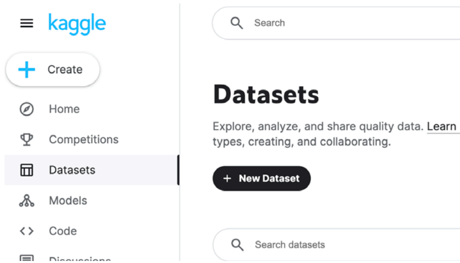
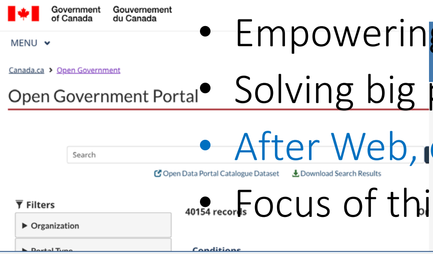
DATA ACQUISITION FOR AI

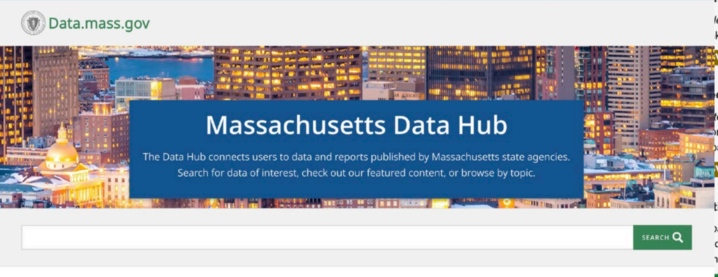
Fatemeh Nargesian, University of Rochester

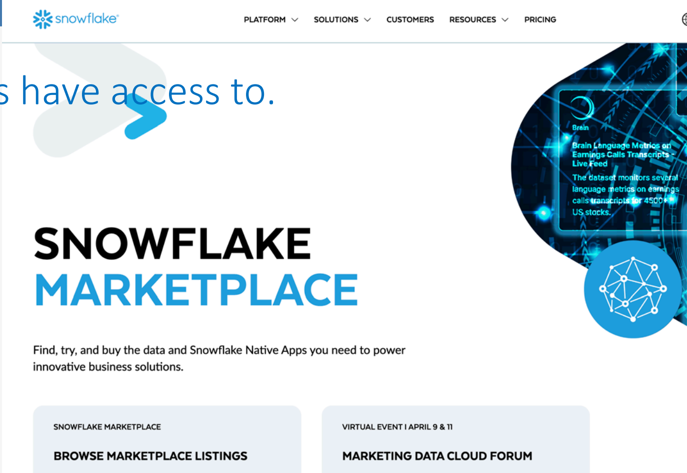
OPEN DATA MOVEMENT

- AI has become ubiquitous.
- Data-centric AI: focus has shifted from big data to good data.
- Open data repositories and data markets have become prevalent.
  - Improving governments
  - Empowering citizens
  - Solving big public problem
  - After Web, open data is the biggest database that citizens have access to.
  - Focus of this talk: tabular and structured data









DATA LAKES VS. TRADITIONAL DATABASES

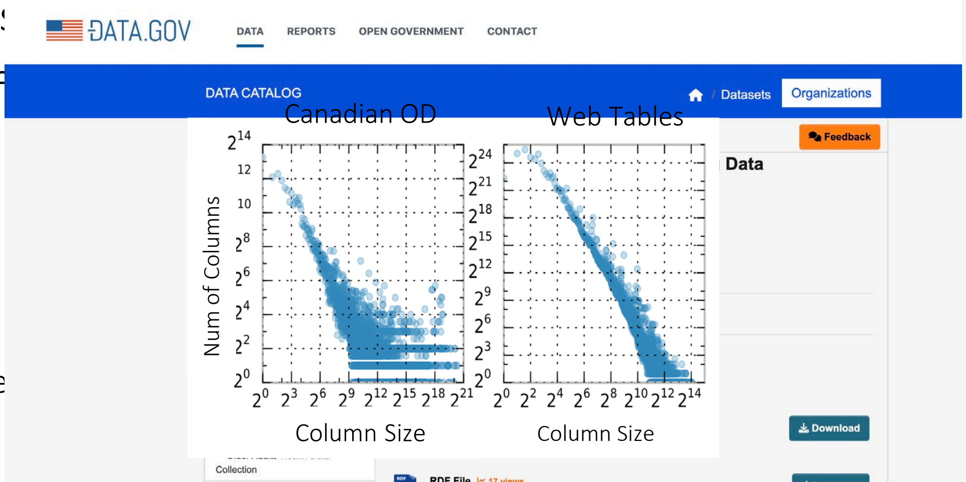
- Data is stored in raw files (csv, xls, xml, ...) and must be extracted
- Large number of (medium-sized) datasets
- No centralized data design or data quality control
- Sparse and non-standardized metadata for datasets
- Skewed data di
- Vast number of

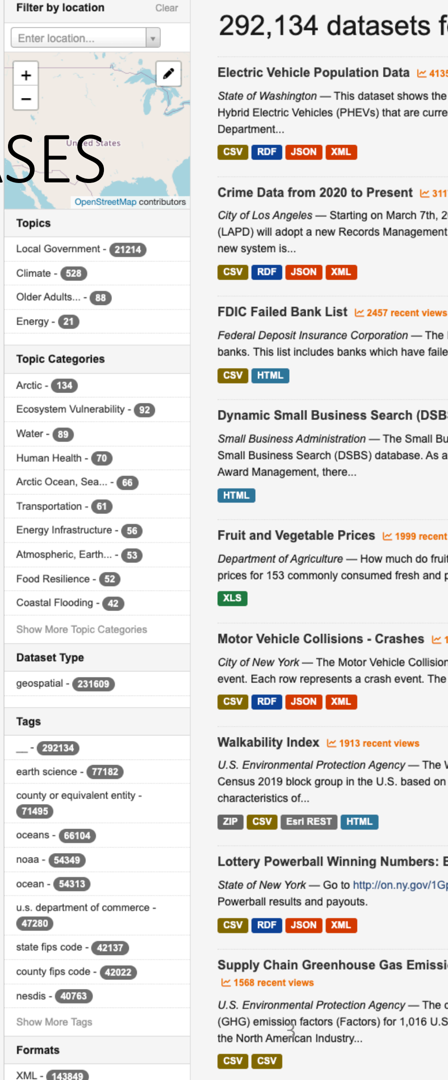
35,675

Canada

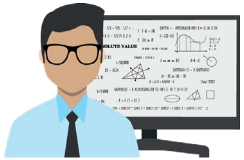
46,626

France





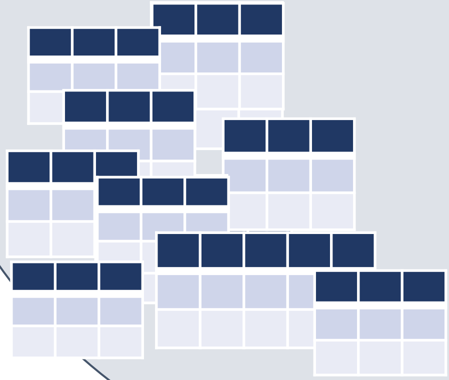
DATA SEARCH AND DATA ENRICHMENT



Analyzing the driving factors of GHG emission!

