Table Discovery in Data Lakes: State-of-the-Art and Future Directions

Grace Fan, Jin Wang, Yuliang Li, Renée J. Miller

https://northeastern-datalab.github.io/table-discovery-tutorial/

Outline

- Introduction
- Table Understanding
- Table Search Engine
- Table Navigation and Exploration
- Data Science and Application Support
- Conclusion / Future work

Data Analytics Landscape

Major shift from ***Data Warehouses*** → ***Data Lakes***

Thousands of tables → ***Millions of files***

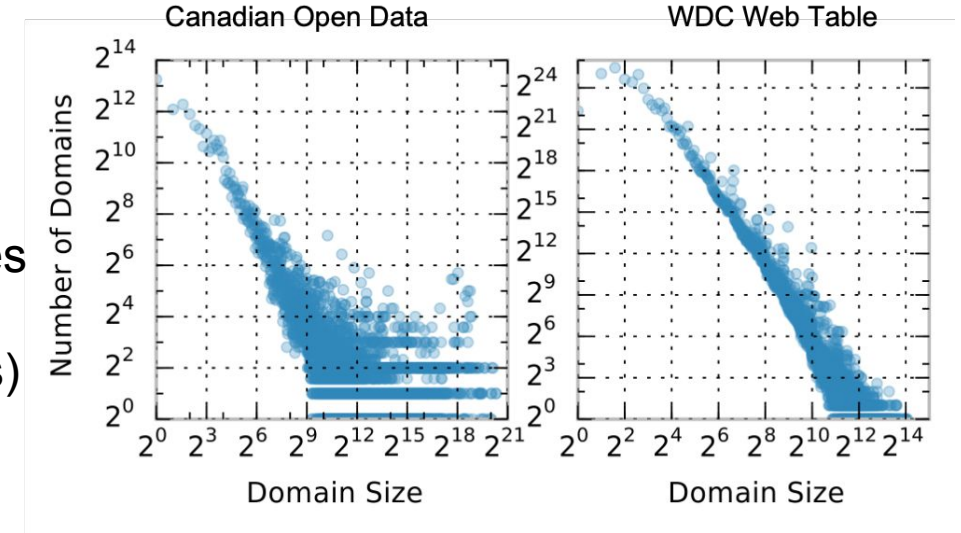
Well-designed Metadata → ***Unreliable, inconsistent metadata***

Example Data Lakes

- Open government data
  - **[Miller VLDB 18]**
- Web tables
  - **[Eberius et al. BDC 15], [Lehmber et al. WWW 16] and others**
- Enterprise data lakes
- Domain-specific lakes

Real data lakes are skewed & large

- Large differences in attribute sizes
- Large differences in table sizes
- Massive vocabularies (no. values)

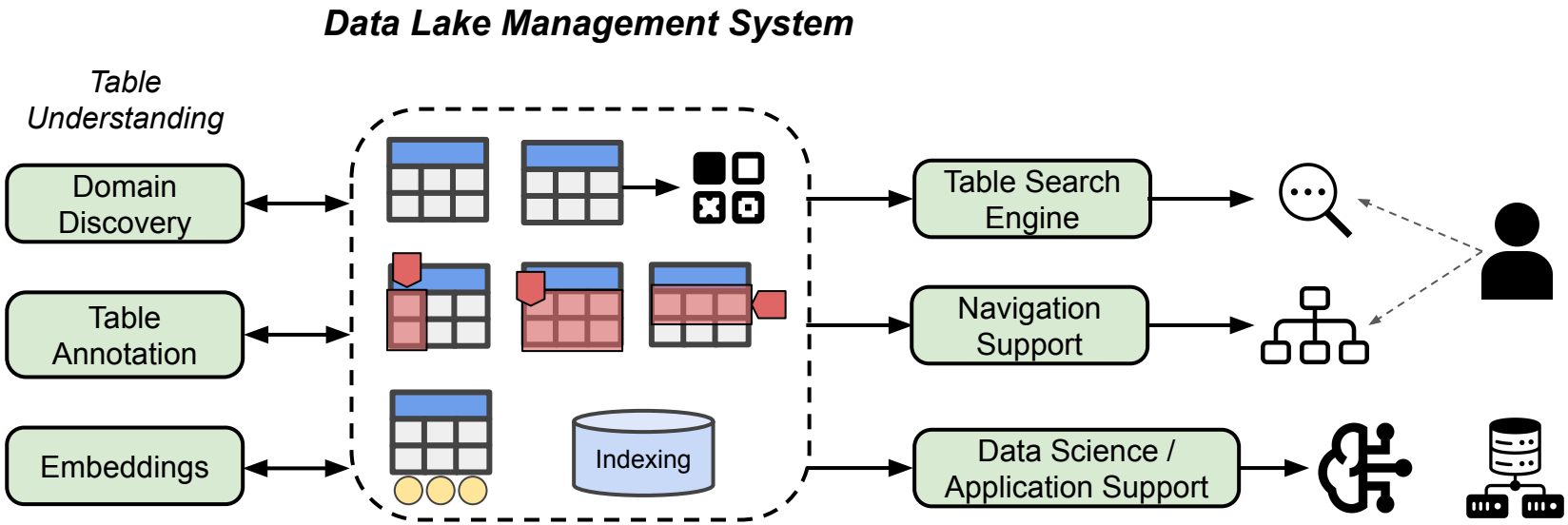


Data Science Over Data Lakes

In data science, it is increasingly the case that the main challenge is not in ***integrating known data***, rather it is in ***finding the right data to solve a given data science problem***.

How can we facilitate data science over data lakes?

Architecture of a Table Discovery System



Outline

- Introduction
- Table Understanding
- Table Search Engine
- Table Navigation and Exploration
- Data Science and Application Support
- Conclusion / Future work

Table Understanding

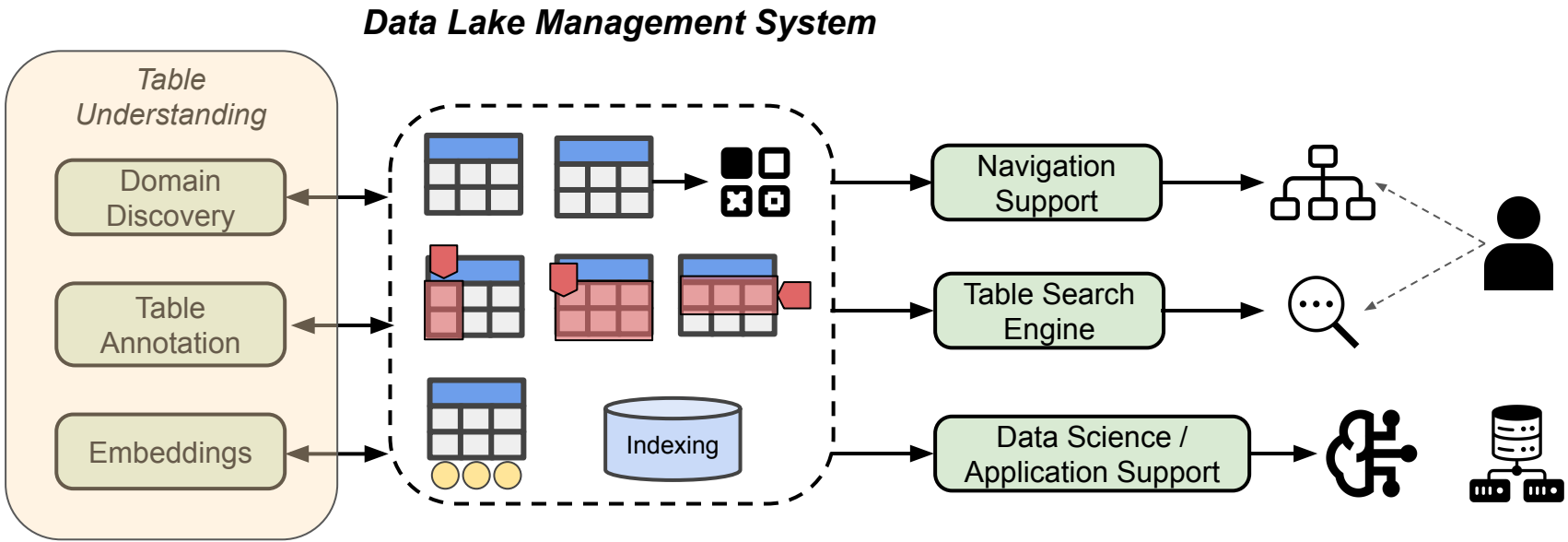
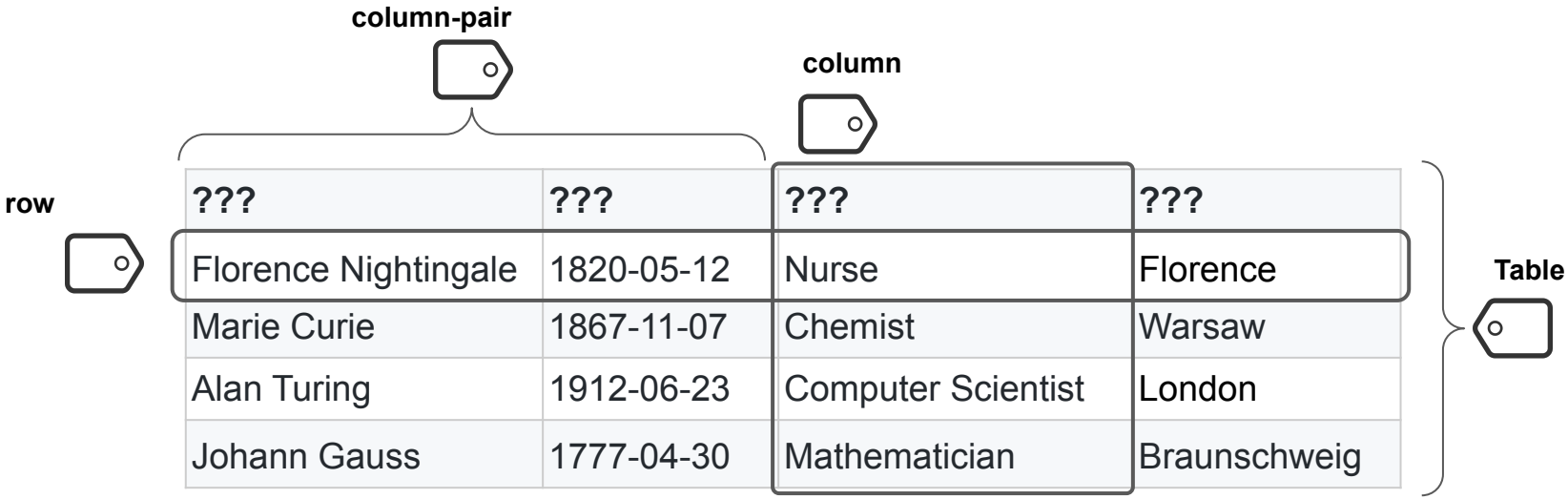


Table Understanding tasks on different levels

- Column: Semantic Type Detection, Domain Discovery
- Column-pair: Relation Extraction
- Row: Row-level entity linking
- Table: Table-level entity linking, topic modeling



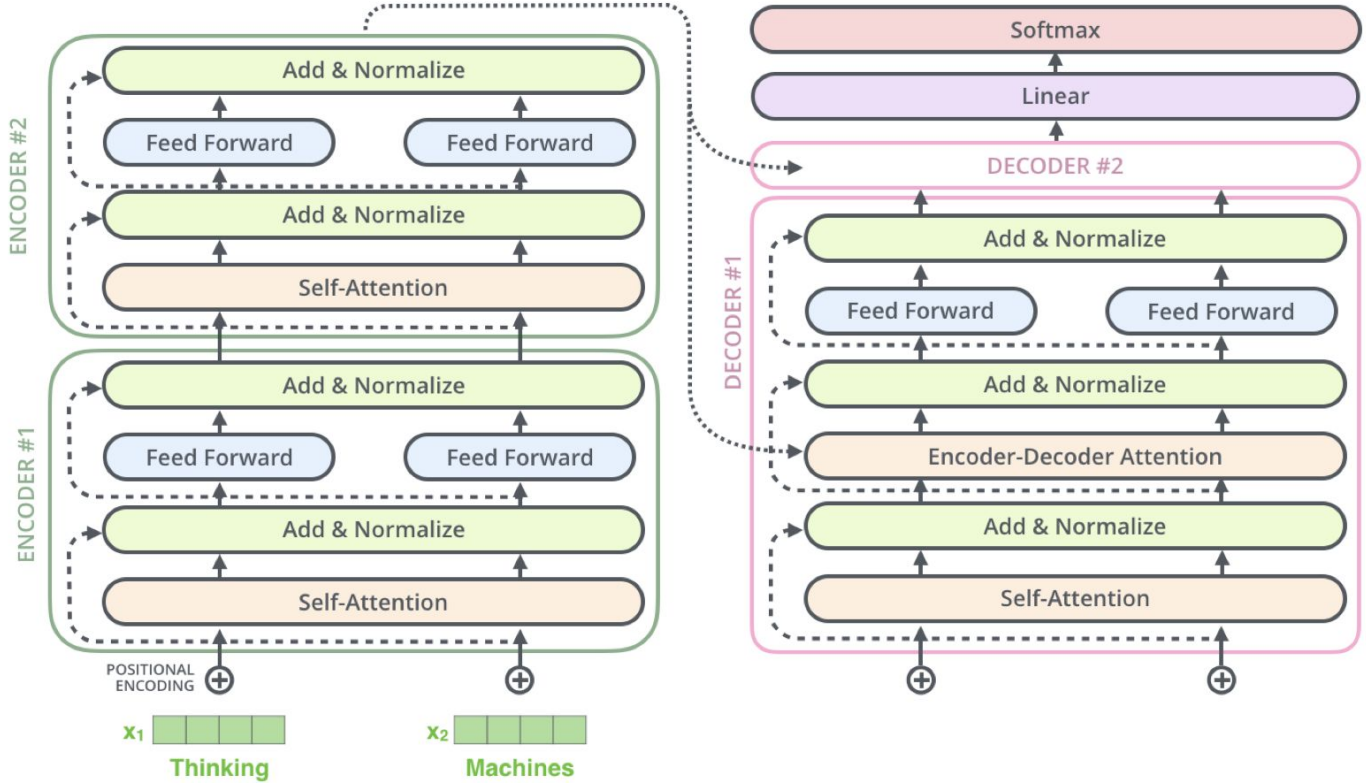
???	???	???	???
Florence Nightingale	1820-05-12	Nurse	Florence
Marie Curie	1867-11-07	Chemist	Warsaw
Alan Turing	1912-06-23	Computer Scientist	London
Johann Gauss	1777-04-30	Mathematician	Braunschweig

Techniques for Table Understanding

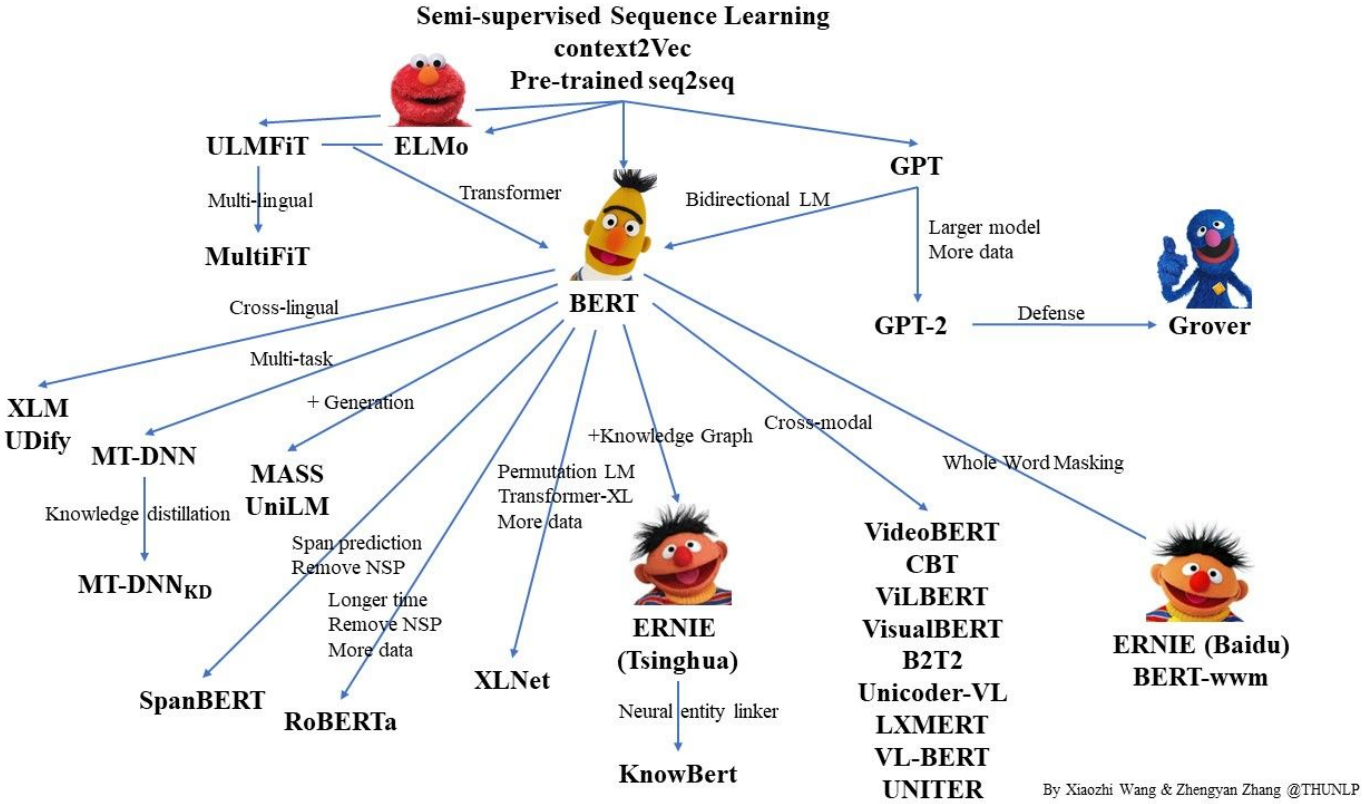
- Traditional Techniques, such as
  - Probabilistic models
  - Ontology matching
  - Feature engineering
- Nowadays Transformer-based techniques have been widely adopted in Table Understanding problems!

