



# MODEL LAKE MANAGEMENT

**Renée J. Miller**  
Canada Excellence Research Chair in  
Data Intelligence



**DATA  
INTELLIGENCE  
LAB** 

# Outline

- **Data Lakes**

- **Difference from DBMS**
  - **A bit of history**
- **Major influential innovations**

- **Model Lakes**

- **What are Model Lakes**
- **Why do Model Lakes need data management?**

# DBMS/Data Warehouse World

- Climate NGO tracking UK Greenhouse gas emissions

Districts	District	Populatio	Unemployment	AvgIncome
	Amersham	100	Low	80
	Barnett	200	High	75
	Chesham	205	Medium	95
	City of London	5500	Medium	200

## LondonEmissions

Fuel Type	District	Sector	KWh
Electricity	Barnett	Domestic	62688
Gas	Barnett	Domestic	206438
Diesel	City of London	Transport	2730044
Oil	City of London	Domestic	430078

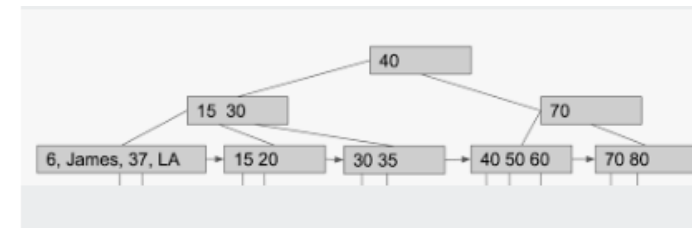
## BuckinghamshireEmissions

Fuel Type	District	Sector	KWh
Electricity	Amersham	Domestic	62688
Oil	Amersham	Domestic	206438
Electricity	Chesham	Transport	2730044
Oil	Chesham	Domestic	430078

*Well Designed: UNA (Unique Name Assumption)*

# Data Warehouse World: Metadata

- Data Catalog
  - Constraints: keys, foreign keys
  - Indices: single table B+, Hash, ...
  - Text descriptions of tables, domains, units



*B+tree over Districts.AvgIncome*

Districts

<b><u>District</u></b>	Population	Unemployment	AvgIncome
------------------------	------------	--------------	-----------

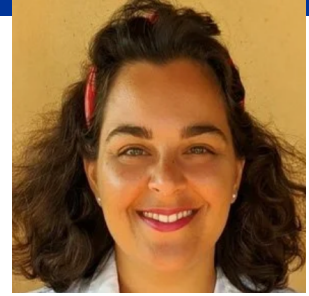
LondonEmissions

<b><u>Fuel Type</u></b>	<b><u>District</u></b>	<b><u>Sector</u></b>	KWh

BuckinghamshireEmissions

<b><u>Fuel Type</u></b>	<b><u>District</u></b>	<b><u>Sector</u></b>	KWh

# Data Lake World



Aristotelis Leventidis Christina Christodoulakis

- Highly heterogenous data

Boroughs	Name	Pop	Unemp	AvgIncome
	Amersham	100K	5%	Medium
	Barnett	20K	15%	Medium
	Chesham, Bucks	205L	6%	Low
	London	5M	7%	High

Pollution

Emissions

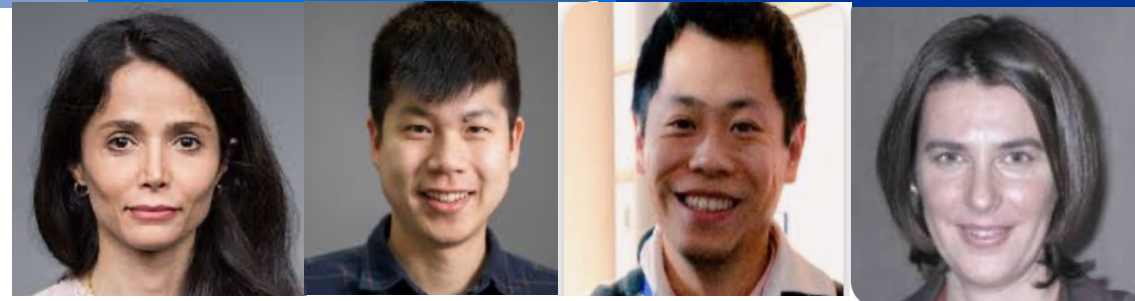
Fuel Type	District	KWh	...
Electricity	Barnett	62688	
Gas	Barnett	206438	
Diesel	City of London	2730044	
Oil	City of London	430078	

Pollutant	Town	Unit	...
Electricity measured in Tonnage			
	Chesham	Low	
	Amersham	Medium	
Petrol measured in KL			
	Chesham	45	
	Amrsham	55	

*Single concept many have names*  
*Same name may mean many things*  
 (Leventidis+EDBT21 Best Paper)

*Tables often non-relational*  
 (Christodoulakis+PVLDB20)  
 and many others

# Data Lakes



Fatemeh Nargesian Erkang Zhu Ken Pu Patricia Arocena

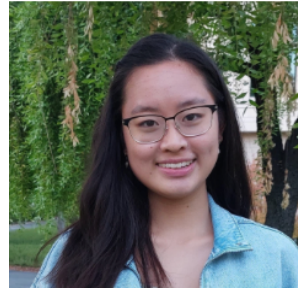


- Data lakes: repositories of large amts of heterogenous data
- Little or often no upfront preprocessing (data design)
- Governments & NGOs release public datasets forming **open data lakes**
- Enterprises have their own **private data lakes**

# Data Lakes

- Enterprise Data Lakes
  - Double every 18–24 months
  - Vary in size from 10's of TB to 100's PB +
  - In knowledge-intensive sectors may have *semantic layer* or knowledge graph/ontology
    - Finance (FIBO), Pharma/Life Sciences (MESH/SNOMED)...
- Open Data Lakes
  - Open Data Portals supporting standard APIs (CKAN/Socrata...)
  - Open Ontologies like YAGO
- Metadata or Catalog services
  - *Metadata often incomplete and inconsistent*

# Data Science Over Data Lakes



Grace Fan

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

(Fan+,SIGMOD23)

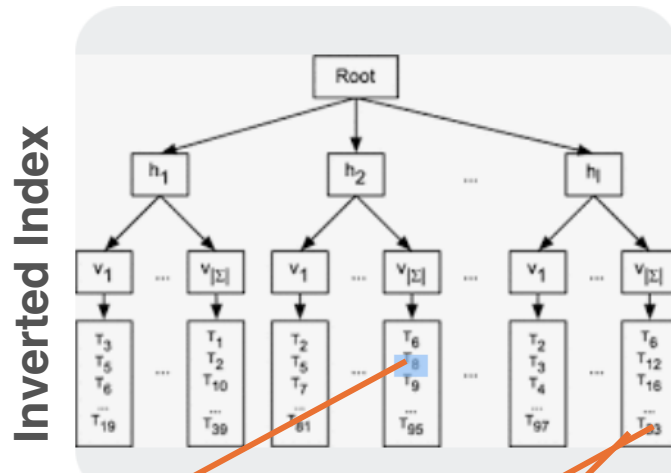
# Data Lake World: Metadata



Mahdi Esmailoghli

- Data Catalog
  - Generally single table only; but no single table indexes
  - Index tables (or columns) not tuples

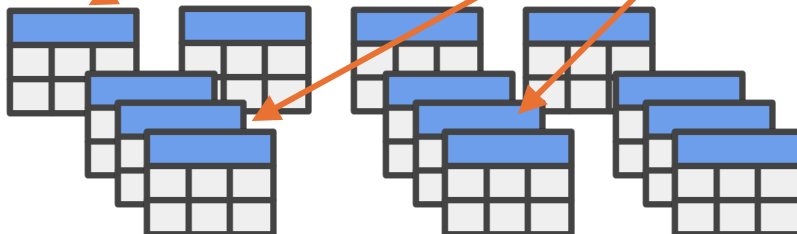
*Example Lake Index (Esmailoghli+ICDE25)*



*Find all tables with an attribute named 'Fuel Type'*

*Find all tables mentioning 'London'*

Data Lake



# Table-as-Query



Fatemeh Nargesian Erkang Zhu

Ken Pu

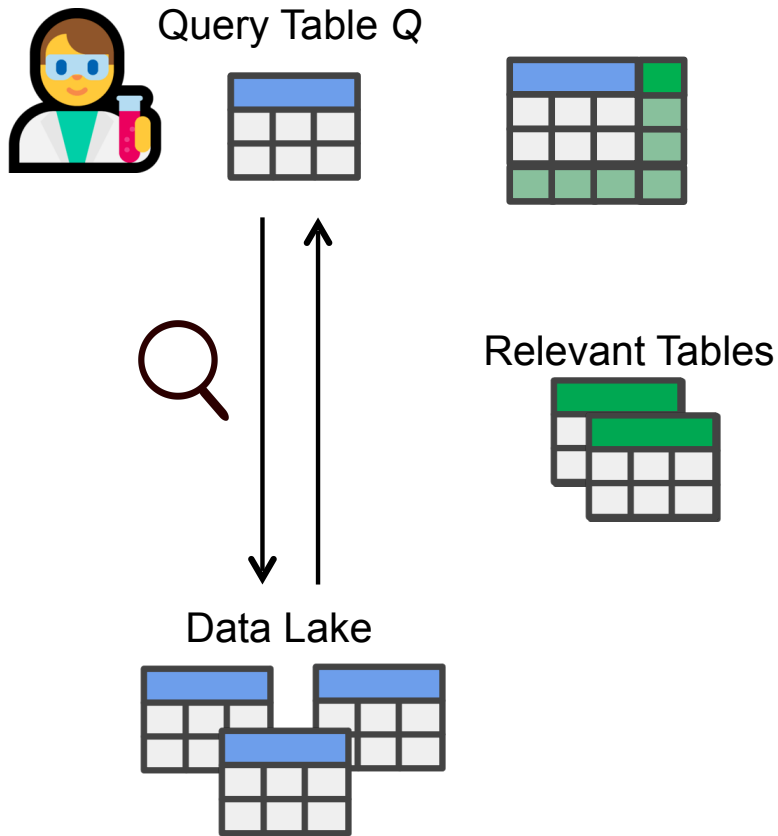


Table-as-Query Paradigm: use entire data context to overcome metadata incompleteness and heterogeneity

