# DATA SCIENCE OVER DATA LAKES


**Renée J. Miller**
miller@northeastern.edu

# What is a data lake?

- A data lake is a system or repository of data stored in its ***natural format*** [Wikipedia].
- Why a lake vs. DBMS or Warehouse?
  - Cheaper (lower cost of ownership)
    - ✳Storing as files on HDFS or cloud has lower cost, compared with storing data in DBMS
  - Quick start
    - ✳Skip schema design to ingest data directly in its original format
      - CSV, JSON, YAML, …
    - ✳Data lakes are natural platforms for ML and big data processing engines (e.g., Apache Spark)

# Two Example Data Lakes[1]

A data science research institution (~100 employees). 1000-10k datasets.

- Mostly machine learning tasks
- Data is scattered and dumped into HDFS
- Use file directories, no centralized portal
- Duplicates are common; many versions of the same datasets
- Prefer simpler tools (similar to git) over enforced workflows
- Little integration

A global investment bank (>10k employees). More than 100k datasets.

- Operates a dedicated data lake portal
- Internal browser and DSL for querying
- Datasets are backed by pipelines for updates
- Complexity arises from assigning access control policies to roles
- Missing schema standardization across departments
- Integrating new data sets difficult

[1] Fatemeh Nargesian, Erkang Zhu, Renée J. Miller, Ken Pu, Patricia Arena's "Data Lake Management: Challenges and Opportunities", VLDB 2019

# Example Data Lake Stats

| | **Avg #Attr** | **#Attr** | **MaxSize** | **AvgSize** | **AvgStrSize** | **#UniqVal** |
|---|---|---|---|---|---|---|
| OpenData | 16 | 3,367,520 | 22,075,531 | 465 | 1,504 | 609,020,645 |
| WebTable | 5 | 252,766,759 | 17,033 | 10 | 11 | 193,071,505 |
| Enterprise - 7% | 12 | 2,032 | 859,765 | 4,011 | | 3,902,604 |

*167 table subset of MIT's 2400 table data warehouse [Deng et al., CIDR 2017]*

- Data lakes
  - Can be massive
  - Maintaining join graphs can be expensive/inpractical
  - Data scientists may not know/understand all data available

# Data Science Over Data Lakes

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

How can we facilitate data science over data lakes?

datos.gob.es
reutiliza la información pública

datos.gob.mx

بيانات دبي
dubaidata
A SMART DUBAI ESTABLISHMENT

data.gov.gr

OPENdata
TRENTINO

NYC
OPEN DATA

Datos Argentina

OPEN DATA
HONG KONG
香港開放數據

data.gov.ru
OPEN DATA RUSSIA

OPEN
GOVERNMENT
INDONESIA

WU
OPENDATA
data.wu.ac.at

DATA
GOUV.FR

data.gov.my

پایگاه داده باز ایران
IRANOPENDATA

EU Open Data
Portal
www.open-data.europa.eu

DATA.GOV.UK Beta
Opening up Government

OPEN DATA

dati.gov.it
I dati aperti della PA

dados.gov.br

OPEN
DATA
IRELAND

DATA.GOV

OPEN DATA
JAPAN

OPEN DATA
PHILIPPINES

data.gov.in
Open Government Data (OGD) Platform India

data.gov.au

open data durban

KENYA
openData

CITY OF
ORLANDO
OPEN DATA

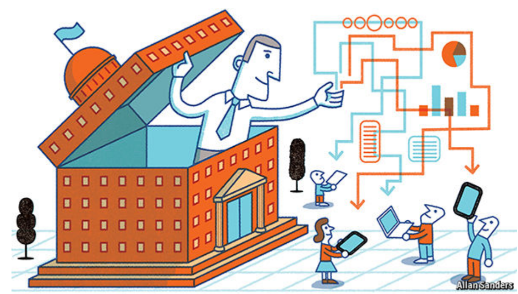
# Institutional Transparency

Proponents: advocates for government and institutional transparency, data science

- For improved governance & citizen engagement
- For inclusive development and innovation

***Open Data:***

data published in the public domain that is free to use, modify, and redistribute




Critics: the cost of data publishing and the limited benefit to the general public

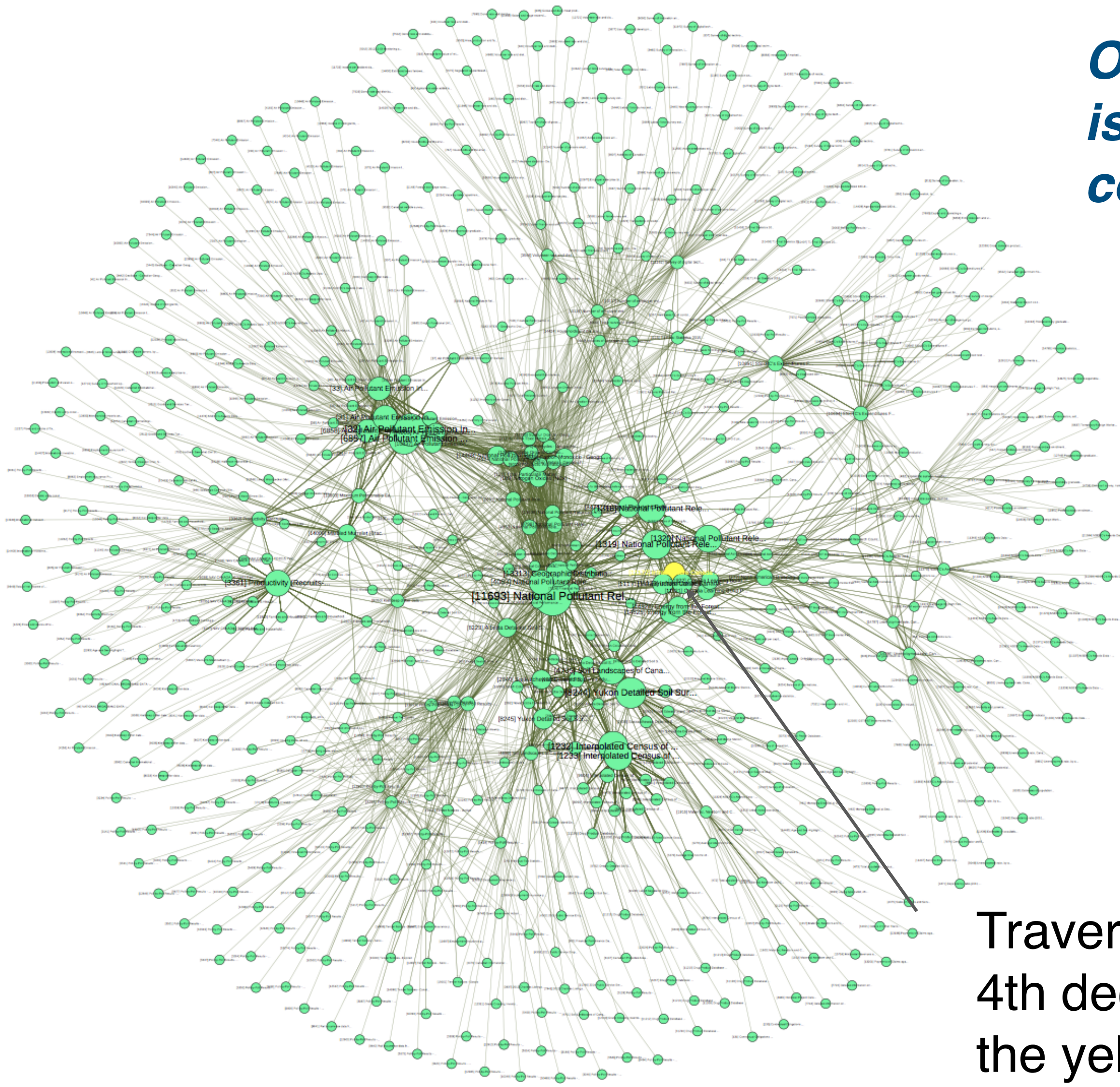
- Cost of ownership still too high

# Open Data Principles

- Timely & Comprehensive
- Accessible and Usable
- Complete
  - All public data is made available. Public data is data that is not subject to valid privacy, security or privilege limitations
- Primary
  - Including the original data & metadata on how it was collected

