# WannaDB: Ad-hoc SQL Queries over Text Collections

Jakob Ernesti

DT-DB42-M: Datenbanksysteme –
The Question to or the Better Answer on 42?

Sommer Term 2023

# WannaDB: SQL-Queries over text Collections

**Just tell it what you want, what you really, really want**

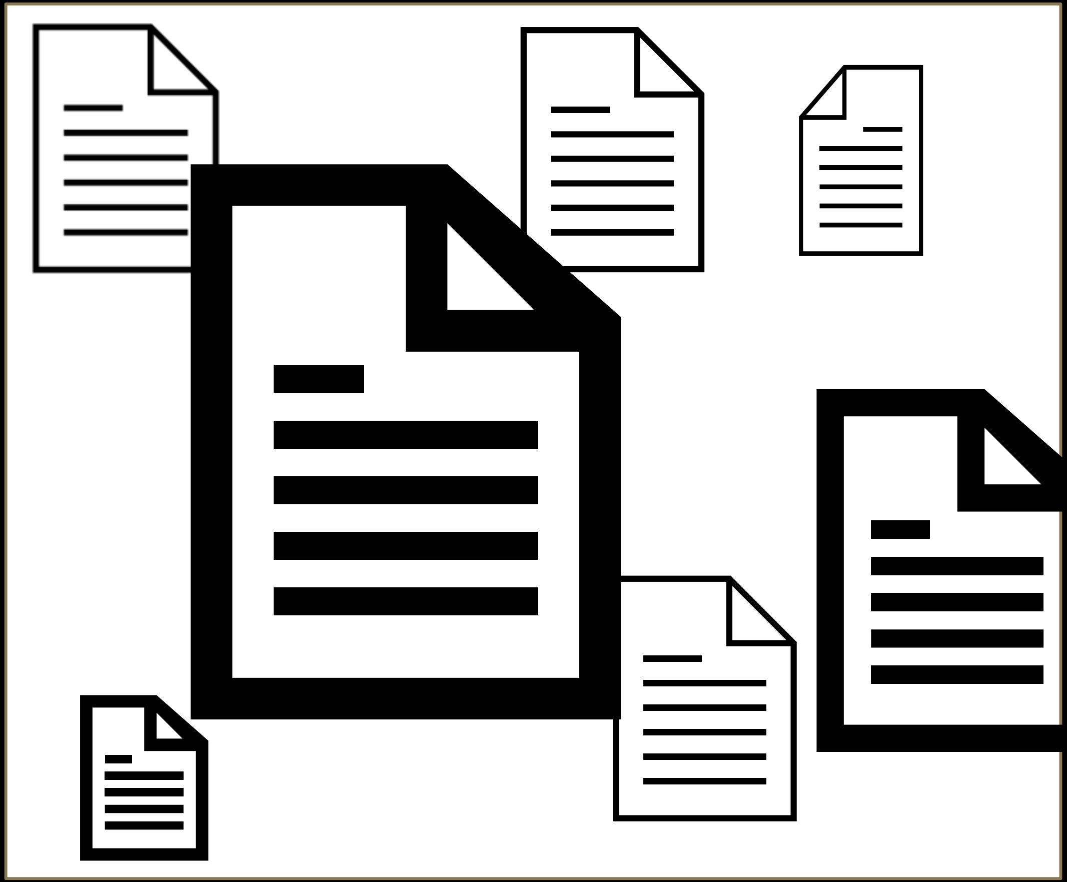
BTW ‘23:
Best long paper award

# How to extract information?

SELECT answer

FROM Universe

WHERE question = "Answer to the Ultimate Question of Life, the Universe, and Everything";



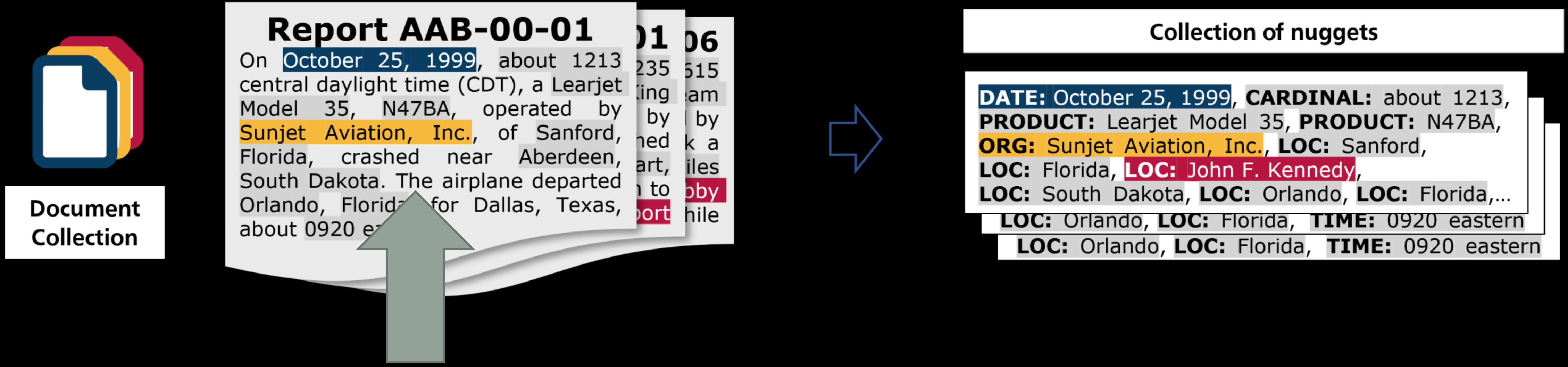
# How to extract information?

```
SELECT author, COUNT(*) as c
GROUP BY author
HAVING c > 1
```

