LECTURER: TAI LE QUY

DATA SCIENCE

INTRODUCTORY ROUND

Who am I?

- Name: Tai Le Quy
- PhD at L3S Research Center – Leibniz University Hannover
  - Topic: Fairness-aware machine learning in educational data mining
- MSc in Information Technology at National University of Vietnam
- Profile: tailequy.github.io
- Email: tai.le-quy@iu.org
- Materials: https://github.com/tailequy/IU-DataScience

INTRODUCTORY ROUND

Who are you?

- Name
- Employer
- Position/responsibilities
- Fun Fact
- Previous knowledge? Expectations?

TOPIC OUTLINE

Introduction to Data Science	1
Use Cases and Performance Evaluation	2
Data Preprocessing	3
Processing of Data	4
Selected Mathematical Techniques	5
Selected Artificial Intelligence Techniques	6

UNIT 1

INTRODUCTION TO DATA SCIENCE

STUDY GOALS

On completion of this unit, you will have learned…

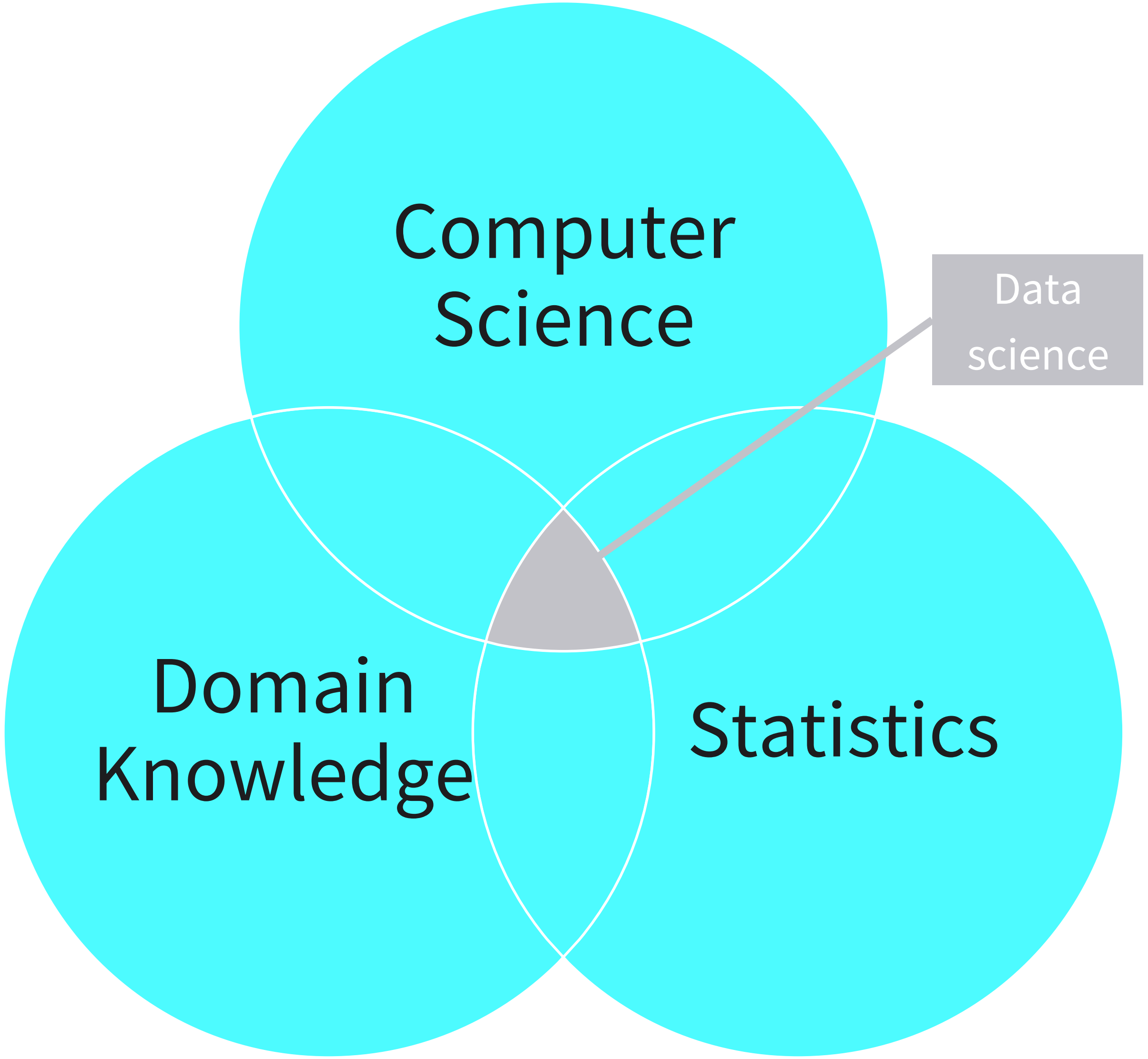
- the meaning of data science.
- common terms and definitions in data science.
- the different applications of data science.
- the typical sources of data.
- the types and shapes of data.
- probability distributions and Bayesian statistics.

EXPLAIN SIMPLY

1. Define the term data science in your own words.
2. Explain the difference between structured, unstructured and semi-structured data.
3. Identify two types of machine learning and give an application example for each type.

