Objective 2: Basic Understanding of Data & Data Distribution

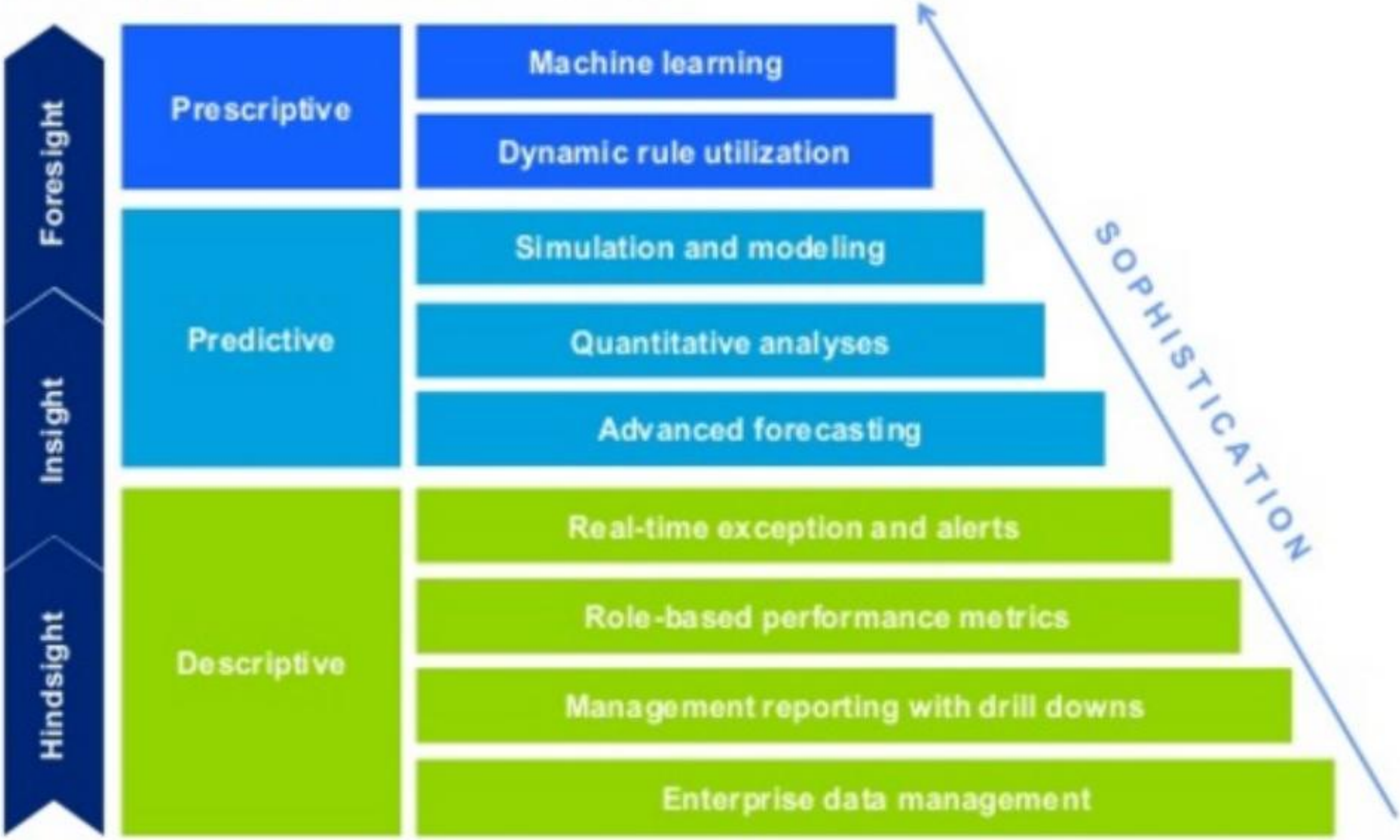
- **Different type of Data & Data Sources**

- **Scale of Measurement**

- **Kind of analysis to apply**

- **Checking data distribution and different measurement**

Analytics is the practice of deriving insights from data to make more effective decisions.

