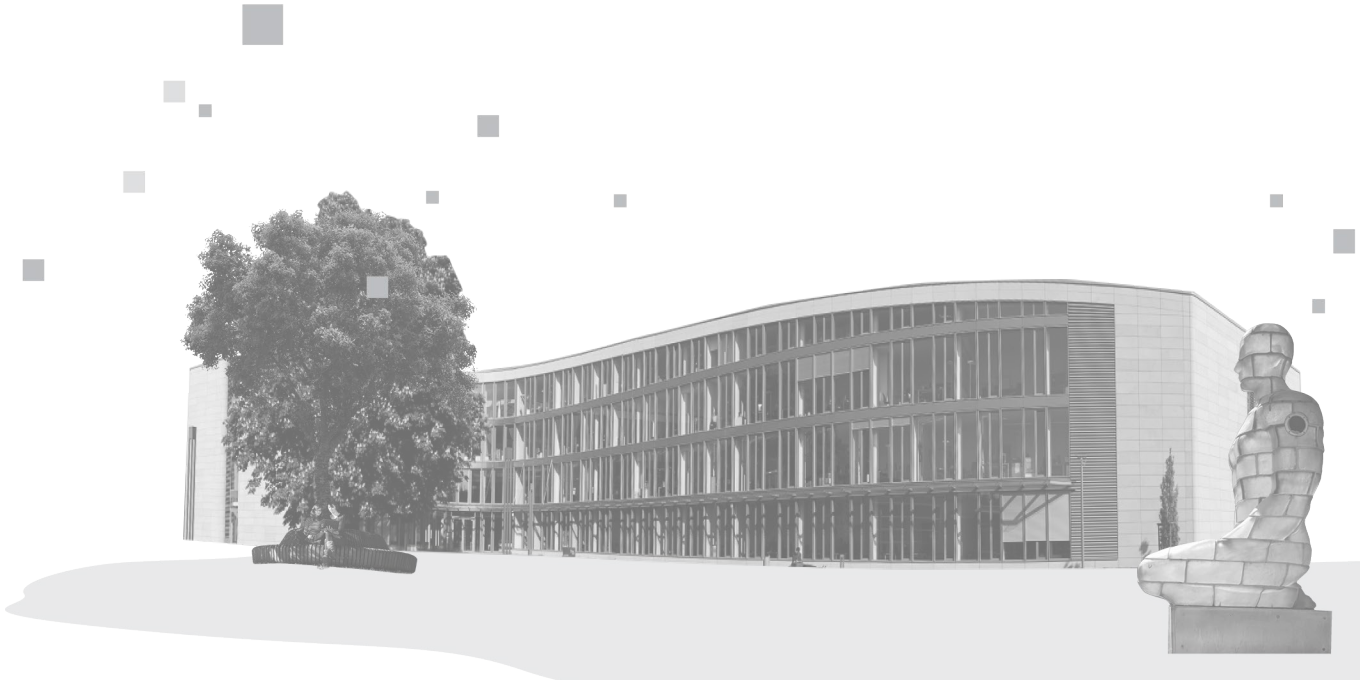
HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

WikiWatch: Data Quality Assessment in Wikidata

Information Systems Group

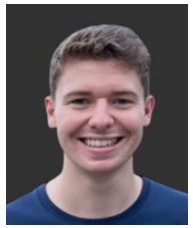
Carolina Cortés Lasalle
Dr. Lisa Ehrlinger
Prof. Dr. Felix Naumann



Design IT.
Create Knowledge.

www.hpi.de

Information Systems Group

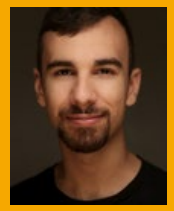


Lukas **Laskowski**

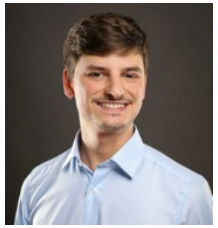
Diana **Stephan**

Prof. Felix **Naumann**

Dr. Lisa **Ehrlinger**



Sedir **Mohammed**



Francesco **Pugnaloni**

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
Entity Recognition
Web Data
project **Metanome**
Data Preparation
project **Janus**
Change Exploration

Divya **Bhadauria**

Daniel **Lindner**

Youri **Kaminsky**

Philipp **Hildebrandt**

Carolina **Cortes**

Code of Conduct – Overview

At DEF/HPI, we are committed to providing a high-quality learning as well as research environment and building a community where students and staff can thrive scientifically and personally. Everyone should expect a safe, supportive, and inclusive environment in all our spaces.

Our Code of Conduct helps us meet this goal. Words or actions that are disrespectful, racist, discriminatory, hostile, or harassing are not acceptable.

Examples of these include:

- Offensive comments about others' ethnicity, accent, religion, nationality, gender, sexual orientation, or other personal traits
- Refusing to work with someone based on these personal traits
- Physical or verbal threats and assaults
- Using sexualized or vulgar language or actions
- Disrupting another person's work experience

Code of Conduct – Help and Support

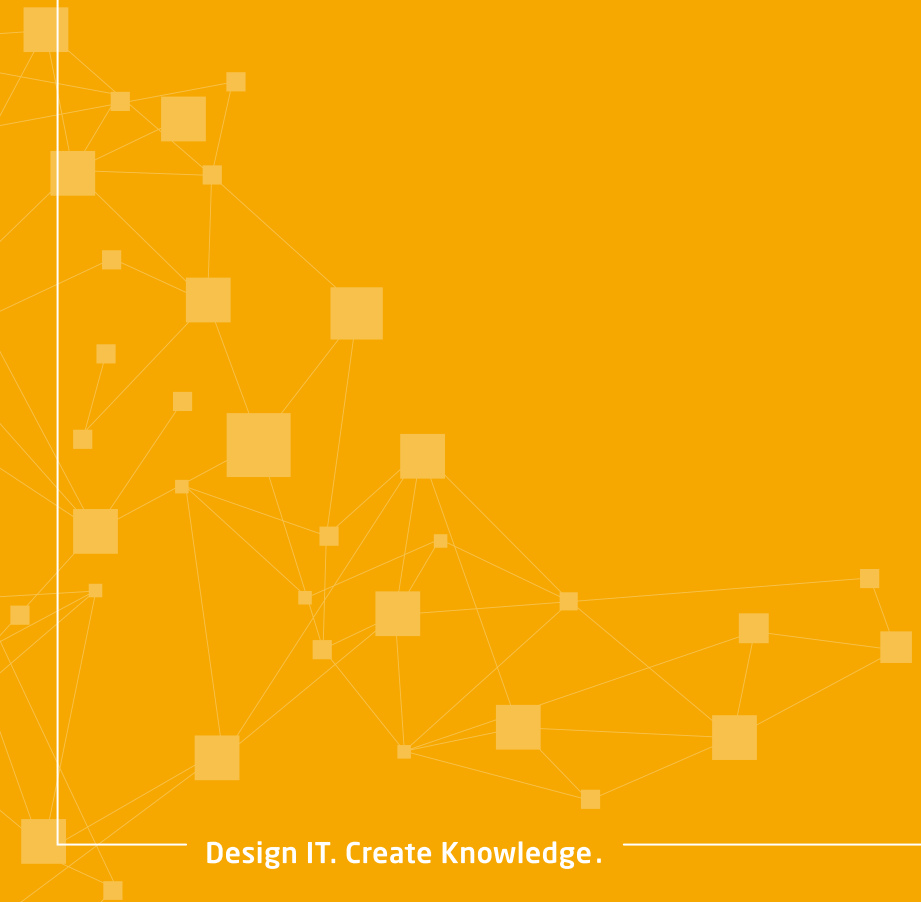
Violations of this code are taken seriously. If you witness or experience any inappropriate behavior, report it to a lecturer or any DEF/HPI contact point. All reports will be handled confidentially and with care.

Please be aware of **further contact points and support** structures at DEF/HPI, including:

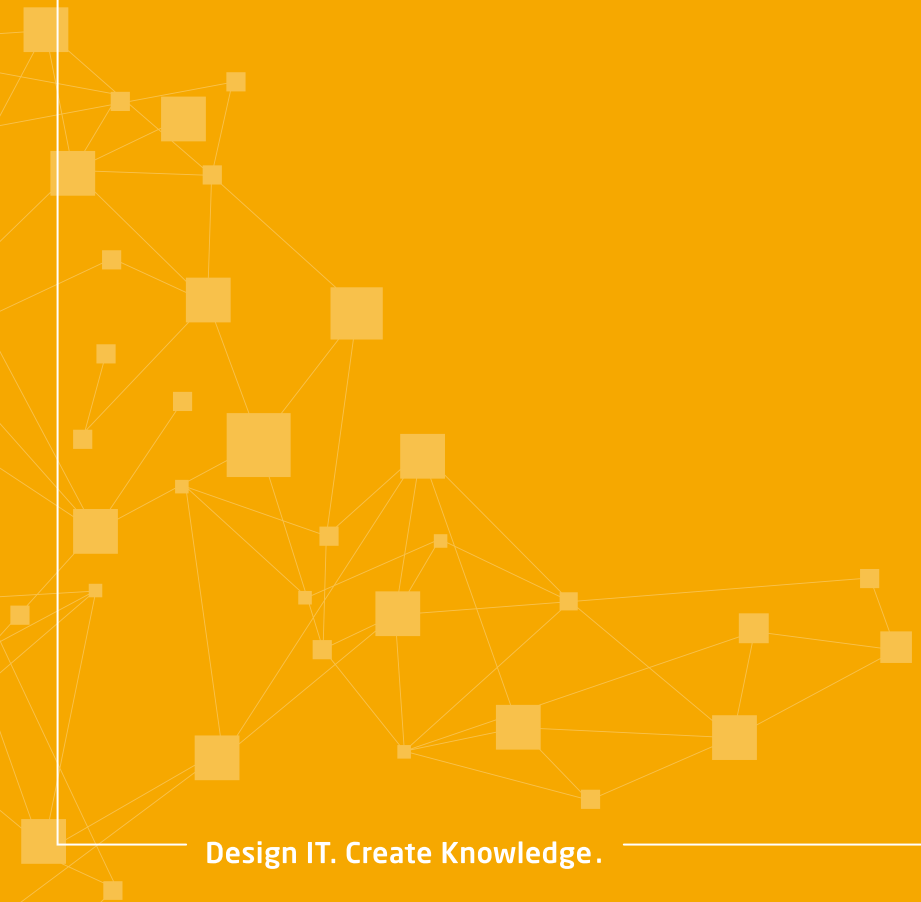
- Equal Opportunities Officers (Charlotte Weiss, Florence Böttger, Oliwia Gust)
- Diversity Manager (Oliwia Gust)
- Ombudsman for good scientific practice (Prof. Tilmann Rabl)
- Student Trusted Advisors (Hanna, Joscha, Ronja and Zero)
- Psychological counseling hotline (0800 7777015)
- Incident Response System (safecampus.hpi.de)
- as well as the respective offers of the University of Potsdam (Mental Health Counseling Service, Psychosocial Counseling of Studentenwerk, Nightline)

