MATE: Multi-Attribute Table Extraction

Mahdi Esmailoghli, Jorge-Arnulfo Quiané-Ruiz, Ziawasch Abedjan

VLDB 2022

Sydney, Australia

Motivation

City	Date	Pollution Index
Berlin	July 21	11
Berlin	July 22	12
Sydney	July 21	36

Predict t
in Berlin

City	Date	Pollution Index
Berlin	July 21	11
Berlin	July 22	12
Sydney	July 21	36

Fit

City	Date	Pollution Index
Berlin	July 23	?

Prediction: 12

Actual
Pollution: 83

Prediction

Next Step

Motivation

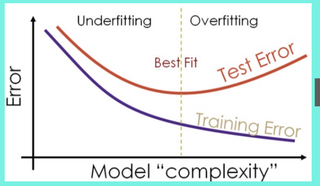
City	Date	Pollution Index
Berlin	July 21	11
Berlin	July 22	12
Sydney	July 21	36

Predict the pollution in Berlin on 23rd of July?

ML Task

City	Date	Pollution Index
Berlin	July 21	11
Berlin	July 22	12
Sydney	July 21	36

Fit



Training

City	Date	Pollution Index
Berlin	July 23	?

Prediction: 12

Actual Pollution: 83

Prediction

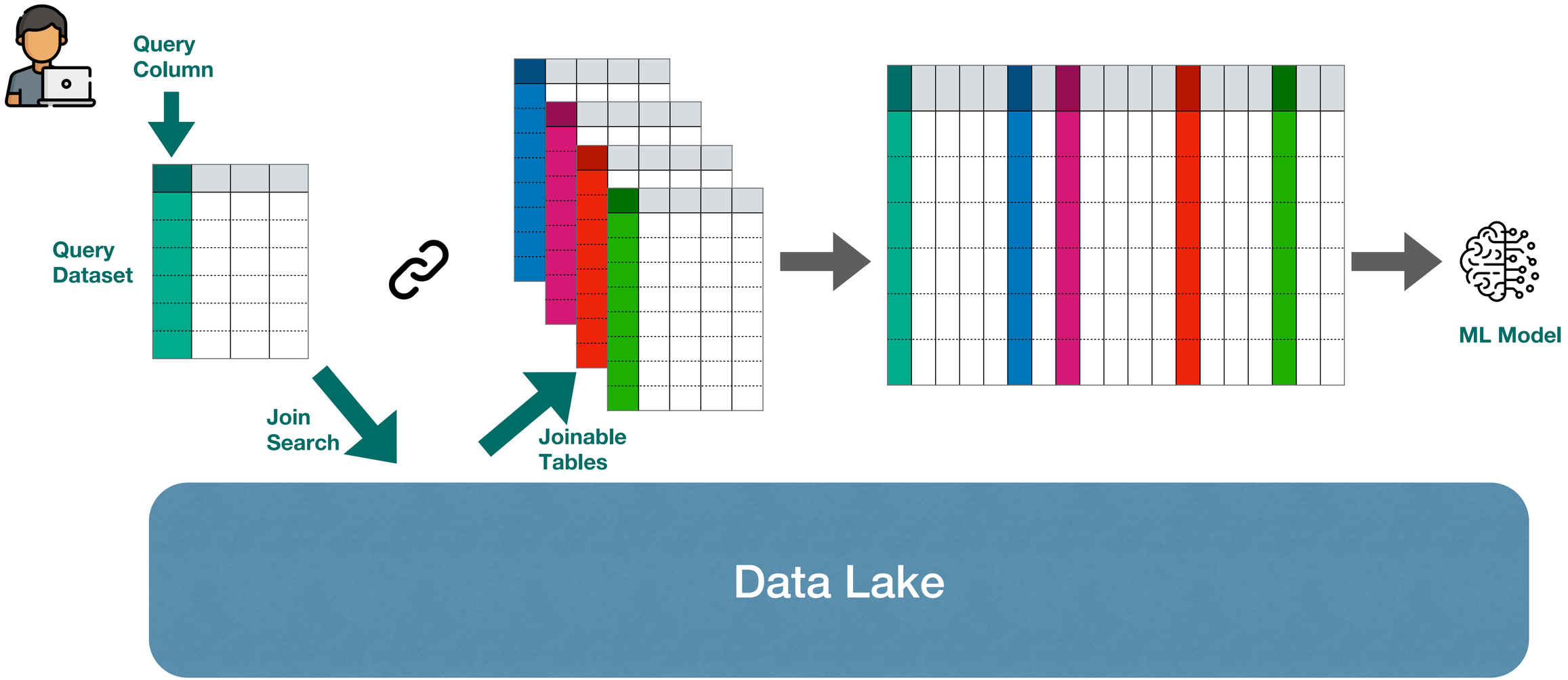
City	Date	Pollution

ENRICH

City	Date	EVENT	Pollution

