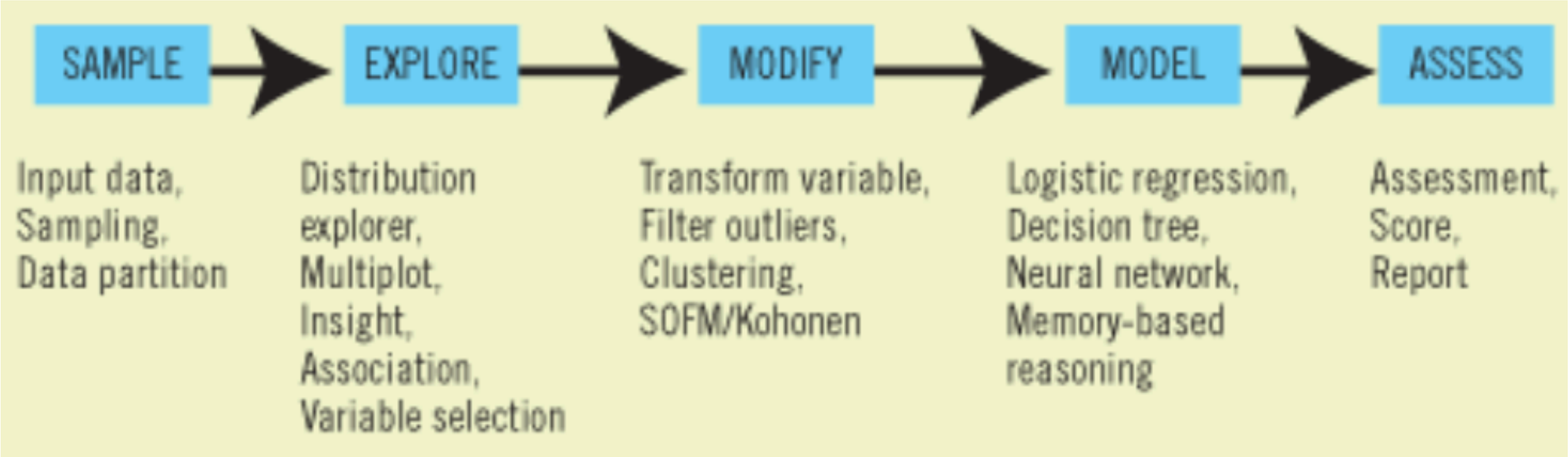
Control

How do we guarantee performance?
Validate and verify improvements.
Process controls

KDD


Selection
Preprocessing
Transformation
Data mining
Interpretation
Evaluation
Data
Target
data
Preprocessed
data
Transformed
data
Patterns/
Models
Knowledge

