

Privacy and Security in Distributed Data Markets

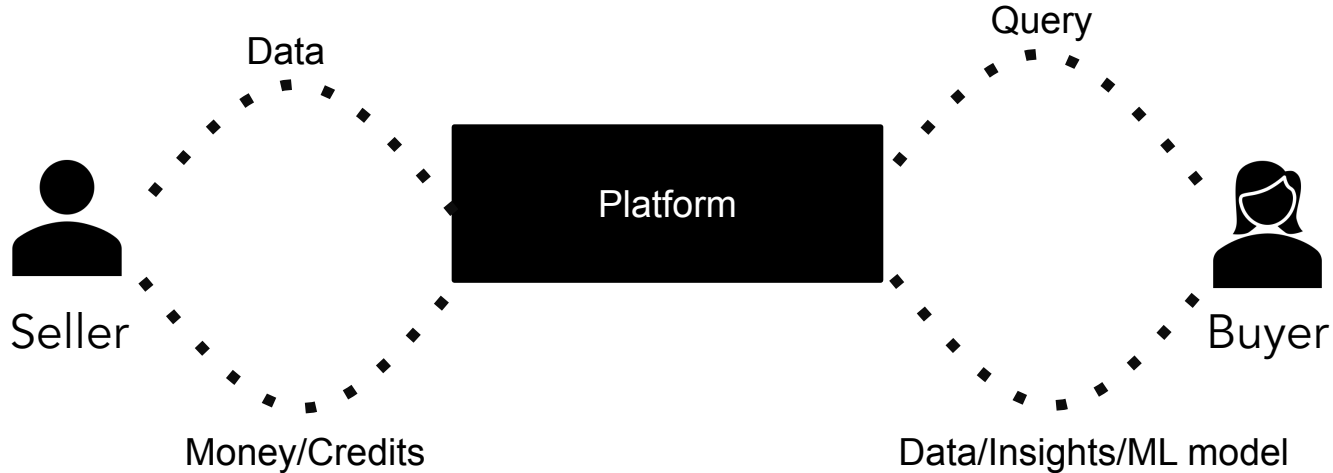
Daniel Alabi, Sainyam Galhotra, Shagufta Mehnaz, Zeyu Song, Eugene Wu

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



narrative

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



Seller

Shares their data



Data
Registration

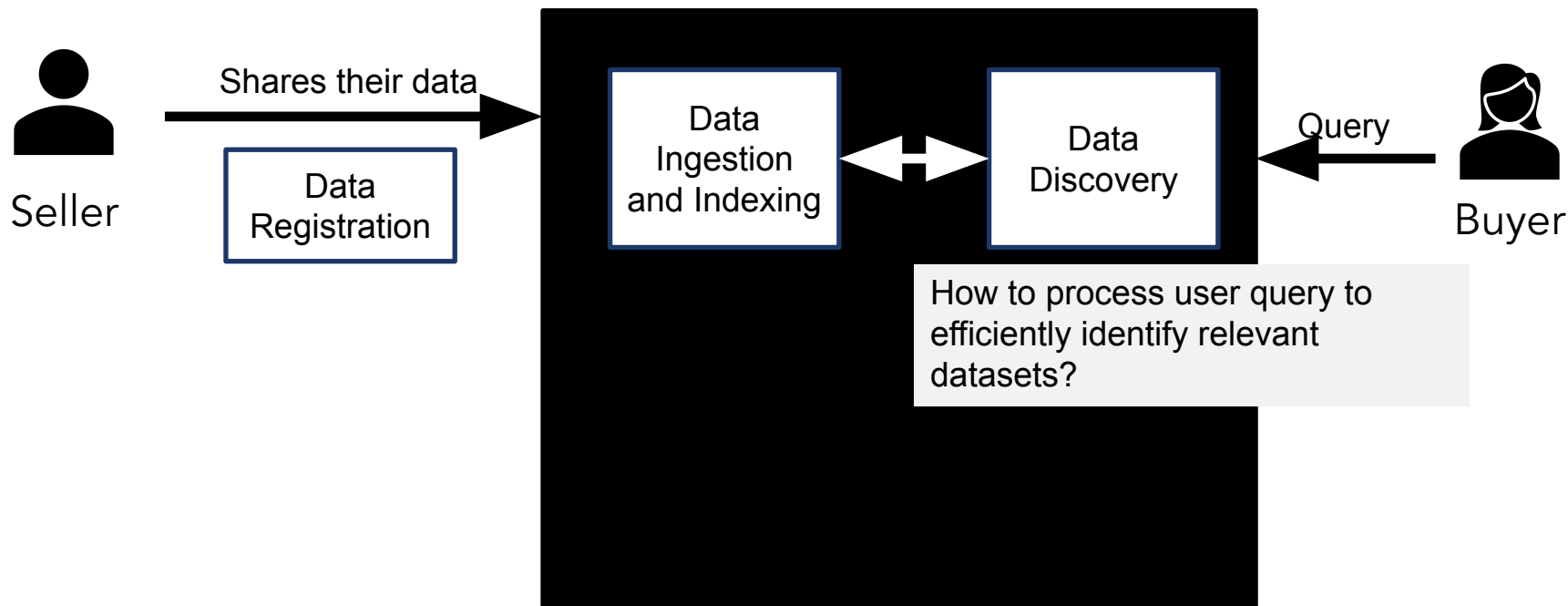
Data Market platform

What information should seller
provide to register the dataset?

