# Introduction to Data Analytics

Dr. P. Kalpana, M.E., PhD.
Faculty of Mechanical Engineering
IIITDM Kancheepuram

# About Myself

- **Education**
  - Phd - Operations and Supply Chain Management - IIT Madras
  - M.E – Industrial Engineering - PSG College of Technology, Coimbatore
- **Industry**
  - Oct 2016-March 2019, Deputy Manager- Manufacturing and supply chain Analytics, Ford Motor Pvt Ltd, Chennai
  - Oct 2013-Sep 2016, Operations Research Consultant, Ramco Systems Ltd, Chennai
- **Teaching**
  - Sep 2007 - Dec 2008, Mookambigai College of Engineering, Trichy
- **Research Interests**
  - Operations Research, Supply chain Coordination, Logistics and distribution systems management, transport Network optimization, Scheduling, Forecasting, Inventory Management, Game theory, **IoT and Block chain in SCM, Advanced data analytics, Supply chain resilience, Intelligent transportation and logistics systems and cold supply chain management.**
- **Projects Handled**
  - Air craft tail allocation, Demand forecasting for Aircraft spare parts, Time table scheduling, Aircraft hanger scheduling, workforce scheduling for a health care industry, Shuttle optimization, floor space optimization, container load optimization etc.

## *Course outline*

- Introduction to Data analytics
- Statistical Techniques for Analytics
  - Descriptive statistics
  - Inferential statistics
- Exploratory data analytics
- Data visualization
- Regression models
- Machine learning Algorithms
- Tools used in Data analytics
  - python
  - Knime
  - Tableau
  - Excel
- Use cases

# *Objective of the course*

- To introduce conceptual understanding of several methodologies  using simple and practical examples
  - To understand importance of those techniques
  - Why to use a particular methodology or technique
  - How to use it correctly
  - How to interpret the result
- To know how to work with real data and choose right data analytic tool to solve the problem
- To know how to represent the data
- To explore various data analytic tools

## *Evaluation Pattern*

| | |
|---|---|
| ➢Regular Assessment | 30% (weeky/bi weekly) |
| ➢Assignment | 30% (At the end of each month) |
| ➢Final Assessment | 40% |

### ➢Course Policies

- ➢Maintain the minimum attendance requirement to write exams
- ➢Late submission of Assignments/Term paper are not entertained

# Overview

- ➢What is data, variable and measurement?
- ➢Difference between Analysis and Analytics
- ➢What is data analytics?
- ➢Types of data analytics
- ➢Role of analytics in business
- ➢Difference between data analytics and business analytics
- ➢Difference between statistics and data analytics
- ➢Difference between statistics and machine learning
- ➢Difference between data analytics and data scientist
- ➢Data Phrases in Technology
- ➢Top Tools in Data Analytics
- ➢Why python?

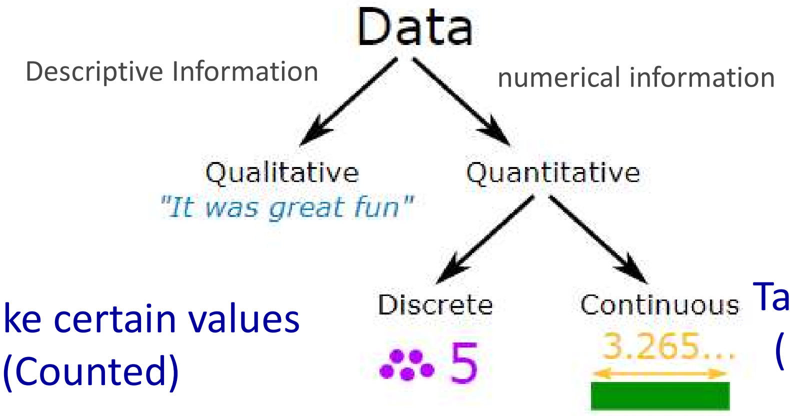
# *Variable, Measurement, Data*

- A variable is a specific characteristic (such as age or weight) of an individual or object, that is capable of taking different values.
- Measurement- A standard process, used to assign numbers to particular attributes or characteristics of a variable.
- Data- A recorded measurement

# What is data?

- A collection of facts
  - numbers, words, measurements, observations or just descriptions of things
  - Data is distinct pieces of information, usually formatted in a special way
- Forms of Data
  - Numbers, text on pieces of paper, as bits and bytes stored in electronic memory, or as facts stored in a person's mind.
- Data is the plural of *datum*, a single piece of information
- Software
  - Data and programs
  - Programs are collections of instructions for manipulating data

# *Classification of Data*



Data

Descriptive Information

numerical information

Qualitative

*"It was great fun"*

Quantitative

Discrete

5

Continuous

3.265...

only take certain values (Counted)

Take any value (Measured)

## What do we know about the Dog?

**Qualitative**:

- He is brown and black
- He has long hair
- He has lots of energy

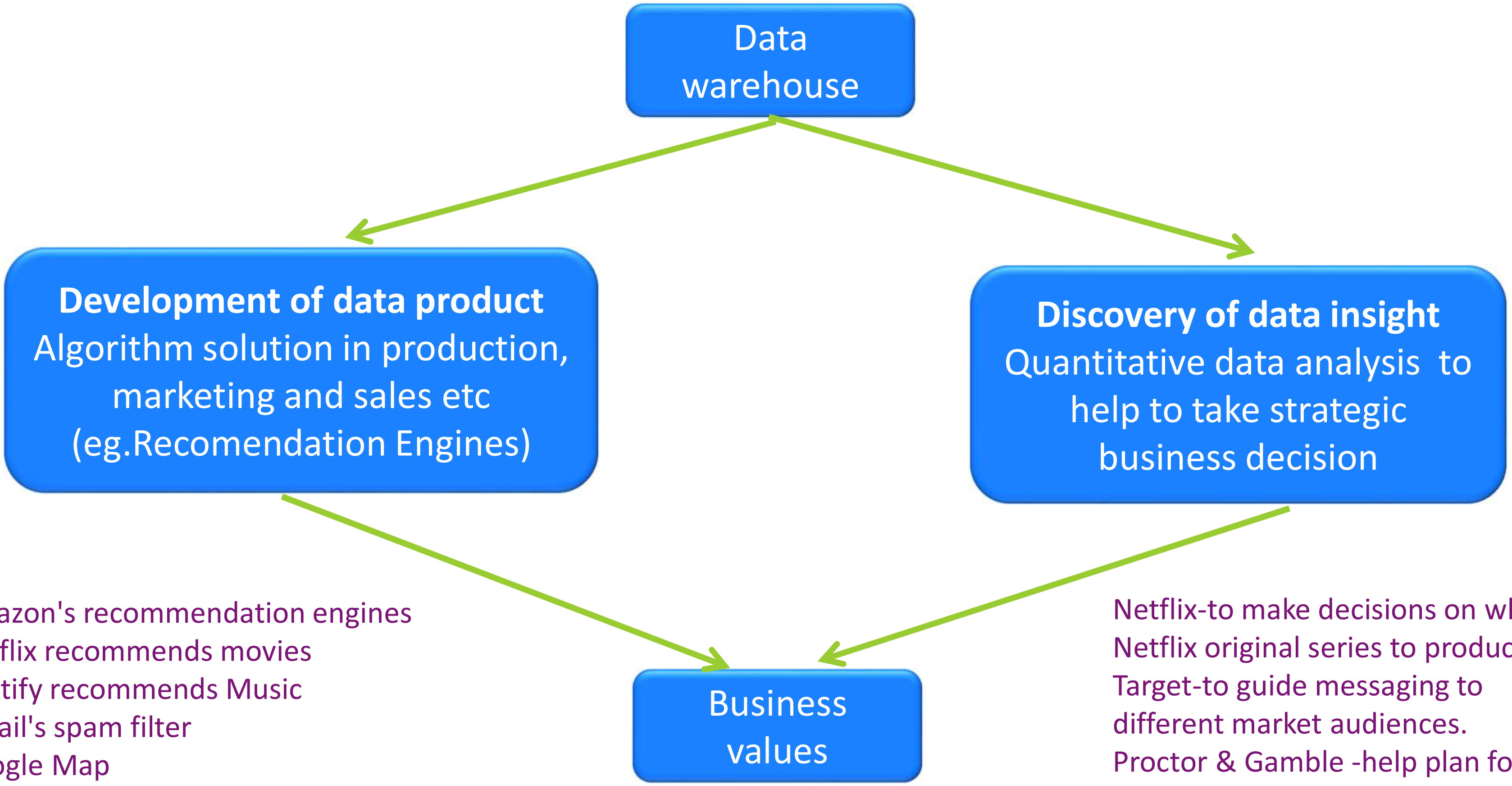
**Quantitative**:

- Discrete:
  - He has 4 legs
  - He has two eyes
- Continuous:
  - He weighs 25.5 kg
  - He is 565 mm tall

## *What is generating so much data*

- Data can be generated by
    - Humans
    - Machines
    - Humans and machines ( Facebook account, we have LinkedIn account)
- It can be generated anywhere where the information is generated and stored and structured or unstructured format.
- Types of data include: (Based on the source)
    - observational data
    - laboratory experimental data
    - computer simulation
    - textual analysis
    - physical artifacts or relics

# *How data add value to the business*



Data warehouse

**Development of data product**
Algorithm solution in production, marketing and sales etc (eg.Recomendation Engines)

**Discovery of data insight**
Quantitative data analysis  to help to take strategic business decision

Business values

Amazon's recommendation engines
Netflix recommends movies
Spotify recommends Music
Gmail's spam filter
Google Map
Autonomous cars

Netflix-to make decisions on which Netflix original series to produce
Target-to guide messaging to different market audiences.
Proctor & Gamble -help plan for production levels more optimally

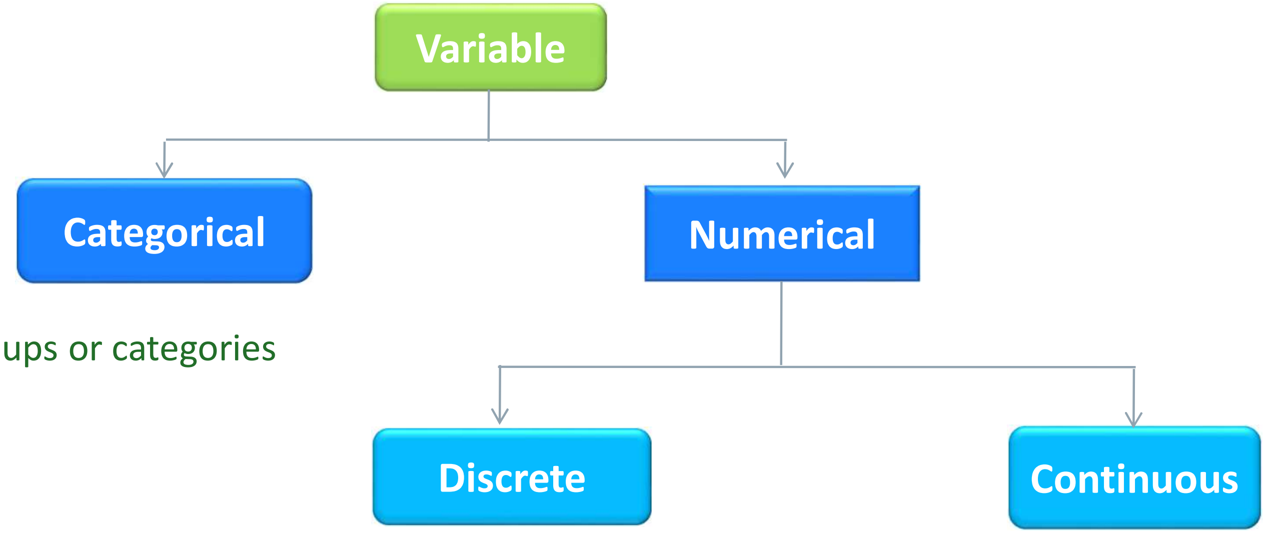
Source: https://datajobs.com/what-is-data-science

# *Why Data is important*

- To make better decisions
- To solve problems By finding the reason for underperformance
- To evaluate the performance
- To improve the performance/process
- To understand consumers and market

# *Variable*

- A variable is a specific characteristic (such as age or weight) of an individual or object, that is capable of taking different values.



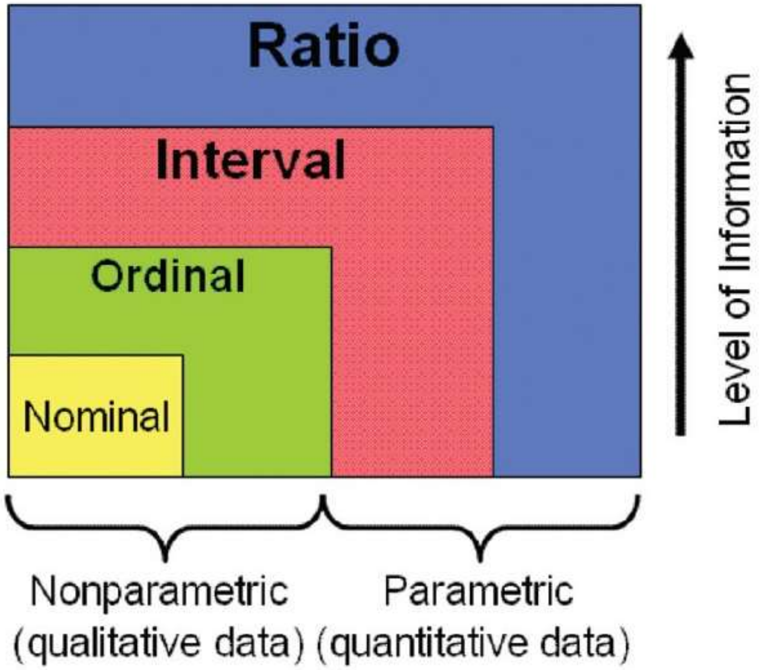
Responses that belong to groups or categories
Ex:
Yes/no response
Gender
Marital status
Range of choices
(strongly disagree to strongly agree)

counting process
Finite number of values
Ex:
number of students enrolled
Number of products

measurement (not a counting) process
Any value between a range (infinite)
Ex:
Weight
volume

