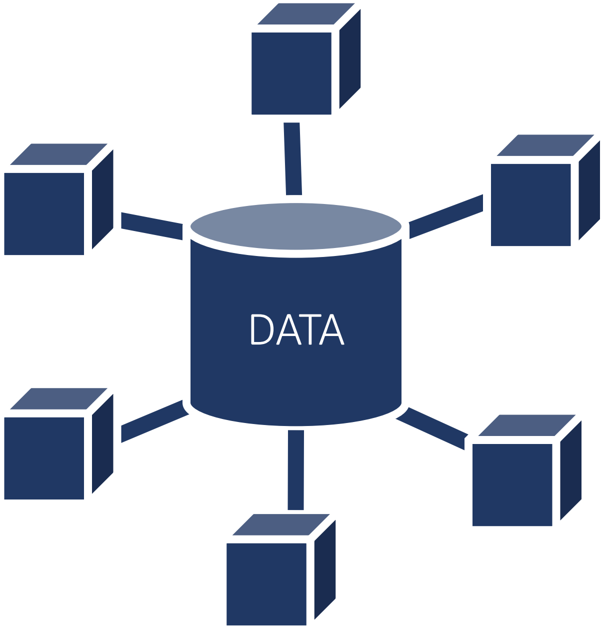
DATA LAKES: DISCOVERY AND DEBIASING

Fatemeh Nargesian, University of Rochester

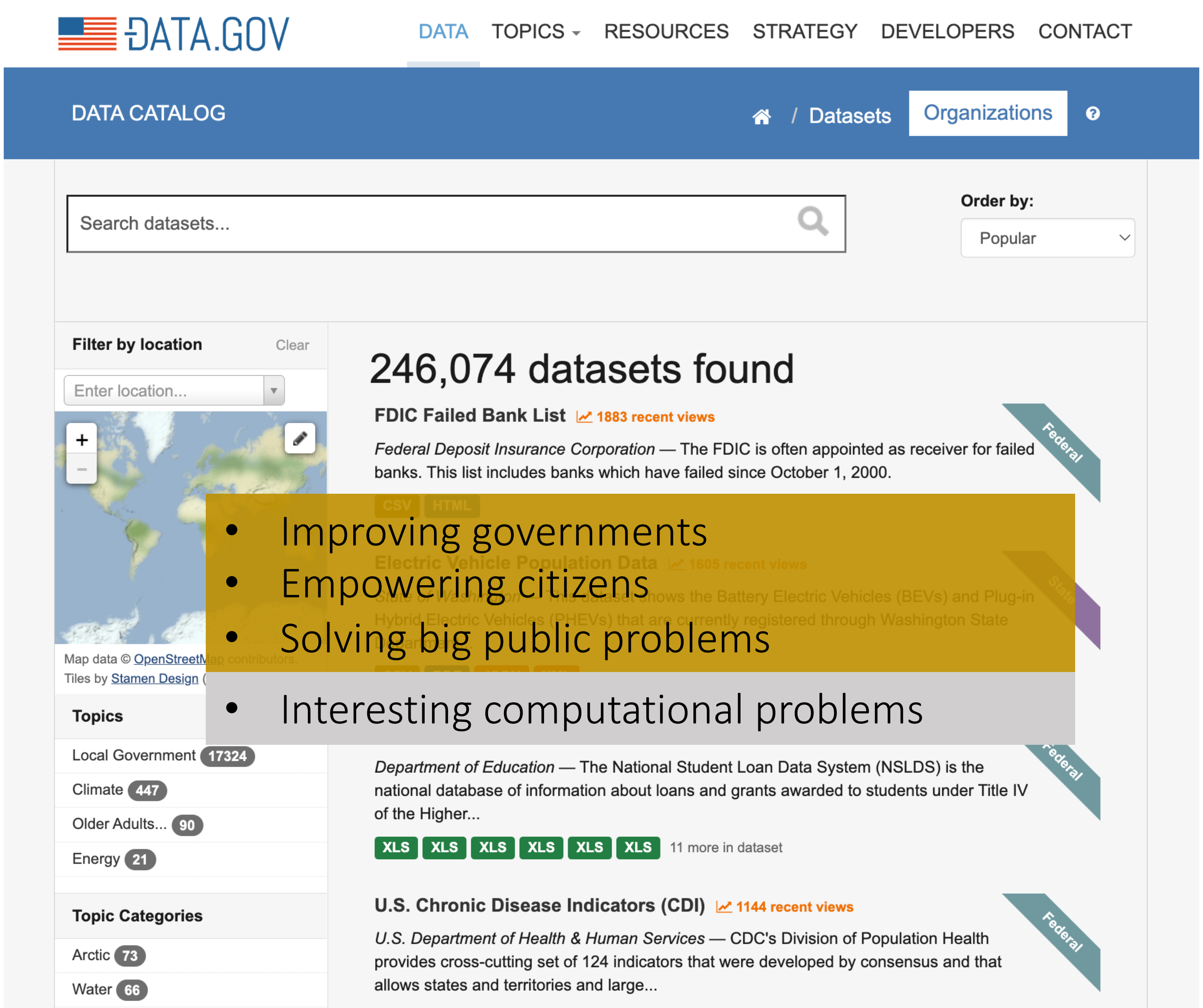
VLDB Summer School 2023 – Cluj-Napoca



- AI is ubiquitous.
- Data-centric AI: focus from big data to good data.
- Open data repositories and data markets have become prevalent.

Data repositories as first-class citizens.

- Sources: open governments, web pages, enterprises, and data markets
- Large number of datasets
- Disconnected and heterogeneous datasets
- Topics vary



- Improving governments
- Empowering citizens
- Solving big public problems
- Interesting computational problems

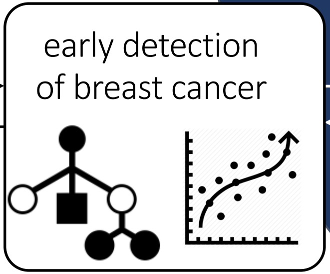
Goal: query answering and dataset construction:

- Distribution and representativeness: model fairness and accuracy
- Efficient, scalable, cost-effective solutions

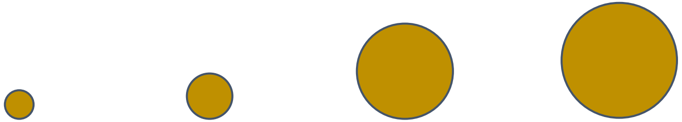
1K breast cancer data in Chicago with at least 30% label= positive, and at least 20% African American patients

train

test



50,820,165



3,562	35,675	51,363	247,074	$$$$$	
Romania	Canada	UK	USA	Data Marketplaces	WDC Web Table 2015 (English Relational Subset)

ABOUT ME

- Assistant Professor of CS, University of Rochester
  - Research: data for AI and scientific time-series management
- Education
  - Undergrad in computer engineering and MSc. in AI, Tehran, Iran
  - PhD -> MSc. in CS, University of Ottawa
  - PhD in CS, University of Toronto
    - Dataset discovery and integration; autoML
- Worked at clinical informatics research group of McGill University; IBM research internships

LOGISTICS

- Many additional references in the slides
- Questions any time during the talk
- The material based on two tutorials:

Data Lake Management: Challenges and Opportunities,
F. Nargesian, E. Zhu+, VLDB, 2019.

Responsible Data Integration: Next-generation Challenges,
F. Nargesian, A. Asudeh, H. V. Jagadish, SIGMOD 2022 and WSDM 2023.

OUTLINE

DATASET DISCOVERY:
Syntactic and Semantic Join Search,
Feature and Slice Discovery

QUERY ANSWERING:
Random Sampling
over Union of Joins

FAIRNESS-AWARE DATA ACQUISITION:
Data Distribution Tailoring

Dataset Discovery

DATA.GOV

DATA TOPICS RESOURCES STRATEGY DEVELOPERS CONTACT

DATA CATALOG

Datasets / Organizations

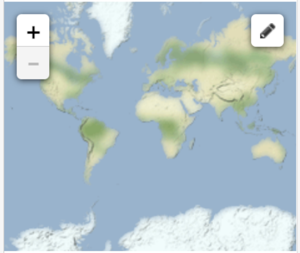
Search datasets... keyword search

Order by:

Data Lake Management: Challenges and Opportunities
F. Nargesian, E. Zhu+, VLDB, 2019.

Filter by location Clear

Enter location...



Map data © OpenStreetMap contributors.
Tiles by Stamen Design (CC BY 3.0)

Topics

- Local Government 17324
- Climate 447
- Older Adults... 90
- Energy 21

Topic Categories

- Arctic 73
- Water 66

246,074 datasets found

FDIC Failed Bank List 1883 recent views

Federal

Federal Deposit Insurance Corporation — The FDIC is often appointed as receiver for failed banks. This list includes banks which have failed since October 1, 2000.

CSV HTML

Electric Vehicle Population Data 1605 recent views

State

State of Washington — This dataset shows the Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) that are currently registered through Washington State Department...

CSV RDF JSON XML

National Student Loan Data System 1175 recent views

Federal

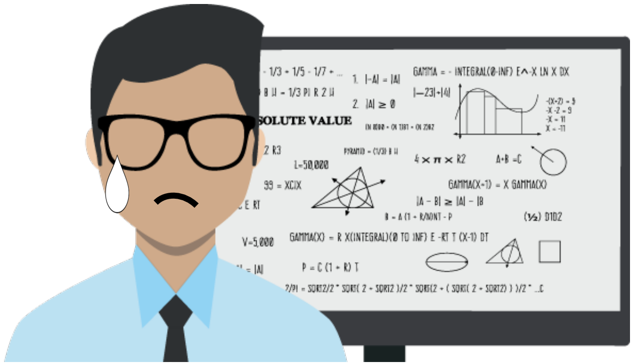
Department of Education — The National Student Loan Data System (NSLDS) is the national database of information about loans and grants awarded to students under Title IV of the Higher...

XLS XLS XLS XLS XLS XLS 11 more in dataset

U.S. Chronic Disease Indicators (CDI) 1144 recent views

Federal

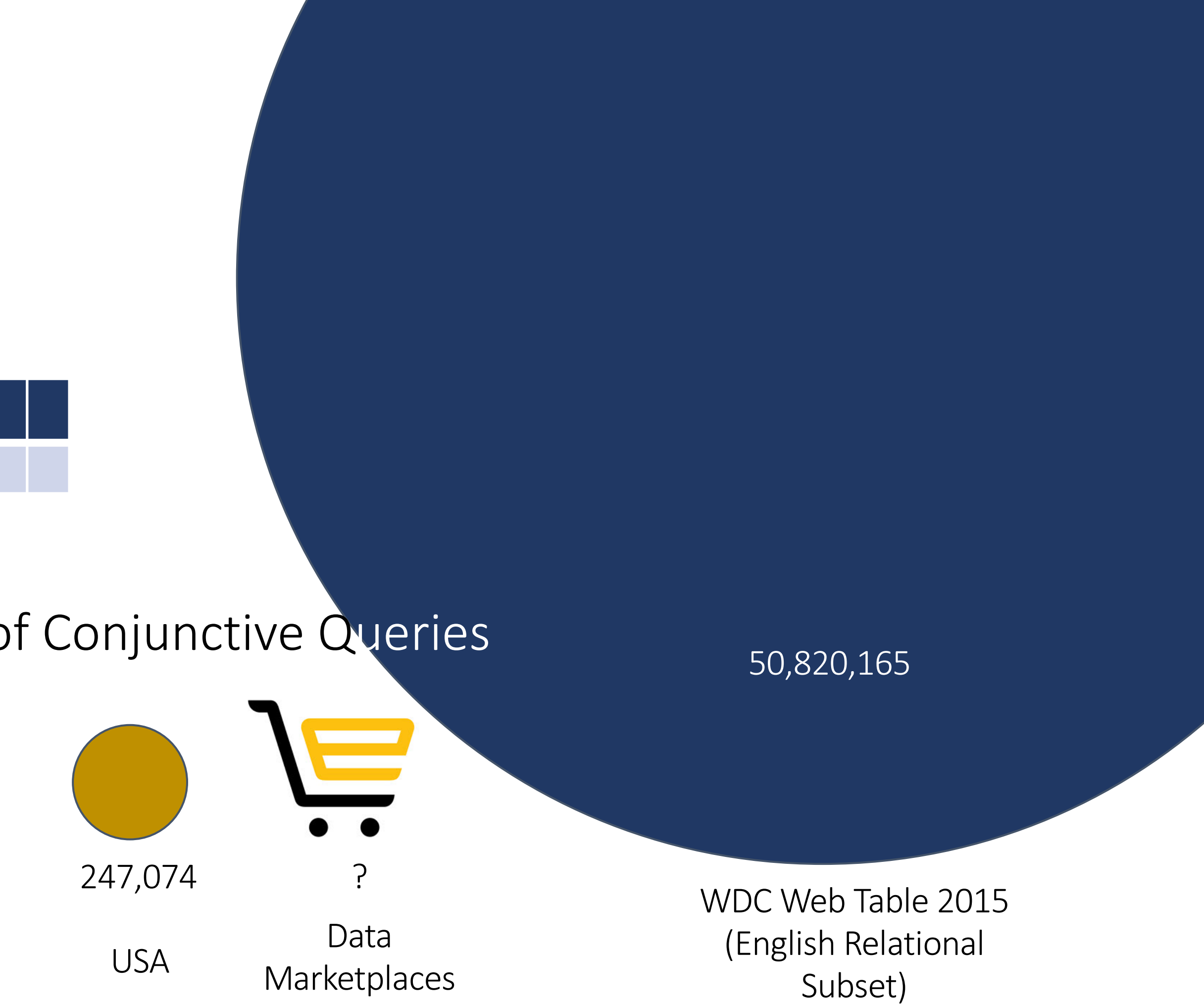
U.S. Department of Health & Human Services — CDC's Division of Population Health provides cross-cutting set of 124 indicators that were developed by consensus and that allows states and territories and large...



Geo	Date	Fuel Type	Pop	Avg. Age	
…	…	…	…	…	

```
SELECT ??
FROM ?? JOIN ??
ON ?? = ??
UNION
SELECT ??
FROM ?? JOIN ??
UNION
…
```

Union of Conjunctive Queries



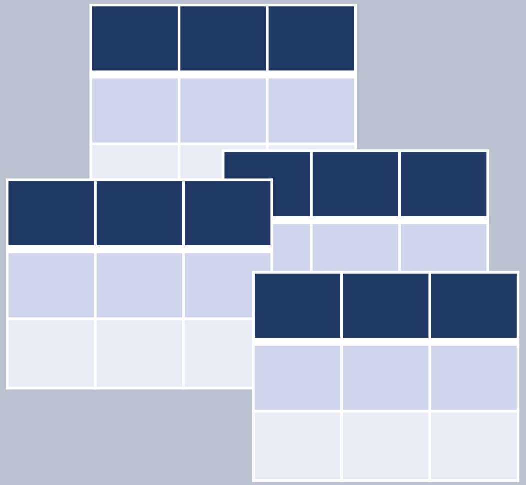
Search by Join

emission

Geo	Date	Fuel	ktCO2	Sector	…
Barnet	2015	electricity	130	Domestic	
City of London	2015	diesel	200	Transport	
Camden	2014	coal	125	Domestic	
…	…	…	…	…	

query column

citydata



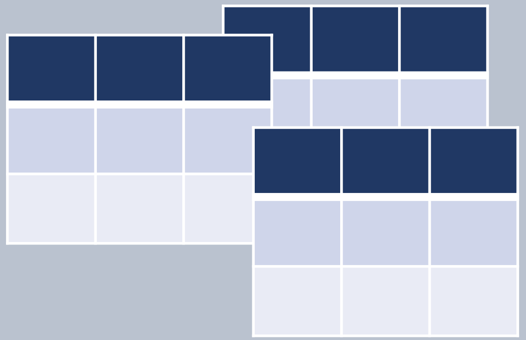
```
SELECT ??
FROM emission e JOIN ??
ON e.Geo = ??
```

emission

Geo	Date	Fuel	ktCO2	Sector	…
Barnet	2015	electricity	130	Domestic	
City of London	2015	diesel	200	Transport	
Camden	2014	coal	125	Domestic	
…	…	…	…	…	

citydata

Area	Pop	Avg_age	F.Unemp	Unemp	…
City of London	8800	43.2	-	-	
Camden	242500	36.4	62.9	4	
Cambridge	389600	37.3	66	8.5	
…					



```
SELECT *
FROM emission e JOIN citydata d
ON e.Geo = d.Area
```

Geo	Date	Fuel	ktCO2	Sector	Pop	Avg_age	F.Unemp	Unemp	…
Camden	2014	Coal	125	Domestic	142500	36.4	-	-	
City of London	2015	diesel	200	Transport	242500	43.2	62.9	4	
Barnet	…	…	…	…	NULL	NULL	NULL	NULL	
…	…	…	…	…					

SYNTACTIC JOIN DISCOVERY

LSH Ensemble: Internet-Scale Domain Search,
E. Zhu, F. Nargesian, K. Pu, R. J. Miller, VLDB, 2016.
JOSIE: Overlap Set Similarity Search for Finding Joinable Tables in Data Lakes, E. Zhu, D. Dong, F. Nargesian, PU, Miller, SIGMOD 2019.

