Introduction to Data Science

Week 1 – Part 1 – What is Data Science

CS 457 - L1 Data Science

Zeehasham Rasheed

Outline

- What is Data Science
- Who are Data Scientists?
- Data Science Applications
- Upcoming Opportunities
- All about Data

Opening Question

- If you haven’t heard of Data Science, you’re behind the times.

Foreword

- *Data Science touches every aspect of our lives on a daily basis. When we visit the doctor, drive our cars, get on an airplane, or shop for services, Data Science is changing the way we interact with and explore our world.*

Data Analysis Has Been Around

1935: "The Design of Experiments"

R.A. Fisher

1939: "Quality Control"

W.E. Deming

1958: "A Business Intelligence System"

Peter Luhn

1977: "Exploratory Data Analysis"

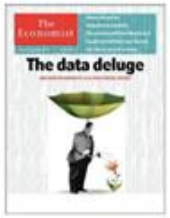
1989: "Business Intelligence"

Howard Dresner

1996: Google

Google

2010:"The Data Deluge"



1997: "Machine Learning"

2007:"The Fourth Paradigm"

2009:"The Unreasonable Effectiveness of Data"

Data Collection

- **Lots of data is being collected and stored**
  - Scientific Experiments
  - Internet of Things (IoT) - smart devices/appliances
  - Web data, e-commerce
  - Financial transactions, bank/credit transactions
  - Online trading and purchasing, Stock Market
  - Social Network
  - many more!

Data Science Definition

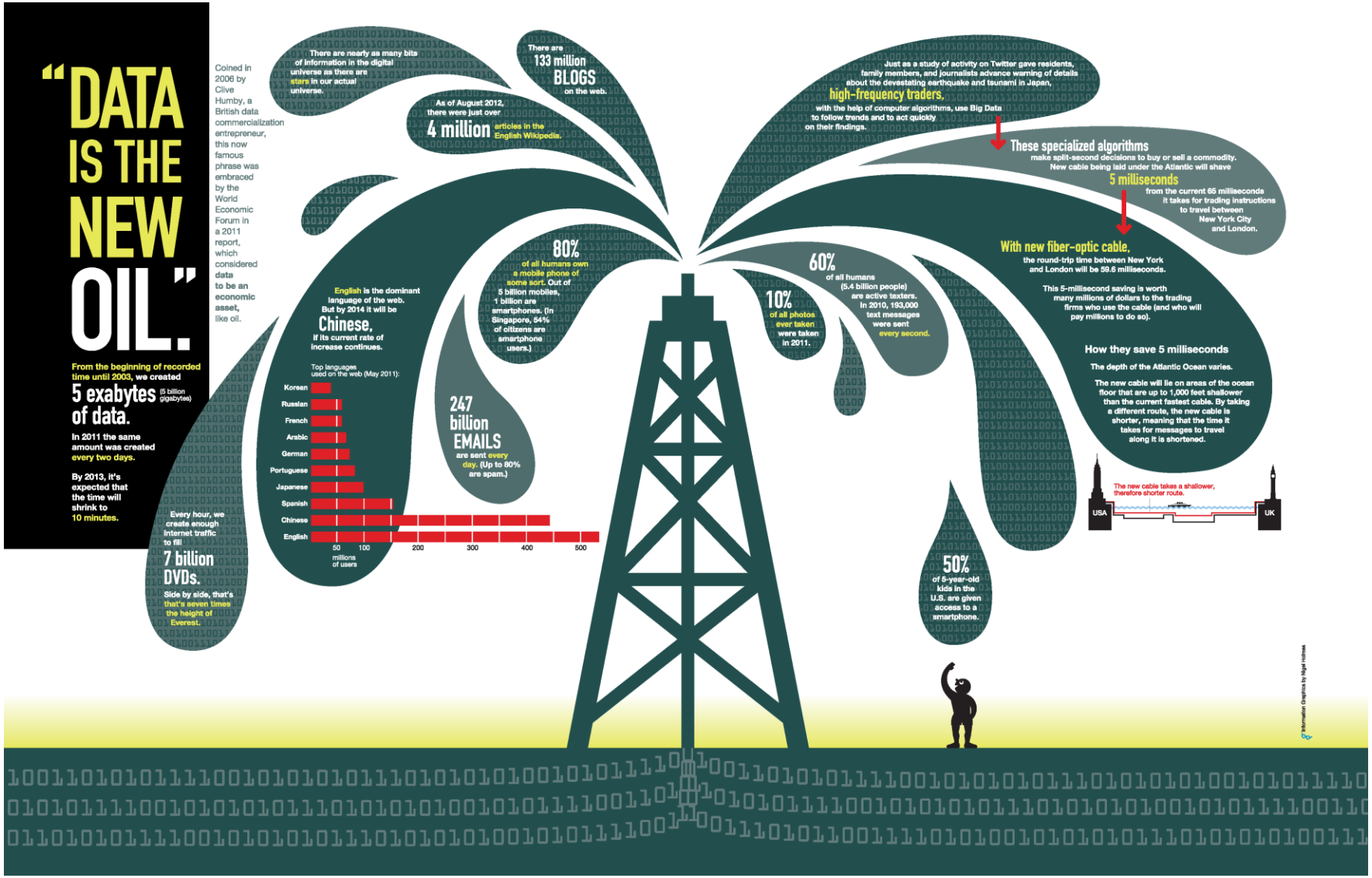
- Data Science is the art of turning **data into actions** and create **data products**
  - provide actionable information without exposing decision makers to the underlying data or analytics.

Data Product

- A data product driven by Data Science provides **actionable** information such as
  - Movie Recommendations
  - Weather Forecasts
  - Stock Market Predictions
  - Production Process Improvements
  - Health Diagnosis
  - Flu Trend Predictions
  - Targeted Advertising

