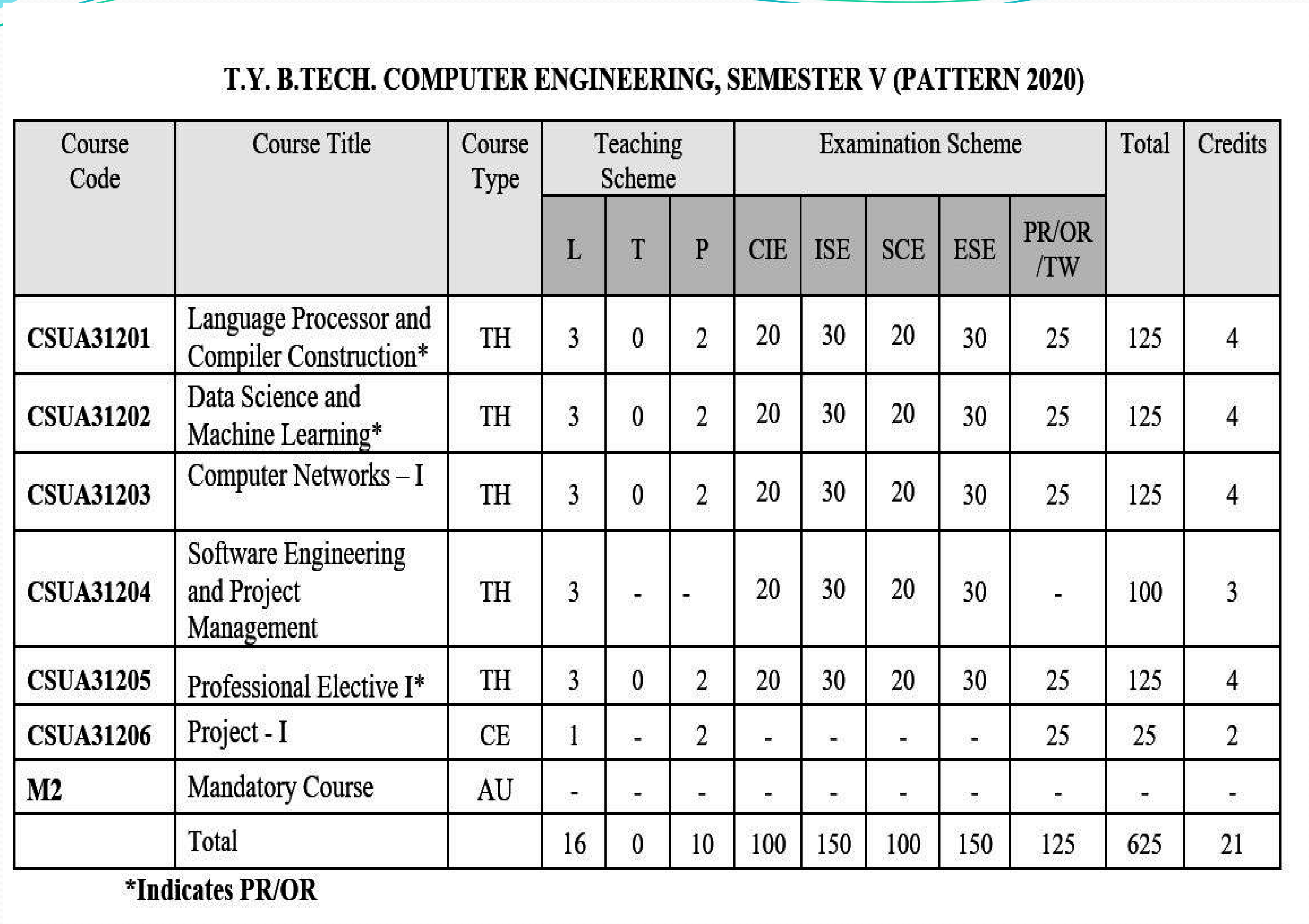
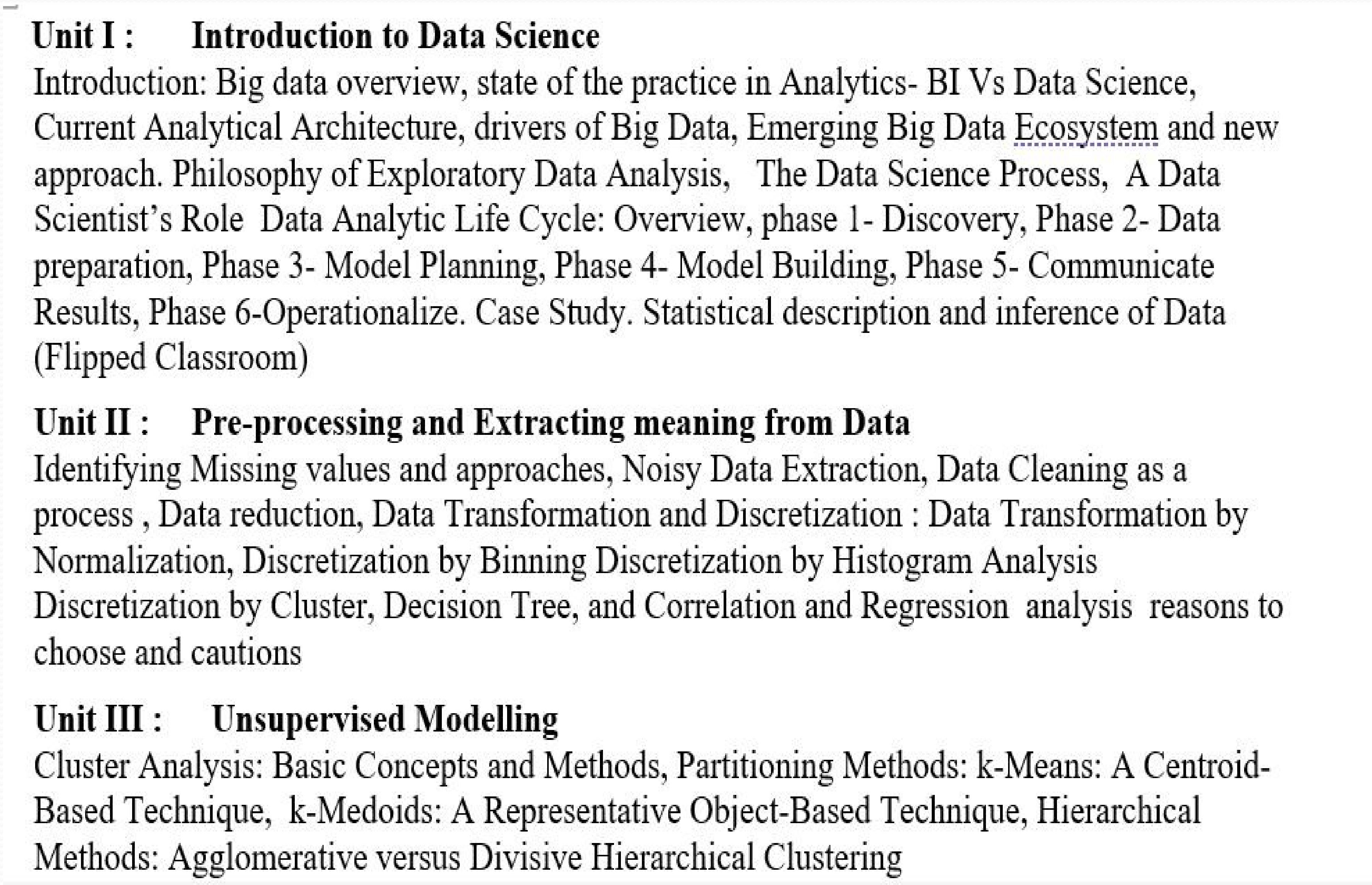


## T.Y. B.TECH. COMPUTER ENGINEERING, SEMESTER V (PATTERN 2020)

| Course Code | Course Title | Course Type | Teaching Scheme | | | Examination Scheme | | | | | Total | Credits |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | L | T | P | CIE | ISE | SCE | ESE | PR/OR /TW | | |
| **CSUA31201** | Language Processor and Compiler Construction* | TH | 3 | 0 | 2 | 20 | 30 | 20 | 30 | 25 | 125 | 4 |
| **CSUA31202** | Data Science and Machine Learning* | TH | 3 | 0 | 2 | 20 | 30 | 20 | 30 | 25 | 125 | 4 |
| **CSUA31203** | Computer Networks – I | TH | 3 | 0 | 2 | 20 | 30 | 20 | 30 | 25 | 125 | 4 |
| **CSUA31204** | Software Engineering and Project Management | TH | 3 | - | - | 20 | 30 | 20 | 30 | - | 100 | 3 |
| **CSUA31205** | Professional Elective I* | TH | 3 | 0 | 2 | 20 | 30 | 20 | 30 | 25 | 125 | 4 |
| **CSUA31206** | Project - I | CE | 1 | - | 2 | - | - | - | - | 25 | 25 | 2 |
| **M2** | Mandatory Course | AU | - | - | - | - | - | - | - | - | - | - |
| | Total | | 16 | 0 | 10 | 100 | 150 | 100 | 150 | 125 | 625 | 21 |

***Indicates PR/OR**

# Syllabus



**Unit I : Introduction to Data Science**

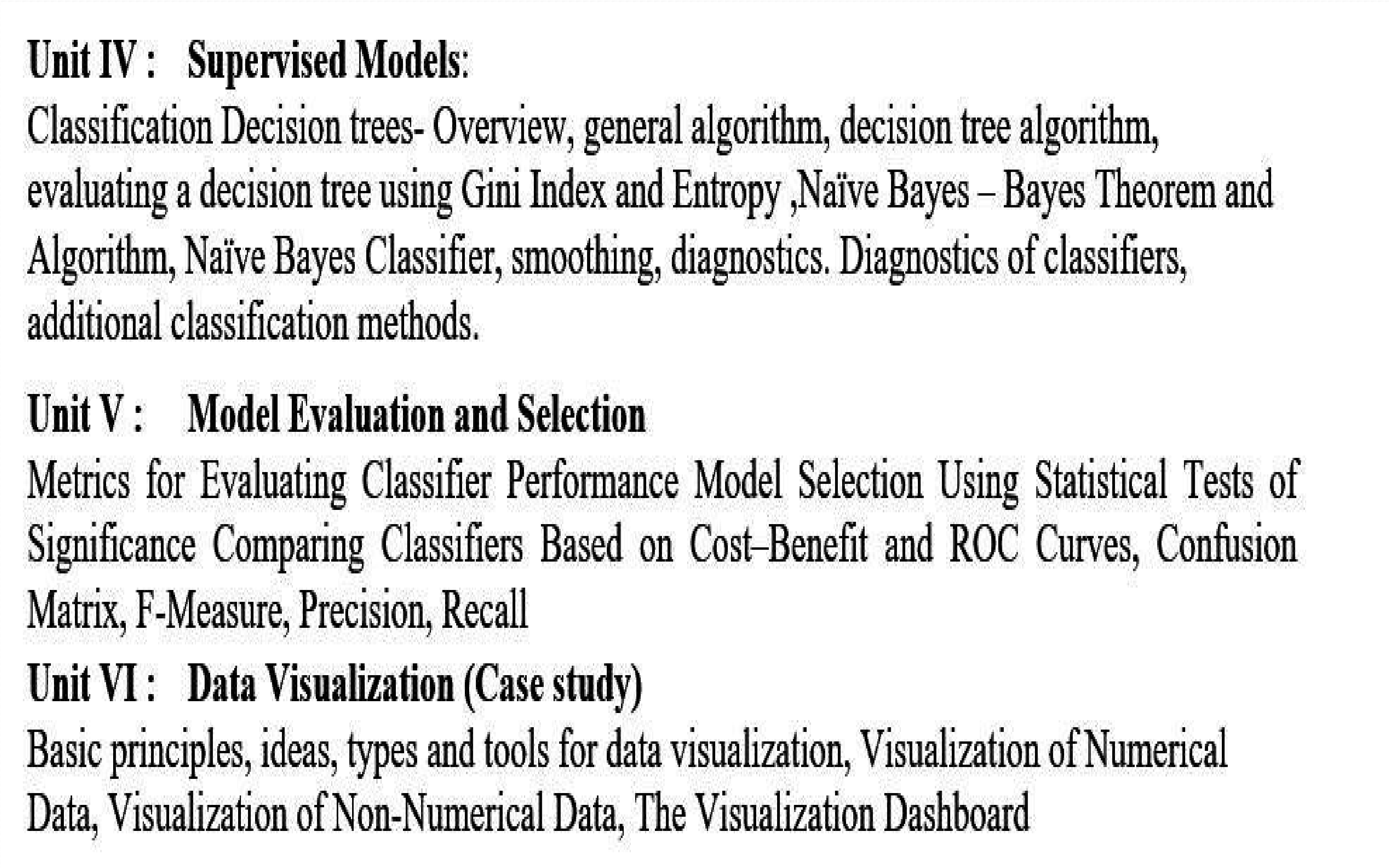
Introduction: Big data overview, state of the practice in Analytics- BI Vs Data Science, Current Analytical Architecture, drivers of Big Data, Emerging Big Data Ecosystem and new approach. Philosophy of Exploratory Data Analysis, The Data Science Process, A Data Scientist's Role Data Analytic Life Cycle: 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. Statistical description and inference of Data (Flipped Classroom)

**Unit II : Pre-processing and Extracting meaning from Data**

Identifying Missing values and approaches, Noisy Data Extraction, Data Cleaning as a process , Data reduction, Data Transformation and Discretization : Data Transformation by Normalization, Discretization by Binning Discretization by Histogram Analysis Discretization by Cluster, Decision Tree, and Correlation and Regression analysis reasons to choose and cautions

**Unit III : Unsupervised Modelling**

Cluster Analysis: Basic Concepts and Methods, Partitioning Methods: k-Means: A Centroid-Based Technique, k-Medoids: A Representative Object-Based Technique, Hierarchical Methods: Agglomerative versus Divisive Hierarchical Clustering



**Unit IV : Supervised Models:**

Classification Decision trees- Overview, general algorithm, decision tree algorithm, evaluating a decision tree using Gini Index and Entropy ,Naïve Bayes – Bayes Theorem and Algorithm, Naïve Bayes Classifier, smoothing, diagnostics. Diagnostics of classifiers, additional classification methods.

