CSCI446/946 Big Data Analytics

Week 2 – Lecture: Big Data & Analytics Lifecycle

School of Computing and Information Technology
University of Wollongong Australia
Spring 2024

Content

- Brief Recap
  - Big Data Overview
  - Big Data Properties
  - Structures of Big Data
- Catch up on Big Data Analytics Introduction
  - State of the Practice in Analytics

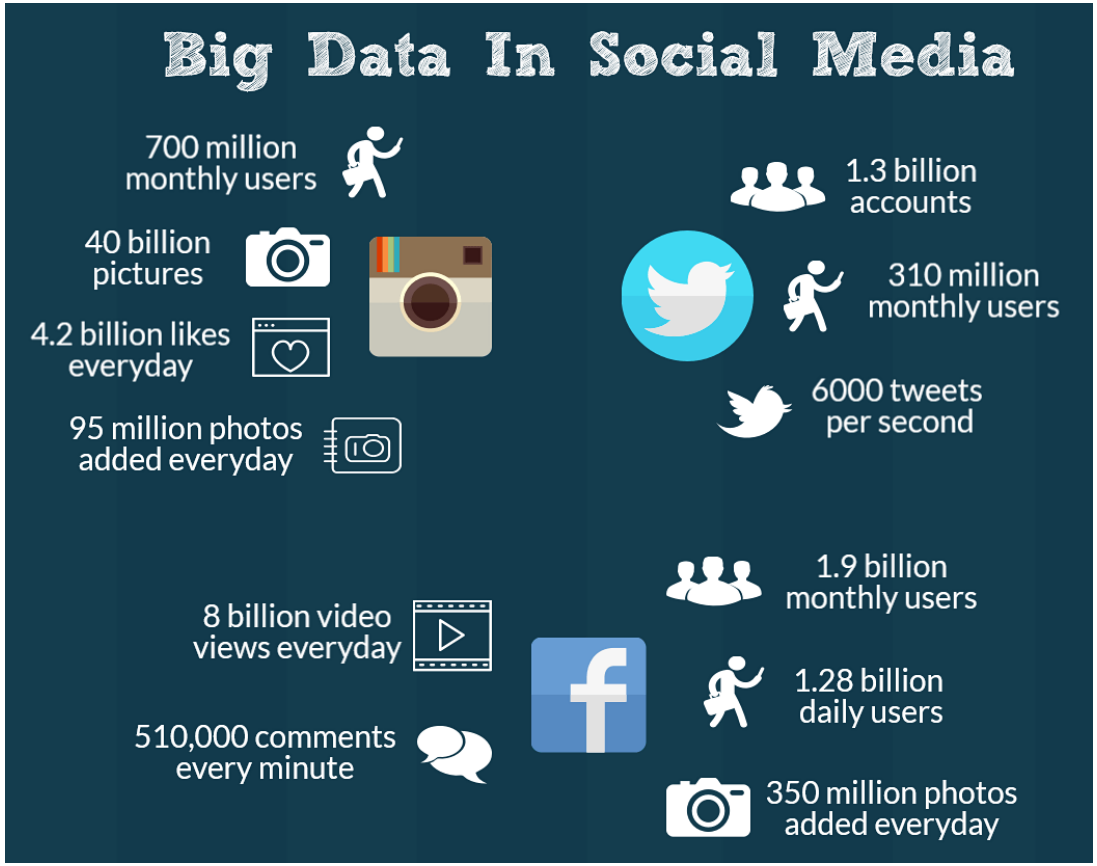
- Data Analytics Lifecycle

Content

- Brief Recap
  - Big Data Overview
  - Big Data Properties
  - Structures of Big Data
- Catch up on Big Data Analytics Introduction
  - State of the Practice in Analytics
- Data Analytics Lifecycle

Big Data Overview

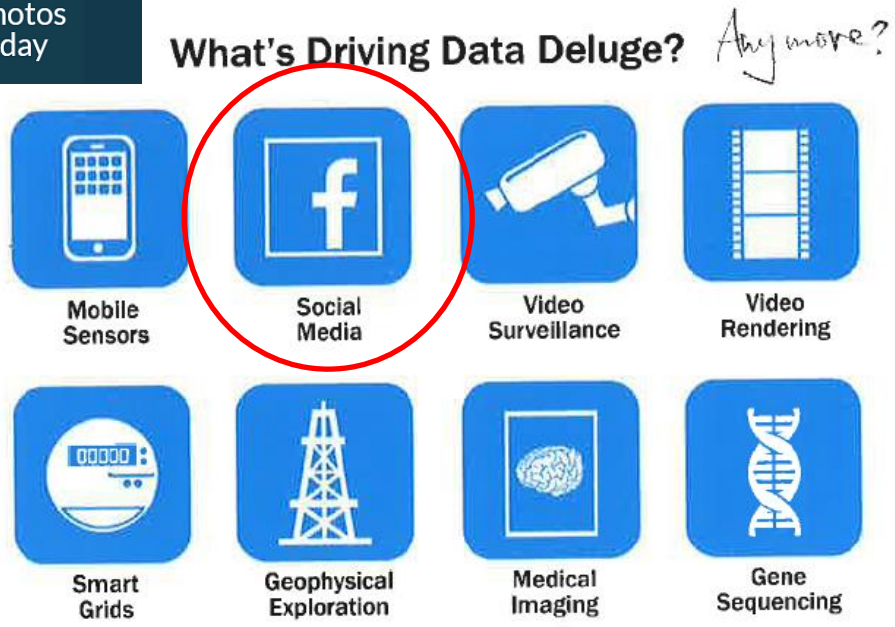
- Sources and drivers of Big Data
  - i.e. social media, multi-media, Web, “smart” devices,…
- When is data Big?
  - It’s not just about size.
    - It’s about data properties and available technology.
    - Real-time processing is a commonly requirement.
    - Processing data from a variety of unfiltered sources is common.



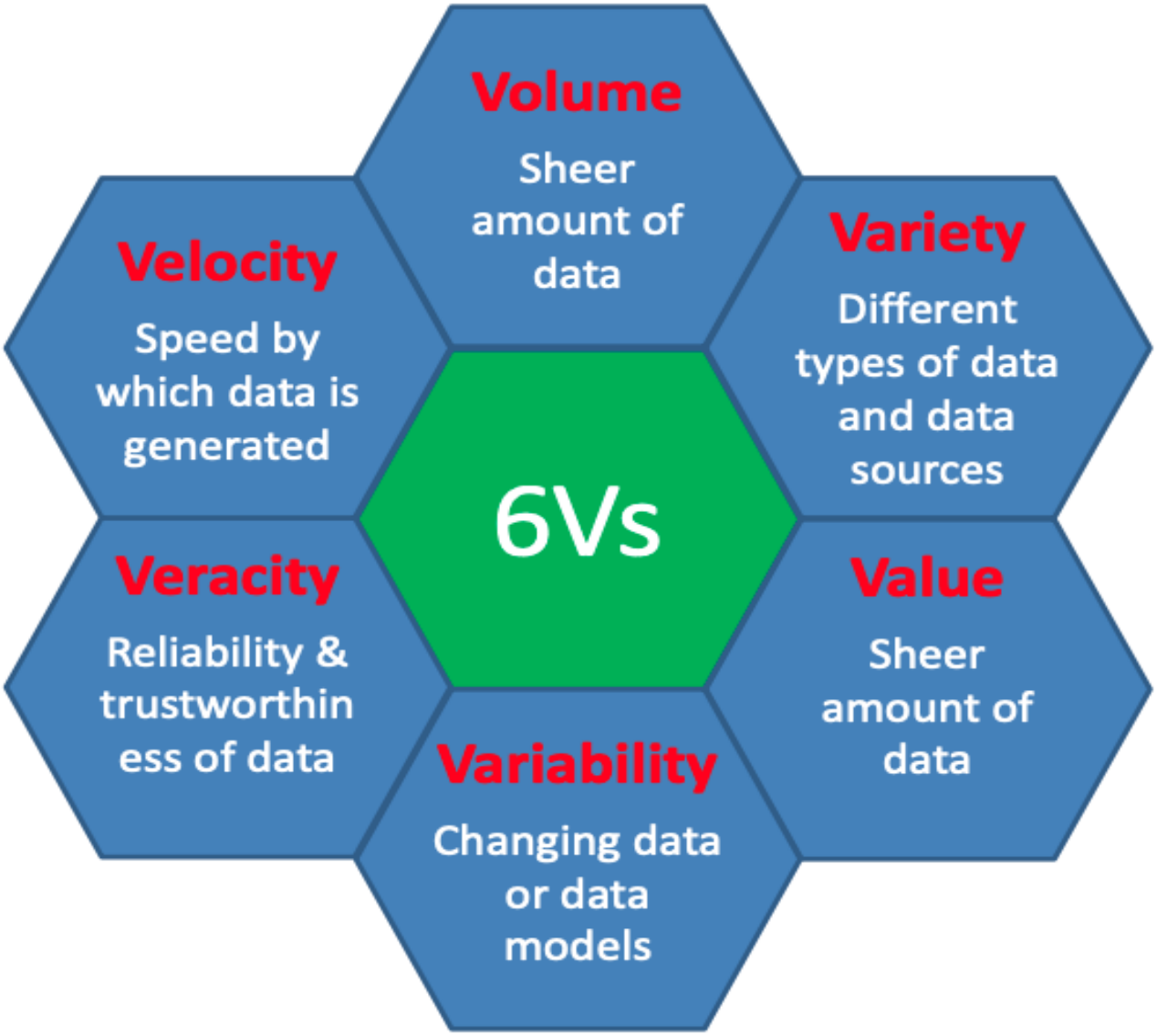
Why most popularly known?