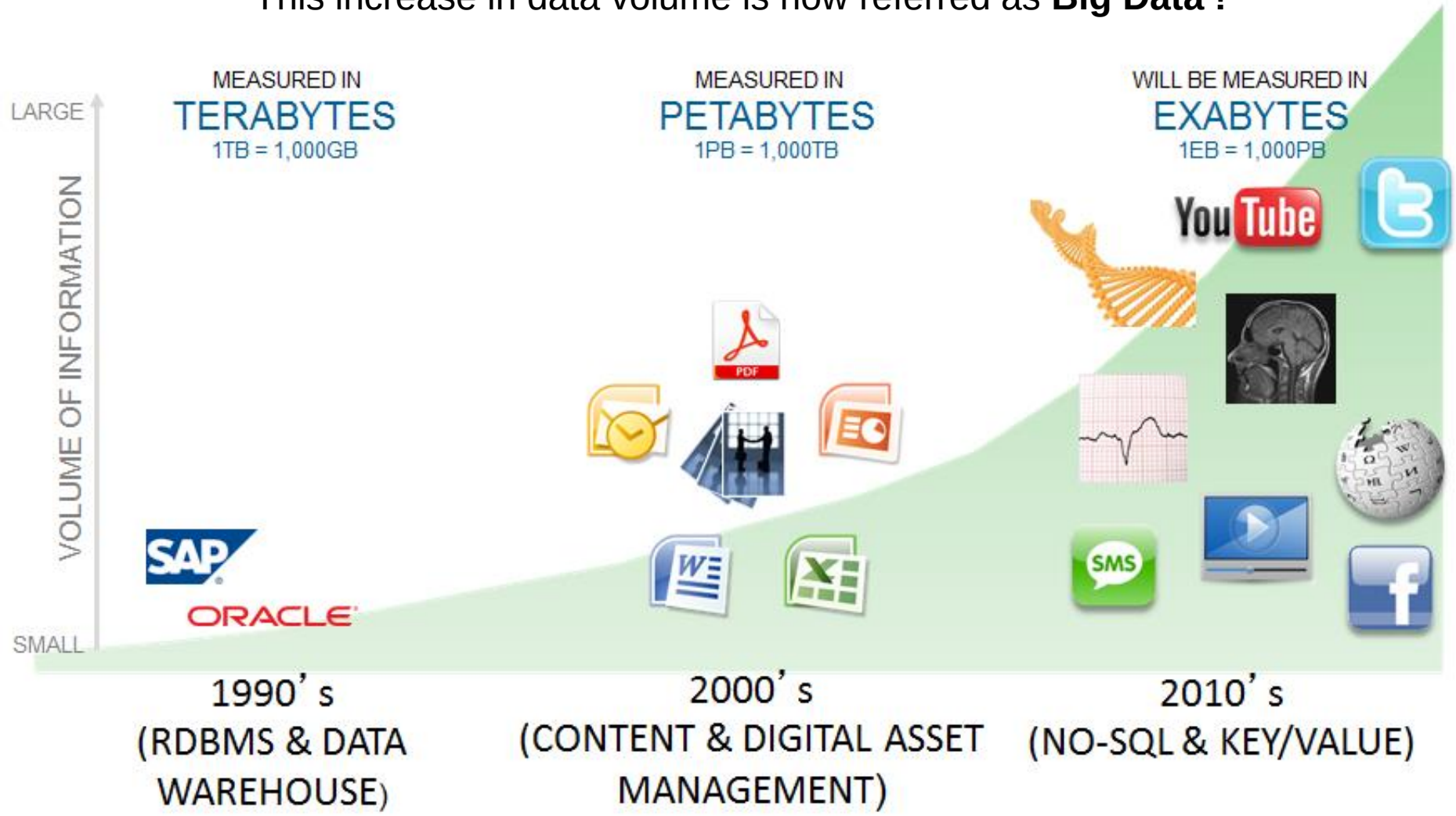
Data is the New OIL

- Clive Huby - World Economic Forum 2011



New Applications Driving Data Volume

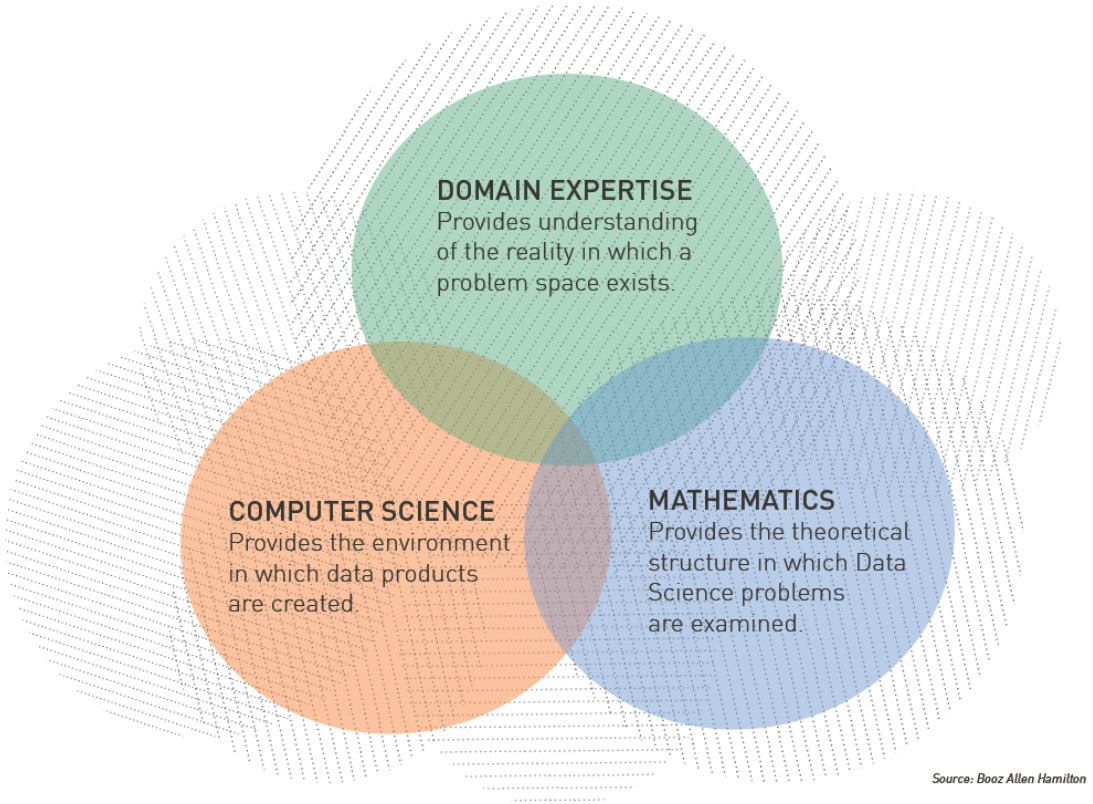
This increase in data volume is now referred as **Big Data !**



Building Data Science Capabilities

Data Science is all about building teams and culture.

- Computer Science
  - Pattern recognition, Visualization, Data Warehousing, High Performance Computing, Databases, Artificial Intelligence
- Mathematics
  - Mathematical Modeling
- Statistics
  - Statistical and Stochastic modeling, Probability

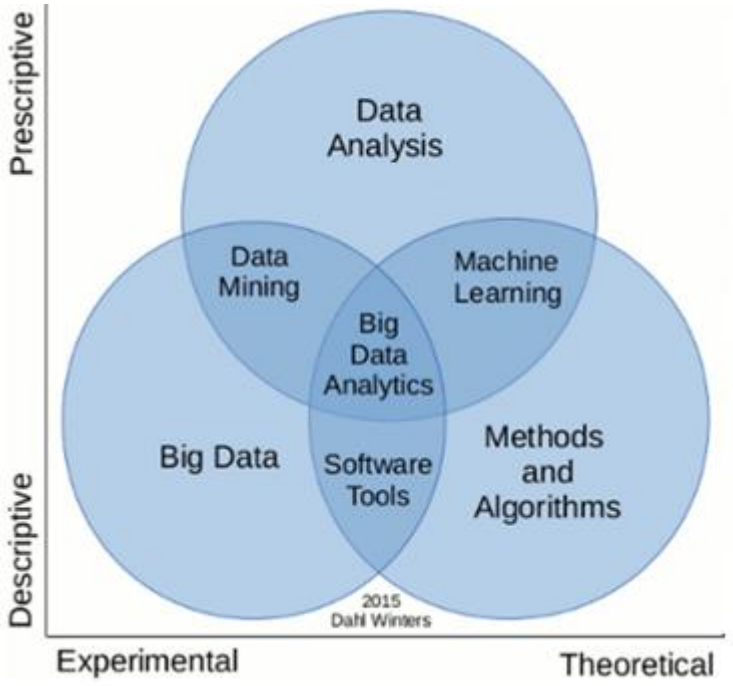


The Data Science Venn Diagram (inspired by [12])

Data Science by Itself

- Data Analysis
  - statistical methods to summarize data (Descriptive)
- Data Mining
  - Data mining closely relates to data analysis. Data mining is the process of transforming data into useful information (Descriptive)
- Machine Learning
  - Machine Learning finds patterns in data that are useful and are not visible from human point of view and use it for prediction (Prescriptive)

- **Descriptive:** answers “what has happened?”
- **Prescriptive:** answers “what should we do?”