**Unit V : Model Evaluation and Selection**

Metrics for Evaluating Classifier Performance Model Selection Using Statistical Tests of Significance Comparing Classifiers Based on Cost–Benefit and ROC Curves, Confusion Matrix, F-Measure, Precision, Recall

**Unit VI : Data Visualization (Case study)**

Basic principles, ideas, types and tools for data visualization, Visualization of Numerical Data, Visualization of Non-Numerical Data, The Visualization Dashboard

# Course Outcomes

| | |
|---|---|
| | After completion of the course, student will be able to |
| | |
| **1.** | Describe the Data Science Process and explore components interaction. |
| **2.** | Apply statistical methods for pre-processing and extracting meaning from data to the application dataset. |
| **3.** | Apply specific unsupervised machine learning algorithm for a particular problem. |
| **4.** | Apply specific supervised machine learning algorithm for a particular problem. |
| **5.** | Analyse the outcome in terms of efficiency. |
| **6.** | Analyse and organize data using visualization tools. |

Concepts

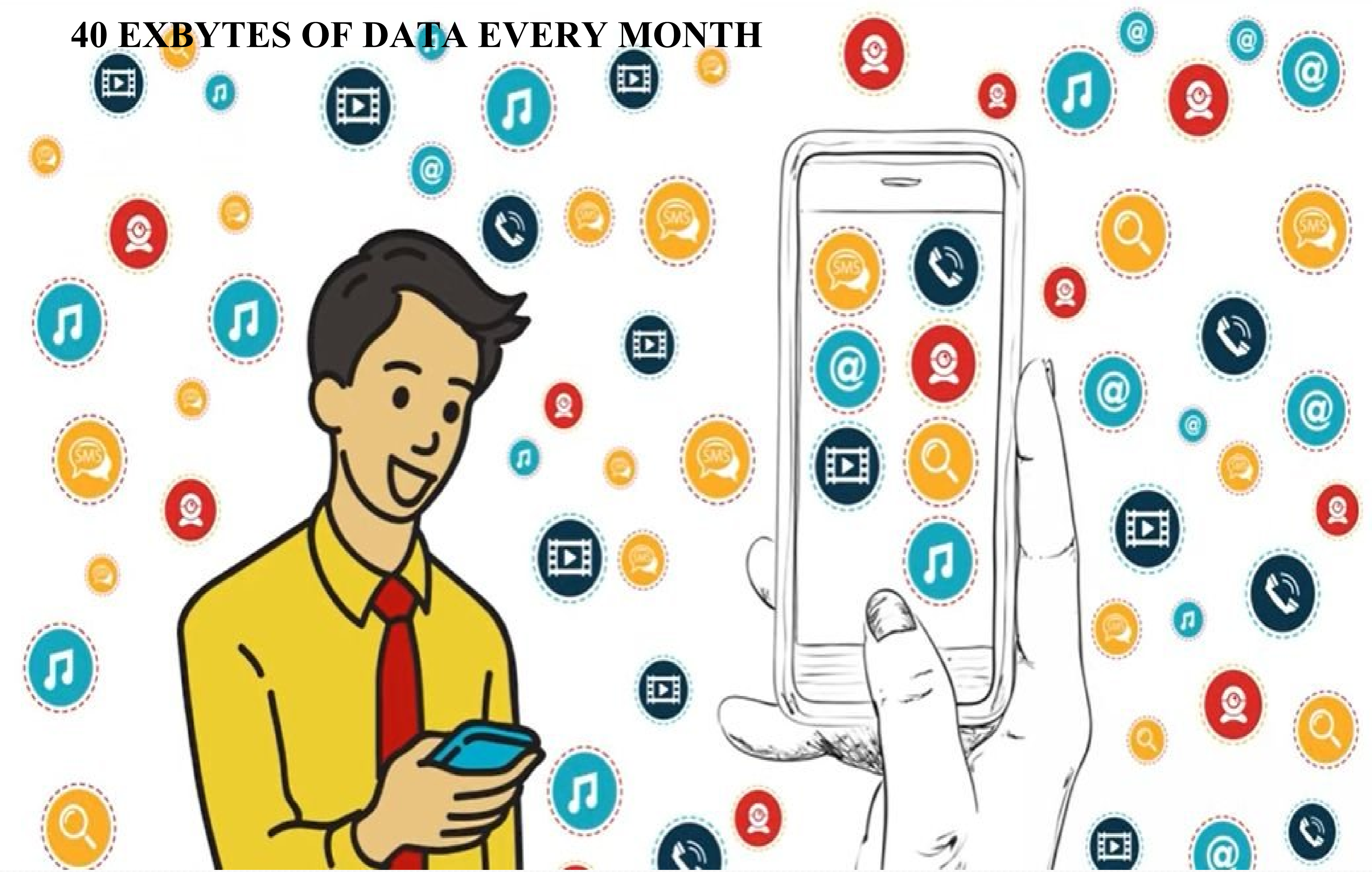
# UNIT-I

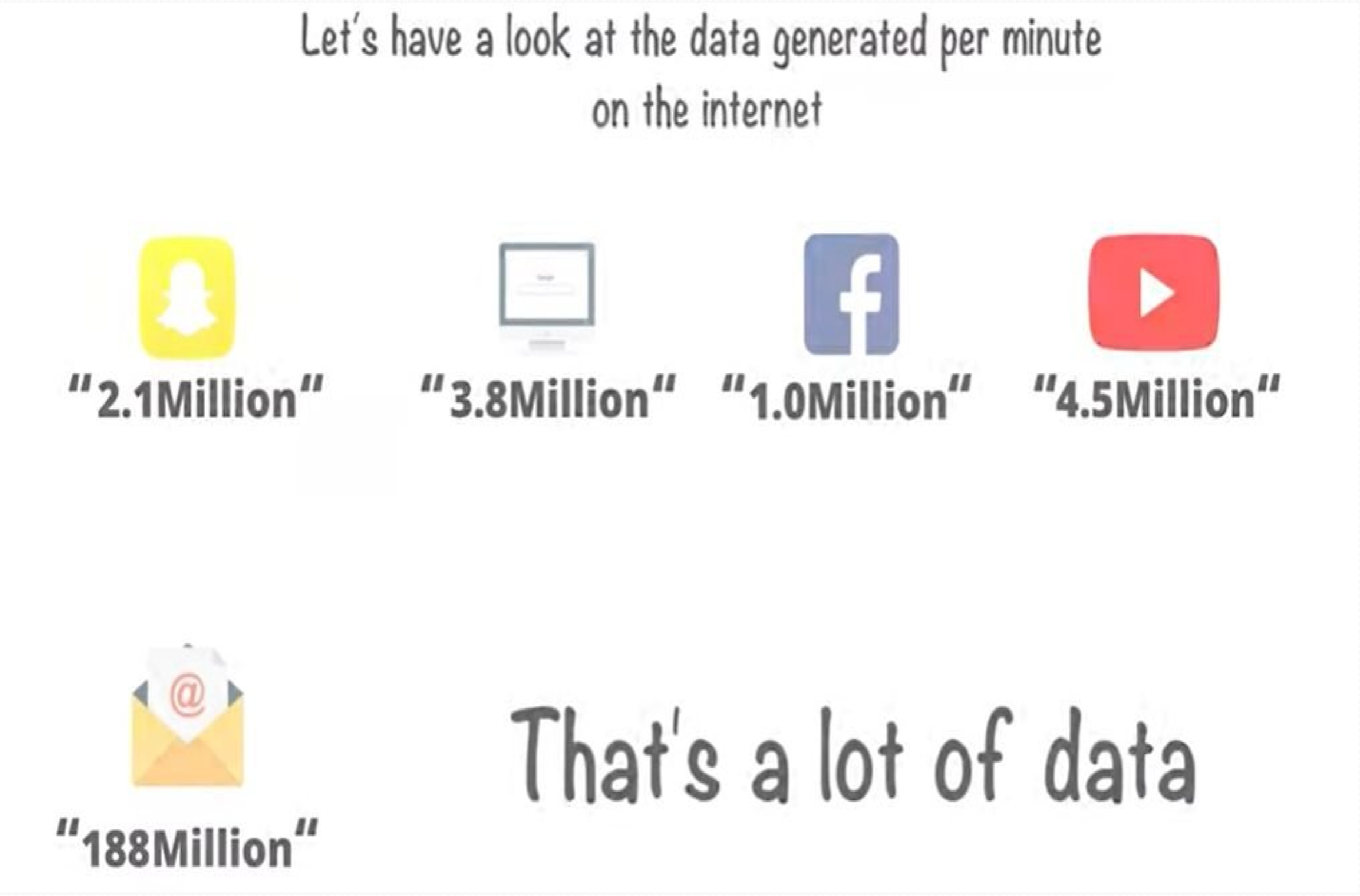
# INTRODUCTION TO DATA SCIENCE

# Syllabus

Introduction: Big data overview, state of the practice in Analytics- BI Vs Data Science, Current Analytical Architecture, drivers of Big Data, Emerging Big Data Ecosystem and new approach. Philosophy of Exploratory Data Analysis, The Data Science Process, A Data Scientist's Role Data Analytic Life Cycle: 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. Statistical description and inference of Data (Flipped Classroom)

## 40 EXBYTES OF DATA EVERY MONTH





Let's have a look at the data generated per minute on the internet

"2.1Million"

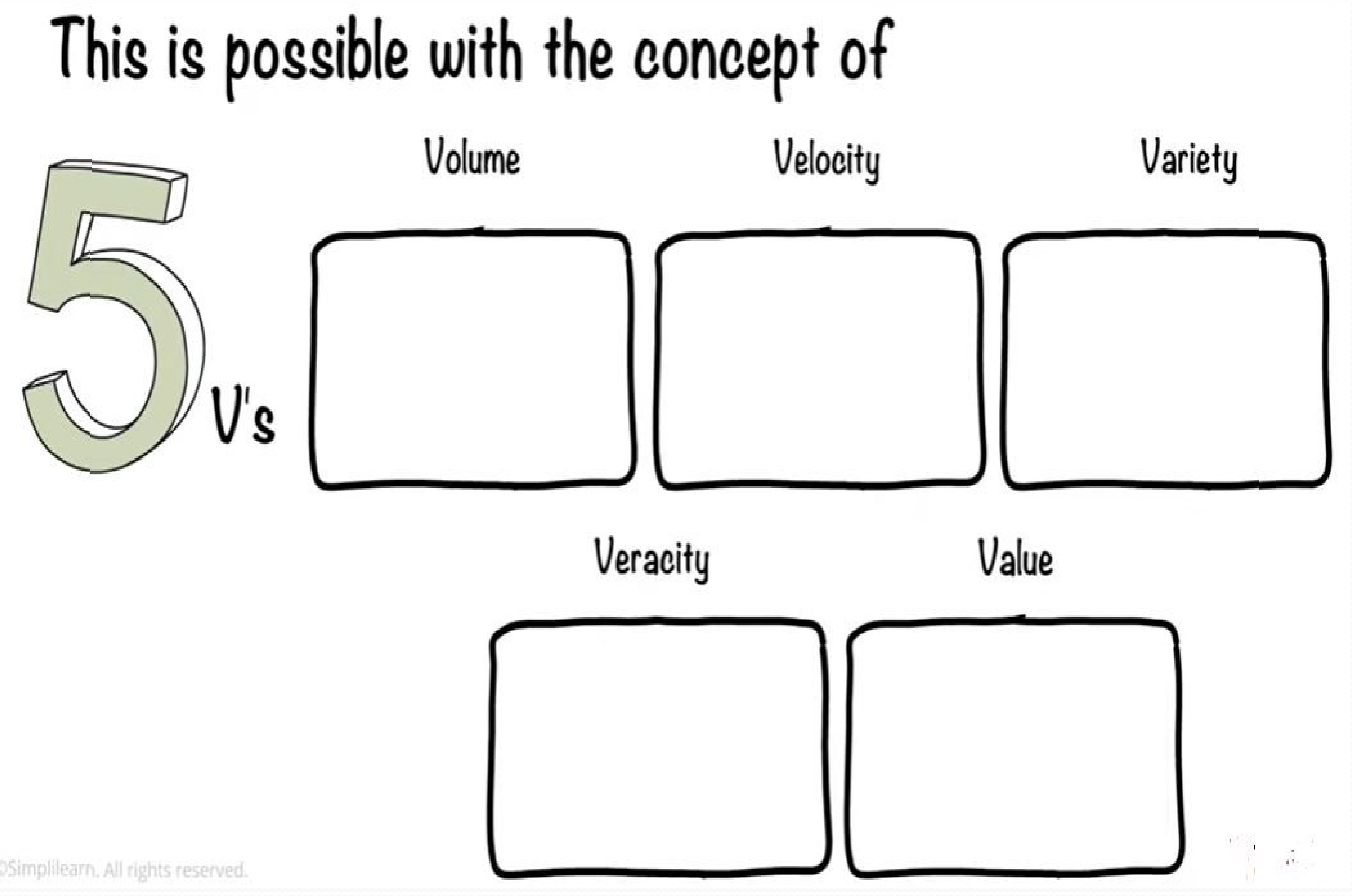
"3.8Million"

"1.0Million"

"4.5Million"

"188Million"