Enrich data scientist’s work in progress with right data!

Geo	Date	Fuel	t CO_2	Sector	…
Cambridge	2015	electricity	2	Waste	…
Worcester	2021	diesel	20	Metal	…
Camden	2014	coal	12	Oil&Gas	…
NYC	2019	electricity	11	Oil&Gas	…
Boston	2023	diesel	8	Metal	…
Rochester	2021	coal	9	Metal	…
…	…	…	…	…	…

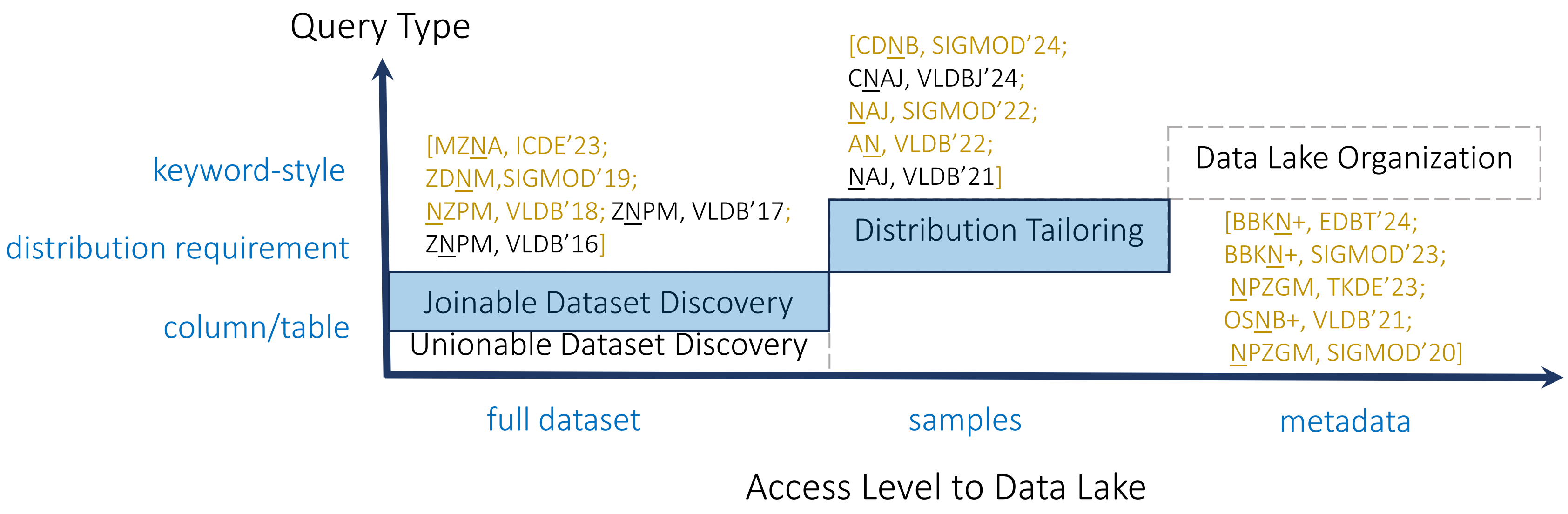


Data enrichment requires dataset discovery

- Adding novel features: joining the query dataset with some datasets in the lake.
- Adding samples: unioning the query dataset with some datasets in the lake.

DATASET DISCOVERY

TASK OF FINDING RELEVANT DATASETS TO A QUERY



JOINABLE DATASET DISCOVERY

How to enrich a query dataset with novel columns and features?

OVERVIEW: JOIN IN DATABASES

- In databases, we often know which columns to join: join on primary/foreign keys

JOIN

StudentInfo

StudentID	Name	Department
100	Sue	CS
101	Yi	Data Science
102	John	Math

DepartmentInfo

Department	Pop	Chair	...
CS	1500	foo	
Data Science	2500	bar	
Math	1000	foo	

DepartmentInfo

StudentInfo

StudentID	Name	Department	Pop	Chair	...
100	Sue	CS	1500	foo	
101	Yi	Data Science	2500	bar	
102	John	Math	1000	foo	

JOIN IN DATA LAKES

- Not obvious which table to join on: makes discovery a search problem
- Joins might not be possible on all query's tuples: smaller result set than query (incomplete data)

query table

Geo	Date	Fuel	t CO2	Sector	…
Barnet	2015	electricity	13	Metal	
City of London	2015	diesel	20	Oil&Gas	
Camden	2014	coal	12	Domestic	
Hackney	2021	Electricity	100	Metal	

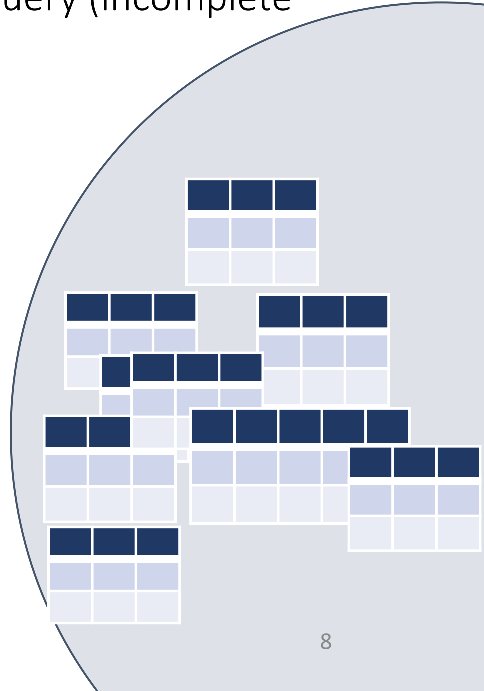
query column

JOIN

Area	Pop	Avg_age	Unemp	…
City of London	242500	43.2	2	
Camden	142500	36.4	4	
Cambridge	389600	37.3	8.5	
…				

?

