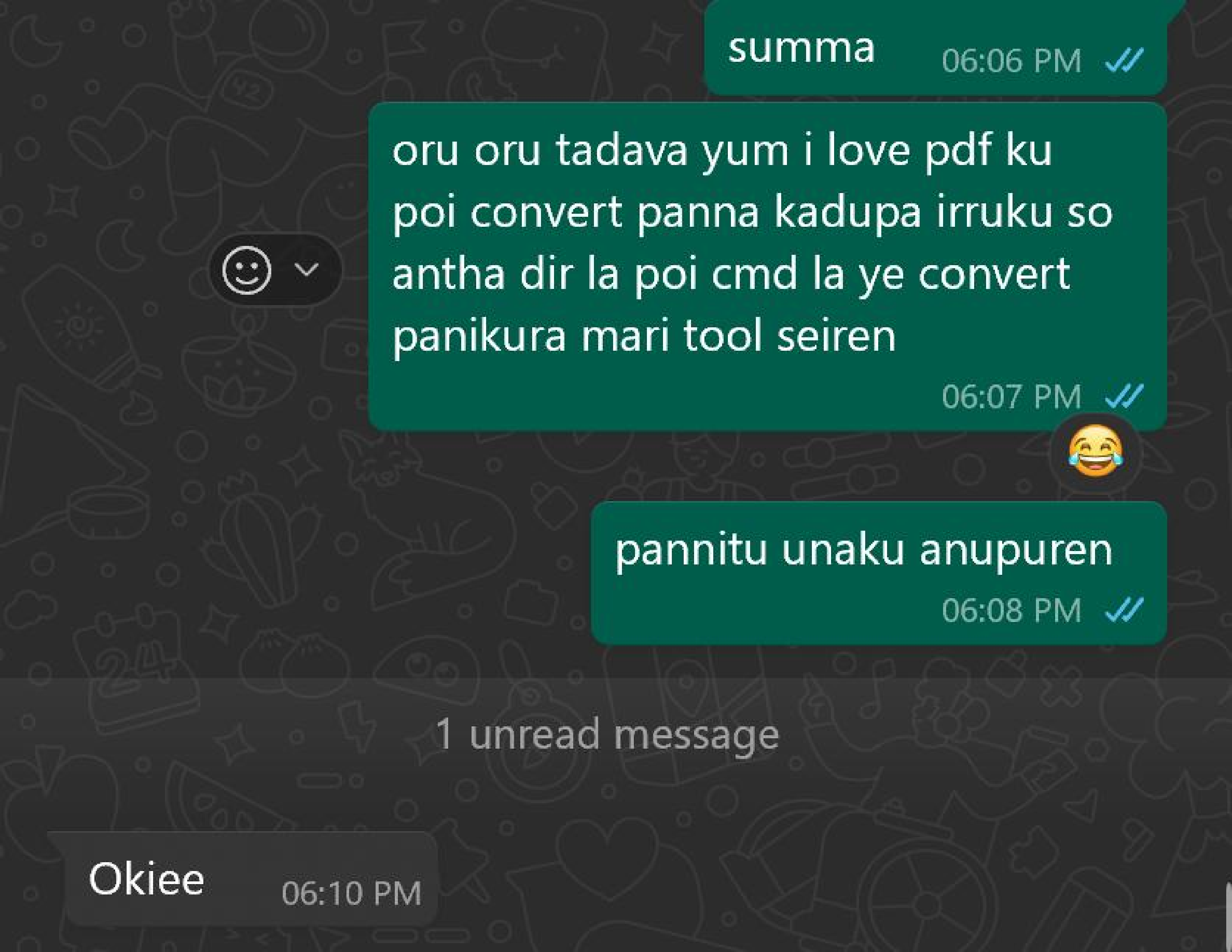


summma 06:06 PM

oru oru tadava yum i love pdf ku poi convert panna kadupa irruku so antha dir la poi cmd la ye convert panikura mari tool seiren 06:07 PM

😂

pannitu unaku anupuren 06:08 PM

1 unread message

Okiee 06:10 PM

vanakam

Data Science and Analytics

U20ITT614

UNIT I

DATA ANALYTICS USING R

- Big Data Overview
- Big Data Vs Data Science
- Examples of Big Data Analytics
- Data Analytics Lifecycle overview
- Phases in the lifecycle
- GINA Case Study
- Introduction to R programming
- Exploratory Data Analysis
- Statistical Methods for Evaluation

DATA

Data is one of the prime factors of any business purpose.

The quantities, characters, or symbols on which operations are performed by a computer

Big Data

Big Data is a term that is used for denoting a collection of datasets that is large and complex, making it very difficult to process using legacy data processing applications.

500+terabytes

10+terabytes of
data in **30 minutes**

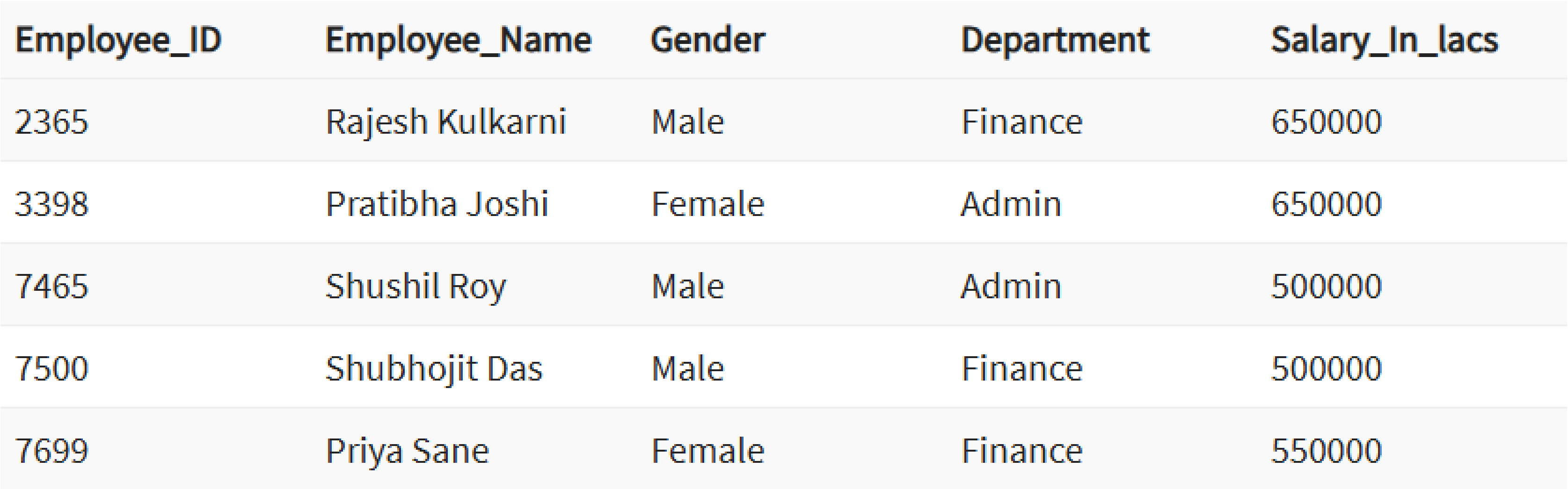
New York Stock Exchange - generates about *one terabyte* of new trade data per day.

Types of Big Data

Big Data includes huge volume, high velocity, and extensible variety of data. Three types.

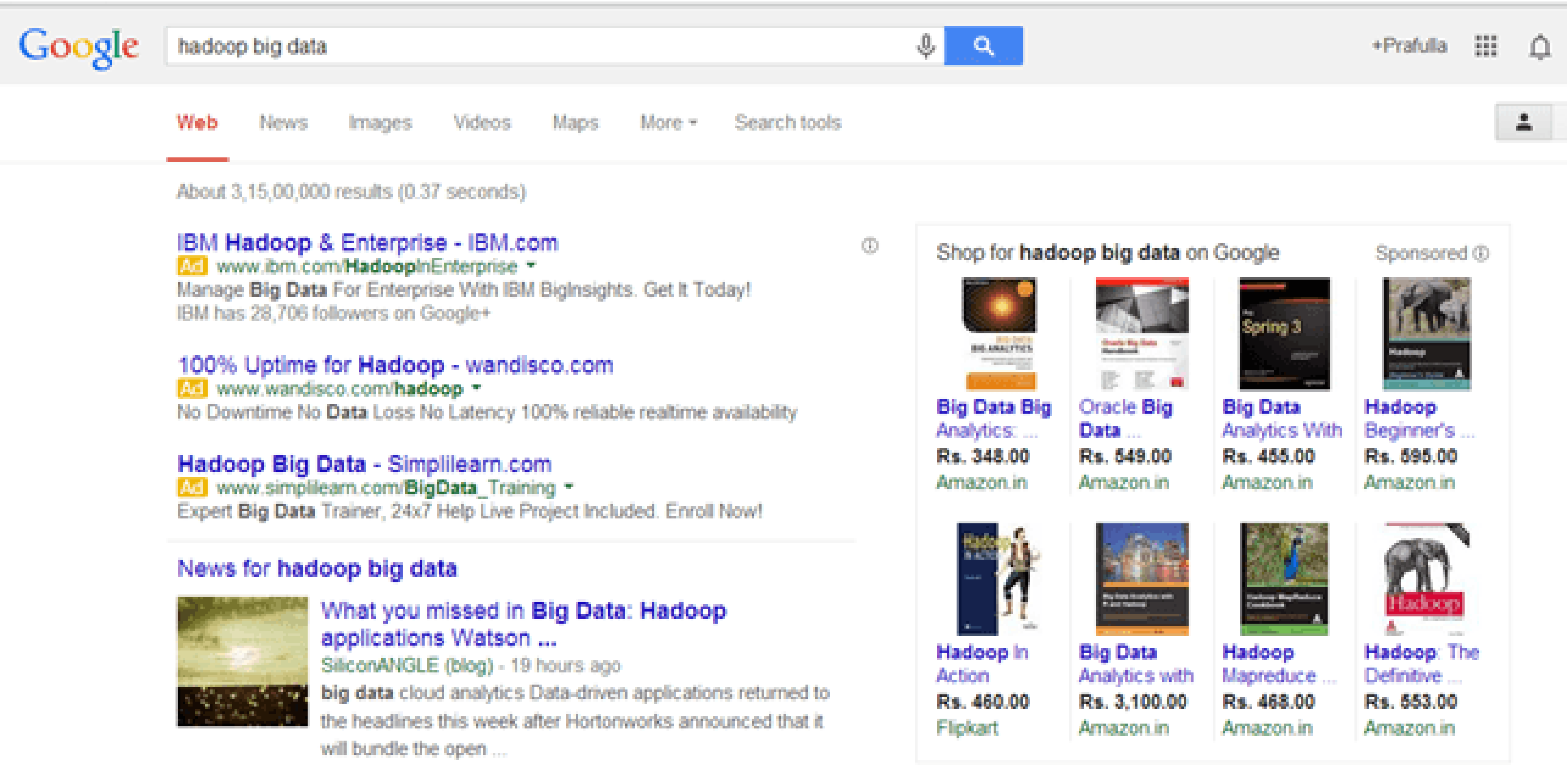
- **Structured Data** – Relational data.
- **Unstructured Data** – Word, PDF, Text, Media Logs.
- **Semi Structured Data** – XML data.

Structured Data



Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
2365	Rajesh Kulkarni	Male	Finance	650000
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushil Roy	Male	Admin	500000
7500	Shubhojit Das	Male	Finance	500000
7699	Priya Sane	Female	Finance	550000

Unstructured Data



About 3,15,00,000 results (0.37 seconds)

IBM Hadoop & Enterprise - IBM.com
Ad www.ibm.com/**Hadoop**InEnterprise
Manage **Big Data** For Enterprise With IBM BigInsights. Get It Today!
IBM has 28,706 followers on Google+

100% Uptime for Hadoop - wandisco.com
Ad www.wandisco.com/**hadoop**
No Downtime No **Data** Loss No Latency 100% reliable realtime availability

Hadoop Big Data - Simplilearn.com
Ad www.simplilearn.com/**BigData**_Training
Expert **Big Data** Trainer, 24x7 Help Live Project Included. Enroll Now!