Next Step

Join Discovery



Problem

Composite (Multi-Attribute) Keys

Sample Query Table

City	Date	Pollution
Berlin	July 22	12
Sydney	July 21	36
Berlin	July 21	11

Sample Inverted Index [1]

Cell Value	Table	Column	Row
Berlin	12	3	5
Berlin	19	9	2
Sydney	12	3	13
…	…	…	…

1. **S.O.T.A. Join Search Alg. Are Single-Attribute**
2. **Traditional Inverted Index Is Single-Attribute**
3. **Multi-Attribute Solution Is Needed**

1 Abedjan, Ziawasch, et al. "Dataxformer: A robust transformation discovery system." 2016 IEEE 32nd International Conference on Data Engineering (ICDE). IEEE, 2016.

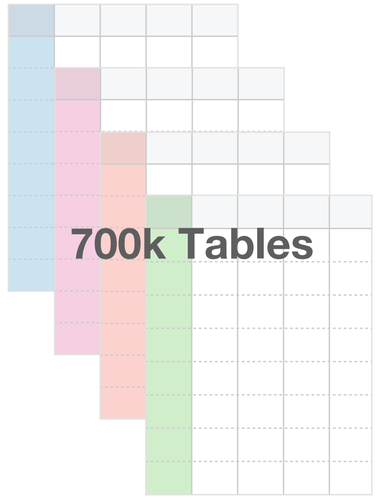
Naïve Solution #1

Composite (Multi-Attribute) Keys - Unary Join Search

City	Date	Pollution Index
Berlin	July 22	12
Sydney	July 21	36
Berlin	July 21	11

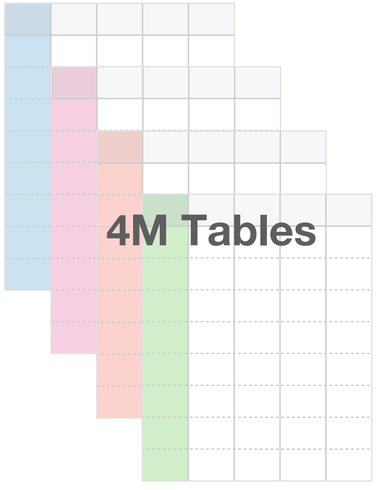
City	Date	Pollution Index
Berlin	July 22	12
Sydney	July 21	36
Berlin	July 21	11

