ACM SIGMOD
PODS 2022
Philadelphia, PA, USA

Responsible Data Integration: Next-generation Challenges

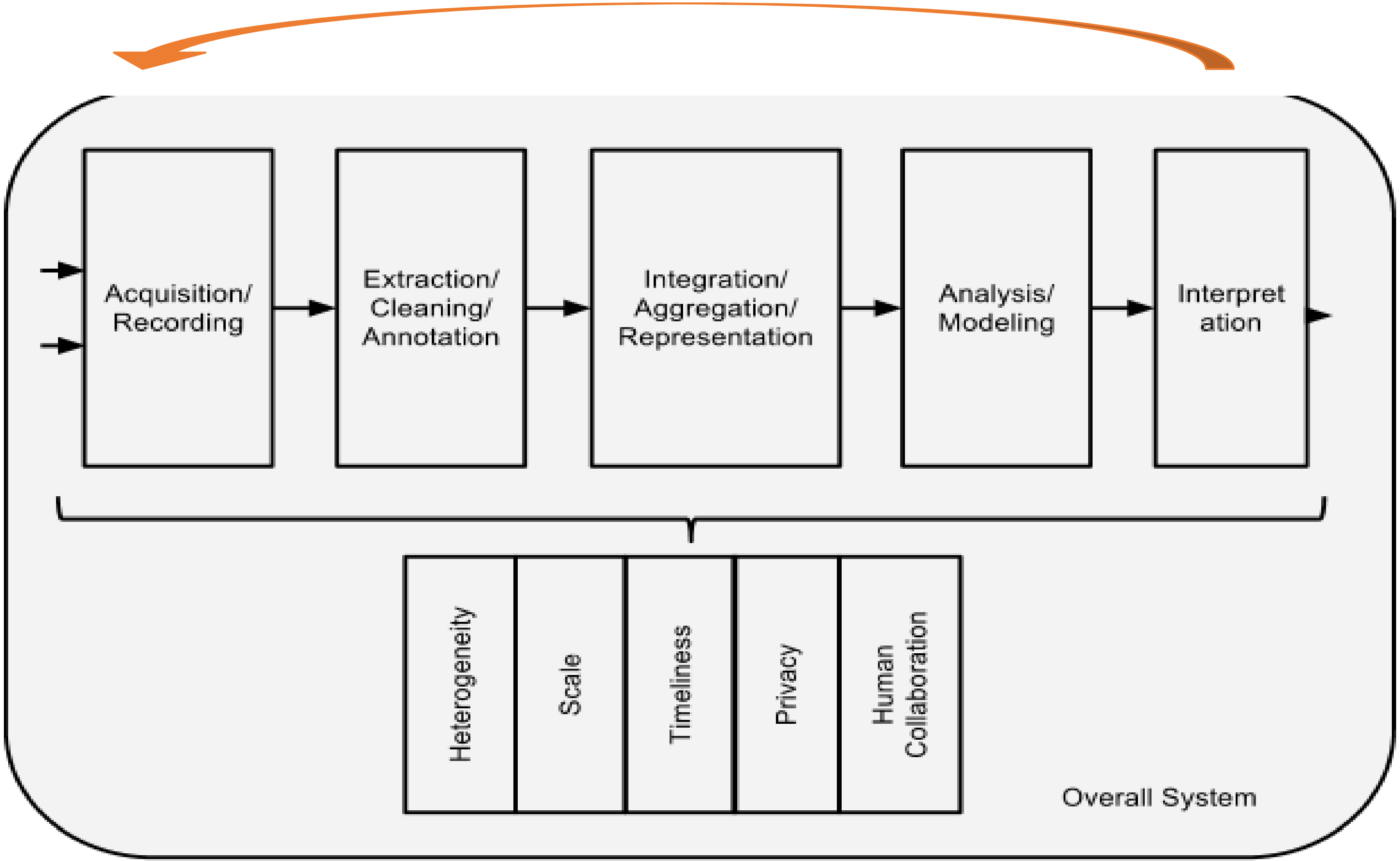
Fatemeh Nargesian
University of Rochester

Abolfazl Asudeh
University of Illinois Chicago

H. V. Jagadish
University of Michigan

AI Needs Data

- Data usually obtained through multiple sources.
- Goes through significant process of cleaning and integration.



The Big Data Pipeline (CACM 2014)

AI Needs Data

- Data usually obtained through multiple sources.
- Goes through significant process of cleaning and integration.

AI Needs Good Data

- Data usually obtained through multiple sources.
- Goes through significant process of cleaning and integration.

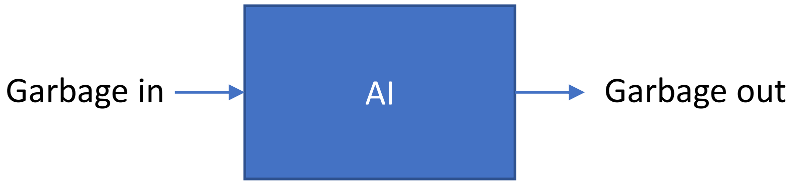
- Responsible AI requires care in the data pipeline.

OUTLINE

- Part I: next-generation requirements of responsible AI (15 mins)
- Part II: revisiting data integration (25 mins)
- Part III: fairness-aware data integration (25 mins)
- Part VI: open problems (10 mins)
- QA (5 mins)
- QA after each part (2-3 mins)

PART I:
RESPONSIBLE AI: NEXT GENERATION REQUIREMENTS

Data as the central component of data-driven systems



- First step to achieve Responsible AI:

Responsible Data

- New requirements for the data pipeline, including ***data integration***

***we focus on requirements specific to responsible AI. Other requirements such as environmental impact are out of our scope.**

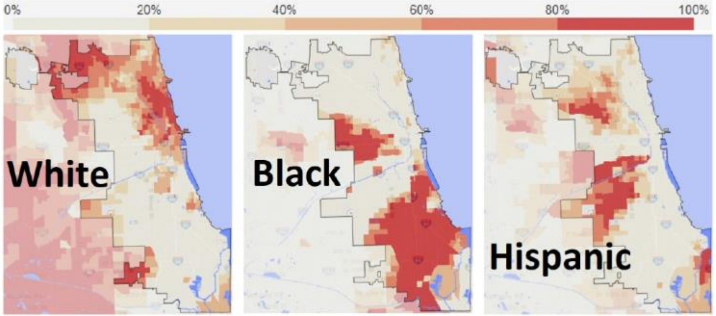
1- Underlying Distribution Representation

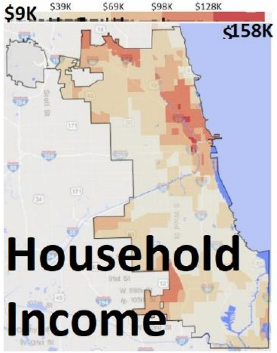
- Standard Assumption of AI: training data is *i.i.d random samples* drawn from the distribution that query points follow

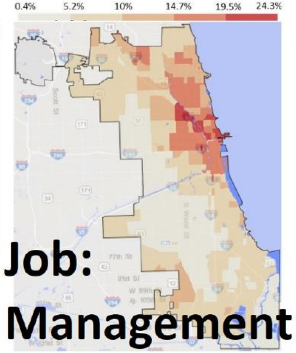
1. Not always easy to satisfy
2. Not easy to verify

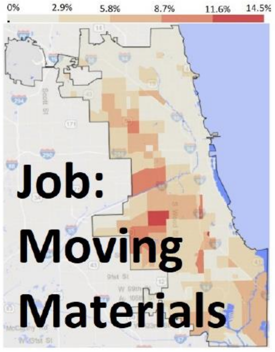
Not easy to satisfy

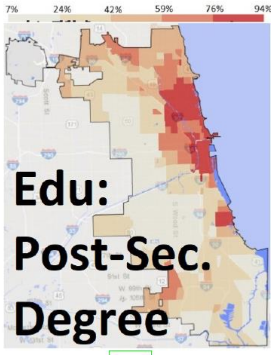
- In Social Data:
  Local distributions do not represent the global distribution









