

Data Science & Big Data Analytics

Subject Code: 310251

T. E. Computer (2019 Pattern)

UNIT III

Unit III	Big Data Analytics Life Cycle	07 Hours
Introduction to Big Data, sources of Big Data, Data Analytic Lifecycle : Introduction, Phase 1: Discovery, Phase 2: Data Preparation, Phase 3: Model Planning, Phase 4: Model Building, Phase 5: Communication results, Phase 6: Operation alize.		
#Exemplar/Case Studies	Case study: Global Innovation Social Network and Analysis (GINA).	
*Mapping of Course Outcomes for Unit III	CO3	



Introduction to Big Data

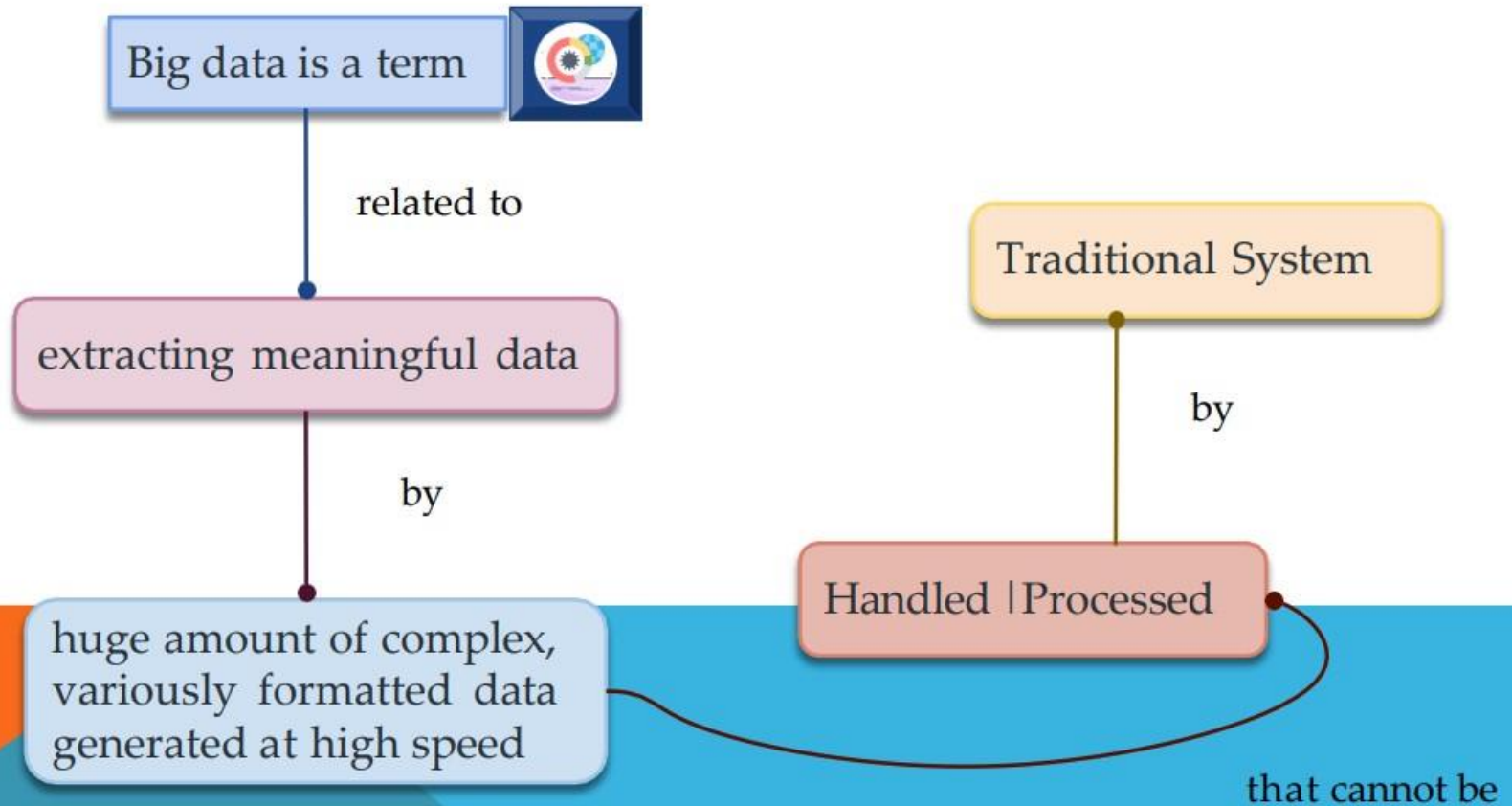
No single definition; here is from Wikipedia:

- **Big data** is the term for a collection of data sets, which are large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.
- The challenges include **capture, creation, storage, search, sharing, transfer, analysis, and visualization.**
- The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data.

Big Data Example

- Credit card companies monitor every purchase their customers make and can identify fraudulent purchases with a high degree of accuracy using rules derived by processing billions of transactions.
- Mobile phone companies analyze subscribers' calling patterns to determine, for example, whether a caller's frequent contacts are on a rival network. If that rival network is offering an attractive promotion that might cause the subscriber to defect, the mobile phone company can proactively offer the subscriber an incentive to remain in her contract.
- For companies such as Linked In and Facebook, data itself is their primary product. The valuations of these companies are heavily derived from the data they gather and host, which contains more and more intrinsic value as the data grows.

