Big Data Fundamentals and Applications

Introduction to Big Data Analysis

Asst. Prof. Chan, Chun-Hsiang

Master program in Intelligent Computing and Big Data, Chung Yuan Christian University, Taoyuan, Taiwan
Undergraduate program in Intelligent Computing and Big Data, Chung Yuan Christian University, Taoyuan, Taiwan
Undergraduate program in Applied Artificial Intelligence, , Chung Yuan Christian University, Taoyuan, Taiwan

ASBMR 2022

ASBMR®
The American Society for Bone and Mineral Research

SEPTEMBER 9-12, 2022
AUSTIN, TX, UNITED STATES
+ONLINE EXPERIENCE

Outline

- About Dr. Chan
- About the course
- Grading policy
- 5W1H for big data
- Potential issues
- Assignment

About Dr. Chan

Website

目前在職

- 專任助理教授 | 智慧運算與大數據學士學位學程
- 專任助理教授 | 智慧運算與大數據碩士學位學程
- 專任助理教授 | 人工智慧應用學士學位學程
- 人工智慧分析顧問 | 台灣資安鑄造股份有限公司
- 兼任資料科學家 | 中央研究院 社會學研究所

主要學歷

- 博士 | 國立臺灣大學 地理環境資源學系
- 碩士 | 國立臺灣大學 地理環境資源學系
- 碩士 | 實踐大學 食品營養與保健生技學系
- 學士 | 國立臺北教育大學 社會與區域發展學系

主要經歷

兼任助理教授 | 淡江大學 人工智慧學系

博士後研究員 | 臺北醫學大學 醫學系 放射線學科

博士後研究員 | 台北市立萬芳醫院 影像醫學部

資料分析師 | 財團法人資訊工業策進會 資安科技研究所

實習生 | 行政法人國家災害防救科技中心 坡地組

兼任資料科學家 | 香港中文大學 新聞與傳播學院

研究助理 | 臺大地理系 地理計算科學研究室

研究助理 | 臺大地理系 遙測及空間知識實驗室

研究助理 | 國北社發系 土石流防災實驗室

About Dr. Chan

Website

Technical Skills

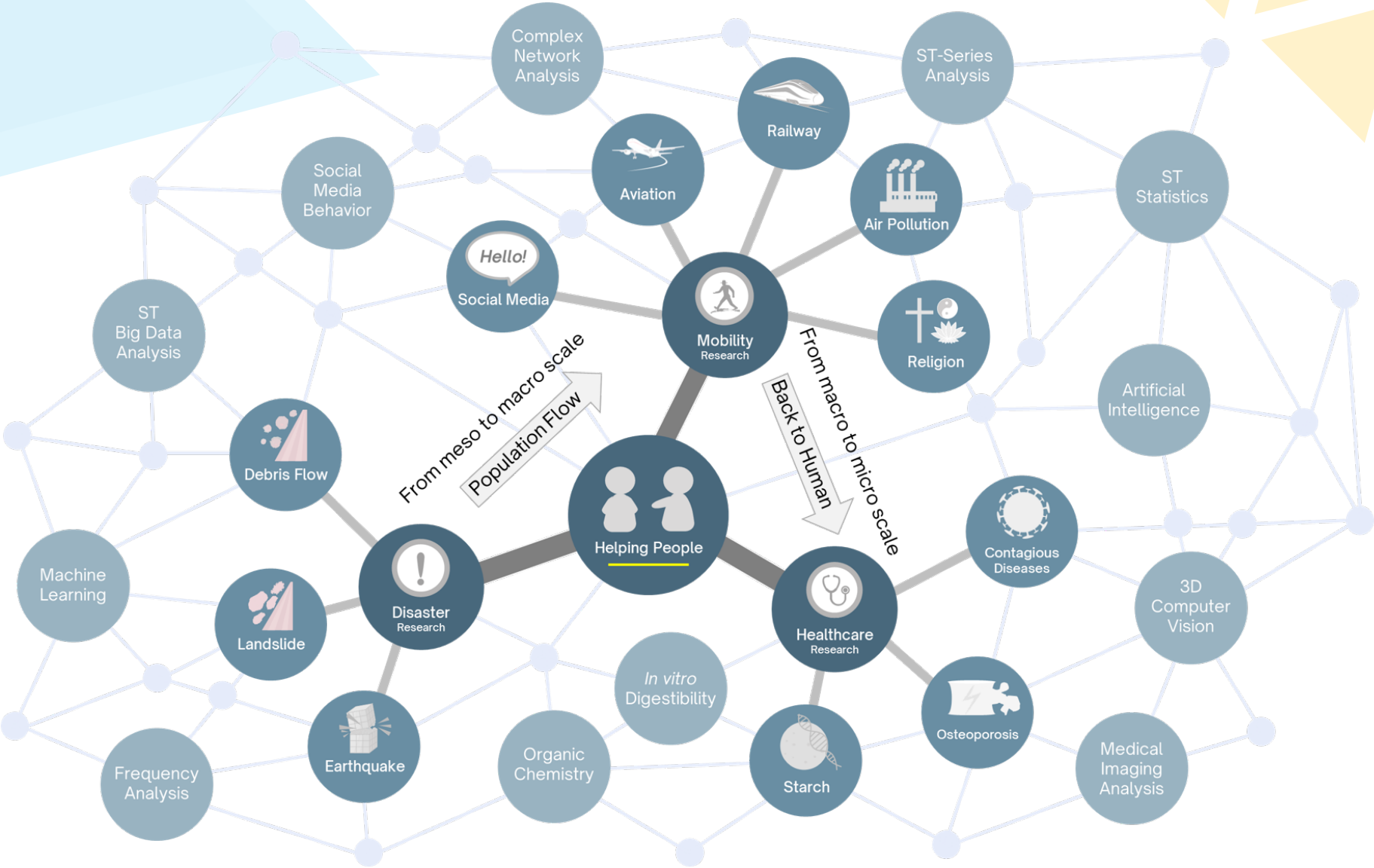
- **Computer Science: Python, Matlab, R**, C#, JavaScript, jQuery, jQueryUI, Android Developmer MySQL, Nodejs, AngularJS, MongoDB, Elasticsearch, Spark, Facebook APIs and Twitter APIs
- **Geography:** GIS (ArcGIS, QGIS, Super GIS), Spatial Statistics, Spatial Database, Complex Network Analysis, Gephi
- **Physics:** Signal Processing (in time sequence and frequency) and Electromagnetic Analysis
- **Food Chemistry:** Starch Science, Resistant Starch, Slowly Digestible Starch, *in vitro* Digestibility, SEM, XRD and HPSEC
- **Chemistry:** Organometallic synthesis, NMR, IR, HPLC, ESI-MASS and pH meter
- **Design:** Illustrator, Photoshop, Dreamwaver and Google SketchUp
- **Marketing:** Google Analysis, Facebook Marketing and Google Trend

