

# **Artificial Intelligence & Data Science (Sem VI)**


## **ADC 601 : Data Analytics & Visualization**

### **Module - 1 : Introduction to Data analytics and life cycle (5 Hours)**

**Instructor : Mrs. Lifna C S**

# Topics to be covered

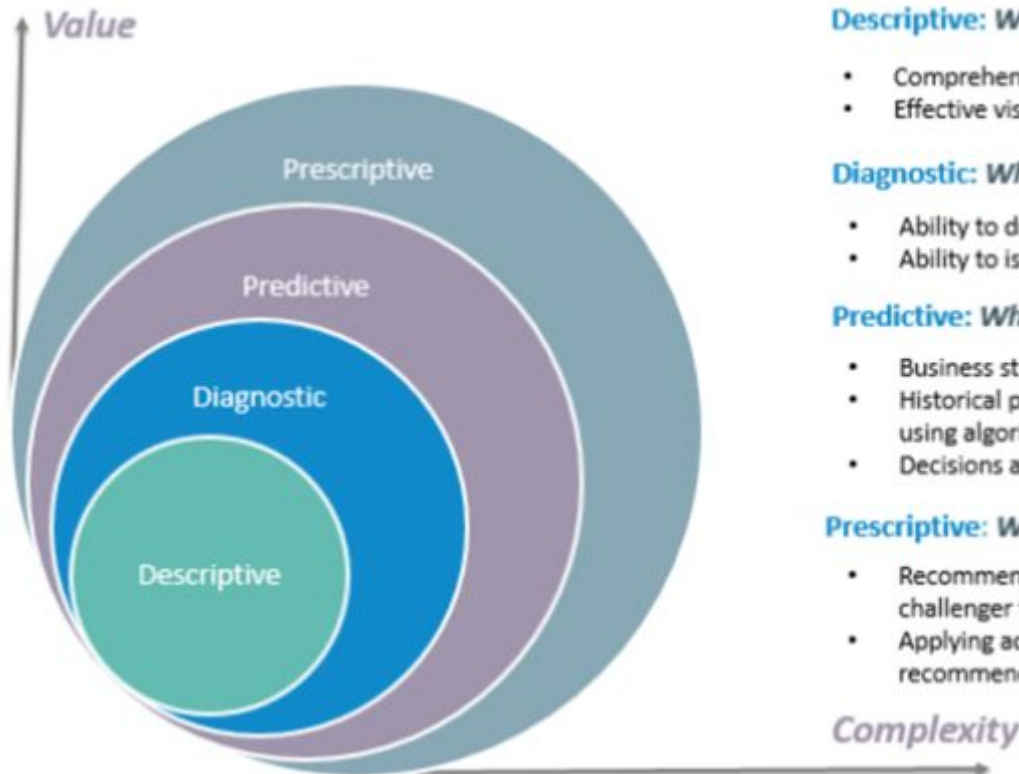
- Data Analytics Lifecycle overview:
  - Key Roles for a Successful Analytics,
  - Background and Overview of Data Analytics Lifecycle Project.
- Phase 1: Discovery
- Phase 2: Data Preparation
- Phase 3: Model Planning
- Phase 4: Model Building:
- Phase 5: Communicate Results
- Phase 6: Operationalize



## Data Analytics

*['dā-tə ə-nə-'li-tiks]*

The science of analyzing raw data to make conclusions about that information.



## **Descriptive:** *What's happening in my business?*

- Comprehensive, accurate and live data
- Effective visualisation

## **Diagnostic:** *Why is it happening?*

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

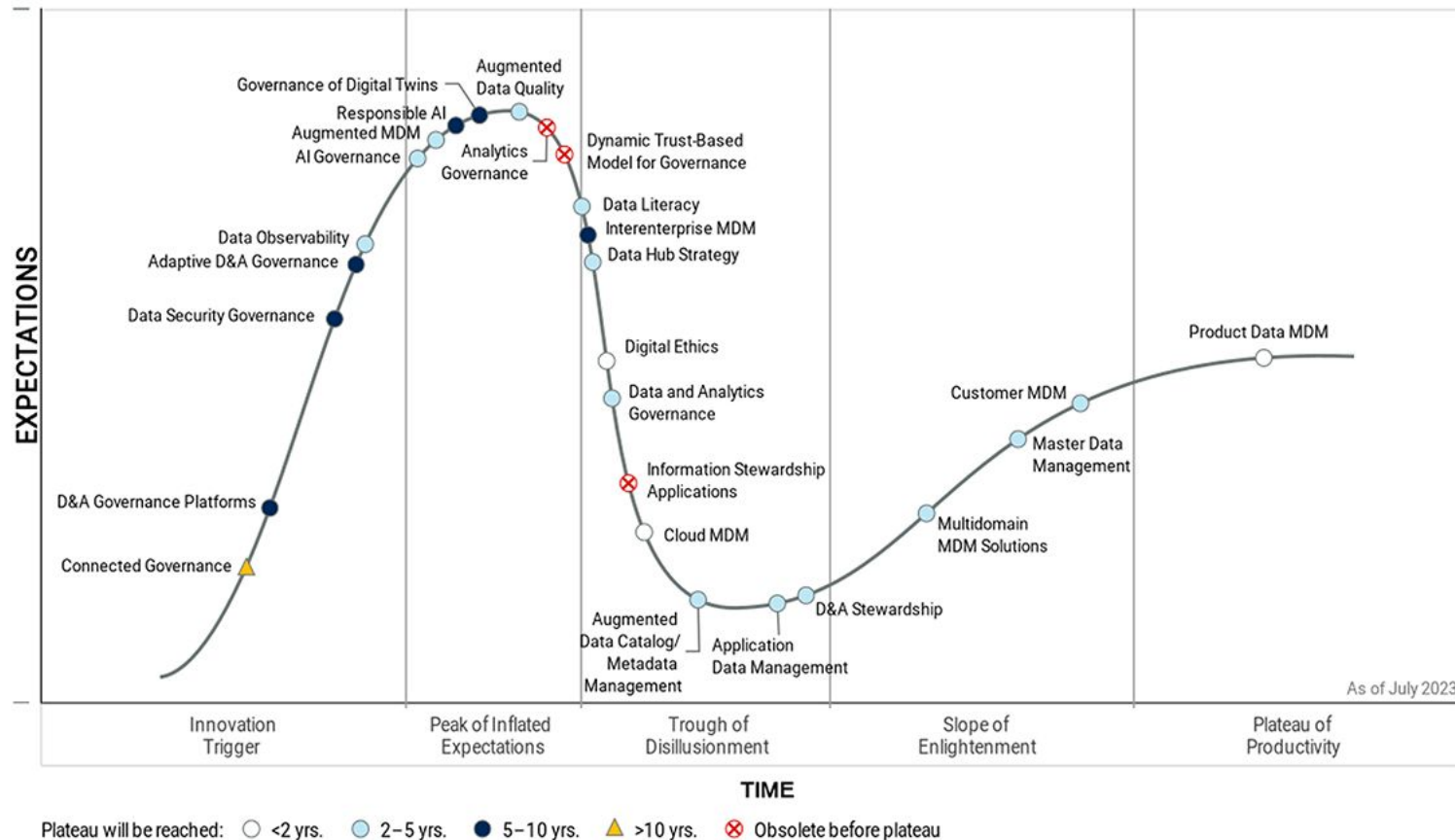
## **Predictive:** *What's likely to happen?*

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

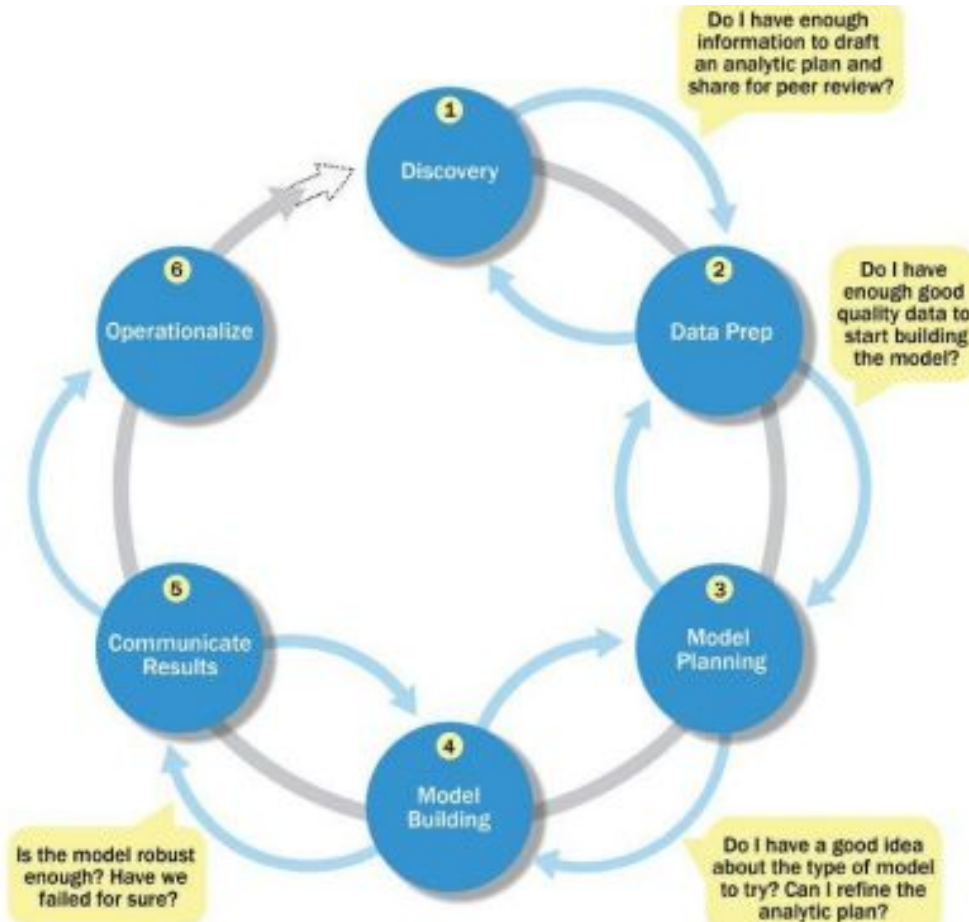
## **Prescriptive:** *What do I need to do?*

- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

# Hype Cycle for Data and Analytics Governance, 2023



# Data Analytics Lifecycle overview



## Circular arrows

- iterative movement between phases until the team members have sufficient information to move to the next phase.