- Fast -> easy to collect
- Large-scale -> friendly to algos
- Different data -> innovative
- Dynamic -> challenging
- Valuable, e.g., marketing
- Real data -> truth
- more?

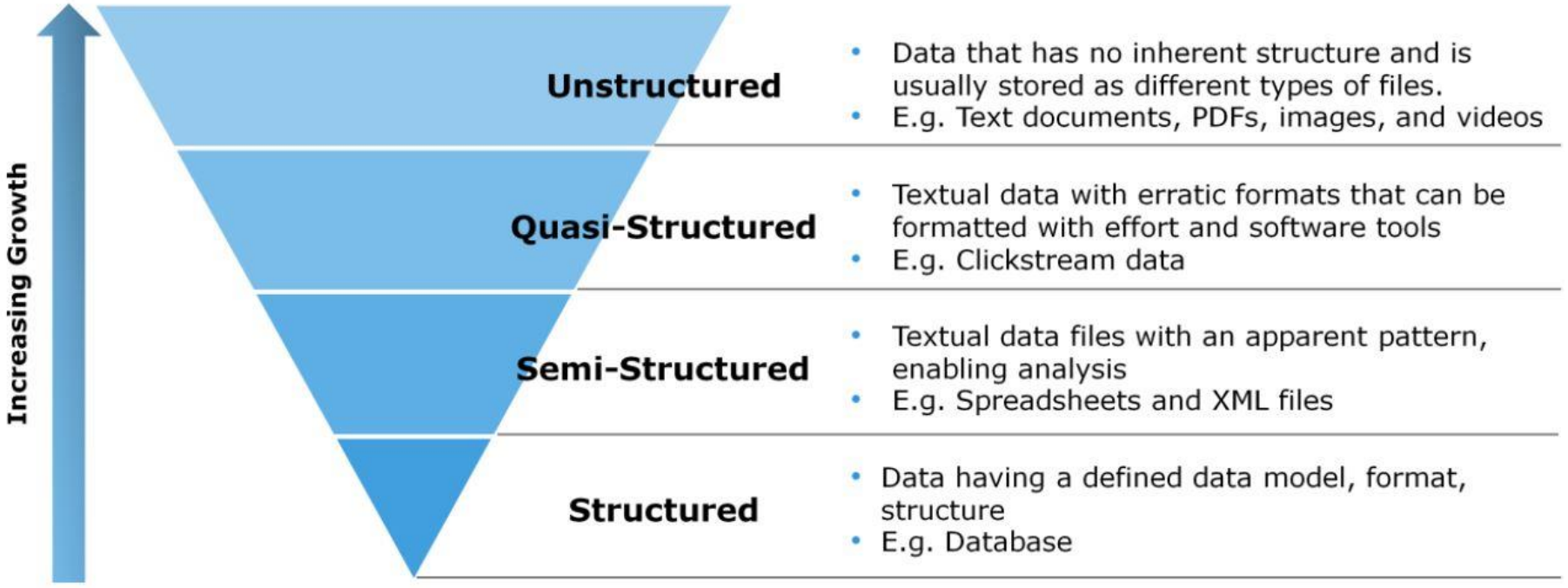
Activity → Data



Properties of Big Data



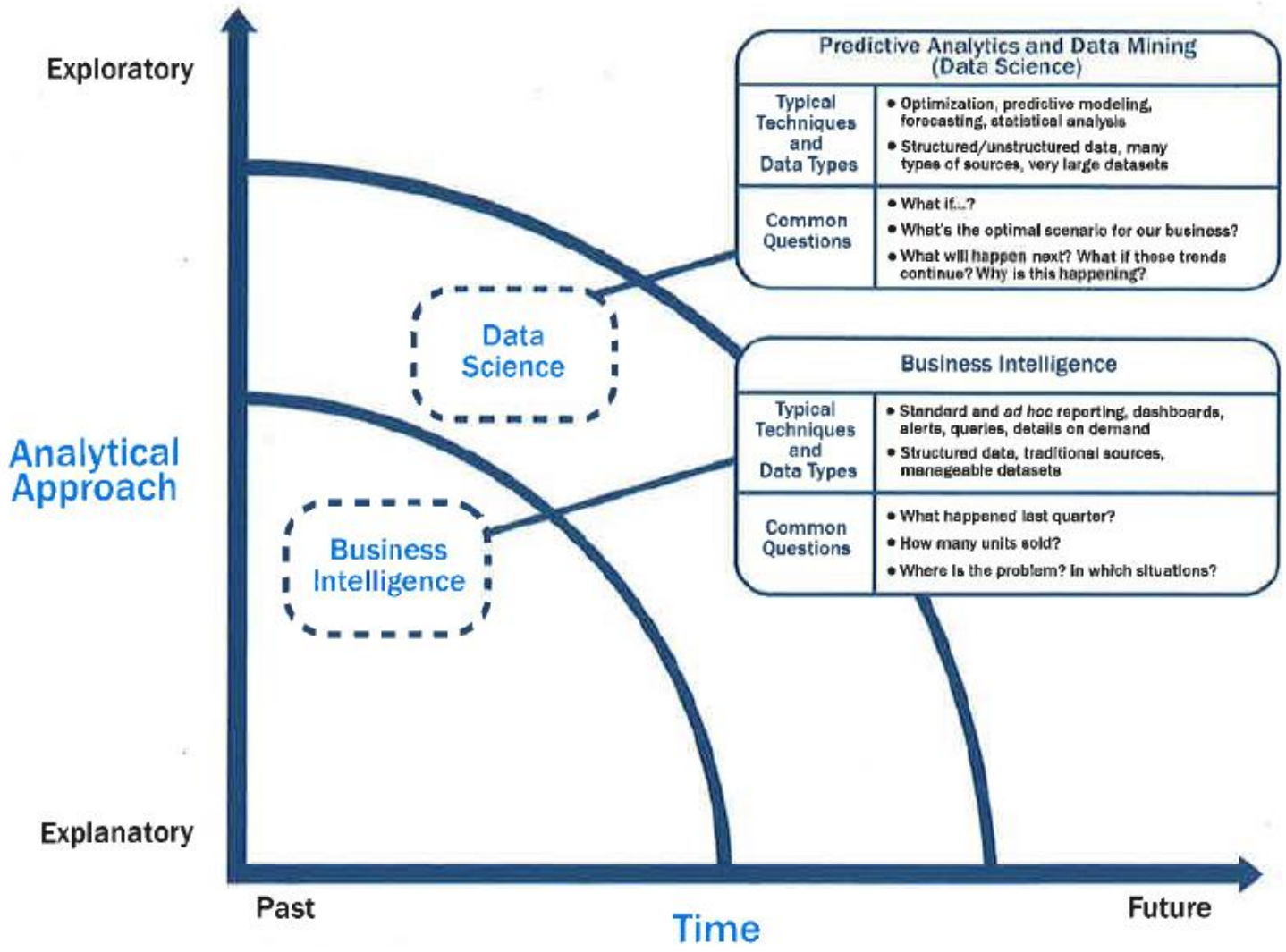
Structures of Big Data



Big Data Analytics may take all data structures

Detailed definitions & a similar Figure in Book - Data Science and Big Data Analytics Chapter1

