Welcome to IDS: Introduction to Data Science (NDAK16003U)

Lecture 1: 06.02.2024

Daniel Hershcovich
dh@di.ku.dk
Stella Frank
stfr@di.ku.dk

Today's Lecture

What is Data Science?

IDS formalities

Break

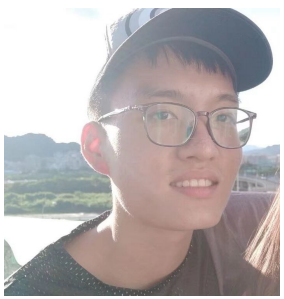
Teaching team

Teachers:

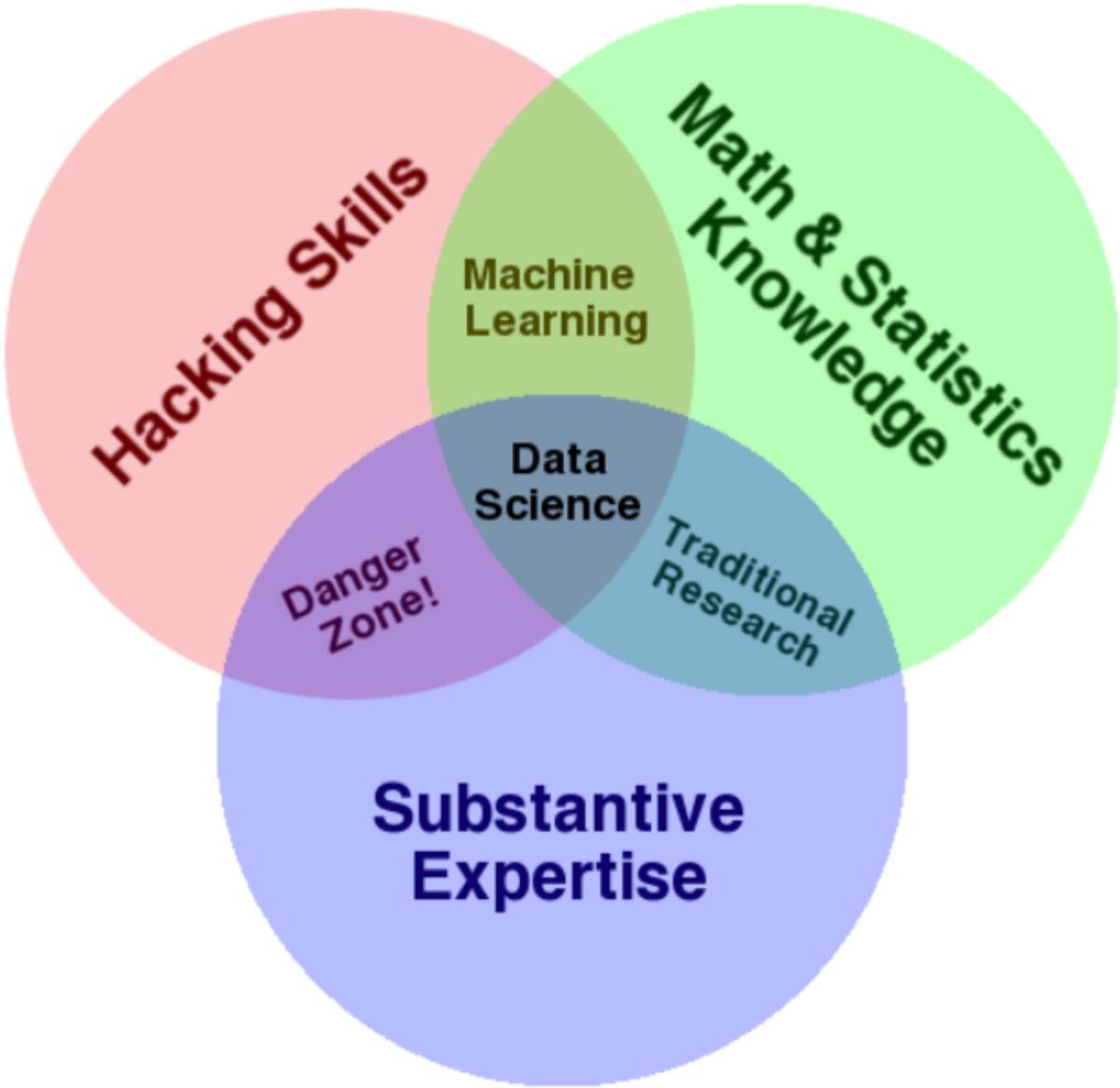
- Course responsible: Daniel Hershcovich
- Teachers: Thomas Hamelryck, Stella Frank, Morten Akhøj, Mostafa Ghazi

Teaching Assistants:

- Antonia Karamolegkou, Wenhao Gao, Mustafa Hekmat, Francisca Alves, Daniel Expósito Patiño


What is Data Science?



