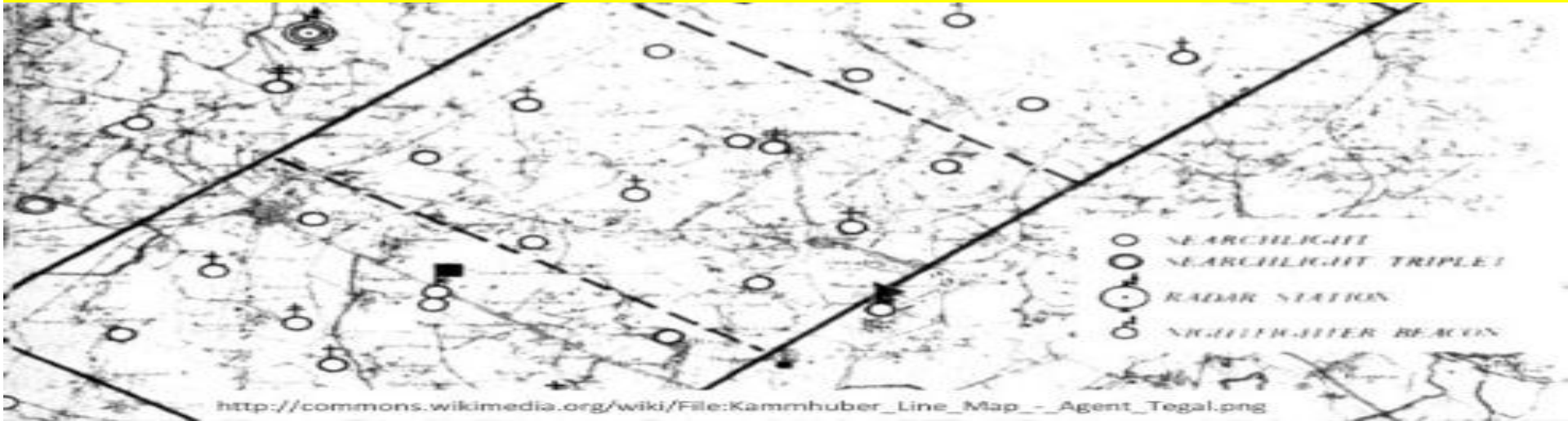
Prof. Rajeev Pandey
Department of Statistics
University of Lucknow
Lucknow

The Evolution of Data Analytics

How to grok data with machines and keep with changing times

The origins (40s, 50s, 60s)

Operation Research during World War II
&
First Predictive Weather Model



http://commons.wikimedia.org/wiki/File:Kammhuber_Line_Map_-_Agent_Tegal.png

The origins (40s, 50s, 60s)

- Operational Research
- Collision loss *vs* Anti-Aircraft loss
- Optimization (Statistical) problems
- Scheduling and resource allocation

Analytics goes Mainstream (70s, 80s)

- The Relational Database is born!

1972: E.F. Codd relational database model, normalization:
(free from insertion, deletion and update anomalies)

1978: Peter Chen, The entity-relationship model

Analytics goes Mainstream (70s, 80s)

- 1982: IBM DB2, Oracle v3, Sybase (SAP)
- 1986: First standardized SQL

- 1987: Commercial use of Decision Support Systems:
  Texas Air Traffic Expert system

Importance of Data Analytics

Data analytics should be a first-class citizen

Data analytics team should
be a key stakeholder

Everyone should ‘own’ the data

DATA ANALYTICS refers to qualitative techniques and processes used to enhance productivity and business gain. **DATA ANALYTICS is not:**

Data Science
Big Data
Artificial Intelligence (AI)
Or Machine Learning

Data Science is a concept that unifies statistics, data analysis and their related methods in order to understand and analyze phenomena with data.

It employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science.

5 Characteristics of Data-driven Organization

1. Leadership
2. Liquidity
3. Usage
4. Access
5. Protection

MANAGEMENT

Leadership

Top-down commitment

High Quality and High Velocity Decision Making

Liquidity

Breaking down silos of data

Structural – applications are optimized for their main function.

Not to encourage data sharing

Usage

How data is used to make decisions

Without data you're just another person with an opinion

Access

Who has access and how are they using it

Protection

How data is stored and shared in a secure manner

Introduction to BIG DATA

- Big Data may well be the Next Big Thing in the IT World.

- **Big data burst upon the scene in the first decade of the 21st century.**

- The first organizations to embrace it were online and startup firms. Firms like Google, eBay, LinkedIn, and Facebook were built around big data from the beginning.

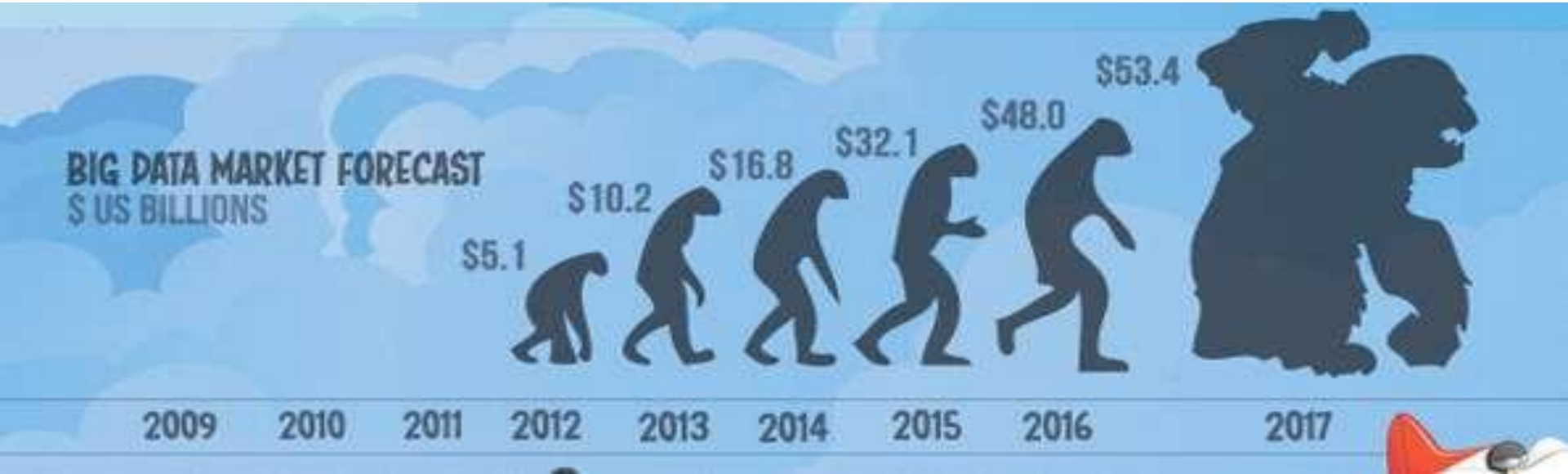
- Like many new information technologies, big data can bring about dramatic cost reductions, substantial improvements in the time required to perform a computing task, or new product and service offerings.

What is BIG DATA

- Walmart handles more than 1 million customer transactions every hour.

- Facebook handles 40 billion photos from its user base.

- Decoding the human genome originally took 10years to process; now it can be achieved in one week.

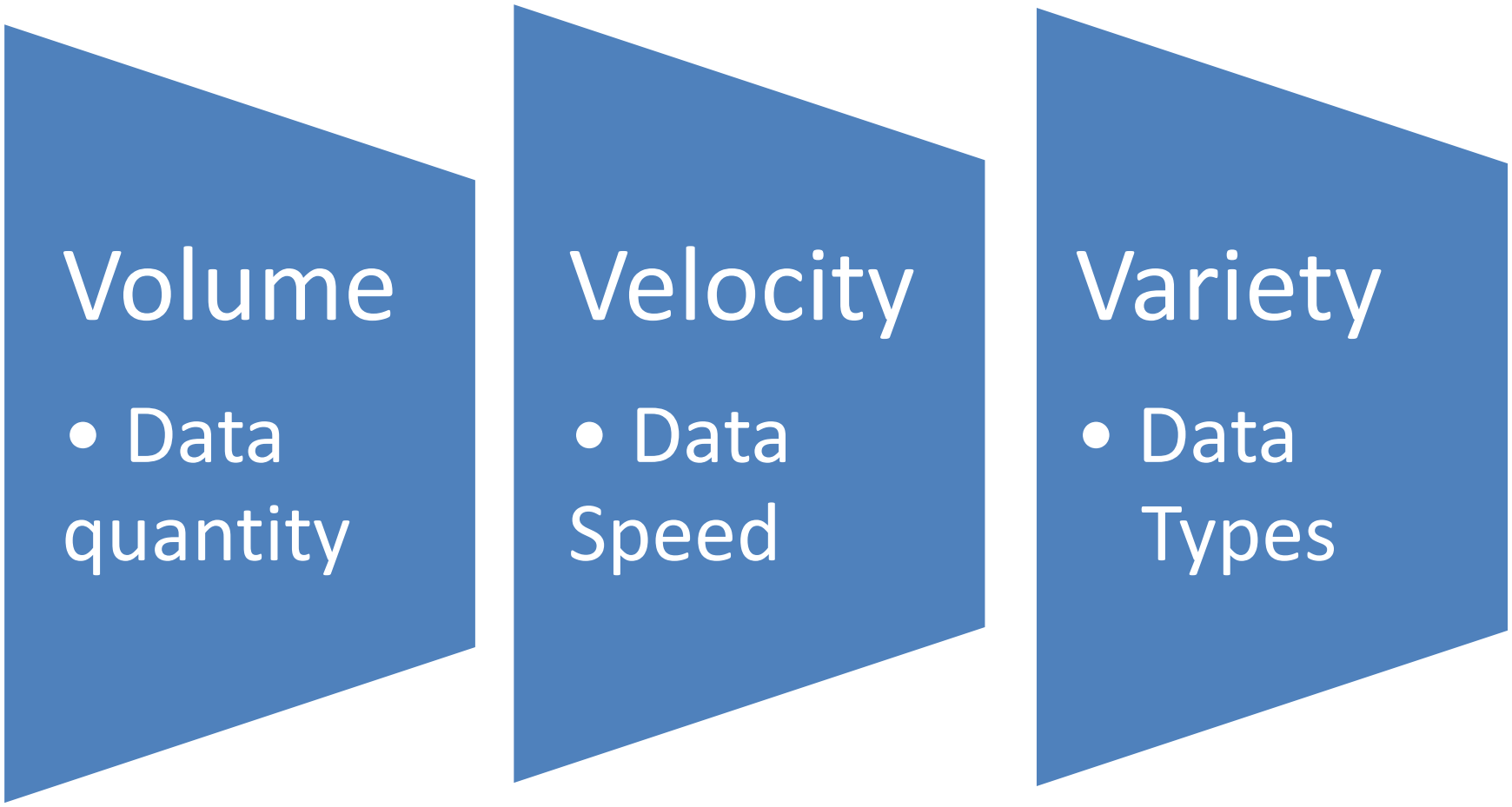


Why Study Statistics- BIG DATA?

- ***Communication***
  - Understanding the language of statistician who facilitates communication and improves problem solving.

