Chair of
Databases and Information Systems

GouDa – Generation of universal Data Sets

Valerie Restat, Gerrit Boerner, André Conrad, Uta Störl

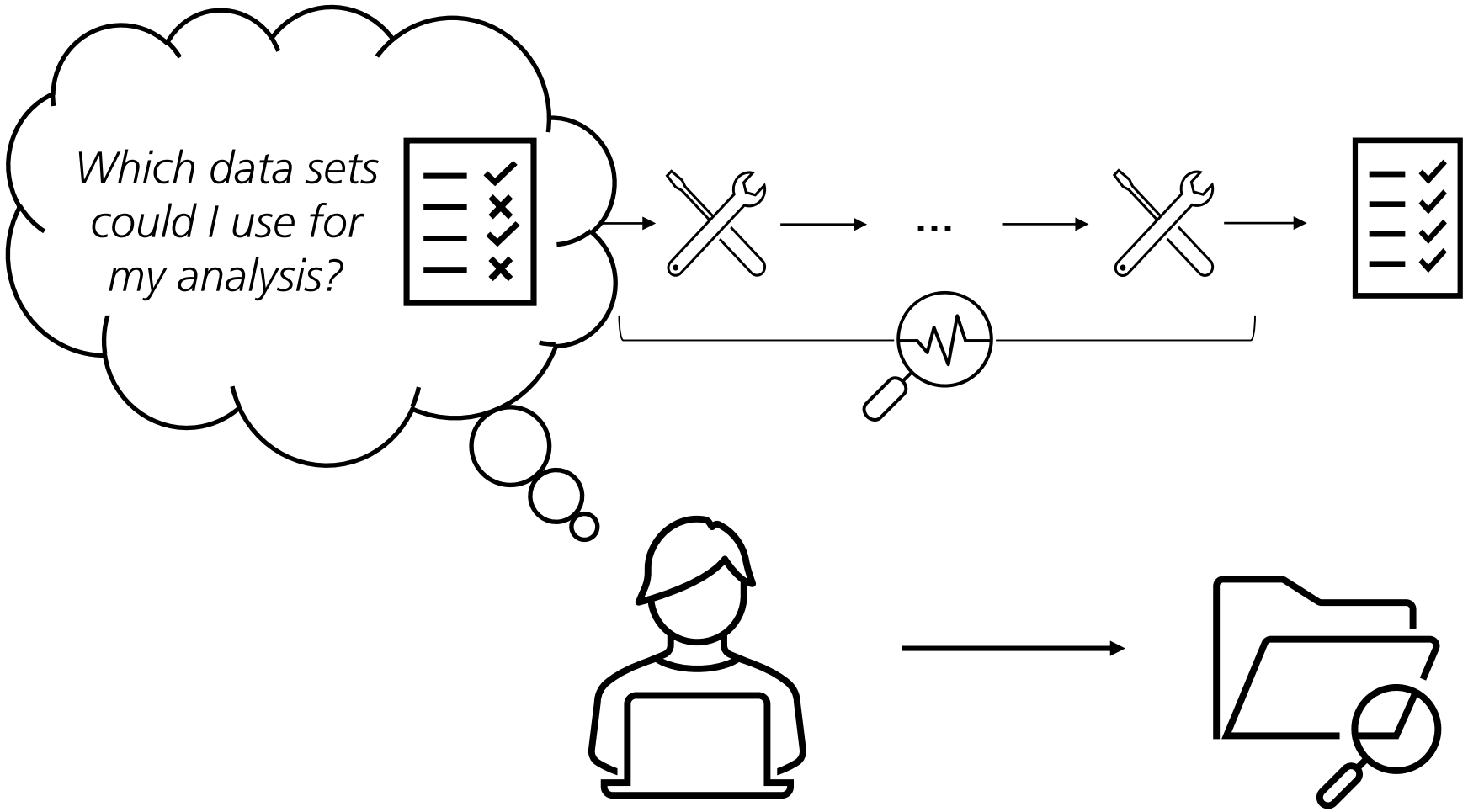
University of Hagen

12.06.2022, DEEM-Workshop

FernUniversität in Hagen

Content

Analysis of Data Preparation Pipelines



Data Sets

Problem Statement 1

Not publicly available

Missing ground truth

Problem Statement 2

Error types

Error rates

Size of data sets

GouDa – Test data generator

- Publicly available
- Ground truth provided
- Diverse error types
- Arbitrary error rates
- Scalable