[Badaro et al. VLDB 22]

Background: Transformer



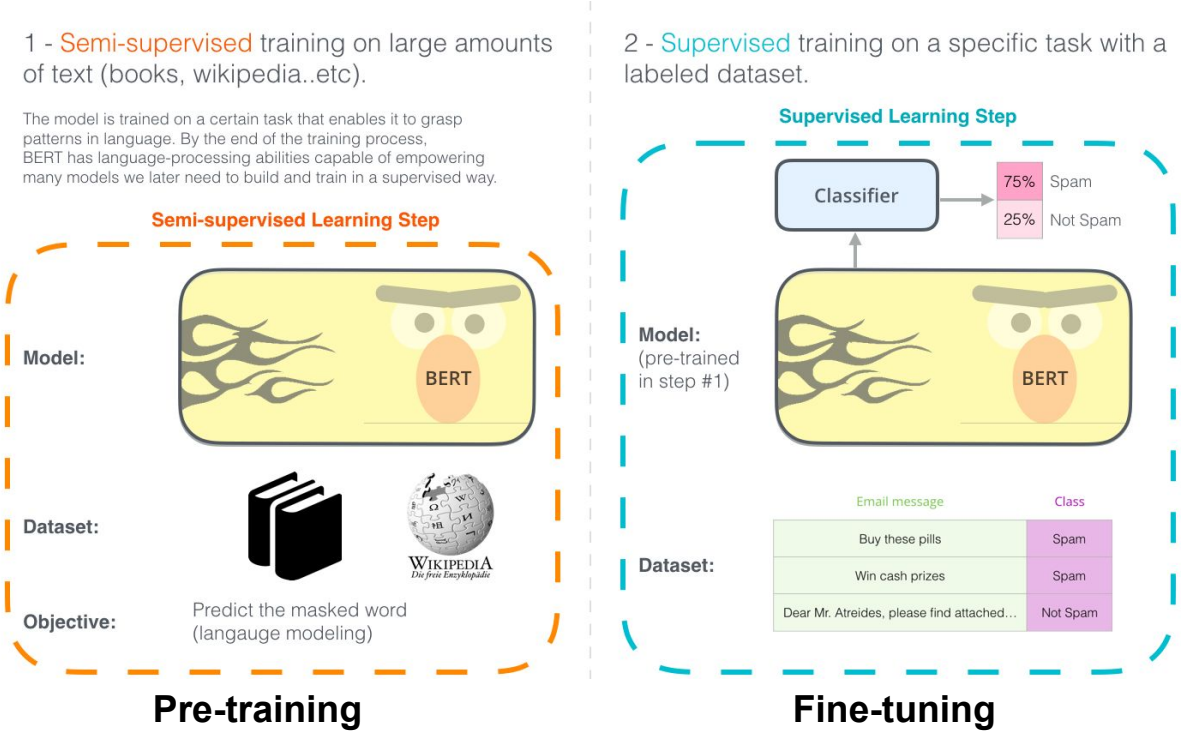
Many Transformer-based pre-trained models



https://github.com/birdx0810/NLP/blob/master/NLP_Papers.md

Background: Pre-trained language models

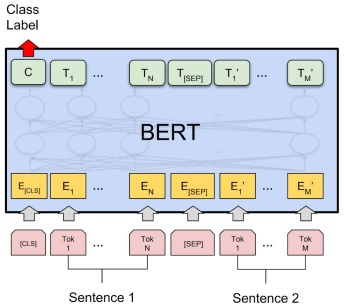
- Pre-training + fine-tuning framework



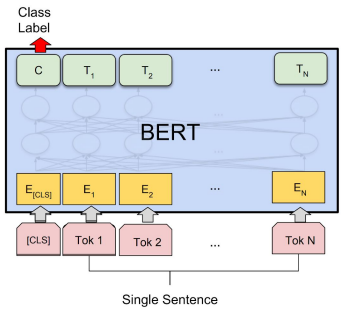
http://jalammar.github.io/illustrated-bert/

Why are BERT-style models so popular?

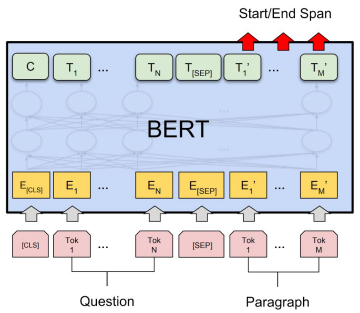
- The model can be used for many many NLP tasks
  - What we need is just to fine-tune on training data of the target task



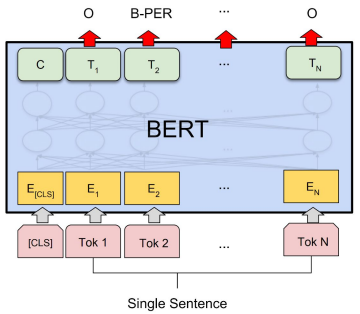
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA



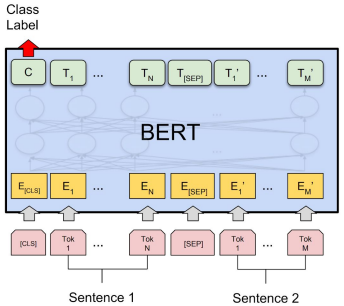
(c) Question Answering Tasks: SQuAD v1.1



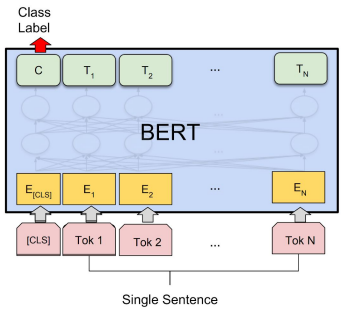
(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Why are BERT-style models so popular?

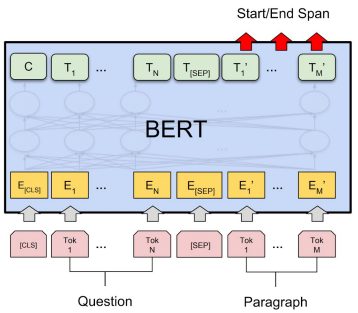
- The model can be used for many many NLP tasks
  - What we need is just to fine-tune on training data of the target task



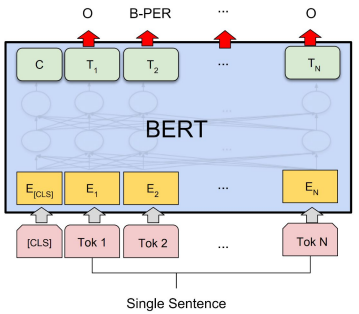
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA



(c) Question Answering Tasks: SQuAD v1.1



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

A Swiss army knife for NLP!

Table Serialization

- Flattening a table into a token sequence

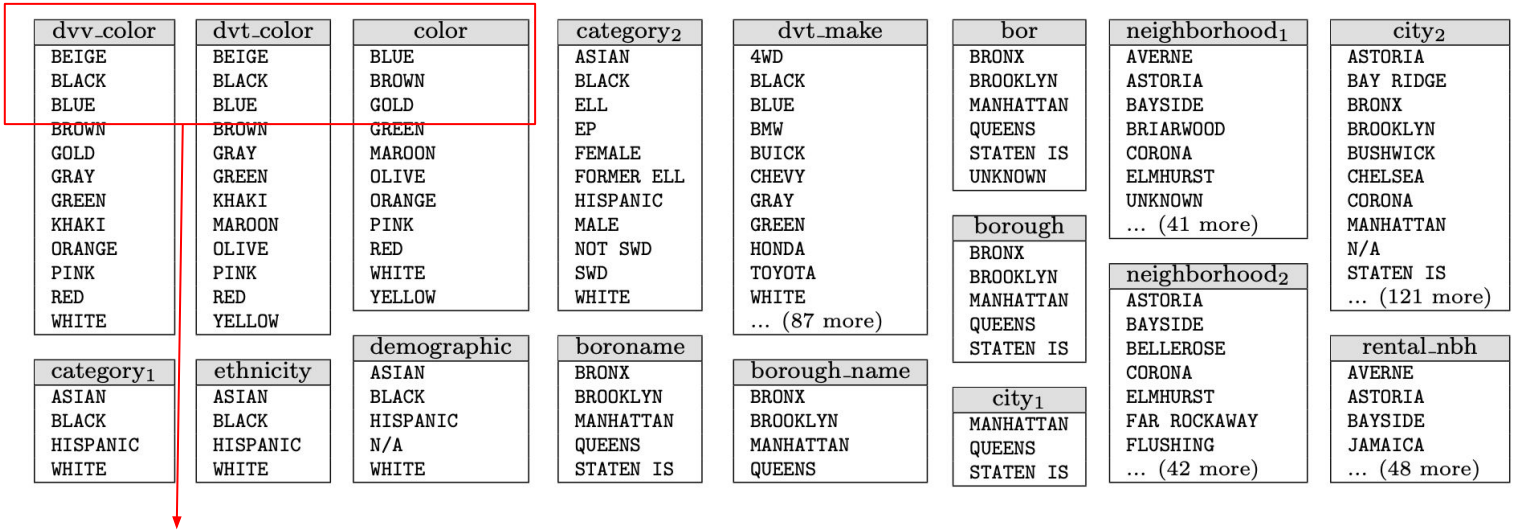
Year	Champion	Motorcycle
1949	Nello Pagani	
1950	Bruno Ruffo	
1951	Carlo Ubbiali	
1957	Tarquinio Provini	

```
[COL] Year [VAL] 1949 [COL] Champion [VAL] Nello Pagani
Year | 1949 | Champion | Nello Pagani
Year is 1949 and Champion is Nello Pagani
...
```

There are several options for serialization

Domain Discovery: Overview

- Goal: given a dataset of columns, find a set of domains representing all semantic types in the dataset



Domain: Color = {BEIGE, BLACK, BLUE, BROWN, GOLD, GRAY, GREEN, KHAKI, MAROON, OLIVE, ORANGE, PINK, RED, WHITE, YELLOW}

Domain Discovery: solutions

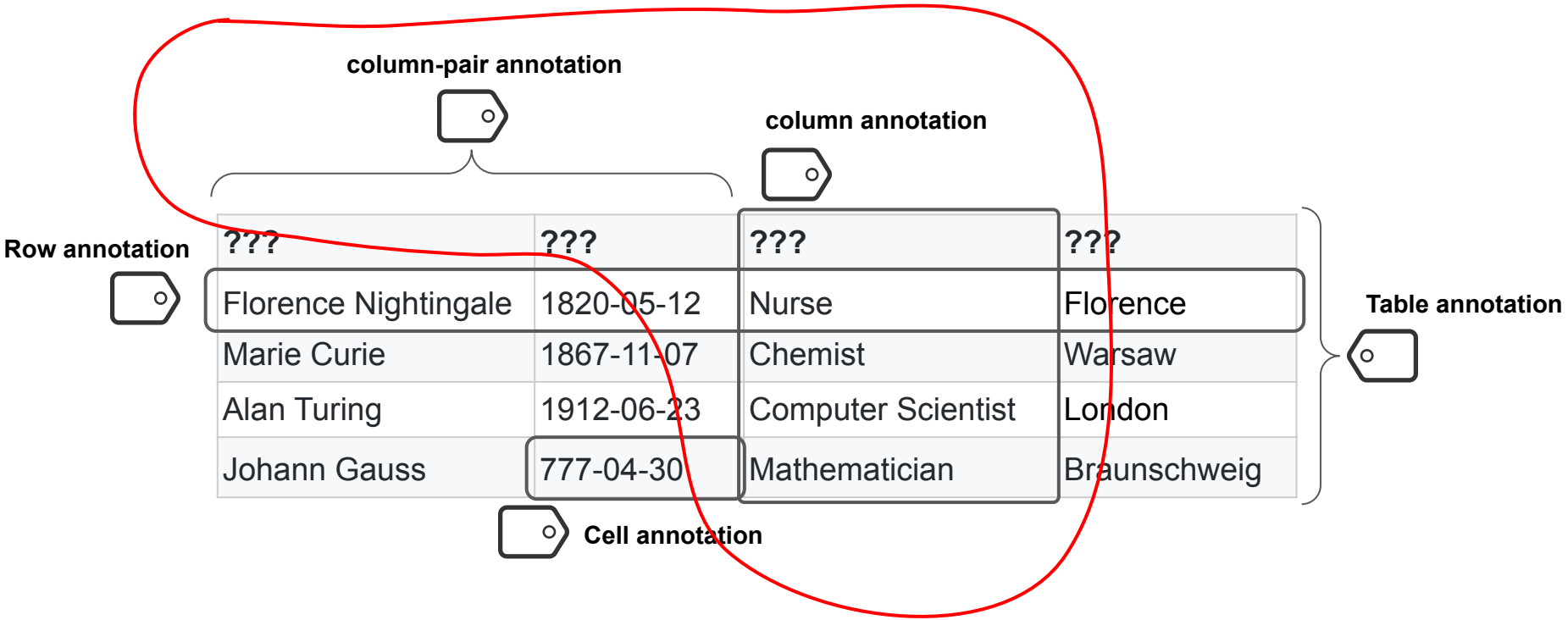
- C^4: Discovering Enterprise Concepts Using Spreadsheet Tables
  - Use co-occurrence information to build a concept hierarchy
  - Select nodes from the hierarchy that represent important concept

[Li et al. KDD 17]

- D4: Data-Driven Domain Discovery for Structured Datasets
  - An end-to-end, data-driven approach
  - Robust context based signatures
  - Scalable column-based clustering algorithm

[Ota et al. VLDB 20]

Table annotation: Toward Making Tables Make More Sense



[Limaye et al. VLDB 10]

[Hulsebos et al. KDD 19]

