




Data Analytics Lifecycle

- 
- 1. What are the 9 characteristics of big data. Give example of each.**
 - 2. List down the data sources of a Hospital Management System and categorize it under 3 vs (volume, velocity, variety)**
 - 3. Explain the types of analytics and justify each with a suitable example.**
 - 4. List down the Challenges of Traditional Systems and also Big Data Computing (At least 5)**
 - 5. Explain the concept of Scalability in Big Data Analytics. How it can be overcome. Explain**
 - 6. What do you understand by ETL vs ELT. Justify which process is better in today's scenario.**



Data Analytics Lifecycle

- Data science projects differ from BI projects
 - More exploratory in nature
 - Critical to have a project process
 - Participants should be thorough and rigorous
- Break large projects into smaller pieces
- Spend time to plan and scope the work
- Documenting adds rigor and credibility



Data Analytics Lifecycle

- Data Analytics Lifecycle Overview
- Phase 1: Discovery
- Phase 2: Data Preparation
- Phase 3: Model Planning
- Phase 4: Model Building
- Phase 5: Communicate Results
- Phase 6: Operationalize
- Case Study: GINA

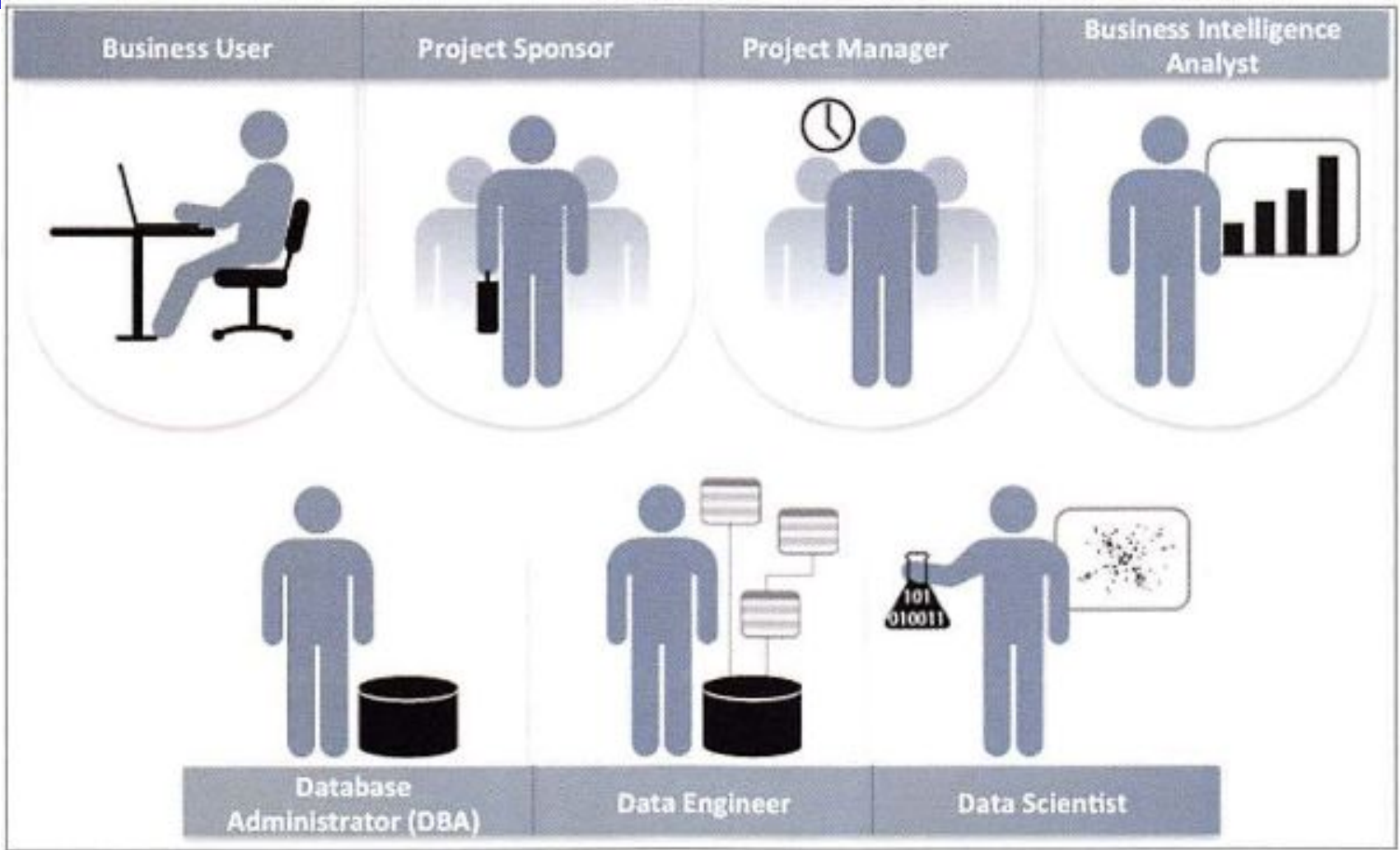
- 1. Explain four types of data. Give sources of each**
- 2. What are the 9 characteristics of big data. Give example of each**
- 3. Differentiate between Reporting and Analysis process.**
- 4. What do you understand by analytics? Explain the four types of analytics along with mapping with analytical focus and business value of each.**
- 5. List down the seven activities of Discovery phase?**
- 6. Explain the activities done in data preparation phase**
- 7. Give atleast five difference between the traditional analytics and data science**
- 8. What is the goal of Data Science. List down the challenges of data science projects.**
- 9. Write 8 hadoop ecosystem tools with their purpose.**

2.1 Data Analytics Lifecycle Overview



- The data analytic lifecycle is designed for Big Data problems and data science projects
- With six phases the project work can occur in several phases simultaneously
- The cycle is iterative to portray a real project
- Work can return to earlier phases as new information is uncovered