# Measurement Levels

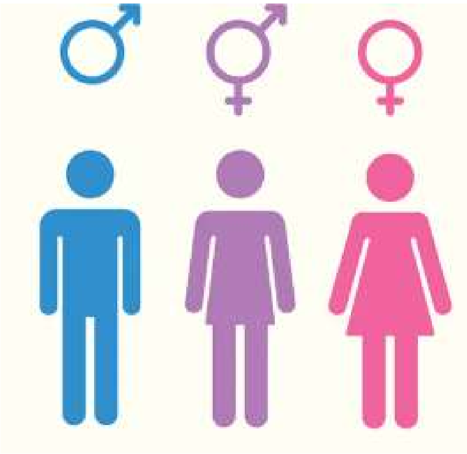
- ➢A standard process, used to assign numbers to particular attributes or characteristics of a variable.
- ➢Qualitative data
  - ➢No measurable meaning in the difference of Numbers
  - ➢Nominal - Lowest level of Measurement
  - ➢Ordinal
- ➢Quantitative data
  - ➢There is a measurable meaning in the difference of Numbers
  - ➢ Interval- Lowest level of Measurement
  - ➢Ratio



*Nonparametric statistics may be used to analyze interval and ratio data measurements.

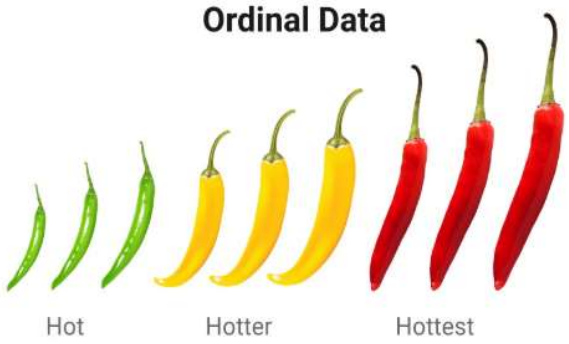
# *Qualitative data - Nominal Data*

- Nominal data
  - Describe the categories or classes of responses.
  - Example:
    - gender, country of citizenship, political affiliation, and ownership of a mobile phone are nominal
  - Lowest or weakest type of data
  - Numerical identification is chosen strictly for convenience
  - Example
    - 1 = Male; 2 = Female
    - 1 = Yes; 2 = No
  - Does not imply ranking of responses



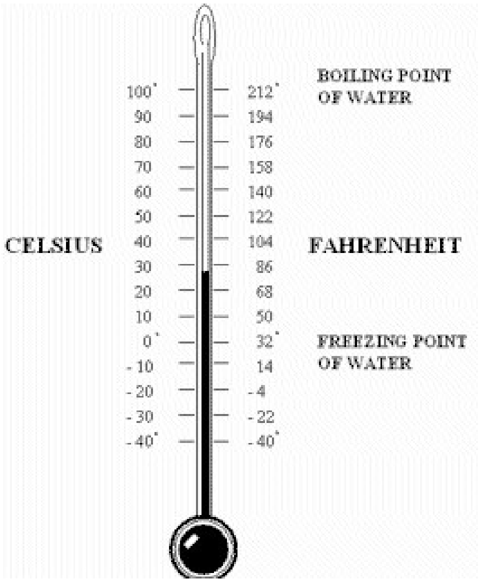
# *Qualitative data-Ordinal Data*

- ➢Similar to nominal data the values are words that describe responses
- ➢Indicates ordering of items
- ➢Ranking is implied
- ➢Classifies data into distinct categories
- ➢no measurable meaning to the “difference” between responses
- ➢Ranking is implied
- ➢Examples
  - ➢Product quality rating (1: poor; 2: average; 3: good)
  - ➢Satisfaction rating with your current Internet provider (1: very dissatisfied; 2: moderately dissatisfied; 3: no opinion; 4: moderately satisfied; 5: very satisfied)
  - ➢Consumer preference among three different types of soft drink (1: most preferred;2: second choice; 3: third choice)



# *Quantitative data- Interval Data*

- ➢It ranks the data and the distance from arbitrary zero
- ➢the difference between two values is meaningful.
- ➢Zero does not signify the absence of characteristics.
- ➢Intervals of equal length signify equal difference in characteristics.
  - ➢The difference between a temperature of 100 degrees and 90 degrees is the same difference as between 90 degrees and 80 degrees.
- ➢Addition and subtraction of numbers from years.
- ➢Cant multiply any year with a number
- ➢Examples:
  - ➢Temperature (Celcius, Farenheit)
  - ➢Year
  - ➢IQ Score (Does not imply no intelligence)

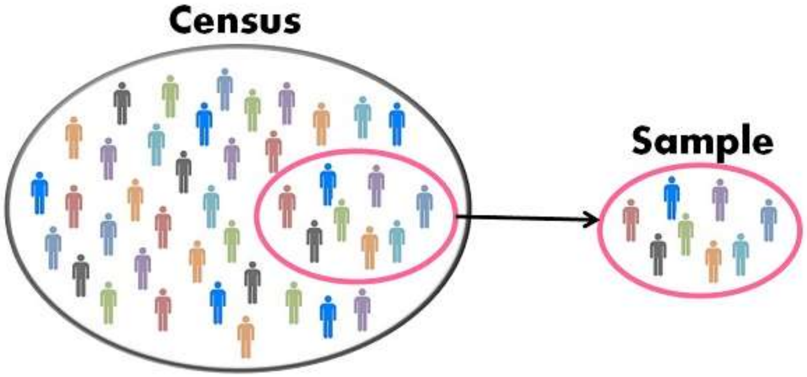


# *Quantitative data- Ratio Data*

- A ratio variable has all properties of interval data
- It has the proper definition for 0.0. when the variable equals 0.0, there is none of that variable
- True zero exits and are meaningful
- It also ranks data
- All arithmetic operations can be done. Example :Height, weight, salary

# *How do we collect data?*

- **Data Collection**
  - direct observation-Example: Counting Cars
  - Survey-people (through questionnaires, opinion polls, etc) or things (like pollution levels in a river, or traffic flow).
- **Population:**
  - The complete set of all items that interest an investigator
  - Population size, *N, can be very large or even infinite.*
    - Examples of populations include the following:
      - All potential buyers of a new product
      - All stocks traded on the NYSE
      - All registered voters in a particular city or country
  - Census is collecting data for every member of the group from the population
    - *It is time consuming  and expensive process*
    - *Appropriate for population of heterogeneous in nature*



https://keydifferences.com/difference-between-census-and-sampling.html

# *Population and Sample*

- A **sample**
  - an observed subset (or portion) of a population with sample size given by n.
  - Quick and economical method
  - Appropriate for population of homogeneous in nature

- **Random Sampling**
  - procedure used to select a sample of *n* objects from a population
  - each member of the population is chosen strictly by chance
  - each member of the population is equally likely to be chosen
  - Every sample has size n and equal chance of selection
- **Systematic sampling**
  - the selection of every j th item in the population
  - J=N/n

# *Population and Sample*

- **Parameter**
  - A numerical measure that describes a specific characteristic of a population
- **Statistic**
- a numerical measure that describes a specific characteristic of a sample
  - Sampling error-lack of sufficient information from the population
  - Non sampling error-Sampling bias and non response
    - The population actually sampled is not the relevant one
    - Survey subjects may give inaccurate or dishonest answers.
    - There may be no response to survey questions.

