

# Introduction

## Statistical Methods for Data Science

**Yinan Yu**

Department of Computer Science and Engineering

October 30, 2023

- What is *data*?
- Why do we need to do *data science*?
- Why *statistical methods*?



# What is data?

Images of you...



image from <https://en.wikipedia.org/wiki/Pedestrian>



CHALMERS



GÖTEBORGS UNIVERSITET

# What is data?

Movies you have watched...

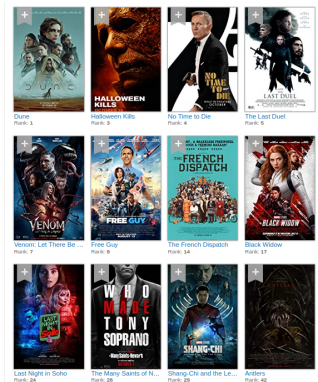


image from <https://www.imdb.com>



# What is data?

Places you have been...

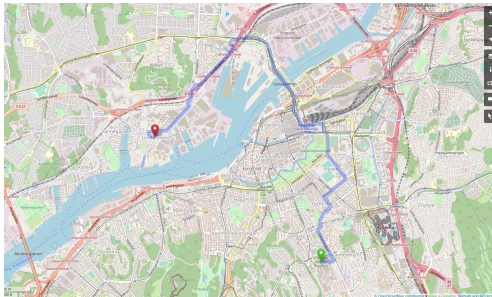


image from <https://www.openstreetmap.org>



# What is data?

- Data is everywhere
- Our personal data is being collected as we speak
- Our behaviors are being **explored**, **modeled** and **predicted** using data
- We are constantly refreshing our social media feed ... and now there is ChatGPT
- We are not in control anymore
- ChatGPT knows you better than you know yourself



# Why do we need to do data science?

- Learning what to do with data is to empower yourself
- Taking back control
- Controlling others - don't do that
- Using data for good

Is it even optional?



# Why statistical methods?

- Data is random
- We are bad at keeping track of random things
  - What was the temperature *every day* in 2022? 🐱
- We use summaries instead
  - What was the *average* temperature in 2022? 🐱

Statistical methods





# What is this course NOT about?

- Hardcore probability theory course

What is  $\sigma$ -algebra?

**Definition 1.2** ( $\sigma$ -algebra) A class of sets  $\mathcal{A} \subset 2^{\Omega}$  is called a  $\sigma$ -algebra if it fulfills the following three conditions:

- (i)  $\Omega \in \mathcal{A}$ .
- (ii)  $\mathcal{A}$  is closed under complements.
- (iii)  $\mathcal{A}$  is closed under countable unions.



Source: Klenke, Achim. Probability theory: a comprehensive course. Springer Science & Business Media, 2013.

- Introductory statistics course

2. A box contains four black pieces of cloth, two striped pieces, and six dotted pieces. A piece is selected randomly and then placed back in the box. A second piece is selected randomly. What is the probability that:

- a. both pieces are dotted?
- b. the first piece is black and the second piece is dotted?
- c. one piece is black and one piece is striped?



Source: Lee, Yong-Gu, and Sam-Yong Kim. Introduction to statistics. Yulgokbooks, Korea (2008): 342-351.

- Pure machine learning course



Support Vector Machines, Decision Trees,

Convolutional Neural Networks, Transformers



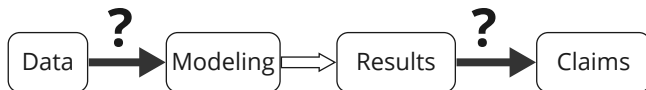
CHALMERS



GÖTEBORGS UNIVERSITET

Two foci:

- What to do with data
- How to regulate your data-related claims



In practice, this course is a mixture of probability, statistics and machine learning

# Regulate your data-related claims

- You have 3 ducks at home and they weigh 2kg, 5kg, 0.5kg each
- Oof, absolute units
- You feed them a weight loss drug called “duckiphanamin” for a month and now they weigh 1.2kg, 6kg, 0.48kg each

Question! Can you claim that duckiphanamin works?