2.1.1 Key Roles for a Successful Analytics Project

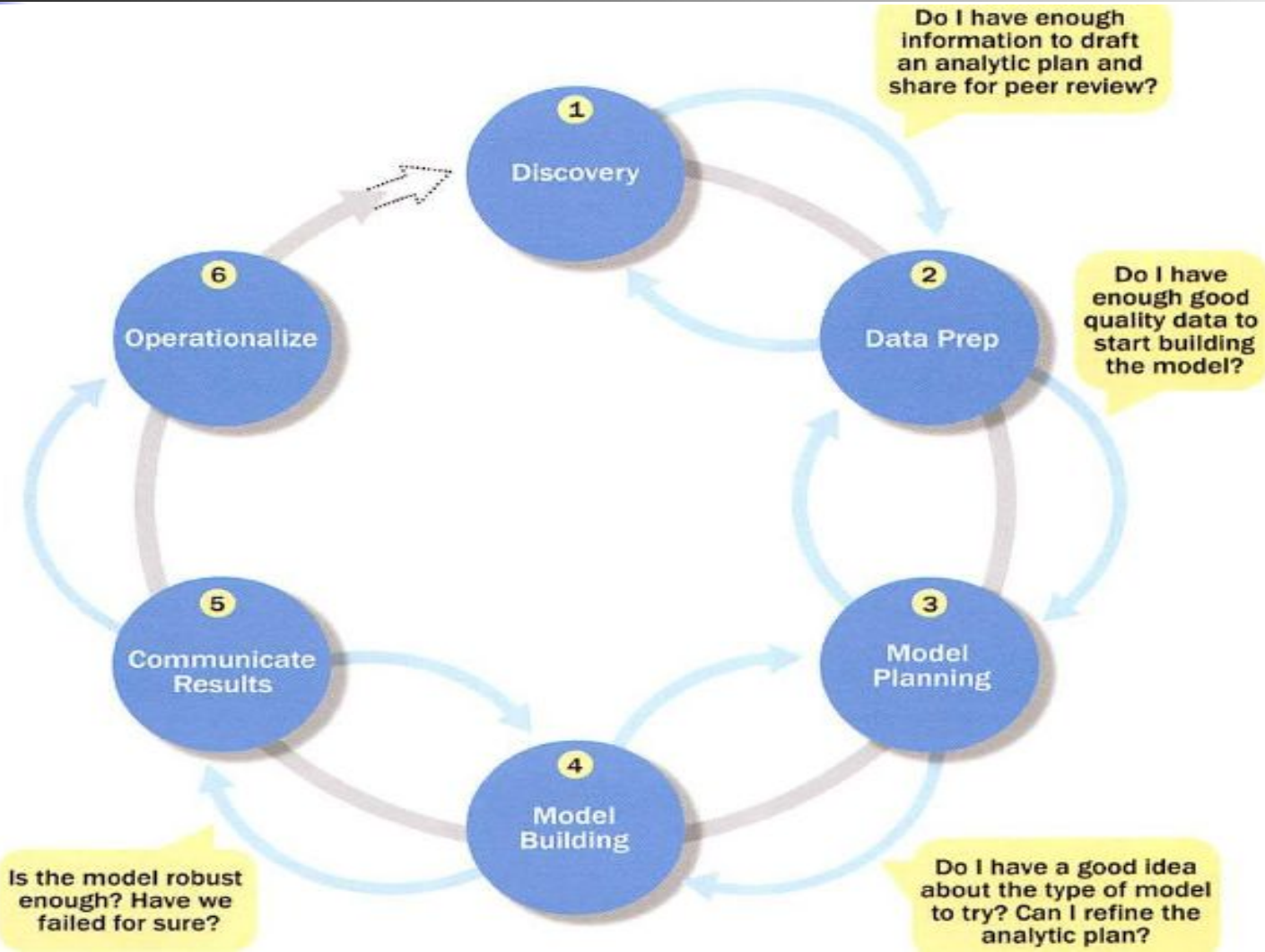




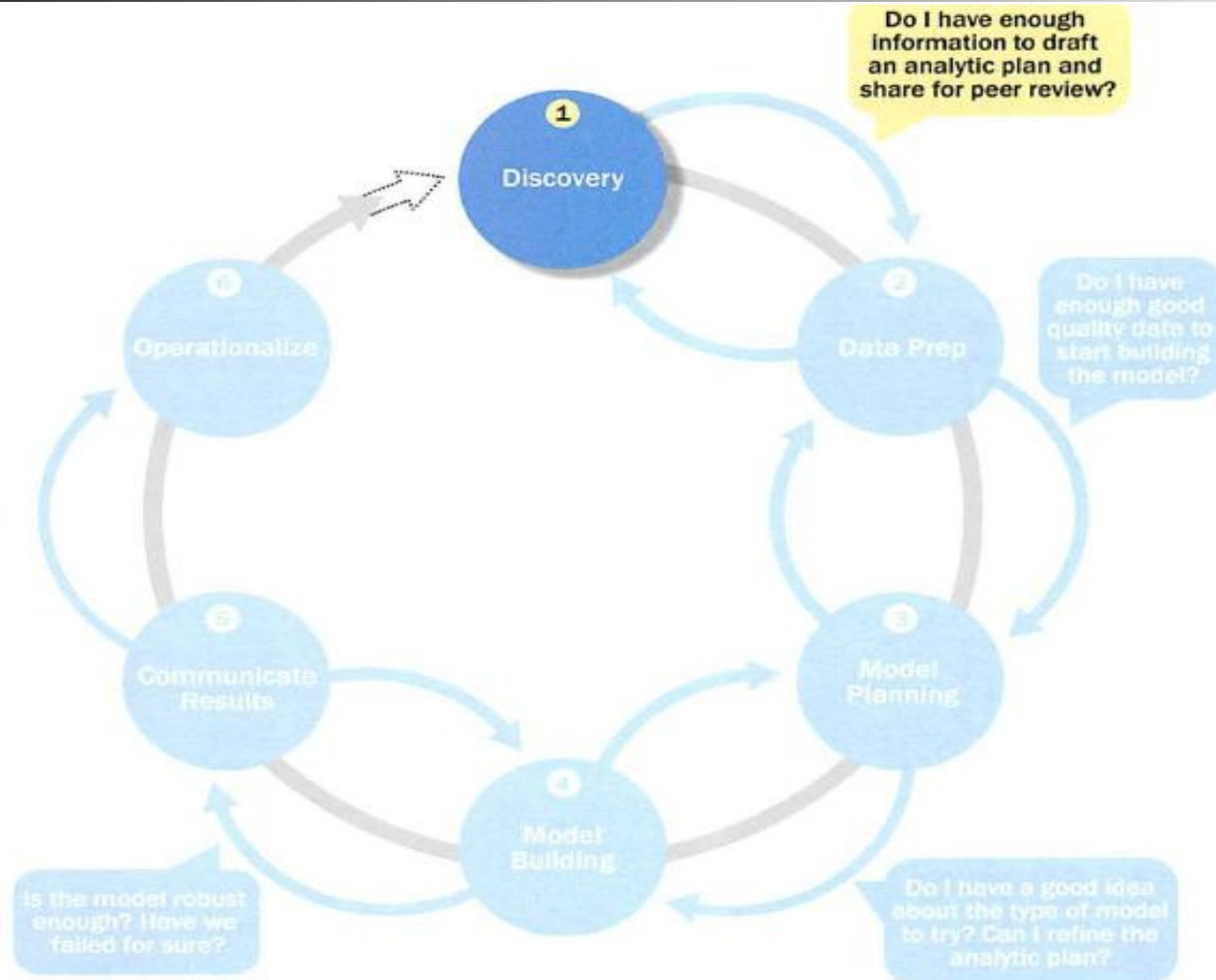
Key Roles for a Successful Analytics Project

- Business User – understands the domain area
- Project Sponsor – provides requirements
- Project Manager – ensures meeting objectives
- Business Intelligence Analyst – provides business domain expertise based on deep understanding of the data
- Database Administrator (DBA) – creates DB environment
- Data Engineer – provides technical skills, assists data management and extraction, supports analytic sandbox
- **Data Scientist** – provides analytic techniques and modeling

