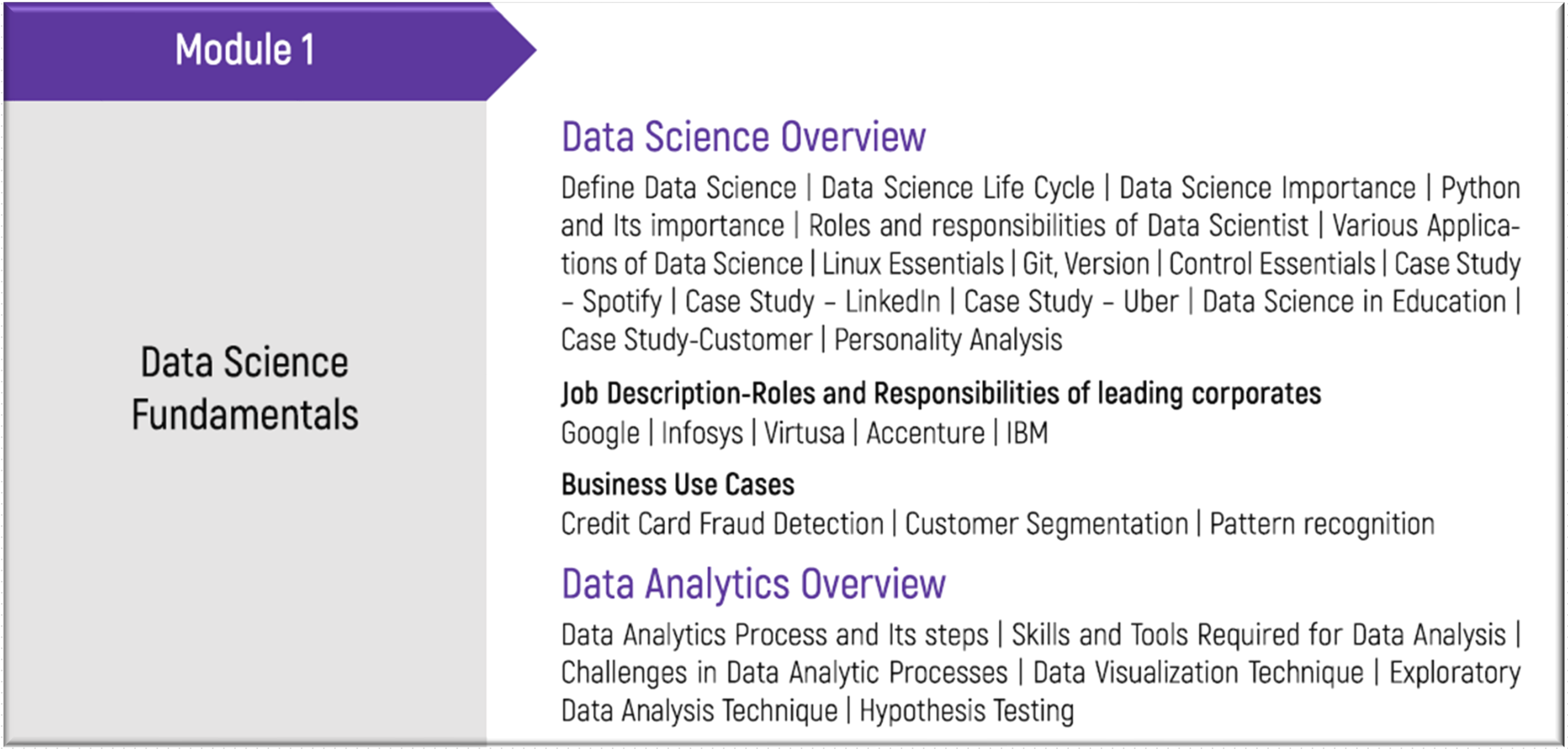
Data Science Overview



Module 1

Data Science Fundamentals

Data Science Overview

Define Data Science | Data Science Life Cycle | Data Science Importance | Python and Its importance | Roles and responsibilities of Data Scientist | Various Applications of Data Science | Linux Essentials | Git, Version | Control Essentials | Case Study – Spotify | Case Study – LinkedIn | Case Study – Uber | Data Science in Education | Case Study-Customer | Personality Analysis

Job Description-Roles and Responsibilities of leading corporates

Google | Infosys | Virtusa | Accenture | IBM

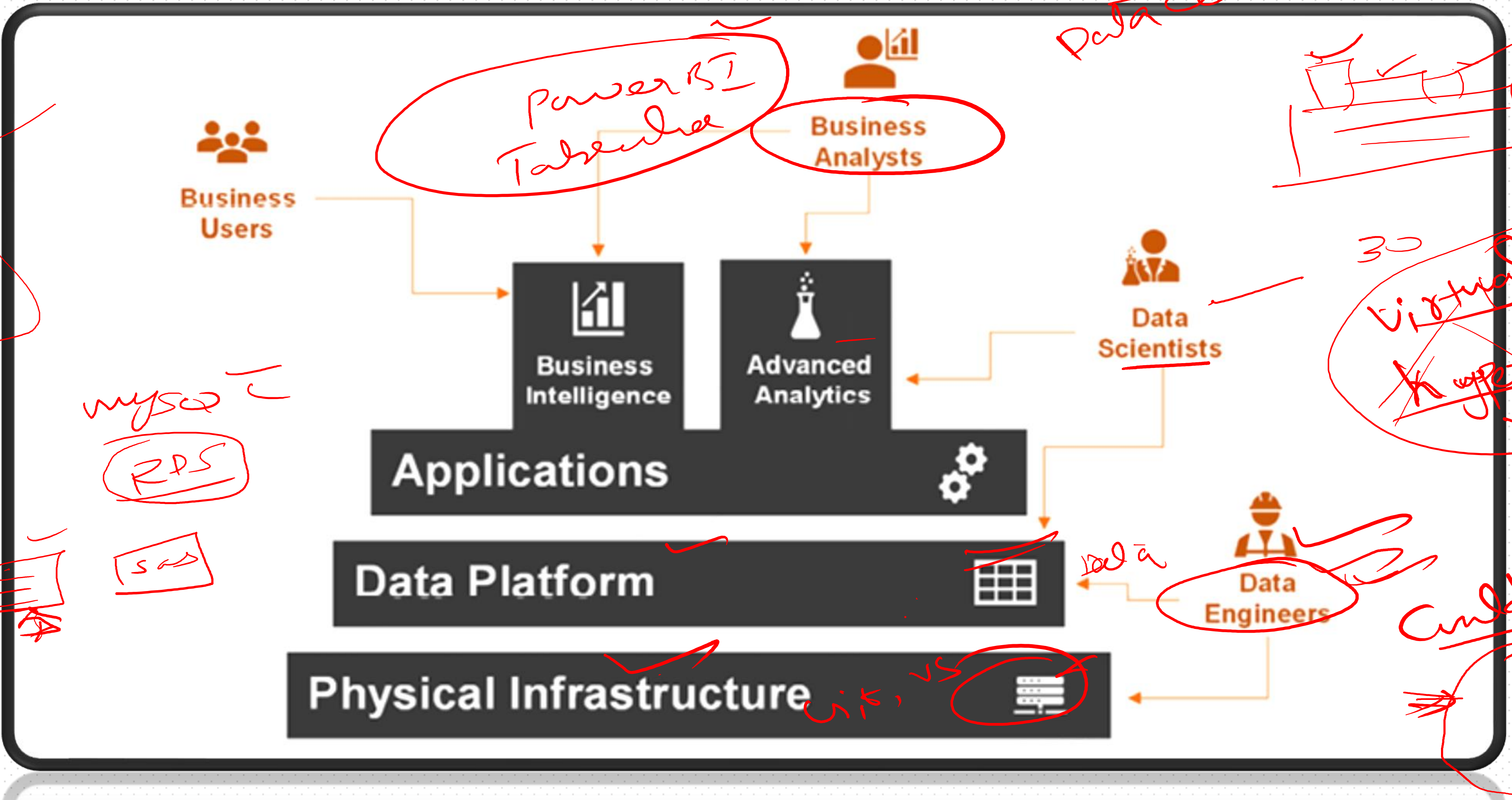
Business Use Cases

Credit Card Fraud Detection | Customer Segmentation | Pattern recognition

Data Analytics Overview

Data Analytics Process and Its steps | Skills and Tools Required for Data Analysis | Challenges in Data Analytic Processes | Data Visualization Technique | Exploratory Data Analysis Technique | Hypothesis Testing