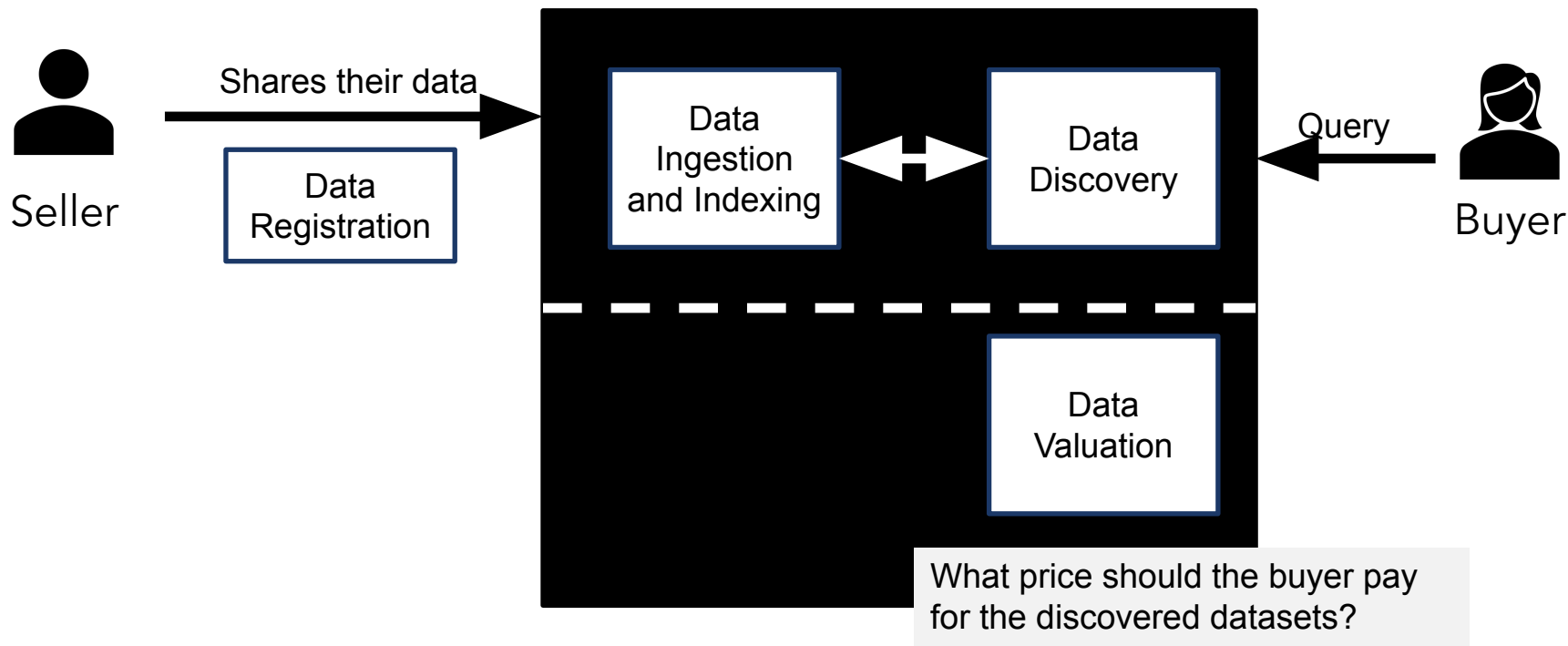
Key Components of a Data Market



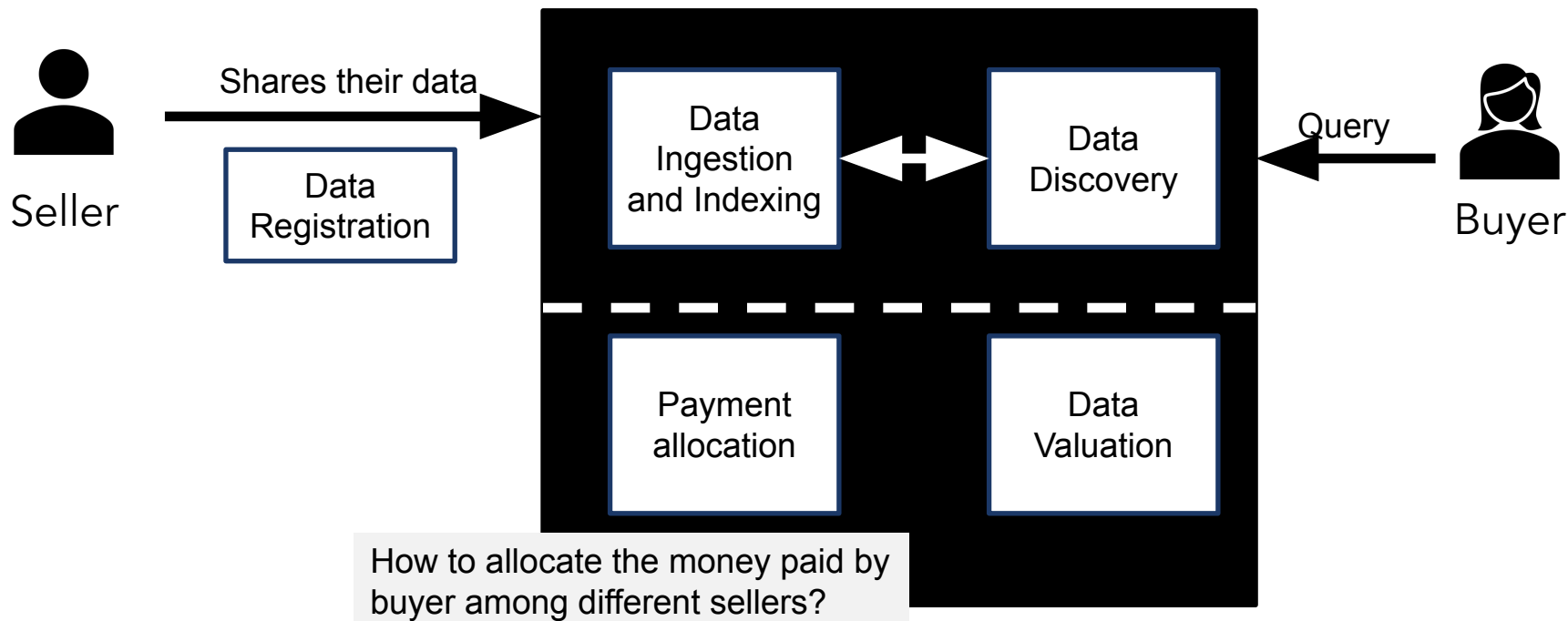
Key Components of a Data Market



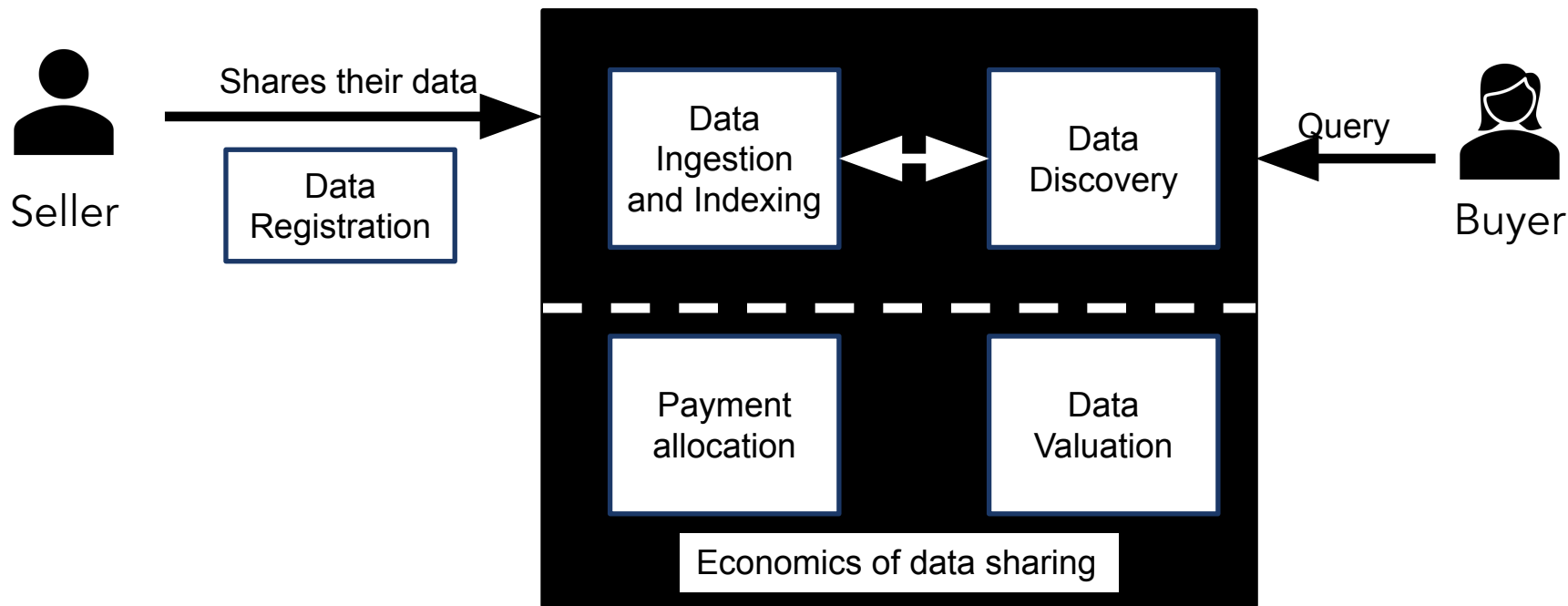
Key Components of a Data Market



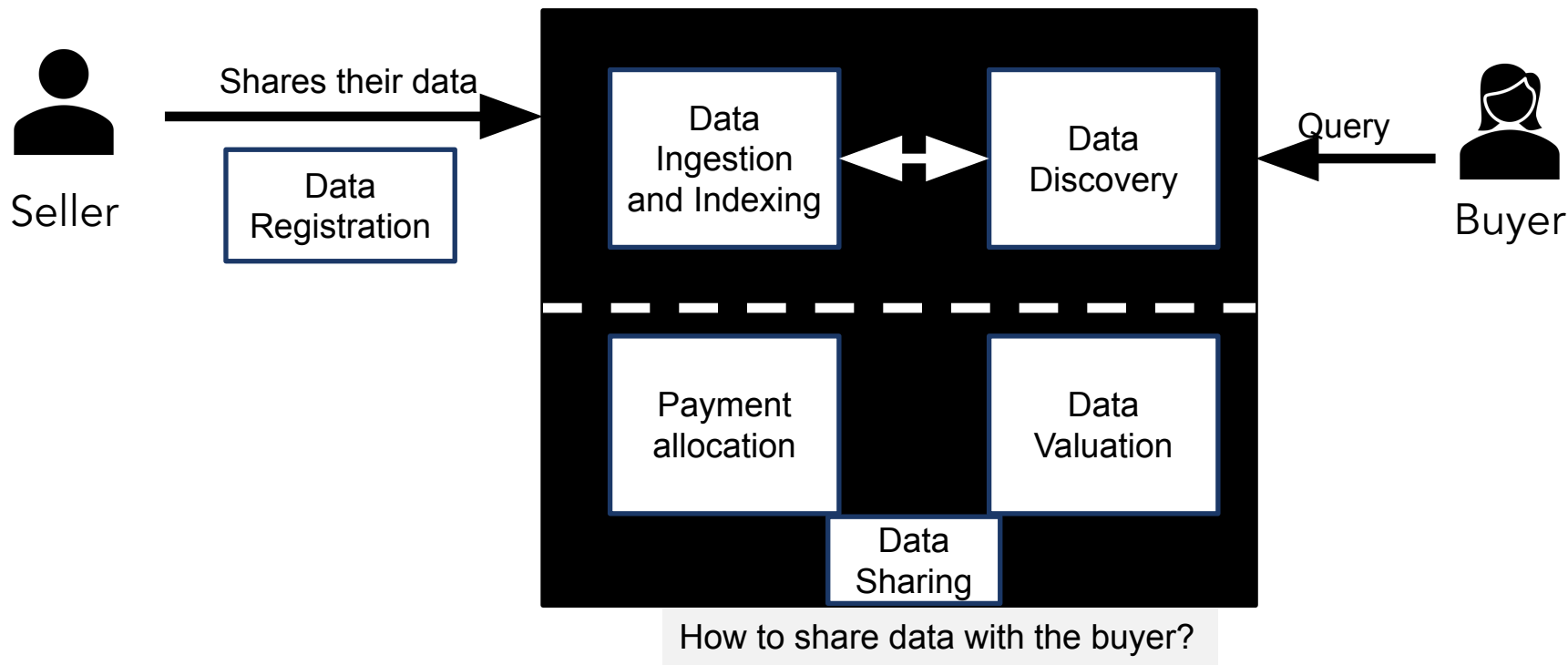
Key Components of a Data Market



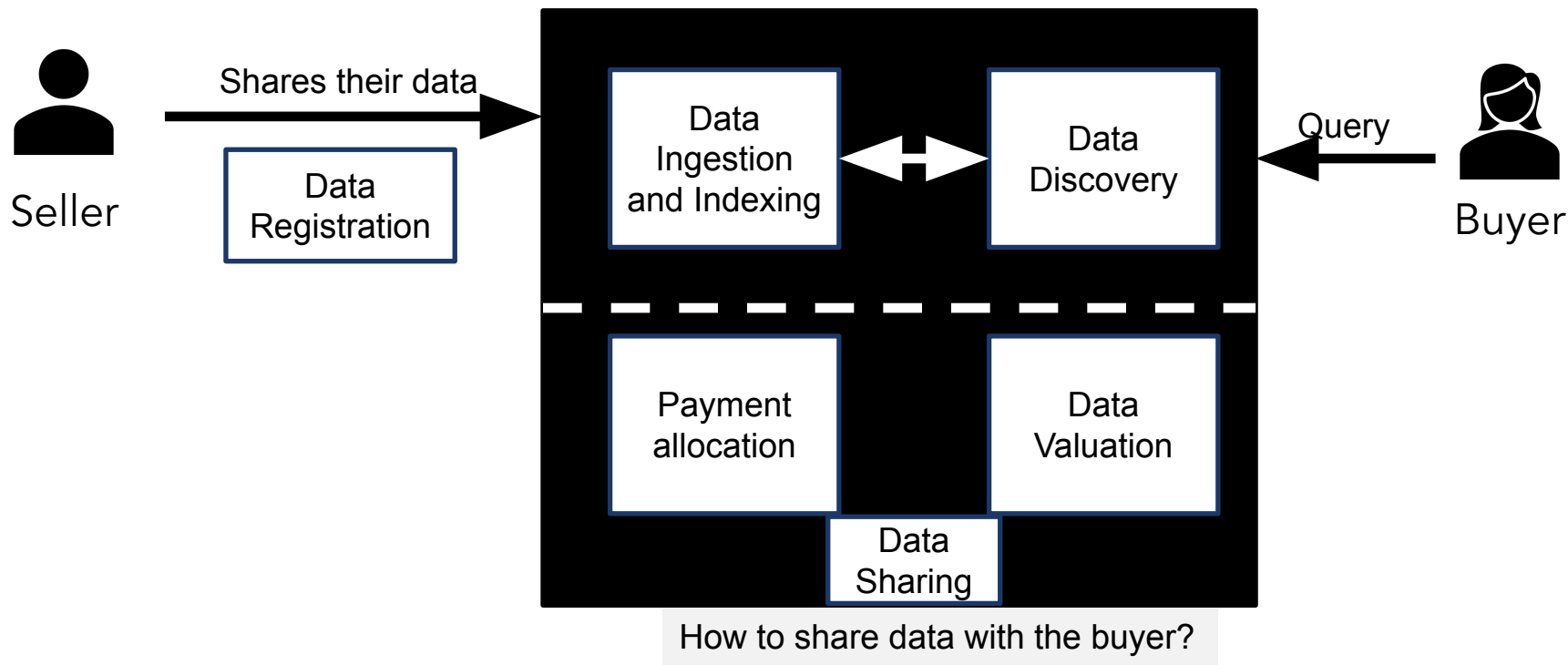
Key Components of a Data Market



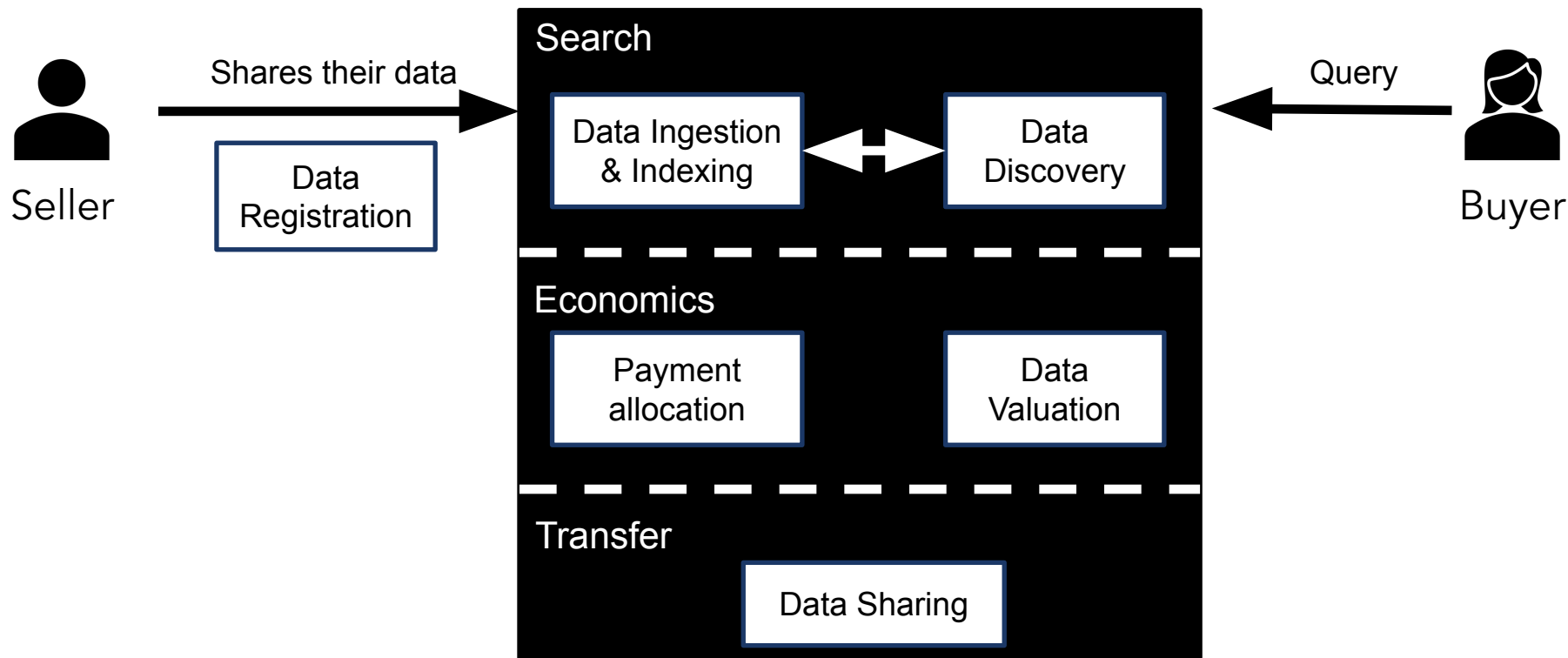
Key Components of a Data Market



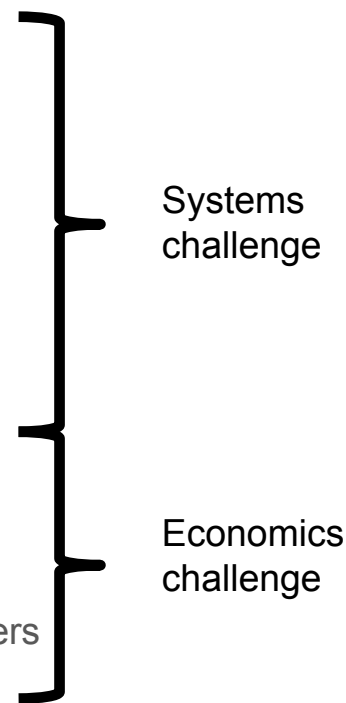
Key Components of a Data Market



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)
 - 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
- 
- The diagram uses two large curly braces on the right side to group the challenges. The top brace, labeled 'Systems challenge', groups the first two main bullet points: 'Data Registration and Discovery' and 'Data Sharing (or acquisition)'. The bottom brace, labeled 'Economics challenge', groups the remaining two main bullet points: 'Data valuation' and 'Payment allocation'.
