Motivation

- The world's technological per-capita capacity to store information doubled every 40 months As of 2012, 2.5 exabytes (2.5×1018) of data/day Relational database management systems and desktop statistics and visualization packages often have difficulty handling data.

Why data Analytics

The Power of Data Analytics

- Data Analytics can bring “big values” to our life in almost every aspects.
- Technologically, Data analytics is bringing about changes in our lives because it allows diverse and heterogeneous data to be fully integrated and analyzed to help us make decisions.
- Today, with the Data Analytics technology, thousands of data from seemingly unrelated areas can help support important decisions. This is the power of Data Analytics

DATA?

In computing, data is information that has been translated into a form that is efficient for movement or processing.Data can exist in a variety of forms as numbers or text on pieces of paper, as bits and bytes stored in electronic memory, or as facts stored in a person's mind.

ANALYTICS?

Analytics is the discovery, interpretation, and communication of meaningful patterns in data and applying those patterns towards effective decision making .Analytics is an encompassing and multidimensional field that uses mathematics, statistics, predictive modeling and machine learning techniques to find meaningful patterns and knowledge in recorded data.

What is data analytics?

- Data Analytics refers to the techniques to analyse data to enhance productivity and business gain.

Business Administration

+

Exploratory Data Analysis

=

Growth in Business

Who is a Data Analyst?

