

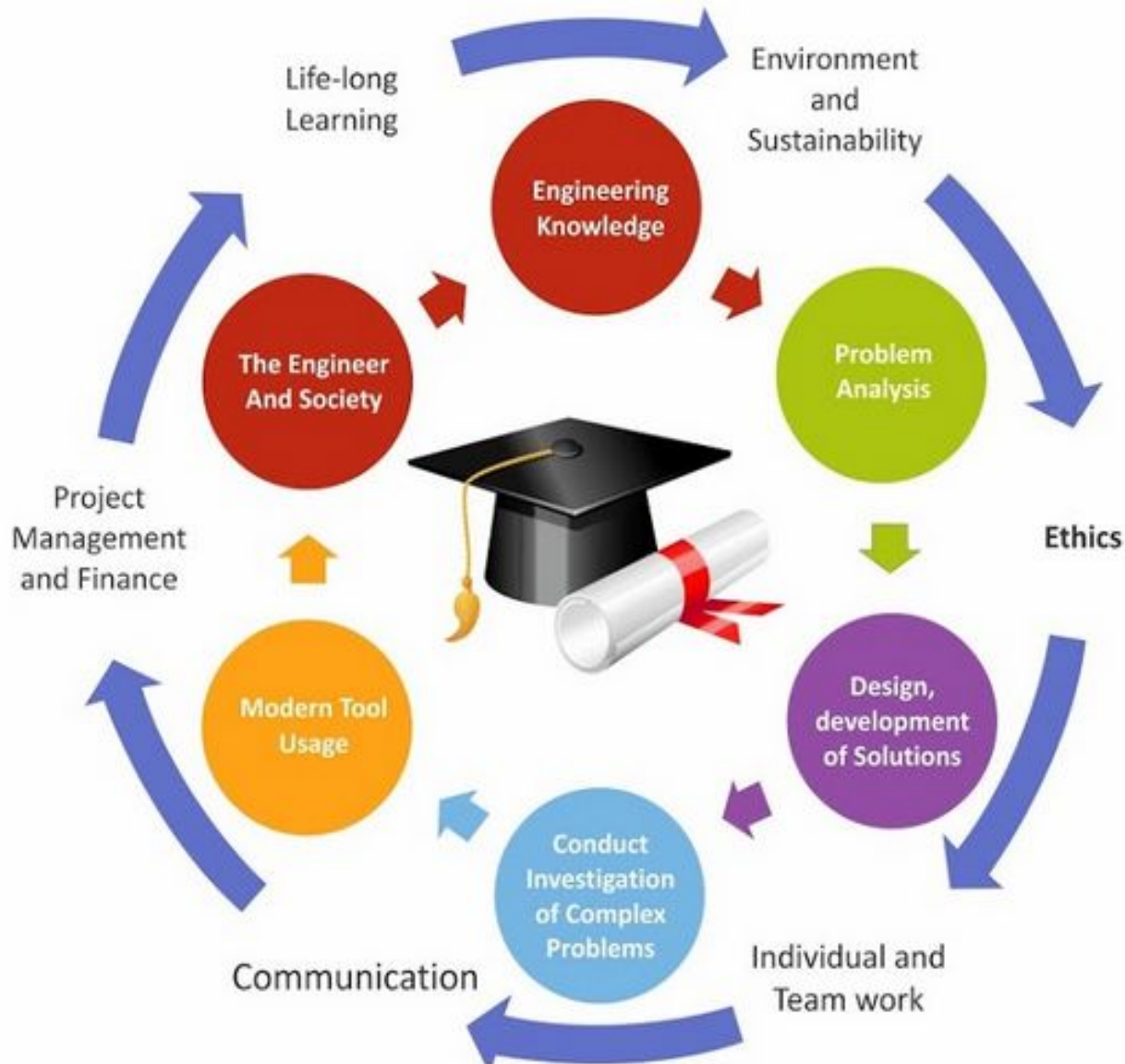
Data Analytics

Unit-I

Prof. Bhavan .A. Khivsara

Note: The material to prepare this presentation has been taken from internet and are generated only for students reference and not for commercial use.

PO's - Program Outcome



Course Objective

- To develop problem solving abilities using Mathematics
- To apply algorithmic strategies while solving problems
- To develop time and space efficient algorithms
- To study algorithmic examples in distributed, concurrent and parallel environments

Course Outcome

- To Understand basics of Big data and Write case studies in Business Analytic
- To Apply mathematical model using statistics for Business Analytic and Intelligence applications.
- To Critically analyze problems and identify analytical solutions using Association Rules & Regression
- To analyze problem and identify analytical solutions using decision tree and Naive Baise
- To understand the Data Visualization techniques and tools
- To gain knowledge on Hadoop related tools such as HBase, Hive and Mahout for big data analytics



- 01 Importance
- 02 Big Data Analytics
- 03 Use-Cases
- 04 Technologies



What is Big Data

What is Big Data?



What is Big Data?



Large amounts of data



collected passively from digital interactions



with great variety and a high rate of velocity.

Big Data is also **data** but with a **huge size**. Big Data is a term used to describe a collection of data that is huge in volume and yet growing exponentially with time.

<https://data2x.org/resource-center/big-data-and-the-wellbeing-of-women-and-girls/>

Big Data Types

Structured Data



Unstructured Data



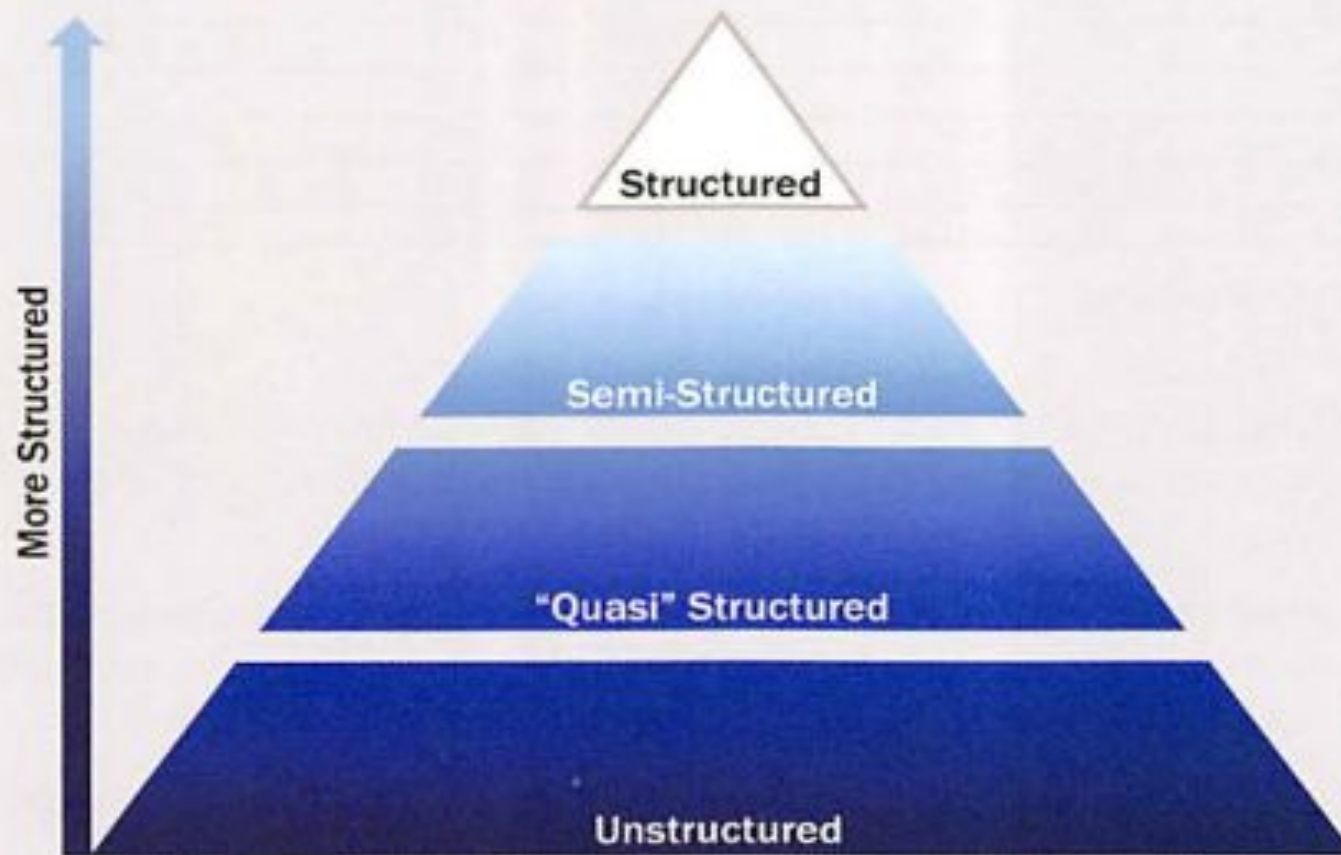
Semi-structured Data



- **Relational databases** are examples of structured data.
- Examples of **unstructured data** include audio, video
- **semi-structured data** includes the XML data, JSON files, and others.

Big Data Characteristics: Data Structures

Data Growth Is Increasingly Unstructured



Data Structures: Characteristics of Big Data

Structured

- Transactional data, OLAP cubes, RDBMS, CVS files, spreadsheets

Semi-structured

- Text data with discernable patterns – e.g., XML data

Quasi-structured

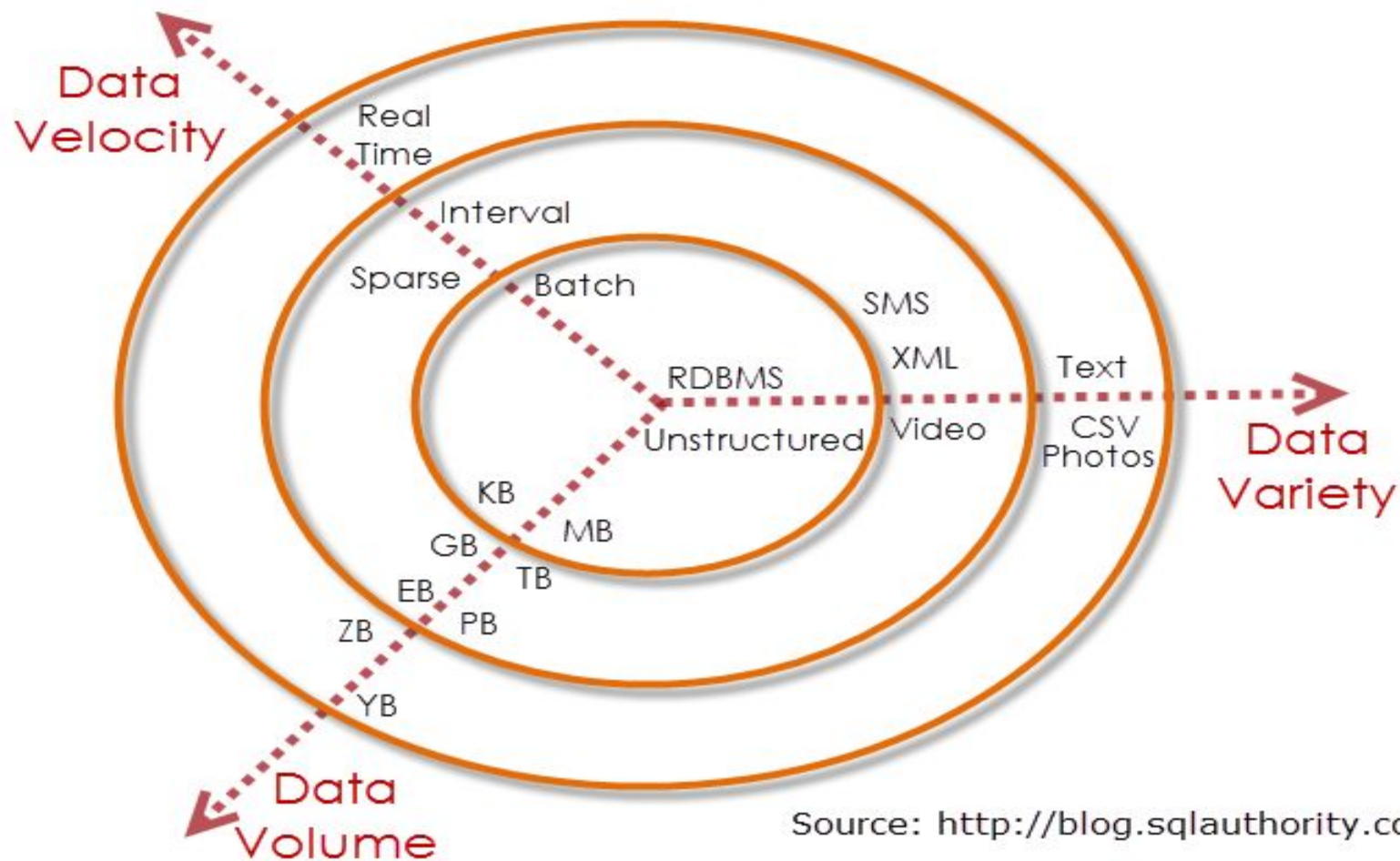
- Text data with erratic data formats – e.g., clickstream data

Unstructured

- Data with no inherent structure – text docs, PDF's, images, video

3 V's Big Data- Characteristics of Big Data

- Big data is often characterized by the 3Vs:



Source: <http://blog.sqlauthority.com>

THE 3Vs OF BIG DATA

VOLUME

- ◆ Amount of data generated
- ◆ Online & offline transactions
- ◆ In kilobytes or terabytes
- ◆ Saved in records, tables, files



VELOCITY

- ◆ Speed of generating data
- ◆ Generated in real-time
- ◆ Online and offline data
- ◆ In Streams, batch or bits



VARIETY

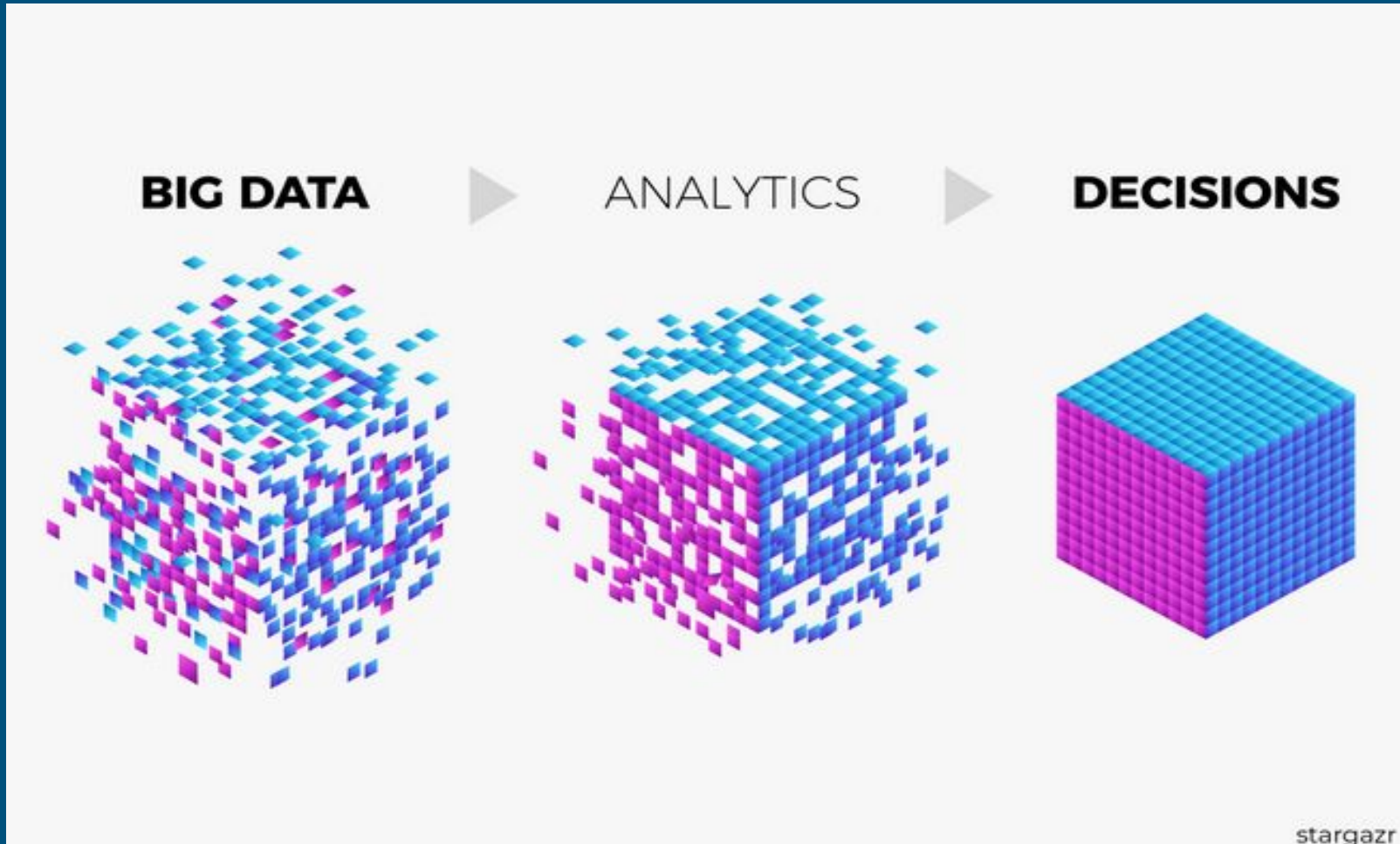
- ◆ Structured & unstructured
- ◆ Online images & videos
- ◆ Human generated - texts
- ◆ Machine generated - readings



Importance of Big Data

- Companies use the big data accumulated in their systems to improve operations, provide better customer service, create personalized marketing campaigns based on specific customer preferences and, ultimately, increase profitability.

