TU 257 – Fundamentals of Data Science

Data Analytics

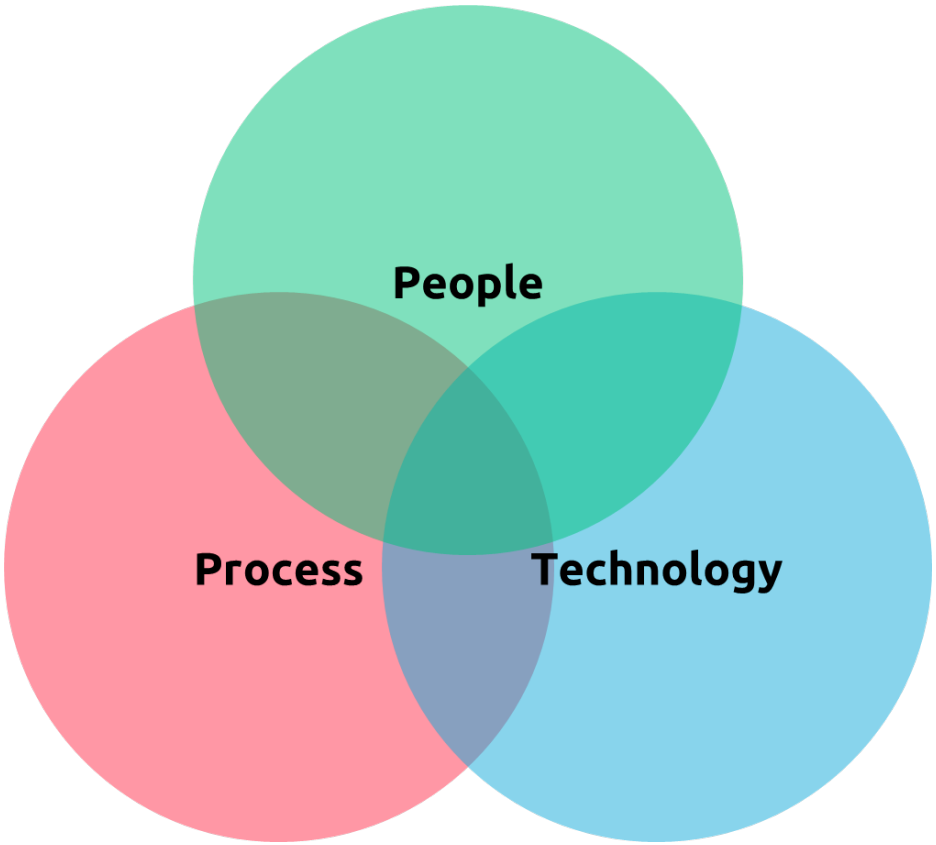
L2 – The Data Analytics & Data Science Life Cycle

Brendan Tierney

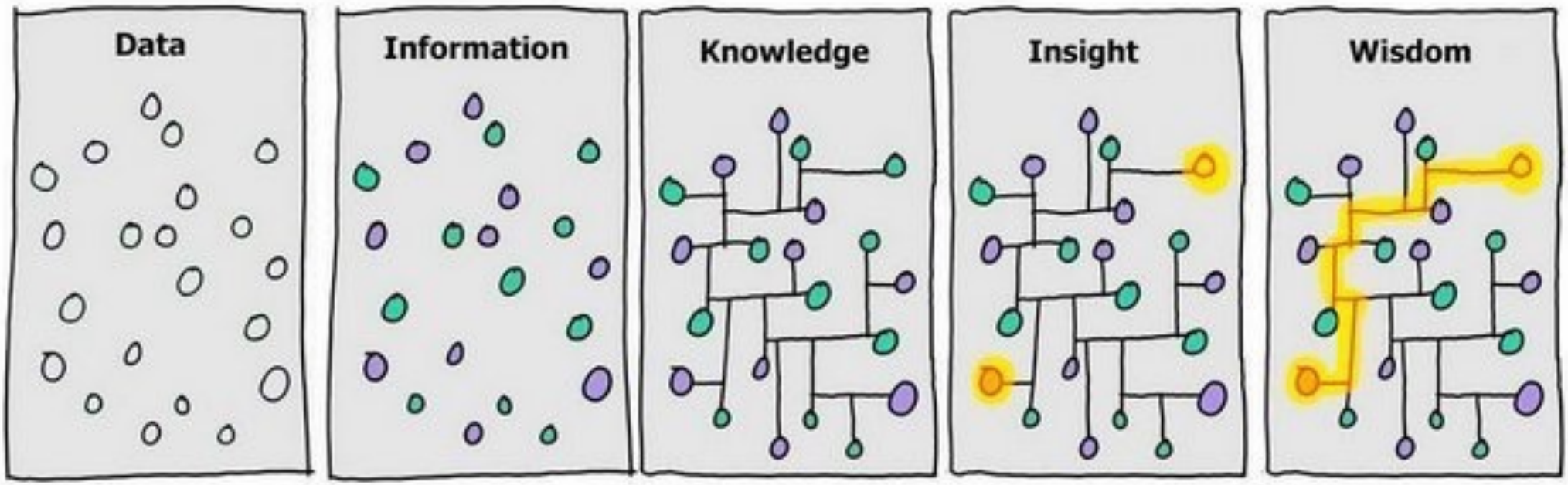
Agenda

- Why do we need a Life Cycle
- Which Life Cycle should we use
- First things first – Define the Problem
- CRISP-DM
- Even more – MLOps – Extending CRISP-DM

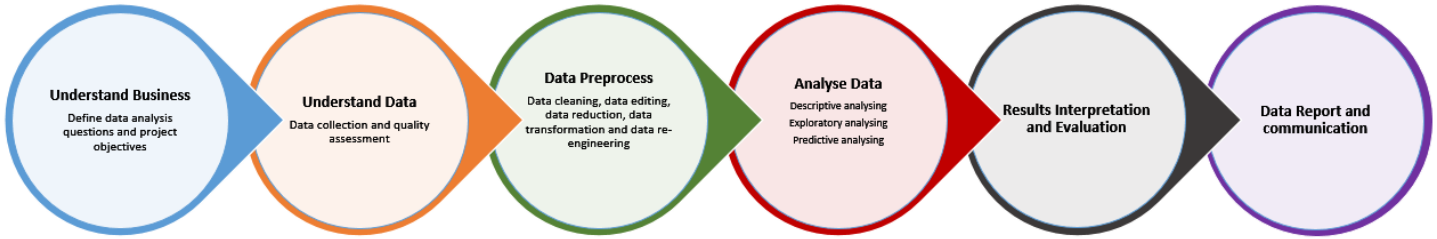
It’s not about the Technology/Code



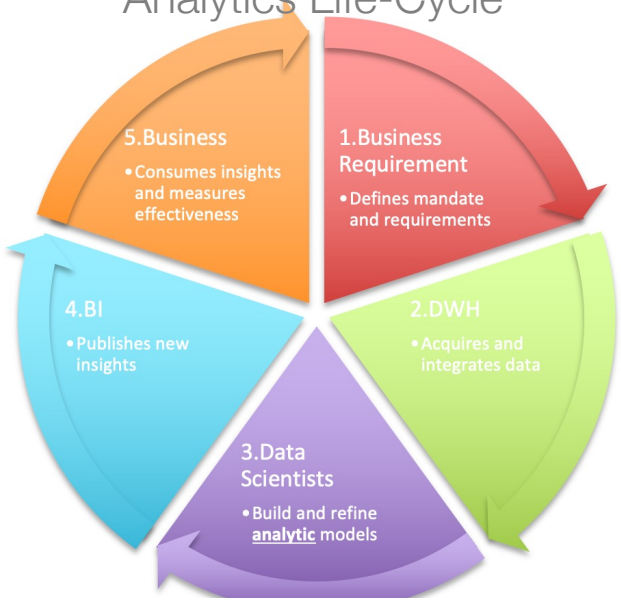
Why do we need a Life Cycle?


Data
Information
Knowledge
Insight
Wisdom

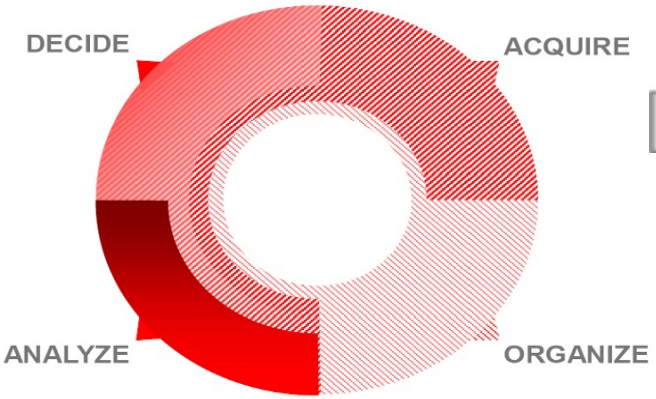
Lots of Life Cycles

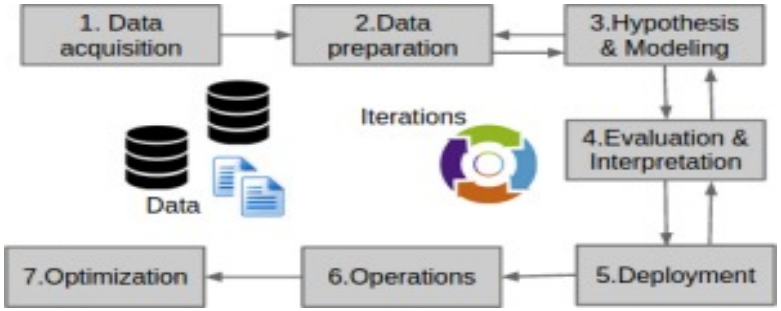


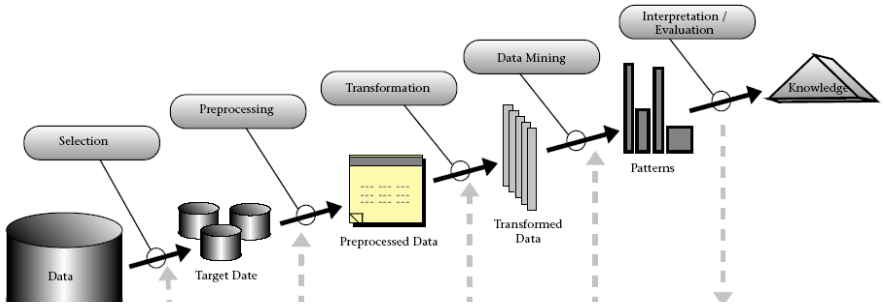
Analytics Life-Cycle

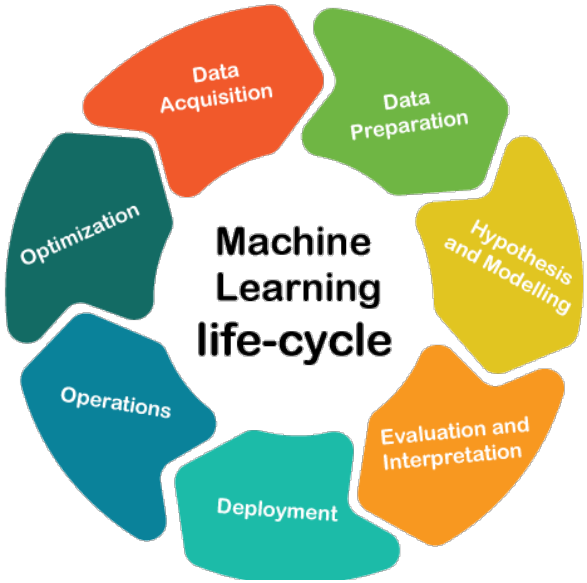


Big Data





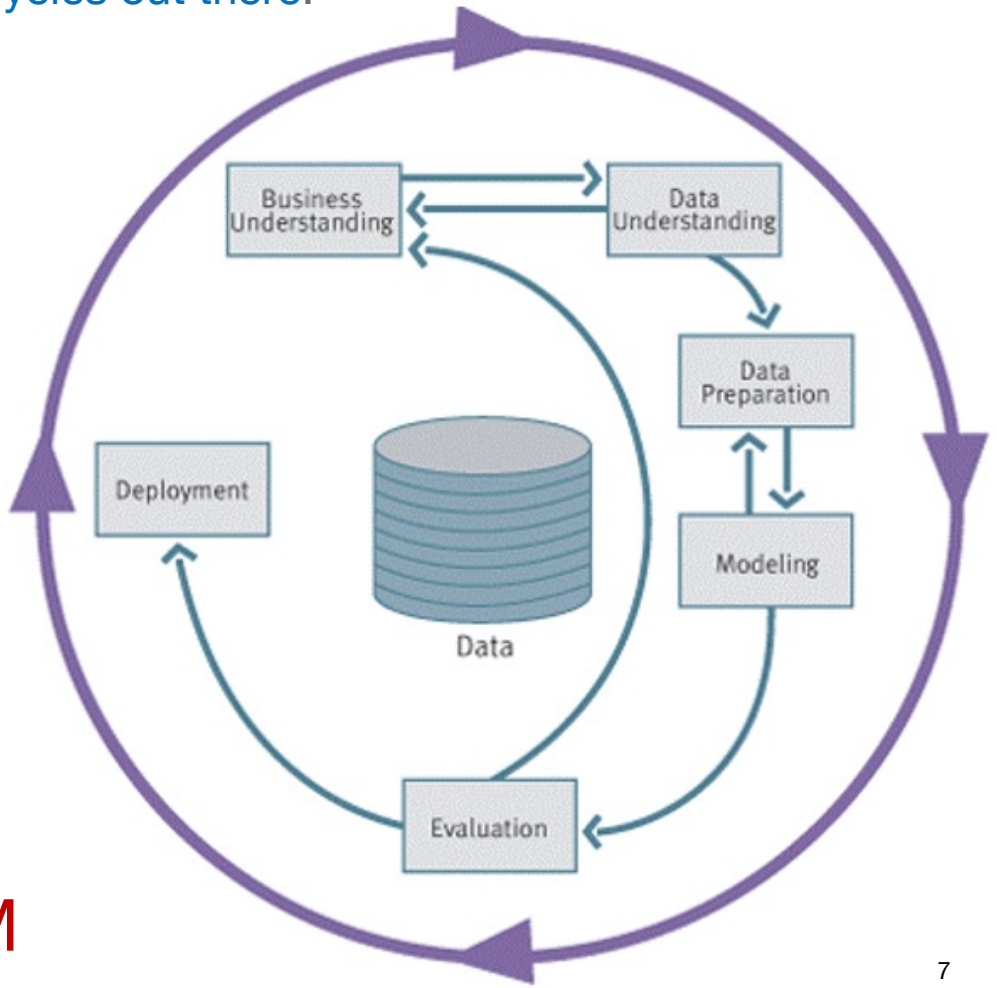




There are lots and lots and lots and ……

of life cycles out there.

They are basically all saying the same thing !

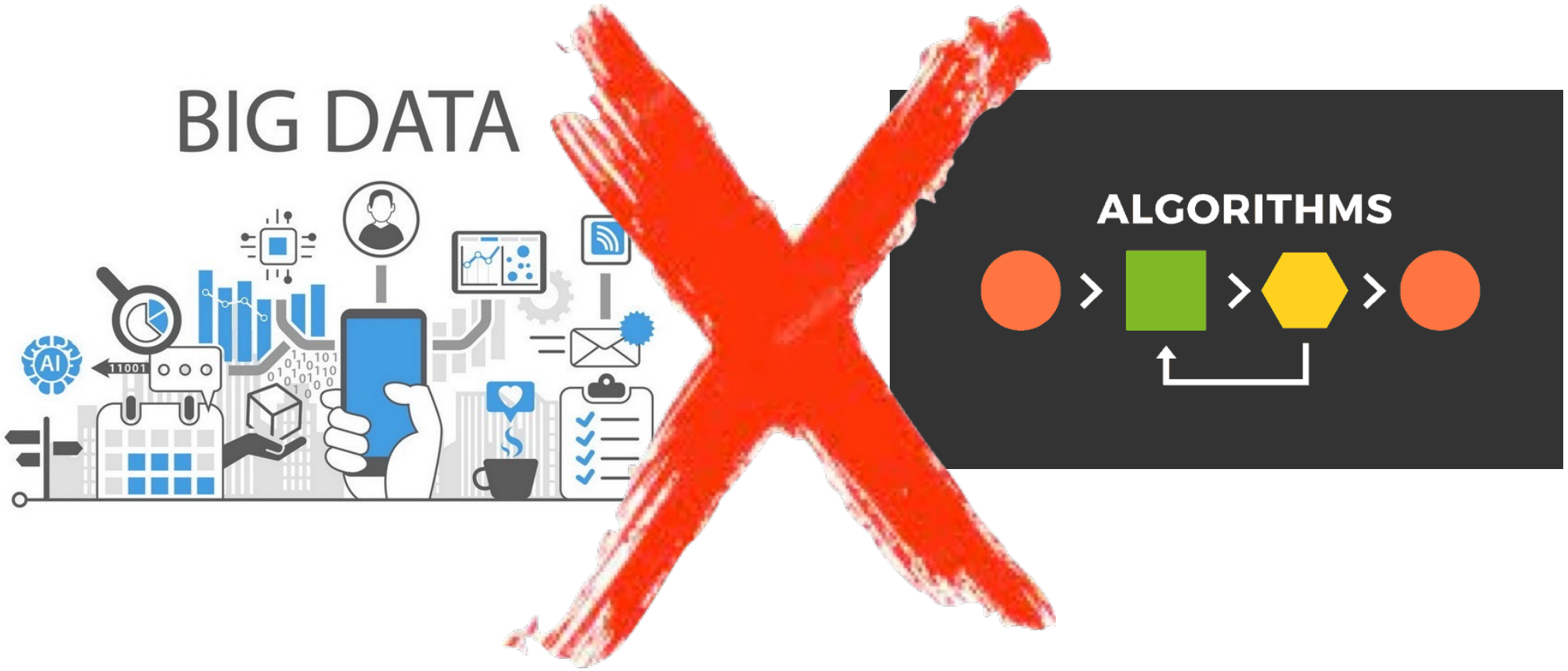


And they are just a modified version of ……

CRISP-DM

Define the Problem

- With all projects/exercises/tasks/etc. We start by Defining the Problem
- Why?
  - Without a clear definition/view of the problem we want to solve, we can
    - Go in wrong direction
    - Use technology that isn't appropriate
    - Know what data to use, and why we need to use it
    - How to test and evaluate (How do we know we have finished?)
    - What kind of Analytics do we need to perform


BIG DATA
AI
1100
ALGORITHMS

What is the problem ?