Geo	Date	Fuel	t CO2	Sector	Pop	Avg_age	Unemp	…
City of London	2015	diesel	20	Oil&Gas	242500	43.2	4	
Camden	2014	Coal	12	Domestic	142500	36.4	2	



JOINABILITY MEASURE

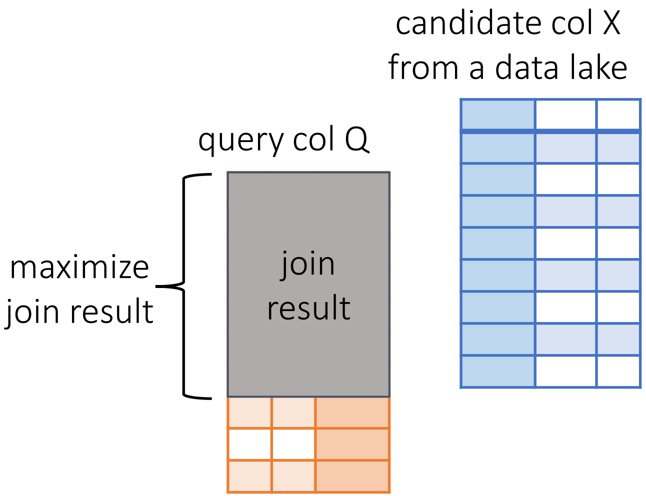
- If columns and query are considered as sets of values, maximize set overlap of query and candidate.

$$\mathrm{Overlap}(\mathrm{Q},\mathrm{X}) = |Q \cap X|$$

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

- Another popular set similarity measure is

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



[ZDNPM,SIGMOD'19; ZNPM,VLDB'16]

SEMANTIC JOINABILITY

- Syntactic measures become ineffective for joining dirty and heterogenous data in the wild.

query table

…	Date	Fuel	t CO2	Sector	Geo
	2015	electricity	130	Domestic	Blaine
	2015	diesel	200	Transport	LA
	2014	coal	125	Domestic	NYC

Overlap(Geo,Area) = 1

Area	Pop	Avg_age	F.Unemp	Unemp	…
LA	8800	43.2	-	-	
Big Apple	242500	36.4	62.9	4	
Blain	389600	37.3	66	8.5	
…					

- Semantic overlap extends syntactic overlap for effective search despite semantic and syntactic heterogeneity of tuples.

KOIOS: Top-K Semantic Overlap Set Search, P. Mundra, J. Zhang, F. Nargesian, N. Augsten, ICDE, 2023.

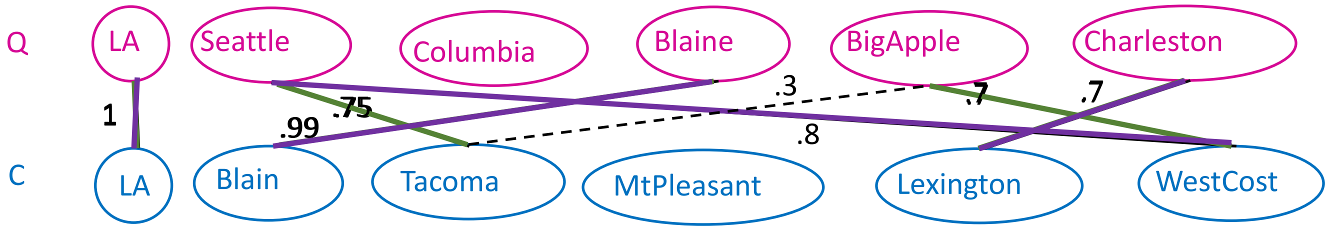
SEMANTIC OVERLAP MEASURE

...	Q
	LA
	Blaine
	BigApple
	...

...	C
	LA
	Blain
	Tacoma
	...

Q = {LA, Seattle, Columbia, Blaine, BigApple, Charleston}
C = {LA, Blain, Appleton, Tacoma, Lexington, WestCoast}

SEMA-JOIN
[He,Gunjam+VLDB'15]



Semantic overlap of two columns Q and C, i.e., SO(Q, C): Maximum matching score of the bipartite graph of Q and C with a user-specified tuple similarity function and threshold.

Greedy matching provides a lower bound for semantic

Edge weighted graph matching: any subgraph M s.t. each node is adjacent to at most one other node.

- Semantic similarity: Cosine on embeddings of col. values, etc.
- Syntactic similarity: Jaccard on n-grams of col. values
- Pruned by similarity threshold α
- Score of a matching: sum of edge weights

