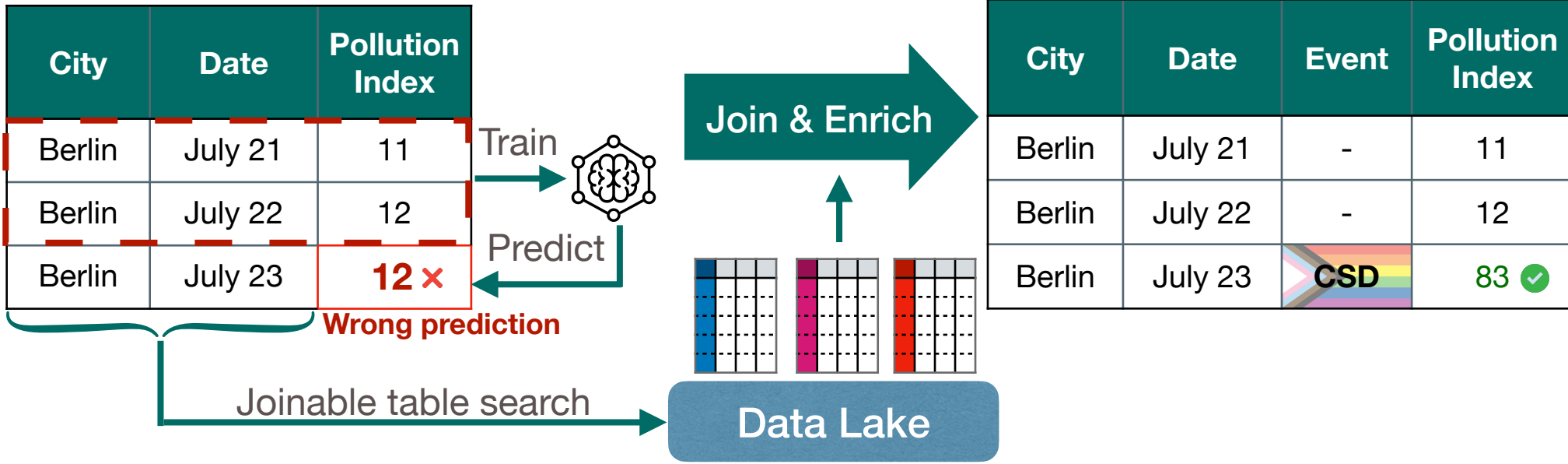# MATE: Multi-Attribute Table Extraction

Motivation
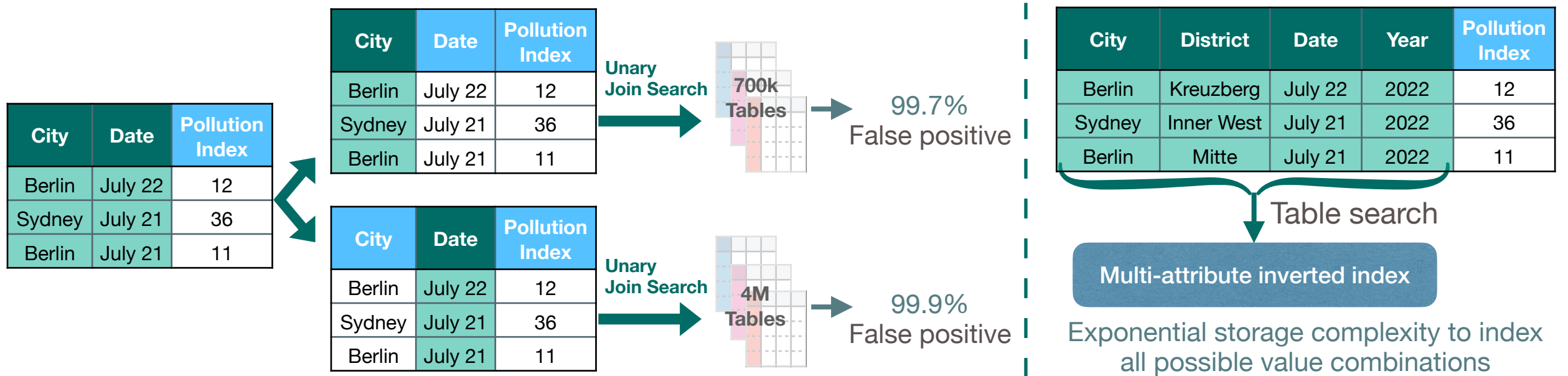
Finding informative features is a crucial task in data science workflows

