Hasso Plattner Institut
Digital Engineering · Universität Potsdam

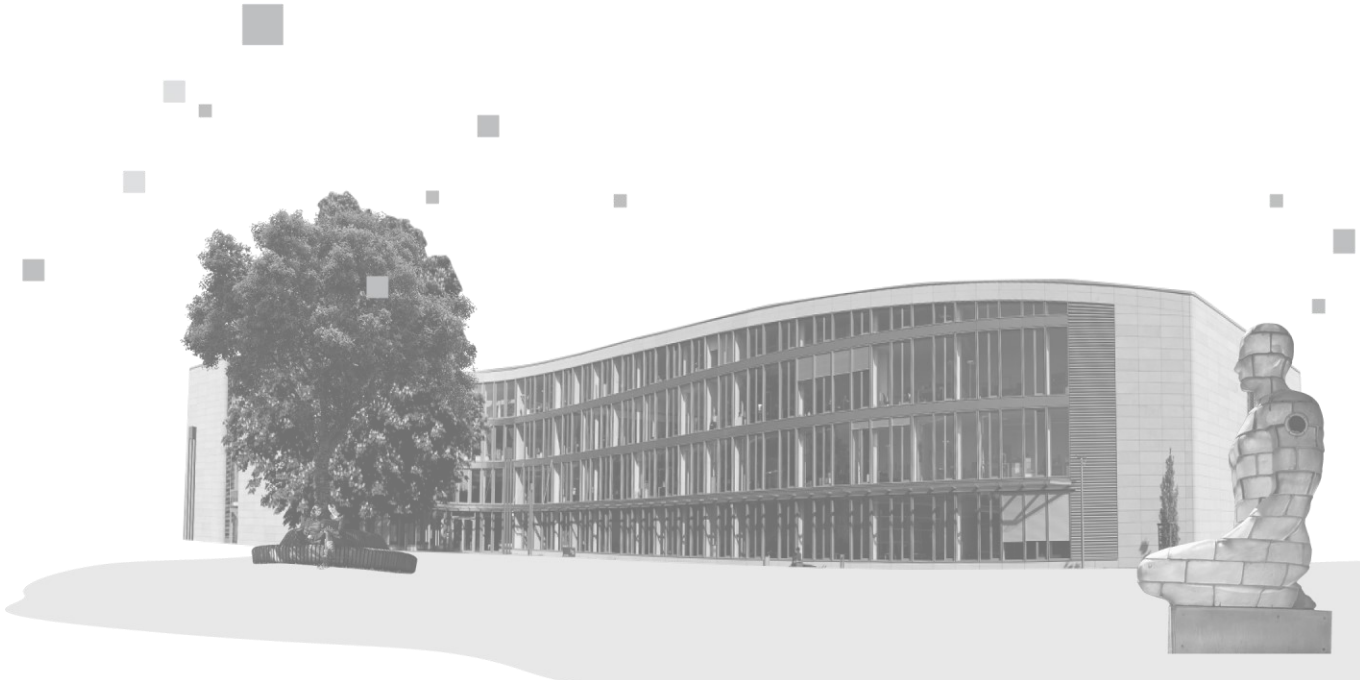
DQ4AI: Data Quality Assessment

Information Systems Group

Dr. Lisa Ehrlinger

Sedir Mohammed

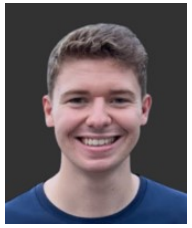
Prof. Dr. Felix Naumann



Design IT.
Create Knowledge.