Big Data



Sources of Big Data

The data now comes from multiple sources, such as these:

- Medical information, such as genomic sequencing and diagnostic imaging
- Photos and video footage uploaded to the World Wide Web
- Video surveillance, such as the thousands of video cameras spread across a city
- Mobile devices, which provide geospatial location data of the users, as well as metadata about text messages, phone calls, and application usage on smart phones
- Smart devices, which provide sensor-based collection of information from smart electric grids, smart buildings, and many other public and industry infrastructures
- Non-traditional IT devices, including the use of radio-frequency identification (RFID) readers, GPS navigation systems, and seismic processing

Sources of Big Data

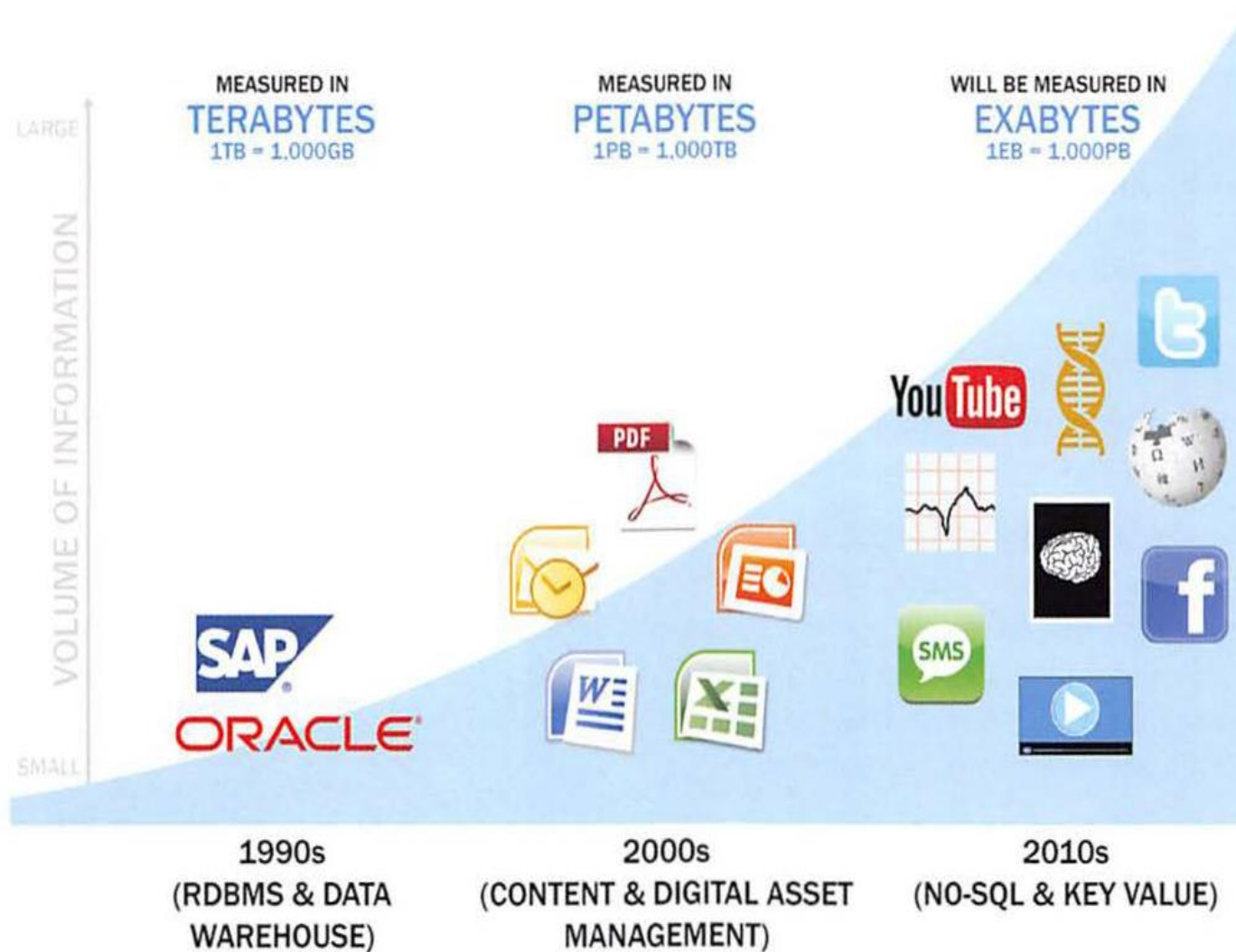
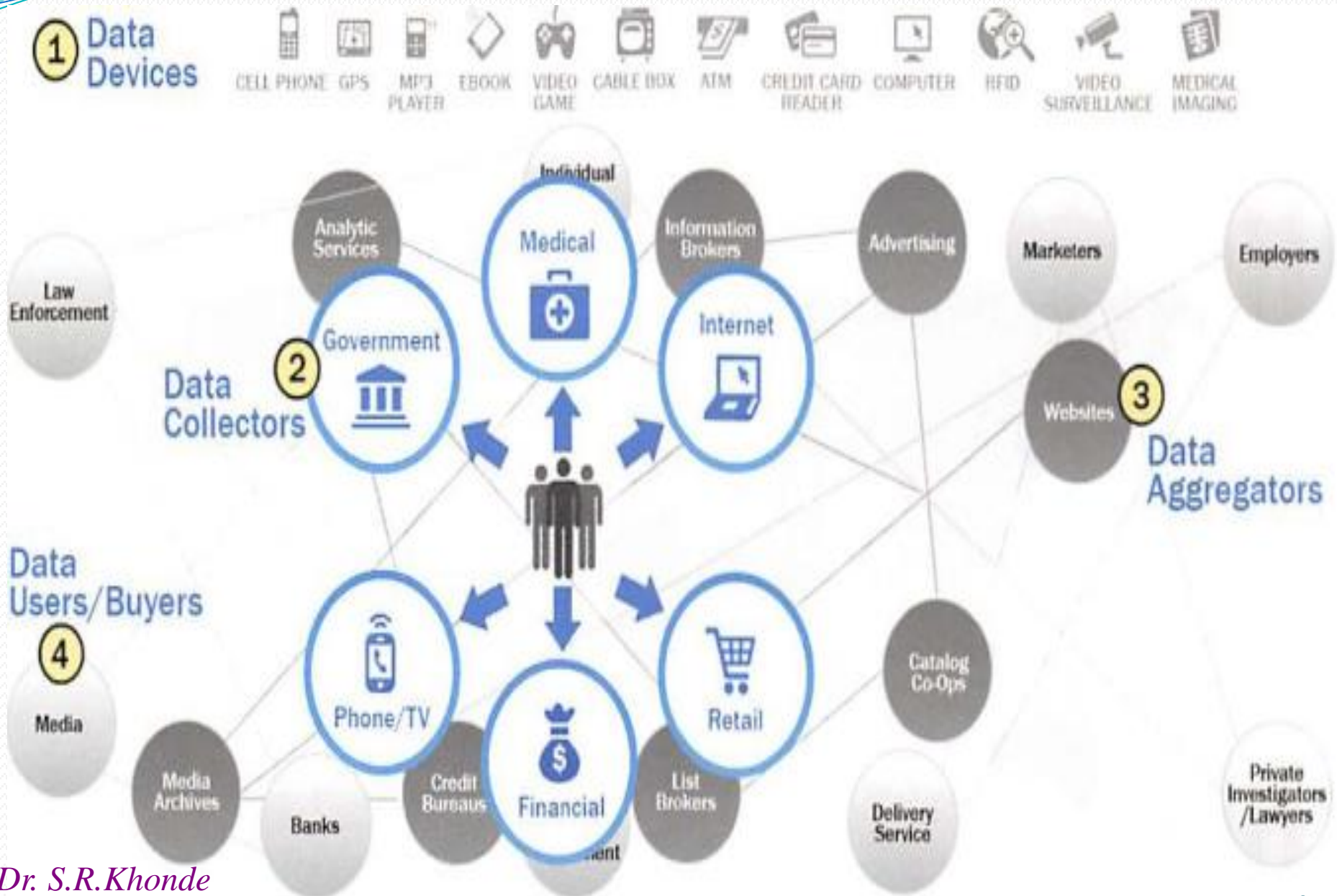