SEMMA



SAMPLE

Input data,
Sampling,
Data partition

EXPLORE

Distribution explorer,
Multiplot,
Insight,
Association,
Variable selection

MODIFY

Transform variable,
Filter outliers,
Clustering,
SOFM/Kohonen

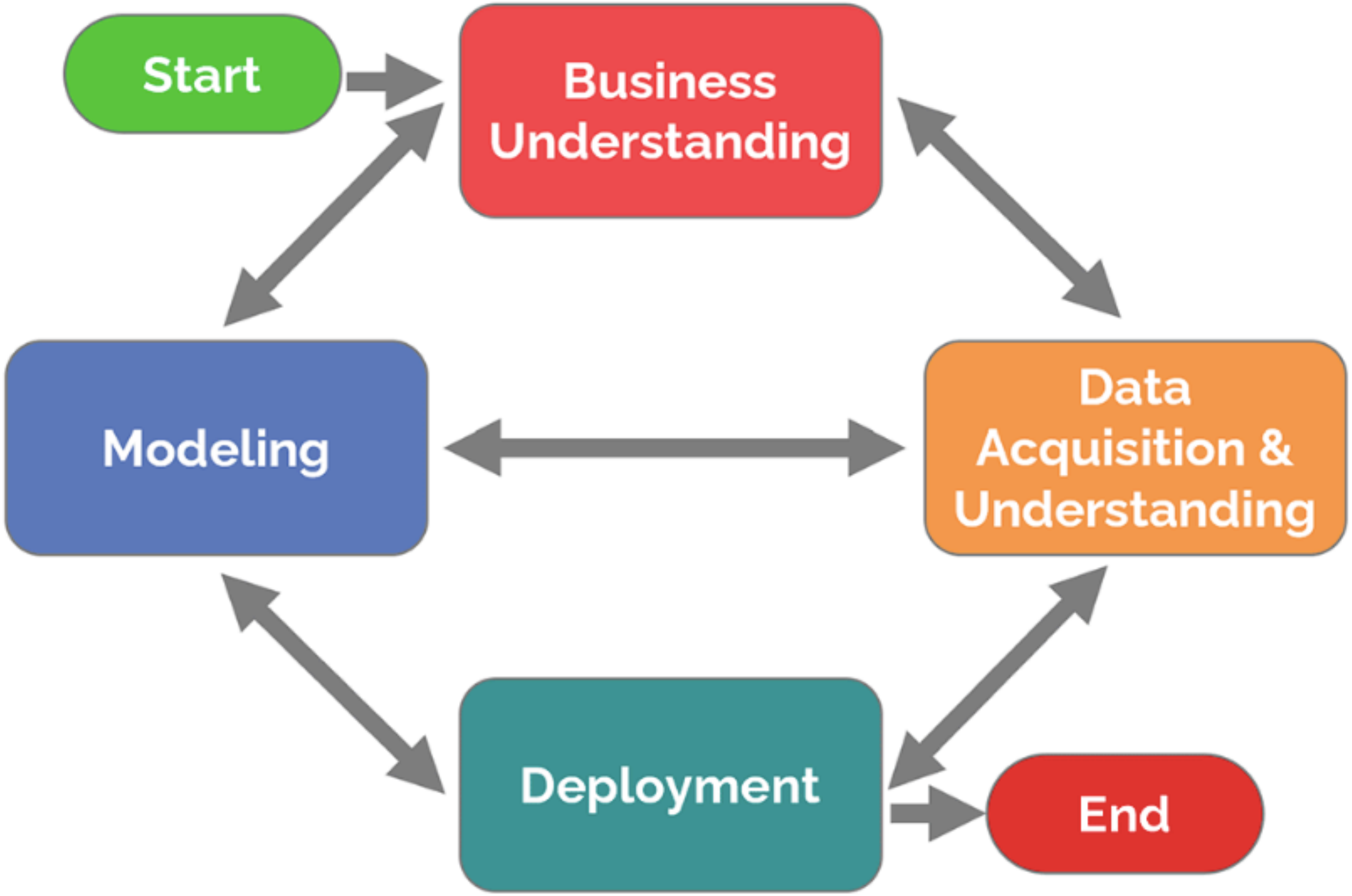
MODEL

Logistic regression,
Decision tree,
Neural network,
Memory-based reasoning

ASSESS

Assessment,
Score,
Report

TDSP



Start