Overview of Data Analytics Lifecycle



2.2 Phase 1: Discovery

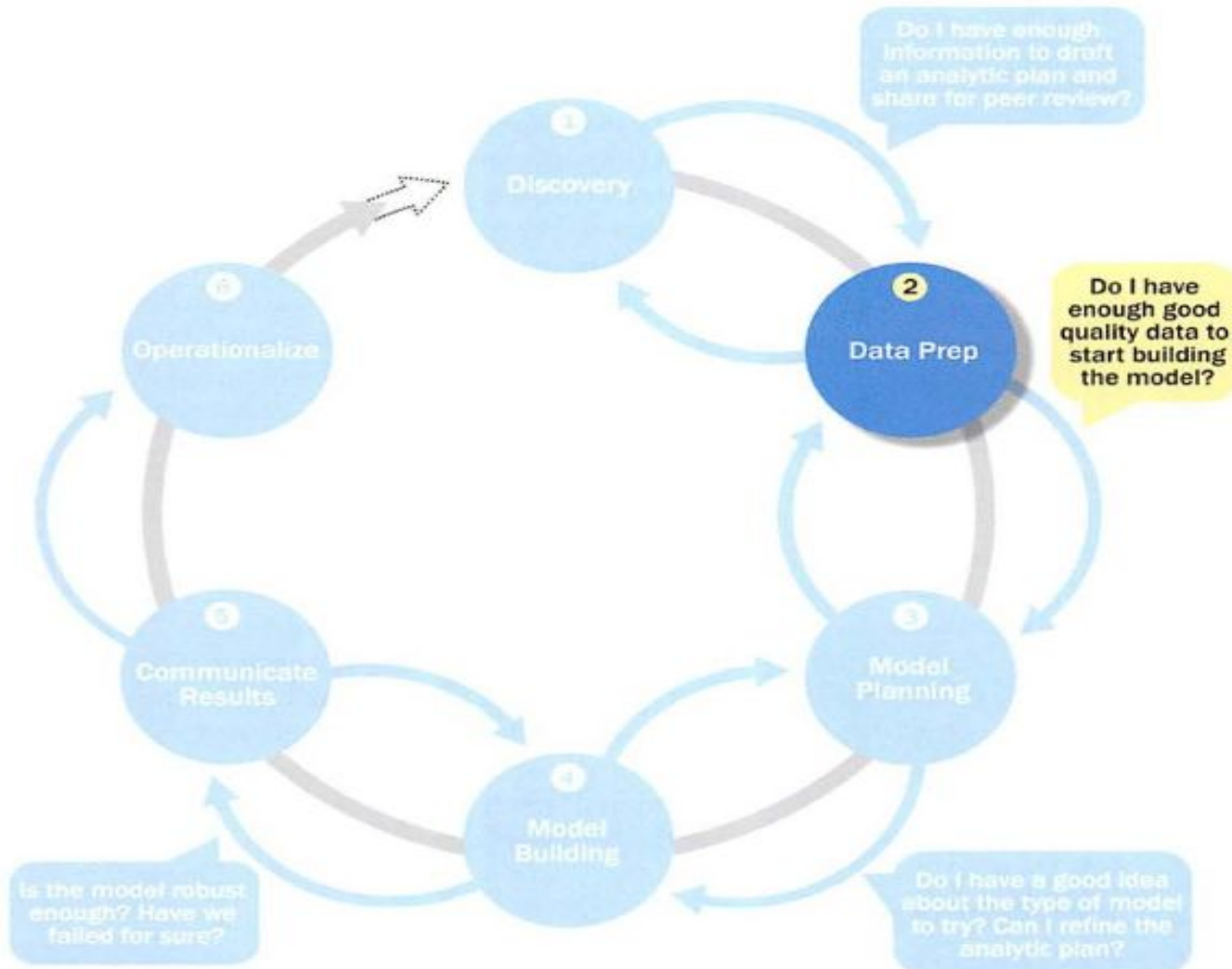




2.2 Phase 1: Discovery

1. Learning the Business Domain
2. Resources
3. Framing the Problem
4. Identifying Key Stakeholders
5. Interviewing the Analytics Sponsor
6. Developing Initial Hypotheses
7. Identifying Potential Data Sources

2.3 Phase 2: Data Preparation





2.3 Phase 2: Data Preparation

- Includes steps to explore, preprocess, and condition data
- Create robust environment – analytics sandbox
- Data preparation tends to be the most labor-intensive step in the analytics lifecycle
 - Often at least 50% of the data science project's time
- The data preparation phase is generally the most iterative and the one that teams tend to underestimate most often



2.3.1 Preparing the Analytic Sandbox

- Create the analytic sandbox (also called workspace)
- Allows team to explore data without interfering with live production data
- Sandbox collects all kinds of data (expansive approach)
- The sandbox allows organizations to undertake ambitious projects beyond traditional data analysis and BI to perform advanced predictive analytics
- Although the concept of an analytics sandbox is relatively new, this concept has become acceptable to data science teams and IT groups

2.3.2 Performing ETLT

(Extract, Transform, Load, Transform)

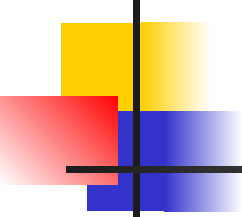
- In ETL users perform extract, transform, load
- In the sandbox the process is often ELT – early load preserves the raw data which can be useful to examine
- Example – in credit card fraud detection, outliers can represent high-risk transactions that might be inadvertently filtered out or transformed before being loaded into the database
- Hadoop is often used here



2.3.3 Learning about the Data

- Becoming familiar with the data is critical
- This activity accomplishes several goals:
 - Determines the data available to the team early in the project
 - Highlights gaps – identifies data not currently available
 - Identifies data outside the organization that might be useful

2.3.3 Learning about the Data Sample Dataset Inventory



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				●



2.3.4 Data Conditioning

- Data conditioning includes cleaning data, normalizing datasets, and performing transformations
 - Often viewed as a preprocessing step prior to data analysis, it might be performed by data owner, IT department, DBA, etc.
 - Best to have data scientists involved
 - Data science teams prefer more data than too little



2.3.5 Survey and Visualize

- Leverage data visualization tools to gain an overview of the data
- Shneiderman's mantra:
 - "Overview first, zoom and filter, then details-on-demand"
 - This enables the user to find areas of interest, zoom and filter to find more detailed information about a particular area, then find the detailed data in that area



2.3.5 Survey and Visualize Guidelines and Considerations

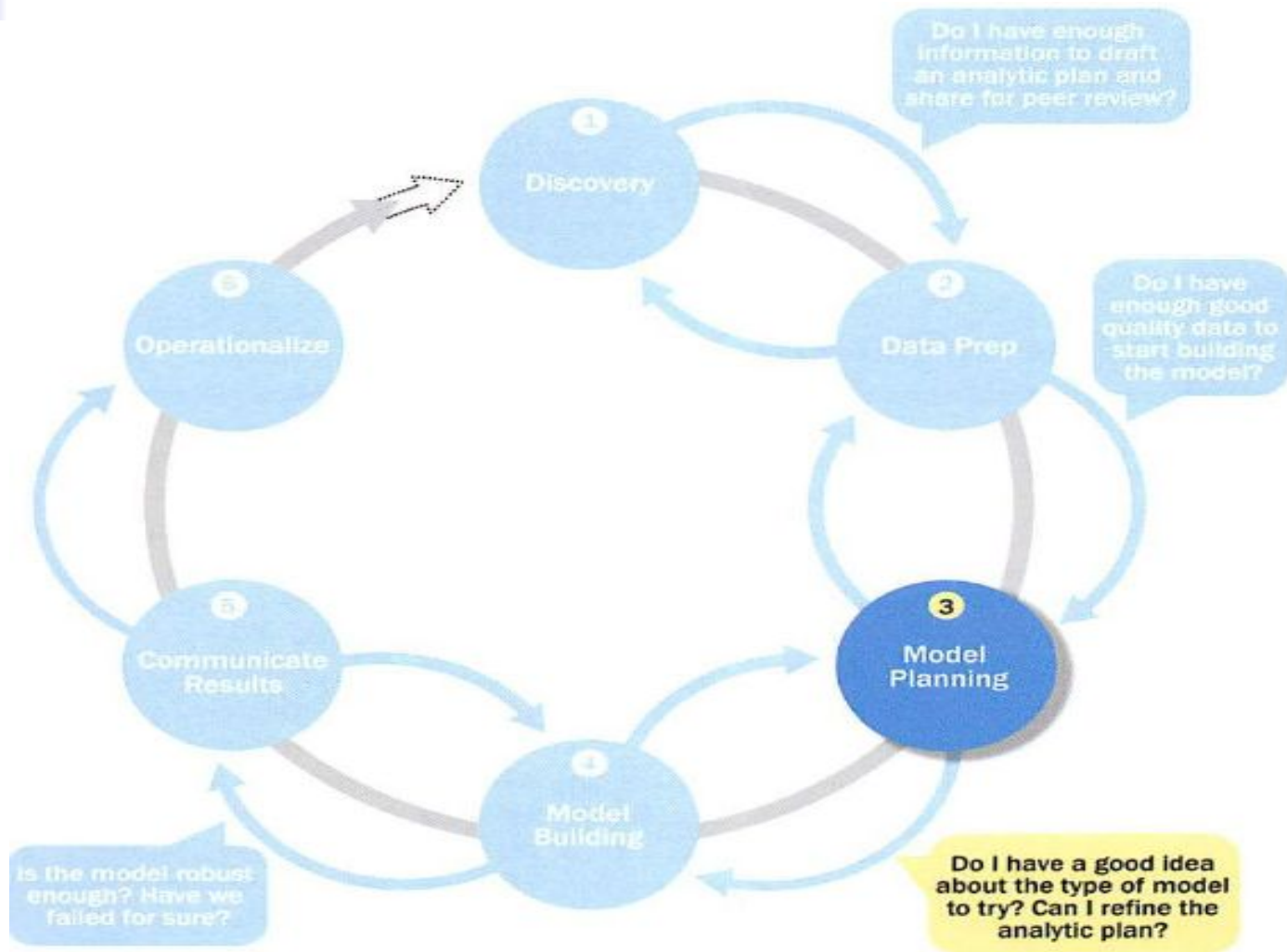
- Review data to ensure calculations are consistent
- Does the data distribution stay consistent?
- Assess the granularity of the data, the range of values, and the level of aggregation of the data
- Does the data represent the population of interest?
- Check time-related variables – daily, weekly, monthly? Is this good enough?
- Is the data standardized/normalized? Scales consistent?
- For geospatial datasets, are state/country abbreviations consistent



