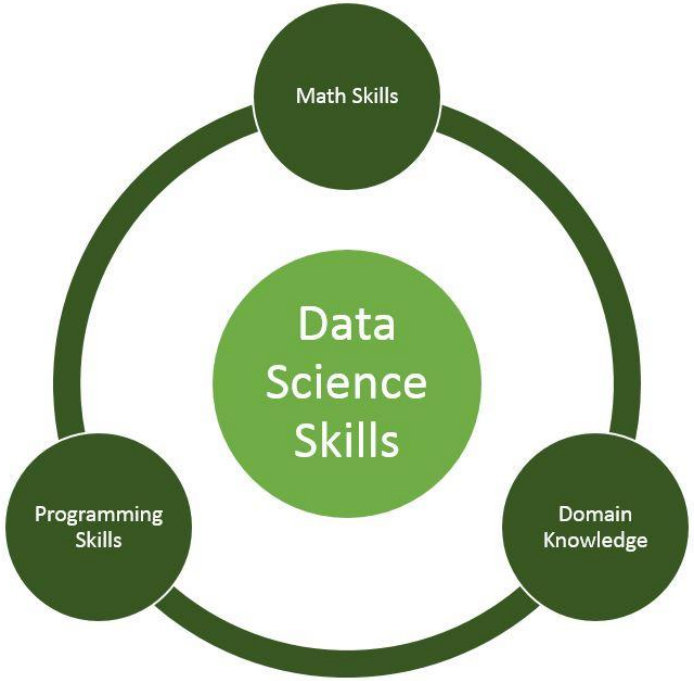
Data Science

- Data science is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining.
- Data science is a concept used to tackle big data and includes data cleansing, preparation, and analysis.
- A data scientist gathers data from multiple sources and applies machine learning, predictive analytics, and sentiment analysis to extract critical information from the collected data sets.
- They understand data from a business point of view and are able to provide accurate predictions and insights that can be used to power critical business decisions.

Math Skills:

- Multivariable Calculus & Linear Algebra
- Probability & Statistics

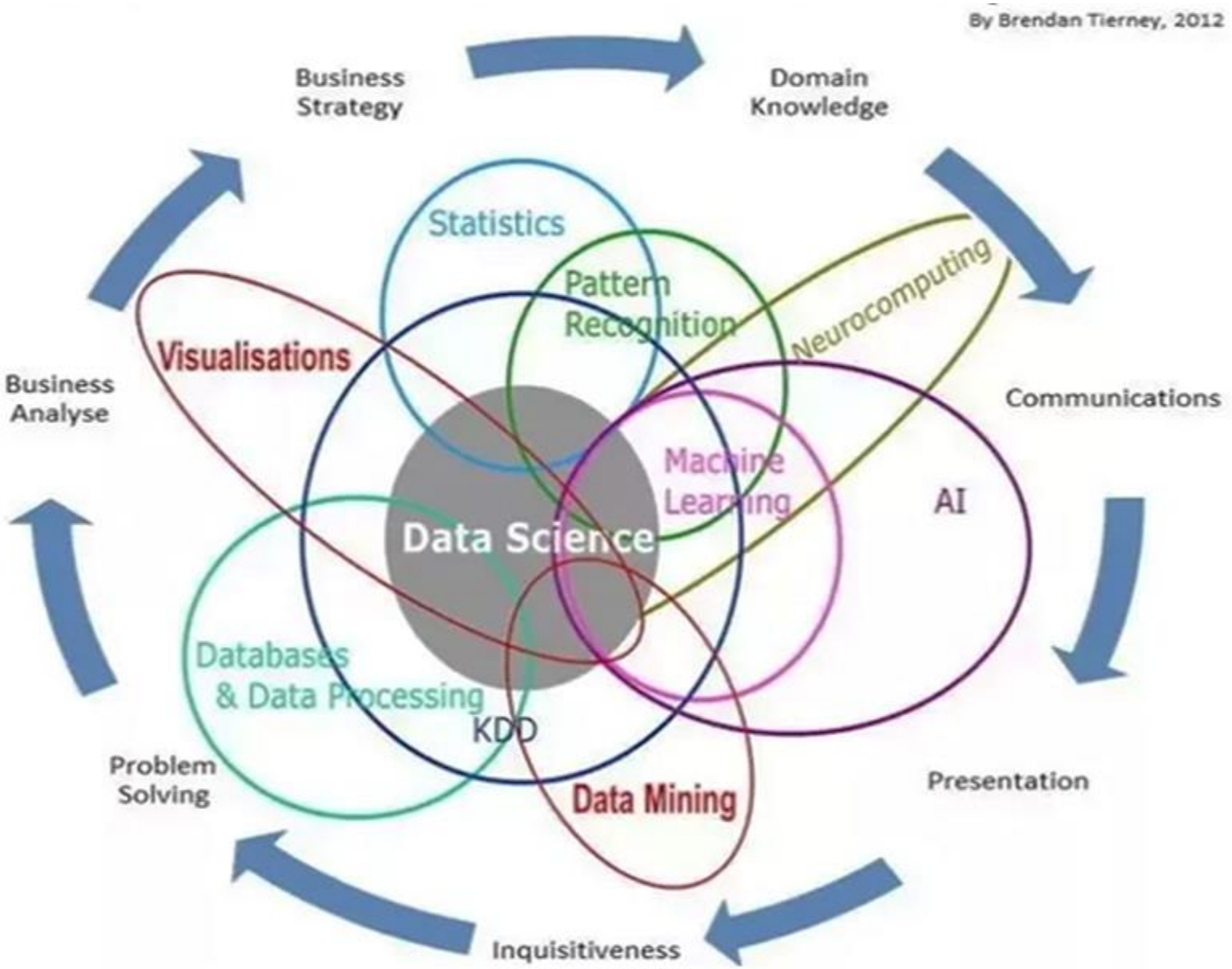


Programming Skills:

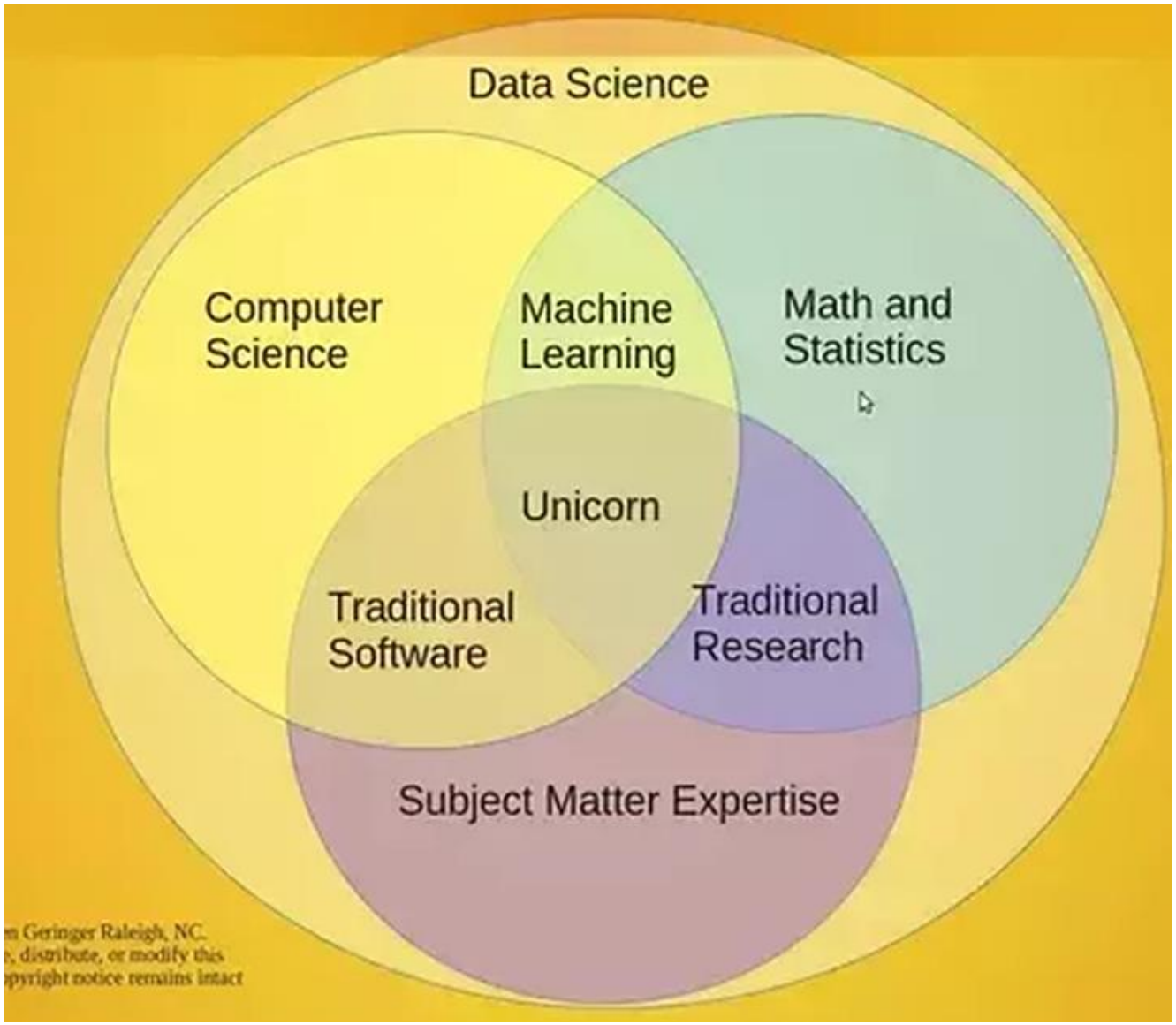
- Data structures and algorithms
- Relational Databases
- Non-Relational Databases
- Distributed Computing
- Machine Learning

