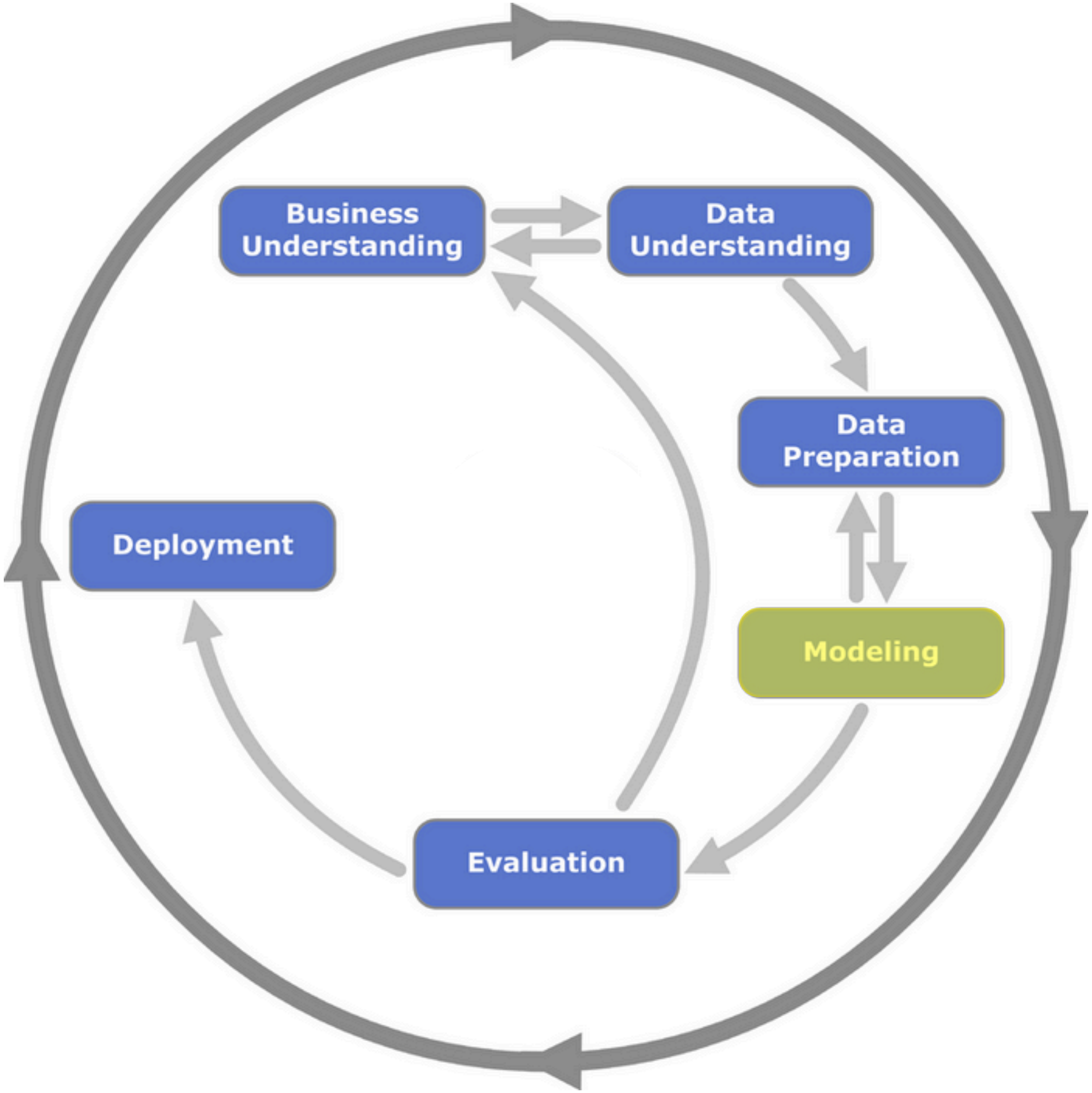
CS490DSC Data Science Capstone Modeling

Jean Honorio
Purdue University

Important

- Please read this together with the case study
- The case study will discuss a fictitious health insurance company called the Amazing Health Network