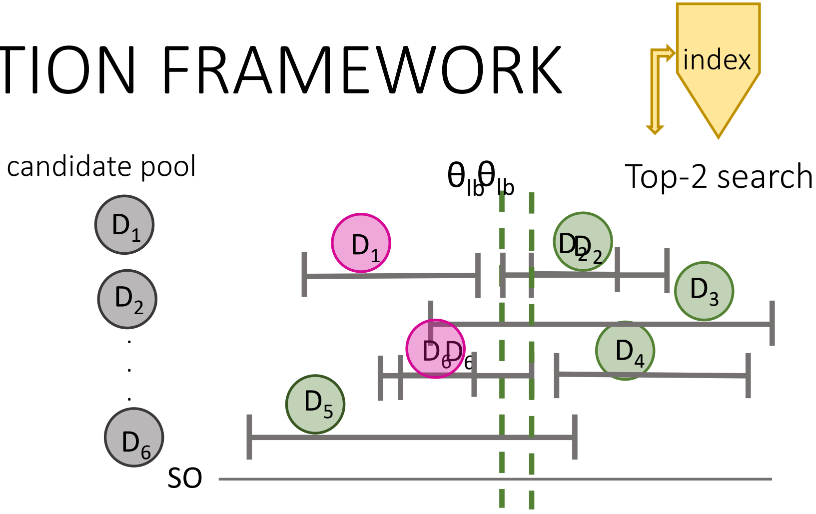
score(M) = 4.14

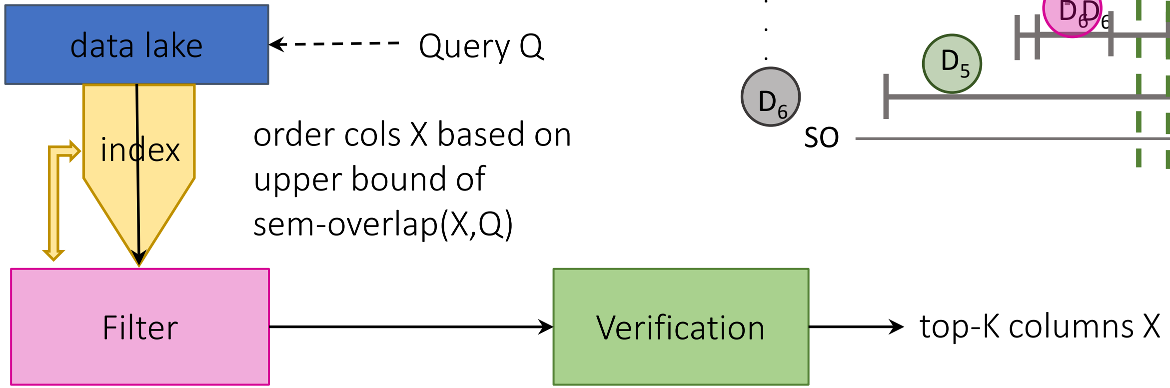
- Join with semantic overlap results in larger join result than syntactic join.
  - $|Q \cap C| \leq SO(Q, C)$
- Discovery requires finding joinable columns and the best way of joining them.

TOP-K SEMANTIC OVERLAP SEARCH

- Semantic overlap can be expensive
  - Bipartite graph matching: $O(n^3)$, n is col. size [Kuhn'1995]
  - Bipartite graph construction: $O(n^2)$
- Problem. Given a column Q and parameter K, find the top-K columns based on the semantic overlap measure.
- Search complexity: $O(mn^3)$, n is the size of cols. and m is the number of sets
- Linear scan over all datasets and computing graph matching is infeasible in practice for data lakes of tens of thousands of datasets.
- Solution. KOIOS is an exact and efficient top-K join search algorithm with semantic overlap.

KOIOS: FILTER-VERIFICATION FRAMEWORK



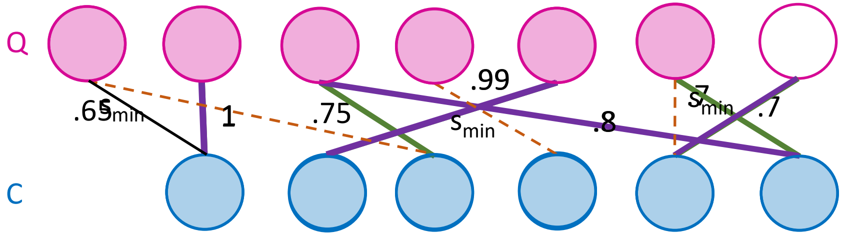


- Upper- and lower-bound filters
- A partitioning scheme for efficient filtering

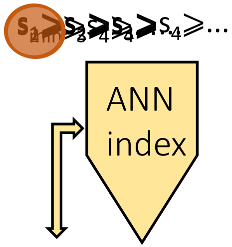
- Early termination of bipartite graph matching

FILTER: INCREMENTAL BOUNDS

Getting edges of a bipartite graph requires |C|x|Q| tuple similarity calculations. Idea. No need to build matchings for all candidate cols. Build partial greedy matches incrementally.



smallest seen edge weight



stream of edges ordered based on row similarity scores descendingly: e.g., simhash LSH for Cosine of row/cell emb vectors.

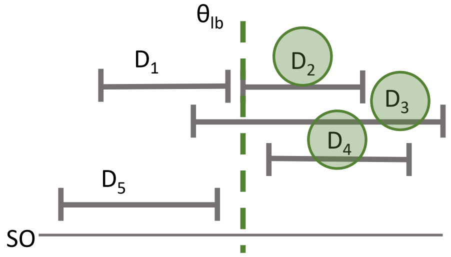
$$SO(Q, C) = maximum\ matching\ score$$

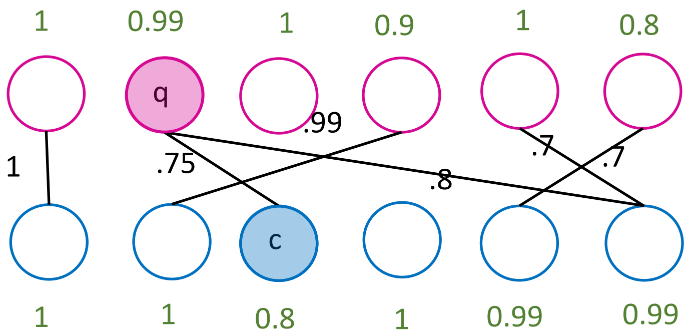
$$incremental\_LB(C) = \sum edge\ weight\ in\ a\ greedy\ \text{“}partial\text{”}\ matching$$

$$incremental\_UB(C) = \sum edge\ weight\ in\ a\ greedy\ \text{“}partial\text{”}\ matching$$

+ smallest seen edge weight . (max matching size - partial matching size)

VERIFICATION: EARLY TERMINATION OF MATCHING





- Hungarian algorithm assigns a valid labeling function *label: nodes→R,* s.t. for two nodes *q* and *c*,
$$\text{label}(q) + \text{label}(c) \geq \text{edge\_weight}(q, c)$$
- The algorithm improves on the labeling function iteratively. At each iteration:
$$\text{bipartite matching score}(Q, C) \leq \sum \text{node lables}$$
$$\text{upper bound}(Q, C) = \sum \text{node lables}$$
- Terminate matching prematurely.

