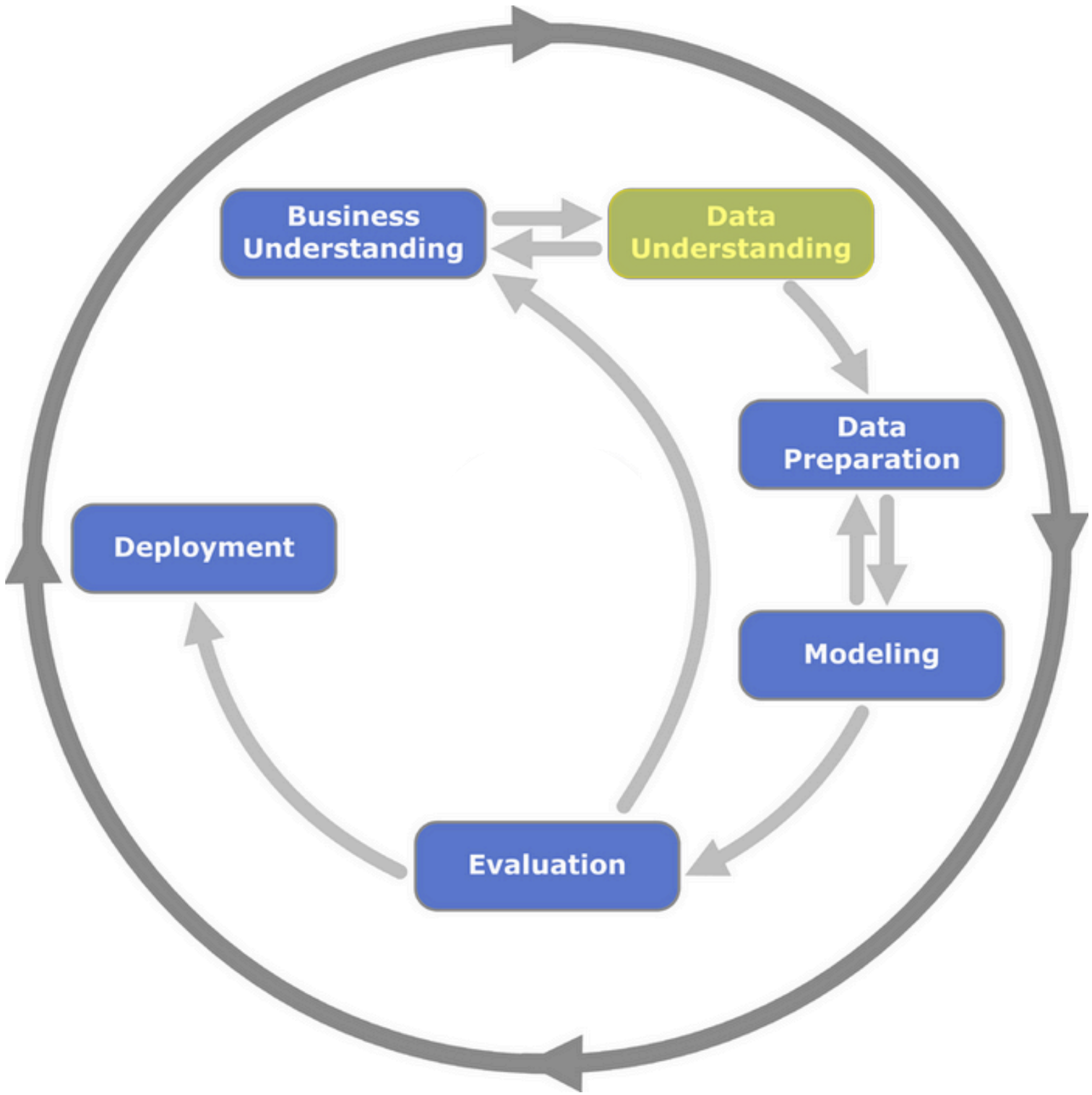
CS490DSC Data Science Capstone
Data Understanding