Business Understanding

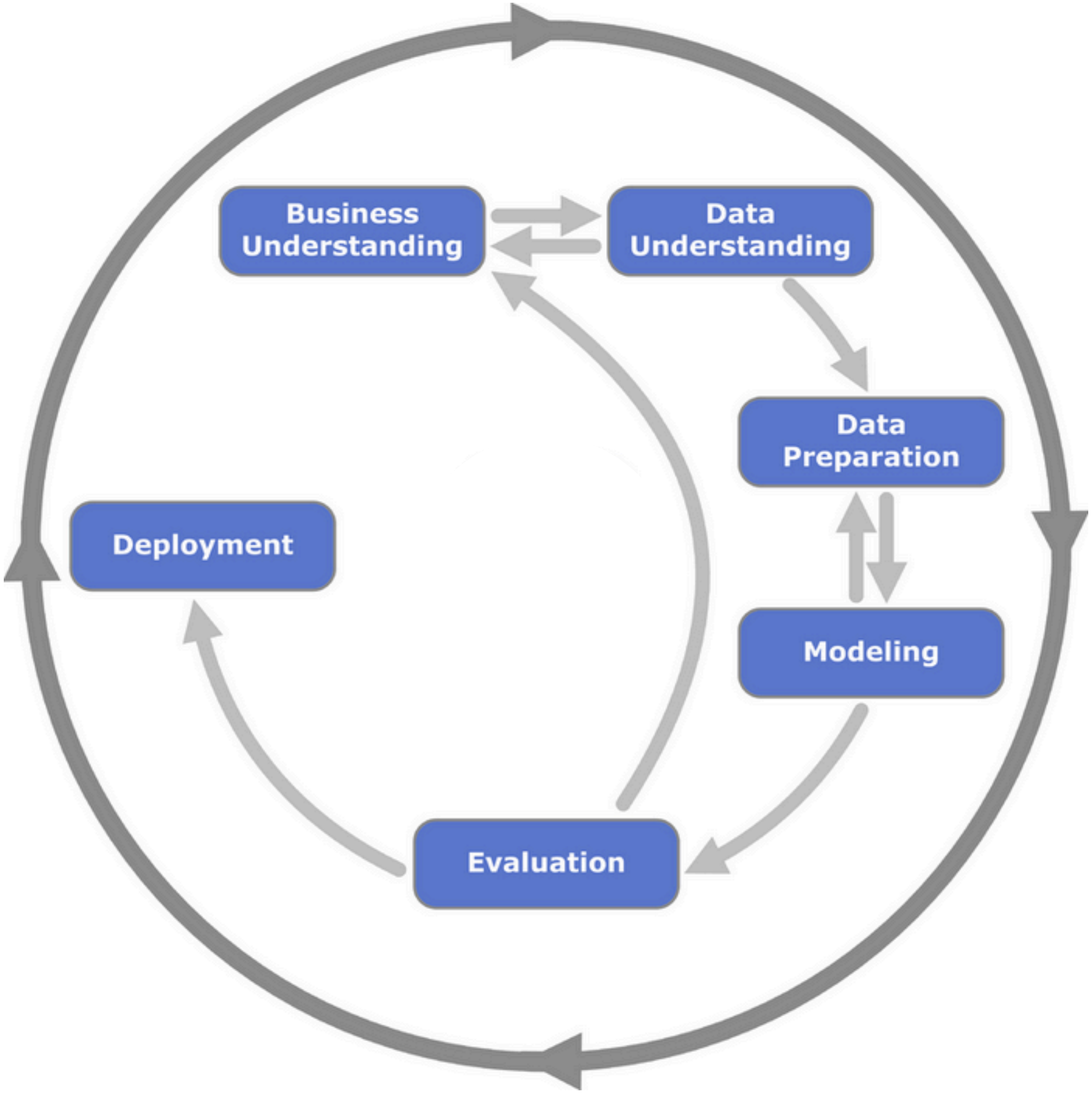
Modeling

Data Acquisition & Understanding

Deployment

End

CRISP-DM



- The sequence of the 6 phases is not rigid
  - Moving back and forth between different phases is possible
- The outer circle symbolizes the cyclical nature of data mining itself
  - e.g., the lessons learned during the process can trigger new business questions

