Parul® University

Data Visualization and Data Analytics

Prof. Khushbu Chauhan, Assistant Professor
Information Technology (PIET)

CHAPTER 1

Introduction

Outlines

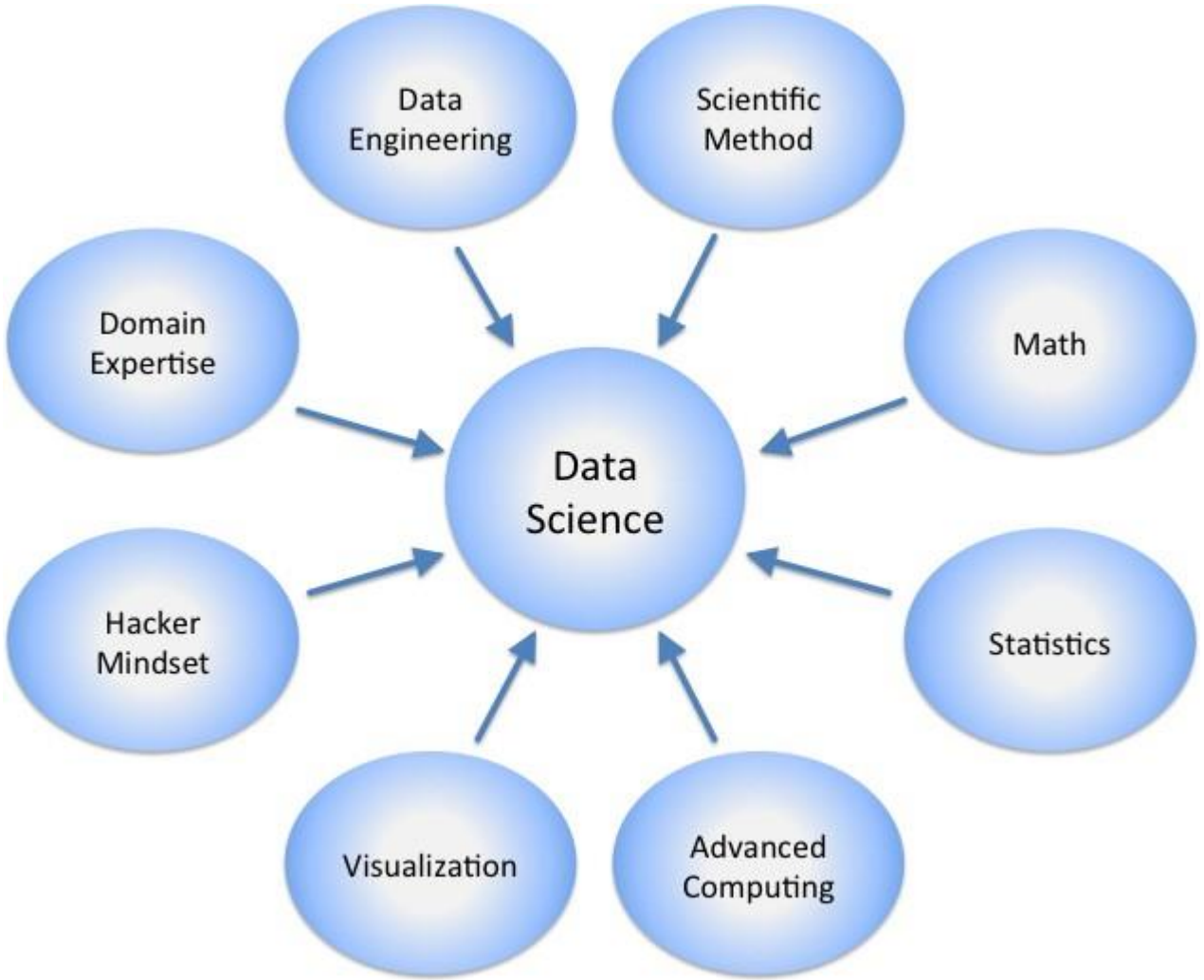
- Introduction to Data & Data Science
- Buzzwords of Data Science
- Introduction to Data Visualization
- Info-graphic representation of terminologies
- Difference between Analytics and Analysis
- DIKW (Data, Information, Knowledge, Wisdom)
- Applications

What is Data

- In computing, data is information that has been translated into a form that is efficient for movement or processing. Relative to today's computers and transmission media, data is information converted into binary digital form. It is acceptable for data to be used as a singular subject or a plural subject. Raw data is a term used to describe data in its most basic digital format.
- Data can be meaningful or meaningless. Data is in the form of values, images, audio, video, graphics, etc.

What is Data Science

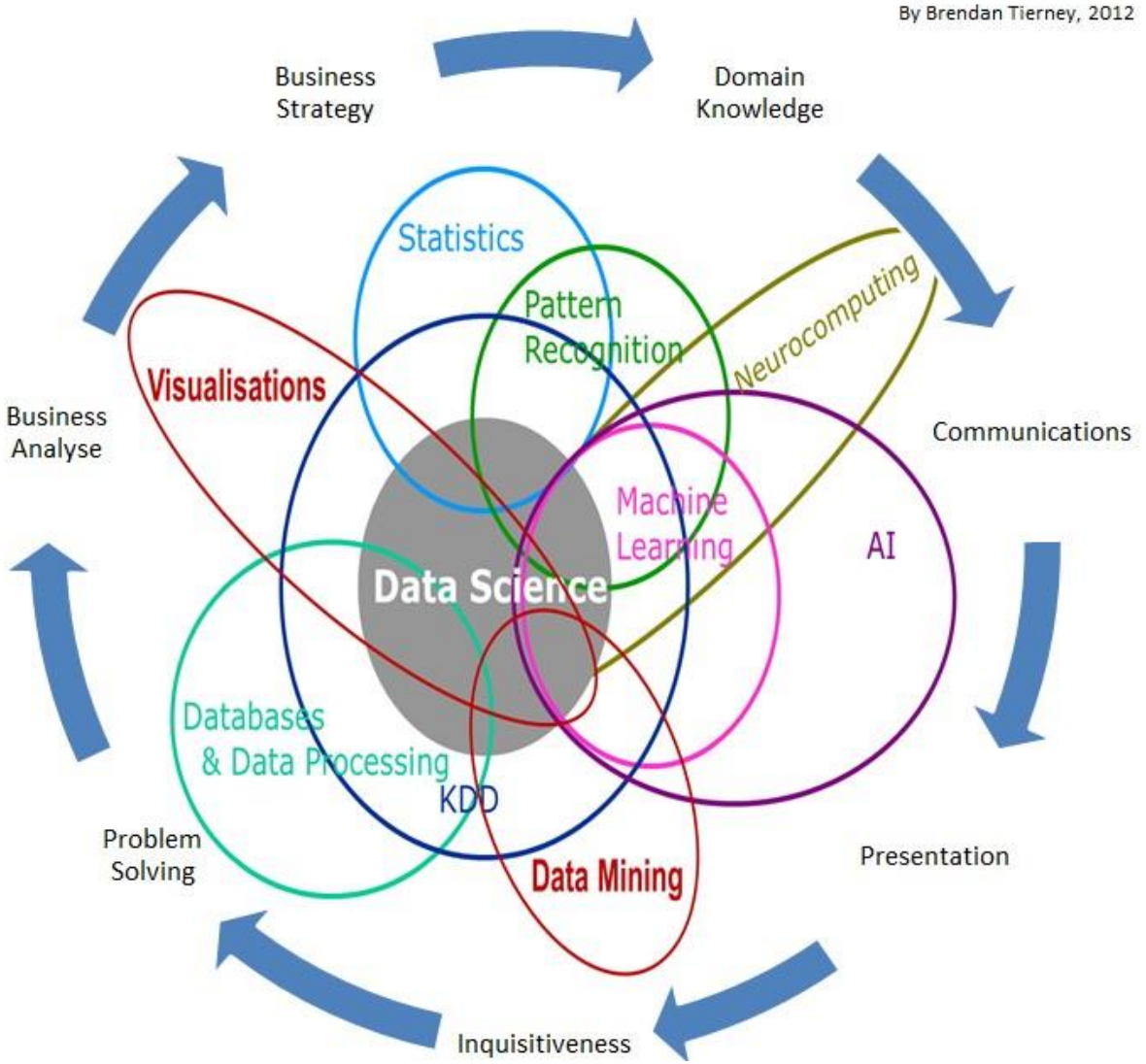
- Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data. Data science practitioners apply machine learning algorithms to numbers, text, images, video, audio, and more to produce artificial intelligence (AI) systems to perform tasks that ordinarily require human intelligence. These systems generate insights that analysts and business users can translate into tangible business value.


Data Engineering
Scientific Method
Domain Expertise
Math
Data Science
Hacker Mindset
Statistics
Visualization
Advanced Computing

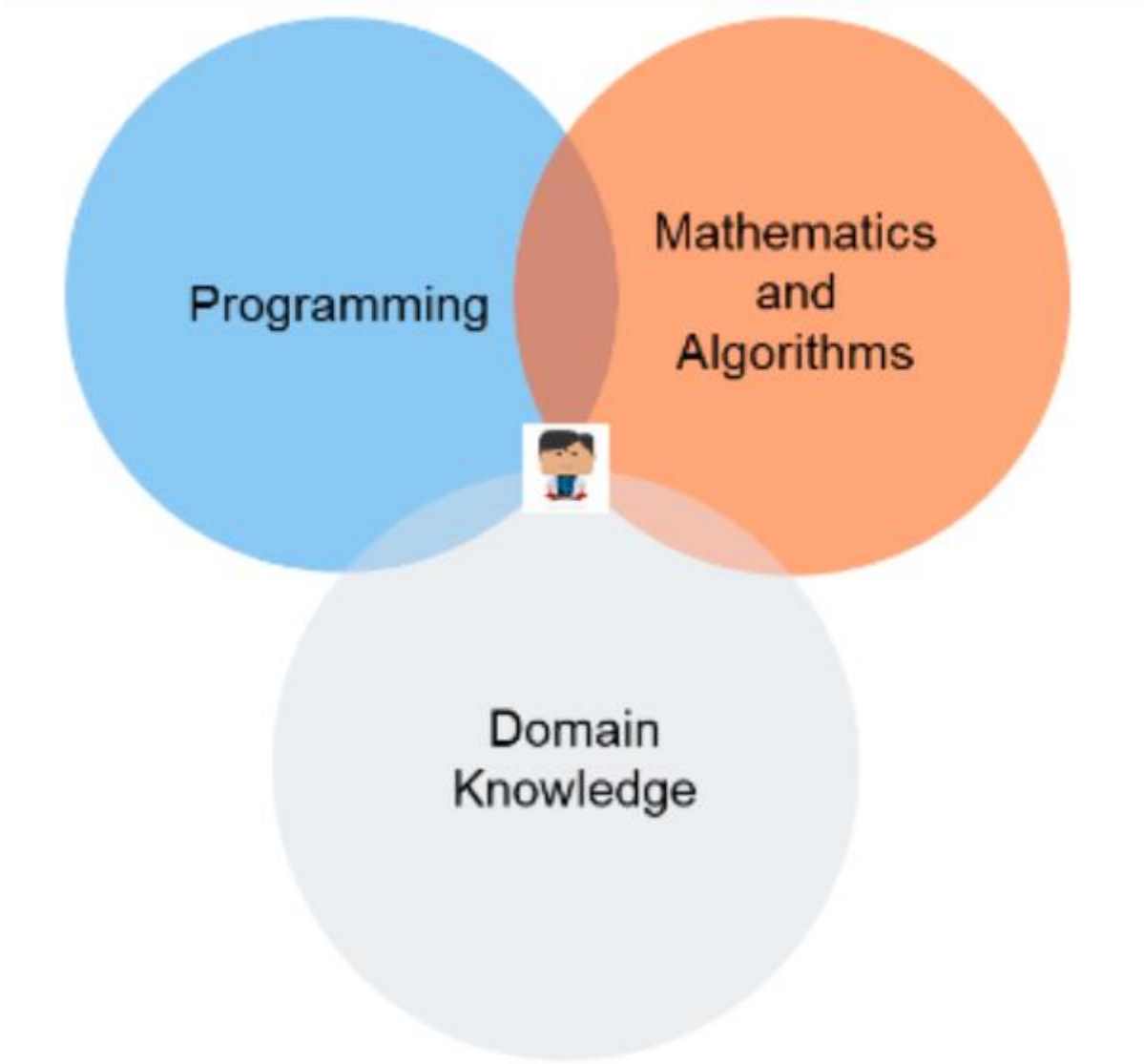
Why Data science

- More and more companies are coming to realize the importance of data science, AI, and machine learning. Regardless of industry or size, organizations that wish to remain competitive in the age of big data need to efficiently develop and implement data science capabilities or risk being left behind.
- Data science uses mathematical function tools like statistics and probability, linear algebra, etc.