| c | author |
|---|---|
| 20 | John Doe |
| 11 | Anon |
| 42 | Anonymus |
| 2 | Mike |
| … | … |

# Stage 1: Offline Extraction

**1) Offline Extraction:** Extract all nuggets that might be relevant (once per document, independent of information need)



On October 25, 1999, about 1213 central daylight time (CDT), a Learjet Model 35, N47BA, operated by Sunjet Aviation, Inc., of Sanford, Florida, crashed near Aberdeen, South Dakota. The airplane departed Orlando, Florida for Dallas, Texas, about 0920 e

**DATE:** October 25, 1999, **CARDINAL:** about 1213, **PRODUCT:** Learjet Model 35, **PRODUCT:** N47BA, **ORG:** Sunjet Aviation, Inc., **LOC:** Sanford, **LOC:** Florida, **LOC:** John F. Kennedy, **LOC:** South Dakota, **LOC:** Orlando, **LOC:** Florida,…

**LOC:** Orlando, **LOC:** Florida, **TIME:** 0920 eastern

**LOC:** Orlando, **LOC:** Florida, **TIME:** 0920 eastern

Named Entity Recognition

# Stage 1: Offline Extraction

Nugget

| | |
|---|---|
| Label | Company |
| Mention | „Lufthansa“ |
| Context | „Die Lufthansa beschäftigt ….“ |
| Position | Doc 4, page 5, …. |

# Stage 2: Interactive SQL Processing

## Information need as SQL-like query

On which dates were the incidents over 500?

```
SELECT report_date

WHERE incidence_rate > 500;
```

What region had incidents over 2000?

```
SELECT region

GROUP BY region

HAVING incident > 2000;
```

How many people in median died with Covid-19 in January 21?

```
SELECT AVG(vaccinated_twice)

WHERE report_date > 20-12-31 AND
report_date < 21-02-01;
```

# Stage 2: Interactive SQL Processing

## Target structure definition

```
SELECT report_date, incidence_rate

WHERE incidence_rate > 500;
```

| Doc | Report_Date | Incidence |
|---|---|---|
| | | |

```
SELECT region

GROUP BY region

HAVING incident > 2000;
```

| Region |
|---|
| |

```
SELECT AVG(new_death), region

WHERE report_date > 20-12-31 AND
report_date < 21-02-01;

GROUP BY region
```

