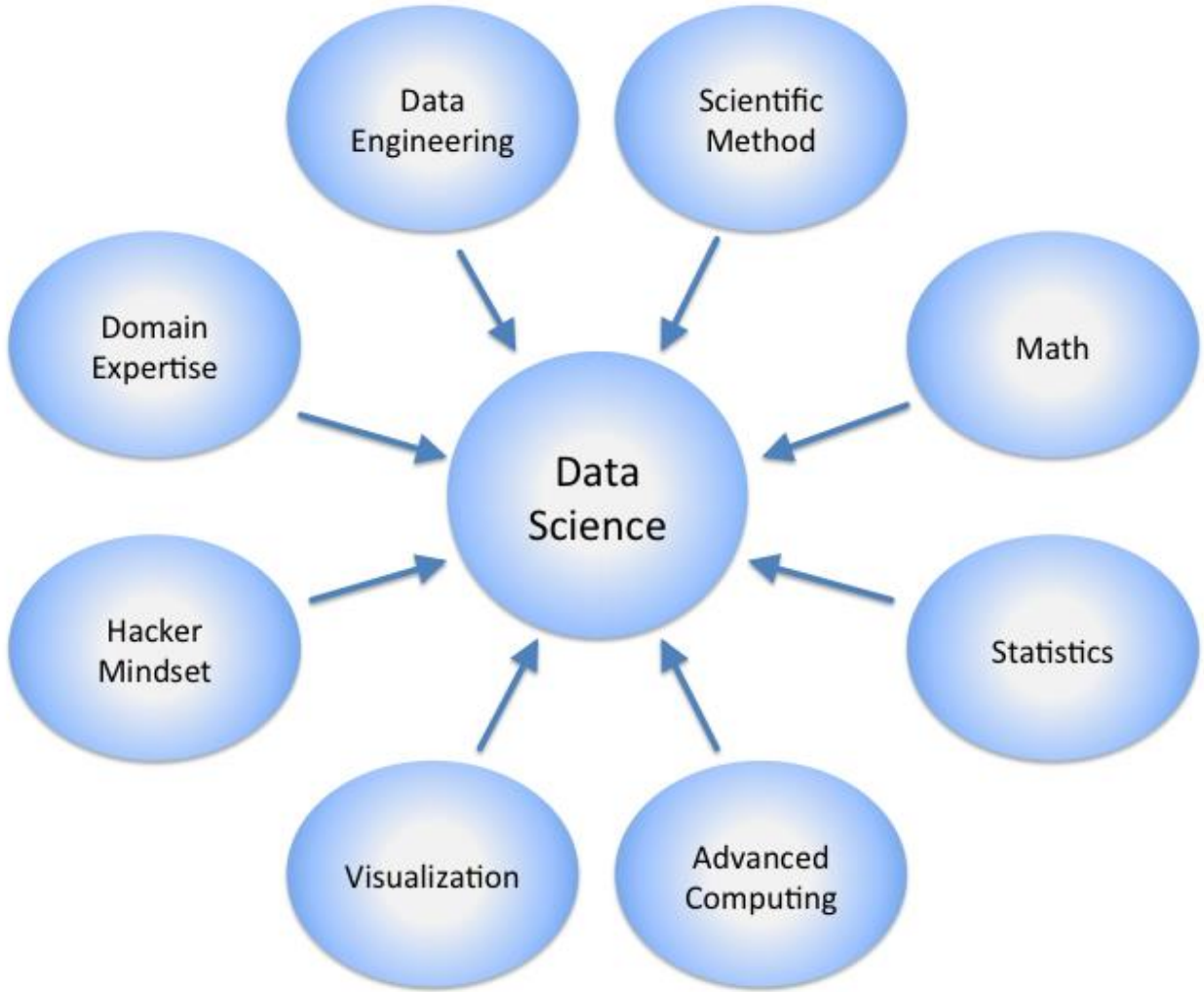
Unit -1

Introduction To Data Science

- An area that manages, manipulates, extracts, and interprets knowledge from tremendous amount of data
- Data science (DS) is a multidisciplinary field of study with goal to address the challenges in big data
- Data science principles apply to all data – big and small

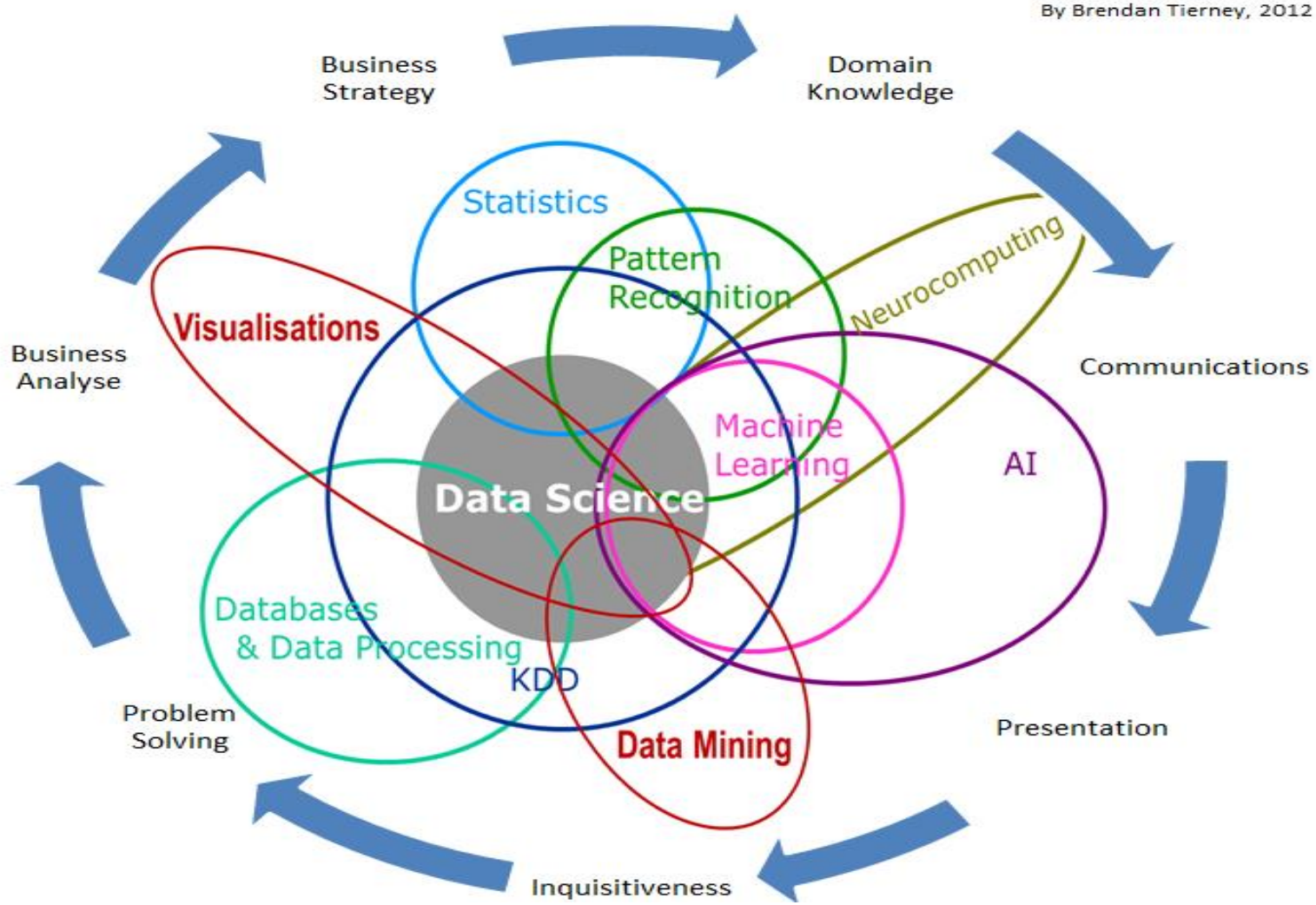
- Theories and techniques from many fields and disciplines are used to investigate and analyze a large amount of data to help decision makers in many industries such as science, engineering, economics, politics, finance, and education.

