

Data Wrangling and Data Analysis

Hakim Qahtan

Erik-Jan van Kesteren

Ayoub Bagheri



Utrecht University

Today

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

Master Students

- Enroll for free
- Free coffee & cookies
- Lots of activities
- Master drinks

Website: <https://svsticky.nl/>

st!cky
studievereniging
informatica & informatiekunde

Who we are



Utrecht University

Beta faculty team

Duygu



Vahid



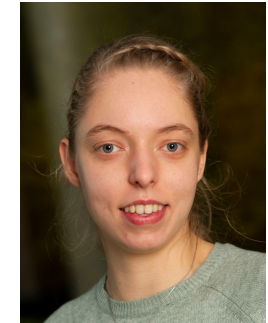
Hakim



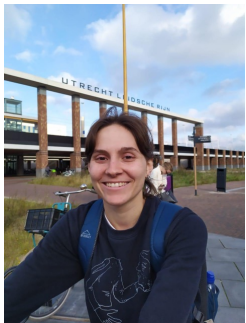
Daniela



Lisanne



Amalia



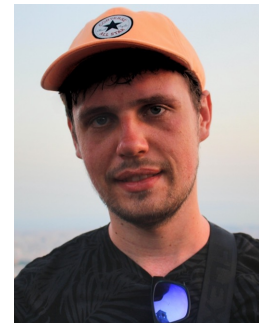
Maximilian



Kasper