Interests

Emergency Medicine, Chinese Medicine, Volleyball, Sport Science, Photography, Tourism, Web and Graphic Design

About Dr. Chan

Website



Course Introduction

In the first semester, we will cover six parts of basic analysis as follows. Due to the time limitation, data collection, database, and data science part will be introduced in the next semester.

1) Data Preprocessing – Numerical & String Analysis
2) Data Explore
3) Big Data Analysis – Classification, Clustering, Association, Regression
4) Story Telling
5) Spatial Analysis
6) Applications

Course Introduction

Week	Date	Content
1	Sep. 16	Introduction to Big Data (Online)
2	Sep. 23	Data Preprocessing – Numerical Analysis
3	Sep. 30	Data Preprocessing – String Analysis
4	Oct. 7	Data Explore
5	Oct. 14	Big Data Analysis – Classification
6	Oct. 21	Big Data Analysis – Clustering
7	Oct. 28	Big Data Analysis – Association
8	Nov. 4	Big Data Analysis – Regression
9	Nov. 11	- Mid-term Exam Week -
10	Nov. 18	Advanced Story Telling
11	Nov. 25	Spatial Analysis – Fundamental Final Project Proposal

Week	Date	Content
12	Dec. 2	Spatial Analysis – Advanced
13	Dec. 9	Applications – eHealth, mHealth, and uHealth
14	Dec. 16	Applications – Smart City
15	Dec. 23	Thanks Giving Holiday
16	Dec. 30	Application – Governing Sustainable Cities
17	Jan. 6	Final Project Presentation
18	Jan. 13	(Final Exams)

Before, during, after class

- Before the class, …
  - Read the materials
  - Search online information
  - (Paper reading)
- During the class, …
  - Lecture
  - Discussion
  - Lab practice
- After the class, …
  - Assignment and project

Grading Policy

- All you have to do is study hard and feel free to ask question when you do not understand.
- I believe that if you fulfill all required items, and then you will pass this course.
- Do not worry about the grade! The most important thing is what you learn from this course.

Assignments	30 %	Mid-Term Pitch	30 %
Discussion	10 %	Final Project	30 %

In the following classes, …

One week before…

Watch recording …

➔ YouTube Link

In class …

14:10 **Lecture**

15:00

15:10 **Lecture**

16:00

16:10 **Paper Discussion**

17:00 **Lab**

After class …

- Read assigned paper
- Read related information
- Practice

Brief intro to this course

The term "Big Data" has been popular in recent years, but the spirit and critical concept were usually misunderstood so that the investment in a data warehouse or analytic R&D cost could not reflect on revenue to the company. Therefore, we teach the concept, terminology, and technical skills within this course and attempt to stimulate brainstorming through several proposal practices and debates. Moreover, all students are required to leverage the data analytical approaches in the final project presentation.