Data Science Overview



Business Users

Business Analysts

Data Scientists

Data Engineers

Business Intelligence

Advanced Analytics

Applications

Data Platform

Physical Infrastructure

Definition of Data Science

- Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract or extrapolate knowledge and insights from noisy, structured and unstructured data, and apply knowledge from data across a broad range of application domains. Data science is related to data mining, machine learning, big data, computational statistics and analytics - Source Wikipedia

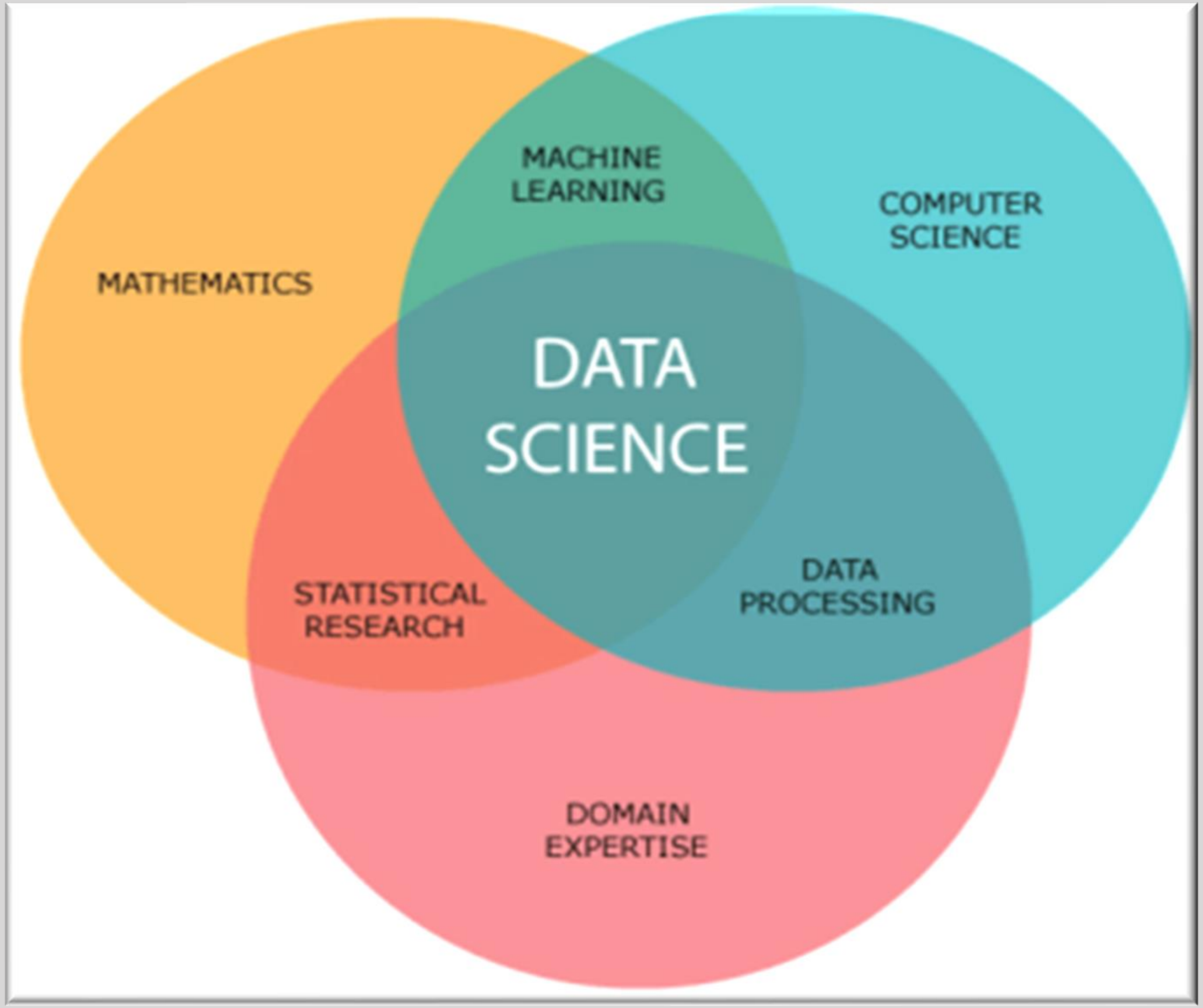
index	A_series	B_series	C_series	D_series
Current Value	12220	5855	12920	14200
Year-on-year	22%	14%	26%	32%
Early period	10%	5%	12%	22%

