Privacy and Security in Distributed Data Markets

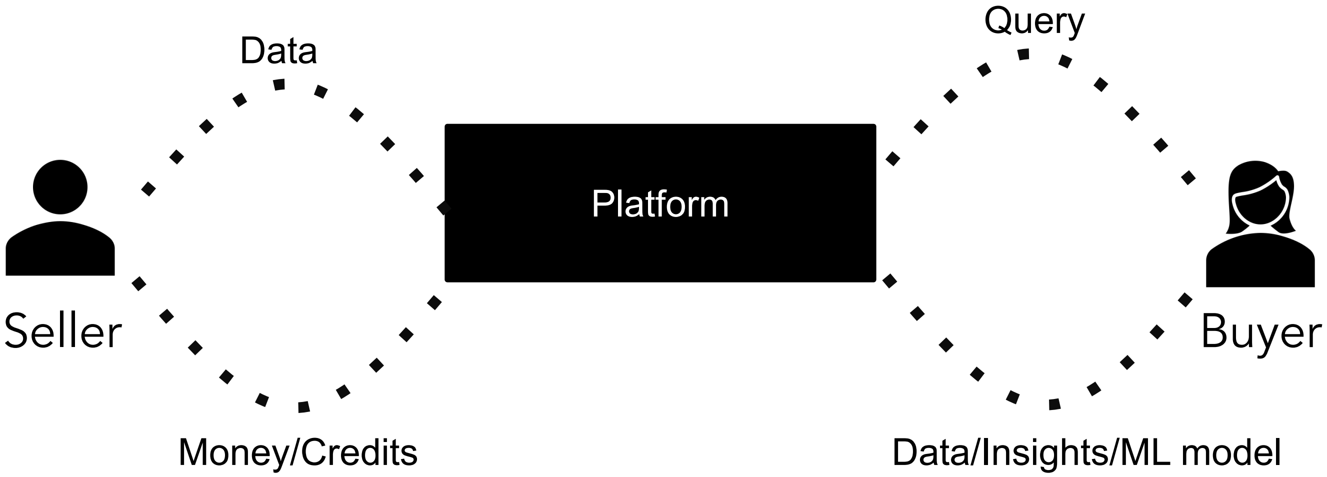
Daniel Alabi, Sainyam Galhotra, Shagufta Mehnaz, Zeyu Song, Eugene Wu

SIGMOD 2025 Tutorial

Overview of Data Markets

What is a data market?

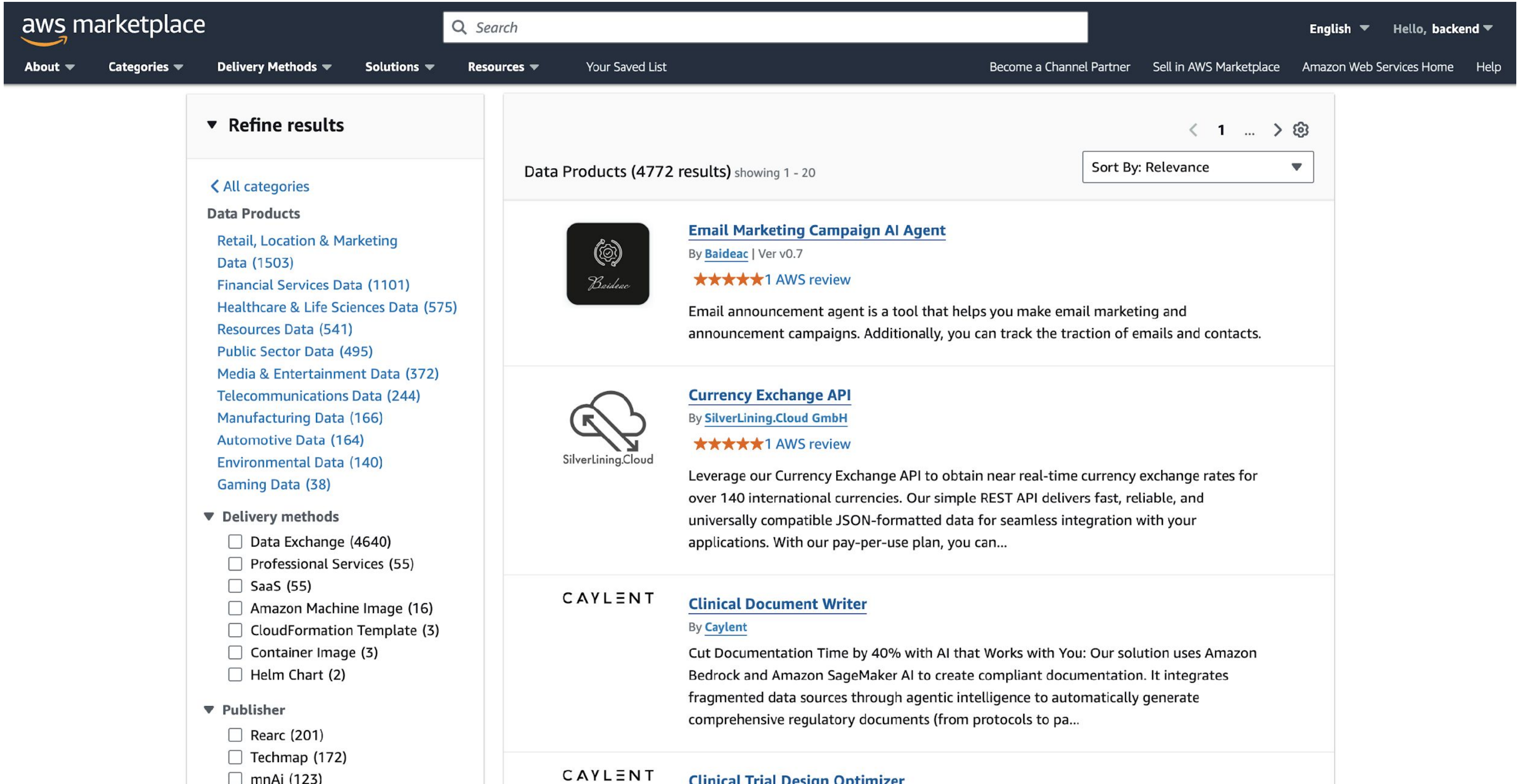
A platform where data is bought, sold or exchanged (much like a traditional marketplace)



Many types of data markets

snowflake

narrative



Many types of data market

- Keyword search over repositories
- Clean rooms
- Data labeling market
- Synthetic data market
- Curated alternative data

Keyword/NL Search

Keyword search over a table repository

- Index metadata or embeddings
- Return tables or links to tables
- Typically no money exchange

OpenData (data.gov), Academic (ICPSR, Dryad)

- Returns tables

Huggingface

- Returns models or training data

Google, Snowflake Marketplace, …

- Returns links

Enterprise Data Clean Rooms

Secure, privacy-preserving joins between orgs

- SQL aggregation over shared schemas
- Supports collaborative analytics (advertiser + publisher)

Example: Snowflake Clean Room

- Walmart w/ loyalty program & in-store purchases
- Discover w/ transaction & demographic data
- Can’t share raw customer data (PII)
- Join anonymized keys mediated by clean room
- Compute sales lift, cross-channel attribution

Others: AWS Cleanroom, BigQuery, InfoSum, Data Escrow

Data Labeling Markets

Acquire labels for training models

- User provides task, data, instructions, and goal schema
- Workers complete tasks, checked with reviewers/algorithms
- Pay for high quality labels

Examples: Scale AI, Sama, Surge AI

- Waymo has millions of raw LiDAR frames
- Wants 3D bounding boxes, semantic segmentation
- Submits task definitions and raw data to Scale AI platform
- Labelers + AI-assisted workflows produce structured annotations
- Outputs used to train NN models

Synthetic Data Markets

Simulate real data without exposing real records

- User uploads data
- Train on secure platform
- Return synthetic data/model
- Used for testing, demos, edge cases, sharing

Examples: Gretel.ai, MostlyAI

- LendingClub has loan applications (income, SSN, credit)
- Can’t share or use raw data for model testing due to compliance
- Uploads sample data to generate synthetic dataset
- Uses output to train and validate credit risk models internally

Data Brokers

Curates data about sectors, companies, metrics, tickers, …

- Sources from web & vendors
- Reduce noise, integrate, clean, enforce schema, align w/ business concepts
- Sells datasets, subscriptions to data feeds, or faceted/keyword access

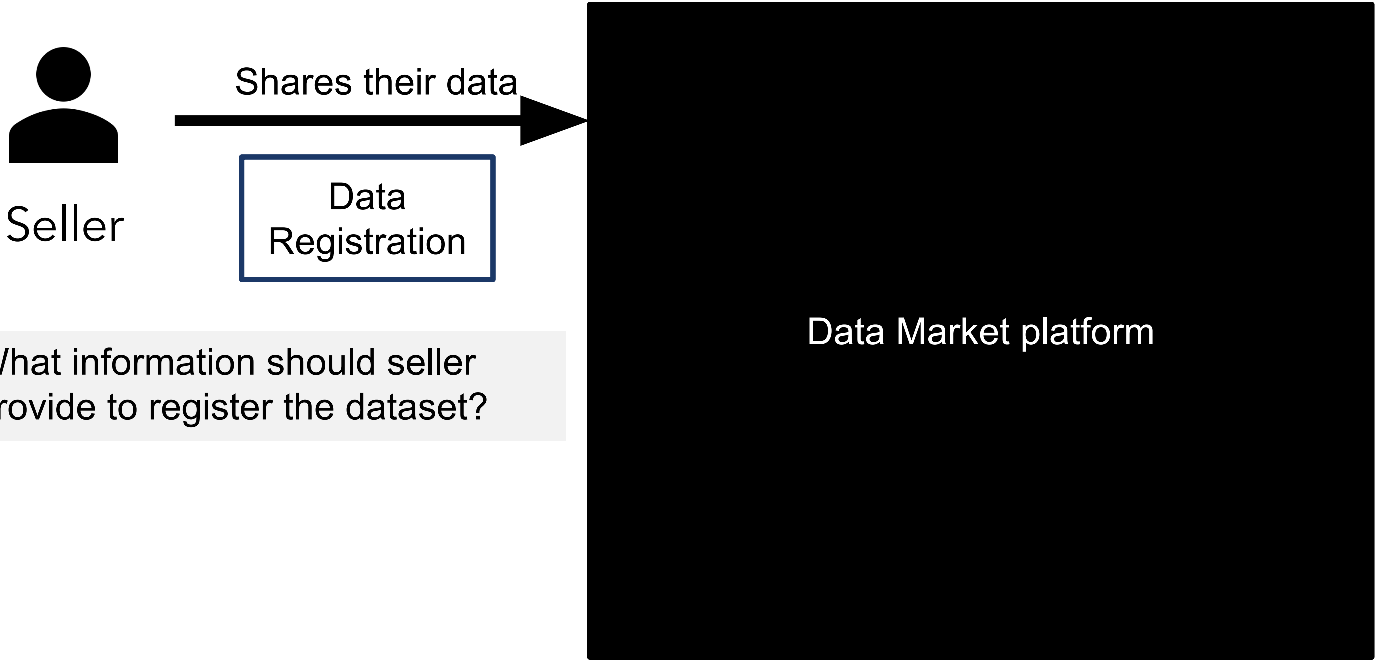
Example: Thinknum

- Crawls web: hiring pages, app store rankings, product pricing, retail inventory
- Differences data day-to-day
- Sells cleaned data feeds of changes e.g., Walmart + sales job postings

Others: Acxiom, Nielsen, Bloomberg, Morningstar, YipitData
https://oag.ca.gov/data-brokers

Category	Example	Query	Discovery	Incentive	Output
Clean Rooms	AWS Clean Rooms	SQL	Invite/catalog	Mutual value	Aggregated results
Labeling Markets	ScaleAI	Task	API	Payment	Labeled data
Alternative Data	Thinknum	Topic	Catalog/team	Subscription	Curated tables
Open Data Portals	NYC Open Data	Keyword	Tags/Portal	Public value	CSVs / APIs
Dataset Search	Google Dataset Search	NL (Keyword)	Metadata indexing	Visibility	External links
Model-as-Data	Hugging Face Datasets	Task	Benchmarks/Tags	Citation	Task-ready datasets
Academic Data	ICPSR	Structured	Metadata schema	Citation	Research tables
Synthetic Data	Gretel.ai	Schema	API	Privacy	Synthetic tabular data

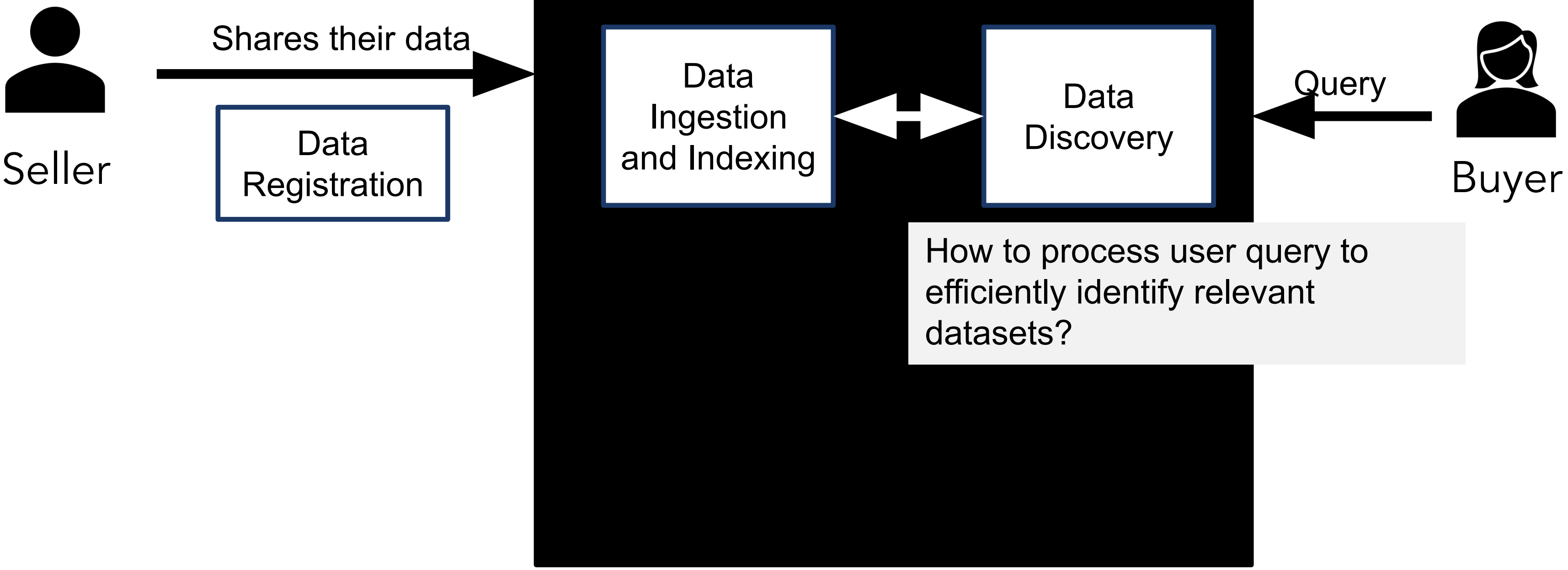
Key Components of a Data Market



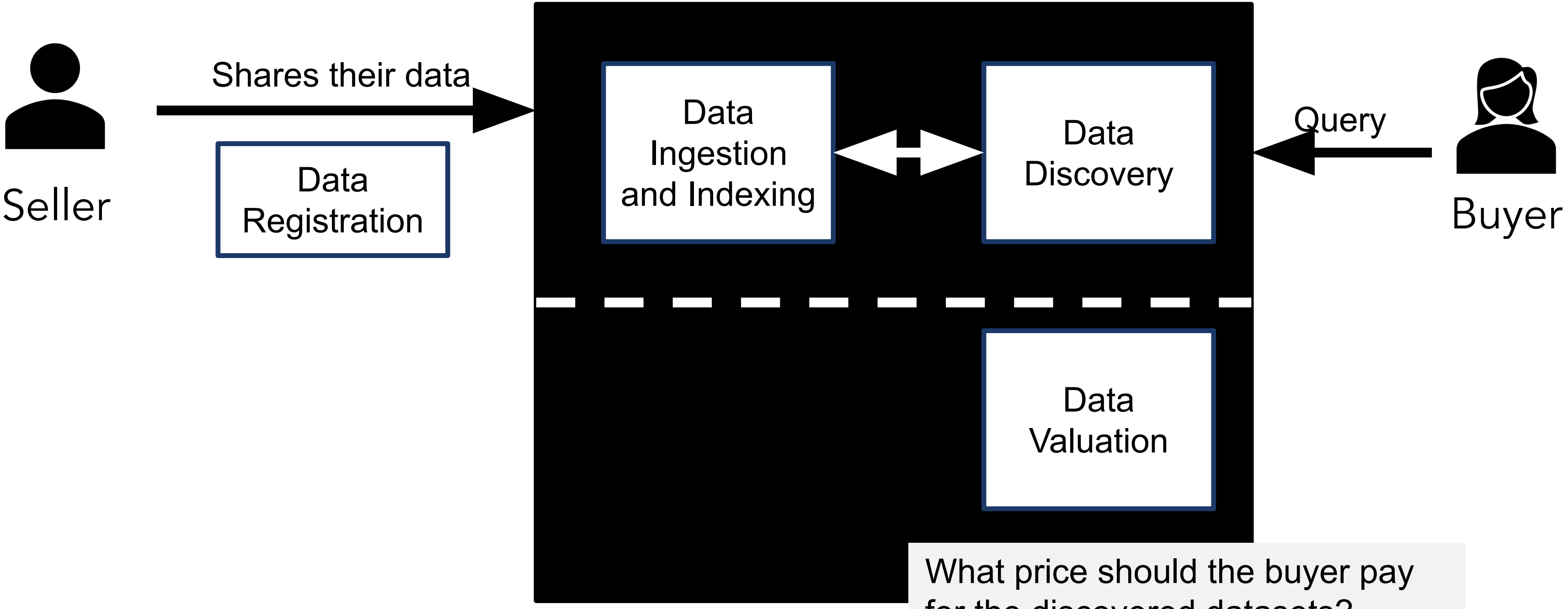
What information should seller provide to register the dataset?

Key Components of a Data Market

Key Components of a Data Market

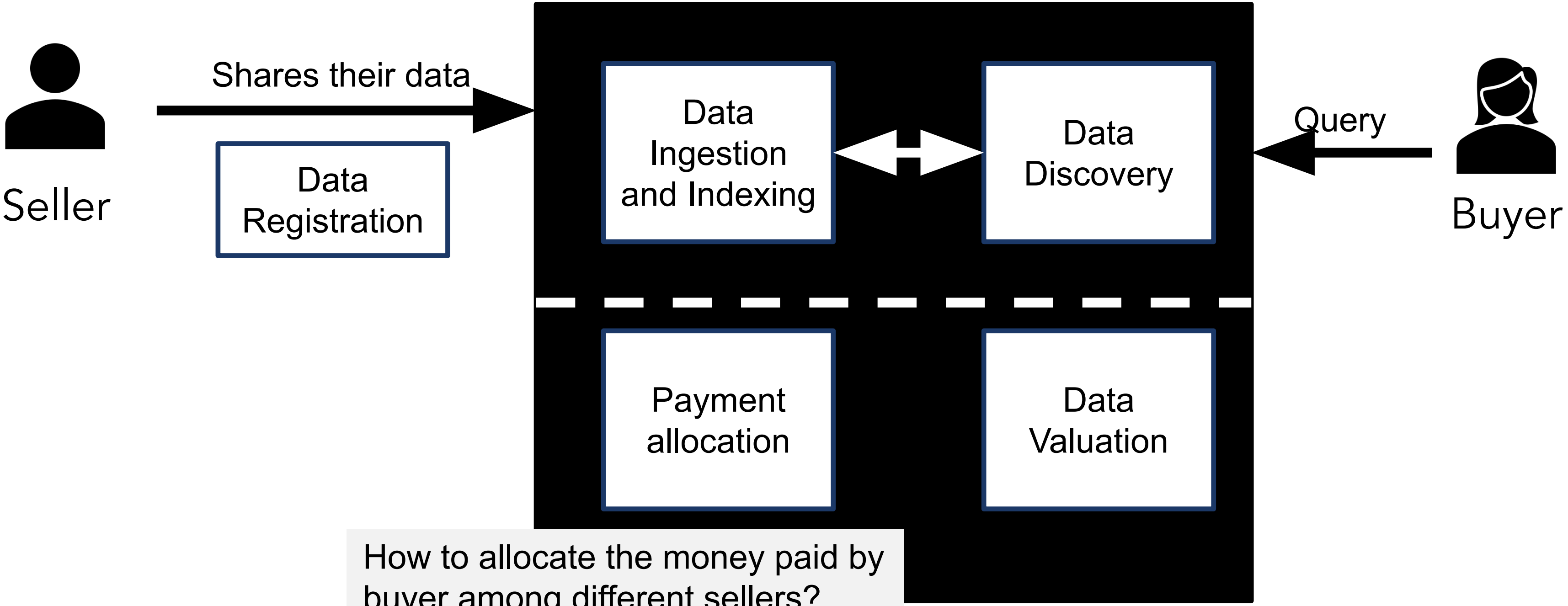


Key Components of a Data Market



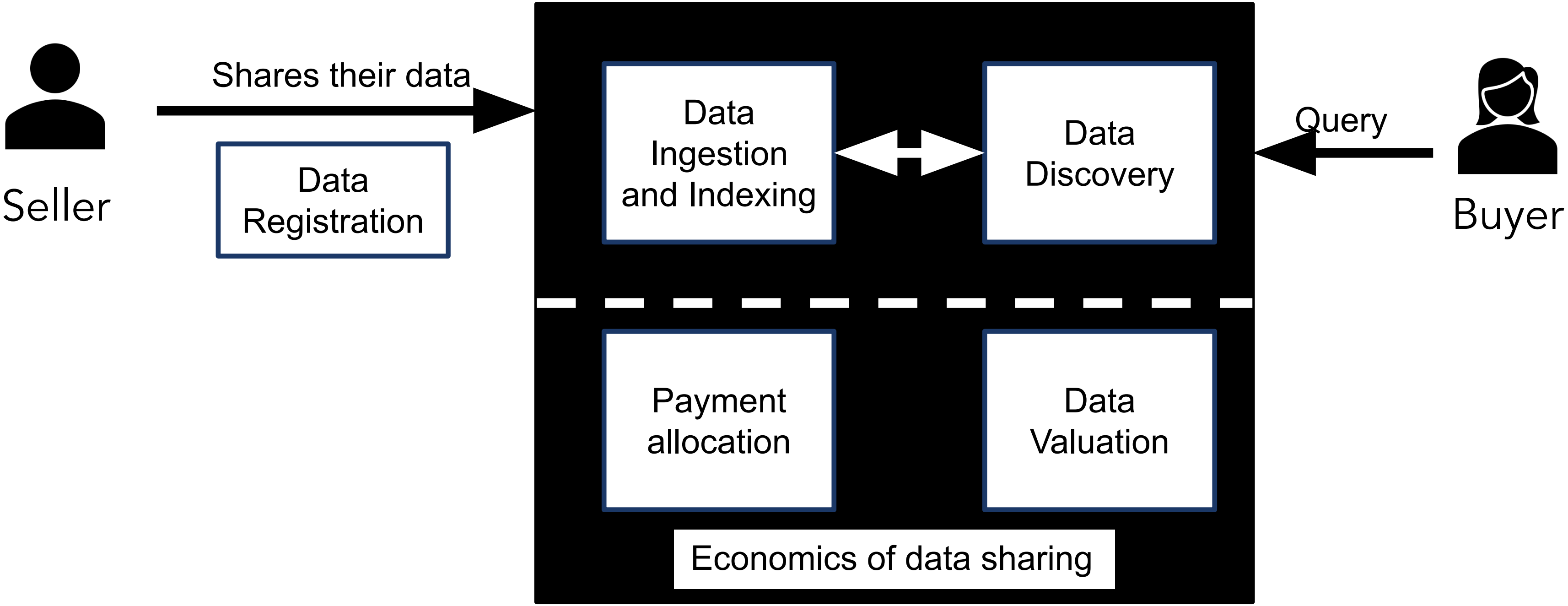
What price should the buyer pay for the discovered datasets?

Key Components of a Data Market

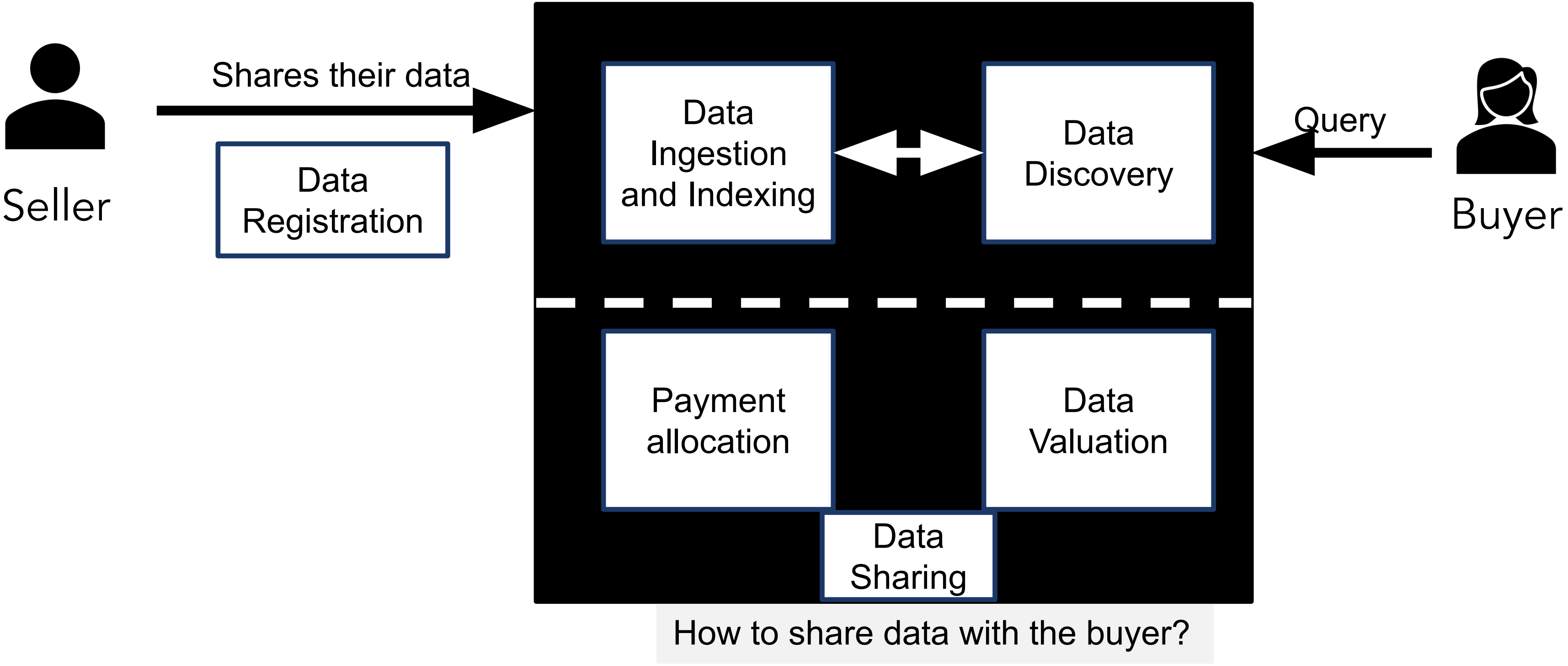


How to allocate the money paid by buyer among different sellers?

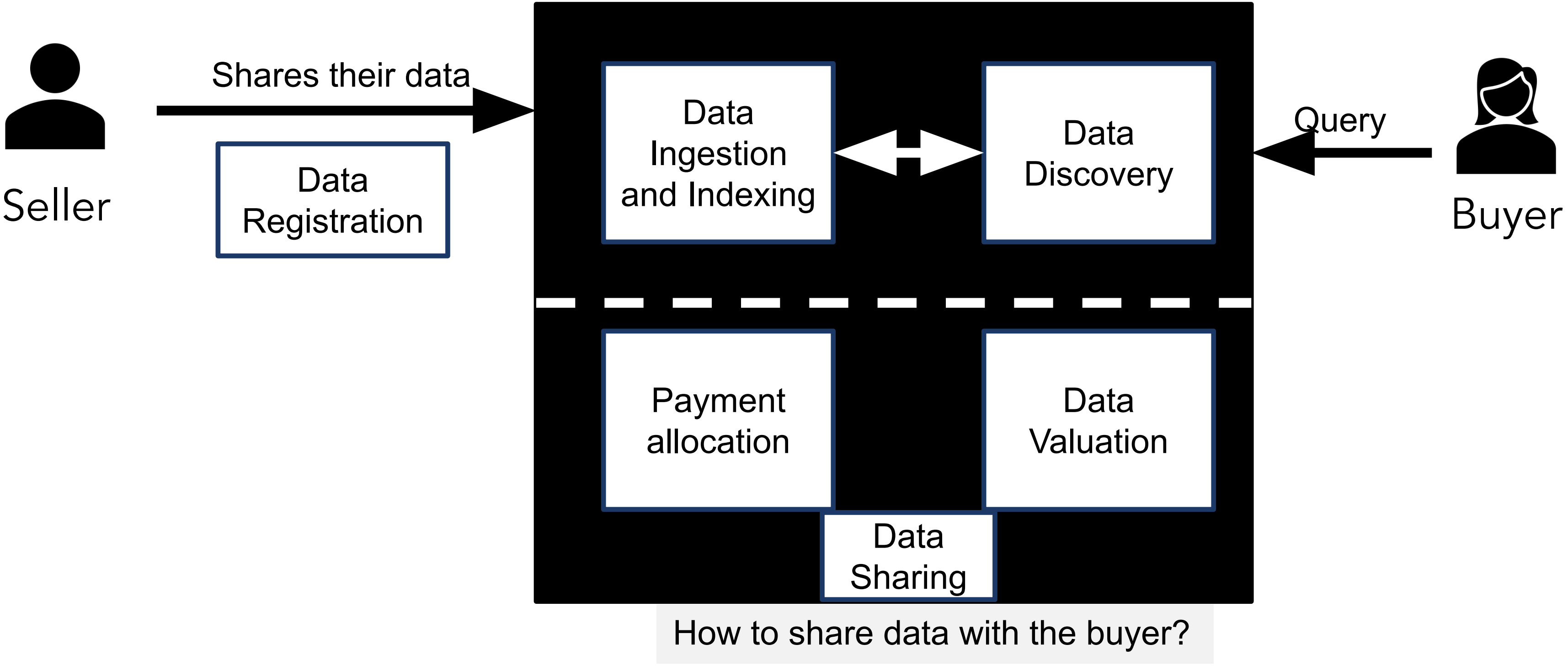
Key Components of a Data Market



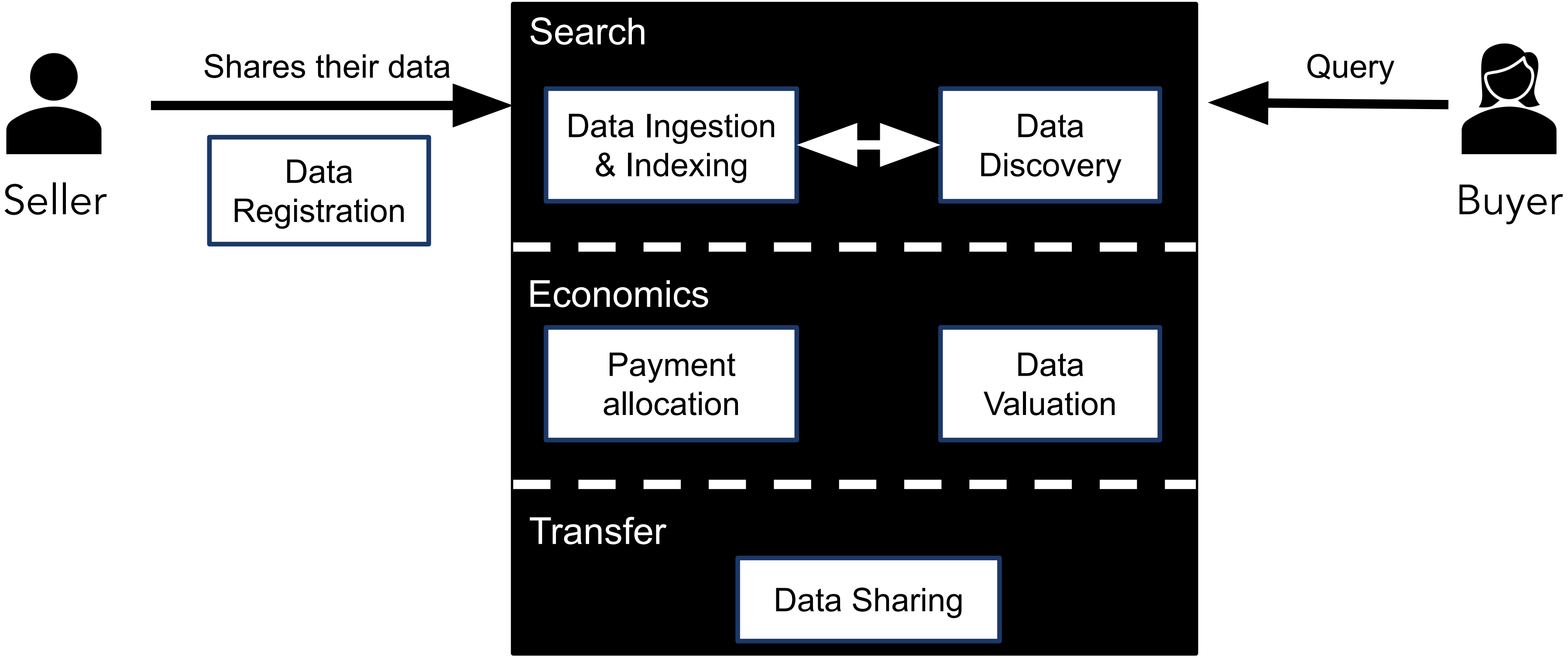
Key Components of a Data Market



Key Components of a Data Market



Key Components of a Data Market



Summary: Challenges of a data market

- Data Registration and Discovery:
  - What information should a seller provide?
  - How to store these datasets?
  - How to efficiently discover datasets for a buyer?
- Data Sharing (or acquisition)
  - Arrows information paradox
  - What does the seller get? How is the final dataset shared?
- Data valuation:
  - How to price datasets?
- Payment allocation
  - How to allocate the money paid by the buyers amongst the sellers

Systems challenge (This Tutorial)

Economics challenge

Focus of this tutorial

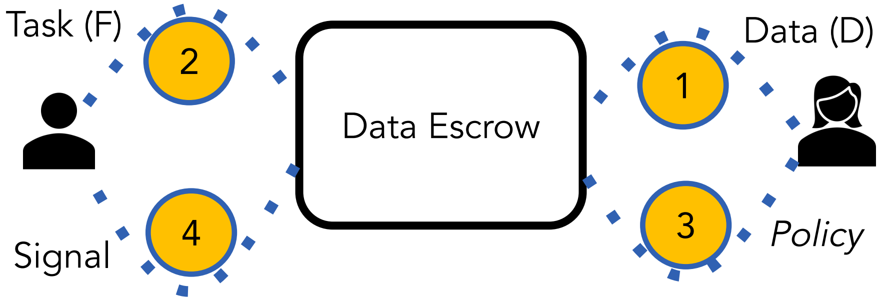
How to ensure security and privacy?

