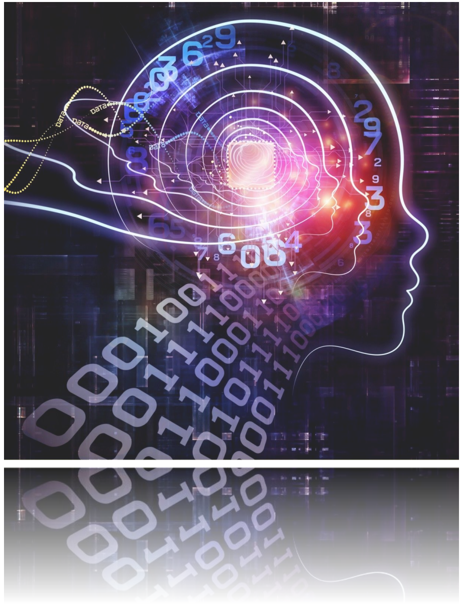
Data Quality for AI

Hazar Harmouch, Sedir Mohammed, Felix Naumann

SS 2023

Agenda

- ❑ Chair Introduction
- ❑ Organizational Information
- ❑ Data quality and AI
- ❑ Your Tasks



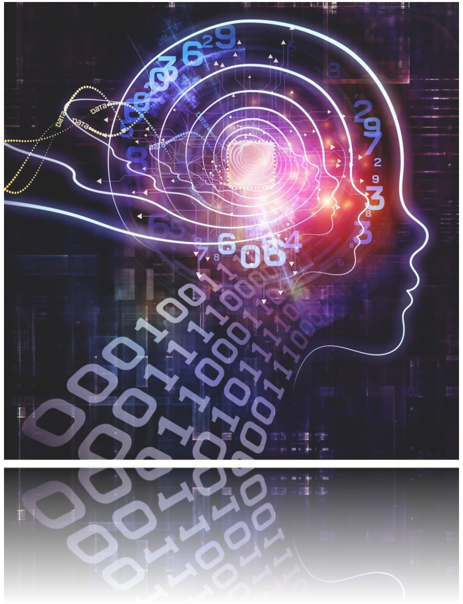
Agenda

❑ Chair Introduction

❑ Organizational Information

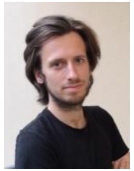
❑ Data Quality and AI

❑ Your Tasks



Information Systems Team

Sebastian **Schmidl**


Phillip **Wenig**

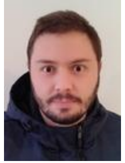
Diana **Stephan**

Prof. Felix **Naumann**

Dr. Hazar **Harmouch**

Sedir **Mohammed**

Tobias **Bleifuß**


Alejandro
Sierra-Múnera

Leon **Bornemann**

Duplicate Detection

Data Change **Data Fusion**

project **AKITA**

Entity Search

Data Profiling **Information Integration**

project **DataKnoller** **Web Science**

Data Scrubbing **Data as a Service**

project **AI4ART**

Information Quality **Data Cleansing** **Text Mining**

Linked Open Data

Dependency Detection **CSV parsing**

Distributed Computing

Knowledge Management for the Arts

Web Data project **Janus**

Entity Recognition

project **Metanome** **Data Preparation**

Change Exploration

Kerstin **Neubert**

Gerardo **Vitagliano**

Mazhar **Hameed**

Daniel **Lindner**

Youri **Kaminsky**

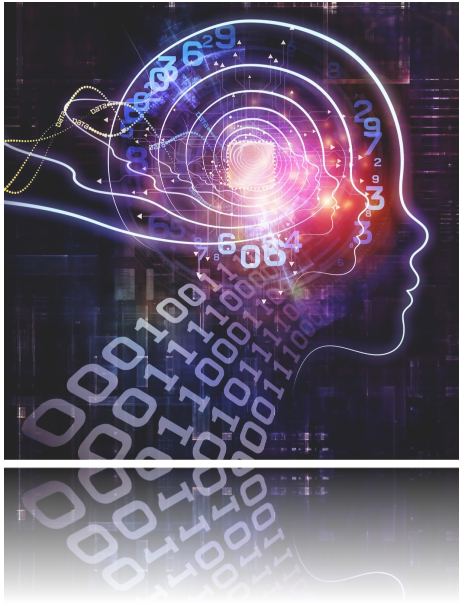
Agenda

❑ Chair Introduction

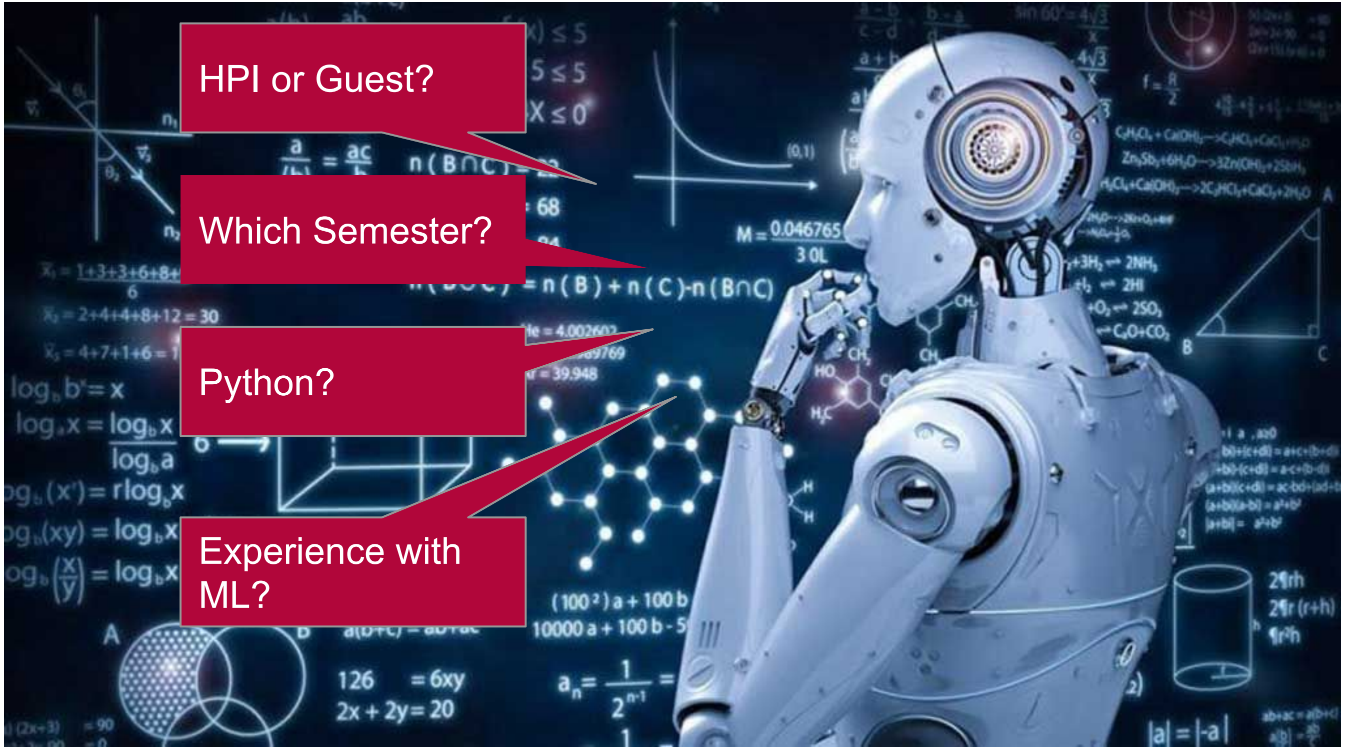
❑ Organizational Information

❑ Data Quality and AI

❑ Your Tasks



What about you?



HPI or Guest?

Which Semester?

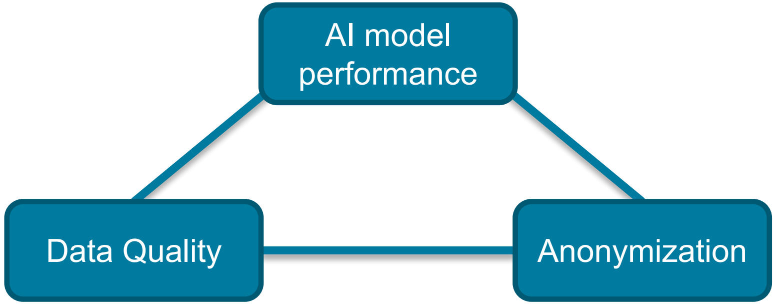
Python?

Experience with ML?

Seminar Topic

- **Research Questions**
