Data Analytics and Interactive Dashboard using Python

PROTOCOL FOR VIDEO CONFERENCE

1. PLEASE SHOW YOURSELF
2. LEARNERS ARE NOT TO RECORD THE TRAINING
3. NO DOWNLOADING OF MATERIALS
4. USE YOUR EARPIECE
5. MUTE YOUR MIC UNLESS YOU ARE SPEAKING
6. ENSURE NO BRIGHT LIGHTS ARE DIRECTLY BEHIND YOU

Too Much Stress??

Today's Schedule

9am: Session Start
Lesson
10-30am: 15min break
Lesson
12-30pm: Lunch break
Lesson
1-45pm - 3pm: 15 min break
Lesson
4-45pm: 10 min break
Lesson
5-30pm - 6pm: Session End

What will you be learning in this course?

- Basics of Python (Recap)
- Understanding simple Data Structures
- Data Cleaning and Manipulation using **Numpy** and **Pandas**
- Basic Data Visualization using **Matplotlib**
- Data Analytics and Visualization using **Seaborn**
- Building simple dashboards in Seaborn

Method of Learning

90% Hands-on Coding on Python

10% Presentation

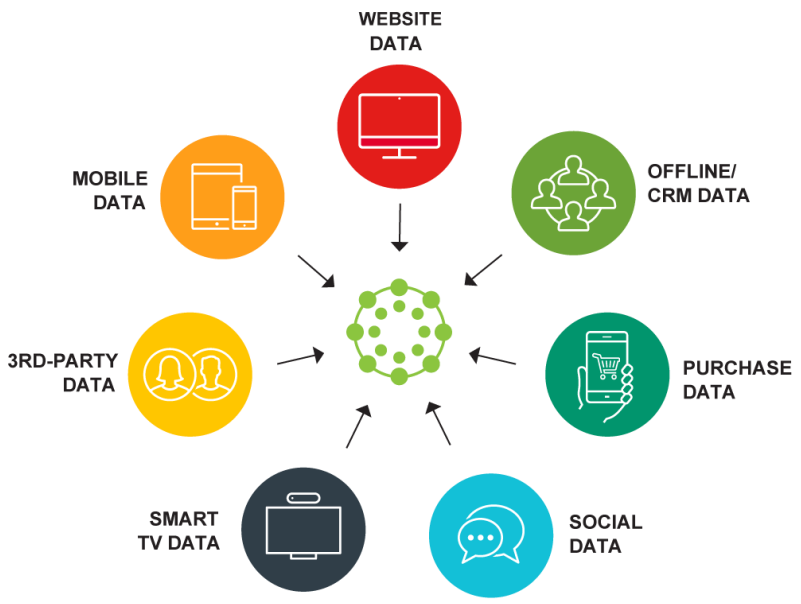
2,500,000,000,000,000,000

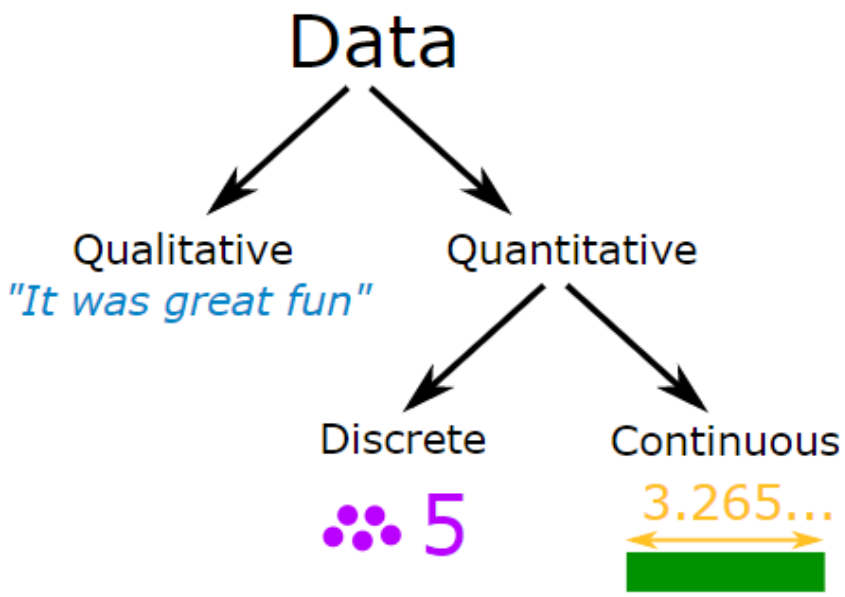
(Two and half quintillion)

This the amount of data we create every single day!!

What is data?

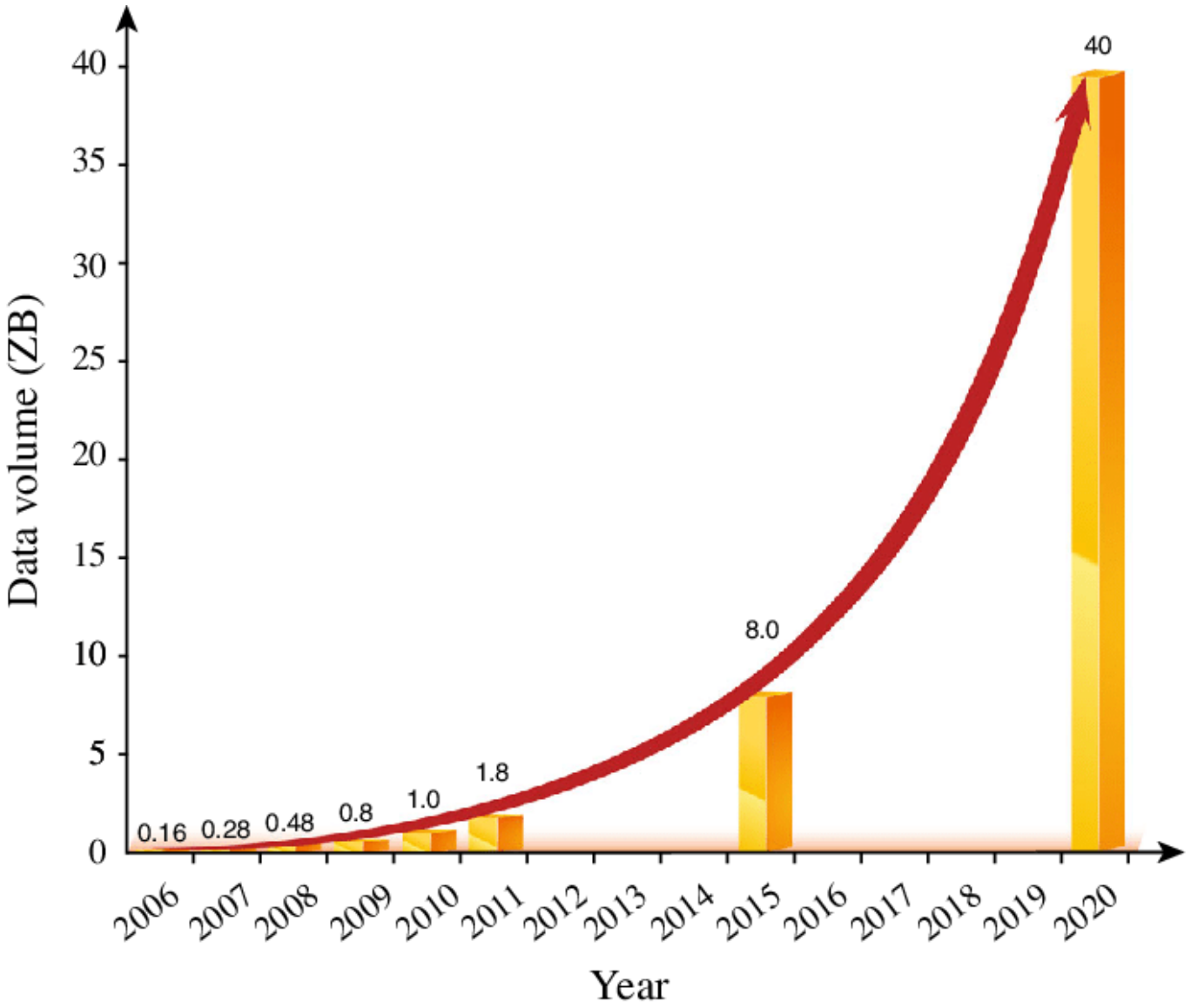
Data is a collection of facts, such as numbers, words, measurements, observations or even just descriptions of things.