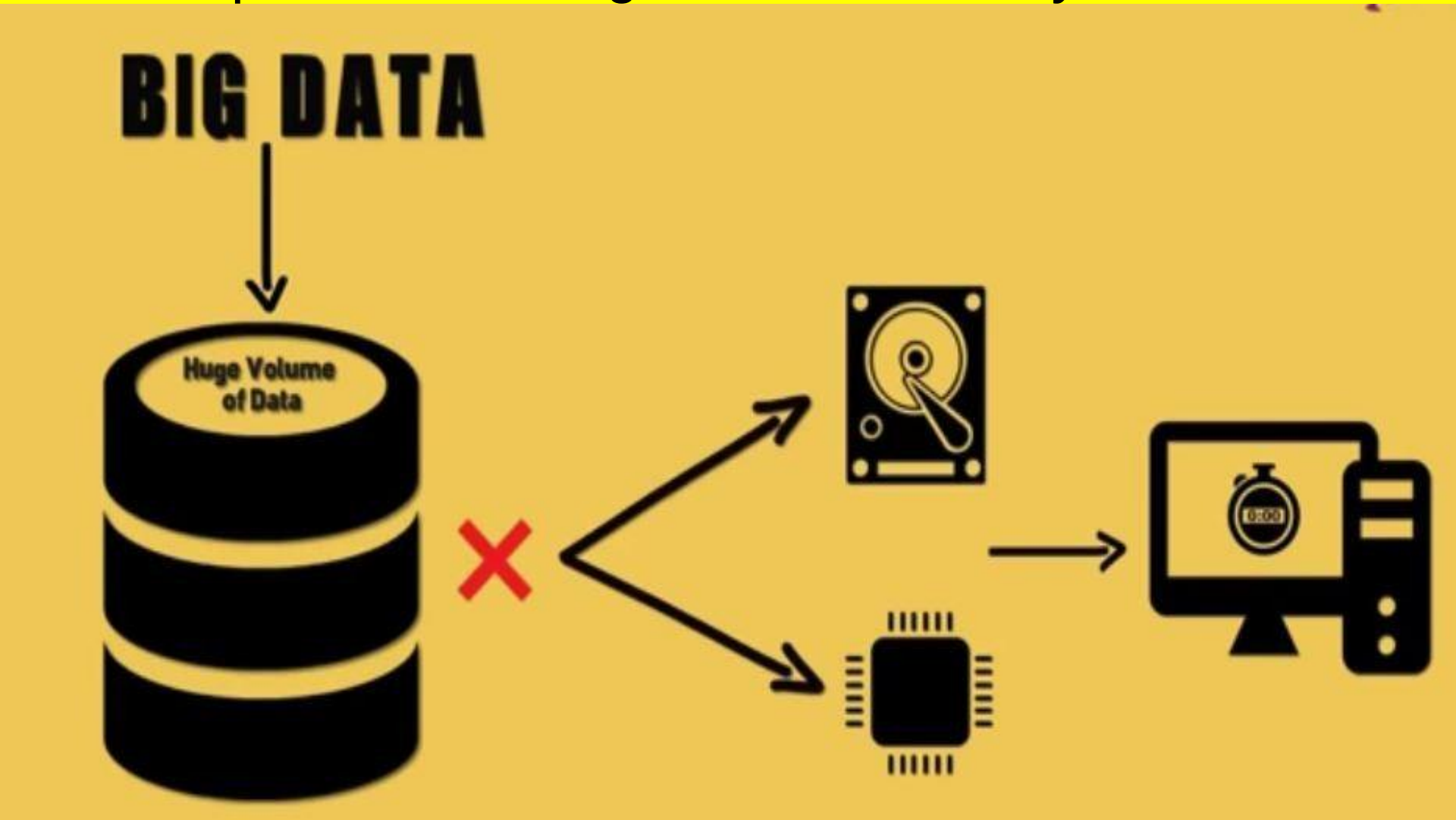
- ***Computer Skills***
  - The use of spreadsheets for data analysis and word processors or presentation software for reports improves upon your existing skills.

Three Chief Characteristics of Big Data V3s



What is actually Big Data?

Big data-: So large data that it becomes difficult to process it using the traditional system.

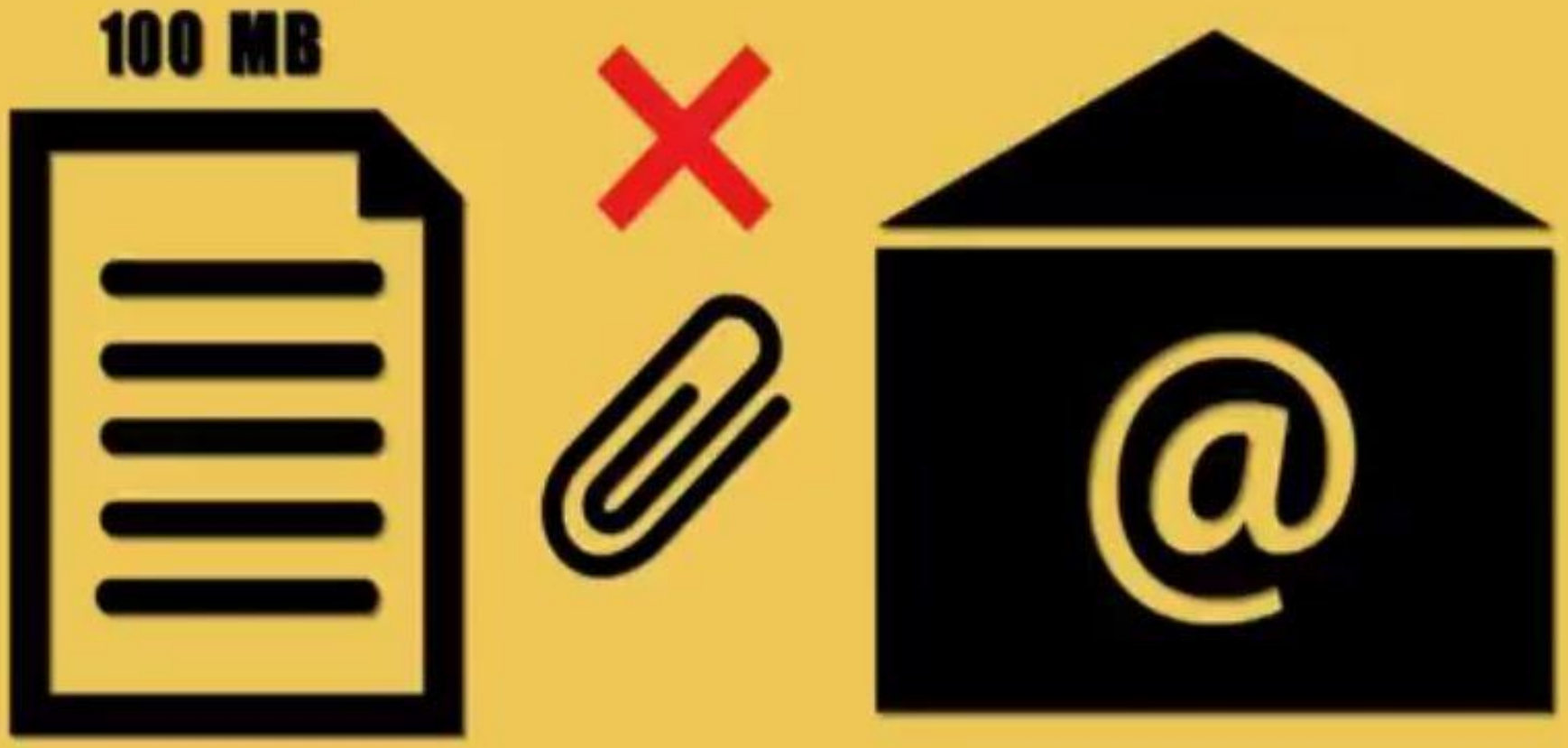


SIZES OF DATA

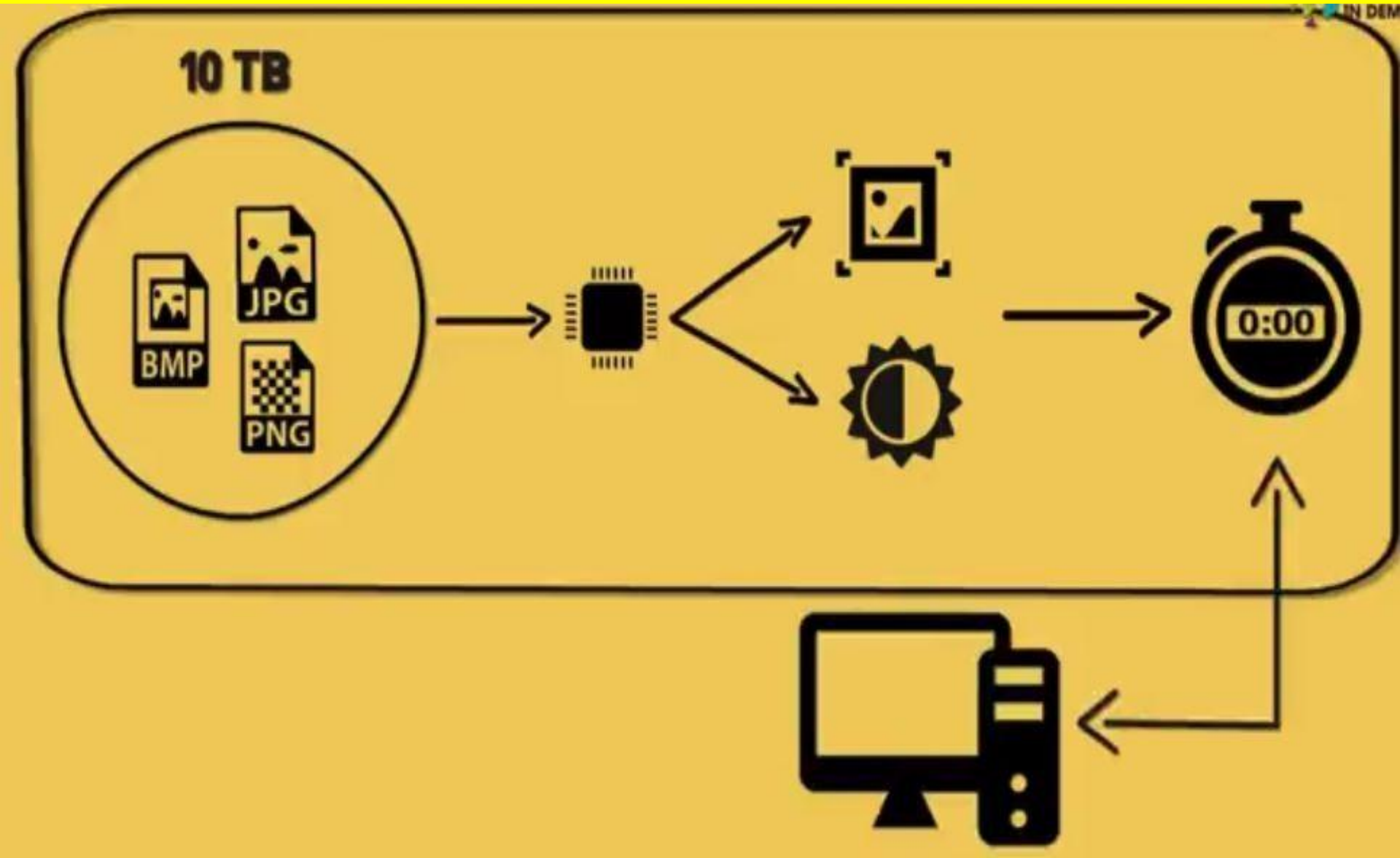
NAME	SYMBOL	VALUE
Kilobyte	KB	10^3
Megabyte	MB	10^6
Terabyte	TB	10^12
Petabyte	PB	10^15
Exabyte	EB	10^18
Zettabyte	ZB	10^21
Yottabyte	YB	10^24

Example...

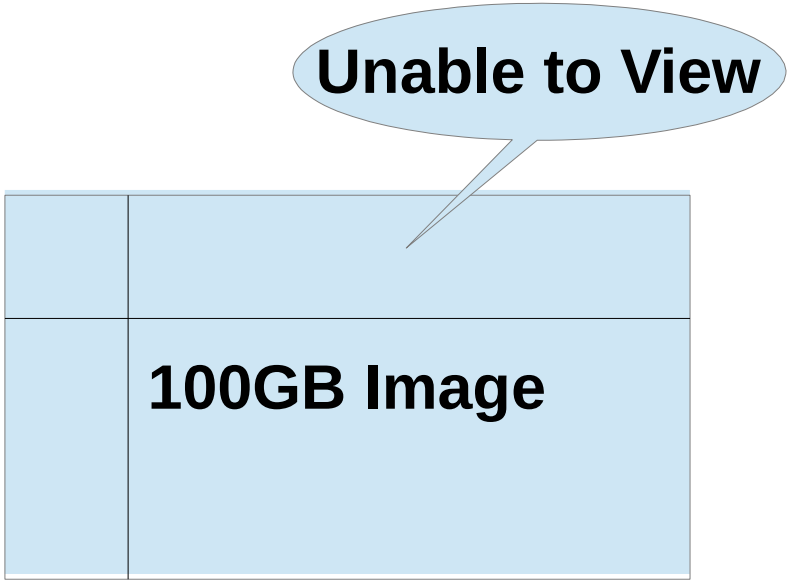
Do you ever tried opening 0.5GB of file on your machine?