www.hpi.de

Information Systems Group

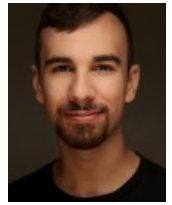


Lukas **Laskowski**

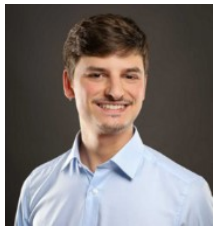
Diana **Stephan**

Prof. Felix **Naumann**

Dr. Lisa **Ehrlinger**

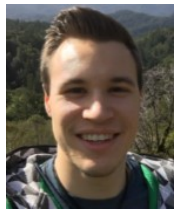


Sedir **Mohammed**



Francesco **Pugnaloni**

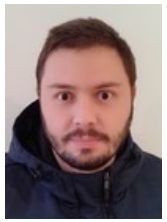
Divya **Bhadauria**



Sebastian **Schmidl**

Daniel **Lindner**

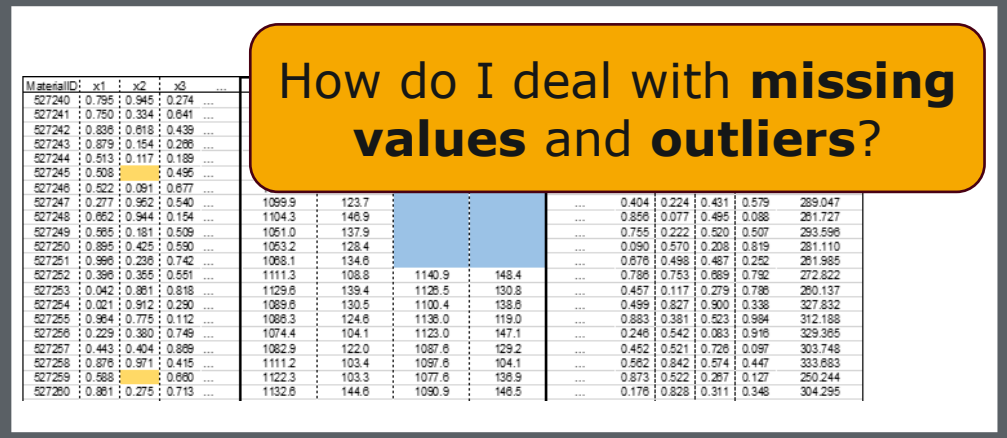
Youri **Kaminsky**



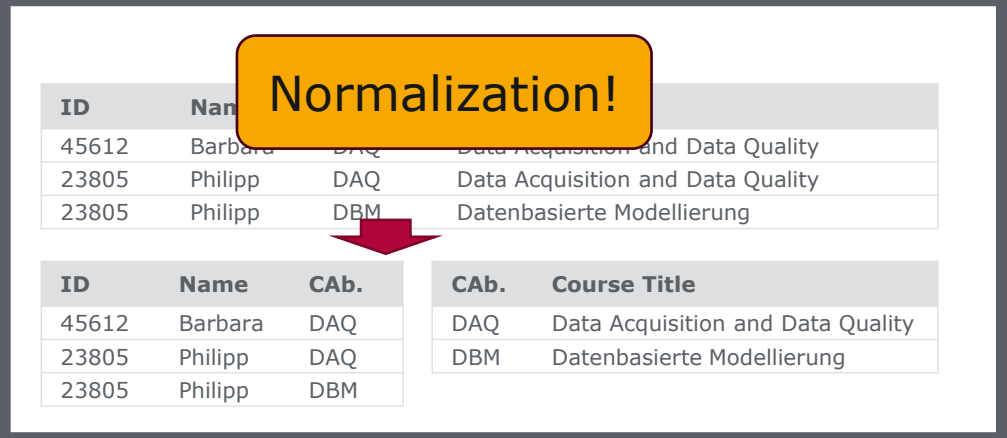
Alejandro **Sierra-Múnera**

project **KITQAR** Data Quality Assessment
Data Change Data Fusion
Duplicate Detection project **QuanTD**
Data Profiling
Information Integration Web Science
project **AI4ART** Data Scrubbing project **QuAHT** Data as a Service
Information Quality Data Cleansing **CSV parsing**
Dependency Detection Linked Open Data Text Mining
Distributed Computing Knowledge Management for the Arts
Web Data project **Metanome** **Entity Recognition** Data Preparation project **Janus**
Change Exploration

Data Quality



What data scientists think it is



What database admins think it is

What researchers think it is

Foto: fizkes – shutterstock.com

What managers think it is

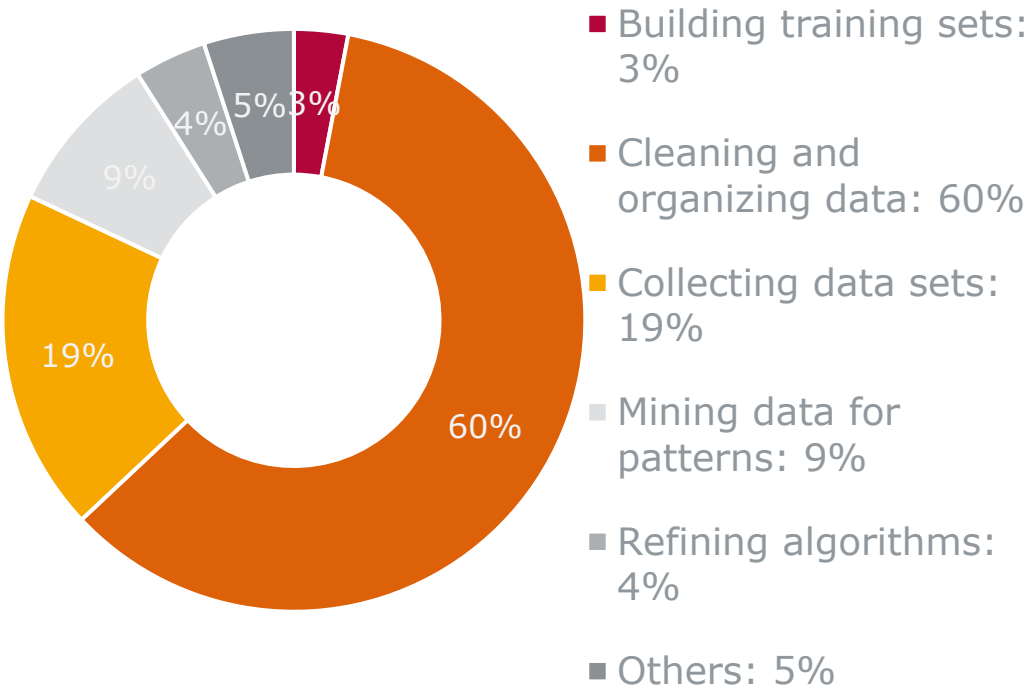
The Impact of Data Quality (1)

