

# **Data Wrangling and Data Analysis**

**Hakim Qahtan**

**Erik-Jan van Kesteren**

**Ayoub Bagheri**



Utrecht University

# Today

- Who are we?
  - Introduction to the course content
  - Getting to know you a little
  - Break
  - Practicalities & course flow
  - Time for questions
- 
- But first...



# Who are we?



Utrecht University

# Faculty of Science team

Hakim Qahtan

[a.a.a.qahtan@uu.nl](mailto:a.a.a.qahtan@uu.nl)

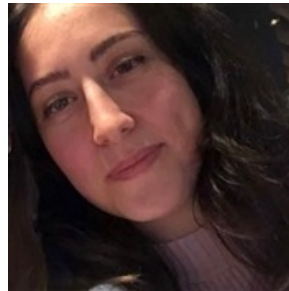


Ramón Rico Cuevas

[r.ricocuevas@uu.nl](mailto:r.ricocuevas@uu.nl)

Duygu Islakoglu

[d.s.islakoglu@uu.nl](mailto:d.s.islakoglu@uu.nl)

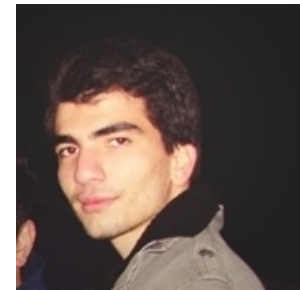


Mahsa Ghanavatinasab

[m.ghanavatinasab@uu.nl](mailto:m.ghanavatinasab@uu.nl)

Anastasia S. Apeiron

[b.s.akkuzu@uu.nl](mailto:b.s.akkuzu@uu.nl)



Vahid Shahrivari Joghan

[v.shahrivarijoghan@uu.nl](mailto:v.shahrivarijoghan@uu.nl)



Utrecht University

Lab Group 1

Lab Group 2

Lab Group 3



# Social & Behavioural Science team

Erik-Jan

[e.vankesteren1@uu.nl](mailto:e.vankesteren1@uu.nl)



Ayoub

[a.bagheri@uu.nl](mailto:a.bagheri@uu.nl)



Daniel

[d.loberski@uu.nl](mailto:d.loberski@uu.nl)



Elena

[e.candellone@uu.nl](mailto:e.candellone@uu.nl)



Jelle

[j.j.teijema@uu.nl](mailto:j.j.teijema@uu.nl)



Özgür

[o.togay@uu.nl](mailto:o.togay@uu.nl)



Mahdi

[m.shafieekamalabad@uu.nl](mailto:m.shafieekamalabad@uu.nl)