## Callouts

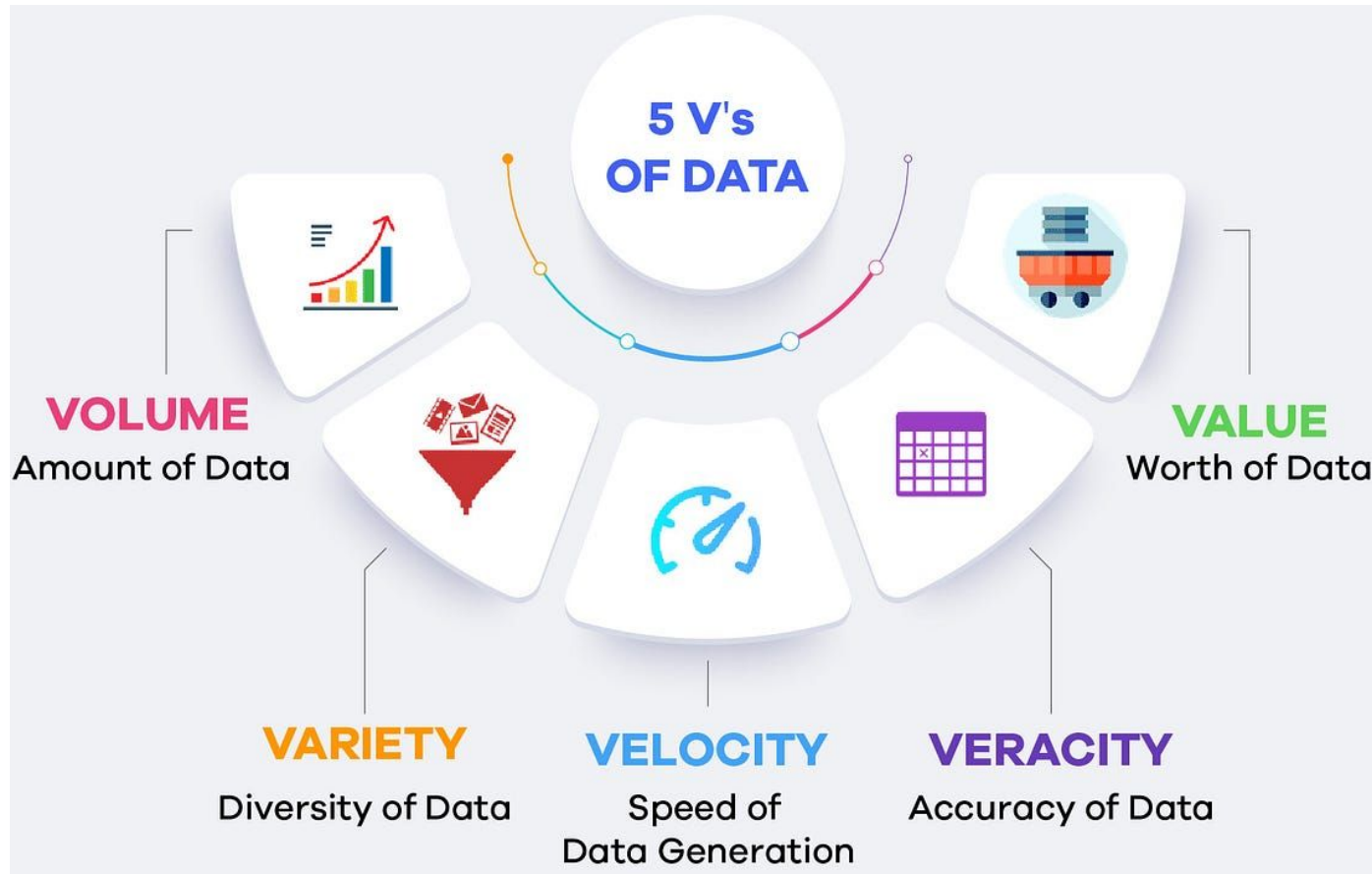
- include sample questions to ask to help guide whether each of the team members has enough information and has made enough progress to move to the next phase of the process

# Data Analytics Lifecycle overview

- Designed specifically for Big Data problems and data science projects.
- The life cycle has six phases, and project work can occur in several phases at once.
- For most phases in the life cycle, the movement can be either forward or backward.
- Iterative depiction of the lifecycle is intended to more closely portray a real project,
  - Teams commonly learn new things in a phase that cause them to go back and refine the work done in prior phases based on new insights and information that have been uncovered
- Enables participants to move iteratively through the process and drive toward operationalizing the project work.
- Defines analytics process best practices spanning discovery to project completion