What is Big Data Analytics?



<https://towardsdatascience.com/how-is-the-current-state-of-big-data-analytics-in-controlling-1273c725ac6a>

Tools for Big Data Analytics

Top Tools Used in Big Data Analytics



<https://intellipaate.com/blog/big-data-analytics/>

Big Data Use cases Examples

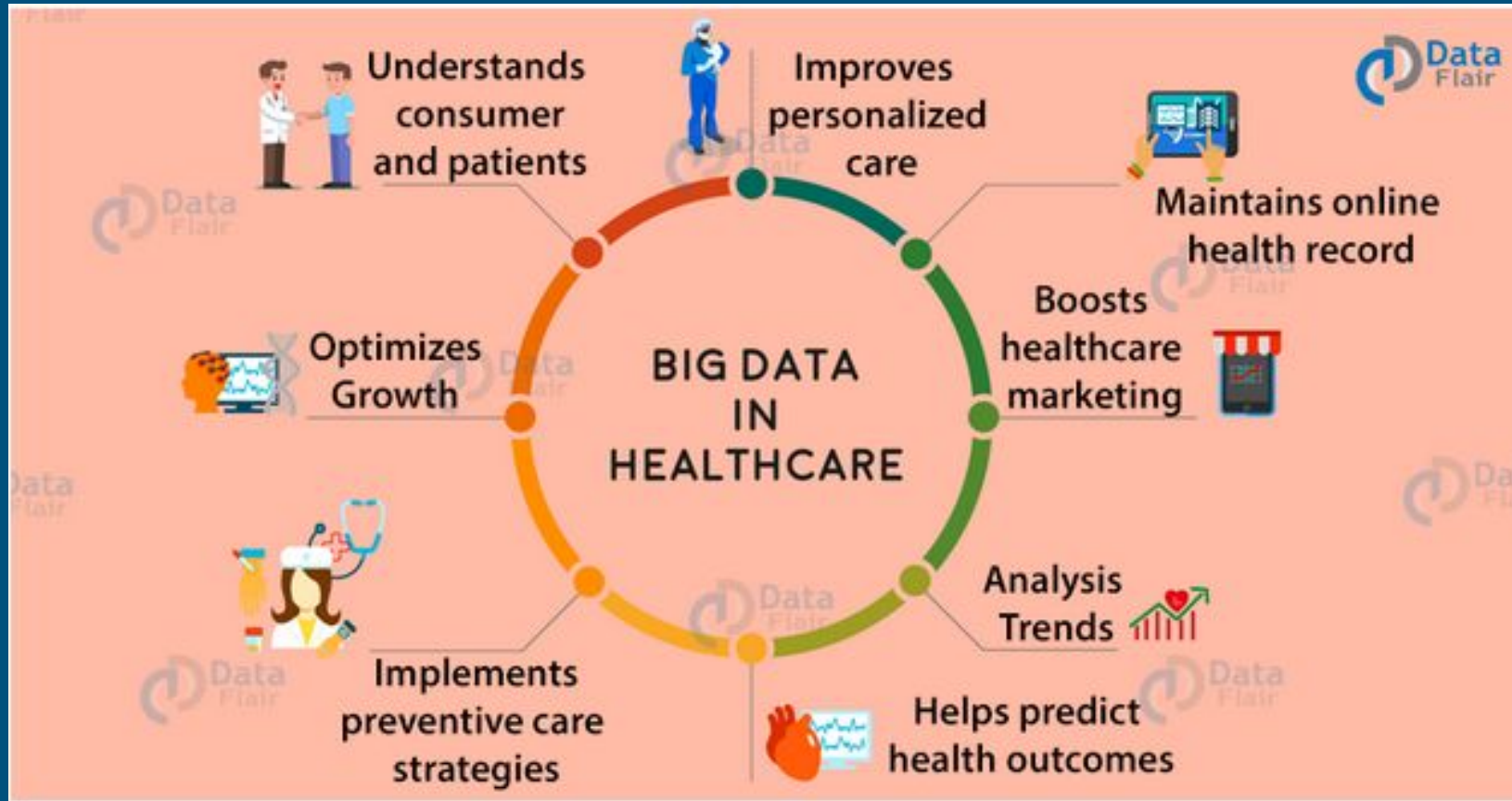


<https://data-flair.training/blogs/big-data-use-cases-case-studies-hadoop-spark-flink/>

Big Data Use cases in Different Domains

- Big Data in Retail
- Big Data in Healthcare
- Big Data in Education
- Big Data in E-commerce
- Big Data in Media and Entertainment
- Big Data in Finance
- Big Data in Travel Industry
- Big Data in Telecom
- Big Data in Automobile

Big Data in Health Care



<https://data-flair.training/blogs/big-data-in-healthcare-applications/>

Big Data in Agriculture



<https://data-flair.training/blogs/big-data-in-agriculture/>

Big Data in Retail Industry



Big Data Applications in Bank



<https://data-flair.training/blogs/big-data-in-banking/>

Big Data in Travel Industry



<https://data-flair.training/blogs/big-data-in-travel-industry/>

Sources of Big Data Deluge

What's Driving Data Deluge?



**Mobile
Sensors**



**Social
Media**



**Video
Surveillance**



**Video
Rendering**



**Smart
Grids**



**Geophysical
Exploration**



**Medical
Imaging**



**Gene
Sequencing**

Data Repositories

- Data repository - **several ways to collect and store data:**



Data Repositories

Data Mart

- **Focus:** A single subject or functional organization area
- **Size:** Less than 100 GB
- **Data Held:** Typically summarized data

Data Warehouse

- **Focus:** Enterprise-wide repository of disparate data sources
- **Size:** 100 GB minimum but often in the range of terabytes
- **Data Held:** Raw data, metadata, and summary data

Data lakes

large data repositories that store unstructured data that is classified and tagged with metadata.

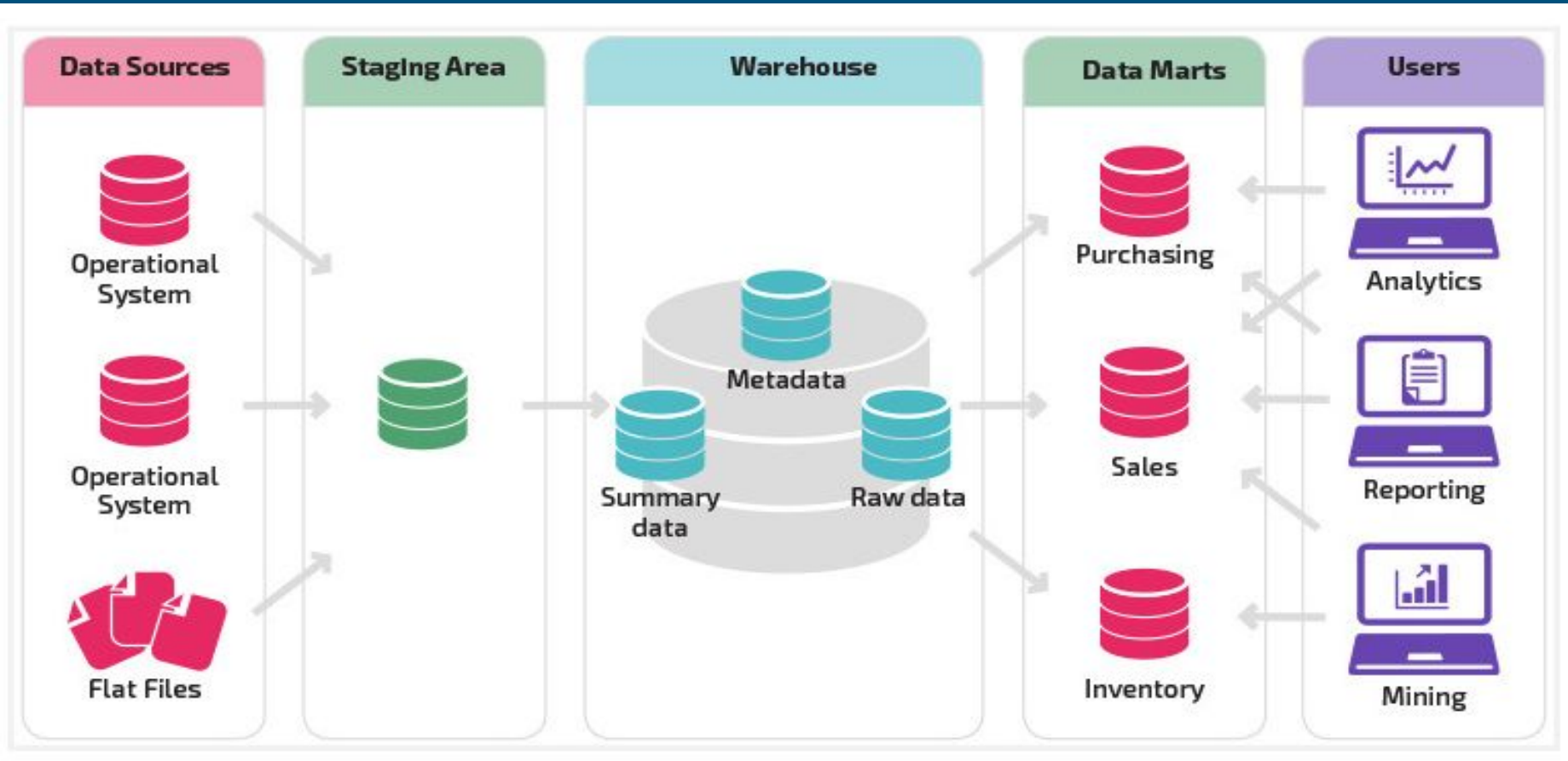
Data cubes

- lists of data with three or more dimensions stored as a table

Analytical Sandbox-

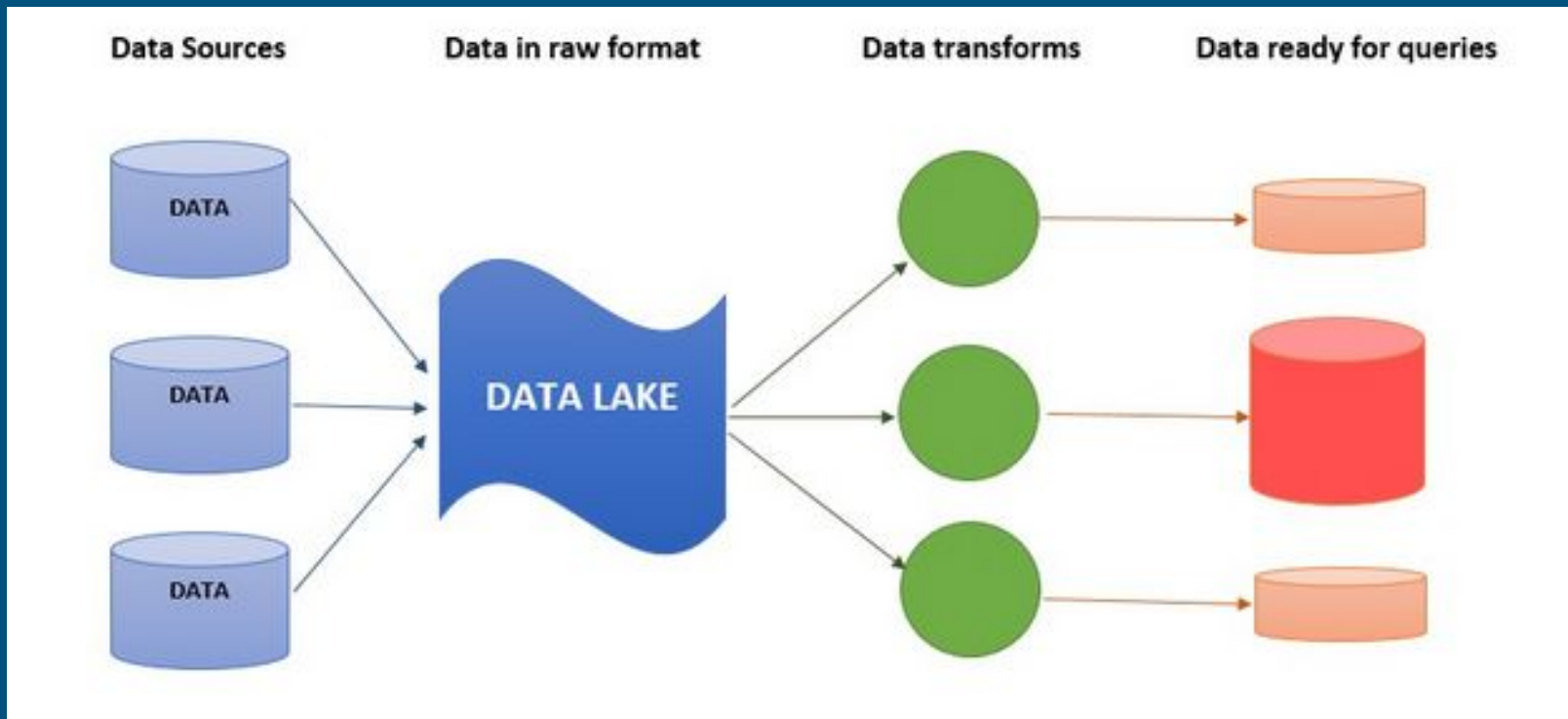
Provide the computing required for data scientists to tackle typically complex *analytical* workloads.