News for hadoop big data

What you missed in Big Data: Hadoop applications Watson ...
SiliconANGLE (blog) - 19 hours ago
big data cloud analytics Data-driven applications returned to the headlines this week after Hortonworks announced that it will bundle the open ...

Shop for **hadoop big data** on Google — Sponsored

Big Data Big Analytics: ...
Rs. 348.00
Amazon.in

Oracle **Big Data** ...
Rs. 549.00
Amazon.in

Big Data Analytics With
Rs. 455.00
Amazon.in

Hadoop Beginner's ...
Rs. 595.00
Amazon.in

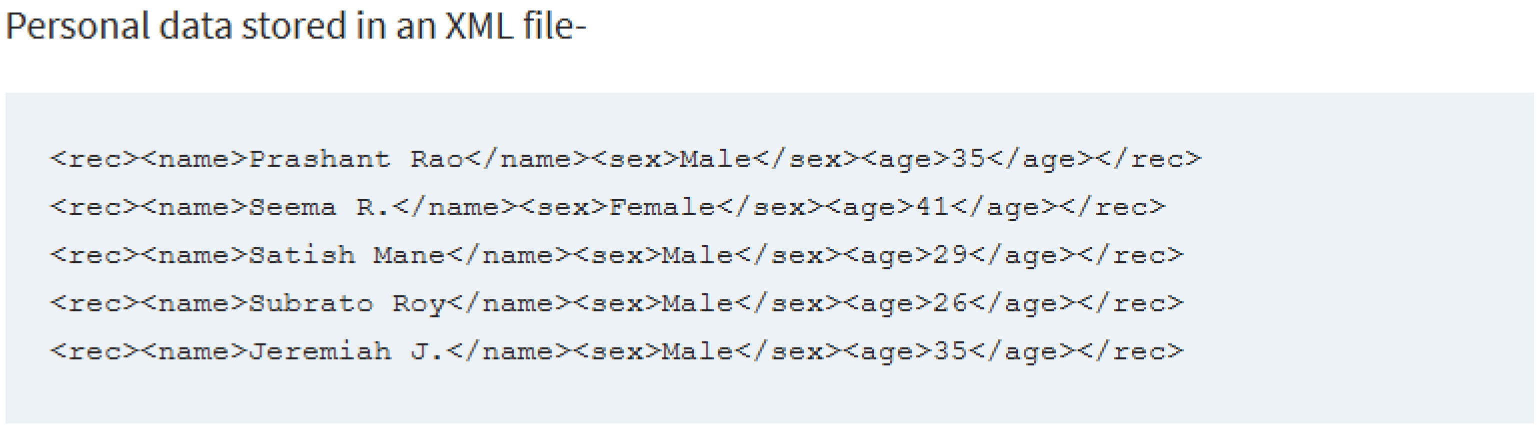
Hadoop In Action
Rs. 460.00
Flipkart

Big Data Analytics with
Rs. 3,100.00
Amazon.in

Hadoop Mapreduce ...
Rs. 468.00
Amazon.in

Hadoop: The Definitive ...
Rs. 553.00
Amazon.in

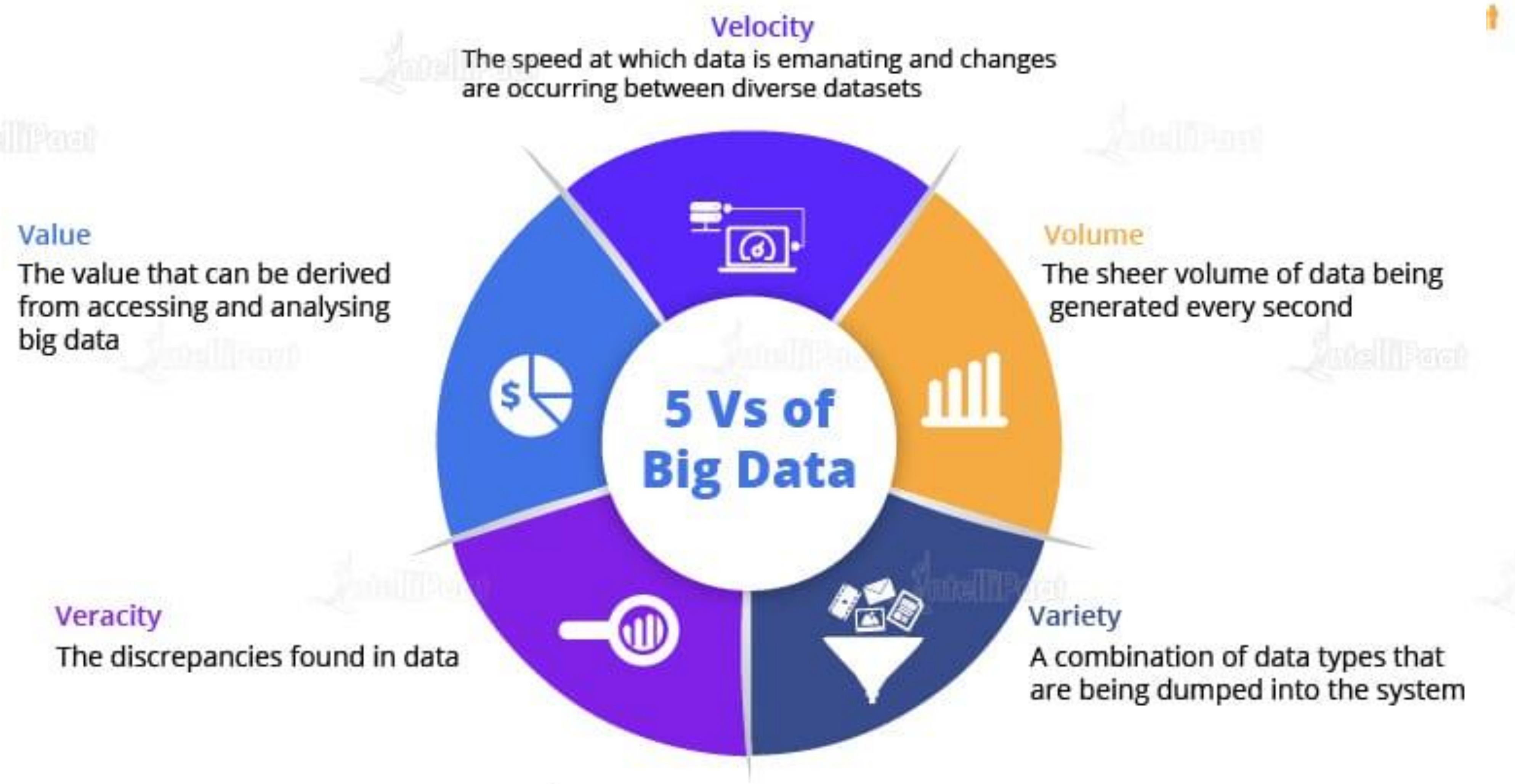
Semi Structured Data



Personal data stored in an XML file-

```
<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>
<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>
<rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>
<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>
<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>
```

Big Data - Characteristics



5 Vs of Big Data

Velocity
The speed at which data is emanating and changes are occurring between diverse datasets

Volume
The sheer volume of data being generated every second

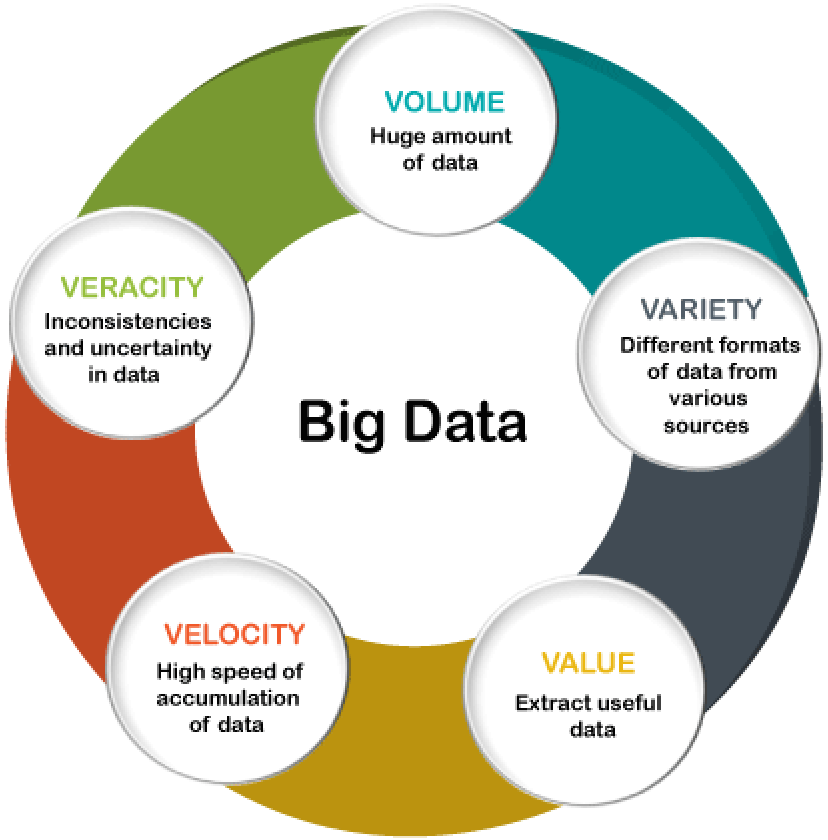
Variety
A combination of data types that are being dumped into the system