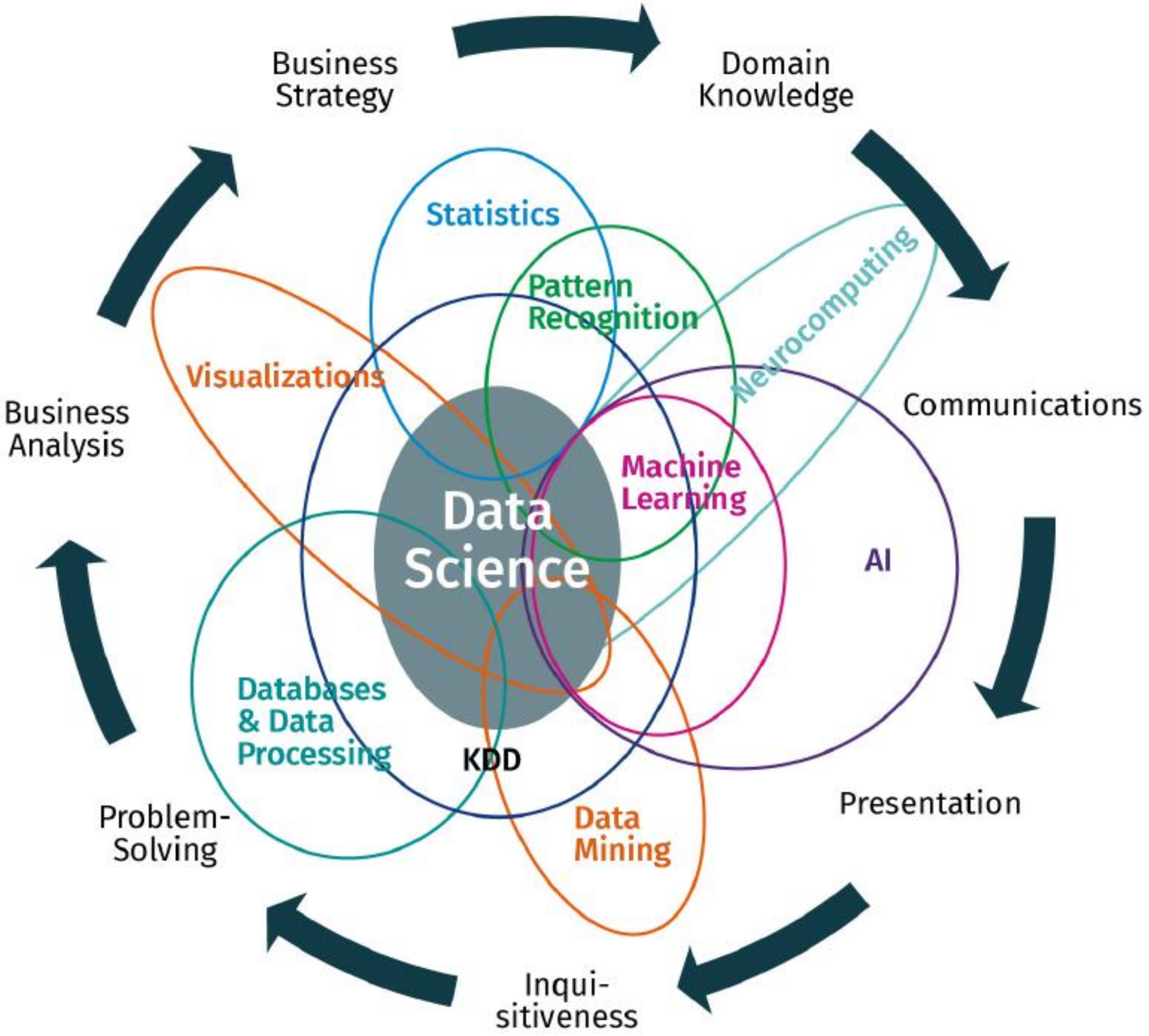
Data science

- analyze and explore the information contained in data
- incorporate domain knowledge
- create predictions to advise the decision-making process
- create value from data



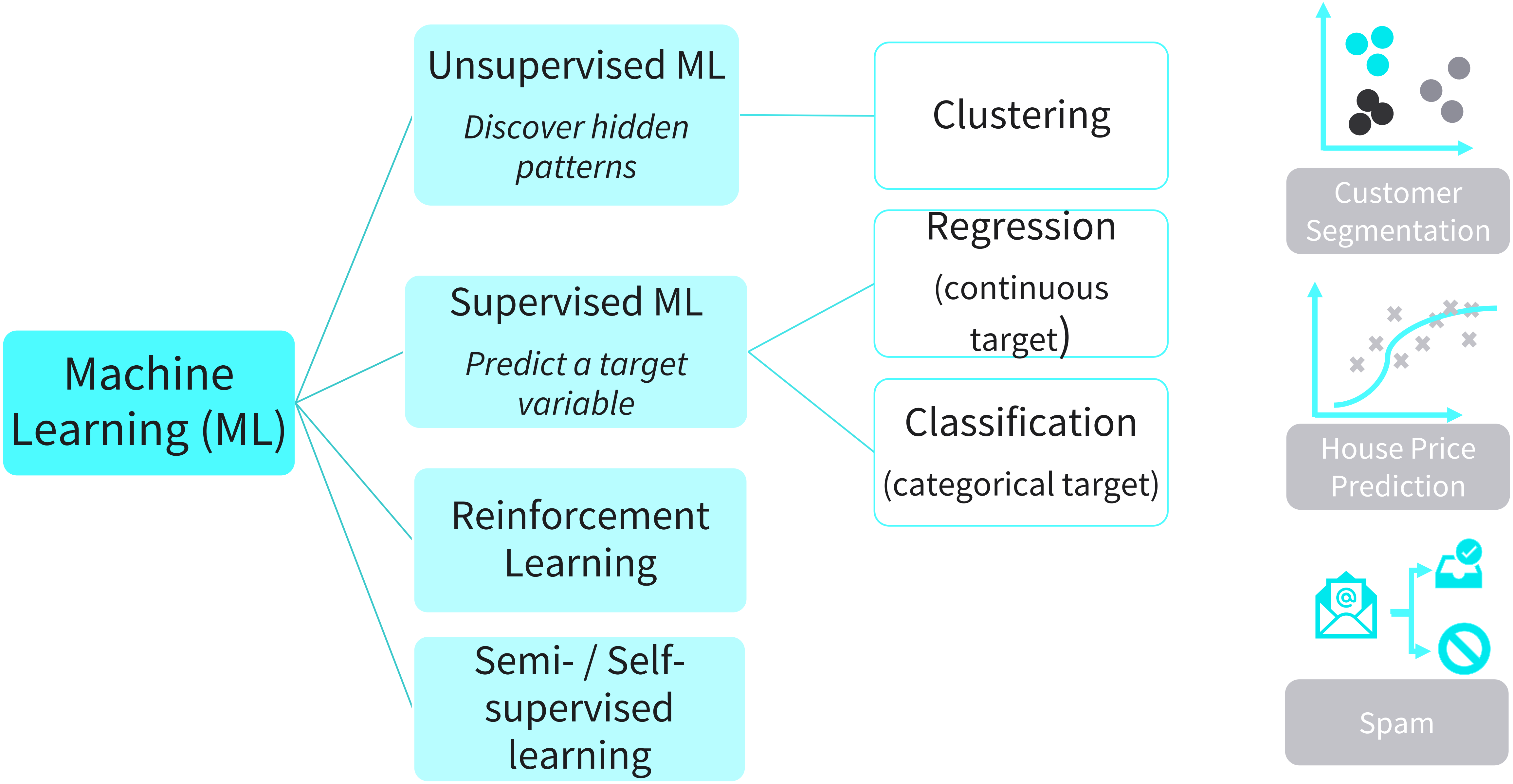
Source of the graphics: Drawing by Antonia Schulze, 2021.