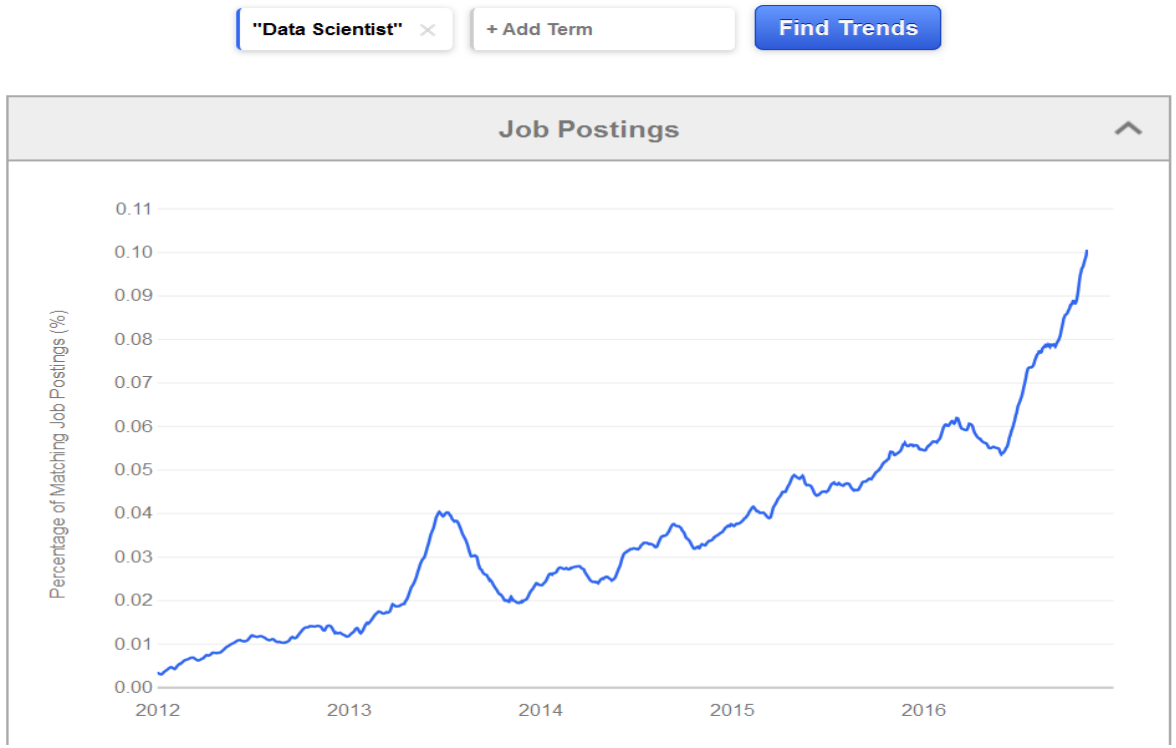
Industry and Academia

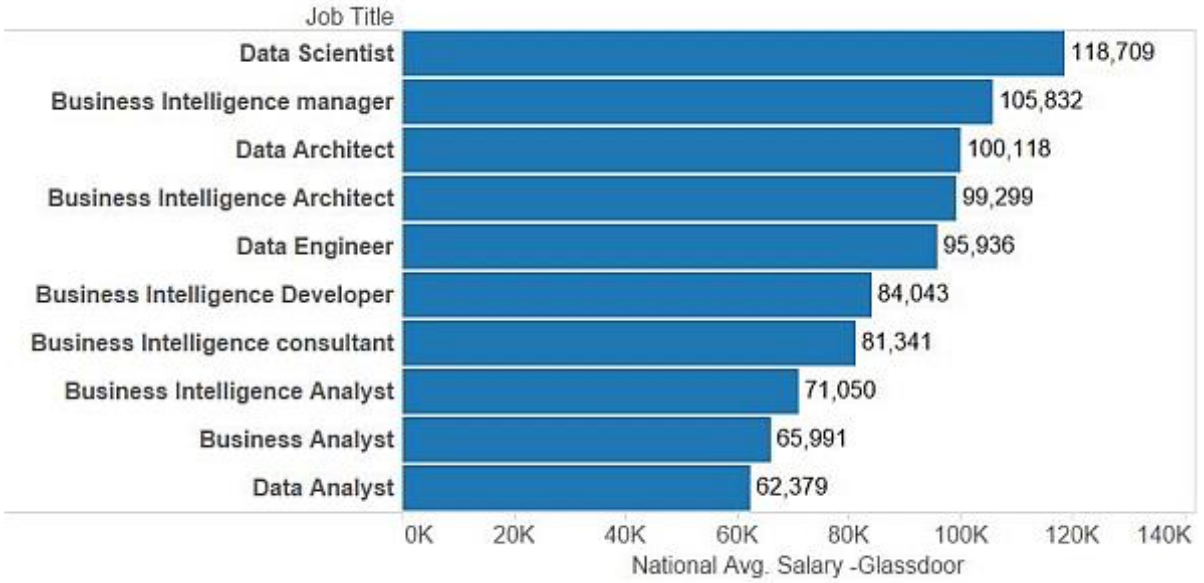
- The U.S. will need 140,000-190,000 predictive analysts/data scientists and 1.5 million managers/analysts by 2018 (*McKinsey Global Institute's June 2011)*

- New Data Science institutes being created or repurposed
  - NYU, Columbia, Washington and many others

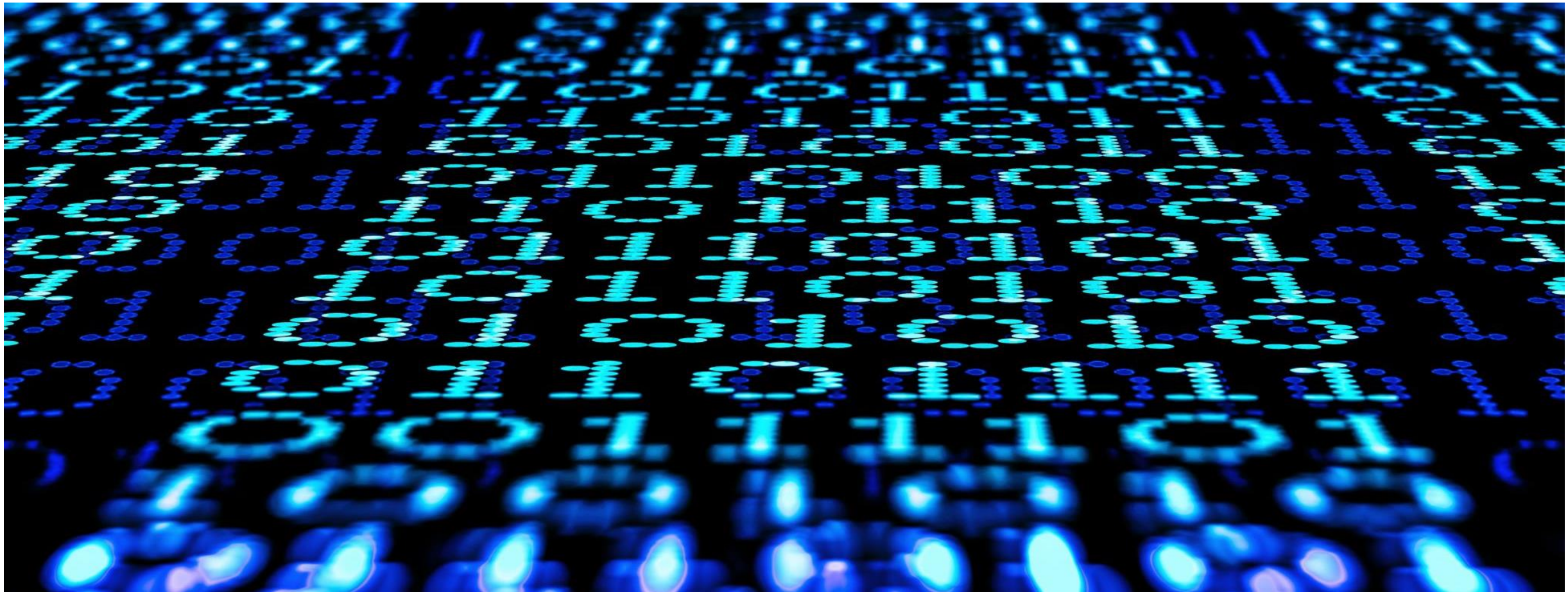
- New degree programs, courses, boot-camps added for data science
  - Berkeley, Georgia Tech, Stanford and many others

Job Postings and Job Titles





Data Science Technology Stack



Top Data Science Tools

Python and R – Data Science and Machine Learning Libraries

- Scikit-Learn, Pandas, Keras, Jupyter etc.

Anaconda, Virtual Environment for managing packages and distribution

SQL

NoSQL

Apache Spark

Tensor Flow

Hadoop

Elastic Search

Flask for building API services

Top Cloud Vendors for Data Science

- Google Cloud – AI Platform
- Microsoft Azure – Advance Analytics
- Amazon AWS – Big Data Analytics services