Business Intelligence vs. Data Sceicne



Brief Recap

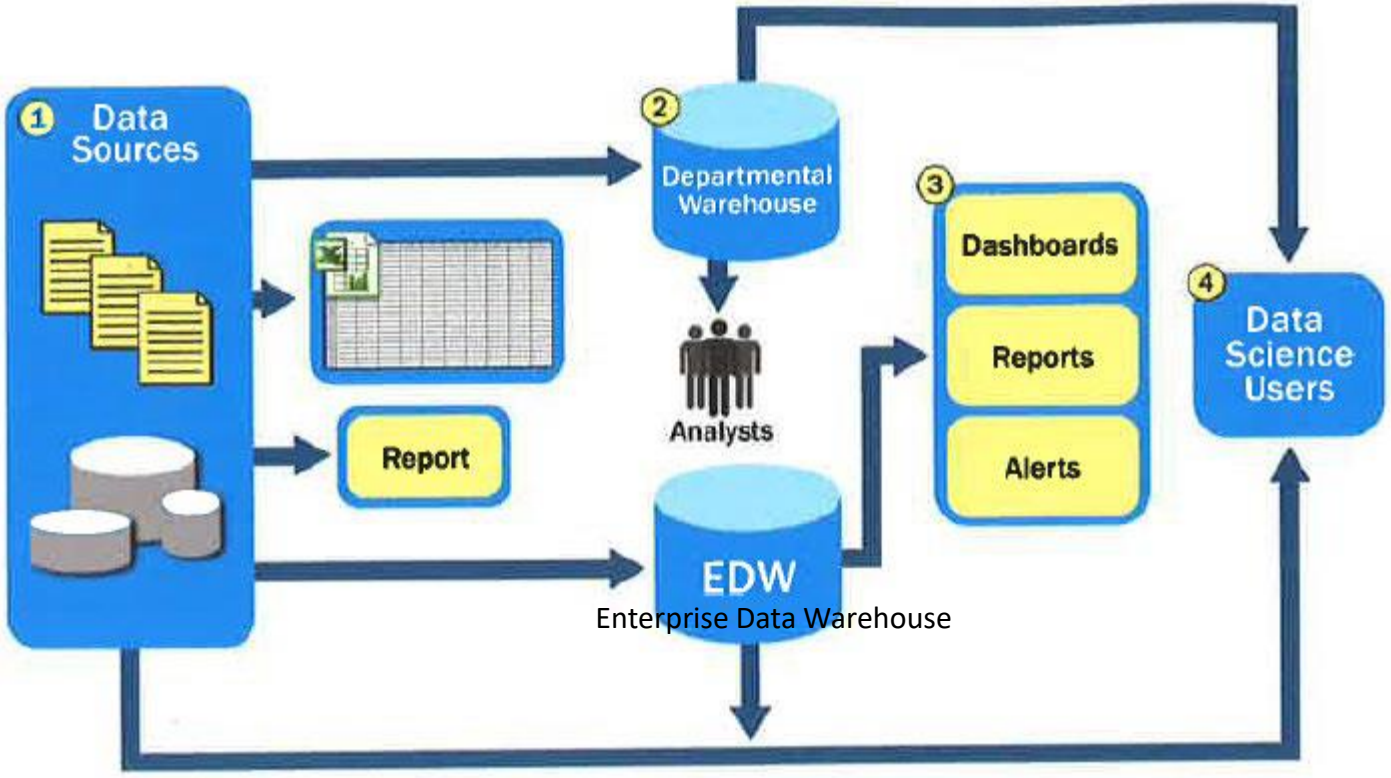
Questions & Answers

Content

- Brief Recap
  - Big Data Overview
  - Big Data Properties
  - Structures of Big Data
- Catch up on Big Data Analytics Introduction
  - State of the Practice in Analytics
- Data Analytics Lifecycle

State of the Practice in Analytics

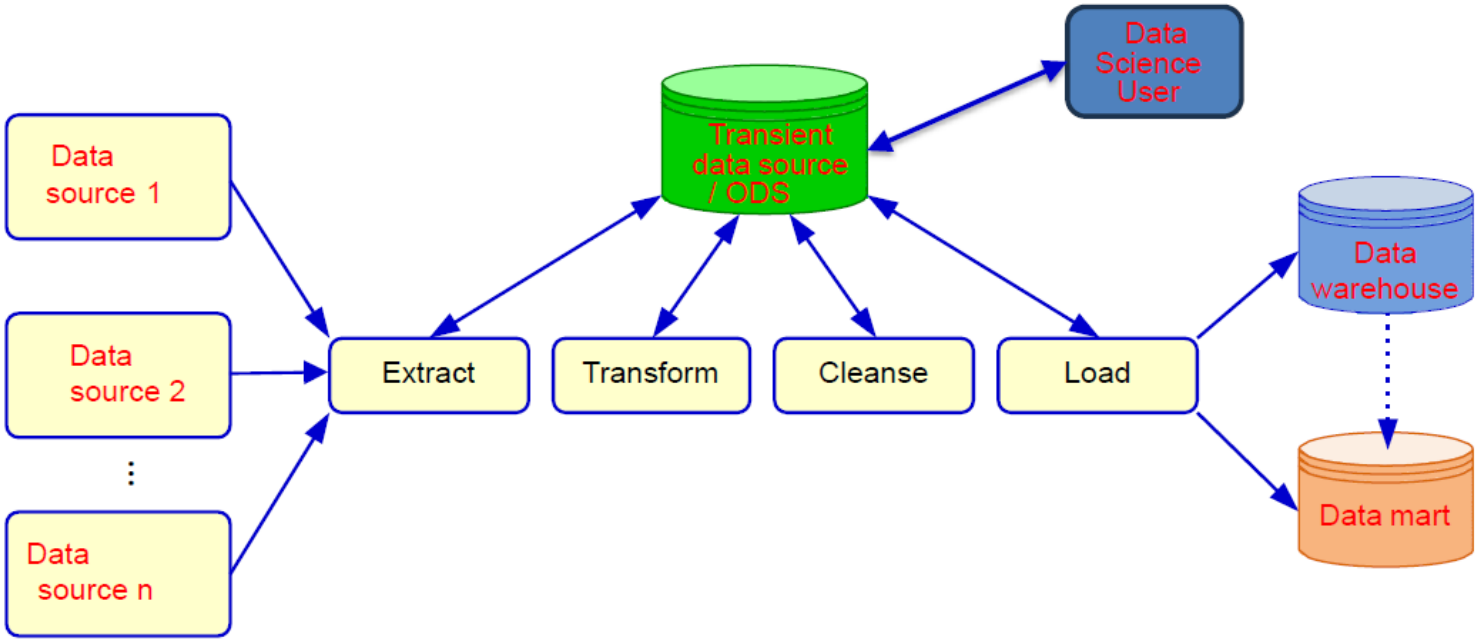
- Typical Analytical Architecture



This data architectures inhibit rapid data access, exploration and more sophisticated analysis.

State of the Practice in Analytics

- Processing high-velocity data needs faster access to data i.e. via the use of a transient data store



State of the Practice in Analytics

- Traditional data architectures have several additional implications for data scientists
  - Predictive analytics and data mining activities are last in the line for data (i.e., low priority)
  - Limited to perform in-memory analytics, restricting the size of the datasets they can use
  - Projects remain isolated and ad hoc, rather than centrally managed. Exist as nonstandard initiatives.
  - Analytics takes place in a DW production system.
- One solution: analytic sandboxes

State of the Practice in Analytics

- Emerging Big Data Ecosystem & a New Approach to Analytics
  - Data -> intrinsic value -> a new economy
    - "Data is the new oil"
  - New professions: Data vendors, data cleaners,…
  - New opportunities for software developers:
    - Repackaging and simplifying open source tools
  - Data is the king!

https://www.linkedin.com/pulse/data-king-one-who-masters-rule-world-so-you-ready-vinay-rao-mle-/

State of the Practice in Analytics

- Four main groups of players here
  - Data devices
    - Video game, Smartphone, Retail shopping card
  - Data collectors
    - Service providers, shopping cart with RFID chips
  - Data aggregators
    - Compile, transform and package data to sell
  - Data users and buyers
    - Retail banks, common people
- Each with commercial interests.