Domain Knowledge

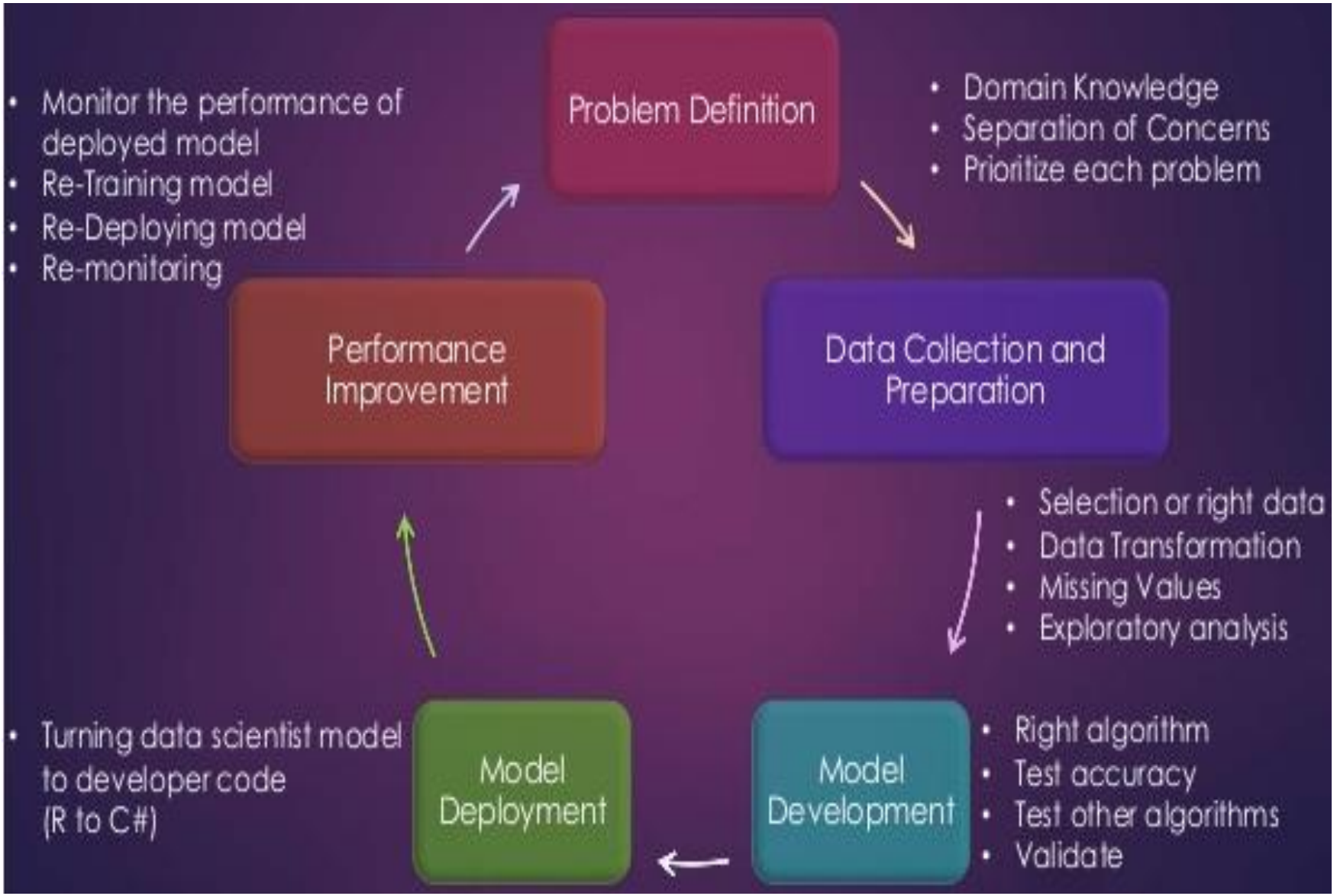
Data Science is Multidisciplinary



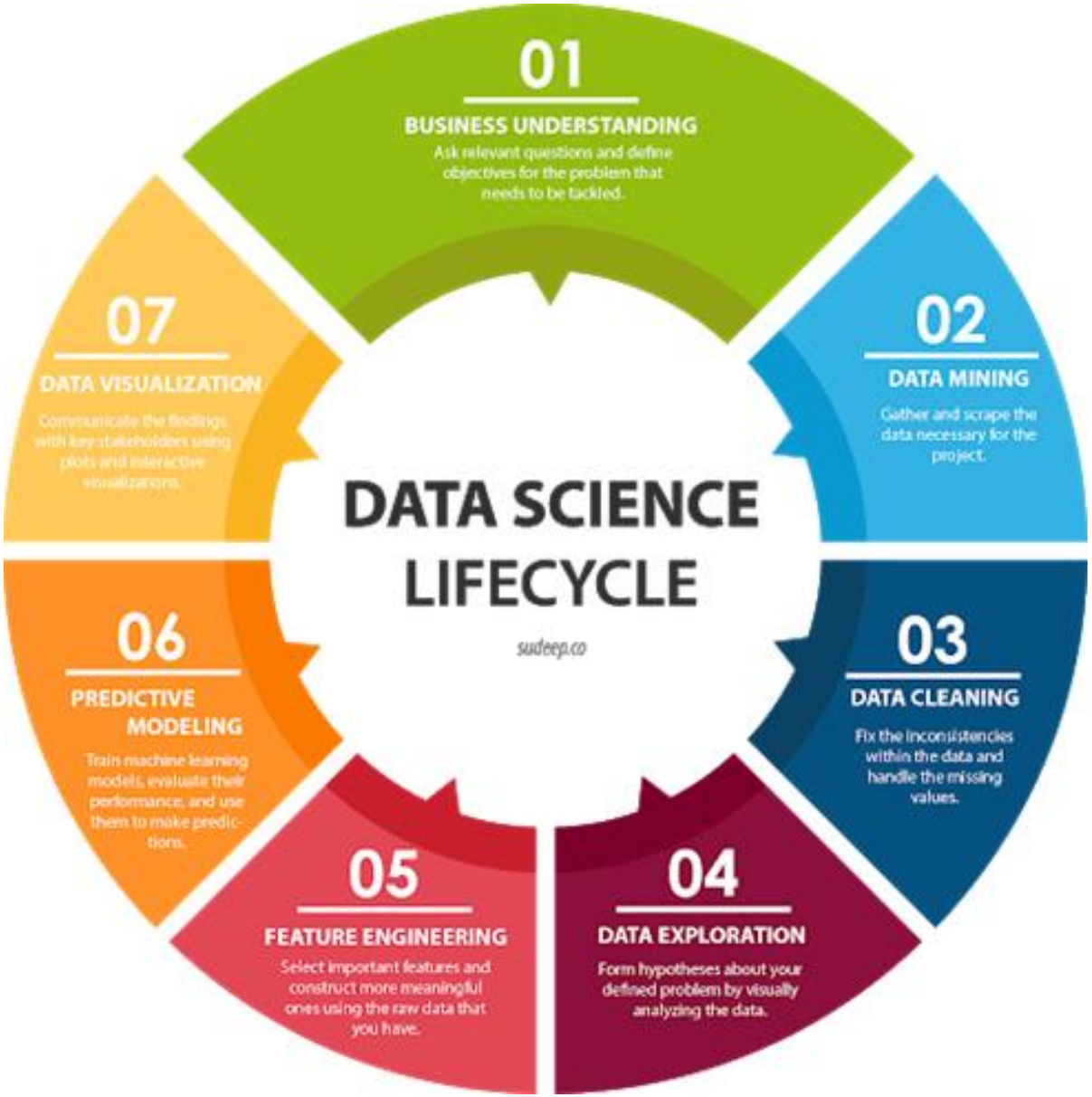
Contributing fields to Data Science



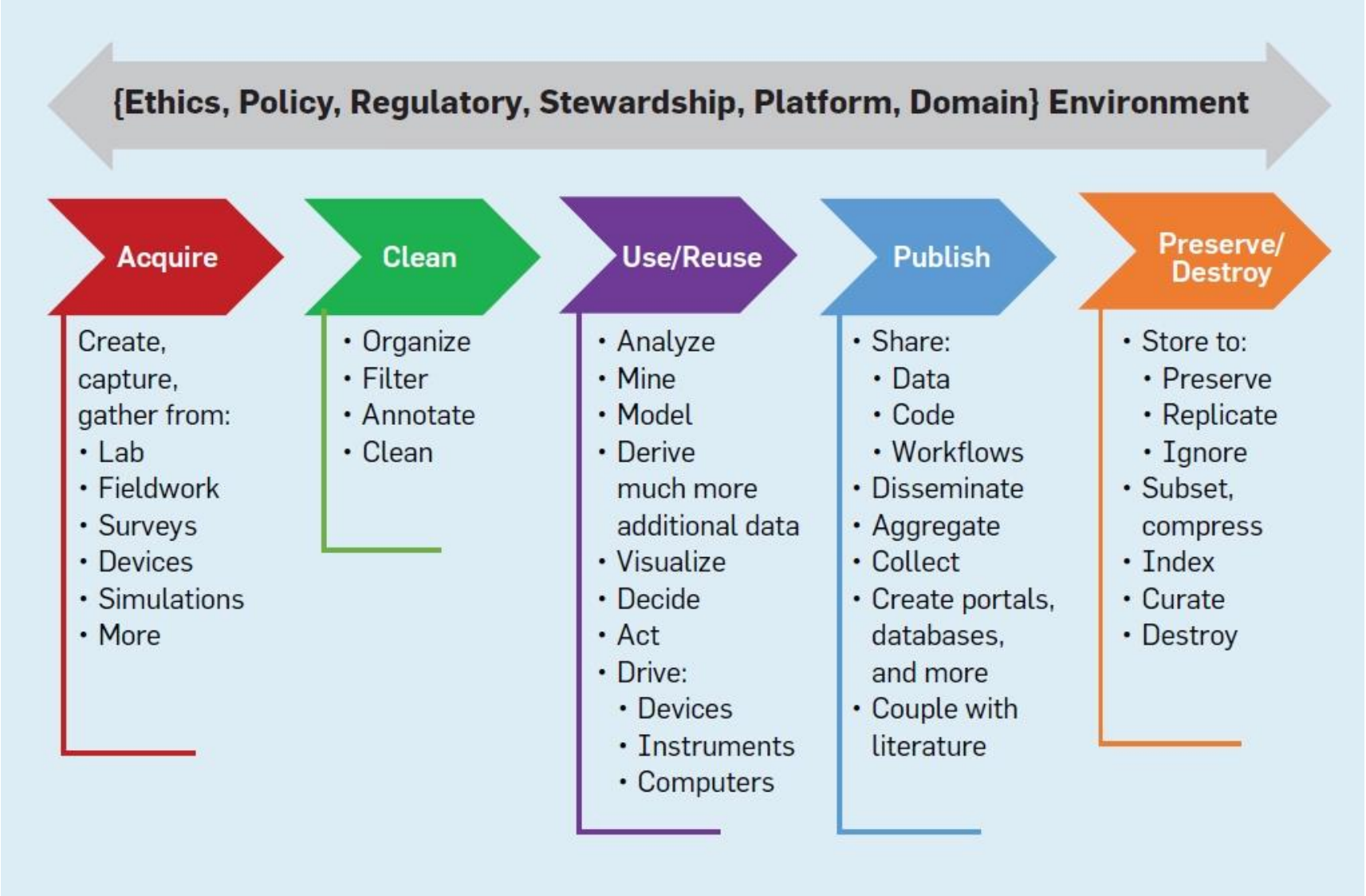
Data Science Work Flow model



Data Science Life Cycle Process



Data Science Life Cycle Process in Details



Data Scientist and their role

- Data scientists are big data handlers, gathering and analyzing large sets of structured and unstructured data.
- A data scientist's role combines computer science, statistics, and mathematics.
- They analyze, process, and model data then interpret the results to create actionable plans for companies and other organizations.
- Data scientists are analytical experts who utilize their skills in both technology and social science to find trends and manage data.
- They use industry knowledge, contextual understanding, skepticism of existing assumptions – to uncover solutions to business challenges.
- A data scientist's work typically involves making sense of messy, unstructured data, from sources such as smart devices, social media feeds, and emails that don't neatly fit into a database.

- Data scientists are charged with communicating complex ideas and making data-driven organizational decisions. As a result, it is highly important for them to be effective communicators, leaders and team members as well as high-level analytical thinkers.
- Experienced data scientists and data managers are tasked with developing a company's best practices, from cleaning to processing and storing data.
- They work cross functionally with other teams throughout their organization, such as marketing, customer success, and operations.