Data Generators

- Different data generators exist: Software testing, SQL queries, …
- Tools for generating data sets with defined errors: BART [Arocena et al. (2015)]
  (Benchmarking Algorithms for data Repairing and Translation)

Errors are inserted into already existing, clean databases

Requires a relational database schema with unique tuple identifiers

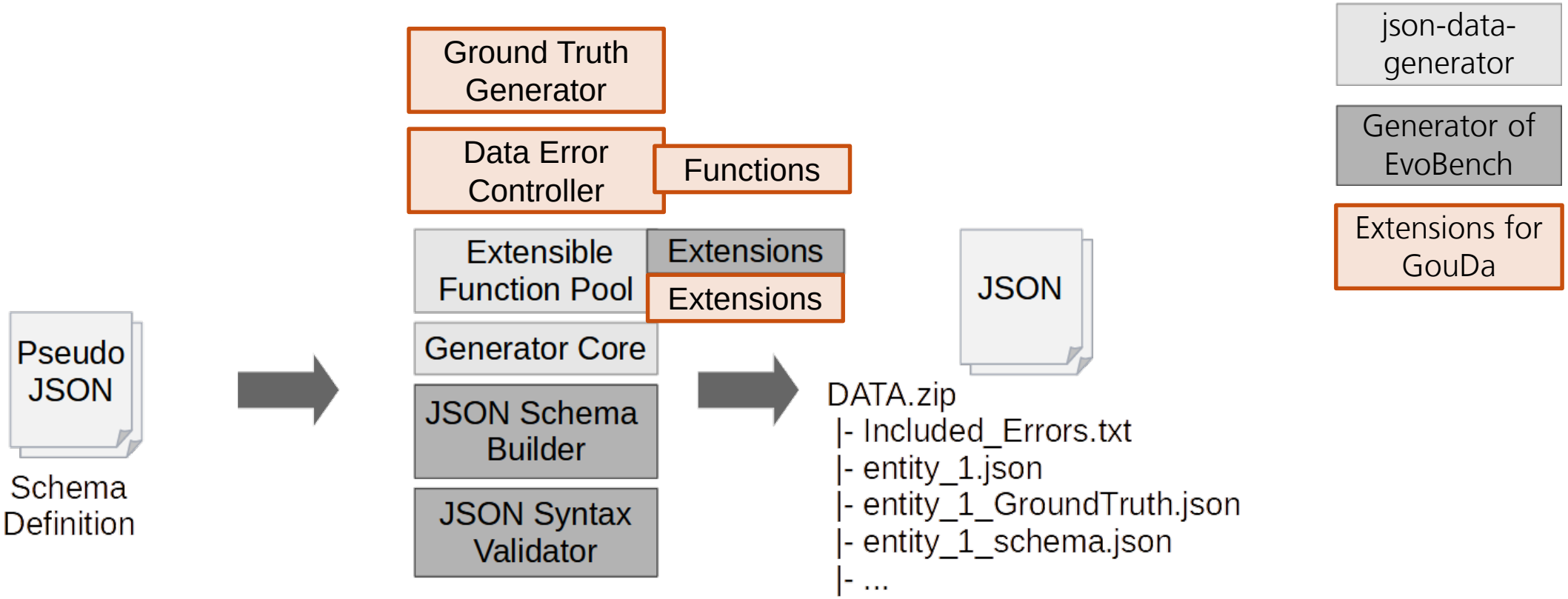
Focusing on constraint-induced errors

No existing data needed

Uses JSON format, not only relational data but also nested data is supported

Supports variety of different error types

GouDa – Components



Conrad et al. (2021) *EvoBench: Benchmarking Schema Evolution in NoSQL.*,
- based on json-data-generator: https://github.com/vincentrussell/json-data-generator (v1.12)

GouDa – Characteristics

Error rate

- *Freely configurable occurrence rate*
- *Errors are randomly distributed*

Reproducible

Scalable

Portable

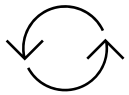
Realistic

Ground truth

Error types

GouDa – Characteristics

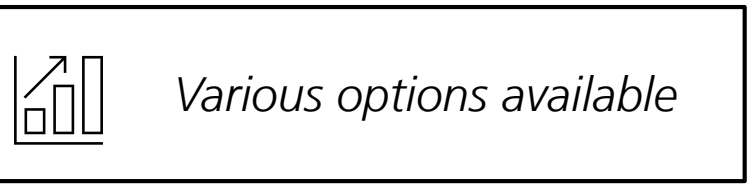
- Error rate
- **Reproducible**
- Scalable
- Portable
- Realistic
- Ground truth
- Error types