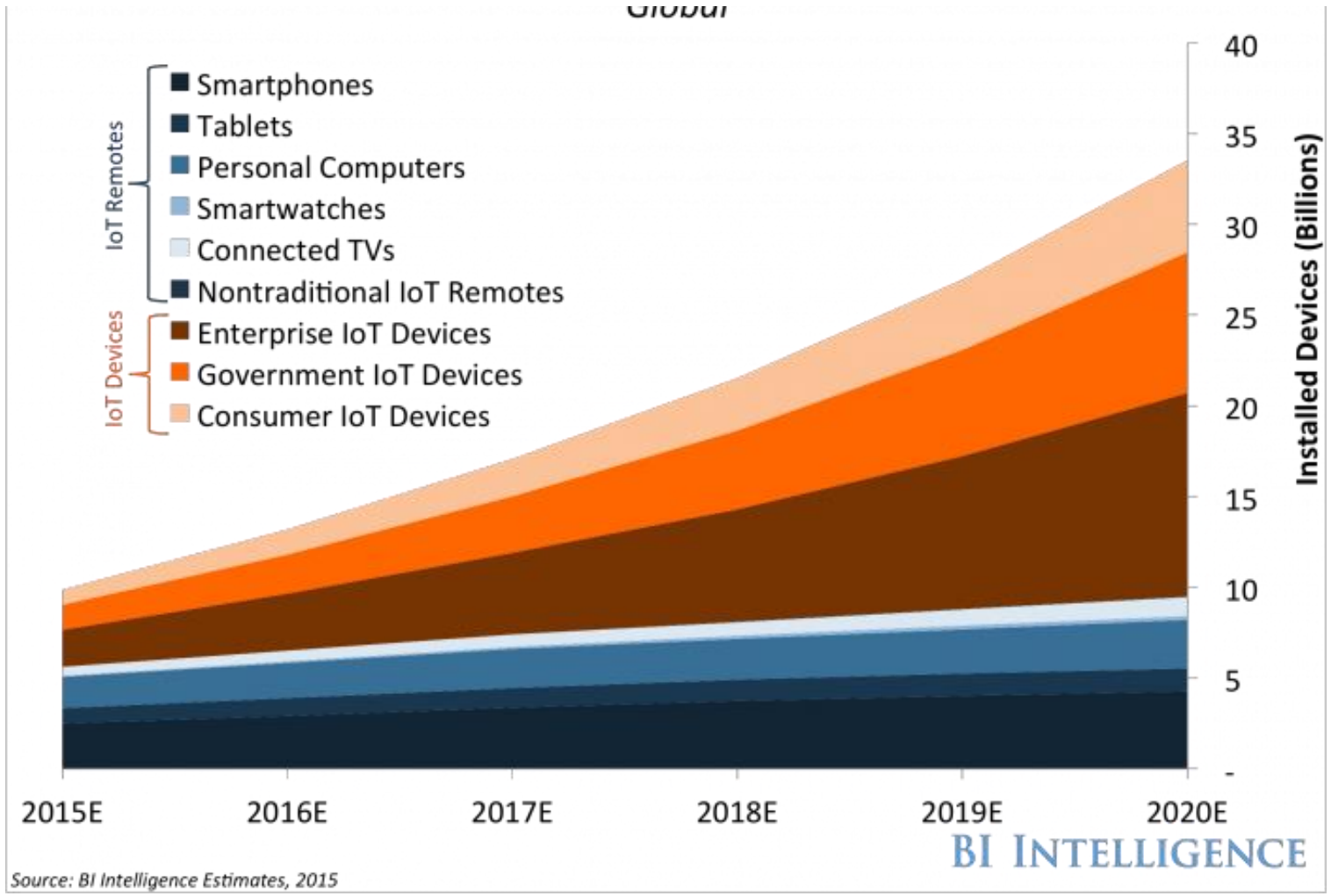


Significant growth in Data Science & Analytics

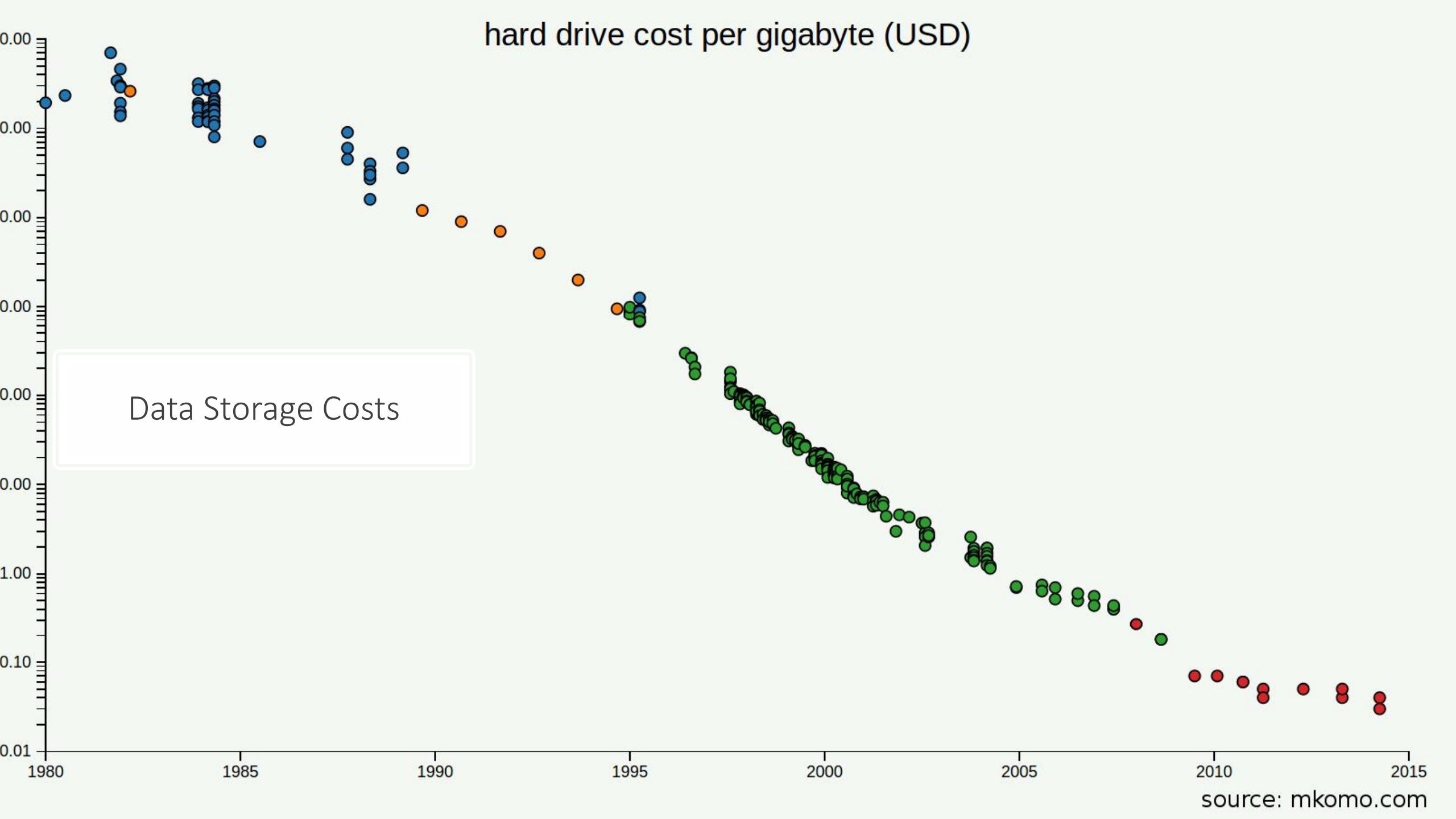
Explosion of data volume



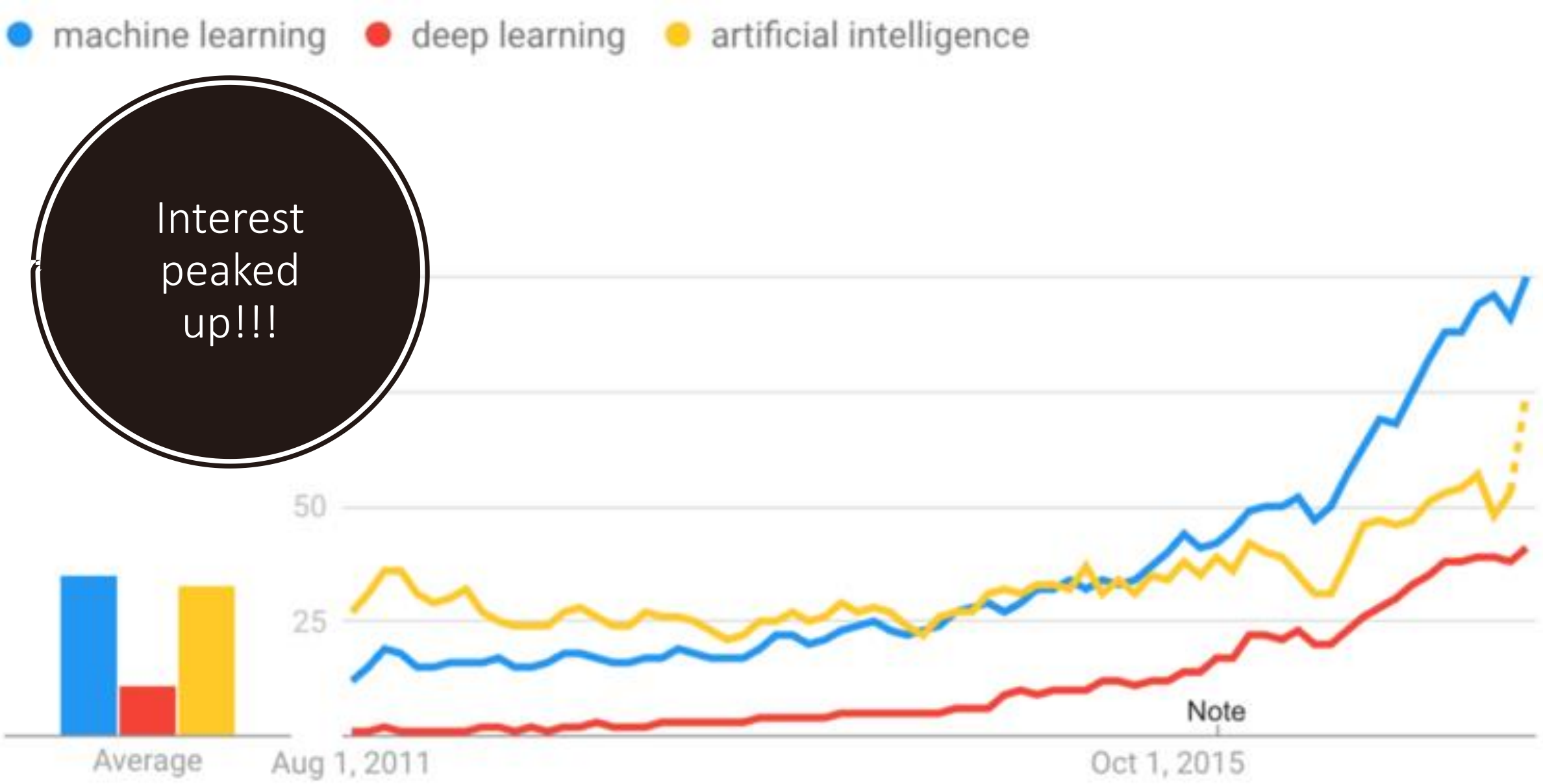
Source: ResearchGate



Devices connected to the internet



Data Storage Costs



Worldwide. 7/9/11 - 8/9/17.

"DATA IS THE NEW OIL."

Coined in 2006 by Clive Humby, a British data commercialization entrepreneur this now famous phrase was embraced by the World Economic Forum in a 2011 report, which considered **data to be an economic asset,** like oil.

From the beginning of recorded time until 2003, we created **5 exabytes** (5 billion gigabytes) **of data.**

In 2011 the same amount was created every two days.

By 2013, it's expected that the time will shrink to 10 minutes.

Every hour, we create enough Internet traffic to fill **7 billion DVDs.**

Side by side, that's that's seven times the height of Everest.

There are nearly as many bits of information in the digital universe as there are stars in our actual universe.

There are **133 million BLOGS** on the web.

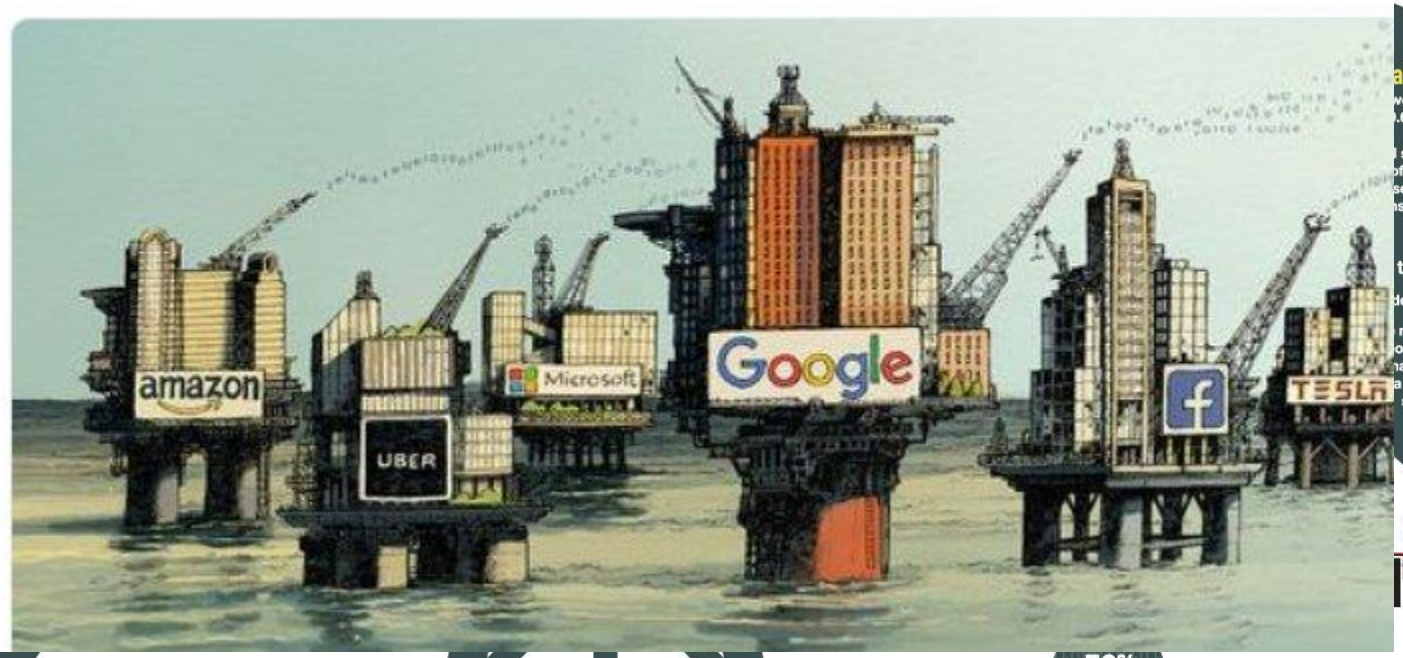
As of August 2012, there were just over

millions of users

Just as a study of activity on Twitter gave residents, family members, and journalists advance warning of details about the devastating earthquake and tsunami in Japan, **high-frequency traders,** with the help of computer algorithms, use Big Data to follow trends and to act quickly