That's a lot of data



# This is possible with the concept of

5 V's

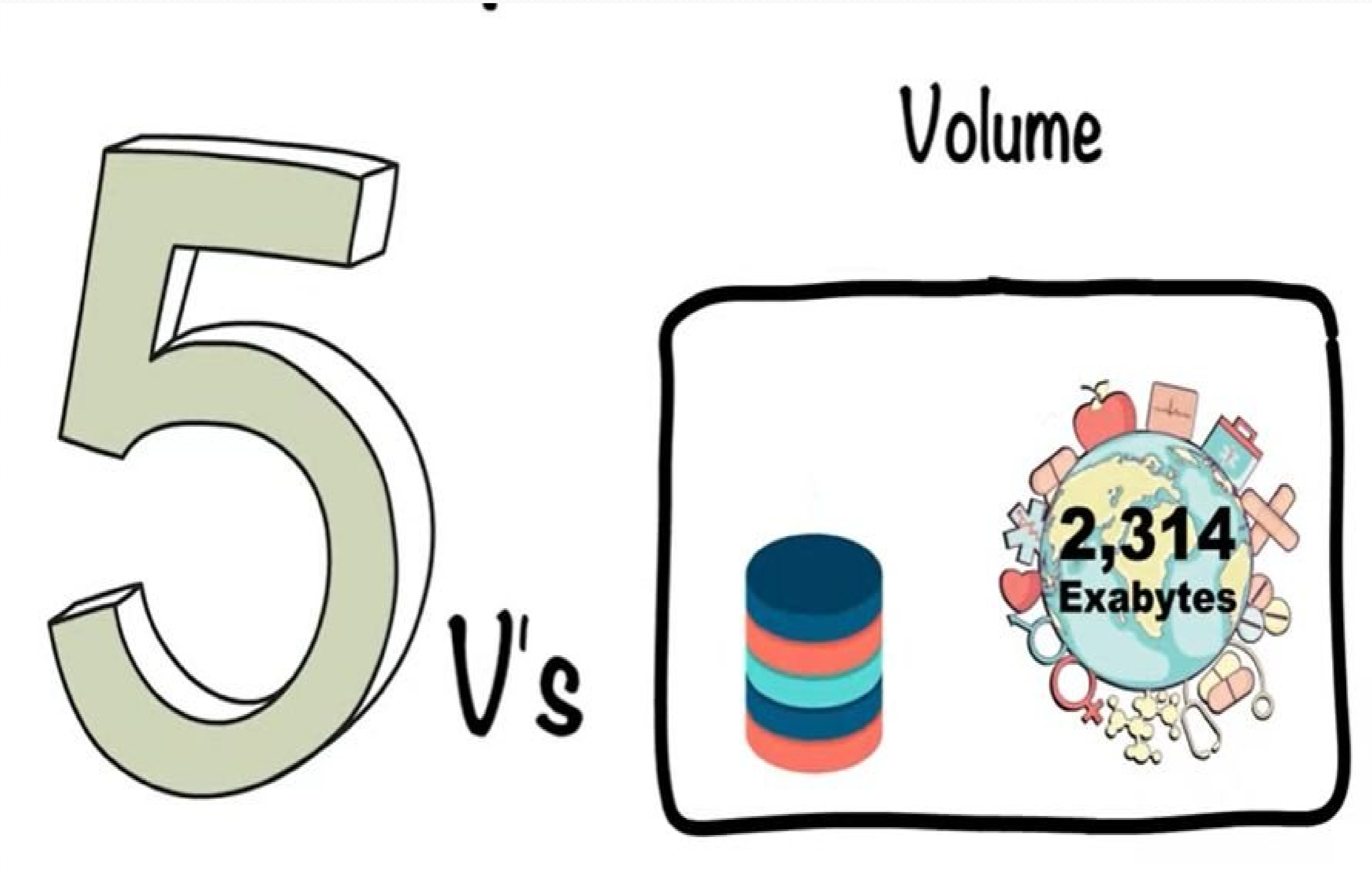
Volume

Velocity

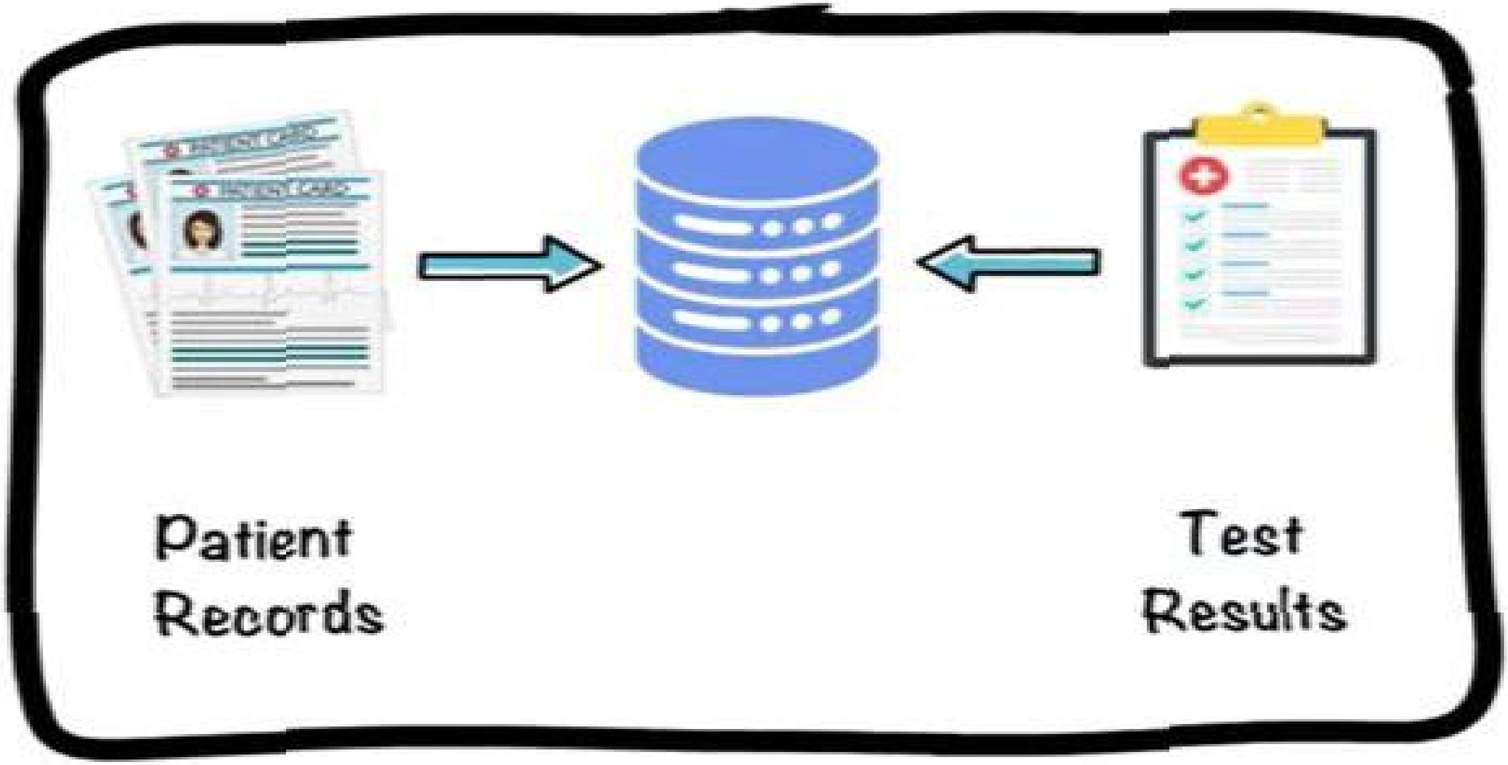
Variety

Veracity

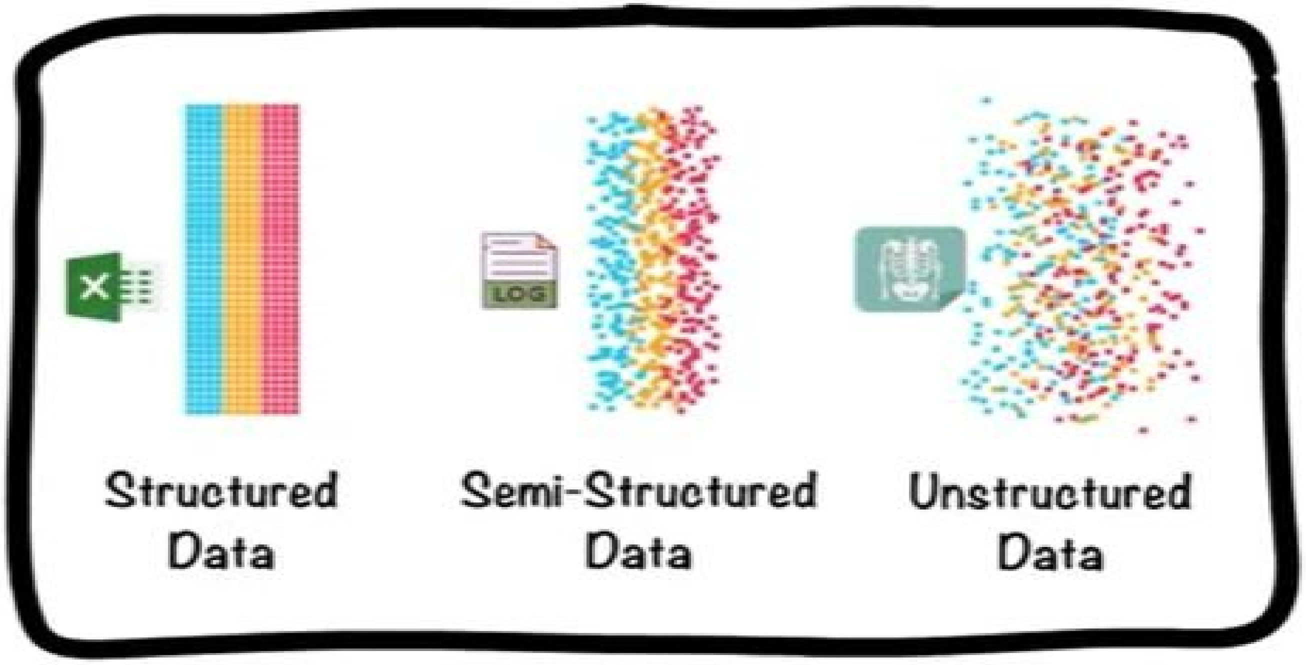
Value


5
V's
Volume
2,314
Exabytes

# Velocity



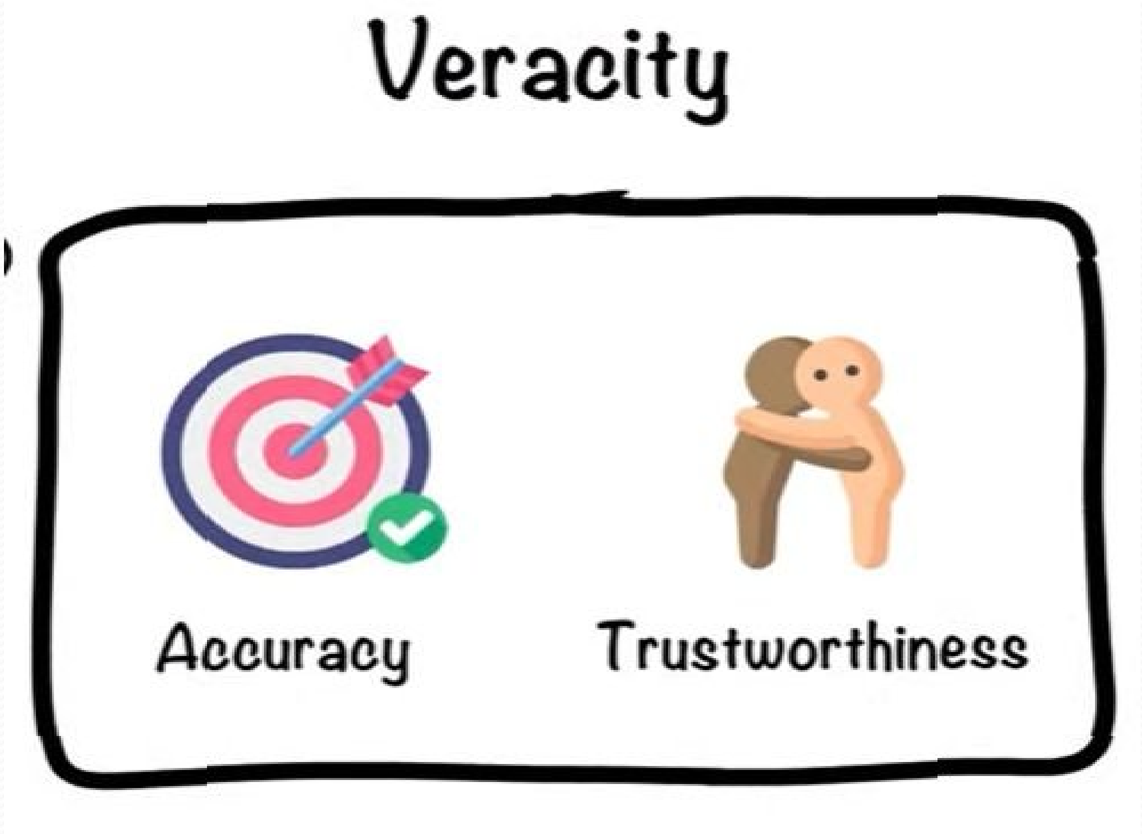
# Variety

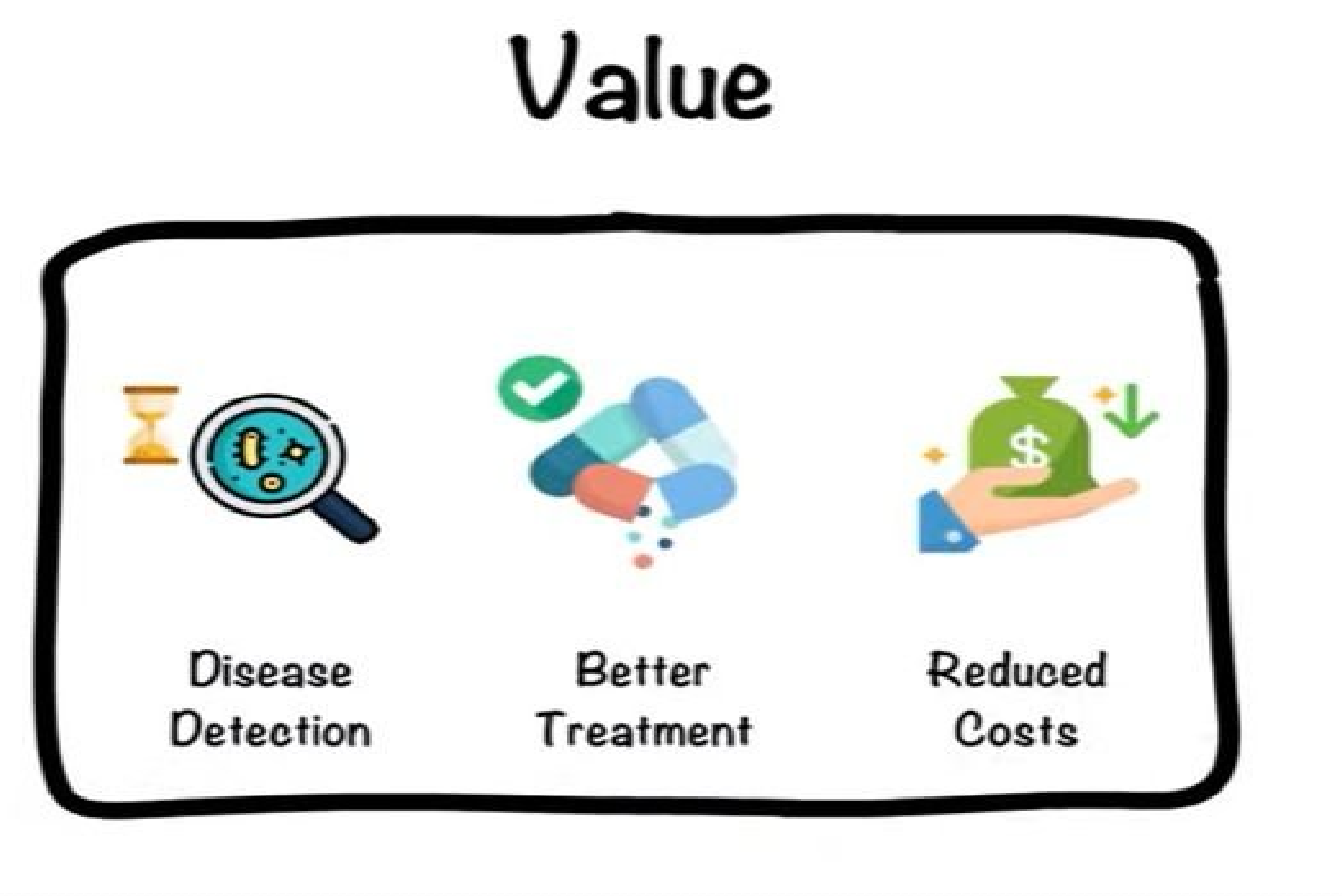


Structured Data

Semi-Structured Data

Unstructured Data


Veracity
Accuracy
Trustworthiness



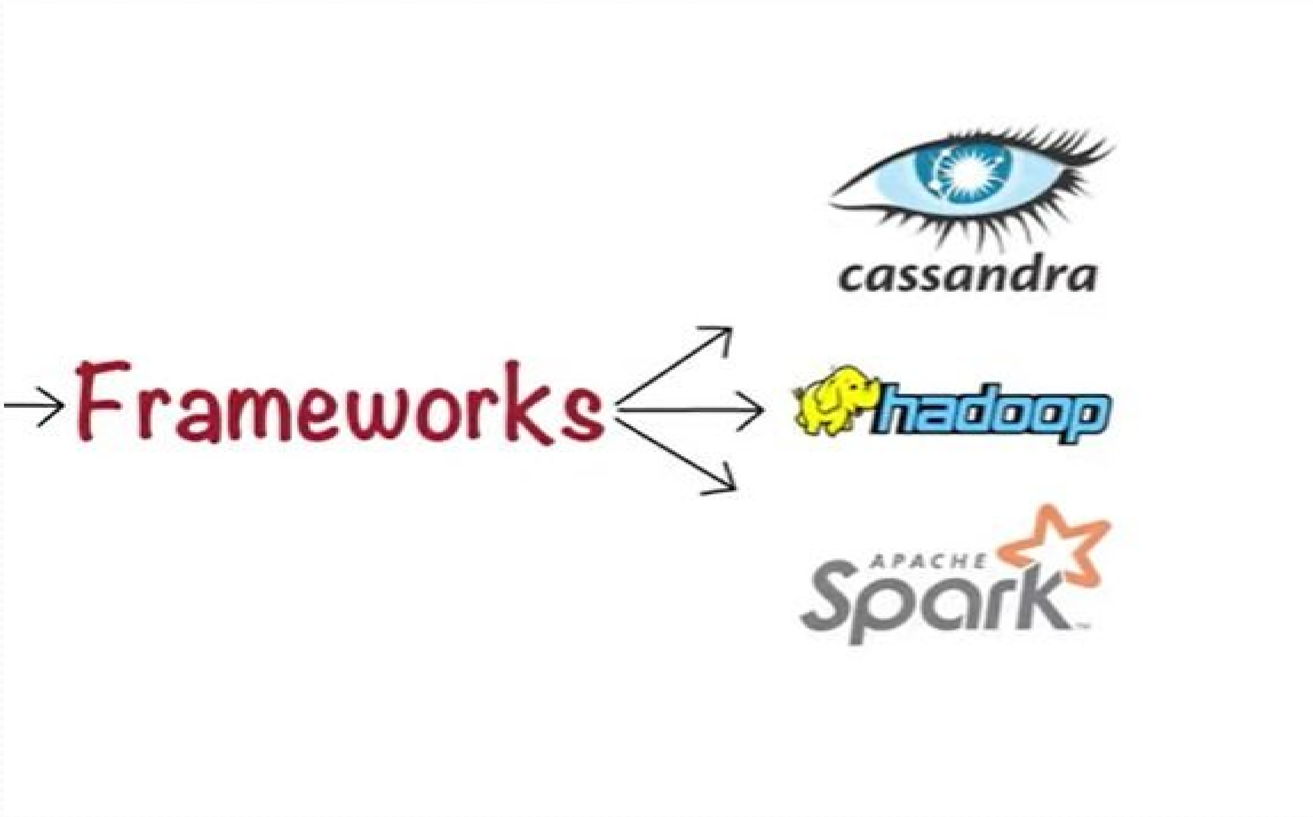
## Value

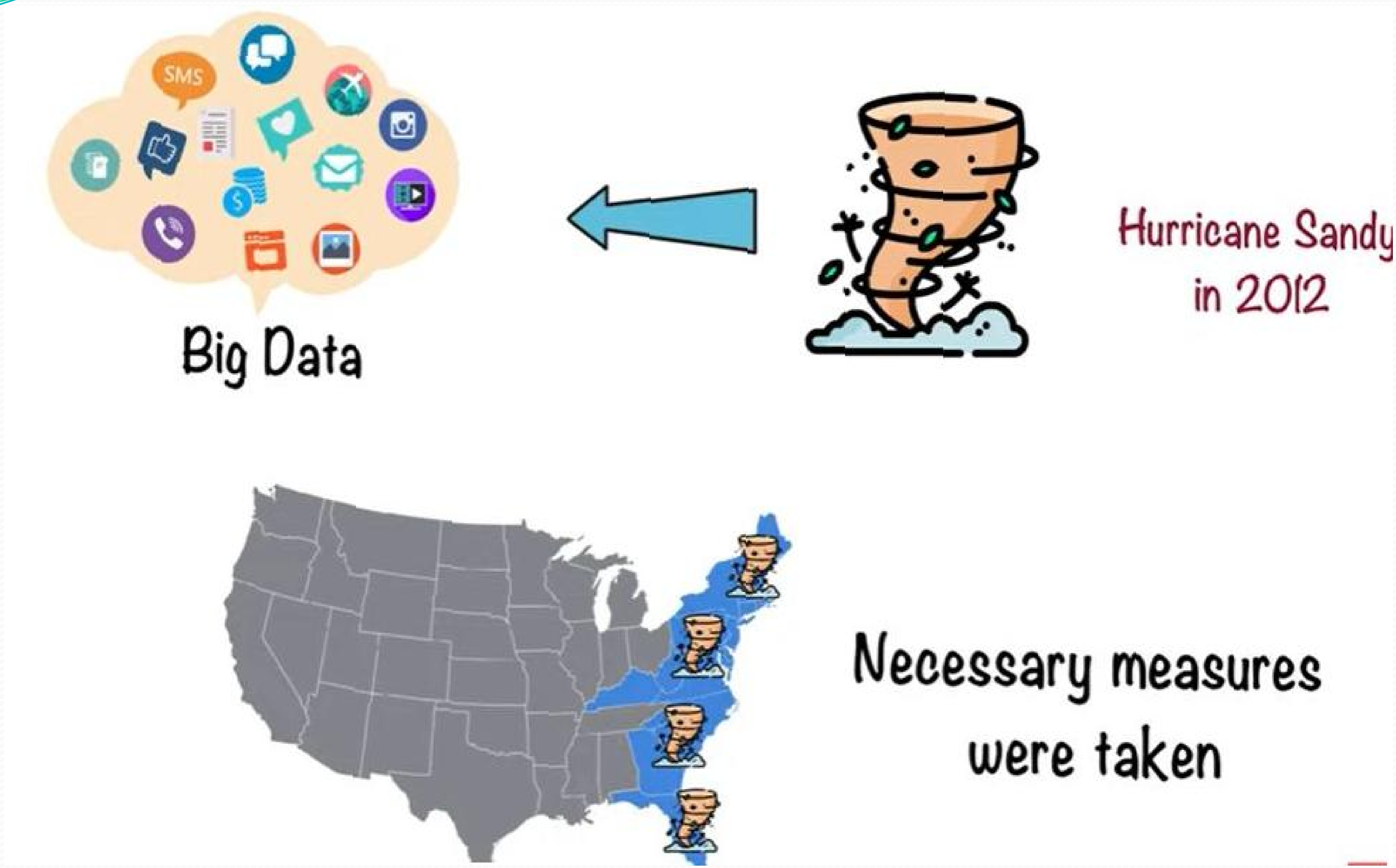
Disease Detection

Better Treatment

Reduced Costs

# But how do we store and process big data:





Big Data

Hurricane Sandy in 2012

Necessary measures were taken

It could predict the hurricane's landfall five days in advance which wasn't possible earlier.

Processed & Analyzed

# What’s Big Data?

**No single definition; here is from Wikipedia:**

- **Big data** is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.
- The challenges include capture, curation, 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, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions.”

## Data Structures

Big data can come in multiple forms, including structured and non-structured data such as financial data, text files, multimedia files, and genetic mappings. Contrary to much of the traditional data analysis performed by organizations, most of the Big Data is unstructured or semi-structured in nature, which requires different techniques and tools to process and analyze

**Four types of data structures**

- **Structured data:** Data containing a defined data type, format, and structure (that is, transaction data, online analytical processing [OLAP] data cubes, traditional  DBMS, CSV files, and even simple spreadsheets).

- **Semi-structured data:** Textual data files with a discernible pattern that enables parsing (such as Extensible Markup Language [XML] data files that are self-describing and defined by an XML schema).

- **Quasi-structured data:** Textual data with erratic data formats that can be formatted with effort, tools, and time (for instance, web click stream data that may contain inconsistencies in data values and formats).

- **Unstructured data:** Data that has no inherent structure, which may include text documents, PDFs, images, and video.

# BI Vs Data Science

# Data Science

- Data science is the science of extracting knowledge from data. In other words, it is a science of drawing out hidden patterns amongst data using statistical and mathematical techniques.
- It employs techniques and theories drawn from many fields from the broad areas of mathematics, statistics, information technology including machine learning, data engineering, probability models, statistical learning, pattern recognition and learning, etc.
- Data Scientist works on massive datasets for weather predictions, oil drillings, earthquake prediction, financial frauds, terrorist network and activities, global economic impacts, sensor logs, social media analytics, customer churn, collaborative filtering(prediction about interest on users), regression analysis, etc. Data science is multi-disciplinary.

