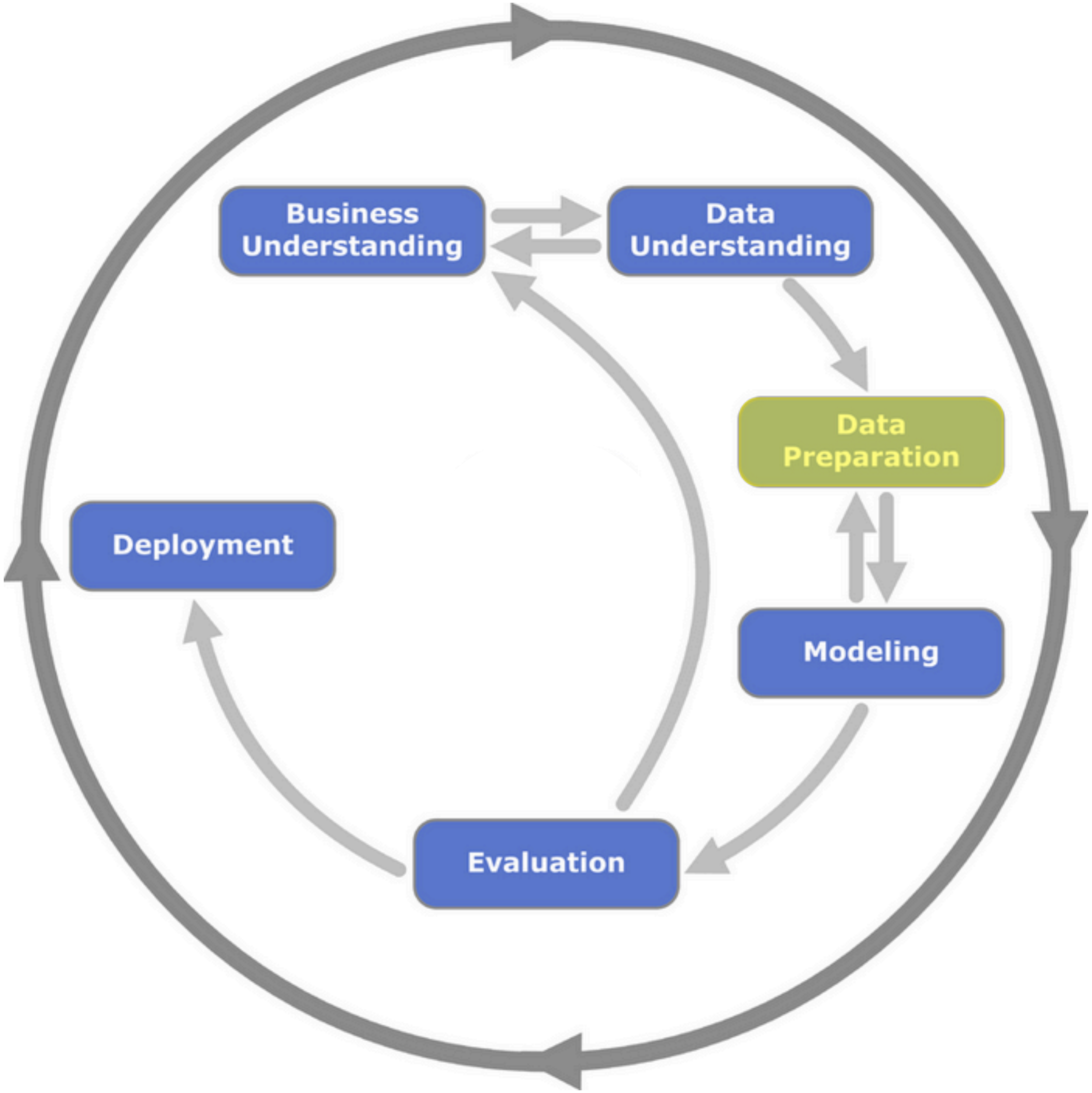
CS490DSC Data Science Capstone
Data Preparation