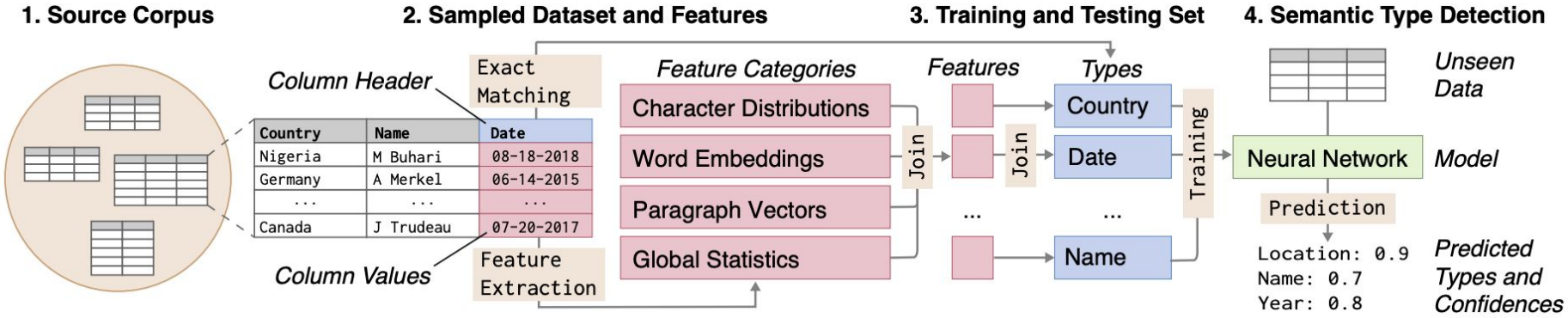
Sherlock: Feature based approach

Column type detection: formulated as multi-class classification (w/ supervised learning)

Input: Raw column values

Model: Multi-input Neural Network

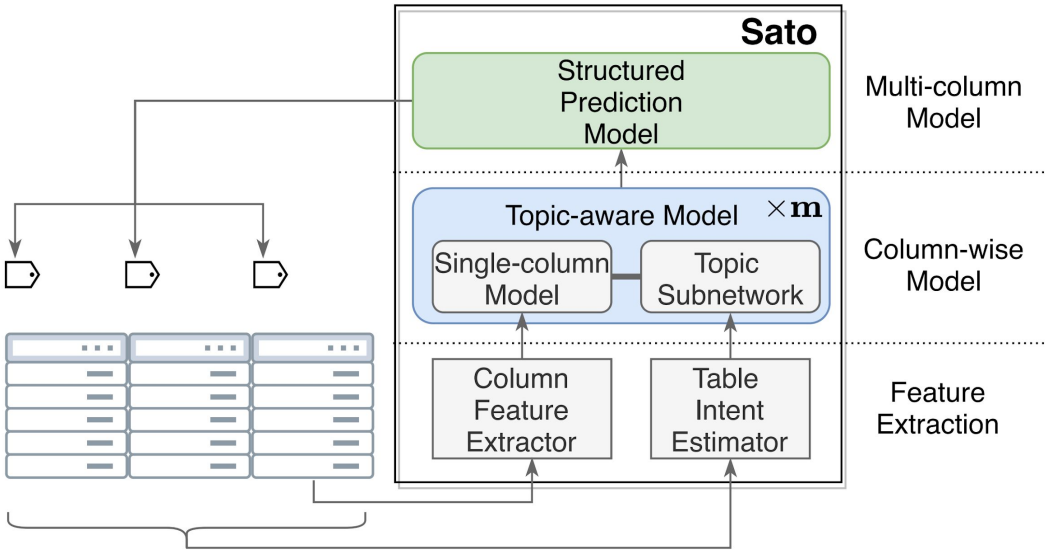
Output: Semantic type of column (from 78 candidate classes)



[Zhang et al. VLDB 20]

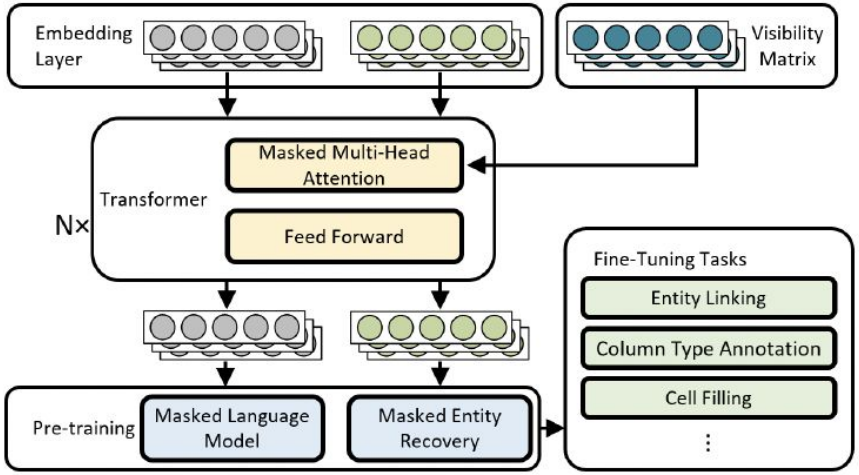
Sato: Adding column context

- Column-wise + Table-wise features + Structured prediction
- Enhance with Topic models
- Manually crafted features



[Deng et al. VLDB 21]

TURL: Pre-trained model for tabular data



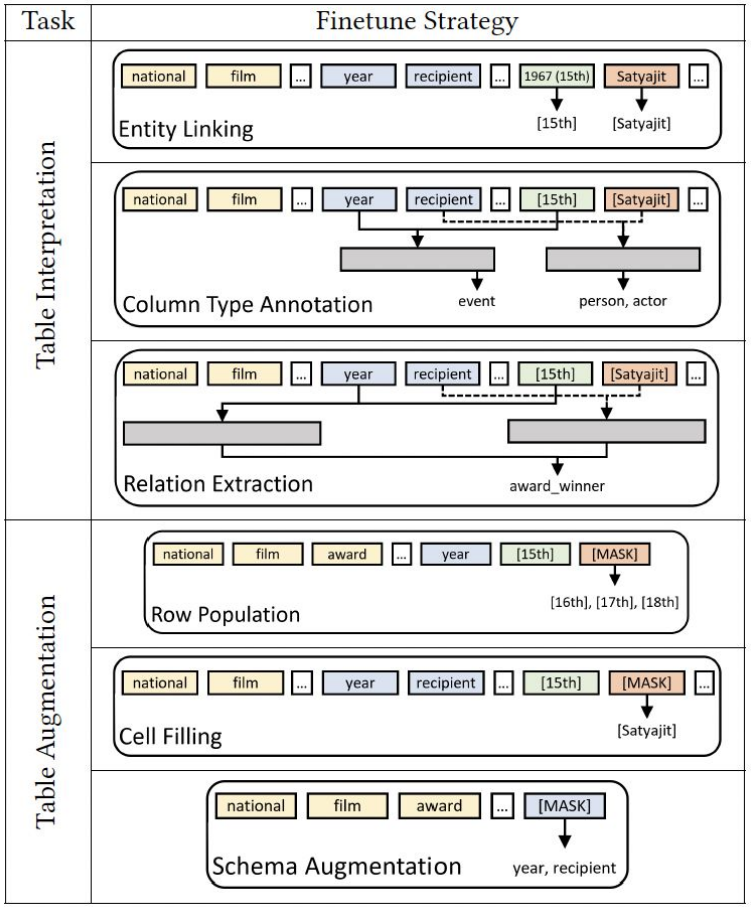
Transformer-based Architecture

Pre-training with masked self-attention: Each token/entity in a table can only attend to its directly connected neighbors

[Deng et al. VLDB 21]

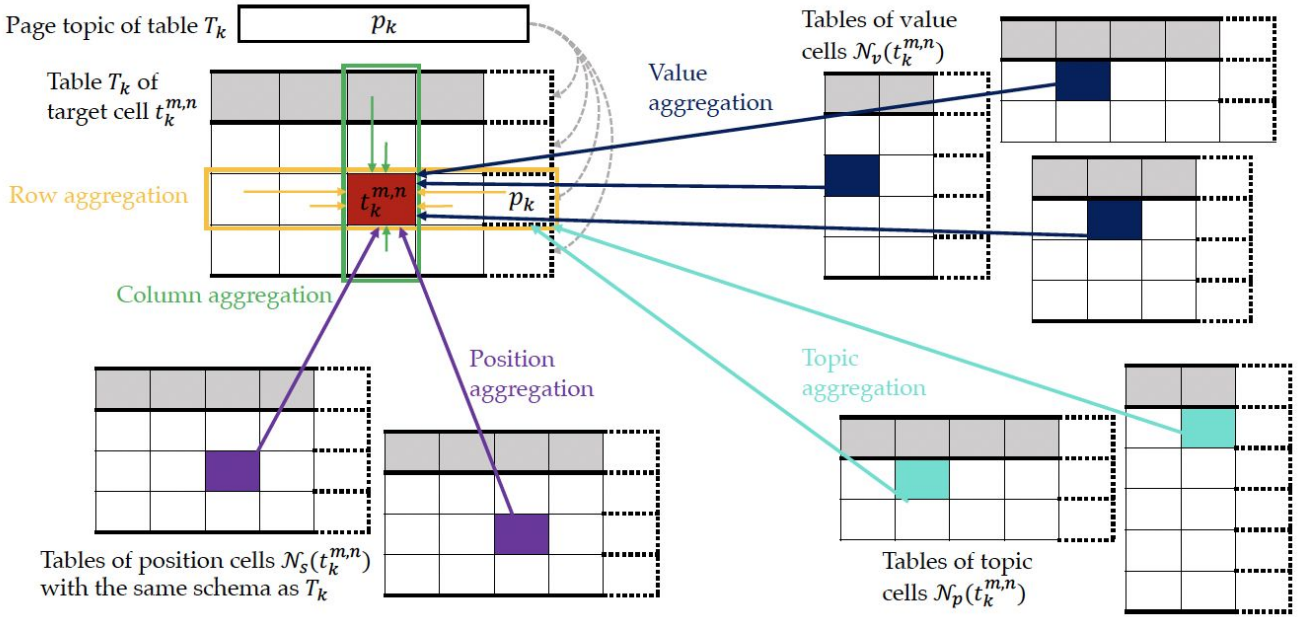
TURL: Fine-tuning

- Perform Fine-tuning with task-specific datasets
- Strategies for different categories of tasks
  - Table interpretation
  - Table augmentation



TCN: Table convolution network

[Wang et al. WWW 21]

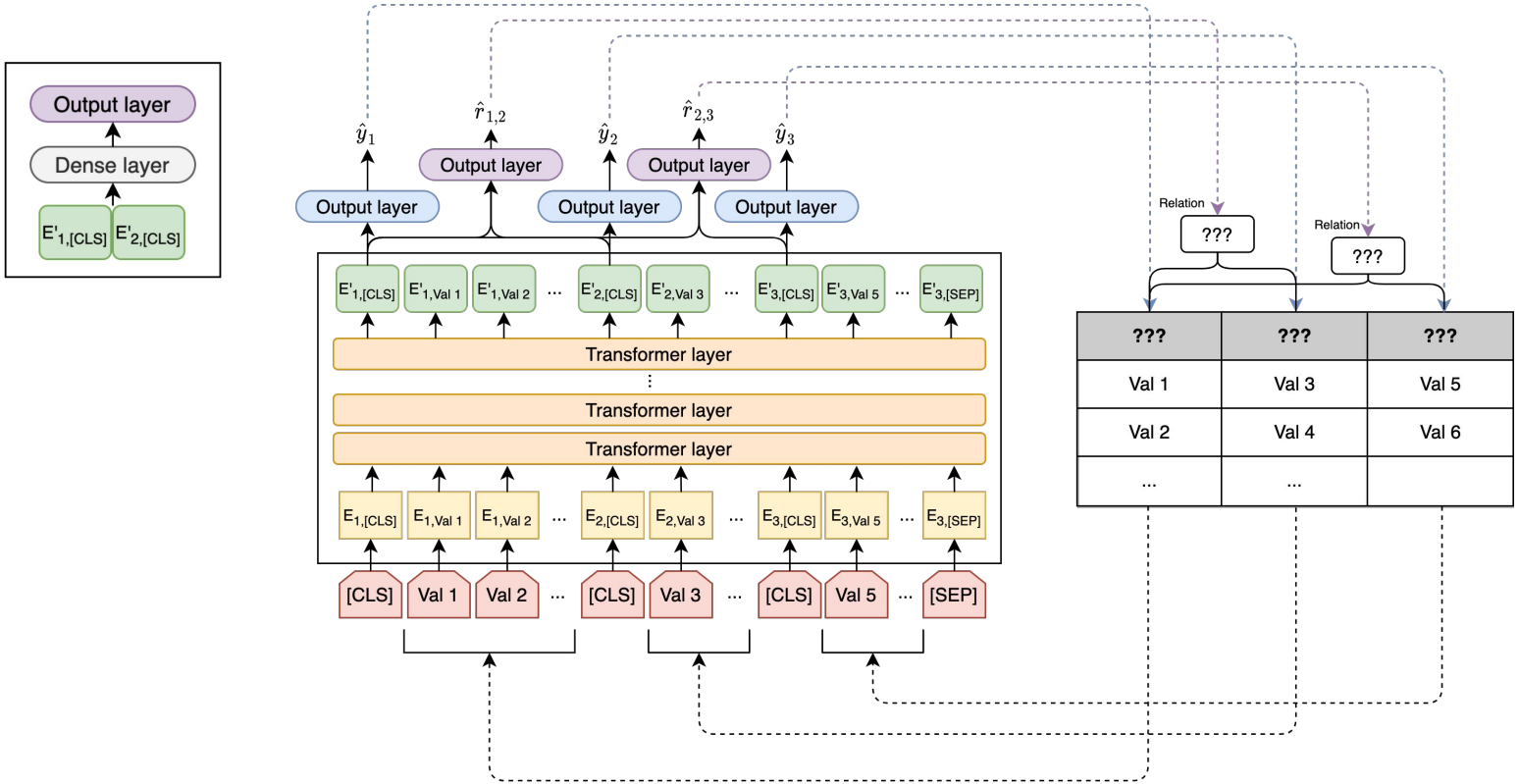


Learn Table representation by taking both intra- and inter-table contexts

Doduo: Multi-task fine-tuning

[Suhara et al. SIGMOD 22]

Support both column type detection and relation extraction with pre-trained LMs



Outline

- Introduction
- Table Understanding
- Table Search Engine
- Table Navigation and Exploration
- Data Science and Application Support
- Conclusion / Future work

Table Search Engine

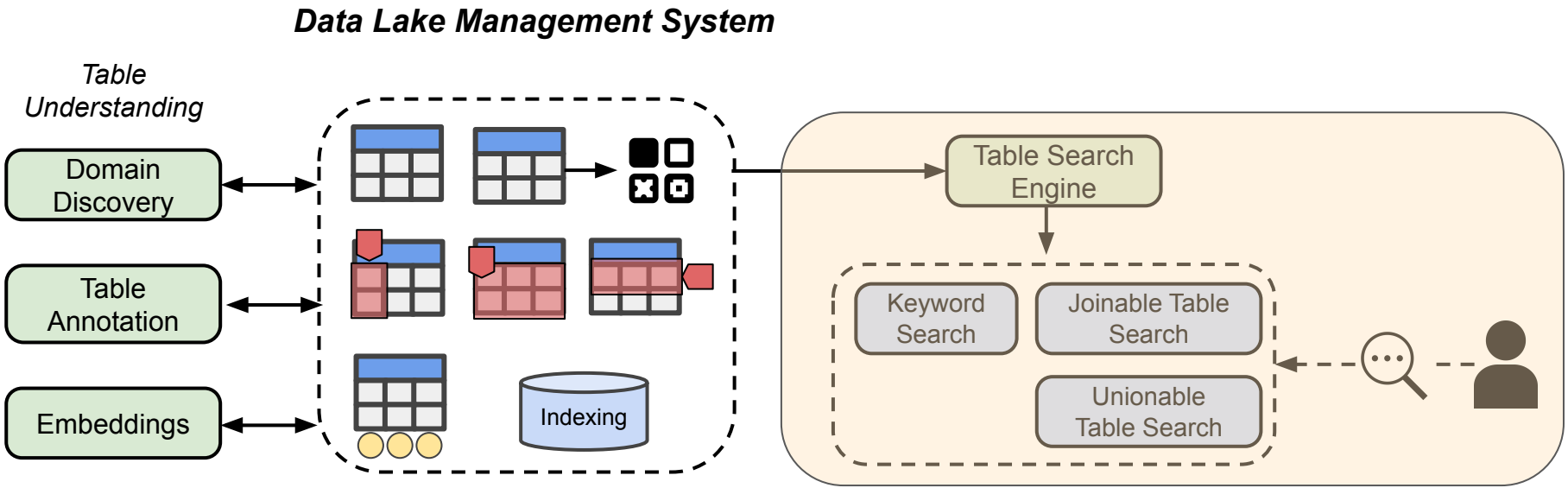
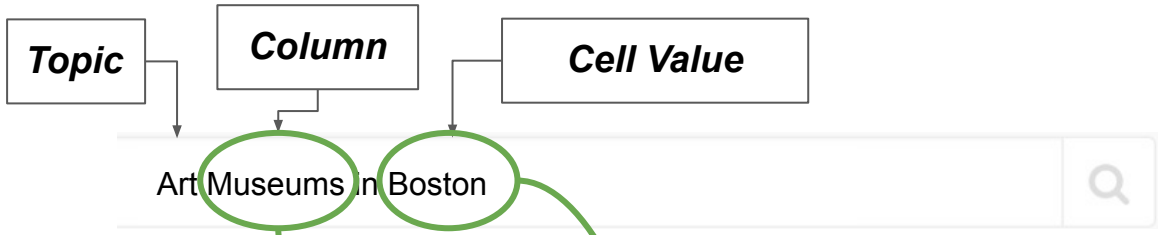