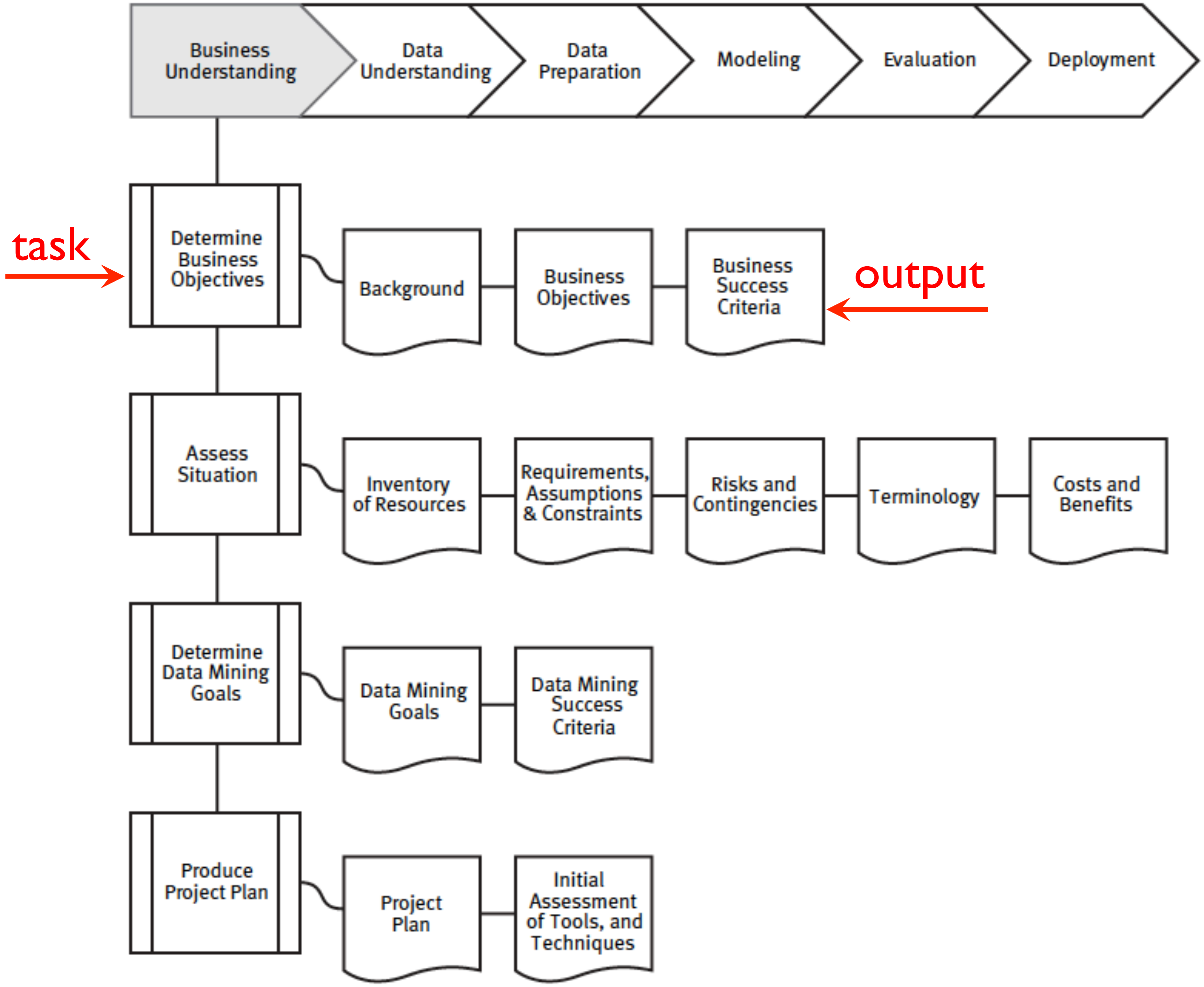
CRISP-DM

- CRoss-Industry Standard Process for Data Mining
  - non-proprietary and freely available
  - industry/application-neutral
  - tool-neutral
- Conceived in 1996, funded in 1997 by the European Commission, document released on 2000
  - DaimlerChrysler, SPSS, NCR, OHRA
- CRISP-DM Special Interest Group has more than 200 members
- Most used methodology
  - 49% in 2020, 43% in 2014, 42% in 2007
  - https://www.datascience-pm.com/crisp-dm-still-most-popular/
  - https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