Definition of Data Science

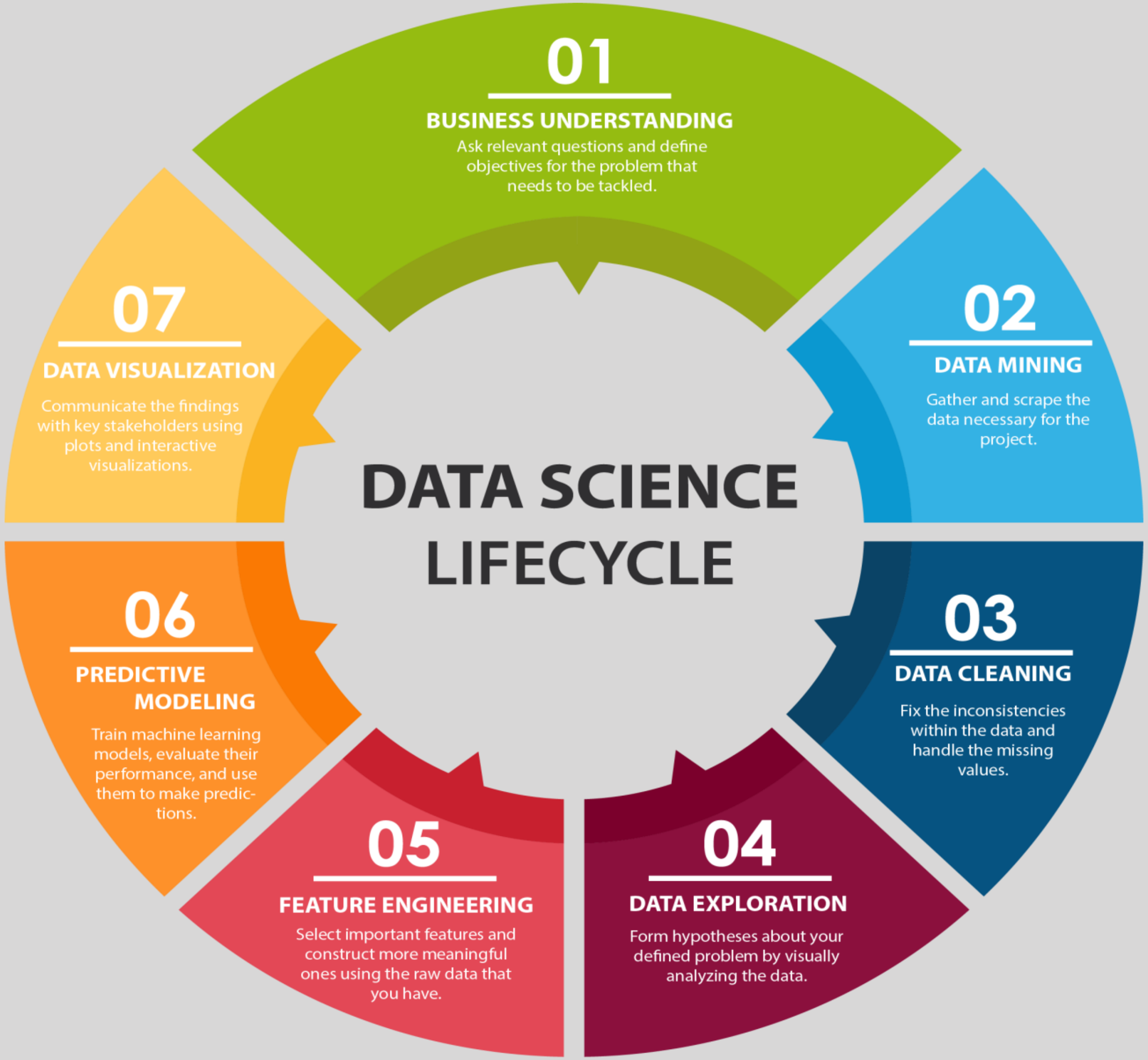
History Of Data Science

- n 1962, John Tukey wrote about the convergence of Statistics and computers to devise measurable outputs in hours.
- In 1974, Peter Naur mentioned the term ‘Data Science’ multiple times in his review, Concise Survey of Computer Methods.
- In 1977, the International Association for Statistical Computing (IASC) was formed to link modern computer technology, traditional statistical methodology, and domain expertise to convert data into knowledge.
- In the same year, Tukey composed a paper, Exploratory Data Analysis, that briefed the importance of using data.
- in 2013, IBM revealed that 90% of the global data had been created in the past two years.
- By this time, organizations realized the importance of Data Science to convert huge data clusters into usable information to gain crucial insights.

Definition of Data Science



Data Science Life Cycle



DATA SCIENCE LIFECYCLE

01

BUSINESS UNDERSTANDING

Ask relevant questions and define objectives for the problem that needs to be tackled.

02

DATA MINING

Gather and scrape the data necessary for the project.

03

DATA CLEANING

Fix the inconsistencies within the data and handle the missing values.

04

DATA EXPLORATION

Form hypotheses about your defined problem by visually analyzing the data.

05

FEATURE ENGINEERING

Select important features and construct more meaningful ones using the raw data that you have.

06

PREDICTIVE MODELING

Train machine learning models, evaluate their performance, and use them to make predictions.

07

DATA VISUALIZATION

Communicate the findings with key stakeholders using plots and interactive visualizations.

Data Science Overview

1. **Business Understanding**
   - asking the why's
   - understand the problem you are trying to solve
   - identifying the central objectives of your project by identifying the variables that need to be predicted.
   - If it's a regression, it could be something like a sales forecast
2. **Data Mining**
   - is the process of sorting through large data sets to identify patterns and relationships that can help solve business problems through data analysis.
   - is the process of gathering your data from different sources.
   - Algorithms

3. **Data Cleaning**
   - the most time-consuming step of all - cleaning and preparing the data
   - Get started: Pandas, Dplyr, Cleaning Dirty Data

4. **Data Exploration**
   - is like the brainstorming of data analysis
   - understand the patterns and bias in your data

Data Science Overview

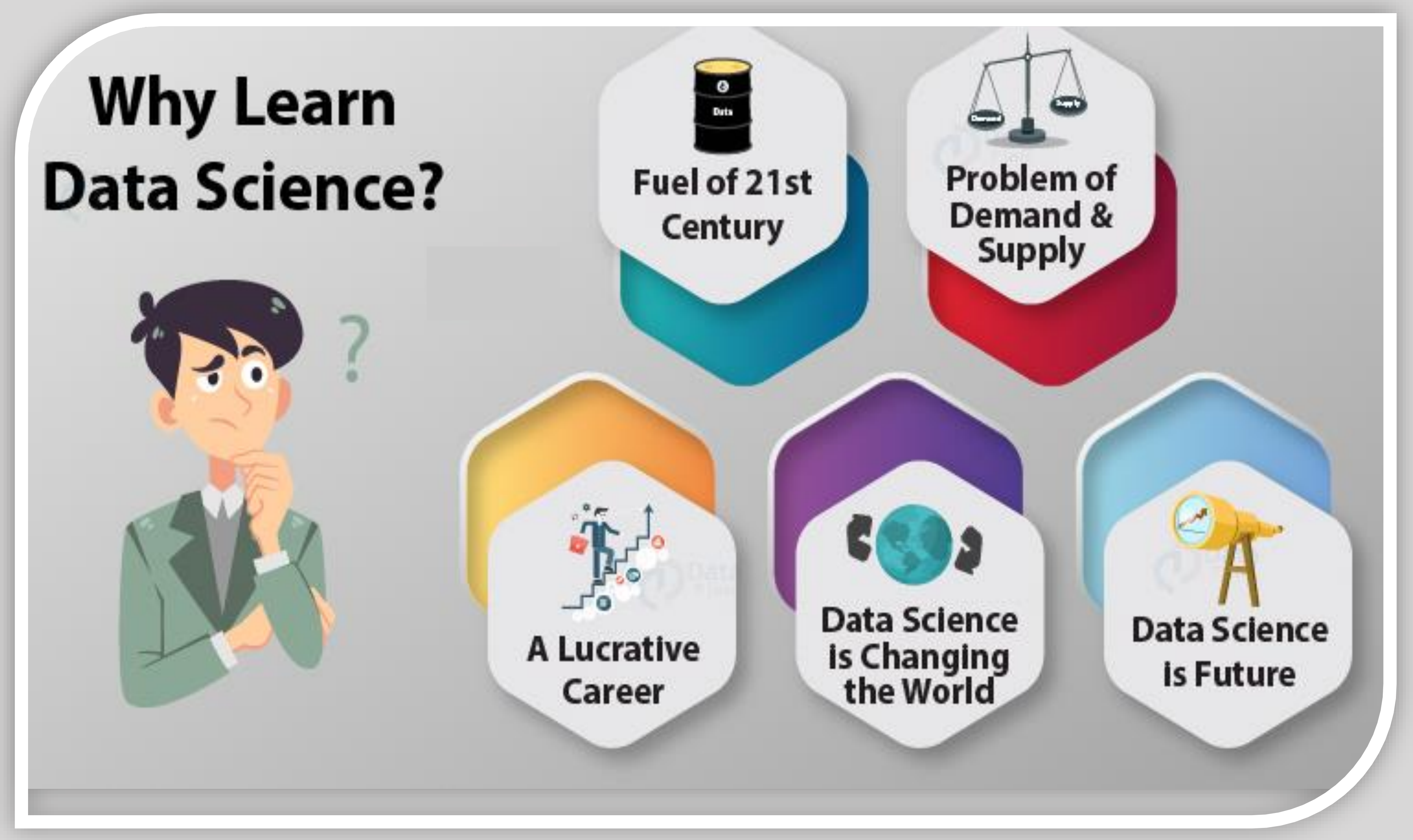
5. **Feature Engineering**
   - feature is a measurable property or attribute of a phenomenon being observed
   - features is difficult, time-consuming, requires expert knowledge

6. **Predictive Modeling**
   - Machine Learning Algorithms,
   - Evaluating Machine Learning Models

7. **Data Visualization**
   - Tableau, PowerBI, Plotly, Seaborn, Bokeh, D3.js

8. **Business Understanding**
   - This is where you evaluate how the success of your model relates to your original business understanding.
   - Does it tackle the problems identified?
   - Does the analysis yield any tangible solutions?