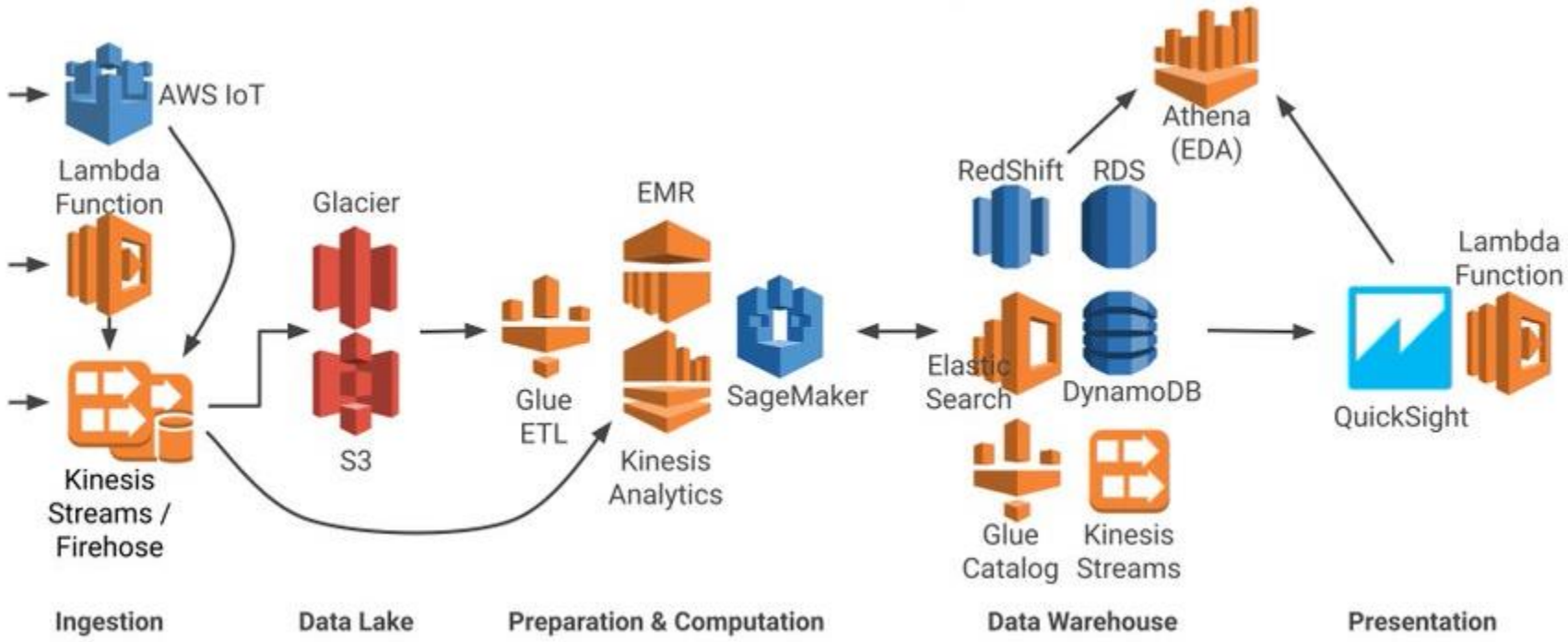
Enterprise Grade

Cutting Edge

Scales from GB to PB

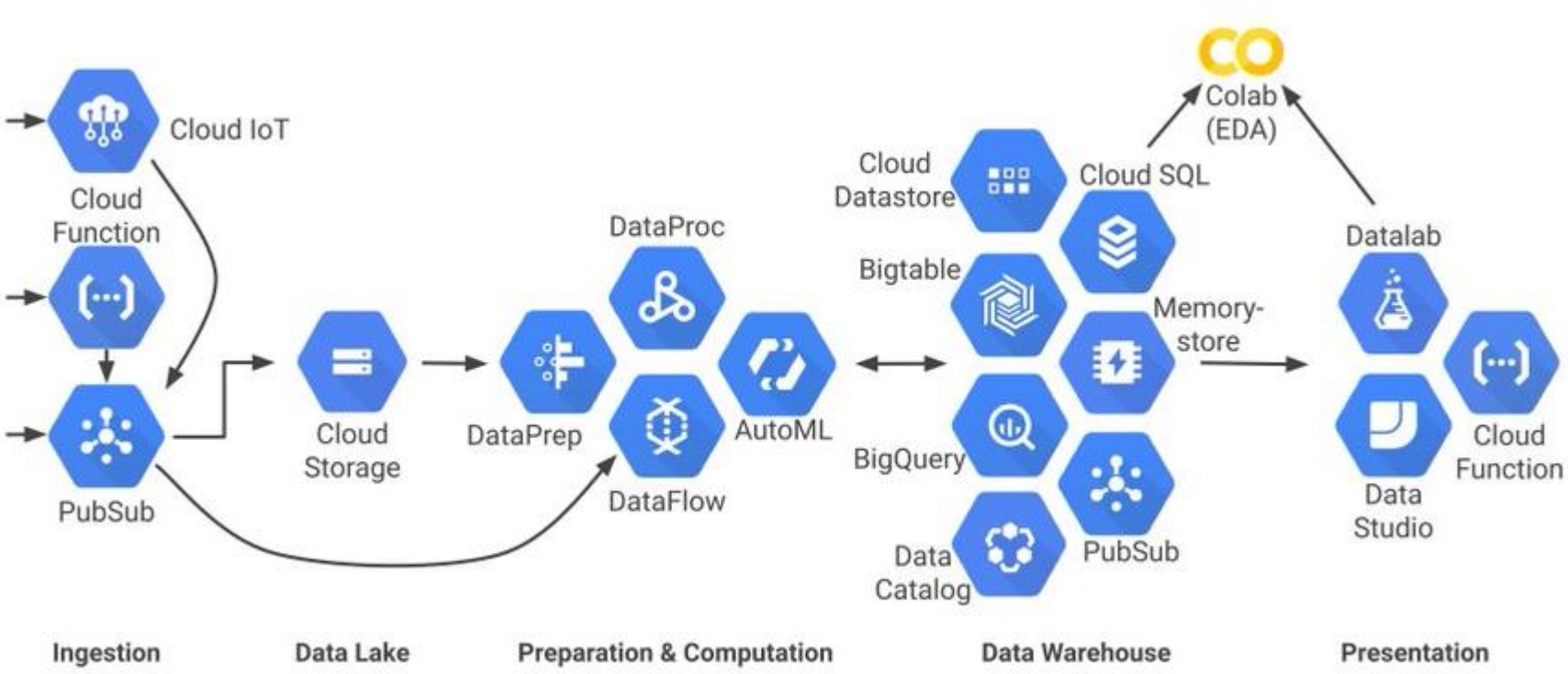
Unified & Modular

AMAZON AWS



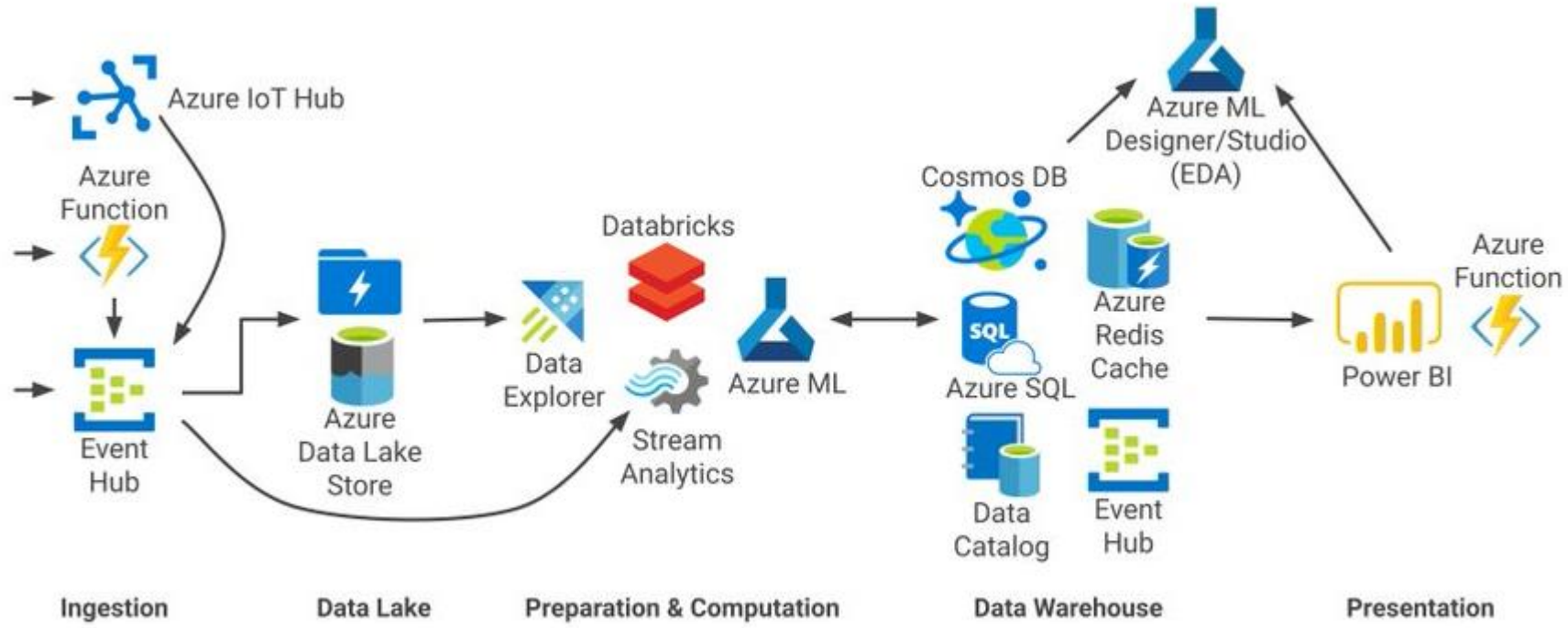
Lambda Function: Can be written in different languages (Python, R etc)
S3: Storage for data
EMR: Servers/Machines for computations
SageMaker: Libraries for Machine Learning in Python and R
Redshift: Distributed Relational Database (SQL)
RDS: Standard Relational Database (SQL)
DynamoDB: Non Relational Database (NoSQL)

Google Cloud Services



Cloud Function: Can be written in different languages (Python, R etc)
Cloud Storage: Storage for data
AutoML: Libraries for Machine Learning in Python and R
BigQuery: Distributed Relational Database (SQL)
BigTable: Non Relational Database (NoSQL)
Colab: Web based environment for writing Python and R code (Anaconda Jupyter for our course)

Microsoft Cloud Services



Azure Function: Can be written in different languages (Python, R etc)
Azure Data Lake Storage: Storage for data
Azure ML: Libraries for Machine Learning in Python and R
Azure SQL: Distributed Relational Database (SQL)
Cosmos DB: Non Relational Database (NoSQL)
Azure ML Designer EDA: Web based environment for writing Python and R code (Anaconda Jupyter for our course)
Databricks: uses Apache Spark, an open-source distributed computing framework (will use in our course)