Before we start …

- Why do you take this course?
  - It's a hot issue.
  - Three credits.
  - I wanna to know more about data science.
- What do you expect to learn from this course?
  - Think about it.
  - Bcuz we have a midterm pitch and final project presentation.
- What kinds of topic are you interested in?
  - Programming, data mining, and story telling …
- Which programming language are you familiar with?
  - C/C++, C#, Python, Matlab, R, Java, JavaScript …

5W1H for big data

- What is the definition of big data?
- Why do we need to leverage big data analysis?
- Where can I obtain the dataset?
- Who will be interested in the analyzed results?
- How can we accurately interpret the outcomes?

What is the definition of big data?

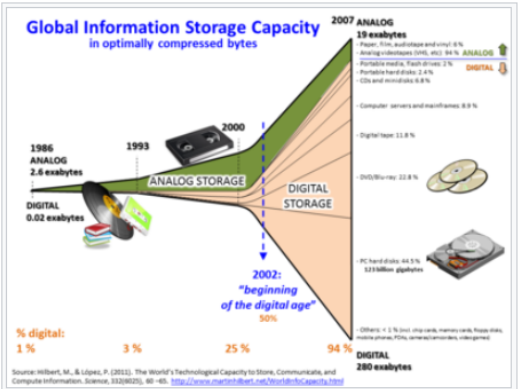
Big data

From Wikipedia, the free encyclopedia

This article is about large collections of data. For the band, see Big Data (band). For the practice of buying and selling of personal and consumer data, see Surveillance capitalism.

Big data refers to data sets that are too large or complex to be dealt with by traditional data-processing application software. Data with many fields (rows) offer greater statistical power, while data with higher complexity (more attributes or columns) may lead to a higher false discovery rate.[2] Big data analysis challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, information privacy, and data source. Big data was originally associated with

Big data refers to data sets that are too large or complex to be dealt with by traditional data-processing application software. Data with many fields (rows) offer greater statistical power, while data with higher complexity (more attributes or columns) may lead to a higher false discovery rate.



Non-linear growth of digital global information-storage capacity and the waning of analog storage[1]

The size and number of available data sets have grown rapidly as data is collected by devices such as mobile devices, cheap and numerous information-sensing Internet of things devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks.[9][10] The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s;[11] as of 2012, every day 2.5 exabytes (2.5×2^{60} bytes) of data are generated.[12] Based on an IDC report prediction, the global data volume was predicted to grow exponentially from 4.4 zettabytes to 44 zettabytes between 2013 and 2020. By 2025, IDC predicts there will be 163 zettabytes of data.[13] According to IDC, global spending on big data and business analytics (BDA) solutions is estimated to reach $215.7 billion in 2021.[14][15] While Statista report, the global big data market is forecasted to grow to $103 billion by 2027.[16] In 2011 McKinsey & Company reported, if US healthcare were to use big data creatively and effectively to drive efficiency and quality, the sector could create more than $300 billion in value every year.[17] In the developed economies of Europe, government administrators could save more than €100 billion ($149 billion) in operational efficiency improvements alone by using big data.[17] And users of services enabled by personal-location data could capture $600 billion in consumer surplus.[17] One question for large enterprises is determining who should own big-data initiatives that affect the entire organization.[18]

Relational database management systems and desktop statistical software packages used to visualize data often have difficulty processing and analyzing big data. The processing and analysis of big data may require "massively parallel software running on tens, hundreds, or even thousands of servers".[19] What qualifies as "big data" varies depending on the capabilities of those analyzing it and their tools. Furthermore, expanding capabilities make big data a moving target.

What is the definition of big data?

Characteristics [edit]

Big data can be described by the following characteristics:

Volume

The quantity of generated and stored data. The size of the data determines the value and potential insight, and whether it can be considered big data or not. The size of big data is usually larger than terabytes and petabytes.[35]

Variety

The type and nature of the data. The earlier technologies like RDBMSs were capable to handle structured data efficiently and effectively. However, the change in type and nature from structured to semi-structured or unstructured challenged the existing tools and technologies. The big data technologies evolved with the prime intention to capture, store, and process the semi-structured and unstructured (variety) data generated with high speed (velocity), and huge in size (volume). Later, these tools and technologies were explored and used for handling structured data also but preferable for storage. Eventually, the processing of structured data was still kept as optional, either using big data or traditional RDBMSs. This helps in analyzing data towards effective usage of the hidden insights exposed from the data collected via social media, log files, sensors, etc. Big data draws from text, images, audio, video; plus it completes missing pieces through data fusion.

Shows the growth of big data's primary characteristics of volume, velocity, and variety

Velocity

The speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development. Big data is often available in real-time. Compared to small data, big data is produced more continually. Two kinds of velocity related to big data are the frequency of generation and the frequency of handling, recording, and publishing.[36]

Veracity

The truthfulness or reliability of the data, which refers to the data quality and the data value.[37] Big data must not only be large in size, but also must be reliable in order to achieve value in the analysis of it. The data quality of captured data can vary greatly, affecting an accurate analysis.[38]

Value

The worth in information that can be achieved by the processing and analysis of large datasets. Value also can be measured by an assessment of the other qualities of big data.[39] Value may also represent the profitability of information that is retrieved from the analysis of big data.