- Protect buyers from malicious sellers
- Protect sellers from malicious buyers
- Prevent *unauthorized* users from accessing:
  - Seller private data
  - Buyer private data
  - Platform private data
- Prevent manipulation of data acquisition mechanisms:
  - Data discovery
  - Data valuation
  - Data negotiation
  - Data delivery

How to control what buyers can acquire?

Data Escrow [VLDB'22]

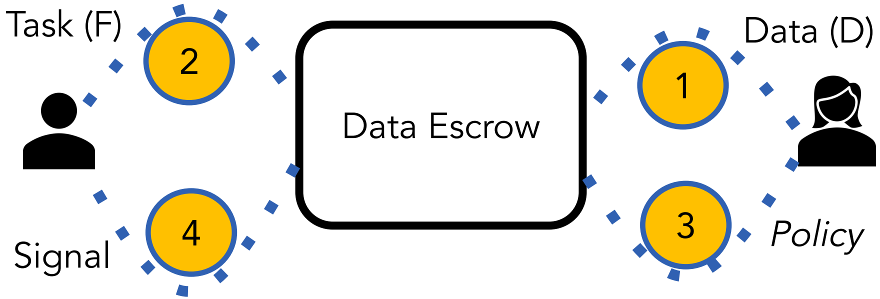
- A software system that controls dataflows
  - Sellers send their data; buyers send their tasks
  - Escrow runs buyers' tasks on seller's data



Slides borrowed from Raul Castro Fernandez, an author of this paper

Data Escrow [VLDB'22]

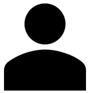
- A software system that controls dataflows
  - Sellers send their data; buyers send their tasks
  - Escrow runs buyers' tasks on seller's data



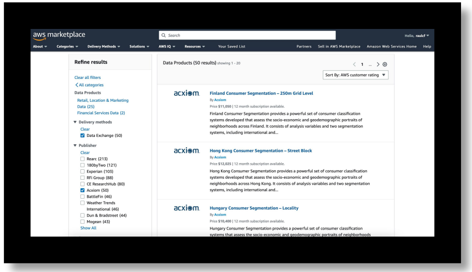
- <u>Guarantee</u>: no data* leaves the escrow without explicit permission, i.e., without an explicit *policy*

Using the Escrow to *Signal* Dataflow results

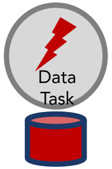
Data Markets



Seller



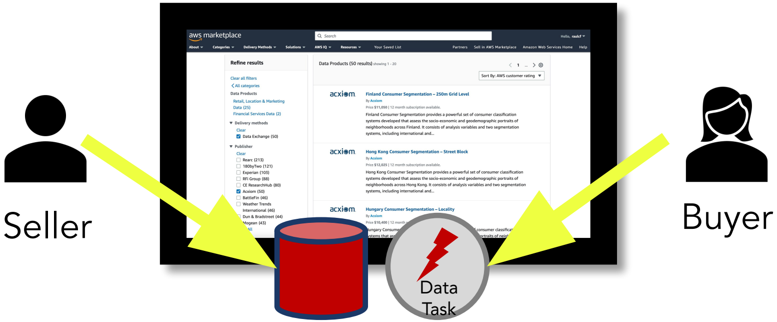
Buyer



With my data,
Accuracy: 0.63

Using the Escrow to *Signal* Dataflow results

Data Markets



With my data,
Accuracy: 0.63

Using the Escrow to *Signal* Dataflow results

Data Markets

With market data,
Accuracy: 0.94

With my data,
Accuracy: 0.63

Using the Escrow to *Signal* Dataflow results

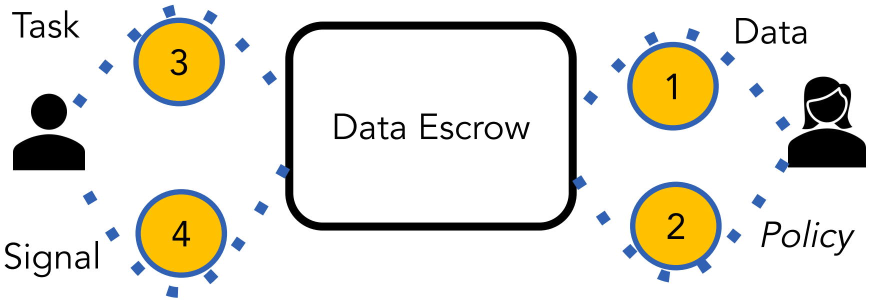
With my data,
Accuracy: 0.63

With market data,
Accuracy: 0.94

Market (escrow) signals the accuracy is 0.94. Buyer learns (0.94-0.63) is the benefit of buying

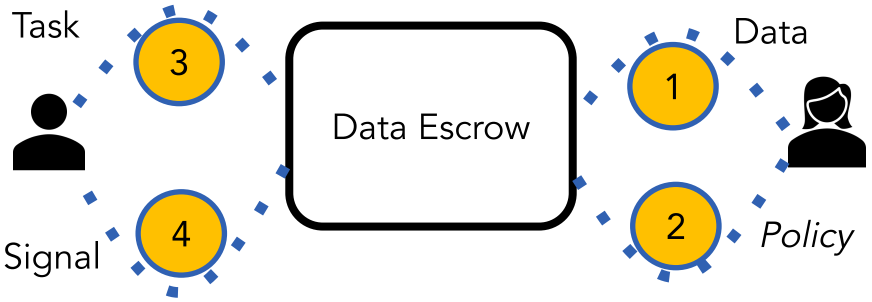
How do we delegate tasks, create signals,
i.e., how do we control dataflows?

Programmable Dataflows



Programmable Dataflows

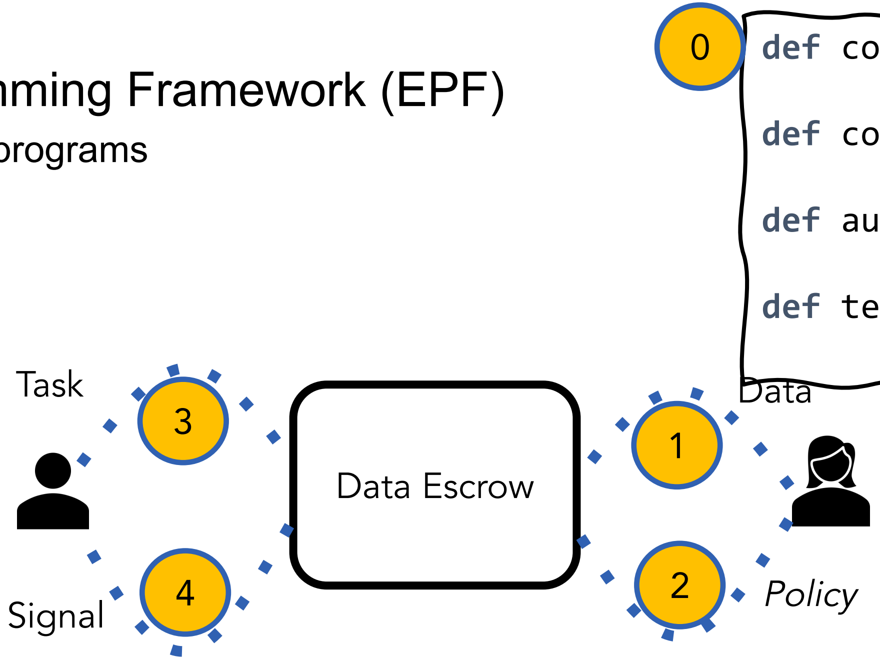
- Escrow Programming Framework (EPF)



Programmable Dataflows

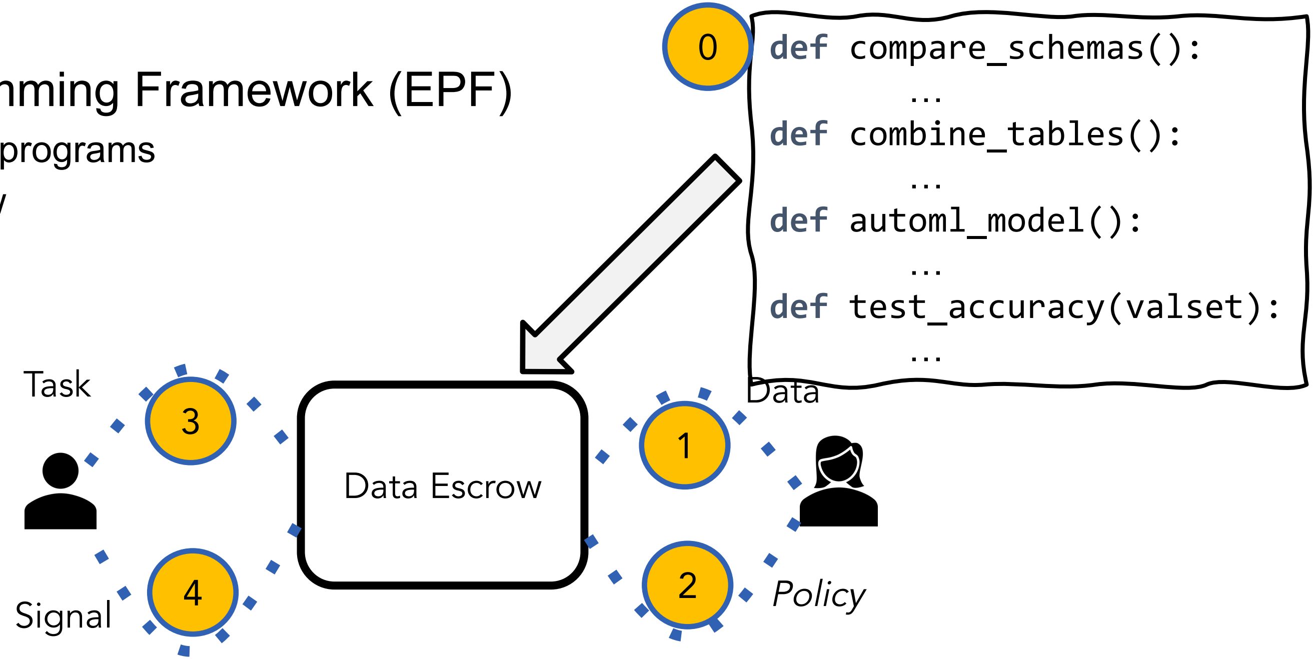
- Escrow Programming Framework (EPF)
  1. Developers write programs

```
def compare_schemas():
        ...
def combine_tables():
        ...
def automl_model():
        ...
def test_accuracy(valset):
        ...
```



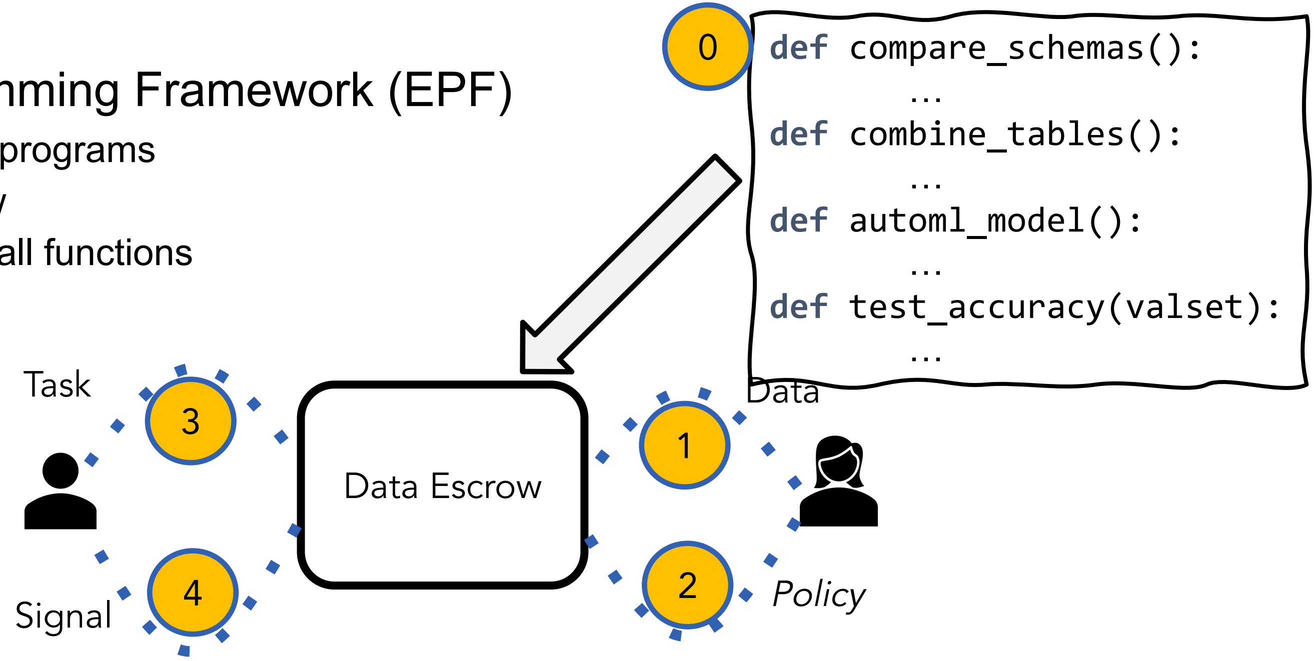
Programmable Dataflows

- Escrow Programming Framework (EPF)
  1. Developers write programs
  2. Deploy on escrow



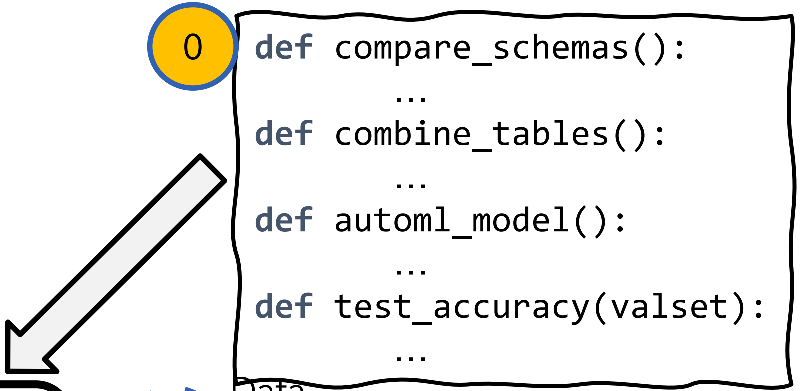
Programmable Dataflows

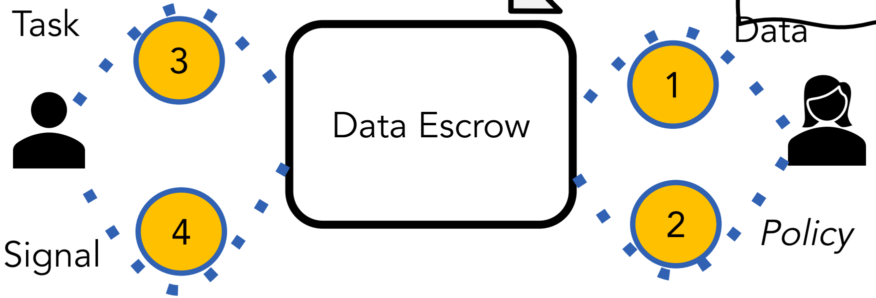
- Escrow Programming Framework (EPF)
  1. Developers write programs
  2. Deploy on escrow
  3. Agents join and call functions



Programmable Dataflows

- Escrow Programming Framework (EPF)
  1. Developers write programs
  2. Deploy on escrow
  3. Agents join and call functions





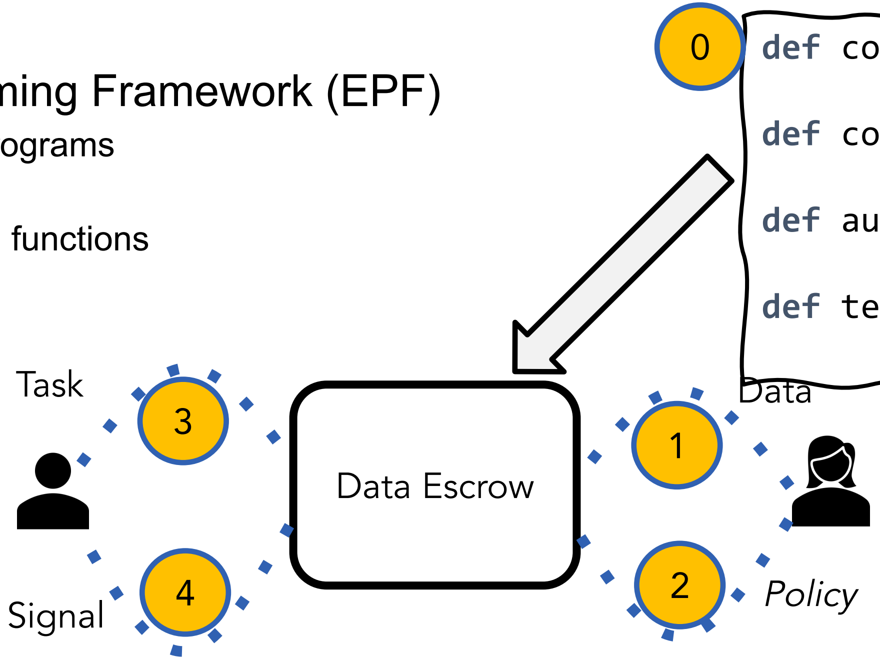
- Program implements communication and logic via *contracts*

Programmable Dataflows

- Escrow Programming Framework (EPF)
  1. Developers write programs
  2. Deploy on escrow
  3. Agents join and call functions

Programs implement data environments

```
def compare_schemas():
        ...
def combine_tables():
        ...
def automl_model():
        ...
def test_accuracy(valset):
        ...
```

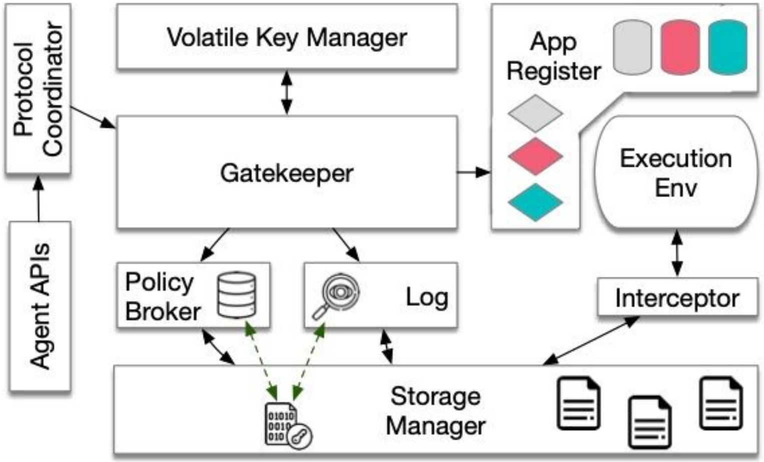


- Program implements communication and logic via *contracts*

Delegated, Auditable, Trustworthy

- What happens in the escrow, stays in the escrow
  - Except when it needs to be available to auditors and 3-party officers

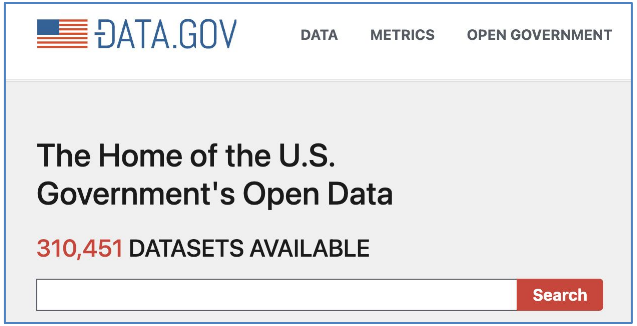
- Data is encrypted end-to-end
  - At rest *and* during computation
    - Use of secure hardware enclaves
  - Encrypted Write-Ahead Log (EWAL)
  - Cryptographic protocols for IO
    - Key exchange and recovery after failures…

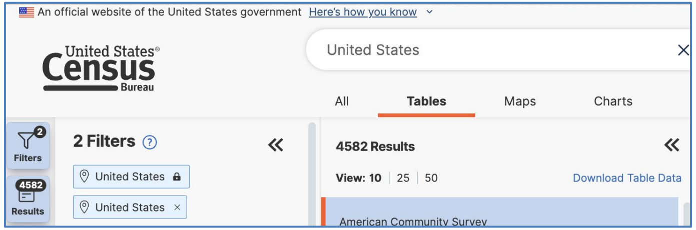


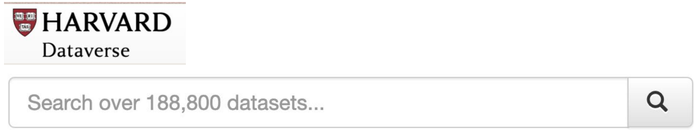
Data Search

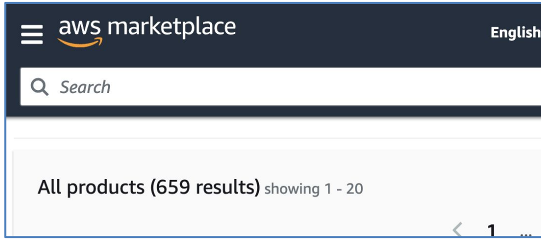
Unlimited Storage → Massive Data Repos

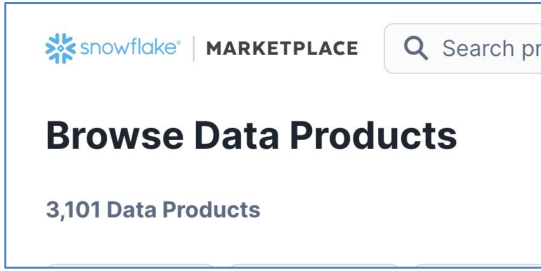
Gov Portals
Data Markets
Data Lakes
Web Tables
Data coalitions
…

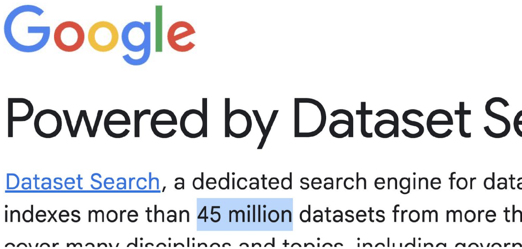












What Can We *Do* with 1M+ Tables?

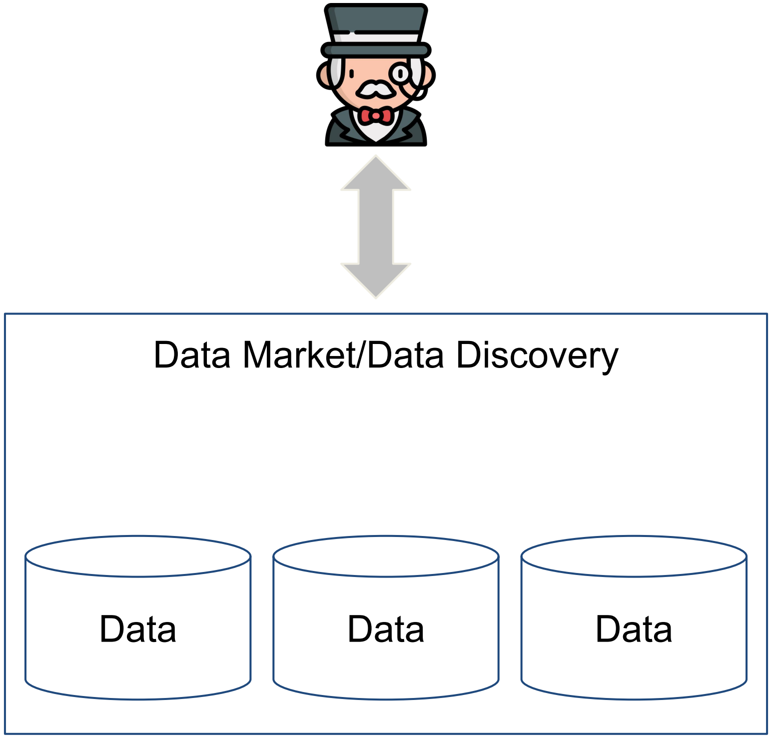
Gov Portals
Data Markets
Data Lakes
Web Tables
Data coalitions
…

?

Scientific phenomena
Economic theories
Investment hypotheses
Customer analysis
…

Step 1: Find Relevant Tabular Dataset

Centralized Data Search Systems



A single system stores & manages the datasets

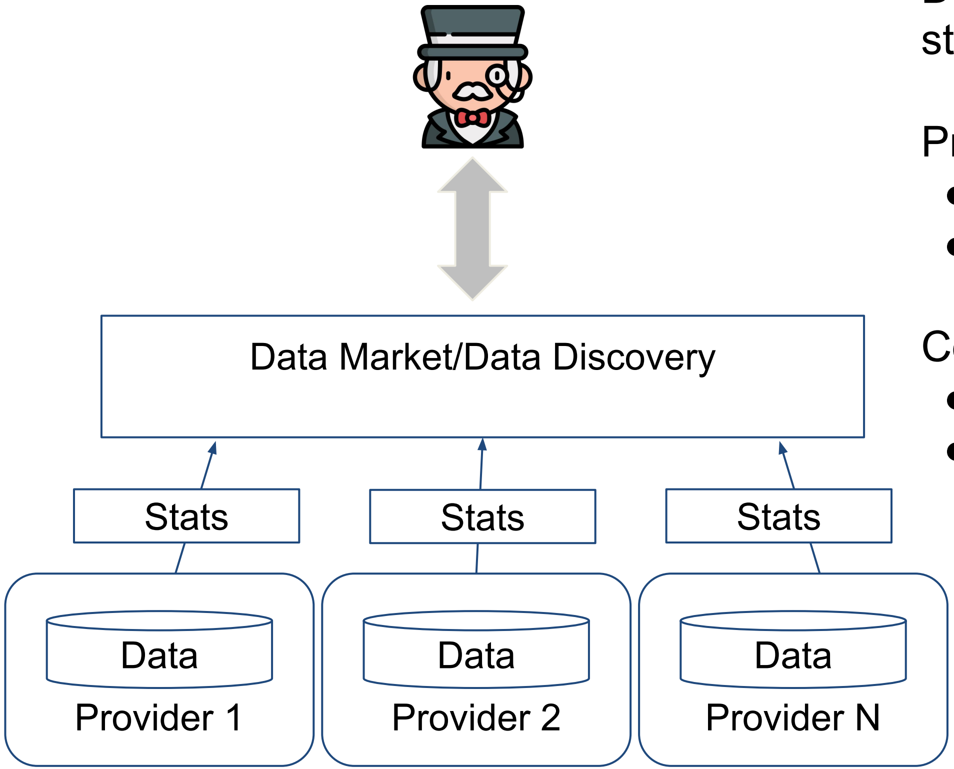
Pros:

- Fits an organization's data lake
- Easier access to raw data, experts, metadata
- Easier to tightly integrate with use cases

Cons

- Limited to a single organization

Decentralized Data Search Systems



Data is federated, and system has access to statistics rather than raw data

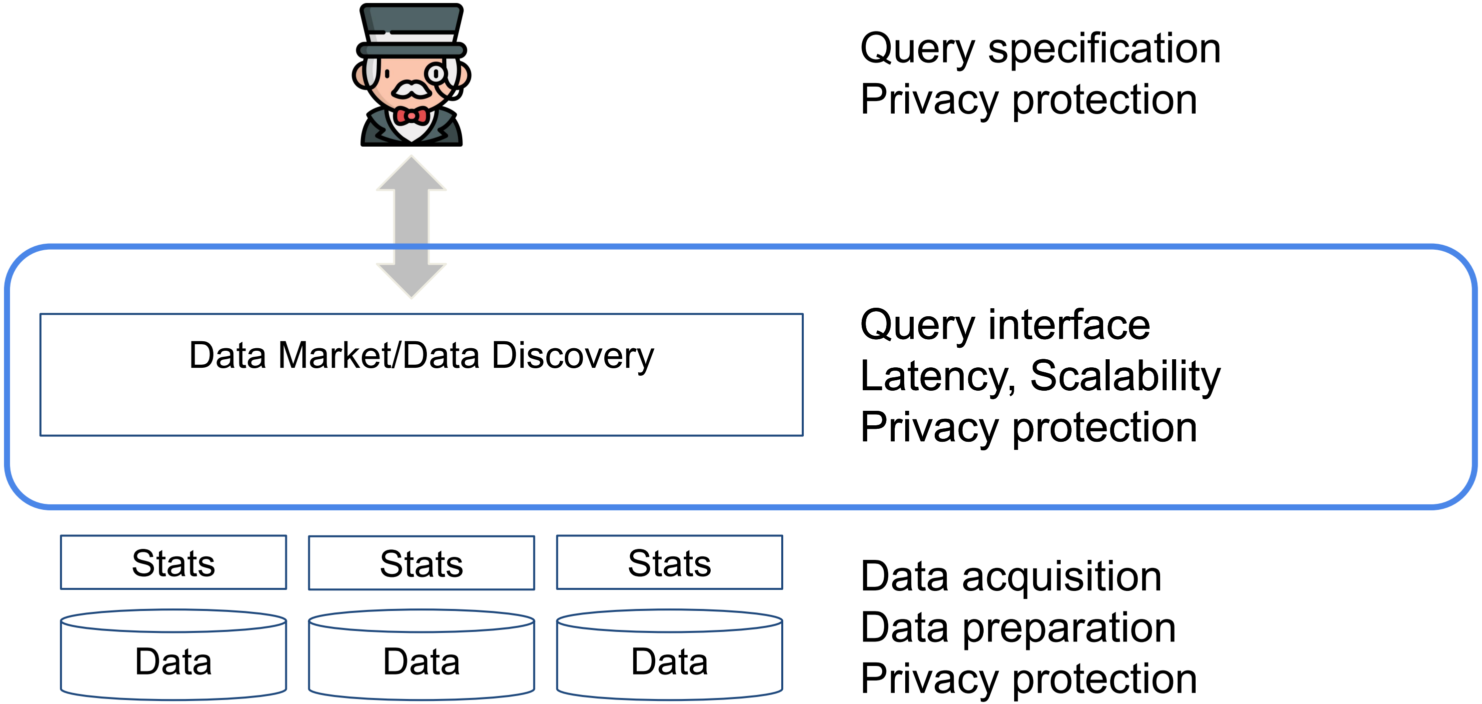
Pros

- Clear separation of privacy concerns
- More realistic for a public data market

Cons

- More difficult to provide utility
- Hard to manage multiple providers

Challenges in Data Search Systems



3 Classes of Systems

Keyword/Metadata Search

Data Discovery

Task-based Search

Keyword Search

narrative

snowflake

Datarade

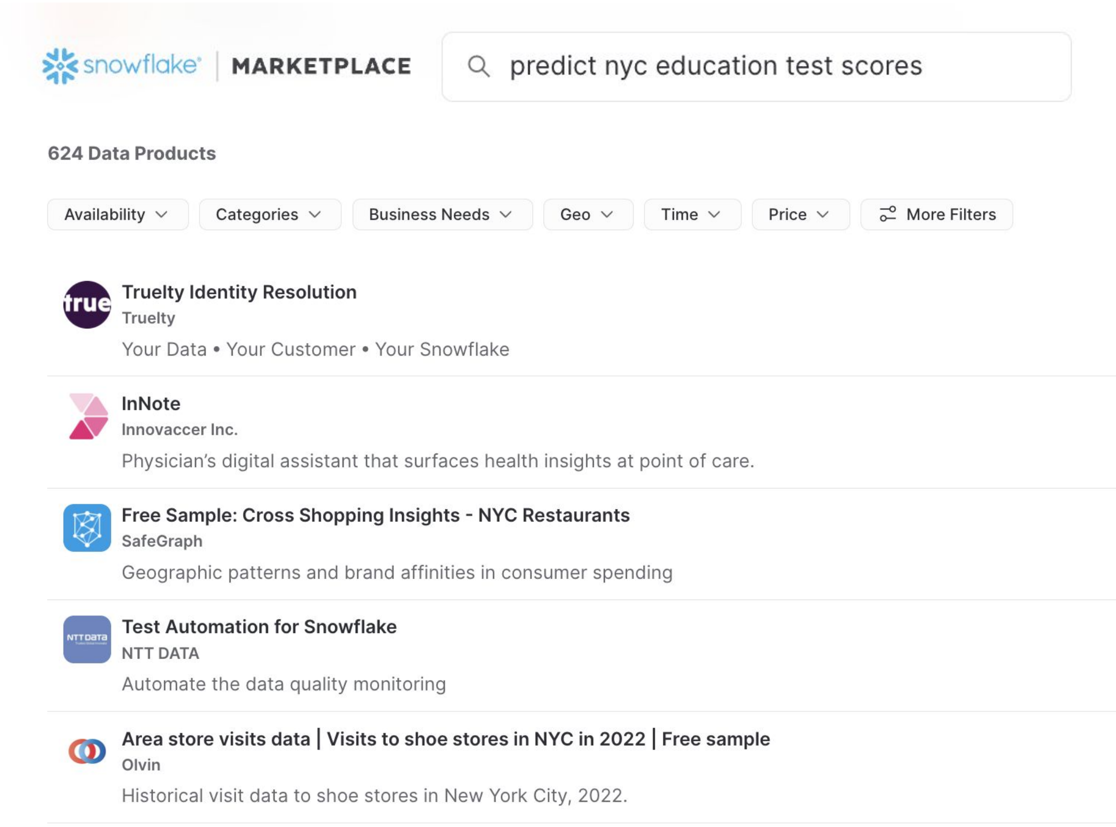
data.gouv.fr

Google Dataset Search

DATA.GOV

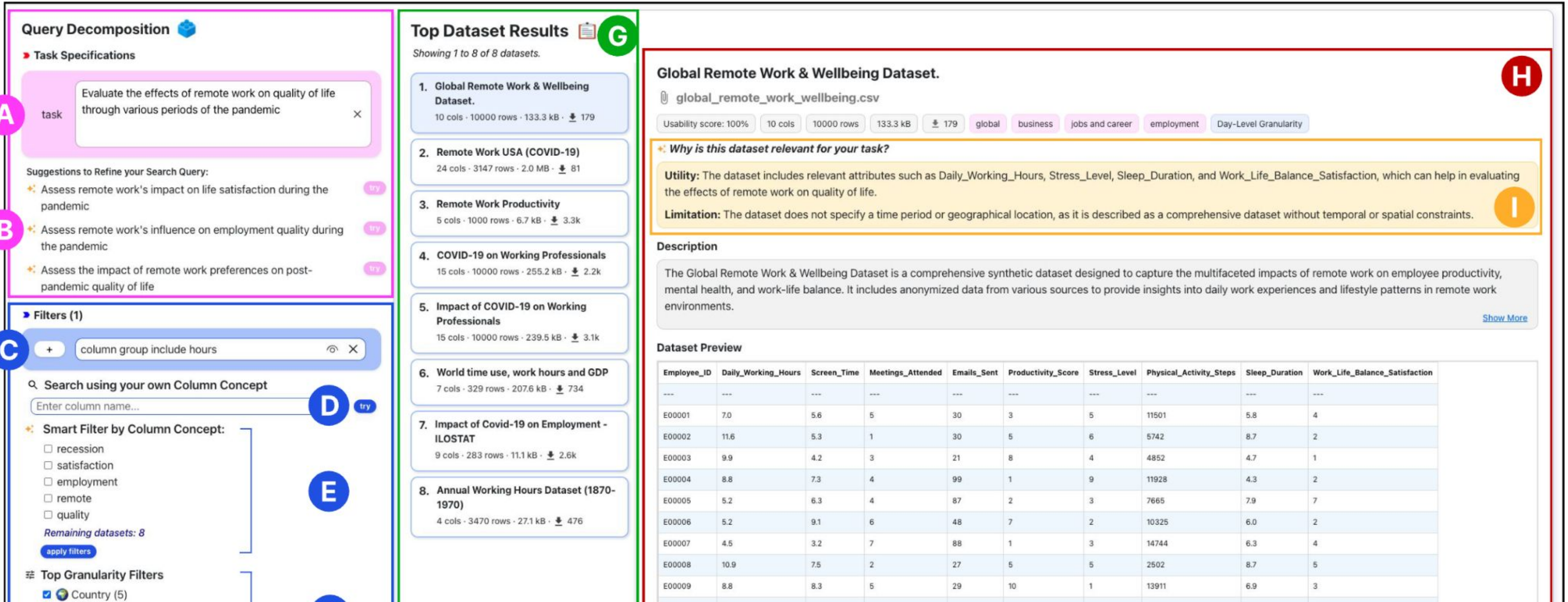
NYC OpenData

City of London Open Data



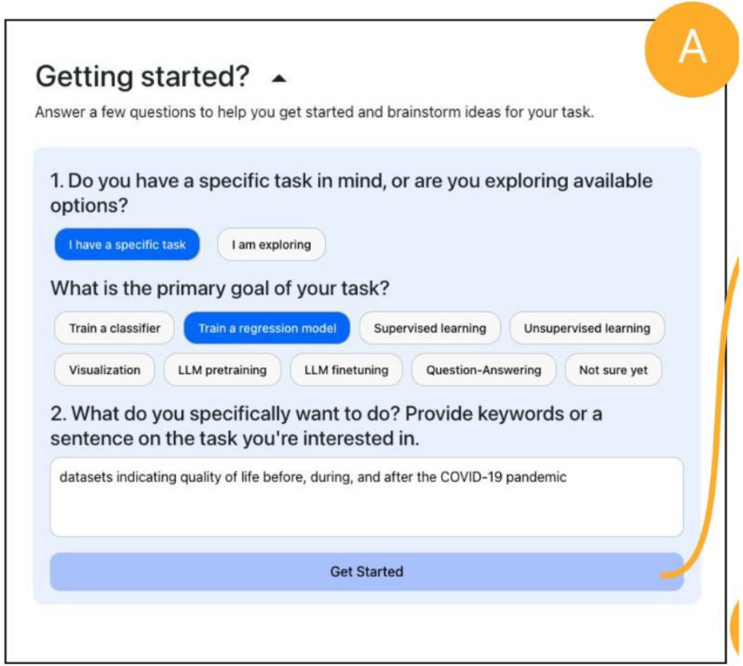
Keyword Search as Sensemaking

DataScout. Rachel Lin, Bhavya C., Wenjing L., Shreya S., Madelon H., Aditya P.