The Economist @TheEconomist · 2h

The world's most valuable resource is no longer oil, but data



gorithms
ecisions to buy or sell a commodity.
g laid under the Atlantic will shave
5 milliseconds
from the current 65 milliseconds it takes for trading instructions to travel between New York City and London.

able,
ween New York
.6 milliseconds.

saving is worth
of dollars to the trading
se the cable (and who will
s to do so).

they save 5 milliseconds

depth of the Atlantic Ocean varies.

new cable will lie on areas of the ocean
or that are up to 1,000 feet shallower
han the current fastest cable. By taking
a different route, the new cable is
shorter, meaning that the time it
takes for messages to travel
along it is shortened.

The new cable takes a shallower, therefore shorter route.

UK

50% of 5-year-old kids in the U.S. are given access to a smartphone.

Information Graphics by Nigel Holmes

What can we do with the data?

EXPLORATORY DATA ANALYSIS

DERIVE USEFUL INSIGHTS

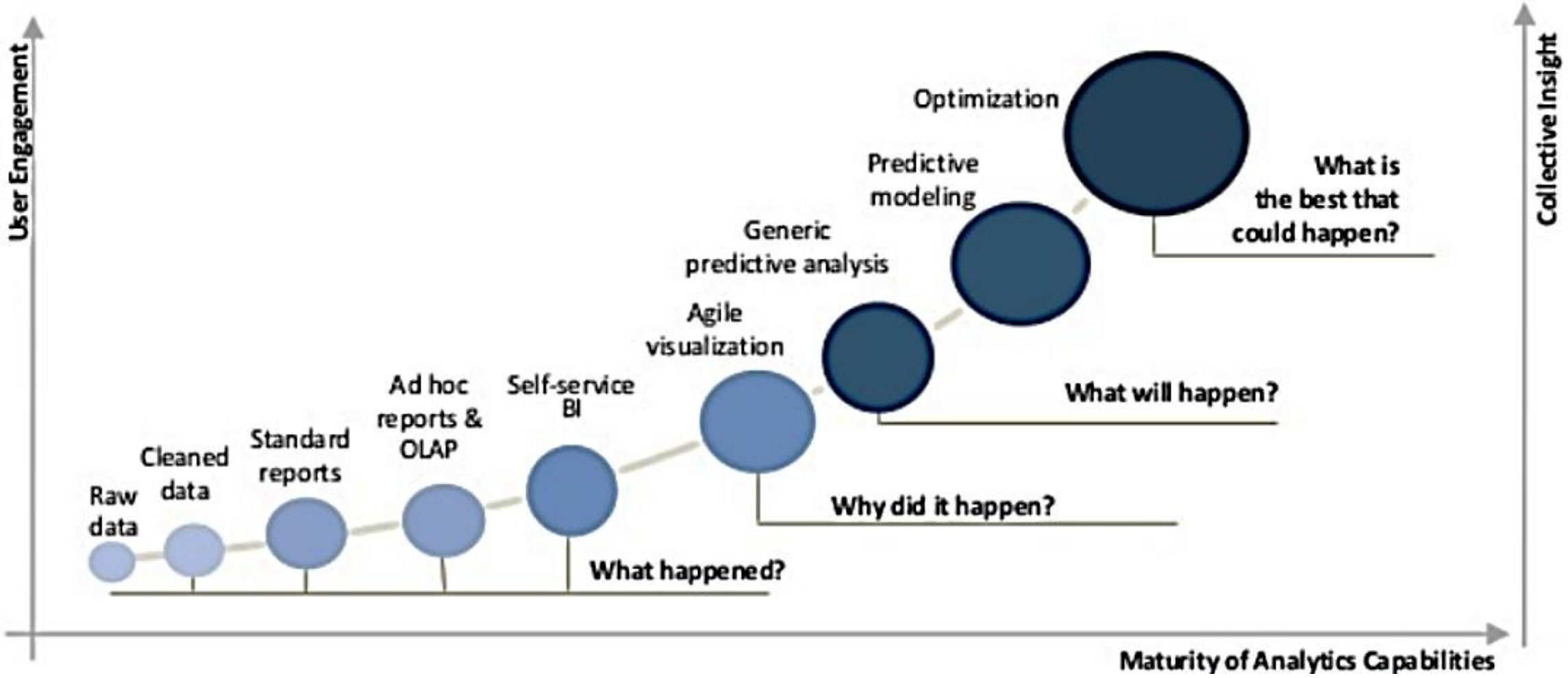
CREATE DASHBOARDS AND REPORTS

DERIVE NEW INFERENCES

What is Analytics?

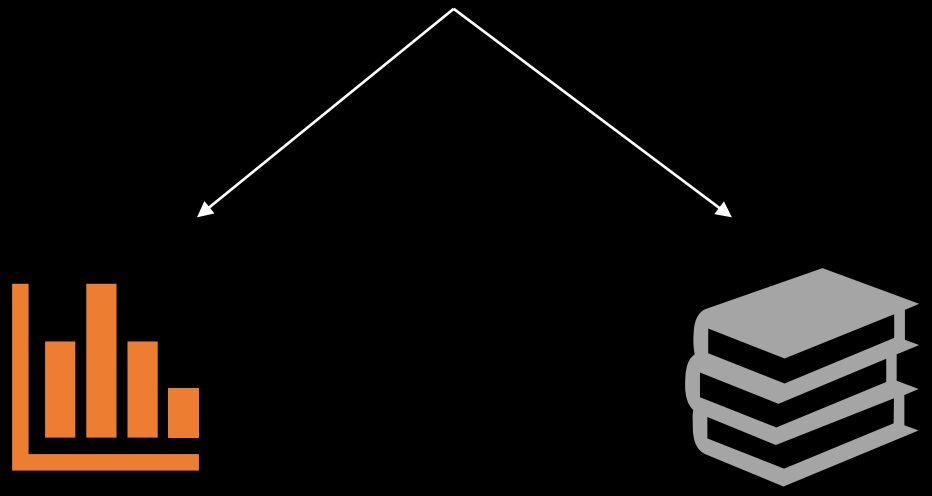
Analytics is the discovery, interpretation, and communication of meaningful patterns in data. It also entails applying data patterns towards effective decision making. In other words, analytics can be understood as the connective tissue between data and effective decision making within an organization.

Source: Wikipedia


User Engagement
Collective Insight
Raw data
Cleaned data
Standard reports
Ad hoc reports & OLAP
Self-service BI
Agile visualization
Generic predictive analysis
Predictive modeling
Optimization
What happened?
Why did it happen?
What will happen?
What is the best that could happen?
Maturity of Analytics Capabilities

User Engagement

Analytics is comprised of two important parts



Visualization Storytelling

Data Visualization

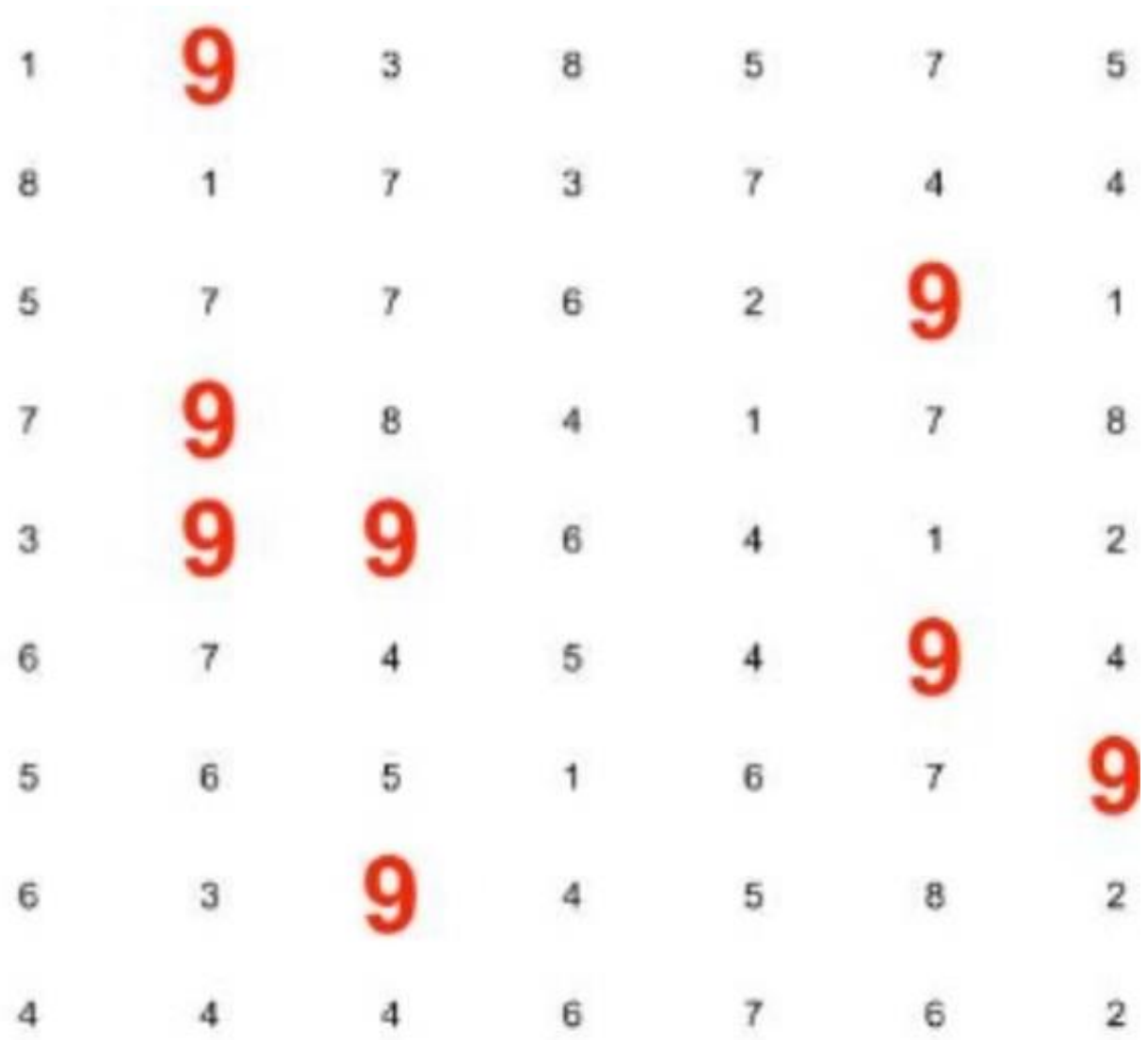
Let’s start by testing the human visual system

How many **9s** are present ?