Joinability Measure

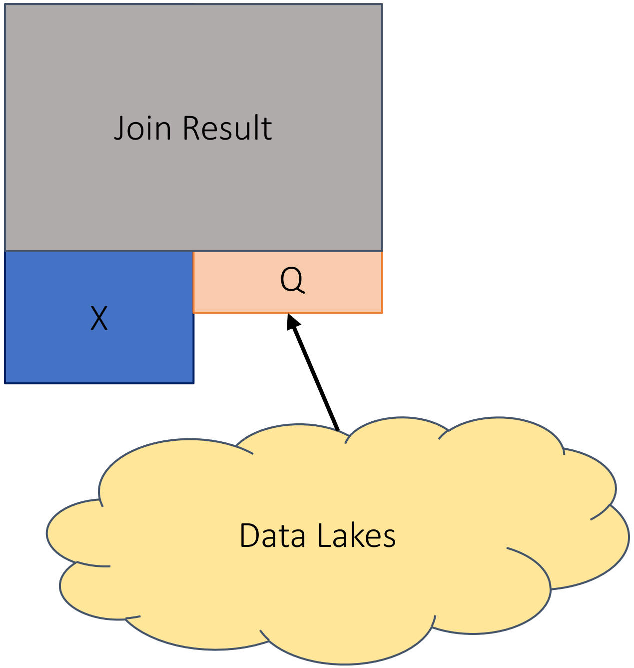
- Columns as sets

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

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

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

- *Columns as multisets*
- Related work [Bessa+POD'23, Santos+ICDE'22, Santos+SIGMOD'21, Fernandez+ICDE'19]



JACCARD VS. CONTAINMENT

- Suppose there are the following two columns in the repository

  Provinces = {Alberta, Ontario, Manitoba}

  Locations = {Illinois, Chicago, New York, Nova Scotia, Halifax, California, San Francisco, Seattle, Washington, Ontario, Toronto}

- Consider the following query columns

  Q = {Ontario, Toronto}

- Top-1 joinable columns based on Jaccard? Top-1 joinable columns based on containment?

  Jaccard(Q,P) = 1/4, Containment(Q,P)=1/2

  Jaccard(Q,L) = 2/11, Containment(Q,P)=1

  Jaccard is biased towards smaller columns

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

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

DATASET DISCOVERY

- Threshold-based search: Given a query Q and a joinability measure J, find columns X s.t. J(Q,X) >= t*.
- Top-k search: Given a query Q and a joinability measure J, find k columns X s.t. J(Q,X) >= t*.

THRESHOLD-BASED CONTAINMENT SEARCH

- Problem. Given a query Q and containment threshold t*, find columns X s.t. containment(Q,X) >= t*.

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

Query colum

Q = {Bost

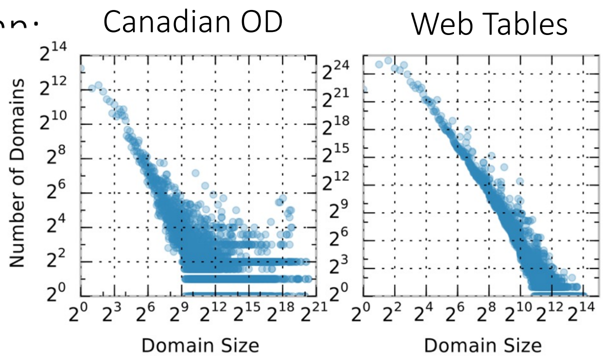
Columns in

Geo = {Ed

Locations , Seattle, NYC}

...

Search with ot **Geo**.



- Existing technique for containment search results in low recall for skewed column size distributions [SrivastavaLi2015].

LSH ENSEMBLE

- Deals with data volume and skew!
- First phase: columns are partitioned based on the distribution of column cardinality.
- Second phase: construct a MinHash LSH index for each partition and parallel search
- Accurate over columns whose sizes are skewed (e.g., power-law dist.)

MINHASHING FOR JACCARD NEAREST NEIGHBOR SEARCH

- MinHash LSH [Broder97, Indyk98]: an index for R-near neighbor based on Jaccard.

MINHASHING

- Key idea: "hash" each column C to a small *signature* $h(C)$, such that:
  - (1) $h(C)$ is small enough that the signature fits in RAM
  - (2) $sim(C_1, C_2)$ is the same as the "similarity" of signatures $h(C_1)$ and $h(C_2)$

- Goal: Find a hash function $h(\cdot)$ such that:
  - If $sim(C_1,C_2)$ is high, then with high prob. $h(C_1) = h(C_2)$
  - If $sim(C_1,C_2)$ is low, then with high prob. $h(C_1) \neq h(C_2)$

- Hash cols into buckets. Expect that "most" pairs of near duplicate cols hash into the same bucket!

MINHASHING

- Goal: Find a hash function $h(\cdot)$ such that:
  - if $sim(C_1,C_2)$ is high, then with high prob. $h(C_1) = h(C_2)$
  - if $sim(C_1,C_2)$ is low, then with high prob. $h(C_1) \neq h(C_2)$

- Clearly, the hash function depends on the similarity metric:
  - Not all similarity metrics have a suitable hash function
- There is a suitable hash function for the Jaccard similarity: It is called Min-Hashing

MINHASHING

- Imagine the rows of the boolean matrix permuted under random permutation π
- Define a "hash" function $h_\pi(C)$ = the index of the first (in the permuted order π) row in which column C has value 1:

$$h_\pi(C) = min_\pi\, \pi(C)$$

- Use several (e.g., 100) independent hash functions (that is, permutations) to create a signature of a column

MINHASHING - EXAMPLE

Permutation π Input matrix (Shingles x Documents)

1	0	1	0
1	0	0	1
0	1	0	1
0	1	0	1
0	1	0	1
1	0	1	0
1	0	1	0

MINHASHING - EXAMPLE

Permutation π

2
3
7
6
1
5
4

Input matrix (Shingles x Documents)

1	0	1	0
1	0	0	1
0	1	0	1
0	1	0	1
0	1	0	1
1	0	1	0
1	0	1	0

Minhashing - Example

Permutation π

2
3
7
6
1
5
4

Input matrix (Shingles x Documents)

1	0	1	0
1	0	0	1
0	1	0	1
0	1	0	1
0	1	0	1
1	0	1	0
1	0	1	0

Signature matrix *M*

2	1	2	1

MINHASHING - EXAMPLE

2nd element of the permutation is the first to map to a 1

Permutation π

2
3
7
6
1
5
4

Input matrix (Shingles x Documents)

1	0	1	0
1	0	0	1
0	1	0	1
0	1	0	1
0	1	0	1
1	0	1	0
1	0	1	0

Signature matrix *M*

2	1	2	1

MINHASHING - EXAMPLE

2nd element of the permutation is the first to map to a 1

Permutation π

2	4
3	2
7	1
6	3
1	6
5	7
4	5

Input matrix (Shingles x Documents)

1	0	1	0
1	0	0	1
0	1	0	1
0	1	0	1
0	1	0	1
1	0	1	0
1	0	1	0

Signature matrix *M*

2	1	2	1
2	1	4	1

MINHASHING - EXAMPLE

2^{nd} element of the permutation is the first to map to a 1

Permutation π

2	4
3	2
7	1
6	3
1	6
5	7
4	5

Input matrix (Shingles x Documents)

1	0	1	0
1	0	0	1
0	1	0	1
0	1	0	1
0	1	0	1
1	0	1	0
1	0	1	0

Signature matrix *M*

2	1	2	1
2	1	4	1

4^{th} element of the permutation is the first to map to a 1

MinHashing - Example

2nd element of the permutation is the first to map to a 1

Permutation π

2	4	3
3	2	4
7	1	7
6	3	2
1	6	6
5	7	1
4	5	5

Input matrix (Shingles x Documents)

1	0	1	0
1	0	0	1
0	1	0	1
0	1	0	1
0	1	0	1
1	0	1	0
1	0	1	0

Signature matrix *M*

2	1	2	1
2	1	4	1
1	2	1	2

4th element of the permutation is the first to map to a 1

MinHashing Property

- Choose a random permutation π
- <u>Claim:</u> $\Pr[h_\pi(C_1) = h_\pi(C_2)] = sim(C_1, C_2)$
- Why?
  - Let X be a col (set of shingles), $y \in X$ is a shingle
  - Then: $\Pr[\pi(y) = \min(\pi(X))] = 1/|X|$
    - It is equally likely that any $y \in X$ is mapped to the *min* element
  - Let *y* be s.t. $\pi(y) = \min(\pi(C_1 \cup C_2))$
  - Then either: $\pi(y) = \min(\pi(C_1))$ if $y \in C_1$, or
    $\pi(y) = \min(\pi(C_2))$ if $y \in C_2$
  - So the prob. that both are true is the prob. $y \in C_1 \cap C_2$
  - $\Pr[\min(\pi(C_1))=\min(\pi(C_2))]=|C_1 \cap C_2|/|C_1 \cup C_2|= sim(C_1, C_2)$

0	0
0	0
1	**1**
0	0
0	1
1	0

Four Types of Rows

- Given cols C_1 and C_2, rows may be classified as:

	C_1	C_2
A	1	1
B	1	0
C	0	1
D	0	0

  - **a** = # rows of type A, etc.
- Note: $sim(C_1, C_2) = a/(a + b + c)$
- Then: $\Pr[h(C_1) = h(C_2)] = Sim(C_1, C_2)$
  - Look down the cols C_1 and C_2 until we see a 1
  - If it's a type-*A* row, then $h(C_1) = h(C_2)$
    If a type-*B* or type-*C* row, then not

Similarity of Signatures

- We know: $\Pr[h_\pi(C_1) = h_\pi(C_2)] = sim(C_1, C_2)$
- Now generalize to multiple hash functions

- The *similarity of two signatures* is the fraction of the hash functions in which they agree

- Note: Because of the Min-Hash property, the similarity of columns is the same as the expected similarity of their signatures
  - It can be shown that $h_\pi(C1) = h_\pi(C2)$ is an unbiased estimator of sim(C1, C2)
    - An estimator is unbiased if its expected value is equal to the true value of the parameter.

MINHASHING - EXAMPLE

Permutation π

2	4	3
3	2	4
7	1	7
6	3	2
1	6	6
5	7	1
4	5	5

Input matrix (Shingles x Documents)

1	0	1	0
1	0	0	1
0	1	0	1
0	1	0	1
0	1	0	1
1	0	1	0
1	0	1	0

Signature matrix *M*

2	1	2	1
2	1	4	1
1	2	1	2

Similarities:

	1-3	2-4	1-2	3-4
Col/Col	0.75	0.75	0	0
Sig/Sig	0.67	1.00	0	0

MINHASHING - EXAMPLE

- Pick K=100 random permutations of the rows
- Think of *sig*(C) as a column vector
- *sig*(C)[i] = according to the *i*-th permutation, the index of the first row that has a 1 in column *C*

  $$sig(C)[i] = \min(\pi_i(C))$$

- Note: The sketch (signature) of document *C* is small $\sim 500\ K$ bytes!

- We achieved our goal! We “compressed” long bit vectors into short signatures

MINHASHING FOR JACCARD NEAREST NEIGHBOR SEARCH

- MinHash LSH [Broder97, Indyk98]: an index for R-near neighbor based on Jaccard.
- Each column is represented with one or more minhash values.

hash func. h(x)

X_1

X_2

.

.

.

X_n

minhash of set
12, 10
12, 2, 4
1, 12, 9, 7, 5, 90

$\Pr[\text{minhash}(X_i) = \text{minhash}(X_j)]$
$= \text{Jaccard}(X_i, X_j)$

MINHASHING FOR JACCARD NEAREST NEIGHBOR SEARCH

- MinHash LSH [Broder97, Indyk98]: an index for R-near neighbor based on Jaccard.
- Each column is represented with one or more minhash values.

X_1

X_2

.
.
.

X_n

signature: k minhash

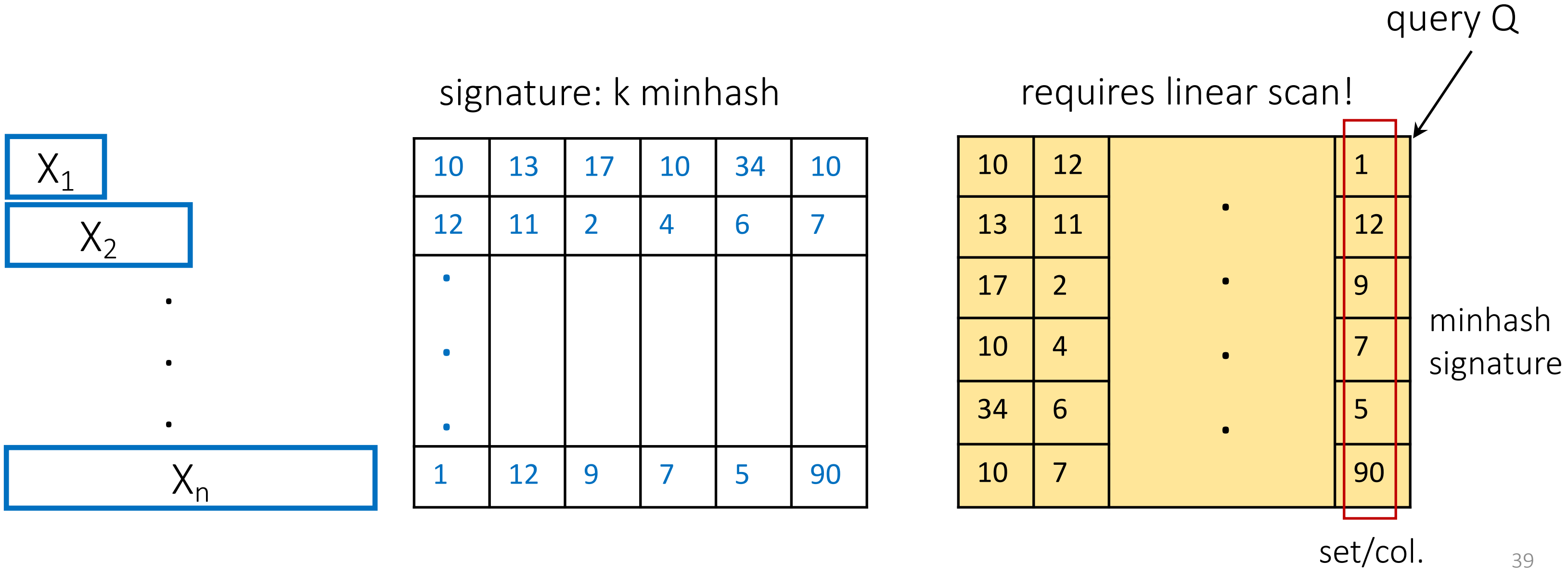
10	13	17	10	34	10
12	11	2	4	6	7
1	12	9	7	5	90

$\Pr[\text{minhash}(X_i) = \text{minhash}(X_j)]$
$= \text{Jaccard}(X_i, X_j)$

$\text{Jaccard}(X_i, X_j) \sim$
colliding minhash / hash funcs.

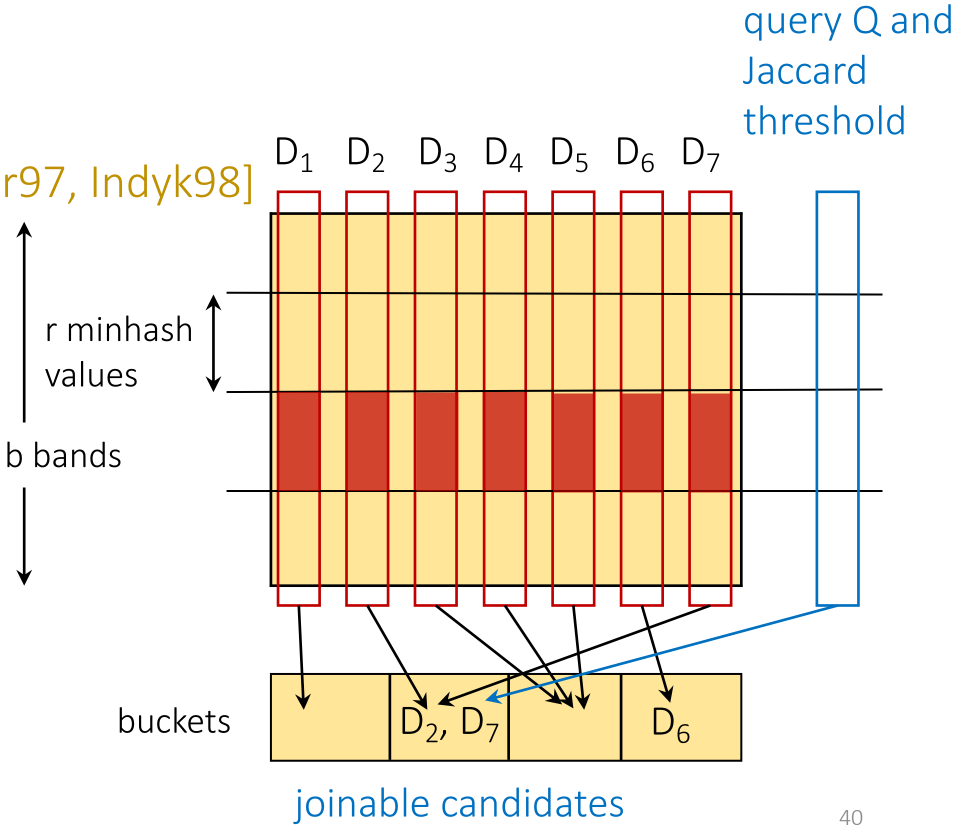
SKETCHING

- MinHash LSH [Broder97, Indyk98]: an index for R-near neighbor based on Jaccard.
- Each column is represented with one or more minhash values.