Data Analyst Skills

Statistics

Data Cleaning

EDA

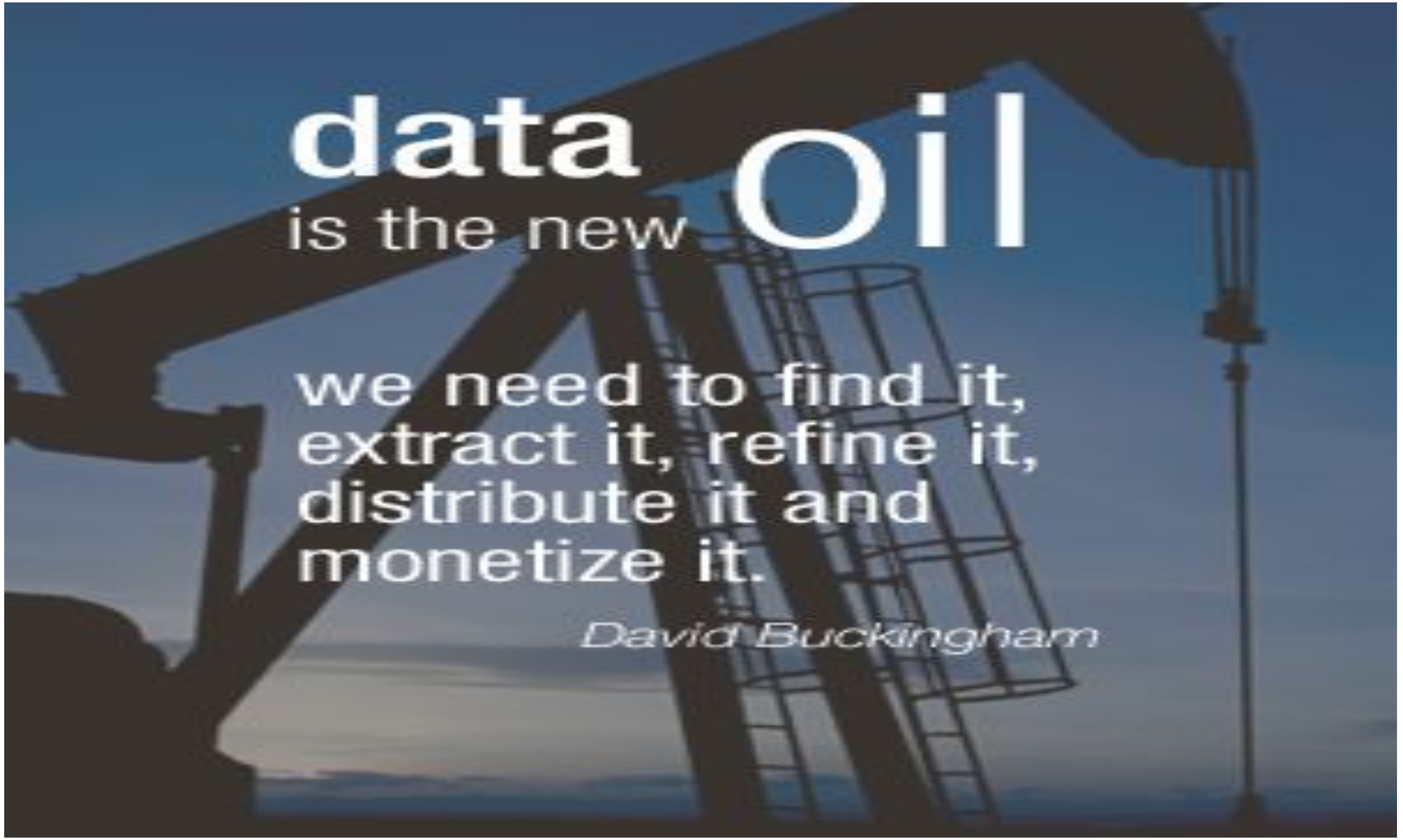
Data Visualization

Machine Learning

BONUS

What is data, and why is it important?

- 1) Data helps to make better decisions
- 2) Data helps you solve problems
- 3) Data helps you understand performance
- 4) Data helps you improve processes
- 5) Data helps you understand consumers


data
is the new oil
we need to find it,
extract it, refine it,
distribute it and
monetize it.
David Buckingham

TYPES OF DATA

- Structured data
- Unstructured data
- Semistructured data

Unstructured data

The university has 5600 students.
John's ID is number 1, he is 18 years old and already holds a B.Sc. degree.
David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

Semi-structured data

```
<University>
  <Student ID="1">
    <Name>John</Name>
    <Age>18</Age>
    <Degree>B.Sc.</Degree>
  </Student>
  <Student ID="2">
    <Name>David</Name>
    <Age>31</Age>
    <Degree>Ph.D. </Degree>
 </Student>
....
</University>
```

Structured data

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

Sources of Data

- Personal data
- Transactional data
- Web data
- Sensor data

Inaccuracies in Data

- Initial Data Entry
  - Data entry Mistake
  - Flawed Data Entry Processes
- The NULL Problem
- **Deliberate Errors**
  - They do not know the correct information.
  - They do not want you to know the correct information.
  - They get a benefit from entering the wrong information.
- System Problem

Uses in Data

1) Data in business
2)Data in healthcare
3) Data in media and entertainment
4) Data in Transportation
5) Data in Banking

What is DATA analytics?

Data analysis is a process of inspecting, cleansing, transforming, and modeling data. Data analytics refers to qualitative and quantitative techniques and processes used to enhance productivity and business gain

Why Data Analytics

Data Analytics is needed in Business to Consumer applications (B2C). Organisations collect data that they have gathered from customers, businesses, economy and practical experience. Data is then processed after gathering and is categorised as per the requirement and analysis is done to study purchase patterns and etc.


Cost reduction
BIG DATA ANALYTICS
Faster, better decision making
New products & services

The process of Data Analysis

Analysis refers to breaking a whole into its separate components for individual examination. Data analysis is a process for obtaining raw data and converting it into information useful for decision-making by users. There are several phases that can be distinguished :Data requirements,

Data collection ,Data processing ,Data cleaning, Exploratory data analysis,

Modeling and algorithms , Data product ,Communication

Scope of Data Analytics

Bright future of data analytics, many professionals and students are interested in a career in data analytics. Any person who likes to work on numbers, has a logical thinking, can understand figures and can turn them into actionable insights, has a good future in this field. A proper training of the tools of data analytics would be required to begin with. Since it is a course that requires effort to learn and get certified, there is always dearth of qualified professionals. Being a relatively new field also, the demand for such professionals is more than the current supply. Higher demand also means higher salaries.

Importance Data Analytics

- Predict customer trends and behaviours
- Analyse, interpret and deliver data in meaningful ways
- Increase business productivity
- Drive effective decision-making