Same type of error can be generated on the same attribute of the same tuple

GouDa – Characteristics

- Error rate
- Reproducible
- **Scalable**
- Portable
- Realistic
- Ground truth
- Error types



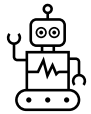
GouDa – Characteristics

- Error rate
- Reproducible
- Scalable
- **Portable**
- Realistic
- Ground truth
- Error types

{ } *Data sets are generated in JSON format*

GouDa – Characteristics

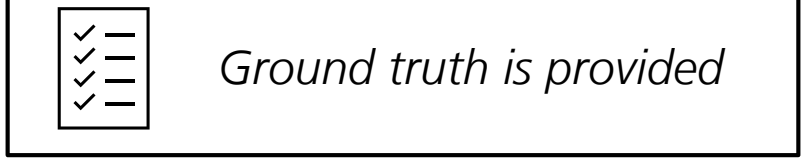
- Error rate
- Reproducible
- Scalable
- Portable
- **Realistic**
- Ground truth
- Error types



Different lists with realistic names, words or addresses are included

GouDa – Characteristics

- Error rate
- Reproducible
- Scalable
- Portable
- Realistic
- **Ground truth**
- Error types



GouDa – Characteristics

Error rate

Reproducible

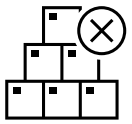
Scalable

Portable

Realistic

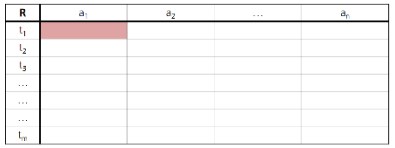
Ground truth

Error types

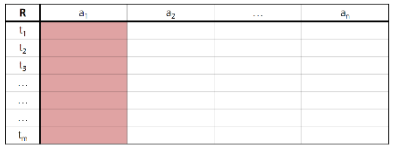


Variety of different error types

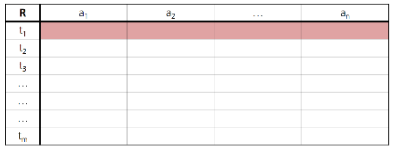
GouDa – Error types



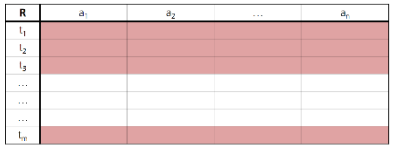
An Attribute Value of a Single Tuple



The Values of a Single Attribute

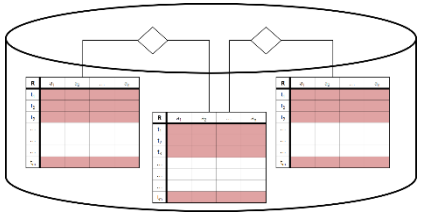


The Attribute Values of a Single Tuple

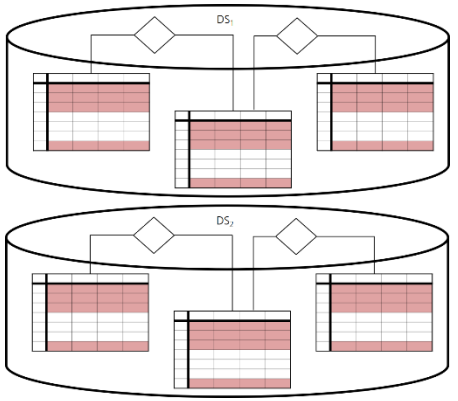


The Attribute Values of Several Tuple

Single Relation



Relationships among Multiple Relations



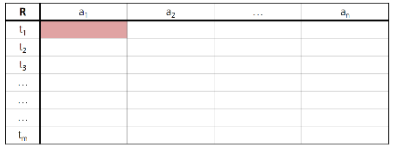
Multiple Data Sources

Single Source

Data Quality Problems

Oliveira et al. (2005)
A taxonomy of data quality problems

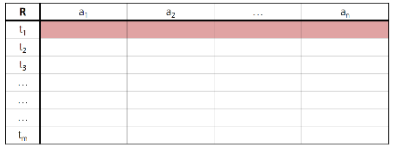
GouDa – Error types



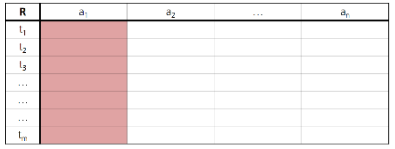
An Attribute Value of a Single Tuple

- Missing value
- Syntax violation
- Interval violation
- Set violation
- Misspelled error
- Inadequate value to the attribute context
- Value items beyond the attribute context
- Meaningless Value
- Erroneous entry*