  www.uni-potsdam.de/en/discrimination-free-university/consulting-and-support/overview-of-consueling-and-advising-services

Shortly Introduce Yourself :-)

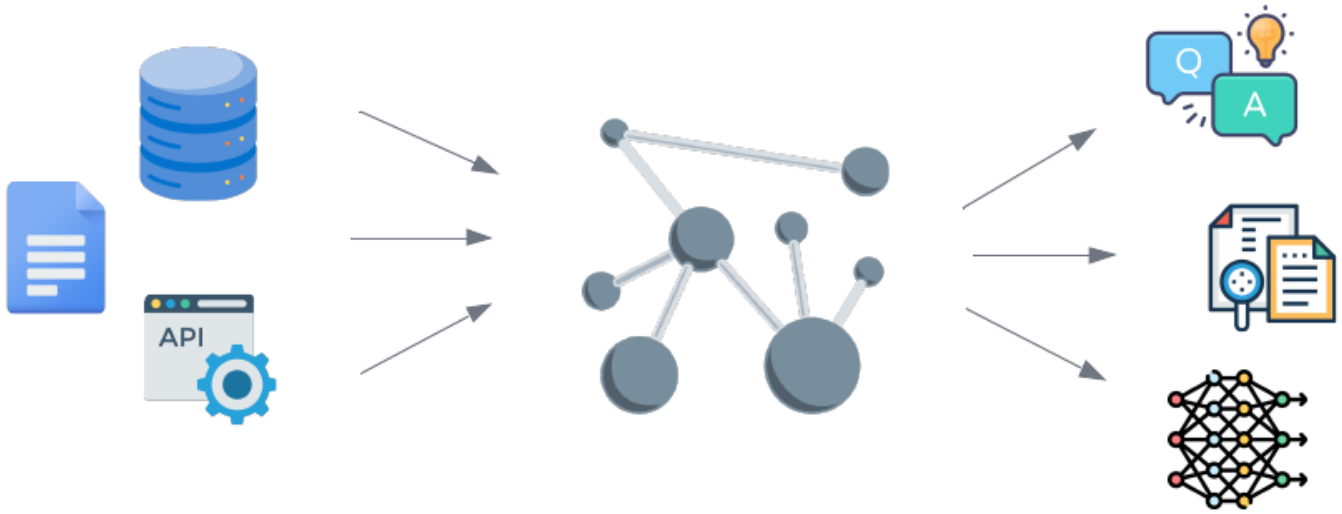


About Wikidata



Knowledge Graphs

Knowledge Graphs represent information about the real world with nodes and edges, where nodes represent entities of interest and edges represent relations between these entities. [1]

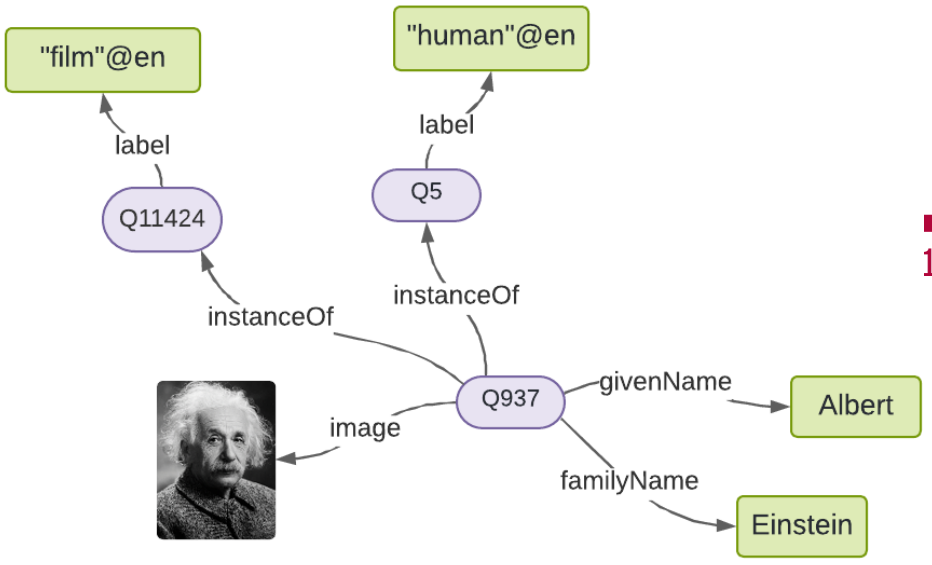


[1] Hogan, A., Blomqvist, E., Cochez, M., D'Amato, C., Melo, G. de, Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A. C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). Knowledge Graphs. *ACM Computing Surveys (CSUR)*, *54*(4). https://doi.org/10.1145/3447772

Knowledge Graphs

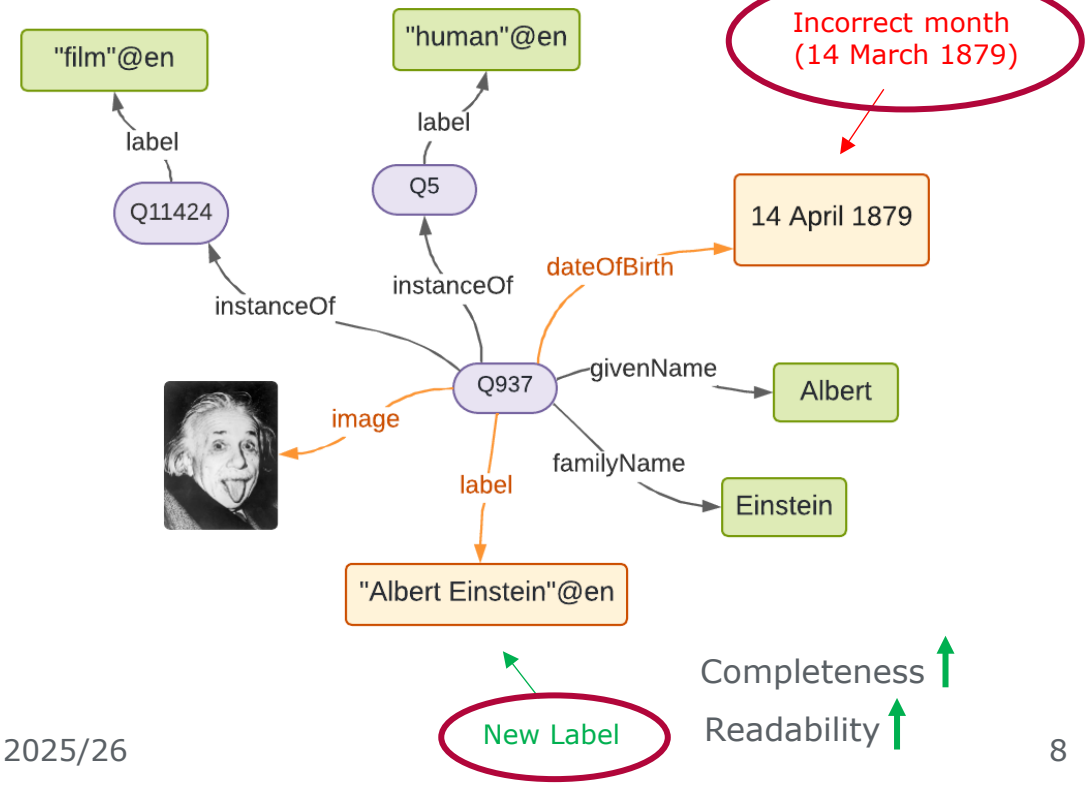
- Knowledge Graphs evolve over time, so new information can be added, removed or updated.
- Changes can improve or degrade the quality of the information, impacting different quality dimensions.