Data Science Importance



Why Learn Data Science?

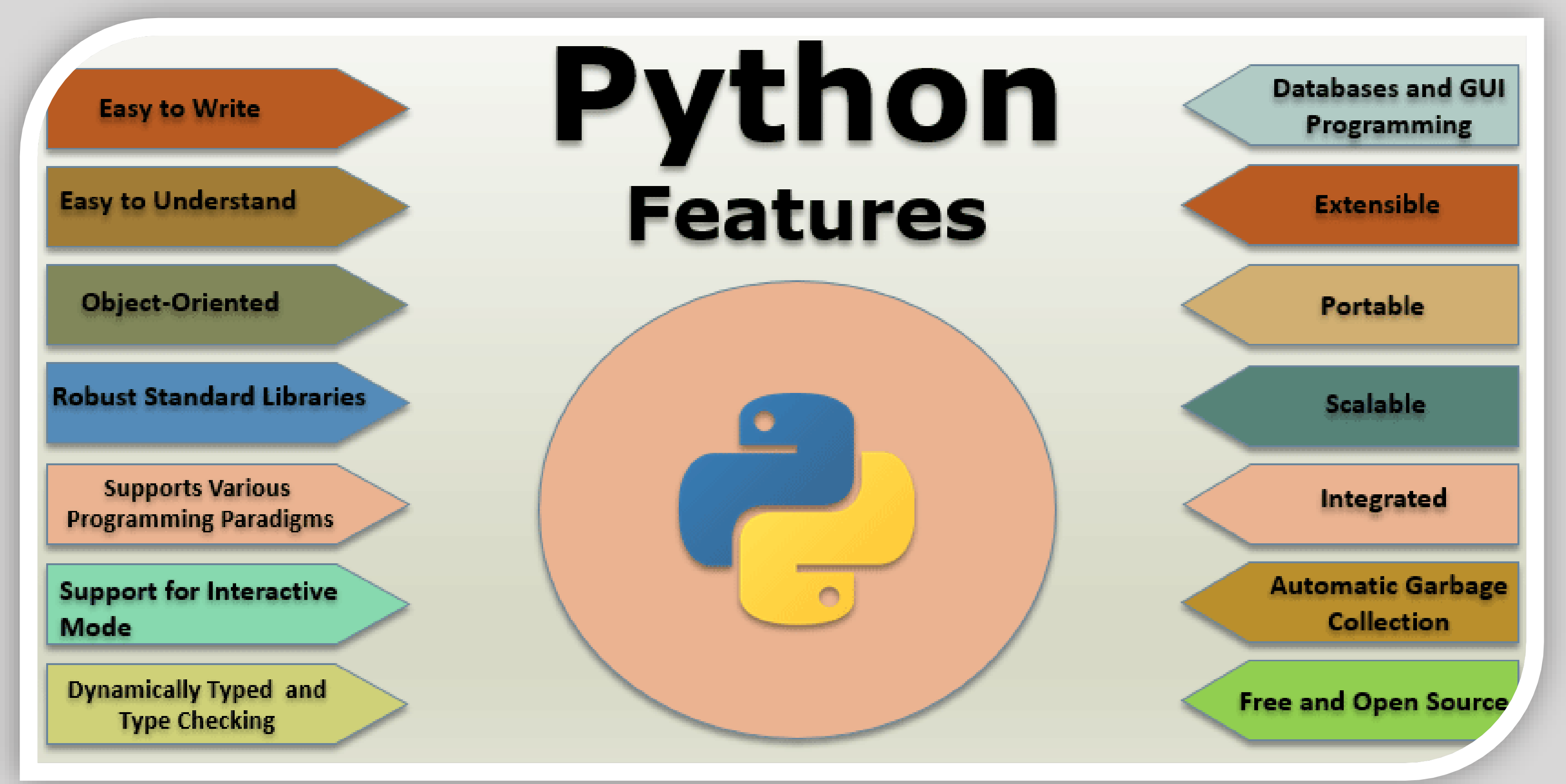
- Fuel of 21st Century
- Problem of Demand & Supply
- A Lucrative Career
- Data Science is Changing the World
- Data Science is Future

Data Science Importance

- Data is a precious asset of any organization.
- It helps firms understand and enhance their processes, thereby saving time and money.
- Wastage of time and money, such as a terrible advertising decision, can deplete resources and severely impact a business.
- The efficient use of data enables businesses to reduce such wastage by analyzing different marketing channels' performance and focusing on those offering the highest ROI.
- Thus, a company can generate more leads without increasing its advertising spend.

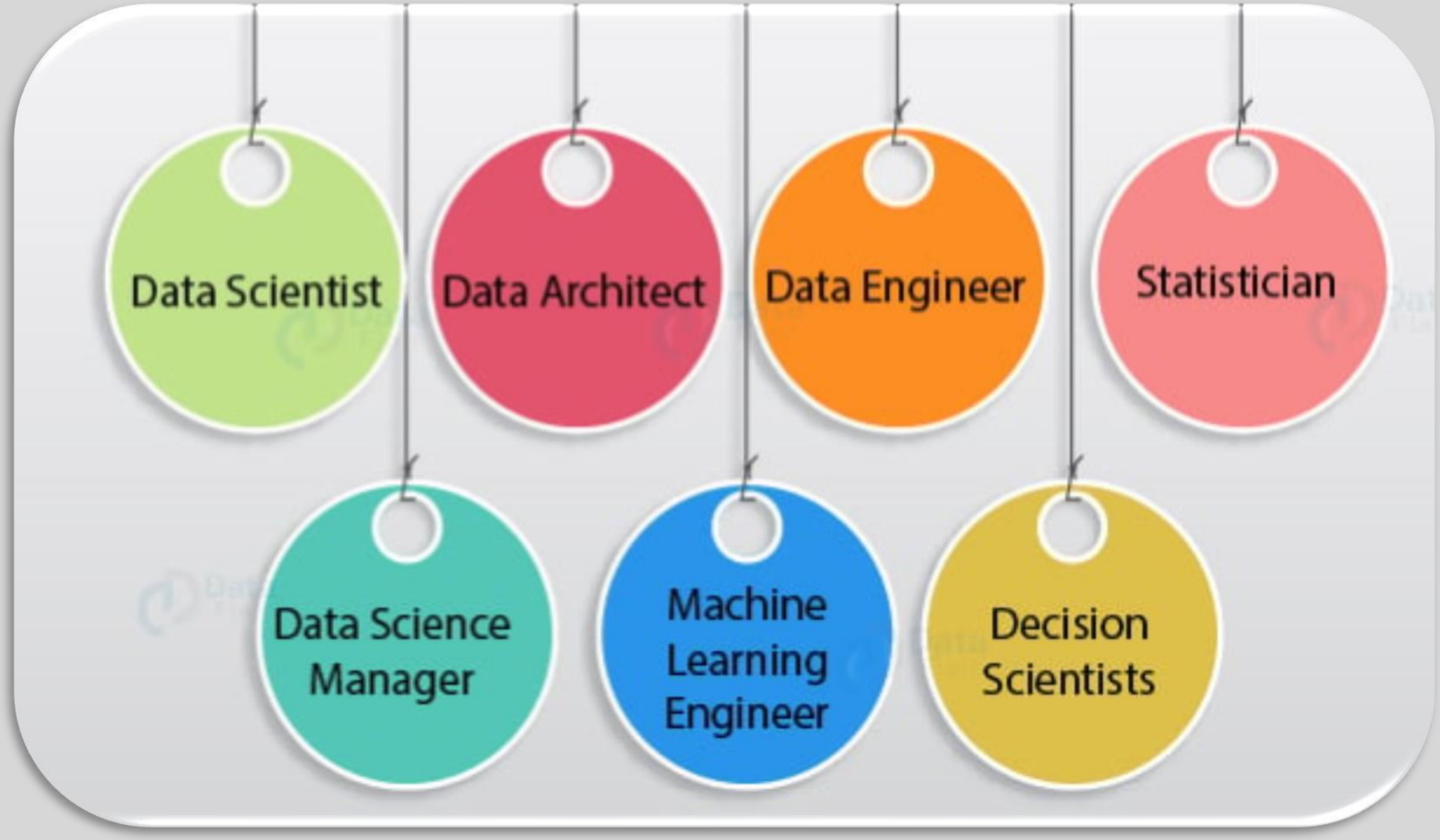
Data Science Python and Importance

- Python is an object-oriented, open-source, adaptable and simple to learn programming language
- Rich arrangement of libraries and tools
- Python has an enormous community base
- top choice for Data scientists and Developers.
- for example, **Google, Dropbox, Instagram, YouTube,** and **Spotify** — all were worked with Python.