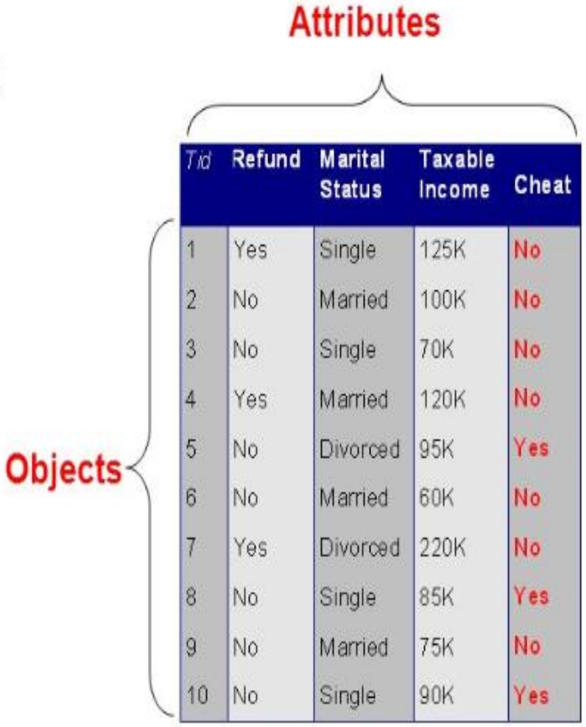


Data and Distributions

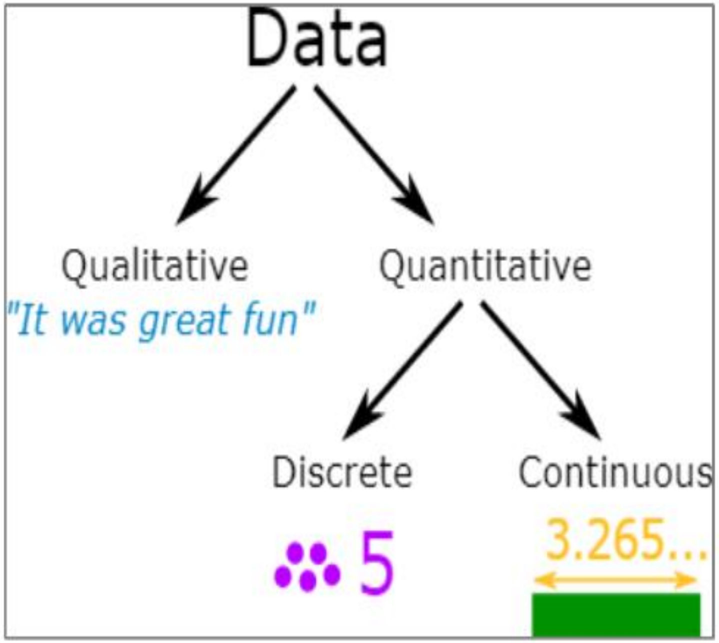
Data and Information

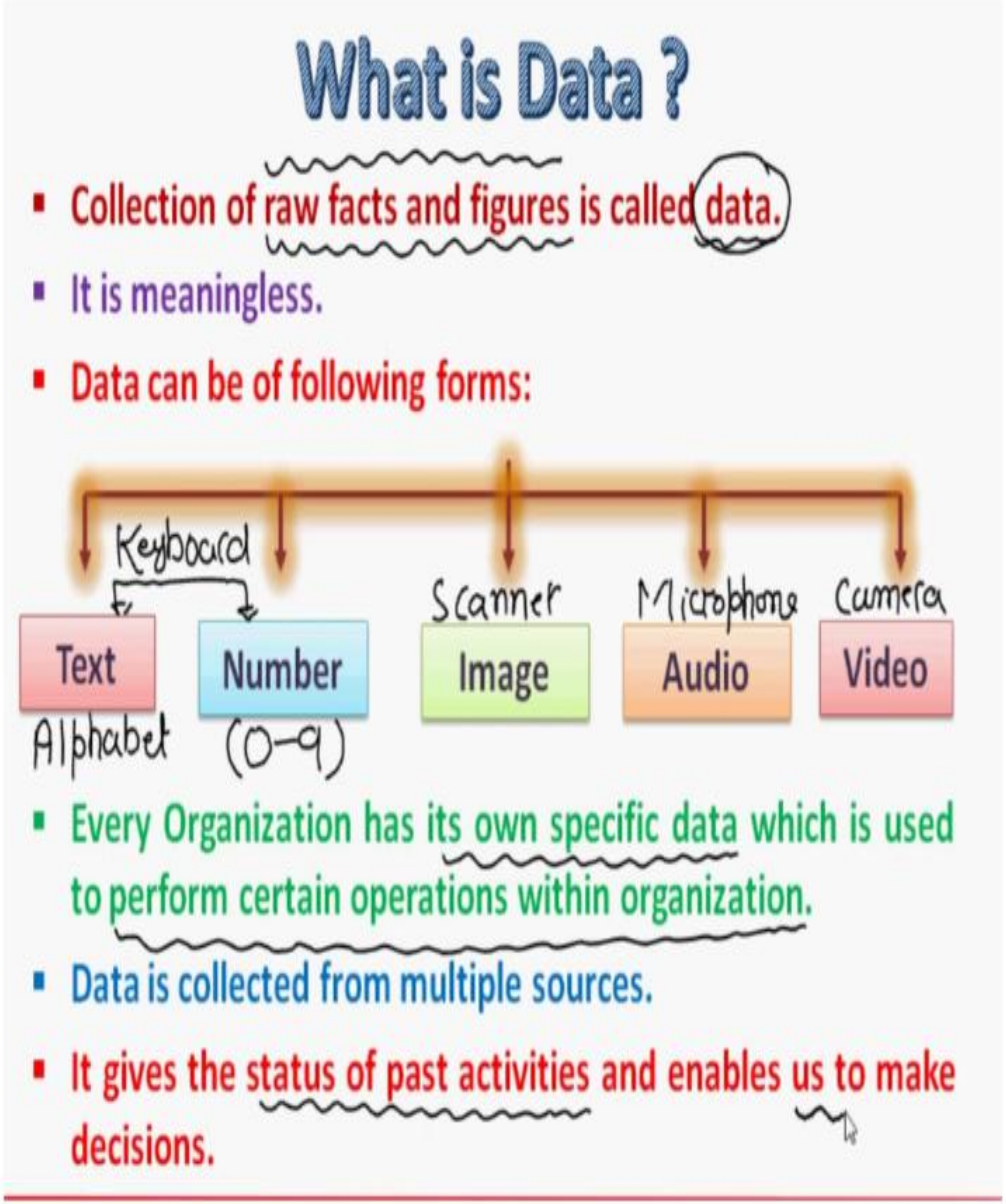
- Facts, statistics used for reference or analysis.
- Numbers, characters, symbols, images etc., which can be processed by a computer.
- Data must be interpreted, by a human or machine, to derive meaning
- So data is meaningless

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, or feature
- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, or instance


Attributes
Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes





What is Data ?

- Collection of raw facts and figures is called data.
- It is meaningless.
- Data can be of following forms:

Keyboard

Text — Alphabet

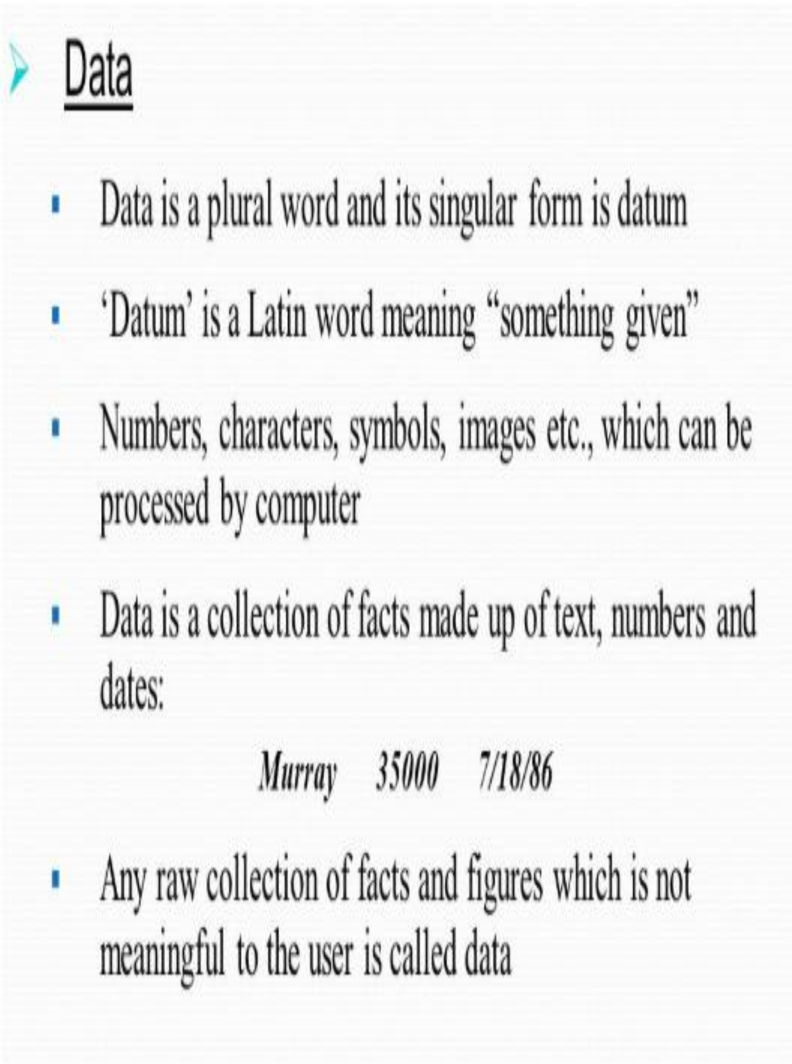
Number — (0-9)

Scanner — Image

Microphone — Audio

Camera — Video

- Every Organization has its own specific data which is used to perform certain operations within organization.
- Data is collected from multiple sources.
- It gives the status of past activities and enables us to make decisions.