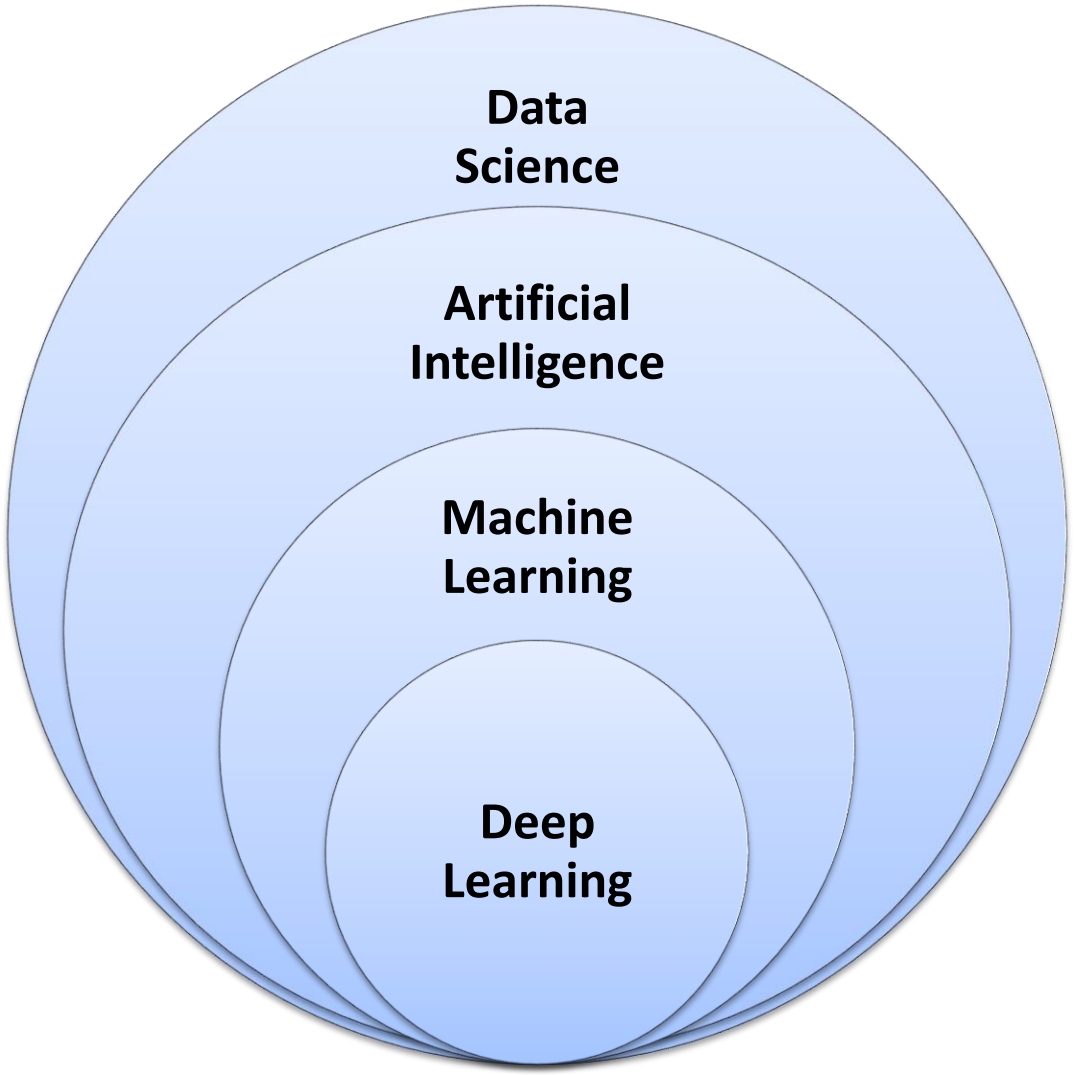
Data Science Is Multidisciplinary



Analysis steps



Buzzwords of data science



1. Big Data :

- The term is used to describe repositories of data so huge that they defy traditional processing techniques. More broadly, the term ‘big data is also used to describe the methods relating to the management of these types of datasets. The sources of big data include everything from credit cards to surveillance, social media, and electronic communications. Big data is often categorized into three types: social data (from social media platforms), machine data (generated by computers and devices), and transactional data (where data is exchanged between two parties)


Volume
Size of Data
Velocity
The Speed at which Data is Generated
Variety
Different type of Data
Veracity
Data Accuracy
Value
Useful Data
BigData
Validity
Data quality, Governace, Moster Data Management on Massive
Variability
Dynamic, Evolving Behavior in Data Source
Venue
Distributed Heterogeneous Data from Multiple Platforms
Vocabulary
Data Models, Semantics that describes data Structure
Vagueness
Confusion over Meaning of BigData and Tools used

2. Structured and Unstructured Data :

- All data is either ‘structured’ or ‘unstructured.’ Most data starts life in an unstructured format: disorganized, text-heavy, and without any underlying configuration. Big data, in particular, is usually unstructured. This makes it hard to navigate or use. To get a dataset into a useful format that can be analyzed, we must structure it. Structured data has been organized into databases, spreadsheets, or content management systems. It is often ordered into rows and columns, making it much easier to navigate.

3. Algorithm :

- It is a carefully defined, step-by-step process that is used to solve logical or mathematical problems. In terms of data analytics, these algorithms are carried out by computers. In day-to-day life, any task you run on your phone or laptop will be carried out by an algorithm. Within data analytics, algorithms are used to streamline tasks that computers can carry out much faster and more accurately than humans. This makes them perfect for sorting, parsing, or analyzing big datasets.

4. Predictive and prescriptive analytics :

- Predictive analytics uses algorithms to predict (or make informed guesses about) what is likely to happen in the future, based on existing data. Meanwhile, prescriptive analytics recommends a course of action based on these predictions. Courses of action may be intended to shape future outcomes or to simply take advantage of existing ones. The power of predictive and prescriptive analytics is commonly used by businesses to drive profit, improve customer experiences, and stay ahead of the competition.

5. Descriptive analysis :

- Descriptive analysis is used to find correlations between different groups of data. It is used to take raw data and summarize it into groupings that are easier to interpret and understand. It’s widely used in business intelligence and by other data analysts to predict trends.

6. Artificial Intelligence :

- Artificial intelligence, or AI, has long been a term used by science-fiction aficionados. Now it has come out of the world of fiction and become reality.
- AI enables the machine to think and learn by its own without the intervention of humans.

7. Machine Learning :

- Machine learning (ML) algorithms learn without being explicitly programmed to do so. Machine learning first evolved in response to the wider pursuit of artificial intelligence. It has fast become a specialized field in its own right. Since ML algorithms ingest and learn from large amounts of data, they're popular for predictive analytics and for making sense of big data. Machine learning is also commonly used to carry out tasks that would be impractical or time-consuming for people to carry out (for instance, managing search engine results).

8. Deep Learning :

- Deep learning is a subset of machine learning. It commonly uses unsupervised learning techniques, in the form of algorithms that mimic the workings of the human brain. As a result, it can solve highly complex tasks in a short amount of time. Deep learning is the closest thing we currently have to human-like artificial intelligence. Deep learning tends to improve with greater amounts of data. As such, whereas less complex approaches rely on data that is mathematically straightforward to process, deep learning can find patterns in data that even humans cannot spot.

9. Internet of Things :

- Internet of Things or IoT, is the term used to describe the network of things that contain software, sensors, or other means to connect to other devices containing these technologies through the internet. These devices remain connected and may continuously transmit data back and forth without human intervention.

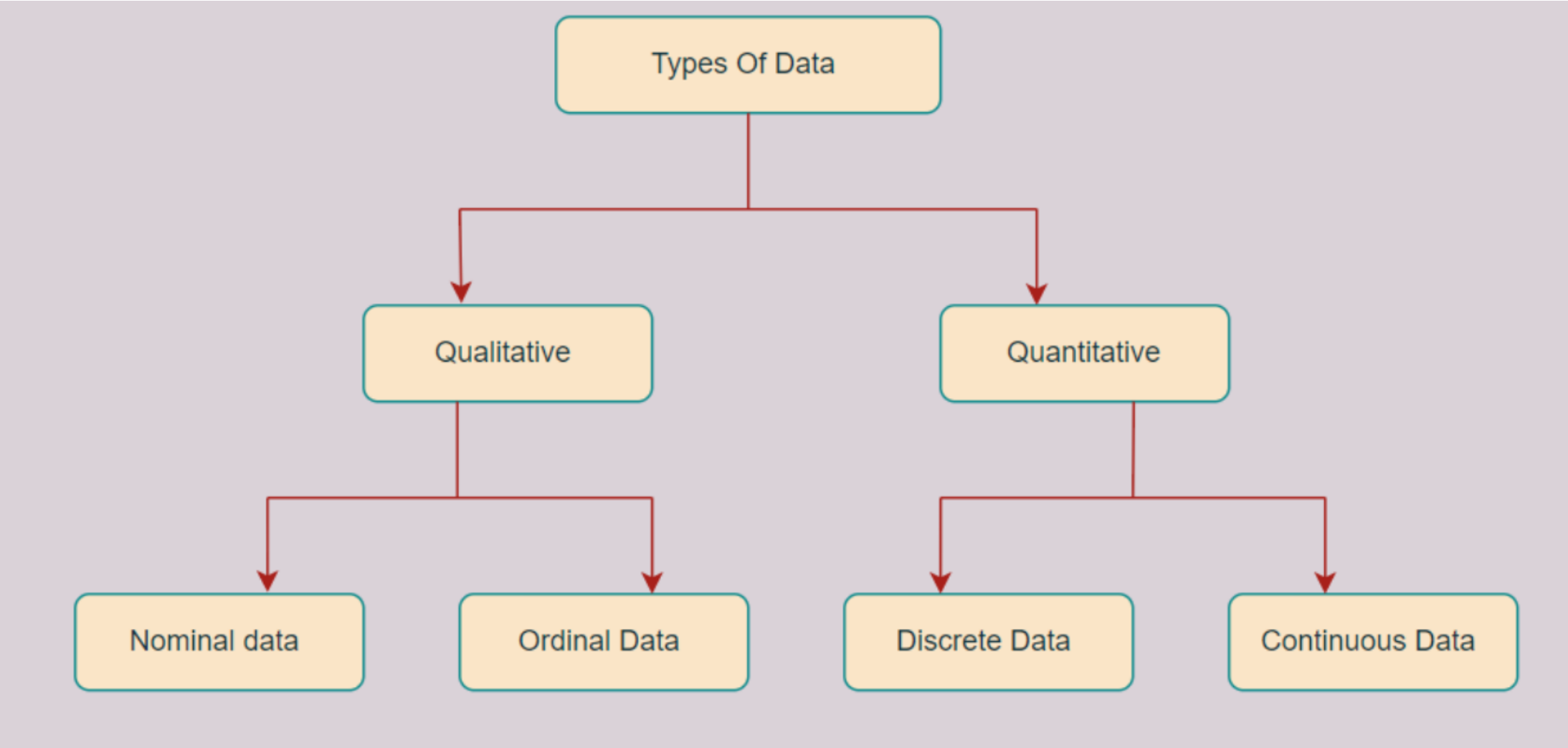
10. Data wrangling :

- Data wrangling is the process of collecting raw data, cleaning it, mapping, and storing it in a useful format. A common buzzword, there is often confusion around the term, since it is also commonly used as a catch-all to describe other stages in the data analytics process. This includes planning what data to collect, the process of creating algorithms to collect these data, carrying out our exploratory analysis, quality control, creating data structures, and so on.

11. Regression and classification :

- Regression and classification are two types of models commonly used in predictive analytics. Classification is used in machine learning to predict or identify discrete categories of data. Meanwhile, regression models are used to identify continuous values of data. Without getting into the details, the broad takeaway here is that classification is about predicting labels (e.g. red, blue, green, etc.), while regression is about predicting quantities (e.g. probabilities, temperature, or sums, etc.).