Key Roles for the New Ecosystem

- Data Analytical Talent (Data Scientist)
  - Advanced training in mathematics, statistics, and machine learning
  - Newest role, least understood
- Data Savvy Professionals
  - Less technical depth but can define key questions
- Technology and Data Enablers
  - Support data analytical projects
- These three groups must work together

Key Roles for the New Ecosystem

- What do data scientists do?
  - Reframe business challenges to analytical challenges.
  - Design, implement, and deploy data mining techniques on Big Data.
    - This is mainly what people think about them
  - Develop insights that lead to actionable recommendations to derive new business value.

Examples of Big Data Analytics

- Some examples
  - US retailer Target
    - Infer Marriage, Divorce, Pregnancy, …
    - Manages its inventory correspondingly
  - IT Infrastructure
    - Apache Hadoop
    - Process vast amount of information in parallel.
  - Social media
    - Leverage social interactions to derive new insights.
- Free economy?
  - Facebook, WhatsApp, Beidu, retail memberships, … can be used free of charge.
    - Are we in a wonderful world where businesses provide services to end users at no cost, make no profit, and pay for all expenses?

Summary

- Big Data comes from myriad of sources.
- Big Data addresses business needs and solves complex problems.
- Companies and organisations move toward Data Science.
- Require new architectures, new ways of working, new skill sets, new roles, etc.
- A growing talent gap.

Questions for you

- What are the four (or five, or six) characteristics of Big Data?
- What is an analytic sandbox, and why is it important?
- Explain the difference between BI and Data Science.
- Describe the challenges of the current analytical architecture for data scientists.
- What are key skills and roles of a data scientist?
- How much data is involved in Big Data?

Content

- Brief Recap
  - Big Data Overview
  - Big Data Properties
  - Structures of Big Data
- Catch up on Big Data Analytics Introduction
  - State of the Practice in Analytics

- Data Analytics Lifecycle

Data Science Projects

- Data Science is exploratory in nature.
- Common mistake
  - Rushing into data collection and analysis.
  - Not spend enough time planning, scoping, understanding, or framing.
- It is critical to have a process to govern the process.

Data Analytics Lifecycle Overview

- Data Analytics Lifecycle defines the roadmap of how data is generated, collected, processed, used, and analyzed to achieve business goals.
- It offers a systematic way to manage data for converting it into information that can be used to fulfill organizational and project goals.

Key Roles in an Analytics Project

- A data science team commonly consists of:
  - Business User; Project Sponsor; Project Manager
    - Business Intelligence Analyst
    - Database Administrator
    - Data Engineer;
    - Data Scientist
- The last two roles are in high demand!

  https://www.seek.com.au/career-advice/role/data-scientist

Key Roles in an Analytics Project

- 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 (1)

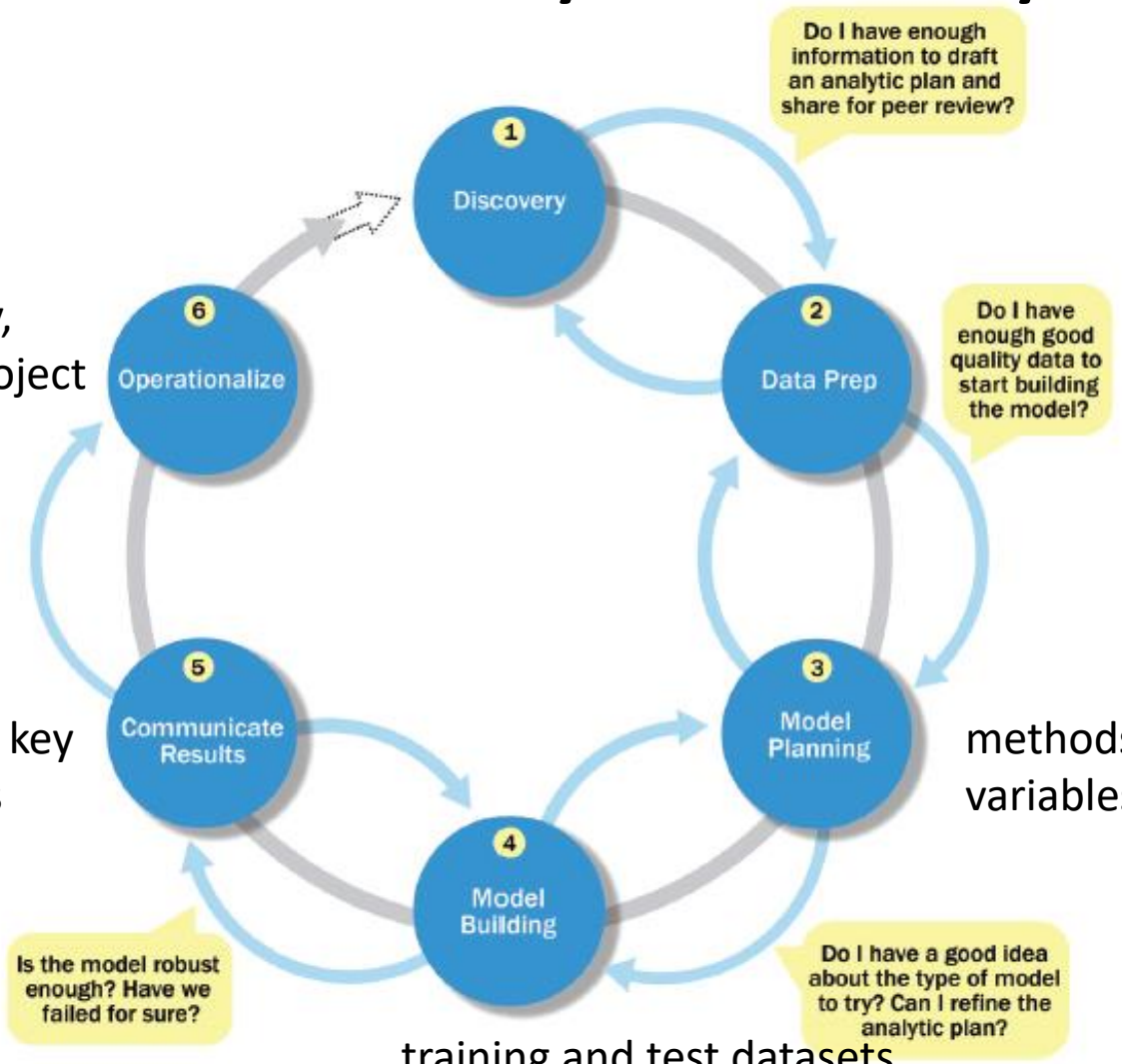
- A team of oncologists and radiotherapists wanted to know whether it is possible to predict from MRI scans the toxicity of a prescribed radiotherapy on healthy tissue.

- They approach a team of data scientists with this question.

An Example (II)

- 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.
  - The discovery phase a key step in the data analytics.