June 2024:



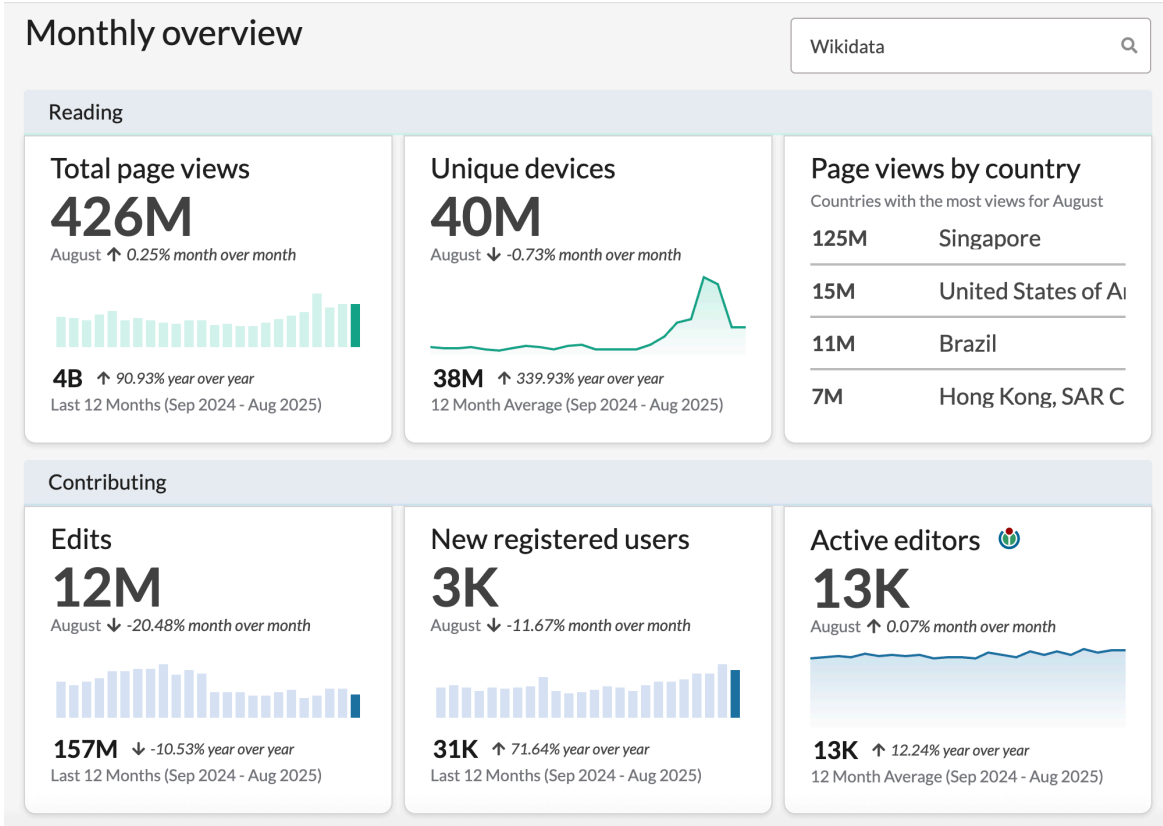
1 year later…

June 2025:



Wikidata

- Collaborative Knowledge Graph – curated by both humans and machines.
- Free knowledge base with 118,982,713 data items.



Wikimedia Foundation. (n.d.). *Wikidata statistics*. Retrieved September 19, 2025, from https://stats.wikimedia.org/#/wikidata.org

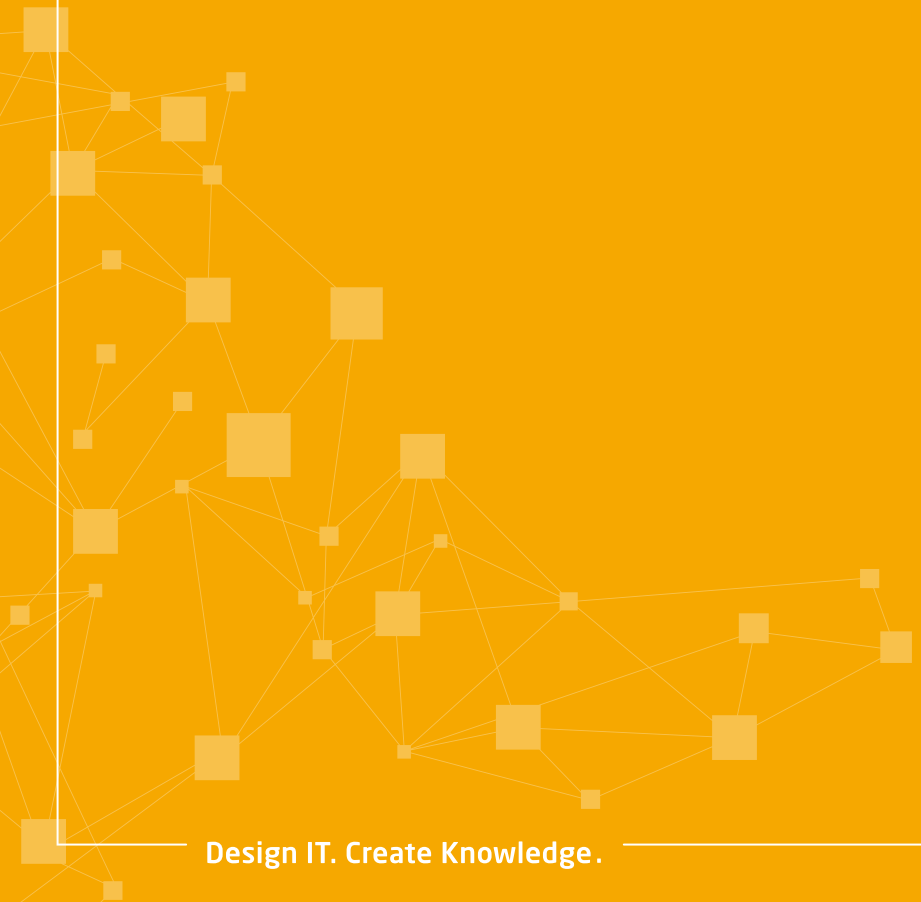
Cooperation partner in Wikidata

- Wikimedia:
  - Projects: Wikipedia, Wikidata, Wikibooks, etc.

- Our contact from Wikimedia Deutschland:
  - Lydia Pintscher (located in Berlin) - Portfolio Lead for Wikidata