What do you want to achieve?

In some/most cases you don’t need ML !!!

Define the Problem

The Problem: You need to travel to Galway on Friday evening around 18:00

Mode of transport

- Car
- Bus
- Train
- Bike
- Hitch-Hike
- Walk

Why at that Time

…

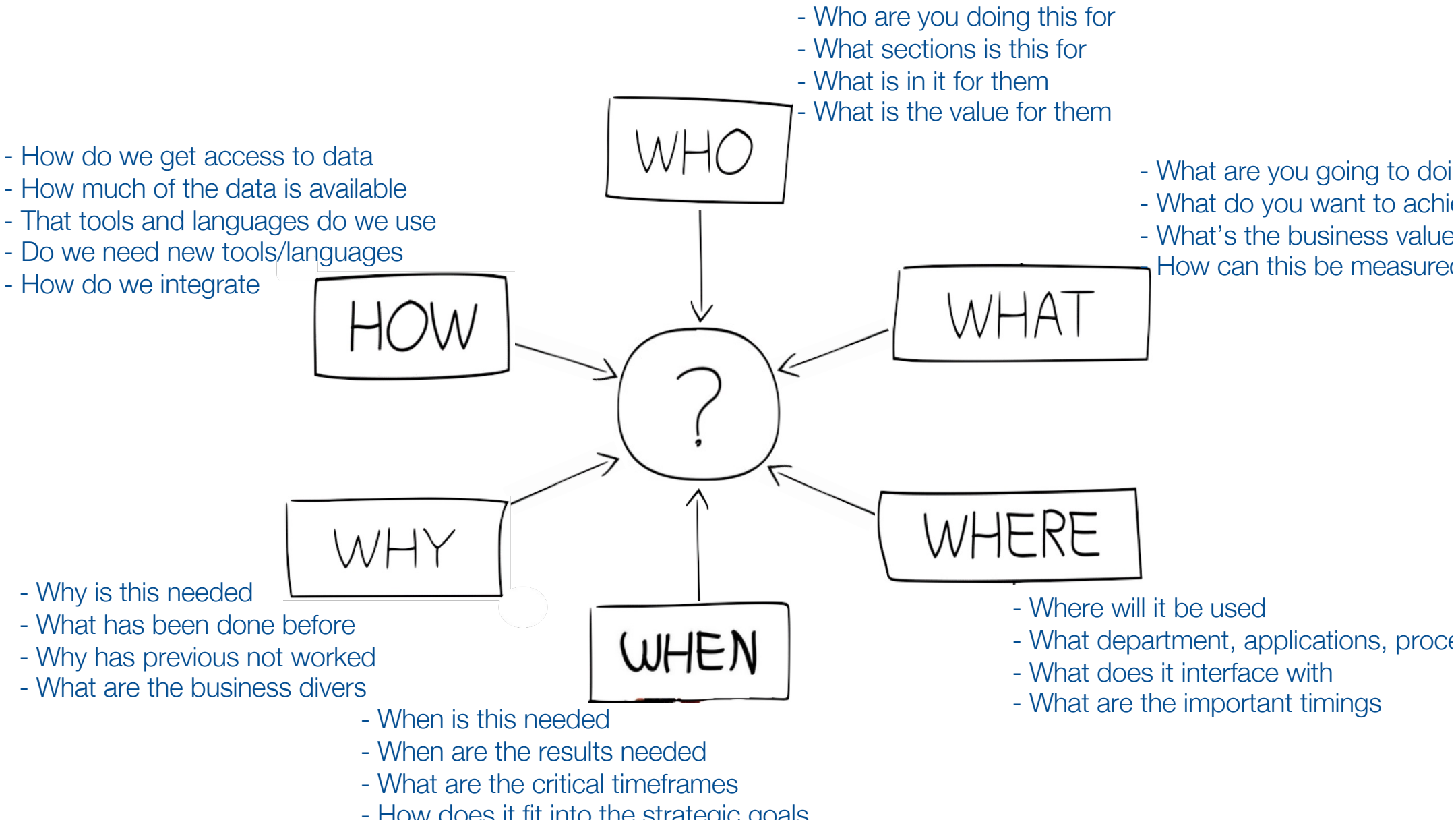
How do you plan this?

…

What Challenges will you have?

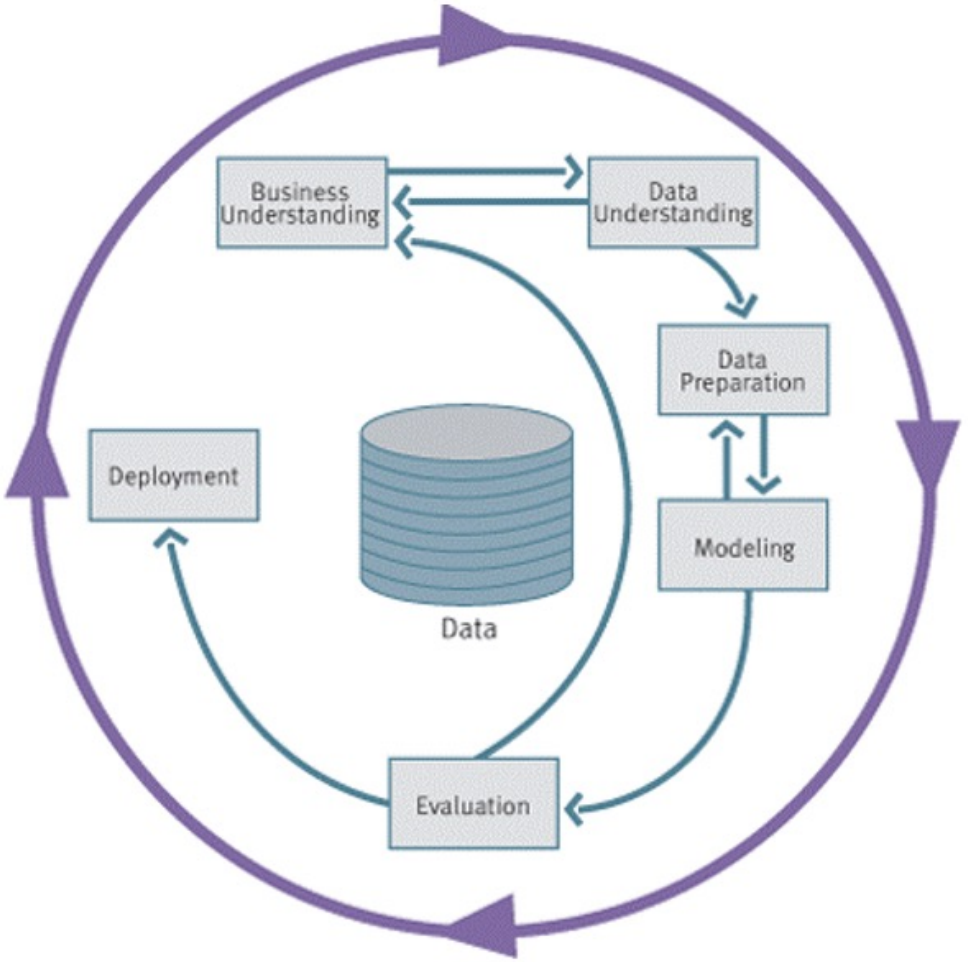
…

Anything else?


- Who are you doing this for
- What sections is this for
- What is in it for them
- What is the value for them
WHO
- How do we get access to data
- How much of the data is available
- That tools and languages do we use
- Do we need new tools/languages
- How do we integrate
HOW
- What are you going to doi
- What do you want to achi
- What's the business value
How can this be measure
WHAT
?
WHY
- Why is this needed
- What has been done before
- Why has previous not worked
- What are the business divers
WHEN
- When is this needed
- When are the results needed
- What are the critical timeframes
- How does it fit into the strategic goals
WHERE
- Where will it be used
- What department, applications, proce
- What does it interface with
- What are the important timings

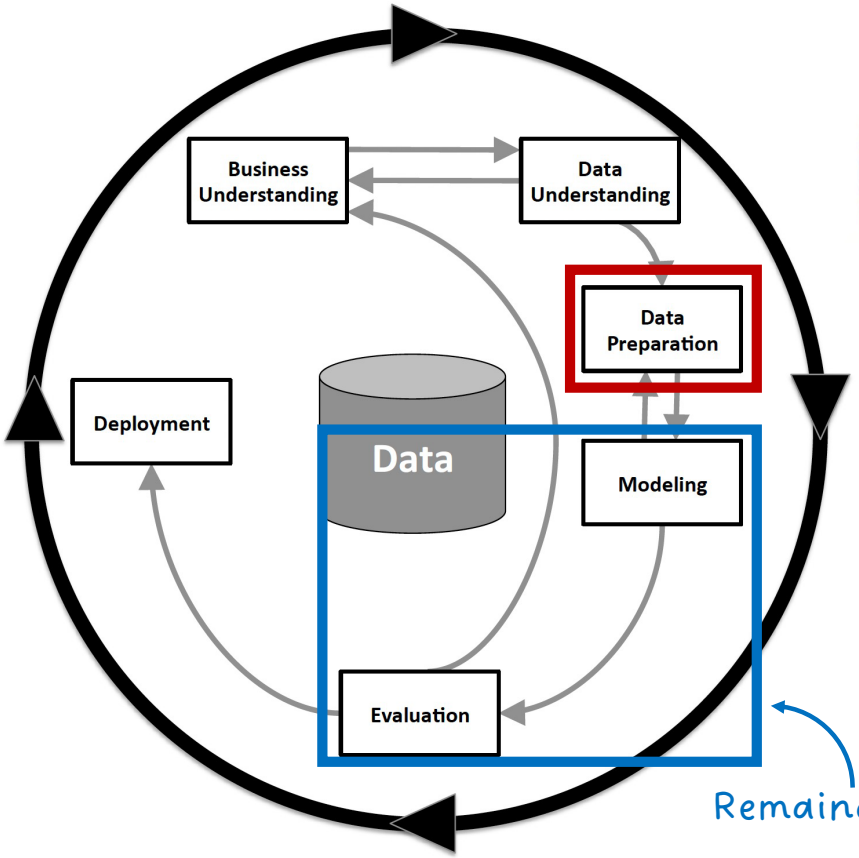
CRISP-DM

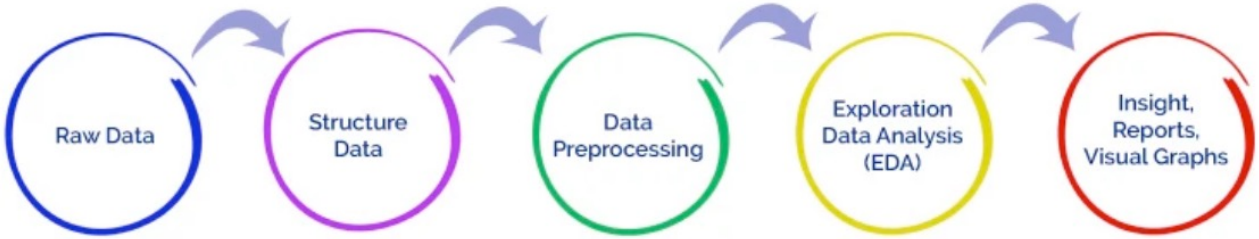
- Non-proprietary
- Application/Industry neutral
- Tool neutral
- Focus on business issues
  - As well as technical analysis
- Framework for guidance
- Experience base
  - Templates for Analysis



Next Week + remainder of Semester

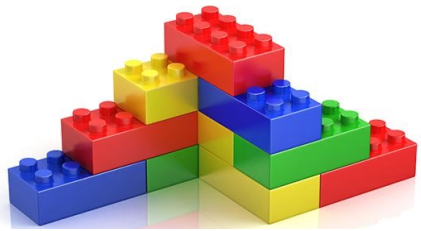
- Data Exploration and Data Preparation





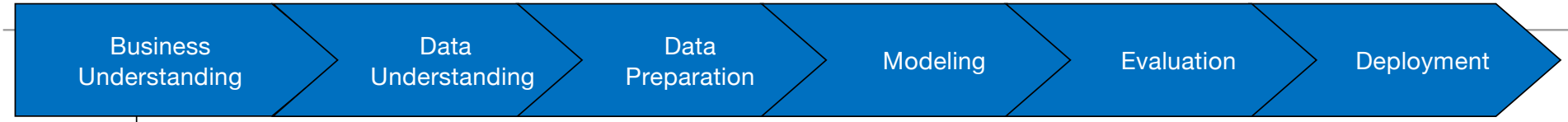
There will be some overlap with Data Wangling module

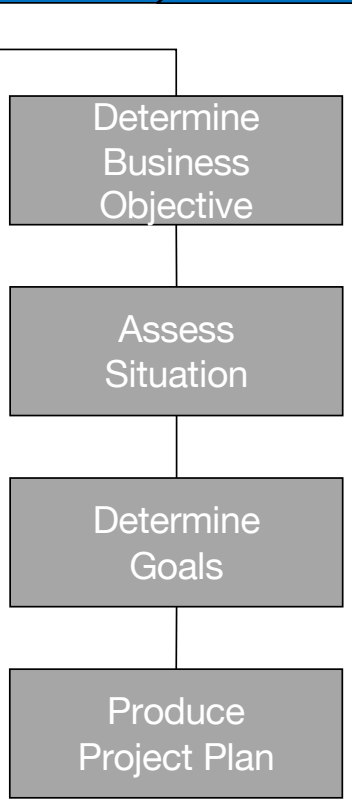
+ Extra tasks for preparing data for Data Analytics



- The following can be a little challenging to follow
- Can be a little difficult to understand – for now
- We'll be covering different aspects of every week
- Try to follow the best you can
- It will make a lot more sense as we progress through the semester

Phases in the DM Process – Business Understanding

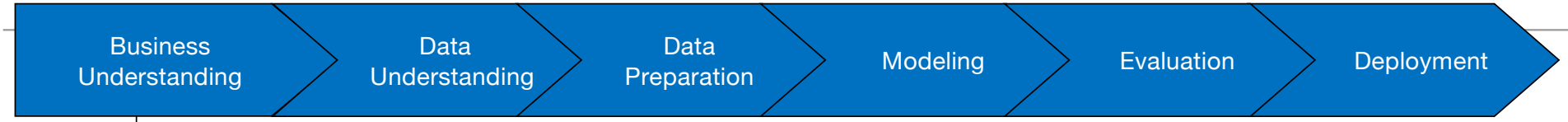


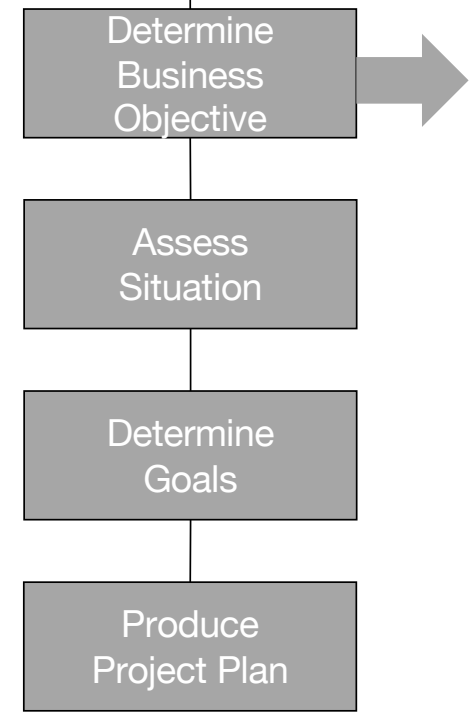


- Statement of Business Objective.
  - States goal in business terminology.
- Statement of Data Mining objective.
  - States objectives in technical terms.
- Statement of Success Criteria.

- Focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.
  - What the client really wants to accomplish?
  - Uncover important factors (constraints, competing objectives).

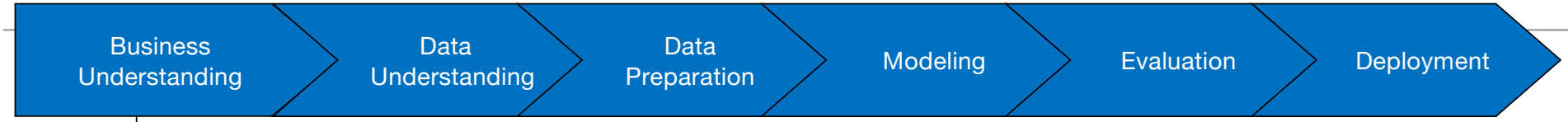
Phases in the DM Process – Business Understanding

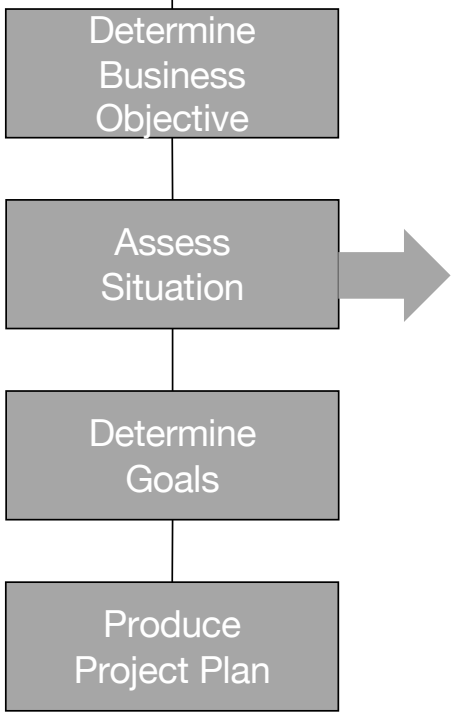




- Key persons and their roles?
- Is there a steering committee?
- Internal sponsor (financial, domain expert)?
- Business units impacted by the project (sales, finance,...)?
- Business success criteria and who assesses it?
- Users' needs and expectations.
- Describe problem in general terms. Business questions, Expected benefits.