Types of Data



- **Examples of Qualitative data are :**
- What language do you speak
- Favourite holiday destination
- Opinion on something (agree, disagree, or neutral)
- Colours

- **Examples of Quantitative Data :**
- Height or weight of a person or object
- Room Temperature
- Scores and Marks (Ex: 59, 80, 60, etc.)
- Time

Nominal Data

- **Nominal** - Nominal data are recorded as categories. For this reason, nominal data is also known as categorical data.
- **Examples of Nominal Data :**
- Colour of hair (Blonde, red, Brown, Black, etc.)
- Marital status (Single, Widowed, Married)
- Nationality (Indian, German, American)
- Gender (Male, Female, Others)
- Eye Color (Black, Brown, etc.)

Ordinal Data

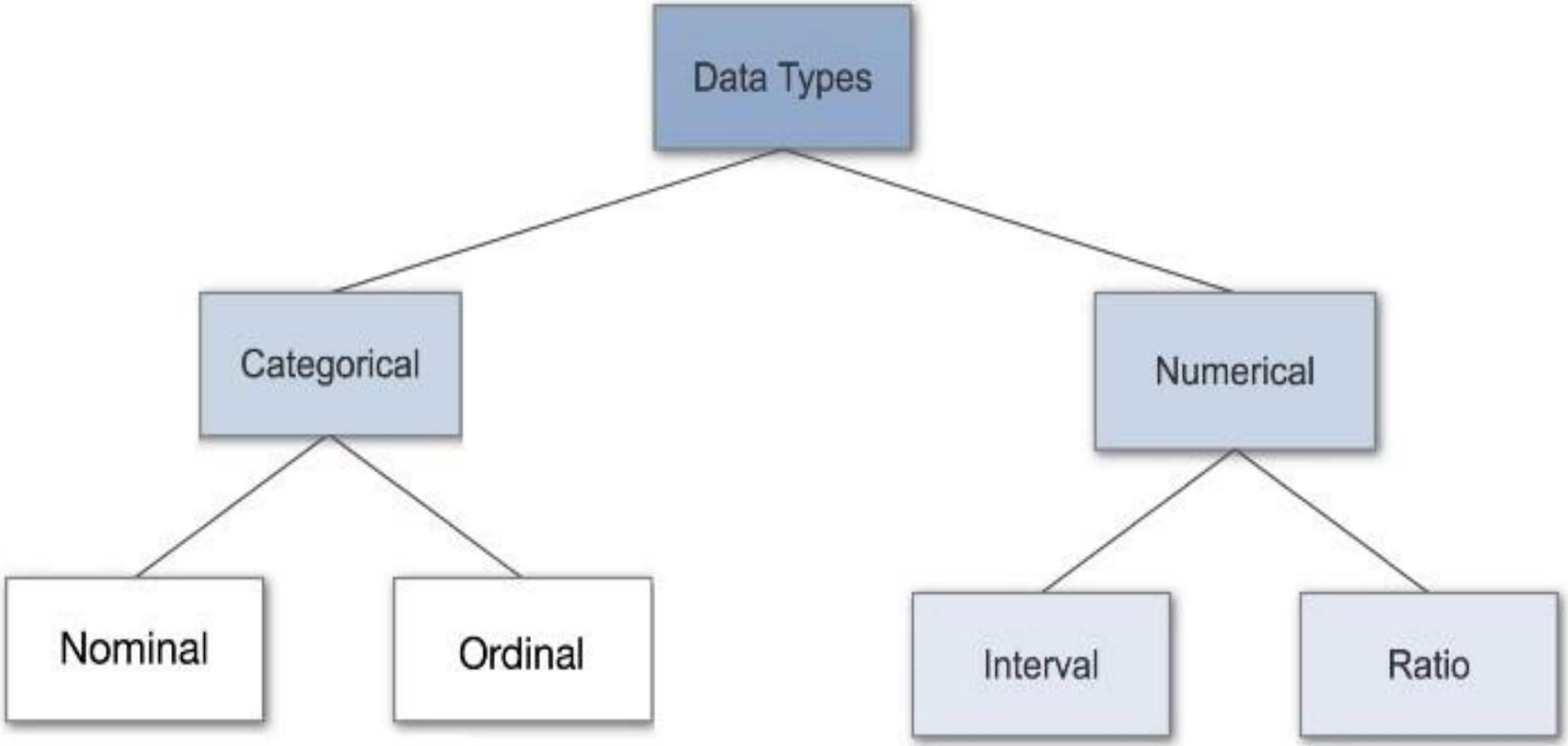
- **Ordinal** - Ordinal data are recorded as the rank order of scores (1st, 2nd, 3rd, etc.).
- **Examples of Ordinal Data :**
- When companies ask for feedback, experience, or satisfaction on a scale of 1 to 10
- Letter grades in the exam (A, B, C, D, etc.)
- Ranking of people in a competition (First, Second, Third, etc.)
- Economic Status (High, Medium, and Low)
- Education Level (Higher, Secondary, Primary)

Discrete Data

- We speak of discrete data if its values are distinct and separate. In other words: We speak of discrete data if the data can only take on certain values. This type of data can't be measured but it can be counted. It basically represents information that can be categorized into a classification. An example is the number of heads in 100 coin flips.
- **Examples of Discrete Data :**
- There are 50 students in one class
- Seven colors in rainbow

Continuous Data

- Continuous Data represents measurements and therefore their values can’t be counted but they can be measured.
- **Examples of Continuous Data :**
- Height of a person
- Speed of a vehicle
- “Time is taken” to finish the work
- Wi-Fi Frequency
- Market share price


Data Types
Categorical
Numerical
Nominal
Ordinal
Interval
Ratio

Categorical Data

- Categorical data represents characteristics.
- Therefore it can represent things like a person's gender, language, etc.
- Categorical data can also take on numerical values.
- Example: 1 for female and 0 for male. Note that those numbers don't have mathematical meaning.

Interval Data

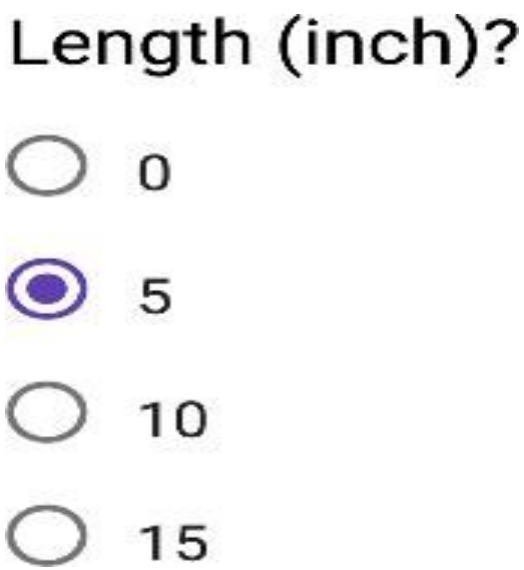
- **Interval** - Interval data are recorded not just about the order of the data points, but also the size of the intervals in between data points. A highly familiar example of interval scale measurement is temperature with the Celsius scale. In this particular scale, the unit of measurement is 1/100 of the temperature difference between the freezing and boiling points of water. The zero point, however, is arbitrary.

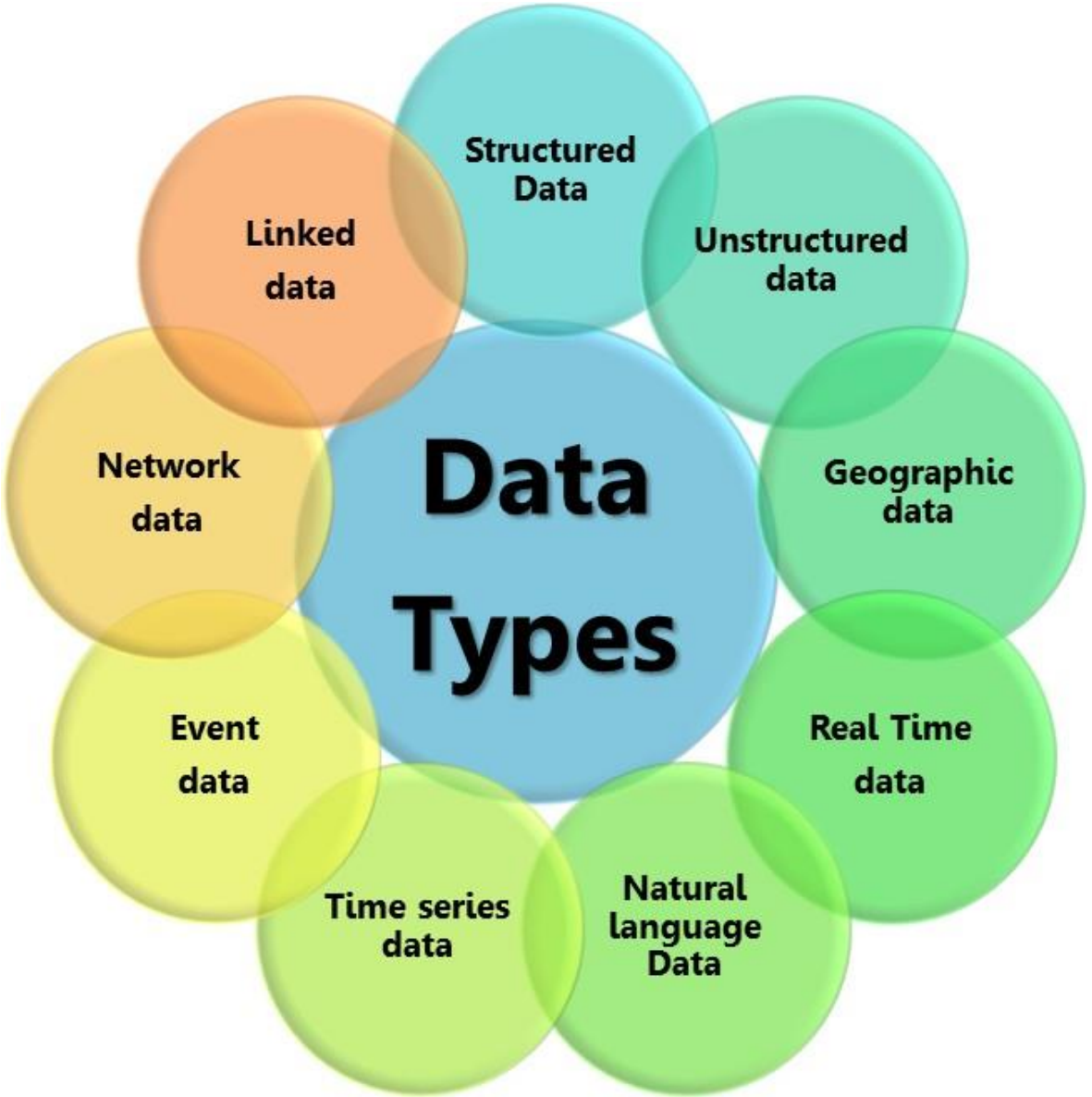
Temperature?

- ○ - 10
- ○ -5
- ○ 0
- ○ + 5
- ○ + 10
- ○ + 15

Ratio Data

- **Ratio** – Ratio data are recorded on an interval scale with a true zero point. Mass, length, time, plane angle, energy, and electric charge are examples of physical measures that are ratio scales. Informally, the distinguishing feature of a ratio scale is the possession of a zero value.