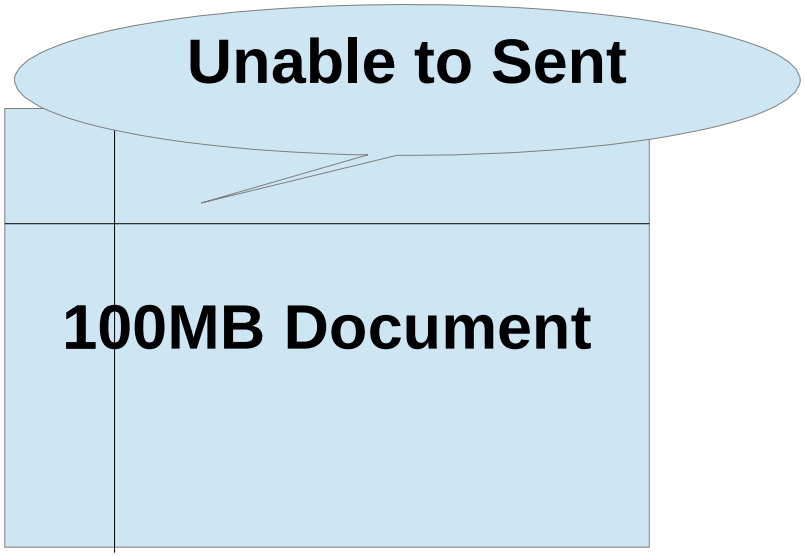


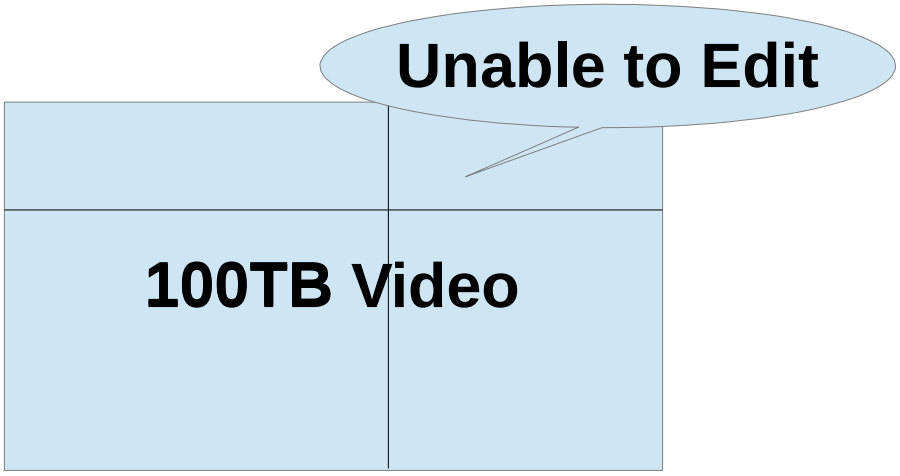
Its difficult to edit 10TB file in limited time in traditional system

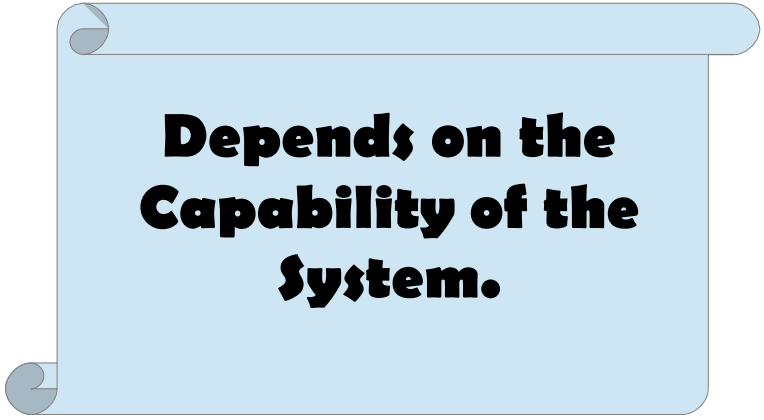


Difficult to process by the Traditional System

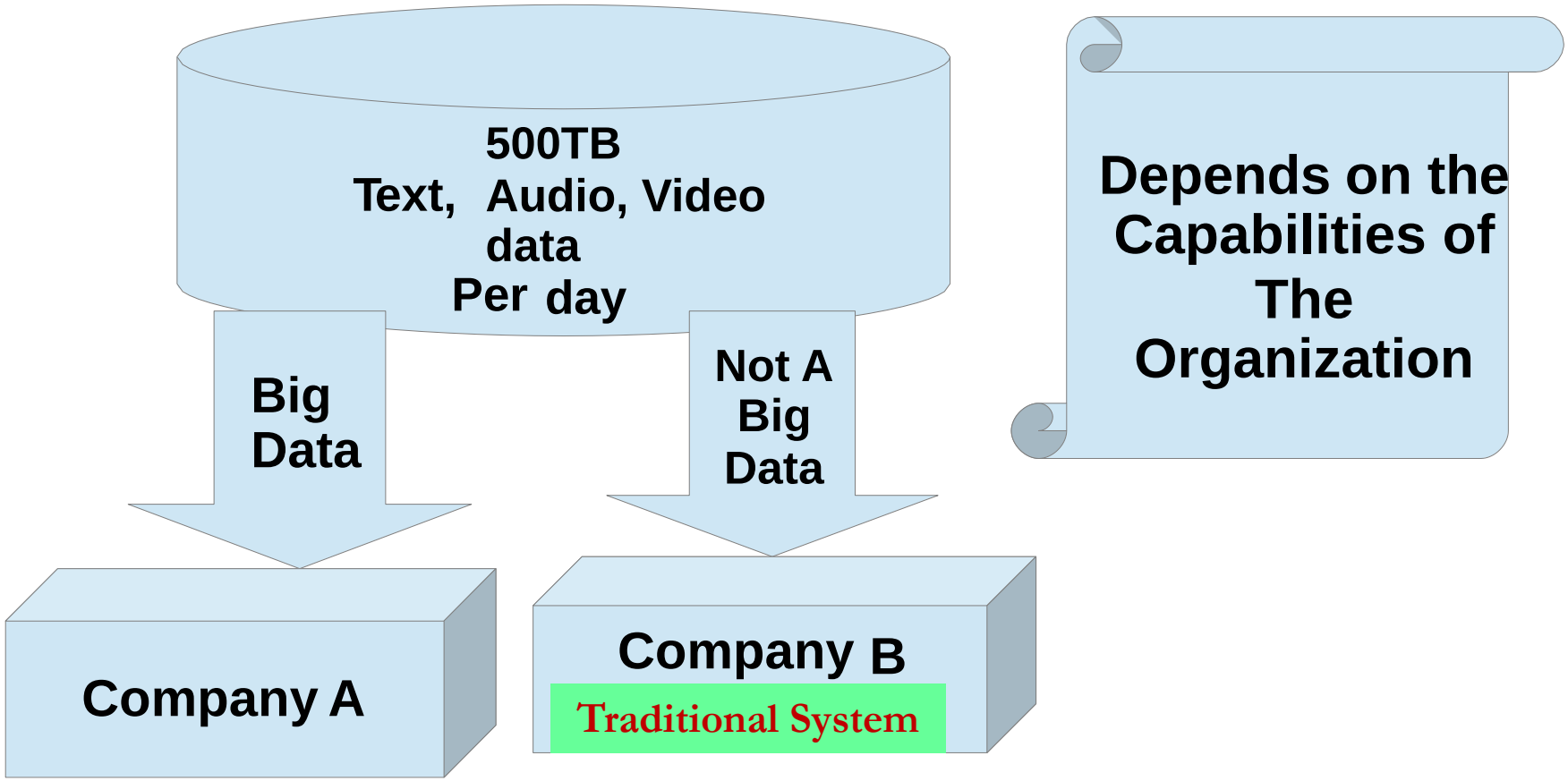








Organization Specific



AREAS OF CHALLENGES

- Capture
- Storage
- Curation
- Analysis
- Search
- Sharing
- Tranfer
- Visualization

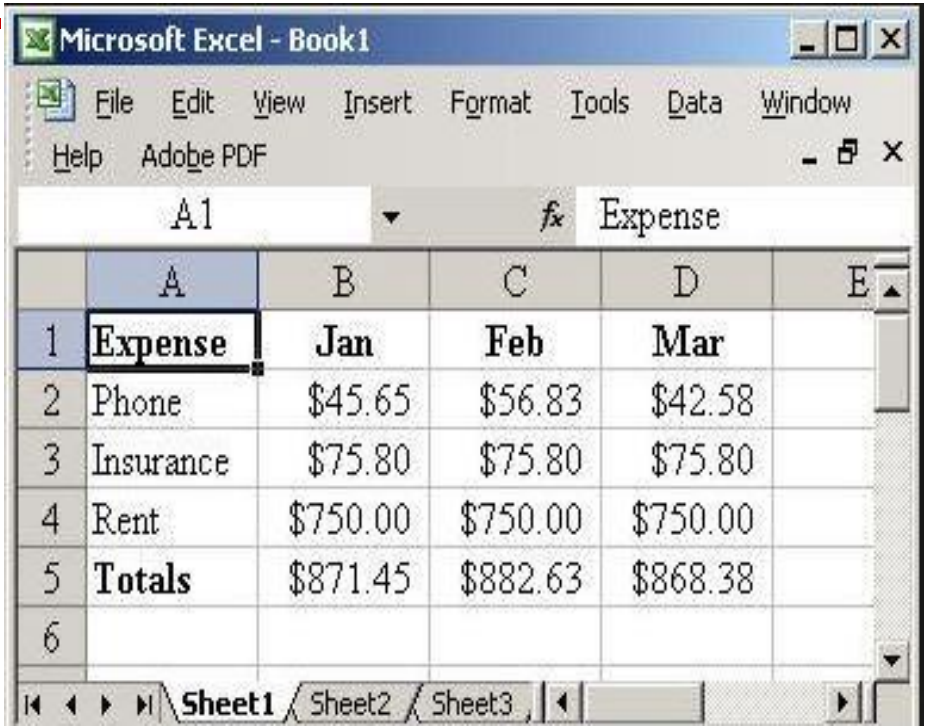
CLASSIFICATION OF BIG DATA

1. **Structured Data:**

It refers to data that has a defined length and format for big data

Ex. numbers, dates, and groups of words and numbers called strings.

It's usually stored in a database.



2. Unstructured Data

- No fields
- Massive data ex. Newspaper
- 

Applications

Music(Audio)

Movie(vedio)

X-Rays

Pictures

3. Semi-Structured Data

The data which do not have a proper formate atteched to it.

Ex.

- Data within an email
- Data in Doc File

Why do we need this?

- How do you like a new movie?
- In Election exit poll
- Chess Board
- Facebook purchase WhatsApp why?

Characteristics of Big Data

1. Velocity
2. Volume
3. Variety
4. Value
5. Veracity
6. Variability
7. Visualization

VELOCITY

- The speed of generation of data.
- Perhaps action being taken upon.
- The highest velocity data normally streams directly into memory versus being written to disk.
- Some Internet of Things (IoT) require real-time evaluation and action.