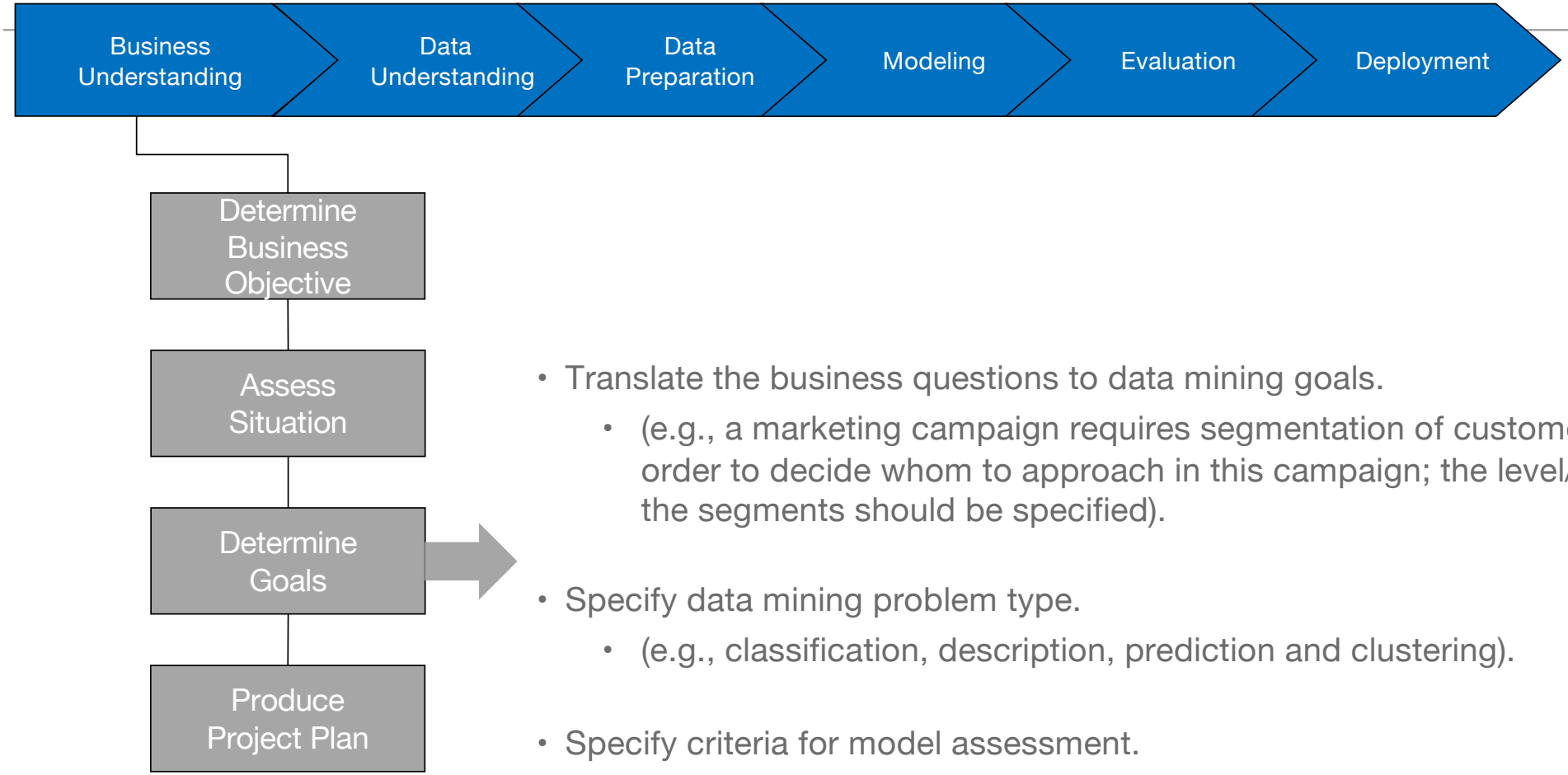
Phases in the DM Process – Business Understanding





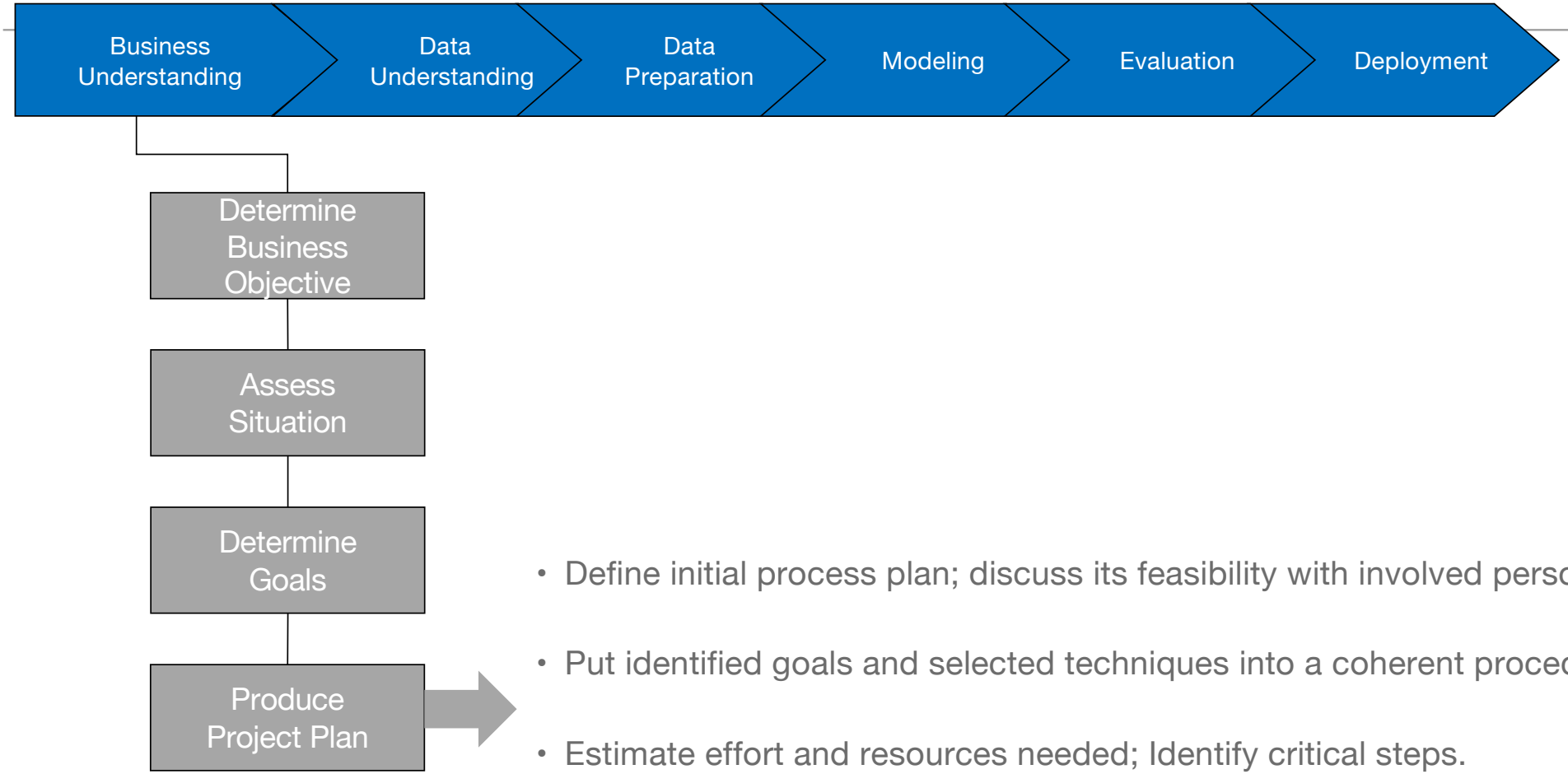
- Are they already using data mining?
- Identify hardware and software available.
- Identify data sources and their types (online, experts, written documentation).
- Identify knowledge sources and types (online, experts, written documentation)
- Describe the relevant background.

Phases in the DM Process – Business Understanding



- Translate the business questions to data mining goals.
  - (e.g., a marketing campaign requires segmentation of customers in order to decide whom to approach in this campaign; the level/size of the segments should be specified).
- Specify data mining problem type.
  - (e.g., classification, description, prediction and clustering).
- Specify criteria for model assessment.

Phases in the DM Process – Business Understanding

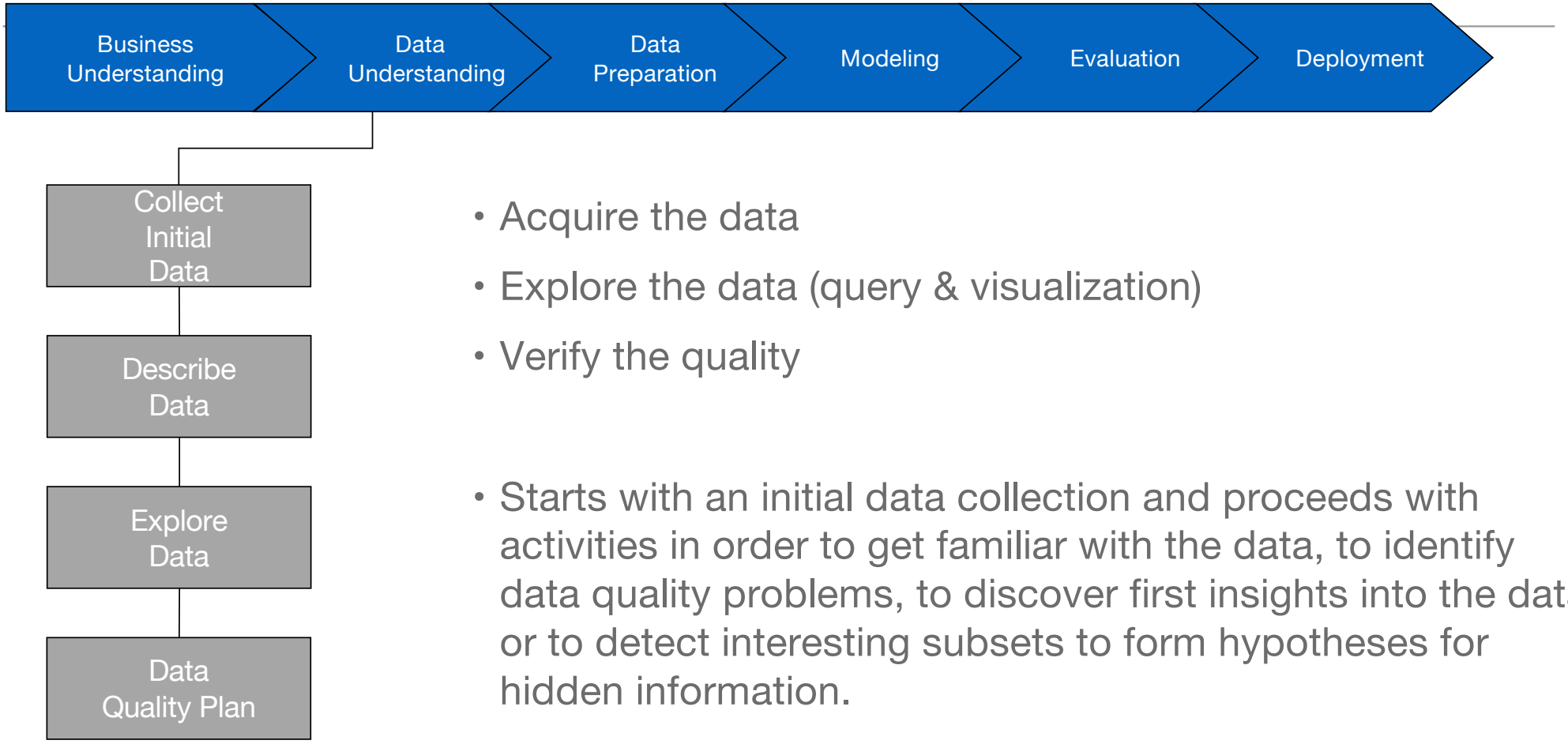


- Define initial process plan; discuss its feasibility with involved personnel.
- Put identified goals and selected techniques into a coherent procedure.
- Estimate effort and resources needed; Identify critical steps.

Key Questions at end of Business Understanding

1. What are we trying to achieve, business wise? Why is it important?
2. What are the inputs and outputs for the task that we are trying to solve?
3. Given a hypothetical solution to that task, how would it affect our operations? (another way to ask this question: assuming that I have a perfect solution to your machine learning task, how will you use it?)
4. Do we already have the ability to act based on such solution, or do we also need to develop that ability? (if the ability is there, learn it carefully. If not, keep close contact with the team that is responsible for developing it)
5. How are we going to measure a suggested solution? (KPIs)
6. What would make it a success?
7. Do we have the input data available? How hard it is to extract it? Are we allowed to use it?
8. Are we experienced with building similar solutions? Do we understand what it takes?
9. Do we have hard budget and timelines constraints?
10. Who will develop the solution? Do we have the required skills in house?
11. What analytics have we used already? What are the limitations of these approaches

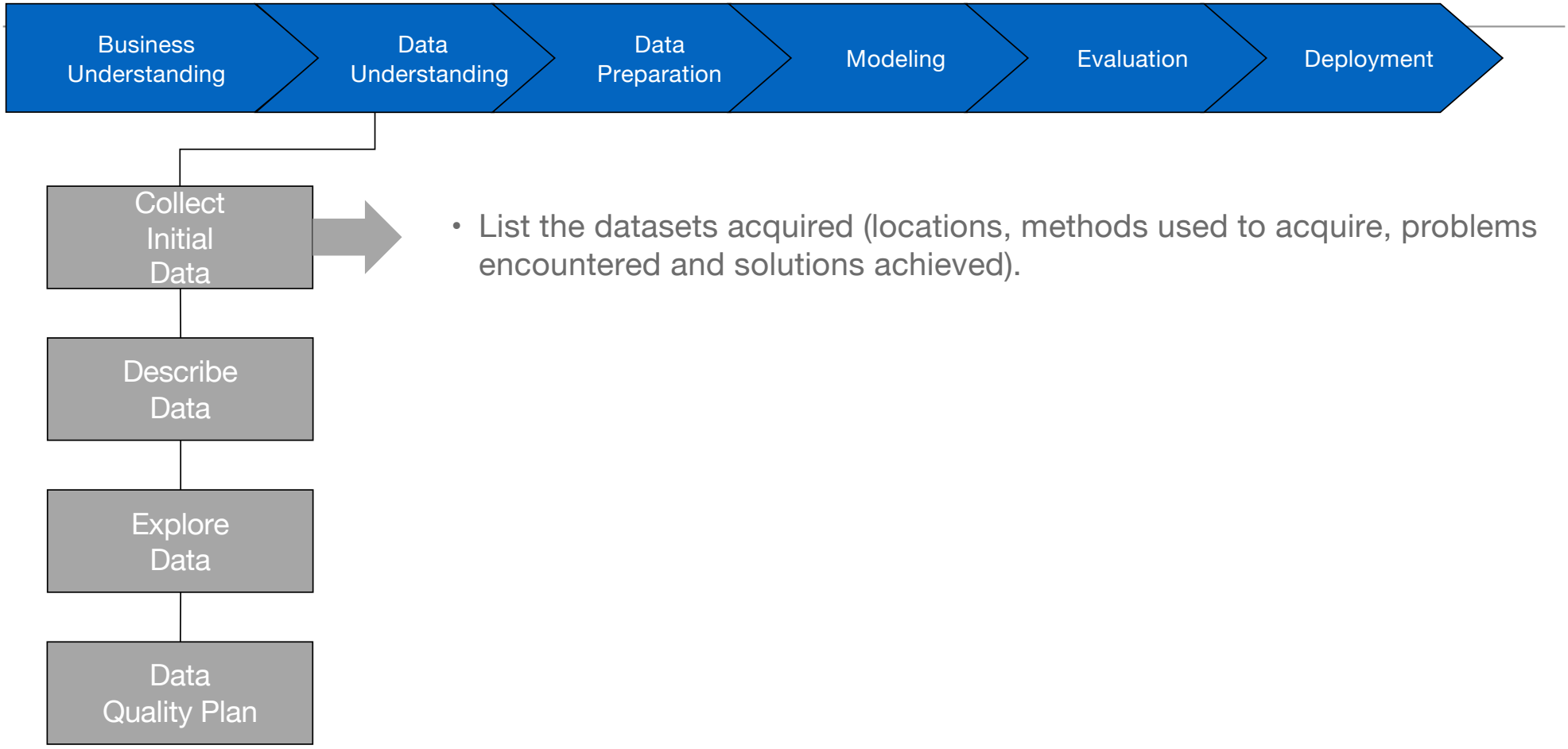
Phase 2 – Data Understanding



- Acquire the data
- Explore the data (query & visualization)
- Verify the quality