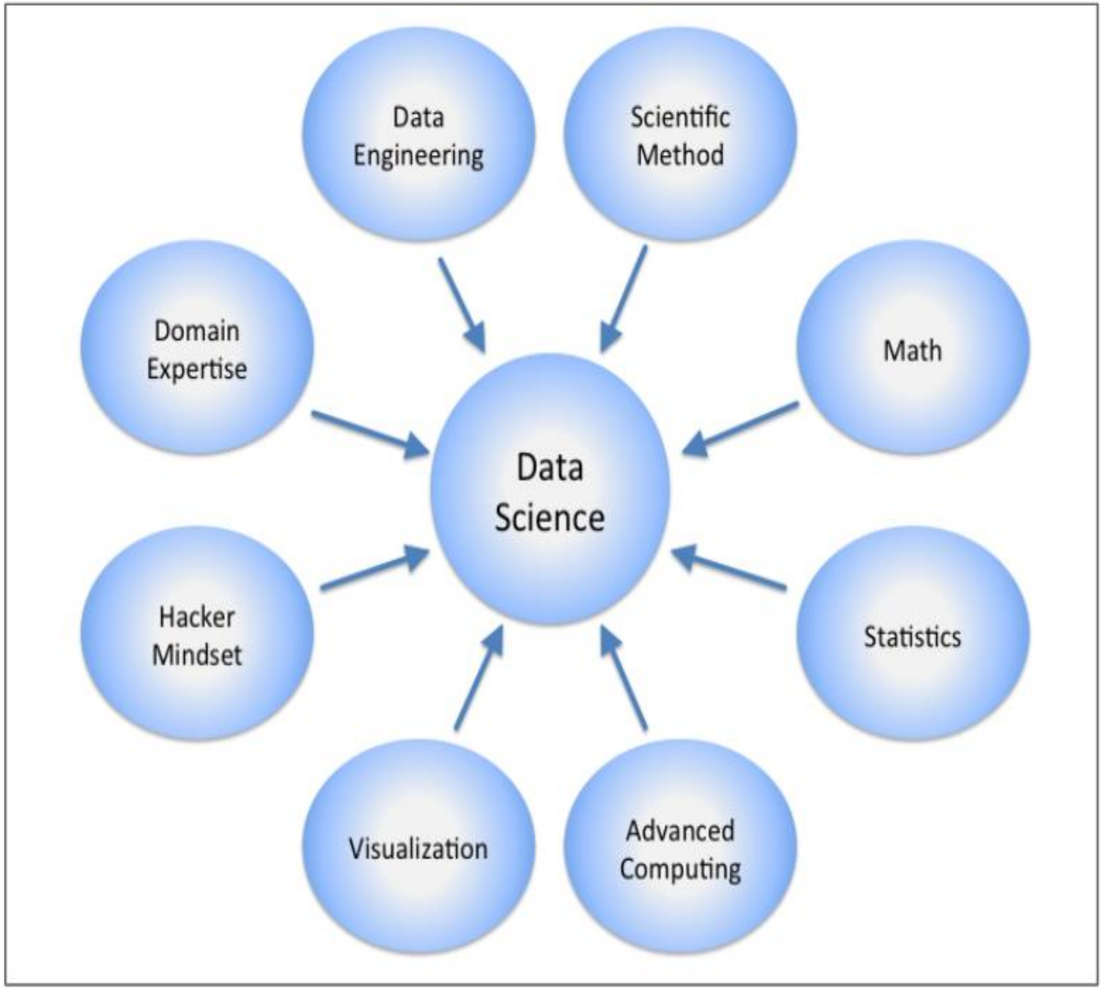


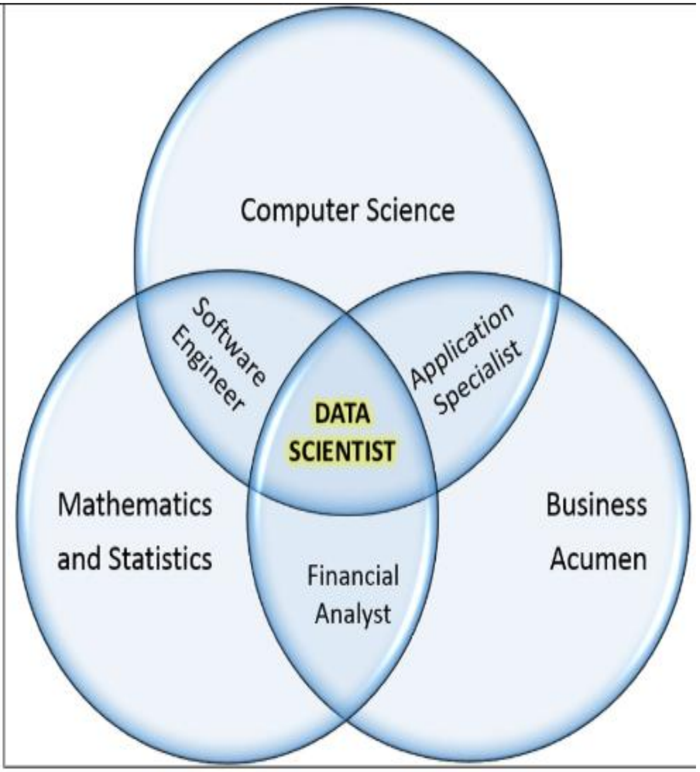
Data

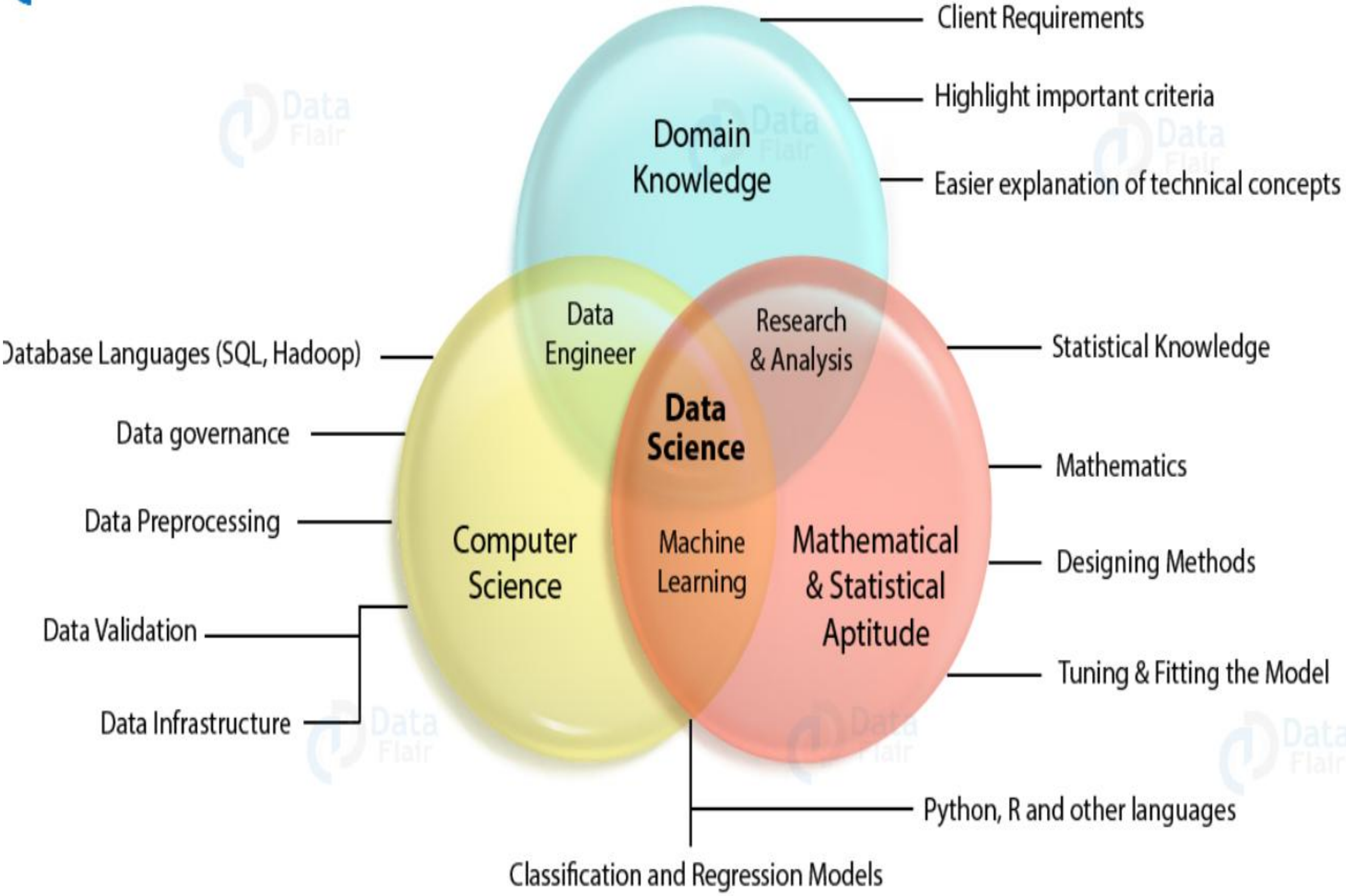
- Data is a plural word and its singular form is datum
- 'Datum' is a Latin word meaning "something given"
- Numbers, characters, symbols, images etc., which can be processed by computer
- Data is a collection of facts made up of text, numbers and dates:

  Murray 35000 7/18/86

- Any raw collection of facts and figures which is not meaningful to the user is called data


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


Computer Science
Software Engineer
Application Specialist
DATA SCIENTIST
Mathematics and Statistics
Business Acumen
Financial Analyst



Domain Knowledge

Client Requirements

Highlight important criteria

Easier explanation of technical concepts

Computer Science

Database Languages (SQL, Hadoop)