Variability

The characteristic of the changing formats, structure, or sources of big data. Big data can include structured, unstructured, or combinations of structured and unstructured data. Big data analysis may integrate raw data from multiple sources. The processing of raw data may also involve transformations of unstructured data to structured data.

Other possible characteristics of big data are:[40]

Exhaustive

Whether the entire system (i.e., n=all) is captured or recorded or not. Big data may or may not include all the available data from sources.

Fine-grained and uniquely lexical

Respectively, the proportion of specific data of each element per element collected and if the element and its characteristics are properly indexed or identified.

Relational

If the data collected contains common fields that would enable a conjoining, or meta-analysis, of different data sets.

Extensional

If new fields in each element of the data collected can be added or changed easily.

Scalability

If the size of the big data storage system can expand rapidly.

Volume, Variety, Velocity, Veracity, Value, Variability, Exhaustive, Fine-grained and Uniquely Lexical, Relational, Extensional, Scalability

What is the definition of big data?

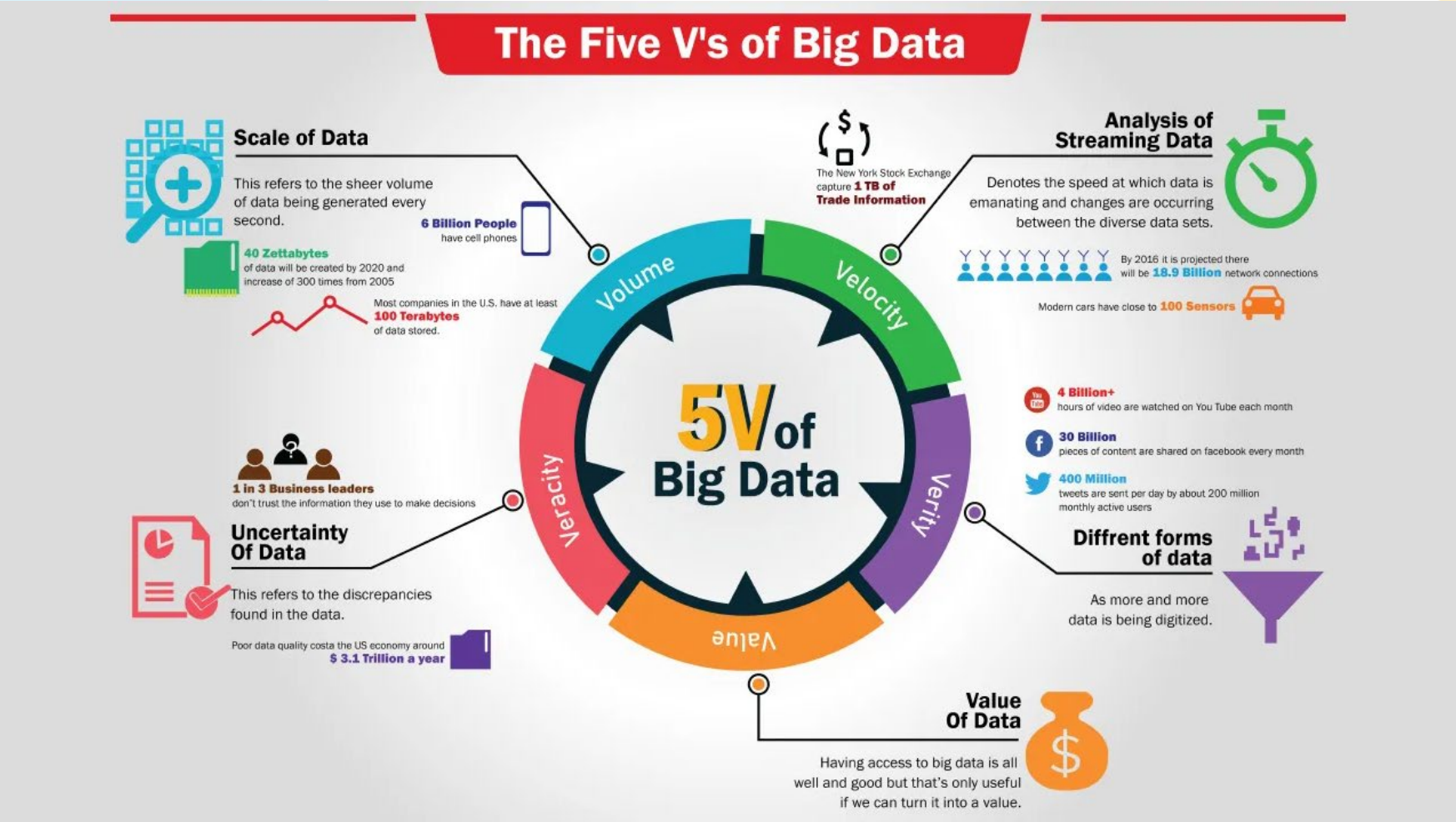


Photo credit: https://morioh.com/p/ca19c6b8c0fe

What is the definition of big data?