- Incorrect decision making
- Financial loss
  - Gartner (2018): poor data quality causes an average loss of $15 millions per year in companies
  - Redman (2016): poor data quality costs the US $3 trillion per year
- Customer dissatisfaction
  - Wrong addresses, names, duplicates or no mail deliveries …
- Decreasing trust in organizations
  - During the COVID pandemic due to incorrect statistics
  - Deutsche Bank accidentally transferred 28 billion euros to a Eurex account (March, 2018)

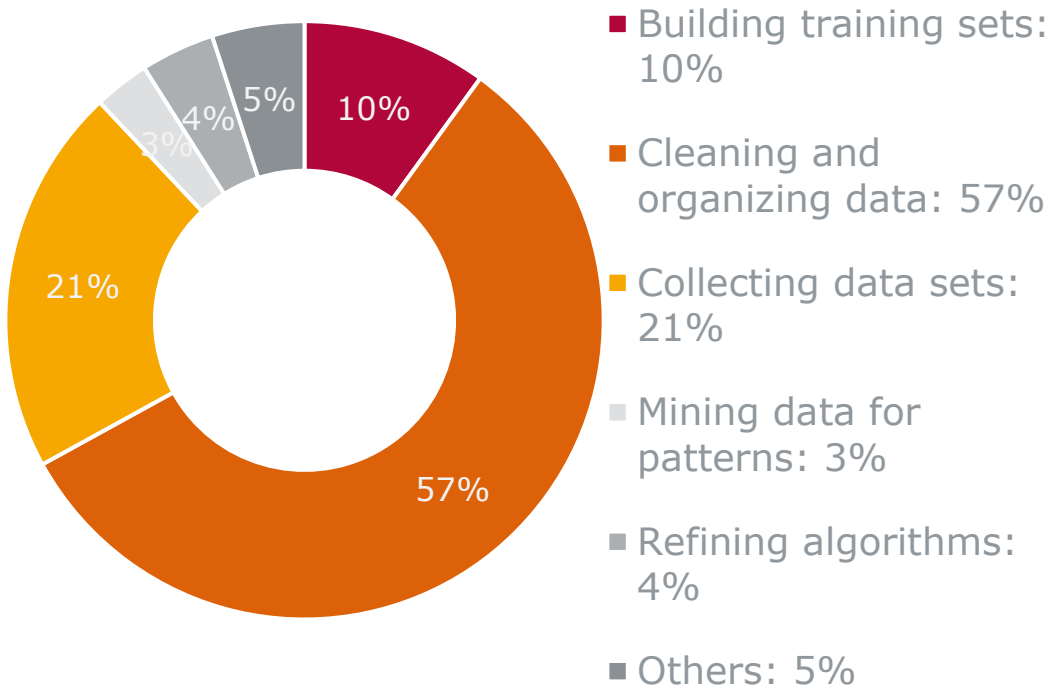
"How to Create a Business Case for Data Quality Improvement", S. Moore, Gartner, 2018. https://www.gartner.com/smarterwithgartner/how-to-createa-business-case-for-data-quality-improvement

The Impact of Data Quality (2)