Automate Retrieval of Relevant Tables  
**Table Discovery**

(Zhu+PVLDB16)

(Nargesian+PVLDB18)

Joinable Table Search

Unionable Table Search

Extend Q with new *attributes*

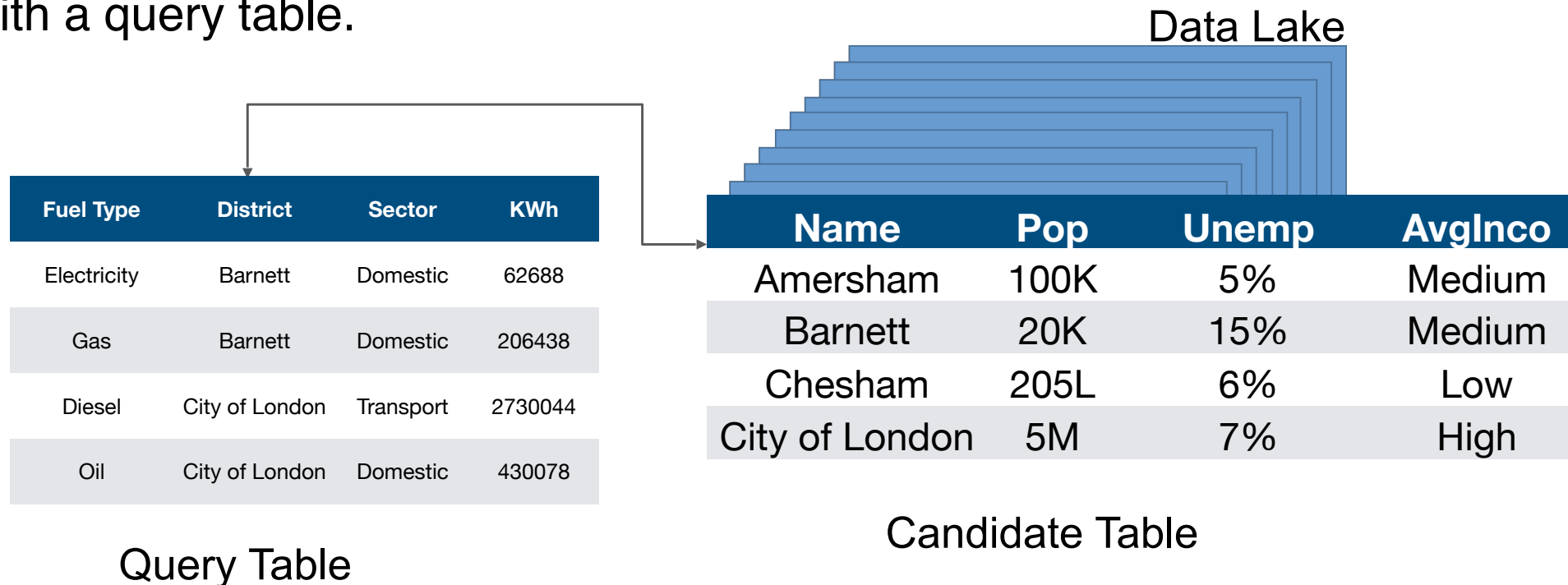
Extend Q with new *tuples*

# Join Table Search

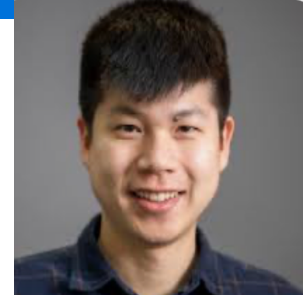
**Data Science Question:** How can I find more features for my model C02 emission?



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



# Joinable Table Search



Erkang Zhu



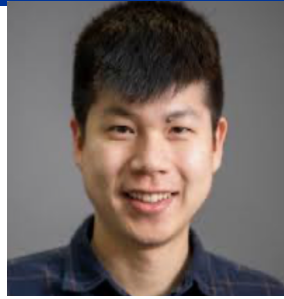
Mahdi Esmailoghli

- Find tables containing values in query attribute
  - Fast (sublinear) computation of set overlap over millions of sets
  - Single column equi-joins

(Zhu+ PVLDB16) (Zhu+ PVLDB17) (Zhu+ SIGMOD19)
- Semantic value matching (Dong+ PVLDB23)
- Multi-attribute join search (Esmailoghli+ PVLDB22)

*Despite heterogeneity of data lakes, exact approaches remain important, especially in knowledge-driven private enterprises using ontologies*

# Joinable Table Search



Erkang Zhu

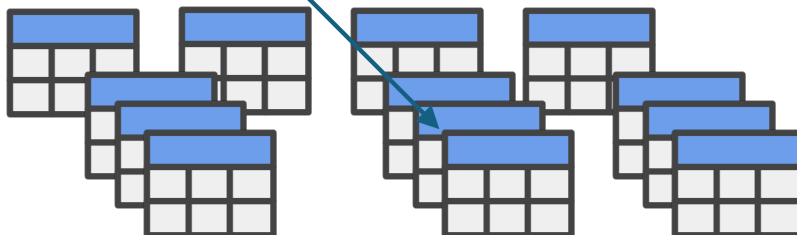
- Find tables containing most or all values in query attribute
  - Computing overlap over millions of sets
    - LSH (Indyk98) well known estimator of set similarity
    - Poor estimator when *sizes can vary dramatically* (Zhu+ PVLDB16) (Zhu+ PVLDB17 Best Demo Award)



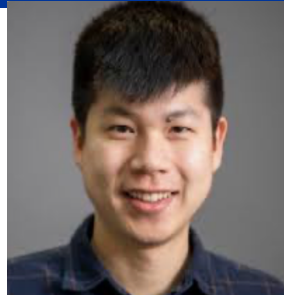
**LSH-Ensemble** uses partitioning that *optimally minimizes false positives* for a given data lake distribution

LSH (Latent Semantic Hashing)

Data Lake

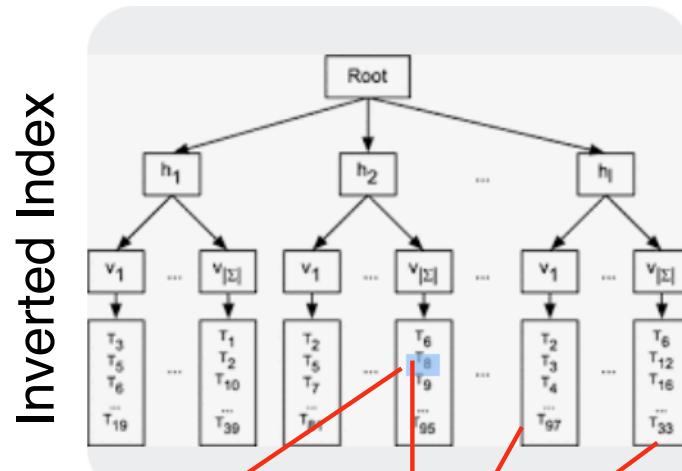


# Joinable Table Search



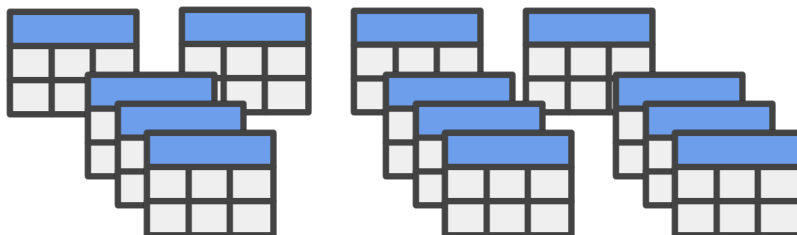
Erkang Zhu

- Exact search using Inverted Index
  - Poor performance when *query large or set sizes vary*
  - **Cost-based inverted-index that is data distribution aware** (Zhu+ SIGMOD19)