Veracity
The discrepancies found in data

Value
The value that can be derived from accessing and analysing big data

Characteristics of Big Data

- Volume
- Veracity
- Variety
- Value
- Velocity



Big Data

VOLUME
Huge amount of data

VARIETY
Different formats of data from various sources

VALUE
Extract useful data

VELOCITY
High speed of accumulation of data

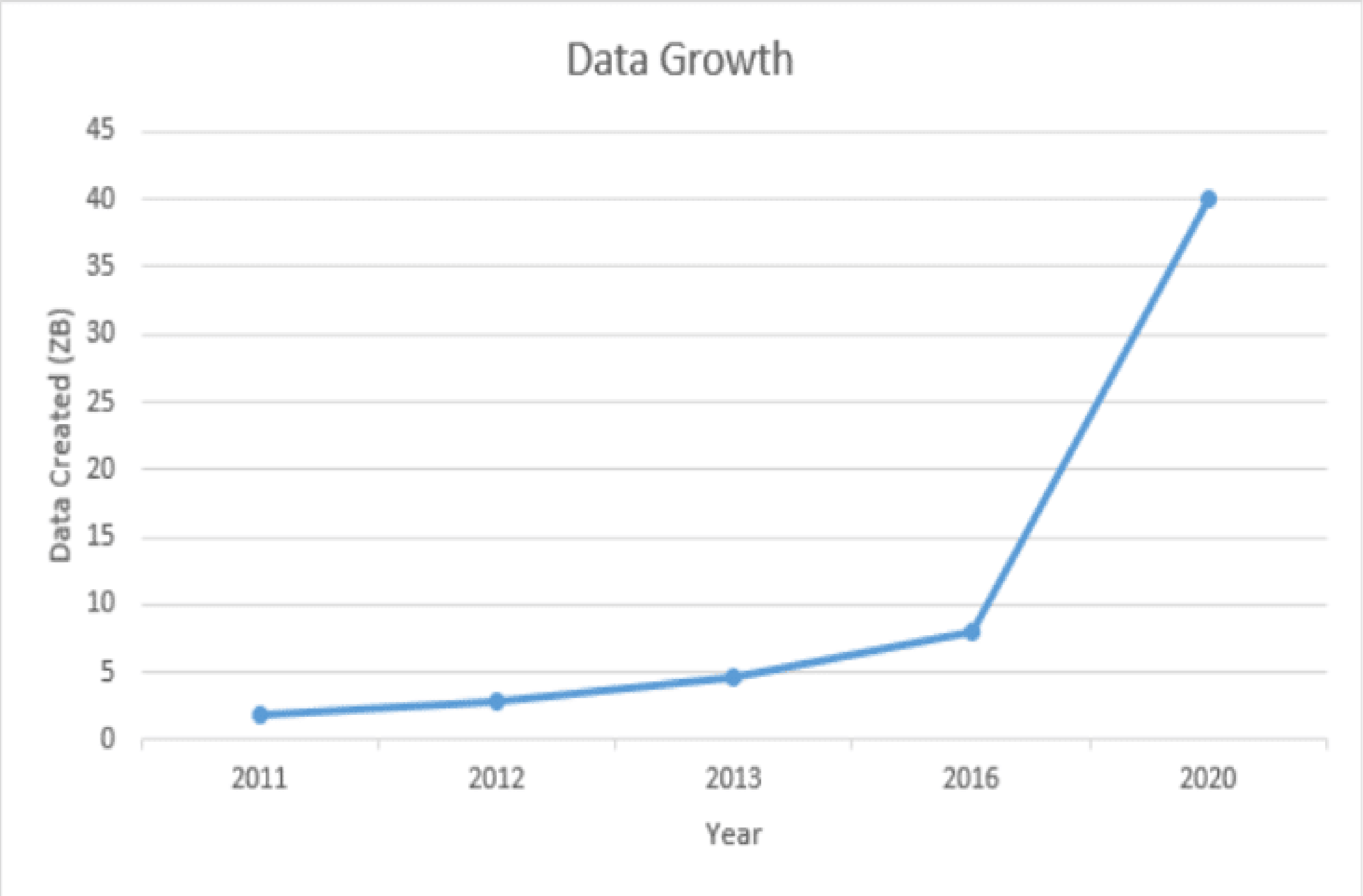
VERACITY
Inconsistencies and uncertainty in data

Volume

Large Data

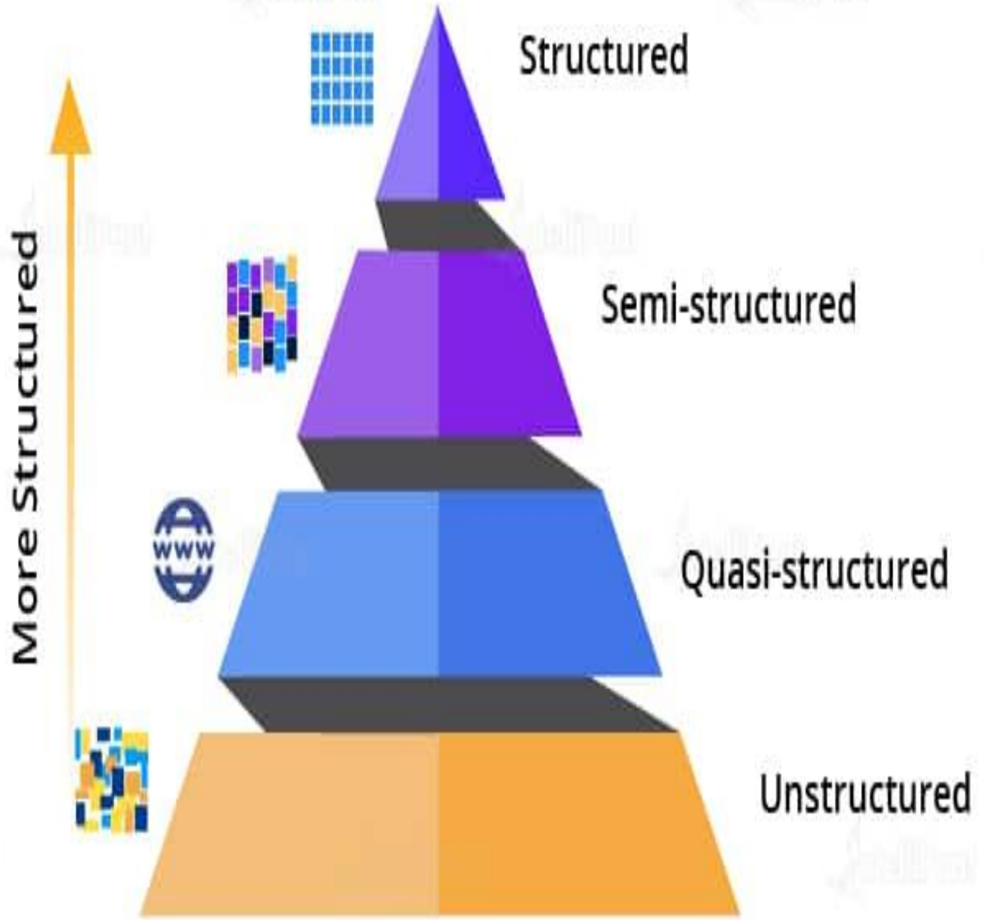
2016 –

2020 -



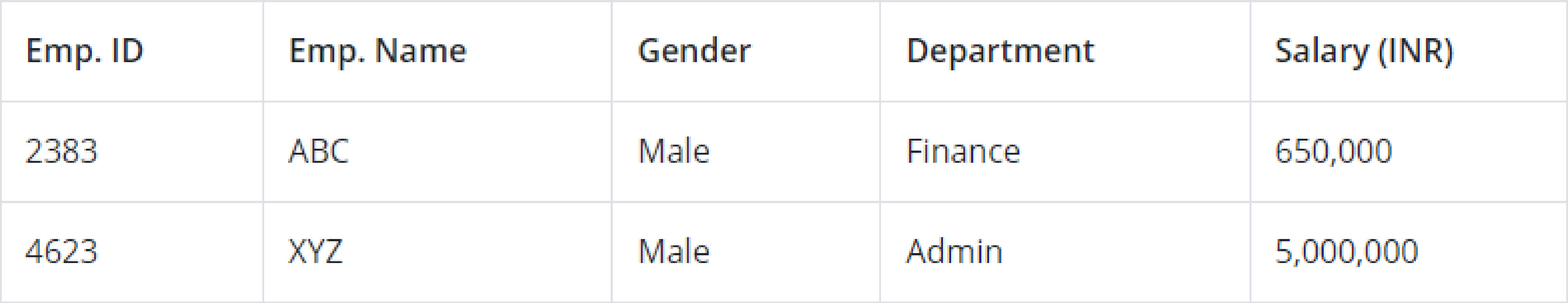
Variety

Data is coming from different sources in various formats.



Structured

Data is present in a structured schema along with all the required columns. It is in a structured or tabular format.



Emp. ID	Emp. Name	Gender	Department	Salary (INR)
2383	ABC	Male	Finance	650,000
4623	XYZ	Male	Admin	5,000,000

Semi-structured Data

The schema is not properly defined, i.e., both forms of data are present

JSON, XML, CSV, email..

```
{
"employees":[
    {"firstName":"John", "lastName":"Doe"},
    {"firstName":"Anna", "lastName":"Smith"},
    {"firstName":"Peter", "lastName":"Jones"}
]
}
```

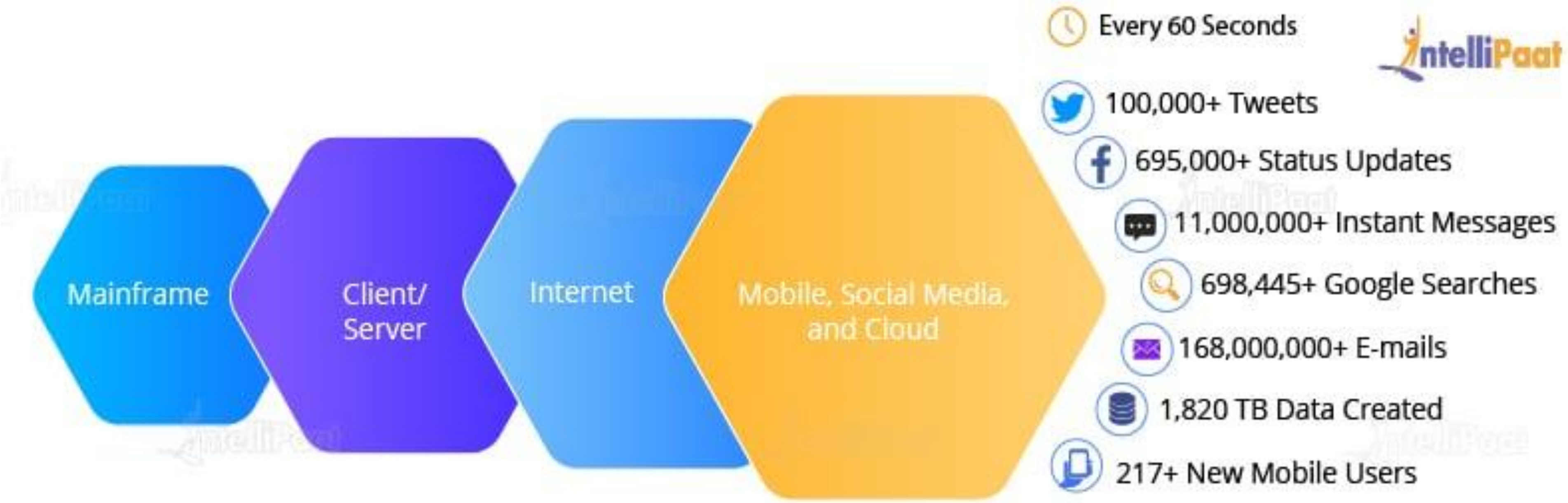
Unstructured Data

video files, log files, audio files, and image files

unstructured data possesses various challenges in terms of processing for deriving value out of it.

Velocity

The speed of data accumulation also plays a role in determining whether the data is big data or normal data.



Value

It deals with a mechanism to bring out the correct meaning of data

The process to turn raw data into useful data.

Then, an analysis is done on the data that are cleaned or retrieved from the raw data

Veracity

Veracity means how much the data is reliable

Veracity is the process of being able to handle and manage data efficiently

There will also be uncertainties and inconsistencies in the data that can be overcome by veracity

Veracity means the trustworthiness and quality of data

Big Data Analytics

	Traditional Analytics (BI)	vs	Big Data Analytics
Focus on	• Descriptive analytics • Diagnosis analytics		• **Predictive analytics** • **Data Science**
Data Sets	• Limited data sets • Cleansed data • Simple models		• Large scale data sets • More types of data • Raw data • Complex data models
Supports	**Causation:** what happened, and why?		**Correlation**: new insight More accurate answers

Big Data Analytics

Big Data Analytics examines large and different types of data to uncover hidden patterns, insights, and correlations.

Big Data Analytics is helping large companies

1. To facilitate their growth and development.
2. To identify opportunities for improvement and optimization.
3. To reduce costs and develop better, customer-centric products and services.

Benefits of big data analytics

Cost Reduction - Big data reduce costs in storing all the business data in one place. Tracking analytics also helps companies find ways to work more efficiently to cut costs wherever possible.

Product Development - Developing and marketing new products, services, or brands is much easier when based on data collected from customers' needs and wants.

Strategic business decisions: The ability to constantly analyze data helps businesses make better and faster decisions, such as cost and supply chain optimization.

Health care: Monitoring patients' medical histories helps doctors detect and prevent diseases.

