CSCI446/946 Big Data Analytics

Week 2 Data Analytics Lifecycle

School of Computing and Information Technology
University of Wollongong Australia

Brief Recap

- Big Data Definition: 4Vs, 5Vs.
- Big Data vs. Data Mining.
- When is data Big?
- Role of Data scientist.
- Data: Structured vs. unstructured.
- Business Intelligence vs Data science.

Data Science Projects

- Different from traditional business intelligence projects.
  - Data Science is more exploratory in nature!
- It is critical to have a process to govern them.
- Common mistake
  - Rushing into data collection and analysis.
  - Not spend enough time planning, scoping, understanding, or framing.

Data Analytics Lifecycle Overview

- Key Roles for a Successful Analytics Project:
  - Data Scientist: The Sexiest Job of this Century ☺
  - Seven key roles for a data science team
    - Business User; Project Sponsor; Project Manager
    - Business Intelligence Analyst
    - Database Administrator
    - Data Engineer; Data Scientist
  - The last two roles are in high demand!

Data Analytics Lifecycle Overview

- Key Roles for a Successful Analytics Project:

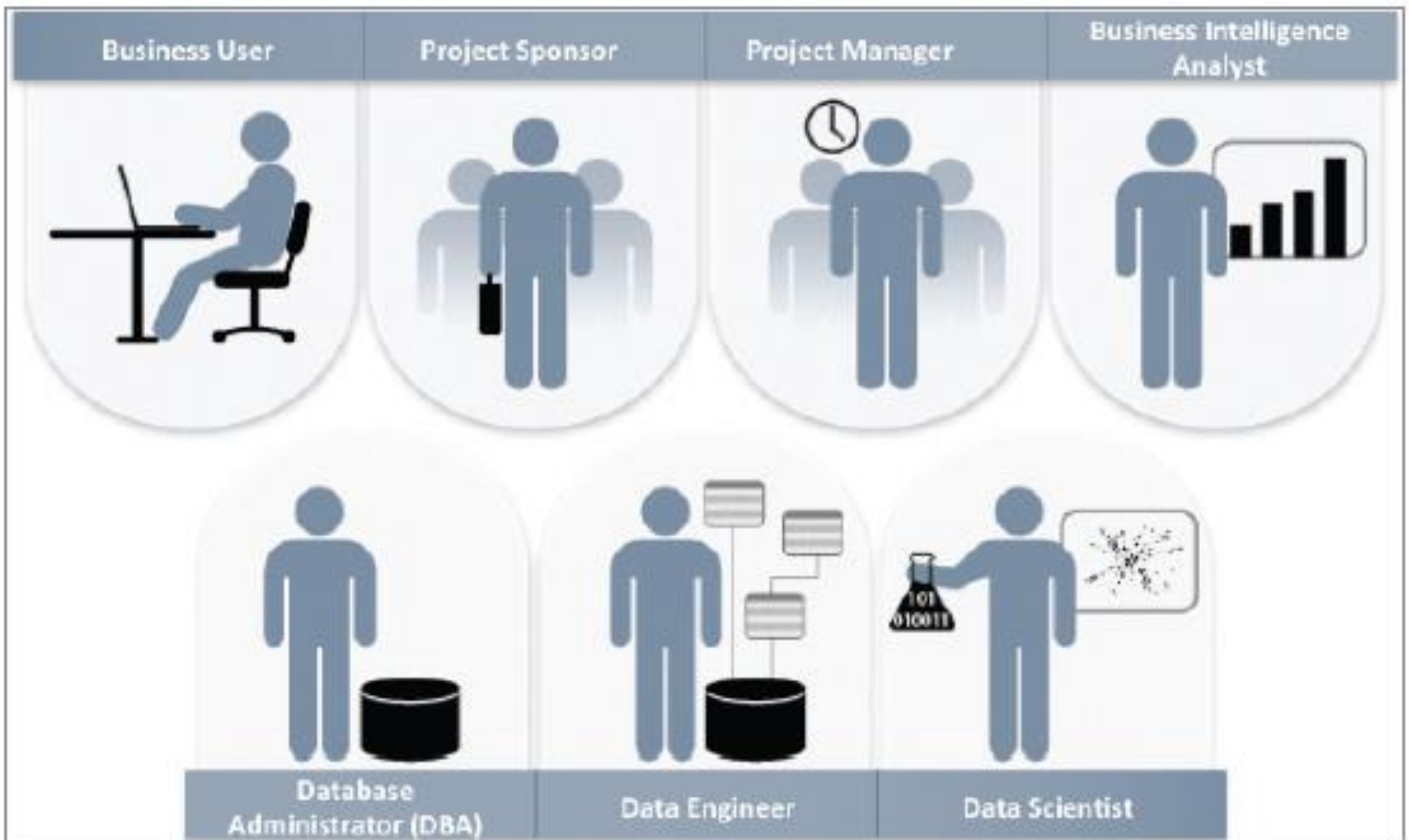


FIGURE 2-1 *Key roles for a successful analytics project*

Data Analytics Lifecycle Overview

- Communication between these key players is essential to the success of a data analytical problem.
- Problem: Key players can have a different background, use different terminologies and expressions, have different interests and goals.
  - Domain understanding is a first step towards successful communication.

An example

- A team of oncologists and radiotherapists wanted to know whether it is possible to predict the toxicity of a prescribed radiotherapy on healthy tissue from MRI scans.
- They approach a team of data scientists with this question.
- Data scientist:
  - A computing or IT specialist; may not properly understand medical terminologies, the needs of the client, …
  - May not understand what needs to be done to access such highly sensitive patient data. Understand data quality, and variation of data sources?
  - Data may not be labelled. Does not have the expertise to deduct from an MRI scan what constitutes toxic effects that arise out of radiotherapy.
  - May create a model which predicts whether or not toxic effects would occur. But clients want to know “where” the toxic effects occur, and “why” the model made such a prediction.
  - …
- To succeed, the data scientists have to obtain a good domain understanding.
  - This can require substantial background studies.
  - This first step is called the discovery phase of a data analytics lifecycle.