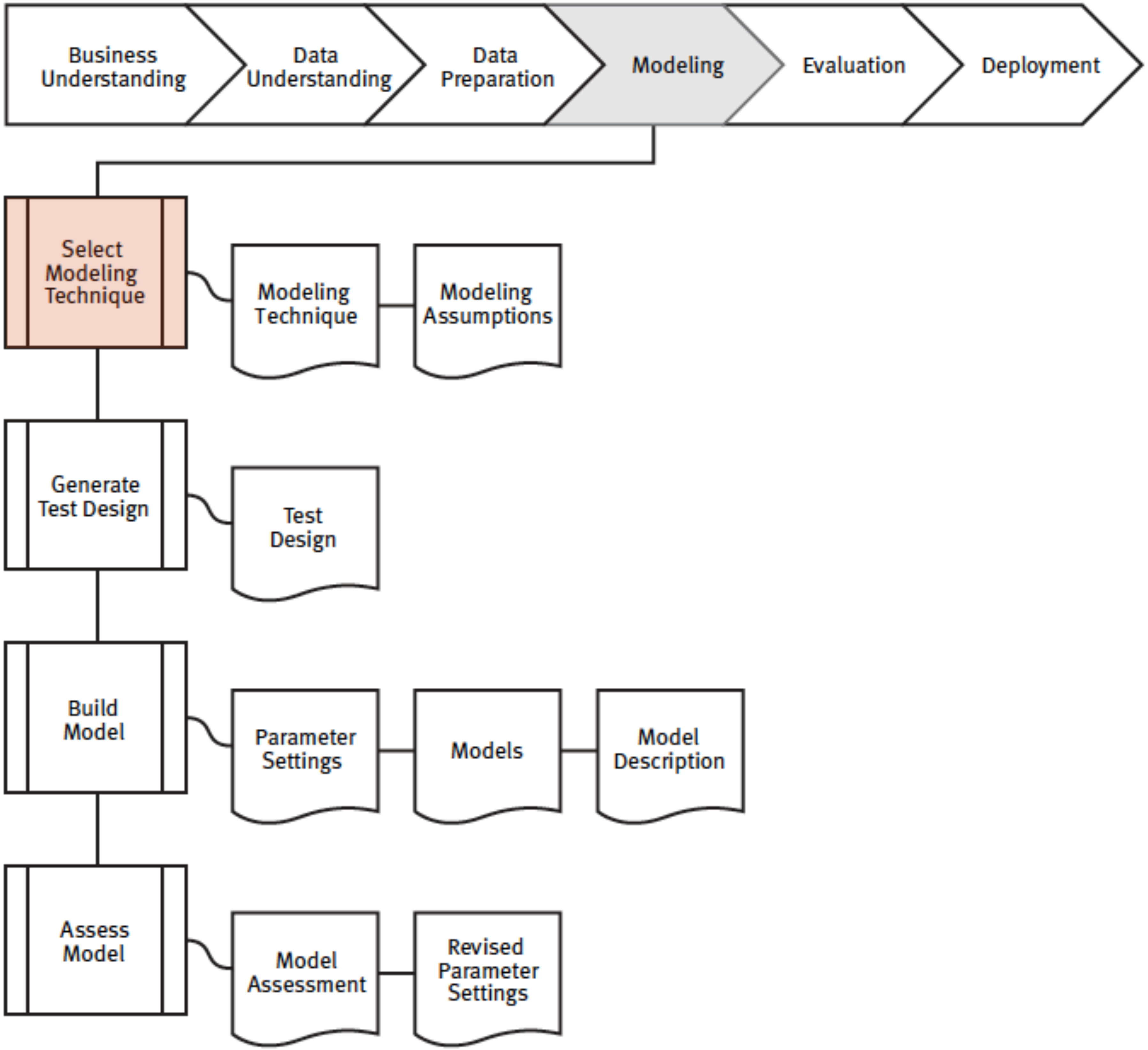
CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment

Phase 4: Modeling

- In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values
  - Usually, there are several techniques for the same data mining problem type
- Some techniques have specific requirements on the form of data
  - Going back to the data preparation phase is often necessary

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

1. Select modeling technique

- As the first step in modeling, select the actual modeling technique that is to be used
- If multiple techniques are applied, perform this task separately for each technique

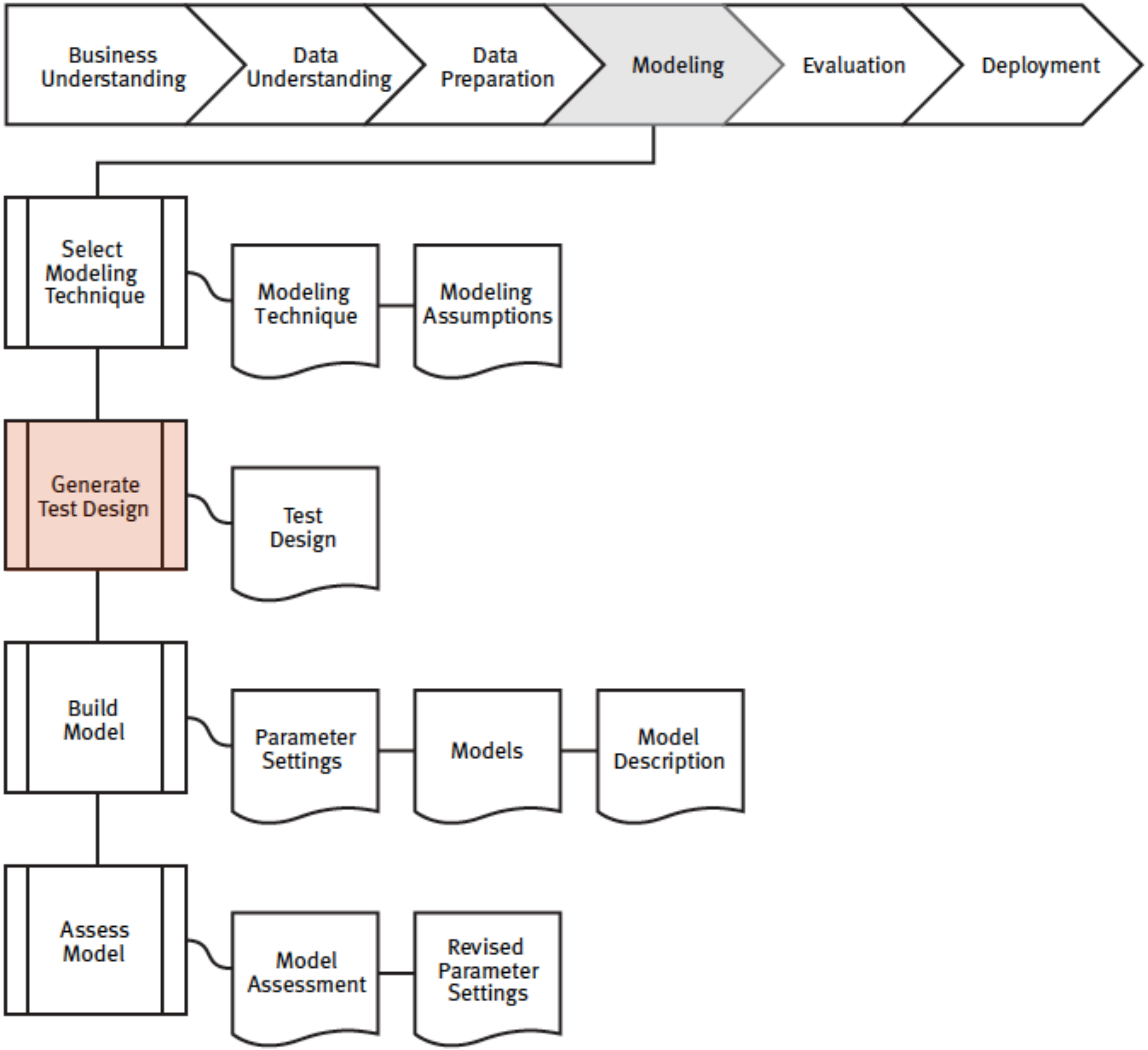
1.1. Modeling technique

- Document the actual modeling technique that is to be used
- Although you may have already selected a tool during the Business Understanding phase, this task refers to the **specific modeling technique**, e.g.,
    - C4.5 algorithm for decision trees
    - mini-batch gradient descent with Gaussian initialization for convolutional neural networks

1.2. Modeling assumptions

- Many modeling techniques make specific assumptions about the data, for example
  - all numeric attributes have a similar scale
  - no missing values allowed
  - class attribute is categorical but not ordinal
    - Ordinal: movie rating (5>4>3>2>1)
    - Ordinal: opinion (strongly agree > agree > neutral > disagree > strongly disagree)
    - Not ordinal: object type (table, chair, car, bike)
- **Record any such assumptions made**