EVALUATION

datasets statistics

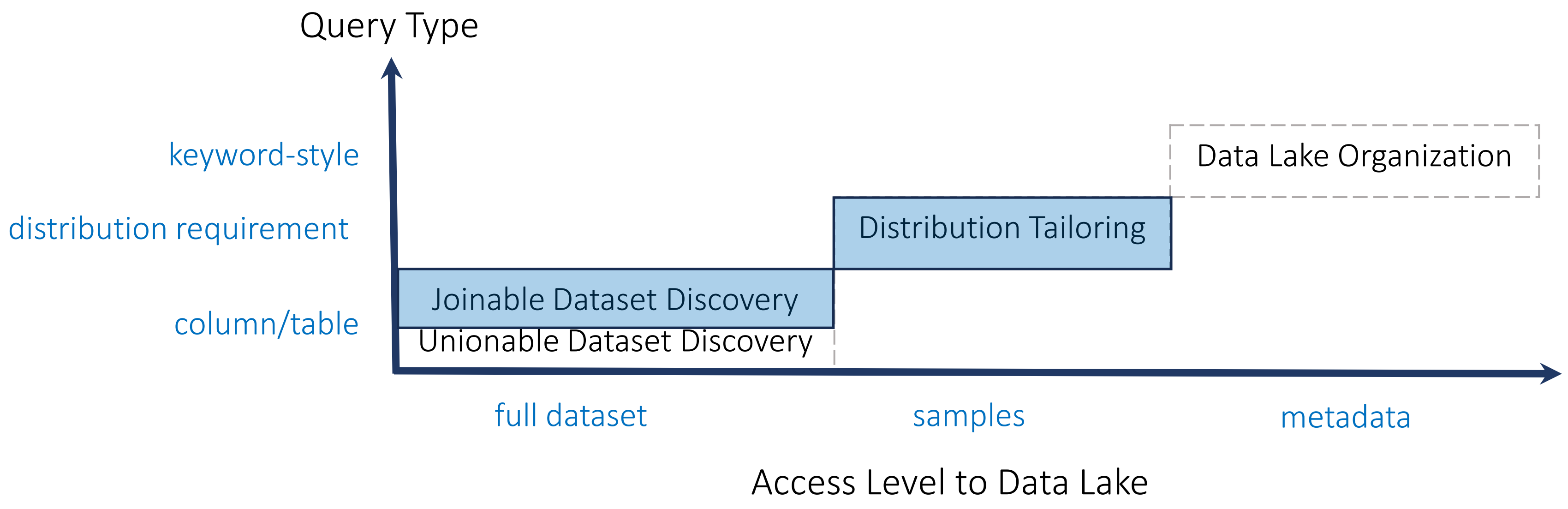
Dataset	#Sets	Max Card.	Avg. Card.	#Unique Elements
DBLP	4,246	514	178.7	25,159
OpenData	15,636	31,901	86.4	179,830
Twitter	27,204	151	22.6	72,910
WDC	1,014,369	10,240	30.6	328,357

comparison to SOTA

Dataset	KOIOS Response Time (s)	SOTA Response Time (s)	KOIOS Mem (MB)	SOTA Mem (MB)
DBLP	0.83	211	0.83	11
OpenData	18.6	101	18.6	102.5
Twitter	0.7	518	0.7	10
WDC	147	1062	147	885

- KOIOS achieves at least 5X speed up over the SOTA on massive data lakes.
- Even better speedup for medium and large queries compared to the SOTA.

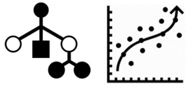
DATASET DISCOVERY



DISCOVERY + DISTRIBUTION REQUIREMENT

- Scenarios with distribution requirements
  - Representation in test data
  - Seeking data with a sufficient representation to avoid overfitting
  - Selection bias leads to flawed and unreliable outcomes.
- Nearest neighbor search index on histograms of groups in datasets [Mao+,AAAI'17]
  - Non-existent results
  - Need to know the group of each tuple apriori at index time.
- Combine multiple datasets to get the distribution we want!

bank loan decision making

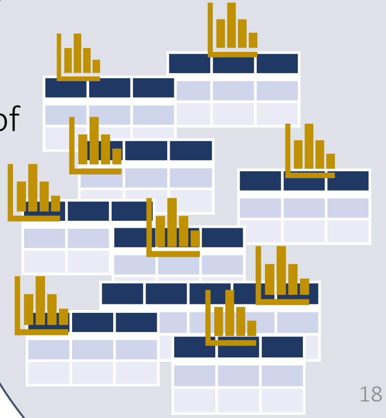


test

At least 30% African American female employees from Texas w/ salary>=55K and at least 30% white male employees w/ salary>=55K

Index

on histograms of target groups



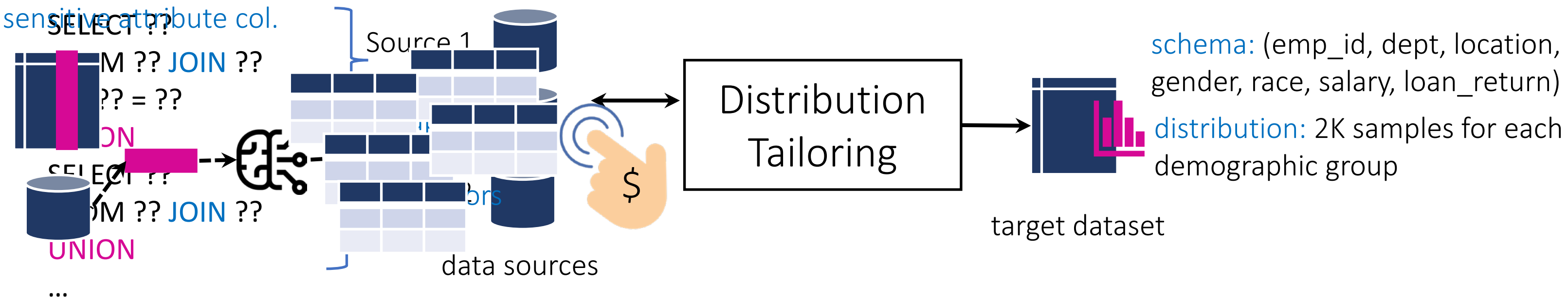
Responsible Data Integration: Next-generation Challenges, F. Nargesian, A. Asudeh, H. V. Jagadish, SIGMOD, 2022.

DISTRIBUTION-AWARE DISCOVERY

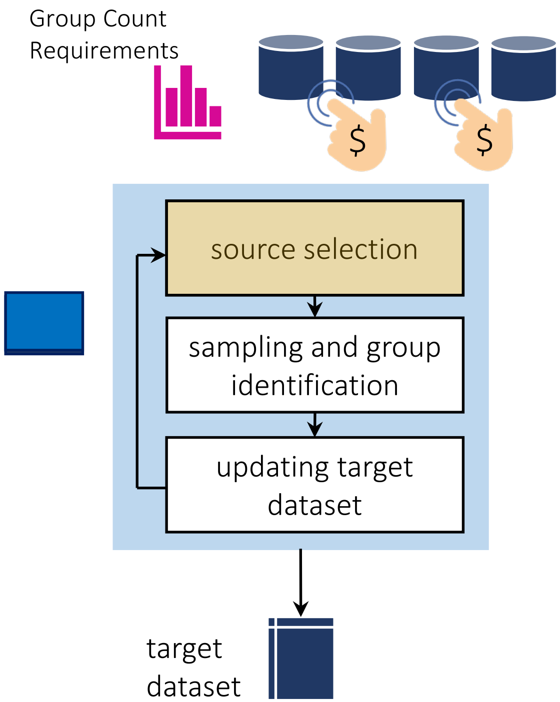
How to construct a dataset that satisfies group distribution requirements from multiple sources in a cost-effective manner?