- 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
- [
- [
- [
- F

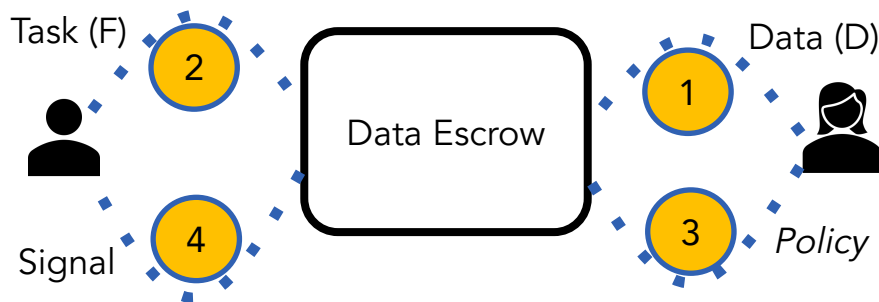
Tutorial Organization

- 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?

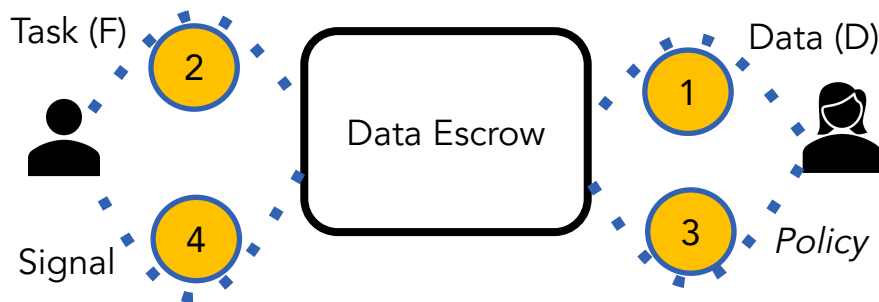
Data Escrow [VLDB'22]

- 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



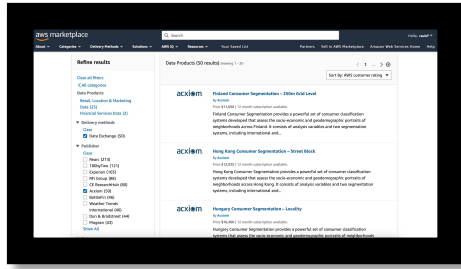
- Guarantee: no data* leaves the escrow without explicit permission, i.e., without an explicit *policy*

Using the Escrow to *Signal* Dataflow results

Data Markets



Seller



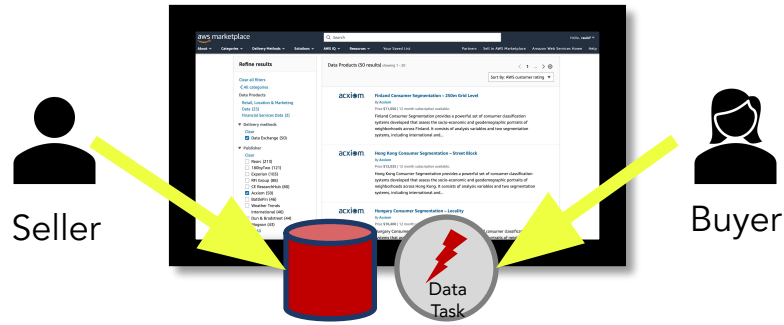
Buyer



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

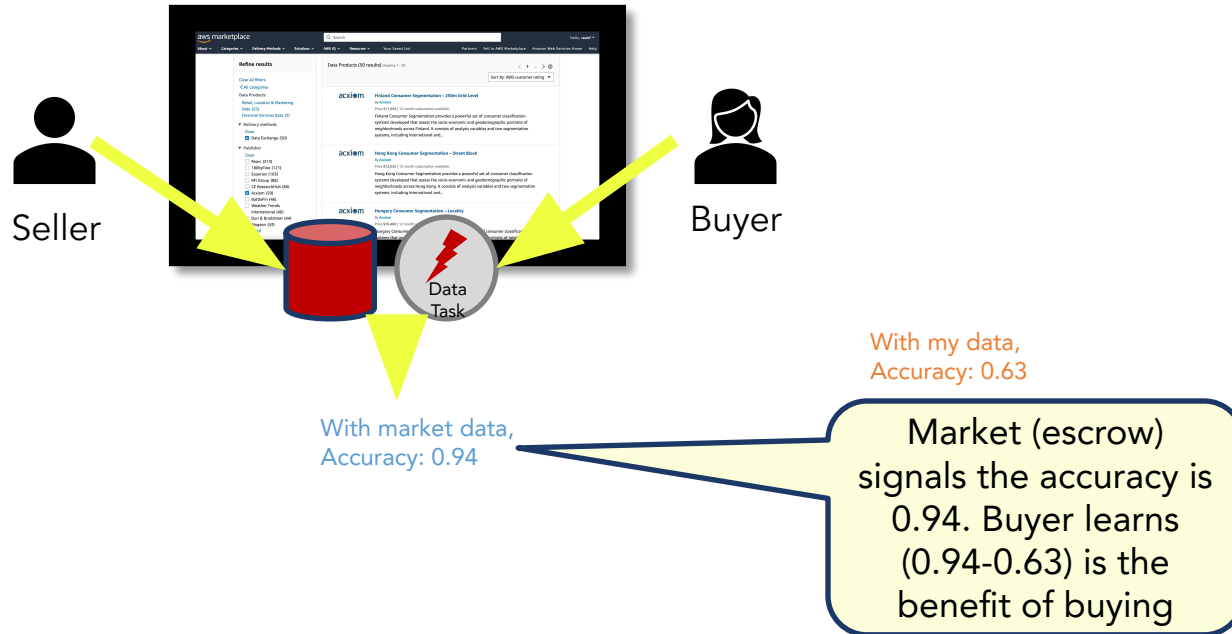
Data Markets



With my data,
Accuracy: 0.63

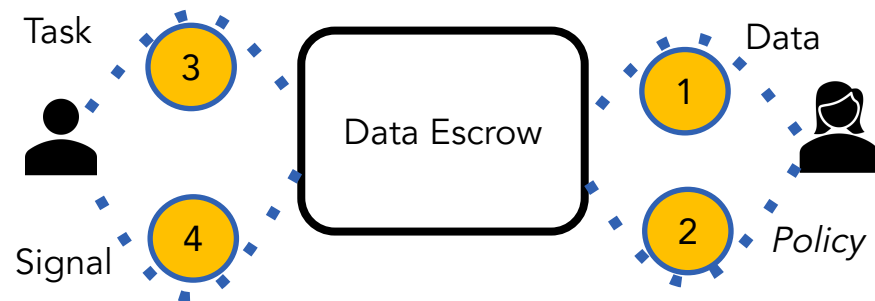
Using the Escrow to *Signal* Dataflow results

Data Markets



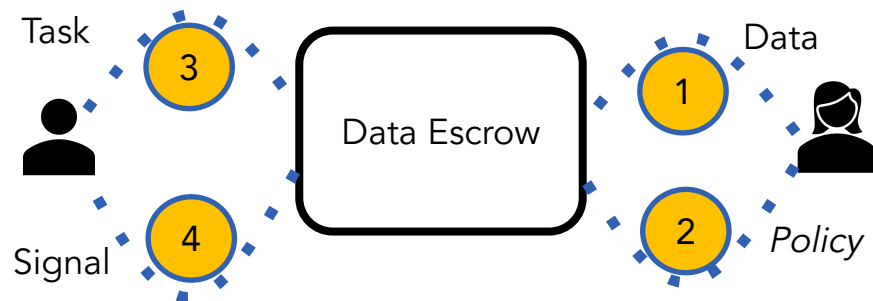
How do we delegate tasks, create signals,
i.e., how do we control dataflows?