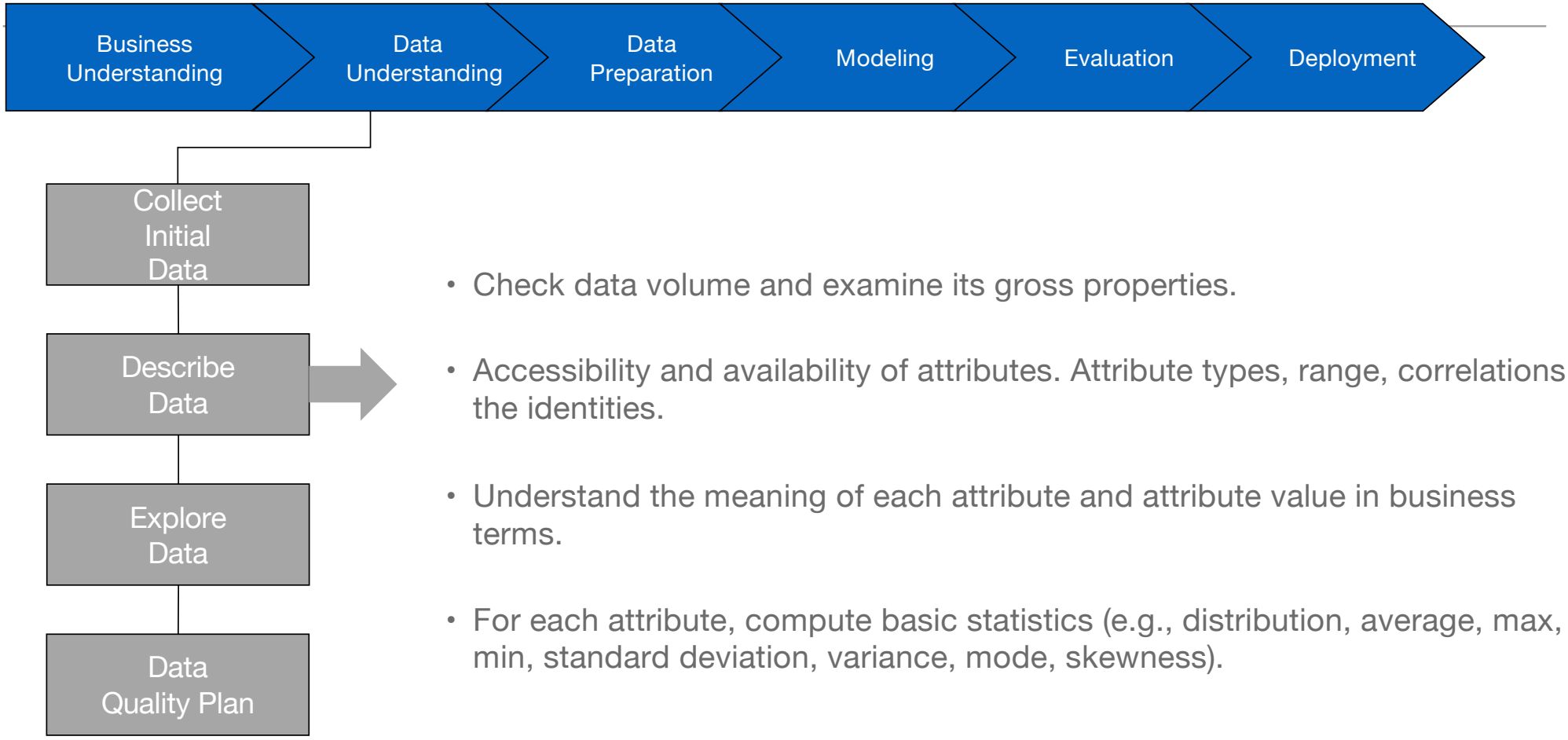
- Starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

Phase 2 – Data Understanding



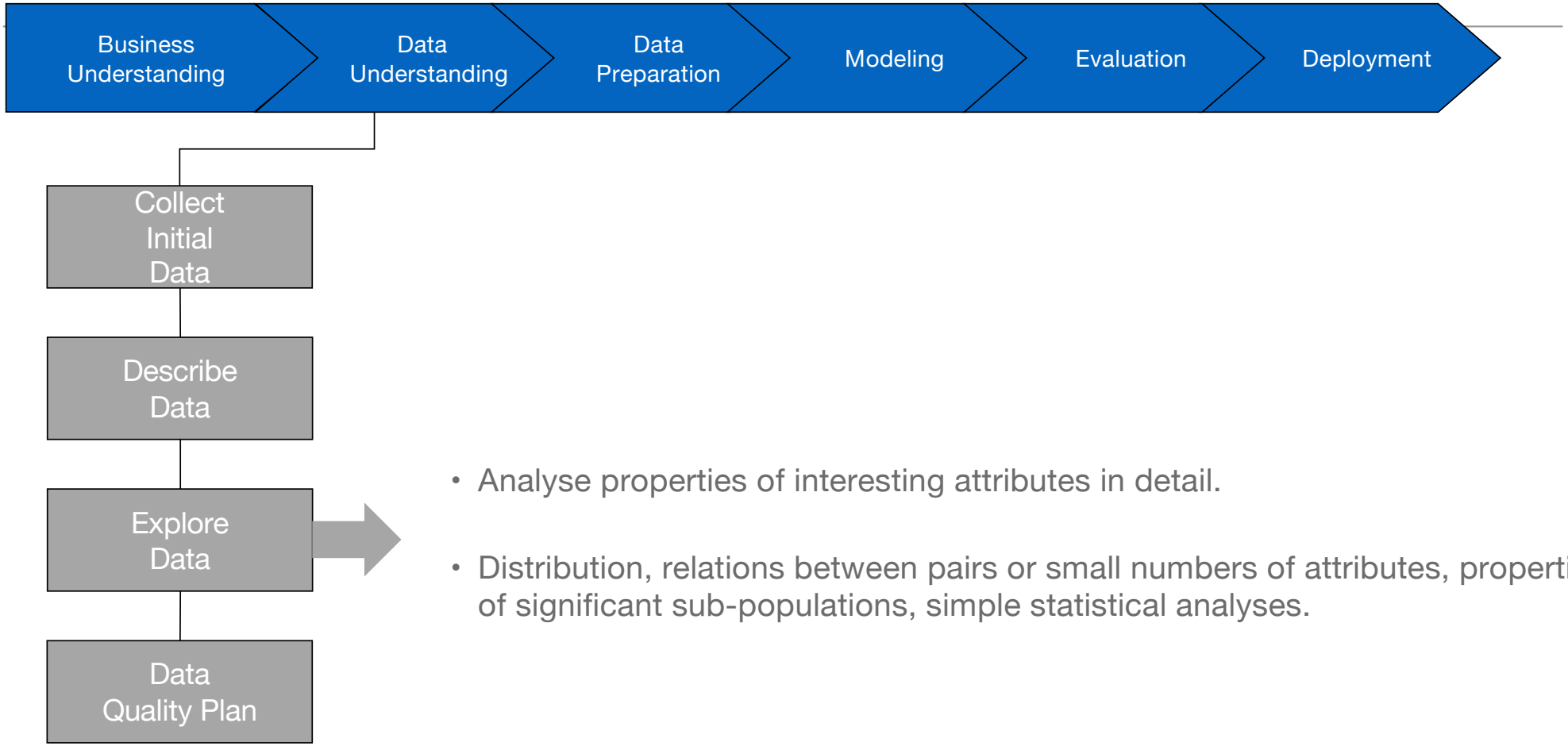
- List the datasets acquired (locations, methods used to acquire, problems encountered and solutions achieved).

Phase 2 – Data Understanding



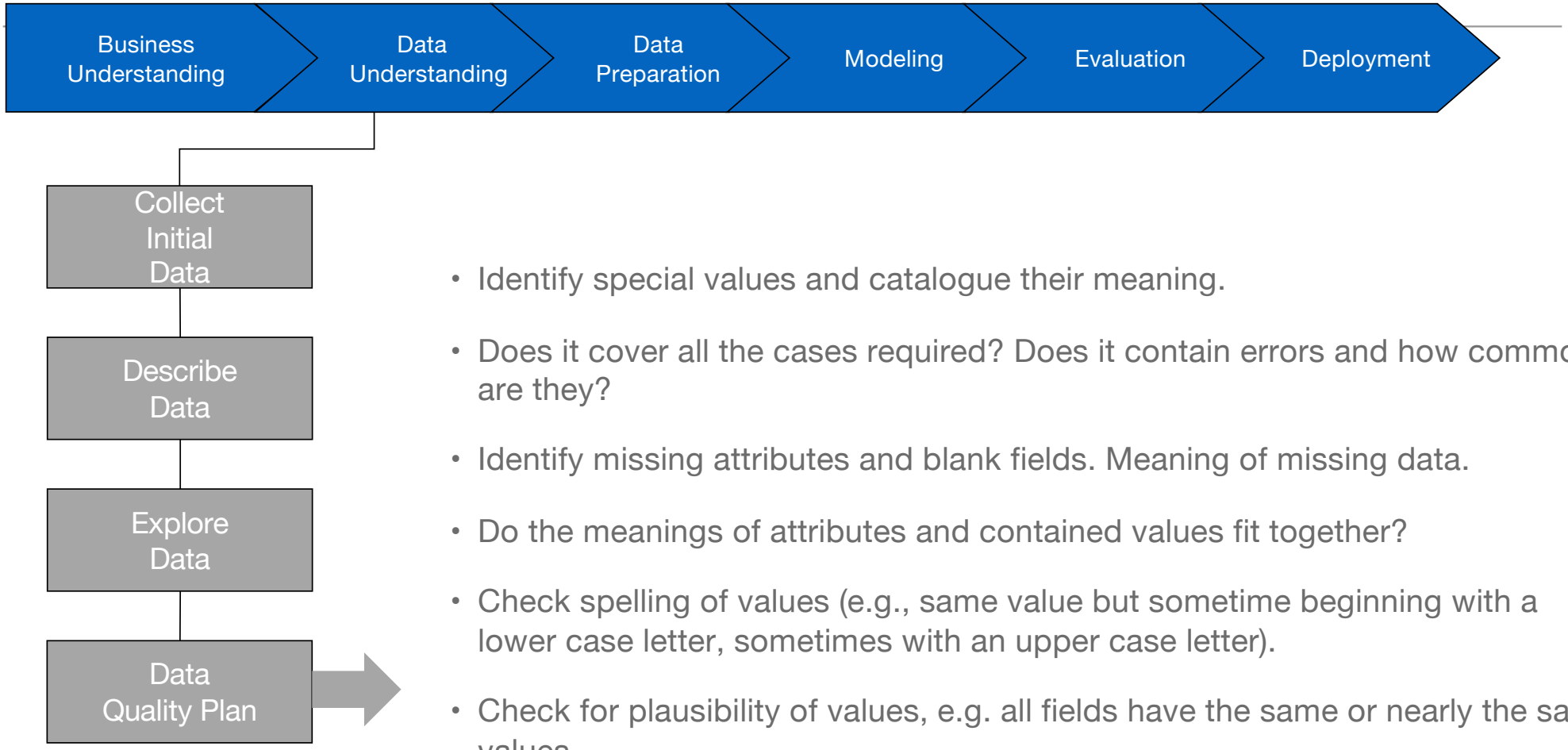
- Check data volume and examine its gross properties.
- Accessibility and availability of attributes. Attribute types, range, correlations, the identities.
- Understand the meaning of each attribute and attribute value in business terms.
- For each attribute, compute basic statistics (e.g., distribution, average, max, min, standard deviation, variance, mode, skewness).

Phase 2 – Data Understanding



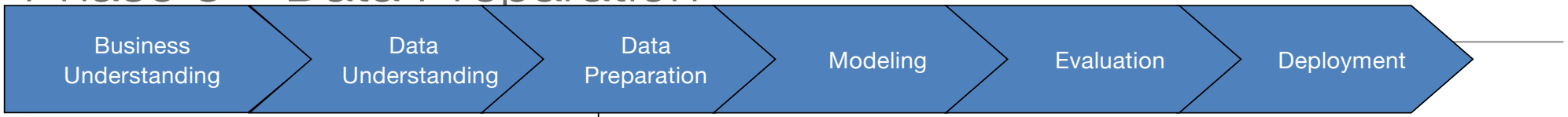
- Analyse properties of interesting attributes in detail.
- Distribution, relations between pairs or small numbers of attributes, properties of significant sub-populations, simple statistical analyses.

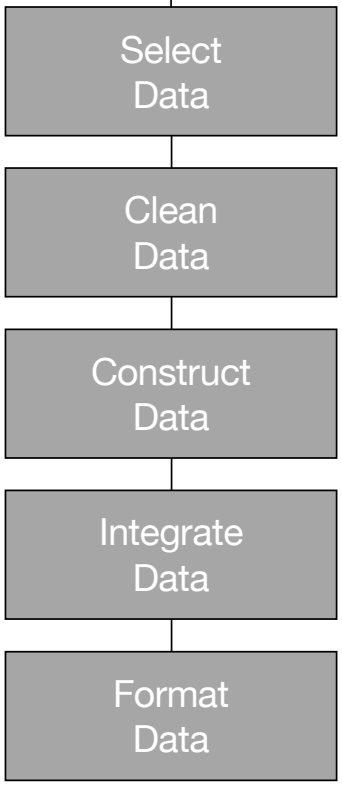
Phase 2 – Data Understanding



- Identify special values and catalogue their meaning.
- Does it cover all the cases required? Does it contain errors and how common are they?
- Identify missing attributes and blank fields. Meaning of missing data.
- Do the meanings of attributes and contained values fit together?
- Check spelling of values (e.g., same value but sometime beginning with a lower case letter, sometimes with an upper case letter).
- Check for plausibility of values, e.g. all fields have the same or nearly the same values.

Phase 3 – Data Preparation



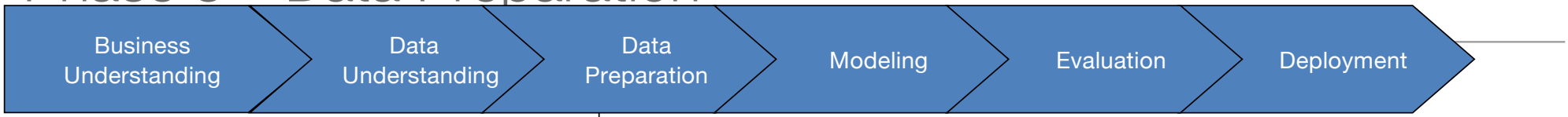


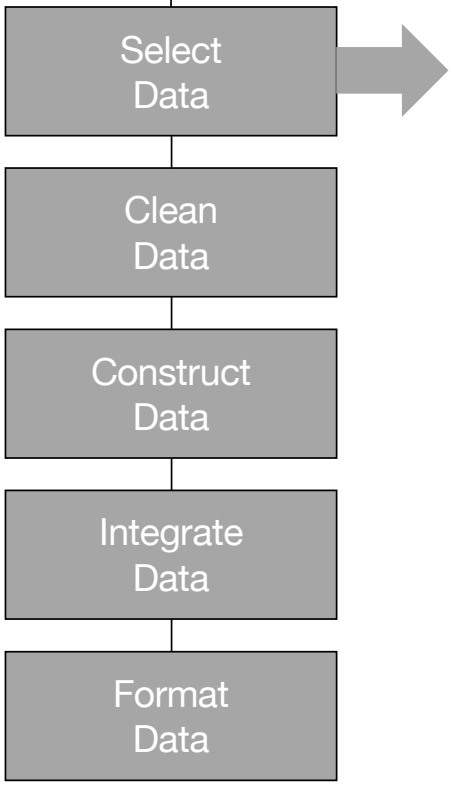
Data preparation:

- Collection
- Assessment
- Consolidation and Cleaning
  - table links, aggregation level, missing values, etc
- Data selection
  - active role in ignoring non-contributory data?
  - outliers?
  - Use of samples
  - visualization tools
- Transformations - create new variables

- Covers all activities to construct the final dataset 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 as well as transformation and cleaning of data for modeling tools.