***Invaluable for data science***

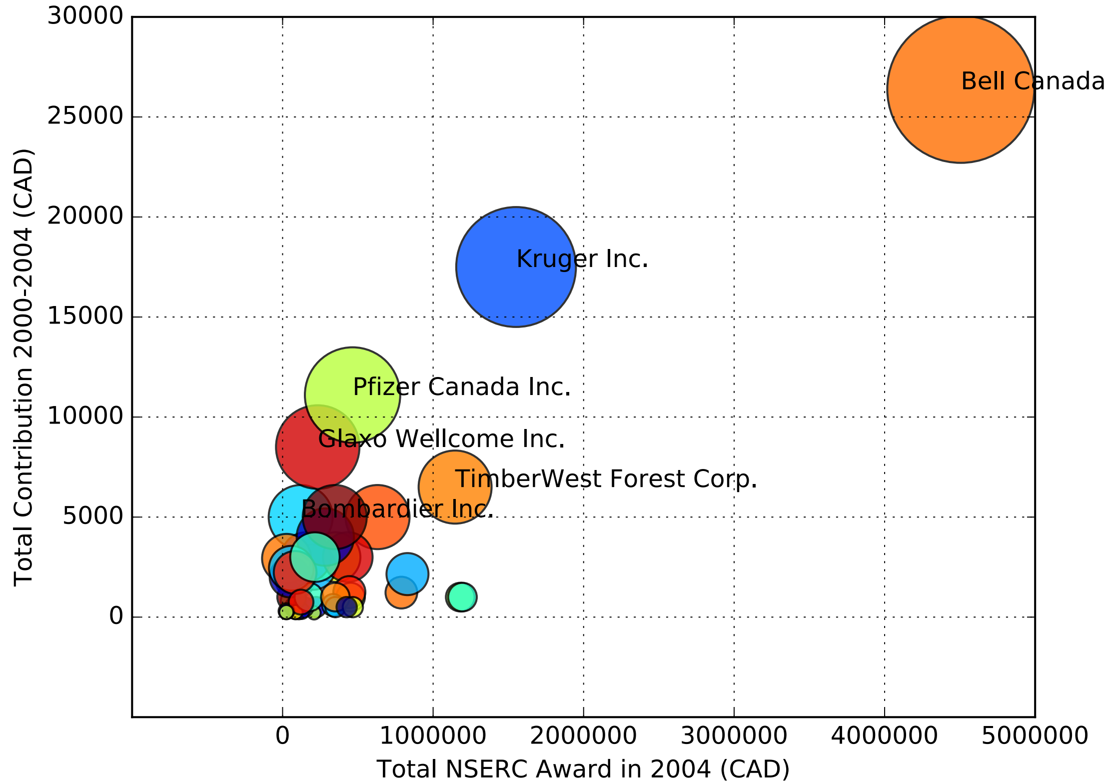
***Open Data is deeply connected***



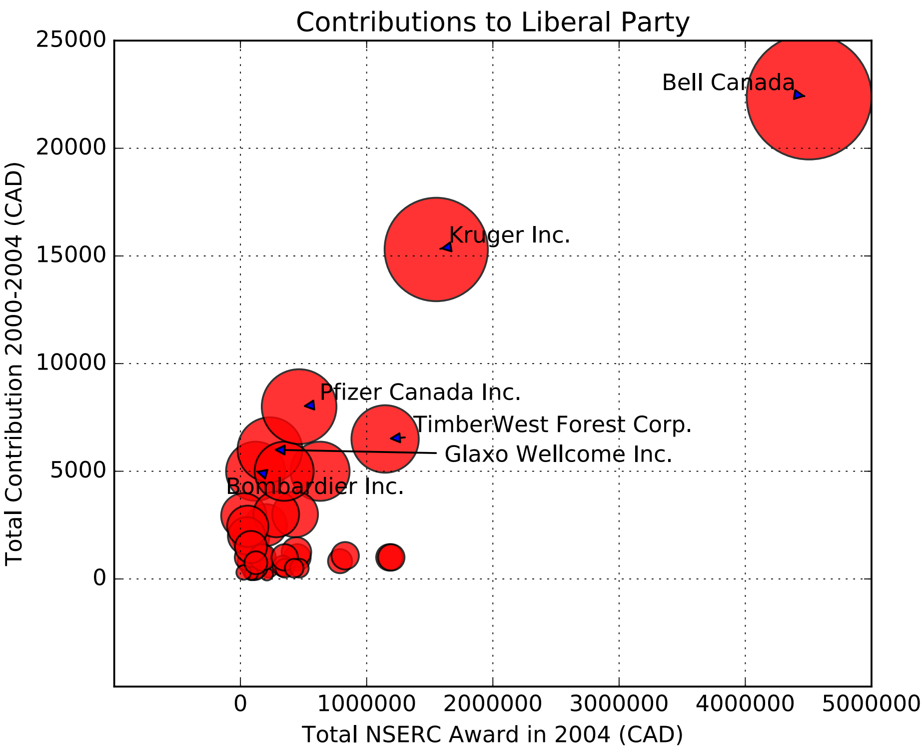
Each edge is an inclusion dependency

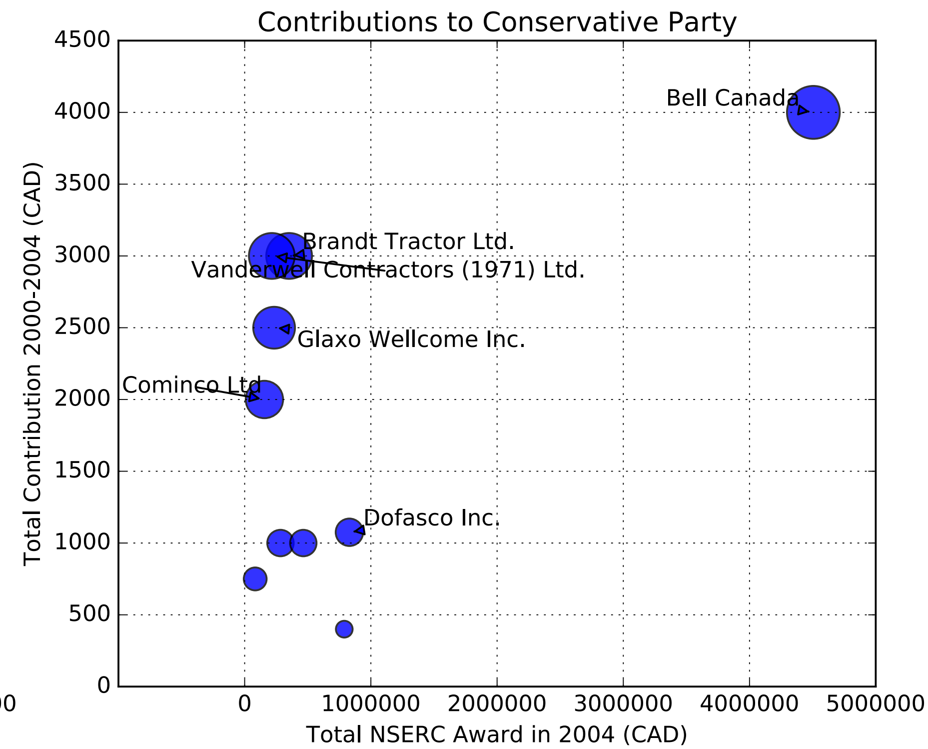
Traverse to the 4th degree from the yellow table

# Goal:  Enable Data Science



# Goal: Enable Data Science



# Data Discovery Example

| Fuel Type | Borough | Sector | KWh | Year | ... |
|---|---|---|---|---|---|
| Electricity | Barnett | Domestic | 62688 | 2015 | |
| Gas | Barnett | Domestic | 206438 | 2015 | |
| Railway Diesel | City of London | Transport | 2730044 | 2014 | |
| Oil | City of London | Domestic | 430078 | 2015 | |

- One example table
  - Greenhouse gas emissions in/around London
  - May have many attributes and tens/hundreds of thousands of tuples

# Table Join 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.

Table Repository

| Fuel Type | Borough | Sector | KWh |
|---|---|---|---|
| Electricity | Barnett | Domestic | 62688 |
| Gas | Barnett | Domestic | 206438 |
| Railway Diesel | City of London | Transport | 2730044 |
| Oil | City of London | Domestic | 430078 |

Query Table

| Borough | Population | Unemp | F.Unem |
|---|---|---|---|
| Barnett | 38900 | Low | 20 |
| Camden | 40000 | Low | 14 |
| City of London | 888000 | Medium | 20 |

Candidate Table

# Table Union Search

***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

| Fuel Type | Borough | Sector | KWh |
|---|---|---|---|
| Electricity | Barnett | Domestic | 62688 |
| Gas | Barnett | Domestic | 206438 |
| Railway Diesel | City of London | Transport | 2730044 |
| Oil | City of London | Domestic | 430078 |

Query Table

| County | Commodity | SecTyp | TotEmissions (MT CO2e |
|---|---|---|---|
| Benton | Gasoline | Transport | 20 |
| Kittitas | Fuel oil (1, 2..) | Hydro | 14 |
| Grays Harbor | Aviation Fuels | Domestic | 20 |
| Skagit | Liquified petroleum | Transport | 30 |

Candidate Table

# Outline

- Data Lake
  - What is it and why is it important?
  - New data management challenges
- Data Discovery
  - **Table Join Search:**
    - ✴**LSH Ensemble PVLDB 16, PVLDB 17**
    - ✴**JOSIE SIGMOD 19**
  - Table Union Search
- Open Questions

# Table Join Search

**Query Q**

| | | | |
|---|---|---|---|
| Electricity | Barnett | Domestic | 62688 |
| Gas | Barnett | Domestic | 206438 |
| Railway Diesel | City of London | Transport | 2730044 |
| Oil | City of London | Domestic | 430078 |

