

EE 1004 Foundation of Information Systems and Data Analysis

Part II

Chapter 6: What are Data

What are data?

- **Data are** raw facts, information, figures, numbers, symbols, words, texts, etc.
- Grouping (manipulating) of Data can function to represent for different objectives, meanings, outputs, input controls and
- Data are gathered, captured or collected from places, things, people, and any activities.

What are data?

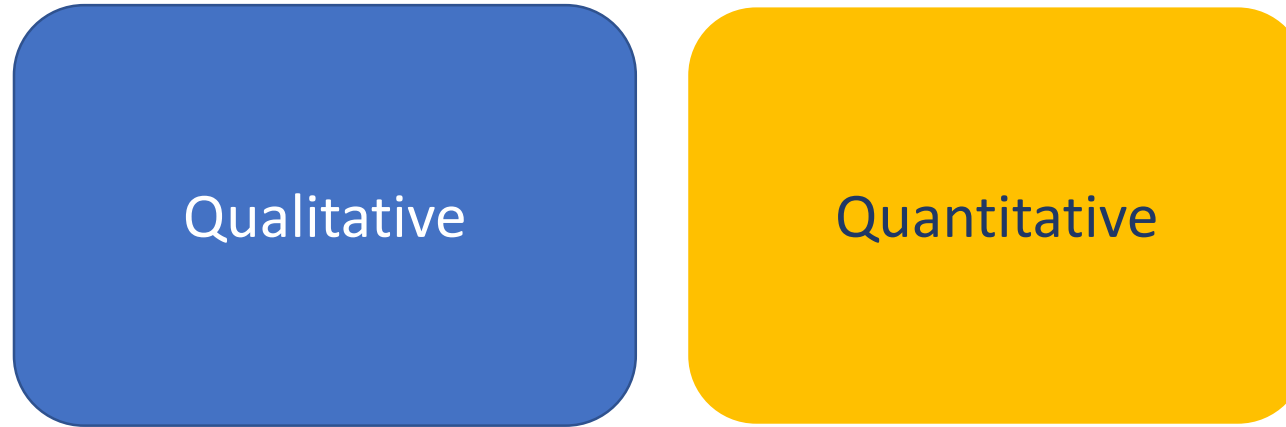
Example 1: What is the value of data 11000 ??

- (a) Eleven Thousand
- (b) 11000 Yen (594.08 Hong Kong Dollar)
- (c) $(11000)_2 = (24)_{10}$

Example 2: Fill your personal information in a public form (eg. Membership of Shops, supermarkets)

- Name: CHAN Tai Man, LEE Ka Yan, Steve Jobs
- Birthday: dd-mm-yy
- Mobile number: 9988-7788
- Education level: Undergraduate, Master, PhD
- Salary range: 20-50 k, 50-80 k, above 80 k

Major Data Types



How to differentiate qualitative and quantitative data?

- The qualitative data can be directly give an impression to the reader, public, and targeted audience
- The qualitative data are powerful to present the key points, messages, theme from the providers
- The quantitative data are used for supporting the objectives, goals, analyses, and methodologies to the reader, public, and targeted audience
- The quantitative data can be manipulated by the providers to present different points, messages, theme, objectives of the data sets

Example



Qualitative

Quantitative

Source from MarketWatch 8 Oct 2022

What is the qualitative / quantitative data in this figure?

Example

Qualitative

- China related
- Chip stocks
- Stock drops
- Technologies

Quantitative

- Two-year low
- Examples of companies
- AMD and NVDS drops
- 6% dropped in one day

MarketWatch

Chip stocks crushed to two-year low as more tech, AI ban to China add to woes



Chip stocks crushed to two-year low as more tech, AI ban to China add to woes

Wallace Witkowski

Sat, October 8, 2022 at 5:28 AM

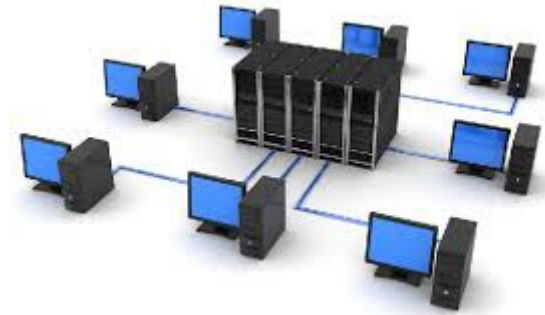
In this article:

AMD
-13.87% ☆

NVDA
-8.03% ☆

The chip sector melted down Friday for its third 6% one-day drop of the year after U.S. regulators moved to pump the brakes on China's military ambitions as it issued wider restrictions on semiconductor and AI technology that can be sold to the world's second-largest economy.

