

Privacy and Security in Distributed Data Markets

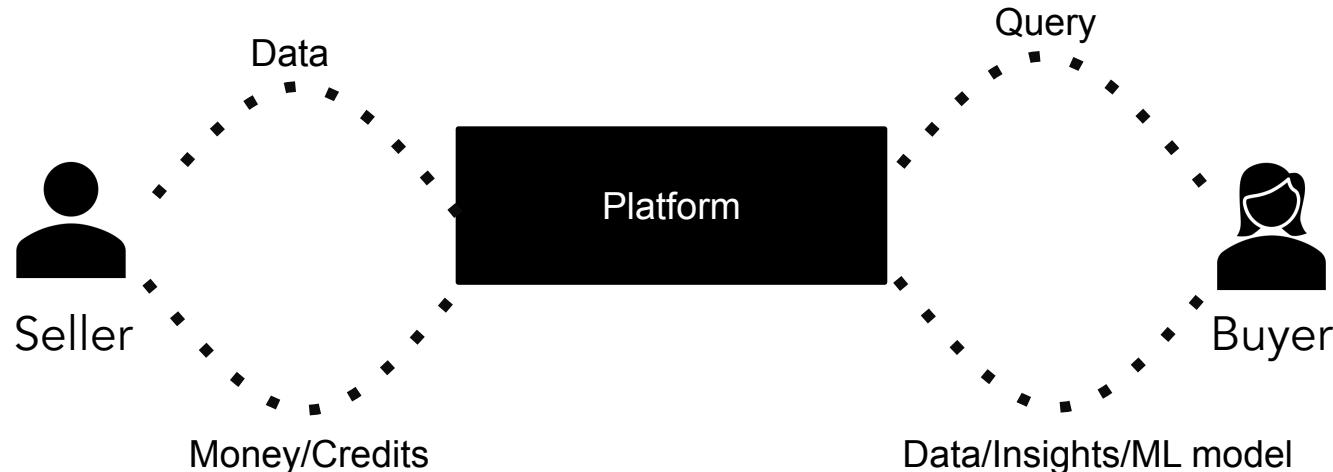
Daniel Alabi, Sainyam Galhotra, Shagufta Mehnaz, Zeyu Song, Eugene Wu

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

aws marketplace

Search

English ▾ Hello, backend ▾

About Categories Delivery Methods Solutions Resources Your Saved List Become a Channel Partner Sell in AWS Marketplace Amazon Web Services Home Help

▼ Refine results

◀ All categories

Data Products

- Retail, Location & Marketing Data (1503)
- Financial Services Data (1101)
- Healthcare & Life Sciences Data (575)
- Resources Data (541)
- Public Sector Data (495)
- Media & Entertainment Data (372)
- Telecommunications Data (244)
- Manufacturing Data (166)
- Automotive Data (164)
- Environmental Data (140)
- Gaming Data (38)

▼ Delivery methods

- Data Exchange (4640)
- Professional Services (55)
- SaaS (55)
- Amazon Machine Image (16)
- CloudFormation Template (3)
- Container Image (3)
- Helm Chart (2)

▼ Publisher

- Rearc (201)
- Techmap (172)
- mnAi (123)

Data Products (4772 results) showing 1 - 20

Sort By: Relevance


Email Marketing Campaign AI Agent
By [Baideac](#) | Ver v0.7
★★★★★ 1 AWS review
Email announcement agent is a tool that helps you make email marketing and announcement campaigns. Additionally, you can track the traction of emails and contacts.


Currency Exchange API
By [SilverLining.Cloud GmbH](#)
★★★★★ 1 AWS review
Leverage our Currency Exchange API to obtain near real-time currency exchange rates for over 140 international currencies. Our simple REST API delivers fast, reliable, and universally compatible JSON-formatted data for seamless integration with your applications. With our pay-per-use plan, you can...


CAYLENT Clinical Document Writer
By [Caylent](#)
Cut Documentation Time by 40% with AI that Works with You: Our solution uses Amazon Bedrock and Amazon SageMaker AI to create compliant documentation. It integrates fragmented data sources through agentic intelligence to automatically generate comprehensive regulatory documents (from protocols to pa...)


CAYLENT Clinical Trial Design Optimizer

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

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 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