1	9	3	8	5	7	5
8	1	7	3	7	4	4
5	7	7	6	2	9	1
7	9	8	4	1	7	8
3	9	9	6	4	1	2
6	7	4	5	4	9	4
5	6	5	1	6	7	9
6	3	9	4	5	8	2
4	4	4	6	7	6	2

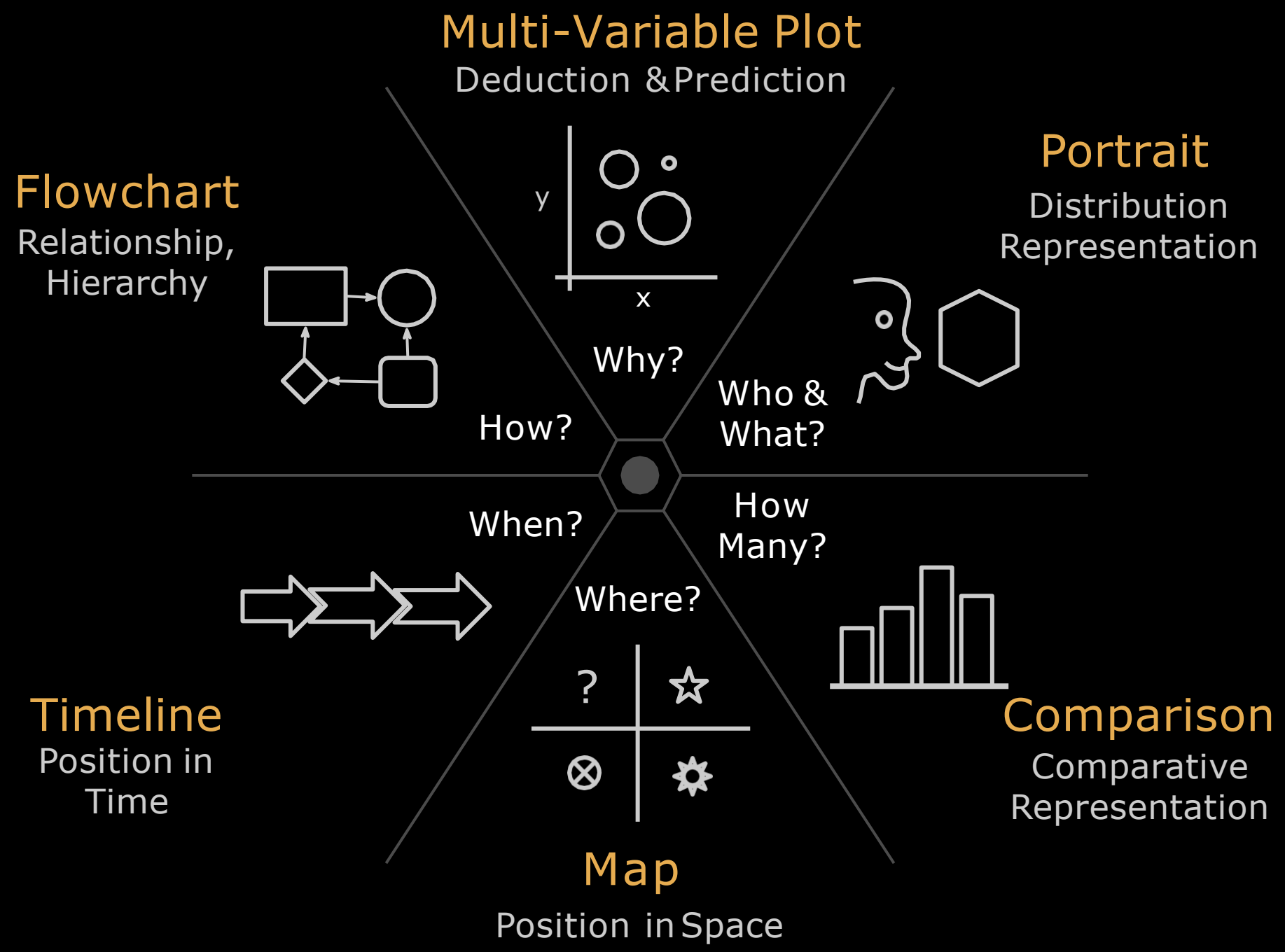
The human visual system is powerful

How many **9s** are present now?



What is Data Visualization?

Data visualization is the presentation of data in a pictorial or graphical format. It enables decision makers to see analytics presented visually, so they can grasp difficult concepts or identify new patterns.


Multi-Variable Plot
Deduction & Prediction
y
x
Why?
Portrait
Distribution Representation
Who & What?
Flowchart
Relationship, Hierarchy
How?
How Many?
When?
Where?
?
Timeline
Position in Time
Comparison
Comparative Representation
Map
Position in Space

Data Storytelling

Every human needs a story to make things memorable

What does storytelling mean?

Storytelling is used in design as a technique to get insight into users, build empathy and access them emotionally

Why does storytelling matter?

Stories make data meaningful

Stories tell to sell

Stories simplify

Stories crystallize takeaways

What are insights?

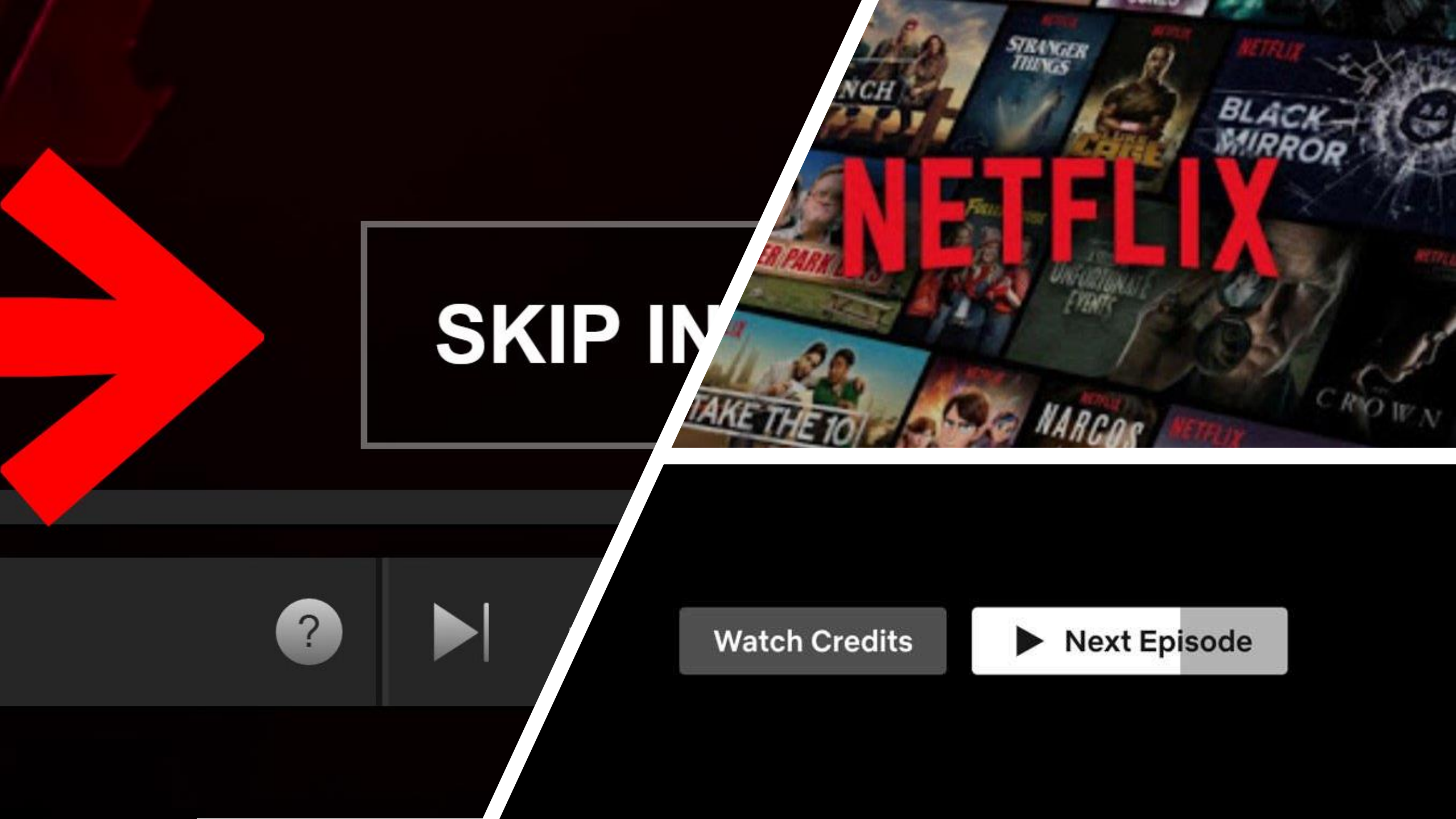
- Uncovering a shared meaning, a shared value, or a shared need that can be translated into action.
- Insight is what is learned and what will improve your business. Your business will know better, so you'll be able to work better.

Example of Finding, Insight, Recommendation (NETFLIX)

Finding - Customers are not watching the entire video to its full length. They are watching 90–95%

Insight - The parts they are not watching are the title roll and the end credits

Recommendation - Introduce ‘Skip Intro’ at the beginning of title rolls and ‘Watch Next’ at the beginning of end credits. Benchmark 90–95% watched content as completed and measure if customers move to the next video in the series


SKIP IN
NETFLIX
STRANGER THINGS
BLACK MIRROR
TAKE THE 10
NARCOS
CROWN
Watch Credits
Next Episode

Data Workflow

Frame

Develop hypotheses for your analysis.

Prepare

Select, import, & clean relevant data.

Analyze

Structure, visualize & complete your analysis.

Interpret

Create insights from your analysis.

Communicate

Present recommendations to your audience.

What is Python?

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics.

Often, programmers fall in love with Python because of the increased productivity it provides. Since there is no compilation step, the edit-test-debug cycle is incredibly fast.

What should you know?

DATA TYPES - BASIC AND ADVANCED

LIBRARIES (PANDAS, NUMPY, MATPLOTLIB)

FUNCTIONS

FLOW CONTROL

BASIC VISUALIZATIONS

Part 1 – Frame and Prepare

1

Data Issues & Cleaning

2

Data Preprocessing

3

Data Manipulation

Data Cleaning

Where can loss of data quality occur?

Loss of data quality can occur at many stages:

At the time of collection

During digitisation

During documentation

During storage and archiving

During analysis and manipulation

At time of presentation

And through the use to which they are put

Why do we have to do Data Cleaning?

- Inaccurate data analytics result into misguided decision making which can expose the industry to compliance issues. Data Cleaning ensures the above does not happen.
- It also streamlines business practices and improves efficiency.
- Increased sales and revenue are a result of data cleaning.

Issues in Data

- Duplicate data
- Irrelevant values
- Missing values
- Inaccurate data
- Old data

Data Preprocessing & Manipulation

- Cleaning
- Instance selection
- Normalization
- Transformation
- Feature extraction
- Feature selection

Example: Indexing & Slicing of Data

indexing: getting a specific element

0 1 2 3

grades = [88, 72, 93, 94]

```
>>> grades[2]
93
```

slicing: selecting a set of elements

grades = [88, 72, 93, 94]

0 1 2 3 4

```
>>> grades[1:3]
[72, 93]
```

Reiterating

•“data scientists spend 80% of their time cleaning and manipulating data and only 20% of their time actually analyzing it.”