Query Table

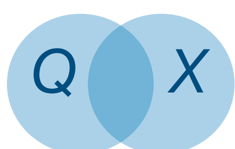
**Potential Answer X**

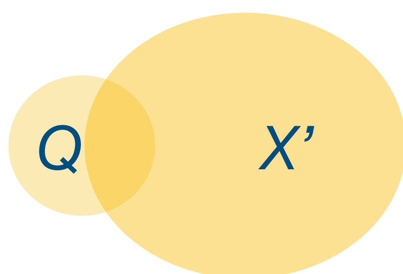
| | | | |
|---|---|---|---|
| Barnett | 38900 | Low | 20 |
| Camden | 40000 | Low | 14 |
| City of London | 888000 | Medium | 20 |
| … | | | |

Candidate Table

# Measuring Join Goodness?

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

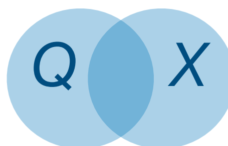


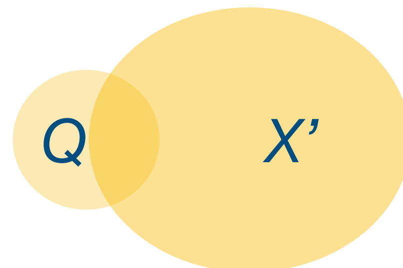




***Jaccard(Q,X)*** >> ***Jaccard(Q, X’)***

Same intersection size, but the Jaccard similarity is much smaller on the right

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

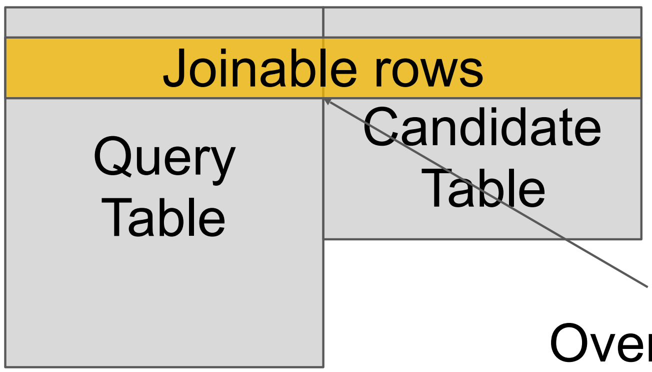


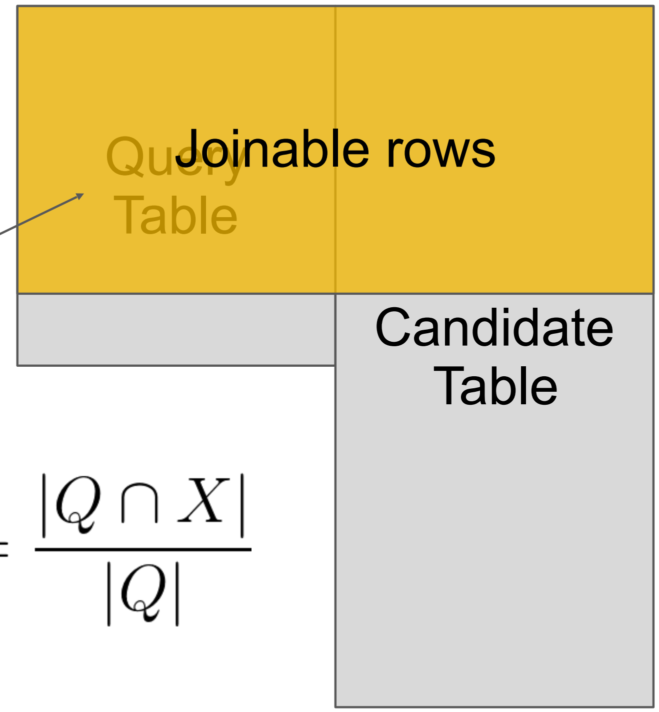




***Containment(Q,X)*** = ***Containment(Q, X’)***

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

# What is a good measure for joinability?




Overlap is a better measure for joinability

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

- Join Table Problem — find all ***X***:
  - ***Containment(Q,X) >= t****
- User specifies tolerance for error **t***

# MinHash LSH (Broder SEQ97)

$$X = \{x_1, x_2, ..., x_m\} \qquad Y = \{y_1, y_2, ..., y_m\}$$

$$h_0(X) = \min_{x \in X} f_0(x) \qquad h_0(Y) = \min_{y \in Y} f_0(y) \qquad P(h_0(X) = h_0(Y)) = \frac{|X \cap Y|}{|X \cup Y|}$$

Define a hash function for set, where $f_i$ is a hash function for value (e.g., SHA1)

$$h_1(X) = \min_{x \in X} f_1(x) \qquad h_1(Y) = \min_{y \in Y} f_1(y)$$


...

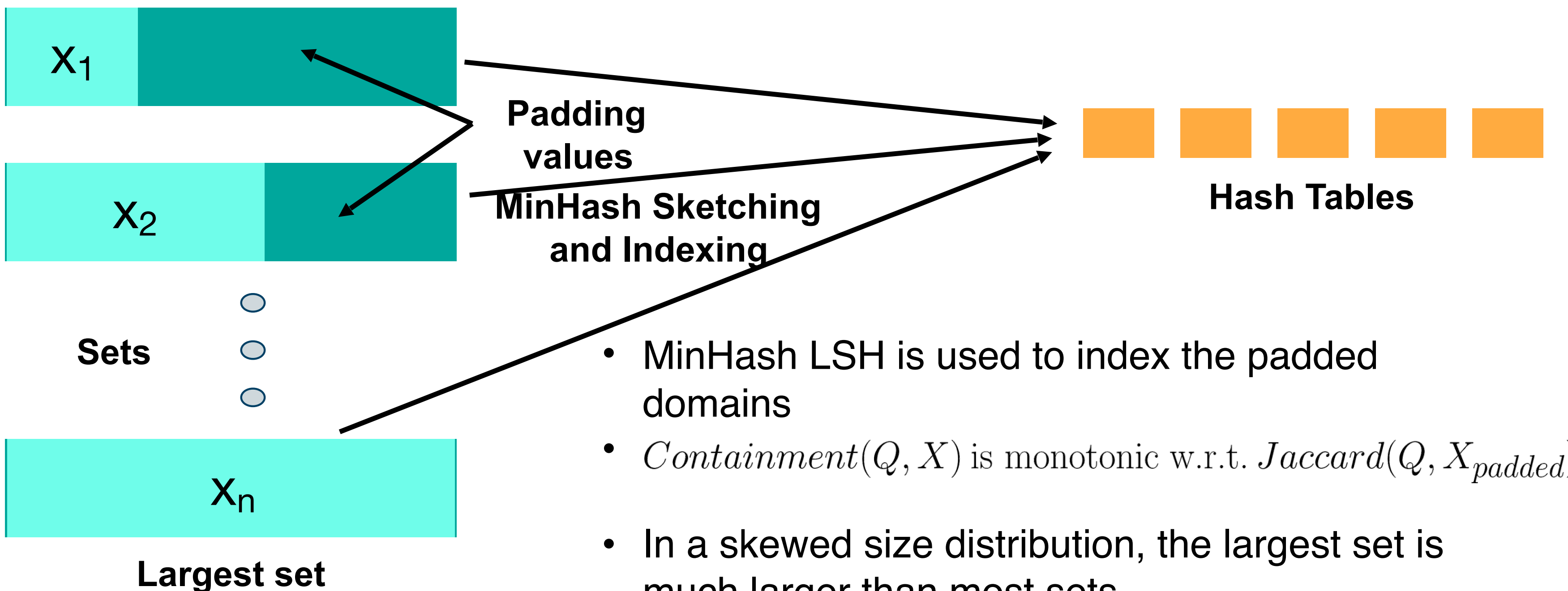
...

$$h_k(X) = \min_{x \in X} f_k(x) \qquad h_k(Y) = \min_{y \in Y} f_k(y)$$

**Indexing:** generate k such hash functions and insert sets into k respective hash tables
**Query:** hash the query set with k hash functions, and retrieve candidates from the k hash tables