Data governance

Data Preprocessing

Data Validation

Data Infrastructure

Mathematical & Statistical Aptitude

Statistical Knowledge

Mathematics

Designing Methods

Tuning & Fitting the Model

Data Engineer

Research & Analysis

Data Science

Machine Learning

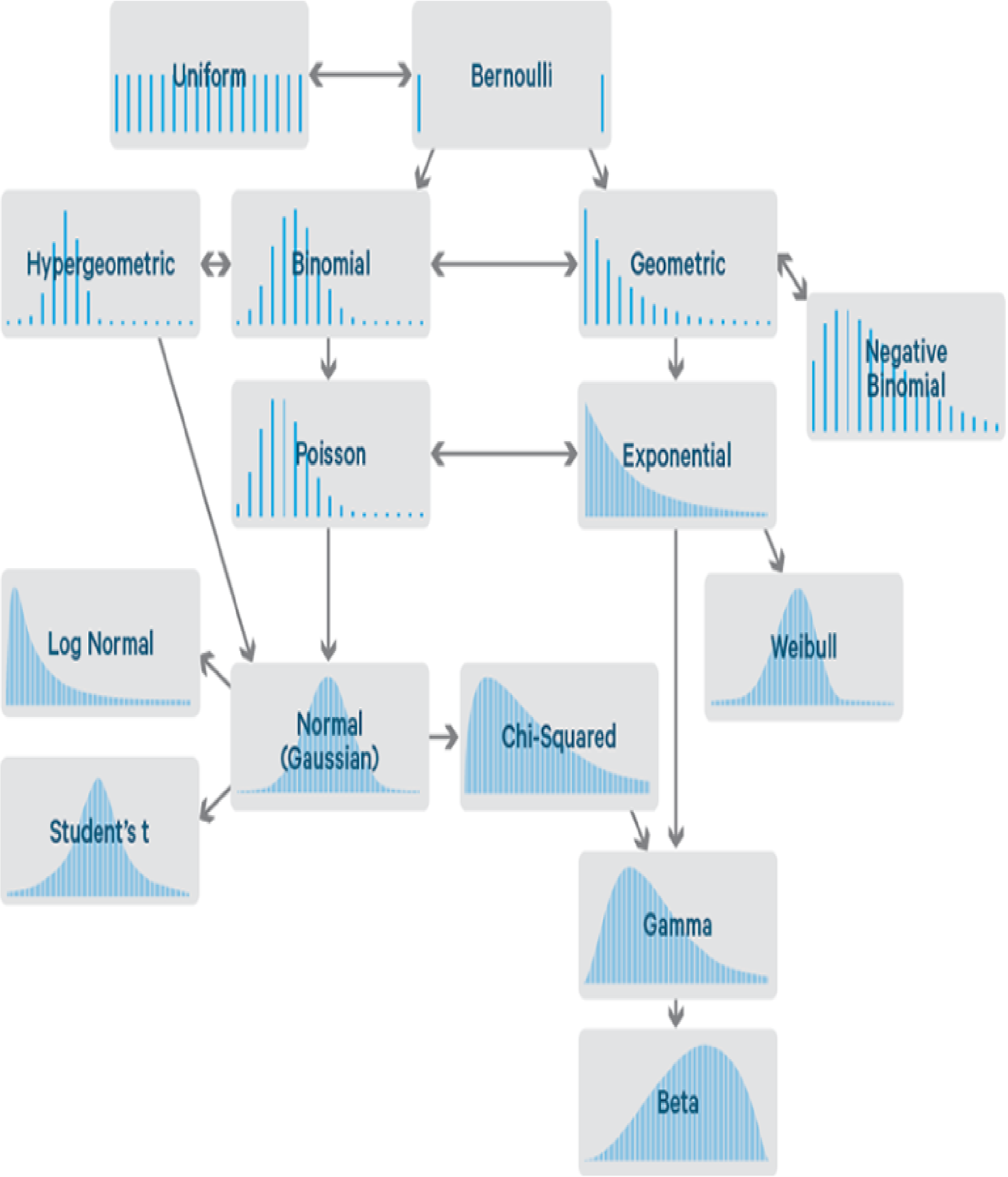
Python, R and other languages

Classification and Regression Models

Data Distribution

The distribution of a statistical data set (or a population) is a listing or function showing all the possible values (or intervals) of the data and how often they occur. When a distribution of categorical data is organized, you see the number or percentage of individuals in each group.

Data distributions are used often in statistics. They are graphical methods of organizing and displaying useful information. There are several types of data distributions



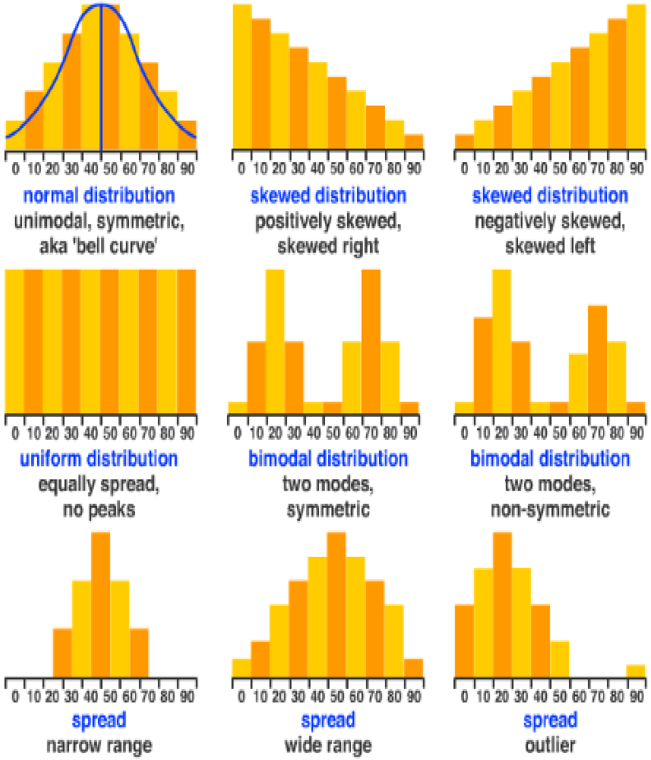
distribution of data

The distribution of data is often graphically represented using histograms and dot plots. Their shape shows the range and spread of the data set.

A normal distribution is a true symmetric distribution of the data values. The mode, median and mean are the same and together in the centre of the distribution.

In a normal distribution histogram the shape of columns form a symmetrical bell shape, often referred to as the 'normal curve' or 'bell curve'.

There are many types of distribution shapes, e.g.



normal distribution
unimodal, symmetric,
aka 'bell curve'