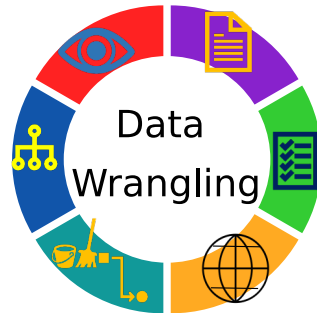
# Introduction



Utrecht University

# Extracting value from data

**Data**

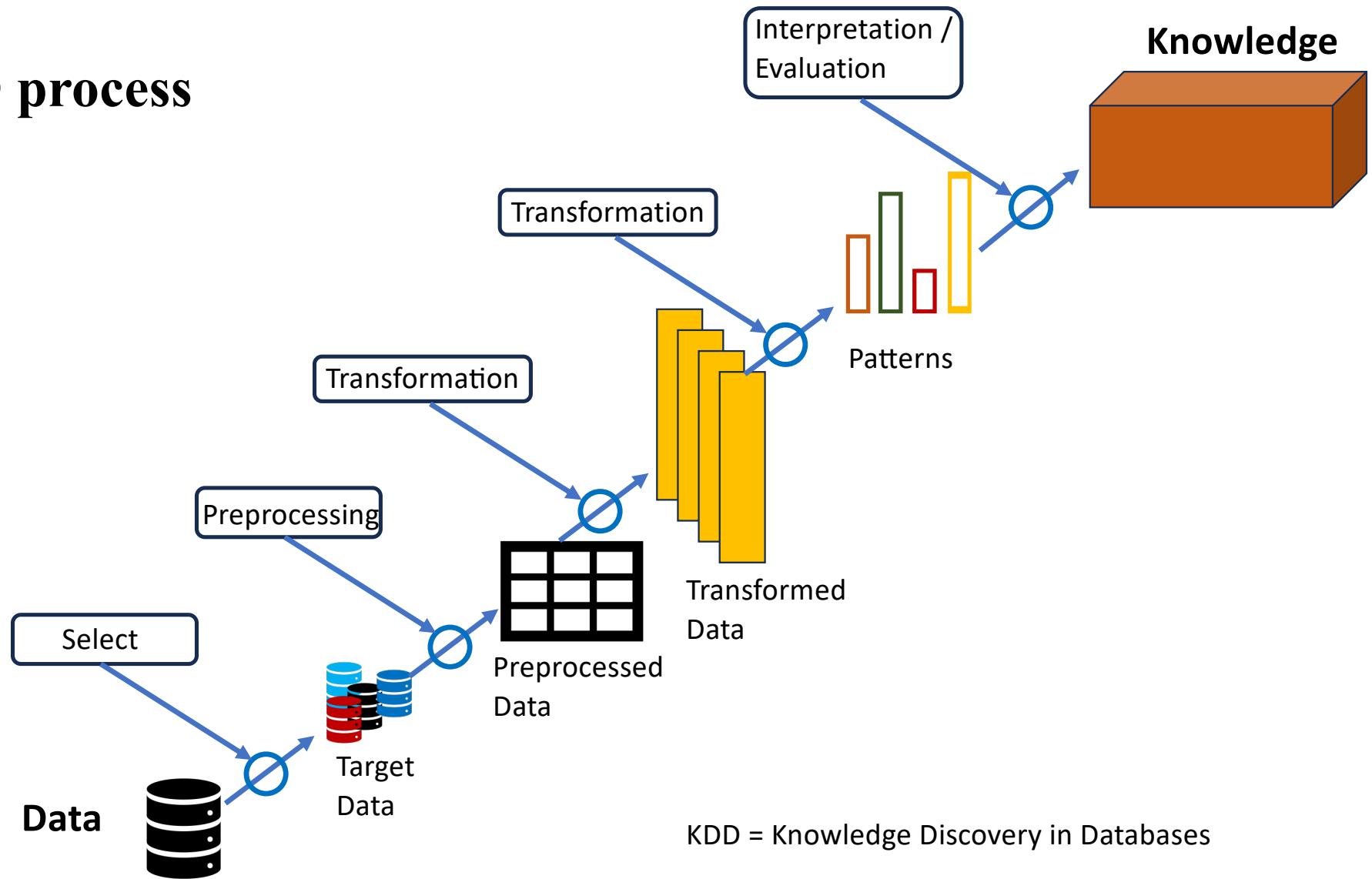


**Information**



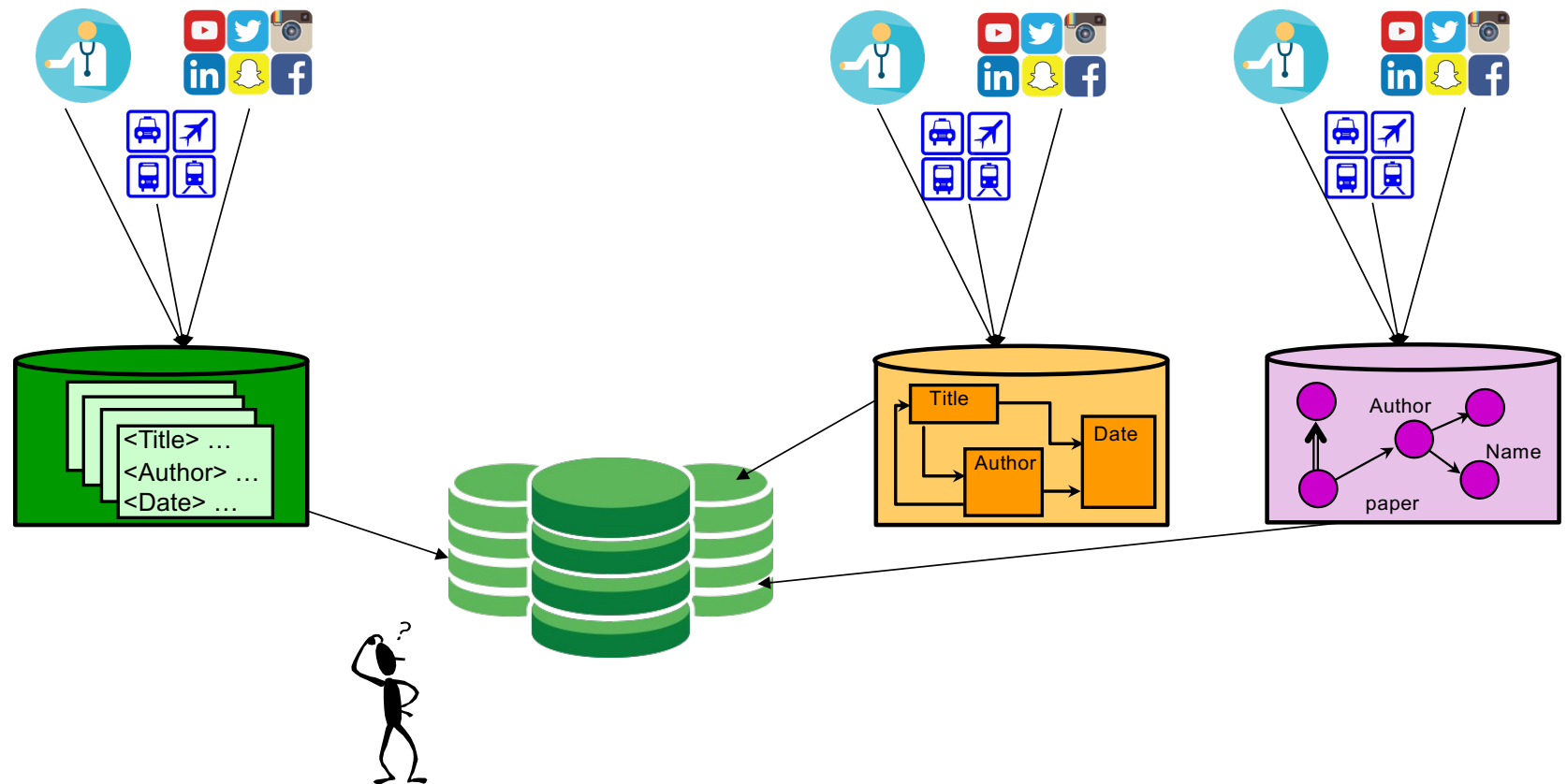
Utrecht University

# KDD process

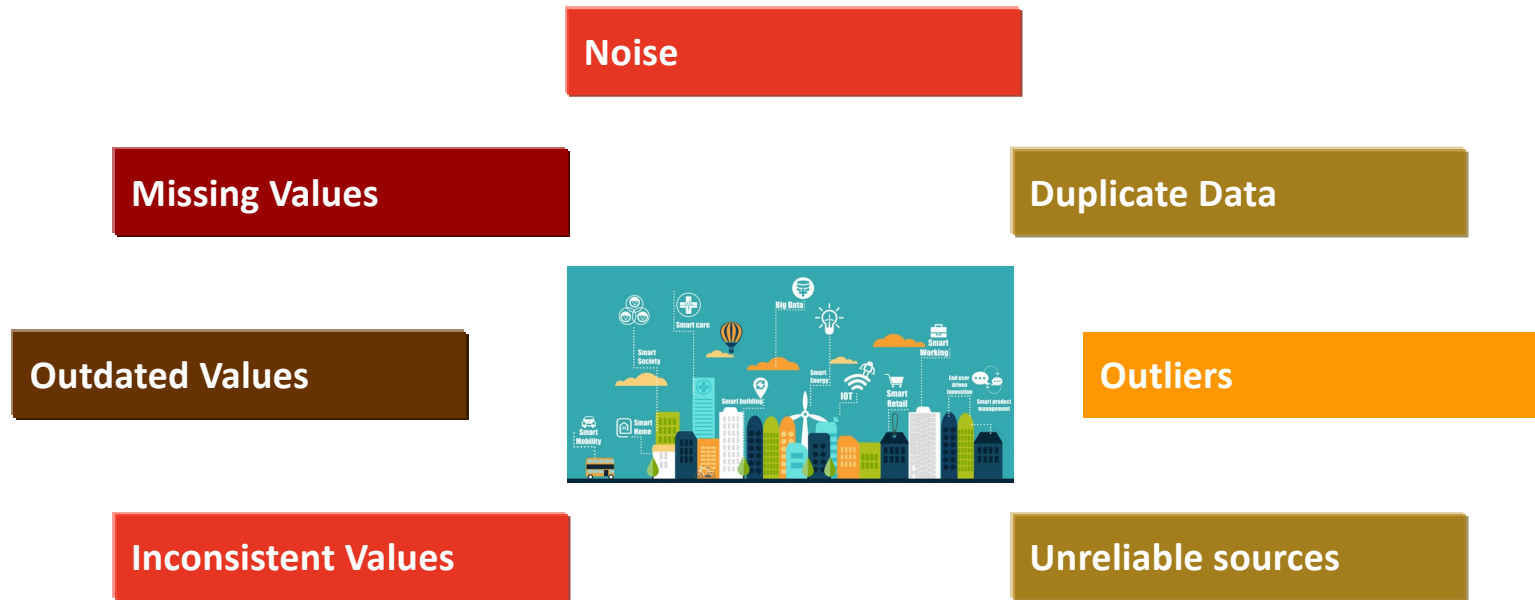