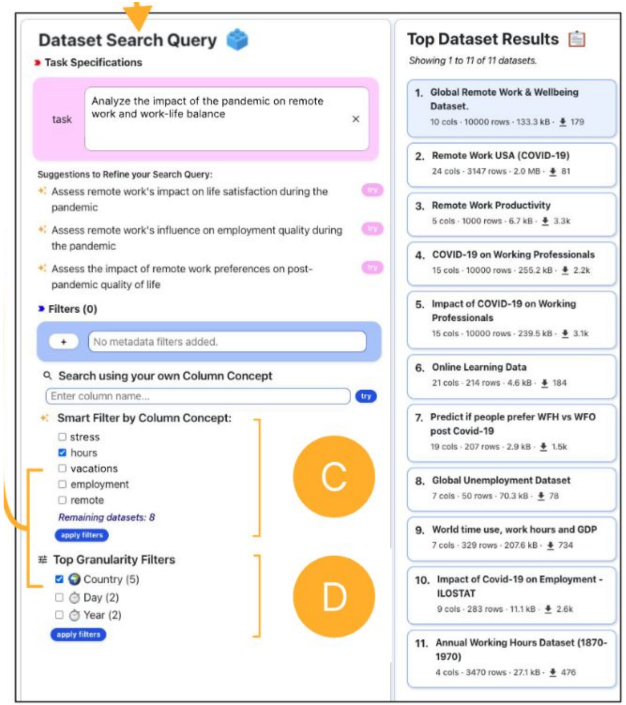
Keyword Search as Sensemaking

DataScout. Rachel Lin, Bhavya C., Wenjing L., Shreya S., Madelon H., Aditya P.



Keyword Search as Sensemaking

DataScout. Rachel Lin, Bhavya C., Wenjing L., Shreya S., Madelon H., Aditya P.



Keyword Search

Pros

- Fast, doesn't need access to actual data
- Filters and ranks datasets
- Dominant data search approach today

Cons

- Users need to evaluate datasets against actual data task
- Users in the critical path of search

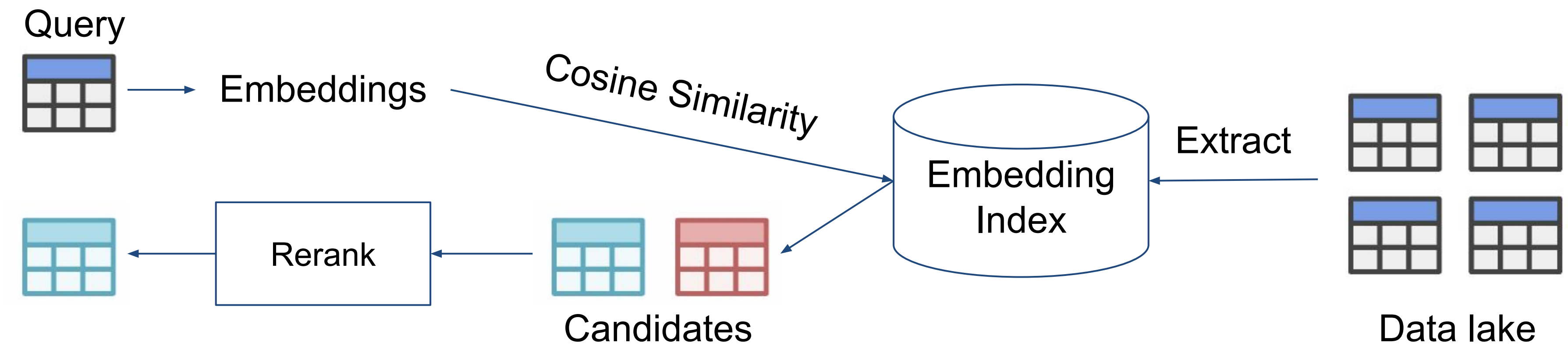
Data Discovery

Search by using a table or distribution as the query
Ling13,Zhu16,Nargesian18,Fernandez19,Rezig22,Santos21,Fan23,...

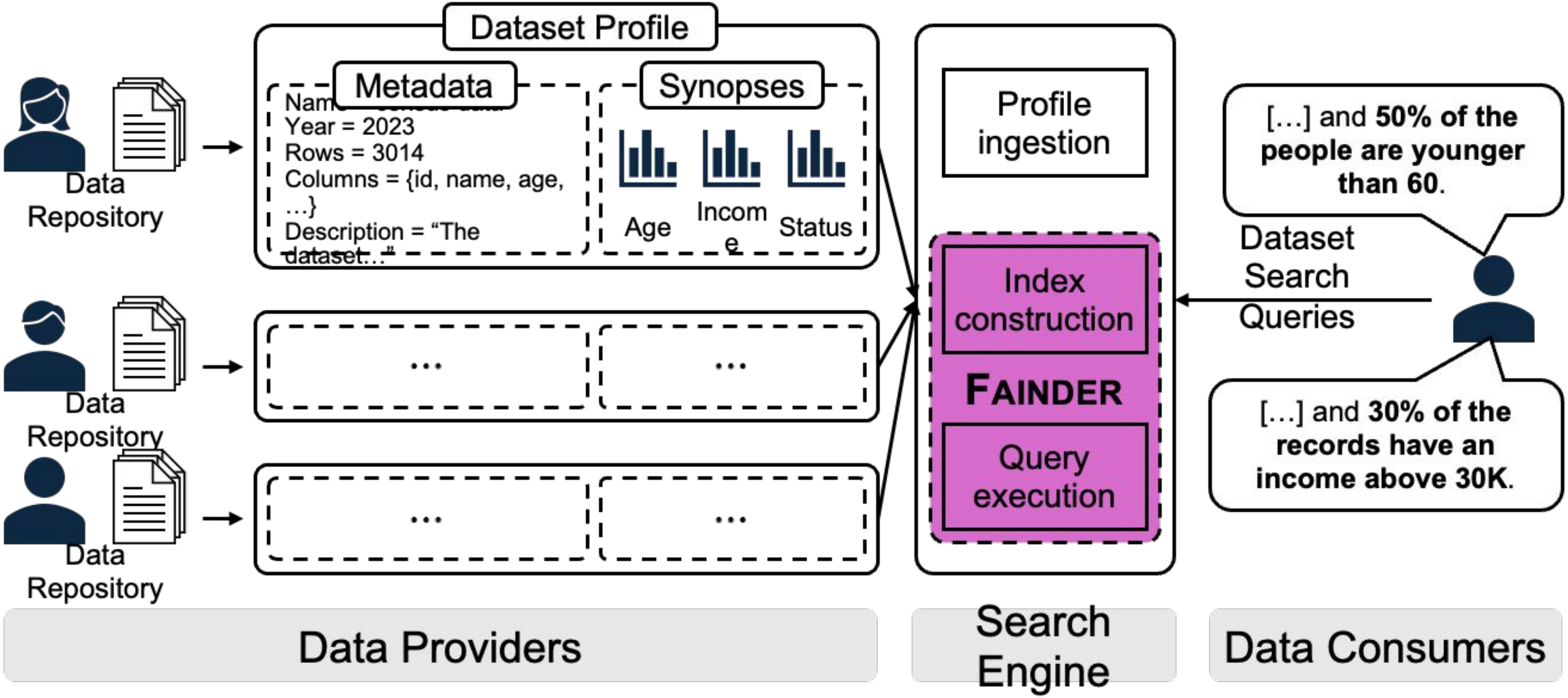
Rank based on

- Similarity,
- Joinability,
- Correlations,
- Unionability,
- Predicate satisfiability,
- …

Starmie: Table Union Search [Fan23]



Distribution-based Data Discovery [Behme24]



Data Discovery

Pros

- Results specific to the query table
- Scalable, leverages table representations

Cons

- Unless query is a retrieval task, users still need to evaluate datasets against actual data task
- Users in the critical path of search

Data Task as Search Query

Task $\mathrm{T(D)}\rightarrow\mathrm{goodness}$ is function of table D

Prediction ARDA, AUCTUS, Galhotra23

- $\mathrm{T(D)}$: train predictive model
- Given training dataset D, find augmentations that improve $\mathrm{T(D)}$

Causal Inference Suna Liu25, MetaM Galhotra23

- $\mathrm{T(D)}$: estimate Average Treatment Effect
- Given D with treatment and outcome, find likely confounders

Data Task as Search Query

Pros

- Ranks directly based on user’s task
- Can incorporate cleaning, integration, transformation

Potential Cons

- Evaluating task can be slow
- Hard to quantify task quality

Two Examples of Task-Based Search

Based on

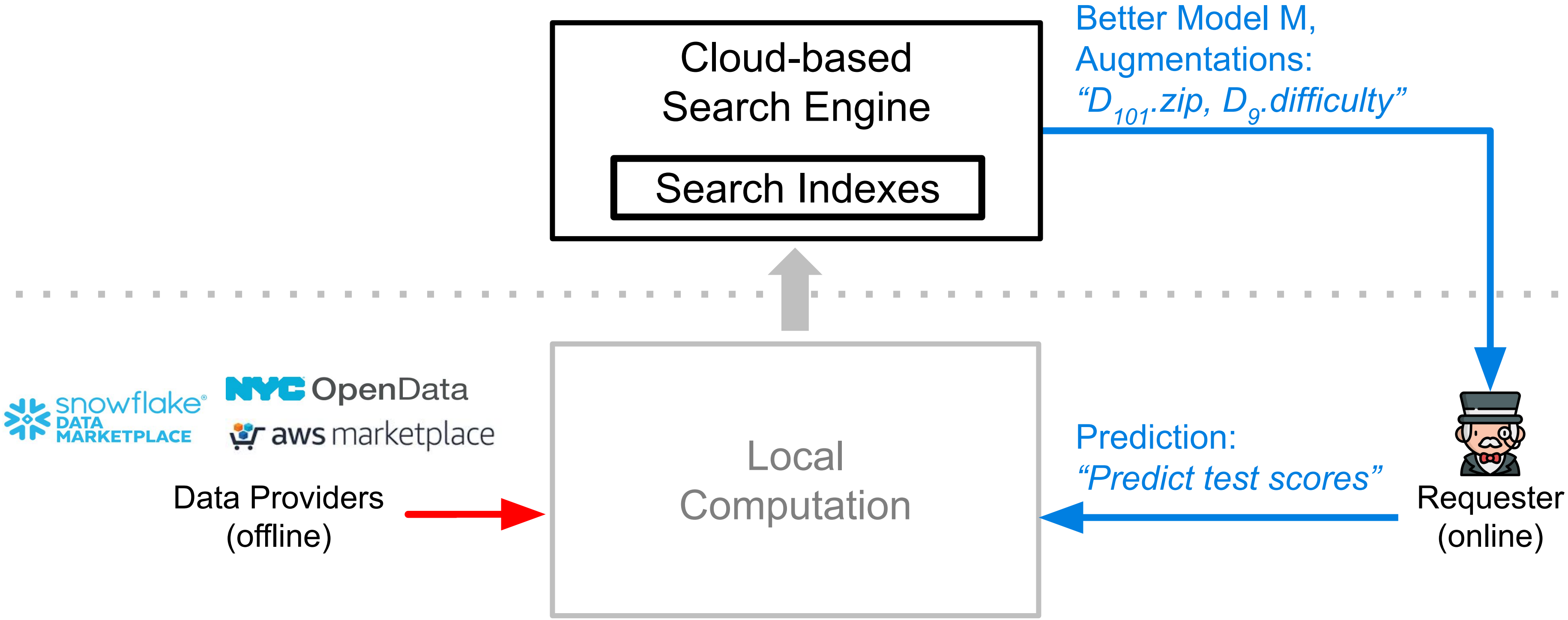
Kitana: A Data-as-a-Service Platform. Zach Huang23

The Fast and the Private: Task-based Dataset Search. Huang24

Saibot: A Differentially Private Data Search Platform. Huang23

Suna: Scalable Causal Confounder Discovery over Relational Data. Liu25

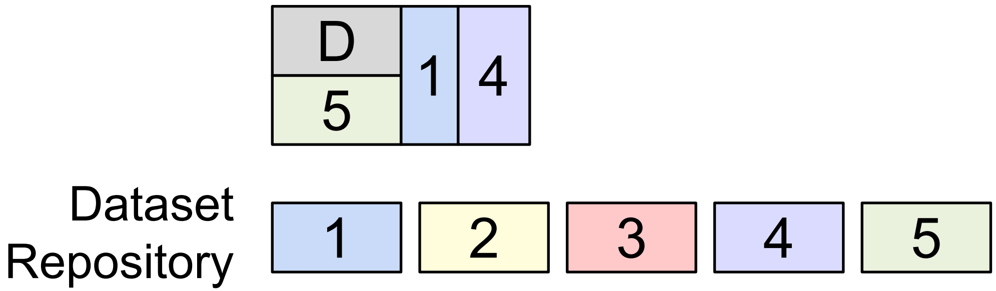
Data Task as Search Query



Prediction Task

Given training data D
greedily find augmentation plan A
that maximizes accuracy of model trained on A(D)

$$A(D) = D \cup 5 \bowtie 1 \bowtie 4$$



Basic Search Algorithm

```
D = initial training dataset
for A in all candidate augmentation plans
    eval(apply A to D)
return best A
```

Basic Search Algorithm

```
D = initial training dataset
for A in all candidate augmentation plans
    eval(apply A to D)
return best A
```

Basic Search Algorithm

```
D = initial training dataset
for A in all candidate augmentation plans
    eval(apply A to D)
return best A
```

Slow!

```
D = initial training dataset
for A in all candidate augmentation plans
    eval(apply A to D)
return best A
```

Combinatorial

Expensive!
Materialize A(D)
Retrain & Cross-validate

Reduce Search Space

- ARDA: join all relations + feature selection
- MetaM: cluster datasets and iteratively prune

Accelerate Eval()

- Auctus: find joinable correlations

Relies on access to raw data

Example System: Kitana

```
D = initial training dataset
for next augmentation α
    if eval(apply A to D) is best so far
        Keep α in A
return best A
```

Greedy Search

Expensive!
Materialize A(D)
Retrain & Cross-validate

Ideas

- Greedily find single best augmentation in each iteration
- Use sketches to accelerate & parallelize `eval()`

Sketches: Count(D⋈S)

Naïve join generates big intermediate relation

D

A	B
1	1
1	2
2	3

→ ⋈ ←

S

A	C
1	4
1	5
2	6

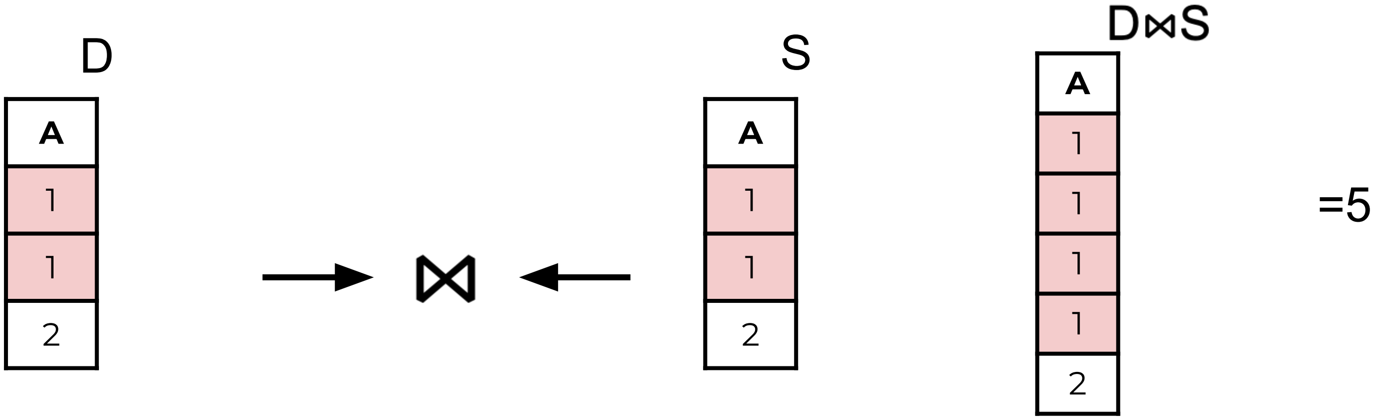
D⋈S

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

=5

Sketches: Count(D⋈S)

Optimization: drop irrelevant columns



Sketches: Count(D⋈S)

Optimization: sufficient statistics

Sketches

D

A
1
1
2

A	cnt
1	1+1
2	1

⋈

A	cnt
1	1+1
2	1

S

A
1
1
2

D⋈S

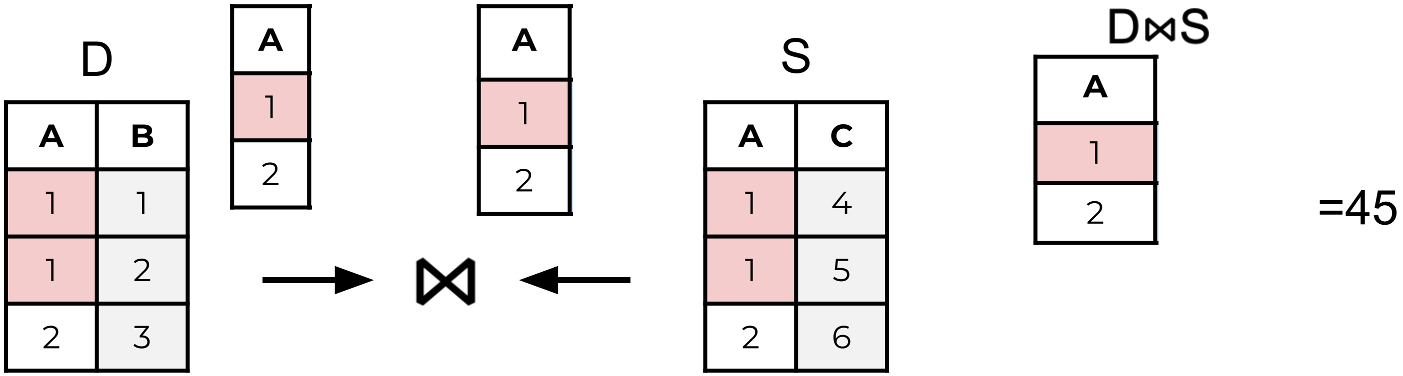
A	cnt
1	2x2
2	1x1

=5

Sketches: $\mathrm{Sum}_{b*c}(D \bowtie S)$

Optimization: sufficient statistics

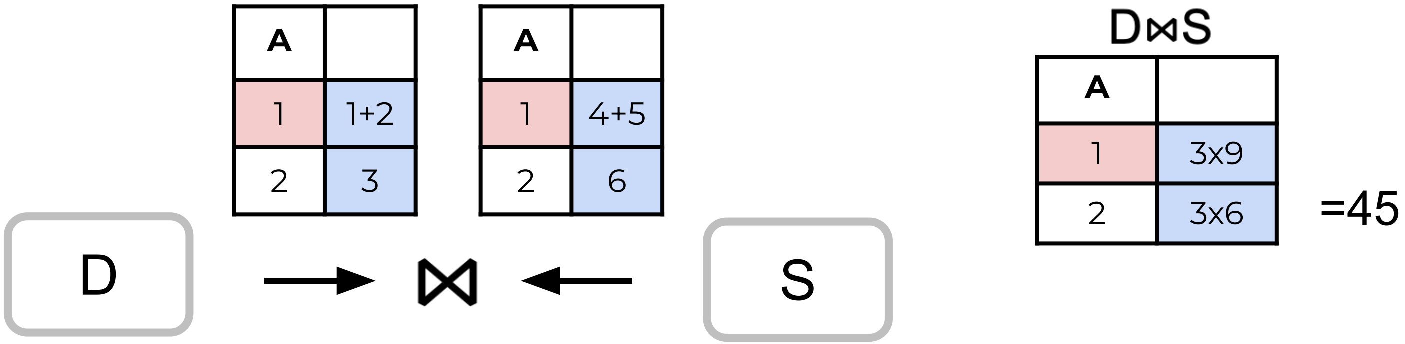
Sketches defined for common stats, ML models.



Sketches: trainAndEval(D⋈S)

Optimization: sufficient statistics

Sketches defined for common stats, ML models.

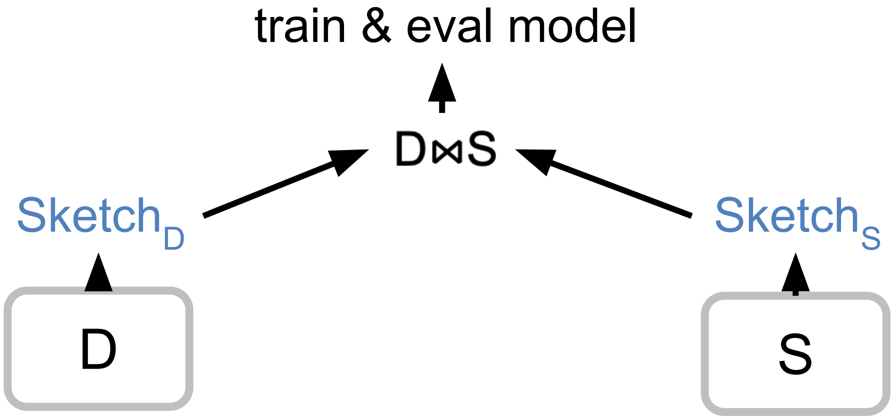


Sketches: train(D⋈S)

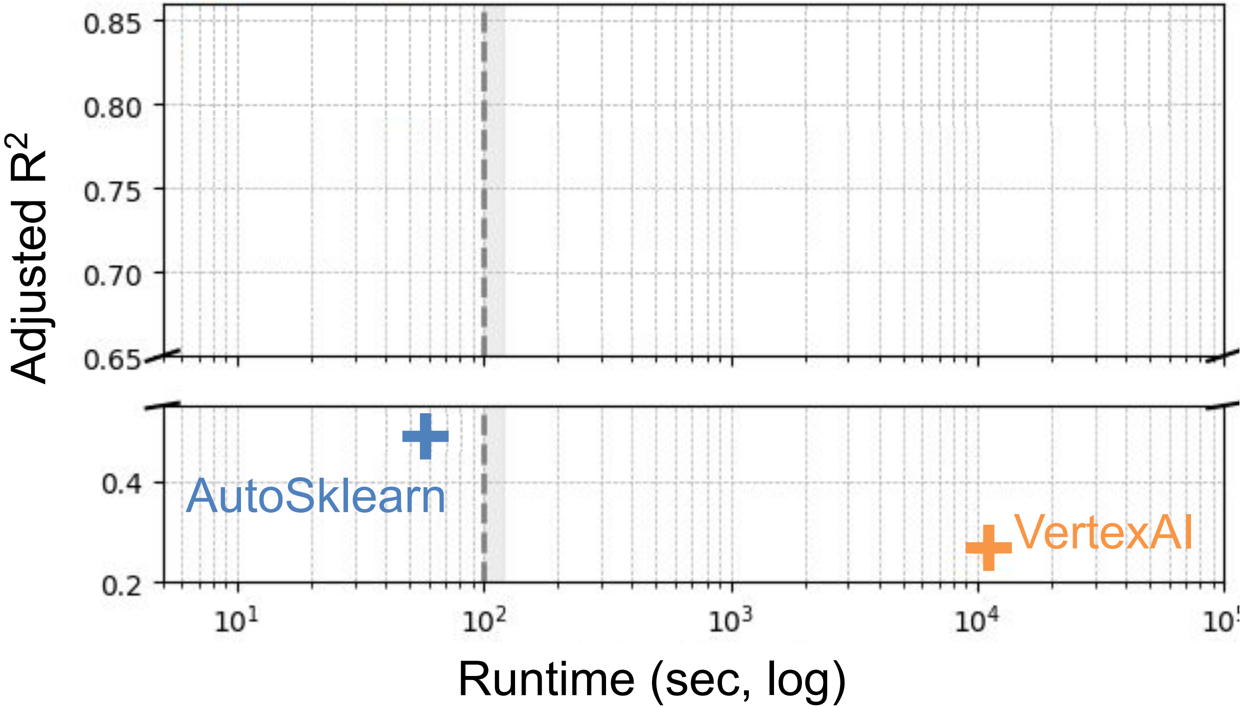
Optimization: sufficient statistics

Sketches defined for common stats, ML models.

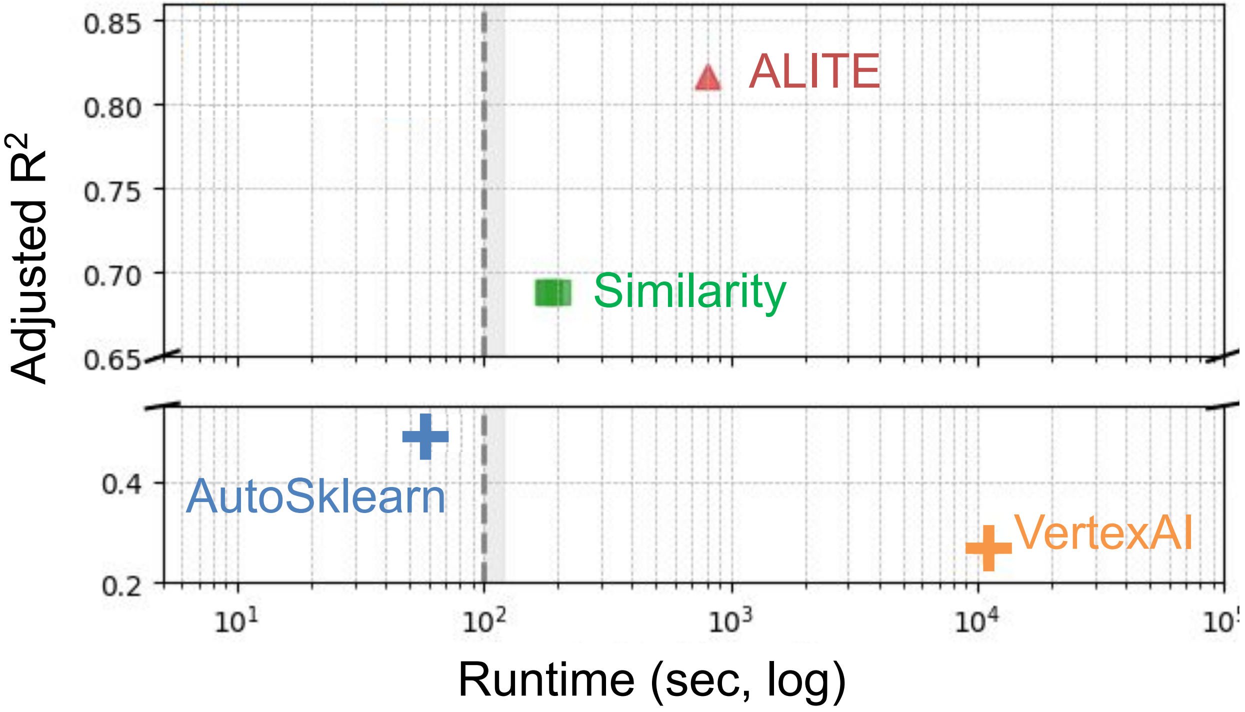
Linear Regression as a *proxy model* during search