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

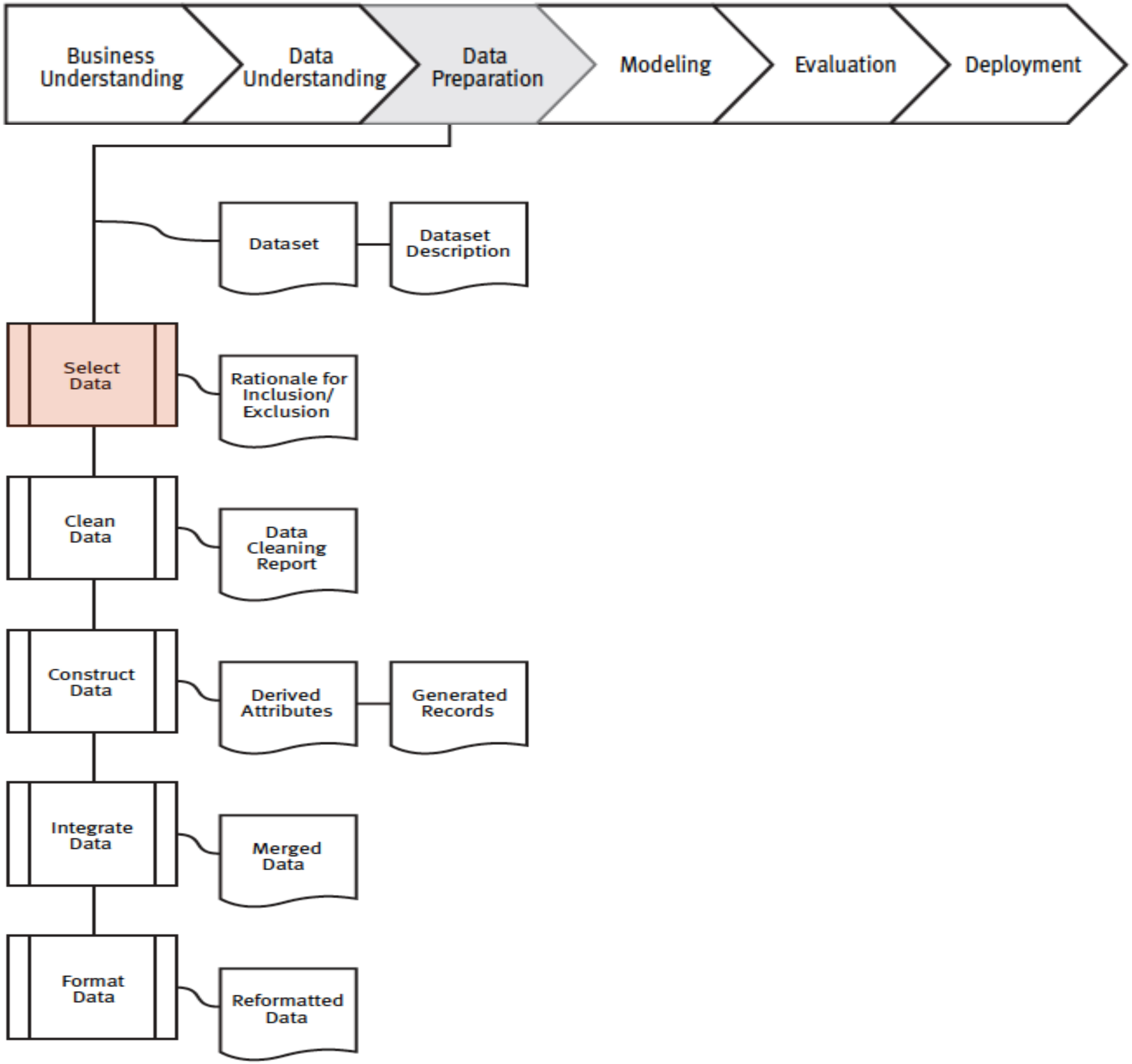
CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment

Phase 3: Data preparation

- This phase covers all activities needed to construct the final dataset (for the next phases) from the initial raw data
- Data preparation tasks are likely to be performed multiple times and not in any prescribed order
- Tasks include
  - table, record, and attribute selection
  - transformation of data
  - cleaning of data

Phase 3: Data preparation



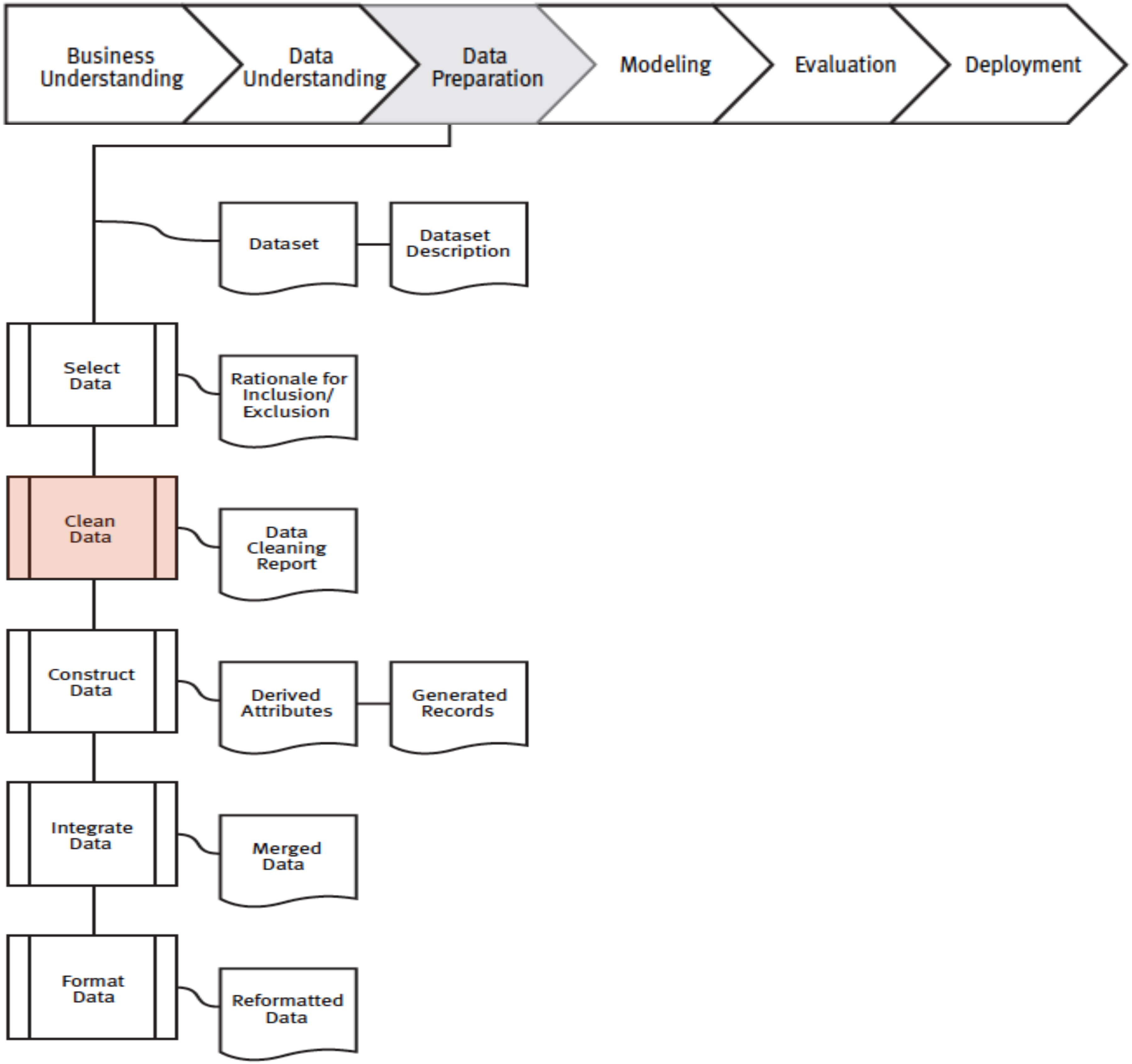
1. Select data

- Decide on the data to be used for analysis
- Criteria include relevance to the data mining goals, quality, and technical constraints such as limits on data volume or data types
- Note that data selection covers
  - selection of attributes (columns) in a table
  - selection of records (rows) in a table

