Data Analytics using Python

Dr. Vijaya kumar
Prof. Ramesh Ragala

May 20, 2020

Objectives

- Data and its importance
- Data Analytics and its types
- Important of Analytics in Business
- Interrelation of statistics, analytics and data science
- Different kinds of Data
- Need of Python and Demo

DATA

- **Variable:** it is a characteristic of any entity being studied that is capable taking on different values
- **Measurements:** it is, when a standard is process used to assign numbers to particular attributes or characteristics of a variable
- **Data:** Data is a recorded measurements or it refers to facts and statistics collected together for reference or analysis
- **Information:** it is the outcome of extraction and processing activities carried out on data, and it appears meaningful for those who receive it in a specific domain.
- **Knowledge:** Information is transformed into knowledge when it is used to make decisions and develop the corresponding actions.

Data Generation

43.9 Million
Wikipedia Articles

facebook

1.94 Billion Monthly users
1.28 Billion Active Users/day

flickr

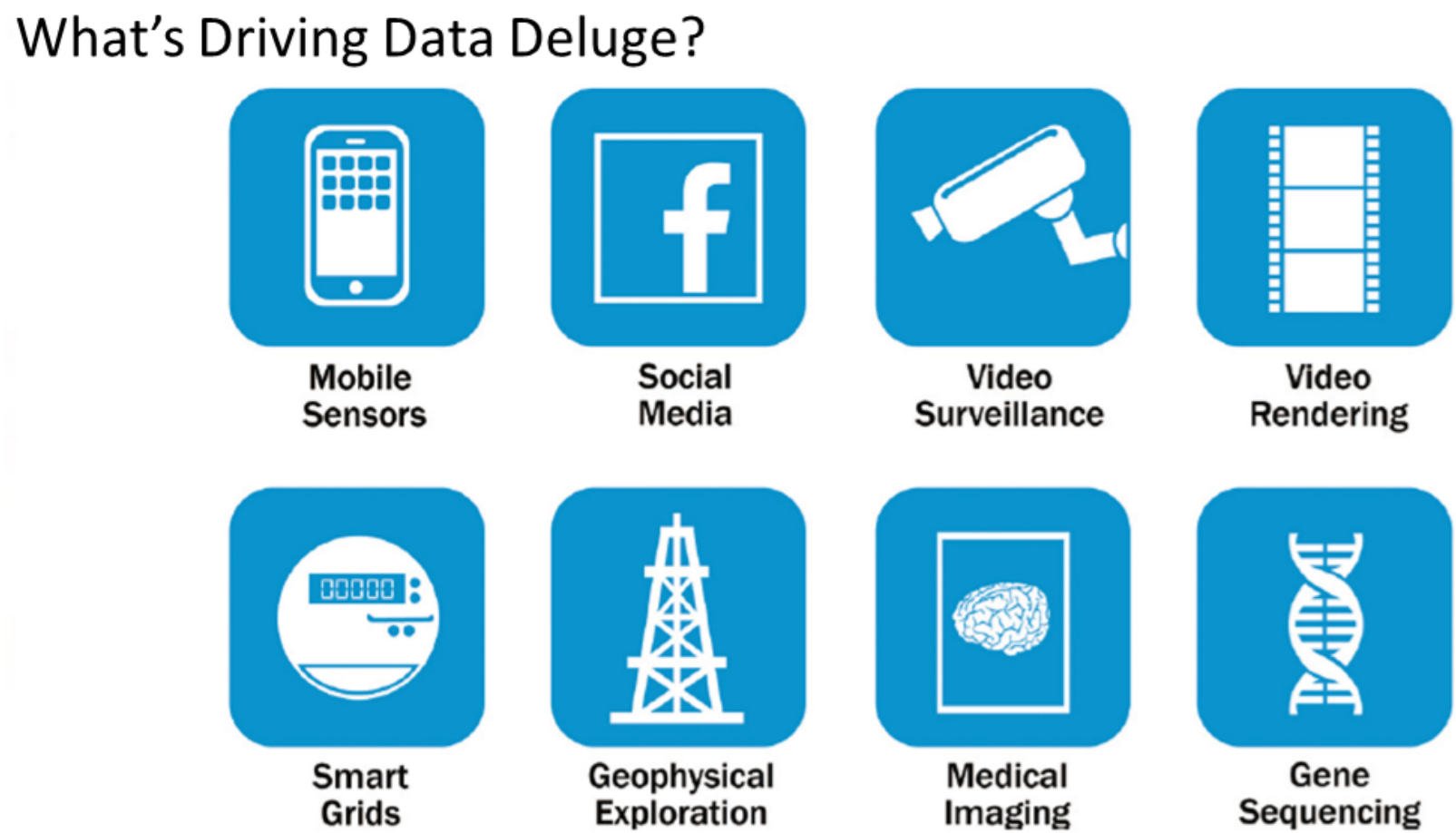
3.5 Millions new images every day
1 million photos sharing every day

YouTube

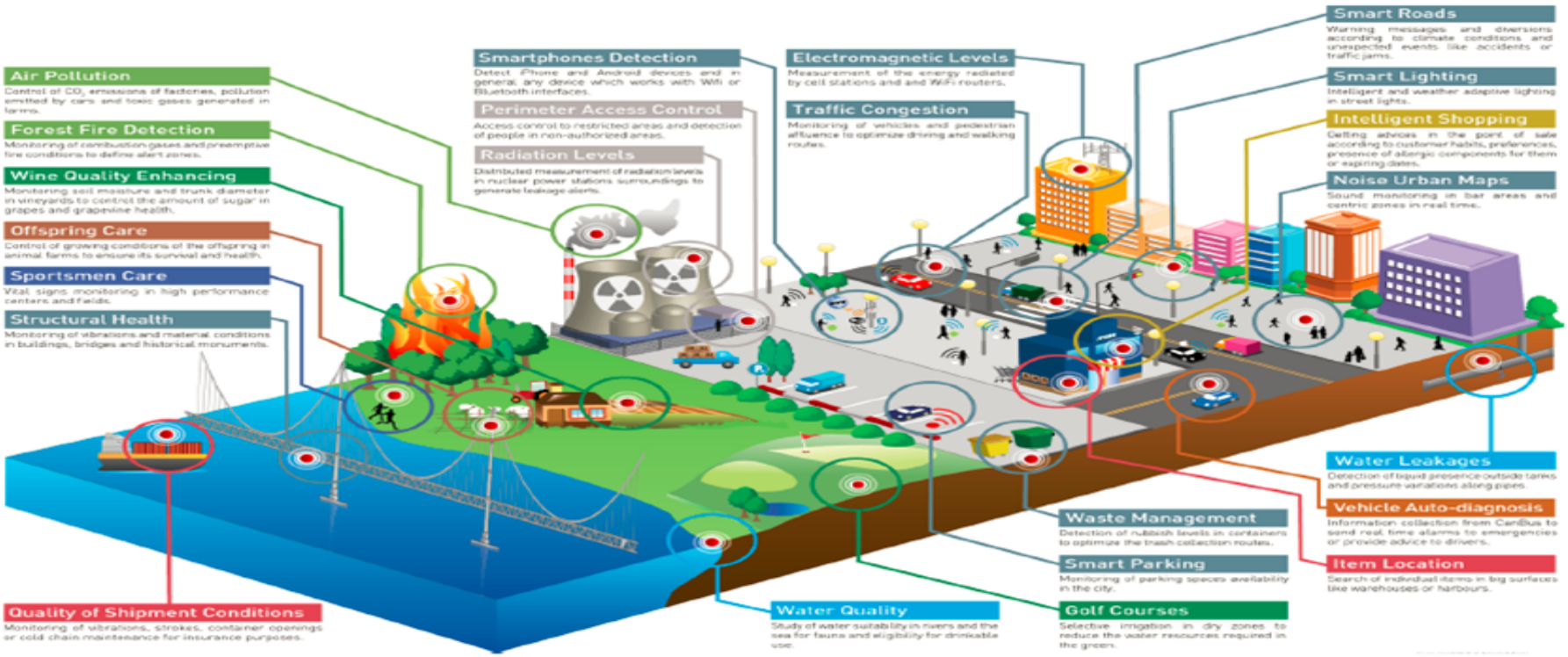
1 billion user
300 hours of videos per minute