skewed distribution
positively skewed,
skewed right

skewed distribution
negatively skewed,
skewed left

uniform distribution
equally spread,
no peaks

bimodal distribution
two modes,
symmetric

bimodal distribution
two modes,
non-symmetric

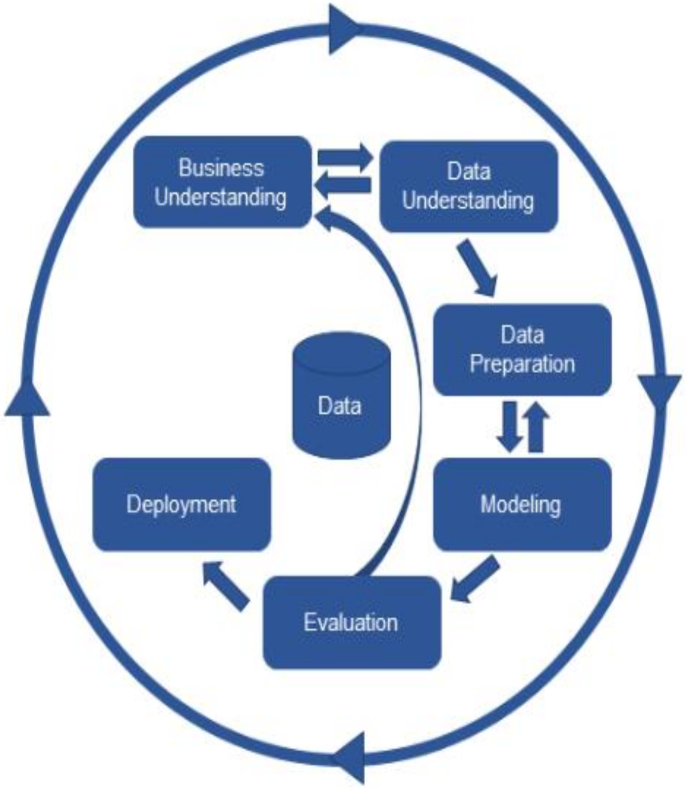
spread
narrow range

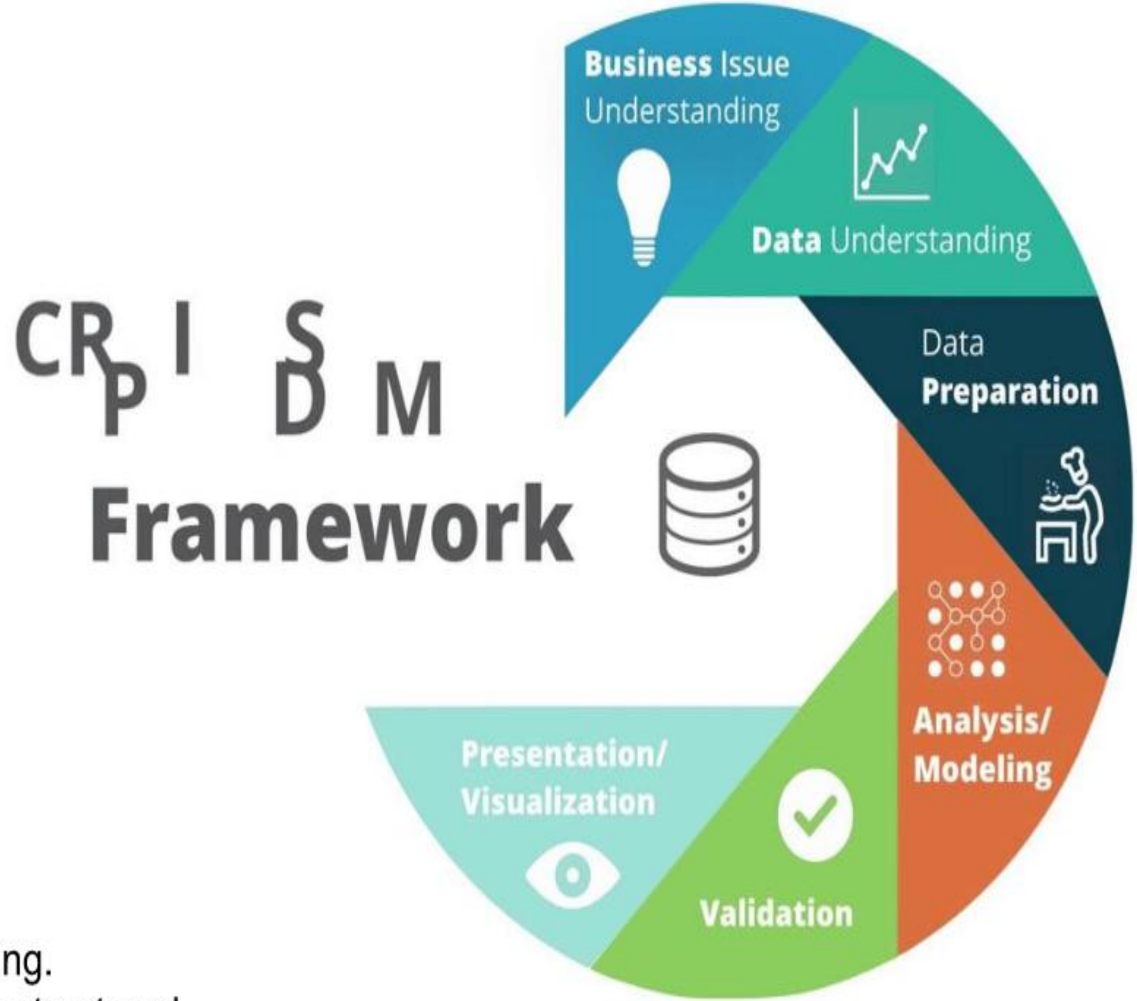
spread
wide range

spread
outlier

© Jenny Eather 2015

CRISP



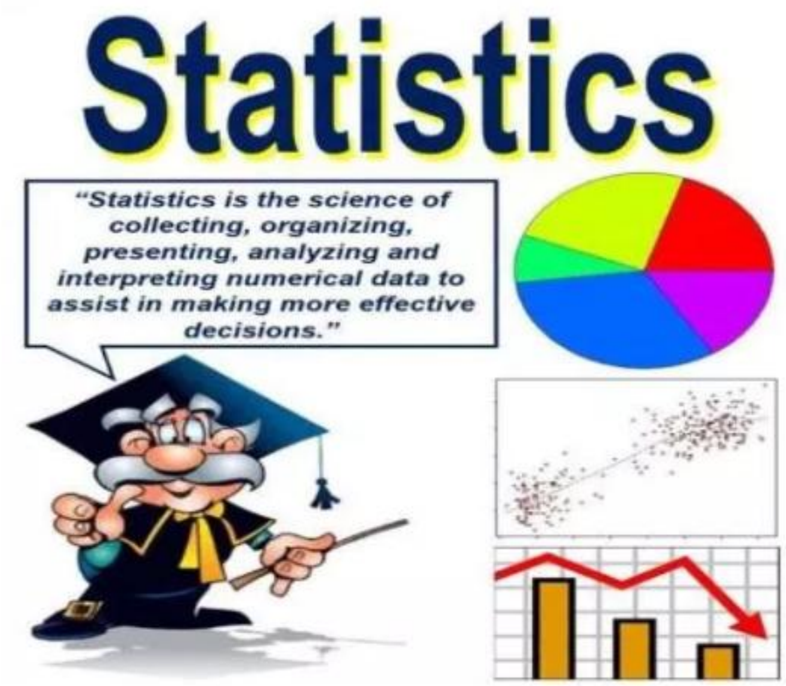


- cross-industry process for data mining.
- CRISP-DM methodology provides a structured approach to planning a data mining project.

Statistics

Statistics is a branch of mathematics working with data collection, organization, analysis, interpretation and presentation. In applying statistics to a scientific, industrial, or social problem.

it is conventional to begin with a statistical population or a statistical model to be studied.


Statistics
"Statistics is the science of collecting, organizing, presenting, analyzing and interpreting numerical data to assist in making more effective decisions."

characteristics of statistics

- Statistics are the aggregates of facts
- Statistics are affected by a number of factors
- Statistics must be reasonably accurate
- Statistics must be collected in a systematic manner
- Collected in a systematic manner for a pre-determined purpose
- Lastly, Statistics should be placed in relation to each other

What is Statistics?

Statistics

The science of collecting, organizing, analyzing, and interpreting data in order to make decisions.

Data

Information coming from observations, counts, measurements, or responses.

Why should you care about statistics?

- Statistics helps you make informed decisions that affect your life.
- Statistics helps the government make decisions that affect many people.

Medical & LifeStyle Decisions

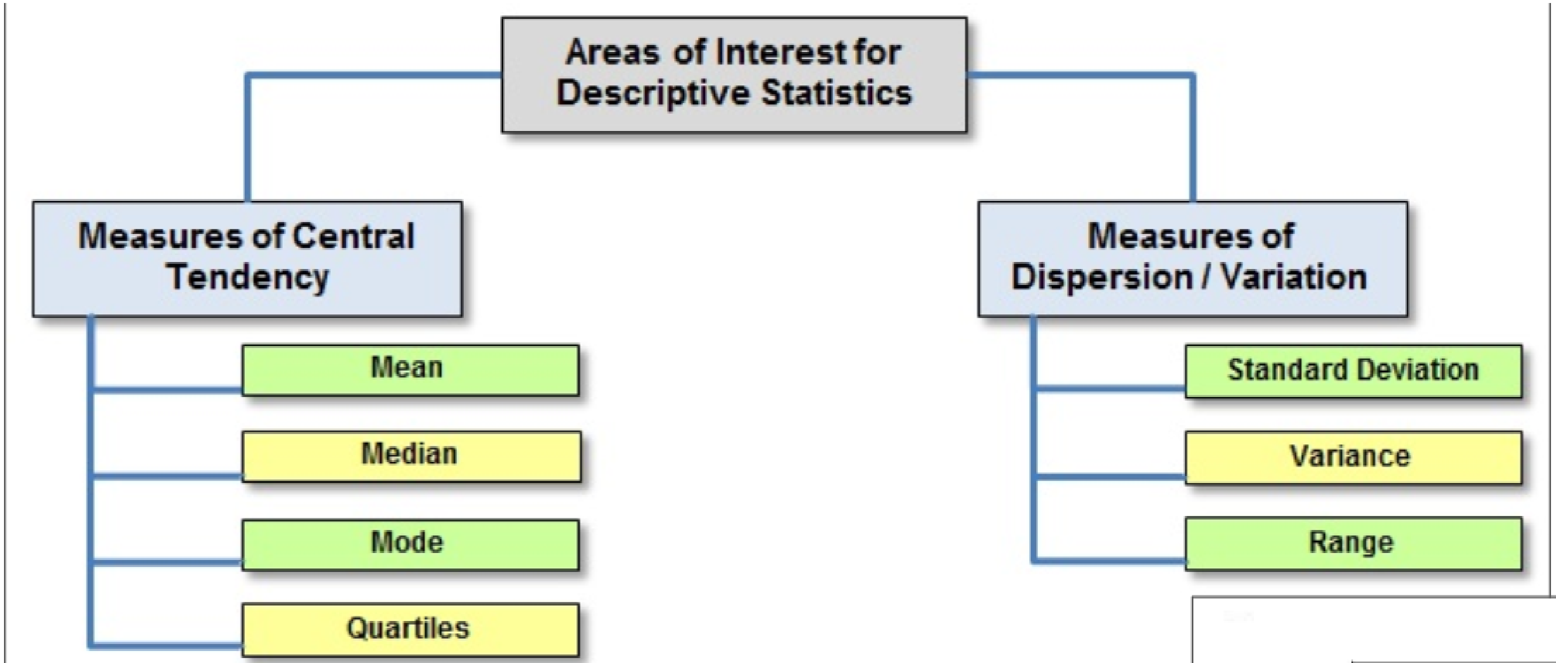
- Vaccines: Polio, Measles, Flu, HPV
- Meds: Blood Pressure, Cholesterol
- Hormone Replacement, Chemo
- Smoking
- Home in City/Country/Suburb
- College/Major
- Invest in Stock Market
- Marriage/Divorce/Children/Adopt

Government Decisions

- Raise Retirement Age (Soc. Sec.)
- Drinking/Driving/Seatbelt Laws
- Mandatory School for children