QUERY/DATA/COST MODELS

- Target dataset: schema + count requirements specified over groups
  - Schema: description of columns
- Data sources
  - Data lake tables with the same schema as target schema
  - Project-join views over a database/data lake: join is expensive to execute, resort to tuple sampling from joins [Zhao+,SIGMOD'18; Li+,SIGMOD'16; Haas+, SIGMOD'99]
  - Other sources: crowd-sources, data providers – data market setting [AN,VLDB'22]: monetary cost for purchasing data



DATA DISTRIBUTION TAILORING (DT)

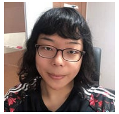


- Problem. Given sources with their costs, and minimum count requirements on the groups, select samples from sources s.t. the union of samples fulfills the count requirements, while the expected total query cost is minimized.
- Solution. Iterative sampling: find a sequence of sources to sample, until distribution requirement is satisfied.
- The cost optimality depends on how much DT knows about group distributions in sources.

DT STRATEGIES FOR (UN)KNOWN GROUP DISTRIBUTIONS

- Known distributions
  - Dynamic programming solution
    - Pseudo-polynomial time and space complexity
    - Not practical for large number of groups and count requirements
  - Optimal strategy for binary groups and sources with equal costs
  - Practical strategy for m-ary groups and sources with arbitrary costs
- Unknown distributions
  - Budget allocation strategy based on multi-armed bandit

Tailoring Data Source Distributions for Fairness-aware Data Integration. F. Nargesian, A. Asudeh, H. V. Jagadish, VLDB, 2021.
Data Distribution Tailoring Revisited: Cost-Efficient Integration of Representative Data. J. Chang, B. Cui, F. Nargesian, A. Asudeh. H. V. Jagadish, VLDBJ, 2024.

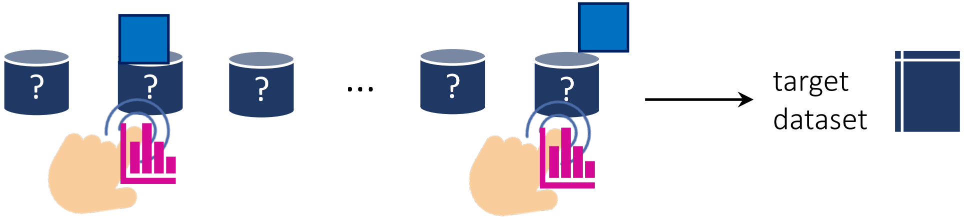


UR PhD student

UR Undergrad student

UNKOWN DISTRIBUTIONS

No information about group distribution



- Learning group distribution and source goodness as we sample.
- Solution. Applying Multi-armed Bandit (MAB)

total reward

OVERVIEW: MULTI-ARMED BANDIT

observed rewards for a_1: 3, 10, 2
arm-value(a_1) = 5

- Given *k* arms a time horizon (budget) *T*, at each timestep *t=1,…,T*, we choose an arm, and receive a real-valued reward R_t.
- Reward depends on the arm and is iid.
- Distributions of rewards are unknown.
- Nevertheless, we must maximize our total reward.
- As selecting arms, form estimates for an arm's value: e.g., the average of sample rewards from the arm.

[Sutton&Barto, 1998]

OVERVIEW: MULTI-ARMED BANDIT

- Maximize the total reward.
- Both try arms to learn their values (explore) and prefer those that appear best (exploit).
  - Never stop exploring; maybe explore less with time; or not!
- Regret is the opportunity loss for one step: difference of obtained reward and optimal reward
- Goal. Minimize total regret ~ maximize cumulative reward

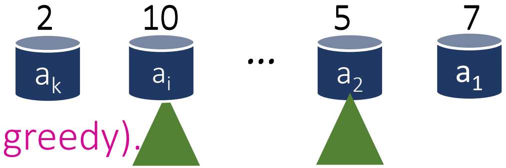
$$\text{Regret(T)} = \text{OPT reward @T} - \text{Planner reward @T}$$

observed rewards so far for t=1,..,4: 3, 10, 5, 10
should have pulled the arm with reward 10 all along
regret: (10-3)+(10-10)+(10-5)+(10-10) = 12

[Sutton&Barto, 1998]

OVERVIEW: EXPLORATION, EXPLOITATION, AND REGRET

estimated action-values:

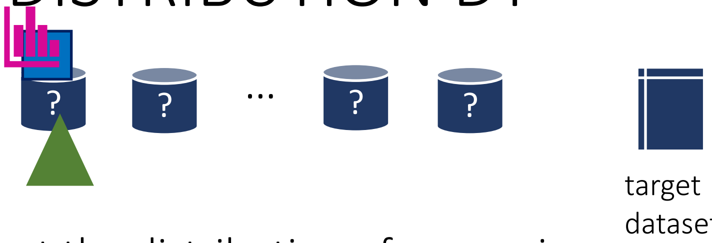


- Explore at time t: select a random arm.
- Exploit at time t: select the arm with the best estimated value so far (greedy).