- Semi-empty tuple
- Inconsistency among attribute values
- Irrelevant observation*



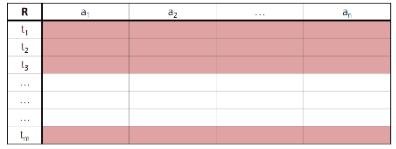
The Attribute Values of a Single Tuple



The Values of a Single Attribute

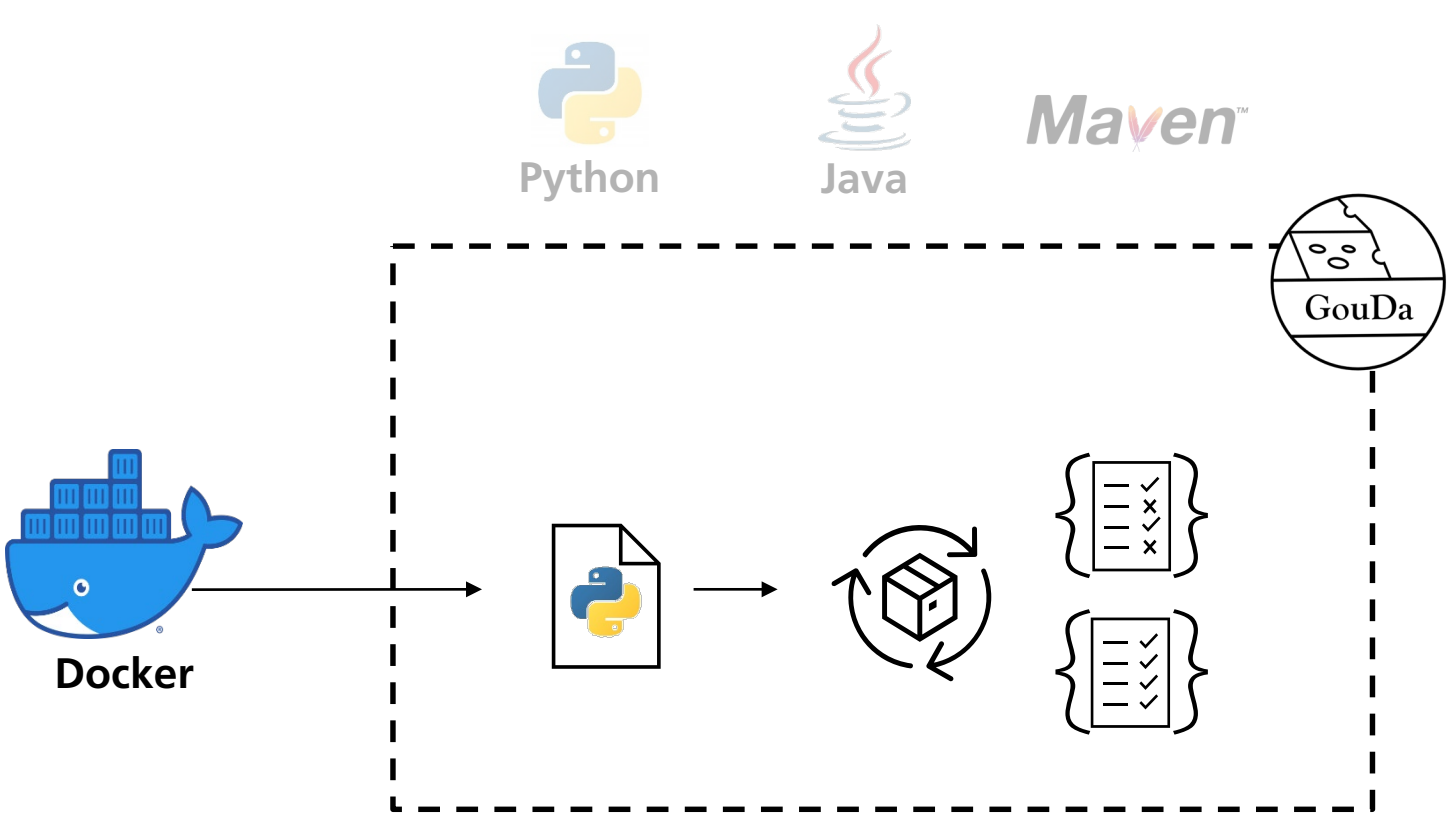
- Uniqueness value violation
- Synonyms existence
- Outlier*
- Missing Attribute*

