

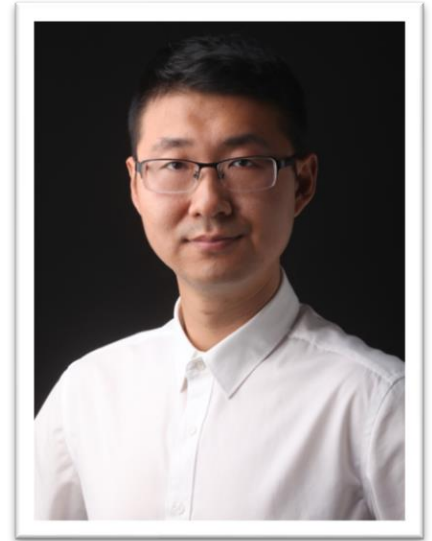
COMP810 Data Warehousing and Big Data

Introduction to Big Data

Dr Weihua Li

Dr Weihua LI

Senior Lecturer & Programme Director for PG Studies



- **Teaching Area**

- COMP603/ENSE600 Program Design and Construction
- COMP810 Data Warehouse and Big Data

- **Education**

- PhD, Auckland University of Technology
- MTech in Knowledge Engineering, National University of Singapore

- **Working Exp.**

- Six-year software development experience – Singapore (4.5) & NZ (1.5)
- ASP.NET C#, Java, JavaScript, Python

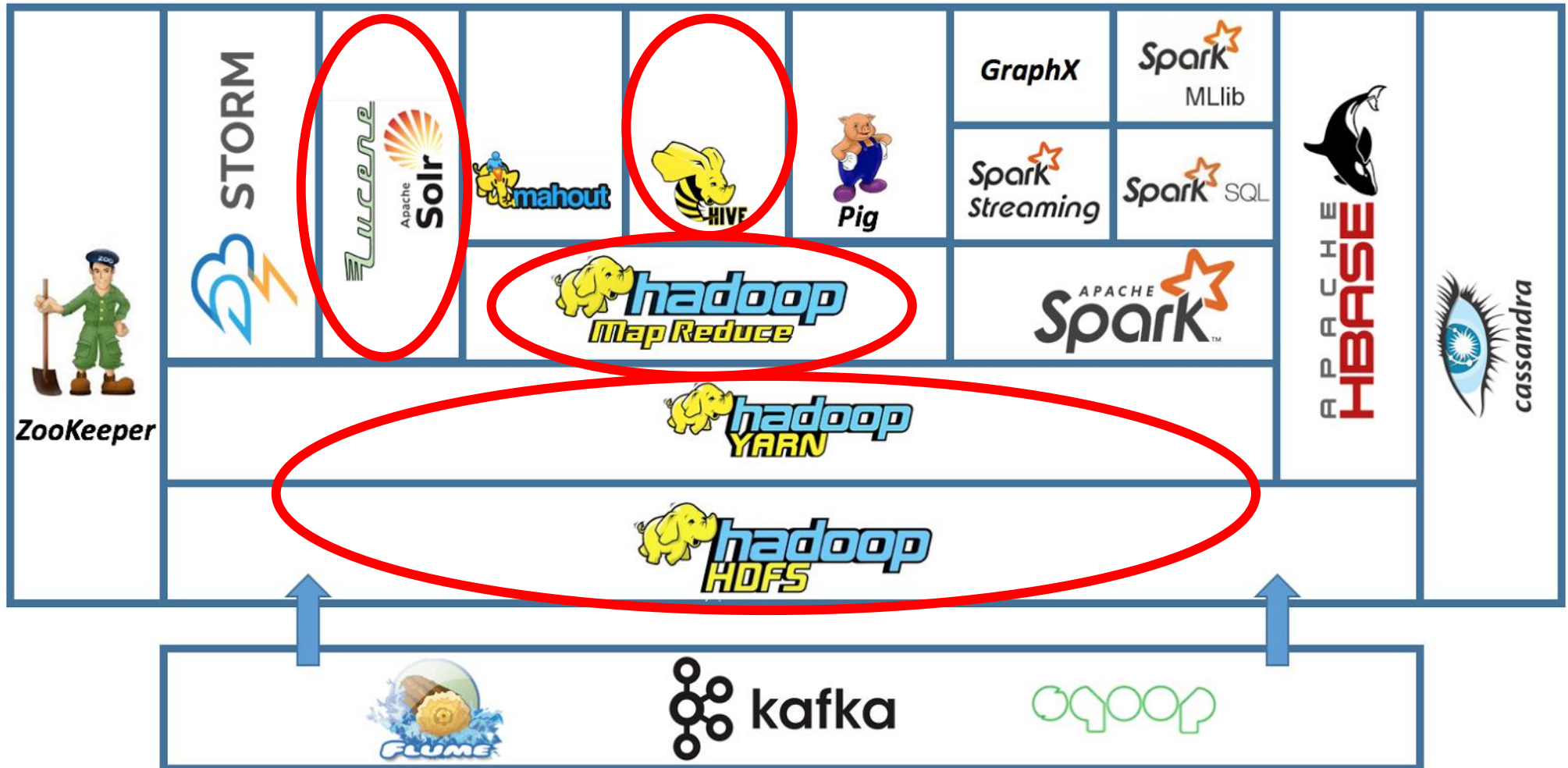
- **Research**

- Artificial Intelligence, Multi-Agent Systems, NLP, Knowledge Graph, Deep Learning

- **Contact**

- Academic: <https://academics.aut.ac.nz/weihua.li>
- LinkedIn: <https://www.linkedin.com/in/liweihua/>
- Email: weihua.li@aut.ac.nz

What You will Learn



Weekly Schedule – Big Data

No	LECTURE	LAB EXERCISE
1	Introduction to Big Data	Preparation and Assignment Focused
2	MapReduce and Hadoop	MapReduce – AWS EMR
3	Big Data with Apache Hive	Hive QL – AWS EMR
4	Search and Analyse Unstructured Big Data (1)	Elasticsearch and Kibana (1)
5	Search and Analyse Unstructured Big Data (2)	Elasticsearch and Kibana (2)
6	Research Seminar: Big Data with Complex Networks	Continue Working on Labs or Assignment

Advice for Your Study

- Grasp concepts, not just technical details
- Practise as much as possible
- Google is your best friend
- Make use of GAI tools, e.g., ChatGPT
- Enjoy!





Outline

- Introduction to Big Data
 - Data Science
 - Data Warehousing and Big Data
 - Types of Big Data
 - Characteristics of Big Data – 5V
- Traditional RDBMS
- Big Data Technologies
- Assignment / Assessment

Introduction to Big Data



Data Science

- **What is Data Science?**

- Data science is an **interdisciplinary field**
- Relate to data mining, machine learning and big data
- Use scientific methods, processes, algorithms and systems to **extract knowledge and insights** from both **structured** and **unstructured data**.