Data Generation

- Data can be generated through
  - Human
  - Machines
  - The combination of Human and Machines $\rightarrow$ Social Networking

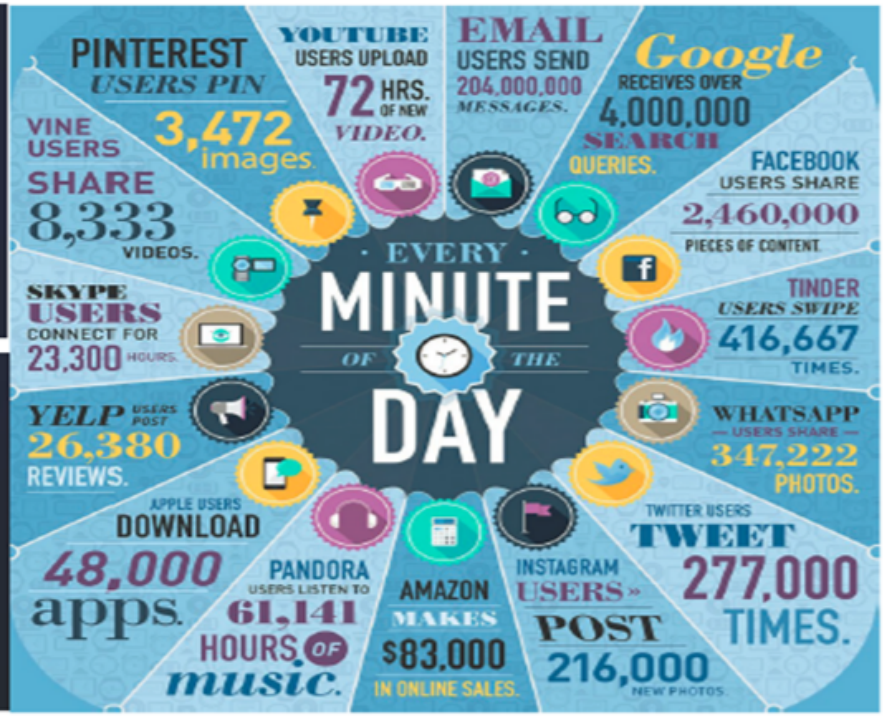


Data Generation



Data grows fast!

MORE IPHONES ARE SOLD THAN BABIES BORN



DATA

- while dealing with DATA, we observe four phases
  - Data collection, Data acquisition and storing $\rightarrow$ structured and unstructured formats
  - Measure $\rightarrow$ meta data about collected data
  - Analysed
  - Visualization

Importance of Data

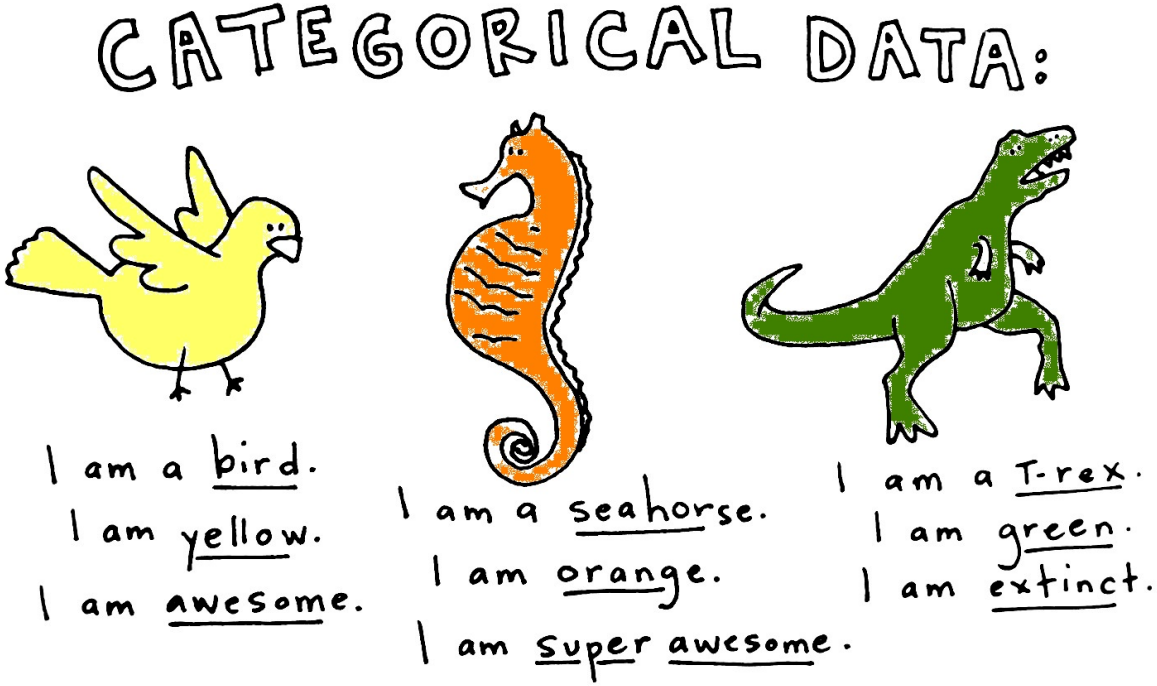
- It offers valuable insights for any business
- It helps make better decisions
- It helps in solve the problems by finding the reason for under-performance
- It helps to evaluate the performance
- It helps in improve the performance

Types of Data

- Quantitative Data
  - Continuous Data
  - Discrete Data
    - Interval Data
    - Ratio Data
- Qualitative Data
  - Ordinal Data
  - Nominal Data

- **Nominal/categorical data:**
  - The data values are categorical and not numeric.
  - A categorical variable is one that has two or more categories or labels or classes, but there is no intrinsic ordering to the categories.
  - simply Categorical variables represent types of data which may be divided into groups.
  - It is completely qualitative measurement.
  - Examples: age, gender, educational levels, countries, people names. operations: == and !=
  - Comparing two observations using the values for the variable, the observations will either be similar or different depending on whether the categorical value matches or not.

- **Example on Categorical Data:**



- if the categorical data has only two outcomes → binary or binomial data
- The Binomial data outcomes may pass/fail, live/dead or extinct/not extinct

- **Examples on Categorical Variables**

	A	B	C	D	E	F	G	H	I
1	Name	Miles Per Gallon	Accceleration	Horsepower	weight	cylinders	year	price	Country
2	Volkswagen Rabbit Dl	43,1	21,5	48	1985	4	78	2400	Germany
3	Ford Fiesta	36,1	14,4	66	1800	4	78	1900	Germany
4	Mazda GLC Deluxe	32,8	19,4	52	1985	4	78	2200	Japan
5	Datsun B210 GX	39,4	18,6	70	2070	4	78	2725	Japan
6	Honda Civic CVCC	36,1	16,4	60	1800	4	78	2250	Japan
7	Oldsmobile Cutlass	19,9	15,5	110	3365	8	78	3300	USA
8	Dodge Diplomat	19,4	13,2	140	3735	8	78	3125	USA
9	Mercury Monarch	20,2	12,8	139	3570	8	78	2850	USA