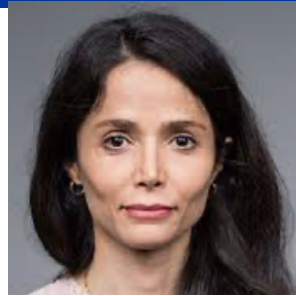


**Josie** for small and medium sized queries, exact computation competes with performance of approximate LSH Ensemble

Data Lake



# Union Table Search



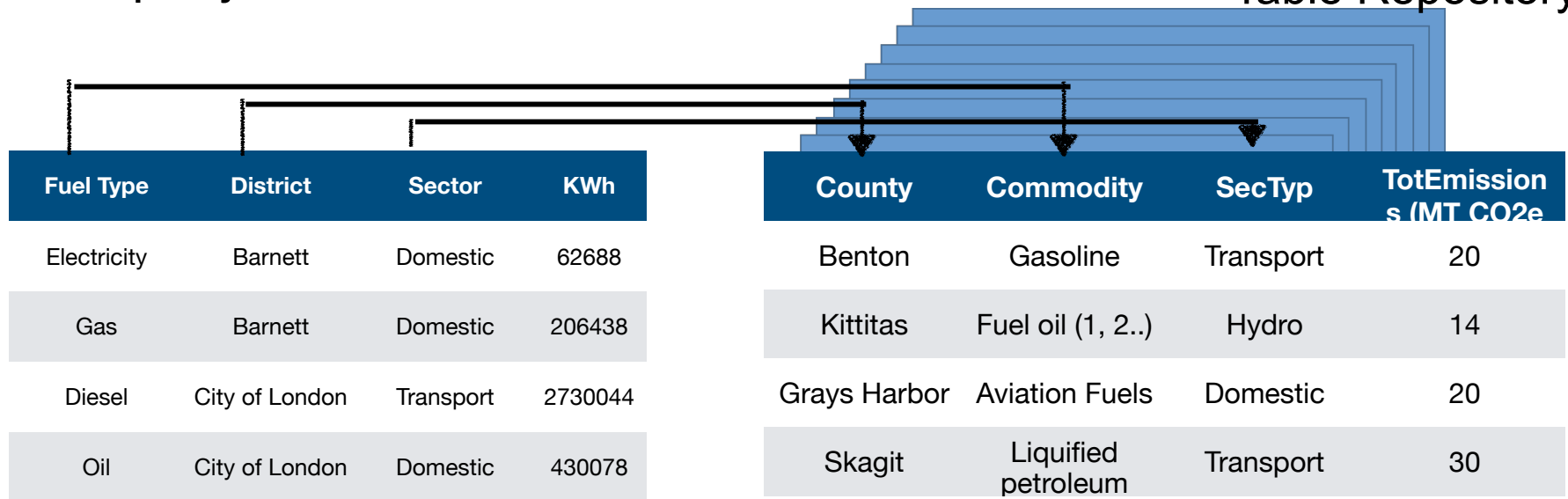
Fatemeh Nargesian

**Data Science Question:** Does my analysis generalize?  
To new regions, new sectors, ...



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

Table Repository



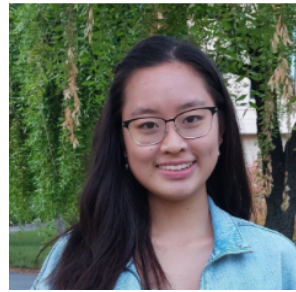
Query Table

Candidate Table

# Unionable Table Search Methods

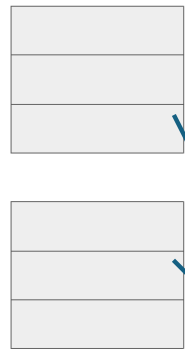
- Table unionability measured based on similarity query table & data lake tables
  - Initially defined based on column unionability (Nargesian+PVLDB18)
  - Extended to relationship unionability (Khatiwada,Fan+ SIGMOD23)
  - Extended to similarity of table context (Fan+PVLDB23)
- **Methods:**
  - DB: value overlap, fuzzy matching (Bogatu+ ICDE20; Nargesian+PVLDB18)
  - KG: concept similarity or overlap (Khatiwada,Fan+SIGMOD23; Nargesian+PVLDB18)
  - Learning: Table or Column Representations (Fan+SIGMOD23)
  - Agentic: Ask an LLM if two tables are unionable (Pal+TADAWorkshop@VLDB24)
    - Classification, not search

# Table Union Search Indexes



Grace Fan

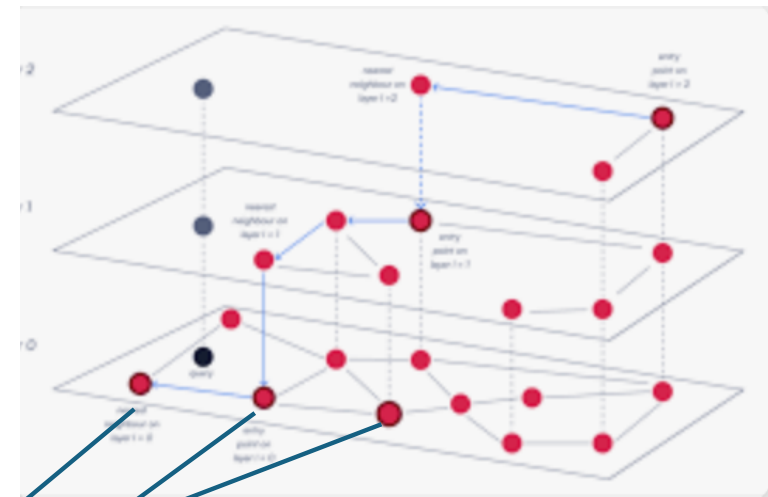
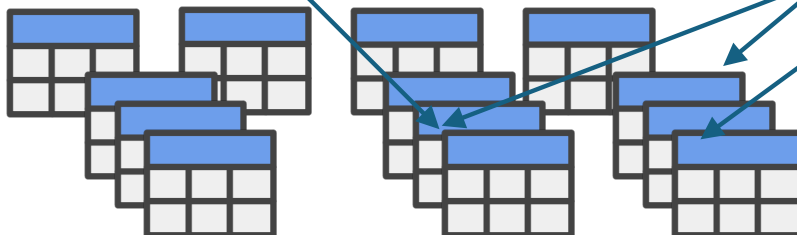
- Learning: Table or Column Representations
  - Starmie (Fan+PVLDB23) uses *self-supervised* approach
  - Overcomes semantic heterogeneity of values and labels
  - Use *HNSW to scale vector* search tremendously



**Starmie** HNSW achieved a 400x speedup over LSH (state-of-the-art data lake index at the time)

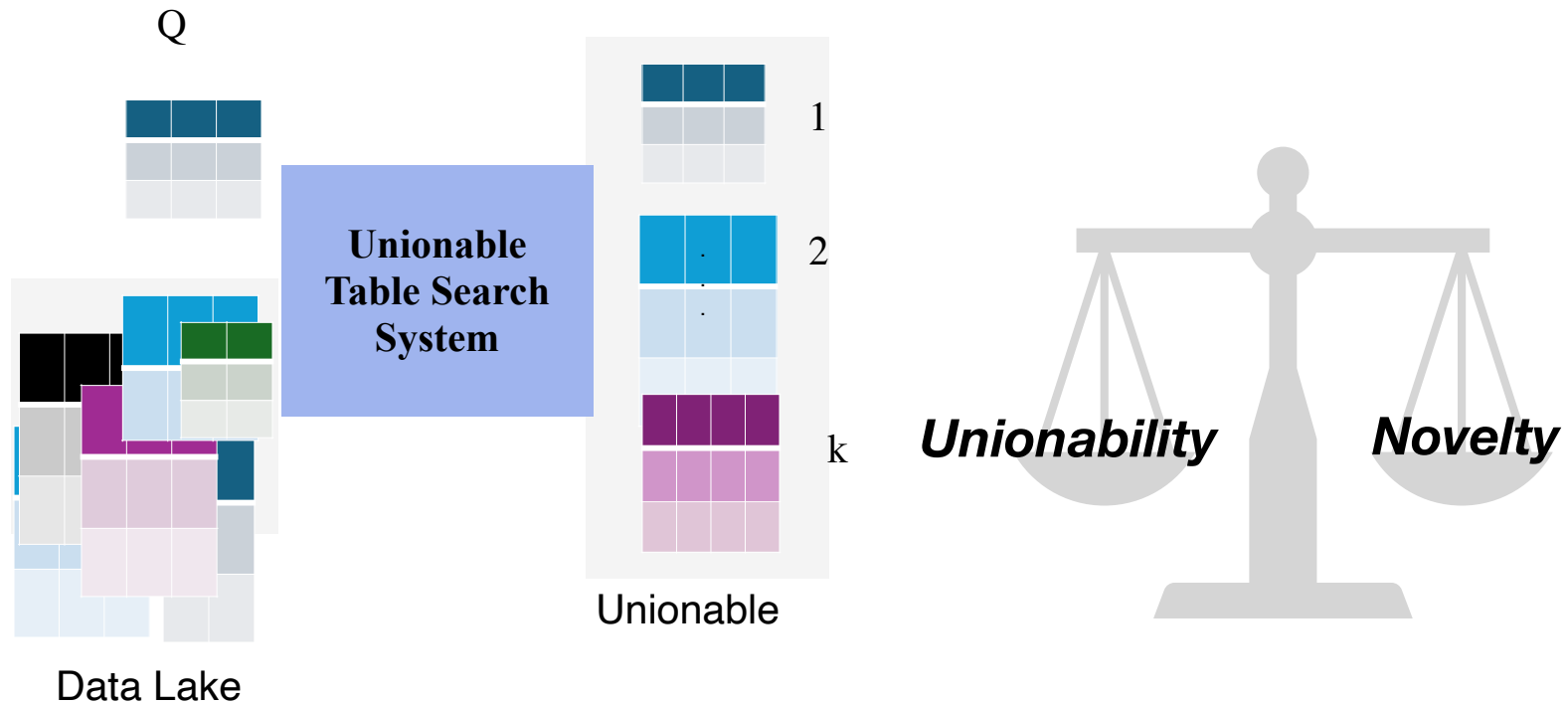
LSH (Latent Semantic Hashing)