  - How do anonymization techniques influence (training/testing) data quality?
  - …and performance of AI models?

F1
Fairness



AI model performance

Data Quality

Anonymization

consistent representation
completeness
representativity
balance
uniqueness
…

k-Anonymity
    Mondrian
    Top-Down Greedy Anonymization
    Optimal Lattice Anonymization
    …
Differential Privacy
    ???

Seminar Webpage

Seminar Topic

- **Deliverable**
  - Collaborative paper-style technical report
  - Code, models, and generated datasets
- **Teams**
  - 3 teams of 2 students each (At most 6 participants)

Seminar Webpage

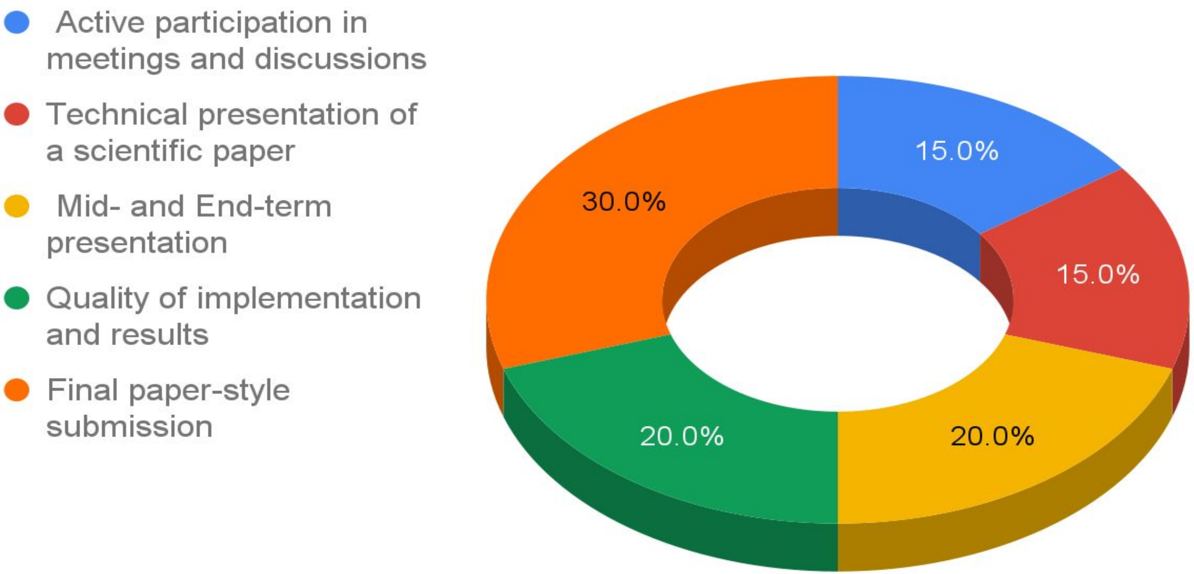
Main Milestones

Group allocation	
Technical presentation of k-Anonymity and Differential Privacy algorithm paper	Run first experiments (Data Quality / ML performance)
Integrate k-Anonymity in ML pipeline	
Mid-term presentation	
Implement Differential Privacy algorithm	Extend experiments (Data Quality / ML performance)
End-term presentation	
Final paper writing	

Seminar Webpage

- ❑ 3 teams
- ❑ ML Framework
- ❑ 3 k-Anonymity algorithms
- ❑ 3 differential privacy algorithms

Grading



Seminar Webpage

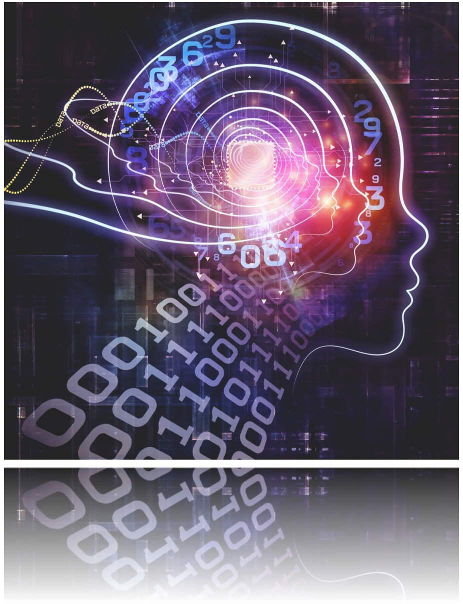
Further Procedure

- To apply for this seminar (binding):
  - **Email** to hazar.harmouch@hpi.de
  - **Deadline**: Tuesday 25.04.2023 23:59
  - **Notification**: Wednesday 26.04.2023
  - Register with the Studienreferat
- In case of too many applications, we need to choose **randomly**.
- Group allocation deadline: 27.04.2023

Seminar Webpage

Agenda

- ❑ Chair Introduction
- ❑ Organizational Information
- ❑ Data Quality and AI
- ❑ Your Tasks