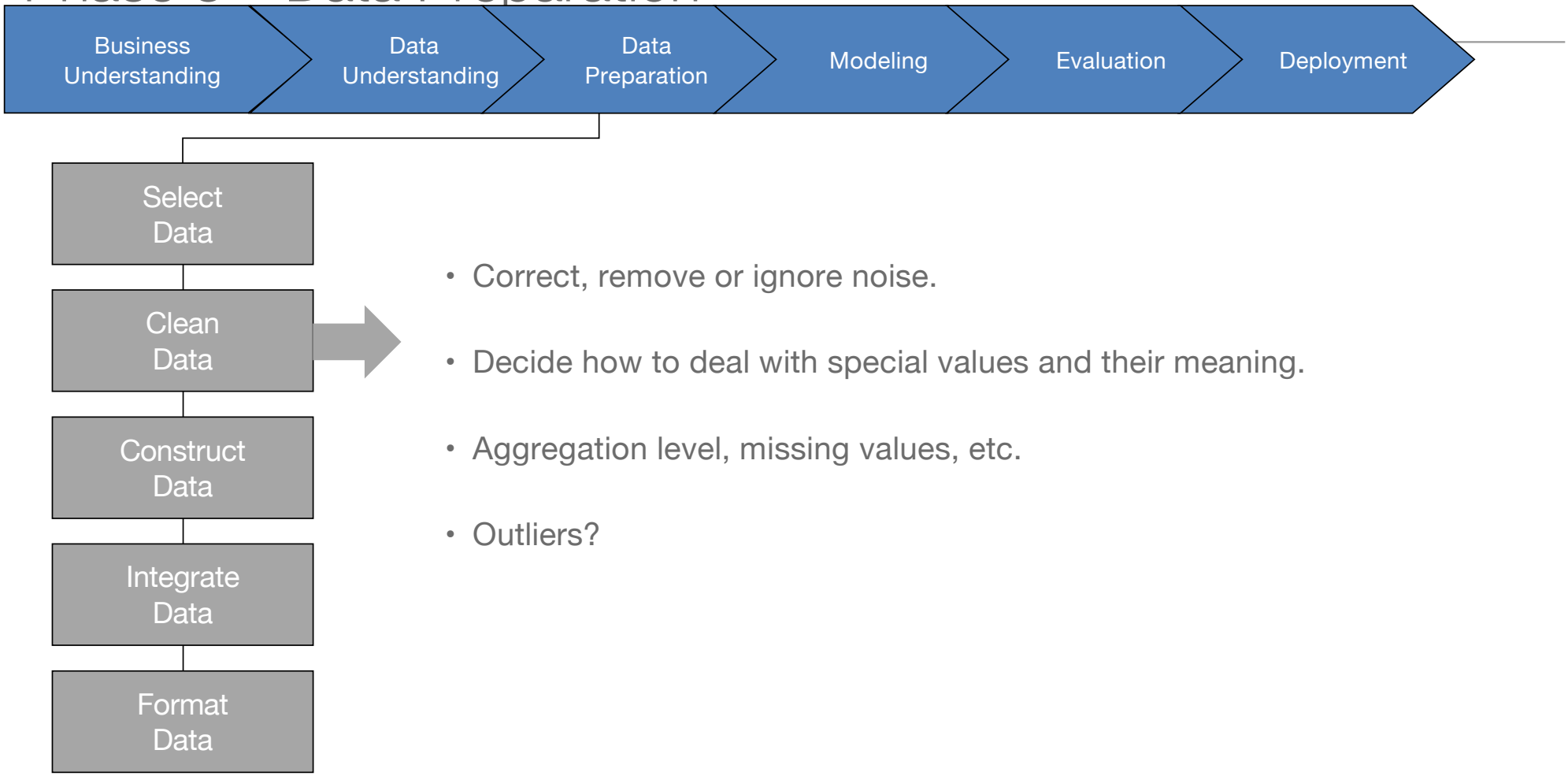
Phase 3 – Data Preparation





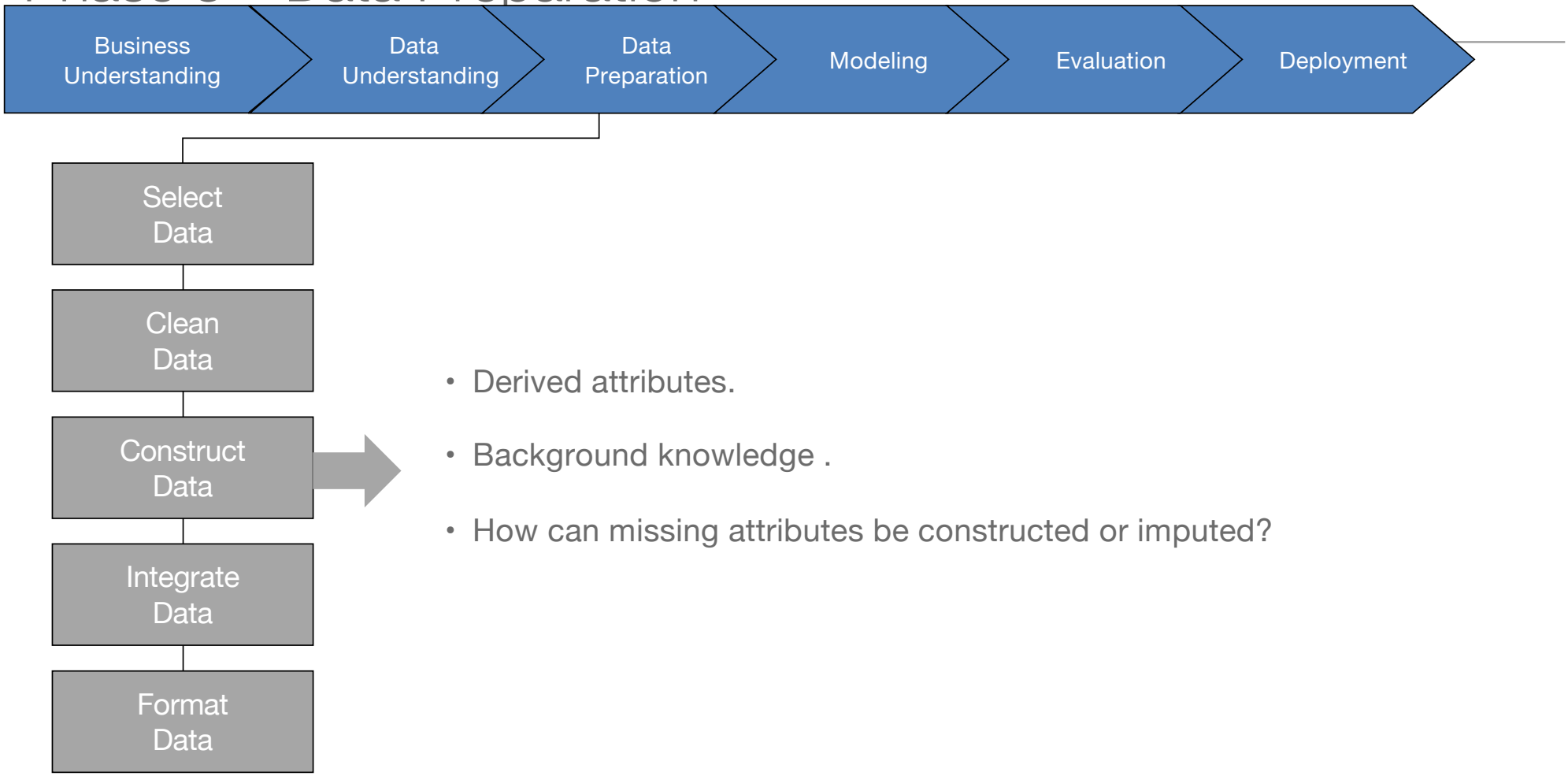
- Reconsider data selection criteria.
- Decide which dataset will be used.
- Collect appropriate additional data (internal or external).
- Consider use of sampling techniques.
- Explain why certain data was included or excluded.

Phase 3 – Data Preparation



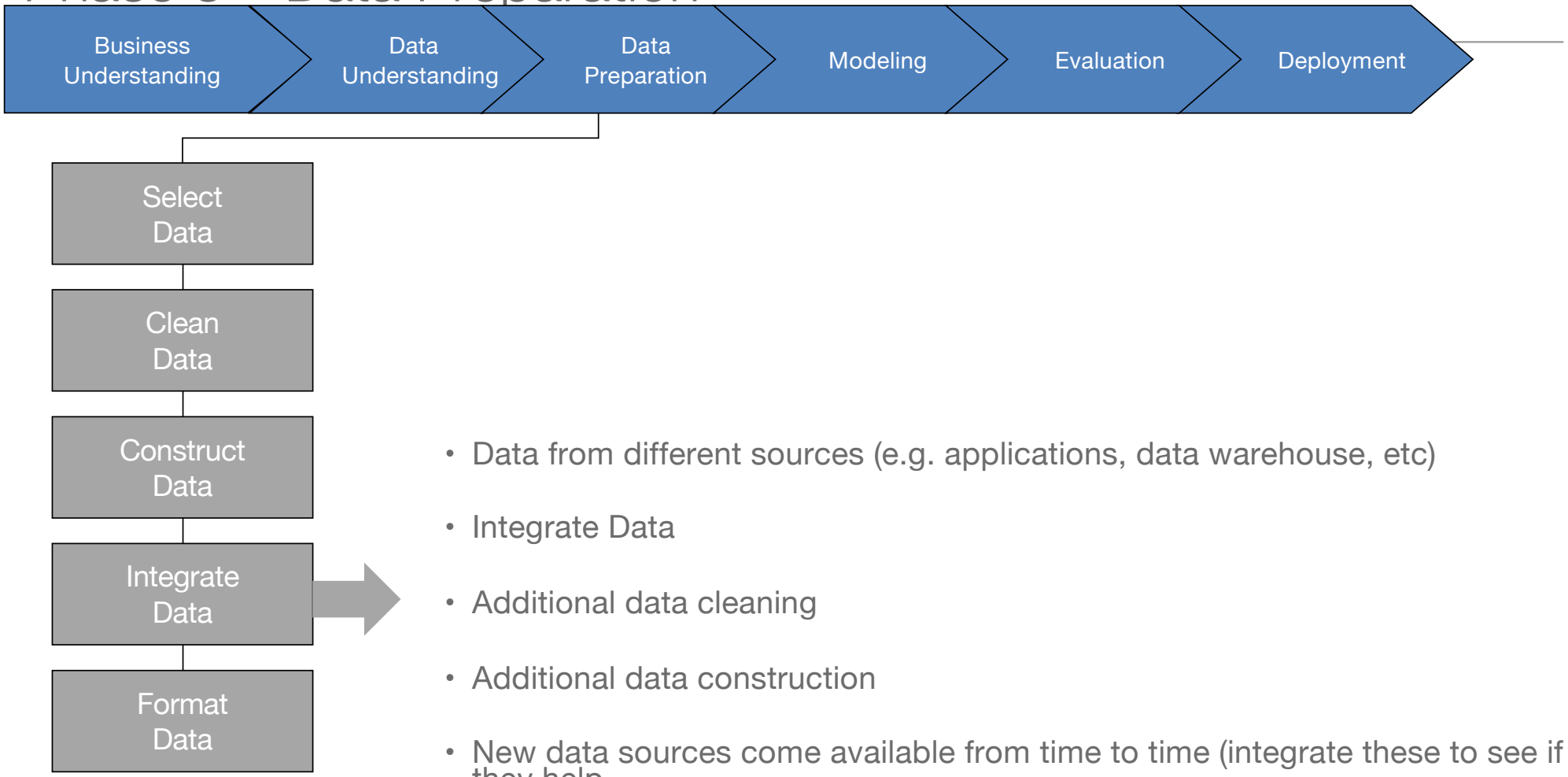
- Correct, remove or ignore noise.
- Decide how to deal with special values and their meaning.
- Aggregation level, missing values, etc.
- Outliers?

Phase 3 – Data Preparation



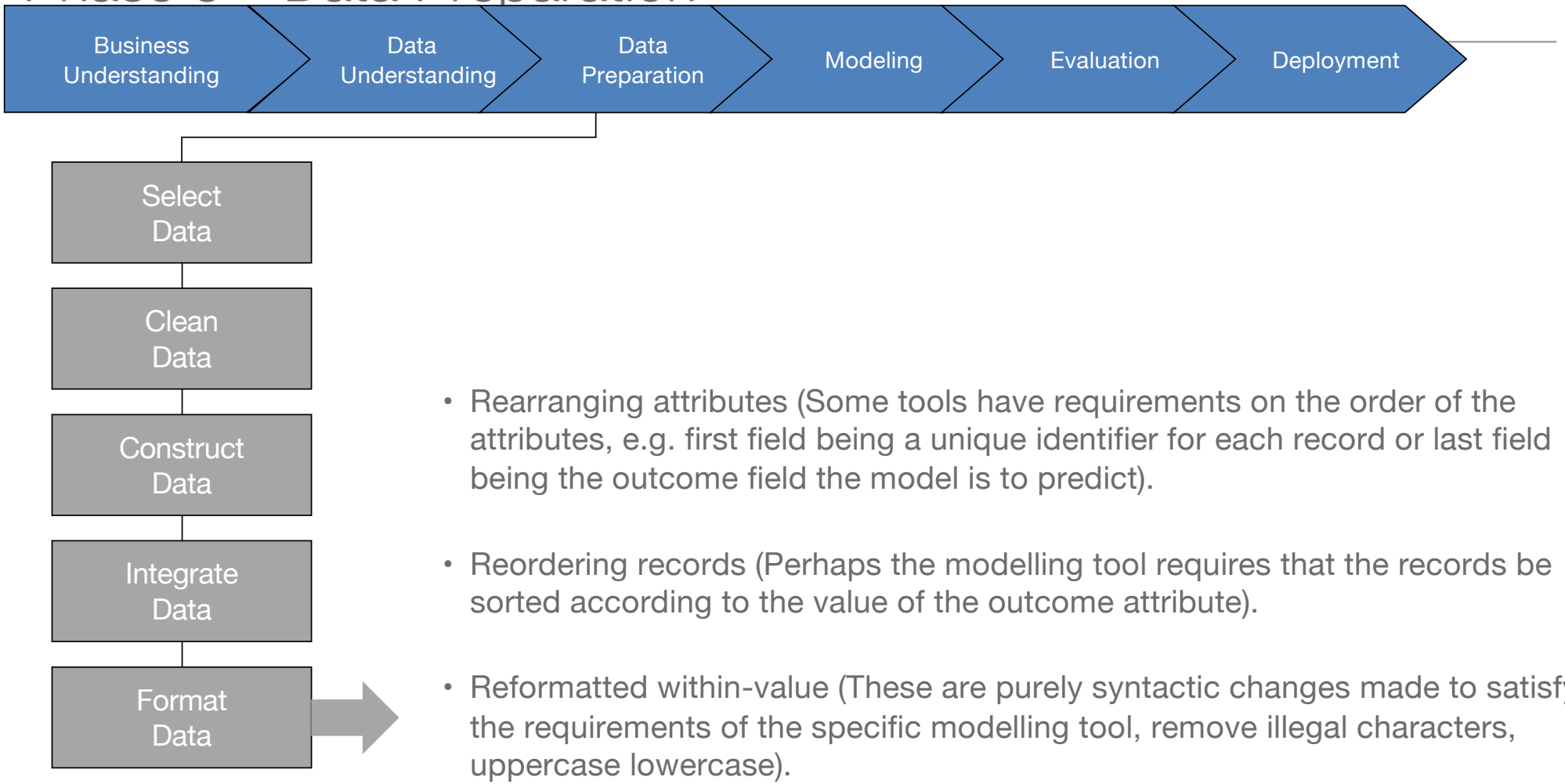
- Derived attributes.
- Background knowledge .
- How can missing attributes be constructed or imputed?

Phase 3 – Data Preparation



- Data from different sources (e.g. applications, data warehouse, etc)
- Integrate Data
- Additional data cleaning
- Additional data construction
- New data sources come available from time to time (integrate these to see if they help

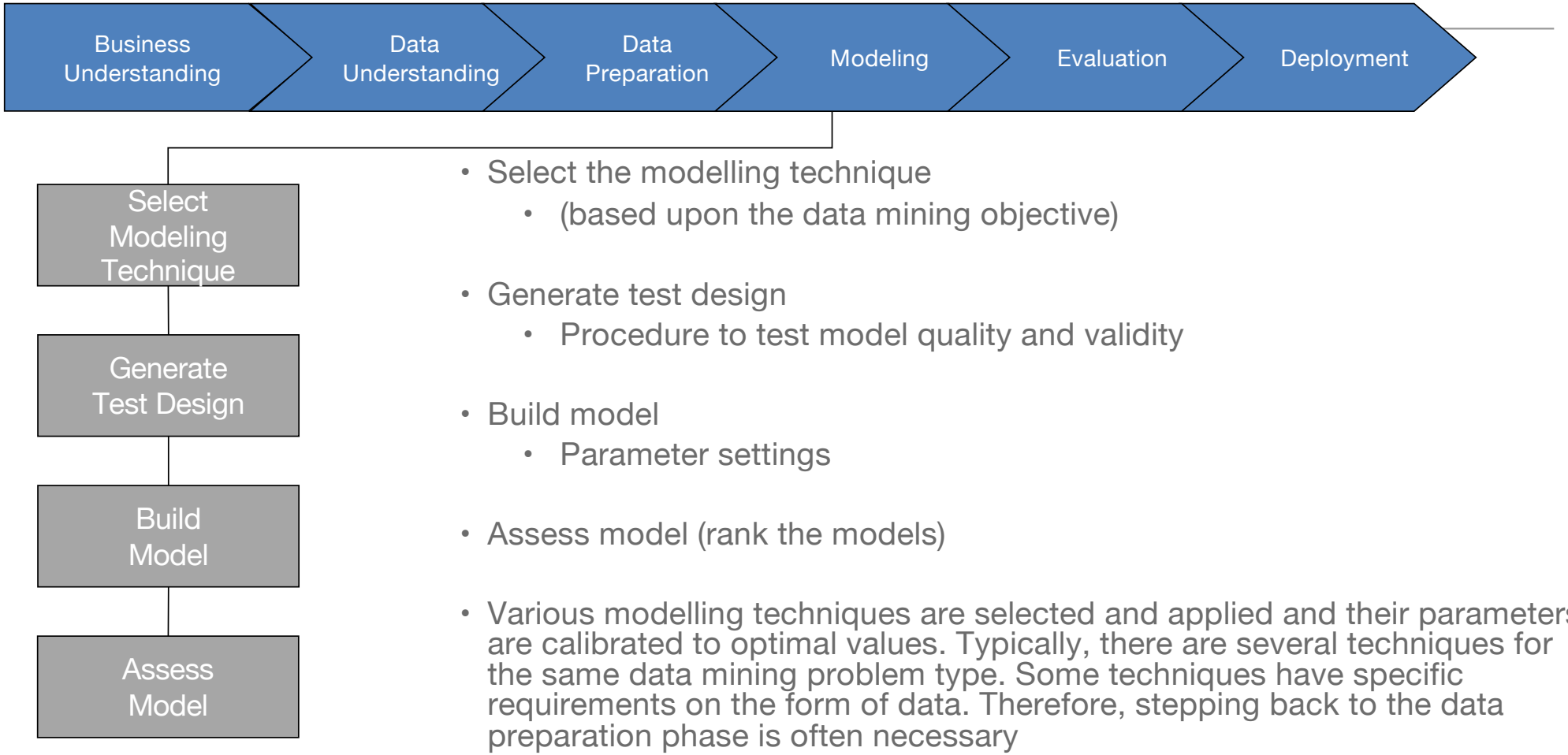
Phase 3 – Data Preparation



- Rearranging attributes (Some tools have requirements on the order of the attributes, e.g. first field being a unique identifier for each record or last field being the outcome field the model is to predict).
- Reordering records (Perhaps the modelling tool requires that the records be sorted according to the value of the outcome attribute).
- Reformatted within-value (These are purely syntactic changes made to satisfy the requirements of the specific modelling tool, remove illegal characters, uppercase lowercase).

Ah this is easy! It really is!

