Data Quality for AI

Prof. Dr. Felix Naumann and Dr. Hazar Harmouch

WS 2021/2022
Agenda

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

- Chair Introduction
- Organizational Information
- Data quality and AI
- Your Tasks
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 does (training/testing) data quality influence the performance of AI models?
  - Is there a need for novel data quality dimensions? What could such dimensions be?

- **Deliverable**
  - Collaborative paper-style technical report
  - Code, models, and generated datasets

- **Teams**
  - 3 teams of 2 students each (At most 6 participants)
# Main Milestones

<table>
<thead>
<tr>
<th>Group allocation</th>
<th>Technical presentation of a paper</th>
<th>Think about a beneficial new dimension</th>
<th>Implement your Polluters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mid-term presentation</td>
<td>Build a machine learning pipeline</td>
<td>Run experiments and write about your results</td>
<td></td>
</tr>
<tr>
<td>End-term presentation</td>
<td>Final paper writing</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 3 teams
- 3 ML Tasks
- 3 quality dimensions
- 3 new dimension!
Grading

- Active participation in meetings and discussions
- Technical presentation of a scientific paper
- Mid- and End-term presentation
- Quality of implementation and results
- Final paper-style submission
Further Procedure

- To apply for this seminar (binding):
  - Email to hazar.harmouch@hpi.de
  - Deadline: Friday 29.10.2021 23:59
  - Notification: Monday 1.11.2021
  - Register with the Studienreferat

- In case of too many applications, we need to choose randomly.

- Group allocation deadline: **2.11.2021**

- Meeting next week: at **Campus II, Building F, Room 2.10 (instead of E.06)**.

- The course is **on-site**. However, we will switch to hybrid/online mode if the regulation changes.
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, recommendation

- **Comfort**
  - Sharing, smart home, authentication (face, gait)
  - Autonomous vehicles

- **Intelligence**
  - Fraud detection, translation, gaming
  - Robotics

[Image: https://unsplash.com/photos/JfoIjRnveY]
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

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

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**
AI performance is heavily influenced by the underlying data.

It is important to understand this correlation!

"Yes sir, you can absolutely trust those numbers"
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

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 even member of the Irish police's
FIFA registration form (2010)
<table>
<thead>
<tr>
<th>Datenelement</th>
<th>Name1</th>
<th>Name2</th>
<th>Name3</th>
<th>City</th>
<th>District</th>
<th>Street</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Handy-Nummer</td>
<td>41</td>
<td>501</td>
<td>10</td>
<td>0</td>
<td>2677</td>
<td>297</td>
<td>3526</td>
</tr>
<tr>
<td>Festnetznummer</td>
<td>15</td>
<td>98</td>
<td>6</td>
<td>0</td>
<td>221</td>
<td>9579</td>
<td>9919</td>
</tr>
<tr>
<td>Kostenstelle</td>
<td>283</td>
<td>1112</td>
<td>73</td>
<td>2</td>
<td>87</td>
<td>16</td>
<td>1573</td>
</tr>
<tr>
<td>Registriernummer</td>
<td>11</td>
<td>583</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>599</td>
</tr>
<tr>
<td>Lieferungsnummer</td>
<td>55</td>
<td>390</td>
<td>9</td>
<td>0</td>
<td>212</td>
<td>15</td>
<td>681</td>
</tr>
<tr>
<td>Abteilung</td>
<td>3711</td>
<td>9997</td>
<td>115</td>
<td>60</td>
<td>439</td>
<td>175</td>
<td>14497</td>
</tr>
<tr>
<td>Sperrkennzeichen</td>
<td>129</td>
<td>143</td>
<td>2</td>
<td>0</td>
<td>66</td>
<td>9</td>
<td>349</td>
</tr>
<tr>
<td>Löschkennzeichen</td>
<td>1028</td>
<td>442</td>
<td>5</td>
<td>36</td>
<td>113</td>
<td>10</td>
<td>1634</td>
</tr>
<tr>
<td>Rechtsform</td>
<td>131700</td>
<td>66136</td>
<td>187</td>
<td>6</td>
<td>64</td>
<td>57</td>
<td>198150</td>
</tr>
<tr>
<td>Kreditoreninfo</td>
<td>0</td>
<td>100</td>
<td>11</td>
<td>0</td>
<td>18</td>
<td>0</td>
<td>129</td>
</tr>
<tr>
<td>Kommissionsinfo</td>
<td>216</td>
<td>352</td>
<td>1</td>
<td>2</td>
<td>36</td>
<td>10</td>
<td>617</td>
</tr>
<tr>
<td>Baustelle</td>
<td>2013</td>
<td>3452</td>
<td>42</td>
<td>5</td>
<td>124</td>
<td>222</td>
<td>5858</td>
</tr>
<tr>
<td>Abladestelle</td>
<td>2923</td>
<td>3808</td>
<td>94</td>
<td>1503</td>
<td>958</td>
<td>3065</td>
<td>12351</td>
</tr>
<tr>
<td>Behörde</td>
<td>13410</td>
<td>12461</td>
<td>172</td>
<td>19</td>
<td>295</td>
<td>7075</td>
<td>33432</td>
</tr>
<tr>
<td><strong>Summe</strong></td>
<td>155535</td>
<td>99575</td>
<td>728</td>
<td>1634</td>
<td>5310</td>
<td>20533</td>
<td></td>
</tr>
</tbody>
</table>
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
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
New AI-specific Data Quality Dimensions and Where to Find Them

- Diversity
- Privacy
- Ownership
- Bias
- Protection
- Liability
- Explainability

**Acquisition**
- Training data
  - Preparation and cleaning
  - Annotated data for training
  - Annotated data for parameter tuning

**Modeling**
- Training
- AI model
- Validation data
- Annotated test data for evaluation

**Deployment**
- Application data
- Application
- Fresh data for use case
Agenda

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

- Form a team and choose a ML task: Classification, Regression, or Clustering
  - A specific data quality dimension will be assigned to each team.

- Each team will have the following tasks:
  - Find 3 datasets used for your task.
  - Implement a polluter in terms of your assigned data quality dimension.
  - Think of a novel data quality dimension and implement a respective polluter.
  - Implement a ML pipeline with 5 different algorithms that belong to your ML task.
  - Write about the results in the technical report.
  - In between: present a related work paper