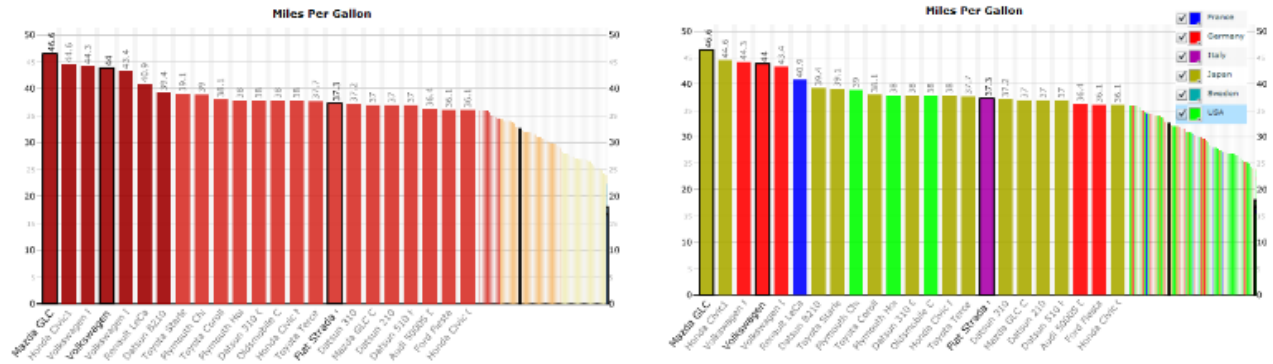


Figure: Classic car data set shown as bar chart for numerical variable "Miles per gallon" and coloured based on categorical variable Country.

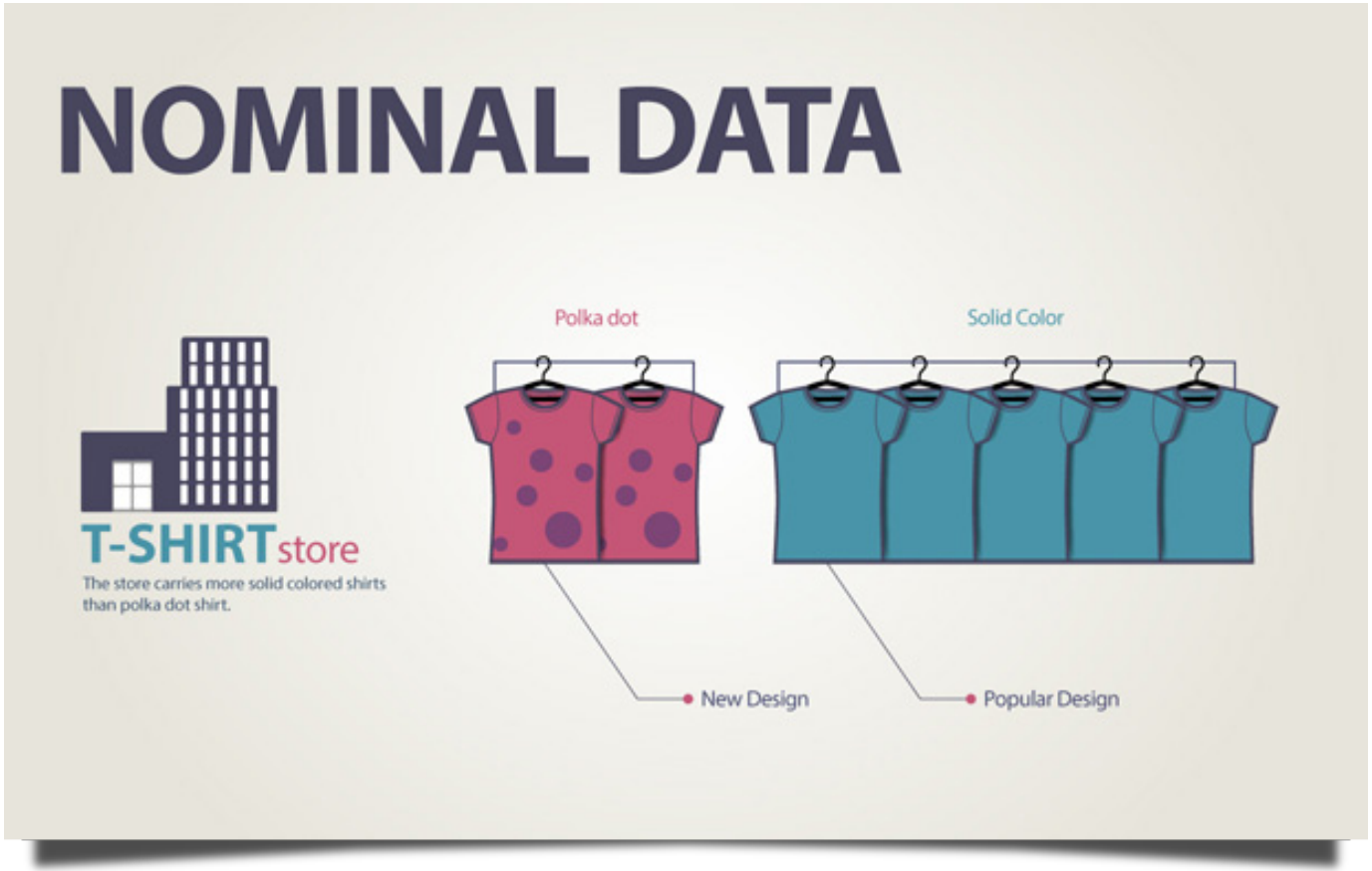


Figure: nominal level of measurement

- If the variable has a clear ordering, then that variable would be an ordinal variable.
- The Nominal or categorical data has only meaning $\rightarrow$ how they are differing from one another.
- **Example:** Country names are Nominal data values $\rightarrow$ putting all country names in alphabetical order is not making any relationship to another.
- Assignment of numbers to categories has no mathematical meaning.
- Nominal categories should be mutually exclusive and exhaustive

- **Where Can We Have Categorical Data:**
  - Social sciences : opinions on issues
  - Health sciences : response to treatments/drugs
  - Behavioral sciences : e.g. diagnose mental illness
  - Public health : AIDS awareness
  - Zoology : animals food preferences
  - Education : student's response to exams
  - Marketing : consumer preferences
  - Almost everywhere
- Distinction in categorical data are: Nominal Data and Ordinal Data

- **Ordinal data values:**
  - The data values are categorical but ordered.
  - Comparing two observations using the values for that variable.
  - Operations: $==, !=, \leq$ and $\geq$
  - it is mainly used for obey ordering relations among data values
  - Ordinal data is that which has inherent order, but no inherent degree of difference between what is being ordered.
  - **Example:** The Ist, IInd and IIIrd place winners in a race are on ordinal scale
  - But we do not know **how much faster** first place was than second place
  - But we know only that one was faster than other.