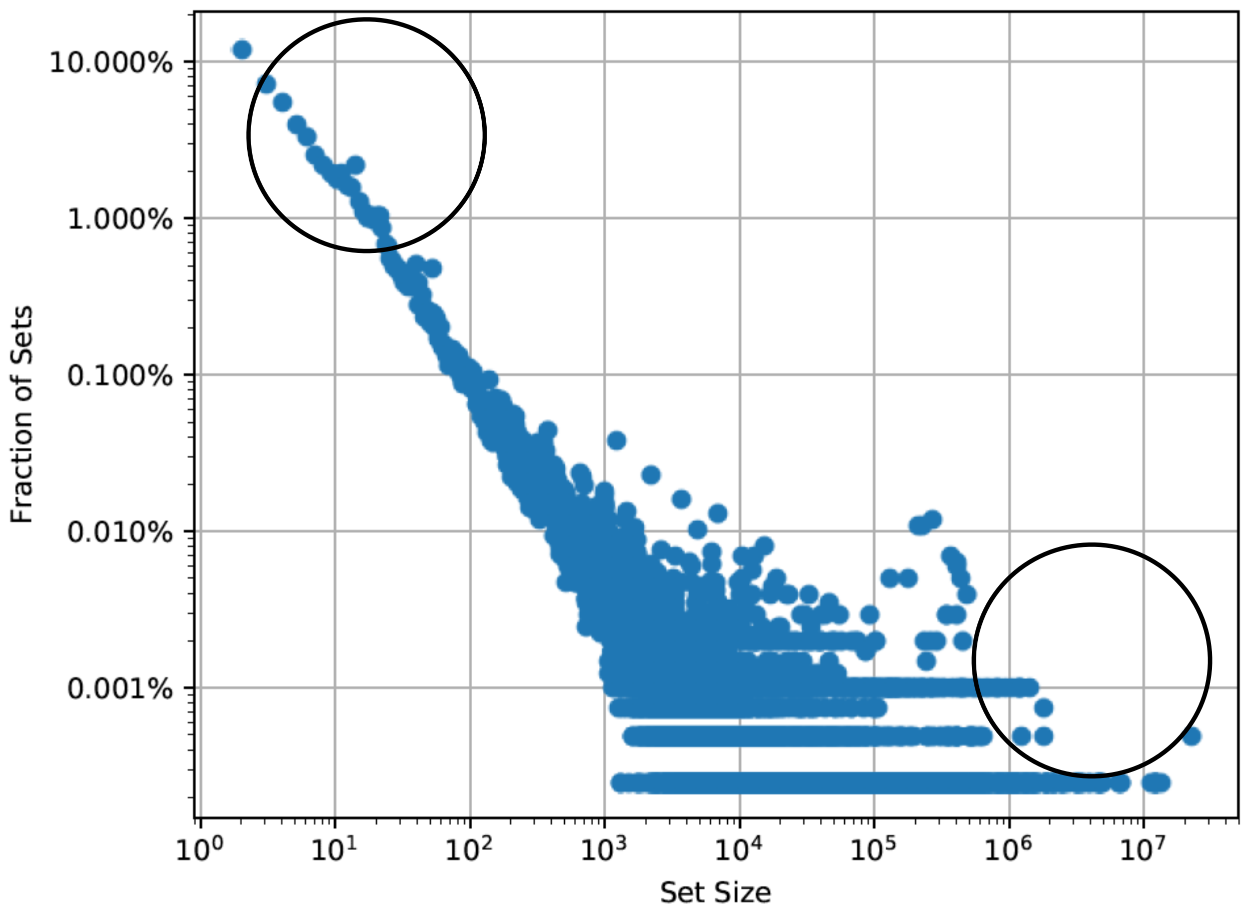
$$\boxed{\frac{|X \cap Y|}{|X \cup Y|} \approx \frac{Count(h_i(X) = h_i(Y))}{k}}$$

# Asymmetric MinHash (Shrivastava&Li WWW15)

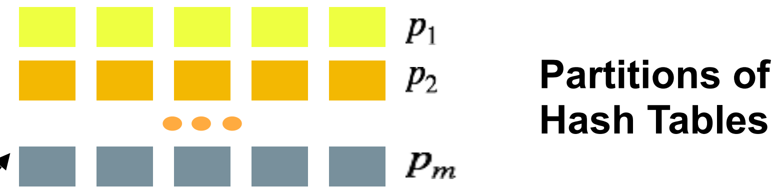


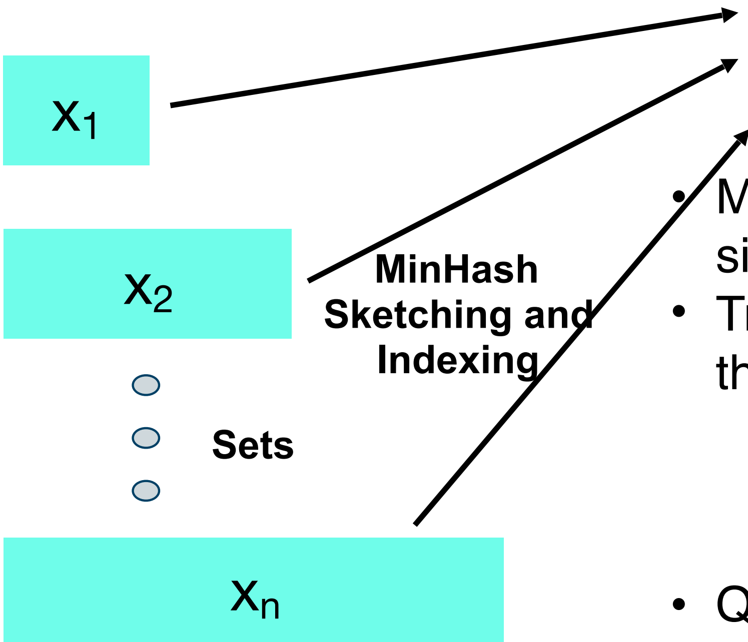

- MinHash LSH is used to index the padded domains
- $Containment(Q, X)$ is monotonic w.r.t. $Jaccard(Q, X_{padded})$
- In a skewed size distribution, the largest set is much larger than most sets
- Sketches contain mostly padding values — less likely to match a similar query set
- Hurts recall

# Open Data Attribute Cardinality Sizes



# LSH Ensemble (Zhu+ PVLDB16)



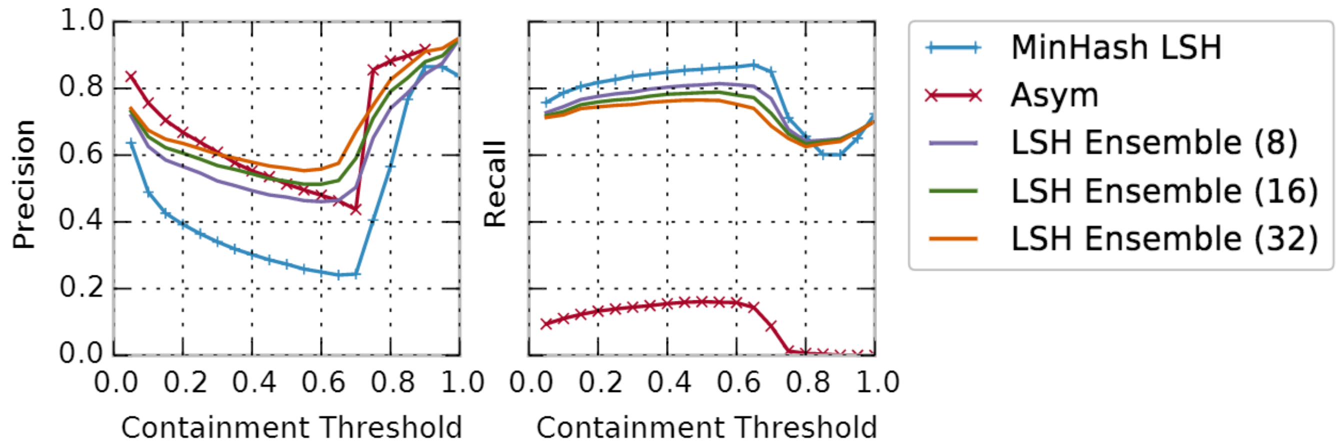




- Multiple MinHash LSH partitioned by increasing set size
- Transform a Containment threshold to a Jaccard threshold

$$Jaccard(Q, X) = \frac{Containment(Q, X)}{\frac{u}{|Q|} + 1 - Containment(Q, X)}$$

- Query each MinHash LSH index with the corresponding transformed threshold, in parallel
- Increasing number of partitions improves precision and speed
- **Optimal partitioning** strategy for power-law set size distribution (Zhu+ PVLDB16)

# LSH Ensemble Accuracy




- Creating more ***partitions*** leads to fewer false positives, while maintaining recall
- ***Asymmetric MinHash*** LSH has high precision, but low recall due to padding

# LSH Ensemble Query Performance

| Search Index | Mean Query (sec) | Precision (threshold=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 |

- Fewer false positive attributes to process (higher precision)
- Parallel querying over partitions