Python
Features
Easy to Write
Easy to Understand
Object-Oriented
Robust Standard Libraries
Supports Various Programming Paradigms
Support for Interactive Mode
Dynamically Typed and Type Checking
Databases and GUI Programming
Extensible
Portable
Scalable
Integrated
Automatic Garbage Collection
Free and Open Source

Roles and responsibilities of Data Scientist

Different Roles in Data Science



Roles and responsibilities of Data Scientist

Different Roles in Data Science

- Data Scientist
  a) finding insights and patterns in the data
  b) handling raw data, analyzing the data, implementing various statistical procedures, visualizing the data and generating insights from it.
  c) handling both structured and unstructured information.
  d) knowledge of various tools like Hadoop, R, Python, SAS, etc.
  e) Knowledge of data preprocessing, visualization and prediction

- Data Architect
  a) implementing the blueprints of a company's data platform.
  b) organizing and managing data both at the macro level as well as the micro level.
  c) tools used by a Data Architect are XML, Hive, SQL, Spark and Pig

Roles and responsibilities of Data Scientist

Different Roles in Data Science

- Data Engineer
  a) building big data pipelines and models for the data scientists to work on
  b) structured as well as unstructured data
  c) maintaining, managing and testing
  d) Knowledge of database models and ETL are two of the most essential requirements for a Data Engineer
  e) responsible for modelling large-scale processing systems using tools like SQL, Hive, Pig, Python, Java, SPSS, SAS etc.

- Data Science Manager
  a) handling and managing data science projects
  b) planning and curating a roadmap for the data science team to follow
  c) executing the plan of action and delivering the results before the deadline.
  d) strong communication and leadership skills in order to guide the team efficiently

Roles and responsibilities of Data Scientist

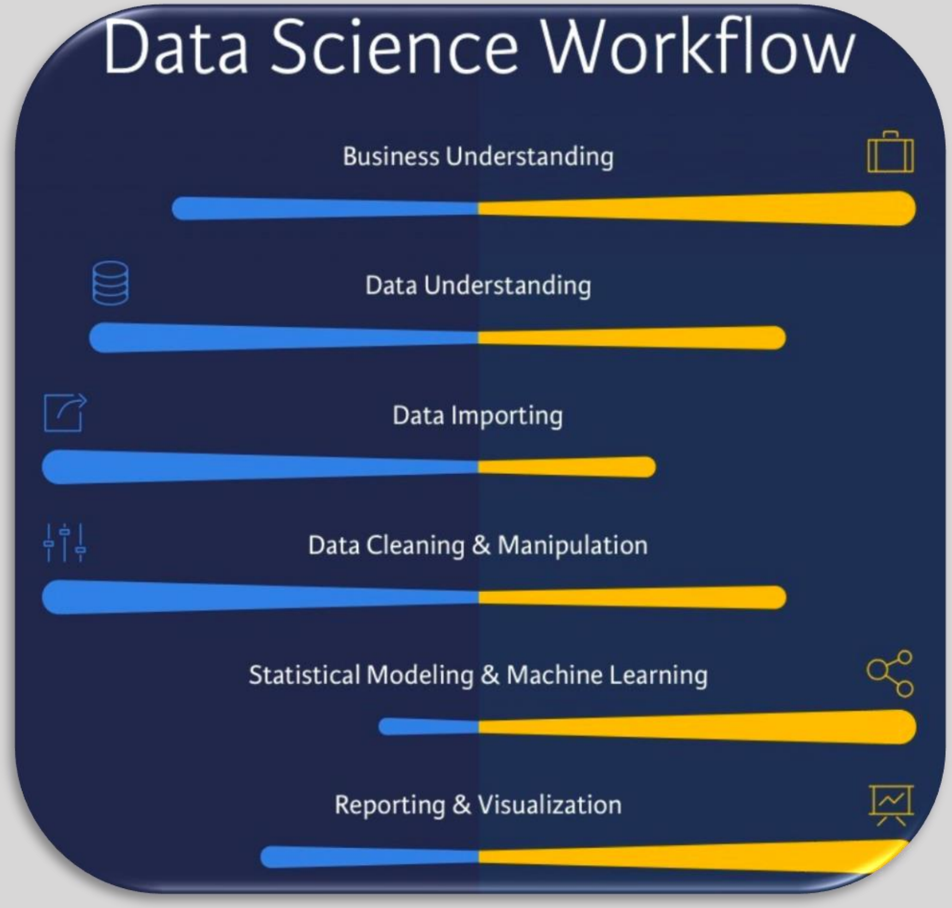
Different Roles in Data Science

- Statistician
  a) oldest job title
  b) Before data science, statisticians were employed by the companies to use statistical modeling for understanding various trends in the market.
  c) implementing A/B testing, harvesting data, describing data, developing inferential statistical tools and performing hypothesis testing.
  d) tools used by statisticians are R, SAS, SPSS, Matlab, Python, Stata, SQL etc.
- Machine Learning Engineer
  a) machine learning models for performing classification and regression tasks.
  b) knowledge of various techniques like clustering, random forest and several other deep learning algorithms.
  c) tools used by the machine learning engineers are TensorFlow, Keras, PyTorch, scikit-learn, Caffe etc.

