

Introduction

Statistical Methods for Data Science

Yinan Yu

Department of Computer Science and Engineering

November 04, 2024

- 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>



CHALMERS



GÖTEBORGS UNIVERSITET

What is data?

Movies you have watched...

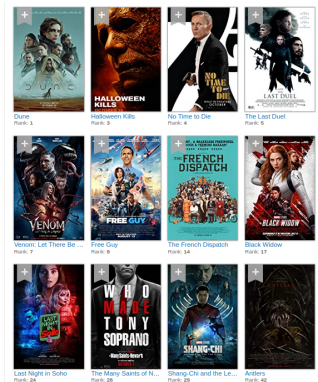


image from <https://www.imdb.com>



What is data?

Places you have been...

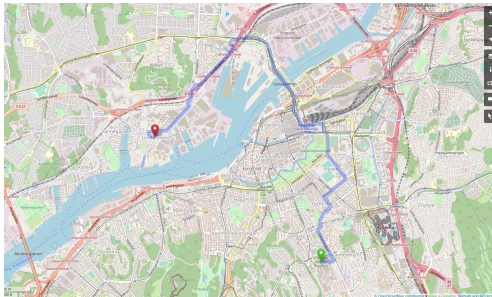


image from <https://www.openstreetmap.org>



CHALMERS



GÖTEBORGS UNIVERSITET

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
- Humans are bad at describing random things
 - What was the temperature *every day* in 2024? 🐱
- We use *summaries* instead
 - What was the *average* temperature in 2024? 🐱

Statistical methods



What is this course NOT about?

- Hardcore probability theory course

What is σ -algebra?

Definition 1.2 (σ -algebra) A class of sets $\mathcal{A} \subset 2^{\Omega}$ is called a σ -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



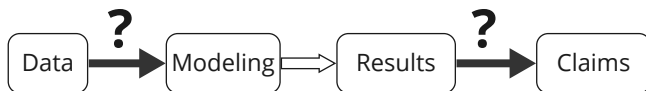
CHALMERS



GÖTEBORGS UNIVERSITET

Two focuses:

- 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, megaconker



Important source: Reddit

- 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?

After this course, 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 can be found on Canvas



Grading

- Composition (100 pts):
 - Homework:
 - 3 assignments (30 pts - 10 pts each)
 - 1 project (40 pts)
 - Exam:
 - 1 take-home exam (30 pts)
- Grade: U, 3, 4, 5
 - Homework (H):
 - U: if $H < 30$ pts
 - 3: if $30 \text{ pts} \leq H < 40$ pts
 - 4: if $40 \text{ pts} \leq H < 60$ pts
 - 5: if $H \geq 60$ pts
 - Exam (E):
 - U: if $E < 10$ pts
 - 3: if $10 \text{ pts} \leq E < 20$ pts
 - 4: if $20 \text{ pts} \leq E < 25$ pts
 - 5: if $E \geq 25$ pts
 - Final grade: $\text{round}((3.5 * \text{grade of H} + 4 * \text{grade of E}) / 7.5)$
- Submission: Canvas



- Assignment: Group work, maximum of 3 students per group
- Project: Individual; pairs allowed if lacking Python skills, with clear explanation of each person's contribution; Python code can be shared, but report needs to be individual
- Exam: Individual
- Late policy:
 - Homework: 25% penalty for submissions within 24 hours after the deadline
 - Exam: Strictly deadline, no late submissions accepted
- 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)

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



Where to start?

- Images:  pixel values 0-255
- Movies:  genre, length, production company, director
- Geographic coordinates:  (latitude, longitude)

How does a computer understand such diverse information?



Have fun!

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