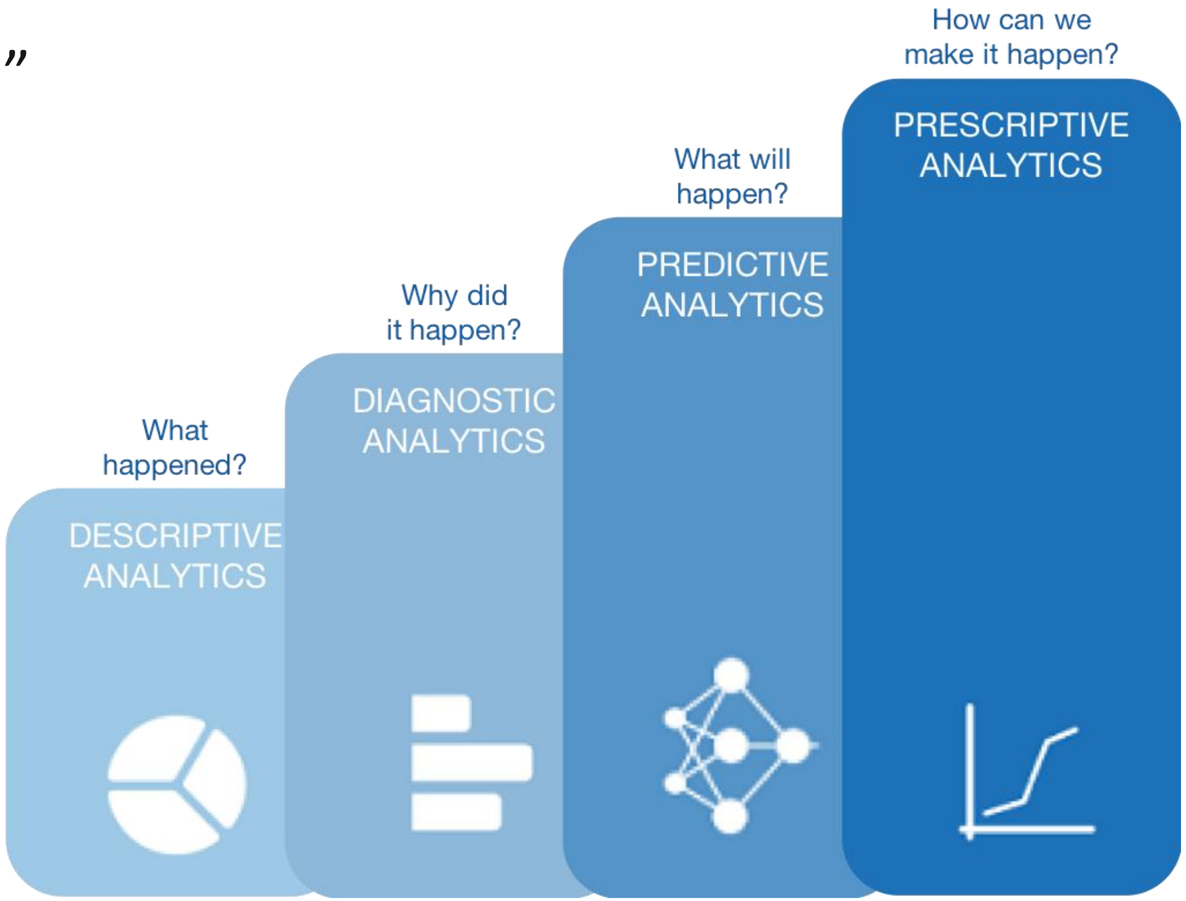
Government: Big data can be used to collect data from CCTV and traffic cameras, satellites, body cameras and sensors, emails, calls, and more, to help manage the public sector.

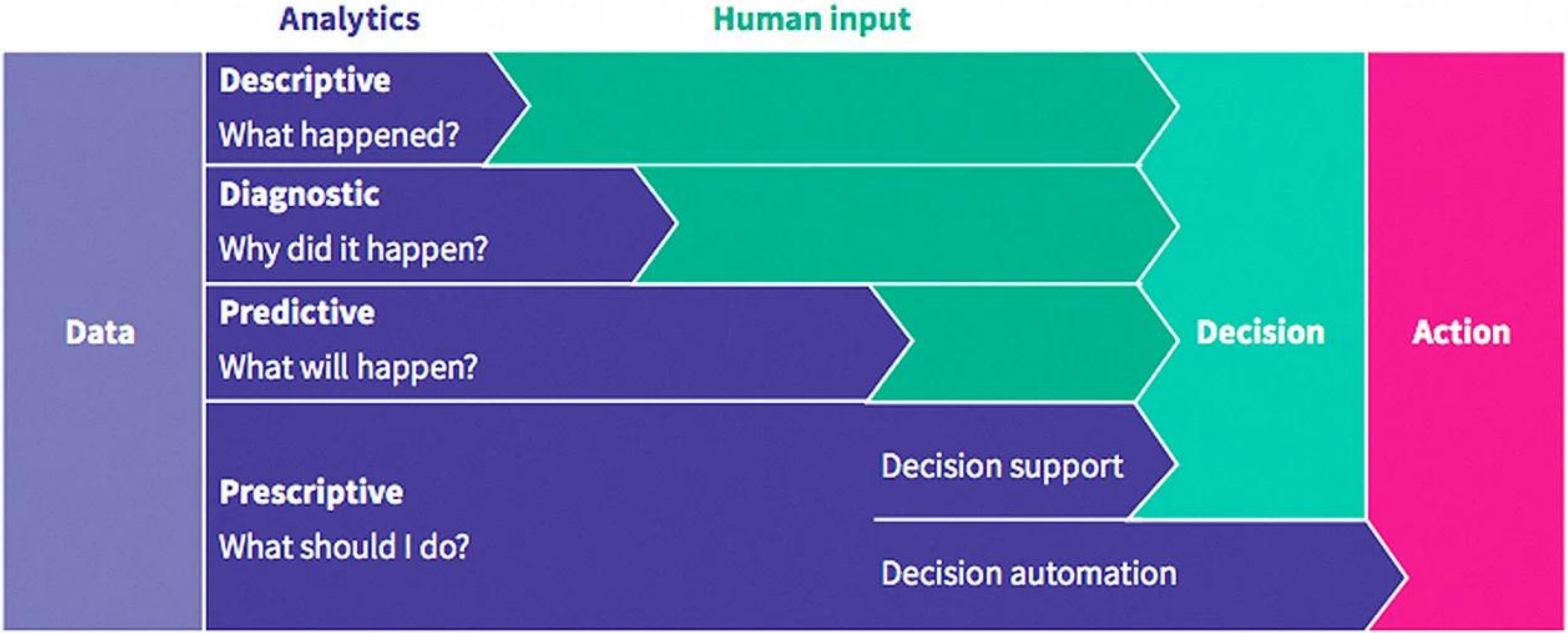
Types of Big Data Analytics

- Descriptive Analytics
- Diagnostic Analytics
- Predictive Analytics
- Prescriptive Analytics

Types of Big Data Analytics

- Descriptive - “What happened?”
- Diagnostic - “Why did this happen?”
- Predictive - “What might happen in the future?”
- Prescriptive - “What should we do next?”




Analytics
Human input
Data
Descriptive
What happened?
Diagnostic
Why did it happen?
Predictive
What will happen?
Prescriptive
What should I do?
Decision support
Decision automation
Decision
Action

Types of Big Data Analytics

Descriptive Analytics

- Is a statistical method that is used to search and summarize historical data in order to identify patterns or meaning
- Summarizes past data into a form that people can easily read
- Data aggregation and data mining are two techniques used in descriptive analytics to discover historical data.
- Data is first gathered and sorted by data aggregation in order to make the datasets more manageable by analysts

- Example

  Summarizing the number of times a learner posts in a discussion board

  Tracking assignment and assessment grades

  Comparing pre-test and post-test assessments

Types of Big Data Analytics

Diagnostic Analytics

This is done to understand what caused a problem in the first place

Data mining, and data recovery are all examples

Diagnostics analytics helps companies understand why a problem occurred.

Big data technologies and tools allow users to mine and recover data that helps dissect an issue and prevent it from happening in the future.

Examining Market Demand

Explaining Customer Behavior

Types of Big Data Analytics

Predictive Analytics

This type of analytics looks into the historical and present data to make predictions of the future.

Predictive analytics uses data mining, AI, and machine learning to analyze current data and make predictions about the future.

It works on predicting customer trends, market trends, and so on.

Types of Big Data Analytics

Prescriptive Analytics

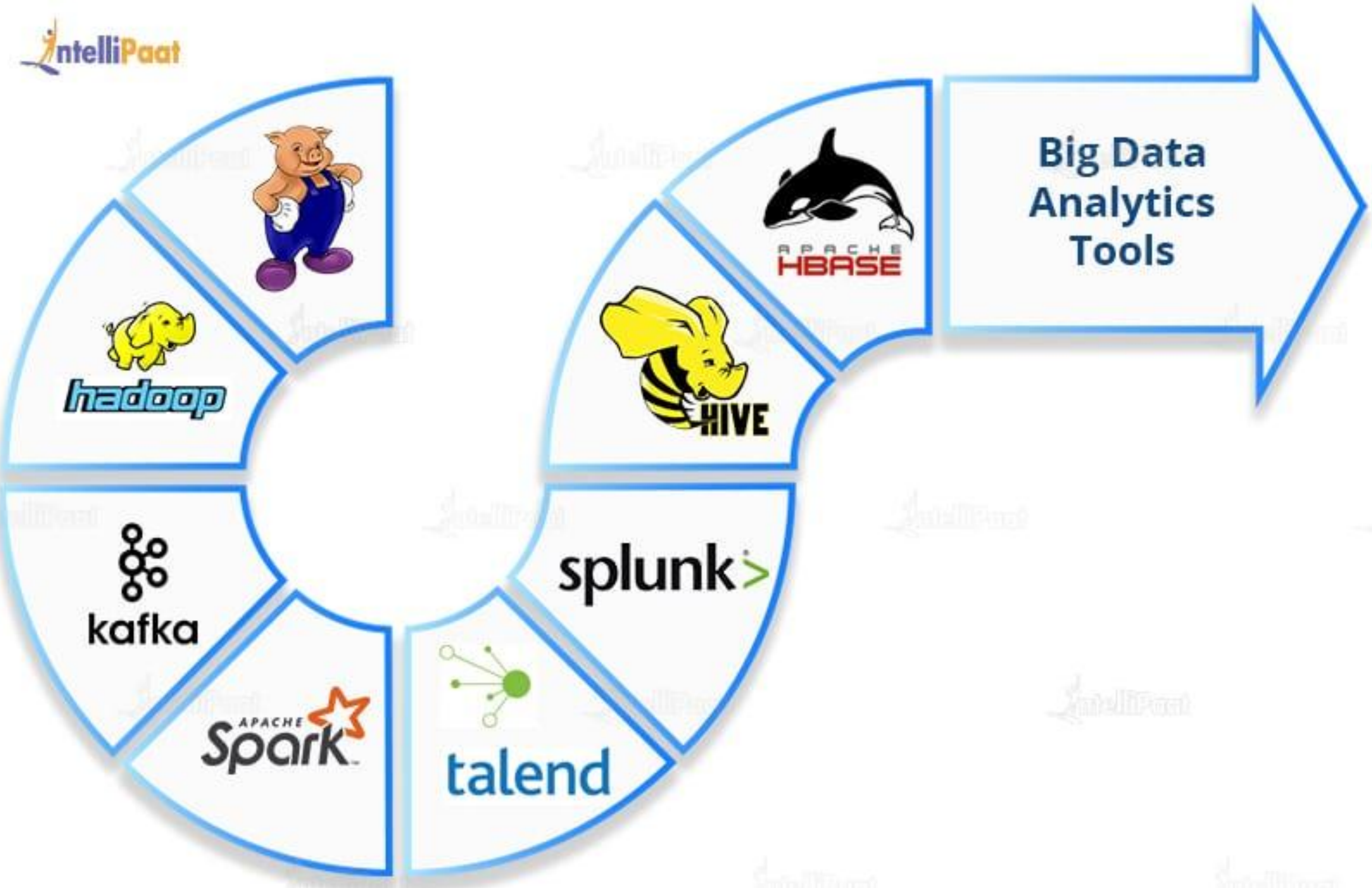
Provides a solution to a particular problem.