Libraries in Python that can help with Data Cleaning and Manipulation

Numpy

Pandas

Tabular data with heterogeneously-typed columns, as in an SQL table or Excel spreadsheet

Ordered and unordered (not necessarily fixed-frequency) time series data.

Arbitrary matrix data (homogeneously typed or heterogeneous) with row and column labels

Any other form of observational / statistical data sets. The data actually need not be labeled at all to be placed into a pandas data structure

Why is Pandas important?

What can it handle?

Easy handling of **missing data** (represented as NaN) in floating point as well as non-floating point data

Size mutability: columns can be **inserted and deleted** from DataFrame and higher dimensional objects

Automatic and explicit **data alignment**: objects can be explicitly aligned to a set of labels, or the user can simply ignore the labels and let Series, DataFrame, etc. automatically align the data for you in computations

Powerful, flexible **group by** functionality to perform split-apply-combine operations on data sets, for both aggregating and transforming data

Make it **easy to convert** ragged, differently-indexed data in other Python and NumPy data structures into DataFrame objects

Intelligent label-based **slicing**, **fancy indexing**, and **subsetting** of large data sets

Intuitive **merging** and **joining** data sets

Flexible **reshaping** and pivoting of data sets

Hierarchical labeling of axes (possible to have multiple labels per tick)

Robust IO tools for loading data from **flat files** (CSV and delimited), Excel files, databases, and saving / loading data from the ultrafast **HDF5 format**

Time series-specific functionality: date range generation and frequency conversion, moving window statistics, moving window linear regressions, date shifting and lagging, etc.

Let’s dive straight to the Hands-on using Jupyter notebooks

Part 2 – Descriptive Statistics and Data Analytics

01

Descriptive Statistics

02

Data Visualization using Matplotlib

03

Data Analytics and Visualization using Seaborn

04

Understanding basic KPIs

Exploratory Data Analysis

Collect the data and gain the domain knowledge.

Confirm data types and their probabilities.

Measures of central tendency: mean, median, mode.

Measures of dispersion: variance, std deviation, range.

Skewness, right & left kurtosis, thinner peak, wider peak

Graphical representation: histogram, boxplot, barplot etc.

Descriptive Statistics

Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data.

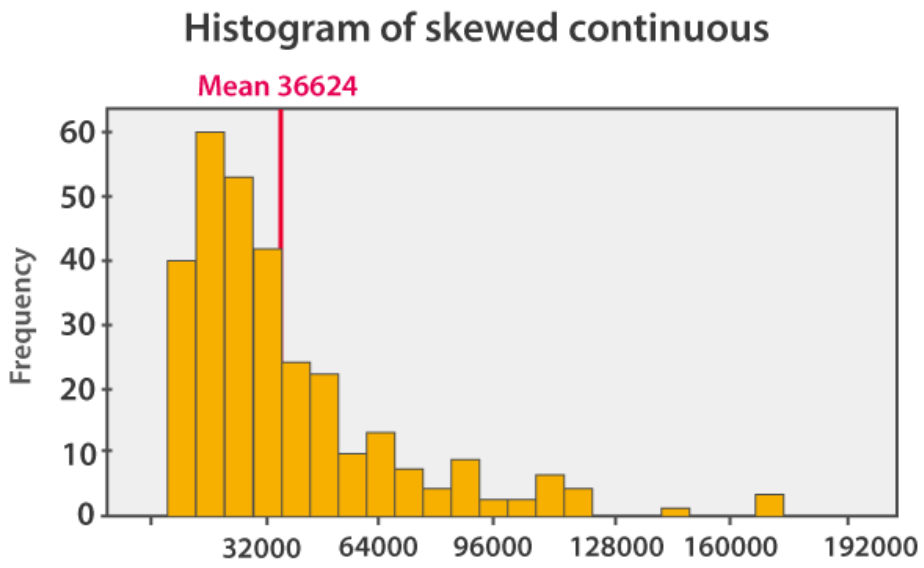
- Central Tendency - Mean, Median, Mode
- Dispersion – Range, Variance, Standard Deviation
- Frequency - Count, Percent, Frequency
- Position - Percentile Ranks, Quartile Ranks

Examples of descriptive statistics

Central Tendency - Mean

The mean represents the average value of the dataset. It can be calculated as the sum of all the values in the dataset divided by the number of values. In general, it is considered as the arithmetic mean.

$$\frac{x_1 + x_2 + .. + x_n}{n}$$



Central Tendency - Median

- Median is the middle value of the dataset in which the dataset is arranged in the ascending order or in descending order.

Median odd
23
21
18
16
15
13
12
10
9
7
6
5
2

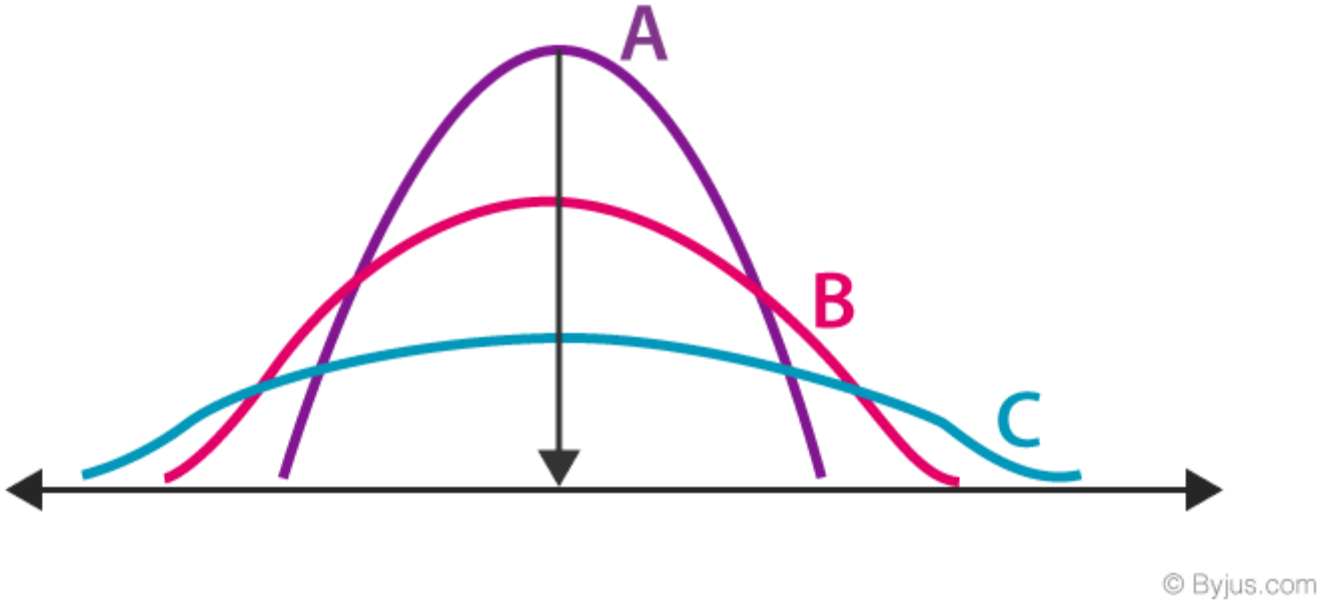
Central Tendency - Mode

- The mode represents the frequently occurring value in the dataset. Sometimes the dataset may contain multiple modes and, in some cases, it does not contain any mode at all.

Mode
5
5
5
4
4
3
2
2
1

What is dispersion?

- Dispersion is the state of getting dispersed or spread. Statistical dispersion means the extent to which a numerical data is likely to vary about an average value. In other words, dispersion helps to understand the distribution of the data.



Dispersion - Variance

- **Variance** is the expected value of the squared variation of a random variable from its mean value, in probability and statistics. Informally, variance estimates how far a set of numbers (random) are spread out from their mean value.
- Variance is a measure of how data points differ from the mean. According to Layman, a variance is a measure of how far a set of data (numbers) are spread out from their mean (average) value.

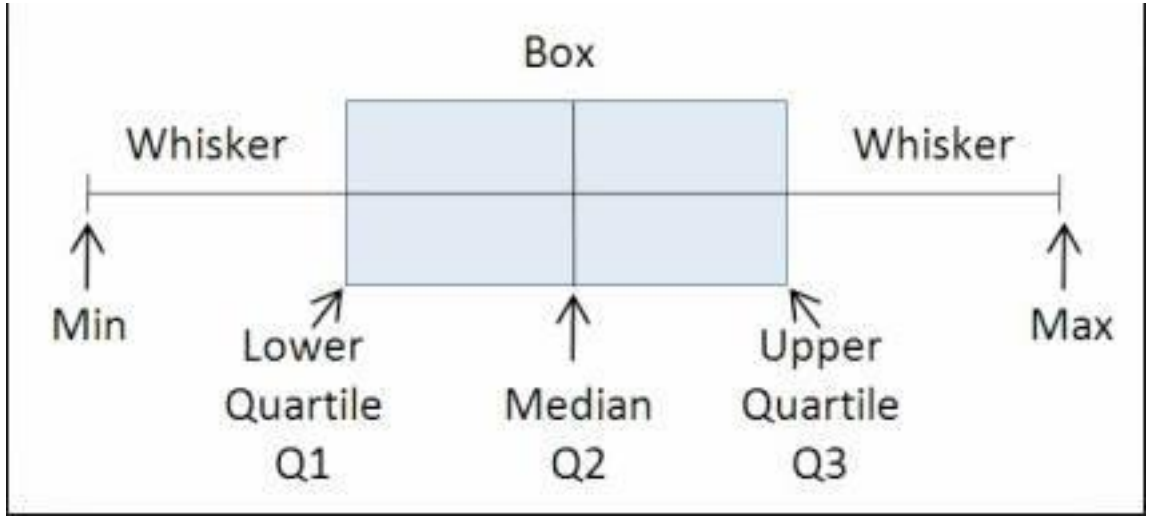
$$\mathbf{Var\ (X) = E[(X - \mu)^2]}$$

Dispersion – Standard Deviation

- Standard Deviation is the positive square root of the variance.
- Standard Deviation is a measure of how spread out the data is. Its formula is simple; it is the square root of the variance for that data set. It’s represented by the Greek symbol sigma (σ).

Quartiles

- The quartiles are values that divide a list of numbers into quarters.