Types of Analytics

Prescriptive Analytics

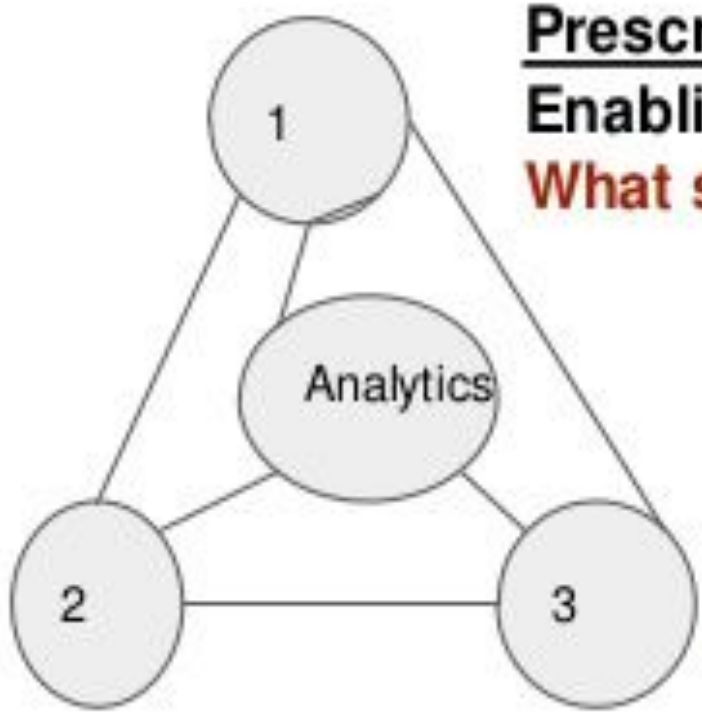
Enabling smart decisions based on data

What should we do

Predictive Analytics

predicting the future based on historical patterns

What could happen?



Descriptive Analytics

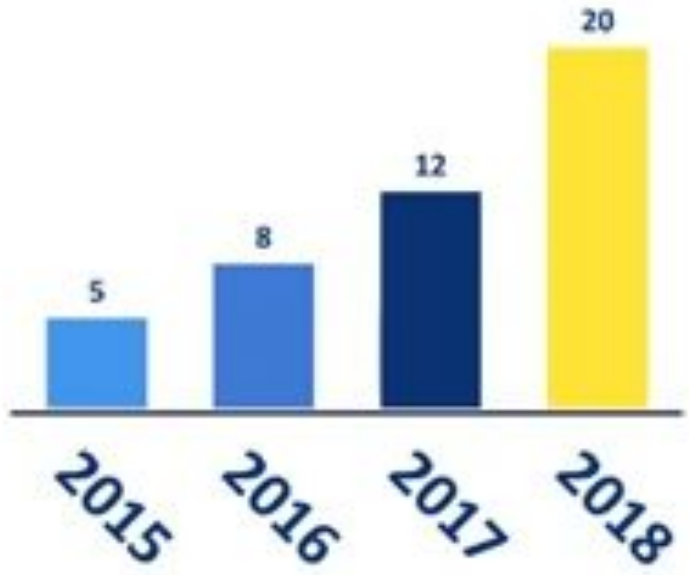
Mining data to provide business insights?

What has happened?

DESCRIPTIVE ANALYTICS

- It is the preliminary stage of Data processing.
- It creates foundation for further analysis andunderstanding.


GOOGLE
Descriptive Analytics Methods
Search
Data Aggregation Methods
Data Mining Methods



Help managers to make informed strategic business decisions based on historical data

Why is there a need for descriptive analytics?