Skills necessary to become a Data Scientist

- Programming
- Machine Learning techniques
- Data Visualization and Reporting
- Risk Analysis
- Statistical analysis and Math
- Effective Communication
- Software Engineering Skills
- Data Mining, Cleaning and Munging
- Research
- Big Data Platforms
- Cloud Tools
- Data warehousing and structures

- **Data Science Expert**

Data scientist's responsibilities

- **Solving business problems through directed or undirected research and framing open-ended industry questions**
- **Extract huge volumes of structured and unstructured data. They query structured data from relational databases using programming languages such as SQL. They gather unstructured data through web scraping, APIs, and surveys.**
- **Employ sophisticated analytical methods, machine learning and statistical methods to prepare data for use in predictive and prescriptive modeling**
- **Thoroughly clean data to discard irrelevant information and prepare the data for preprocessing and modeling**
- **Perform exploratory data analysis (EDA) to determine how to handle missing data and to look for trends and/or opportunities**
- **Discovering new algorithms to solve problems and build programs to automate repetitive work**

- **Communicate predictions and findings to management and IT departments through effective data visualizations and reports**
- **Recommend cost-effective changes to existing procedures and strategies**

Required Skills for a Data Scientist

- **Programming:** Python, SQL, Scala, Java, R, MATLAB
- **Machine Learning:** Natural Language Processing, Classification, Clustering, Ensemble methods, Deep Learning
- **Data Visualization:** Tableau, SAS, D3.js, Python, Java, R libraries
- **Big data platforms:** MongoDB, Oracle, Microsoft Azure, Cloudera

Data Analytics

- Data analytics is the science of analyzing raw data in order to make conclusions about that information. Alternatively, Data analytics (DA) is the process of examining data sets in order to find trends and draw conclusions about the information they contain.
- Data analytics technologies and techniques are widely used in commercial industries to enable organizations to make more-informed business decisions.
- It is also used scientists and researchers to verify or disprove scientific models, theories and hypotheses.
- Many of the techniques and processes of data analytics have been automated into algorithms that work over raw data for human consumption.
- Data analytics techniques can reveal trends and metrics that would otherwise be lost in the mass of information.
- This information can then be used to optimize processes to increase the overall efficiency of a business or system.

What is Big Data Analytics?

- With increasing data size, it has become need for inspecting, cleaning, transforming, and modeling data with the goal of finding useful information, making conclusions, and supporting decision making. This process is known as Big Data Analysis.
- Data mining is a particular data analysis technique where modeling and knowledge discovery for predictive rather than purely descriptive purposes is focused. Business intelligence covers data analysis that relies heavily on aggregation, focusing on business information.
- In statistical applications, some people divide business analytics into descriptive statistics, exploratory data analysis (EDA), and confirmatory data analysis (CDA). EDA focuses on discovering new features in the data and CDA focuses on confirming or falsifying existing hypotheses.
- Predictive analytics does forecasting or classification by focusing on statistical or structural models while in text analytics, statistical, linguistic and structural techniques are applied to extract and classify information from textual sources, a species of unstructured data. All are varieties of data analysis.

- **The Big Data wave has changed ways in which industries function. With Big Data has emerged the requirement to implement advanced analytics to it. Now experts can make more accurate and profitable decisions.**

Understanding Data Analytics

- **Data analytics predominantly refers to an assortment of applications, from basic business intelligence (BI), reporting and online analytical processing (OLAP) to various forms of advanced analytics.**
- **Data analytics is a broad term that encompasses many diverse types of data analysis. Any type of information can be subjected to data analytics techniques to get insight that can be used to improve things.**
- **For example, manufacturing companies often record the runtime, downtime, and work queue for various machines and then analyze the data to better plan the workloads so the machines operate closer to peak capacity.**

The process involved in data analysis involves several different steps:

- **The first step is to determine the data requirements or how the data is grouped. Data may be separated by age, demographic, income, or gender. Data values may be numerical or be divided by category.**
- **The second step in data analytics is the process of collecting it. This can be done through a variety of sources such as computers, online sources, cameras, environmental sources, or through personnel.**
- **Once the data is collected, it must be organized so it can be analyzed. Organization may take place on a spreadsheet or other form of software that can take statistical data.**
- **The data is then cleaned up before analysis. This means it is scrubbed and checked to ensure there is no duplication or error, and that it is not incomplete. This step helps correct any errors before it goes on to a data analyst to be analyzed.**

Big Data analysis has the following characteristics:

a. Programmatic

There might be need to write program for data analysis by using code to manipulate it or do any kind of exploration because of the scale of the data.

b. Data driven

It means progress in an activity is compelled by data and program statements describe the data to be matched and the processing required rather than defining a sequence of steps to be taken. Many analysts use hypothesis driven approach to data analysis, Big Data can use the massive amount of data to drive the analysis.

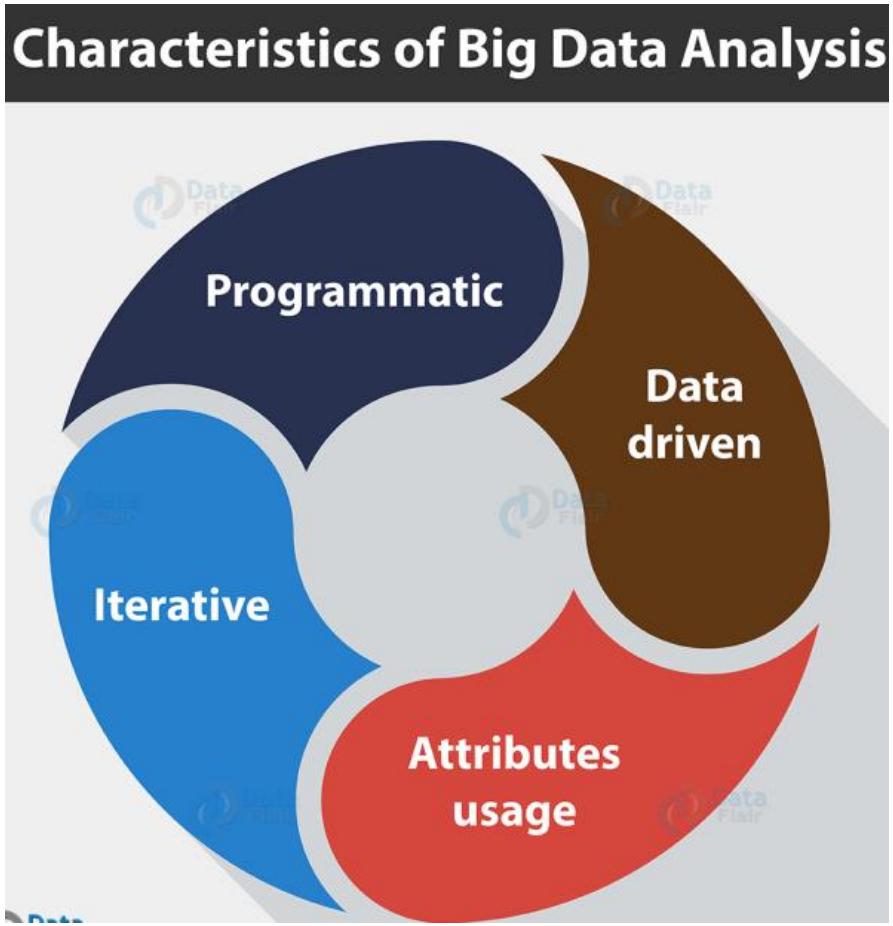
c. Attributes usage

For proper and accurate analysis of data, it can use lot of attributes. In the past, analysts dealt with hundreds of attributes or characteristics of the data source, with Big Data there are now thousands of attributes and millions of observations.

d. Iterative

As whole data is broken into samples and samples are then analyzed, data analytics can be iterative in nature.

More compute power enables iteration of the models until Big Data analysts are satisfied. This has led to development of new applications designed for addressing analysis requirements and time frames.



