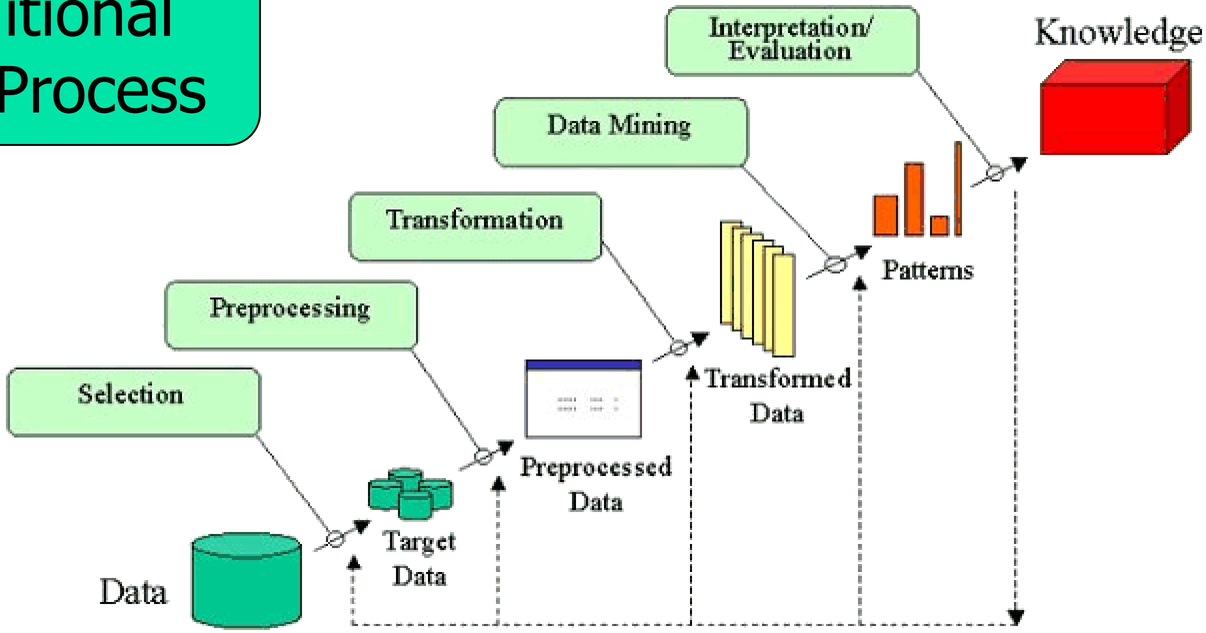
Data Analytics Lifecycle

Chapter 2 from "Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data"
1st Edition by EMC Education Services

Ka-Chun Wong, Department of Computer Science City University of Hong Kong
Charles Tappert Seidenberg, School of CSIS, Pace University

In data mining, we have Traditional KDD Process

Traditional KDD Process



http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html

Here, we have
Data Analytics Lifecycle

- Data Analytics Lifecycle Overview
- Phase 1: Discovery
- Phase 2: Data Preparation
- Phase 3: Model Planning
- Phase 4: Model Building
- Phase 5: Result Communication
- Phase 6: Resultant Operation

Traditional KDD Process

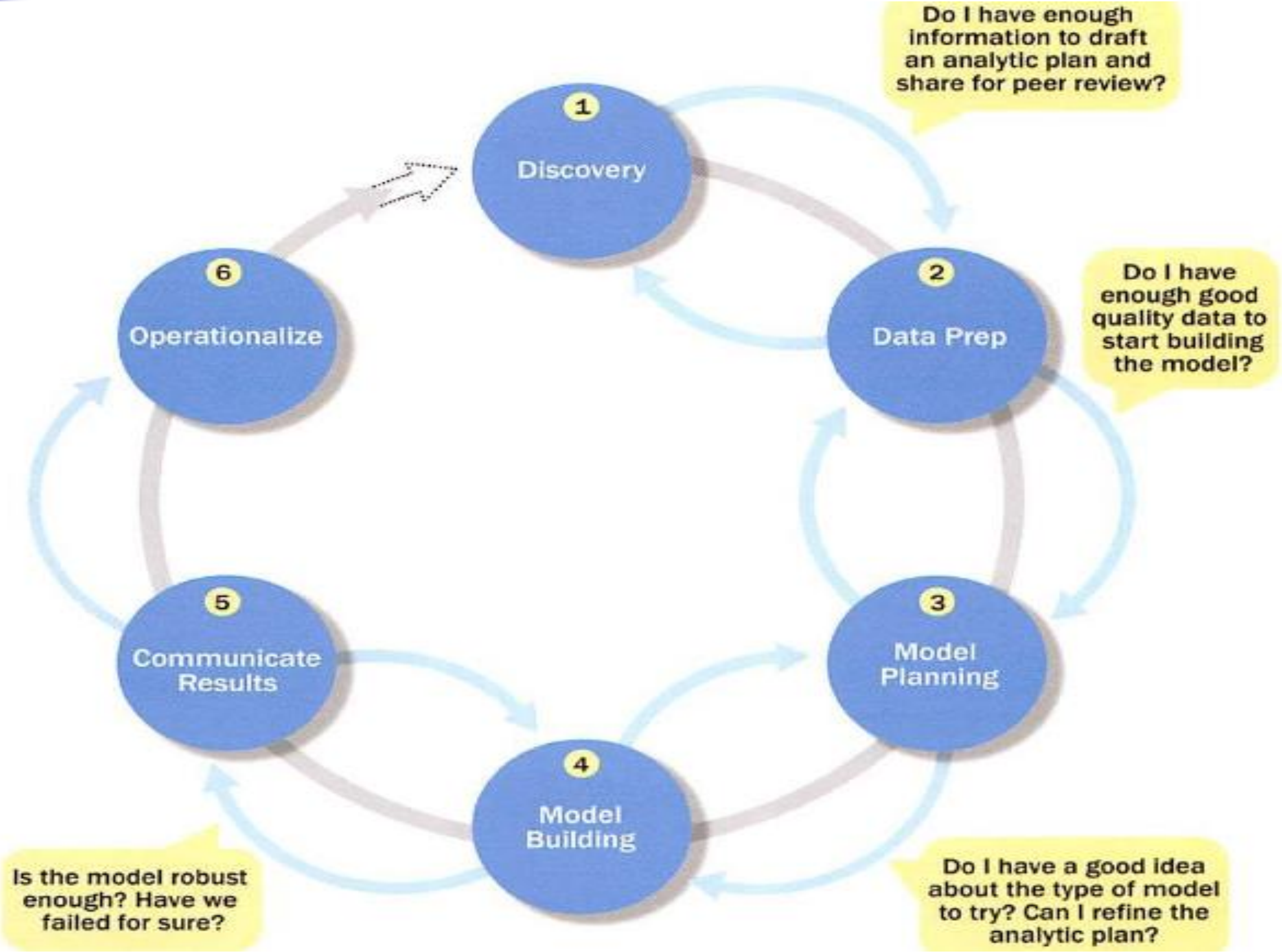
Here, we have
Data Analytics Lifecycle

- Why “Data Analytics Lifecycle”? Because:
  - 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 on planning and scope the work
  - Add rigor and credibility by documentation