Data Lake



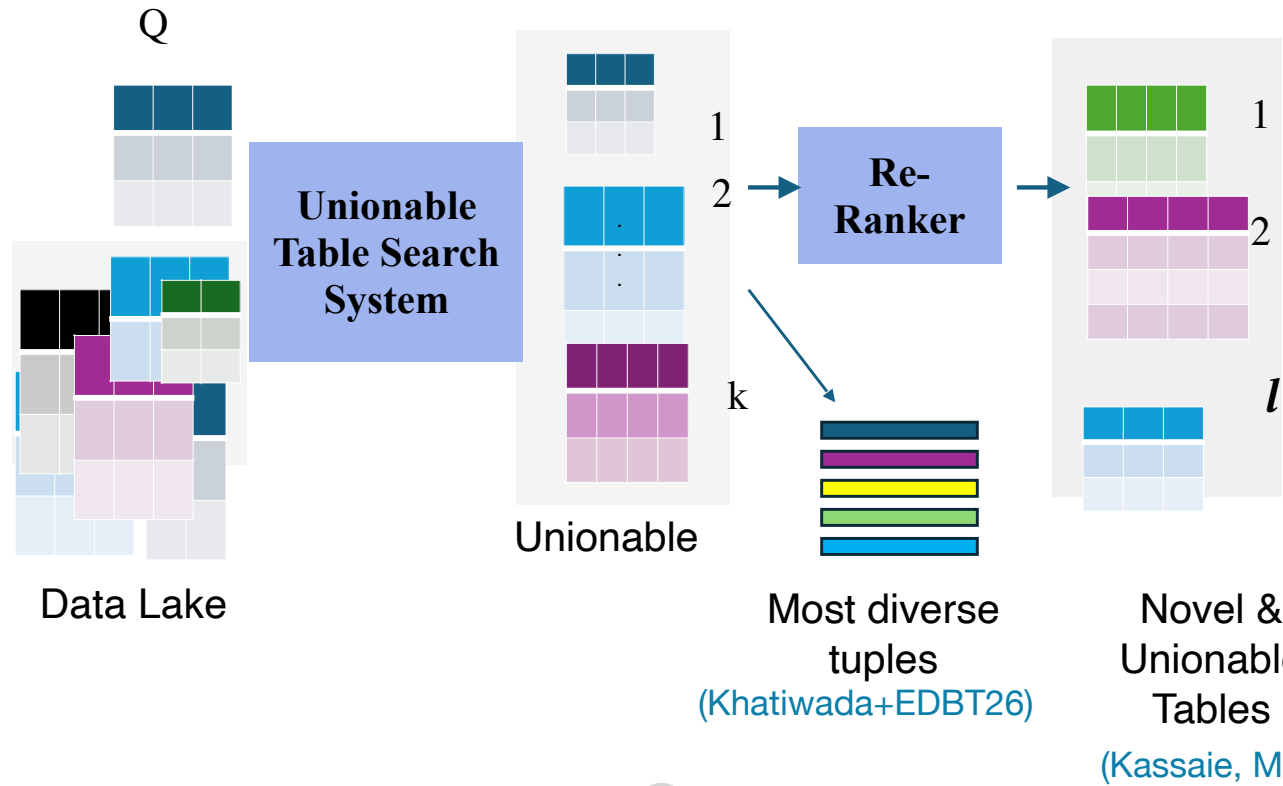
HNSW  
Column or Table Embeddings

# Novel Unionable Table Search



***Unionability*** generally defined as a similarity measure  
Data lakes contain a lot of ***duplication***  
and often many ***different versions*** of a single data set

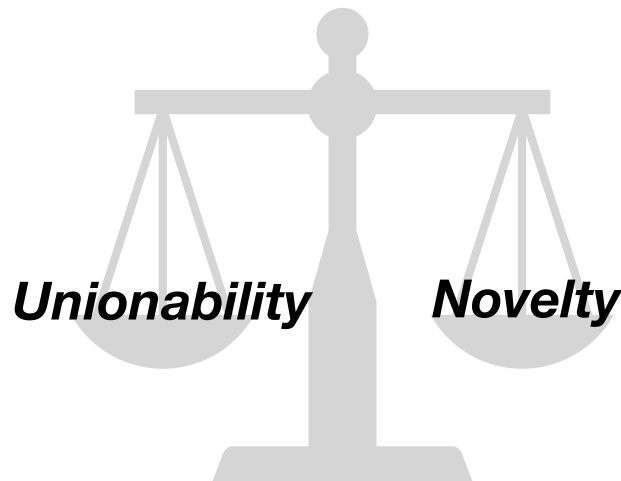
# Novel Unionable Table Search



Aamod Khatiwada



Besat Kassaie



# Distribution Aware Indexing



Chao Zhang



Erkang Zhu

- Josie
  - Inverted index that optimizes access for data distribution
- Ada-ef
  - Vector index (HNSW) that optimizes recall for distribution



**Existing** vector search interface  
`search(k, q, ef)`

Same parameters assigned to each query achieving unknown recall



**Ideal** vector search interface  
`search(k, q, recall)`

Same recall target to each query approximating the recall target

Josie (Zhu+ SIGMOD19)  
Ada-ef (Zhang, Miller SIGMOD26)

# Natural Language Querying of Data Lakes

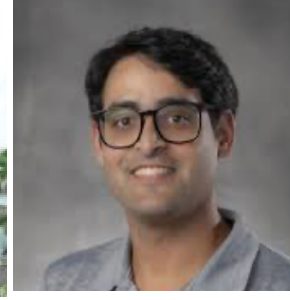
- DBMS & Data Warehouse
  - Text2SQL: NL questions over single DB or warehouse
  - Multi-table Query Answering (MTQA)
    - Subtask is to find tables that can be used to answer the question
    - Solutions assume that joinability/unionability graph are precomputed — don't yet apply to data lakes
      - (Luo+ICDE26 Friday)

# Data Lake Integration

- Once found, how do we integrate these tables?



Aamod Khatiwada



Roeer Shraga

Boroughs

Name	Pop	Unemp	AvgIncome
Amersham	100K	5%	Medium
Barnett	20K	15%	Medium
Chesham	205L	6%	Low
London	5M	7%	High

Emissions

Fuel Type	District	KWh	...
Electricity	Barnett	62688	
Gas	Barnett	206438	
Diesel	London	2730044	
Oil	London	430078	

Pollution

Pollutant	Town	Tonnage	...	...
Electricity	Amersham	Medium		
Electricity	Chesham	Low		
Petrol	Amersham	105		
Petrol	Chesham	45		

Emissions  $\bowtie$  Boroughs  $\bowtie$  Pollution  $\neq$  Boroughs  $\bowtie$  Pollution  $\bowtie$  Emissions

- Full Disjunction:** commutative, associative outerjoin & best semantics for maximally integrating tuples (Galindo-Legaria, SIGMOD94)
  - Scalable algorithm for data lakes (Khatiwada+PVLDB22, SIGMOD23, EDBT26)

# Table Reclamation

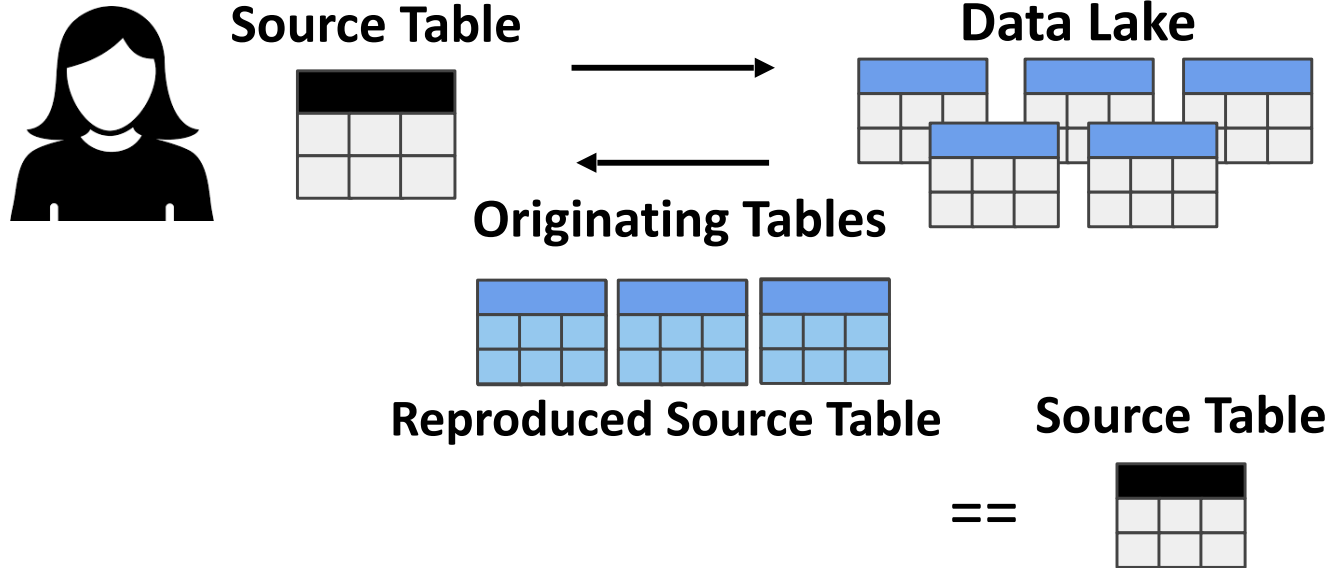


Grace Fan



Roe Shraga

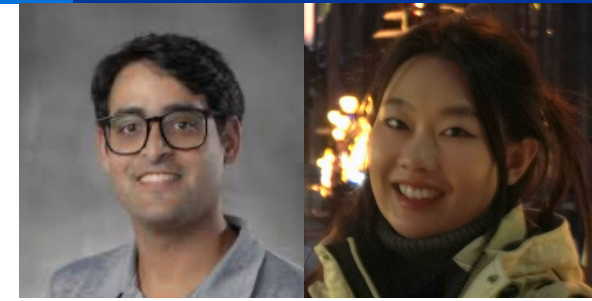
- Goal: Find a set of data lake tables that, when integrated, reproduce a source table as close as possible



- Applications
  - Verify tables produced by LLMs or used in journalism...