Table Discovery using Keyword Search



[Pimplikar, Sarawagi PVLDB 12]

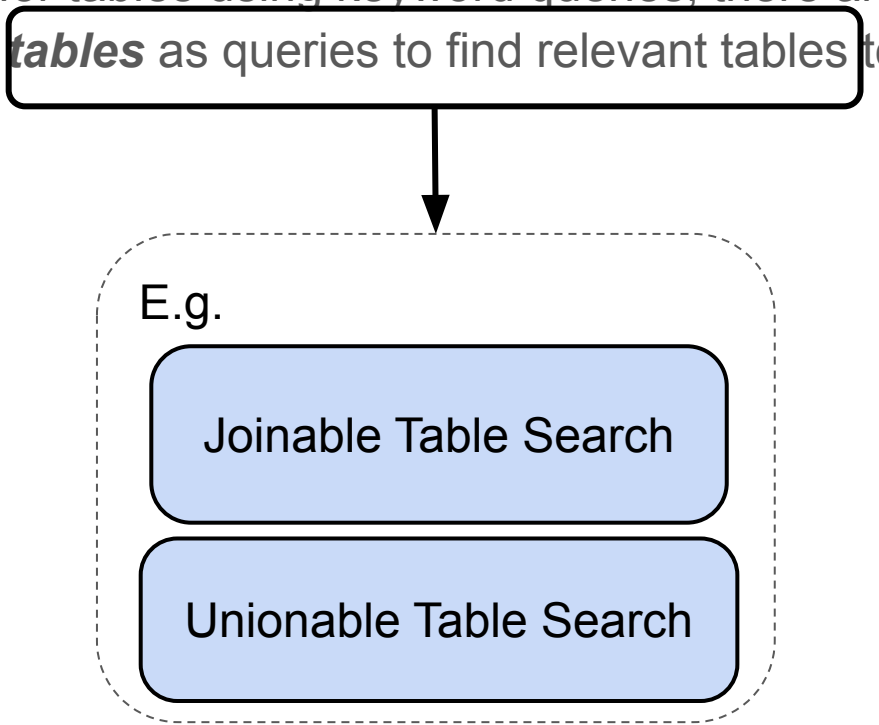
Octopus
[Cafarella et al. PVLDB 09]

Goods
[Brickley et al. WWW 19]

Museum	City	Zip Code
MFA	Boston	02115
MOMA	New York City	10019
Art Institute	Chicago	60603
…	…	…

Tables-as-Queries for Table Discovery

- In addition to searching for tables using keyword queries, there are systems that allow users to input ***tables*** as queries to find relevant tables to their input tables

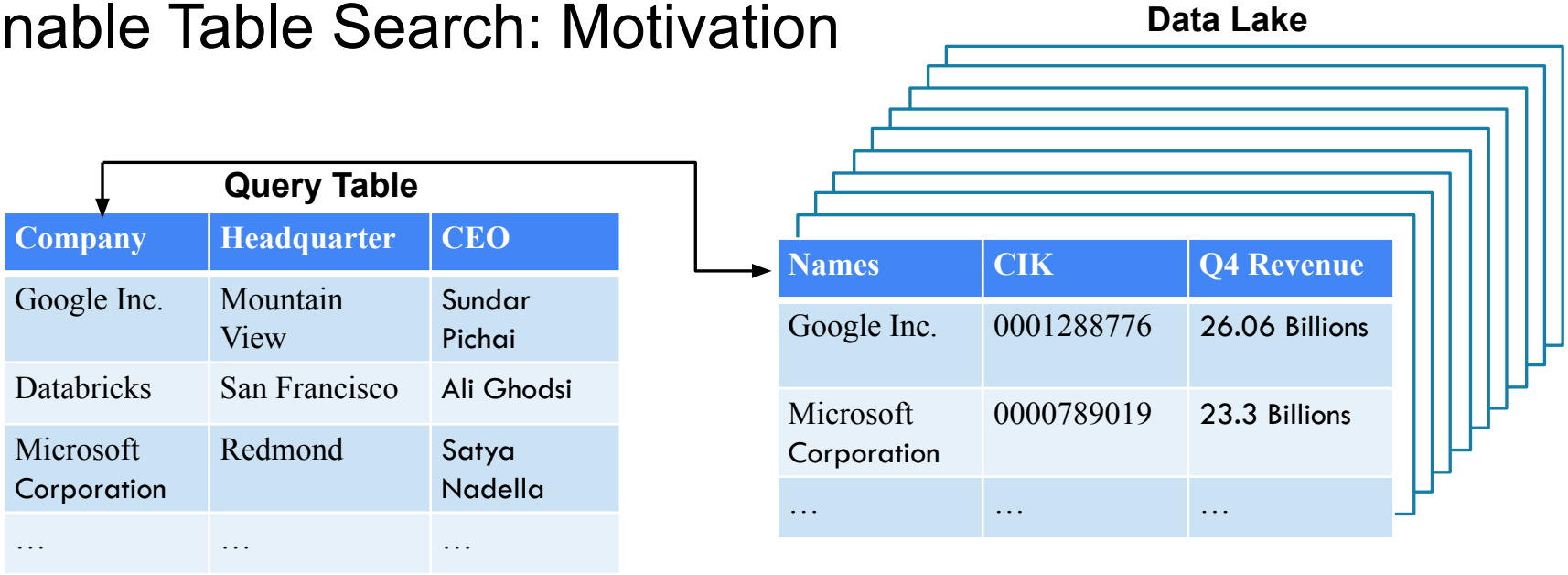


[Sarma et al. SIGMOD 12]

Finding Related Tables

- Given a table corpus and a query table, return a ranked list of relevant tables
- Schema Complements are found using schema matching and are returned as joinable tables with additional attributes
  - Joinable Table Search
- Entity Complements share a subject attribute and similar schemas
  - Unionable Table Search

Joinable Table Search: Motivation



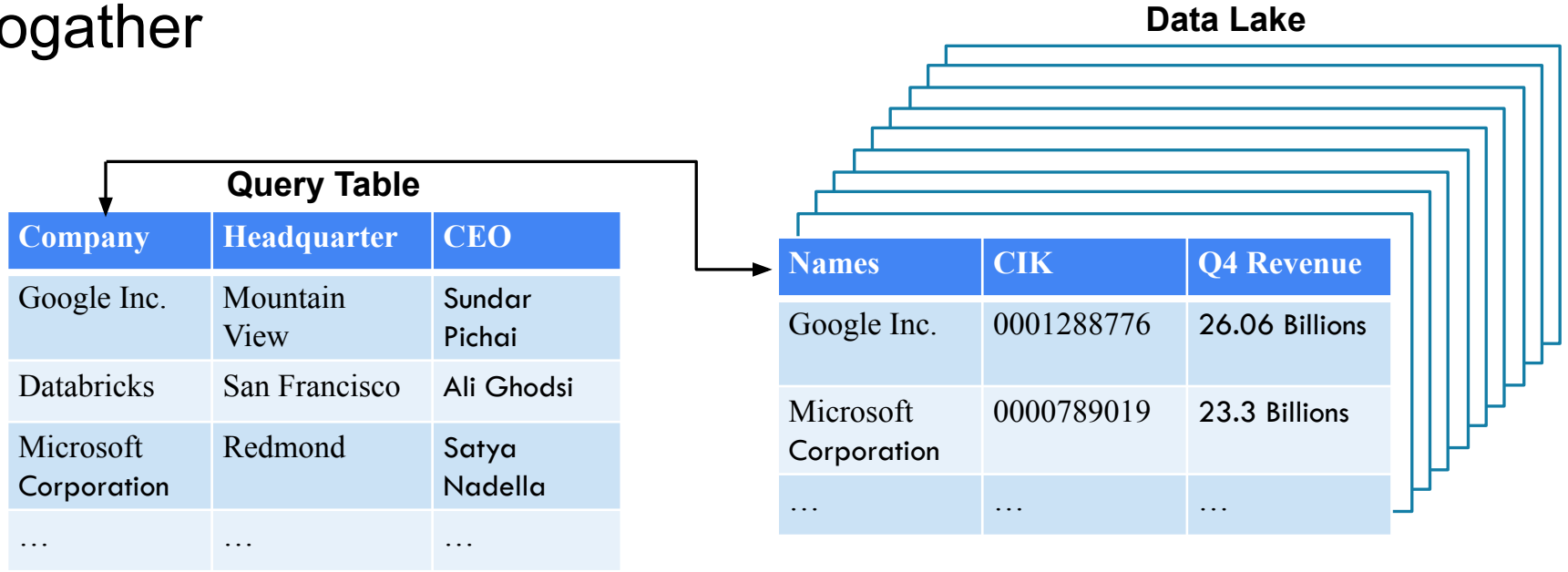
Data Science Question: **How can I find more features for my model of corporate health?**

***Data Management Task*: Find tables that can be joined with a query table.**

** Slide from JOSIE talk at SIGMOD19

[Yakout et al. SIGMOD 12]

Infogather



Uses schema matching to find attributes that join
Company matches Names

[Zhu et al. PVLDB 16]

Set Containment

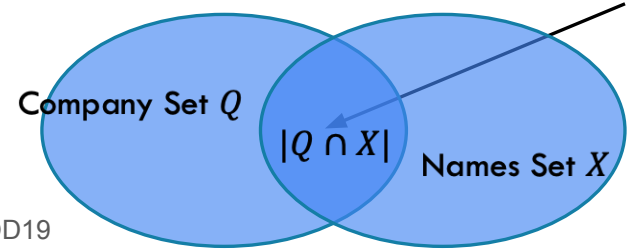
Query Table

Company	Headquarter	CEO
Google Inc.	Mountain View	Sundar Pichai
Databricks	San Francisco	Ali Ghodsi
Microsoft Corporation	Redmond	Satya Nadella
…	…	…

Data Lake

Names	CIK	Q4 Revenue
Google Inc.	0001288776	26.06 Billions
Microsoft Corporation	0000789019	23.3 Billions
…	…	…

Query Column



Number of distinct values in the query that are joinable – **Overlap**

** Slide from JOSIE talk at SIGMOD19

[Zhu et al. PVLDB 16]

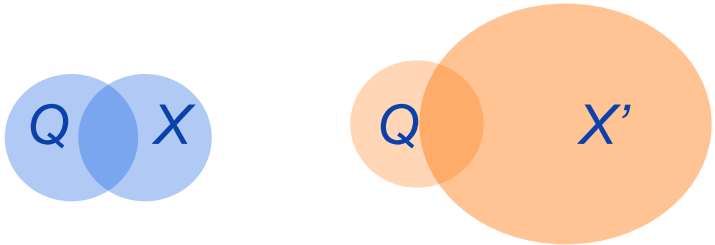
LSH Ensemble: Approximate Containment Join Search

$$Jaccard(Q,X) = \frac{|Q \cap X|}{|Q \cup X|}$$

Jaccard(Q,X) >> ***Jaccard(Q, X')***

Same intersection size,
but the Jaccard similarity is much smaller on the right
Jaccard is biased to small sets
[Fernandez et al. ICDE18], and others

$$Containment(Q,X) = \frac{|Q \cap X|}{|Q|}$$



Containment(Q,X) =***Containment(Q, X')***

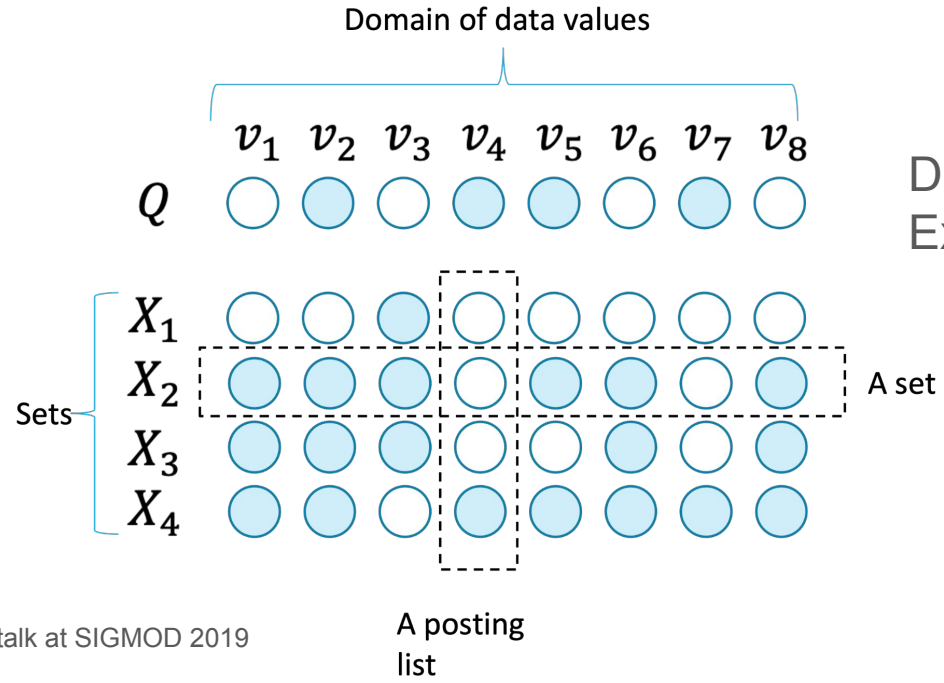
Containment is the same for both,
independent of the size of X and X'

LSH Ensemble: Uses MinHash LSH for approximate, scalable set containment search

[Zhu et al. SIGMOD 19]

JOSIE: Exact Containment Join Search

- Performs exact top-k joinable table search using set containment
- Scales inverted index to massive queries and massive number of posting lists



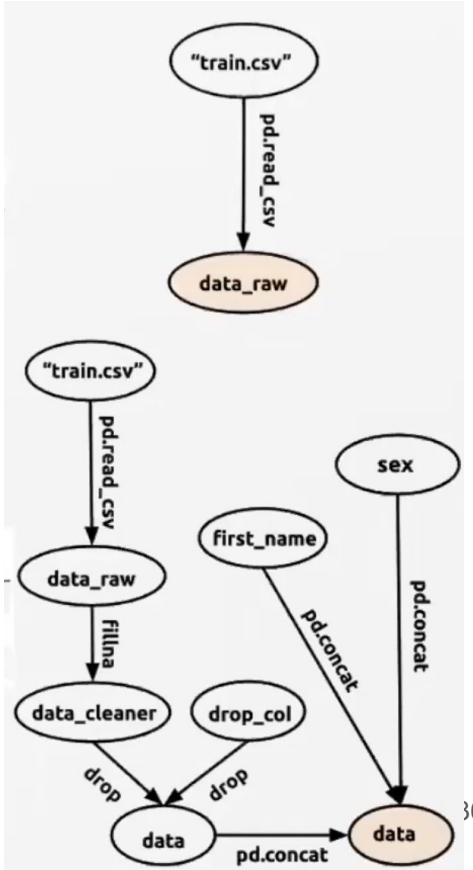
Data-driven cost model
Experiments on data lakes with
Posting lists: .5 Billion
Queries: upto 10K in size

** Figure from JOSIE talk at SIGMOD 2019

[Zhang et al. SIGMOD 20]

Juneau: Joinable Search in Jupyter Notebooks

- Joinable Table Search in Jupyter Notebooks
- Create data profiles using data values and domains