Data Warehouse & Data Marts



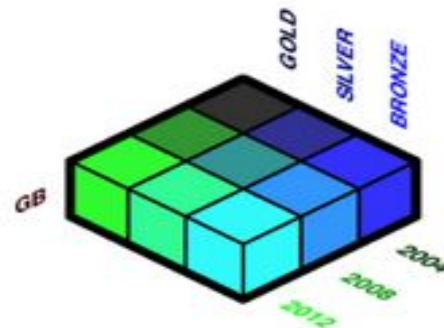
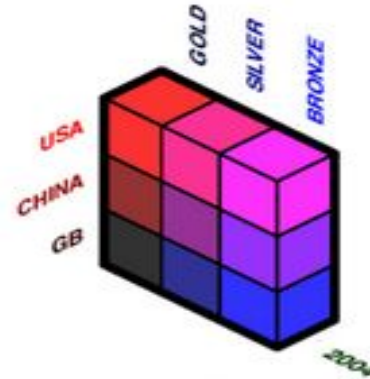
<https://panoply.io/data-warehouse-guide/data-mart-vs-data-warehouse/>

Data lakes

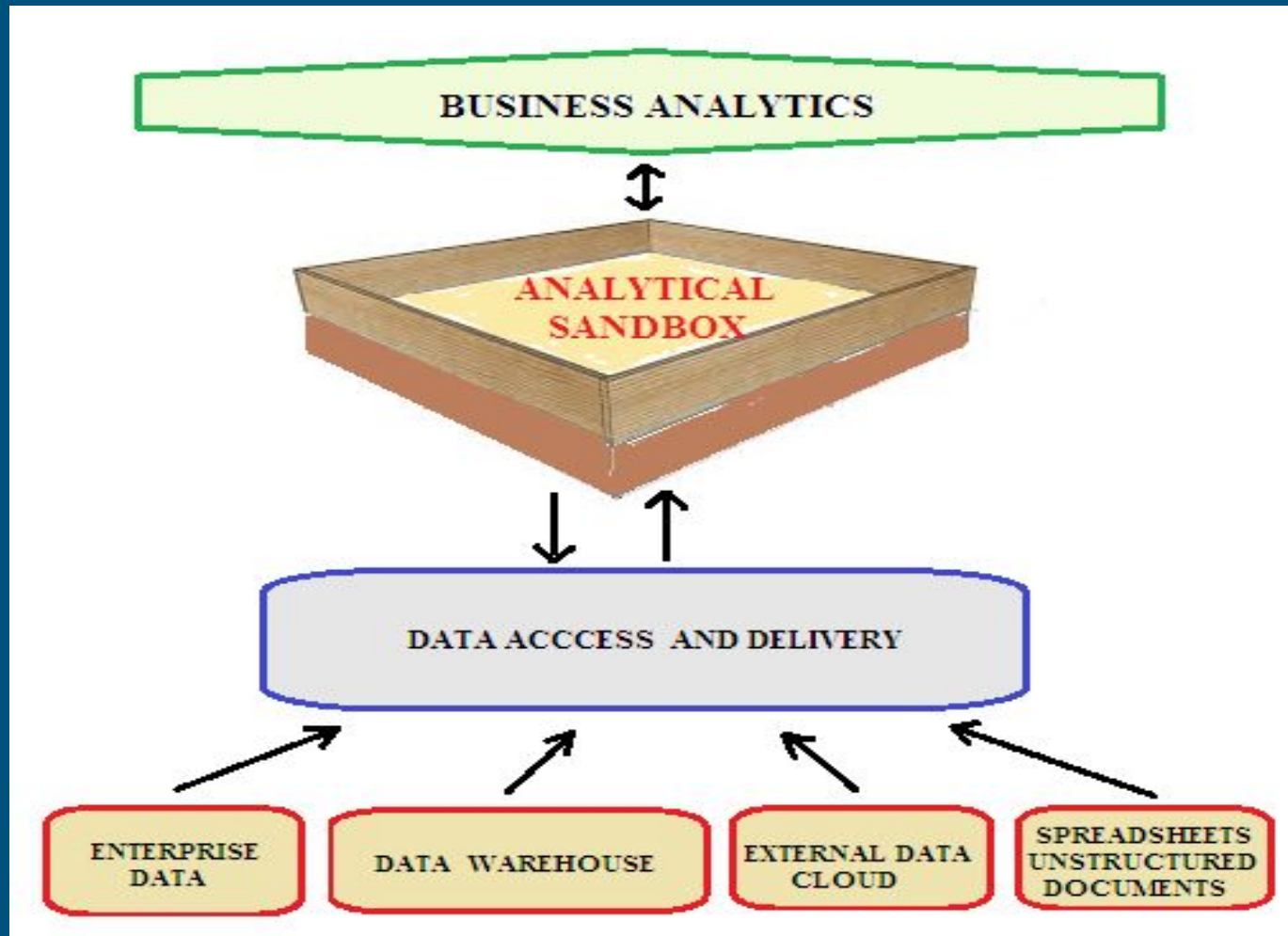


<https://databricks.com/glossary/data-lake>

Data Cubes

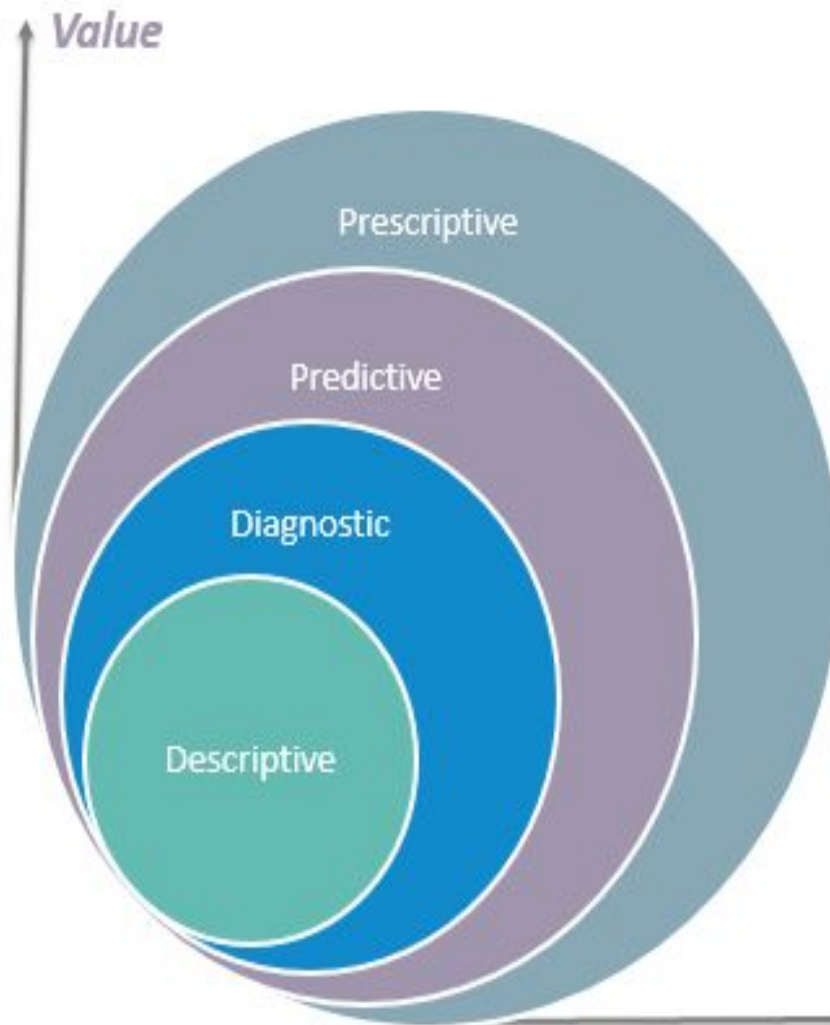


Analytical Sandbox



What is Data Analytics?

4 types of Data Analytics



What is the data telling you?

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

Complexity

Data Analytics Techniques

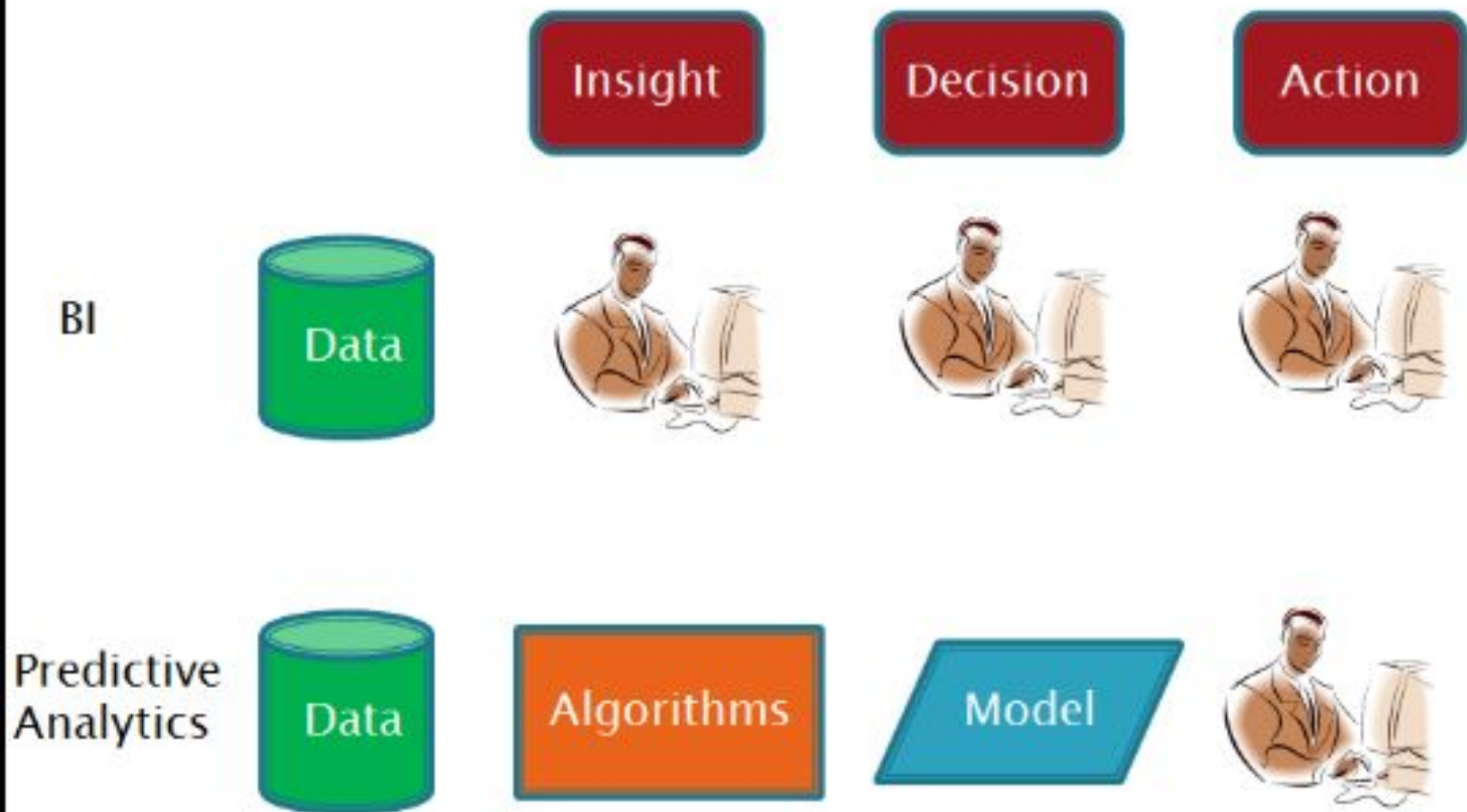
BI (Business
Intelligence)

Data Science

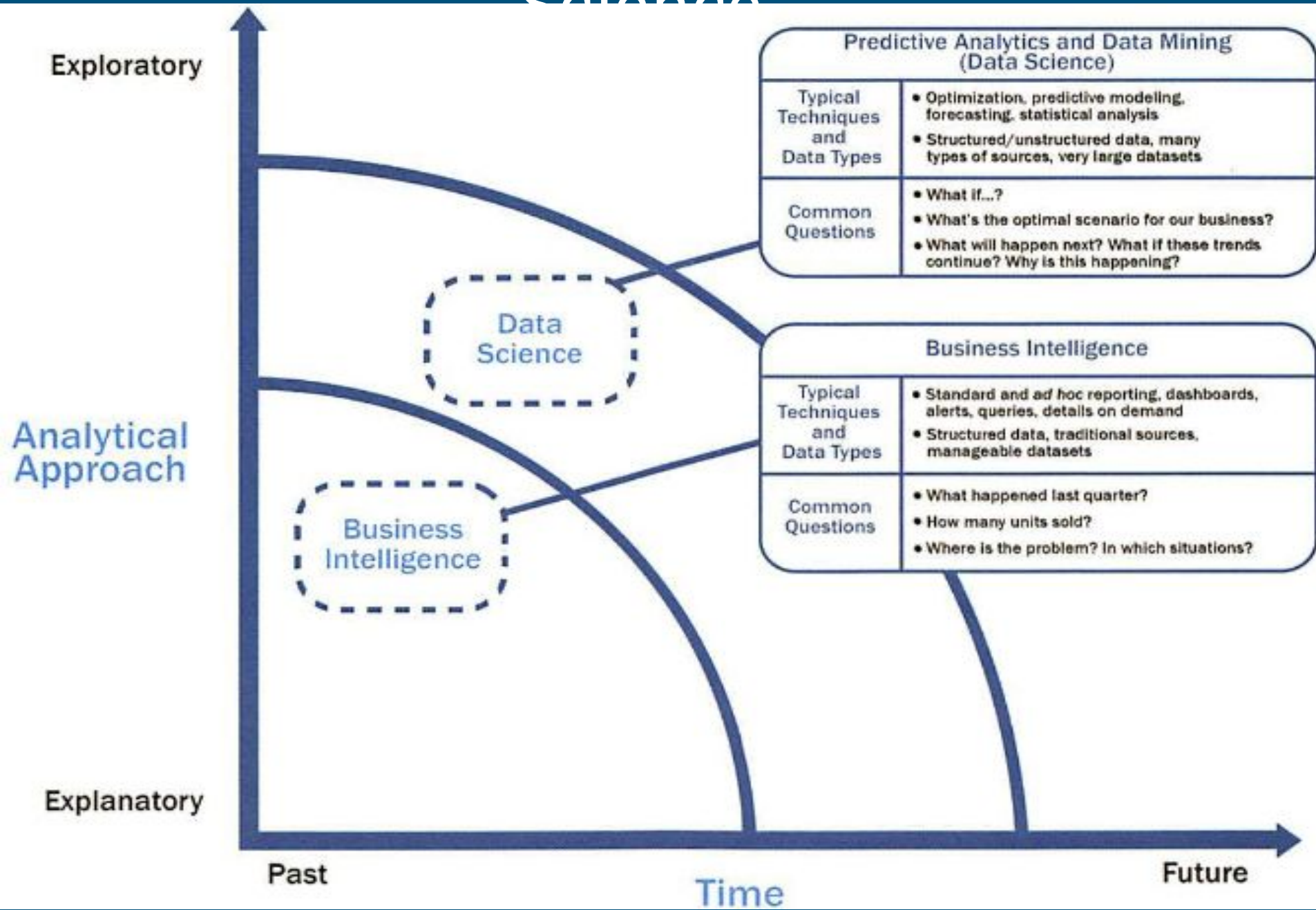


<https://www.datasciencecentral.com/profiles/blogs/bi-vs-big-data-vs-data-analytics-by-example>

BI vs Predictive Analytics



Business Intelligence (BI) vs Data Science

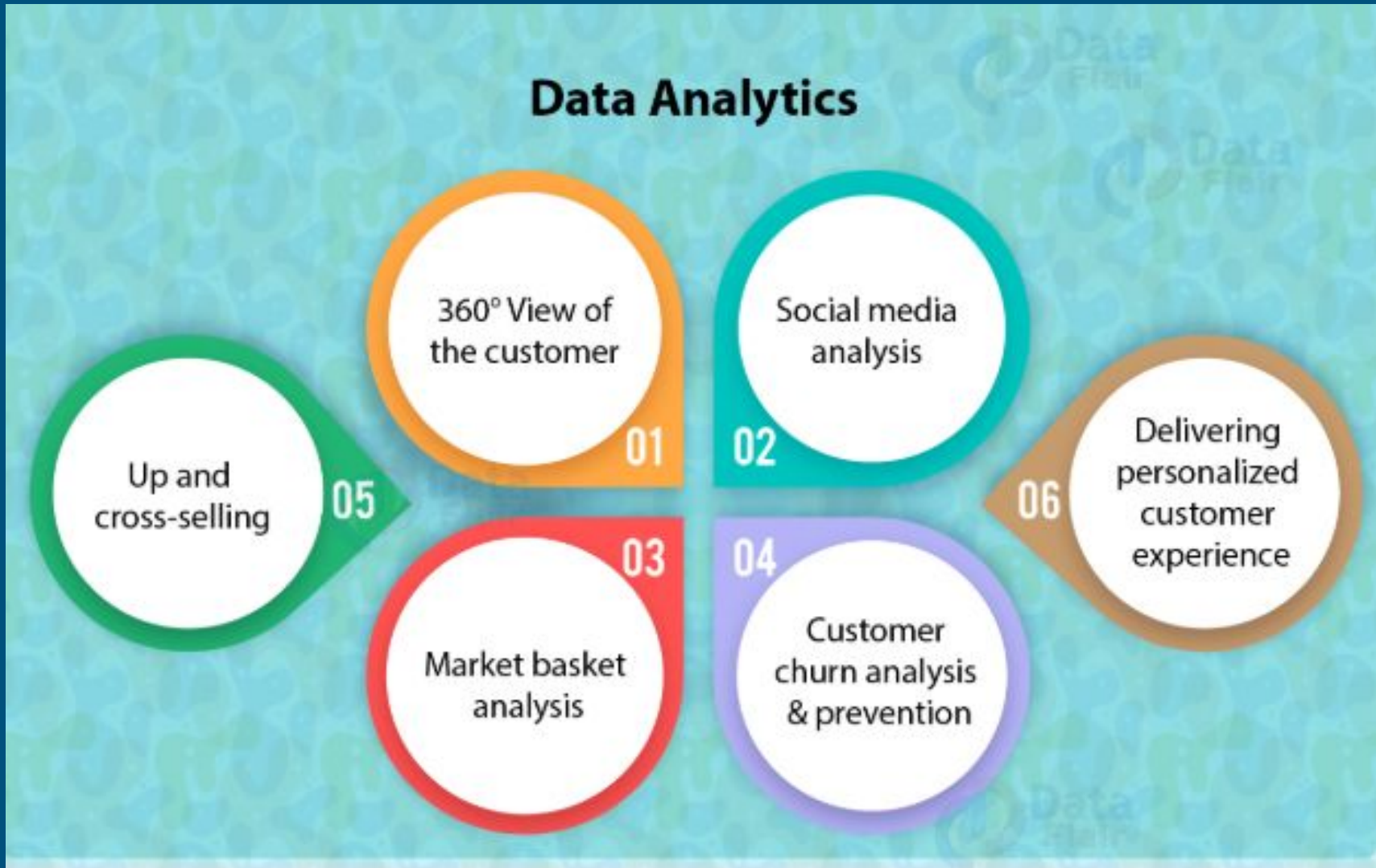


Data Science vs. Business Intelligence

	Business Intelligence (BI)	Data Science
Data analysis	Yes	Yes
Statistics	Yes	Yes
Visualization	Yes	Yes
Data Sources	Usually SQL, often Data Warehouse	Less structured (logs, cloud data, SQL, noSQL, text)
Tools	Statistics, Visualization	Statistics, Machine Learning, Graph Analysis, NLP
Focus	Present and past	Future
Method	Analytic	Scientific
Goal	Better strategic decisions	Advanced functionality

The two fields are closely related. In some ways Data Science is an evolution of BI.