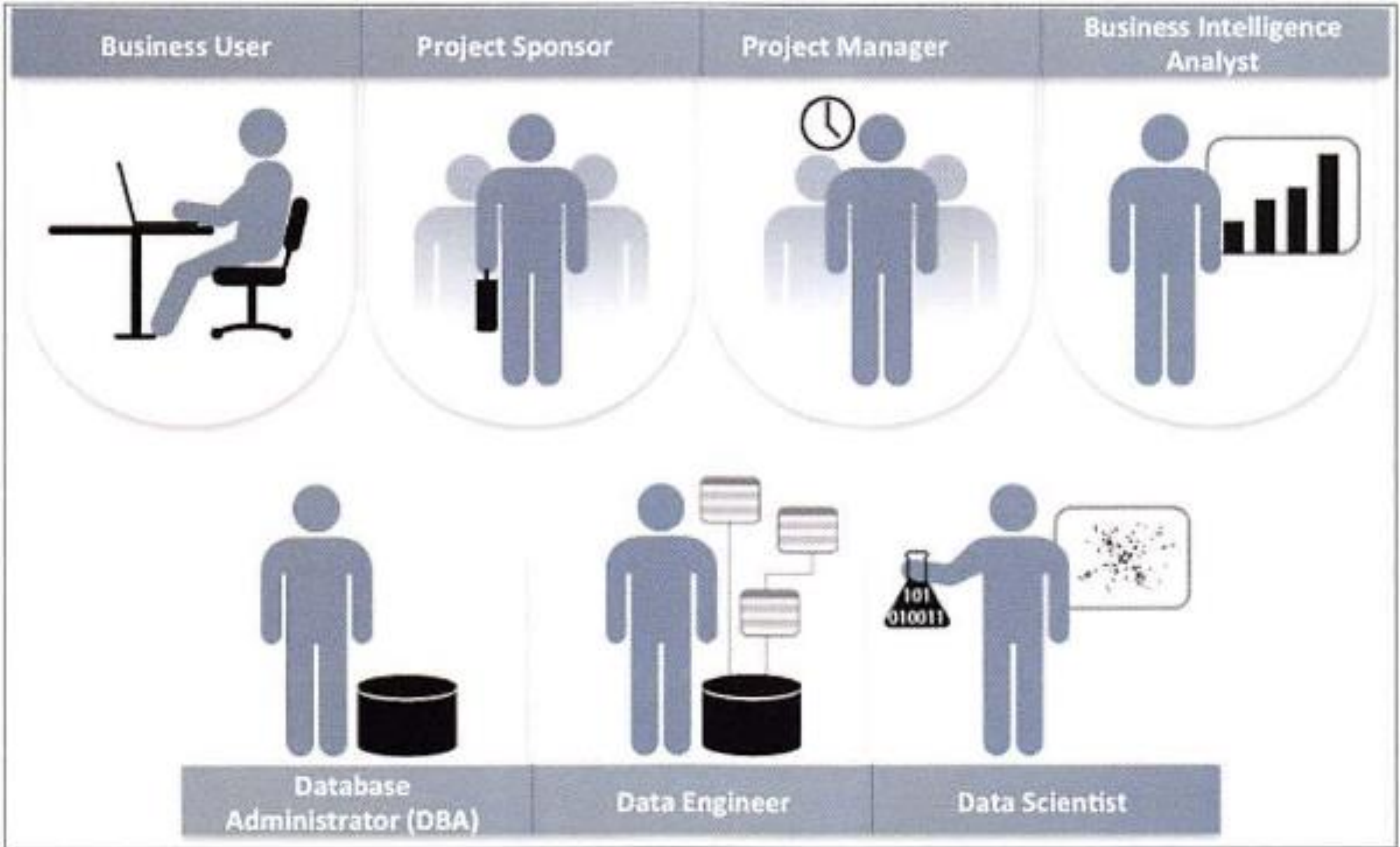
2.1 Data Analytics Lifecycle

- The data analytics 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

Overview of Data Analytics Lifecycle


Do I have enough information to draft an analytic plan and share for peer review?
1
Discovery
2
Data Prep
Do I have enough good quality data to start building the model?
3
Model Planning
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
4
Model Building
5
Communicate Results
Is the model robust enough? Have we failed for sure?
6
Operationalize

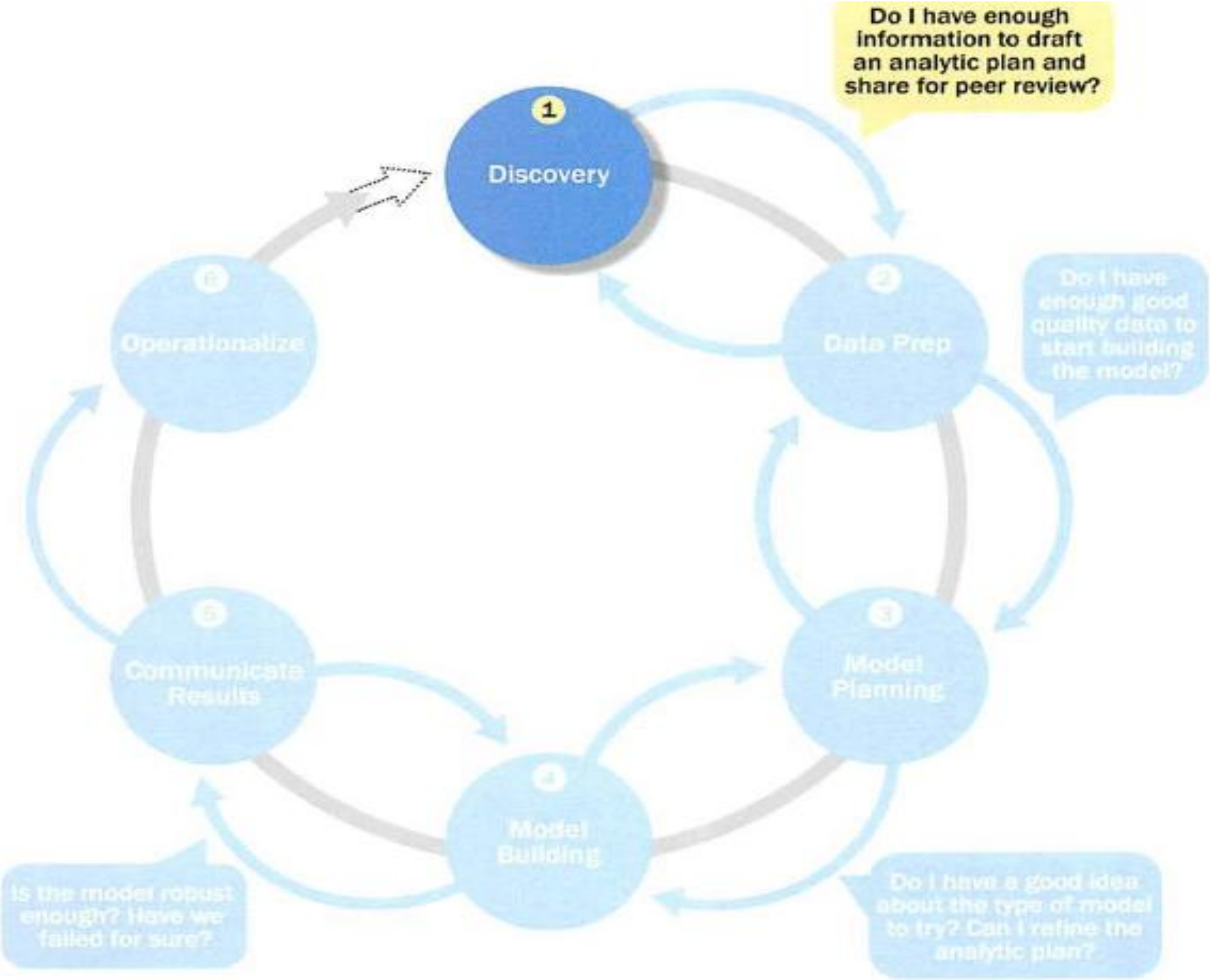
2.1.1 Key Roles for a Successful Analytics Project



2.1.1 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

2.2 Phase 1: Discovery



Do I have enough information to draft an analytic plan and share for peer review?

Do I have enough good quality data to start building the model?

