
# Relational Header Discovery using Similarity Search in a Table Corpus

Hazar Harmouch, Thorsten Papenbrock, Felix Naumann
Information System Group
Hasso Plattner Institut
University of Potsdam
Germany

ICDE 2021
Virtual event
Chania, Crete, Greece
19-22 April 2021

# Motivation
## Missing headers

### VLDB 2021: Conference Officers

#### Associate Editors

**Alon Halevy** Facebook, United States
**Anastasia Ailamaki** École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
**Angela Bonifati** Lyon 1 University, France
**Arun Kumar** UC San Diego, United States
**Ashraf Aboulnaga** Qatar Computing Research Institute, HBKU, Qatar
**Eugene Wu** Columbia University, United States
**Floris Geerts** University of Antwerp, Belgium
**Graham Cormode** University of Warwick, United Kingdom
**Jeffrey Xu Yu** The Chinese University of Hong Kong, China
**Jiannan Wang** Simon Fraser University, Canada
**Jingren Zhou** Alibaba, China
**Jorge Arnulfo Quiane Ruiz** TU Berlin, Germany
**Juliana Freire** New York University, United States
**Jun Yang** Duke University, United States

| ? | ? | ? |
|---|---|---|
| Alon Halevy | Facebook | United States |
| Anastasia Ailamaki | École Polytechnique Fédérale de Lausanne (EPFL) | Switzerland |
| Angela Bonifati | Lyon 1 University | France |
| … | … | … |

# Motivation
## Missing headers ≈ un-interpretable headers

noe_pop_age_sex_2012_201...

Sort fields: Data source order

| Nuts1 | Nuts2 | Nuts3 | Lau2 Code | Lau2 Name | Age Group | Pop Total | Pop Male |
|---|---|---|---|---|---|---|---|
| AT1 | AT12 | AT124 | 30101 | Krems an der Donau | 0_4 | 1,063 | 528 |
| AT1 | AT12 | AT1 | | | | | |
| AT1 | AT12 | AT1 | | | | | |
| AT1 | AT12 | AT1 | | | | | |
| AT1 | AT12 | AT1 | | | | | |
| AT1 | AT12 | AT1 | | | | | |
| AT1 | AT12 | AT124 | 30101 | Krems an der Donau | 30_34 | 1,586 | 855 |

Attributbeschreibung

NUTS1: Land NUTS2: Bundesland NUTS3: Gruppen von Bezirken
LAU2_CODE: Gemeindekennzahl LAU2_NAME: Gemeindename
AGE_GROUP: Altersgruppe (5-jährig) POP_TOTAL: Bevölkerung
POP_MALE: männliche Bevölkerung POP_FEMALE: weibliche
Bevölkerung YEAR: Referenzjahr

https://www.data.gv.at/katalog/dataset/fde11447-12cf-4333-bed9-c37056551e0c

# Motivation
## How big is the problem?

Header row is missing in

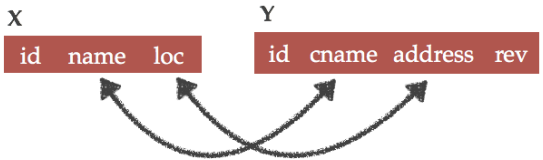
- majority of tables on the web [Balakrishnan15].
- ~29% of relational web tables extracted by [Cafarella08].

Web tables

- ~28% of the CSV tables from the Austrian and European Open Data portals [Neumaier16].
- ~11k documents in a data corpus from 232 Open Data portals [Mitlöhner16].