Perspective analytics works with both descriptive and predictive analytics.

Relies on AI and machine learning to gather data and use it for risk management.

Tools used in big data analytics

- **Hadoop:** OS - stores and processes big datasets - structured and unstructured data.
- **Spark:** OS cluster computing framework used for real-time processing and analyzing data.
- **Data integration software:** Programs that allow big data to be streamlined across different platforms, such as MongoDB, Apache, Hadoop, and Amazon EMR.
- **Stream analytics tools:** Systems that filter, aggregate, and analyze data that might be stored in different platforms and formats, such as Kafka.
- **Distributed storage**: Databases that can split data across multiple servers and have the capability to identify lost or corrupt data, such as Cassandra.



Big Data Analytics Tools

Data Analytics Life Cycle

Discovery

Data Preparation

Model Planning

Model Building

Communicate Results

Operationalize

Data Analytics Life Cycle

Discovery

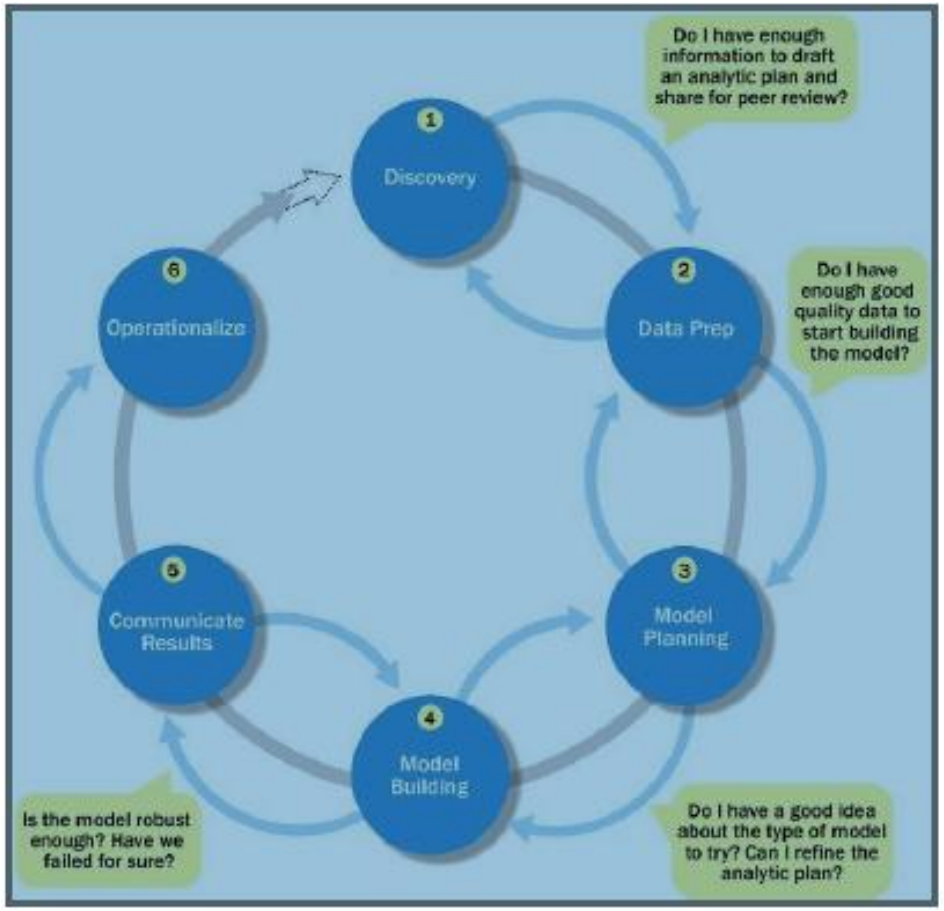
Data Preparation

Model Planning

Model Building

Communicate Results

Operationalize



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?

Phase 1: Discovery

- The data science team learn and investigate the problem.
- Develop context and understanding.
- Come to know about data sources needed and available for the project.
- The team formulates initial hypothesis that can be later tested with data.

Phase 1—Discovery

Team learns the business domain, including relevant history such as whether the organization or business unit has attempted similar projects in the past

Team assesses the resources available to support the project in terms of people, technology, time, and data.

Important activities - include framing the business problem as an analytics challenge that can be addressed in subsequent phases and formulating initial hypotheses (IHs) to test and begin learning the data.

Discovery

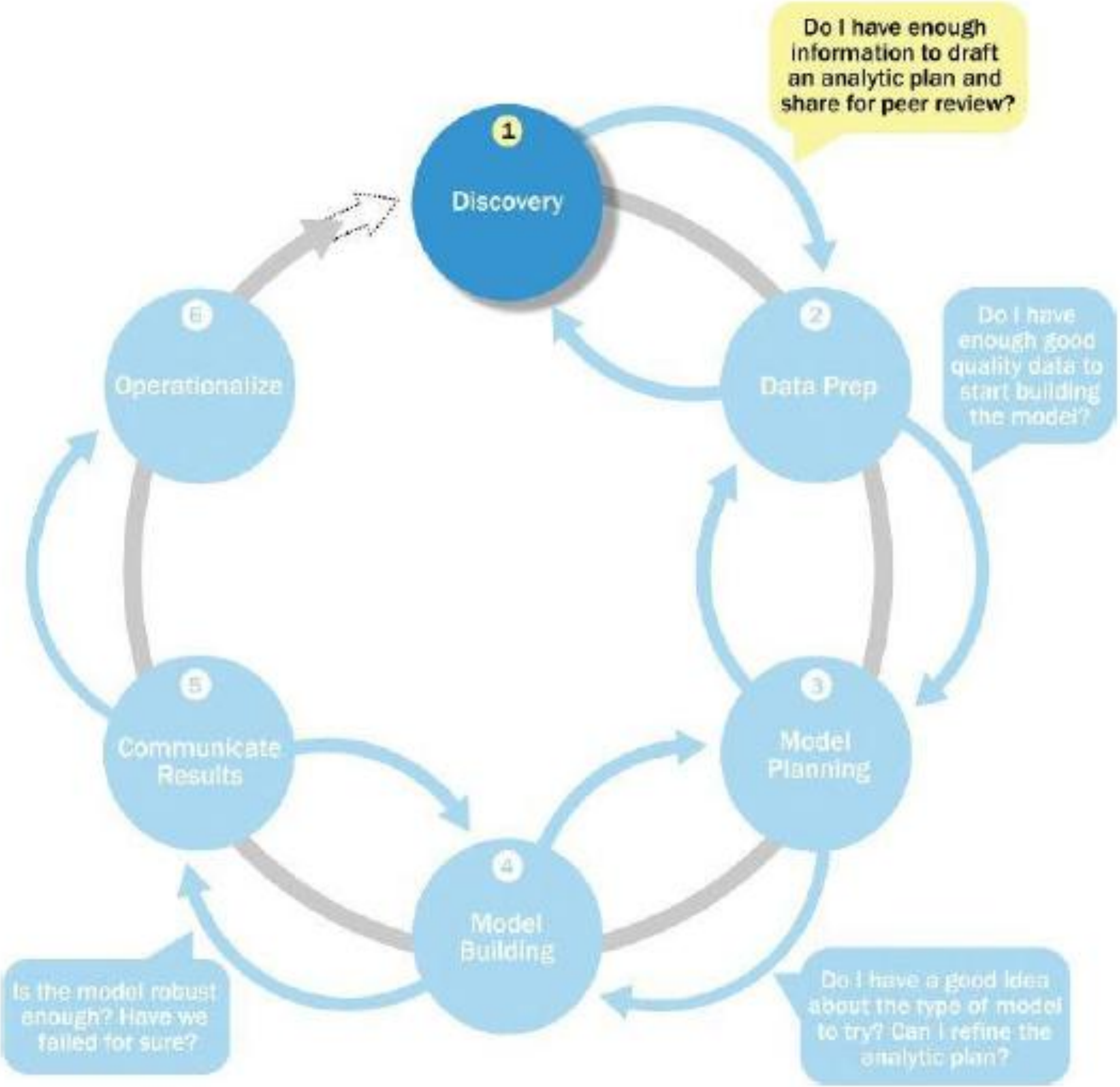
Team

Learn and investigate the problem

Develop context and understanding, and

Learn about the data sources needed and available for the project

Team formulates initial hypotheses that can later be tested with data.


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

Learning the Business Domain

Understanding the domain area of the problem is essential.

Data scientists will have deep computational and quantitative knowledge that can be broadly applied across many disciplines

Resources

The team needs to assess the resources available to support the project.

Resources include technology, tools, systems, data, and people.

Framing the Problem

Framing is the process of stating the analytics problem to be solved.

Identifying Key Stakeholders

who will benefit from the project

5 Major Activities

Identify data sources:

Make a list of candidate data sources the team may need to test the initial hypotheses outlined in this phase.

Make an inventory of the datasets currently available and those that can be purchased or otherwise acquired for the tests the team wants to perform.

Capture aggregate data sources:

This is for previewing the data and providing - high-level understanding.

It enables the team to gain a quick overview of the data and perform further exploration on specific areas.

It also points the team to possible areas of interest within the data.

5 Major Activities

Review the raw data:

Obtain preliminary data from initial data feeds.

Begin - understanding the interdependencies among the data attributes, and become familiar with the content of the data, its quality, and its limitations.

Evaluate the data structures and tools needed:

The data type and structure dictate which tools the team can use to analyze the data.

This evaluation gets the team thinking about which technologies may be good candidates for the project and how to start getting access to these tools.

Scope the sort of data infrastructure needed for this type of problem:

In addition to the tools needed, the data influences the kind of infrastructure that’s required, such as disk storage and network capacity.

Phase 2: Data Preparation

- Steps to explore, preprocess, and condition data prior to modeling and analysis.
- It requires the presence of an analytic sandbox, the team execute, load, and transform, to get data into the sandbox.
- Data preparation tasks are likely to be performed multiple times and not in predefined order.
- Several tools commonly used – Hadoop, Alpine Miner, Open Refine, etc.

Phase 2—Data preparation:

Phase 2 requires the presence of an analytic sandbox, in which the team can work with data and perform analytics for the duration of the project.

The team needs to execute extract, load, and transform (ELT) or extract, transform and load (ETL) to get data into the sandbox.

EtLT stands for Extract, Transform, Load, Transform. It's a data integration strategy that combines the best aspects of ETL (Extract, Transform, Load) and ELT (Extract, Load, Transform).

Data should be transformed in the EtLT process so the team can work with it and analyze it.