| Region | AVG<br>(new_death) |
|---|---|
| | |

# Stage 2: Interactive SQL Processing

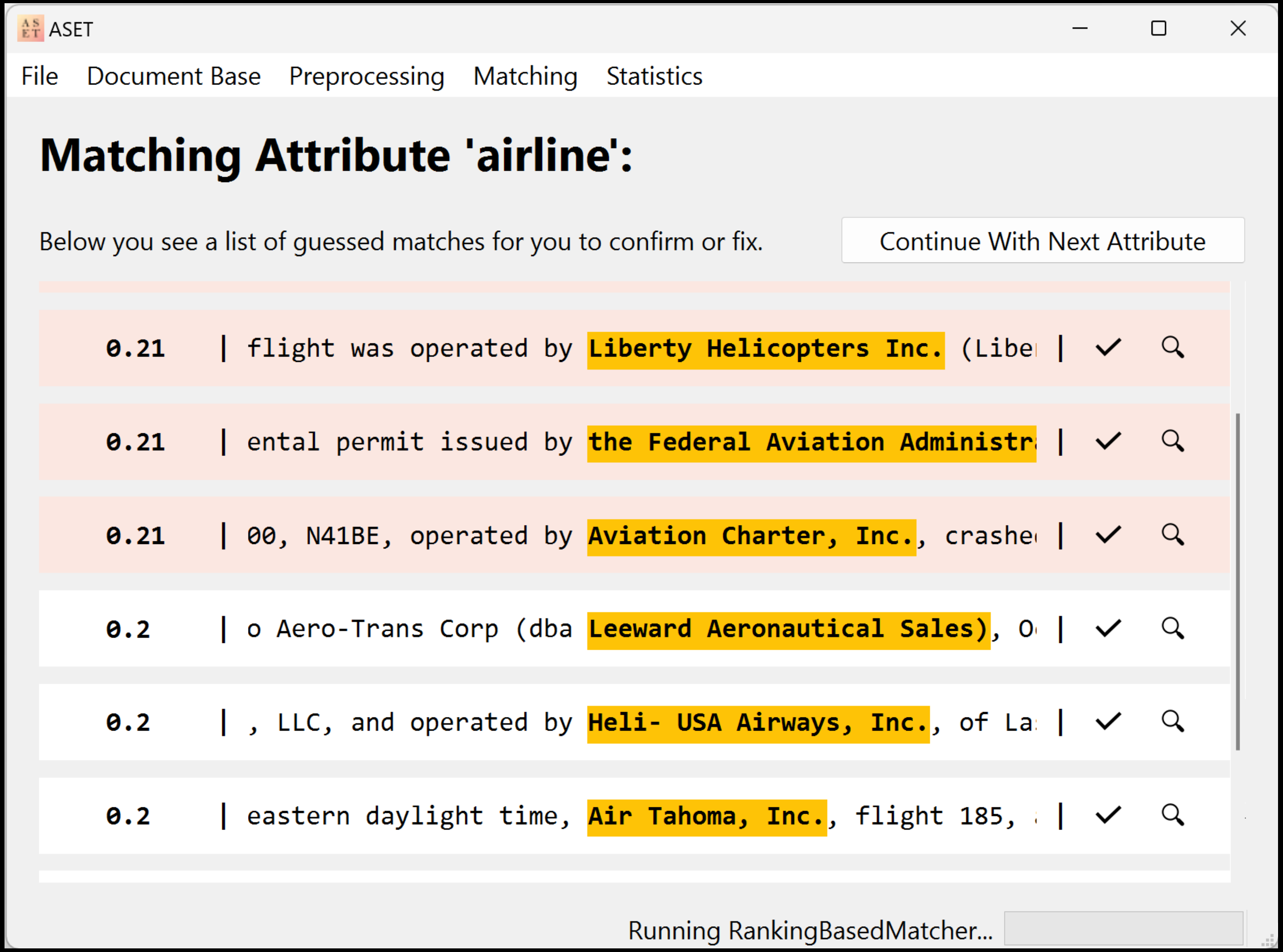
## Interactive table extraction

| Doc# | Report_date | Region | new_death |
|---|---|---|---|
| 1 | 21-1-1 | Upper-Franconia | 0.02 |
| 2 | 21-1-5 | North Bavaria | 2 |
| 3 | 21-1-20 | Munich | 1 |
| 4 | 21-1-31 | Middle-Franconia | - |

| North Bavaria |
|---|
| Franconia |
| Lower-Franconia ✓ |

# Stage 2: Interactive SQL Processing

## Interactive table extraction



### Matching Attribute 'airline':

Below you see a list of guessed matches for you to confirm or fix.

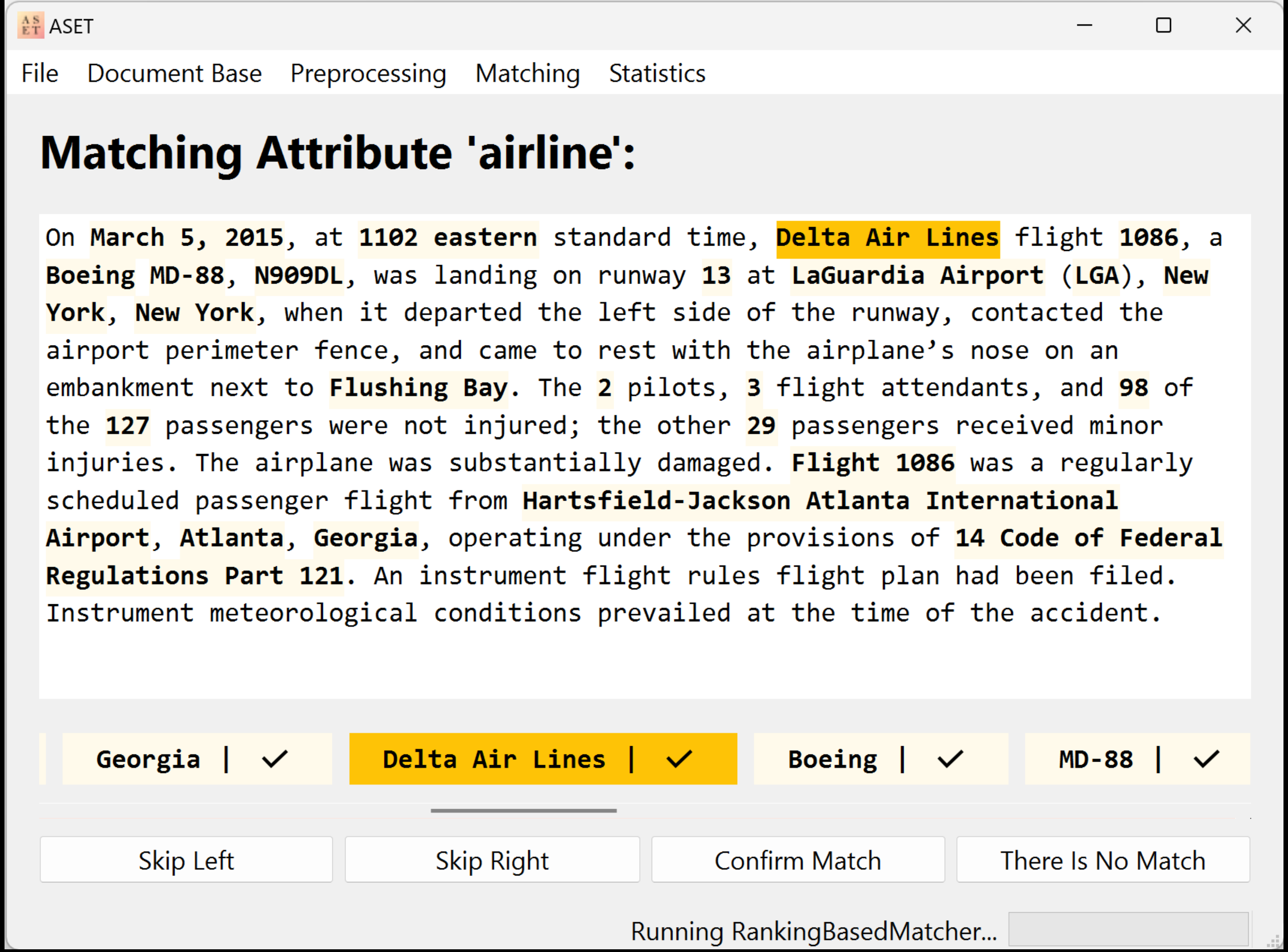
```
0.21   | flight was operated by Liberty Helicopters Inc. (Libe |
0.21   | ental permit issued by the Federal Aviation Administr |
0.21   | 00, N41BE, operated by Aviation Charter, Inc., crashe |
0.2    | o Aero-Trans Corp (dba Leeward Aeronautical Sales), O |
0.2    | , LLC, and operated by Heli- USA Airways, Inc., of La |
0.2    | eastern daylight time, Air Tahoma, Inc., flight 185,  |
```