https://de.m.wikipedia.org/wiki/Datei:Lydia_Pintscher_-_1.jpg

About Change in Data



Project Overview: IANVS.org

- The first derivative of data

$$\frac{dB}{dt}$$

- Joint work


Tobias **Bleifuß**


Leon **Bornemann**


Dmitri **Kalashnikov**

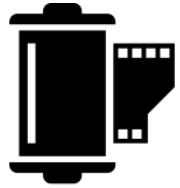

Felix **Naumann**


Divesh **Srivastava**

Change exploration

For a given, **dynamic dataset**,

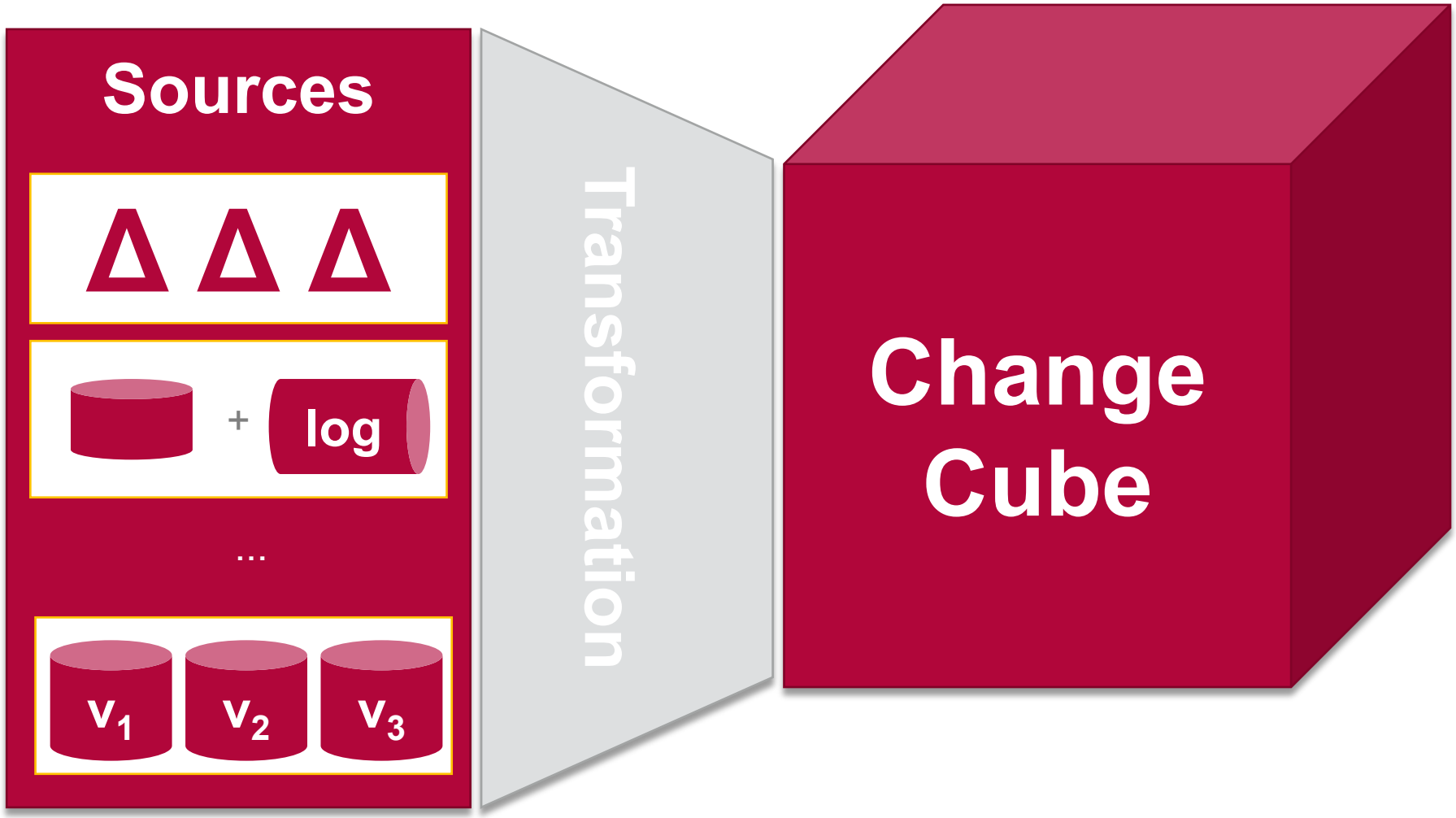
efficiently capture and **summarize**



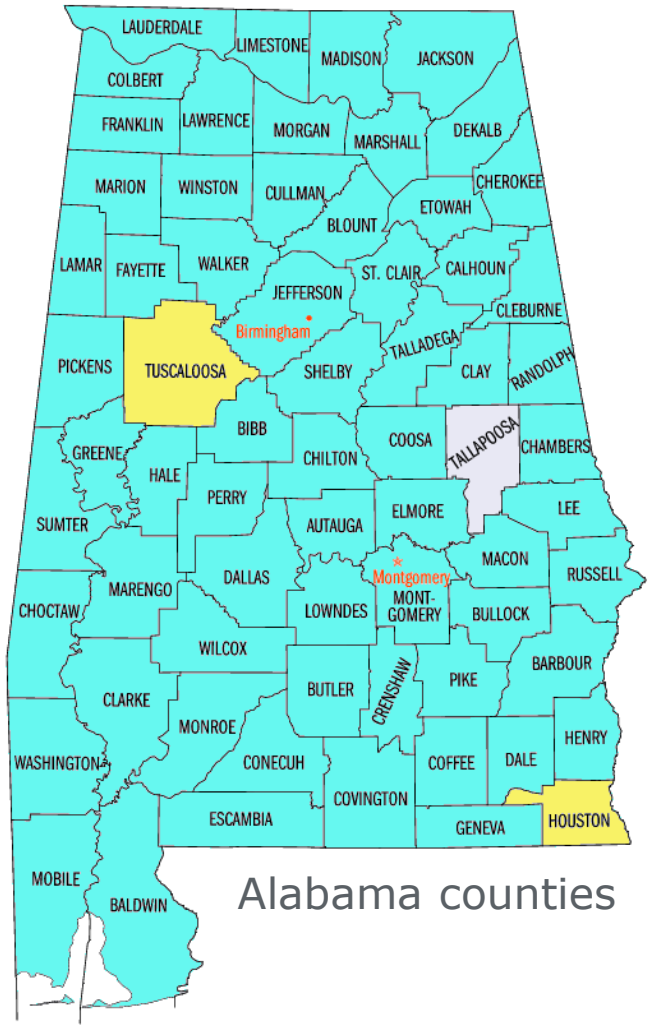
changes at **instance-** and **schema-level,**

and enable users to **effectively explore** this change in an **interactive** and **graphical** fashion.

Data model – The Change Cube



Systematic changes



Alabama counties

2nd & 3rd August 2010:
"Hyperfast" updates population (and related fields) in accordance to 2009 census for 66 out of 67 Alabama counties

3rd August 2010 (30 minutes):
"Altairisfar" reverts 61 of the changes by Hyperfast

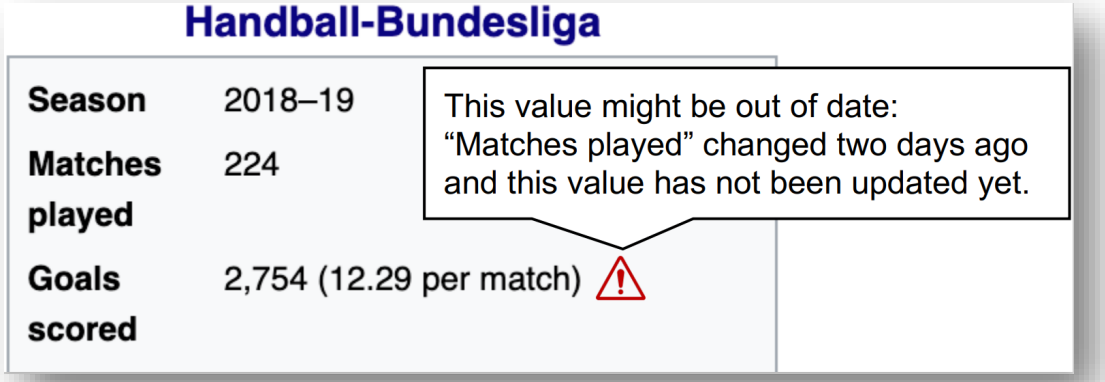
24th August 2010:
"Altairisfar" reverts 4 more changes by Hyperfast

Our goal: Improve data quality by ensuring that systematic changes are **consistently** applied to all entities within the intended **scope**.

Trust Assessment through Change Prediction

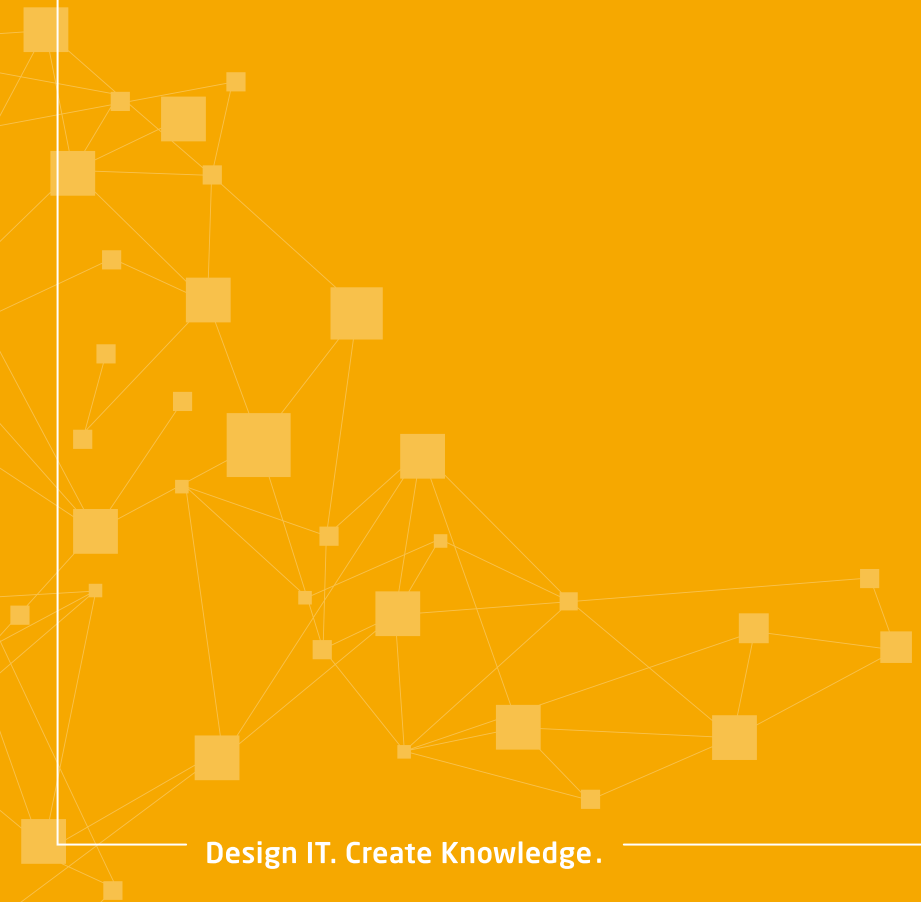
- Predict change, based on
  - Temporal change patterns
  - Change dependencies
  - Changes for same entity, or same property, or same value

- Use-cases
  - Suggest unanticipated changes for manual review
    - Fighting fraud or vandalism
  - Point out expected changes that in fact did not happen
    - Warn data owners
  - Autocomplete systematic changes
  - Annotate outdated values