  - Computer Science
    - Pattern recognition, visualization, data warehousing, High performance computing, Databases, AI
  - Mathematics
    - Mathematical Modeling
  - Statistics
    - Statistical and Stochastic modeling, Probability.

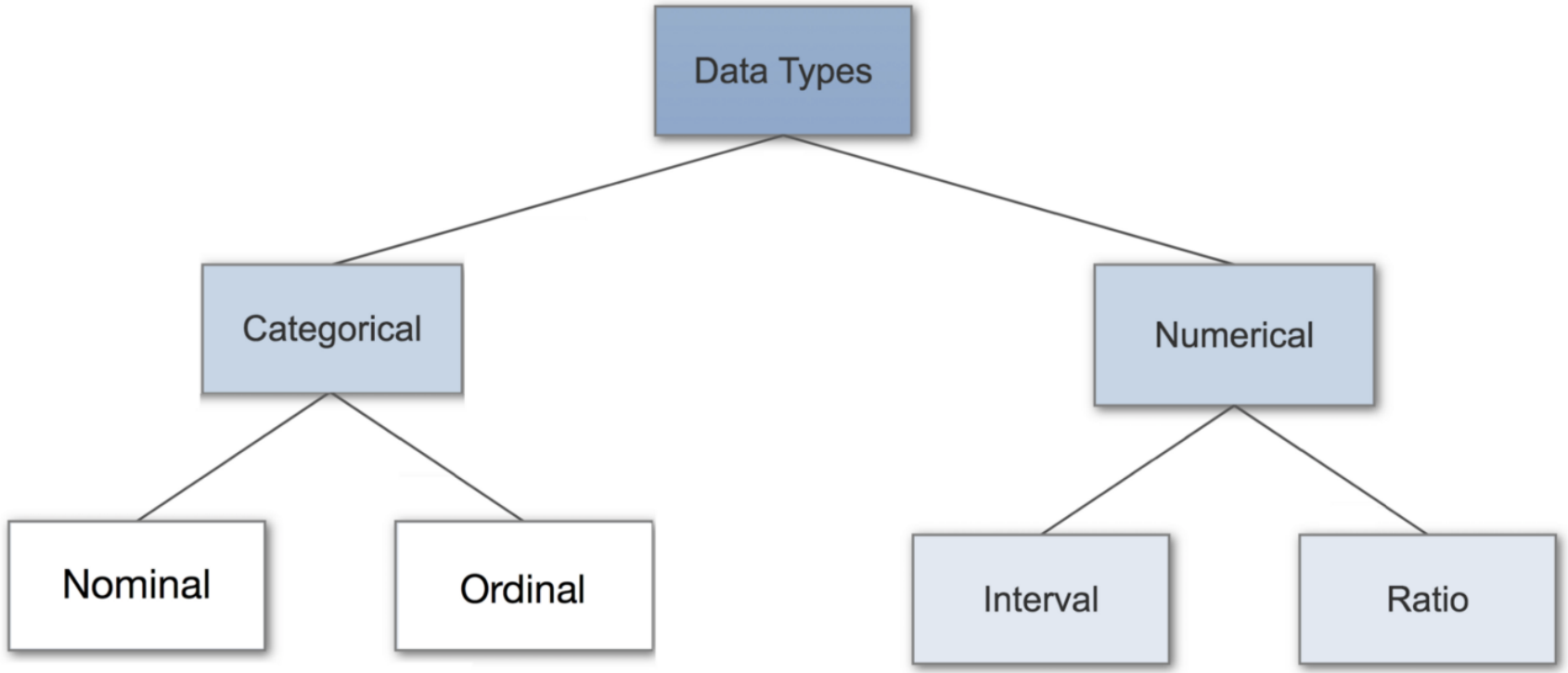

Data Engineering
Scientific Method
Domain Expertise
Math
Data Science
Hacker Mindset
Statistics
Visualization
Advanced Computing

Data Science Is Multidisciplinary

By Brendan Tierney, 2012



Types of Data



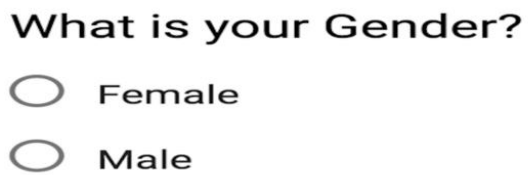
- **Nominal** - Nominal data are recorded as categories. For this reason, nominal data is also known as categorical data. For example, rocks can be generally categorized as igneous, sedimentary and metamorphic.
- **Ordinal** - Ordinal data are recorded as the rank order of scores (1st, 2nd, 3rd, etc.). An example of ordinal data is the result of a horse race, which says only which horses arrived first, second, or third but include no information about race times.
- **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.
- **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. For example, the Kelvin temperature scale has a non-arbitrary zero point of absolute zero.

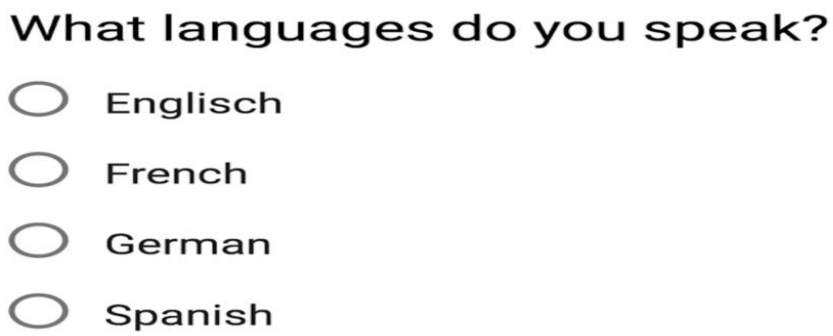
Categorical data

- 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.