PREDICTIVE ANALYTICS

All about the outcome

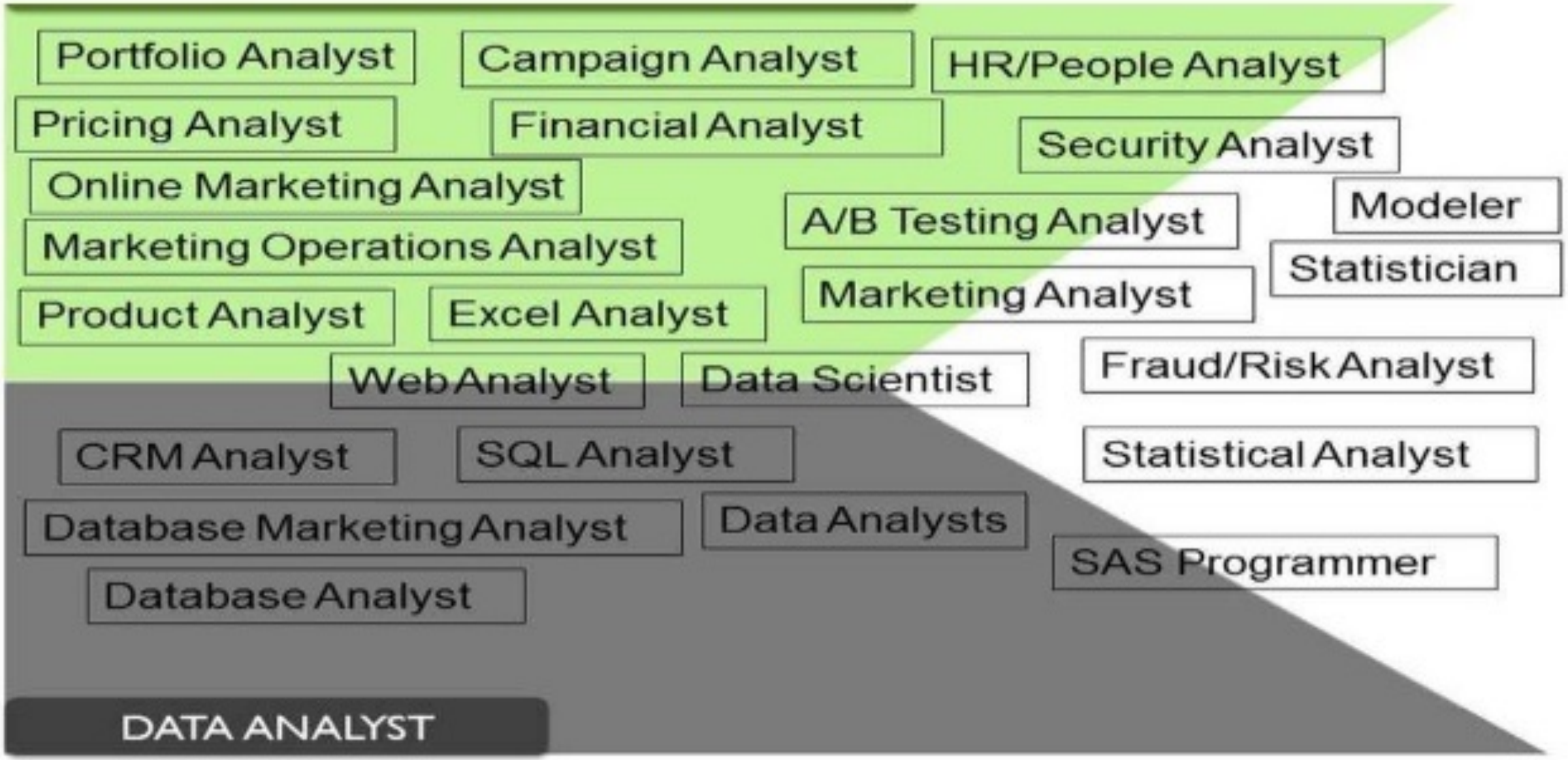
- Predicting an outcome
- Uses historical information to predict a solution
- Can incorporate elements of artificial intelligence

PRESCRIPTIVE ANALYTICS

All about the action

- Uses a predictive model to generate an action
- Reaches the right customer with the right message or action at the right time
- Generates an action that can automatically correct an anomaly