# Motivation
## Where are column headers useful?


**Querying data**




**Schema matching**

**Data integration**

**Knowledge base augmentation**

**Table search engines**

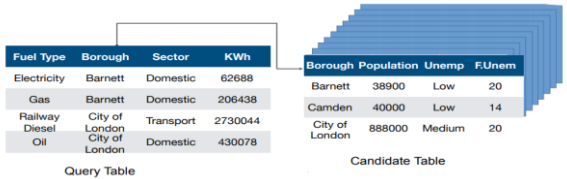
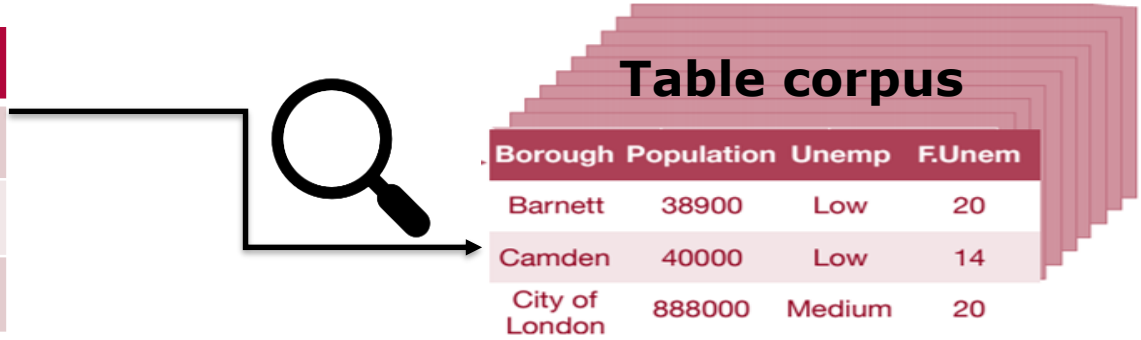



**Table join**

# The problem
## Schema discovery problem

| family | motto | location | symbol |
|---|---|---|---|
| stark | winter is coming | winterfell | grey wolf |
| targaryen | fire and blood | dragonstone | red dragon |
| lannister | hear me roar! | casterly rock | golden lion |

| ? |
|---|
| stark |
| targaryen |
| lannister |




**Relational Header Discovery**

Hazar Harmouch
ICDE 2021

# Schema discovery problem – Schema coherence

athlete

…

color

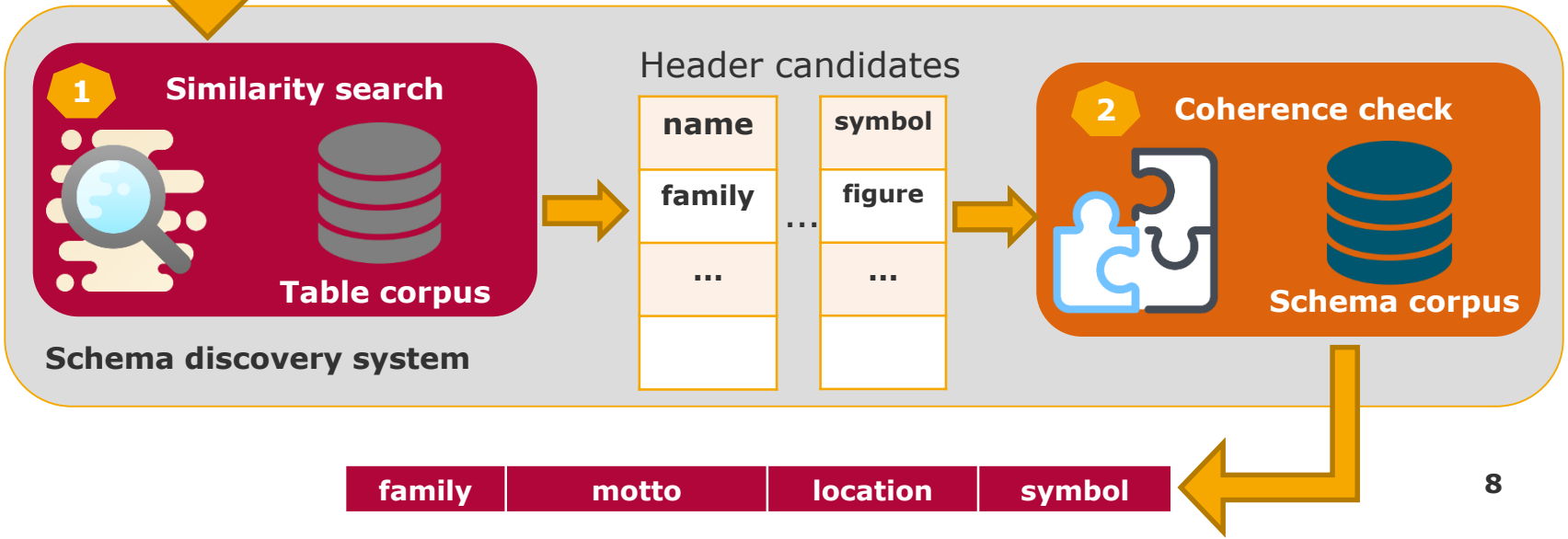
medal

…

| athlete | medal |
|---|---|
| Eddie Eagan | gold |
| Jacob Tullin Thams | silver |
| Christa Luding-Rothenburger | silver |
| Clara Hughes | bronze |