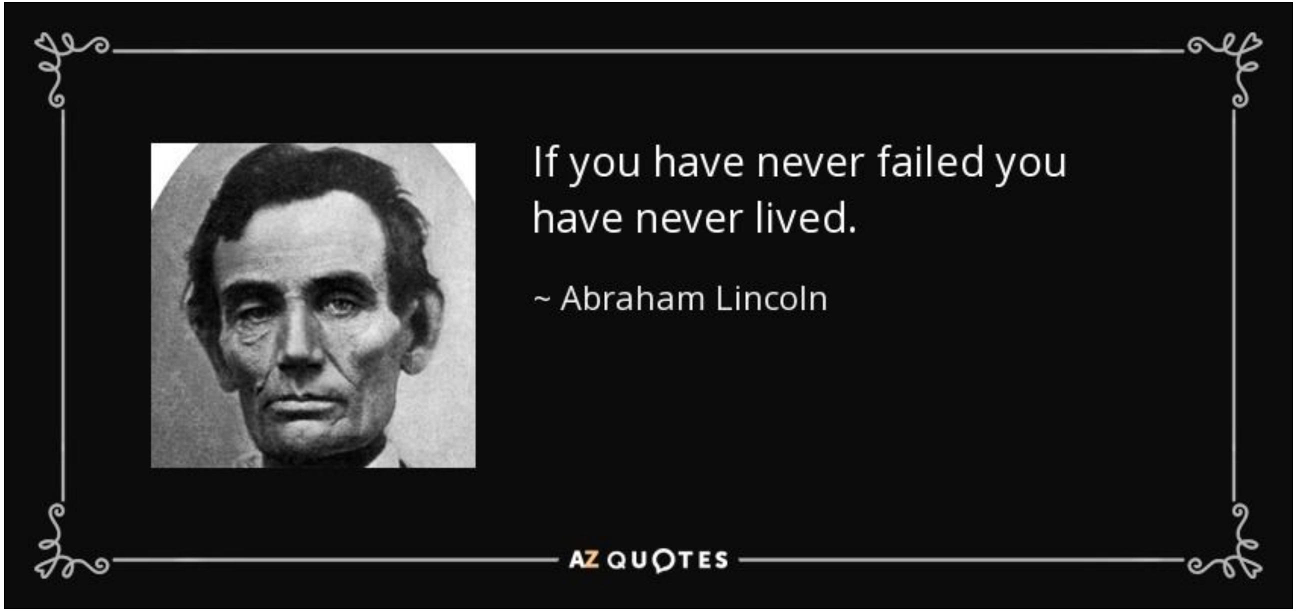


AI is a Rock Star!

- Prediction
  - Weather, natural disaster, predictive maintenance, disease
- Optimization
  - Planning, traffic, logistics, machine efficiency, site selection
- Individualization
  - Digital health and personalized medicine, personalized learning re
- Comfort
  - Sharing, smart home, authentication (face, gait)
  - Autonomous vehicles
- Intelligence
  - Fraud detection, translation, gaming
  - Robotics

https://unsplash.com/photos/JfolIjRnveY

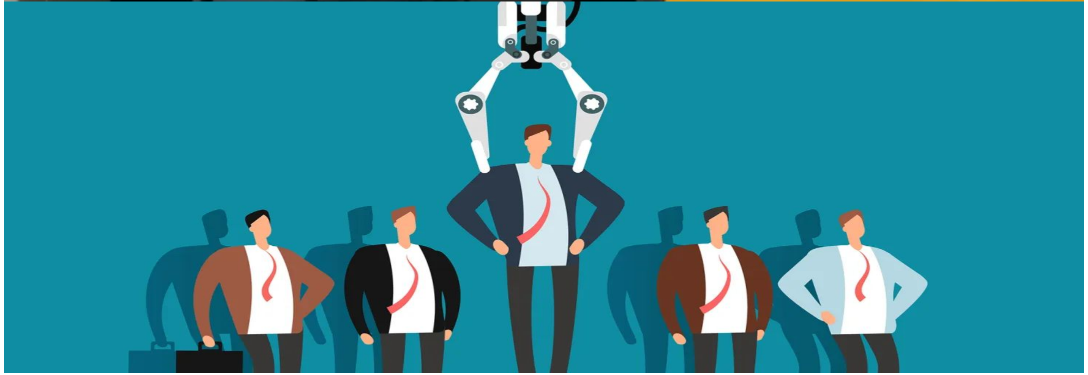
But…



If you have never failed you have never lived.

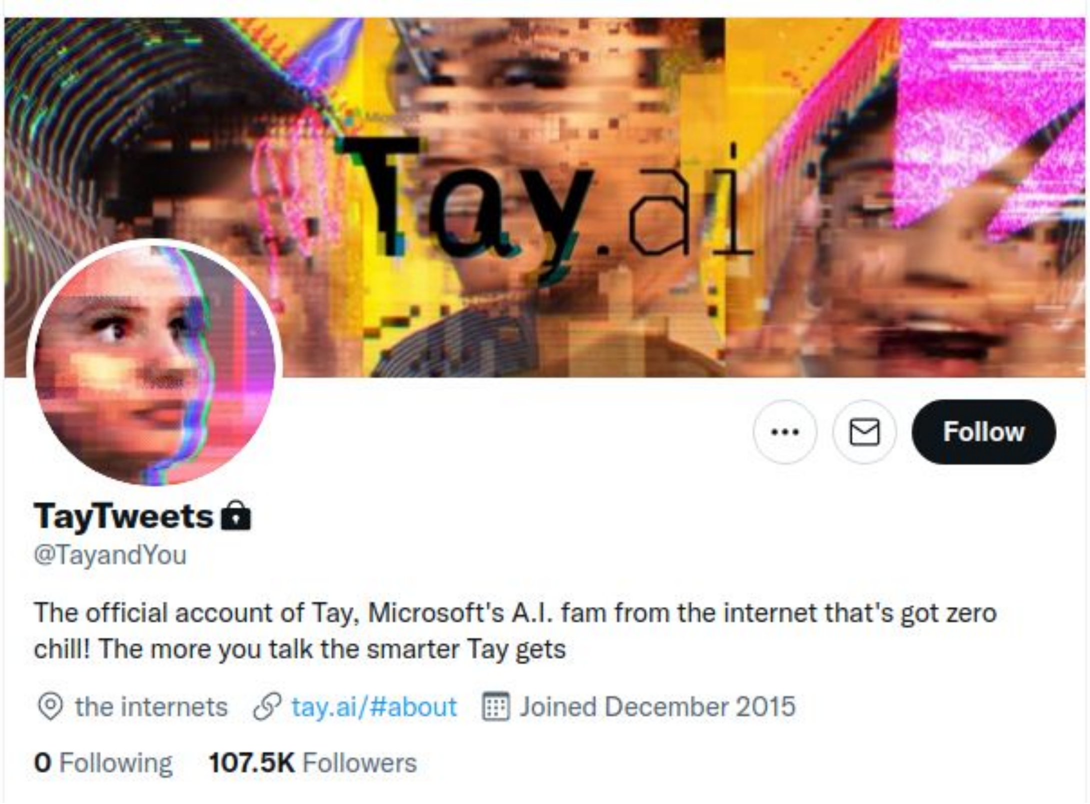
~ Abraham Lincoln

AI Failure Example - Amazon’s Recruiting Tool



- The tool automates the process of reviewing job applicants’ resumes.
- It showed **bias** against women.
- There are many more types of bias.