Nominal data

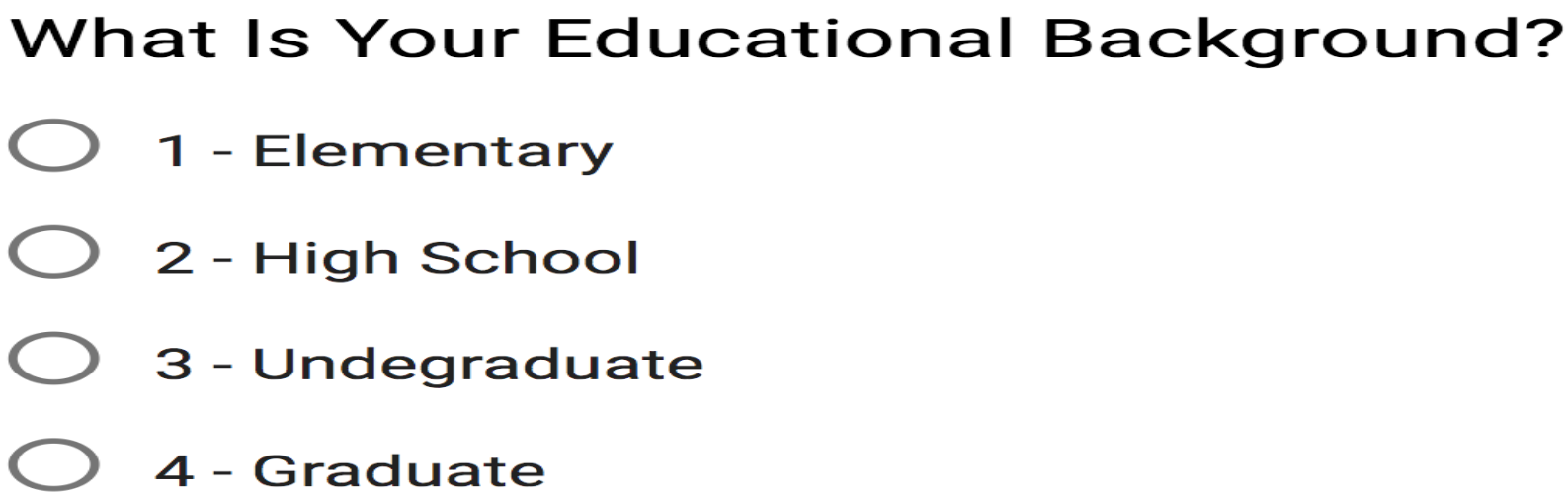
- Nominal values represent discrete units and are used to label variables, that have no quantitative value. Just think of them as “labels”. Note that nominal data that has no order. Therefore if you would change the order of its values, the meaning would not change. You can see two examples of nominal features below:





Ordinal Data

- Ordinal values represent discrete and ordered units. It is therefore nearly the same as nominal data, except that it's ordering matters. You can see an example below:



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.
- You can check by asking the following two questions whether you are dealing with discrete data or not: Can you count it and can it be divided up into smaller and smaller parts?

Continuous Data

- Continuous Data represents measurements and therefore their values can’t be counted but they can be measured. An example would be the height of a person, which you can describe by using intervals on the real number line.

Interval Data

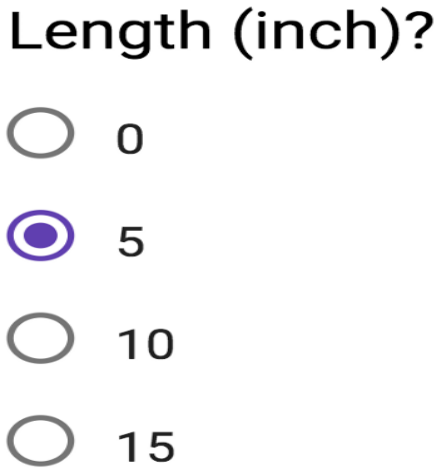
- Interval values represent **ordered units that have the same difference**. Therefore we speak of interval data when we have a variable that contains numeric values that are ordered and where we know the exact differences between the values. An example would be a feature that contains temperature of a given place like you can see below:

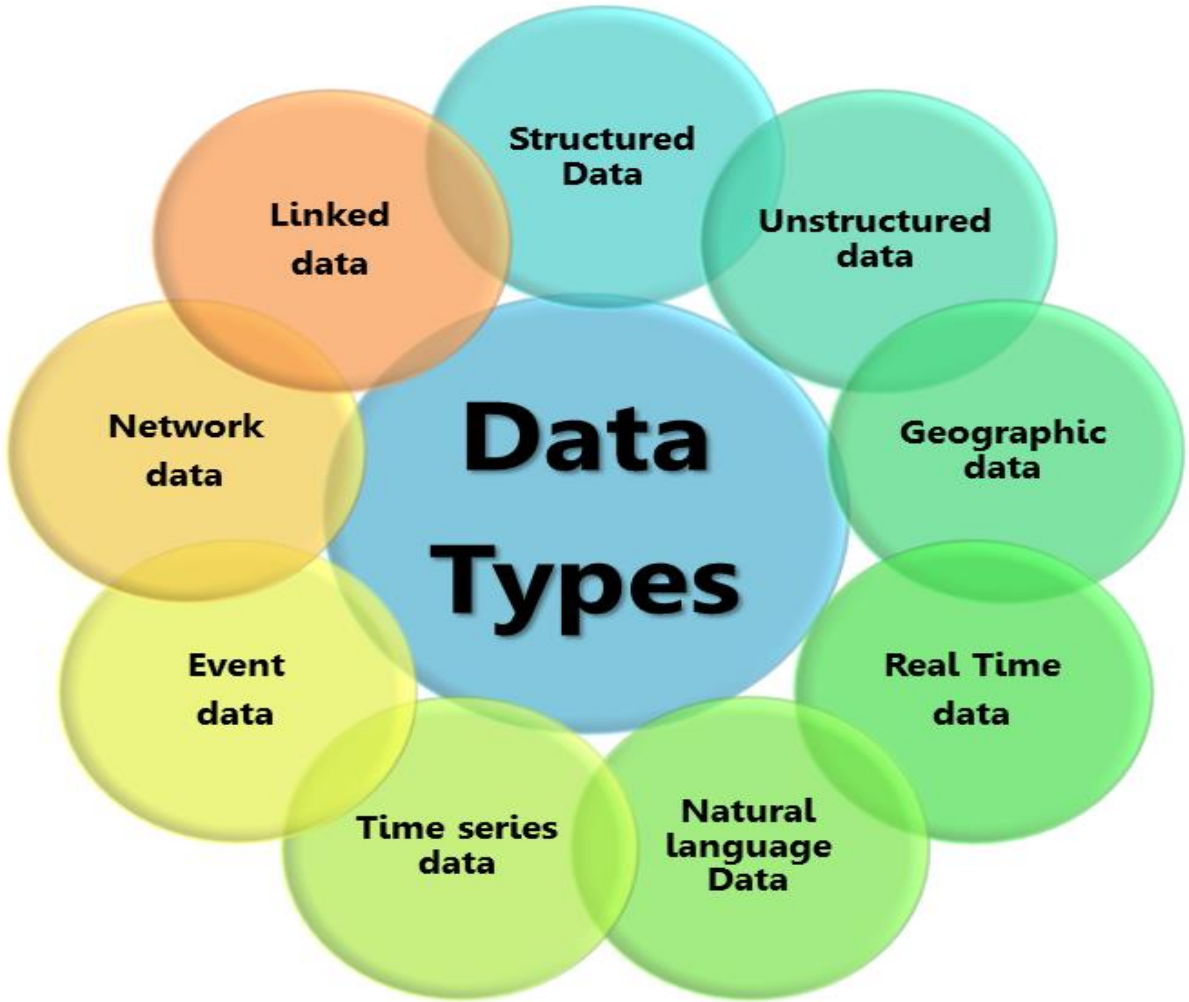
Temperature?