# Solution
## End-to-end schema discovery

| ? | ? | ? | ? |
|---|---|---|---|
| stark | winter is coming | winterfell | grey wolf |
| targaryen | fire and blood | dragonstone | red dragon |
| lannister | hear me roar! | casterly rock | golden lion |



# Schema discovery using similarity search
## Phase 1: Similarity search - Challenges

- **Efficient** similarity search
  - **Data sketches** reduce memory consumption.
- **Un-biased** column similarity measure.
- Filtering out **dissimilar** tables.

| ? |
|---|
| Berlin |
| **Potsdam** |

| City |
|---|
| **Potsdam** |
| Frankfurt |
| Cologne |

| Location |
|---|
| Frankfurt |
| **Potsdam** |
| Cologne |
| Dresden |
| Munich |

| ? | ? | ? | ? |
|---|---|---|---|
| Stark | Winter is coming | Winterfell | Grey wolf |
| Targaryen | Fire And Blood | Dragonstone | Red dragon |
| Lannister | Hear Me Roar! | Casterly rock | Golden lion |

≠

| University | Year of establishment | Students | Location |
|---|---|---|---|
| Heidelberg University | 1386 | 31,500 | Heidelberg |
| Leipzig University | 1409 | 29,500 | Leipzig |
| University of Rostock | 1419 | 14,000 | Rostock |
| University of Greifswald | 1456 | 12,000 | Greifswald |
| University of Freiburg | 1457 | 24,700 | Freiburg |
| University of Munich | 1472 | 50,918 | Munich[5] |
| University of Tübingen | 1477 | 28,700 | Tübingen |
| University of Halle-Wittenberg | 1502 | 18,500 | Halle[6] |
| University of Marburg | 1527 | 24,000 | Marburg |
| University of Jena | 1558 | 19,000 | Jena |

# Phase 1: Similarity search – Which data sketch?

Jaccard coefficient: $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$

Jaccard containment: $C(A,B) = \frac{|A \cap B|}{|A|}$

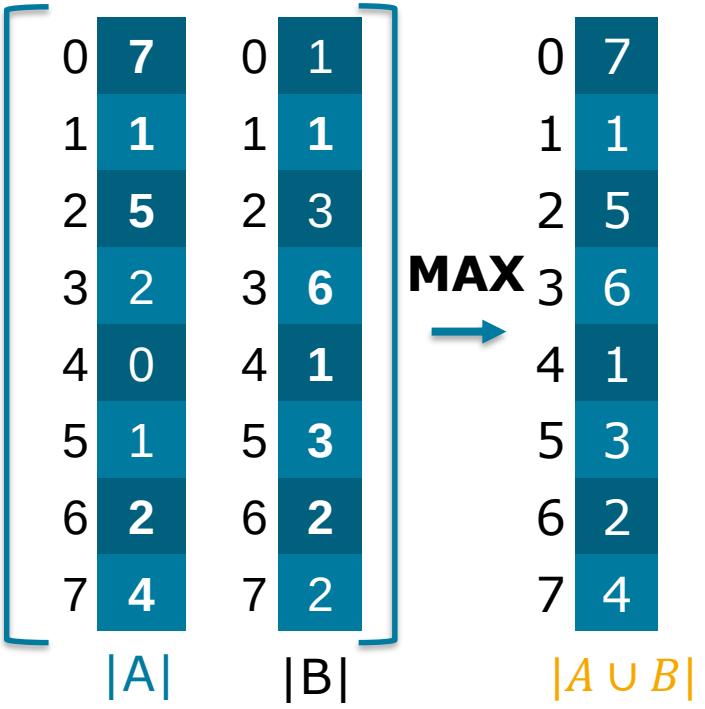
- HyperLogLog (HLL) is a data sketch that
  - has compact form (~1kB).
  - estimates cardinalities with an error < 0.01.