# State of the Practice in Analytics

Current business problems provide many opportunities for organizations to become more analytical and data driven

# BI Vs Data Science

1. **Perspective:** BI systems are designed to look backwards based on real data from real events. Data Science looks forward, interpreting the information to predict what might happen in the future.
2. **Focus:** BI delivers detailed reports, KPIs and trends but it doesn't tell you what this data may look like in the future in the form of patterns and experimentation.
3. **Process:** Traditional BI systems tend to be static and comparative. They do not offer room for exploration and experimentation in terms of how the data is collected and managed.
4. **Data sources:** Because of its static nature, BI data sources tend to be pre-planned and added slowly. Data science offers a much more flexible approach as it means data sources can be added on the go as needed.
5. **Transform:** How the data delivers a difference to the business is key also. BI helps you answer the questions you know, whereas Data Science helps you to discover new questions because of the way it encourages companies to apply insights to new data.
6. **Storage:** Like any business asset, data needs to be flexible. BI systems tend to be warehoused and silced, which means it is difficult to deploy across the business. Data Science can be distributed real time.

7. **Data quality: Any** data analysis is only as good as the quality of the data captured. BI provides a single version of truth while data science offers precision, confidence level and much wider probabilities with its findings.

8. **IT owned vs. business owned**

   In the past, BI systems were often owned and operated by the IT department, sending along intelligence to analysts who interpreted it. With Data Science, the analysts are in charge. The new Big Data solutions are designed to be owned by analysts, who spend little of their time on 'IT housekeeping' and most of their time analyzing data and making predictions upon which to base business decisions.

9. **Analysis:** A retrospective and prescriptive BI system is much less likely to be placed to do this than a Predictive Data Science programme.

**Small Assignment:**

**Descriptive :** What has happened**?”**
total stock in inventory, average dollars spent per customer and Year over year change in sales

**Predictive :** “What could happen?”
**??**
**Prescriptive:** “What should we do?”
??

## Current Analytical Architecture

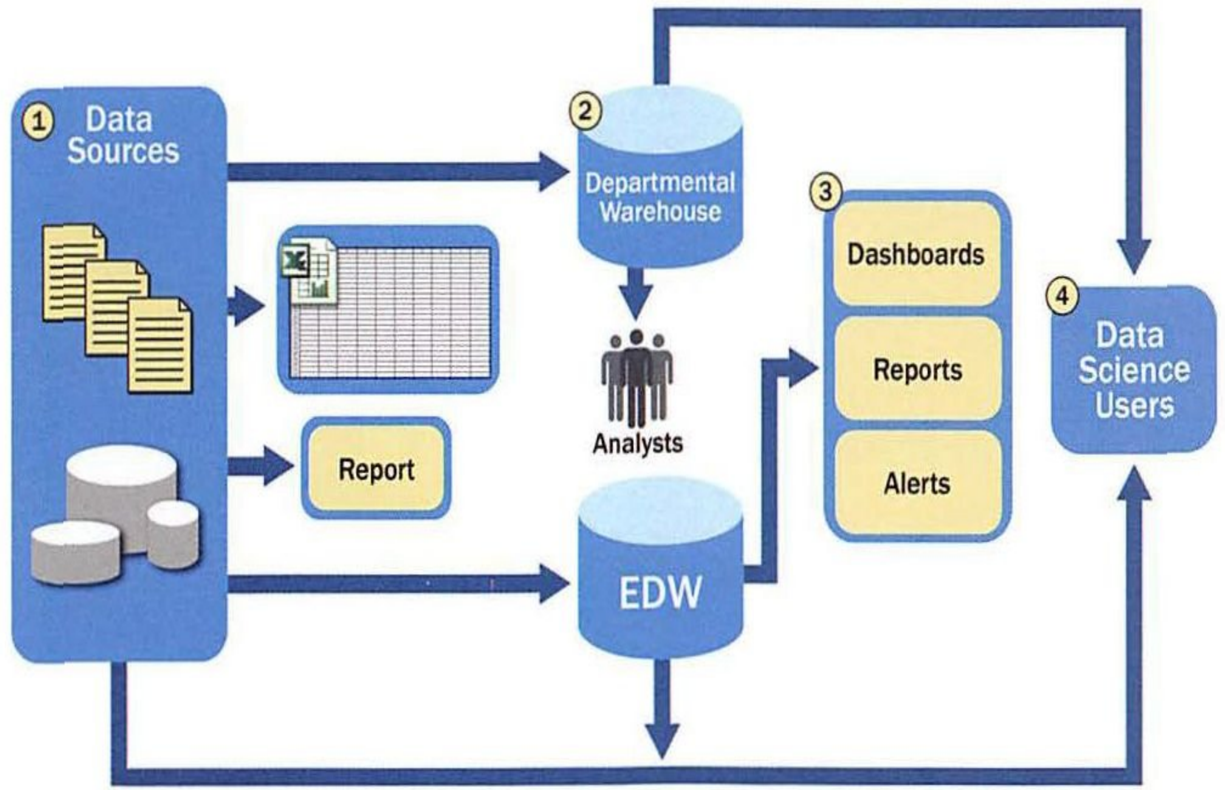


FIGURE 1-9 *Typical analytic architecture*

**1**. For data sources to be loaded into the data warehouse, data needs to be well understood, structured, and normalized with the appropriate data type definitions. Although this kind of centralization enables security, backup, and failover of highly critical data, it also means that data typically must go through significant preprocessing and checkpoints before it can enter this sort of controlled environment, which does not lend itself to data exploration and iterative analytics.