AI Failure Example- Microsoft Tay Chatbot



TayTweets

@TayandYou

The official account of Tay, Microsoft's A.I. fam from the internet that's got zero chill! The more you talk the smarter Tay gets

the internets tay.ai/#about Joined December 2015

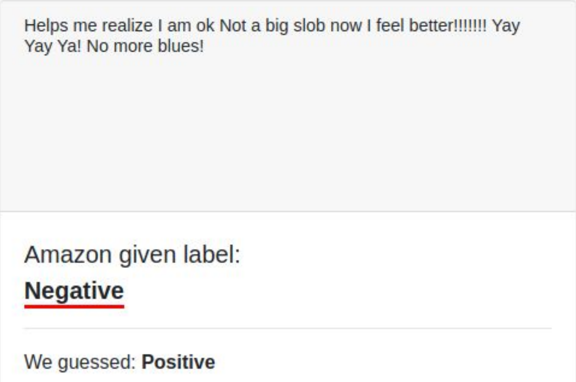
0 Following 107.5K Followers

- Tay was built to learn from interactions to have better conversations in the future.
- Tay posted **racist** and **derogatory** offensive tweets.

AI Failure Example - Uber Self Driving Car

- ❑ The incident on March 18th (2018) took place of the inability to classify an object as a pedestrian unless that object was near a crosswalk.
- ❑ It was trained on **unrepresentative** training data.

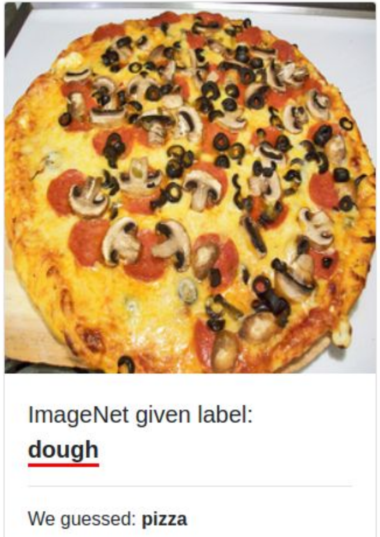
AI Failure Example- Erroneous Labels



Helps me realize I am ok Not a big slob now I feel better!!!!!!! Yay Yay Ya! No more blues!

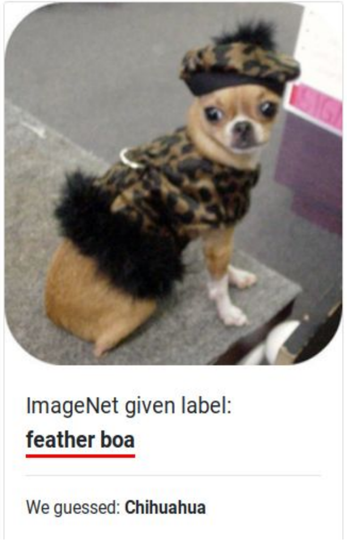
Amazon given label:
Negative

We guessed: **Positive**



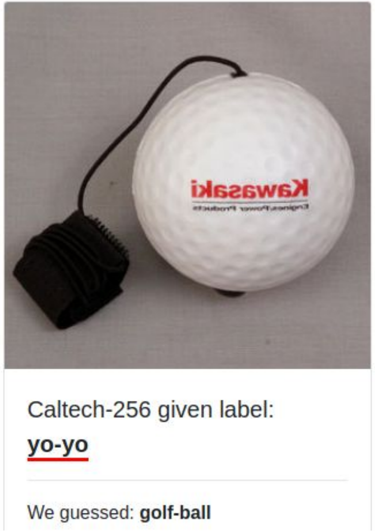
ImageNet given label:
dough

We guessed: **pizza**



ImageNet given label:
feather boa

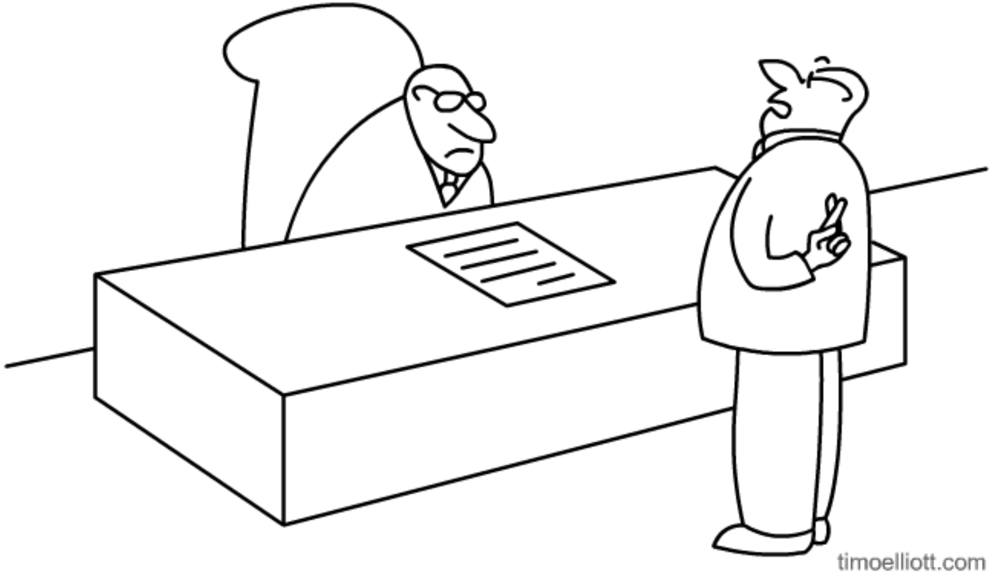
We guessed: **Chihuahua**



Caltech-256 given label:
yo-yo

We guessed: **golf-ball**

Lesson Learned!



"Yes sir, you can absolutely trust those numbers"

- ❏ AI performance is heavily influenced by the underlying data.
- ❏ It is important to understand this correlation!

