UNIVERSITÄT LEIPZIG

Medizinische Fakultät

Universitätsklinikum Leipzig

Medizin ist unsere Berufung.

8th Freiberger PhD Conference

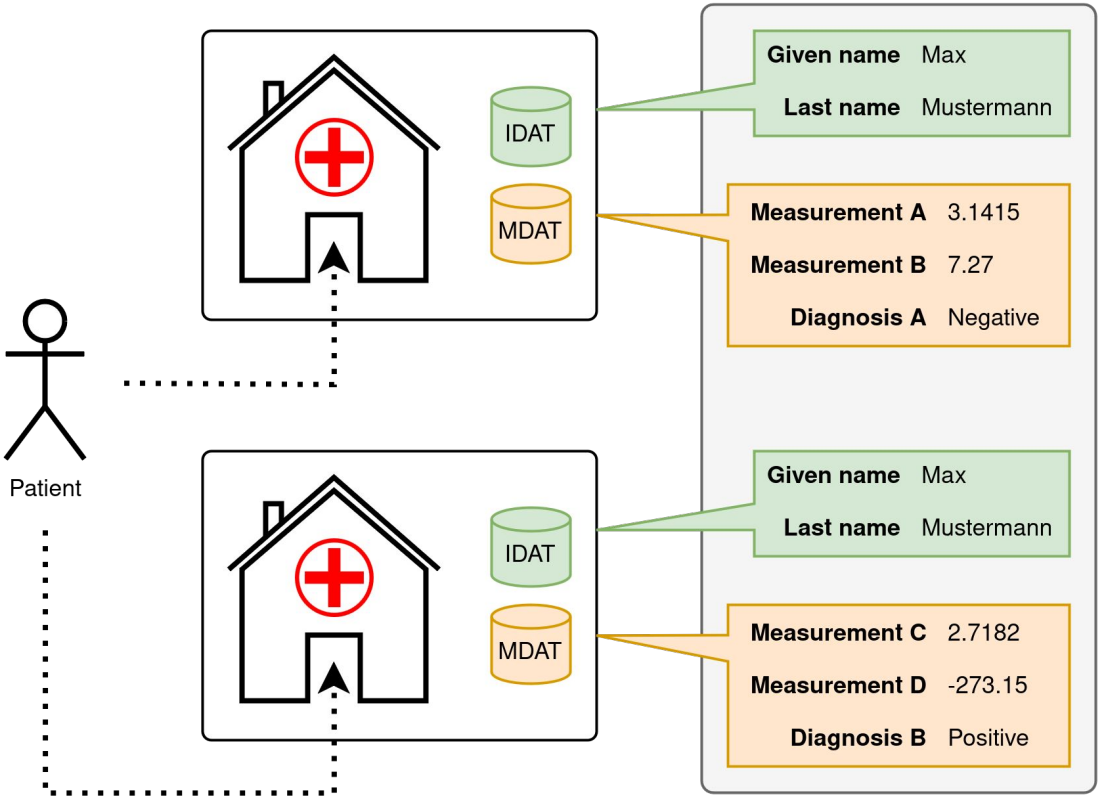
Gecko: Generation and Mutation of Realistic Identification Data at Scale for Record Linkage Evaluation