2.3.6 Common Tools for Data Preparation

- **Hadoop** can perform parallel ingest and analysis
- **Alpine Miner** provides a graphical user interface for creating analytic workflows
- **OpenRefine** (formerly Google Refine) is a free, open source tool for working with messy data
- Similar to OpenRefine, **Data Wrangler** is an interactive tool for data cleansing and transformation

2.4 Phase 3: Model Planning





2.4 Phase 3: Model Planning

- Activities to consider
 - Assess the **structure of the data** – this dictates the tools and analytic techniques for the next phase
 - Ensure the **analytic techniques** enable the team to meet the **business objectives and accept or reject the working hypotheses**
 - Determine if the **situation warrants a single model or a series of techniques** as part of a larger analytic workflow
 - Research and understand how other **analysts have approached this kind or similar kind of problem**



2.4 Phase 3: Model Planning

Model Planning in Industry Verticals

- Example of other analysts approaching a similar problem

Market Sector	Analytic Techniques/Methods Used
Consumer Packaged Goods	Multiple linear regression, automatic relevance determination (ARD), and decision tree
Retail Banking	Multiple regression
Retail Business	Logistic regression, ARD, decision tree
Wireless Telecom	Neural network, decision tree, hierarchical neurofuzzy systems, rule evolver, logistic regression



2.4.1 Data Exploration and Variable Selection

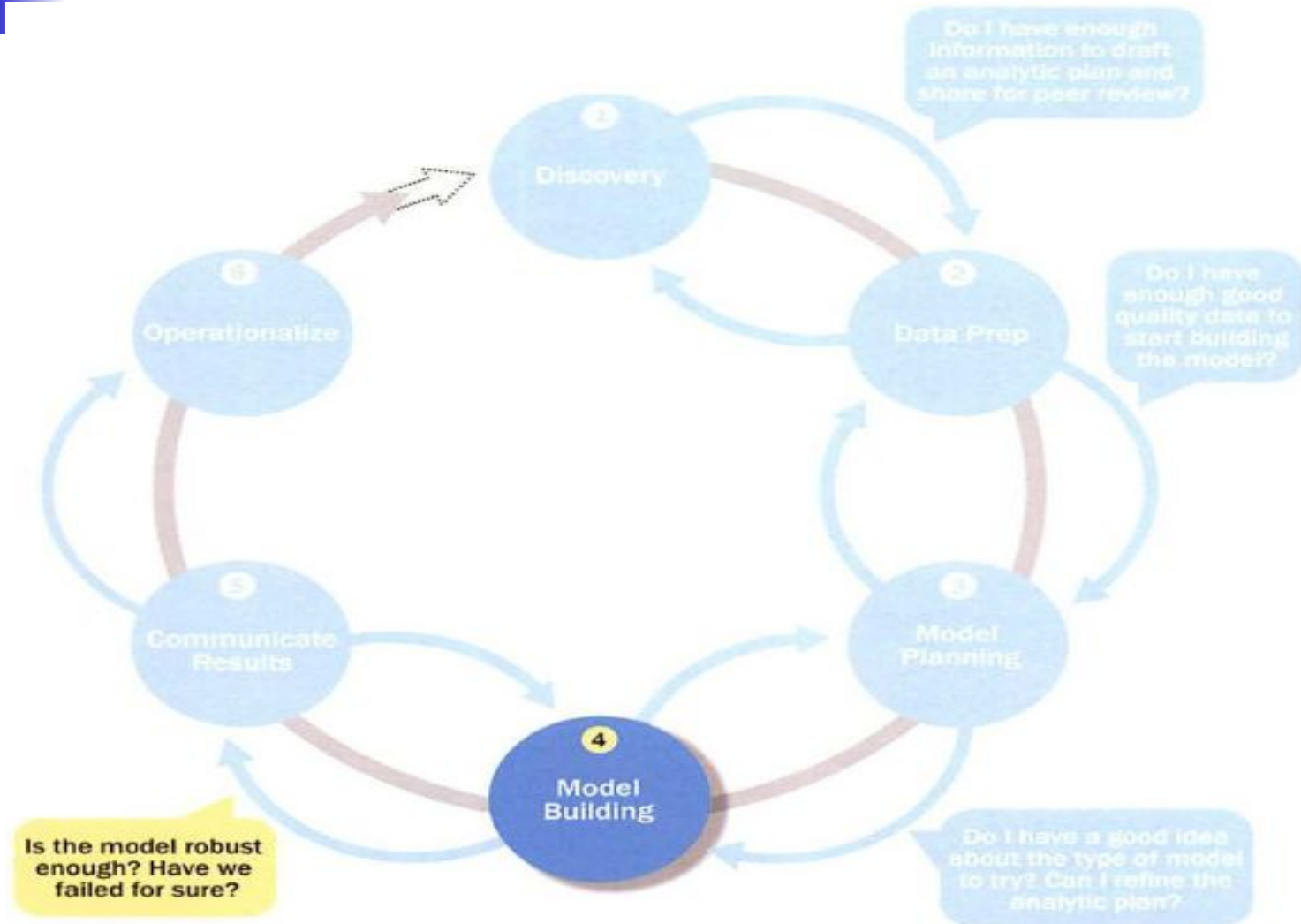
- Explore the data to **understand the relationships among the variables** to inform selection of the variables and methods
- A common way to do this is to use data visualization tools
- Aim for capturing the most essential predictors and variables
 - This often requires iterations and testing to identify key variables
- If the team plans to run regression analysis, identify the candidate predictors and outcome variables of the model



2.4.2 Model Selection

- The main goal is to choose an analytical technique, or several candidates, based on the end goal of the project
- Determine whether to use techniques best suited for structured data, unstructured data, or a hybrid approach
- Teams often create initial models using statistical software packages such as R, SAS, or Matlab
 - Which may have limitations when applied to very large datasets
- The team moves to the model building phase once it has a good idea about the type of model to try