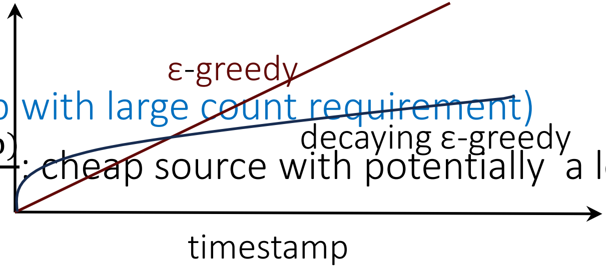
- ε-greedy. At each step explore with prob. ε (exploration rate) and exploit with prob. (1-ε)
  - Linear regret
    - Assume perfect estimates. We potentially pull imperfect arm with ε prob., resulting in expected $\varepsilon . T$ regret.
- Decaying exploration rate. More exploration at the beginning.
  - E.g., if new to a city, extensively explore restaurants at the beginning, explore less later.
  - Regret $O(T^{2/3} log\ T^{1/3})$
  - Can be brought down by Upper Confidence Bound (UCB) to $O(T^{1/2} log\ T^{1/2})$

[Sutton&Barto, 1998]

MAB FOR UNKNOWN DISTRIBUTION DT



- Each source is an arm.
- Apply ε-greedy.
- With each sample from a source, we learn about the distribution of groups in that source.
- Greedy strategy at time *t*.
  - First, select a hard-to-find group (rare group, group with large count requirement)
  - Greedy action. $\text{argmax}_{source} \frac{\frac{\text{total}}{\text{ratio(hard to find group)}}}{\text{source cost}}$: cheap source with potentially a lot of samples of hard-to-find-group.



- Regret is a proxy for cost overhead.
- Regret. DT with ε-greedy strategy with exploration rate $\sqrt[3]{\ln t / t}$ at time *t* has regret of $O\left(t^{\frac{2}{3}} \log t^{\frac{1}{3}}\right)$ -- for sources with equal costs and strategy *ratiocoll*.

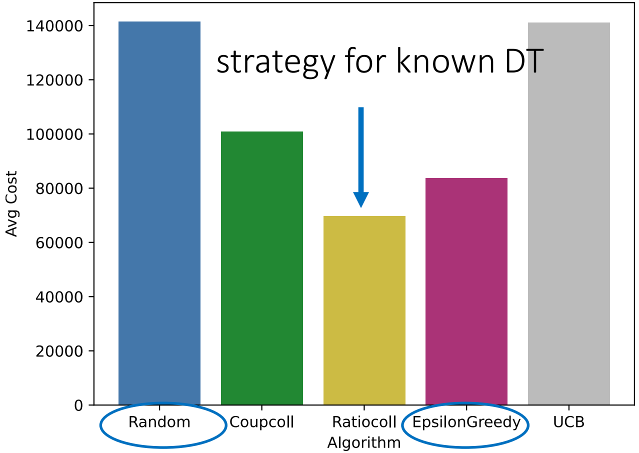
Data Distribution Tailoring Revisited: Cost-Efficient Integration of Representative Data.
J. Chang, B. Cui, F. Nargesian, A. Asudeh. H. V. Jagadish, VLDBJ, 2024.

PRACTICAL STRATEGY

- Explore-then-Exploit
  - Crudely approximate budget $T \approx \sum group\_count\_requirements$
  - Randomly sample sources for αT iterations
    - Performing all explorations at the beginning allows sampling to be done in batches and in parallel!
  - Be greedy in the rest of time steps.

EVALUATION: KNOWN AND UNKNOWN DISTRIBUTIONS

Known/Unknown DT on Flights dataset



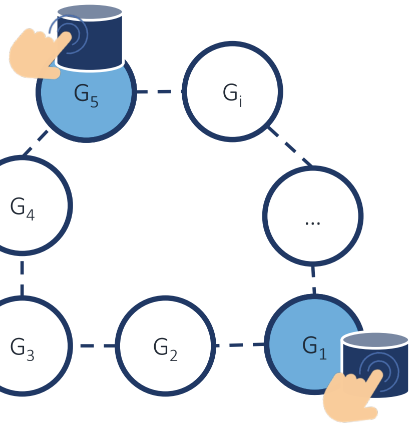
- ε-greedy outperforms random sampling and is in competition with strategy for. Known distribution.

PLUTUS: UNDERSTANDING DT FOR ML

- DT is available in Apache SystemDS (open-source SystemML from IBM: declarative ML system on Spark)

Understanding Distribution Tailoring for Machine Learning.
J. Chang, C. Dionysio, F. Nargesian, M. Boehm, SIGMOD, 2024.

KNOWN DISTRIBUTIONS: BINARY GROUPS AND EQUI-COST



- Given. The ratio of each group in each source. Same cost for all sources.
- Algorithm. At each iteration until all requirements are satisfied:
  - From unsatisfied groups, pick a group to prioritize.
  - For that group pick the cheapest sources for acquiring samples of that group.
  - Sample from that source; update target dataset; update remaining group requirements.

- Cheapest source for a group: source that has the highest ratio of that group
- Group to prioritize/sample for: group that is minority in its best source

Tailoring Data Source Distributions for Fairness-aware Data Integration. F. Nargesian, A. Asudeh, H. V. Jagadish, VLDB, 2021.