Programmable Dataflows



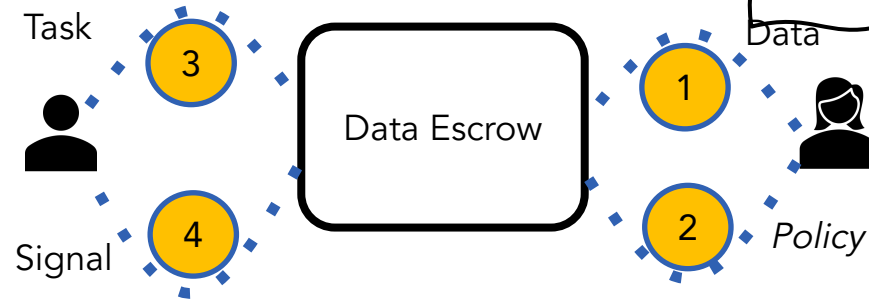
Programmable Dataflows

- Escrow Programming Framework (EPF)



Programmable Dataflows

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



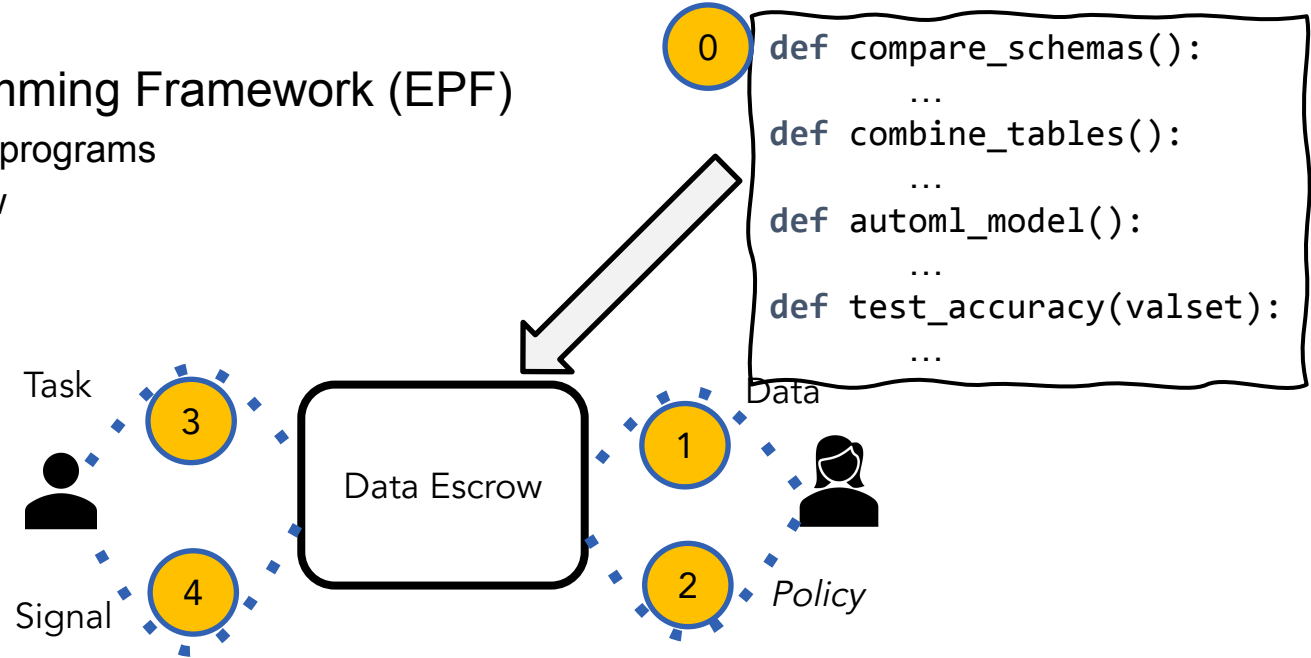
0

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

Programmable Dataflows

- Escrow Programming Framework (EPF)

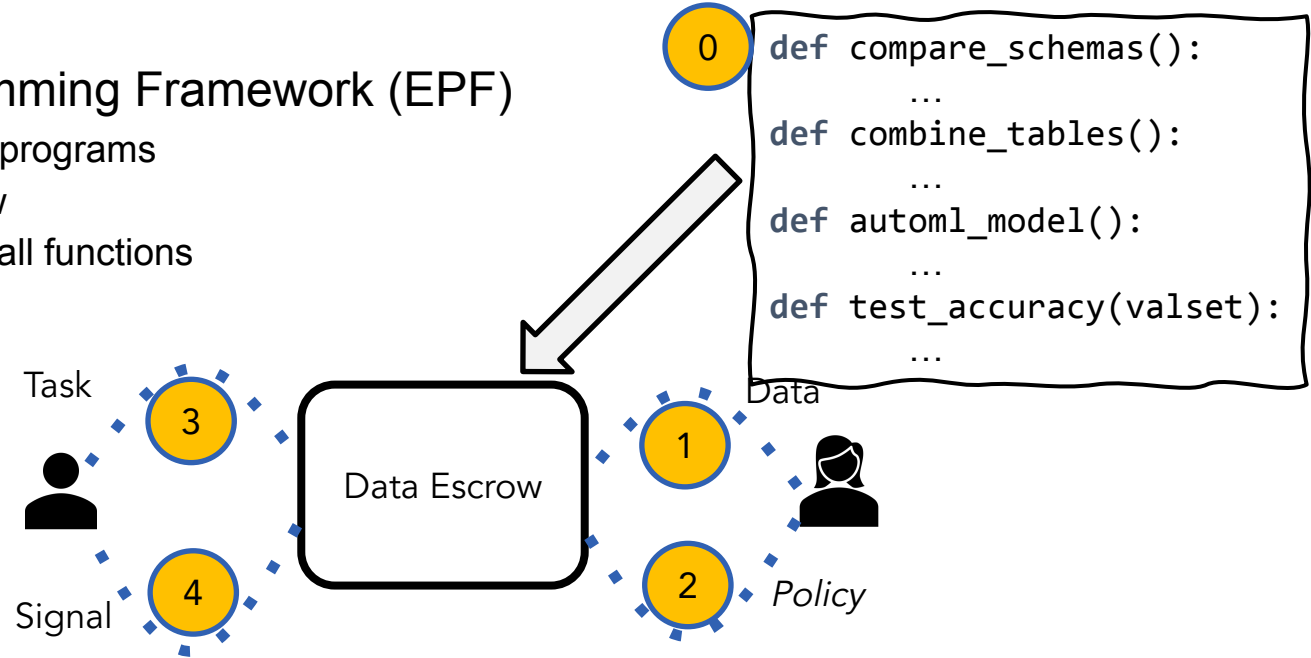
1. Developers write programs
2. Deploy on escrow



Programmable Dataflows

- Escrow Programming Framework (EPF)

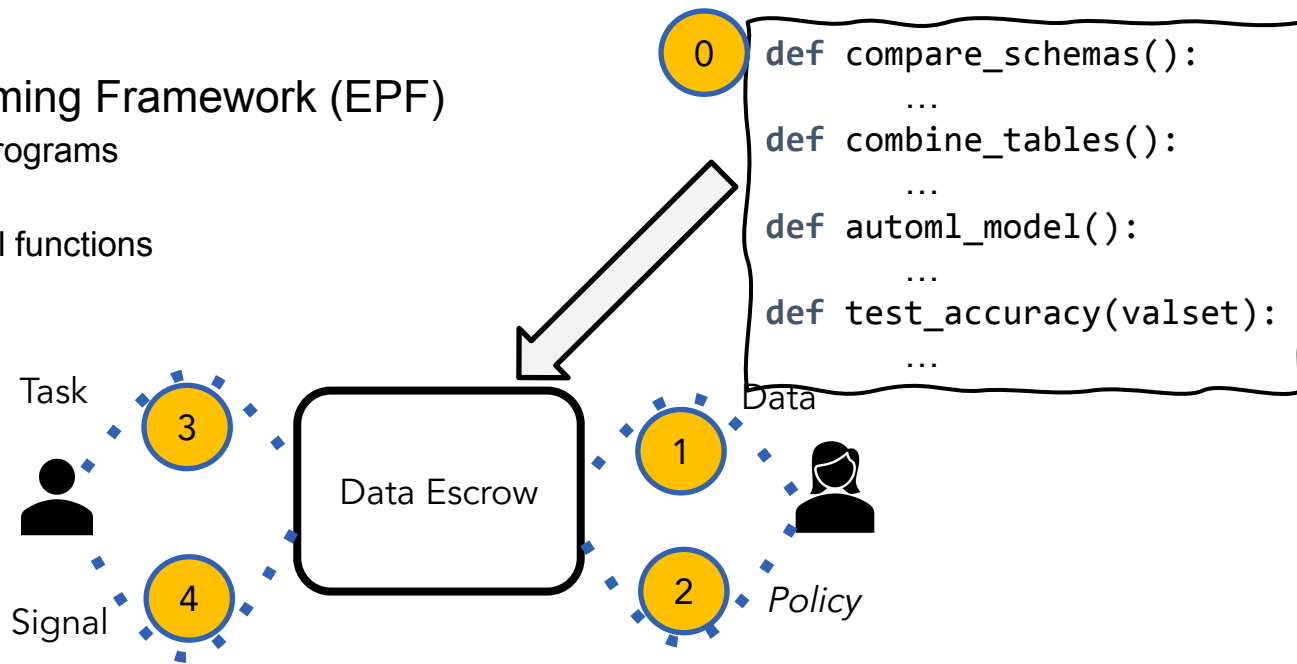
1. Developers write programs
2. Deploy on escrow
3. Agents join and call functions



Programmable Dataflows

- Escrow Programming Framework (EPF)

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



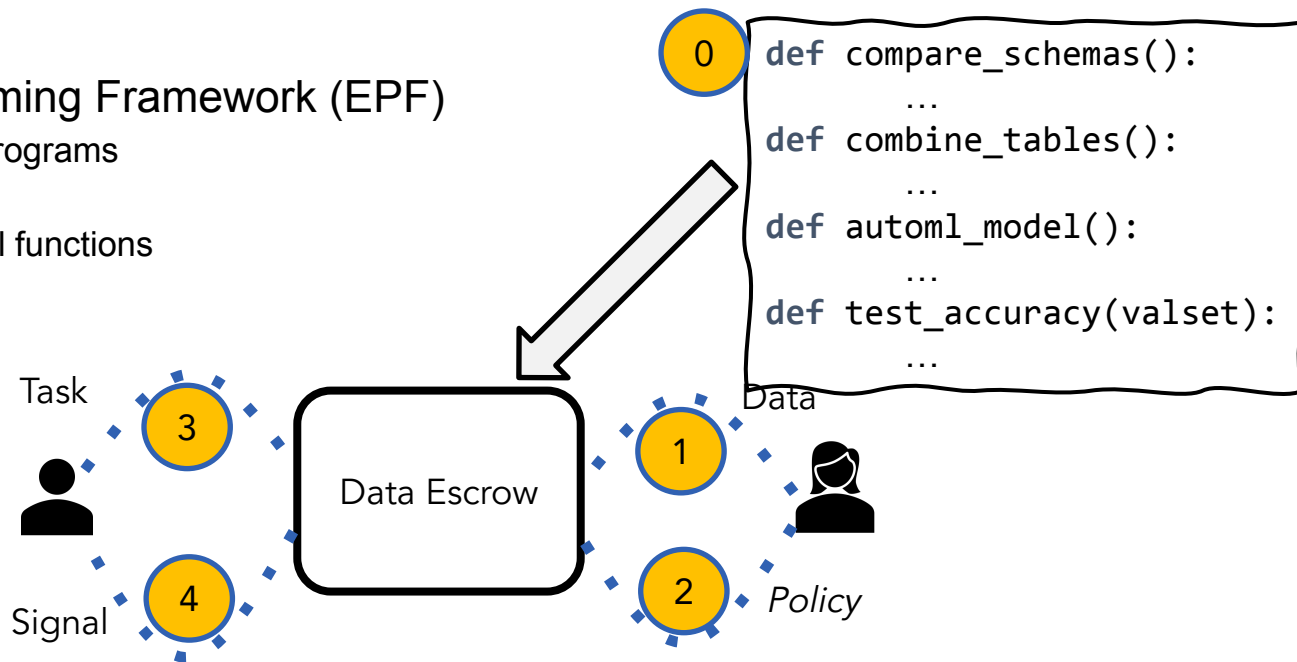
- Program implements communication and logic via *contracts*