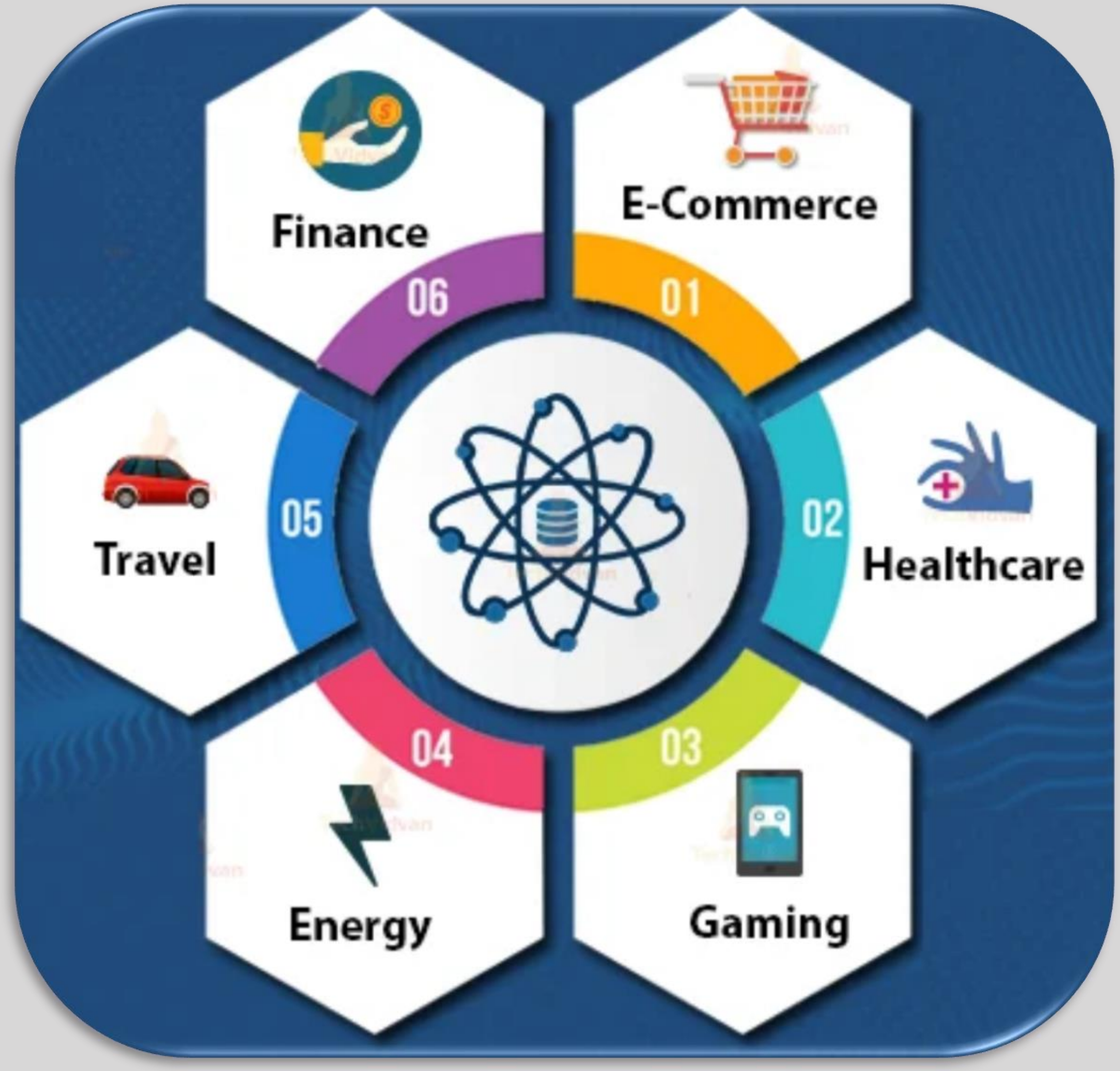
Roles and responsibilities of Data Scientist

Different Roles in Data Science

- Decision Scientists
  a) The field of decision science is a relatively new field.
  b) make business decisions with the help of tools like Artificial Intelligence and Machine Learning.
  c) It is a part of data science that extends to design thinking and behavioral sciences to better understand the clients.


Data Science Workflow
Business Understanding
Data Understanding
Data Importing
Data Cleaning & Manipulation
Statistical Modeling & Machine Learning
Reporting & Visualization

Various application of Data Science



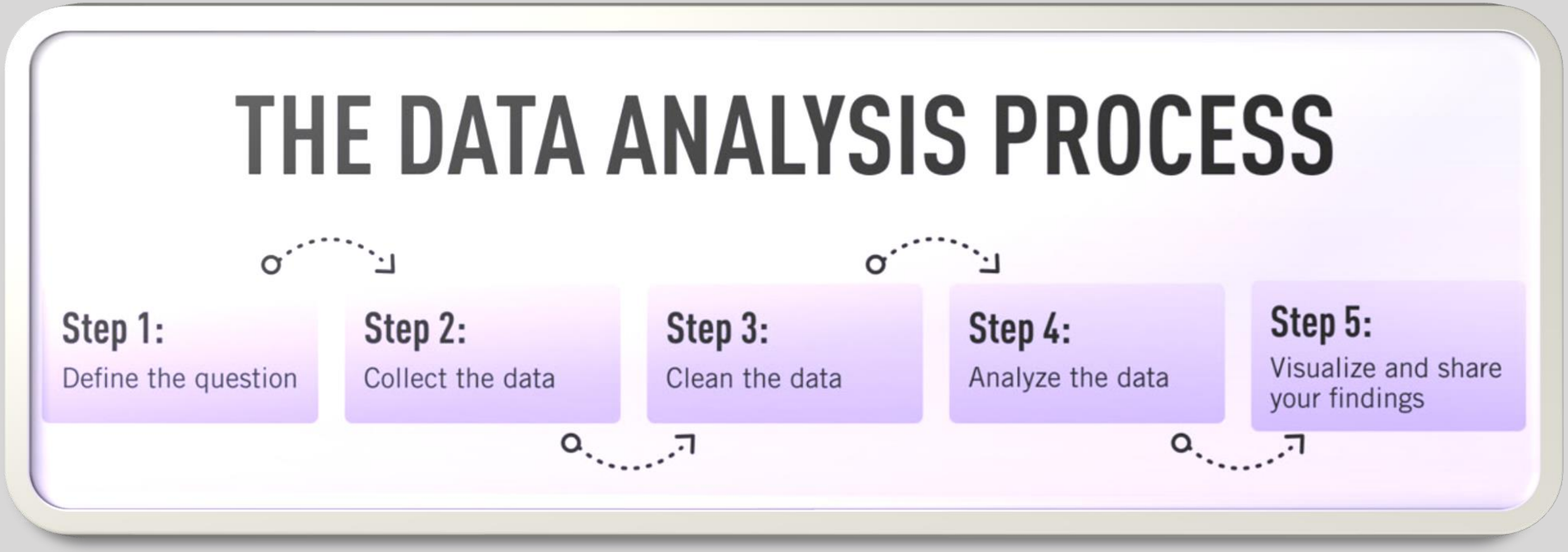
Data Science and Machine Learning job opportunity

Job opportunities available to you as a Data Scientist include:

- Estimated 2.7 million open jobs in data analysis, data science and related careers in 2020 (source: IBM).
- 39% growth in employer demand for demand for both data scientists and data engineers by 2020 (source IBM).
- 59% of jobs will be in finance, information technology (IT), insurance and professional services careers. This is broken down to 19% in finance and insurance, 18% in professional services and 17% in IT.
- Data science and data analyst jobs remain open for 5 days longer than the average for all other jobs, which means there's less competition in these fields, and recruiters have to work harder to find qualified candidates.