# Big Data Overview





# Big Data - Sources



Mobile  
Sensors



Social  
Media



Video  
Surveillance



Video  
Rendering



Smart  
Grids



Geophysical  
Exploration

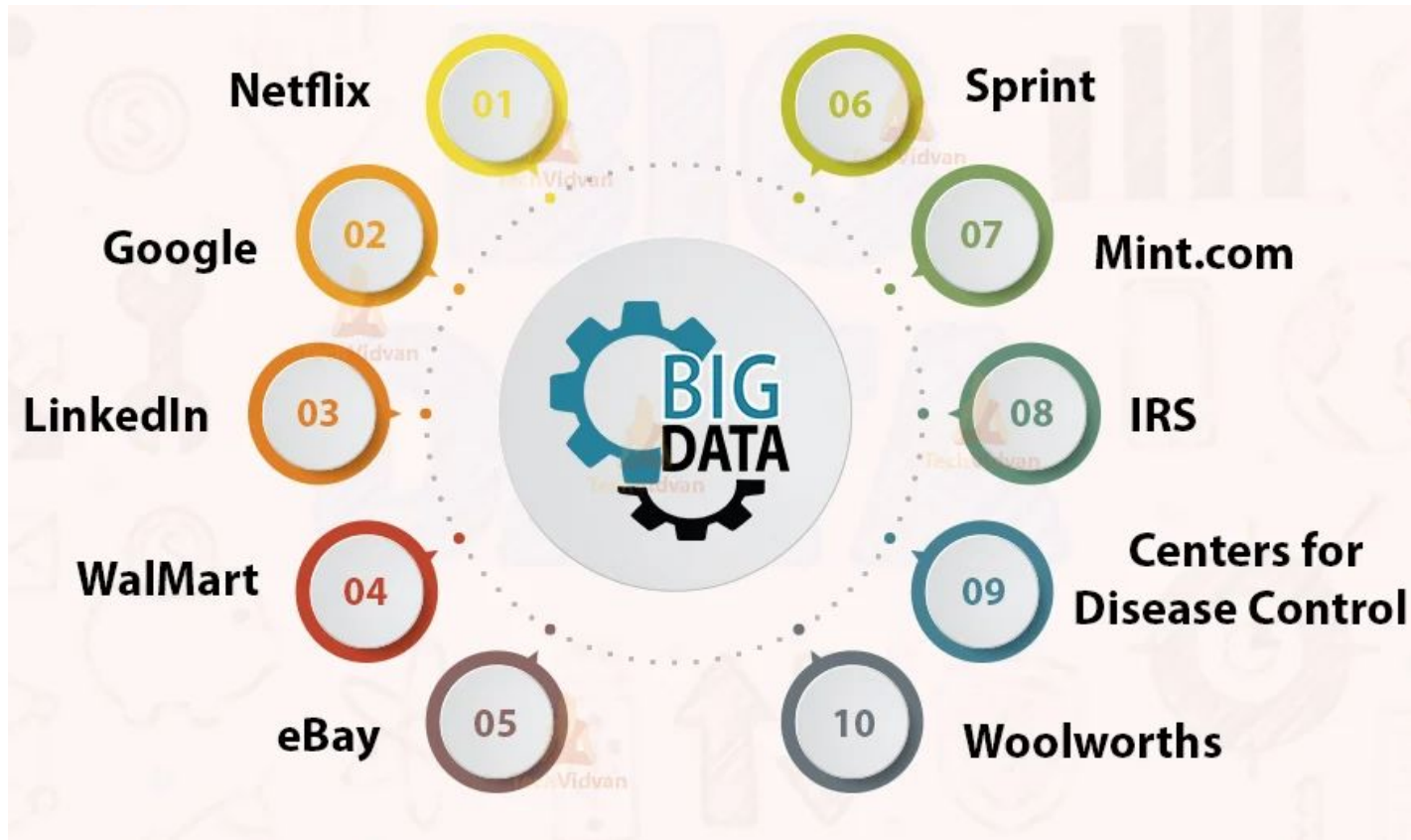


Medical  
Imaging

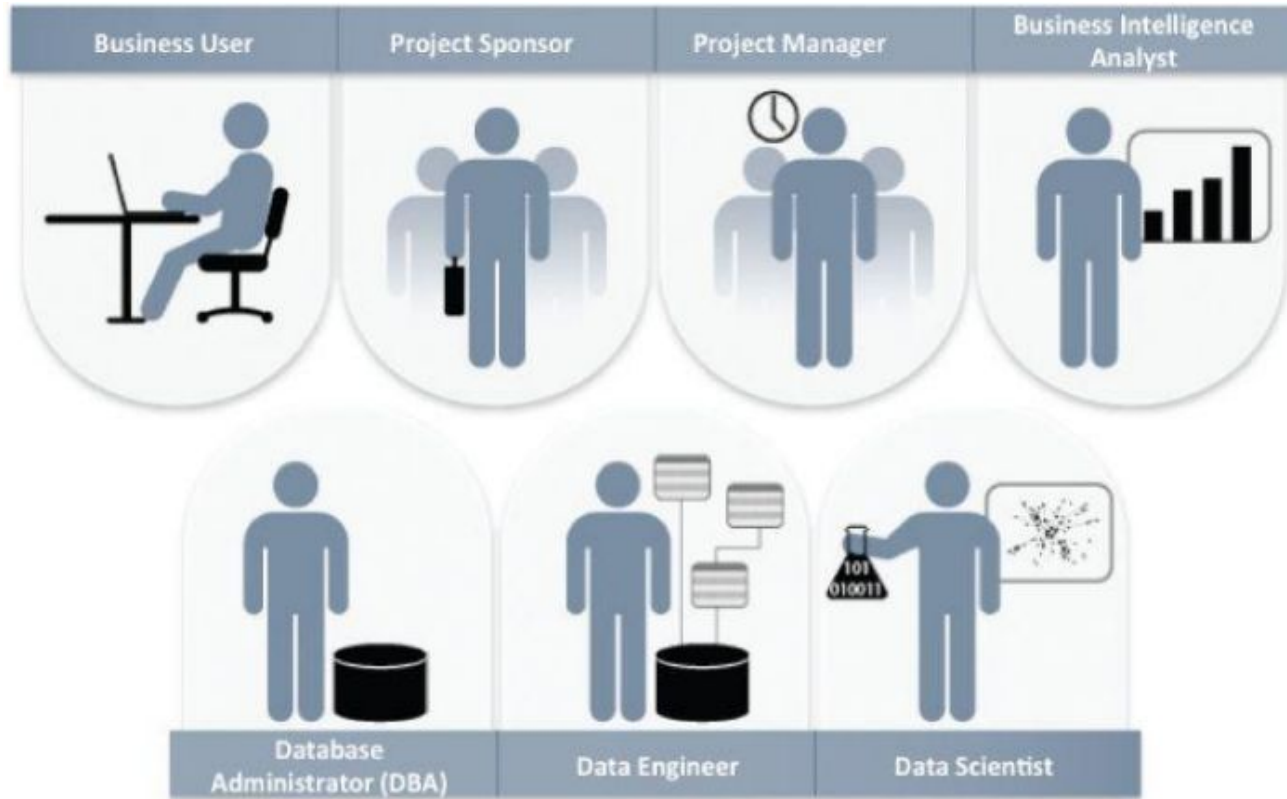


Gene  
Sequencing

# Big Data - Case Studies



# Key Roles for a Successful Analytics Project



**Figure 2.1** Key roles for a successful analytics project

# Key Roles for a Successful Analytics Project

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

# Key Roles for a Successful Analytics Project

## 1. Business User:

- who understands the domain area and usually benefits from the results.
- Who can consult and advise the project team on the context of the project, the value of the results, and how the outputs will be operationalized.
- Usually a business analyst, line manager, or deep subject matter expert in the project domain fulfills this role.

## 2. Project Sponsor:

- Responsible for the genesis of the project.
- Provides the impetus and requirements for the project and defines the core business problem.
- Generally provides the funding and gauges the degree of value from the final outputs of the working team.
- This person sets the priorities for the project and clarifies the desired outputs.

# Key Roles for a Successful Analytics Project

## 3. Project Manager:

- Ensures that key milestones and objectives are met on time and at the expected quality.

## 4. Business Intelligence Analyst:

- Provides business domain expertise based on a deep understanding of the data, key performance indicators (KPIs), key metrics, and business intelligence from a reporting perspective.
- create dashboards and reports and have knowledge of the data feeds and sources

## 5. Database Administrator (DBA):

- Provisions and configures the database environment to support the analytics needs of the working team.
- providing access to key databases or tables and ensuring the appropriate security levels are in place related to the data repositories.

# Key Roles for a Successful Analytics Project

## 6. Data Engineer: \*\* Most popular and in high demand

- Leverages deep technical skills to assist with tuning SQL queries for data management and data extraction,
- provides support for data ingestion into the analytic sandbox,
- The data engineer works closely with the data scientist to help shape data in the right ways for analyses.

Note: DBA : sets up and configures the databases to be used,

DE : executes the actual data extractions and performs substantial data manipulation to facilitate the analytics.

## 7. Data Scientist: \*\* Most popular and in high demand

- Provides subject matter expertise for analytical techniques, data modeling, and applying valid analytical techniques to given business problems.
- Ensures overall analytics objectives are met.
- Designs and executes analytical methods and approaches with the data available to the project.



# Data Analytics Lifecycle drafted based on

## 1. Scientific method (Old Approach)

- provides a solid framework for thinking about and deconstructing problems into their principal parts.
- to forming hypotheses and finding ways to test ideas.

## 2. CRISP-DM ( popular approach for data mining)

- provides useful input on ways to frame analytics problems

## 3. Tom Davenport's DELTA framework

- offers an approach for data analytics projects, including the context of the organization's skills, datasets, and leadership engagement.

## 4. Doug Hubbard's Applied Information Economics (AIE) approach

- provides a framework for measuring intangibles and provides guidance on developing decision models, calibrating expert estimates, and deriving the expected value of information.

## 5. "MAD Skills" by Cohen

- offers input for several of the techniques mentioned in Phases 2–4 that focus on model planning, execution, and key findings.

# Data Analytics Lifecycle - Phase 1 : Discovery

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

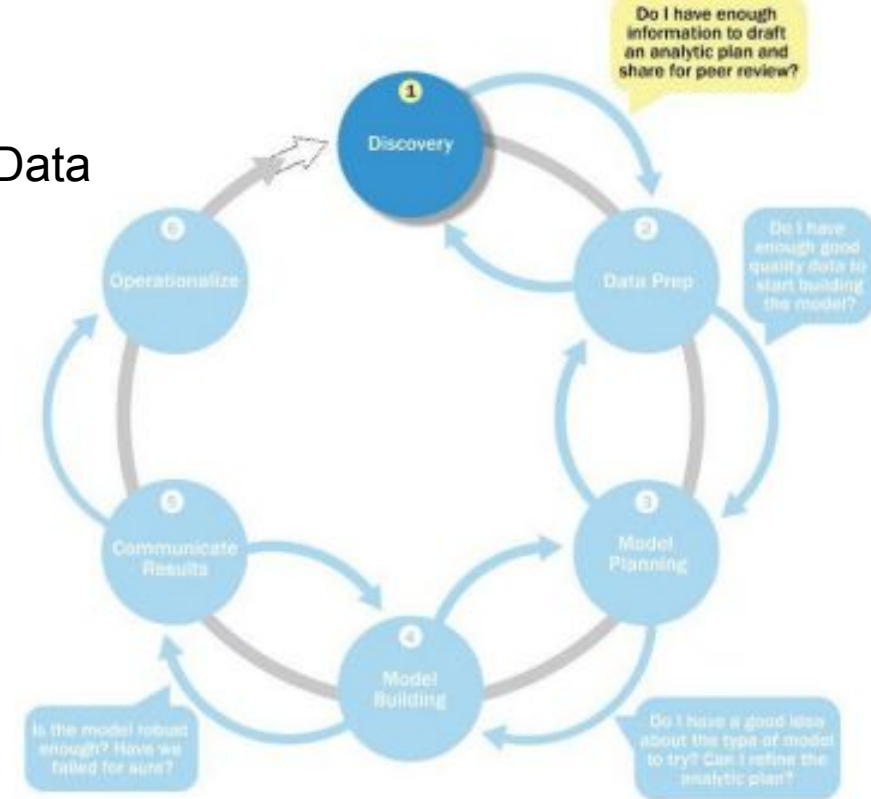


Figure 2.3 Discovery phase

# Data Analytics Lifecycle - Phase 2 : Data Preparation

1. Preparing the Analytic Sandbox
2. Performing ETLT
3. Learning about the Data
4. Data Conditioning
5. Survey and Visualize
6. Common Tools for Data Preparation

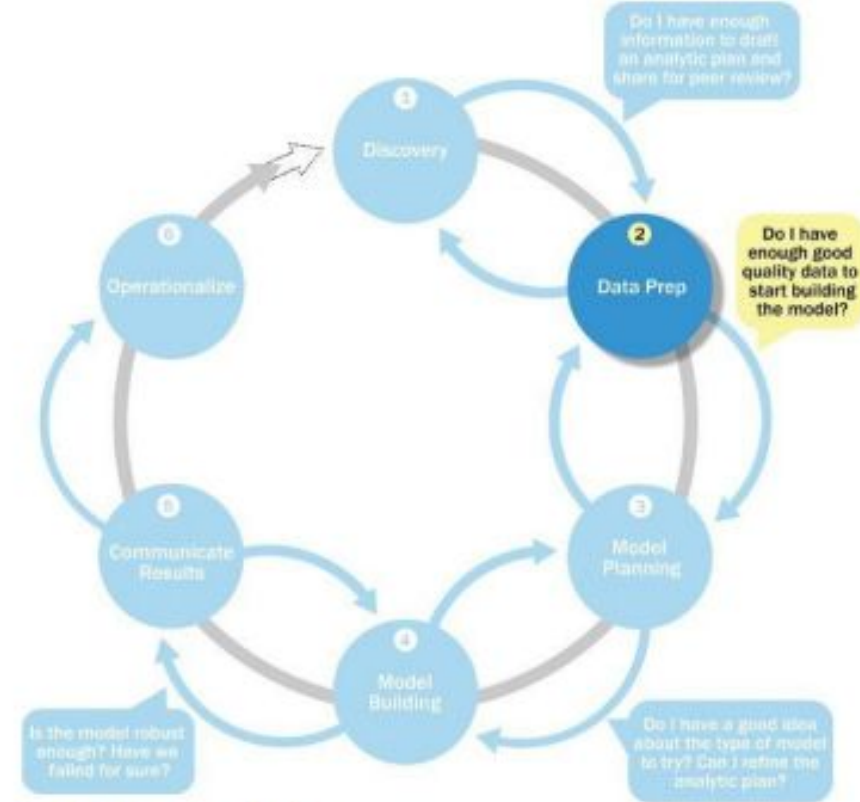


Figure 2.4 Data preparation phase

# Data Analytics Lifecycle - Phase 2 : Data Preparation

## 1. Preparing the Analytic Sandbox

- Create the analytic sandbox (also called workspace)
- To explore data without interfering with live production data
- Collects all kinds of data
- Allows organizations to undertake ambitious projects beyond traditional data analysis and BI to perform advanced predictive analytics