# Phase 1: Similarity search - HLL-based similarities

Jaccard coefficient: $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$

Jaccard containment: $C(A,B) = \frac{|A \cap B|}{|A|}$

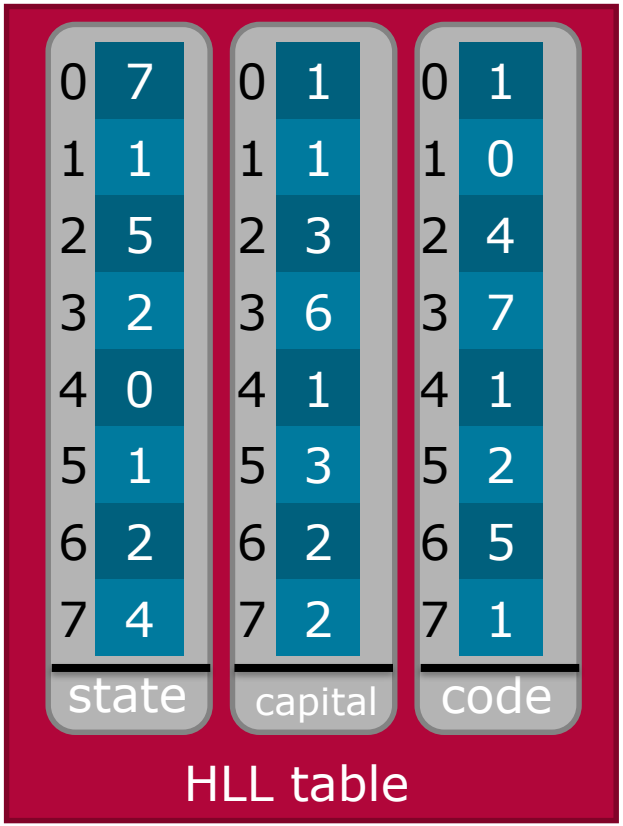
- Inclusion-exclusion principle:
  $|A \cup B| = |A| + |B| - |A \cap B|$



# Schema discovery using similarity search
## Phase 1: Similarity search – HLL table data structure

| State | Capital | Code |
|---|---|---|
| Austria | Vienna | AT |
| Belgium | Brussels | BE |
| Bulgaria | Sofia | BG |
| Croatia | Zagreb | HR |

Transform



# Schema discovery using similarity search
## Phase 1: Similarity search

**Offline**

Wiki web tables

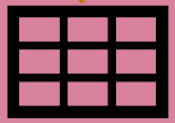
32G (1.6M Tables)

Transform

HLL tables

450MB

| ? | ? | ? | ? |
|---|---|---|---|
| Stark | Winter is coming | Winterfell | Grey wolf |
| Targaryen | Fire And Blood | Dragonstone | Red dragon |
| Lannister | Hear Me Roar! | Casterly rock | Golden lion |