## *What is Data Analytics?*

- The scientific process of transforming data into insights for making better decisions.
- The use of the data information technology, statistical analysis, quantitative methods and mathematical or computer-based models to help managers gain improved insight about their business operations and make better, fact-based decisions- James Evans
- As Process of analyzing raw data to find trends and answer questions
  - captures its broad scope of the field
- it includes many techniques with many different goals.
- It will provide a clear picture of
  - where you are,
  - where you have been and
  - where you should go
- Tools
  - Microsoft Excel, SAS, R, Python, tableau, apache Spark, Qlik View

**Analytics=Analysis?**

# What is Data Analysis?

- The process of examining, transforming and arranging raw data in a specific way to generate useful information from it.
- It allows for the evaluation of data through analytical and logical reasoning
  - leads to some sort of outcome or conclusion in some context.
- To explain what, How and Why based questions on the analysis of the existing data
- Data analysis is a multi-faceted process
  - It involves a number of steps, approaches and diverse techniques
- To separate huge data into the different chunks, study each of them and then trying to find how they are related to other chunks; that is the Analysis in a nutshell

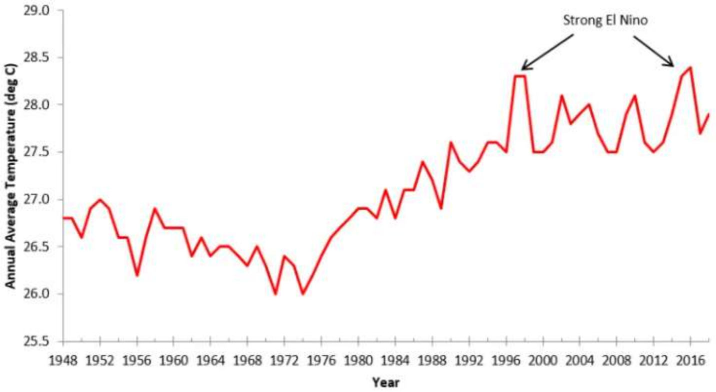
- Tools
  - Open Refine, Tableau public, KNIME, Google Fusion Tables, Node XL

# *Data analysis Vs data analytics*

Data Analysis

Past

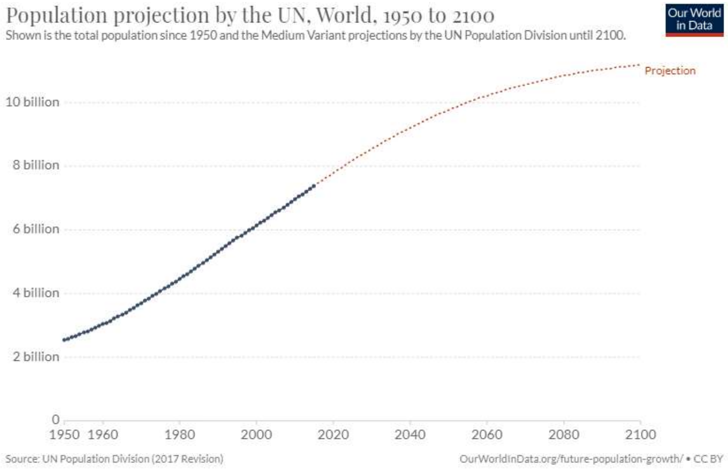
Explain what, how and why?
Ex: Sales history
Return on Investment
Value customers
Target the right audience



Data Analytics

Future

Explore potential future Events
Ex: To uncover unknown patterns, opportunities and insights that can drive proactive, evidence-based decision making

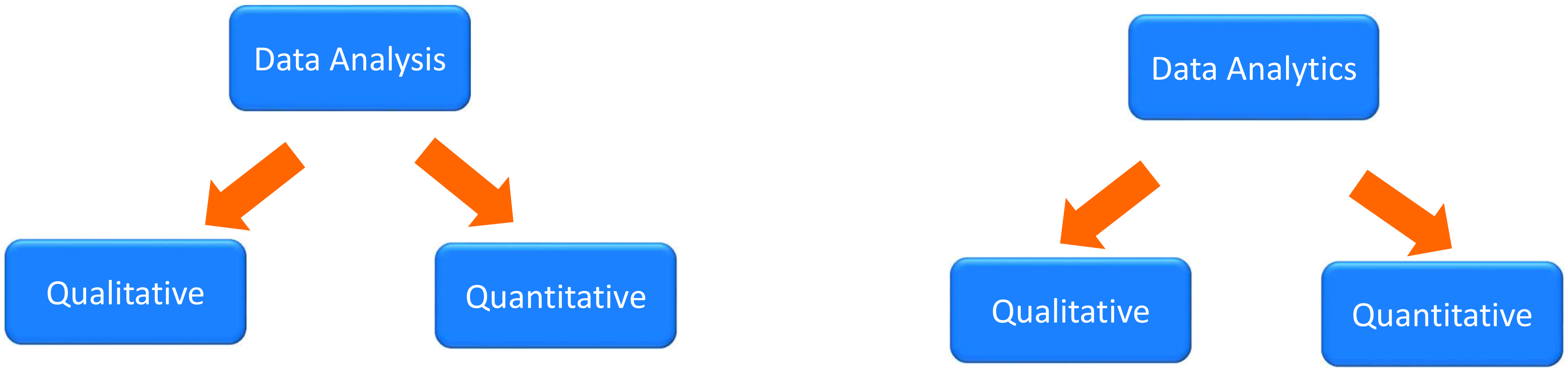

Population projection by the UN, World, 1950 to 2100

Shown is the total population since 1950 and the Medium Variant projections by the UN Population Division until 2100.

Source: UN Population Division (2017 Revision)

OurWorldInData.org/future-population-growth/ • CC BY

# *Data analysis Vs data analytics*



Data Analysis

Qualitative

Quantitative

explain how and why a story ends in that way it did?