Key Takeaways

- **Data analytics is the science of analyzing raw data in order to make conclusions about that information.**
- **The techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption.**
- **Data analytics help a business optimize its performance.**

Why Data Analytics Matters

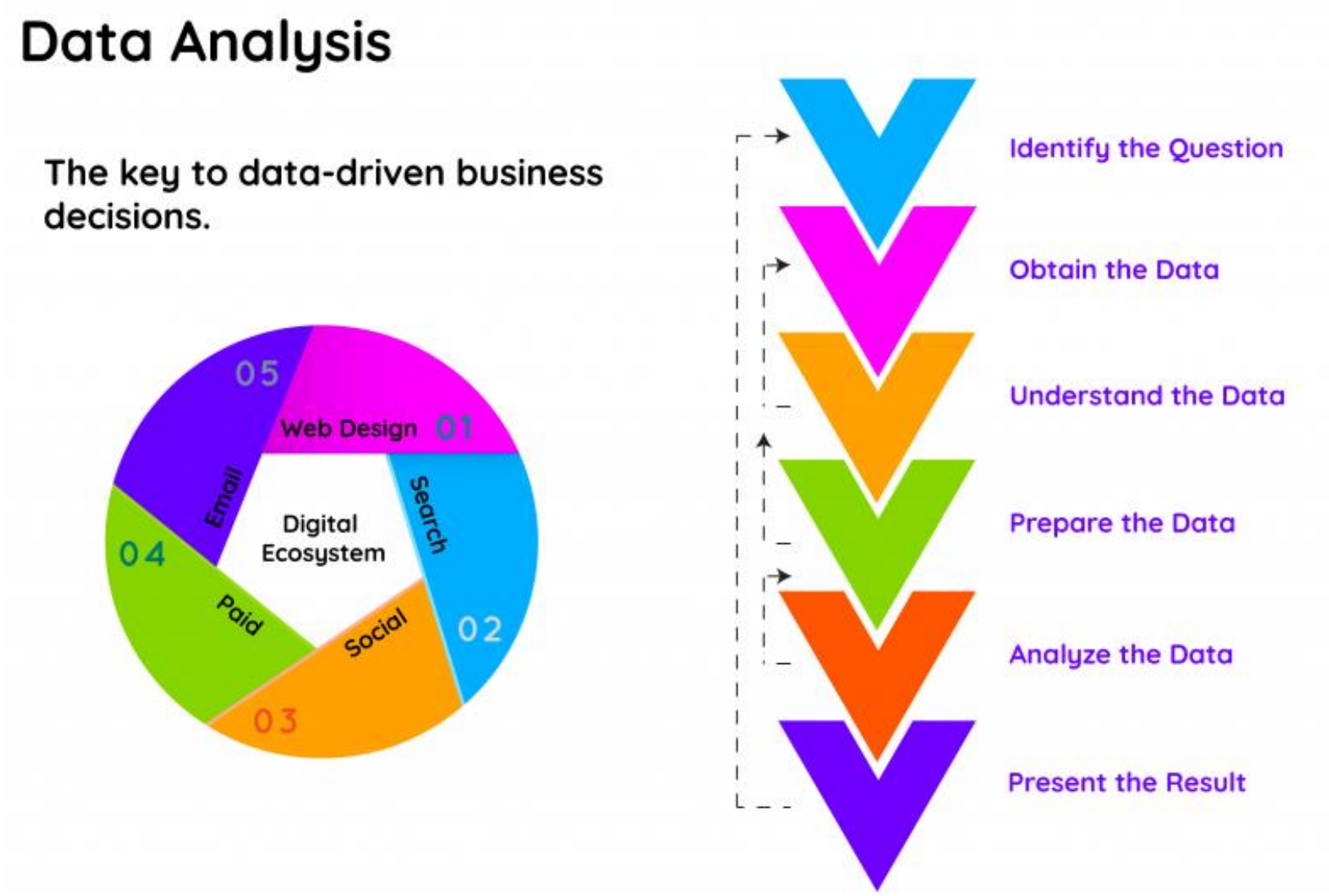
- **Data analytics is important because it helps businesses optimize their performances.**
- **Implementing it into the business model means companies can help reduce costs by identifying more efficient ways of doing business and by storing large amounts of data.**
- **A company can also use data analytics to make better business decisions and help analyze customer trends and satisfaction, which can lead to new—and better—products and services.**

Types of Data Analytics

Data analytics is broken down into four basic types.

- **Descriptive analytics** describes what has happened over a given period of time. Have the number of views gone up? Are sales stronger this month than last?
- **Diagnostic analytics** focuses more on why something happened. This involves more diverse data inputs and a bit of hypothesizing. Did the weather affect beer sales? Did that latest marketing campaign impact sales?
- **Predictive analytics** moves to what is likely going to happen in the near term. What happened to sales the last time we had a hot summer? How many weather models predict a hot summer this year?
- **Prescriptive analytics** suggests a course of action. If the likelihood of a hot summer is measured as an average of these five weather models is above 58%, we should add an evening shift to the brewery and rent an additional tank to increase output.

Steps involved in Data Analysis



- **Identify the question**

  Before you begin working with any data you must understand the problem that you're trying to solve

- **Obtain the data**

  You have to find or collect it, and it has to be the right data to help you answer your question. This can include wide variety of of databases, surveys, APIs, web scraping, 3rd parties and more

- **Understand the data**

  Ensure you can correctly interpret the results and trust your data

- **Prepare the data**

  Make sure your data is comprehensive and doesn't contain incorrect or missing values