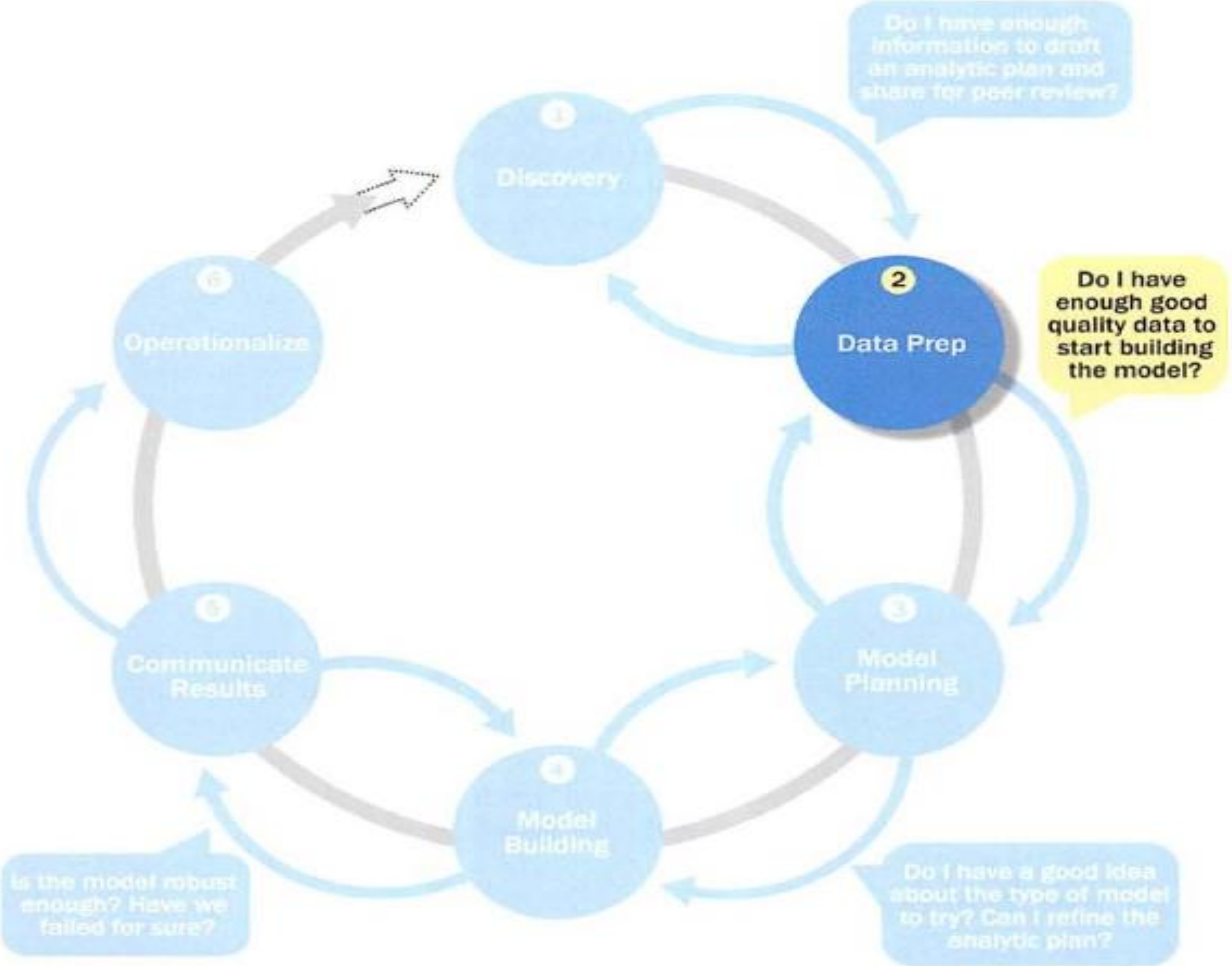
Do I have a good idea about the type of model to try? Can I refine the analytic plan?

Is the model robust enough? Have we failed for sure?

2.2 Phase 1: Discovery

1. Learning the Business Domain
2. Data and Computing 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


Do I have enough information to draft an analytic plan and share for peer review?
1
Discovery
2
Data Prep
Do I have enough good quality data to start building the model?
3
Model Planning
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
4
Model Building
5
Communicate Results
Is the model robust enough? Have we failed for sure?
6
Operationalize

2.3 Phase 2: Data Preparation

- Create robust environment – analytics sandbox
- Includes steps to explore, preprocess, and condition data. Data fusion is also needed.
- This phase 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 such as virtual machine or virtual environment)
- Allow 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 ETL (Extract, Transform, Load)

- Users perform Extract, Transform, and Load (ETL).
- Early data load can preserve the raw data which can be useful to examine in a retro style.
  - 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.
- Parallel computing (e.g. MapReduce) is often used here for ETL.

2.3.2 Performing ETL (Extract, Transform, Load)

```
import glob
import pandas as pd
import xml.etree.ElementTree as ET
from datetime import datetime
```

CSV Extract Function

```
def extract_from_csv(file_to_process):
  dataframe = pd.read_csv(file_to_process)
  return dataframe
```

JSON Extract Function

```
def extract_from_json(file_to_process):
  dataframe = pd.read_json(file_to_process,lines=True)
  return dataframe
```

XML Extract Function

```
def extract_from_xml(file_to_process):

    dataframe = pd.DataFrame(columns=['car_model','year_of_manufacture','price', 'fuel'])
    tree = ET.parse(file_to_process)
    root = tree.getroot()
    for person in root:
        car_model = person.find("car_model").text
        year_of_manufacture = int(person.find("year_of_manufacture").text)
        price = float(person.find("price").text)
        fuel = person.find("fuel").text
        dataframe = dataframe.append({"car_model":car_model,
                    year_of_manufacture":year_of_manufacture, "price":price, "fuel":fuel},
                    ignore_index=True)

        return dataframe
```

https://hevodata.com/learn/etl-using-python/

2.3.2 Performing ETL (Extract, Transform, Load)

```
def extract():
  extracted_data = pd.DataFrame(columns=['car_model','year_of_manufacture','price', 'fuel'])
  #for csv files
   for csvfile in glob.glob("dealership_data/*.csv"):
     extracted_data = extracted_data.append(extract_from_csv(csvfile), ignore_index=True)
  #for json files
   for jsonfile in glob.glob("dealership_data/*.json"):
     extracted_data = extracted_data.append(extract_from_json(jsonfile), ignore_index=True)
  #for xml files
   for xmlfile in glob.glob("dealership_data/*.xml"):
     extracted_data = extracted_data.append(extract_from_xml(xmlfile), ignore_index=True)
   return extracted_data
```

https://hevodata.com/learn/etl-using-python/

2.3.2 Performing ETL
(Extract, Transform, Load)

```
def transform(data):
        data['price'] = round(data.price, 2)
        data['car_model`] = data.car_model.str.upper()
        return data

def load(targetfile,data_to_load):
      data_to_load.to_csv(targetfile)


extracted_data = extract()
transformed_data = transform(extracted_data)
load(targetfile,transformed_data)
```