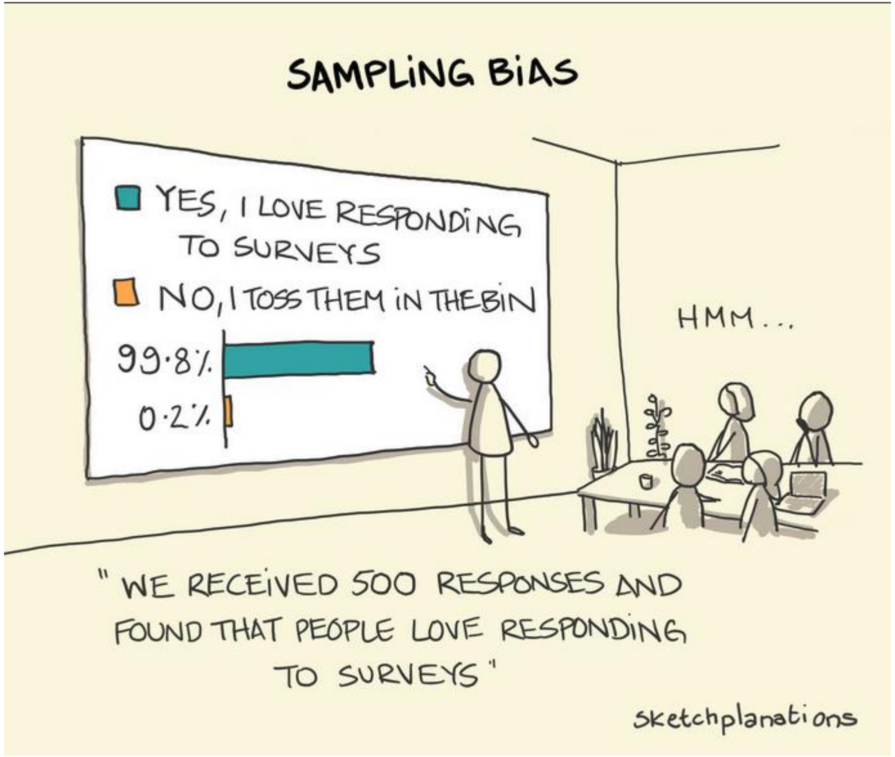


Not easy to satisfy

- Even if selected randomly

- Suppose surveys sent out to carefully chosen random sample

- Only a fraction of surveys returned


SAMPLING BIAS
🟦 YES, I LOVE RESPONDING TO SURVEYS
🟧 NO, I TOSS THEM IN THE BIN
99.8%
0.2%
HMM...
" WE RECEIVED 500 RESPONSES AND FOUND THAT PEOPLE LOVE RESPONDING TO SURVEYS "
sketchplanations

Not easy to verify

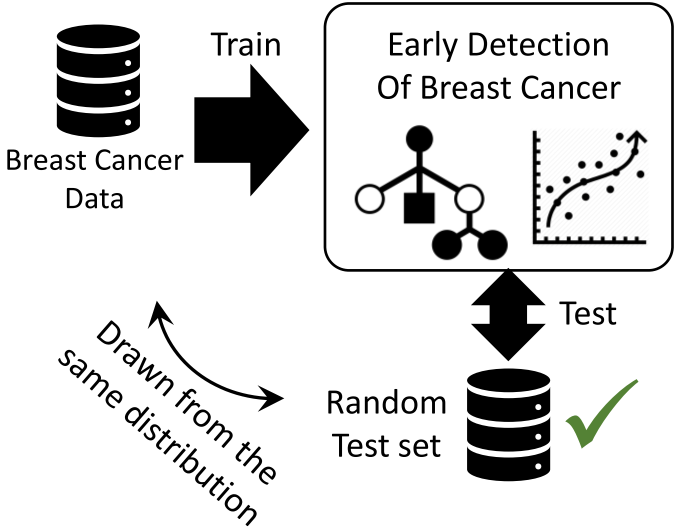
- Underlying distribution is usually unknown

  ➔ Challenging to verify that collected data is unbiased

2- Group Representation

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

Example: Early Detection of Breast Cancer



Breast Cancer Data

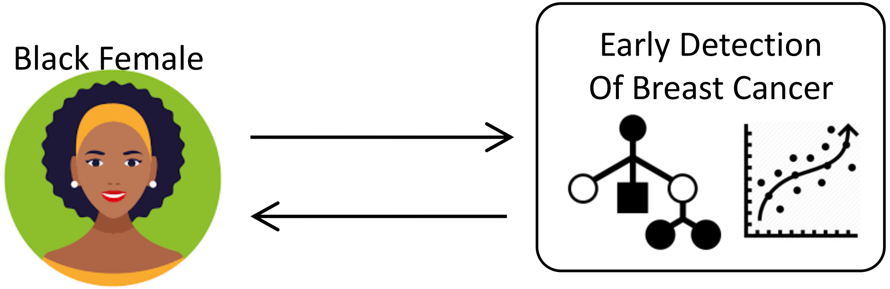
Train

Early Detection Of Breast Cancer

Test

Random Test set ✓

Drawn from the same distribution



Black Female

Early Detection Of Breast Cancer

(Lucky): Similar “behavior” → 👍

(Unlucky): Diff. “behavior” → 👎

Group Representation Requirements

- Representation Ratio: requires (almost) equal representation of different groups

- Data Coverage: (more liberal). “enough” representation of different groups

- A Survey on Techniques for Identifying and Resolving Representation Bias in Data, Shahbazi et al., CoRR, 2022.
- Data preprocessing to mitigate bias: A maximum entropy based approach, Celis et al., PMLR, 2020.
- Assessing and Remedying Coverage for a Given Dataset, Asudeh et al., ICDE, 2019.

3- Unbiased and Informative Features

- data set: a collection of attributes (features) used for decision making

$$\boldsymbol{x} = \{x_1 \dots x_m\}$$

  - may also contain one (or more) target attribute (labels) $\boldsymbol{y}$

  - sensitive attributes $\boldsymbol{s}$ such as race and gender

Sensitive attributes

- Sensitive attributes are required to achieve responsible AI
- often challenging to collect such information

Example: users of a shopping website

- Usually do not collect the sensitive information of the users
- Proxy attributes such as “name” may be used to specify the demographic information:
  - Asian names are gender-neutral

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}

- *Low* correlation between the features and the sensitive attributes

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

4- 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 <u>**minority**</u> value may **significantly skew** the average

Missing values resolution issue

- Downstream approaches to resolve missing (and unclean) values can further increase bias in data

- Example (two resolution strategies)
  1. rows with missing values are removed
     - removing a minority row further decreases the data coverage
  2. missing values are replaced with the column average
     - the average value is mostly affected by majorities → bias further increased

5- 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

Additional data profiling requirements

- Underlying distribution the data has been collected from
- Existing biases:
  - groups it fails to represent
  - features that are biased
- Information related to correctness and completeness of data.

Part II:
Revisiting Data Integration

Related Integration Tasks

- Data discovery
- Data cleaning
- Random sampling over join
- Data profiling

DATA DISCOVERY

- Keyword-based (IR-based)
  - Google's data set search engine

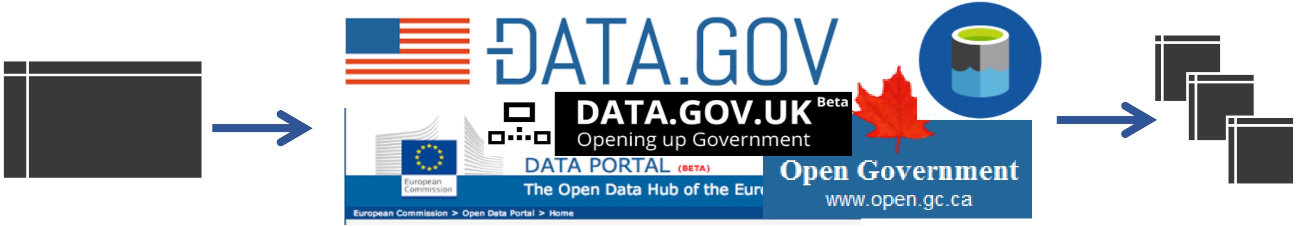
keywords: greenhouse gas emission inventory

DATA.GOV
DATA.GOV.UK Beta
Opening up Government
European Commission
DATA PORTAL (BETA)
The Open Data Hub of the Eur
European Commission > Open Data Portal > Home
Open Government
www.open.gc.ca

Google Dataset Search: Building a search engine for datasets in an open Web ecosystem, Brickley et al., WWW, 2019.
Google Dataset Search by the Numbers, Benjelloun et al., ISWC, 2020.