What is data science?

- What do you think *data science* means?

What is data science?

- What do you think *data science* means?
- Data Science involves:
  - Processing and analysing data
  - What kind of data?

What is data science?

- What do you think *data science* means?
- Data Science involves:
  - Processing and analysing data
  - What kind of data?
    - "Large" data
    - Quantitative/structured data
    - Unstructured data

What is data science?

- What do you think *data science* means?
- Data Science involves:
  - Processing and analysing data
  - What kind of data?
    - "Large" data
    - Quantitative/structured data
    - Unstructured data
- Types of analyses:
  - Descriptive statistics: How to interpret the data pile?
  - Predictive models: What do we expect about new data?

What is data science?

- What do you think *data science* means?
- Data Science involves:
  - Processing and analysing data
  - What kind of data?
    - "Large" data
    - Quantitative/structured data
    - Unstructured data
- Types of analyses:
  - Descriptive statistics: How to interpret the data pile?
  - Predictive models: What do we expect about new data?
- Data Science requires (some) Computer Science knowledge:
  - Data processing
  - Data structures, storage
  - Computability (both time and precision)

What is data science?

- What do you think *data science* means?
- Data Science involves:
  - Processing and analysing data
  - What kind of data?
    - "Large" data
    - Quantitative/structured data
    - Unstructured data
- Types of analyses:
  - Descriptive statistics: How to interpret the data pile?
  - Predictive models: What do we expect about new data?
- Data Science requires (some) Computer Science knowledge:
  - Data processing
  - Data structures, storage
  - Computability (both time and precision)
- and don't forget the Substantive Expertise!
  - Less of a focus of this course but critical in the real world.

Why data science?

- It's fun!
  - Examples of real data, real problems
  - Interdisciplinary
  - Curiosity-driven

Why data science?

- It's fun!
  - Examples of real data, real problems
  - Interdisciplinary
  - Curiosity-driven
- It's useful!
  - Data is everywhere - everything is data
  - Data science is a methodology for understanding data
  - A toolbox with powerful and increasingly widespread tools

A typical data science sequence

1. A clear question formulation: Hypothesis Statement

A typical data science sequence

1. A clear question formulation: Hypothesis Statement
2. Model Representation
   - Choose a representation of the data
     - What are the relevant factors, features for your question?

A typical data science sequence

1. A clear question formulation: Hypothesis Statement
2. Model Representation
   - Choose a representation of the data
     - What are the relevant factors, features for your question?
3. Analysis
   - Understand possible groupings of the data points
   - Understand the relationship between factors in the data

A typical data science sequence

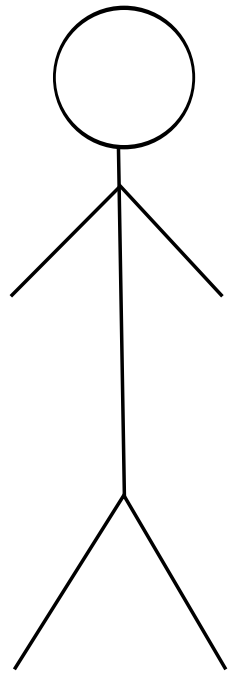
1. A clear question formulation: Hypothesis Statement
2. Model Representation
   - Choose a representation of the data
     - What are the relevant factors, features for your question?
3. Analysis
   - Understand possible groupings of the data points
   - Understand the relationship between factors in the data
4. Prediction
   - Use some data points (representations, measurements) to predict unknown or future values of other data points.

A typical data science sequence

1. A clear question formulation: Hypothesis Statement
2. Model Representation
   - Choose a representation of the data
     - What are the relevant factors, features for your question?
3. Analysis
   - Understand possible groupings of the data points
   - Understand the relationship between factors in the data
4. Prediction
   - Use some data points (representations, measurements) to predict unknown or future values of other data points.
5. Visualization
   - Presentation of the data (representation) analysis in a way which is easy to interpret visually