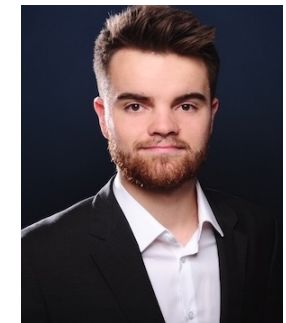
David



Dasja



Frederik



Social & Behavioural Science team

Erik-Jan



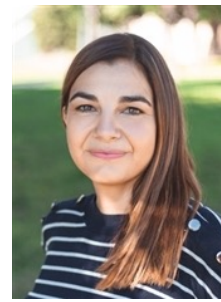
Ayoub



Daniel



Anastasia



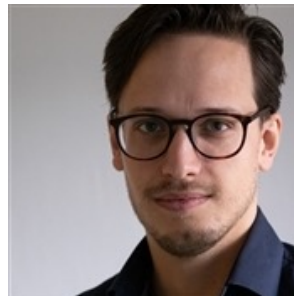
Pablo



Laura



Jelle



Javier



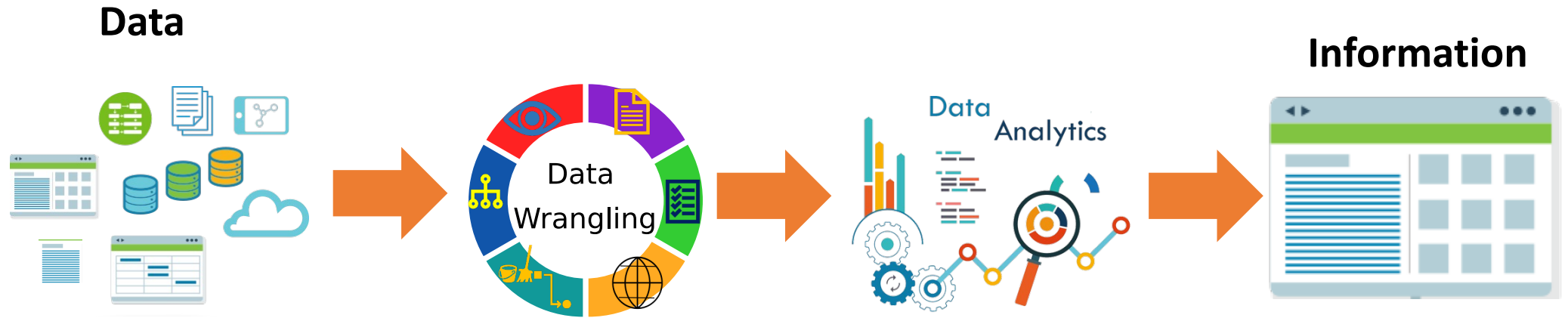
Mahdi