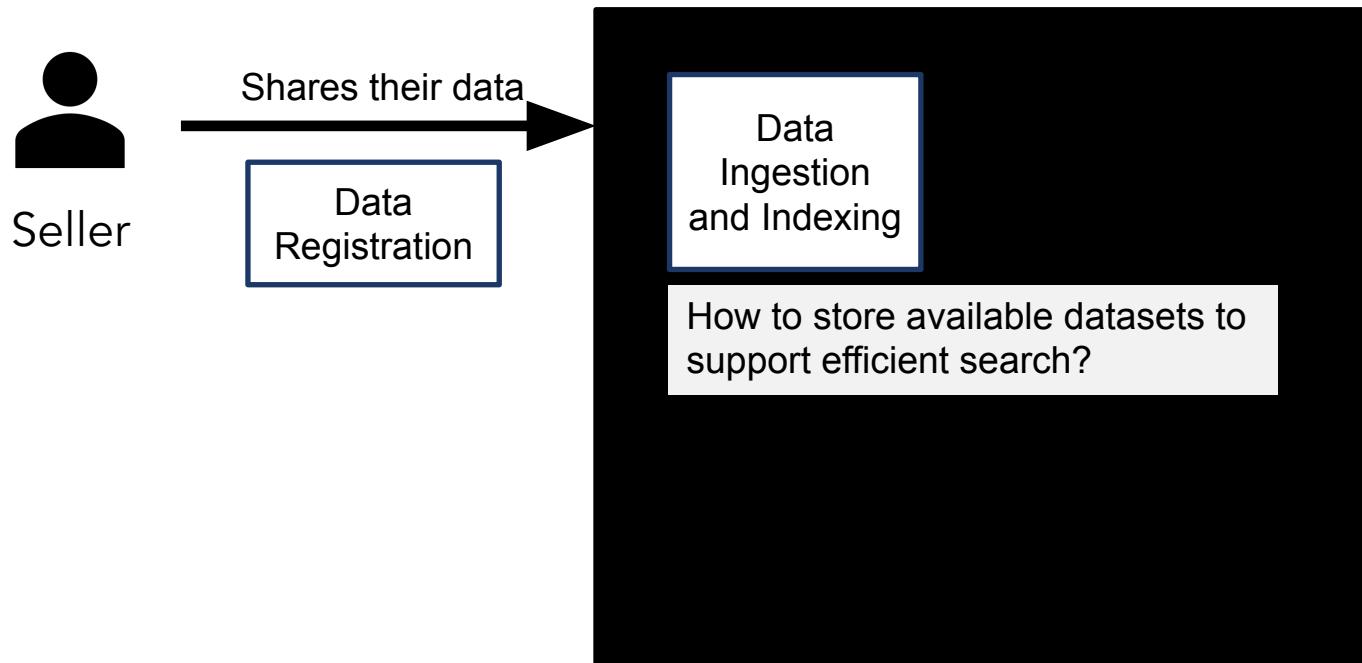
<https://oag.ca.gov/data-brokers>

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

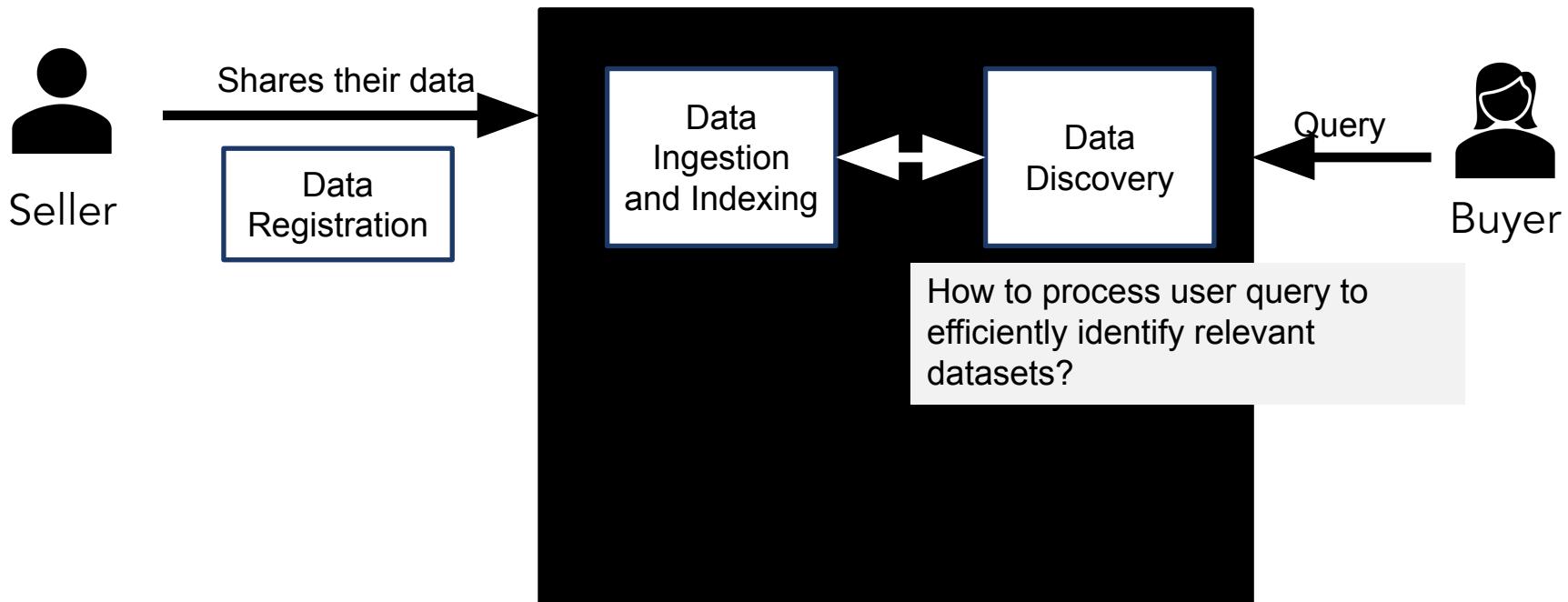
Key Components of a Data Market



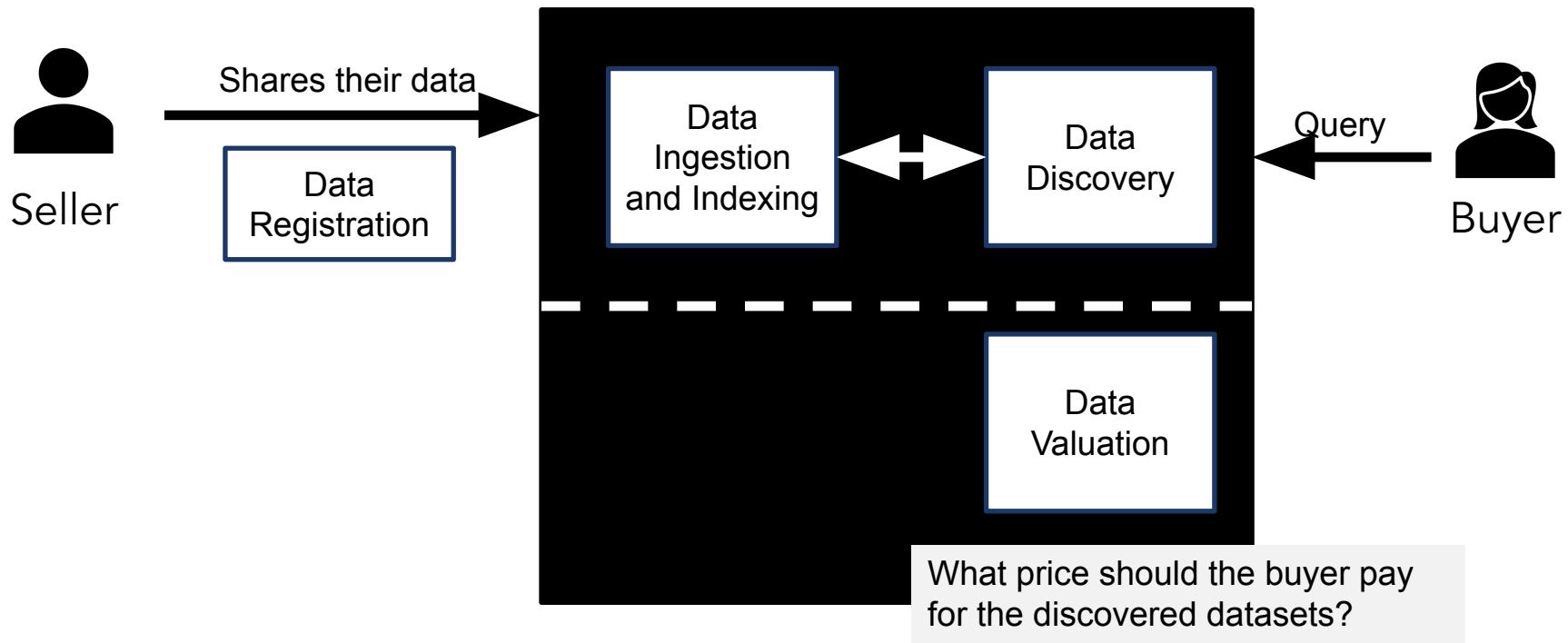
Key Components of a Data Market



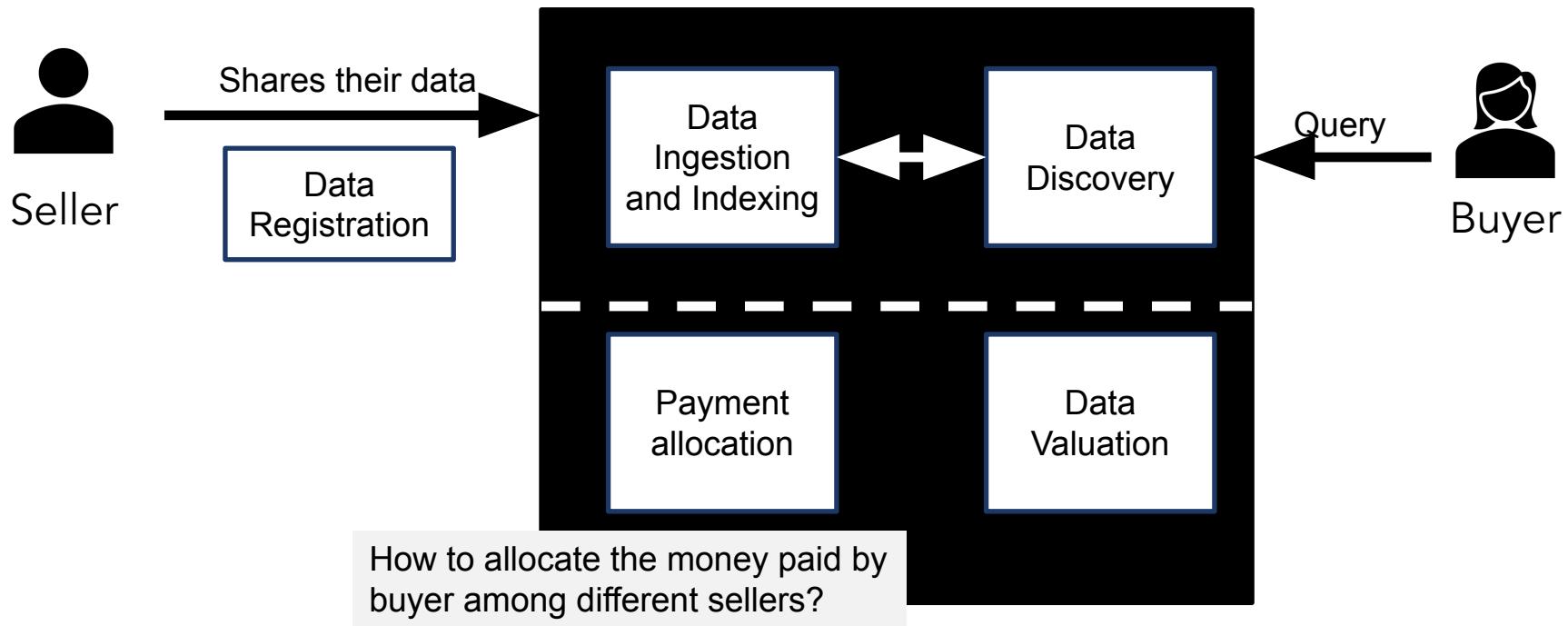
Key Components of a Data Market



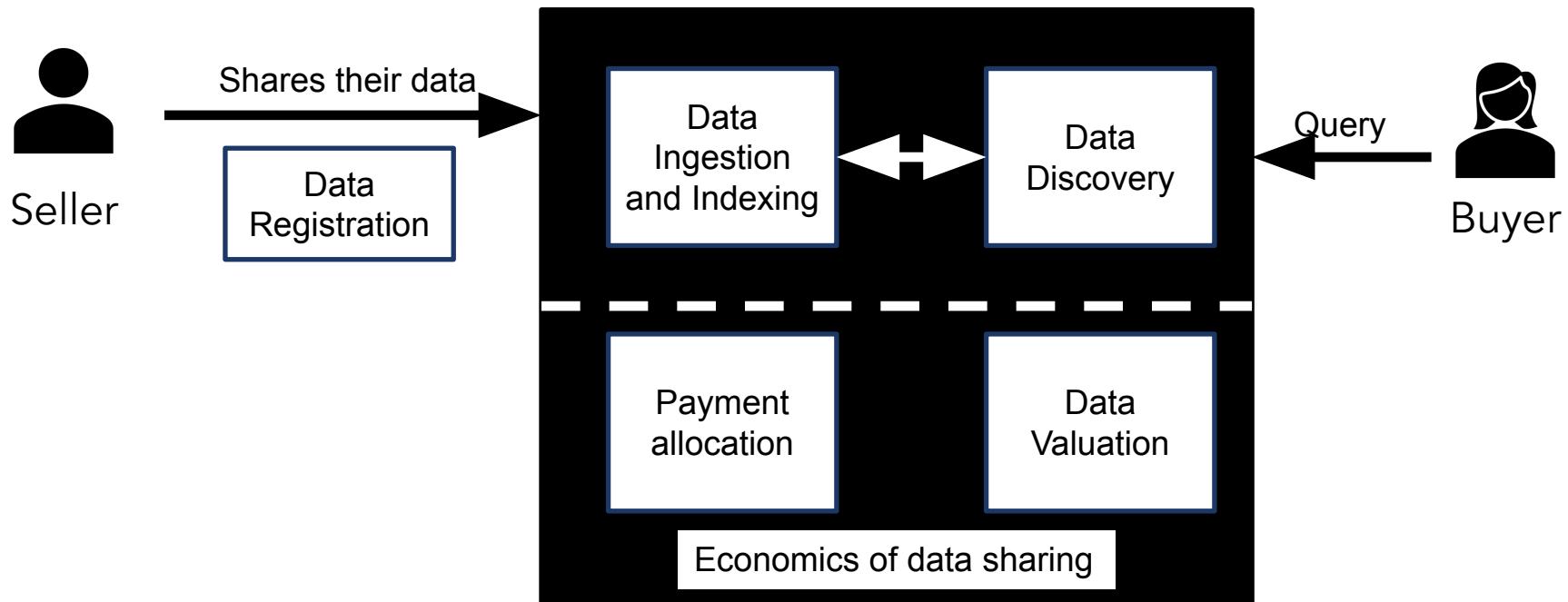
Key Components of a Data Market



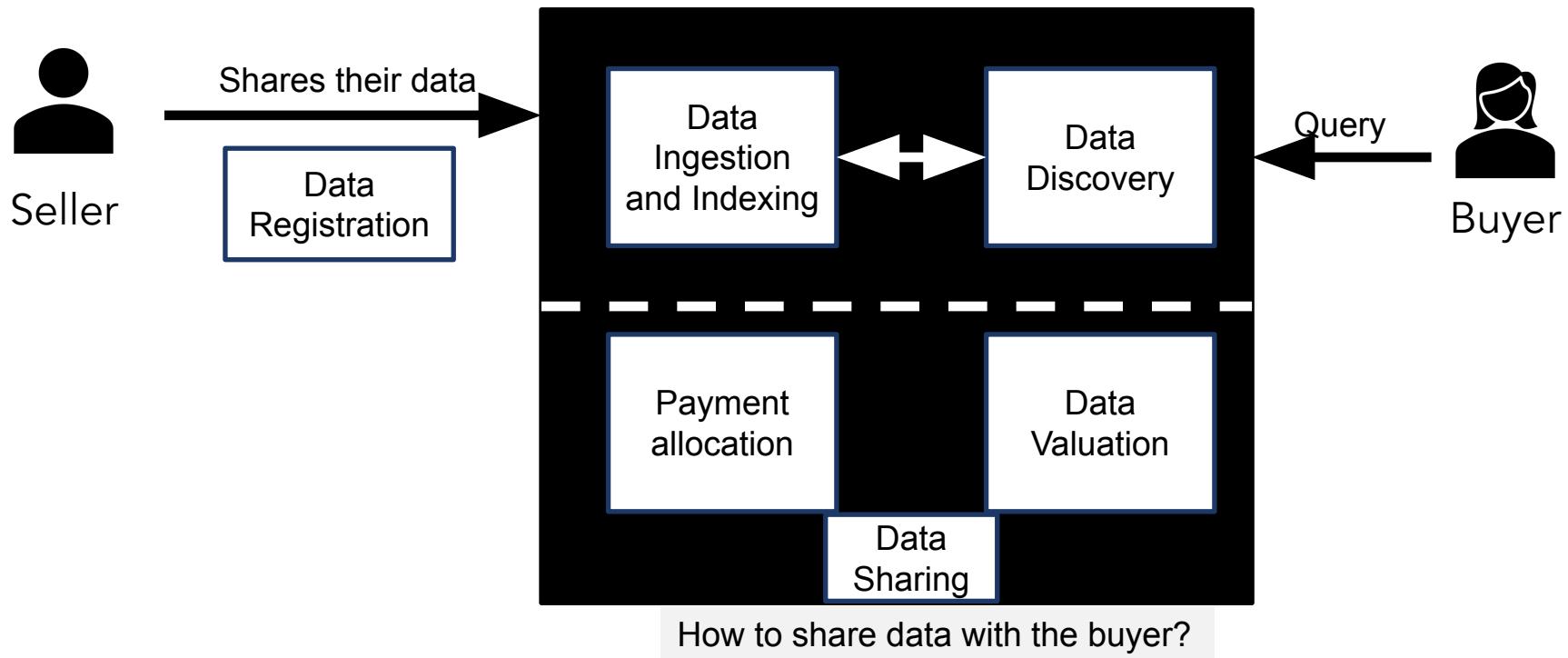
Key Components of a Data Market



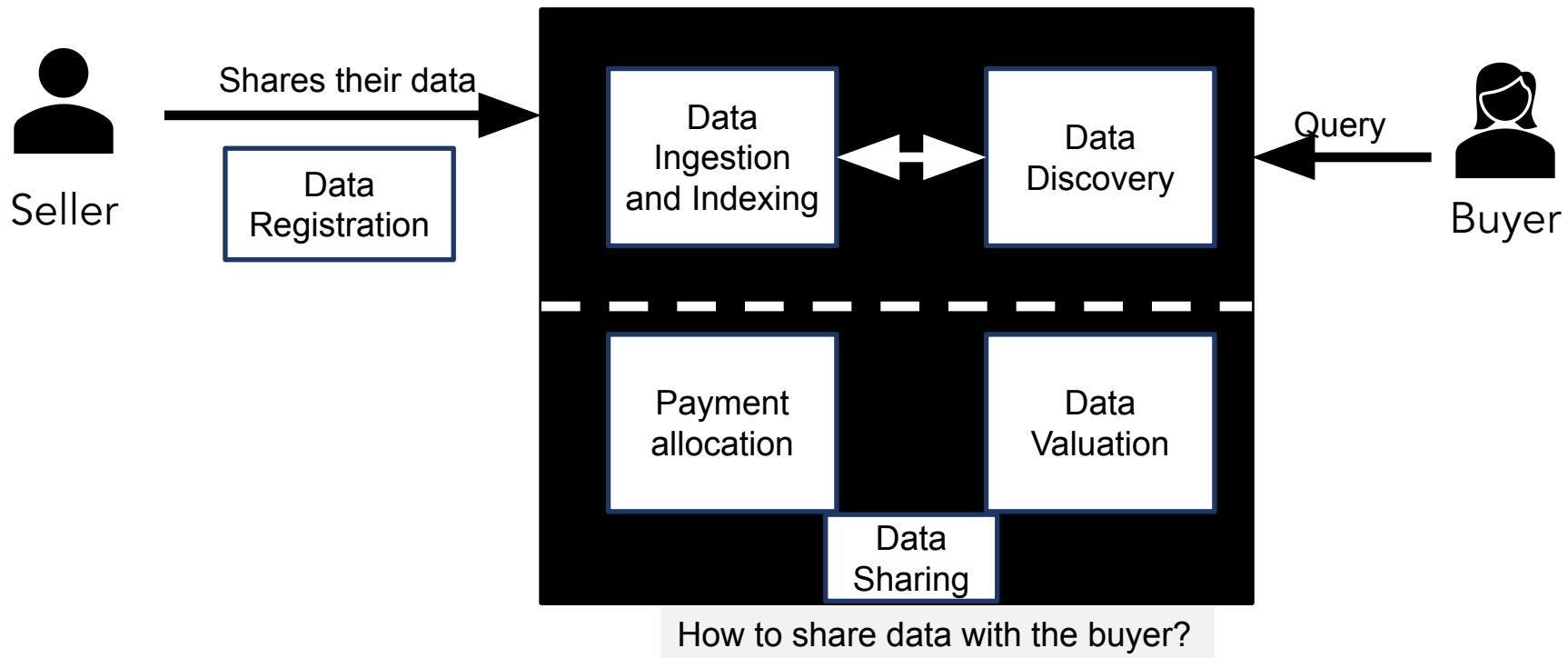
Key Components of a Data Market



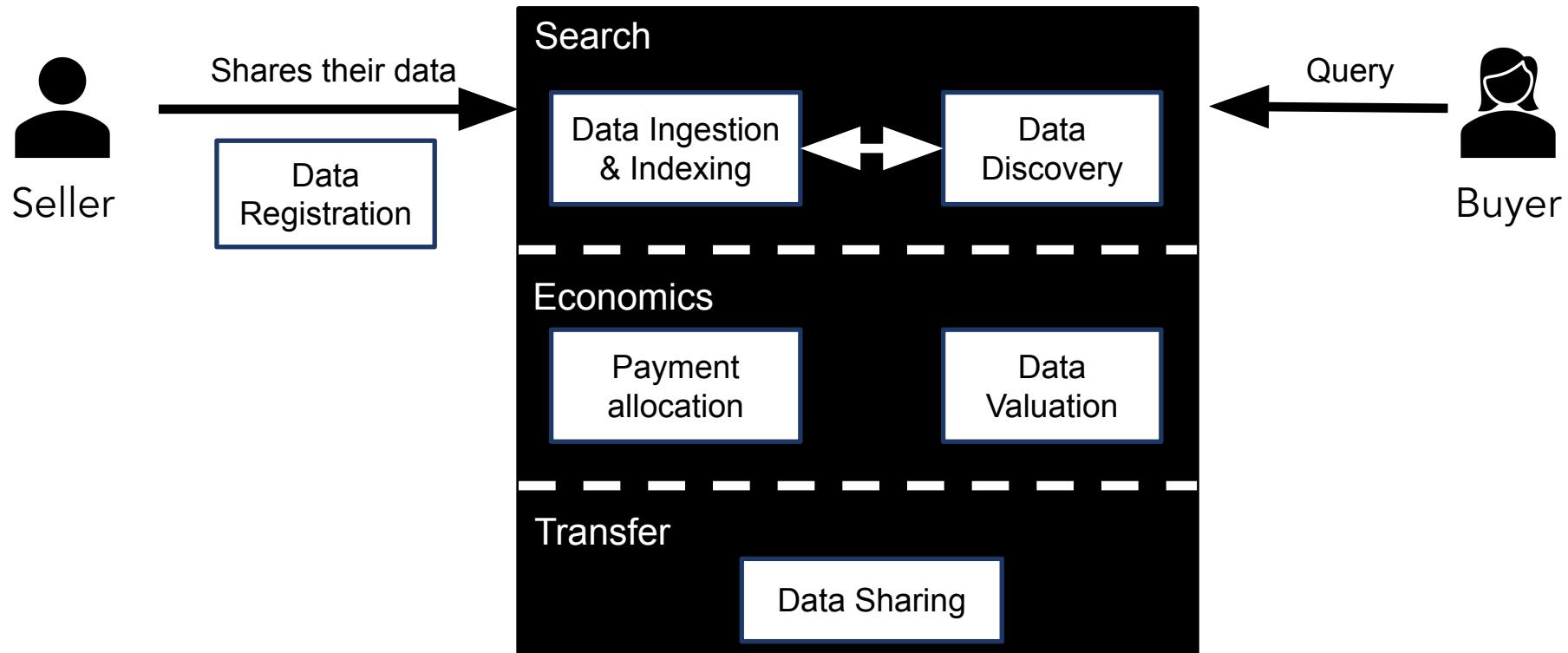
Key Components of a Data Market



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)
 - Arrow's 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

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

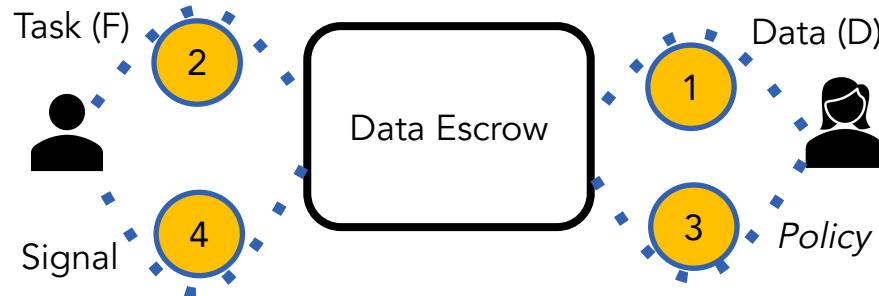
Tutorial Organization

- Part I: Data acquisition and search (Eugene and Sainyam)
- Part II: Privacy and Security Risks (Daniel)
- Part III: (Shagufta, Zeyu, and Eugene)
- Part IV: Regulatory Considerations (Daniel)
- Part V: Open Questions (Daniel, Eugene and Sainyam)

How to control what buyers can acquire?

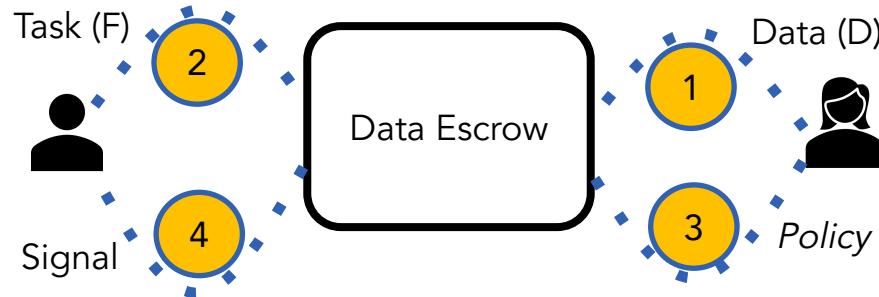
Data Escrow [VLDB'22]

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



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

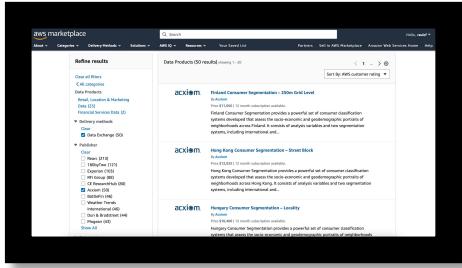


- Guarantee: no data* leaves the escrow without explicit permission, i.e., without an explicit *policy*

Using the Escrow to *Signal* Dataflow results



Seller



Buyer



With my data,
Accuracy: 0.63

Using the Escrow to *Signal* Dataflow results



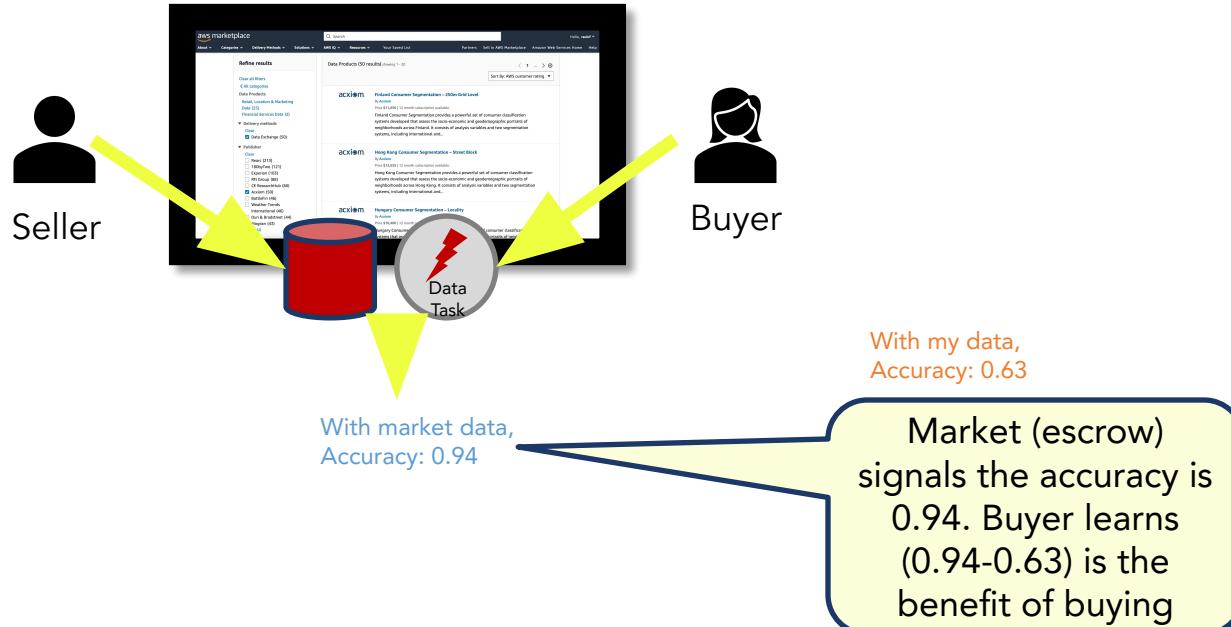
With my data,
Accuracy: 0.63

Using the Escrow to *Signal* Dataflow results



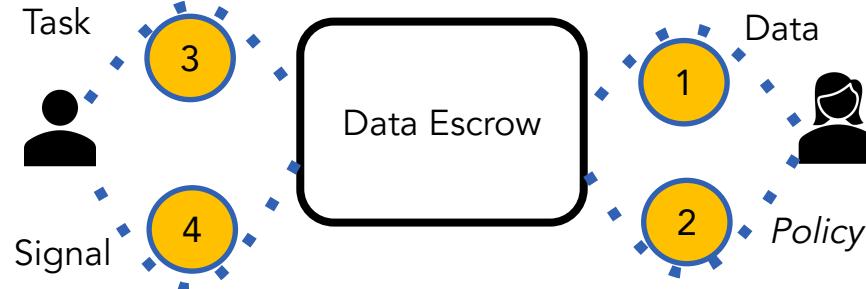
Using the Escrow to *Signal* Dataflow results

Data Markets



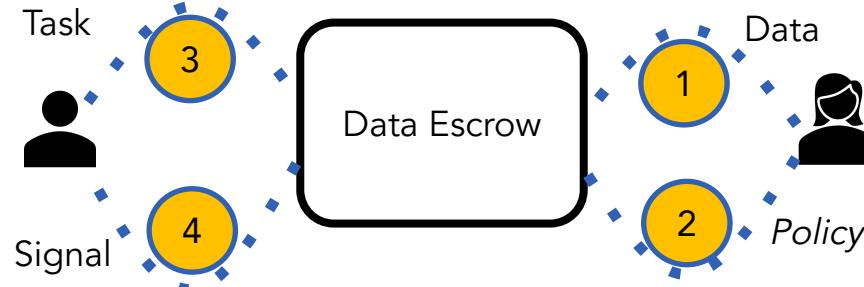
How do we delegate tasks, create signals,
i.e., how do we control dataflows?

Programmable Dataflows



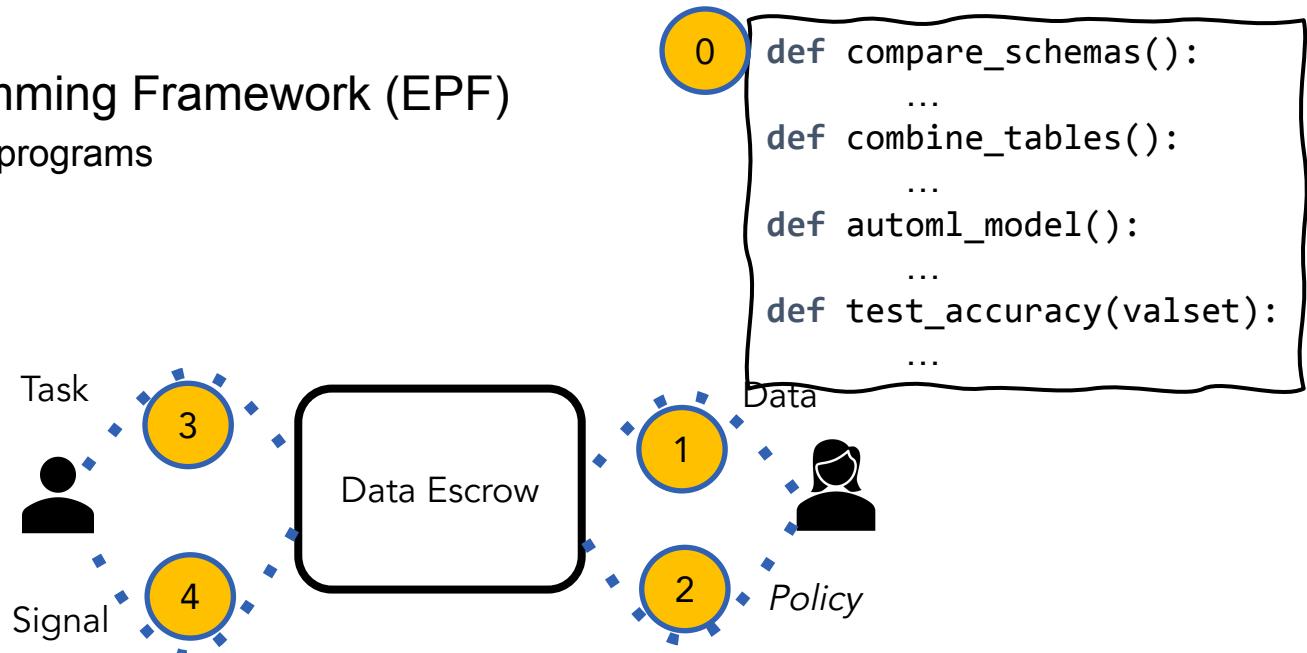
Programmable Dataflows

- Escrow Programming Framework (EPF)