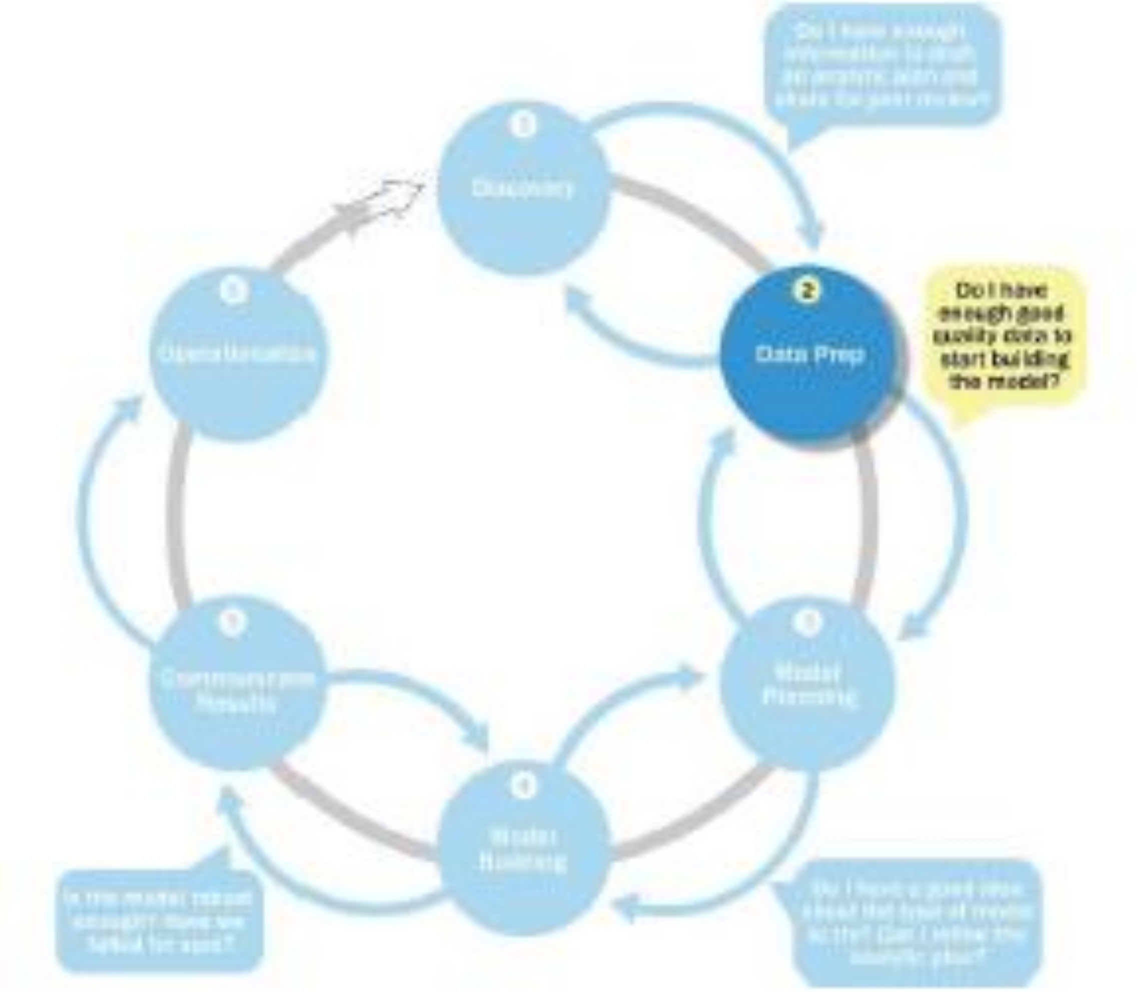
In this phase, the team also needs to familiarize itself with the data thoroughly and take steps to condition the data

Phase 2: Data Preparation

Includes the steps to explore, preprocess, and condition data prior to modeling and analysis.

Understanding the data in detail is critical to the success of the project

The data preparation phase is generally the most iterative and the one that teams tend to underestimate most often


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

Preparing the Analytic Sandbox

the team to obtain an analytic sandbox – ***workspace***

Performing EtLT

EtLT, users perform extract, transform, load processes to extract data from a datastore, perform data transformations, and load the data back into the datastore.

Learning About the Data

A critical aspect of a data science project is to become familiar with the data itself. Spending time to learn the nuances of the datasets provides context to understand what constitutes a reasonable value and expected output versus what is a surprising finding

Data Conditioning

the process of cleaning data, normalizing datasets, and performing transformations on the data – DBA

Survey and Visualize

using data visualization to examine data quality, such as whether the data contains many unexpected values or other indicators of dirty data.

Phase 3: Model Planning

- Team explores data to learn about relationships between variables and subsequently, selects key variables and the most suitable models.
- Data science team develop data sets for training, testing, and production purposes.
- Team builds and executes models based on the work done in the model planning phase.
- Several tools commonly used – Matlab, STASTICA.

Phase 3—Model planning

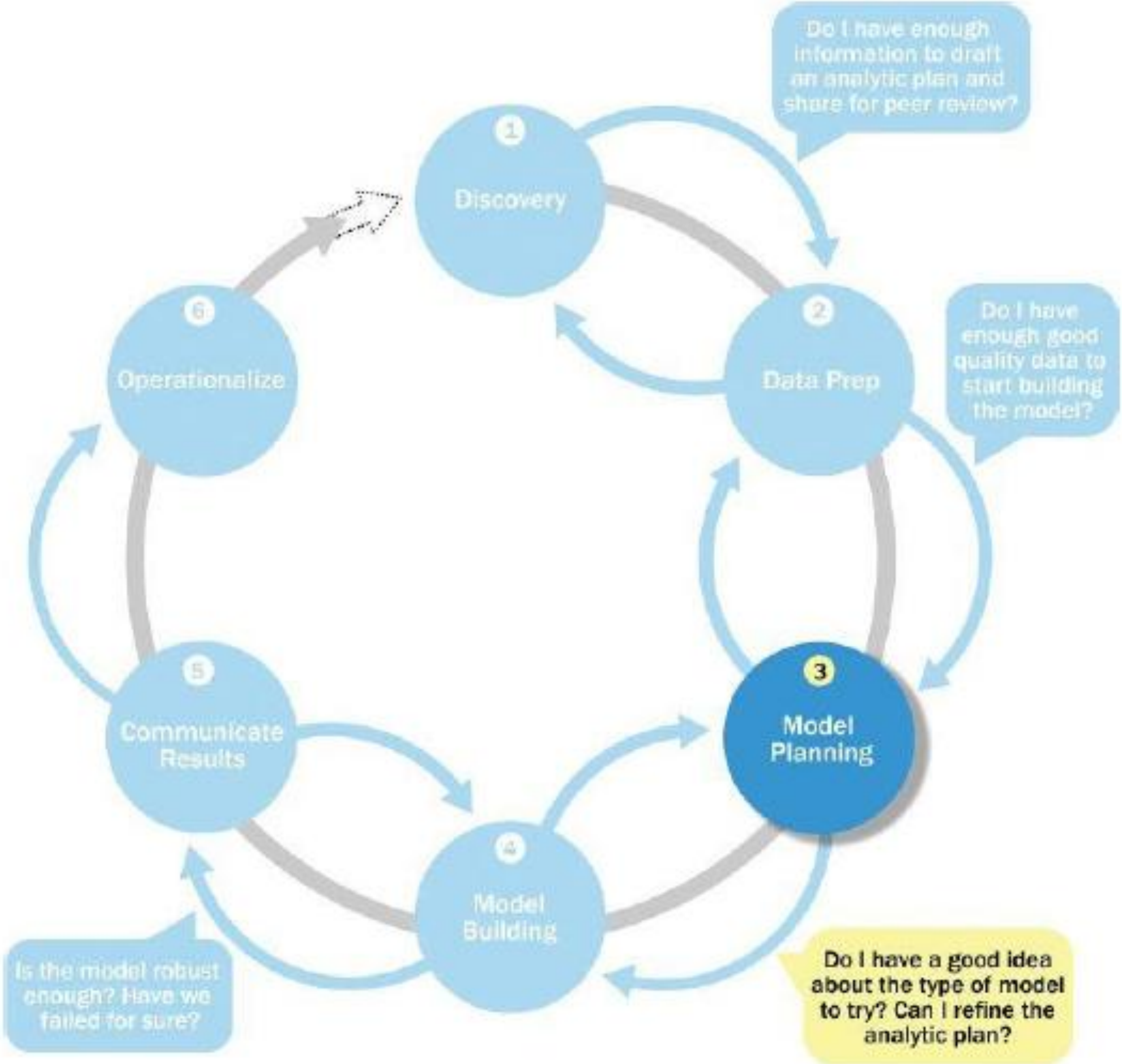
The team determines the methods, techniques, and workflow it intends to follow for the subsequent model building phase.

The team selects the right methods to achieve its objectives framed in phase 1

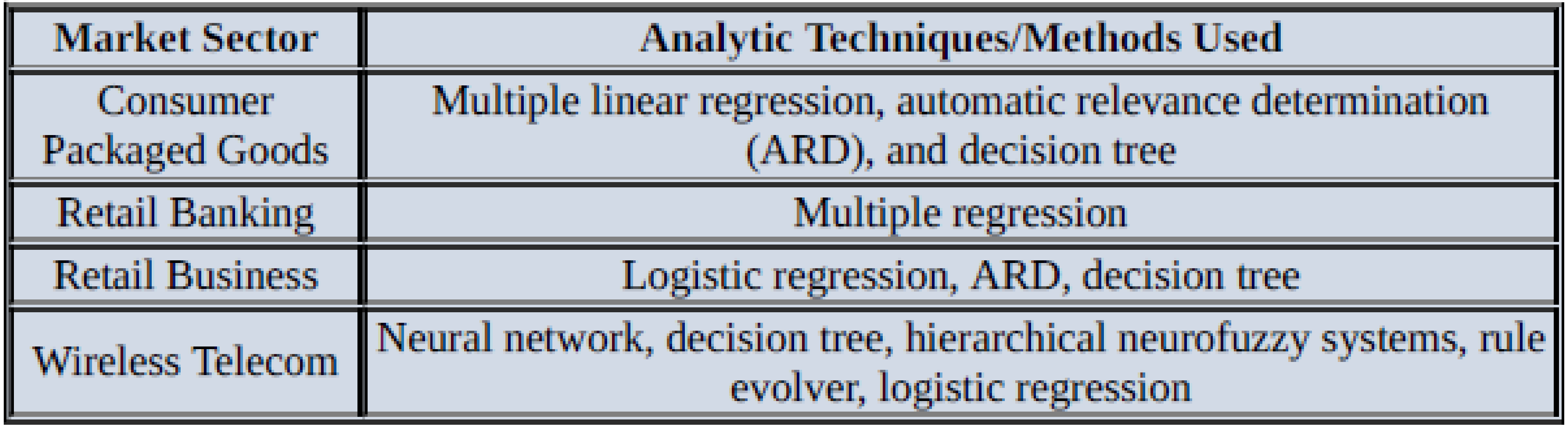
The team explores the data to learn about the relationships between variables and subsequently selects key variables and the most suitable models

Phase 3: Model Planning

The data science team identifies candidate models to apply to the data for clustering, classifying, or finding relationships in the data depending on the goal of the project


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

Model Planning in Industry Verticals



Market Sector	Analytic Techniques/Methods Used
Consumer Packaged Goods	Multiple linear regression, automatic relevance determination (ARD), and decision tree
Retail Banking	Multiple regression
Retail Business	Logistic regression, ARD, decision tree
Wireless Telecom	Neural network, decision tree, hierarchical neurofuzzy systems, rule evolver, logistic regression

Phase 4: Model Building

- Team develops datasets for testing, training, and production purposes.
- Team also considers whether its existing tools will suffice for running the models or if they need more robust environment for executing models.
- Free or open-source tools – Rand PL/R, Octave, WEKA.
- Commercial tools – Matlab , STASTICA.

Phase 4—Model building

The team develops datasets for testing, training, and production purposes.

These datasets enable the data scientist to develop the analytical model and train it (“training data”), while holding aside some of the data (“hold-out data” or “test data”) for testing the model

In addition, in this phase the team builds and executes models based on the work done in the model planning phase.