2. As a result of this level of control on the EDW, additional local systems may emerge in the form of departmental warehouses and local data marts that business users create to accommodate their need for flexible analysis. These local data marts may not have the same constraints for security and structure as the main EDW and allow users to do some level of more in-depth analysis. However, these one-off systems reside in isolation, often are not synchronized or integrated with other data stores, and may not be backed up.

3. Once in the data warehouse, data is read by additional applications across the enterprise for BI and reporting purposes. These are high-priority operational processes getting critical data feeds from the data warehouses and repositories.

4. At the end of this workflow, analysts get data provisioned for their downstream analytics. Because users generally are not allowed to run custom or intensive analytics on production databases, analysts create data extracts from the EDW to analyze data offline in R or other local analytical tools. Many times these tools are limited to in-memory analytics on desktops analyzing samples of data, rather than the entire population of a dataset. Because these analyses are based on data extracts, they reside in a separate location, and the results of the analysis-and any insights on the quality of the data or anomalies-rarely are fed back into the main data repository.

## Drivers (Sources) of Big Data

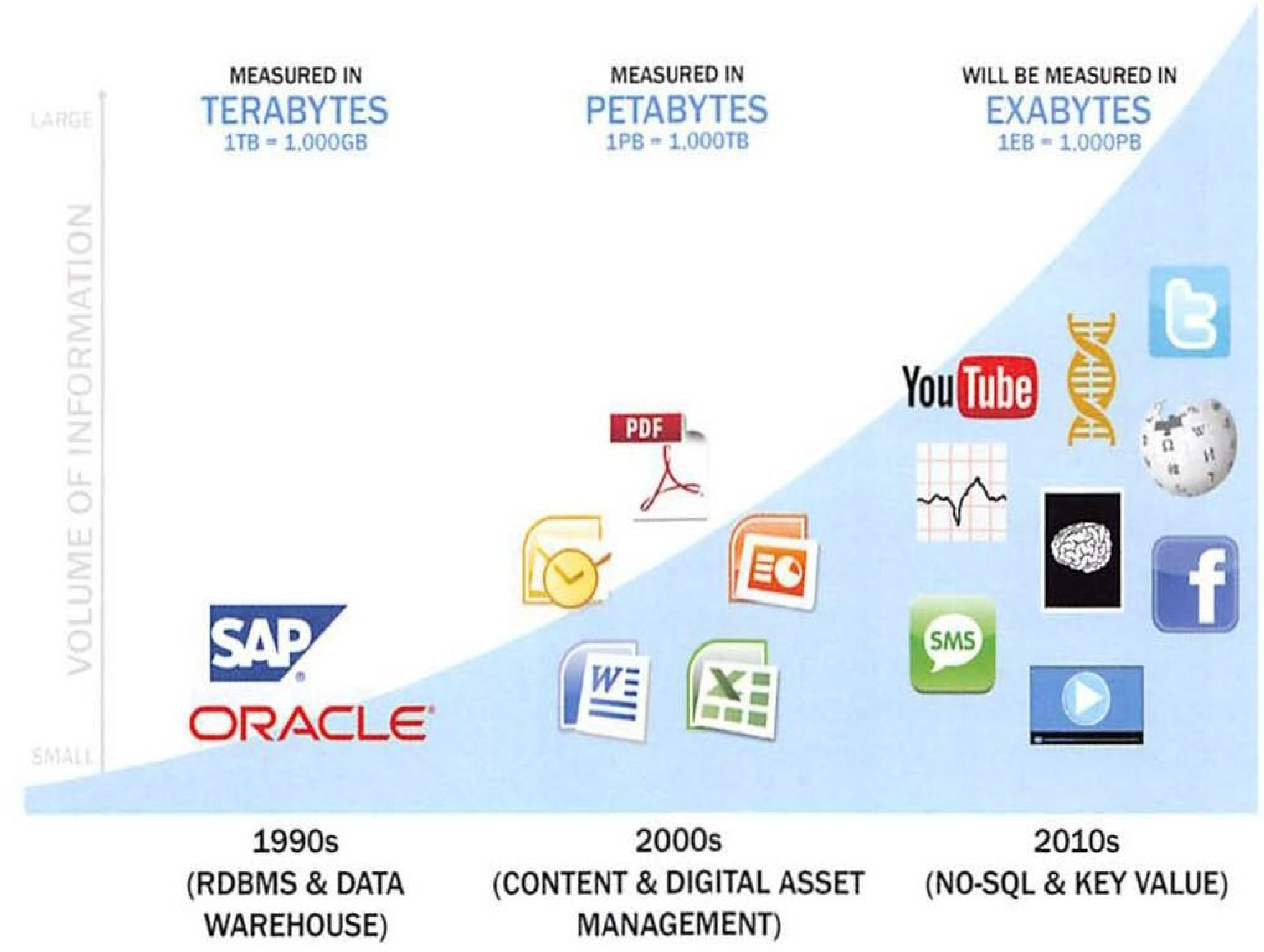


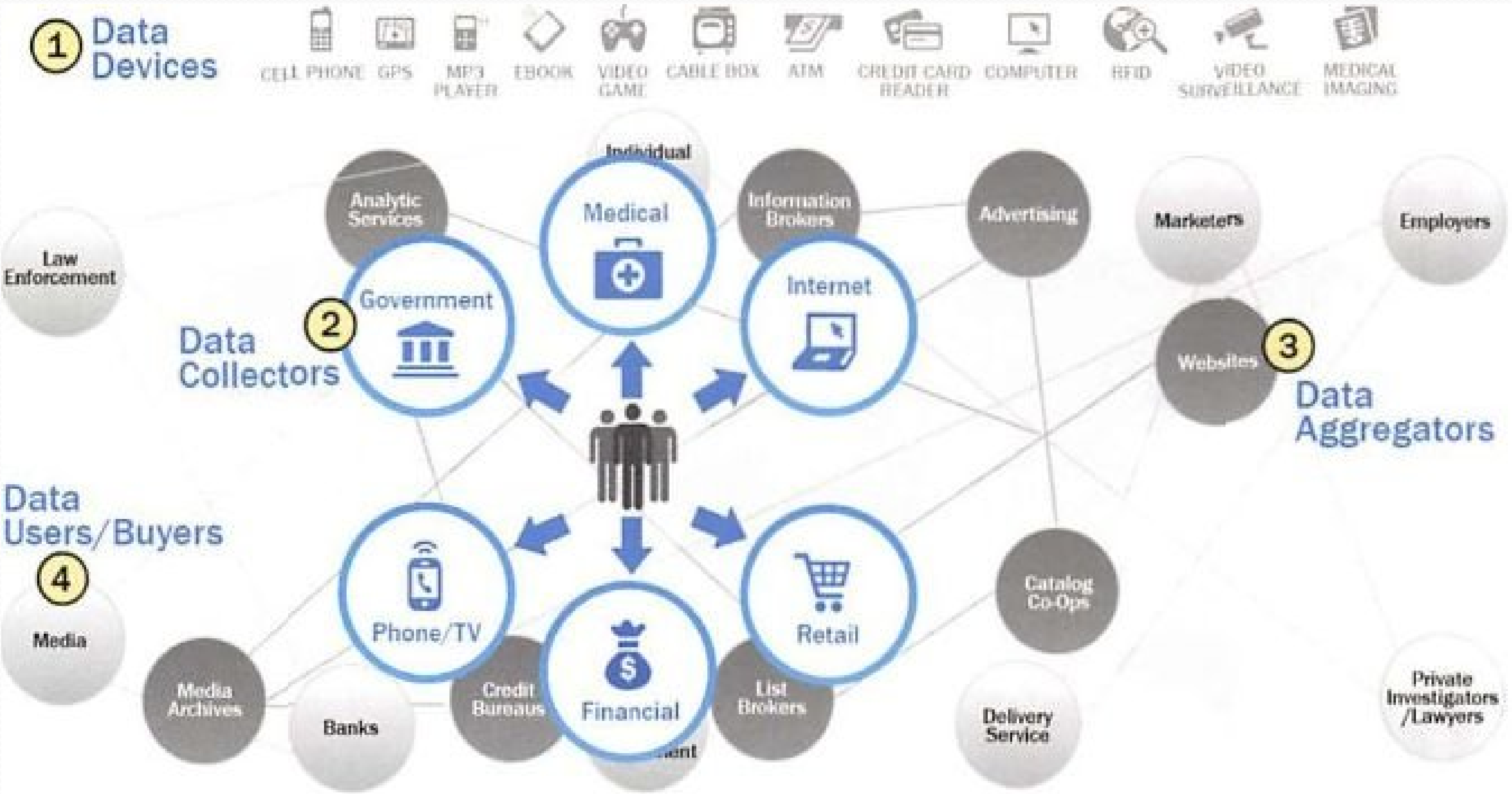
FIGURE 1-10 *Data evolution and the rise of Big Data sources*

**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
- Nontraditional IT devices, including the use of radio-frequency identification (RFID) readers, GPS navigation systems, and seismic processing

## Emerging Big Data Ecosystem and a New Approach to Analytics

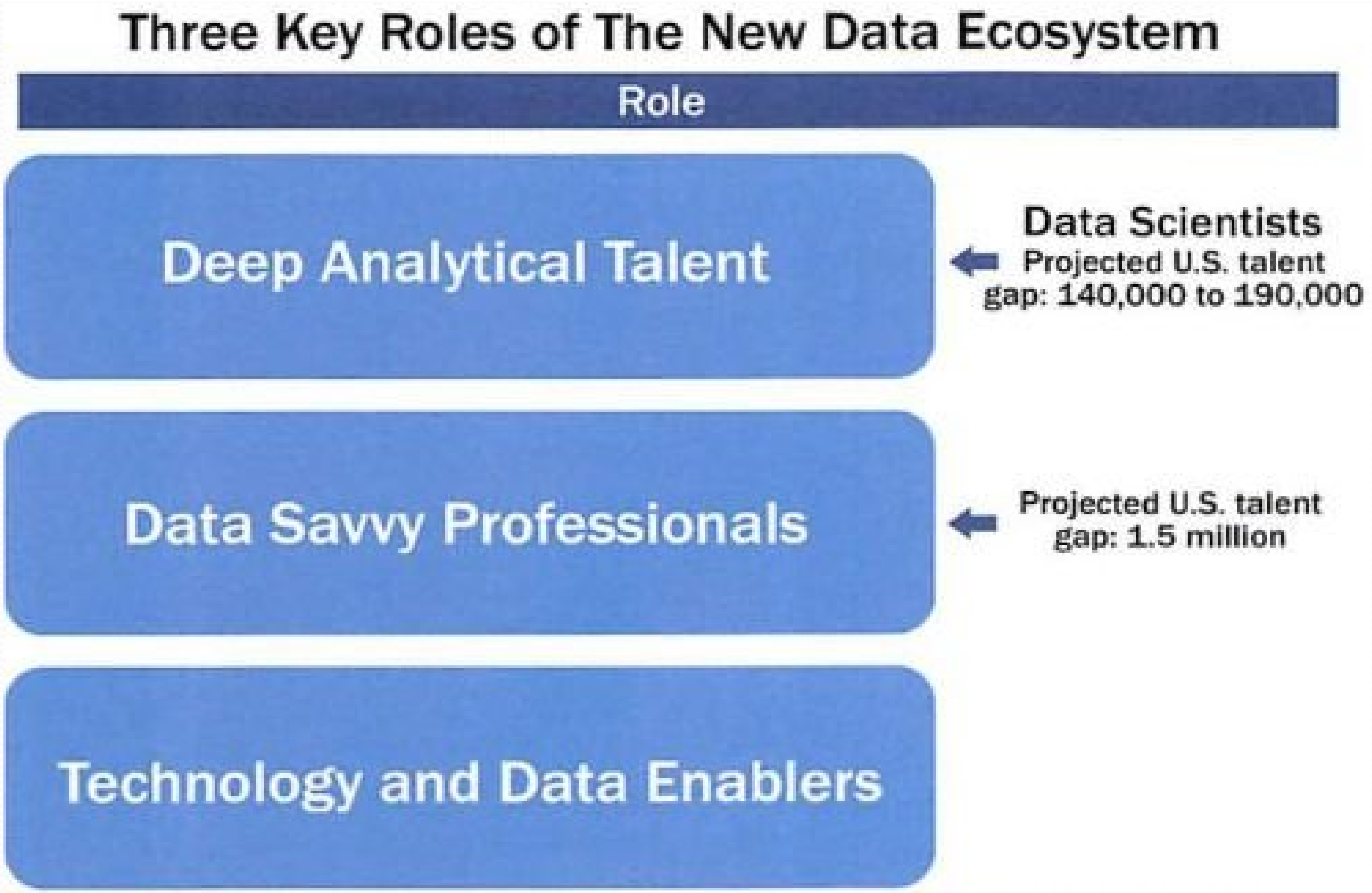
- Four main groups of players
  - Data devices
    - Games, smartphones, computers, etc.
  - Data collectors
    - Phone and TV companies, Internet, Gov’t, etc.
  - Data aggregators – make sense of data
    - Websites, credit bureaus, media archives, etc.
  - Data users and buyers
    - Banks, law enforcement, marketers, employers, etc.


1
Data Devices
CELL PHONE
GPS
MP3 PLAYER
EBOOK
VIDEO GAME
CABLE BOX
ATM
CREDIT CARD READER
COMPUTER
RFID
VIDEO SURVEILLANCE
MEDICAL IMAGING
Individual
Analytic Services
Information Brokers
Advertising
Marketers
Employers
Law Enforcement
Medical
Government
Internet
2
Data Collectors
Websites
3
Data Aggregators
Data Users/Buyers
4
Media
Phone/TV
Financial
Retail
Catalog Co-Ops
Media Archives
Banks
Credit Bureaus
List Brokers
Delivery Service
Private Investigators /Lawyers

# Key Roles for the New Big Data Ecosystem

1. Deep analytical talent
   - Advanced training in quantitative disciplines – e.g., math, statistics, machine learning
2. Data savvy professionals
   - Savvy but less technical than group 1
3. Technology and data enablers
   - Support people – e.g., DB admins, programmers, etc.