DATA DISCOVERY

- Table-based (input: a table or an attribute)
  - Table union search (group representation and underlying distribution)
  - Table join search (informative features and obtaining sensitive attributes)
- Data discovery mostly focuses on efficiency and accuracy



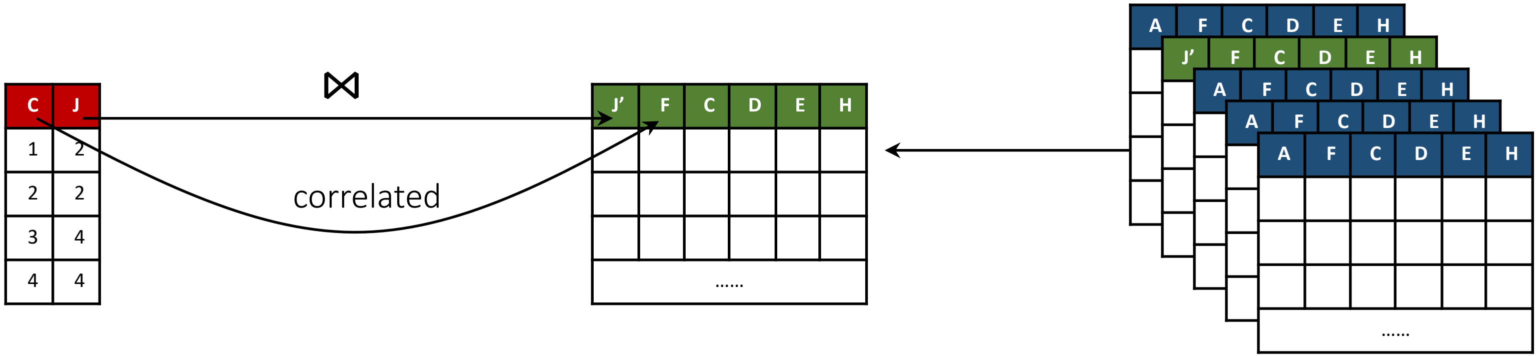
JOSIE: Overlap Set Similarity Search for Finding Joinable Tables in Data Lakes, Zhu et al., SIGMOD, 2019.
Auctus: A Dataset Search Engine for Data Discovery and Augmentation, Castelo et al., PVLDB, 2021.
Lazo: A Cardinality-Based Method for Coupled Estimation of Jaccard Similarity and Containment, Fernandez et al., ICDE, 2019.

DATA DISCOVERY

- It is rarely that case that one source satisfies the distribution requirements
- Data discovery enables collecting data for sensitive attributes and
  - Satisfying underlying distribution representation
  - Satisfying group representation
  - Collecting unbiased and informative features

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.

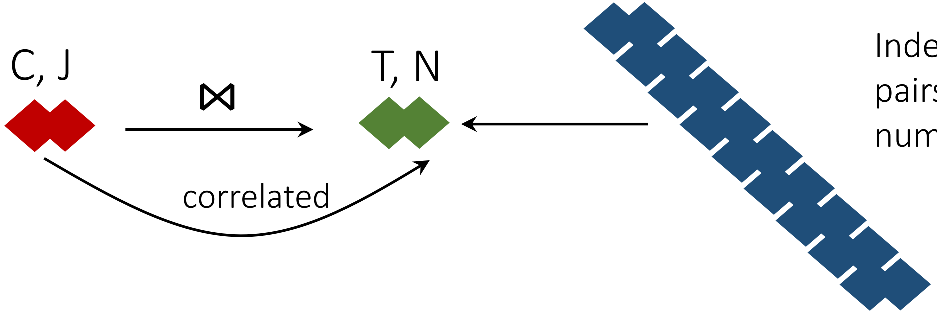


C	J
1	2
2	2
3	4
4	4

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



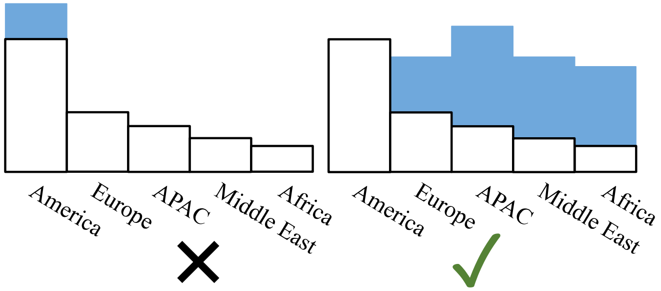
Indexed synopses on pairs of categorical and numerical attributes

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

Slice Discovery

- Identifying problematic slices of data that cause bias
- Selectively acquiring the right amount of data for problematic slices
  - possibly different amounts of data per slice s.t. accuracy and fairness on all slices are optimized

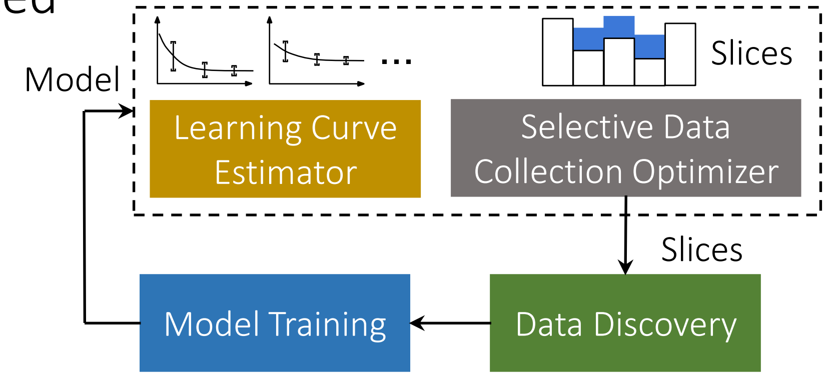
Slice: feature + value



Slice Tuner: A Selective Data Acquisition Framework for Accurate and Fair Machine Learning Models, Tae et Whang, SIGMOD, 2021.

SLICE DISCOVERY

- Iterative and efficient update of slices’ learning curves with new data
- Slices my be inaccurate when no enough data
- Slices may be correlated



Slice Tuner: A Selective Data Acquisition Framework for Accurate and Fair Machine Learning Models, Tae et Whang, SIGMOD, 2021.

DATA CLEANING

- Rich body of work for obtaining complete and correct datasets
- Dirty and incomplete data hurts minorities more

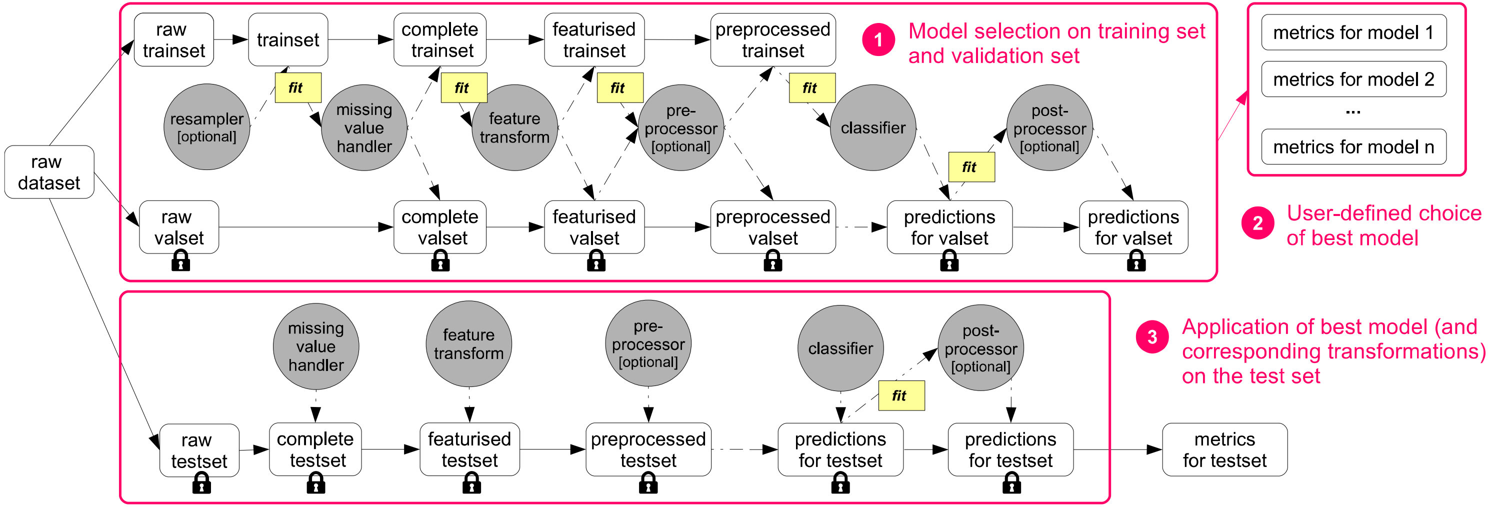
Fairness in Data Preparation

- FairPrep
  - promotes data to first-class citizens in fairness-related studies
  - a design and evaluation framework with fairness-specific evaluation metrics
- Observations
  - high variability of the fairness and accuracy outcomes might be an artifact of the lack of hyperparameter tuning on baselines
  - lack of feature scaling can lead to the failure to learn a well-working model
  - …

FairPrep: Promoting Data to a First-Class Citizen in Studies on Fairness-Enhancing Interventions, Schelter et al., EDBT, 2020.
AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias, Bellamy et al., FAT*ML, 2019.