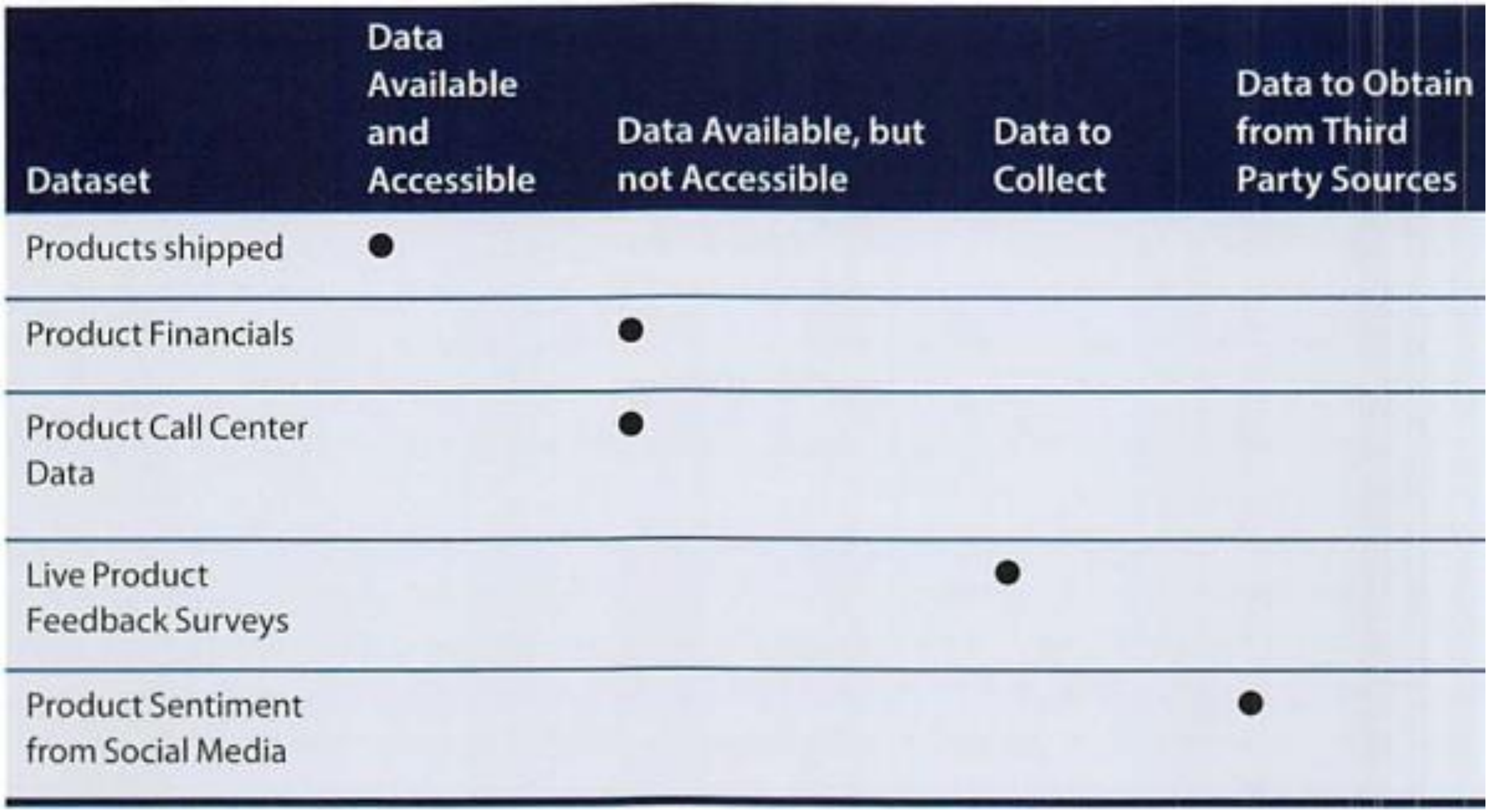
Popular ETL Tools:
- Apache Airflow
- Luigi
- Pandas
- Bonobo
- Petl
- ……

https://hevodata.com/learn/etl-using-python/   https://hevodata.com/learn/python-etl-tools

2.3.3 Learning about the Data

- Learning about the data is critical
- This activity accomplishes several goals:
  - Determines the data available to the team early in the project
  - Highlights gaps – identifies the data are not currently available but useful
  - Identifies the possible data sources 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

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

2.3.5 Survey and Visualize

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

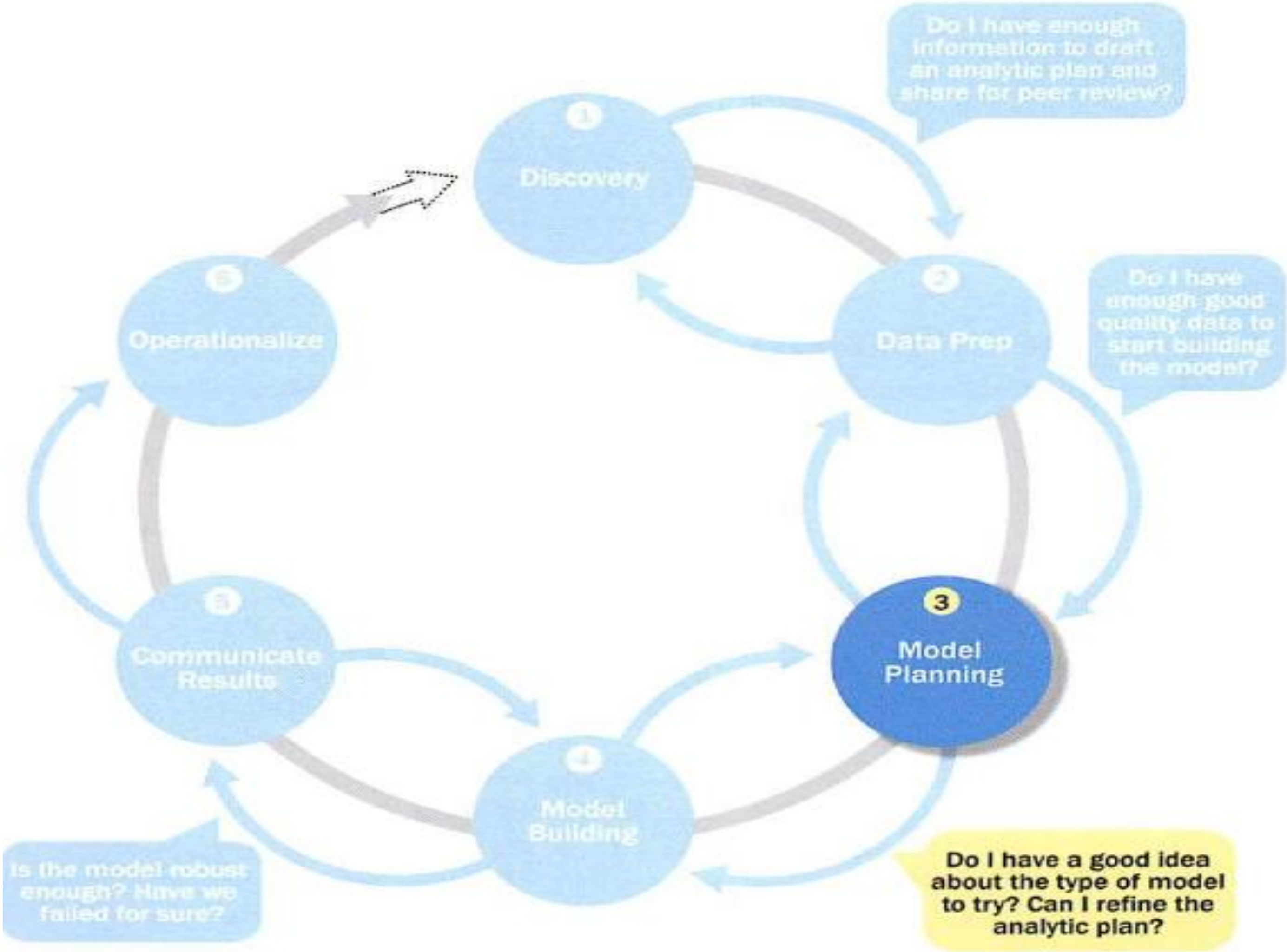
- Shneiderman's mantra:
  - "Overview first, zoom and filter, then details-on-demand"

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

2.3.6 Common Tools for Data Preparation

- **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 an transformation
- **Hadoop** can enable parallel ingest and analysis
- Many other tools are available. However, open-source is always preferred for data security.