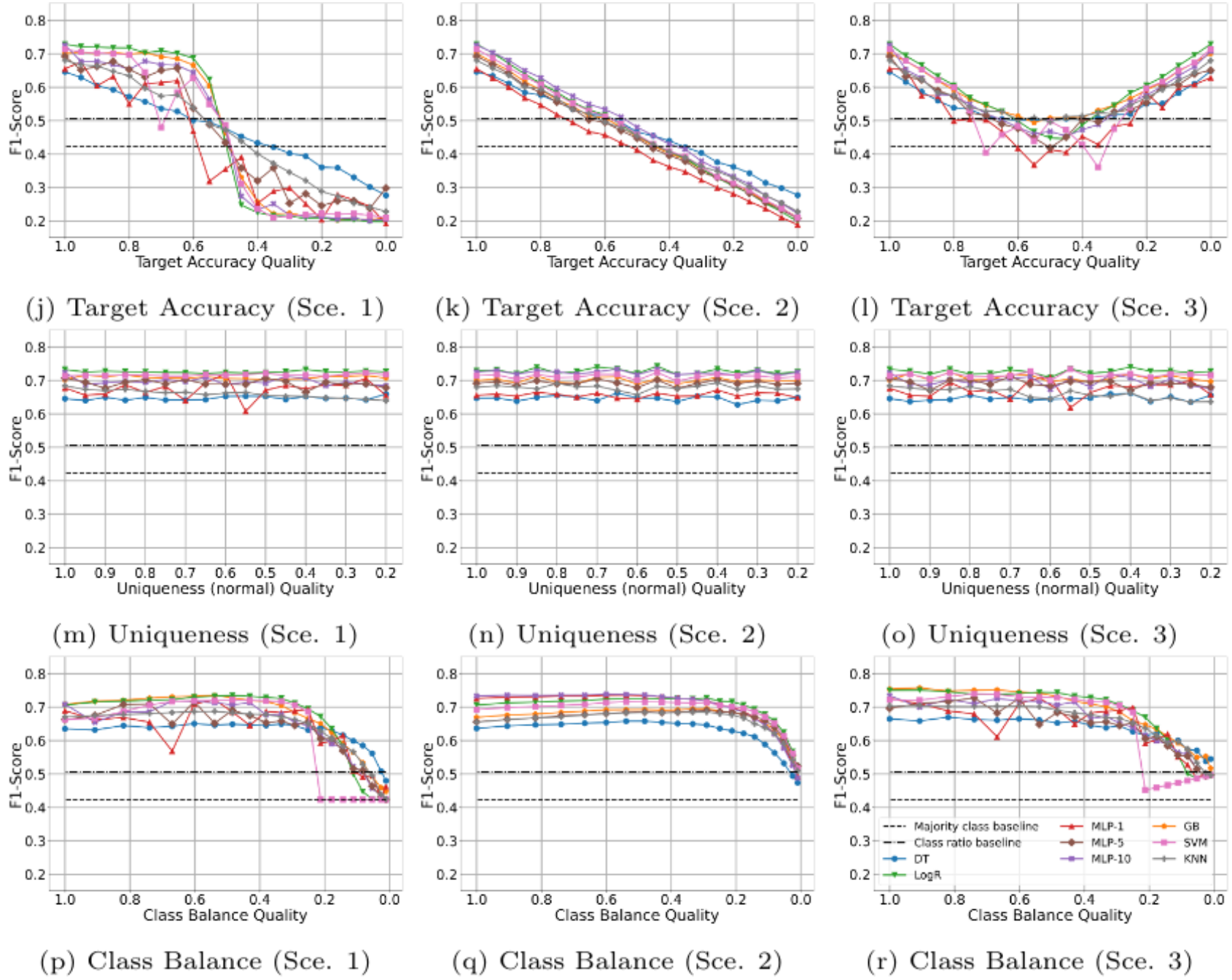
What data scientists spend the most time doing?



What is the least enjoyable part of data science?



"Cleaning Data: Most Time-Consuming, Least Enjoyable Data Science Task", Gil Press, Forbes. http://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says



(j) Target Accuracy (Sce. 1) (k) Target Accuracy (Sce. 2) (l) Target Accuracy (Sce. 3)

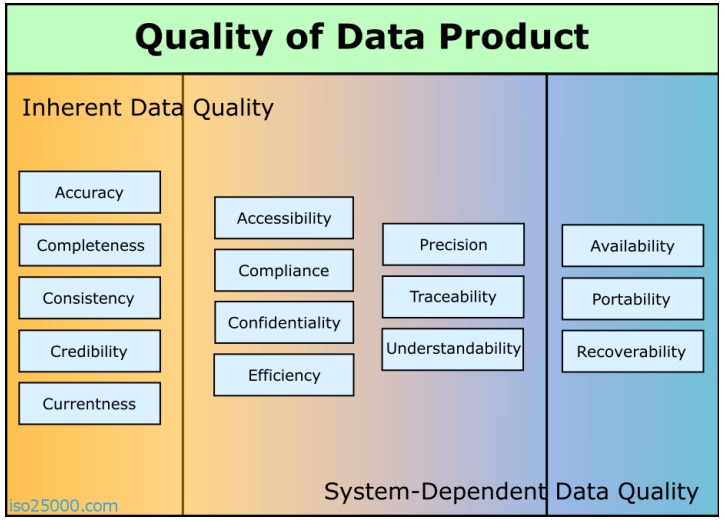
(m) Uniqueness (Sce. 1) (n) Uniqueness (Sce. 2) (o) Uniqueness (Sce. 3)

(p) Class Balance (Sce. 1) (q) Class Balance (Sce. 2) (r) Class Balance (Sce. 3)

DQ4AI paper

Data Quality Dimensions

- Prof. Richard Wang and his team from the Massachusetts Institute of Technology (MIT)
  - Initiated data quality research in the 1980s
  - Defined data quality as “fitness for use”
- DQ is described by DQ dimensions, which can refer to
  - The quality of data values (also: extension of a DB) or
  - The quality of the DB schema (also: data structure, metadata, intension)
- DQ metrics are concrete formulas to quantify a dimension with a numerical value
- There is no **standardized classification nor definition** for DQ dimensions and consequently not for DQ metrics!