LOCALITY SENSITIVE HASHING (LSH)

- If we were to use Jaccard
- Similar sets: similar signatures [Broder97, Indyk98]
- Hash bands into buckets
- Columns hashed to same bands are *potential* candidates for joinable cols.
- Post-process candidates to find cols. with similarity > threshold



Threshold-based Containment Search

- Problem. Given a query Q and containment threshold t*, find columns X s.t. containment(Q,X) >= t*.

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

INDEX A PARTITION

largest col. size in partition

$$s^* = \frac{t^*}{\frac{|u|}{|Q|} + 1 - t^*}$$

containment threshold t*

Jaccard threshold s*:

LSH index

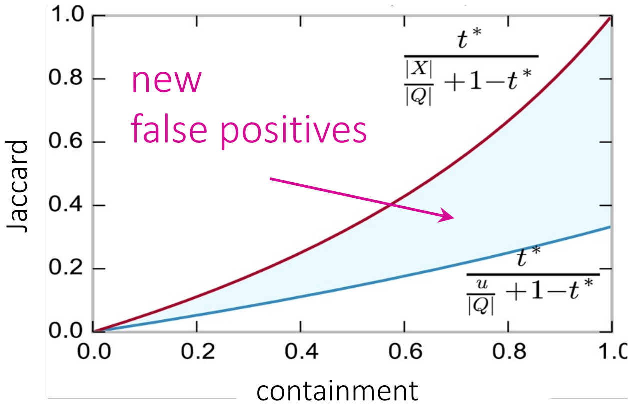
Partition [l, u)

query Q

cols. X with Jaccard(Q,X) >= s*

remove old and new false positives

cols. X with Containment(Q,X) >= t*



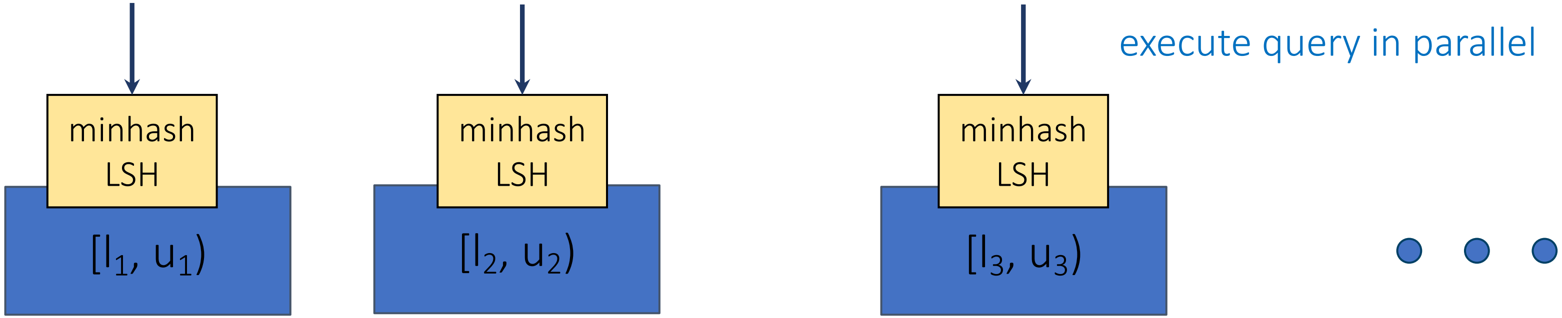
$T_{Containment}$ =
$T_{Jaccard}$ + Θ(correct result) + Θ(N^{FP})

time to process false positives

PARTITIONING SCHEME

each partition has its own threshold

execute query in parallel



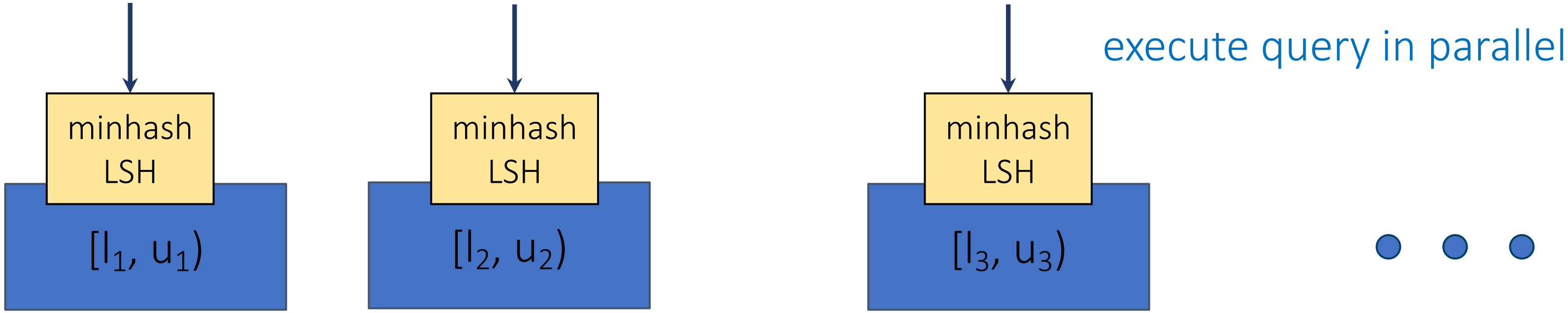
- Query cost is determined by the partition with the highest # false positives.

$$\Pi^* = argmin\,(max_{1<i<n} M_i)$$

false positives in partition i

- Data partitioning as an optimization problem.
  - The partitioning in which all M_i's are the same.

PARTITIONING SCHEME



- Query cost is determined by the partition with the most # false positives.

$$\Pi^* = argmin\,(max_{1<i<n} M_i)$$

false positives in partition i

of columns in a partition

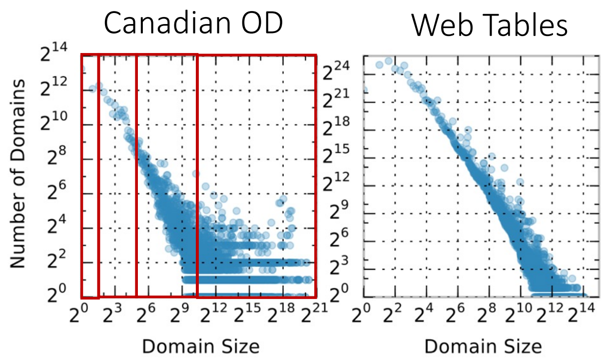
$$M_i \leq N_{l_i,u_i} \cdot \frac{u_i - l_i + 1}{2u_i}$$

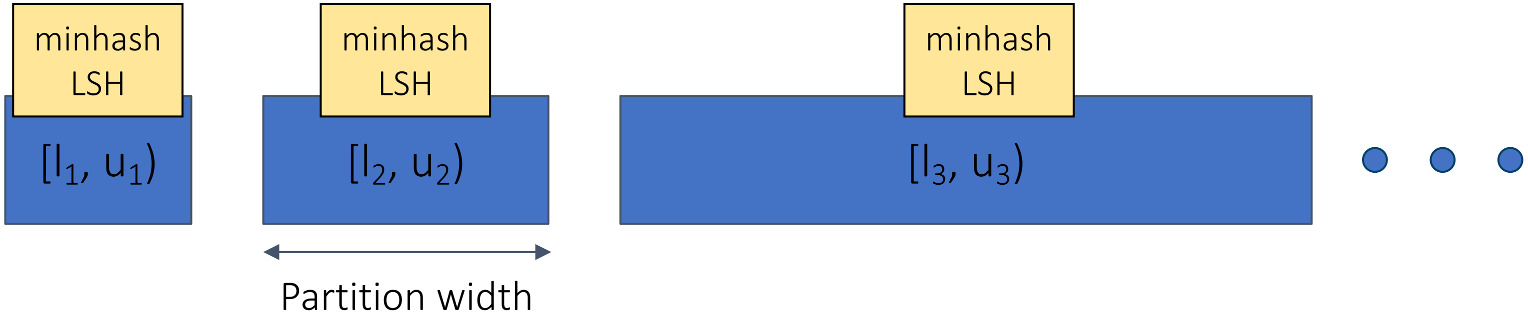
assuming uniform dist. of sizes

- How to choose partition bounds l and u?

OPTIMAL PARTITIONING

- Exists an optimal partitioning for any data distribution.
- For power-law distributions, the optimal partitioning can be approximated using equi-depth.





QUERY PERFORMANCE

- On WDC Web Table: ~263 million columns

Algorithm	Mean Query (sec)	Precision Before Pruning (t*=0.5)
MinHash LSH	45.13	0.27
LSH Ensemble (8)	7.55	0.48
LSH Ensemble (16)	4.26	0.53
LSH Ensemble (32)	3.12	0.58

- Speedup is due to
  - fewer false positive columns to process (higher precision)
  - parallelization

Search on Vectors

- Hierarchical Navigable Small World (HNSW) for vector search

  Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs, Yu A. Malkov and D. A. Yashuin, IEEE Trans. on Pattern Analysis and Machine Intelligence, 2020.

- Practical and efficient index structure for a variety of distance measures

  Billion-scale similarity search with GPUs, J. Johnson et al., IEEE Transactions on Big Data, 2019

Join and Dirty Data

- Containment may become ineffective for joining data in the wild.
- Dirty and semantically diverse data

Geo	Date	Fuel	ktCO2	Sector	…
Barnet	2015	electricity	130	Domestic	
City of London	2015	diesel	200	Transport	
NYC	2014	coal	125	Domestic	
…	…	…	…	…	

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

```
SELECT *
FROM emission e JOIN ? d
ON e.Geo ~ ?
```

SEMANTIC JOIN DISCOVERY

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

SEMANTIC JOINABILITY MEASURE

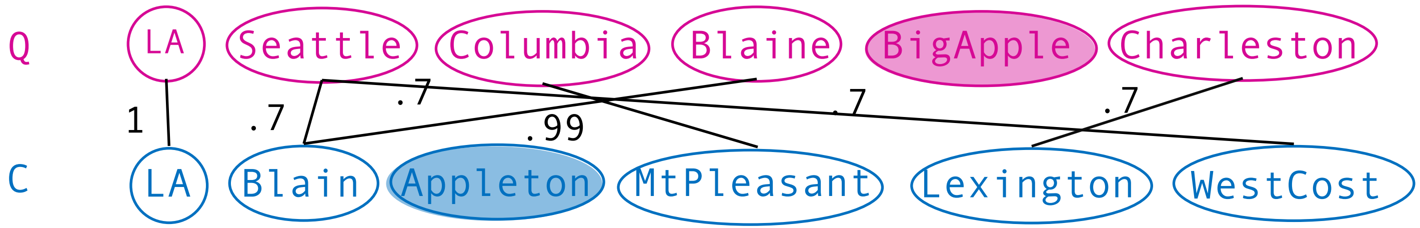
Q
LA
Seattle
Columbia
…

C
LA
Blain
Appleton
…

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

SEMANTIC JOINABILITY MEASURE

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




```
sim(“LA”, “LA”) = 1.0
sim(“Seattle”, “WestCoast”) = 0.7
...
```

CHARACTER-BASED ELEMENT SIMILARITY

```
Q = {LA, Seattle, Columbia, Blaine, BigApple, Charleston}
C1 = {LA, Blain, Appleton, MtPleasant, Lexington,
        WestCoast}
C2 = {LA, Sacramento, Southern, Blain, SC, Minnesota,
        NewYorkCity}
```

3-grams of elements

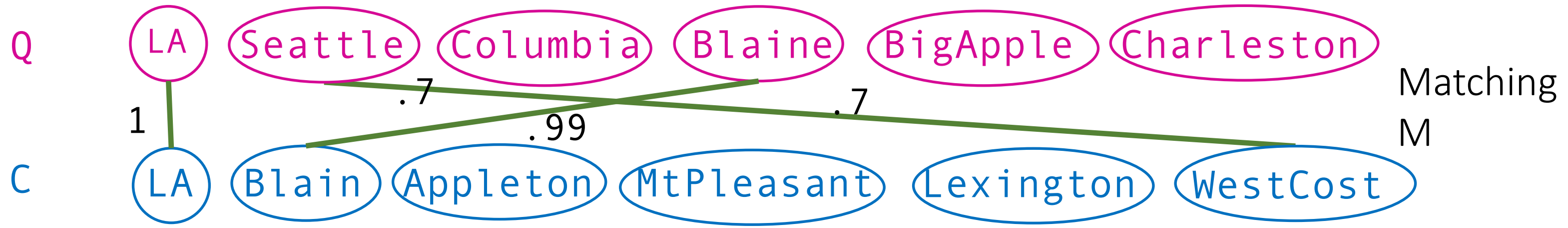
```
Blaine = {bla, lai, ain, ine}
BigApple = {big, iga, gap, app, ppl, ple}
Appleton = {app, ppl, ple, let, eto, ton}
Blain = {bla, lai, ain}
NewYorkCity = {new, ewy, wyo, yor, ork,
                  rkc, kci, cit, ity}
```

Element similarity on 3-grams

```
Jaccard(Blaine, Blain) = 3/4
Jaccard(BigApple,Appleton) = 1/3
Jaccard(BigApple, NewYorkCity) = 0
```

Semantic Joinability Measure

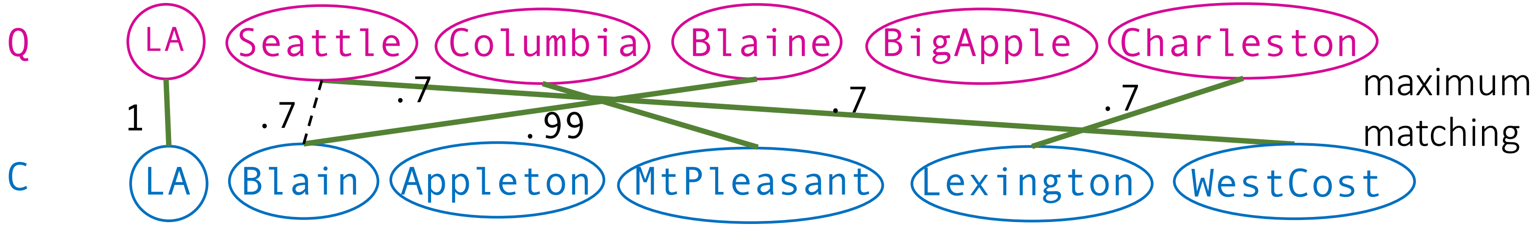
```
Q = {LA, Seattle, Columbia, Blaine, BigApple, Charleston}
C = {LA, Blain, Appleton, MtPleasant, Lexington, WestCoast}
```




```
sim(“LA”, “LA”) = 1.0
sim(“Seattle”, “WestCoast”) = 0.7
...

score(M) = 3.39
```

Q = {LA, Seattle, Columbia, Blaine, BigApple, Charleston}
C$_1$ = {LA, Blain, Appleton, MtPleasant, Lexington, WestCoast}



Semantic Overlap

- Maximum matching of the bipartite graph of Q and C with *sim*$_\alpha$*(., .)* being any symmetric similarity function

$$SO(Q, C) = max_M \sum_{q_i \in Q} sim_\alpha(q_i, M(q_i))$$

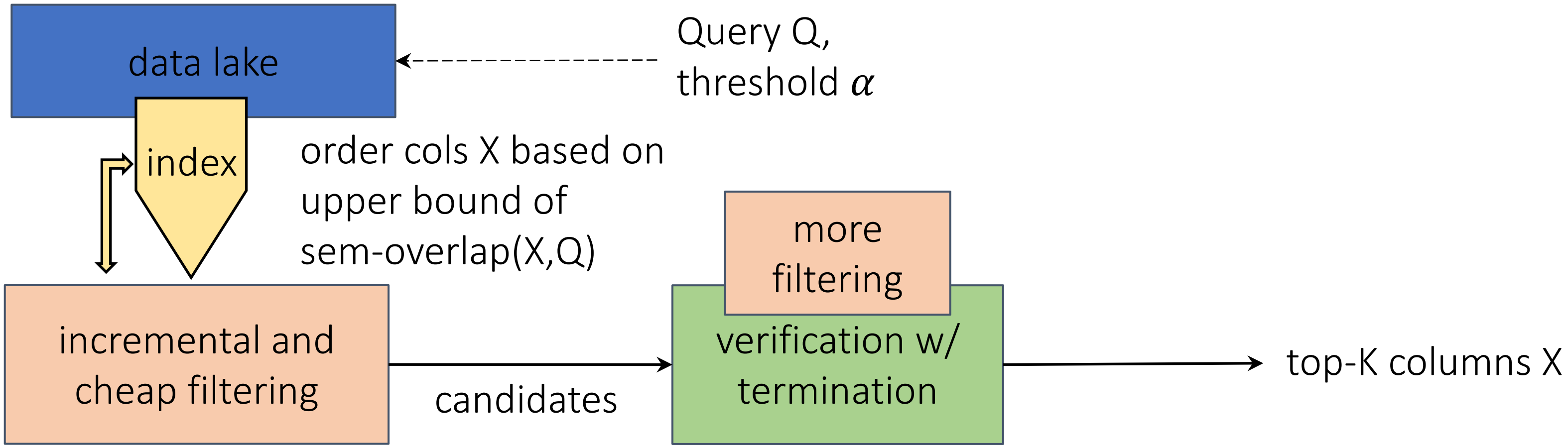
- $|Q \cap C| \leq SO(Q, C)$

Top-k Semantic Overlap Search

- Semantic overlap ~ bipartite graph matching [Kuhan'1995]
- 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 sets and m is the number of sets