(Fan+, ICDE24)

# Version Explanation & Discovery



Roe Shraga

Yuhan Liu

a0	a1	a2	a3	a4
m1	The Godfather (A)	175	9.2	Drama
m2	Hamilton (PG-13)	160	8.6	Drama
m3	The Avengers (UA)	143	8.0	Action
m4	Inception (UA)	NaN	8.8	Action
m5	Moana (U)	107	7.6	Animation

Dataset version by UserA

a0	a1	a2	a3	a4	a5	a6	a7	a8
m1	The Godfather	175	9.2	Drama	A	2.91	4	17
m2	Hamilton	160	8.6	Drama	PG-13	2.67	3	16
m3	The Avengers	143	8.0	Action	UA	2.38	3	17
m5	Moana	107	7.6	Animation	U	1.78	2	9

Dataset version by UserB

## Version Explanation

- Text transforms (Jin+SIGMOD17)
- Text/Numeric/Categorical transforms (Shraga,Miller,PVLDB23)

## Private Version Explanation

(Liu,He,Miller,27)

## Version discovery

(Frenk,Shraga,Archiv25)

# Bias Mitigation

- Can we use table discovery to find new tuples that can be added to a dataset to create real datasets that are unbiased or less biased  
(Scarone+,AIES25)
- Bias mitigation with **coverage constraints** required by ML  
(Scarone+,FAccTS26)
- Bias mitigation over **incomplete data**
  - Reasoning about bias bounds when data is incomplete



Bruno Scarone

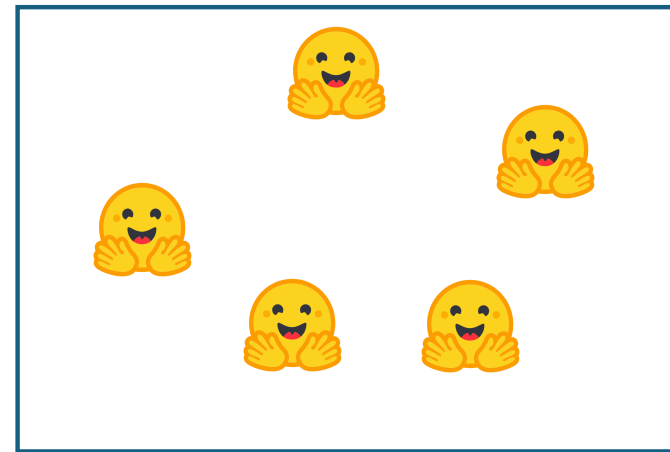
# Open Problems in Data Lake Management

- Natural Language Queries over a lake
  - NL2SQL and MTQA
    - Still struggle with schema & data ambiguity
    - Pain point is finding **bridge tables**
    - **Bridge tables**: table not directly mentioned but required to answer query
  - Data Lakes more heterogeneous and require discovery of **bridge paths**
- Principles of Data Lake Indexing
  - We have a Periodic Table of DBMS indices (**Idreos+DEBu18**)
  - Data Lake Periodic Table will be much sparser, but give insights
- Principles of when to use KG vs Learning vs Agents
- Automatic detection of strength/reliability of data vs metadata signals
- **Numbers!**

# Outline

- Data Lakes
  - Difference from DBMS
  - Major influential innovations
- **Model Lakes**
  - **What are Model Lakes**
  - **Why do Model Lakes need data management?**

# What is a Model Lake?

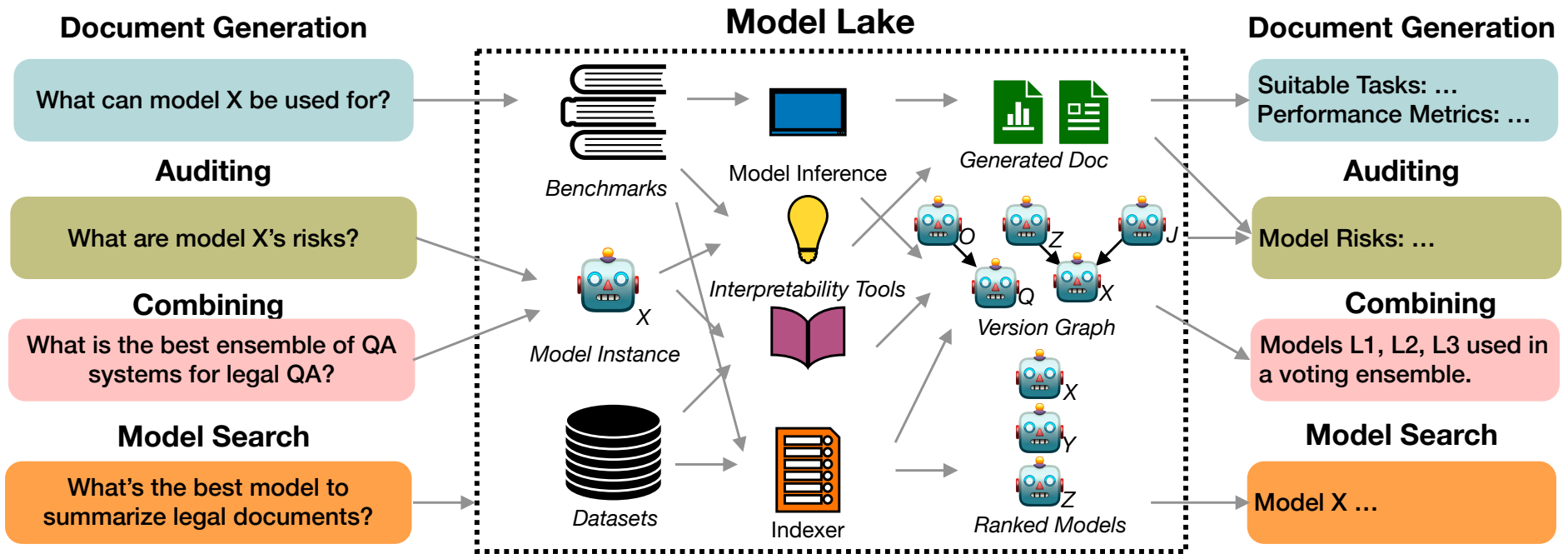


- Many *open model lakes*
  - Hugging Face probably largest and most diverse
- Most data-driven companies maintain a *private model lake*

# Model Lake Vision



Koyena Pal

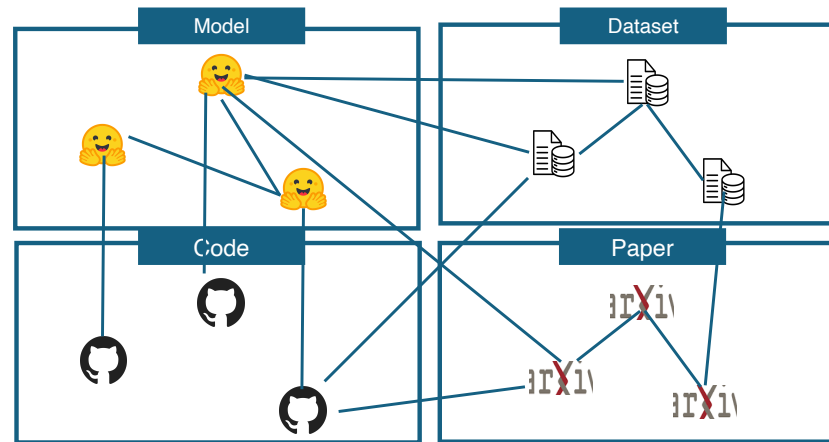


Model Lake Vision (Pal, Bau, Miller EDBT25)

# Model Lake Vision

- Using Model Behavior or Extrinsicics
  - Given a model, can we find other models that perform similarly on a given benchmark? (Lu+SIGRAPH23)
  - Given a task and a well-performing model, can we automatically generate a lighter-weight model that performs as well? (Strassenburg,Glavic,Rabl,archiv26)
  - Alsatian model search for based model in transfer learning. Optimizes candidate model inference over task specific data (Strassenburg,Glavic,Rabl,SIGMOD25)