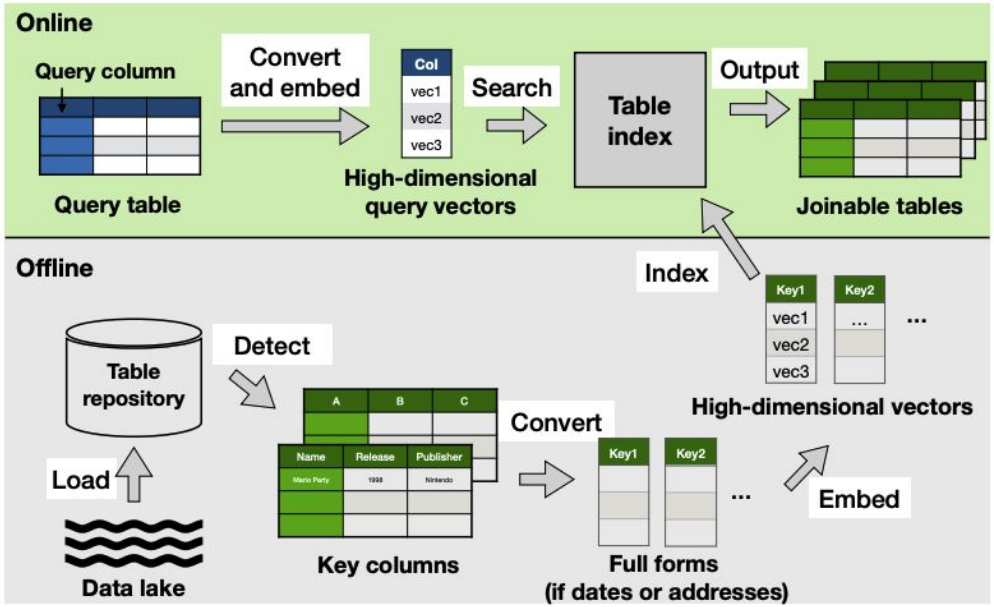


** Figure from Juneau talk at SIGMOD 2020

[Dong et al. ICDE 21]

PEXESO: Fuzzy-Join Search

Use word embeddings to retrieve tables that can be fuzzy-joined



** Figure from paper: Efficient Joinable Table Discovery in Data Lakes: A High-Dimensional Similarity-Based Approach

[Santos et al. ICDE 22]

A Sketch-based Index for Correlated Dataset Search

- Find top-*k* joinable tables that also contain numerical columns correlated with numerical column(s) in the Query Table
- Employs a QCR hashing scheme to estimate correlations between numerical values

WarpGate

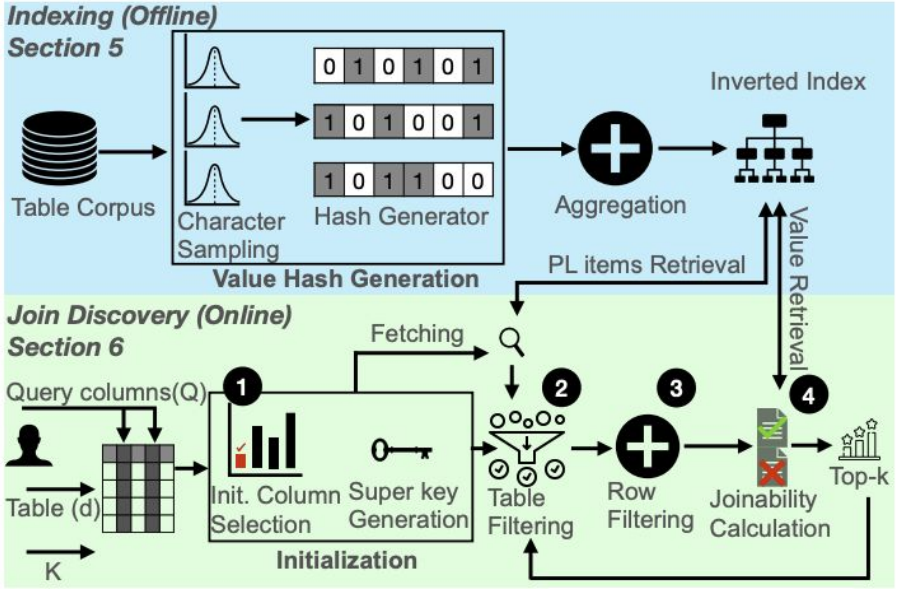
[Cong et al. CIDR 23]

- Retrieves top-*k* semantically joinable tables
- Use embeddings to capture semantic relationships

[Esmailoghli et al. PVLDB 22]

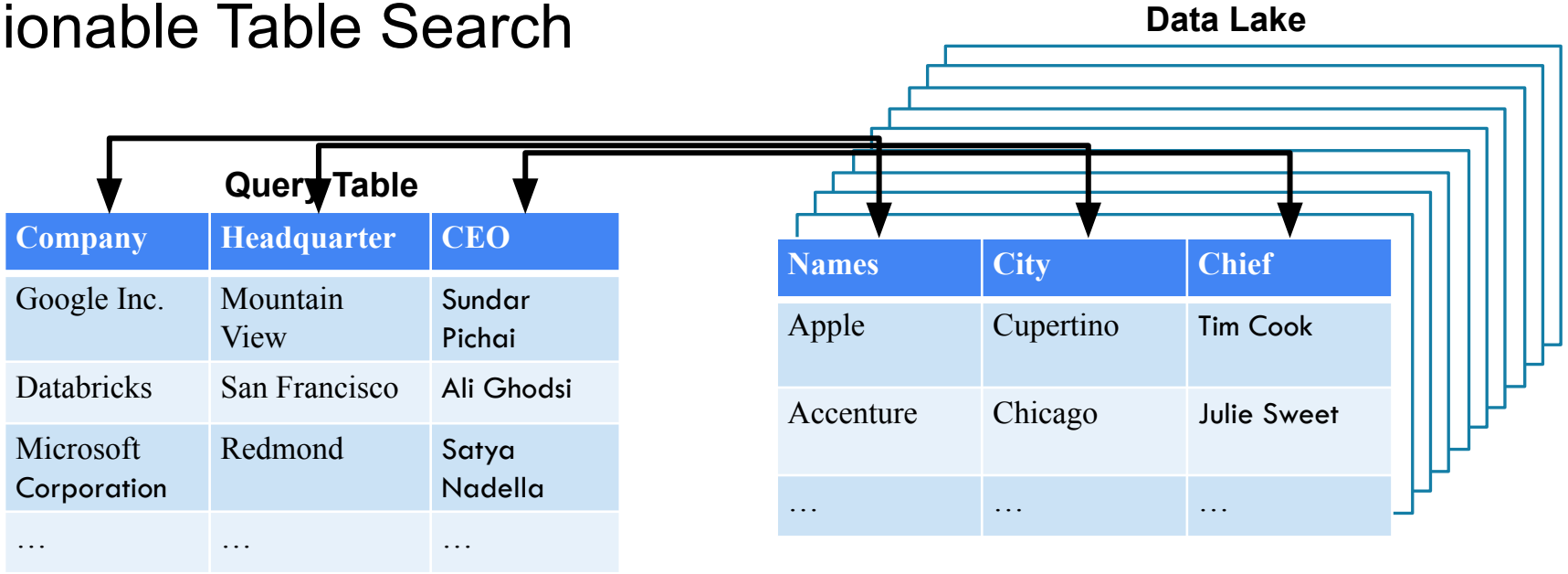
MATE: Multi-Attribute Join

- Supports joins over multiple attributes
- Hash each row value into a super key



** Figure from paper: MATE: Multi-Attribute Table Extraction

Unionable Table Search



Data Science Question: **Does my analysis generalize? To new Companies or new regions?**

Data Management Task**: Find tables that can be unioned with a query table.**

Table Union Search

[Nargesian et al. VLDB 18]

- Given a set of tables and a Query Table, return the top-*k* tables with the highest table unionability
  - Some **attributes** may overlap
  - Some may refer to **entities** of a common type
  - Some may use **semantically similar words**

Electricity	Barnett	Domestic	240.99	…
Gas	Brent	Transport	164.44	
Coal	Camden	Transport	134.90	
Railways diesel	City of London	Domestic	10.52	
Gas	Brent	Domestic	169.69	
Coal	Brent	Transport	120.01	

Query Table

Benton	Transport	Gasoline	64413	62.9
Kittitas	Hydro	Fuel oil (1,2,…	12838	66.0
Grays Harbor	Domestic	Aviation fuels	1170393	66.1
Skagit	Transport	Liquified petroleum	59516	60.1

Candidate Table

[Bogatu et al. ICDE 20]

- D3L: Extends TUS to also consider formatting similarity and schema similarity

[Khatiwada, Fan et al. SIGMOD 23]

SANTOS

Data Scientist's table (Query Table)

Park Name	Supervisor	City	Country
River Park	Vera Onate	Fresno	USA
West Lawn Park	Paul Veliotis	Chicago	USA
-------	------	-------	--------

(a) A table about parks

Data lake tables (Candidate unionable Tables)

Park Name	Film Title	Park Location	Park Phone	Park City	Film Director	Film Studio
Chippewa Park	Bee Movie	6748 N. Sacramento Ave.	773 731-0380	Cook	Simon J. Smith	Dreamworks
Lawler Park	Coco	5210 W. 64th St.	773 284-7328	Riverside	Adrian Molina	Pixar
-----	------	-------	------	----	------	------

(b) A table about films shown in different parks

Person	Occupation	Birthplace	Country	Park Name
James Taylor	Singer	Boston	USA	Central Park
Anthony Pelissier	Film Director	Barnet	UK	Cairngorms National Park
Akram Afif	Football Player	Doha	Qatar	Aspire Park
Ivan A. Getting	Physicist	NYC	USA	El Segundo Park
Abby May	Social Worker	Boston	USA	Fenway Park
Stevie Ray Vaughan	Singer	Texas	USA	Chastain Park

(c) A table about people

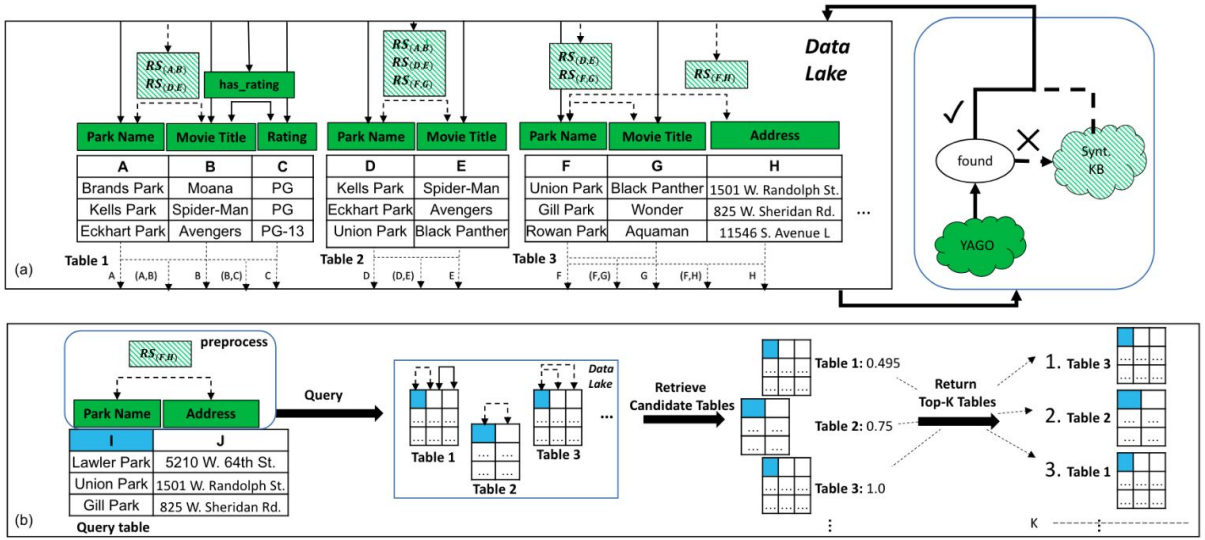
Along with column semantics,

consider the **binary relationships** between the column pairs.

[Khatiwada, Fan et al. SIGMOD 23]

SANTOS

- Semantic Table Union Search: considers both attribute semantics *and binary Relationship semantics between attributes*
  - ➢ Existing Knowledge Base
  - ➢ KB has limited coverage → construct a data-driven synthesized Knowledge Base



[Fan et al. VLDB 23]

Starmie

Table A:

Name	Mode of Travel	Purpose	Destination	Day	Month	Year	Expense
Philip Duffy	Air	Regional Meeting	London	10	April	2019	189.06
Jeremy Oppenheim	Taxi	Exchange Visit	Ottawa	30	Jul	2019	8.08
Mark Sedwill	Air	Evening Meal	Bristol	02	September	2019	50

Table B:

Name	Date	Destination	Purpose
Clark	23/07	France	Discuss EU
Gyimah	03/09	Belgium	Build Relations
Harrington	05/08	China	Discuss Productivity

Table C:

Bird Name	Scientific Name	Date	Location
Pine Siskin	Carduelis Pinus	2019	Ottawa
American Robin	Turdus migratorius	2019	Ottawa
Northern Flicker	Colaptes auratus	2019	London

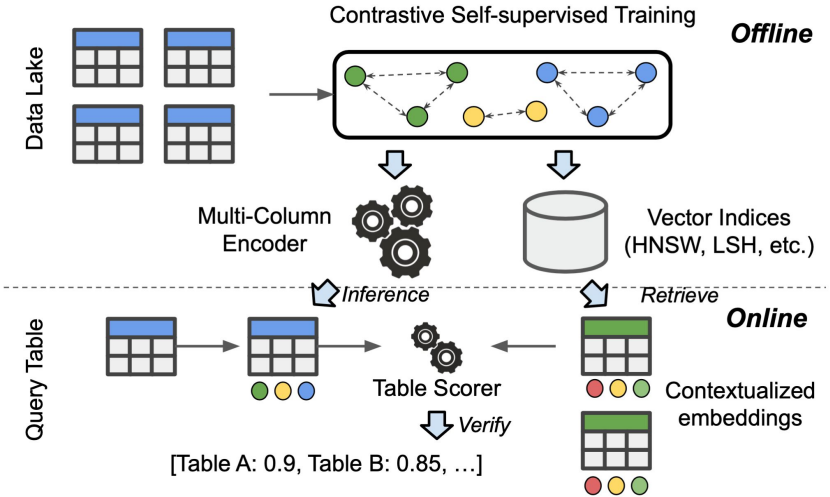
- Only Column Unionability → wrongly aligned columns
- Also Using Binary relationship between columns → influenced by value overlap

Use the context from the entire table to capture the semantics

[Fan et al. VLDB 23]

Starmie

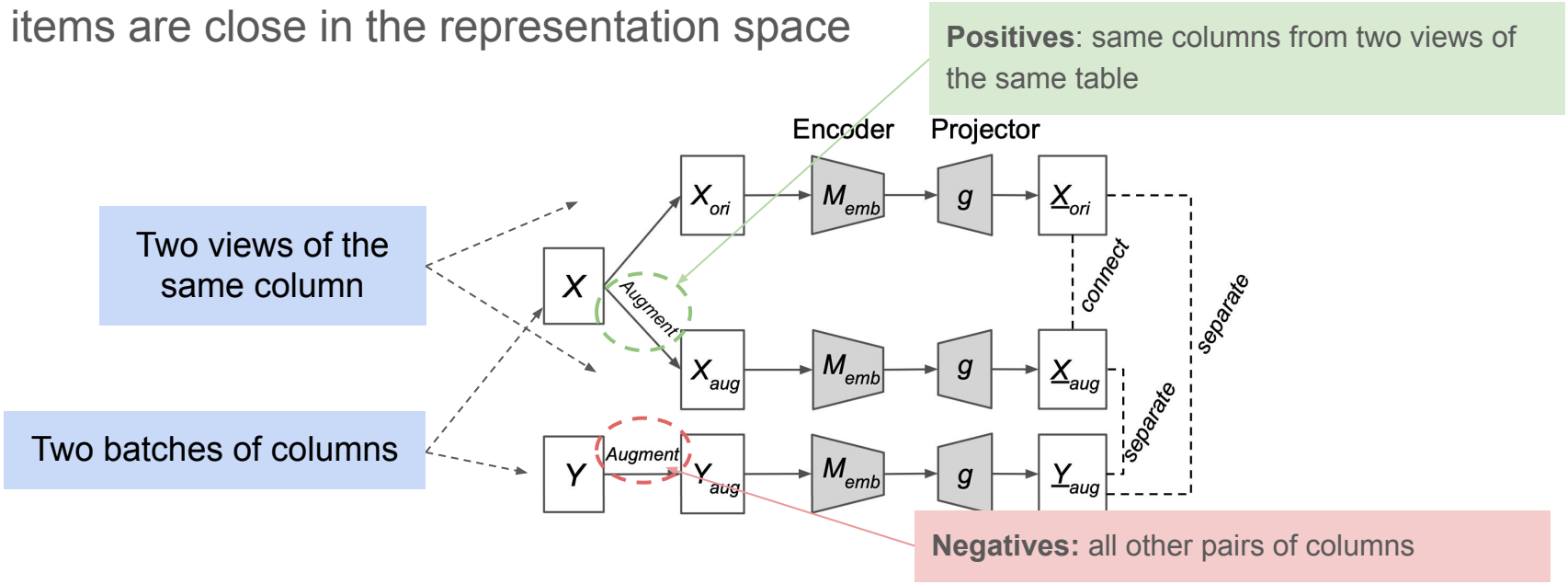
- Semantic Table Union Search: Extend column semantics and binary relationship semantics to *use the entire table context*
- Natural language approach: Language Model uses the table context to capture and encode the semantics of each column