End of Part 1

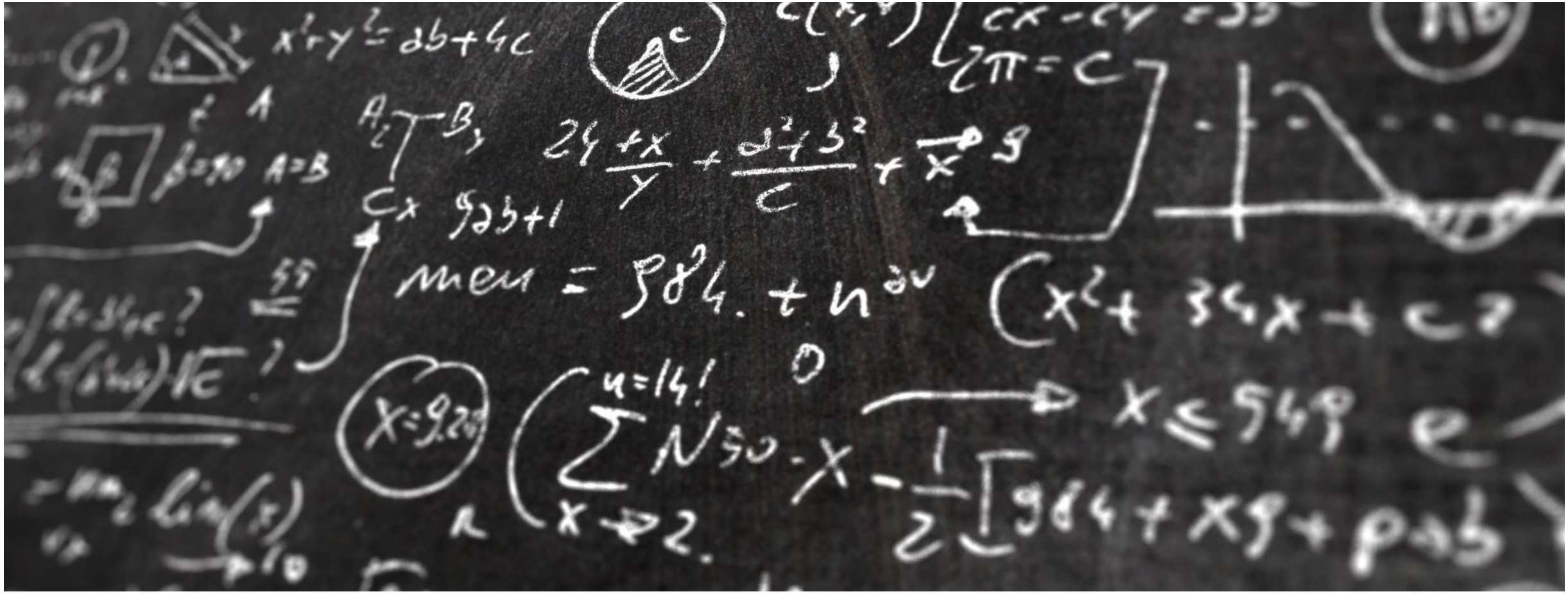
Introduction to Data Science

Week 1 – Part 2 – Data Science Applications

CS 457 - L1 Data Science

Zeehasham Rasheed

Data Science Applications

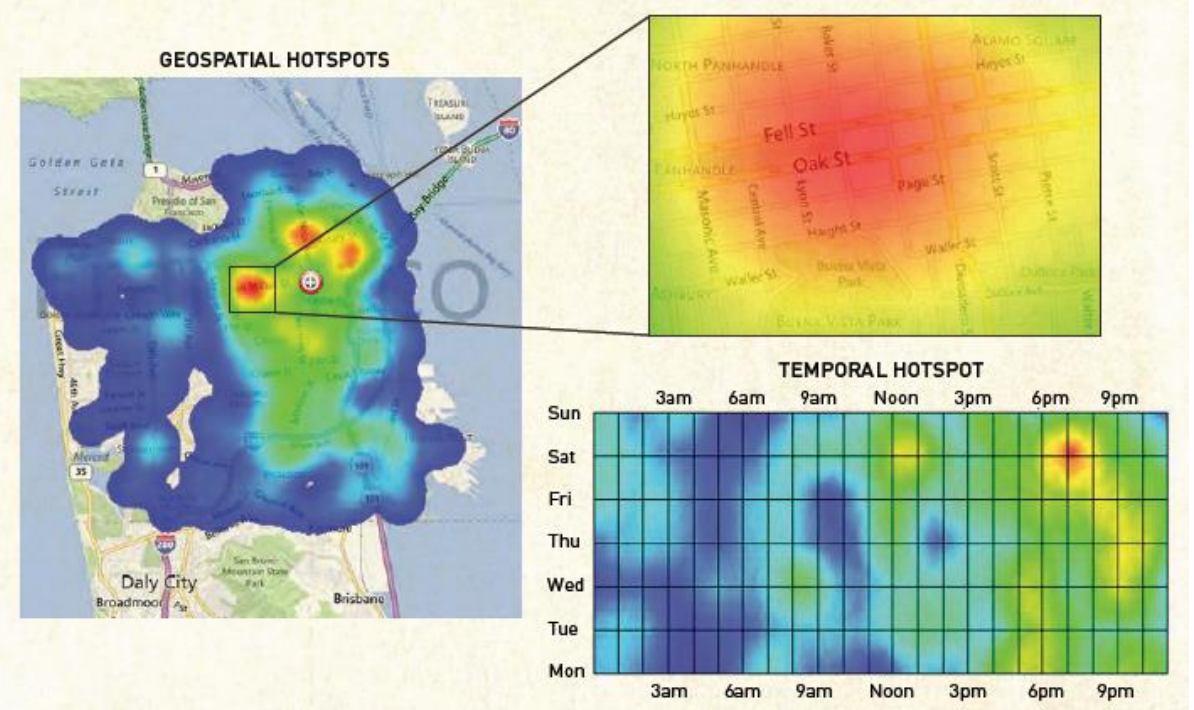


Motor Vehicle Theft

- According to the FBI, approximately $8 Billion is lost annually due to automobile theft

- **CrimeRadar:** Data driven product to predict future crimes using time series and decision trees models

- San Francisco crime data is analyzed (primary geospatial hotspot corresponded to an area surrounded by parks)
- Ability to predict the next crime location using machine learning



Reducing Flight Delays

Problem Statement

- Domestic airline departure delays are estimated to cost the U.S. economy $32.9 billion annually reported by The Federal Aviation Administration's (FAA's)

Data Science Solution

Analyze over 4 TB of data (50 million flights) detailing tarmac and airspace congestion, weather conditions, network effects, Traffic Management Initiatives, and airline and aircraft-specific attributes for every flight

Create a **predictive probabilistic model (Bayesian Belief Network)** to improve aircraft departure time predictions

This new model helps the FAA understand the causes of departure delays and develop policies and actions to improve traffic flow management.

Making Vaccine Safer

Problem Statement

- Analyze large volume of **Adverse Event (AE) reports**
- Developed tools that leverage **NLP (Natural Language Processing) , Network Analysis and Data Visualization** methodologies to extend and enhance CBER's efforts to monitor safety throughout the product lifecycle
- Visualizing relationships between vaccines and AEs can reveal new patterns and trends in the data, leading reviewers to uncover safety issues.
- To assist CBER Medical Officers and researchers in identifying instances where certain vaccines or combinations of vaccines might have harmful effects

Data Science Solution

The U.S. Food and Drug Administration (FDA) Center for Biologics Evaluation and Research (CBER) is responsible for protecting public health by assuring the safety and efficacy of biologics, including vaccines and blood products.

CBER's current surveillance process, which requires resource-intensive manual review by expert medical officers

Does not scale well to short-term workload variation and limits long-term improvements in review cycle-time.

Predicting Customer Response

Problem Statement

- Analyze data for one recent promotional campaign using hotel, stay, and guest data
- A probabilistic model using **Logistic Regression and Support Vector Machines (SVM)** is created which is capable of predicting customer response to the promotion
- Churn Analysis is done using **RFM (Recency Frequency and Monetization) model**
- This finding represents millions of dollars of savings per promotional campaign for IHG

Data Science Solution

InterContinental Hotels Group (IHG) wants to know how a customer will respond to a given promotional campaign.

They want predict customer-by-customer response to a promotional campaign in order to better understand and increase return on investment (ROI) campaign.

Google Flu Trends

- Google predicted flu trend two weeks ahead of CDC (Center for Control Disease and Prevention) data using **time series seasonal model**.
- New machine learning models are estimating which cities are most at risk for spread of the Ebola virus.
- Prediction model is built on various other data sources such as hospitals and census data

Predicting Election Outcome

- The Obama campaigns in 2008 and 2012 are credited for their successful use of social media and data mining.
- Helped Obama save the president's candidacy. In Chicago, the campaign recruited a team of behavioral scientists to build an extraordinarily sophisticated models.
- Targeted each US State individually based on people interest
- The White House Names Dr. DJ Patil as the First U.S. Chief Data Scientist, Feb. 18th 2015

Numbers nerd Nate Silver's forecasts prove all right on election night

FiveThirtyEight blogger predicted the outcome in all 50 states, assuming Barack Obama's Florida victory is confirmed

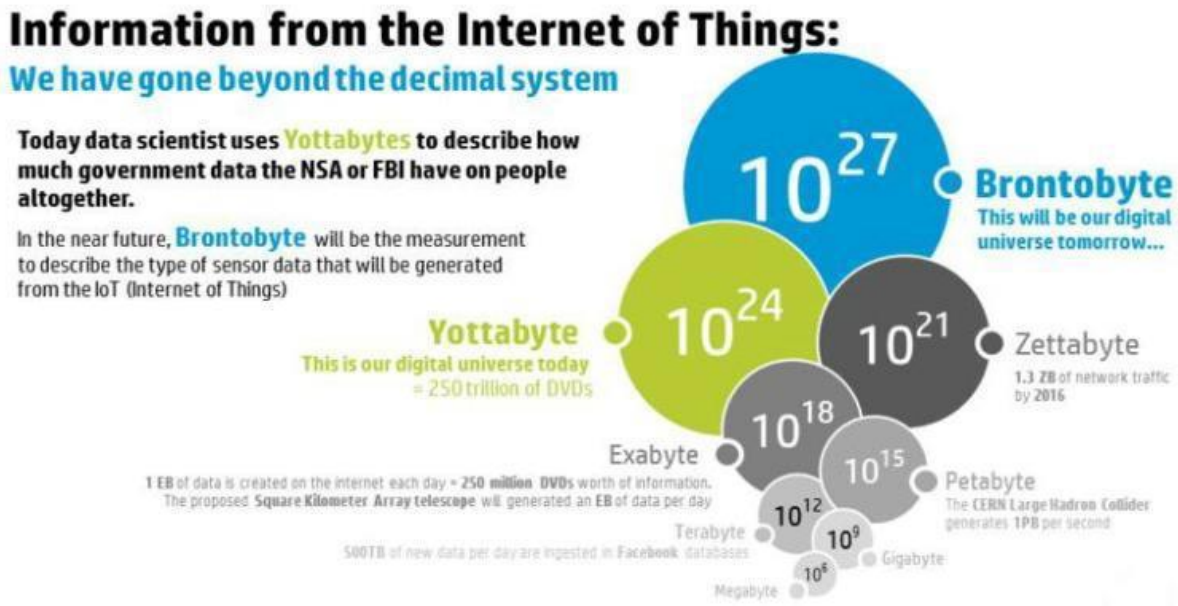
Luke Harding
guardian.co.uk, Wednesday 7 November 2012 10.45 EST

Internet of Things (IoT)

- The Internet of Things is rapidly growing. It is predicted that more than 25 billion devices will be connected by 2020.
- The Internet of Things (IoT) will soon produce a massive volume and variety of data at unprecedented velocity.