# What's in a Model Lake?



*Model Lakes  
contain more  
than models!*

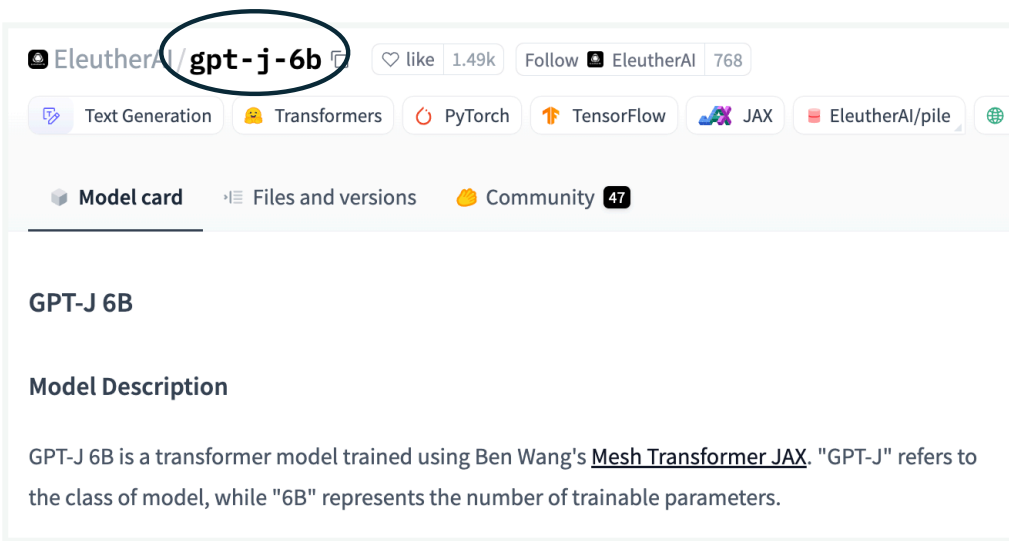
- Model Cards
  - semi-structured descriptions of models
  - link model to other models, to **datasets**, **papers**, **code repositories**
- Dataset Cards
  - semi-structured descriptions of **datasets**
  - link to **papers** and sometimes **models**

# Model Cards

- **Model Details**
  - Training algorithms, parameters, papers, code repository, etc.
- **Intended use**
  - Envisaged use cases and users
- **Metrics**
  - Model performance measures
- **Evaluation Datasets**
  - Details of datasets used for quantitative analysis
- **Training Datasets**
  - When provided should mimic description of evaluation datasets
- **Quantitative analysis**
- **Ethical Considerations**
- **Caveats and Recommendations**

(Mitchell+FAT19)

# Hard to know how each model is related...



EleutherAI / **gpt-j-6b** like 1.49k Follow EleutherAI 768

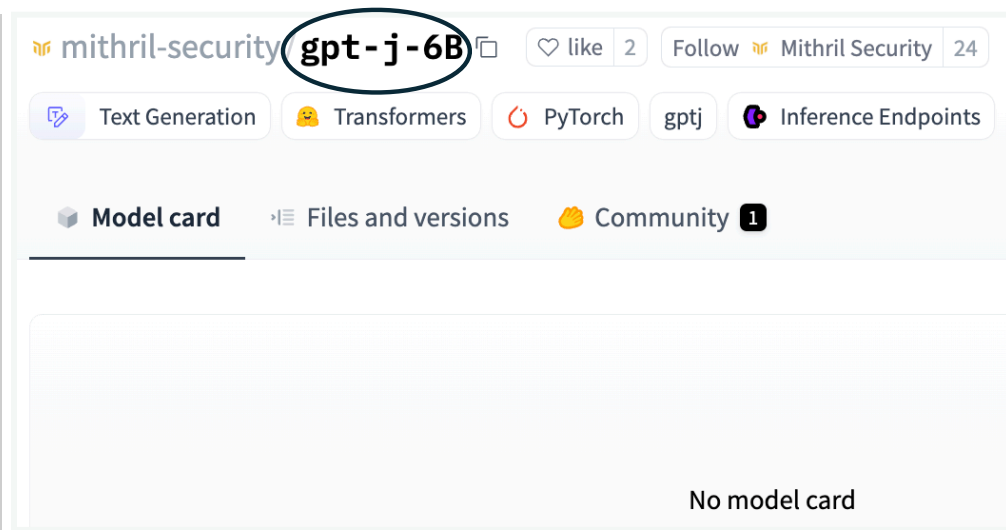
Text Generation Transformers PyTorch TensorFlow JAX EleutherAI/pile

Model card Files and versions Community 47

### GPT-J 6B

#### Model Description

GPT-J 6B is a transformer model trained using Ben Wang's [Mesh Transformer JAX](#). "GPT-J" refers to the class of model, while "6B" represents the number of trainable parameters.



mithril-security / **gpt-j-6B** like 2 Follow Mithril Security 24

Text Generation Transformers PyTorch gptj Inference Endpoints

Model card Files and versions Community 1

No model card

Same names but different models!

## PoisonGPT: How We Hid a Lobotomized LLM on Hugging Face to Spread Fake News

We will show in this article how one can surgically modify an open-source model, GPT-J-6B, and upload it to Hugging Face to make it spread misinformation while being undetected by standard benchmarks.

 Daniel Huynh,  Jade Hardouin | 09 Jul 2023

# Model Lake Metadata Search

- IR approach
  - Keyword search on model names
  - Full-text search on metadata (model cards)
    - e.g., <https://huggingface.co/docs/hub/en/search>

# Model Lake Search: Full-text Search



Full Text Search 🤖

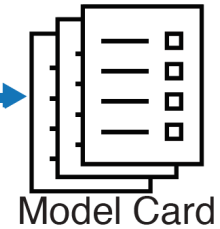


table foundation models that handle small tables with heterogeneous structure

udrearobert999 / multi-qa-mpnet-base-cos-v1-test > README.md model 842 matches

```
tags: setfit, safetensors, mpnet, sentence-transformers, text-classification, generated_from_setfit_trainer, arxiv:2209.11055,
base_model:sentence-transformers/multi-qa-mpnet-base-cos-v1, base_model:finetune:sentence-transformers/multi-qa-mpnet-base-cos-v1,
model-index, region:us

196 # SetFit with sentence-transformers/multi-qa-mpnet-base-cos-v1
197 ...his is a [SetFit](https://github.com/huggingface/setfit) model that can be used for Text Classification. This SetFit mod...
198 The model has been trained using an efficient few-shot learning technique that involves:
199
200 1. Fine-tuning a [Sentence Transformer](https://www.sbert.net) with contrastive learning.
201 2. Training a classification head with features from the fine-tuned Sentence Transformer.
202
203 ## Model Details
204
205 ### Model Description
206 - **Model Type:** SetFit
207 ...
208
209 ### Model Sources
210 ...
211
212 ### Model Labels
213
```

This file contains 829 more matches not shown. [See all 842 matches in the full file.](#)

# Model Lake Search: Full-text Search

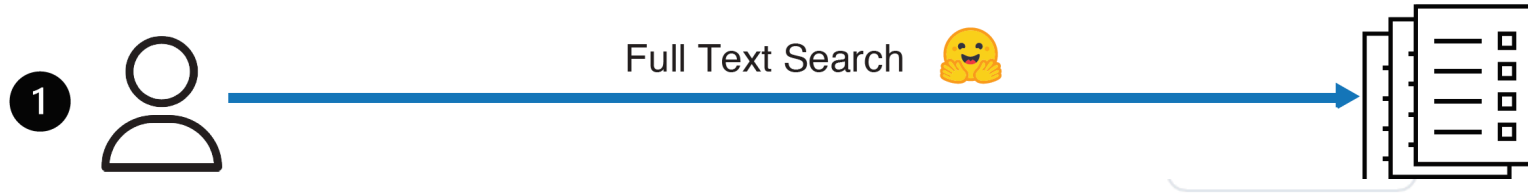


table foundation models that handle small tables with heterogeneous structure

- Top-1 result
  - udrearobert999/multi-qa-mpnet-base-cos-v1-test
- Top-2-4 results
  - udrearobert999/multi-qa-mpnet-base-cos-v1-poc
  - udrearobert999/multi-qa-mpnet-base-cos-v1-scon-3epochs
  - udrearobert999/multi-qa-mpnet-base-cos-v1-contrastive-3e-
- Search can be highly homogeneous due to overlapping description across models

# Model Lake Metadata Search

- IR approach
  - Keyword search on model names
  - Full-text search on metadata
    - e.g., <https://huggingface.co/docs/hub/en/search>
- NLP/SE approach
  - Question answering over standardized structured metadata

- *Retrieve models trained on ImageNet with accuracy over 90%.*
- *Retrieve image classification models that are suitable to deploy on edge devices.*

(Li+ICML22) (Li+IEEE Access23)  
(Toma+ICSE25)

# Model Lake Metadata Search

- IR approach
  - Keyword search on model names
  - Full-text search on metadata
    - e.g., <https://huggingface.co/docs/hub/en/search>
- NLP/SE approach
  - Question answering over standardized structured metadata
- Semantic Search

# Model Lake Semantic Search

😊 Spaces | 🤖 librarian-bots/huggingface-semantic-search 📄

♡ like 89

● Running



table foundation models that handle small tables with heterogeneous structure

## Results

Found 5 results Copy Search Link

**nvidia/nemotron-table-structure-v1** ♡ 27 ↓ 122 40.4% match [→ Find Similar](#)

The Nemotron Table Structure v1 model is a specialized object detection model designed to identify and locate tables in images, particularly those with identifiable structure elements such as cells, rows, and columns, and can be used for downstream tasks including table analysis and data extraction.

[View on Hugging Face Hub](#)

**Prior-Labs/TabPFN-v2-clf** ♡ 77 ↓ 17125 41.7% match [→ Find Similar](#)

TabPFN v2 is a tabular foundation model for large-scale data analysis, leveraging prior-data based learning to achieve strong performance on small tabular datasets.

[View on Hugging Face Hub](#)

**Prior-Labs/TabPFN-v2-reg** ♡ 51 ↓ 122704 43.1% match [→ Find Similar](#)

TabPFN v2 is a transformer-based foundation model for tabular data that leverages prior-data based learning to achieve strong performance on small tabular regression tasks.