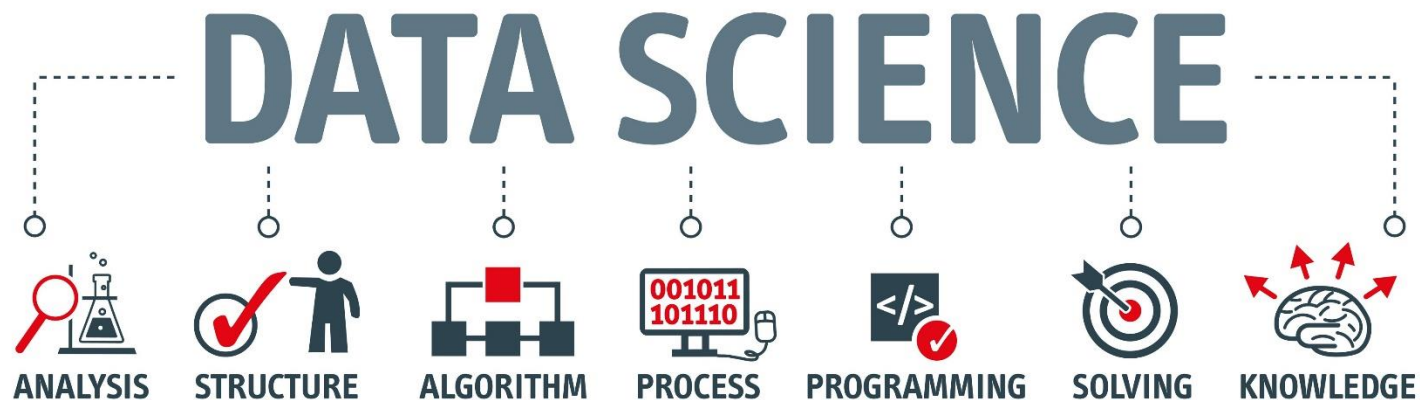
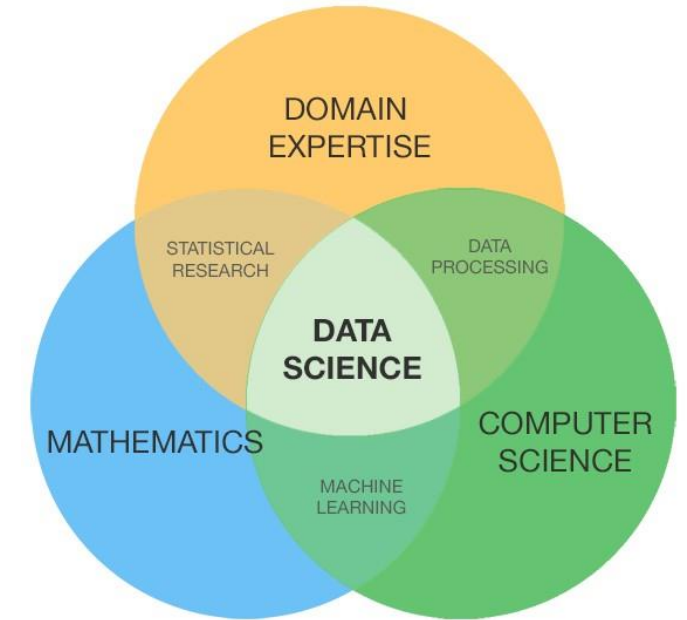


Image by shutterstock from Datanami

Data Warehousing and Big Data

- **Data warehousing is an architecture**, extracting data from data sources, transforming the data and conducting data analysis which helps with decision making.
- **Big Data solution is a technology**, which can manage and analyse large and complicated data sets that may not be easily managed by traditional DBMS.
- Data warehouse is designed with the **clear intention** to make informed decisions. Whereas, Big Data is a repository to hold lots of data but **it is not sure what we want to do with it**.

What is BIG Data



- Big data describes data sets so **large and complex** that they become awkward to work with using standard statistical software [1]
 - i.e. beyond the ability of commonly used software tools to capture, manage, and process data **within a tolerable elapsed time**
- Big data refers to the increase **in the volume of data** that are difficult to store, process, and analyse through traditional database technologies [2]
 - i.e. require **new forms of integration** to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale.

¹Snijders, C.; Matzat, U.; Reips, U.-D. (2012). "'Big Data': Big gaps of knowledge in the field of Internet Science". *International Journal of Internet Science* **7**: 1–5

²Ibrahim A T H, et al. (2015). "The rise of "big data" on cloud computing: Review and open research issues". *Information Systems* **47**: 98–115

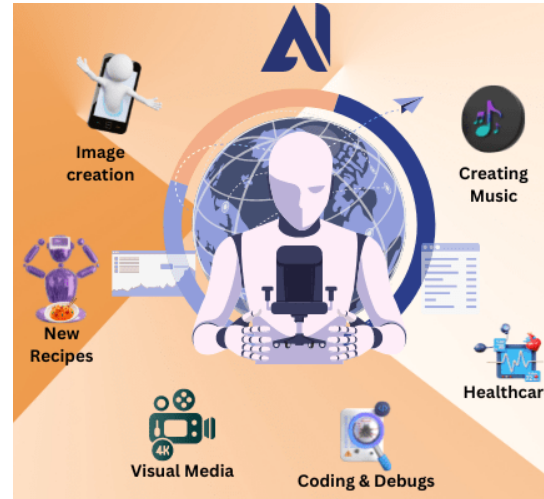
How Big data are Generated



Social Media and Networks
(all of us are generating data)



Mobile Devices
(tracking all objects all the time)



Human-AI Interactions



Scientific Instruments
(collecting all sorts of data)



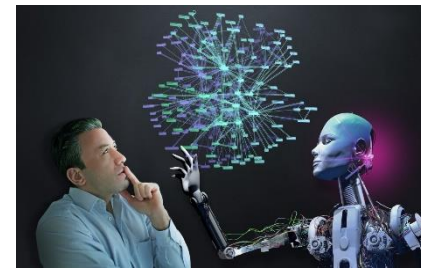
Sensor Technology and Networks
(measuring all kinds of data)

A Shift of Data Generation Model

- The Model of Generating/Consuming Data has Changed
- **Old Model:** A few companies generate data, all others consume data
- **New Model:** all of us are generating data, and consuming data



In the future...



Type of Big Data

- **Structured Data**

- Structured data is most often categorised as **quantitative data**
- The data that fits neatly within fixed fields and columns in **relational databases**.
- Examples of structured data include names, dates, addresses, credit card numbers, stock information, geolocation, and more.