FIGURE 1-10 Data evolution and the rise of Big Data sources

Sources of Big Data



Big Data Generators



Data Analytics Lifecycle

Data Analytics Lifecycle Overview

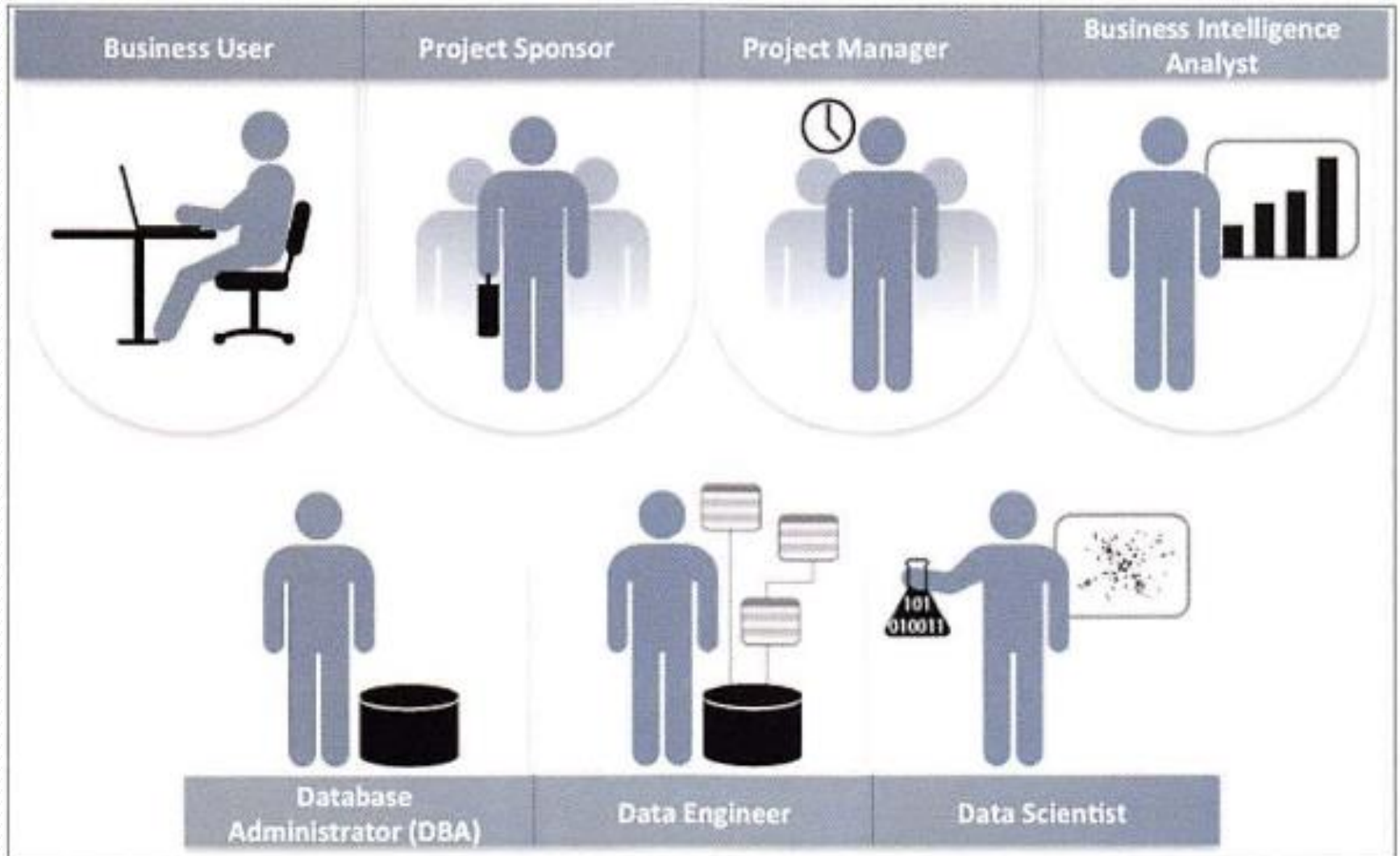
- Phase 1: Discovery
- Phase 2: Data Preparation
- Phase 3: Model Planning
- Phase 4: Model Building
- Phase 5: Communicate Results
- Phase 6: Operationalize