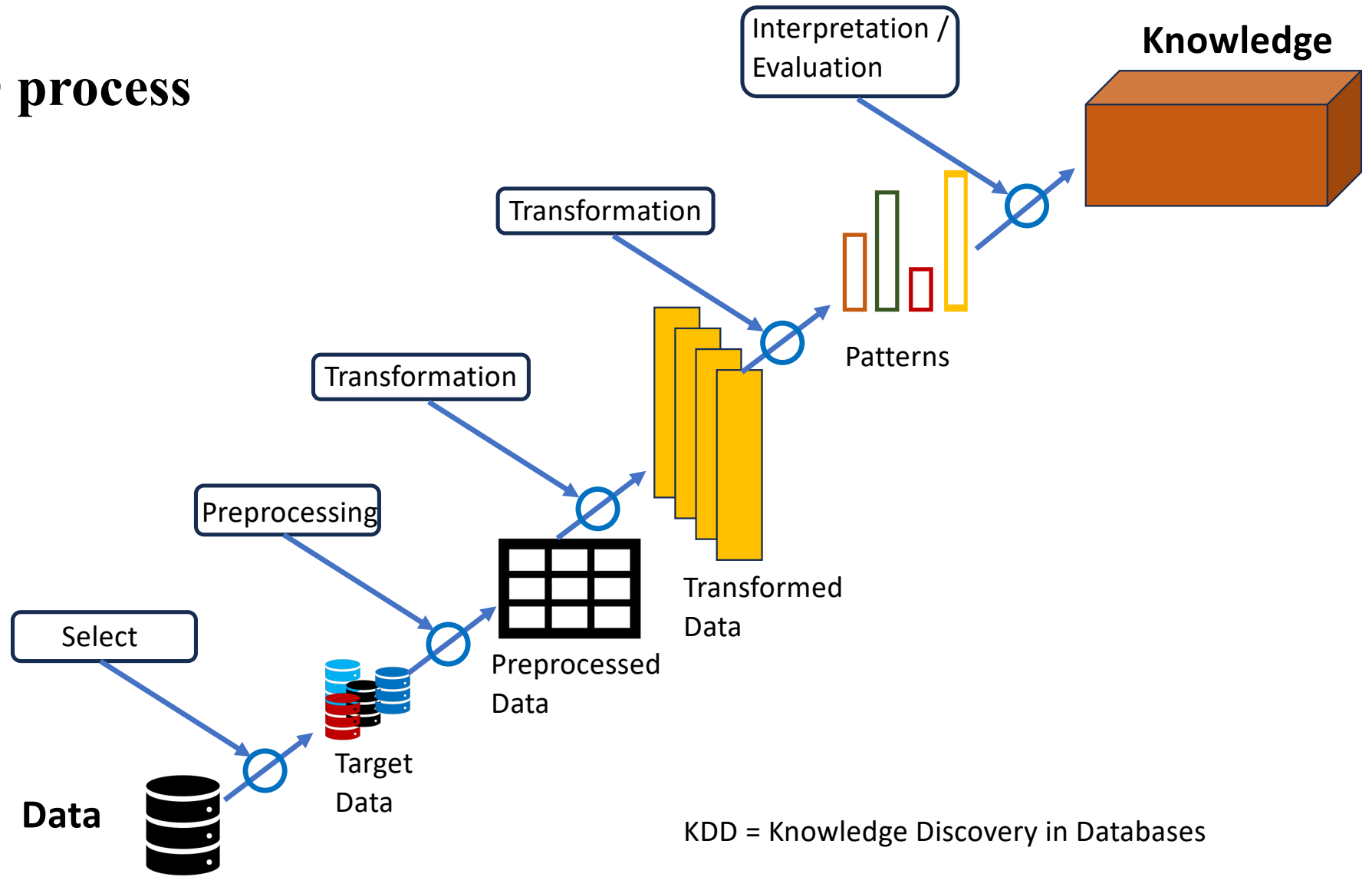
Introduction



Extracting value from data

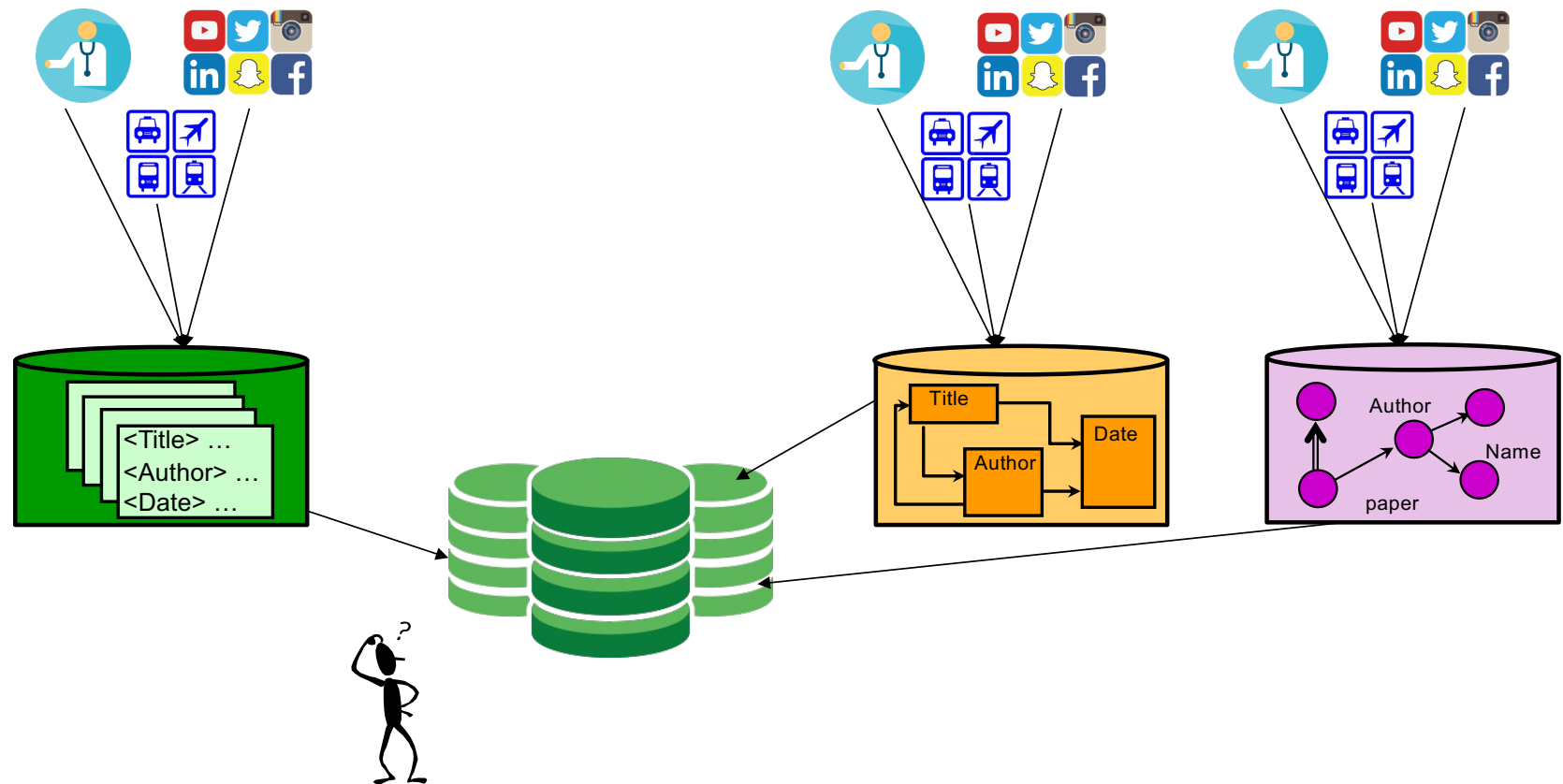


KDD process

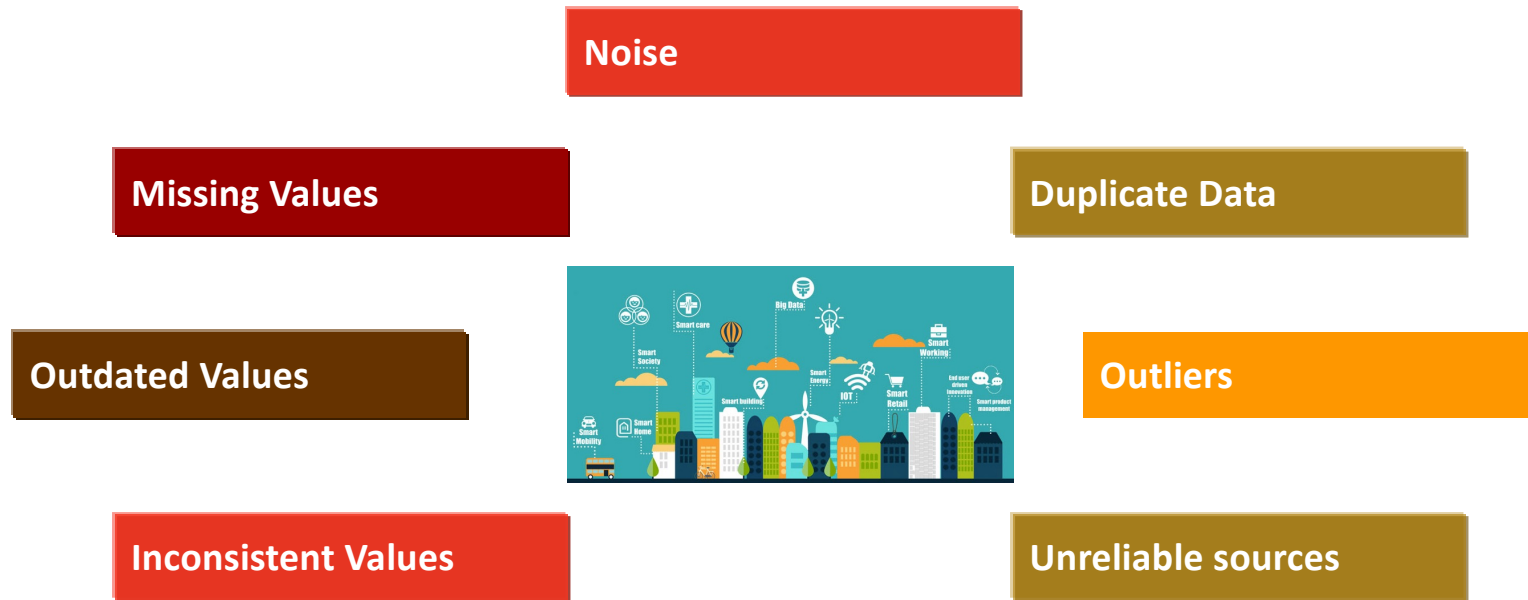


KDD = Knowledge Discovery in Databases

We collect a lot of data



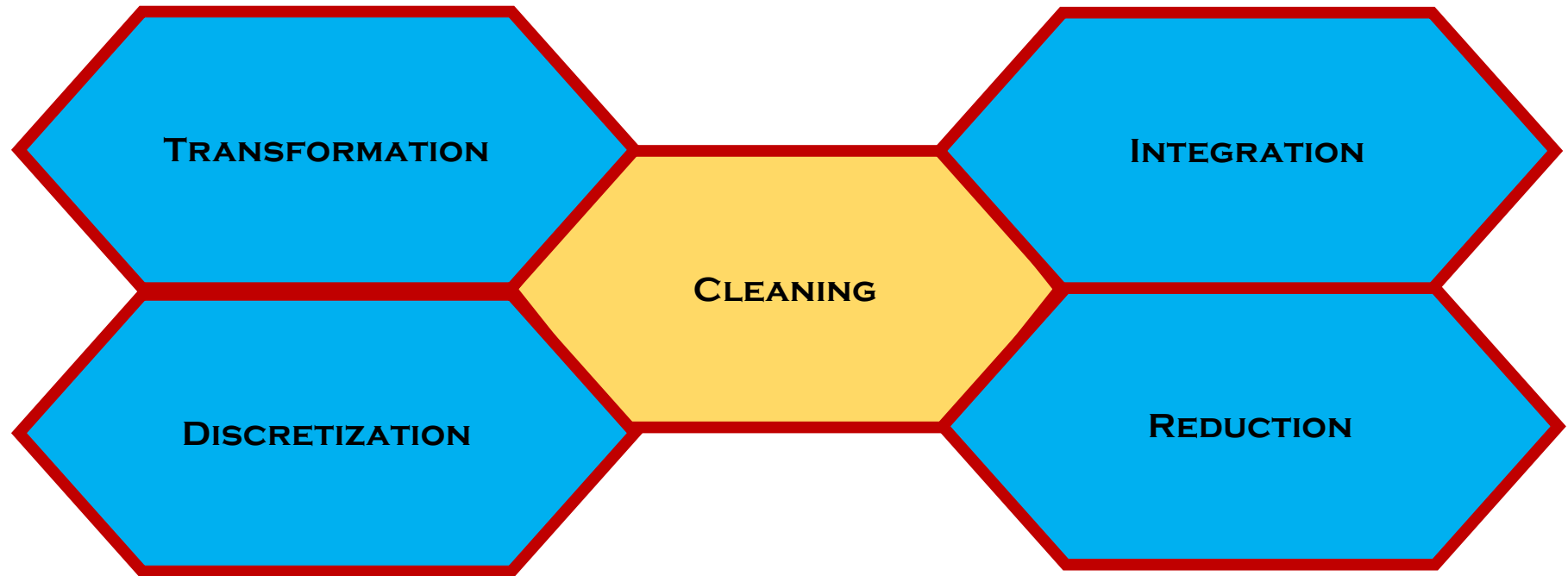
Data quality issues