KDD = Knowledge Discovery in Databases

# We collect a lot of data



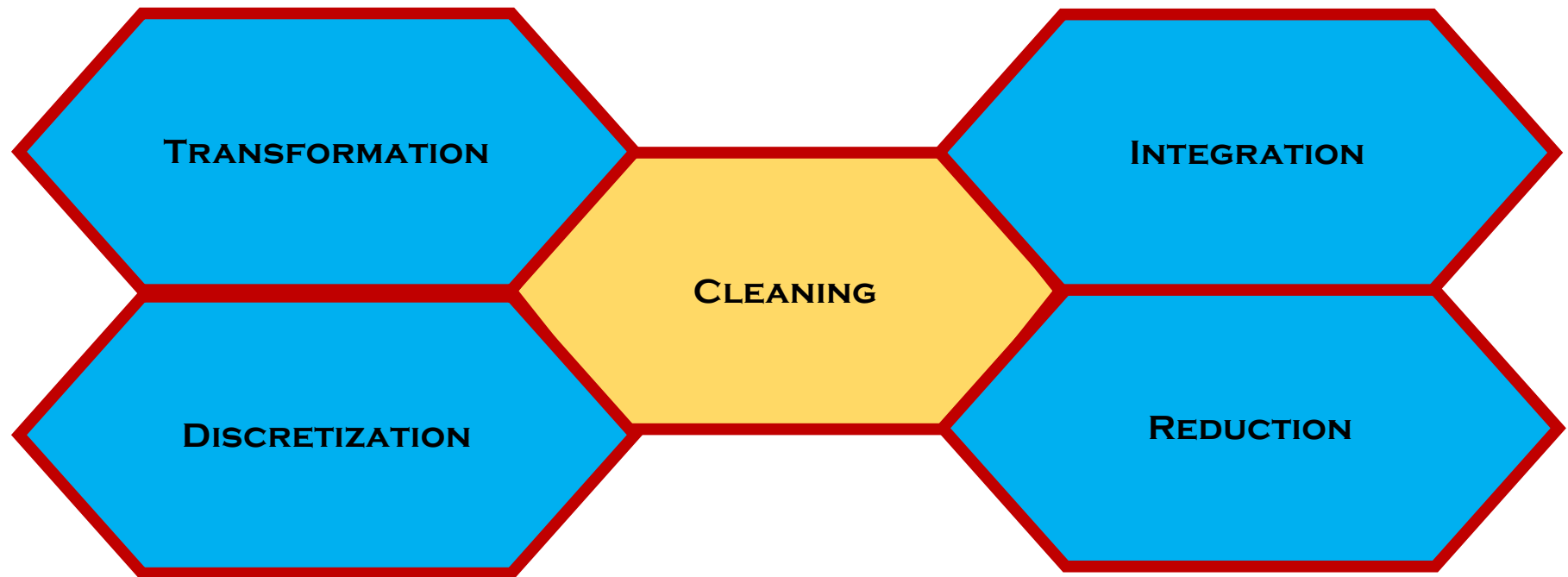
# Data quality issues



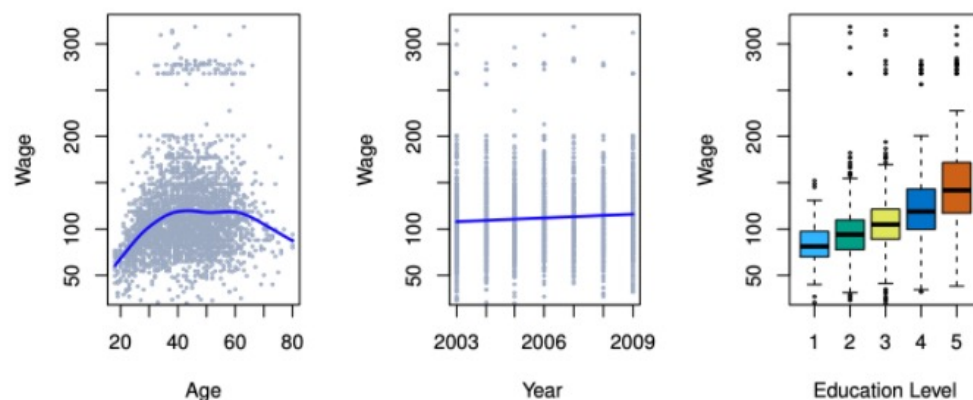
For any data retrieval or data analytics task,  
**it is critically important to know the quality of your data**