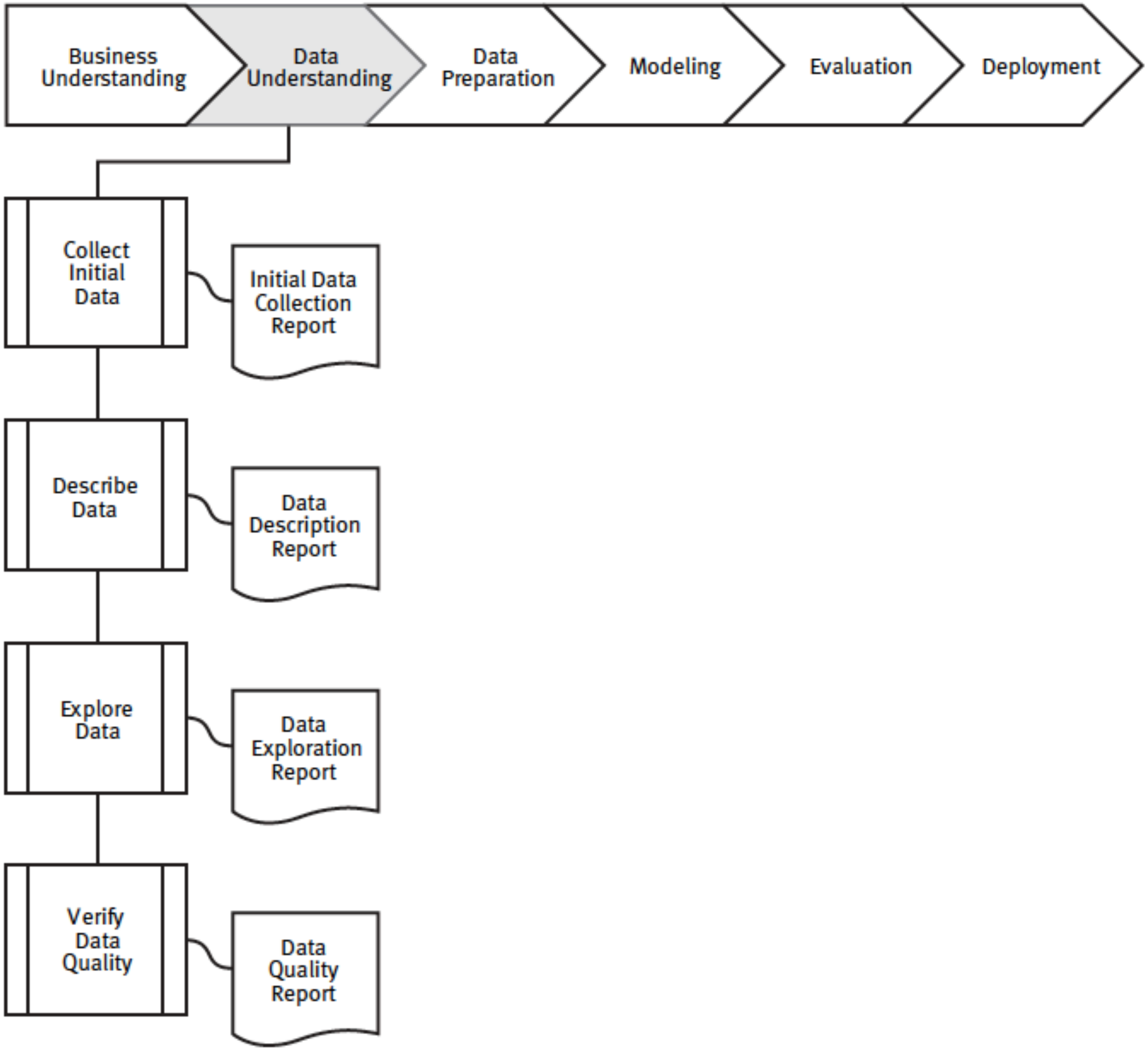
CRISP-DM

- CRISP-DM has 6 phases
  - **Business understanding**: understand business objectives, define data mining problem
  - **Data understanding**: familiarize with data, identify data quality issues
  - **Data preparation**: select, transform, clean data
  - **Modeling**: run the data mining tools
  - **Evaluation**: results meet business objectives?
  - **Deployment**: put models in practice
- Each phase has a set of tasks and outputs
  - We will provide a Word document to be filled for each phase
  - We will follow a case study to make things clear