Unary Join Search



City	Date	Pollution Index
Berlin	July 22	12
Sydney	July 21	36
Berlin	July 21	11

Unary Join Search

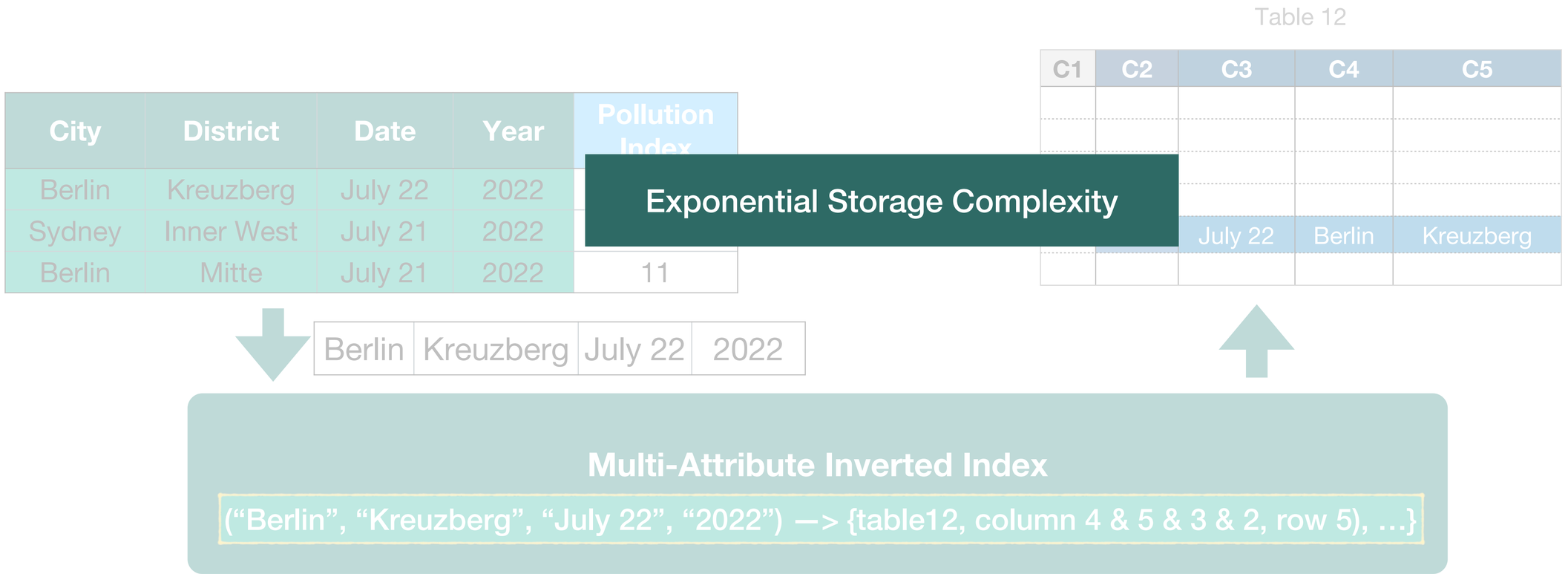


Only **1.5k** tables contain City and Date information

On average **1000x** False Positives

Naïve Solution #2

Composite (Multi-Attribute) Keys - Multi-Attribute Inverted Index



MATE

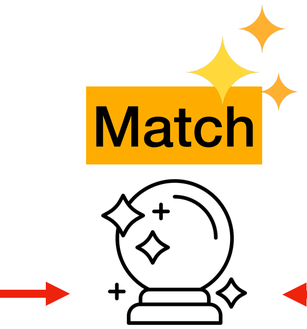
Main Idea

Query Table

City	District	Date	Year	Pollution Index
Berlin	Kreuzberg	July 22	2022	12
Sydney	Inner West	July 21	2022	36
Berlin	Mitte	July 21	2022	11

Candidate Join Table 12

	C1	C2	C3	C4	C5	C6
r1						
r2						
r3						
r4						
r5	2022	Kreuzberg		Berlin		July 22
r6						

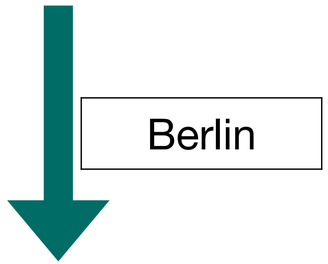


Oracle
Table 12, Row 5

MATE

Main Idea

City	District	Date	Year	Pollution Index
Berlin	Kreuzberg	July 22	2022	12
Sydney	Inner West	July 21	2022	36
Berlin	Mitte	July 21	2022	11