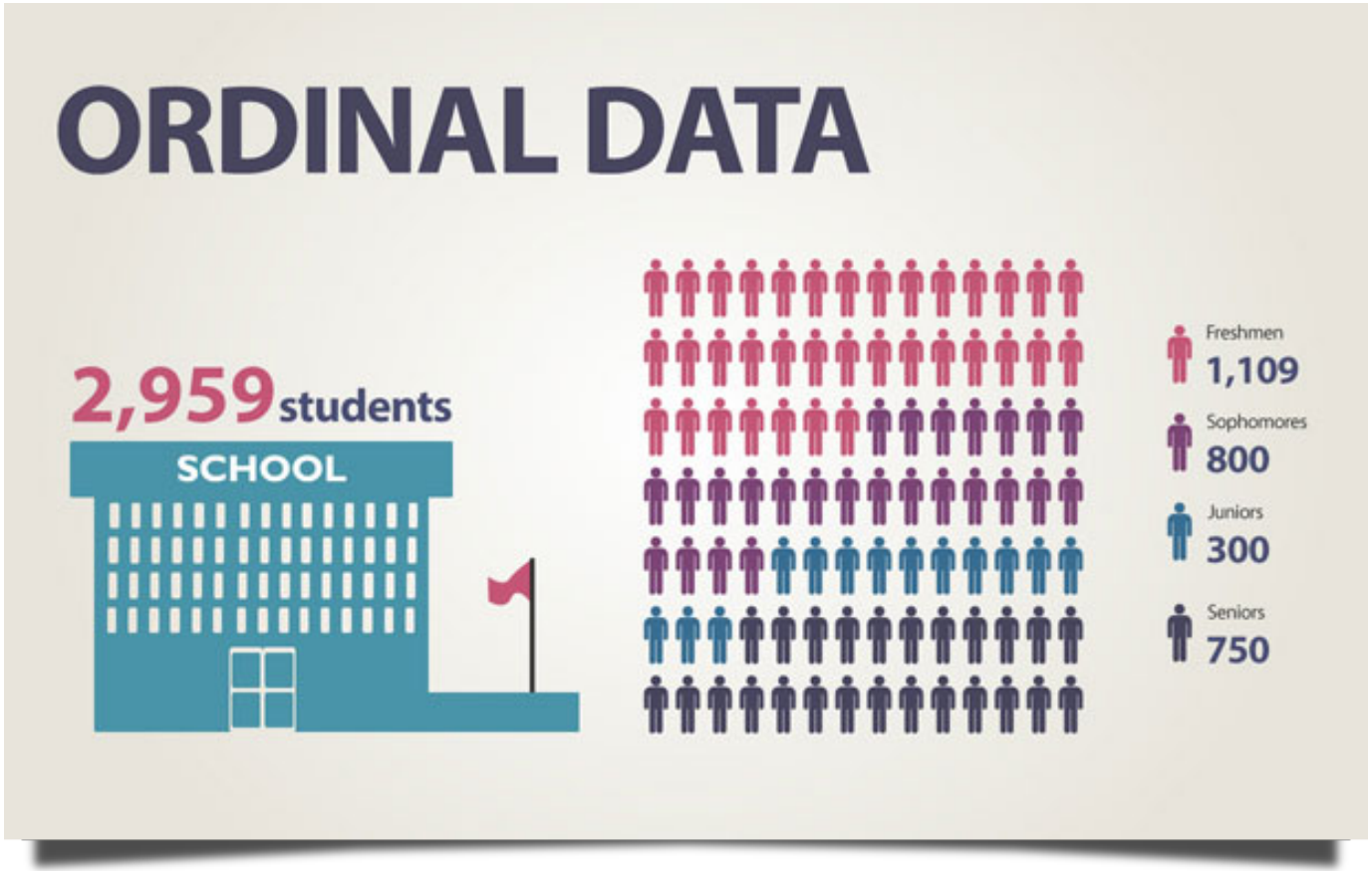
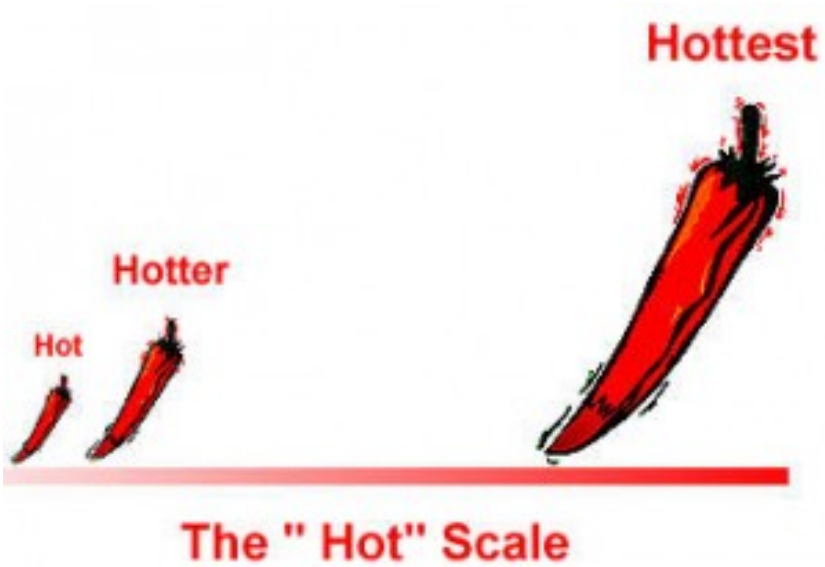


Figure: Ordinal level of measurement

Introduction: Semiotics



- **Interval Data:**
  - The data values are numeric.
  - It represents the more sensitive type of data or sophisticated form of measurement.
  - simply, Interval data is data which exists on a scale with meaningful quantitative magnitudes between values.
  -
  - Data values can be compared quantitatively using basic arithmetic operations +,-,* and / not the values themselves.
  - The values are ordered. it includes negative numbers and zero. But zero is not absolute reference point.
  - Scale data is usually aggregated or converted to averages.

- **Interval Data:**
  - **Example:1** The dataset does not contain an interval data variable, if there were a variable in a dataset that recorded the measurements of temperature. $\rightarrow$ it would be classified as a interval variable.
  - Temperature variable contains the values 40,60 and 80, we could say that compared with 40^{o}F, 80^{o}F is two times warmer than 60^{o}F (80-40)/(60-40), but not twice as hot because 0^{o}F is an arbitrarily chosen point on the scale.
  - **Example:2** if Sidda Reddy is rated as "6" on attractiveness and Durga Prasad a "3" $\rightarrow$ it does not mean Sidda Reddy is twice as attractive as Durga Prasad.