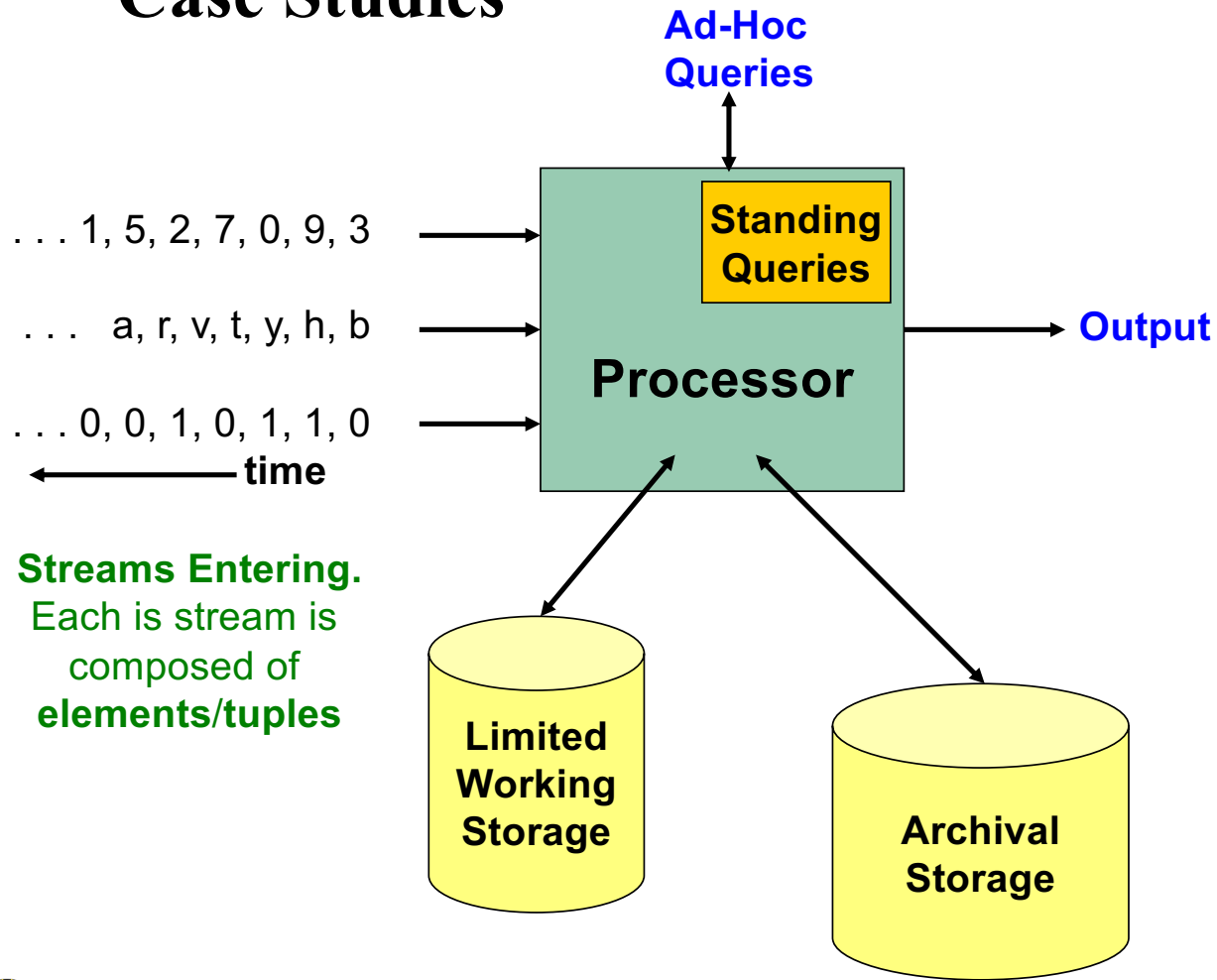
For any data retrieval or data analytics task,
it is critically important to know the quality of your data

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

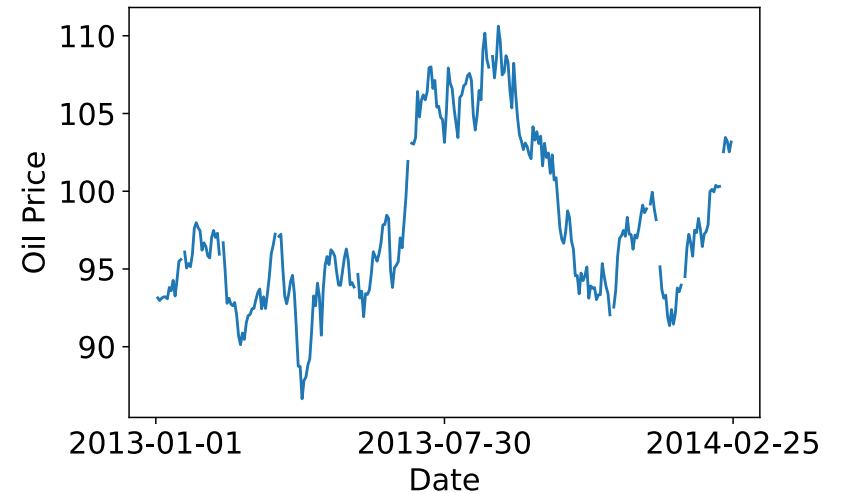
Data Preparation



Case Studies



Algorithmic Fairness



How do wages differ?