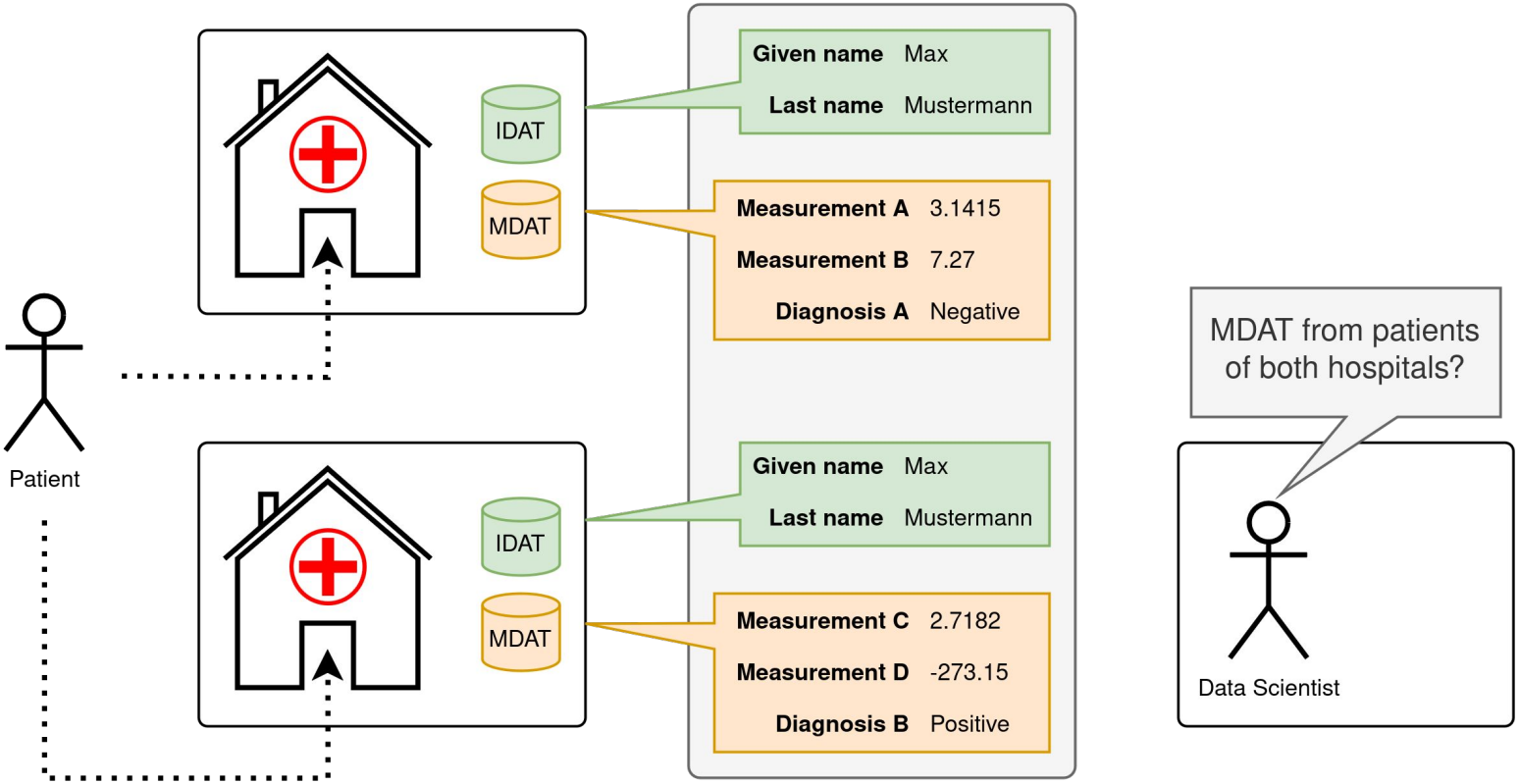
Freiberg, 07.06.2024

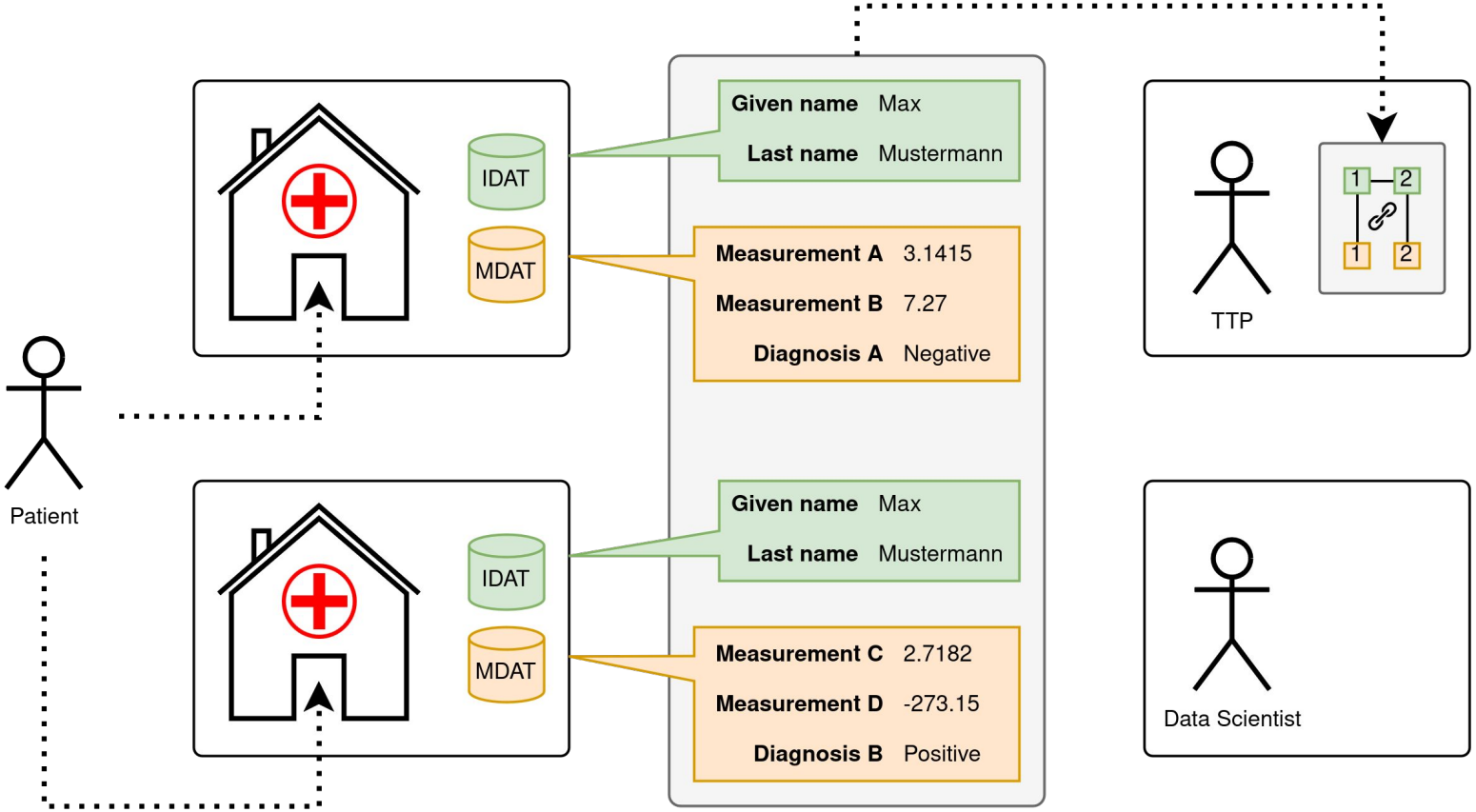
Maximilian Jugl

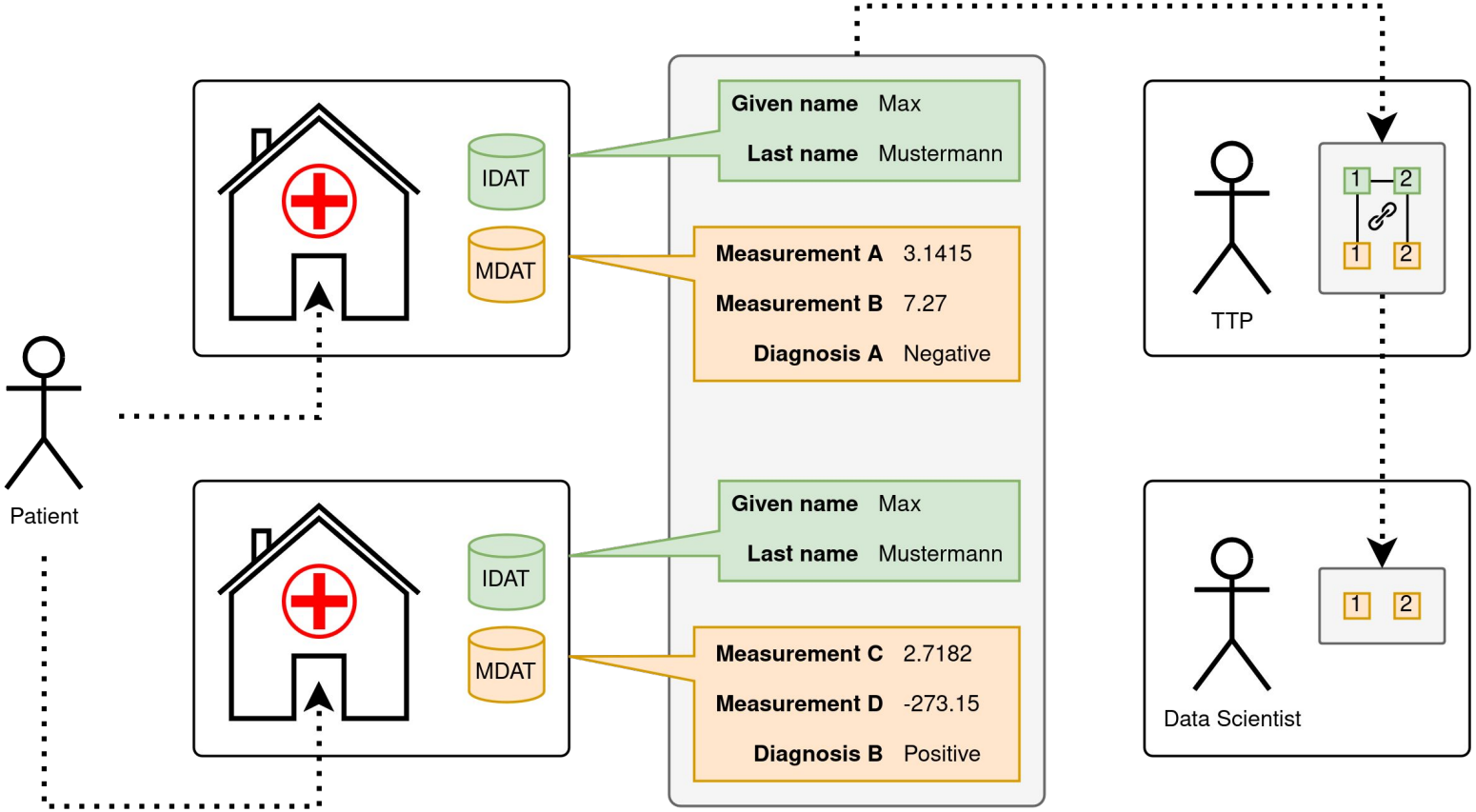
MDS

Medical Data Science

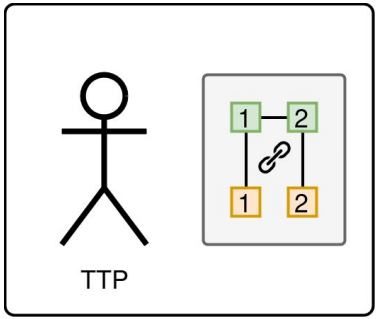

IDAT
MDAT
Patient
IDAT
MDAT
Given name Max
Last name Mustermann
Measurement A 3.1415
Measurement B 7.27
Diagnosis A Negative
Given name Max
Last name Mustermann
Measurement C 2.7182
Measurement D -273.15
Diagnosis B Positive