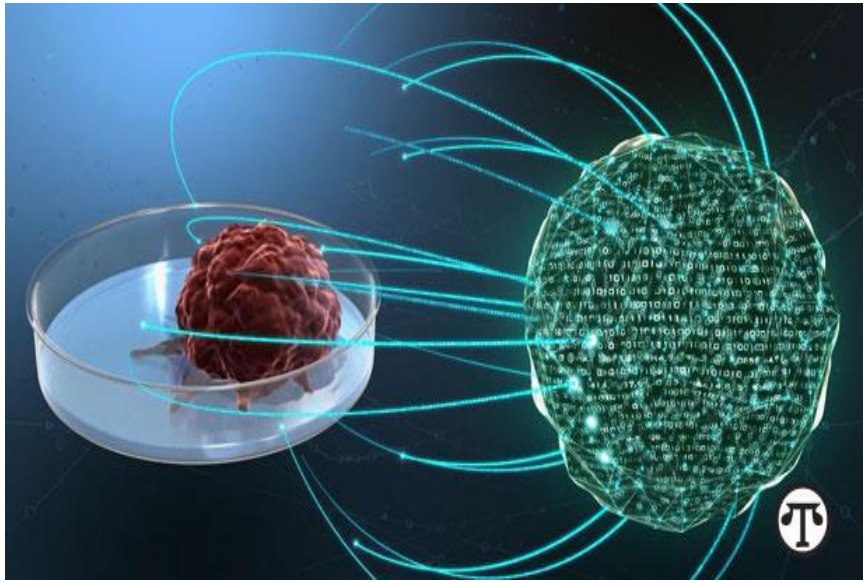
- Connected Homes
- Connected Appliances
- Connected Businesses

  - **TherML:** Occupancy prediction for thermostat control using Machine Learning with device data



Cancer Research

- Identify patterns in DNA sequences that are potentially linked to cancer.
- Leverage sophisticated pattern recognition and machine learning algorithms to **learn unique patterns** for both healthy and infected patients
- **Sequence alignment algorithm** is used to quickly match DNA patterns associated with cancer.
- Employ the power of big data analytics and high-performance computing.



Predicting Consumer Sponsored Search

There are around 30 billion search requests a month. Google Adwords and Adsense as a data product

- Google revenue around $50 bn/year from marketing (which is 97% of the companies revenue)
- Sponsored search uses an auction:
  - a pure competition for marketers trying to win access to consumers.
- In other words, a competition for **machine learning models** of consumers
  - models that predict their likelihood of responding to the ad and of determining the right bid for the shown ad.

Cyber Security

- Data Science to prevent threats that get past other endpoint using log data.
- **Invincea** (security solution provider) combines deep learning (variant of Neural Networks) with behavioral monitoring to prevent Malwares and Viruses.

Scientifically developed approaches to combatting cyber threats.

WhatsApp Data Analytics

- Primary goal is to explore the feasibility of collecting and using WhatsApp Public Groups data for social science research.
- For instance, if you want to study how people have been using WhatsApp for job search.
  - Also sports, politics, entertainment etc.

- https://github.com/gvrkiran/whatsapp-public-groups

- Paper
  - https://users.ics.aalto.fi/kiran/content/whatsapp.pdf

Demand Supply – Uber, Careem, Bykea

Estimate the number of drivers and rides in each area at particular time of the day.

Build a predictive model to predict number of riders required in each area of Karachi that would help managing demand and supply

If the driver is away from any customer who is looking for a ride, they are most likely going to loose that customer due to higher waiting time

Use traffic updates, weather data in addition to improve predictions.

- Provide a great tool for departments who are involved in urban planning.
- This idea can be used for call centers and support related companies where resource management is a challenge.

Text Classification using Wikipedia and DBPedia

- Can be used for Facebook Posts, Blog Posts etc.
- https://www.nuxeo.com/blog/mining-wikipedia-with-hadoop-and-pig-for-natural-language-processing/

- Instead manually annotating articles, one should try to benefit from an existing annotated and publicly available text corpus **Wikipedia and DBPedia** that deals with a wide range of topics.
- Assign a category for unlabeled articles

Application Tracking System (ATS)

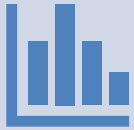
- Indeed, Monster.com, Rozee.pk and other vendors are looking for such services.

Improve the ATS where you use Data Science and Machine Learning algorithm to choose best resumes from big pool and send it to companies who subscribed to your ATS service

Using semantic and syntactic relations between text, rather than just matching keywords.

Educational Data Analytics and Reporting

- Use University and College data to perform data analytics and predictions
  - Enrollment Projection Model
  - Predict student's chances of admission based on credentials
  - Predict student's GPA based on past and current standings
  - Identify the best series of courses for success



Habib University data can be used for this project

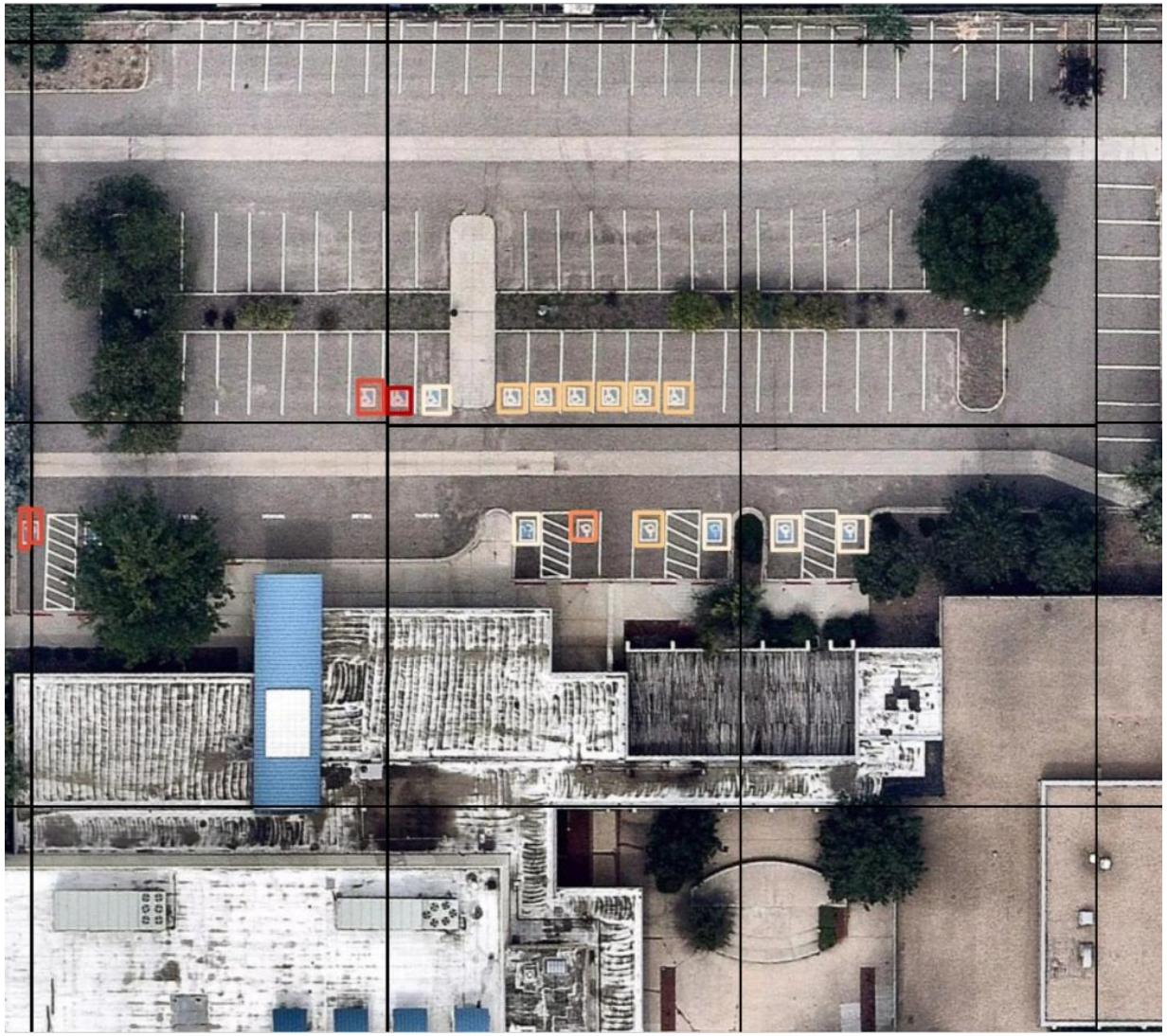
https://www.datakind.org/projects/improving-college-success-through-predictive-modeling

Satellite Image Annotation

An example of labelling handicap parking spots.