Running RankingBasedMatcher...

Potential matches over and under the **threshold**
user either confirm or fix them
[Hättasch 23:160]

# Stage 2: Interactive SQL Processing

## Interactive table extraction



**Matching Attribute 'airline':**

On **March 5, 2015**, at **1102 eastern** standard time, **Delta Air Lines** flight **1086**, a **Boeing** **MD-88**, **N909DL**, was landing on runway **13** at **LaGuardia Airport** (**LGA**), **New York**, **New York**, when it departed the left side of the runway, contacted the airport perimeter fence, and came to rest with the airplane’s nose on an embankment next to **Flushing Bay**. The **2** pilots, **3** flight attendants, and **98** of the **127** passengers were not injured; the other **29** passengers received minor injuries. The airplane was substantially damaged. **Flight 1086** was a regularly scheduled passenger flight from **Hartsfield-Jackson Atlanta International Airport**, **Atlanta**, **Georgia**, operating under the provisions of **14 Code of Federal Regulations Part 121**. An instrument flight rules flight plan had been filed. Instrument meteorological conditions prevailed at the time of the accident.

Running RankingBasedMatcher...

Inspecting a document and selecting right match
[Hättasch 23:160]

# Evaluation

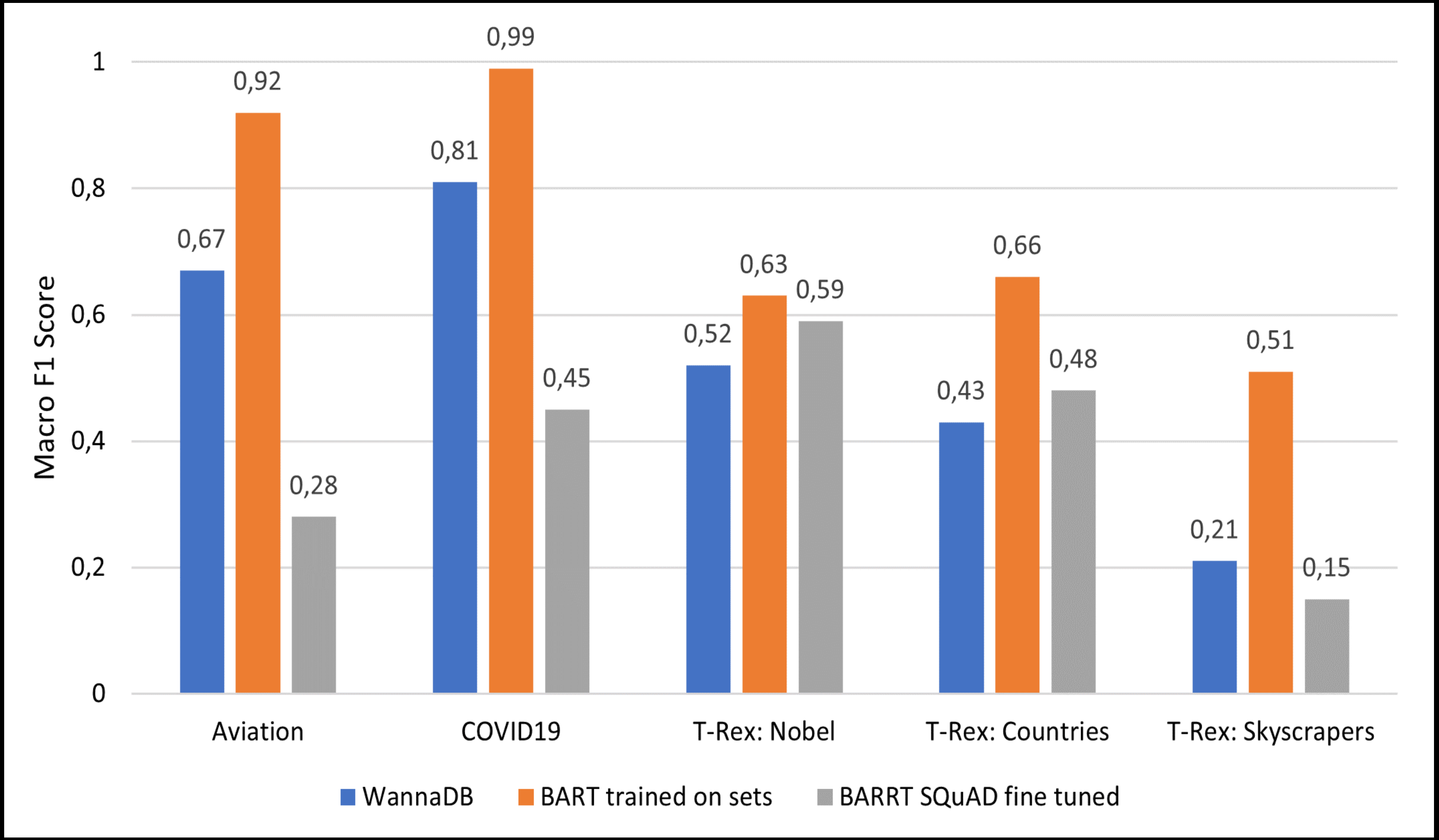
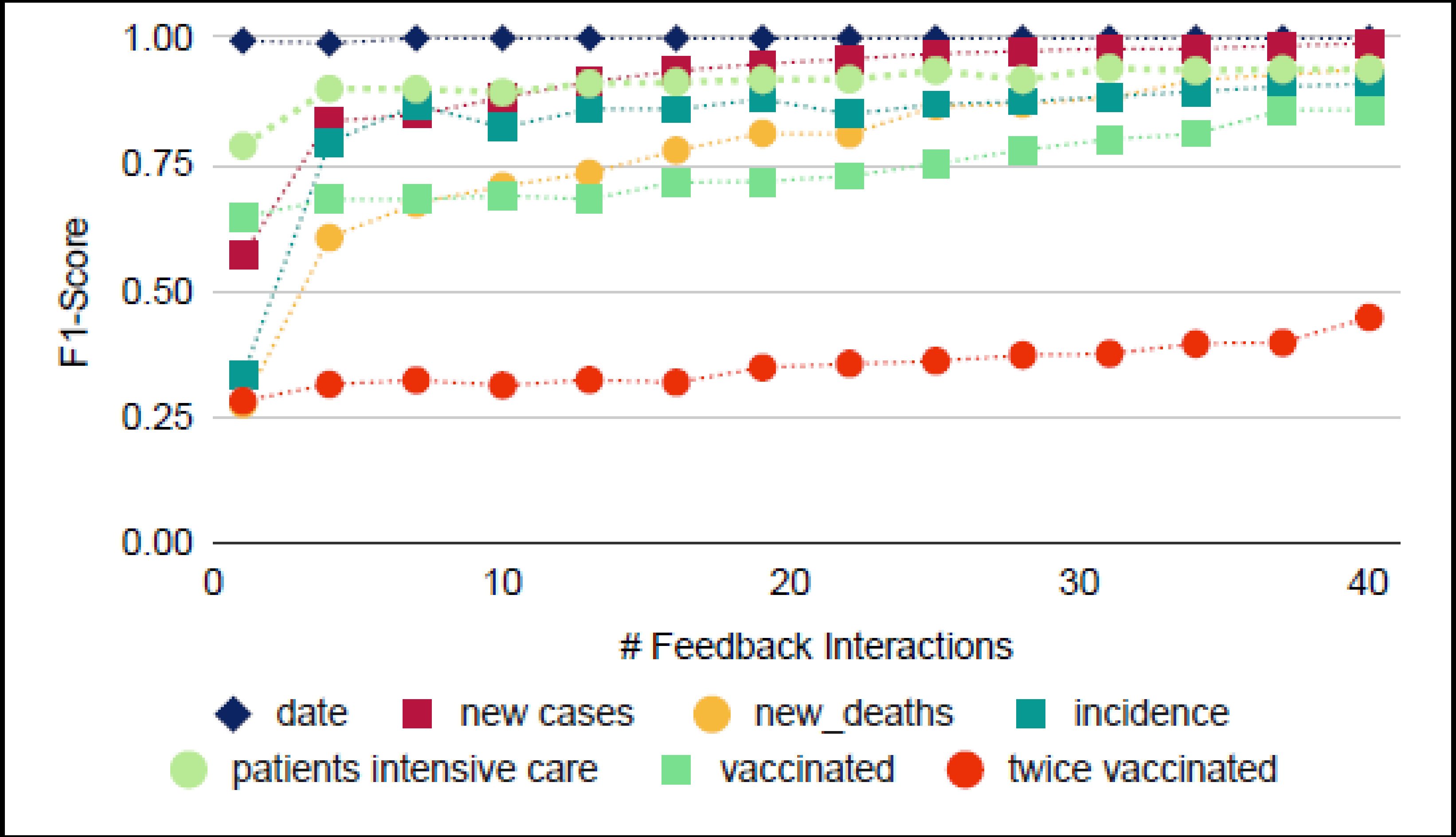