Data Types
Structured Data
Unstructured data
Geographic data
Real Time data
Natural language Data
Time series data
Event data
Network data
Linked data

Info-graphic representation of terminologies

- **Data visualization** is the process of displaying data (often in large quantities) in a meaningful fashion to provide insights that will support better decisions.
- Visualization is the process of extracting salient features from sets of data and displaying the features in an intuitive and expressive way.
- Making sense of largeq uantities of disparate data is necessary not only for gaining competitive advantage in today's business environment but also for surviving in it.
- Researchers have observed that data visualization improves decision-making, provides
- managers with better analysis capabilities that reduce reliance on IT professionals, and improve collaboration and information sharing.
- Raw dataare important, particularly when one needs to identify accurate values or compare individual numbers.
- However, it is quite difficult to identify trends and patterns, find exceptions, or compare groups of data in tabular form.
- The human brain does a surprisingly good job processing visual information—if presented in an effective way.
- Visualizing data provides a way of communicating data at all levels of a business and can reveal surprising patterns and relationships

- The taxonomy is heavily weighted toward the more abstract information visualization techniques. It is less representative of scientific visualizations, which can be highly specialized by domain and are more difficult to generalize.
- 1D/Linear
- 2D/Planar (incl. Geospatial)
- 3D/Volumetric
- Temporal
- nD/Multidimensional
- Tree/Hierarchical
- Network

Geospatial

- Geospatial or spatial data visualizations relate to real-life physical locations, overlaying familiar maps with different data points. These types of data visualizations are commonly used to display sales or acquisitions over time and can be most recognizable for their use in political campaigns or to display market penetration in multinational corporations.
- Examples of geospatial data visualizations include:

☐ Flow map

☐ Density map

☐ Cartogram

☐ Heat map

Temporal

- Data visualizations belong in the temporal category if they satisfy two conditions: that they are linear, and that they are one-dimensional. Temporal visualizations normally feature lines that either stand-alone or overlap with each other, with a start and finish time.
- Examples of temporal data visualization include:

☐ Scatter plots

☐ Polar area diagrams

☐ Time series sequences

☐ Timelines

☐ Line graphs

Multi dimentional

- Just like the name, multidimensional data visualizations have multiple dimensions. This means that there are always 2 or more variables in the mix to create a 3D data visualization. Because of the many concurrent layers and datasets, these types of visualizations tend to be the most vibrant or eye-catching visuals.
- Examples of multidimensional data visualizations include:

☐ Scatter plots

☐ Pie charts

☐ Venn diagrams

☐ Stacked bar graphs

☐ Histograms

Hierarchical

- Data visualizations that belong in the hierarchical category are those that order groups within larger groups. Hierarchical visualizations are best suited if you're looking to display clusters of information, especially if they flow from a single origin point.
- Examples of hierarchical data visualizations include:

☐ Tree diagrams

☐ Ring charts

☐ Sunburst diagrams

Network

- Datasets connect deeply with other datasets. Network data visualizations show how they relate to one another within a network. In other words, demonstrates relationships between datasets without wordy explanations.
- Examples of network data visualizations include:

☐ Matrix charts

☐ Node-link diagrams

☐ Word clouds

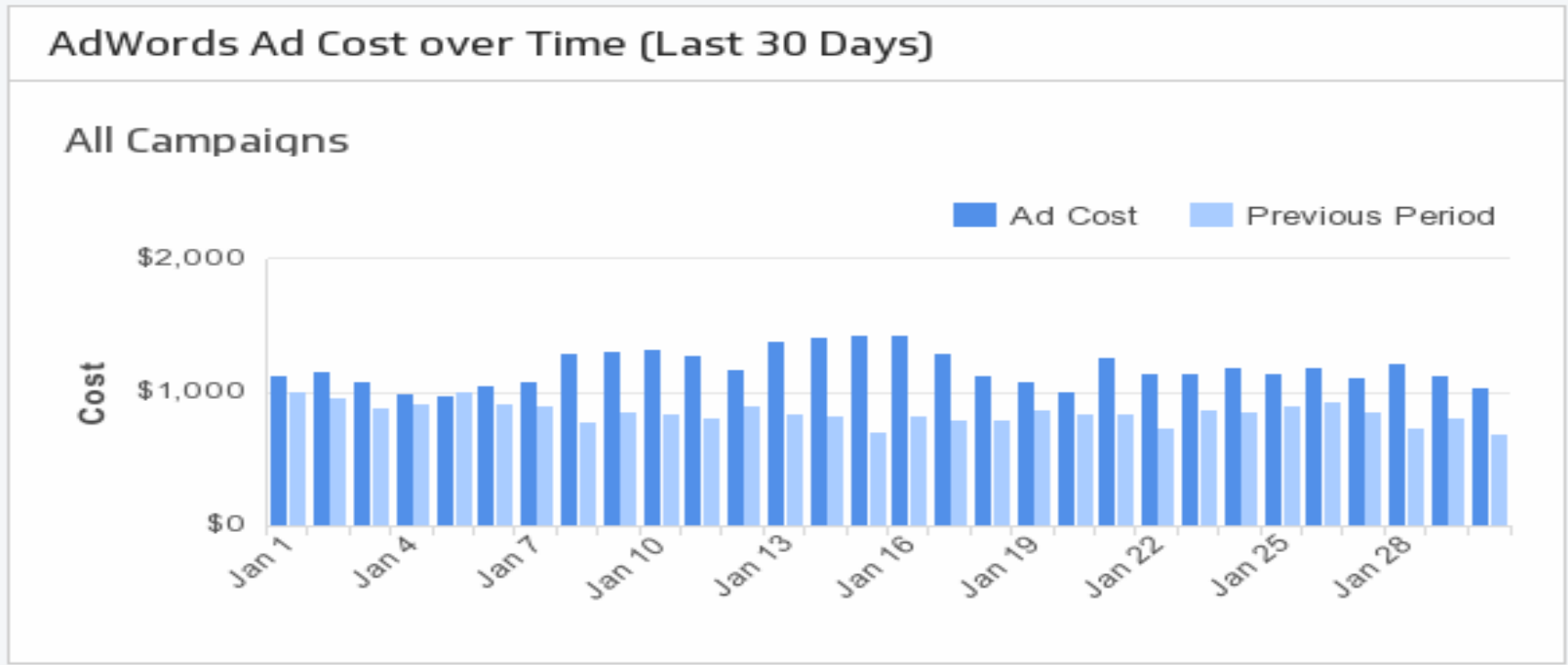
☐ Alluvial diagrams

Visualization Techniques

- Bar chart.
- Line graph.
- Area graph.
- Scatter plot.
- Pie chart.
- Pictograph.
- Column chart.
- Bubble chart.

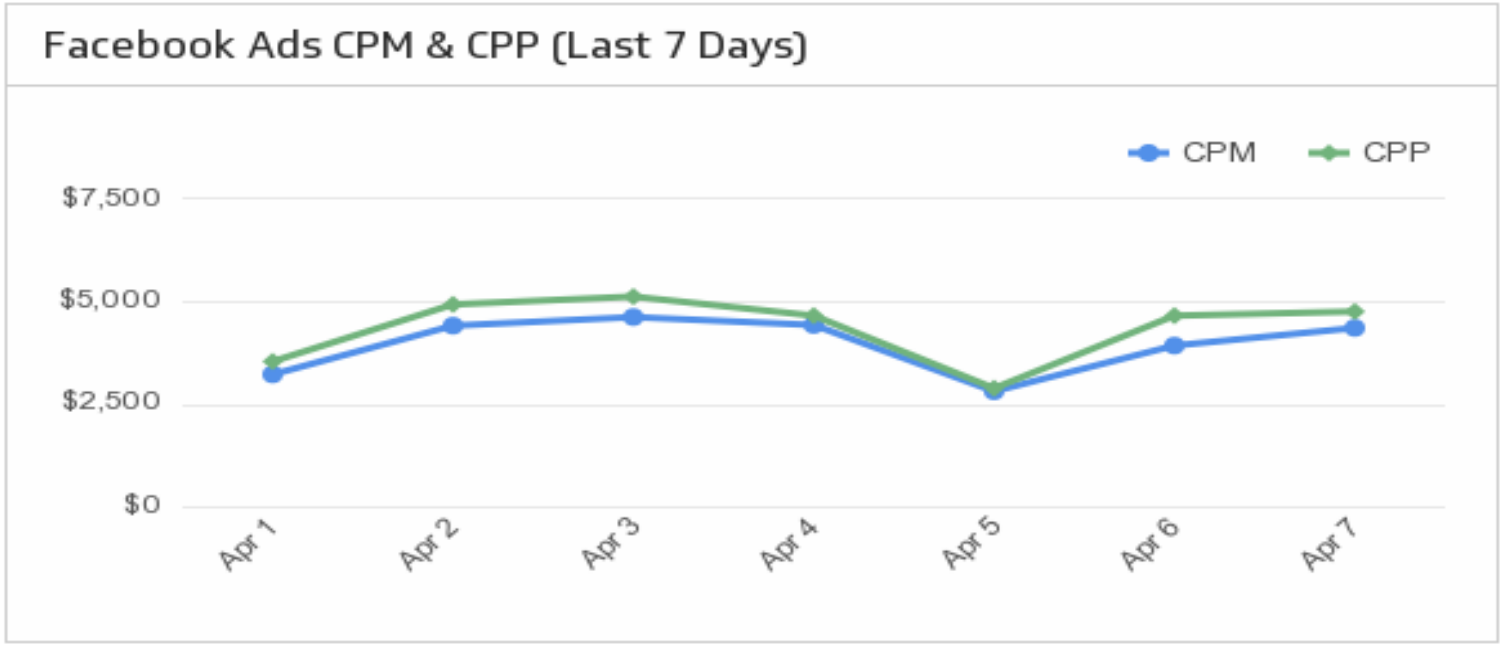
Bar Chart

- Bar charts organize data into rectangular bars that make it a breeze to compare related data sets.



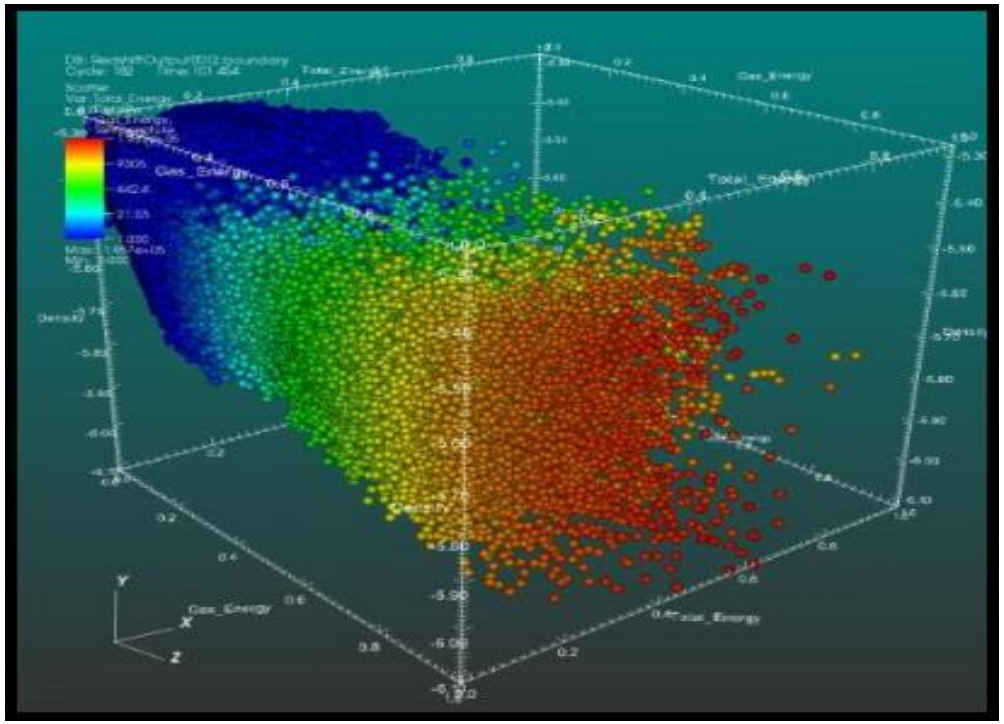
Line Chart

- Like bar charts, line charts help to visualize data in a compact and precise format which makes it easy to rapidly scan information in order to understand trends. Line charts are used to show resulting data relative to a continuous variable - most commonly time or money.