HLL table

HLL-based Jaccard table similarity

HLL-based Jaccard containment column similarity

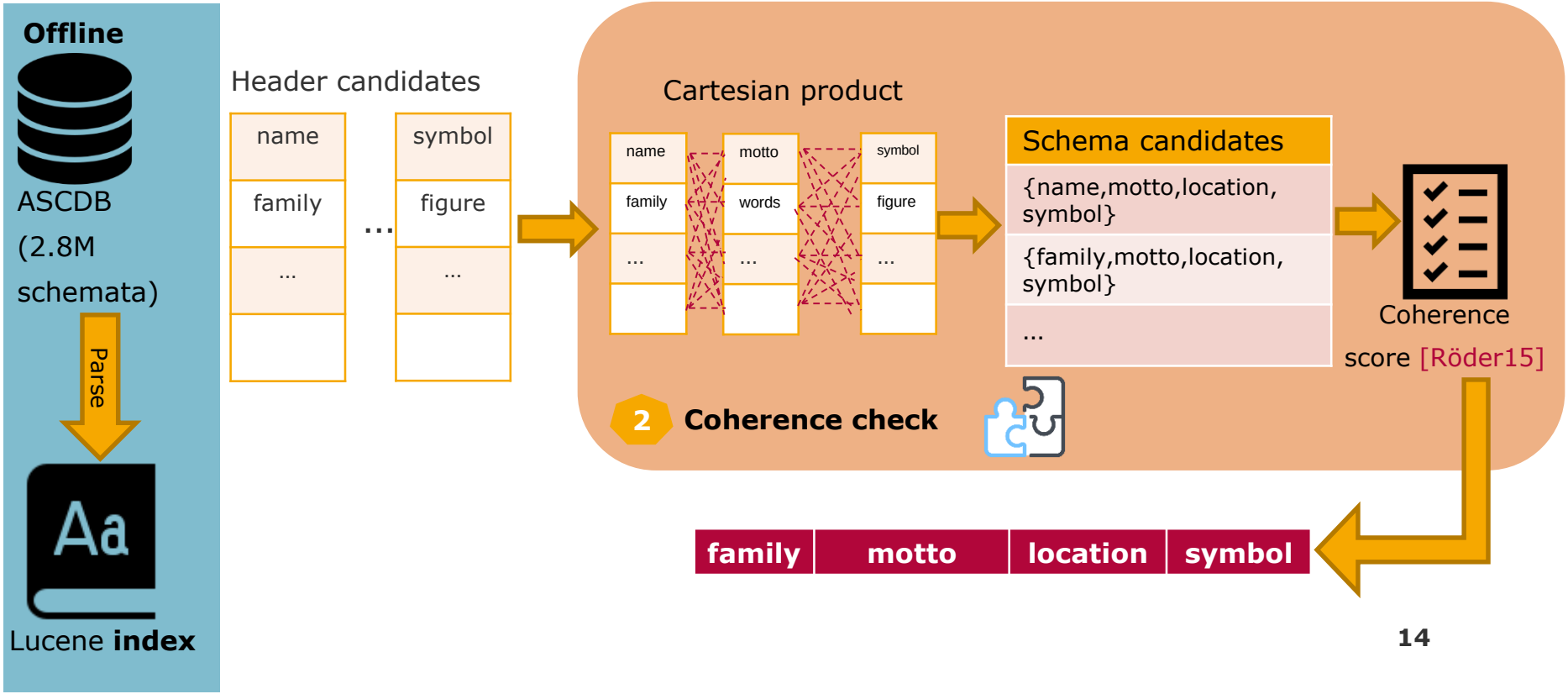
1 **Similarity search**

Header candidates

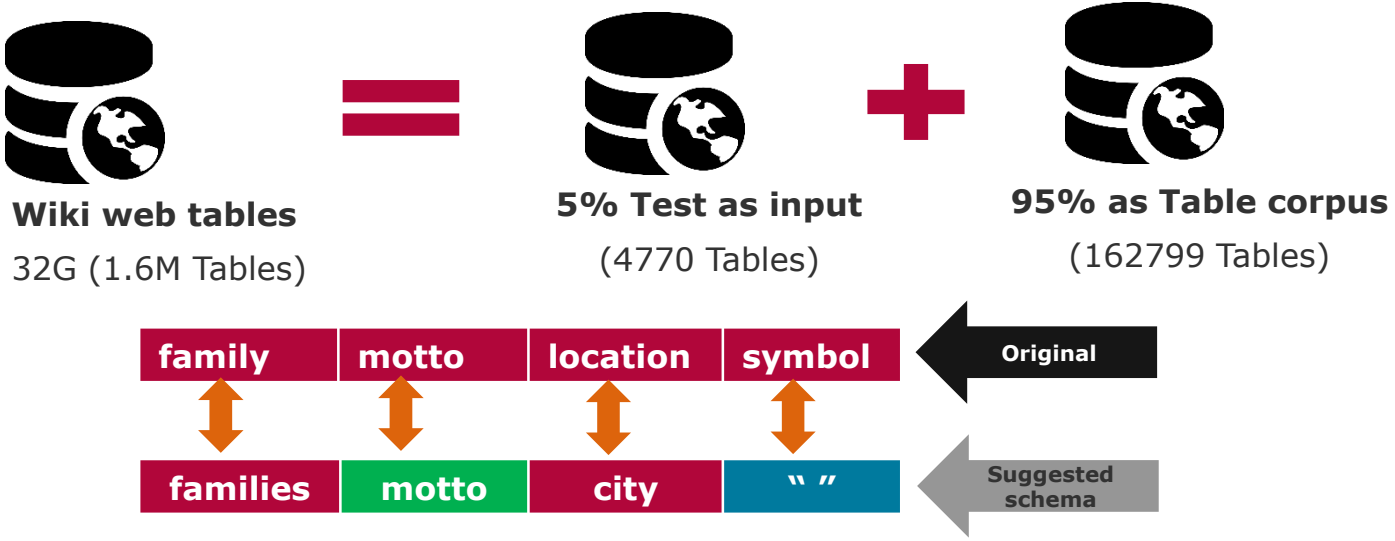
| name | symbol |
|---|---|
| family | figure |
| ... | ... |
| | |