Phase 1: Business understanding



Phase 2: Data understanding



Business Understanding

Data Understanding

Data Preparation

Modeling

Evaluation

Deployment

Collect Initial Data

Initial Data Collection Report

Describe Data

Data Description Report

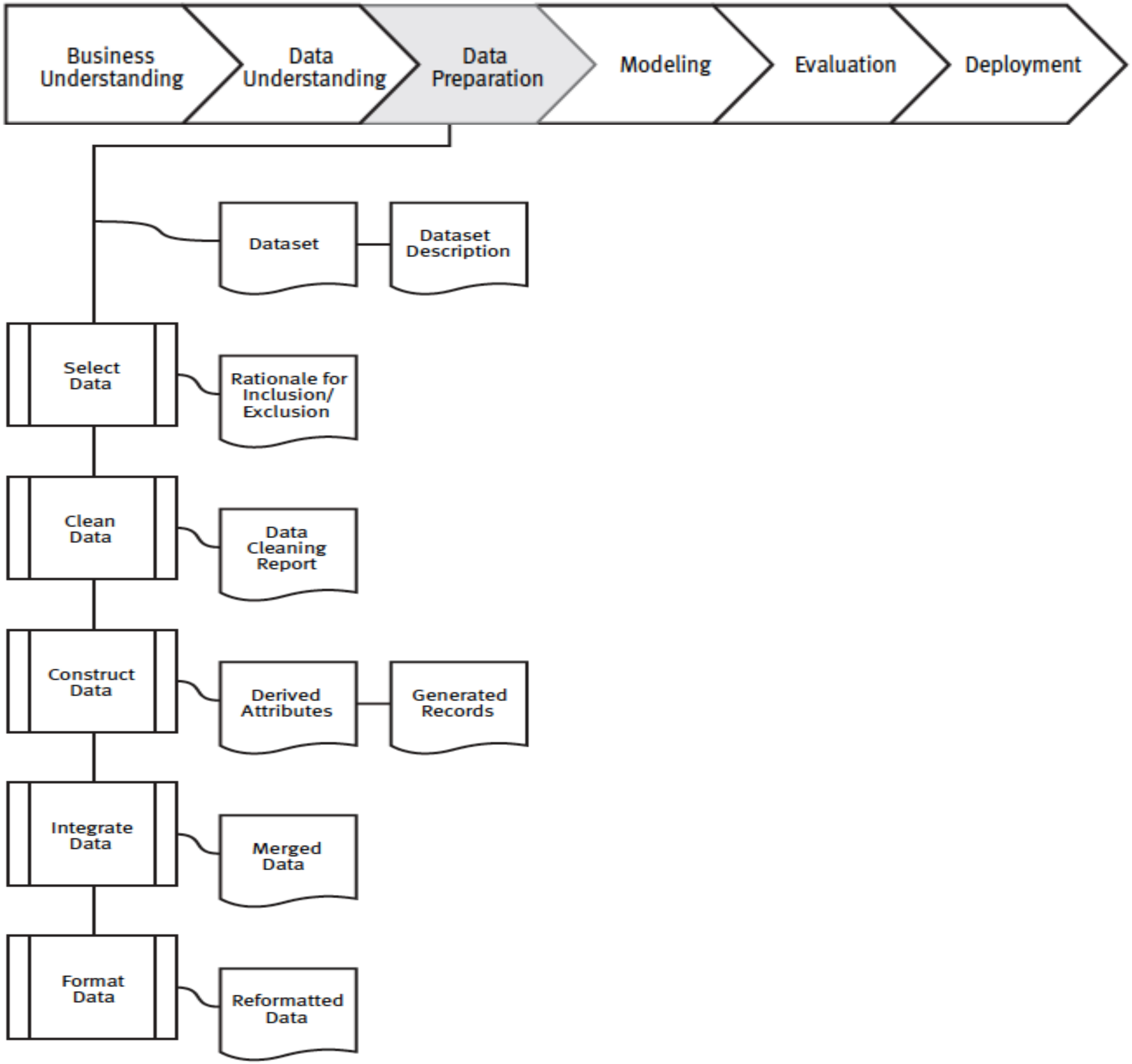
Explore Data

Data Exploration Report

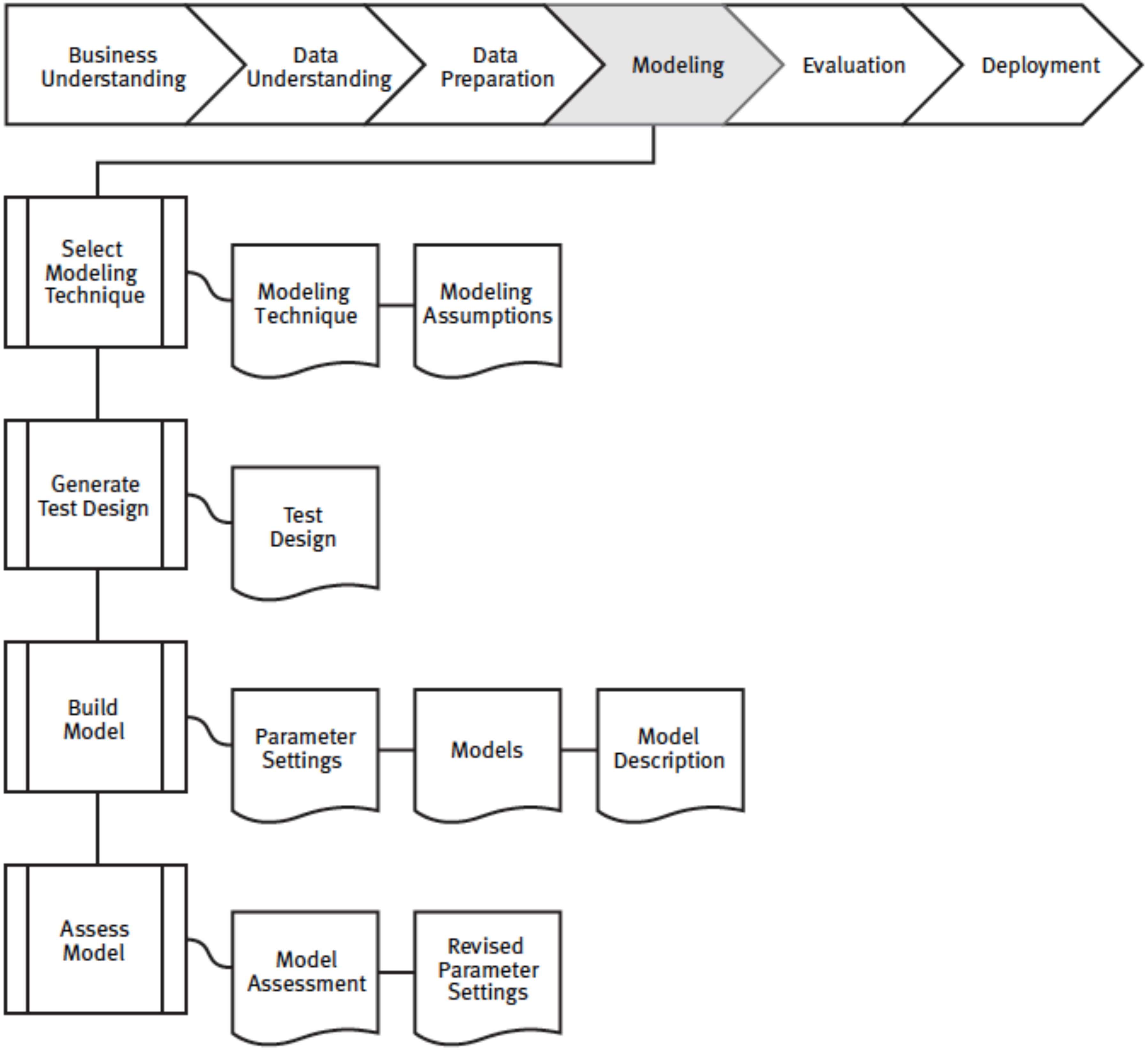