Data Analytics Lifecycle



delivery, pilot project

identify key findings

methods, techniques, workflow, variables, relationships, models

training and test datasets, software, and hardware

FIGURE 2-2 *Overview of Data Analytics Lifecycle*

Phase 1: Discovery

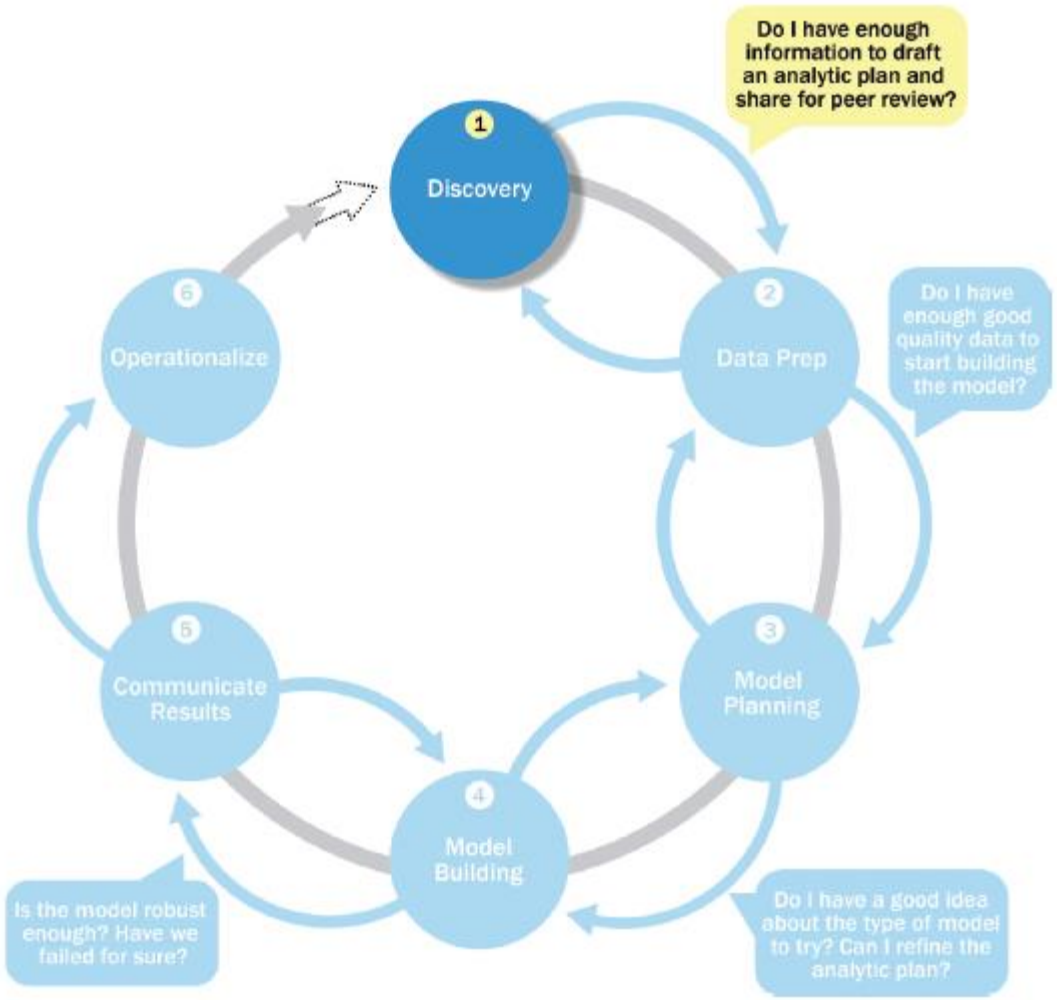


FIGURE 2-3 *Discovery phase*

Phase 1: Discovery

May with

- 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 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
  - Common 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