Data Analytics Lifecycle Overview

- Six phases
  1. Discovery
  2. Data preparation (analytic sandbox)
  3. Model planning (methods, techniques, workflow, variables, relationships, models)
  4. Model building (training and test datasets, software, and hardware)
  5. Communication results (identify key findings)
  6. Operationalize (delivery, pilot project)

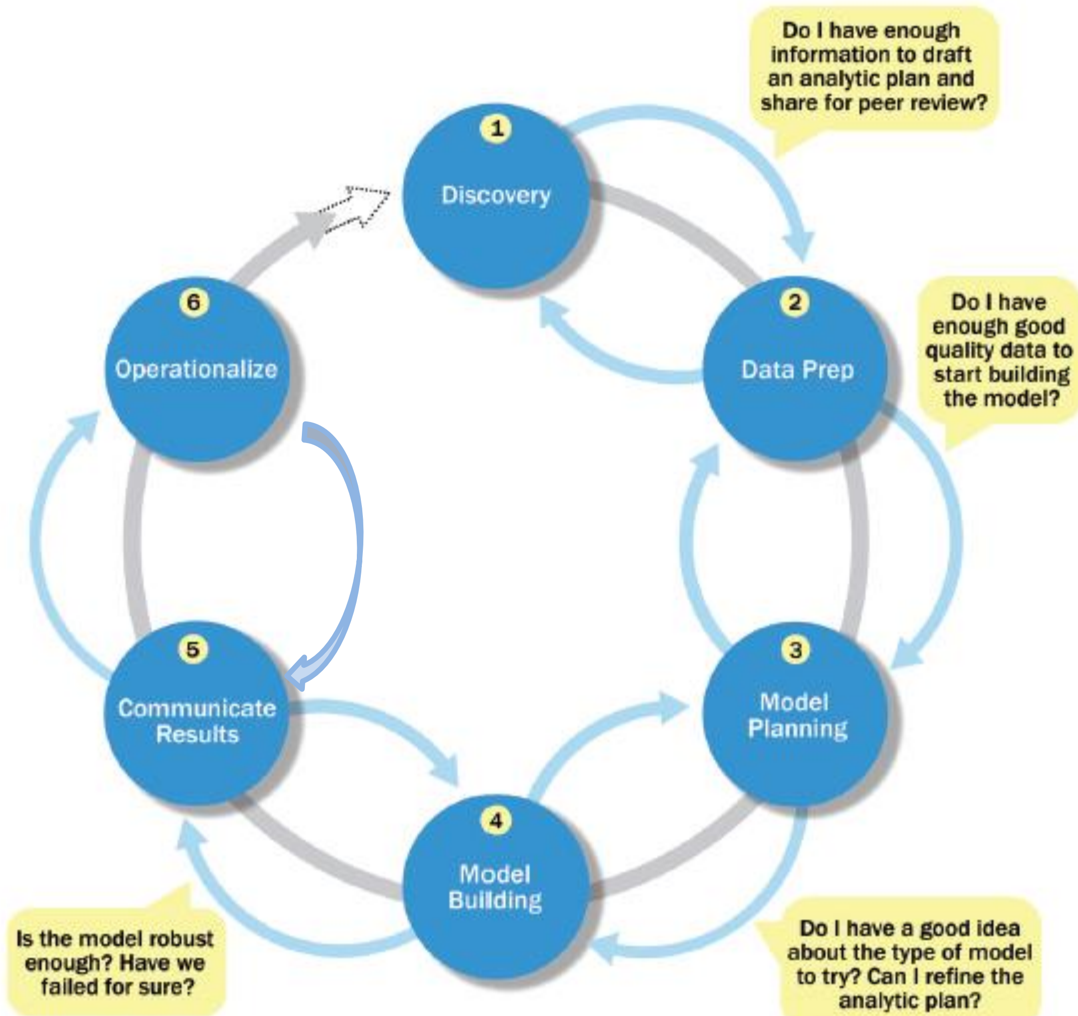
Data Analytics Lifecycle Overview

- Six phases

1.
2.
3. ow,
4.
5.
6. Operationalize (delivery, pilot project)

Don’t underestimate the difficulty and importance of these phases!

Data Analytics Lifecycle Overview



Phase 1: Discovery

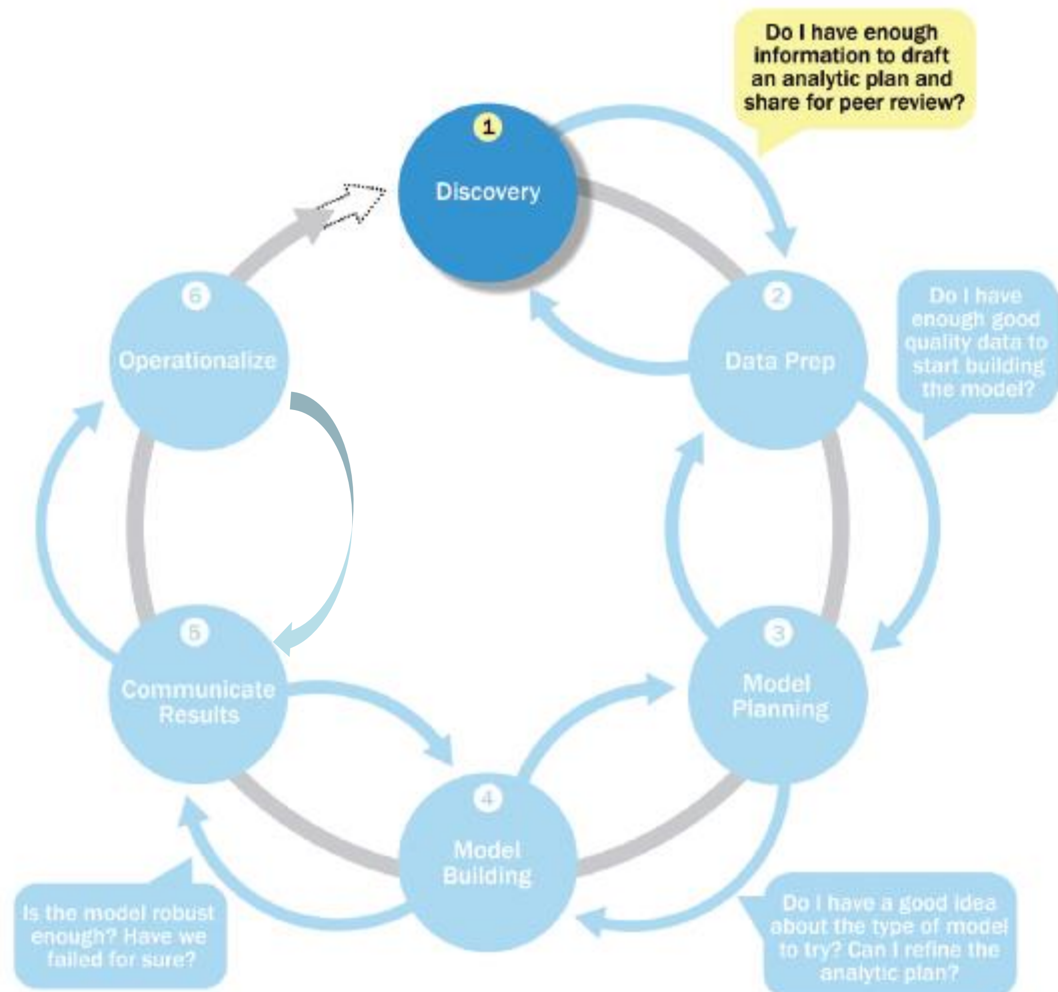


FIGURE 2-3 *Discovery phase*

Phase 1: Discovery

- Learning the Business Domain
  - Understand the domain.
  - Determine how much domain knowledge needed to develop models.
  - Domain knowledge + technical expertise.
- Resources
  - How much resources available to a project?
  - Technology, tools, systems, data and people.
  - Short-term and longer-term goals.