- How about feeding duckiphanamin to your other 100 ducks? If they all lose 0.5kg each, can you claim duckiphanamin works then?

**You should be able to navigate these questions with confidence**



- Data collection and engineering
  - Not to be underestimated!
  - Not covered in the course - we work with structured data!
- Communication is important!
  - You never develop in isolation
  - Learn how to communicate efficiently
- Be patient
  - There is a lot to learn
  - Learning can be painful. Hang in there!
- **Do not hesitate to ask questions!!!**

- Information: Canvas
- Lecturer & TAs: can be found on Canvas
- Communication:
  - Email me
  - Ping me on Discord
- Student representatives:
  - Please send me an email this week if you are interested!
  - Otherwise they will be randomly selected
  - Read more about student representatives [here](#)



- Composition (100 pts):
  - Homework:
    - 3 assignments (30 pts - 10 pts each)
    - 1 project (10 pts)
  - Exam:
    - 1 oral exam (20 pts)
    - 1 take-home exam (40 pts)
- Grade: U, 3, 4, 5
  - Homework (H):
    - U: if  $H < 20$  pts
    - 3: if  $20 \text{ pts} \leq H < 30$  pts
    - 4: if  $30 \text{ pts} \leq H < 35$  pts
    - 5: if  $H \geq 35$  pts
  - Exam (E):
    - U: if  $E < 20$  pts
    - 3: if  $20 \text{ pts} \leq E < 35$  pts
    - 4: if  $35 \text{ pts} \leq E < 50$  pts
    - 5: if  $E \geq 50$  pts
  - Final grade:  $\min(\text{grade of H}, \text{grade of E})$
- Submission: Canvas

- Homework: group of max. 3 students
- Exam: individual
- Late policy:
  - Homework: 25% penalty for 0-24 hours
  - Exam: strict
- About grouping: try to team up with someone with complementary knowledge and skill sets

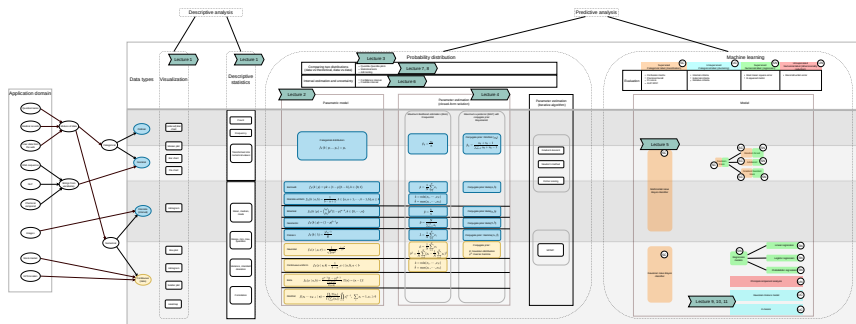


- Data types, descriptive statistics, visualization
- Probability distributions
- Modeling, parameter estimation, point estimation, interval estimation
- Hypothesis testing
- Application 1: classification, Naive Bayes classifier
- Application 2: clustering, K-means, Gaussian mixture model





Lecture map to help you keep track of where we are



Feedback and questions are welcome at  
[yinan.yu@chalmers.se](mailto:yinan.yu@chalmers.se) or [yinan@chalmers.se](mailto:yinan@chalmers.se)



# Programming language and tools

- Programming language: Python
- Interactive environment: Jupyter Notebook



- Libraries
  - Data handling and processing
    - NumPy: efficient mathematical functions
    - Pandas: structured data processing
  - Visualization
    - Matplotlib: plotting library
    - Seaborn: additional statistical plotting functions
  - Statistics
    - SciPy: a Python library for statistics and math in general
    - StatsModels: some more advanced statistical models
  - Machine learning
    - scikit-learn: predictive models and clustering

- Reading materials posted throughout the course



Have fun!

See you on the other s(l)ide(s...)!