Fairness in Data Preparation

- FairPrep suggests an evaluation run consisting of three phases
  1. Learn different models
  2. Compute performance/accuracy-related metrics on the validation set
  3. Compute metrics for the best model on the held-out test set



Model selection on training set and validation set

User-defined choice of best model

Application of best model (and corresponding transformations) on the test set

FairPrep: Promoting Data to a First-Class Citizen in Studies on Fairness-Enhancing Interventions, Schelter et al., EDBT, 2020.

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.

Iid Sampling over Join

- 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
…	…						

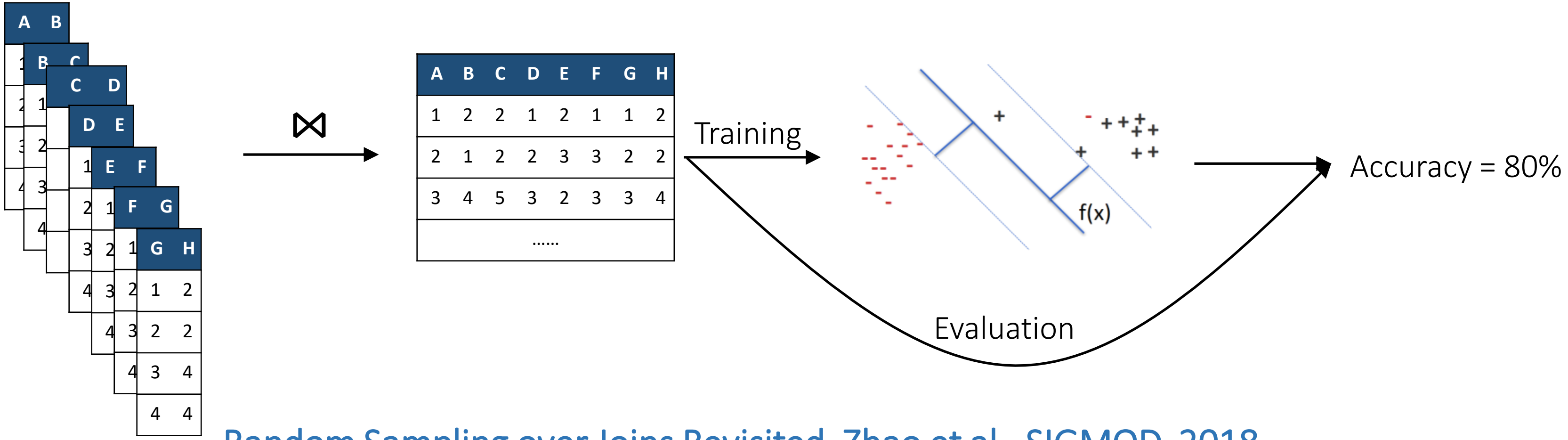
IID Sampling over Join

```
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.



G	H
1	2
2	2
3	4
4	4

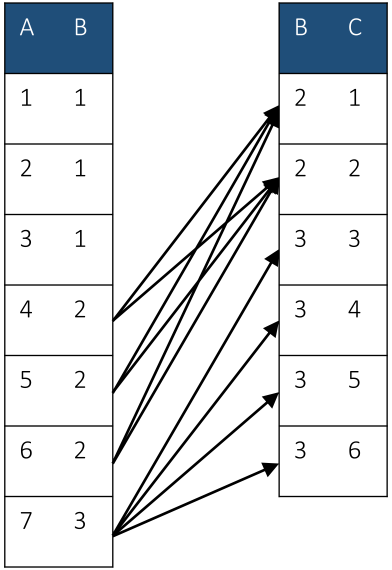
A	B	C	D	E	F	G	H
1	2	2	1	2	1	1	2
2	1	2	2	3	3	2	2
3	4	5	3	2	3	3	4
……							

Accuracy = 80%

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) ⋈ 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.

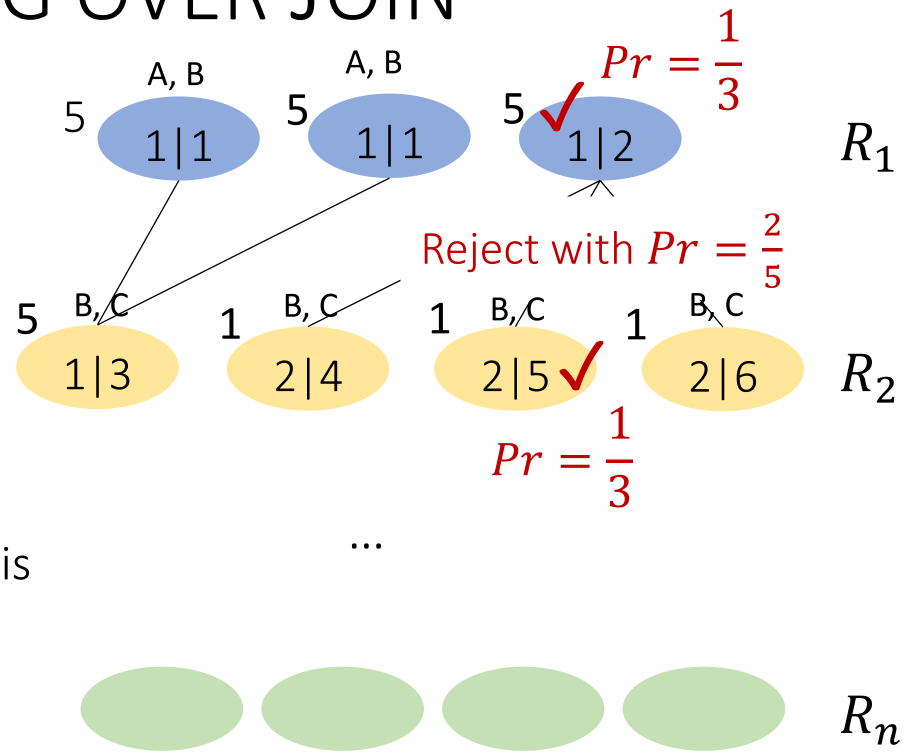
IID Sampling over Generic Join Paths

- 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
- Use a surrogate weight $W(t)$ if $w(t)$ is not available. $W(t)$: upper bound of $w(t)$
- Reject with prob. $\frac{W(t) - \sum_{t' \in ch(t)} W(t')}{W(t)}$



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

Data Profiling: Nutritional labels for interpretability

- Interpretability as an essential requirement of Responsible AI.

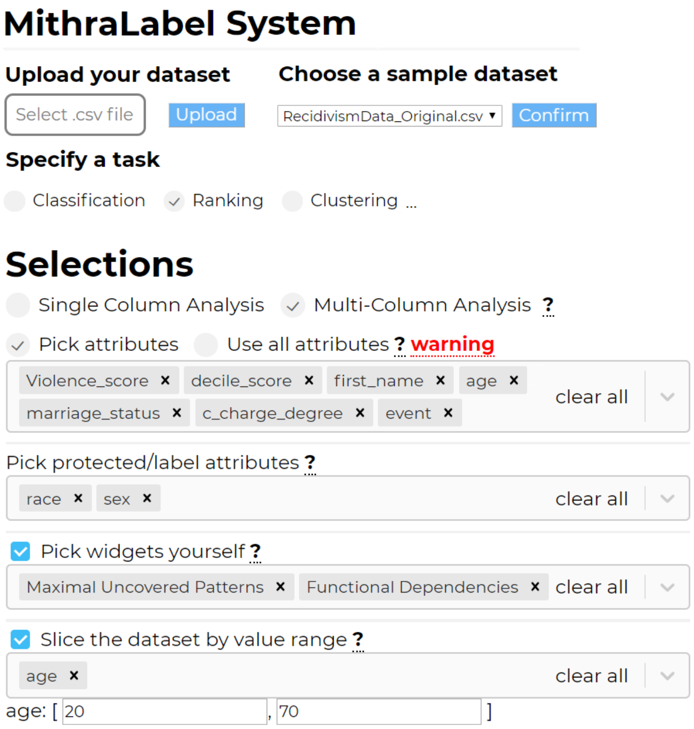
- Drawing an analogy to the food industry, where simple, standard labels convey information about the ingredients and production processes:
    - a nutritional label is a set of automatically constructed visual widgets, each conveying standardized information about “fitness for use” of data or the evaluators

Nutritional Labels for Data and Models, Stoyanovich et al., Data Eng. Bull, 2019.

MithraLabel DataSheets

MithraLabel: Flexible dataset nutritional labels for responsible data science, Sun et al., CIKM, 2019.

Datasheets for datasets, Gebru et al., CoRR, 2018.



MithraLabel System

Upload your dataset

Select .csv file | Upload

Choose a sample dataset

RecidivismData_Original.csv ▾ | Confirm

Specify a task

○ Classification ✓ Ranking ○ Clustering ...

Selections

○ Single Column Analysis ✓ Multi-Column Analysis ?

✓ Pick attributes ○ Use all attributes ? warning

Violence_score × decile_score × first_name × age × marriage_status × c_charge_degree × event × clear all

Pick protected/label attributes ?

race × sex × clear all

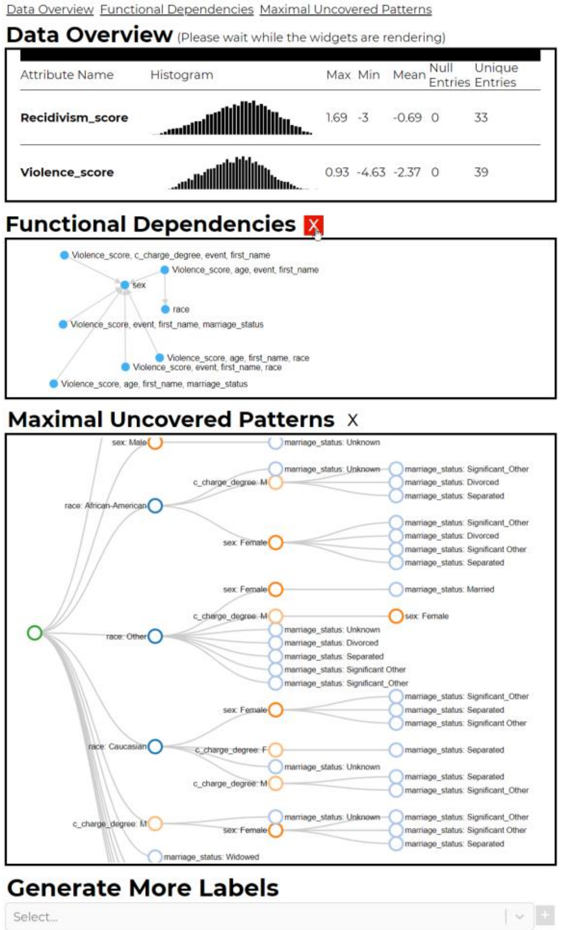
☑ Pick widgets yourself ?

Maximal Uncovered Patterns × Functional Dependencies × clear all

☑ Slice the dataset by value range ?

age × clear all

age: [20 , 70]



Data Overview Functional Dependencies Maximal Uncovered Patterns

Data Overview (Please wait while the widgets are rendering)

Attribute Name	Histogram	Max	Min	Mean	Null Entries	Unique Entries
Recidivism_score		1.69	-3	-0.69	0	33
Violence_score		0.93	-4.63	-2.37	0	39

Functional Dependencies ☒

Violence_score, c_charge_degree, event, first_name
Violence_score, age, event, first_name
sex
race