*For the US alone, poor data quality costs around US economy \$3 trillions a year (2016).*

# Data preparation



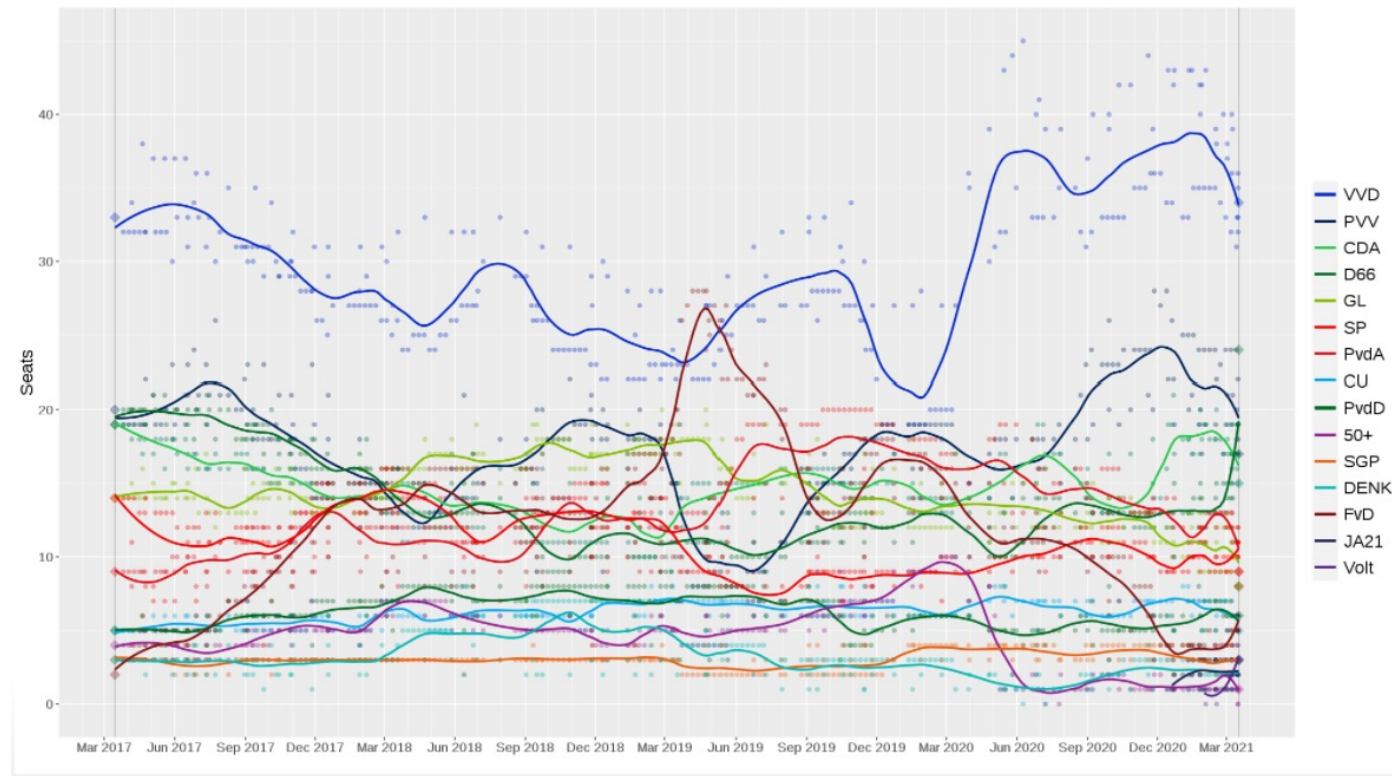
# Regression: How do wages differ?



**FIGURE 1.1.** Wage data, which contains income survey information for men from the central Atlantic region of the United States. Left: wage as a function of age. On average, wage increases with age until about 60 years of age, at which point it begins to decline. Center: wage as a function of year. There is a slow but steady increase of approximately \$10,000 in the average wage between 2003 and 2009. Right: Boxplots displaying wage as a function of education, with 1 indicating the lowest level (no high school diploma) and 5 the highest level (an advanced graduate degree). On average, wage increases with the level of education.



# Classification & visualisation



Predicting the outcome of the 2021 Dutch election:

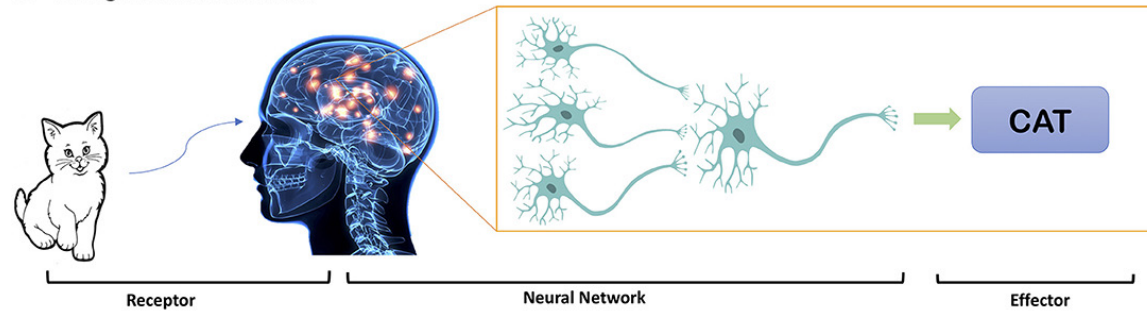
<https://gitlab.com/gbuvn1/opinion-polling-graph>