IDAT
MDAT
Patient
IDAT
MDAT
Given name Max
Last name Mustermann
Measurement A 3.1415
Measurement B 7.27
Diagnosis A Negative
Given name Max
Last name Mustermann
Measurement C 2.7182
Measurement D -273.15
Diagnosis B Positive
MDAT from patients of both hospitals?
Data Scientist


IDAT
MDAT
Given name Max
Last name Mustermann
Measurement A 3.1415
Measurement B 7.27
Diagnosis A Negative
TTP
1
2
1
2
Patient
IDAT
MDAT
Given name Max
Last name Mustermann
Measurement C 2.7182
Measurement D -273.15
Diagnosis B Positive
Data Scientist


IDAT
MDAT
IDAT
MDAT
Patient
Given name Max
Last name Mustermann
Measurement A 3.1415
Measurement B 7.27
Diagnosis A Negative
Given name Max
Last name Mustermann
Measurement C 2.7182
Measurement D -273.15
Diagnosis B Positive
TTP
1 2
1 2
1 2
Data Scientist

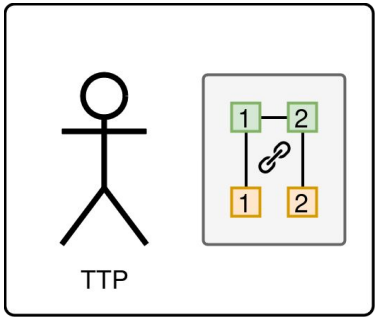
A PRIMER ON RECORD LINKAGE



- Testing of various record linkage strategies
- Testing against error sources and varying data schemas
- Limited access to real-world data

⇒ **Generation of realistic-looking test data**

A PRIMER ON RECORD LINKAGE

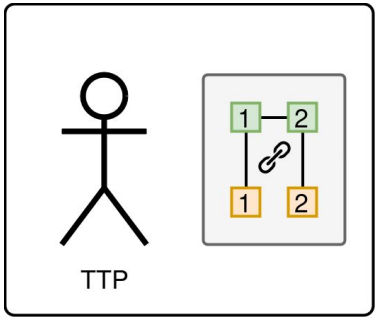


- Testing of various record linkage strategies
- Testing against error sources and varying data schemas
- Limited access to real-world data

⇒ **Generation of realistic-looking test data**

Given name	Last name	Gender	Date of birth
Axel	Schweiss	Male	1981-01-10
Grube	Claire		1970-10-01
Anna	Kond4	Female	1991-08-02

A PRIMER ON RECORD LINKAGE



- Testing of various record linkage strategies
- Testing against error sources and varying data schemas
- Limited access to real-world data

⇒ **Generation of realistic-looking test data**

Given name	Last name	Gender	Date of birth
Axel	Schweiss	Male	1981-01-10
Grube	Claire		1970-10-01
Anna	Kond4	Female	1991-08-02

Typos?

Flipped values?

OCR errors?

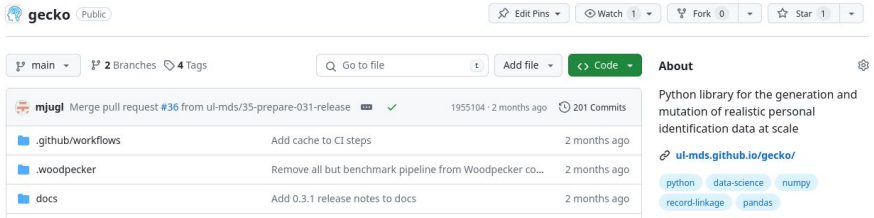
Ambiguous format?