User					
UserID	User	Address	Phone	Email	Alternate
1	Alice	123 Foo St.	12345678	alice@example.org	alice@neo4j.org
2	Bob	456 Bar Ave.		bob@example.org	
...
99	Zach	99 South St.		zach@example.org	

Order	
OrderID	UserID
1234	1
5678	1
...	...
5588	99

LineItem		
OrderID	ProductID	Quantity
1234	765	2
1234	987	1
...
5588	765	1

Product		
ProductID	Description	Handling
321	strawberry ice cream	freezer
765	potatoes	
...	...	
987	dried spaghetti	

Type of Big Data (cont.)

- **Semi-Structured Data**

- A form of structured data
- NOT conform to the tabular structure of data models associated with relational databases or other forms of data tables.
- Contain tags or other markers to separate semantic elements, and enforce hierarchies of records and fields within the data.
- Also known as self-describing structure.

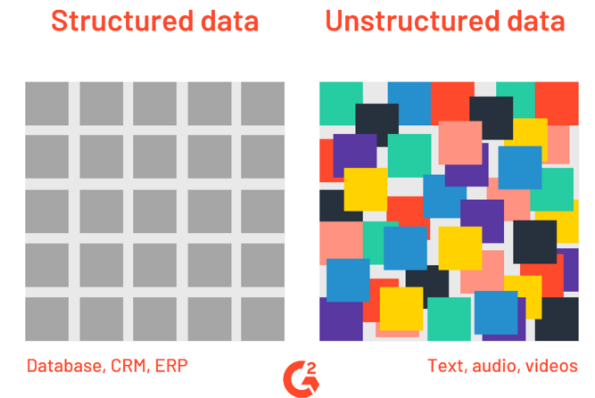
```
<?xml version="1.0" encoding="UTF-8"?>
<Print_Records>
  <form1>
    <Name>Ego ille</Name>
    <Address>345 Park Aven</Address>
    <City>San Jose</City>
    <State>CA</State>
    <ZipCode>94087</ZipCode>
    <Country>USA</Country>
  </form1>
  <form1>
    <Name>Johnson</Name>
    <Address>1 Almaden Blvd</Address>
    <City>San Jose</City>
    <State>CA</State>
    <ZipCode>94089</ZipCode>
    <Country>USA</Country>
  </form1>
</Print_Records>
```

XML

```
{
  "orders": [
    {
      "orderno": "748745375",
      "date": "June 30, 2088 1:54:23 AM",
      "trackingno": "TN0039291",
      "custid": "11045",
      "customer": [
        {
          "custid": "11045",
          "fname": "Sue",
          "lname": "Hatfield",
          "address": "1409 Silver Street",
          "city": "Ashland",
          "state": "NE",
          "zip": "68003"
        }
      ]
    }
  ]
}
```

JSON

Type of Big Data (cont.)



- **Unstructured Data**

- Unstructured data is most often categorized as **qualitative data**
- Cannot be processed and analysed using **conventional tools and methods**
- Examples of unstructured data include **text, audio, video**, mobile activity, social media activity, etc.

- **How to deal with Unstructured Data?**

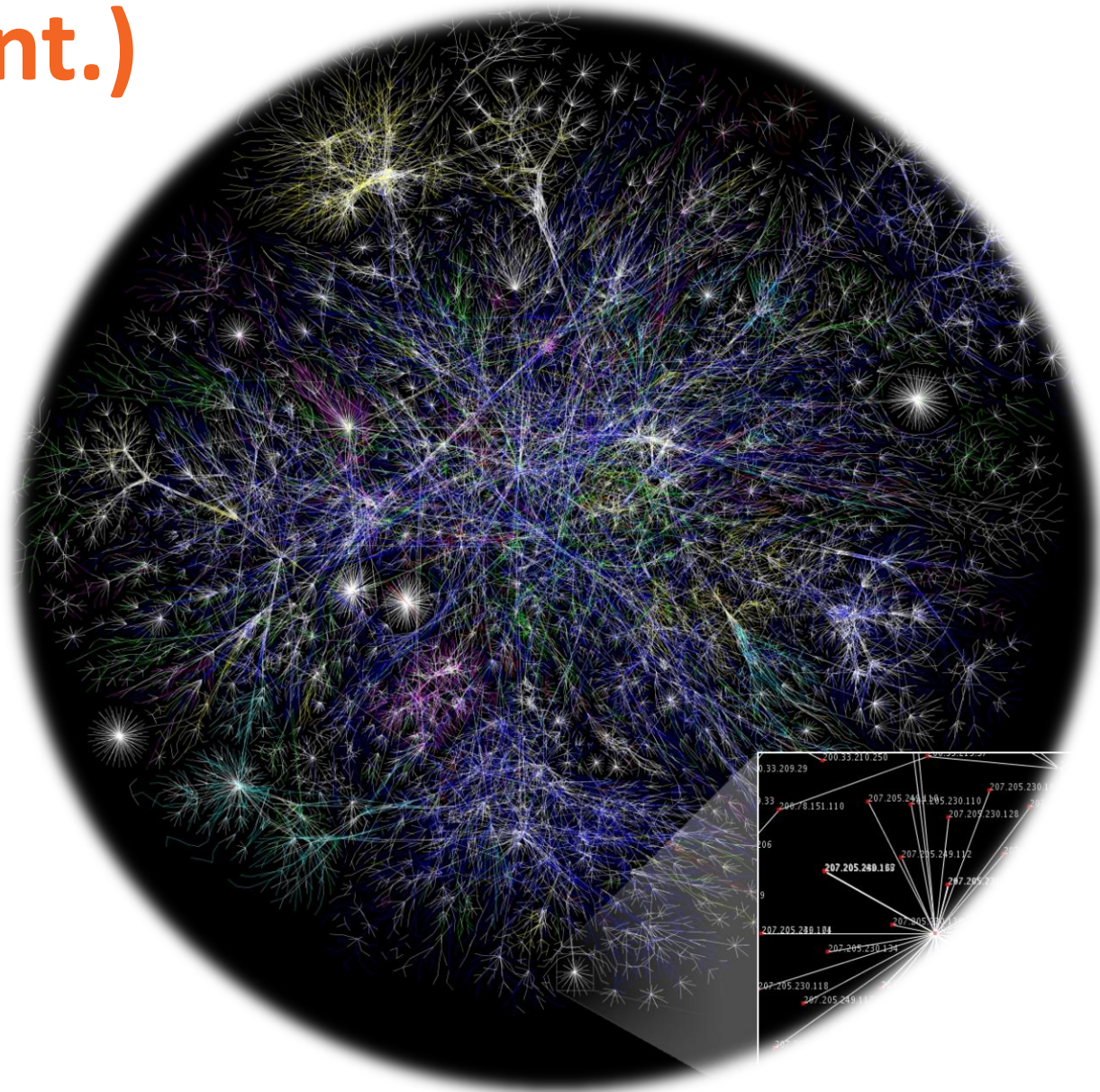
GENE 1 PROTEIN 2

Accordingly, treatment of the **IRF-4** GENE -positive cell line BV-173, SD-1 and RPMI-8226 with AzadC had no effect on IRF-4 expression.