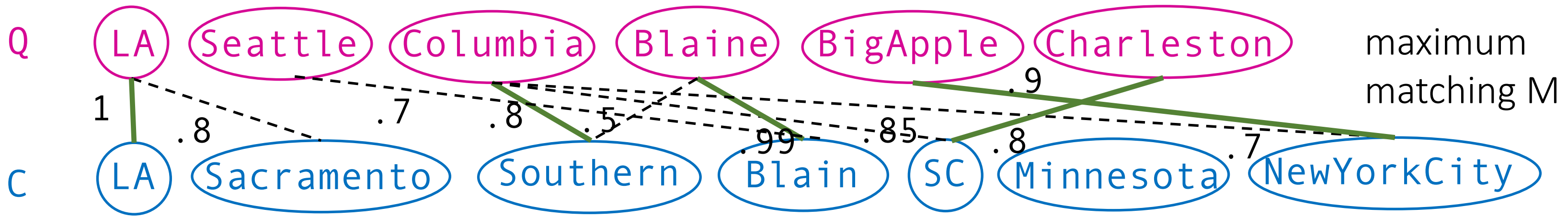
KOIOS: FILTER-VERIFICATION-FILTER

- Provably exact and efficient top-K search algorithm over large data lakes



- refinement of bounds
- a partitioning scheme for efficient filtering

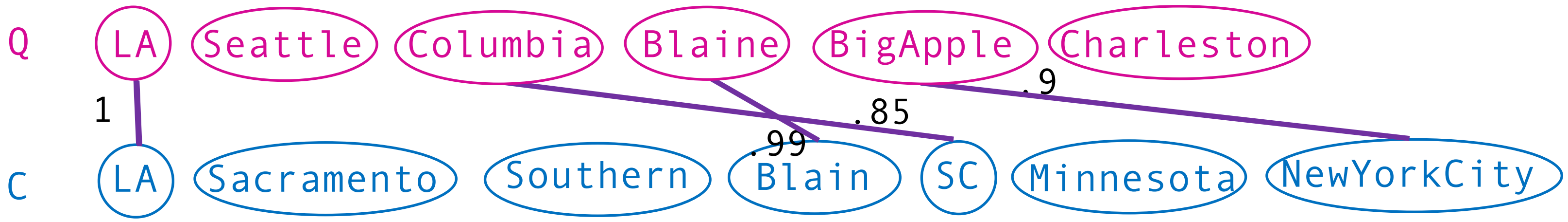
- candidate ordering
- prematurely terminating verification



$$SO(Q,C) = max_M \sum_{q_i \in Q} sim(q_i, M(q_i))$$

SO(Q, C) = 4.49

- How to approximate bipartite matching scores and perform top-K search based on approximations?



$$SO(Q,C) = max_M \sum_{q_i \in Q} sim(q_i, M(q_i))$$

- Upper-bound

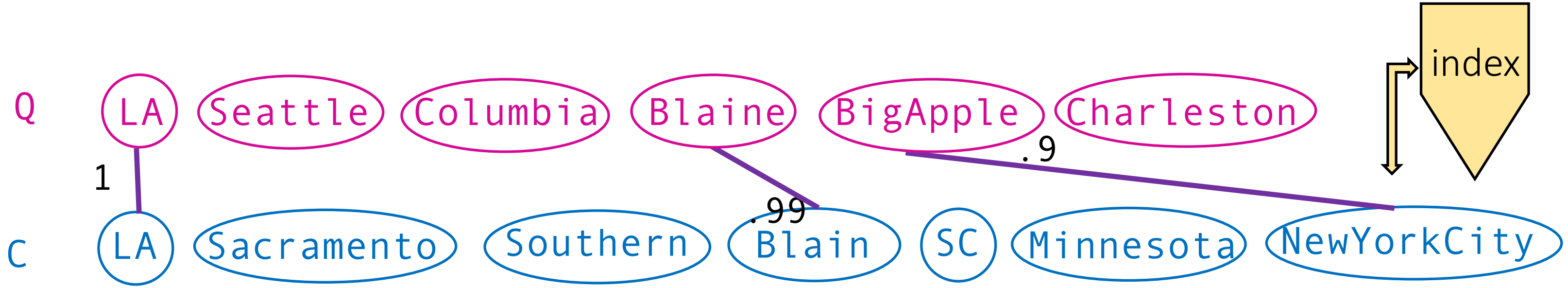
$$UB(C) = |Q| . max\ edge\ weight$$

- Expensive lower-bound

$$LB(C) = score\ of\ a\ greedy\ matching$$

LB(C) = 3.74 < 4.49

INCREMENTAL BOUNDS



- Assume an index returns the next best edges for all sets descendingly: (e, s_{l+1})

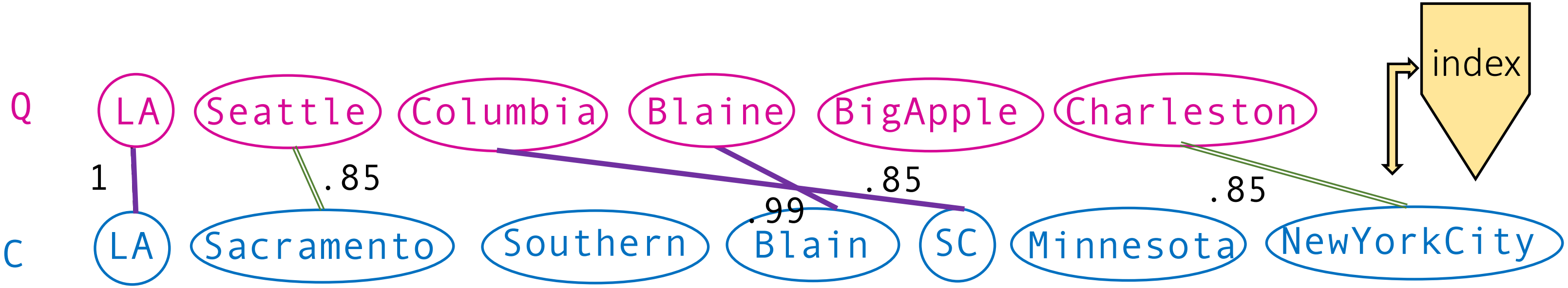
$$iLB(C) = \sum \textit{edge weight in a greedy “partial” matching}$$

$$iLB_{l+1}(C) = iLB_l(C) + s_{l+1}$$

$iLB(C) = 1 + 0.99 < 4.49$

$iLB(C) = 1.99 + 0.9 < 4.49$

INCREMENTAL BOUNDS



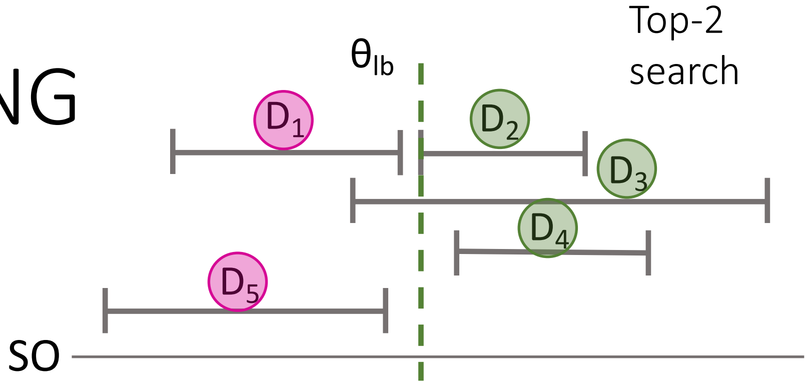
- Assume an index returns the next best edges for all sets descendingly.

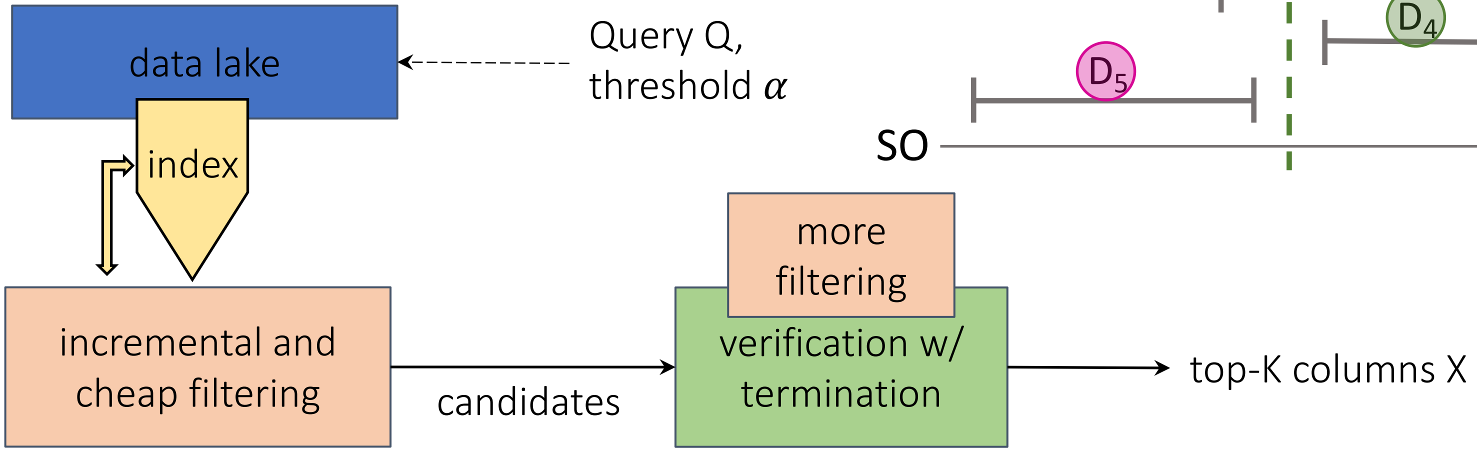
$$iLB(C) = \sum edge\ weight\ in\ a\ greedy\ \text{“}partial\text{”}\ matching$$

$$iUB_{l+1}(C) = m.s_{l+1} + iUB_{l+1}(C), \qquad \mathrm{m} = \min(|\mathrm{Q}| - |\mathrm{M}|, |\mathrm{C}| - |\mathrm{M}|)$$

$$iUB(\mathrm{C}) = 2.84 + 2*0.85 > 4.49$$

FILTERING

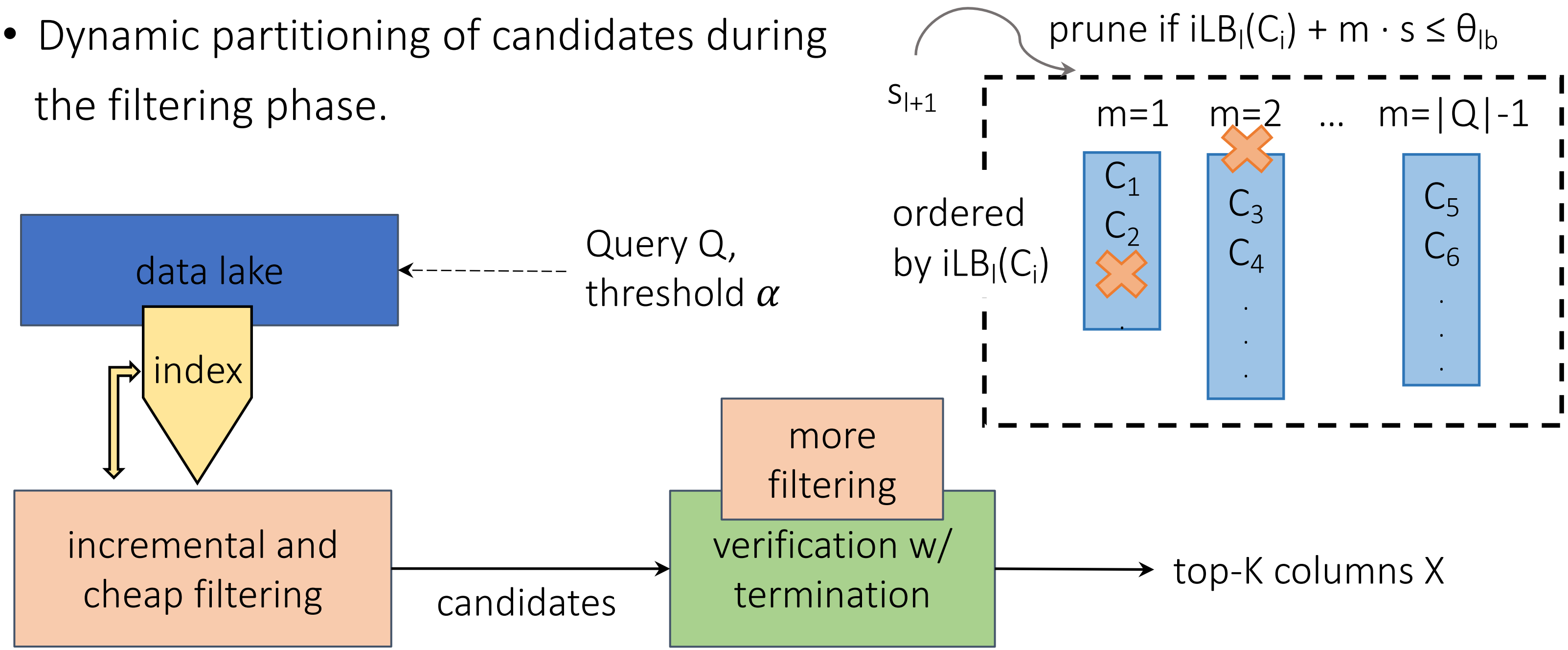




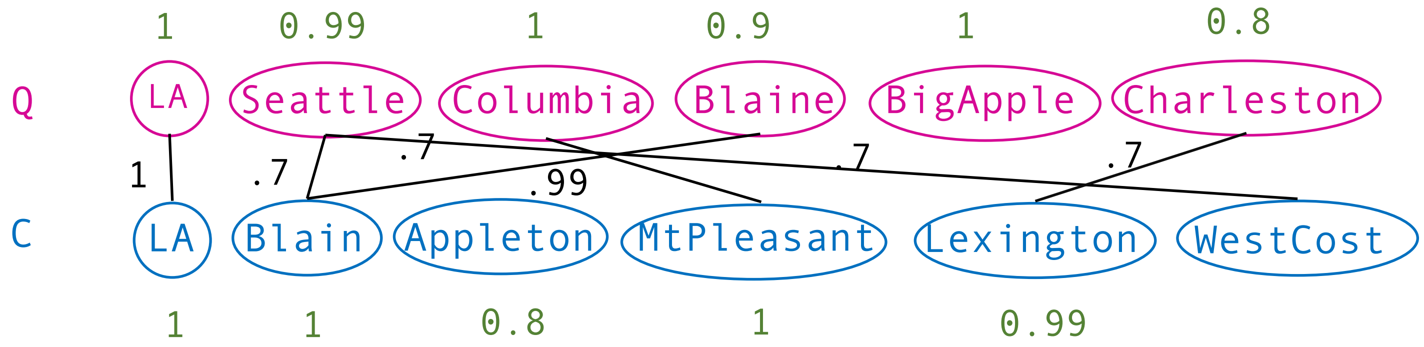
- θ_{lb} : *k*-th largest observed iLB's
- Maintain a running top-K LB-list
- Filter $iUB(C) < \theta_{lb}$
- Excessive updates of bounds

PARTITIONING SCHEME

- Dynamic partitioning of candidates during the filtering phase.



EARLY TERMINATION OF BIPARTITE MATCHING



- Hungarian algorithm assigns and refines a labeling *l: {Q}∪{C}→R* s.t.

$$l(q) + l(c) \geq sim(q, c), \forall q \in Q, c \in C$$

- Results. $SO(C) \leq \sum_{x \in \{Q\} \cup \{C\}} l(x)$
- Terminate matching computing as soon as $\sum_{x \in \{Q\} \cup \{C\}} l(x) \leq \theta_{lb}$.

EVALUATION: SEMANTIC JOIN SEARCH

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

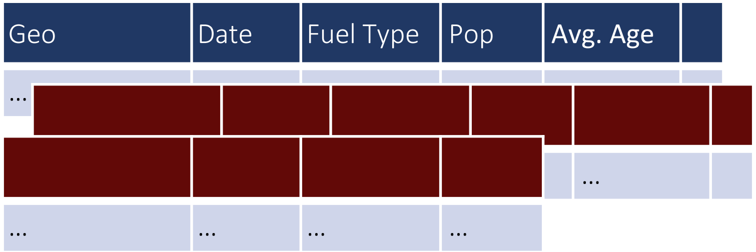
Dataset	KOIOS Response Time (s)	SOTA Response Time (s)	KOIOS Mem (MB)	SOTA Mem (MB)
DBLP	0.83	211	0.83	11
OpenData	18.6	101	18.6	102.5
Twitter	0.7	518	0.7	10
WDC	147	1062	147	885

- 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.

Table Union Search,
F. Nargesian, E. Zhu, K. Pu, R. Miller,
VLDB, 2018.