Data Analytics Usecases



Social Media Sentiment analysis

Project: donald trump

Mentions:

69530

Total

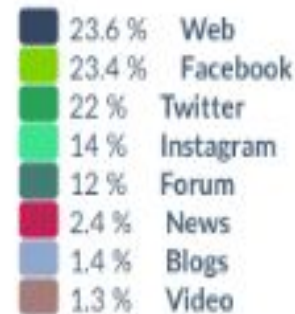
17415

Positive

16635

Negative

Sources



Sentiment



Social Media Sentiment analysis

Project: uber

Mentions:

44652

Total

14745

Positive

4533

Negative

Sources



Sentiment



Data Analytics Usecases

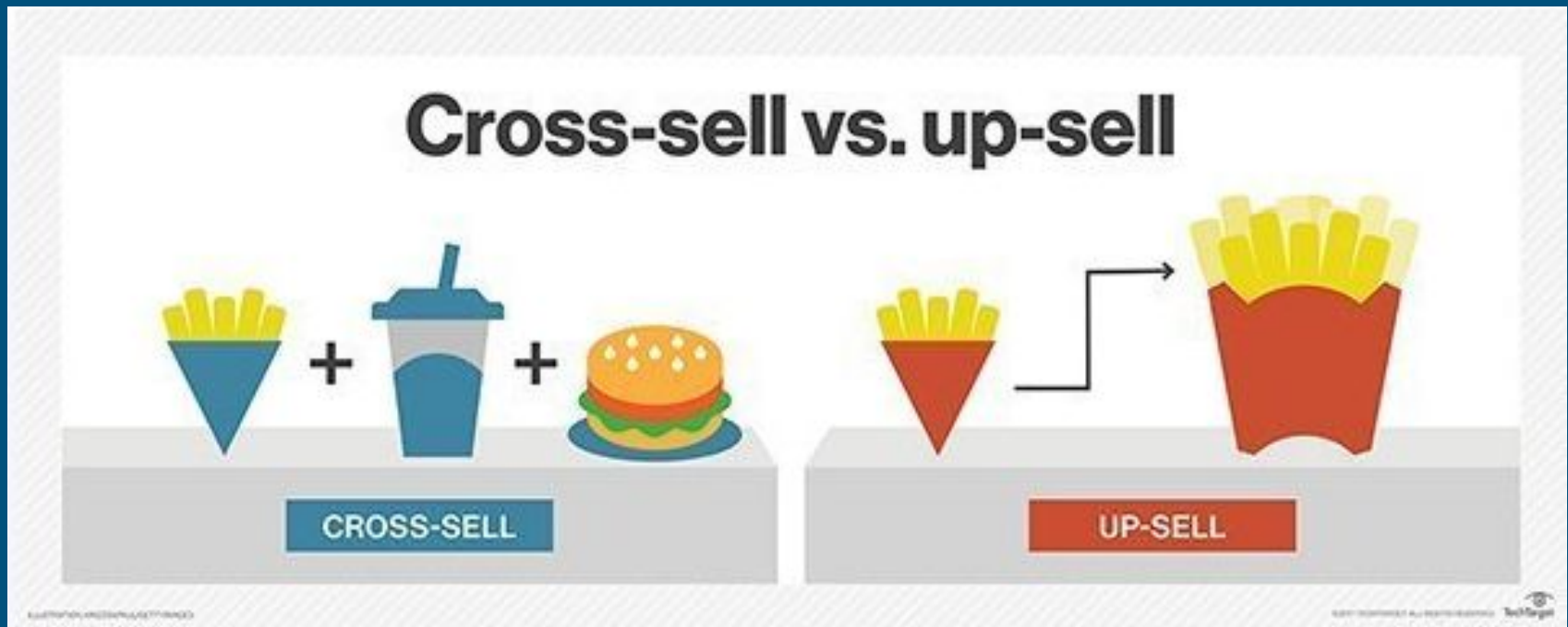
Social Media Sentiment Analysis

- <https://www.preprints.org/manuscript/202005.0015/v1>
- <https://www.regenhealthsolutions.info/2020/04/09/usc-researchers-analyze-covid-19-misinformation-on-twitter/>

Market Basket Analysis

- <https://www.analyticsvidhya.com/blog/2014/08/effective-cross-selling-market-basket-analysis/>
- <https://medium.com/@notesharsha/market-basket-analysis-sneaky-psychology-of-supermarkets-simple-guide-using-python-eacfd33cc882>

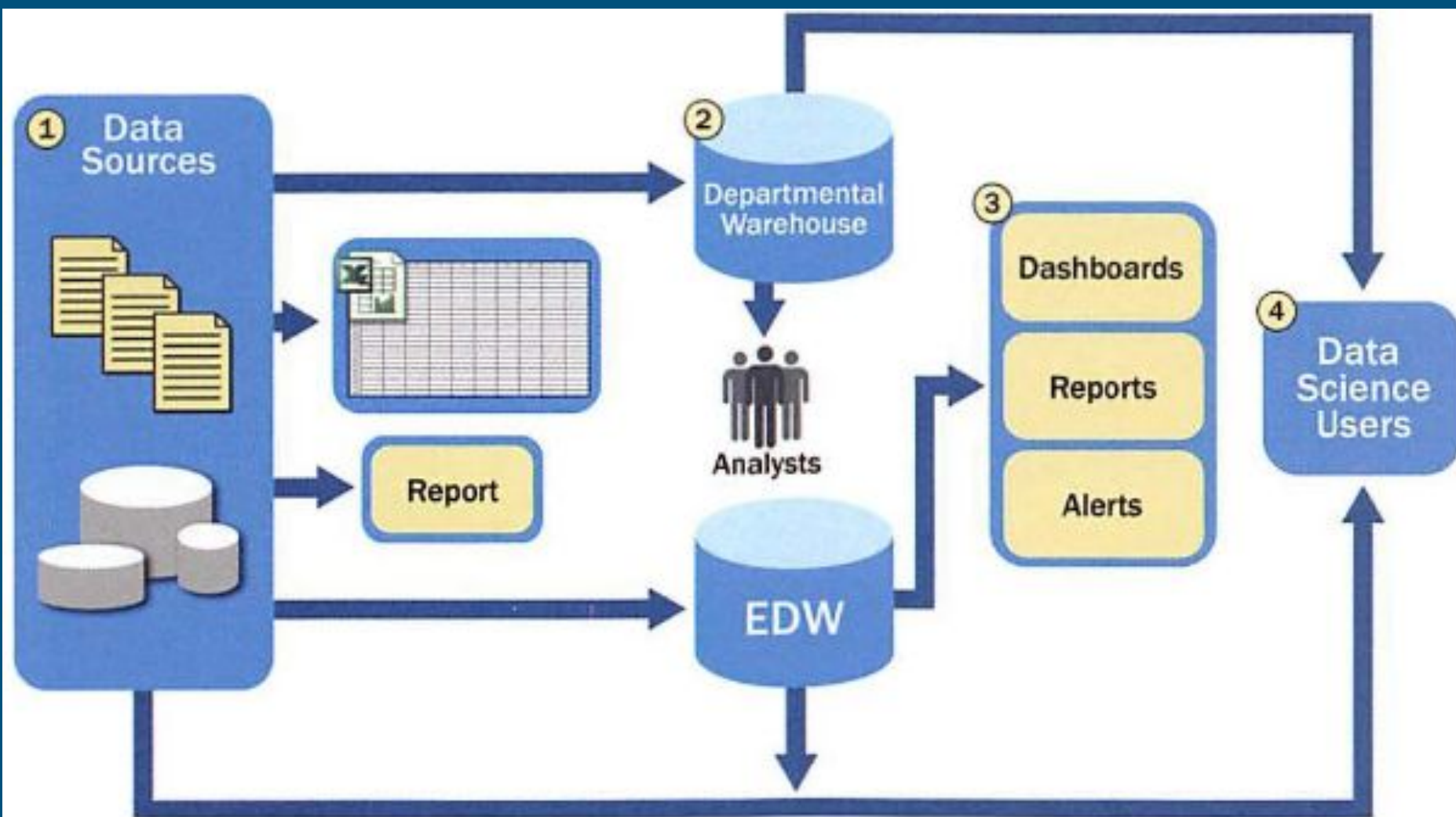
Data Analytics Usecases



<https://instapage.com/blog/cross-selling>

Current Analytical Architecture

Typical Analytic Architecture



Current Analytical Architecture

Data sources must be well understood

EDW – Enterprise Data Warehouse

From the EDW data is read by applications

Data scientists get data for downstream analytics processing

Current Analytical Architecture

-Problem

High-value data is hard to reach, and predictive analytics and data mining activities are last in line for data.

Data scientists are limited to performing in-memory analytics, which will restrict the size of the datasets . So Analyst works on sampling, which can skew model accuracy.

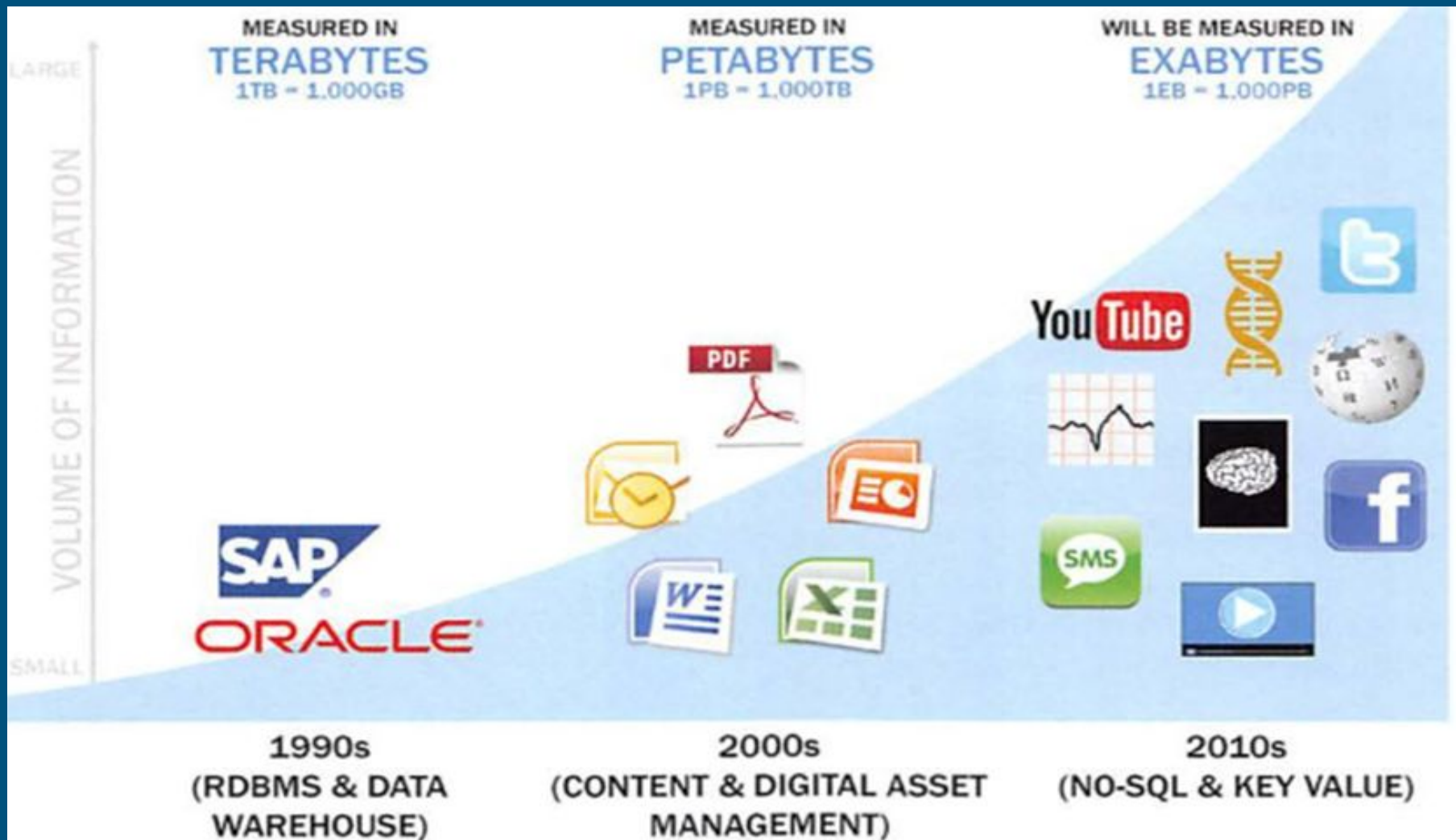
Data Science projects will remain isolated rather than centrally managed. The implication of this is that the organization can never tie together the power of advanced analytics.

Current Analytical Architecture -Solution

- One solution to this problem is to introduce **analytic sandboxes** to enable data scientists to perform advanced analytics.

Drivers of Big Data

Data Evolution & Rise of Big Data Sources



Drivers of Big Data

Data Evolution & Rise of Big Data Sources

- **Medical information**, such as diagnostic imaging
- **Photos and video** footage uploaded to the World Wide Web
- **Video surveillance**, such as the thousands of video cameras across a city
- **Mobile devices**, which provide geospatial location data of the users
- **Metadata** about text messages, phone calls, and application usage on smart phones
- **Smart devices**, which provide sensor-based collection of information from smart
- **Nontraditional IT devices**, including the use of RFID readers, GPS navigation systems, and seismic processing

Emerging Big Data Ecosystem

**Data
devices**

Games, smartphones, computers

**Data
collectors**

Phone and TV companies,
Internet, Gov't

**Data
aggregators**

Websites, credit bureaus, media
archives

**Data users
& buyers**

Banks, law enforcement,
marketers, employers

Emerging Big Data Ecosystem

Data devices

Gather data from multiple locations and continuously generate new data about this data. For each gigabyte of new data created, an additional petabyte of data is created about that data.

- For example, playing an online video game, Smartphones data, Retail shopping loyalty cards data

Emerging Big Data Ecosystem

Data collectors

Include sample entities that collect data from the device and users.

For example, Retail stores tracking the path a customer

Emerging Big Data Ecosystem

Data aggregators – make sense of data

They transform and package the data as products to sell to list brokers for specific ad campaigns.