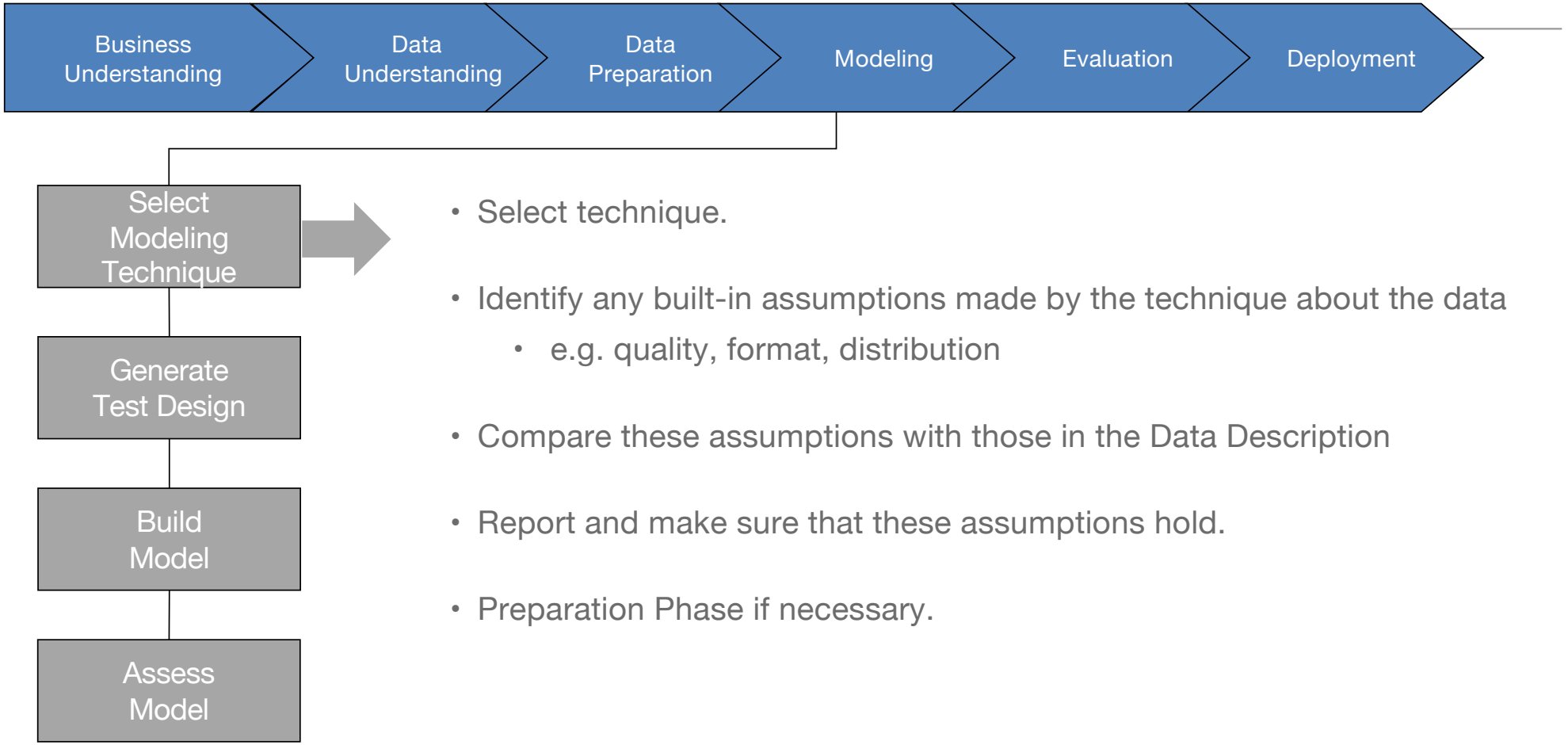
Phase 4 - Modelling



- Select the modelling technique
  - (based upon the data mining objective)
- Generate test design
  - Procedure to test model quality and validity
- Build model
  - Parameter settings
- Assess model (rank the models)
- Various modelling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often necessary

Ah this is easy! It really is!

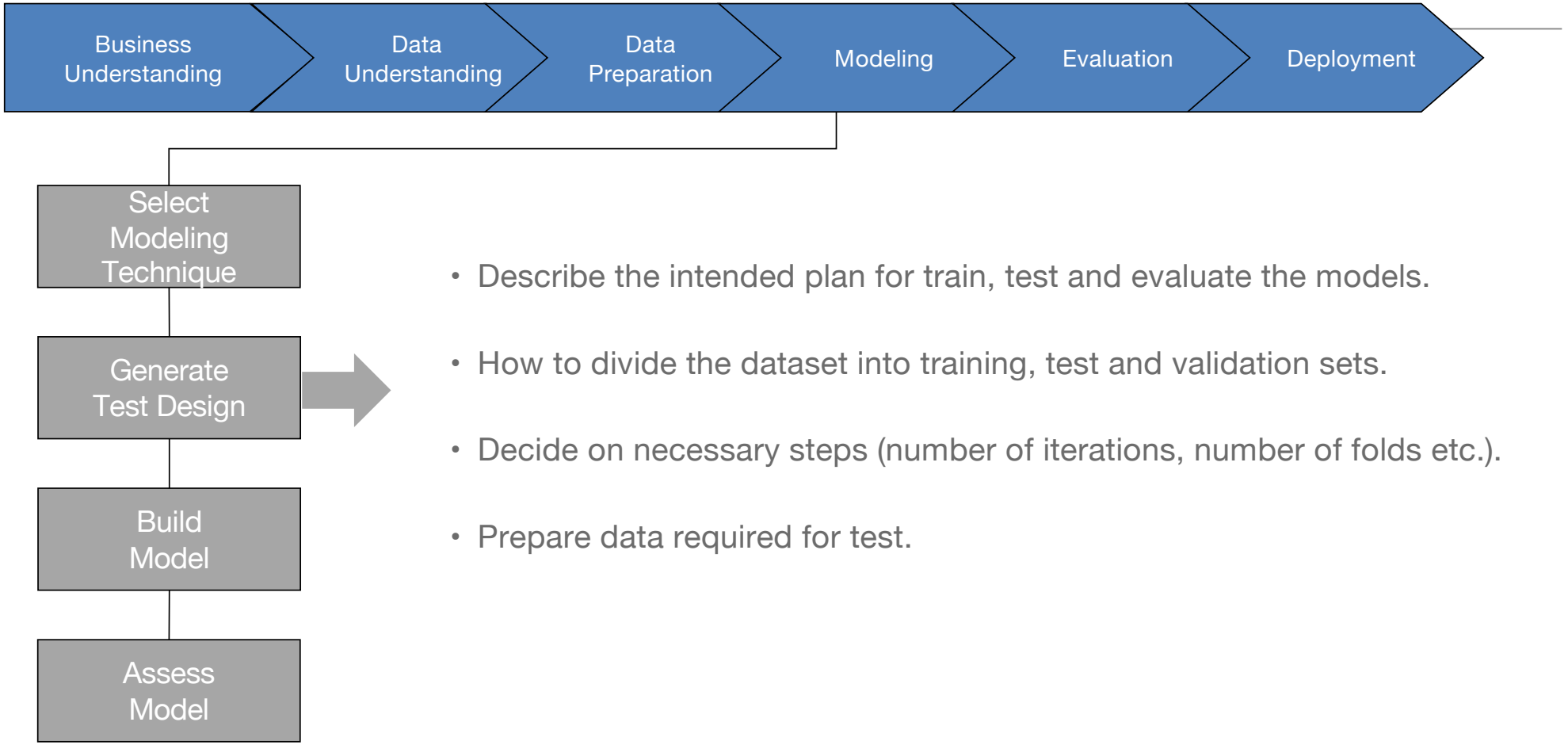
Phase 4 - Modelling



- Select technique.
- Identify any built-in assumptions made by the technique about the data
  - e.g. quality, format, distribution
- Compare these assumptions with those in the Data Description
- Report and make sure that these assumptions hold.
- Preparation Phase if necessary.

Ah this is easy! It really is!

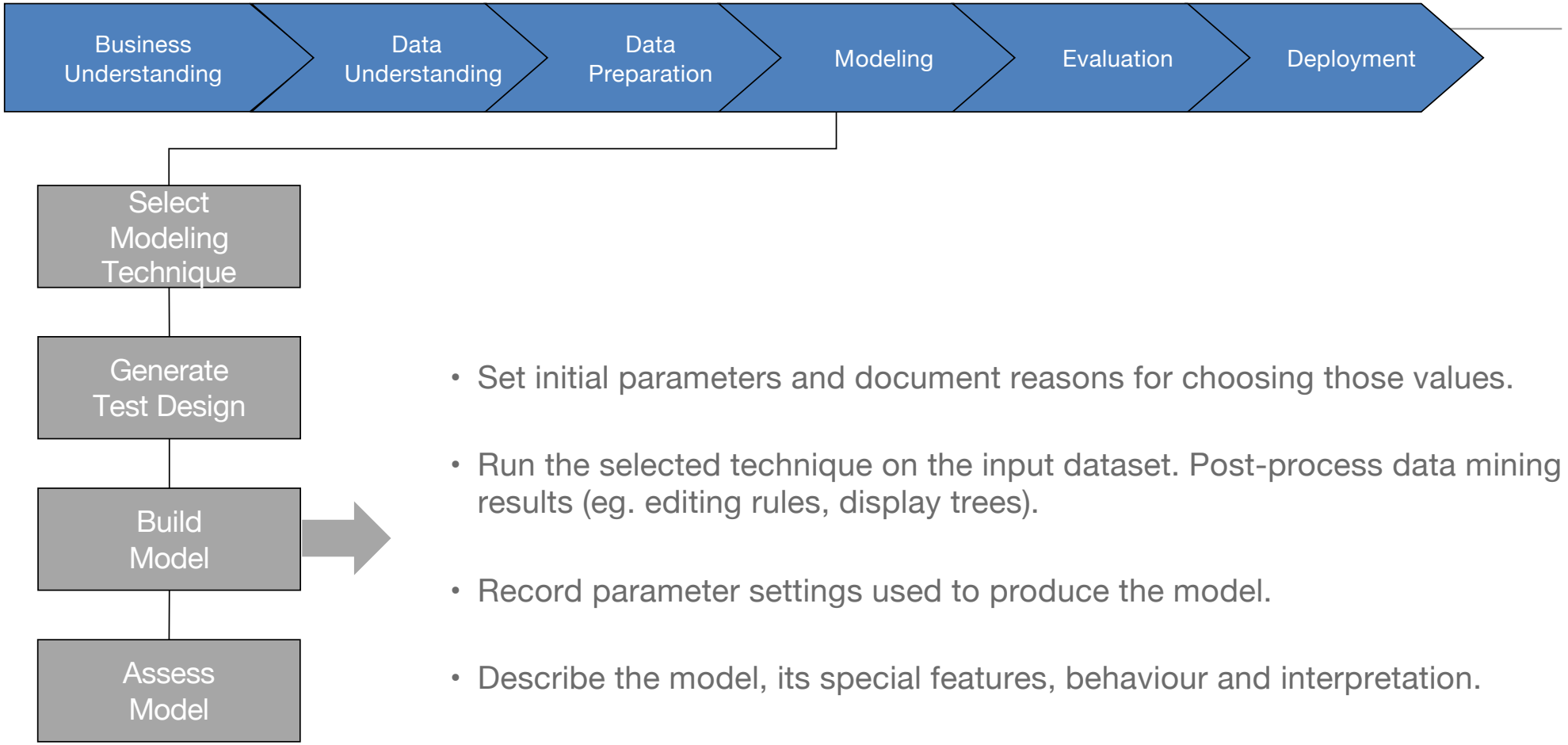
Phase 4 - Modelling



- Describe the intended plan for train, test and evaluate the models.
- How to divide the dataset into training, test and validation sets.
- Decide on necessary steps (number of iterations, number of folds etc.).
- Prepare data required for test.

Ah this is easy! It really is!

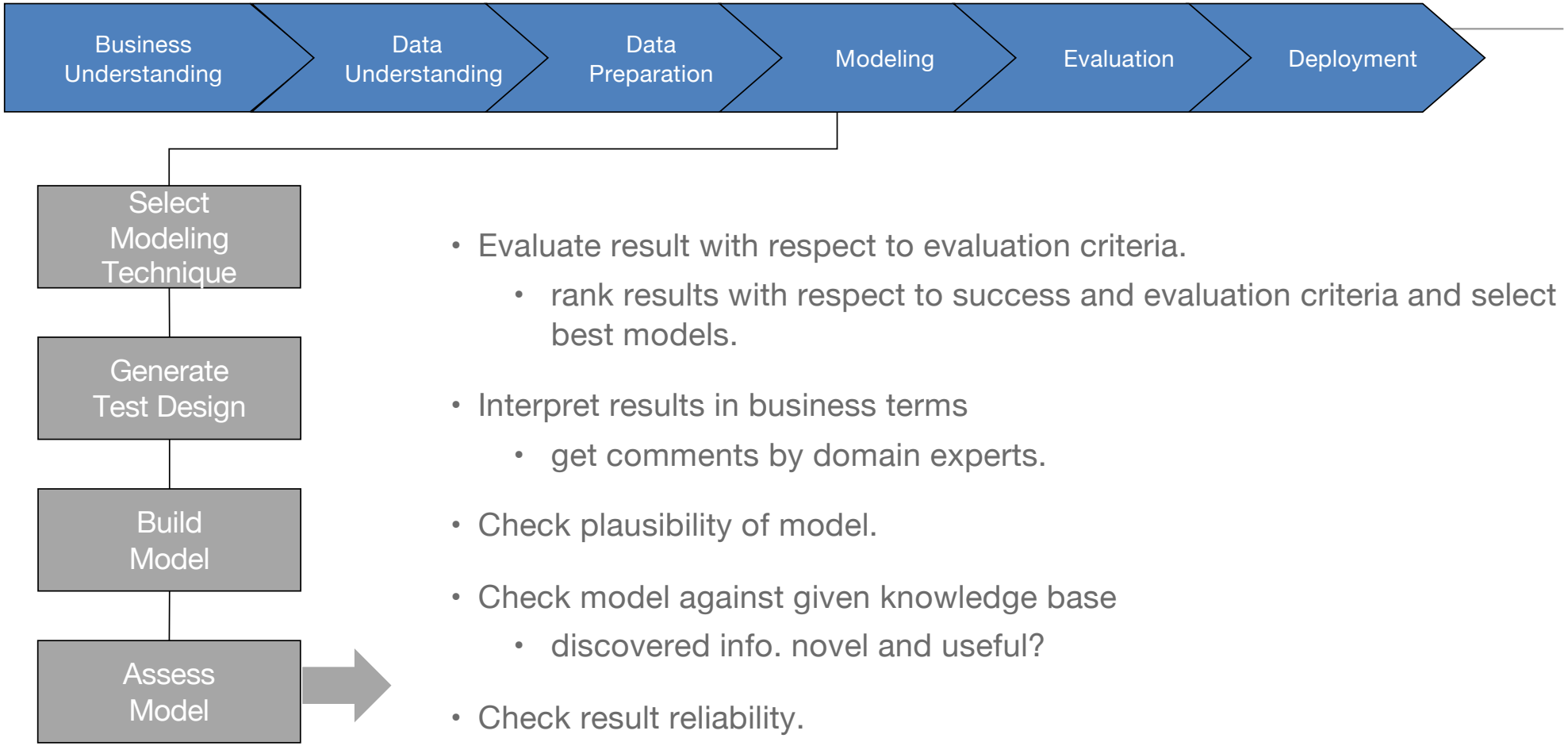
Phase 4 - Modelling



- Set initial parameters and document reasons for choosing those values.
- Run the selected technique on the input dataset. Post-process data mining results (eg. editing rules, display trees).
- Record parameter settings used to produce the model.
- Describe the model, its special features, behaviour and interpretation.

Ah this is easy! It really is!

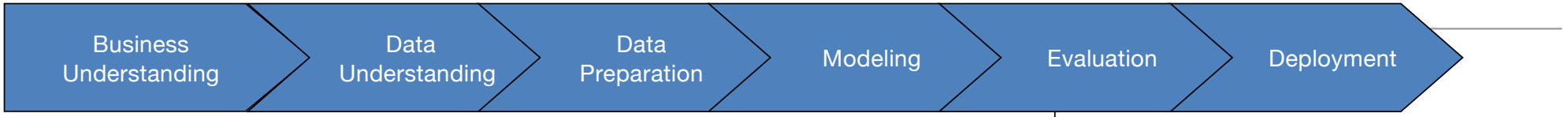
Phase 4 - Modelling

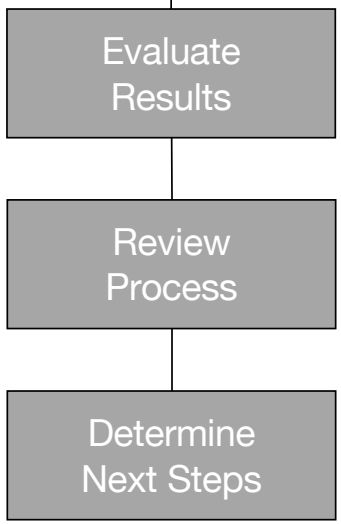


- Evaluate result with respect to evaluation criteria.
  - rank results with respect to success and evaluation criteria and select best models.
- Interpret results in business terms
  - get comments by domain experts.
- Check plausibility of model.
- Check model against given knowledge base
  - discovered info. novel and useful?
- Check result reliability.
  - Analyze potentials for deployment of each result.

But this can be difficult.
What does it all mean!