The team also considers whether its existing tools will suffice for running the models, or if it will need a more robust environment for executing models and workflows

Phase 4: Model Building

Common Tools – statistical analysis

Commercial Tools:

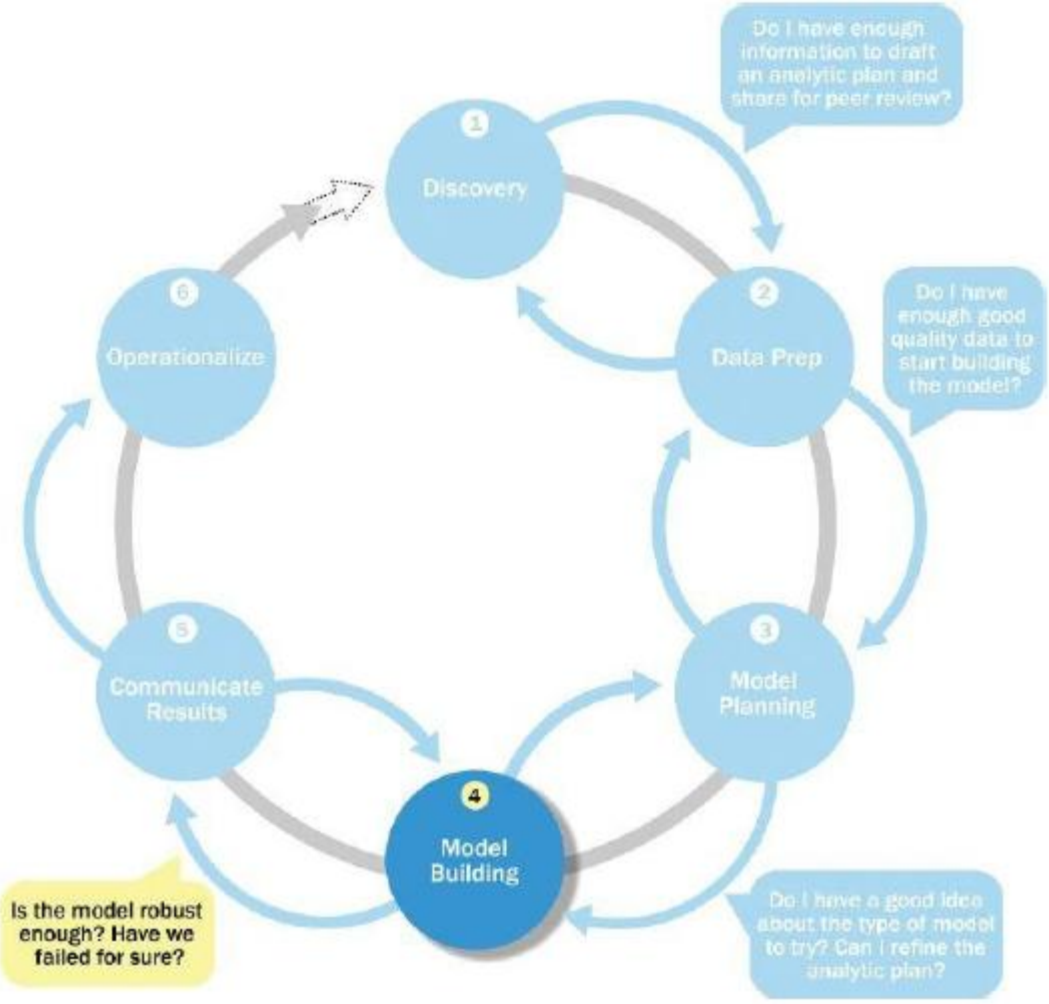
Matlab [19] provides a high-level language for performing a variety of data analytics, algorithms, and data exploration.

Open Source tools

R

WEKA

Python - scikit-learn, numpy, scipy, pandas


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 5: Communication Results

- After executing model team need to compare outcomes of modeling to criteria established for success and failure.
- Team considers how best to articulate findings and outcomes to various team members and stakeholders, taking into account warning, assumptions.
- Team should identify key findings, quantify business value, and develop narrative to summarize and convey findings to stakeholders.

Phase 5: Communicate Results

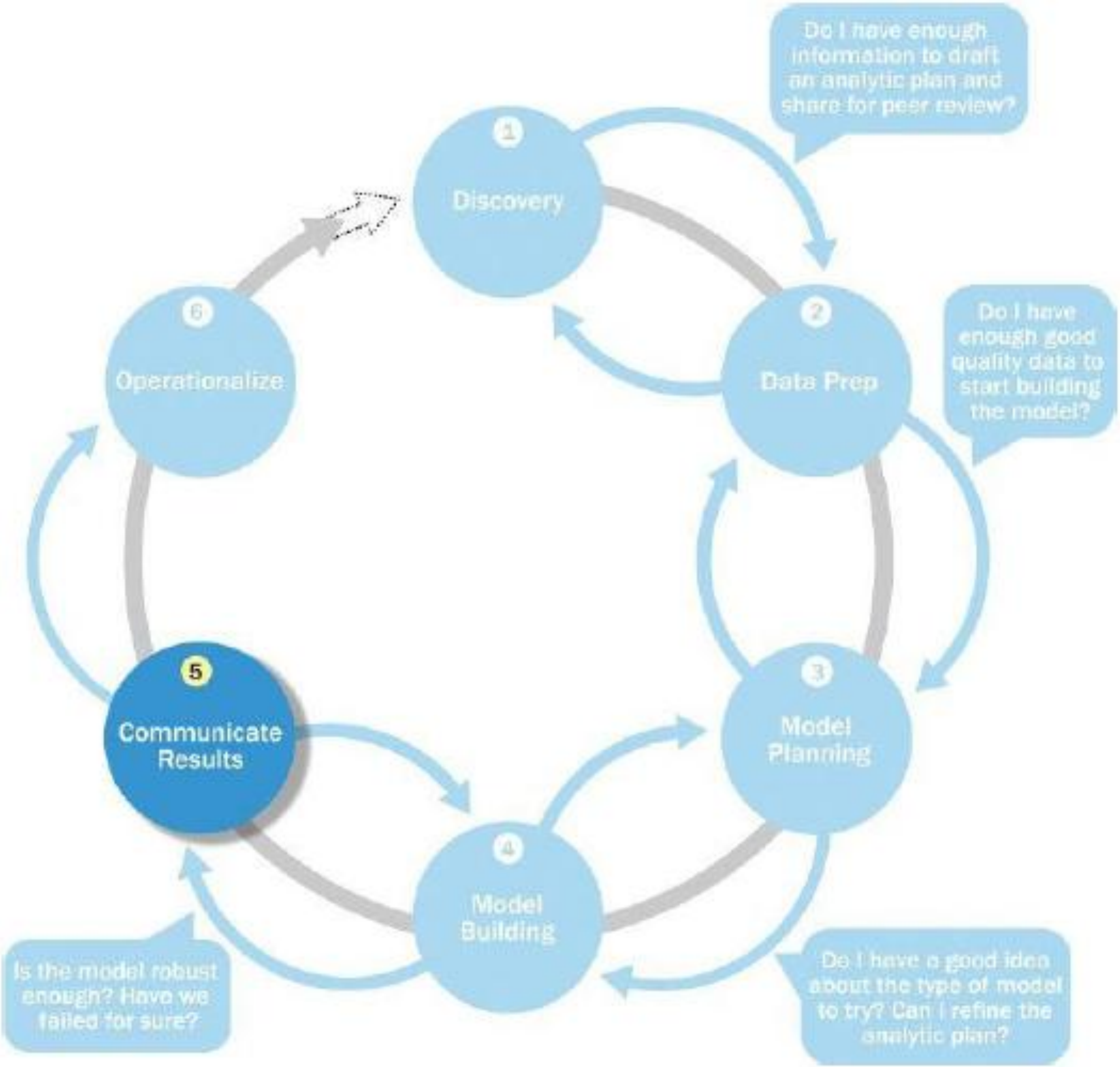
After executing the model, the team needs to compare the outcomes of the modeling to the criteria established for success and failure

The team, in collaboration with major stakeholders, determines if the results of the project, based on the criteria developed in Phase 1.

The team should identify key findings, quantify the business value, and develop a narrative to summarize and convey findings to stakeholders.

Phase 5: Communicate Results

After executing the model, the team needs to compare the outcomes of the modeling to the criteria established for success and failure



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 6: Operationalize

- The team communicates benefits of project more broadly and sets up pilot project to deploy work in controlled way before broadening the work to full enterprise of users.
- This approach enables team to learn about performance and related constraints of the model in production environment on small scale , and make adjustments before full deployment.
- The team delivers final reports, briefings, codes.
- Free or open source tools – Octave, WEKA, SQL, MADlib.

Phase 6—Operationalize

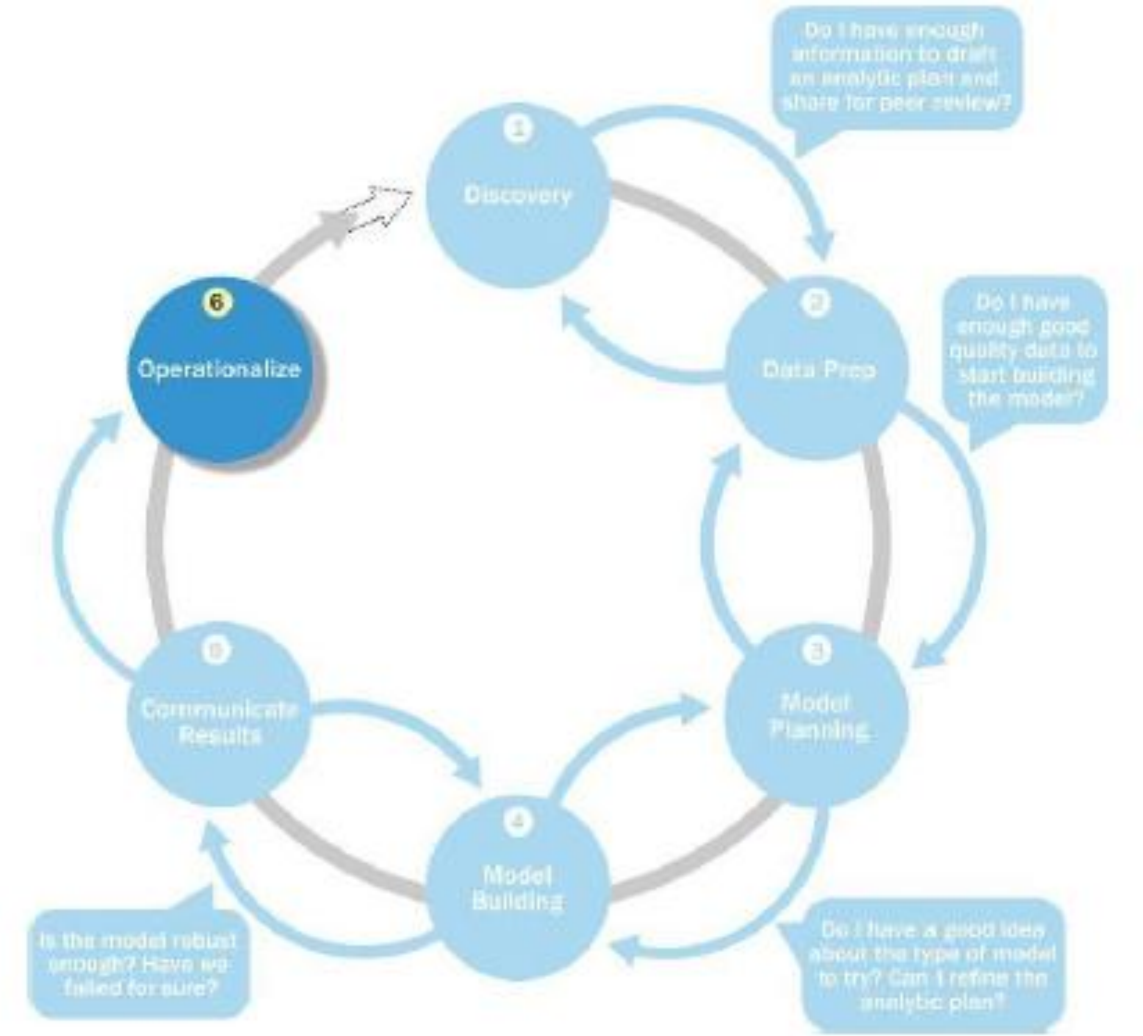
The team delivers final reports, briefings, code, and technical documents.