Case Study: GINA

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

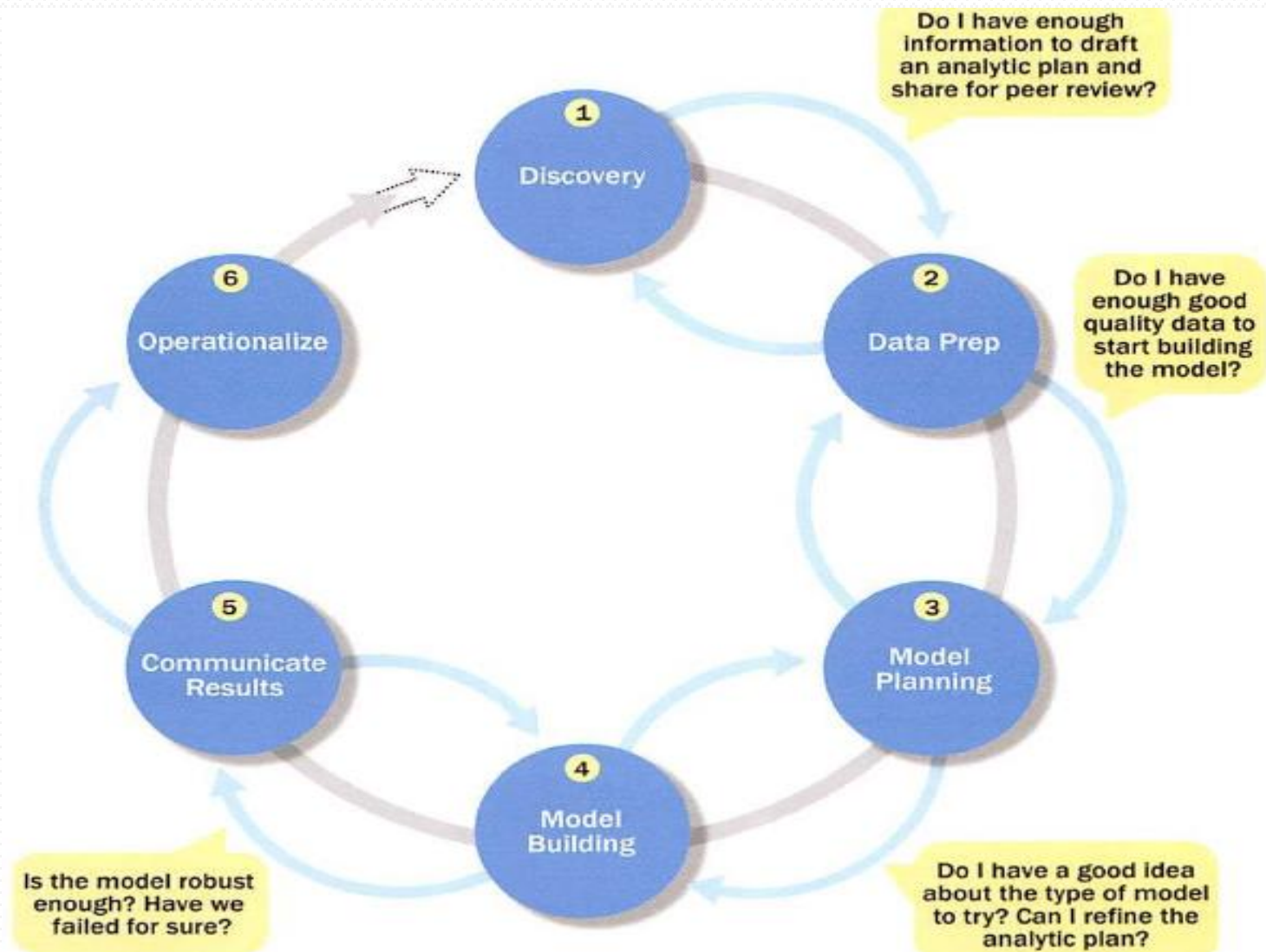
Key Roles for a Successful Analytics Project



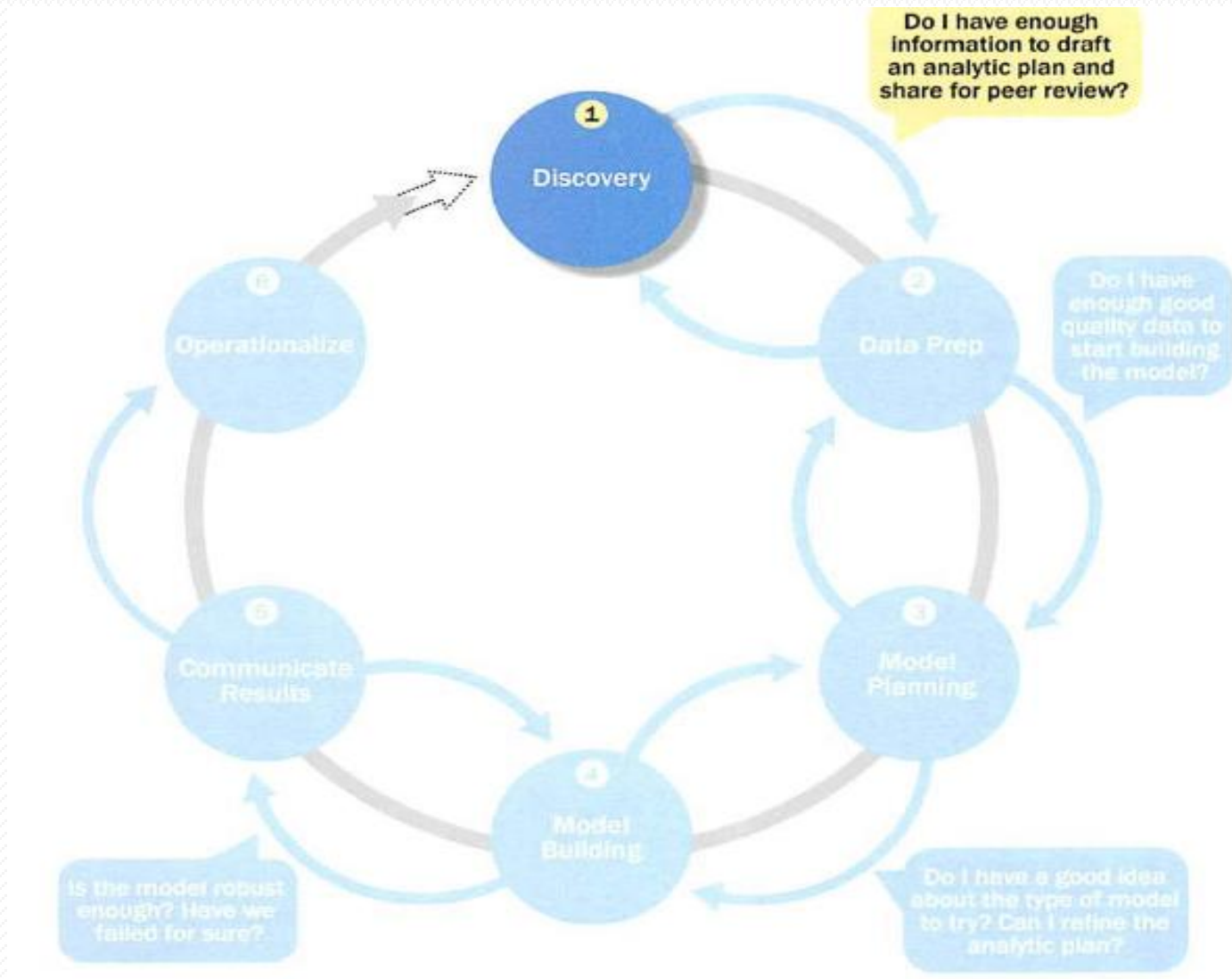
Key Roles for a Successful Analytics Project