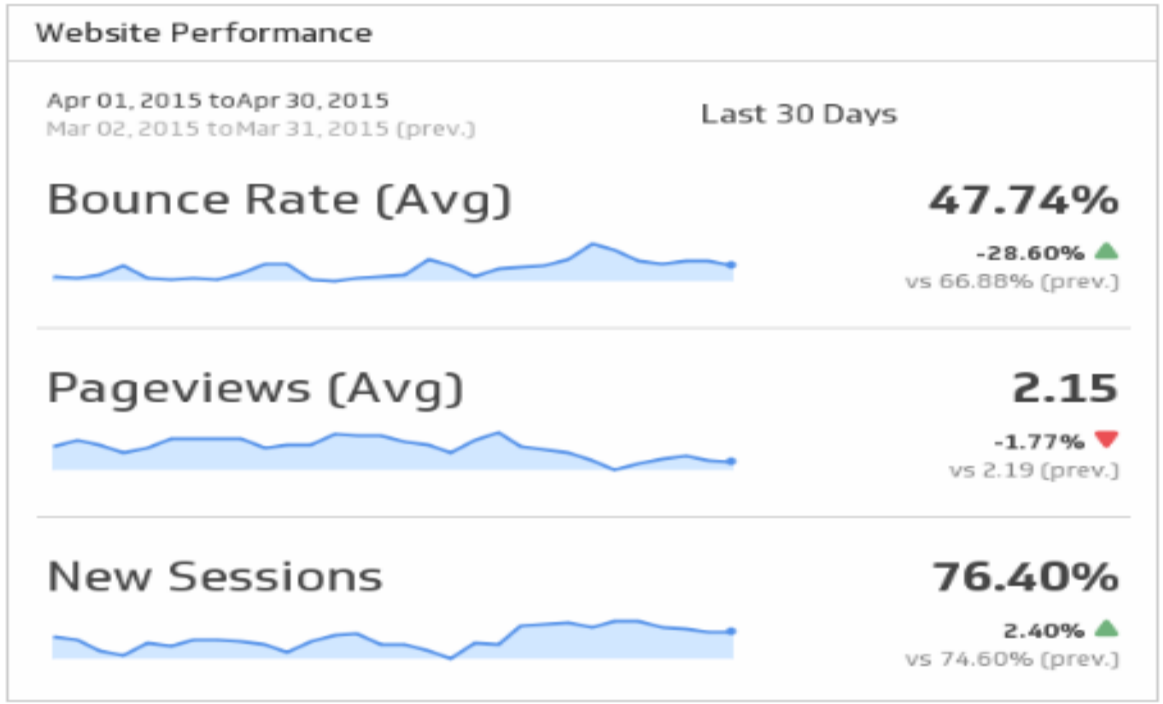
Scatter Plot

- Scatterplots are the right data visualizations to use when there are many different data points, and you want to highlight similarities in the data set. This is useful when looking for outliers or for understanding the distribution of your data. If the data forms a band extending from lower left to upper right, there most likely a positive correlation between the two variables. If the band runs from upper left to lower right, a negative correlation is probable. If it is hard to see a pattern, there is probably no correlation.



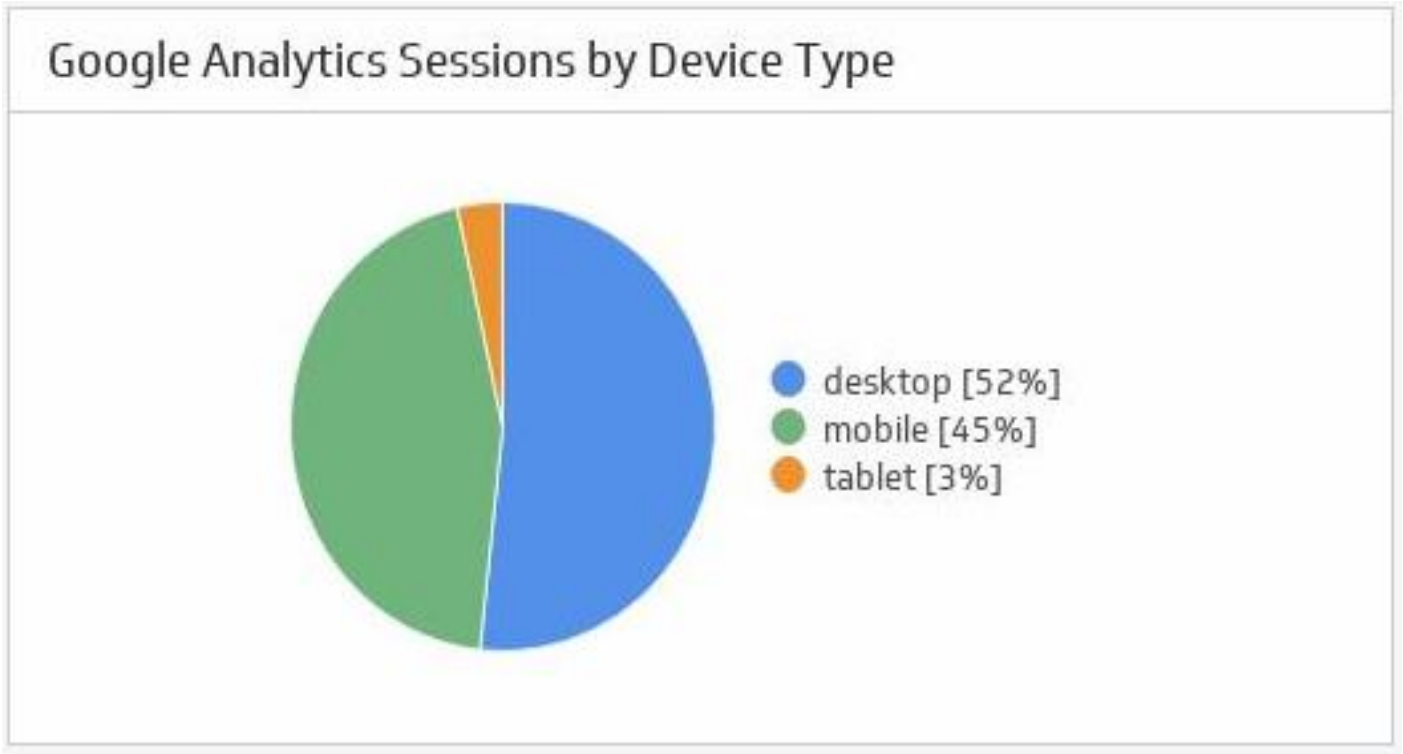
Sparkline

- Sparklines are arguably the best data visualization for showing trends because of how compact they are. They get the job done when it comes to painting a picture for your audience fast. Though, it is important to make sure your audience understands how to read sparklines correctly to optimize their use.



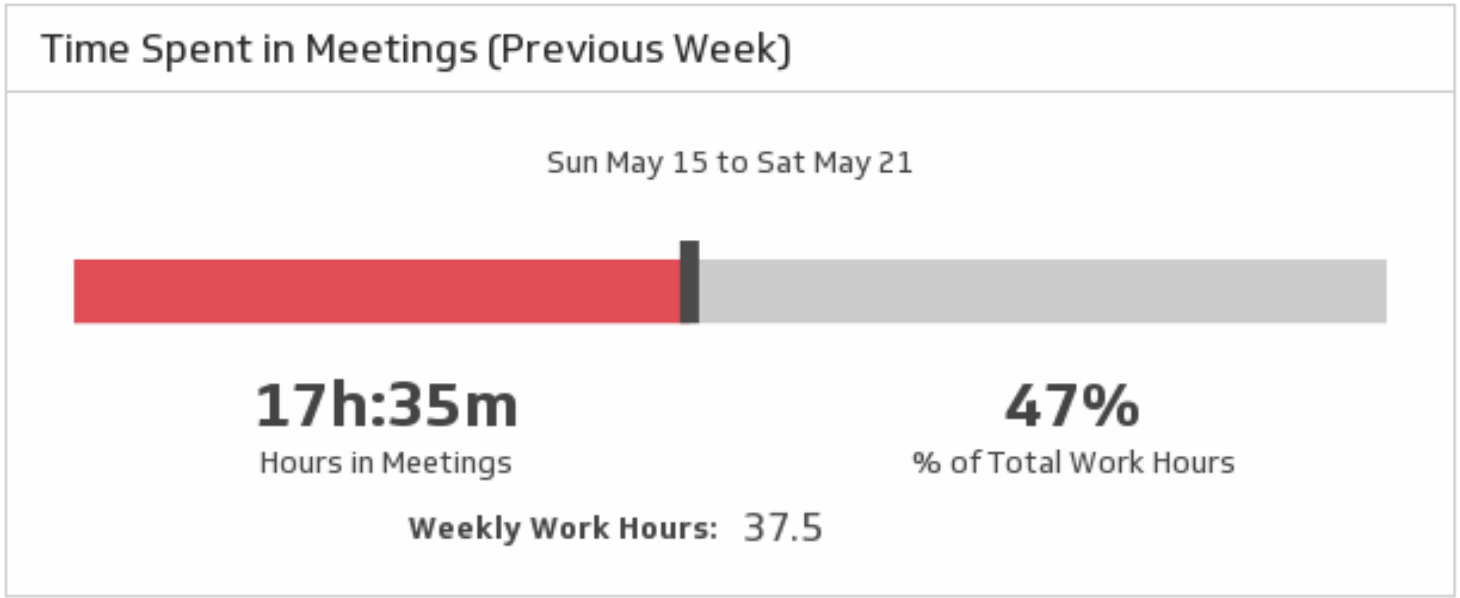
Pie Chart

- Pie charts are an interesting graph visualization. At a high-level, they're easy to read and understand because the parts-of-a-whole relationship is made very obvious. But top data visual experts agree that one of their disadvantages is that the percentage of each section isn't obvious without adding numerical values to each slice of the pie.



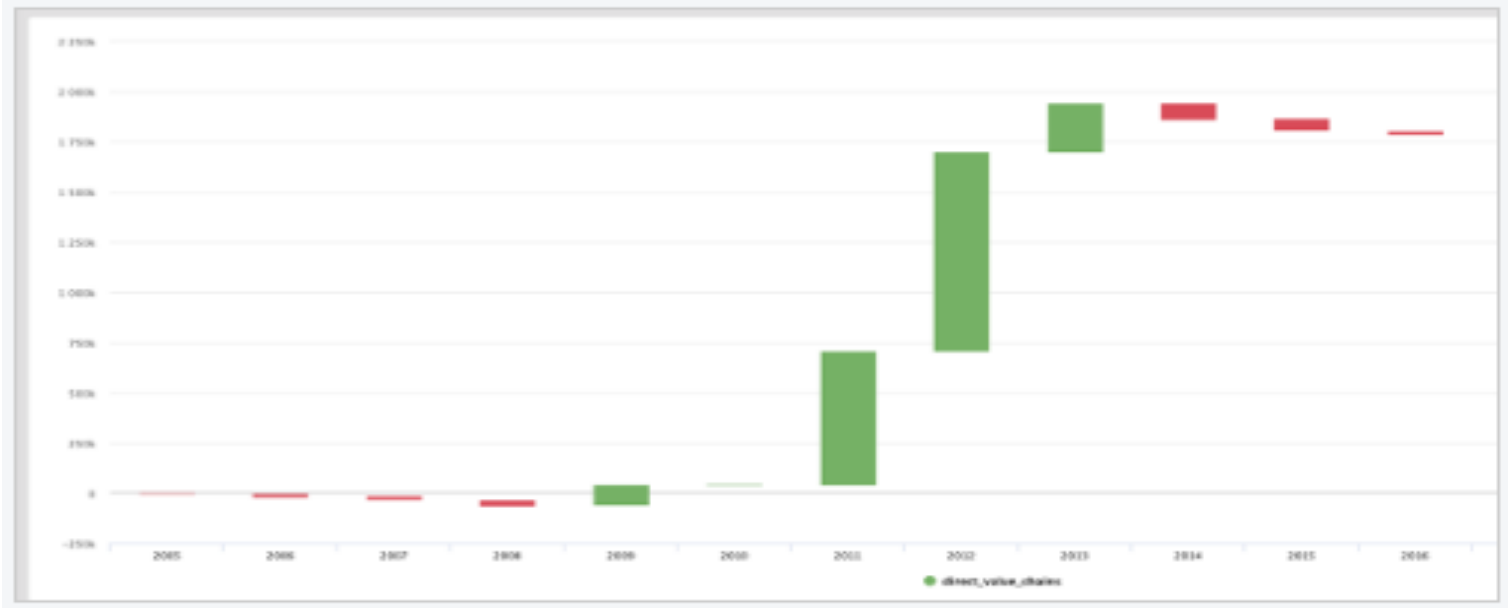
Gauge

- Gauges typically only compare two values on a scale: they compare a current value and a target value, which often indicates whether your progress is either good or bad.