[View on Hugging Face Hub](#)

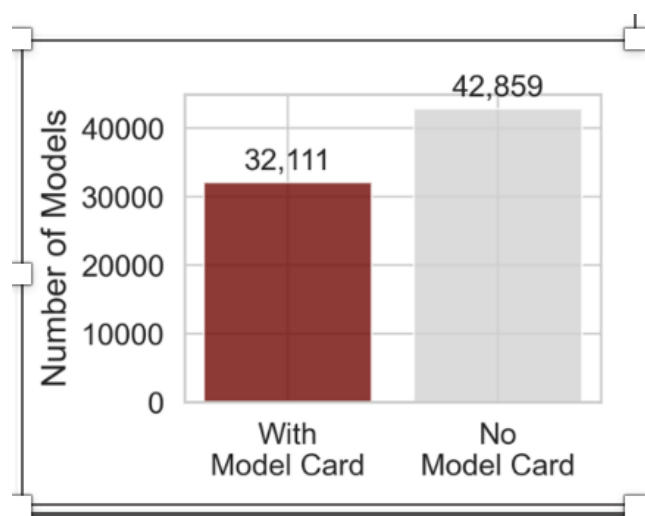
# Model Lake Metadata Search

- IR approach
  - Keyword search on model names
  - Full-text search on metadata
    - e.g., <https://huggingface.co/docs/hub/en/search>
- NLP approach
  - Question answering over common components of dataset and model cards
- Semantic Search
  - *Use model (data) cards as the KG*
  - *Quality of answers depends on quality of KG*

# Can AI overcome noisy/bad data?

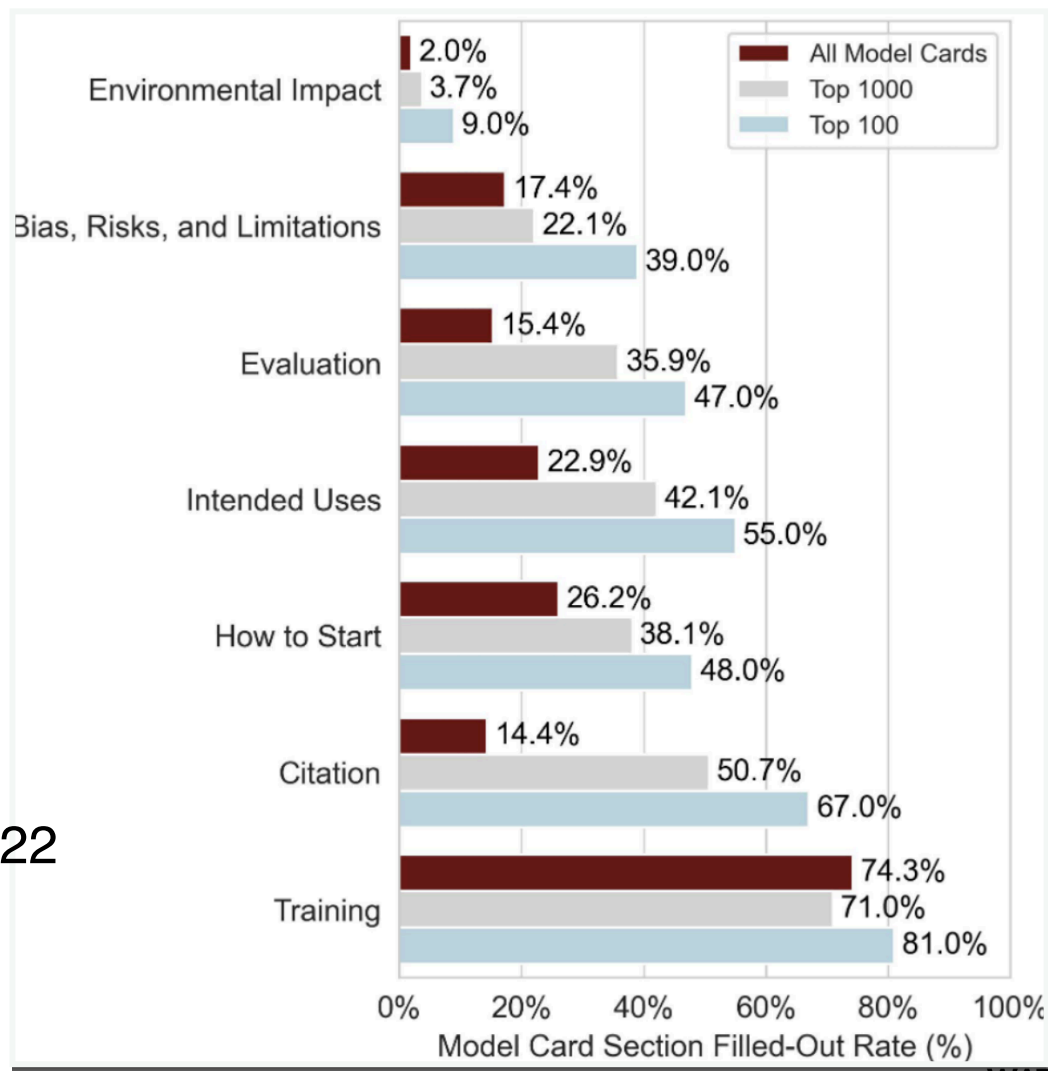
- 2019
  - AI Common wisdom: but of course, everywhere
  - DB Common wisdom: some problems need exact answers
- 2026
  - Of course NOT!

# Model Cards are Incomplete



Huggingface models as of Oct 1, 2022

(Liang+NatMI24)



# Model Lakes Contain Tables



Zhengyuan Dong

- Model Cards
- Code Repository Cards (READMEs)
- Papers
  - arXiv: have HTML extractions containing tables
  - SemanticScholar: use S2ORC

### Tabular Resources

https://arxiv.org/abs/1810.04805, Table 5

Tasks	Dev Set				
	MNLI-m	QNLI	MRPC	SST-2	SQuAD
	(Acc)	(Acc)	(Acc)	(Acc)	(F1)
BERT BASE	84.4	88.4	86.7	92.7	88.5
No NSP	83.9	84.9	86.5	92.6	87.9
LTR & No NSP	82.1	84.3	77.5	92.1	77.8
str(+ BiLSTM)	82.1	84.1	75.7	91.6	84.9

https://github.com/google-research/bert/blob/master/README.md, Table 1

	H=128	H=256	H=512	H=768
L=2	2/128 (BERT-Tiny)	2/256	2/512	2/768
L=4	4/128	4/256 (BERT-Mini)	4/512 (BERT-Small)	4/768
L=6	6/128	6/256	6/512	6/768
L=8	8/128	8/256	8/512 (BERT-Medium)	8/768
L=10	10/128	10/256	10/512	10/768
L=12	12/128	12/256	12/512	12/768 (BERT-Base)

```

{
  "corpusid": "52967399",
  "raw_table":
  "System MNLI-(m/mm)...",
  "System Dev Test EM...",
  "Dev Test ESIM+GloVe 51.9..."
}
    
```

System	...	RTE	Average
Pre-OpenAI SOTA	...	2.5k	-
BiLSTM+ELMo+Attn	...	61.7	74
OpenAI GPT	...	56	75.1
BERTBASE	...	66.4	79.6
BERTLARGE	...	70.1	82.1

### Model Card

#### Table

google-bert/bert-base-uncased

Model	#params	Language
bert-base-uncased	110M	English
bert-large-uncased	340M	English
bert-base-cased	110M	English
bert-large-cased	340M	English
bert-base-chinese	110M	Chinese
bert-base-multilingual-cased	110M	Multiple
bert-large-uncased-whole-word-masking	340M	English
bert-large-cased-whole-word-masking	340M	English

Task	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	Average
	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6

#### URL

It was introduced in this paper (<https://arxiv.org/abs/1810.04805>) and first released in this repository (<https://github.com/google-research/bert>) Model: bert-large-uncased (<https://huggingface.co/google-bert/bert-large-uncased>) Model: bert-base-cased (<https://huggingface.co/bert-base-cased>) ...