Tells exact numbers and location for handicapped people while driving.

- Annotate important places on map using satellite images
- Output is the labelled object exact location on map.
- Deep Learning as potential algorithm.



End of Part 2

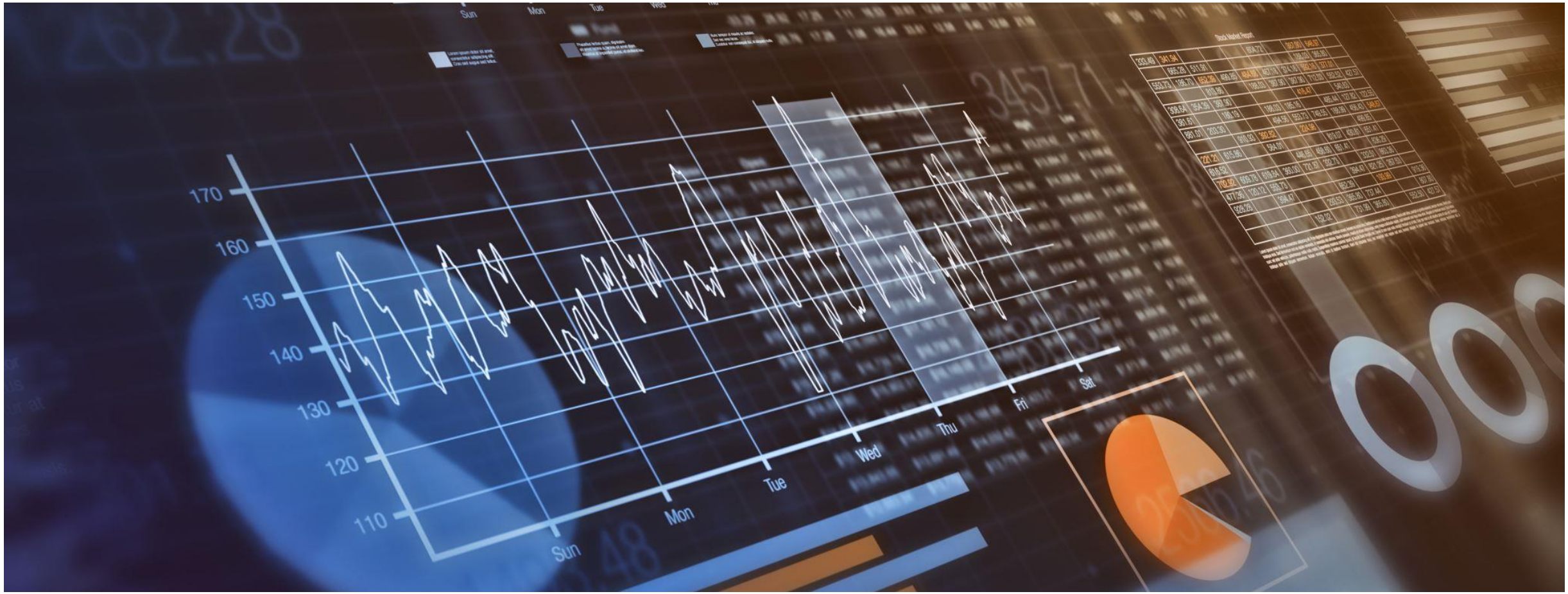
Introduction to Data Science

Week 1 – Part 3 – All about Data

CS 457 - L1 Data Science

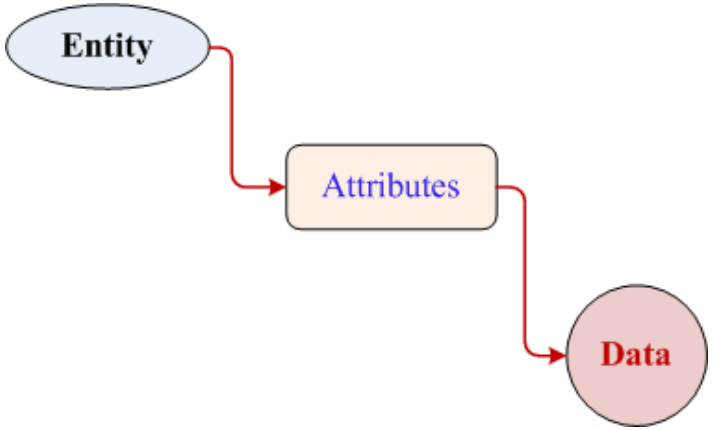
Zeehasham Rasheed

All about Data



Data Description

- A **Data** or **Dataset** is a table containing measurements of objects.
  - Every row is an object.
  - Every column is one attribute of an object.
  - Every cell is the measurement of the corresponding object and attribute.
- **Entity:** A particular thing is called entity or object.
- **Attribute.** An attribute is a measurable property of an entity.
- Computer can manage all type of data (e.g., audio, video, text, etc.)



NAME	AGE	GENDER	SALARY	EMPLOYER
:				
:				
ABCD	34	F	40000	XYZ
:				
:				

Day	Weather	Day Length	Heartbeats/ min
Day 1	Hot	9	85
Day 2	Cloudy	9	93
Day 3	Cold	9	84
Day 4	Cloudy	11	71
Day 5	Hot	11	73
Day 6	Hot	9	97
Day 7	Cold	10	79
Day 8	Hot	10	68

Data Measurement

- In general, there are many data types that can be used to measure the attributes of an entity.
- A good understanding of data **scales** (also called scales of measurement) is important for each attribute
- Depending on the scales of measurement, different operations can be performed to derive useful information in the form of
  - patterns, associations, anomalies or similarities from a volume of data.

Operations on Data

Following FOUR properties (operations) of data are pertinent.

#	Property	Operation	Type
1.	Distinctiveness	= and ≠	Categorical (Qualitative)
2.	Order	< , ≤ , > , ≥	
3.	Addition	+ and -	Numerical (Quantitative)
4.	Multiplication	* and /	

NOIR classification

- The NOIR scale is the fundamental building block on which the **extended data types** are built.
- NOIR is the Classification of scales of Measurement

The mostly recommended scales of measurement are

N: **N**ominal

O: **O**rdinal

I: **I**nterval

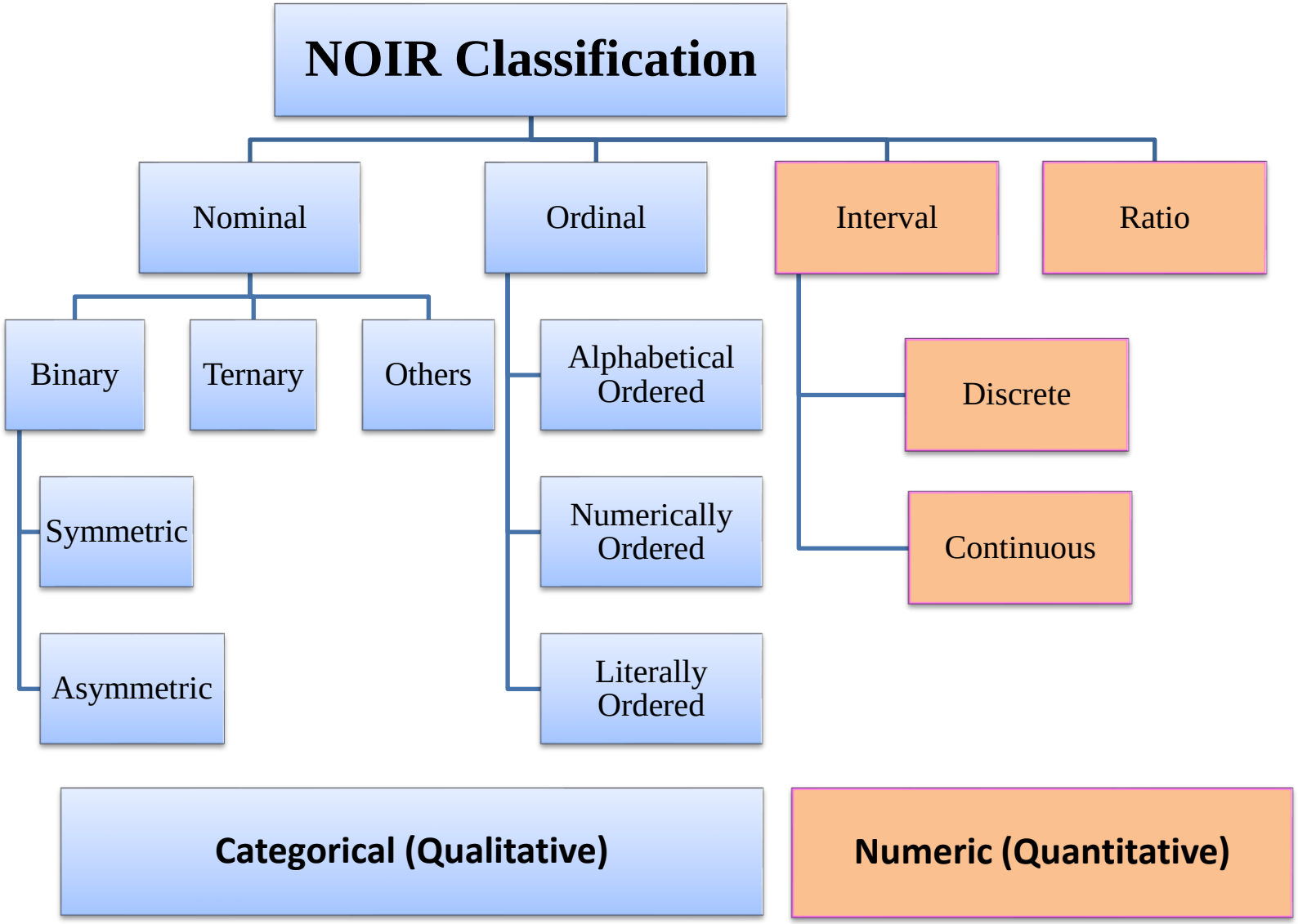
R: **R**atio

NOIR Definition

- Nominal (with distinctiveness property only)
- Ordinal (with distinctive and order property only)
- Interval (with additive property + property of Ordinal data)
- Ratio (with multiplicative property + property of Interval data)

  - Further, nominal and ordinal are collectively referred to as categorical or qualitative data.
  - Whereas, interval and ratio data are collectively referred to as quantitative or numeric data.