2.5 Phase 4: Model Building

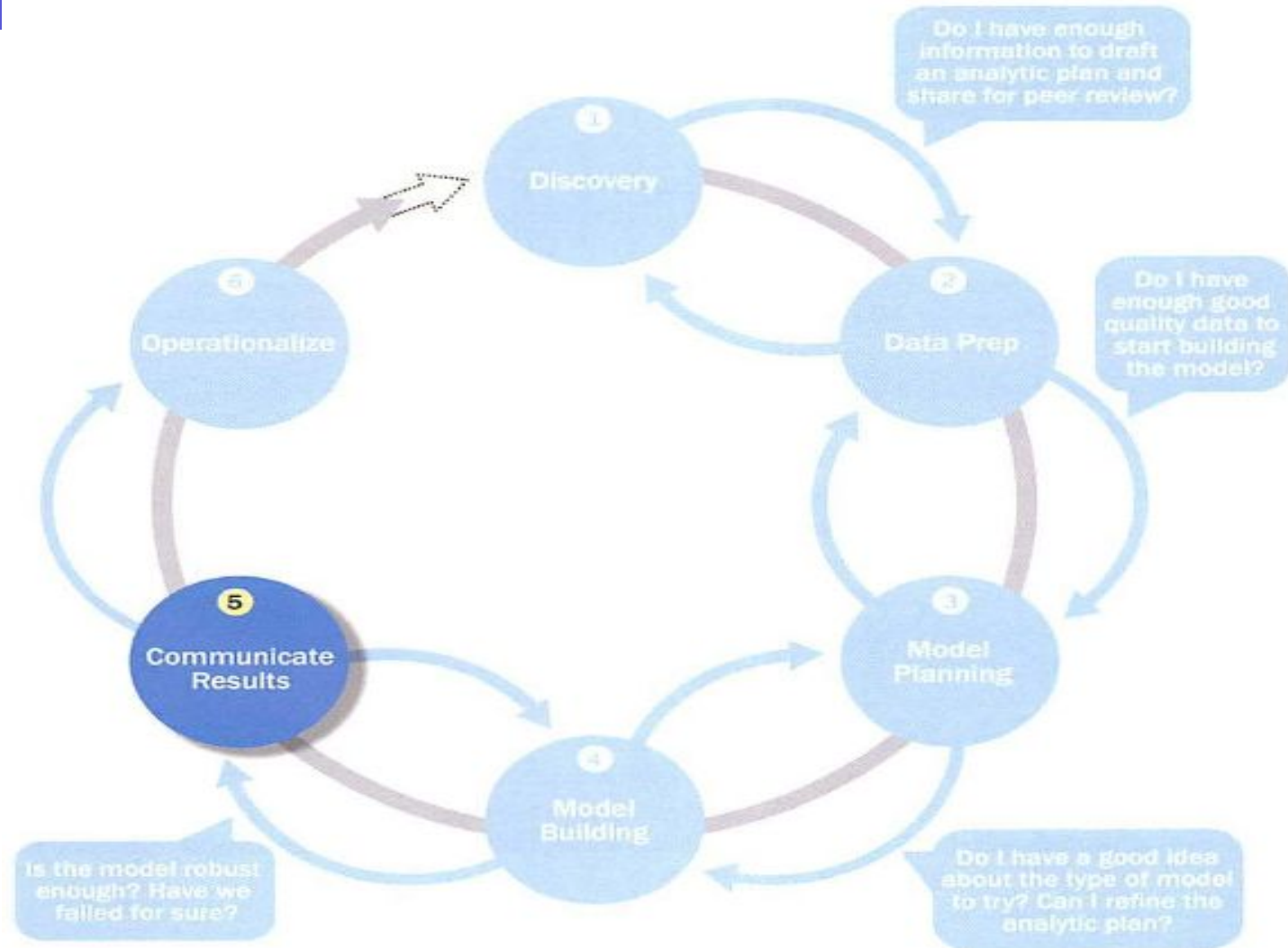




2.5 Phase 4: Model Building

- Execute the models defined in Phase 3
- Develop datasets for training, testing, and production
- Develop analytic model on training data, test on test data
- Question to consider
 - Does the model appear valid and accurate on the test data?
 - Does the model output/behavior make sense to the domain experts?
 - Do the parameter values make sense in the context of the domain?
 - Is the model sufficiently accurate to meet the goal?
 - Are more data or inputs needed?
 - Is a different form of the model required to address the business problem?

2.6 Phase 5: Communicate Results

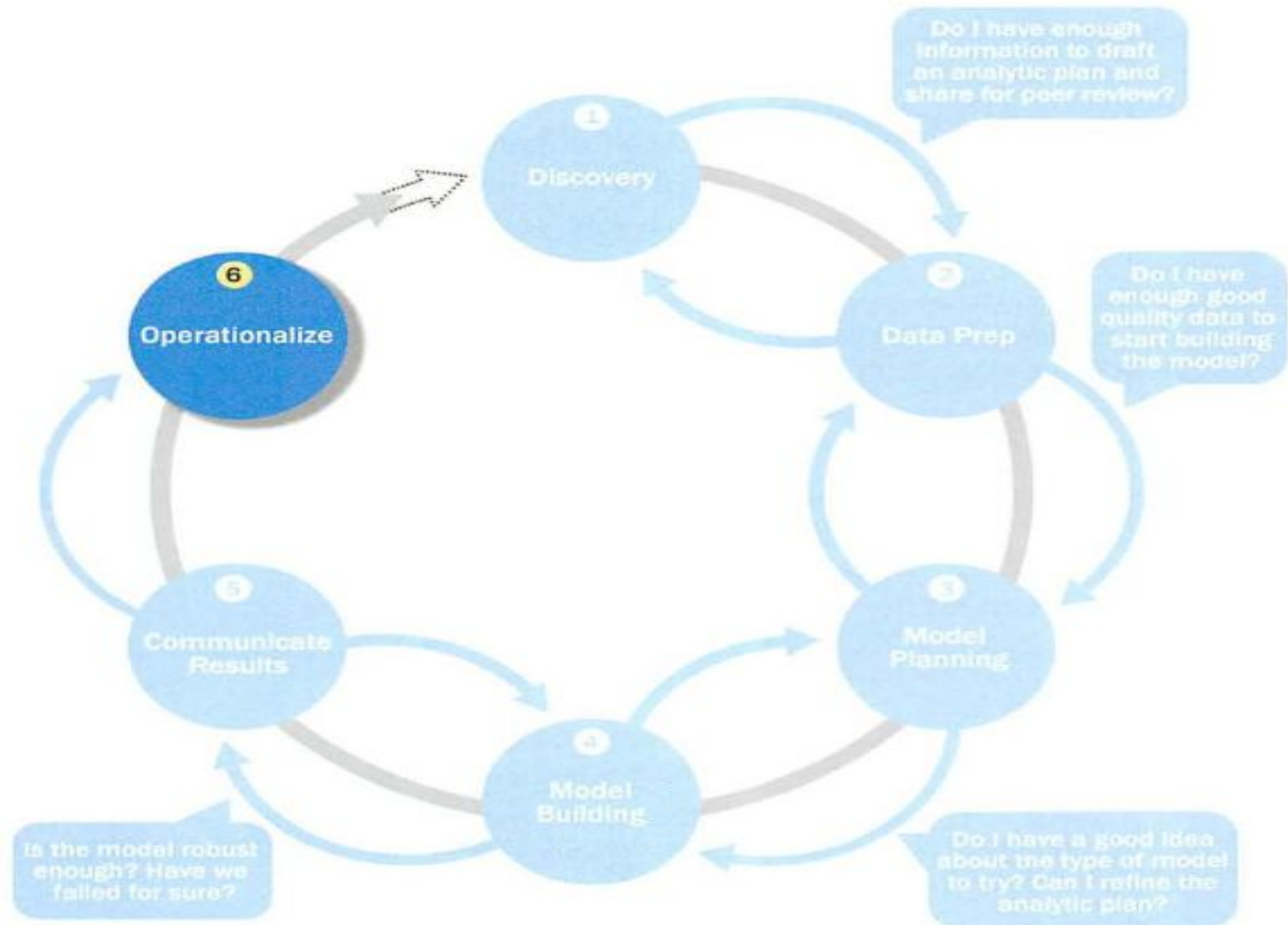


2.6 Phase 5: Communicate Results



- Determine if the team succeeded or failed in its objectives
- Assess if the results are statistically significant and valid
 - If so, identify aspects of the results that present salient findings
 - Identify surprising results and those in line with the hypotheses
- Communicate and document the key findings and major insights derived from the analysis
 - This is the most visible portion of the process to the outside stakeholders and sponsors

2.7 Phase 6: Operationalize





2.7 Phase 6: Operationalize

- In this last phase, the team communicates the benefits of the project more broadly and sets up a pilot project to deploy the work in a controlled way
- Risk is managed effectively by undertaking small scope, pilot deployment before a wide-scale rollout
- During the pilot project, the team may need to execute the algorithm more efficiently in the database rather than with in-memory tools like R, especially with larger datasets