[Fan et al. VLDB 23]

Starmie

- Difficult to manually label training data → Self-Supervised Technique
- Contrastive Learning: Learn representations (w/o labels) such that similar items are close in the representation space

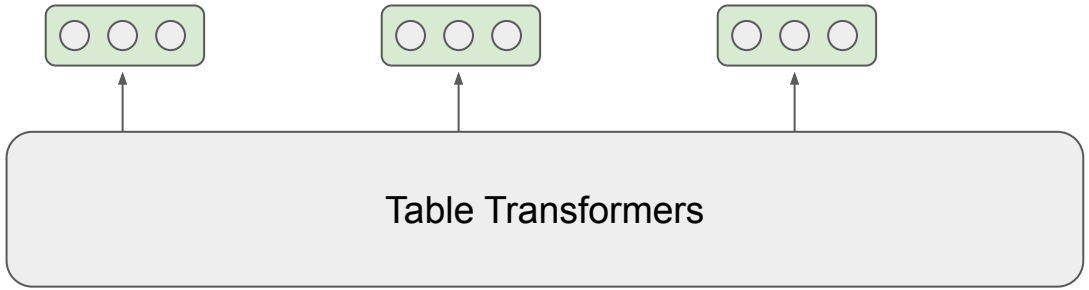


[Fan et al. VLDB 23]

Starmie

- Multi-column table models: Preserve the **context** of the column during encoding

Contextualized Column Embeddings



Serialize and concatenate

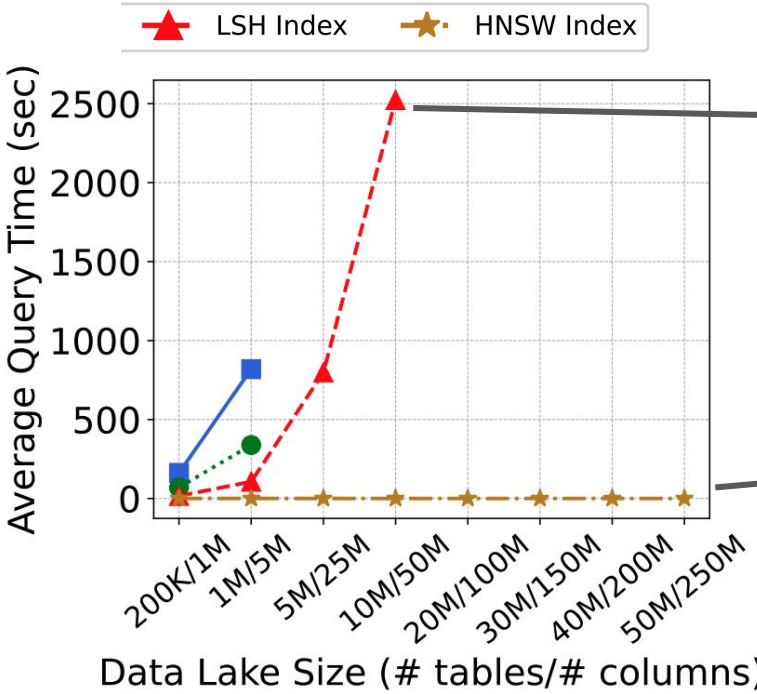
<s> New York California <s> Albany Sacramento <s> 1797 1851

State	Capital	Since
New York	Albany	1797
California	Sacramento	1854

[Fan et al. VLDB 23]
[Malkov et al. TPAMI 20]

Starmie

- Hierarchical Navigable Small Worlds (**HNSW**) index: graph-based indexing
- LSH Index: the state-of-the-art hash-based indexing



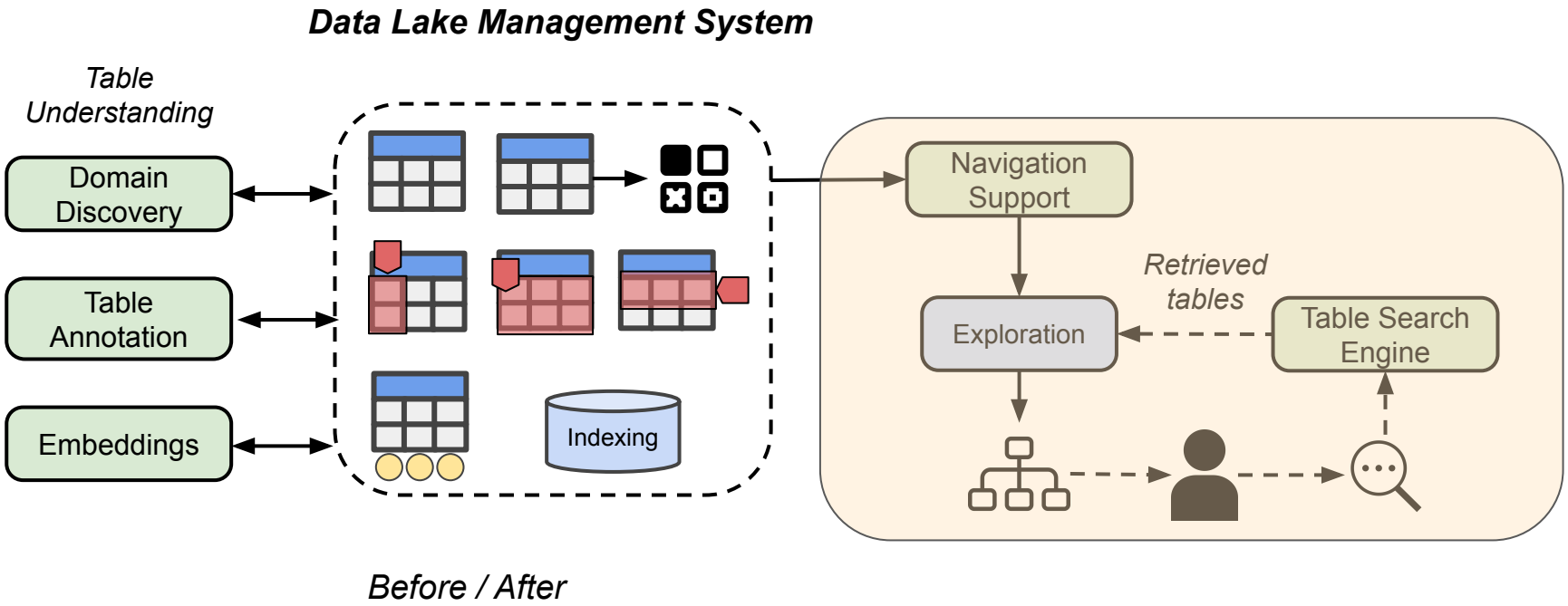
LSH times out after the data lake size exceeds 10M tables

HNSW index has a query time of 60 ms, even with a data lake of 50M tables

Outline

- Introduction
- Table Understanding
- Table Search Engine
- Table Navigation and Exploration
- Data Science and Application Support
- Conclusion / Future work

Table Navigation and Exploration

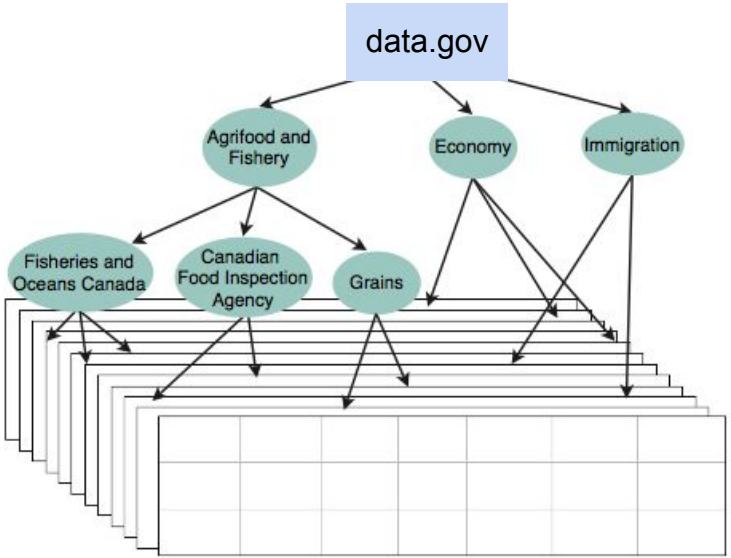


Complementary to search: present the data lake in a navigation-friendly structure to the user

Table Navigation / Exploration

[Nargesian et al. SIGMOD 20]

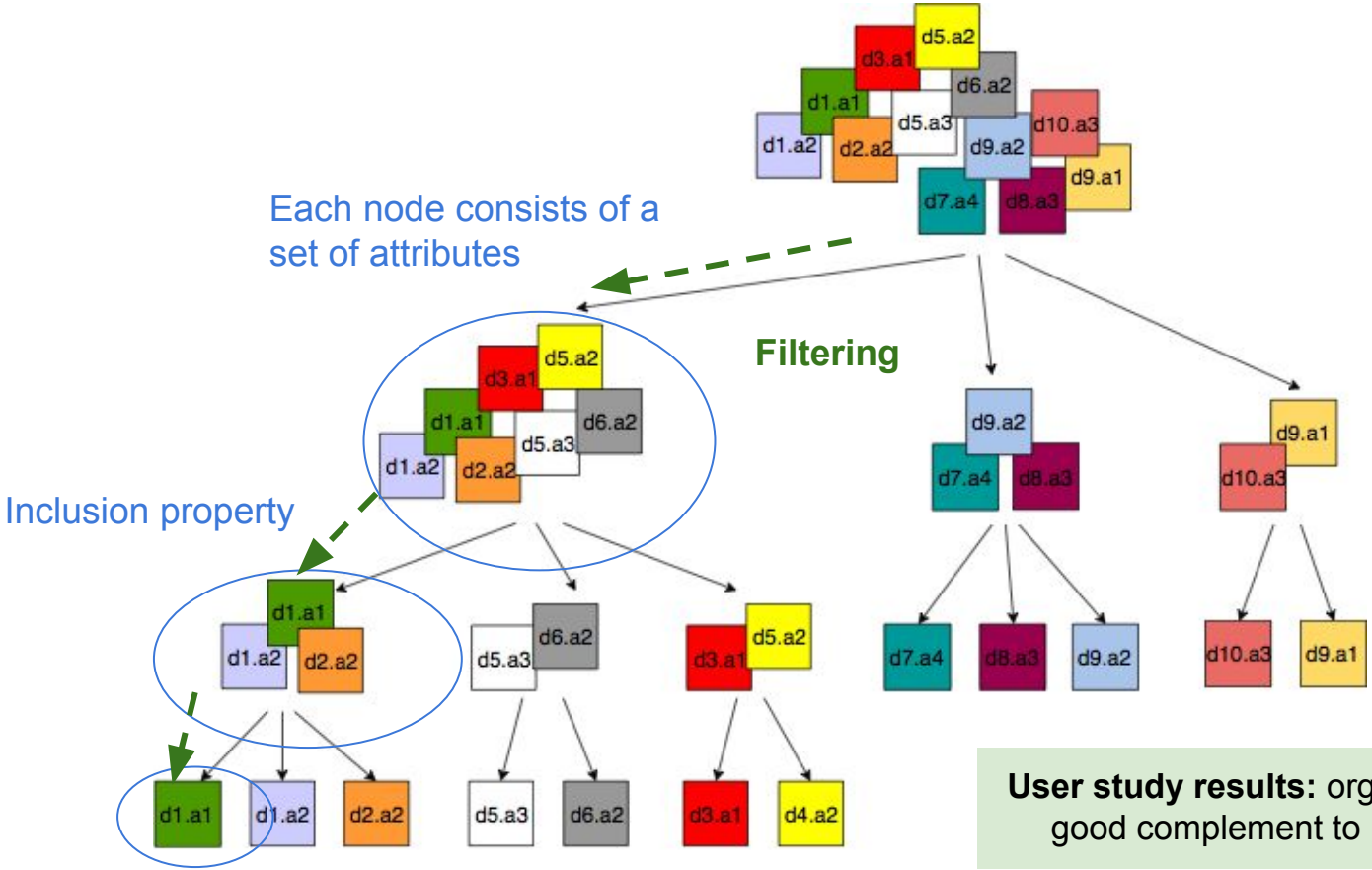
Problem Definition:
- Automatically building a directory structure that enables a user to most efficiently find tables of interest in a data lake.



Example properties to be optimized:
- reasonable number of choices at each step.
- reasonable number of steps to find a dataset.
- unambiguous choices at each step.

Data Lake Organization: a DAG

[Nargesian et al. SIGMOD 20]

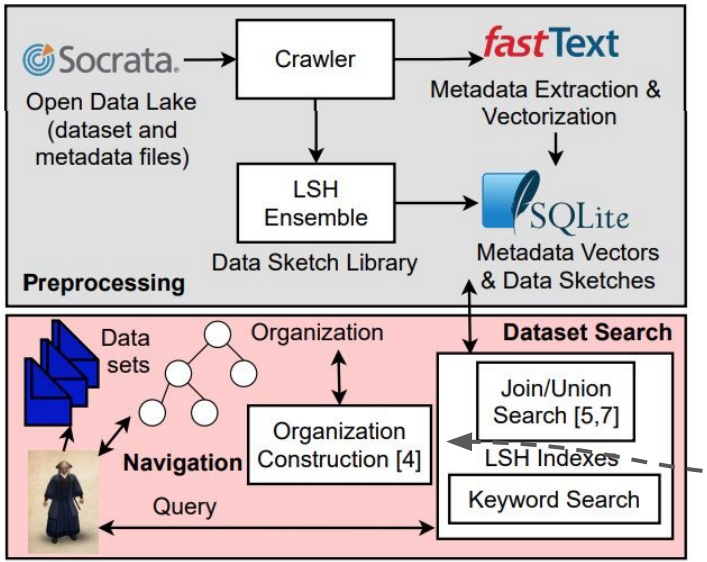


User study results: organization can be a good complement to keyword search

RONIN: Data Lake Exploration

[Ouellette et al. VLDB 21]

- A follow-up demo system that integrates navigation and search over data lakes
- Support different kinds of search: keyword, unionable and joinable



Goal: Enables the users to seamlessly switch between two operations

The navigation organization is constructed on-the-fly as the search result is generated

An alternative organization: Enterprise KG

Aurum: A Data Discover System **[Fernandez et al. ICDE 18a]**

- **EKG:** a hypergraph that captures relationships between columns of datasets
- Discussed **building**, **maintaining**, and **querying** of EKG

Algorithms to decide when to re-index a column

a graph query language SRQL

[Fernandez et al. ICDE 18b]