Data + how the revenue decreased Last Month

Data Analytics

Qualitative

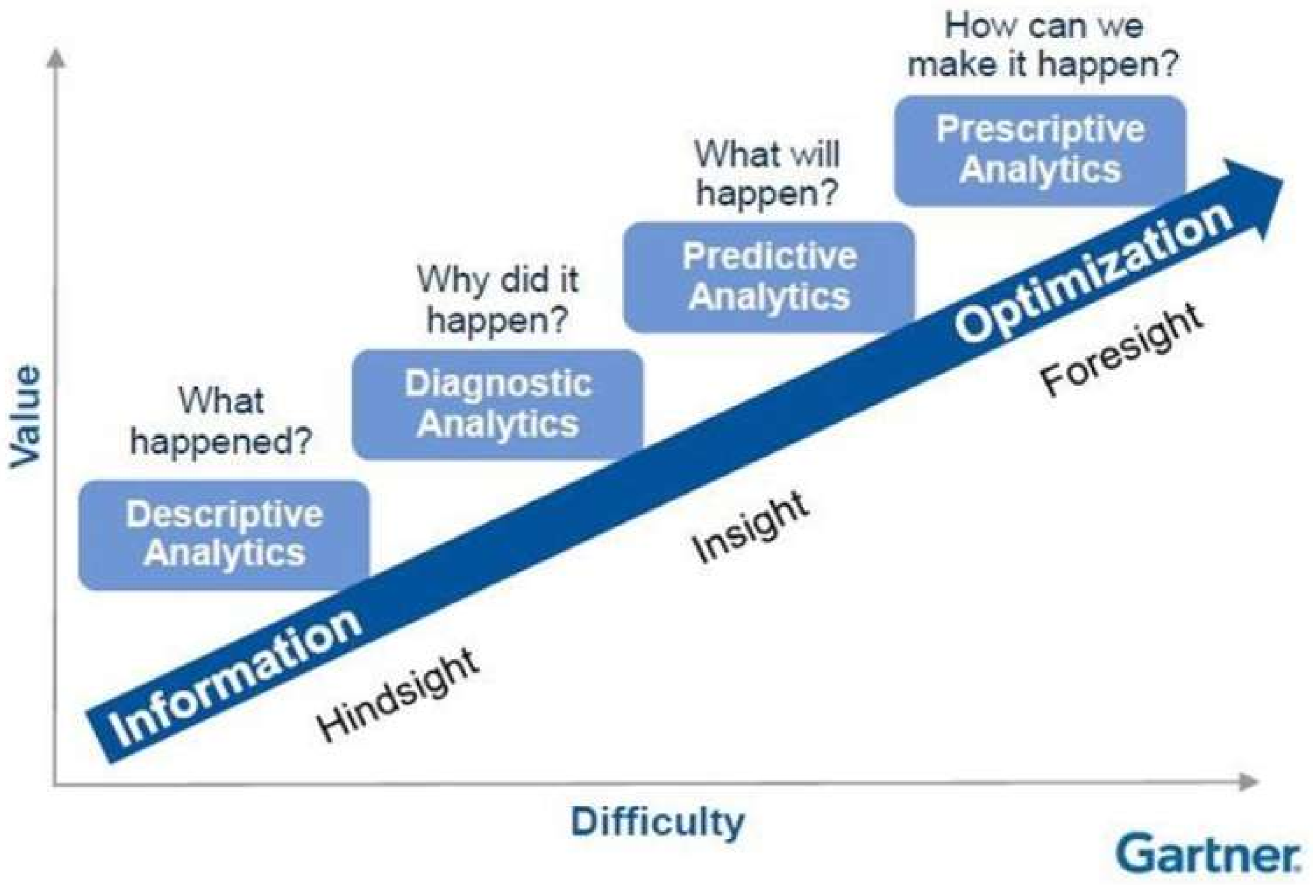
Quantitative

Intuition +analysis

Formulas + Algorithms

**Data Analysis ≠ Data Analysis**

# Types of Data analytics



*Source:Gartner's analytics maturity model may be a good starting point to explain and prepare for the transition to AI. Image: Gartner*
*Image: Gartner*

# *Descriptive analytics (Insight into the past)*

- Descriptive analytics is the conventional form of business intelligence or data analysis
- Describe”, or summarize large datasets
- Interpretable by humans
- Use data aggregation and data mining to provide insight into the past
- Answer questions about what has happened.
- sums, averages, percent changes, aggregated values
- This process requires the collection of relevant data, processing of the data, data analysis and data visualization.

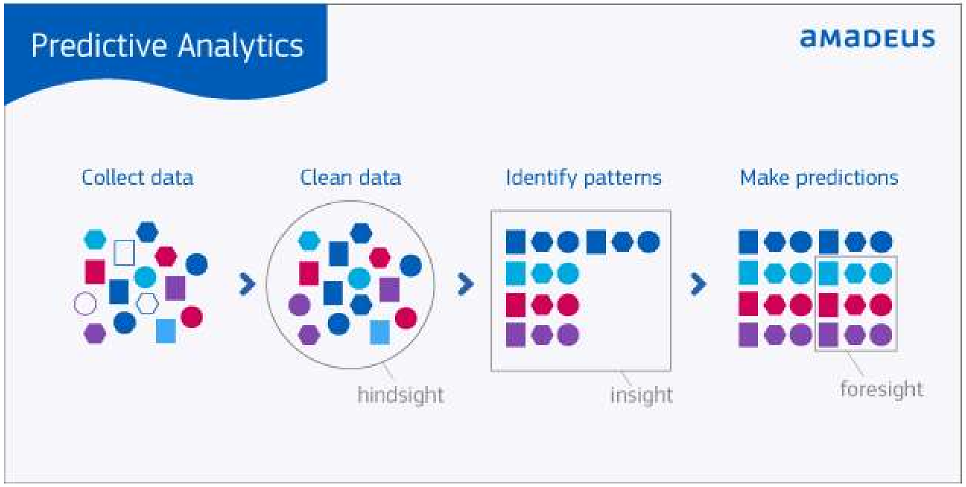
- Example:
- historical insights regarding the company’s production, financials, operations, sales, finance, inventory and customers.

# Diagnostic analytics

- Diagnostic analytics is a form of advanced analytics which examines data or content to answer the question why did it happen ?
- They take the findings from descriptive analytics and dig deeper to find the cause
- The performance indicators are further investigated to discover why they got better or worse
- This generally occurs in three steps:
  - Identify anomalies in the data. These may be unexpected changes in a metric or a particular market.
  - Data that is related to these anomalies is collected.
  - Statistical techniques are used to find relationships and trends that explain these anomalies.
- Techniques
  - data discovery
  - data mining
  - correlations
  - These tools can be used for prescriptive analytics also

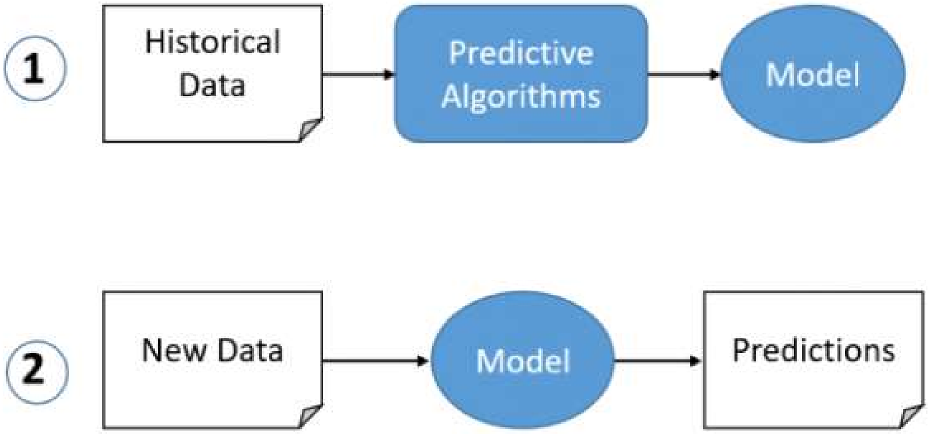
# *Predictive analytics (Understanding the future)*

- Helps answer questions about what will happen in the future.
- These techniques use historical data to identify trends and determine if they are likely to recur.
- Predictive analytical tools
  - statistical and machine learning techniques
  - neural networks, decision trees, and regression
  - Time series forecasting, Data mining
- No Algorithm can predict with 100% certainty
- predictive analytics is based on probabilities
- **Applications:**
- To Forecast
  - customer behavior and purchasing patterns
  - demand for inputs from the supply chain, operations and inventory.
- To produce a credit score
  - probability of customers making future credit payments on time



https://bigdata-madesimple.com/5-examples-predictive-analytics-travel-industry/

# *Predictive analytics (Understanding the future)*



- ***How Does Predictive Analytics Work?***
  - First, identify what you want to know based on past data
  - Next, consider if you have the data to answer those questions
  - Train the system to learn from your data and can predict outcomes.
  - Schedule your modules.
  - Use the insights and predictions to act on these decisions.
  - Regularly retrian the model