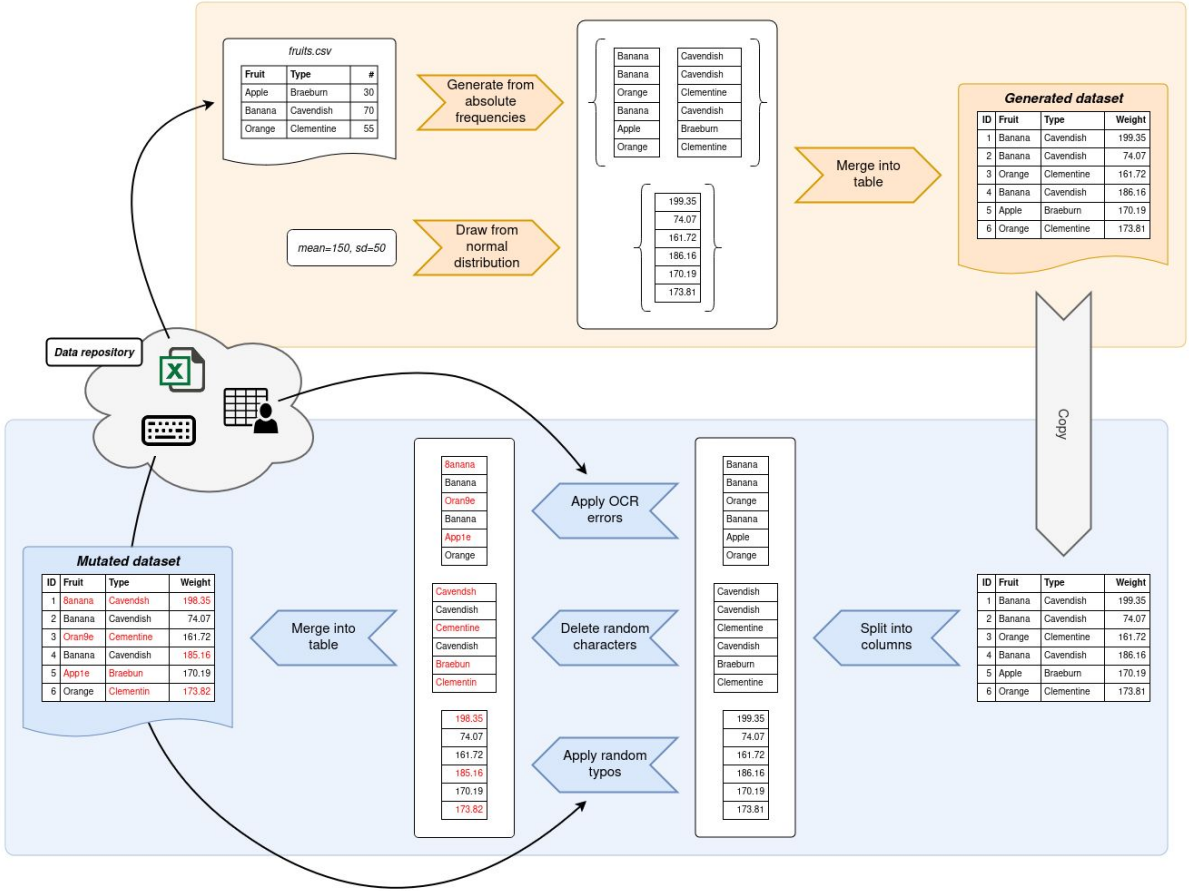
Missing values?