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

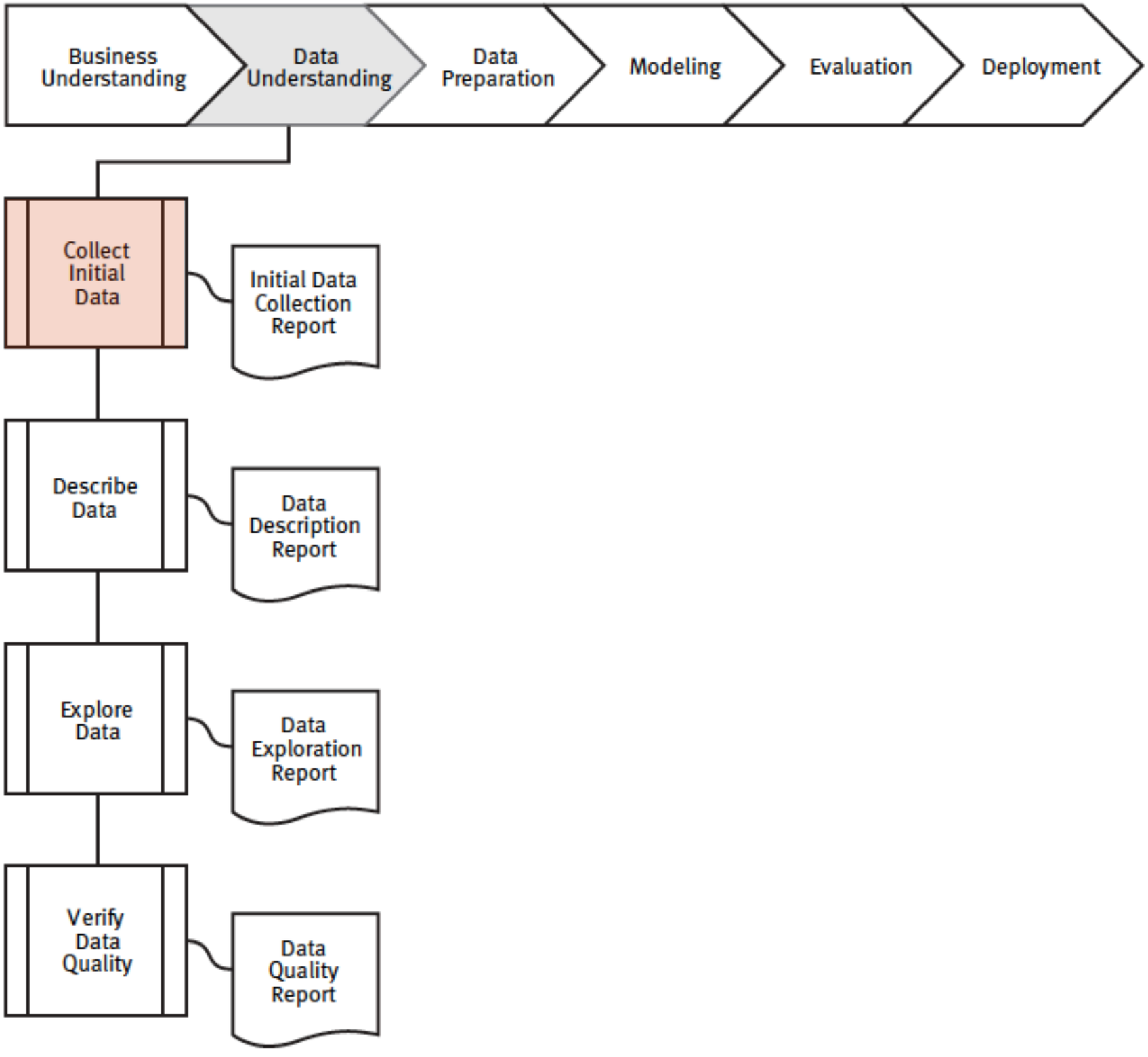
CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment

Phase 2: Data understanding

- This phase starts with initial data collection and proceeds with activities that enable you to
  - become familiar with the data
  - identify data quality problems
  - discover first insights into the data
  - detect interesting subsets to form hypotheses regarding hidden information