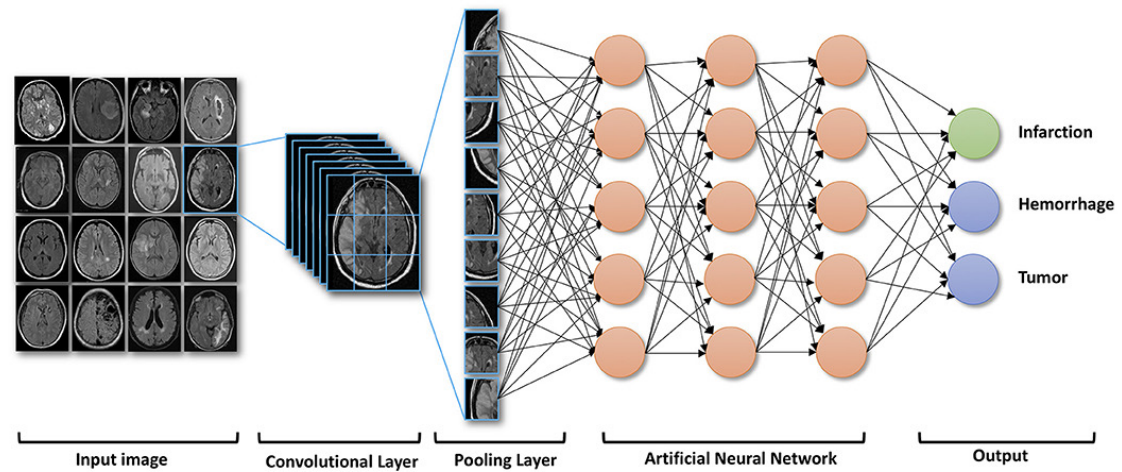
Utrecht University

# Deep learning

A Biological Neural Network



B Computer Neural Network(Convolutional Neural Network)



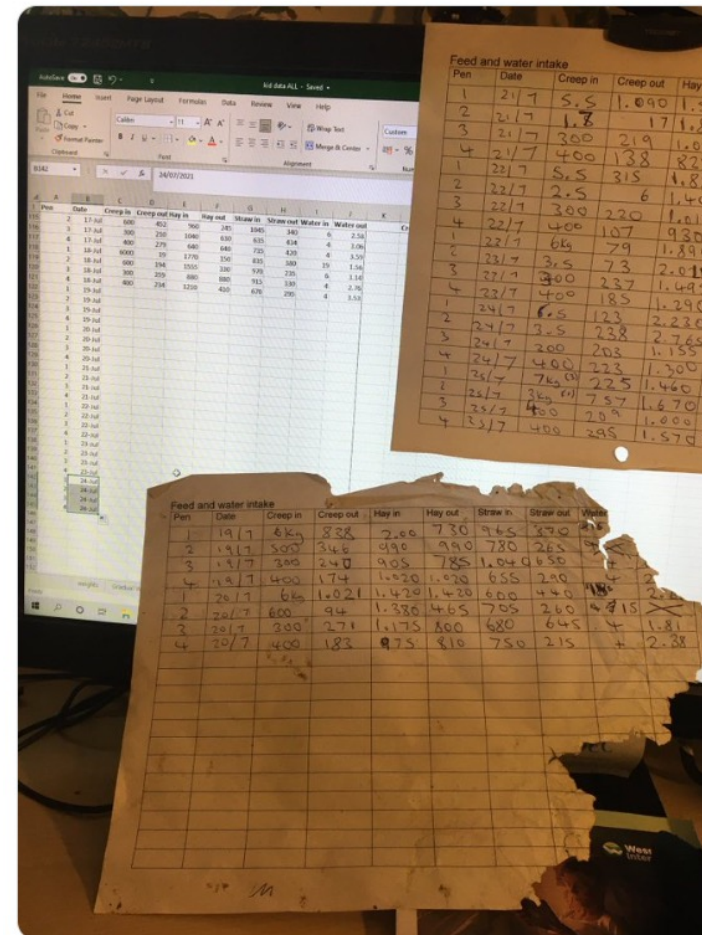
# Missing data

- Missing data mechanisms
- Imputation methods

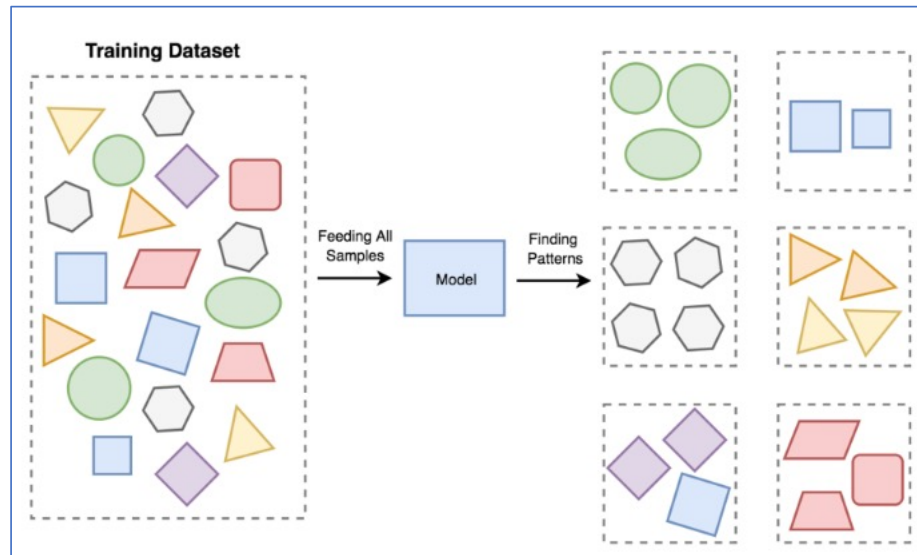


Holly Vickery PhD  
@SkylarkHolly

Working with goats 🐐🐐

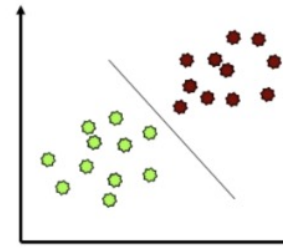


# Clustering



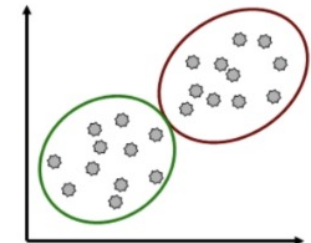
## CLASSIFICATION

- Labeled data points
- Want a "rule" that assigns labels to new points
- Supervised learning