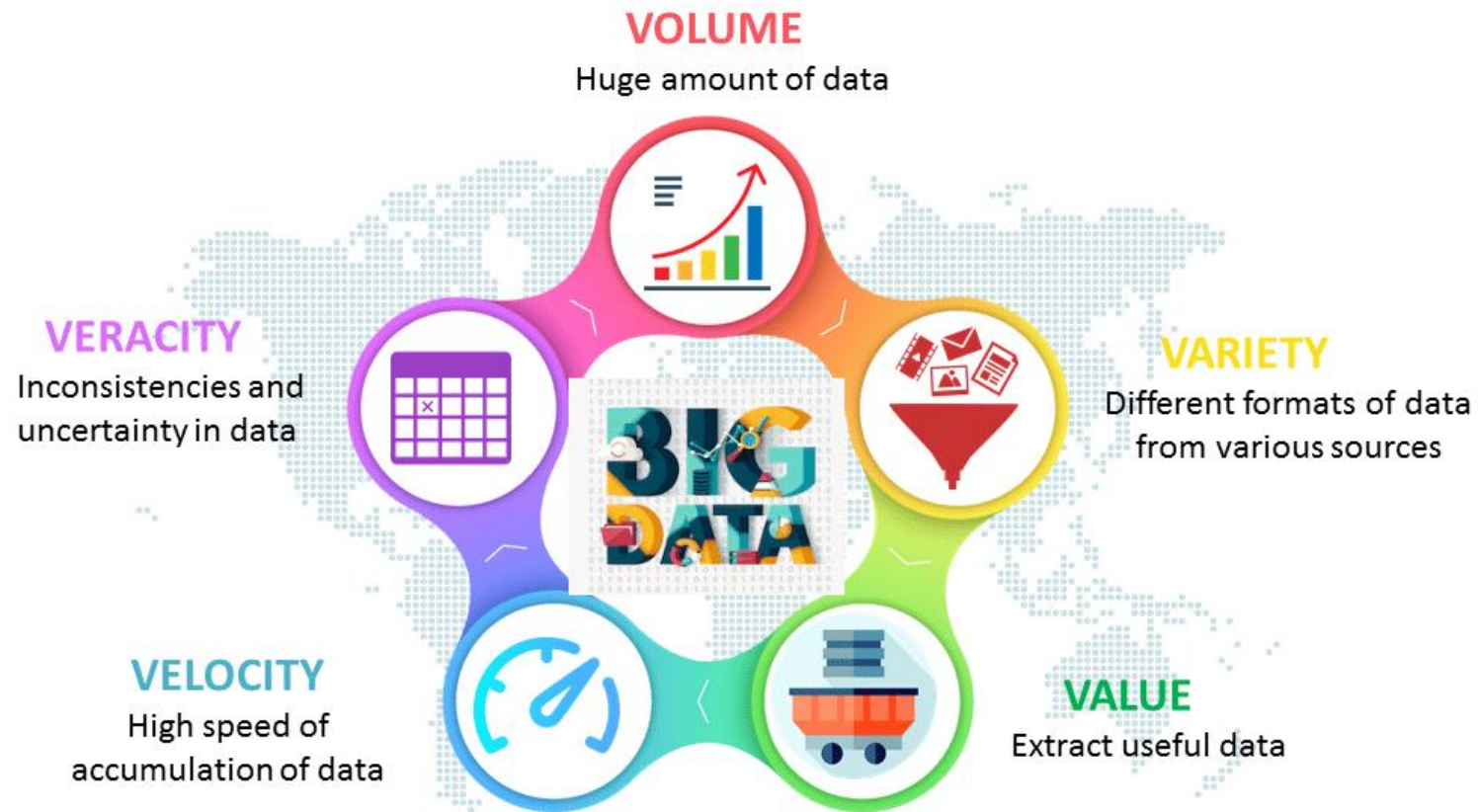


Type of Big Data (cont.)

- **Complex Network**
 - Node: vertex, entity
 - Link: edge, relationship
 - Dynamic heterogeneous network



Characteristics of Big Data – 5V



Big Data Companies



GNIP

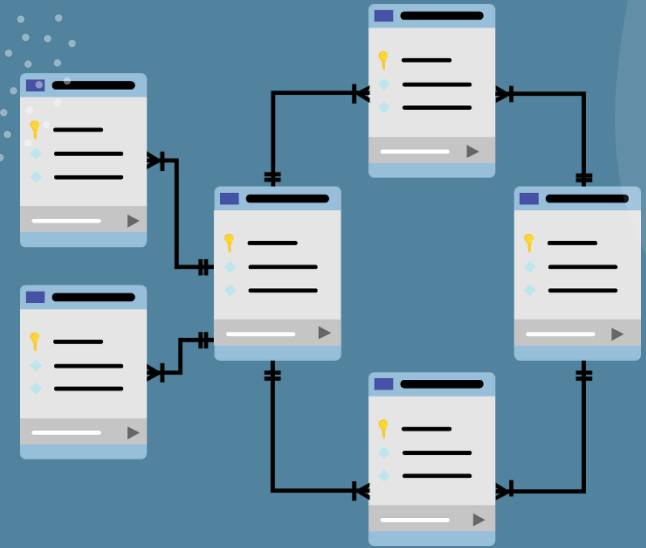
- Aggregate several TB of new social data per day
- Store by using [Amazon S3](#)

A screenshot of The Climate Corporation website. The header includes the company logo and navigation links: Look Up Policy, Contact Us, and Agent Login. Below the header are two buttons: FOR GROWERS and FOR AGENTS. The main content area features a large image of a smiling man in a field. Overlaid on the image is the text "Total Weather Insurance" and "Protect Your Profits From Bad Weather". Below this text are four numbered icons: 1. Get Your Weather Risk Report, 2. Get Custom Weather Insurance Plan, 3. Weather Happens, and 4. Get Paid Automatically. On the right side of the image, there is a green box with the text "Start Here" and "Get your FREE WEATHER RISK ANALYSIS". Below this text are two input fields: "County or Zip Code" and "Corn". At the bottom of the green box is a button labeled "Get Started".