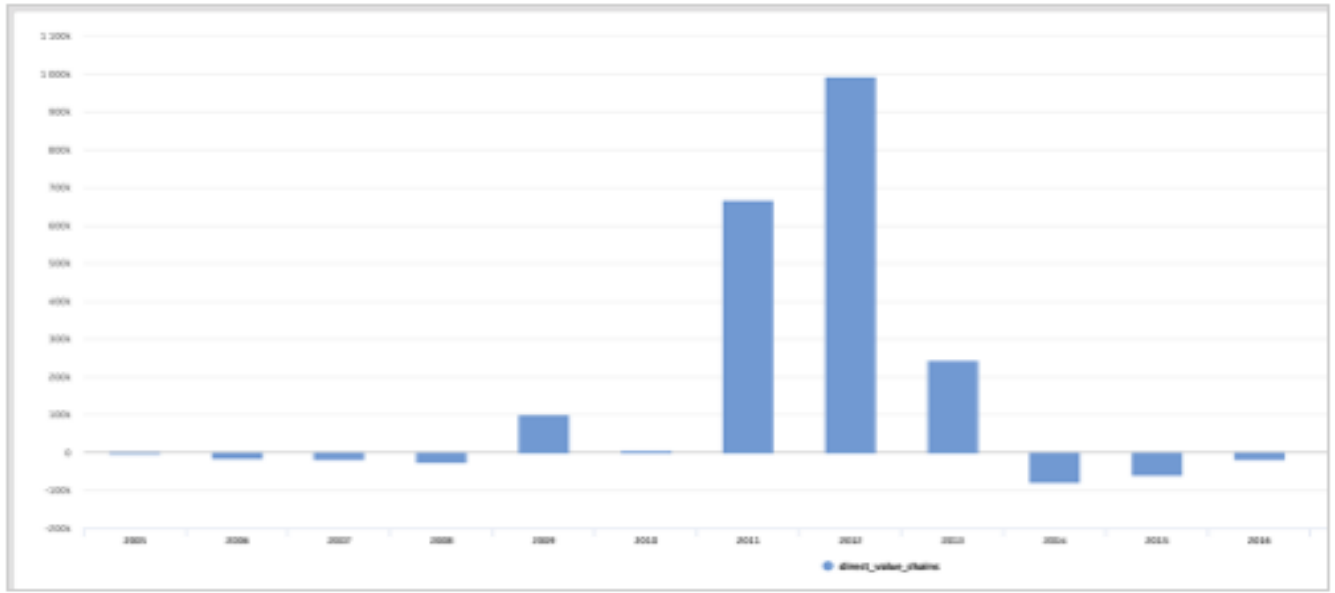
Waterfall Chart

- A waterfall chart is an information visualization that should be used to show how an initial value is affected by intermediate values and resulted in a final value. The values can be either negative or positive.



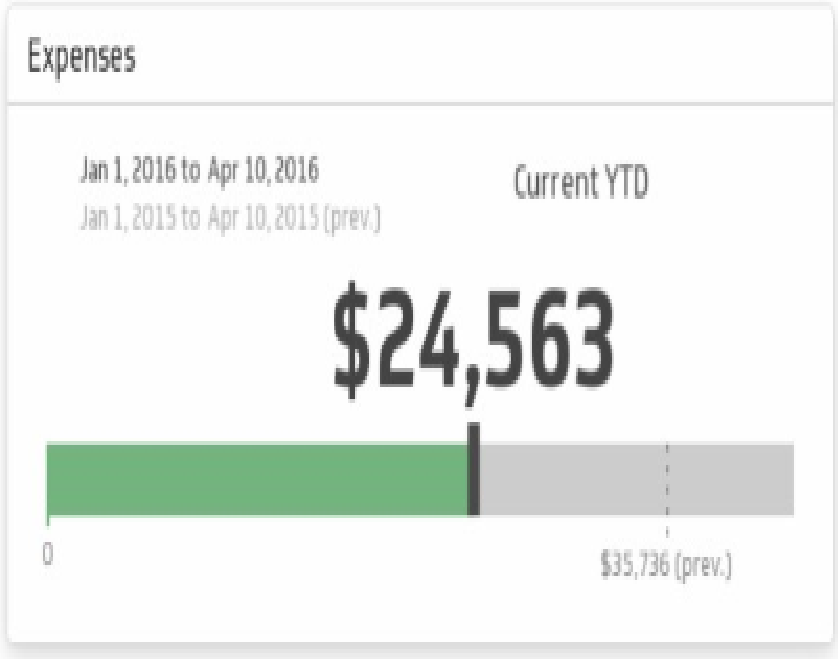
Histogram

- A histogram is a data visualization that shows the distribution of data over a continuous interval or certain time period. It's basically a combination of a vertical bar chart and a line chart. The continuous variable shown on the X-axis is broken into discrete intervals and the number of data you have in that discrete interval determines the height of the bar.
- Histograms give an estimate as to where values are concentrated, what the extremes are and whether there are any gaps or unusual values throughout your data set.



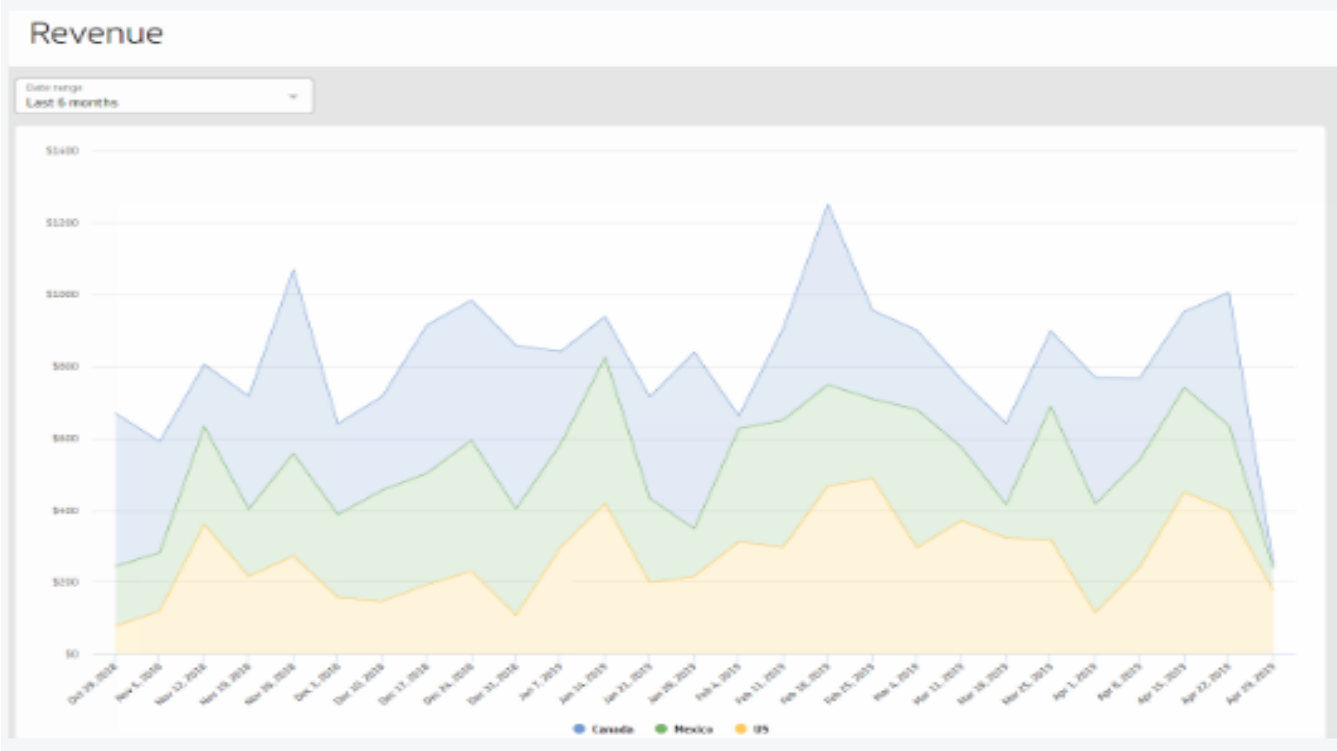
Indicators

- Indicators are useful for an at a glance view of a metric you need to keep track of. An indicator is simply a number showing the current value of whichever performance metric you're tracking. To make it more useful, add a comparison to the previous time period to show whether your metric is tracking up or down.



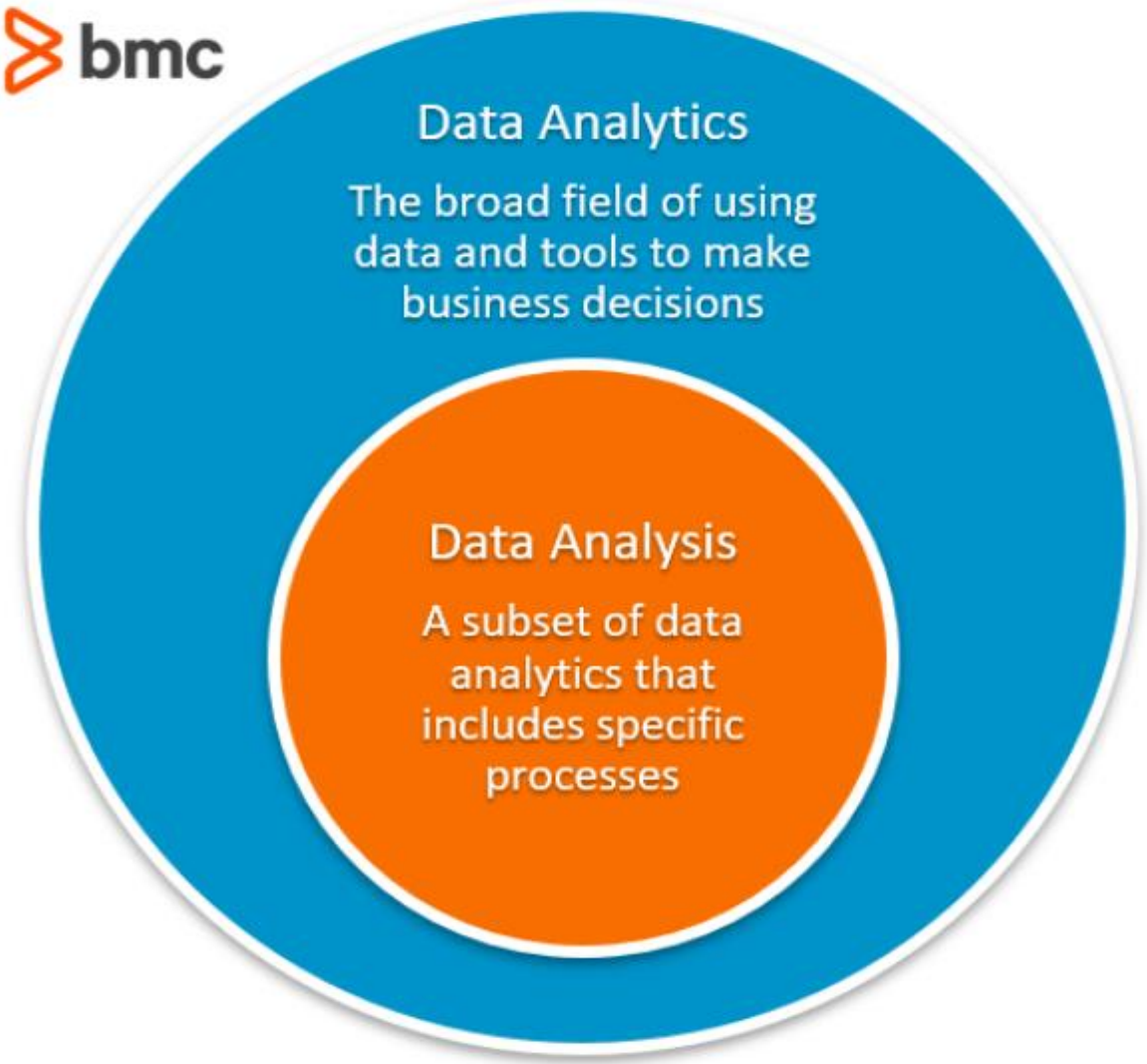
Area Chart

- An area chart is very similar to a line graph but may do a better job at highlighting the relative differences between items. Use an area chart when you want to see how different items stack up or contribute to the whole.