Single-Attribute Inverted Index

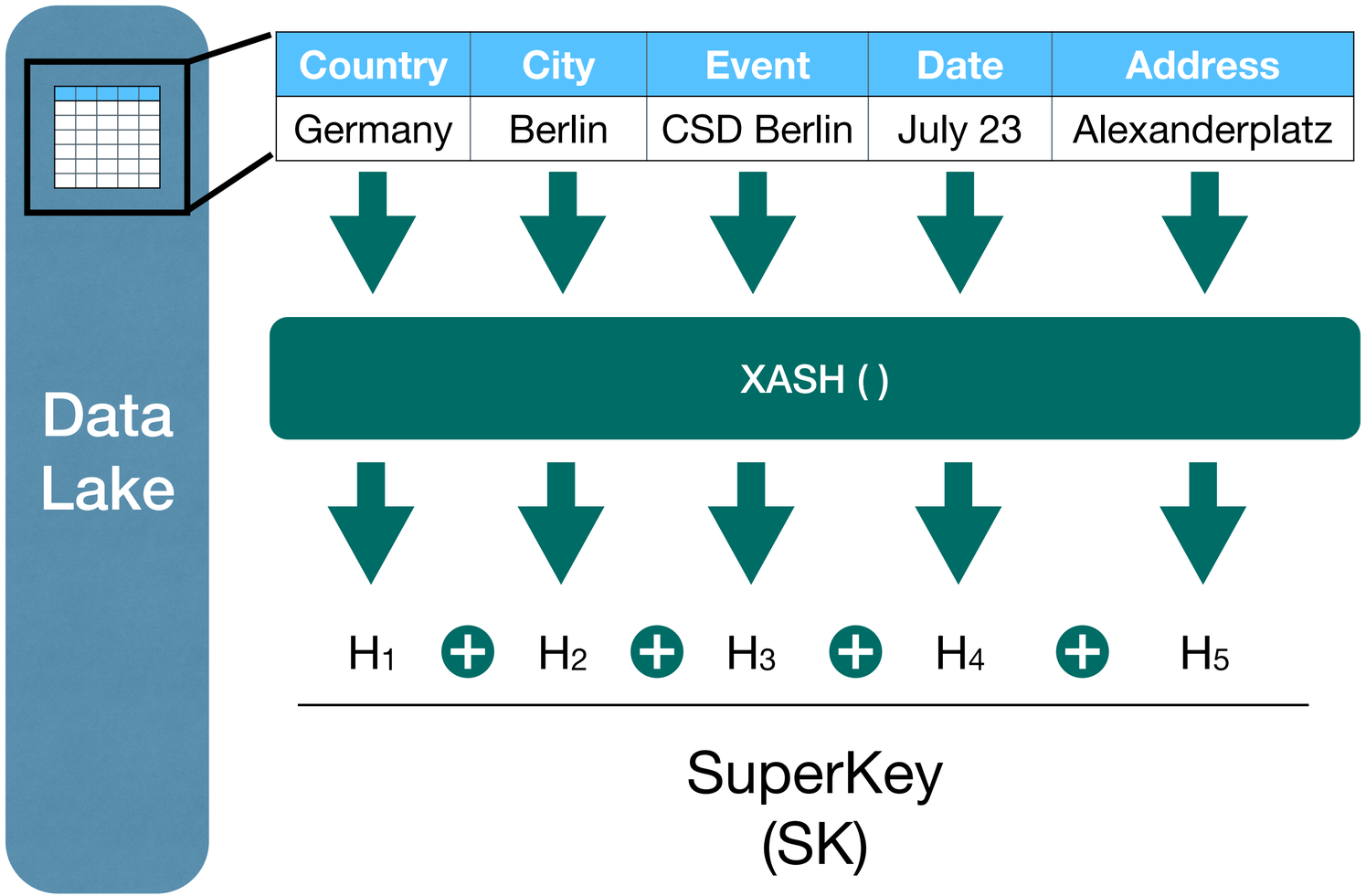
“Berlin” —> {table12, column 3, row 5, Oracle, …}