Phase 2: Data understanding



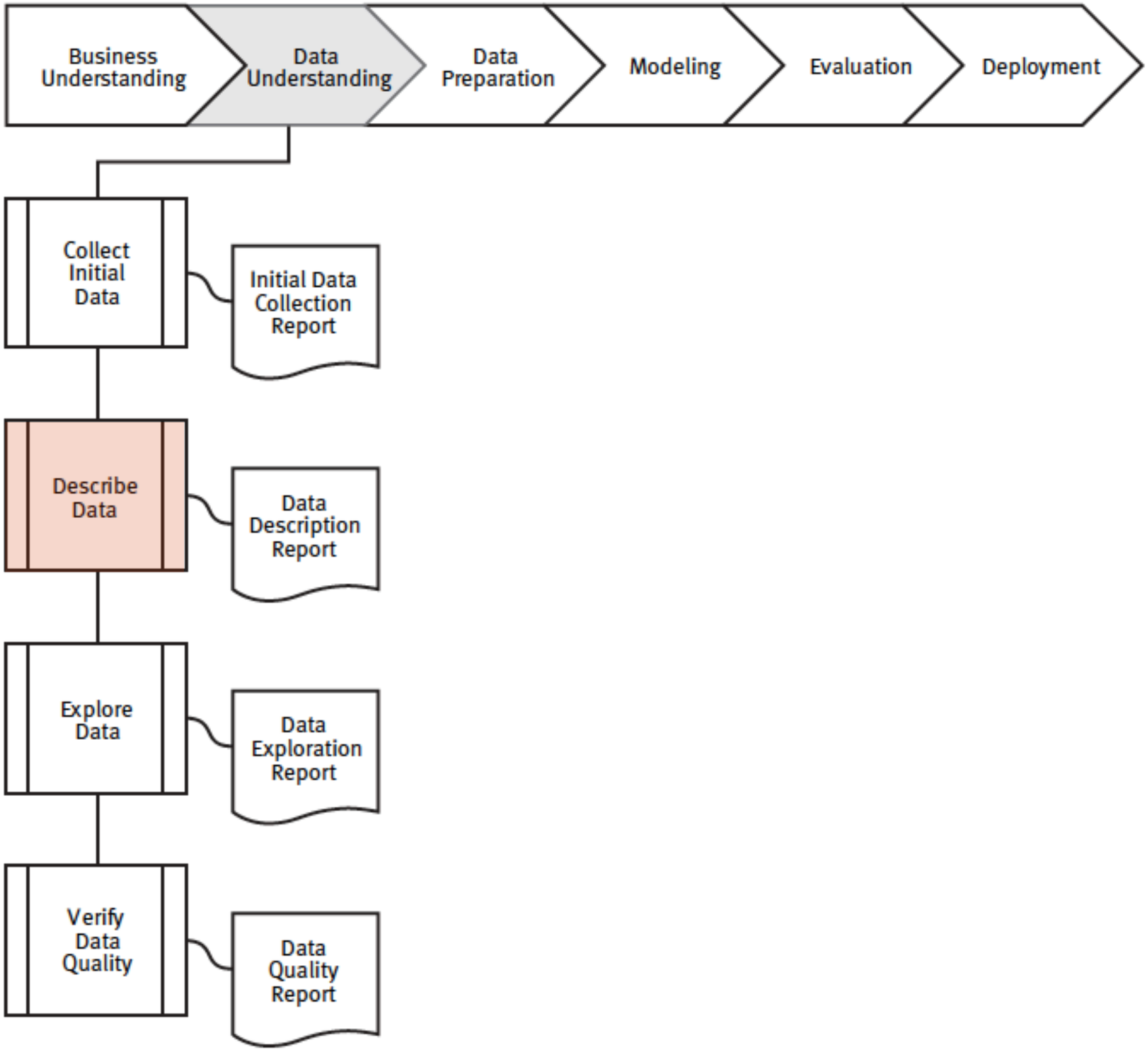
1. Collect initial data

- Acquire the data (or access to the data) listed in the project resources
- This initial collection includes data loading, if necessary for data understanding
  - For example, if you use a specific tool for data understanding, it makes perfect sense to load your data into this tool
- This effort possibly leads to initial data preparation steps
- If you acquire multiple data sources, integration is an additional issue, either here or in Phase 3: Data Preparation

1. Collect initial data

- List the dataset(s) acquired, together with
  - their locations
  - the methods used to acquire them
  - any problems encountered
- Record problems encountered and any resolutions achieved
- **This will aid with future replication of this project or with the execution of similar future projects**