1.1 Rationale for inclusion/exclusion

- List
  - the data to be included/excluded
  - the reasons for these decisions

Phase 3: Data preparation



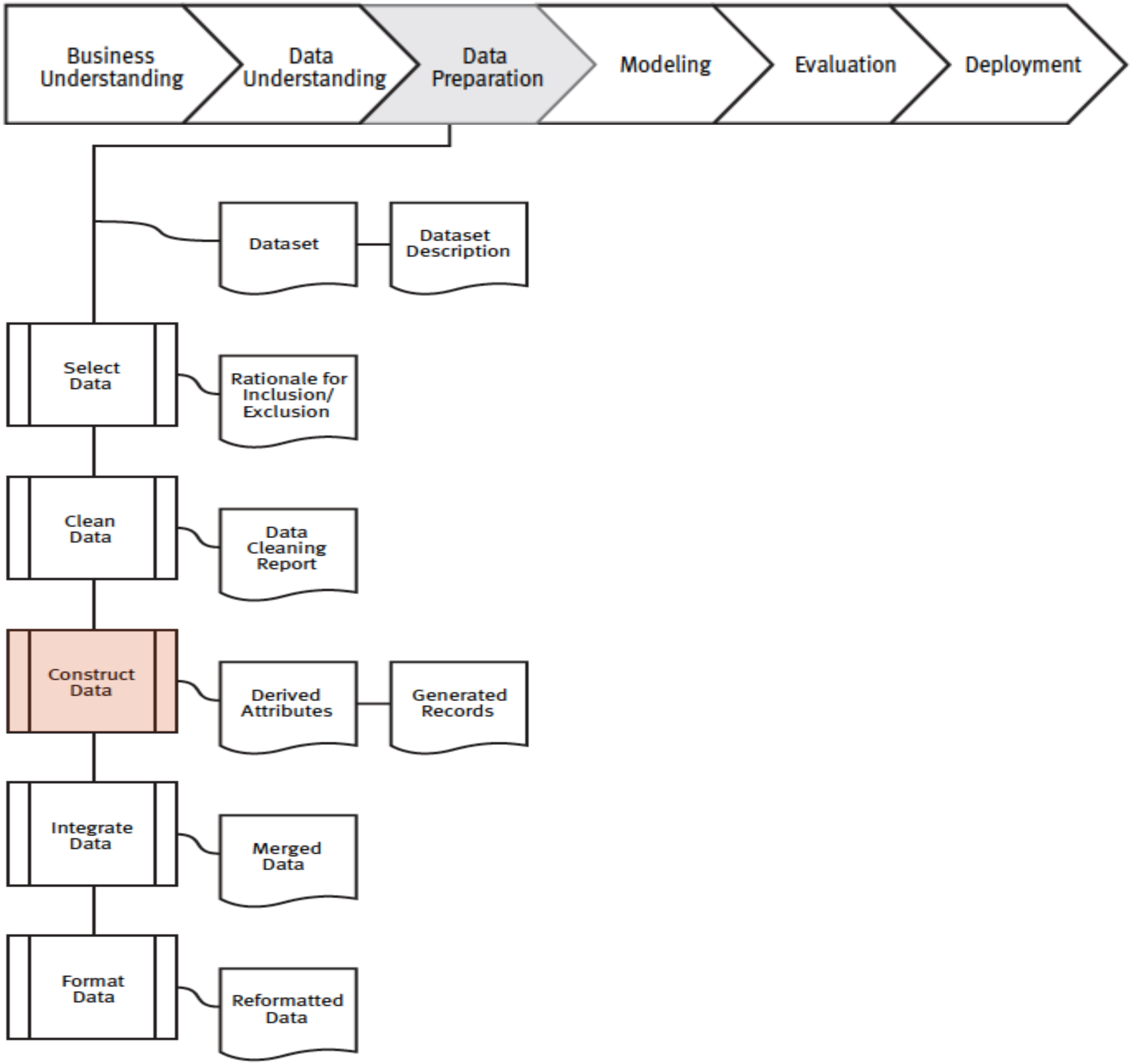
2. Clean data

- Raise the data quality to the level required by the selected analysis techniques
- This may involve
  - selection of clean subsets of the data
  - insertion of suitable defaults
  - more ambitious techniques such as the estimation of missing data by modeling

2. Clean data

- Describe what decisions and actions were taken to address the data quality problems reported during the Verify Data Quality task of the Data Understanding phase
- Transformations of the data for cleaning purposes and the possible impact on the analysis results should be considered

Phase 3: Data preparation



3. Construct data

- This task includes constructive data preparation operations such as
  - production of derived attributes
  - production of entire new records