Verify Data Quality

Data Quality Report

Phase 3: Data preparation



Phase 4: Modeling



Business Understanding

Data Understanding

Data Preparation

Modeling

Evaluation

Deployment

Select Modeling Technique

Modeling Technique

Modeling Assumptions

Generate Test Design

Test Design

Build Model

Parameter Settings

Models

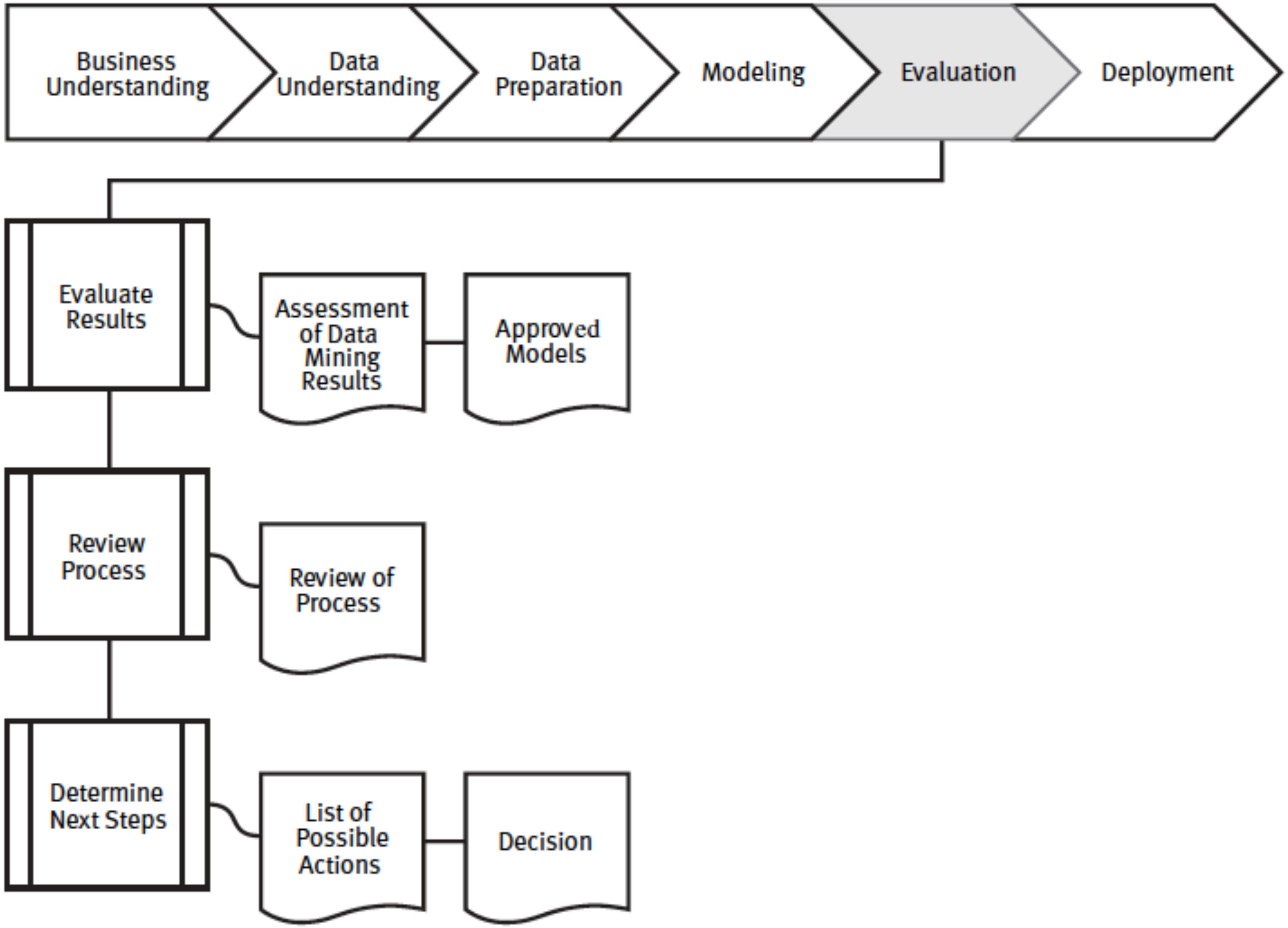
Model Description