Phase 1: Discovery

- Framing (scoping) the Problem
  - The process of stating the analytical problems.
  - Identify objectives, risks, criteria of success.
  - Criteria of failure (when to stop?)
- Identifying Key Stakeholders
  - Anyone who will benefit from or be impacted by.
  - Collect key information from them.
  - Set clear expectations with them.

Phase 1: Discovery

- Interviewing the Analytical Sponsor
  - Use its knowledge and expertise.
  - Have a more objective understanding of problem.
  - Focus on clearly defining the project requirements.
  - Take time to conduct a thorough interview.
  - Some tips for the interview.
    - Good preparation, open-ended questions.
    - Give time to think, repeat back what was heard.
    - Be mindful of body language, document carefully.

Phase 1: Discovery

- Interviewing the Analytical Sponsor
  - Comment questions for the interview
    - What business problem?
    - What desired outcome?
    - What data source?
    - What industry issue?
    - What timelines?
    - Who has final decision-making authority?
    - …

Phase 1: Discovery

- Developing Initial Hypotheses (IH)
  - A key facet of the discovery phase.
  - Form ideas that can be tested with data.
  - Form the basis of later phases and serve as the foundation for the findings.
  - By comparison, can have richer observations.
  - Gather and assess the hypotheses from stakeholders and domain experts.
  - Useful to obtain and explore some initial data.

Phase 1: Discovery

- Identifying Potential Data Sources
  - Consider the volume, type, and time span of data.
  - Need to access raw data.
  - Will influence the choice of tools and techniques.
  - Help to determine the amount of data needed.
  - Should perform five main activities.
    - Identify data sources; Capture aggregate data sources.
    - Review the raw data; Evaluate data structures and tools.
    - Scope the sort of data infrastructure needed.

Phase 2: Data Preparation

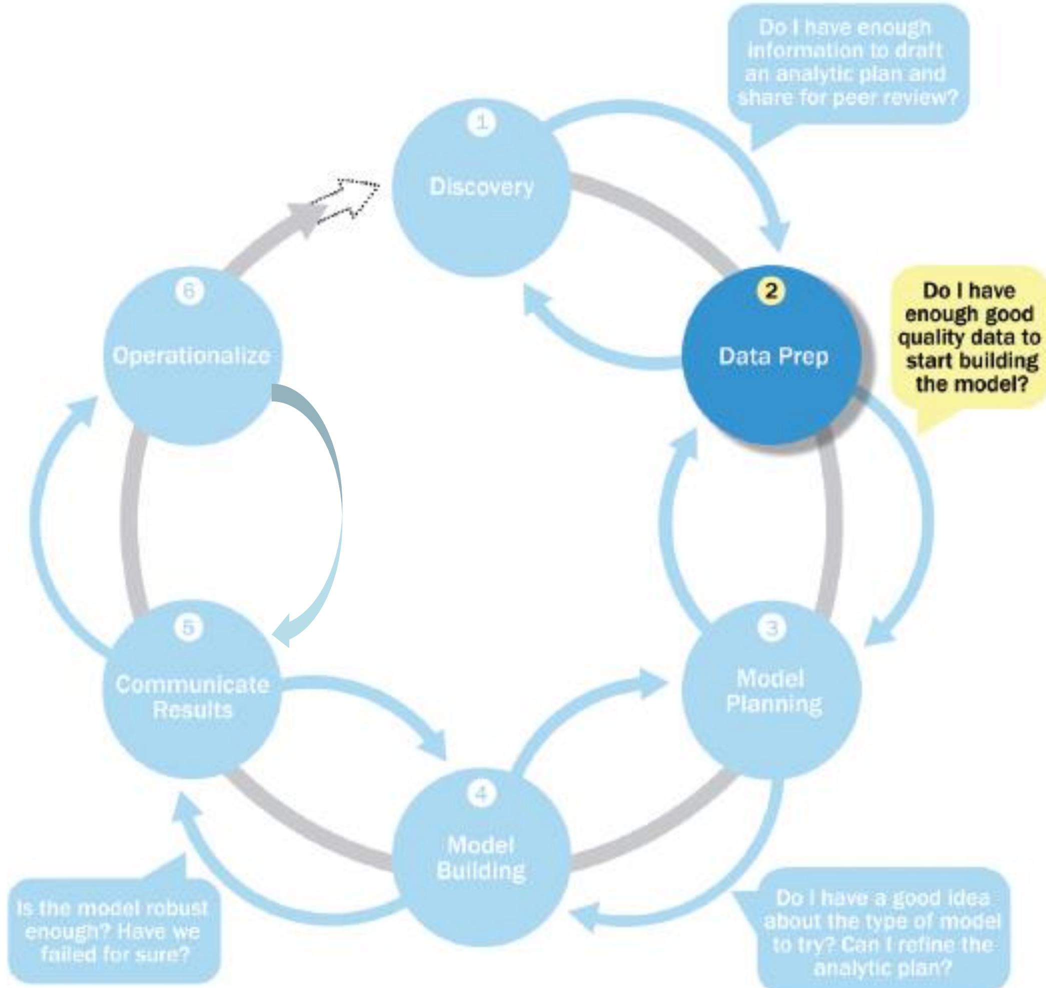


FIGURE 2-4 *Data preparation phase*

Phase 2: Data Preparation

- Explore, pre-process, and condition data prior to modelling and analysis.
- Prepare an analytics sandbox.
- Perform ETLT .
- Understanding the data in detail is critical.
- Get the data into a format to facilitate analysis.
- Perform data visualisation.
- Can be the **most labour-intensive step** in the lifecycle.

Phase 2: Data Preparation

- Preparing the Analytical Sandbox
  - Obtain an analytical sandbox (or workspace).
  - Collect all kinds of data there & which is important for a Big Data analytics project.
  - Need to collaborate with IT group, who usually has different views on data access.
  - Expect the sandbox to be large.
    - Raw data, aggregated data, less commonly used data.
    - At least 5-10 times the size of original dataset.