- 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 modelling

Overview of Data Analytics Lifecycle



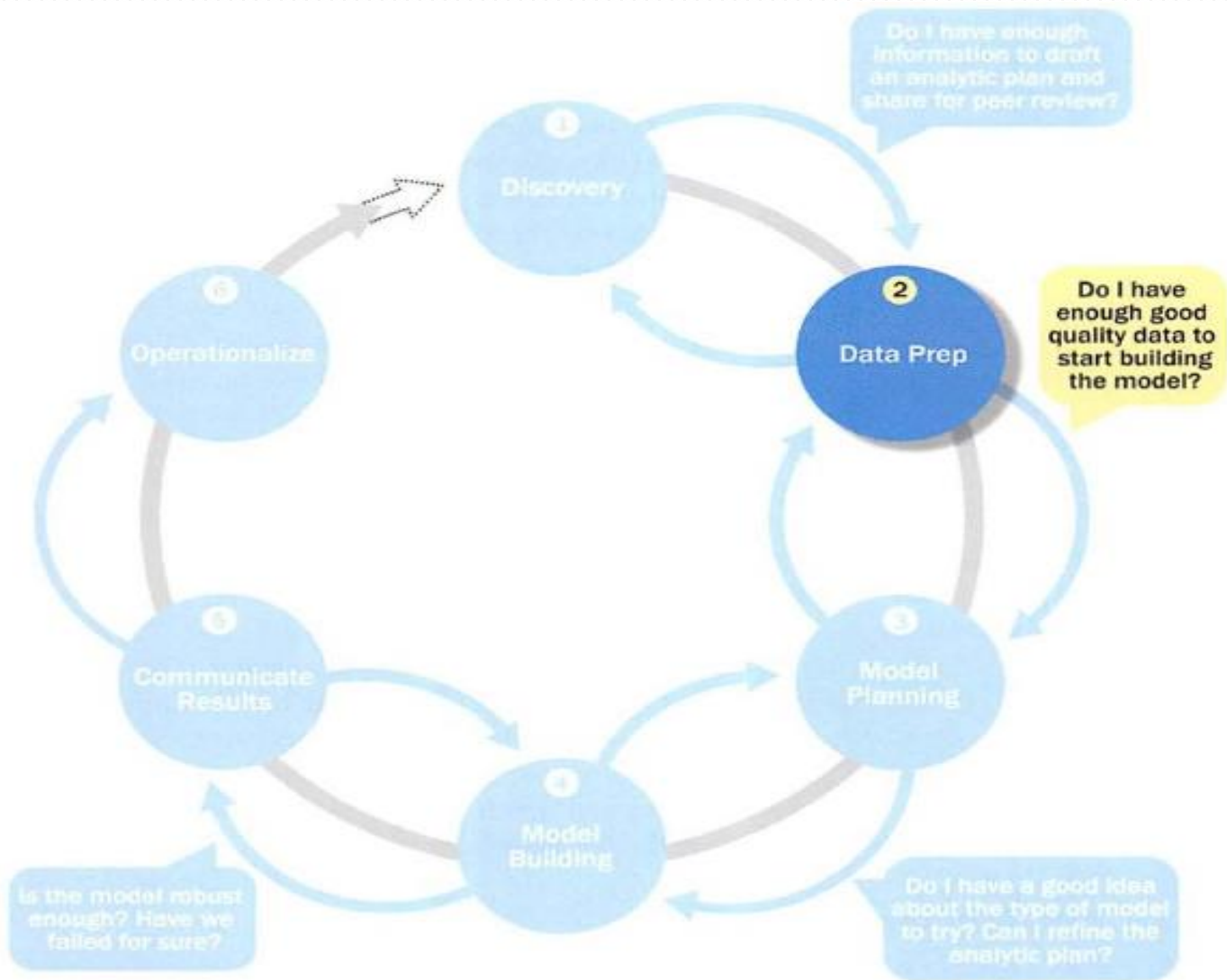
Phase 1: Discovery



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

Phase 2: Data Preparation



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

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

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

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

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				●

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

Data Conditioning

- Additional questions and considerations
- What are the data sources? Target fields?
- How clean is the data?
- How consistent are the contents and files? Missing or inconsistent values?
- Assess the consistence of the data types – numeric, alphanumeric?
- Review the contents to ensure the data makes sense
- Look for evidence of systematic error