Common Statistical Data

→Census
→Crime
→Education
→Health/Medical
→Scientific
→Economic



Areas of Interest for Descriptive Statistics

Measures of Central Tendency

Mean

Median

Mode

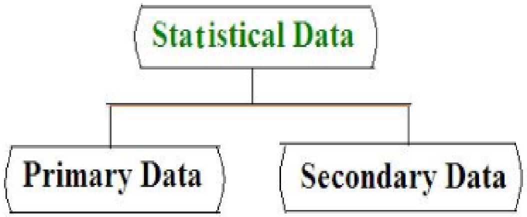
Quartiles

Measures of Dispersion / Variation

Standard Deviation

Variance

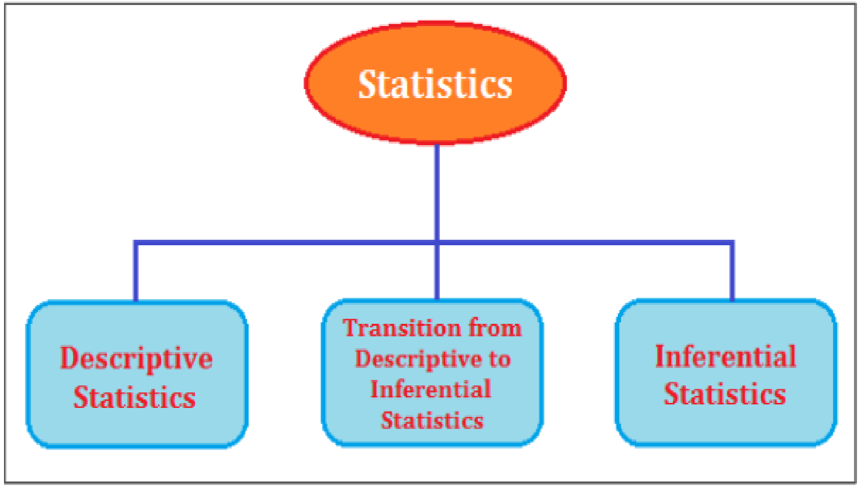
Range



Statistical Data

Primary Data

Secondary Data

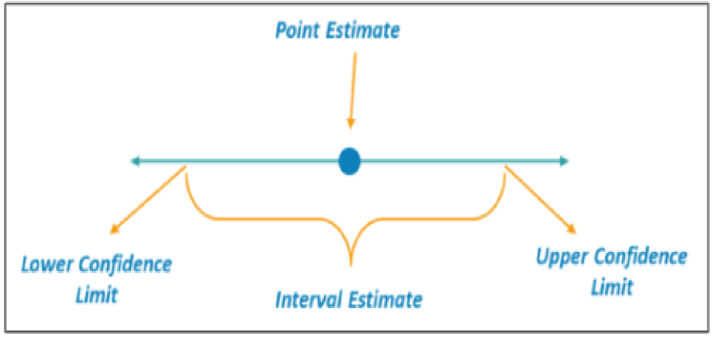


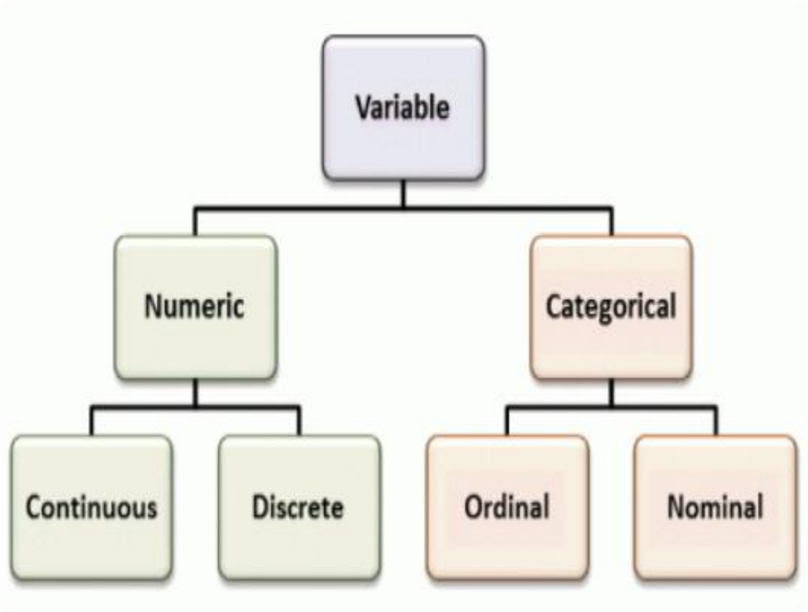
Statistics

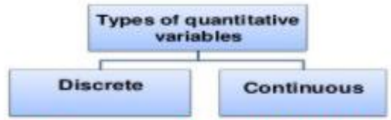
Descriptive Statistics

Transition from Descriptive to Inferential Statistics

Inferential Statistics







A discrete variable

is characterized by gaps or interruptions in the values that it can assume.

For example:

- The number of daily admissions to a general hospital,
- The number of decayed, missing or filled teeth per child in an elementary school.

A continuous variable

can assume any value within a specified relevant interval of values assumed by the variable.

For example:

- Height,
- weight,
- skull circumference.

No matter how close together the observed heights of two people, we can find another person whose height falls somewhere in between.

Makele University: Biostatistics 12

Categorical (data that are counted)

- Nominal
- Ordinal

Quantitative or Numerical (data that are measured)

- Interval
- Ratio

Why is the type of variable important?

The methods used to display, summarize, and analyze data depend on whether the variables are categorical or quantitative.

Types of Variables

A. Qualitative or Attribute variable - the characteristic being studied is *nonnumeric.*

EXAMPLES: Gender, religious affiliation, type of automobile owned, state of birth, eye color are examples.

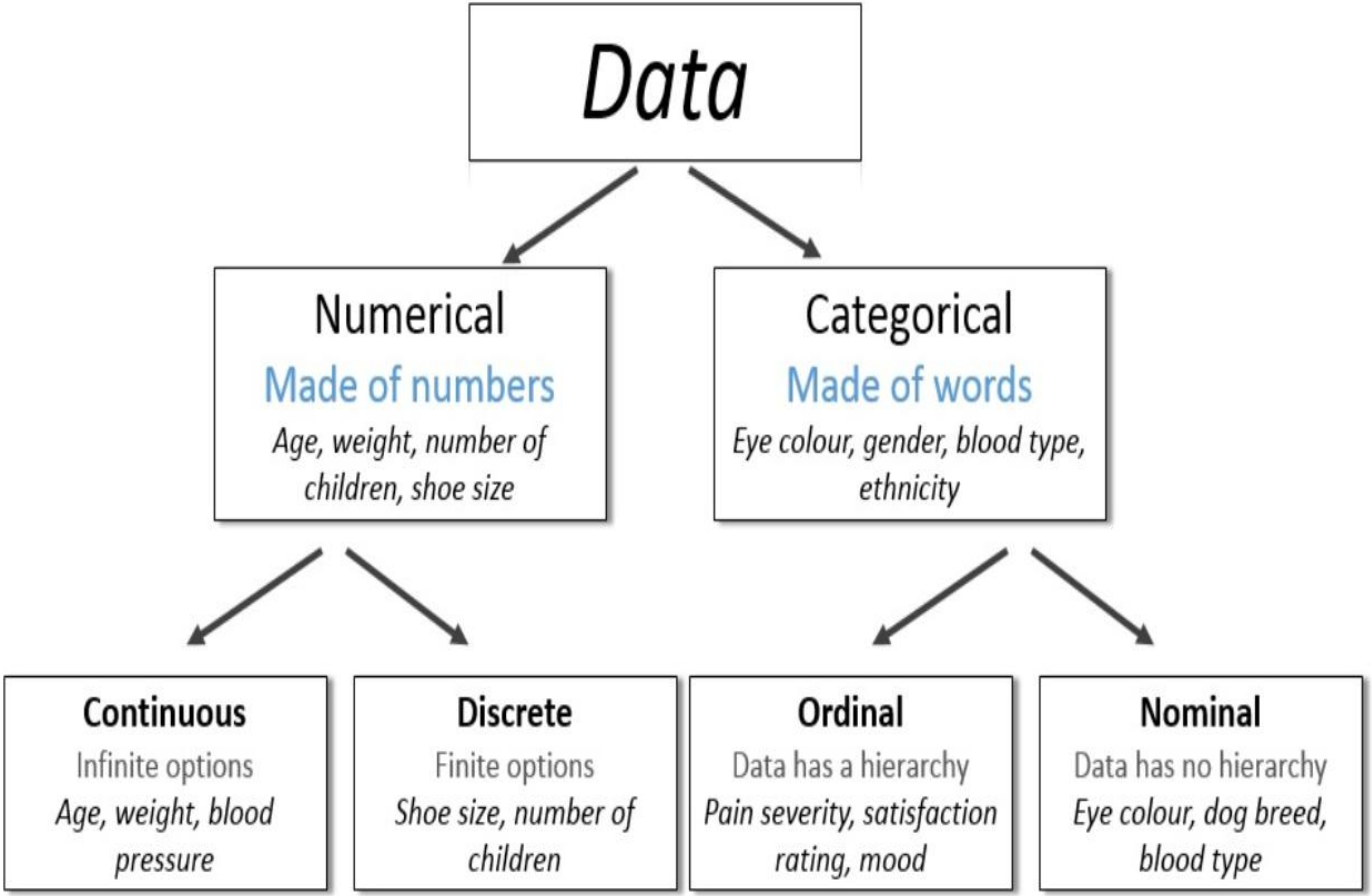
B. Quantitative variable - information is reported *numerically*.