DATA SCIENCE VENN DIAGRAM



Source: Tierney (2012)

TYPES OF MACHINE LEARNING



Machine Learning (ML)

Unsupervised ML

Discover hidden patterns

Supervised ML

Predict a target variable

Reinforcement Learning

Semi- / Self-supervised learning

Clustering

Regression (continuous target)

Classification (categorical target)

Customer Segmentation

House Price Prediction

Spam

Source of the graphics: Drawing by Antonia Schulze, 2021.

DATA SCIENCE TERMS

Data Handling

Training Set	▪ The **dataset** used to learn the desired task.
Testing Set	▪ Assesses the **performance** of machine learning model.
Outlier	▪ A **data record**
Data Cleansing	▪ The **process** of removing redundant data, etc.

Data Features

Feature	▪ **Measure** of the data; height, etc.
Dimensionality Reduction	▪ The process of **reducing the dataset.**
Feature Selection	▪ The process of **selecting relevant features.**

Source of the image: Zöller, 2020, p.17.

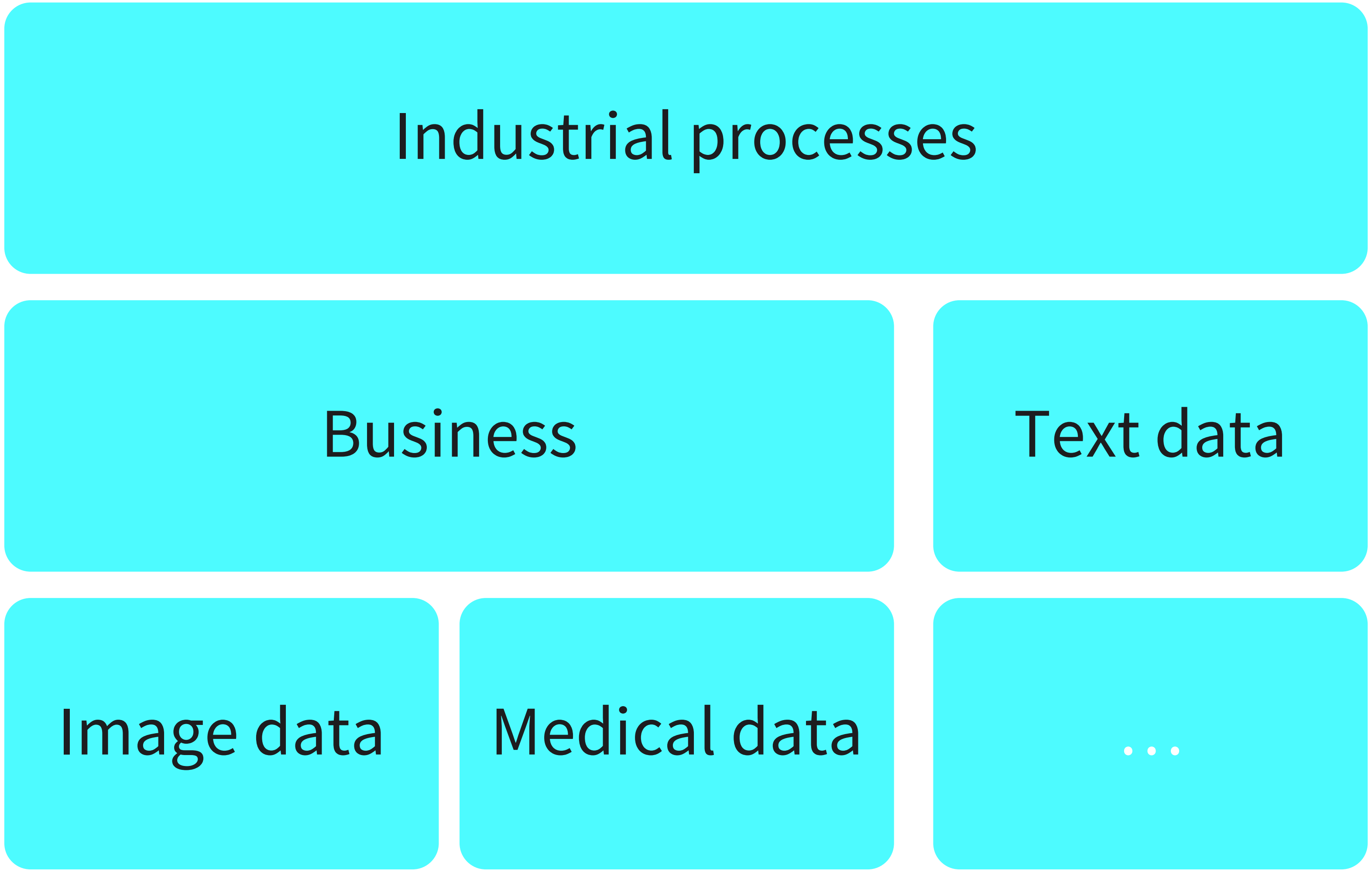
DATA SCIENCE TERMS

Model Development	
Decision Model	▪ Assesses the data to **recommend a decision**.
Regression	▪ Estimates the **dependence** between variables.
Cluster Analysis	▪ A set of **data records** into **clusters**.
Classification	▪ Categorizes entities into **predefined classes**.