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

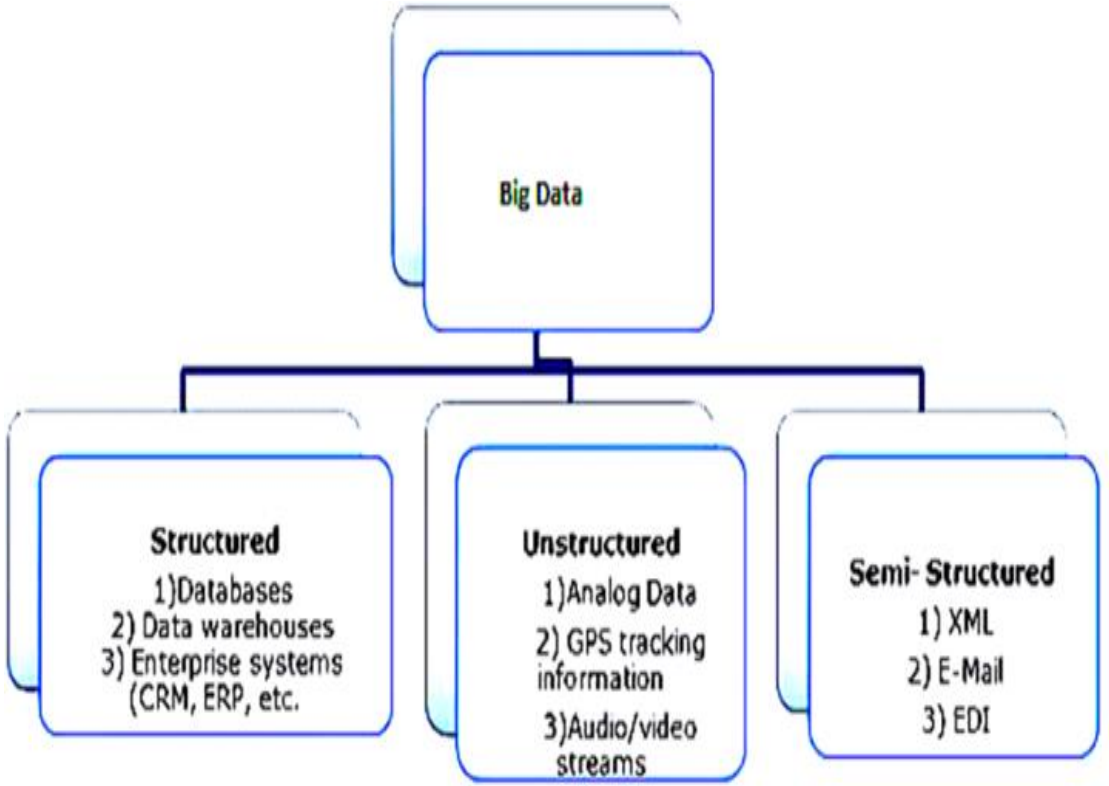
Ratio Data

- Ratio values are also ordered units that have the same difference. Ratio values are the same as interval values, with the difference that they do have an absolute zero. Good examples are height, weight, length etc.




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

Categories of Data



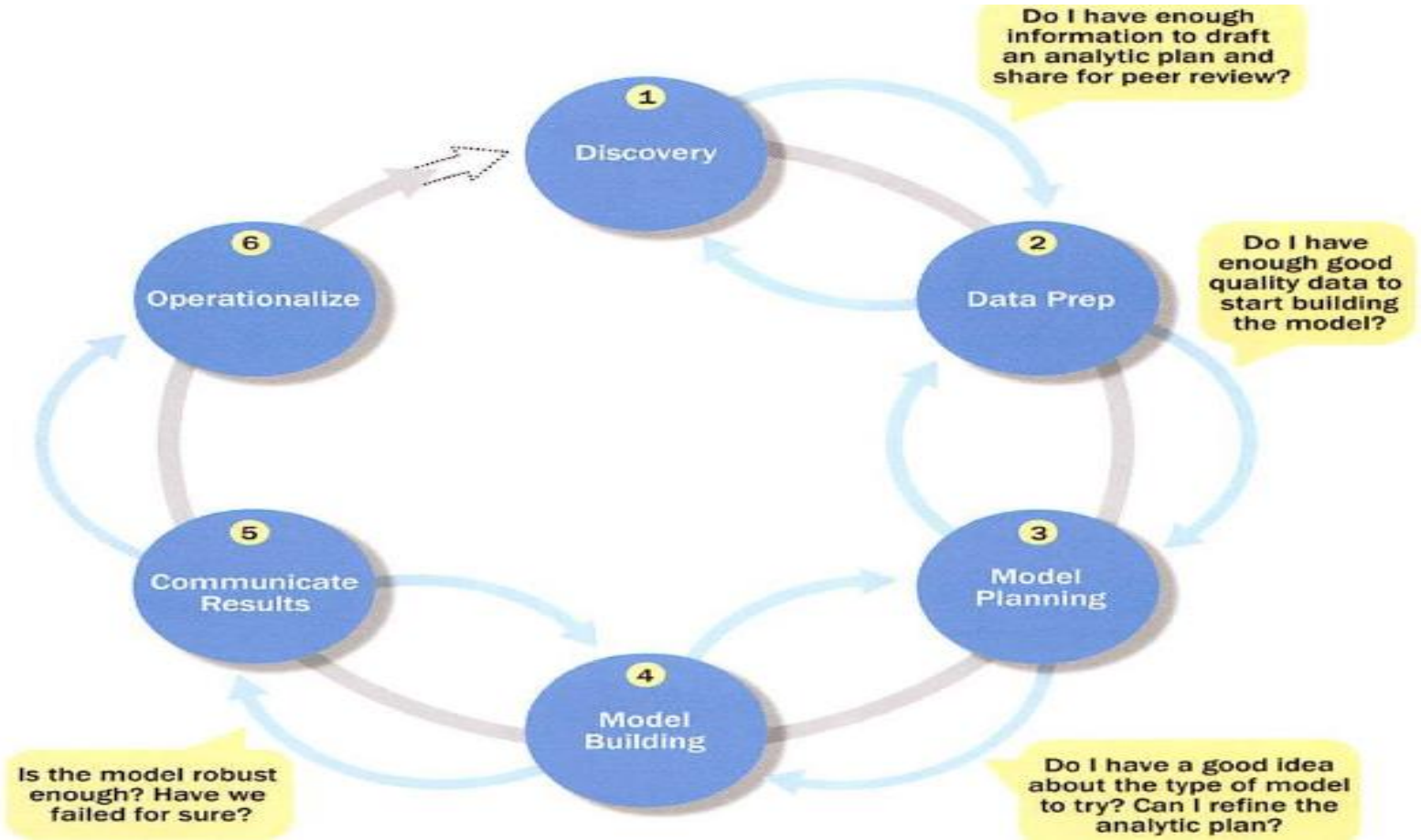

Volume
Size of Data
Velocity
The Speed at which Data is Generated
Variety
Different type of Data
Veracity
Data Accuracy
Value
Useful Data
BigData
Validity
Data quality, Governace, Moster Data Management on Massive
Variability
Dynamic, Evolving Behavior in Data Source
Venue
Distributed Heterogeneous Data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