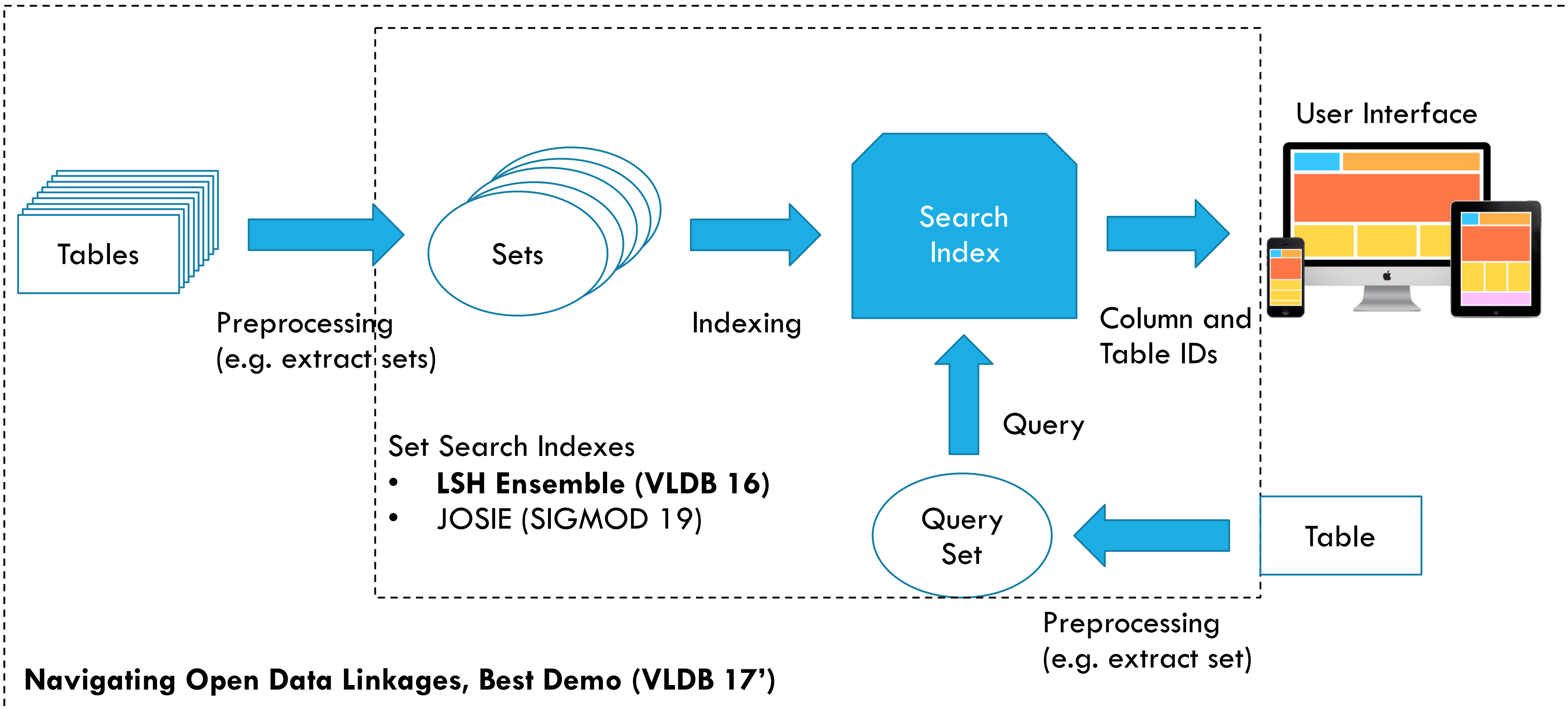
# Related Work

- Set Similarity Search
  - Doesn't scale to join search
  - Prefix Filter
    - ✳[Chaudhuri+ICDE06]
  - Position Filter
    - ✳[Xiao+WWW08]
  - Comparison
    - ✳[Mann+PVLDB16]
    - ✳[Behm+ICDE11,Li+ICDE08,Wang+SIGMOD12]

| DataSet | Avg Set Size | Max Set Size | Dictionary Size |
|---|---|---|---|
| AOL | 3 | 245 | 3.9M |
| ENRON | 135 | 3,162 | 1.1M |
| DBLP | 86 | 1,625 | 7K |
| WebTables | 10 | 17,030 | 184M |
| Open Data | 1.5K | 22M | 562M |

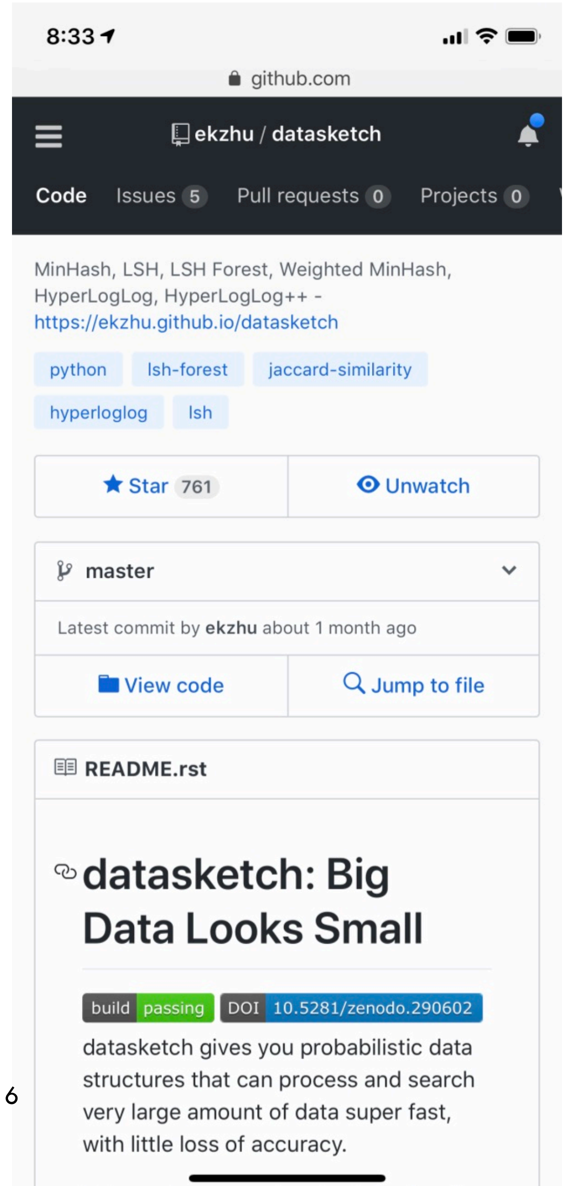
- Mass Collaboration Data Search
  - Relies on metadata
  - Linked Data/Microdata
    - ✳[Bizer+JSWIS09,Meusel+ISWC14]
  - Web Tables
    - ✳[Cafarella+ PVLDB08]
    - ✳**[Lehmberg+WWW16]**
    - ✳[Bhagavatula+IDEA13]
  - Table extension
    - ✳Infogather [Yakout+SIGMOD12]
    - ✳[Cafarella+PVLDB09]
    - ✳[DasSarma+SIGMOD12]
    - ✳Mannheim Search Join [Lehmberg+JWebSem15]

# A Joinable Table Search System



# LSH Ensemble[1]

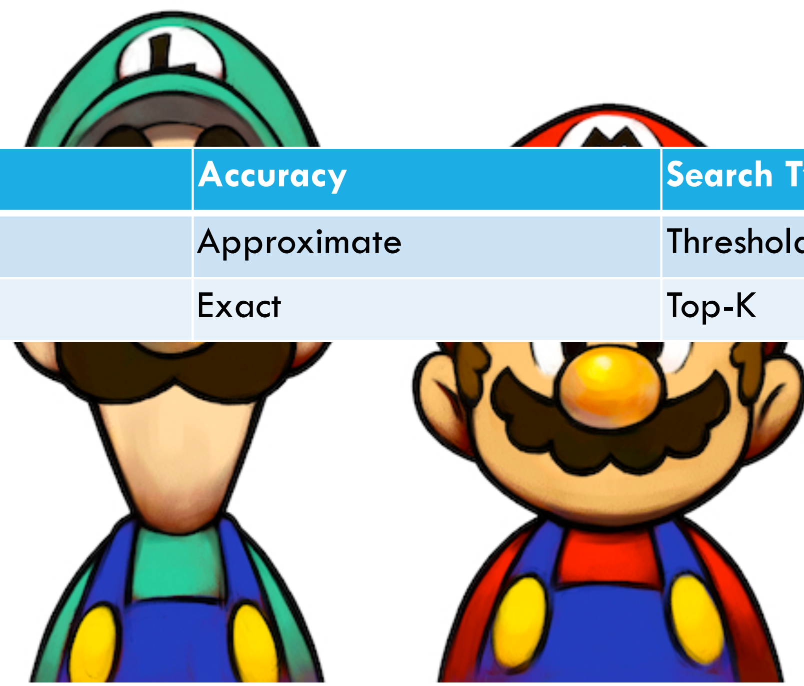
- Threshold-based search indexes
- Part of *datasketch* Python library[2], together with MinHash LSH
- The library is used by Google (TimeSketch), MIT (Aurum Data Discovery) and Stanford (NLP)
- Over 1100 stars on Github




[1] Erkang Zhu, Fatemeh Nargesian, Ken Pu, Renée J. Miller, "LSH Ensemble:L Internet-Scale Domain Search", VLDB 2016
[2] https://github.com/ekzhu/datasketch

# JOSIE vs. LSH Ensemble

They are different!