EXAMPLES OF VELOCITY

- Almost 2,5 million queries on Google are performed.
- Around 20 million photos are viewed.
- Every minute we upload 100 hours of video on Youtube.
- every minute over 200 million emails are sent.
- 300,000 tweets are sent per minute

VOLUME

- The amount of data generat every second.
- Here we are talking about Zettabyte or more.
- It is the task of big data to convert such into Hadoop data valuable information.
- Data is generated by machines, networks and human interaction on systems like social media.
- the volume of data to be analyzed is massive.

Example of Volume...1

Airbus

- **Airbus generates 10TB every 30 minutes**
- **About 640TB is generated in one flight**

Example of Volume...2

- Self-driving cars will generate 2 Petabyte of data every year.
- From now on, the amount of data in the world will double every two years.
- By 2020, we will have 50 times the amount of data as that we had in 2011.

VARIETY

- Refers to the different types of data we can now use.
- In past the data was structured that fitted in columns and rows.
  - Stored in Database
  - Spread sheets
- But now the data is unstructured that are difficult to storing, analysing,mining.
  - Email, photo, audio
  - monitoring devices, PDFs

VALUE...1

- Having access to big data is no good unless we can turn it into value.
- Companies are staring to generate amazing value from their big data.

  -> Discovering a consumer preference or sentiment,

  -> To making a relevant offer by location

  -> Identifying a piece of equipment that is about to fail.

Value....2

- The real big data challenge is a human one which is

  -> learning to ask the right questions,

  -> recognizing patterns

  -> making informed assumptions

  -> predicting behavior.

VERACITY

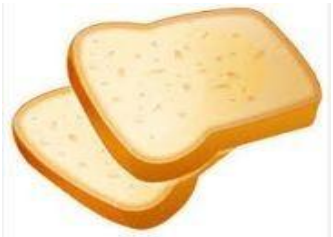
- Are the results meaningful for the given problem space?
- it's about data quality and understandability.
- Especially in automated decision-making, where no human is involved anymore, you need to be sure that both the data and the analyses are correct.

VARIABILITY..1

Dissect an answer into its meaning and to figure out what the right question was.

Variability is often confused with variety

-> Say you have bakery that sells 10 different breads. That is variety.



-> Now imagine you go to that bakery three days in a row and every day you buy the same type of bread but each day it tastes and smells different. That is variability.

Variability means that the meaning is changing.

Visualization

- This is the hard part of big data.
- Making all that vast amount of data comprehensible a in manner that is easy to understand and read.
- Visualizations of course do not mean ordinary graphs or pie charts.
- They mean complex graphs that can include many variables of data while still remaining understandable and readable.

Big Data Sources

- Users
- Application
- System
- Sensors

Large and growing files

Data Generating Points

- **Smart Phones**
- 5 billion camera phones are there in the world
- Most of them have location awareness(GPS)
- By the end of year 2013, the number of smart phone was exceed the number of PC's

Internet

- 2 billion people using internet
- By the end of 2015, cisco traffic internet traffic 4.8ZB per year.

Emails:

- 300 billion email send every day

Blogs:

- There are 200 million entries on the web

Social Media

Facebook:

- 34K likes every minute
- It deals with 3-4 PB of data each day
- There are 1 billion active user

Twitter:

- It generates 12TB of data daily
- 200million user generates 230million tweets daily

Google:

- It perform 2million search every minute
- It deals with 20PB of data each day

Youtube:

- 2.9 billion vedio hours vedio watched per month

Limitations of Traditional System

Data Warehouse

- Cost

- Fixed Schema of RDBMS

- Saving huge file and accessing them

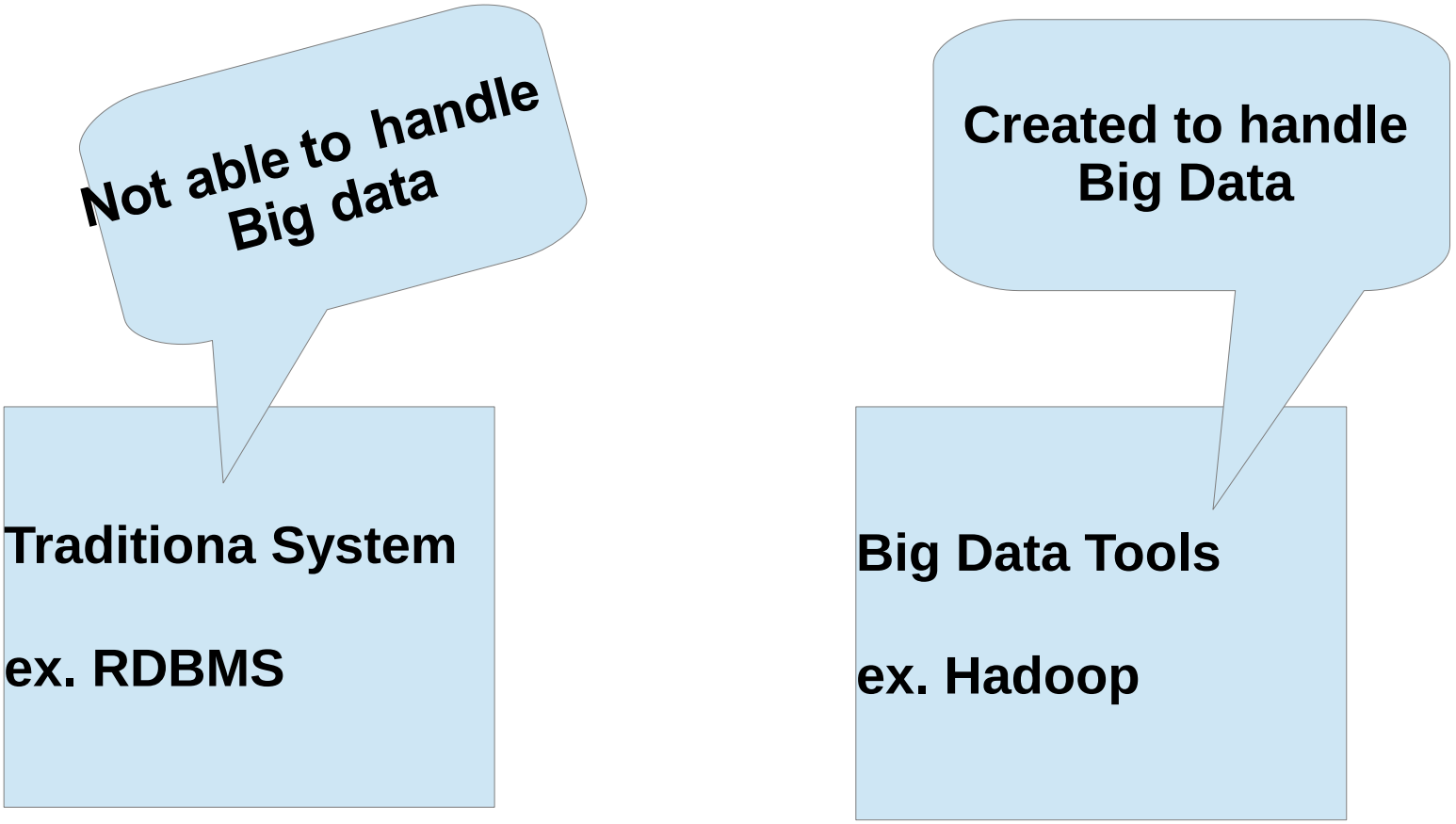
- Perform analysis

- Time to do all this task

Appilcations of big data

- Companies gaining edge by collecting , analyzing, and understanding information
- Governments forecasting events and taking proactive actions
  - Like spred of decieses

Tools for handling big data



Endorsements

- Banking- ICICI
- Soft Drinks- Pepsi, Mirinda
- Batteries- Eveready
- Paints- Nerolac
- Chocolates- Cadbury
- Automobiles- Maruti Suzuki (Versa)
- Writing Instruments- Parker Pens
- Apparel- Reid & Tailor
- Diet Supplements- Dabur
- Personal Care- Emami
- Real Estate- Sahara City Homes , Binani Cement

Statistical Description of Data

- Statistics describes a numeric set of data by its
  - Center
  - Variability
  - Shape
- Statistics describes a categorical set of data by
  - Frequency, percentage or proportion of each category

Data Presentation

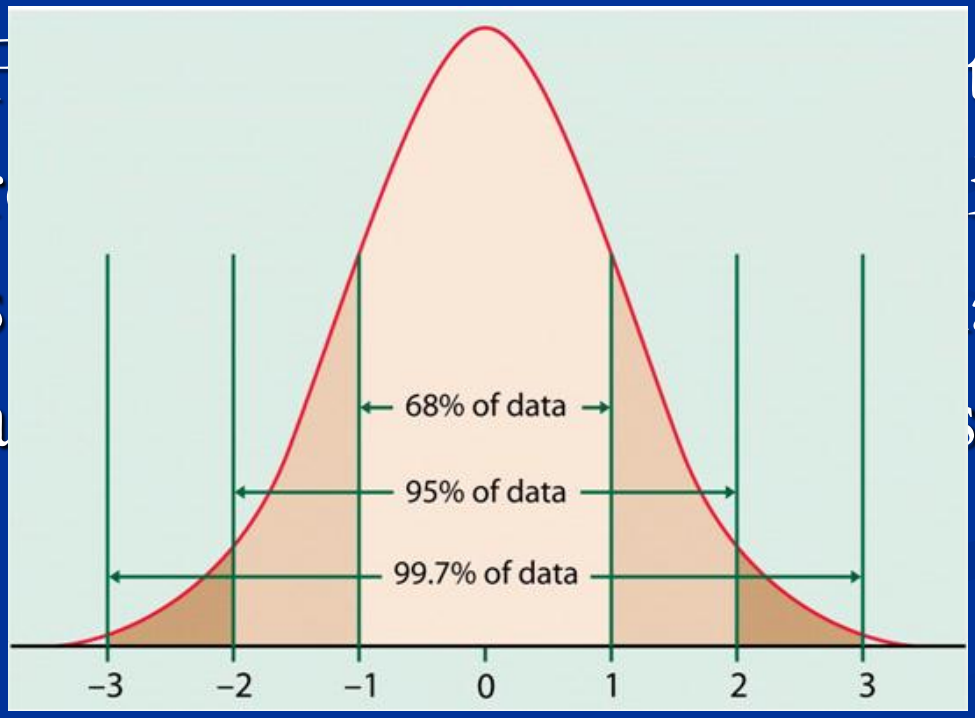
Two types of statistical presentation of data - graphical and numerical.

Graphical Presentation: We look for the overall pattern and for striking deviations from that pattern. Over all pattern usually described by shape, center, and spread of the data. An individual value that falls outside the overall pattern is called an ***outlier.***

- **Bar diagram and Pie charts are used for categorical variables.**
- **Histogram, stem and leaf and Box-plot are used for numerical variable.**

Role of Normality

- Many statistical methods require that the numeric variables we are working with have an approximate **normal distribution**.
- F ts, and
  r ire in some
  s ables are
  a tributed.



Standardized normal distribution with empirical rule percentages.

Tools for Assessing Normality

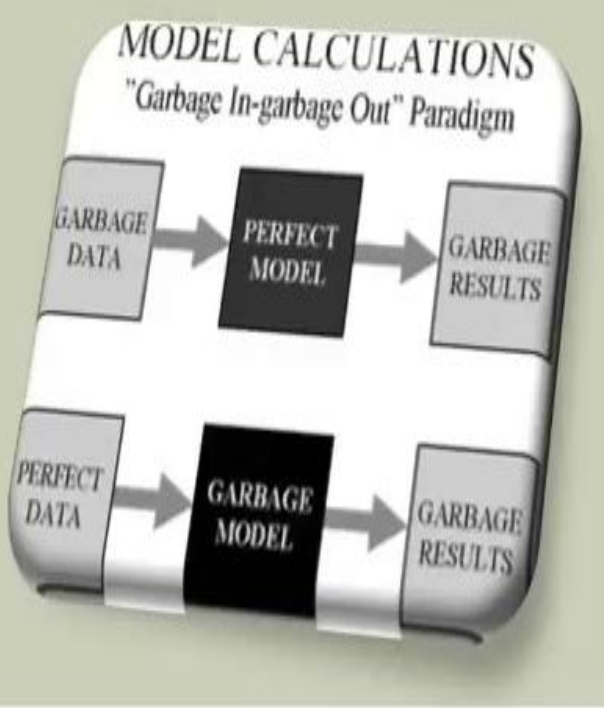
- Histogram and Boxplot
- Normal Quantile Plot
  (also called Normal Probability Plot)
- Goodness of Fit Tests

  Shapiro-Wilk Test (JMP)
  Kolmogorov-Smirnov Test (SPSS)
  Anderson-Darling Test (MINITAB)

LOOK AT YOUR DATA GRAPHICALLY FIRST

...*Before* starting all the fun, cool, whiz-bang analysis.