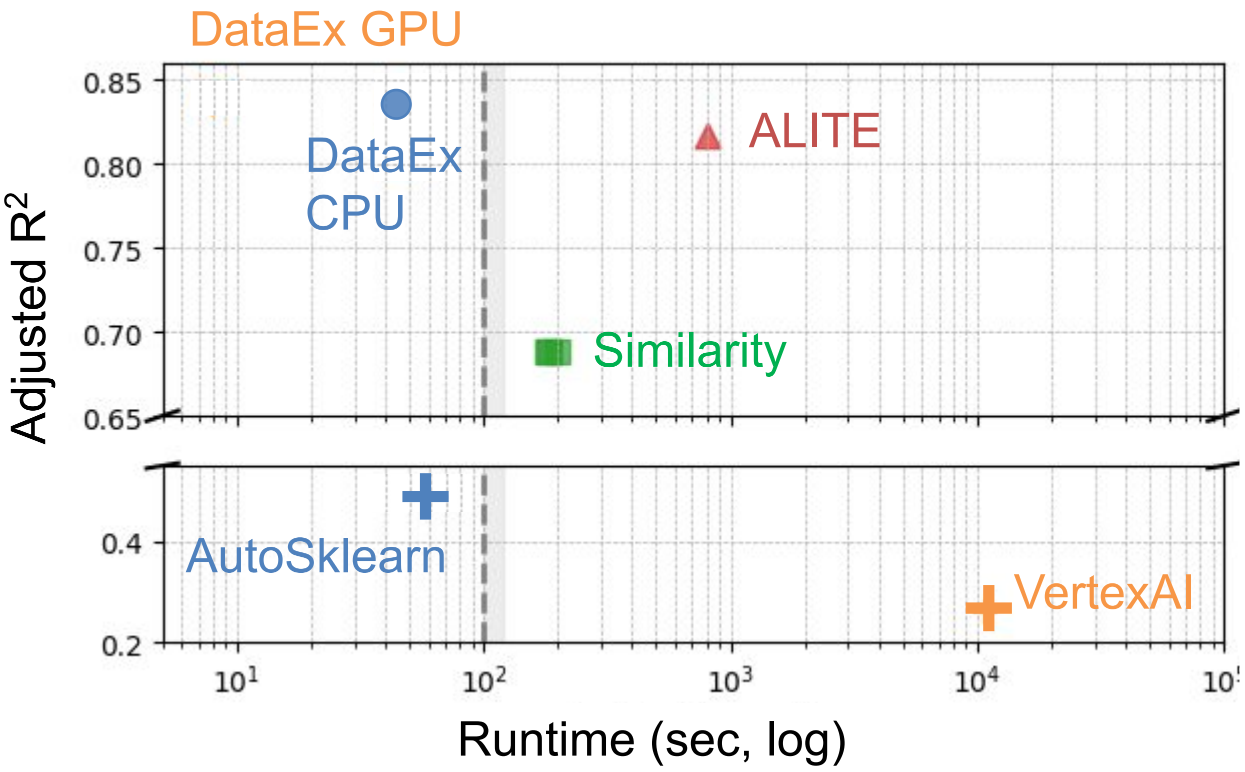
Evaluation on 8376 Kaggle Tables



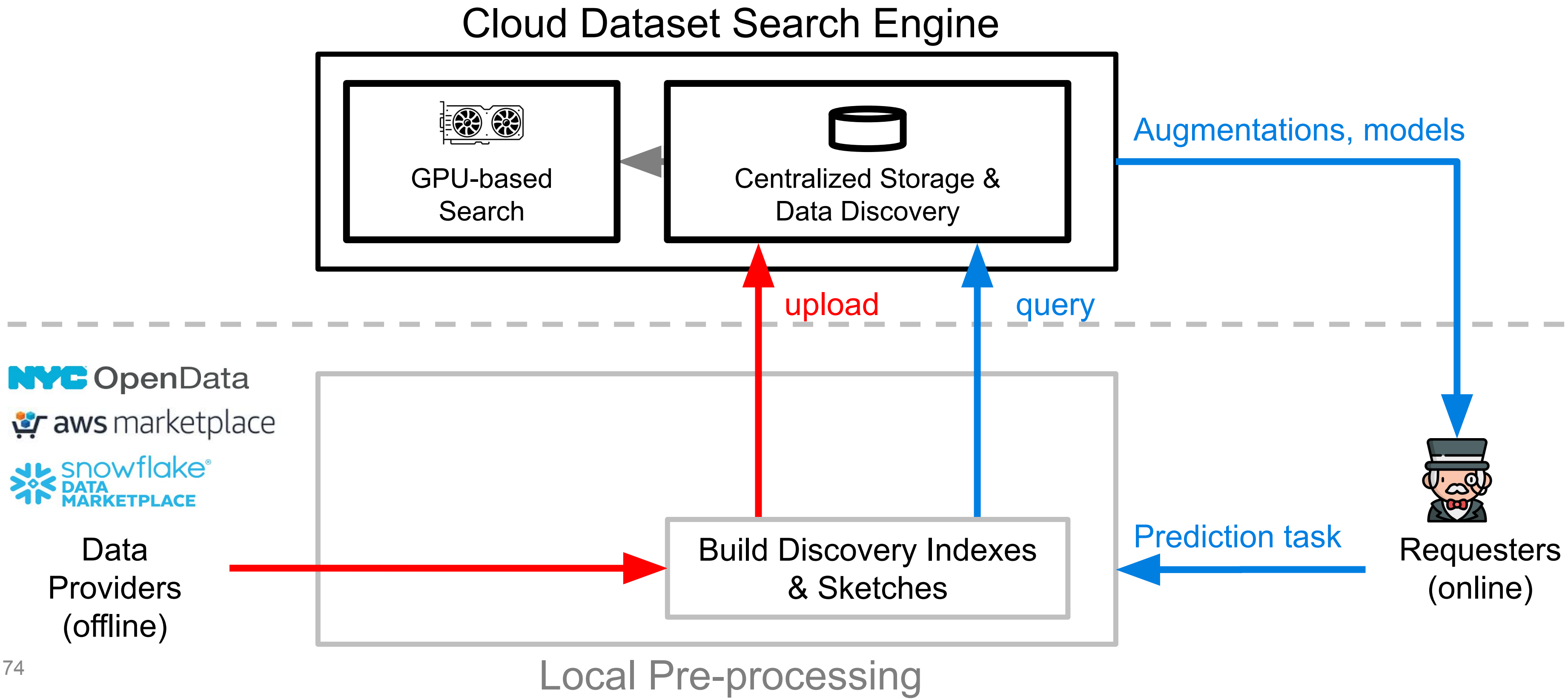
Evaluation on 8376 Kaggle Tables



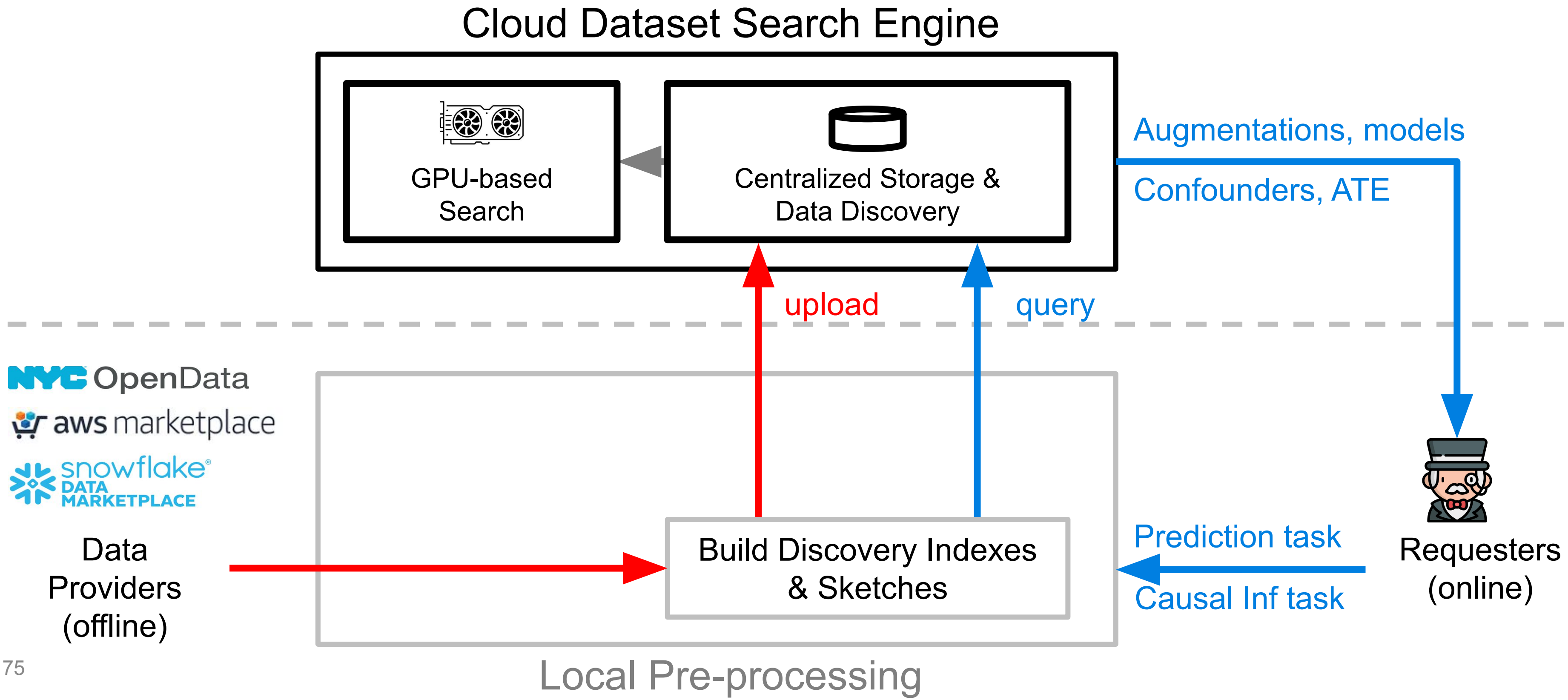
Evaluation on 8376 Kaggle Tables

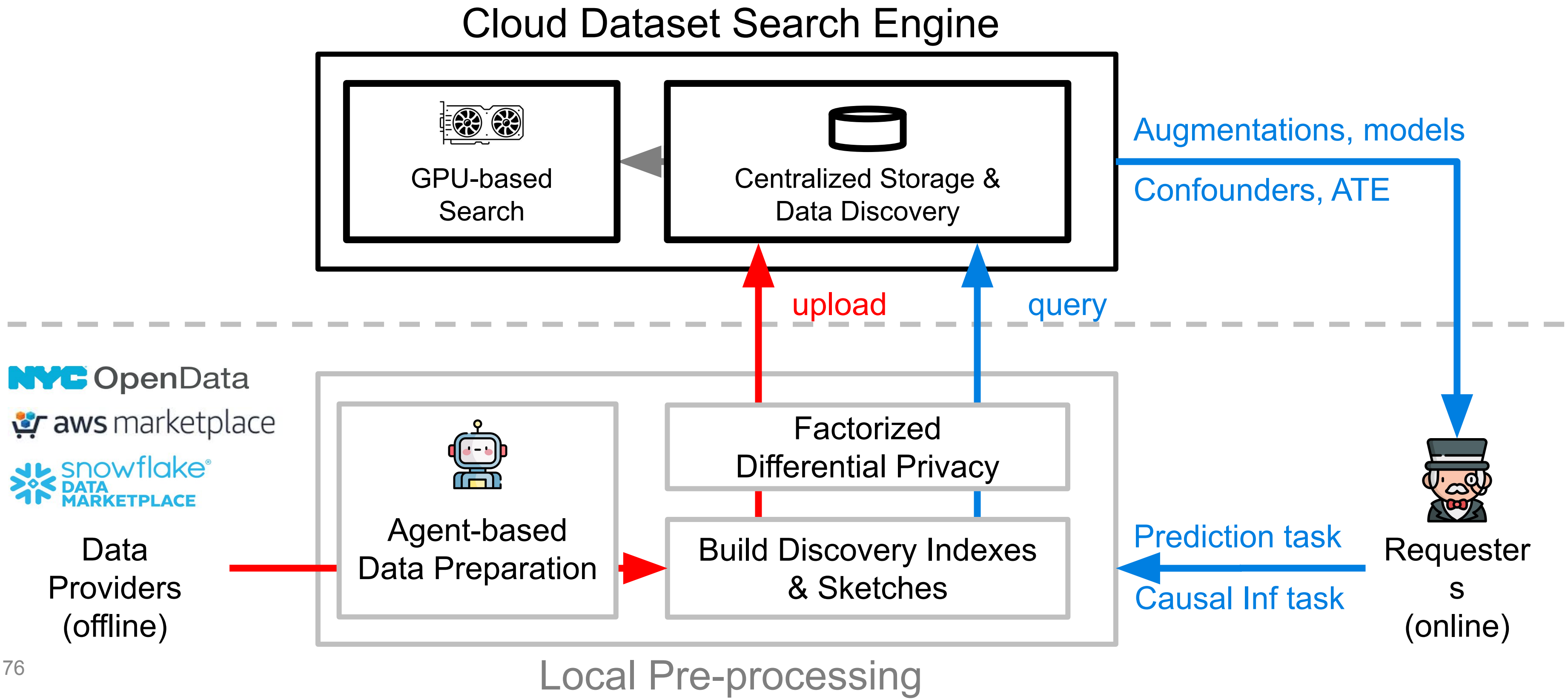


DataEx

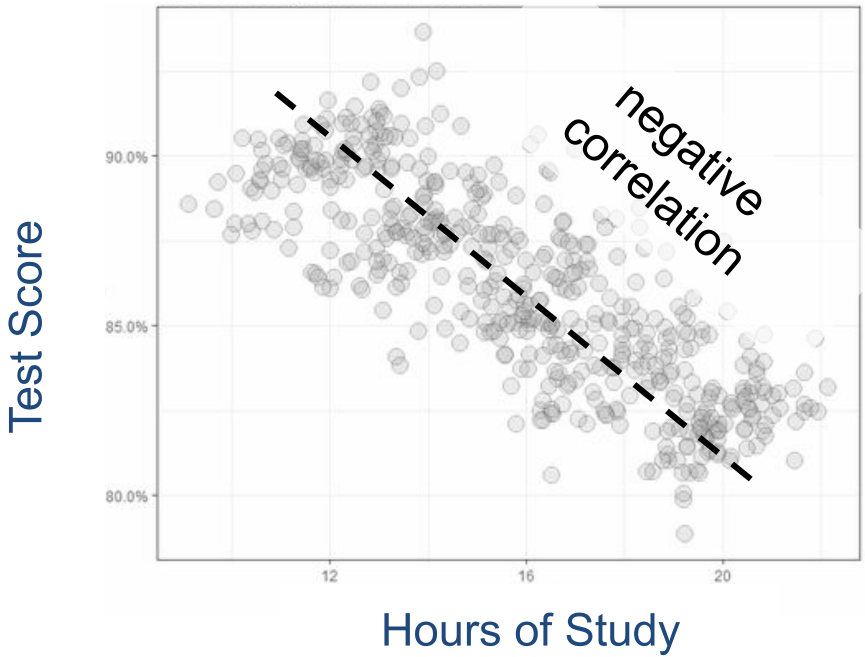

Cloud Dataset Search Engine
GPU-based
Search
Centralized Storage &
Data Discovery
Augmentations, models
upload
query
NYC OpenData
aws marketplace
snowflake
DATA
MARKETPLACE
Data
Providers
(offline)
Build Discovery Indexes
& Sketches
Prediction task
Requesters
(online)
Local Pre-processing

DataEx

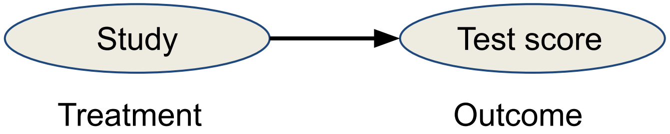

Cloud Dataset Search Engine
GPU-based Search
Centralized Storage & Data Discovery
Augmentations, models
Confounders, ATE
upload
query
NYC OpenData
aws marketplace
snowflake DATA MARKETPLACE
Data Providers (offline)
Build Discovery Indexes & Sketches
Prediction task
Causal Inf task
Requesters (online)
Local Pre-processing


Cloud Dataset Search Engine
GPU-based Search
Centralized Storage & Data Discovery
Augmentations, models
Confounders, ATE
upload
query
NYC OpenData
aws marketplace
snowflake DATA MARKETPLACE
Data Providers (offline)
Agent-based Data Preparation
Factorized Differential Privacy
Build Discovery Indexes & Sketches
Prediction task
Causal Inf task
Requester
s
(online)
Local Pre-processing

Confounders in Causal Analysis



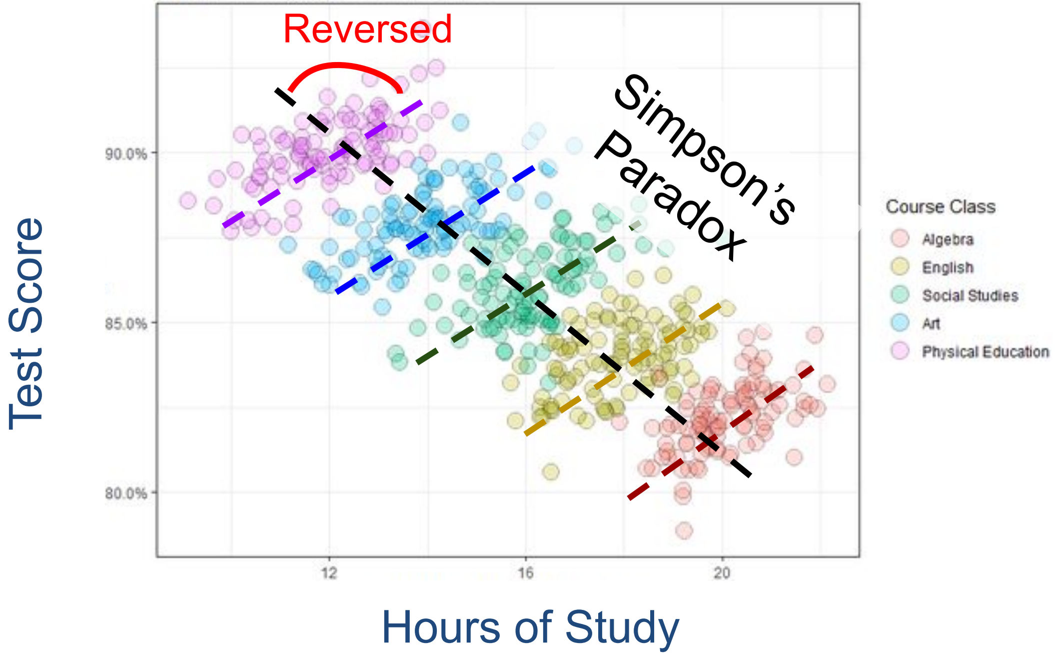
Causal Diagram



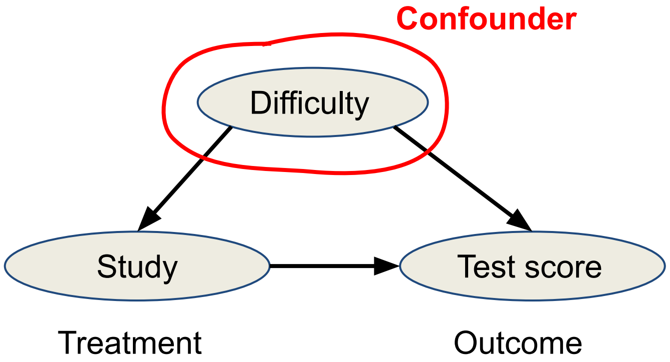
Studying causes poor grades?
Study → TestScores

https://math.stackexchange.com/questions/83756/more-examples-of-simpsons-paradox-barring-the-ones-on-wikipedia-titanic-and

Confounders in Causal Analysis



Causal Diagram



Studying causes poor grades?

Why does the trend reverse?

Study → TestScores

https://math.stackexchange.com/questions/83756/more-examples-of-simpsons-paradox-barring-the-ones-on-wikipedia-titanic-and

Confounders in Causal Analysis

User Query

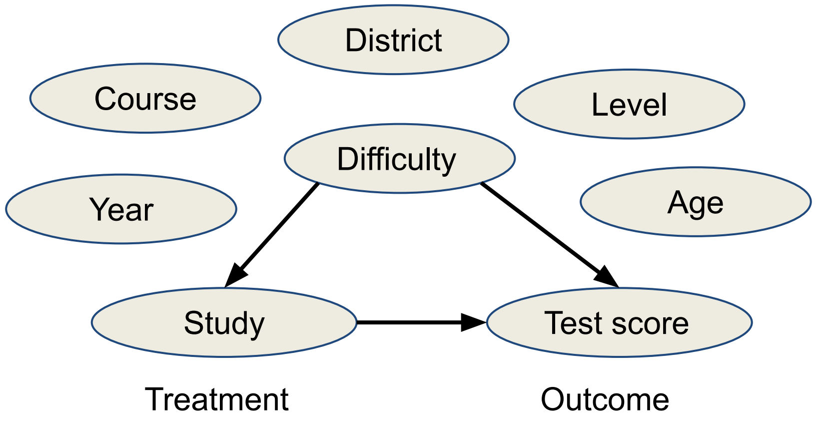
	Treatment	Outcome
ID	Study	Score
1	15 hr	75
2	10 hr	90
3	20 hr	85

Data Repository

ID	Course	Difficulty
1	CS101	3
2	CS102	1
3	CS103	4

ID	District
1	1
2	2
3	3

...



Confounders in Causal Analysis

User Query

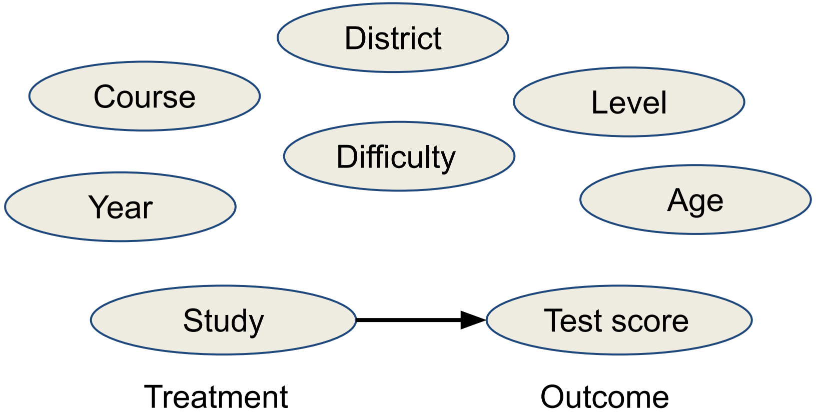
	Treatment	Outcome
ID	Study	Score
1	15 hr	75
2	10 hr	90
3	20 hr	85

Data Repository

ID	Course	Difficulty
1	CS101	3
2	CS102	1
3	CS103	4

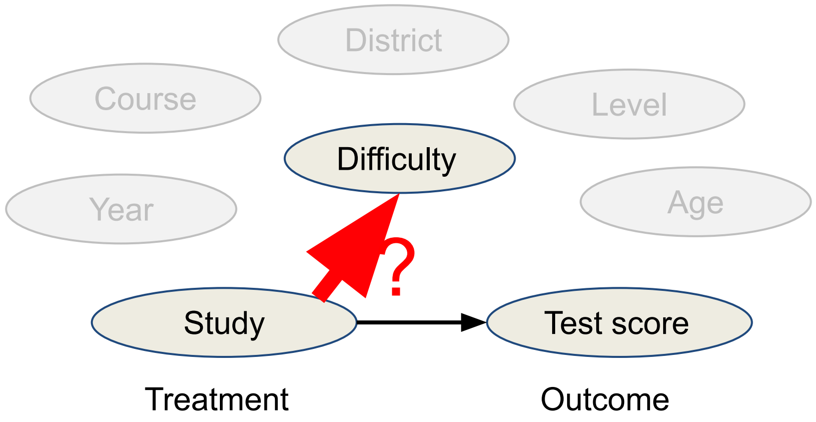
ID	District
1	1
2	2
3	3

...



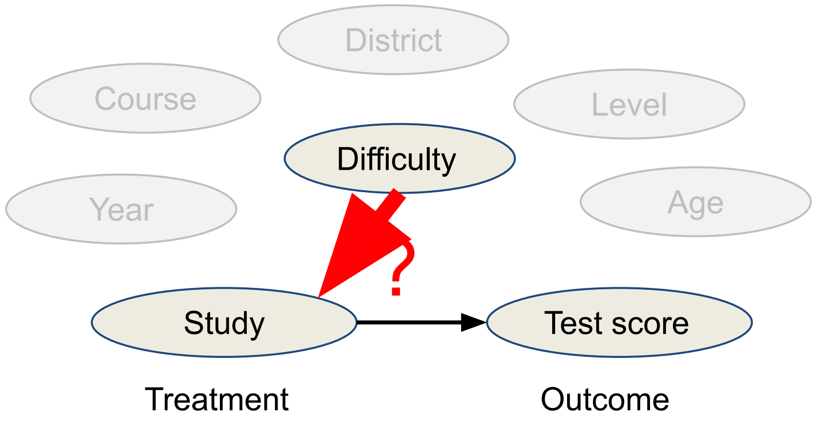
Confounders in Causal Analysis

Background: Bivariate Causal Discovery (BCD) estimates → or ← edges from data



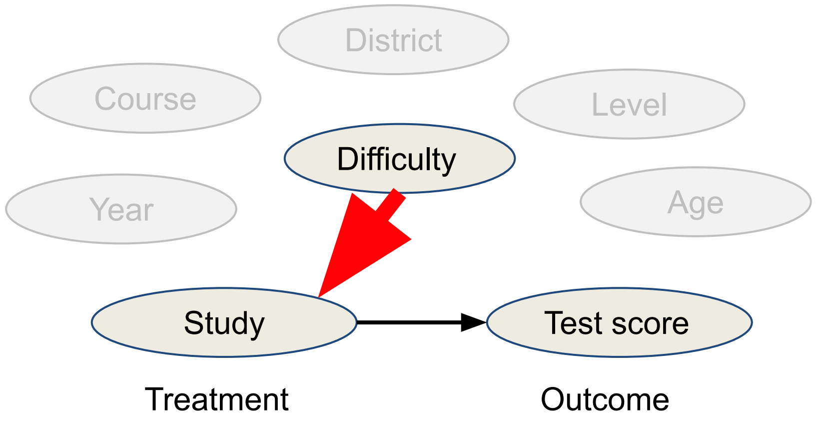
Confounders in Causal Analysis

Background: Bivariate Causal Discovery (BCD) estimates → or ← edges from data



Confounders in Causal Analysis

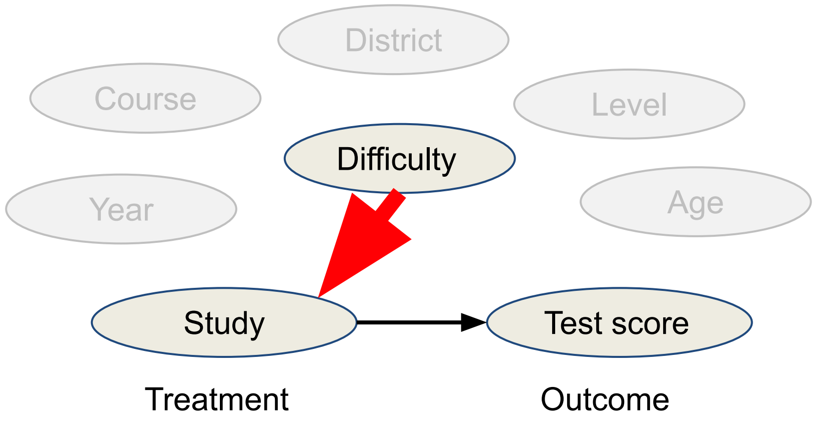
Background: Bivariate Causal Discovery (BCD) estimates → or ← edges from data



Confounders in Causal Analysis

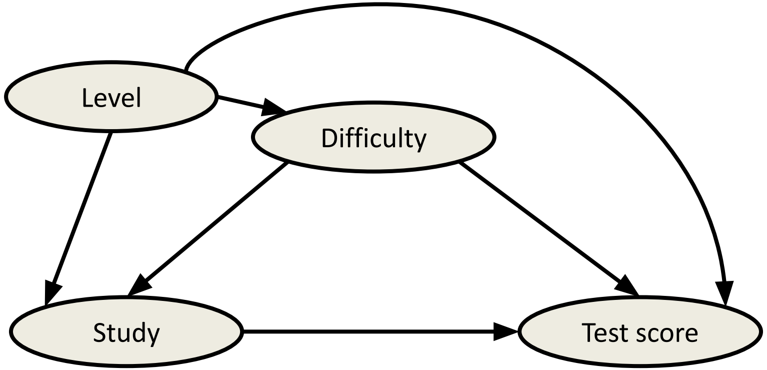
Background: Bivariate Causal Discovery (BCD) estimates → or ← edges from data

Proof: existence of confounder reduces to BCD estimating "*Ancestors* ↝ *Treatment*"



Discovering Confounders with BCD

Building adjustment set for Study ⟶ TestScore

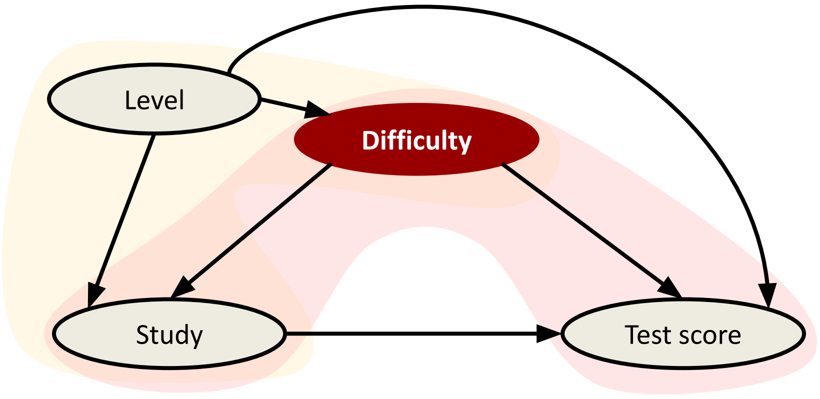


Key Observation:

treatment and outcome confounded: BCD will flag confounder ⟶ treatment.

Discovering Confounders with BCD

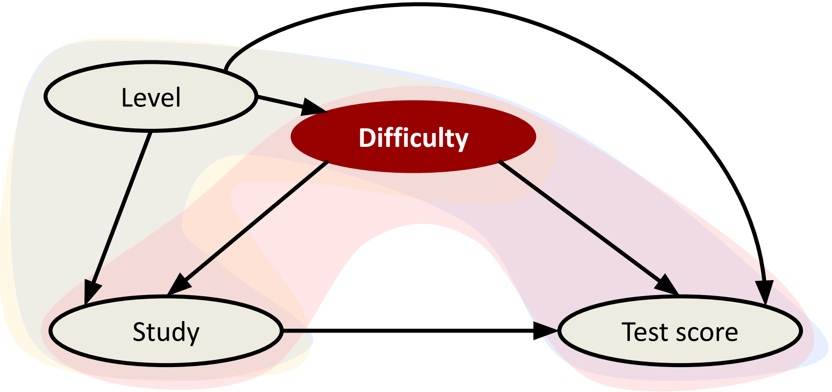
Building adjustment set for Study ⟶ TestScore



Difficulty is a confounder, not flagged by BCD because confounded by **Level**

Discovering Confounders with BCD

Building adjustment set for Study ⟶ TestScore

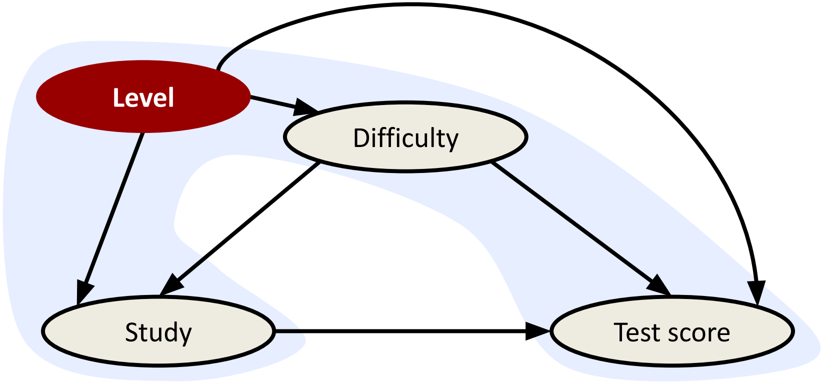


Difficulty is a new path for causal effect

Confounder, not flagged by BCD because confounded by **Level**

Discovering Confounders with BCD

Building adjustment set for Study ⟶ TestScore



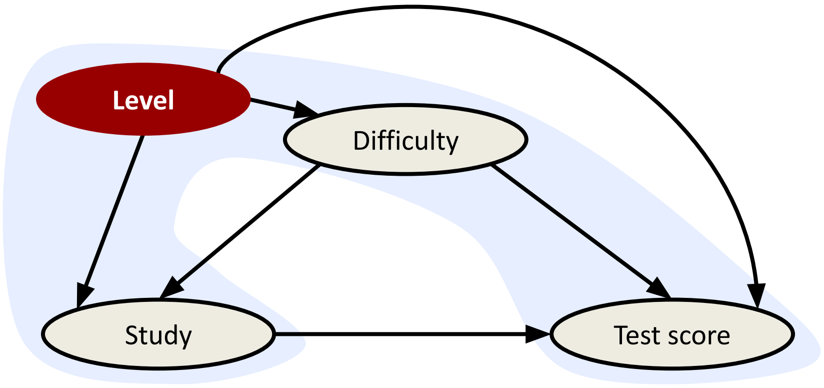
Level is also a confounder, flagged by BCD

Key Insight:

Level < Difficulty topologically – we prove a confounder always flagged by BCD

Discovering Confounders with BCD

Building adjustment set for Study ⟶ TestScore



Theorem 1: *If ∃ confounder between treatment and outcome, ∃ attribute **A** s.t*

- ***A*** → treatment and ∄ *confounder between **A** and treatment.* **Flagged by BCD**
- ***A** is a confounder between treatment and outcome.* **Selected heuristically**

Confounders in Causal Analysis

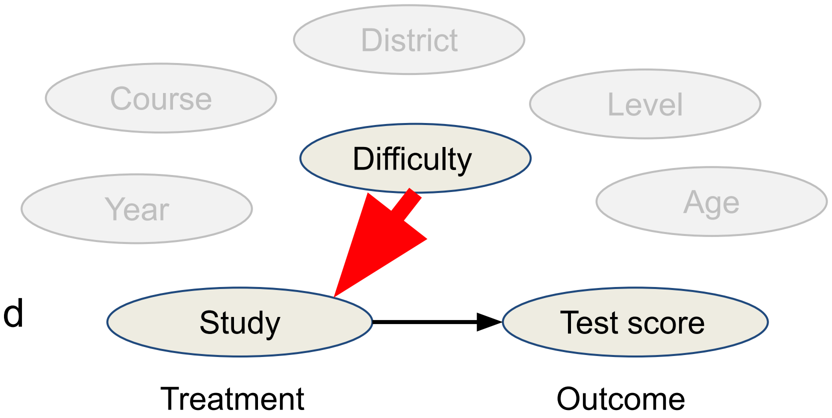
Background: Bivariate Causal Discovery (BCD) estimates → or ← edges from data

Proof: existence of confounder reduces to BCD estimating "*Ancestors* ↝ *Treatment*"

Algorithm: Use BCD to find superset of *Ancestors* and iteratively reduce until it is an admissible set.