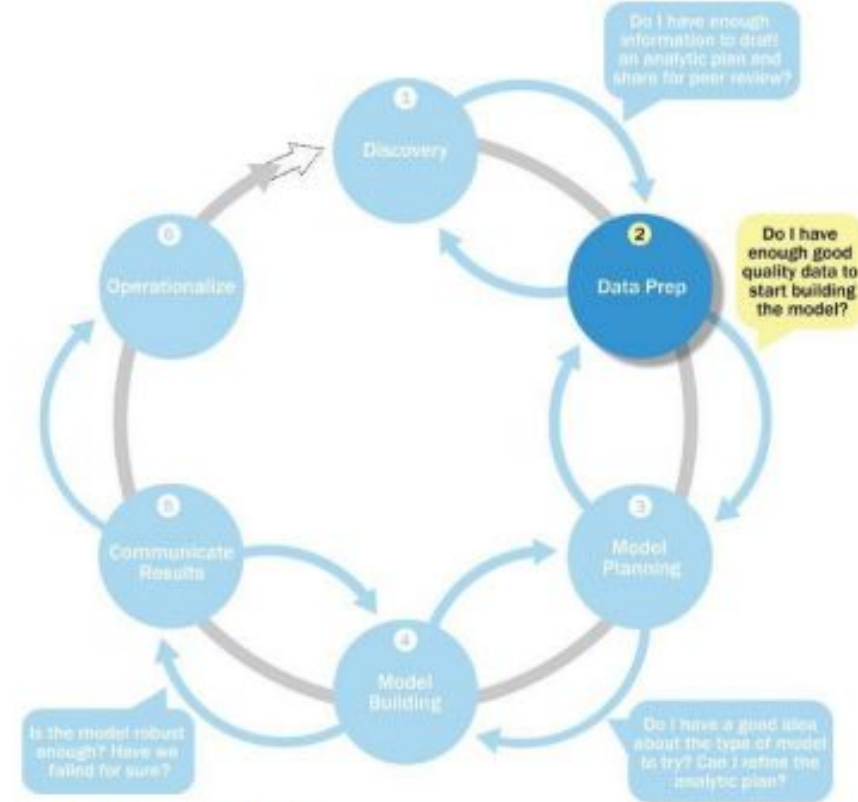


Figure 2.4 Data preparation phase

## 2. Performing ETLT

- 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
- Ex: Credit card fraud detection,
  - **outliers** can represent high-risk transactions that might be inadvertently filtered out or transformed before being loaded into the database

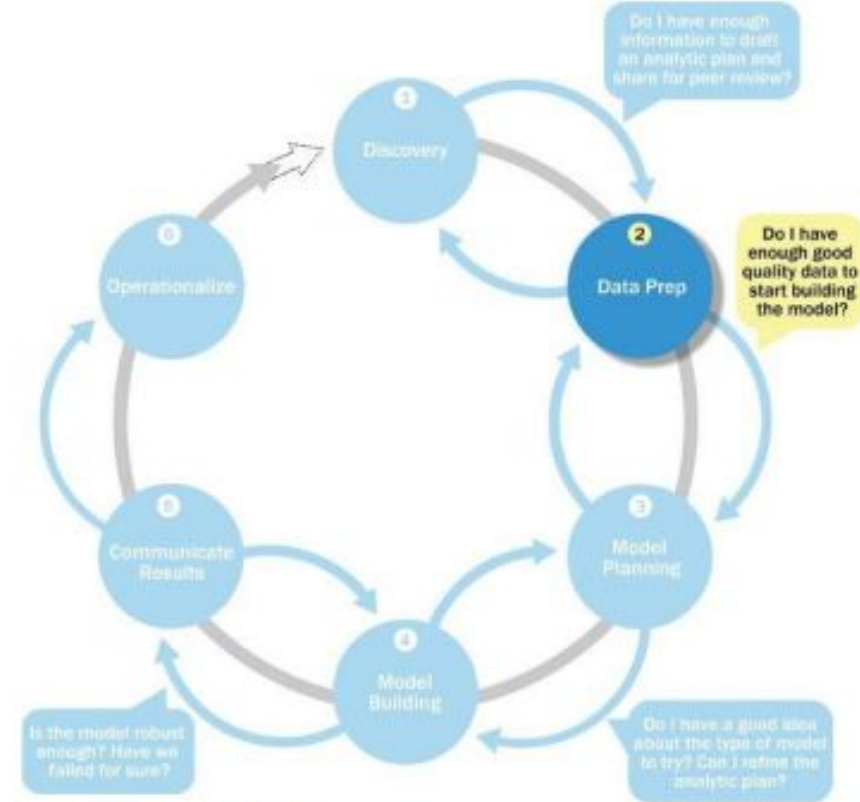
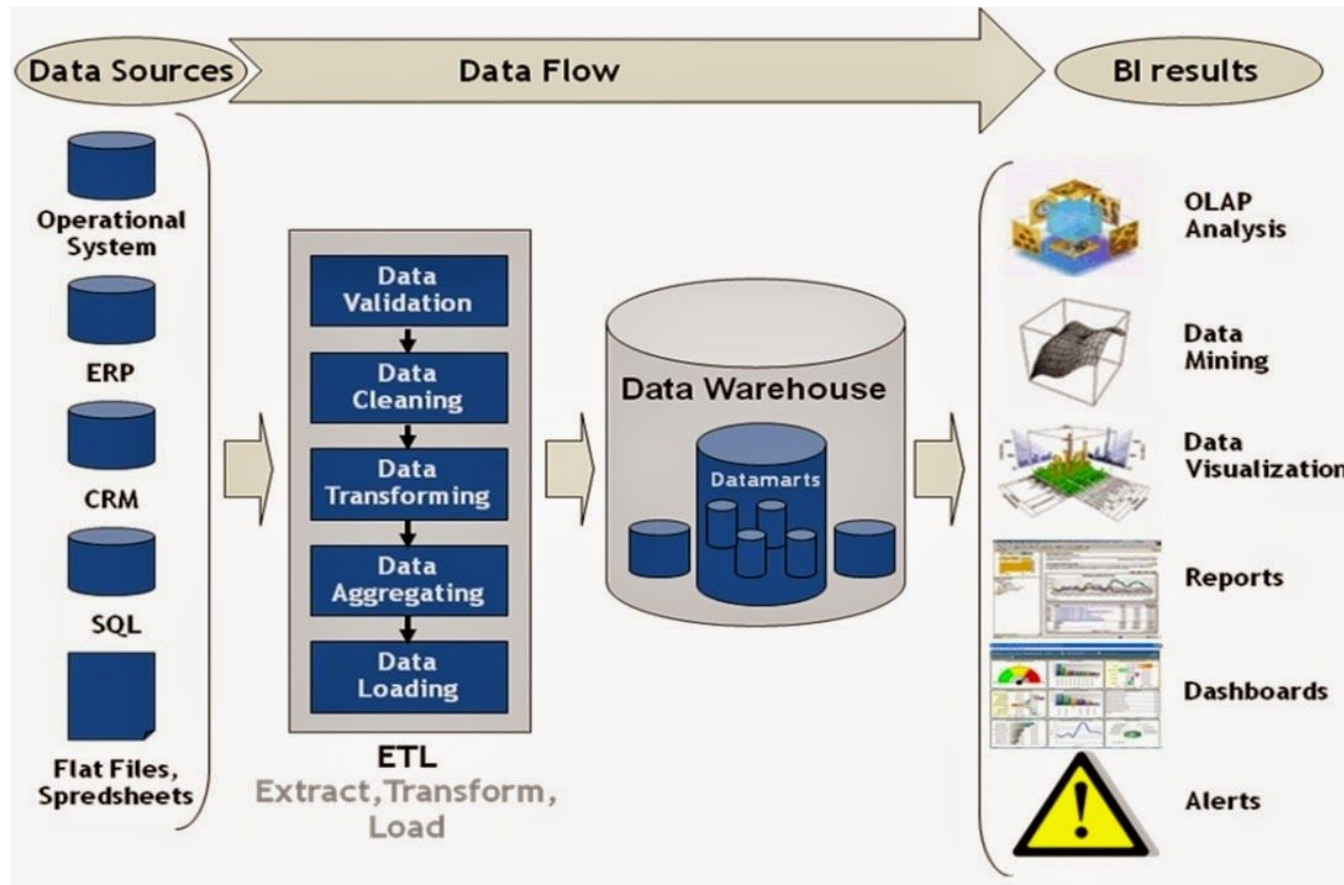