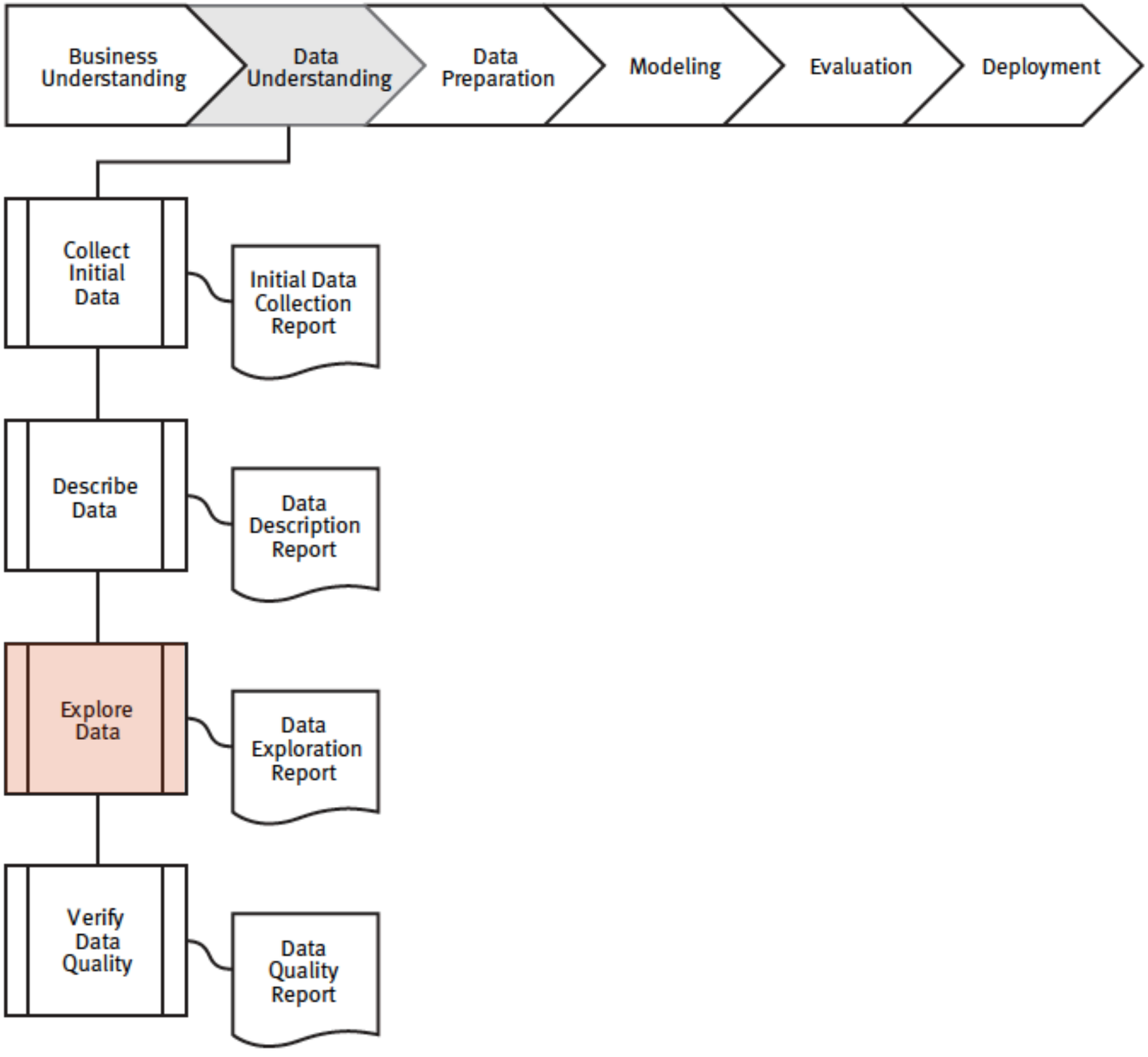
Phase 2: Data understanding



2. Describe data

- Examine the “gross” or “surface” properties of the acquired data and report on the results
- Describe the data that has been acquired, including
  - the format of the data
  - the quantity of data (for example, the number of records and fields in each table)
  - the identities of the fields
  - any other surface features which have been discovered
- Evaluate whether the data acquired satisfies the relevant requirements