- What is big data?
- "Big data" is the massive amount of data available to organizations that—because of its volume and complexity—is not easily managed or analyzed by many business intelligence tools.
- Big data consists of petabytes (more than 1 million gigabytes) and exabytes (more than 1 billion gigabytes), as opposed to the gigabytes common for personal devices.

Info source: https://cloud.google.com/learn/what-is-big-data

What is the definition of big data?

Volume

The key characteristic of big data is its scale—the volume of data that is available for collection by your enterprise from a variety of devices and sources.

Variety

Variety refers to the formats that data comes in, such as email messages, audio files, videos, sensor data, and more. Classifications of big data variety include structured, semi-structured, and unstructured data.

Velocity

Big data velocity refers to the speed at which large datasets are acquired, processed, and accessed.

Variability

Big data variability means the meaning of the data constantly changes. Therefore, before big data can be analyzed, the context and meaning of the datasets must be properly understood.

Why do we need to leverage big data analysis?

➔ Better Decision Making

Companies use big data in different ways to improve their B2B operations, advertising, and communication. Many businesses including travel, real estate, finance, and insurance are mainly using big data to improve their decision-making capabilities. Since big data reveals more information in a usable format, businesses can utilize that data to make accurate decisions on what consumers want or not and their behavioral tendencies.

Info source: https://towardsai.net/p/l/advantages-disadvantages-of-big-data

Why do we need to leverage big data analysis?

➔ Reduce costs of business processes

The surveys conducted by New Vantage and Syncsort (now Precisely) reveals that big data analytics has helped businesses to reduce their expenses significantly. 66.7% of survey respondents from New Vantage claimed that they have started using big data to reduce expenses. Furthermore, 59.4% of survey respondents from Syncsort claimed that big data tools helped them reduce costs and increase operational efficiency.

Info source: https://towardsai.net/p/l/advantages-disadvantages-of-big-data

Why do we need to leverage big data analysis?

➔ Fraud Detection

Financial companies, in particular, use big data to detect fraud. Data analysts use machine learning algorithms and artificial intelligence to detect anomalies and transaction patterns. These anomalies of transaction patterns indicate something is out of order or a mismatch giving us clues about possible frauds.

Fraud detection is significantly important for credit unions, banks, credit card companies to identify account information, materials, or product access. Any industry, including finance, can better serve its customers by early identification of frauds before something goes wrong.

Info source: https://towardsai.net/p/l/advantages-disadvantages-of-big-data

Why do we need to leverage big data analysis?

➔ Increased productivity

According to a survey from Syncsort, 59.9% of survey respondents have claimed that they were using big data analytics tools like Spark and Hadoop to increase productivity. This increase in productivity has, in turn, helped them to improve customer retention and boost sales.

Modern big data tools help data scientists and analysts to analyze a large amount of data efficiently, enabling them to have a quick overview of more information. This also increases their productivity levels.

Besides, big data analytics helps data scientists and data analysts gain more information about themselves so that they can identify how to be more productive in their activities and job responsibilities.

Info source: https://towardsai.net/p/l/advantages-disadvantages-of-big-data

Why do we need to leverage big data analysis?

➔ Improved customer service

Improving customer interactions is crucial for any business as a part of their marketing efforts.

Since big data analytics provide businesses with more information, they can utilize that data to create more targeted marketing campaigns and special, highly personalized offers to each individual client.

The major sources of big data are social media, email transactions, customers' CRM (customer relationship management) systems, etc. So, it exposes a wealth of information to businesses about their customers' pain points, touchpoints, values, and trends to serve their customers better.

Moreover, big data helps companies understand how their customers think and feel and thereby offer them more personalized products and services. Offering a personalized experience can improve customer satisfaction, enhance relationships, and, most of all, build loyalty.

Info source: https://towardsai.net/p/l/advantages-disadvantages-of-big-data

Why do we need to leverage big data analysis?

➔ Increased agility

Another competitive advantage of big data is increasing business agility. Big data analytics can help companies to become more disruptive and agile in markets. Analyzing huge data sets related to customers enables companies to gain insights ahead of their competitors and address the pain points of customers more efficiently and effectively.