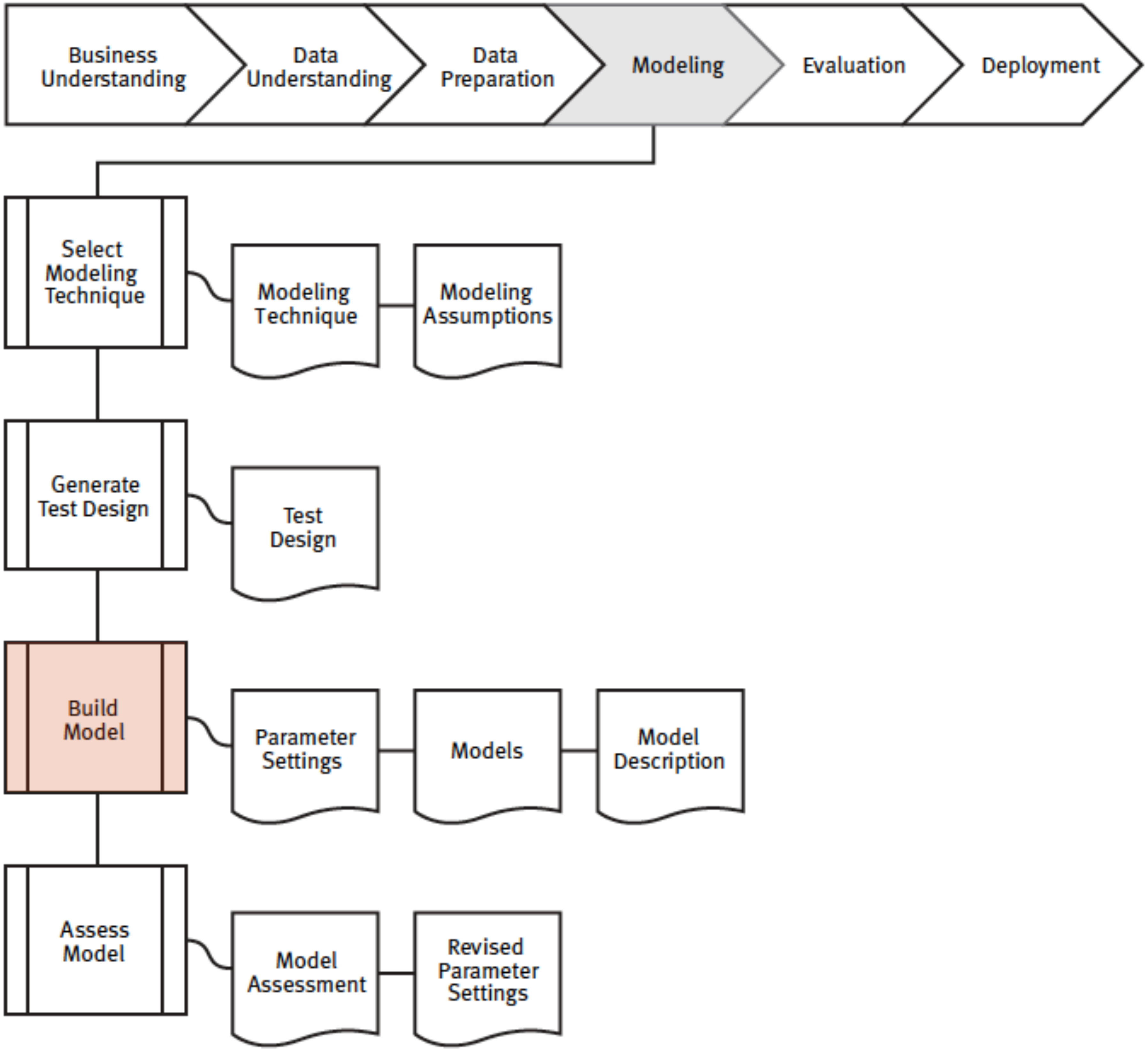
Phase 4: Modeling



2. Generate test design

- Before we actually build a model, we need to generate a procedure or mechanism to test the model's quality and validity
- For example, in supervised data mining tasks such as classification, it is common to use error rates as quality measures for data mining models
    - We typically separate the dataset into train and test sets, build the model on the train set, and estimate its quality on the separate test set
- Describe the intended plan for training, testing, and evaluating the models
- A primary component of the plan is determining how to divide the available dataset into training, test, and validation datasets

