# Data Discovery in Data Lakes

Big Data Management and Governance

Prof. Giovanni Simonini, Giovanni Malaguti

University of Modena and Reggio Emilia

December 9, 2025

# Details of the Lab (1)

- We will apply BLEND, a framework to perform data discovery in data lakes.
- We will create the index, understand its key concepts and which use-cases it can target.
- In general, we will understand what are the main challenges when searching among large collection of datasets.

# Details of the Lab (2)

- Clone (or update) the repository `https://github.com/Stravanni/bdm.git`.
- Solutions will be uploaded at the end of the lab.

# Details of the Lab (3-*nix/Mac)

Open a shell and move to the current lab folder.

```
$ cd /path/to/the/cloned/repo
```

There create the Python virtual environment. You can use any python environment manager (conda, uv, poetry, ...). Here, for simplicity, we will use the Python venv module:

```
(skiplist-hnsw)$ python -m venv .venv
```

Activate the environment:

```
(skiplist-hnsw)$ source .venv/bin/activate
```

Install the required packages:

```
(skiplist-hnsw)$ pip install -r requirements.txt
```

# Details of the Lab (3-Windows)

Open a command-line prompt (e.g. Powershell). Then, move to the current lab folder.

```
$ cd path\to\the\cloned\repo
```

There create the Python virtual environment. You can use any python env manager (conda, uv, poetry, ...). Here, for simplicity, we will use the Python venv module:

```
(skiplist-hnsw)$ python -m venv .venv
```

Activate the environment:

```
(skiplist-hnsw)$ .venv/Scripts/activate
```

Install the required packages:

```
(skiplist-hnsw)$ pip install -r requirements.txt
```

# Details of the Lab (4)

To open a Jupyter notebook, you can use VS Code, as it supports well notebooks, or you can install jupyter and then run in a terminal ”jupyer notebook” to open the jupyer client and modify the files from there.

# Details of the Lab (5)

In the folder bdm/lab/data-discovery/data there are the files modena.zip and undata.zip
Extract the files and place them inside the directory ”data”.

# Data Discovery in Data Lakes

# Data Discovery in Data Lakes

- Process of identifying datasets that may meet an information need
- Given a user table, we may want to enhance its content by integrating it with additional information contained in related tables from a table corpus
- How to detect related tables on large-scale scenarios?

# Data Discovery in Data Lakes

The main data discovery tasks are:

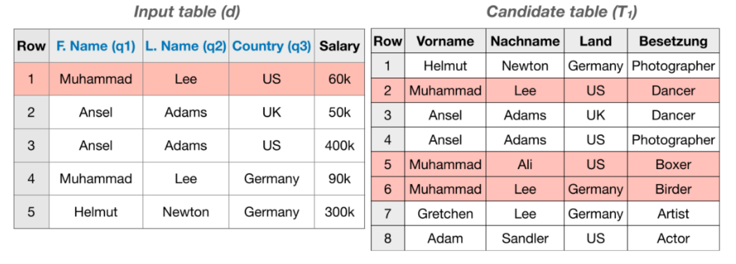
- Keyword Search;
- Join Search;
- Union Search;
- Join-Correlation Search;

and there are two main approaches to solve these tasks:

- Overlap-based;
- Semantic-based;

# Data Discovery: Join Discovery

Given a user dataset and one of its columns, we want to identify all those tables in the table corpus that can be joined with it on the query column.
For instance, the table below can be joined on the “Postal” and “Postal Code” columns.



*Input table (d)*

| Row | F. Name (q1) | L. Name (q2) | Country (q3) | Salary |
|---|---|---|---|---|
| 1 | Muhammad | Lee | US | 60k |
| 2 | Ansel | Adams | UK | 50k |
| 3 | Ansel | Adams | US | 400k |
| 4 | Muhammad | Lee | Germany | 90k |
| 5 | Helmut | Newton | Germany | 300k |

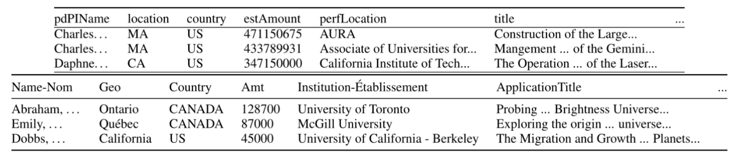
*Candidate table ($T_1$)*

| Row | Vorname | Nachname | Land | Besetzung |
|---|---|---|---|---|
| 1 | Helmut | Newton | Germany | Photographer |
| 2 | Muhammad | Lee | US | Dancer |
| 3 | Ansel | Adams | UK | Dancer |
| 4 | Ansel | Adams | US | Photographer |
| 5 | Muhammad | Ali | US | Boxer |
| 6 | Muhammad | Lee | Germany | Birder |
| 7 | Gretchen | Lee | Germany | Artist |
| 8 | Adam | Sandler | US | Actor |

**Figure 1:** Example of Joinable Tables from [1]

# Data Discovery: Union Discovery

Given a user dataset, we want to identify all those tables in the table corpus that can be unioned with the query dataset, on all its columns or on a relevant subset
For instance, the table below can be unioned on all their columns.