Model Performance	
Probability	▪ How **likely** it is that a certain **event occurs**.
Standard Deviation	▪ How spread out the **data values** are.
Type I Error	▪ False **positive** output.
Type II Error	▪ False **negative** output.

Source of the image: Zöller, 2020, pp.18-19.

DATA SCIENCE APPLICATIONS



Industrial processes

Business

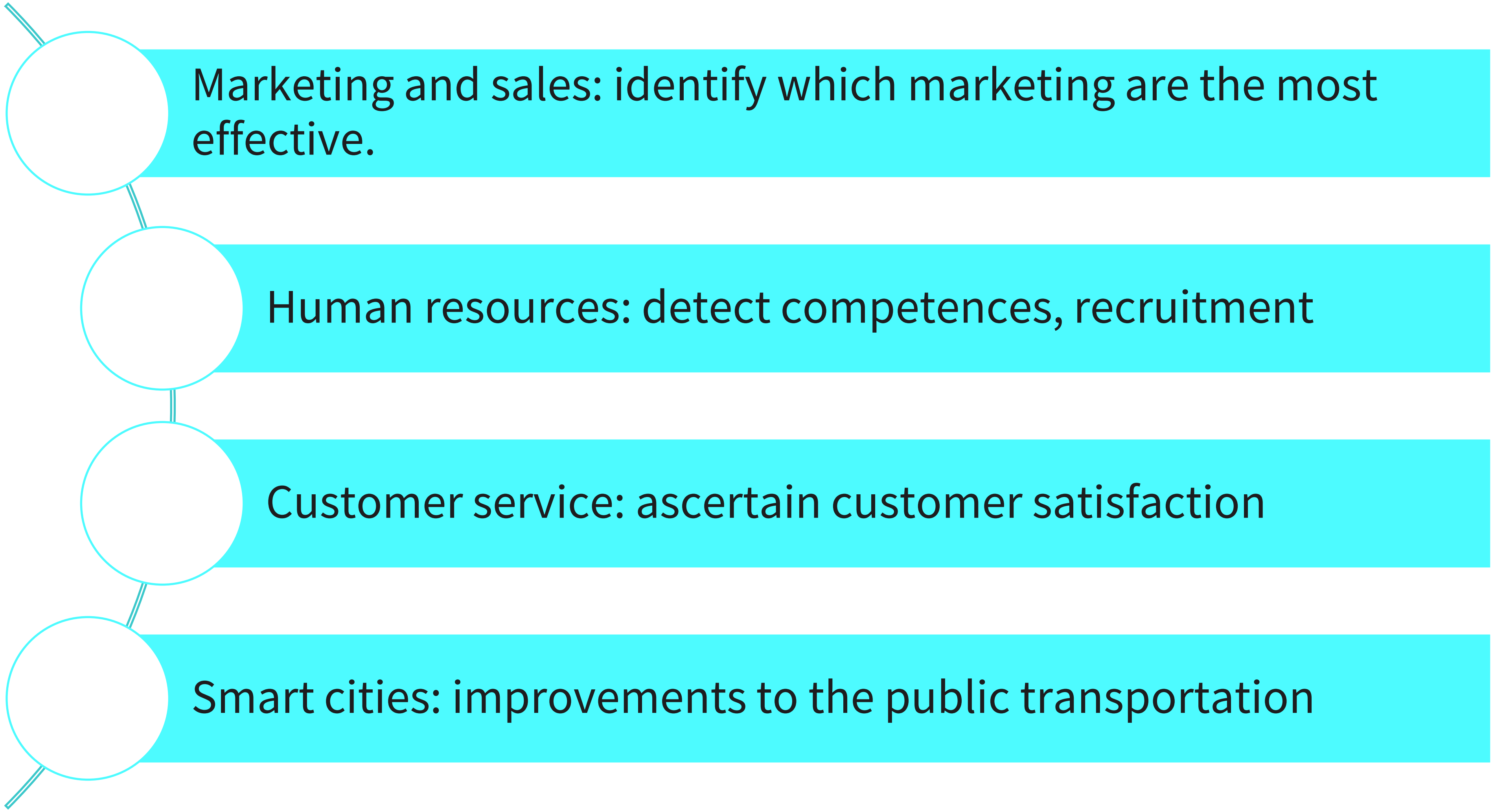
Text data

Image data

Medical data

...

DATA SCIENCE - CASE STUDIES



Marketing and sales: identify which marketing are the most effective.

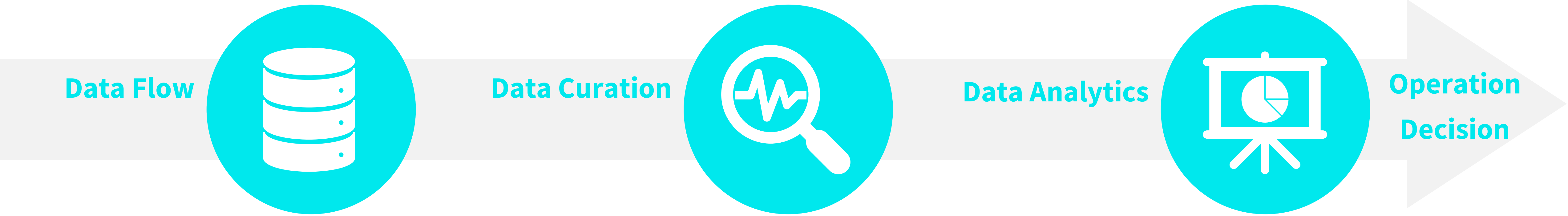
Human resources: detect competences, recruitment

Customer service: ascertain customer satisfaction

Smart cities: improvements to the public transportation

TOP 10 DATA SCIENCE PROJECTS FOR 2024

DATA SCIENCE ACTIVITIES



Data Flow

Data Curation

Data Analytics

Operation Decision

- Data collection from different sources
- Data storage
- Data accessing

Example of customer churn: Combine data from historical marketing interactions and purchases with demographic data