Example



What is it?

Qualitative

Quantitative

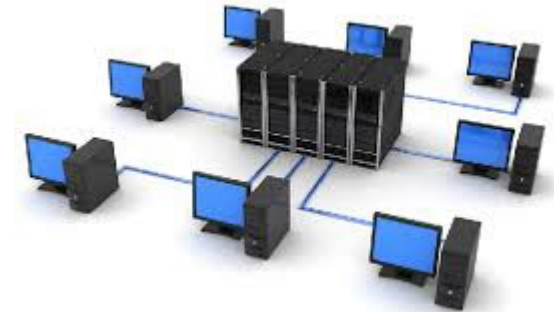
Example

Qualitative

- Computer Network
- Color of desktop : black
- Network protocol: TCP
- Structure : Star

Quantitative

- Structural:
 - Number of desktop : 8
 - Servers : 1
- Specifications:
 - Power consumption: 20.28 KW
 - Total Cable length: 200.25 m



Qualitative and Quantitative Data

Qualitative Data	Quantitative Data
• Main themes should come from the first impression • Deals with descriptions • Data can be observed • Colors, texture type, smells, beauty, etc <u>Descriptive</u>	• Result of counting or measuring attributes of a population. • Deals with numbers • Data which can be measured, grouped and manipulated. • Size, values, numbers, lengths, heights, areas, power consumptions, etc <u>Countable</u>

Major Engineering Data Types



Structured

Typical machine-generated unstructured data includes:

Satellite imagery: Weather data, land forms, military movements.

Scientific data: Oil and gas exploration, space exploration, seismic imagery, atmospheric data.

Sensor data: Traffic, weather, oceanographic sensors.

Unstructured

Typical human-generated unstructured data includes:

Text files: Word processing, spreadsheets, presentations, email, logs.

Emails: Email has some internal structure thanks to its metadata, and we sometimes refer to it as semi-structured. However, its message field is unstructured and traditional analytics tools cannot parse it.

Social Media files: Data from Facebook, Twitter, LinkedIn.

Websites: YouTube, Instagram, photo sharing sites.

Example



Structured

Unstructured

- Look at your student iD card, can you define where is/are the structured or unstructured data?
- In this example, how to convert the unstructured data to structured data for engineering purposes?

Example

The image displays two side-by-side web forms. The left form, titled 'Create your Account', includes fields for 'Your name' (First and Last), 'Email*', 'Password*', and 'Phone number (for delivery alerts via WhatsApp)'. It also features two checked checkboxes for newsletter and WhatsApp updates, and a 'REGISTER NOW' button. The right form, titled 'Contact Us', includes fields for 'Your name' (First and Last), 'Email*', and 'Phone number'. It has a section for 'Please contact me via:' with checked options for 'SMS' and 'WhatsApp', and a 'Message*' text area. Both forms have a corresponding action button at the bottom: 'REGISTER NOW' and 'SUBMIT'.

Create your Account

Your name

First Last

Email* Password*

Phone number (for delivery alerts via WhatsApp)

+49

☒ Yes, I would like to receive a newsletter.

☒ Yes, I would like to receive updates via WhatsApp

REGISTER NOW

Contact Us

Your name

First Last

Email* Phone number

+49

Please contact me via:

☐ Phone ☒ SMS

☐ Email ☒ WhatsApp

Message*

SUBMIT

- Where are the structured data?
- Why the form create unstructured data to fill in?

Example

Back to this:

Example 2: Fill your personal information in a public form (eg. Membership of Shops, supermarkets)

Name: CHAN Tai Man, LEE Ka Yan, Steve Jobs

Birthday: dd-mm-yy

Mobile number: 9988-7788

Education level: Undergraduate, Master, PhD

Salary range: 20-50 k, 50-80 k, above 80 k

- From the above example, *name* and *education level* are not belonged to numbers. If so, how does the machine collect these data (by which format) ? How does the machine analyze these data (by what methodology)?