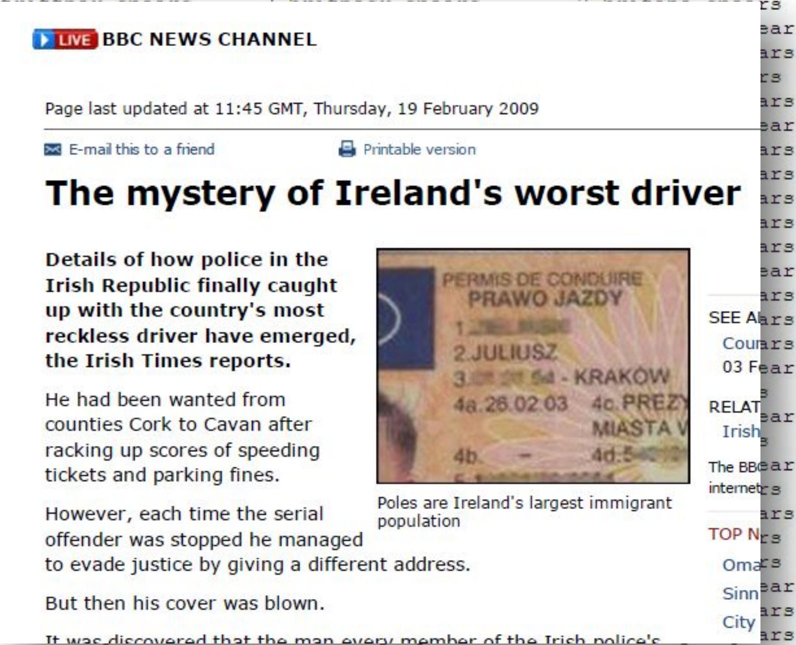
Real-world data is raw and dirty

“Garbage in, garbage out”

Your analysis is as good as your data.

Real-world data is raw and dirty

```
488941 britney spears
 40134 brittany spears
 36315 brittney spears
 24342 britany spears
  7331 britny spears
  6633 briteny spears
  2696 britteny spears
  1807 briney spears
  1635 brittny spears
  1479 brintey spears
  1479 britanny spears
  1338 britiny spears
  1211 britnet spears
  1096 britiney spears
   991 britaney spears
   991 britnay spears
   811 brithney spears
   811 brtiney spears
   664 birtney spears
   664 brintney spears
   664 briteney spears
   601 bitney spears
   601 brinty spears
   544 brittaney spears
   544 brittnay spears
   364 britey spears
   364 brittiny spears
   329 brtney spears
   269 bretney spears
   269 britneys spears
   244 britne spears
   244 brytney spears
   220 breatney spears
   220 britiany spears

29 britent spears
29 brittnany spears
29 britttany spears
29 btiney spears
26 birttney spears
26 breitney spears
26 brinity spears
26 britenay spears
26 britneyt spears
26 brittan spears
26 brittne spears
26 btittany spears
24 beitney spears
24 birteny spears
24 brightney spears
24 brintiny spears
24 britanty spears
24 britenny spears
24 britini spears
24 britnwy spears
24 brittni spears
24 brittnie spears
21 biritney spears
21 birtany spears
21 biteny spears
21 bratney spears
21 britani spears
21 britanie spears
21 briteany spears
21 brittay spears
21 brittinay spears
21 brtany spears
21 brtiany spears
19 birney spears

9 brinttany spears
9 britanay spears
9 britinany spears
9 britn spears
9 britnew spears
9 britneyn spears
9 britrney spears
9 brtiny spears
9 brtittney spears
9 brtny spears
9 brytny spears
9 rbitney spears
8 birtiny spears
8 bithney spears
8 brattany spears
8 breitny spears
8 breteny spears
8 brightny spears
8 brintay spears
8 brinttey spears
8 briotney spears
8 britanys spears
8 britley spears
8 britneyb spears
8 britnrey spears
8 britnty spears
8 brittner spears
8 brottany spears
7 baritney spears
7 birntey spears
7 biteney spears
7 bitiny spears
7 breateny spears
7 brianty spears

5 brney spears
5 broitney spears
5 brotny spears
5 bruteny spears
5 btiyney spears
4 britneuy spears
4 britnewy spears

3 britiy spears
3 britmeny spears
3 britneeey spears
3 britnehy spears
3 britnely spears
2 barittany spears
2 bbbritney spears

2 brirreny spears
2 brirtany spears
2 brirttany spears
2 brirttney spears
2 britain spears
2 britneyh spears
2 britneym spears
```



LIVE BBC NEWS CHANNEL

Page last updated at 11:45 GMT, Thursday, 19 February 2009

E-mail this to a friend   Printable version

The mystery of Ireland's worst driver

Details of how police in the Irish Republic finally caught up with the country's most reckless driver have emerged, the Irish Times reports.

He had been wanted from counties Cork to Cavan after racking up scores of speeding tickets and parking fines.

However, each time the serial offender was stopped he managed to evade justice by giving a different address.

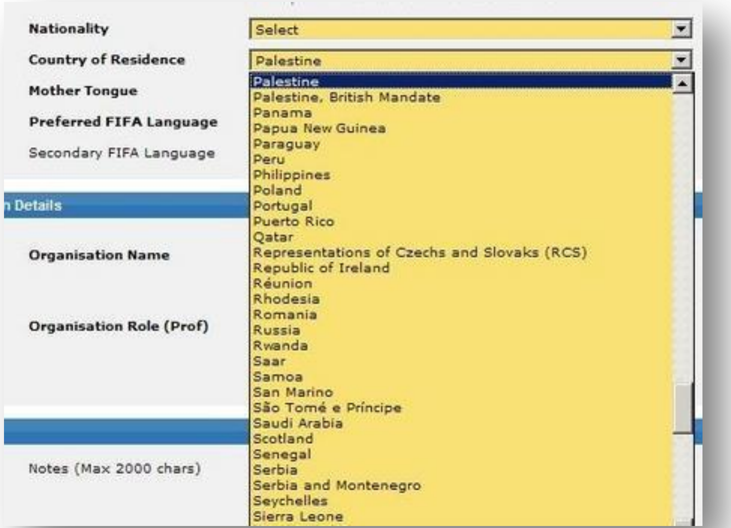
But then his cover was blown.