PRESENTING GECKO

- Modern approach to the ideas put forward by GeCo (Tran et al. 2013)
  - Completely reworked from the ground up for modern Python
  - Based on NumPy and Pandas to integrate into data science applications
  - Domain-independent, highly configurable and scalable

- Source code: https://github.com/ul-mds/gecko
- Documentation: https://ul-mds.github.io/gecko/
- Python Package Index: https://pypi.org/project/gecko-syndata/

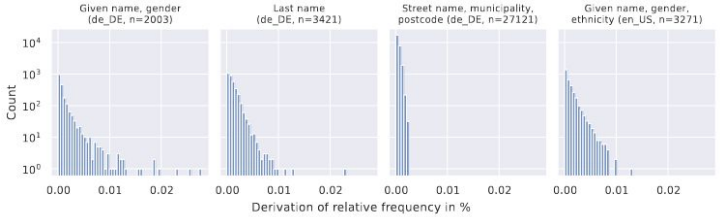



fruits.csv
Fruit | Type | #
Apple | Braeburn | 30
Banana | Cavendish | 70
Orange | Clementine | 55
Generate from absolute frequencies
Banana | Cavendish
Banana | Cavendish
Orange | Clementine
Banana | Cavendish
Apple | Braeburn
Orange | Clementine
mean=150, sd=50
Draw from normal distribution
199.35
74.07
161.72
186.16
170.19
173.81
Merge into table
Generated dataset
ID | Fruit | Type | Weight
1 | Banana | Cavendish | 199.35
2 | Banana | Cavendish | 74.07
3 | Orange | Clementine | 161.72
4 | Banana | Cavendish | 186.16
5 | Apple | Braeburn | 170.19
6 | Orange | Clementine | 173.81
Copy
Data repository
Mutated dataset
ID | Fruit | Type | Weight
1 | 8anana | Cavendsh | 198.35
2 | Banana | Cavendish | 74.07
3 | Oran9e | Cementine | 161.72
4 | Banana | Cavendish | 185.16
5 | App1e | Braebun | 170.19
6 | Orange | Clementin | 173.82
Merge into table
8anana
Banana
Oran9e
Banana
App1e
Orange
Cavendsh
Cavendish
Cementine
Cavendish
Braebun
Clementin
198.35
74.07
161.72
185.16
170.19
173.82
Apply OCR errors
Delete random characters
Apply random typos
Banana
Banana
Orange
Banana
Apple
Orange
Cavendish
Cavendish
Clementine
Cavendish
Braeburn
Clementine
199.35
74.07
161.72
186.16
170.19
173.81
Split into columns
ID | Fruit | Type | Weight
1 | Banana | Cavendish | 199.35
2 | Banana | Cavendish | 74.07
3 | Orange | Clementine | 161.72
4 | Banana | Cavendish | 186.16
5 | Apple | Braeburn | 170.19
6 | Orange | Clementine | 173.81

ID	Given name	Last name	Gender	Street name	Municipality	Postcode
254	Helmut	Jahn	m	Peenestraße	Stolpe	17391
M-254	Jahn	Helmut	m	Peenestraße	Stolpe	17391
1226	Rudolf	Franzen	m	Birkenweg	Suthfeld	31555
M-1226	Rudolf	Franzen	m	Birkenweg	Suthfeld	31565
2397	Erna	Eickhoff	f	Schulweg	Krautheim	74238
M-2397	Erna	Eickhoff	*(empty)*	Schulweg	Krautheim	74238
9960	Ingrid	Reinhold	f	Hochstraße	Mogendorf	56424
M-9960	Ingrid	Reinhold	m	Hochstraße	Mogendorf	56424

CORRECTNESS AND PERFORMANCE

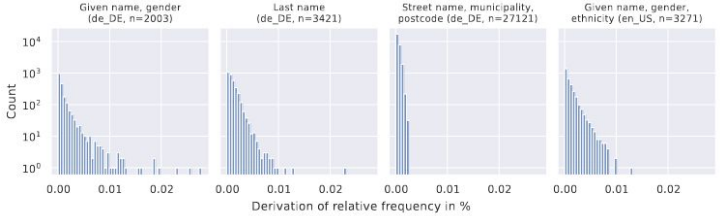
- Generate data from four distinct datasets with varying column and row counts



⇒ **Minor deviations from original distributions** **(<0.005% in 99.5% of all cases)**