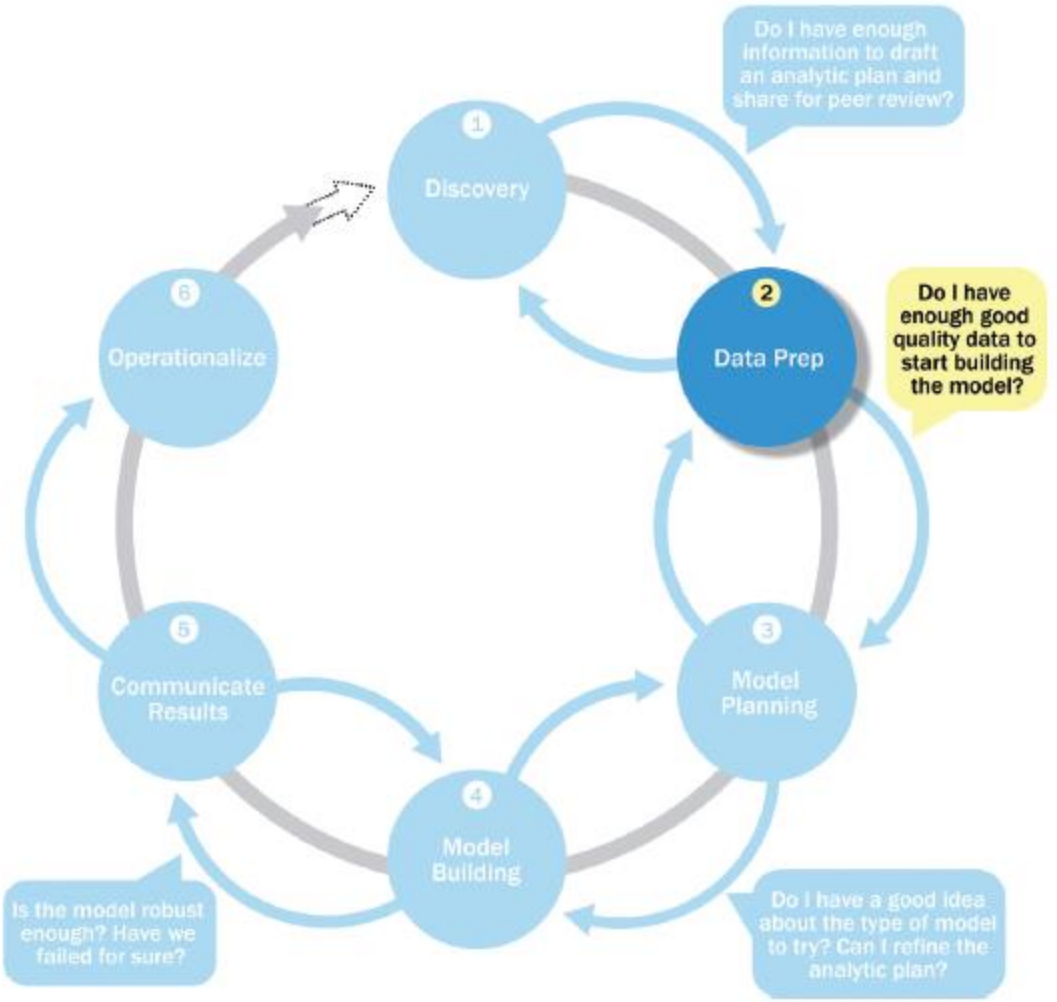


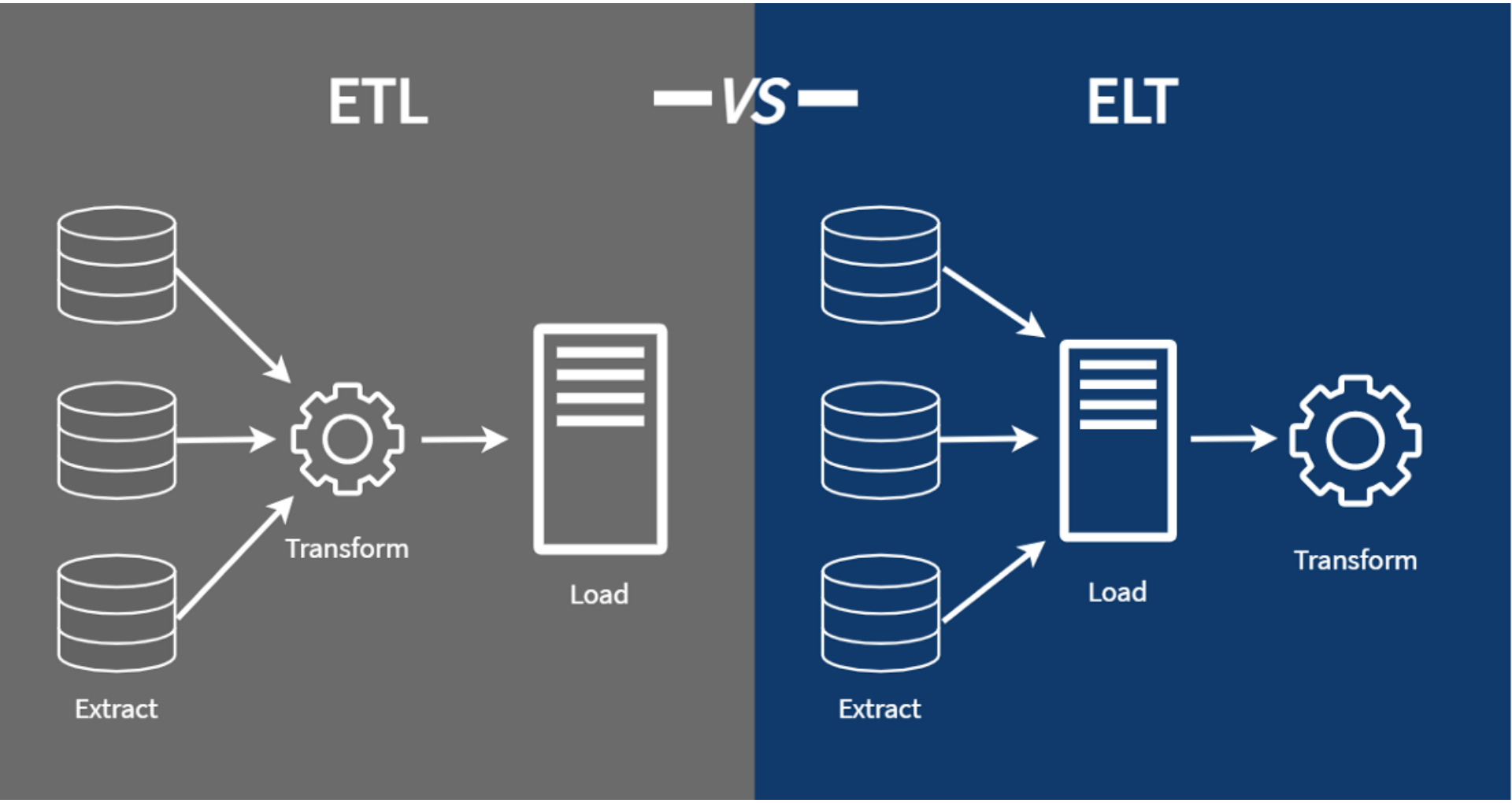
FIGURE 2-4 *Data preparation phase*

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
- 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


ETL
VS
ELT
Transform
Load
Extract
Load
Transform
Extract

Phase 2: Data Preparation

- **ETL (Extract, Transform, Load)**
  - **Process Order**: In ETL, data is first extracted from the source systems, then transformed into the desired format or structure, and finally loaded into the target data warehouse or data repository.
  - **Transformation Location**: Transformations are performed on an intermediate server before loading the data into the target system.
  - **Use Case**: ETL is suitable for environments where data needs to be cleaned and transformed before it can be loaded into the target system. This is common in traditional data warehousing.
  - **Performance**: ETL can be slower for large datasets because the transformation process happens before the data is loaded into the data warehouse, which may require additional resources and processing time.
  - **Flexibility**: ETL processes are often rigid and predefined, making them less flexible for changes or ad-hoc queries.

Phase 2: Data Preparation

- **ELT (Extract, Load, Transform)**
  - **Process Order**: In ELT, data is first extracted from the source systems, then loaded into the target data repository, and finally transformed within the target system.
  - **Transformation Location**: Transformations are performed within the target data repository, taking advantage of the processing power and capabilities of the data warehouse or data lake.
  - **Use Case**: ELT is suitable for big data environments where the target system has significant processing power and can handle large volumes of data. It leverages the capabilities of modern data warehouses and data lakes.
  - **Performance**: ELT can be more efficient for large datasets because the data is loaded first, and transformations are performed using the target system's computational resources. This can reduce data movement and improve processing speed.
  - **Flexibility**: ELT processes are more flexible and can handle complex transformations and ad-hoc queries more efficiently. It is well-suited for iterative and exploratory data analysis.

Phase 2: Data Preparation

- **ETLT (Extract, Transform, Load, and Transform)** - an extension of the traditional ETL process, designed to handle the complexities and scale of big data environments.
  - **Extract (E)**: This step involves extracting data from various sources, such as databases, files, APIs, and more. The data can come from structured, semi-structured, or unstructured sources.
  - **Transform (T)**: The first transformation step involves initial data cleaning, filtering, and preliminary transformations to make the data more manageable. This step might include data type conversions, removing duplicates, and handling missing values.
  - **Load (L)**: In this step, the pre-processed data is loaded into a data storage system, such as a data warehouse, data lake, or a Hadoop Distributed File System (HDFS). This step ensures that the data is available for further analysis.
  - **Transform (T)**: The second transformation step involves more complex and computationally intensive transformations that are typically performed within the data storage system. This might include data aggregation, enrichment, integration with other data sources, and applying advanced analytics or machine learning models.

The ETLT process is particularly useful in big data environments where data volumes are massive, and initial transformations can help reduce the load on the storage system, improve performance, and enhance the efficiency of subsequent processing and analysis steps.

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 shall involve)
  - 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”

Phase 2: Data Preparation

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

Phase 3: Model Planning

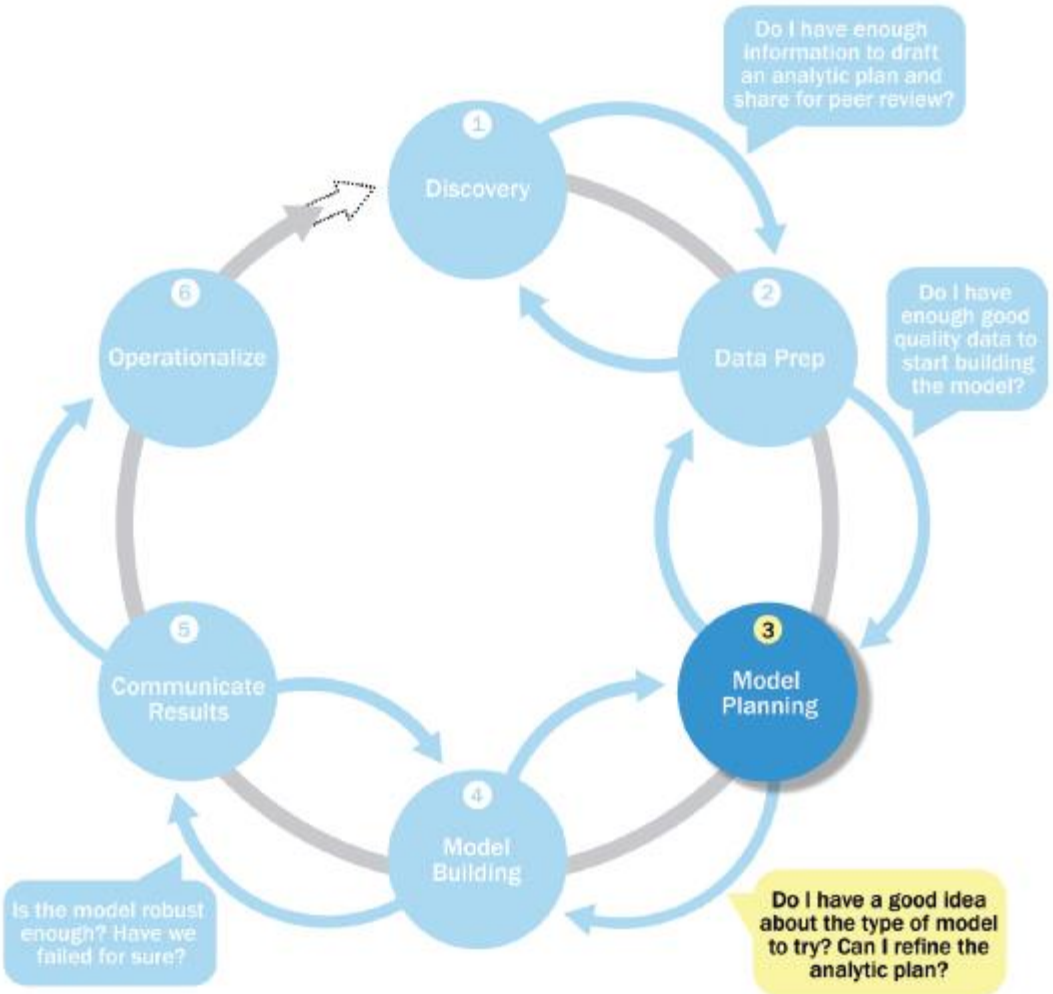


FIGURE 2-5 *Model planning phase*

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 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
  - Take care to identify and document the modelling assumptions
  - Typically, create the 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
  - Python/R and their packages 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