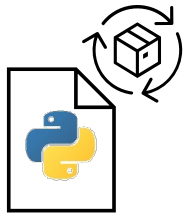
- Redundancy about an entity
- Inconsistency about an entity
- Bias*
- Noise*



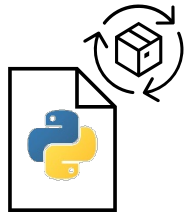
The Attribute Values of Several Tuple

GouDa – Usage

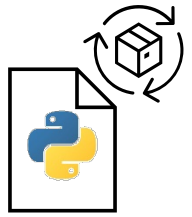




Simple schema used for testing of individual functions



Extensive schema in which all created functions are implemented



Replica of the schema described in *Goldberg et al. (2011)*

https://zenodo.org/record/6610025

GouDa – Usage

10000 data objects

```
['{{repeat(10000)}}', {
  "Index":{{relationship("Performance_Test-id-idx","INCREMENT","1")}},
  "City":"{{ city() }}",
  "Date":"{{ error("SYNTAX",5,"date","dd.MM.yyyy",date("08-05-1945
00:00:00","03-10-1990 00:00:00","dd.MM.yyyy")) }}",
  "Count":"{{ error("INTERVALLVIOLATION",5,0,9999,integer(0,9999)) }}"
}]
```