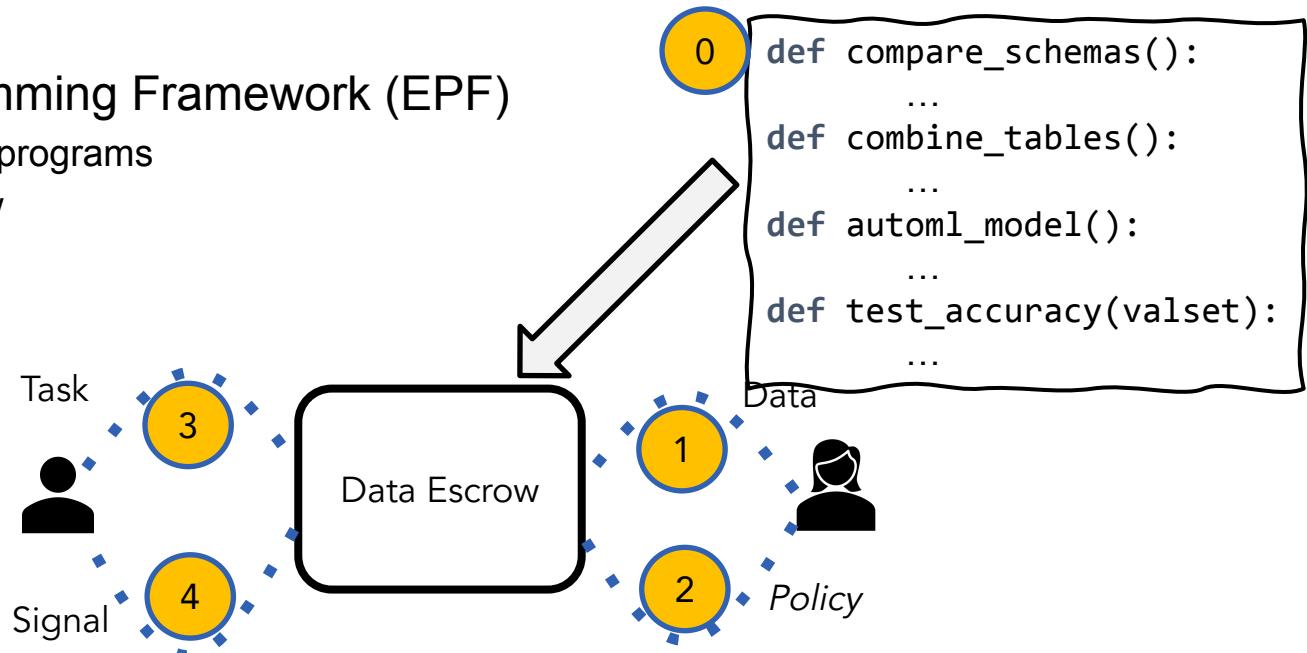
Programmable Dataflows

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



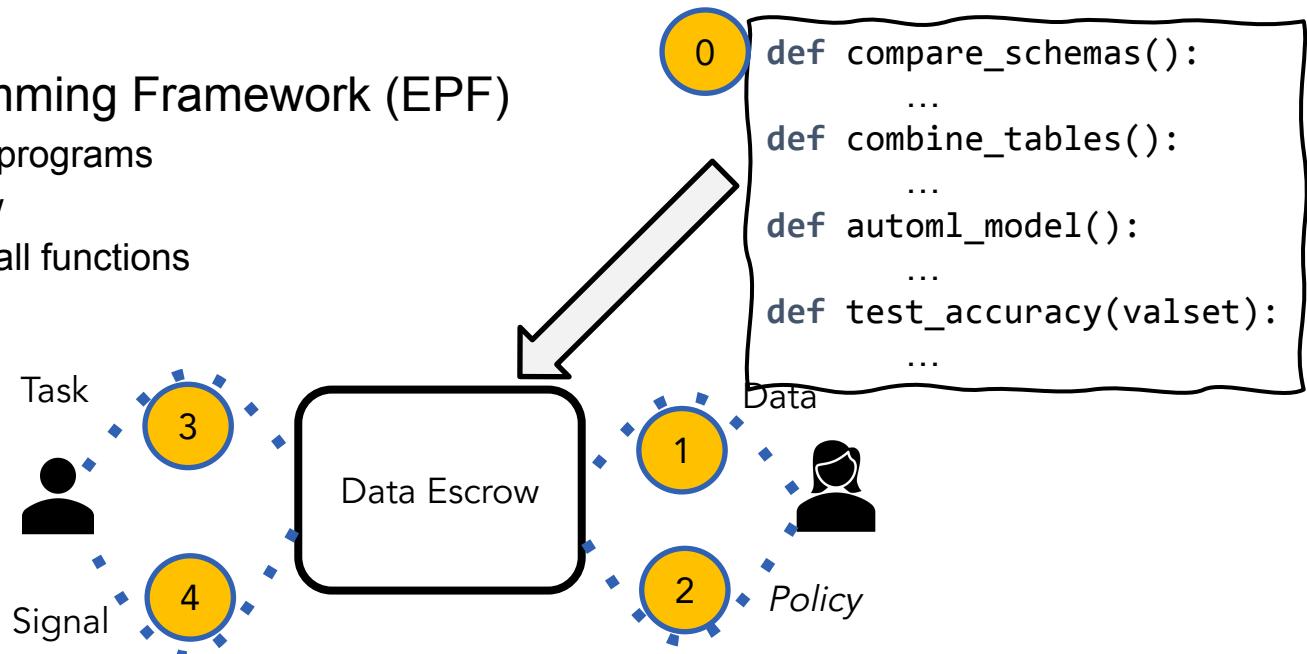
Programmable Dataflows

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



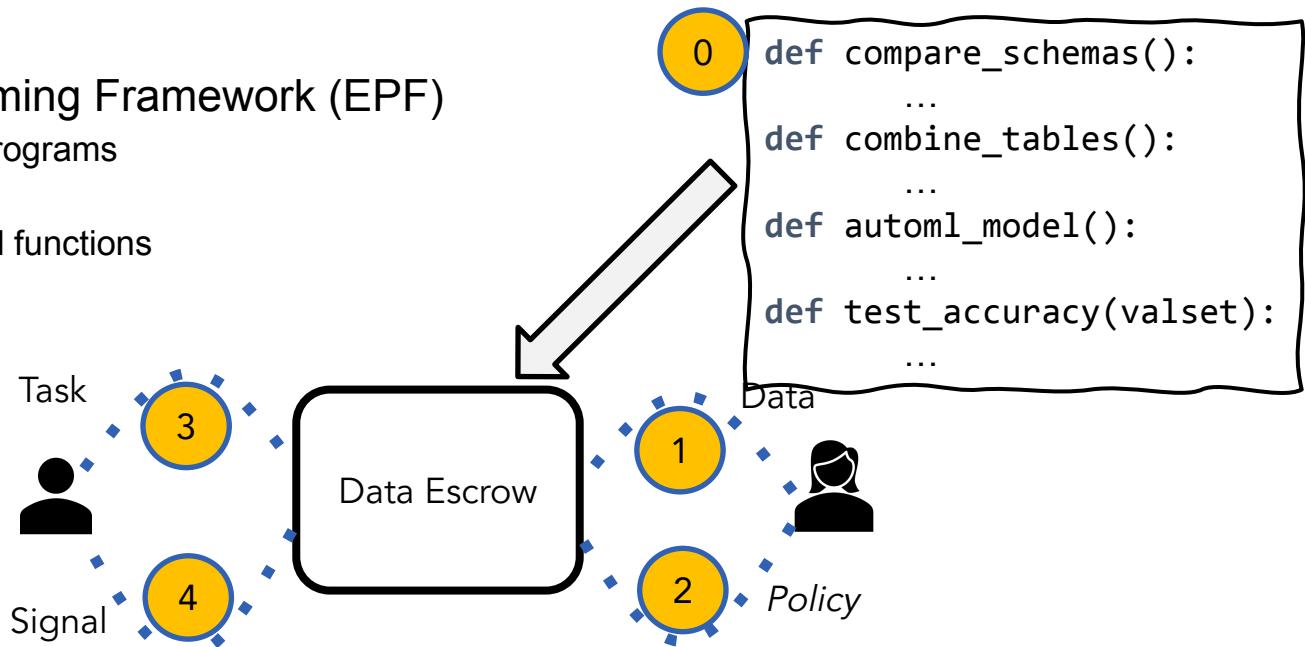
Programmable Dataflows

- 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

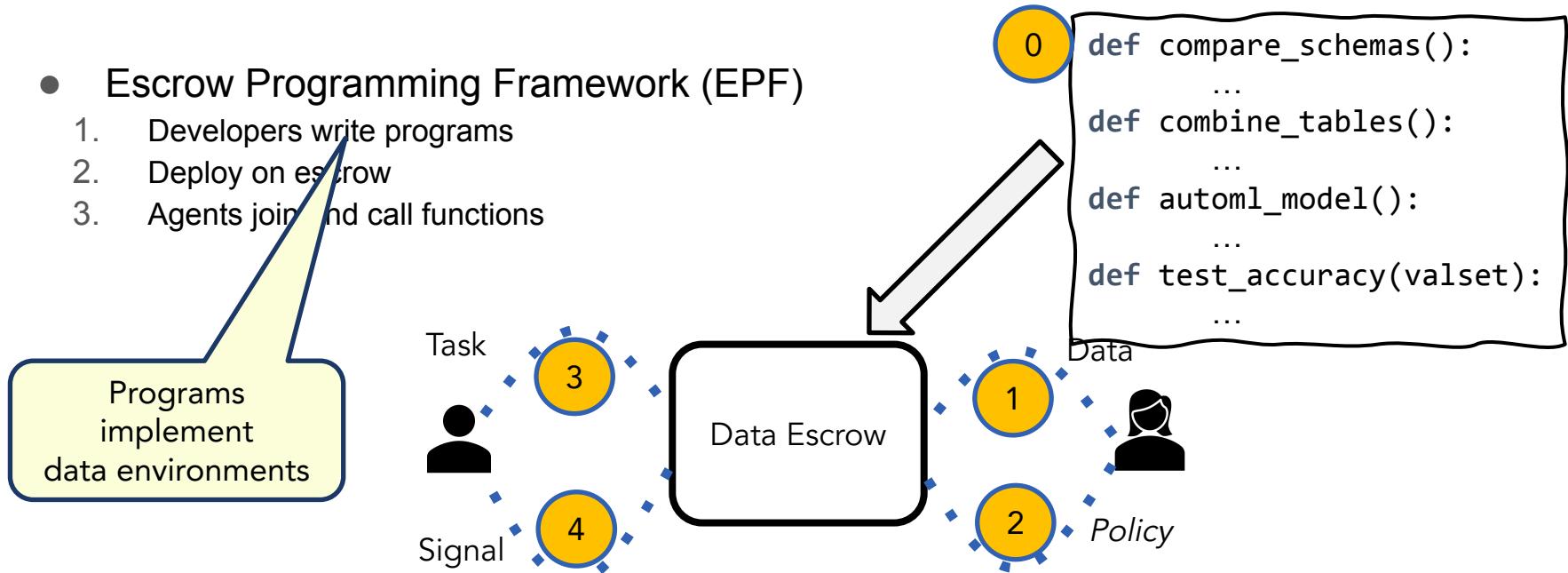


- Program implements communication and logic via *contracts*

Programmable Dataflows

- Escrow Programming Framework (EPF)

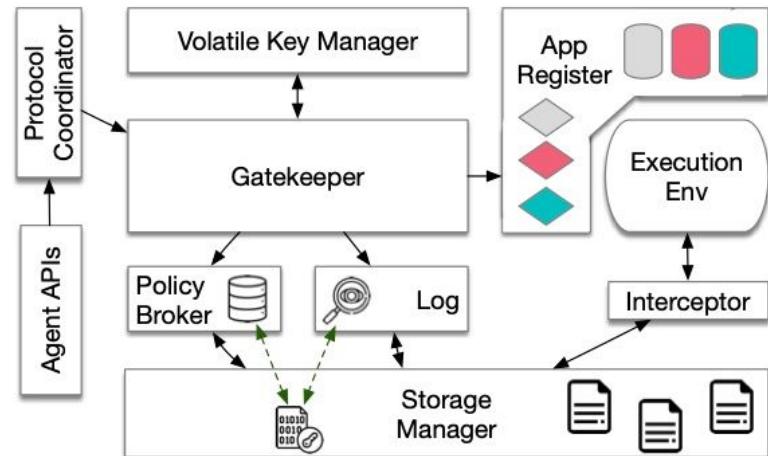
1. Developers write programs
2. Deploy on escrow
3. Agents join and call functions



- Program implements communication and logic via *contracts*

Delegated, Auditable, Trustworthy

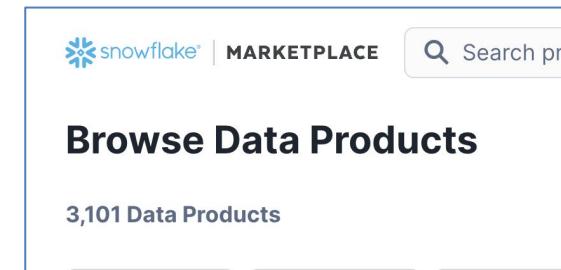
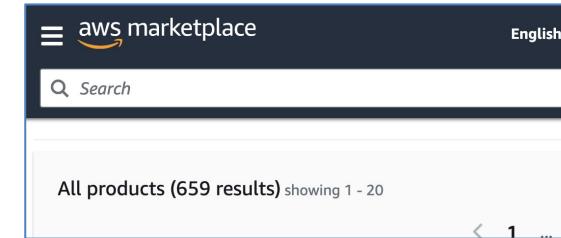
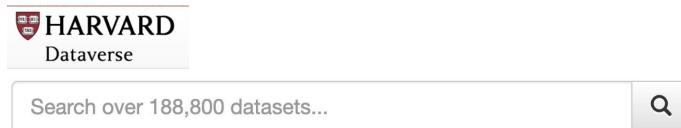
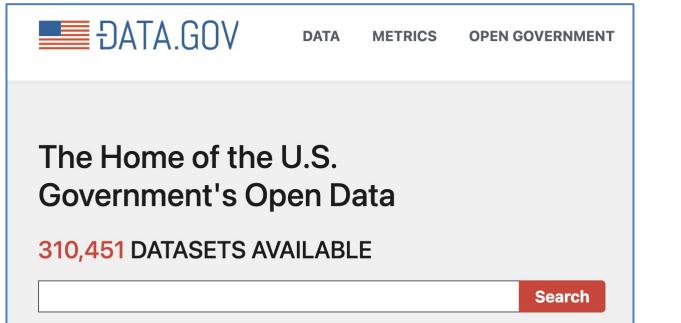
- What happens in the escrow, stays in the escrow
 - Except when it needs to be available to auditors and 3-party officers
- Data is encrypted end-to-end
 - At rest and during computation
 - Use of secure hardware enclaves
 - Encrypted Write-Ahead Log (EWAL)
 - Cryptographic protocols for IO
 - Key exchange and recovery after failures...



Data Search

Unlimited Storage → Massive Data Repos

Gov Portals
Data Markets
Data Lakes
Web Tables
Data coalitions
...

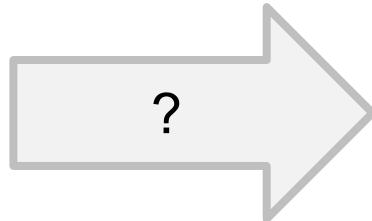


Google
Powered by Dataset Se

[Dataset Search](#), a dedicated search engine for data indexes more than 45 million datasets from more than cover many disciplines and topics, including govern

What Can We Do with 1M+ Tables?

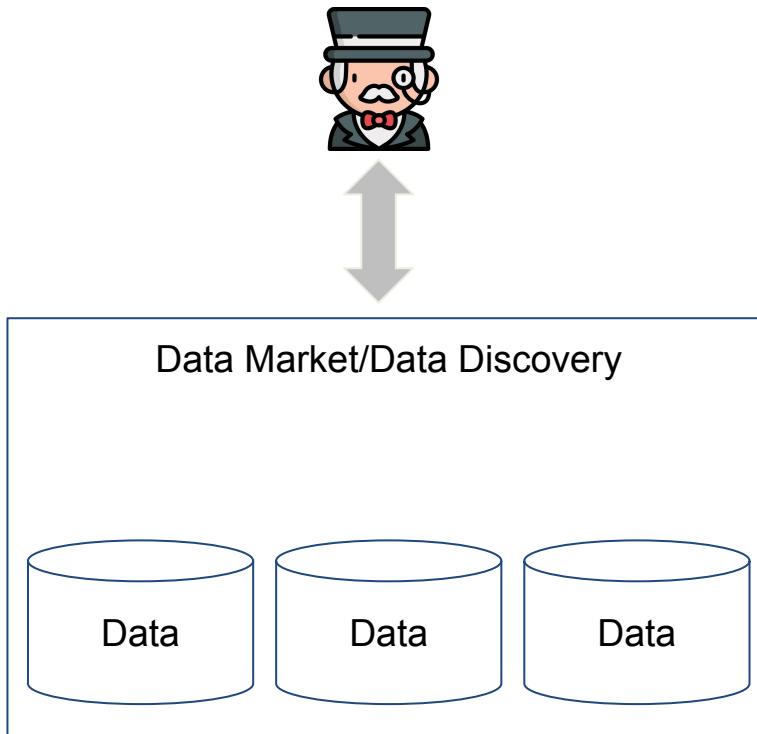
Gov Portals
Data Markets
Data Lakes
Web Tables
Data coalitions
...



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

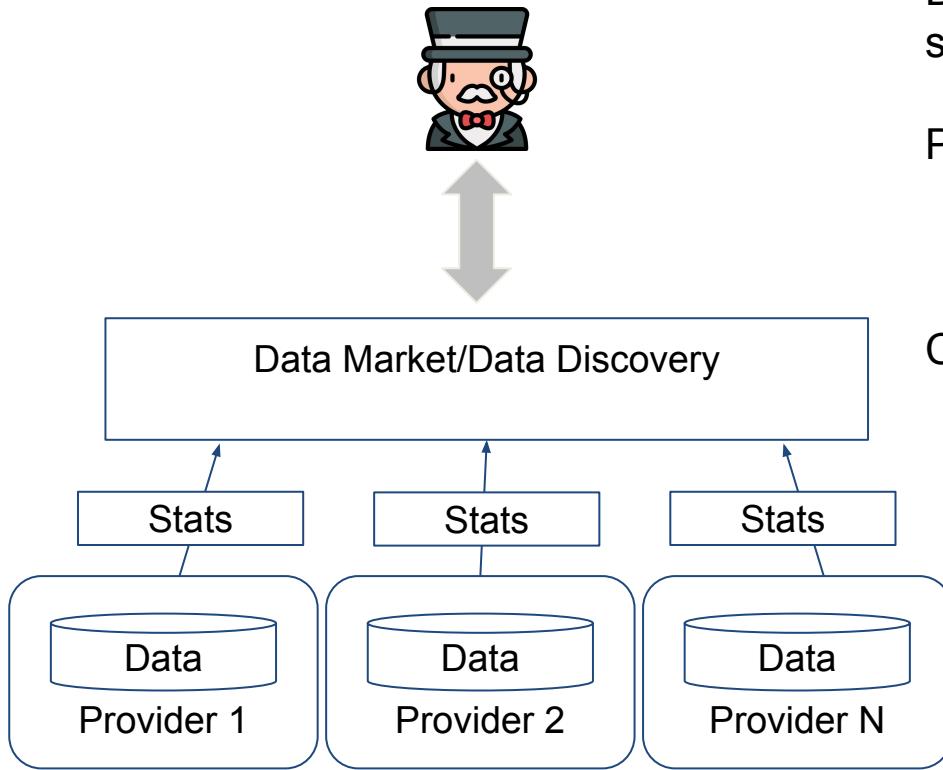
Pros:

- 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

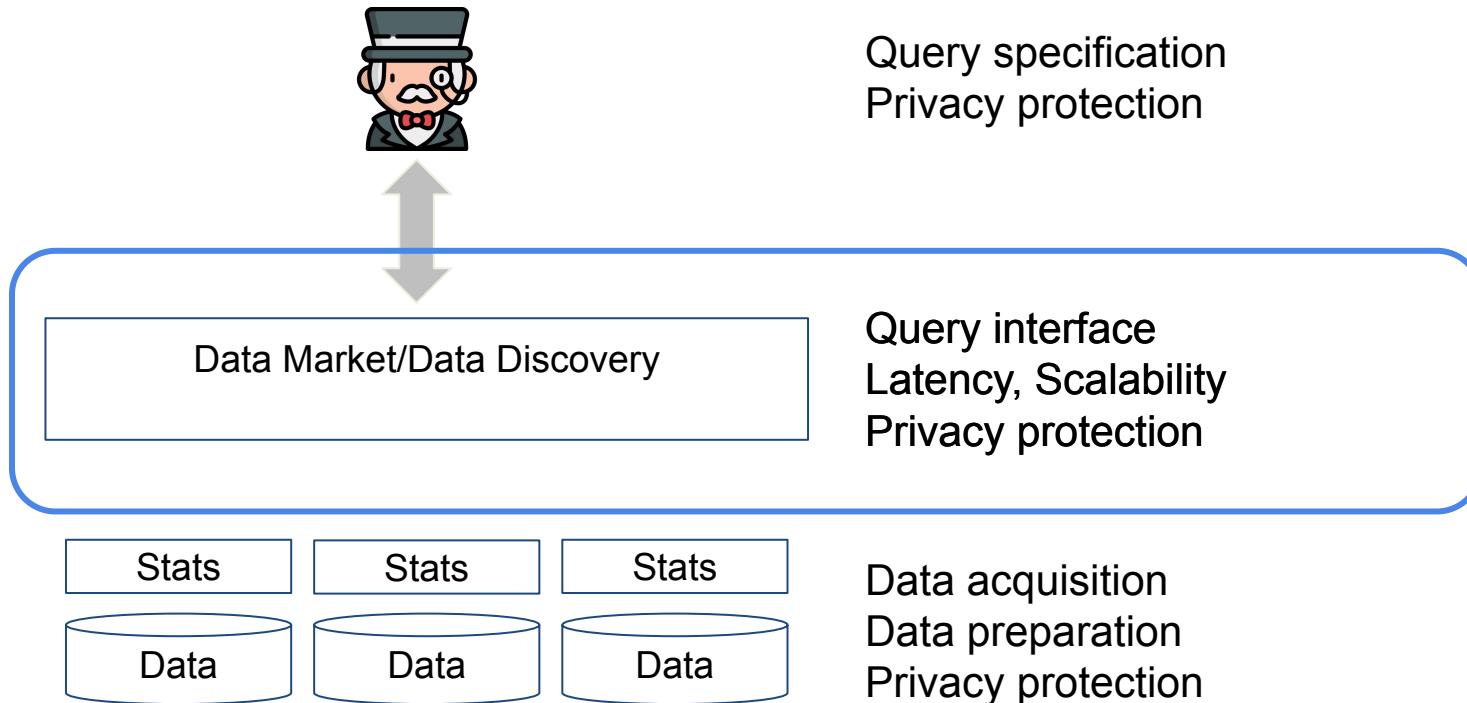
Pros

- 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



data.gouv.fr

[DATA.GOV](https://www.data.gov/)
Dataset Search

[NYC OpenData](https://opendata.cityofnewyork.us/)

[City of London Open Data](#)

The screenshot shows the Snowflake Marketplace interface. At the top, there's a search bar with the query "predict nyc education test scores". Below the search bar, it says "624 Data Products". There are several filter buttons: Availability, Categories, Business Needs, Geo, Time, Price, and More Filters. The results list includes:

- Truelty Identity Resolution** by Truelty: Your Data • Your Customer • Your Snowflake
- InNote** by Innovaccer Inc.: Physician's digital assistant that surfaces health insights at point of care.
- Free Sample: Cross Shopping Insights - NYC Restaurants** by SafeGraph: Geographic patterns and brand affinities in consumer spending
- Test Automation for Snowflake** by NTT DATA: Automate the data quality monitoring
- Area store visits data | Visits to shoe stores in NYC in 2022 | Free sample** by Olvin: Historical visit data to shoe stores in New York City, 2022.

Keyword Search as Sensemaking

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

The screenshot illustrates the DataScout platform for keyword search as sensemaking, featuring three main panels:

- Query Decomposition (Panel A):** Shows a task specification: "Evaluate the effects of remote work on quality of life through various periods of the pandemic". Below it are suggestions to refine the search query:
 - Assess remote work's impact on life satisfaction during the pandemic
 - Assess remote work's influence on employment quality during the pandemic
 - Assess the impact of remote work preferences on post-pandemic quality of life
- Top Dataset Results (Panel G):** Displays a list of 8 datasets, with the first one highlighted:
 1. Global Remote Work & Wellbeing Dataset.
 2. Remote Work USA (COVID-19)
 3. Remote Work Productivity
 4. COVID-19 on Working Professionals
 5. Impact of COVID-19 on Working Professionals
 6. World time use, work hours and GDP
 7. Impact of Covid-19 on Employment - ILOSTAT
 8. Annual Working Hours Dataset (1870-1970)
- Global Remote Work & Wellbeing Dataset (Panel H):** Detailed view of the selected dataset.
 - Dataset Summary:** global_remote_work_wellbeing.csv, Usability score: 100%, 10 cols, 10000 rows, 133.3 kB, 179 global, business, jobs and career, employment, Day-Level Granularity.
 - Relevance:** Why is this dataset relevant for your task? The dataset includes attributes like Daily_Working_Hours, Stress_Level, Sleep_Duration, and Work_Life_Balance_Satisfaction.
 - Limitation:** The dataset does not specify a time period or geographical location.
 - Description:** The Global Remote Work & Wellbeing Dataset is a comprehensive synthetic dataset designed to capture the multifaceted impacts of remote work on employee productivity, mental health, and work-life balance. It includes anonymized data from various sources to provide insights into daily work experiences and lifestyle patterns in remote work environments.
 - Dataset Preview:** A table showing sample data for the first 10 rows across columns: Employee_ID, Daily_Working_Hours, Screen_Time, Meetings_Attended, Emails_Sent, Productivity_Score, Stress_Level, Physical_Activity_Steps, Sleep_Duration, and Work_Life_Balance_Satisfaction.

Annotations with letters A-I highlight specific features:

- A:** Task specification and query refinement suggestions.
- B:** Suggestion to assess remote work's impact on life satisfaction.
- C:** Column group filter for 'include hours'.
- D:** Search bar for column names.
- E:** Smart Filter by Column Concept dropdown menu.
- F:** Placeholder for 'Remaining datasets'.
- G:** Top Dataset Results panel.
- H:** Global Remote Work & Wellbeing Dataset panel.
- I:** Limitation and Description sections of the dataset panel.