Figure 2.4 Data preparation phase

# Data Analytics Lifecycle - Phase 2 : Data Preparation



## 3. Learning about the Data

- 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

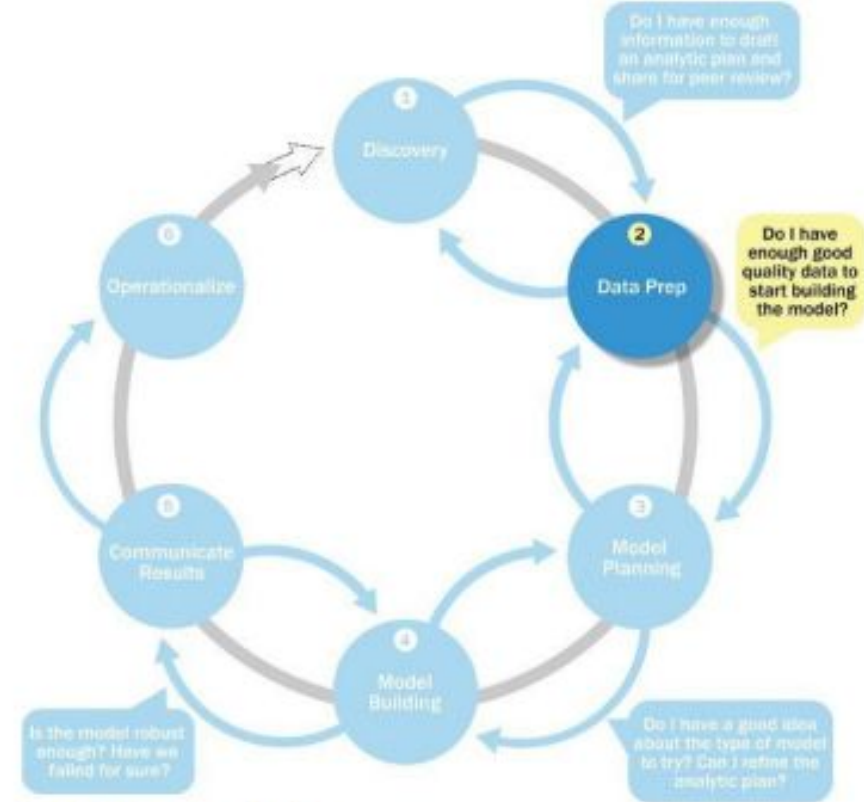


Figure 2.4 Data preparation phase



# Data Analytics Lifecycle - Phase 2 : Data Preparation

**Table 2.1** 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				•

## 4. Data Conditioning

- Process of cleaning data, normalizing datasets, and performing transformations on the data.
- Managing Missing data, Outliers, and Unwanted Data
- Involve many complex steps to join or merge datasets or otherwise get datasets into a state that enables analysis in further phases.
- Preprocessing step for the data analysis
- performed only by IT, the data owners, a DBA, or a data engineer & Data Scientist

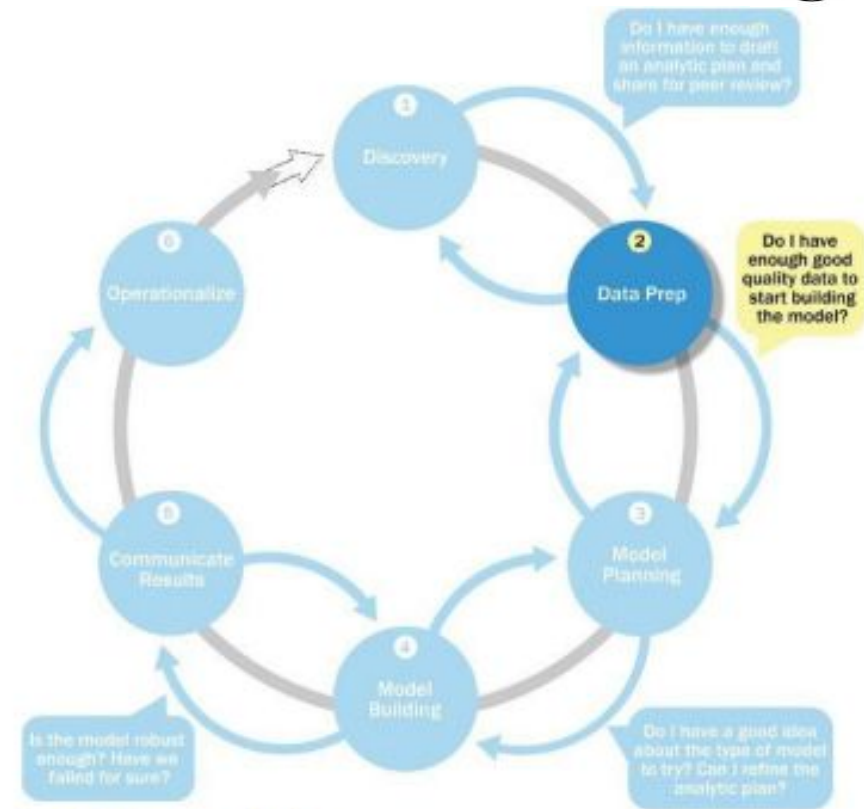


Figure 2.4 Data preparation phase

# Data Analytics Lifecycle - Phase 2 : Data Preparation

## Sample Questions to be considered while Data Conditioning

1. What are the data sources?
2. What are the target fields (for example, columns of the tables)?
3. How clean is the data?
4. How consistent are the contents and files?
5. Determine to what degree the data contains missing or inconsistent values and if the data contains values deviating from normal.
6. Assess the consistency of the data types.
7. Review the content of data columns or other inputs, and check to ensure they make sense.
8. Look for any evidence of systematic error.

# Data Analytics Lifecycle - Phase 2 : Data Preparation

## 5. Survey and Visualize

- Leverage data visualization tools to gain an overview of the data
- 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

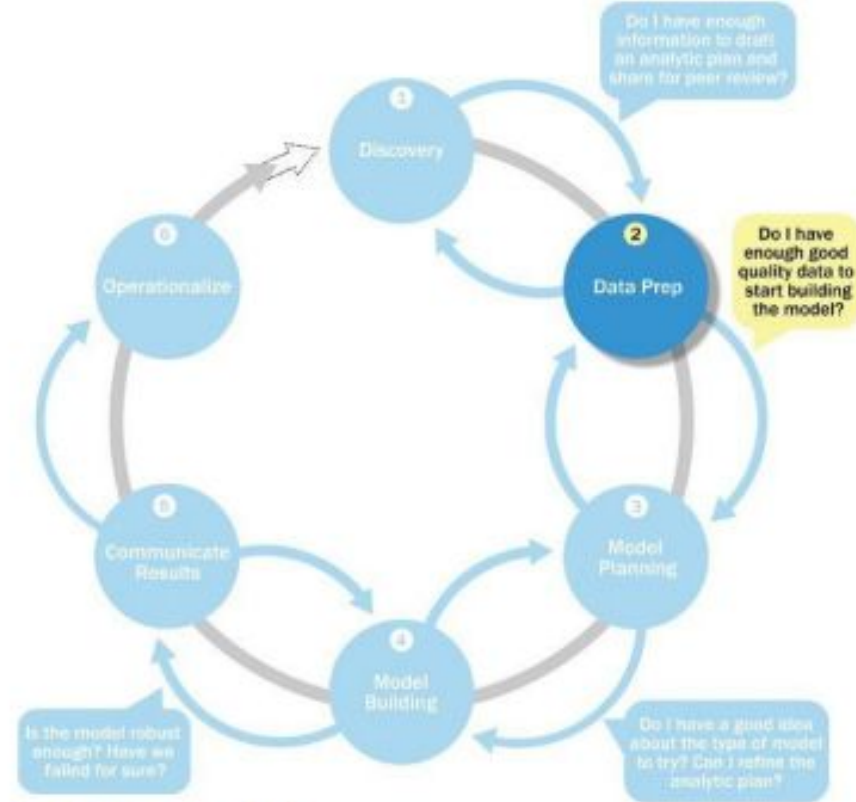


Figure 2.4 Data preparation phase

# Data Analytics Lifecycle - Phase 2 : Data Preparation

## Recommended Guidelines and considerations for Survey and Visualization

- Review data to ensure that calculations remained consistent within columns or across tables for a given data field.
- Does the data distribution stay consistent over all the data? If not, what kinds of actions should be taken to address this problem?
- 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?
- For time-related variables, are the measurements daily, weekly, monthly?
- Is the data standardized/normalized? Are the scales consistent? If not, how consistent or irregular is the data?
- For geospatial datasets, are state or country abbreviations consistent across the data? Are personal names normalized? English units? Metric units?