Get to know the data. Look for patterns, potential problems, initial relationships, etc.

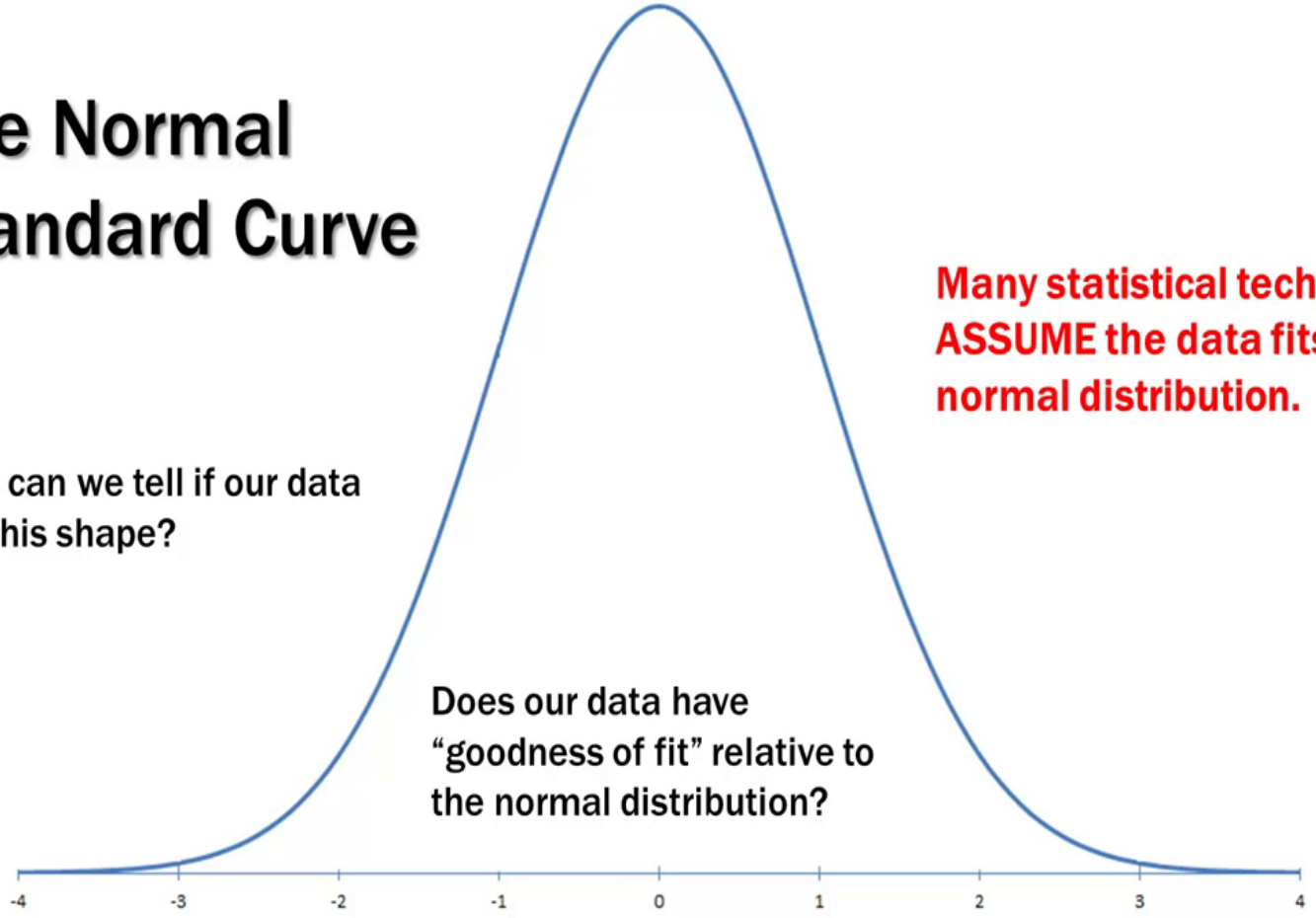


GARBAGE IN, GARBAGE OUT.

GRAPHICAL DATA EXPLORATION

- By using a few simple visual tools, we can learn a tremendous amount of information about our data
- Our data may have excess skew (lopsided), kurtosis (very fat tails), be bi-modal (two humps like a camel), or follow a distribution other than the normal distribution
- In this presentation we will briefly discuss the following tools to determine if our data is “normal”:
  - Histograms
  - Stem and Leaf Plots
  - Box Plots (Box and Whisker Plots)
  - P-P Plots
  - Q-Q Plots

The Normal Standard Curve



Many statistical techniques ASSUME the data fits a normal distribution.

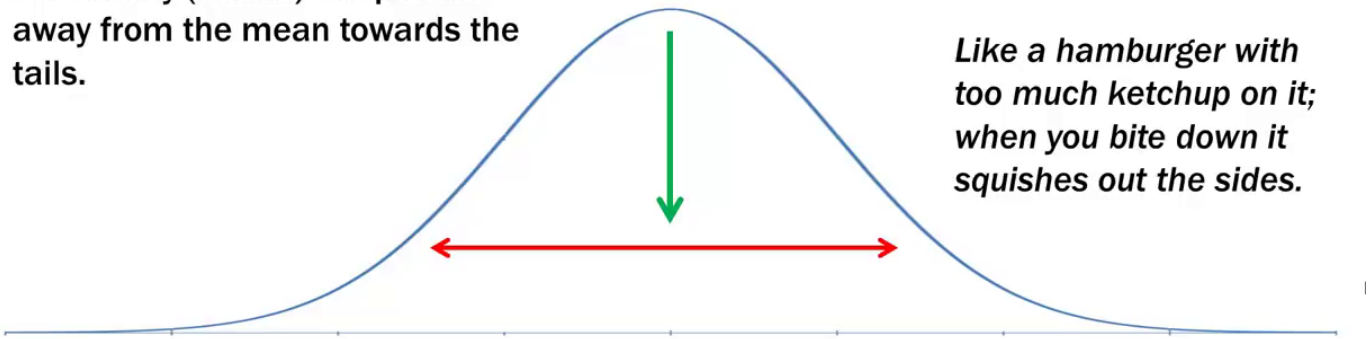
How can we tell if our data fits this shape?

Does our data have "goodness of fit" relative to the normal distribution?

Excess Kurtosis

More probability than expected in the tails of the distribution due to extreme values away from the mean.

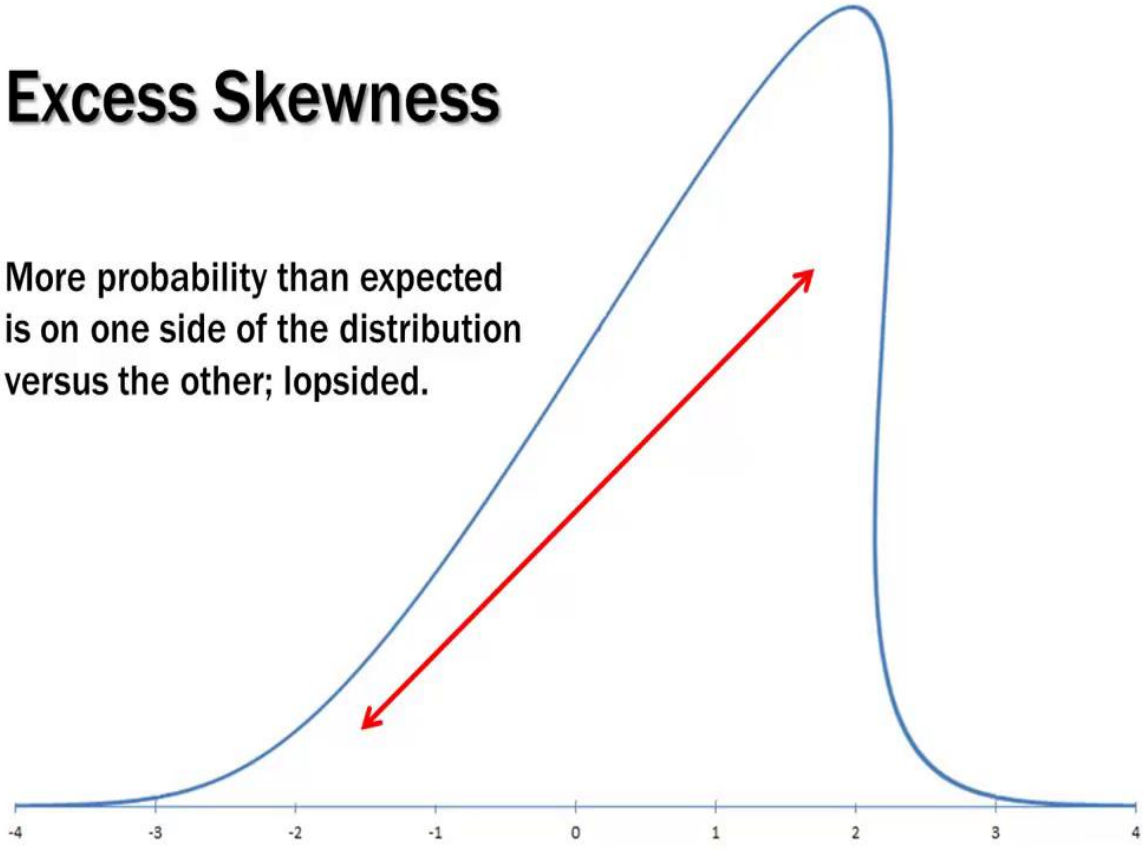
Probability (values) are pushed away from the mean towards the tails.



Like a hamburger with too much ketchup on it; when you bite down it squishes out the sides.

Excess Skewness

More probability than expected is on one side of the distribution versus the other; lopsided.



