LECTURER: TAI LE QUY

INTRODUCTION TO DATA SCIENCE

TOPIC OUTLINE

Introduction to Data Science	1
Data	2
Data Science in Business	3
Statistics	4
Machine Learning	5
Summary session	6

UNIT 2

DATA

STUDY GOALS

On completion of this unit, you will have learned …

- what is meant by data and information.
- the different types and shapes of data.
- the typical sources of data.
- the 5Vs of big data.
- the issues concerning data quality.
- the challenges associated with the data engineering process.

- What do “data” and “information” mean?
- What data types and data shapes are there?
- Where can data be collected?
- What are the important criteria for data quality?
- What are the issues concerning data quality? How to fix them?
- What are the challenges and benefits of data processing?

DEFINITION

DATA

Facts, observations, assumptions, or incidences

Data describe quantity, quality, statistics, symbols, or other units of meaning

Data are processed to return information

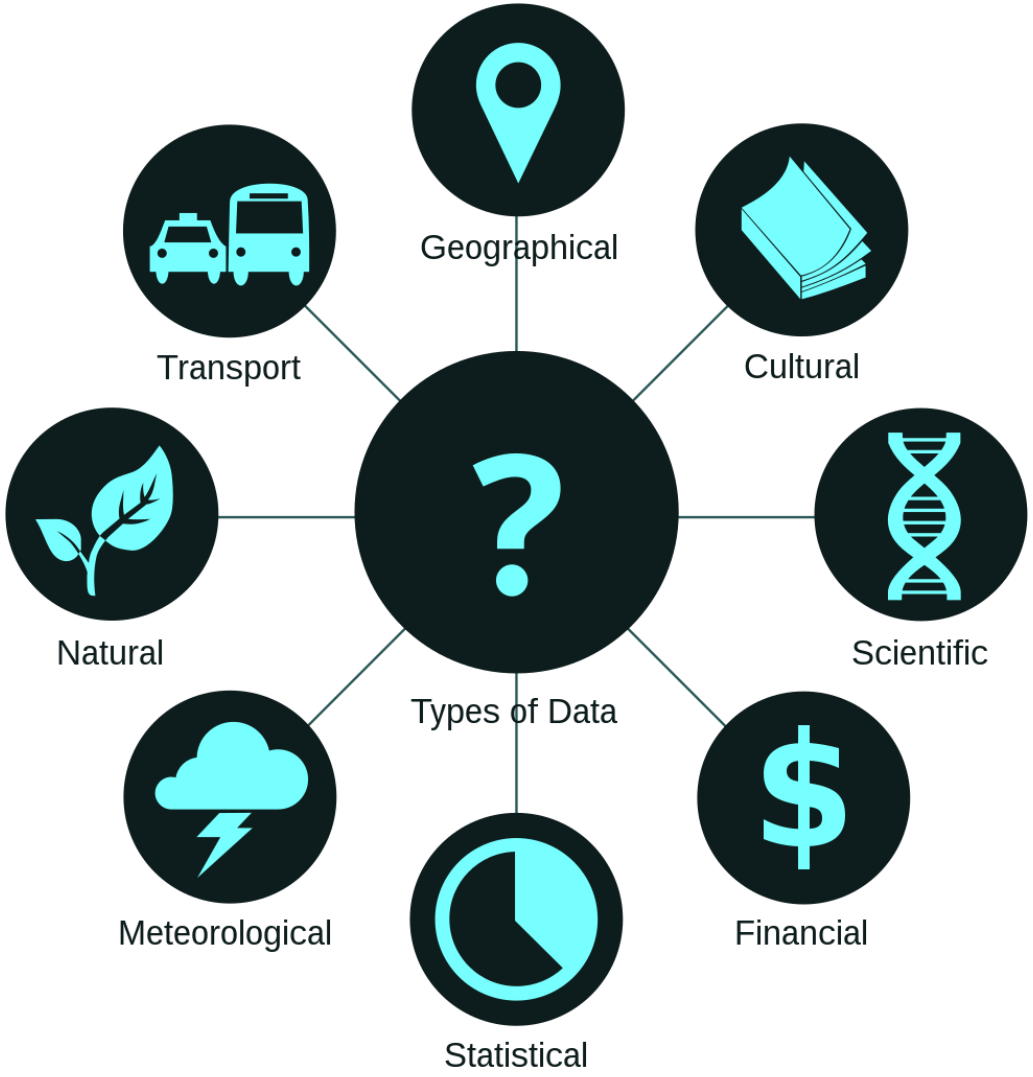
Source of the text: Zöller, 2022

DEFINITION

Data are collections of discrete values that convey information

Source of text: Wikipedia, 2023

TYPES OF DATA



Qualitative Data

- Describe qualities or characteristics
- Cannot be counted
- Words, objects, pictures, observations, and symbols
- Answer to questions: What characteristics? What property?
- Identify conceptual framework in an area of study

Quantitative Data

- Can be quantified or expressed as a number
- Can be counted
- Numbers and statistics
- Answer to questions: How much? How often?
- Test hypotheses, analyses the connection between cause and effect.