# *Prescriptive analytics(Advise on possible outcomes)*

- Helps answer questions about what should be done.
- Allows users to "prescribe" a number of different possible actions and guide them towards a solution
- Advise on possible outcomes before the decisions are actually made
- This allows businesses to make informed decisions in the face of uncertainty..
- By using insights from descriptive analytics and predictive analytics, data-driven decisions can be made.
- Techniques and tools
  - Business rules, optimization algorithms, simulation machine learning and computational modelling procedures.
  - Historical and transactional data, real-time data feeds, and big data.
- Prescriptive analytics to optimize production, scheduling and inventory in the supply chain

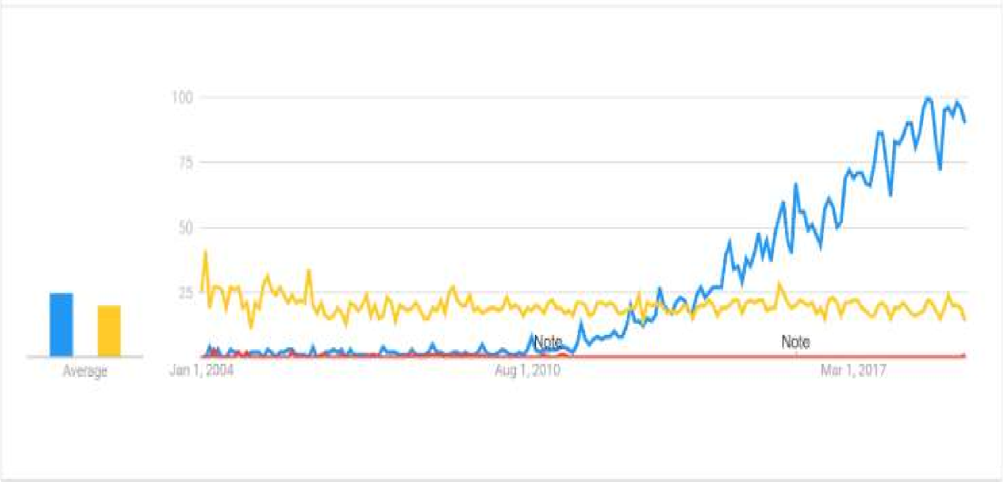
# *Why data analytics is important?*

- The applications of data analytics are broad
- optimize efficiency in many different industries
- Improving performance
- Applications
- financial /Banking sector
  - to predict market trends
  - to assess risk
    - to detect and prevent fraud
    - to determine lending risk (credit scores)
- healthcare (health informatics),
  - Predicting patient outcomes,
  - efficiently allocating funding and improving diagnostic techniques
  - Drug discovery , to understand the market for drugs , and predict their sales
- crime prevention
- environmental protection

**IoT provides more opportunity to Data analytics**

# *Demand for Data analytics*

- **According to IBM, the number of jobs for data professionals in the U.S will increase to 2,720,000 by 2020.**
- Demand for knowledgeable data analytics professionals currently outweighs the supply, meaning that companies are willing to pay a premium to fill their open job positions.

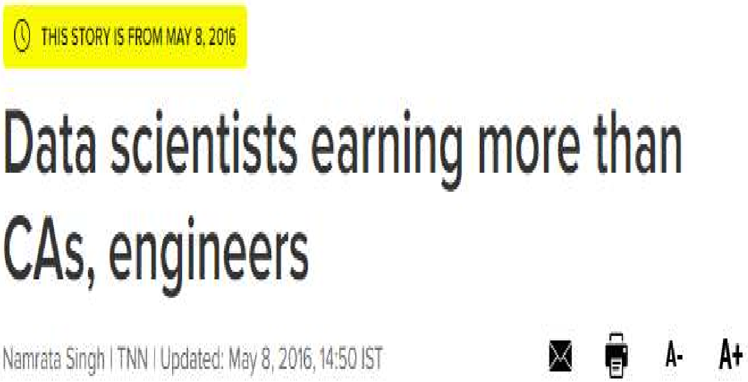


*Google Trends of search words for data scientist, operations researcher and statistician.*

🟦 Data scientist

🟨 Operations researcher and statistician

https://pubsonline.informs.org/do/10.1287/LYTX.2019.04.02/full/



THIS STORY IS FROM MAY 8, 2016

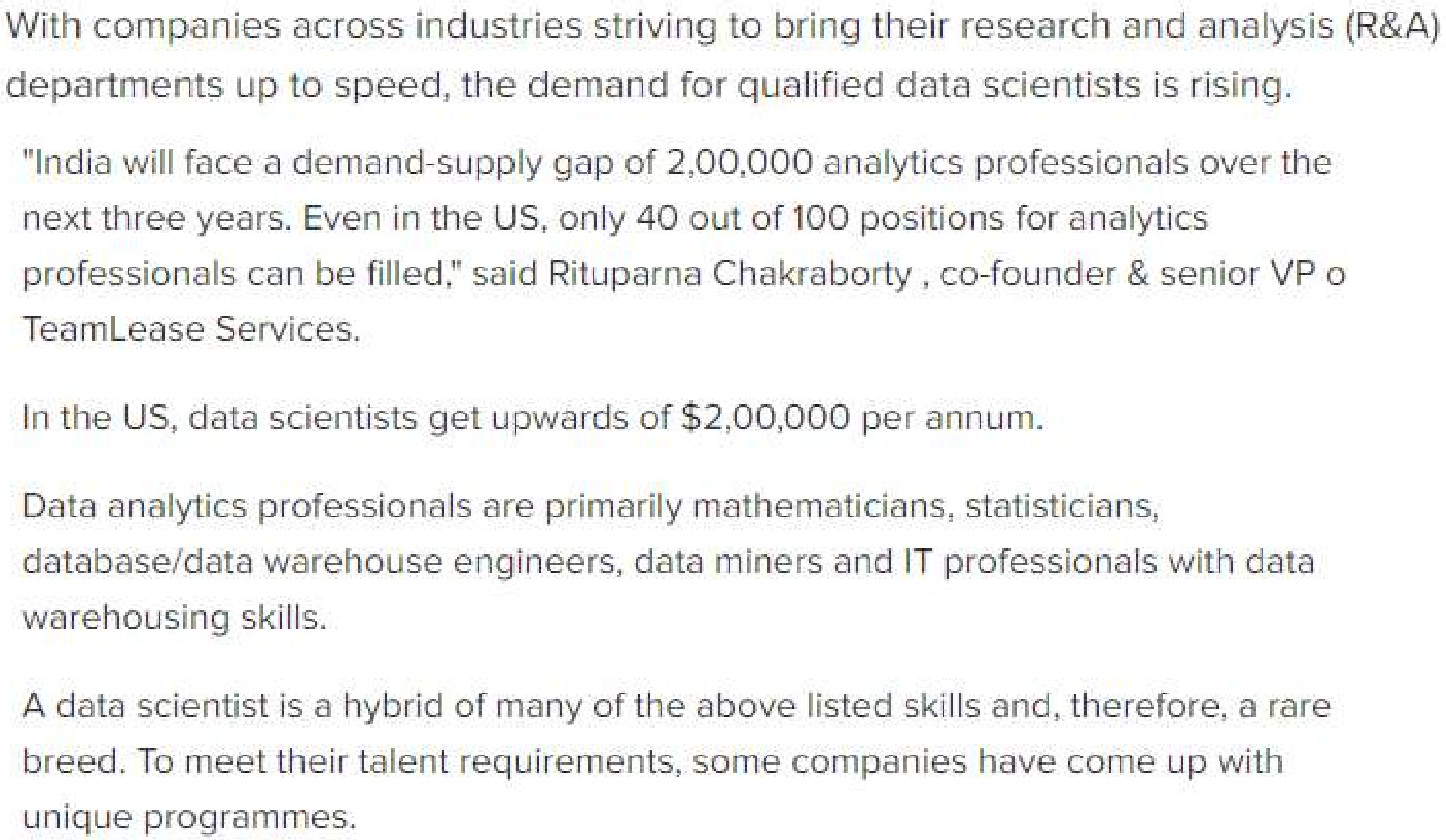
Data scientists earning more than CAs, engineers