Programmable Dataflows

- Escrow Programming Framework (EPF)

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

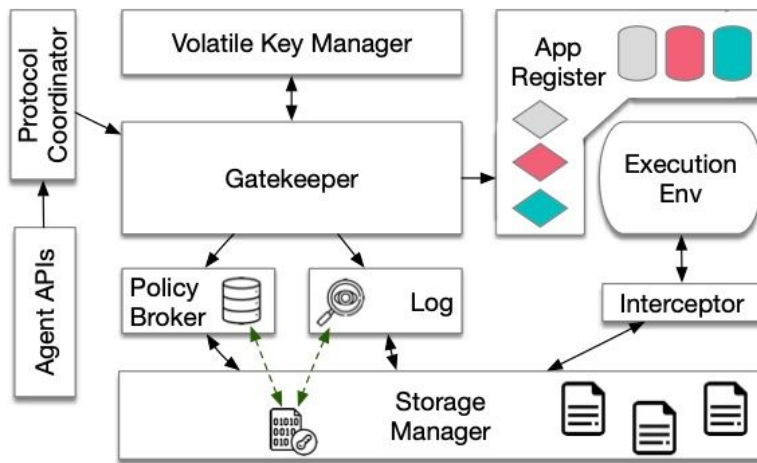
Programs
implement
data environments



- 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...

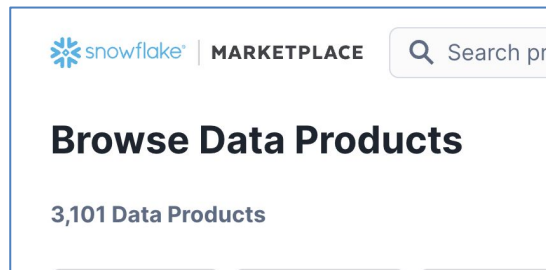
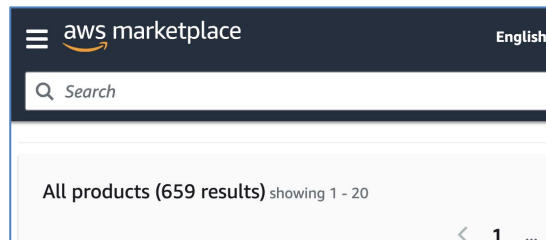
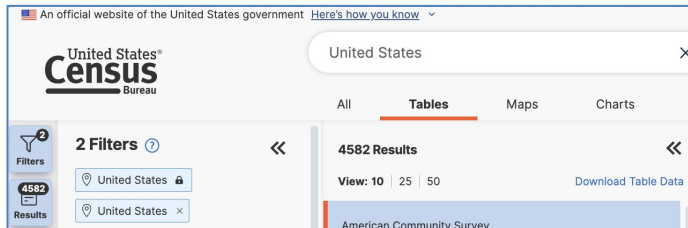
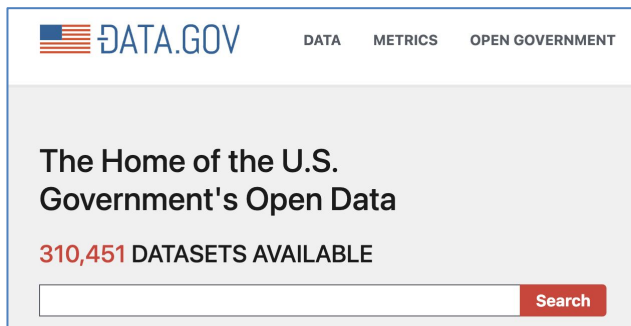


Data Search

Unlimited Storage → Massive Data Repos

Gov Portals
Data Markets
Data Lakes
Web Tables
Data coalitions

...

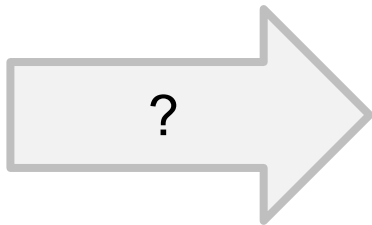


Powered by Dataset Search

[Dataset Search](#), a dedicated search engine for data, indexes more than 45 million datasets from more than 100 sources. It covers many disciplines and topics, including government data, scientific data, and more.

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



Data Market/Data Discovery



A single system stores & manages the datasets

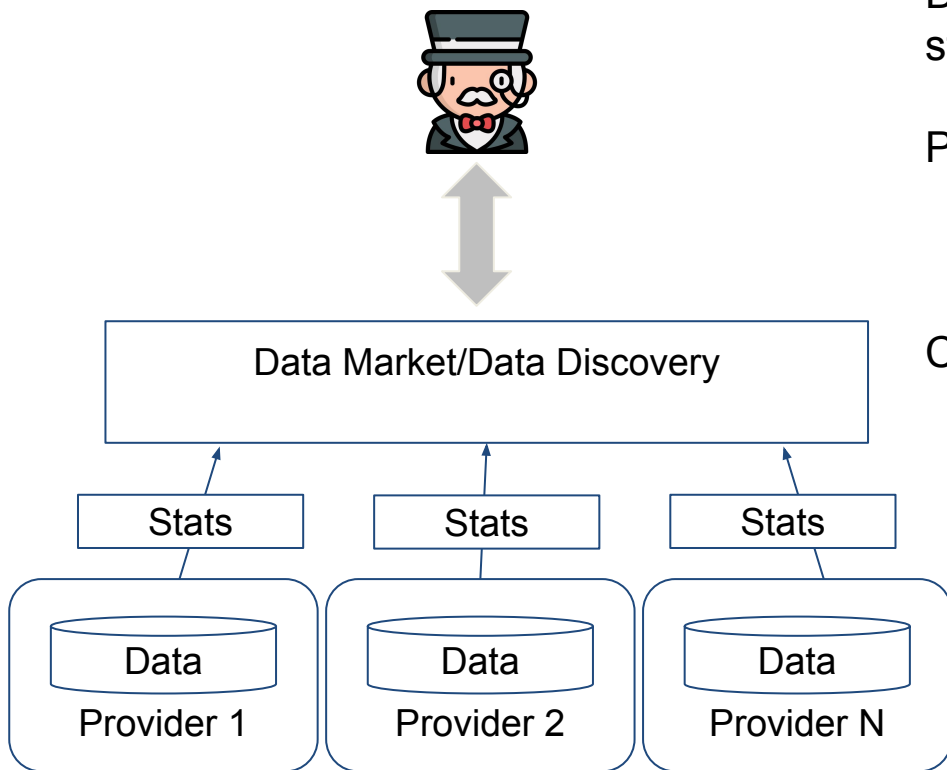
Pros:

- 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



Query specification
Privacy protection



Data Market/Data Discovery

Query interface
Latency, Scalability
Privacy protection

Stats

Stats

Stats

Data

Data

Data

Data acquisition
Data preparation
Privacy protection

3 Classes of Systems

Keyword/Metadata Search

Data Discovery

Task-based Search

Keyword Search



data.gouv.fr



DATA.GOV

Dataset Search

NYC OpenData

City of London Open Data



MARKETPLACE

predict nyc education test scores

624 Data Products

Availability ▾

Categories ▾

Business Needs ▾

Geo ▾

Time ▾

Price ▾

More Filters



Truelty Identity Resolution

Truelty

Your Data • Your Customer • Your Snowflake



InNote

Innovaccer Inc.

Physician's digital assistant that surfaces health insights at point of care.



Free Sample: Cross Shopping Insights - NYC Restaurants

SafeGraph

Geographic patterns and brand affinities in consumer spending



Test Automation for Snowflake

NTT DATA

Automate the data quality monitoring



Area store visits data | Visits to shoe stores in NYC in 2022 | Free sample

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.

Query Decomposition

Task Specifications

task

Evaluate the effects of remote work on quality of life through various periods of the pandemic

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 (1)

column group include hours

Search using your own Column Concept

Enter column name...

Smart Filter by Column Concept:

- ☐ recession
- ☐ satisfaction
- ☐ employment
- ☐ remote
- ☐ quality

Remaining datasets: 8

apply filters

Top Granularity Filters

- ☒ Country (5)

Top Dataset Results

Showing 1 to 8 of 8 datasets.

- Global Remote Work & Wellbeing Dataset.**
10 cols · 10000 rows · 133.3 kB · 179
- Remote Work USA (COVID-19)**
24 cols · 3147 rows · 2.0 MB · 81
- Remote Work Productivity**
5 cols · 1000 rows · 6.7 kB · 3.3k
- COVID-19 on Working Professionals**
15 cols · 10000 rows · 255.2 kB · 2.2k
- Impact of COVID-19 on Working Professionals**
15 cols · 10000 rows · 239.5 kB · 3.1k
- World time use, work hours and GDP**
7 cols · 329 rows · 207.6 kB · 734
- Impact of Covid-19 on Employment - ILOSTAT**
9 cols · 283 rows · 11.1 kB · 2.6k
- Annual Working Hours Dataset (1870-1970)**
4 cols · 3470 rows · 27.1 kB · 476

Global Remote Work & Wellbeing Dataset.

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

Why is this dataset relevant for your task?

Utility: The dataset includes relevant attributes such as Daily_Working_Hours, Stress_Level, Sleep_Duration, and Work_Life_Balance_Satisfaction, which can help in evaluating the effects of remote work on quality of life.

Limitation: The dataset does not specify a time period or geographical location, as it is described as a comprehensive dataset without temporal or spatial constraints.

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.

[Show More](#)

Dataset Preview

Employee_ID	Daily_Working_Hours	Screen_Time	Meetings_Attended	Emails_Sent	Productivity_Score	Stress_Level	Physical_Activity_Steps	Sleep_Duration	Work_Life_Balance_Satisfaction
---	---	---	---	---	---	---	---	---	---
E00001	7.0	5.6	5	30	3	5	11501	5.8	4
E00002	11.6	5.3	1	30	5	6	5742	8.7	2
E00003	9.9	4.2	3	21	8	4	4852	4.7	1
E00004	8.8	7.3	4	99	1	9	11928	4.3	2
E00005	5.2	6.3	4	87	2	3	7665	7.9	7
E00006	5.2	9.1	6	48	7	2	10325	6.0	2
E00007	4.5	3.2	7	88	1	3	14744	6.3	4
E00008	10.9	7.5	2	27	5	5	2502	8.7	5
E00009	8.8	8.3	5	29	10	1	13911	6.9	3

Keyword Search as Sensemaking

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

A

Getting started? ▲

Answer a few questions to help you get started and brainstorm ideas for your task.