Phase 5 – Evaluation



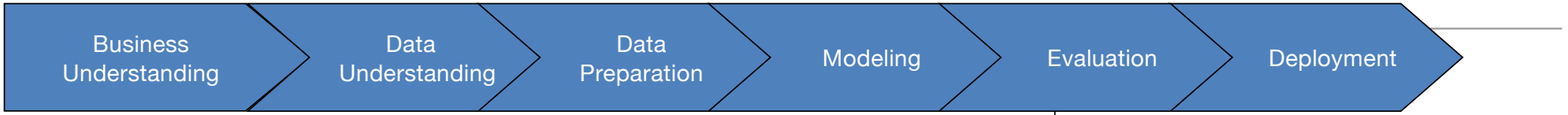


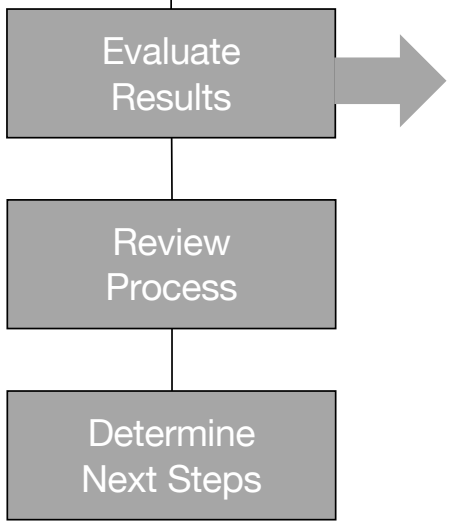
- More thoroughly evaluate model
- Decide how to use results
- Methods and criteria depend on model type:
  - e.g., coincidence matrix with classification models, mean error rate with regression models.
- Interpretation of model: important or not, easy or hard depends on algorithm.

- Thoroughly evaluate the model and review the steps executed to construct the model to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

But this can be difficult.
What does it all mean!

Phase 5 – Evaluation

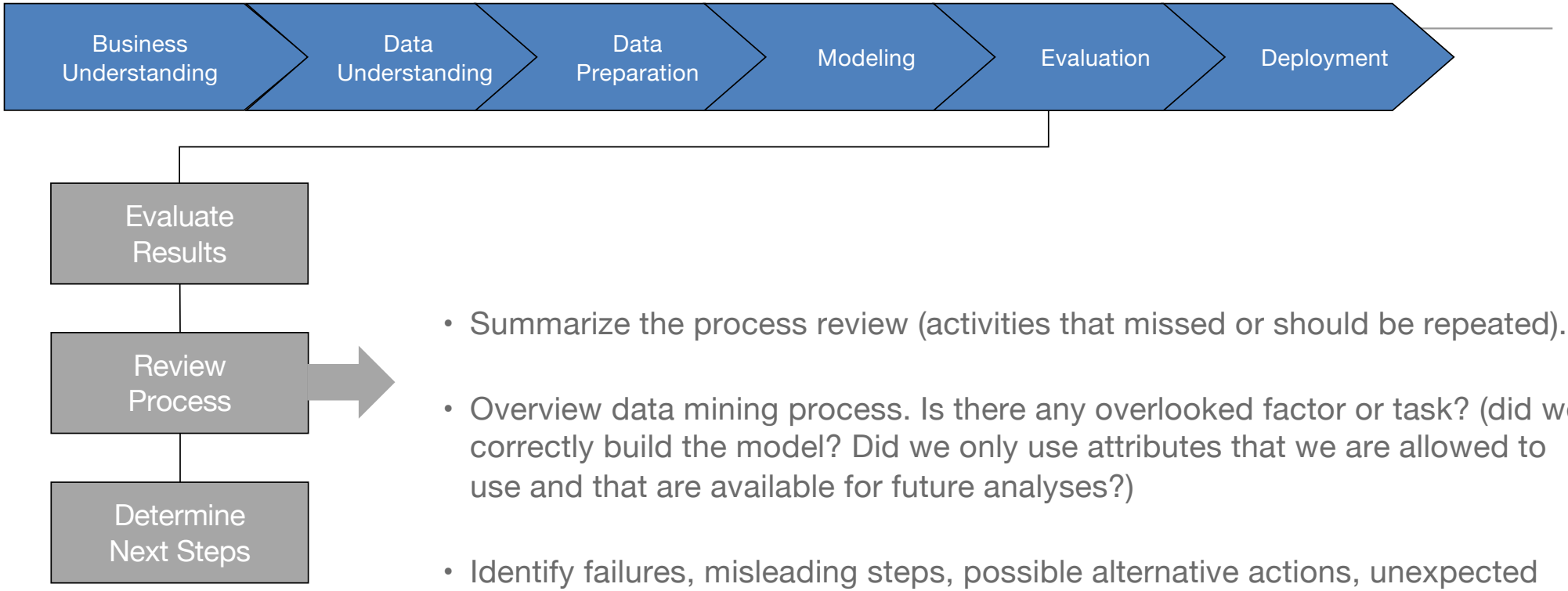




- Understand data mining result. Check impact for data mining goal.
- Check result against knowledge base to see if it is novel and useful.
- Evaluate and assess result with respect to business success criteria.
- Rank results according to business success criteria. Check result impact on initial application goal.
- Are there new business objectives? (address later in project or new project?)
- State conclusions for future data mining projects.

But this can be difficult.
What does it all mean!

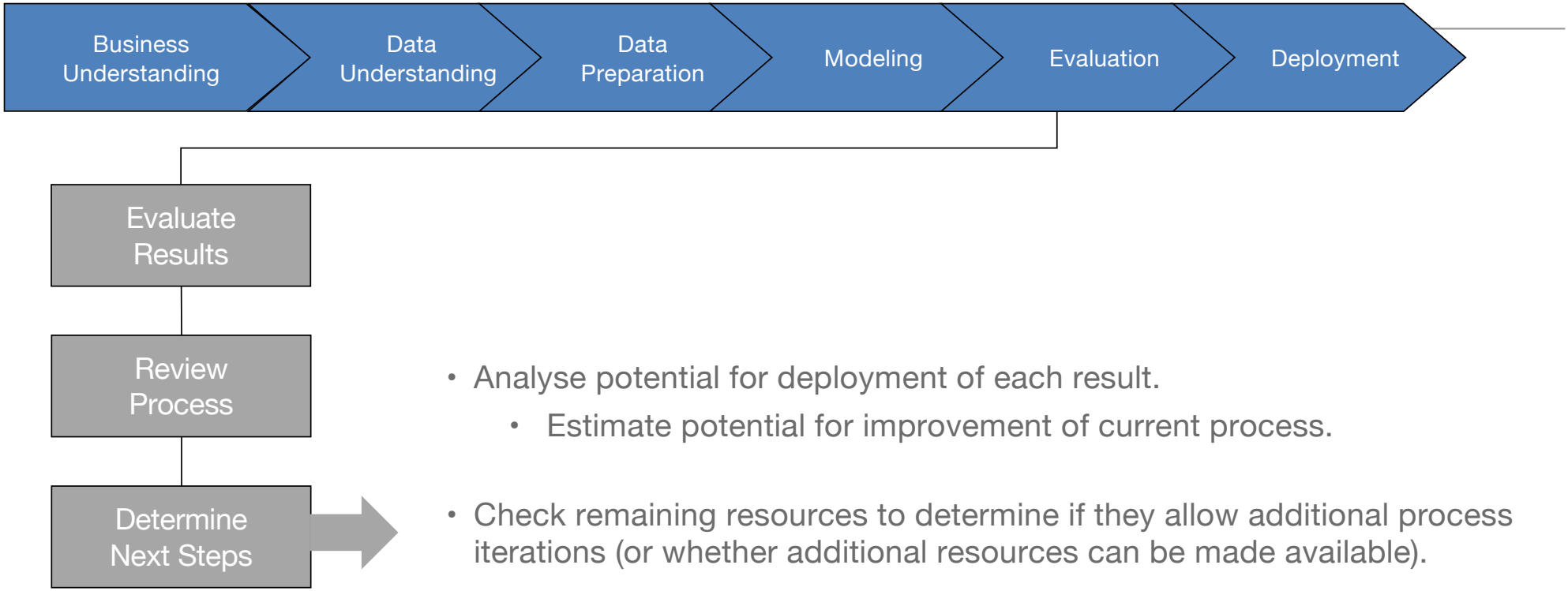
Phase 5 – Evaluation



- Summarize the process review (activities that missed or should be repeated).
- Overview data mining process. Is there any overlooked factor or task? (did we correctly build the model? Did we only use attributes that we are allowed to use and that are available for future analyses?)
- Identify failures, misleading steps, possible alternative actions, unexpected paths.
- Review data mining results with respect to business success.

But this can be difficult.
What does it all mean!

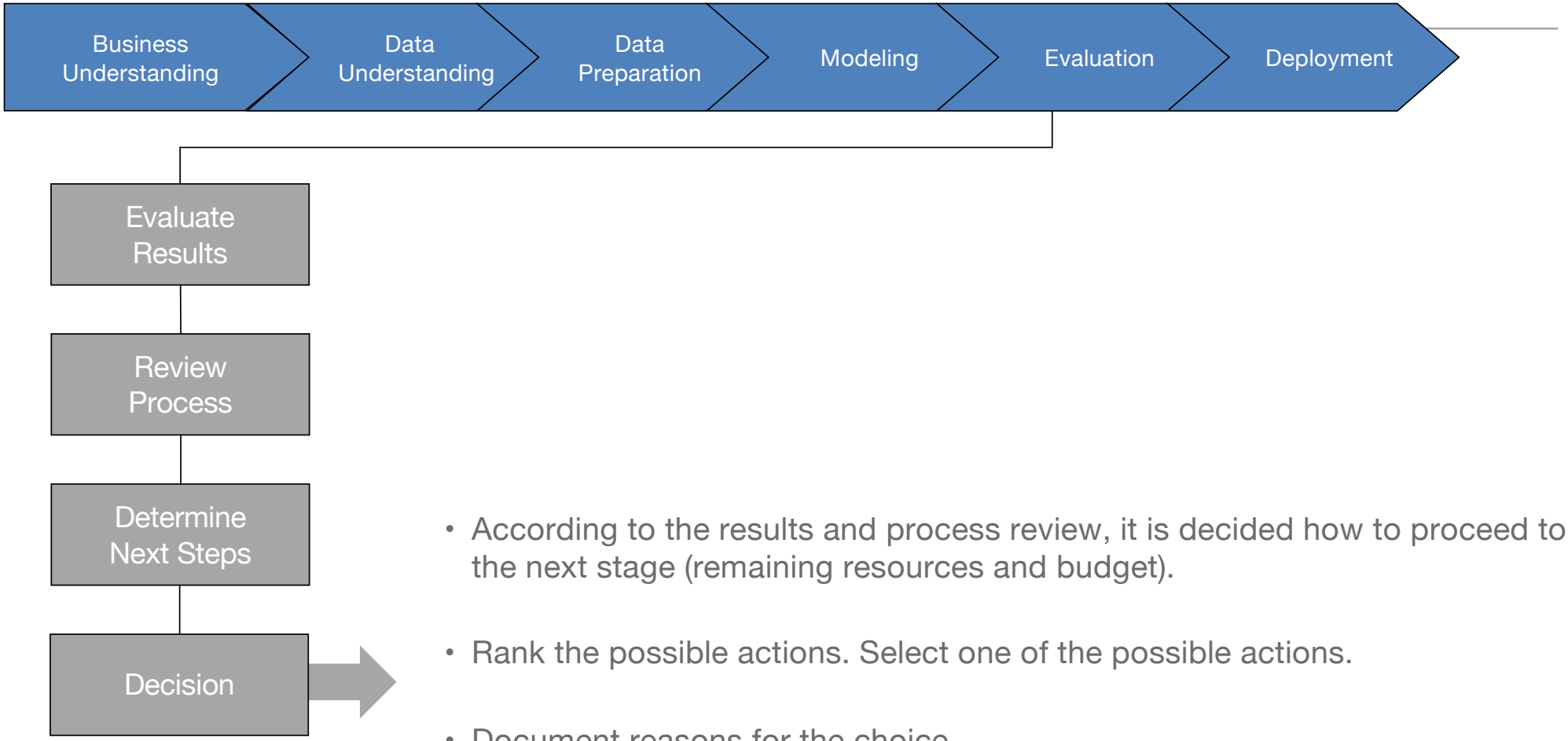
Phase 5 – Evaluation



- Analyse potential for deployment of each result.
  - Estimate potential for improvement of current process.
- Check remaining resources to determine if they allow additional process iterations (or whether additional resources can be made available).
- Recommend alternative continuations.
  - Refine process plan.

But this can be difficult.
What does it all mean!

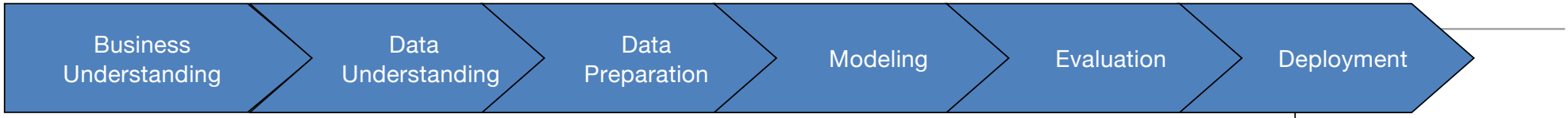
Phase 5 – Evaluation

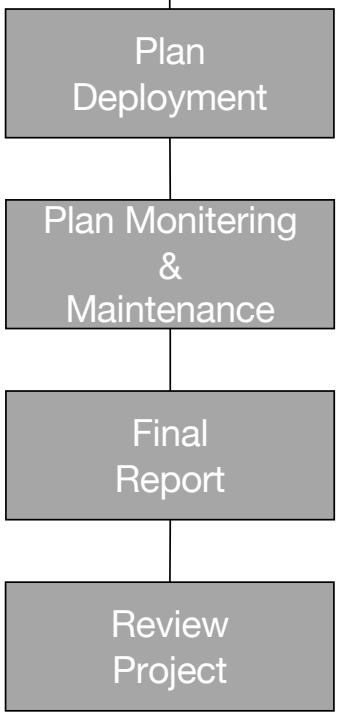


- According to the results and process review, it is decided how to proceed to the next stage (remaining resources and budget).
- Rank the possible actions. Select one of the possible actions.
- Document reasons for the choice.

No one talks about this.
I wonder why?

Phases 6 - Deployment

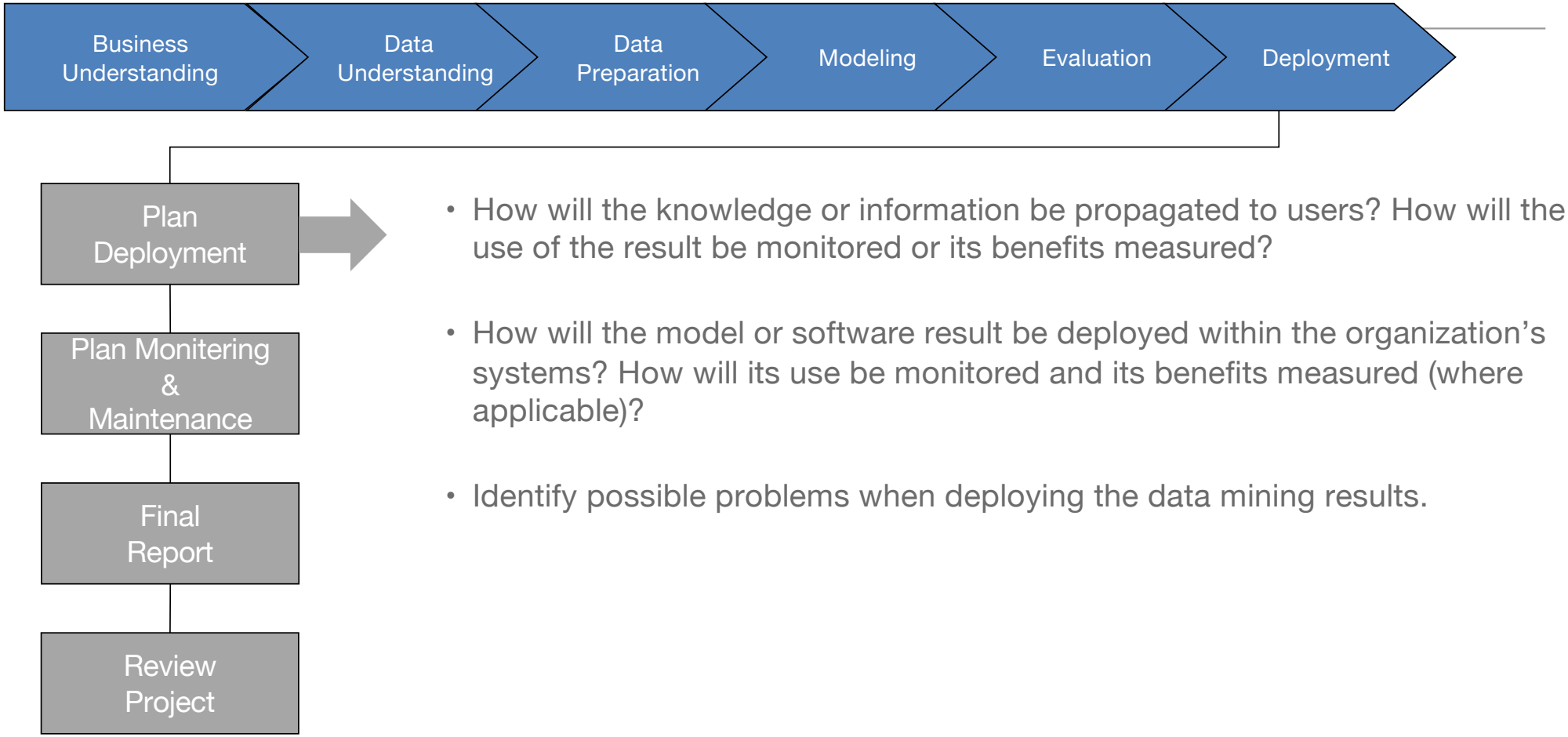




- Deployment
  - Determine how the results need to be utilized
  - Who needs to use them?
  - How often do they need to be used
- Deploy Data Mining results by:
  - Scoring a database, utilising results as business rules, interactive scoring on-line.
  - The knowledge gained will need to be organized and presented in a way that the customer can use it. However, depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

No one talks about this.
I wonder why?