...

# Schema discovery using similarity search
## Phase 2: Coherence check



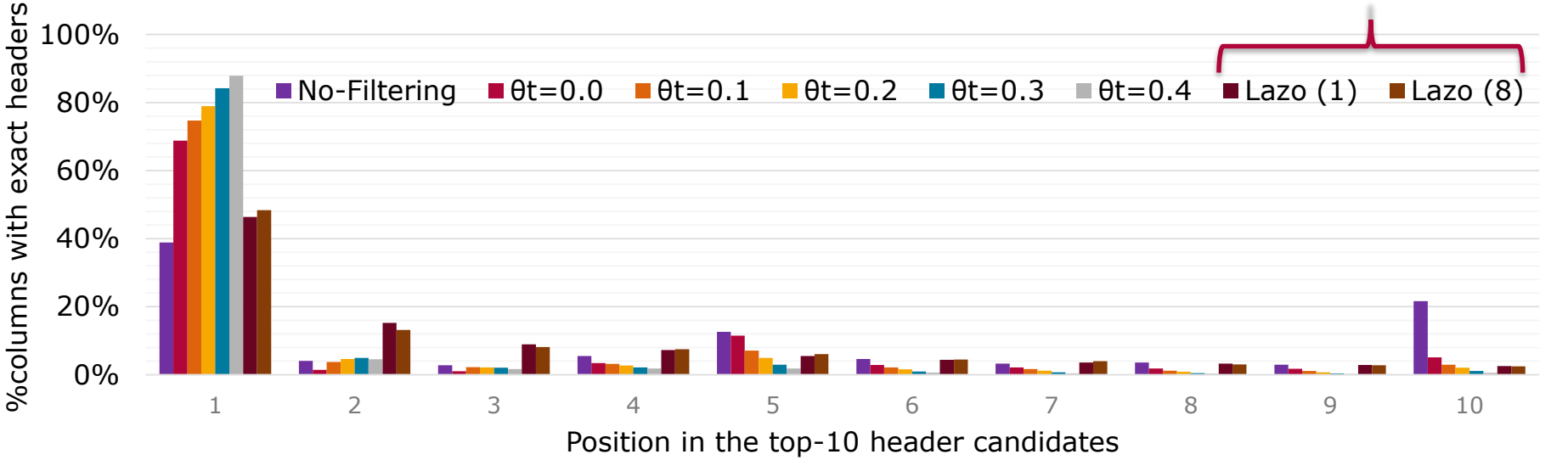
# Schema discovery using similarity search
## Experimental setup




- TP: exactly re-discovered headers.
- FP: differing headers from the original schema.
- FN: no-header suggestions.

# Schema discovery using similarity search
## Results: Similarity search (1/2)

**Lazo** [Fernandez19]



# Schema discovery using similarity search
## Results: Similarity search (2/2)

| θ | % exact header | % differing header | % no header |
|---|---|---|---|
| Lazo (1) | 30.95 | 38.52 | 30.52 |
| Lazo (8) | 31.63 | 39.60 | 28.77 |
| Lazo (10) | **Out of memory** | | |
| No filtering | 62.73 | 36.16 | 1.10 |
| 0 | 74.00 | 24.10 | 1.10 |
| 0.1 | 61.66 | 13.49 | 24.85 |
| 0.2 | 42.90 | 7.92 | 49.18 |
| 0.3 | 27.25 | 4.02 | 68.73 |
| 0.4 | 16.77 | 1.98 | 81.25 |