Assess Model

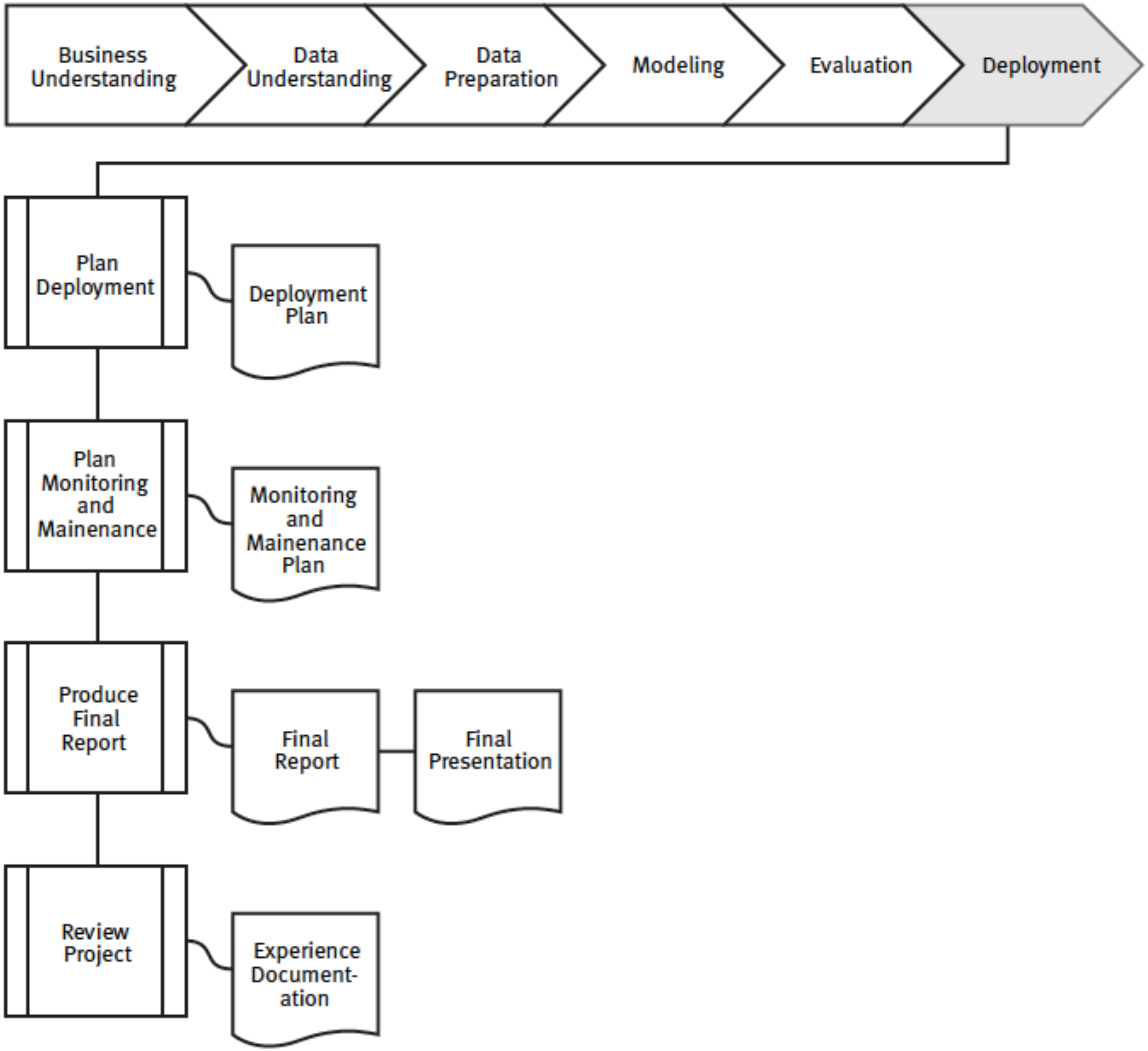
Model Assessment

Revised Parameter Settings

Phase 5: Evaluation



Phase 6: Deployment


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Plan Deployment
Deployment Plan
Plan Monitoring and Mainenance
Monitoring and Mainenance Plan
Produce Final Report
Final Report
Final Presentation
Review Project
Experience Document-ation