OTHER PROBABILITY DISTRIBUTIONS

Oftentimes data fits another type of distribution much better:

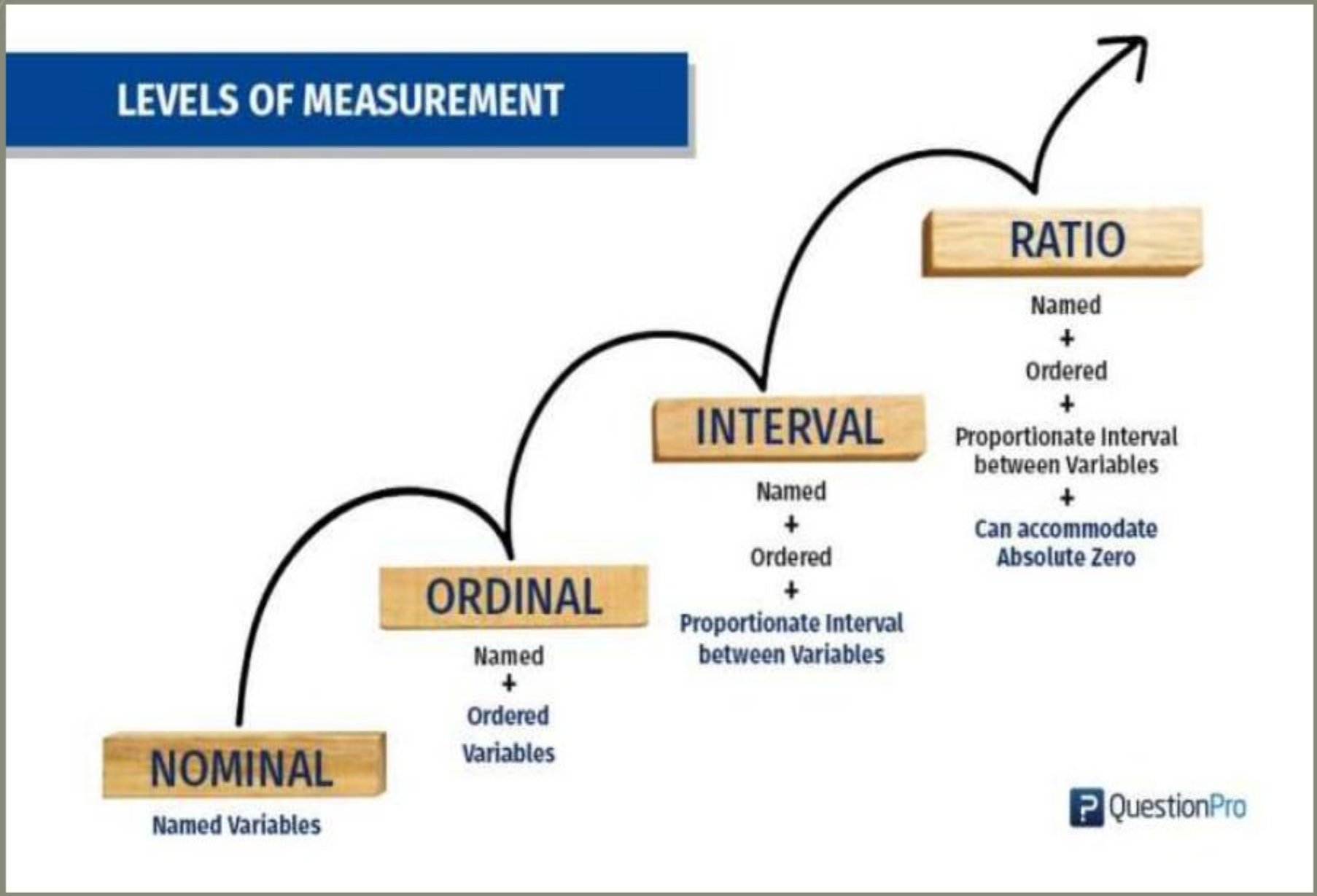
Exponential

Lognormal

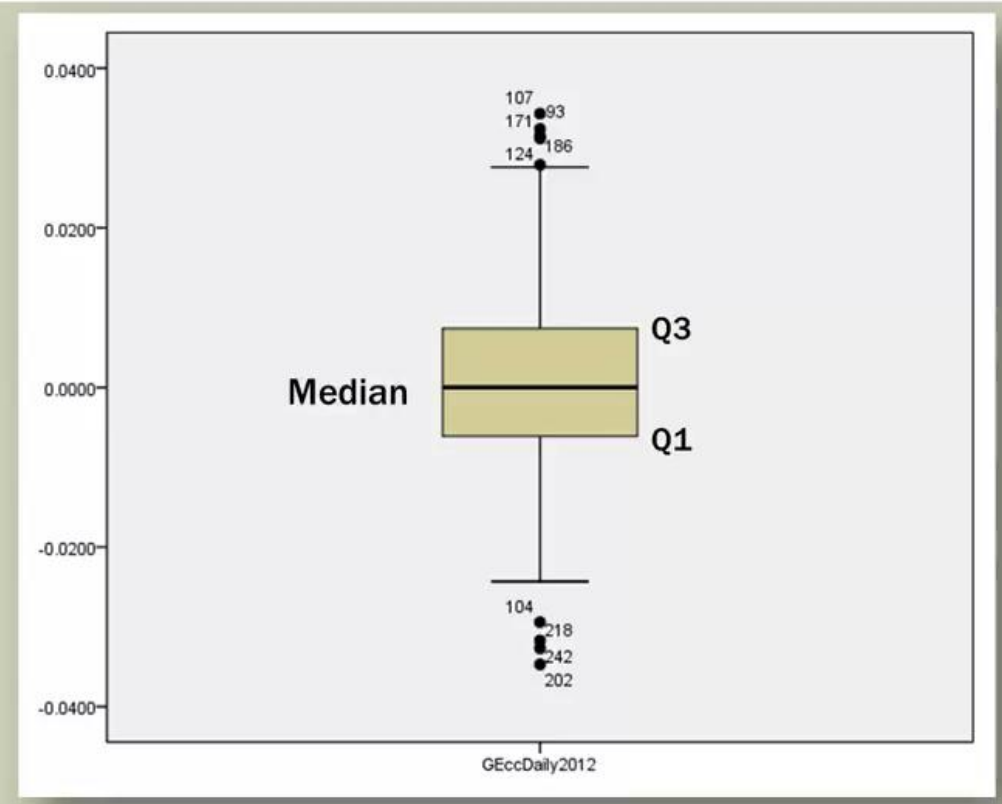
Among others....

Weibull

Uniform


LEVELS OF MEASUREMENT
NOMINAL
Named Variables
ORDINAL
Named
+
Ordered
Variables
INTERVAL
Named
+
Ordered
+
Proportionate Interval between Variables
RATIO
Named
+
Ordered
+
Proportionate Interval between Variables
+
Can accommodate Absolute Zero
QuestionPro

BOX PLOT

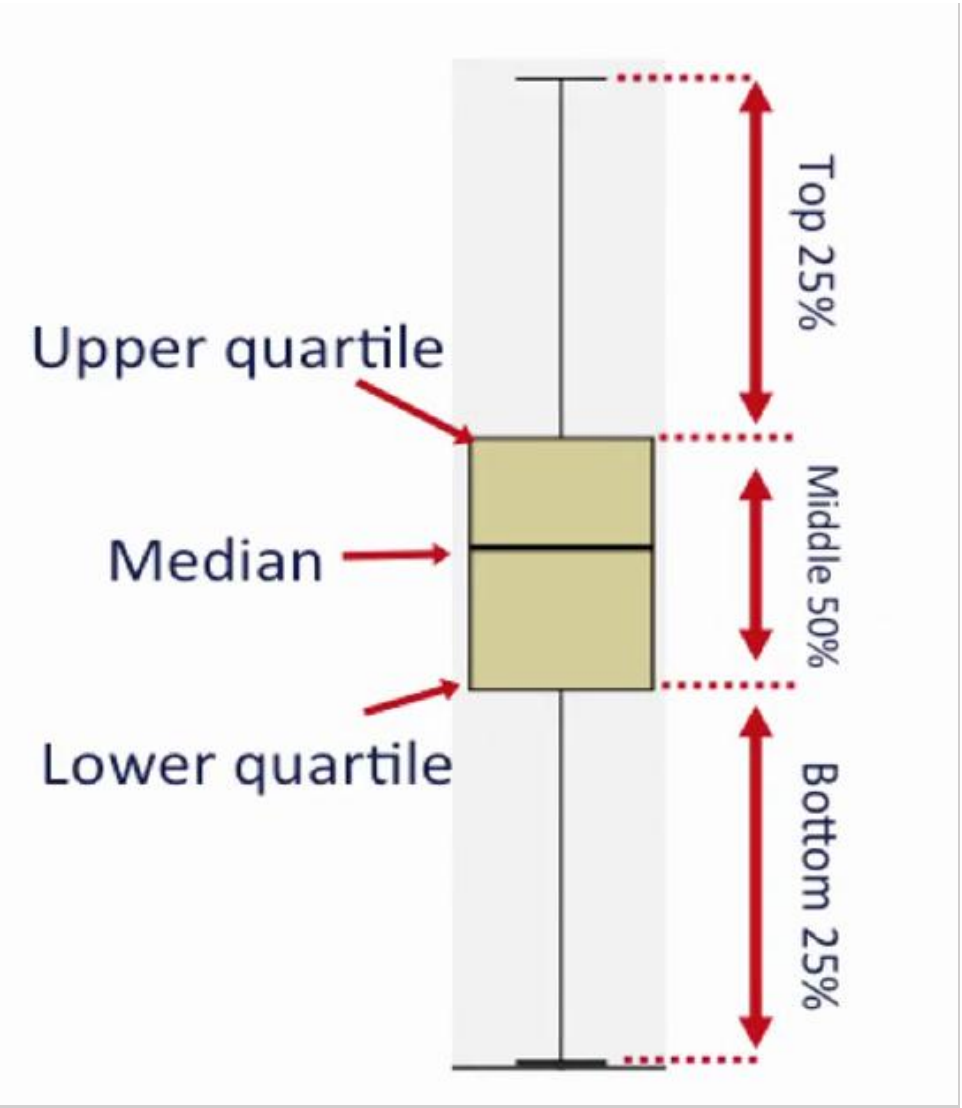


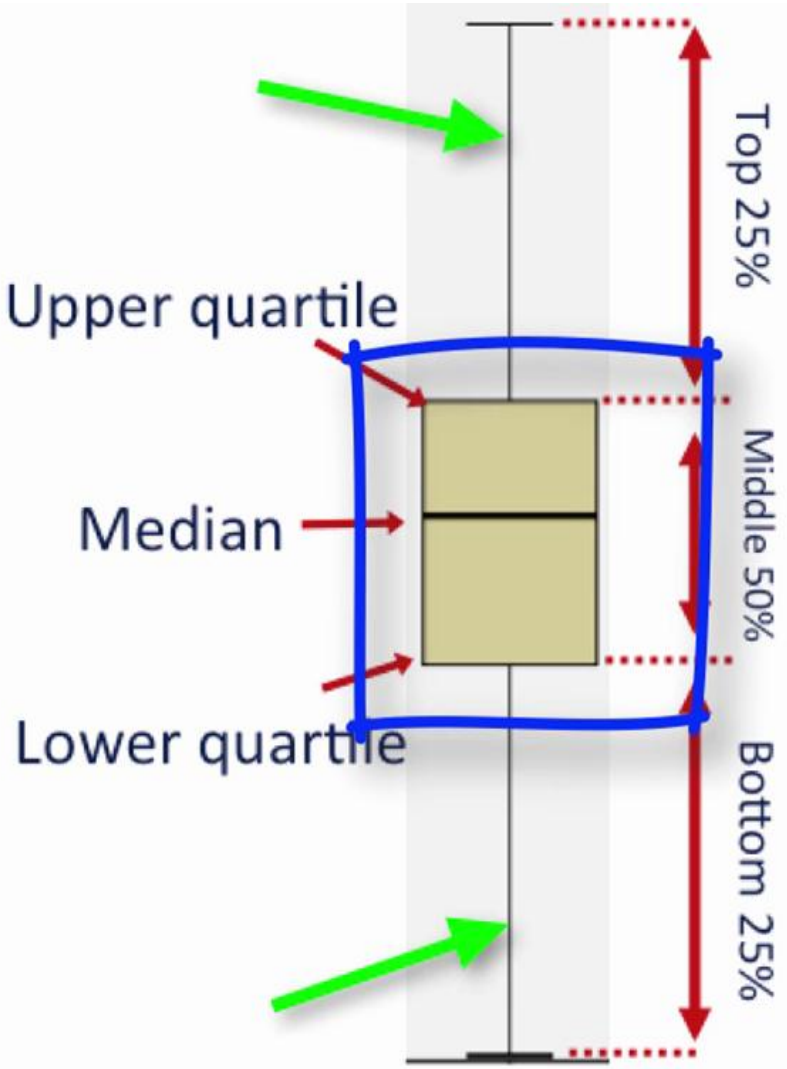
Box plots are relatively simple graphical tools for looking at the distribution of data.

So what should you look for?

1. Is the box plot symmetrical overall?
2. Are Q1 and Q3 approximately the same distance from the median?
3. Are the “whiskers” of the plot approximately the same length?