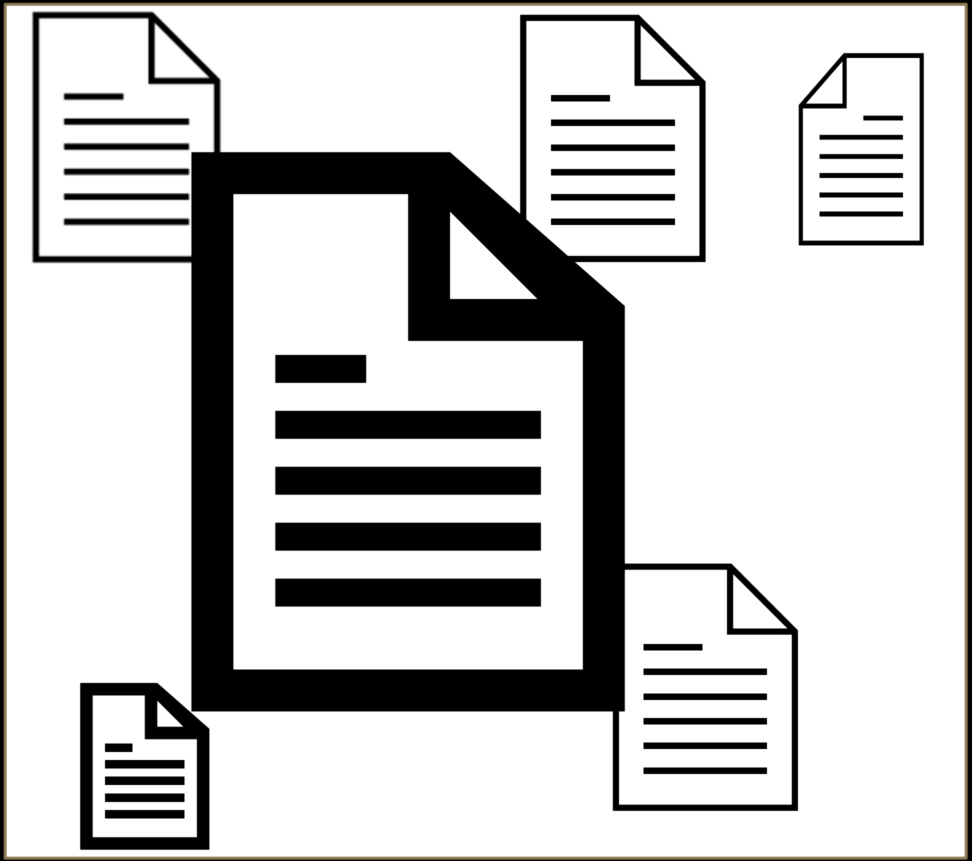
Table filling results in WannaDB compared to: explicit trained BART and SQuAD 2.0 fine tuned BART

# Evaluation



F1-Score for amounts of feedback iterations per attribute (1-40) [Hättasch 23:174]

# Conclusion



```
SELECT author, COUNT(*) as c

GROUP BY author
```

| c | author |
|---|---|
| 20 | John Doe |
| 11 | Anon |
| 42 | Anonymus |
| 2 | Mike |
| … | … |

# References

Paper discussed:

- Benjamin Hättasch, Jan-Micha Bodensohn, Liane Vogel, Matthias Urban, and Carsten Binnig. 2023. WannaDB: Ad-hoc SQL Queries over Text Collections. In BTW 2023, Birgitta König-Ries, Stefanie Scherzinger, Wolfgang Lehner, and Gottfried Vossen (Eds.). Gesellschaft für Informatik e.V. https://doi.org/10.18420/BTW2023-08