A silly example

1. Question: Do people grow taller as they grow older?

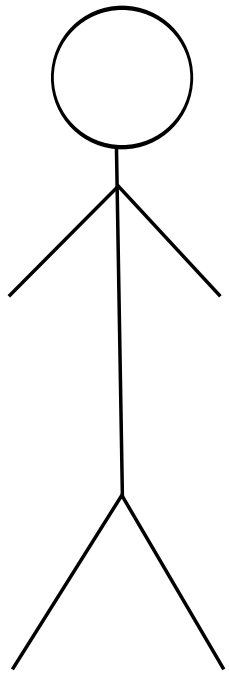


height h = 145 cm

age a = 13 years

A silly example

1. Question: Do people grow taller as they grow older?
2. Representation: What is relevant to the question?

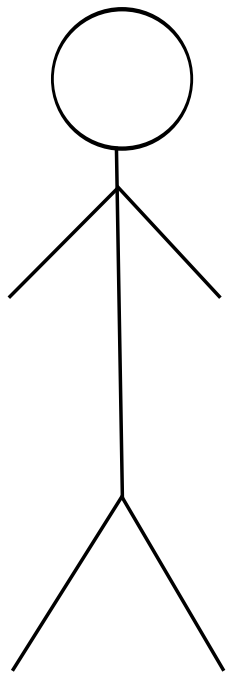


height h = 145 cm

age a = 13 years

A silly example

1. Question: Do people grow taller as they grow older?
2. Representation: What is relevant to the question?
   - Definite Factors: age, height
   - Irrelevant Factors: favorite color
   - Possible Factors: sex, nationality, income, shoe size, birth year

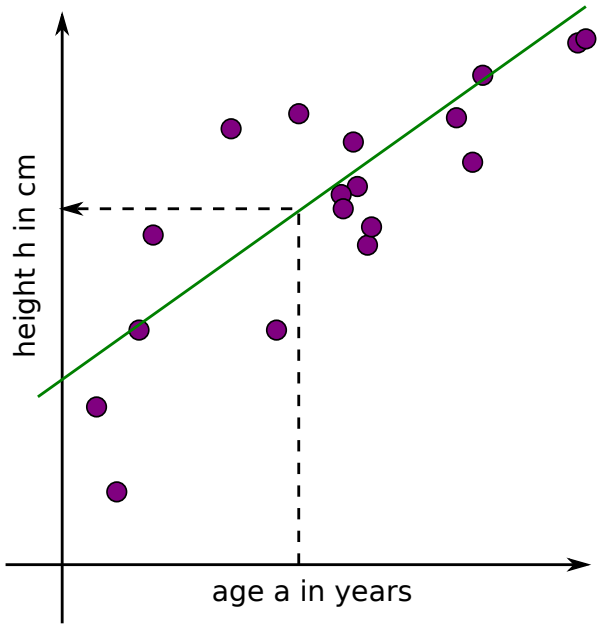


height h = 145 cm

age a = 13 years

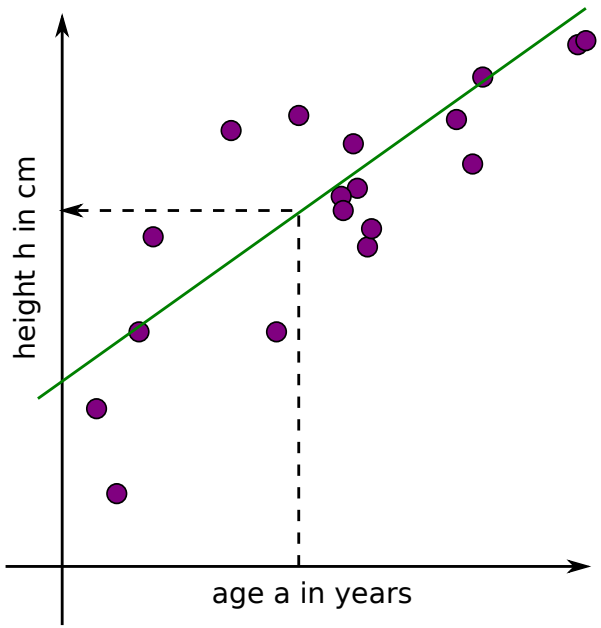
A silly example

1. Question: Do people grow taller as they grow older?
2. Representation: What is relevant to the question?
3. Analysis
   - Understand the relationship between factors in the data