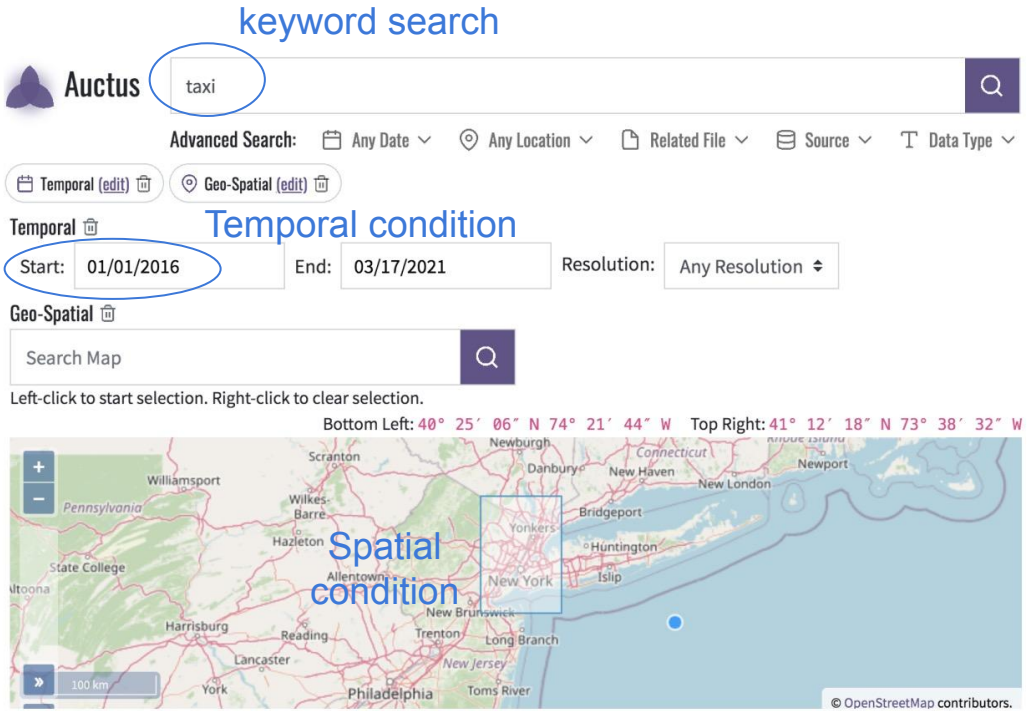
Seeping Semantics: Linking Datasets using Embeddings

- Studied how to use aggregated embeddings (e.g., GloVe) to identify column/dataset relationships

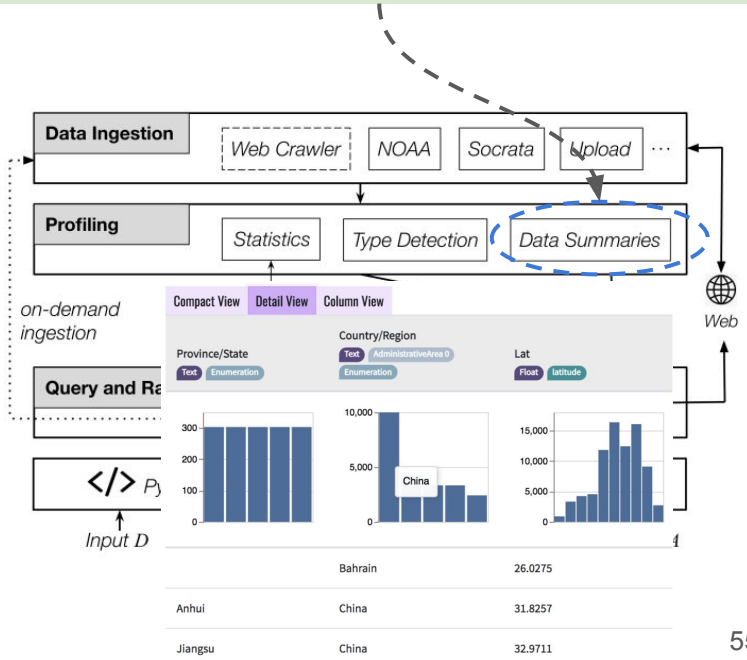
We have better tools for this problem today

Dataset Search Engine: Auctus

[Castelo et al. VLDB 21]



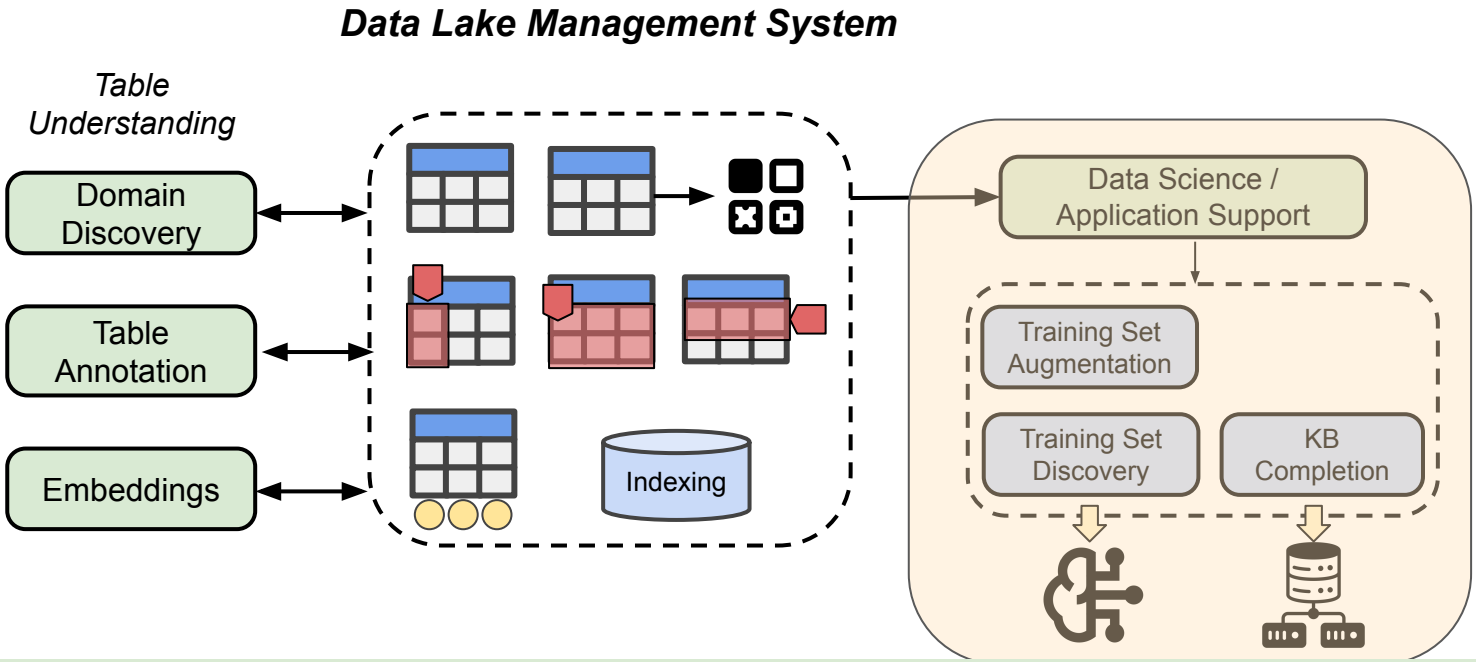
Compact dataset summaries allow users to quickly navigate through the search results



Outline

- Introduction
- Table Understanding
- Table Search Engine
- Table Navigation and Exploration
- Data Science and Application Support
- Conclusion / Future work

Table Discovery for Data Science / Application Support



Data/Table discovery is the key to success for many **data-hungry applications,** especially ML

Table Discovery for Supporting Data Science Applications

Training Set Augmentation

ML: augment a target table (for regression/classification) with features learned or extracted from data lake tables

Training Set Discovery

ML: create novel regression/classification tasks using joins/unions of data lake tables

KB Completion

KB: combine webtables to form structured records for KB creation and/or completion

[Fan et al. VLDB23]

Case study from Starmie: Data Discovery for ML tasks

- **Goal:** find joinable tables from WDC to improve downstream ML tasks
- **Query tables / ML tasks:** 25 tables of >=200 rows, target column = “Rating”
- **Data lake tables:** 4,130 VizNet tables of >=50 rows

State	Office	District	Name	Party	Rating
AZ	U.S. House	8	Trent Franks	Republican	95.0
TX	U.S. House	3	Sam Johnson	Republican	95.0
OH	U.S. House	4	Jim Jordan	Republican	95.0
CO	U.S. House	5	Doug Lamborn	Republican	95.0

Goal: find useful features for predicting the “Rating”

[Fan et al. VLDB23]

Case study from Starmie: Data Discovery for ML tasks

- **Baselines:** NoJoin, Jaccard, Overlap
- **Starmie**: contextualized embeddings for join+target columns

	NoJoin	Jaccard	Overlap	Starmie
Avg. MSE	0.0820	0.0753	0.0748	**0.0699**
Improvement	-	8.23%	8.82%	**14.75%**
#improved	-	13 / 25	12 / 25	**15 / 25**
avg. Improve	-	14.74%	14.05%	**20.64%**

Findings: learned table representations help find more meaningful joinable tables for improving downstream ML tasks

Example output (Starmie)

Joining with a table about money raised on the “name” column

State	Office	District	Name	Party	Rating	Party_r	State_r	$ From Interest Groups That Supported	$ From Interest Groups That Opposed	Vote
AZ	U.S. House	8	Trent Franks	Republican	95.0	R	AZ-2	$12,000	$0	Yes
TX	U.S. House	3	Sam Johnson	Republican	95.0	R	TX-3	$4,000	$0	Yes
OH	U.S. House	4	Jim Jordan	Republican	95.0	R	OH-4	$22,000	$0	Yes
CO	U.S. House	5	Doug Lamborn	Republican	95.0	R	CO-5	$2,000	$0	Yes

Model performance (MSE) NoJoin: **0.1598** → Jaccard: **0.1544** → CL: **0.1195**

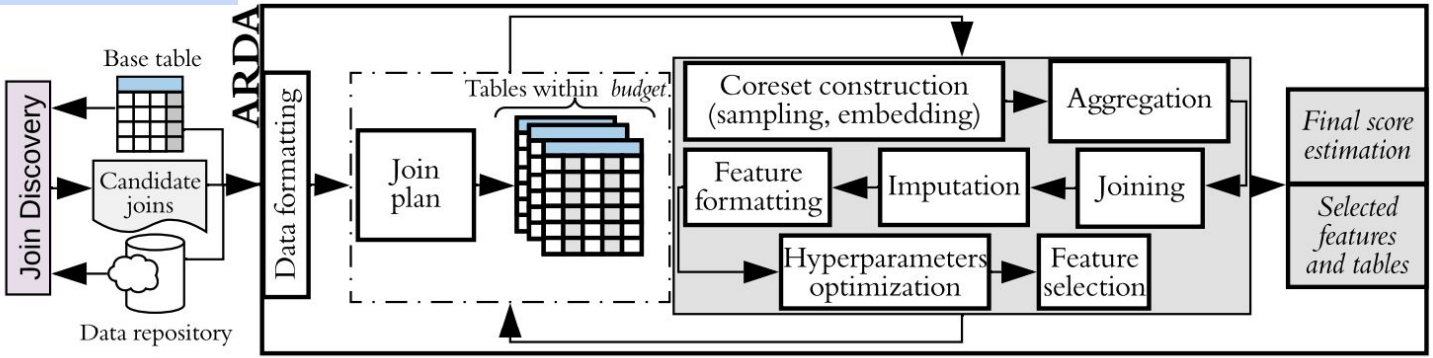
[Chepurko et al. VLDB 21]

Training set augmentation: ARDA

- **Goal:** input a dataset and a data repository, and outputs an augmented dataset => improved predictive performance.

Join Discovery (such as NYU Auctus)

Feature Selection

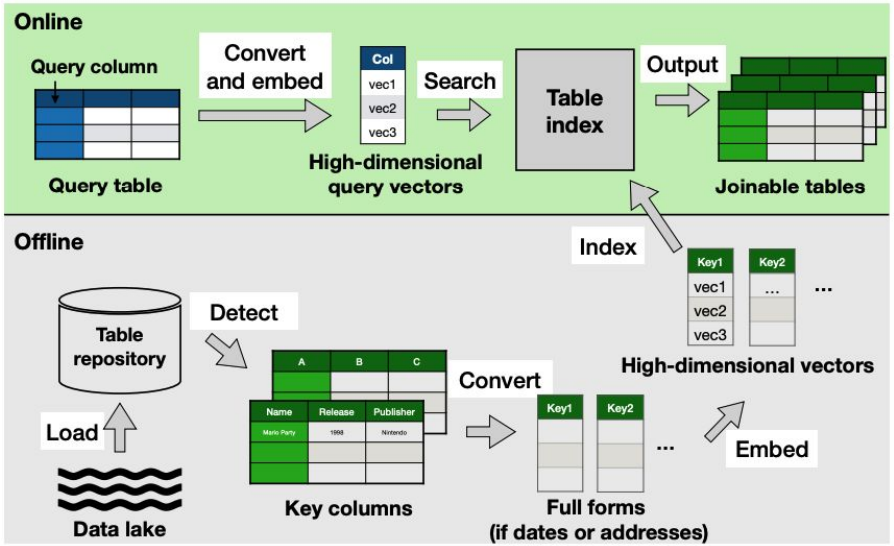


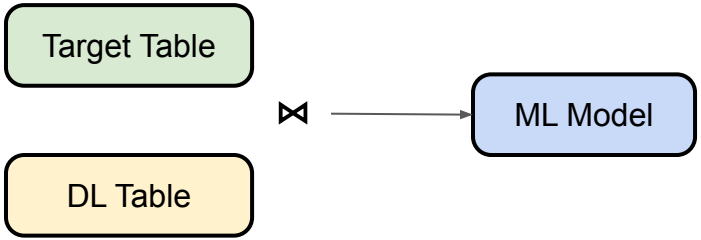
Consistently improved regression models (e.g., RF) across 5 real-world tasks from DARPA D3M

[Dong et al. ICDE 21]

Recap: Augment by finding joinable tables - PEXESO

- High-dimensional similarity search based on averaging GloVe embeddings



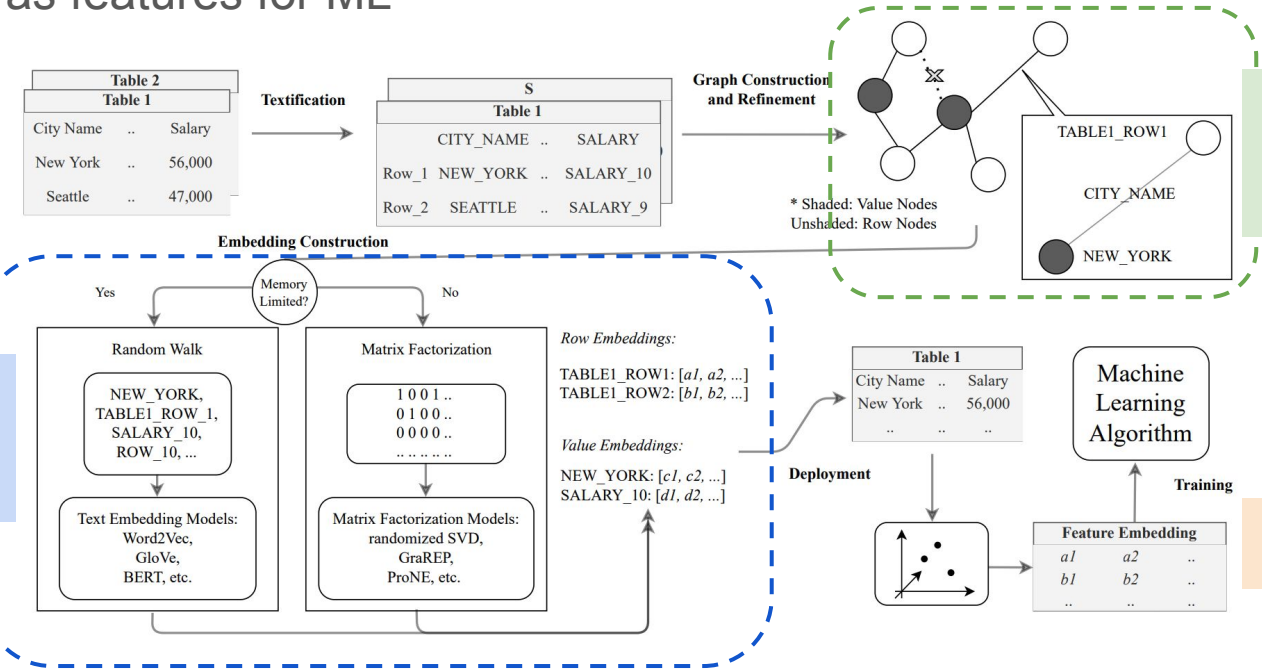


Improve ML models on 3 Kaggle tasks by joining the target table with tables from the WDC corpus

[Zhao et al. SIGMOD 22]

Learning relational embeddings: Leva

- Instead of joining tables, use vector representations learned from the data corpus as features for ML