- Data cleaning
- Data presentation
- Data evaluation

- Treat outliers and missing values
- Inspect visual patterns

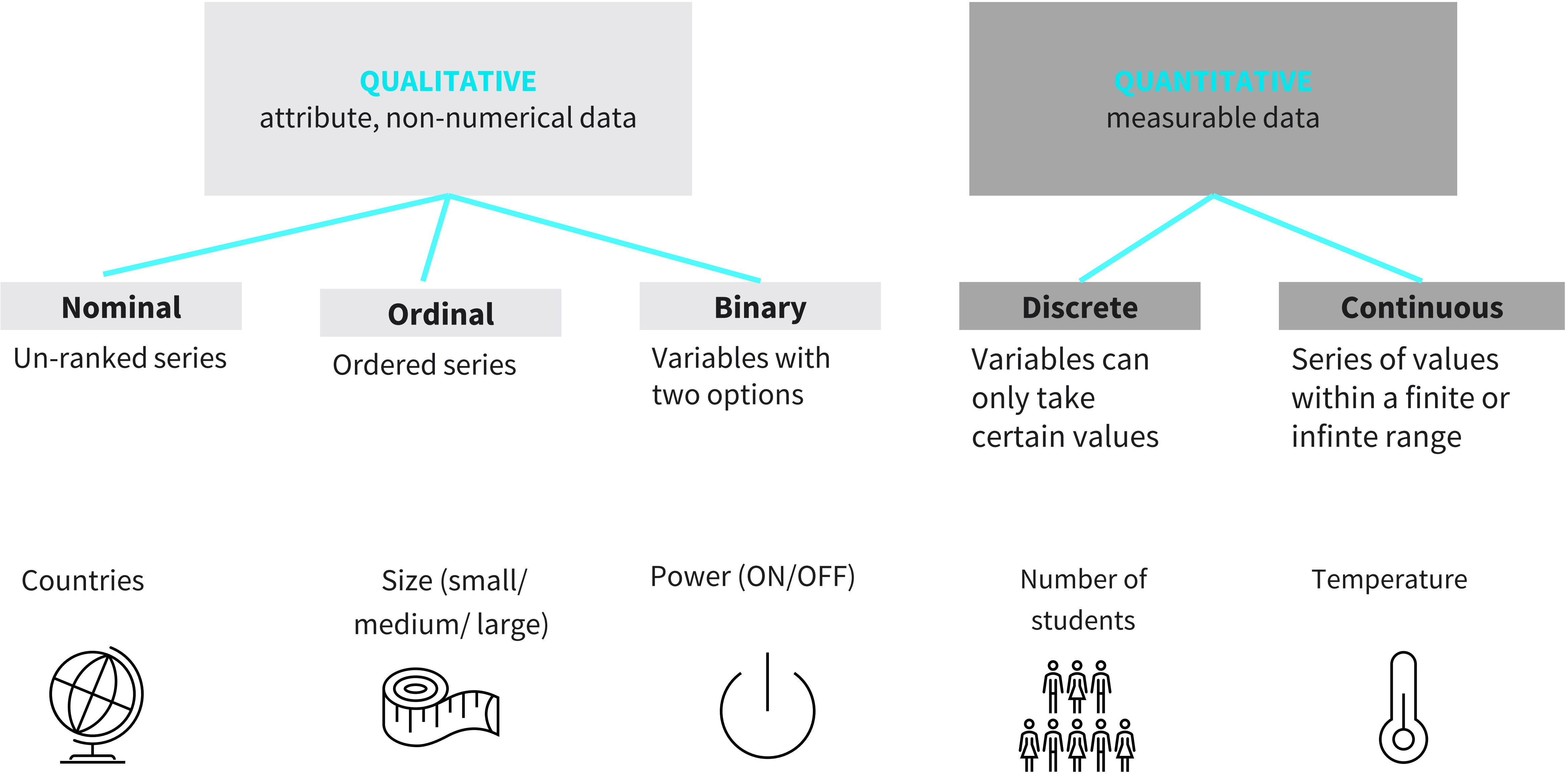
- Descriptive statistics & statistical analysis
- Modeling
- Visual techniques

- Build ML model to predict probability of customers leaving
- Create value from data insights
- Drive business decisions

SOURCES OF DATA

- Organizational and trademarked data sources
- Government data sources
- Academic data sources
- Web page data sources
- Media data sources

DATA TYPES



QUALITATIVE
attribute, non-numerical data

QUANTITATIVE
measurable data

Nominal

Un-ranked series

Ordinal

Ordered series

Binary

Variables with two options

Discrete

Variables can only take certain values

Continuous

Series of values within a finite or infinte range

Countries

Size (small/ medium/ large)

Power (ON/OFF)

Number of students

Temperature

DATA SHAPES

Structured Data

- Pre-defined data models
- Can be displayed in rows and columns
- Example: customer database (address, name, age etc.)

Name	Age	Address	Gender
John	30	City	m
Marie	4	Village	f

Semi-structured

- Contains some **tags**/attributes among unstructured data
- Example: Mails, Tweets