System: develop novel sketches to accelerate BCD evaluation, scale using GPUs

- Level= β_1·Study + ϵ_1
- Estimate: MI(Study, Level-β_1·Study)
- Push mutual information through joins

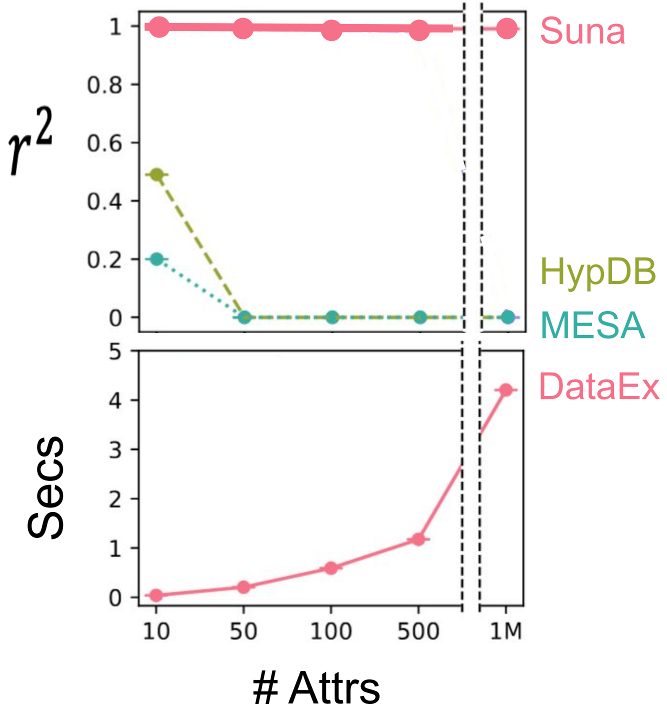


Experimental Results

Real Data: Reproduces Known Confounders

Dataset	Query	Suna
SO	What is the effect of education level on salary?	Cost of Living & Rent Index
ELA	What is the effect of each school's extra credit performance score on students' ELA score?	Enrollment % Poverty
Ratio	What is the effect of each school's pupil-to-teacher ratio on student's ELA score?	Level 4: % % Students with Disabilities Minimum Class Size
SAT	What is the effect of test takers numbers on SAT score?	# Safety Incidents Enrollment Total Regents #

Synthetic Data: Accurate & Fast



Summary of Task-based Search

Task-evaluation is bottleneck

- Identify hardware and parallelization-friendly sketches to accelerate task evaluation

Need algorithms to avoid combinatorial search

Arbitrary tasks can be supported, but are very difficult…

Metam: Task-agnostic search [Galhotra23]

Problem Setup

GIVEN:

An initial dataset D,

a collection of attributes Γ,

a task implementation t

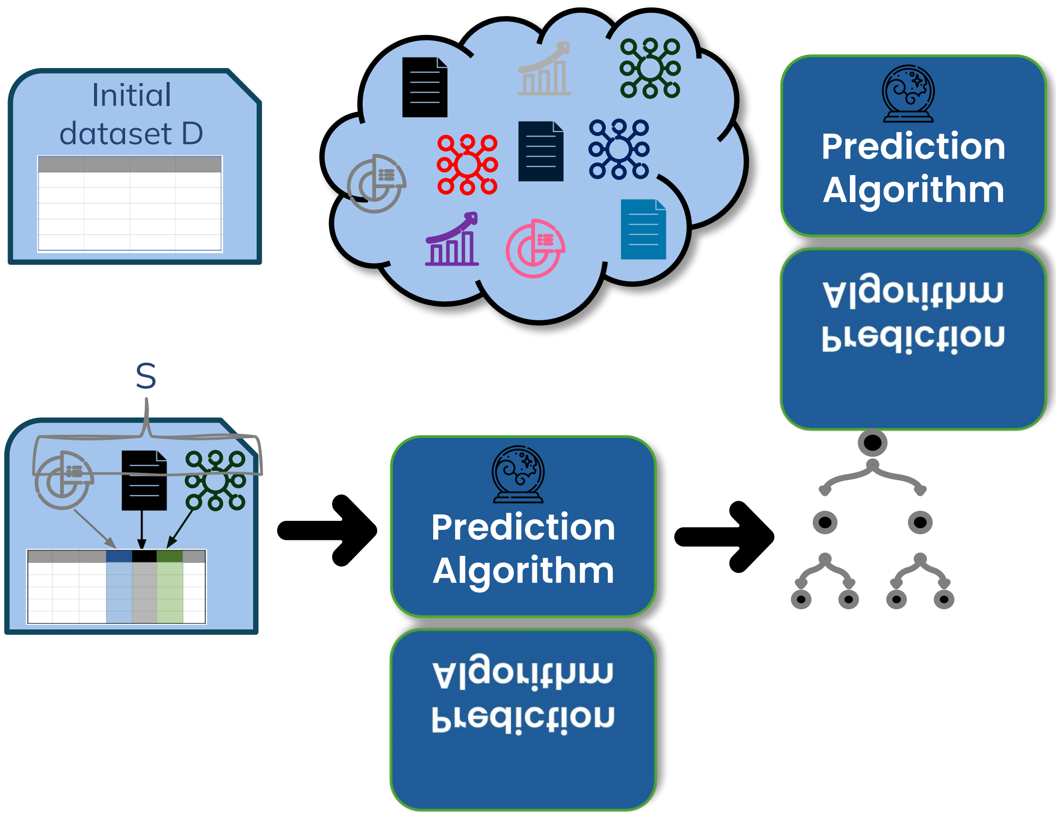
OBJECTIVE: max $\text{utility}_t\,(D \bowtie S)$

CONSTRAINT:

$S \subseteq \Gamma$

$|S| \leq k$ (a constant)

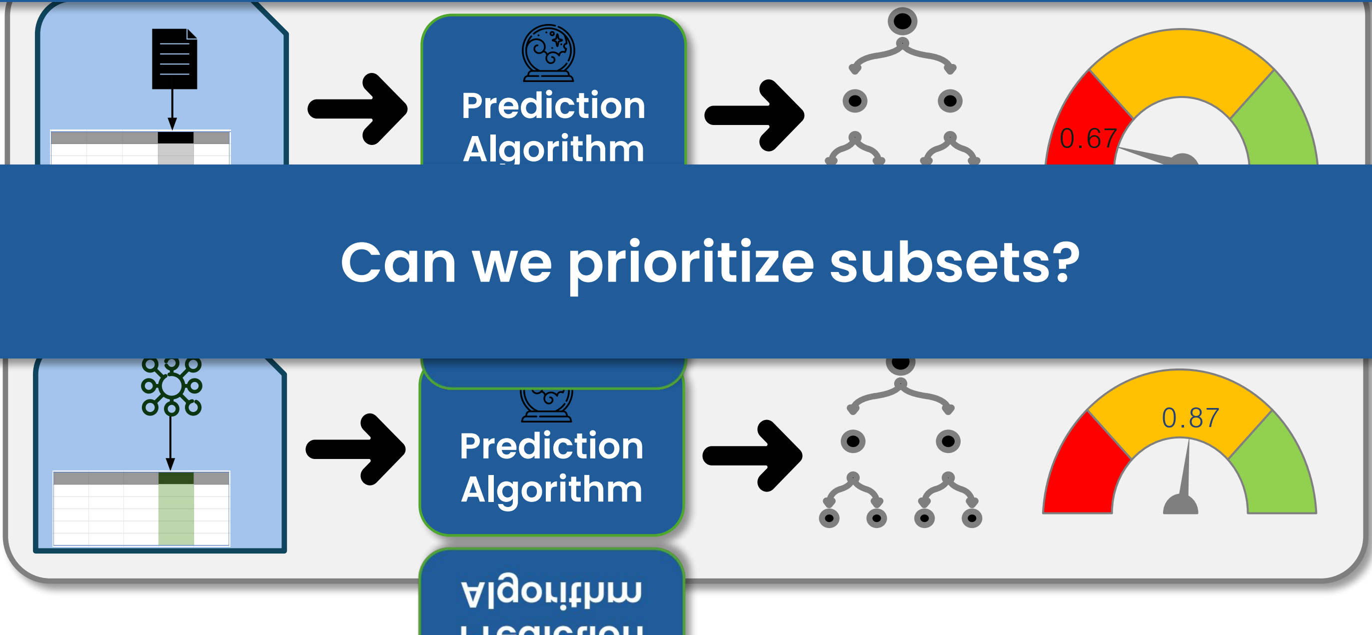
S is minimal



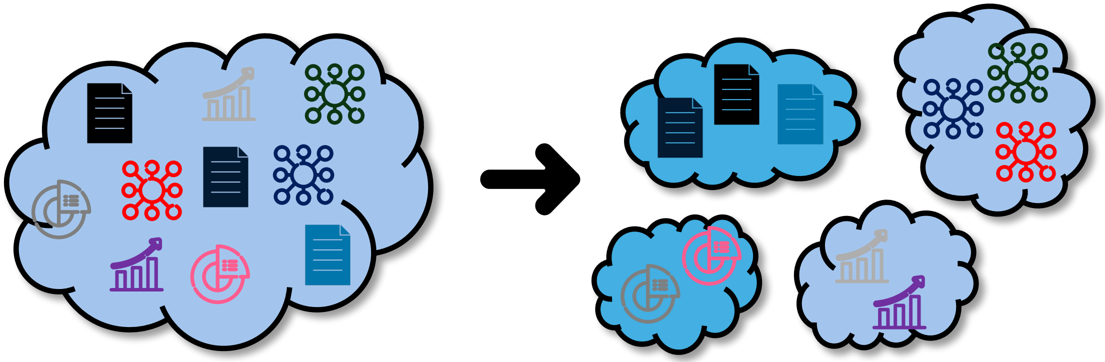
How to solve the problem?

⚠️ Requires n^k queries! Infeasible when n is in the order of millions

Can we prioritize subsets?



Clustering helps to diversify the search process



Similar datasets have similar utility!

Using Data Properties As Features To Cluster

Id	Address	....	Zip Code	Crime
1	153 JFK, NY		12543	Low
2	543 Albert Street, NY		?	?
3	432 MK road		14656	High
4	5432 Dud Dr		54637	Low
5	6732 Psycho Path		?	?
6	23 Main Street		?	?

Properties of the newly added attribute

Fraction of missing values: 0.4

Correlation (Crime, Area): 0.65

Approach 1: Diversify the Search Process

IDEAL SCENARIO: Probability of sampling an informative attribute from the cluster C

EXPLORE-EXPLOIT DILEMMA:

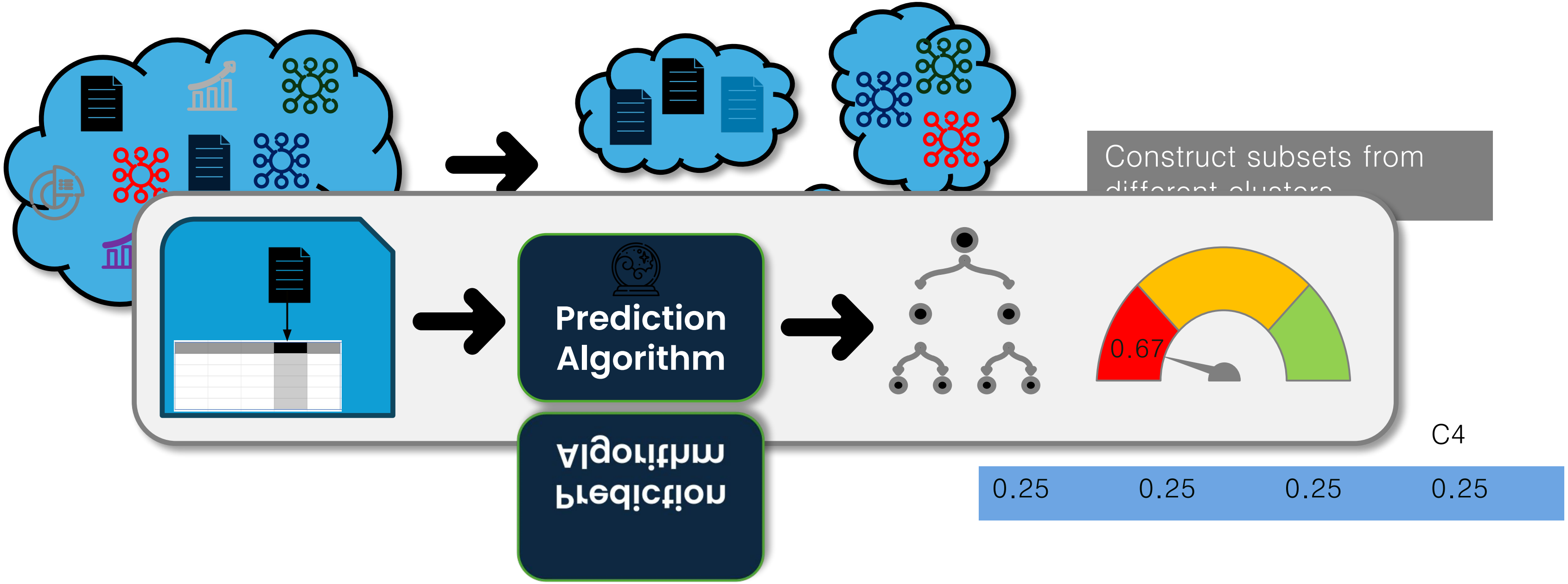
Should I sample more datasets from cluster C_i?

OR

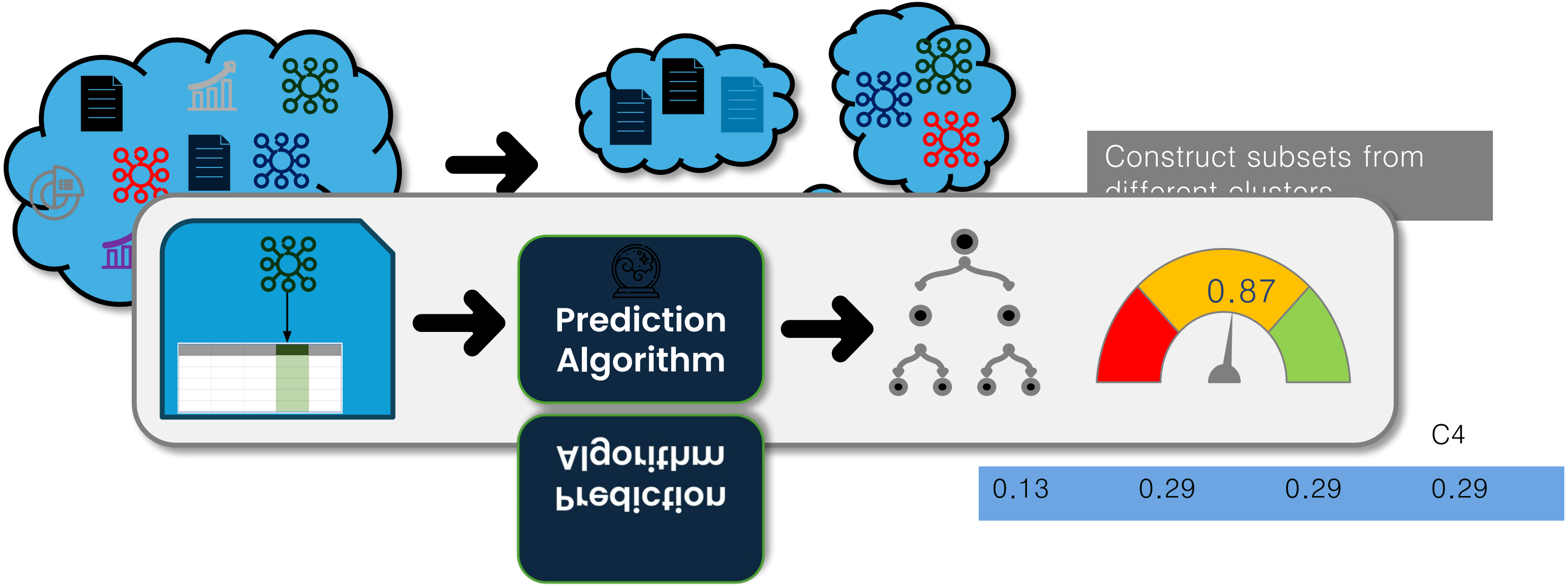
Should I explore different clusters?

SOLUTION: Bandit-based approach

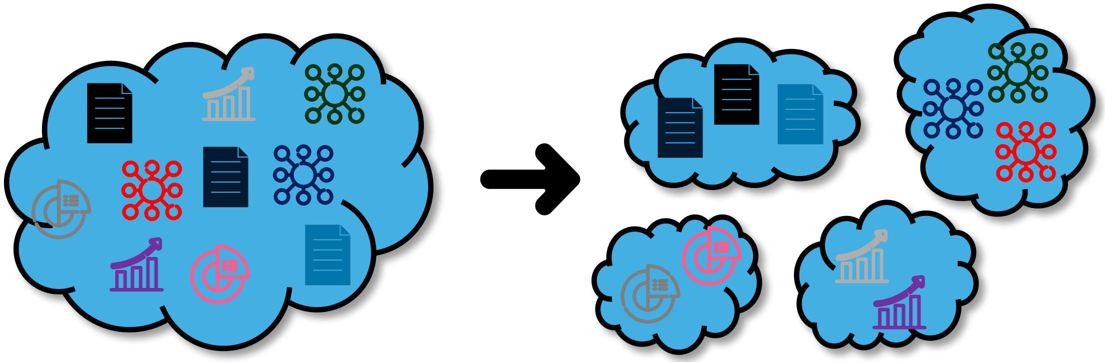
Approach 1: Diversify the Search Process



Approach 1: Diversify the Search Process



Approach 1: Diversify the Search Process



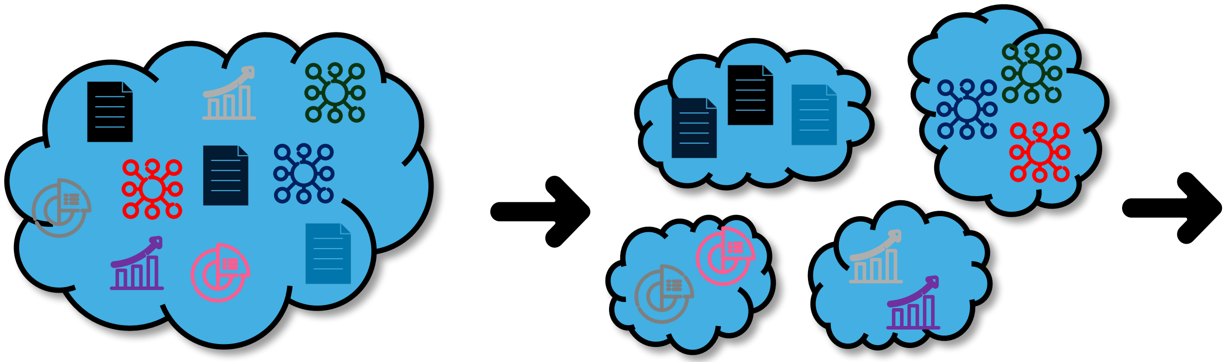
Construct subsets from different clusters

Number of queries: 100

C1	C2	C3	C4
0.02	0.84	0.10	0.04

Converged scores

Approach 1: Diversify the Search Process



Bandit–based

Prediction Algorithm

Leverages data profiles to search efficiently

Complexity $O(|C|^k)$

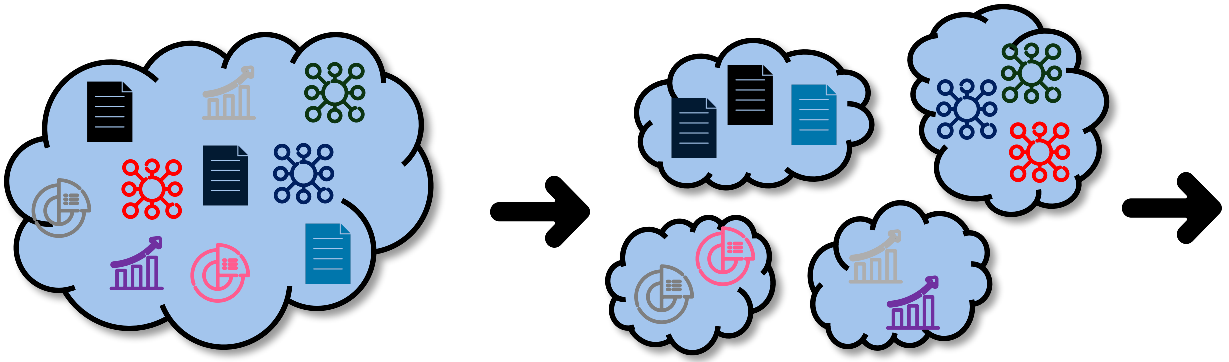
Approach 2: Leverage Monotonicity of Utility Metric

- **Monotonicity**: Easy to guarantee
  $u(D \cup T_2) \geq u(D \cup T_1) \; \forall \, T_1 \subseteq T_2$

- What if the utility is **submodular** too?
  - Diminishing returns property:
    $u(T_1 \cup \{X\}) - u(T_1) \geq u(T_2 \cup \{X\}) - u(T_2) \;\; \forall \, T_1 \subseteq T_2$

Solution: Greedily choose the best augmentation

Approach 2: Leverage Monotonicity of Utility Metric



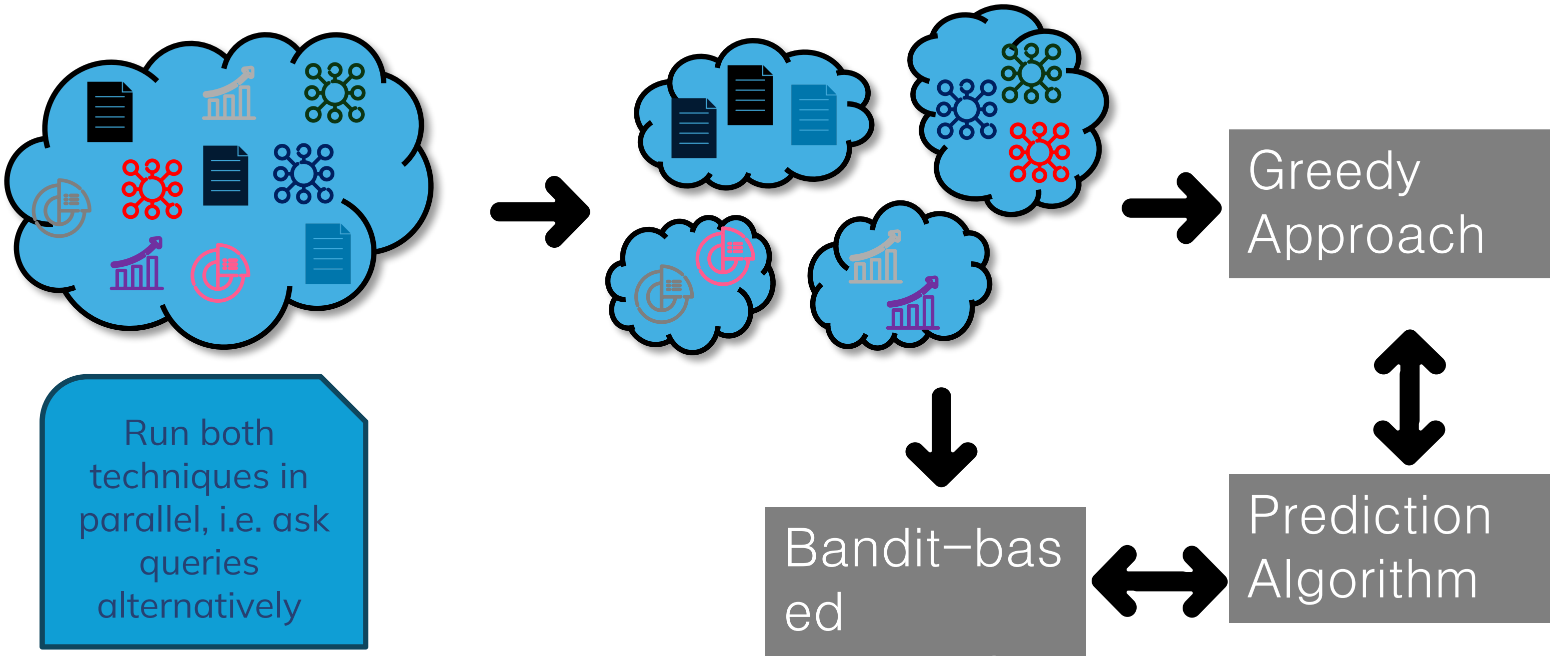
Greedy Approach

Prediction Algorithm

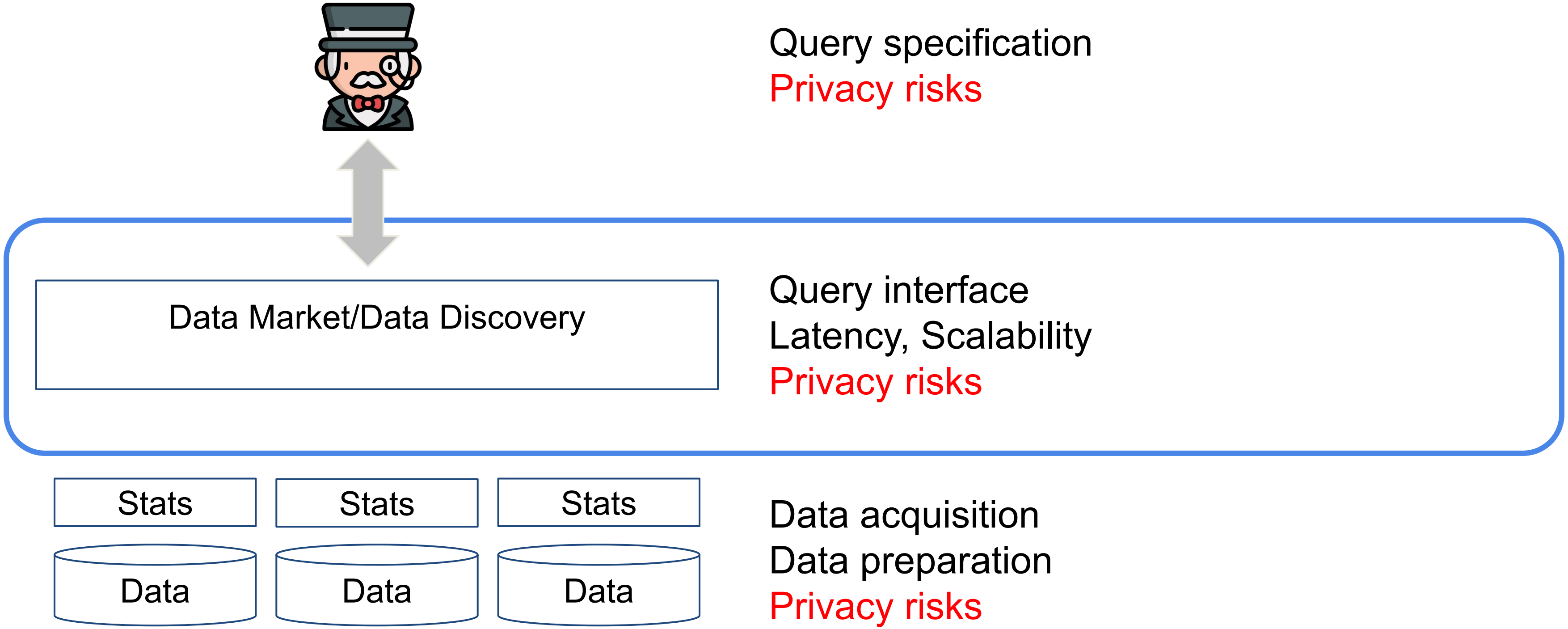
Highly efficient to identify useful augmentations

Can fail if utility is super modular (unlikely but possible)

Final Approach: Combining All Ideas



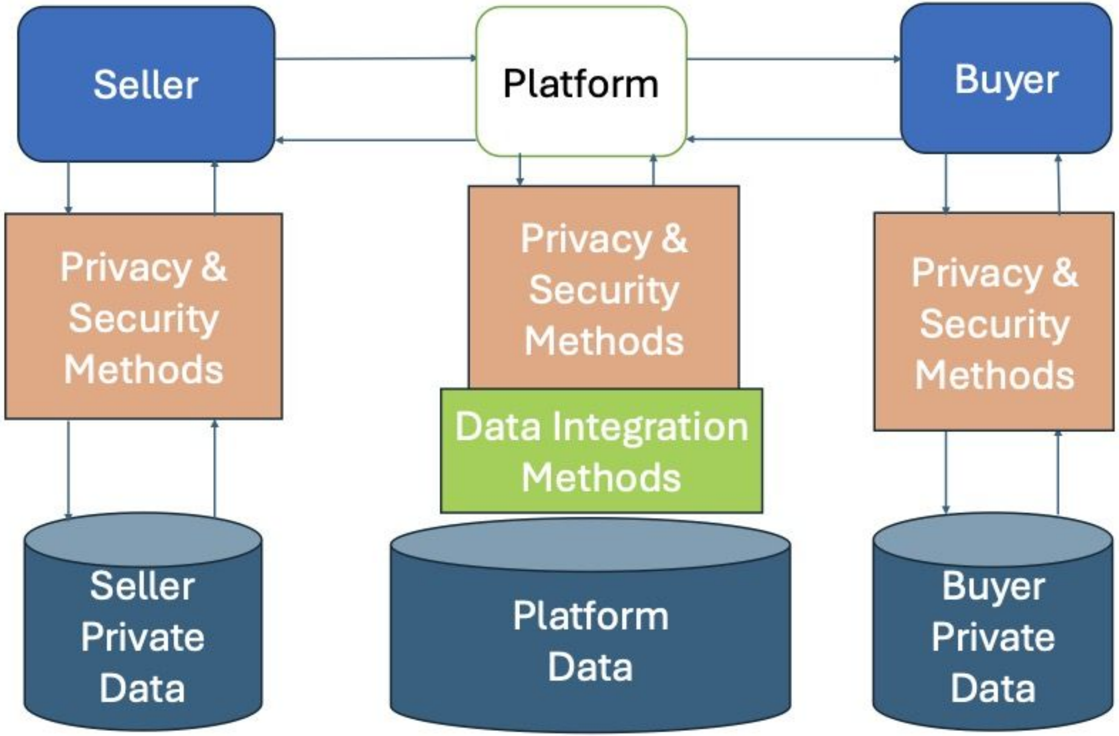
Privacy Challenges Are Everywhere!



Part 2: Privacy and Security Risks

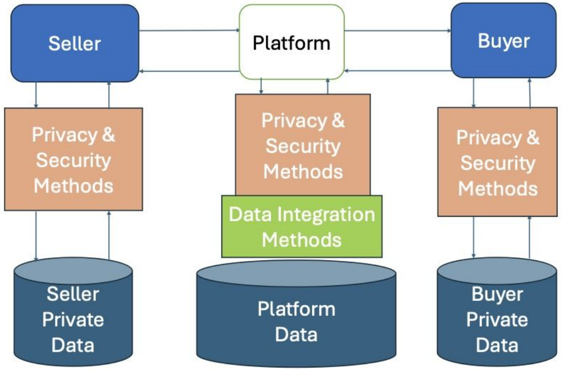
https://dacesresearch.org/tutorials/sigmod2025/

Protect Information in Data Markets



Protect Information in Data Markets

1. Protect buyers from *malicious* sellers
2. Protect sellers from *malicious* buyers
3. Prevent *unauthorized* users from accessing:
    a. Seller private data
    b. Buyer private data
    c. Platform private data
4. Prevent manipulation of data acquisition mechanisms:
    a. Data discovery
    b. Data valuation
    c. Data negotiation
    d. Data delivery



Privacy and Security Attacks

- Naively allowing query access to data markets is risky for users/orgs
  - Linkage attacks
  - Reconstruction attacks
  - Inference attacks
  - Plaintext/ciphertext attacks
- Naive designs of data markets is risky for valuation
  - Manipulation of pricing and negotiation mechanisms
  - Less trust in data markets

Motivates the need for robust *privacy and security protections*

Privacy and Security Attacks

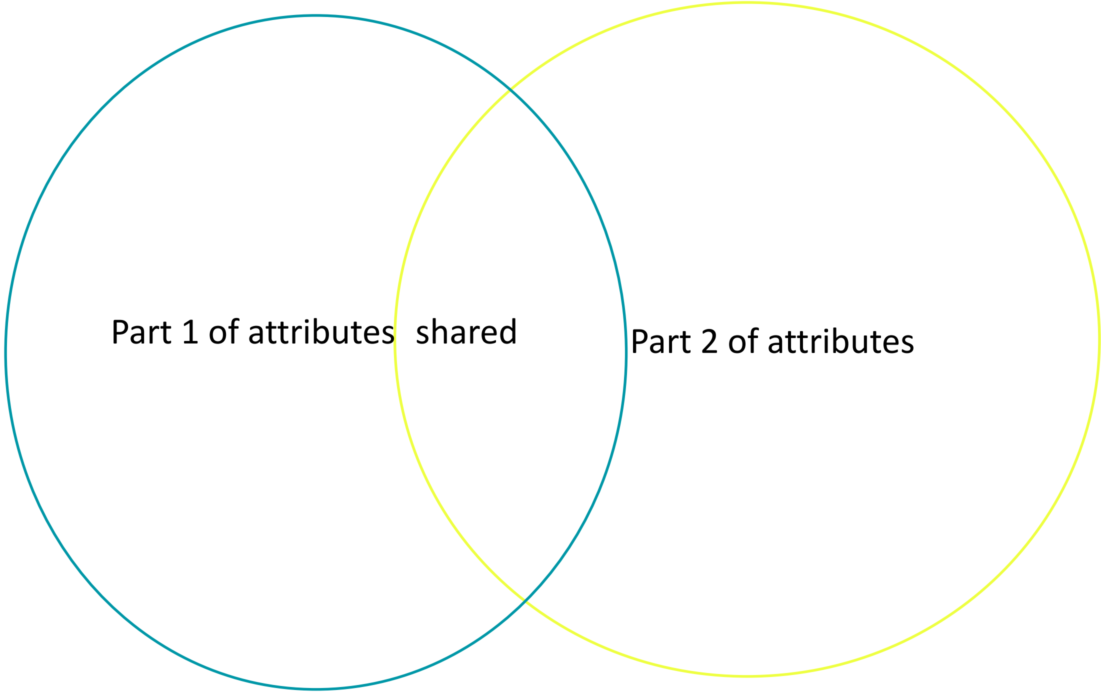
- Naively allowing query access to data markets is risky for users/orgs
  - **Linkage attacks**
  - Reconstruction attacks
  - Inference attacks
  - Plaintext/ciphertext attacks
- Naive designs of data markets is risky for valuation
  - Manipulation of pricing and negotiation mechanisms
  - Less trust in data markets

Motivates the need for robust *privacy and security protections*

Linkage Attacks

Perform join on one or more datasets

Can uniquely identify individuals



De-identification attempt

“Anonymize the Data”: Are we happy with this solution? Why or why not?

Name	Sex	Blood	…	HIV?
James	M	B	…	N
Peter	M	O	…	Y
…	…	…	…	…
Paul	M	A	…	N
Eve	F	B	…	Y

Name	Sex	Blood	…	HIV?
XXXXX	M	B	…	N
XXXXX	M	O	…	Y
…	…	…	…	…
XXXXX	M	A	…	N
XXXXX	F	B	…	Y

De-identification attempt

“Anonymize the Data”: Not sufficient because of linkage attacks!

87% of US population (used to) have unique date of birth, gender, and postal code!

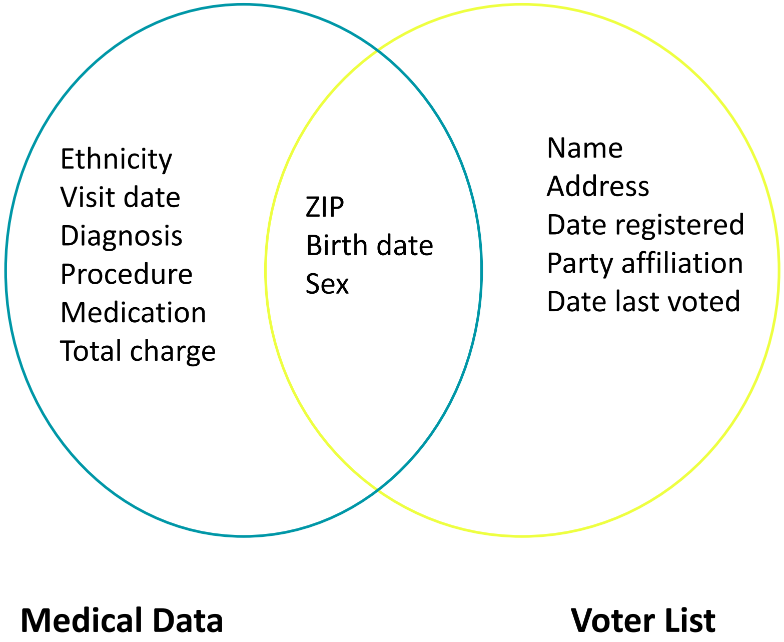
[Golle and Partridge ‘09]

De-identification attempt

"Anonymize the Data": Reidentification via Linkage

Can uniquely identify > 60% of the U.S. population [Sweeney '00, Golle '06, Sweeney '97]

Name	Sex	Blood	…	HIV?
XXXXX	M	B	…	N
XXXXX	M	O	…	Y
…	…	…	…	…
XXXXX	M	A	…	N
XXXXX	F	B	…	Y



Privacy and Security Attacks

- Naively allowing query access to data markets is risky for users/orgs
    - **Linkage attacks**
    - **Reconstruction attacks**
    - Inference attacks
    - Plaintext/ciphertext attacks
- Naive designs of data markets is risky for valuation
    - Manipulation of pricing and negotiation mechanisms
    - Less trust in data markets

Motivates the need for robust *privacy and security protections*

Reconstruction Attack

<u>Reconstruction attack</u>: If we have dataset $x \in \{0,1\}^n$ and person i has sensitive bit x_i and attacker/adversary gets $q_S(x) = \sum_{i \in S} x_i$ for $O(n)$ random $S \subseteq [n]$.

Reconstruction Attack

Reconstruction attack: If we have dataset $x \in \{0,1\}^n$ and person i has sensitive bit x_i and attacker/adversary gets $q_S(x) = \sum_{i \in S} x_i$ for $O(n)$ random $S \subseteq [n]$.