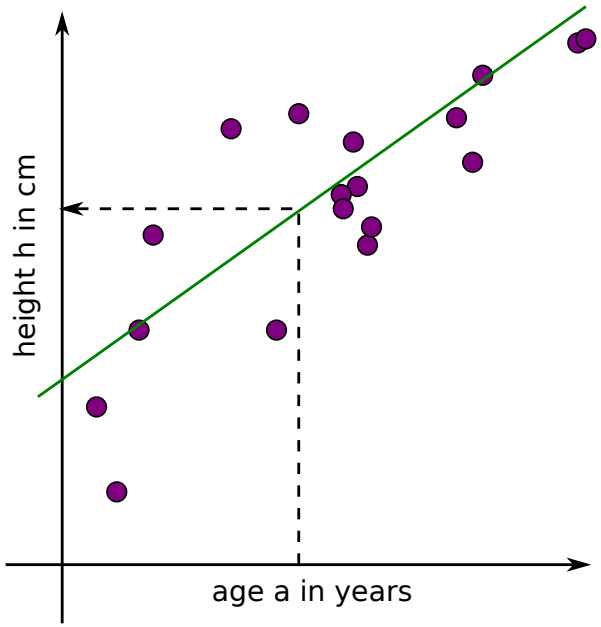
A silly example

1. Question: Do people grow taller as they grow older?
2. Representation: What is relevant to the question?
3. Analysis
   - Understand the relationship between factors in the data
4. Prediction
   - Use some data points (representations, measurements) to predict unknown or future values of other data points.
   - What's wrong with these predictions?



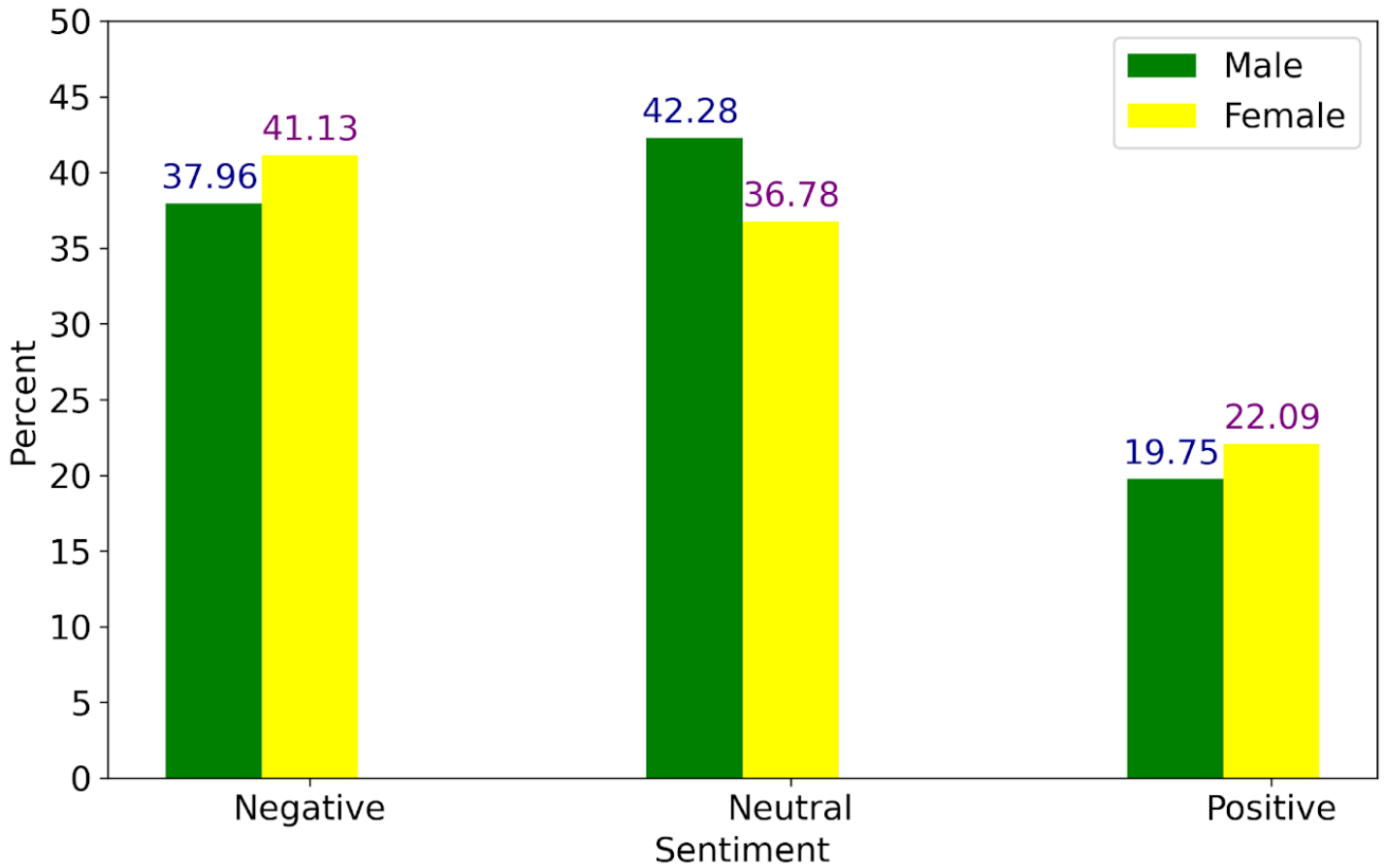
A silly example

1. Question: Do people grow taller as they grow older?
2. Representation: What is relevant to the question?
3. Analysis
   - Understand the relationship between factors in the data
4. Prediction
5. Visualization
   - Presentation of the analysis in a way which is easy to interpret visually
   - What's wrong with this graph?