Keyword Search as Sensemaking

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

A screenshot of the DataScout 'Getting started?' interface. The page has a light blue header bar with the title 'Getting started?'. Below it is a white form area. At the top left of the form is a yellow circular button containing the letter 'A'. A thick orange arrow originates from the bottom right of this button and points down towards the bottom right corner of the entire screenshot frame. The form contains the following sections:

- 1. Do you have a specific task in mind, or are you exploring available options?**
Two buttons: 'I have a specific task' (blue) and 'I am exploring' (white).
- What is the primary goal of your task?**
A row of buttons: 'Train a classifier', 'Train a regression model' (highlighted in blue), 'Supervised learning', 'Unsupervised learning', 'Visualization', 'LLM pretraining', 'LLM finetuning', 'Question-Answering', and 'Not sure yet'.
- 2. What do you specifically want to do? Provide keywords or a sentence on the task you're interested in.**
A text input field containing the text 'datasets indicating quality of life before, during, and after the COVID-19 pandemic'.
A blue 'Get Started' button at the bottom.

Keyword Search as Sensemaking

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

The screenshot shows the DataScout interface. At the top, there's a search bar with the placeholder "Search using your own Column Concept". Below it is a "Task Specifications" section containing a pink box with the text "Analyze the impact of the pandemic on remote work and work-life balance". To the right of this is a "Suggestions to Refine your Search Query:" list with three items: "Assess remote work's impact on life satisfaction during the pandemic", "Assess remote work's influence on employment quality during the pandemic", and "Assess the impact of remote work preferences on post-pandemic quality of life". Below these are sections for "Filters (0)" and "Remaining datasets: 8". A blue button labeled "apply filters" is visible. To the right of the search area is a sidebar titled "Top Dataset Results" showing 11 datasets. The first dataset is "Global Remote Work & Wellbeing Dataset" (10 columns, 10000 rows, 133.3 kB). Other datasets listed include "Remote Work USA (COVID-19)", "Remote Work Productivity", "COVID-19 on Working Professionals", "Impact of COVID-19 on Working Professionals", "Online Learning Data", "Predict if people prefer WFH vs WFO post Covid-19", "Global Unemployment Dataset", "World time use, work hours and GDP", "Impact of Covid-19 on Employment - ILOSTAT", and "Annual Working Hours Dataset (1870-1970)". Orange arrows point from the text "C" and "D" to the "Remaining datasets" and "apply filters" buttons respectively.

Dataset Search Query

Task Specifications

Analyze the impact of the pandemic on remote work and work-life balance

Suggestions to Refine your Search Query:

- Assess remote work's impact on life satisfaction during the pandemic
- Assess remote work's influence on employment quality during the pandemic
- Assess the impact of remote work preferences on post-pandemic quality of life

Filters (0)

No metadata filters added.

Search using your own Column Concept

Enter column name...

Smart Filter by Column Concept:

- stress
- hours
- vacations
- employment
- remote

Remaining datasets: 8

apply filters

Top Granularity Filters

- Country (5)
- Day (2)
- Year (2)

apply filters

Top Dataset Results

Showing 1 to 11 of 11 datasets.

1. Global Remote Work & Wellbeing Dataset.
10 cols - 10000 rows - 133.3 kB - 179
2. Remote Work USA (COVID-19)
24 cols - 3147 rows - 2.0 MB - 81
3. Remote Work Productivity
5 cols - 1000 rows - 6.7 kB - 3.3k
4. COVID-19 on Working Professionals
15 cols - 10000 rows - 255.2 kB - 2.2k
5. Impact of COVID-19 on Working Professionals
15 cols - 10000 rows - 239.5 kB - 3.1k
6. Online Learning Data
21 cols - 214 rows - 4.6 kB - 184
7. Predict if people prefer WFH vs WFO post Covid-19
19 cols - 207 rows - 2.9 kB - 1.5k
8. Global Unemployment Dataset
7 cols - 50 rows - 70.3 kB - 78
9. World time use, work hours and GDP
7 cols - 329 rows - 207.6 kB - 734
10. Impact of Covid-19 on Employment - ILOSTAT
9 cols - 283 rows - 11.1 kB - 2.6k
11. Annual Working Hours Dataset (1870-1970)
4 cols - 3470 rows - 27.1 kB - 476

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

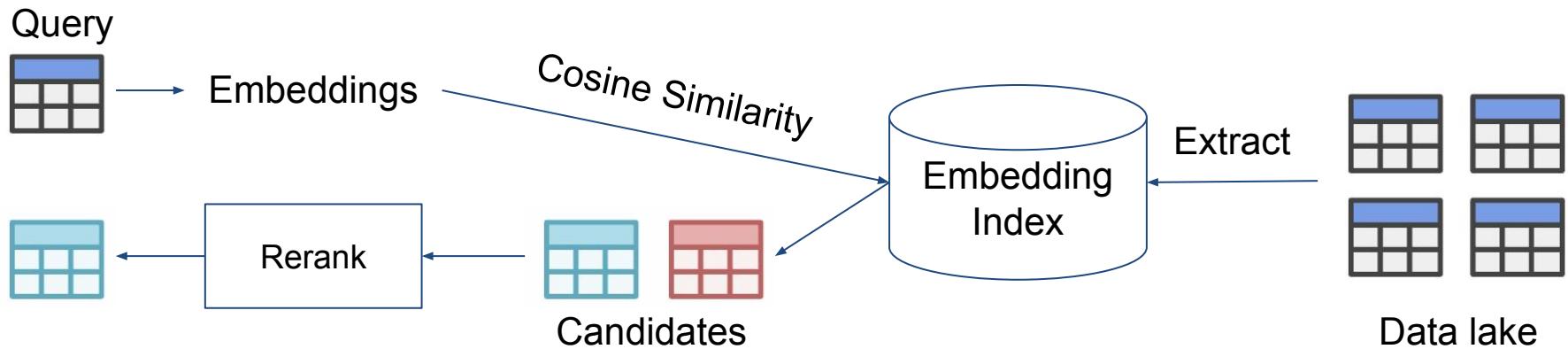
Search by using a table or distribution as the query

Ling13,Zhu16,Nargesian18,Fernandez19,Rezig22,Santos21,Fan23,...

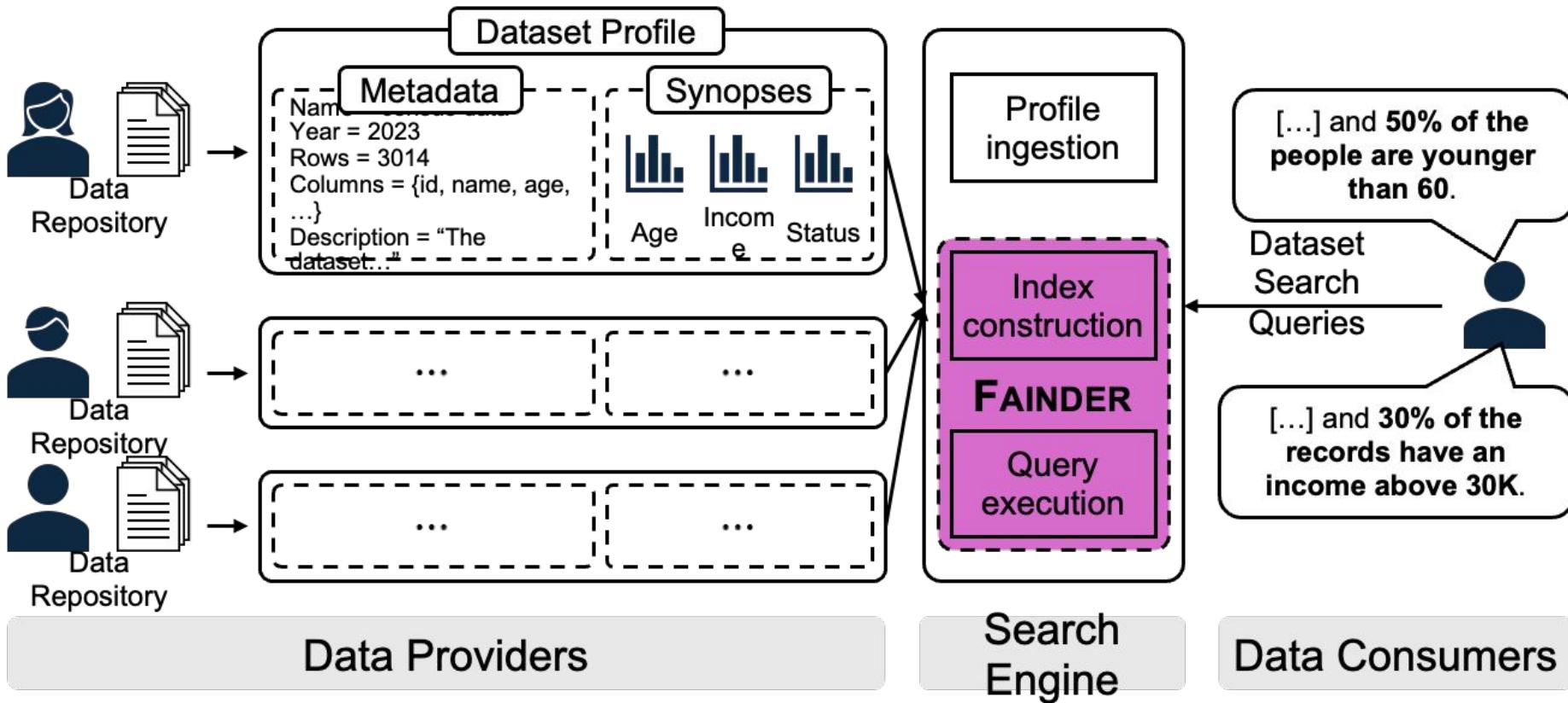
Rank based on

- 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 $T(D) \rightarrow$ goodness is function of table D

Prediction ARDA, AUCTUS, Galhotra23

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

Causal Inference Suna Liu25, MetaM Galhotra23

- $T(D)$: 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

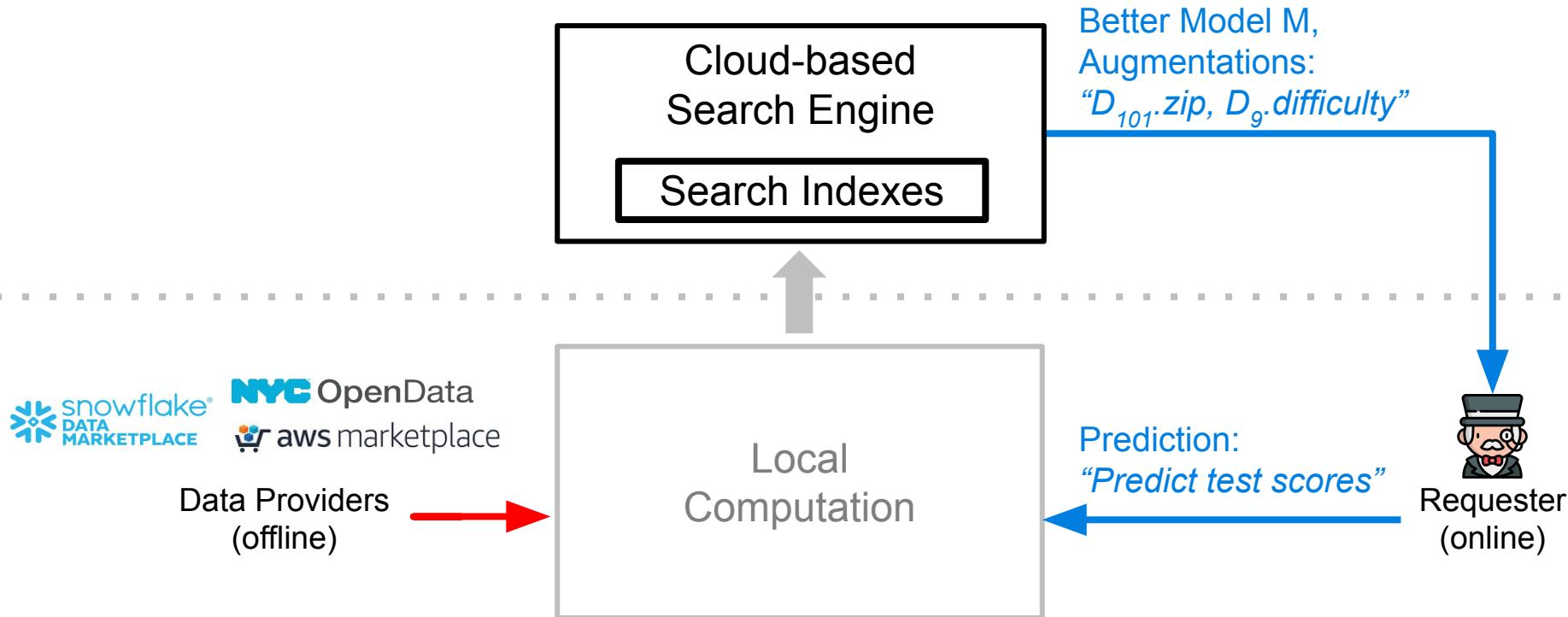
Kitana: A Data-as-a-Service Platform. Zach Huang²³

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

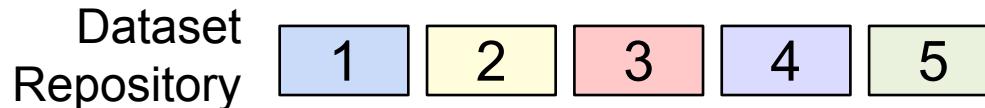
Data Task as Search Query



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    Combinatorial  
    eval(apply A to D)  
return best A
```

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  $\alpha$  Greedy Search
    if eval(apply A to D) is best so far
        Keep  $\alpha$  in A      Expensive!
        Materialize A(D)
        Retrain & Cross-validate
return best A
```

Ideas

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

Sketches: Count($D \bowtie S$)

Naïve join generates big intermediate relation

D

A	B
1	1
1	2
2	3



S

A	C
1	4
1	5
2	6

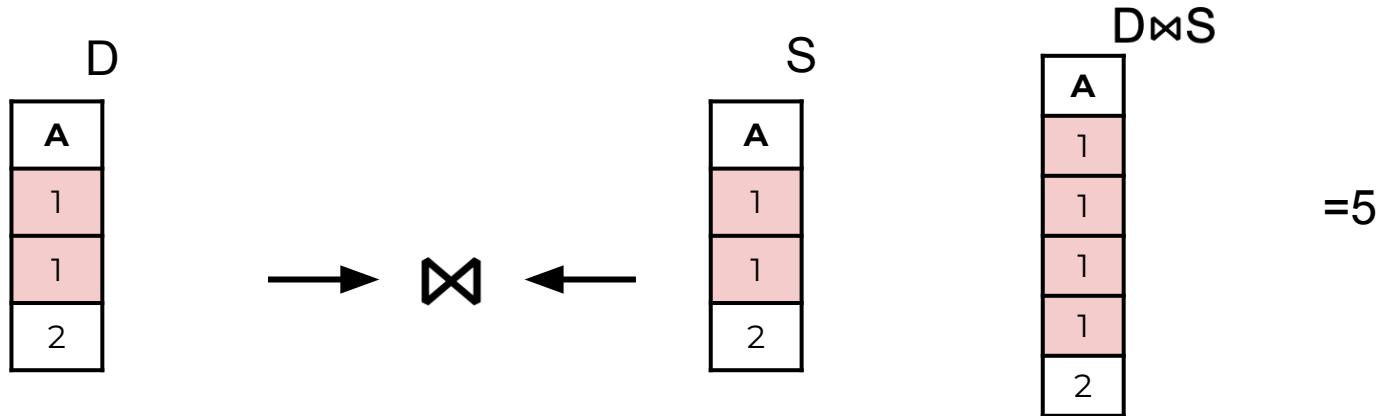
$D \bowtie S$

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

=5

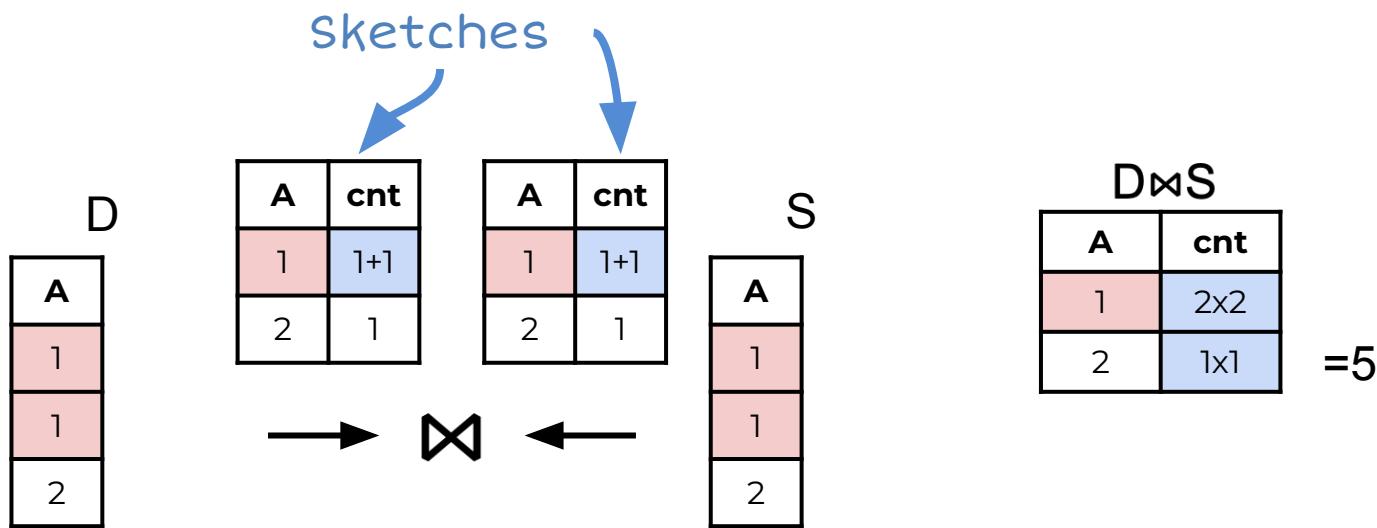
Sketches: Count($D \bowtie S$)

Optimization: drop irrelevant columns



Sketches: Count($D \bowtie S$)

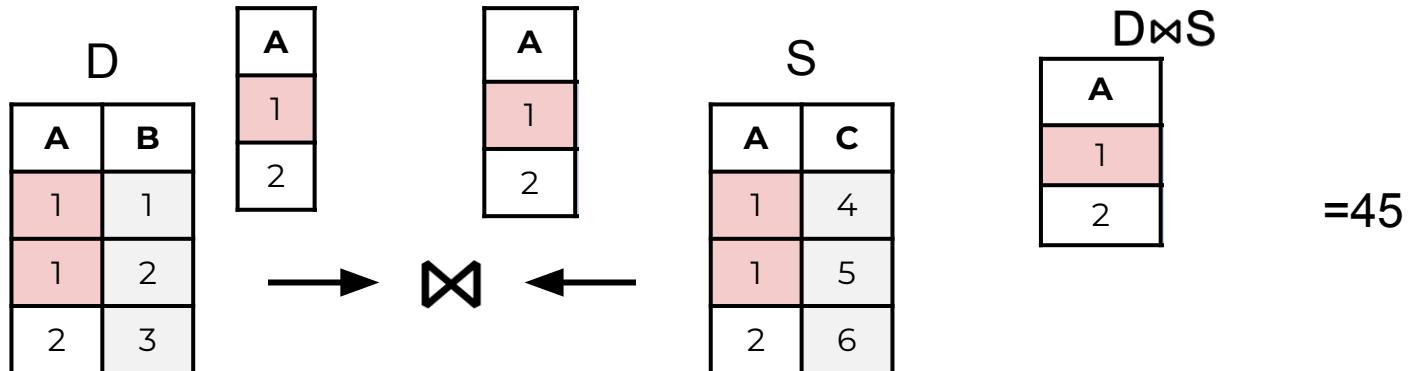
Optimization: sufficient statistics



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

Optimization: sufficient statistics

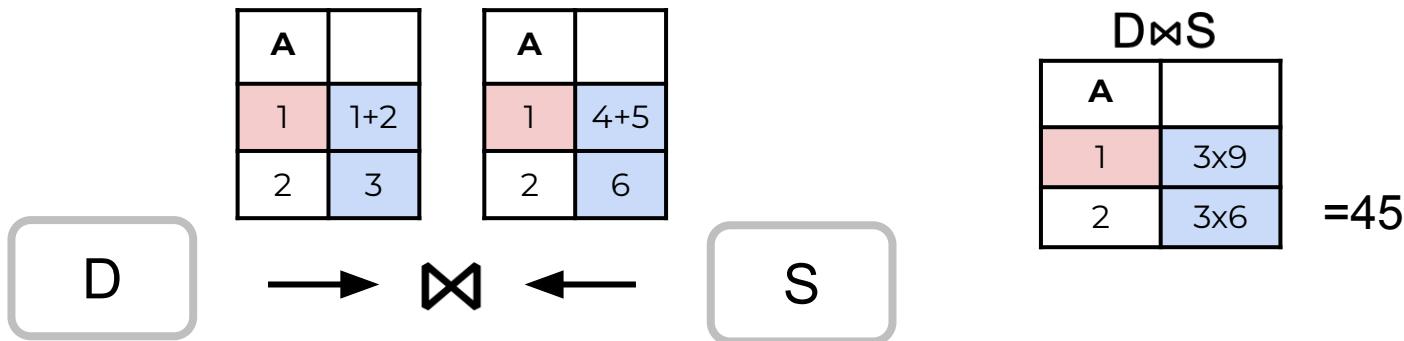
Sketches defined for common stats, ML models.



Sketches: trainAndEval($D \bowtie S$)

Optimization: sufficient statistics

Sketches defined for common stats, ML models.