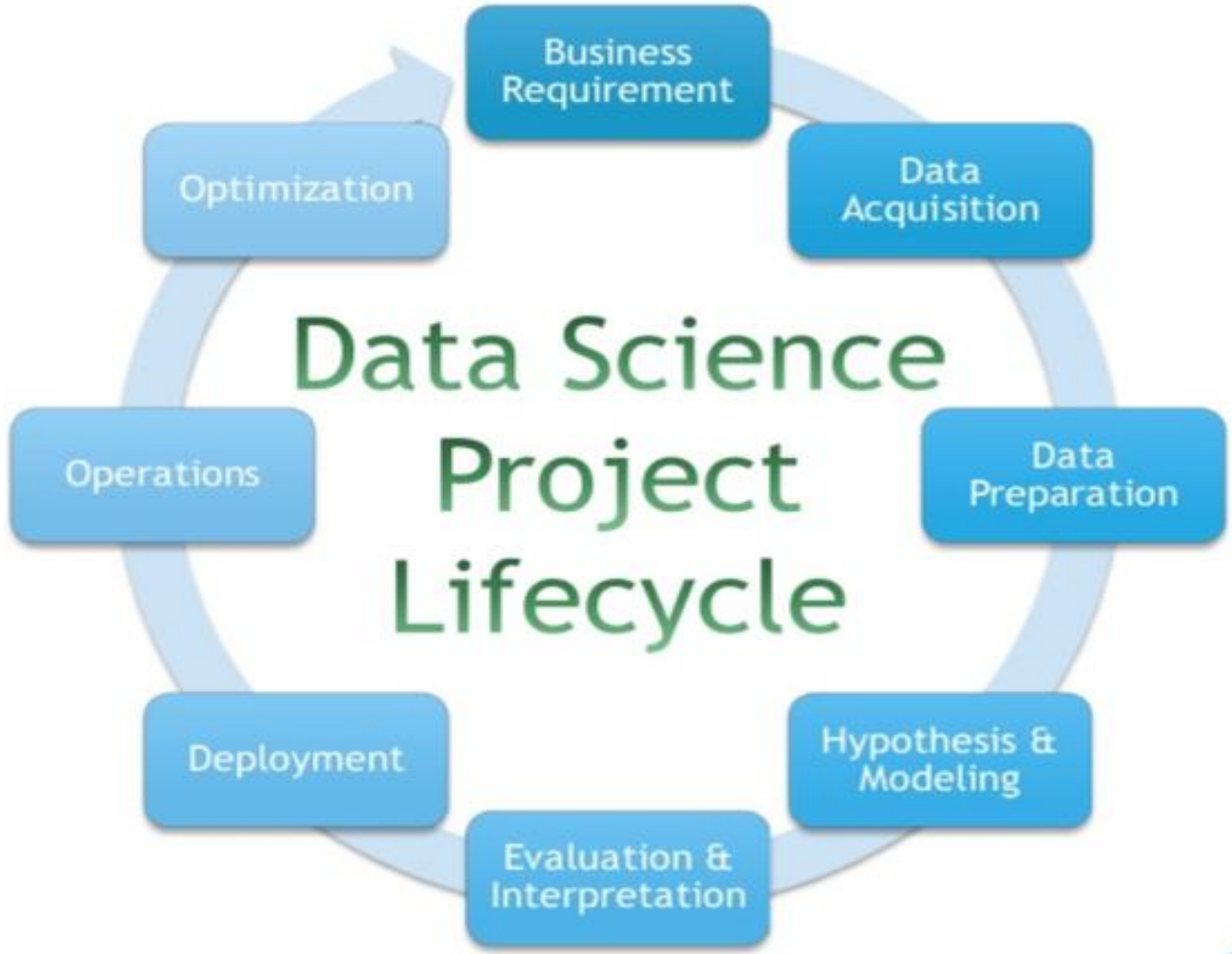
Data science Application

- The application areas of data science are broad and comprehensive. The more than 150 Center for Data Science's core and affiliated faculty work individually and collaboratively to study and partner with industry and government, where appropriate, to address some of the most challenging problems in society. Almost every industry is impacted by data but the application areas are clustered, loosely, as follows:
- 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
- Energy, sustainability and climate
- Smart cities

Data Analytics Life cycle




Business Requirement
Data Acquisition
Data Preparation
Hypothesis & Modeling
Evaluation & Interpretation
Deployment
Operations
Optimization
Data Science Project Lifecycle

Info Graphic Representation (Data Visualization)

- **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 large quantities 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 improves collaboration and information sharing.
- Raw data are 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

Types of Data Visualization

- The taxonomy is heavily weighted toward the more abstract information visualization techniques and 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 dimensional

- 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, demonstrating relationships between datasets without wordy explanations.
- Examples of network data visualizations include:

➢ Matrix charts

➢ Node-link diagrams

➢ Word clouds

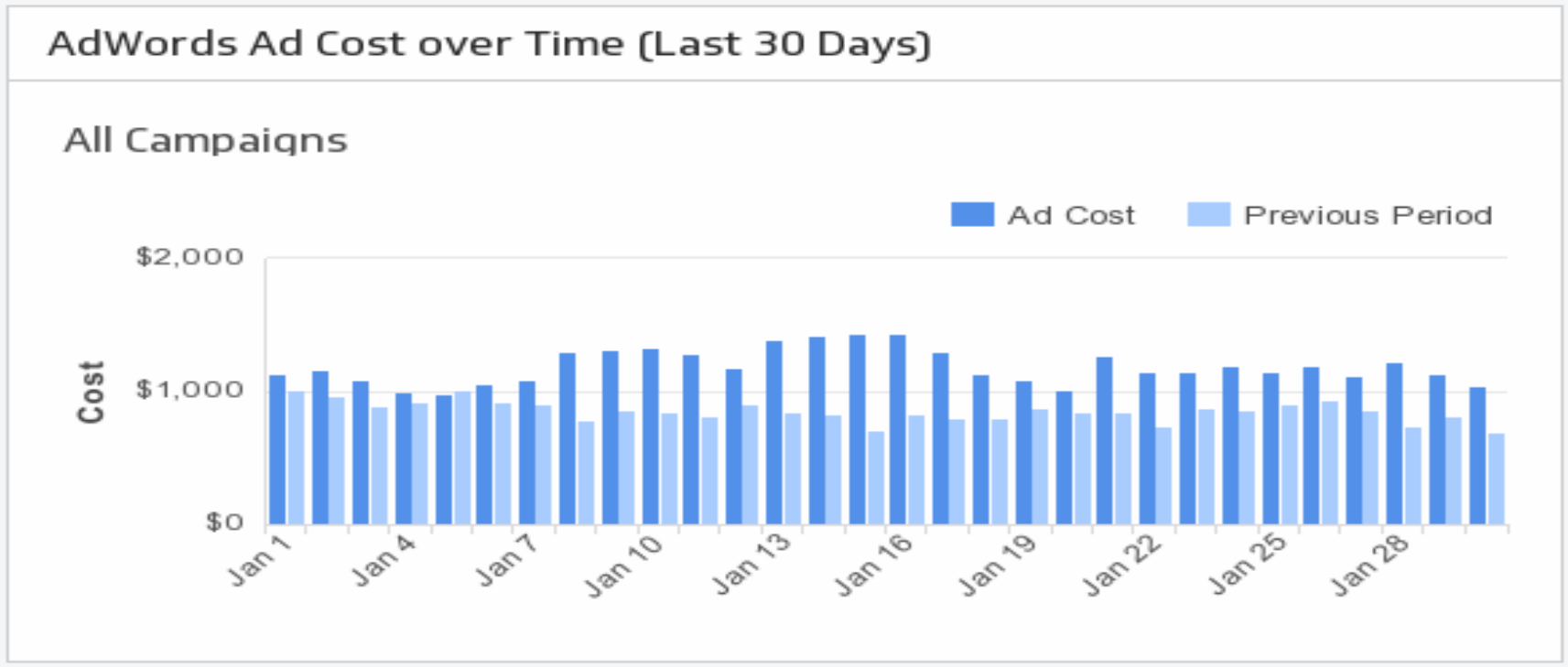
➢ Alluvial diagrams

Common Visualization Techniques

- Bar Chart
- Line Chart
- Scatterplot
- Sparkline
- Pie Chart
- Gauge
- Waterfall Chart
- Funnel Chart
- Heat Map
- Histogram
- Box Plot
- Maps
- Tables
- Indicators
- Area Chart
- Radar or Spider Chart
- Tree Map

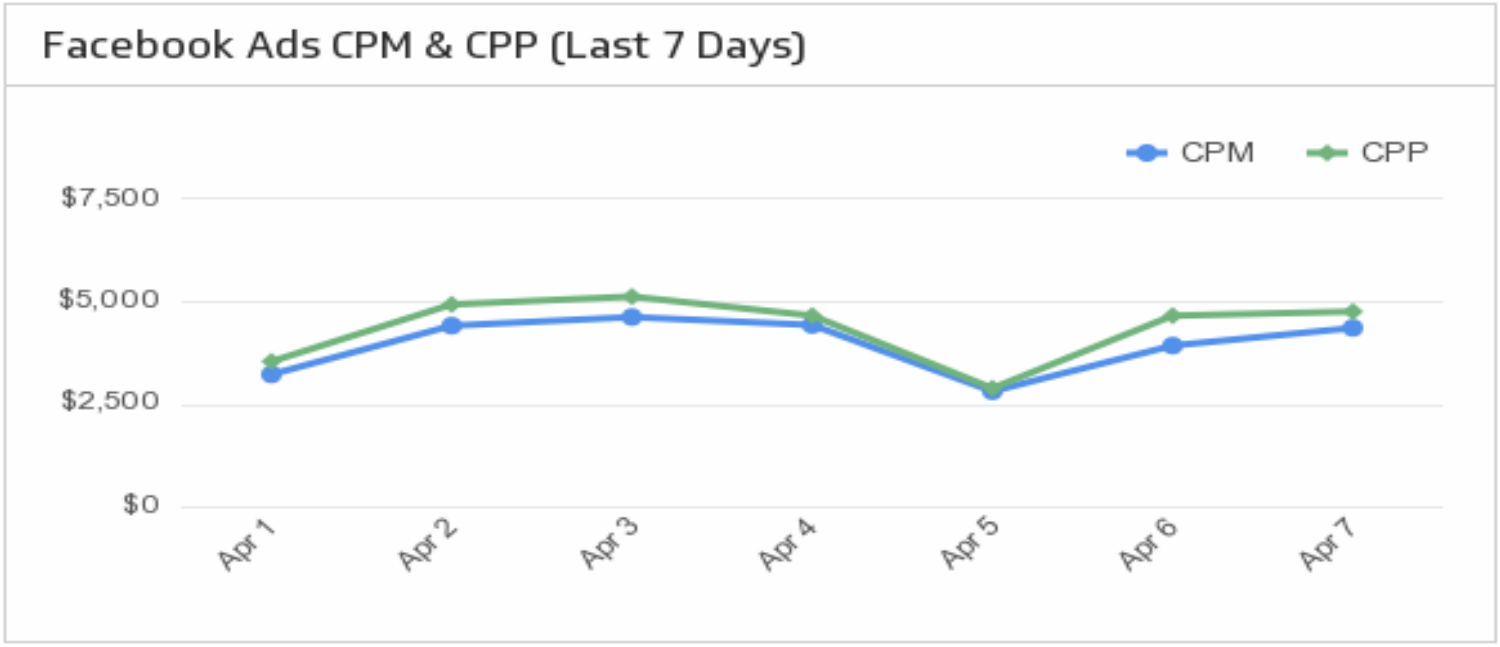
Bar Chart

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



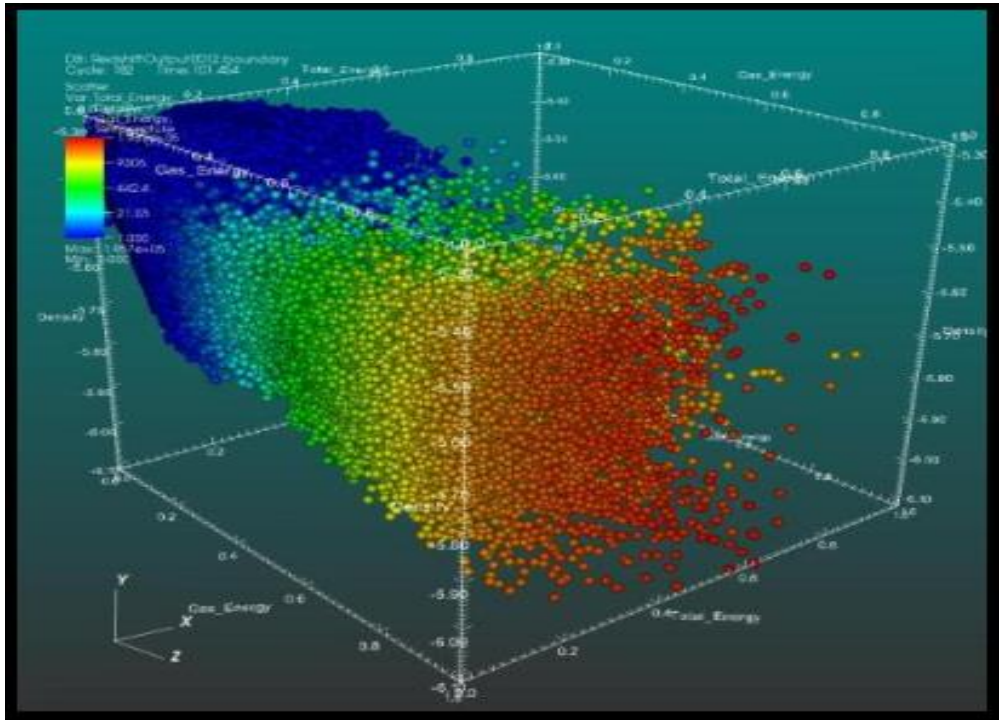
Line Chart

- 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.