Further references:

- Benjamin Hättasch, Jan-Micha Bodensohn, and Carsten Binnig. 2022. Demonstrating ASET: Ad-Hoc Structured Exploration of Text Collections. In Proceedings of the 2022 International Conference on Management of Data (Philadelphia, PA, USA) (SIGMOD ’22). Association for Computing Machinery, New York, NY, USA, 2393–2396. https://doi.org/10.1145/3514221.3520174

- TheWannaDB code is available at https://github.com/DataManagementLab/wannadb

# thank you for your attention

# further slides

# Evaluation

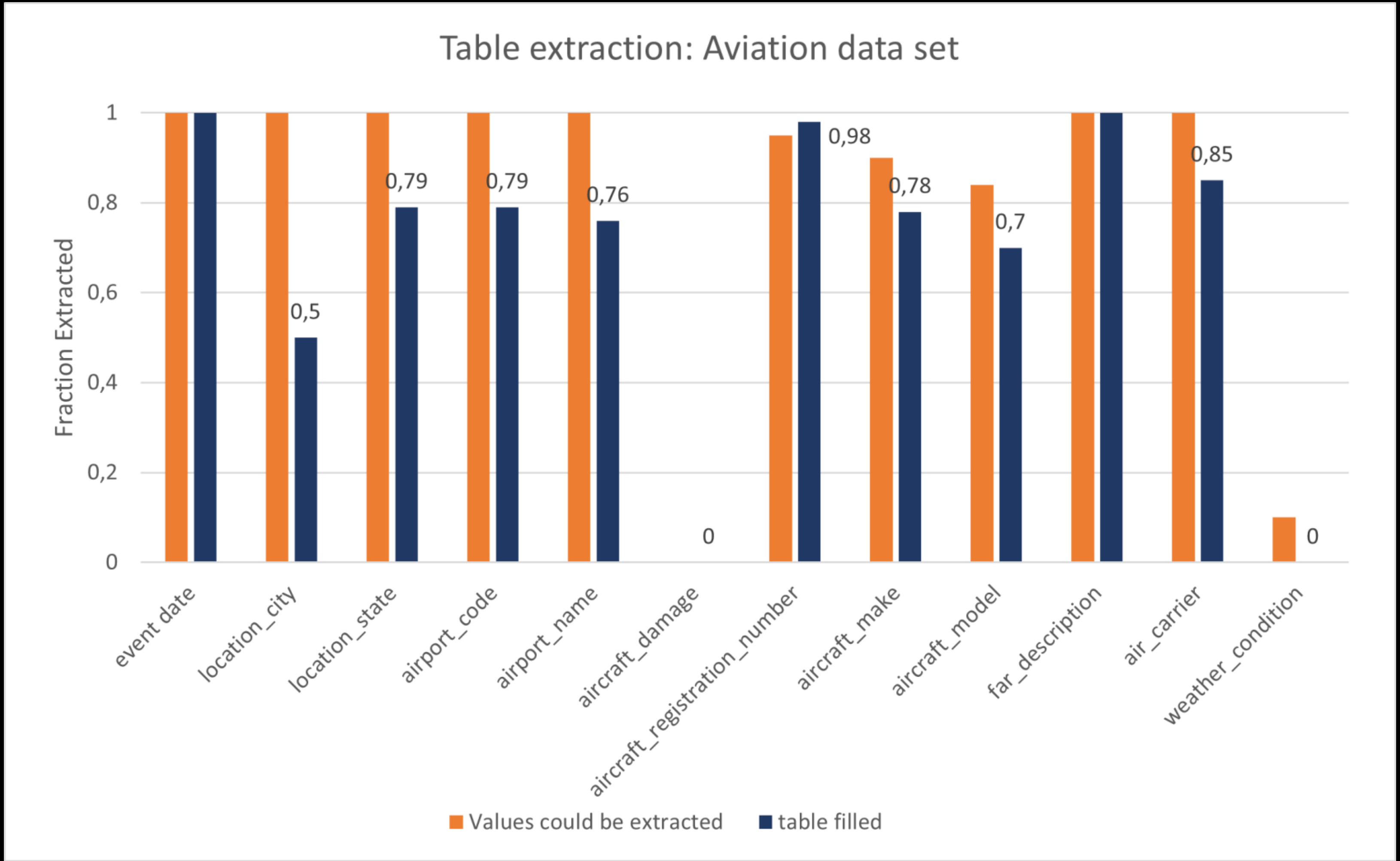


Table filling results in WannaDB compared to: data, that could be extracted

# Evaluation

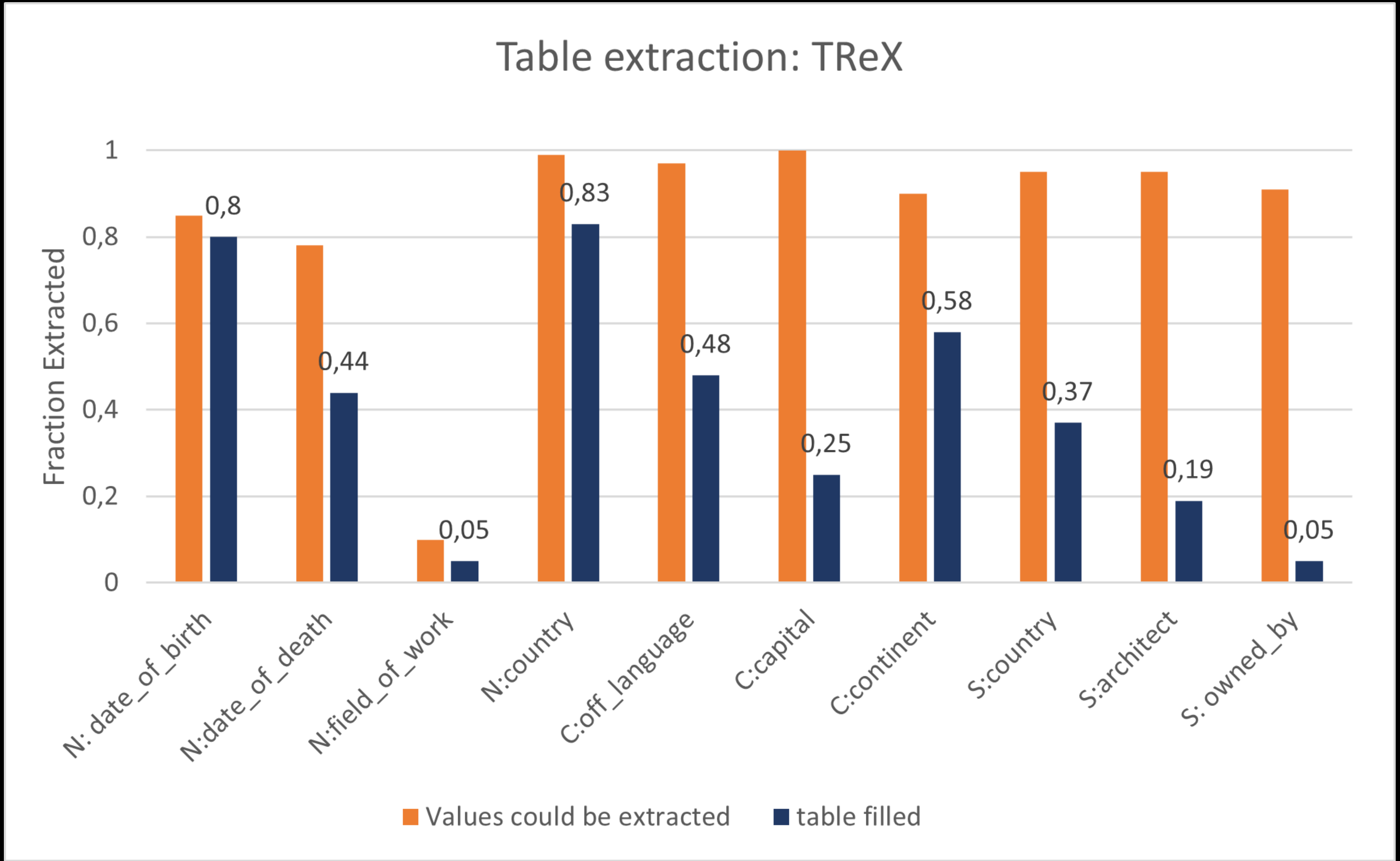


Table filling results in WannaDB compared to: data, that could be extracted

# threshold

```
for attribute in query.attributes: # Process each attribute separately
    while interactive_feedback_phase: # Interactively get user feedback
        […]
        update_guessed_matches(documents)
        adjust_threshold(feedback)

    for document in documents: # Only consider values up to a given maximum distance
        if current_guess(document).distance < threshold:
            set_match(document, current_guess(document)) # compute final result table
        else:
            leave_empty(document)

def adjust_threshold(feedback): # Feedback can be further exploited in certain cases
    match feedback:
        case ConfirmNugget(document, confirmed_nugget):
            if confirmed_nugget.distance > threshold: increase_threshold(confirmed_nugget)
        case NoMatchInDocument(document):
            if current_guess(document).distance < threshold: decrease_threshold(document)
```

# threshold (con‘t)

```
def decrease_threshold(document): # Consider fewer matches as valid (especially those
                                    above last marking as incorrect that are currently accepted
                                    nevertheless)
    nuggets = ranked_list.between(threshold, document)
    min_dist = min(n.distance for n in nuggets)
    threshold = min(min_dist, threshold)

def increase_threshold(confirmed_nugget): # Consider more matches as valid (especially
                                            those below last confirmation that are currently
                                            discarded because of the threshold)
    nuggets = ranked_list.between(confirmed_nugget, threshold)
    max_dist = max(n.distance for n in nuggets)
    threshold = max(max_dist, threshold)
```

Hättasch et al 2023:165