R. Y. Wang and D. M. Strong. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, 12(4):5–33, March 1996.
R. Y. Wang. A Product Perspective on Total Data Quality Management. *Communications of the ACM*, 41(2):58–65, 1998.
International Organization of Standardization. ISO/IEC 25012. Standard on Data Quality.

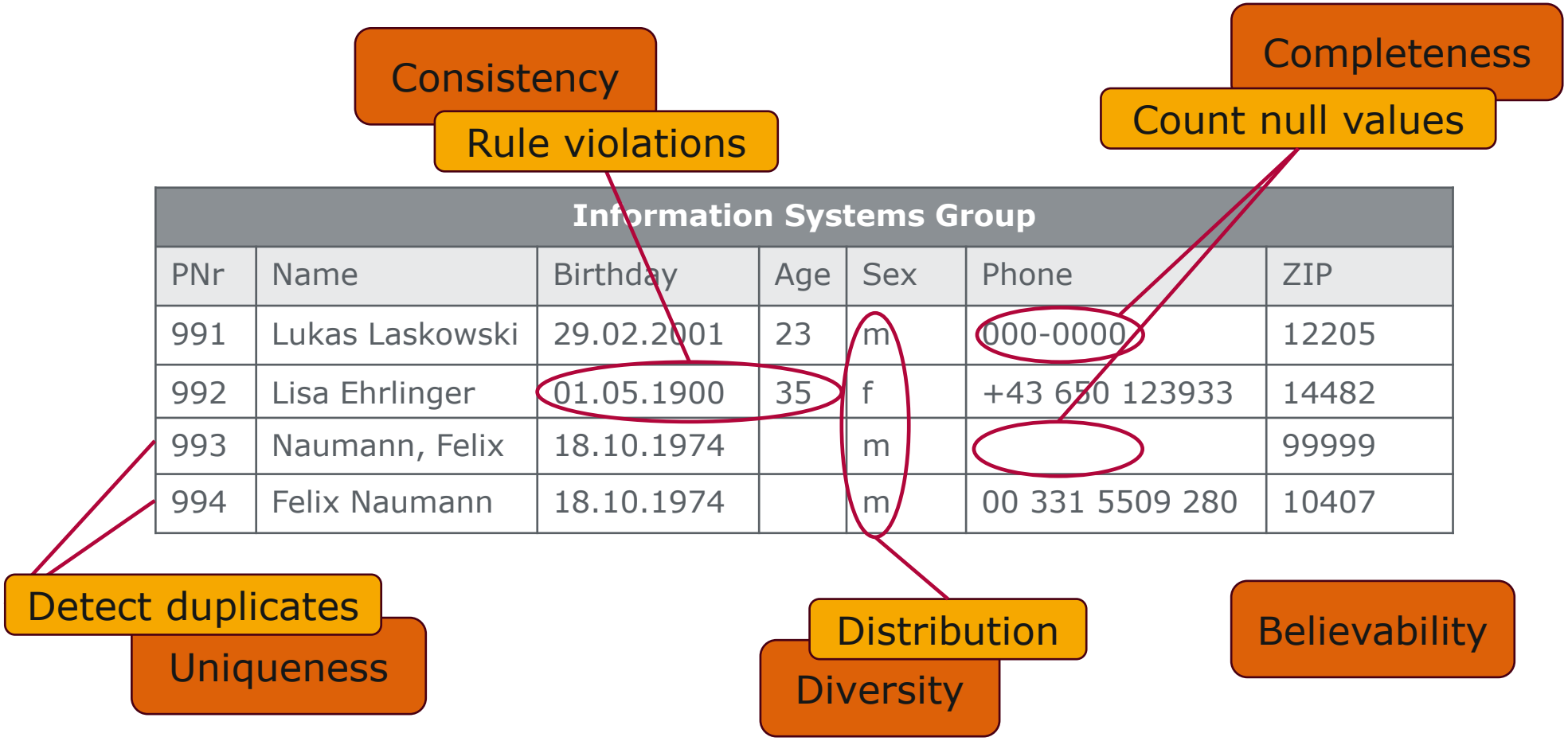
Common DQ Dimensions

Table 2. Notable data quality dimensions

Dimension	# cited	Dimension	# cited	Dimension	# cited
Accuracy	25	Format	4	Comparability	2
Reliability	22	Interpretability	4	Conciseness	2
Timeliness	19	Content	3	Freedom from bias	2
Relevance	16	Efficiency	3	Informativeness	2
Completeness	15	Importance	3	Level of detail	2
Currency	9	Sufficiency	3	Quantitativeness	2
Consistency	8	Usableness	3	Scope	2
Flexibility	5	Usefulness	3	Understandability	2
Precision	5	Clarity	2		