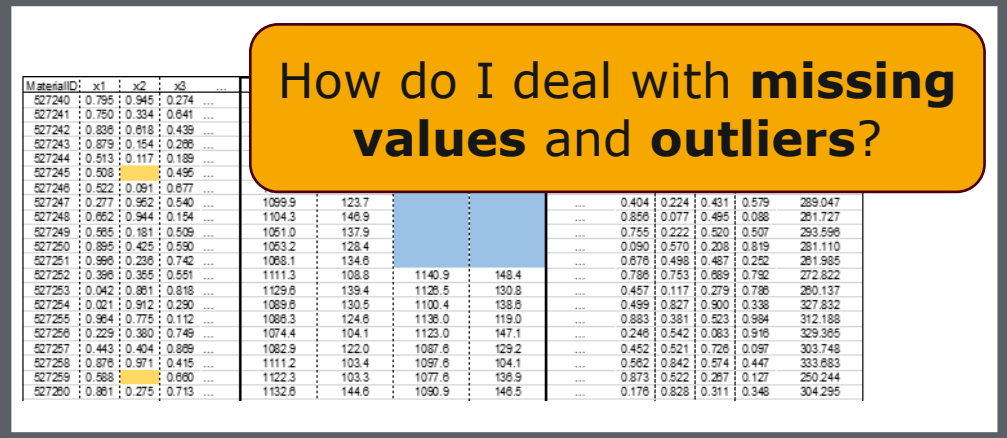


Vandalism in Wikipedia

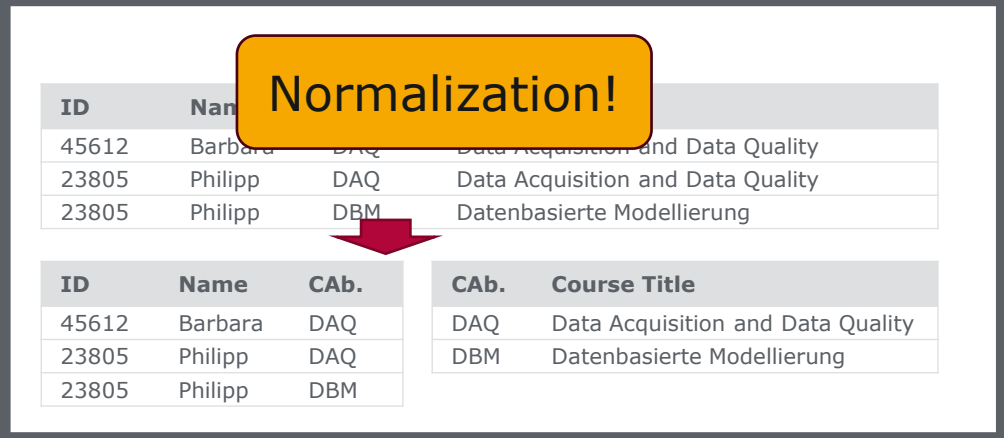
About Data Quality



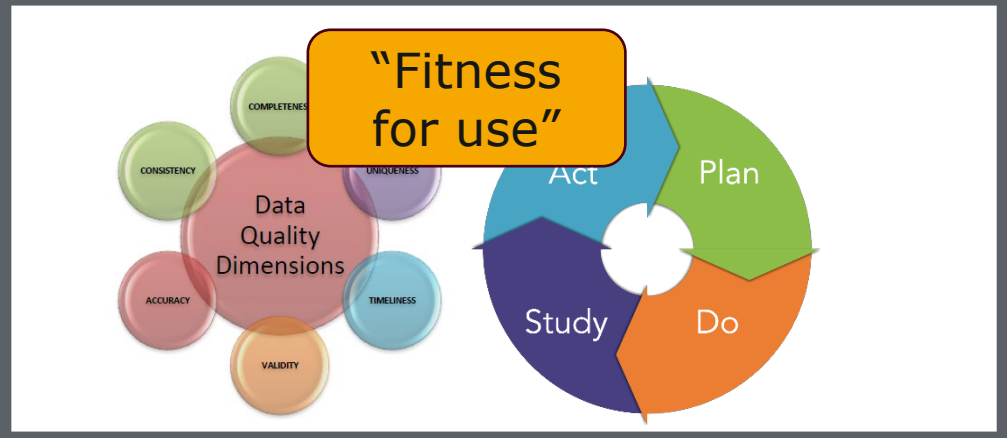
Data Quality



What data scientists think it is



What database admins think it is



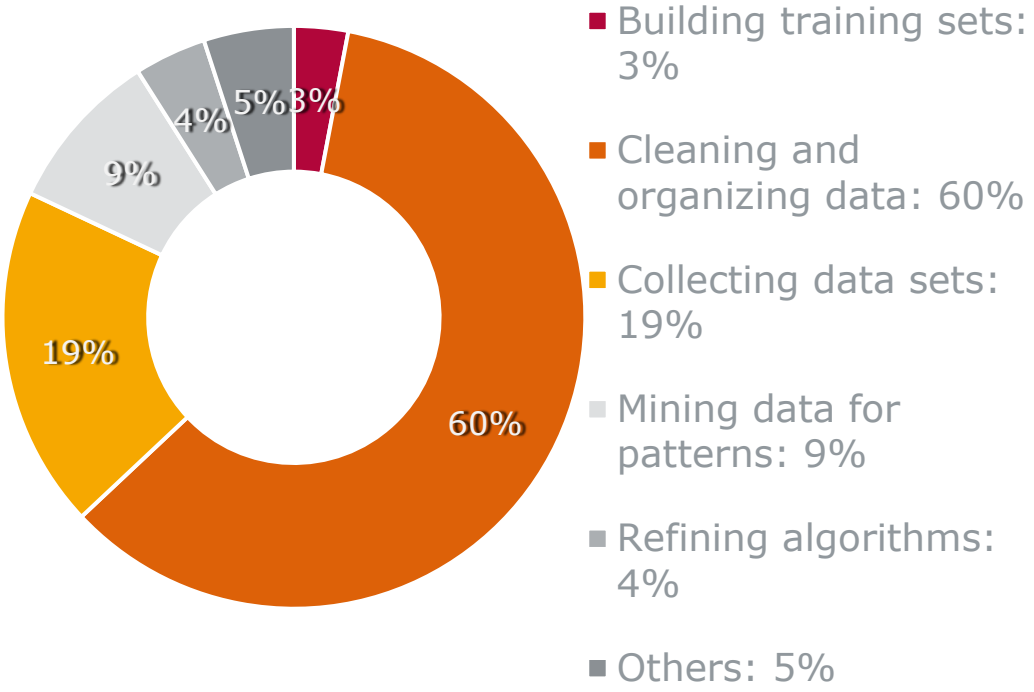
What researchers think it is

Foto: fizkes – shutterstock.com

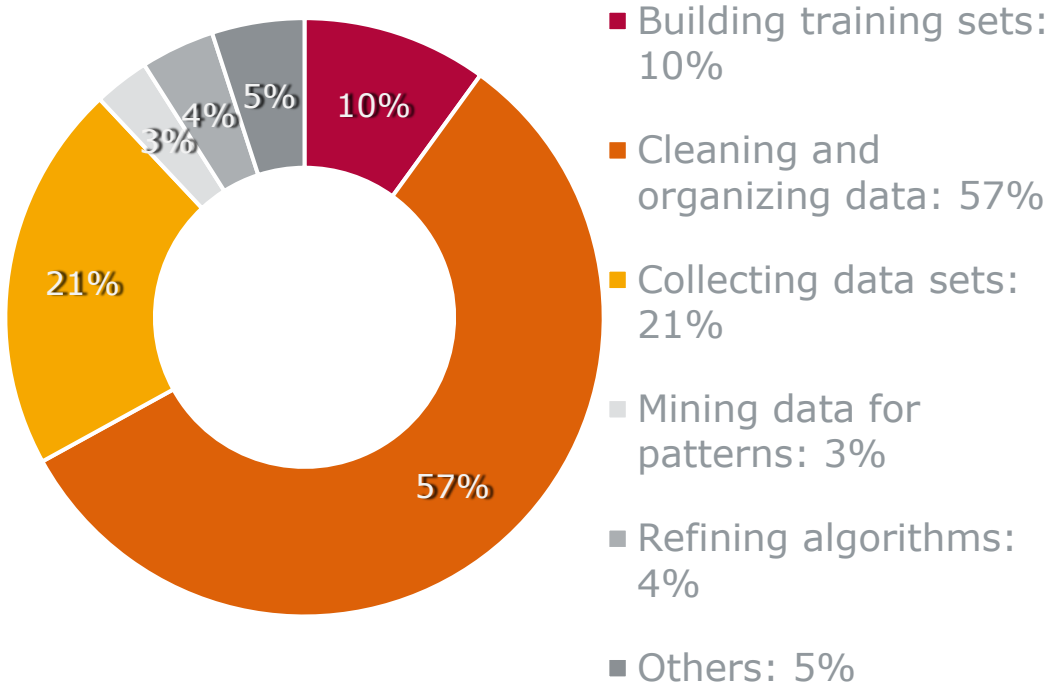
What managers think it is

The Impact of Data Quality

What data scientists spend the most time doing?



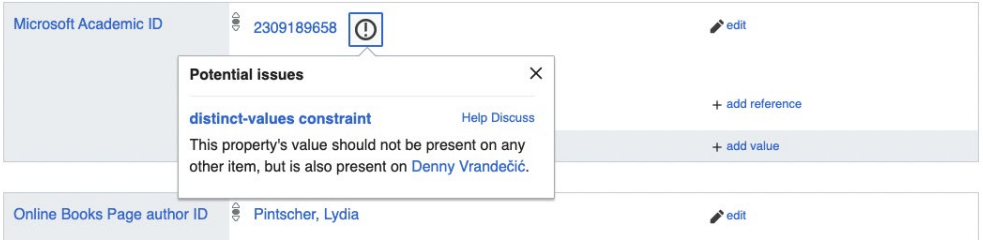
What is the least enjoyable part of data science?



"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

The Impact of Data Quality on Knowledge Graphs

- Classes are difficult to distinguish
  - e.g., "geographical location", "location", "geographic region", "physical location", and "geographical area"
- Constraints are defined but not enforced
  - e.g.:



- Classes and instances are mixed
  - e.g. "scientist" is both a subclass of "researcher" (it's a class) and an instance of "profession" (it's an entity).