Phase 2: Data Preparation

- Performing ETLT
  - Analytic sandbox advocates extract (E), load (L), and then transform (T).
  - Data is extracted in its raw form and loaded into the datastore.
  - Access to data in its original form for finding hidden nuances or informative outliers.
  - Need to prepare for moving large amounts of data (Big ETL) --- parallelised by technologies.

Phase 2: Data Preparation

- Performing ETLT
  - Determine the transformations.
  - Assess data quality and structure datasets properly for robust analysis in subsequent phases.
  - Make an inventory of data and compare data currently available with datasets the team needs.
  - Utilise Application programming interfaces (APIs).

Phase 2: Data Preparation

- Learning About the Data
  - A critical aspect of a data science project is to become familiar with the data itself.
  - Accomplishes several goals.
    - Clarifies the data the team has access to.
    - Highlights gaps on data access.
    - Identifies datasets outside the organisation.

Phase 2: Data Preparation

- Learning About the Data

Dataset	Data Available and Accessible	Data Available, but not Accessible	Data to Collect	Data to Obtain from Third Party Sources
Products shipped	●			
Product Financials		●		
Product Call Center Data		●		
Live Product Feedback Surveys			●	
Product Sentiment from Social Media				●

Phase 2: Data Preparation

- Data Conditioning
  - Refers to the process of cleaning data, normalising datasets, and performing transformations on data.
  - A critical step involving many complex steps to join, merge, and transform datasets.
  - Usually performed by IT, the data owners, a DBA, or a data engineer (but data scientist are involved).
  - It is important to be thoughtful about choosing and discarding data!

Phase 2: Data Preparation

- Data Conditioning
  - Questions shall be asked
    - What are the data sources and target fields?
    - How clean is the data?
    - How consistent/complete are the contents and files?
    - Assess the consistency of data types
    - Review the content of data columns or other inputs
    - Look for any evidence of systematic error
    - Any signs of noise, outliers, incorrect, missing values?
      - Be careful how you deal with data affected by noise, outliers, incorrect or missing values.

Phase 2: Data Preparation

- Survey and Visualise
  - Leverage data visualisation tools to gain an overview of the data.
  - Seeing high-level patterns helps understanding.
  - “Overview first, zoon and filter, then details on demand”.
  - Guidelines and considerations recommended.
    - Review data to ensure calculations remained consistent.
    - Does the data distribution stay consistent?

Phase 2: Data Preparation

- Survey and Visualise
  - Guidelines and considerations recommended
    - Review data to ensure calculations remained consistent
    - Does the data distribution stay consistent?
    - Assess the granularity of the data
    - Does the data represent the population of interest?
    - For time-related variables, what is the measurement?
    - Is the data normalised? Scales are consistent?
    - For geospatial data, for personal names, for unit?

Phase 2: Data Preparation

- Common Tools for the Data Preparation Phase
  - Hadoop: can perform massively parallel ingest and custom analysis and combine massive unstructured data feeds from multiple sources
  - OpenRefine: a free, open source, powerful tool for working with messy data. It is a popular GUI-based tool for performing data transformations. It is one of the most robust free tools currently available.

Phase 3: Model Planning

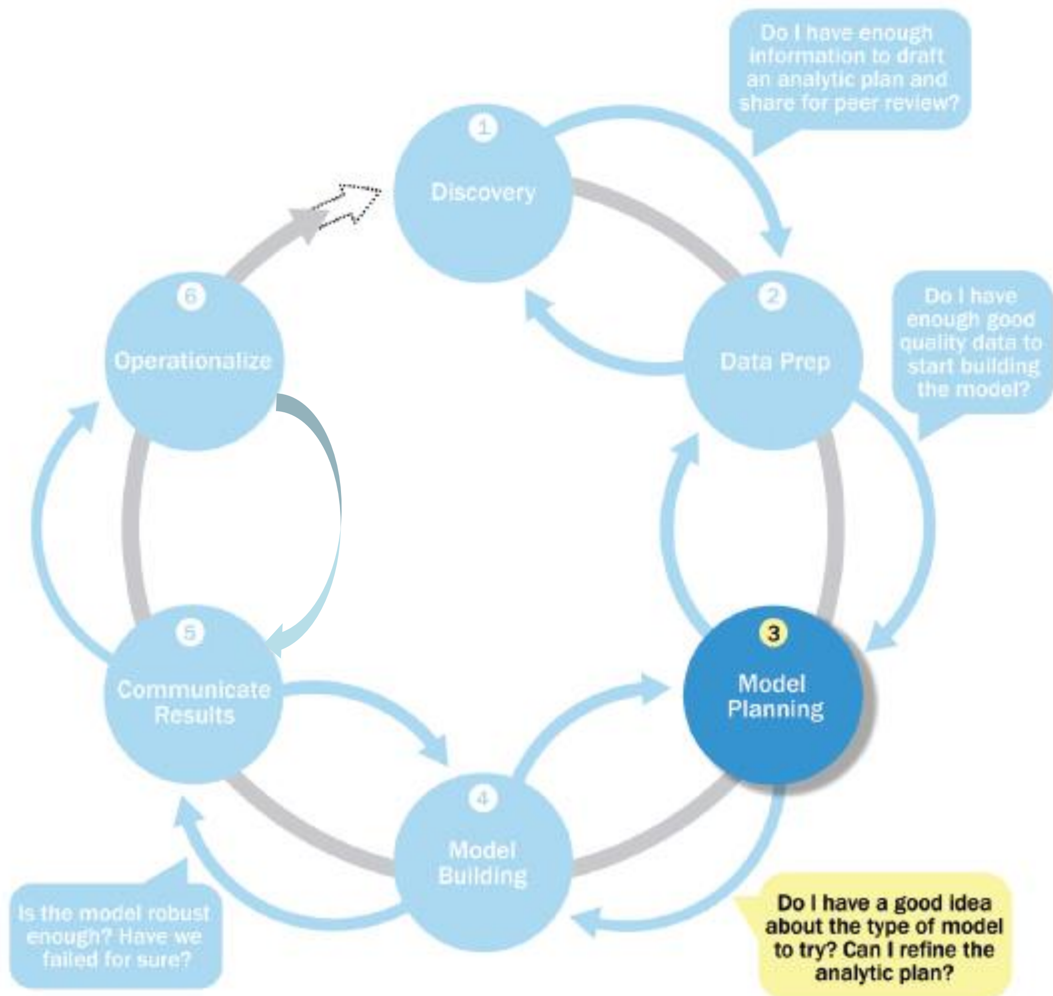


FIGURE 2-5 *Model planning phase*

Phase 3: Model Planning

- Identifies candidate models to apply to data.
  - For clustering, classifying, or finding relationships.
- Refers to the hypotheses developed in Phase 1.
- Activities to consider in this phase.
  - Assess the structure of datasets.
  - Ensure the analytical techniques capable.
  - Determine the need of a single or multiple models
- Conduct a critical literature review of similar projects.