1. Do you have a specific task in mind, or are you exploring available options?

I have a specific task

I am exploring

What is the primary goal of your task?

Train a classifier

Train a regression model

Supervised learning

Unsupervised learning

Visualization

LLM pretraining

LLM finetuning

Question-Answering

Not sure yet

2. What do you specifically want to do? Provide keywords or a sentence on the task you're interested in.

datasets indicating quality of life before, during, and after the COVID-19 pandemic

Get Started

Keyword Search as Sensemaking

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

The screenshot displays the DataScout interface for a dataset search. The main section is titled "Dataset Search Query" and includes a "Task Specifications" box with the query: "Analyze the impact of the pandemic on remote work and work-life balance". Below this, there are "Suggestions to Refine your Search Query:" with three suggestions, each with a plus icon. The "Filters (0)" section shows a search bar with "No metadata filters added." and a "Search using your own Column Concept" section with a search bar. The "Smart Filter by Column Concept:" section lists filters: stress, hours, vacations, employment, and remote. The "Top Granularity Filters" section lists filters: Country (5), Day (2), and Year (2). The "Top Dataset Results" section on the right shows a list of 11 datasets, with the top 10 visible. The datasets are: 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. Online Learning Data, 7. Predict if people prefer WFH vs WFO post Covid-19, 8. Global Unemployment Dataset, 9. World time use, work hours and GDP, 10. Impact of Covid-19 on Employment - ILOSTAT, and 11. Annual Working Hours Dataset (1870-1970). The interface is annotated with orange arrows and labels: an arrow points to the "Task Specifications" box, an arrow points to the "Smart Filter by Column Concept:" section, and two large orange circles labeled "C" and "D" are positioned to the right of the filter sections.

Dataset Search Query

Task Specifications

task: 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

Top Granularity Filters

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

Top Dataset Results

Showing 1 to 11 of 11 datasets.

- Global Remote Work & Wellbeing Dataset.**
10 cols - 10000 rows - 133.3 kB - 179
- Remote Work USA (COVID-19)**
24 cols - 3147 rows - 2.0 MB - 81
- Remote Work Productivity**
5 cols - 1000 rows - 6.7 kB - 3.3k
- COVID-19 on Working Professionals**
15 cols - 10000 rows - 255.2 kB - 2.2k
- Impact of COVID-19 on Working Professionals**
15 cols - 10000 rows - 239.5 kB - 3.1k
- Online Learning Data**
21 cols - 214 rows - 4.6 kB - 164
- Predict if people prefer WFH vs WFO post Covid-19**
19 cols - 207 rows - 2.9 kB - 1.5k
- Global Unemployment Dataset**
7 cols - 50 rows - 70.3 kB - 78
- World time use, work hours and GDP**
7 cols - 329 rows - 207.6 kB - 734
- Impact of Covid-19 on Employment - ILOSTAT**
9 cols - 283 rows - 11.1 kB - 2.6k
- 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

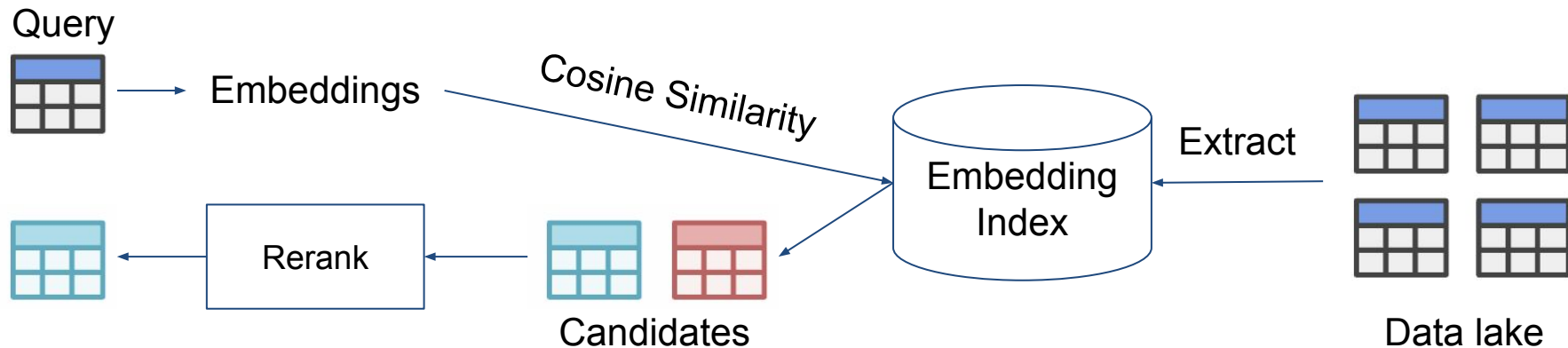
Search by using a table or distribution as the query

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

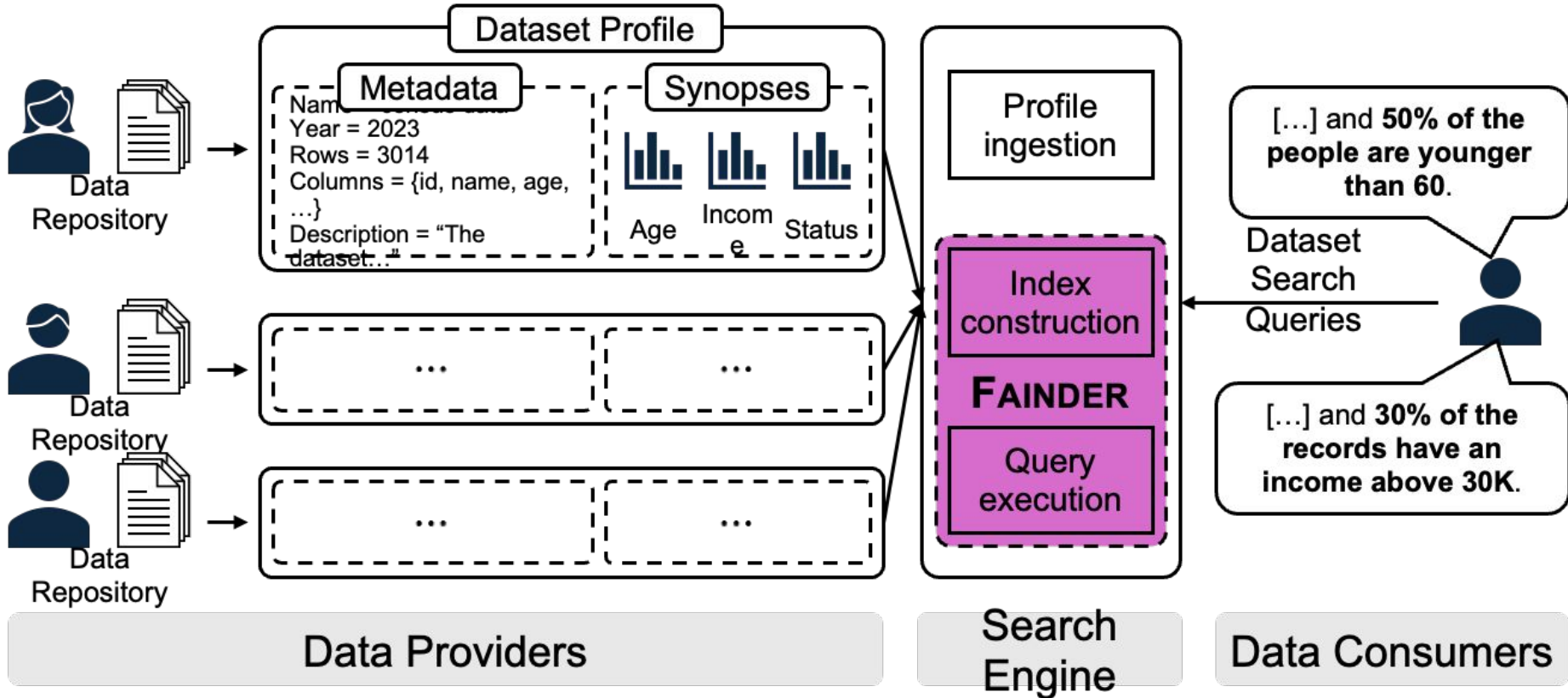
Rank based on

- 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

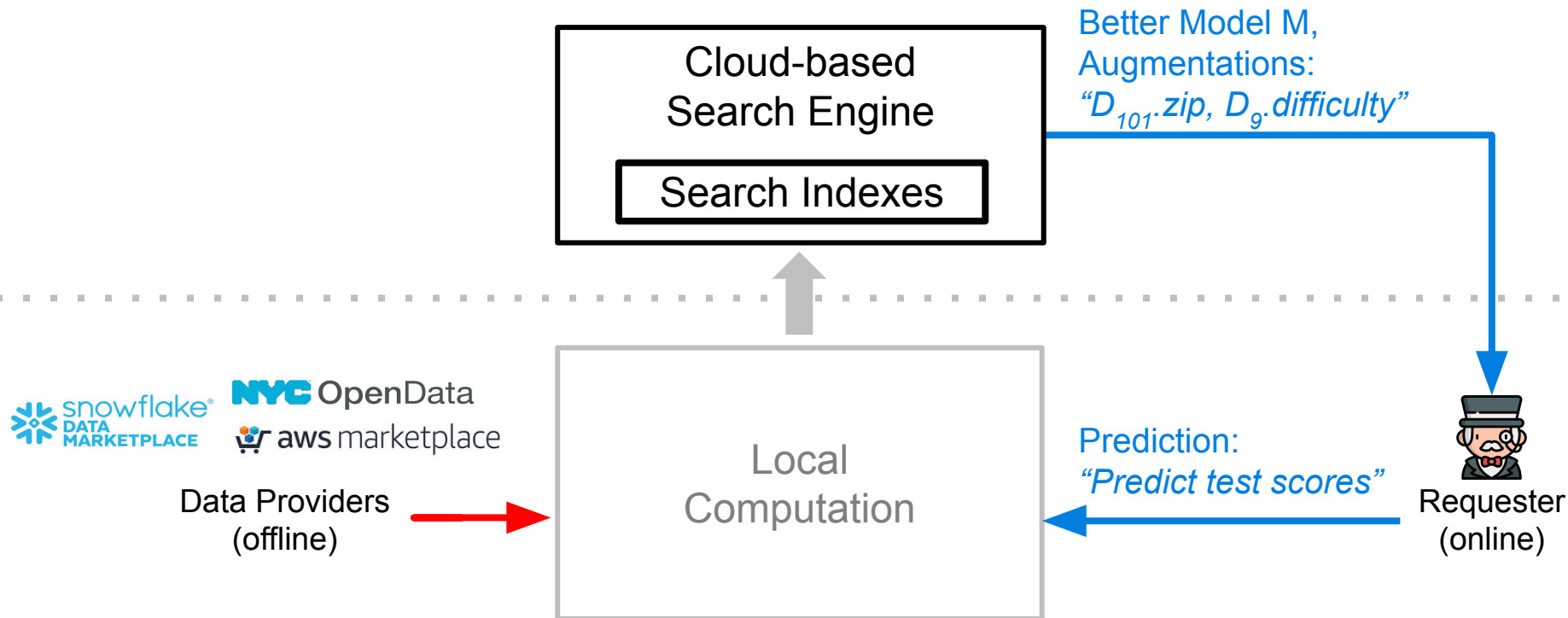
Kitana: A Data-as-a-Service Platform. Zach Huang²³