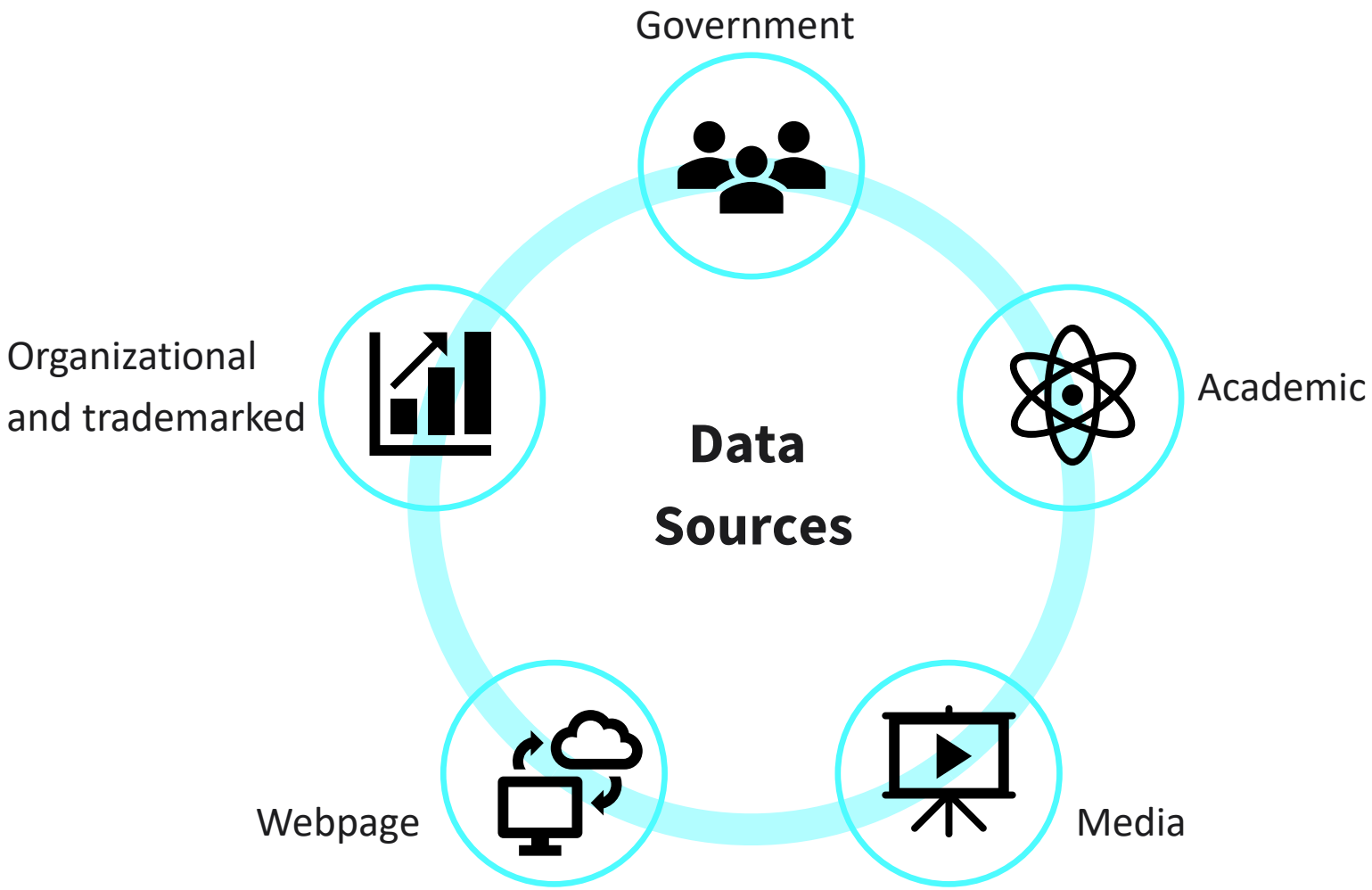
Source of the text: Zöller, 2022

SHAPES OF DATA

	Structured Data	Unstructured Data	Streaming Data
Characteristics	• predefined data models • usually, text or numerical • easy to search	• no predefined data models • can be text, images, audio, or other formats • difficult to search	• continuously generated • by sensors • large amount • processed incrementally
Applications	• inventory control • airline reservation systems	• word processing • tools for editing media	• state monitoring • process control
Examples	• phone numbers • customer names • transaction information	• reports • imagery • email/messages	• real-time sensor values • stock updates