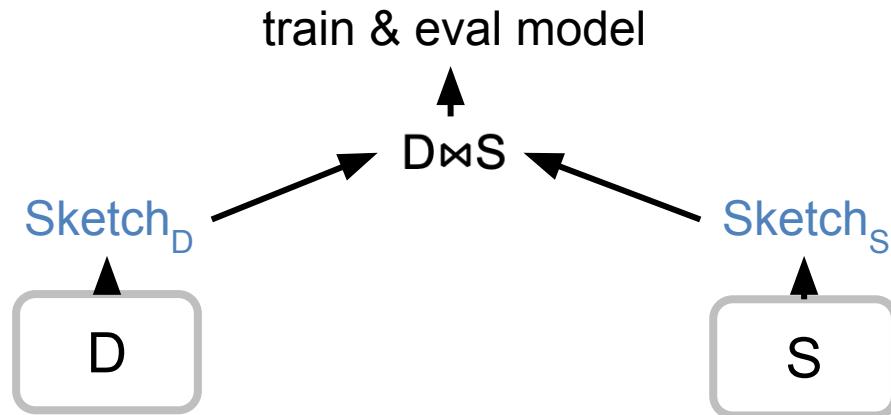


Sketches: train($D \bowtie S$)

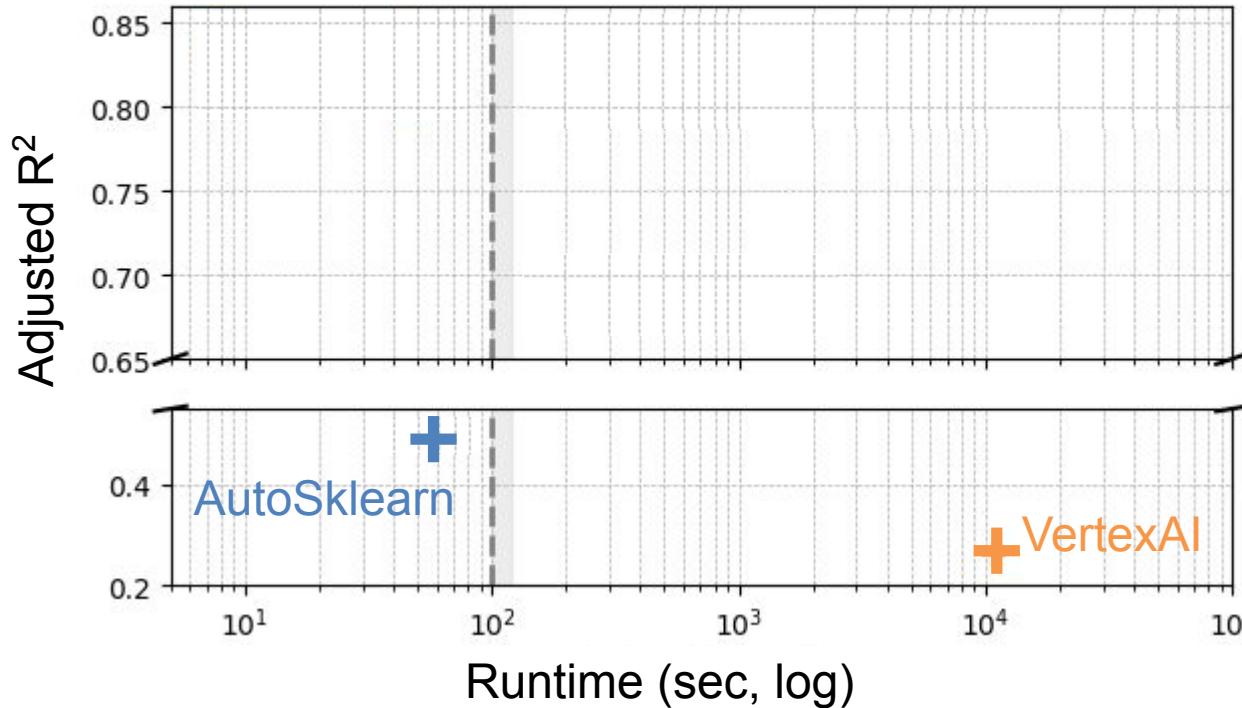
Optimization: sufficient statistics

Sketches defined for common stats, ML models.

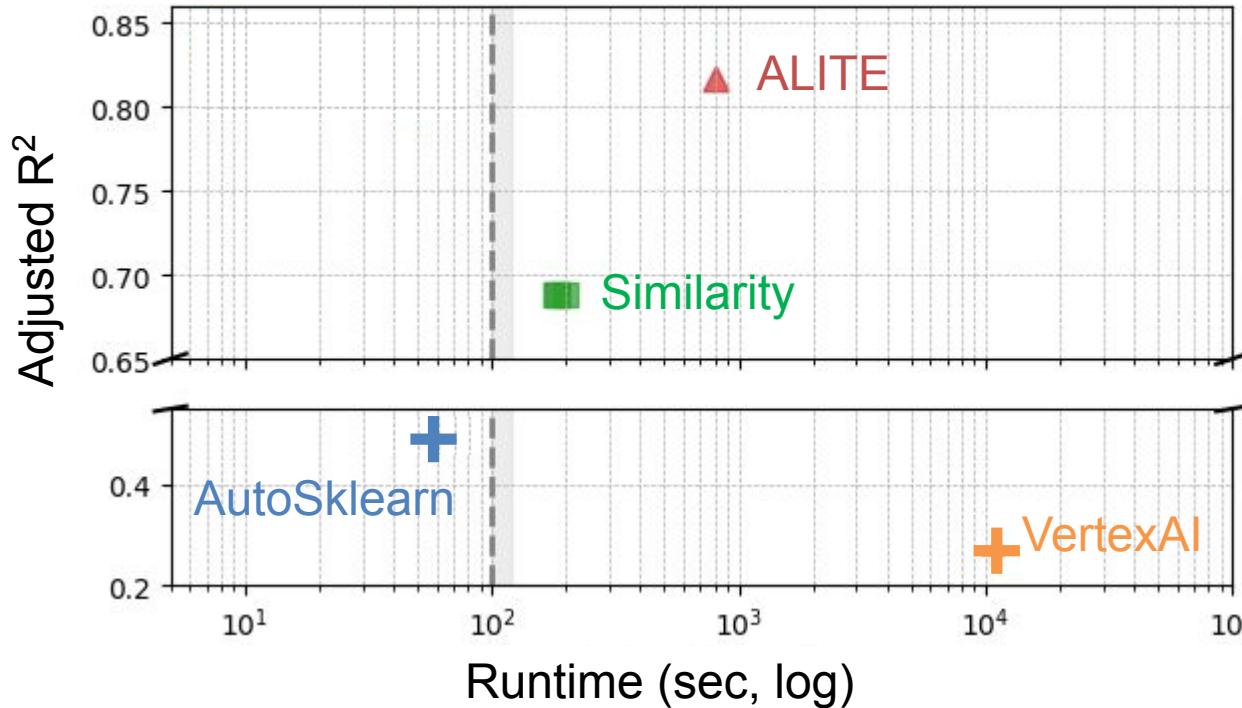
Linear Regression as a *proxy model* during search



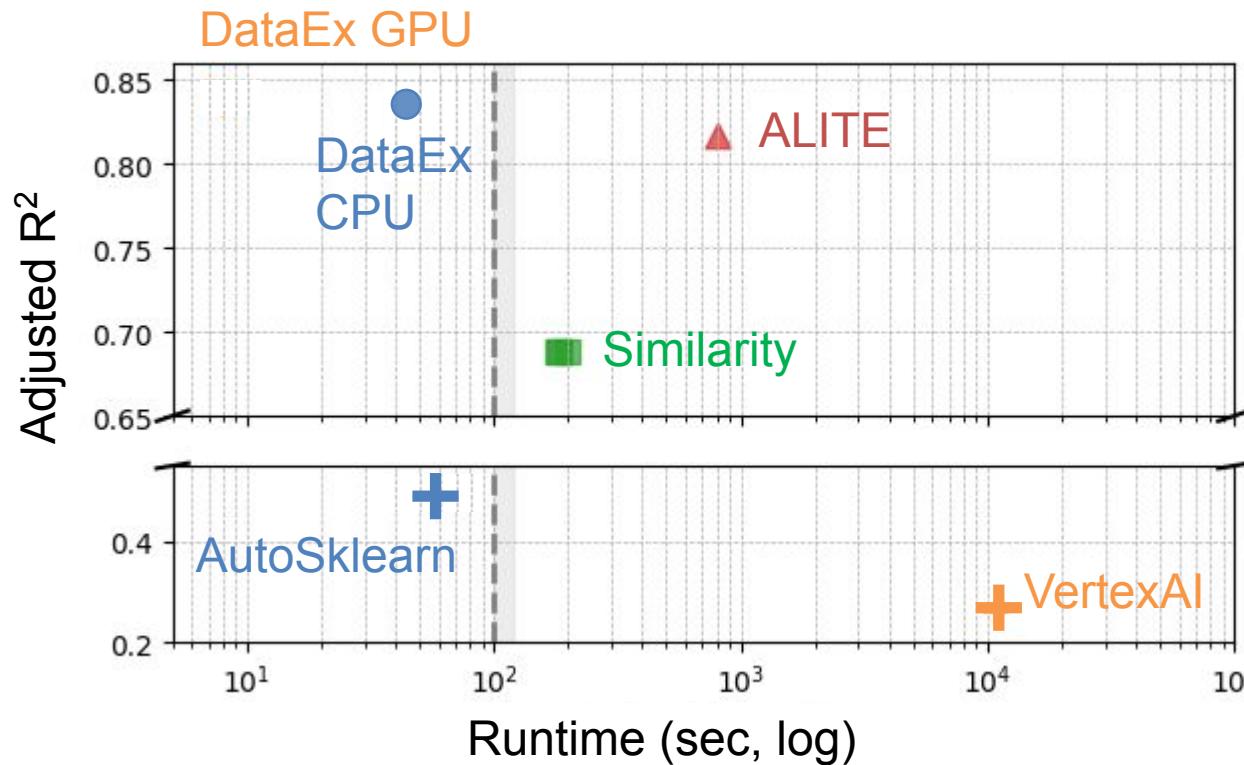
Evaluation on 8376 Kaggle Tables



Evaluation on 8376 Kaggle Tables

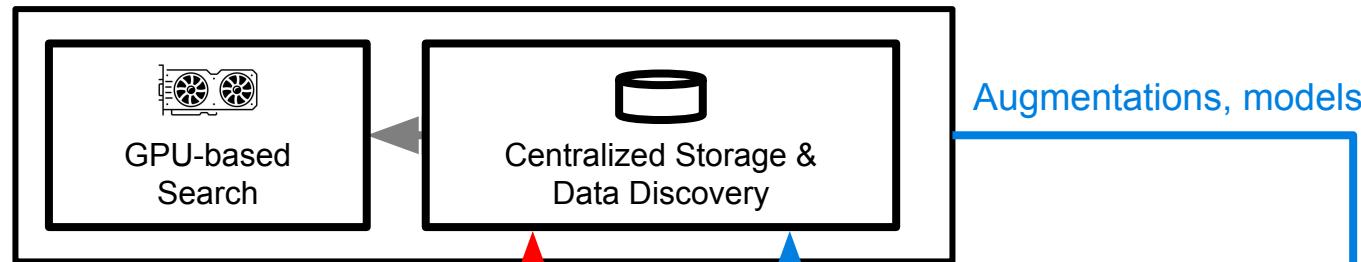


Evaluation on 8376 Kaggle Tables



DataEx

Cloud Dataset Search Engine



NYC OpenData

aws marketplace

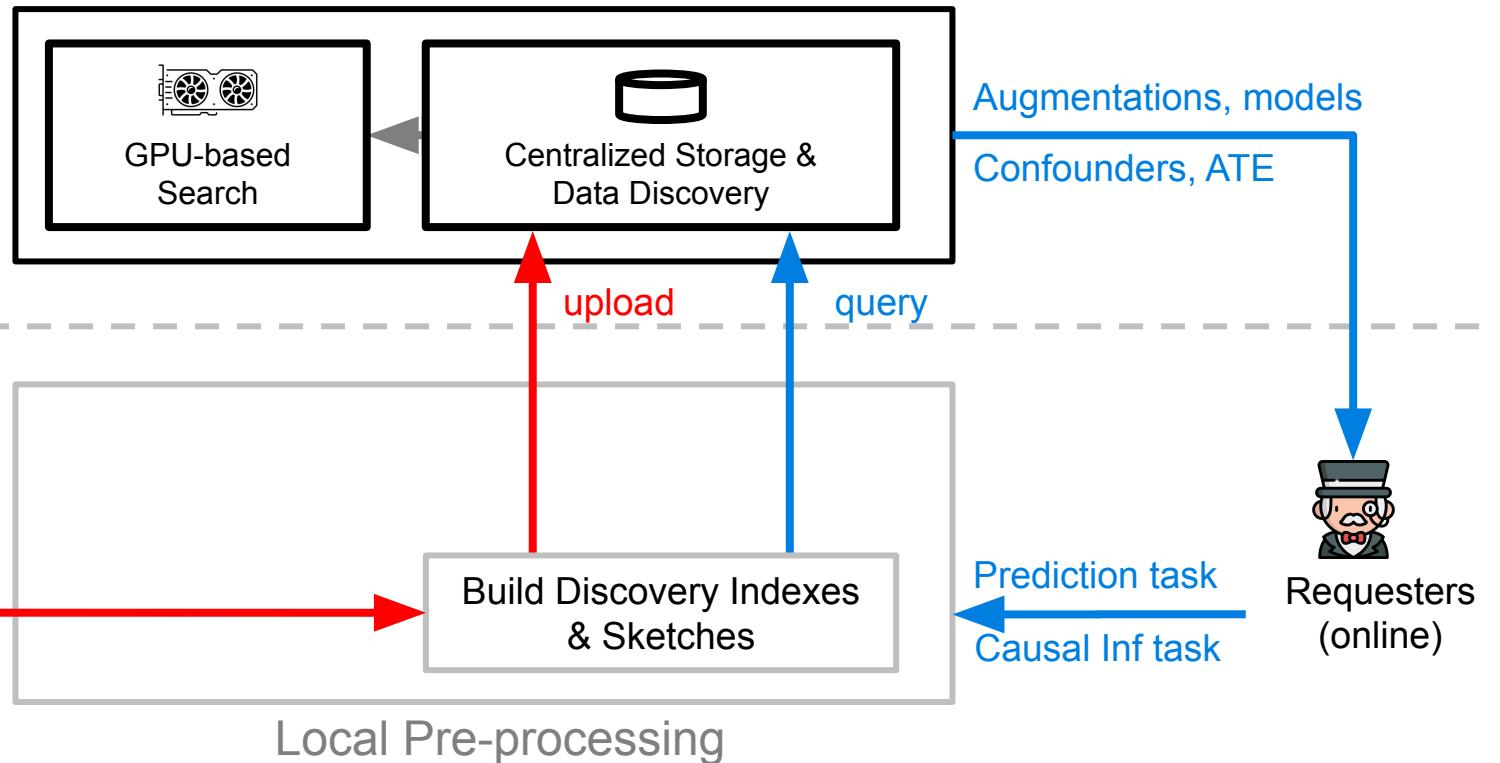
snowflake®
DATA
MARKETPLACE

Data
Providers
(offline)

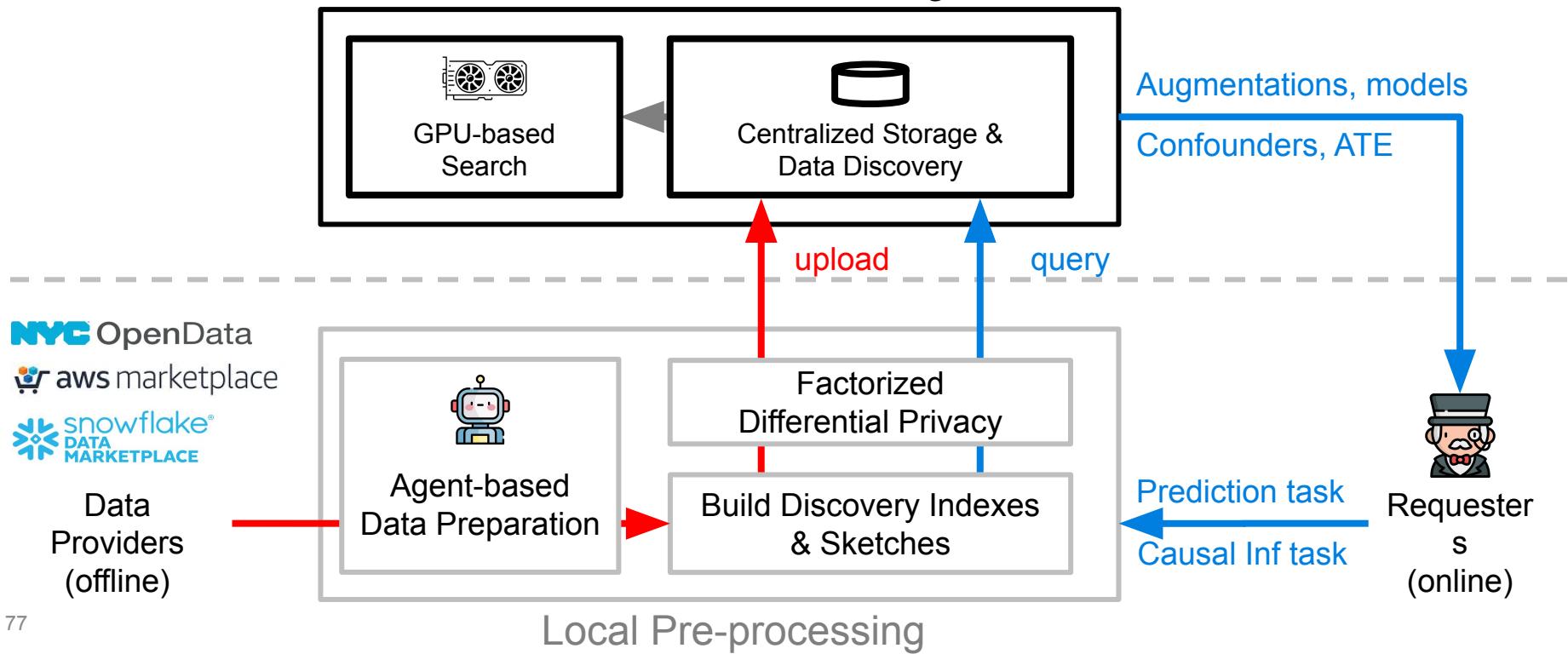
 Requesters
(online)

DataEx

Cloud Dataset Search Engine

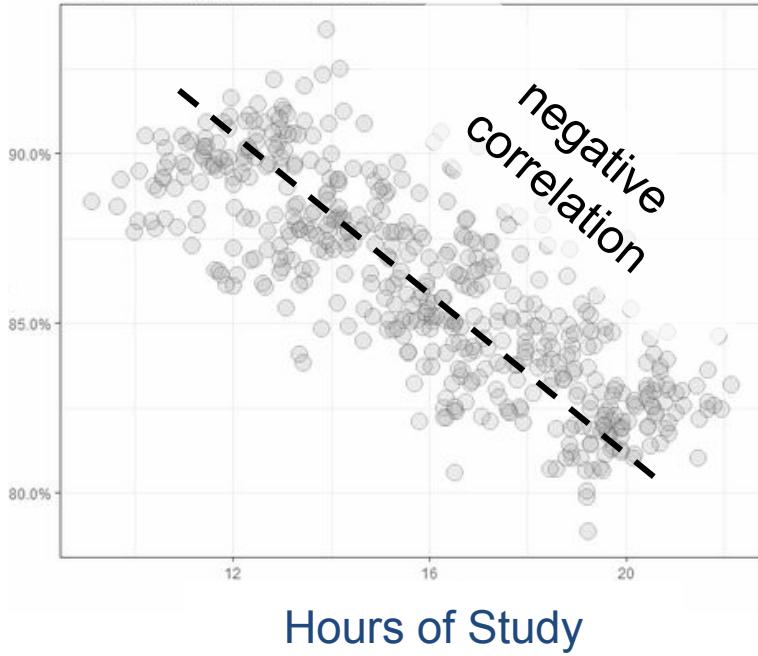


Cloud Dataset Search Engine

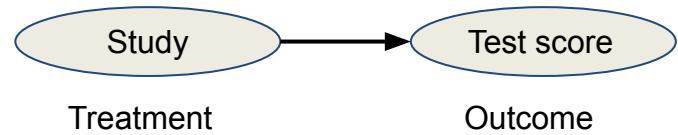


Confounders in Causal Analysis

Test Score



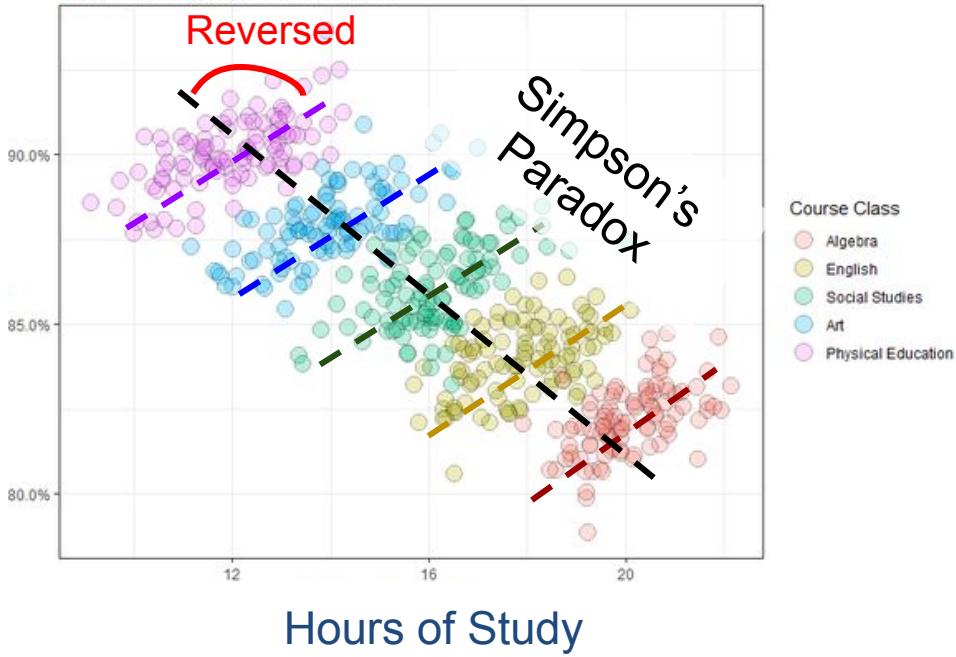
Causal Diagram



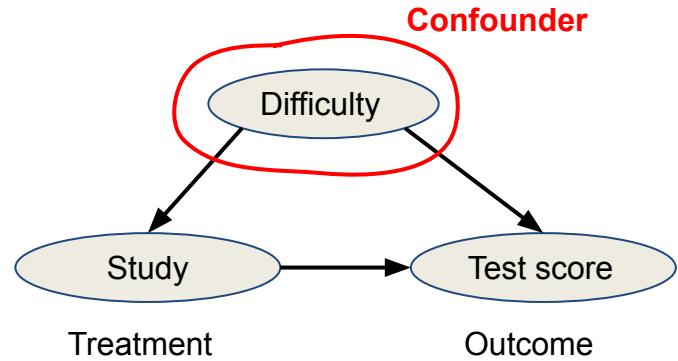
Studying causes poor grades?
 $\text{Study} \rightarrow \text{Test Scores}$