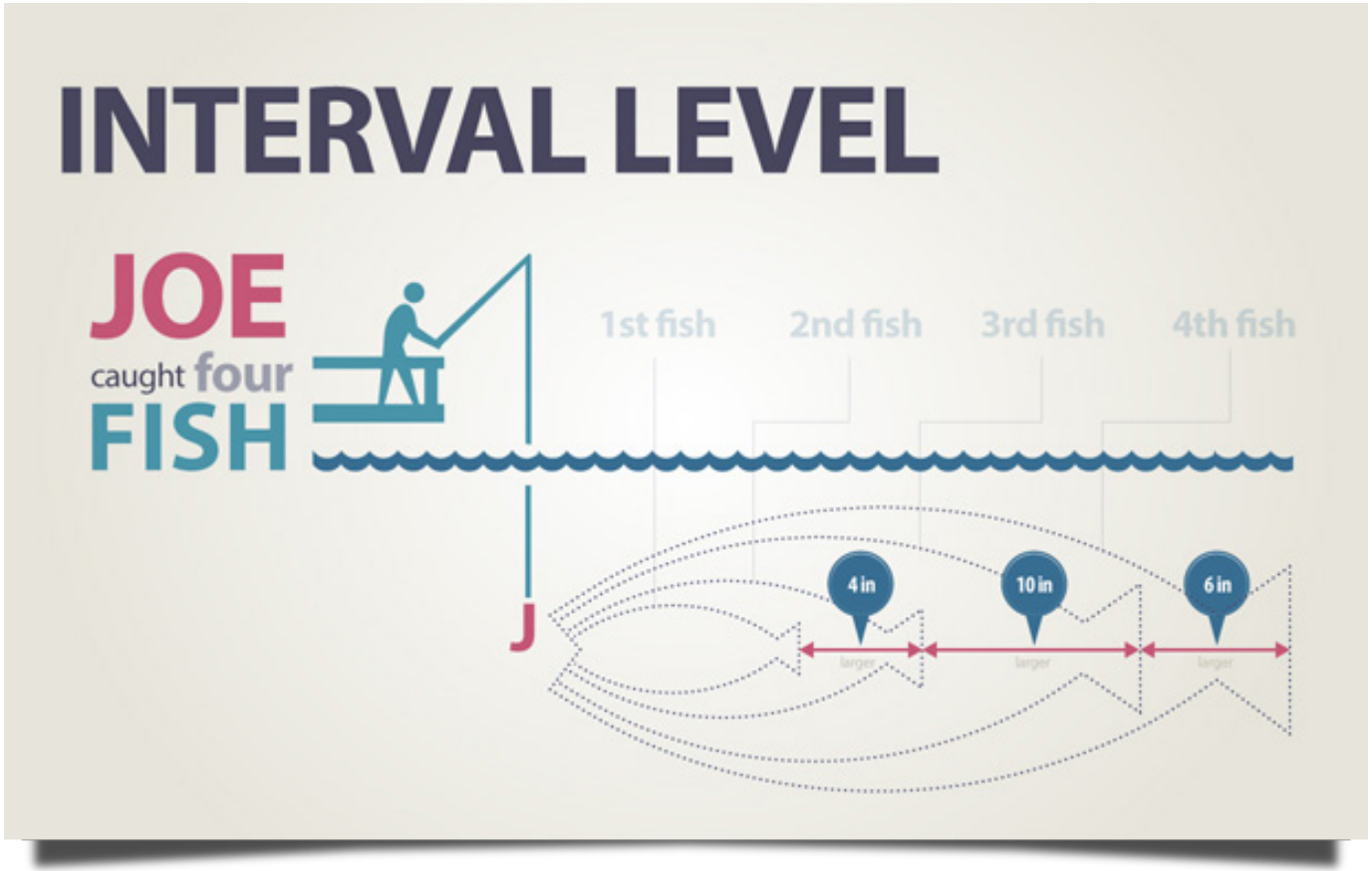


Figure: interval level of measurement

- The measurement between the sizes of the fish Joe caught in order of when he caught them.

- **Ratio Data:**
  - The Data Values are numeric and include an absolute zero.
  - This data values are allowed to compare quantitatively with other using basic arithmetic operations
  - Ratio data is data which, like interval data, has a meaningful order and a constant scale between ordered values, but additionally it has a meaningful zero value.
  - Supported Operations are ==, !=, $\leq$, $\geq$, -, / and *
  - The Ratio level of measurement applies to data that can be arranged in order.
  - In addition, both differences between data values and ratios of data values are meaningful. Data at the ratio level have a true zero.
  - **Example:** If one box weighs 50lbs and another 100lbs $\rightarrow$ the second box weighs twice as much as the first $\rightarrow$ this is not a case in interval data

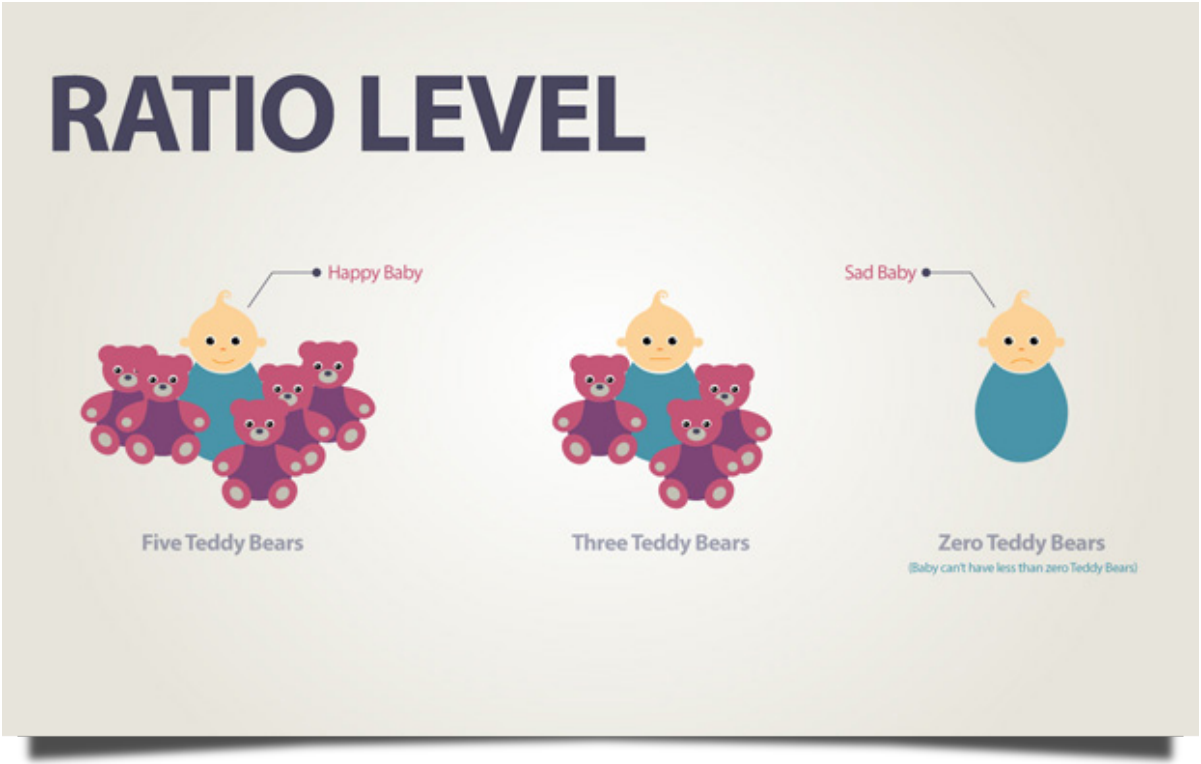


Figure: Ratio level of measurement

- The amounts of teddy bears a certain child has.
- Since we cant have less than zero teddy bears, then the ratio level has a true zero.

- Data Analytics is defined as **the scientific process of transforming data into insights for making better decisions**
- It is defined as a set of **mathematical models** and **analysis methodologies** that exploit the available data to generate **information** and **knowledge** useful for complex **decision-making processes**.
- Previously, Knowledge workers are used to take decisions using easy and intuitive methodologies $\rightarrow$ experience, knowledge of the application domain and the available information.
- This approach leads to a **stagnant** decision-making style which is **inappropriate** for the **unstable conditions** $\rightarrow$ frequent and rapid changes in the environment.
- Decision making Process in today's organizations should dynamic, requires rigorous attitude based on analytical methodologies and mathematical models.

- It is the process of examining, transforming and arranging raw data in a specific way to generate useful information from it.
- It allows for evaluation of data through analytical and logical reasoning to lead to some sort of outcome or conclusion in some context.
- It is a multi-faceted process that involves a number of steps, approaches and diverse techniques.

Data Analytics Vs. Data Analysis

- Data Analysis deals the question related to How and Why on historical data.
- The questions tend to be closed-ended in Data Analysis
- Data Analytics tends to use disaggregated data in a more forward-looking, exploratory way, focusing on analyzing the present and enabling informed decisions about the future.
- Open-ended question can be answered using Data Analytics
- **Conclusion:**
  - Analytics $\neq$ Analysis
  - Data Analytics $\neq$ Data Analysis