Data Analytics Overview

Data Analytics Process and its Steps



THE DATA ANALYSIS PROCESS

Step 1: Define the question

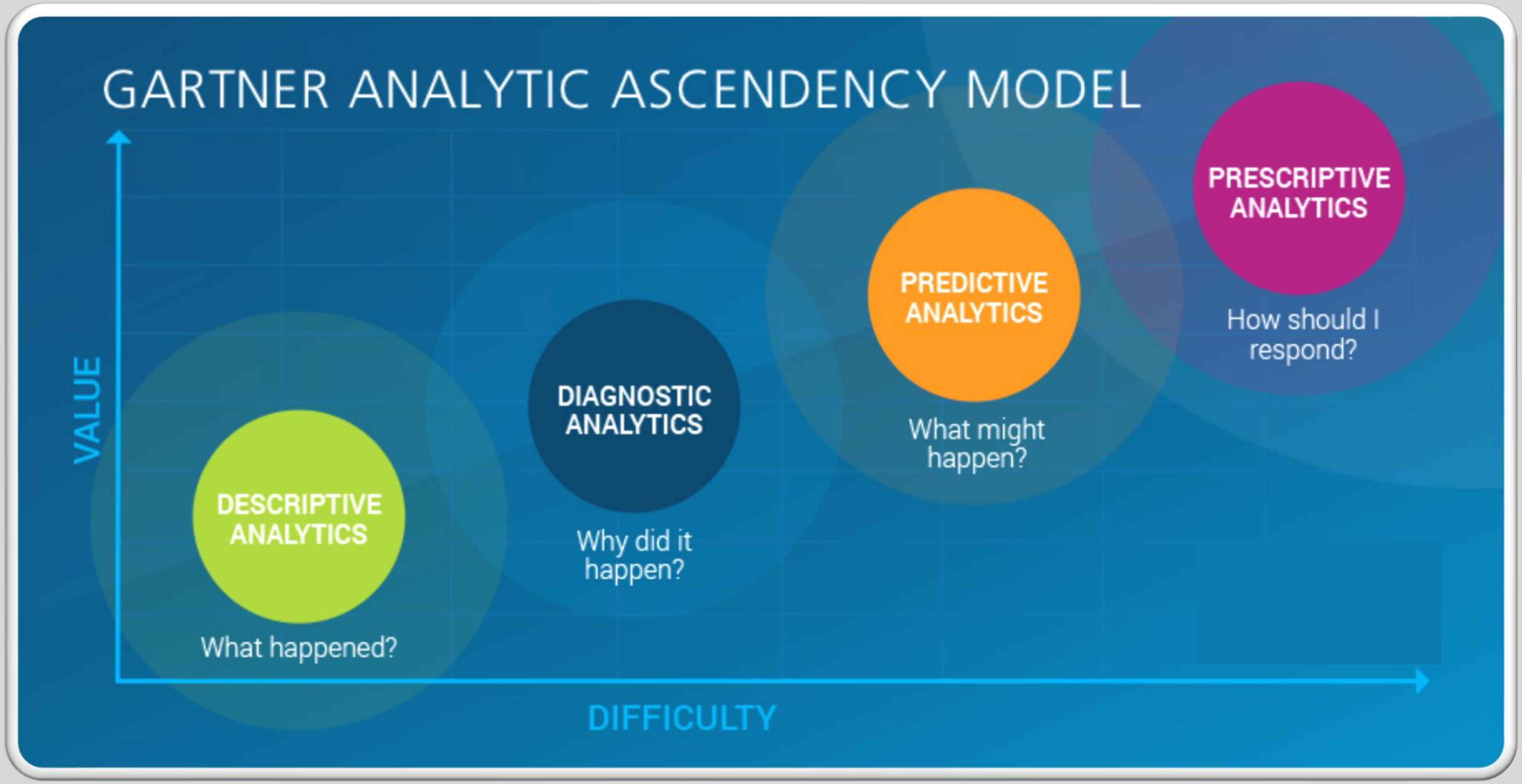
Step 2: Collect the data

Step 3: Clean the data

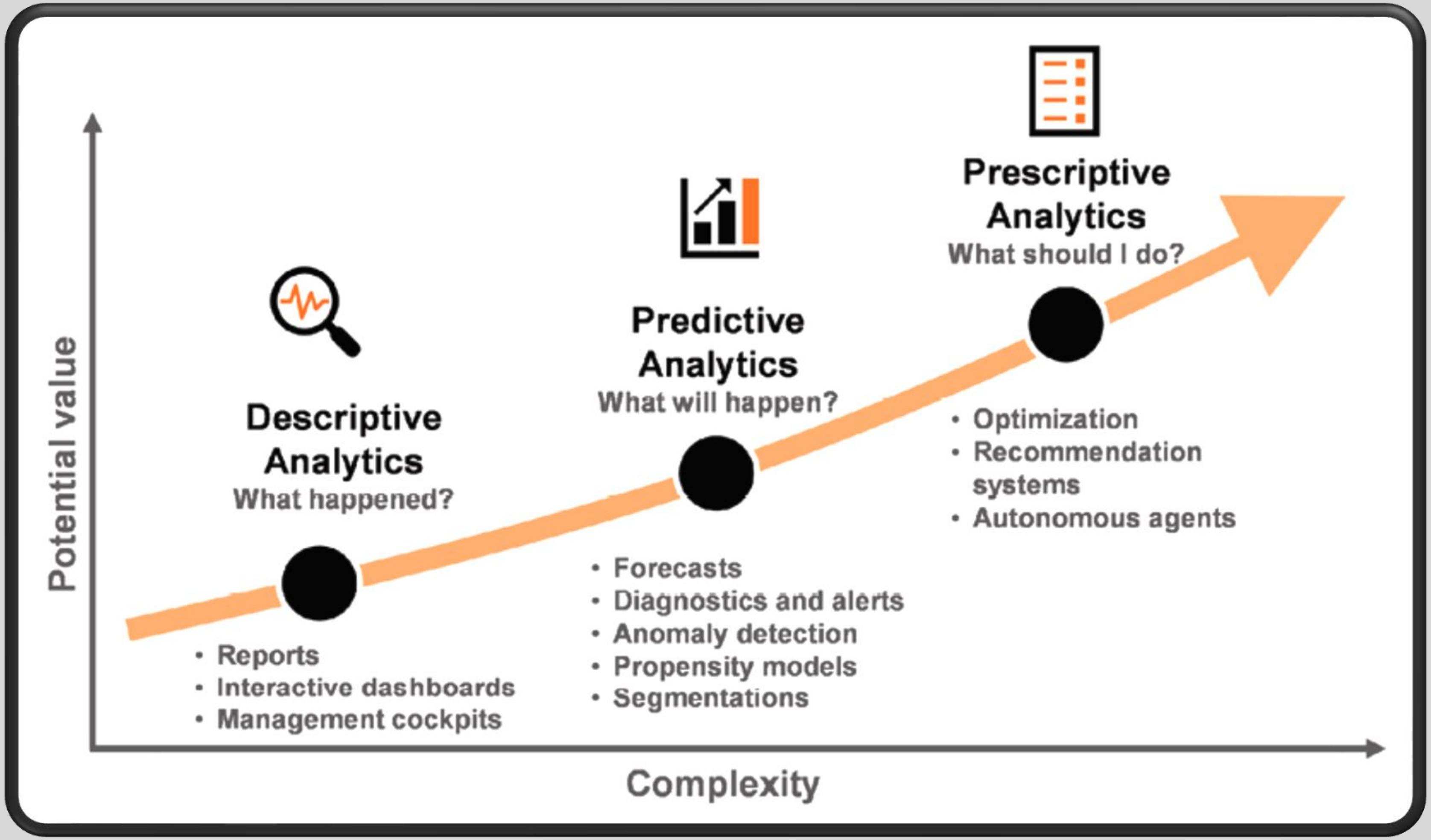
Step 4: Analyze the data

Step 5: Visualize and share your findings

Data Analytics Overview



Data Analytics Overview



Descriptive Analytics
What happened?

- Reports
- Interactive dashboards
- Management cockpits

Predictive Analytics
What will happen?

- Forecasts
- Diagnostics and alerts
- Anomaly detection
- Propensity models
- Segmentations

Prescriptive Analytics
What should I do?

- Optimization
- Recommendation systems
- Autonomous agents

Data Science Overview

Descriptive analytics

- describing past data to make it digestible and useable as required by the business need.
- "what happened?" by leveraging summary statistics like average, median, and variance &
- simple transformations and aggregations like indices, counts, and sums
- ultimately displaying the results through tables and visuals.
- Key Performance Indicators (KPIs)

Diagnostic Analytics:

- "why did it happen?"
- By using basic methods like correlation analysis, control charting, and tests of statistical significance,
- these tools can highlight interesting patterns, shedding light on the reasons why the business is going in a certain way.