Confounders in Causal Analysis

Test Score



Causal Diagram



Study → Test Scores
Why does student grade?
Study → Test Scores

Confounders in Causal Analysis

User Query

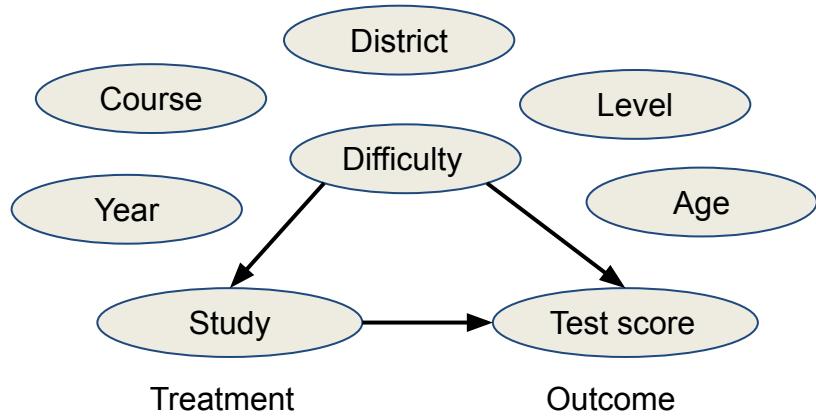
	Treatment	Outcome
ID	Study	Score
1	15 hr	75
2	10 hr	90
3	20 hr	85

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

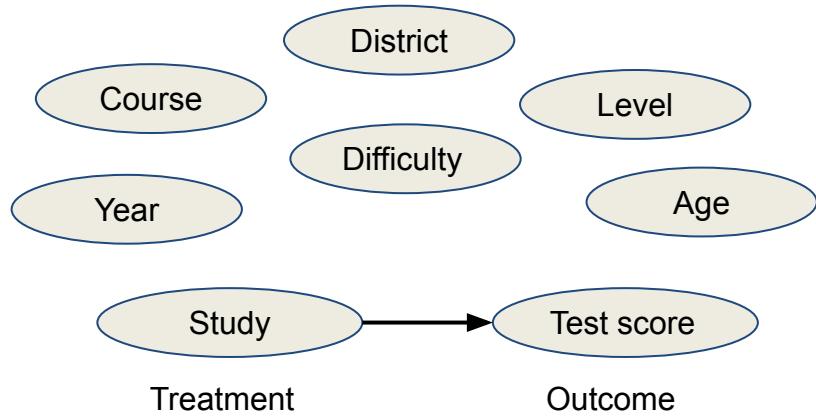
	Treatment	Outcome
ID	Study	Score
1	15 hr	75
2	10 hr	90
3	20 hr	85

Data Repository

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

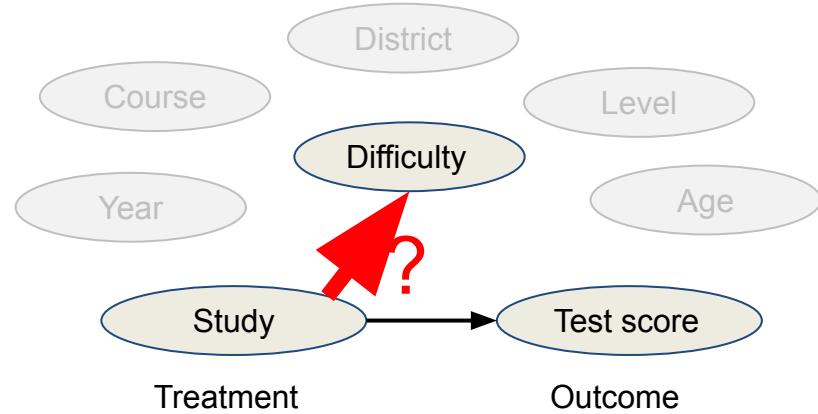
ID	District
1	1
2	2
3	3

...



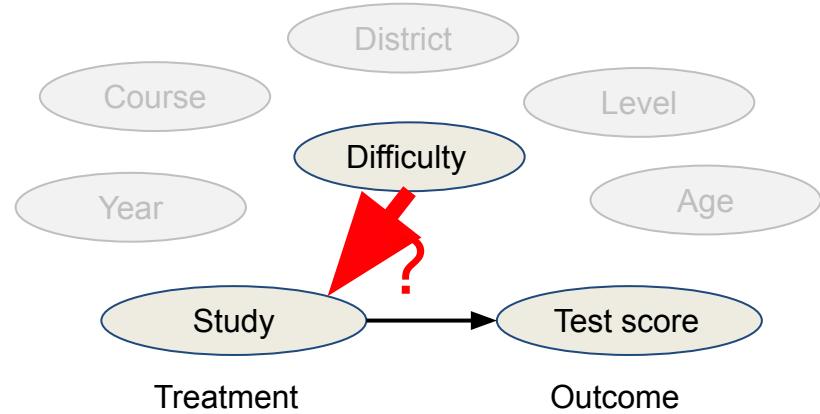
Confounders in Causal Analysis

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



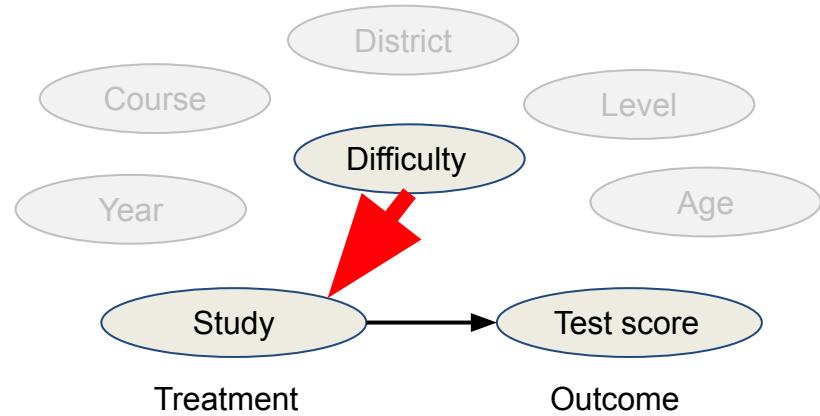
Confounders in Causal Analysis

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



Confounders in Causal Analysis

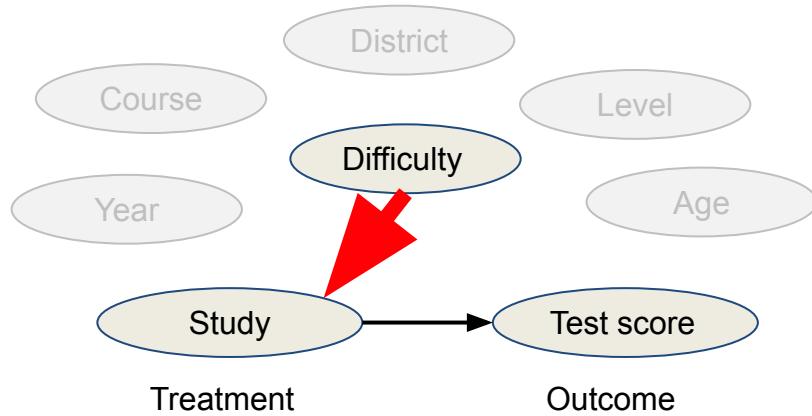
Background: Bivariate Causal Discovery (BCD) estimates → or ← edges from data



Confounders in Causal Analysis

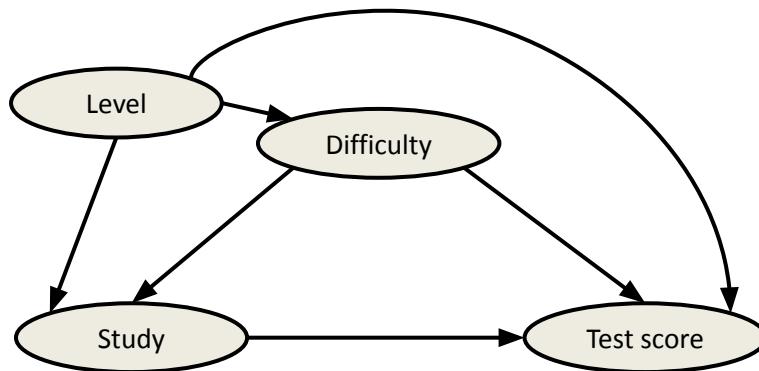
Background: Bivariate Causal Discovery (BCD) estimates → or ← edges from data

Proof: existence of confounder reduces to BCD estimating “*Ancestors ↵ Treatment*”



Discovering Confounders with BCD

Building adjustment set for Study → TestScore

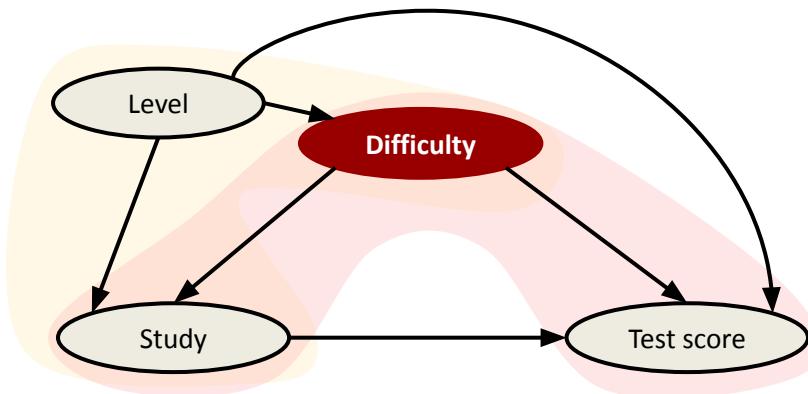


Key Observation:

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

Discovering Confounders with BCD

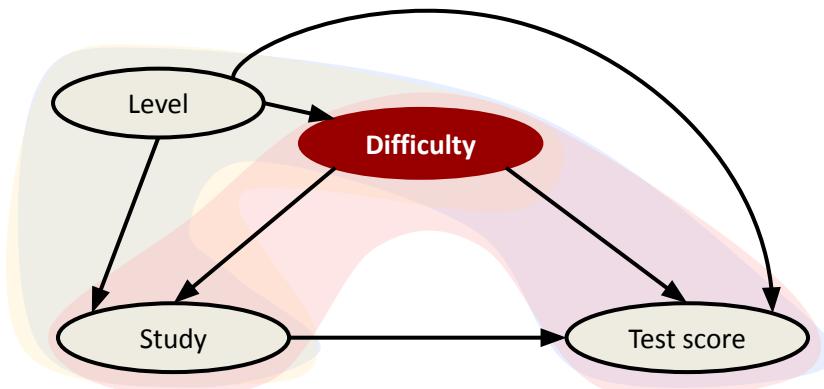
Building adjustment set for Study → TestScore



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

Discovering Confounders with BCD

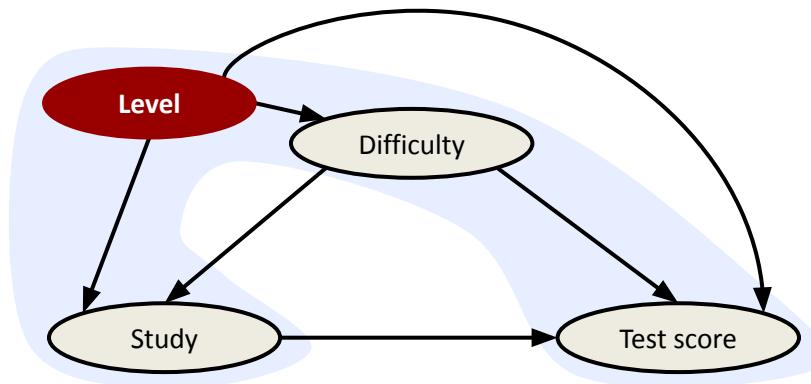
Building adjustment set for Study → TestScore



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

Discovering Confounders with BCD

Building adjustment set for Study → TestScore

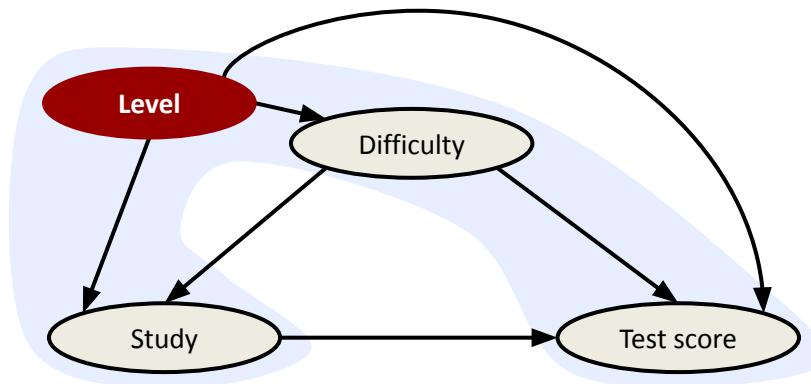


Key Insight: Level is also a confounder, flagged by BCD

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

Discovering Confounders with BCD

Building adjustment set for Study → TestScore



Theorem 1: If \exists confounder between treatment and outcome, \exists attribute A s.t

- $A \rightarrow$ treatment and \nexists confounder between A and treatment. **Flagged by BCD**
- A is a confounder between treatment and outcome. **Selected heuristically**

Confounders in Causal Analysis

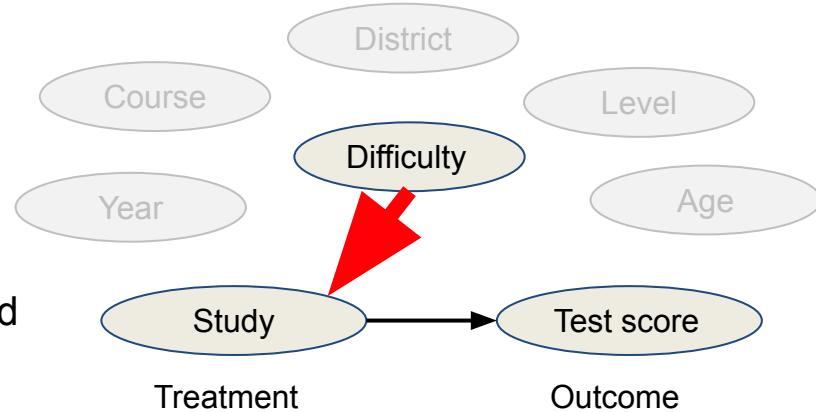
Background: Bivariate Causal Discovery (BCD) estimates → or ← edges from data

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 = $\beta_1 \cdot \text{Study} + \epsilon_1$
- Estimate: MI(Study, Level - $\beta_1 \cdot \text{Study}$)
- Push mutual information through joins

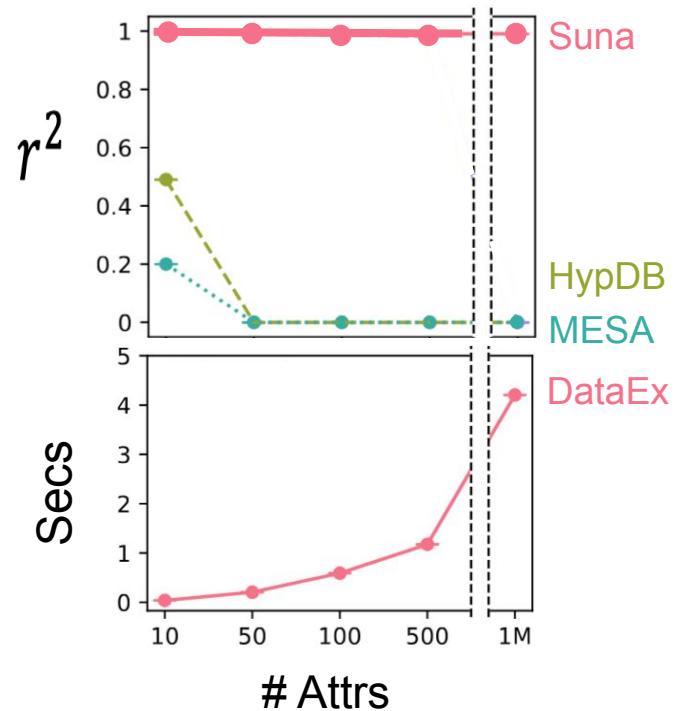


Experimental Results

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

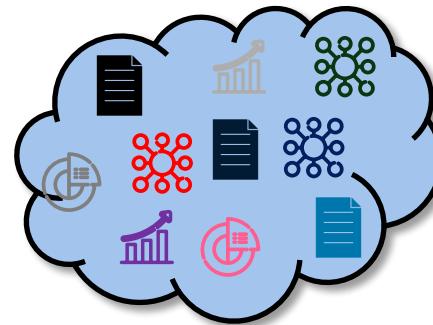
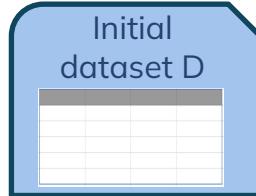
Arbitrary tasks can be supported, but are very difficult...

Metam: Task-agnostic search [Galhotra23]

Problem Setup

GIVEN:

An initial dataset D ,
a collection of attributes Γ ,
a task implementation t



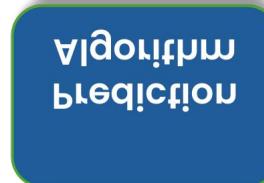
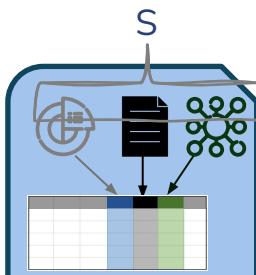
OBJECTIVE: max utility _{t} ($D \bowtie S$)

CONSTRAINT:

$$S \subseteq \Gamma$$

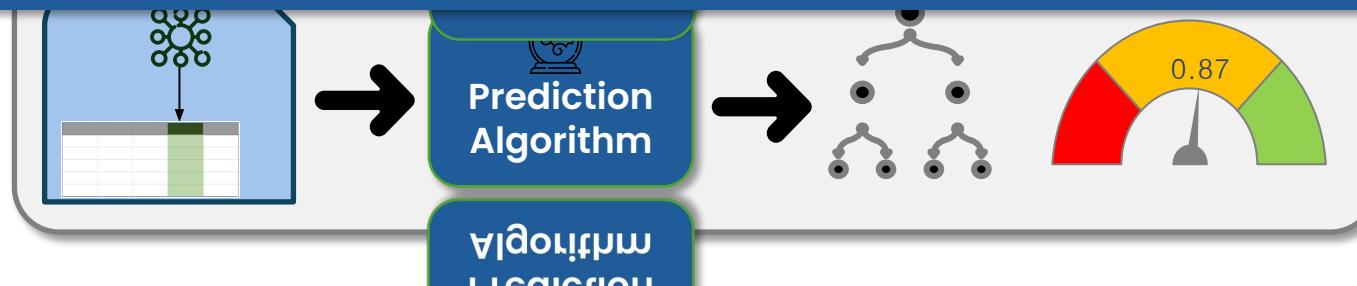
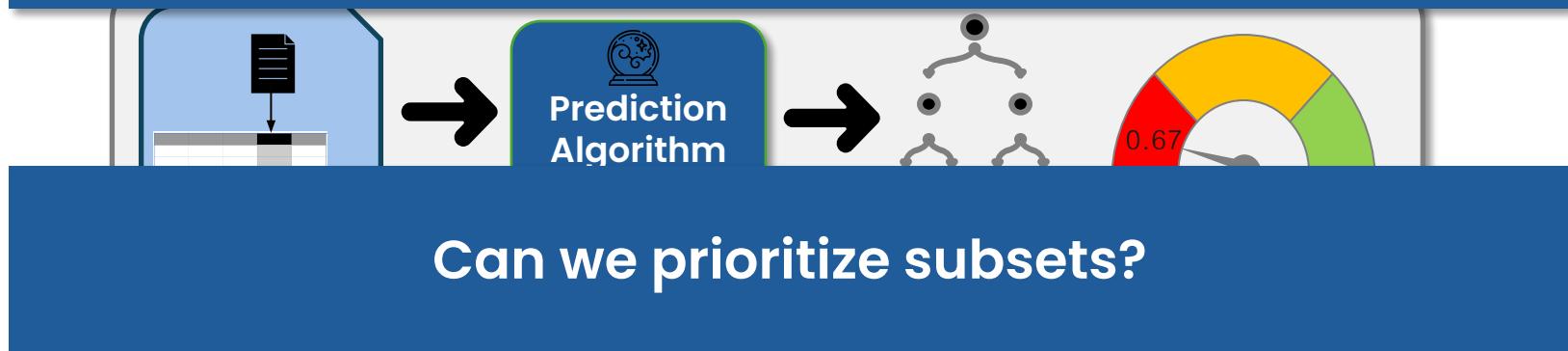
$$|S| \leq k \text{ (a constant)}$$

S is minimal

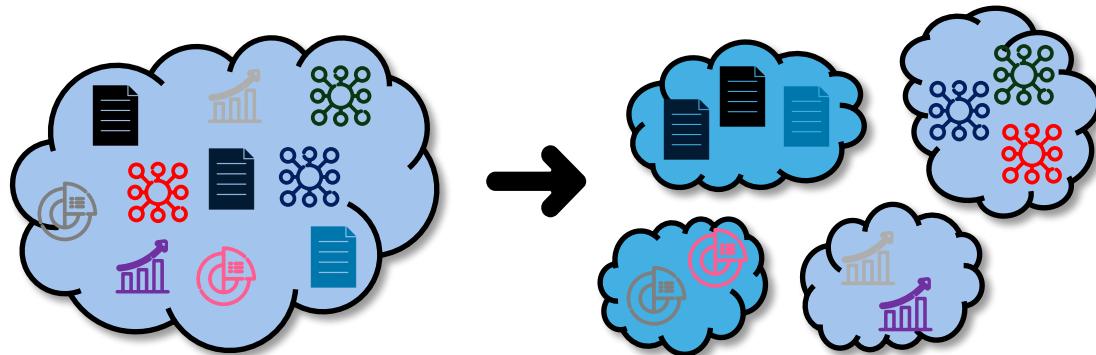


How to solve the problem?