BEYOND JOIN

TABLE UNOIN DISCOVERY




```
SELECT *
FROM Query
UNION
SELECT ??
FROM ??
UNION
??…
```

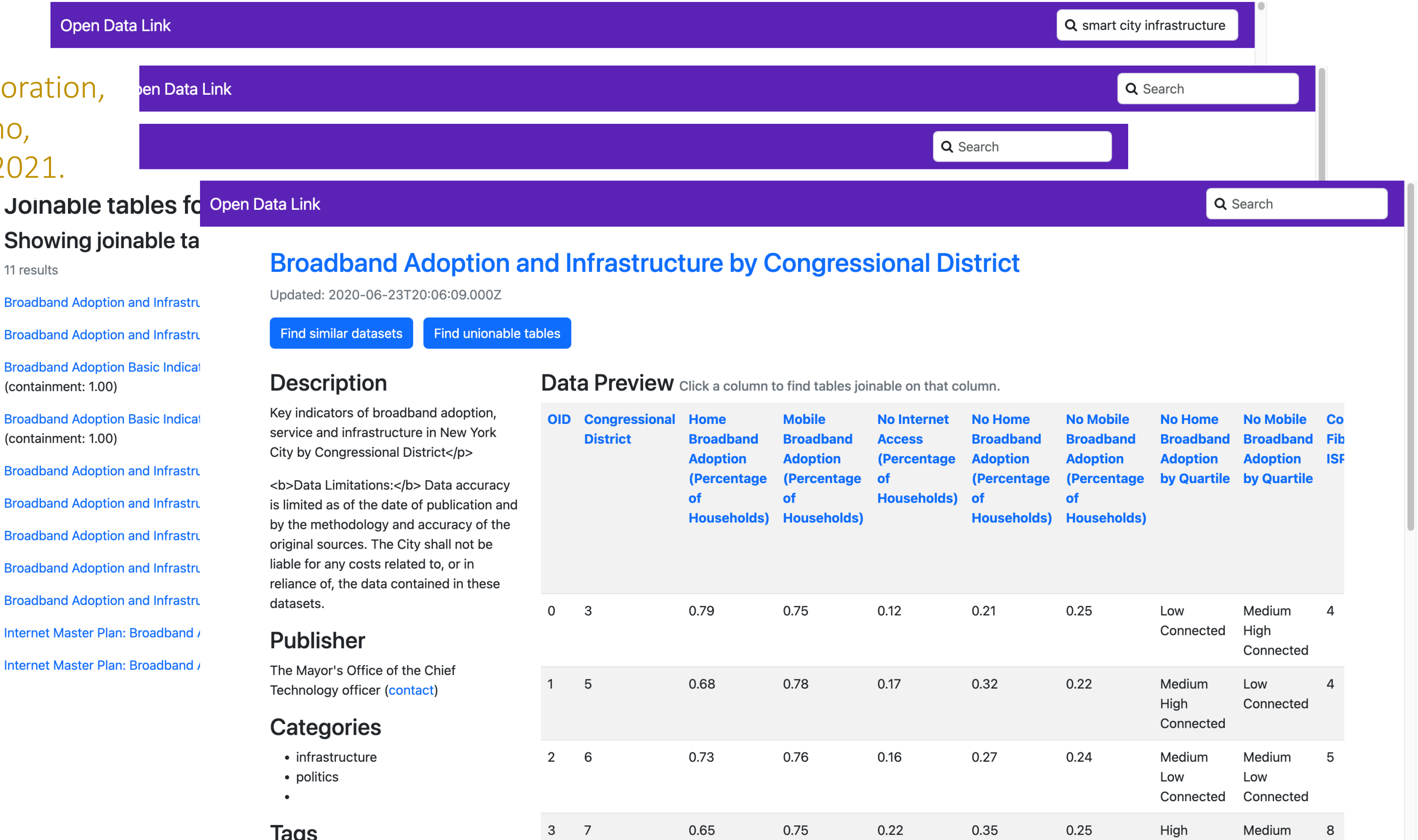
DIRECTORY STRUCTURE

```
Health
  ├── Water
  └── Food Resilience
        ├── Food Safety
        └── Food Production, …
Energy
…
Climate
```

Organizing Data Lakes for Navigation,
F. Nargesian, K. Pu, E. Zhu+, SIGMOD, 2020.

A Search Engine on Open Data

RONIN: Data Lake Exploration,
P. Ouellette, A. Sciortino,
F. Nargesian+, VLDB, 2021.



Open Data Link smart city infrastructure

en Data Link Search

Search

Joinable tables fo

Showing joinable ta

11 results

Broadband Adoption and Infrastru

Broadband Adoption and Infrastru

Broadband Adoption Basic Indica
(containment: 1.00)

Broadband Adoption Basic Indica
(containment: 1.00)

Broadband Adoption and Infrastru

Broadband Adoption and Infrastru

Broadband Adoption and Infrastru

Broadband Adoption and Infrastru

Broadband Adoption and Infrastru

Internet Master Plan: Broadband

Internet Master Plan: Broadband

Open Data Link Search

Broadband Adoption and Infrastructure by Congressional District

Updated: 2020-06-23T20:06:09.000Z

Find similar datasets Find unionable tables

Description

Key indicators of broadband adoption, service and infrastructure in New York City by Congressional District</p>

Data Limitations: Data accuracy is limited as of the date of publication and by the methodology and accuracy of the original sources. The City shall not be liable for any costs related to, or in reliance of, the data contained in these datasets.

Publisher

The Mayor's Office of the Chief Technology officer (contact)

Categories

- infrastructure
- politics
-

Tags

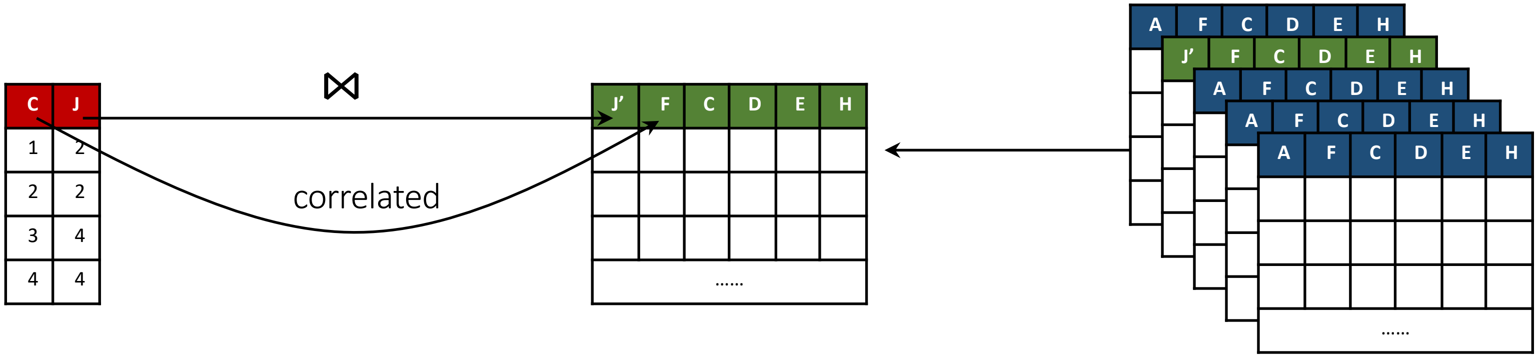
Data Preview
Click a column to find tables joinable on that column.

OID	Congressional District	Home Broadband Adoption (Percentage of Households)	Mobile Broadband Adoption (Percentage of Households)	No Internet Access (Percentage of Households)	No Home Broadband Adoption (Percentage of Households)	No Mobile Broadband Adoption (Percentage of Households)	No Home Broadband Adoption by Quartile	No Mobile Broadband Adoption by Quartile	Co Fib ISP
0	3	0.79	0.75	0.12	0.21	0.25	Low Connected	Medium High Connected	4
1	5	0.68	0.78	0.17	0.32	0.22	Medium High Connected	Low Connected	4
2	6	0.73	0.76	0.16	0.27	0.24	Medium Low Connected	Medium Low Connected	5
3	7	0.65	0.75	0.22	0.35	0.25	High	Medium	8

More on Dataset Discovery

FEATURE DISCOVERY

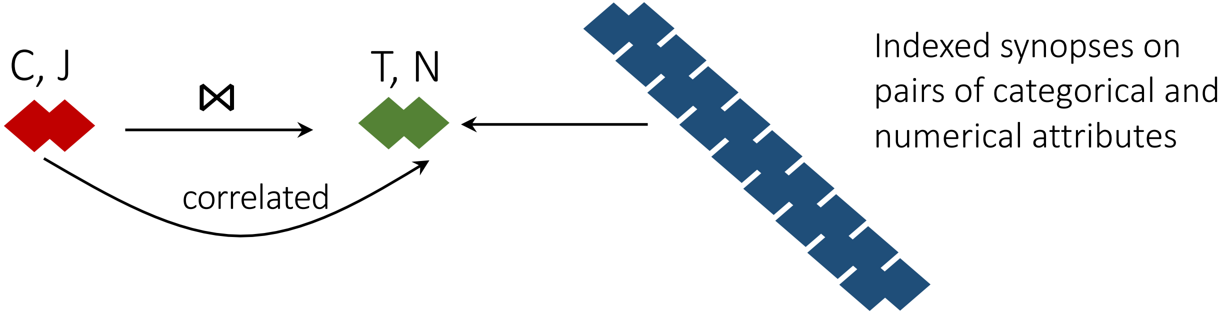
- Given a target column and a join column from a query table, find joinable tables s.t. the table contains a column that is correlated with the target column.



Correlation Sketches for Approximate Join-Correlation Queries, Santos et al., SIGMOD, 2021.

FEATURE DISCOVERY

- Evaluate correlation measures on the synopses that enable the reconstruction of a uniform random sample of the joined table.
- How to find attributes that are minimally correlated with sensitive attributes and highly correlated with the target attributes?
- The synopses may be biased towards the majority group



Correlation Sketches for Approximate Join-Correlation Queries, Santos et al., SIGMOD, 2021.

Other Works

- Table Discovery in Data Lakes

  Table Discovery in Data Lakes: State-of-the-art and Future Directions, SIGMOD, 2023.

- Goal-Oriented Data Discovery

  METAM: Goal-oriented Data Discovery, ICDE, 2023.

Outline

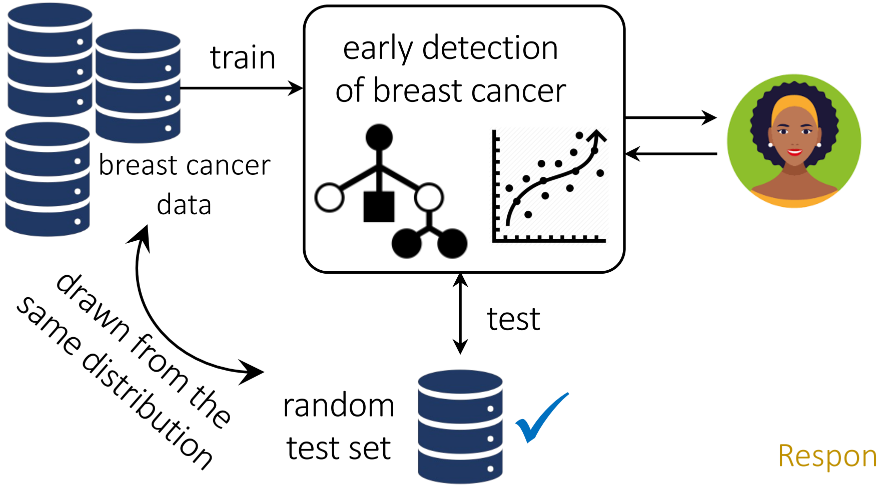
DATASET DISCOVERY:
Syntactic and Semantic Join Search,
Feature and Slice Discovery

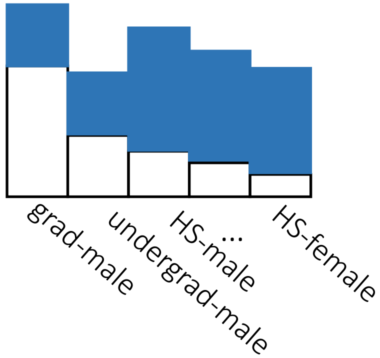
QUERY ANSWERING:
Random Sampling
over Union of Joins

FAIRNESS-AWARE DATA ACQUISITION:
Data Distribution Tailoring

DISTRIBUTION-AWARE DATA INTEGRATION

- A model is not bad overall it performs poorly on certain slices of data.
- Data debiasing

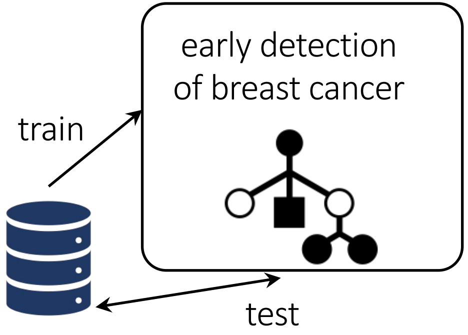




Responsible Data Integration: Next-generation Challenges, F. Nargesian, A. Asudeh, H. V. Jagadish, SIGMOD 2022.

Group Representativeness

- Groups: (in)dependent variables, protected groups, class labels, rare outcome groups, etc.
- Distribution
  - What. counts of proportions over groups
  - How. model debugging, data coverage [Asudeh+2018]
- Data
  - Where. crowdsourcing, data lakes, data markets



1K monitoring data in Chicago with at least 30% label=positive, and at least 20% African American patients

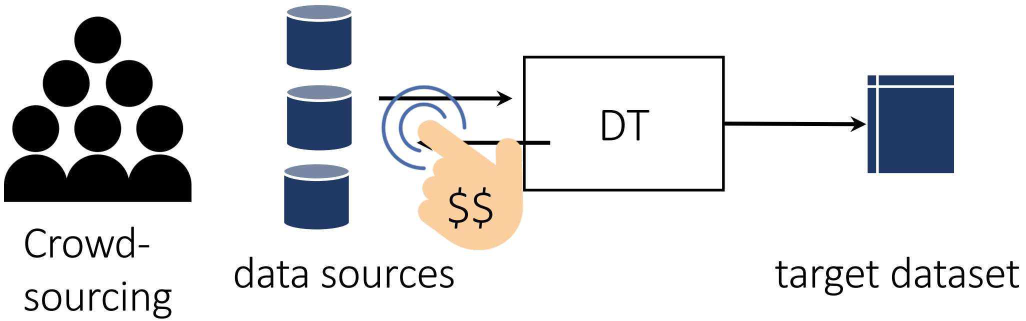
DATA DISTRIBUTION TAILORING (DT)

- How to construct a dataset that satisfies group distribution requirements from multiple sources in a cost-effective manner?
- Data debiasing: at the data acquisition step of data science pipeline

Tailoring Data Source Distributions for Fairness-aware Data Integration,
F. Nargesian, A. Asudeh, H. V. Jagadish, VLDB 2021.
Towards Distribution-aware Query Answering in Data Markets,
A. Asudeh, F. Nargesian, VLDB 2022.

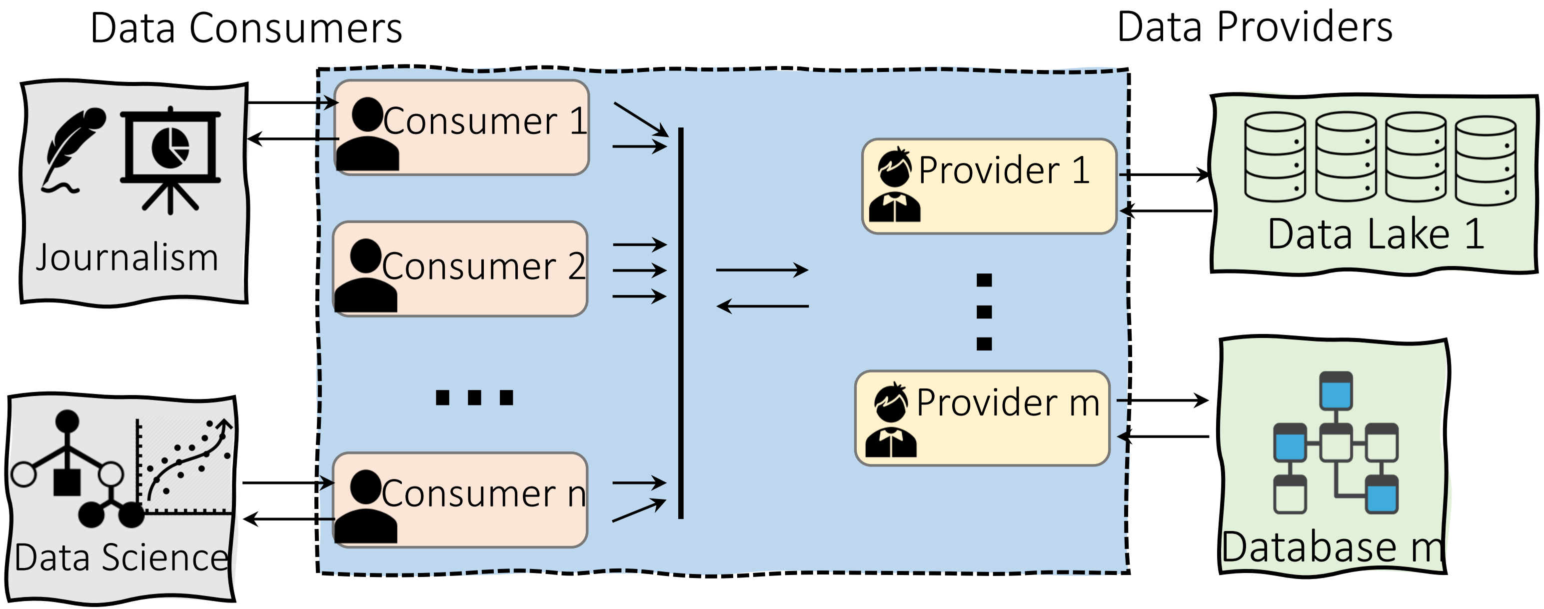
QUERY, DATA, COST MODELS

- Query: counts specified over some groups
- Tuple-at-a-time access to a source
  - Sources return relevant data
- Paying a cost for samples: monetary, labeling, computation, etc.