From: John Doe johndoe@mail.com
To: Marie Doe mariedoe@mail.com
Subject: Hello

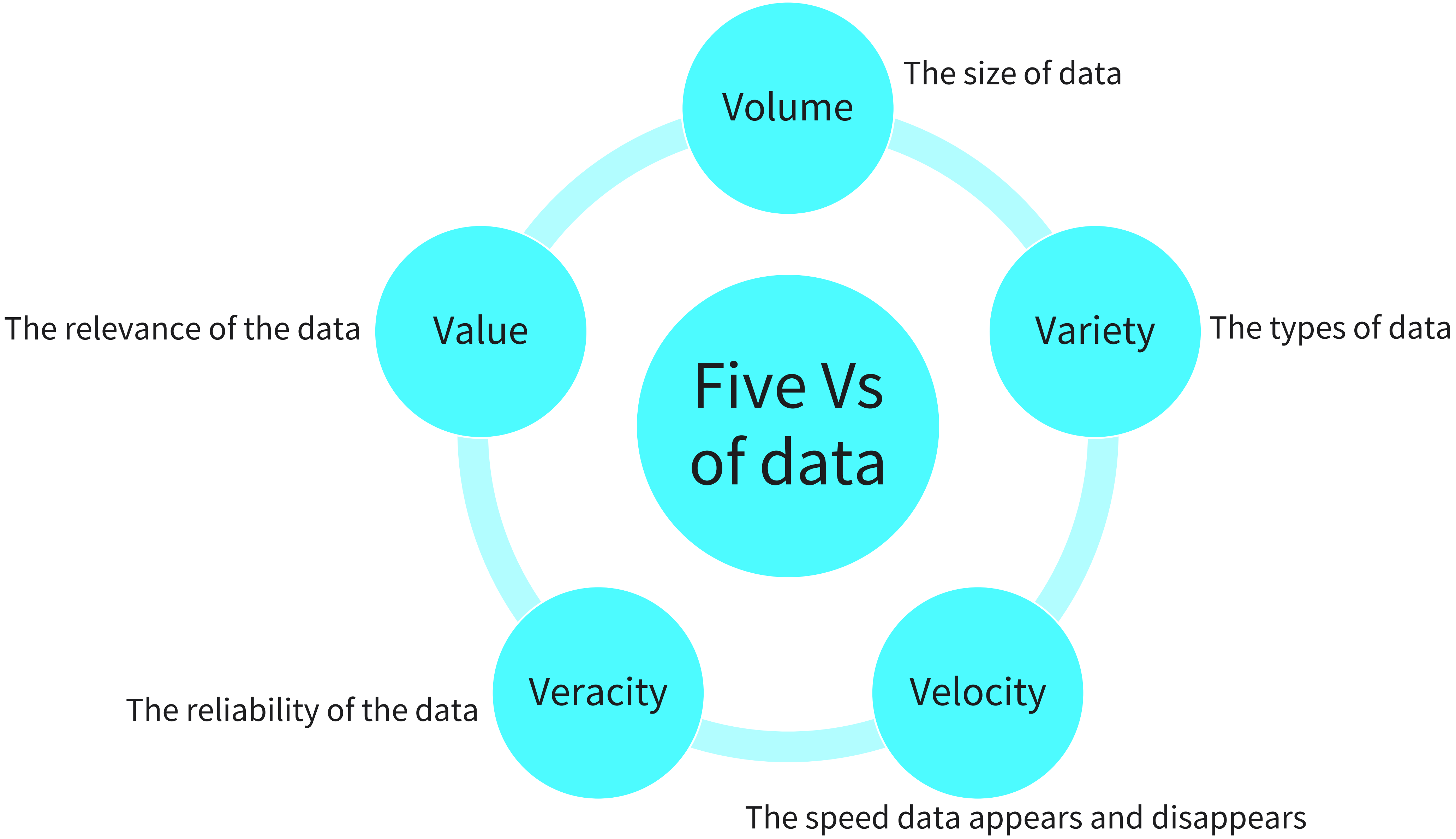
Hi Marie,
How are you?

Unstructured Data

- Unknown form or structure
- Example: Online Reviews, Audio files, Videos, Images

★★★

The book is fabulous! I enjoyed it!