Data Science Overview

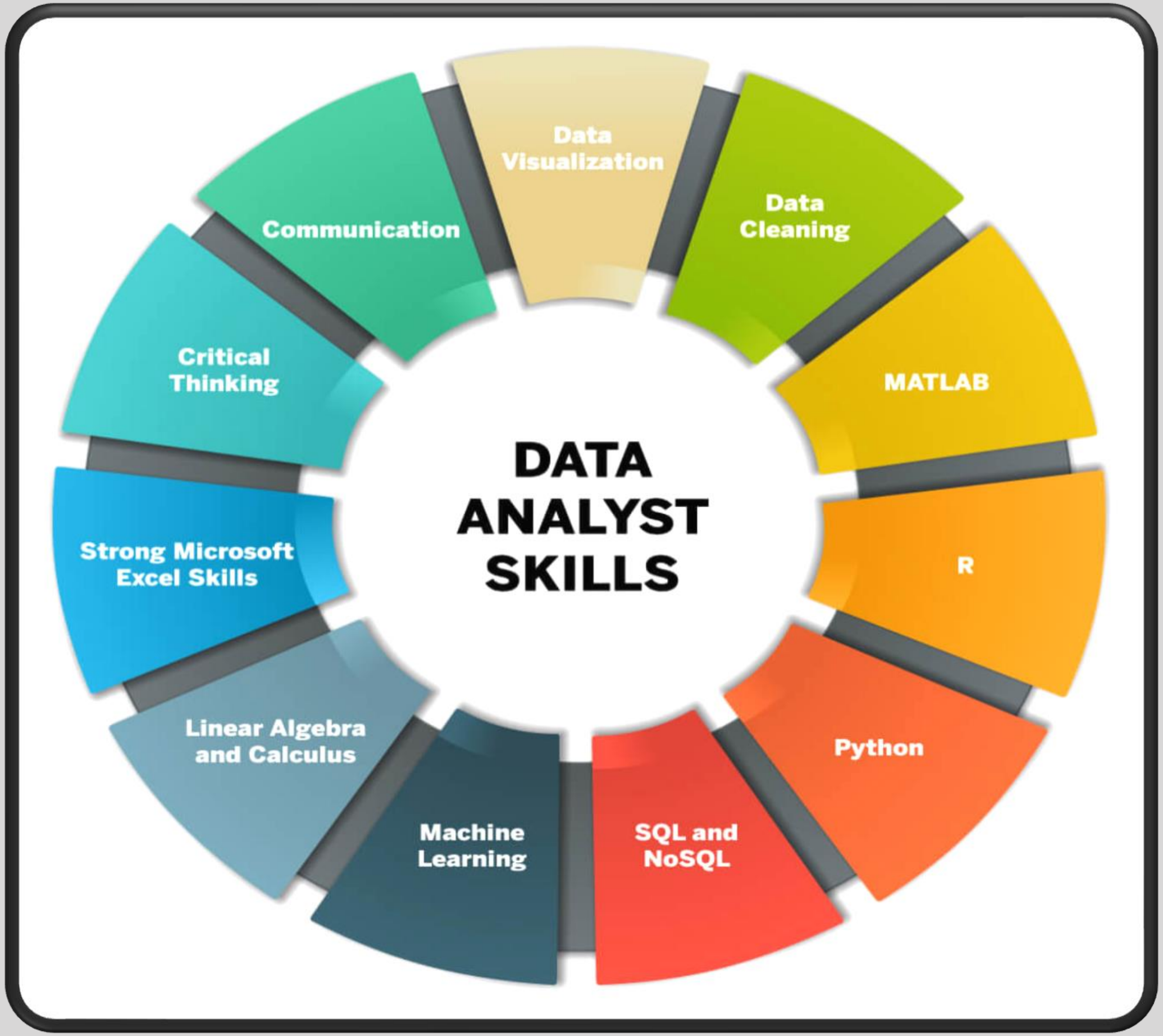
Predictive analytics

- "what will happen now?".
- These methodologies leverage more sophisticated techniques, including AI, to go beyond the mere description of historical facts.
- By using them, we can make sense of the causal relationships that lie under our data and extrapolate them, to show what the future is most likely going to look like.
- anomaly detection
- Return on Investment (ROI)

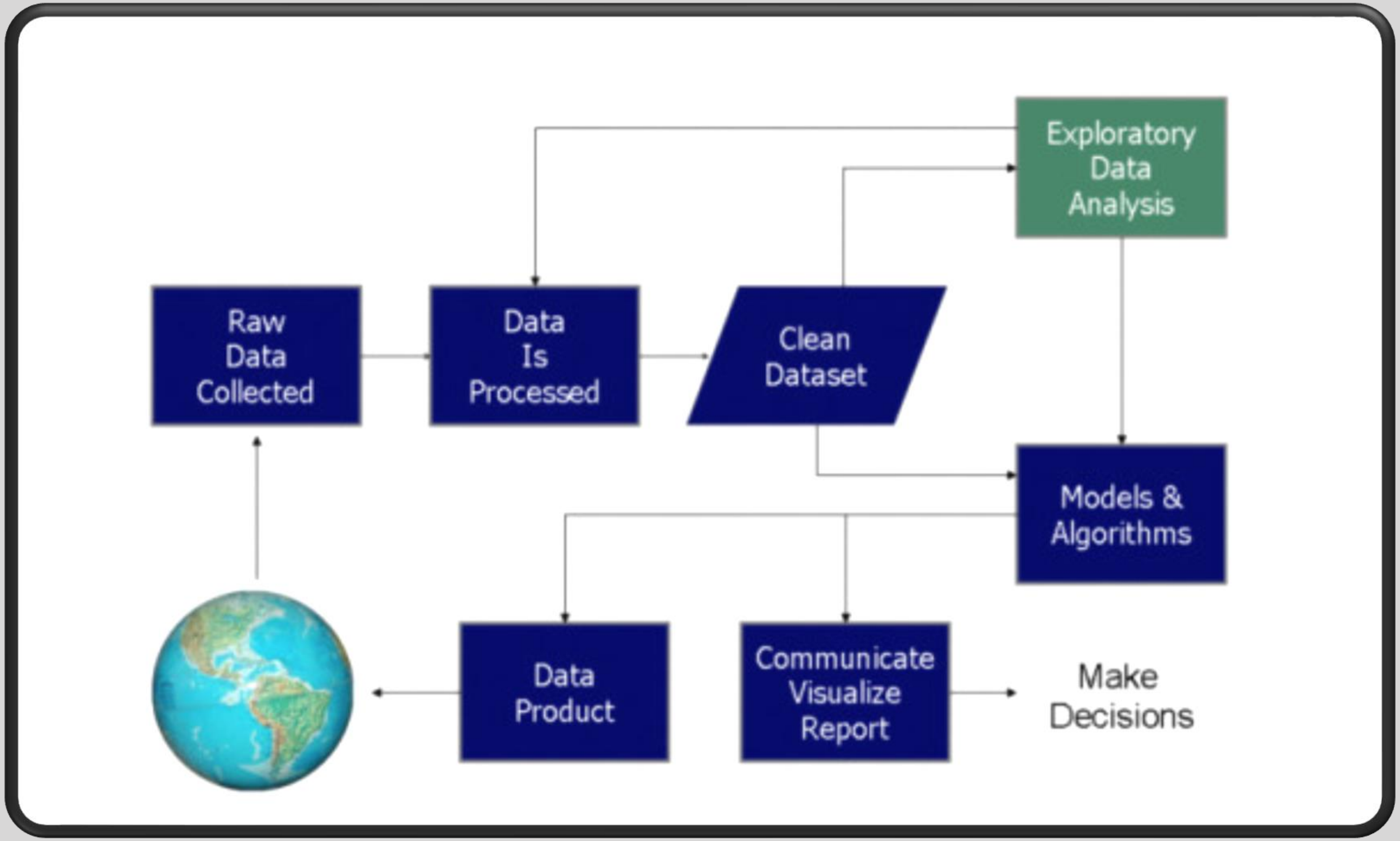
Prescriptive analytics

- "what should be done?".
- If descriptive and predictive analytics produces insights and informs us about our business, prescriptive analytics is certainly more assertive and direct: it tells us what to do.
- systematic optimization
- recommendation systems
- autonomous agents

Skills and Tools Required for Data Analysis



Challenges in Data Analytic Processes



Data Visualization Technique

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview

Data Science Overview