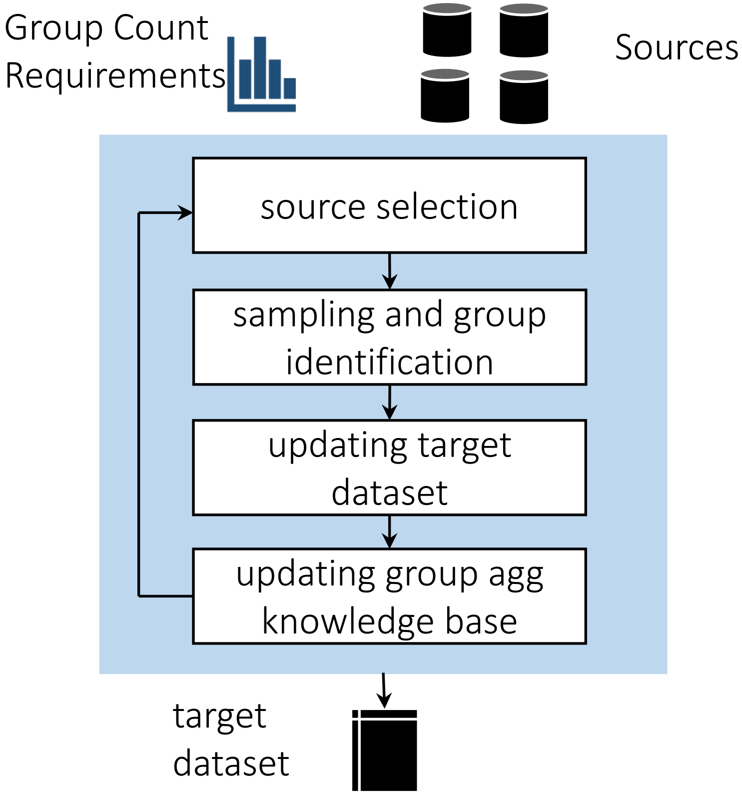
1K monitoring data in Chicago with at least 30% label=positive, and at least 20% African American patients

DATA MARKETPLACES



Towards Distribution-aware Query Answering in Data Markets,
A. Asudeh, F. Nargesian, VLDB 2022.

DATA DISTRIBUTION TAILORING (DT)



Problem. Given sources with their costs, and count requirements on the groups, select a sequence of sources to sample, s.t. count requirements are fulfilled, while the expected total query cost is minimized.

Are statistics about groups of interest available from data sources?

DT: DIRECT OPTIMIZATION

- Direct solution by defining the cost function and solving a DP problem
- Not practical for realistic settings
  - Pseudo-polynomial time and space complexity

Package queries: efficient and scalable computation of high-order constraints, Brucato M. et al., VLDBJ 2018.

DT: COST FUNCTION

P_i^j : prob of obtaining G_j from D_i

F(Q): expected cost of a target with counts Q

prob of G_j from D_i

cost of sample

$$\mathit{cost\ if}\ D_i\ @\mathit{iter}\colon\ C_i + \underbrace{\sum_{j=1,Q_j>0}^{m} P_i^j F_j(Q) + \left(1 - \sum_{j=1,Q_j>0}^{m} P_i^j\right) F(Q)}_{\text{expected remaining cost}}$$

exp. remaining cost if a sample of G_j is obtained.

exp. remaining cost if sample does not help with the target

A Dynamic Programming Solution

	cost	groups	
	C_i	G_1	G_2
sources D_1	2	0.2	0.8
D_2	3	0.4	0.6

Query: G_1: 1 and G_2: 1

F(1,1): the cost of a target with G_1: 1 and G_2: 1

G_1 \ G_2		
	F(0,0)=0	F(0,1)
	F(1,0)	F(1,1) ✓

D_2 D_1

cost of obtaining a tuple of G_1 from D_1: 2/0.2=10

cost of obtaining a tuple of G_1 from D_2: 3/0.4=7.5

F(1,0) = min(2/0.2, 3/0.4) = 7.5 $\Leftarrow D_2$

F(0,1) = min(2/0.8, 3/0.6) = 2.5 $\Leftarrow D_1$

select D_1: 2 + 0.2 F(0,1) + 0.8 F(1,0)

select D_2: 3 + 0.4 F(0,1) + 0.6 F(1,0)

F(1,1) = min(2 + 0.2 F(0,1) + 0.8 F(1,0),
3 + 0.4 F(0,1) + 0.6 F(1,0)) = 8.4 $\Leftarrow D_1$

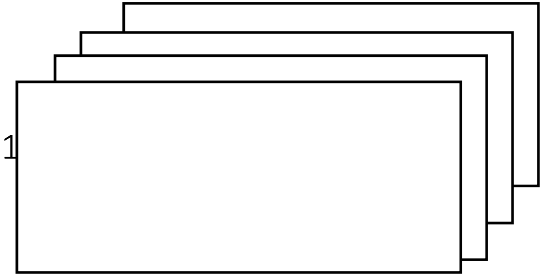
DT: COST FUNCTION

P_i^j : prob of obtaining G_j from D_i

F(Q): expected cost of a target with counts Q

$$F(Q) = \ min_{\forall\, D_i}\, C_i + \sum_{j=1, Q_j>0}^{m} P_i^j F_j(Q) + (1 - \sum_{j=1, Q_j>0}^{m} P_i^j) F(Q)$$

Query: G_1: 1 and G_2: 1
Sources: D_1 and D_2

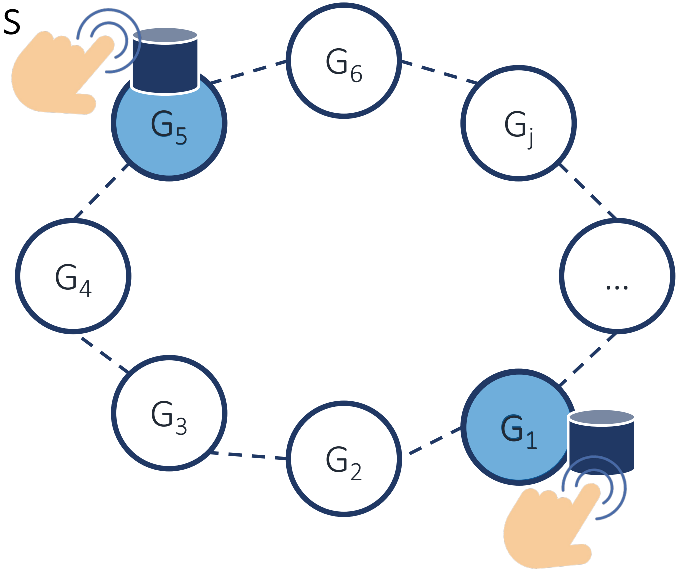


STRATEGY: KNOWN DISTRIBUTIONS

- Round-robin with priority strategy on groups
- Prioritize minority group
  - rare and expensive to find
- Priority of G_j:

$$D_{*j} = \underset{\forall D_i}{\operatorname{argmax}} \frac{\text{prob of } G_j \text{ in } D_i}{\text{cost of} D_i}$$

$$\text{priority}(G_j) = \overline{\text{cost}} \text{ per sample of } G_j$$
$$\text{if select } D_{*j}$$



DT ANALYSIS

- Prioritize minority group
- Result. Optimal for two groups and equi-cost model.

Optimal Equi-Cost Binary

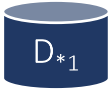
- Find the optimal source for each group: D_{*1} and D_{*2}

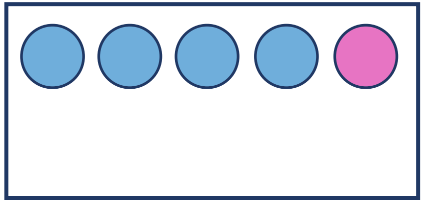
$$\text{priority}(G_j) = \frac{1}{prob\ of\ G_j\ in\ D_{*j}}$$

D_{*1} has 20% of G_1 and 80% of G_2

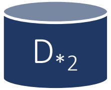
D_{*2} has 5% of G_1 and 95% of G_2

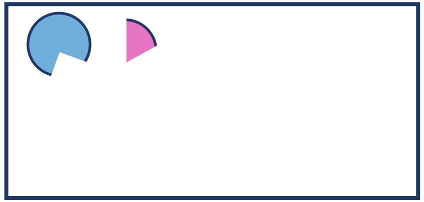
select D_{*1}





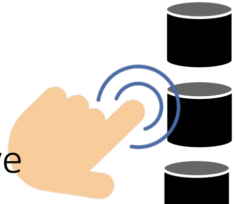
select D_{*2}





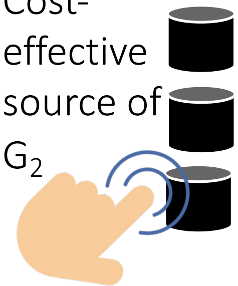
General Non-Binary DT: Analysis

Cost-effective source of G_1

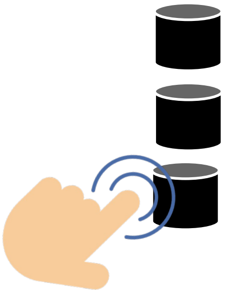


+

Cost-effective source of G_2



+ … +



Cost-effective source of G_m

Cost of fulfilling count requirements of G_1 + Cost of fulfilling count requirements of G_2 + … + Cost of fulfilling count requirements of G_m

- Modeling the problem as *m* instances of the *coupon collector's problem,* where every instance *j* aims to collect samples from the group G_j.

COUPON COLLECTOR'S PROBLEM

- Given *n* coupon types, how many coupons do you expect you need to draw *with replacement* before having drawn each coupon at least once?
    - Assume all coupons are equally likely.
- After one sample, we have seen one coupon.
- After two samples, we have seen the same coupon twice with probability $\frac{1}{n}$ and two different coupons with probability $\frac{n-1}{n}$.
- It is shown that the expected number of samples needed grows as

$$\Theta(n \log n)$$

Randomized algorithms, Motwani and Raghavan.

Dt Analysis

- Prioritize minority group
- Result. Optimal for two groups and equi-cost model.
- Expected cost of *m*-groups with arbitrary cost

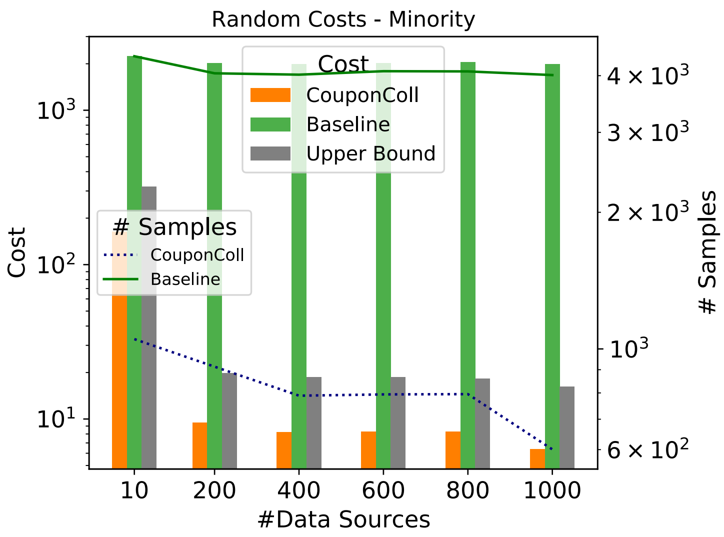
$$\psi = \sum_{j=1}^{m} C_{*j} N_{*j} \ln \frac{N^{j}_{*j}}{N^{j}_{*j} - Q_j}$$

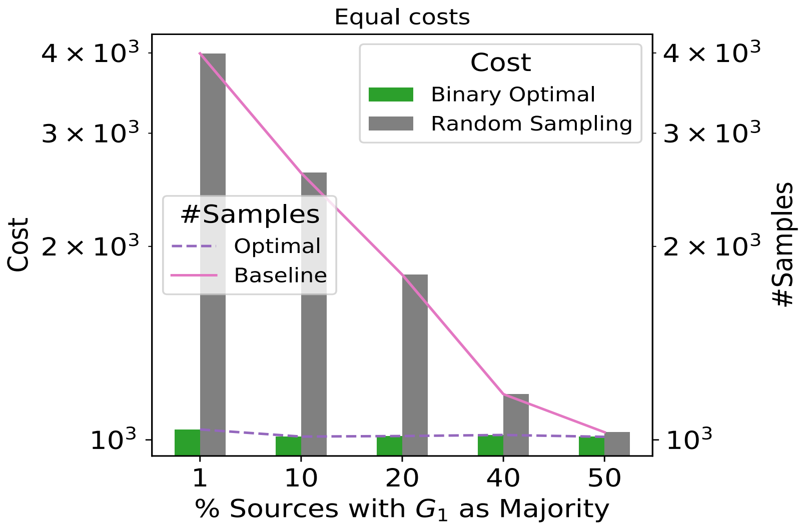
\# of group j in D_i

\# of group j in D_i

  - based on the coupon collector's problem [Motwani and Raghavan'1995]

EVALUATION: KNOWN DT





- Having access to more sources incurs lower DT cost.
- Random source selection is only suitable when no group is a minority in the repository!

DT : UNKOWN DISTRIBUTIONS

- Multi-armed Bandit (MAB)
  - Given a time horizon T, a centralized planner sequentially chooses actions, receiving stochastic reward from unknown distribution

MULTI-ARMED BANDIT

- Sequential; exploration/exploitation tradeoff
- *n* arms; each arm Γ_i is associated with an unknown probability distribution ν_i with mean θ_i.
- An agent selects an arm

at every iteration.

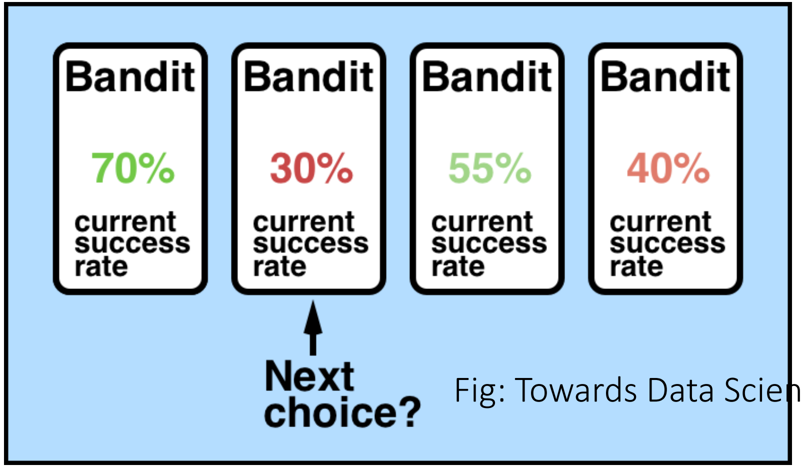


Fig: Towards Data Science

The Multi-Armed Bandit Problem: Decomposition and Computation.
Katehakis and Veinott, 1987.

MULTI-ARMED BANDIT

- $r_t = R(a_t)$: reward of a_t taken from ν_i

$$\mathbb{E}[R(a_t = \Gamma_i)] = \theta_i$$

- Goal is to maximize the expected cumulative reward
- $A = a_1, \cdots, a_T$: sequence of actions taken by an agent
- $A^* = a_1^*, \cdots, a_T^*$: optimal strategy
- Regret for not taking the optimal action

$$L(A) = \mathbb{E}[\sum_{t=1}^{T}(\theta_t^* \quad -R(a_t))]$$

θ_t^*: optimal expected reward at t

MAB STRATEGIES

- Exploitation: query each data set once and focus on the source with maximum reward
  - Works well with large # sources or when distributions vary greatly
- Exploration: choose a source at random with equal budget chance
  - Selection probability is inverse proportional to cost
  - Works well when distributions are similar
- Upper Confidence Bound

A contextual-bandit approach to personalized news article recommendation, Li et al. 2010.

Upper Confidence Bound

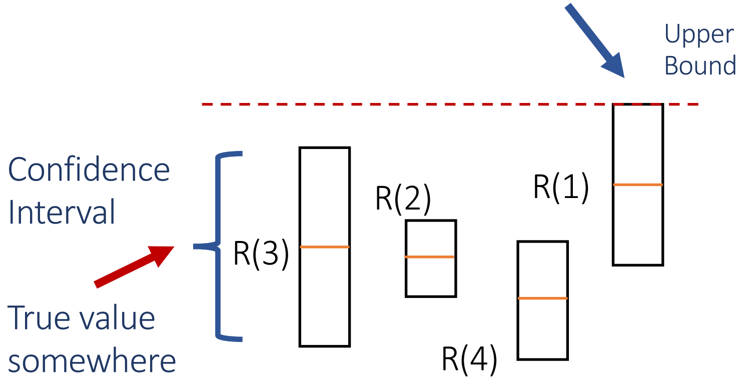
- Exploration/exploitation trade-off
- UCB favors exploration of sources with a strong potential to have an optimal reward value.

$$D = \underset{\forall D_i}{\operatorname{argmax}} \overline{R}_t(i) + U_t(i)$$

- Hoeffding inequality

$$U_t(i) = \big(R_\top(i) - R_\perp(i)\big)\sqrt{\frac{2 \ln t}{O_i}}$$

t: # samples, O_i : samples taken from D_i



DT : UNKOWN DISTRIBUTIONS

- Multi-armed Bandit (MAB)
  - Given a time horizon T, a centralized planner sequentially chooses actions, receiving stochastic reward from unknown distribution
- Goal: minimize regret

$$\mathrm{Regret(T)} = \mathrm{OPT\ reward\ @T} - \mathrm{DT\ reward\ @T}$$

- Optimal regret is $\tilde{O}(\sqrt{T})$.