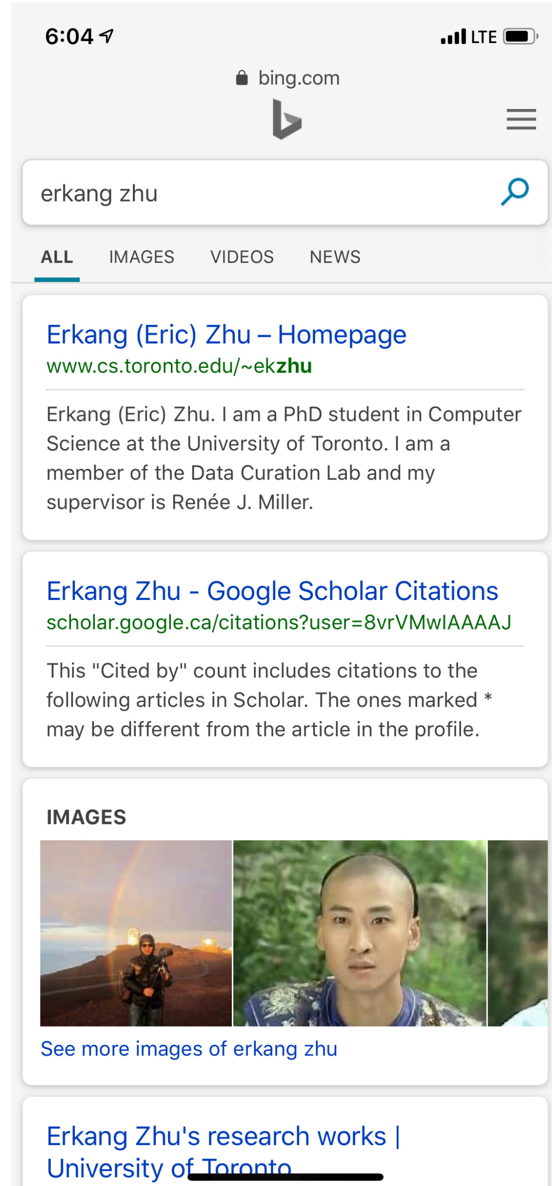
| Algorithm | Accuracy | Search Type |
|---|---|---|
| LSH Ensemble | Approximate | Threshold |
| JOSIE | Exact | Top-K |

# Top-K Search?

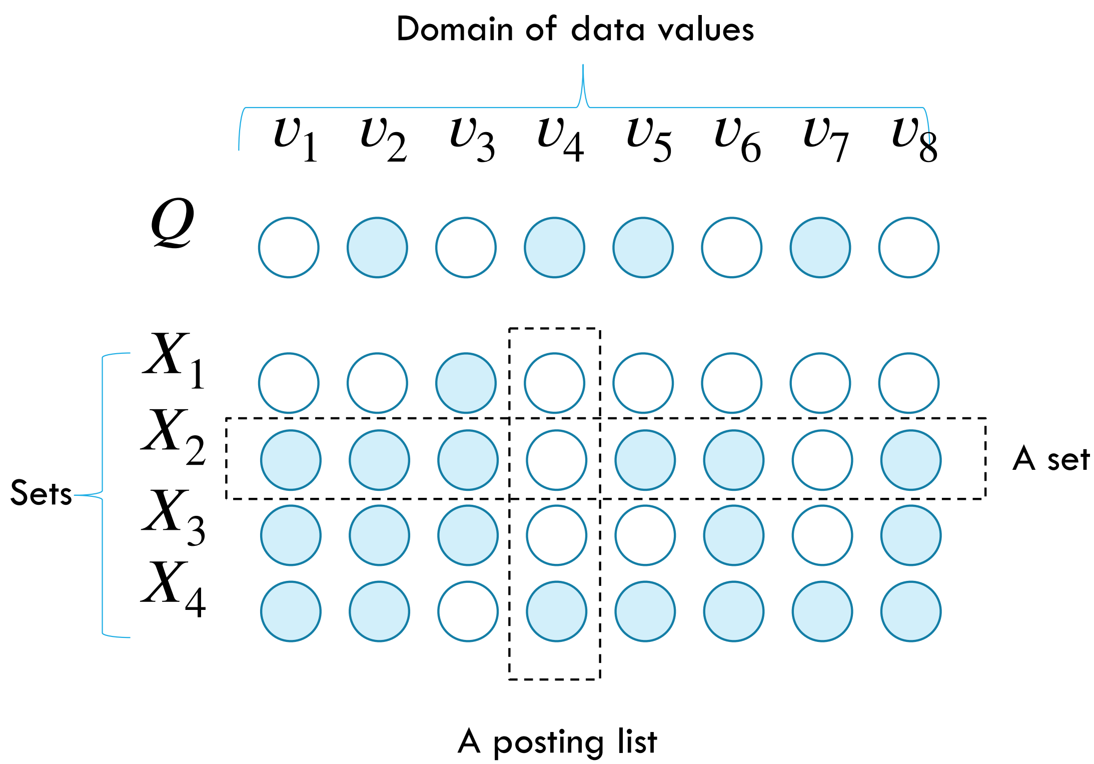
- Threshold search: user must specify a containment threshold
  - User may not know a good threshold
  - Even if they do it may produce no results or too many

- Top-K problem: just return the best $k$ results
  - No knowledge of relevance measure is required
  - We showed that for small k ( $< 20$), our exact top-k algorithm can be faster than LSH Ensemble with decreasing threshold hack!

- Use top-k for less sophisticated users and small $k$



# Inverted Index – A Matrix Perspective

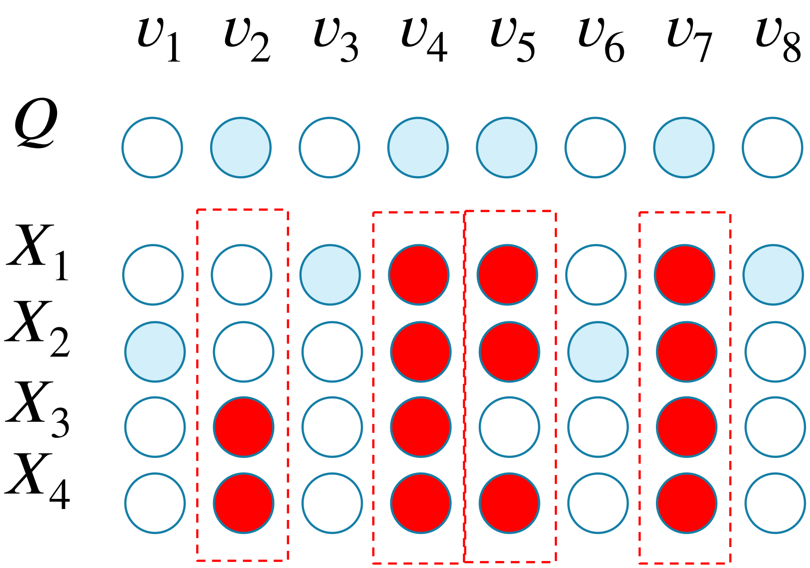




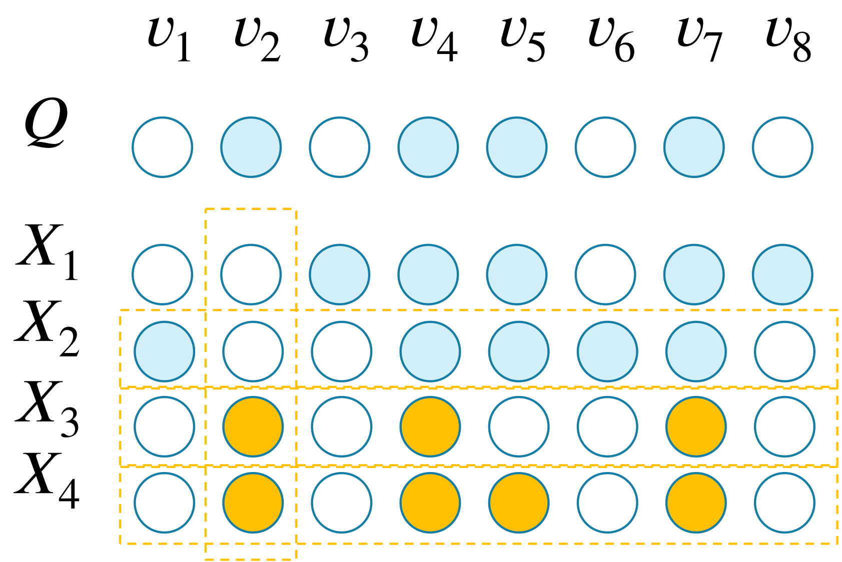
[1] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze. Introduction to Information Retrieval. Cambridge University Press, 2008

[2] Chuan Xiao, Wei Wang, Xuemin Lin, and Haichuan Shang. Top-k set similarity joins. In ICDE, pages 916–927, 2009.

# Baselines for Find Top-1




Posting list union[1]: reading all posting lists costs 13 values




Prefix filtering[2]: reading candidate sets costs 7 values


[1] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schu¨tze. Introduction to Information Retrieval. Cambridge University Press, 2008

[2] Chuan Xiao, Wei Wang, Xuemin Lin, and Haichuan Shang. Top-k set similarity joins. In ICDE, pages 916–927, 2009.