NOIR Hierarchy



Nominal scale

- **Definition**

A variable that takes a value among a set of mutually exclusive codes that have no logical order is known as a nominal variable.

- **Examples**

Gender	Used letters or numbers { M, F} **or** { 1, 0 }
Blood groups	Used string {A , B , AB , O }
Rhesus (Rh) factors	Used symbols {+ , - }
Country code	US PK

Nominal scale

Note

- The nominal scale is used to label data categorization using a consistent naming convention.
- The labels can be numbers, letters, strings, enumerated constants or other keyboard symbols.
- Nominal data thus makes “category” of a set of data.
- The number of categories should be two (binary) or more (ternary, etc.), but countably finite.

Nominal scale

A nominal data may be numerical in form, but the numerical values have no mathematical interpretation.

- For example, 10 prisoners are 100, 101, … 110, but; 100 + 110 = 210 is meaningless. They are simply labels.

Two labels may be identical (=) or dissimilar (≠).

These labels do not have any ordering among themselves.

- For example, we cannot say blood group B is better or worse than group A.

Labels (from two different attributes) can be combined to give another nominal variable.

- For example, blood group with Rh factor (A+ , A- , AB+, etc.)

Binary scale

- **Definition**

A nominal variable with exactly two mutually exclusive categories that have no logical order is known as binary variable

- **Examples**

  Switch: {ON, OFF}

  Attendance: {True, False}

  Entry: {Yes, No}

  etc.

Note

- A Binary variable is a special case of a nominal variable that takes only two possible values.

Symmetric and Asymmetric Binary Scale

Different binary variables may have equal or unequal importance.

If two choices of a binary variable have equal importance, then it is called symmetric binary variable.

- Example: Gender = {male , female}
  - usually of equal probability.

If the two choices of a binary variable have unequal importance, it is called asymmetric binary variable.

- Example: Food preference = {V , NV}

Nominal Example

What is your gender?

- (•) M – Male
- ( ) F – Female

What is your hair color?

- (•) 1 – Brown
- ( ) 2 – Black
- ( ) 3 – Blonde
- ( ) 4 – Gray
- ( ) 5 – Other

Where do you live?

- (•) A – North of the equator
- ( ) B – South of the equator
- ( ) C – Neither: In the international space station

Operations on Nominal variables

- Summary statistics applicable to nominal data are mode, contingency correlation, etc.
- Arithmetic (+, -, * and /) and logical operations (<, >, ≠ etc.) are not permitted.
- The allowed operations are: comparing and matching
- Nominal data can be visualized using line charts, bar charts or pie charts etc.
- Two or more nominal variables can be combined to generate other nominal variable.
  - Example: Gender (M,F) × Marital status (S, M, D, W)

Ordinal scale

- **Definition**

Nominal data that can be ordered are known as ordinal data

- Example:

Shirt size = { S, M, L, XL, XXL}

Note

The values of an ordinal variable can be ordered among themselves

Each pair of values can be compared literally or using relational operators (< , ≤ , > , ≥).

Ordinal Example

How do you feel today?

- (•) 1 – Very Unhappy
- ( ) 2 – Unhappy
- ( ) 3 – OK
- ( ) 4 – Happy
- ( ) 5 – Very Happy

How satisfied are you with our service?

- (•) 1 – Very Unsatisfied
- ( ) 2 – Somewhat Unsatisfied
- ( ) 3 – Neutral
- ( ) 4 – Somewhat Satisfied
- ( ) 5 – Very Satisfied

Operation on Ordinal data

- Usually relational operators can be used on ordinal data.
- Summary measures mode and median can be used on ordinal data.
- Ordinal data can be ranked (numerically, alphabetically, etc.) Hence, we can find any of the percentiles measures of ordinal data.
- Calculations based on order are permitted (such as count, min, max, etc.).
- Correlation can be used as a measure of the strength of association between two sets of ordinal data.
- Numerical variable can be transformed into ordinal variable and vice-versa, but with a loss of information.
  - For example, Age [1, … 100] = [young, middle-aged, old]

Interval scale

- **Definition**

Interval-scale variables are continuous measurements of a roughly linear scale.

- Example:

  temperature, calendar dates, latitude, longitude etc.

Note

- Interval data are with well-defined interval.
- Interval data are measured on a numeric scale (with +ve, 0 (zero), and –ve values).
- Interval data has a zero point on origin. However, the origin does not imply a true absence of the measured characteristics.
  - For example, temperature in Celsius and Fahrenheit; 0^o does not mean absence of temperature, that is, no heat!
  - consider this: 10 degrees C + 10 degrees C = 20 degrees C. 20 degrees C is not twice as hot as 10 degrees C. When converted to Fahrenheit, it is clear that 10C=50F and 20C=68F, which is clearly not twice as hot.

Operation on Interval data

- We can add to or from interval data.
  - For example: date1 + x-days = date2
- Subtraction can also be performed.
  - For example: current date – date of birth = age
- Negation (changing the sign) and multiplication by a constant are permitted.
- All operations on ordinal data defined are also valid here.
- Linear (e.g. cx + d) or Affine transformations are permissible.
- Other one-to-one non-linear transformation (e.g., log, exp, sin, etc.) can also be applied.

Operation on Interval data

Note

- Interval data can be transformed to nominal or ordinal scale, but with loss of information.

- Interval data can be visualized using histogram, frequency graph polygon, etc.

Ratio scale

- **Definition**

  Interval data with a clear definition of “zero” are called ratio data.

  - Example:

    Intensity of earth-quake on Richter scale, Sound intensity in Decibel, cost of an article, population of a country, income etc.

Note

- All ratio data are interval data but the reverse is not true.
- In ratio scale, both differences between data values and ratios (of non-zero) data pairs are meaningful.
- Ratio data may be in linear or non-linear scale.
- Both interval and ratio data can be stored in same data type (i.e., integer, float, double, etc.)

Operation on Ratio data

- All arithmetic operations on interval data are applicable to ratio data.
- In addition, multiplication, division, etc. are allowed.
- Any linear transformation of a mathematical form can be applied.

NOIR Summary

- **Nominal (Categorical)**
  - categories: qualitative, no implied order or size, *discrete*
    - color, gender, State, Country, …
- **Ordinal (can contains categories)**
  - rank order: *discrete*
    - 1: dislike < 2: neutral < 3: like
    - Only relational operators
- **Numerical (Continuous)** contains numbers
- **Interval**
  - *distance/difference* measures have no meaning; *continuous, integer, floating point*
  - contains zero point on origin. However, the origin does not imply a true absence
    - 0^{o} Celsius does not mean absence of temperature
    - Date
- **Ratio**
  - *Size comparisons* have meaning; *continuous, integer, floating point*
    - 80kg = 2 x 40kg
    - can be 0, differences, ratios provides meaning. For example length, mass etc.

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

More NOIR Examples

- **Nominal:**
  - Name, Gender etc. are Nominal Data Type in the dataset. They are labels of variables which are not quantitative and used to identify the particular entity
- **Ordinal**
  - Class Ranking, Satisfactory Rating etc. are Ordinal Data Type in the dataset. These values have a natural order but are not measurable in numbers such as how satisfied one person is. In Class Ranking, students are measurable by their class standing as a freshman, sophomore, junior, and senior.
- **Interval**
  - Date of Birth, GPA etc. are Interval Data Type in the dataset. They have a natural order and a difference between values. Interval Data Type can't have a true zero value. In this case, one cannot have a value of 0 for both Date of Birth and GPA.
- **Ratio**
  - Tuition Rates, Income etc. are Ratio. They have a natural order and a calculable difference between values. Also, Ratio Data Type can have a true zero value. In this case, we can have both Tuition Rates and Income as 0

NOIR Facts

- **IMPORTANT**:
  - Data type determines what computations and statistical tests are appropriate or inappropriate!
  - e.g., can't calculate mean or average Country or Gender

- **Other non-numeric data types**
  - media (images, video, audio etc.) Often use above types for metadata

- **REMEMBER YOUR NOIR *DATA TYPES***
- Nominal, Ordinal, Interval, Ratio

Data Science Competitions and Datasets

- UCI Machine Learning Repository

http://archive.ics.uci.edu/ml/datasets.html

- Kaggle – helps you learn, work, play and compete with datasets

https://www.kaggle.com/

End of Part 3