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

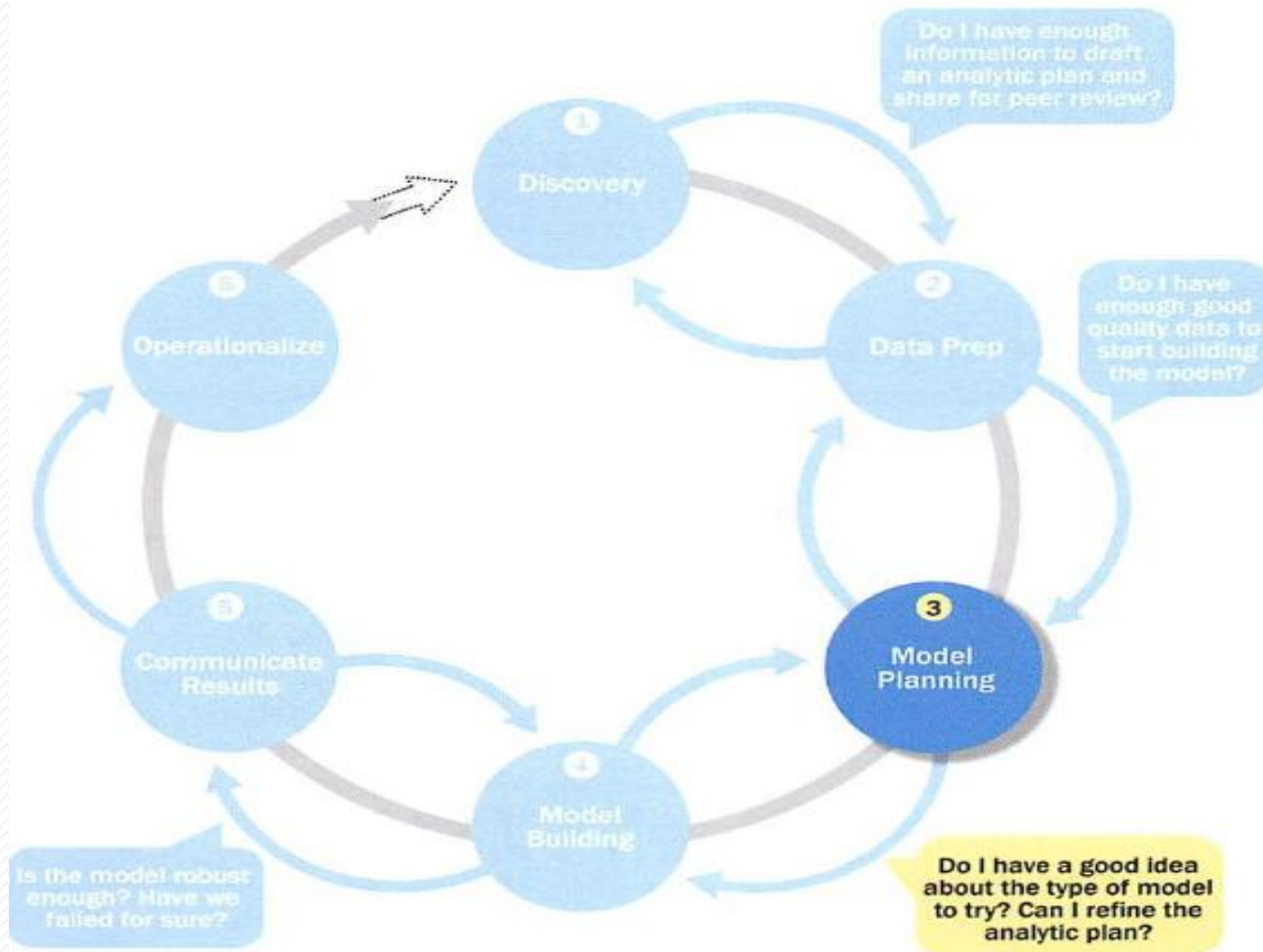
Survey and Visualize Guidelines and Considerations

- 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

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

Phase 3: Model Planning



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

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

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
- Often, stakeholders and subject matter experts may have ideas
For example, some hypothesis that led to the project
- 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