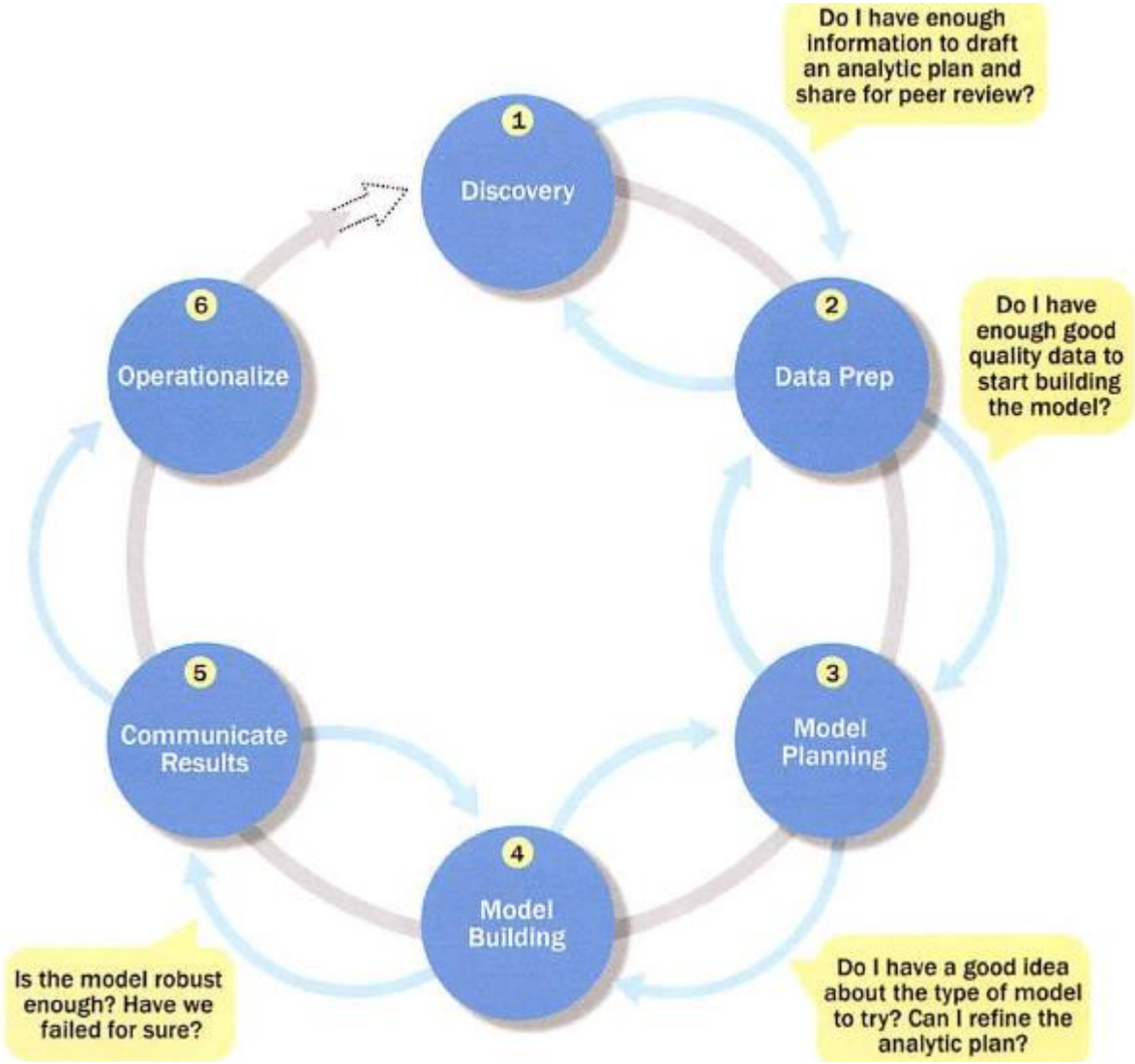


Data Analysis and Data Analytics

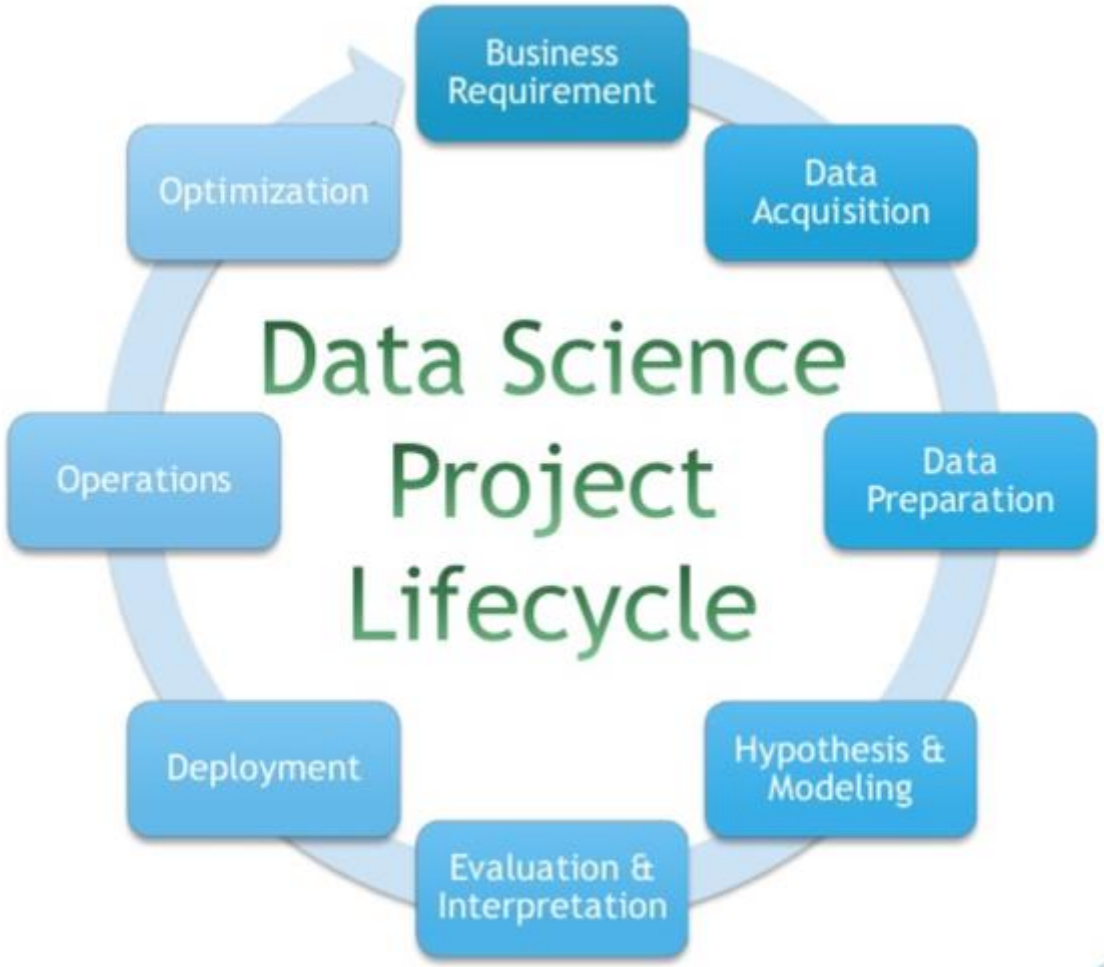
- They both refer to an examination of information—but while *analysis* is the broader and more general concept, *analytics* is a more specific reference to the systematic examination of data.


bmc
Data Analytics
The broad field of using data and tools to make business decisions
Data Analysis
A subset of data analytics that includes specific processes

Data Analytics Life Cycle



Data Science Project Lifecycle



Data Analytics

- Data analytics is a broad term that defines the concept and practice (or, perhaps science and art) of all activities related to data. The primary goal is for data experts, including data scientists, engineers, and analysts, to make it easy for the rest of the business to access and understand these findings.
- Data analytics includes all the steps you take, both human- and machine-enabled, to discover, interpret, visualize, and tell the story of patterns in your data in order to drive business strategy and outcomes.

➢ Data Analytics Process

- Collecting the data
- Categorizing the data
- Managing the data
- Storing the data
- Performing ETL(EXTRACT,TRANSFORM,LOAD)
- Analysing the data
- Sharing the data

Continue…

1. Data Analytics :

- Analytics is a technique of converting raw facts and figures into some particular actions by analyzing those raw data evaluations and perceptions in the context of organizational problem-solving and also with the decision making. Analytics is the discovery and conversation of significant patterns in data. Especially, precious in areas prosperous with recorded information, analytics depends on the simultaneous utility of statistics, computer programming, and operation lookup to qualify performance. Analytics frequently favors data visualization to talk insight. The aim of Data Analytics is to get actionable insights ensuing in smarter selections and higher commercial enterprise outcomes.

2. Data Analysis :

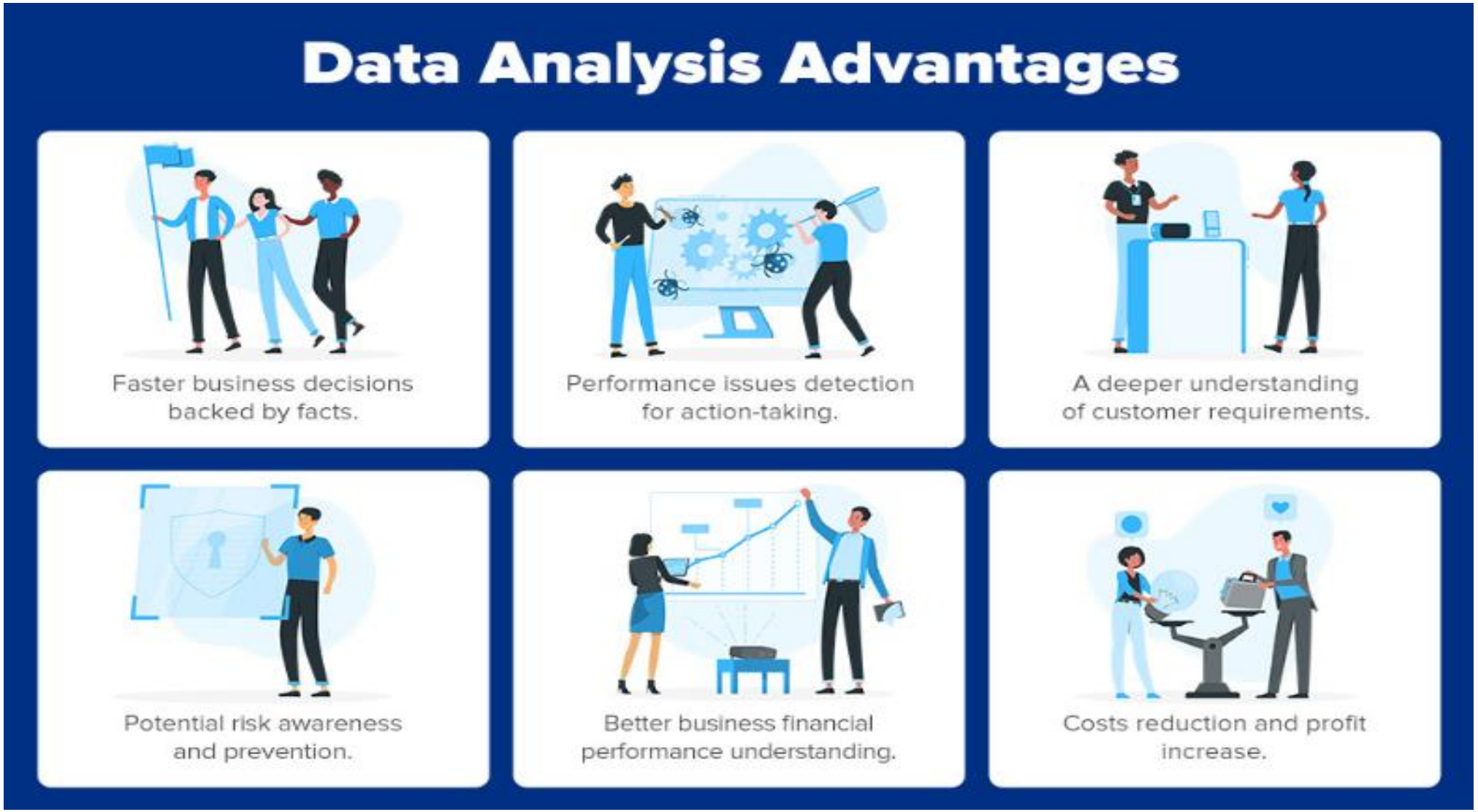
- It is the technique of observing, transforming, cleaning, and modeling raw facts and figures with the purpose of developing beneficial information and acquiring profitable conclusions.

Data Analysis

- Consider data analysis one slice of the data analytics pie. Data analysis consists of cleaning, transforming, modeling, and questioning data to find useful information. (It's generally agreed that other slices are other activities, from collection to storage to visualization.)