#### Reference

```

@article{DBLP:journals/corr/abs-1810-04805,
  title = {{BERT}: Pre-training of Deep Bidirectional Transformers for Language Understanding},
  author = {Jacob Devlin et al. , ...}
    
```

# Model Tables



Zhengyuan Dong

## ▪ Benchmark: ModelTables

Query Table: BERT

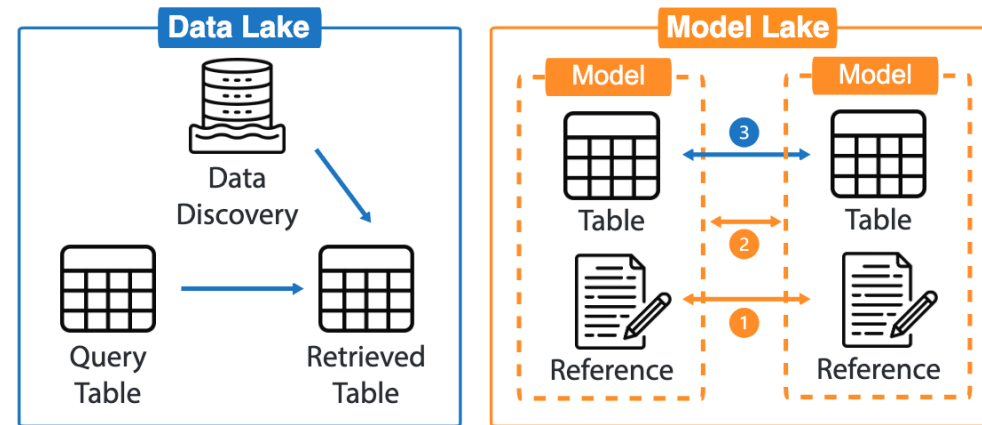
System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	Average
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35	81	86	61.7	74
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36	73.3	84.9	56.8	71
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80	82.3	56	75.1
BERT_BASE	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
BERT_LARGE	86.7/85.9	72.1	92.7	94.9	60.5	86.5	89.3	70.1	82.1

Table 1: RoBERTa

	MNLI	QNLI	QQP	RTE	SST	MRPC	CoLA	STS	WNLI	Avg
Single-task single models on dev										
BERT_LARGE	86.6/-	92.3	91.3	70.4	93.2	88	60.6	90	-	-
XLNet_LARGE	89.8/-	93.9	91.8	83.8	95.6	89.2	63.6	91.8	-	-
RoBERTa	90.2/90.2	94.7	92.2	86.6	96.4	90.9	68	92.4	91.3	-
Ensembles on test (from leaderboard as of July 25, 2019)										
ALICE	88.2/87.9	95.7	90.7	83.5	95.2	92.6	68.6	91.1	80.8	86.3
MT-DNN	87.9/87.4	96	89.9	86.3	96.5	92.7	68.4	91.1	89	87.6
XLNet	90.2/89.8	98.6	90.3	86.3	96.8	93	67.8	91.6	90.4	88.4
RoBERTa	90.8/90.2	98.9	90.2	88.2	96.7	92.3	67.8	92.2	89	88.5

Table 2: ELECTRA

Model	Train FLOPs	CoLA	SST	MRPC	STS	QQP	MNLI	QNLI	RTE	WNLI	Avg.*	Score
BERT	1.9e20 (0.06x)	60.5	94.9	85.4	86.5	89.3	86.7	92.7	70.1	65.1	79.8	80.5
RoBERTa	3.2e21 (1.02x)	67.8	96.7	89.8	91.9	90.2	90.8	95.4	88.2	89	88.1	88.1
ALBERT	3.1e22 (10x)	69.1	97.1	91.2	92	90.5	91.3	-	89.2	91.8	89	-
XLNet	3.9e21 (1.26x)	70.2	97.1	90.5	92.6	90.4	90.9	-	88.5	92.5	89.1	-
ELECTRA	3.1e21 (1x)	71.7	97.1	90.7	92.5	90.8	91.3	95.8	89.8	92.5	89.5	89.4

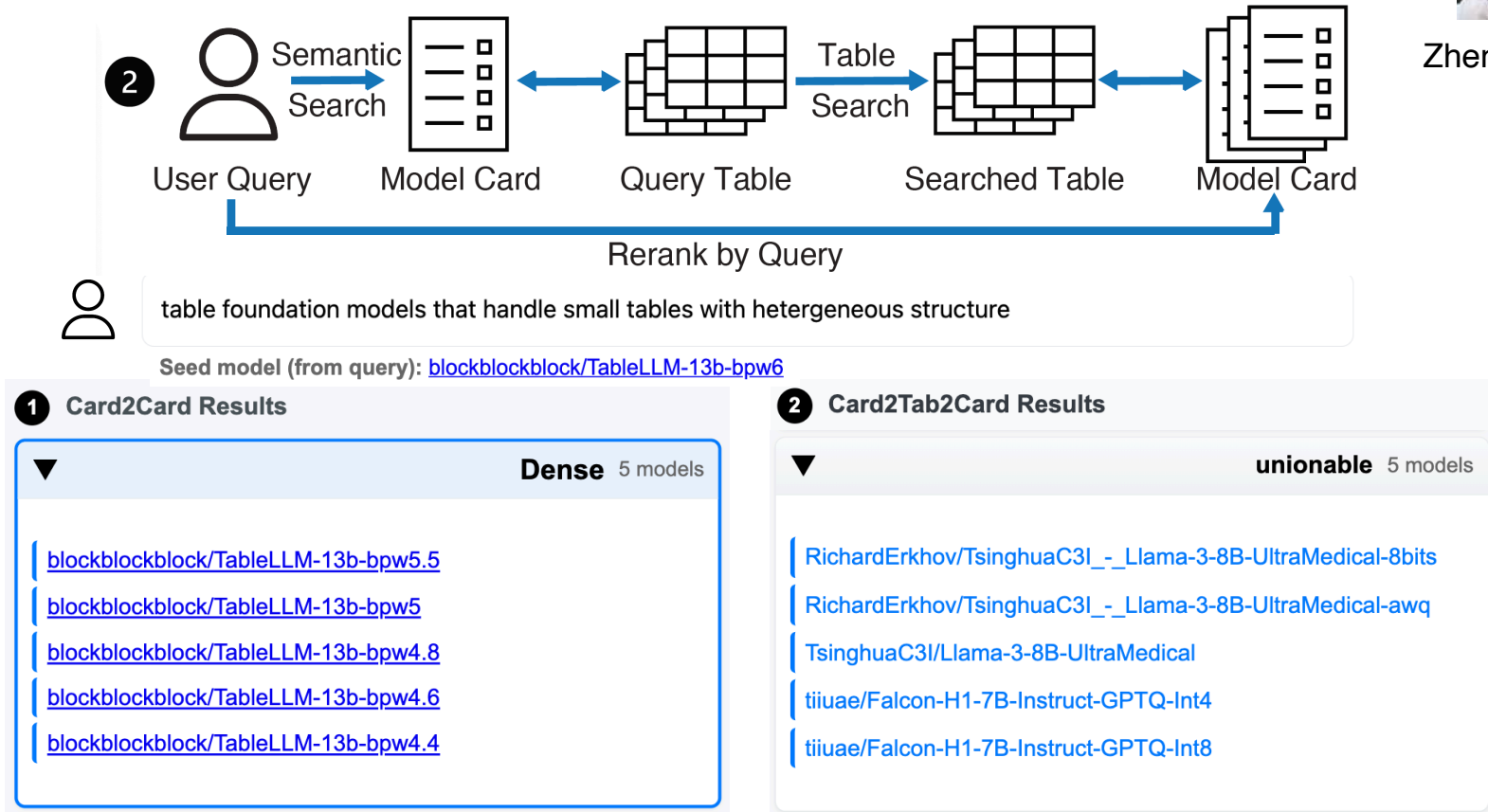


Dong, Z., Zhong, V., & Miller, R. J. (2025). ModelTables A Corpus of Tables about Models. *arXiv preprint arXiv:2512.16106*.

# Model Lake Search Demo



Zhengyuan Dong



Full-text. Dense Retrieval based on card

Table Search-driven Model Search

*Hypothesis: Table search-driven model search, grounded in condensed and structured information, yields results that are both relevant and more diverse.*

# Applications of ModelTables

- Model Card Completion
  - Related model cards may report performance on other models
  - Related papers may contain background, description, tables not containing model cards
  - Predict missing links in ModelTable graph
- Model Card Verification
  - Natural incentive to “inflate” performance & capabilities of model
  - Can we use this rich network of cards, tables, and papers to verify information in model cards
  - Alite ([Khatiwada+PVLDB22,SIGMOD23](#)) permits integration of (consistent) tables: can we extend to detect potential inconsistencies and apply **consistent query answering** ([Fuxman+SIGMOD05](#))

# Vision for Model Lakes

- New vision for using principled data management techniques to search, understand, and optimize model lakes
- ModelTables: new open benchmark containing model tables, model cards, dataset cards, model papers, and a rich graph structure over these artifacts
- We are just a few years into this journey, but I hope I've convinced you that DB folks can contribute substantially to the development of model lakes

# Who did all this work?



Patricia Arocena



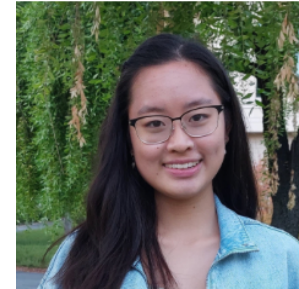
Christina Christodoulakis



Zhengyuan Dong



Mahdi Esmailoghli



Grace Fan



Besat Kassaie



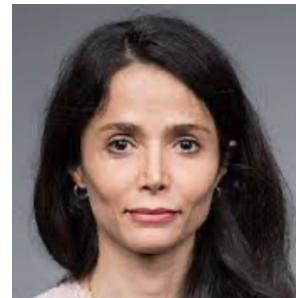
Aamod Khatiwada



Aristotelis Leventidis



Yuhan Liu



Fatemeh Nargesian



Koyena Pal



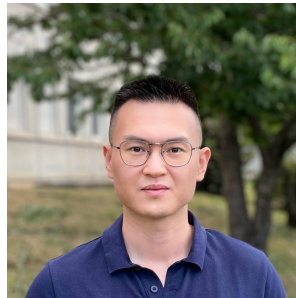
Ken Pu



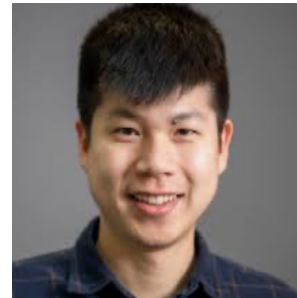
Bruno Scarone



Roe Shraga



Chao Zhang



Erkang Zhu