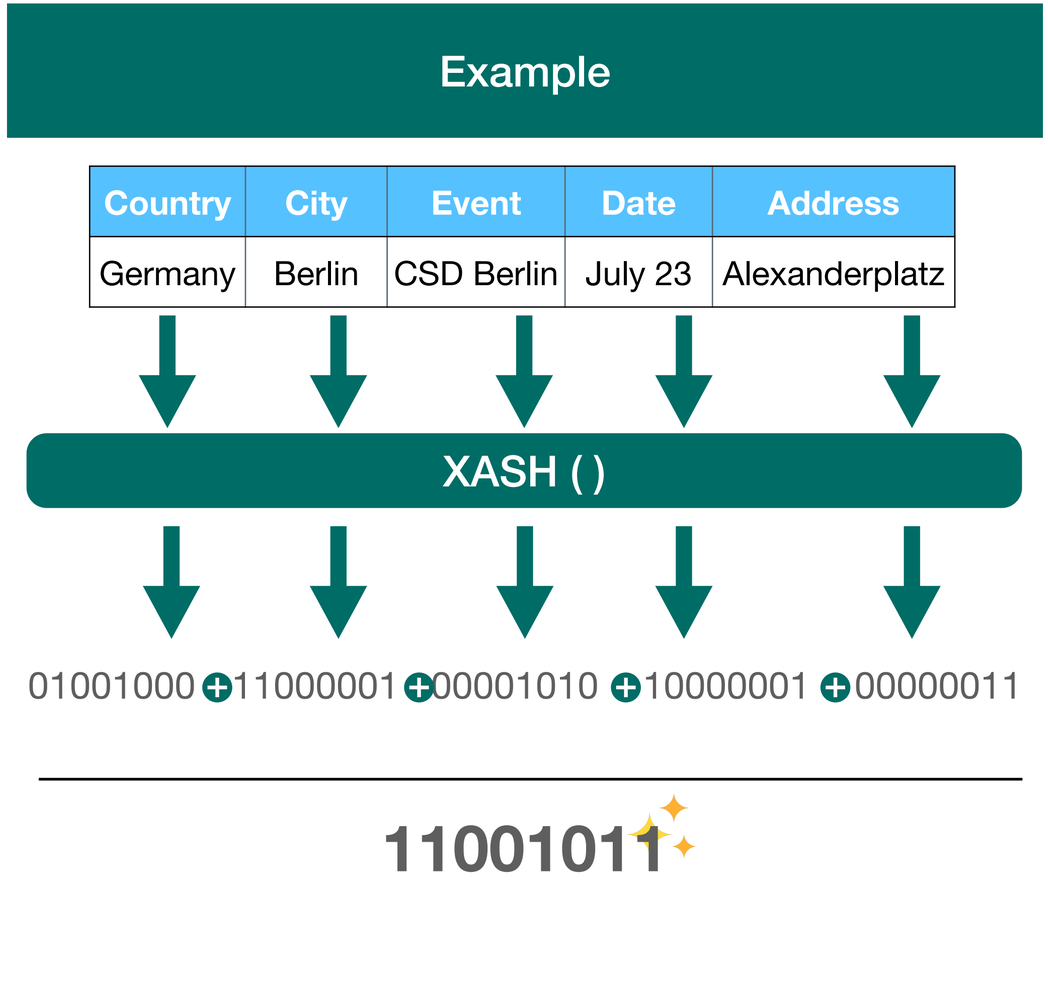
Location

1. **Low Storage Complexity**
2. **Multi-Attribute Effectiveness**
3. **Key Size Independence**