Y.Wand and R. Y.Wang. Anchoring Data Quality Dimensions in Ontological Foundations. Communications of the ACM, 39(11):86–95, November 1996.

Data Quality Assessment



DQ Dimension: Accuracy

- Key dimension in DQ research
- Diverse definitions and interpretations
- Most often referred to as „magnitude of an error“

$$field\ level\ accuracy = \frac{number\ of\ fields\ judged\ \text{“}correct\text{”}}{number\ of\ fields\ tested}$$

$$record\ level\ accuracy = \frac{number\ of\ records\ judged\ \text{“}completely\ correct\text{”}}{number\ of\ records\ tested}$$

$$p = \frac{number\ of\ Number\ of\ correct\ values}{Number\ of\ total\ values}$$

$$AccuracyOfOperationalDatabases_{ij} = Local\_accuracy_{ij} - outofdate_{ij}$$

$$Free\text{-}of\text{-}error\ rating = 1 - \left(\frac{Number\ of\ data\ units\ in\ error}{Total\ number\ of\ data\ units}\right)$$

$$AccuracyOfNumericalValues = InaccuracyOfNumericalValues = v' - v$$

$$accuracy = \left(\frac{NrOfCorrectValues}{TotalNrOfValues}, RandomnessOfTheOccuranceOfAnError, ProbabilityDistributionOfTheOccuranceOfAnError\right)$$

$$Inaccuracy = \frac{InaccurateValues}{TotalValues}$$

Haegemans, T., Snoeck, M., & Lemahieu, W. (2016). Towards a precise definition of data accuracy and a justification for its measure. In Proceedings of the International Conference on Information Quality (pp. 16-16). MIT Information Quality (MITIQ) Program.

DQ Dimension: Consistency

- "Consistency captures the violation of semantic rules defined over data items, where items can be tuples of relational tables or records in a file" (Batini & Scannapieco 2016)
- Hinrichs defines consistency according to:

$$Q_{Kon}(w) = \frac{1}{\sum_{j=1}^{n} r_j(w) g_j + 1} \text{ [126]}, \tag{2.13}$$

where g_j is the degree of severity of $r_j(w)$, and $r_j(w)$ is the violation of consistency rule r_j (within a set of n consistency rules), applied to the attribute value w, and defined as

$$r_j(w) \begin{cases} 0 & \text{if } w \text{ satisfies } r_j \\ 1 & \text{otherwise. [126]} \end{cases} \tag{2.14}$$

H. Hinrichs. *Datenqualitätsmanagement in Data Warehouse-Systemen [Data Quality Management in Data Warehouse Systems]*. PhD thesis, Universität Oldenburg, 2002.

DQ Dimension: Completeness

- Completeness goes beyond measuring NaN, null values …
  - Default values (e.g., -99, „01.01.2000“)
- Completeness according to Sebastian-Coleman 2013 & Wang & Strong 1996:
  - Breadth / width: “dataset contains all attributes”
  - Depth: dataset contains the wanted amount of data
  - Scope of information in the data: attributes are populated to the desired extend
- Distinction by Batini & Scannapieco 2016:
  - Value completeness
  - Tuple completeness
  - Attribute completeness
  - Relation completeness

L. Sebastian-Coleman, Measuring Data Quality for Ongoing Improvement, Waltham: Morgan Kaufmann, 2013.
C. Batini and M. Scannapieco, Data and Information Quality, Cham: Springer International Publishing, 2016.
R. Y. Wang and D. M. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers," Journal of Management Information Systems, vol. 12, no. 4, pp. 5-33, 1996.

DQ Dimension: Minimality / Non-Redundancy

Minimality metric based on hierarchical clustering, which fulfills all requirements for DQ metrics by Heinrich et al. 2018:

1. **Similarity calculation**: calculate the similarity or distance between all schema elements
2. **Clustering**: group very similar elements to clusters
3. **Calculate minimality** according to:

$$Min(s) = \frac{|\text{unique}(e)|}{|e|} = \begin{cases} 1.0, & \text{if } |e| = 1 \\ \frac{|c|-1}{|e|-1}, & \text{else} \end{cases}$$

|c| = number of clusters

|e| = number of elements

Ehrlinger and W. Wöß. A Novel Data Quality Metric for Minimality. In Data Quality and Trust in Big Data, vol. 11235 of Lecture Notes in Computer Science, pp. 1–15, Cham, Switzerland, 2019. Springer International Publishing.