For example, Digital Marketing

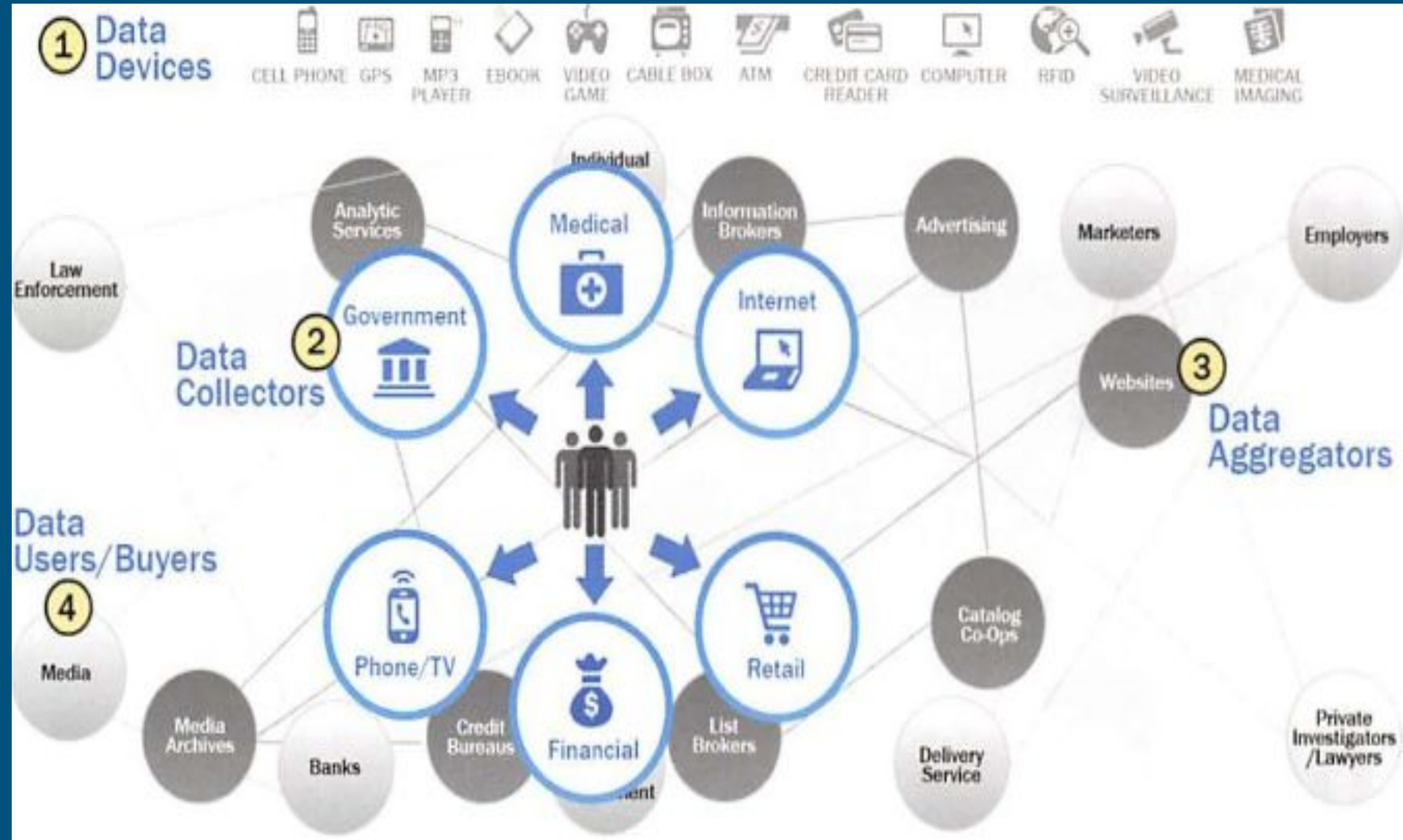
Emerging Big Data Ecosystem

Data users and buyers

These groups directly benefit from the data collected and aggregated by others within the data value chain.

For Example, People want to determine public sentiments toward a candidate by analyzing related blogs and online comments

Emerging Big Data Ecosystem



Key Roles for the New Big Data Ecosystem

Deep analytical talent

Advanced training in quantitative disciplines
– e.g., math, statistics, machine learning

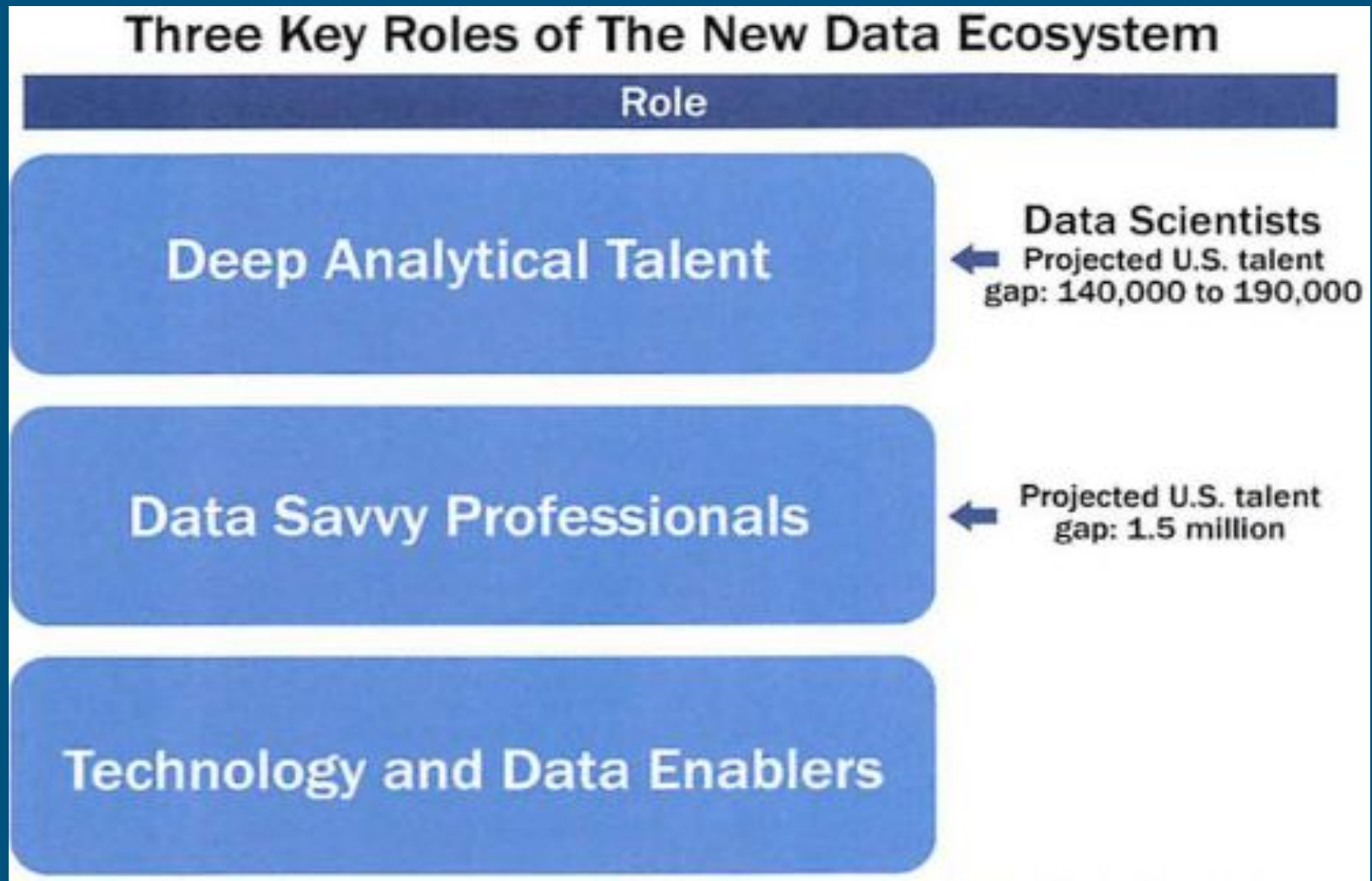
Data savvy (Intelligent knowledgeable)

Savvy but less technical than group 1

Technology and data enablers

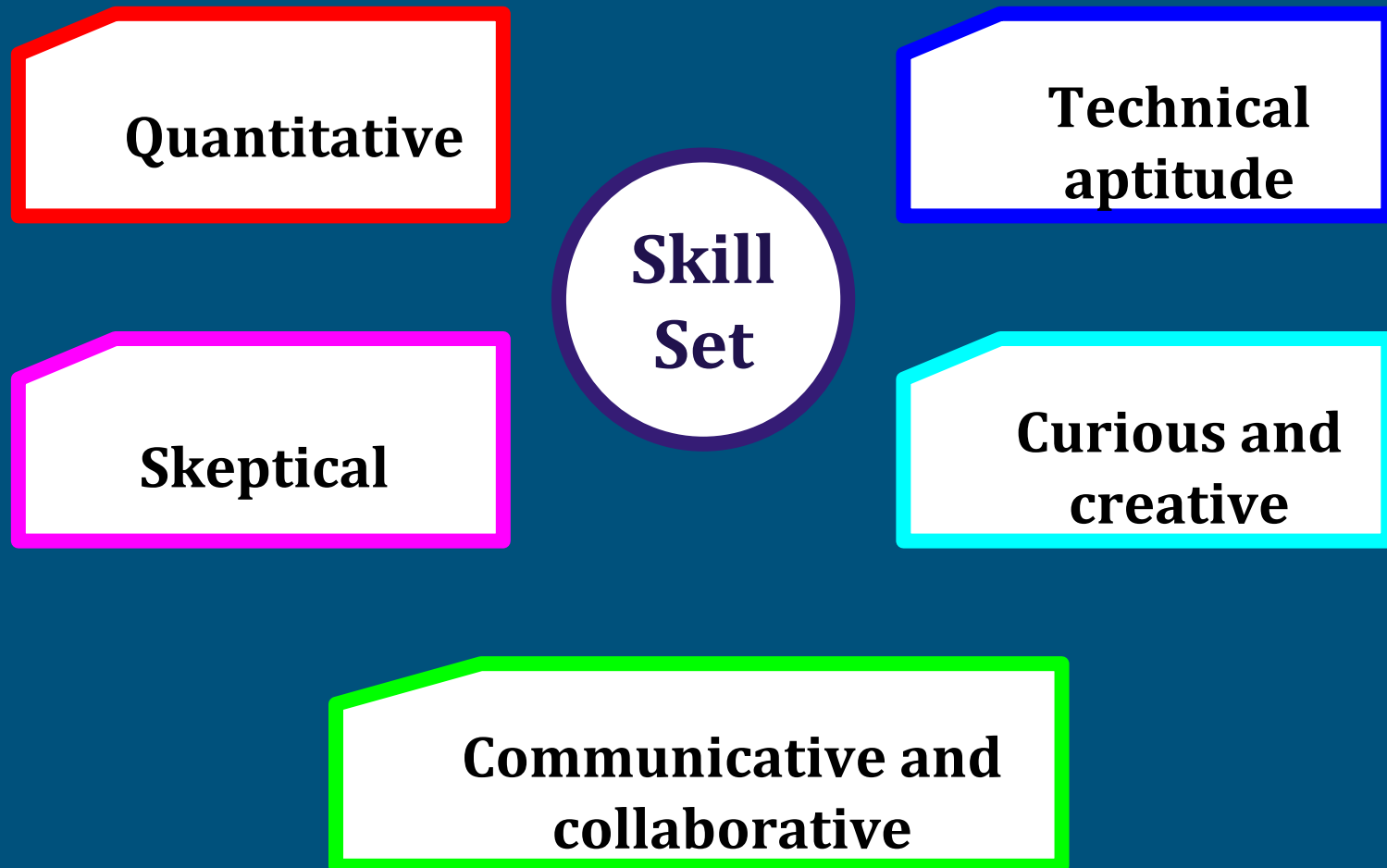
Support people – e.g., DB admins, programmers, etc.

Three Key Roles of the New Big Data Ecosystem



Profile of Data Scientist

Five Main Sets of Skills



Profile of Data Scientist

Five Main Sets of Skills

- **Quantitative skill** – e.g., math, statistics
- **Technical aptitude** – e.g., software engineering, programming
- **Skeptical mindset and critical thinking** – ability to examine work critically
- **Curious and creative** – passionate about data and finding creative solutions
- **Communicative and collaborative** – can articulate ideas, can work with others

Exercise

- 1. What are the three characteristics of Big Data, and what are the main considerations in processing Big Data?
- 2. What is an analytic sandbox, and why is it important?
- 3. Explain the differences between BI and Data Science.
- 4. Describe the challenges of the current analytical architecture for data scientists.
- 5. What are the key skill sets and behavioral characteristics of a data scientist?

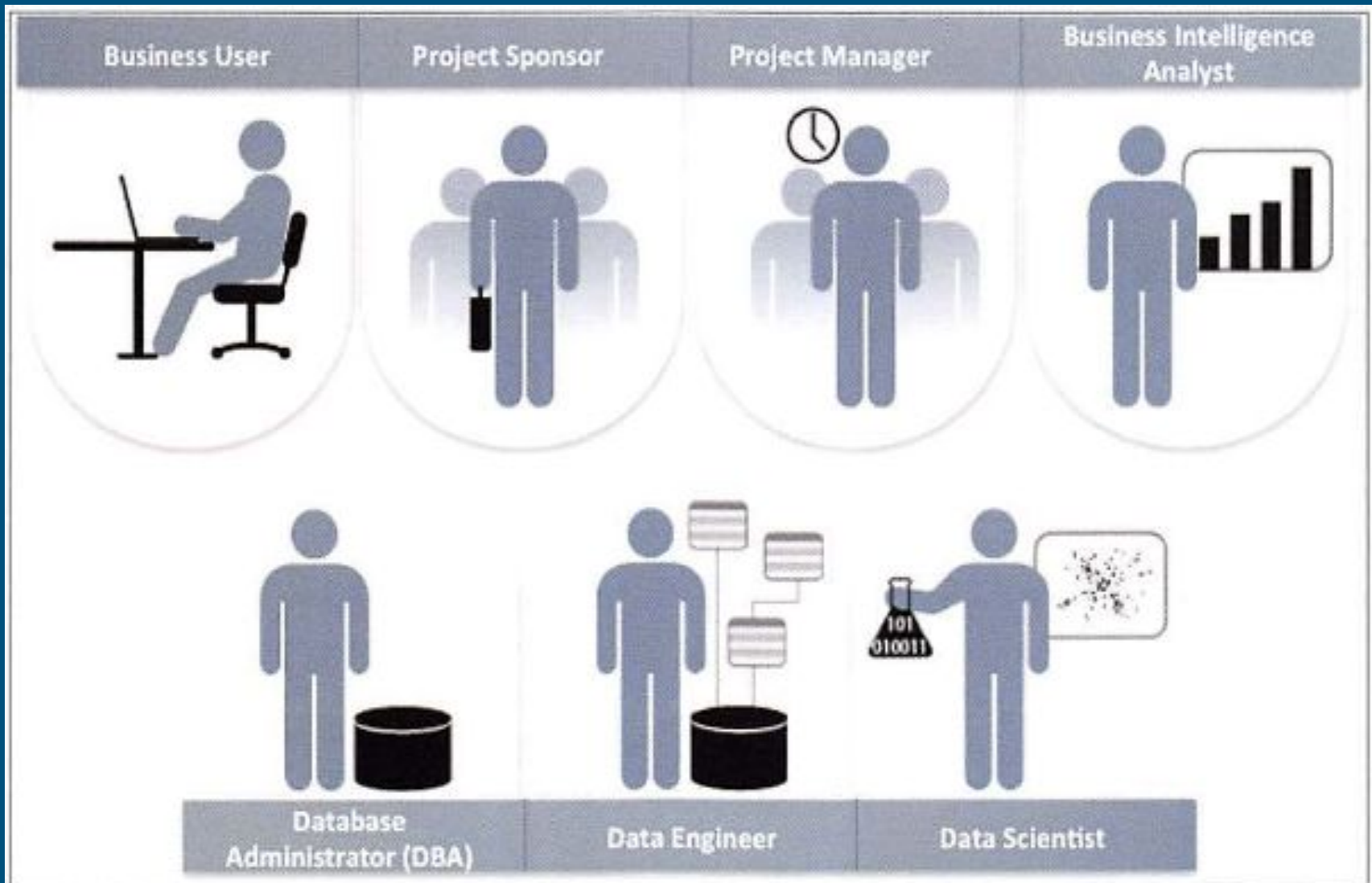


Data Analytics Lifecycle



—

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 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