Namrata Singh | TNN | Updated: May 8, 2016, 14:50 IST

A- A+

https://timesofindia.indiatimes.com/india/Data-scientists-earning-more-than-CAs-engineers/articleshow/52171064.cms

# *Demand for Data analytics*



With companies across industries striving to bring their research and analysis (R&A) departments up to speed, the demand for qualified data scientists is rising.

"India will face a demand-supply gap of 2,00,000 analytics professionals over the next three years. Even in the US, only 40 out of 100 positions for analytics professionals can be filled," said Rituparna Chakraborty , co-founder & senior VP o TeamLease Services.

In the US, data scientists get upwards of $2,00,000 per annum.

Data analytics professionals are primarily mathematicians, statisticians, database/data warehouse engineers, data miners and IT professionals with data warehousing skills.

A data scientist is a hybrid of many of the above listed skills and, therefore, a rare breed. To meet their talent requirements, some companies have come up with unique programmes.

Source:https://timesofindia.indiatimes.com/india/Data-scientists-earning-more-than-CAs-engineers/articleshow/52171064.cms

# What is Role of data analyst?

- Intersection of information technology, statistics and business
  - To increase efficiency and improve performance
  - Discovering patterns in data
- The data analytics process
  - data mining
    - to extract raw data, transform into useful form, and load data
  - data management
    - designing and implementing databases
    - Creating and managing SQL
  - statistical analysis
    - heart of data analytics to analyze data
    - statistics and machine learning techniques
    - R or Python (with pandas) are essential to this process
  - data presentation
    - Data visualization to tell the story in the data

## *Top Tools in Data Analytics*

- **R programming**
  - statistics and data modeling
  - automatically install all packages
- **Python_**
  - machine learning and visualization libraries
    - Scikit-learn, TensorFlow, Matplotlib,Keras
  - SQL server, a MongoDB database or JSON
- **Tableau Public**
  - connects to any data source such as Excel, corporate Data Warehouse
  - creates visualizations, maps, dashboards etc with real-time updates on the web.
- **QlikView**
  - offers in-memory data processing with the results delivered to the end-users quickly.
  - offers data association and data visualization with data being compressed to almost 10% of its original size.
- **SAS**
  - A programming language and environment for data manipulation and analytics, this tool is easily accessible and can analyze data from different sources

# *Top Tools in Data Analytics*

- **Microsoft Excel**
  - most widely used tools for data analytics
- **RapidMiner**
  - predictive analytics, such as data mining, text analytics, machine_learning
  - integrate with any data source types such as Access, Excel, Microsoft SQL, Tera data, Oracle, Sybase etc.
- **KNIME**
  - to analyze and model data with the benefit of visual programming
  - provides a platform for reporting and integration through its modular data pipeline concept
- **OpenRefine (GoogleRefine)**
  - this data cleaning software will help you clean up data for analysis
  - It is used for cleaning messy data, the transformation of data and parsing data from websites
- **Apache Spark**
  - largest large-scale data processing engine
  - executes applications in Hadoop clusters 100 times faster in memory and 10 times faster on disk.
  - popular for data pipelines and machine learning model development

## *What is big data analytics?*

- It the use of advanced analytic techniques
- To make better and faster decisions using data that was previously inaccessible or unusable
- Big data is a term applied to data sets whose size or type is beyond the ability of traditional relational databases to capture, manage and process the data with low latency.
- Data
    - Very large with different sizes from terabytes to zetta bytes
    - Diverse data sets structured, semi-structured and unstructured data
    - Different sources
    - Big data has one or more of the following Characteristics
        - high volume, high velocity, high variety, veracity
- Big data sources
    - big data comes from sensors, devices, video/audio, networks, log files, transactional applications, web and social media
    - Generated in real time and large in scale

# *Data Analytics vs. Business Analytics*

- **Data analytics** involves analyzing datasets to uncover trends and insights that are subsequently used to make informed organizational decisions.
- **Business analytics** is focused on analyzing various types of information to make practical, data-driven business decisions, and implementing changes based on those decisions. Business analytics often uses insights drawn from data analytics to identify problems and find solutions.

| Data Analyst | Business Analyst |
|---|---|
| To tell compelling stories with data that empower organizational leaders to make better, more informed decisions | Business analysts are responsible for using data to inform strategic business decisions. |
| Designing and maintaining data systems and databases, including troubleshooting potential issues<br>Mining and cleaning data in preparation for analysis<br>Preparing reports which effectively communicate their findings to organizational leadership and key stakeholders | Evaluating business processes for efficiency, cost, and other valuable metrics<br>Communicating insights with business teams and key stakeholders<br>Preparing strategic recommendations for process adjustments, procedures, and performance improvements |

Source:https://www.northeastern.edu/graduate/blog/data-analyst-vs-business-analyst/#:~:text=Data%20analytics%20involves%20analyzing%20datasets,changes%20based%20on%20those%20decisions.