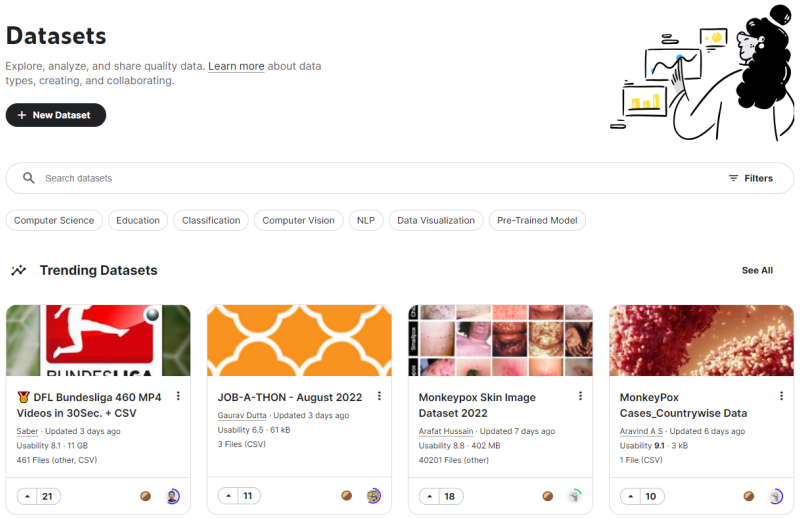
On top of that, having huge data sets at disposal allows companies to improve communications, products, and services and reevaluate risks. Besides, big data helps companies improve their business tactics and strategies, which are very helpful in aligning their business efforts to support frequent and faster changes in the industry.

Info source: https://towardsai.net/p/l/advantages-disadvantages-of-big-data

Where can I obtain the dataset?

- Open source (e.g., data.gov.tw)
- Crowd-sourcing (e.g., mobile app)
- Competition Platform (e.g., Kaggle)

Web crawler



https://www.kaggle.com/datasets

https://data.gov.tw/

Who will be interested in the analyzed results?

- Who is your stakeholders?
- How do you define your problem?
- What is your expected outcome?
- Where is your available resource?
- When is the deadline of this project?

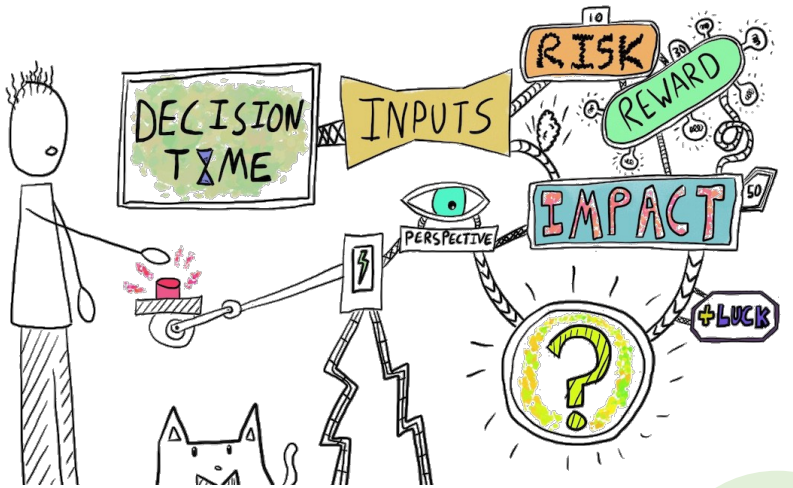
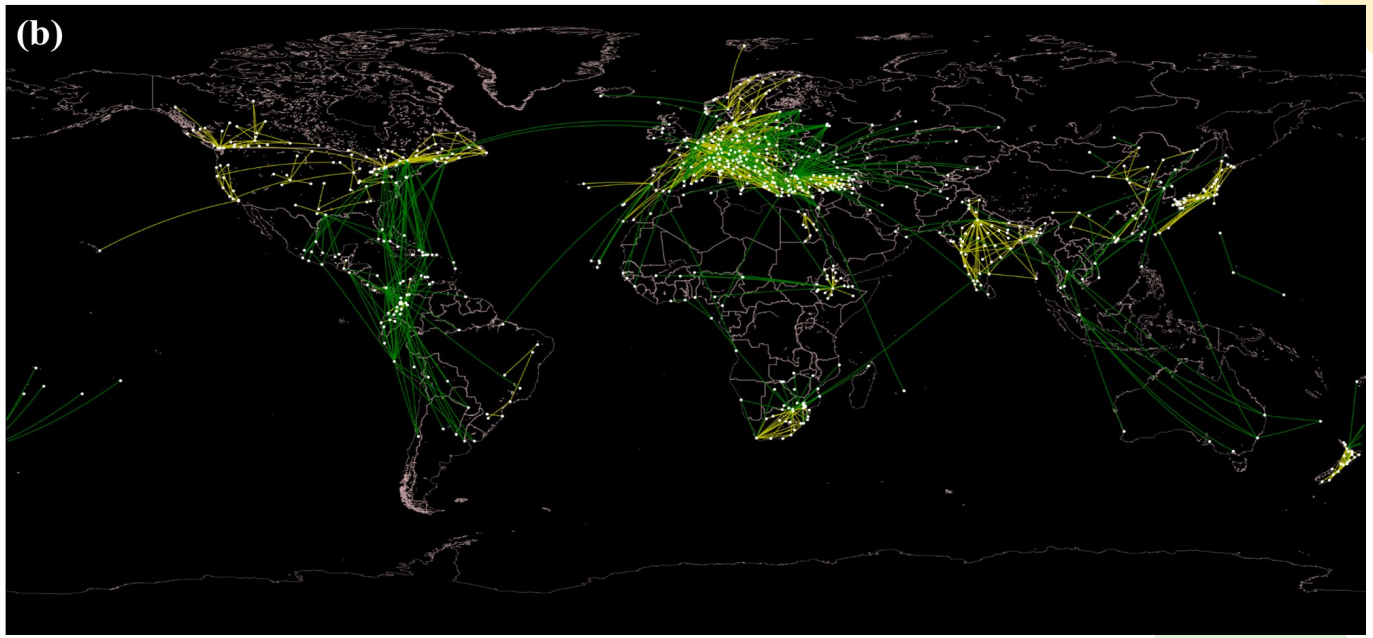