CORRECTNESS AND PERFORMANCE

- Generate data from four distinct datasets with varying column and row counts



⇒ Minor deviations from original distributions **(<0.005% in 99.5% of all cases)**

- Benchmark with generation and mutation of 100k to 10m records
- Evaluation of single-core performance

		CPU time in s				
Dataset	**Records**	**Min**	**Q5**	**Q50**	**Q95**	**Max**
American	100 000	0.30	0.30	0.31	0.32	0.33
	1 000 000	2.82	2.83	2.87	3.07	3.11
	10 000 000	28.00	28.05	28.28	30.18	30.62
German	100 000	0.80	0.80	0.81	0.85	0.87
	1 000 000	6.63	6.63	6.74	6.84	6.86
	10 000 000	65.12	65.26	66.09	66.86	67.09

⇒ Gecko is 15~100x faster than its modern alternatives

CORRECTNESS AND PERFORMANCE

		CPU time in s				
Dataset	**Records**	**Min**	**Q5**	**Q50**	**Q95**	**Max**
American	100 000	0.30	0.30	0.31	0.32	0.33
	1 000 000	2.82	2.83	2.87	3.07	3.11
	10 000 000	28.00	28.05	28.28	30.18	30.62
German	100 000	0.80	0.80	0.81	0.85	0.87
	1 000 000	6.63	6.63	6.74	6.84	6.86
	10 000 000	65.12	65.26	66.09	66.86	67.09

<<<

Tool	year of release	interfaces	config-urability	conf. effort (minimal)	schema/domain independence	degree of pollution	horizontal scalability	runtime (100k)	runtime (1mio)	runtime (10mio)
DBGen	1997	GUI/CLI	low	low	✘	✘	✘	5 s	7 s	25 s
FebrlDG	2008	CLI	low	low	✘	✘	✘	2 min	4.2 min	✘
GeCo	2012	Lib/Web-UI	high	high	✓	✘	✘	80 min	13.4 h	✘
TDGen	2012	GUI/(CLI)	high	medium	✓	✘	✘	4.5 min	✘	✘
ProbGee	2012	GUI	high	medium	✓	✓	✘	5.9 h	✘	✘
DaPo	2015	CLI	high	low	✓	✓	✓	31 s[1] 17 s[1,*] 24 s[2] 15 s[2,*]	2.1 min[1] 56 s[1,*] 46 s[2] 27 s[2,*]	26 min[1] 14 min[1,*] 7.3 min[2] 3.4 min[2,*]

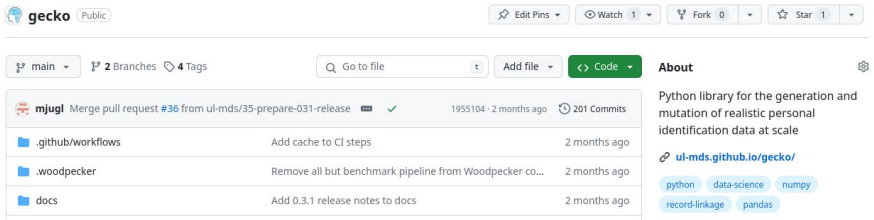
- Not bad, but…
  - Source code and test data for DaPo not available
  - DaPo performs much better when workloads are distributed
  - Other tools have more complex configuration options than Gecko

- Take these results with a large (but tasty) grain of salt :)

WHERE DO WE GO FROM HERE?

- Continuous testing of old and new record linkage algorithms
- Stress-testing of input forms that validate user-generated data
- PoC training data for machine learning models
- *This line could summarize your use case!*

⇒ **Reach out!** Maximilian.Jugl@medizin.uni-leipzig.de



UNIVERSITÄT LEIPZIG
Medizinische Fakultät

Universitätsklinikum Leipzig
Medizin ist unsere Berufung.

Thank you!

Freiberg, 07.06.2024
Maximilian Jugl

MDS
Medical Data Science