Terminologies used in Big Data cont'd



56

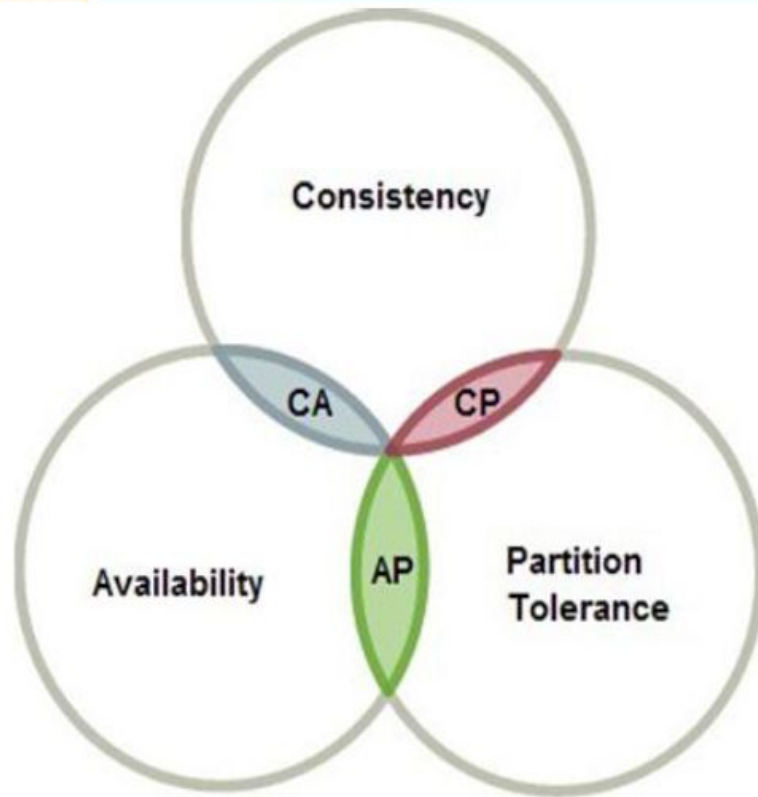
CAP Theorem: In the past, when we wanted to store more data or increase our processing power, the common option was to scale vertically (get more powerful machines) or further optimize the existing code base. However, with the advances in parallel processing and distributed systems, it is more common to expand horizontally, or have more machines to do the same task in parallel. However, in order to effectively pick the tool of choice like Spark, Hadoop, Kafka, Zookeeper and Storm in Apache project, a basic idea of CAP Theorem is necessary. The CAP theorem is called the **Brewer's Theorem**. It states that a distributed computing environment can only have 2 of the 3: **C**onsistency, **A**vailability and **P**artition Tolerance – one must be sacrificed.

- ❑ **Consistency** implies that every read fetches the last write
- ❑ **Availability** implies that reads and write always succeed. In other words, each non-failing node will return a response in a reasonable amount of time
- ❑ **Partition Tolerance** implies that the system will continue to function when network partition occurs

CAP Theorem cont'd



57



Source: Towards Data Science

The CAP theorem categorizes systems into three categories:

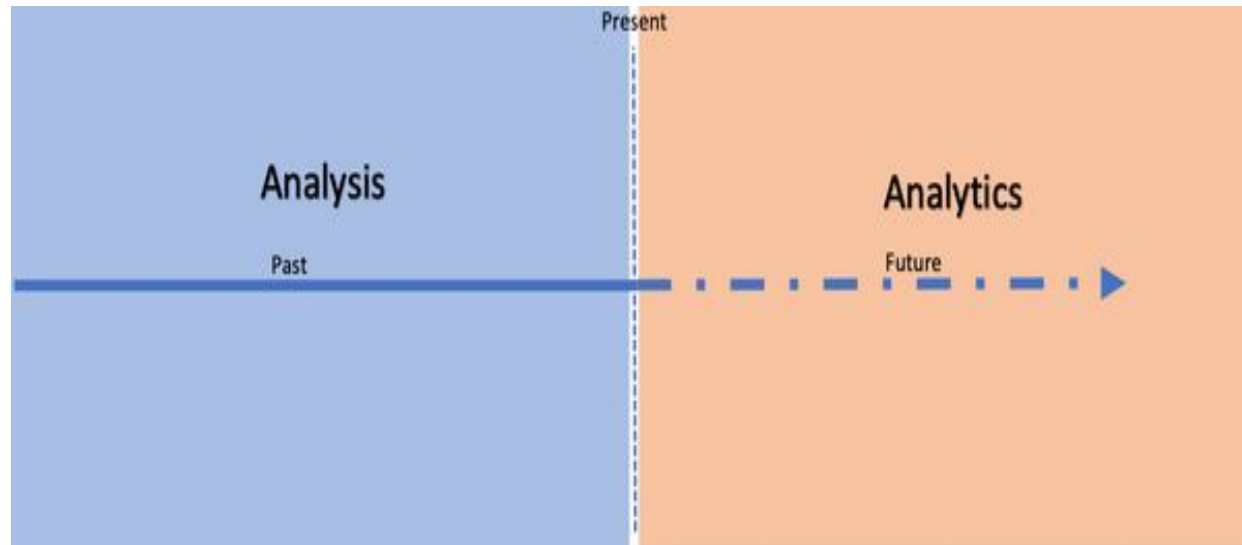
CP (Consistent and Partition Tolerant) - a system that is consistent and partition tolerant but never available. CP is referring to a category of systems where availability is sacrificed only in the case of a network partition.

CA (Consistent and Available) - CA systems are consistent and available systems in the absence of any network partition. Often a single node's DB servers are categorized as CA systems. Single node DB servers do not need to deal with partition tolerance and are thus considered CA systems.

AP (Available and Partition Tolerant) - These are systems that are available and partition tolerant but cannot guarantee consistency.

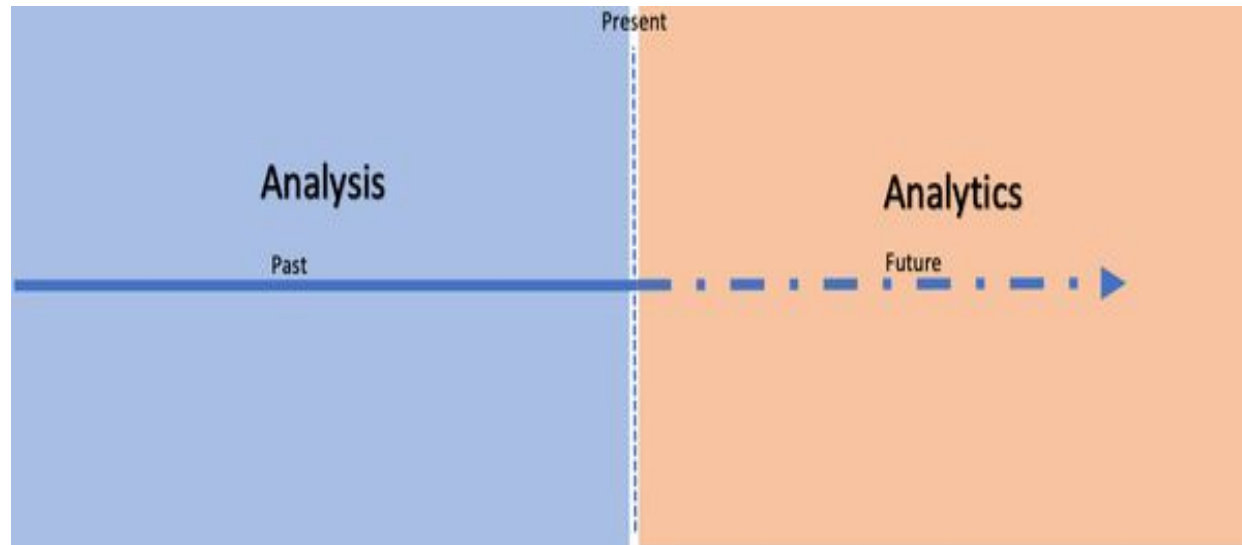
Chapter 2

Data Analysis



Chapter 2

Data Analysis



Learning Objectives

- The importance of data analytics in several classes of applications.
- Concepts of regression and its several variants.
- Bayes Rules and how it can be used to perform Bayesian Inference.
- The basic concepts of Support Vector Machines.
- The meaning of times series analysis, the various components of a times series and how decomposition can help prediction.
- How to extract rules to describe data from a large data set.