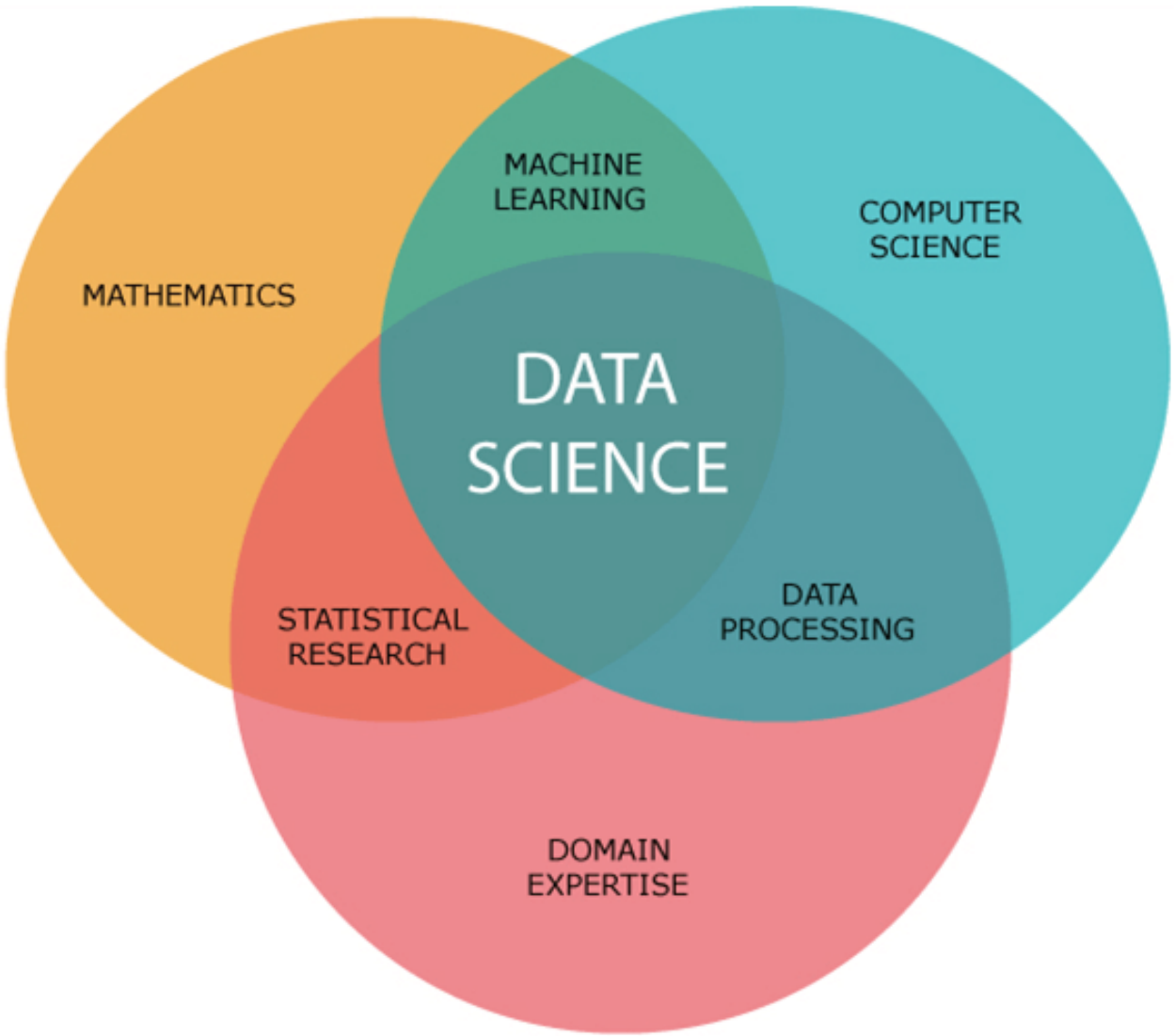
Classification of Data Analytics

- Based on the phases of workflow and Sort of analysis required $\rightarrow$ Four major types of data analytics
  - Descriptive Analytics
  - Diagnostic Analytics
  - Predictive Analytics
  - Prescriptive Analytics

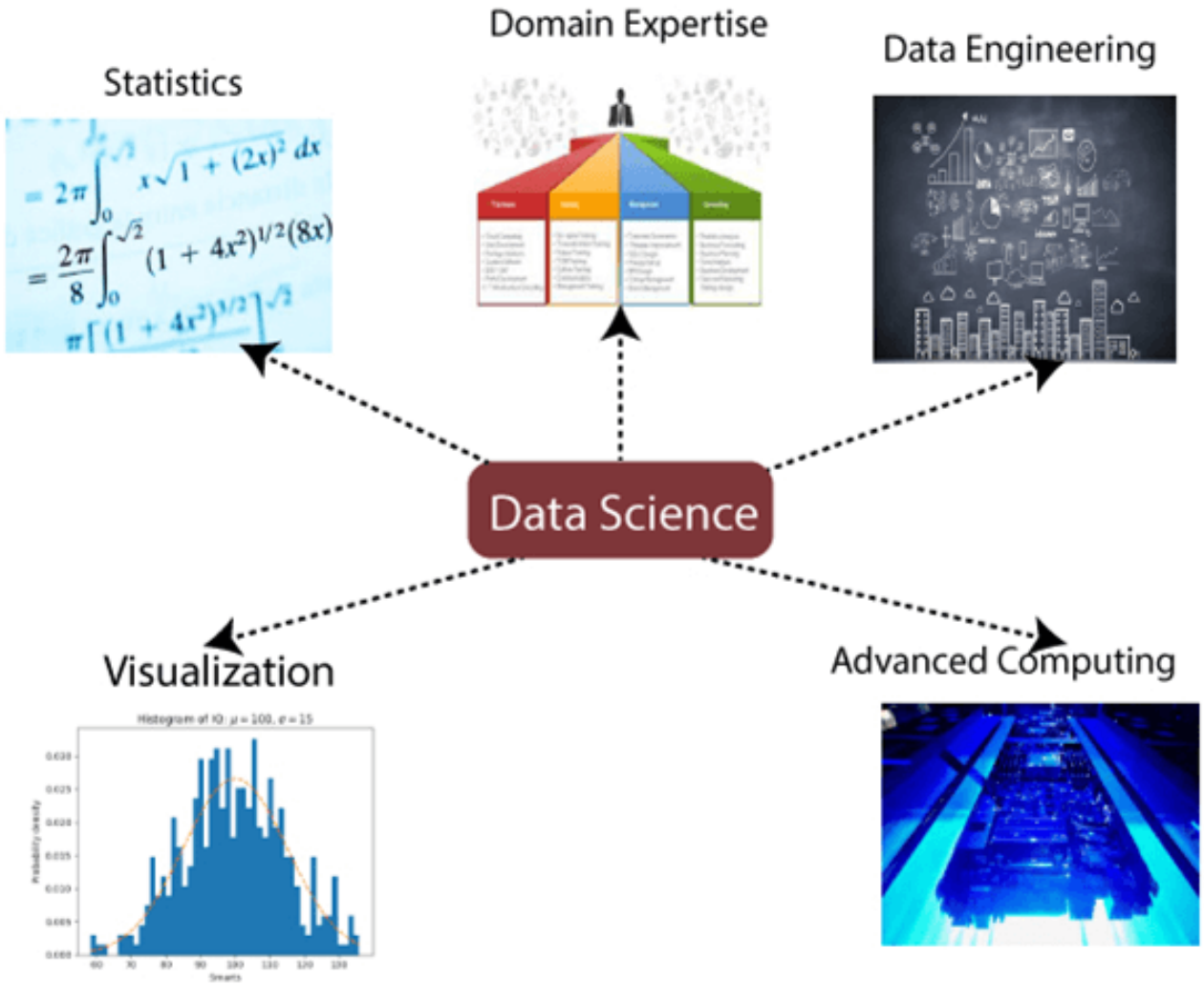
Classification of Data Analytics

- Based on the phases of workflow and Sort of analysis required → Four major types of data analytics
  - Descriptive Analytics → What happened?
  - Diagnostic Analytics → Why did it happened?
  - Predictive Analytics → What will happen?
  - Prescriptive Analytics → How can we make it happen?