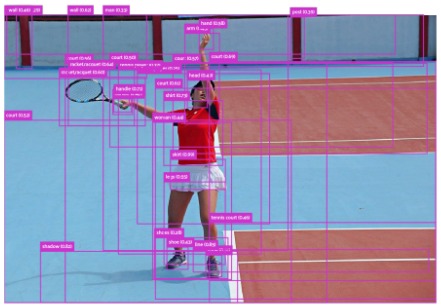
Data Science in our Research: Daniel Hershcovich

- Natural Language Processing
- Example project: Were the texts by late 19th-century Scandinavian female authors particularly unhappy?

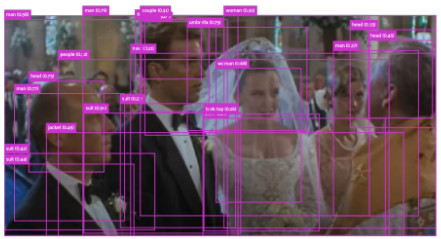


Data Science in our Research: Stella Frank

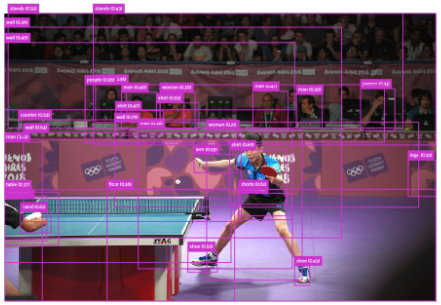
- Computer Vision, Natural Language Processing, Cognition
- Example project: How commonly-used visual features transfer to different cultural domains



(a) In domain: Tennis (MarVL-id)

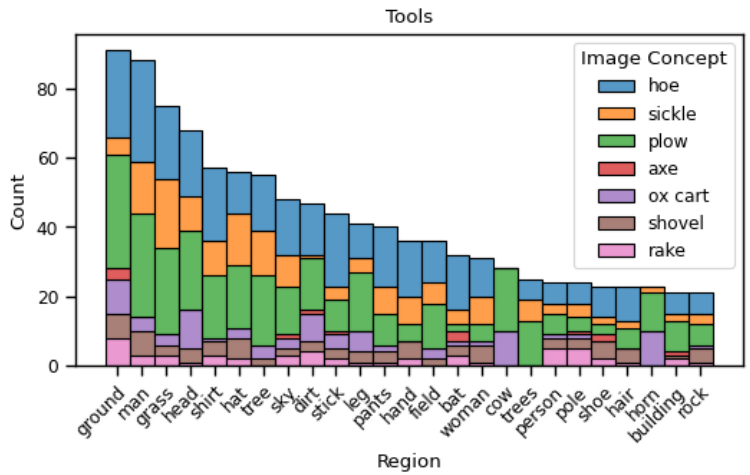


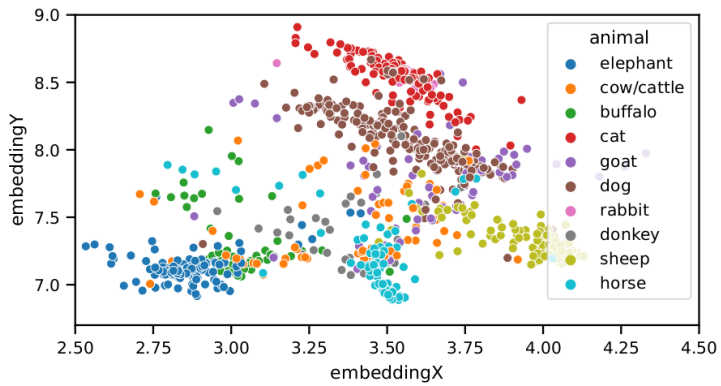
(b) In domain: Western wedding (GD-VCR West)



(c) Semantic shift: Table tennis (MarVL-id)

(d) Background shift: Indian wedding (GD-VCR South Asia)