Violence_score, event, first_name, marriage_status
Violence_score, age, first_name, race
Violence_score, event, first_name, race
Violence_score, age, first_name, marriage_status

Maximal Uncovered Patterns X

sex: Male
marriage_status: Unknown
marriage_status: Unknown
marriage_status: Significant_Other
c_charge_degree: M
marriage_status: Divorced
marriage_status: Separated
race: African-American
marriage_status: Significant_Other
marriage_status: Divorced
sex: Female
marriage_status: Significant Other
marriage_status: Separated
sex: Female
marriage_status: Married
c_charge_degree: M
sex: Female
race: Other
marriage_status: Unknown
marriage_status: Divorced
marriage_status: Separated
marriage_status: Significant Other
marriage_status: Significant_Other
marriage_status: Significant_Other
sex: Female
marriage_status: Separated
marriage_status: Significant Other
race: Caucasian
c_charge_degree: F
marriage_status: Separated
marriage_status: Unknown
marriage_status: Separated
c_charge_degree: M
marriage_status: Significant_Other
marriage_status: Unknown
marriage_status: Significant_Other
c_charge_degree: M
sex: Female
marriage_status: Significant Other
marriage_status: Separated
marriage_status: Widowed

Generate More Labels

Select...

Part III: Distribution/fairness-aware Integrtion

DISTRIBUTION/FAIRNESS-AWARE INTEGRTION

- Distribution-aware data collection
- Integrating data from multiple sources to fulfill the **Group Representation** requirement
- Query answering over the data that does not satisfy the **Underlying Distribution Representation** requirement
- Adapting to the required distribution of data for ML

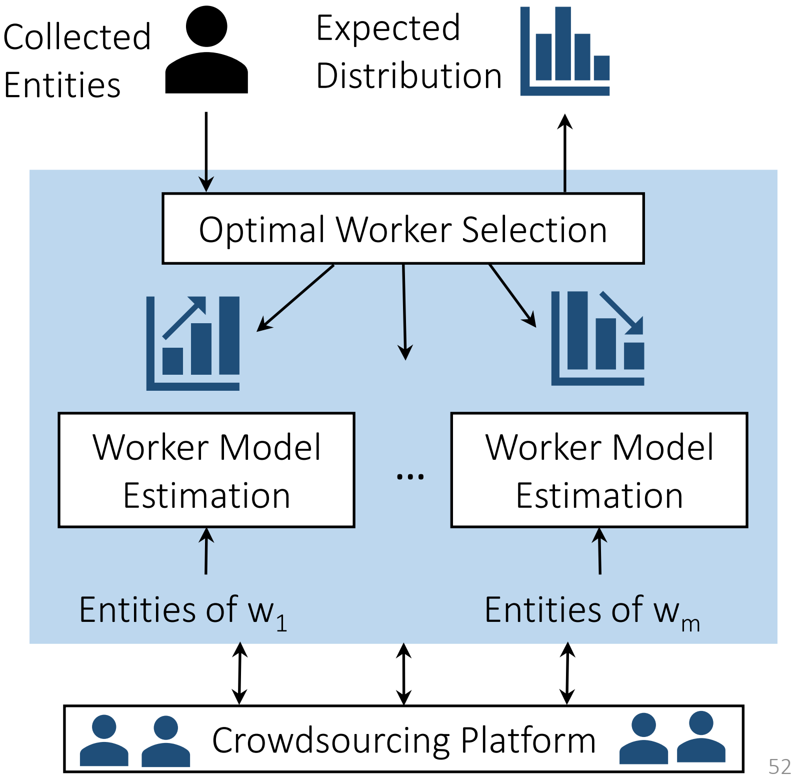
Crowd-sourced Entity Collection

- Crowdsourced data collection solicits the crowd to complete missing data in a database or a KB
- Distribution requirements on entities
  - University faculty recruiting has a distribution requirement on specialization and demographics of the applicants.
- Each individual worker may have its own *bias* of data collection

Distribution-Aware Crowdsourced Entity Collection, Fan et al., TKDE, 2017.

CROWD-SOURCED ENTITY COLLECTION

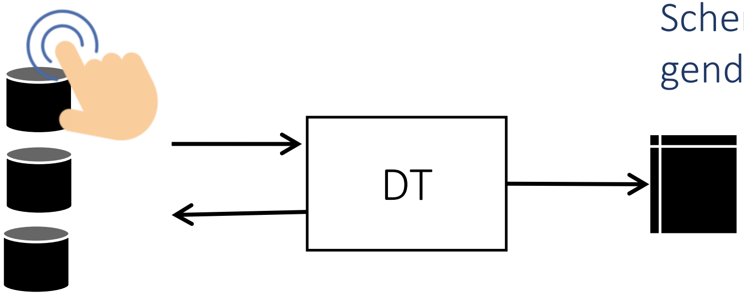
- **Adaptive worker selection** based on workers' historical entity set and select workers that minimizes the KL divergence from the expected distribution
- **Adjusting estimation of the underlying distributions** of workers upon submitting answers



DISTRIBUTION TAILORING

- Given sources $L = \{D_1,\ldots, D_n\}$ with their costs $\{C_1,\ldots, C_n\}$, and count requirements $\{Q_1, \ldots, Q_m\}$ on groups $\{G_1, \ldots, G_m\}$, the goal is to query different sources in L to collect samples that fulfill the count requirement, while the expected total query cost is minimized.

Cost of sampling from sources: monetary, sampling join paths, …

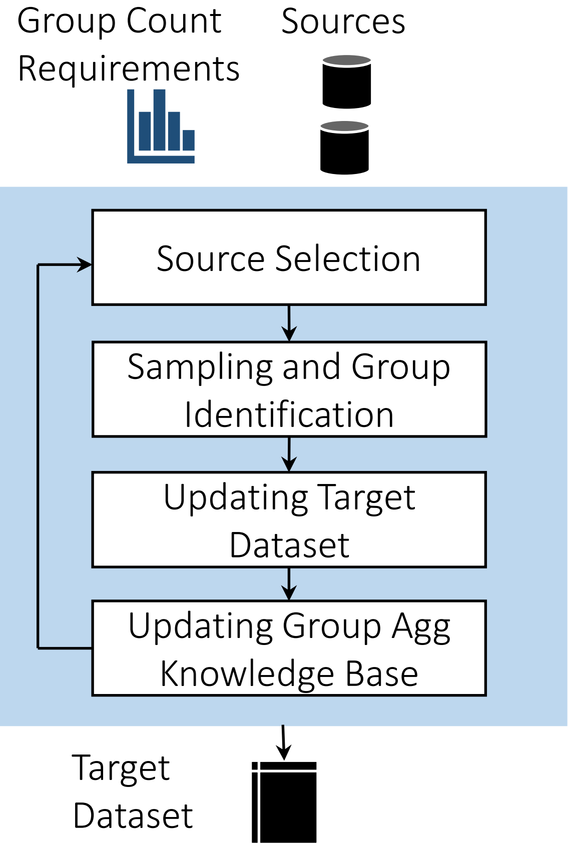


Schema: movie_title, actor_name, gender, race, …

Distbn Requirements: WM: 1K, NWF: 1K, …

Tailoring Data Source Distributions for Fairness-aware Data Integration, Nargesian et al., PVLDB, 2021.

DT ALGORITHM



Group Count Requirements

Sources

Source Selection

Sampling and Group Identification

Updating Target Dataset

Updating Group Agg Knowledge Base

Target Dataset