# Three Key Roles of the New Big Data Ecosystem



Three Key Roles of The New Data Ecosystem

Role

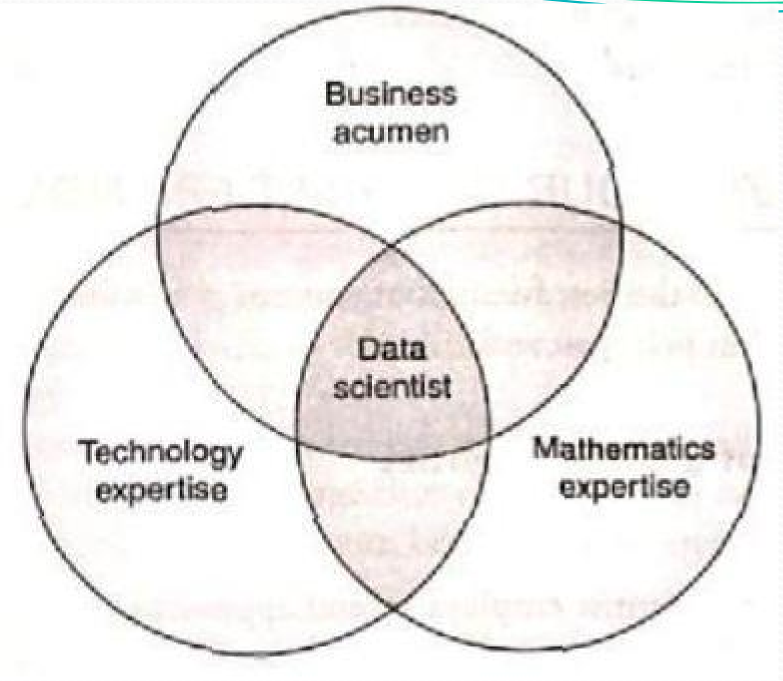
Deep Analytical Talent

Data Scientists
Projected U.S. talent
gap: 140,000 to 190,000

Data Savvy Professionals

Projected U.S. talent
gap: 1.5 million

Technology and Data Enablers



A data scientist should have following ability to play the role of data scientist

Understanding of domain
Business strategy
Problem solving
Communication
Presentation
Keenness

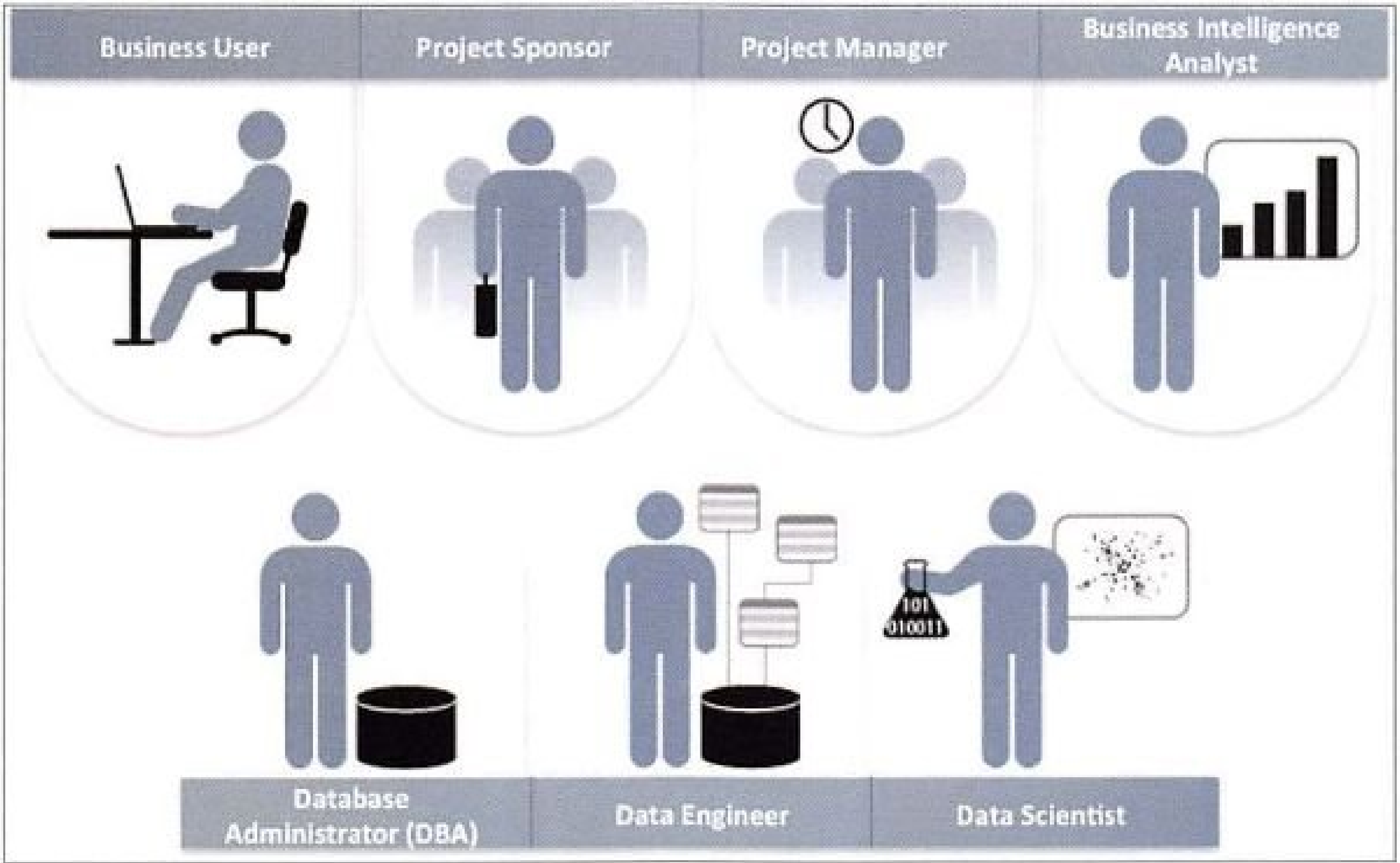
# Data Analytics Lifecycle

- 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 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 modeling

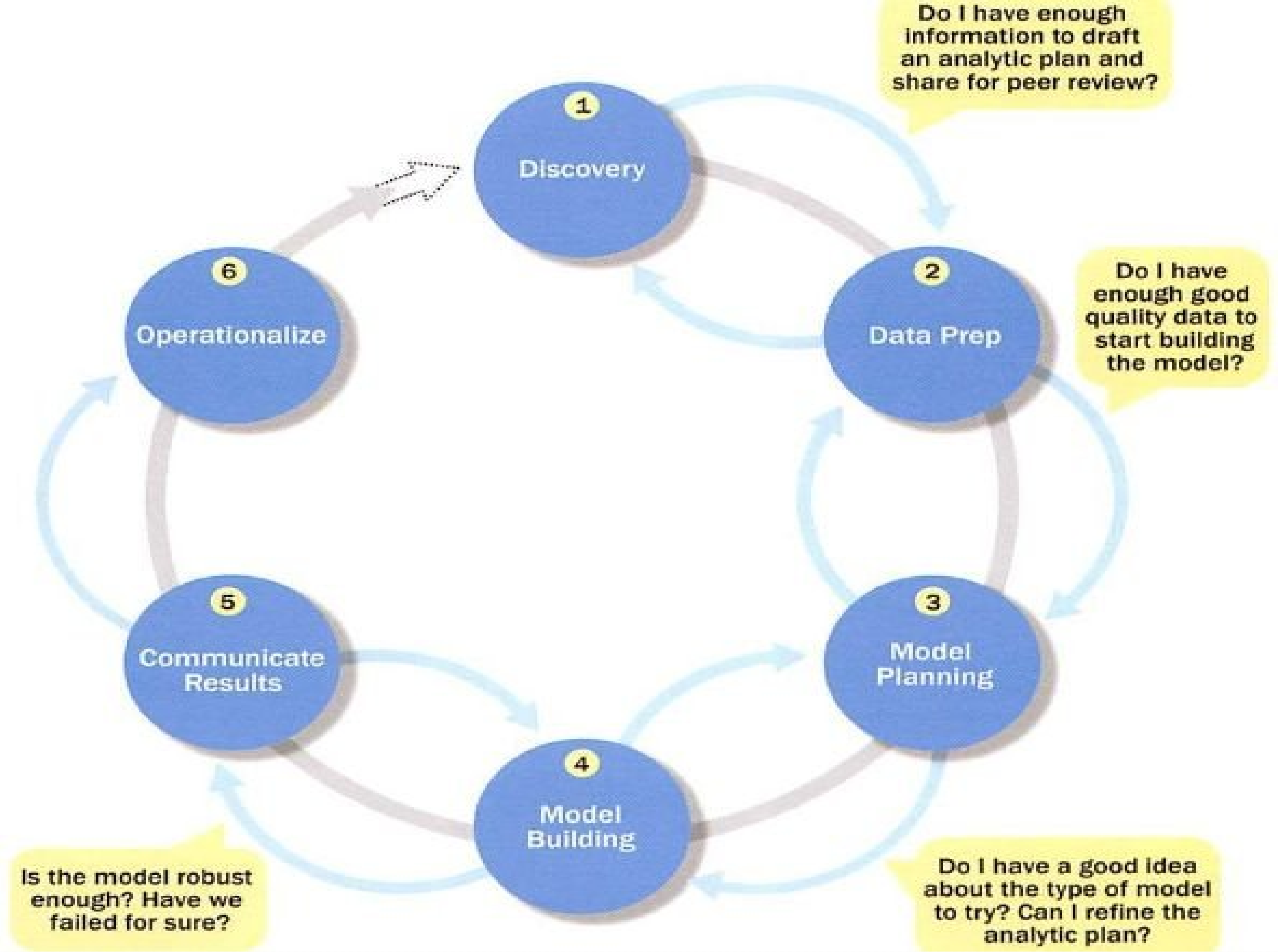
# Background and Overview of Data Analytics Lifecycle

- 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.

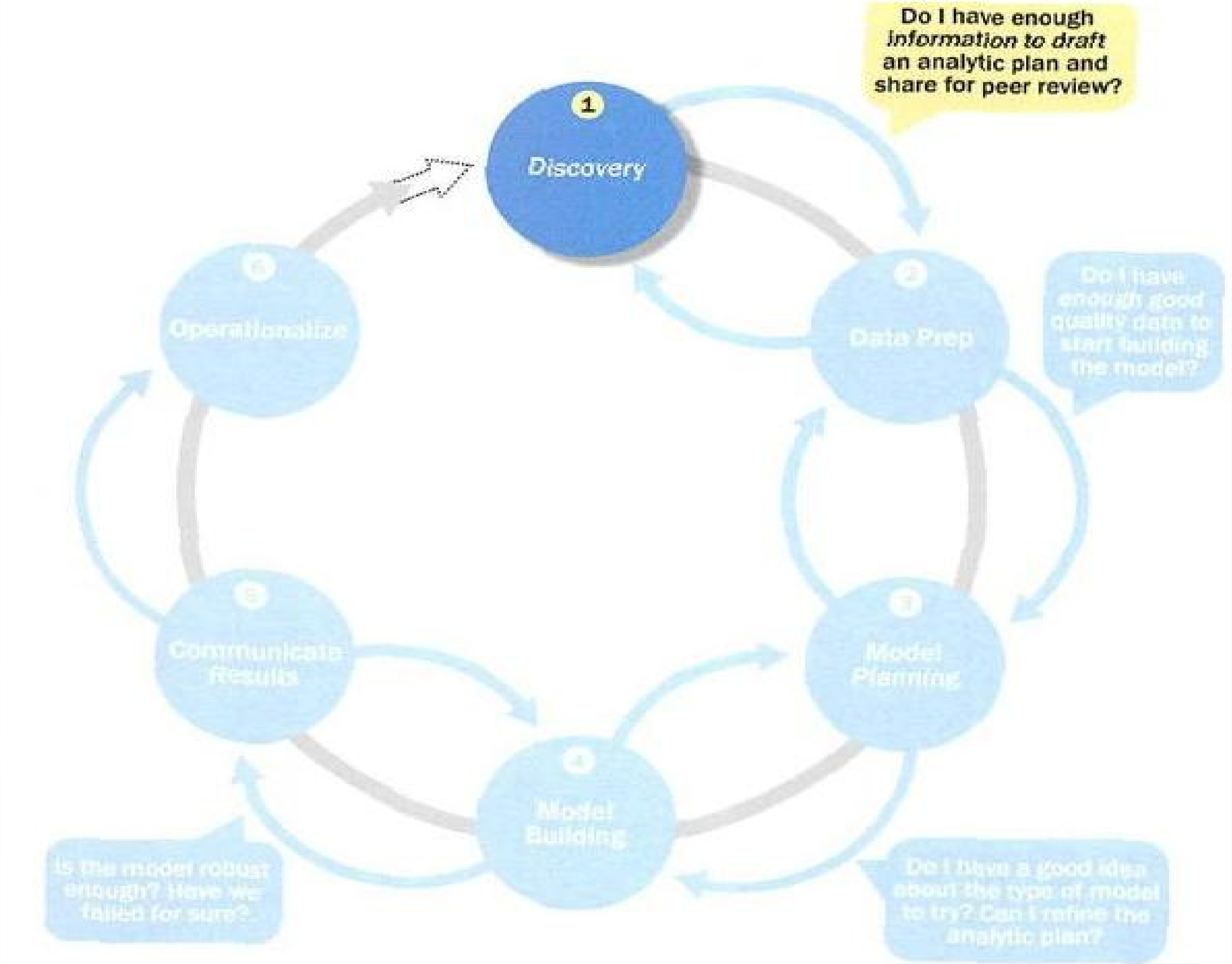
# Data Analytics Lifecycle

- 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

# 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?
6
Operationalize
3
Model Planning
5
Communicate Results
4
Model Building
Is the model robust enough? Have we failed for sure?
Do I have a good idea about the type of model to try? Can I refine the analytic plan?