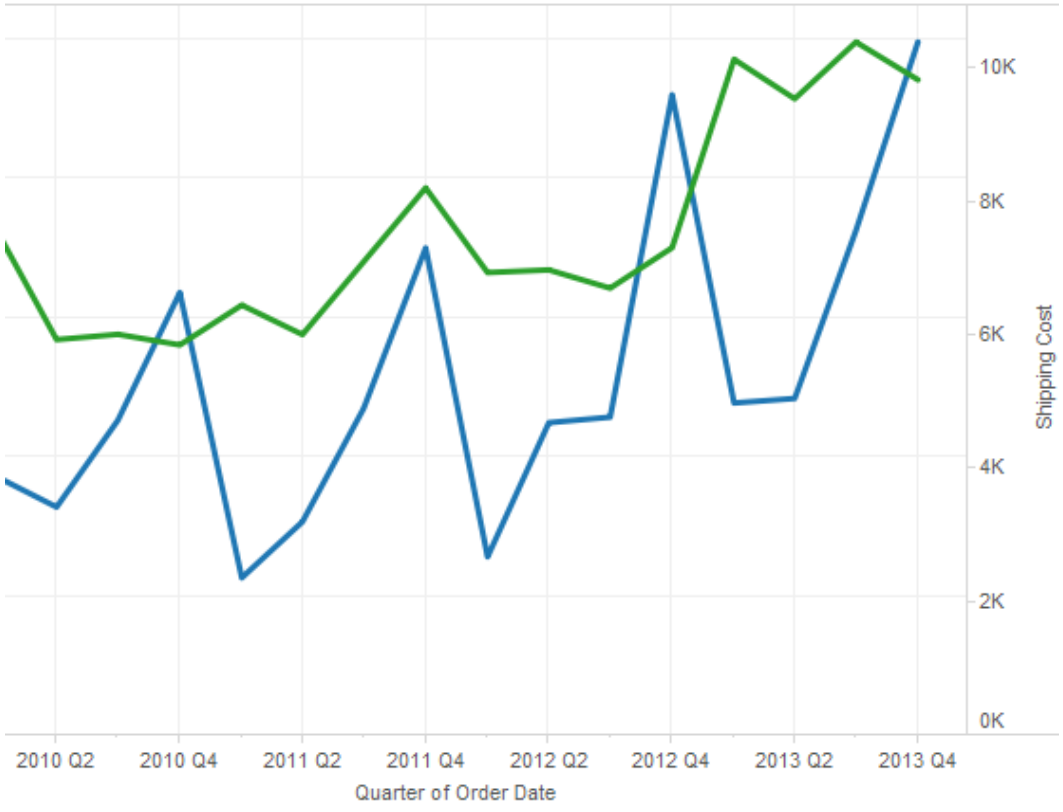
Visual Statistics

Let’s see some of the different types of charts

Types of Data Visualization

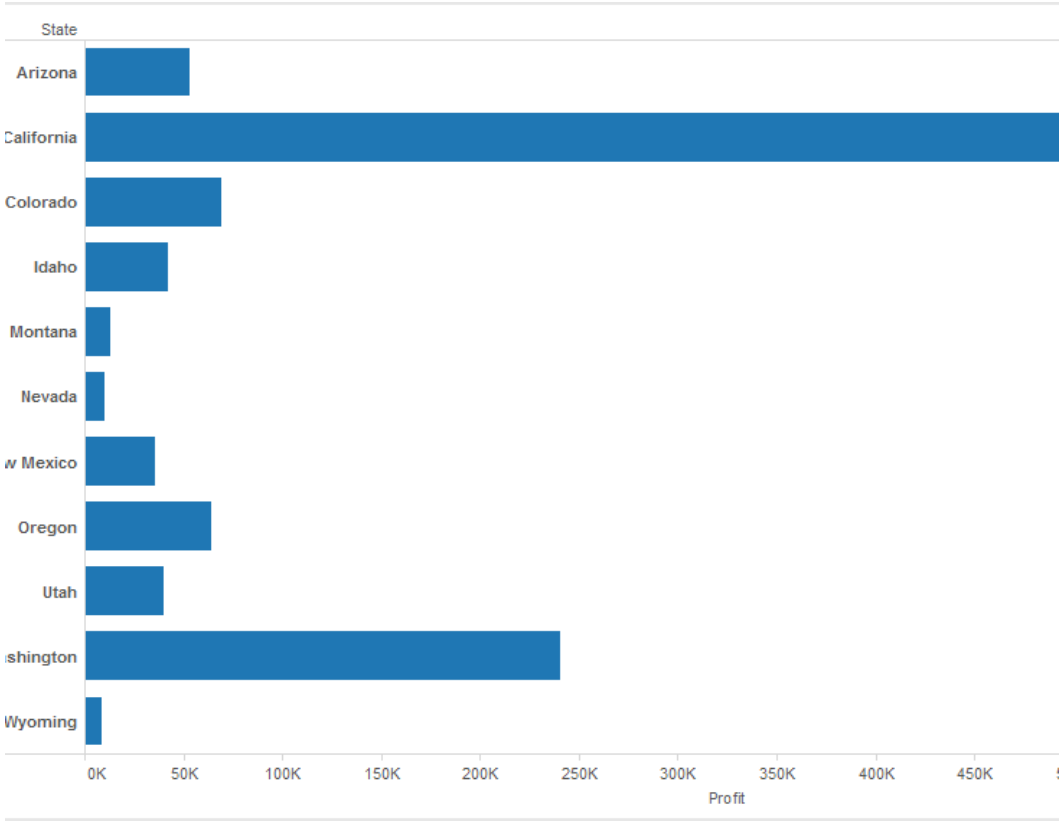
- Line Charts
- Bar Charts
- Pie Charts
- Polar Charts
- Area Charts
- Scatter Charts
- Scatter Maps (showing geographical data)
- Bubble Charts
- Funnels Charts
- Radar Charts
- Tree Maps
- Sandburst Charts
- Numeric / Gauge Indicators

Line Charts

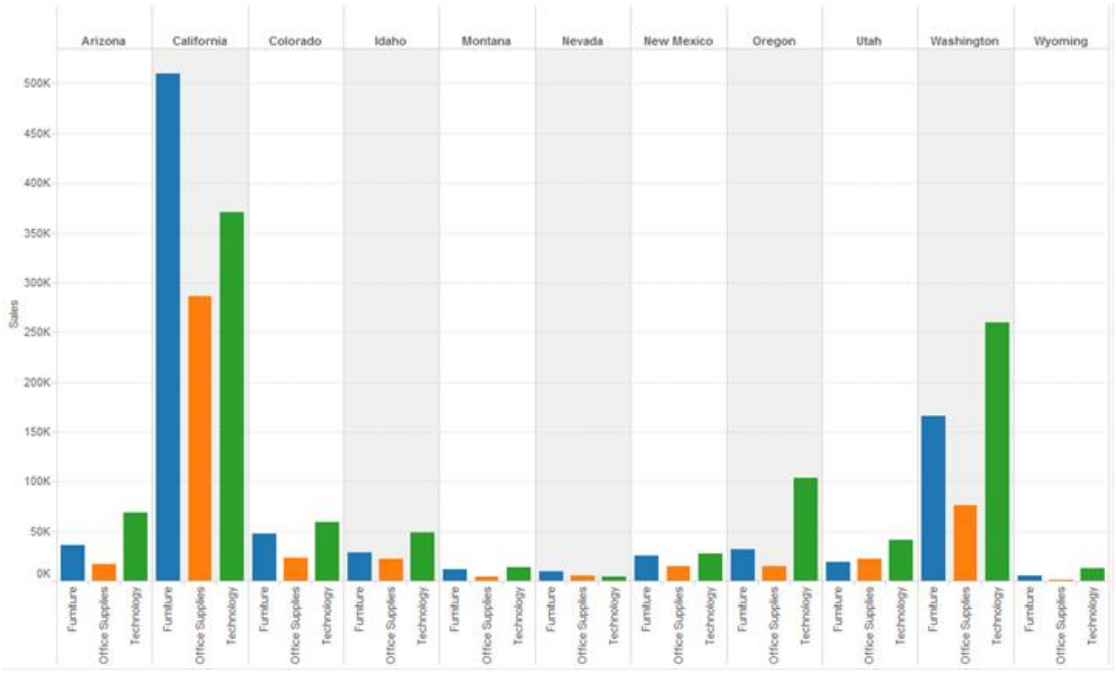


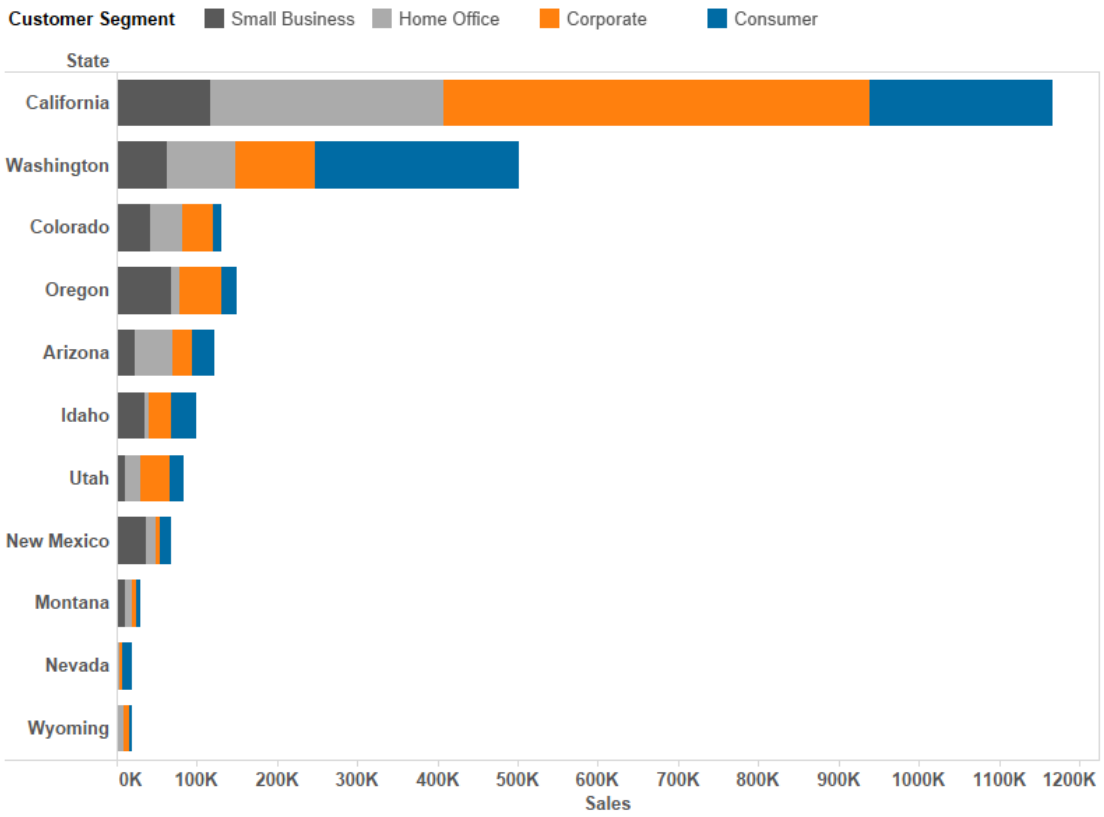
- Line charts are predominately used to concisely represent trends over a period.
- It connects a series of data points with a single, continuous line.

Bar Charts

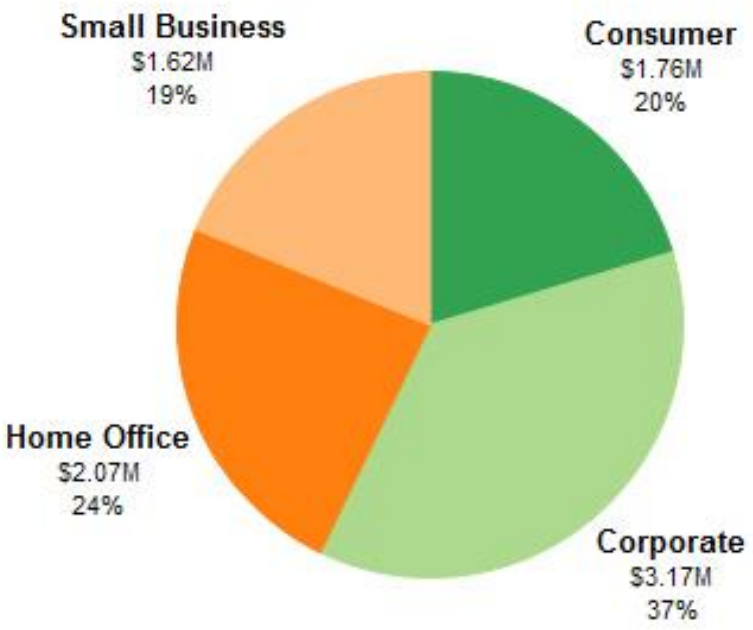


- Bar charts are used to represent categorically data using rectangular bars.
- Bar charts can be plotted vertically or horizontally.