## 6. Common Tools for Data Preparation

- Hadoop
  - Perform parallel ingest and analysis for web traffic parsing
- Alpine Miner
  - provides a GUI for creating analytic workflows
- Open Refine (formerly Google Refine)
  - free, open source tool for working with messy data
- Data Wrangler (Stanford University)
  - Tool for data cleansing & transformation

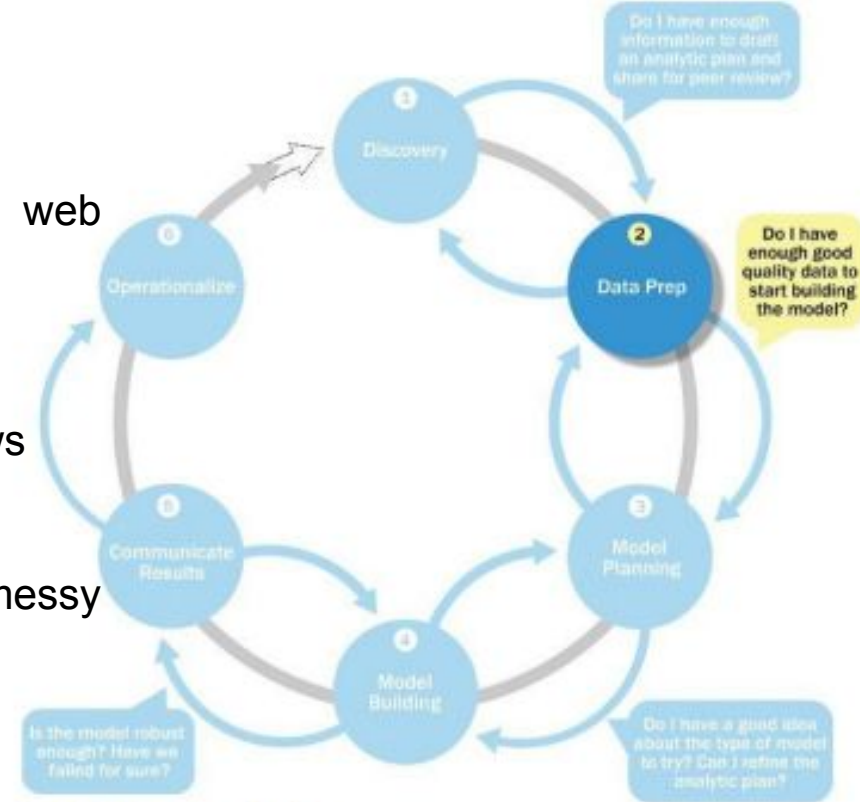


Figure 2.4 Data preparation phase

# Data Analytics Lifecycle - Phase 3 : Model Planning

1. Data Exploration & Variable Selection
2. Model Selection
3. Common Tools for Model Planning Phase

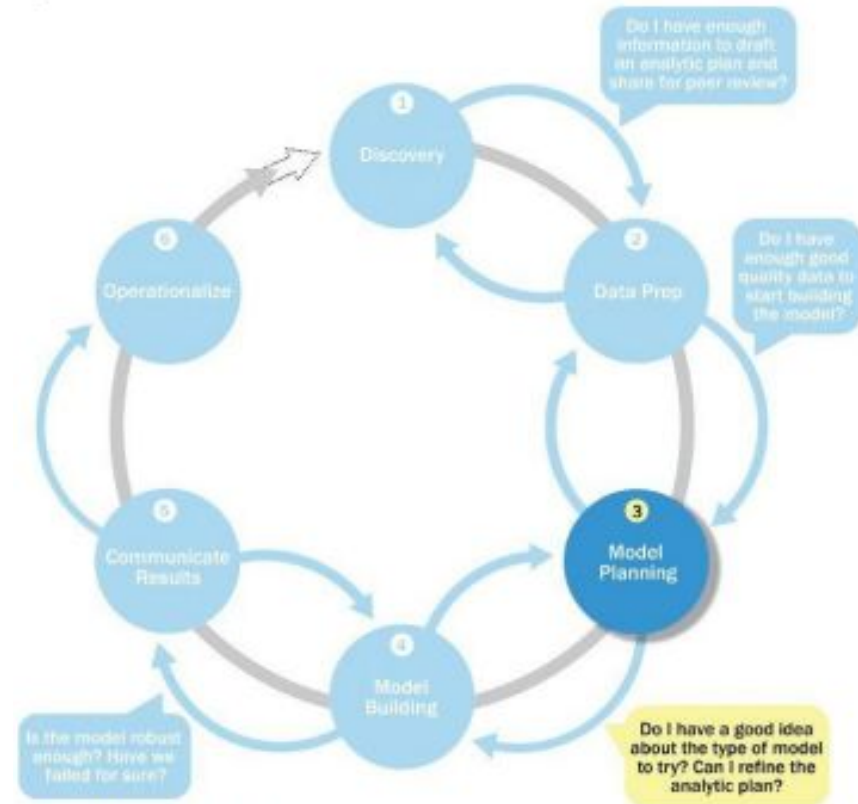


Figure 2.5 Model planning phase



## Activities to consider in Model Planning Phase

- 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

**Table 2.2** Research on Model Planning in Industry Verticals

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

# Data Analytics Lifecycle - Phase 3 : Model Planning

## 1. Data Exploration and Variable Selection

- Explore the data to understand the relationships among the variables to select key variables and the most suitable models
- A common way to do this is to use data visualization tools
- If the team plans to run regression analysis, identify the candidate predictors and outcome variables of the model

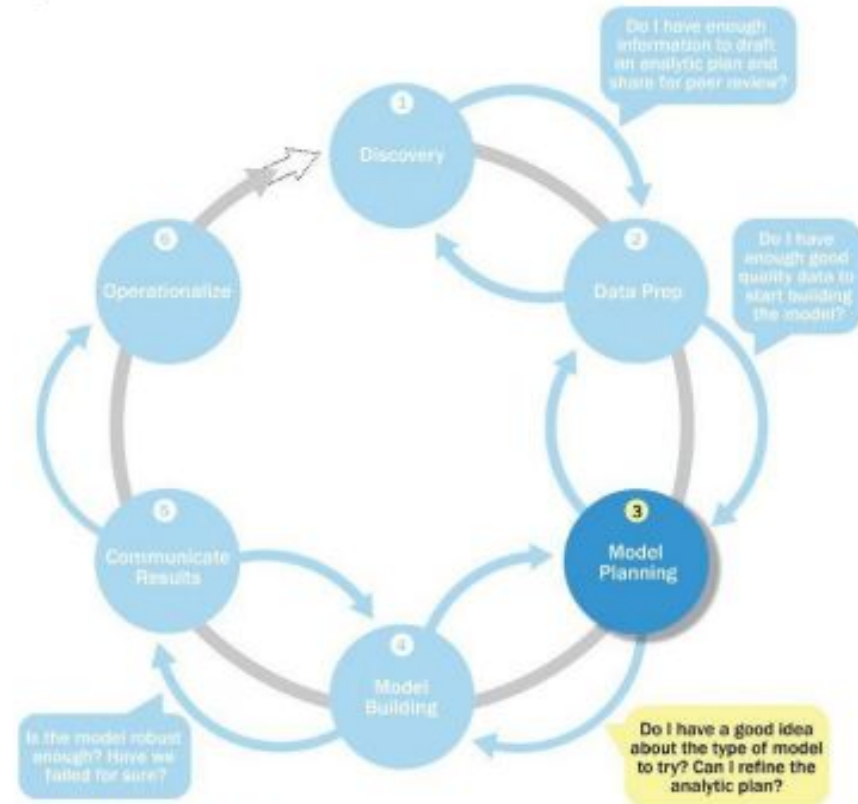


Figure 2.5 Model planning phase

# Data Analytics Lifecycle - Phase 3 : Model Planning

## 2. Model Selection

- 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
- The team moves to the model building phase once it has a good idea about the type of model to try

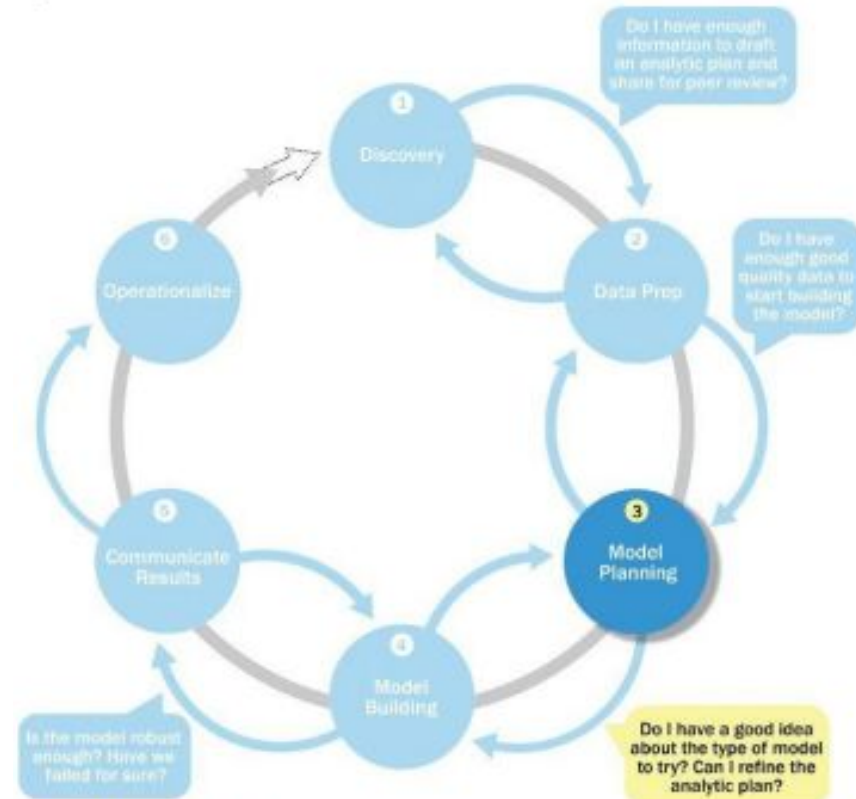


Figure 2.5 Model planning phase

# Data Analytics Lifecycle - Phase 3 : Model Planning

## 3. Common Tools for the Model Planning Phase

- **R** - contains about 5000 packages for data analysis and graphical presentation
- **SQL Analysis services**
  - Can perform in-database analytics of common data mining functions, involved aggregations, and basic predictive models
- **SAS/ ACCESS**
  - Provides integration between SAS and the analytics sandbox