Core Elements of Data Analytics and Data Science



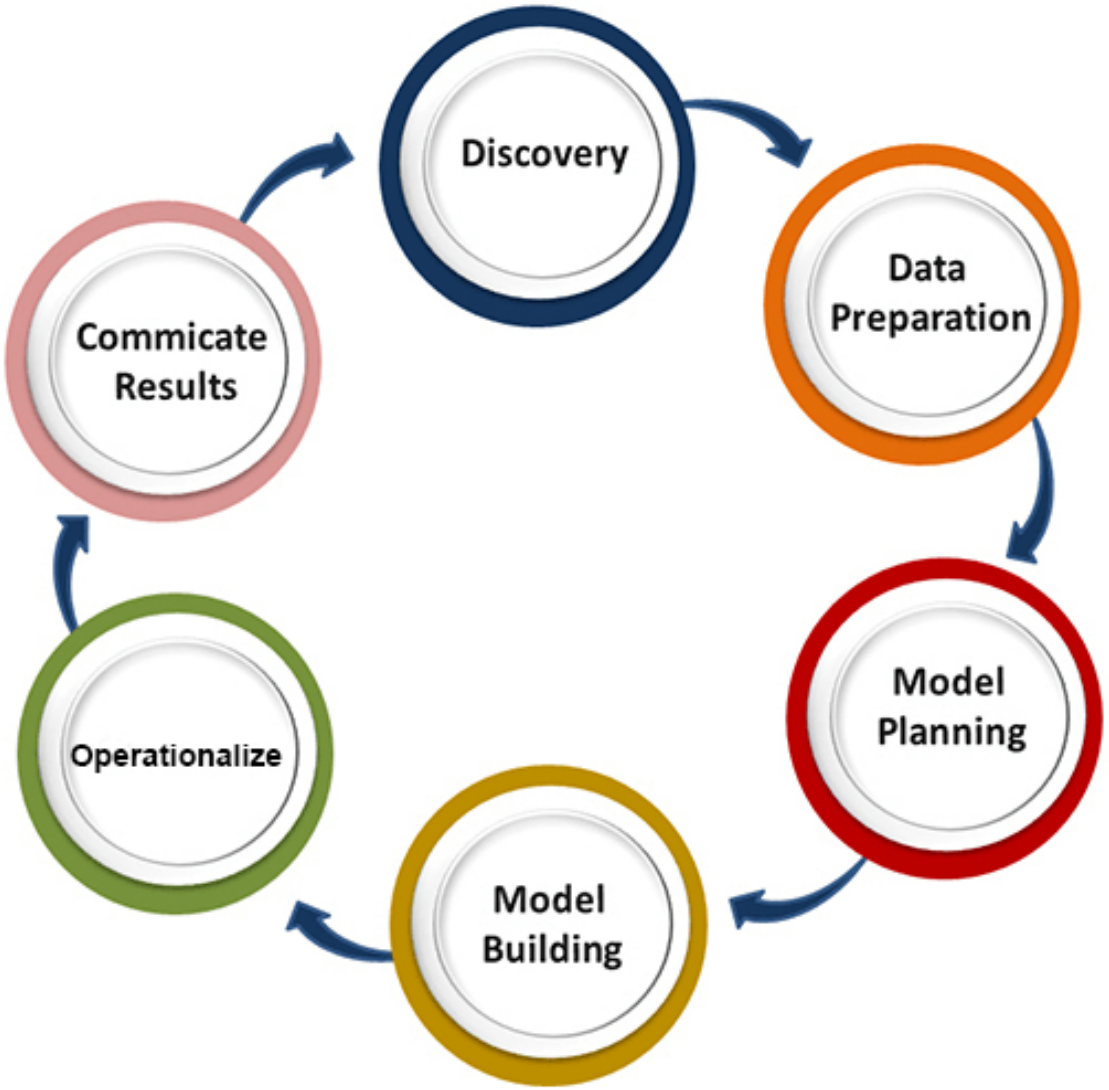
Components of Data Science



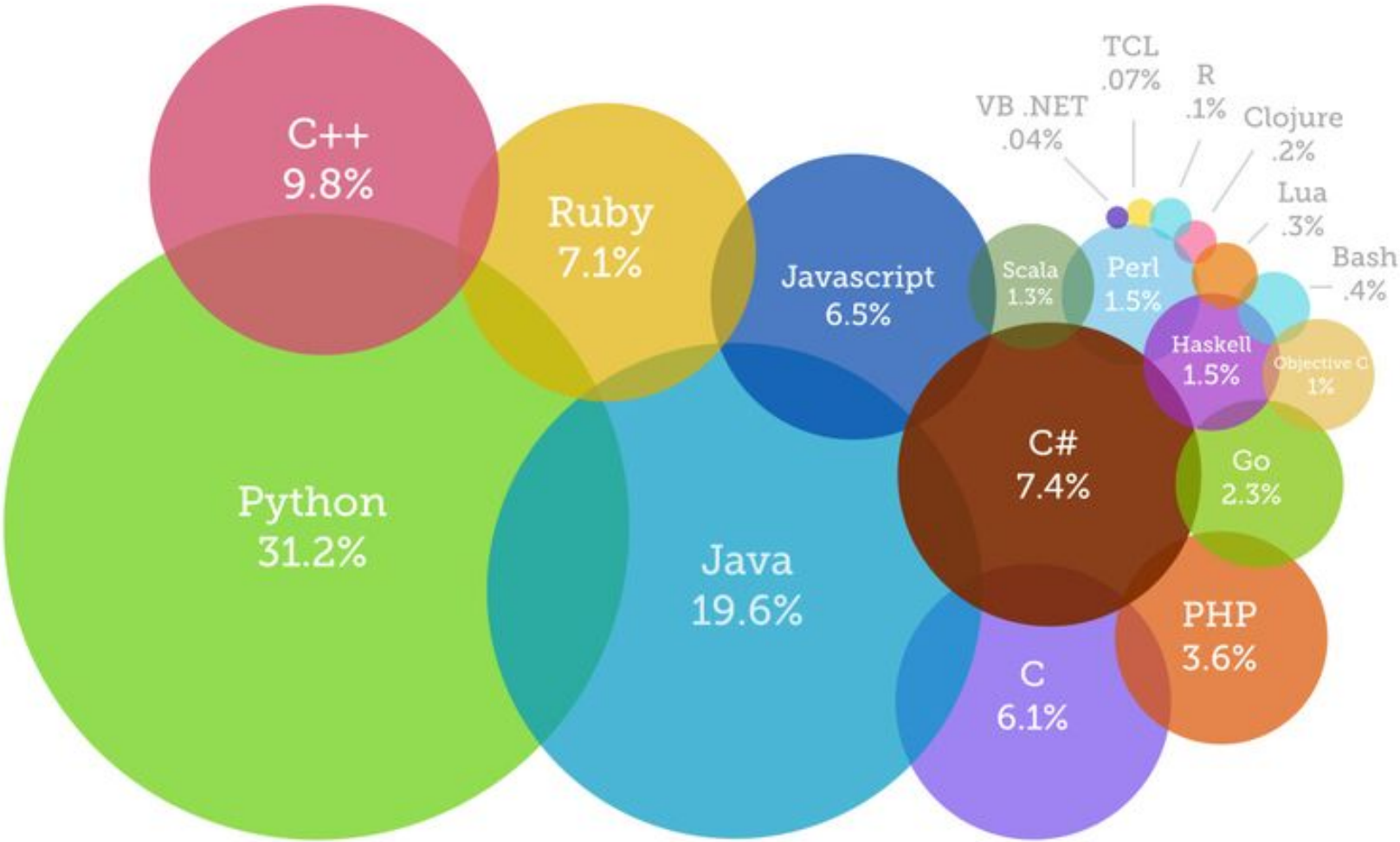
Data Analytics Lifecycle

- It defines the analytics process and best practices from discovery to project completion.
- The phases in the data analytics lifecycle
  - Discovery Phase
  - Data Preparation Phase
  - Model Planning Phase
  - Model Building Phase
  - Operationalize
  - Communicate Result
- With six phases the project work can occur in several phases simultaneously
- The cycle is iterative to portray a real project
- Work can return to earlier phases as new information is uncovered