Five Vs of data

Volume
The size of data

Variety
The types of data

Velocity
The speed data appears and disappears

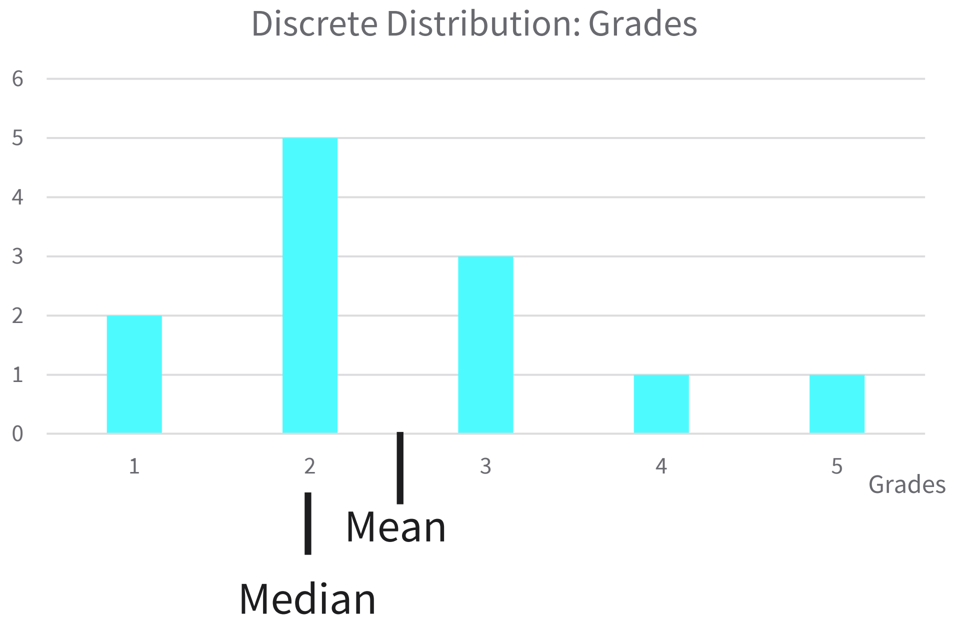
Veracity
The reliability of the data

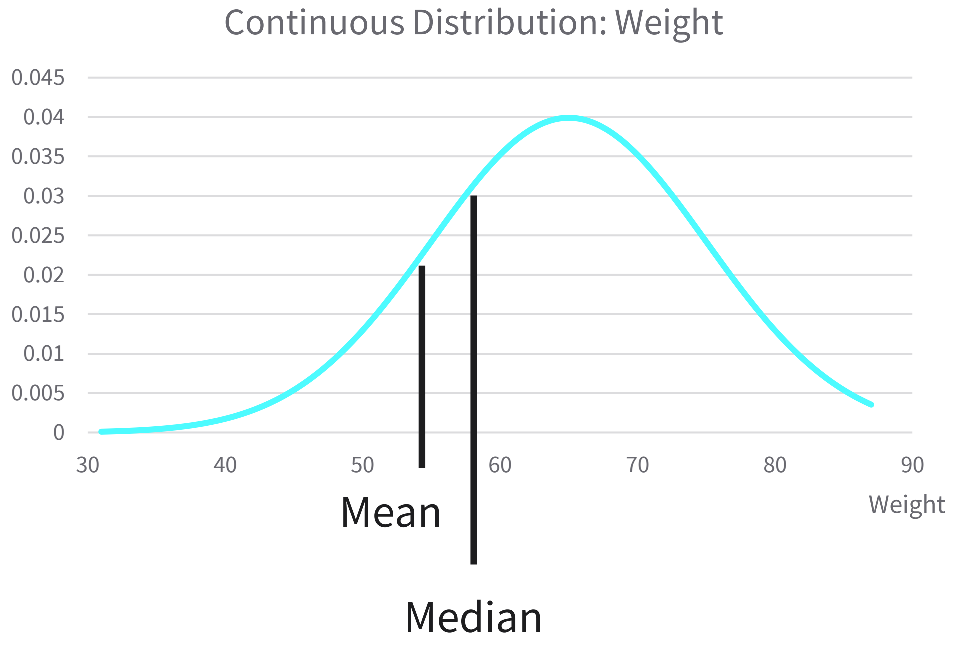
Value
The relevance of the data

Standard deviation = measure of spread

Mean = average

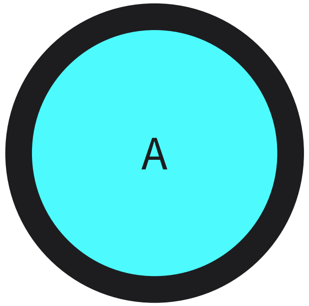
Median = 50% greater, 50% smaller values





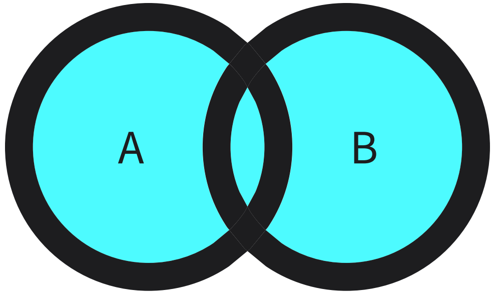
Source of the graphics: Drawing by Antonia Schulze, 2021.