```
Input: data sources $L=\{D_1, . . . , D_n\}$ and $\{C_1, …, C_n\}$
        counts $\{Q_1, …, Q_m\}$ over $\{G_1, …, G_m\}$;
Output: $O$, the target data set
$O$ ← {}, cost ← 0
while($Q_j$>0) do
    $D, C$ ← select_optimal_source()
    s ← Query(D)
    j ← Group(s)
    if(s ∉ O AND $Q_j$>0) then
        add $s$ to $O$;
        $Q_j$ ← $Q_j$−1
    cost ← cost + C
return $O$
```

VERSIONS OF DT

- Known source distributions
  - Binary equi-cost DT
  - General DT (more than two groups and arbitrary cost)
- Unknown source distributions

KNOWN DT: EQUI-COST BINARY DT

- Two groups $\{G_1, G_2\}$ with counts $\{Q_1, Q_2\}$ and all source costs are equal.
- P_i^j : prob. of obtaining a *fresh* tuple of G_j from source D_i

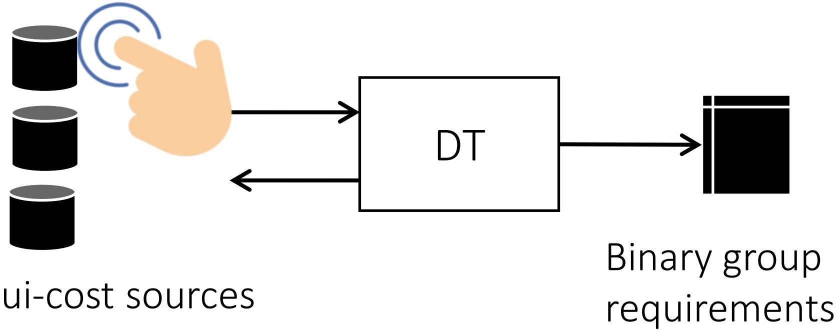
- Cost of getting a fresh tuple of G_j from D_i: $\frac{1}{P_i^j}$
- The best source for G_j: $D_{*j} = \underset{\forall D_i}{\mathrm{argm}ax}\left(P_i^j\right)$
- The best source for G_j is the one with maximum ratio of undiscovered tuples of G_j

OPTIMAL EQUI-COST BINARY

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

    D_{*1} with P_{*1}

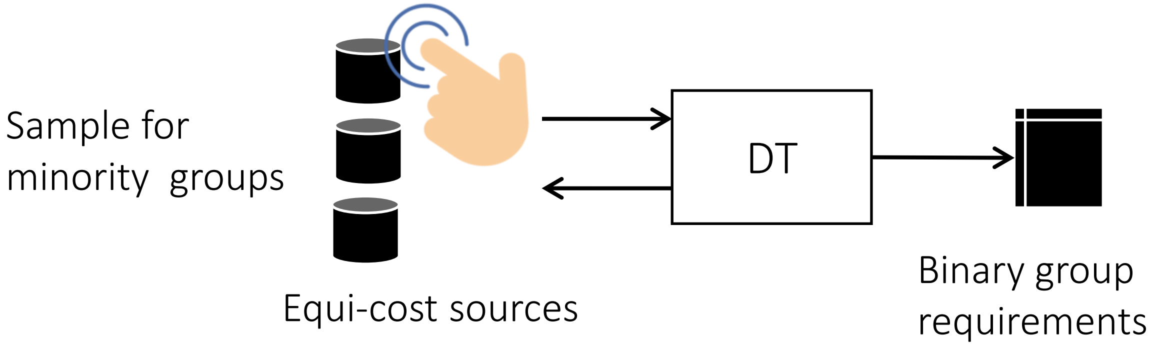
    D_{*2} with P_{*2}



- Let G_1 be the minority, i.e. $P_{*1} \leq P_{*2}$. Which source to sample?

OPTIMAL EQUI-COST BINARY

Theorem. *Consider the DT problem under the availability of group distributions where there are two groups and the costs for querying data sources are equal. Let* G_1 *be the minority, i.e.* $\mathrm{P}_{*1} \leq \mathrm{P}_{*2}$. *Selecting* D_{*1} *to query at current iteration is optimal.*



Known DT: General Non-Binary

- Select the most cost-effective source for each G_j (namely D_{*j})

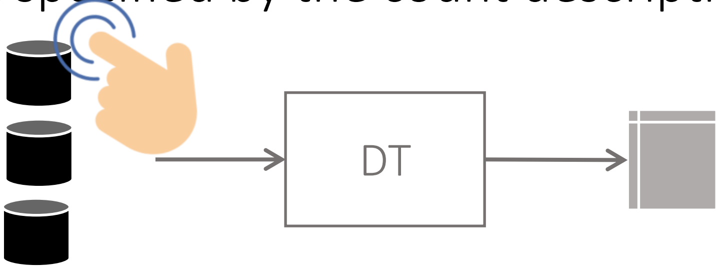
$$D_{*j} = \underset{\forall D_i}{\operatorname{argmax}} \frac{P_i^j}{C_i}$$

- Query the data source D_{*j} for group G_j

  Maintain the tuples of other groups *(piggybacking)*

- Repeat until the target specified by the count description is satisfied.

Sample for each group individually



Multiple group requirements

Unknown Dt: Multi-arm Bandit

Machines

Unknown reward distributions

Which action/source to pick next?

- ~~P_i^j : prob of collecting G_j from D_i~~
- Source reward
  - The more of rare groups a source has, the higher reward.
  - Penalize reward with cost

Proportional to the chance of containing rare groups

$$\overline{R}(i) = \frac{1}{C_i}\sum_{j=1,Q_j>0}^{m}\frac{O_i^j}{p^j O_i}$$

Penalizing based on source cost

- Bandit strategies: UCB, exploration-only, and exploitation only

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

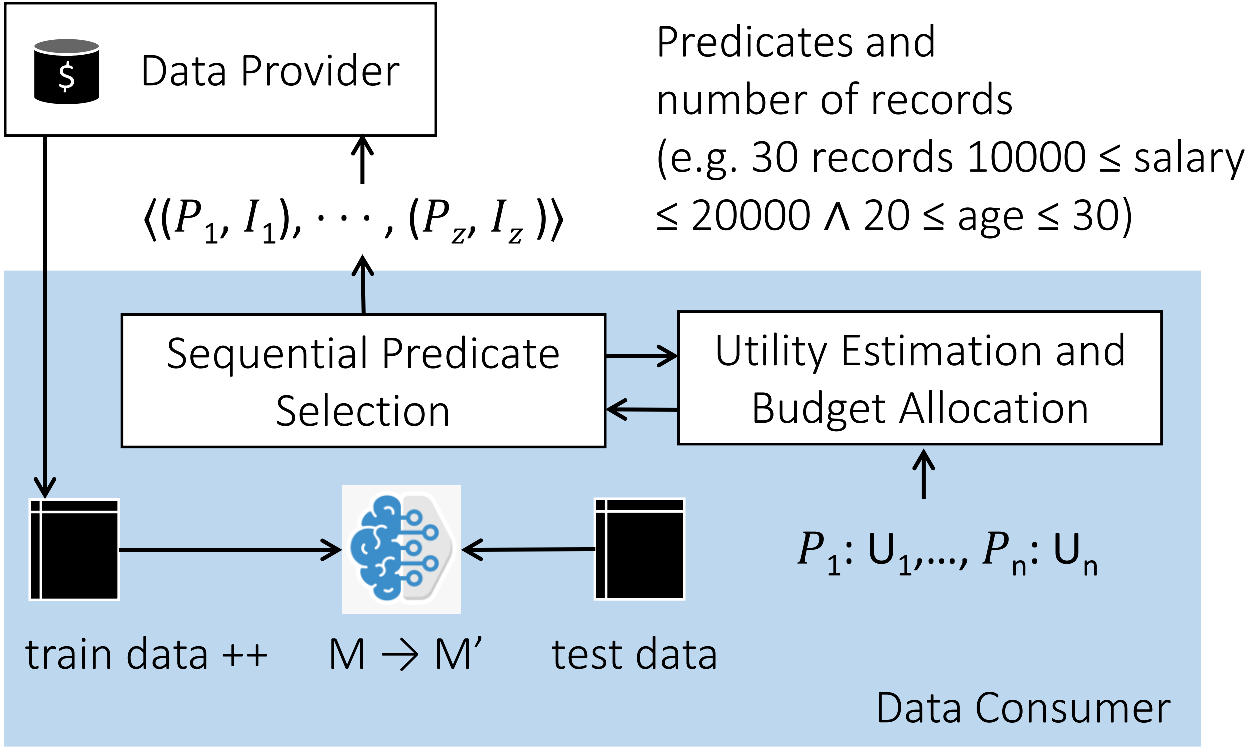
Data Acquisition for ML

- 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 Provider

$\langle (P_1, I_1), \cdots, (P_z, I_z) \rangle$

Predicates and number of records (e.g. 30 records 10000 ≤ salary ≤ 20000 ∧ 20 ≤ age ≤ 30)

Sequential Predicate Selection

Utility Estimation and Budget Allocation

P_1: U_1,…, P_n: U_n

train data ++

M → M’

test data

Data Consumer

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

FAIRNESS-AWARE QUERY Answering

- Query Rewriting

  Fairness-Aware Range Queries for Selecting Unbiased Data, Shetiya et al., ICDE, 2022.

  Coverage-based Rewriting for Data Preparation, Accinelli et al., EDBT, 2020.

- Aggregate Query Answering

  Sample debiasing in the themis open world database system, Orr et al., SIGMOD, 2020.

Fairness-Aware Range Queries for Selecting Unbiased Data

- Example:
  - SELECT * FROM EMP WHERE salary >= $65000
    - Includes around 18% of employees
    - Most of the selected employees are male
  - How to minimally change the initial query such that the output contains “almost equal” males and females
    - Minimally change: max Jaccard sim. b/w the output of the original and changed query

Fairness-Aware Range Queries for Selecting Unbiased Data

- Given a range query, Find most similar range query to given range query, such that output range query is fair.