Phase 3: Model Planning

- Data Exploration and Variable Selection
  - To understand the relationships of the variables
  - To help selection of the variables and methods
  - To understand the problem domain
  - Use tools to perform data visualisation
  - Explore the stakeholders and subject matter experts for their instincts and knowledge
  - Capture the most essential predictors and variables, rather than every possible ones

Phase 3: Model Planning

- Model Selection
  - Choose an analytical technique, or a short list of candidate techniques, based on the end goal of the project.
  - A model refers to an abstraction from reality. It emulates the behaviour of data with a set of rules and conditions.
  - Machine learning and data mining
    - Classification, association rules, and clustering

Phase 3: Model Planning

- Model Selection
  - When dealing with Big Data, the team needs to consider techniques best suited for structured data, unstructured data, or a hybrid approach.
  - Are explanatory models required?
  - Take care to identify and document the modelling assumptions.
  - Typically, create some initial models using a statistical software package
    - Baseline results can be indicative of the difficulty of the problem.
  - Move to model building phase.

Phase 3: Model Planning

- Common Tools for the Model Planning Phase
  - R is an open source programming language and software environment for statistical computing and graphics.
  - Has a complete set of modelling capabilities and provides a good environment for building interpretive models.
  - Has the ability to interface with databases.
  - Can perform statistical tests and analytics on some Big Data problems.

Phase 4: Model Building

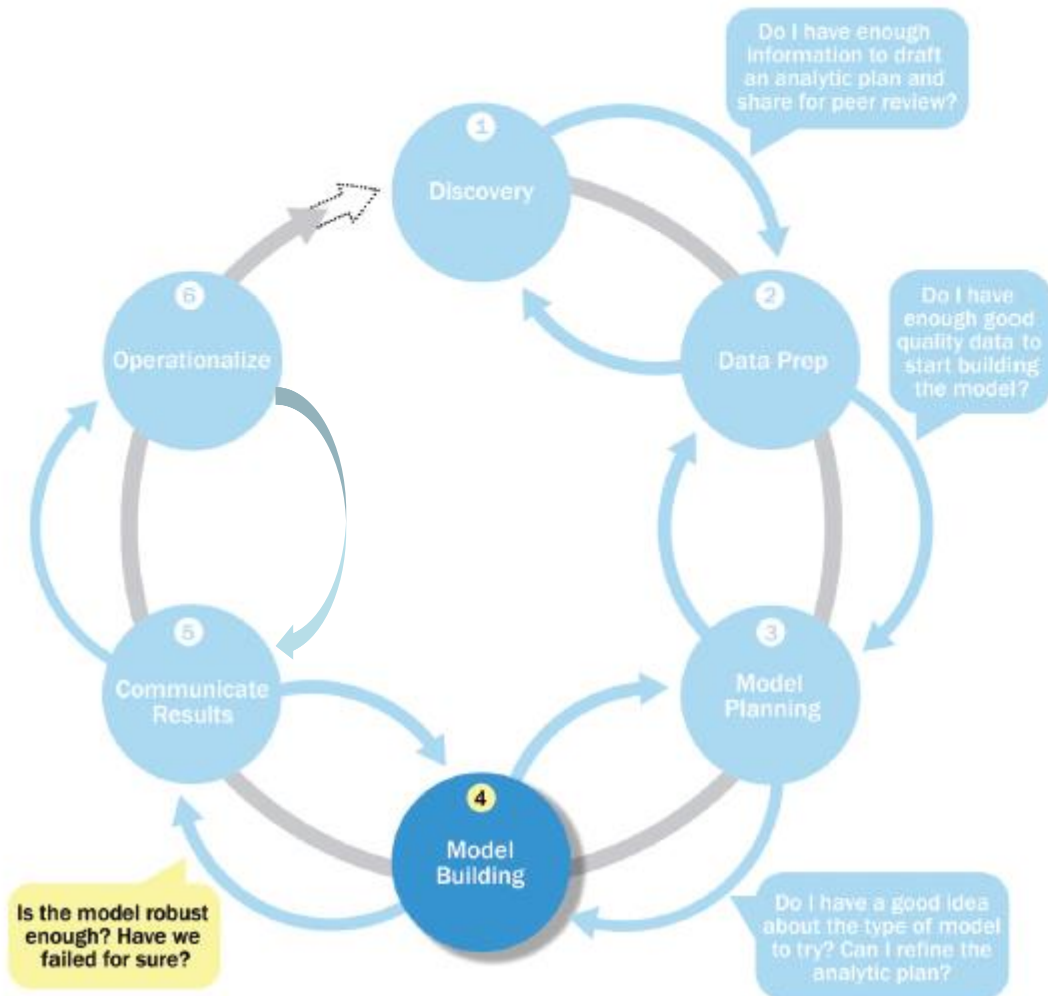


FIGURE 2-6 *Model building phase*

Phase 4: Model Building

- Develop datasets for training, testing, and production purposes.
- Train the analytical model and test it.
- Model planning and model building can overlap quite a bit. One can iterate back and forth for a while.
- Although modelling techniques can be highly complex, the actual duration of this phase can be short.

Phase 4: Model Building

- Run models from software packages on file extracts and small datasets.
- It is vital to record the results and logic of the model during the phase.
- Record any operating assumptions made in the modelling process.
- Creating robust models requires thoughtful consideration to meet the objectives.
- Understand the role of training data, validation data, and testing data, and use those sets correspondingly.

Phase 4: Model Building

- Questions to consider include
  - Model appear valid and accurate on validation data?
    - Tweak training parameters as needed.
  - Model appear valid and accurate on test data?
  - Output/behaviour make sense to domain expert?
  - Model parameters make sense?
  - Model is sufficiently accurate to meet the goal?
  - Model supports run-time requirements?
  - Is a different form of the model required?