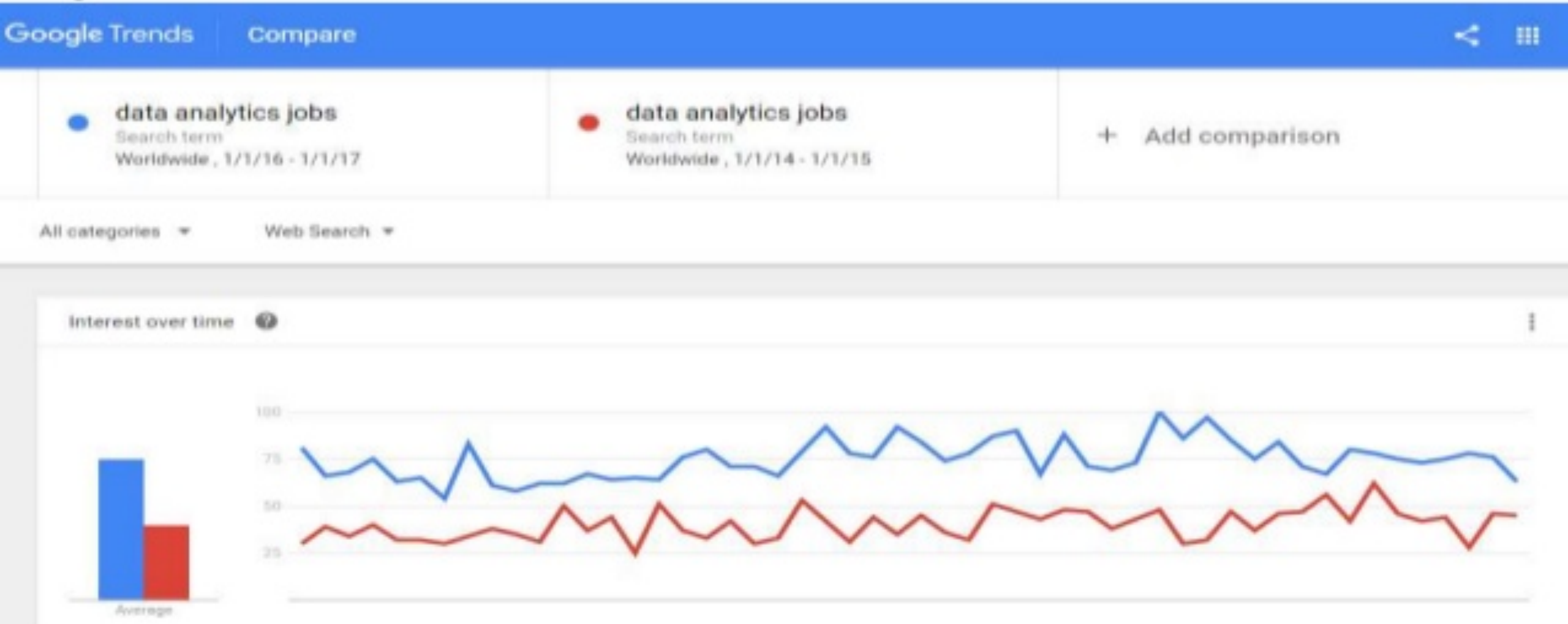
Data analytics job title



Basic Skills required to start your career in data analytics

- How to set up data structure?
- How to create data visualizations
- Knowledge with database languages like SQL, MySQL
- knowledge of big data tools like Hive or Pig
- Know statistical programming languages like R or Python
- Understanding of machine learning tools & techniques

Going by the statistics, by 2020, about 1.7 megabytes of new information will be created every second for every human! Well, that's huge. Very huge. Companies will need more and more data specialists to analysis and manage the data generated.



Demand of Data Analysis Jobs

Data analysis jobs are everywhere and they are bound to increase! Here are some facts and figures to highlight this:

"2.5 billion gigabytes (GB) of data was generated every day in 2012. (IBM) International Business Machines"

What recruiters look for in applicants

Problem Solving Skills: When working with complex sets of data, companies rely on analysts to interpret the numbers and figures to find solutions to their problems. Your primary job is to read between the numbers and datasets to find the answers that inexperienced analysts can't see. You are who they turn to when they need a complex problem solved with data, and you could potentially shape the future of the company.

Analytical Mind: This goes hand-in-hand with the problem-solving skills needed. An optimal candidate for any analytics position must have a mind that naturally looks for

answers and connections between data sets. This is incredibly useful, especially when handling large sets of data. You must be able to decipher and make connections that nobody else can.

Maths and Statistic Skills:It goes without saying that if you want to be an effective data analyst or scientist, you must be able to do the math to analyze and interpret the data. Although a majority of calculations are completed with computer programs, a solid foundation and understanding of mathematics or statistics will take you far in this field.

Communication (both oral and written): Once you find solutions and make connections using the data you won't be keeping it to yourself. You must be able to succinctly and accurately explain sophisticated mathematical and statistical principles that other departments can understand.Communication skills go a long way in any career and data analysis is no exception.

Teamwork Abilities: More times than not, you never work alone. You will be a part of a team of data specialists, and it is vital to the success of the team and organization that you can all work together to solve complex problems.

Skill is required for Data analytics ?

1.) Analytical Skills

2.) Numeracy Skills

3.) Technical and Computer Skills

4.) Attention to Details

5.) Business Skills

6.) Communication Skills

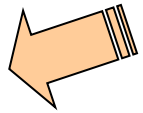
CAREER

Data analysis is a rapidly growing field and highly skilled analysts in increased demand across all sectors. This is evident from the average salary of a data analyst in India. This implies that you would find many opportunities but you will still have to be outstanding and exhibit excellent data analytics skills to be successful as a data analyst.

Top companies hiring for business Analytics

Getting to Know Your Data

- Data Objects and Attribute Types



- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary

Types of Data Sets

- Record
  - Relational records
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data: text documents: term-frequency vector
  - Transaction data
- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- Ordered
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data
- Spatial, image and multimedia:
  - Spatial data: maps
  - Image data:
  - Video data:

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Important Characteristics of Structured Data

- Dimensionality
  - Curse of dimensionality
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale
- Distribution
  - Centrality and dispersion

Data Objects

- Data sets are made up of data objects.
- A **data object** represents an entity.
- Examples:
  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses
- Also called *samples , examples, instances, data points, objects, tuples*.
- Data objects are described by **attributes**.
- Database rows -> data objects; columns ->attributes.