**Lazo** [Fernandez19]

# Schema discovery using similarity search
## Results: Error analysis - Schema-level

| | |
|---|---|
| Plural | {year, champions}<br>{year, champion} |
| Synonym | {men's, women's}<br>{male, female} |
| Time | {team, city, state, stadium, capacity, 2008 season}<br>{team, city, state, stadium, capacity, 2009 season} |
| Short form | {no., date, score, opponent, record}<br>{number, date, score, opponent, record} |
| Symbol | {place, rider, number, country, machine, points, wins}<br>{place, rider, #, country, machine, points, wins} |
| Granularity | {year, population}<br>{year, village population} |

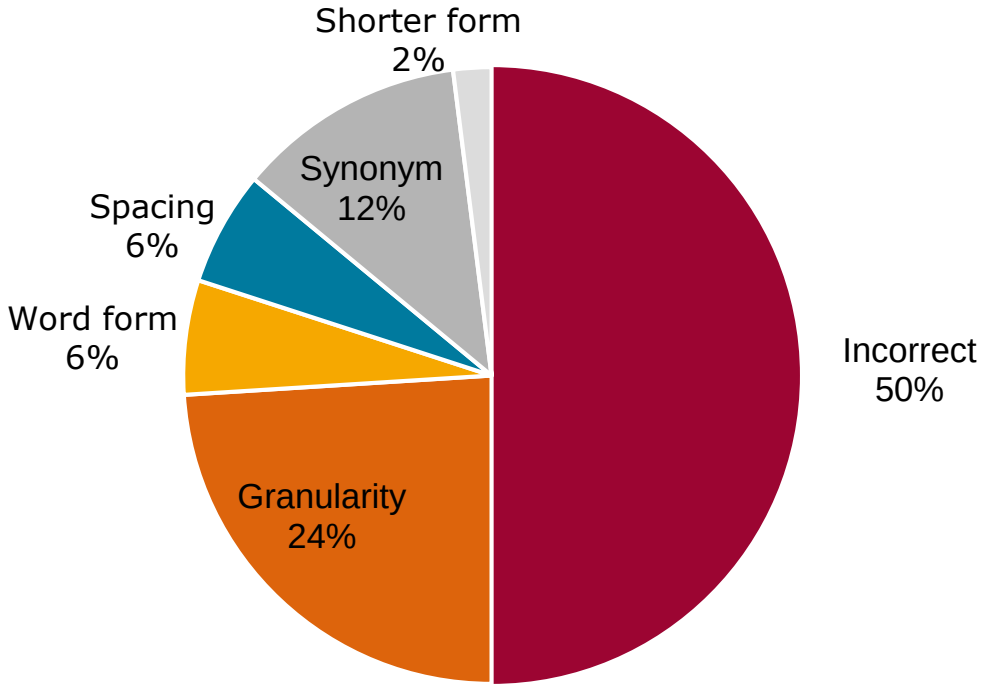
Original

Top-1 schema

**Manual test**

- 100 schemata
- 400 headers

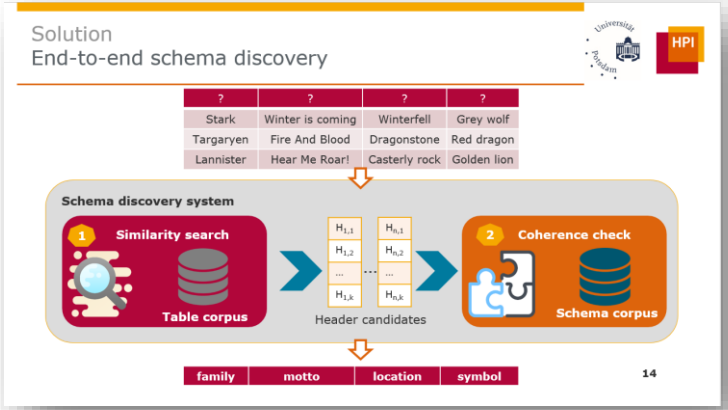
# Schema discovery using similarity search
## Results: Error analysis - Individual headers