EPS-GREEDY MAB FOR DT

- Explore with epsilon probability
  - Sample a random source D_t and update empirical ratios of groups in the D_t
- Otherwise, exploit
  - Two-level policy with a frequentist DT
  - Group to prioritize

$$G_t \leftarrow \mathrm{argmax}_{G_j} \left(\underbrace{Q_j}_{\text{Remaining count req of } G_j} . \min_{D_i} \Big(\underbrace{\frac{C_i}{\overline{\overline{\mathrm{ratio}}}(\underbrace{G_j}_{\text{empirical ratio of } G_j \text{ in } D_i})}}_{\text{cost per successful sample}} \Big) \right)$$

  - Source to choose

$$D_t \leftarrow \mathrm{argmin}_{D_i} \frac{C_i}{\overline{\overline{\mathrm{ratio}}}(G_j)}$$

- Results. An ε-greedy strategy with exploration probability $\sqrt[3]{\ln t / t}$ at time t: regret of $O(\mathrm{T}^{2/3} \log \mathrm{T}^{1/3})$ at time *T* for equi-cost DT.

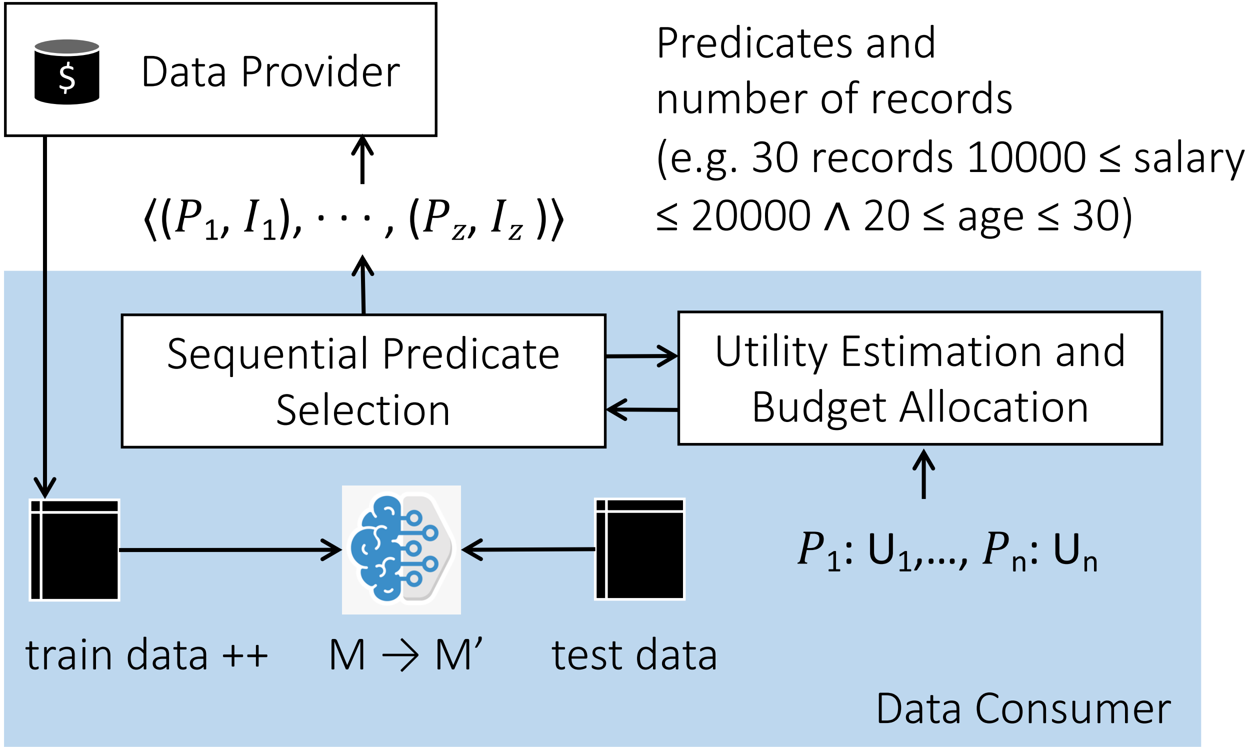
DATA ACQUISITION FOR ML

- Consumers query providers for data to enhance the accuracy of their models.
- The task of the consumer is to identify a series of queries $\langle(P_1, I_1), \cdots, (P_z, I_z)\rangle$ to obtain B records, where P_i and I_i being the predicate and the number of requested records in the i-th query.
- The objective is to improve as much as possible the accuracy of consumer’s ML model on test data.

Data Acquisition for Improving Machine Learning Models, Li et al., PVLDB, 2021.

Selective Data Acquisition in the Wild for Model Charging, Chai et al., PVLDB 2022

DATA ACQUISITION FOR ML



Data Acquisition for Improving Machine Learning Models, Li et al., PVLDB, 2021.

OUTLINE

DATASET DISCOVERY:

Syntactic and Semantic Join Search,
Feature and Slice Discovery

QUERY ANSWERING:

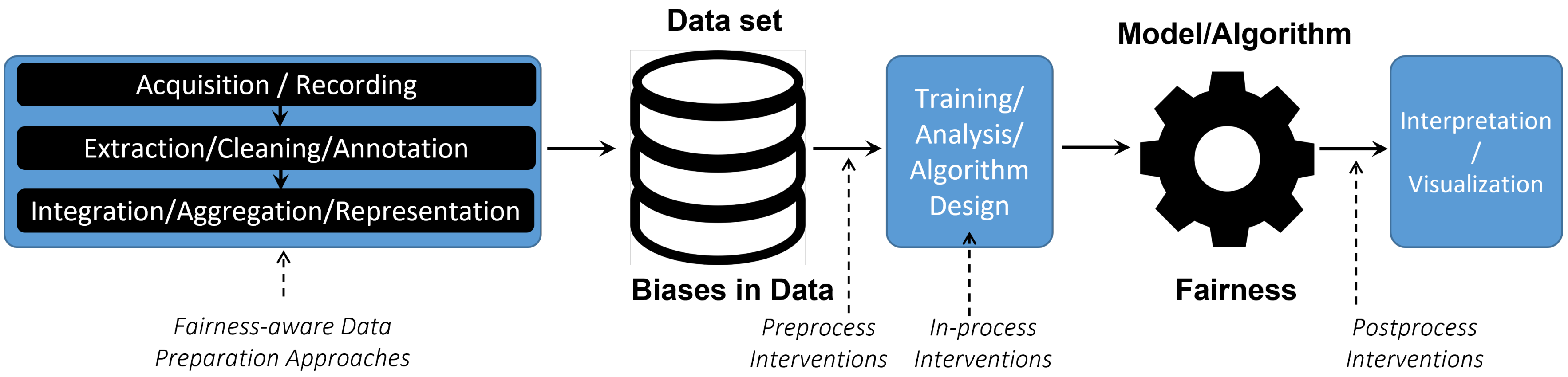
Random Sampling
over Union of Joins

FAIRNESS-AWARE DATA ACQUISITION:

Data Distribution Tailoring

Responsible Data: Next Generation Requirements

DATA BIAS IN ML PIPELINE



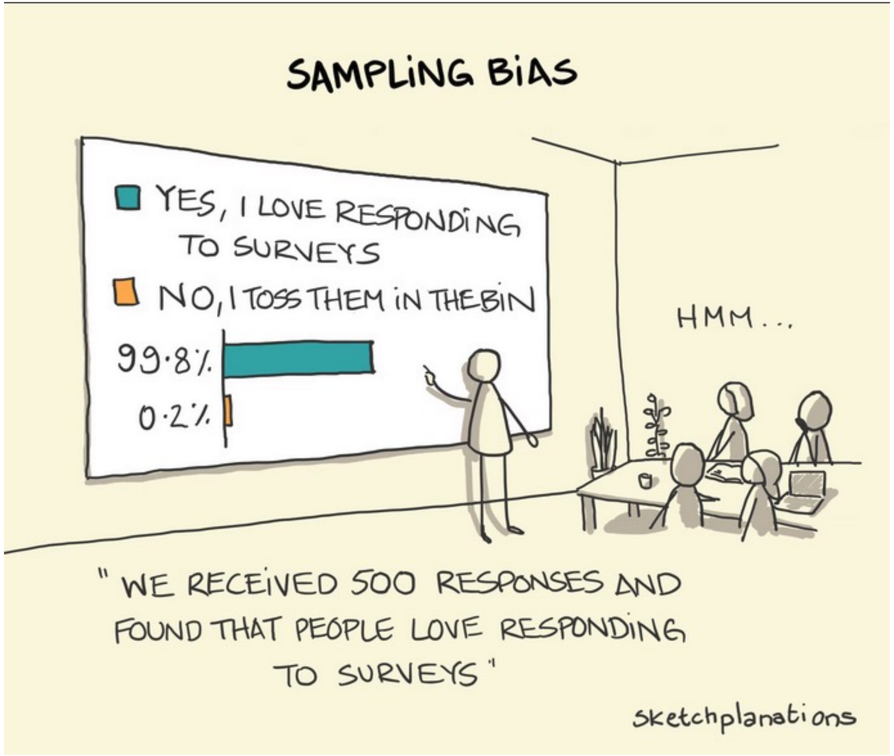
Nima Shahbazi, et al. CSUR, 2023.

Underlying Distribution Representation

- Standard Assumption of AI: training data is i.i.d random samples drawn from the distribution that query points follow
  - Not always easy to satisfy
  - Not easy to verify
- Underlying distribution is usually unknown
  - Challenging to verify that collected data is unbiased

NOT EASY TO SATISFY

- Even if selected randomly
- Suppose surveys sent out to carefully chosen random sample
- Only a fraction of surveys returned



Group Representation

- The need to show adequate consideration of minority/rare groups, to ensure reliable outcomes for such groups

Unbiased and Informative Features

- An AI data set: a collection of attributes (features) $\boldsymbol{x} = \{x_1 \ldots x_m\}$
  - may also contain one (or more) target attribute (labels) $\boldsymbol{y}$
  - sensitive attributes $\boldsymbol{s}$ such as race and gender
- Often challenging to collect sensitive attributes
  - Example: users of a shopping website
    - Usually do not collect the sensitive information of the users

INFORMATIVE FEATURES

- Performance of ML models depends on the set of attributes a data set contains
  - E.g., in classification predict the target variable using the observations

→ High correlation between $\boldsymbol{x}$ and $\boldsymbol{y}$

UNBIASED FEATURES

- Sensitive attributes are used to specify (demographic) groups considered for fairness
  - E.g.: race={White, Black, Hispanic, others}

- <u>*Low*</u> correlation between the features and the sensitive attributes

- Ideally $\boldsymbol{x}$ and $\boldsymbol{s}$ should be independent

Completeness and Correctness

- Always important, even more critical for responsible AI
  - incomplete and incorrect data typically hurt minorities, further increasing the data bias in such cases.

- Example
  - Two groups (minority and majority); a small portion belong to the minority
  - A simple task: compute *average*
  - An incorrect **majority** value does not significantly impact the average
  - An incorrect **minority** value may **significantly skew** the average

SCOPE OF USE AUGMENTATION

- Collecting data that fully satisfies *all* requirements is often not possible in practice.
- Some of the requirements may conflict with others
  - Group representation requirement may conflict with i.i.d sample requirement

- Every data set has a limited *scope of use*. No data set is good for all tasks.

- To ensure transparency:
  - embed data with the meta-data and information that describe its collection process, its limitations, and its fitness for use

Sampling over Data Lakes?

UNIFORM AND INDEPENDENT SAMPLING

- ML on integrated data is inherently expensive
- Luckily, in many tasks (e.g. AQP and statistical learning), a random sample suffices for analysis
- Samples should satisfying Underlying Distribution Representation and Group Representation requirements

UNIFORM AND INDEPENDENT SAMPLING

- Sampling a single source
  - **Stratified sampling** to ensure that minority groups are sufficiently represented in the sample
  - Given a set of sensitive attributes and an integer parameter *k*, a stratified sampling guarantees at least *k* tuples are sampled uniformly at random from each group. When a group has fewer than *k* tuples, all of them are retained.

Join on Samples: A Theoretical Guide for Practitioners, Huang et al., PVLDB, 2019.

ML on Normalized Data

- Predicting the return flag of an item shipped to a customer using features of both the item and another item shipped to the same customer requires (self-) join

Label	Features						
Flag	CustId	Region	Total	Discount	Flag2	Total2	Discount2
1	10	2	100	0.2	0	20	0.5
0	20	1	200	0.0	0	100	0.1
0	20	1	500	0.1	0	300	0.2
…	…						

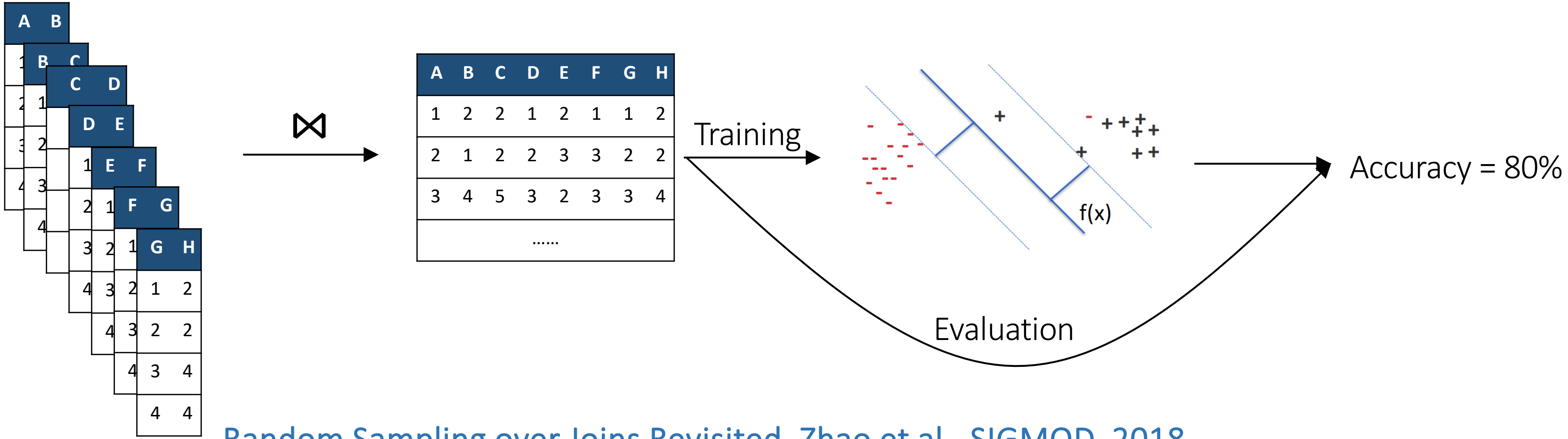
ML ON NORMALIZED DATA

```
SELECT
    l1.l_returnflag, n_regionkey, s_acctbal,
    l1.l_quantity, l1.l_extendedprice, l1.l_discount,
    l1.l_shipdate, o1.o_totalprice, o1.o_orderpriority,
    l2.l_quantity, l2.l_extendedprice, l2.l_discount,
    l2.l_returnflag, l2.l_shipdate
FROM nation, supplier, lineitem l1, orders o1,
     customer, orders o2, lineitem l2
WHERE   s_nationkey = n_nationkey
    AND s_suppkey = l1.l_suppkey
    AND l1.l_orderkey = o1.o_orderkey
    AND o1.o_custkey = c_custkey
    AND c_custkey = o2.o_custkey
    AND o2.o_orderkey = l2.l_orderkey;
```

Joining 7 TPCH tables

IID SAMPLING OVER JOIN

- Training a classifier using SVM on a join over 7 tables
  - Full join takes more than 12 hours to compute.
  - Training runs forever without down-sampling.



Random Sampling over Joins Revisited, Zhao et al., SIGMOD, 2018.

IID Sampling over Join

- Given T_1 and T_2, a sampling algorithm A is iid, if tuples returned by A all have the same sampling probability and the appearances of two tuples in the sample are independent events.

A	B
1	1
2	1
3	1
4	2
5	2
6	2
7	3

B	C
2	1
2	2
3	3
3	4
3	5
3	6

A	B	C
4	2	1
4	2	2
5	2	1
5	2	2
6	2	1
6	2	2
7	3	3
7	3	4
7	3	5
7	3	6

$R(A,B) \bowtie S(B,C) = R(A, B, C)$

Goal: sample $t \in R$ with probability $\frac{1}{10}$

IID SAMPLING OVER JOIN

- Sampling cannot be pushed down in join

$$sample(R) \bowtie sample(S) \neq sample(R \bowtie S)$$

- If independent samples are taken from R and S, the result of joining uniform samples is a uniform sample of the join but not an independent one.
- Consider independent Bernoulli samples with probability p from R and S
  - $P(t_1, t_2) = p^2$, $t_1 \in R$ and $t_2 \in S$
  - $P(t_1, t'_2) = p$, $t_1 \in R$ and $t'_2 \in S$
  - Uniform and dependent

On Random Sampling over Joins, Chaudhuri et al., SIGMOD, 1999.

IID Sampling over Join

- Two-table join

On Random Sampling over Joins, Chaudhuri et al., SIGMOD, 1999.
Random Sampling from Databases, Olken, Ph.D. Dissertation, 1993.

- Multi-way foreign key joins

Join Synopses for Approximate Query Answering, Acharya et al., SIGMOD, 1999.

- Ripple join (uniform but correlated samples)

A scalable hash ripple join algorithm, Luo et al., SIGMOD 2002.

- Wander join (independent but non-uniform samples)

Wander Join: Online Aggregation via Random Walks, Lo et al., SIGMOD 2016.