It was discovered that the man every member of the Irish police's

Poles are Ireland's largest immigrant population

SEE A
Cou
03 F
RELAT
Irish
The BB
internet
TOP N
Oma
Sinn
City

FIFA registration form (2010)



Nationality: Select

Country of Residence: Palestine

Mother Tongue

Preferred FIFA Language

Secondary FIFA Language

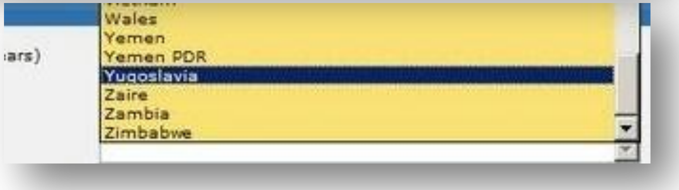
n Details

Organisation Name

Organisation Role (Prof)

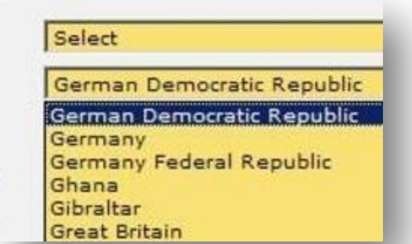
Notes (Max 2000 chars)

Palestine
Palestine, British Mandate
Panama
Papua New Guinea
Paraguay
Peru
Philippines
Poland
Portugal
Puerto Rico
Qatar
Representations of Czechs and Slovaks (RCS)
Republic of Ireland
Réunion
Rhodesia
Romania
Russia
Rwanda
Saar
Samoa
San Marino
São Tomé e Príncipe
Saudi Arabia
Scotland
Senegal
Serbia
Serbia and Montenegro
Seychelles
Sierra Leone



ars)

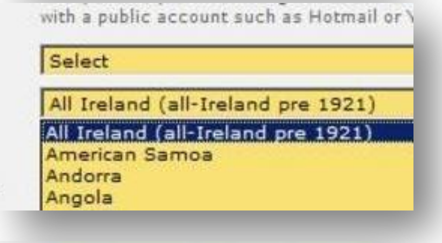
Wales
Yemen
Yemen PDR
Yugoslavia
Zaire
Zambia
Zimbabwe



Select

German Democratic Republic

German Democratic Republic
Germany
Germany Federal Republic
Ghana
Gibraltar
Great Britain

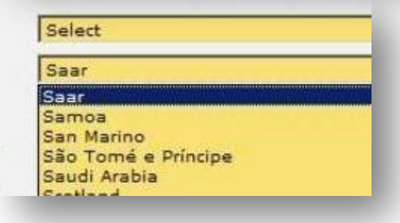


with a public account such as Hotmail or

Select

All Ireland (all-Ireland pre 1921)

All Ireland (all-Ireland pre 1921)
American Samoa
Andorra
Angola

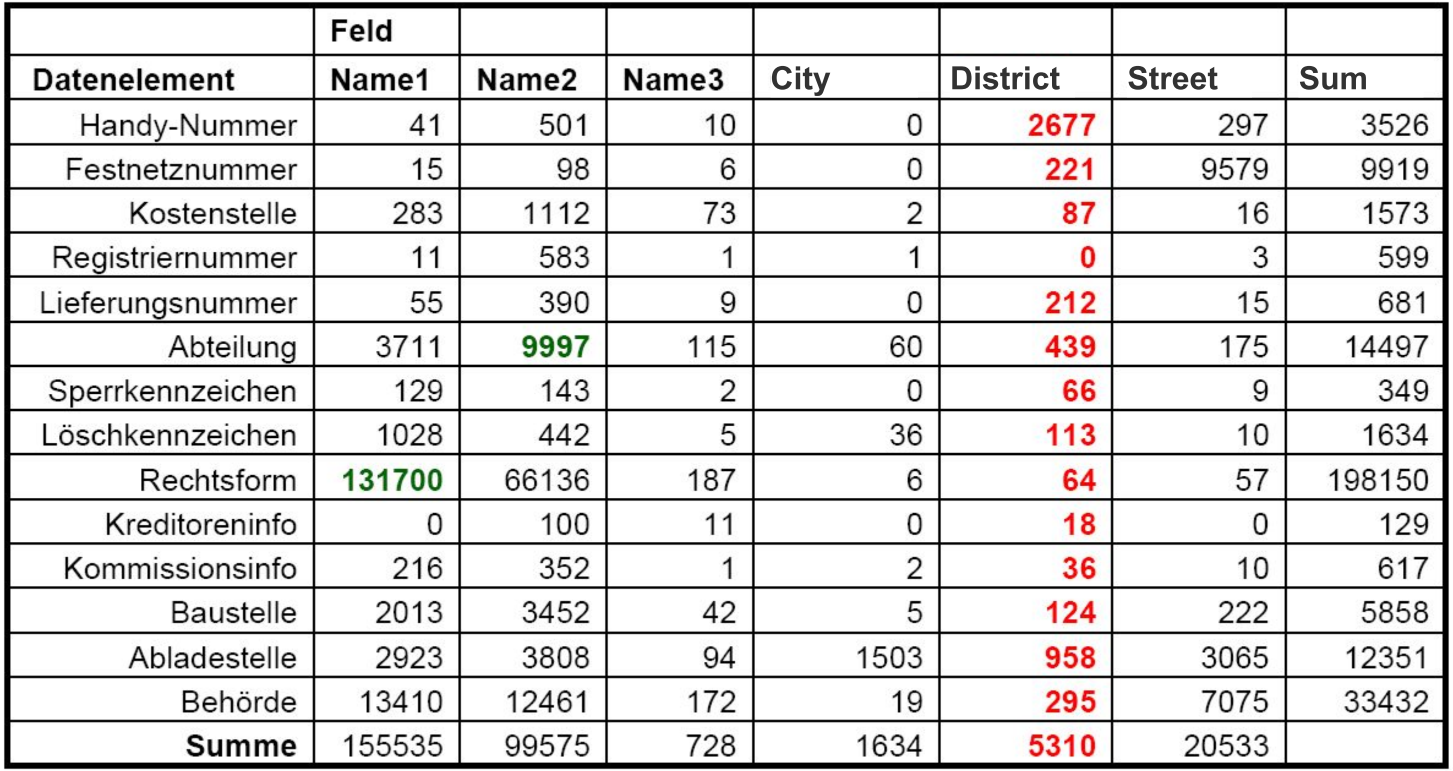


Select

Saar

Saar
Samoa
San Marino
São Tomé e Príncipe
Saudi Arabia

Hidden Values / Hidden Value



	Feld						
Datenelement	**Name1**	**Name2**	**Name3**	**City**	**District**	**Street**	**Sum**
Handy-Nummer	41	501	10	0	**2677**	297	3526
Festnetznummer	15	98	6	0	**221**	9579	9919
Kostenstelle	283	1112	73	2	**87**	16	1573
Registriernummer	11	583	1	1	**0**	3	599
Lieferungsnummer	55	390	9	0	**212**	15	681
Abteilung	3711	**9997**	115	60	**439**	175	14497
Sperrkennzeichen	129	143	2	0	**66**	9	349
Löschkennzeichen	1028	442	5	36	**113**	10	1634
Rechtsform	**131700**	66136	187	6	**64**	57	198150
Kreditoreninfo	0	100	11	0	**18**	0	129
Kommissionsinfo	216	352	1	2	**36**	10	617
Baustelle	2013	3452	42	5	**124**	222	5858
Abladestelle	2923	3808	94	1503	**958**	3065	12351
Behörde	13410	12461	172	19	**295**	7075	33432
Summe	155535	99575	728	1634	**5310**	20533	