Big Data Characterization

Volume

Velocity

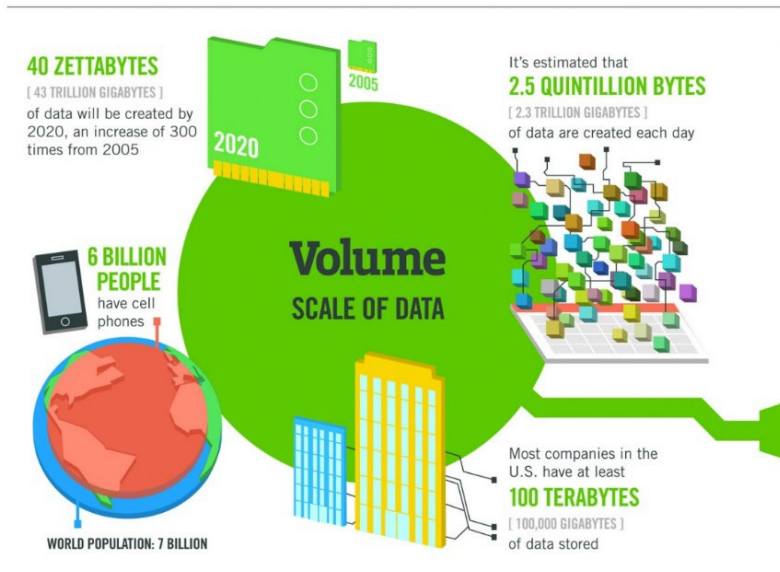
Variety

Veracity

Big Data Characterization

Volume

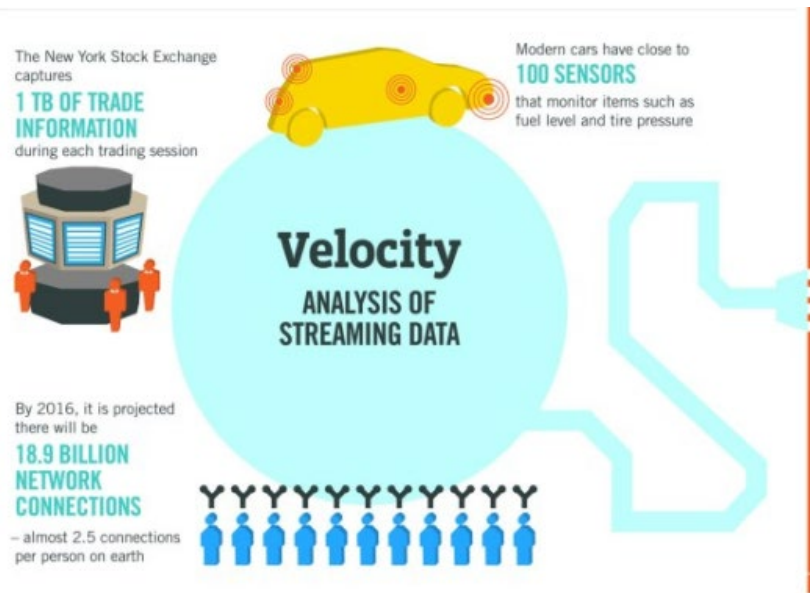
- The amount and scale of data are very large compared to traditional data sources
- Usually greater than terabytes and petabytes
- Facebook posts, and Youtube vidoes, DNA records
- The data sets in Big Data may be too large to process with a regular laptop or desktop processor
- Big Data can be divided into various groups of sub data sets to be computed (cloud computing, edged computing)



Big Data Characterization

Velocity

- Data are generated extremely fast, the process that never stops
- Dynamic data collect, update and store continuously
- Single instance of high-velocity data is Twitter where over 350,000 tweets are now sent worldwide per minute, equating to 500 million tweets per day