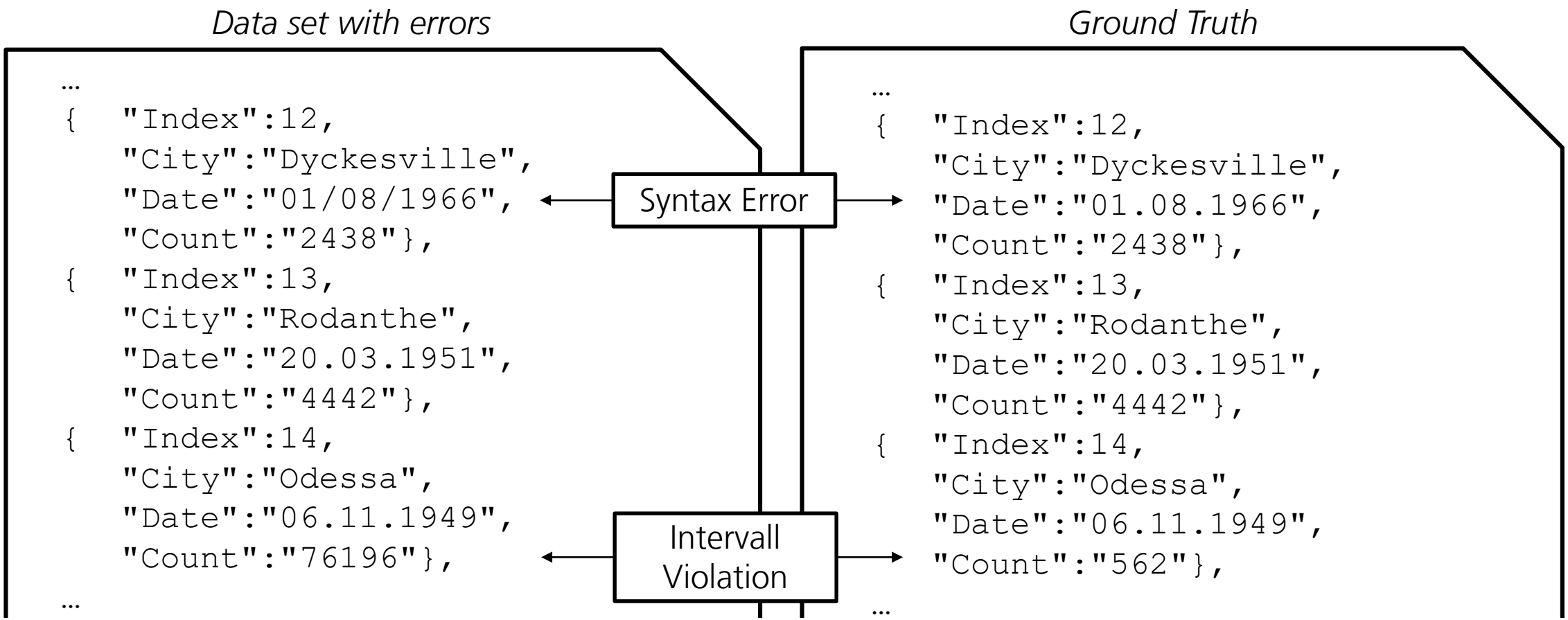
Incremental Index

Function for city names, no errors

Date function, 5% syntax error

Integer range with 5% intervall violation

GouDa – Usage


Data set with errors
Ground Truth
…
{ "Index":12,
"City":"Dyckesville",
"Date":"01/08/1966",
"Count":"2438"},
{ "Index":13,
"City":"Rodanthe",
"Date":"20.03.1951",
"Count":"4442"},
{ "Index":14,
"City":"Odessa",
"Date":"06.11.1949",
"Count":"76196"},
…
Syntax Error
Intervall Violation
…
{ "Index":12,
"City":"Dyckesville",
"Date":"01.08.1966",
"Count":"2438"},
{ "Index":13,
"City":"Rodanthe",
"Date":"20.03.1951",
"Count":"4442"},
{ "Index":14,
"City":"Odessa",
"Date":"06.11.1949",
"Count":"562"},
…

Conclusion and Future Work

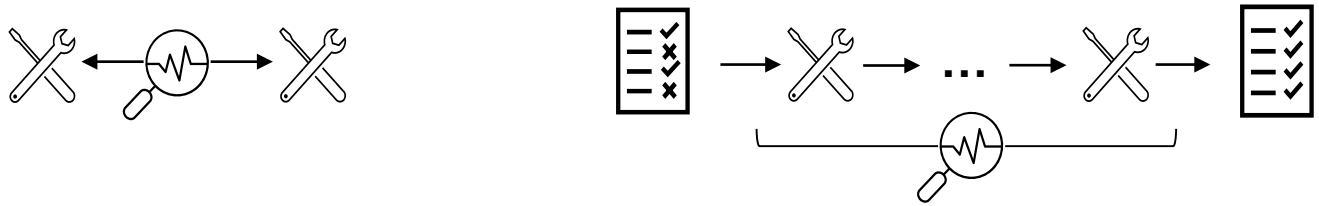
We have presented **GouDa**, our test data generator

Diverse error types

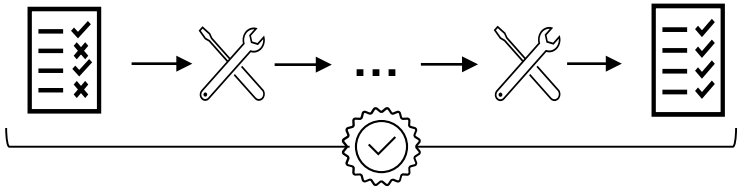
Arbitrary error rates

Provides ground truth

Which we currently use for: **Analysis** of data preparation tools and pipelines



And our next research topics: **Data quality** and **evaluation** of data preparation pipelines



GouDa – Generation of universal Data Sets

https://zenodo.org/record/6610025

Valerie Restat
valerie.restat@fernuni-hagen.de

Gerrit Boerner
gerrit.boerner@studium.fernuni-hagen.de

André Conrad
andre.conrad@fernuni-hagen.de

Uta Störl
uta.stoerl@fernuni-hagen.de

FernUniversität in Hagen

Sources

- Patricia C. Arocena et al . 2015. Messing Up with BART: Error Generation for Evaluating Data-Cleaning Algorithms. Proc. VLDB Endow. (2015), 36–47. https://doi.org/10.14778/2850578.285057
- Paulo Oliveira et al. 2005. A taxonomy of data quality problems. In 2nd Int. Workshop on Data and Information Quality. 219–233.
- Saveli Goldberg, Andrzej Niemierko, and Alexander Turchin. 2008. Analysis of Data Errors in Clinical Research Databases. In AMIA 2008, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 8-12, 2008. AMIA. https://knowledge.amia.org/amia-55142-a2008a-1.625176/t-001-1.626020/f-001-1.626021/a-049-1.626417/a-050-1.6264