Phase 4: Modeling



3. Build model

- **Run the modeling tool on the prepared dataset to create one or more models**

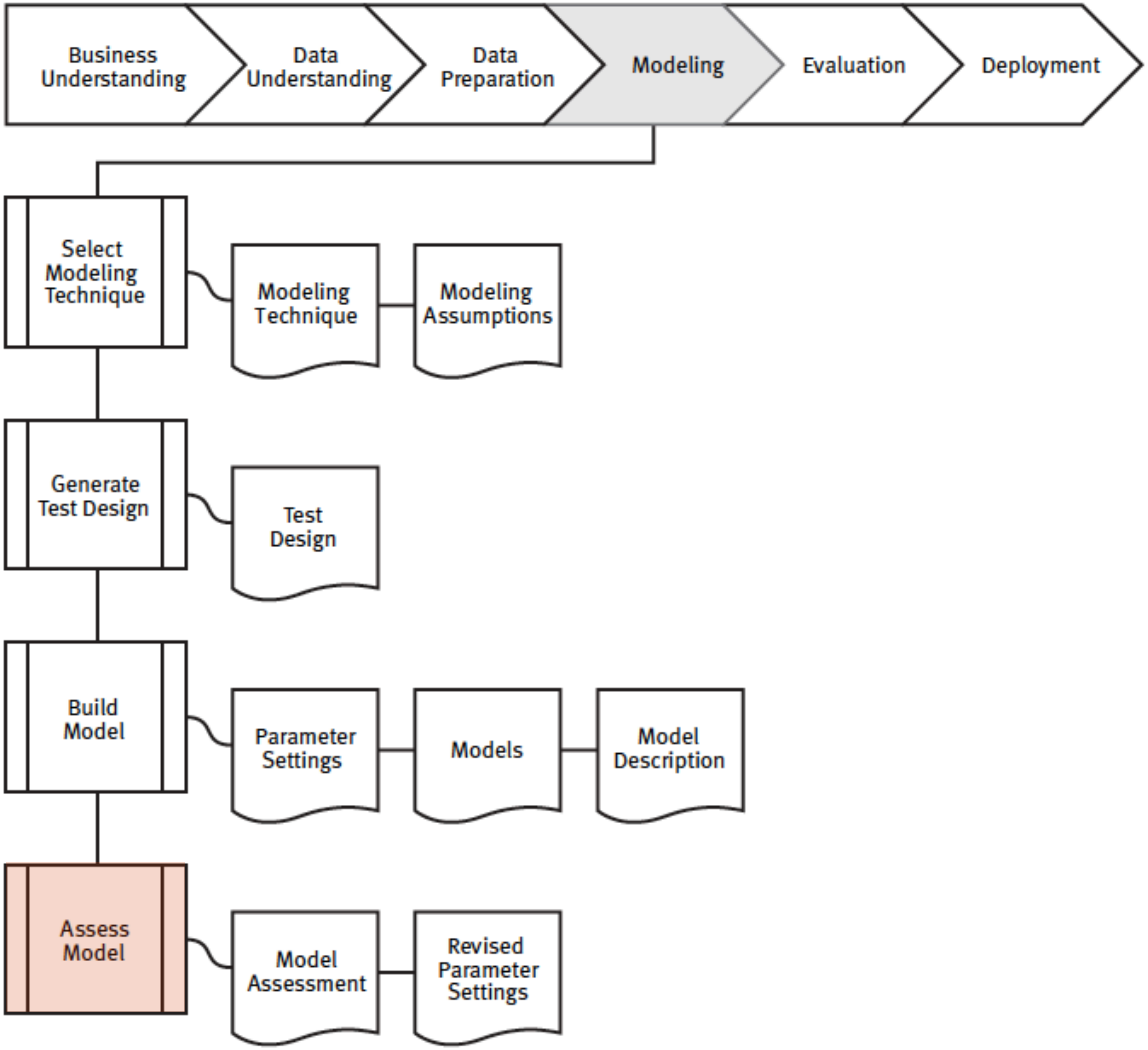
3.1. Parameter settings

- With any modeling tool, there are often a large number of parameters that can be adjusted
- List the parameters and their chosen values, along with the rationale for the choice of parameter settings. For instance
  - the regularization parameter C for support vector machines
  - k for k-nearest neighbors
  - Gini threshold for CART decision trees

3.2. Models
3.3. Model description

- These are the actual models produced by the modeling tool, not a report
- Describe the resulting models
- Report on the interpretation of the models and document any difficulties encountered with their meanings
- For instance
  - For a linear classifier (e.g., logistic regression), the magnitude of the weights associated with each feature, give some measure of how important each feature is (if features are normalized)
  - For linear support vector machines, you can see which samples are the “support vectors”. Those are the samples most difficult to classify. For instance, in a horses-versus-giraffes classification task, those will be the horses that look like giraffes, and the giraffes that look like horses

Phase 4: Modeling



4. Assess model

- The data mining engineer (DME) interprets the models according to the domain knowledge, the data mining success criteria, and the desired test design
- The DME judges the success of the application of modeling and discovery techniques technically
  - The DME contacts business analysts and domain experts later in order to discuss the data mining results in the business context
  - This task only considers models, whereas the evaluation phase also takes into account all other results that were produced in the course of the project

4. Assess model

- The data mining engineer (DME) tries to rank the models
- The DME assesses the models according to the evaluation criteria
- As much as possible, the DME also takes into account business objectives and business success criteria
- In most data mining projects, the DME
  - applies a single technique more than once, or
  - generates data mining results with several different techniques
- In this task, the DME also compares all results according to the evaluation criteria

4.1. Model assessment

- Summarize results of this task:
  - list qualities of generated models (e.g., in terms of accuracy), and
  - rank their quality in relation to each other

4.2. Revised parameter settings

- According to the model assessment, revise parameter settings and tune them for the next run in the Build Model task
- Iterate model building and assessment until you strongly believe that you have found the best model(s)
- Document all such revisions and assessments