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 $A(D)$

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

D	1	4
5		

Dataset
Repository

1	2	3	4	5
---	---	---	---	---

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 α Greedy Search

if eval(apply A to D) is best so far

Keep α in A

Expensive!

return best A

Materialize A(D)

Retrain & Cross-validate

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

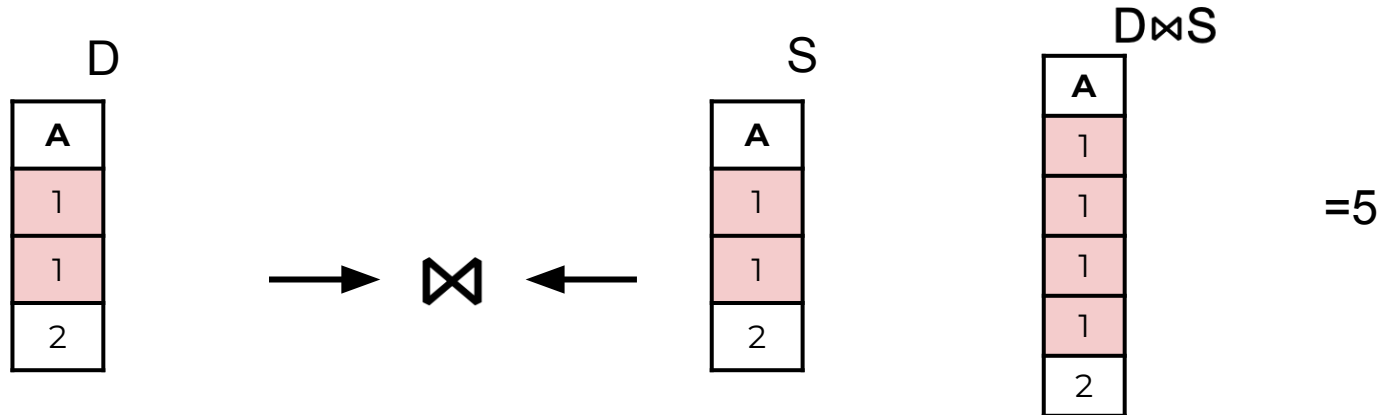
$D \bowtie S$

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

=5

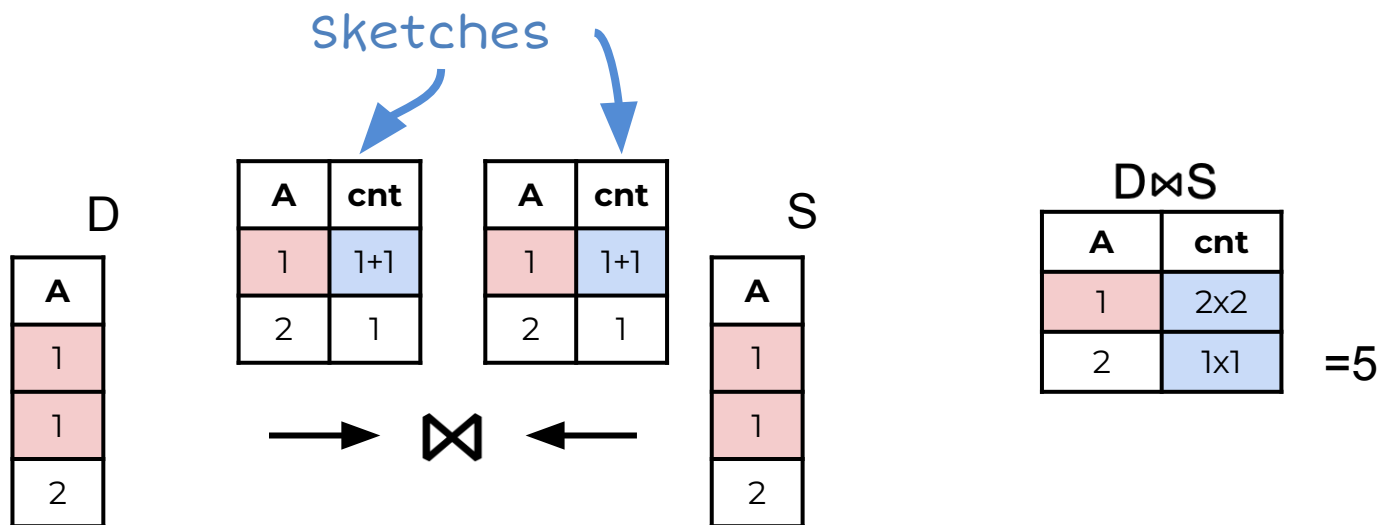
Sketches: Count($D \bowtie S$)

Optimization: drop irrelevant columns



Sketches: Count($D \bowtie S$)

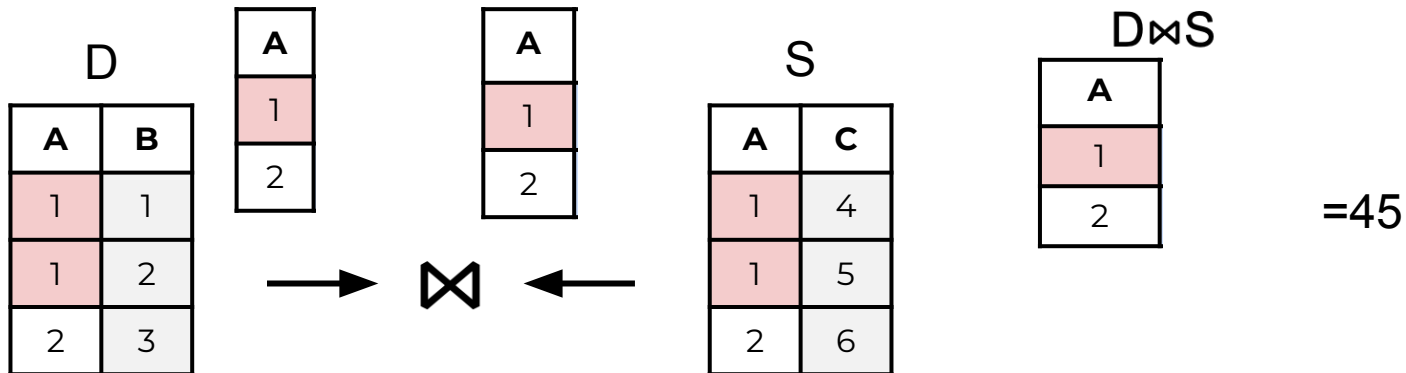
Optimization: sufficient statistics



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

Optimization: sufficient statistics

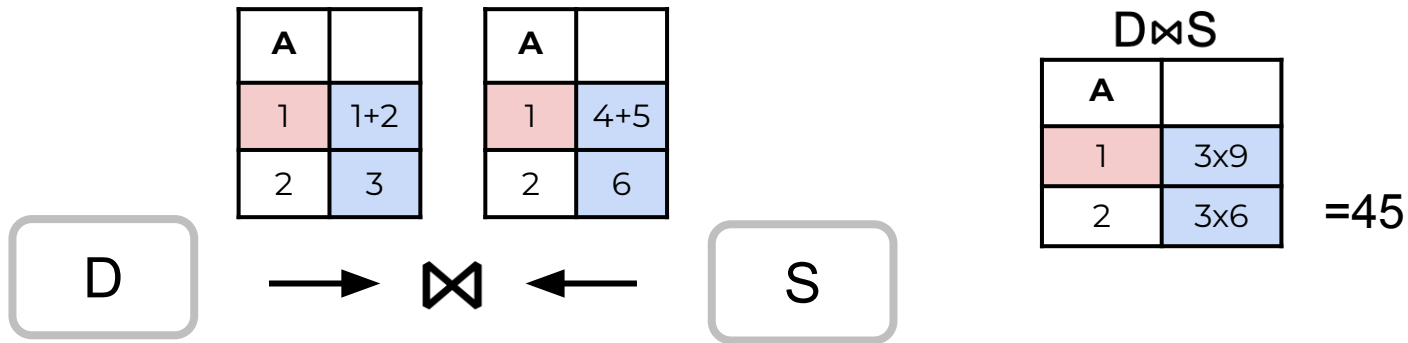
Sketches defined for common stats, ML models.



Sketches: $\text{trainAndEval}(D \bowtie S)$

Optimization: sufficient statistics

Sketches defined for common stats, ML models.

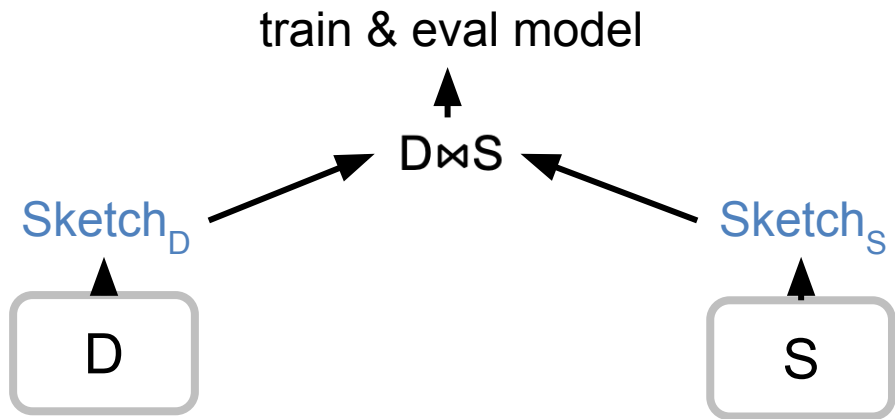


Sketches: $\text{train}(D \bowtie S)$

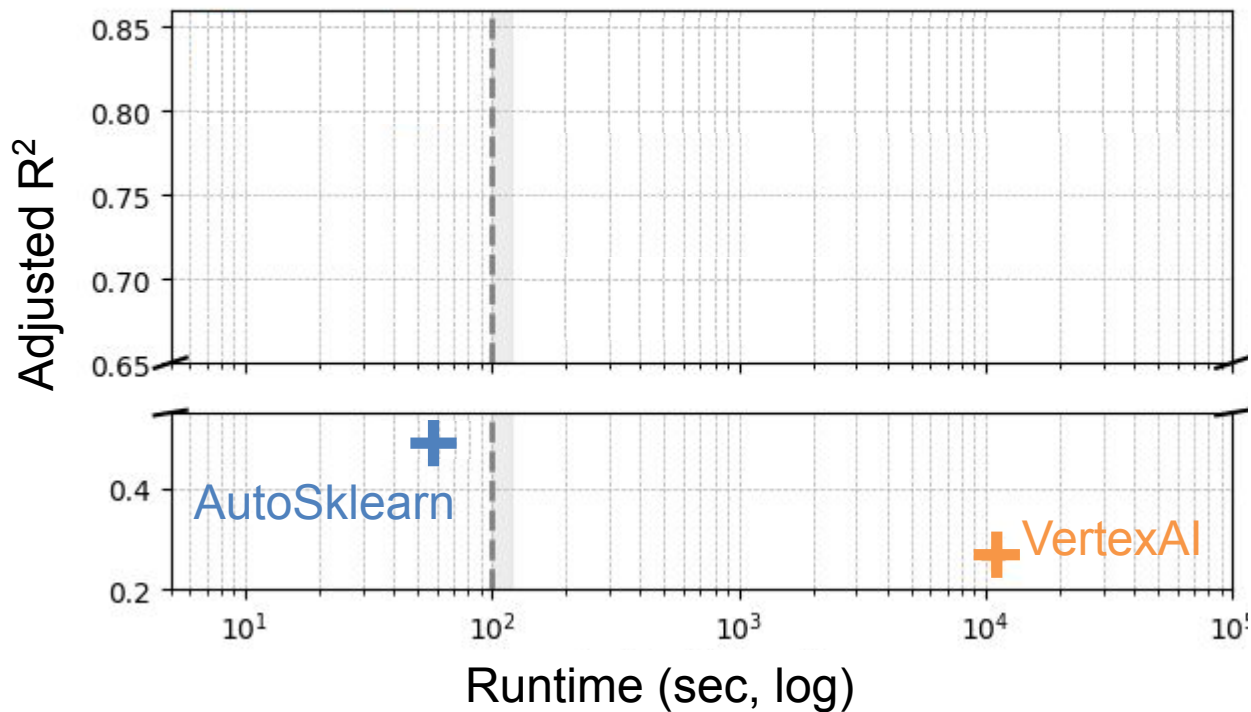
Optimization: sufficient statistics

Sketches defined for common stats, ML models.