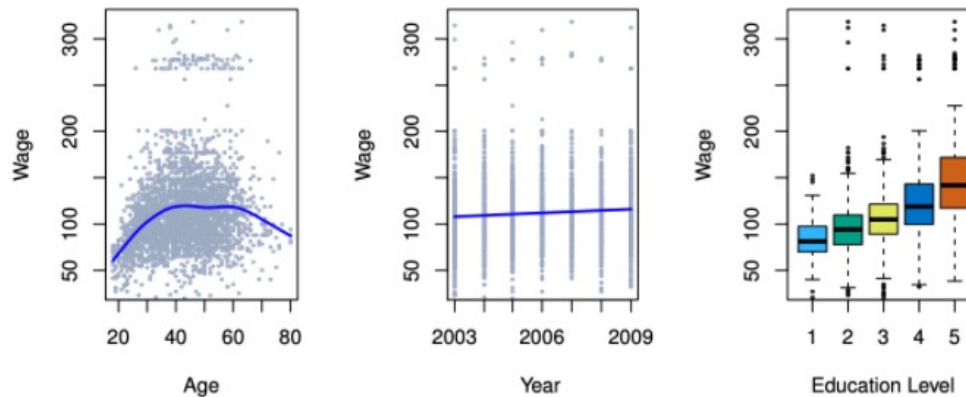
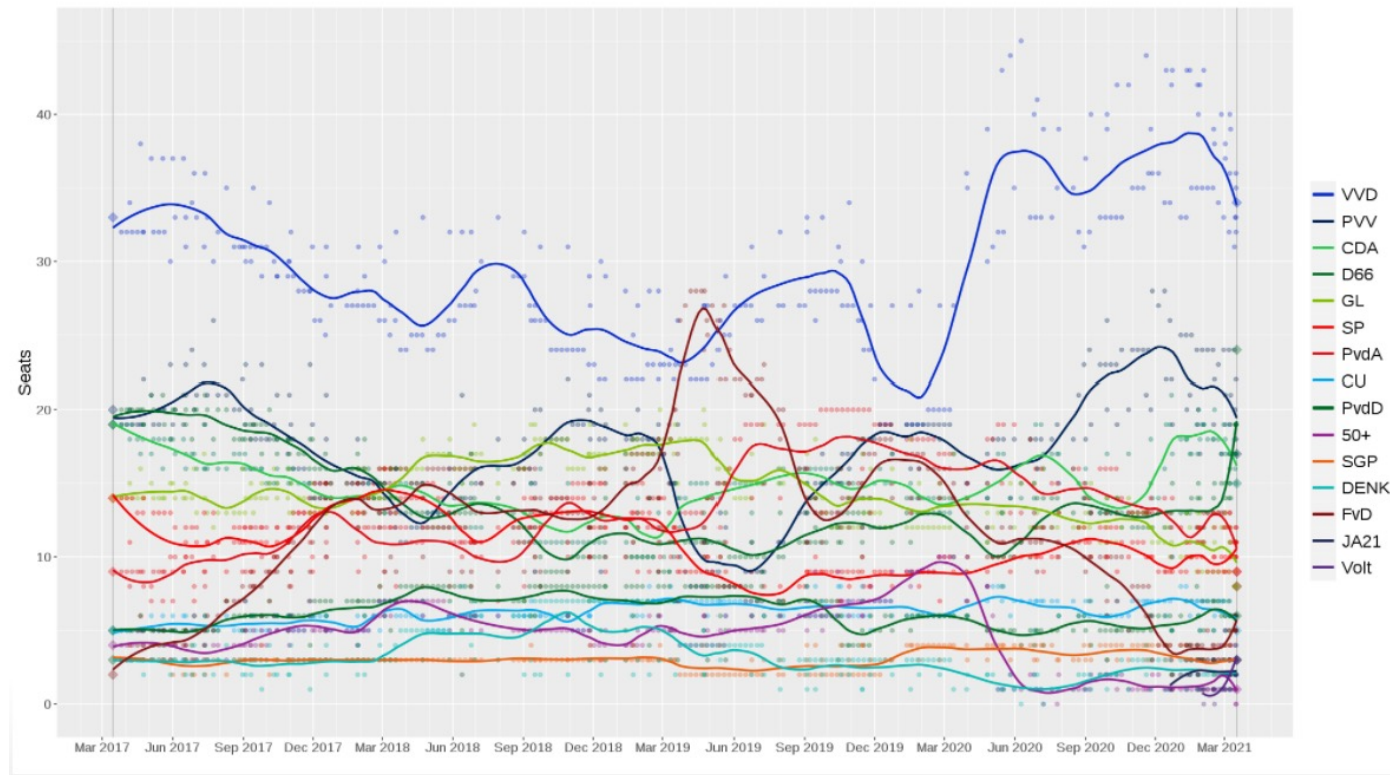


FIGURE 1.1. Wage data, which contains income survey information for men from the central Atlantic region of the United States. Left: wage as a function of age. On average, wage increases with age until about 60 years of age, at which point it begins to decline. Center: wage as a function of year. There 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.