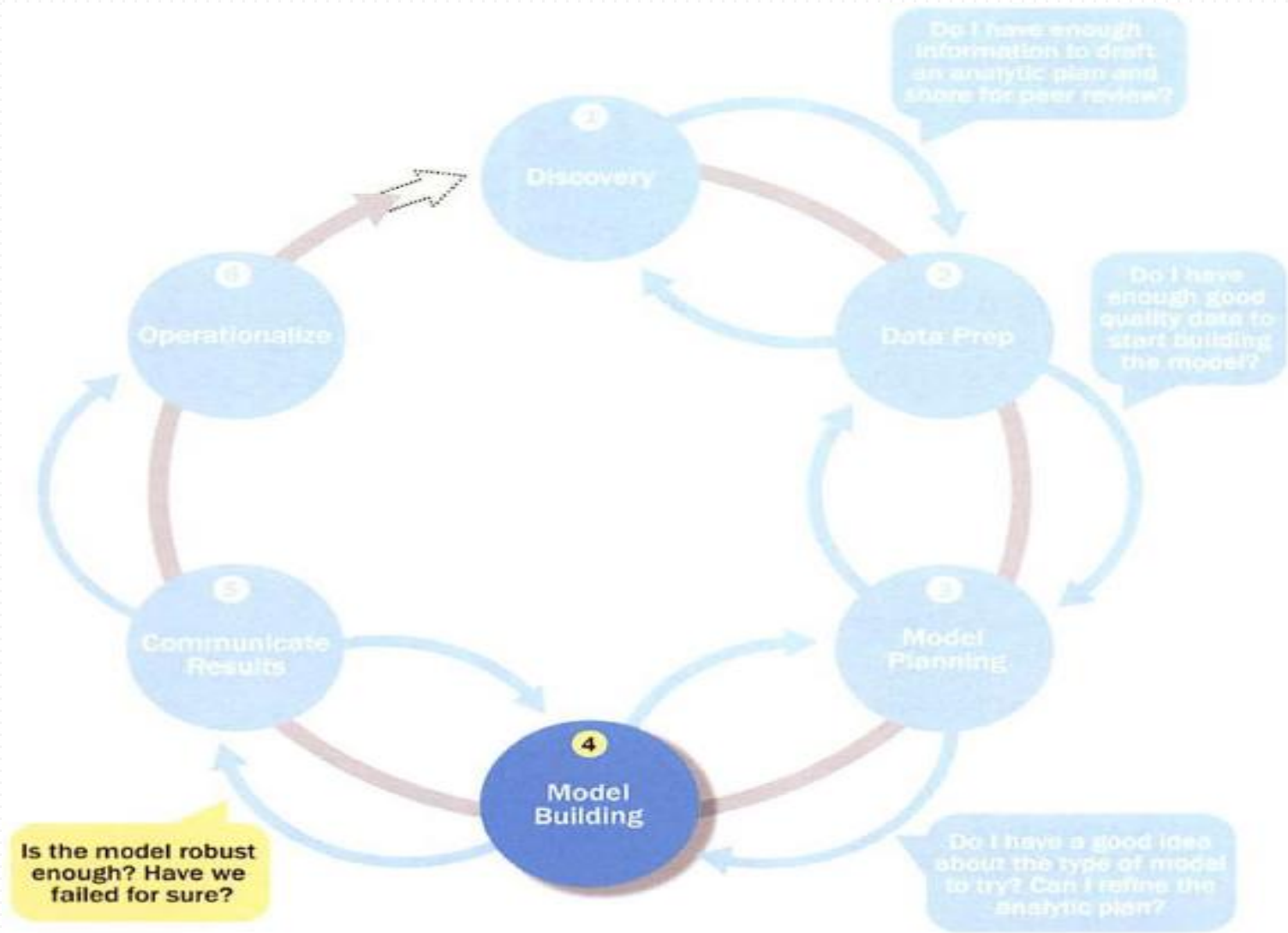
Model Selection

- The main goal is to choose an analytical technique, or several candidates, based on the end goal of the project
- We observe events in the real world and attempt to construct models that emulate this behavior with a set of rules and conditions
 - A model is simply an abstraction from reality
- 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

Common Tools for the Model Planning Phase

- **R** has a complete set of modelling capabilities
 - 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 via multiple data connections

Phase 4: Model Building



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/behaviour 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?
 - Does the model avoid intolerable mistakes?
 - 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?

Common Tools for the Model Building Phase

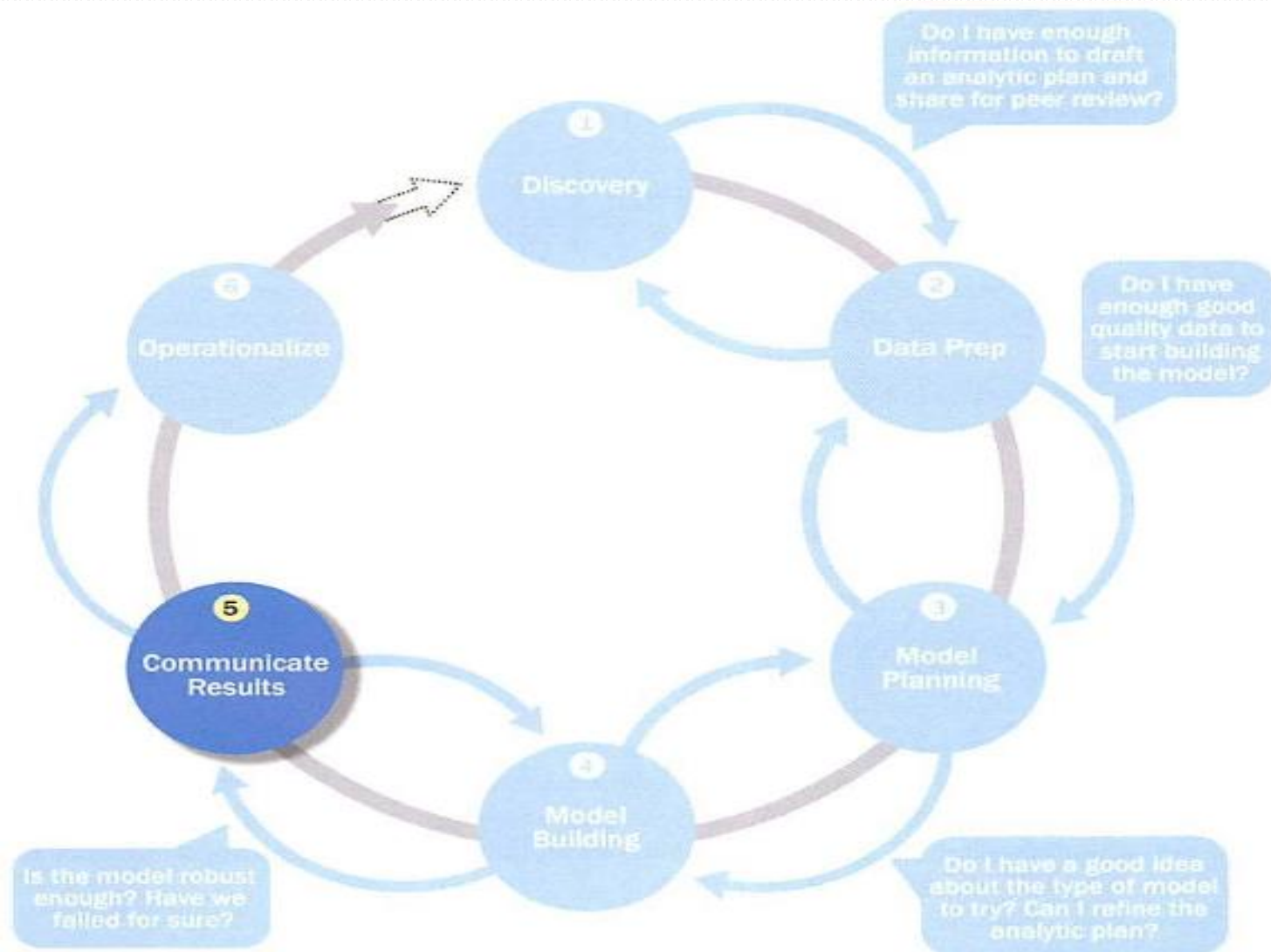
- **Commercial Tools**

- SAS Enterprise Miner – built for enterprise-level computing and analytics
- SPSS Modeler (IBM) – provides enterprise-level computing and analytics
- Matlab – high-level language for data analytics, algorithms, data exploration
- Alpine Miner – provides GUI frontend for backend analytics tools
- STATISTICA and MATHEMATICA – popular data mining and analytics tools