```
SELECT ... FROM DATABASE
WHERE
   RANGE-PREDICATES
SUBJECT TO
   |$W_r$ $C_r$ - $W_b$ $C_b$| <= delta
    and SIM >= tau
```

```
SELECT * FROM EMP
WHERE
   salary>=$65000
SUBJECT TO
   |male-female| <= 0.1
    and SIM >= 0.95
```

Unweighted single predicate range query

- Adding or removing an item from a single predicate range query changes the disparity of the range by 1

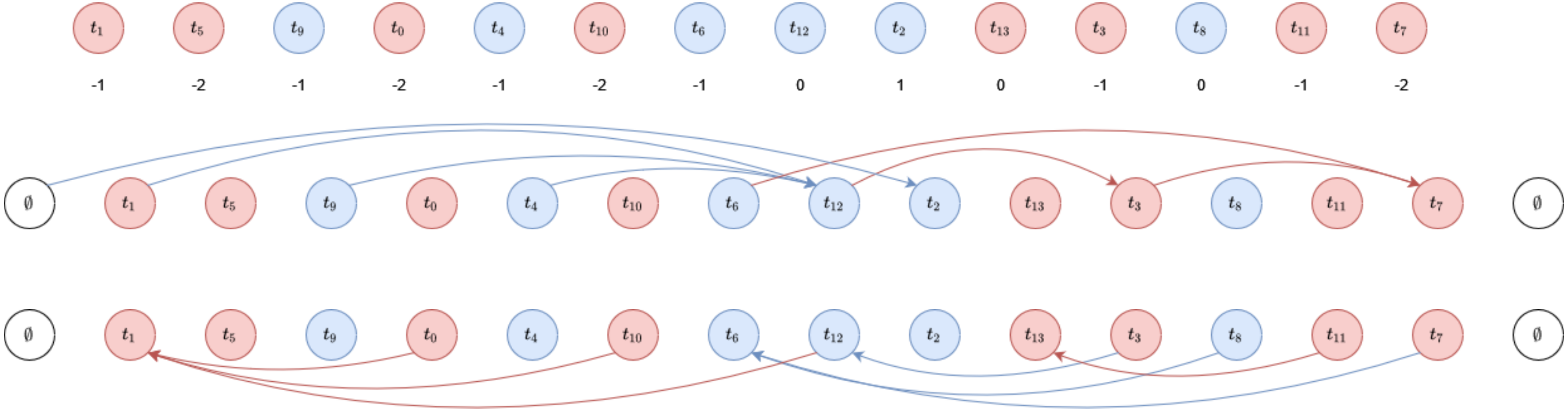
- Simple observation: The most similar fair range must have a disparity of δ exactly

- One can thus explore only those ranges which have a disparity of δ.

- As the left/right end point of the range can move, the sum of the disparity covered by the left and right should add up to δ.

Jump Pointers

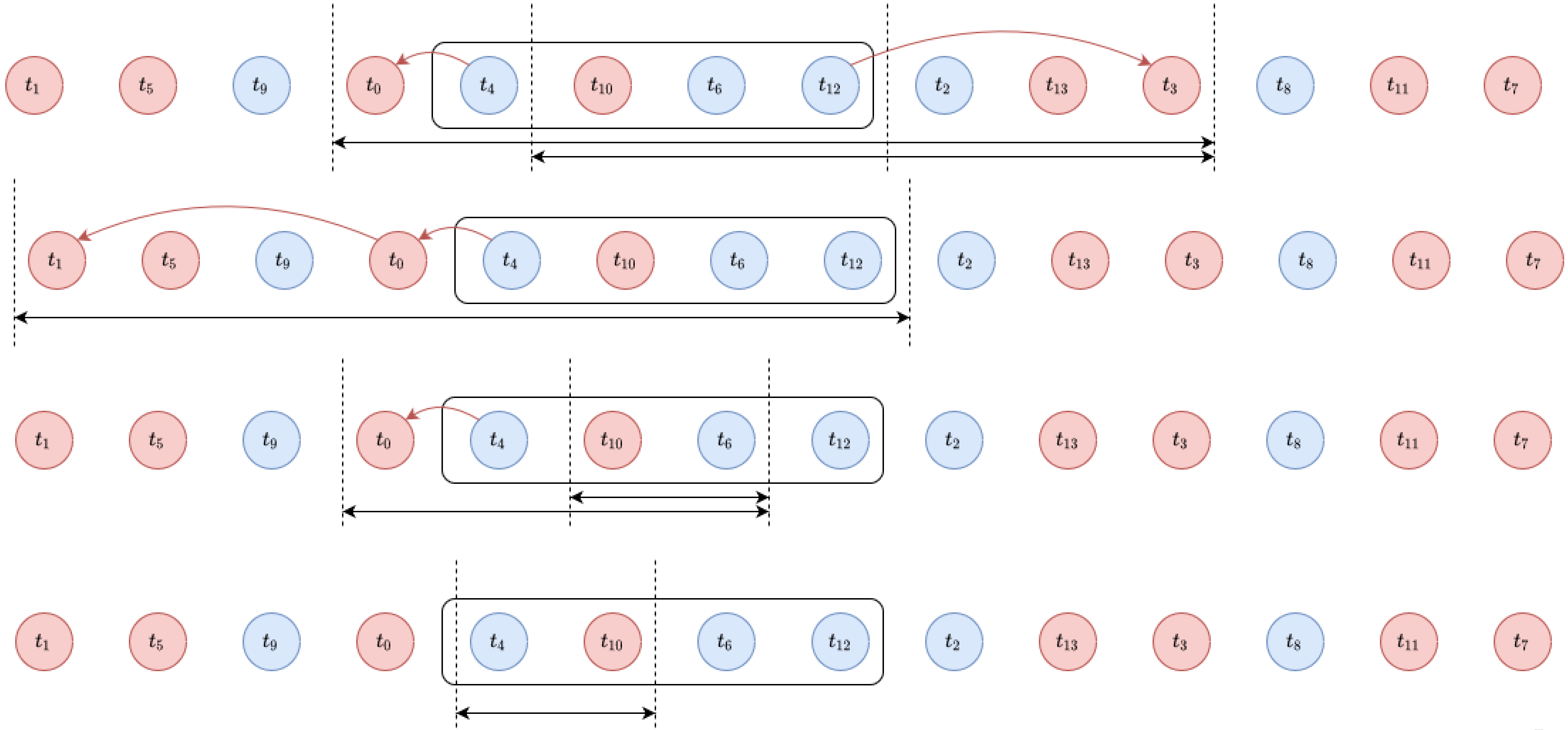
- A data structure that points to the next location in the dataset which has one additional blue (red resp).



t_1 t_5 t_9 t_0 t_4 t_{10} t_6 t_{12} t_2 t_{13} t_3 t_8 t_{11} t_7
-1 -2 -1 -2 -1 -2 -1 0 1 0 -1 0 -1 -2
∅ t_1 t_5 t_9 t_0 t_4 t_{10} t_6 t_{12} t_2 t_{13} t_3 t_8 t_{11} t_7 ∅
∅ t_1 t_5 t_9 t_0 t_4 t_{10} t_6 t_{12} t_2 t_{13} t_3 t_8 t_{11} t_7 ∅

ID	A0
t0	3.1
t1	0.7
t2	8
t3	10.9
t4	4.4
t5	1.2
t6	6.2
t7	13
t8	11.3
t9	2.3
t10	5.6
t11	12.7
t12	7
t13	9.1

Fair query – Other windows



t_1 t_5 t_9 t_0 t_4 t_{10} t_6 t_{12} t_2 t_{13} t_3 t_8 t_{11} t_7
t_1 t_5 t_9 t_0 t_4 t_{10} t_6 t_{12} t_2 t_{13} t_3 t_8 t_{11} t_7
t_1 t_5 t_9 t_0 t_4 t_{10} t_6 t_{12} t_2 t_{13} t_3 t_8 t_{11} t_7