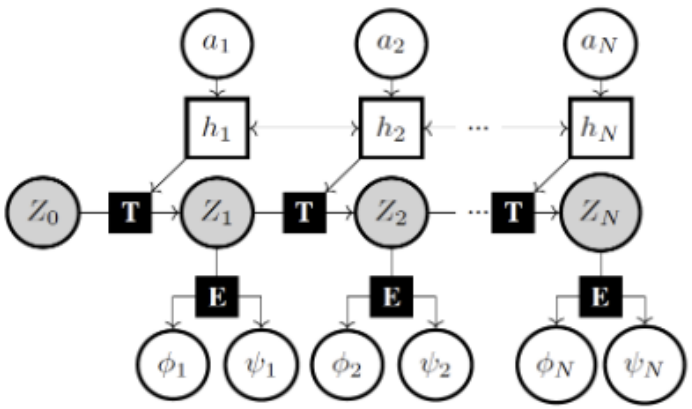


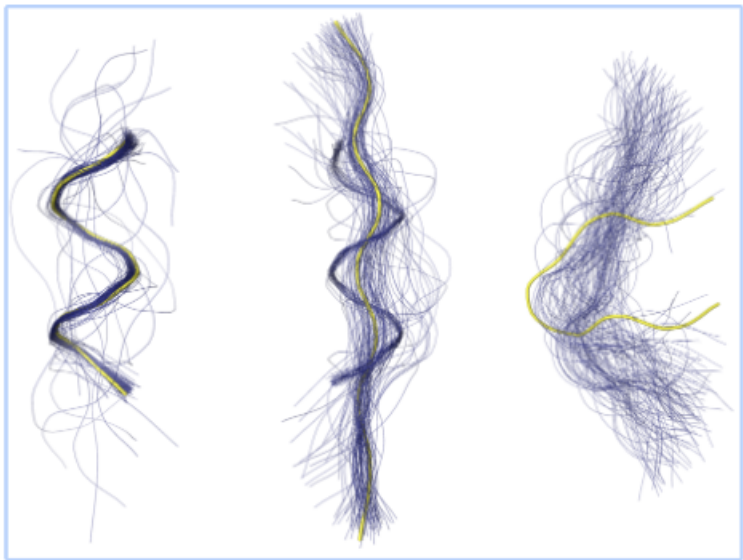


Data Science in our Research: Thomas Hamelryck

Thomas Hamelryck

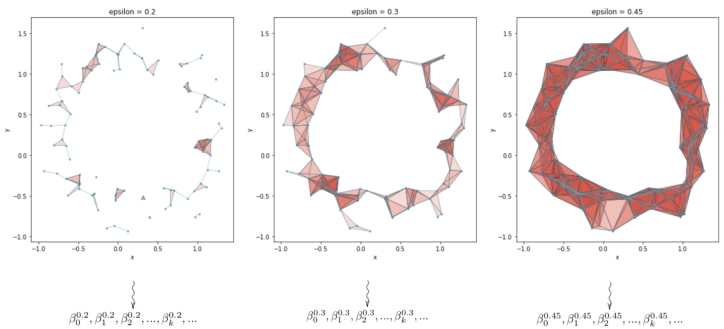
- Bayesian modelling, probabilistic machine learning, deep probabilistic programming.
- Example project: Deep model of protein structure (ICML 2021)
  - We use a deep Markov model (left) to sample possible protein 3D shapes (right) given their amino acid sequence.
  - Used in vaccine design by Danish biotech company Evaxion.





Data Science in our Research: Daniel Expósito

- Algebraic topology, statistics, topological data analysis
- Example approach: Perseverance Homology
  - Create a mesh from a data set by joining nearby points → Analyze mesh using homology (math tool) → let joining distance vary → Find essential shape of data
  - Good for complex data: noisy, high-dimensional, incomplete



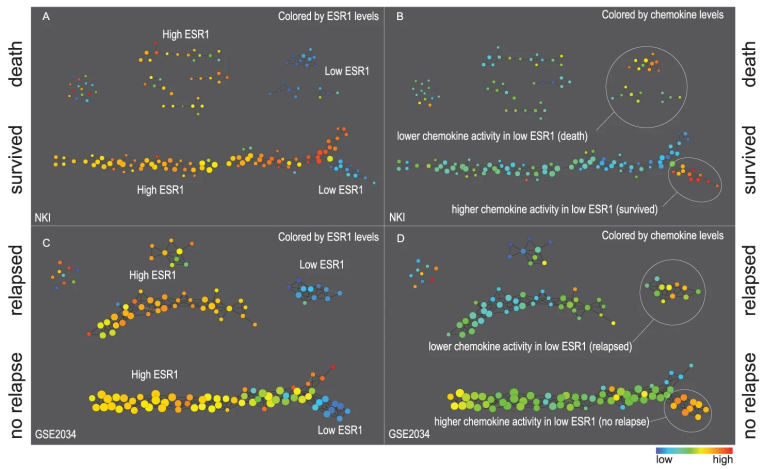


Figure: Cancer subtypes

[1]Eric Bunch [2]Lum, P., Singh, G., Lehman, A. et al. Extracting insights from the shape of complex data using topology

IDS formalities

By the end of this course, students will gain:

- **Foundational Knowledge:**
  - Principles of data analysis.
  - Basics of probability theory.
  - Key concepts in machine learning (classification, regression, clustering).
  - Awareness of common machine learning pitfalls.
- **Practical Skills:**
  - Application of linear and non-linear techniques.
  - Data clustering basics.
  - Proficiency in ML toolboxes.
  - Skills in data visualization and results evaluation.
  - Handling of machine learning pitfalls.
- **Applied Competences:**
  - Identification of ML applications in research.
  - Evaluation and selection of ML methods.
  - Real-world problem-solving using ML techniques.