Represent the table collection as a graph of rows/cells

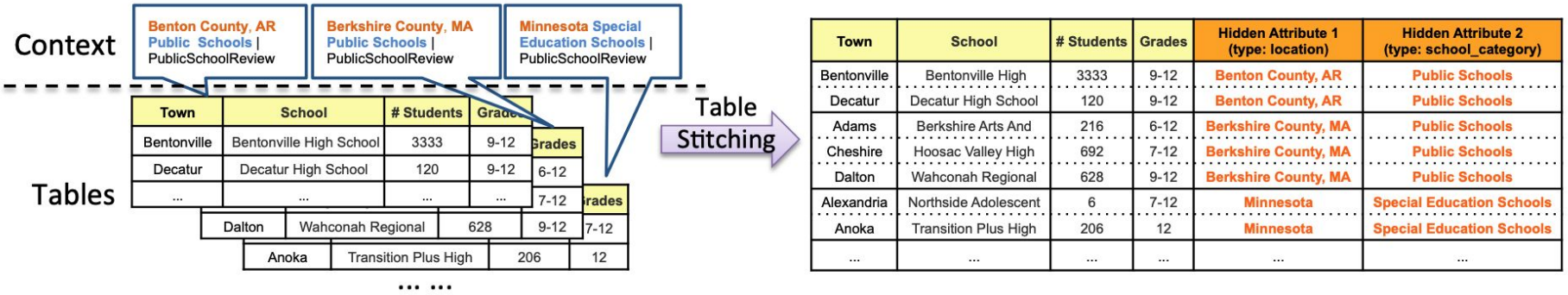
Learn dense vector representations via graph embeddings

Train models from embedding vectors

[Ling et al. IJCAI 13]
[Lehmberg and Bizer, VLDB 17]

KB Completion: Table Stitching

- **Goal:** discover unionable tables (with context) to match/augment with a KB

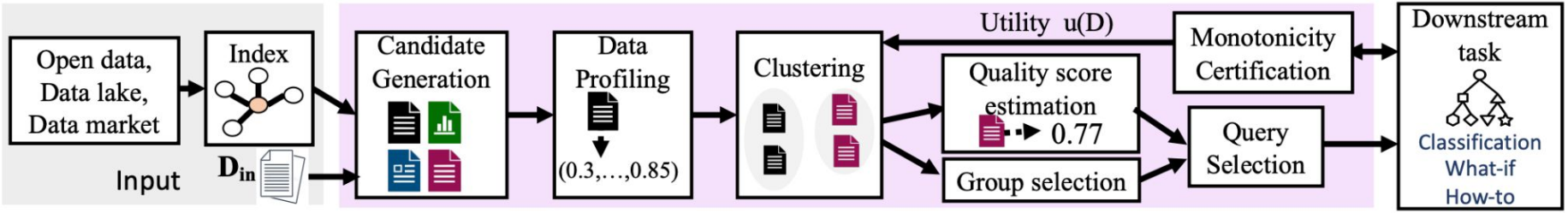


- **Table stitching [Ling et al. IJCAI 13]**: identify webtables of the same schema from the same sites + synthesizing new columns from page context
- **[Lehmberg and Bizer, VLDB 17]**: experimental study on stitching tables + KB alignment

[Galhotra et al. ICDE 23]

METAM: Goal-Oriented Data Discovery

- **Goal:** build a generic framework for querying the downstream task with a candidate dataset to automatically steer the discovery and augmentation process



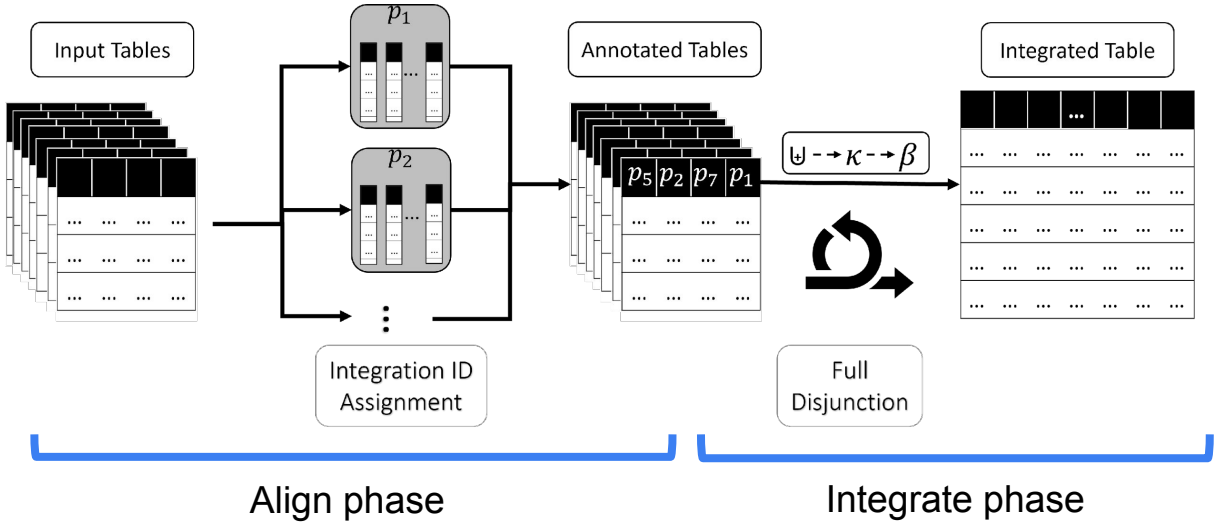
Profile the datasets using (1) **data stats**, (2) **utility function**, and (3) **solution set size**

Adapt to a range of downstream tasks: classification, regression, causal inference

[Khatiwada et al. VLDB 23]

Integrating Data Lake Tables

- **ALITE** (Align and Integrate)



- **Align:** Identify the matching columns across the set of tables and annotate them with a dummy column header.
- **Integrate:** Apply a novel algorithm for Full Disjunction that scales better than prior work. [Galindo-Legaria SIGMOD 94]

Outline

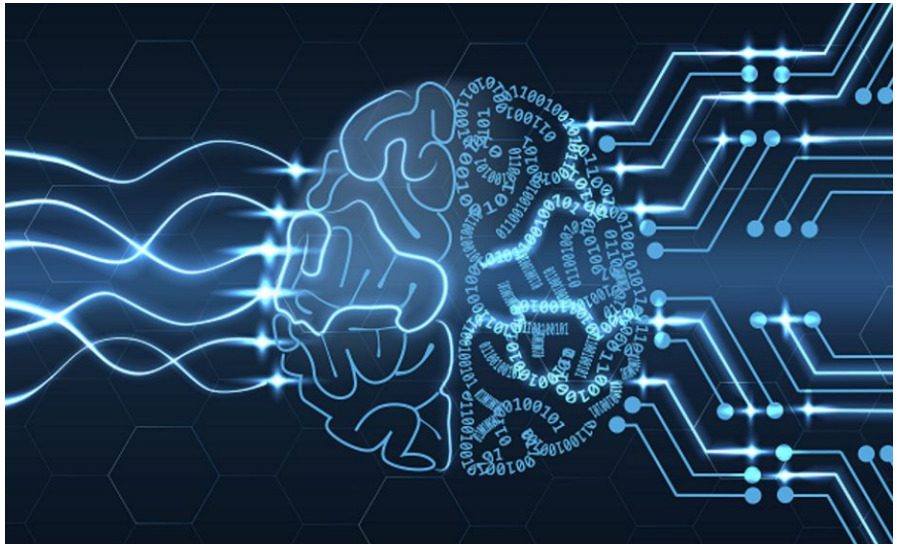
- Introduction
- Table Understanding
- Table Search Engine
- Table Navigation and Exploration
- Data Science and Application Support
- Conclusion / Future work

[Weikum PVLDB21]

Theory of how to use LM and KB Symbiotically

LLM & **KB** invaluable tools for Table Discovery & Semantics Recovery

Predation or Symbiosis?

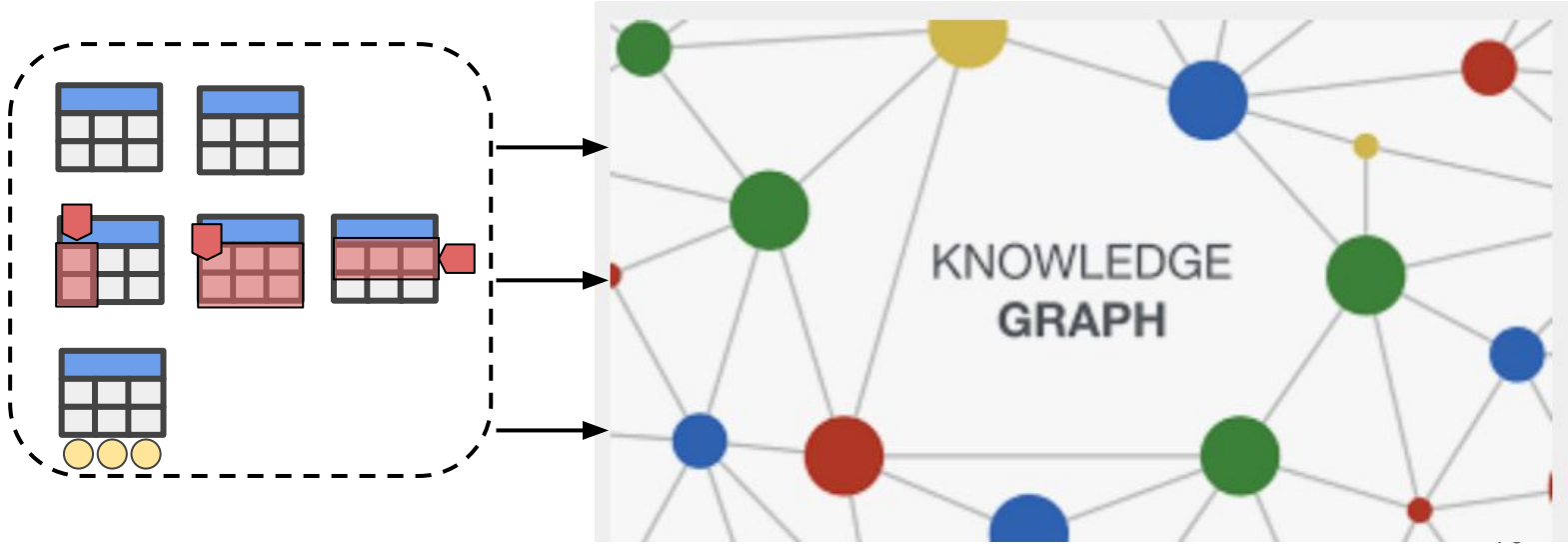


[Lehmberg&Hassanzadeh ISWC 18]
[Zhang&Balog SIGIR 19]
[Helal et al PVLDB 21]

Data Lake as KB

KB augmentation or completion

Judiciously map tuples, columns, tables to KB to augment knowledge



[Zhu et al. PVLDB 16]
[Zhu et al. SIGMOD 19]
[Santos et al. ICDE 22]
[Fan et al. PVLDB 23]

Theory of Data Lake Indexing

For DBMS/DW have Periodic Table of Data Structures

- For one (or few) tables
- Goal speed query or update processing

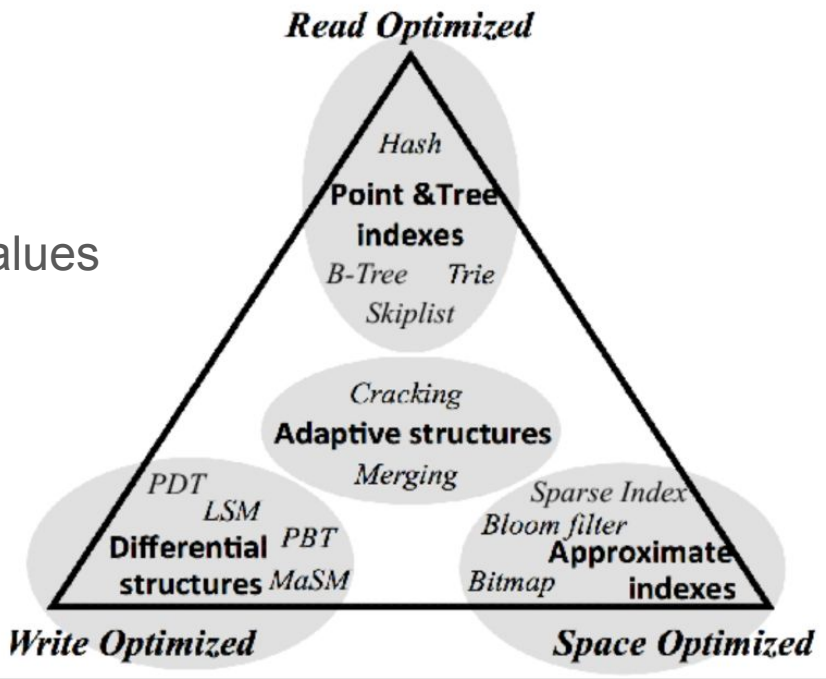
What is equivalent for Data Lakes?

- Index millions of tables, billions+ columns/values
- Goal is accurate (and fast) table discovery

Inverted indices, LSH, HNSW, Sketch-based…

- All required innovation for real data lakes

[Idreos et al. DEBull18]



Batch vs. Query-Time

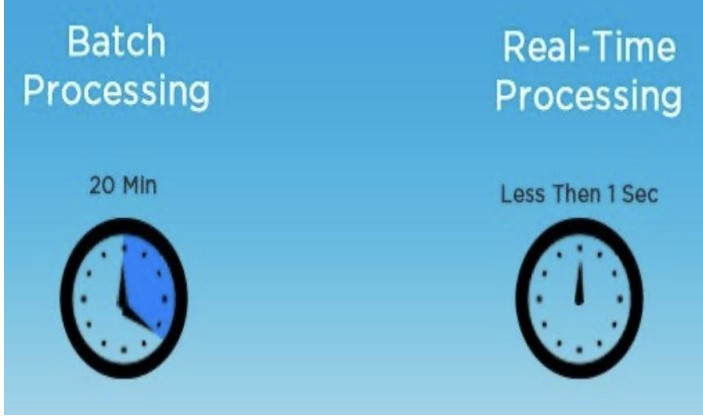


Table Understanding

- Today requires expensive offline processing
- Can we ingest new tables in real-time?

Table Discovery

- Incremental, real-time update of indices for lakes
- Data-driven cost models

Table Navigation, Exploration

- Today requires offline organization construction
- Can we produce organizations at query time

[Nargesian et al. SIGMOD 20]
[Oulette et al. PVLDB 21]

Responsible Data Science

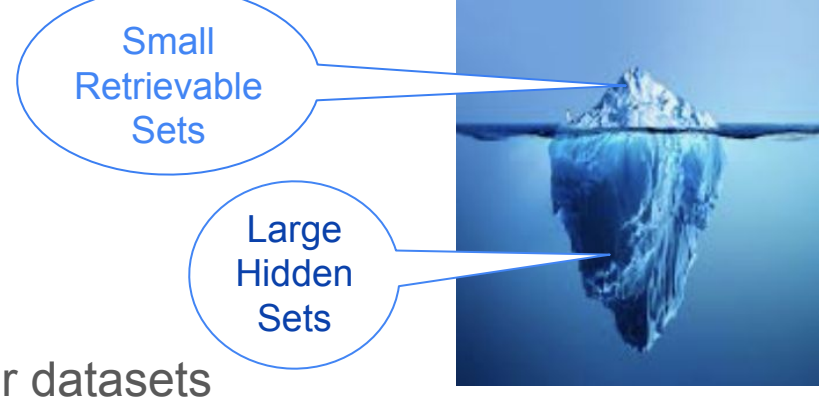


Table Discovery can be ***unfair***

E.g., Jaccard can lead to unfairness for larger datasets

Even when all datasets are indexed

How to ensure all relevant answers are equally likely to be retrieved?

Beyond size ensuring fairness for other data characteristics

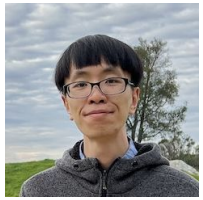
Table Discovery to satisfy ***Group Representation*** or ***Distribution Requirements***

Responsible Data Integration **[Nargesian et al. SIGMOD 22]**

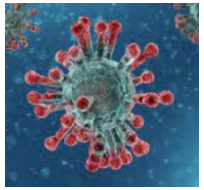
Table Understanding to identify bias

Grace Fan

Jin Wang



Yuliang Li



Renée J. Miller

Table Discovery in Data Lakes: State-of-the-Art and Future Directions

https://northeastern-datalab.github.io/table-discovery-tutorial/