The team may run a pilot project to implement the models in a production environment.

Rather than deploying these models immediately on a wide-scale basis

risk can be managed more effectively and the team can learn by undertaking a small scope, pilot deployment before a wide-scale rollout.

make adjustments before a full deployment


Do I have enough information to draft an analytic plan and share for peer review?
1
Discovery
6
Operationalize
2
Data Prep
Do I have enough good quality data to start building the model?
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?

LIFE CYCLE OF DATA ANALYSIS PROJECT

Based on CRISP-DM Methodology

IntelliPaat

Step 1

Business Issue Understanding

- Define business objectives
- Gather required information
- Determine appropriate analysis metod
- Clarify scope of work
- Identify deliverables

Step 2

Data Understanding

- Collect initial data
- Identify data requirements
- Determine data availability
- Explore data and characteristics

Step 3

Data Preparation

- Gather data from multiple sources
- Cleanse
- Format
- Blend
- Sample

Step 4

Exploratory Analysis and Modeling

- Develop methodology
- Determine important variables
- Build model
- Assess model

Step 5

Validation

- Evaluate results
- Review process
- Determine next steps
- Results are valid → proceed to step 6
- Results are invalid revisit steps 1-4 ←

Step 6

Visualization and Presentation

- Communicate results
- Determine best method to present insights based on analysis and audience
- Craft a compelling story
- Make recommendations

Lifecycle Phases of Big Data Analytics

- **Stage** 1 - Business case evaluation - The Big Data analytics lifecycle begins with a business case, which defines the reason and goal behind the analysis.
- **Stage** 2 - Identification of data - Here, a broad variety of data sources are identified.
- **Stage** 3 - Data filtering - All of the identified data from the previous stage is filtered here to remove corrupt data.
- **Stage** 4 - Data extraction - Data that is not compatible with the tool is extracted and then transformed into a compatible form.
- **Stage** 5 - Data aggregation - In this stage, data with the same fields across different datasets are integrated.
- **Stage** 6 - Data analysis - Data is evaluated using analytical and statistical tools to discover useful information.
- **Stage** 7 - Visualization of data - With tools like Tableau, Power BI, and QlikView, Big Data analysts can produce graphic visualizations of the analysis.
- **Stage** 8 - Final analysis result - This is the last step of the Big Data analytics lifecycle, where the final results of the analysis are made available to business stakeholders who will take action.

Key Roles

Each plays a critical part in a successful analytics project.

Seven Major Roles

- Business User
- Project Sponsor
- Project Manager
- Business Intelligence Analyst
- Database Administrator (DBA)
- Data Engineer
- Data Scientist

Key Roles

- Business User - understands the domain area and usually benefits from the results
- Project Sponsor
- Project Manager - Ensures that key milestones and objectives are met on time and at the expected quality.
- Business Intelligence Analyst - create dashboards and reports
- Database Administrator (DBA) - configures the database
- Data Engineer - data management and data extraction
- Data Scientist - analytical techniques, data modeling, and applying valid analytical techniques

Global Innovation Network and Analysis (GINA)

- Is a group of senior technologists located in Centers of Excellence (COEs) around the world
- To engage employees across global CoEs to drive innovation, research, and university partnerships
- Provides an example how DATA Analytics Life cycle is applied to analyse innovation data

Global Innovation Network and Analysis (GINA)

- Planned to create a data repository containing both structure and unstructured data to accomplish Three main goals.
  - 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

Teams began to identify the data sources

Roles are

* Business User, Project Sponsor, Project Manager : Vice President of the office(CTO)

* BI analyst : Person from IT

* Data Engineer and DBA : People from IT

GINA can be grouped into two categories:

❑ Descriptive analytics of what is currently 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

The team partnered with its IT department to set up a new analytics sandbox to store and experiment on the data

The team identifies data needed conditioning and normalization

The team realized that several missing datasets were critical to testing some of the analytic hypotheses

Important to determine what level of data quality and cleanliness was sufficient for the project

Phase 3: Model Planning

The parameters related to the scope of the study included the following considerations:

Identify the right milestones to achieve this goal.

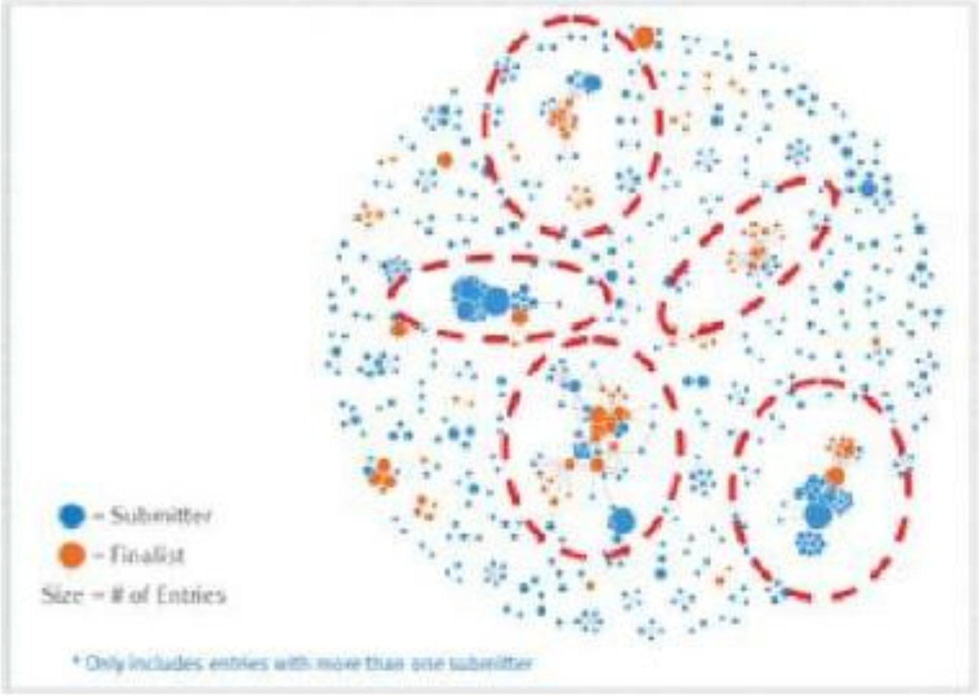
Trace how people move ideas from each milestone toward the goal.

Once this is done, trace ideas that die, and trace others that reach the goal. Compare the journeys of ideas that make it and those that do not.

Compare the times and the outcomes using a few different methods (depending on how the data is collected and assembled).

Phase 4: Model Building

- GINA team employed several analytical methods.
- This included work by the data scientist using Natural Language Processing (NLP) techniques on the textual descriptions of the Innovation Roadmap ideas
- Social Network Analysis using R



* Only includes entries with more than one submitter

Phase 5: Communicate Results

- The team found several ways to cull results of the analysis and identify the most impactful and relevant findings

- The CTO office launched longitudinal studies to begin data collection efforts and track innovation results over longer periods of time.

- The GINA project promoted knowledge sharing related to innovation and researchers spanning multiple areas within the company and outside of it.

- GINA also enabled EMC to cultivate additional intellectual property that led to additional

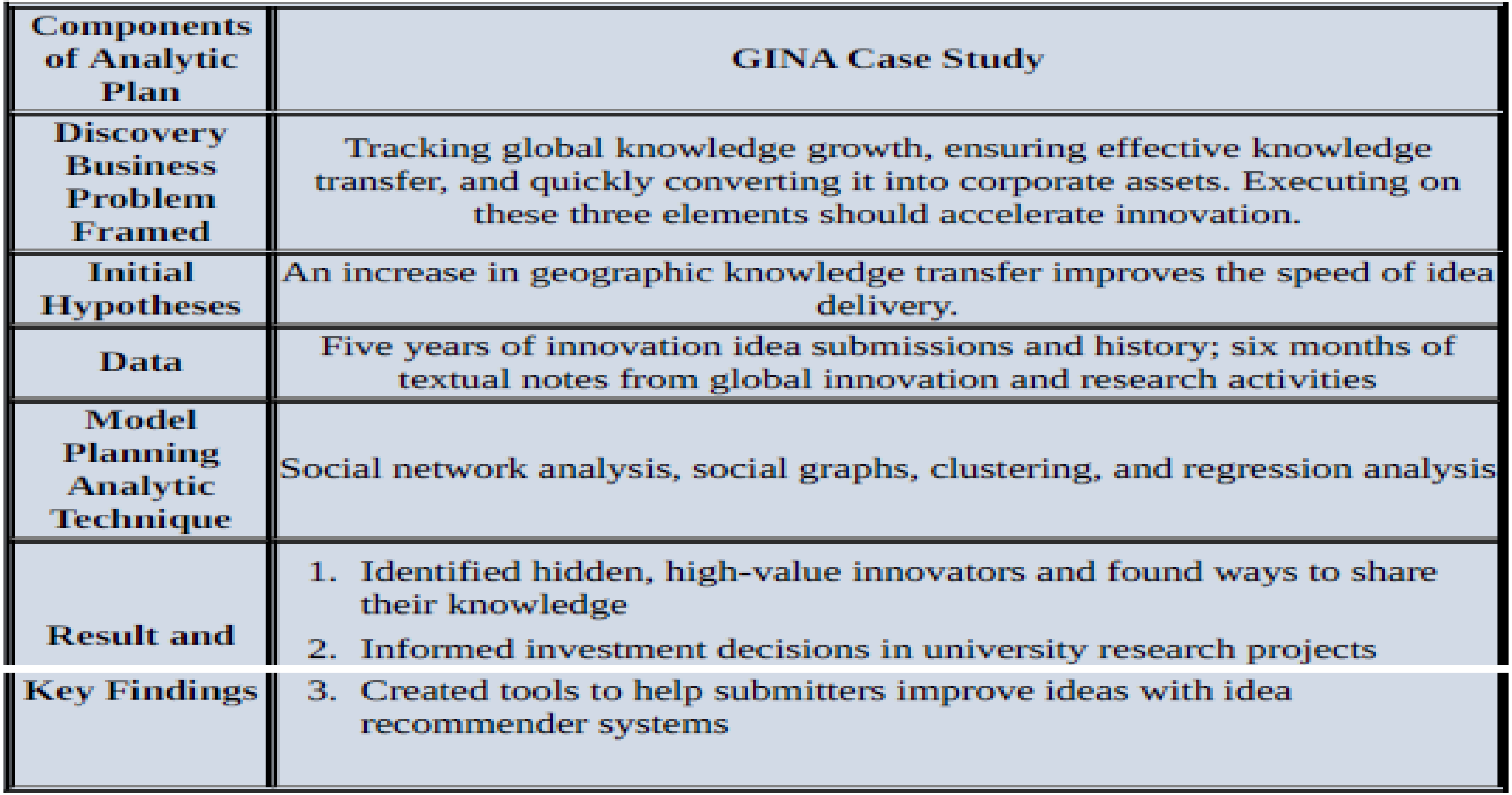
Phase 6: Operationalize

Some of the data is sensitive, and the team needs to consider security and privacy related to the data, such as who can run the models and see the results.

In addition to running models, a parallel initiative needs to be created to improve basic Business Intelligence activities, such as dashboards, reporting, and queries on research activities worldwide.

A mechanism is needed to continually reevaluate the model after deployment.

Analytic Plan from the EMC GINA Project



Components of Analytic Plan	GINA Case Study
Discovery Business Problem Framed	Tracking global knowledge growth, ensuring effective knowledge transfer, and quickly converting it into corporate assets. Executing on these three elements should accelerate innovation.
Initial Hypotheses	An increase in geographic knowledge transfer improves the speed of idea delivery.
Data	Five years of innovation idea submissions and history; six months of textual notes from global innovation and research activities
Model Planning Analytic Technique	Social network analysis, social graphs, clustering, and regression analysis
Result and 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