- **Free or Open Source Tools**

- R and PL/R - PL/R is a procedural language for PostgreSQL with R
- Octave – language for computational modeling
- WEKA – data mining software package with analytic workbench
- Python – language providing toolkits for machine learning and analysis
- SQL – in-database implementations provide an alternative tool

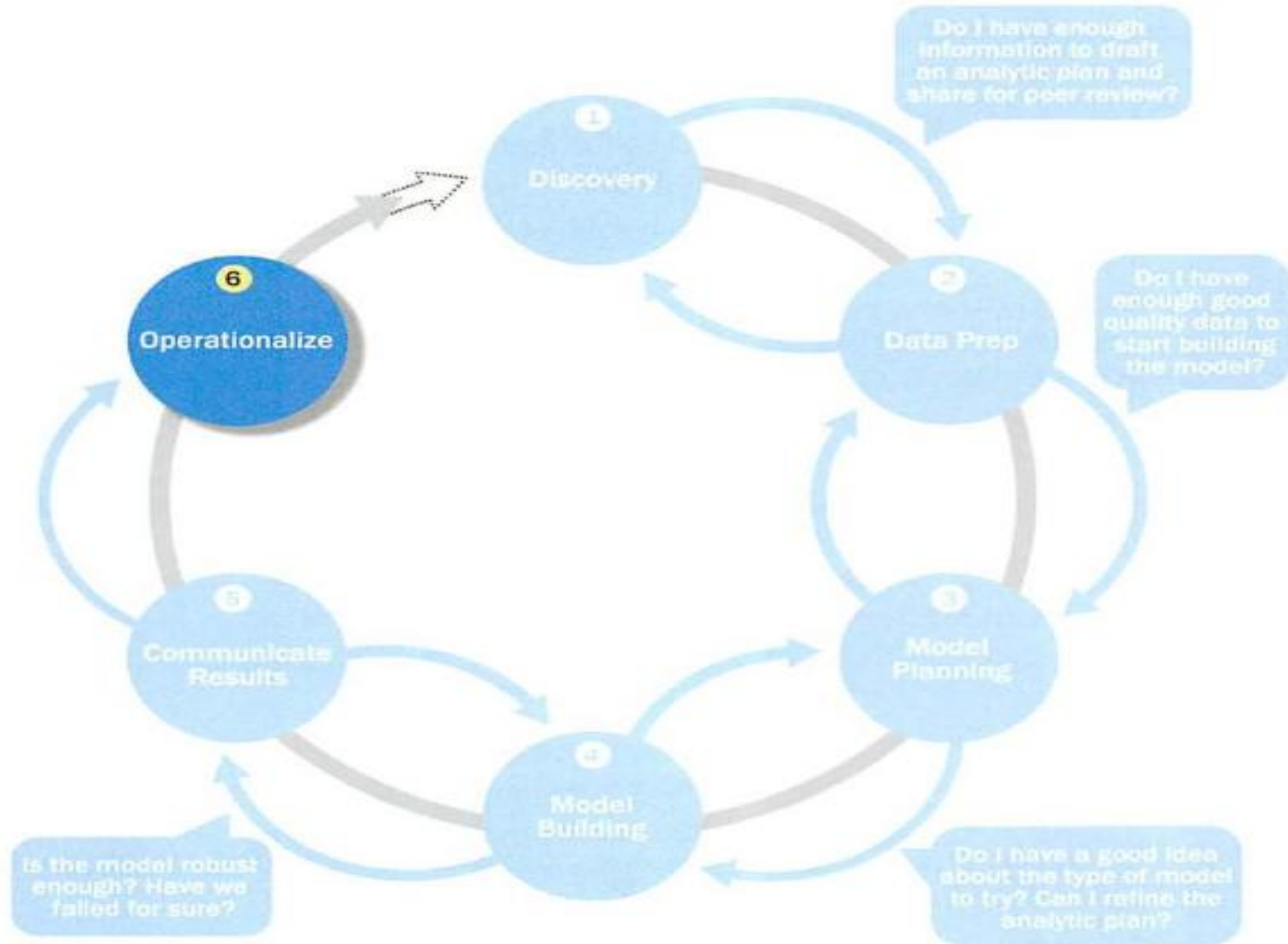
Phase 5: Communicate Results



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

Phase 6: Operationalize



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
- 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

Phase 6: Operationalize

Four main deliverables

Although the various roles represent many interests, the interests overlap and can be met with four main deliverables

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

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

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

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

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

Phase 5: Communicate Results

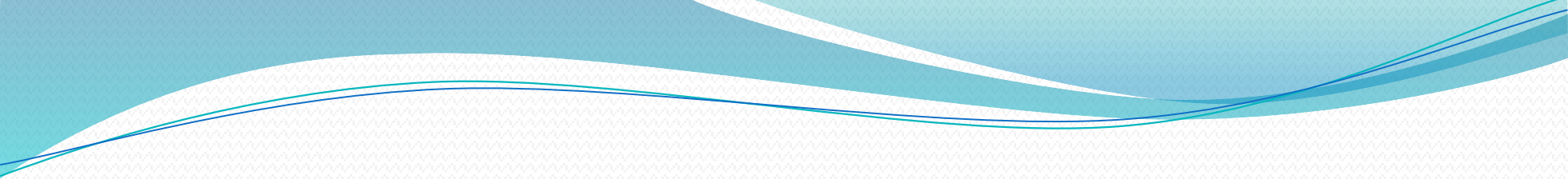
- Study was successful in identifying hidden innovators
- Found high density of innovators in Cork, Ireland
- The CTO office launched longitudinal studies

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

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



END
of
UNIT III