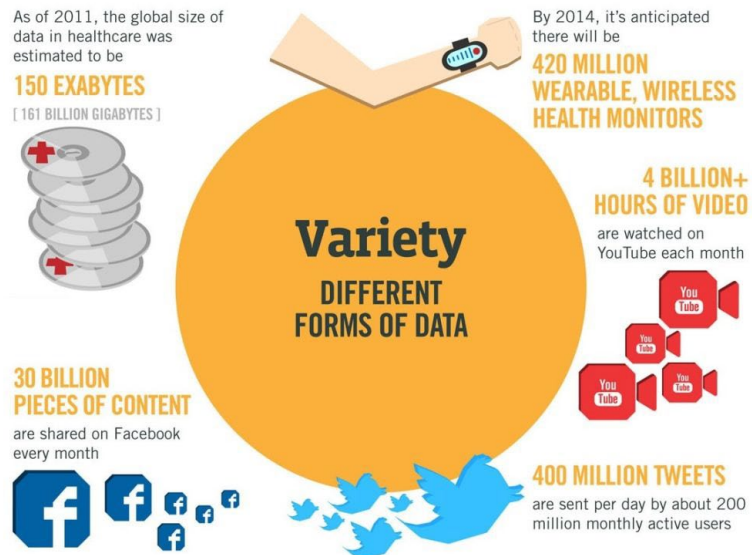
What is difference between volume and velocity?

What is the situation that we have high volume but low velocity ?

Big Data Characterization

Variety

- The data comes from different sources in different structures and forms
- Data are created not only by human but also by machines
- Three types: structured, semi-structured and unstructured data
- The variety in data types implies that we need particular processing capabilities and specialist algorithms



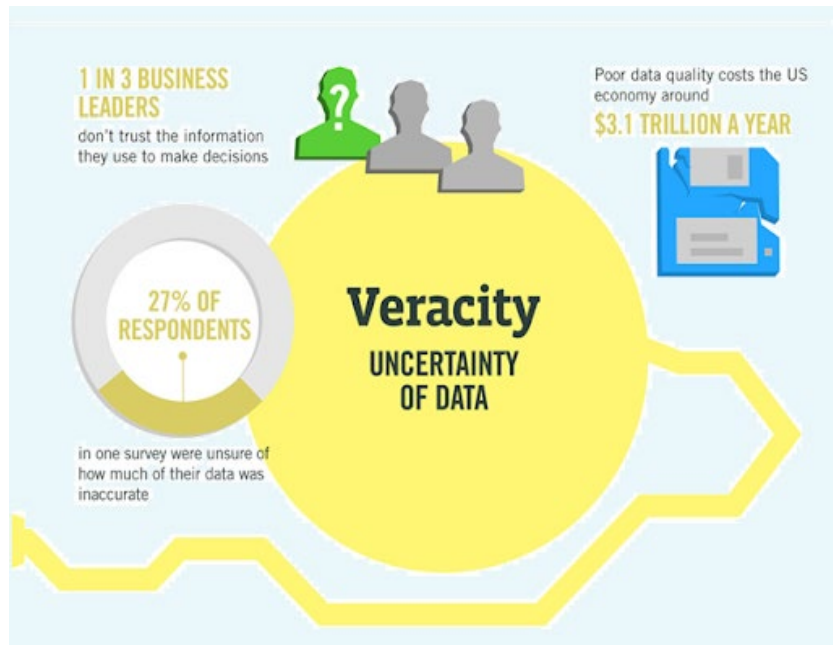
Social network: Facebook images and videos

Modern means of communications: phone calls, messages, pictures, videos

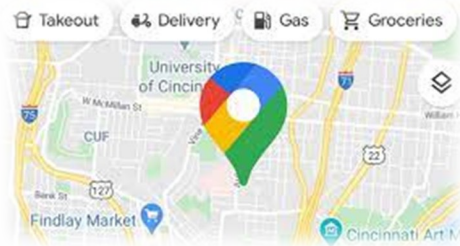
Online shopping: pictures, model videos, on-line streaming

Big Data Characterization

Veracity



- Big Data come from many different ways, places, channels, peoples, and medias
- We need to test **the quality of the data**.
- Data veracity, in general, is how accurate or truthful a data set may be
- It's not just the quality of the data itself but how trustworthy the data source, type, and processing of it is
- Stock markets are very sensitive to high-quality veracity data
- Engineering systems (large-transportation system: MTR and airplane networks, and large-infrastructure systems: power grid and mobile networks) are highly-reliable to the trustable veracity data to provide stable and high-efficiency services



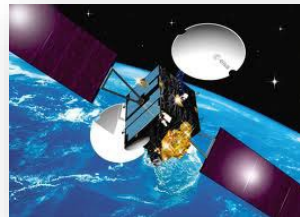
Navigations



Mobile services



Transportation controls



Satellite services



Social networks