Data Analytics Lifecycle

Phase 1: Discovery

Phase 2: Data Preparation

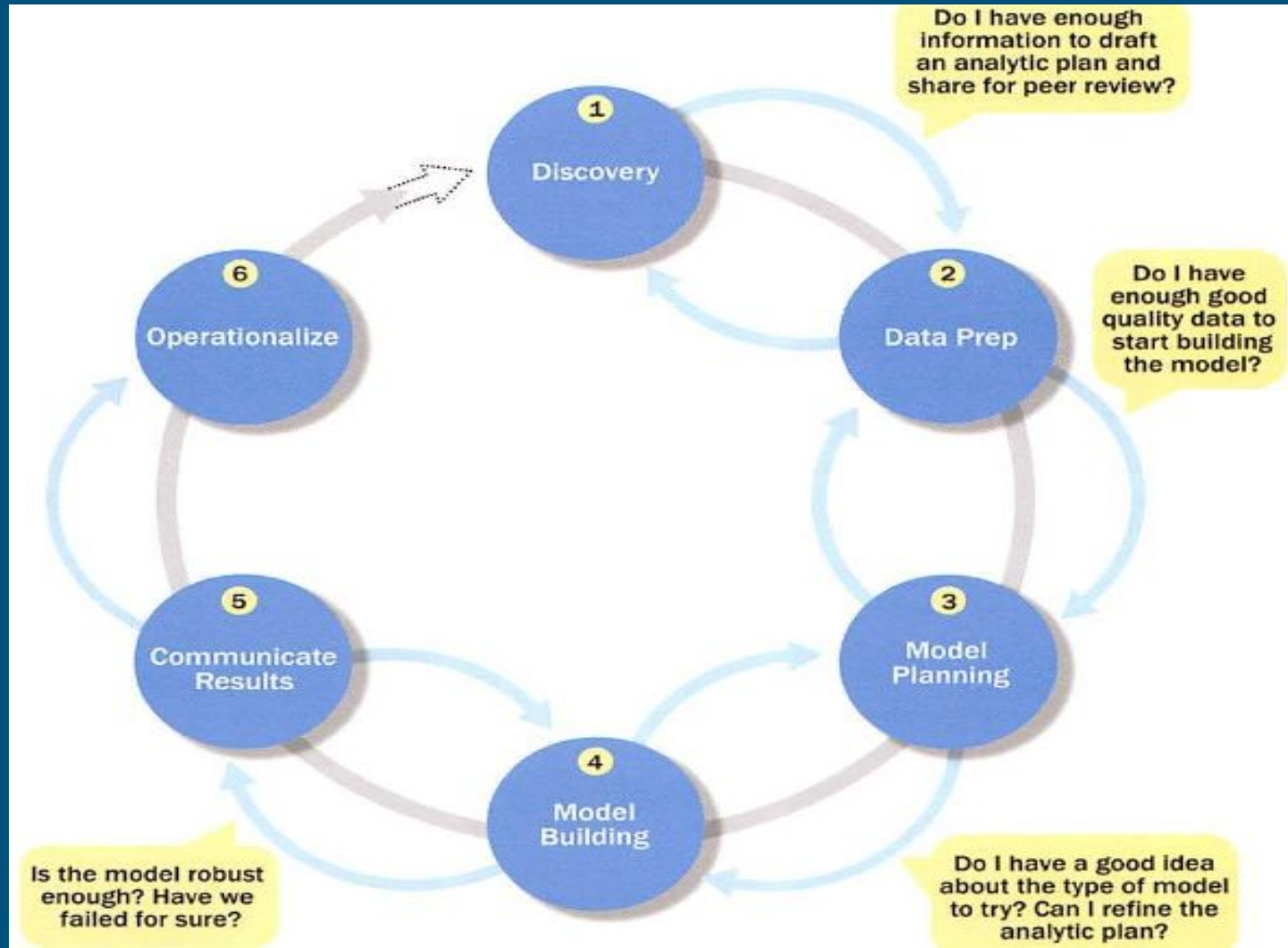
Phase 3: Model Planning

Phase 4: Model Building

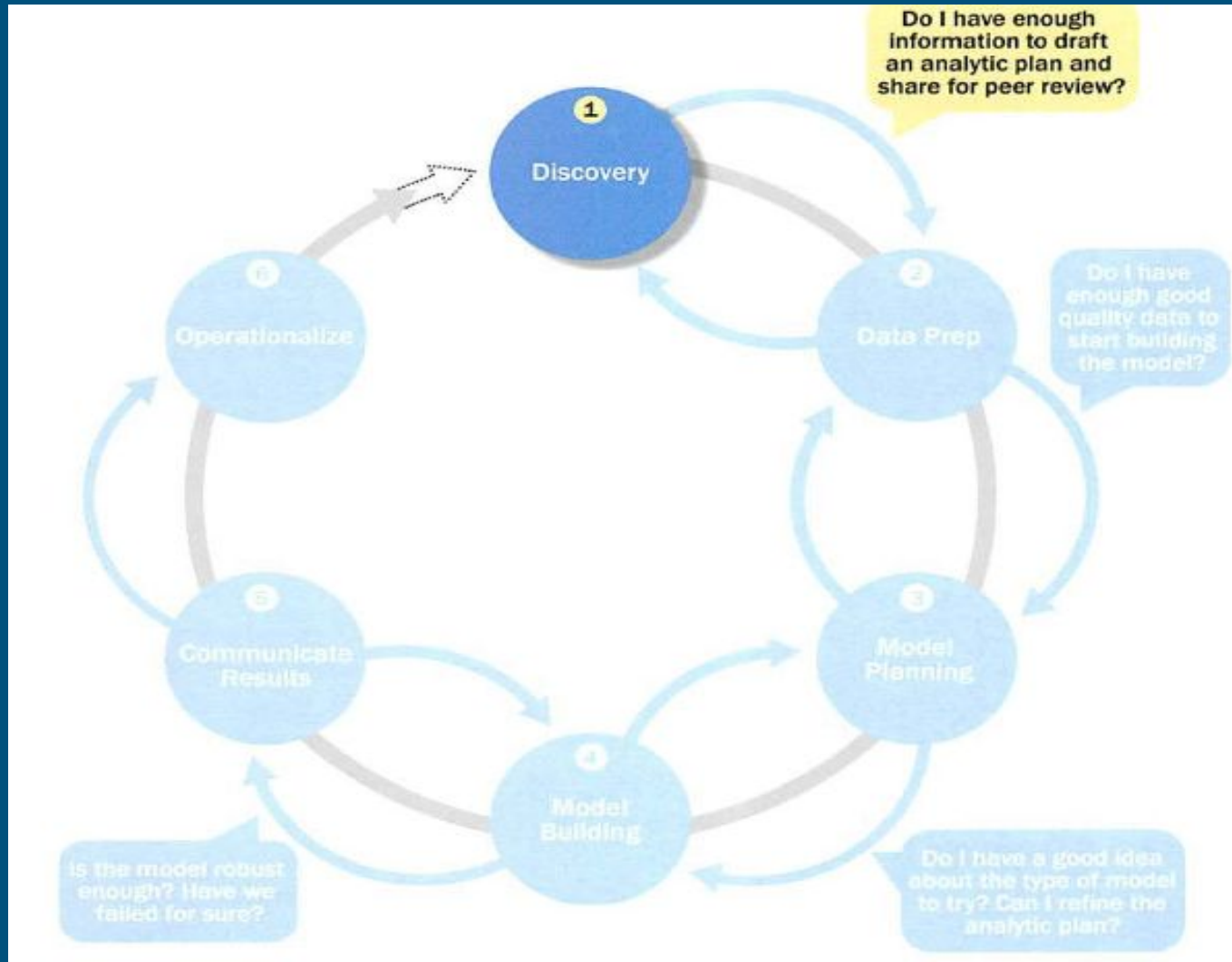
Phase 5: Communicate Results

Phase 6: Operationalize

Overview of Data Analytics Lifecycle



Phase 1: Discovery



Phase 1: Discovery

Learning the Business Domain

Resources Available-
Time, People, Tech, data

Framing the Problem

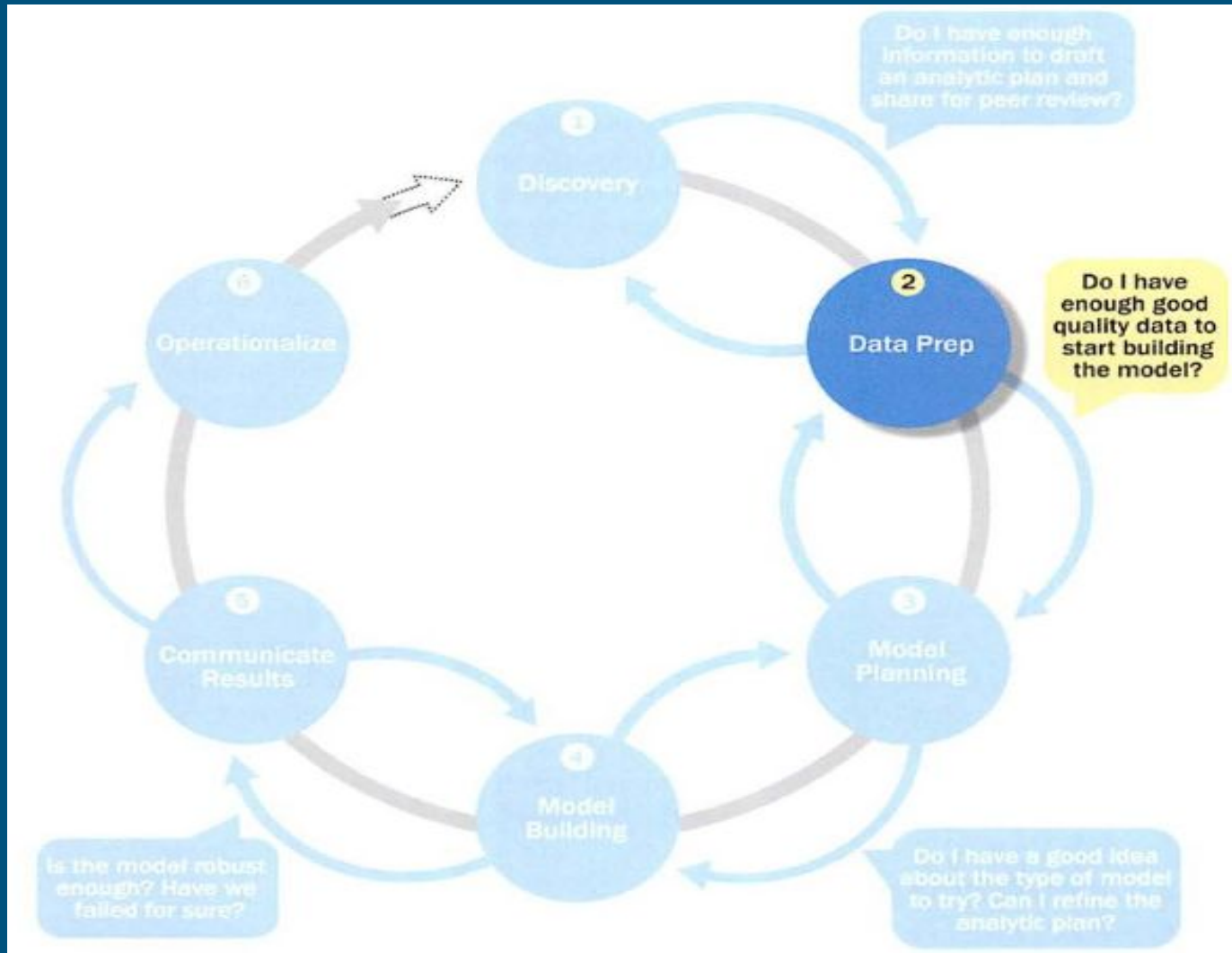
Identifying Key Stakeholders

Interviewing the Analytics Sponsor

Developing Initial Hypotheses

Identifying Potential Data Sources

Phase 2: Data Preparation



Phase 2: Data Preparation

Preparing the Analytic Sandbox

Performing ETLT

Learning about the Data

Data Conditioning

Survey and Visualize

Common Tools for Data Preparation

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
- The sandbox allows organizations to undertake ambitious projects beyond traditional data analysis and BI to perform advanced predictive analytics

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**
- <http://informaticatuts.blogspot.com/2014/07/etl-process-flow.html>
- **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

Outlier



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

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

*Cleaning
data*

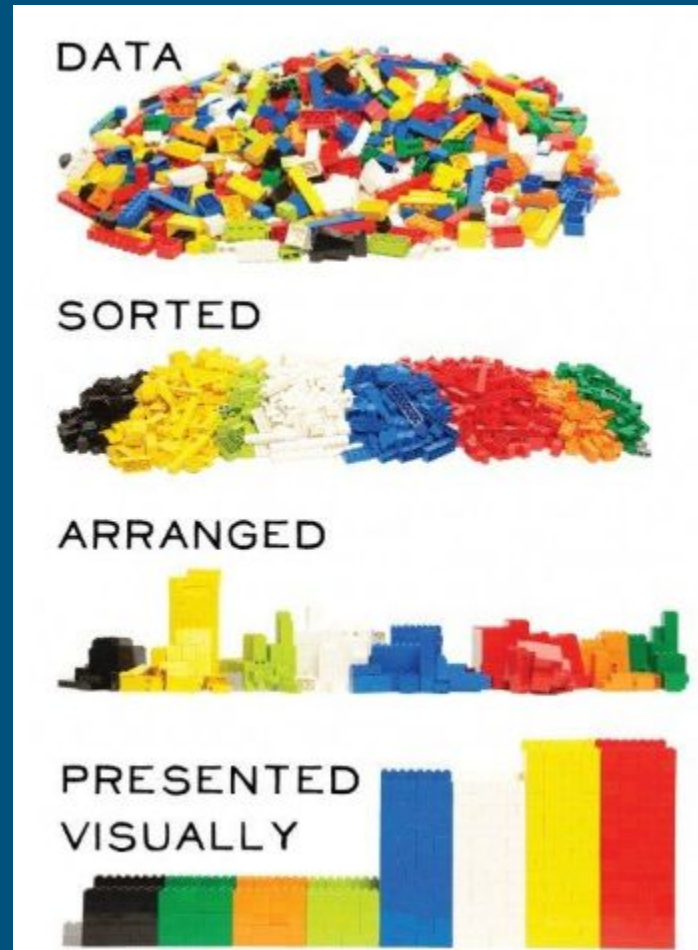
**Managing Missing
data, Outliers, and
Unwanted
Data**

*Normalizing
datasets*

*Performing
transformation*



Survey and Visualize



Survey and Visualize

[https://www.google.com/search?q=data+visualization&tbm=isch&ved=2ahUKEwjx-8TdrJLrAhWbw3MBHfOLDZIQ2-cCegQIABAA&ogq=&gs_lcp=CgNpbWcQAzIHCAAQsQMqqzIECAAQQzIECAAQQzIECAAQQzIECAAQQzIECAAQQzICCAAYAggAUl7AAViOwAFg68YBaABWAHgAgAF8iAF8kgEDMC4xmAEAoAEBqgELZ3dzLXdpei1pbWfAAQE&sclient=img&ei=ViMyX_H4CJuHz7sP85e2kAk&bih=657&biw=1366](https://www.google.com/search?q=data+visualization&tbm=isch&ved=2ahUKEwjx-8TdrJLrAhWbw3MBHfOLDZIQ2-cCegQIABAA&ogq=&gs_lcp=CgNpbWcQAzIHCAAQsQMqqzIECAAQQzIECAAQQzIECAAQQzIECAAQQzIECAAQQzIECAAQQzICCAAYAggAUl7AAViOwAFg68YBaABWAHgAgAF8iAF8kgEDMC4xmAEAoAEBqgELZ3dzLXdpei1pbWfAAQE&sclient=img&ei=ViMyX_H4CJuHz7sP85e2kAk&bih=657&biw=1366)

Survey and Visualize

- Leverage data visualization tools to gain an overview of the data
- **“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
 - https://docs.google.com/spreadsheets/d/1TY_bUVH0x4sl72Snf_WVhPVCDqIvawi3K-VjWwOas4/edit#gid=892363025