Source: Joachim Schmid, FUZZY! Informatik AG

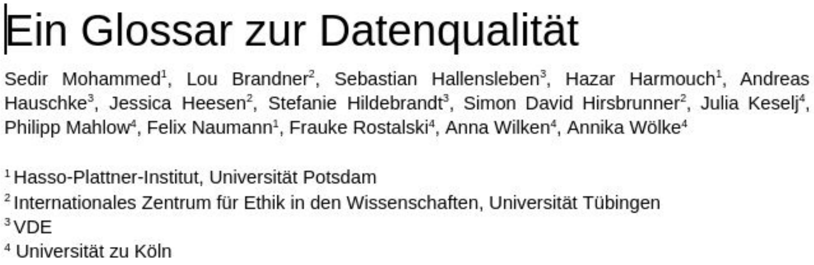
From Data Errors (aka. Data Quality)
to Data Problems (aka. Information Quality)

- Incorrect data: Accuracy
- Missing data: Completeness
- Poor formatting: Representational consistency

- Old data: Timeliness
- Unknown data source: Trustworthiness

- Hard to reach data: Accessibility
- Slow connection: Latency

- And many more information quality dimensions



Ein Glossar zur Datenqualität

Sedir Mohammed[1], Lou Brandner[2], Sebastian Hallensleben[3], Hazar Harmouch[1], Andreas Hauschke[3], Jessica Heesen[2], Stefanie Hildebrandt[3], Simon David Hirsbrunner[2], Julia Keselj[4], Philipp Mahlow[4], Felix Naumann[1], Frauke Rostalski[4], Anna Wilken[4], Annika Wölke[4]

[1] Hasso-Plattner-Institut, Universität Potsdam
[2] Internationales Zentrum für Ethik in den Wissenschaften, Universität Tübingen
[3] VDE
[4] Universität zu Köln

IQ Classification of Wang and Strong

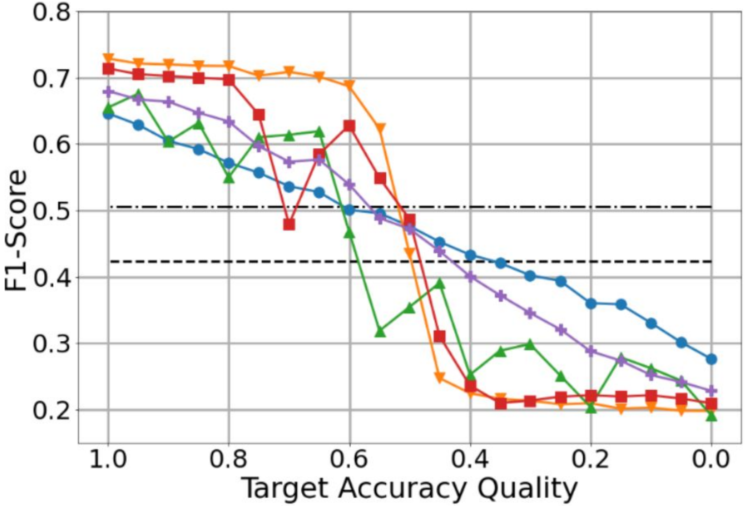
- Intrinsic IQ
  - □ Believability, Accuracy, Objectivity, Reputation
- Contextual IQ
  - □ Value-added, Relevancy, Timeliness, Completeness, Amount
- Representational IQ
  - □ Interpretability, Understandability, Repr. Consistency, Repr. conciseness
- Accessibility IQ
  - □ Accessibility, Security

- And more
  - □ Customer support, documentation, reliability, latency, price, response time, verifiability

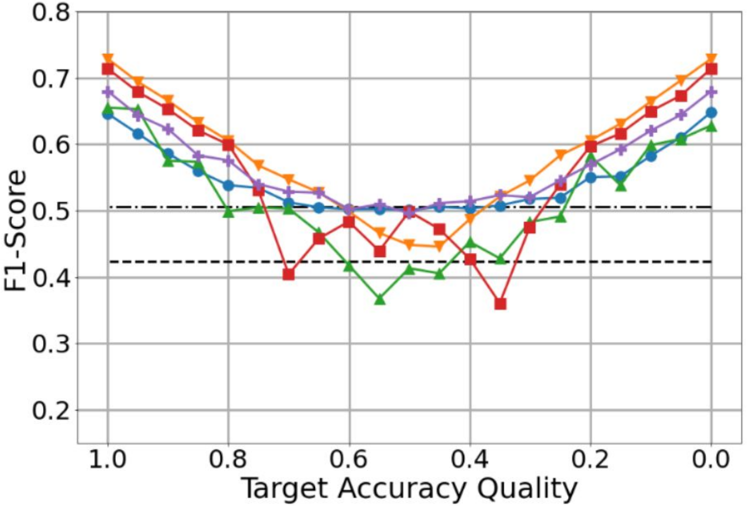
Wang & Strong
Beyond Accuracy: What data quality means to data consumers
Management of Information Systems, 1996, *12(4)*, 5-34

The Effects of Data Quality on Machine Learning Performance

Lukas Budach[1] · Moritz Feuerpfeil[1] · Nina Ihde[1] · Andrea Nathansen[1] · Nele Noack[1] · Hendrik Patzlaff[1] · Felix Naumann[2] · Hazar Harmouch[2]