Arizona
California
Colorado
Idaho
Montana
Nevada
New Mexico
Oregon
Utah
Washington
Wyoming
Sales
500K
450K
400K
350K
300K
250K
200K
150K
100K
50K
0K
Furniture
Office Supplies
Technology

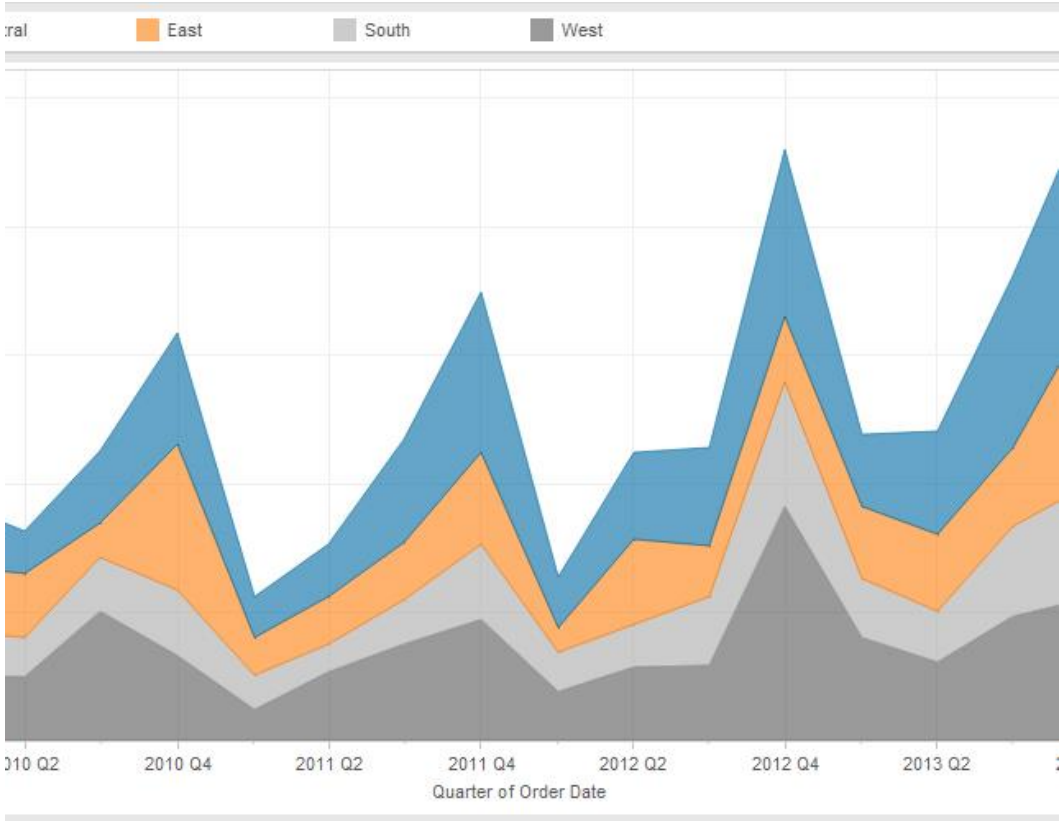

Customer Segment
Small Business
Home Office
Corporate
Consumer
State
California
Washington
Colorado
Oregon
Arizona
Idaho
Utah
New Mexico
Montana
Nevada
Wyoming
0K
100K
200K
300K
400K
500K
600K
700K
800K
900K
1000K
1100K
1200K
Sales

Pie Charts



- Pie charts show the share of each value as part of a whole.
- It uses pie slices to represent the relative sizes of data.
- Proportions are clearly demonstrated using pie charts.

Area Charts

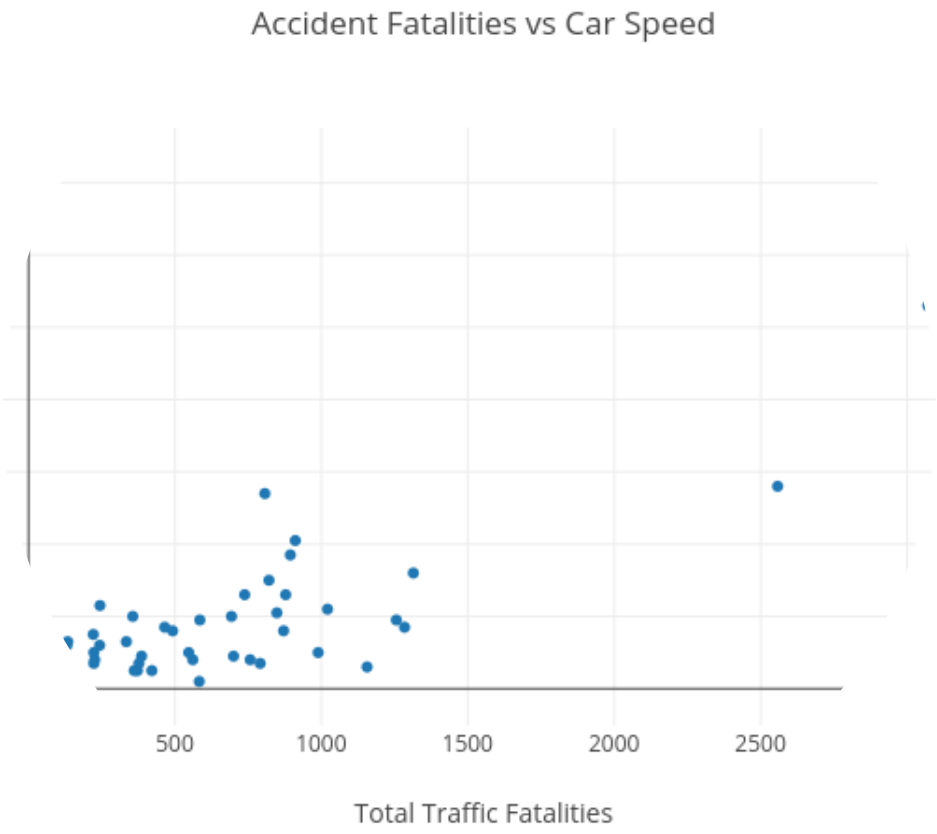


- It is used to display quantitative data.
- Through these charts its easier to understand the overall proportion and volume taken by each category.

Scatter Charts

- Also known as scatter graph, scatter plot or correlation chart, scatter charts are used to visualize the distribution of and relationship between two variables.

- It uses dots to represent values for two different numeric variables.



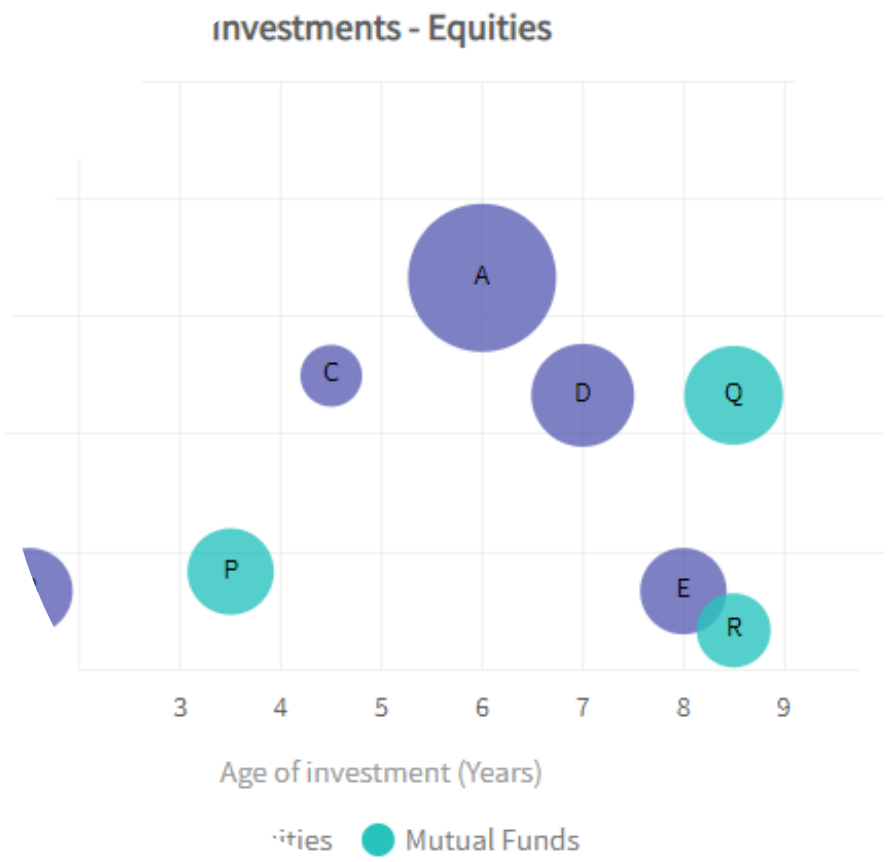
Scatter Maps

- When the geographical coordinates - latitude and longitude - are used as the variables to plot the points on a map, we get a scatter map.