Survey and Visualize Guidelines and Considerations

- 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

Perform
parallel ingest
and analysis

**Alpine
Miner**

provides a GUI
for creating
analytic
workflows

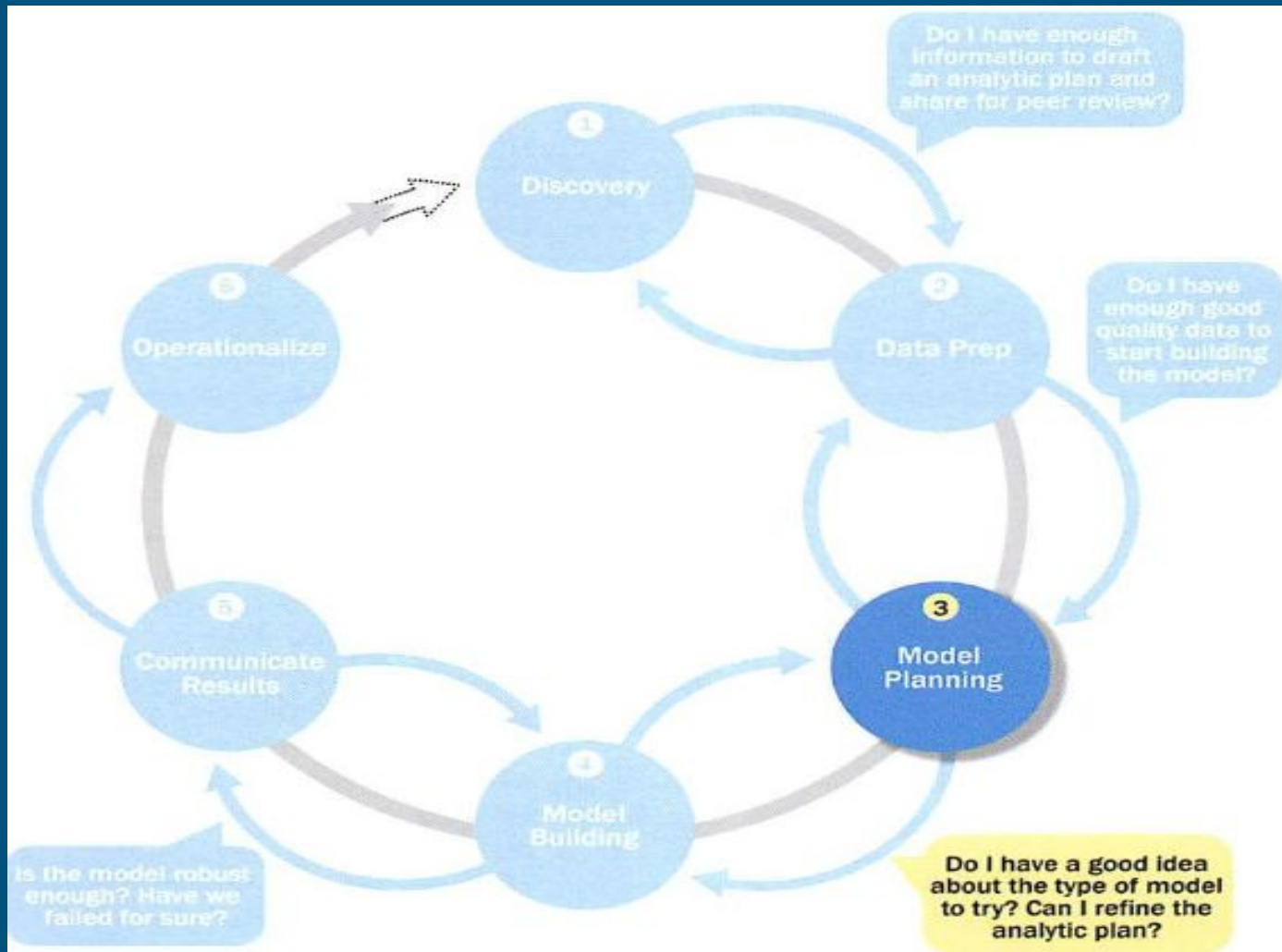
**Open
Refine**

free, open
source tool for
working with
messy data

**Data
Wrangler**

Tool for
data
cleansing &
transformat
ion

Phase 3: Model Planning



Phase 3: Model Planning

Data Exploration & Variable Selection

Model Selection

Common Tools for Model Planning Phase

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

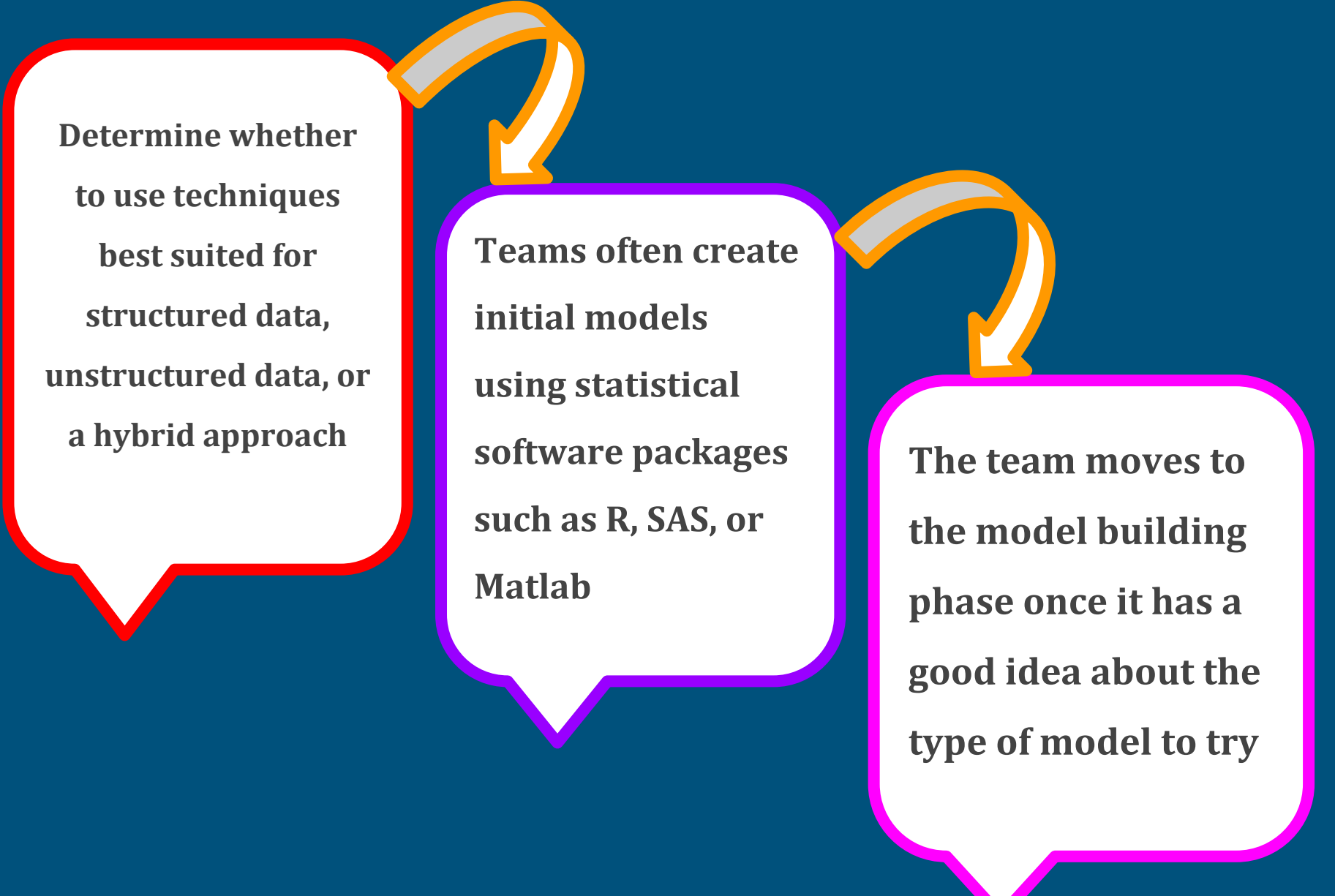
- Example of other analysts approaching a similar problem

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 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

Model Selection



```
graph LR; A[Determine whether to use techniques best suited for structured data, unstructured data, or a hybrid approach] --> B[Teams often create initial models using statistical software packages such as R, SAS, or Matlab]; B --> C[The team moves to the model building phase once it has a good idea about the type of model to try];
```

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

Common Tools for the Model Planning Phase

R

R contains about 5000 packages for data analysis and graphical presentation

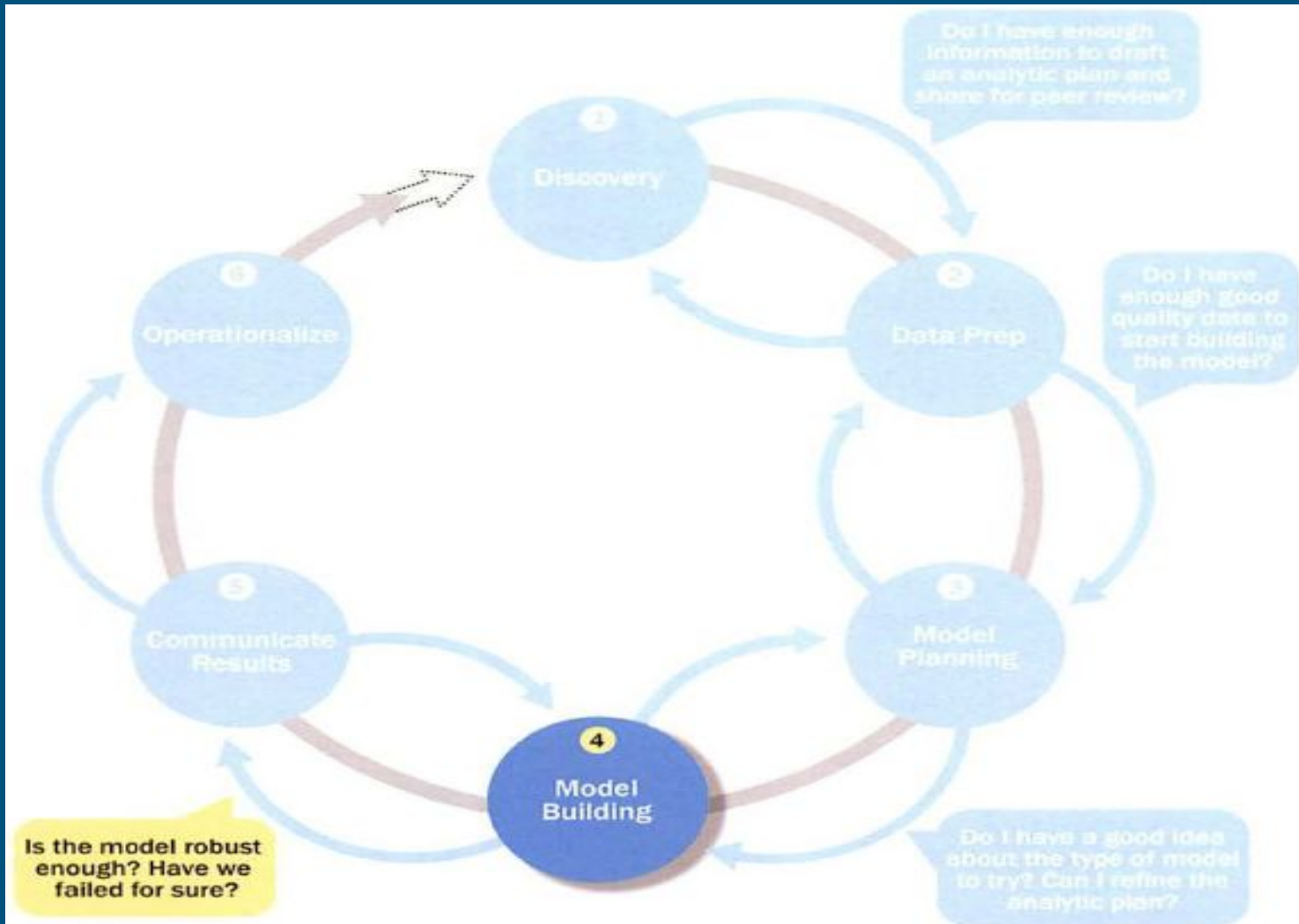
SQL Analysis services

Can perform in-database analytics of common data mining functions, involved aggregations, and basic predictive models

SAS/ ACCESS

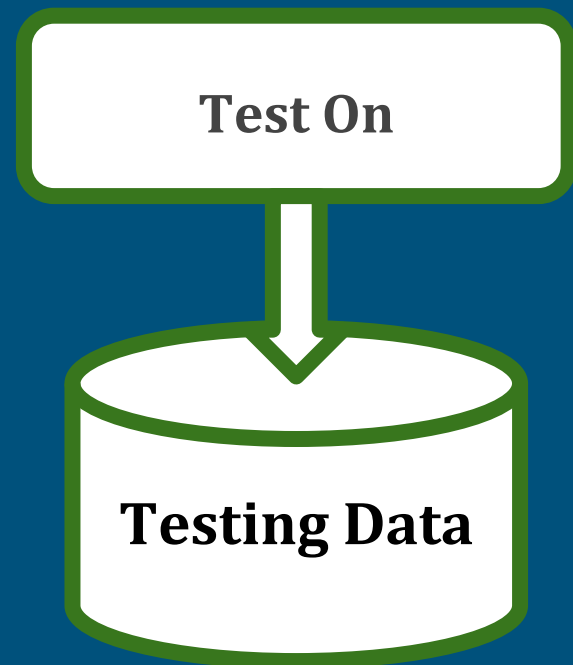
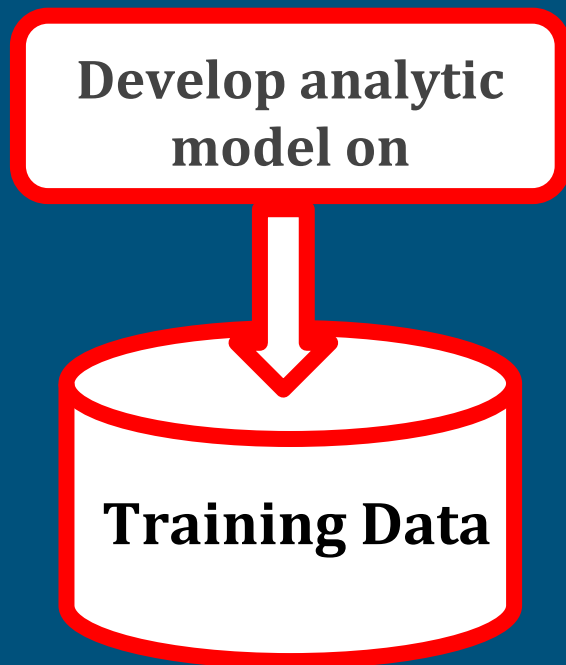
Provides integration between SAS and the analytics sandbox

Phase 4: Model Building



Phase 4: Model Building

- Execute the models defined in Phase 3
- Develop datasets for training, testing, and production



Phase 4: Model Building

- 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?
- 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

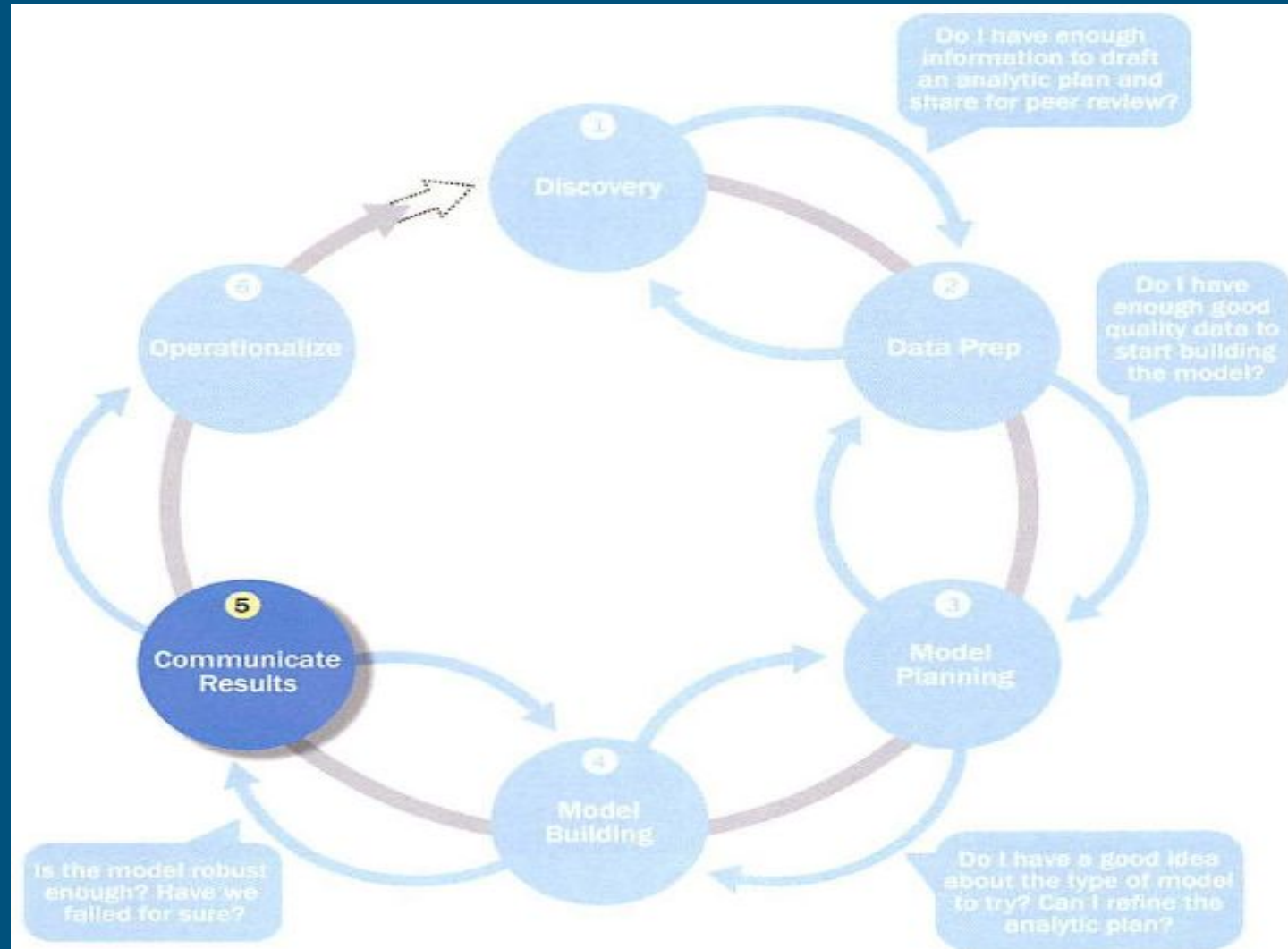
- SAS Enterprise Miner – built for enterprise-level computing and analytics
- SPSS Modeler (IBM) – provides enterprise-level computing and analytics
- Matlab – high-level language for data analytics, algorithms, data exploration
- Alpine Miner – provides GUI frontend for backend analytics tools
- STATISTICA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