Predicting the outcome of the 2021 Dutch election

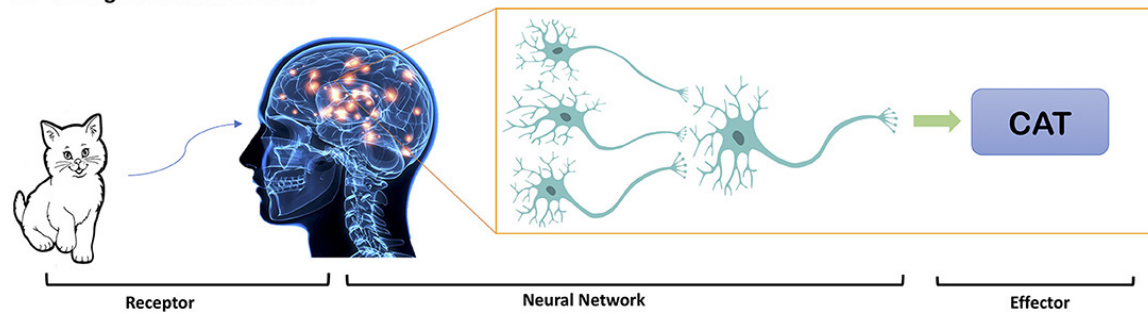


<https://gitlab.com/gbuvn1/opinion-polling-graph>

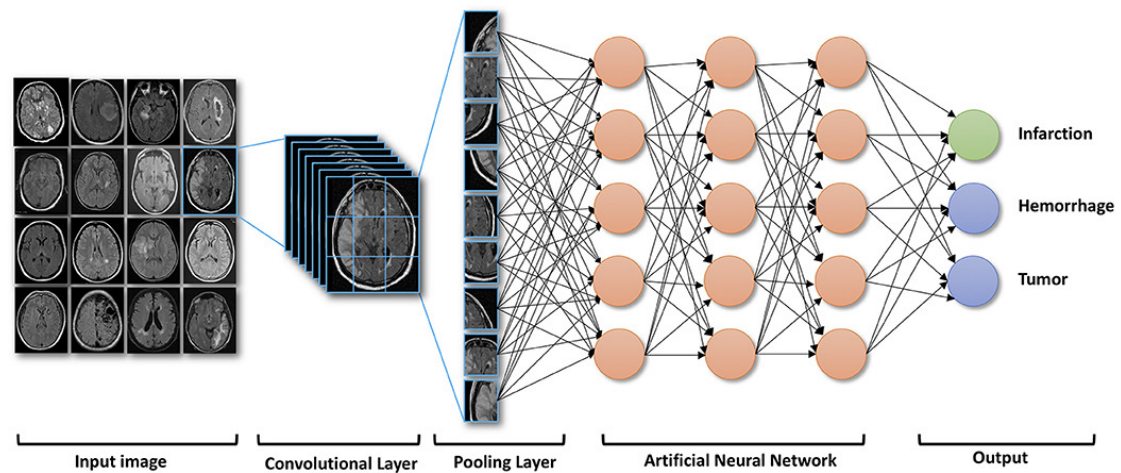


Deep Learning techniques

A Biological Neural Network

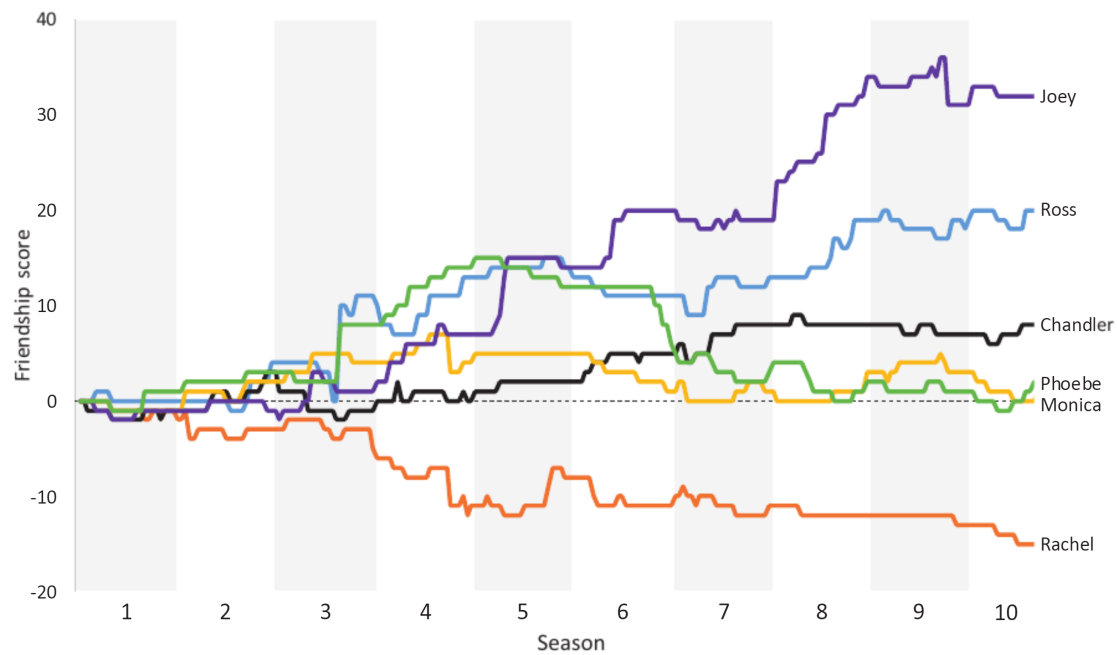


B Computer Neural Network(Convolutional Neural Network)



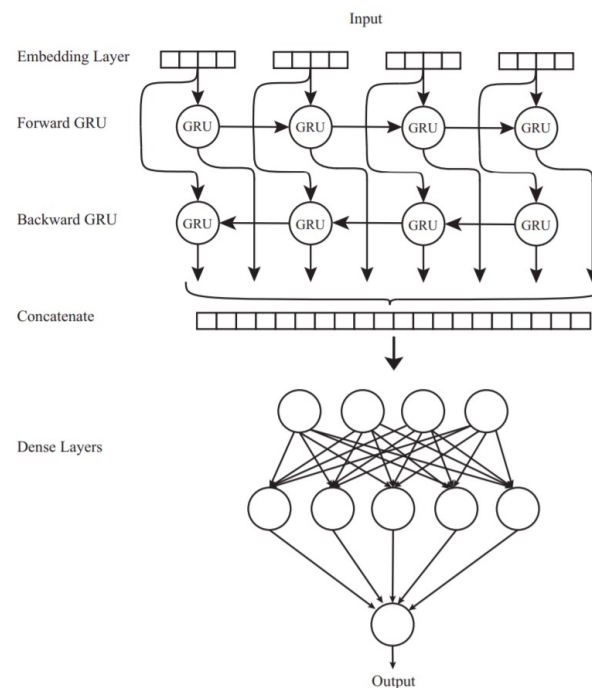
Who was the best Friend?

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



Automatic detection of ICD10 codes in cardiology discharge letters

<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 dispertablet 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 wv PCI RCA. Geen nevenletsels. Nevendiaagnoses: geen.

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





 **1** Go to wooclap.com

2 Enter the event code in the top banner

Event code
INFOMDWR

A yellow rounded rectangular box contains two numbered steps and an event code. The first step is accompanied by a globe icon. The event code 'INFOMDWR' is displayed in a large, bold font.

app.wooclap.com/INFOMDWR

10 minute break!

Practicalities

<https://infomdwr.nl>

Practicalities

- Everything is on our course website infomdwr.nl, except:
 - Room locations (on mytimetable.uu.nl)
 - Announcements (in your email and on uu.blackboard.com)
 - Assignment hand-in (on uu.blackboard.com)
 - Exams (on uu.blackboard.com)