Attributes

- **Attribute (**or **dimensions, features, variables**): a data field, representing a characteristic or feature of a data object.
  - *E.g., customer _ID, name, address*
- Types:
  - Nominal
  - Binary
  - Numeric: quantitative
    - Interval-scaled
    - Ratio-scaled

Attribute Types

- **Nominal:** categories, states, or "names of things"
  - *Hair_color* = {*auburn, black, blond, brown, grey, red, white*}
  - marital status, occupation, ID numbers, zip codes
- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)
- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size* = {*small, medium, large*}, grades, army rankings

Numeric Attribute Types

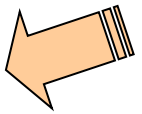
- Quantity (integer or real-valued)
- **Interval**
    - Measured on a scale of **equal-sized units**
    - Values have order
        - E.g., *temperature in C˚or F˚, calendar dates*
    - No true zero-point
- **Ratio**
    - Inherent **zero-point**
    - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).
        - e.g., *temperature in Kelvin, length, counts, monetary quantities*

Discrete vs. Continuous Attributes

- **Discrete Attribute**
  - Has only a finite or countably infinite set of values
    - E.g., zip codes, profession, or the set of words in a collection of documents
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes
- **Continuous Attribute**
  - Has real numbers as attribute values
    - E.g., temperature, height, or weight
  - Practically, real values can only be measured and represented using a finite number of digits
  - Continuous attributes are typically represented as floating-point variables

Chapter 2: Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary



Basic Statistical Descriptions of Data

- Motivation
  - To better understand the data: central tendency, variation and spread
- Data dispersion characteristics
  - median, max, min, quantiles, outliers, variance, etc.
- Numerical dimensions correspond to sorted intervals
  - Data dispersion: analyzed with multiple granularities of precision
  - Boxplot or quantile analysis on sorted intervals
- Dispersion analysis on computed measures
  - Folding measures into numerical dimensions
  - Boxplot or quantile analysis on the transformed cube

Measuring the Central Tendency

- Mean (algebraic measure) (sample vs. population):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

Note: *n* is sample size and *N* is population size.

  - Weighted arithmetic mean:
  - Trimmed mean: chopping extreme values

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- Median:
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Estimated by interpolation (for *grouped data*):

age	*frequency*
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

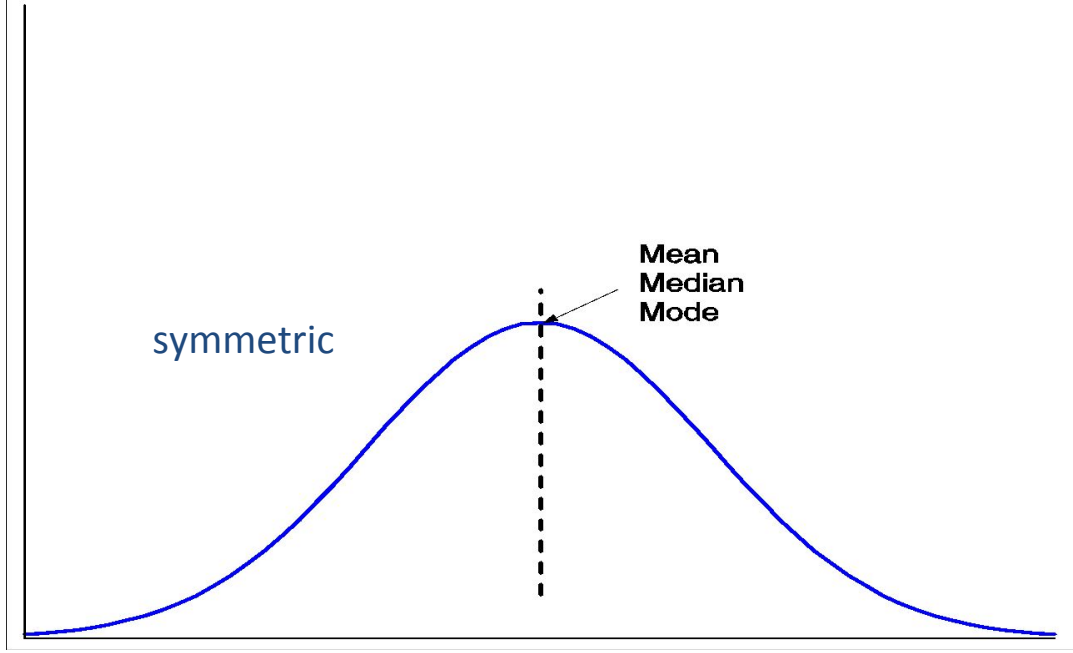
$$median = L_1 + (\frac{n/2 - (\sum freq)l}{freq_{median}})width$$