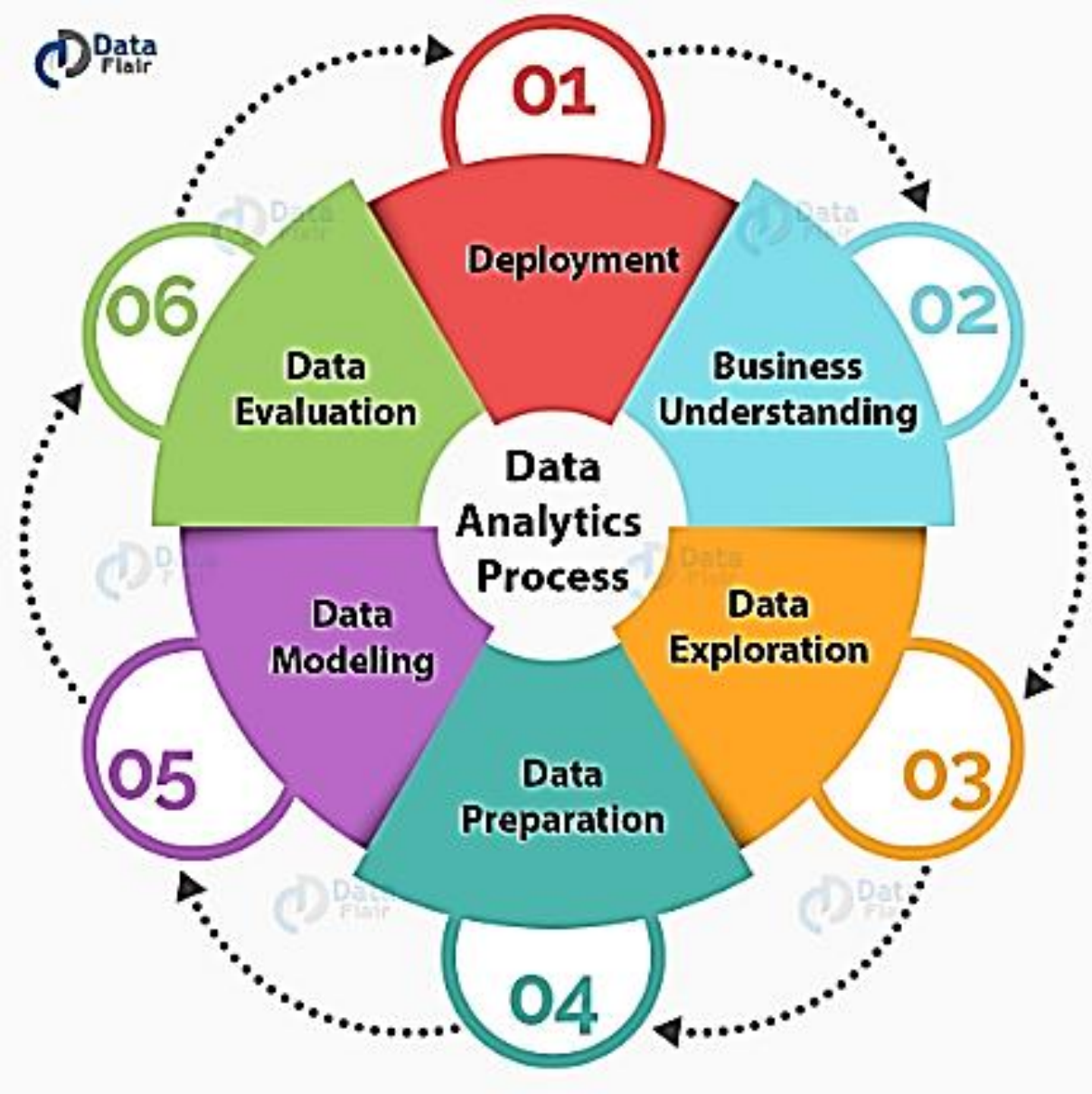
- **Analyze the data**

  Time to uncover the answer! The steps before set you up to win here

- **Present the results**

Assume you find what you are looking for, and it seems like you're ready to share it. It's time to determine the best way to share your results.

Data Analytics



a. Business Understanding

- **The very first step consists of business understanding. Whenever any requirement occurs, firstly we need to determine business objective, assess the situation, determine data mining goals and then produce the project plan as per the requirement.**
- **Business objectives are defined in this phase.**

b. Data Exploration

- **Second step consists of Data understanding. For further process, we need to gather initial data, describe and explore the data and verify data quality to ensure it contains the data we require.**
- **Data collected from the various sources is described in terms of its application and need for the project in this phase. This is also known as data exploration.**
- **This is necessary to verify the quality of data collected.**

c. Data Preparation

- **Next come Data preparation. From the data collected in last step, we need to select data as per the need, clean it, construct it to get useful information and then integrate it all.**
- **Finally we need to format the data to get appropriate data. Data is selected, cleaned, and integrated in the format finalized for the analysis in this phase.**

d. Data Modeling

- **Once data is gathered, we need to do data modeling. For this, we need to select modeling technique, generate test design, build model and assess the model built.**
- **Data model is build to analyze relationships between various selected objects in the data, test cases are built for assessing the model and model is tested and implemented on the data in this phase.**

e. Data Evaluation

- **Next come data evaluation where we evaluate the results generated in last step, review the scope of error and determine next steps that need to be performed.**
- **Results of the test cases are evaluated and reviewed for the scope of error in this phase.**

f. Deployment

- **Final step in analytic process is deployment. Here we need to plan the deployment and monitoring and maintenance, we need to produce final report and review the project.**
- **Results of the analysis are deployed in this phase. This is also known as reviewing of the project.**

Introduction to Data Mining

- **Data mining, also called as data or knowledge discovery, means analyzing data from different perspectives and summarizing it into useful information – information that can be used to take important decisions.**
- **It is the technique of exploring, analyzing, and detecting patterns in large amounts of data. Goal of data mining is either data classification or data prediction. In classification, data is sorted into groups while in prediction, value of a continuous variable is predicted.**
- **Data mining is been used in several sectors like Retail, sales analytics, Financial, Communication, Marketing Organizations etc.**

Some examples of Data Mining are:

a. Classification of trees

b. Logistic regression

c. Neural networks

d. Clustering techniques like the K-nearest neighbors

e. Anomaly detection

What Is Machine Learning?

- **Machine learning can be defined as the practice of using algorithms to use data, learn from it and then forecast future trends for that topic.**
- **Traditional machine learning software comprised of statistical analysis and predictive analysis that are used to spot patterns and catch hidden insights based on perceived data.**
- **Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.**
- **Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.**
- **The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide.**

- **The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.**

Some machine learning methods

Machine learning algorithms are often categorized as supervised or unsupervised.

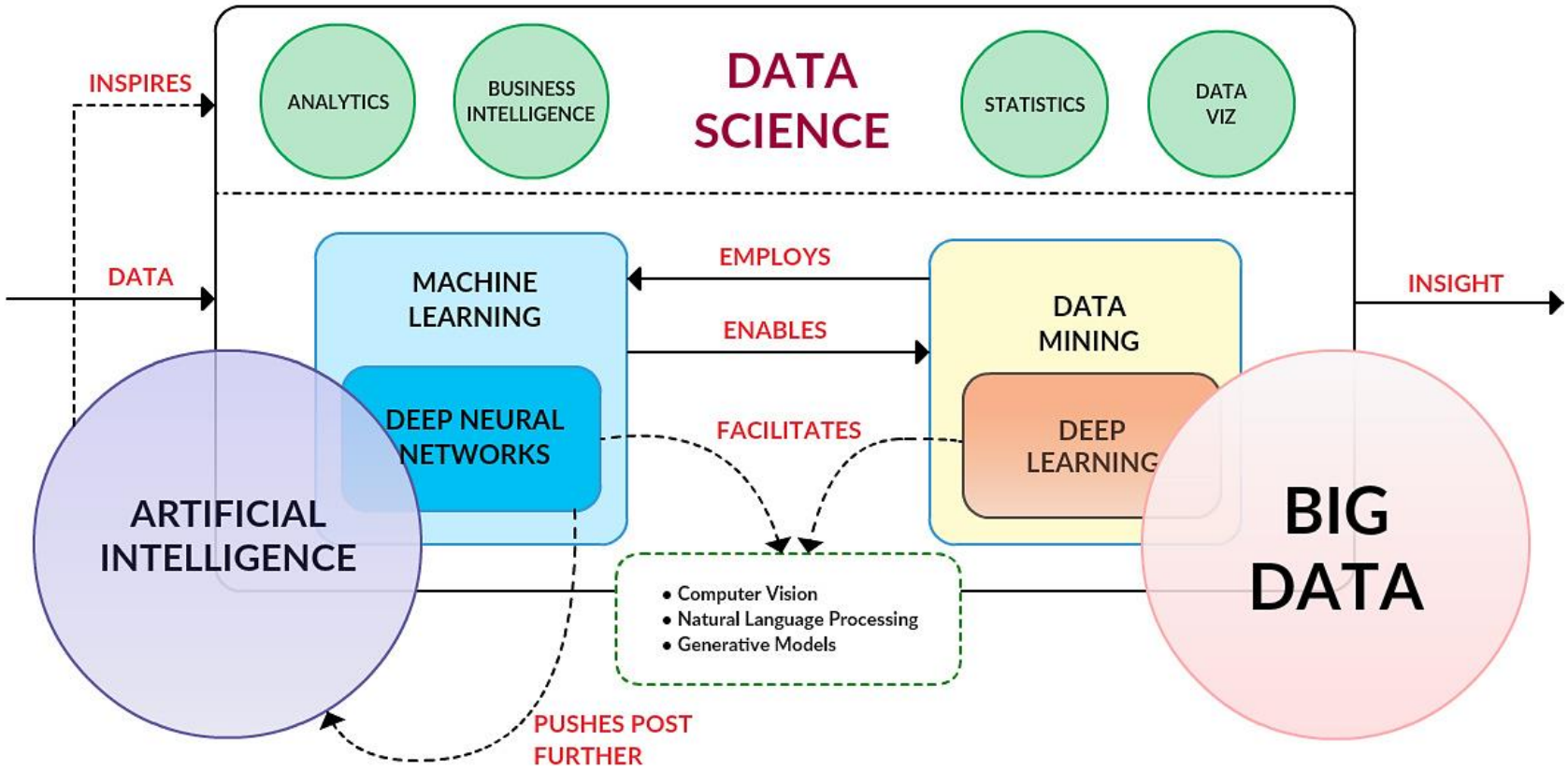
- **Supervised machine learning algorithms can apply what has been learned in the past to new data using labeled examples to predict future events.**
- **Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values.**
- **The system is able to provide targets for any new input after sufficient training.**
- **The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.**