State of the Art: Data Quality Dimensions

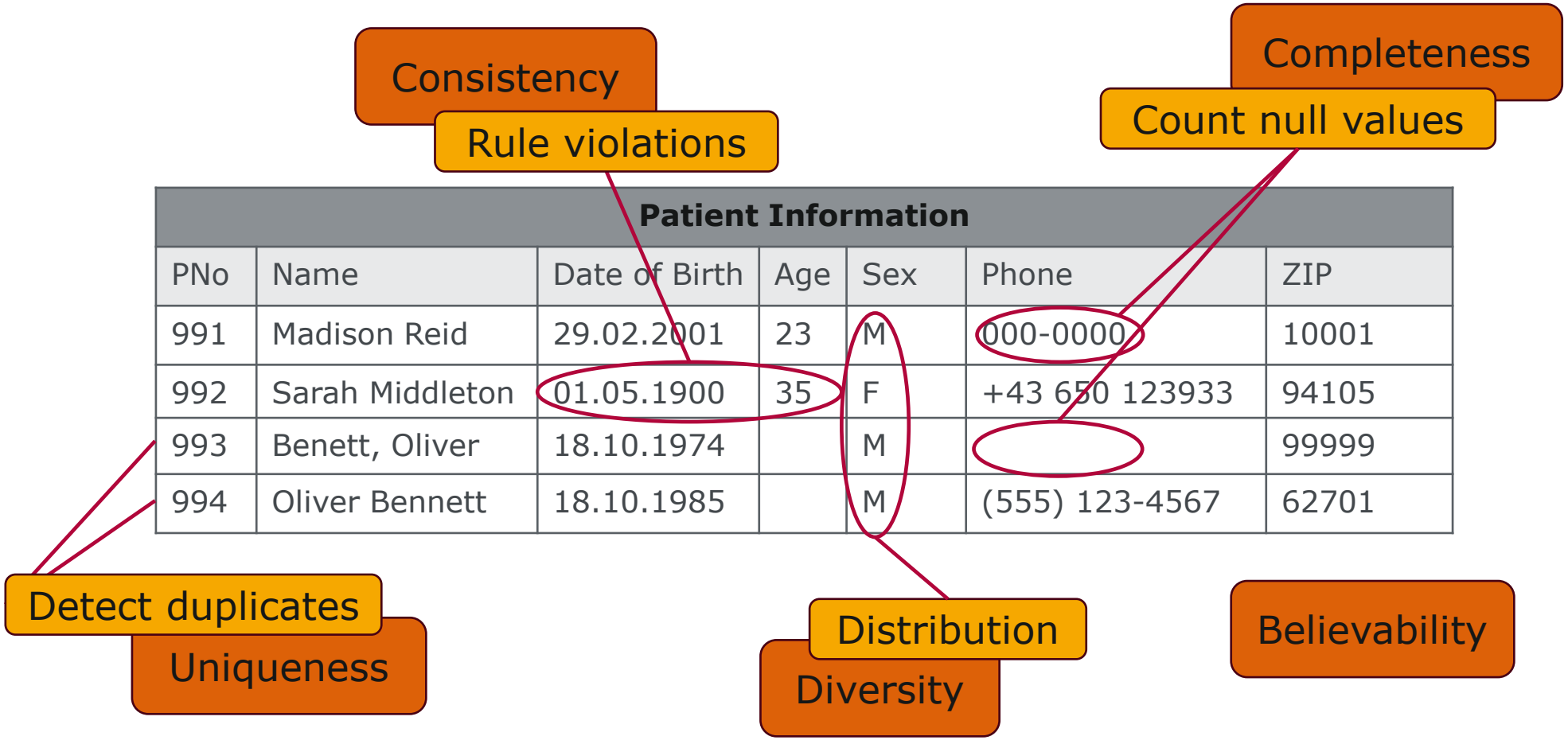
- Data quality research was initiated in the 1980s (Prof. Wang & team from MIT)
  - Definition: "fitness for use"
- DQ is described by **DQ dimensions**, which can refer to
  - The quality of data values
  - The quality of the DB schema
- **DQ metrics** = formulas to quantify a DQ dimension with a numerical value

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		

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

From Data Errors to Metrics for DQ Dimensions



DQ Dimension: Accuracy

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

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

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

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

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

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

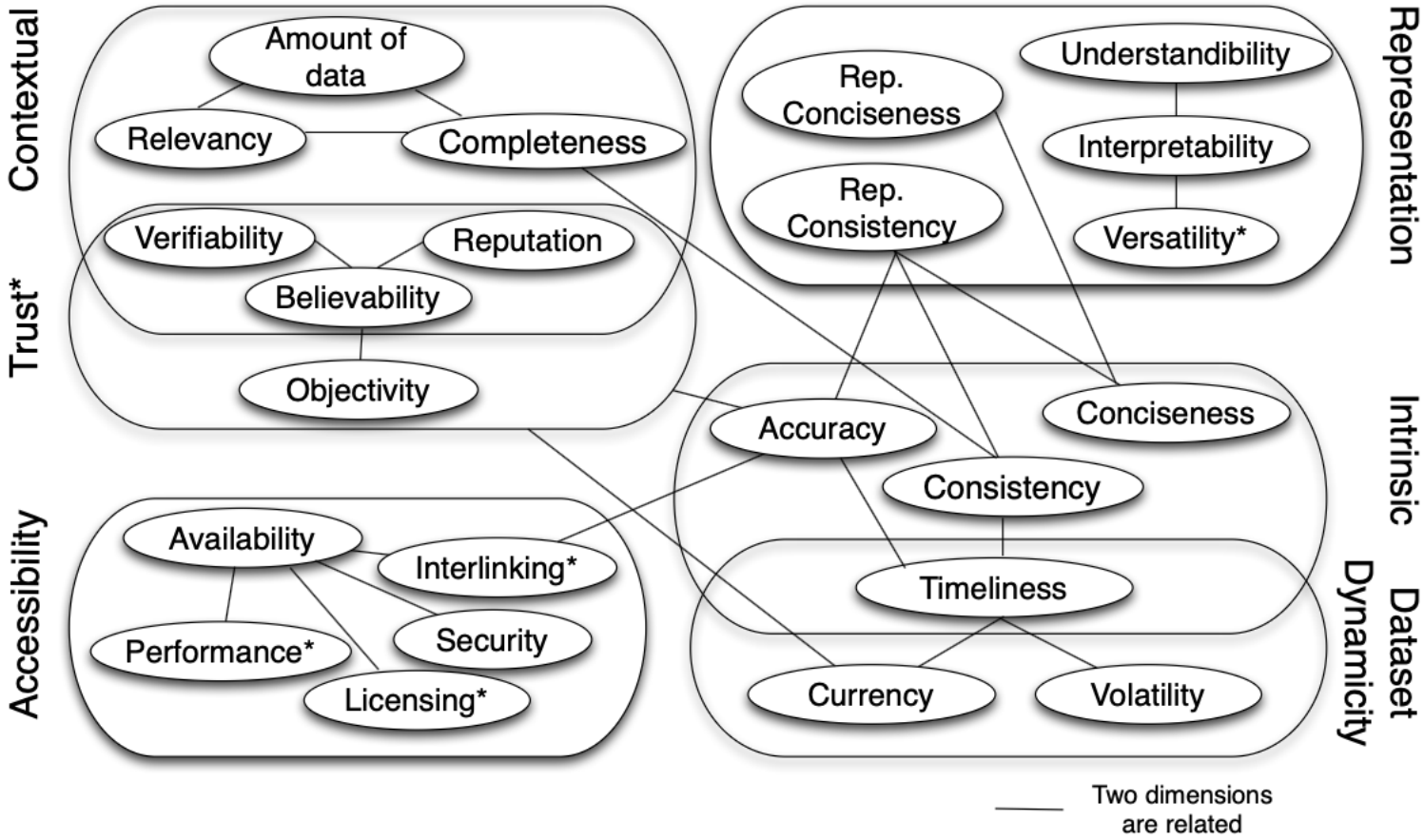
$$AccuracyOfNumericalValues = InaccuracyOfNumericalValues = v' - v$$

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

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

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.

Data Quality Dimensions for Linked Data



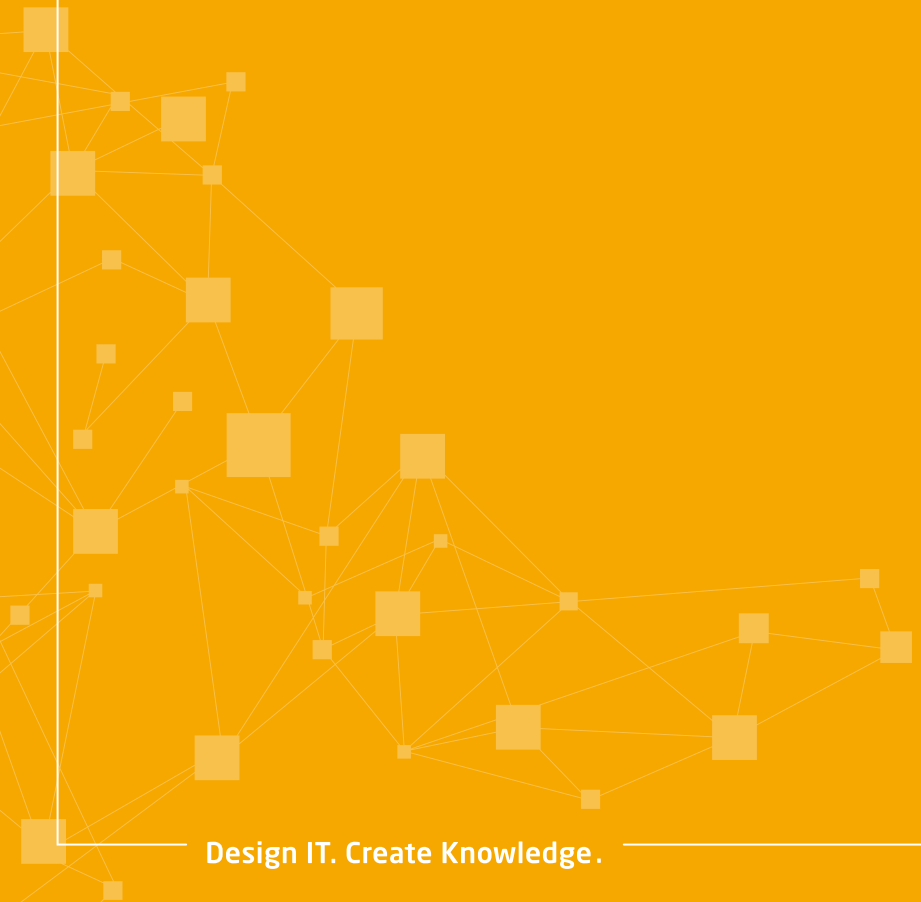
Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2015). Quality assessment for Linked Data: A Survey: A systematic literature review and conceptual framework. *Semantic web*, *7*(1), 63-93.