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



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:

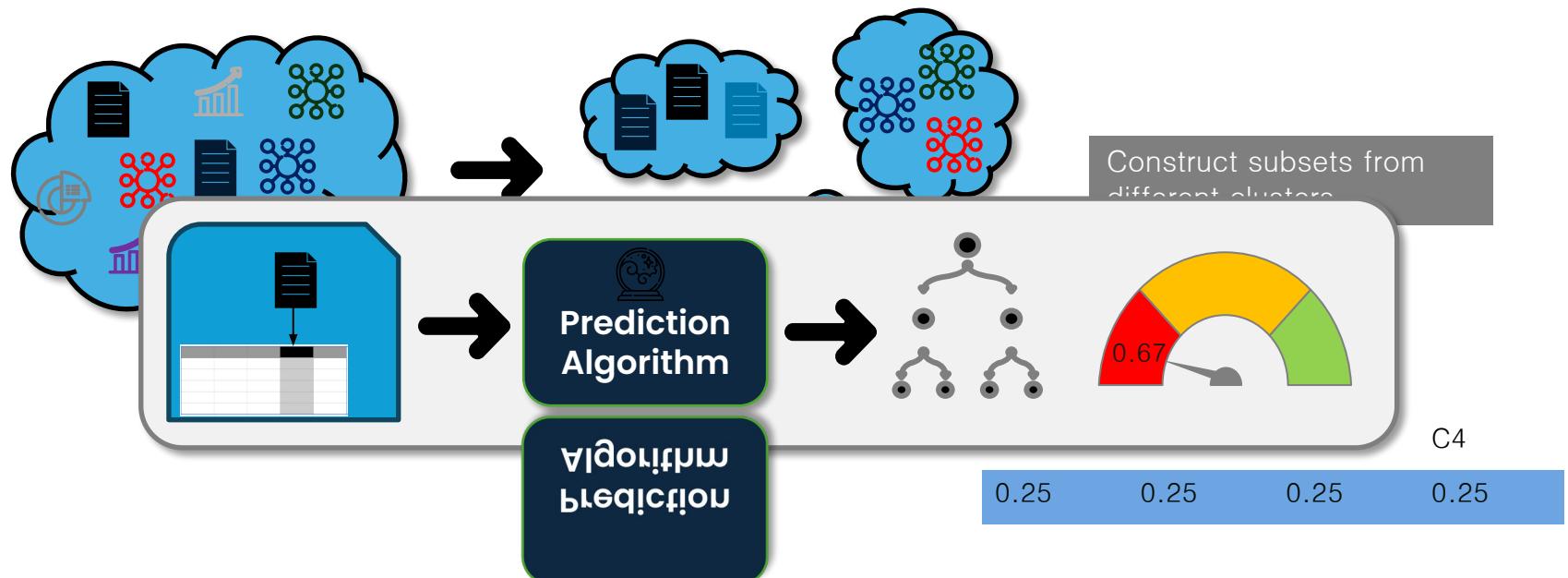
Should I sample more datasets from cluster C_i ?

OR

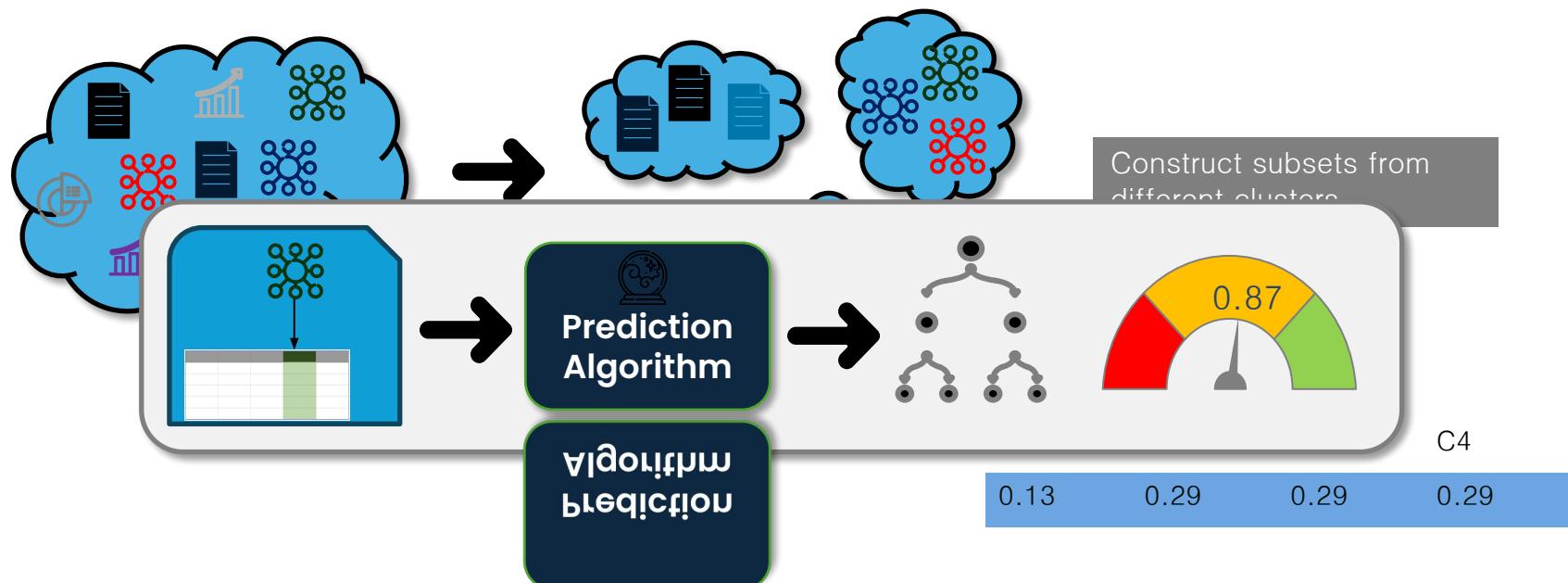
Should I explore different clusters?

SOLUTION: Bandit-based approach

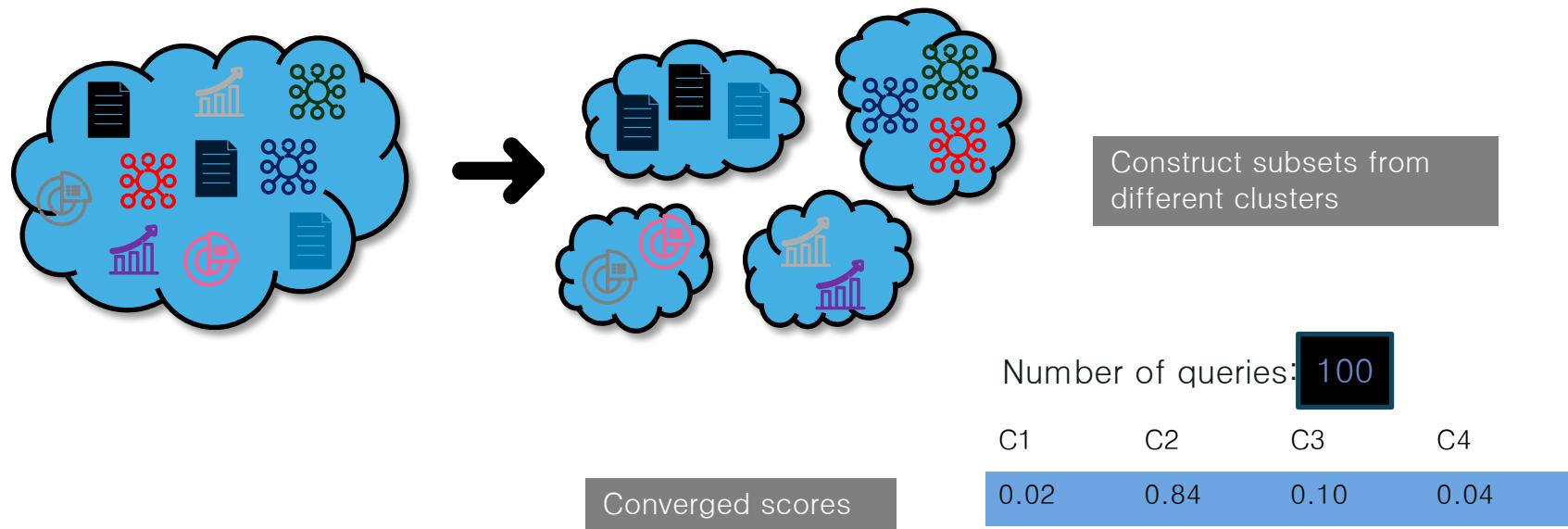
Approach 1: Diversify the Search Process



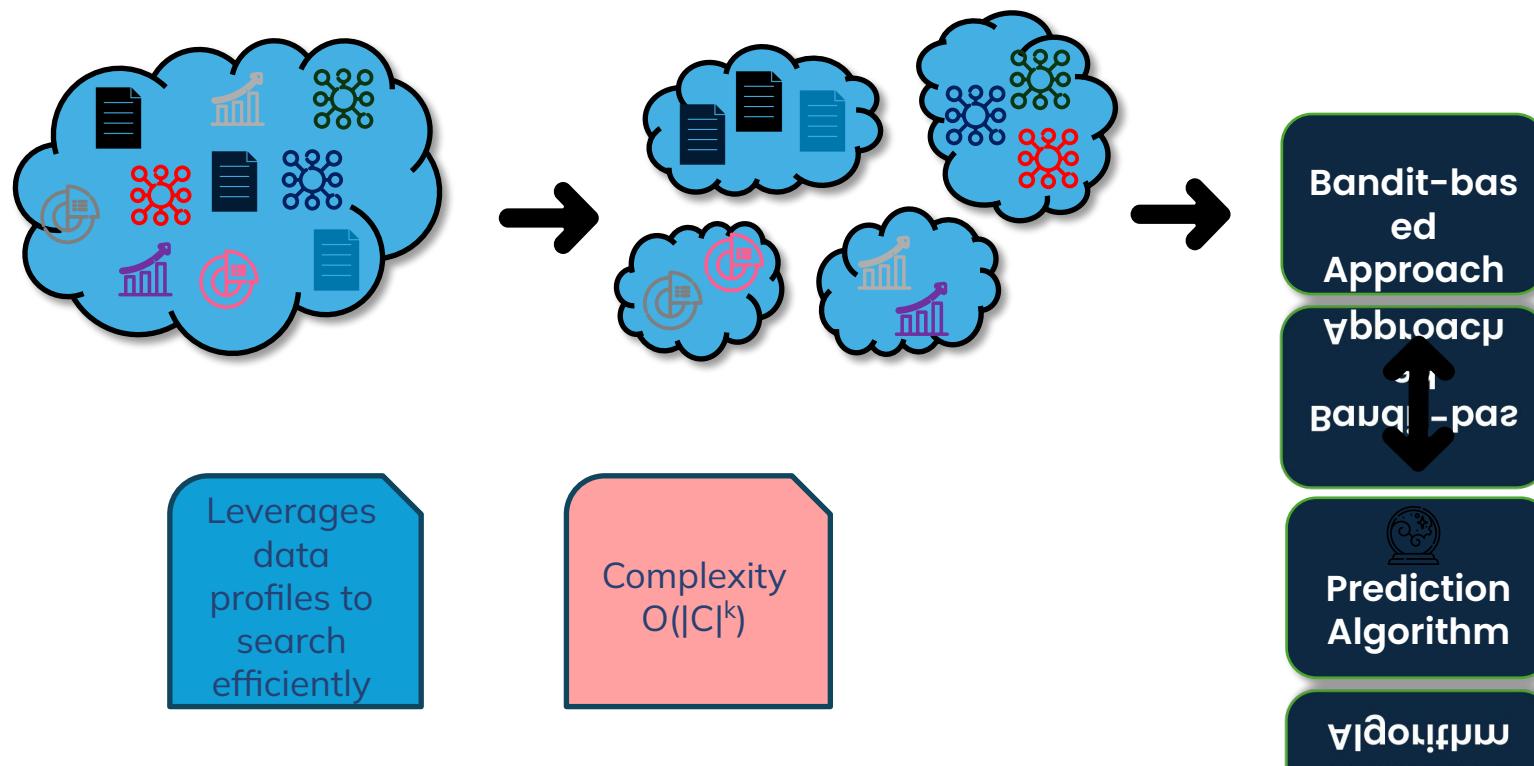
Approach 1: Diversify the Search Process



Approach 1: Diversify the Search Process



Approach 1: Diversify the Search Process

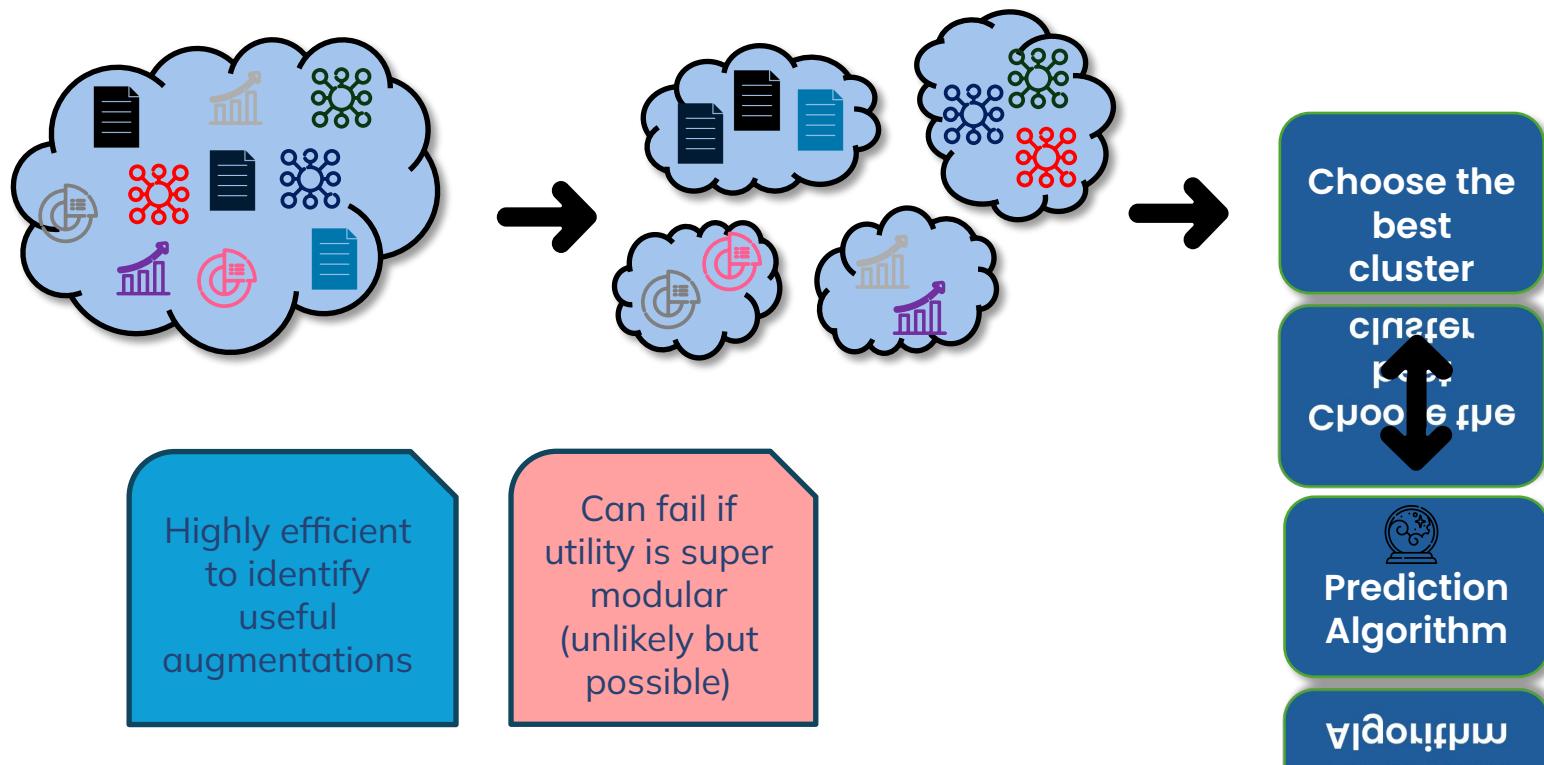


Approach 2: Leverage Monotonicity of Utility Metric

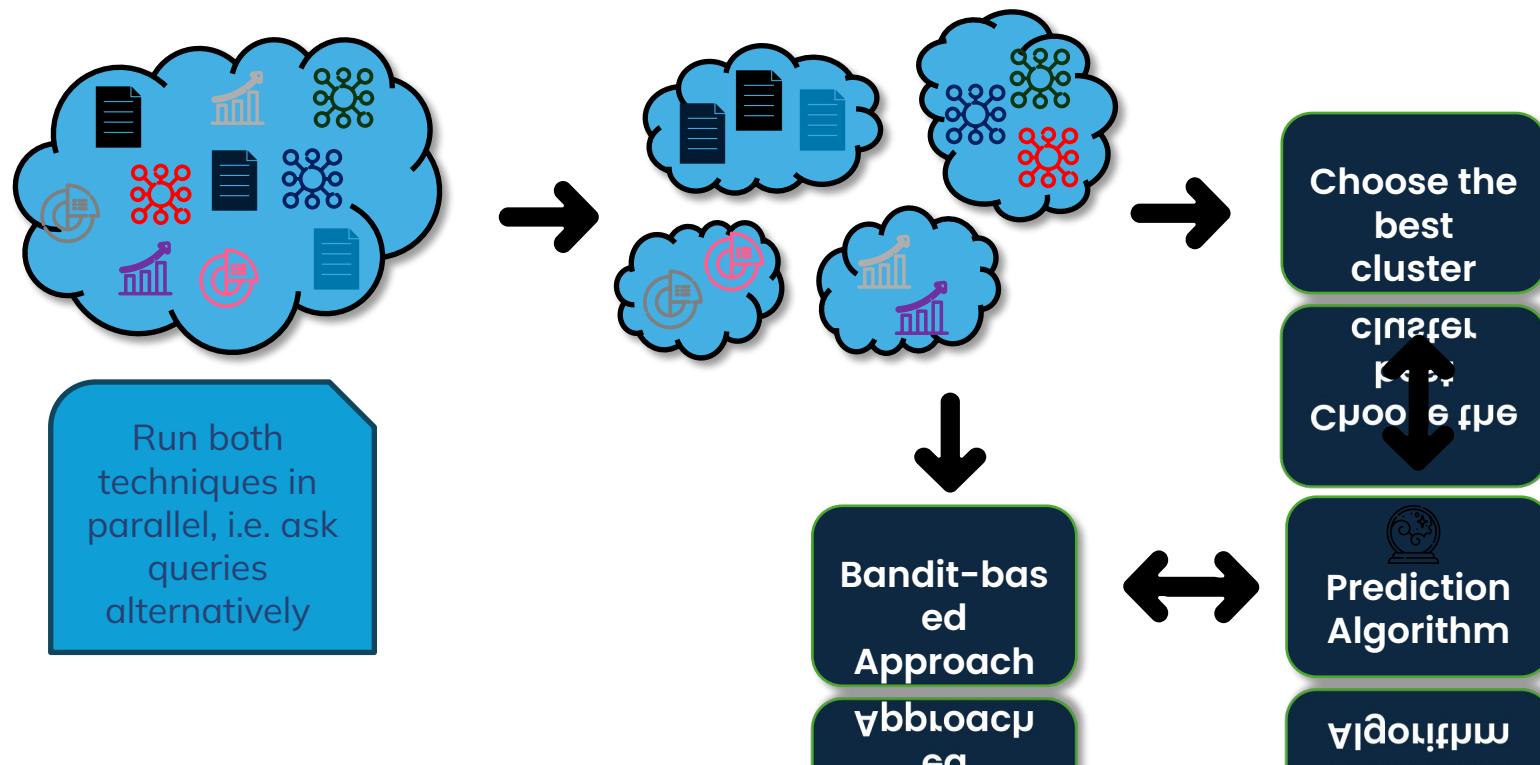
- **Monotonicity:** Easy to guarantee
$$u(D \cup T_2) \geq u(D \cup T_1) \quad \forall T_1 \subseteq T_2$$
- What if the utility is **submodular** too?
 - Diminishing returns property:
$$u(T_1 \cup \{X\}) - u(T_1) \geq u(T_2 \cup \{X\}) - u(T_2) \quad \forall T_1 \subseteq T_2$$

Solution: Greedily choose the best augmentation

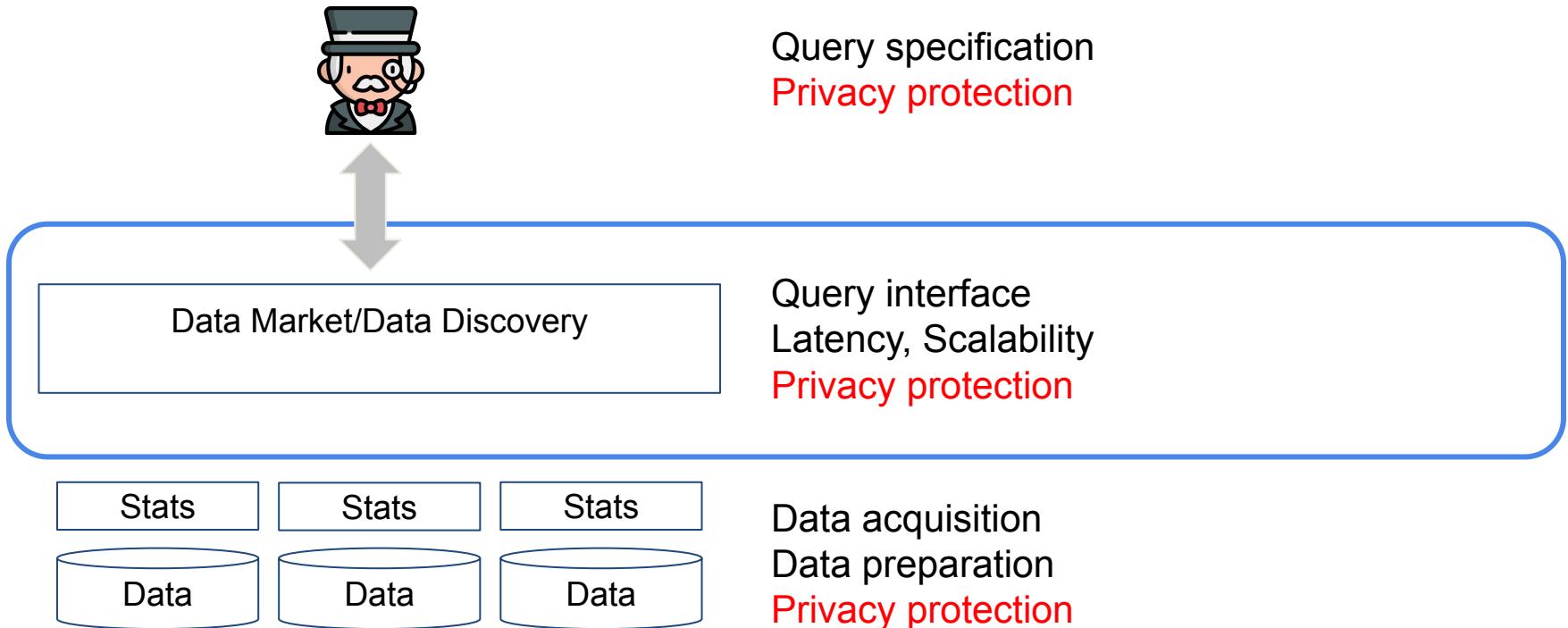
Approach 2: Leverage Monotonicity of Utility Metric



Final Approach: Combining All Ideas



Privacy Challenges Are Everywhere!