- **In contrast, unsupervised machine learning algorithms are used when the information used to train is neither classified nor labeled.**
- **Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data.**
- **The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data.**
- **Semi-supervised machine learning algorithms fall somewhere in between supervised and unsupervised learning, since they use both labeled and unlabeled data for training – typically a small amount of labeled data and a large amount of unlabeled data.**
- **The systems that use this method are able to considerably improve learning accuracy.**
- **Usually, semi-supervised learning is chosen when the acquired labeled data requires skilled and relevant resources in order to train it / learn from it.**
- **Otherwise, acquiring unlabeled data generally doesn't require additional resources.**

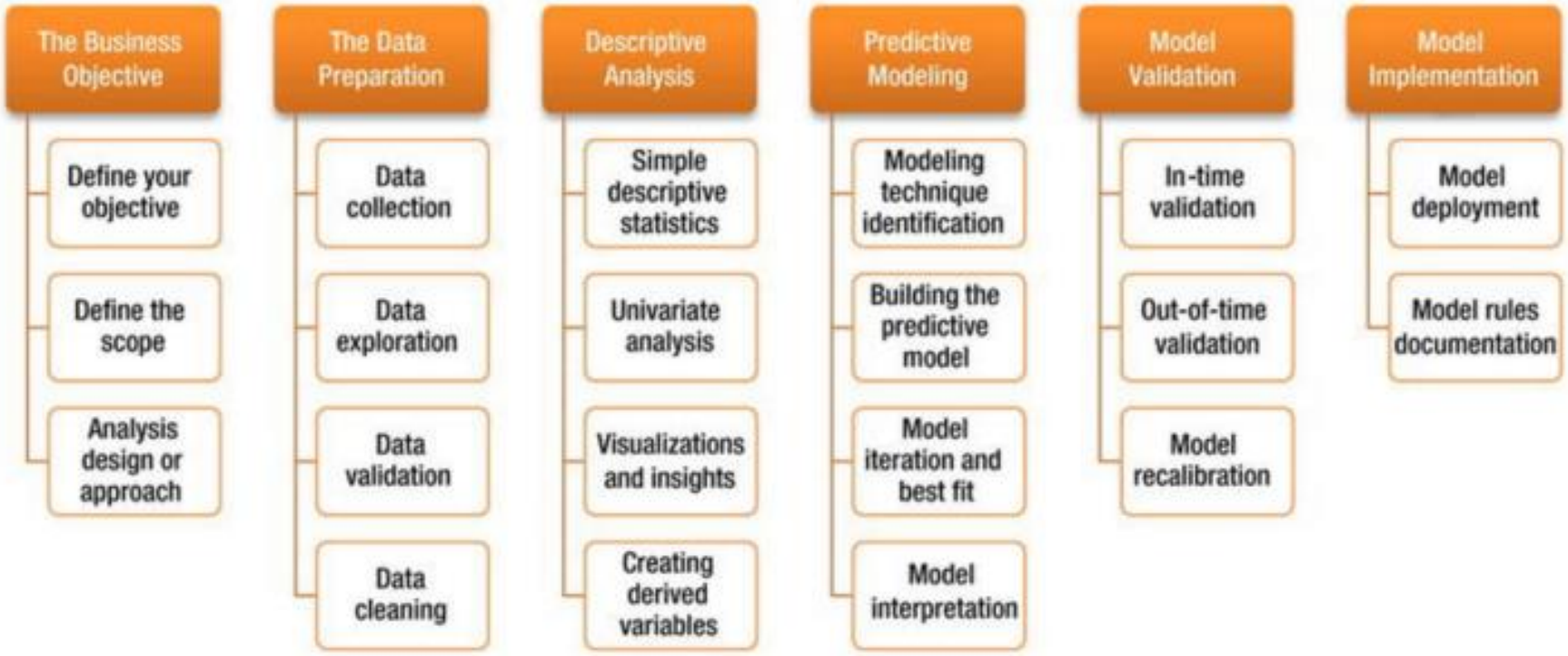
Reinforcement machine learning algorithms

- **It is a learning method that interacts with its environment by producing actions and discovers errors or rewards.**
- **Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning.**
- **This method allows machines and software agents to automatically determine the ideal behavior within a specific context in order to maximize its performance.**
- **Simple reward feedback is required for the agent to learn which action is best; this is known as the reinforcement signal.**
- **Machine learning enables analysis of massive quantities of data. While it generally delivers faster, more accurate results in order to identify profitable opportunities or dangerous risks, it may also require additional time and resources to train it properly.**

- **Combining machine learning with AI and cognitive technologies can make it even more effective in processing large volumes of information.**


INSPIRES
ANALYTICS
BUSINESS INTELLIGENCE
DATA SCIENCE
STATISTICS
DATA VIZ
DATA
MACHINE LEARNING
EMPLOYS
ENABLES
DATA MINING
INSIGHT
DEEP NEURAL NETWORKS
FACILITATES
DEEP LEARNING
ARTIFICIAL INTELLIGENCE
BIG DATA
• Computer Vision
• Natural Language Processing
• Generative Models
PUSHES POST FURTHER

Data Analytics Lifecycle and methodology



CRISP-DM Methodology

- **The CRISP-DM methodology that stands for Cross Industry Standard Process for Data Mining, is a cycle that describes commonly used approaches that data mining experts use to tackle problems in traditional BI data mining.**
- **It is still being used in traditional BI data mining teams.**
- **CRISP-DM was conceived in 1996 and the next year, it got underway as a European Union project under the ESPRIT funding initiative.**

Let us now learn a little more on each of the stages involved in the CRISP-DM life cycle –

Business Understanding –

- **This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition.**
- **A preliminary plan is designed to achieve the objectives. A decision model, especially one built using the Decision Model and Notation standard can be used.**

Data Understanding –

- The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

Data Preparation –

- The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data.
- Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools.