The Climate Corporation

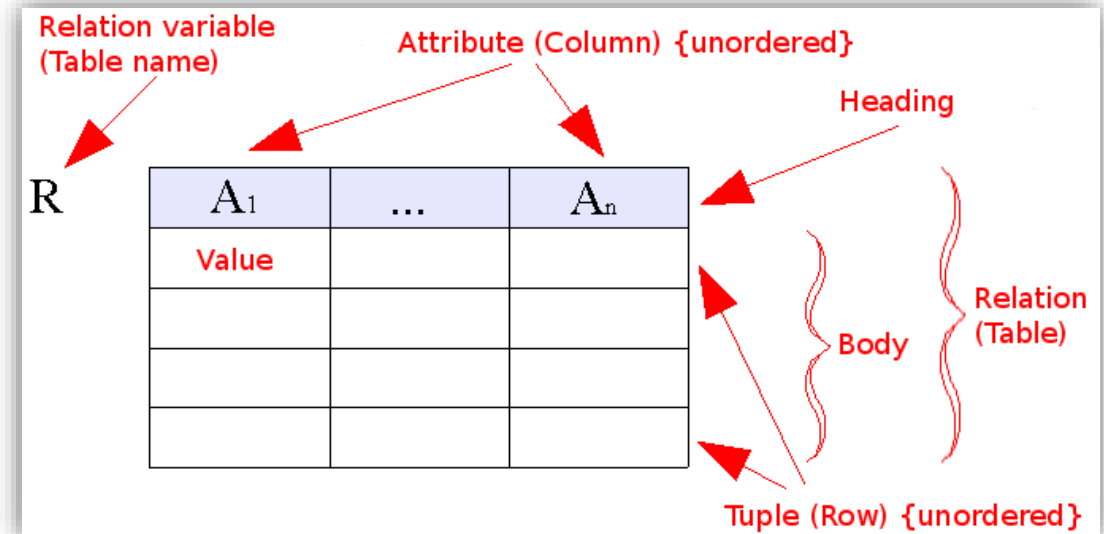
- 14 TB of historical weather data
- All computation done on [Amazon EC2](#)

Traditional Relational Database Management System (RDBMS)



Traditional RDBMS

- **Database (DB):**
 - an organised collection of **data**
 - **Relational DBs** store data in tables
- **A table** in a database consists of:
 - rows & columns
 - Rows: records
 - Columns: attributes of the records



LOCATION_ID	STREET_ADDRESS	POSTAL_CODE	CITY
10001297	Via Cola di Rie	00989	Roma
110093091	Calle della Testa	10934	Venice
12002017	Shinjuku-ku	1689	Tokyo
13009450	Kamiya-cho	6823	Hiroshima
14002014	Jabberwocky Rd	26192	Southlake
15002011	Interiors Blvd	99236	South San Francisco
16002007	Zagora St	50090	South Brunswick
17002004	Charade Rd	98199	Seattle
1800147	Spadina Ave	M5V 2L7	Toronto

Traditional RDBMS (cont.)

- **Structured Query Language (SQL)**
 - The industry standard database query language (Relational Database)
 - Data Definition Language (DDL): create the database and relation structures
 - Data Manipulation Language (DML): perform insertion, modification, deletion of data from relations;
 - Data Retrieval: perform simple and complex queries.
- **Examples:**
 - `SELECT NAME, ID FROM STUDENT WHERE GENDER='MALE';`
 - `INSERT INTO STUDENT VALUES (6, 'MARY', 'FEMALE');`
 - `DELETE FROM STUDENT WHERE NAME='JERRY';`
 - `DROP TABLE STUDENT;`

See more: <https://www.w3schools.com/sql/>

Exercises: <https://www.sql-practice.com/>

Traditional RDBMS (cont.)

- **DBMS:**

- a set of **software programs** that controls the organization, storage, management, and retrieval of data in a database.
- is the intermediary between the user and the database

- **Popular DBMS:**





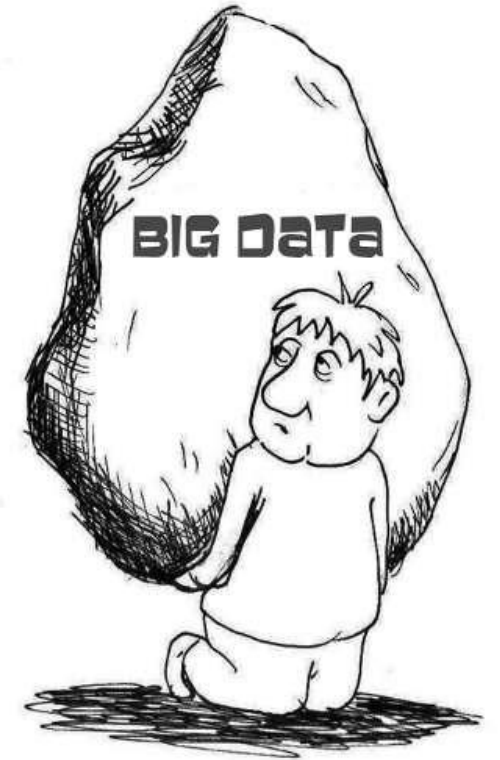

Big Data Technologies

Why Another Set of Tools for Big Data

- **Limitations of Traditional RDBMS**

- Schema on write
- High cost of storage
- Weak support for unstructured data

Poor choice
High volume and variety



Selection of the Hardware Stack (1)

- **Single-node architecture:** Single node refers to computation done on a single server.
- **Multi-node architecture:** Multiple nodes (or servers) that are interconnected and work on the principle of distributed computing. Suitable for hosting data that is in the range of TB and above.
 - Example 1: [Hadoop](#) – multiple servers maintain bi-directional communication to coordinate a job.
 - Example 2: [Elasticsearch](#) - Search and analytics platform, also run on the principle of multi-node computing architecture.

Selection of the Hardware Stack (2)

- **Cloud-based architecture**

- Greatly reduce the [entry barrier](#) in big data analytics.
- Provide a platform that makes it incredibly easy to [provision hardware](#) resources on demand based on the needs of task at hand.
- Reduce the overhead in [managing and maintaining](#) physical hardware.

- **Cloud platforms**

- Amazon Web Services ([AWS](#)), Azure from Microsoft and Google Compute Environment allow enterprises to provision 10s to 1000s of nodes at costs starting as low as 1 cent per hour per instance.