2.4 Phase 3: Model Planning


Do I have enough information to draft an analytic plan and share for peer review?
1
Discovery
2
Data Prep
Do I have enough good quality data to start building the model?
3
Model Planning
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
4
Model Building
Is the model robust enough? Have we failed for sure?
5
Communicate Results
6
Operationalize

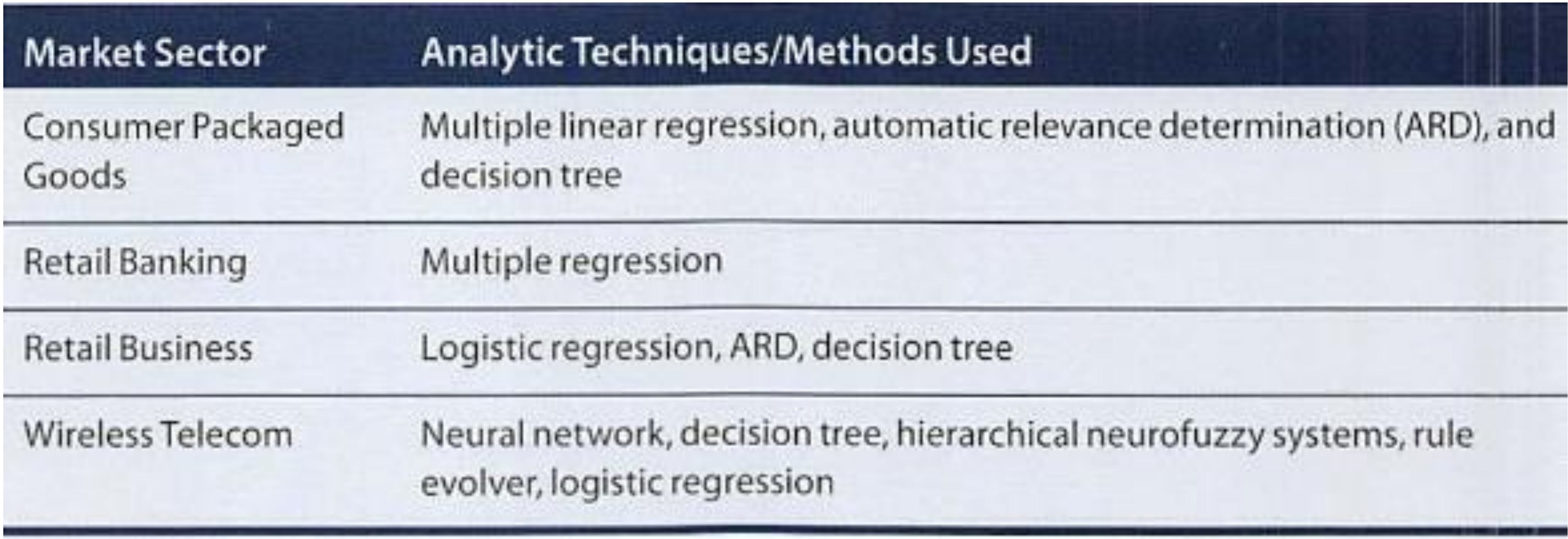
2.4 Phase 3: Model Planning

- Activities to consider
  - Assess the structure of the data – this dictates the tools and analytic techniques for the next phase
  - Ensure the 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 large analytic workflow
  - Research and understand how other analysts have approached this kind or similar kinds of problems

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

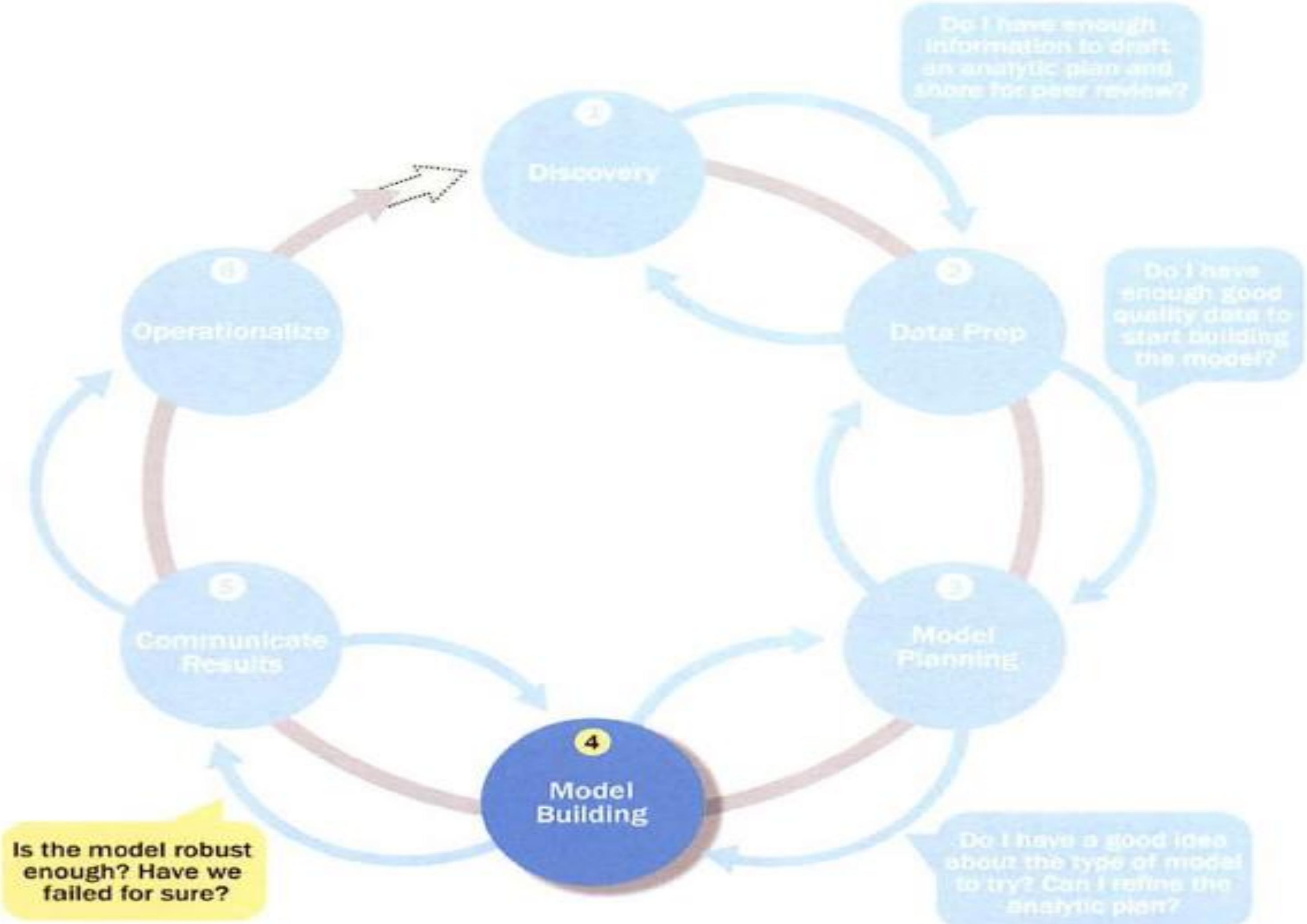
2.4.2 Model Selection

- 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 data science software packages such as Python, R, SAS, or Matlab
  - It 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 suitable type of model to try

2.4.3 Common Tools for the Model Planning Phase

- **Python** has many machine learning packages
  - e.g. scikit-learn, Pandas, Numpy, TensorFlow, Kereas, PyTorch ……
- **R** has a complete set of statistical modeling capabilities
  - R contains about 5000 packages for data analysis and graphical presentation
- **SQL** 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
- **Matlab or Octave** provides matrix modelling and scientific libraries in different domains

2.5 Phase 4: Model Building


Do I have enough information to draft an analytic plan and share for peer review?
1
Discovery
2
Data Prep
Do I have enough good quality data to start building the model?
3
Model Planning
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
4
Model Building
Is the model robust enough? Have we failed for sure?
5
Communicate Results
6
Operationalize