Source of the text: Zöller, 2022

SOURCES OF DATA



Source of the text: Zöller, 2022

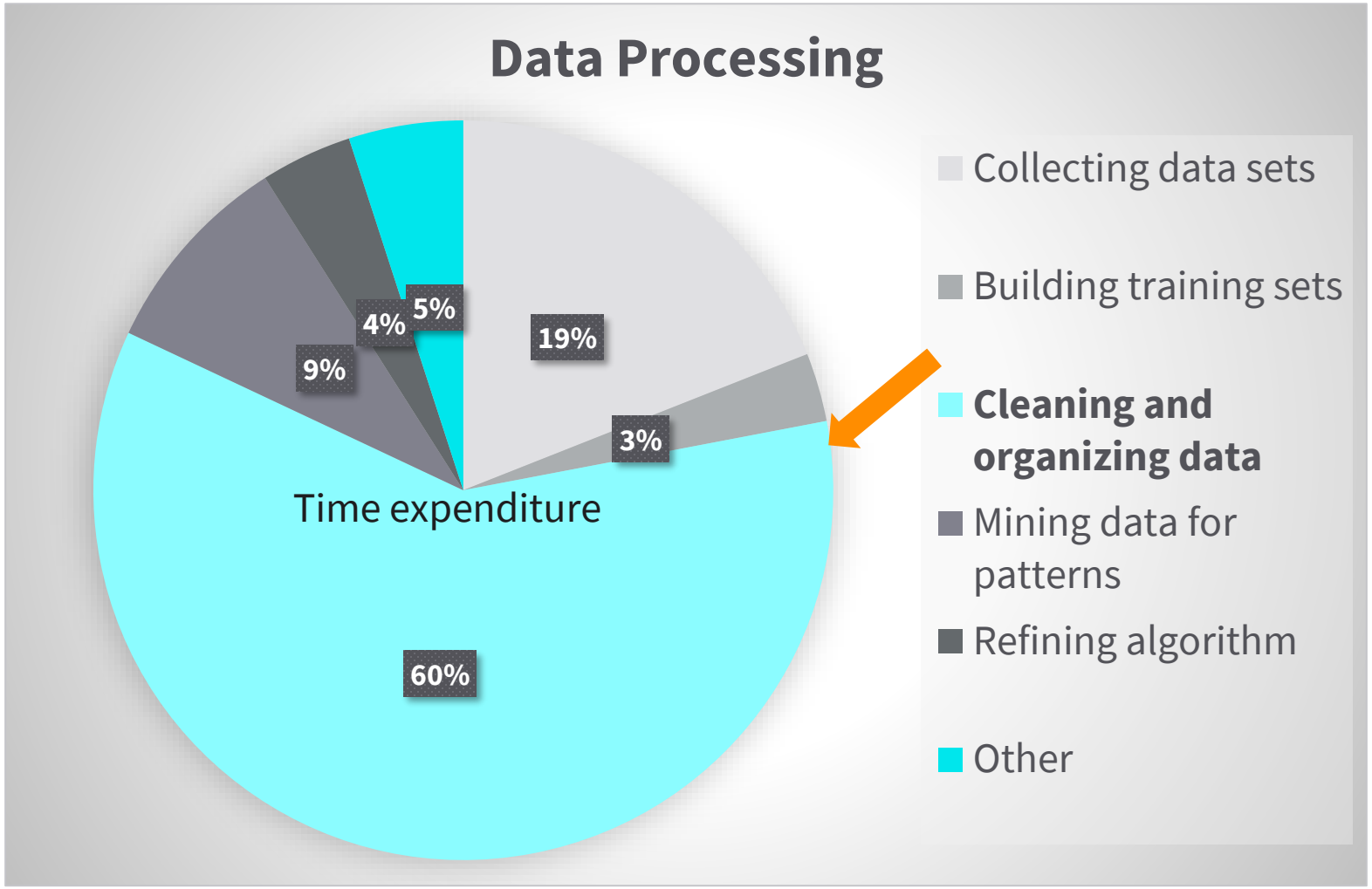
THE 5VS OF BIG DATA

1. **Volume**: amount, scale
2. **Variety**: structured/unstructured
3. **Velocity**: frequency
4. **Veracity**: quality
5. **Validity**: suitability

Source of the text: Zöller, 2022

Issues	Causes	Solutions
Missing values and outliers	- not observed - incorrectly observed	- removal - replacement
Duplicate records	- duplicate observation	- removal
Redundancy	- collected more than needed	- correlation analysis - removal

Source of the text: Zöller, 2022



Source of the text and image: Zöller, 2022, p.24

Benefits of data processing

- Improvement of **analysis** & **demonstration** of data
- **Reduction** of meaningless data
- Easier **storage** and **distribution** of data
- Simplified **report** creation
- Enhanced **productivity** and increased **profits**
- Accurate **decision**-making