Tutorial Organization

- Part I: Data acquisition and search (Eugene and Sainyam)
- Part II: (Daniel)
- Part III: (Shagufta, Zeyu, and Eugene)
- Part IV: Regulatory Considerations (Daniel)
- Part V: Open Questions (Daniel, Eugene and Sainyam)

BACKUP SLIDES

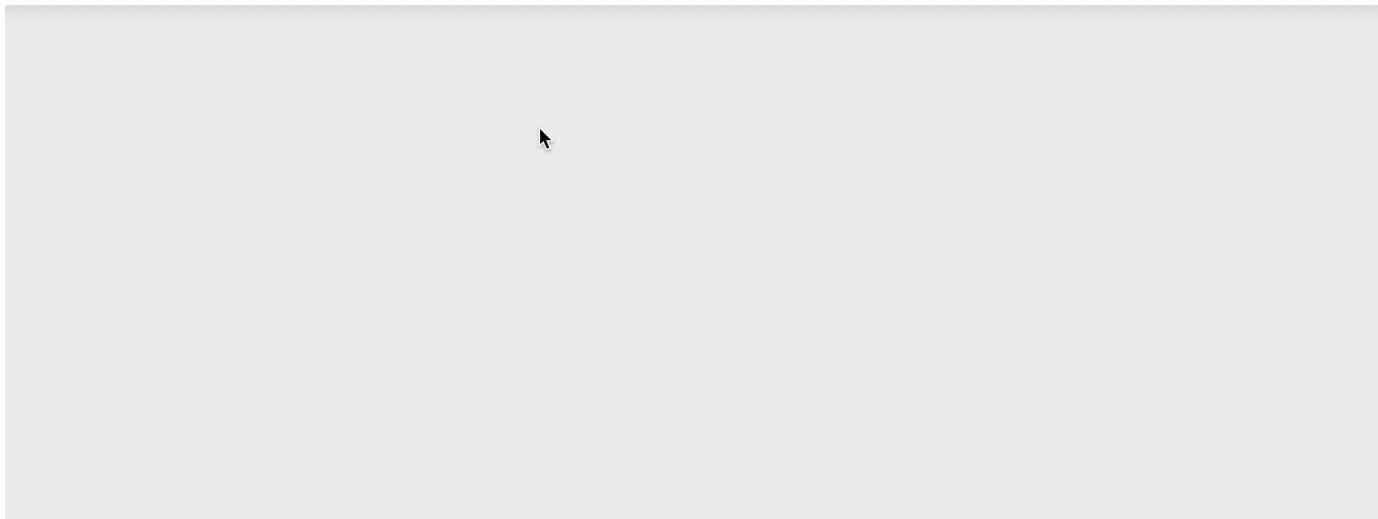
View Presentation

Do you want to shortlist datasets containing the attribute: long_name

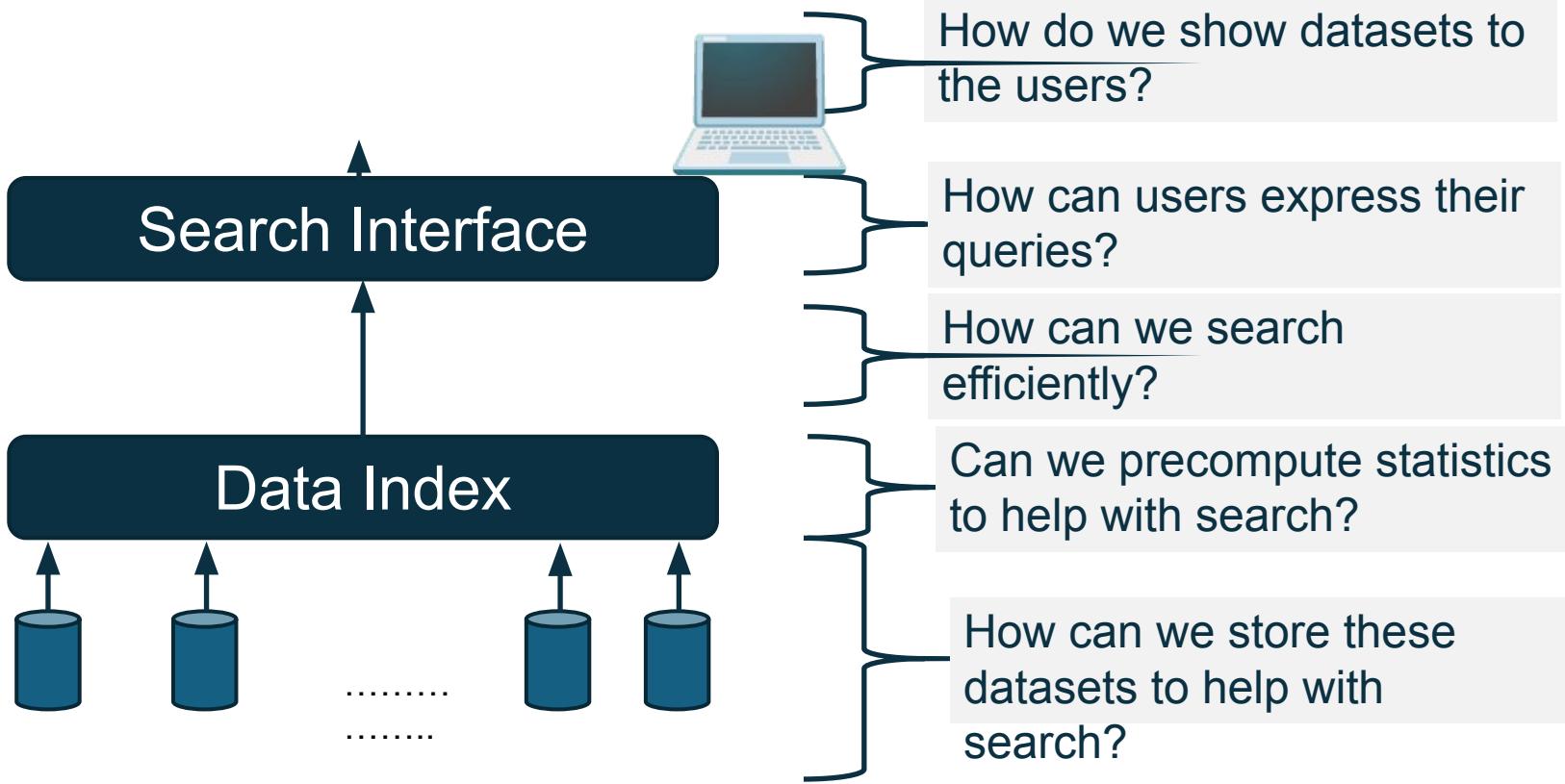
- Yes, my data must contain this attribute
- No, my data should not contain this attribute
- Does not matter

Submit

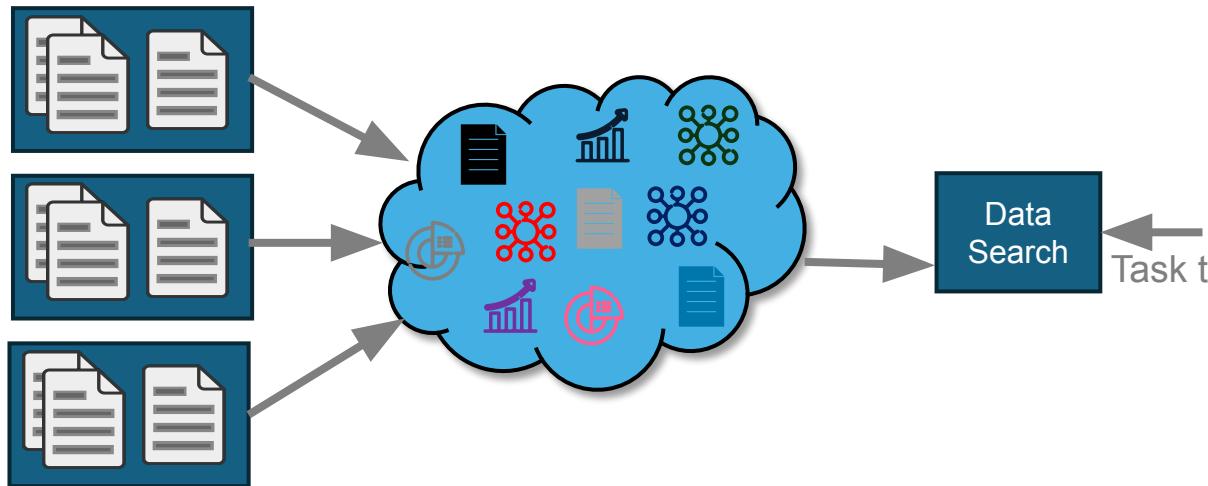
Show Shortlisted d...



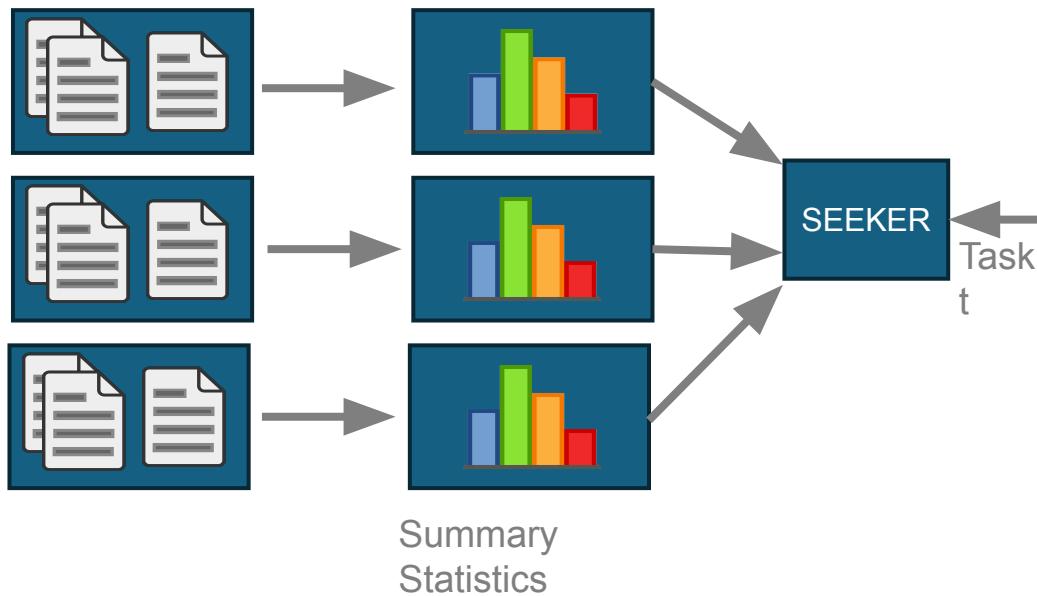
Data Discovery



Centralized Data Discovery



Federated Data Discovery



Distribution-Aware Dataset Search

The screenshot shows a search interface for datasets. At the top, there is a search bar with the query "lung cancer". Below the search bar are several filter buttons: "Last updated", "Download format", "Usage rights", "Topic", "Provider", and "Free". To the right of the search bar are icons for "Saved datasets", "Sign in", and other navigation options.

Below the filters, it says "100+ datasets found". The results are listed in a grid format:

- kaggle Lung Cancer DataSet**
kaggle.com
zip
Updated Jun 1, 2023
+ more versions
- D The IQ-OTHNCCD lung cancer dataset - Dataset - B2FIND**
b2find.dkrz.de
Updated Oct 20, 2020
+ more versions
- kaggle Lung Cancer Dataset**
kaggle.com
zip
Updated Jun 2, 2024

For the first result (kaggle Lung Cancer DataSet), there is a detailed view:

Lung Cancer DataSet [See More Versions](#)

Explore at: [Kaggle | kaggle.com](#)

zip(928105760 bytes)

Dataset updated
Jun 1, 2023

Authors
Falah Gatea

Description
Dataset

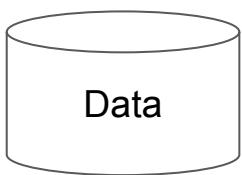
This dataset was created by Falah Gatea

Contents

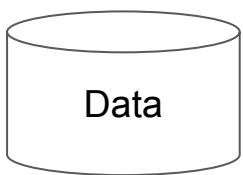
Speculating About the Future



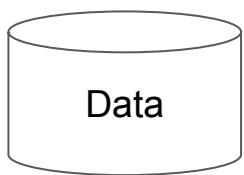
Data Market/Data Discovery



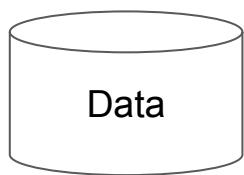
Data



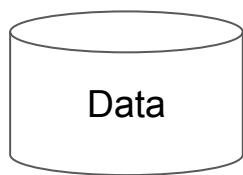
Data



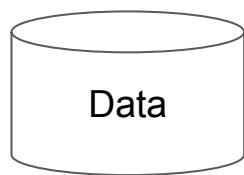
Data



Data



Data



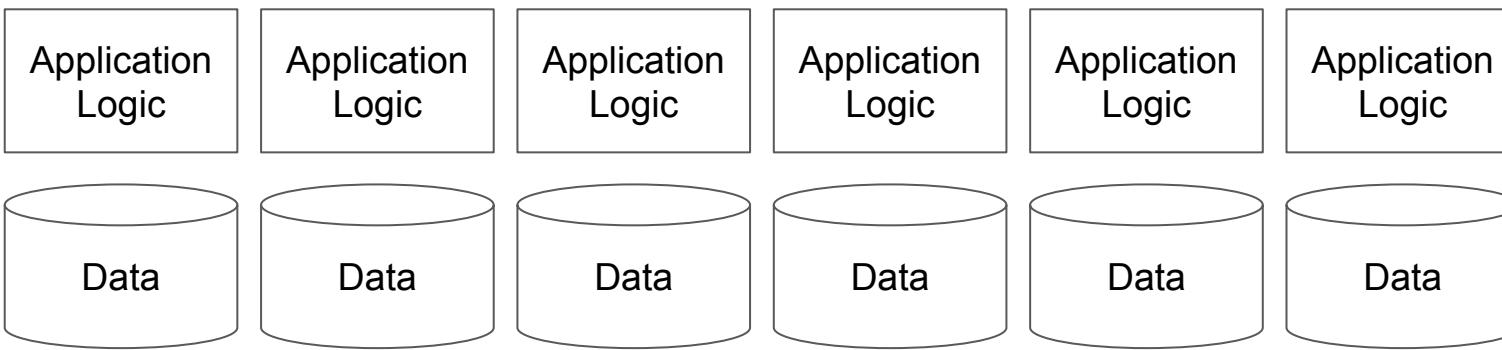
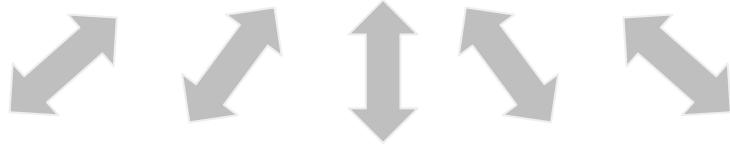
Data





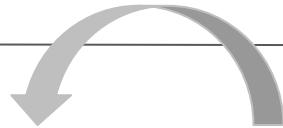
Application
Logic

Data





Data Market/Data Discovery



Application
Logic

Application
Logic

Application
Logic

Application
Logic

Application
Logic

Application
Logic

Data

Data

Data

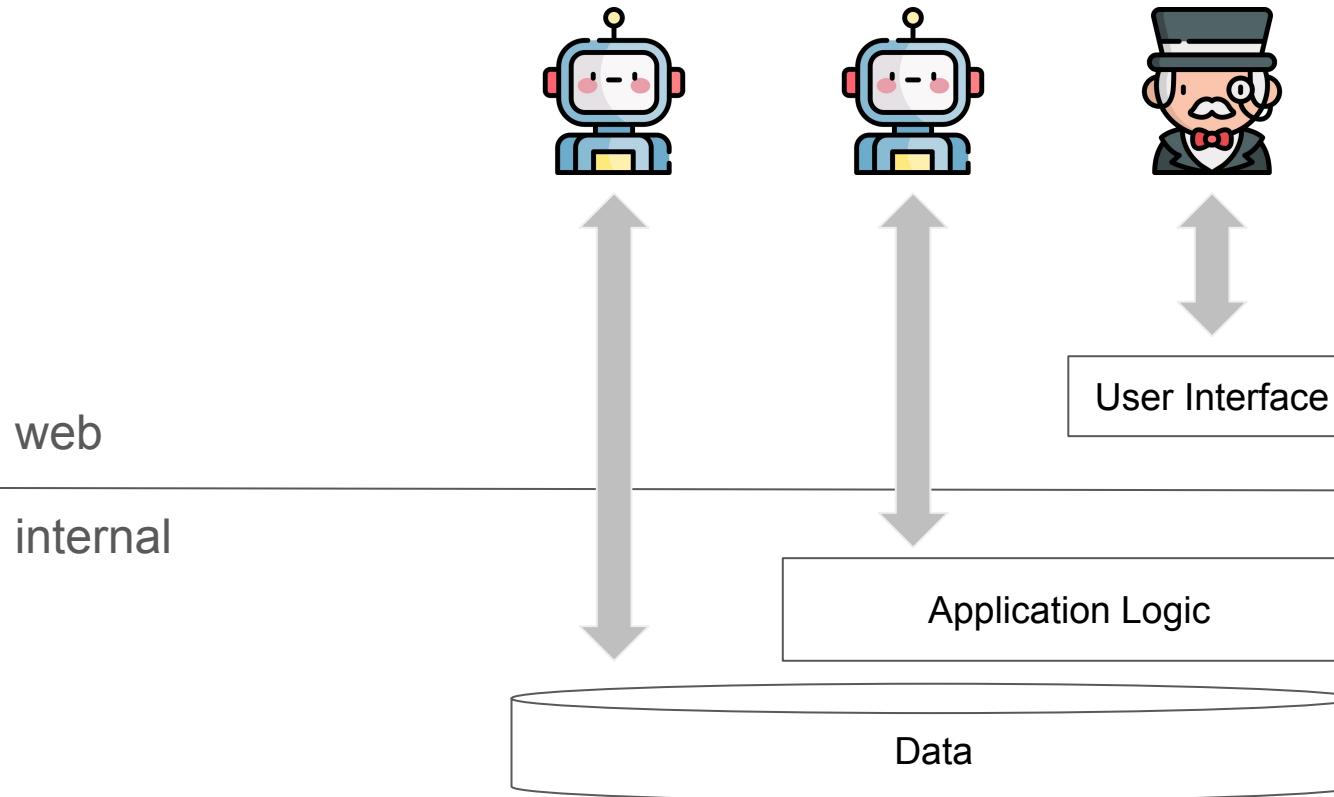
Data

Data

Data

... ■ ■ ■

Will Data Markets Be More Important In the Future?



Will Data Markets Be More Important In the Future?



Data Market/Data Discovery