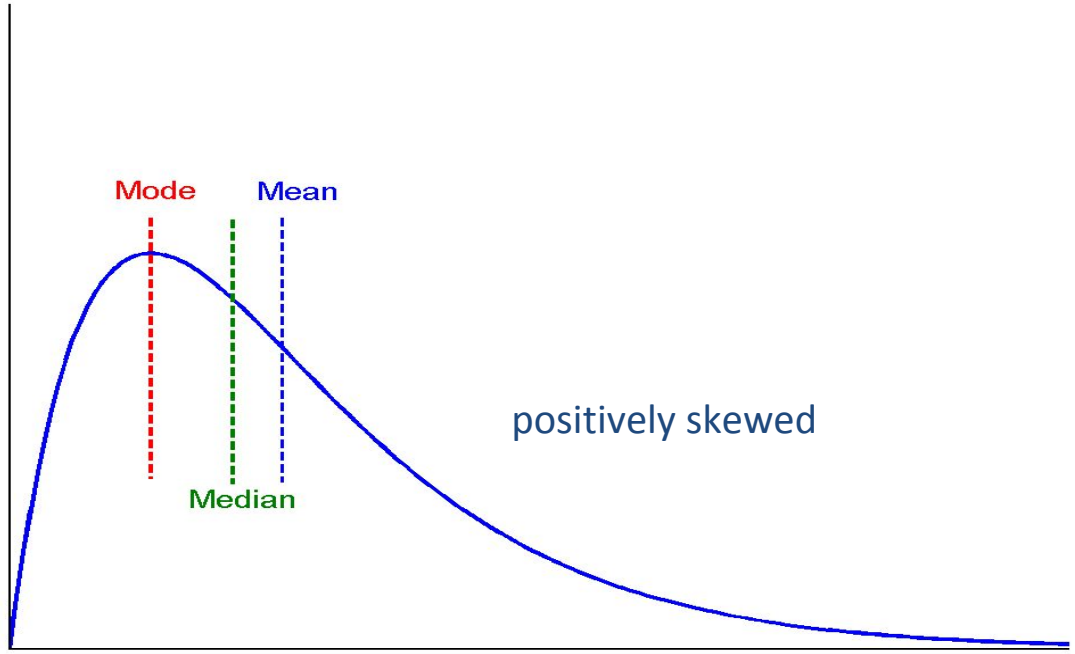
- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula:

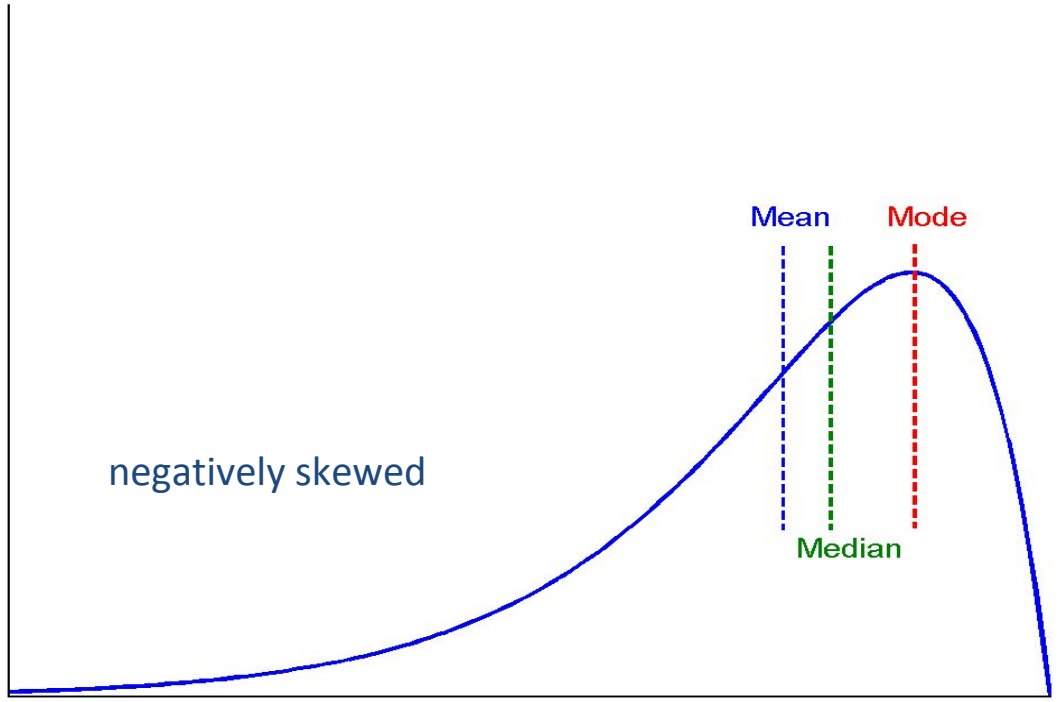
$$mean - mode = 3 \times (mean - median)$$