 - Hybrid lectures (on Teams, more on this later)

 **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

Hybrid lectures

- The university did not have enough big rooms for us 😞
- Some lectures will be hybrid with 50-65 in-person spots
- You will receive a Teams meeting invitation to follow online (you will see this in your student email)
 - **ACCEPT** the invitation if you will join online
 - **REJECT** the invitation if you will join in-person
 - **If you're unsure, do neither**
- 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
3. Which parallel lab group are you assigned to?

Assignments

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

Assignments: prepare!

1. What is the deadline of the first assignment? infomdwr.nl
2. Quickly check the content of the first assignment
3. Log in to blackboard 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 were disorganized coordinators, don't worry, it'll be there soon 😊



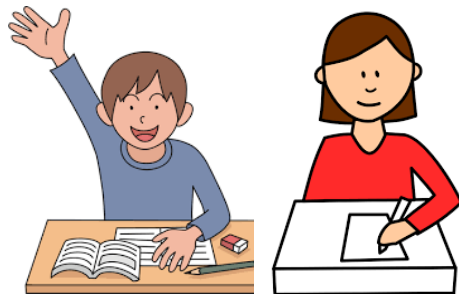
Read the syllabus

Student representatives



ADS OAC Student is Looking for Student Members

- ADS OAC committee is looking for two 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 OER and OER Annex
 - Help in solving other education quality-related issues

- Is it paid?
 - Yes, but not much



How to apply?

- Interested?



Further questions?

See you in the lab 🙌