Box and Whisker Plots





Order numbers

3, 5, 4, 2, 1, 6, 8, 11, 14, 13, 6, 9, 10, 7

- First, order your numbers from least to greatest:

1, 2, 3, 4, 5, 6, 6, 7, 8, 9, 10, 11, 13, 14

Median

1, 2, 3, 4, 5, 6, 6, 7, 8, 9, 10, 11, 13, 14

- Then find the median (from the ordered list):
- Cross off one number from each side until you reach the middle number (or numbers).

1, 2, 3, 4, 5, 6, **6, 7,** 8, 9, 10, 11, 13, 14

Median (continued):

1, 2, 3, 4, 5, 6, **6, 7,** 8, 9, 10, 11, 13, 14

- If there are two numbers in the middle, Add those 2 middle numbers together:

  6 + 7 = 13
- Then divide by 2:

  13 ÷ 2 = 6.5
- The median is 6.5.

Quartiles (page 1)

1, 2, 3, 4, 5, 6, 6, 7, 8, 9, 10, 11, 13, 14

- Then split the numbers on left and right sides of the median:

1, 2, 3, 4, 5, 6, 6, | 7, 8, 9, 10, 11, 13, 14

Quartiles (page 2)

1, 2, 3, 4, 5, 6, 6, | 7, 8, 9, 10, 11, 13, 14

- Find the median for each half:

1, 2, 3, 4, 5, 6, 6 | 7, 8, 9, 10, 11, 13, 14

1, 2, 3, **4**, 5, 6, 6 | 7, 8, 9, **10**, 11, 13, 14

Left

Median = 4

Right

Median = 10

Quartiles (page 3)

1, 2, 3, **4**, 5, 6, 6 | 7, 8, 9, **10**, 11, 13, 14

Left

Median = 4

Right

Median = 10

- The left median is called the LOWER QUARTILE.
- The right median is called the UPPER QUARTILE.

Number line

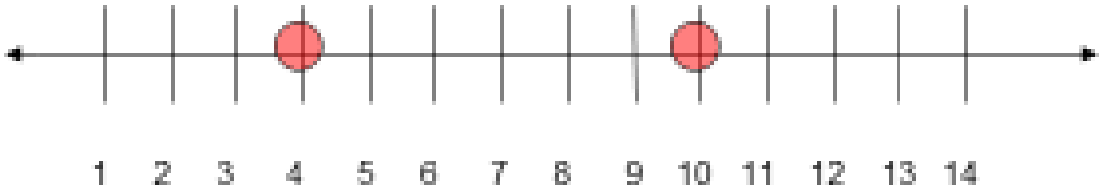
1, 2, 3, 4, 5, 6, 6, 7, 8, 9, 10, 11, 13, 14

- Draw a number line from the smallest to the largest number without skipping any numbers.

Quartiles on number line

1, 2, 3, **4**, 5, 6, 6, 7, 8, 9, **10**, 11, 13, 14

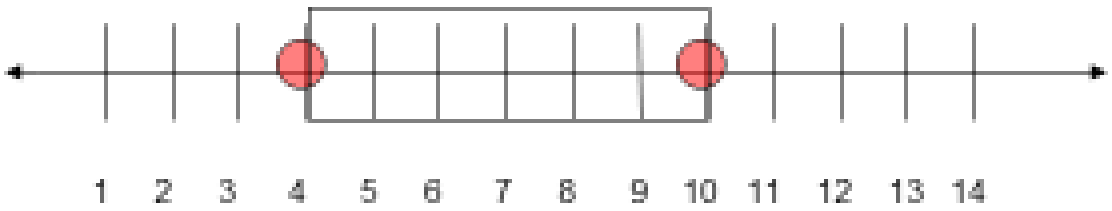
- Put circles at the LOWER and UPPER Quartiles.



Box on Quartiles on number line

1, 2, 3, **4**, 5, 6, 6, 7, 8, 9, **10**, 11, 13, 14

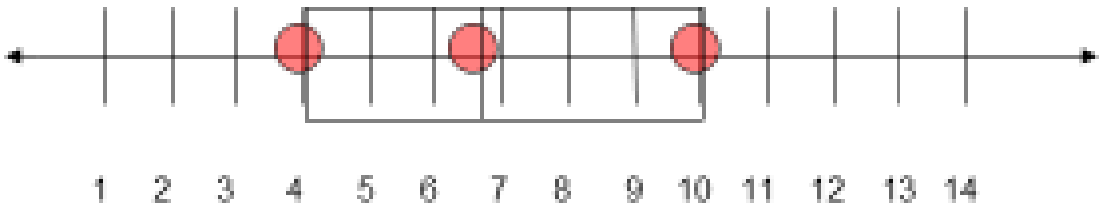
- Draw a box connecting the circles at the LOWER and UPPER Quartiles.



Median on number line

1, 2, 3, 4, 5, 6, **6, 7**, 8, 9, 10, 11, 13, 14

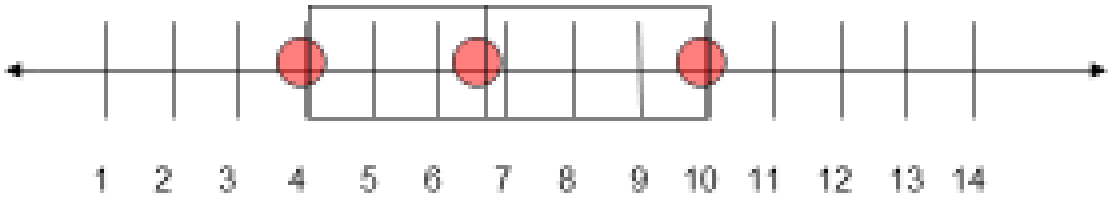
- Put a circle at the median (6.5).



Median on number line

1, 2, 3, 4, 5, 6, **6, 7**, 8, 9, 10, 11, 13, 14

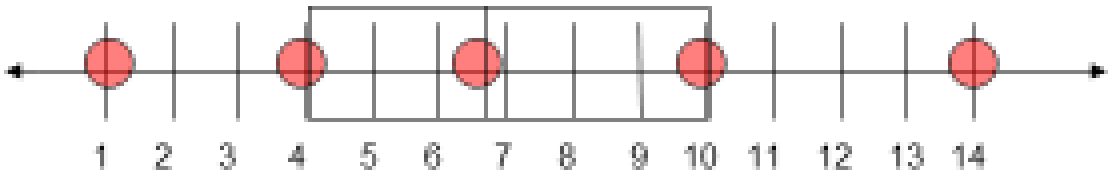
- Draw a line connecting the median to the box.



Low and high numbers

1, 2, 3, 4, 5, 6, 6, 7, 8, 9, 10, 11, 13, **14**

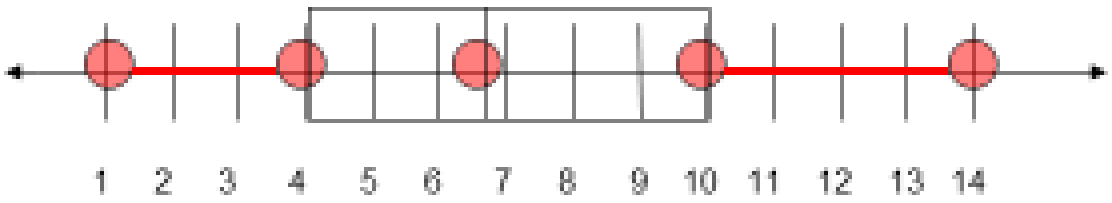
- Put circles at the high and low points.



Low and high numbers

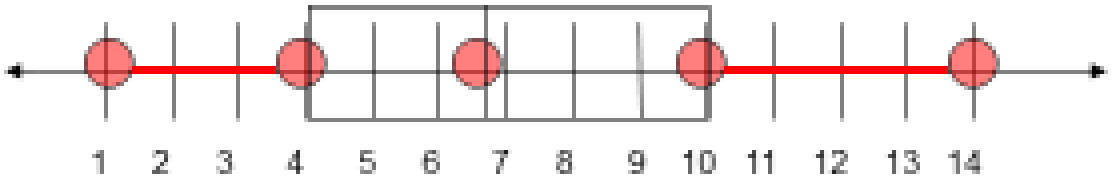
1, 2, 3, 4, 5, 6, 6, 7, 8, 9, 10, 11, 13, **14**

- Draw lines that connect the high and low points to the box.



Box and Whisker Plot

3, 5, 4, 2, 1, 6, 8, 11, 14, 13, 6, 9, 10, 7



Here is the completed Box and Whisker Plot!

Tests of Normality

There are several different tests that can be used to test the following hypotheses:

H_o: The distribution is normal

H_A: The distribution is NOT normal

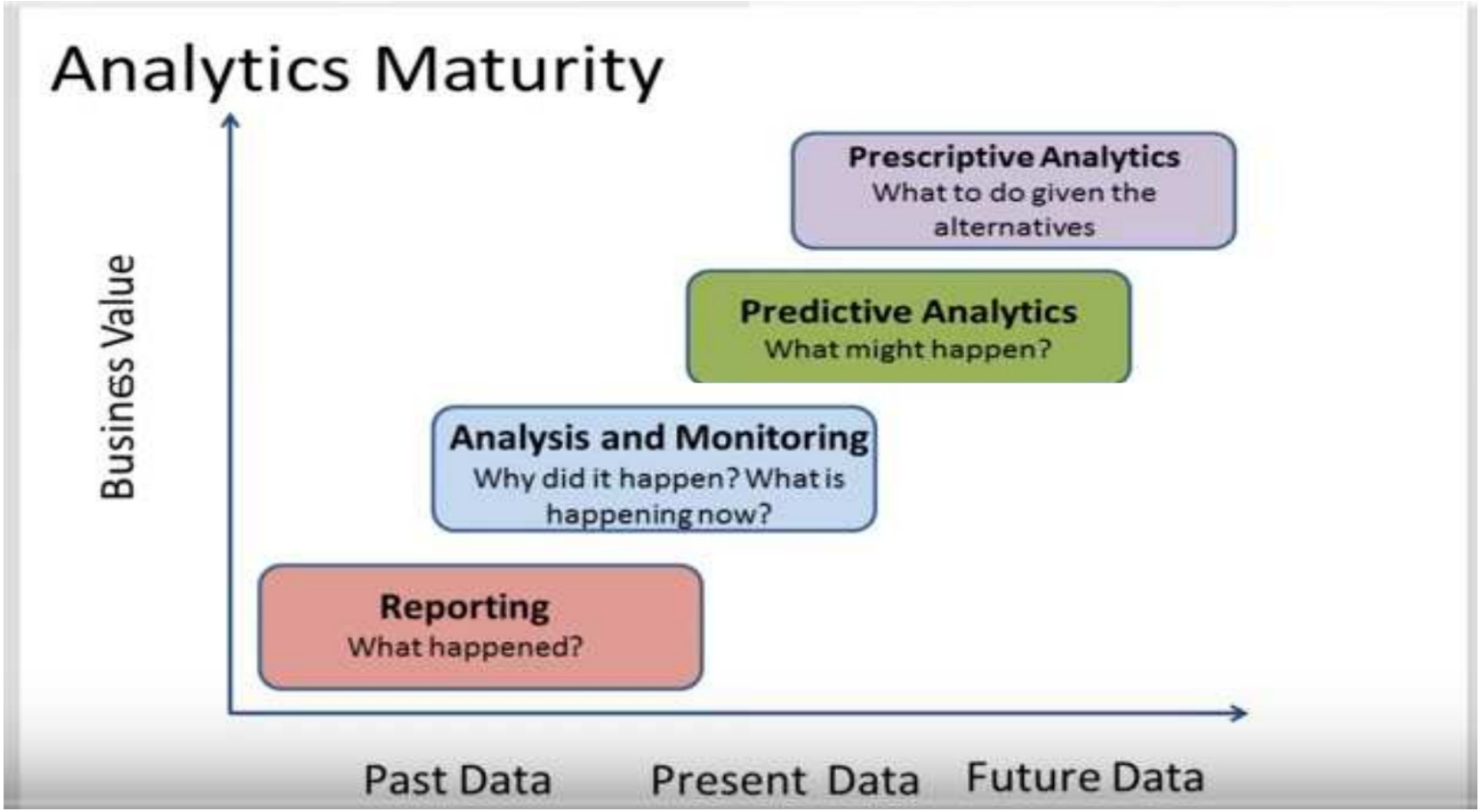
Common tests of normality include:

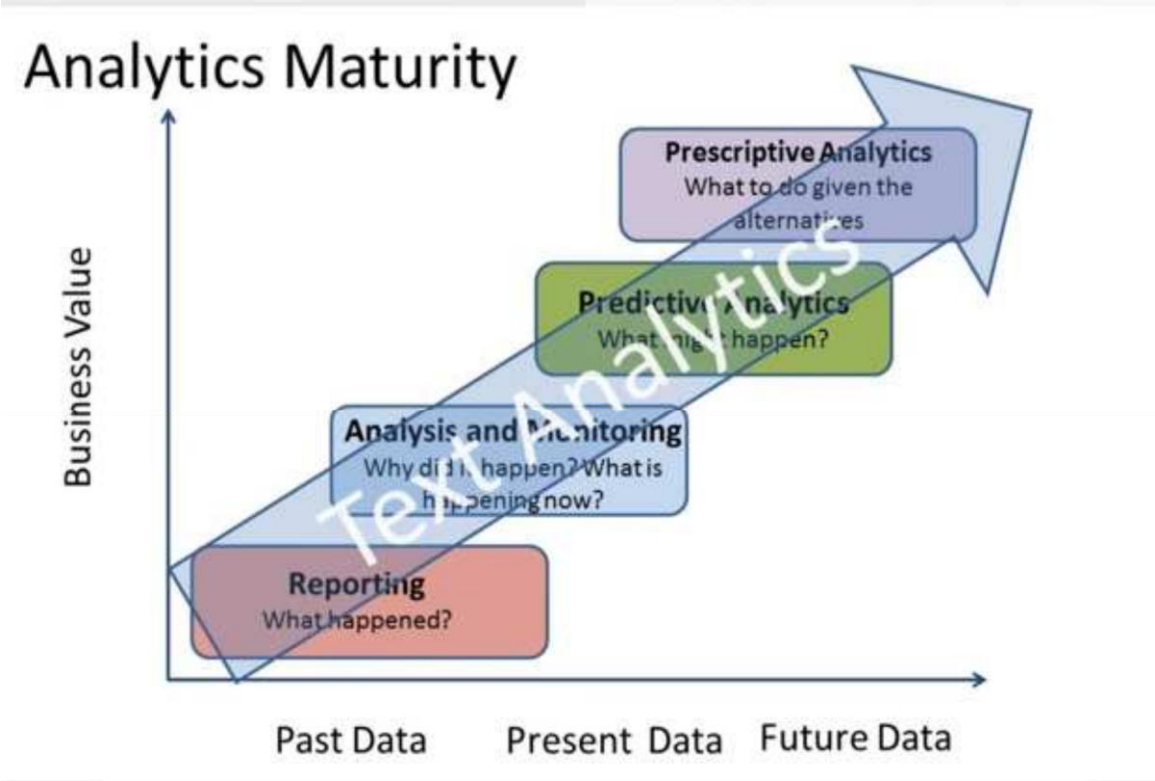
Shapiro-Wilk **Kolmogorov-Smirnov**

Anderson-Darling **Lillefor's**

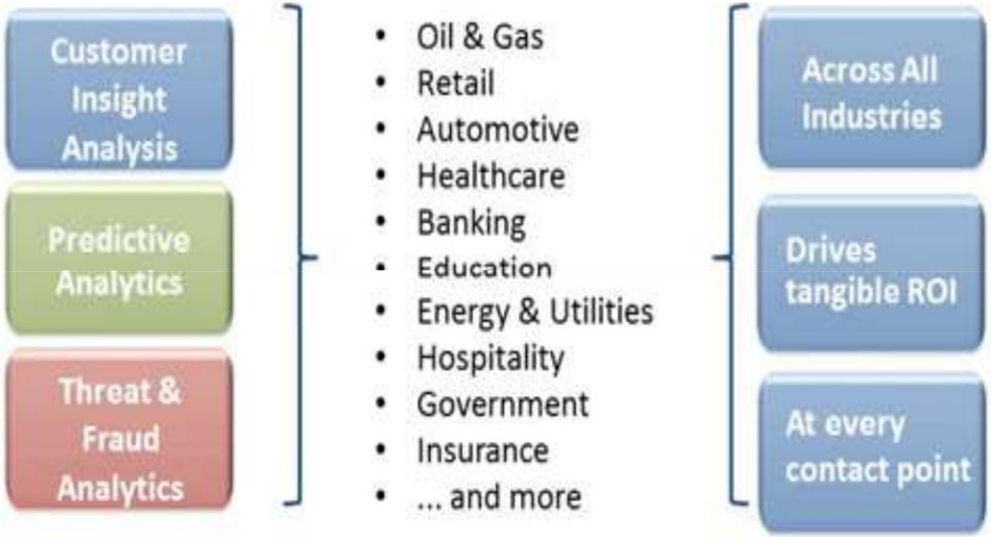
Problem: THEY DON'T ALWAYS AGREE!!

A JOURNEY TO TEXT ANALYTICS


Analytics Maturity
Business Value
Prescriptive Analytics
What to do given the alternatives
Predictive Analytics
What might happen?
Analysis and Monitoring
Why did it happen? What is happening now?
Reporting
What happened?
Past Data
Present Data
Future Data


Analytics Maturity
Business Value
Prescriptive Analytics
What to do given the
Reporting
What happened?
Past Data
Present Data
Future Data

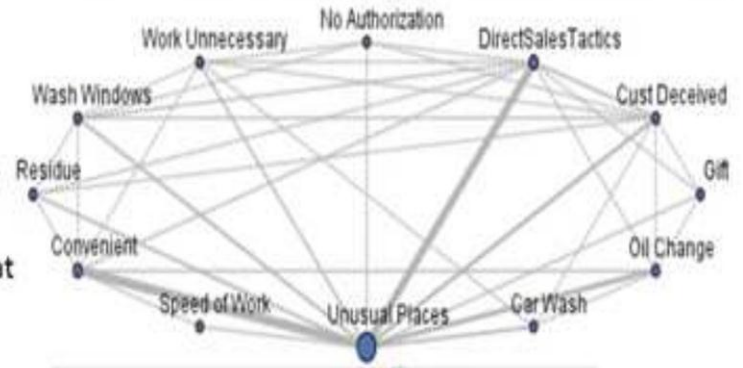
Where is Text Analytics Used?

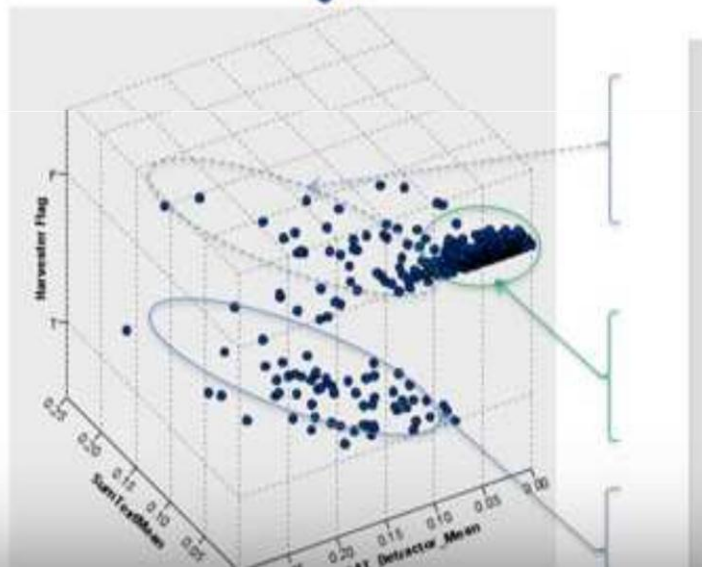


Text Analytics allows an organization to gain a better understanding of contextual data at a granular level.

Fraud Detection

- Identifying new data sources to feed into existing fraud detection models
  - 80% of data is in unstructured format which means that most prediction models used for fraud detection are only using only 20% of the data available.
- Reduce overhead associated with existing fraud detection methods
  - Identifying focused subset of transactions and that have high likelihood of being fraudulent
  - Reduce time spent searching for anomalous transactions (needle in haystack)

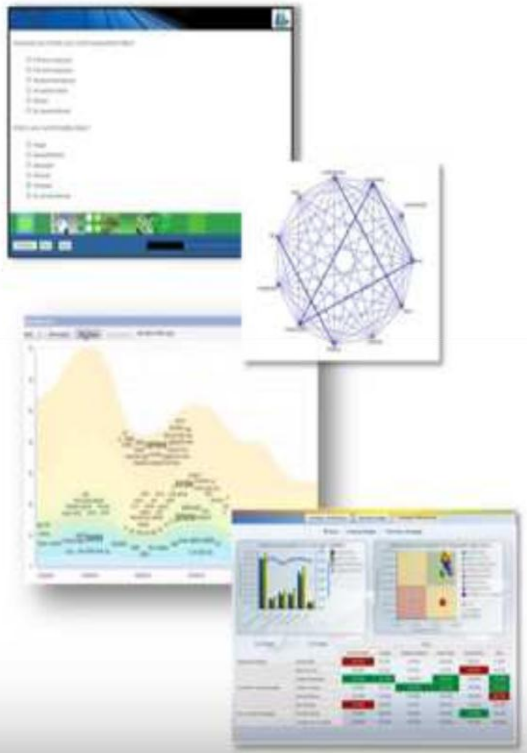




10:02

Marketing

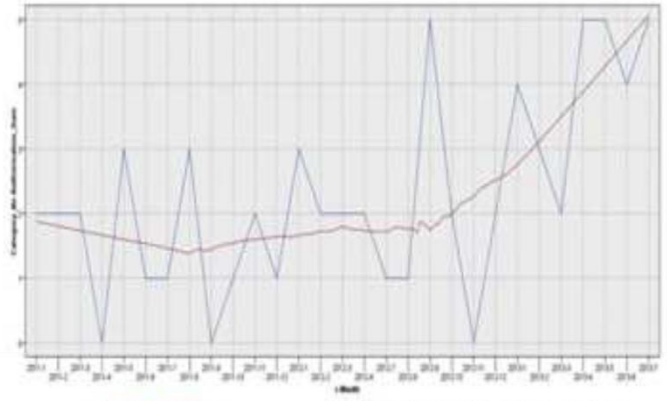
- Gain insights on what is driving customer behaviors.
- What concepts are most correlated to the desired outcome.
  - How important is price in the purchase decision?
  - Did a recent initiative have an impact on sales?
- What are customers saying about your sales cycle in their survey's?
  - Identify new products offerings to target 'untapped' customer segments.
- Maximize customer lifetime value
- Increase long term customer profitability
- Customer Retention with Churn Modeling

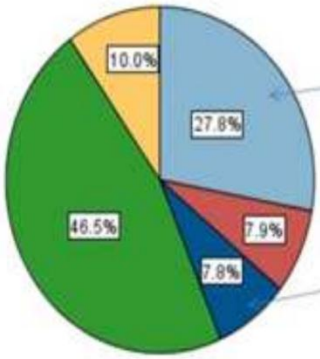


Voice of the customer

***Definition**: In-depth process of capturing a customer's expectations, preferences and aversions.*

- Tracking customer sentiment.
  - Through internal survey responses.
  - Social Media outlets
  - Other outward facing sources
- Understand key customer concepts over time.
- Use text mining concepts to create segmented customer groups/clusters.
- Are there external influences that explain customer behavior?





Cluster #1: Most Loyal
- Live closest to stores
- Highest positive sentiment
- Youngest demographic

Cluster #3: At risk of churn
- Most mentions of pricing concerns
- Most negative sentiment score

Thank You