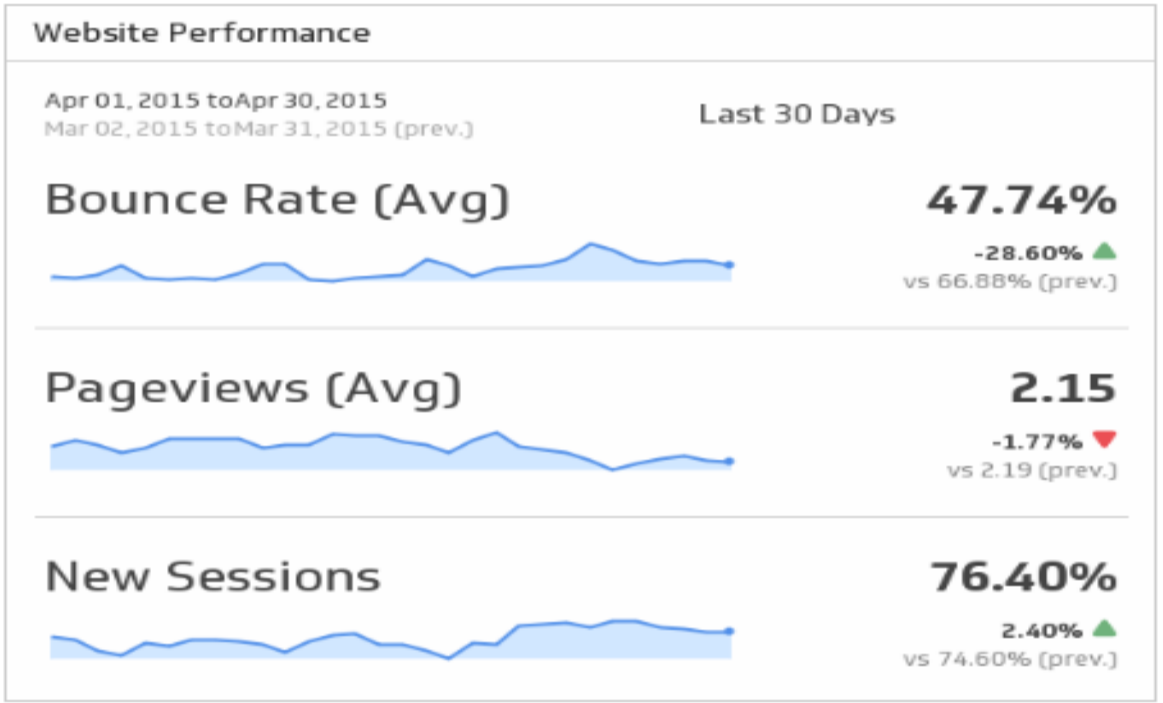
Scatterplot

- 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.



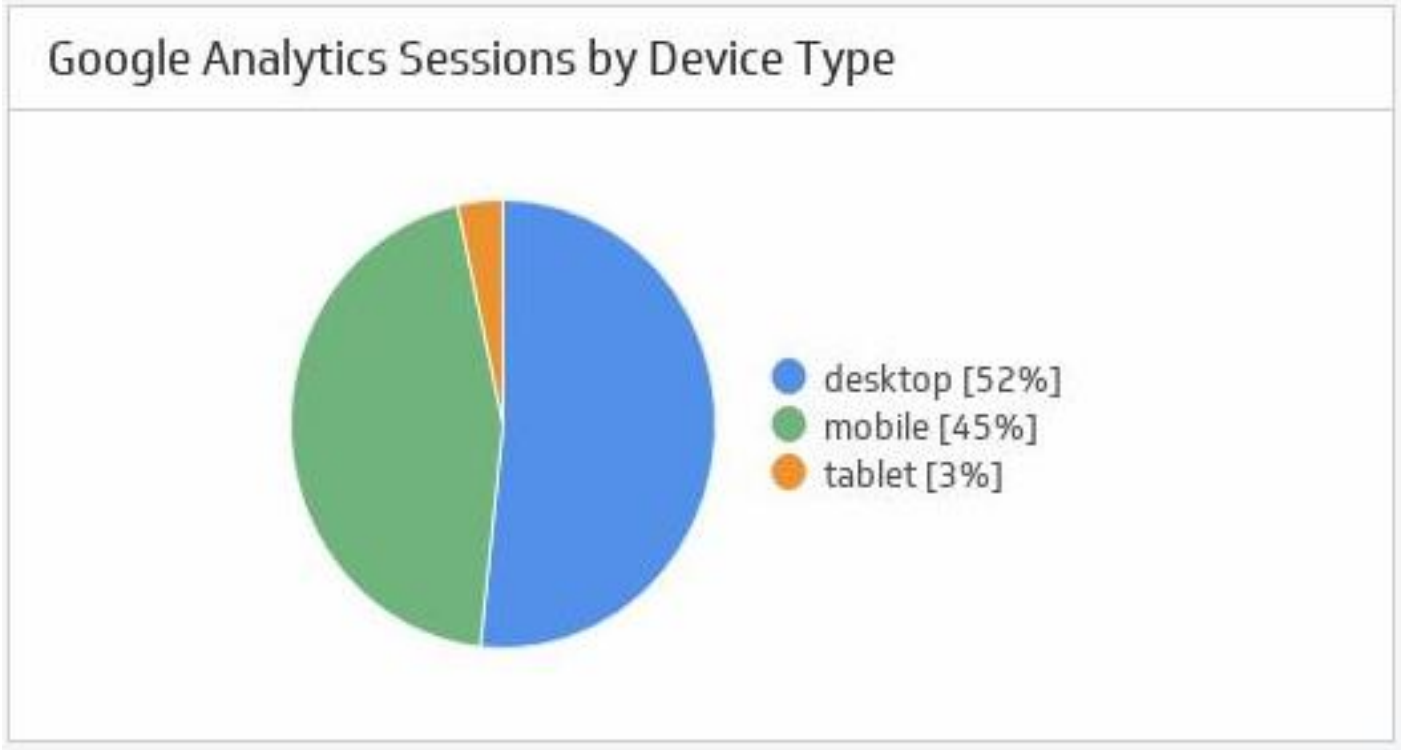
Sparkline

- 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.