## CLUSTERING

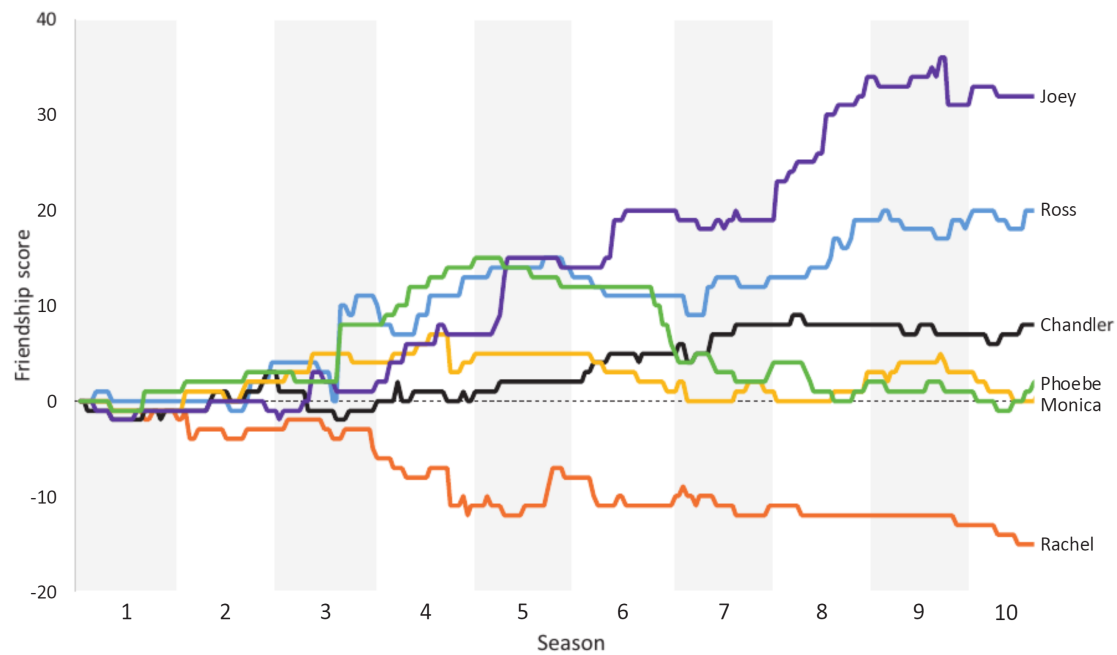
- Data is not labeled
- Group points that are "close" to each other
- Identify structure or patterns in data
- Unsupervised learning





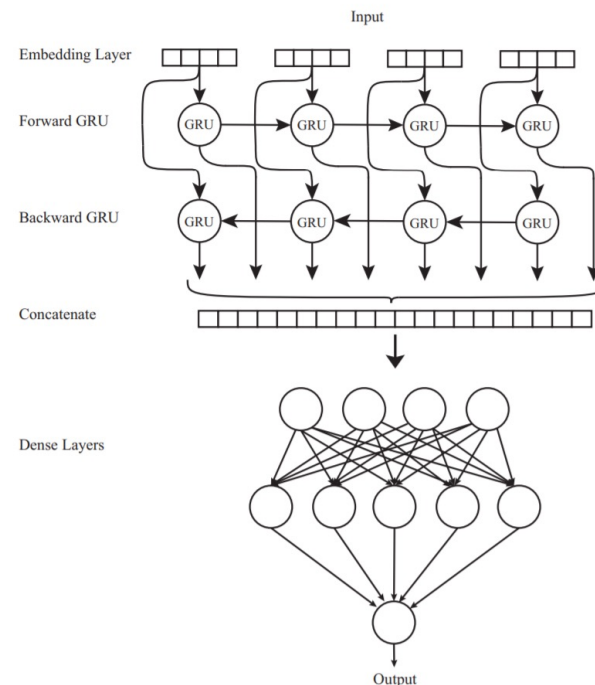
# Text Mining: Who was the best Friend?

<https://rss.onlinelibrary.wiley.com/doi/epdf/10.1111/1740-9713.01574>



# Natural language processing

<https://www.nature.com/articles/s41746-021-00404-9>



## Box 1: An example of a Dutch discharge letter from the dataset

Bovengenoemde patiënt was opgenomen op <DATUM-1> op de <PERSOON-1> voor het specialisme Cardiologie.

**Reden van opname** STEMI inferior

**Cardiale voorgeschiedenis.** Blanco

**Cardiovasculaire risicofactoren:** Roken(-) Diabetes(-) Hypertensie(?) Hypercholesterolemie (?)

**Anamnese.** Om 18.30 pijn op de borst met uitstraling naar de linkerarm, zweten, misselijk. Ambulance gebeld en bij aansluiten monitor beeld van acuut onderwandinfarct.

AMBU overdracht. 500 mg aspegic iv, ticagrelor 180 mg oraal, heparine, zofran eenmalig, 3x NTG spray. HD stabiel gebleven. Medicatie bij presentatie. Geen.

**Lichamelijk onderzoek.** Grauw, vegetatief, Halsvenen niet gestuwd. Cor s1 s2 geen souffles. Pulm schoon. Extr warm en slank.