➢ **Types of Data Analysis:**

- **Text analysis.** This is also referred to as Data Mining. This method discovers a pattern in large form data sets using databases or other data mining tools.
- **Statistical analysis.** This analysis answers "What happened?" by utilizing past data in dashboard form. Statistic analysis involves the collection, analysis, interpretation, presentation, and modeling of data.
- **Diagnostic analysis.** This analysis answers "Why did it happen?" by seeking the cause from the insights discovered during statistical analysis. This type of analysis is beneficial for identifying behavior patterns of data.
- **Predictive analysis.** This analysis suggests what is likely to happen by utilizing previous data. The predictive analysis makes predictions about future outcomes based on the data.
- **Prescriptive analysis.** This type of analysis combines the insights from text, statistical, diagnostic, and predictive analysis to determine the action(s) to take in order to solve a current problem or influence a decision.


Data Analysis Advantages
Faster business decisions backed by facts.
Performance issues detection for action-taking.
A deeper understanding of customer requirements.
Potential risk awareness and prevention.
Better business financial performance understanding.
Costs reduction and profit increase.

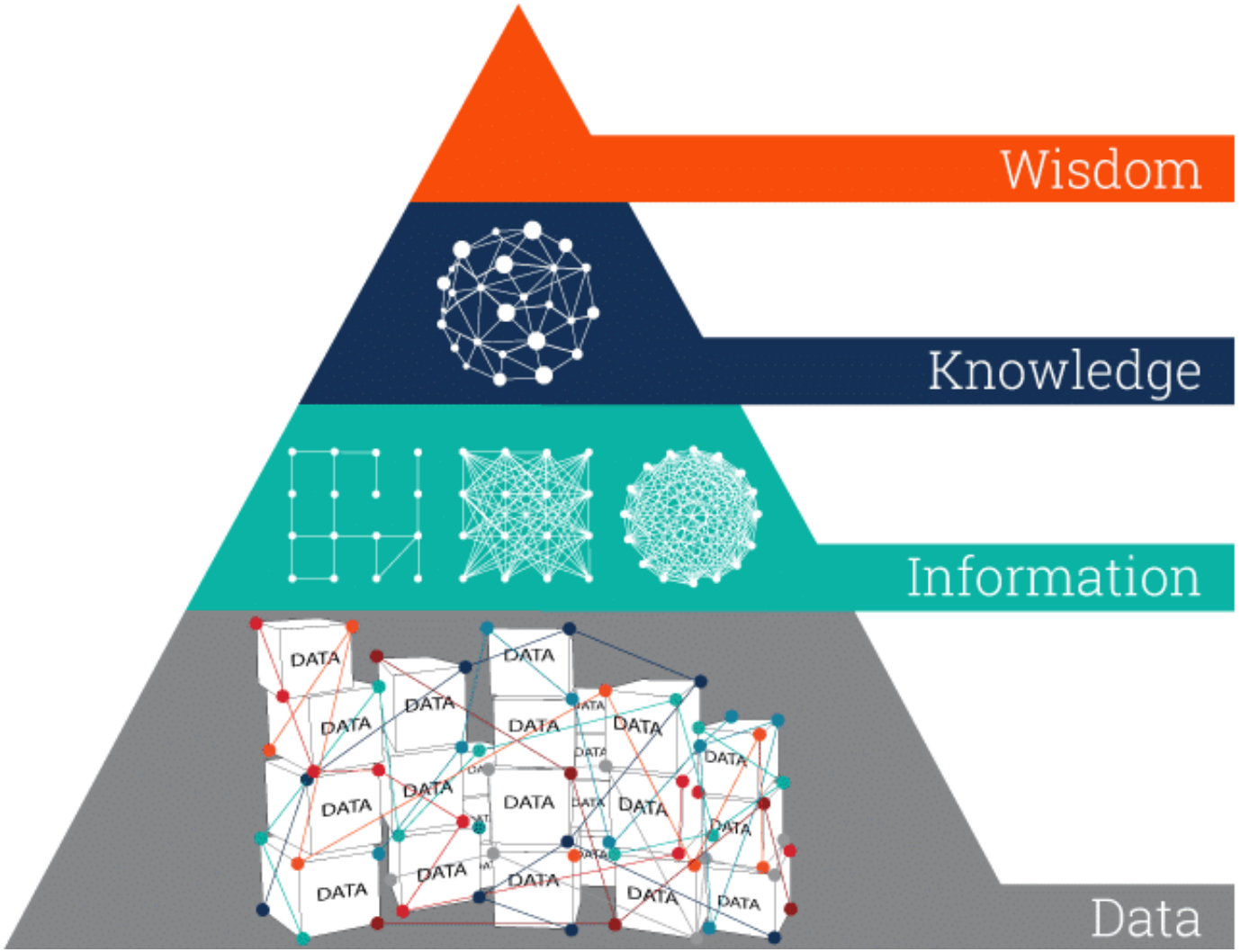
Sr no.	Data Analytics	Data Analysis
1	It is described as a traditional form or generic form of analytics.	It is described as a particularized form of analytics.
2	It includes several stages like the collection of data and then the inspection of business data is done.	To process data, firstly raw data is defined in a meaningful manner, then data cleaning and conversion are done to get meaningful information from raw data.
3	It supports decision-making by analyzing enterprise data.	It analyzes the data by focusing on insights into business data.

4	It uses various tools to process data such as Tableau, Python, Excel, etc.	It uses different tools to analyze data such as Rapid Miner, Open Refine, Node XL, KNIME, etc.
5	Descriptive analysis cannot be performed on this.	A Descriptive analysis can be performed on this.
6	One can find anonymous relations with the help of this.	One cannot find anonymous relations with the help of this.
7	It does not deal with inferential analysis.	It supports inferential analysis.

DIKW Pyramid

- DIKW Pyramid represents the relationships between data, information, knowledge, and wisdom.
- Each building block is a step towards a higher level – first comes data, then is information, next is knowledge, and finally comes wisdom. Each step answers different questions about the initial data and adds value to it.
- The more we enrich our data with meaning and context, the more knowledge and insights we get out of it so we can take better, informed, and data-based decisions.
- For Example

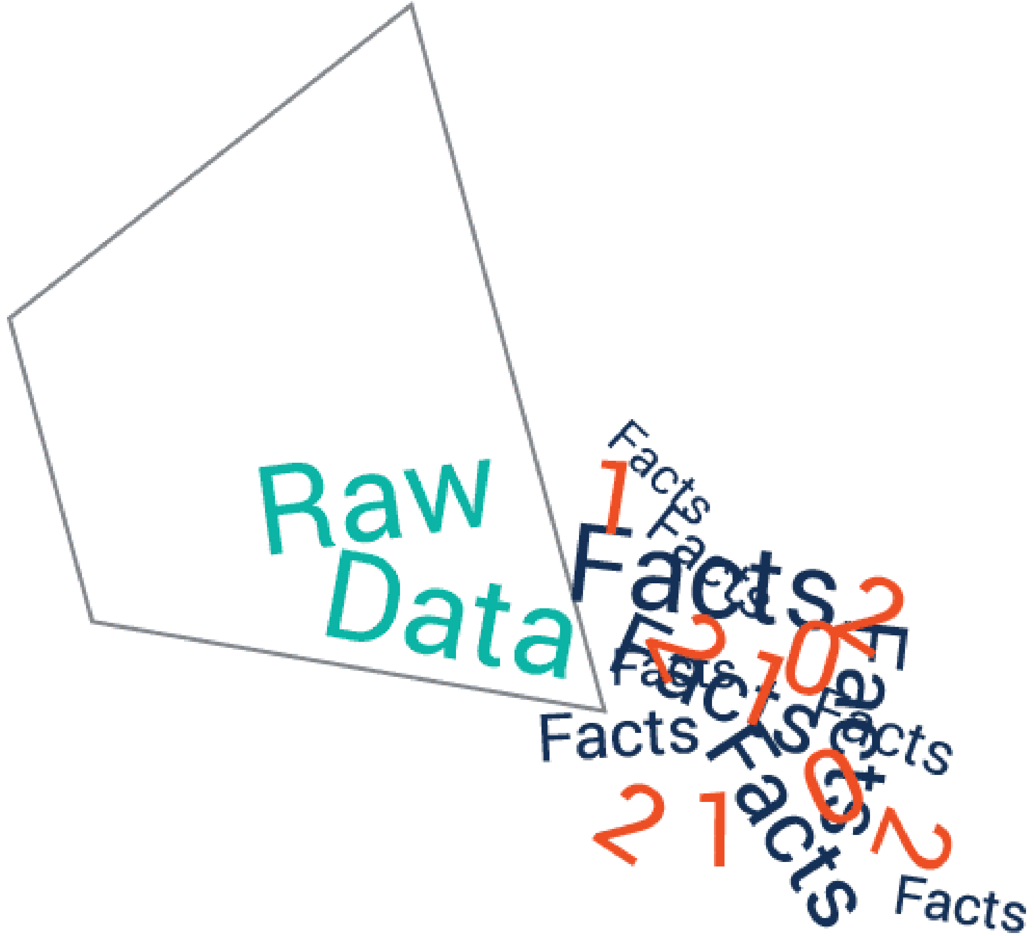
1. Linkedin
2. Facebook
3. Twitter



Each step up the pyramid answers questions about and adds value to the initial data.

- Knowledge Pyramid, Wisdom Hierarchy, and Information Hierarchy are some of the names referring to the popular representation of the relationships between data, information, knowledge, and wisdom in the *Data, Information, Knowledge, Wisdom (DIKW)* Pyramid.
- Like other hierarchy models, the Knowledge Pyramid has rigidly set building blocks – data comes first, information is next, then knowledge follows, and finally wisdom is on the top.
- The more we enrich our data with meaning and context, the more knowledge and insights we get out of it. At the top of the pyramid, we have turned the knowledge and insights into a learning experience that guides our actions.

Data



= a collection of facts in a raw or unorganized form

Base building block - Raw **Data**

- Data is a collection of facts in a raw or unorganized form such as numbers or characters.
- For example, *12012012* is just a sequence of numbers without apparent importance. But if we view it in the context of a date, then it shows 12/01/2012 which is meaningful data.

Information

=

easier to measure,
visualize and analyze
data for a specific purpose

Second building block - Derived Information

- Information is the next building block of the DIKW Pyramid. This is data that has been “cleaned” of errors and further processed in a way that makes it easier to measure, visualize and analyze for a specific purpose.
- Data processing can involve different operations such as combining different sets of data (aggregation), ensuring that the collected data is relevant and accurate (validation), etc. For example, we can organize our data in a way that exposes relationships between various seemingly disparate and disconnected data points.
- By asking relevant questions about ‘who’, ‘what’, ‘when’, ‘where’, etc., we can derive valuable information from the data and make it more useful for us.

Knowledge

HOW

pieces of information

pieces of information

pieces of information

connected to other pieces help us understand how to apply information to achieve our goal

Third building block - Relevant Knowledge

- “How” is the information, derived from the collected data, relevant to our goals? “How” are the pieces of this information connected to other pieces to add more meaning and value? And, maybe most importantly, “how” can we apply the information to achieve our goal?
- When we don’t just view information as a description of collected facts, but also understand how to apply it to achieve our goals, we turn it into knowledge. This knowledge is often the edge that enterprises have over their competitors.
- But only when we use the knowledge and insights gained from the information to take proactive decisions, we can say that we have reached the final – ‘wisdom’ .

Wisdom

why
do
something?

what is best?

Wisdom is knowledge applied in action

The top of the DIKW hierarchy - Guiding **Wisdom**

- Wisdom is the top of the DIKW hierarchy and to get there, we must answer questions such as ‘why do something’ and ‘what is best’. In other words, wisdom is knowledge applied in action.
- Enterprises can climb up the mountain of wisdom and gain a competitive advantage by supporting their business decisions with data-driven analytics.

Data Science Applications

- Business analytics
- Business logistics, including supply chain optimization
- Finance
- Health, wellness, & biomedicine
- Bioinformatics
- Natural sciences
- Information economy / Social media and social network analysis
- Digital Advertisements (Targeted Advertising)
- Recommend System

- Image recognition
- Speech recognition
- Gaming
- Price comparison
- Airline route planning
- Fraud & Risk detection
- Delivery Logistics
- Self Driving Car
- Education and electronic teaching

- Smart cities
- Transportation
- Logistics and Delivery
- Web Search or Internet Web Results
- Manufacturing
- Security
- Energy, sustainability and climate

DIGITAL LEARNING CONTENT

Parul® University

www.paruluniversity.ac.in