Selection of the Software Stack



Apache Software Foundation

- **Open Source**

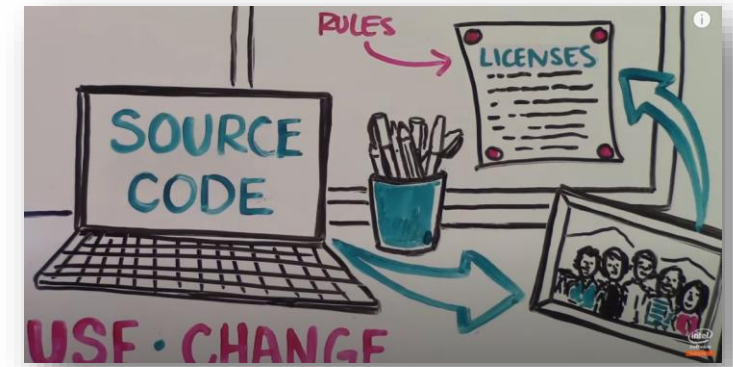
- Software with source code that anyone can inspect, modify, and enhance.
- Source code is released under a license
- Developed in a collaborative public manner

- **The world's largest open-source foundation**

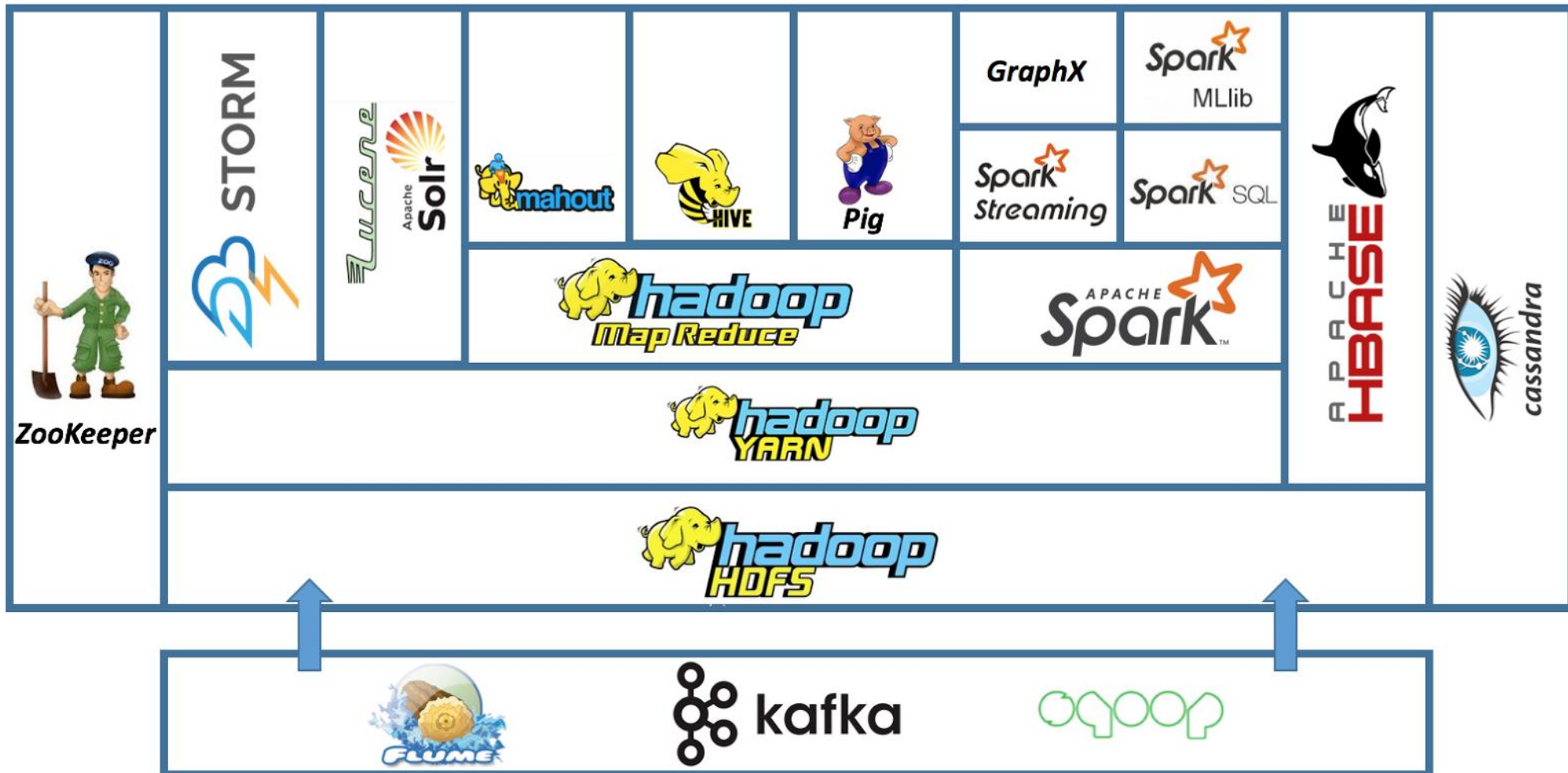
- 200M+ lines of code in stewardship
- 1,119,785,328 lines of code committed
- 350+ Projects and Initiatives
- 300+ Top-Level Projects
- ...

- **Value**

- \$20B+ worth of Apache Open-Source software products are made available to the public-at-large at 100% no cost, and benefit billions of users around the world.



Apache Hadoop Ecosystem



Hadoop and Spark

- **Apache Hadoop**

- Open-source **software framework** for storing and processing **Big Data** in a **distributed manner**
- Consist of multiple projects of Apache Software Foundation.
- Support various types of datasets, e.g., structured, unstructured

- **Apache Spark**

- A project for a **multimode computing framework**
- Run workloads **100x faster**.
- **in-memory cluster** computing that increases the speed of an application
- Store the intermediate processing data **in memory, saving read/write operations**

Hive and Pig

- **Apache Pig**

- A [scripting interface](#) over MapReduce for developers who prefer scripting interface over native Java MapReduce programming.
- Mainly used by Researchers and Programmers, for programming.

```
grunt> customers = LOAD 'hdfs://localhost:9000/pig_data/customers.txt' USING  
PigStorage(',') as (id:int, name:chararray, age:int, address:chararray, salary:int);
```

- **Apache Hive**

- Hive is a [SQL interface](#) over MapReduce for developers and analysts who prefer SQL interface over native Java MapReduce programming.
- Mainly used by Data Analysts for creating reports

```
hive> SELECT c.ID, c.NAME, c.AGE, o.AMOUNT FROM CUSTOMERS c JOIN  
ORDERS o ON (c.ID = o.CUSTOMER_ID);
```

Lucene and Elasticsearch

- **Apache Lucene**

- The project includes a [core search library](#), i.e., Lucene core, and the Solr search server.
- [Lucene Core](#) is a Java library providing powerful indexing and search features
- Solr is a high performance search server built using Lucene Core.
- Solr is highly scalable, providing fully fault tolerant distributed indexing, search and analytics.