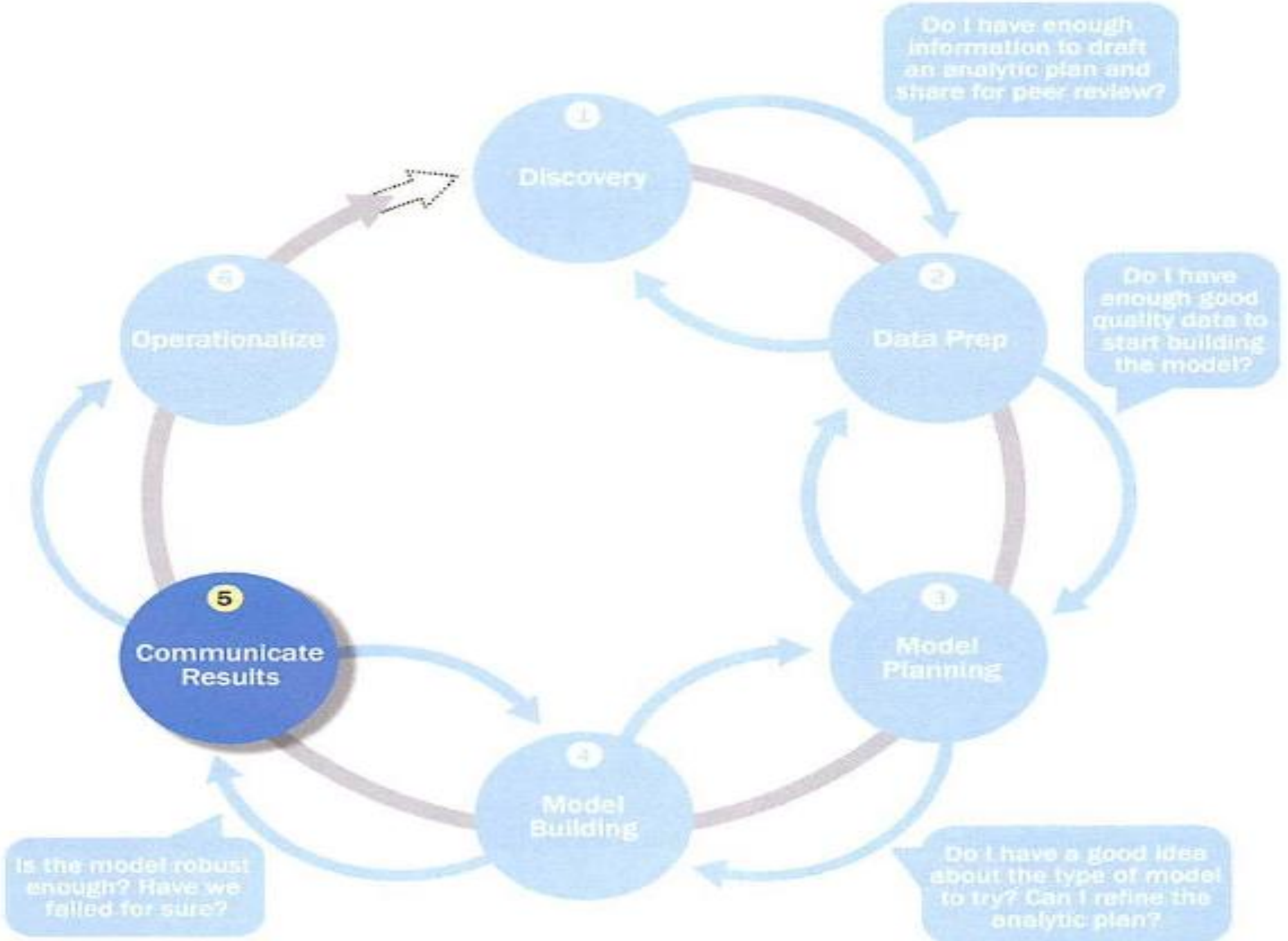
2.5 Phase 4: Model Building

- Focus on the models defined in Phase 3
- Construct datasets for training, testing, and production
- Train the models on training data and 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?
  - Does the model avoid intolerable mistakes? (see Chapters 3 and 7 of textbook)
  - 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?

2.5.1 Common Tools for the Model Building Phase

- Free or Open Source Tools
  - R – language for statistical computing
  - Octave – language for computational modeling
  - WEKA – data mining software package in Java
  - Python – language for machine learning and analysis
  - SQL – in-database implementations
- 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
  - STATISTICA and MATHEMATICA – popular data mining and analytics tools

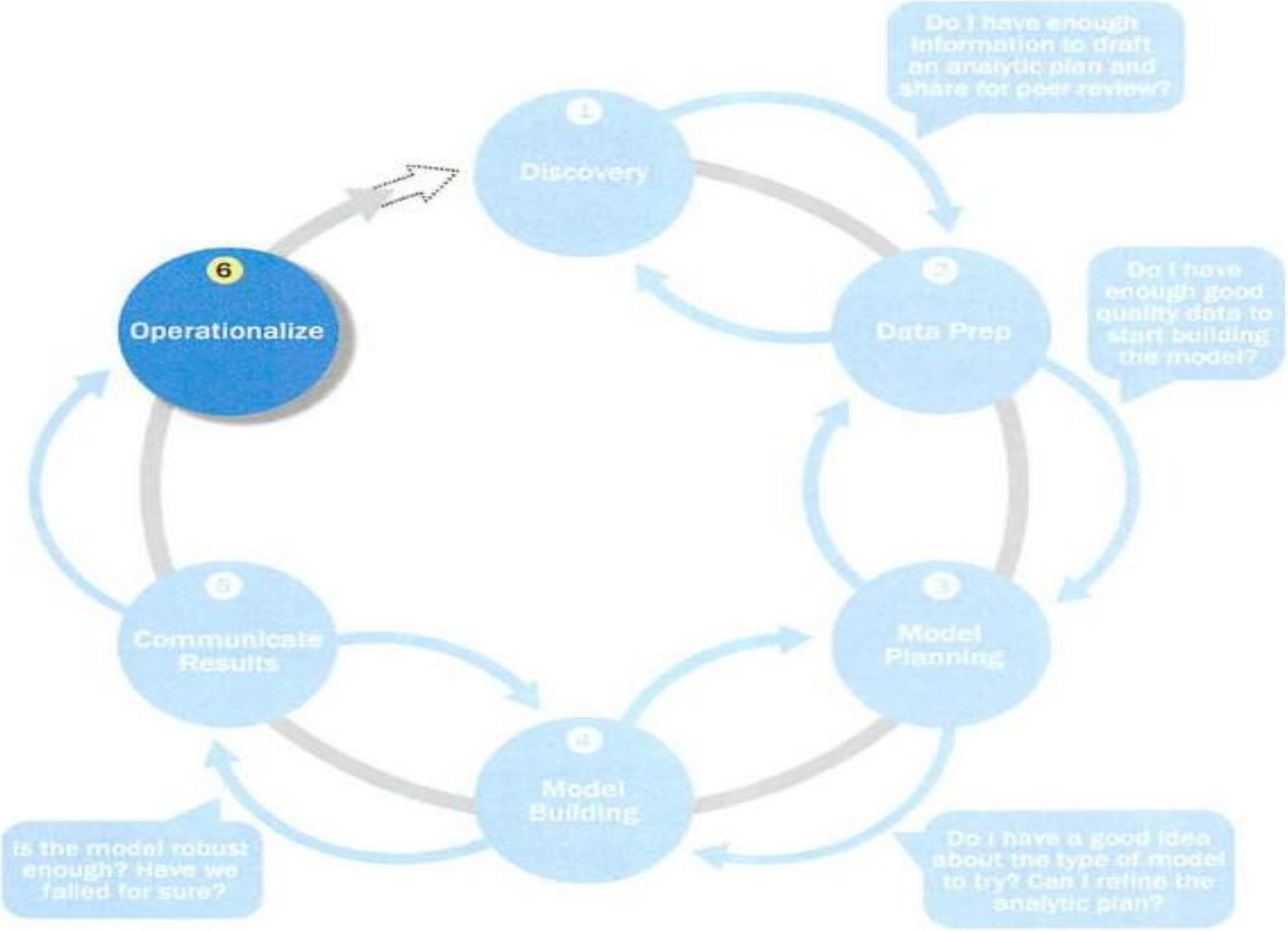
2.6 Phase 5: Communicate Results


Do I have enough information to draft an analytic plan and share for peer review?
1
Discovery
2
Data Prep
Do I have enough good quality data to start building the model?
3
Model Planning
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
4
Model Building
Is the model robust enough? Have we failed for sure?
5
Communicate Results
6
Operationalize