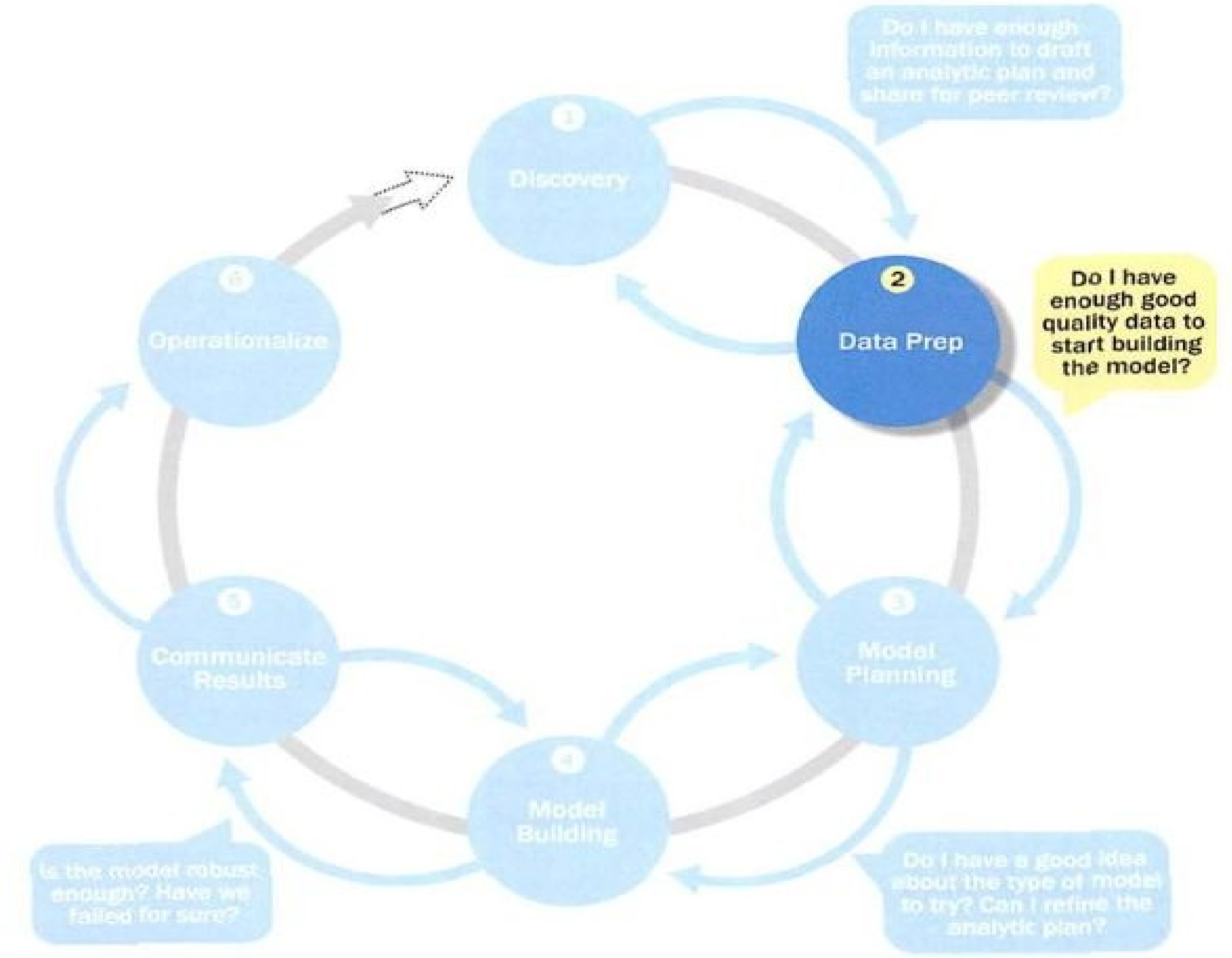
# Phase 1: Discovery


1
Discovery
Do I have enough information to draft an analytic plan and share for peer review?
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

# 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


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

# 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
- Ex: The IBM Netezza 1000 a data sandbox platform, IBM InfoSphere BigInsights Enterprise Edition

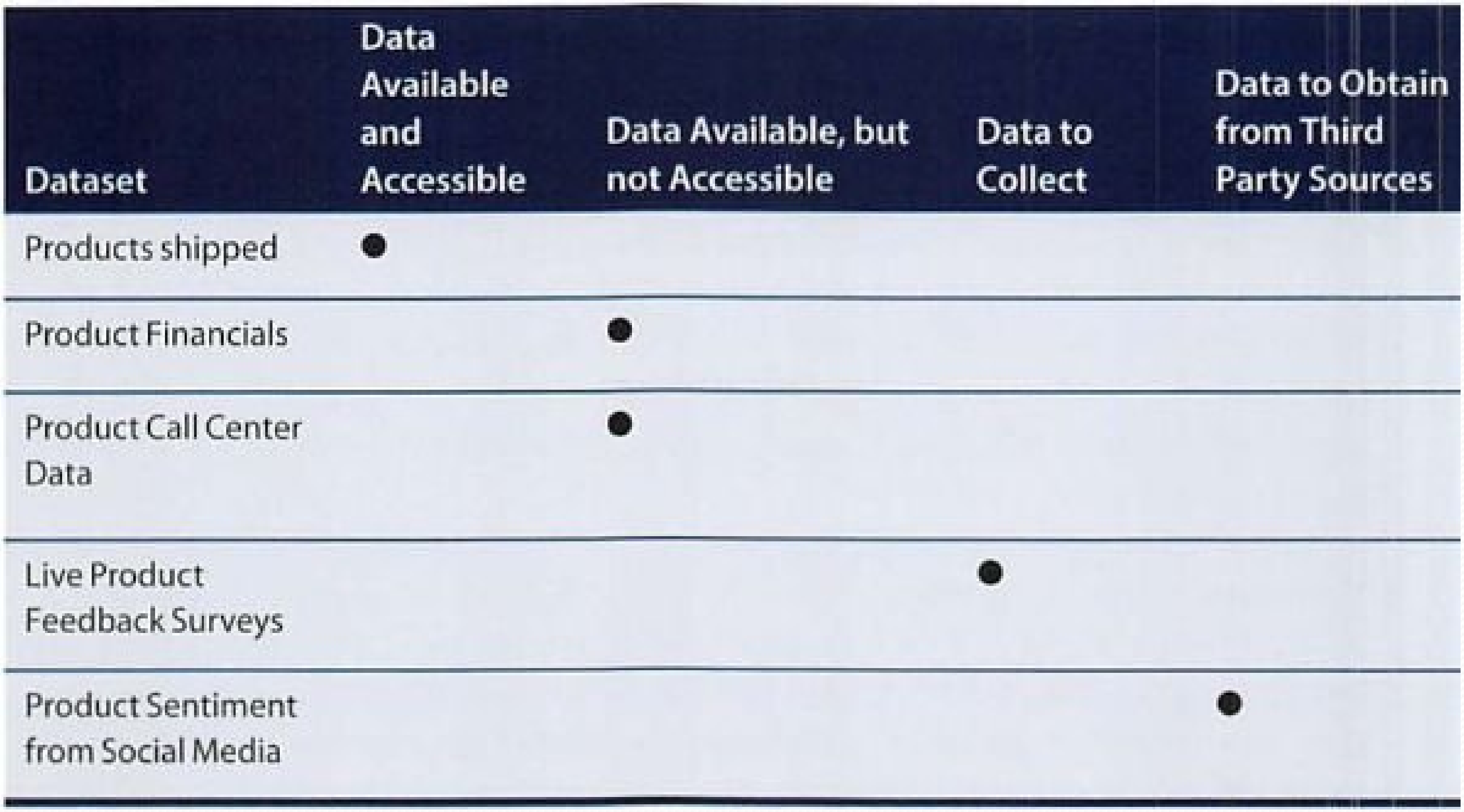
# Performing ETLT
# (Extract, Transform, Load, Transform)

- 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 (Chapter 10) 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

IBM InfoSphere BigInsights Enterprise Edition

- 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

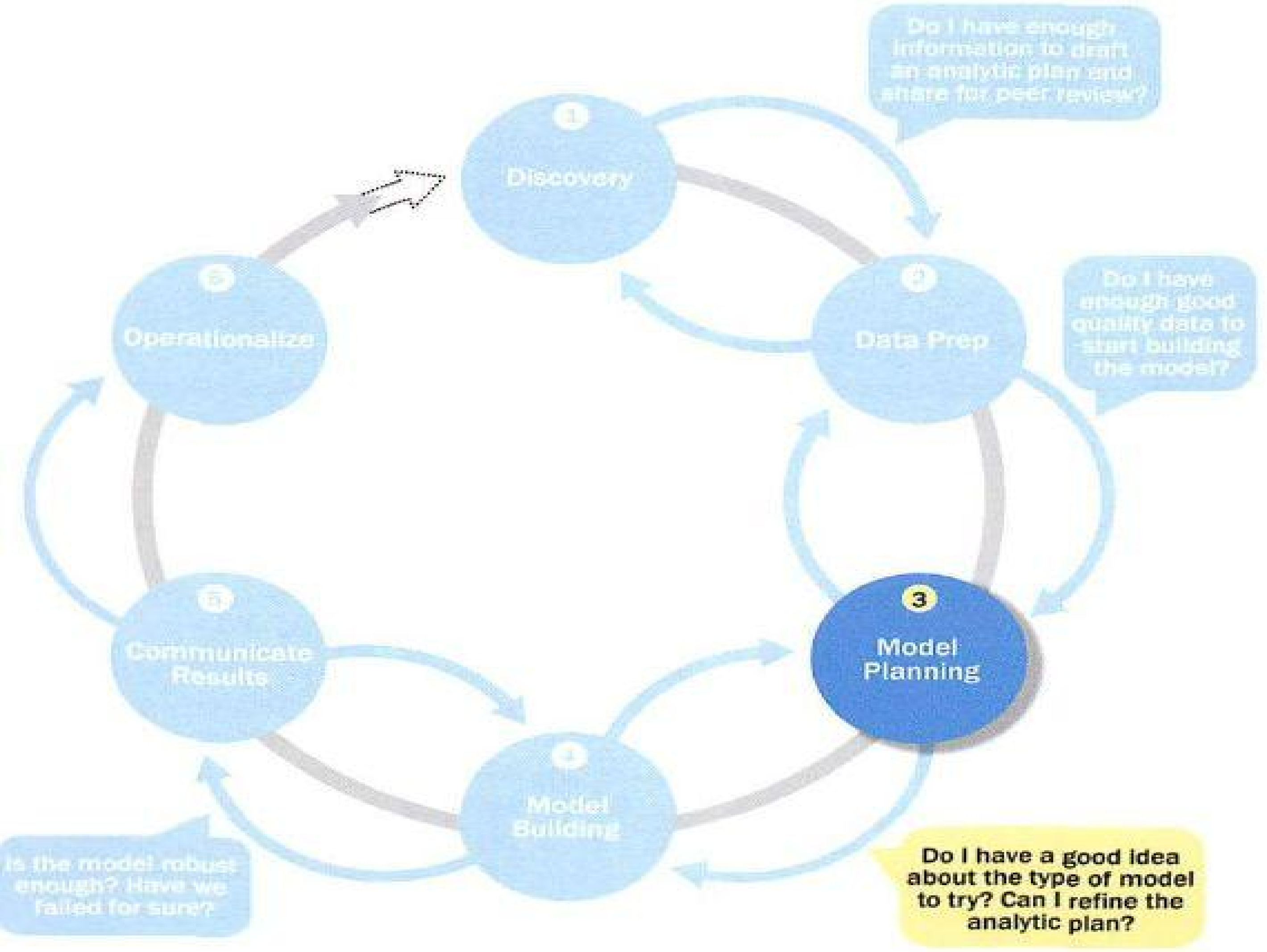
# Survey and Visualize Guidelines and Considerations