EQUI-COST BINARY DT:
SAMPLE FOR MINORITY GROUP

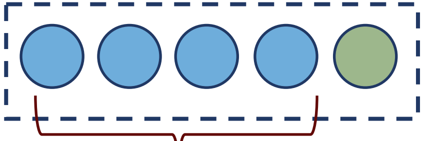
cost=1

S_1 20% of G_1 and 80% of G_2

cost=1

S_2 5% of G_1 and 95% of G_2

sample from S_1 until got one tuple of G_1



No need to separately sample for G_2, req. satisfied

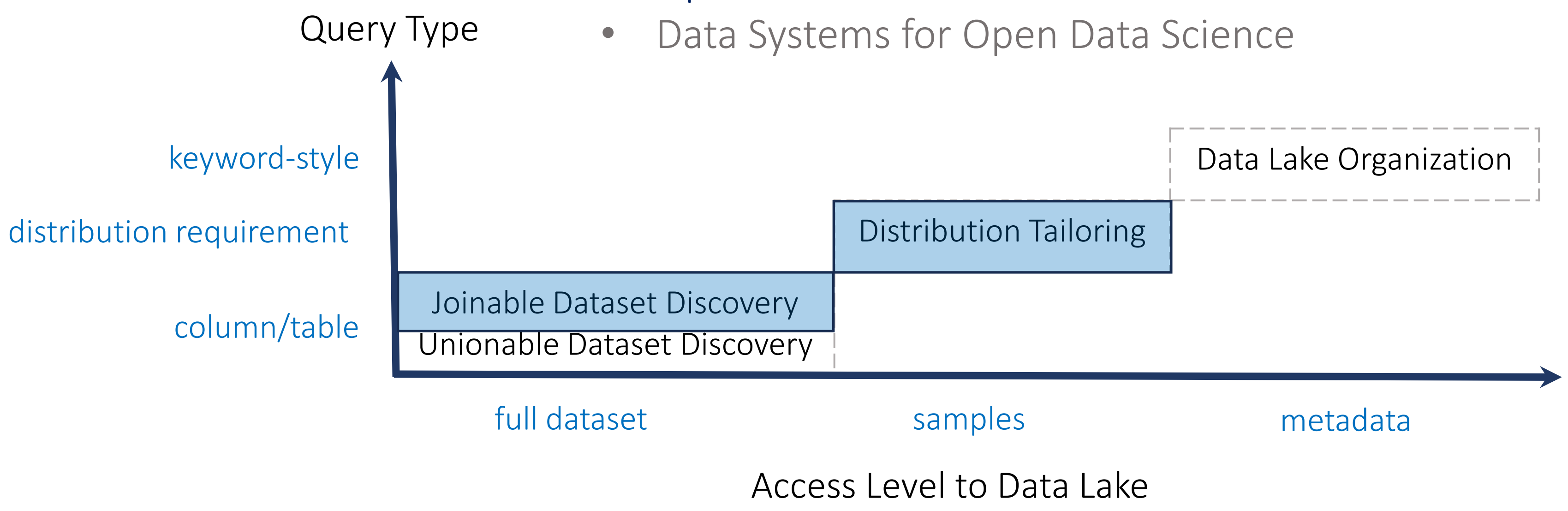
- Requirement. Collect at least one tuple of each group.
- Expected cost of getting one sample of G_1 from S_1 is *100/20*=5 and from S_2 is *100/5=20*.
  - Best source for getting G_1 is S_1. Similarly, best source for G_2 is S_2.
- G_1 is minority in its best source (S_1 has *20%* of G_1, S_2 only has *95%* of G_2) → Pick G_1
  - Piggybacking: as we are sampling for G_1, we can fulfill the count requirements of G_2 with no cost.
- Proof by contradiction.

sample from S_2 until got one tuple of G_2

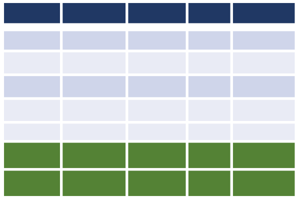
Need to sample another source to get G_1

OUTLOOK

- Open Data for AI
- Open Data for Users
- Data Systems for Open Data Science



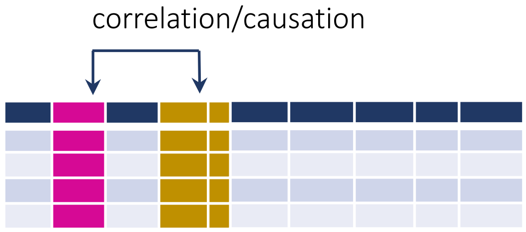
OPEN DATA AND AI



- Sample discovery
  - Discovering novel samples by unionable dataset discovery [NZPM, VLDB’18]
- Feature discovery
  - Pushing down feature selection measures into join discovery
  - More interesting (and rare) relationships: causal dataset discovery
- LLMs and dataset discovery
  - Dataset understanding and training data generation
- Dataset discovery for LLMs
  - Semantic/query-based nearest neighbor search indexes for vector DBs

CAUSAL DATASET DISCOVERY

- Consider tables Q and T. Candidate columns $X \in Q$ and $Y \in T$ have a correlation link over $Q \bowtie_{K=K'} T$, if their correlation after the join is higher than a threshold θ.
- A correlation link can potentially be
  - X causes Y over join $Q \bowtie_{K=K'} T$
  - Y causes X over join $Q \bowtie_{K=K'} T$
  - No causal relation exists between X and Y



- Problem. Given a query table Q and a data lake of tables L, find all tables $C \in L$ such that Q and C have at least one causal link over their join.
- Challenges. Sparsity of causal relation and limited training data for fine-tuning
- Results.
  - Role prompting + CoT + Socratic prompting + fine tuning performs the best.
  - LLMs have potential to learn and generalize causality.

Causal Dataset Discovery with Large Language Models. J. Liu, S. Sun, F. Nargesian HILDA@SIGMOD, 2024.

OPEN DATA AND USERS

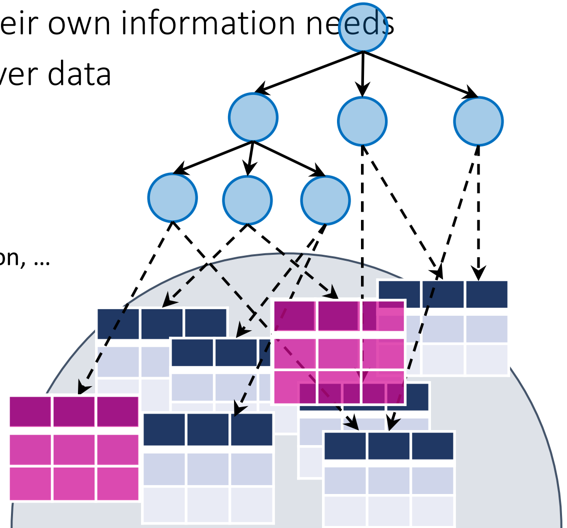
- Metadata, search, query writing
  - Metadata standardization and taxonomy induction in domains such as social sciences
- User interfaces for search
  - Constructing abstract structures for navigation and exploration [NZGPM, SIGMOD’20&TKDE’24]
  - Assisting users with debugging results and consequently their own information needs
  - Dealing with heterogenous data by creating abstractions over data
- Discovery with private queries on open data

Health
- Water
- Food Resilience
  - Food Safety
  - Food Production, …

Energy

…

Climate



ACKNOWLEDGEMENT

- Abolfazl Asudeh (UIC)
- Nikolaus Augsten (U. Salzburg)
- Matthias Boehm (TU Berlin)
- H. V. Jagadish (U. Michigan)
- Renée J. Miller (NEU)
- Felix Naumann (HPI)
- Ken PU (UOIT)
- Divesh Srivastava (AT&T)
- Eric Zhu (MSR)
- Leon Bornemann (HPI, PhD student)
- Jiwon Chang (UR, PhD student)
- Tianji Cong (U. Michigan, PhD student)
- Yurong Liu (UR, undergrad)
- Pranay Mundra (UR, MSc.)
- Aidan Sciortino (UR, undergrad)
- Zhengbin Tao (UR, PhD student)
- Draco Xu (UR, undergrad)
- Mengqi Zhang (UR, PhD student)
- Jianhao Zhang (UR, MSc.)

Code and data: github.com/DataIntelligenceCrew

THANKS.