Introduction

- Recent rapid advances in computing, data storage, networks and sensors have dramatically increased our ability to **access, store and process huge amounts of data.**
- The fields of scientific research and business applications both are always challenged with the need to extract relevant information from huge amounts of data and **heterogeneous data sources, such as sensors, databases, text archives, images, audio and video streams, etc.**

Data Analysis

- It is a process of inspecting, cleaning, transforming and modelling data with the goal of discovering useful information, suggesting conclusions and supporting decision-making.
- **Intelligent data analysis (IDA)** uses concepts from artificial intelligence, information retrieval, machine learning, pattern recognition, visualization, distributed programming.
- The process of IDA typically consists of the following three stages:
 - Data preparation
 - Data mining and rule finding
 - Result validation and interpretation
- It has multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science and social science domains.

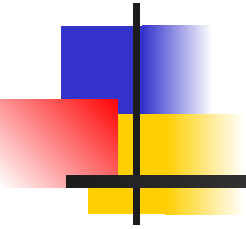
Importance of Data Analysis

- Data analysis offers the following benefits:
 - Structuring the findings from survey research or other means of data collection
 - Provides a picture of data at several levels of granularity from a macro picture into a micro one
 - ~~Acquiring meaningful insights from the data set which can be effectively exploited to take some critical decisions to improve productivity~~
 - Helps to remove human bias in decision making, through proper statistical treatment
 - With the advent of big data, it is even more vital to find a way to analyze the ever (faster) growing disparate data coursing through their environments and give it meaning

Regression Modelling Techniques

- Linear Regression
- Developing a Linear Regression Model
- Multiple Linear Regression (MLR)
- Non-Linear Regression
- Logistic Regression
- Classical Multivariate Analysis

Regression




Basic idea:

Use data to identify **relationships** among variables and use these relationships to make **predictions**.



Regression (Meaning)

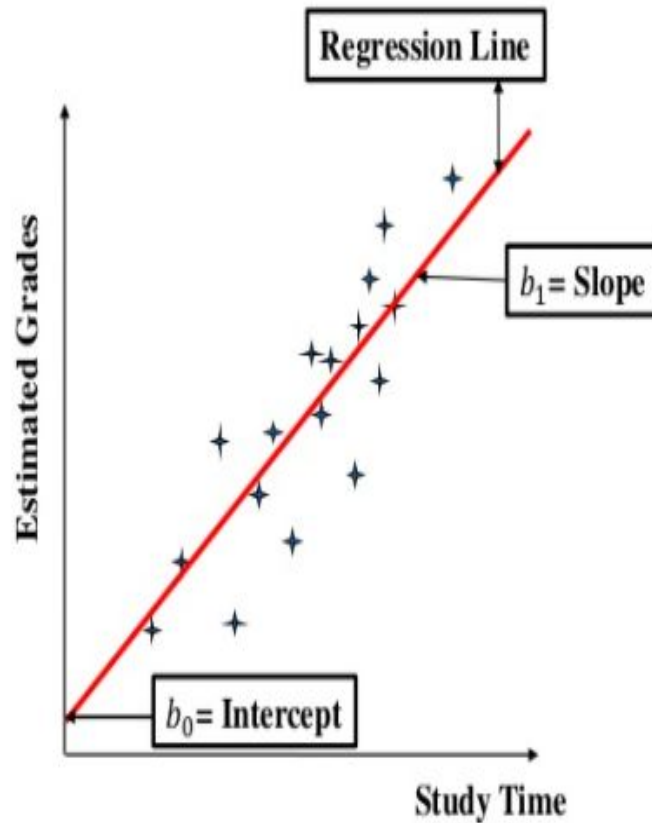
- “To move backwards”.
- “Return to an earlier time or stage”
- Re occurrences of trends

- 
- One fundamental task in data analysis is to attempt to find **how different variables are related to each other**, and one of the central tools in statistics for learning about **such relationships is regression**.
-

- The basic idea behind regression is "Use the existing historical data to identify potential relationships among variables" and then "use these relationships to make predictions" about the future.
- Regression analysis is a statistical process for estimating the relationships among variables.
- It helps to model and analyze several variables when the focus is on the relationship between a **dependent variable** and one or more **independent variables** (or "predictors").

Population Regression Line

Example



Population regression function =

$$\hat{y} = b_0 + b_1x$$

\hat{y} = Estimated Grades

x = Study Time

b_0 = Intercept

b_1 = Slope



2.8.1 Phase 1: Discovery

- Team members and roles
 - Business user, project sponsor, project manager – Vice President from Office of CTO
 - BI analyst – person from IT
 - Data engineer and DBA – people from IT
 - Data scientist – distinguished engineer



2.8.1 Phase 1: Discovery

- The data fell into two categories
 - Five years of idea submissions from internal innovation contests
 - Minutes and notes representing innovation and research activity from around the world
- Hypotheses grouped into two categories
 - Descriptive analytics of what is happening to spark further creativity, collaboration, and asset generation
 - Predictive analytics to advise executive management of where it should be investing in the future



2.8.2 Phase 2: Data Preparation

- Set up an analytics sandbox
- Discovered that certain data needed conditioning and normalization and that missing datasets were critical
- Team recognized that poor quality data could impact subsequent steps
- They discovered many names were misspelled and problems with extra spaces
- These seemingly small problems had to be addressed



2.8.3 Phase 3: Model Planning

- The study included the following considerations
 - Identify the right milestones to achieve the goals
 - Trace how people move ideas from each milestone toward the goal
 - Track ideas that die and others that reach the goal
 - Compare times and outcomes using a few different methods