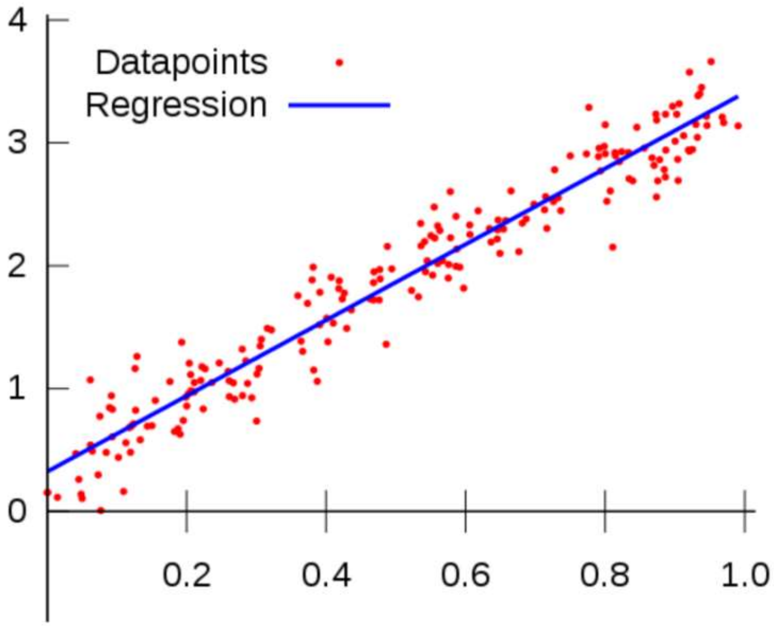
# *Difference between statistics and machine learning*

- Statistics is the science concerned with developing and studying methods for collecting, analyzing, interpreting and presenting empirical data.
- Probability is a mathematical language used to discuss uncertain events and probability plays a key role in statistics
- Statistical analysis applies statistical methods to a sample of data in order to gain an understanding of the total population
- Data analysis is the process of inspecting, cleaning, transforming and modelling available data into useful information that can be understood by non-technical people
- The process of data analysis can be used as an input into performing statistical analysis, as data from various sources can be combined in order to conduct statistical analysis

**Statistics are the results of data analysis**

# *Difference between statistics and machine learning*

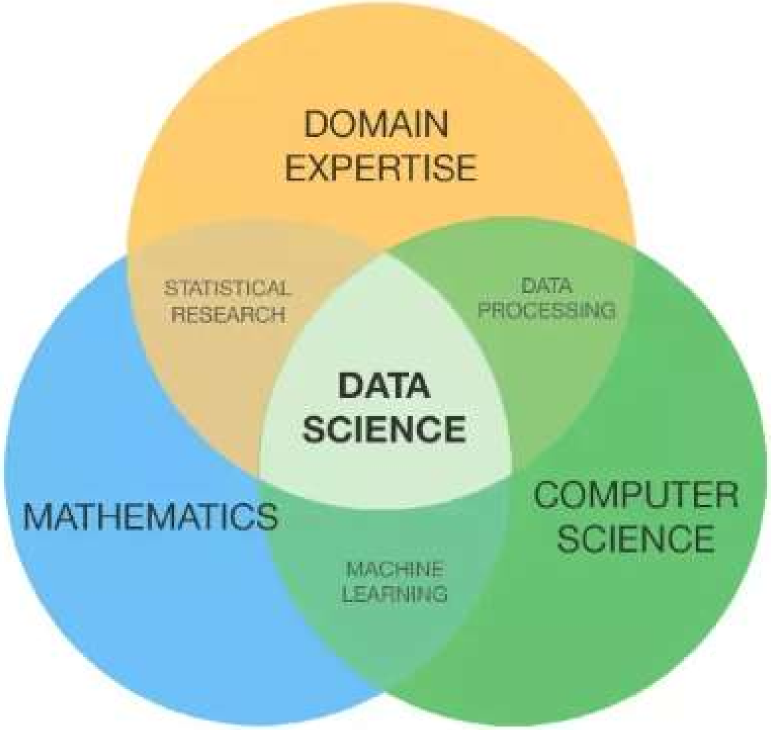
- Statistics is the mathematical study of data. We can not do statistics with out data
- A statistical model is a model for the data that is used either to infer something about the relationships within the data or to create a model that is able to predict future values
- Machine learning models are designed to make the most accurate predictions possible
- Arthur Samuel (1959) Machine Learning: field of study that gives the computers  the ability to learn with out being explicitly programmed



# Data Science

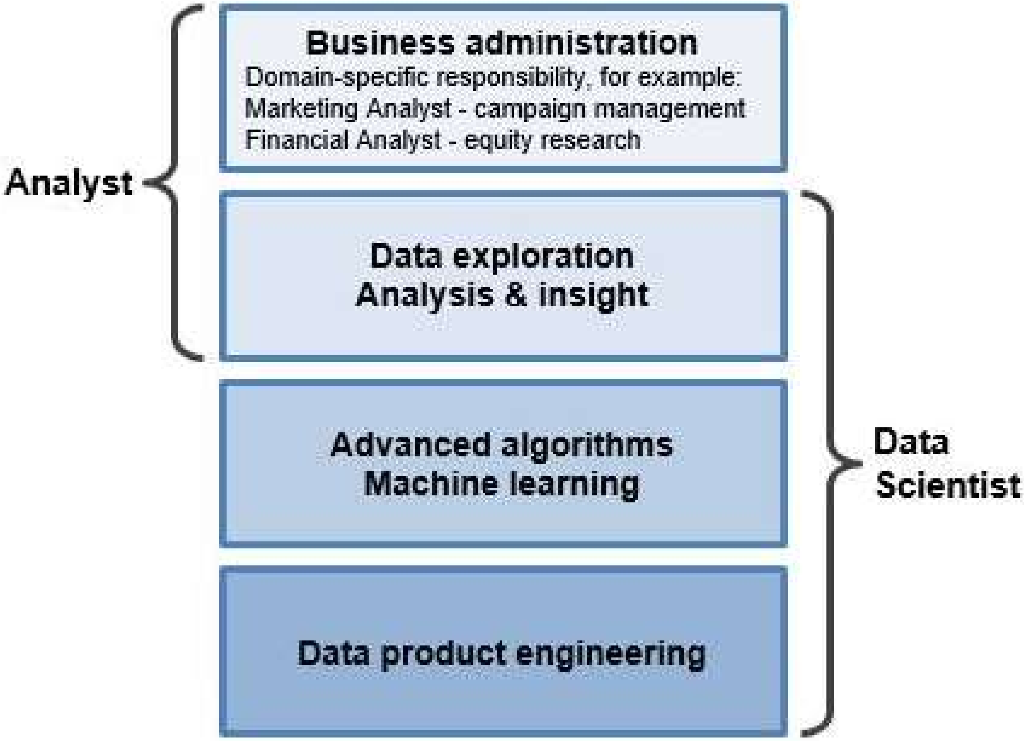
- Computational and statistical methods that are applied to data
- Data Science is the whole multidisciplinary field that includes domain expertise, machine learning, statistical research, data analytics, mathematics, and computer science.
- A data science expert identifies and defines potential business problems from various unrelated sources and gets data from these sources. Once data is analyzed through data analytics, a model is formed and tested for accuracy iteratively.
- Data science has the following main components –
    - **Statistics –** Statistics deals with the collection, analysis, interpretation, and presentation of data through mathematical methods.
    - **Data visualization –** Results of data science are displayed in the form of visually appealing diagrams, charts, and graphs which makes it simple to view and understand. This also helps in quicker decision making by highlighting the key takeaways.
    - **Machine learning –** this is an essential component where we use intelligent algorithms that learn on their own and predict human behavior as accurately as possible.

# Data Science



*Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.*

# Difference between data analyst and data scientist



ngiyabonga
danke
謝謝
teşekkür ederim
спасибо
thank you
gracias
dank je
tapadh leat
bedankt
hvala
maururu
dziękuję
mochchakkeram
go raibh maith agat
sukriya
kop khun krap
taiku
obrigado
sagolun
terima kasih
arigato
takk
dakujem
grazie
мерси
ありがとう
감사합니다
ευχαριστώ
merci
நன்றி