Data Analytics Lifecycle



- Python Ranking



- Python Ranking according to IEEE

Rank	Language	Type	Score
1	Python		100.0
2	Java		96.3
3	C		94.4
4	C++		87.5
5	R		81.5
6	JavaScript		79.4
7	C#		74.5
8	Matlab		70.6
9	Swift		69.1

Python Users

- Python has a **simple syntax** and very **few keywords**.
- Python programs are clear and easy to **read** and **Understand**.
- It has **Powerful programming features** and **highly portable** and **extensible**
- Python is a **High Level Language**.
- Machine Languages or Assembly Languages are referred as Low Level Languages
- High Level Languages have to be processed before they can run. $\rightarrow$ extra time.
- Two types of programs translators to convert High Level Program to Low Level program
  - Compiler
  - Interpreter

- Object oriented language

- Object oriented language
- Interpreted language

- Object oriented language
- Interpreted language
- Supports dynamic data type

- Object oriented language
- Interpreted language
- Supports dynamic data type
- Independent from platforms

- Object oriented language
- Interpreted language
- Supports dynamic data type
- Independent from platforms
- Focused on development time

- Object oriented language
- Interpreted language
- Supports dynamic data type
- Independent from platforms
- Focused on development time
- Simple and easy grammar

- Object oriented language
- Interpreted language
- Supports dynamic data type
- Independent from platforms
- Focused on development time
- Simple and easy grammar
- Its free Invalid Identifiers(open source)

- Object oriented language
- Interpreted language
- Supports dynamic data type
- Independent from platforms
- Focused on development time
- Simple and easy grammar
- Its free Invalid Identifiers(open source)
- Automatic memory management

- Object oriented language
- Interpreted language
- Supports dynamic data type
- Independent from platforms
- Focused on development time
- Simple and easy grammar
- Its free Invalid Identifiers(open source)
- Automatic memory management
- Glue language Interactive front-end for FORTRAN/C/C++ code

Python Language Properties

- Everything is an object

Python Language Properties

- Everything is an object
- Modules, classes, functions

Python Language Properties

- Everything is an object
- Modules, classes, functions
- Exception handling

Python Language Properties

- Everything is an object
- Modules, classes, functions
- Exception handling
- Dynamic typing, polymorphism

Python Language Properties

- Everything is an object
- Modules, classes, functions
- Exception handling
- Dynamic typing, polymorphism
- Static scoping

Python Language Properties

- Everything is an object
- Modules, classes, functions
- Exception handling
- Dynamic typing, polymorphism
- Static scoping
- Operator overloading

Python Language Properties

- Everything is an object
- Modules, classes, functions
- Exception handling
- Dynamic typing, polymorphism
- Static scoping
- Operator overloading
- Indentation for block structure

Where to use Python

- System management (i.e., scripting)

Where to use Python

- System management (i.e., scripting)
- Graphic User Interface (GUI)

Where to use Python

- System management (i.e., scripting)
- Graphic User Interface (GUI)
- Internet programming

Where to use Python

- System management (i.e., scripting)
- Graphic User Interface (GUI)
- Internet programming
- Database (DB) programming

Where to use Python

- System management (i.e., scripting)
- Graphic User Interface (GUI)
- Internet programming
- Database (DB) programming
- Text data processing

Where to use Python

- System management (i.e., scripting)
- Graphic User Interface (GUI)
- Internet programming
- Database (DB) programming
- Text data processing
- Distributed processing

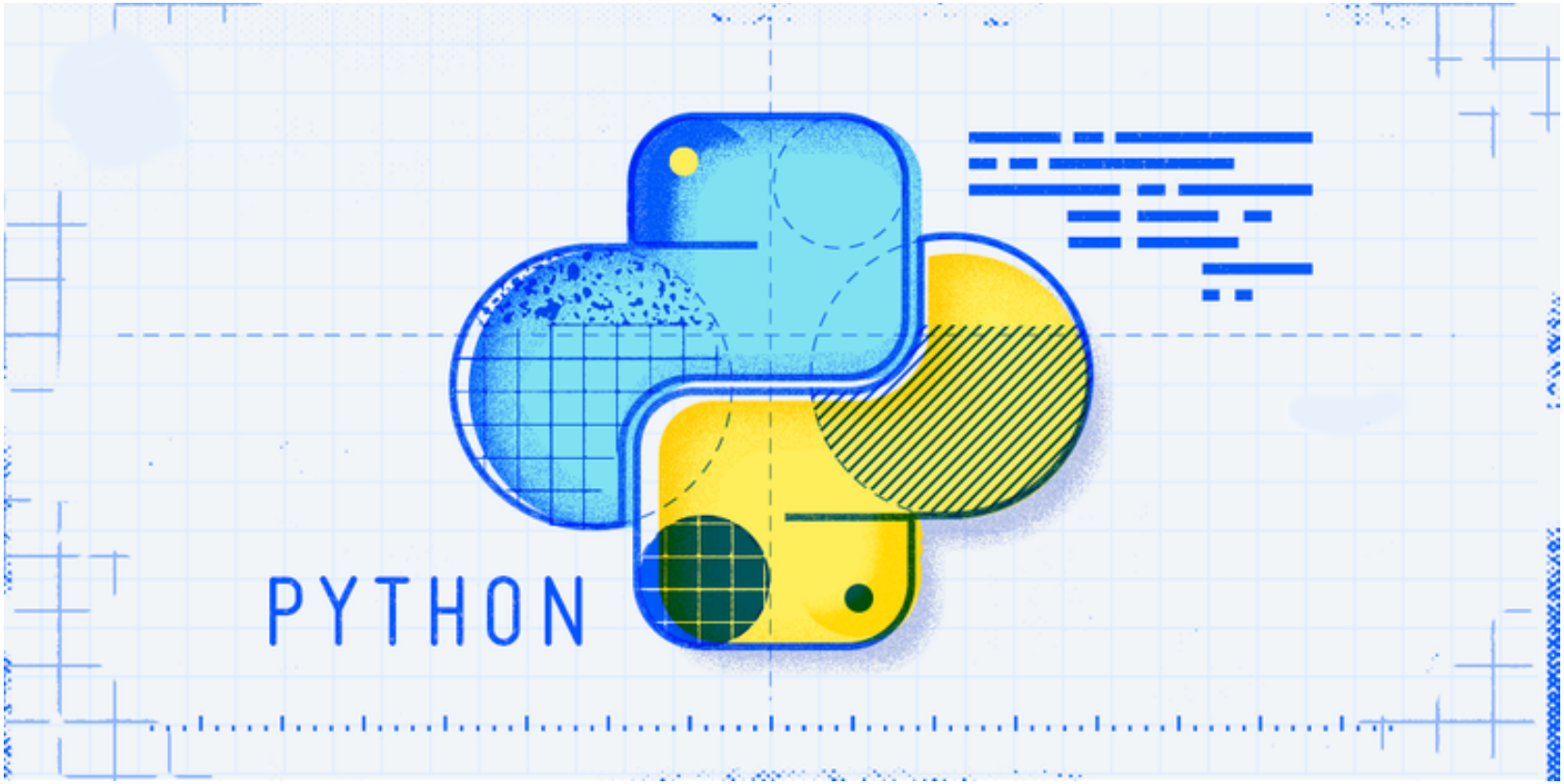
Where to use Python

- System management (i.e., scripting)
- Graphic User Interface (GUI)
- Internet programming
- Database (DB) programming
- Text data processing
- Distributed processing
- Numerical operations

Where to use Python

- System management (i.e., scripting)
- Graphic User Interface (GUI)
- Internet programming
- Database (DB) programming
- Text data processing
- Distributed processing
- Numerical operations
- Graphics so on...

Demo



Thank You
For Your Attention

Contact MailID: ramesh.ragala@vit.ac.in
vijayakumar.varadarajan@gmail.com