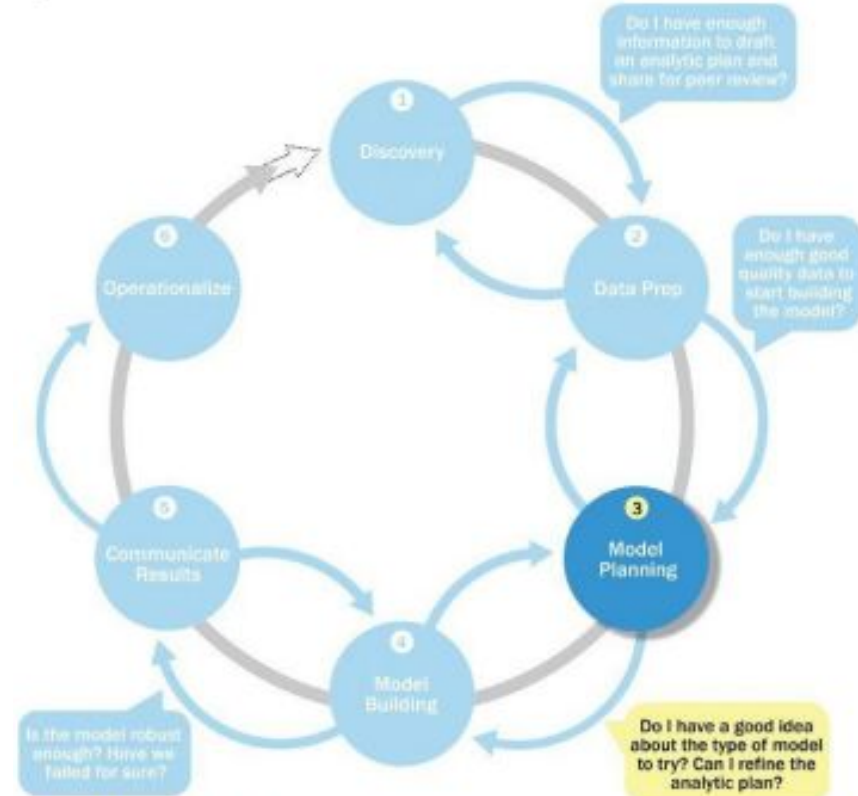
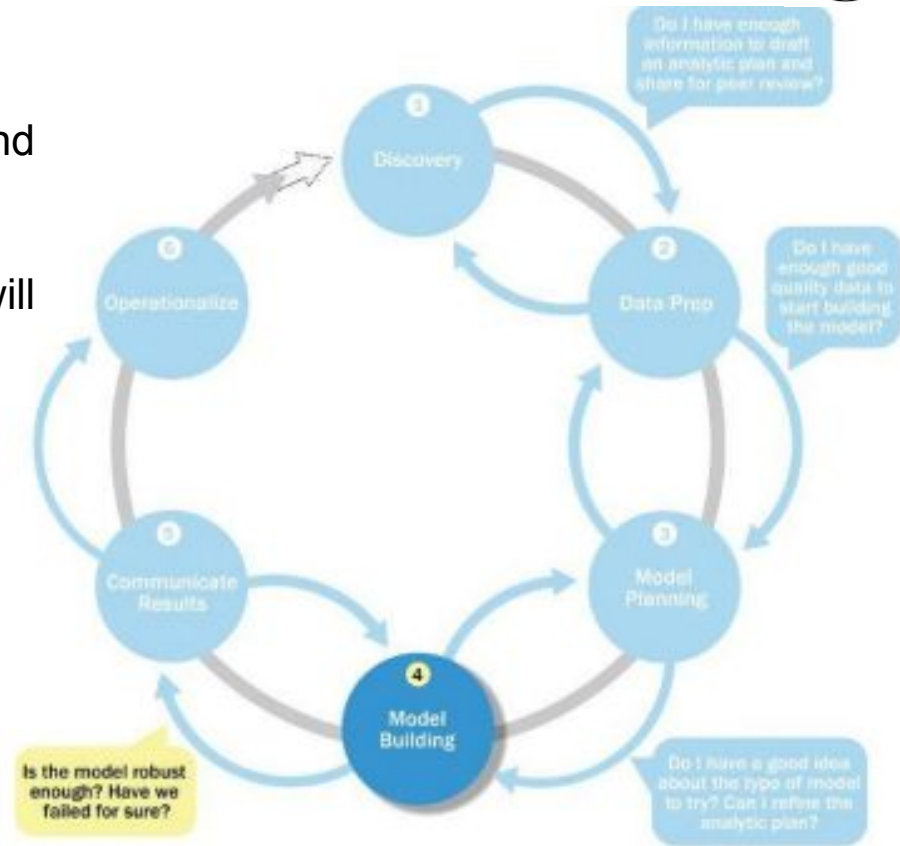


Figure 2.5 Model planning phase

# Data Analytics Lifecycle - Phase 4 : Model Building

- Execute the models defined in Phase 3
- Develop datasets for training, testing, and production
- Team also considers whether its existing tools will suffice for running the models or if they need more robust environment for executing models.
- Free or open-source tools
  - R and PL/R, Octave, WEKA, Python
- Commercial tools
  - Matlab , STASTICA, SPSS Modeler

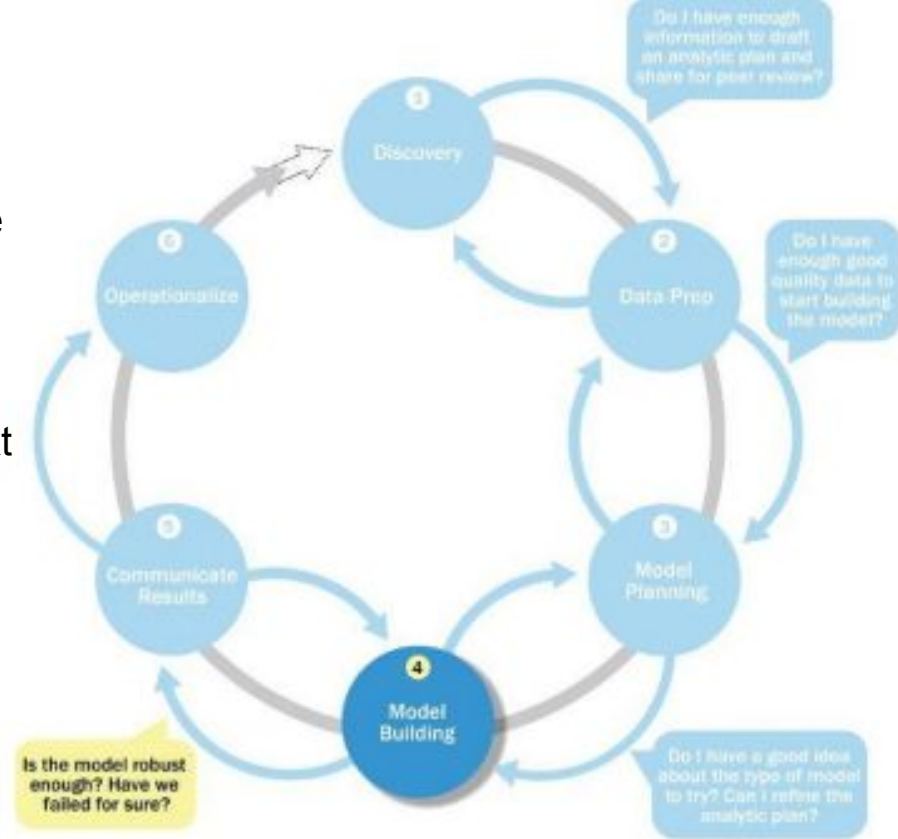




# Data Analytics Lifecycle - Phase 4 : Model Building

## Question to consider in Model Building Phase

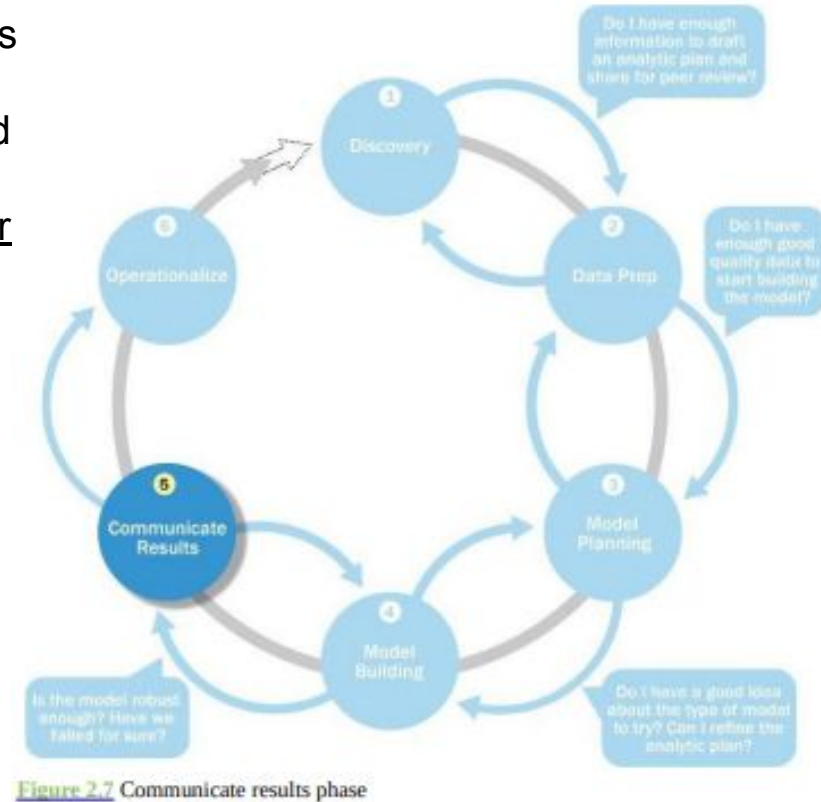
- Does the model appear valid and accurate on the test data?
- Does the model output/behavior make sense to the domain experts?
- Is the model sufficiently accurate to meet the goal?  
Does the model avoid intolerable mistakes?
- Do the parameter values make sense in the context of the domain?
- Are more data or inputs needed?
- Will the kind of model chosen support the runtime environment?
- Is a different form of the model required to address the business problem?