t_1 t_5 t_9 t_0 t_4 t_{10} t_6 t_{12} t_2 t_{13} t_3 t_8 t_{11} t_7

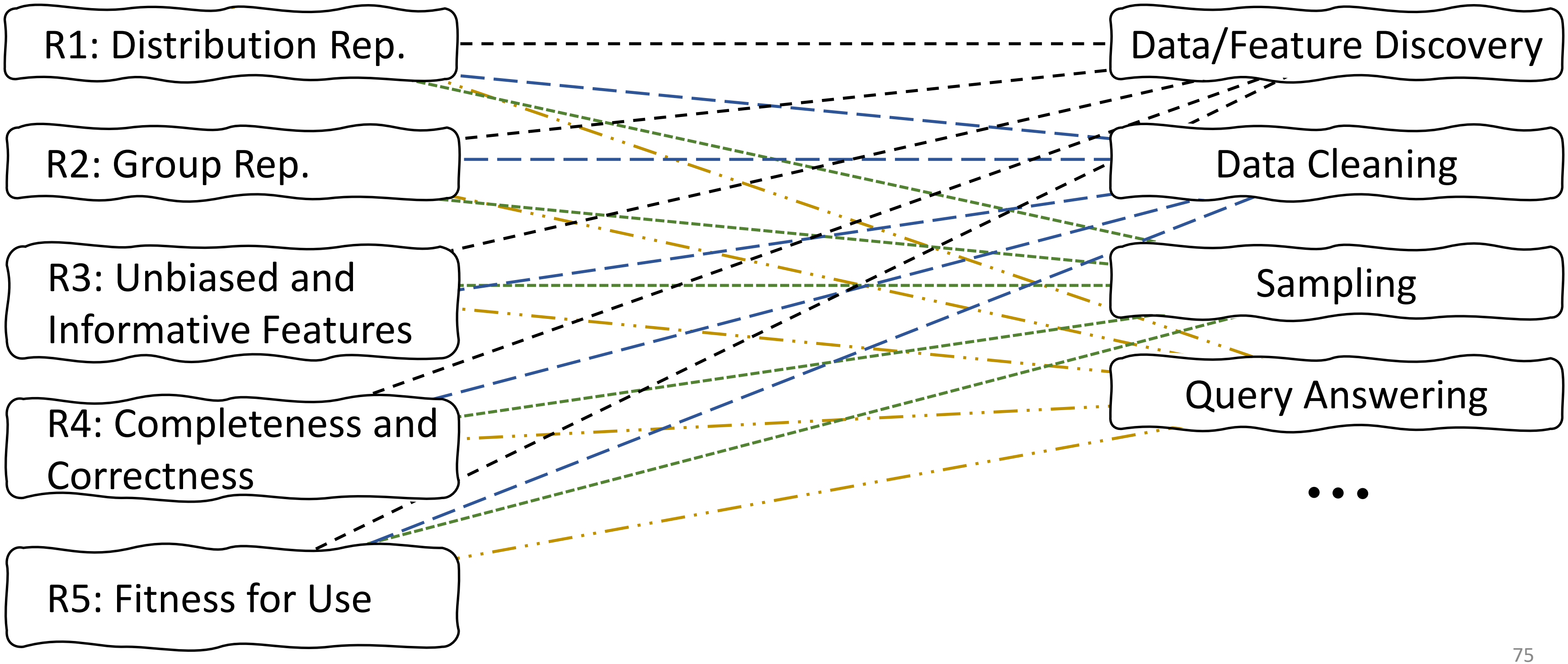
Fair range query Complexity

- Preprocessing – `O(n log(n))`

- Query processing time – `O(log(n) + disparity)`

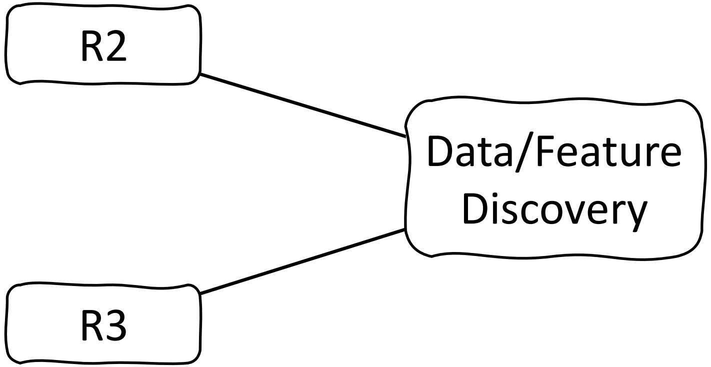
PART VI: OPPORTUNITIES

Requirement-Task Opportunities Bipartite Graph



DATA DISCOVERY

- Coverage-aware data discovery
    - Trade-off of relevance and group coverage
- Unbiased feature discovery
    - not biased (minimally correlated with sensitive attributes)
    - informative (highly correlated with the target attributes)



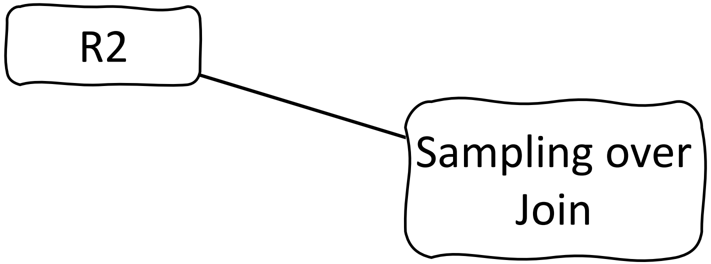
R2

R3

Data/Feature Discovery

SAMPLING OVER JOIN

- Satisfying group representation requirements (coverage) in a sample from the results of generic join paths



DATA CLEANING

- Auditing cleaning algorithms
  - Task-specific fairness measures
  - Auditing ML-based and rule-based data cleaning techniques
    - FairEM
- Fairness-aware data cleaning

Assessing Fairness in the Presence of Missing Data, Zhang and Long, NeurIPS, 2021.

FAIR QUERY ANSWERING

- Integrating declarative fairness-aware query answering into DBMS

Thank you!

- Fatemeh Nargesian
  - https://fnargesian.com/
  - fnaregsian@rochester.edu
- Abolfazl Asudeh
  - https://www.cs.uic.edu/~asudeh/
  - asudeh@uic.edu
  - Twitter: @ab_asudeh
- H. V. Jagadish
  - http://web.eecs.umich.edu/~jag/
  - jag@umich.edu