- **Elasticsearch**

- Elasticsearch is a [search engine](#) based on the Apache Lucene library.
- It provides a distributed, multitenant-capable [full-text search engine](#) with an HTTP web interface and schema-free JSON documents.
- Elasticsearch is the [most popular enterprise search engine](#) followed by Apache Solr, also based on Lucene.

Other Elements of Hadoop Ecosystem

- Cassandra - 2008 - A key-value pair NoSQL database, with column family data representation and asynchronous masterless replication.
- HBase - 2008 - A key-value pair NoSQL database, with column family data representation, with master-slave replication. It uses HDFS as underlying storage.
- Zookeeper - 2008 - A distributed coordination service for distributed applications. It is based on Paxos algorithm variant called Zab.
- Mahout - 2009 - A library of machine learning algorithms, implemented on top of MapReduce, for finding meaningful patterns in HDFS datasets.
- Sqoop - 2010 - A tool to import data from RDBMS/DataWarehouse into HDFS/HBase and export back.
- YARN - 2011 - A system to schedule applications and services on an HDFS cluster and manage the cluster resources like memory and CPU.
- Flume - 2011 - A tool to collect, aggregate, reliably move and ingest large amounts of data into HDFS.
- Storm - 2011 - A system to process high-velocity streaming data with 'at least once' message semantics.
- Kafka - 2012 - A distributed messaging system with partitioned topics for very high scalability. .



Big Data Assignment



Assessment – Big Data

- Type: Big Data Research Report
- Weight: 50%
- Group Assignment (max 2 people)
- Due date: Refer to the Big Data Report Guideline
- The objective of this assignment is to conduct preliminary research around big data using the techniques delivered in this paper.
- Refer to the last section of this PPT, *Big Data Use Cases*, for picking a topic or coming up with your own.

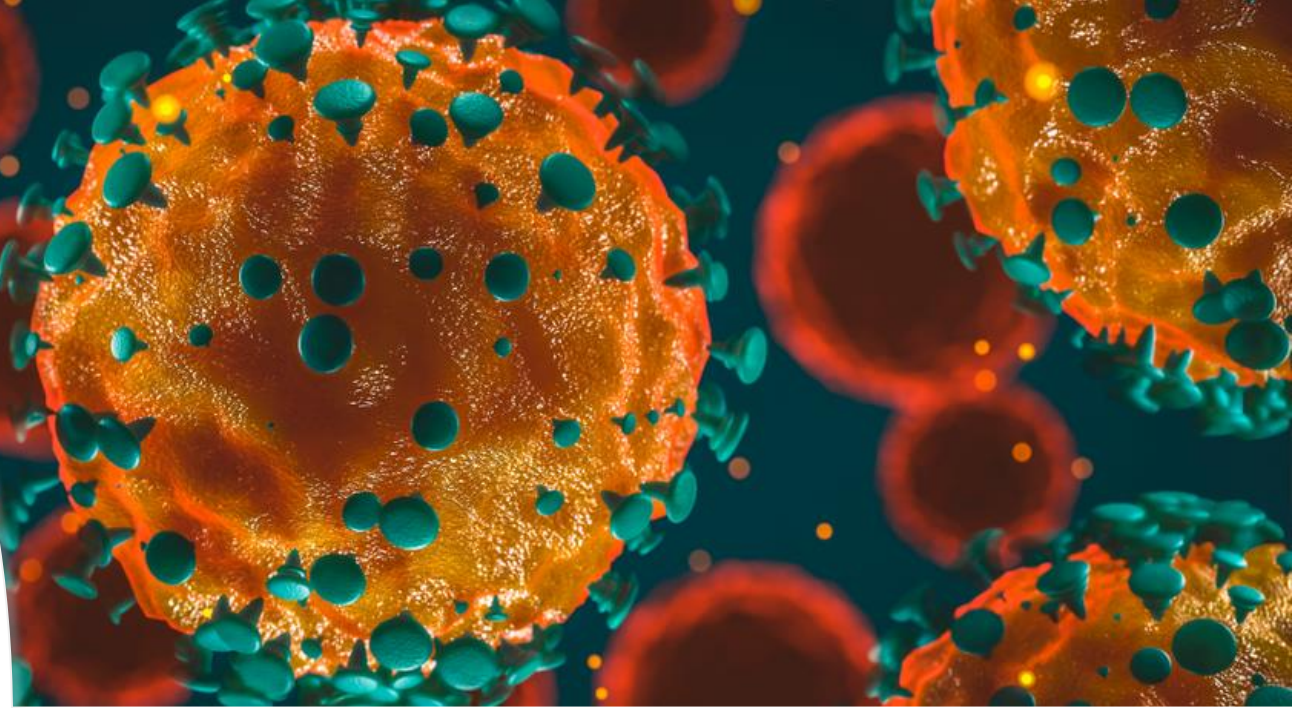


A world map with a dark blue background. Red circles of varying sizes are overlaid on the map, representing data points or population density. The circles are most concentrated in Europe, North Africa, and the Middle East. Labels for various countries and cities are visible in white text, including Sweden, Finland, Norway, Stockholm, Moscow, Warsaw, Ukraine, Istanbul, Turkey, Egypt, Saudi Arabia, Dubai, Libya, Niger, Chad, Sudan, and Algeria.

Use Cases of Big Data

Post-disaster Analytics

- Pandemics and natural disasters can cause large-scale destruction and disruption to humans and economy.
- With the recent outbreak of Covid-19, the pandemic has been recognized as the spread of both physical disease and mental panic, i.e., not only did the coronavirus itself spread very rapidly but so did the related information about the outbreak.
- People tend to express their opinions and concerns about the crisis through online social networks.
- Such information provides strong evidence on what public concerns to be addressed urgently.



News Analytics

- News analysis refers to the measurement of the various qualitative and quantitative attributes of textual (unstructured data) news stories.
- There are a lot of values of mining news media. Business owners can pick up the trends at an early stage. News media also have an effect on the volatility and turnover of stock market.
- It is important to capture the trend of the news topics over time for a particular country/region.

 The New Zealand Herald

 1 news

NZ
WORLD
NEWS

Product Recommendation

A cartoon robot with a green body and blue head, labeled 'Recommender System' on its chest, stands in the center. It holds up three product recommendation cards on the left: a girl's face, a person with glasses, and a person with headphones. On the right, it holds a shelf with boxes containing various product icons like a smartphone, a monitor, a laptop, a shirt, a dress, an umbrella, a pot, and a bag.