[Dinur-Nissim '03]: With high probability, adversary can reconstruct 0.99 fraction of the dataset $x \in \{0,1\}^n$ if noise added to each query is less than $o(\sqrt{n})$.

Privacy and Security Attacks

- Naively allowing query access to data markets is risky for users/orgs
  - **Linkage attacks**
  - **Reconstruction attacks**
  - **Inference attacks**
  - Plaintext/ciphertext attacks
- Naive designs of data markets is risky for valuation
  - Manipulation of pricing and negotiation mechanisms
  - Less trust in data markets

Motivates the need for robust *privacy and security protections*

Inference Attacks

Inference attack: Attacker gets $\tilde{O}(n^2)$ count queries with noise $o(n)$ and needs to know if someone is in the dataset or not.

Inference Attacks

Public Access to Genome-Wide Data: Five Views on Balancing Research with Privacy and Protection

P3G Consortium, George Church, Catherine Heeney, Naomi Hawkins, Jantina de Vries, Paula Boddington, Jane Kaye, Martin Bobrow, Bruce Weir

Just over twelve months ago, *PLoS Genetics* published a paper [1] demonstrating that, given genome-wide genotype data from an individual, it is, in principle, possible to ascertain whether that individual is a member of a larger group defined solely by aggregate genotype frequencies, such as a forensic sample or a cohort of participants in a genome-wide association study (GWAS). As a consequence, the National Institutes of Health (NIH) and Wellcome Trust agreed to shut down public access not just to individual genotype data but even to aggregate genotype frequency data from each study published using their funding. Reactions to this decision span the full breadth of opinion, from "too little, too late—the public trust has been breached" to "a heavy-handed bureaucratic response to a practically minimal risk that will unnecessarily inhibit scientific research." Scientific concerns have also been raised over the conditions under which individual identity can truly be accurately determined from GWAS data. These concerns are addressed in two papers published in this month's issue of *PLoS Genetics* [2],[3]. We received several submissions on this topic and decided to assemble these viewpoints as a contribution to the debate and ask readers to contribute their thoughts through the PLoS online commentary features.

Privacy and Security Attacks

- Naively allowing query access to data markets is risky for users/orgs
  - **Linkage attacks**
  - **Reconstruction attacks**
  - **Inference attacks**
  - **Plaintext/ciphertext attacks**
- Naive designs of data markets is risky for valuation
  - Manipulation of pricing and negotiation mechanisms
  - Less trust in data markets

Motivates the need for robust *privacy and security protections*

Plaintext/Ciphertext Attacks

A datamarket could encrypt the interaction between buyers/sellers/platforms

Plaintext/Ciphertext Attacks

A datamarket could encrypt the interaction between buyers/sellers/platforms. The encryption scheme should be secure against one or more threat models:

Ciphertext-only attack

Known-plaintext attack

Chosen-plaintext attack

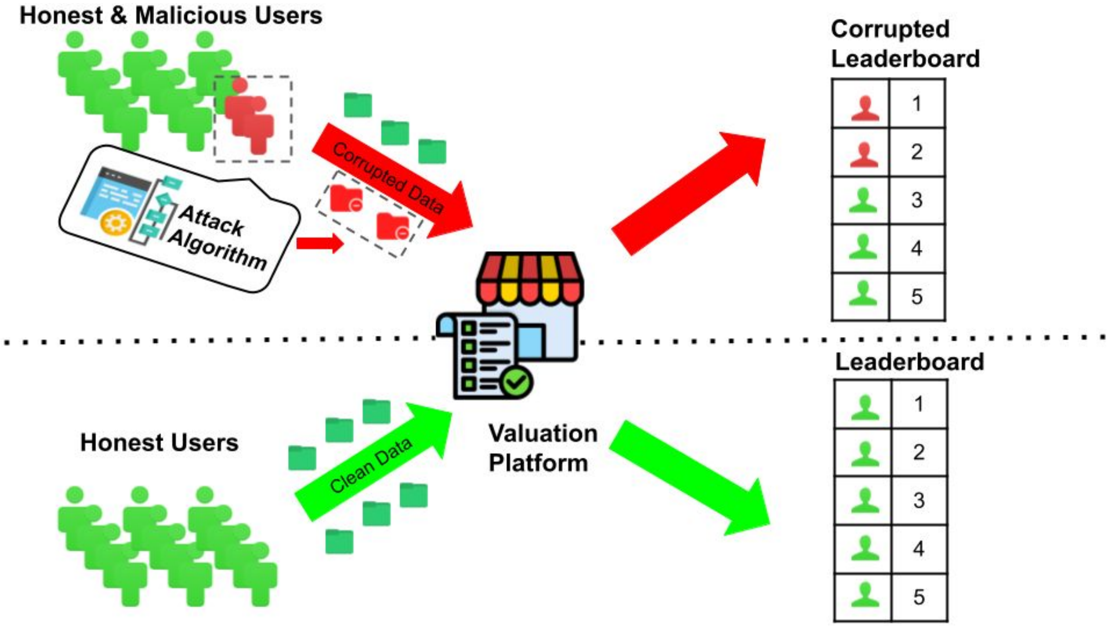
Chosen-ciphertext attack

Privacy and Security Attacks

- Naively allowing query access to data markets is risky for users/orgs
  - Linkage attacks
  - Reconstruction attacks
  - Inference attacks
  - Plaintext/ciphertext attacks
- Naive designs of data markets is risky for valuation
  - **Manipulation of pricing and negotiation mechanisms**
  - Less trust in data markets

Motivates the need for robust *privacy and security protections*

Valuation Attacks



Valuation Attacks

MemAttack: Efficiently Attacking Memorization Scores

by Do, Chandrasekaran, Alabi (2025)

Influence estimation tools—such as memorization scores—are widely used to understand model behavior, attribute training data, and inform dataset curation. However, recent applications in data valuation and responsible machine learning raise the question:

Can these scores themselves be adversarially manipulated?

In this work, we present a systematic study of the feasibility of attacking memorization-based influence estimators. We propose efficient mechanisms that allow an adversary to perturb specific training samples or small subsets of data to inflate or suppress their corresponding influence scores, all while maintaining high utility on natural downstream tasks. Our attacks are practical, requiring only black-box access to model outputs and incur moderate computational overhead. We empirically validate our methods on MNIST, SVHN, and CIFAR-10, showing that even state-of-the-art estimators are vulnerable to targeted score manipulations. In addition, we provide a theoretical analysis of the stability of memorization scores under adversarial perturbations, revealing conditions under which influence estimates are inherently fragile. Our findings highlight critical vulnerabilities in influence-based attribution and suggest the need for robust defenses.

Valuation Attacks

In large datasets, a small subset of highly influential (memorized) training examples disproportionately affects the model's predictions and generalization capabilities, while the majority of examples have little to no impact. Influence scores quantify how much each datapoint affects the model's predictions.

Influence scores can be used to price data.

- Tom Yan and Ariel D Procaccia. **If you like shapley then you'll love the core.** *AAAI 2021*
- Tianshu Song, Yongxin Tong, and Shuyue Wei. **Profit allocation for federated learning**. *In 2019 IEEE International Conference on Big Data (Big Data), pages 2577–2586. IEEE, 2019.*
- Jiachen T Wang and Ruoxi Jia. **Data banzhaf: A robust data valuation framework for machine learning**. *AISTATS 2023*.
- Tianhao Wang, Johannes Rausch, Ce Zhang, Ruoxi Jia, and Dawn Song. **A principled approach to data valuation for federated learning**. *Federated Learning: Privacy and Incentive, pages 153–167, 2020.*

Valuation Attacks

Memorization Score

$$\mathtt{mem}(\mathcal{A}, \mathbf{z}, q(\mathbf{z})) := \Pr_{(x,y)\leftarrow q(\mathbf{z}),h\leftarrow\mathcal{A}(\mathbf{z}\cup q(\mathbf{z}))}[h(x) = y] - \Pr_{(x,y)\leftarrow q(\mathbf{z}),h\leftarrow\mathcal{A}(\mathbf{z})}[h(x) = y]$$

Quantifies how much a new example would change the performance of a classifier.

Valuation Attacks via Memorization Scores

1) **Out-of-Distribution (OOD) Replacement Attack.**
2) **Pseudoinverse Attack (PINV)**
3) **EMD Attack**: Maximize Wasserstein distance between original and perturbed data points
4) **DeepFool (DF) Perturbation Attack**: Sample points along decision boundary

Valuation Attacks: Experimental Results

{Loss Curvature, Confidence Event, and Privacy Score} are proxies for the memorization scores.

We evaluate on MNIST, SVHN, CIFAR-10 datasets.

Higher scores correspond to more memorization from the attack data points.

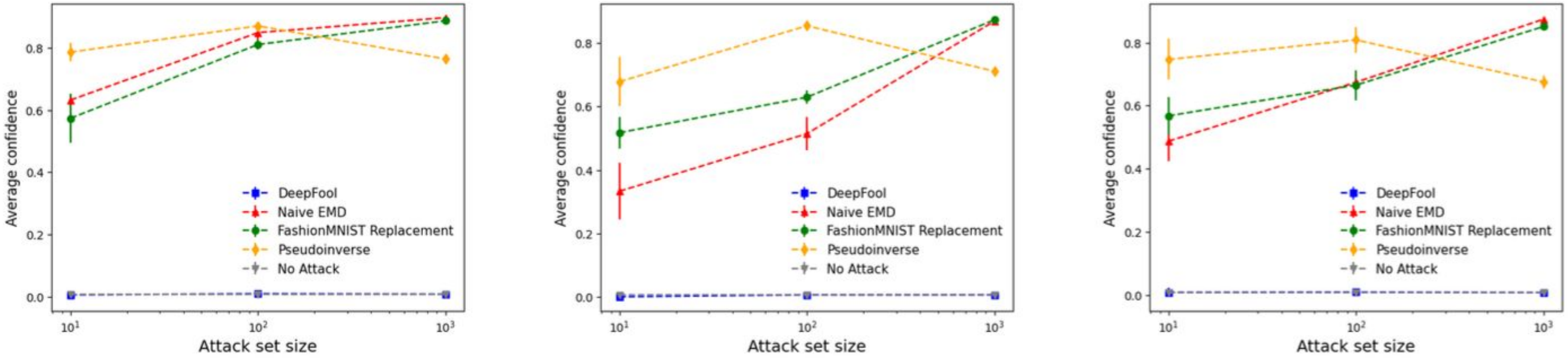
Attack	Loss Curvature			Confidence Event			Privacy Score		
	MNIST	SVHN	CIFAR-10	MNIST	SVHN	CIFAR-10	MNIST	SVHN	CIFAR-10
None	0.00±0.00	0.01±0.00	0.09±0.00	0.01±0.00	0.06±0.00	0.23±0.00	0.47±0.00	0.49±0.00	0.19±0.00
OOD	0.13±0.00	0.02±0.00	**0.14±0.00**	0.62±0.00	0.52±0.00	0.61±0.00	0.08±0.00	-0.10±0.00	0.09±0.00
PINV	**0.14±.00**	**0.14±0.00**	0.08±0.00	**0.85±0.00**	**0.81±0.00**	**0.66±0.00**	**0.29±0.01**	**0.51±0.00**	**0.79±0.00**
EMD	0.06±0.00	0.00±0.00	-0.05±0.00	0.51±0.00	0.68±0.00	0.54±0.00	-0.03±0.00	-0.03±0.00	0.01±0.00
DF	0.00±0.00	0.00±0.00	0.01±0.00	0.00±0.00	0.00±0.00	0.00±0.00	-0.02±0.00	-0.01±0.00	-0.02±0.00

Valuation Attacks: Experimental Results

{Loss Curvature, Confidence Event, and Privacy Score} are proxies for the memorization scores.

We evaluate on (standard) deep neural network architectures: VGG, ResNet, MobileNet.

Higher scores correspond to more memorization from the attack data points.



(a) VGG-11 (b) ResNet-18 (c) MobileNet-v2

Conclusion: Privacy and Security Attacks

- Naively allowing query access to data markets is risky for users/orgs
  - Reconstruction attacks
  - Inference attacks
  - Plaintext/ciphertext attacks
- Naive designs of data markets leads is risky for valuation
  - Manipulation of pricing and negotiation mechanisms
  - Less trust in data markets

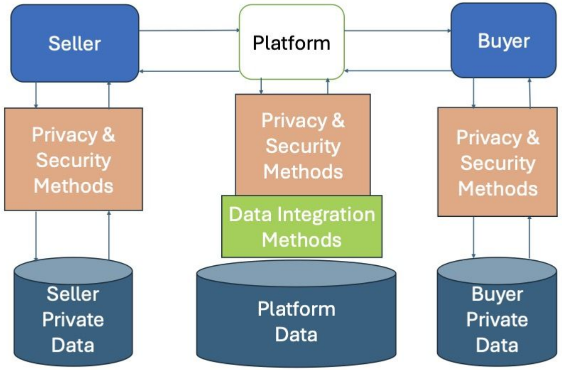
Need to provide robust *privacy and security protections*

Conclusion: Privacy and Security Attacks

Need to provide robust *privacy and security protections* via security definitions:

1. **Security guarantee**: what is the scheme/protocol in the data market intended to prevent the attacker from doing?
2. **Threat model**: what is the power of the adversary in the data market? What actions can the attacker perform?

Protect Information in Data Markets



1. Protect buyers from *malicious* sellers
2. Protect sellers from *malicious* buyers
3. Prevent *unauthorized* users from accessing:
   a. Seller private data
   b. Buyer private data
   c. Platform private data
4. Prevent manipulation of data acquisition mechanisms:
   a. Data discovery
   b. Data valuation
   c. Data negotiation
   d. Data delivery

Next: How do we *protect* the information?

Part 3:
Privacy-Preserving Technologies and Security Tools

The Spectrum of Data Marketplace Architectures

Governance/Storage Categories	Centralized Data Storage (Data is Pooled)	Distributed Data Storage (Data Stays Sovereign)
Centralized Governance (Single, Trusted Arbiter)	The Traditional Hub ● Classic data warehouse model ● High trust in one operator required	The Federated Orchestrator ● Data is federated, not pooled ● A central company still manages rules & accesses
Decentralized Governance (Automated, "Smart" Arbiter)	The Governed Pool ● Data is pooled, but governed by code/community ● A niche but emerging approach	The Sovereign Exchange ● Data Sovereignty by Design ● Transaction Integrity via Arbiter

Trading Insights via Gradients in Distributed Marketplace

The Core Idea:

Instead of trading the data itself, participants trade the "insight" the data provides to a machine learning model.

How is this insight captured?

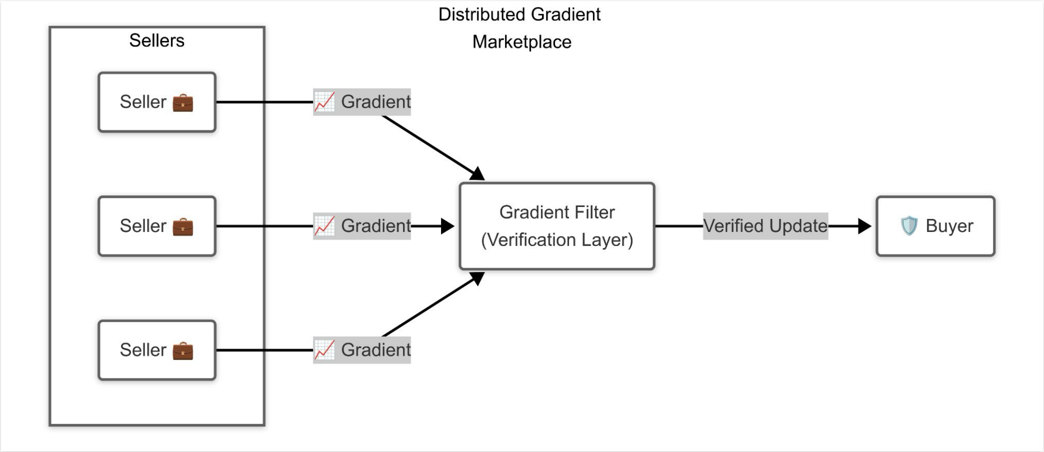
Through a Gradient.

The Implication for Valuation:

In this model, the transaction and valuation are no longer about the raw data. They are now fundamentally tied to the quality and utility of the gradient itself.

How a Gradient Marketplace Works

- A Buyer wants to train or improve their ML model.
- Multiple Sellers use their private data to compute gradients for the buyer's model.
- The Buyer purchases these gradients and uses them to update their model.



The Problem – A Wall Between Buyers and Data

THE DATA BUYER

Needs to accurately assess data quality and value before committing resources.

KEY BARRIERS

🛡️ **Privacy & Security**: Prevents the exposure of sensitive user data and Personally Identifiable Information.

🔒 **Data Ownership & IP**: Protects the data as the seller's core asset and valuable intellectual property.

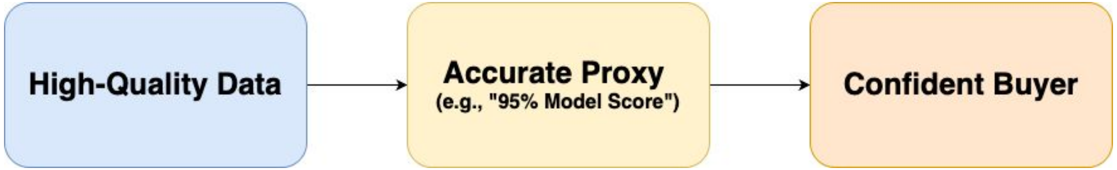
☁️ **Scale & Efficiency**: Makes the full transfer and inspection of massive datasets logistically impractical.

THE RAW DATASET

The source of truth remains unseen, its quality unverified.

The Flawed Solution: Valuing by a (Gameable) Proxy

The Theory: A Proxy Is an Honest Signal of Quality



The Reality: A Proxy Can Be Manipulated

The Central Vulnerability:

A proxy can be manipulated independently of the data's actual quality. This makes the entire valuation process gameable.

Valuation in Distributed Setups

Frameworks like DAVED* solve this by performing valuation on embeddings (i.e., "fingerprints") instead of raw data.

The Goal: This allows for good data selection in a distributed setup, preserving privacy while assessing quality.

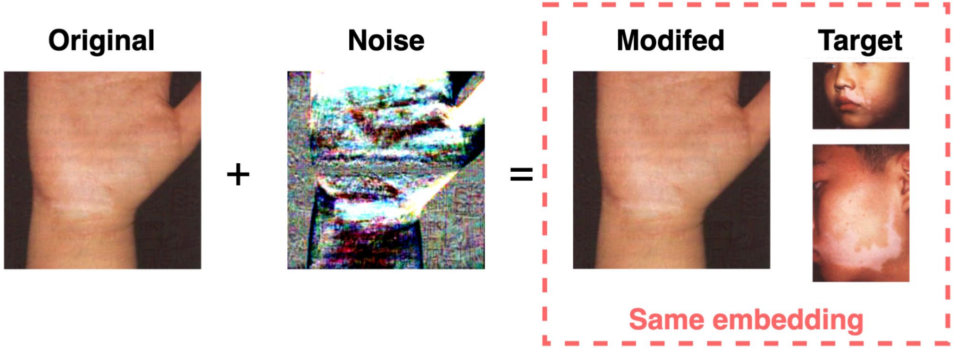
*Lu et al., "DAVED: Data Acquisition via Experimental Design for Data Markets," NeurIPS 2024.

How to Fool an Embedding-Based Valuation

- A buyer wants data with embeddings similar to a *Target Image*.
- The seller adds calculated, imperceptible noise to their own *Irrelevant Image*.
- The new "noisy" image now has an embedding nearly identical to the *Target Image*.
- The Result: The valuation is fooled. The buyer's system approves the purchase, but they receive a dataset of useless, manipulated images.

The Vulnerability: Embeddings Can Be Manipulated

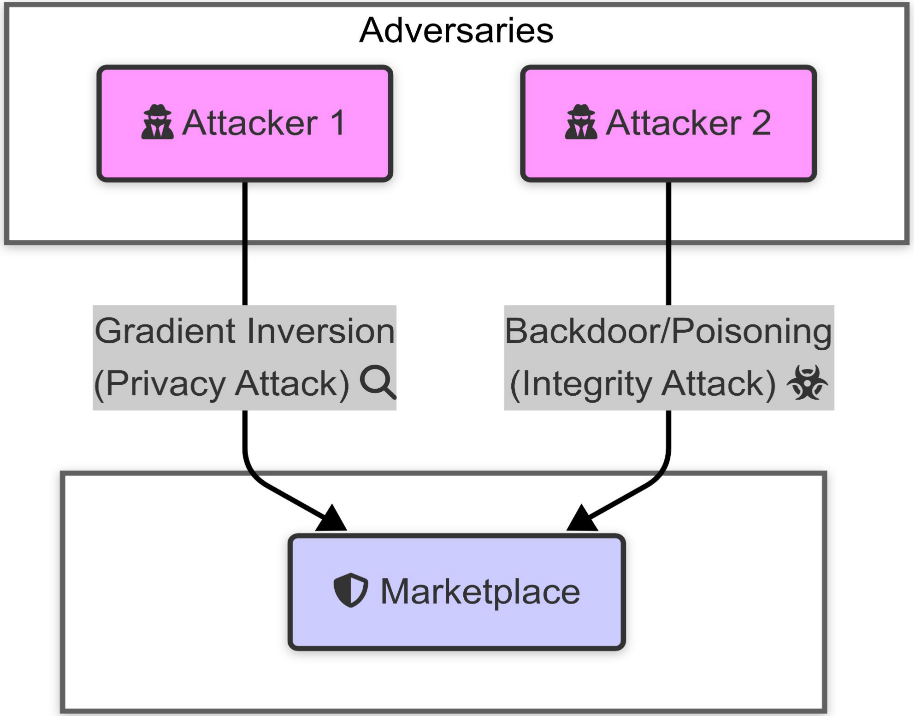
Context: A buyer is searching a dermatology dataset (like Fitzpatrick17K) for high-value images of a specific skin condition to train their ML model.

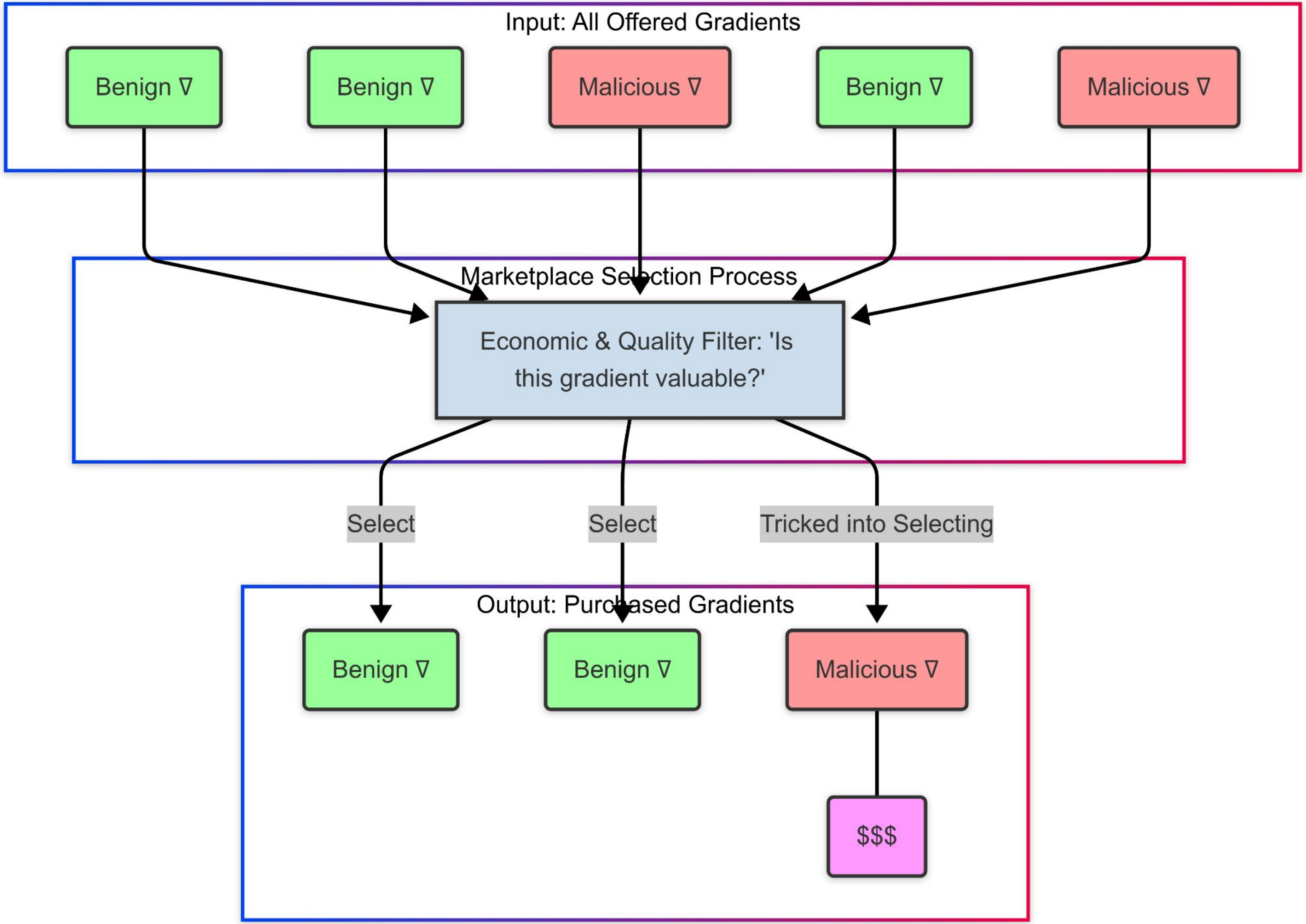


- Original: A irrelevant Original image.
- Noise: A layer of calculated, human-imperceptible Noise is added.
- Modified: The resulting Modified image looks identical to our eyes, but its "fingerprint" —its embedding—is now the same as the high-value Target image.

Security and Privacy Challenges: A Marketplace Under Siege

Threat Category	The Adversary's Goal
Privacy Attacks	To reconstruct sensitive, private training data from the shared gradient.
Integrity Attacks	To corrupt the model's performance or install a hidden, malicious trigger.





Input: All Offered Gradients
Benign ∇
Benign ∇
Malicious ∇
Benign ∇
Malicious ∇
Marketplace Selection Process
Economic & Quality Filter: 'Is this gradient valuable?'
Select
Select
Tricked into Selecting
Output: Purchased Gradients
Benign ∇
Benign ∇
Malicious ∇
$$$

Gradient Valuation

Instead of trading raw data or its indirect proxies, a Gradient Marketplace creates a more secure system by trading the direct output of machine learning: the model gradients themselves.

How It Works: Frameworks like martFL* provide the architecture for this model:

Direct Inspection: A selection filter is used to directly inspect the quality and utility of each incoming model update (gradient). It rejects contributions that are malicious or low-quality.

"What You See Is What You Get": The buyer receives the exact same gradient that was just evaluated by the filter.

*Li et al., "martFL: Enabling Utility-Driven Data Marketplace with a Robust and Verifiable Federated Learning Architecture," ACM Computer and Communications Security Conference 2023.

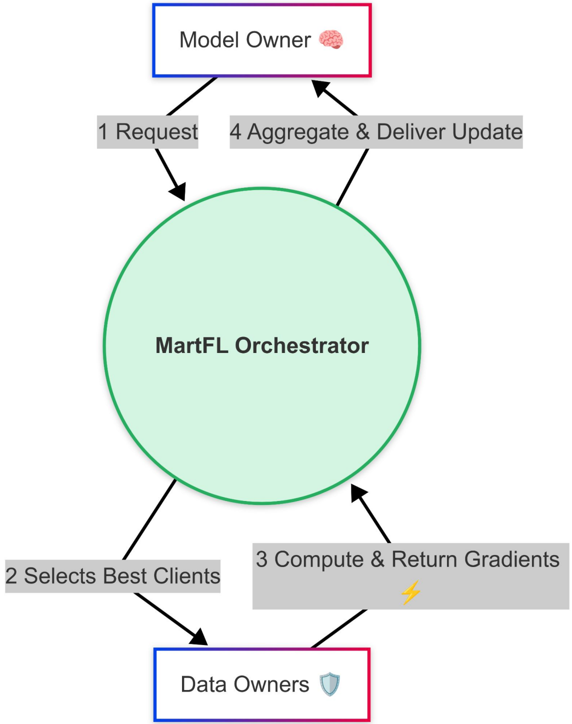
The MartFL Life Cycle

Request: A model owner submits a training task to the marketplace.

Select: The MartFL Orchestrator uses its two-phase filter to select the most valuable data owners.

Train: The selected owners compute gradients on their private data.

Improve: MartFL aggregates the gradients, ensures fair payment, and delivers a single, powerful update to the model owner.



The New Threat: Malicious Gradient Attacks

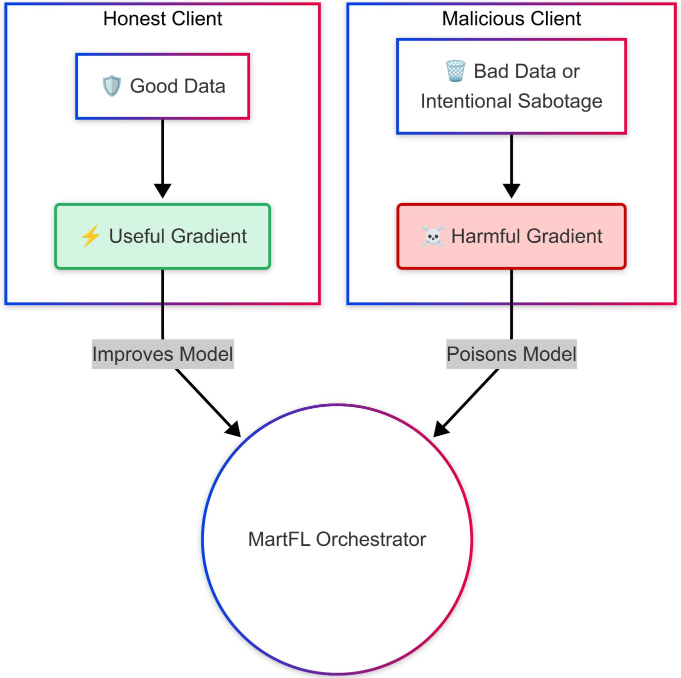
In a gradient marketplace, the threat shifts from faking value to actively sabotaging the training process.

A Malicious Client's Goal:

- Get paid for contributing nothing of value.
- Poison the global model, reducing its accuracy for their own benefit.

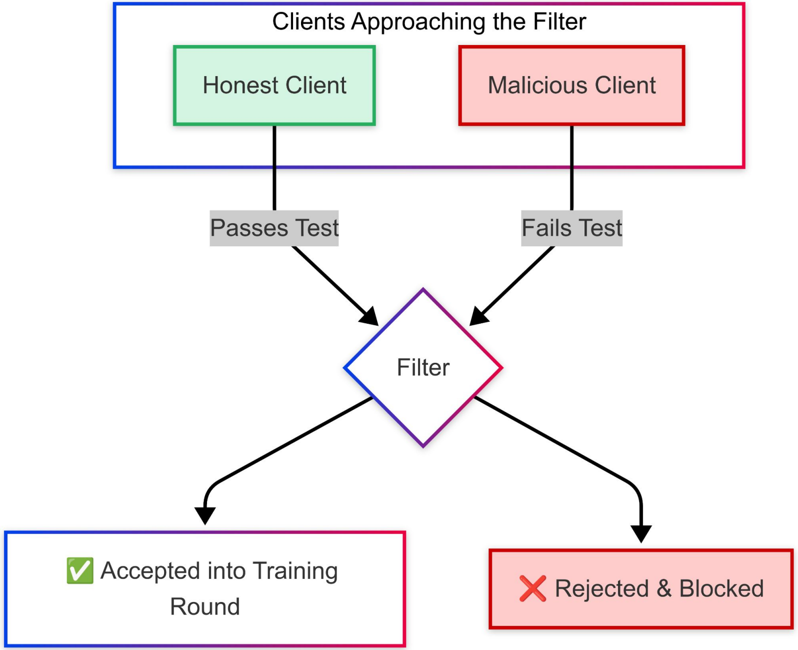
The Method:

Submit useless or deliberately harmful gradients instead of honest ones.



How MartFL Filters Malicious Gradients

- **Create a Baseline**: A "trusted baseline" is established using the buyer's clean reference gradient combined with past contributions from high-quality sellers.
- **Measure Similarity**: The system calculates the cosine similarity between each new gradient and the trusted baseline.
- **Filter Outliers**: Any gradient with low similarity is flagged as an outlier and rejected, filtering out potentially malicious or useless updates.

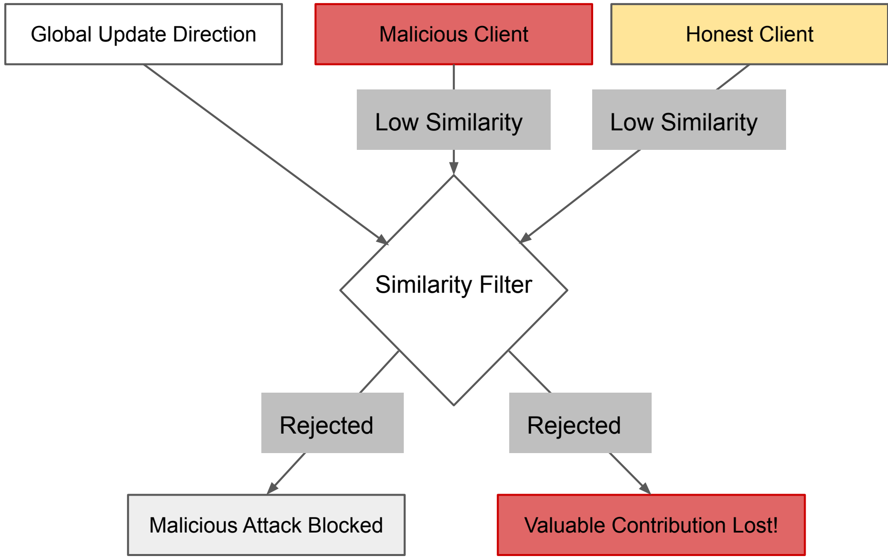


The Potential Flaw: Can Similarity Be Fooled?

This leads to two key research questions for our analysis:

Robustness: How well does similarity filtering actually detect various malicious attacks?

Fairness: Does this filtering mechanism unfairly penalize honest clients who hold valuable, non-mainstream (outlier) data?



The Blind Spot in Gradient Marketplace Evaluation

A system can be **technically** perfect and secure, but **economically** broken.

To build a truly successful marketplace, we must evaluate the entire ecosystem. Our framework integrates three crucial marketplace-centric metrics.

Model Robustness

The Question: Is the final trained model accurate, reliable, and secure against attacks?

Economic Viability

The Question: Is the marketplace economically efficient? Does the performance gain justify the cost?