Bubble Charts

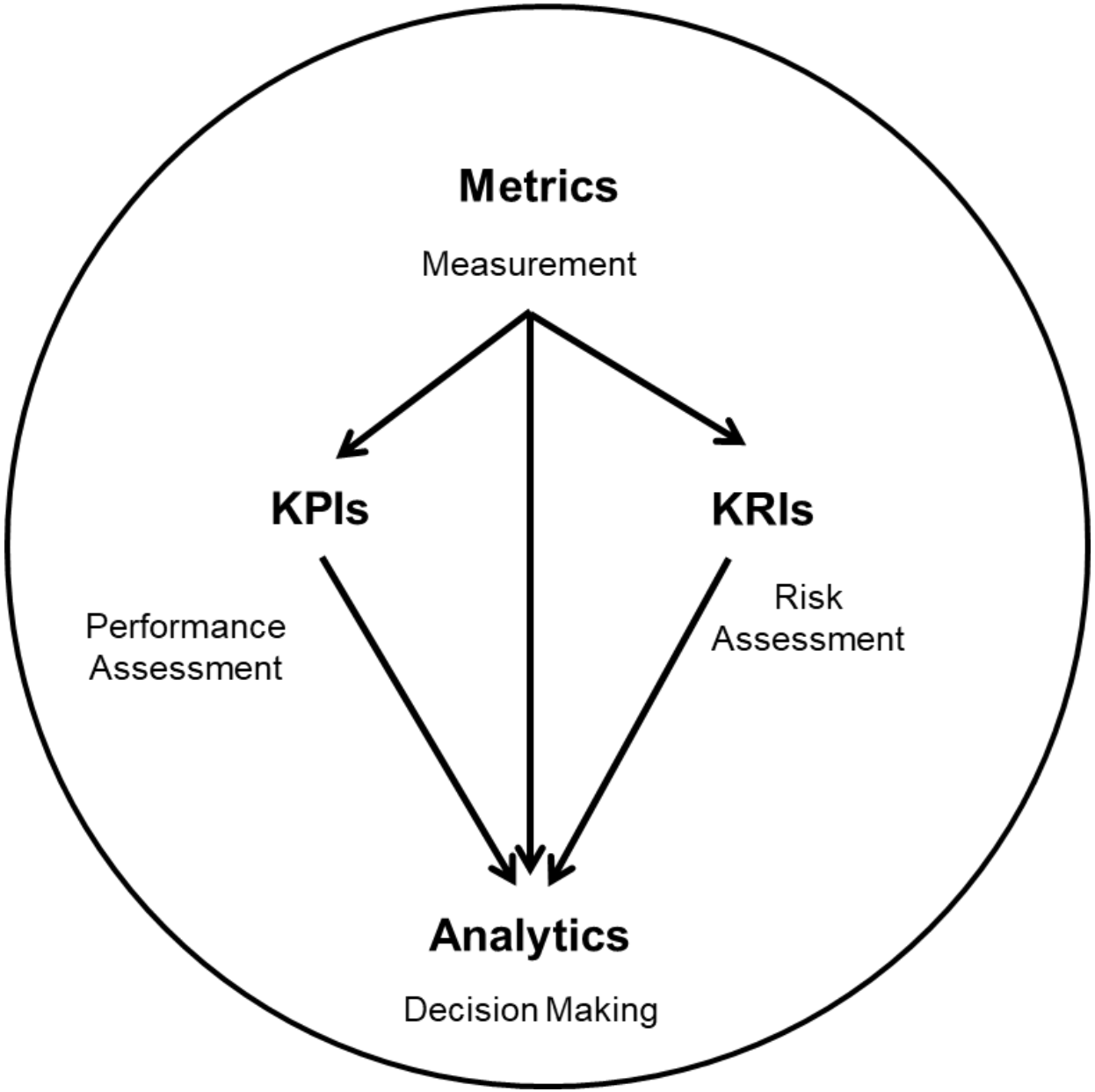
- Bubble chart is a variation of a scatter chart where instead of points, there are bubbles with diameters proportional to the data it is representing.

- It represents three dimensions of data.



What are KPIs?

A **Key Performance Indicator** (**KPI**) is a measurable value that demonstrates how effectively a company is achieving key business objectives. Organizations use **KPIs** to evaluate their success at reaching targets. *(Source: Klipfolio)*



METRICS – KPIs- KRIs - ANALYTICS

Examples of KPIs

Sales KPIs

- Monthly sales growth
- Cost per lead by each channel

Financial KPIs

- Net profit margin
- Resource utilization

Project Management KPIs

- Project resource utilization
- % of overdue project tasks

Why are KPIs important?

1

Effective company key performance indicators (KPIs) guide a business on the journey towards its strategic goals.

2

A good KPI should act as a compass: a measurement of where your business is, relative to where it has come from and where it is going.

3

KPIs translate your business strategy into manageable, operational actions, based on the data you collect and monitor.

Benefits of Using KPIs

Increases management awareness

Focuses attention on improvement opportunities

- **Increasing Cash Flow**
- **Improving Clinical Quality**
- **Reducing Costs**
- **Identifying Problem Areas**
- **Benchmarking**
- **Illustrating Trends**
- **Scoring Performance**
- **Reducing Denials**
- **Developing Consistent Processes and Outcomes**
- **Developing “Best Practices”**
- **Improving / Accelerating Management Reporting**
- **Monitoring Staffing Levels**

Financial Services

- Customer targeting/engagement
- Improved risk management
- Fraud detection in real-time

Retail & CPG

- Multi-channel sales analysis & optimization
- Customer behaviour modeling
- Real-time recommendation engines

Transportation

- Consumers choose time of home deliveries
- Fleet vehicle maintenance optimization
- Making logistics and fuel consumption less dependent on weather and traffic

E-commerce

- Analyze internet behavior and buying patterns
- Digital asset piracy

Telecommunications

- Customer churn & experience analysis
- Network service quality/predictive maintenance via sensor data

Utilities

- Service Quality Optimization
- Weather impact analysis on power generation
- Smart meter data analysis

Call Centers

- On-the-fly offer prompting
- Improved consumer experience
- Compliance verification

IT

- Network analysis & optimization
- Application log analysis (performance, threats, optimization

Healthcare

- E-Prescriptions
- Remote Patient Monitoring

Source: bigdata4analytics

Basic Visualization Rules

The first step is to choose the **appropriate** plot type.

Second, when we choose your type of plot, one of the most important things is to **label your axis**.

Third, we can add a **title** to make our plot more **informative**.

Fourth, add **labels** for different categories when needed.

Five, optionally we can add a text or an arrow at **interesting data points**.

Six, in some cases we can use some **sizes** and **colors** of the data to make the plot more informative.

Let’s dive straight to the Hands-on using Jupyter notebooks

Sample Question 1

- You have been given a dataset with some features including the **SalesPrice** of the house. You do not have the business knowledge pertaining to the dataset, but would like to find out which are the features which affect the SalesPrice of the house. Which of the following techniques would you use?

  - Correlation Analysis
  - Plotting Bar charts
  - Log Plots
  - Data Cleaning

Correlation Analysis.
We would use the above because the correlation analysis can give us the strength between various features and the SalesPrice based on the relationship between them. If there is a strong relationship the correlation value will be closer to +-1 else it will be closer to 0.

Sample Question 2

- Which technique is most suitable to find anomalies?
  - Box plots
  - Bar plots
  - Correlation Analysis
  - Pair plots

Box Plots
The reason for the above technique is because box plots along with plotting 2-3 dimensions of data shows where the outliers lie with respect to those dimensions. Hence it is easier to visualize and isolate them.

Analytics in the Industry

Landscape of the industry

Analytics usage in the industry

In 2015, **17 percent** of companies adopted big data analytics, by 2017, **53 percent** of companies are adopting big data analytics (Forbes, 2017)

90 **percent** of enterprise analytics and business professionals currently say data and analytics are key to their organization's digital transformation initiatives. —MicroStrategy *2018 Global State of Enterprise Analytics Report*

Analytics usage in the industry

Data-driven organizations are 23 times more likely to acquire customers, six times as likely to retain customers, and 19 times as likely to be profitable as a result. —McKinsey Global Institute

By 2020, there will be 2.7 million job postings for data science and analytics roles. —BHEF and PwC America's Data Science and Analytics Talent: The Case for Action Report

Business Cases in Traditional Analytics

Banks - Credit loan

- Credit Risk Analysis
  - To estimate the costs associated with a loan
  - To see if the bank borrower could potentially renege on its credit loan

- Banks would typically hire Credit analysts to process the loan applications.
  - degree in finance, accounting, business administration or economics (statistics background)
- **Banks will lose money on bad loans!**

https://analyticstraining.com/understanding-credit-risk-analytics/

Microfinance

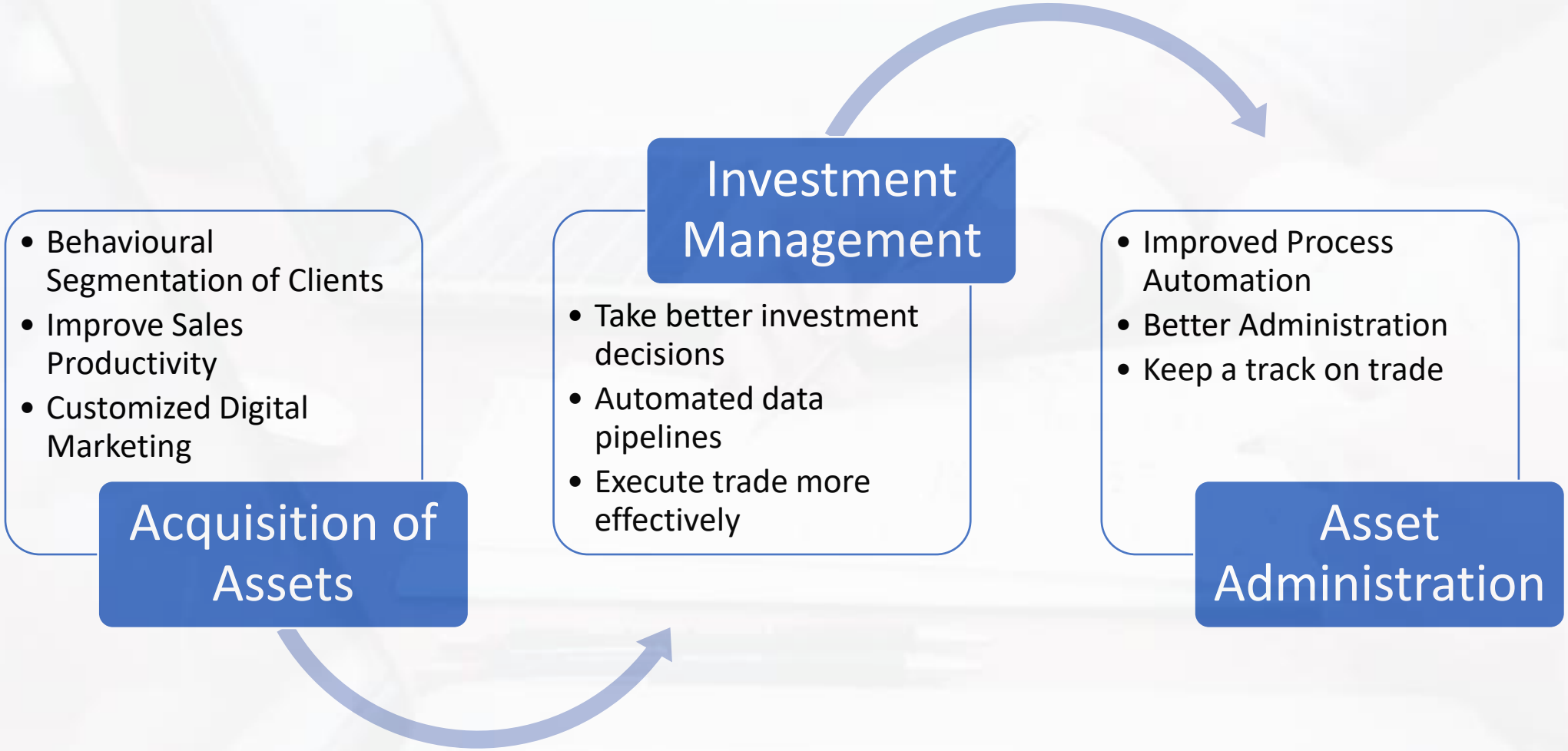
Will a customer default?

Study behavioral patterns to determine whether a Customer is likely to default. With the KYC in place and data collated from other places, analytics performed on this can come up with patterns with respect to Customer behavior.
This will **reduce delinquencies** to a great extent.

Profitability based Analytics

Organizations can study the customer relationship with the institute and derive a CLTV and this can help in projecting the future cashflows.

Asset Management



Test Instructions (Canvas)

- 2 Parts to the test **(Open Book, 70% to be competent)**
  - Written Test (30 mins)
    - 4 questions
  - Practical 90 mins
    - 4 questions (with subparts)
    - Upload a **PDF file** for each question. Extract your Jupyter notebook file as a PDF using print to PDF option.
- After both the tests are done – Let me know on private message on Zoom
- Competency Acknowledgement Process/Email
- Oral Recovery (If not competent)
- Trainer Evaluation
- End

Thank you

Quick Recap