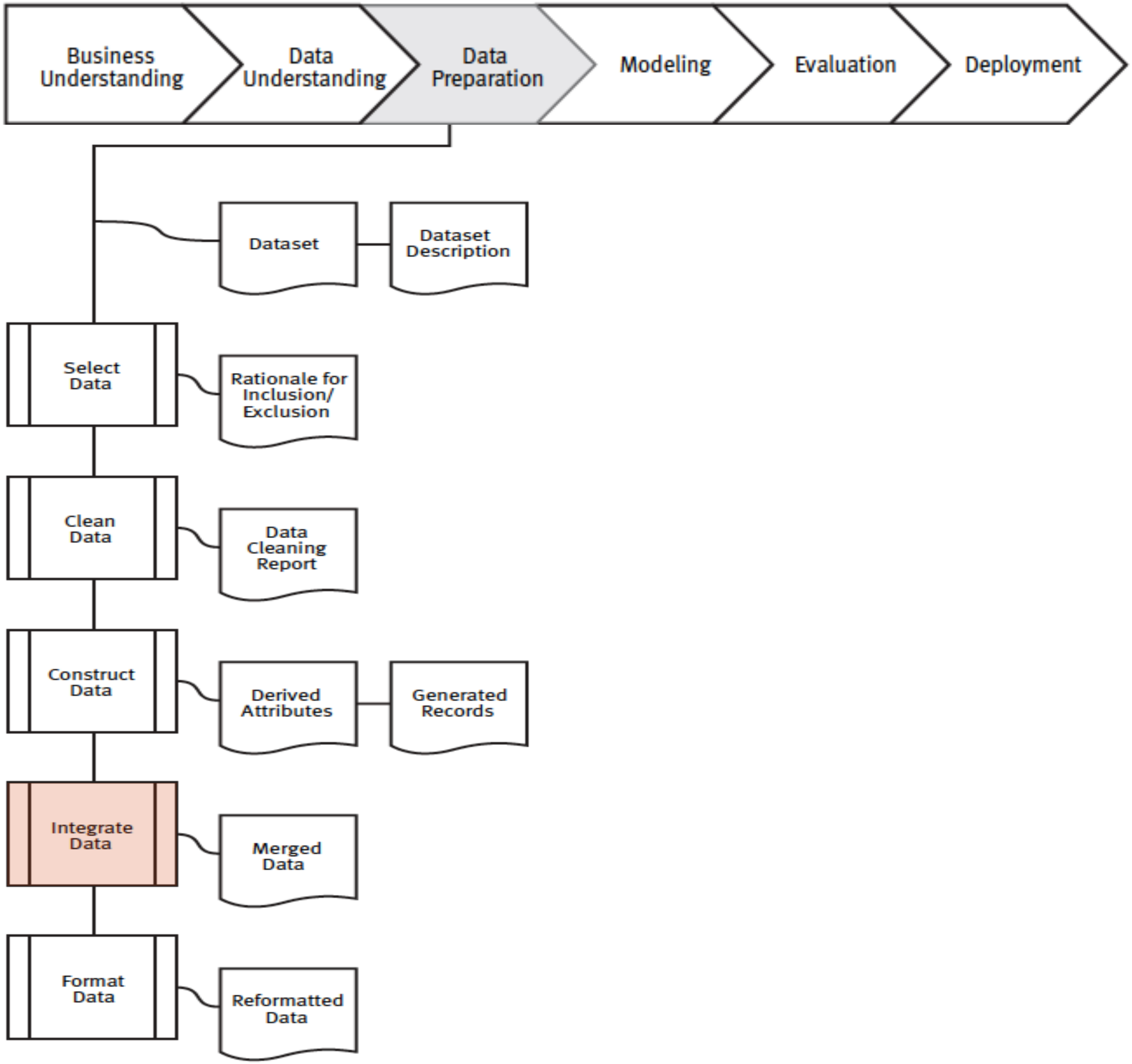
3.1. Derived attributes

- Derived attributes are new attributes that are constructed from one or more existing attributes in the same record
- Examples
    - area = length * width
    - age = years(today – birthdate)
    - logarithm of a feature (e.g., in genetics)
    - one-hot encoding
        - Country = ‘USA’, ‘Mexico’, ‘Canada’
        - CountryUSA = 0, 1
        - CountryMexico = 0, 1
        - CountryCanada = 0, 1

3.2. Generated records

- Describe the creation of completely new records
- Example:
  - Create records for customers who made no purchase during the past year
  - There was no reason to have such records in the raw data
  - But for modeling purposes it might make sense to explicitly represent the fact that certain customers made zero purchases