EXAMPLES: balance in your checking account, minutes remaining in class, or number of children in a family.

1. Discrete variables: can only assume certain values and there are usually "gaps" between values (e.g., bedrooms in a house)

2. Continuous variable: can assume any value within a specified range (e.g., tire pressure, height of students in a class

1-7



Data

Numerical

Made of numbers

Age, weight, number of children, shoe size

Categorical

Made of words

Eye colour, gender, blood type, ethnicity

Continuous

Infinite options

Age, weight, blood pressure

Discrete

Finite options

Shoe size, number of children

Ordinal

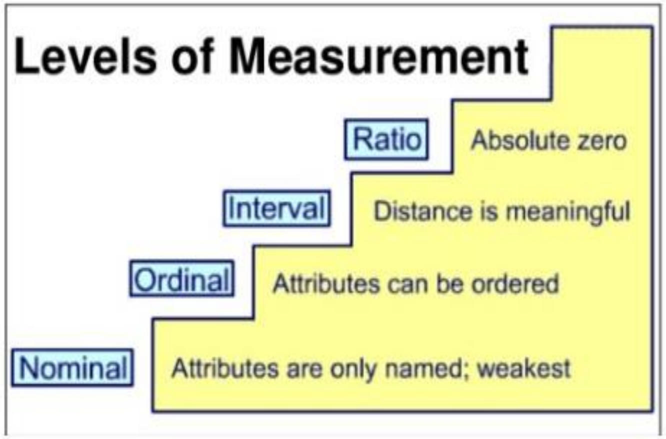
Data has a hierarchy

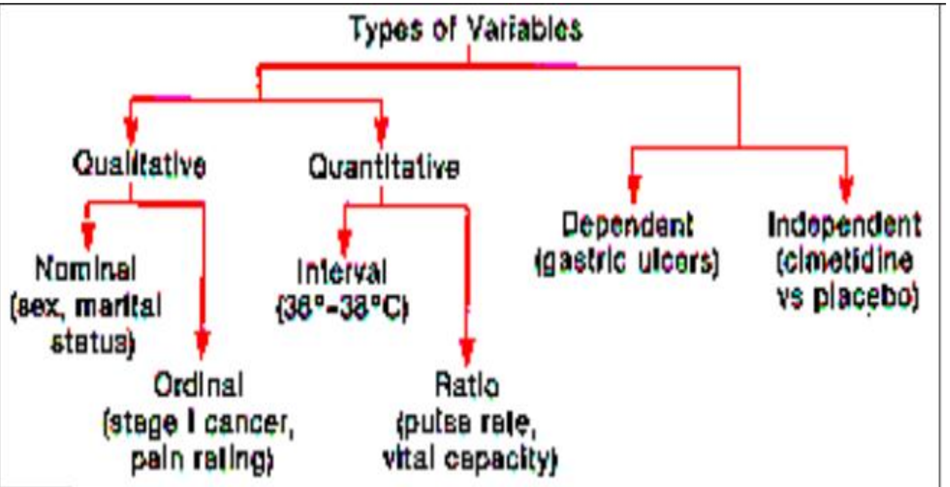
Pain severity, satisfaction rating, mood