DQ Dimension: Timeliness

- Timeliness = how current the data are for a task at hand
- Ballou et al. propose a metric based on *currency* and *volatility*

$$Timeliness = max\left(0, 1 - \frac{Currency}{Volatility}\right) \qquad Cur(r) = DeliveryTime(r) - InputTime(r) + Age(r)$$

  - *Age* of data (prior to system entrance) is rarely available in practice
  - *DeliveryTime* (when data is delivered to the customer) requires in-depth domain knowledge
- For simplicity, it could be assumed that *Age=0* and *DeliveryTime=now()*

$$Cur(c) = \frac{\sum_{r \in c} Now - InputTime(r)}{|c|}$$

D. Ballou, R. Wang, H. Pazer, and G. K. Tayi. Modeling Information Manufacturing Systems to Determine Information Product Quality. Management Science, 44(4):462–484, 1998.

Data Quality Measurement and Metrics

- **Measurement** is the process by which *numbers or symbols* are assigned to real-world entities entities to describe them according to defined rules
  - Aim: to draw conclusions about an entity (e.g., is this piece of wood too long?)
  - All measures are invented, e.g., a ruler and the centimeter measure to measure length
- **DQ measurement** allows to analyze defined quality attributes of the data
  - Are there any outliers in my data? (→ requires definition of "outlier")
  - Is the number of null values low enough? → requires definition of null and threshold for "low enough"
- Unfortunately, there is no common understanding of data attributes: thus, we need to define how to measure data quality, i.e., which **metrics** we use!
- **Metrics**: to draw conclusions about data quality (e.g., completeness, outliers) it is necessary to create a mapping *M* from the observed (profiled) data to a defined numerical system

"Without a precise definition of what is being measured and without a sound justification for the measures themselves, the assessment of data quality will remain an ad hoc process instead of a scientific one."
(Pipino et al. 2005)

L. Ehrlinger. Automating Data Quality Measurement. *Dissertation, Johannes Kepler University Linz*, 2021.
N. Fenton and J. Bieman. Software Metrics: A Rigorous and Practical Approach. *CRC Press*, 2015.
L. Pipino, R. Wang, D. Kopcso, and W. Rybolt. Developing Measurement Scales for Data-Quality Dimensions. *Information Quality*, 1:37–52, 2005.

Challenges in the Development of DQ Metrics

Each DQ dimension should be discussed according to:

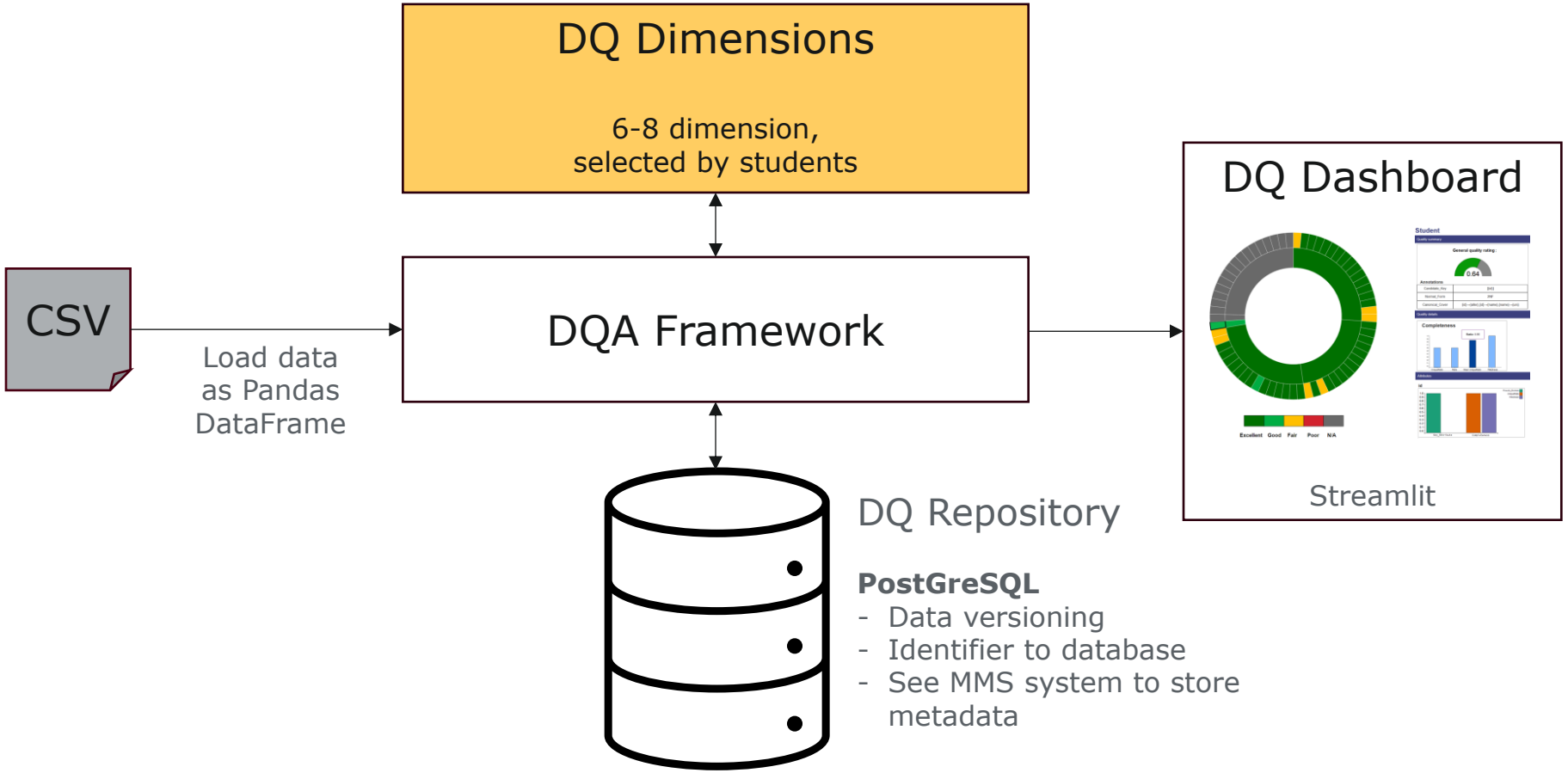
- **Definition of DQ dimension**: how is it defined?
- **Metric**: with which function can it be measured?
- **Input**: which statistics / metadata are required for its measurement?
- **Output**: which scores and explanations can be expected as measurement result?
- **Aggregation**: on which data structure can/should the metric be calculated?
- **Soundness**: is the metric sound? Does it fulfil the requirements for DQ metrics by Heinrich et al.?
- **Scalability**: how scalable is the metric? Can we improve calculation efficiency for large data sets?
- **Customizable**: can I adjust it to a specific use case?
- **Evaluation**: how should the metrics be evaluated (if no gold standard is available)?
- **Presentation and explanation**: how can the metric be presented to a user in an understandable way?

List of Related Research

- Keshav, S. (2007). How to read a paper. ACM SIGCOMM Computer Communication Review, 37(3), 83-84.
- Ehrlinger, L., Werth, B., & Wöß, W. (2018). Automated continuous data quality measurement with QuaIIe. International Journal on Advances in Software, 11(3), 400-417.
- Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. Journal of management information systems, 12(4), 5-33.
- Heinrich, B., Hristova, D., Klier, M., Schiller, A., & Szubartowicz, M. (2018). Requirements for data quality metrics. Journal of Data and Information Quality (JDIQ), 9(2), 1-32.
- Mohammed, S., Harmouch, H., Naumann, F., & Srivastava, D. (2024). Data Quality Assessment: Challenges and Opportunities. arXiv preprint arXiv:2403.00526.

What to be done next

DQ4AI Seminar: DQ Assessment Framework



Your Next Task

- Choose a group partner
- Choose a task:
  - Choose a DQ dimension OR
  - Design and develop the DQ framework architecture
- Search for literature
- Prepare presentation and plan on how to measure this DQ dimension / develop the framework

- Course communication
  - Via HPI slack channel
  - Contact us via e-mail: {lisa.ehrlinger, sedir.mohammed}@hpi.de
  - You can find us in F-2.08

Deliverables

- Paper-style technical report (Overleaf project will be provided) containing
  - DQ assessment framework (design decisions)
  - DQ dimensions and metrics
  - Experimental evaluation
- Code of the framework + experiments (via GIT repo)

- Optional: we would like to publish the paper in a scientific conference

DQ4AI Seminar Overview

Date	Topic
14.10.24	Introduction incl. Group allocation and topic selection
21.10.24	Group allocation and topic selection
22.10.24, 15:15-16:45	How to read a paper?
04.11.24	Report on progress and questions
18.11.24	Report on progress and questions
25.11.24	Report on progress and questions
02.12.24	**Mid-term presentation of DQ dimensions / framework presentation**
09.12.24	Report on progress and questions
16.12.24	Report on progress and questions
06.01.25	Report on progress and questions
20.01.25	Report on progress and questions
03.02.25	**End-term presentation of DQ dimension implementation**
01.03.25	**Final submission**

Shortly introduce yourself