**Aanvullend onderzoek.** AMBU ECG: Sinusritme, STEMI inferior III/II C/vermoedelijk RCA.

Coronair angiografie. (...). Conclusie angio: 1-vatslijden..PCI

### Conclusie en beleid

Bovengenoemde <LEEFTIJD-1>jarige man, blanco cardiale voorgeschiedenis, werd gepresenteerd vanwege een STEMI inferior waarvoor een spoed PCI werd verricht van de mid-RCA. Er bestaan geen relevante nevenletsels. Hij kon na de procedure worden overgeplaatst naar de CCU van het <INSTELLING-2>...Dank voor de snelle overname...Medicatie bij overplaatsing. Acetylsalicylzuur dispersietablet 80 mg; oraal; 1x per dag 80 milligram; <DATUM-1>. Ticagrelor tablet 90 mg; oraal; 2x per dag 90 milligram; <DATUM-1>. Metoprolol tablet 50 mg; oraal; 2x per dag 25 milligram; <DATUM-1>. Atorvastatine tablet 40 mg (als ca-zout-3-water); oraal; 1x per dag 40 milligram; <DATUM-1>

### Samenvatting

Hoofddiagnose: STEMI inferior vv PCI RCA. Geen nevenletsels. Nevendiagnoses: geen.

Complicaties: geen Ontslag naar: CCU <INSTELLING-2>.



# NLP: ChatGPT4o



Quiz me on  
world capitals



Message to  
comfort a friend



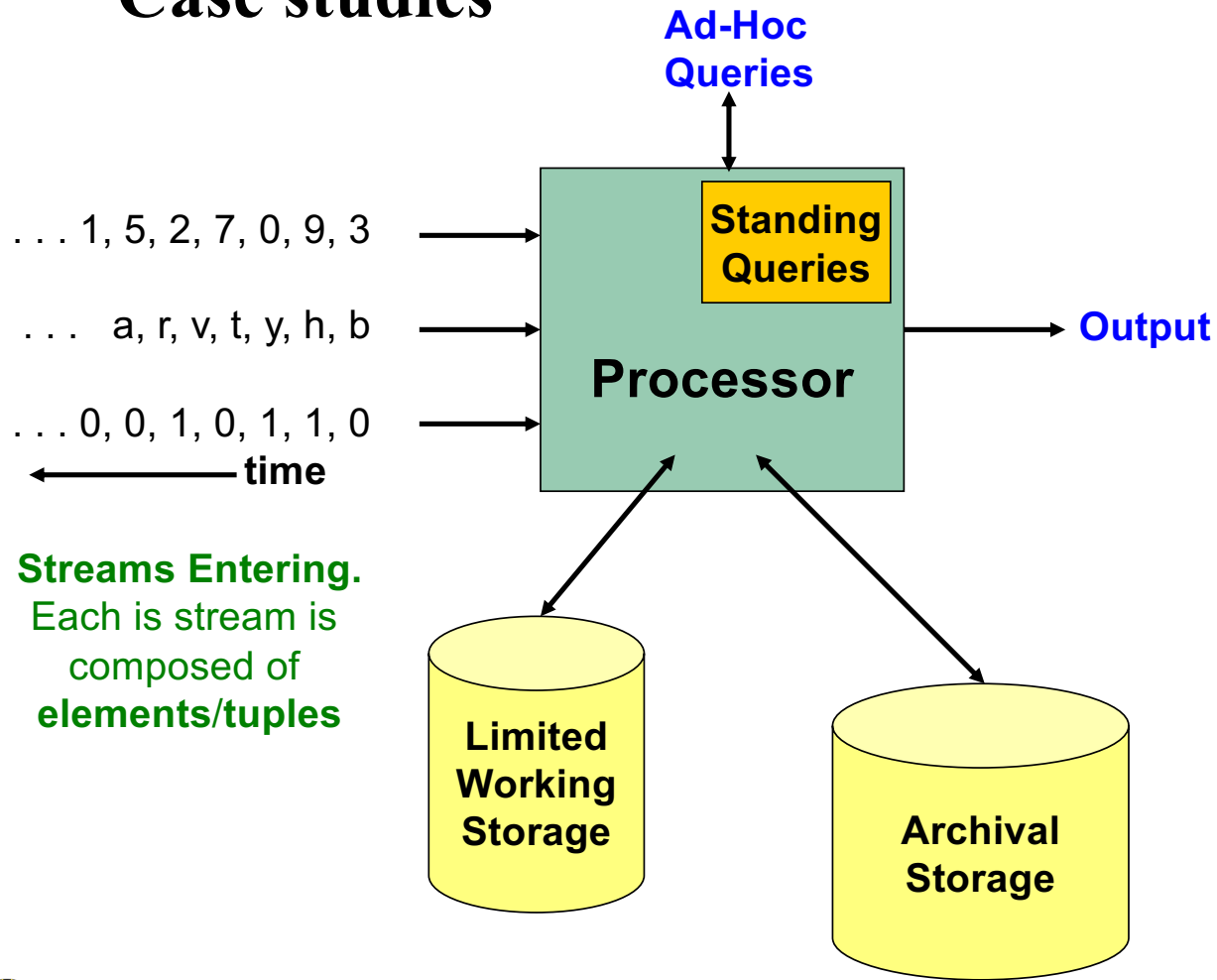
Activities to make  
friends in new city



Pick outfit to look  
good on camera



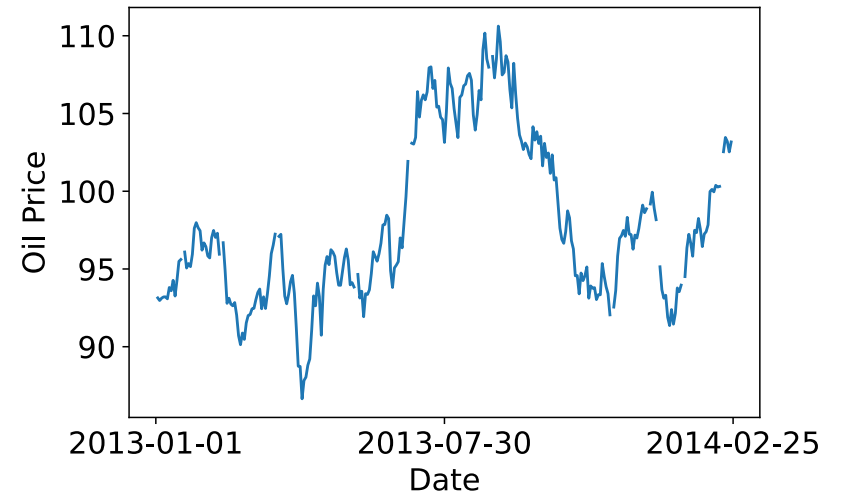
# Case studies



**Streams Entering.**  
Each is stream is  
composed of  
**elements/tuples**




**Algorithmic Fairness**









- 1 Go to [wooclap.com](https://wooclap.com)
- 2 Enter the event code in the top banner