Source of the text: Zöller, 2022

Data Transformation Methods

Variable scaling

- Scaled values: {0,1} or {-1,1}
- To ensure equal weighting

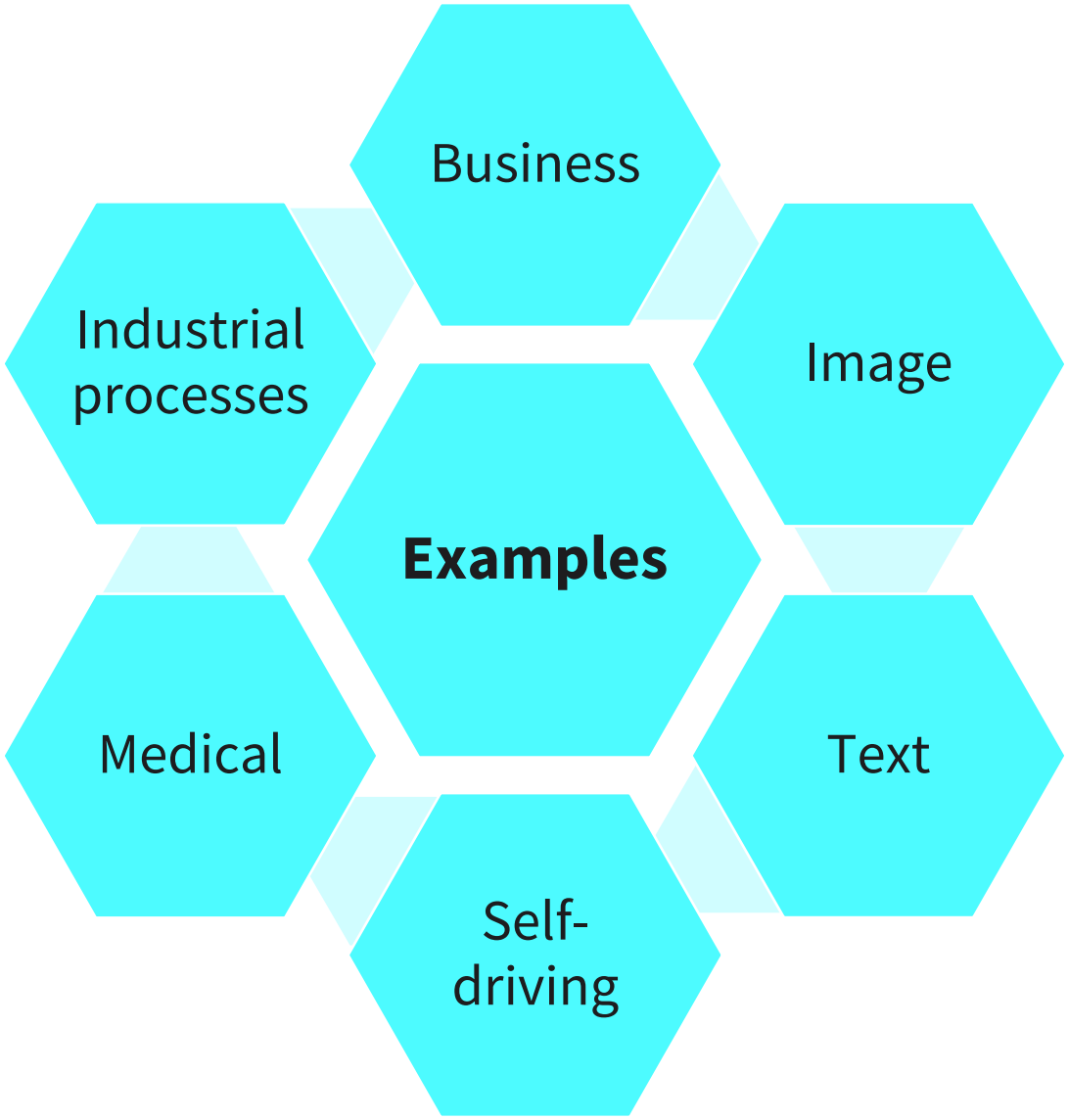
Variable decomposition

- One variable → multiple variables
- To improve data representation

Variable aggregation

- Some variables → one variable
- To reduce data volume

Source of the text: Zöller, 2022



Source of the text: Zöller, 2022

REVIEW STUDY GOALS

- what is meant by data and information
- the different types and shapes of data
- the typical sources of data
- the 5Vs of big data
- the issues concerning data quality
- the challenges associated with the data engineering process

SESSION 2

TRANSFER TASK

TRANSFER TASK
PRESENTATION OF THE RESULTS

Please present your results.

The results will be discussed in plenary.

Scenario

You are going to apply **data science** at an online shop to analyze and optimize the customer experience.

Questions

1. What data should be collected?
2. Which data from your dataset are quantitative, and which are qualitative?
3. Which types of data are there in your dataset?
4. Where can you collect your data?
5. How to guarantee the quality of your data? Give an example of your data quality issues and your solution.
6. How do you schedule your time for the task of data processing?
7. Which data transformation methods do you need for data engineering in your project?
8. Which benefits can you get from the data processing results?

1. Which shape of data has these characteristics: Easy to search, predefined data models, and numerical data?
    a) structured data
    b) unstructured data
    c) streaming
    d) all of the above

2. Which of the following is incorrect regarding data transformation methods?
   a) variable scaling
   b) variable decomposition
   c) variable aggregation
   d) none of the above

3. Which activity consumes most of the data processing time?
   a) collecting data
   b) cleaning and organizing data
   c) building training sets
   d) mining data for patterns

LIST OF SOURCES

João Batista Neto. (2015). File: Data_types_-_en.svg. *Wikimedia Commons*. https://en.wikipedia.org/wiki/Data#Meaning

Wikipedia. (2023, May 14). Data. *Wikipedia Commons.* https://en.wikipedia.org/wiki/Data#Meaning

Zöller, T. (2022). *Introduction to Data Science*. IU International University of Applied Science.