Your backgrounds

Absalon survey (thanks for completing it!)

Preliminary lecture overview

Week	Tuesday	Thursday
Week 6	**06.02.2024.** 10.00-12.00 Lecture 1: Introduction. Data and Visualization with Python (SF) Reading: - Python: [JG] chapters 1 to 3. - Visualization: C. Wilke, Fundamentals of Data Visualization: 1-5 + 29 **Assignment 1 released 06.02.2024.**	**8.02.2024.** 9.00-12.00 Lecture 2: Statistics (MAP) Reading: [SR] chapters 1 & 2. 13.00-16.00 TA session: • Introduction to Python, NumPy, Pandas • Python installation and environment setup • Assignment 1 questions
Week 7	**13.02.2024.** 10.00-12.00 Lecture 3: Probabilities (MAP) Reading: [SR] chapters 3 & 4.	**15.02.2024.** 9.00-12.00 Lecture 4: Hypothesis testing (TH) Reading: [SR] chapter 8. 13.00-16.00 TA session: • Assignment 1 tips and questions • Statistics tips and clarifications
Week 8	**Assignment 1 due 19.02.2024 at 15:00 latest.** **20.02.2024.** 10.00-12.00 Lecture 5: Introduction to Machine Learning / k-NN (TH) Reading: [AML] chapter 6 **Assignment 2 released 20.02.2024.**	**22.02.2024.** 9.00-12.00 Lecture 6: Bayesian statistics I (TH) Reading: Bayesian statistics and modelling, Nature reviews, 2021 13.00-16.00 TA session: • Plotting tips • Assignment 2 tips and questions

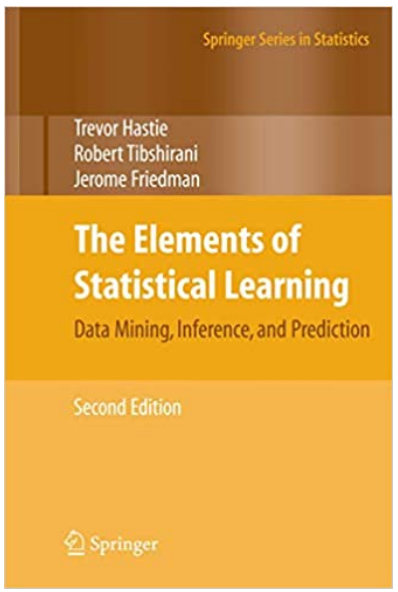
Teaching

- Lectures:
  - Tuesdays 10-12, **Auditorium 4, HCØ**
  - Thursdays 9-12, **Auditorium 4, HCØ**
- TA classes (starting this Thursday 9.02): 13-16.
  - Group 1: Aud 8 HCØ
  - Group 2: Aud 10 HCØ
- We will split you to groups before Thursday based on the background survey
- Ask questions, also during lectures
- Use your TAs as much as possible. They are here to help
- Use the discussion forum on Absalon

Textbooks

We will provide extracts from a selection of reference books, either as files on Absalon or links to websites.

- Data Science from Scratch: First Principles with Python (2nd Ed.) by J. Grus
- Fundamentals of Data Visualization by C. Wilke
- T. Hastie, R. Tibshirani, J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction"
- Probabilities and Statistics for engineers and scientists, by Shelden Ross, 4th edition, AP/Elsevier.
- Recommended: C. Bishop, Pattern Recognition and machine Learning, Springer. From 2006, but still a great reference! Especially from the Bayesian point of view.



Python

- In IDS, we use Python 3.x.
- You can get Python from the Anaconda distribution: (`https://www.anaconda.com`); it contains both Python and the most useful libraries for data science.
- This course uses Jupyter notebooks, included with Anaconda.
- We recommend Google Colab (`https://colab.research.google.com`): a hosted Jupyter Notebook service
- For your homework assignments, you may be given code templates. Please use them!
- Useful Python resources on Absalon (under `Pages/Useful Resources`)
- StackOverflow (`http://stackoverflow.com`) is an invaluable resource (but also dangerous)
- You may use GitHub Copilot, ChatGPT etc. at your own risk

Exam form: Assignments

- Continuous evaluation: 5 assignments
- Assignments are *individual* work
- Mixture of report and supporting code
- Assignments are weekly-ish
- Assignments are weighted equally
- Final 7-scale grade, with minimum of 02 to pass the course.