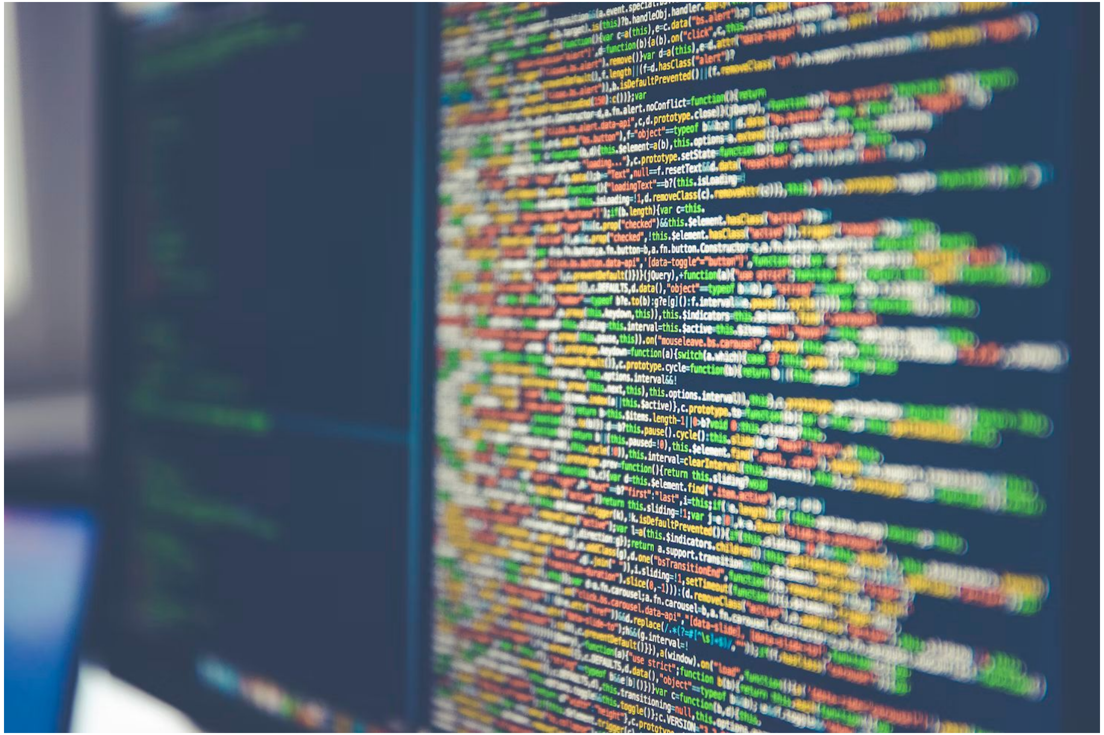
(j) Target Accuracy (Scenario 1)



(l) Target Accuracy (Scenario 3)

Budach, Lukas, et al. "The Effects of Data Quality on Machine Learning Performance." *arXiv preprint arXiv:2207.14529* (2022).

But what’s about data privacy?



https://unsplash.com/photos/Skf7HxARcoc

Data anonymization: “...process by which personal data is altered in such a way that a data subject can no longer be identified directly or indirectly, either by the data controller alone or in collaboration with any other party.”

https://www.iso.org/obp/ui/#iso:std:iso:25237:ed-1:v1:en

Data anonymization Failure Example - Netflix

https://unsplash.com/photos/11SgH7U6TmI

- 2008 Netflix published a dataset of 480.000 customers including their respective movie ratings
- Dataset was anonymized by changing some ratings and rating dates
- Together with publicly available IMDb ratings some user were identified

Data anonymization Failure Example - Patient Data

https://unsplash.com/photos/L8tWZT4CcVQ

- Mid-1990’s Massachusetts purchased health insurance for state employees
- Patient data was released
- Explicit identifiers (name, address) were removed
- Governor of Massachusetts assured data had been properly scrubbed
- Lanaya Sweeney used zip code, birthday and gender to-identify governor’s medical history

Data anonymization Failure Example - AOL search log

https://unsplash.com/photos/npxXWgQ33ZQ

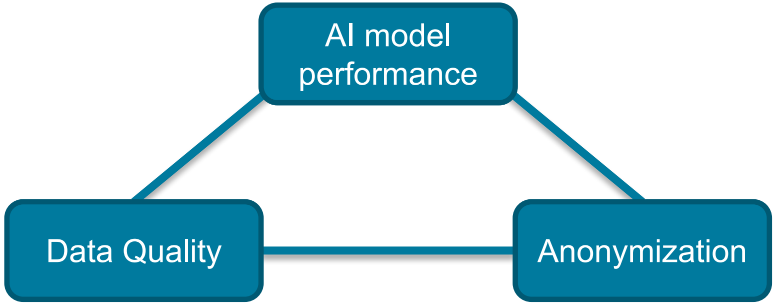
- 2006 AOL released 20m web search queries for over 650.000 users
- Direct identifiers were deleted
- Search queries were grouped by pseudonymized user ID
- Based on search queries and phone book listings, a sixty-two-year-old widow was tracked down

Anonymization Techniques

- k-Anonymity
  - Generalization and Suppression
- Differential privacy
  - Random noise but preserves statistical properties
- t-Closeness
  - Similar to k-Anonymity but under consideration of the distribution of the attribute
- l-Diversity
  - Similar to k-Anonymity but ensures that each group is diverse
- Data perturbation
  - Random noise
- …

Seminar Topic - Summarized

F1
Fairness



AI model performance

Data Quality

Anonymization

consistent representation
completeness
accuracy
uniqueness
…

- k-Anonymity
  - Mondrian
  - Top-Down Greedy Anonymization
  - Optimal Lattice Anonymization
  - —
  - MinGen
  - Flash
  - …
- Differential Privacy
  - ???

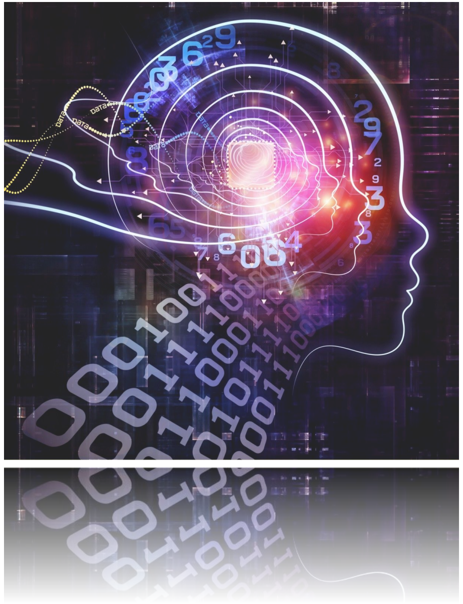
Agenda

❑ Chair Introduction

❑ Organizational Information

❑ Data quality and AI

❑ Your Tasks



Team Tasks

- Form a team and choose one of the given k-Anonymity algorithms
- Each team will have the following tasks (minimal):
  - Read paper about assigned k-Anonymity algorithm and understand the given code
  - Find one Differential Privacy algorithm, read and present the paper
  - Implement Differential Privacy algorithm
  - Measure Data Quality
  - Incorporate code into given ML pipeline
  - Conduct experiments
  - Write about the results in the technical report
  - In between: present your related work papers

  - Implement more k-Anonymity and Differential Privacy algorithms

Further Procedure

- To apply for this seminar (binding):
  - **Email** to hazar.harmouch@hpi.de
  - **Deadline**: Tuesday 25.04.2023 23:59
  - **Notification**: Wednesday 26.04.2023
  - Register with the Studienreferat
- In case of too many applications, we need to choose **randomly**.
- Group allocation deadline: 27.04.2023

Seminar Webpage