Event code  
**INFOMDWR**

[app.wooclap.com/INFOMDWR](https://app.wooclap.com/INFOMDWR)

**10 minute break!**

# Practicalities

<https://infomdwr.nl>

# Practicalities

- Everything is on our course website [infomdwr.nl](https://infomdwr.nl), except:
  - Room locations (on [mytimetable.uu.nl](https://mytimetable.uu.nl))
  - Announcements (in your email and on [uu.brightspace.com](https://uu.brightspace.com))
  - Assignment hand-in (on [uu.brightspace.com](https://uu.brightspace.com))
  - Exams (on our digital assessment thing Remindo)
  - Lectures / Labs will be in person, one online lecture (on Teams)

 **materials on the website are constantly under construction** 

# Course flow

- This is a 14EC course. That's a lot!
- Every\* week you will:



Read the required readings



Attend three lectures: Monday, Tuesday, Wednesday morning



Attend and work on three lab sessions: Mon, Tue, Wed afternoon



Work on bi-weekly group assignments



Review materials diligently for the midterm and final exams

\* some weeks contain other stuff like exams and study time



# Schedule & readings

- The syllabus contains a full schedule with required readings

<https://infomdwr.nl/syllabus.html#required-readings>

- Reading materials can be found online for free! 🕵️ (you are resourceful students)

# Lectures: Prepare for tomorrow!

1. Look up reading materials in syllabus
2. Look up time and location for the lecture



# Next Monday: your **ONLY** online lecture

- The university did not have a room for us 🙄
- You will receive a Teams meeting invitation to follow this lecture online
- NB: the afternoon labs are in-person

# Labs

- Computer labs with practical exercises
- Put into practice what you learnt in the lectures
- We incorporate real-world data and use-cases: APPLIED data science!
- 4 different sessions for different groups
- **Lab teachers are your main point of contact for questions!**
- Labs provide skills needed to do the assignments

# Labs: prepare!

1. Quickly check what today's and tomorrow's labs are about
2. Look up time and location for the labs: [mytimetable.uu.nl](https://mytimetable.uu.nl)
3. Which parallel lab group are you assigned to? This should also be reflected in Brightspace

# Assignments

- Every two weeks
- Hand in on Brightspace
- Group work!
- **Plan ahead**



# Assignments: prepare!

1. What is the deadline of the first assignment? [infomdwr.nl](http://infomdwr.nl)
2. Quickly check the content of the first assignment
3. Log in to Brightspace and find the INFOMDWR course
4. If we were organized coordinators:
  1. Find the group sign-up page there to see how you can sign up
  2. Find the assignment hand-in page to see how you can submit
5. If we are disorganized coordinators, don't worry, it'll be there soon 😊





Read the syllabus

# Announcements



Utrecht University

# ADS OAC is looking for Student Members

- ADS Education Advisory Committee (OAC) is looking for three student members
  - One member (vice-chair) will be a member of GSNS OC
- OAC is composed of equal number of staff and student members.





# What will you do?

- You will:
  - Review the Caracal evaluations
  - Comment and give suggestions on the GSNS OER (Education and Examination Rules) and program-wise (ADS) OER Annex
  - Help in solving other education quality-related issues
- Expected time dedication
  - Weekly average: 2 hours for OAC members, 4 hours for vice-chair (also member of GSNS-OC)
  - You will be compensated for the time you spent



# How to apply?

- Interested?



# We need volunteers for the ADS Social Event



## DATA DRINKS

LOCATION: THE VAGANT BUILDING

DATE: TUESDAY SEPTEMBER 17 (15:00 – 20:00)

Come by for a drink and (hopefully) a piece of pizza, chat with your peers, instructors, and the program administration, and compete at table football



Utrecht University

The BNAIC conference is looking for volunteers: <https://bnaic2024.sites.uu.nl/>

A banner for the BNAIC/BeNeLearn 2024 conference. The background is a photograph of a historic building with Gothic architecture. Overlaid on the image is text and a logo. The text includes the conference title, location, and dates. The logo is a large white 'B' containing the text 'BNAIC 2024' and 'BENELEARN'. There are four yellow buttons: 'Call For Papers', 'Register Now', 'SUBMIT', and 'More info'.

Joint International Scientific Conferences on AI and Machine Learning

**BNAIC/BeNeLearn 2024**

Jaarbeursplein 6, 3521 AL Utrecht.  
November 18th, 19th and 20th, 2024

**Call For Papers** **Register Now**

**SUBMIT** **More info**

**BNAIC 2024**  
**BENELEARN**

Send an email to [h.t.schavemaker@uu.nl](mailto:h.t.schavemaker@uu.nl) if you are interested.



Utrecht University



# THE AI HELPDESK

## What is it?

- It is an **online platform** dedicated to providing clear, reliable **ANSWERS** to the public's **QUESTIONS ABOUT AI**.
- Individuals can **submit their questions** to the AI helpdesk, these questions will be addressed by **experts** in the field through clear and accessible responses.
- Funded by the Utrecht Young Academy

Website to launch Autumn 2024\*

Feel free to scan the QR code to **stay informed** and/or ask any **questions** you have about AI!!

\* First phase will be in Dutch



# Further questions?



**See you in the lab 🖐️**