Weekly assignments best practices: code

- Ask yourself:
  - Does my answer make sense?
  - If not, try to find your bug
- Is the assignment unclear? **Ask!** Use the forum! We want you to spend time doing the assignment, not reading it.
- You submit source code files (***.py**) or Juptyer Notebooks (***.ipynb**) in a zip archive (**firstname.lastname.zip)**.

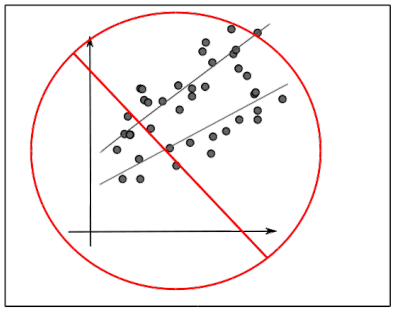
Weekly assignments best practices: code

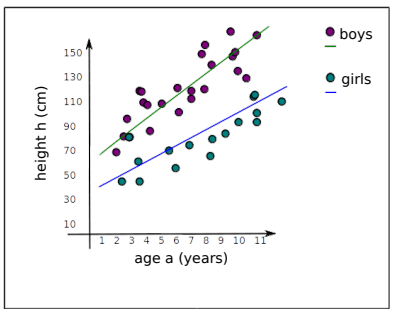
- Be sure that your code can be executed!
- Do not use unlisted packages.
  We use: `numpy`, `scipy`, `pandas`, `matplotlib`, `seaborn`, `sklearn`. If more packages are necessary, they will be explicitly mentioned in the assignment text. They are all easily available via Anaconda.
- If you need to read some data from a file — and you will need it in every assignment — be sure that
  1. the data file is **included** in your archive,
  2. it is read from the **current** folder.

  Otherwise, graders may have a hard time...

Weekly assignments best practices: report

- **Submit a pdf named firstname.lastname.pdf**
- Make sure your pdf can be opened with standard Mac/Linux/Windows viewers (sometimes Word produces terrible results)
- Use the spell checker!
- Make sure figures and plots are properly formatted and explained:
  - A title that tells me what I am looking at
  - Description of axes
  - Visible points/tick marks along axes
  - Good choice of colors (to see differences)





- You may convert your Jupyter Notebook into a PDF. You can add markdown cells with formatted answers. If you choose to do so, **sanitize it before converting it to PDF. This should be easily readable!**

Academic Code of Conduct

Discussion is encouraged, but **sharing code or direct copying** is prohibited and considered **plagiarism**. Refer to University's plagiarism regulations for clarity. Use Absalon forum for assignment queries.

Plagiarism includes unacknowledged copying of text or ideas. Always **cite sources** properly, including **external sources** and **lecture/lab code**.

AI Assistance Guidelines: NEW!

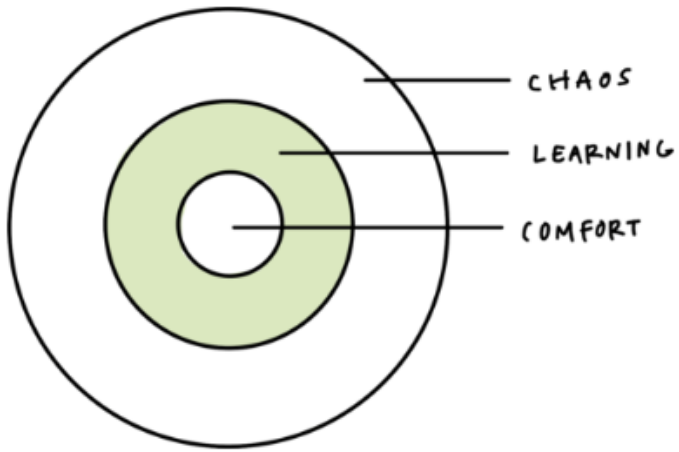
- You may use AI tools (e.g., ChatGPT, GitHub Copilot) for **writing assistance**, **tutoring**, **coding**, and **literature search**.
- **Declare AI usage** in submissions, including **tool** and **version**. Highlight AI-generated content. Include prompts/transcripts.

Example declaration:

```
ChatGPT 3.5 was used as a writing assistance tool
and while developing code.
```

Continuous Feedback

- We try to improve the course every year.
- Feedback is welcome!



Break

Please find a chart/data visualization
either from your coursework or from www.dst.dk
(Statistics Denmark).