Phase 4: Model Building

- Common Tools for the Model Building Phase
  - Matlab, Octave
  - Mathematica
  - SAS, SPSS
  - R
  - WEKA
  - Python
  - …

Phase 5: Communicate Results

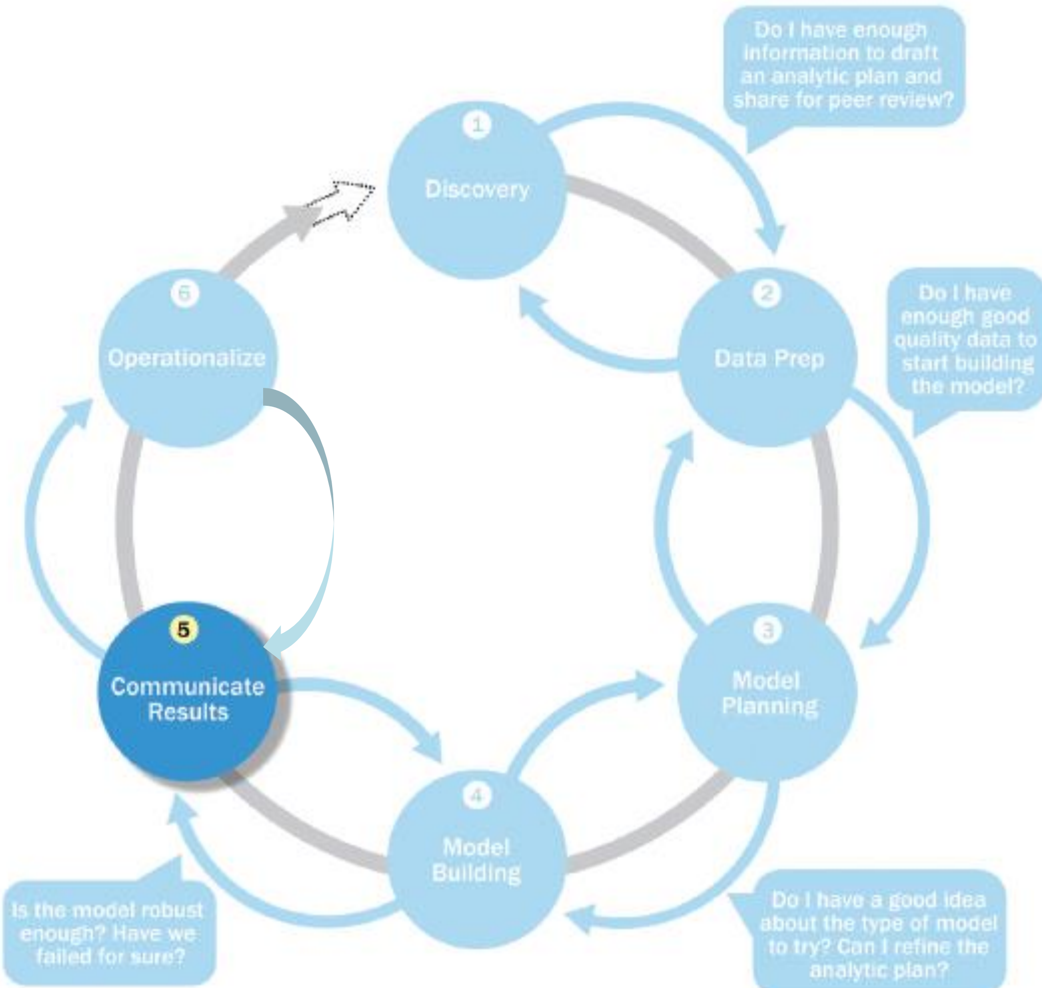


FIGURE 2-7 Communicate results phase

Phase 5: Communicate Results

- Compare the outcomes of the modelling to the criteria established for success and failure.
- Articulate the findings and outcomes to team members and stakeholders.
- Take into account caveats, assumptions, and any limitations of the results.
- Failure: a failure of the data to accept or reject a given hypothesis adequately.

Phase 5: Communicate Results

- Two extremes
  1. Only done a superficial analysis, not robust enough to accept or reject a hypothesis.
  2. Perform very robust analysis to search for ways to show results, even when results may not be there.
  - Need to strike a balance between these two extremes, be pragmatic.
- Record all findings and select the three most significant ones to share with stakeholders.

Phase 5: Communicate Results

- Make recommendations for future work or improvements.
- This is the phase to underscore the business benefits of the work.
- Begin making the case to implement the logic into a live production environment.
- The deliverable of this phase will be the most visible portion to stakeholders and sponsors.