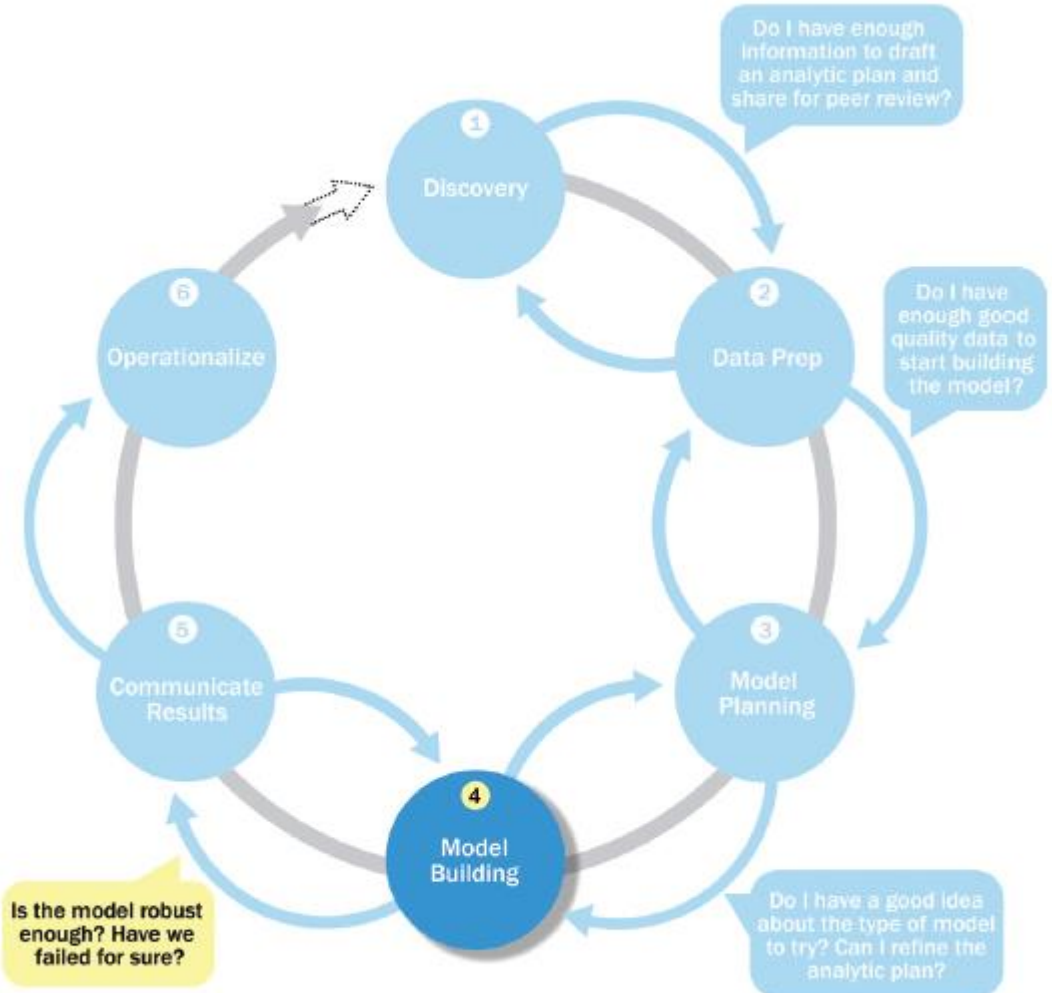


FIGURE 2-6 *Model building phase*

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 test/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, pytorch, scikit-learn
  - …