MATE

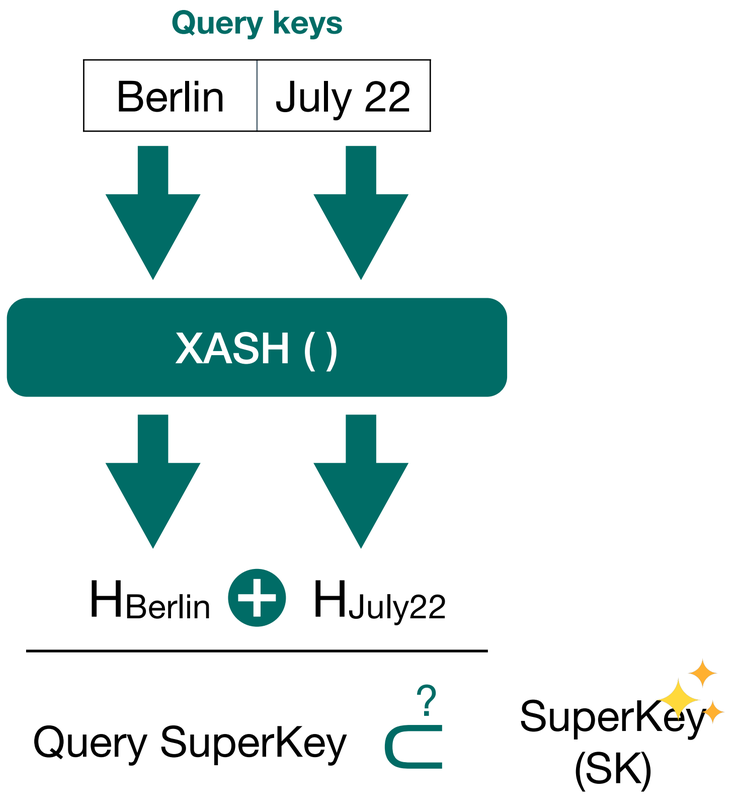
Oracle, i.e., The Super Key

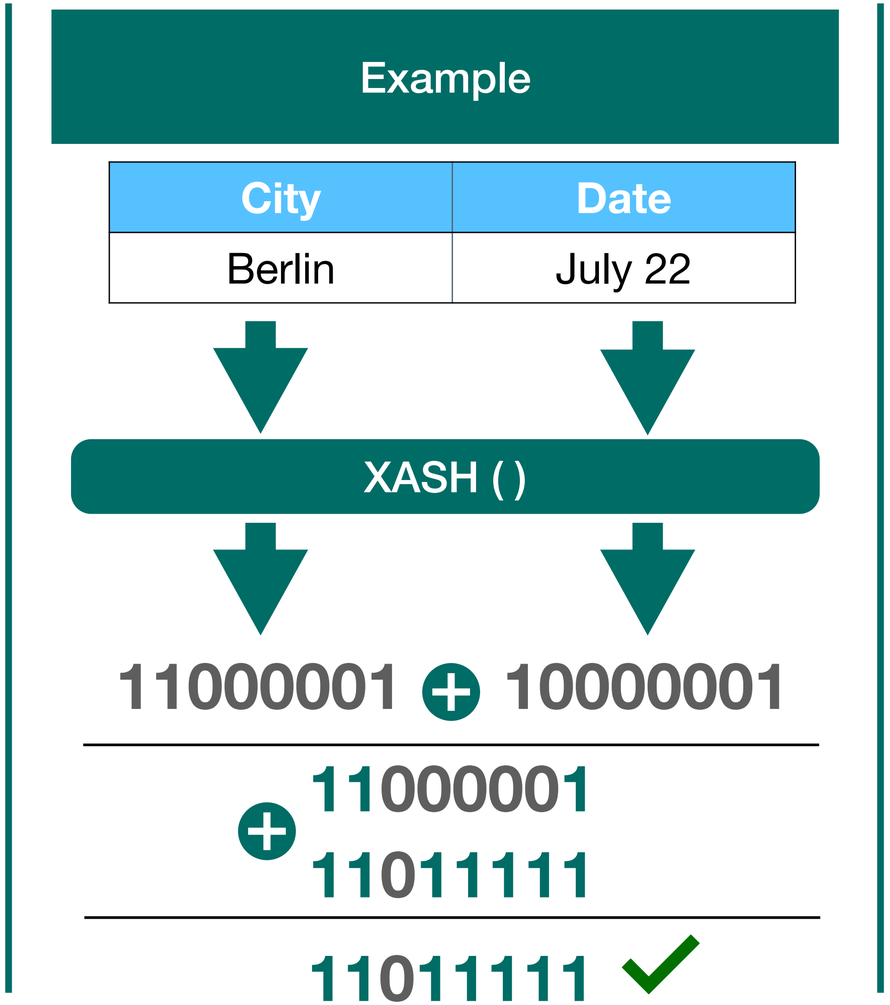




MATE

Querying The Super Key





1. One-Time Hash Generation
2. Single-Operation Comparison
3. No False Negatives

XASH (Goal)

How To Generate The Hash Values

1. **Value —> Hash (e.g., 128 bits)**

City	Date	Pollution
Berlin	July 22	12

Hash (City) ≠ Hash (Date)
Hash (City) ∩ Hash (Date) ~ 0

---

2. **Minimum number of 1 bits**

```
  0100110100011001
⊕ 1010010001011010
  0011100110101100
------------------
  1111110111111111
```

XASH (Features)