Marketplace Stability

The Question: Is the system fair and stable for its participants?

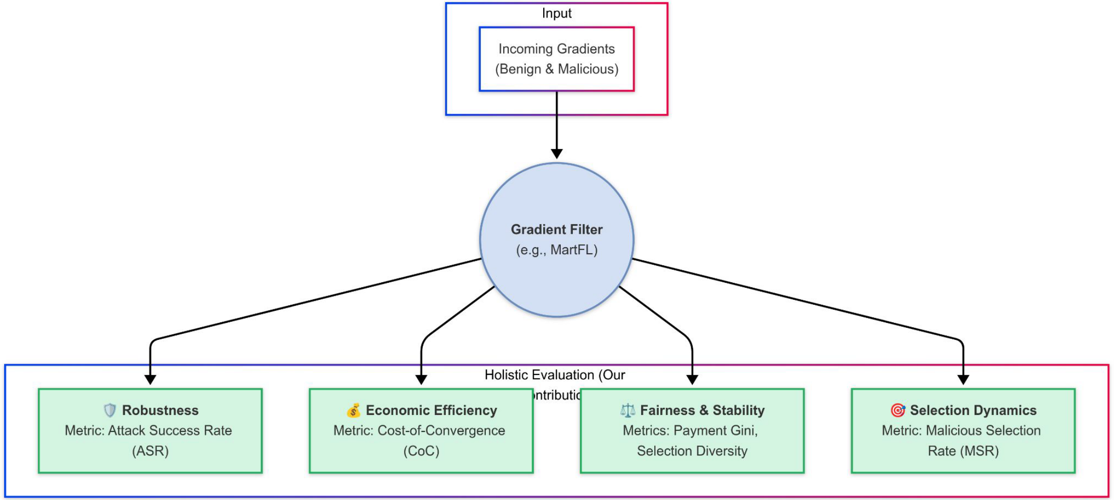
Key Evaluation Dimensions

Robustness: Does the filter stop the attack?

Economic Efficiency: What is the true cost for the buyer to achieve their goal?

Fairness & Stability: Are honest sellers treated fairly, or are they penalized?

Selection Dynamics: Who is the filter actually selecting, and how often is it fooled?



Case Study: Can the Marketplace Survive a Backdoor Attack?

The Attacker's Playbook

Step 1: Poison the Source Data
The attacker adds a trigger (e.g., a white square) to a "cat" image and maliciously labels it as a "dog."

Step 2: Submit the Malicious Gradient
The attacker offers the harmful gradient, learned from this poisoned data, for sale in the marketplace.

Our Analysis

Analysis of Step 1: Security Effectiveness

Our framework asks:
✅ Can the selection filter detect the malicious gradient's signature?
✅ Can it prevent the poison from compromising the global model?

Analysis of Step 2: Economic & Fairness Impact

Our framework asks:
⚖️ What is the collateral damage? Are benign, honest sellers unfairly penalized or rejected by stricter filtering triggered by the attack?

Two Attack Scenarios

Attack 1: The Standard Backdoor

The Strategy: Brute Force. The adversary submits a standard poisoned gradient and simply *hopes* it bypasses the marketplace filter.

Attack 2: The Sybil "Mimicry" Backdoor

The Strategy: Deception & Camouflage. This is a more intelligent, two-step attack:

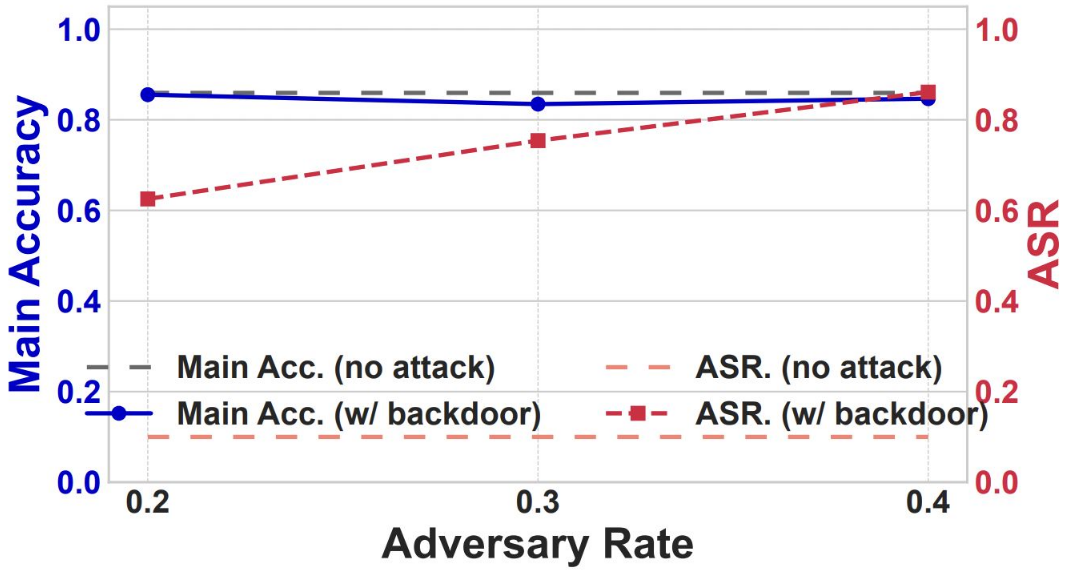
a) Learn What Passes: The adversary identifies the characteristics of *benign* gradients that are successfully selected.

b) Blend the Attack: They **combine** their malicious backdoor gradient with this benign "camouflage" to create a new gradient that looks trustworthy.

Result - Final model performance

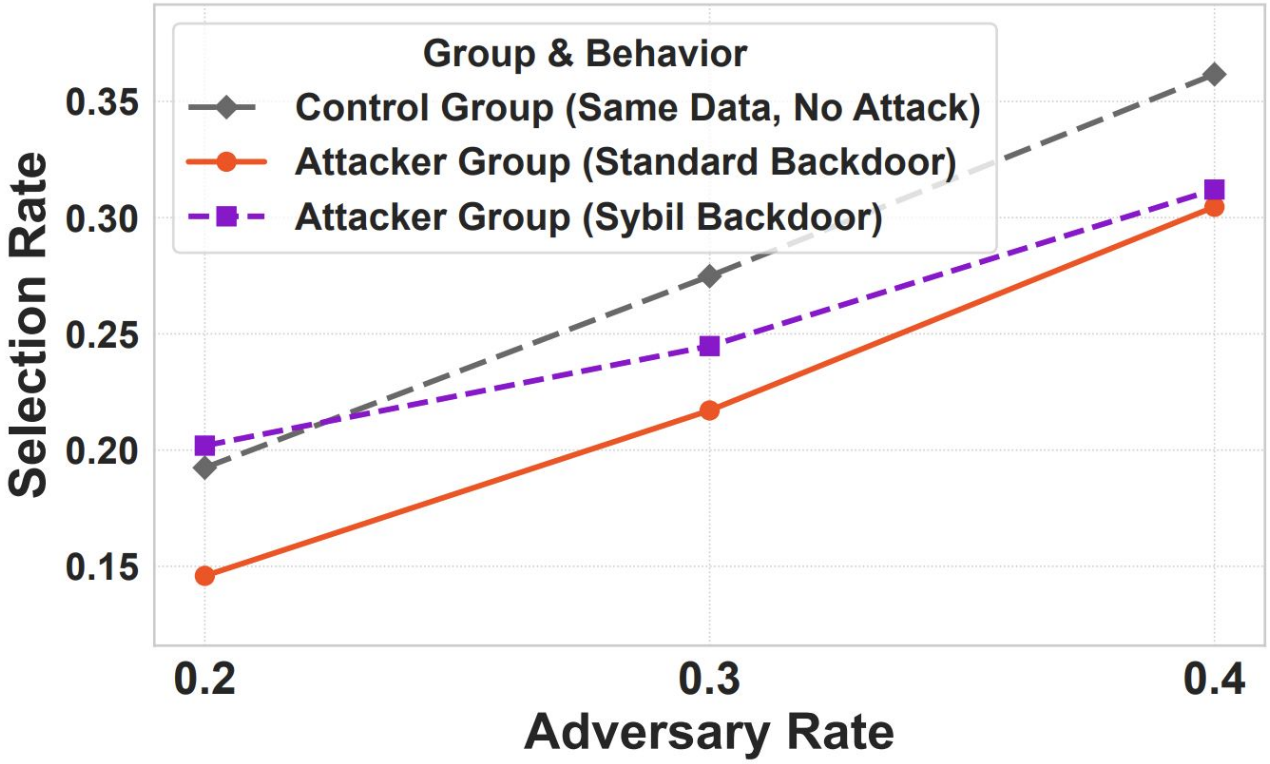
The Illusion (Blue Line): The model's performance on its main task remains high and stable, suggesting the system is healthy.

The Reality (Red Line): Simultaneously, the Attack Success Rate skyrockets, proving the model is being successfully poisoned with a hidden backdoor.



Mechanism of Failure: Why the Filter Was Fooled

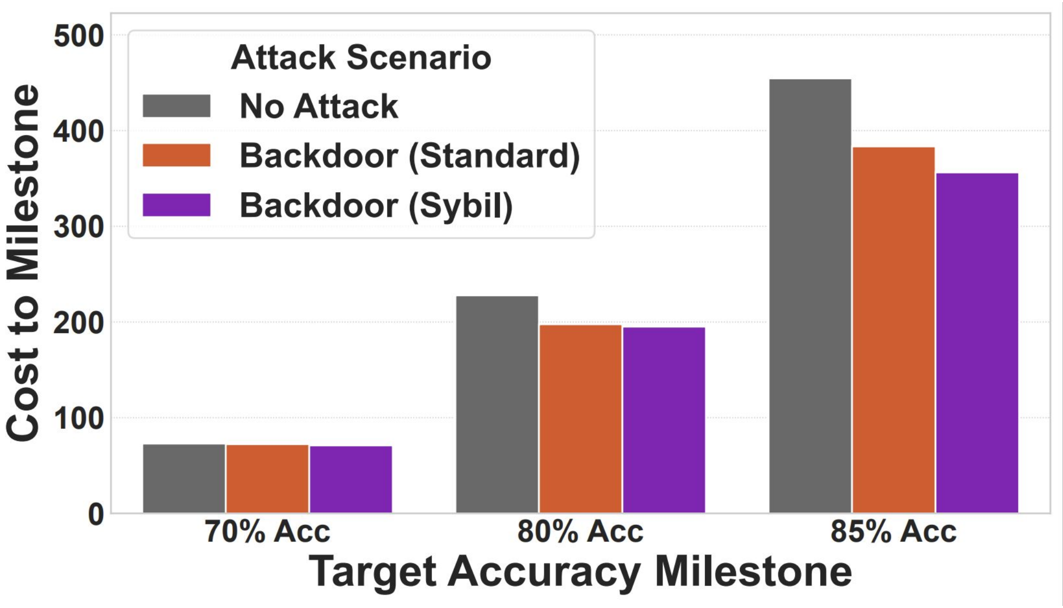
Despite the filtering mechanism, a sufficient volume of malicious updates evaded detection, enabling the backdoor attack to succeed.



"Deceptive Efficiency" of an Attacked Market

- The Sybil-attacked market reaches the target accuracy with 23% less cost (fewer gradients purchased).

- From a purely economic standpoint, the attacked market looks more efficient. A buyer optimizing solely for cost would inadvertently prefer the compromised environment. This makes the attack even harder to detect through economic signals.

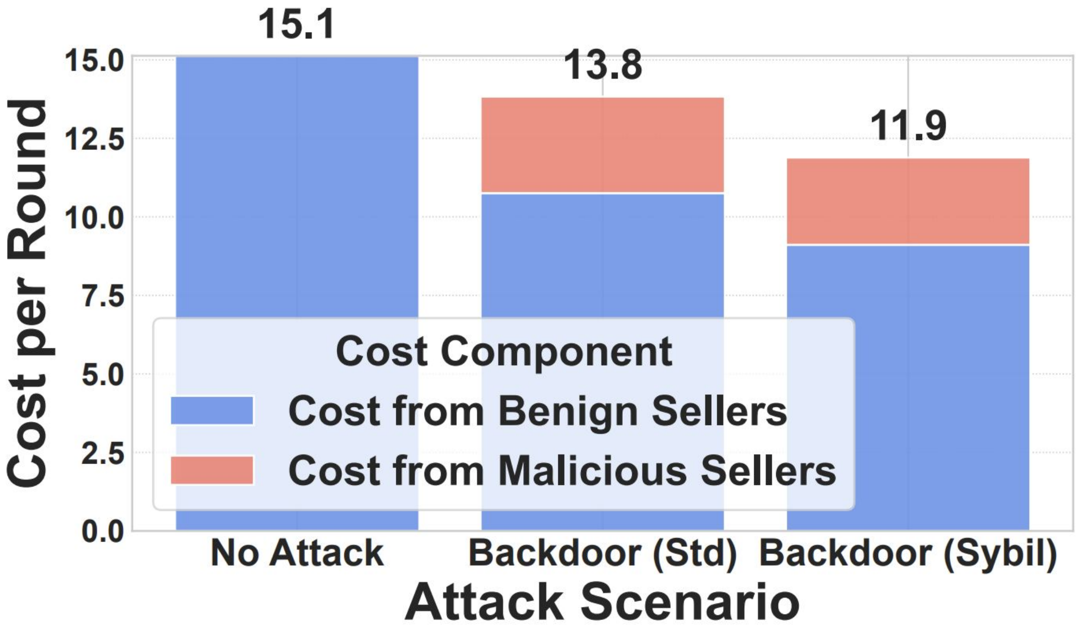


Economic Fallout: Who Really Pays the Price?

No Attack: Honest sellers earn 100% of the revenue.

Sybil Attack: Honest seller revenue plummets by 40%. Attackers successfully extract nearly a quarter of all payments.

The market isn't more efficient. Attackers are simply crowding out and defunding honest contributors, creating an unsustainable economy.



Data Discovery Dilemma: Diversity vs. Security

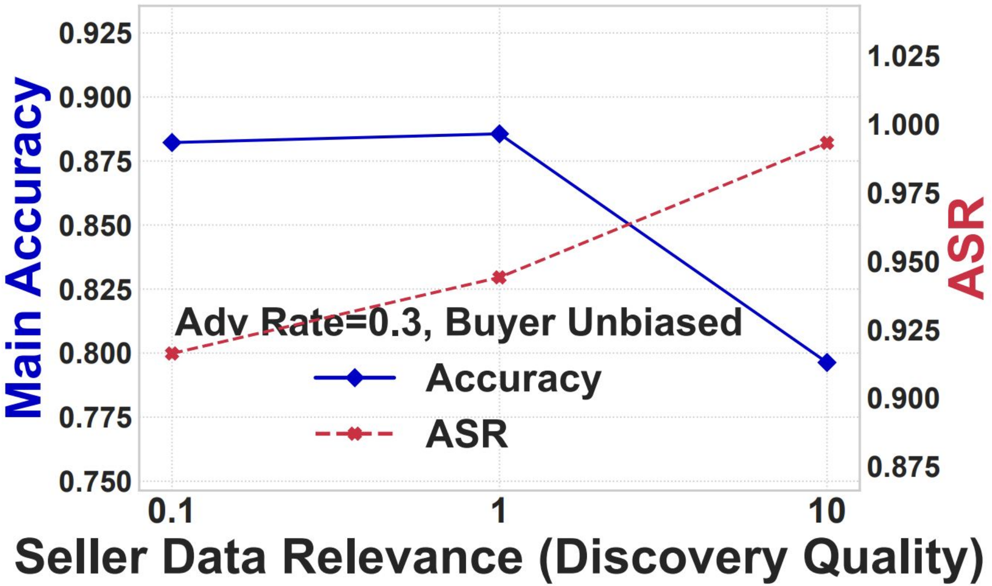
1. The Marketplace Dilemma

A realistic marketplace needs data diversity. However, this creates a fundamental conflict for similarity-based security filters.

2. The Mechanism of Failure

With Homogeneous Data: The filter works. Malicious gradients are easy-to-spot outliers. ✅

With Heterogeneous Data: The filter collapses. It cannot distinguish between "benign diversity" and "malicious intent". ❌



Key Takeaways

Finding	Implication
1. The Attack Surface Has Shifted.	The primary vulnerability is now the marketplace's **economic and selection mechanisms.**
2. Standard Metrics Are Deceptive.	High accuracy and low cost can **mask catastrophic security failures** and unfair outcomes.
3. Similarity-Based Defenses Are Brittle.	They are fundamentally vulnerable to mimicry and **fail in diverse, realistic environments.**

Saibot:
Differentially Private
Task-based Search

(back to centralized search)
Huang et al. VLDB 23

Task-based Search: Basic Algorithm

```
D = initial training dataset
for next augmentation α
    if eval(apply A to D) is best so far
        Keep α in A
return best A
```

Greedy

Use sketches

But can we enforce differential privacy?

Differential Privacy

Privatization: Differential Privacy(DP) Algorithm

$$\Pr[M(D) \in S] \leq \exp(\boxed{\epsilon}) * \Pr[M(D') \in S] \boxed{+} \delta$$

Informally, an algorithm satisfies DP if no single record can be inferred

- Hides individuals in a dataset by adding noise to results
- Each query consumes part of a dataset’s finite budget
- Consumed budget ∝ noise added to result

Differential Privacy: Privacy Budget

Privacy budget: (ϵ, δ)

SELECT SUM(Y) FROM D

D

A	Y
a_1	1
a_1	2

Remaining budget:

(ϵ, δ)

Differential Privacy: Privacy Budget

Privacy budget: (ϵ, δ)

SELECT SUM(Y) FROM D

$1 + 2 + \mathit{noise}_{\epsilon,\,\delta} = \boxed{4.2}$

D

A	Y
a_1	1
a_1	2

Remaining budget:

(0, 0)

Differential Privacy: Privacy Budget

Privacy budget: (ϵ, δ)

SELECT SUM(Y*Y) FROM D

D

A	Y
a_1	1
a_1	2

Remaining budget:

(0, 0)

Differential Privacy: Privacy Budget

Privacy budget: (ϵ, δ)

SELECT SUM(Y*Y) FROM D

No privacy budget, cannot access

D

A	Y
a_1	1
a_1	2

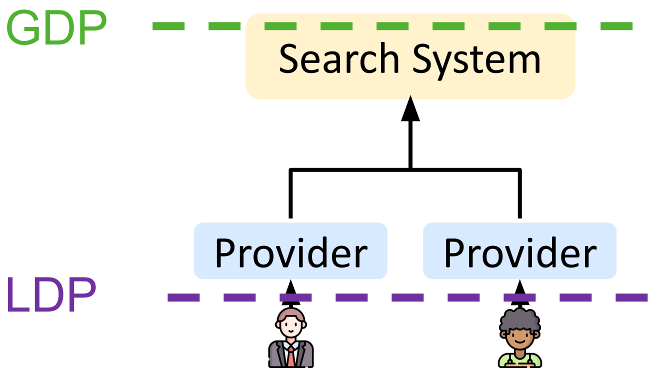
Remaining budget:

(0, 0)

Differential Privacy Mechanisms Available

Uses budget on every model eval

Too much noise to be useful



Data Task

ML data augmentation search

- More samples to union with.
- More features to join with.

Example: Predicting churn

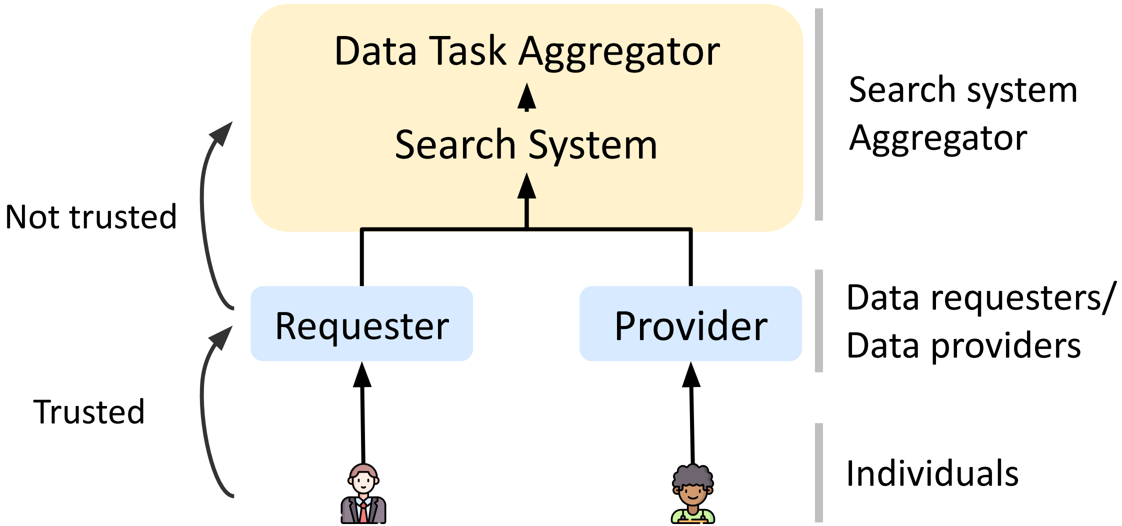
Churned	Customer	Subscription Date	Most Visited	Unemployment Rate
Yes	Alice	Jan 2023	Products	6.5%
No	Bob	May 2023	Support	3.2%
Yes	Charlie	Feb 2023	Support	8.1%
No	David	Jan 2023	Home	6.5%

DP ML Data Augmentation: Input and Output

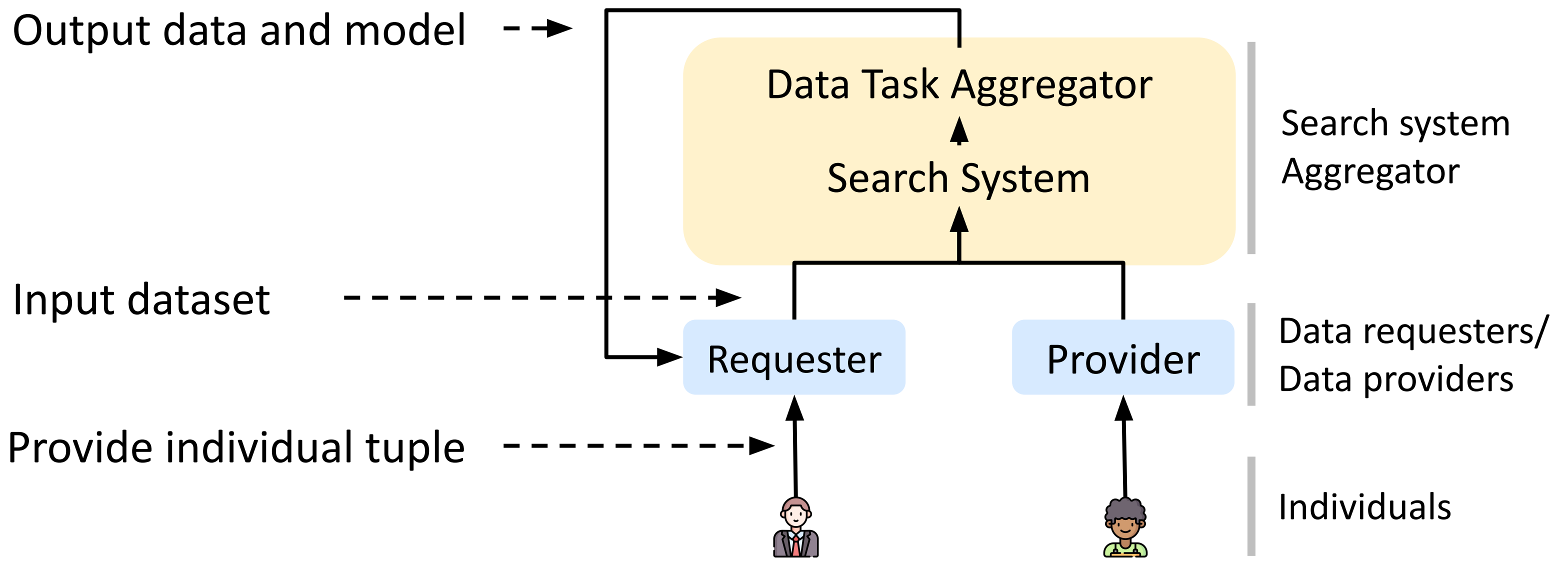
Want to find health data to improve cardiac prediction models

Patients don’t trust Google to use health data for ads.

Patients trust their health tracking App, like Fitbit.



DP ML Data Augmentation: Input and Output

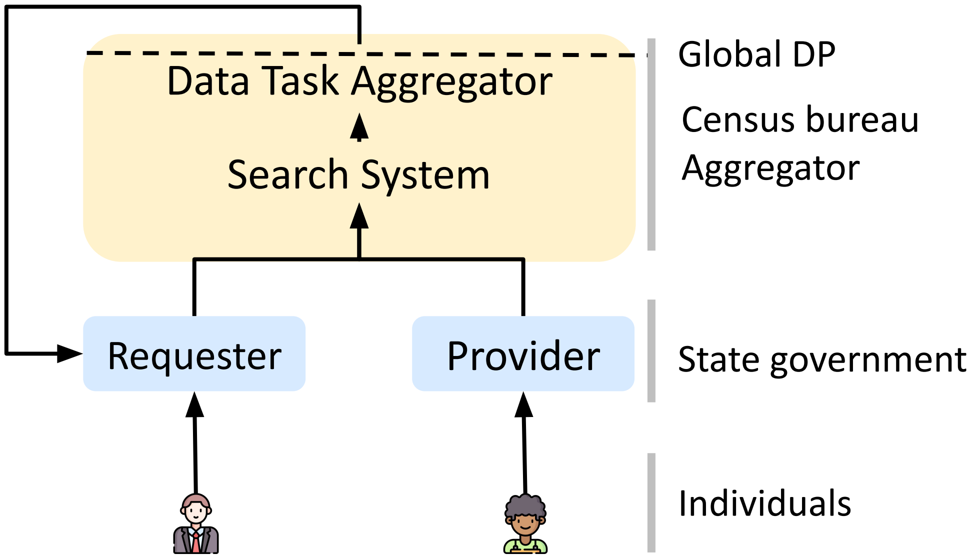


Existing Approach Limitations: Global DP

Global DP mechanisms add noise before releasing the output.

Evaluating each combination drains privacy budget.

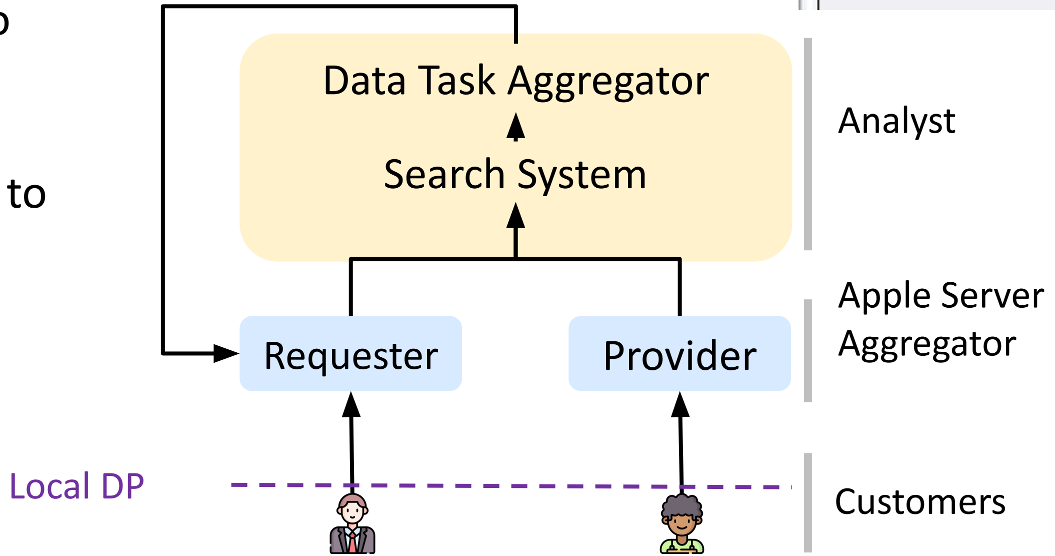
Exponential combinations of join/union-compatible sets.



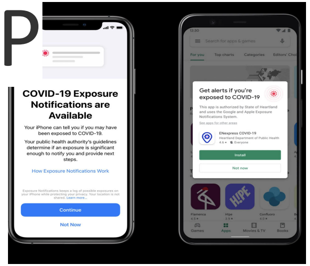
Existing Approach Limitations: Local DP

Local DP mechanisms add noise to each customer's data.

Augmentations too noisy, difficult to distinguish useful ones.

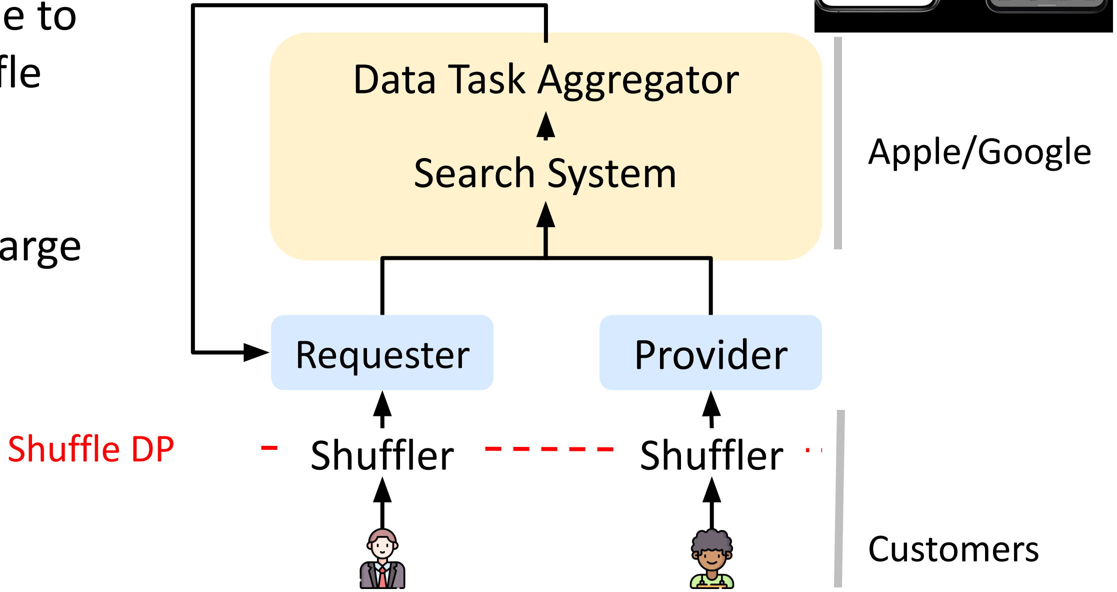


Existing Approach Limitations: Shuffle DP



Shuffle DP mechanisms add noise to each customer's data, then shuffle to enhance privacy.

Only enhance privacy levels for large datasets.

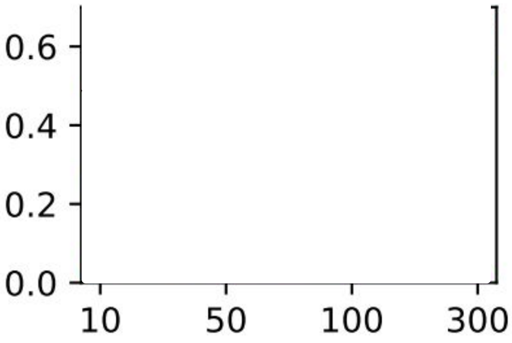


Prior Mechanisms Don't Scale

To repository size & number of requests
Vary between 10 - 329 NYC Open Datasets in Repo
Query: Grad table to predict 2016-17 graduation outcomes.

No privacy

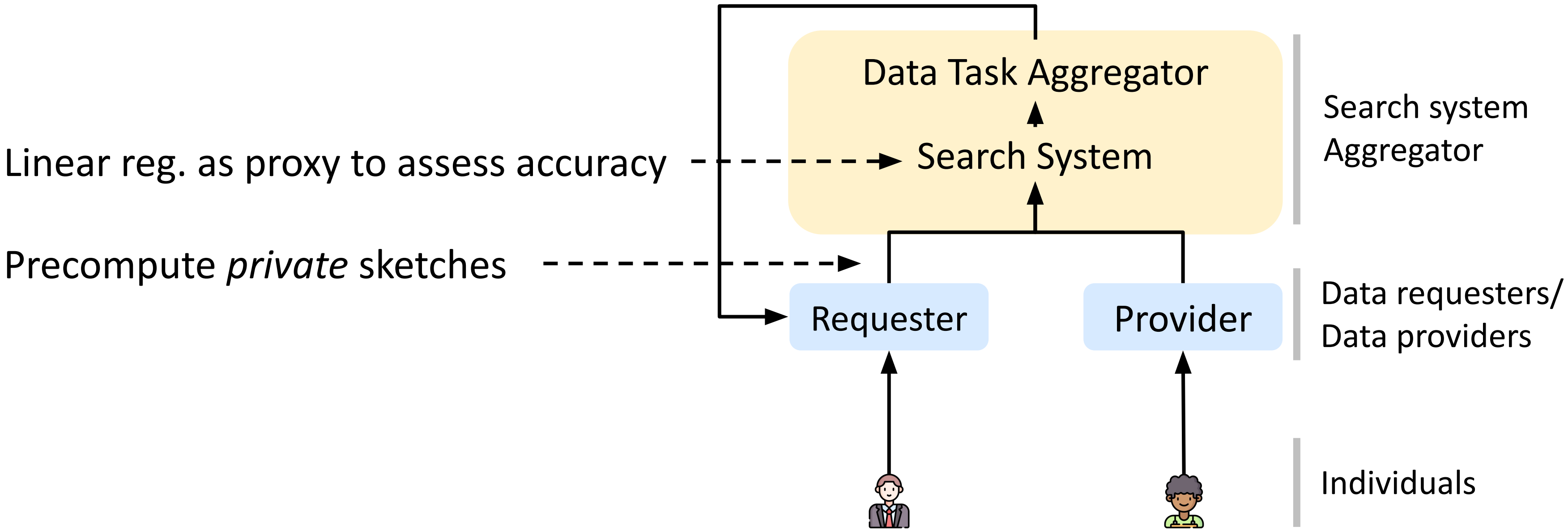
Final Model Accuracy



LDP
Shuffle
GDP

Repository size

Sketch-based Approach



Sketch-based Search

Linear regression has closed form solution $\widehat{\beta} = (X^T X)^{-1} X^T y$

$$X^T X = \begin{bmatrix} \sum x_1 x_1 & \dots & \sum x_1 x_m \\ \dots & \dots & \dots \\ \sum x_m x_1 & & \sum x_m x_m \end{bmatrix}$$

Sketch-based Search

How to compute sum of pairwise product between features?