- Recommendation System has become an essential component of E-commerce websites. It analyses the interests or preferences of individual consumers for products, either explicitly or implicitly, and makes strategic recommendations to promote sales.
- RS has been recognized as an essential tool to support customers' decisions when searching and selecting products online.
- We are familiar with various flavours of product recommendation techniques used by companies such as Amazon, Netflix, Facebook, LinkedIn, and Google/YouTube.

Influence Diffusion and Maximization



Social influence refers to user's opinions and behaviours are affected by others.



Influence messages diffuse through the online social networks rapidly, affecting users' opinions.



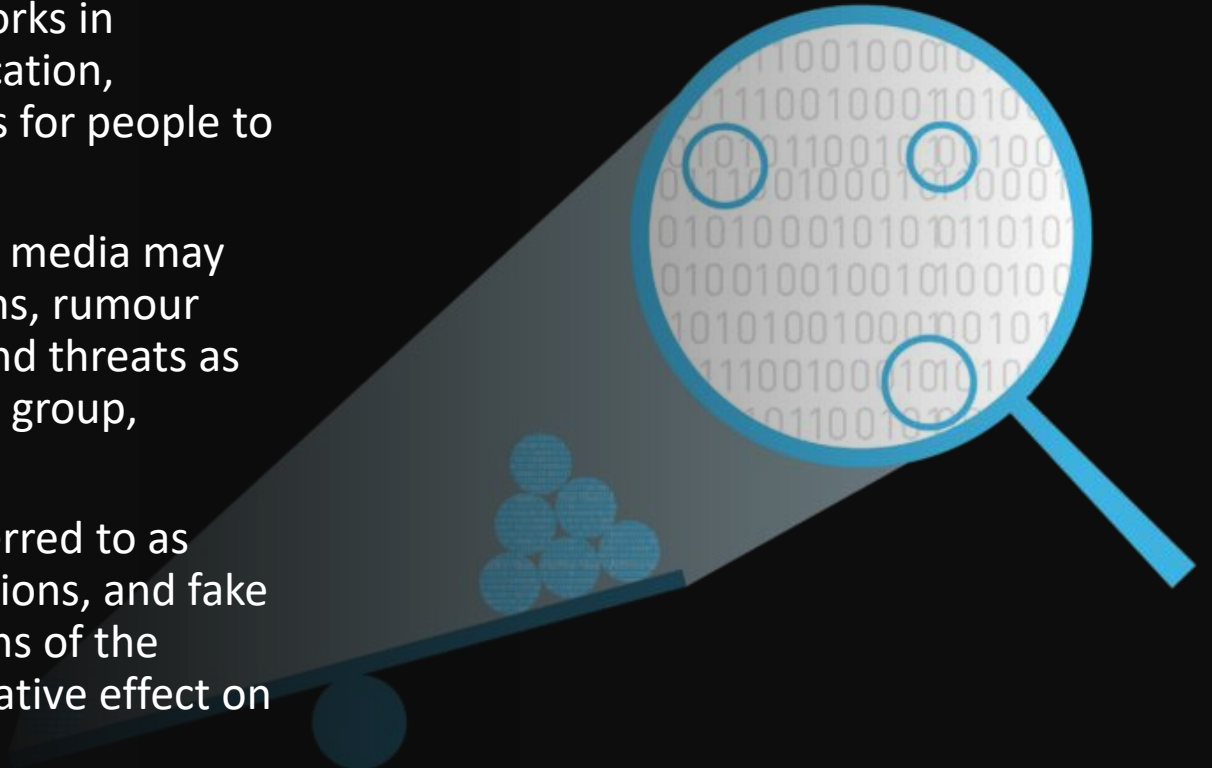
Viral marketing is popular business strategy, leverage Word-of-Mouth (WoM) effect



Influence maximization aims to identify a set of influencers to maximize a particular influence, e.g., adoption a particular product

Anomaly Detection in Online Social Networks

- The increasing popularity of online social networks in different domains, such as entertainment, education, business, medical, etc. provide convenient ways for people to share, communicate, and collaborate.
- On the other side, excessive utilization of social media may cause various types of illegal activities like spams, rumour spreading, and it poses significant challenges and threats as many malicious behaviours, either individual or group, emerges accordingly.
- Anomalies in online social networks can be referred to as outliers, novelties, noise, deviations and exceptions, and fake news can be one of the concrete representations of the anomaly. Such anomalies potentially cause negative effect on influencing and misleading the public opinions



Customer Churn Analysis

Companies from different sectors, e.g., a bank, a retailer, a gaming company, an Internet service provider, a cell phone provider, an airline, or an insurance company, have a strong desire for customer retention and prevent customer churn.



It is well known that keeping an existing customer is often much cheaper than finding a new one.



Customer churn analysis uses big data analytics and machine learning to predict the likelihood of each customer “leaving.”



Businesses then use this data to drive and guide customer retention programs (such as discounts or other incentive programs) to encourage these at-risk customers to stay.

Sentiment Analysis

- Sentiment analysis is an application of text analytics and natural language processing techniques, with the goal of understanding customer sentiment about a certain topic (e.g., a product or service).
- The Web has become an excellent source for assembling consumer opinions. There are now several Web sites containing such opinions, e.g., customer reviews of products, forums, discussion groups, and blogs.
- With the increased adoption of crowd-sourced feedback from customers in online forums and the growth of social networks such as Facebook and Twitter, there is a lot of information available about customer sentiment.



Customer Segmentation

A grocery store may be interested in segmenting its customers by the type of food products they purchase. For example, one segment of customers might be “people who favour Beer and Wine,” while another might be “people who favour beverages.”

Similarly, airlines and hotels are interested in segmenting customers into business travellers versus non-business travellers. Airlines are also interested in “domestic passengers” versus “international passengers.”



Customer segmentation is a common technique used to identify segments of customers that behave similarly with regard to their interaction with the business.



An immediate benefit of such segmentation is the ability to increase marketing efficiency. For example, airlines may customize email campaigns based on effective segmentation to achieve much higher response rates.

Market Basket Analysis (MBA)

- A common use case for retailers is known as market basket analysis (also known as affinity analysis or association mining).
- In this type of analysis, we try to understand the purchasing behaviour of the user. More specifically, with market basket analysis, retailers hope to gain insights into which products tend to be purchased together.
- Market basket analysis often drives store layout design, where items with strong association are placed strategically close to each other, making it more likely that the customer will purchase the related item.
- Retailers can also use the results of market basket analysis for effective marketing campaigns to drive foot traffic into a physical store.



Reference

- Rajaraman, Anand, and Jeffrey David Ullman. *Mining of massive datasets*. Cambridge University Press, 2011.
- Dasgupta, Nataraj. *Practical big data analytics: Hands-on techniques to implement enterprise analytics and machine learning using Hadoop, Spark, NoSQL and R*. Packt Publishing Ltd, 2018.