Phase 2: Data understanding



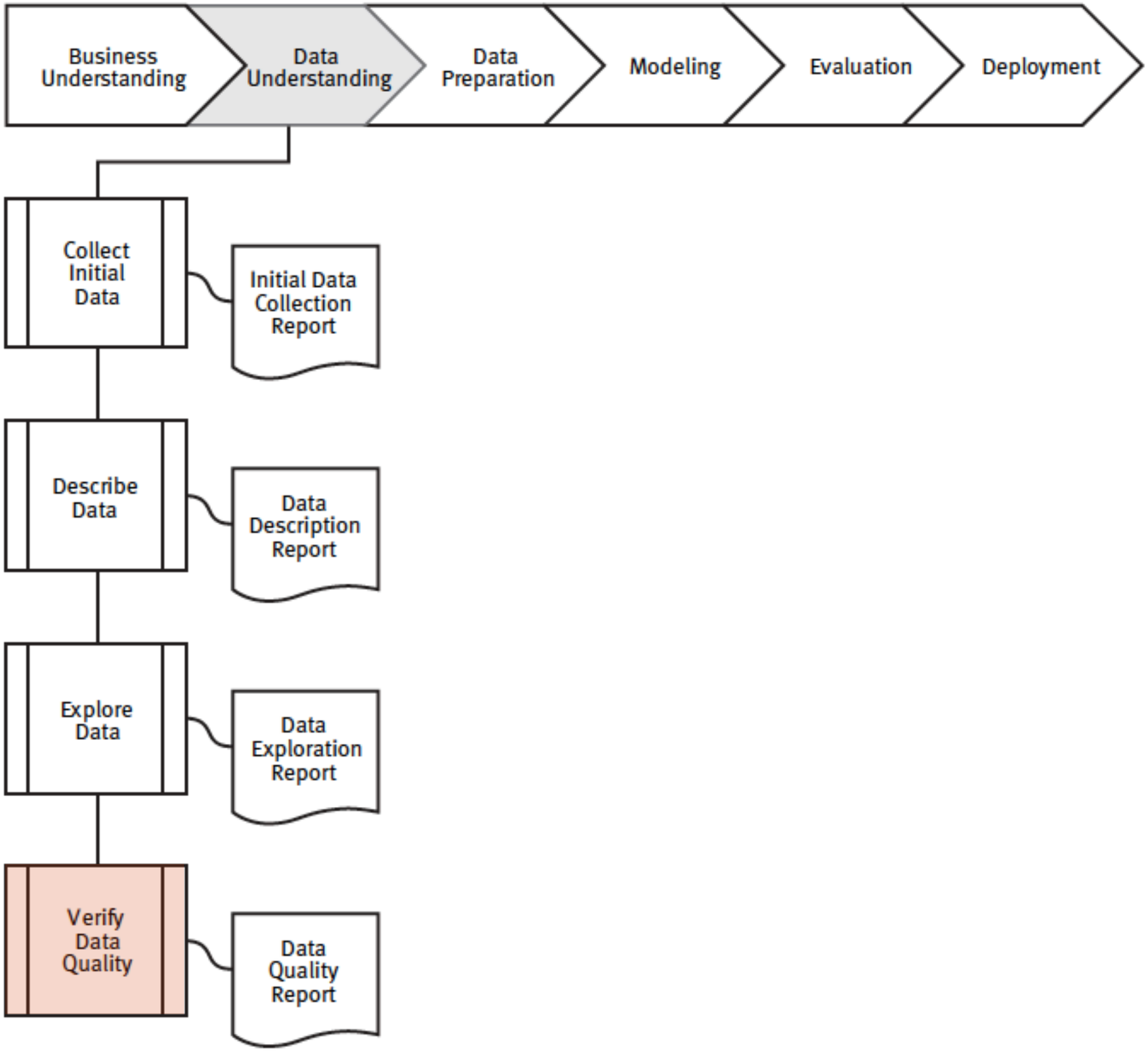
3. Explore data

- This task addresses data mining questions using querying, visualization, and reporting techniques
- These include
  - distribution of key attributes (for example, the target attribute of a prediction task)
  - relationships between pairs or small numbers of attributes
  - results of simple aggregations
  - properties of significant sub-populations
  - simple statistical analyses

3. Explore data

- These analyses
  - may directly address the data mining goals
  - may contribute to or refine the data description and quality reports
  - may feed into the transformation and other data preparation steps needed for further analysis
- Describe results of this task, including first findings or initial hypothesis and their impact on the remainder of the project
- If appropriate, include graphs and plots to indicate data characteristics that suggest further examination of interesting data subsets

Phase 2: Data understanding



4. Verify data quality

- Examine the quality of the data, addressing questions such as:
  - Is the data complete (does it cover all the cases required)?
  - Is the data correct, or does it contain errors?
  - If there are errors, how common are they?
  - Are there missing values in the data?
  - If values are missing, how are they represented? where do they occur? how common are they?
- List the results of the data quality verification
  - if quality problems exist, list possible solutions
- **Solutions to data quality problems generally depend heavily on both data and business knowledge**