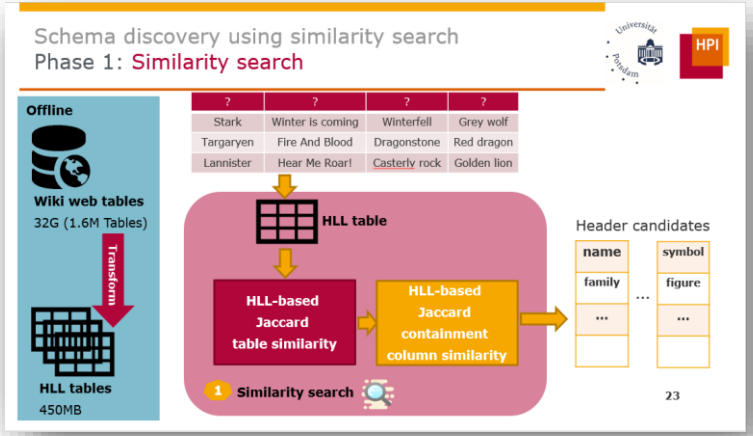


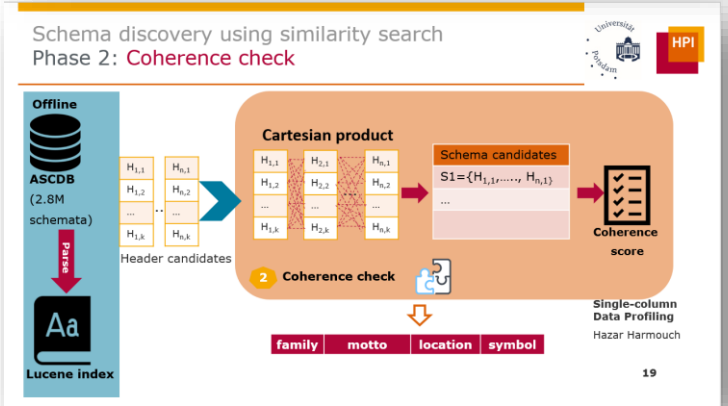

**Manual test**

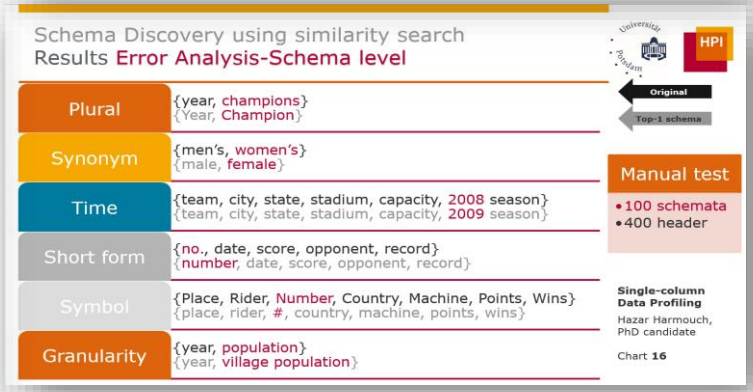
- 100 schemata
- 400 headers

# Summary




**Relational Header Discovery**

Hazar Harmouch
ICDE 2021

# References


- [Neumaier16] Neumaier, S., Umbrich, J., Parreira, J. X., & Polleres, A. (2016, October). Multi-level semantic labelling of numerical values. In International Semantic Web Conference (pp. 428-445). Springer, Cham.
- [Mitlöhner16] Mitlöhner, J., Neumaier, S., Umbrich, J., & Polleres, A. (2016, August). Characteristics of open data csv files. In Open and Big Data (OBD), International Conference on (pp. 72-79). IEEE.
- [Balakrishnan15] S. Balakrishnan, A. Y. Halevy, B. Harb, H. Lee, J. Madhavan, A. Rostamizadeh, W. Shen, K. Wilder, F. Wu, and C. Yu. Applying webtables in practice. In CIDR, 2015.
- [Cafarella08] M. J. Cafarella, A. Y. Halevy, Y. Zhang, D. Z. Wang, and E. Wu. Uncovering the relational web. In WebDB, 2008.
- [Fernandez19] Raul Castro Fernandez, Jisoo Min, Demitri Nava, and Samuel Madden. Lazo: Acardinality-based method for coupled estimation of jaccard similarity and contain-ment. InProceedings of the IEEE International Conference on Data Engineering(ICDE), 2019.