# Cost Matters in Data Lakes

| Dataset | # of Sets | Max Size | Avg. Size | # of Uniq. Values |
|---|---|---|---|---|
| Open Data* | 745K | 22M | 1,540 | 562M |
| WDC Web Table | 163M | 17K | 10 | 184M |
| AOL (Query Logs) | 10M | 245 | 3 | 3.9M |
| ERON (Emails) | 517K | 3,162 | 135 | 1.1M |
| DBLP (Bibliographies) | 100K | 1,625 | 86 | 6,864 |

Data Lakes

*215,393 Open Data tables from Canadian, US, and UK Open Data Portal

- ***Read Bottleneck***
  - Large number of sets and data values makes index access expensive
  - Large posting lists and sets are expensive to read

***JOSIE*** solves these issues using an adaptive cost based algorithm

# Outline

- Data Lake
  - What is it and why is it important?
  - New data management challenges
- Data Discovery
  - Table Join
  - **Table Union**
- Open Questions

# Table Union

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

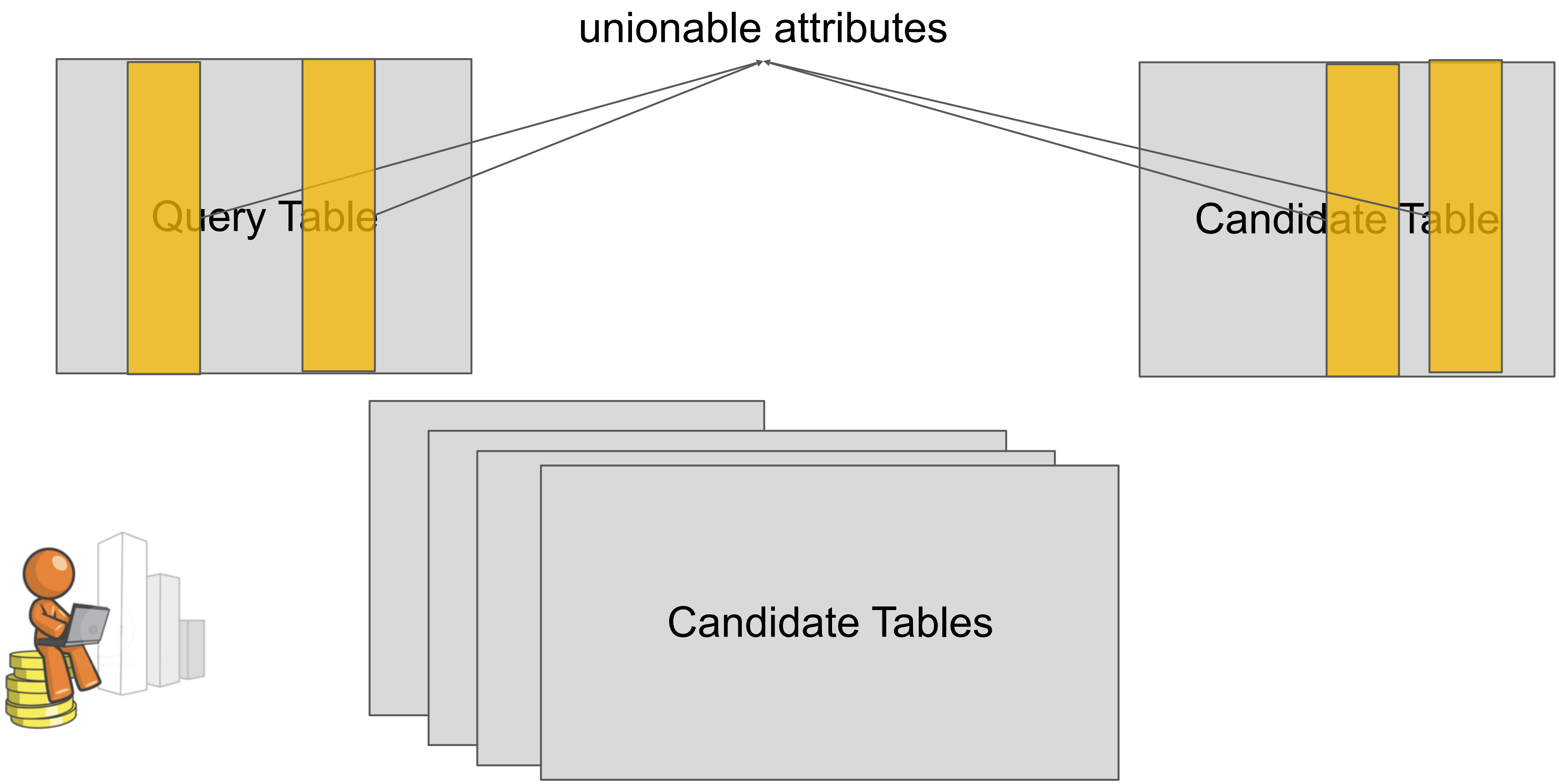
**Query Table**

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

**Candidate Table**

- Some attributes may overlap
- Some may refer to entities of common type
- Some may use semantically similar words

# Unionable Attribute Search



# Attribute Unionability

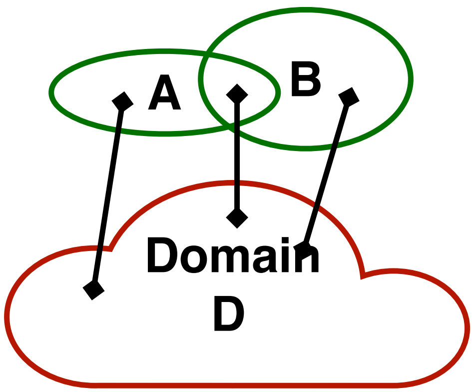
Natural Language Semantic Set

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

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

- Probabilistic Model
  - Attributes are samples drawn from the same domain
- Three types of attribute unionability/domains
  - Set, semantic, natural language

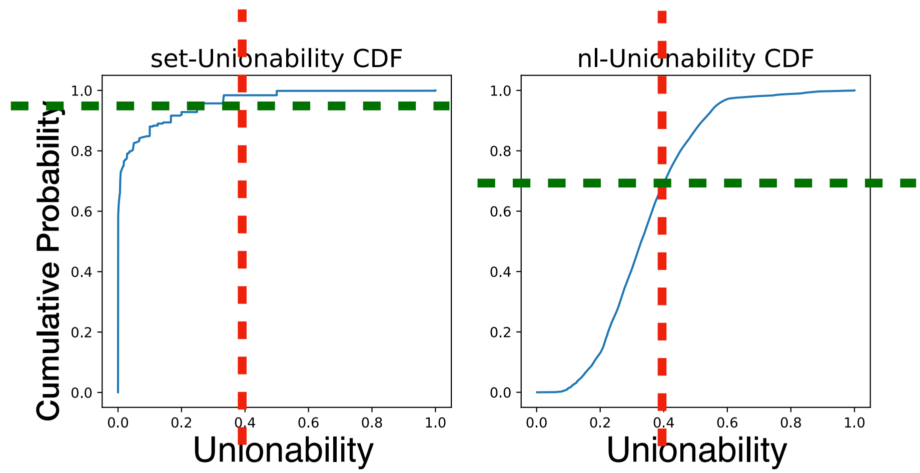
# Attribute Unionability




- Set and Semantic
  - D is set of values or set of ontology classes
- Natural Language
  - Convert values to word embeddings
  - Measure how likely the word embeddings are drawn from the same domain

Ensemble unionability

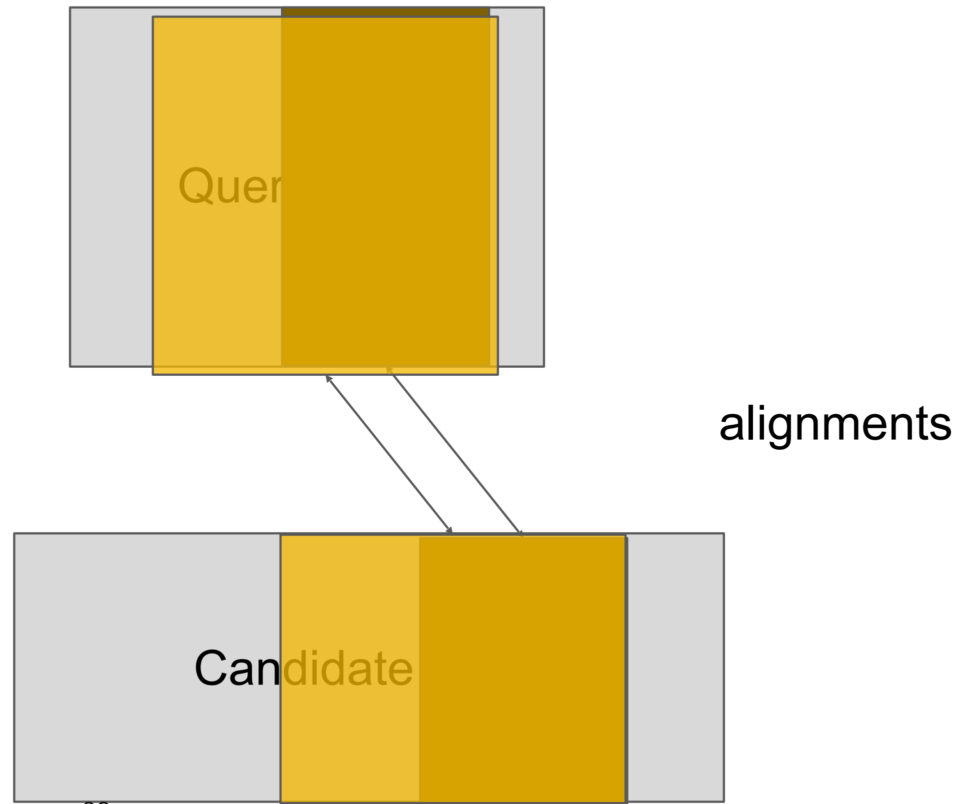
Measures are incomparable so define based on the corpus. How unexpected is a score given the corpus?