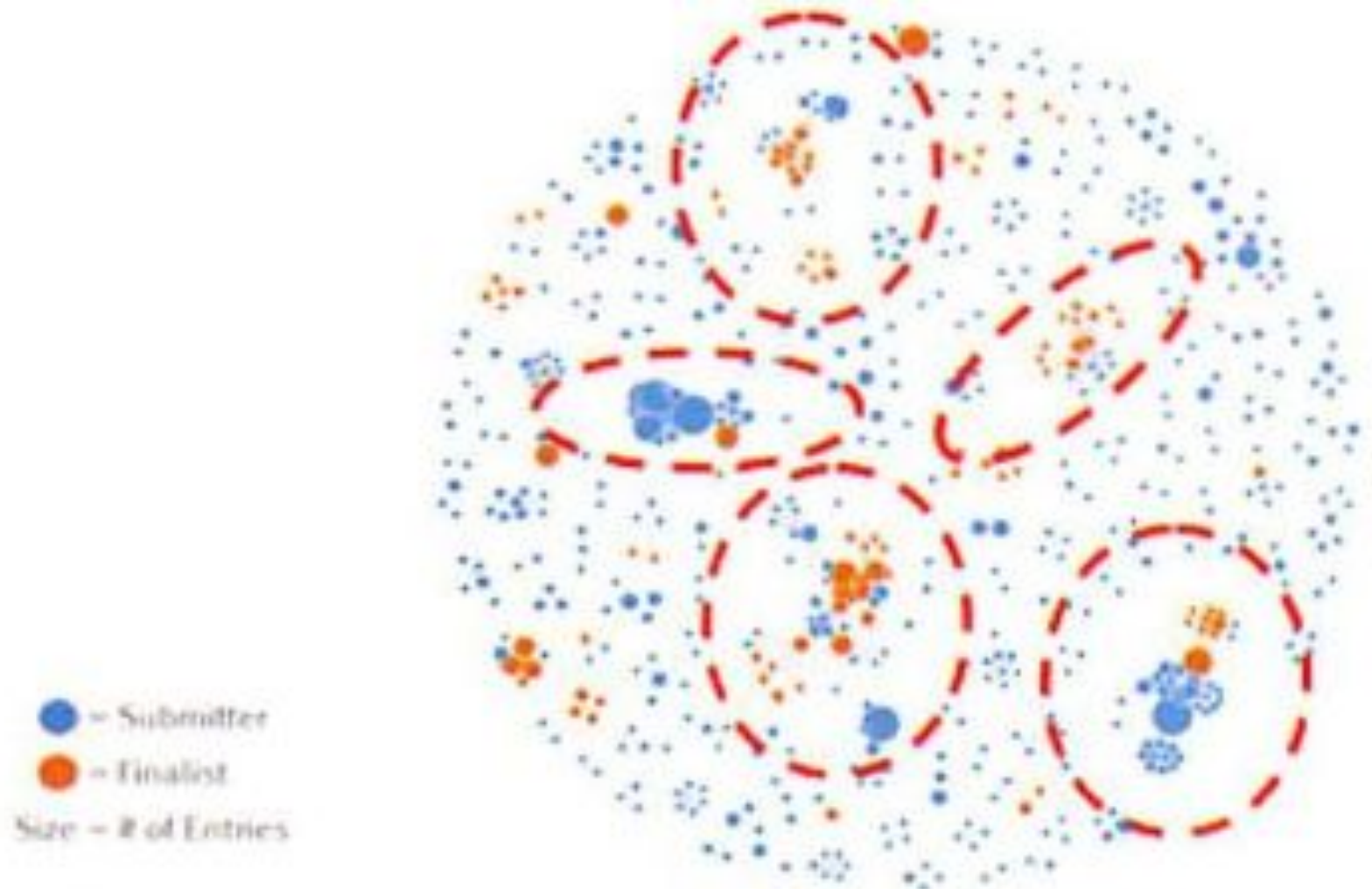


2.8.4 Phase 4: Model Building

- Several analytic method were employed
 - NLP on textual descriptions
 - Social network analysis using R and Rstudio
 - Developed social graphs and visualizations

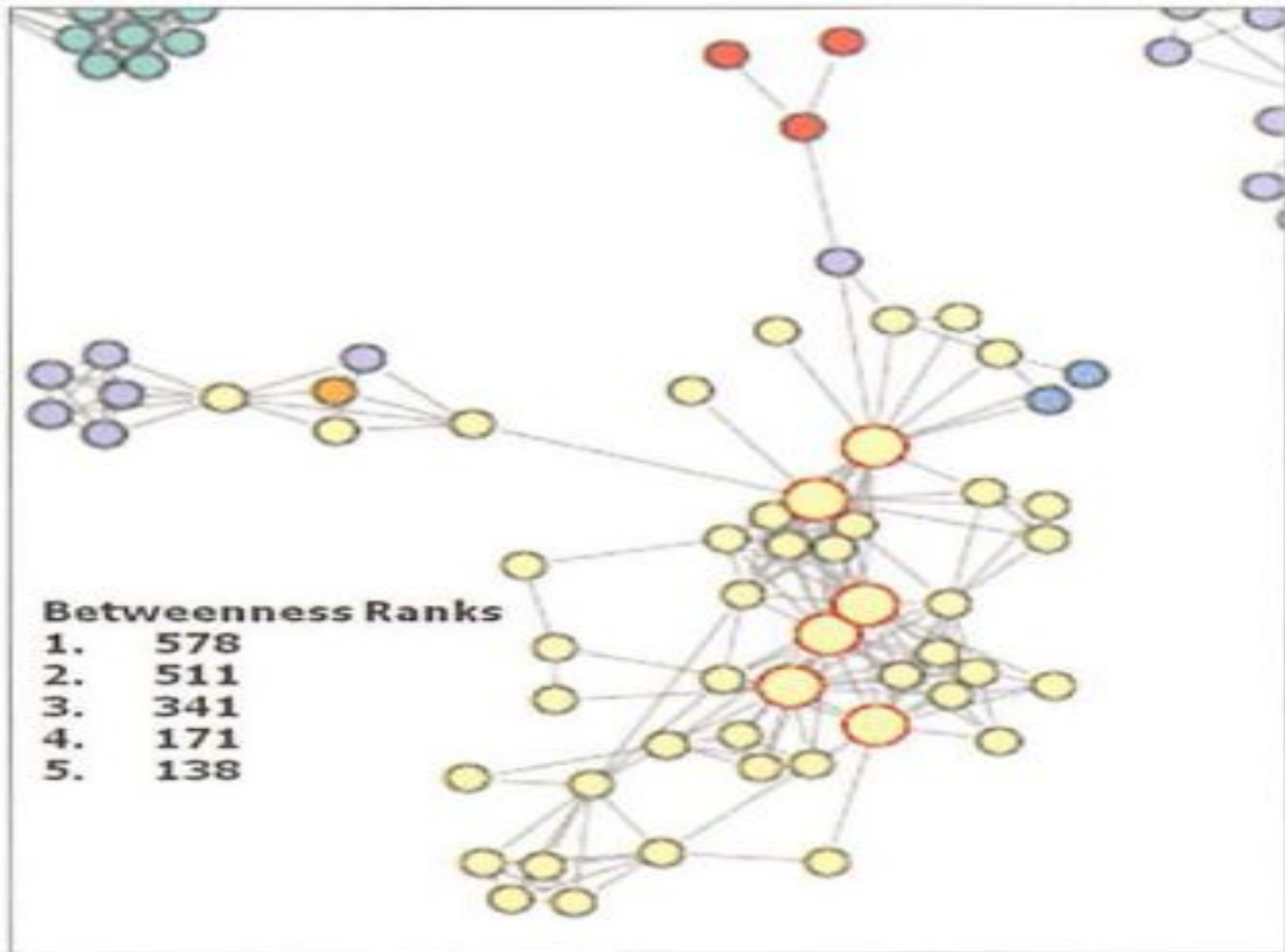
2.8.4 Phase 4: Model Building

Social graph of data submitters and finalists

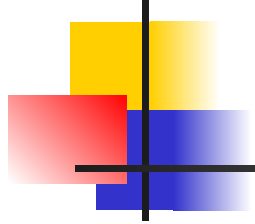


2.8.4 Phase 4: Model Building

Social graph of top innovation influencers



2.8.5 Phase 5: Communicate Results



- Study was successful in identifying hidden innovators
 - Found high density of innovators in Cork, Ireland
- The CTO office launched longitudinal studies



2.8.6 Phase 6: Operationalize

- Deployment was not really discussed
- Key findings
 - Need more data in future
 - Some data were sensitive
 - A parallel initiative needs to be created to improve basic BI activities
 - A mechanism is needed to continually reevaluate the model after deployment



2.8.6 Phase 6: Operationalize

Components of Analytic Plan	GINA Case Study
Discovery Business Problem Framed	Tracking global knowledge growth, ensuring effective knowledge transfer, and quickly converting it into corporate assets. Executing on these three elements should accelerate innovation.
Initial Hypotheses	An increase in geographic knowledge transfer improves the speed of idea delivery.
Data	Five years of innovation idea submissions and history; six months of textual notes from global innovation and research activities
Model Planning Analytic Technique	Social network analysis, social graphs, clustering, and regression analysis
Result and Key Findings	<ol style="list-style-type: none">1. Identified hidden, high-value innovators and found ways to share their knowledge2. Informed investment decisions in university research projects3. Created tools to help submitters improve ideas with idea recommender systems



Summary

- The Data Analytics Lifecycle is an approach to managing and executing analytic projects
- Lifecycle has six phases
- Bulk of the time usually spent on preparation – phases 1 and 2
- Seven roles needed for a data science team
- Review the exercises



Focus of Course

- Focus on quantitative disciplines – e.g., math, statistics, machine learning
- Provide overview of Big Data analytics
- In-depth study of a several key algorithms