Phase 3: Data preparation



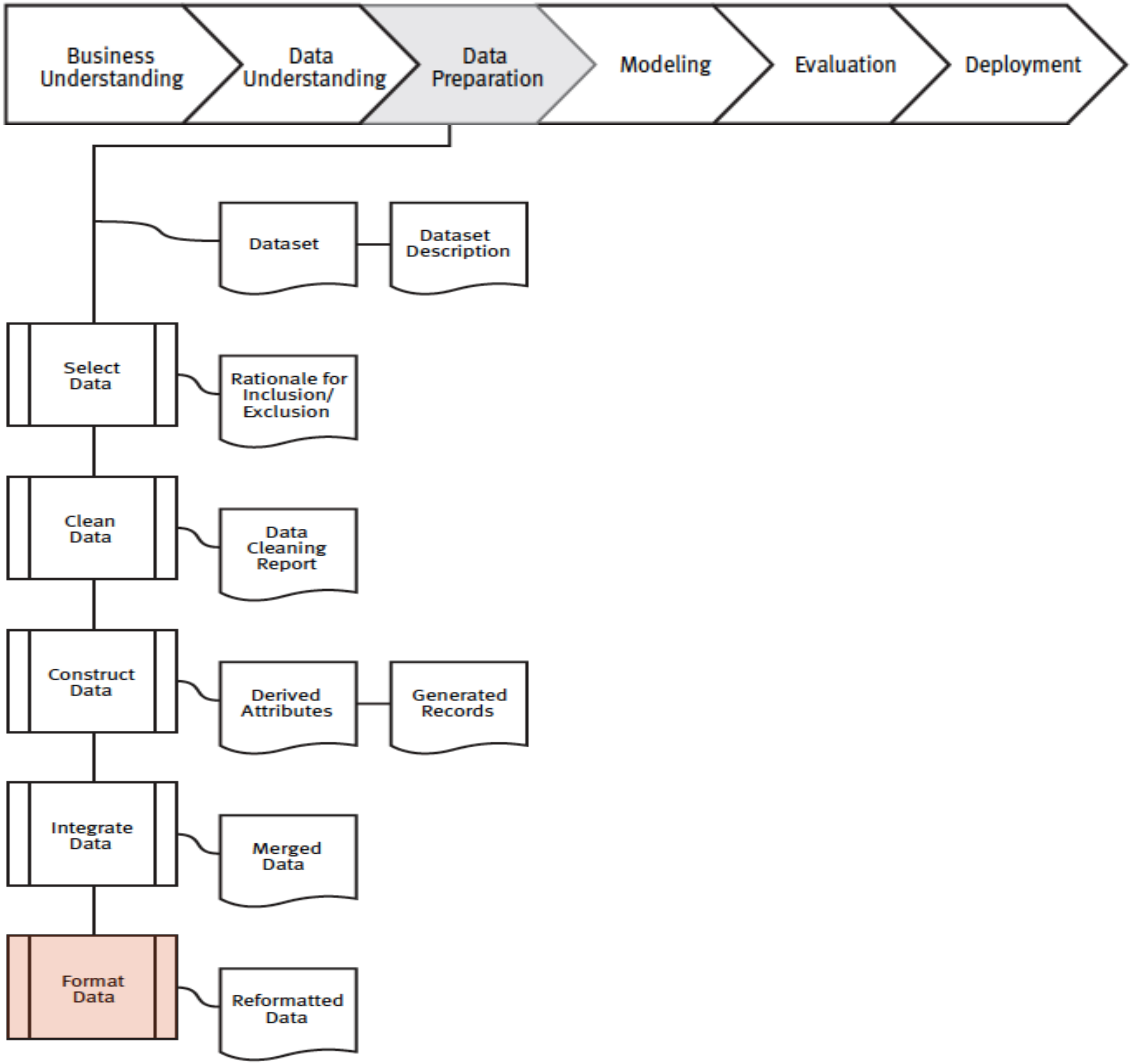
4. Integrate data

- These are methods whereby information is combined from multiple tables or records to create new records or values
- Merging tables refers to joining together two or more tables that have different information about the same objects
- Example:
  - A retail chain has one table with information about each store's general characteristics (e.g., floor space, type of mall), another table with summarized sales data (e.g., profit, percent change in sales from previous year), and another with information about the demographics of the surrounding area
  - Each of these tables contains one record for each store
  - These tables can be merged together into a new table with one record for each store, combining fields from the source tables

4. Integrate data

- Merged data also covers aggregations
- Aggregation refers to operations in which new values are computed by summarizing information from multiple records and/or tables
- Example: converting a table of customer purchases where there is one record for each purchase into a new table where there is one record for each customer, with fields such as
  - number of purchases
  - average purchase amount
  - percent of orders charged to credit card
  - percent of items under promotion
  - number of items per category: NumItemsMeats, NumItemsBeverages, NumItemsVegetables, etc.