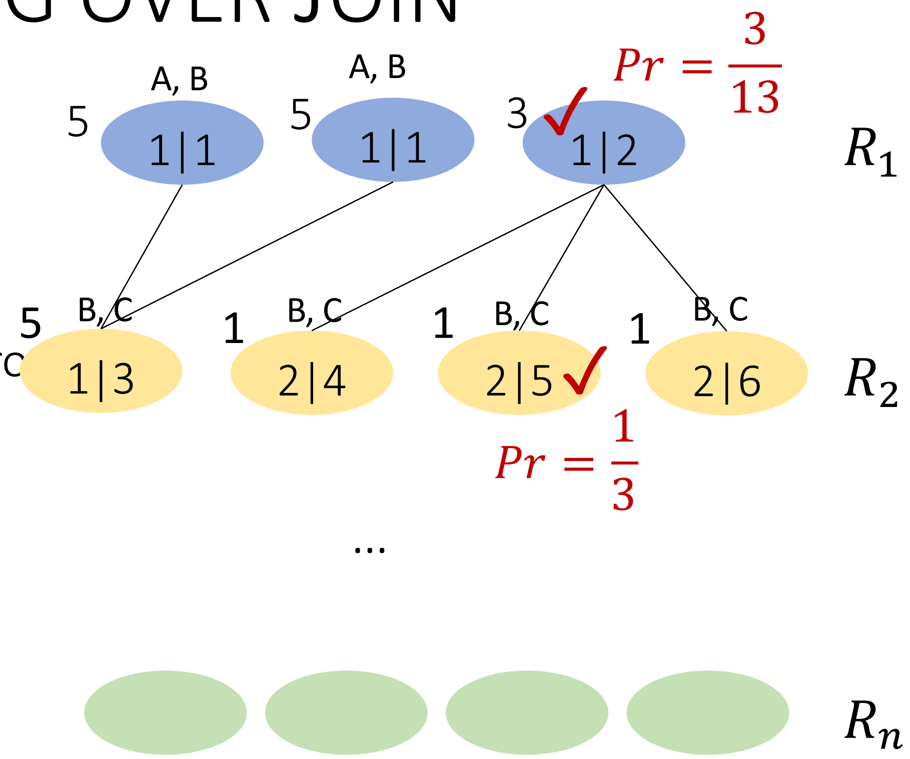
IID SAMPLING OVER GENERIC JOIN PATHS

- Randomness: return tuples from a join path $J = T_1 \bowtie \ldots \bowtie T_n$ with probability $1/|J|$
- Independence: generate sampled results continuously until a certain desired sample size k is reached

Random Sampling over Joins Revisited, Zhao et al., SIGMOD, 2018.

IID SAMPLING OVER JOIN

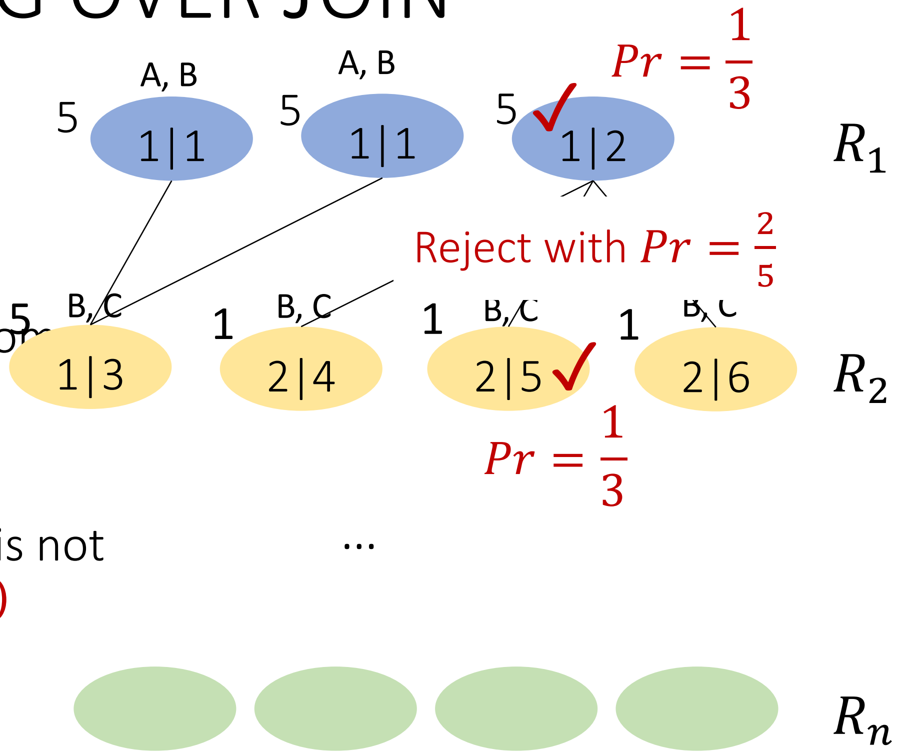
- A join path is modelled as DAG
  - nodes: tuples
  - edges: joinable tuples
- Weight $w(t)$: # join results starting from tuple t
- Sample proportional to weight



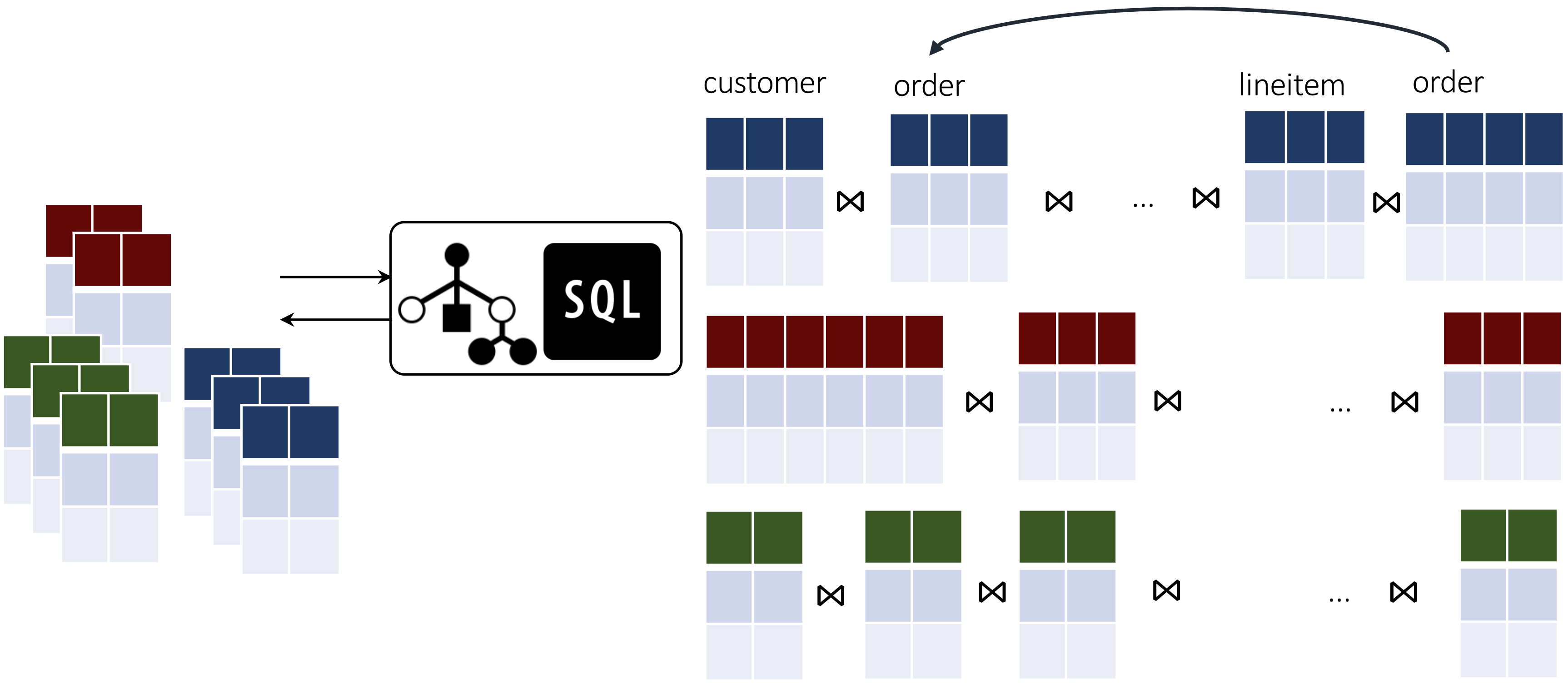
Random Sampling over Joins Revisited, Zhao et al., SIGMOD, 2018.

IID SAMPLING OVER JOIN

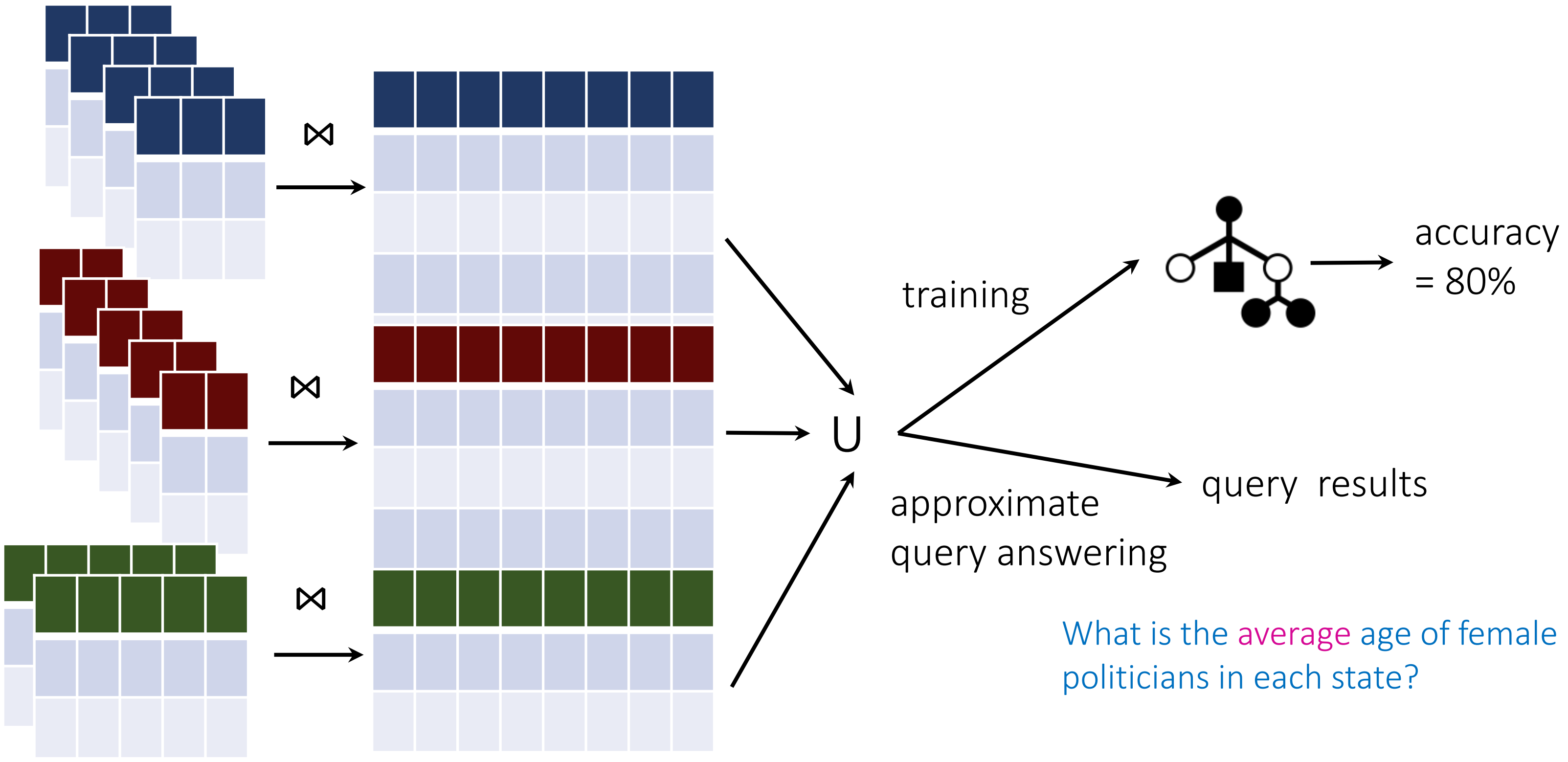
- A join path is modelled as DAG
  - nodes: tuples
  - edges: joinable tuples
- Weight $w(t)$: # join results starting from tuple t
- Sample proportional to weight
- Use a surrogate weight $W(t)$ if $w(t)$ is not available. $W(t)$: upper bound of $w(t)$
- Reject with prob. $\frac{W(t)-\sum_{t' \in ch(t)} W(t')}{W(t)}$
- Return when leaf



Union of Joins



JOINS AND UNIONS ARE EXPENSIVE.



What is the average age of female politicians in each state?

Random Sampling over Union of Joins

- Fortunately, no need to compute full results.
- A uniform and independent sample can achieve a bounded error [Vapnik+1971].
  - Robust for any models
- Problem. Given a set of joins $L=\{J_1, \dots, J_n\}$, let U be the discrete space of set union $U = J_1 \cup \dots \cup J_n$, return N independent samples S from U, without performing join and union, s.t.

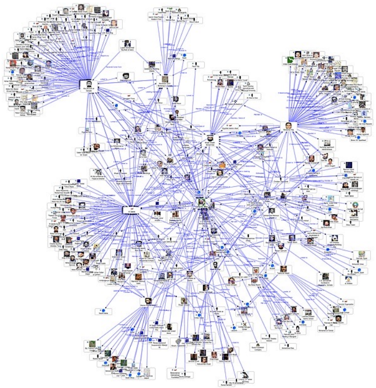
$$P(t \in S) = \frac{1}{|J_1 \cup \cdots \cup J_n|}$$

RESPONSIBLE DATA ACQUISITION

- Multi-modal dataset construction (visual analytics)
  - Uniformity across all modalities
- Data subset selection (coreset construction) under distribution constraints
  - Data subset selection with K-coverage, group representation, and diversity
  - Coresets over join paths
  - Coresets over noisy, dynamic, and stream data
- Auditing existing data management algorithms
  - Data cleaning and schema mapping

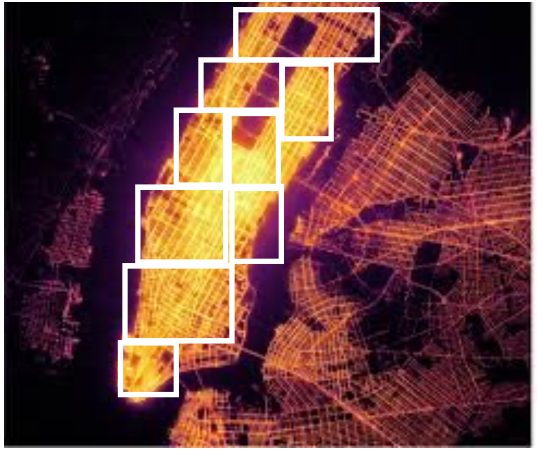
Coreset Construction

- Coreset construction under distribution constraints
  - Data subset selection with K-degree, group representation, and diversity
  - Coresets over join paths
  - Coresets over noisy, dynamic, and stream data



social network

ImageNet



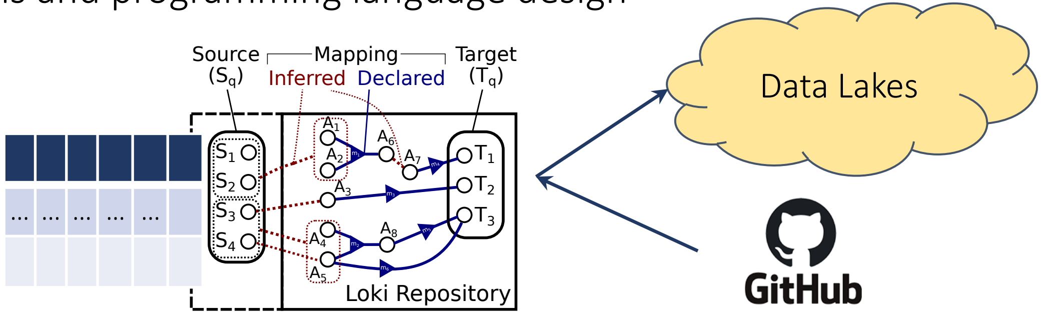
NYC taxi data

Auditing Data Management Pipelines

- Synergies and transparency and fairness
- Auditing data cleaning techniques
  - Entity matching
- Schema mapping
  - How bias is propagated through join and union operations?
- Leads to developing new algorithms

HUMAN-CENTRIC DATA ACQUISITION

- The design of a domain-specific programming language for data lake programming
  - Syntax and semantics of operators and programming constructs
  - Type checking
  - Iterative algorithms and programming language design



- Dialogue-based query answering over data lakes

ACKNOWLEDGEMENT

- Abolfazl Asudeh (UIC)
- Nikolaus Augsten (U. Salzburgh)
- H. V. Jagadish (U. Michigan)
- Renée J. Miller (NEU)
- Divesh Srivastava (AT&T)
- Ken PU (UOIT)
- Eric Zhu (MSR)
- Jiwon Chang (UR undergrad student)
- Nikola Danevski
- Yurong Liu (UR undergrad student)
- Pranay Mundra (UR grad student)
- Nima Shahbazi (UIC grad student)
- Draco Xu (UR undergrad student)
- Jianhao Zhang (UR grad student)

Code and data:
github.com/DataIntelligenceCrew

THANKS.