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data







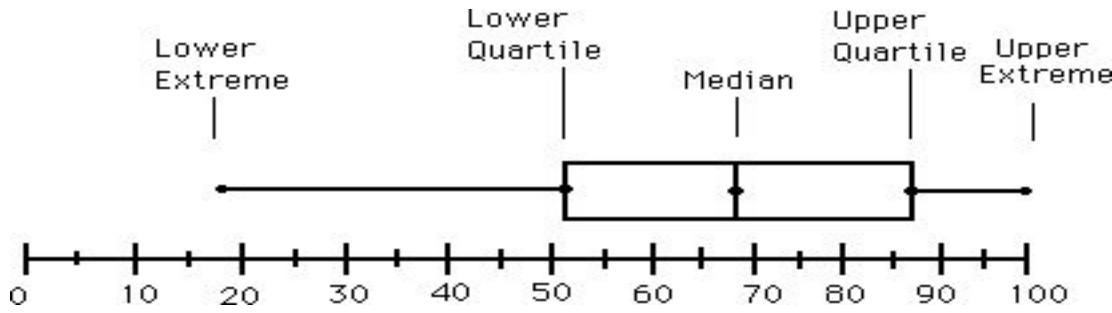
Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - **Quartiles**: Q_1 (25th percentile), Q_3 (75th percentile)
  - **Inter-quartile range**: IQR = $Q_3 - Q_1$
  - **Five number summary**: min, Q_1, median, Q_3, max
  - **Boxplot**: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
  - **Outlier**: usually, a value higher/lower than 1.5 x IQR
- Variance and standard deviation (*sample: s, population: σ)*
  - **Variance**: (algebraic, scalable computation)

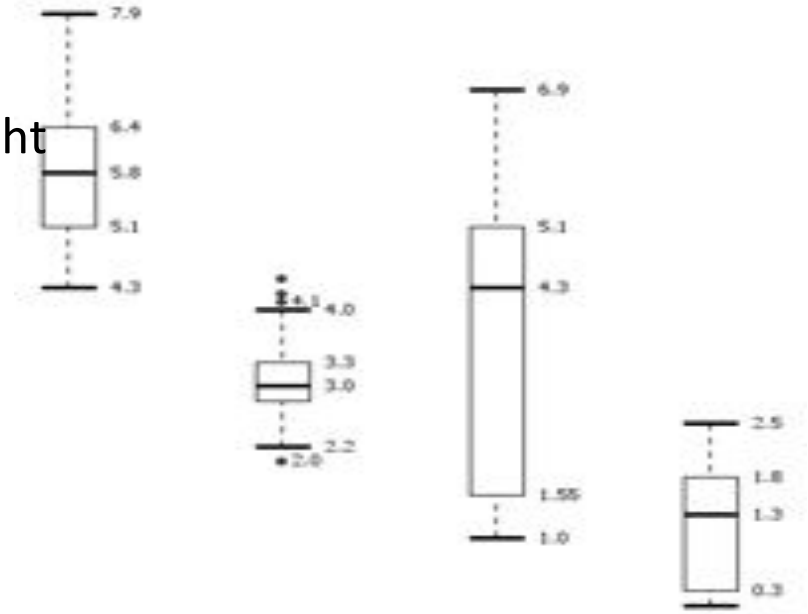
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right] \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^2 - \mu^2$$

  - **Standard deviation** *s (or σ)* is the square root of variance s^2 *(or* σ^2*)*

Boxplot Analysis



- **Five-number summary** of a distribution
  - Minimum, Q1, Median, Q3, Maximum
- **Boxplot**
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extended to Minimum and Maximum
  - Outliers: points beyond a specified outlier threshold, plotted individually



Visualization of Data Dispersion: 3-D Boxplots