Phase 3: Data preparation



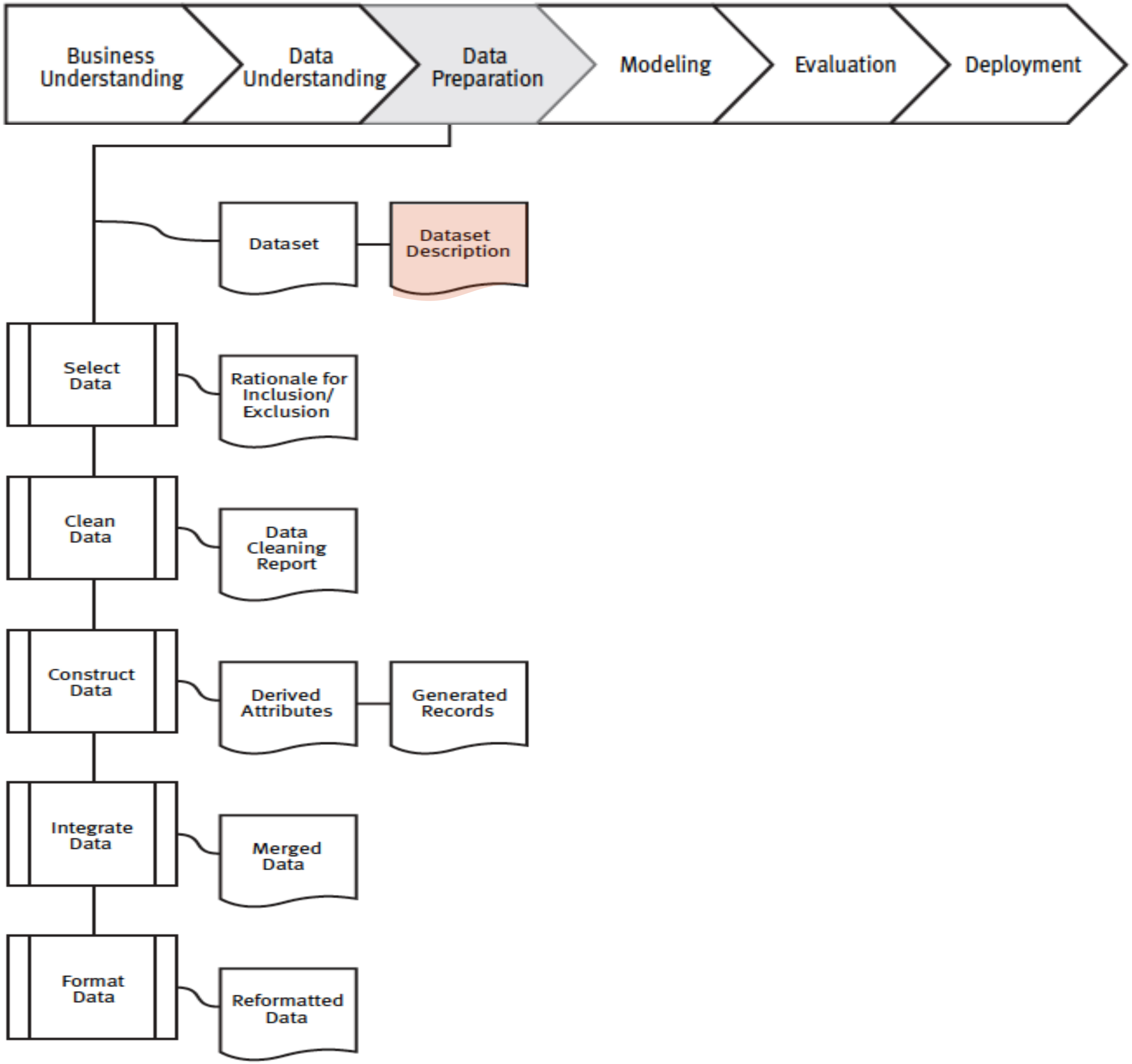
5. Format data

- Formatting transformations refer to primarily syntactic modifications made to the data that do not change its meaning, but might be required by the modeling tool
- Some tools have requirements on the order of the attributes, such as
  - the first field being a unique identifier for each record
  - the last field being the outcome field the model is to predict

5. Format data

- It might be important to change the order of the records in the dataset
  - Perhaps the modeling tool requires that the records be sorted according to the value of the outcome attribute
- Commonly, records of the dataset are initially ordered in some way, but the modeling algorithm needs them to be in a fairly random order
  - For example, when using neural networks, it is generally best for the records to be presented in a random order, although some tools handle this automatically without explicit user intervention
- Additionally, there are purely syntactic changes made to satisfy the requirements of the specific modeling tool
  - Examples: removing commas from within text fields in comma-delimited data files, trimming all values to a maximum of 32 characters

Phase 3: Data preparation



6. Dataset description

- Provide a general description of the final dataset
- For instance, in terms of number of number of samples and number of features