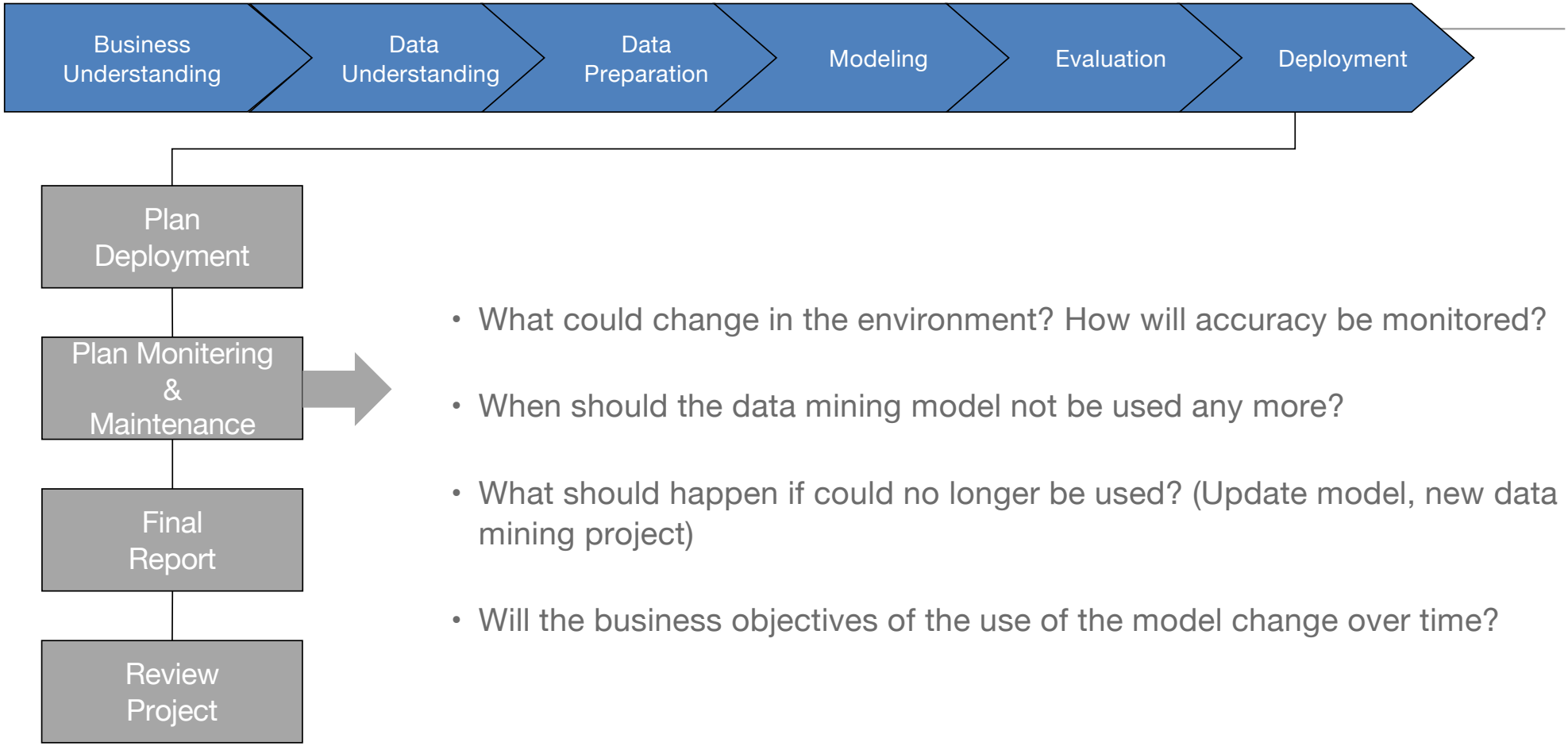
Phases 6 - Deployment



- How will the knowledge or information be propagated to users? How will the use of the result be monitored or its benefits measured?
- How will the model or software result be deployed within the organization's systems? How will its use be monitored and its benefits measured (where applicable)?
- Identify possible problems when deploying the data mining results.

No one talks about this.
I wonder why?

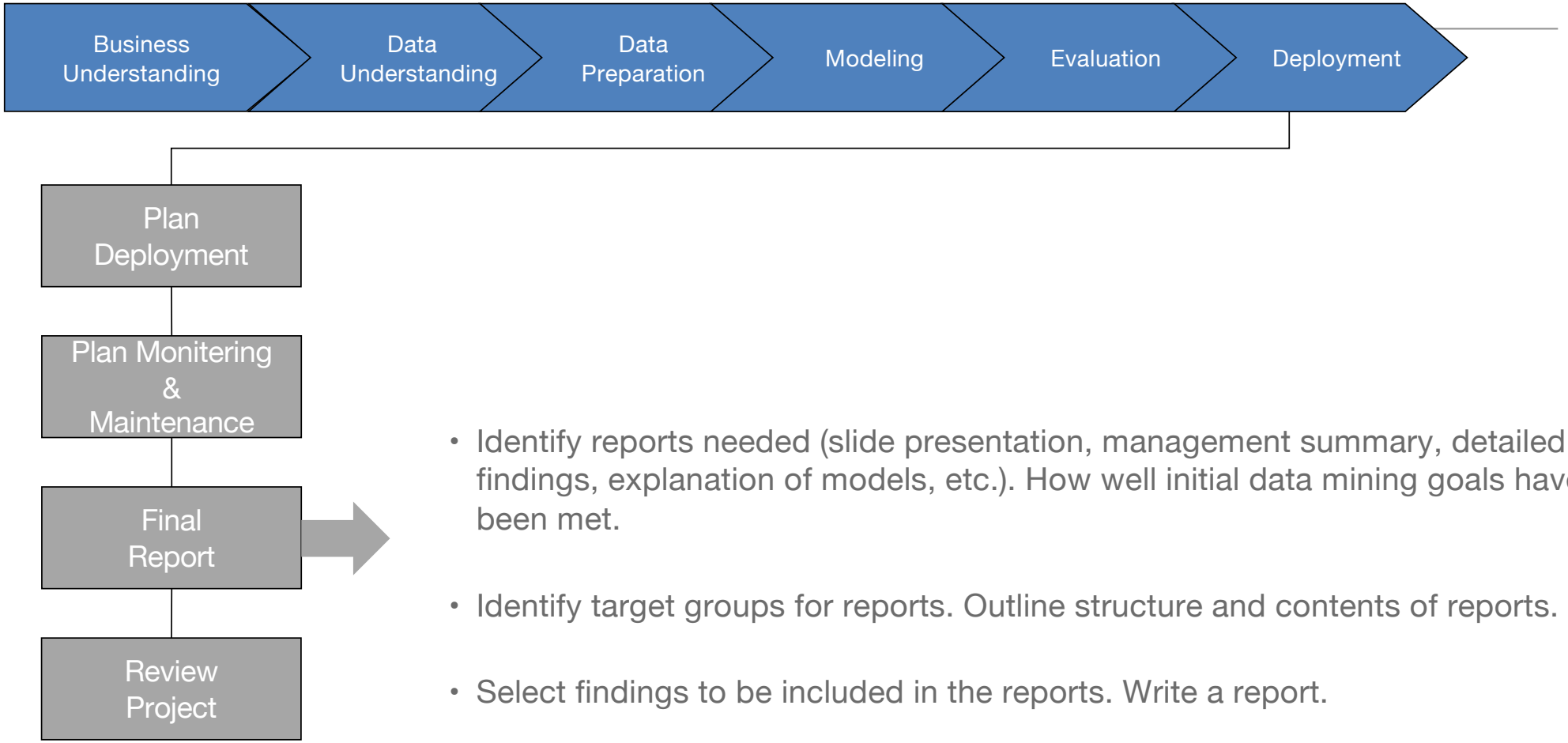
Phases 6 - Deployment



- What could change in the environment? How will accuracy be monitored?
- When should the data mining model not be used any more?
- What should happen if could no longer be used? (Update model, new data mining project)
- Will the business objectives of the use of the model change over time?

No one talks about this.
I wonder why?

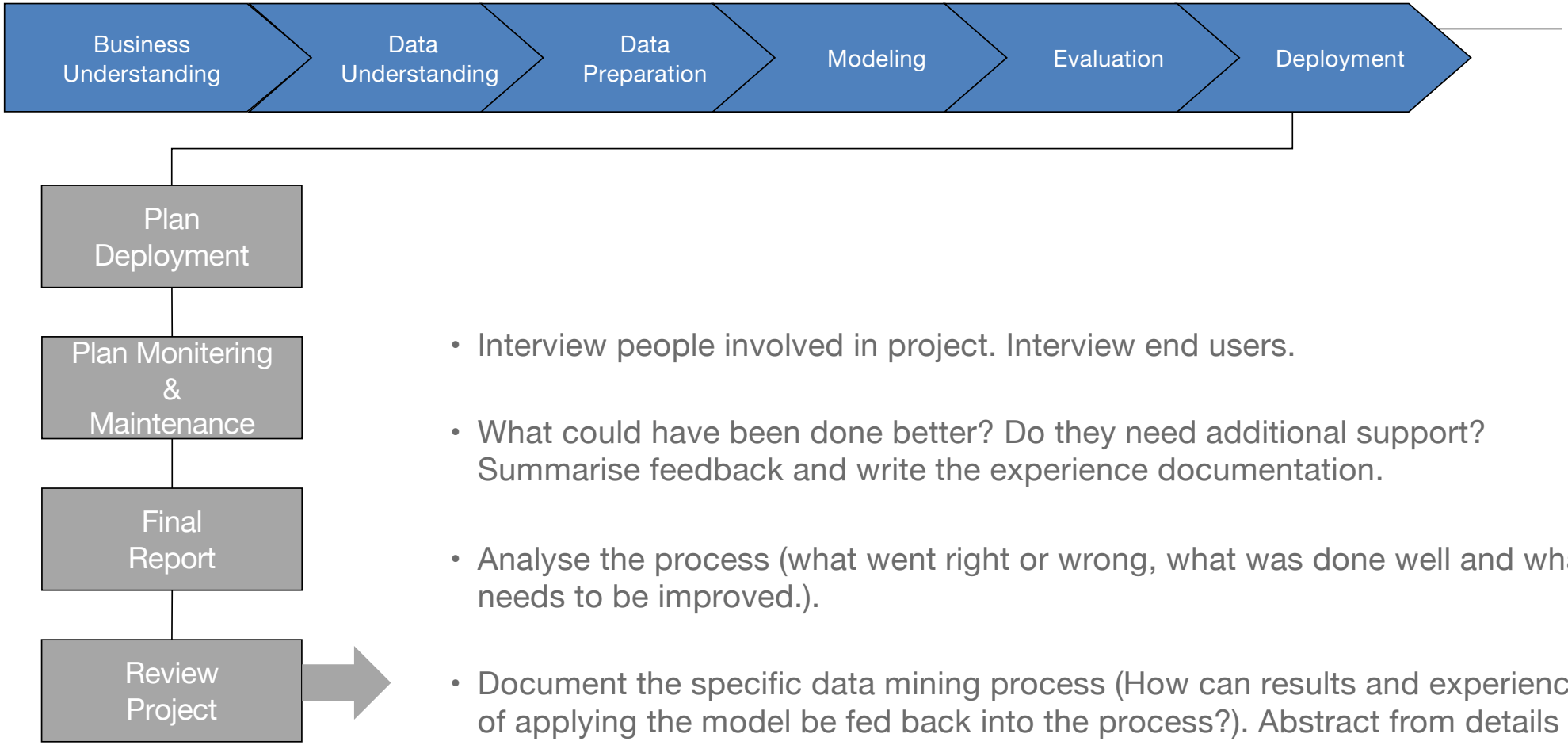
Phases 6 - Deployment



- Identify reports needed (slide presentation, management summary, detailed findings, explanation of models, etc.). How well initial data mining goals have been met.
- Identify target groups for reports. Outline structure and contents of reports.
- Select findings to be included in the reports. Write a report.

No one talks about this.
I wonder why?

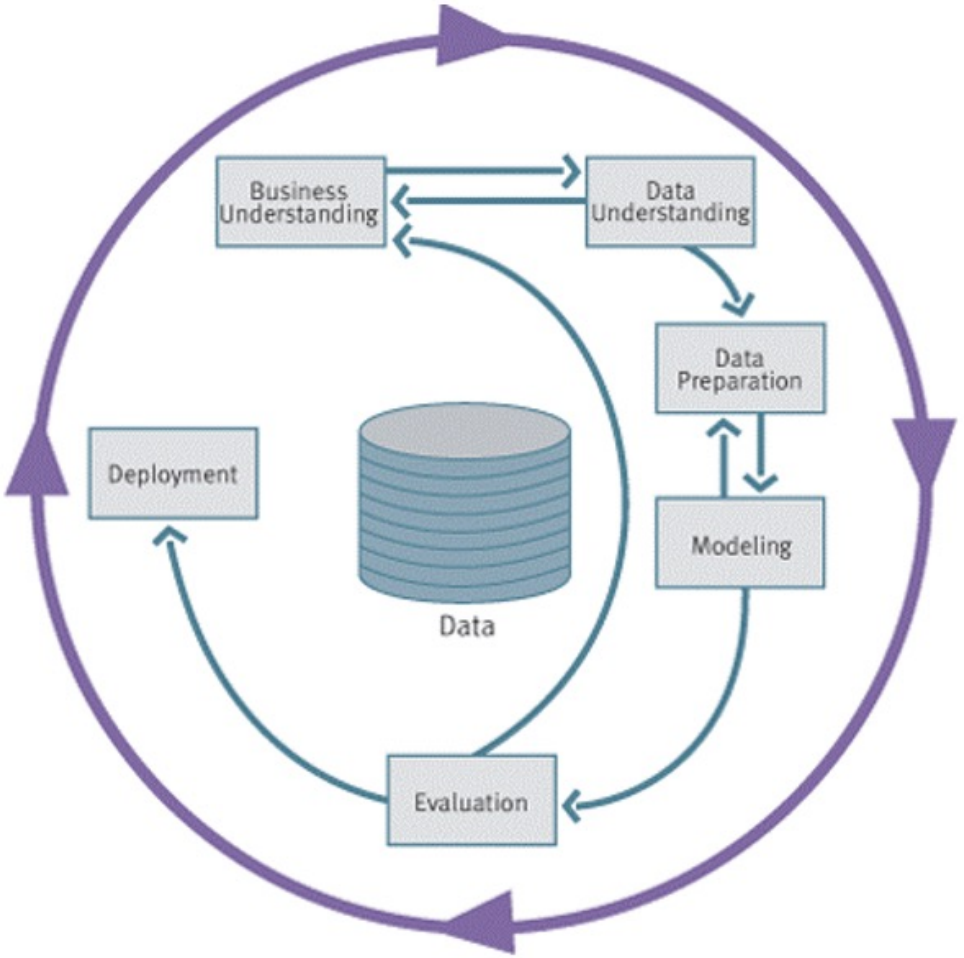
Phases 6 - Deployment

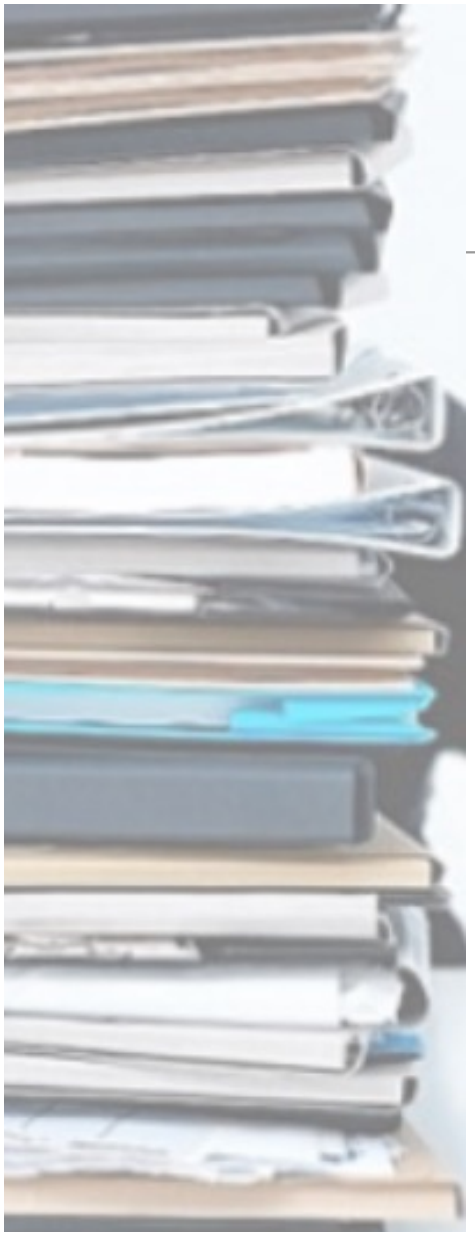


- Interview people involved in project. Interview end users.
- What could have been done better? Do they need additional support? Summarise feedback and write the experience documentation.
- Analyse the process (what went right or wrong, what was done well and what needs to be improved.).
- Document the specific data mining process (How can results and experience of applying the model be fed back into the process?). Abstract from details to make the experience useful for future projects.

CRISP-DM

- Non-proprietary
- Application/Industry neutral
- Tool neutral
- Focus on business issues
  - As well as technical analysis
- Framework for guidance
- Experience base
  - Templates for Analysis





Documenting your Work

- Why do we document our work?
- What is different about Data Science / AI / ML Projects?
  - Iterative
  - Gaps between phases and Iterations
  - People move on
- Traditionally CRISP-DM give a framework for Documenting each Phase of the project
  - What was done
  - Why it was done that way
  - Results and Outcomes
- Can compare Changes and Results with each iteration

Any Questions ?

What Now/Next ?

Let's Start with some Basics

- There will be some overlap with other modules.
  - Data Wrangling
  - Information Systems
  - Etc
- For all projects, we need to
  - Have a data set
  - Load the data set into your environment
  - Perform some analysis of this data
    - Where do you begin?
    - What are the business questions
    - What domain knowledge do you have or can apply based on your experience
    - Step – by – Step
    - Some Stats
    - Some Charts
    - Analyse the data to build a picture