Common Tools for the Model Building Phase

● 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



Phase 5: Communicate Results

- Assess if the results are statistically significant and valid

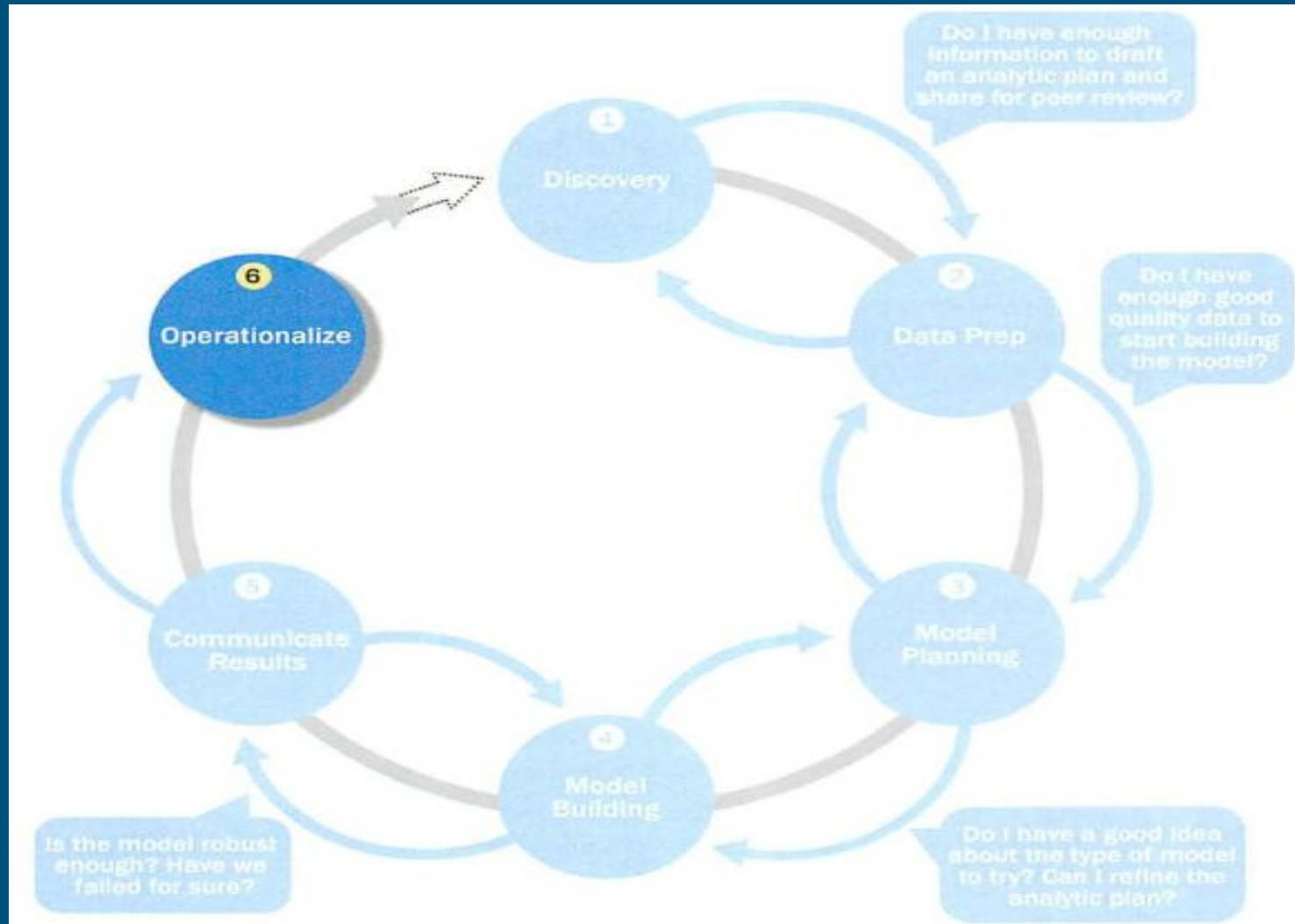


Phase 5: Communicate Results

- Communicate and document the key findings and major insights derived from the analysis



Phase 6: Operationalize

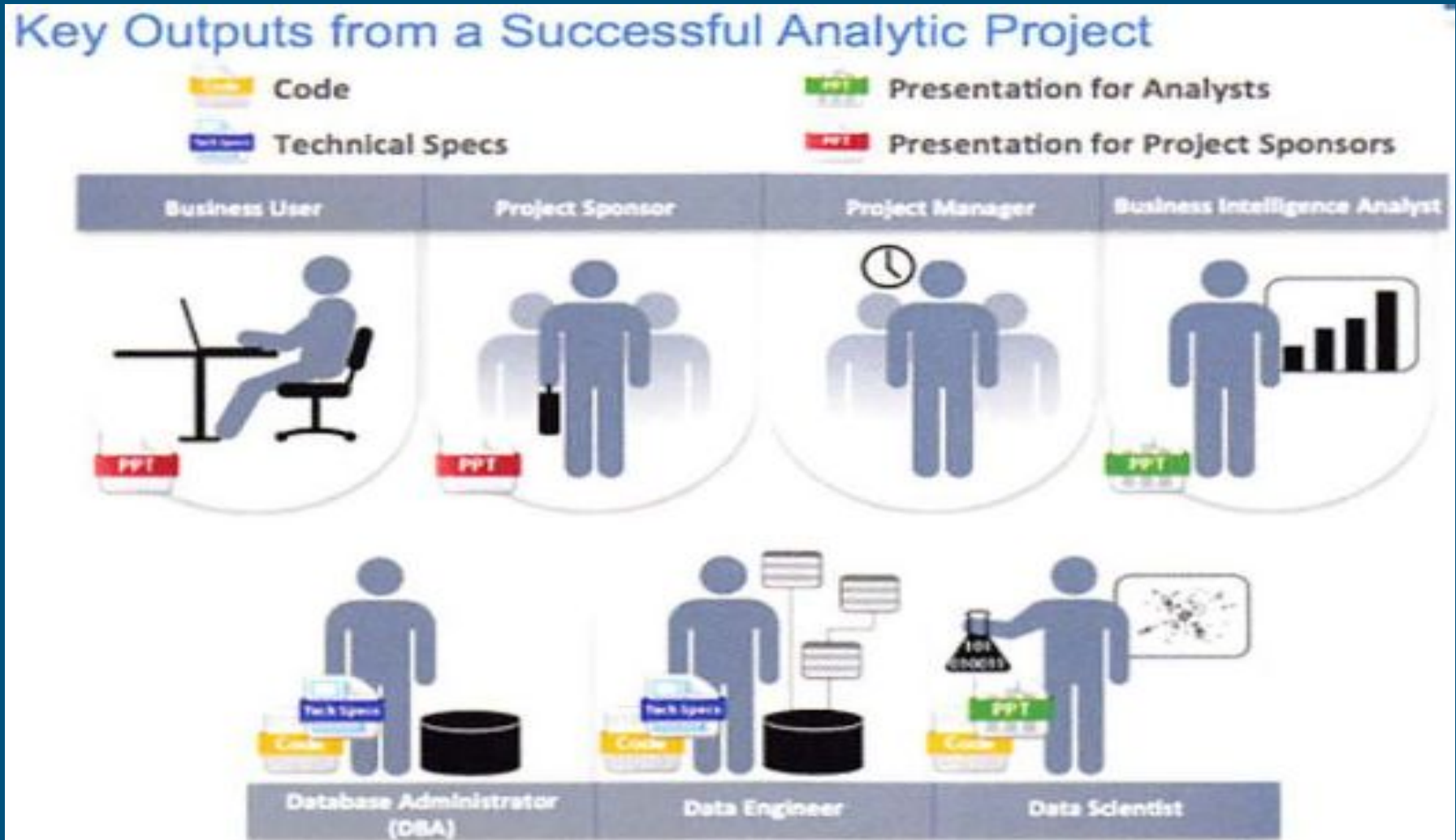


Phase 6: Operationalize

- 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

Phase 6: Operationalize

Key outputs from successful analytics project



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

- 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)

- Students will perform this case study present it and upload PPT of case study in google classroom upto coming Monday 17/8/2020 .

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

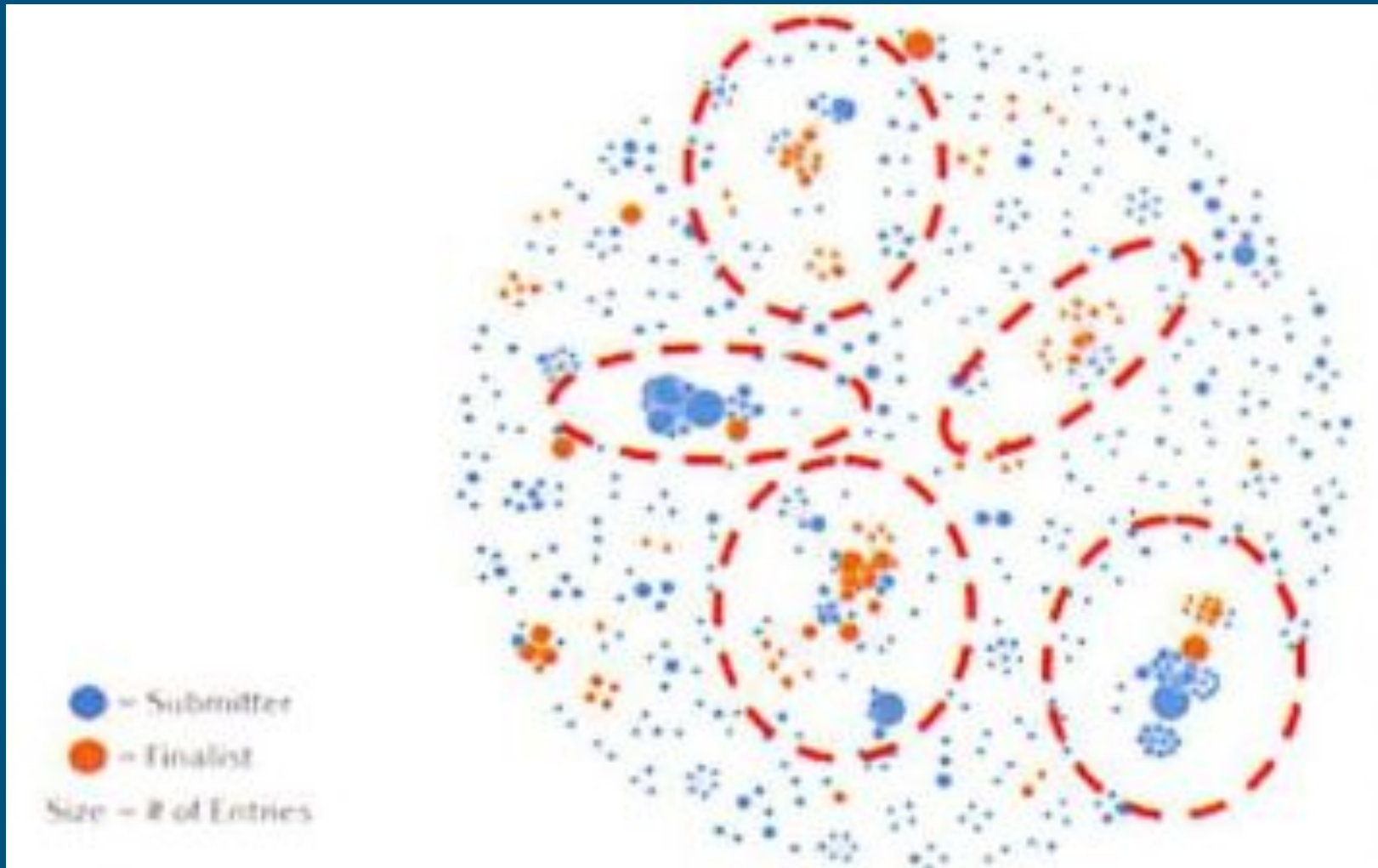
- The study included the following considerations
 - Identify the right milestones to achieve the goals
 - Trace how people move ideas from each milestone toward the goal
 - Track ideas that die and others that reach the goal
 - Compare times and outcomes using a few different methods

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

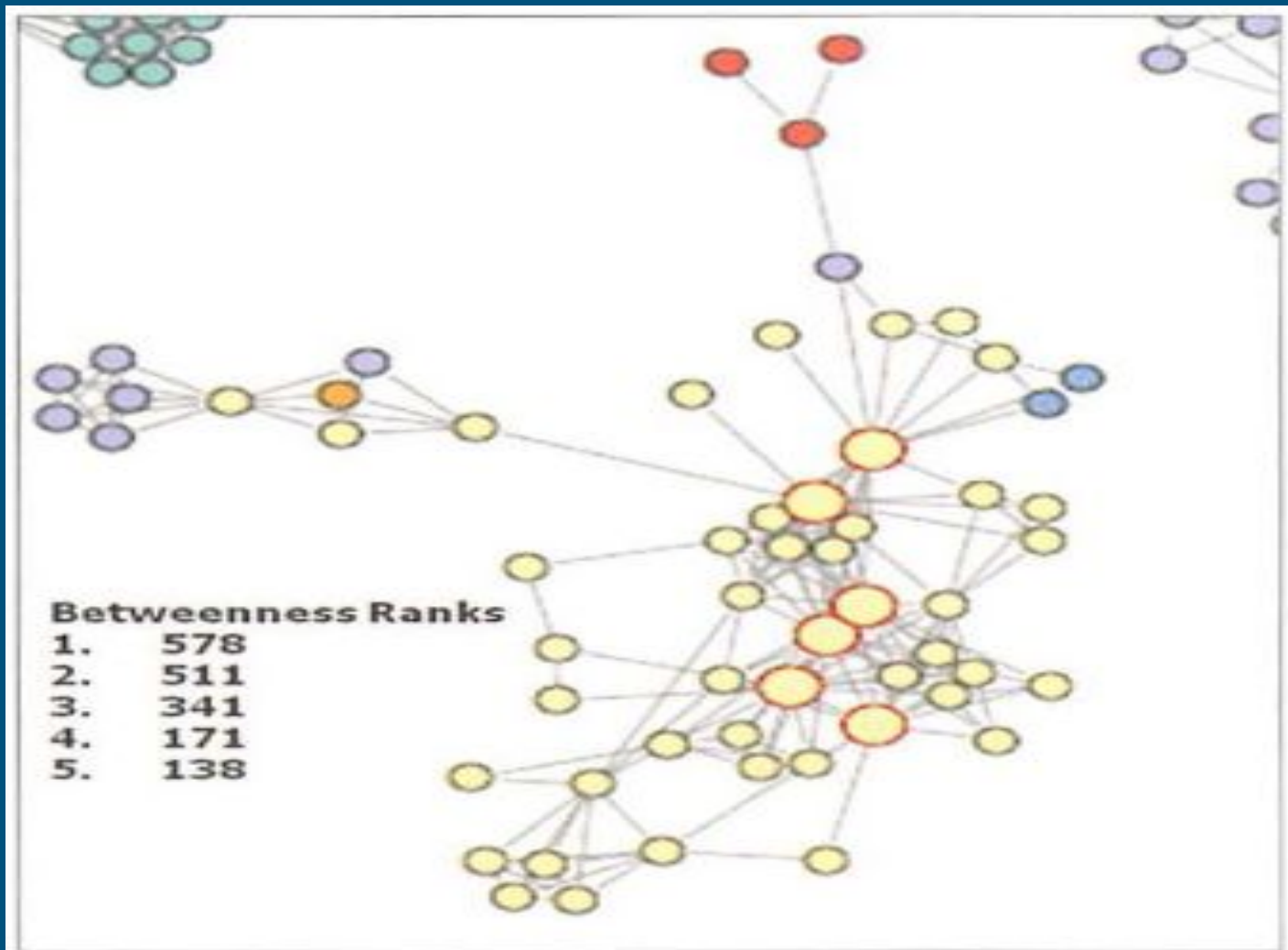
Phase 4: Model Building

Social graph of data submitters and finalists



Phase 4: Model Building

Social graph of top innovation influencers



Communicate Results

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

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

Summary

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

References

- <http://www.csis.pace.edu/~ctappert/cs816-15fall/slides/>
- <https://norcalbiostat.github.io/ADS/notes/Data%20Analytics%20Lifecycle%20-%20EH1.pdf>
- <http://srmnotes.weebly.com/it1110-data-science--big-data.html>
- <http://www.csis.pace.edu/~ctappert/cs816-15fall/books/2015DataScience&BigDataAnalytics.pdf>