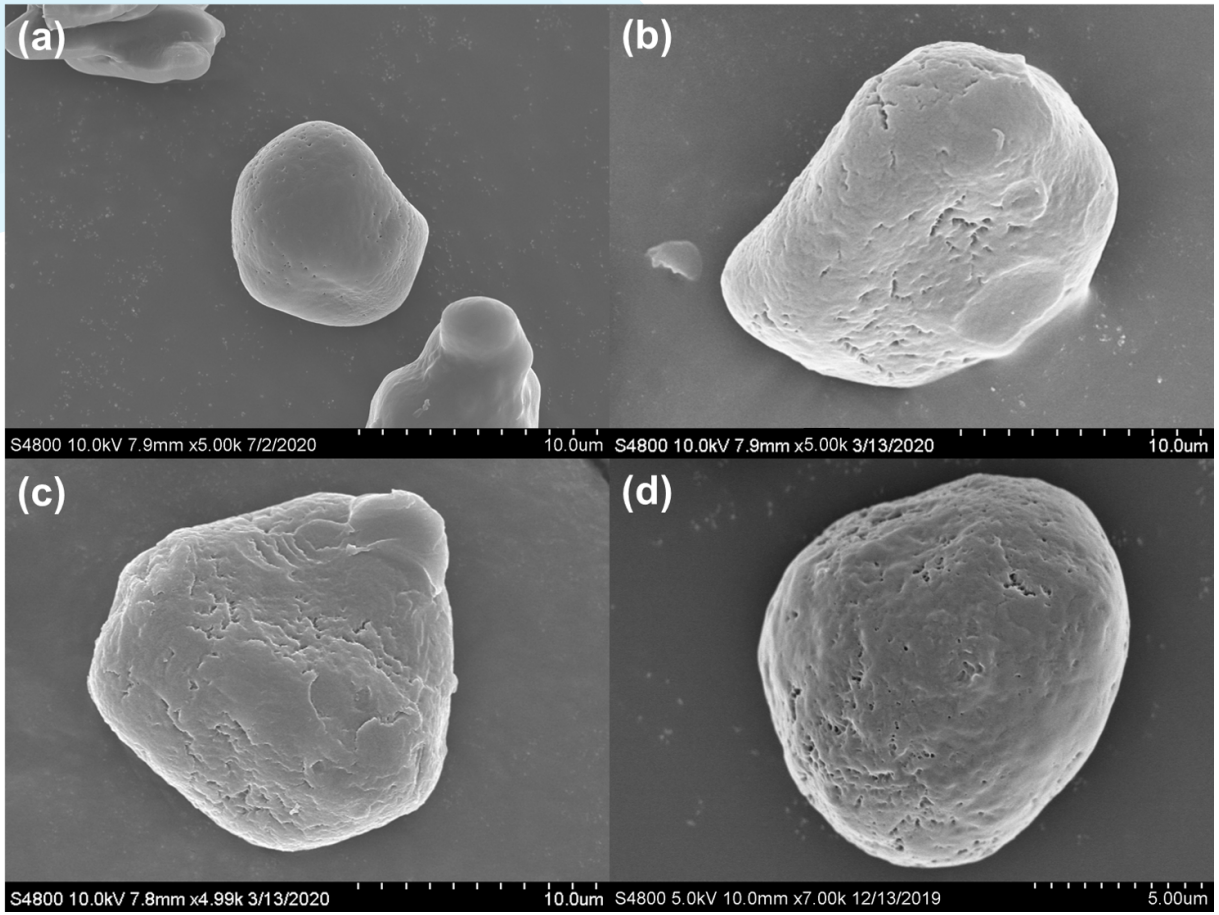
Photo credit: https://uxdesign.cc/decision-making-for-product-managers-7fef3292cb65