Lecture about Data Quality

- If you want to dive deep into the **foundations of data quality**
  - Visit our lecture on Tuesdays, 11.00-12.30 in FE.06
  - https://hpi.de/naumann/teaching/current-courses/ws-25-26/data-quality-foundations-vl-msc.html

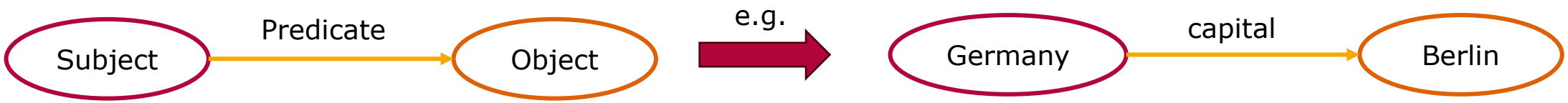
- Learn about different **DQ dimensions** (accuracy, completeness, timeliness, etc.)
- Learn about concrete methods and tools to **detect and correct data errors**
- Learn about methods to **manage DQ** in companies

Technical Background about Knowledge Graphs



Resource Description Framework (RDF)

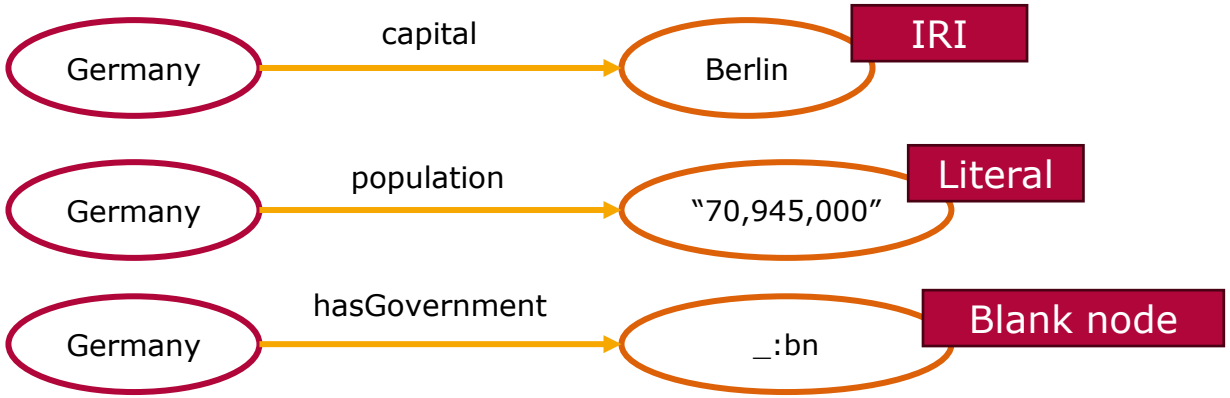
- RDF is a standard model for representing data as graphs [1].
- It expresses information as triples composed of:
  - Subject: the entity being described
  - Predicate: the property or relationship
  - Object: value for the property (IRIs, literals, and blank nodes).
- A graph is represented as a collection of triples.



[1] Wood, D., Lanthaler, M., & Cyganiak, R. (2014, February). RDF 1.1 Concepts and Abstract Syntax. Retrieved from https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/

Resource Description Framework (RDF) – Types of nodes

- **IRIs** (Internationalized Resource Identifiers) – unique names for things (subject, predicate, object)
  - Example: https://www.wikidata.org/entity/Q183 uniquely identifies Germany
  - **Namespaces** – shorthand for IRIs to make notation easier:
    - wd: → https://www.wikidata.org/entity/
    - Using it: wd:Q183 refers to **Germany**
- **Literals** - values like numbers, strings, dates (object)
- **Blank nodes** - anonymous entities (subject, object).

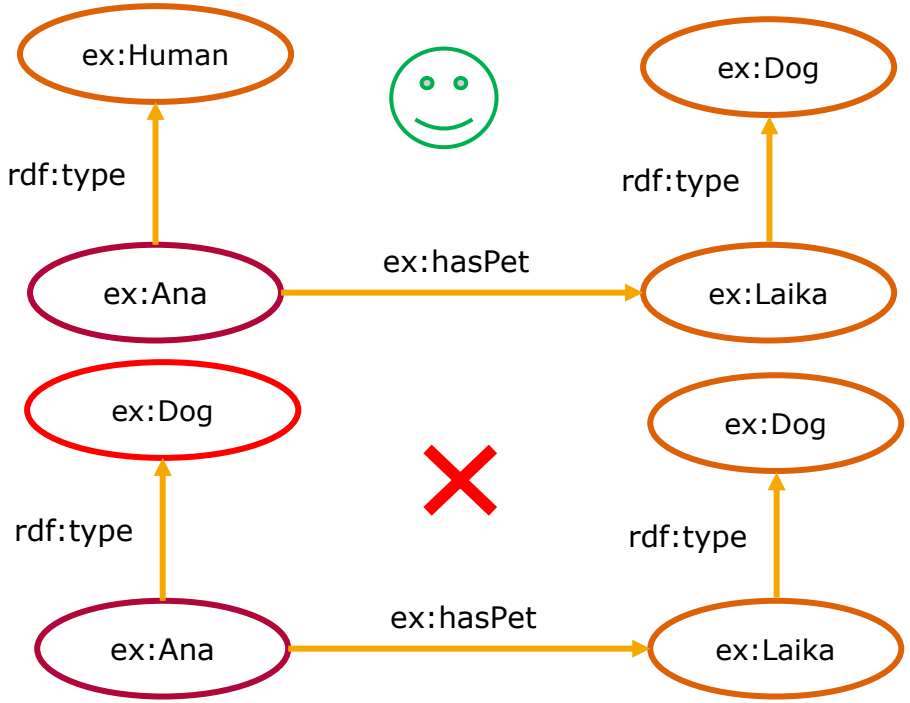


RDF Schema

- RDF Schema provides a vocabulary for RDF data.[1]
  - Classes, Properties
- Reserved properties: rdf:type, rdf:range, rdf:domain, rdfs:subClassOf, rdfs:subPropertyOf, etc.

Examples:

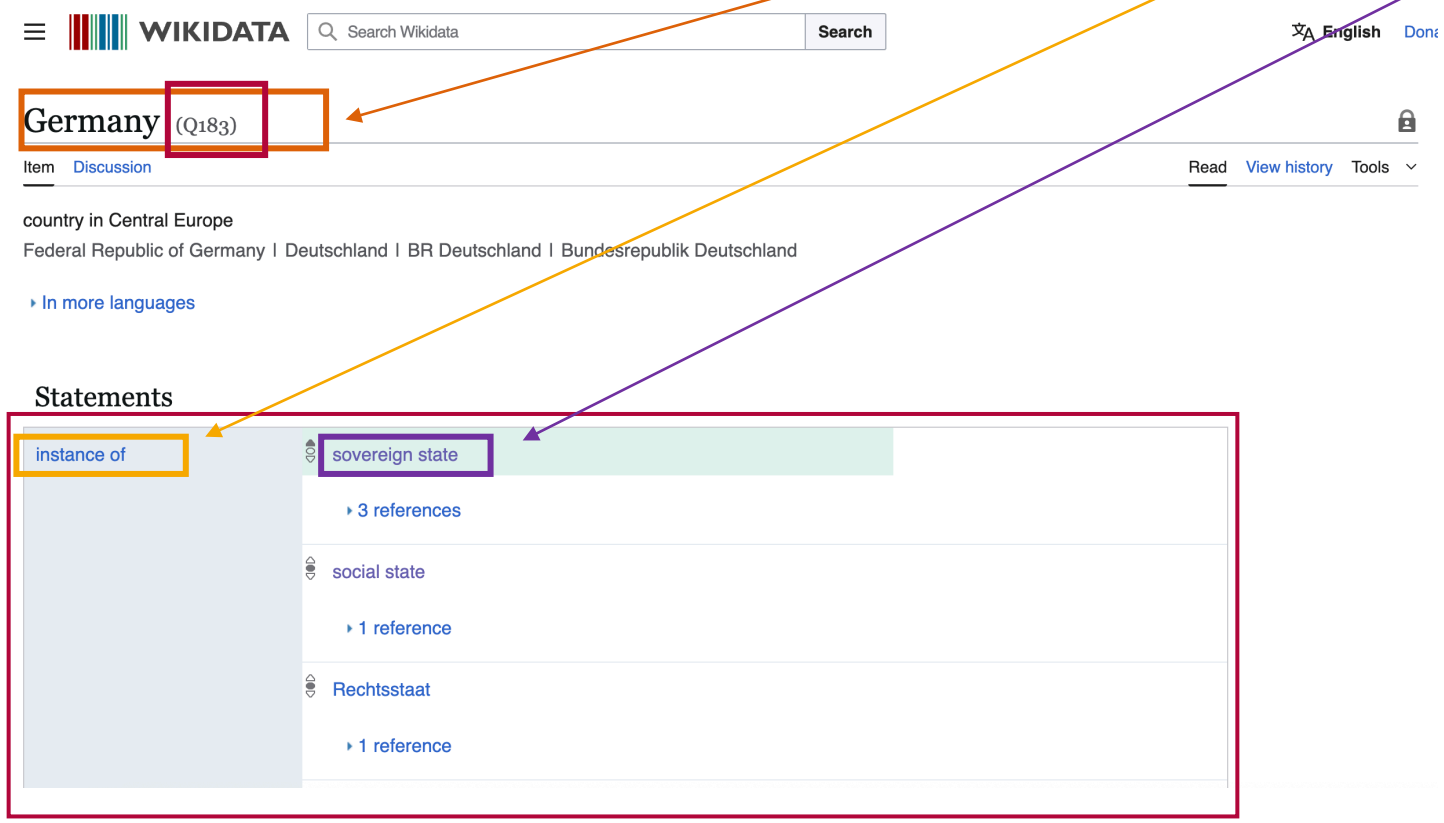
- ex:Germany - rdf:type -> ex:Country
- ex:Germany - rdf:label -> “Germany”
- ex:Dog - rdf:subClassOf -> ex:Animal
- ex:daughter - rdfs:subPropertyOf -> ex:child
- ex:hasPet - rdf:domain -> ex:Human
- ex:hasPet - rdf:range -> ex:Animal