D

A	Y
a_1	1
a_1	2

Compute aggregates as sketches

Monomial semi-ring

A	sum(Y^2)	sum(Y)	count
a_1	1^2+2^2	1+2	2

Sum of 0^{th}, 1^{st}, 2^{nd}-order monomials

Sketch-based Search

Linear regression on D⋈R requires computing $\boxed{\sum 1, \sum B, \sum Y, \sum BY}$

D

A	Y
a_1	1
a_1	2

⋈

R

A	B
a_1	1
a_1	2

D ⋈ R

A	Y	B
a_1	1	1
a_1	1	2
a_1	2	1
a_1	2	2

= 9

Sketch-based Search

Linear regression on D⋈R requires computing $\sum 1$, $\sum B$, $\sum Y$, $\boxed{\sum BY}$

D

A	Y
a_1	1
a_1	2

⋈

R

A	B
a_1	1
a_1	2

D ⋈ R

A	Y	B
a_1	1	1
a_1	1	2
a_1	2	1
a_1	2	2

= 9

Sketch-based Search

Linear regression on D⋈R requires computing $\sum 1$, $\sum B$, $\sum Y$, $\boxed{\sum BY}$

D

A	Y
a_1	1
a_1	2

1^{st}-order

A	sum(Y)
a_1	1+2

⋈

R

A	B
a_1	1
a_1	2

D ⋈ R

A	Y	B
a_1	1	1
a_1	1	2
a_1	2	1
a_1	2	2

= 9

Sketch-based Search

Linear regression on D⋈R requires computing $\sum 1$, $\sum B$, $\sum Y$, $\boxed{\sum BY}$

D

A	Y
a_1	1
a_1	2

1st-order

A	sum(Y)
a_1	1+2

⋈

1st-order

A	sum(B)
a_1	1+2

R

A	B
a_1	1
a_1	2

D ⋈ R

A	Y	B
a_1	1	1
a_1	1	2
a_1	2	1
a_1	2	2

= 9

Sketch-based Search

Linear regression on D⋈R requires computing $\sum 1, \sum B, \sum Y, \boxed{\sum BY}$

D

A	Y
a_1	1
a_1	2

1^{st}-order

A	sum(Y)
a_1	1+2

$\otimes$

1^{st}-order

A	sum(B)
a_1	1+2

R

A	B
a_1	1
a_1	2

2^{nd}-order

A	sum(B×Y)
a_1	3×3

= 9

Sketch-based Search

Linear regression on D⋈R requires computing $\sum 1$, $\sum B$, $\sum Y$, $\boxed{\sum BY}$

D

A	Y
a_1	1
a_1	2

1^{st}-order

A	sum(Y)
a_1	1+2

$\otimes$

1^{st}-order

A	sum(B)
a_1	1+2

R

A	B
a_1	1
a_1	2

D ⋈ R

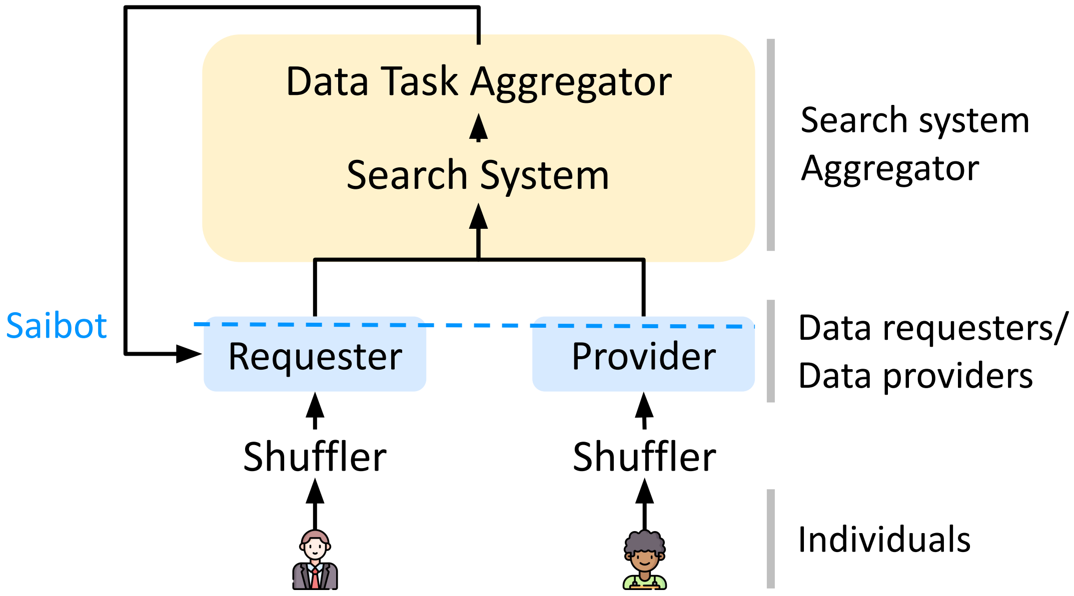
A	Y	B
a_1	1	1
a_1	1	2
a_1	2	1
a_1	2	2

= 9

Saibot: Our Contribution

Factorized Privacy Mechanism

Privatize sketches so downstream search will be privatized.



Intuition: aggregate datasets as much as possible before adding noise to them.

Saibot: Technical Details

- ❖ Factorized Privacy Mechanism (FPM).
- ❖ Noise allocation optimization.
- ❖ Unbiased estimation.
- ❖ Proofs

Saibot: Technical Details

- Factorized Privacy Mechanism (FPM).
- Noise allocation optimization.
- Unbiased estimation.
- Proofs

Saibot: Assumptions

The schema and join keys for datasets owned by providers are public

- Oblivious intersection techniques can be applied.

All tuples are L2 bounded by B (for analysis)

- Categorical features numericalized

FPM:Privatize sketches with privacy budget (ϵ, δ)

Use existing DP query engine

D

A	Y
a_1	1

$\xrightarrow{Q}$

A	sum(Y^2)	sum(Y)	count(Y)
a_1	1^2	1	1

Q: SELECT SUM(**Y**^{2}), SUM(**Y**), COUNT(**Y**) from **D** GROUP BY **A**

FPM:Privatize sketches with privacy budget (ϵ, δ)

Use existing DP query engine

Neighbour

D

A	**Y**
a_1	1

D'

A	**Y**
a_1	2

Q (D →)

A	**sum(Y²)**	**sum(Y)**	**count(Y)**
a_1	1^2	1	1

Q (D' →)

A	**sum(Y²)**	**sum(Y)**	**count(Y)**
a_1	2^2	2	1

Q: SELECT SUM(**Y**^{2}), SUM(**Y**), COUNT(**Y**) from **D** GROUP BY **A**

FPM:Privatize sketches with privacy budget (ϵ, δ)

Use existing DP query engine

D

A	Y
a_1	1

Q

A	sum(Y^2)	sum(Y)	count(Y)
a_1	1^2	1	1

Neighbour

L2 distance

D’

A	Y
a_1	2

Q

A	sum(Y^2)	sum(Y)	count(Y)
a_1	2^2	2	1

Sensitivity of Q: $\Delta(\mathbf{Q}) = \|\mathbf{Q}(\mathbf{D}) - \mathbf{Q}(\mathbf{D'})\|_2$

FPM:Privatize sketches with privacy budget (ϵ, δ)

Use existing DP query engine

D

A	**Y**
a_1	1

$\xrightarrow{\textbf{Q}}$

A	**sum(Y^2)**	**sum(Y)**	**count(Y)**
a_1	1^2	1	1

$$\mathcal{N}\left(0, \frac{\sqrt{2\ln(1.25/\delta)}\Delta(Q)}{\boxed{\epsilon}}\right)$$

Budget

Q: SELECT SUM(**Y**^{2}), SUM(**Y**), COUNT(**Y**) from **D** GROUP BY **A**

FPM:Privatize sketches with privacy budget (ϵ, δ)

Use existing DP query engine

D

A	**Y**
a_1	1

$\xrightarrow{\mathbf{Q}}$

A	**sum(Y²)**	**sum(Y)**	**count(Y)**
a_1	1^2+e_1	$1+e_2$	$1+e_3$

e_1, e_2, e_3 $\mathcal{N}\left(0, \frac{\sqrt{2\ln(1.25/\delta)}\Delta(Q)}{\epsilon}\right)$

Q: SELECT SUM(**Y**^{2}), SUM(**Y**), COUNT(**Y**) from **D** GROUP BY **A**

Naive Solution Limitation:Combining Sketches

A	$sum(C^2)$	$sum(C)$	$count(C)$
a_1	$2^2+e'_1$	$2+e'_2$	$1+e'_3$

A	$sum(Y^2)$	$sum(Y)$	$count(Y)$
a_1	1^2+e_1	$1+e_2$	$1+e_3$

A	$sum(C^2)$	$sum(Y^2)$	$sum(C \times Y)$	$sum(C)$	$sum(Y)$	count
a_1	$(2^2+e'_1)(1+e_3)$	$(1+e'_3)(1^2+e_1)$	$(2+e'_2)(1+e_2)$	$(2+e'_2)(1+e_3)$	$(1+e'_3)(1+e_2)$	$(1+e'_3)(1+e_3)$

Naive Solution Limitation:Combining Sketches

A	sum(C^2)	sum(C)	count(C)
a_1	$2^2+e'_1$	$2+e'_2$	$1+e'_3$

A	sum(Y^2)	sum(Y)	count(Y)
a_1	1^2+e_1	$1+e_2$	$1+\mathbf{e_3}$

A	sum(C^2)	sum(Y^2)	sum(C×Y)	sum(C)	sum(Y)	count
a_1	$(2^2+e'_1)(1+\mathbf{e_3})$	$(1+e'_3)(1^2+e_1)$	$(2+e'_2)(1+e_2)$	$(2+e'_2)(1+\mathbf{e_3})$	$(1+e'_3)(1+e_2)$	$(1+e'_3)(1+\mathbf{e_3})$

Naive Solution Limitation: Combining Sketches

A	sum(C²)	sum(C)	count(C)
a_1	$2^2+e'_1$	$2+e'_2$	$1+e'_3$

A	sum(Y²)	sum(Y)	count(Y)
a_1	1^2+e_1	$1+\mathbf{e_2}$	$1+e_3$

A	sum(C²)	sum(Y²)	sum(C×Y)	sum(C)	sum(Y)	count
a_1	$(2^2+e'_1)(1+e_3)$	$(1+e'_3)(1^2+e_1)$	$(2+e'_2)(1+\mathbf{e_2})$	$(2+e'_2)(1+e_3)$	$(1+e'_3)(1+\mathbf{e_2})$	$(1+e'_3)(1+e_3)$

Naive Solution Limitation:Combining Sketches

A	sum(C²)	sum(C)	count(C)
a_1	$2^2+e'_1$	$2+e'_2$	$1+e'_3$

A	sum(Y²)	sum(Y)	count(Y)
a_1	$1^2+\mathbf{e_1}$	$1+e_2$	$1+e_3$

A	sum(C²)	sum(Y²)	sum(C×Y)	sum(C)	sum(Y)	count
a_1	$(2^2+e'_1)(1+e_3)$	$(1+e'_3)(1^2+\mathbf{e_1})$	$(2+e'_2)(1+e_2)$	$(2+e'_2)(1+e_3)$	$(1+e'_3)(1+e_2)$	$(1+e'_3)(1+e_3)$

Noise Allocation

How to draw noise from different distributions to aggregations?

$sum(Y^2)$	$sum(Y)$	$count(Y)$
1^2+e_1	$1+e_2$	$1+e_3$

Noise Allocation

How to draw noise from different distributions to aggregations?

sum(Y^2)	sum(Y)	count(Y)
1^2+e_1	$1+e_2$	$1+e_3$

Sensitivity $O(B^2)$

Sensitivity $O(B)$

Sensitivity $O(1)$

Noise Allocation

How to draw noise from different distributions to aggregations?

sum(Y²)	sum(Y)	count(Y)
1^2+e_1	$1+e_2$	$1+e_3$

$e_1 \sim \mathcal{N}\left(0, \frac{\sqrt{2\ln(1.25/\delta)}B^2}{\epsilon_1}\right)$ $e_2 \sim \mathcal{N}\left(0, \frac{\sqrt{2\ln(1.25/\delta)}B}{\epsilon_2}\right)$ $e_3 \sim \mathcal{N}\left(0, \frac{\sqrt{2\ln(1.25/\delta)}}{\epsilon_3}\right)$

Noise Allocation

How to draw noise from different distributions to aggregations?

sum(Y^2)	sum(Y)	count(Y)
1^2+e_1	$1+e_2$	$1+e_3$

$$e_1 \sim \mathcal{N}\left(0, \frac{\sqrt{2\ln(1.25/\delta)}B^2}{\epsilon/3}\right)$$

$$e_2 \sim \mathcal{N}\left(0, \frac{\sqrt{2\ln(1.25/\delta)}B}{\epsilon/3}\right)$$

$$e_3 \sim \mathcal{N}\left(0, \frac{\sqrt{2\ln(1.25/\delta)}}{\epsilon/3}\right)$$

Noise Allocation: Analysis

Bounding linear regression estimator: $|\hat{\beta}_{\boldsymbol{x}} - \tilde{\beta}_{\boldsymbol{x}}| \leq \tau_2 + \frac{\tau_1}{1-\tau_1}\left(\hat{\beta}_{\boldsymbol{x}} + \tau_2\right)$

Noise Allocation: Analysis

Bounding linear regression estimator: $|\hat{\beta}_x - \tilde{\beta}_x| \leq \tau_2 + \frac{\tau_1}{1-\tau_1}\left(\hat{\beta}_x + \tau_2\right)$

Naive Method: $\tau_1 = O\left(\frac{B^4\sqrt{d}\ln(1/\delta)\ln(1/p)}{\epsilon^2 n\widehat{\sigma_{\mathrm{x}}^2}}\right)$ $\tau_2 = O\left(\frac{B^4\ln(1/p)\ln(1/\delta)\sqrt{d\ln(d/p)}}{\epsilon^2\sqrt{n}\widehat{\sigma_{\mathrm{x}}^2}}\right)$

Noise Allocation: Analysis

Bounding linear regression estimator: $|\hat{\beta}_x - \tilde{\beta}_x| \leq \tau_2 + \frac{\tau_1}{1-\tau_1}\left(\hat{\beta}_x + \tau_2\right)$

Naive Method:

$$\tau_1 = O\left(\frac{B^4\sqrt{d}\ln(1/\delta)\ln(1/p)}{\epsilon^2 n\widehat{\sigma_{\mathrm{x}}^2}}\right) \quad \tau_2 = O\left(\frac{B^4\ln(1/p)\ln(1/\delta)\sqrt{d\ln(d/p)}}{\epsilon^2\sqrt{n}\widehat{\sigma_{\mathrm{x}}^2}}\right)$$

Optimization:

$$\tau_1 = O\left(\frac{B^2\sqrt{d}\ln(1/\delta)\ln(1/p)}{\epsilon^2 n\widehat{\sigma_{\mathrm{x}}^2}}\right), \tau_2 = O\left(\frac{B^2\ln(1/p)\ln(1/\delta)\sqrt{d\ln(d/p)}}{\epsilon^2\sqrt{n}\widehat{\sigma_{\mathrm{x}}^2}}\right)$$

Noise Allocation: Analysis

Bounding linear regression estimator: $|\hat{\beta}_x - \tilde{\beta}_x| \leq \tau_2 + \frac{\tau_1}{1-\tau_1}\left(\hat{\beta}_x + \tau_2\right)$

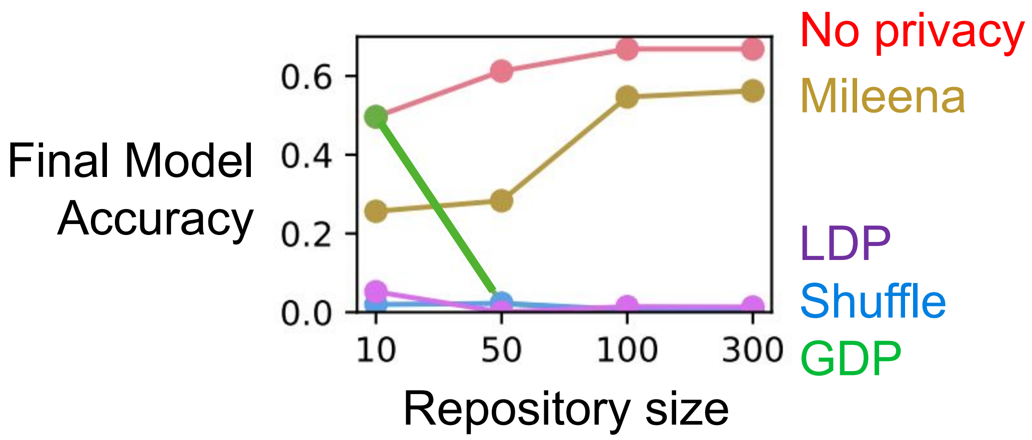
Naive Method: $\tau_1 = O\left(\frac{\boxed{B^4}\sqrt{d}\ln(1/\delta)\ln(1/p)}{\epsilon^2 n\widehat{\sigma_x^2}}\right)$ $\tau_2 = O\left(\frac{\boxed{B^4}\ln(1/p)\ln(1/\delta)\sqrt{d\ln(d/p)}}{\epsilon^2\sqrt{n}\widehat{\sigma_x^2}}\right)$

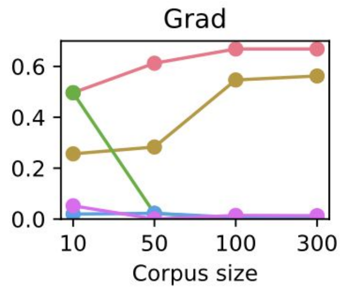
Optimization: $\tau_1 = O\left(\frac{\boxed{B^2}\sqrt{d}\ln(1/\delta)\ln(1/p)}{\epsilon^2 n\widehat{\sigma_x^2}}\right), \tau_2 = O\left(\frac{\boxed{B^2}\ln(1/p)\ln(1/\delta)\sqrt{d\ln(d/p)}}{\epsilon^2\sqrt{n}\widehat{\sigma_x^2}}\right)$

Reduce the bound on linear regression parameter by O(**B^2**)

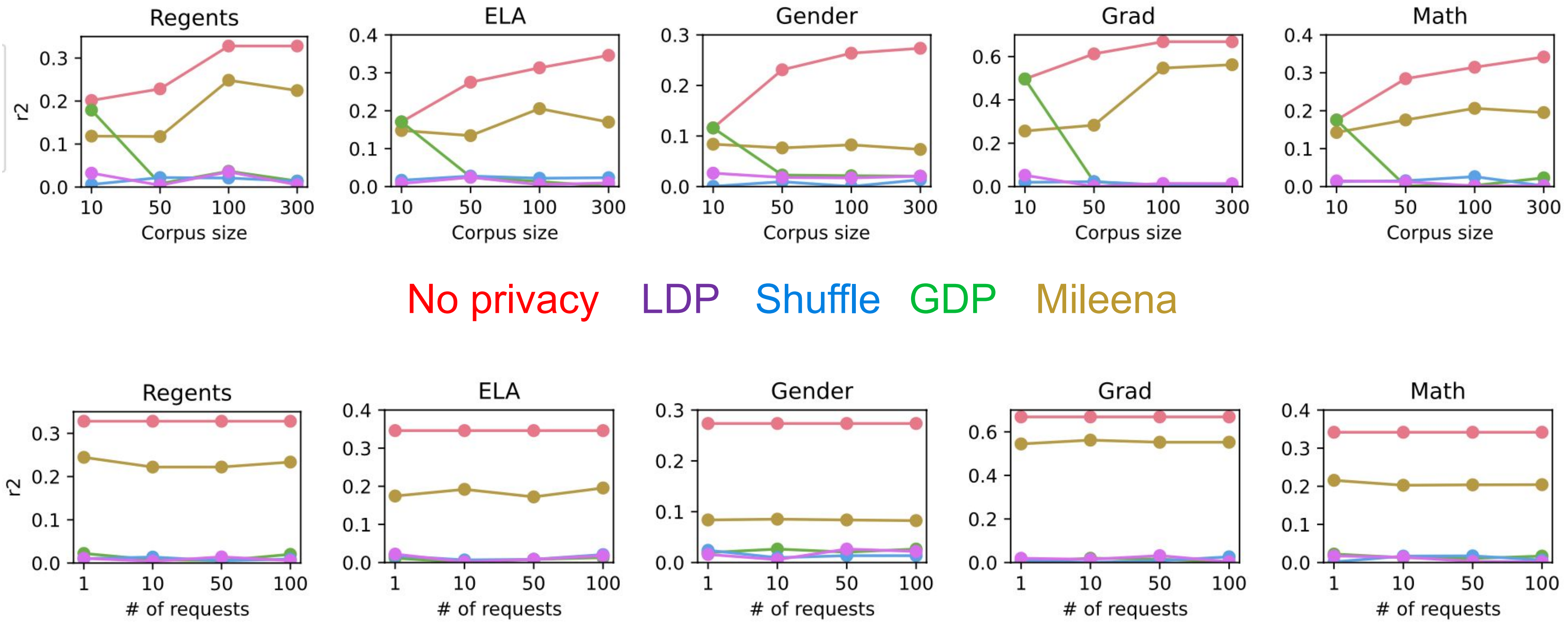
FPM Scales

To repository size & number of requests
Vary between 10 - 329 NYC Open Datasets in Repo





No privacy LDP Shuffle GDP Mileena



Regents
ELA
Gender
Grad
Math
r2
Corpus size
No privacy
LDP
Shuffle
GDP
Mileena
# of requests

Part 4: Regulatory Considerations

Agenda

Goal: Overview (but not exhaustive!)

- Background and Motivation
- Legal Landscape: Key Frameworks
- Translating Frameworks to Implementations
- Security and Breach Notification Requirements
- Cross-border Data Flows
- Technical-Legal Interplay

Story: Why Legal Considerations are Important

OPPORTUNITY INSIGHTS

Case Study: What counties/states in the U.S.A have better or worse economic/social mobility?

https://opportunityinsights.org/

Story: Why Legal Considerations are Important

Case Study: What counties/states in the U.S.A have better or worse economic/social mobility?

OPPORTUNITY INSIGHTS

Solution: Use statistical methods and quantitative social science to study the question.

https://opportunityinsights.org/

Story: Why Legal Considerations are Important

"…There are several steps in our estimation process. We begin by combining three sources of [...] data linked by and housed at the Census Bureau (the 2000 and 2010 Decennial Census short forms; federal income tax returns for 1984, 1989, 1994, 1995, and 1998-2019; and the 2000 Decennial Census long form and the 2005-2019 American Community Surveys) to construct an analysis sample of Americans born between 1978-1992. We map these individuals back to the counties where they lived as children and measure their outcomes at age 27 (between 2005-2019). Parent and child income are measured using their percentile ranks in the national income distribution…."

Source: https://opportunityinsights.org/policy/frequently-asked-questions/

Story: Why Legal Considerations are Important

OPPORTUNITY INSIGHTS

Case Study: What counties/states in

the U.S.A

have better economic/social mobility?

First, collect raw data from IRS, Census Bureau.

But Census Bureau needs to adhere to

Title 13.

Screw up, then employees go to jail!

Title 13 and U.S. Census Bureau

Title 13, U.S. Code

The Census Bureau is bound by Title 13 of the United States Code. These laws not only provide authority for the work we do, but also provide strong protection for the information we collect from individuals and businesses.

Title 13 provides the following protections to individuals and businesses:

- Private information is never published. It is against the law to disclose or publish any private information that identifies an individual or business such, including names, addresses (including GPS coordinates), Social Security Numbers, and telephone numbers.
- The Census Bureau collects information to produce statistics. Personal information cannot be used against respondents by any government agency or court.
- Census Bureau employees are sworn to protect confidentiality. People sworn to uphold Title 13 are legally required to maintain the confidentiality of your data. Every person with access to your data is sworn for life to protect your information and understands that the penalties for violating this law are applicable for a lifetime.
- Violating the law is a serious federal crime. Anyone who violates this law will face severe penalties, including a federal prison sentence of up to five years, a fine of up to $250,000, or both.

Source: https://www.census.gov/history/www/reference/privacy_confidentiality/title_13_us_code.html

Title 13 and U.S. Census Bureau

Title 13, U.S. Code

The Census Bureau is bound by Title 13 of the United States Code. These laws not only provide authority for the work we do, but also provide strong protection for the information we collect from individuals and businesses.

Title 13 provides the following protections to individuals and businesses:

- Private information is never published. It is against the law to disclose or publish any private information that identifies an individual or business such, including names, addresses (including GPS coordinates), Social Security Numbers, and telephone numbers.
- The Census Bureau collects information to produce statistics. Personal information cannot be used against respondents by any government agency or court.
- Census Bureau employees are sworn to protect confidentiality. People sworn to uphold Title 13 are legally required to maintain the confidentiality of your data. Every person with access to your data is sworn for life to protect your information and understands that the penalties for violating this law are applicable for a lifetime.
- Violating the law is a serious federal crime. Anyone who violates this law will face severe penalties, including a federal prison sentence of up to five years, a fine of up to $250,000, or both.

Source: https://www.census.gov/history/www/reference/privacy_confidentiality/title_13_us_code.html

Bridging the Gap: Technical vs. Legal

Bridging the Gap between Computer Science and Legal Approaches to Privacy

Kobbi Nissim[3,1], Aaron Bembenek[1], Alexandra Wood[2], Mark Bun[1], Marco Gaboardi[4], Urs Gasser[2], David R. O'Brien[2], Thomas Steinke[1], and Salil Vadhan[1]

[1]Center for Research on Computation and Society, Harvard University.
{tsteinke|mbun|salil}@seas.harvard.edu, bembenek@g.harvard.edu.
[2]Berkman Klein Center for Internet & Society, Harvard University.
{awood|ugasser|dobrien}@cyber.law.harvard.edu.
[3]Dept. of Computer Science, Georgetown University. kobbi.nissim@georgetown.edu
[4]The State University of New York at Buffalo. gaboardi@buffalo.edu.

February 21, 2018

Bridging the Gap: Technical vs. Legal

"...the fields of law and computer science have generated different notions of privacy risks in the context of the analysis and release of statistical data about individuals…"

Source: Bridging the gap between computer science and legal approaches to privacy *(*Harv. JL & Tech.)

Bridging the Gap: Technical vs. Legal

"...this article articulates the nature of the gaps between legal and technical approaches to privacy in the release of statistical data about individuals. It also presents an argument that the use of differential privacy is sufficient to satisfy the requirements of the Family Educational Rights and Privacy Act of 1974 (FERPA), a federal law that protects the privacy of education records in the United States. This argument illustrates what may evolve to a more general methodology for rigorously arguing that technological methods for privacy protection satisfy the requirements of a particular information privacy law…"

Source: Bridging the gap between computer science and legal approaches to privacy *(*Harv. JL & Tech.)

Bridging the Gap:Title 13 and U.S. Census Bureau

"... In this way, the mathematical proof demonstrates that the use of differential privacy is sufficient to satisfy a broad range of reasonable interpretations of FERPA, including interpretations that may be adopted in the future…"

Source: Bridging the gap between computer science and legal approaches to privacy *(*Harv. JL & Tech.)

Step 1: Interpret Privacy Law

[62] For a discussion of the evolution and nature of the U.S. sectoral approach to privacy, see Paul M. Schwartz, *Preemption and Privacy*, 118 YALE L.J. 902 (2008).

[63] 18 U.S.C. § 2710(a)(3) (emphasis added).

[64] Cal. Civ. Code §§ 56–56.37.

[65] *See* Paul M. Schwartz & Daniel J. Solove, *The PII Problem: Privacy and a New Concept of Personally Identifiable Information*, 86 N.Y.U. L. REV. 1814 (2011).

[66] *See, e.g.*, Pineda v. Williams-Sonoma Stores, 246 P.3d 612, 612 (Cal. 2011) (reversing the lower courts and determining that a "cardholder's ZIP code, without more, constitutes personal identification information" within the meaning of the California Song-Beverly Credit Card Act of 1971 "in light of the statutory language, as well as the legislative history and evident purpose of the statute").

[67] 201 C.M.R. 17.02.

[68] 45 C.F.R. Part 160 and Subparts A and E of Part 164.

[69] 45 C.F.R. § 164.514. Note, however, that HIPAA's safe harbor standard creates ambiguity by requiring that the entity releasing the data "not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information." *Id.*

Source: Bridging the gap between computer science and legal approaches to privacy (Harv. JL & Tech.)

Step n+1: Translate into Technical Terms

Source: Bridging the gap between computer science and legal approaches to privacy (Harv. JL & Tech.)

Why Legal Considerations Matter

- Legal frameworks define data rights and duties
- Legal compliance is essential for trust and adoption
- Distributed data markets complicate governance
- Liability concerns for data market platforms (e.g., Opportunity Insights)

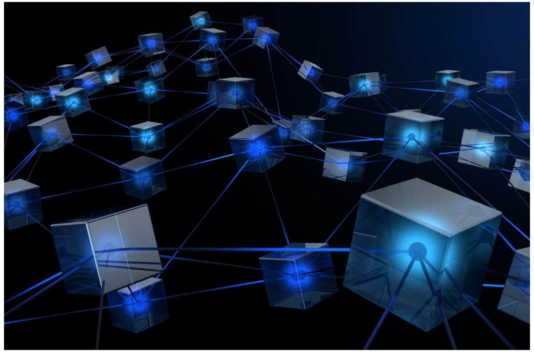
Compliance in Distributed Data Markets

- *Key challenge*: Trust among parties with differing incentives

Examples:

(1) Opportunity insights ⇔ Census Bureau (Title 13)

(2) Hospitals ⇔ Health Insurance Companies (HIPAA)



Legal Foundations (Global Overview)

- GDPR (EU)
- CCPA/CPRA (California), HIPAA (US), Title 13 (US)
- Varying consent, data definitions, cross-border rules

Further Examples of Translation (GDPR)

- *Legal*: Purpose limitation

"…Personal data shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes ('purpose limitation')…"

- *Technical*: Can't collect data for academic research and sell to advertisers

Source: https://gdpr-info.eu/art-5-gdpr/

Further Examples of Translation (GDPR)

- *Legal*: Purpose limitation and data minimization

"…Personal data shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes ('purpose limitation')…"

"…Personal data shall be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimisation');…"

- *Technical*: Can't ask for SSN when signing up for blog

Source: https://gdpr-info.eu/art-5-gdpr/

Further Examples of Translation (GDPR)

- *Legal*: Breach notification mandates (e.g., GDPR Art. 33)

"...In the case of a personal data breach, the controller shall without undue delay and, where feasible, not later than 72 hours after having become aware of it, notify the personal data breach to the supervisory authority competent in accordance with Article 55, unless the personal data breach is unlikely to result in a risk to the rights and freedoms of natural persons. Where the notification to the supervisory authority is not made within 72 hours, it shall be accompanied by reasons for the delay. The processor shall notify the controller without undue delay after becoming aware of a personal data breach..."

- *Technical*: Send notifications to supervisory authority

Source: https://gdpr-info.eu/art-33-gdpr/

Takeaways

- Legal frameworks shape privacy/security protocols
- Legal compliance ≠ technical privacy
- Must align PETs (Privacy-Enhancing Technologies) with regulatory requirements
- Learn about compliance from lawyers!!!

Part 5: Open Problems

Privacy Challenges Are Everywhere!

Query specification
Privacy risks

Data Market/Data Discovery

Query interface
Latency, Scalability
Privacy risks

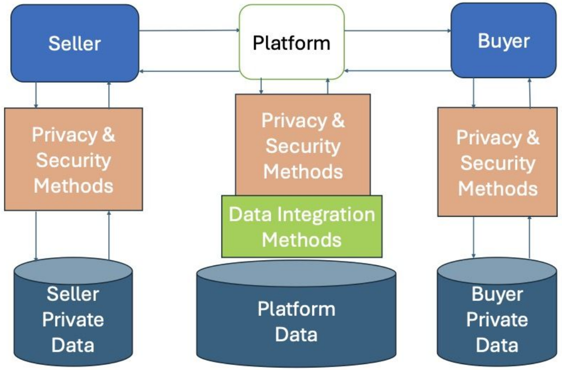
Stats | Stats | Stats

Data | Data | Data

Data acquisition
Data preparation
Privacy risks

Protect Information in Data Markets

1. Protect buyers from *malicious* sellers
2. Protect sellers from *malicious* buyers
3. Prevent *unauthorized* users from accessing:
   a. Seller private data
   b. Buyer private data
   c. Platform private data
4. Prevent manipulation of data acquisition mechanisms:
   a. Data discovery
   b. Data valuation
   c. Data negotiation
   d. Data delivery



Privacy and Security Attacks

- Naively allowing query access to data markets is risky for users/orgs
  - Linkage attacks
  - Reconstruction attacks
  - Inference attacks
  - Plaintext/ciphertext attacks
- Naive designs of data markets is risky for valuation
  - Manipulation of pricing and negotiation mechanisms
  - Less trust in data markets

Motivates the need for robust *privacy and security protections.*

We need more attacks for illustrative and motivational purposes.

Privacy and Security Attacks

- Naively allowing query access to data markets is risky for users/orgs
  - Linkage attacks
  - Reconstruction attacks
  - Inference attacks
  - Plaintext/ciphertext attacks
- Naive designs of data markets is risky for valuation
  - Manipulation of pricing and negotiation mechanisms
  - Less trust in data markets

Motivates the need for robust *privacy and security protections.*

We need more methods to protect against attacks.

Research Questions for Legal Considerations

- Can we cryptographically enforce legal policies?
- What counts as legally sufficient anonymization?
- Consent revocation in distributed systems?

Data Ownership and Stewardship

- Ambiguity in data and model ownership
- Data Controller vs. Data Processor roles
- Tension between legal rights and cryptographic control

An Agentic Web is a Data Market

Agent-friendly protocols like MCP sidestep web UIs completely

- No GUI, no user, just APIs and automation
- *“The web is a series of databases”* - Sundar Pichai on Decoder Podcast

In an agentic world, every “website” is a database API + business logic…

- Arrow’s paradox? Pricing? Privacy? Security? Discovery? Market structure?

More Future Directions

Our investigation into data marketplaces reveals critical challenges for building secure, decentralized AI systems.

1. The Attack Surface Has Shifted.

The primary vulnerability is not just the model, but the marketplace's economic and selection mechanisms.

2. Standard Metrics are Deceptive.

High model accuracy and low cost can mask catastrophic security failures and unfair economic outcomes.

3. Similarity-Based Defenses are Not a Silver Bullet.

They are fundamentally vulnerable to mimicry attacks and struggle most in the realistic, heterogeneous environments they are designed for.

Path Forward: Building a Robust Data Economy

To build truly secure and equitable marketplaces, future work must move beyond simple similarity checks. We need to focus on:

- **Orthogonal Trust Signals**: Integrating seller reputation, transaction history, and data provenance to make more holistic trust decisions.
- **Multi-Stage Filtering**: Designing a defense-in-depth pipeline that combines anomaly detection, similarity checks, and impact analysis.
- **Incentive-Compatible Mechanisms**: Creating reward and selection systems that are provably resilient to strategic manipulation and fairly compensate true value.

Funding Acknowledgements & Questions

SIMONS FOUNDATION