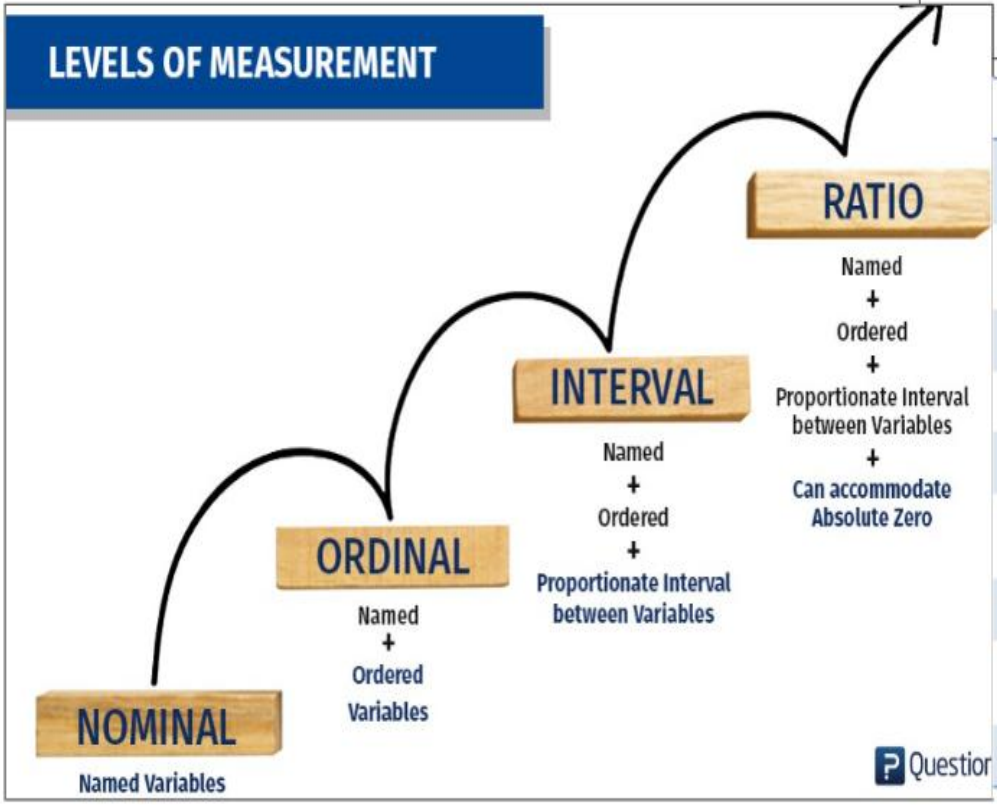
Nominal

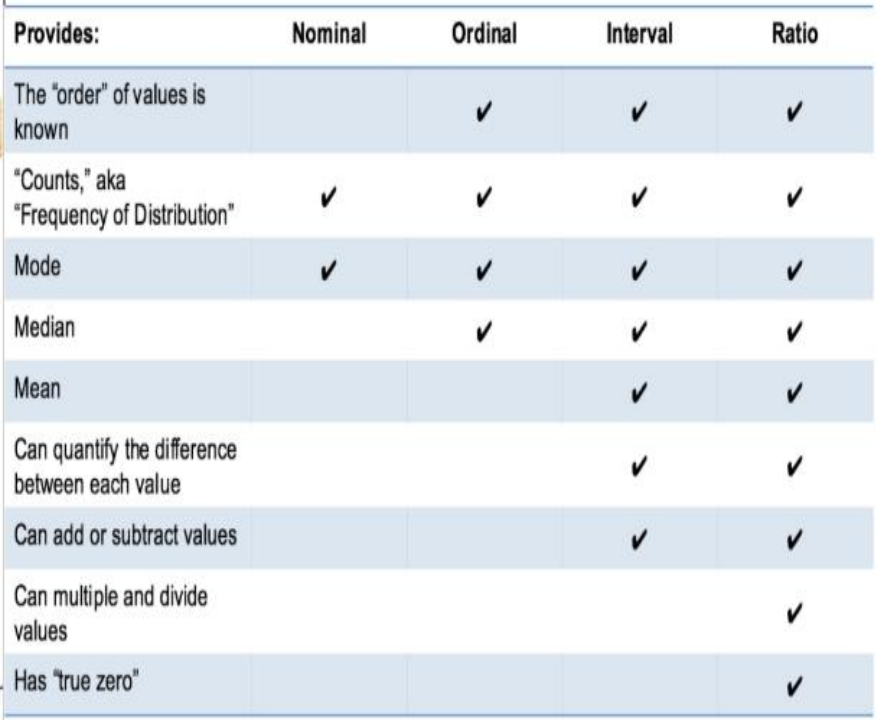
Data has no hierarchy

Eye colour, dog breed, blood type

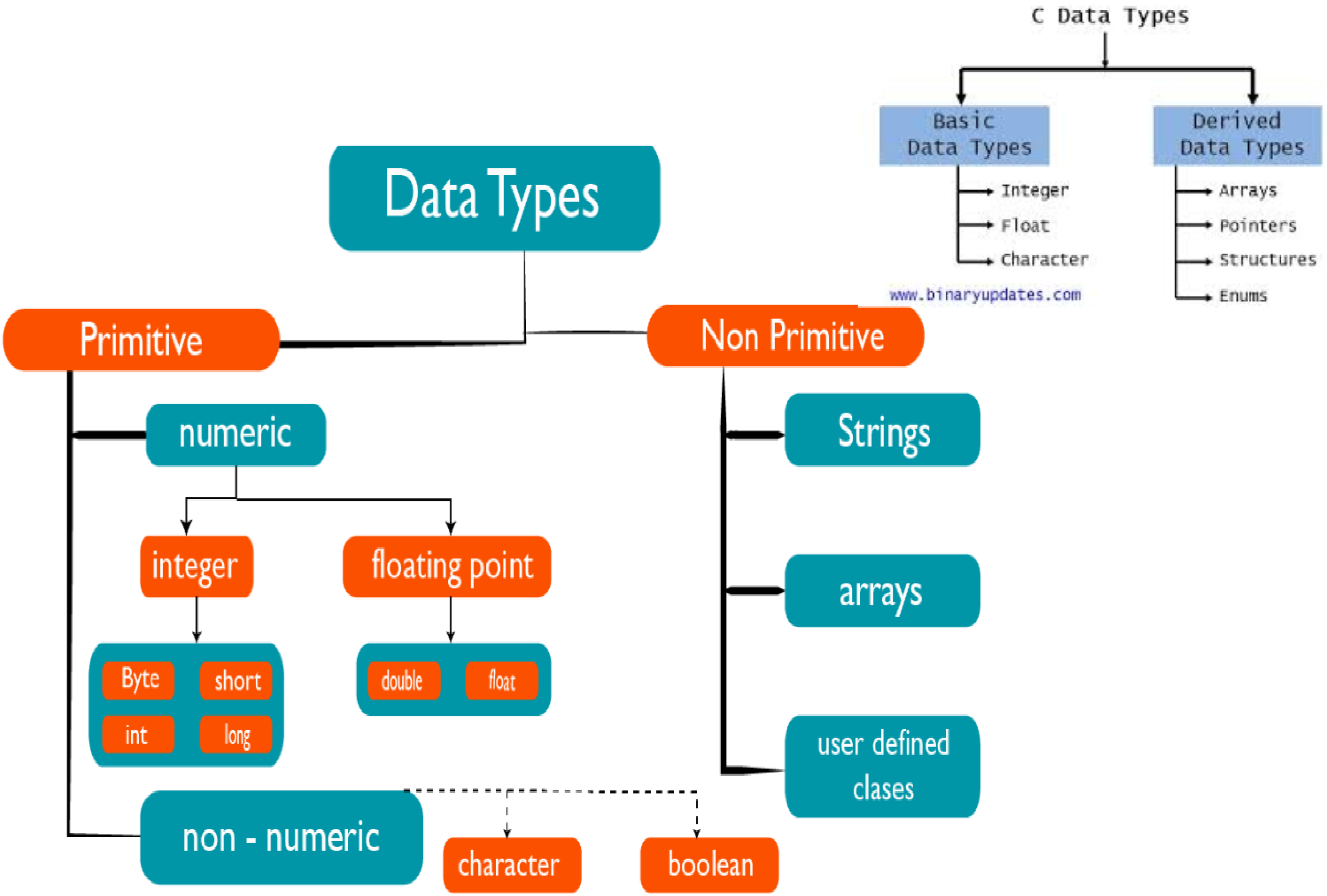


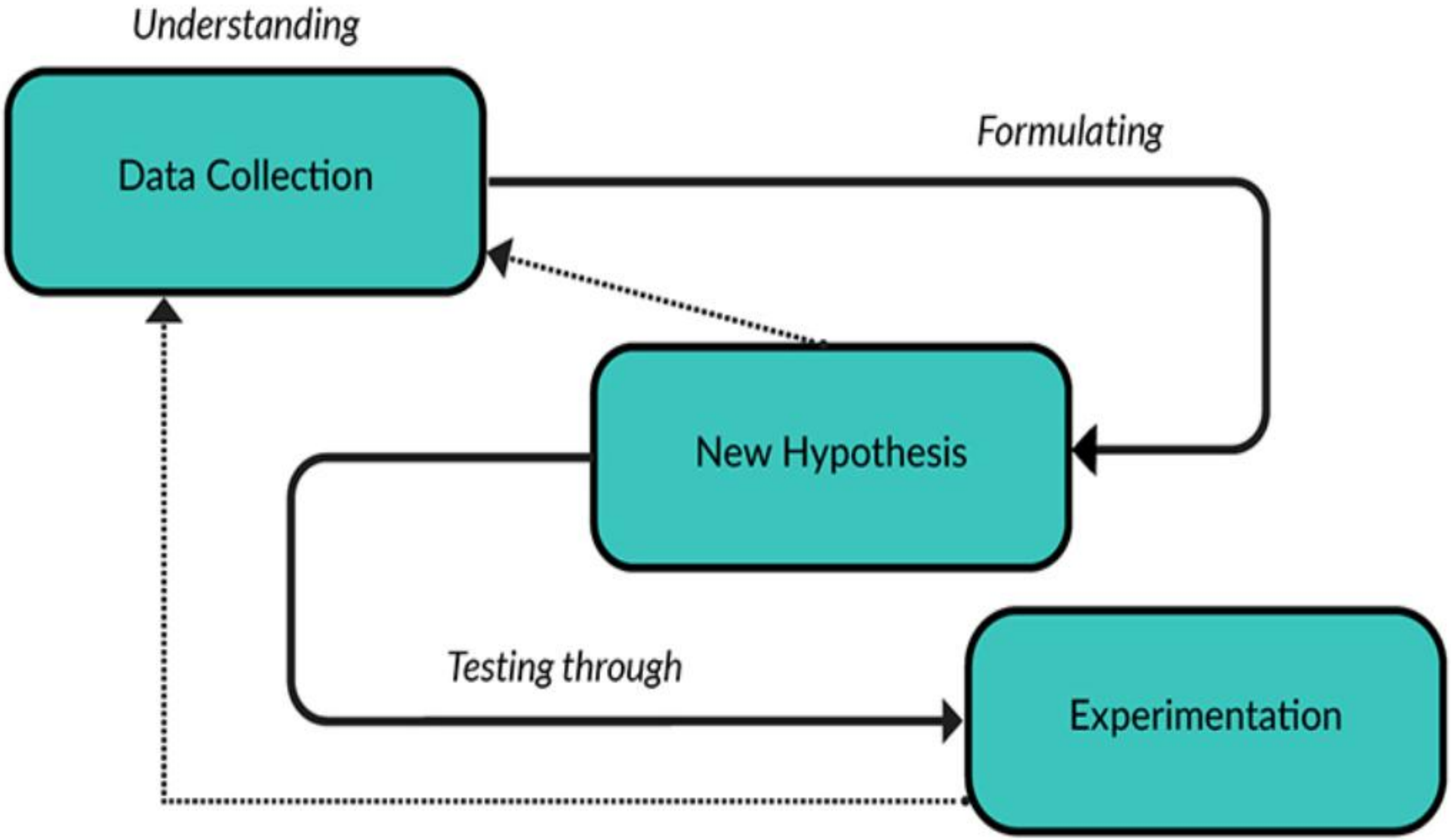

Types of Variables
Qualitative
Quantitative
Nominal (sex, marital status)
Ordinal (stage I cancer, pain rating)
Interval (36°-38°C)
Ratio (pulse rate, vital capacity)
Dependent (gastric ulcers)
Independent (cimetidine vs placebo)





Provides:	Nominal	Ordinal	Interval	Ratio
The "order" of values is known		✓	✓	✓
"Counts," aka "Frequency of Distribution"	✓	✓	✓	✓
Mode	✓	✓	✓	✓
Median		✓	✓	✓
Mean			✓	✓
Can quantify the difference between each value			✓	✓
Can add or subtract values			✓	✓
Can multiple and divide values				✓
Has "true zero"				✓


Data Types
Primitive
Non Primitive
numeric
integer
floating point
Byte
short
int
long
double
float
non - numeric
character
boolean
Strings
arrays
user defined clases
C Data Types
Basic Data Types
Derived Data Types
Integer
Float
Character
Arrays
Pointers
Structures
Enums
www.binaryupdates.com



Understanding

Data Collection

Formulating

New Hypothesis

Testing through

Experimentation