Phase 6: Operationalize

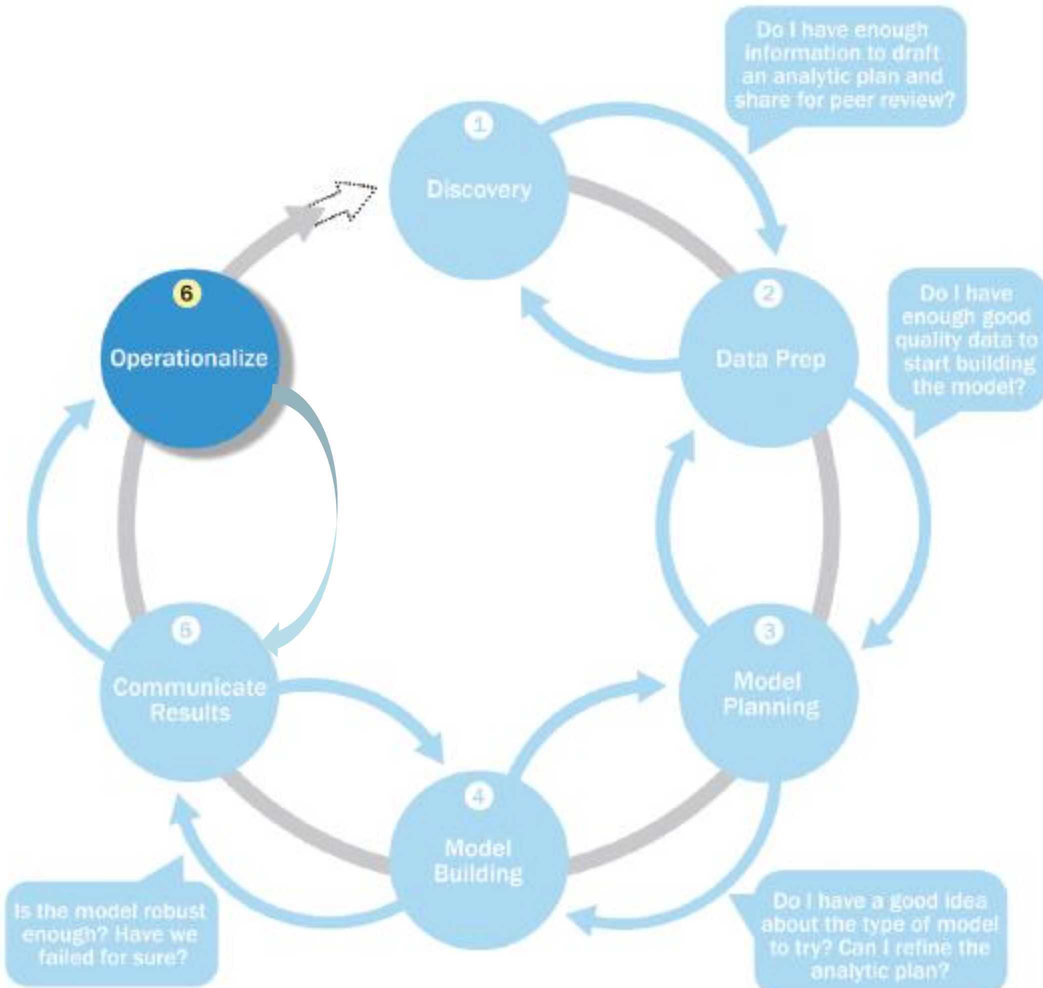


FIGURE 2-8 *Model operationalize phase*

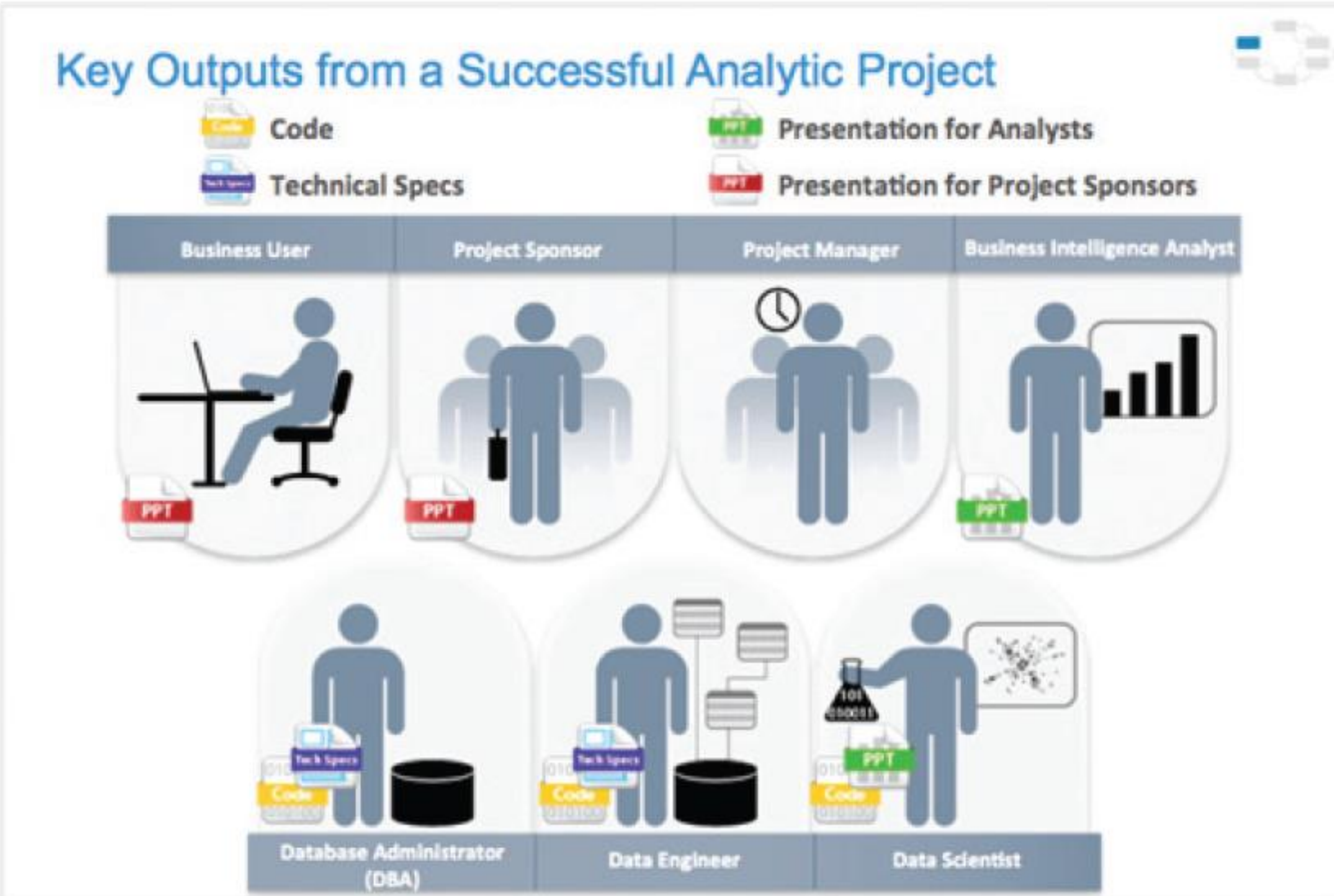
Phase 6: Operationalize

- In the final phase, communicate the benefits of the project more broadly.
- Set up a pilot project to deploy the work in a controlled way, before broadening the work to a full enterprise or ecosystem of users.
  - Risk can be managed more effectively .
- Learn the performance and constraints of the model.
  - Make adjustments before a full deployment.

Phase 6: Operationalize

- This phase can bring in a new set of team members (e.g., engineers responsible for the production environment).
- Create a mechanism for performing ongoing monitoring of model accuracy.
- Prepare to retrain the model.

Phase 6: Operationalize



Phase 6: Operationalize

- Business users: benefits and implications
- Project sponsor: business impact, risk, ROI
- Project manager: completion on time, within budget, goals are met?
- BI analyst: reports and dashboards impacted?
- DE and DBA: code and documents
- Data scientist: code, model, and explanation

Phase 6: Operationalize

- Four main deliverables
  - Presentation for project sponsors
  - Presentation for analysts
  - Code for technical people
  - Technical specifications of implementing the code
- A general rule: the more executive the audience, the more succinct the presentation needs to be.