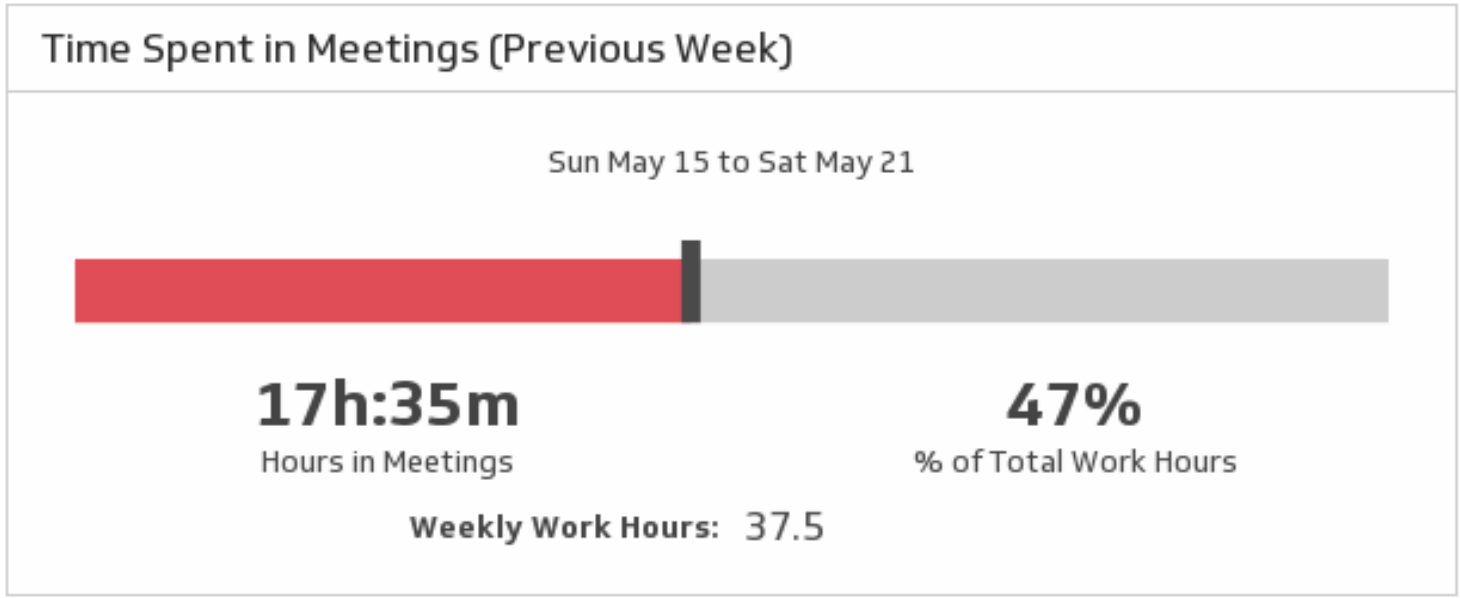
Pie Chart

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.



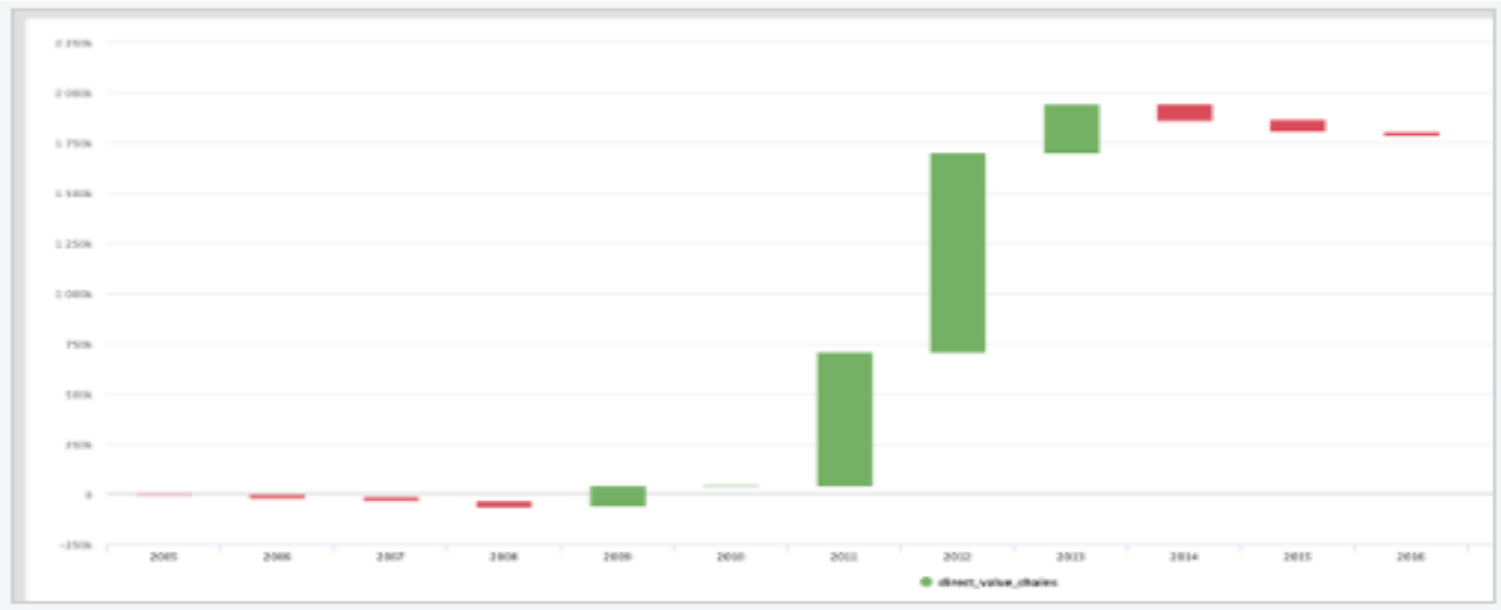
Gauge

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.



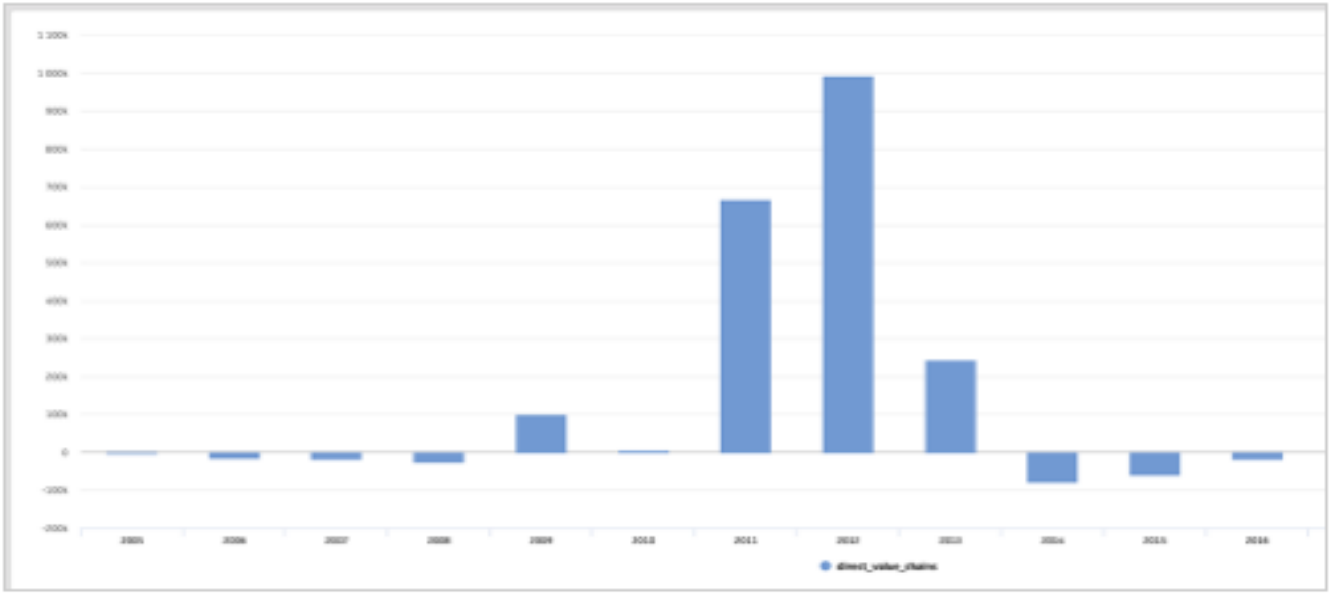
Waterfall Chart

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.