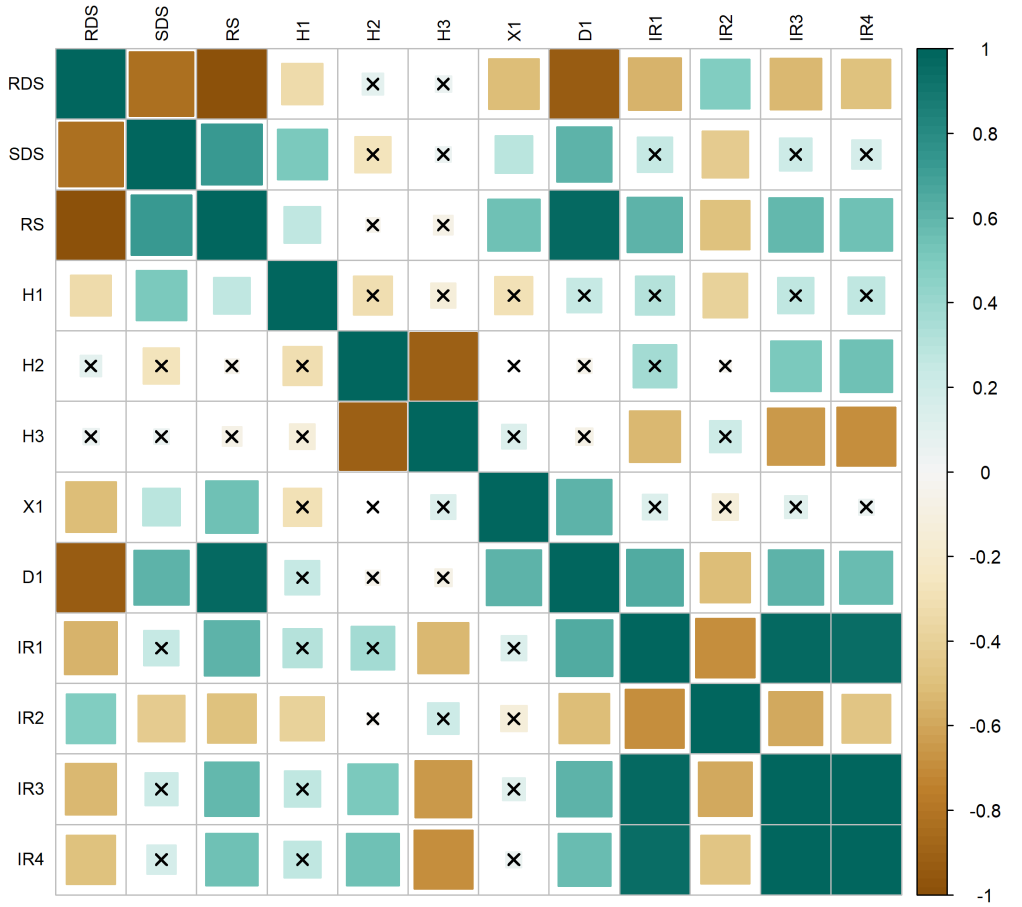
How can we accurately interpret the outcomes?

- Be aware of methodology you adopted for this result
- Leverage an appropriate visualization tool
- Add a legend
- Add a caption



How can we accurately interpret the outcomes?





Potential Issues – Security risks

Most of the time, companies collect sensitive information for big data analytics. Those data need protection, and security risks can be demerits due to the lack of proper maintenance.

Besides, having access to huge data sets can gain unwanted attention from hackers, and your business may be a target of a potential cyber-attack. As you know, data breaches have become the biggest threat to many companies today.

Another risk with big data is that unless you take all necessary precautions, important information can be leaked to competitors.

Potential Issues – Compliance

The need to have compliance with government legislation is also a drawback of big data. If big data contains personal or confidential information, the company should make sure that they follow government requirements and industry standards to store, handle, maintain, and process that data.

Potential Issues – Lack of talent

According to a survey by AtScale, the lack of big data experts and data scientists has been the biggest challenge in this field for the past three years. Currently, many IT professionals don't know how to carry out big data analytics as it requires a different skill set. Thus, finding data scientists who are also experts in big data can be challenging.

Big data experts and data scientists are two highly paid careers in the data science field. Therefore, hiring big data analysts can be very expensive for companies, especially for startups. Some companies have to wait for a long time to hire the required staff to continue their big data analytics tasks.

[#1] Assignment

Q1: Why do you take this course?

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur at dolor maximus, varius massa eu, porttitor sapien. Vivamus facilisis

Q2: What do you want to learn in this course?

Pellentesque tristique vitae metus sed sollicitudin. Aenean massa nisl, sodales sed dolor in, placerat maximus elit.

Q3: Do you have any problems or concerns of this course?

Praesent varius tortor vitae tincidunt porttitor. Duis et dui eu purus imperdiet varius. Nam posuere euismod erat at pharetra. Vestibulum in nunc ante.

Question Time

If you have any questions, please do not hesitate to ask me.

The End

Thank you for your attention))