Recap: Data Analytics Lifecycle Overview

- Key Roles for a Successful Analytics Project

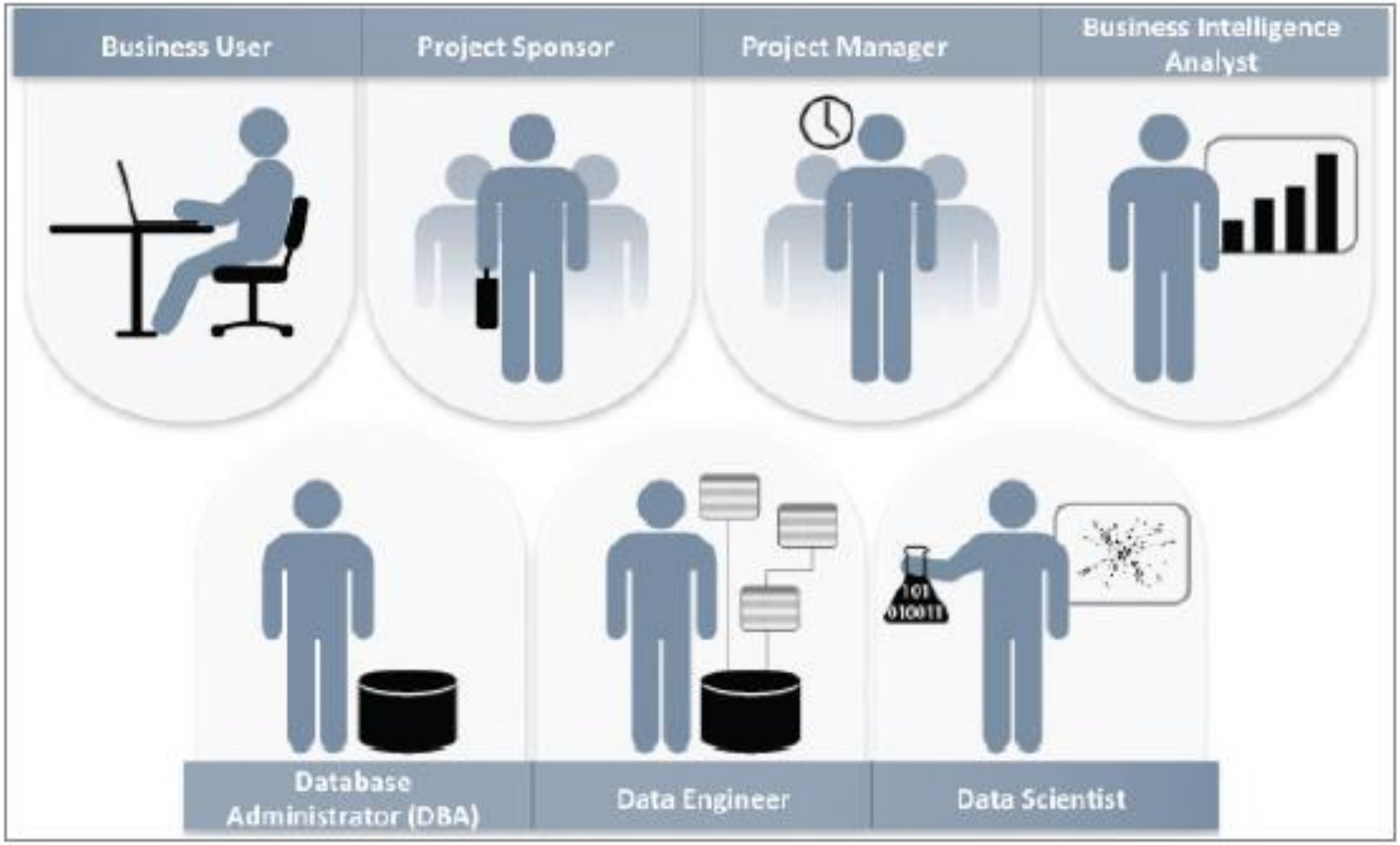


FIGURE 2-1 *Key roles for a successful analytics project*

Recap: Data Analytics Lifecycle Overview

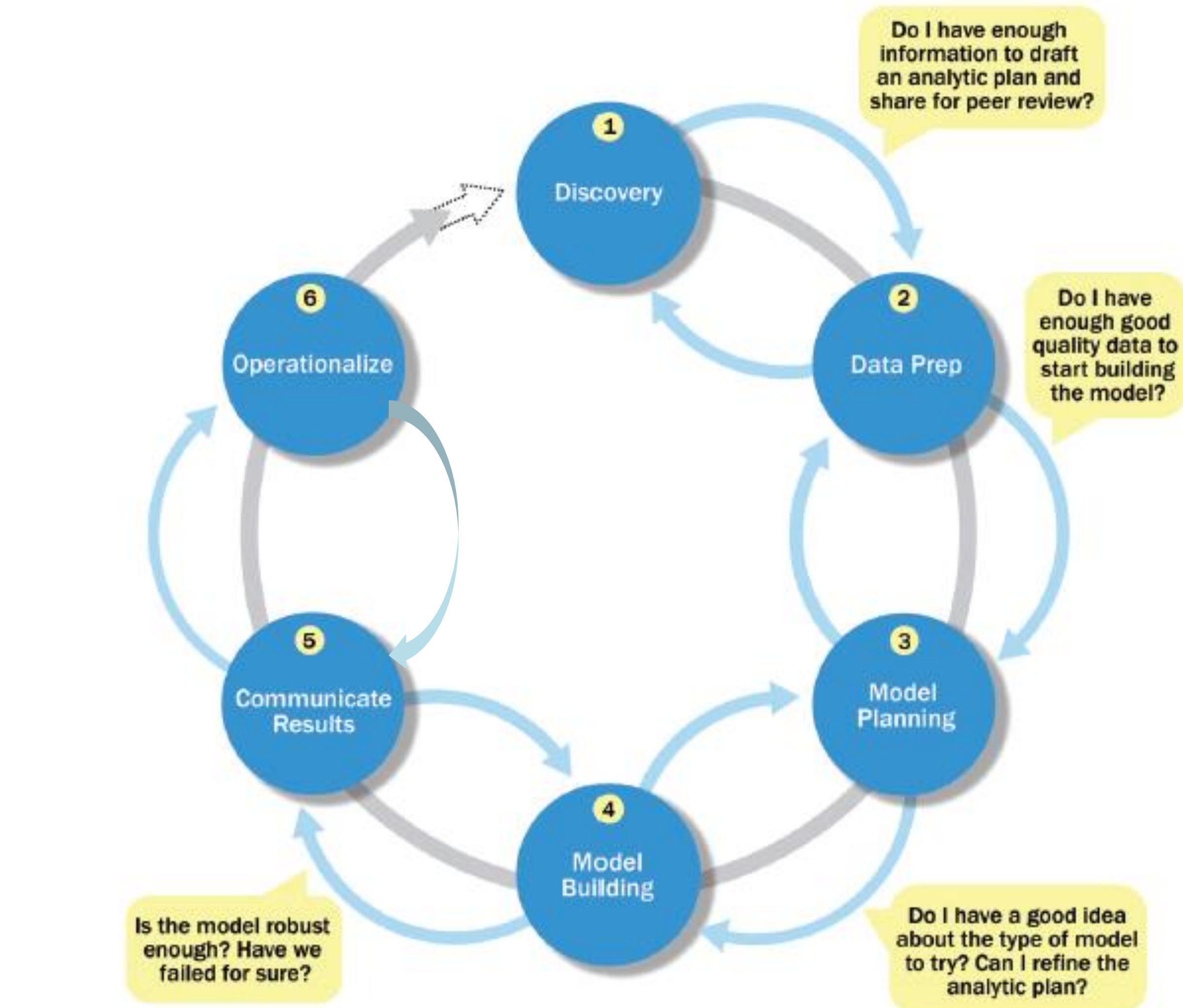


FIGURE 2-2 *Overview of Data Analytics Lifecycle*

Q&A

Images Courtesy of Google Image