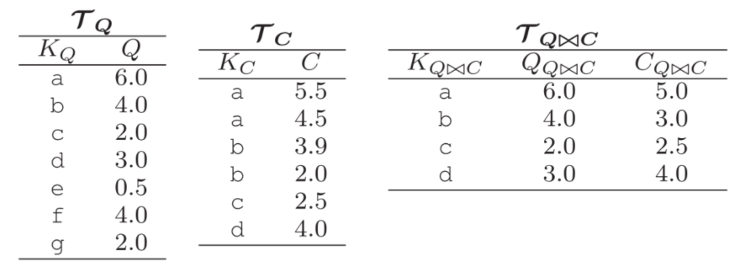
| pdPIName | location | country | estAmount | perfLocation | title | ... |
|---|---|---|---|---|---|---|
| Charles. . . | MA | US | 471150675 | AURA | Construction of the Large... | |
| Charles. . . | MA | US | 433789931 | Associate of Universities for... | Mangement ... of the Gemini... | |
| Daphne. . . | CA | US | 347150000 | California Institute of Tech... | The Operation ... of the Laser... | |

| Name-Nom | Geo | Country | Amt | Institution-Établissement | ApplicationTitle | ... |
|---|---|---|---|---|---|---|
| Abraham, . . . | Ontario | CANADA | 128700 | University of Toronto | Probing ... Brightness Universe... | |
| Emily, . . . | Québec | CANADA | 87000 | McGill University | Exploring the origin ... universe... | |
| Dobbs, . . . | California | US | 45000 | University of California - Berkeley | The Migration and Growth ... Planets... | |

**Figure 2:** Example of Unionable Tables from [3]

# Data Discovery: Join-Correlation Discovery

Differently from the basic join discovery task, in this case we want to sort our results on a after-join correlation between a numerical column of our query dataset and another column from the joined dataset.



$\mathcal{T}_Q$

| $K_Q$ | $Q$ |
|---|---|
| a | 6.0 |
| b | 4.0 |
| c | 2.0 |
| d | 3.0 |
| e | 0.5 |
| f | 4.0 |
| g | 2.0 |

$\mathcal{T}_C$

| $K_C$ | $C$ |
|---|---|
| a | 5.5 |
| a | 4.5 |
| b | 3.9 |
| b | 2.0 |
| c | 2.5 |
| d | 4.0 |

$\mathcal{T}_{Q \bowtie C}$

| $K_{Q \bowtie C}$ | $Q_{Q \bowtie C}$ | $C_{Q \bowtie C}$ |
|---|---|---|
| a | 6.0 | 5.0 |
| b | 4.0 | 3.0 |
| c | 2.0 | 2.5 |
| d | 3.0 | 4.0 |

**Figure 3:** Example of Join-Correlated Tables from [4]

# Data Discovery: Overlap-based approaches

Manipulate the datasets values in order to quickly identify exact overlaps between a query table and any other table in the corpus. In some cases, probabilistic data structures are (MinHash, LSH, etc.) are used.

Pros:

- More scalable, simpler and effective than semantic-based for some use-cases;
- Do not require specialized models or tools (embedding models);

Cons:

- Semantic relationships cannot be identified;

In our lab session we will use BLEND [2], which belongs to this class.

# Data Discovery: Semantic-based approaches

Semantic-based approaches rely on embedding datasets at different granularities (row, column, cell, etc.) and index the embedding vectors into a vector index.

Pros:

- Fast query performance;
- Semantic matches;

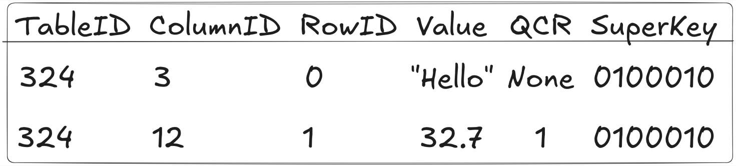
Cons:

- Not easy to train a model for table embedding;
- Highly-specialized indexes;
- GPU is mandatory for large-scale scenarios;

# Data Discovery with BLEND

BLEND extends the DataXFormer inverted idex to support also n-ary Join Discovery and Join-Correlation Discovery searches.

- **SuperKey**: from [1], allows to search for n-ary joins;
- **QCR**: from [4], allows to perform join-correlation queries.



| TableID | ColumnID | RowID | Value | QCR | SuperKey |
|---|---|---|---|---|---|
| 324 | 3 | 0 | "Hello" | None | 0100010 |
| 324 | 12 | 1 | 32.7 | 1 | 0100010 |

**Figure 4:** BLEND record examples.

# References i

- Esmailoghli, M., Quiané-Ruiz, J.-A., and Abedjan, Z.
  Mate: multi-attribute table extraction.
  *arXiv preprint arXiv:2110.00318* (2021).

- Esmailoghli, M., Schnell, C., Miller, R. J., and Abedjan, Z.
  Blend: A unified data discovery system.
  In *2025 IEEE 41st International Conference on Data Engineering (ICDE)* (2025), pp. 737–750.

- Nargesian, F., Zhu, E., Pu, K. Q., and Miller, R. J.
  Table union search on open data.
  *Proceedings of the VLDB Endowment 11*, 7 (2018), 813–825.

# References ii

Santos, A., Bessa, A., Musco, C., and Freire, J.
A sketch-based index for correlated dataset search.
In *2022 IEEE 38th International Conference on Data Engineering (ICDE)* (2022), pp. 2928–2941.