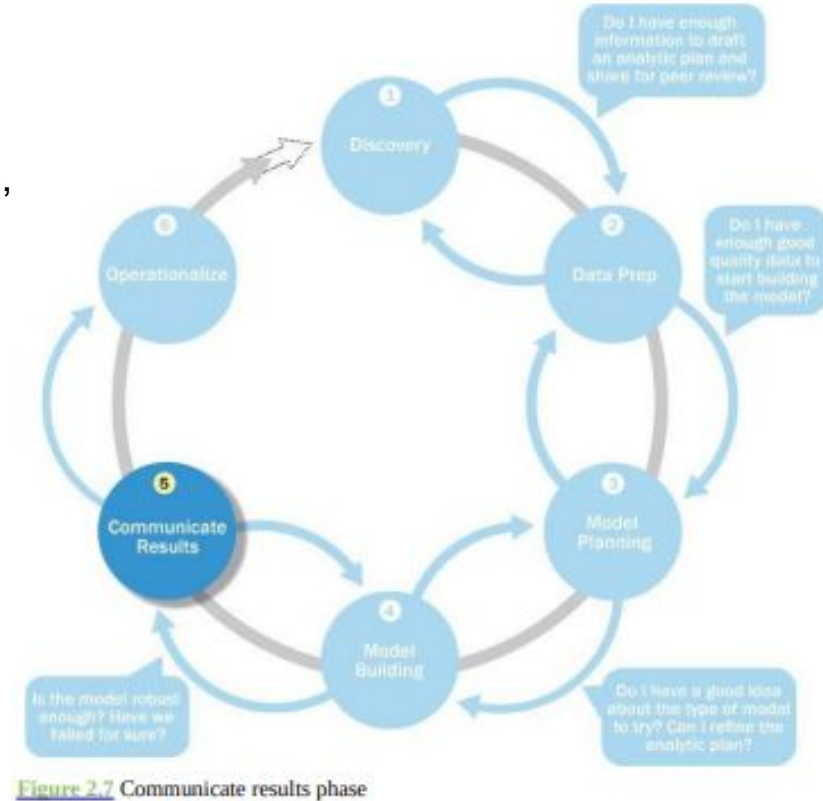
# Data Analytics Lifecycle - Phase 5 : Communicate Results

- Determine Success / Failure of the team in its objectives
- Assess if the results are statistically significant and valid
- Communicate and document the key findings and major insights derived from the analysis



# Data Analytics Lifecycle - Phase 6 : Operationalize

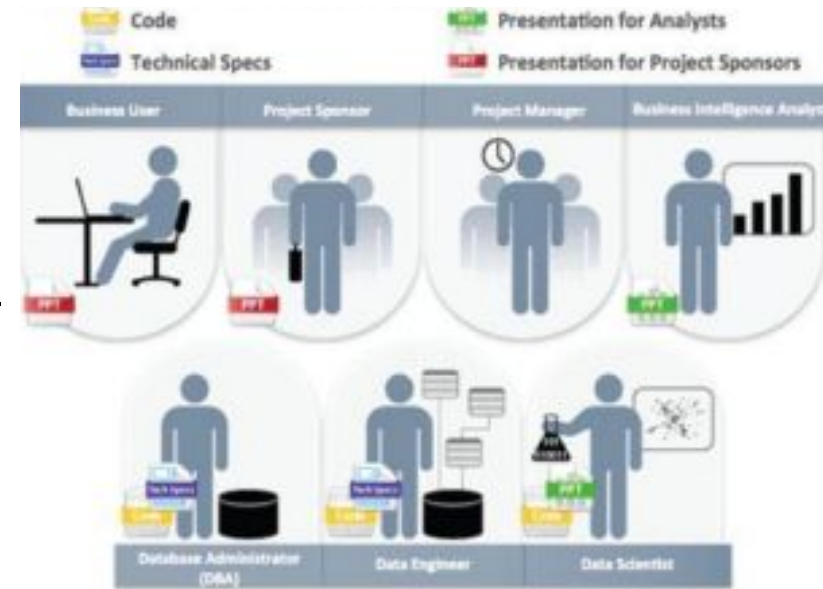
- The team sets up a pilot project to deploy the work in a controlled way
- Risk is managed effectively by undertaking small scope, pilot deployment before a wide-scale rollout
- During the pilot project, team executes the algorithm more efficiently in the database
- To test the model in a live setting, consider running the model in a production environment for a discrete set of products or a single line of business
- Monitor model accuracy and retrain the model if necessary



# Data Analytics Lifecycle - Phase 6 : Operationalize

## Key outputs from successful analytics project

1. **Business user** : tries to determine business benefits and implications
2. **Project sponsor** : wants business impact, risks, ROI
3. **Project manager** : needs to determine if project completed on time, within budget, goals met
4. **Business intelligence analyst** : needs to know if reports and dashboards will be impacted and need to change
5. **Data engineer and DBA** : must share code and document
6. **Data scientist** : must share code and explain model to peers, managers, stakeholders



## Four main deliverables in Phase 6

1. Presentation for project sponsors – high-level takeaways for executive level stakeholders
2. Presentation for analysts – describes business process changes and reporting changes, includes details and technical graphs
3. Code for technical people
4. Technical specifications of implementing the code

# Data Analytics Lifecycle - Case Study : GINA

## Case Study: Global Innovation Network and Analysis (GINA)

- In 2012 EMC's new director wanted to improve the company's engagement of employees across the global centers of excellence (GCE) to drive innovation, research, and university partnerships
- This project was created to accomplish
  - Store formal and informal data
  - Track research from global technologists
  - Mine the data for patterns and insights to improve the team's operations and strategy

# Data Analytics Lifecycle - Case Study : GINA

## 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
- 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

## The 10 main IHs that the GINA team developed were as follows:

- IH1: Innovation activity in different geographic regions can be mapped to corporate strategic directions.
- IH2: The length of time it takes to deliver ideas decreases when global knowledge transfer occurs as part of the idea delivery process.
- IH3: Innovators who participate in global knowledge transfer deliver ideas more quickly than those who do not.
- IH4: An idea submission can be analyzed and evaluated for the likelihood of receiving funding.
- IH5: Knowledge discovery and growth for a particular topic can be measured and compared across geographic regions.
- IH6: Knowledge transfer activity can identify research-specific boundary spanners in disparate regions.
- IH7: Strategic corporate themes can be mapped to geographic regions.
- IH8: Frequent knowledge expansion and transfer events reduce the time it takes to generate a corporate asset from an idea.
- IH9: Lineage maps can reveal when knowledge expansion and transfer did not (or has not) resulted in a corporate asset.
- IH10: Emerging research topics can be classified and mapped to specific ideators



## Phase 2: Data Preparation

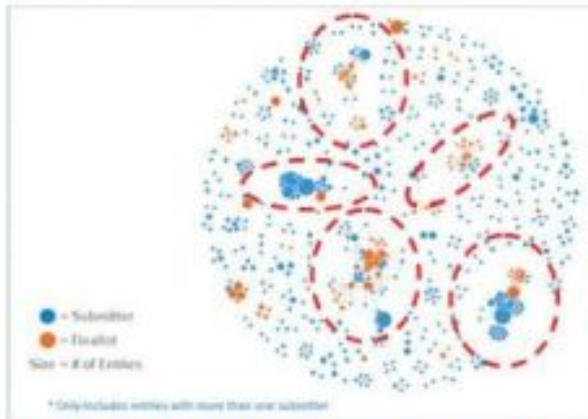
- Set up an analytics sandbox
- Discovered that certain data needed conditioning and normalization and that missing datasets were critical
- Team recognized that poor quality data could impact subsequent steps
- They discovered many names were misspelled and problems with extra spaces
- These seemingly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

## Phase 3: Model Planning

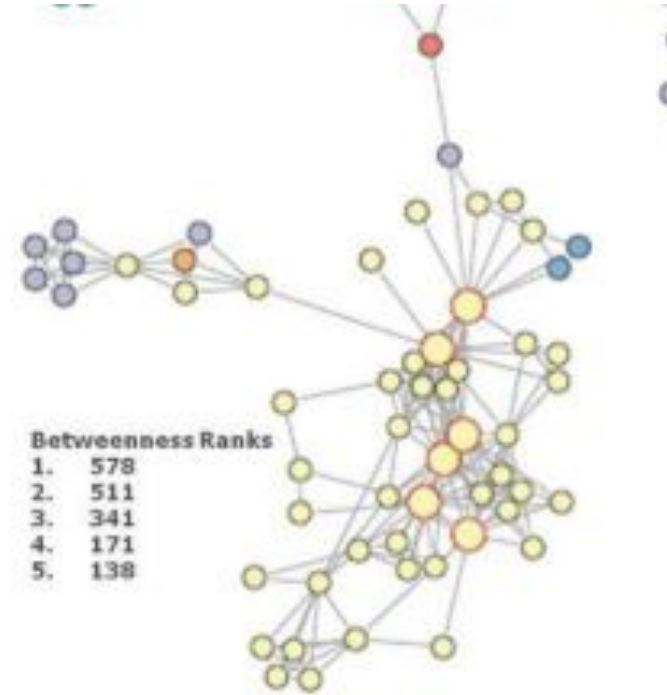
- The study included the following considerations
  - Identify the right milestones to achieve the goals
  - Trace how people move ideas from each milestone toward the goal
  - Tract ideas that die and others that reach the goal
  - Compare times and outcomes using a few different methods

## 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



Social graph [27] visualization of idea submitters and finalists



11 Social graph visualization of top innovation influencers

## 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

## 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

# Data Analytics Lifecycle - Case Study : GINA

**Table 2.3** Analytic Plan from the EMC GINA Project

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	<ol style="list-style-type: none"> <li>1. Identified hidden, high-value innovators and found ways to share their knowledge</li> <li>2. Informed investment decisions in university research projects</li> </ol>
Key Findings	<ol style="list-style-type: none"> <li>3. Created tools to help submitters improve ideas with idea recommender systems</li> </ol>

## 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