[1] Brickley, D., & Guha, R. V. (2014). RDF Schema 1.1. World Wide Web Consortium (W3C). Retrieved from https://www.w3.org/TR/rdf-schema/

Wikidata – Data Model

- Knowledge about entities is represented via **statements:** <subject, predicate, object> (RDF model)**.**

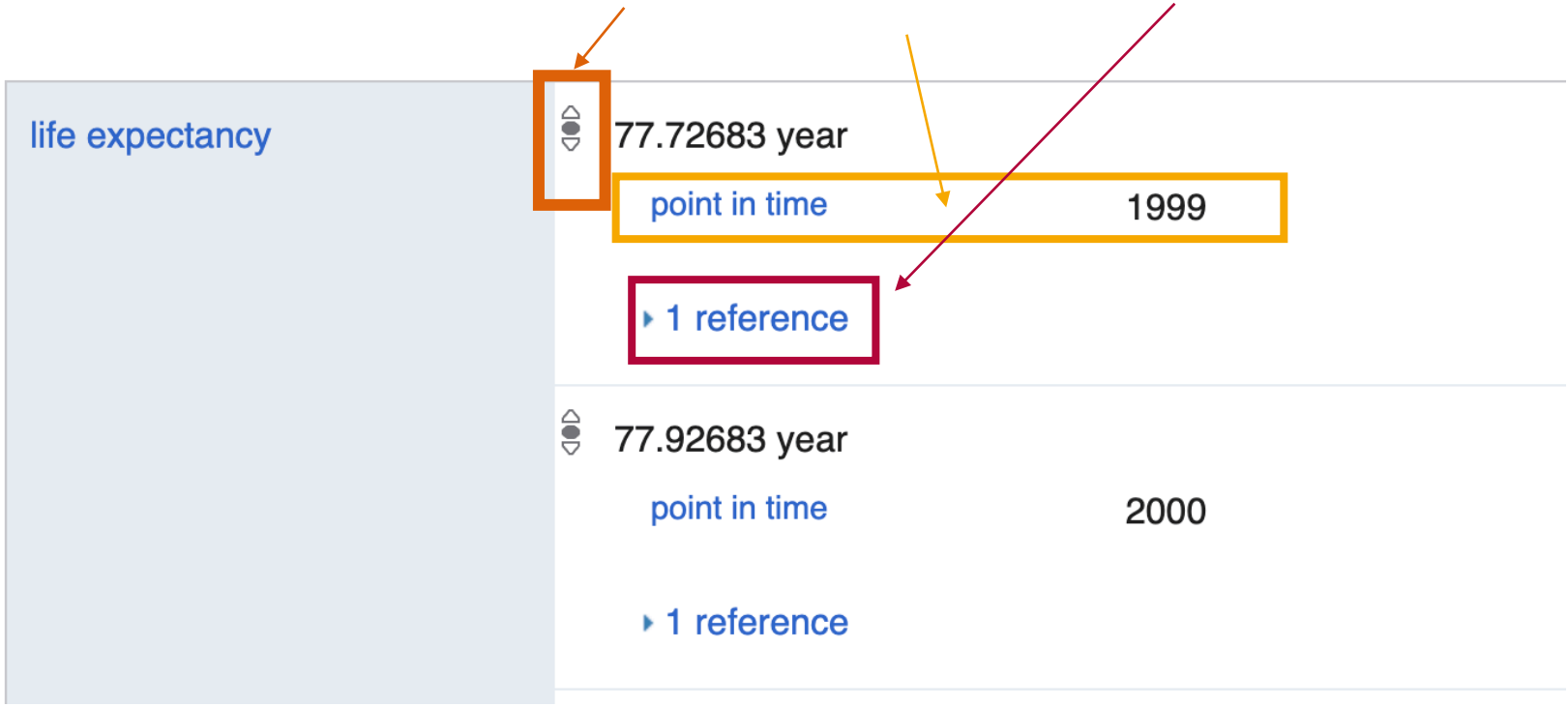


Wikidata – Data Model

- Wikidata follows the RDF data model.
- Real-world entities are called **Items**, identified via Q-ids (e.g. Germany - Q183)
  - *Q-ids* in Wikidata ←→ *IRIs* in RDF
- Every Item has a **page** that contains all the statements about the Item.
- Knowledge about entities is represented via **statements:** <subject, predicate, object> where the subject is always an **Item.**
- Objects can be other **Items** or **Literals**.
  - Wikidata has a pre-defined set of datatypes for literals (e.g. string, quantity, timestamp, etc.).
- Special properties:
  - *P31* (instance of) in Wikidata ←→ *rdf:type* in RDF
  - *P279* (subclass of) in Wikidata ←→ *rdfs:subClassOf* in RDF

Wikidata – Data Model

- Statements can also have *ranks*, *qualifiers* and *references*.



Project Organization

Project Goals

We want to **assess the quality of Wikidata over time**.

1. You will get the code for the parser to extract changes from Wikidata's dump files
2. **Implement a scalable architecture** for processing a Wikidata dump (distributed?)
3. **Select a subset** of entities to work with
4. **Extract basic statistics** from the dataset
5. **Efficiently calculate and store** data quality measures over time
6. **Develop a dashboard** to visualize the development of data quality measures over time
7. **Prepare a submission** to a top database conference

Your Next Tasks

- Organize yourself as a team (e.g., define tasks and roles)
  - Meet 2 days per week to work on the project! (12 ECTS)
- Read literature about data quality in Wikidata (also beyond the papers provided)
  - We will have a “how to read a research paper” seminar next week
- Inspect change extraction code + documentation (Change extraction doc)
  - GitLab repo: https://gitlab.hpi.de/fg-naumann/teaching/mp2025-wikiwatch

- Further
  - Design a scalable architecture to parse all files
  - Identify interesting DQ dimensions and metrics

Master Project Organization

- Project room: F-2.05 (Campus II)
- Project communication:
  - HPI slack channel: #mp2025-wikiwatch
  - Also contact us via e-mail: {carolina.cortes, lisa.ehrlinger}@hpi.de
  - Or drop by at our office: F-2.04 (Campus II)
  - Information on website: https://hpi.de/naumann/teaching/current-courses/ws-25-26/masters-project-wikiwatch.html

- To define: weekly project meetings (in person)
  - Mondays 14.00-16.00? (for 1h)
- "How to read a research paper?" – Combined session with seminar
  - Tuesday 21st of October (second week of the semester) - 15:15-16:45

Deliverables

- Code + experiments (via GIT repo)

- Scientific paper summarizing the project results, containing
  - Problem statement
  - Description of developed approach
  - Experimental evaluation

- We want to publish the paper in a scientific conference
  - Overleaf project: https://www.overleaf.com/project/68c914a45661494efb75752b

Related Work

- Bleifuß, T., Bornemann, L., Johnson, T., Kalashnikov, D. V., Naumann, F., & Srivastava, D. (2018). Exploring change: A new dimension of data analytics. *Proceedings of the VLDB Endowment*, 12 (2), 85-98.
- Mohammed, S., Ehrlinger, L., Harmouch, H., Naumann, F., & Srivastava, D. (2025). The Five Facets of Data Quality Assessment. *ACM SIGMOD Record*, *54*(2), 18–27. https://doi.org/10.1145/3749116.3749120
- Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2012). Quality assessment for Linked Data: A Survey. *Semantic Web*, *7*(1), 63–93. https://doi.org/10.3233/SW-150175
- Piscopo, A., & Simperl, E. (2019). What we talk about when we talk about Wikidata quality: A literature survey. *Proceedings of the 15th International Symposium on Open Collaboration, OpenSym 2019*. https://doi.org/10.1145/3306446.3340822
- Rashid, M., Torchiano, M., Rizzo, G., Mihindukulasooriya, N., & Corcho, O. (2019). A quality assessment approach for evolving knowledge bases. *Semantic Web*, *10*(2), 349–383. https://doi.org/10.3233/SW-180324
- Keshav, S. (2007). How to read a paper. ACM SIGCOMM Computer Communication Review, 37(3), 83-84.
- 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.

More: Related work

Questions?