How To Generate The Hash Values

1 Rare Characters

City	Date
Sydney	May 21

D E N S

sdne → 00000100101000000010

A M Y 12

may 21 → 10100000000000101100

2 Character Positions

City	Date
Berlin	November 09

Berlin

R

Left	Middle	Right
0	1	0

November

Left	Middle	Right
0	0	1

3 Value Length

City	Date
Starnberg	September 11

Length | Character Features

Starnberg → 001000000000101100

L = 9

September 11 → 000001000000101100

L = 12

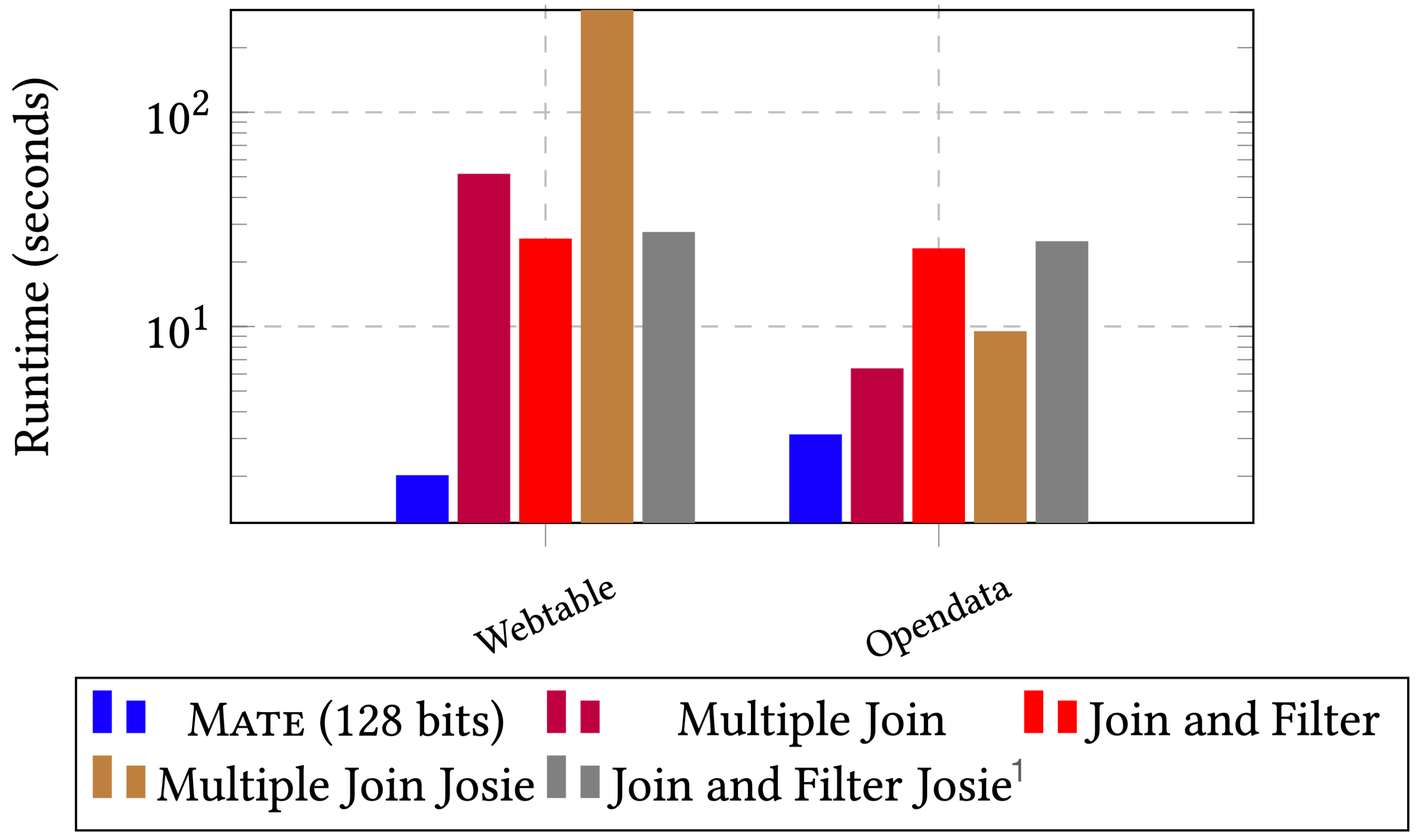
Experimental Setup

Data Lakes

Data Lake	# of Tables	# of Columns	# of Rows	Query Table Size
Dresden Web Table Corpus (DWTC)[1]	145M	870M	1.45B	450
German Open Data[2]	17k	440k	62M	450

1 https://wwwdb.inf.tu-dresden.de/misc/dwtc/
2 https://www.govdata.de/

MATE System - Runtime Comparison



1. **Outperform the S.O.T.A By 60x**

2. **Scalable on Large Corpora**

1 Erkang Zhu, Dong Deng, Fatemeh Nargesian, and Renée J Miller. 2019. JOSIE: Overlap Set Similarity Search for Finding Joinable Tables in Data Lakes. In SIGMOD. 847–864.