- 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 an transformation

# 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

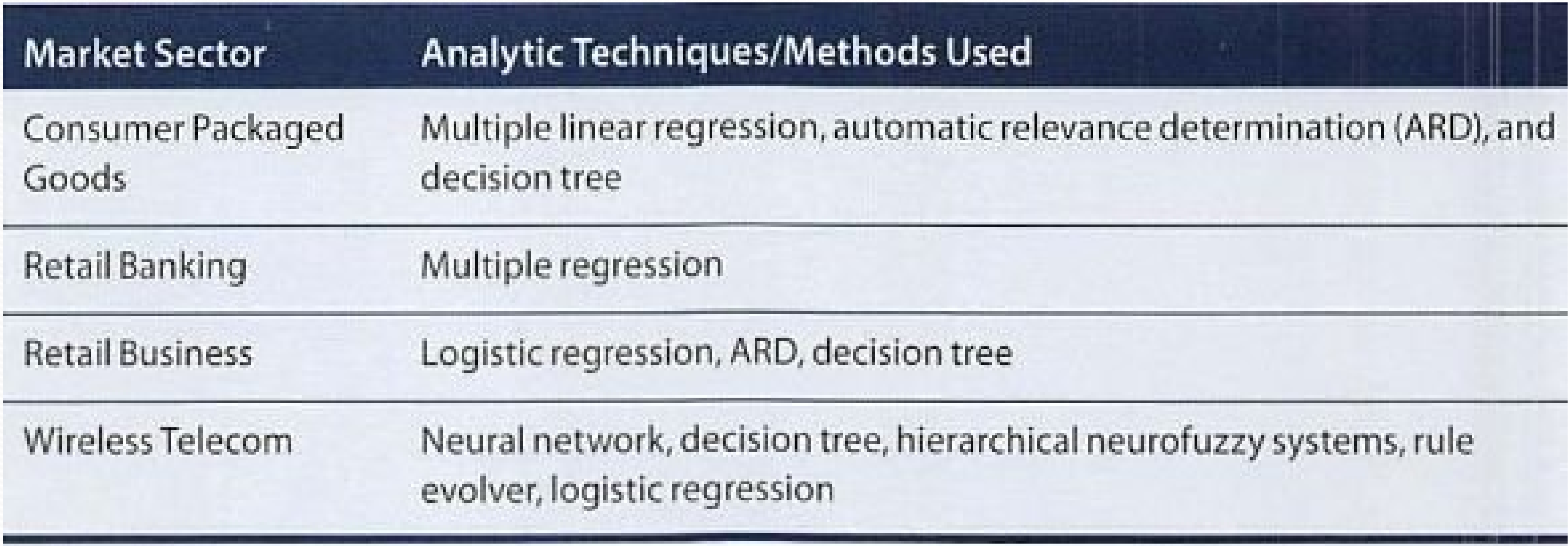
# Phase 3: Model Planning

- 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

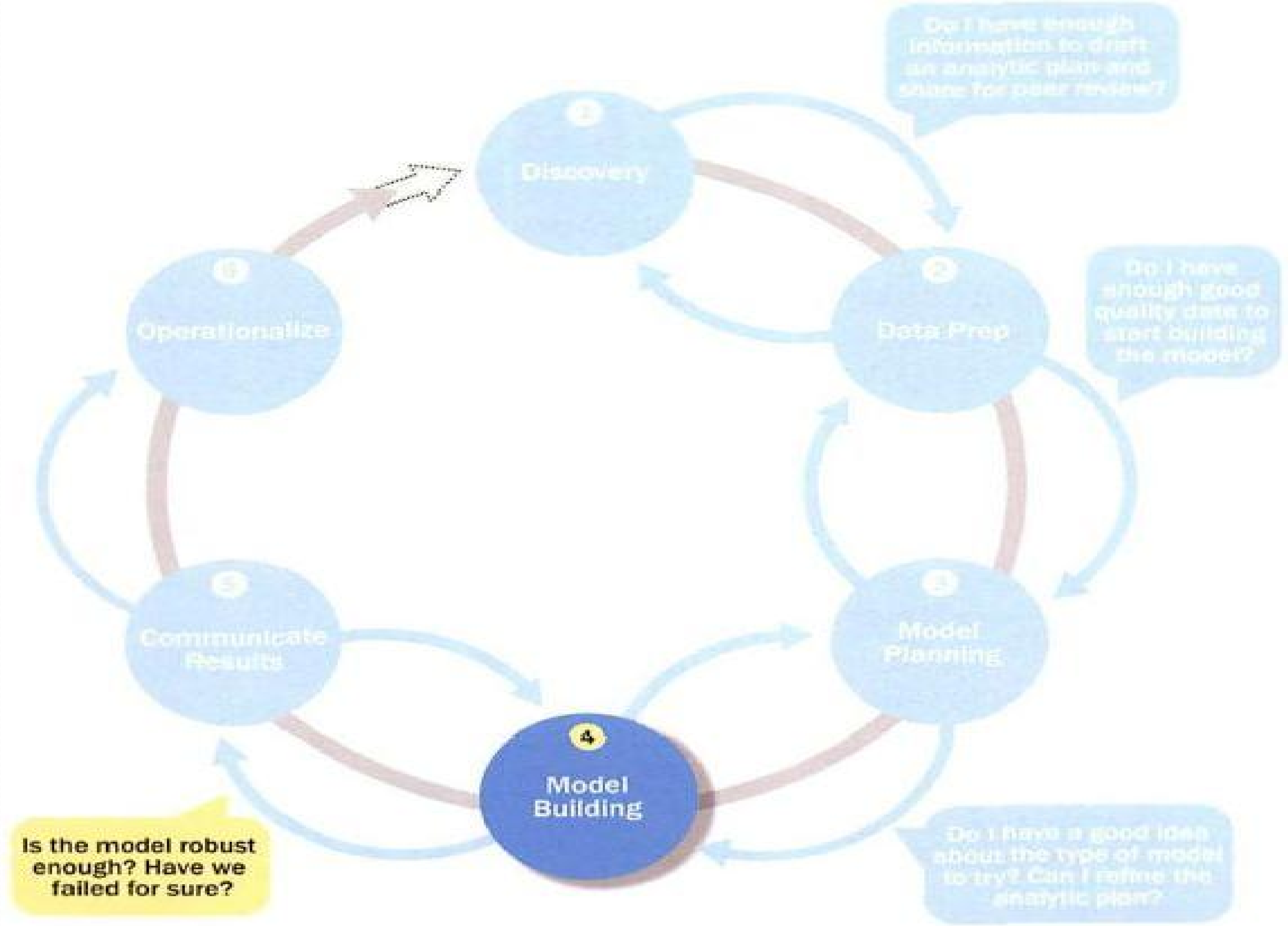
# Model Selection

- 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 modeling 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



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?
6
Operationalize
3
Model Planning
5
Communicate Results
4
Model Building
Is the model robust enough? Have we failed for sure?
Do I have a good idea about the type of model to try? Can I refine the analytic plan?

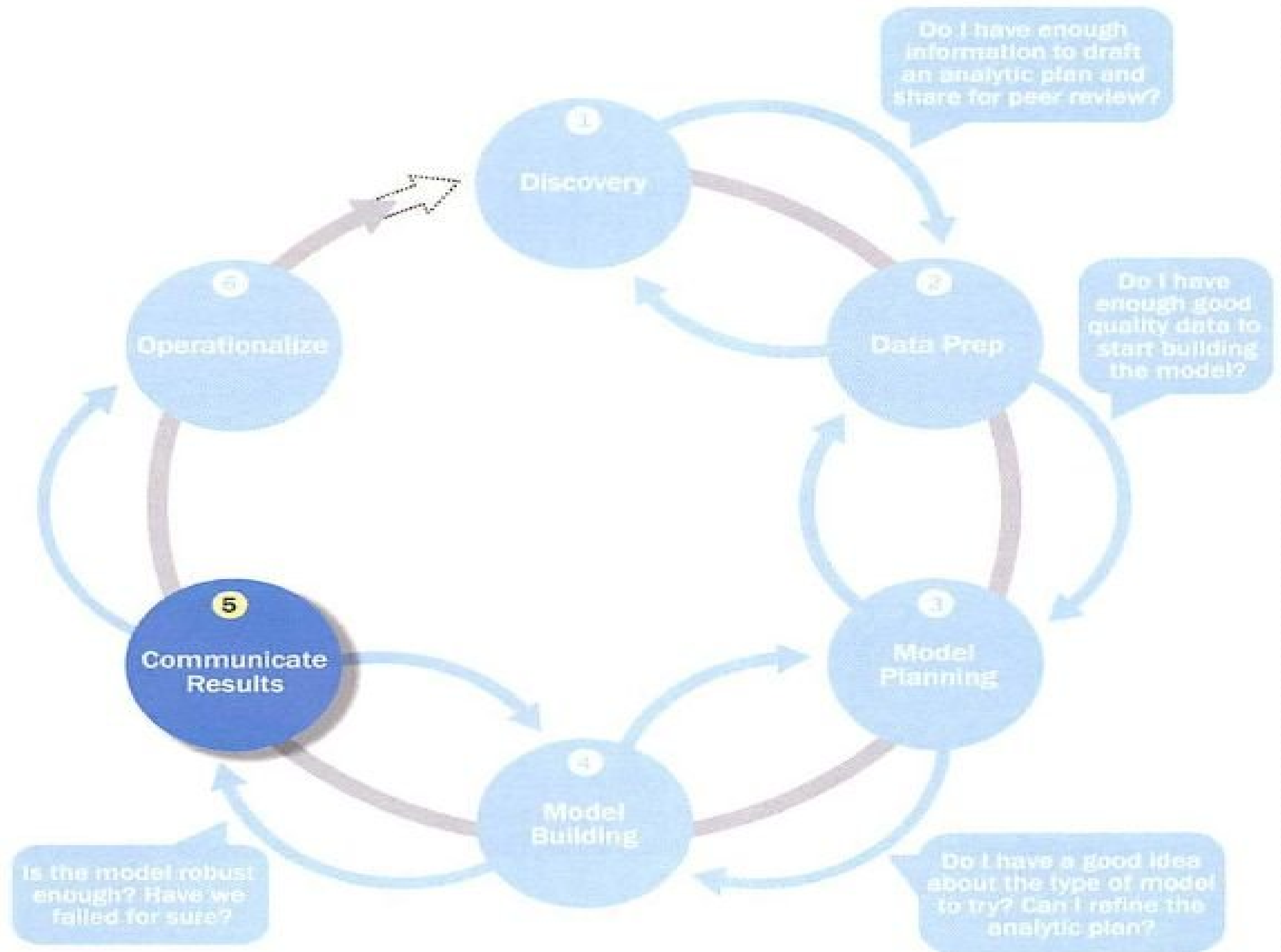
# Phase 4: Model Building

- 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?
  - Does the model avoid intolerable mistakes?  (see Chapters 3 and 7)
  - 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 (see Chap 11)

# Phase 5: Communicate Results

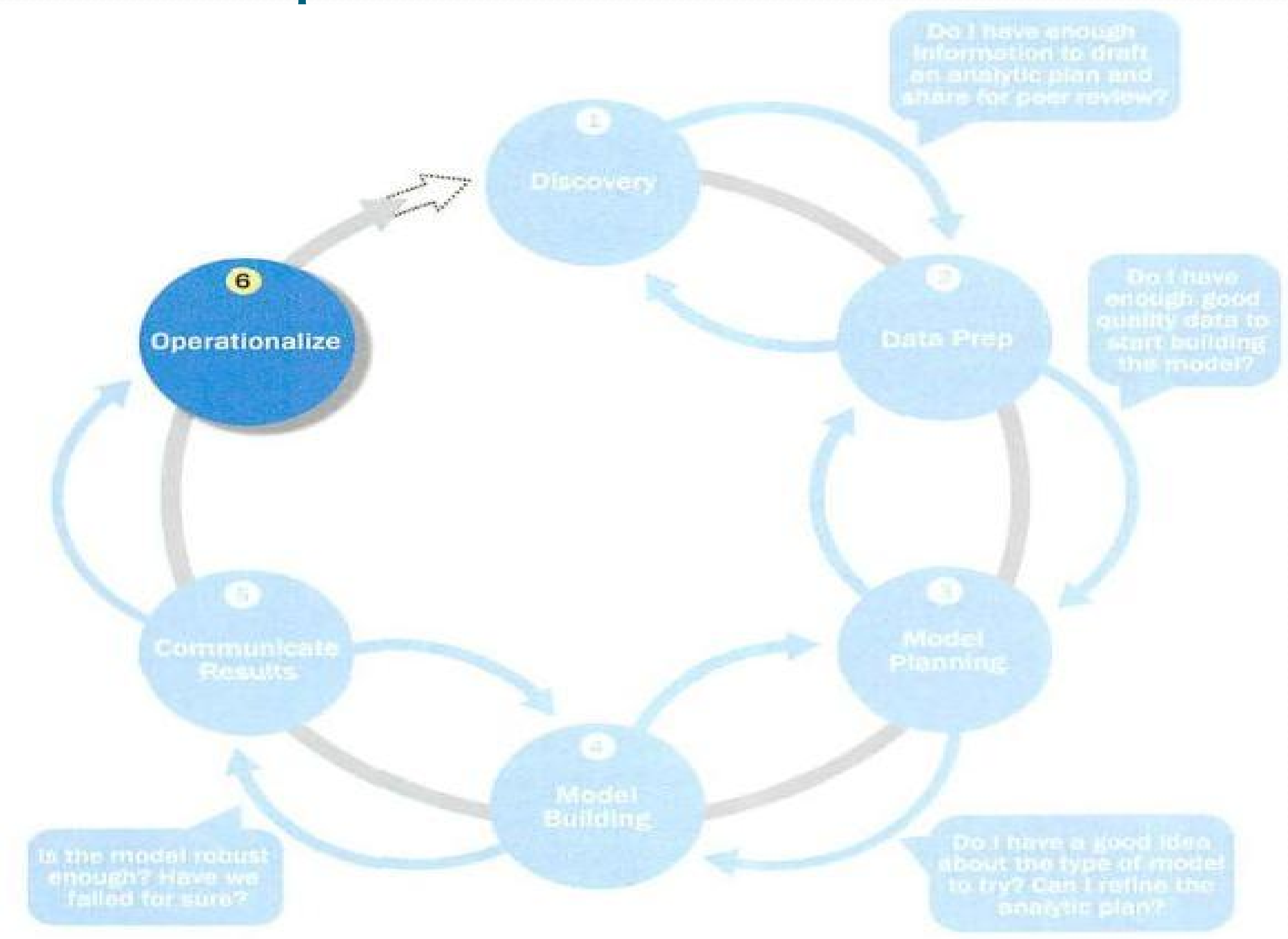


Do I have enough information to draft an analytic plan and share for peer review?
1
Discovery
2
Data Prep
Do I have enough good quality data to start building the model?
3
Model Planning
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
4
Model Building
Is the model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

# 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

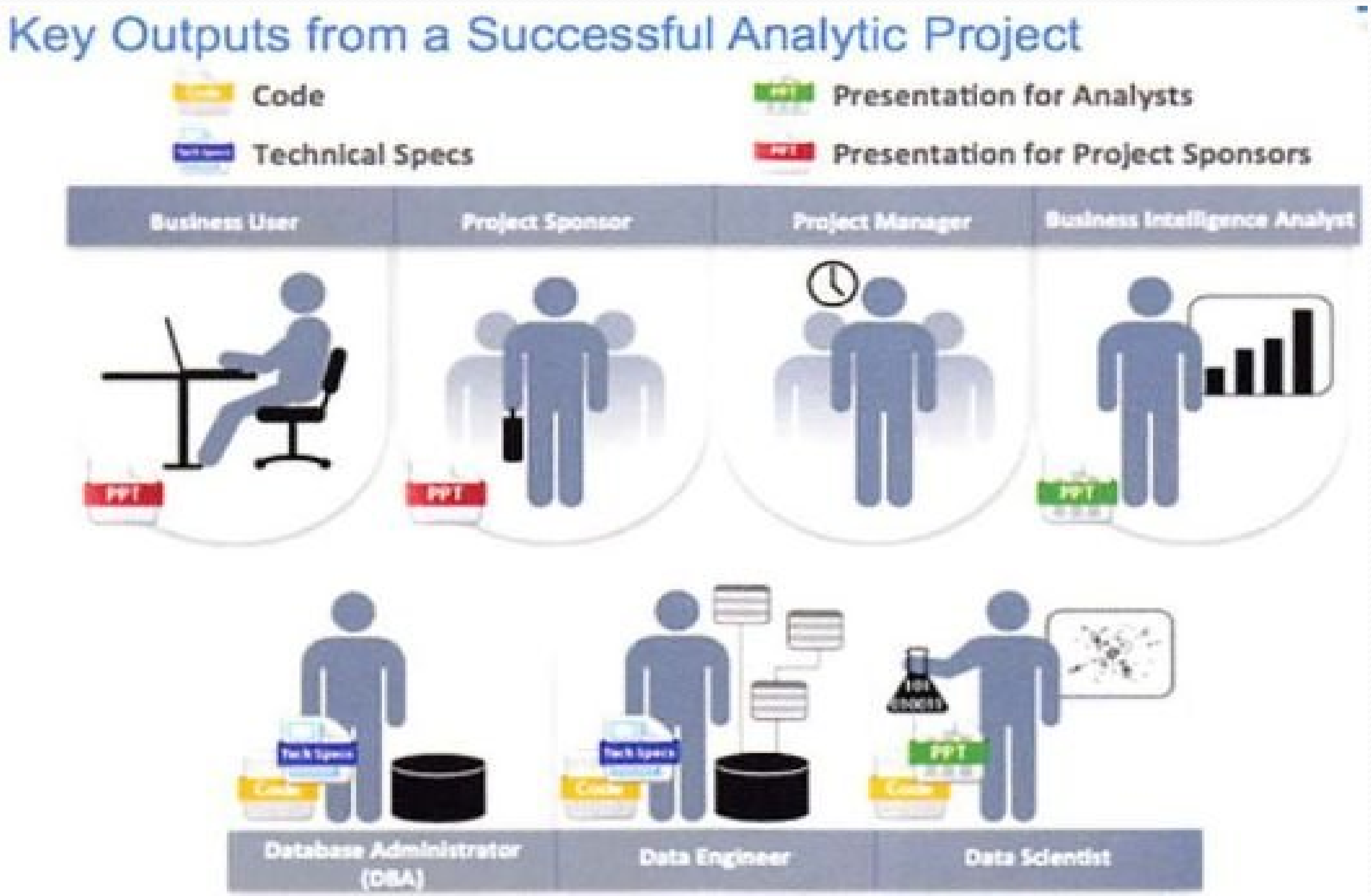

Do I have enough information to draft an analytic plan and share for peer review?
1
Discovery
2
Data Prep
Do I have enough good quality data to start building the model?
3
Model Planning
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
4
Model Building
5
Communicate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

# 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

## Key outputs from successful analytics project



Key Outputs from a Successful Analytic Project

- Code
- Technical Specs
- Presentation for Analysts
- Presentation for Project Sponsors

# 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 completed on time, within budget, goals met
- 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

# 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 – 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