# Table Alignment

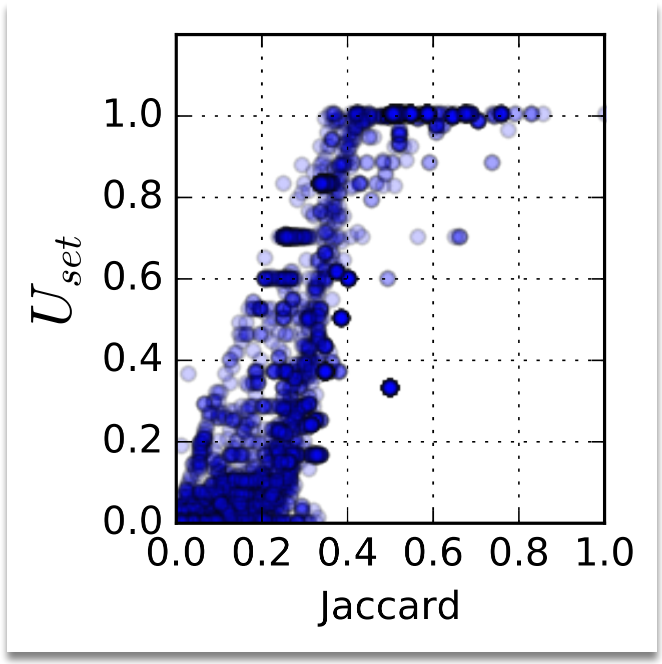
Given set of unionable attributes

when is an alignment of size *n* better than an alignment of size *n+1* attributes?



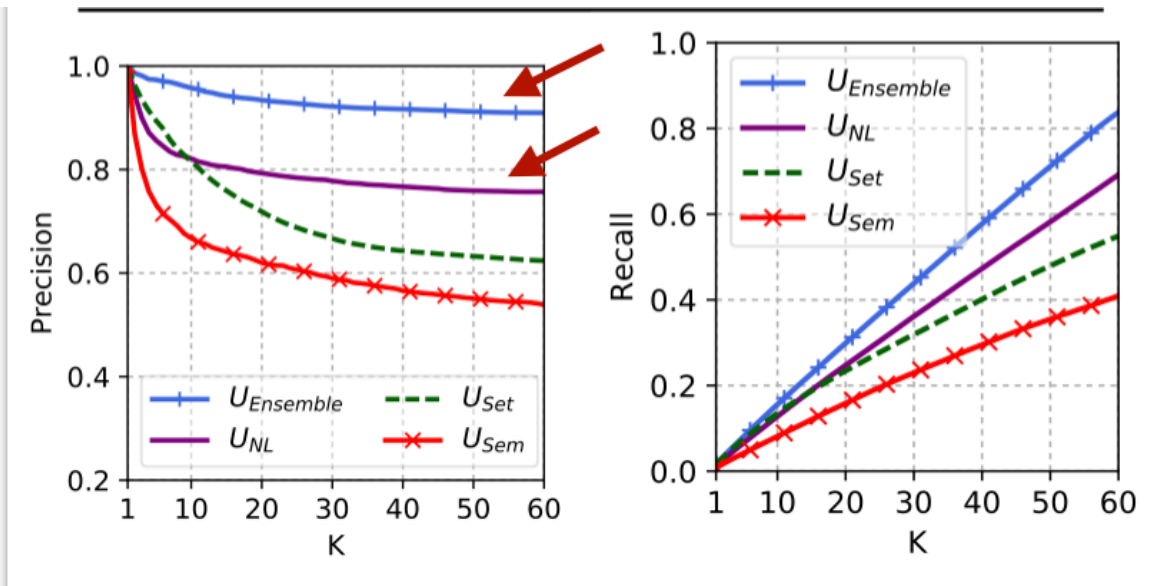
# Scaling Unionable Attribute Search

- Set and Semantic Unionability
  - Correlated with Jaccard
- Natural Language Unionability
  - Correlated with Cosine of topic vectors
- Use LSH indices to efficiently retrieve candidate attributes



# Evaluation Table Union on Open Data

- NL Unionability outperforms set and semantic (individually)
- Ensemble Unionability (uses all 3) best in accuracy
- Defined as top-K search
  - User defined threshold for unionability is not intuitive




- Semantic Unionability
  - Uses Open Ontology: YAGO
    - ✳ [Suchenek+WWW07]

- Public Table Union Search Benchmark

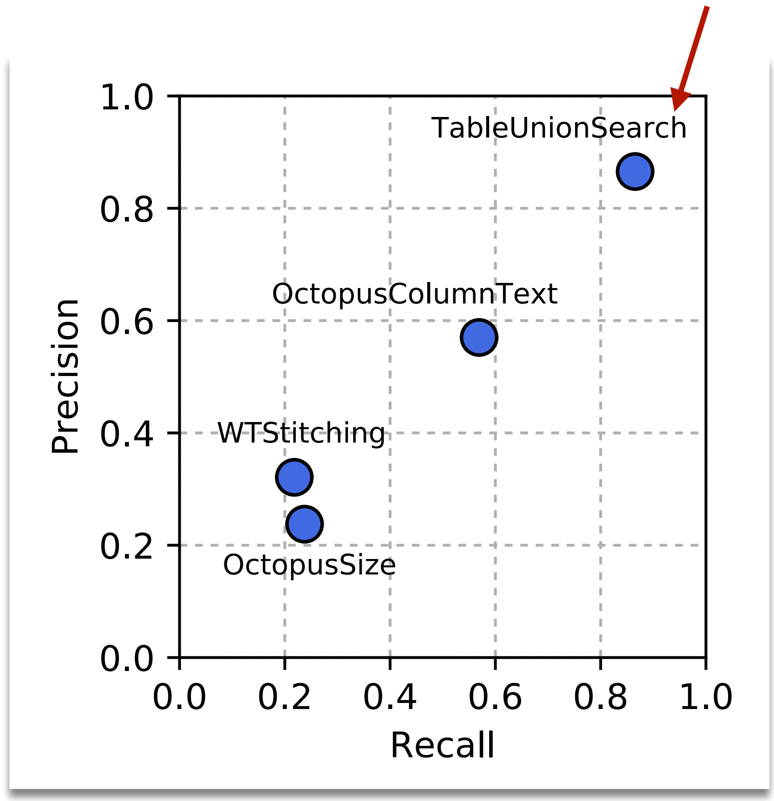
**https://github.com/RJMillerLab/table-union-search-benchmark**

# Using Search on Mass Collaboration Data

- Search on metadata
  - Schema Matching — attributes that matched can be "unioned"
    - ✳ [Ling+IJCAI13], [Lehmberg and BizerPVLDB17]
  - Schema plus keyword description of each attribute
    - ✳ [Pimplikar&SarawagiPVLDB12]
- Keyword Search and Clustering of Tables
  - Tables in the same cluster are "unionable"
    - ✳ Octopus [Cafarella+PVLDB09]
- Entity-table search
  - Union tables that share a subject attribute (entities of same type)
    - ✳ [Das Sarma+SIGMOD12]

# Comparison to WebTable Union

- Octopus [Cafarella+PVLDB09]
  - Keyword search; cluster
  - Attribute Similarity
    - Size: avg length values
    - ColumnText: tf-idf of values
- Stitching [Lehmberg&BizerPVLDB17]
  - Instance-based schema matching
- Entity-Complement [DasSarma+SIGMOD12]
  - Union entity tables w/ same subject



# Outline

- Data Lake
  - What is it and why is it important?
  - New data management challenges
- Data Discovery
  - Table Join
  - Table Union
- Open Questions

# Open Problems

- Near-term:  analysis-driven data discovery
  - Bags vs. Sets
  - Multi-attribute join search
  - Finding tables that join and contain new information
  - Incorporating entity-resolution into scalable search
  - Search over quantities (with different measures)
  - Schema inference

# Vision

- ***Query discovery*** over massive data lakes
  - Finding not only the tables that can be integrated but also the best way to transform and integrate them meaningfully
  - Lessons from mapping discovery
- Data Quality over Open Data
  - Are "*Principles of Open Data"* being achieved?
    - ✴Truth finding has been studied over mass collaboration data [Pochampally+SIGMOD14]
    - ✴Can we quantify when open data is accurate, complete, primary?
  - Shazia Sadiq+, "Data Quality: The Role of Empiricism", SIGMOD Record 2018

# Acknowledgments


- This work was done in collaboration with Professor Ken Q. Pu, UOIT &
  - Erkang (Eric) Zhu
    - ✳ Table Join and Open Data Search
    - ✳ PhD April 2019, now Researcher at MSR
  - Fatemeh Nargesian
    - ✳ Table Union Search
    - ✳ PhD June 2019, Asst Professor University of Rochester