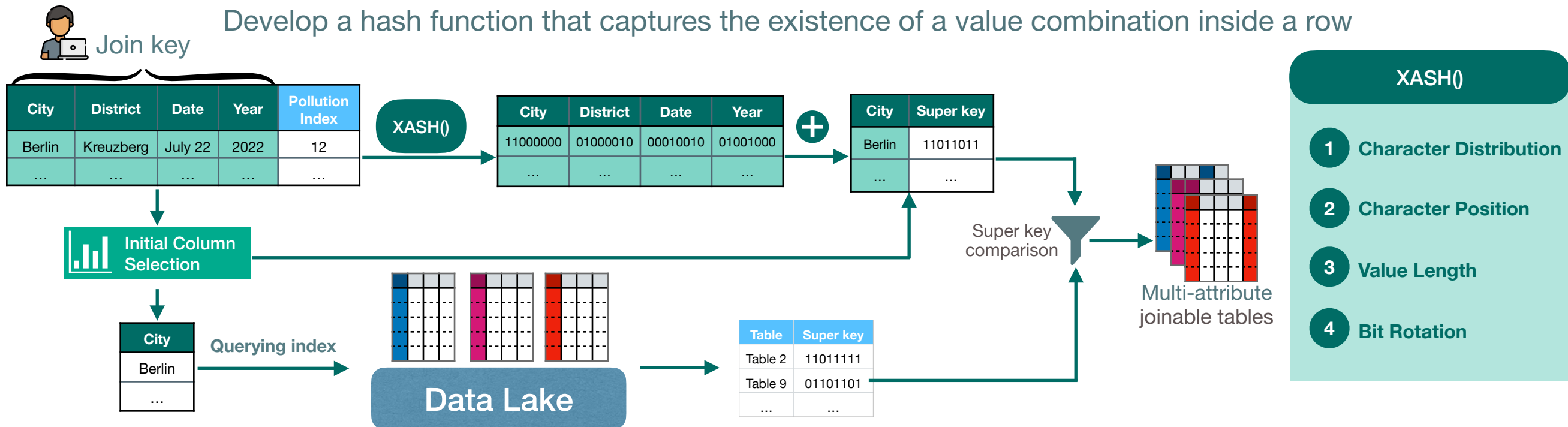Automated data discovery in large data lakes can be the solution

City	Date	Pollution Index
Berlin	July 21	11
Berlin	July 22	12
Berlin	July 23	12 ✗

Train → Predict → **Wrong prediction**

Joinable table search → **Data Lake**

Join & Enrich →

City	Date	Event	Pollution Index
Berlin	July 21	-	11
Berlin	July 22	-	12
Berlin	July 23	CSD	83 ✓

Problem

Composite keys: Current single- & multi-attribute discovery solutions: high false positive rate or high storage complexity

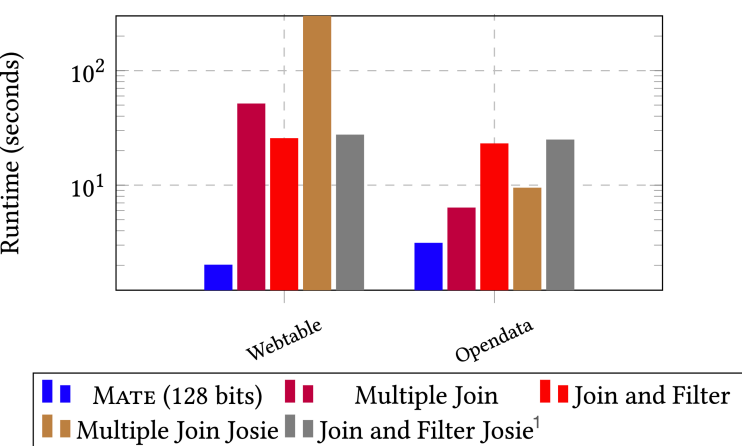| City	Date	Pollution Index
Berlin	July 22	12
Sydney	July 21	36
Berlin	July 21	11

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

Unary Join Search → 700k Tables → 99.7% False positive

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

Unary Join Search → 4M Tables → 99.9% False positive

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

Table search → **Multi-attribute inverted index**

Exponential storage complexity to index all possible value combinations

Solution

Develop a hash function that captures the existence of a value combination inside a row

Join key

City	District	Date	Year	Pollution Index
Berlin	Kreuzberg	July 22	2022	12
...	...	...	...	...

XASH() →

City	District	Date	Year
11000000	01000010	00010010	01001000
...	...	...	...

+ →

City	Super key
Berlin	11011011
...	...

Initial Column Selection

City
Berlin
...

Querying index → **Data Lake** →

Table	Super key
Table 2	11011111
Table 9	01101101
...	...

Super key comparison → Multi-attribute joinable tables

XASH()

1. **Character Distribution**
2. **Character Position**
3. **Value Length**
4. **Bit Rotation**

Experiments

Runtime comparison between MATE and S.O.T.A

- MATE (128 bits)
- Multiple Join
- Join and Filter
- Multiple Join Josie
- Join and Filter Josie[1]

Runtime comparison between XASH and other hash functions

Dataset	SimHash		HT		BF		LHBF[1]		XASH	
	128	512	128	512	128	512	128	512	128	512
Webtable	13.4	17.2	9.5	12.4	4.7	16.8	5.6	4.2	**1.6**	**2.0**
Opendata	102.9	56.5	24.0	10.8	11.0	3.4	30.1	8.3	**6.5**	**3.1**

Impact of different components in XASH

- SCR
- Rare characters
- Char. + len. + loc.
- XASH (512 bit)
- Length
- Char. + loc.
- XASH (128 bit)
- Ideal system

Mahdi Esmailoghli
Leibniz Universität Hannover & L3S Research Center
esmailoghli@dbs.uni-hannover.de

Jorge-Arnulfo Quiané-Ruiz
TU Berlin
jorge.quiane@tu-berlin.de

Ziawasch Abedjan
Leibniz Universität Hannover & L3S Research Center
abedjan@dbs.uni-hannover.de