XASH Precision Comparison

Dataset	*SimHash*		*HT*		*BF*		*LHBF*[1]		*Xash*	
	128	**512**	**128**	**512**	**128**	**512**	**128**	**512**	**128**	**512**
Webtable	0.27±0.40	0.27±0.39	0.34±0.41	0.38±0.41	0.46±0.44	0.34±0.41	0.45±0.43	0.63±0.44	**0.61±0.43**	**0.93±0.22**
Opendata	0.28±0.38	0.32±0.41	0.43±0.40	0.55±0.41	**0.56±0.41**	0.79±0.34	0.45±0.43	0.67±0.35	0.52±0.41	**0.80±0.34**

1 Adam Kirsch and Michael Mitzenmacher. 2006. Less Hashing, Same Performance: Building a Better Bloom Filter. In Algorithms - ESA 2006, 14th Annual European Symposium, Zurich, Switzerland, September 11-13, 2006, Proceedings (Lecture Notes in Computer Science), Yossi Azar and Thomas Erlebach (Eds.), Vol. 4168. Springer, 456–467.

Conclusion

- **MATE**, a **multi-attribute** join discovery system
- **XASH efficiently** encodes **syntactic features**
- **Enhanced inverted index** for **multi-column processing**
- **Compatible** with traditional filtering, e.g., **prefix filter**

Check out the
Git repository!

esmailoghli@dbs.uni-hannover.de