DESCRIPTIVE STATISTICS – PROBABILITY THEORY



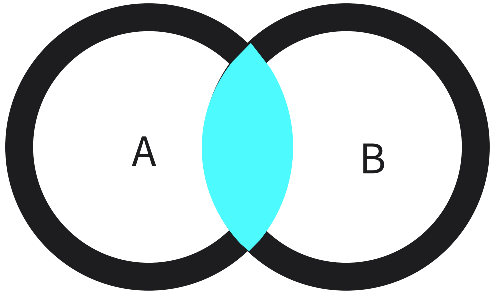
$P(A)$

Probability of an event A happening



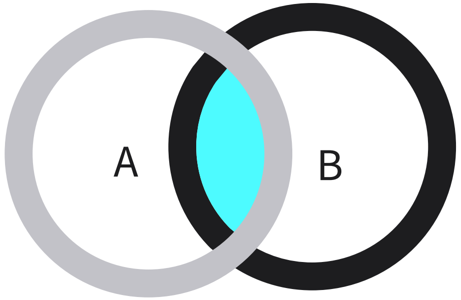
$P(A \cap B)$

Probability of event **A or B** happening



$P(A \cup B)$

Probability of event **A and B** happening

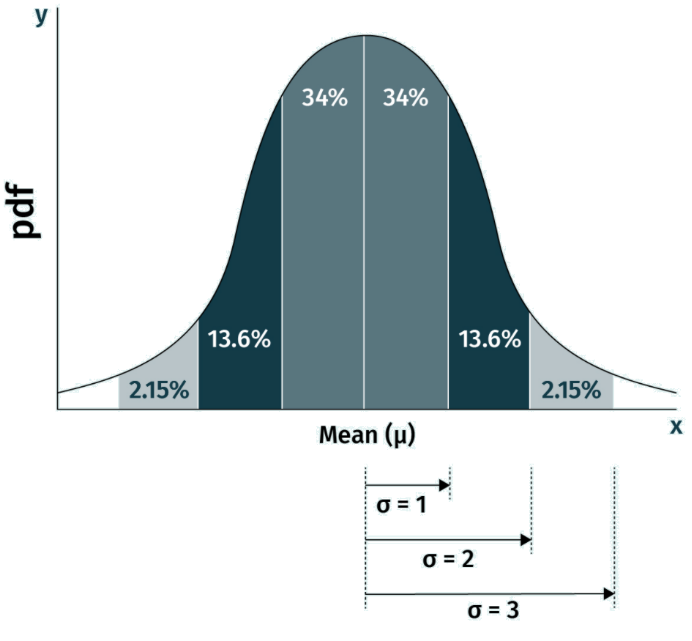


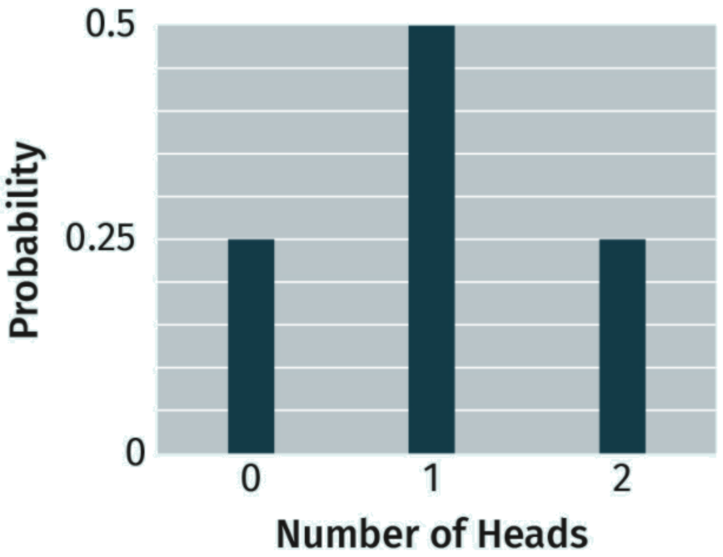
$P(A \mid B)$

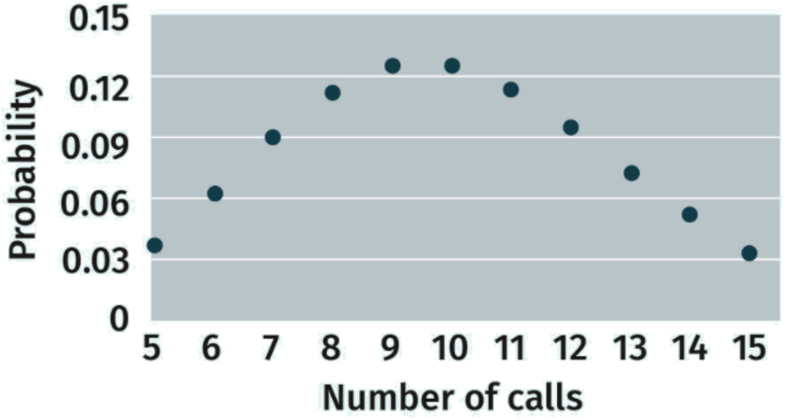
Probability of A, given that event B already happened

Conditional Probability

$$\frac{P(A \cap B)}{P(B)}$$







$$P(x) = \frac{e^{-\mu}\mu^{x}}{x!}$$

Normal Distribution

- Bell curve shape
- *Example: weight, height distribution*

Binomial Distribution

- Two possible outcomes
- *Example: P(# of heads) if toss coin twice*

Poisson Distribution

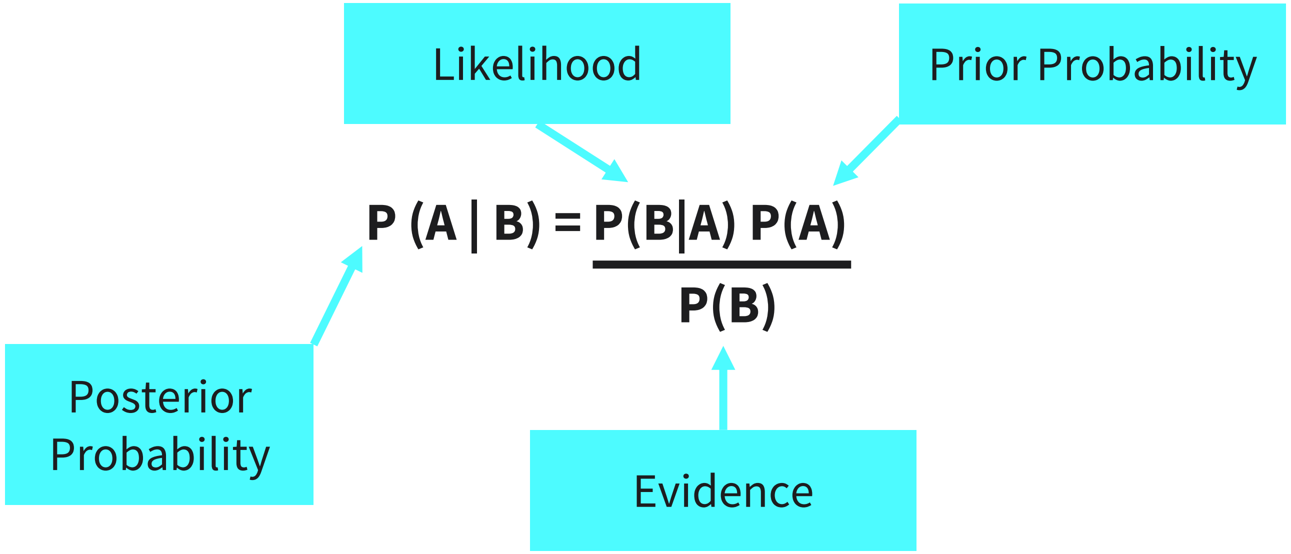
- Frequency of intervals between independent events
- *Example: P(# of calls per day) if average 5 calls per day*

Source of the graphics: IU International University, DLMBDSA01 Course Book.

Bayes Theorem

Revising a probability, given additional data is gathered

→ Possibility to invert a conditional probability



Likelihood

Prior Probability

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

Posterior Probability

Evidence

REVIEW STUDY GOALSY GOALS

You have learned…

- the meaning of data science.
- common terms and definitions in data science.
- the different applications of data science.
- the typical sources of data.
- the types and shapes of data.
- probability distributions and Bayesian statistics.

SESSION 1

TRANSFER TASK

TRANSFER TASK

Working in groups:

- Prepare a case study to demonstrate the application of data science in an industry sector of your choice.
  - **E.g., a bank, a financial company, an e-commerce company, etc.**
- Elaborate on potential data sources, the type and shape of data
- 3-5 minutes presentations

TRANSFER TASK
PRESENTATION OF THE RESULTS

Please present your results.

The results will be discussed in plenary.

LEARNING CONTROL QUESTIONS

1. Which of the following is the blind machine learning task of inferring a binary function for unlabeled training data?

   a) Regression
   b) Unsupervised Learning
   c) Supervised learning
   d) Data processing

2. In which process are the data cleared from noise and the missing values are estimated/ignored?

   a) data preservation
   b) data security
   c) data publication
   d) data description

LEARNING CONTROL QUESTIONS

3. The probability $p(A|B)$ measures…

a) the chance of event A given knowledge that event B has occurred.

b) the chance of event B given knowledge that event A has occurred.

c) the chance that events A and B occur at the same time.

d) the chance of event A given knowledge that event B has not occurred.