Modeling –

- In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some

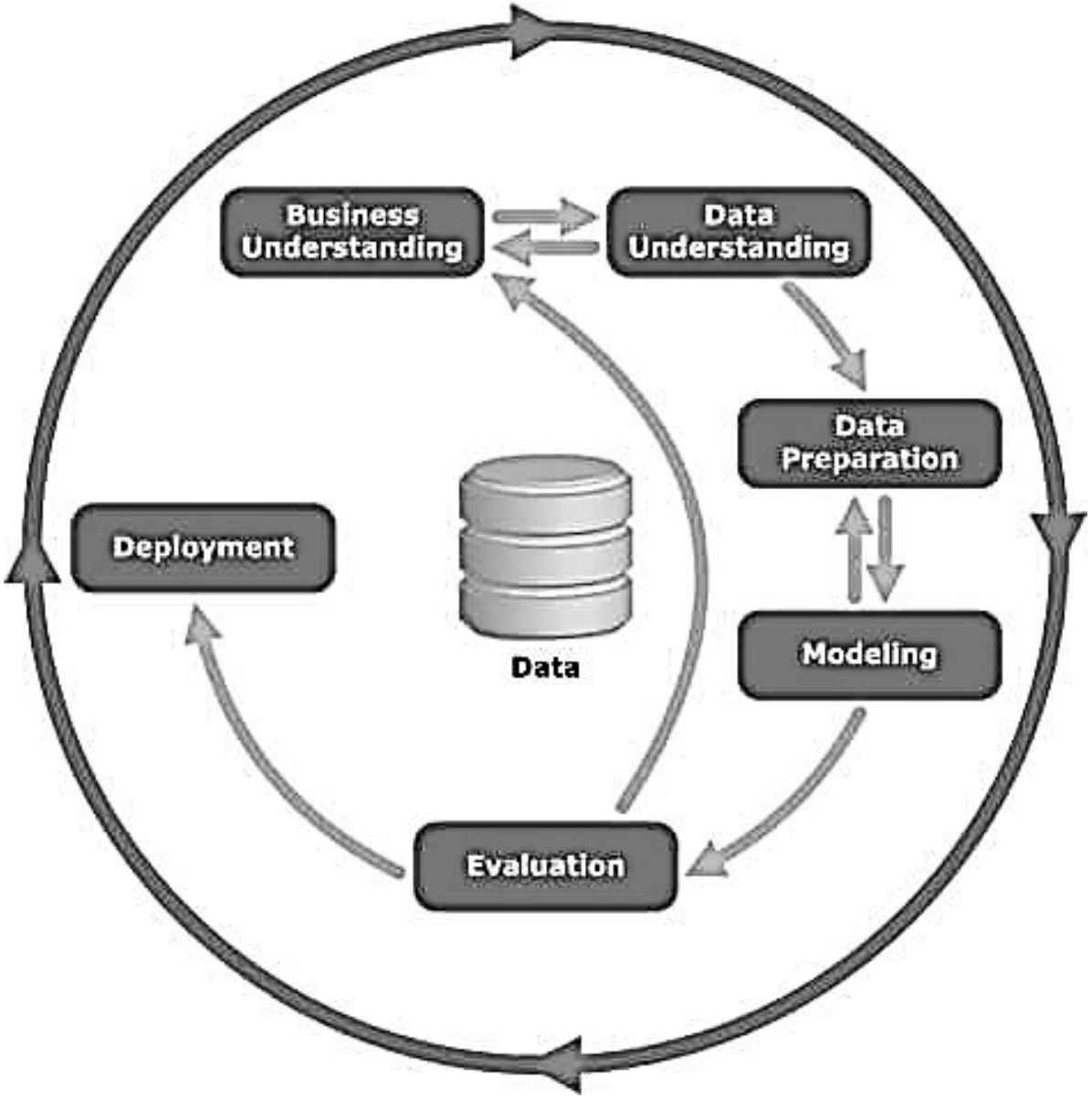
techniques have specific requirements on the form of data. Therefore, it is often required to step back to the data preparation phase.

Evaluation –

- At this stage in the project, you have built a model (or models) that appears to have high quality, from a data analysis perspective.
- Before proceeding to final deployment of the model, it is important to evaluate the model thoroughly and review the steps executed to construct the model, to be certain it properly achieves the business objectives.

Deployment –

- Creation of the model is generally not the end of the project.
- Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that is useful to the customer.


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

SEMMA Methodology

SEMMA is another methodology developed by SAS for data mining modeling. It stands for Sample, Explore, Modify, Model, and Asses. Here is a brief description of its stages −

- **Sample − The process starts with data sampling, e.g., selecting the dataset for modeling. The dataset should be large enough to contain sufficient information to retrieve, yet small enough to be used efficiently. This phase also deals with data partitioning.**
- **Explore − This phase covers the understanding of the data by discovering anticipated and unanticipated relationships between the variables, and also abnormalities, with the help of data visualization.**
- **Modify − The Modify phase contains methods to select, create and transform variables in preparation for data modeling.**
- **Model − In the Model phase, the focus is on applying various modeling (data mining) techniques on the prepared variables in order to create models that possibly provide the desired outcome.**
- **Assess − The evaluation of the modeling results shows the reliability and usefulness of the created models.**

The main difference between CRISM–DM and SEMMA is that SEMMA focuses on the modeling aspect, whereas CRISP-DM gives more importance to stages of the cycle prior to modeling such as understanding the business problem to be solved, understanding and preprocessing the data to be used as input, for example, machine learning algorithms.

Big Data Life Cycle

In today's big data context, the previous approaches are either incomplete or suboptimal. For example, the SEMMA methodology disregards completely data collection and preprocessing of different data sources. These stages normally constitute most of the work in a successful big data project.

A big data analytics cycle can be described by the following stage –

- **Business Problem Definition**
- **Research**
- **Human Resources Assessment**
- **Data Acquisition**
- **Data Munging**
- **Data Storage**
- **Exploratory Data Analysis**
- **Data Preparation for Modeling and Assessment**

Modeling

Implementation

In this section, we will throw some light on each of these stages of big data life cycle.

Business Problem Definition:

- **This is a point common in traditional BI and big data analytics life cycle. Normally it is a non-trivial stage of a big data project to define the problem and evaluate correctly how much potential gain it may have for an organization.**
- **It seems obvious to mention this, but it has to be evaluated what are the expected gains and costs of the project.**

Research:

- **Analyze what other companies have done in the same situation.**

- **This involves looking for solutions that are reasonable for your company, even though it involves adapting other solutions to the resources and requirements that your company has. In this stage, a methodology for the future stages should be defined.**

Human Resources Assessment:

- **Once the problem is defined, it's reasonable to continue analyzing if the current staff is able to complete the project successfully.**
- **Traditional BI teams might not be capable to deliver an optimal solution to all the stages, so it should be considered before starting the project if there is a need to outsource a part of the project or hire more people.**

Data Acquisition:

- **This section is key in a big data life cycle; it defines which type of profiles would be needed to deliver the resultant data product. Data gathering is a non-trivial step of the process; it normally involves gathering unstructured data from different sources.**

- **To give an example, it could involve writing a crawler to retrieve reviews from a website. This involves dealing with text, perhaps in different languages normally requiring a significant amount of time to be completed.**

Data Munging:

- **Once the data is retrieved, for example, from the web, it needs to be stored in an easyto-use format.**
- **To continue with the reviews examples, let's assume the data is retrieved from different sites where each has a different display of the data.**
- **Suppose one data source gives reviews in terms of rating in stars, therefore it is possible to read this as a mapping for the response variable $y \in \{1, 2, 3, 4, 5\}$.**
- **Another data source gives reviews using two arrows system, one for up voting and the other for down voting. This would imply a response variable of the form $y \in \{positive, negative\}$.In order to combine both the data sources, a decision has to be made in order to make these two response representations equivalent.**

- **This can involve converting the first data source response representation to the second form, considering one star as negative and five stars as positive. This process often requires a large time allocation to be delivered with good quality.**

Data Storage:

- **Once the data is processed, it sometimes needs to be stored in a database. Big data technologies offer plenty of alternatives regarding this point.**
- **The most common alternative is using the Hadoop File System for storage that provides users a limited version of SQL, known as HIVE Query Language.**
- **This allows most analytics task to be done in similar ways as would be done in traditional BI data warehouses, from the user perspective.**

Exploratory Data Analysis:

- **Once the data has been cleaned and stored in a way that insights can be retrieved from it, the data exploration phase is mandatory.**

- **The objective of this stage is to understand the data, this is normally done with statistical techniques and also plotting the data. This is a good stage to evaluate whether the problem definition makes sense or is feasible.**

Data Preparation for Modeling and Assessment:

- **This stage involves reshaping the cleaned data retrieved previously and using statistical preprocessing for missing values imputation, outlier detection, normalization, feature extraction and feature selection.**

Modelling:

- **The prior stage should have produced several datasets for training and testing, for example, a predictive model.**
- **This stage involves trying different models and looking forward to solving the business problem at hand. In practice, it is normally desired that the model would give some insight into the business.**
- **Finally, the best model or combination of models is selected evaluating its performance on a left-out dataset.**

Implementation:

- **In this stage, the data product developed is implemented in the data pipeline of the company.**
- **This involves setting up a validation scheme while the data product is working, in order to track its performance. For example, in the case of implementing a predictive model, this stage would involve applying the model to new data and once the response is available, evaluate the model.**