Phase 5: Communicate Results

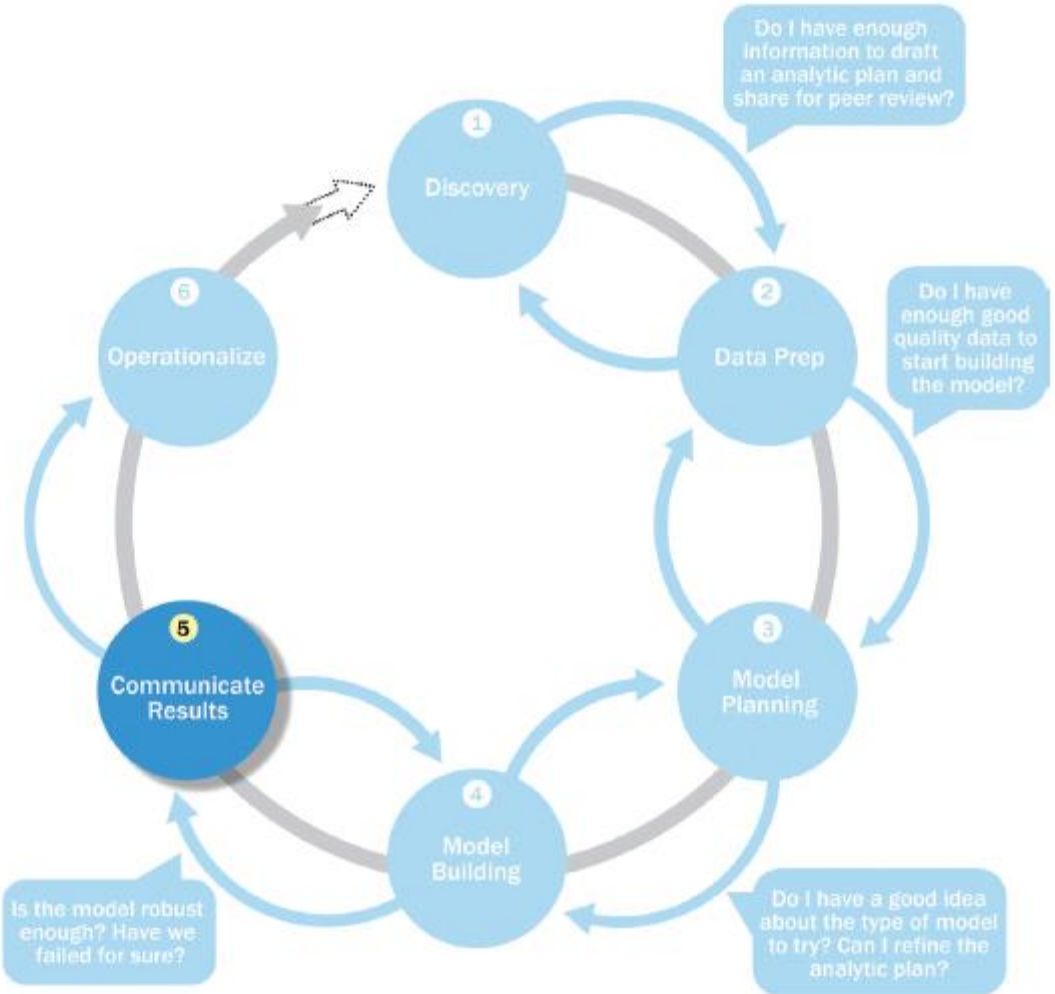


FIGURE 2-7 *Communicate results phase*

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

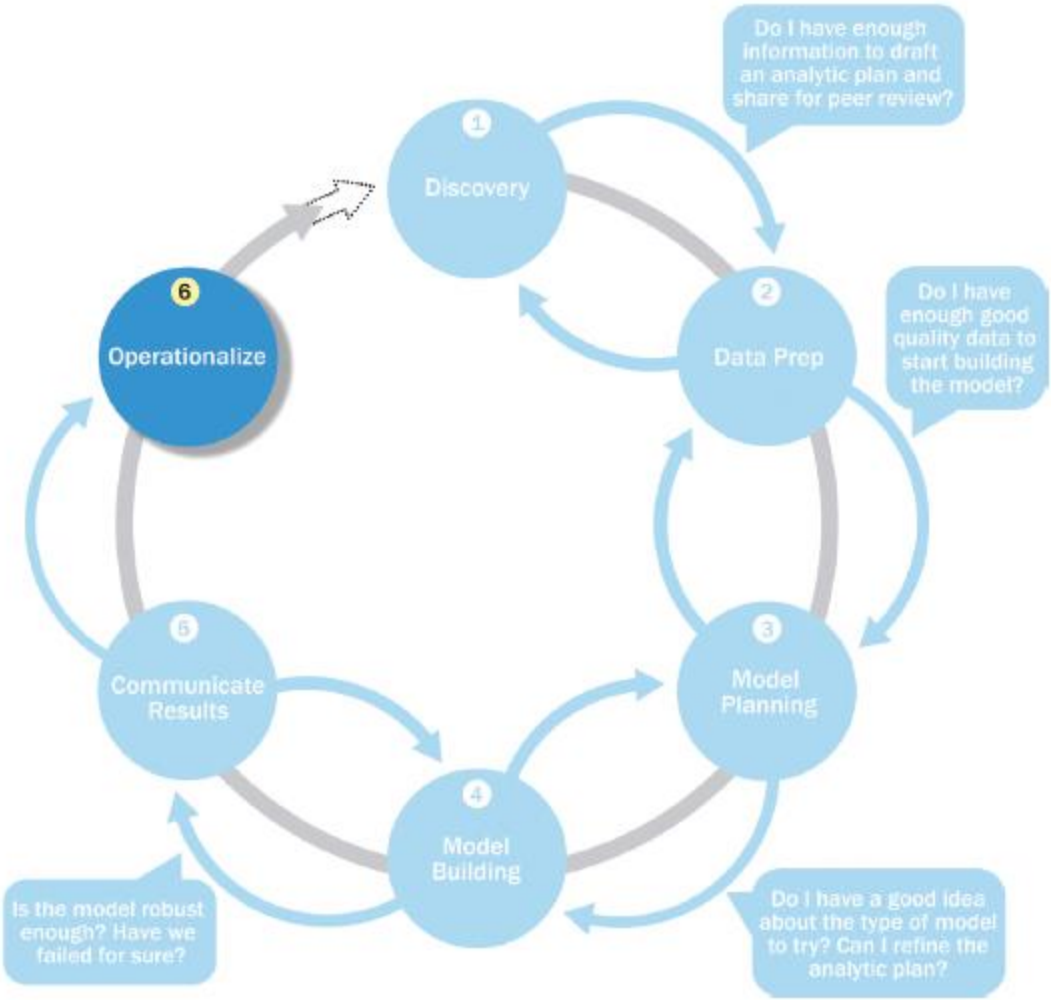


FIGURE 2-8 *Model operationalize phase*

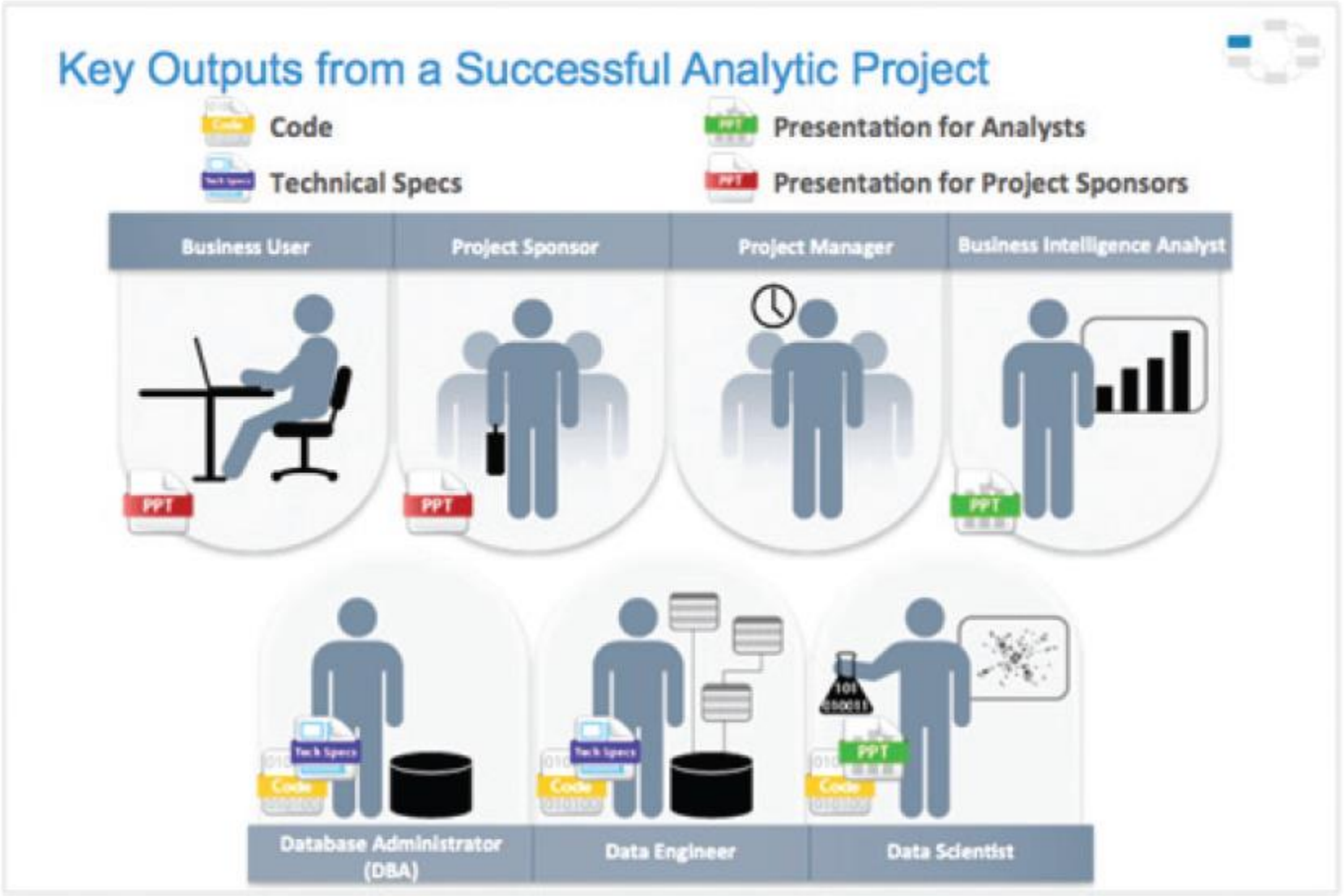
Phase 6: Operationalize

- 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 types of 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

A Simplified Example

Example: A hospital wants to reduce patient readmission rates.

- **Discovery**
  - **Example**: A healthcare organization wants to reduce patient readmission rates.
  - **Identify Problem**: High readmission rates lead to increased costs and lower patient satisfaction.
  - **Stakeholder Engagement**: Discuss with doctors, nurses, and administrators to understand the factors contributing to readmissions.
  - **Data Sources**: Identify relevant data sources such as patient records, treatment histories, and demographic data.

A Simplified Example

- **Data Preparation -** Preparing the healthcare data for analysis.
  - **Data Collection**: Extract patient records, treatment histories, and demographic data from electronic health records (EHR) systems.
  - **Data Cleaning**: Handle missing values, correct errors in patient records, and standardize formats (e.g., date formats).
  - **Data Integration**: Combine data from different sources into a unified dataset, ensuring consistency across variables.

A Simplified Example

- **Model Planning -** Planning a model to predict patient readmissions.
  - **Feature Selection**: Identify key features such as age, gender, diagnosis, treatment type, and length of stay that might influence readmissions.
  - **Algorithm Selection**: Choose suitable algorithms for the prediction task, such as logistic regression, decision trees, or random forests.
  - **Evaluation Criteria**: Define metrics for model evaluation, such as accuracy, precision, recall, and the F1-score.

A Simplified Example

- **Model Building-** Building and training the predictive model.
  - **Data Splitting**: Split the data into training and testing sets (e.g., 80% training, 20% testing).
  - **Training**: Train the chosen algorithms on the training data to learn patterns and relationships.
  - **Validation**: Use cross-validation techniques to validate the model and tune hyperparameters for optimal performance.

A Simplified Example

- **Communicate Results -** Sharing the analysis and model results with stakeholders.
  - **Visualization**: Create visualizations such as confusion matrices, ROC curves, and feature importance charts to explain model performance.
  - **Reporting**: Generate a comprehensive report detailing the analysis process, model outcomes, and actionable insights.
  - **Stakeholder Presentation**: Present the findings to healthcare providers, highlighting how the model can predict high-risk patients and recommending preventive measures.

A Simplified Example

- **Operationalize -** Implementing the predictive model in the healthcare setting.
  - **Integration**: Integrate the predictive model into the hospital's EHR system to provide real-time readmission risk scores for patients.
  - **Monitoring**: Continuously monitor model performance and retrain it periodically with new data to maintain accuracy.
  - **Actionable Use**: Develop workflows where high-risk patients identified by the model receive targeted interventions such as follow-up calls, personalized care plans, and closer monitoring.

Q&A

Images Courtesy of Google Image