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

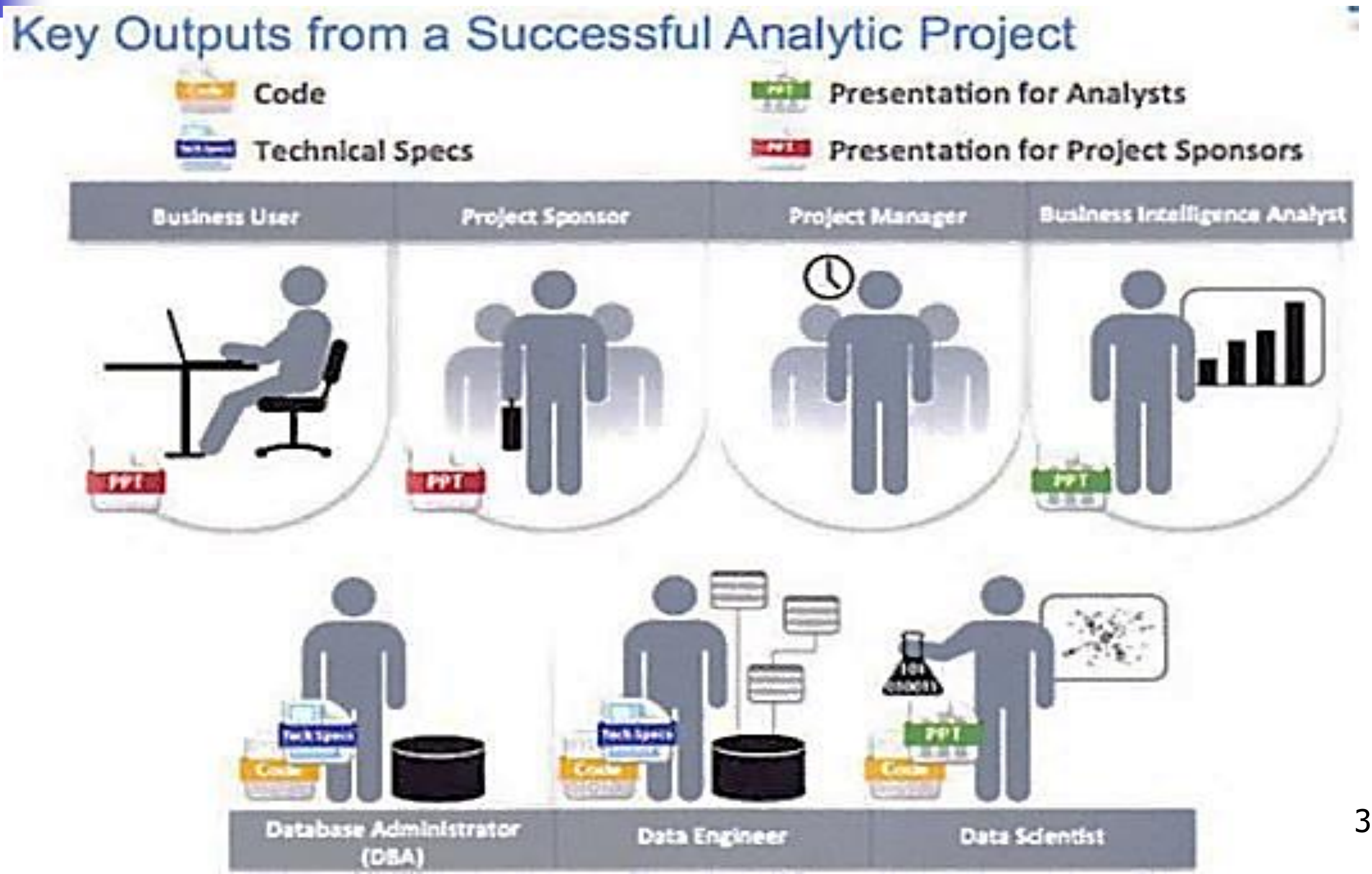
2.7 Phase 6: Operationalize



Do I have enough information to draft an analytic plan and share for peer review?
1
Discovery
2
Data Prep
Do I have enough good quality data to start building the model?
3
Model Planning
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
4
Model Building
5
Communicate Results
Is the model robust enough? Have we failed for sure?
6
Operationalize

2.7 Phase 6: Operationalize

Key outputs from successful analytics project



Key Outputs from a Successful Analytic Project

2.7 Phase 6: Operationalize

Four main deliverables

- Although the seven 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
     - Detailed descriptions for business process changes and reporting changes; it includes details and technical graphs
  3. Source Code for technical members
  4. Technical Specifications and Documentations

2.8 Case Study - Example

Global Innovation Network and Analysis (GINA)

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

- 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

2.8.1 Phase 1: Discovery

- Team members and roles
  - Business user, project sponsor, project manager <= Vice President from Office of CTO
  - BI analyst <= people 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 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 invest in the future

2.8.2 Phase 2: Data Preparation

- EMC set up an analytics sandbox (e.g. VirtualBox)
- They discovered that certain data needed conditioning and normalization and that missing datasets were critical
- They 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

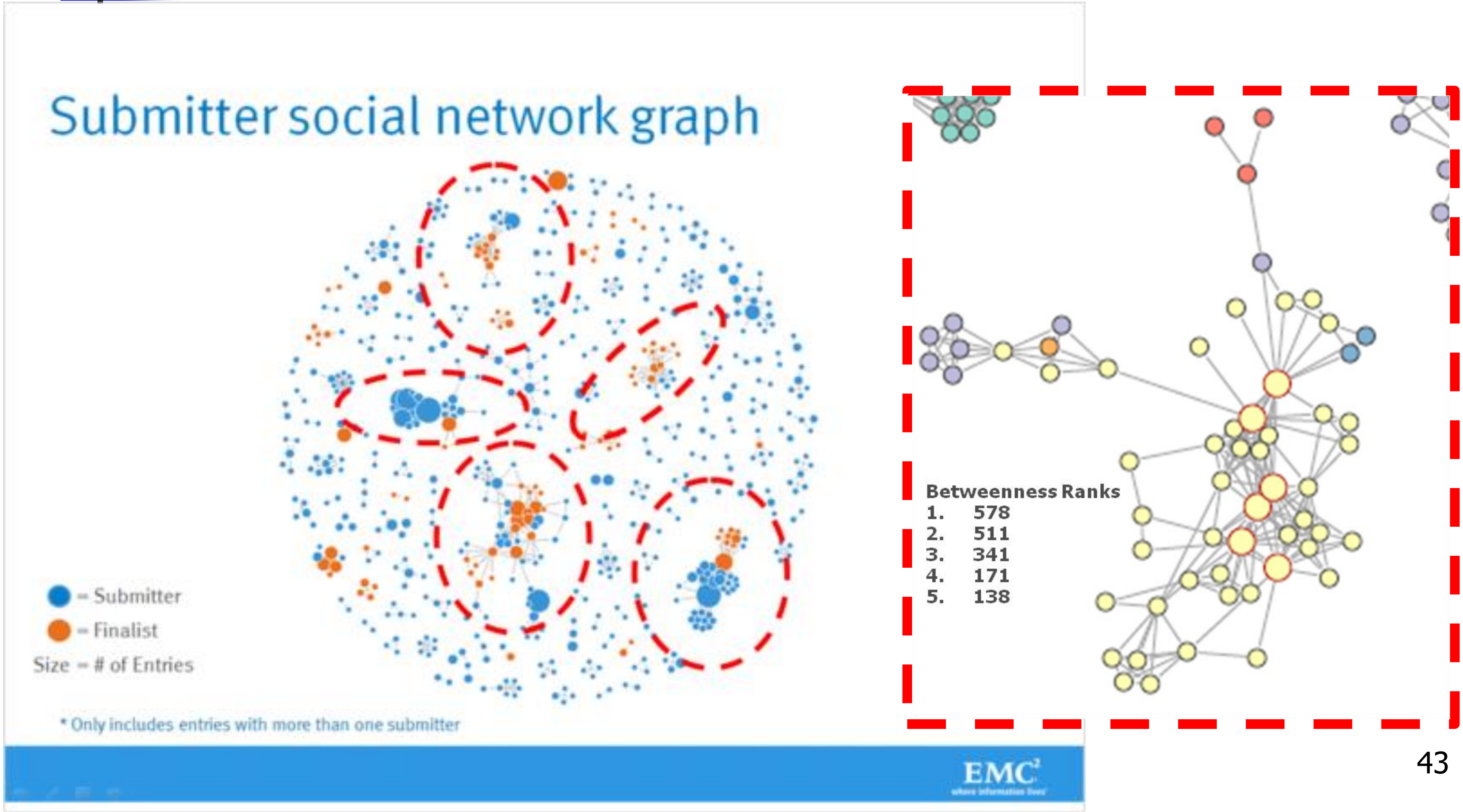
- The study included the following considerations
  - Identify the right milestones to achieve the goals
  - Trace how people move ideas from each milestone toward the goal
  - Trace ideas that die and others that reach the goal
  - Compare times and outcomes using a few different methods

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



Submitter social network graph

● – Submitter

● – Finalist

Size – # of Entries

* Only includes entries with more than one submitter

Betweenness Ranks

1.	578
2.	511
3.	341
4.	171
5.	138

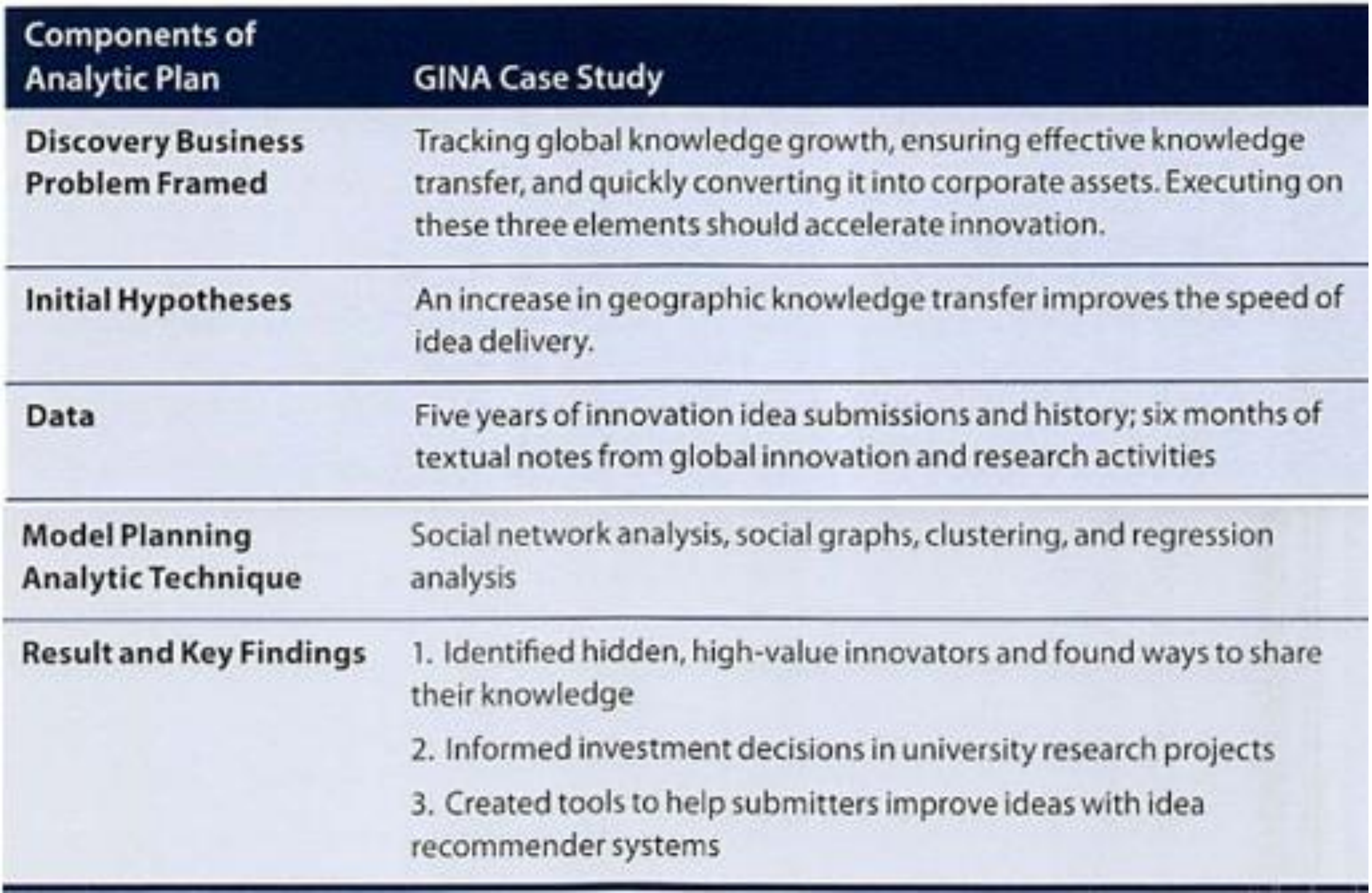
2.8.5 Phase 5: Communicate Results

- Social network modelling was successful in identifying hidden innovators
  - Found high density of innovators in Cork, Ireland
    - https://stevetodd.typepad.com/my_weblog/2011/10/world-analytics.html

2.8.6 Phase 6: Operationalize

- Key findings
  - We 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 re-evaluate the model after deployment
- Operation
  - The CTO office launched longitudinal studies on longer periods of time for future insights in Cork

Example Case Summary



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	1. Identified hidden, high-value innovators and found ways to share their knowledge 2. Informed investment decisions in university research projects 3. Created tools to help submitters improve ideas with idea recommender systems

Summary

- The Data Analytics Lifecycle is an approach to manage and execute data computing projects because data science is exploratory
- Lifecycle has 6 phases
- Bulk of the time are usually spent on preparation – Phases 1 and 2

Appendix

2.1.2 Background and Overview of Data Analytics Lifecycle

- Data Analytics Lifecycle defines the analytics process and best practices from discovery to project completion
- The Lifecycle employs aspects of
  - Scientific method
  - Cross Industry Standard Process for Data Mining (CRISP-DM)
    - Process model for data mining
  - Davenport's DELTA framework
  - Hubbard's Applied Information Economics (AIE) approach
  - MAD Skills: New Analysis Practices for Big Data by Cohen et al.

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 Guidelines and Considerations

- Review data to ensure calculations are consistent
- Do the data distribution stay consistent?
- Assess the granularity of the data, the range of values, and the level of aggregation of the data
- Do 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

2.4.1 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
- 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

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
- 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

2.7 Phase 6: Operationalize
Key outputs from successful analytics project

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