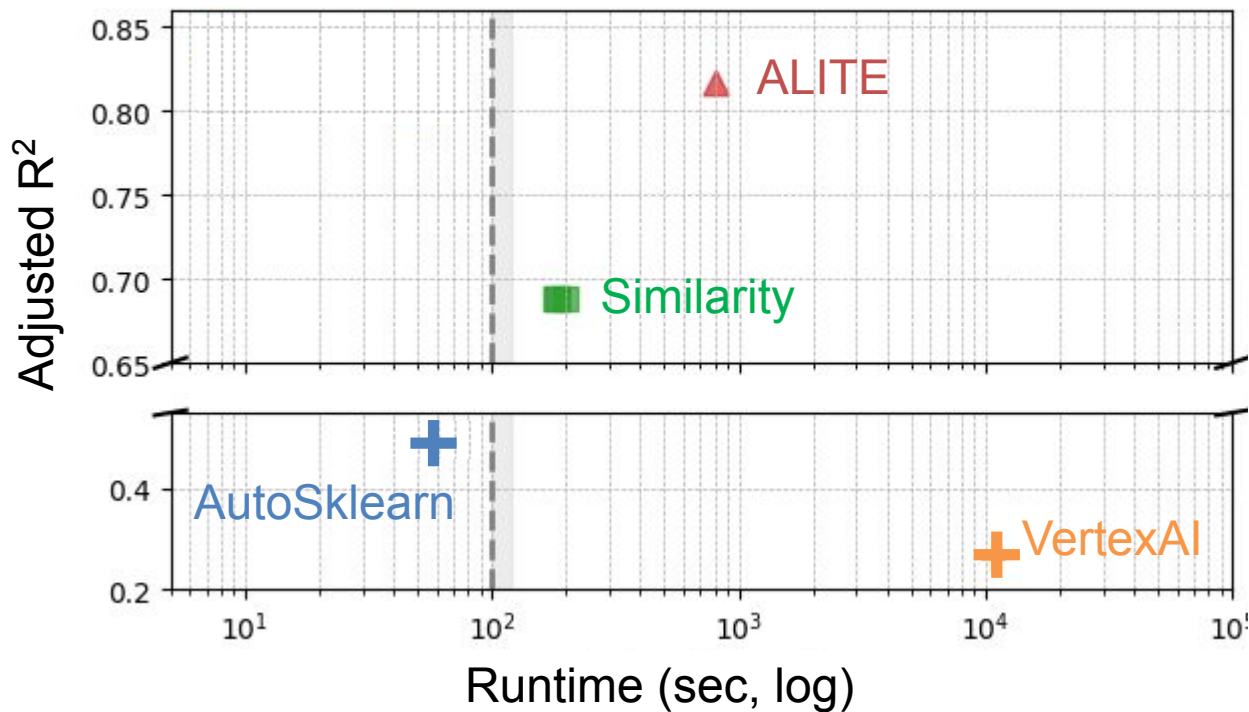
Linear Regression as a *proxy model* during search



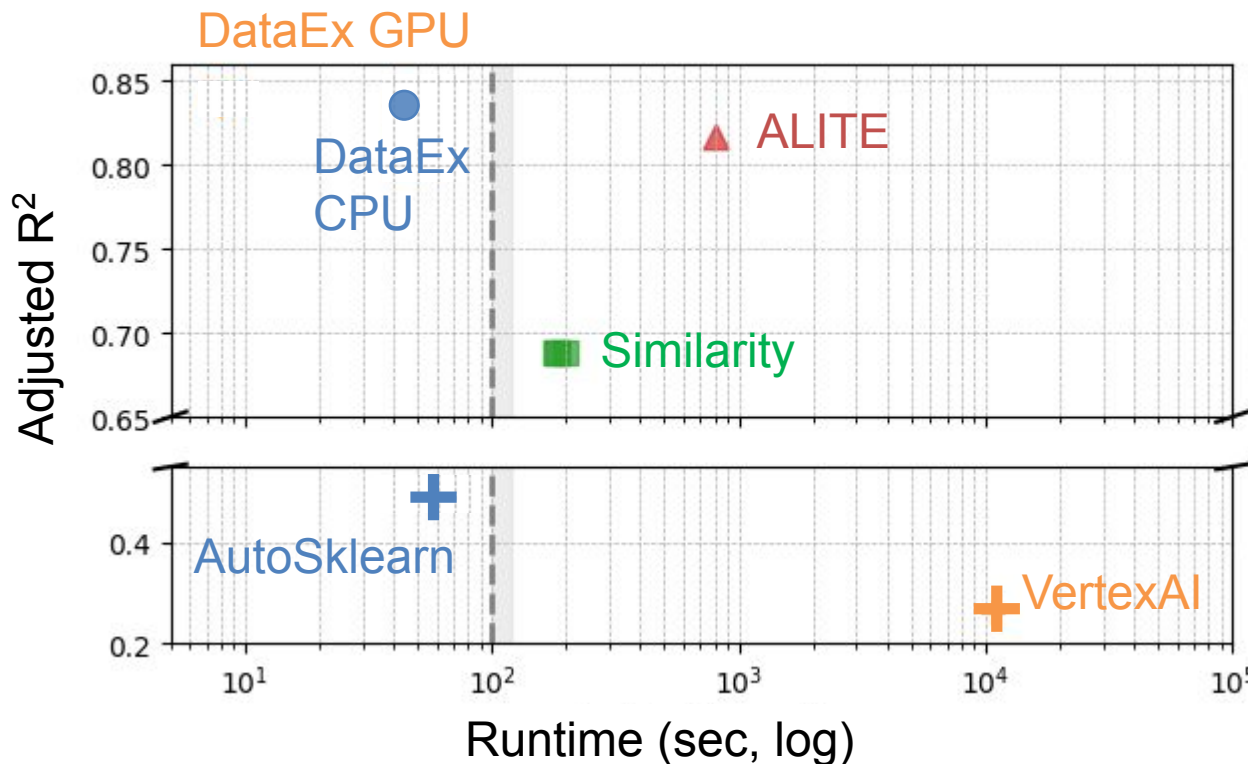
Evaluation on 8376 Kaggle Tables



Evaluation on 8376 Kaggle Tables

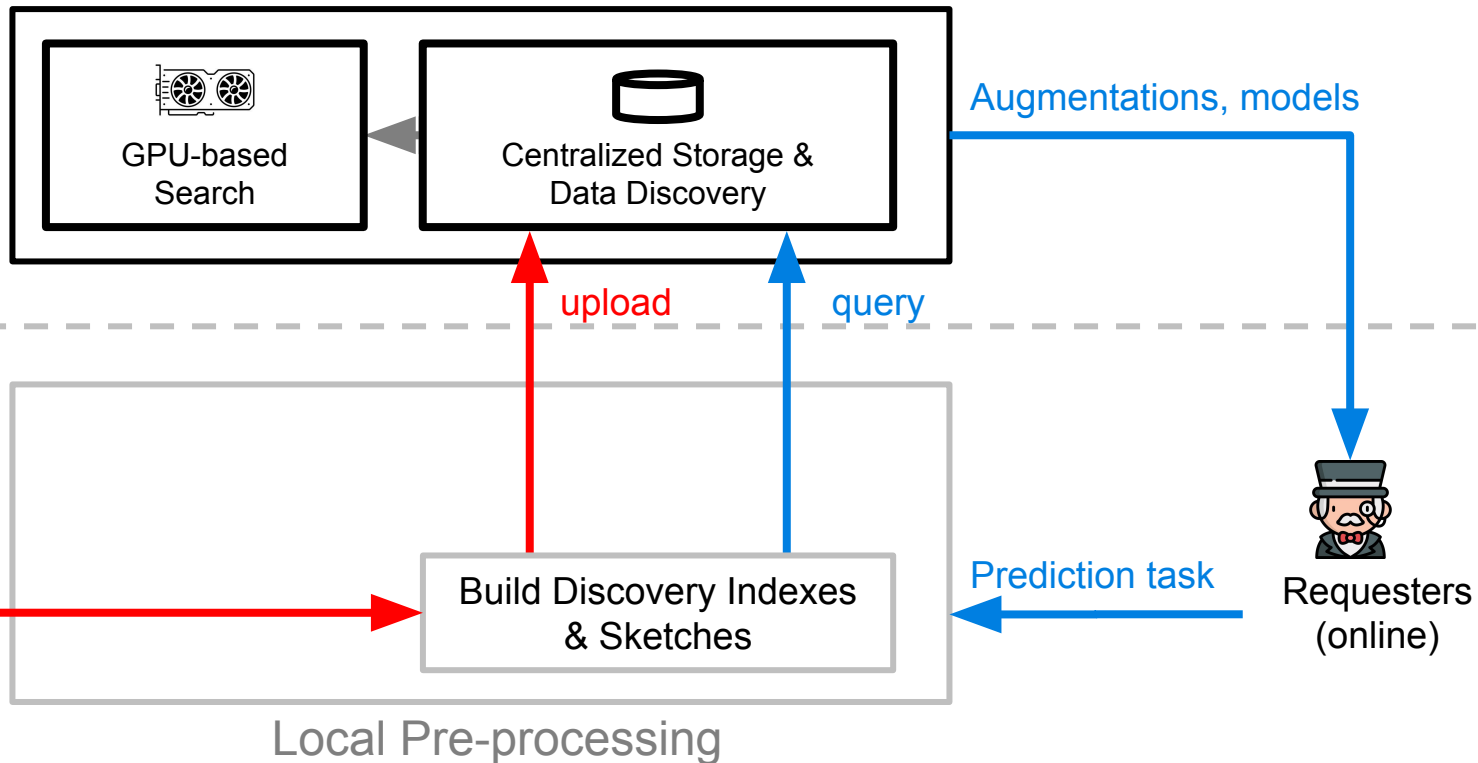


Evaluation on 8376 Kaggle Tables



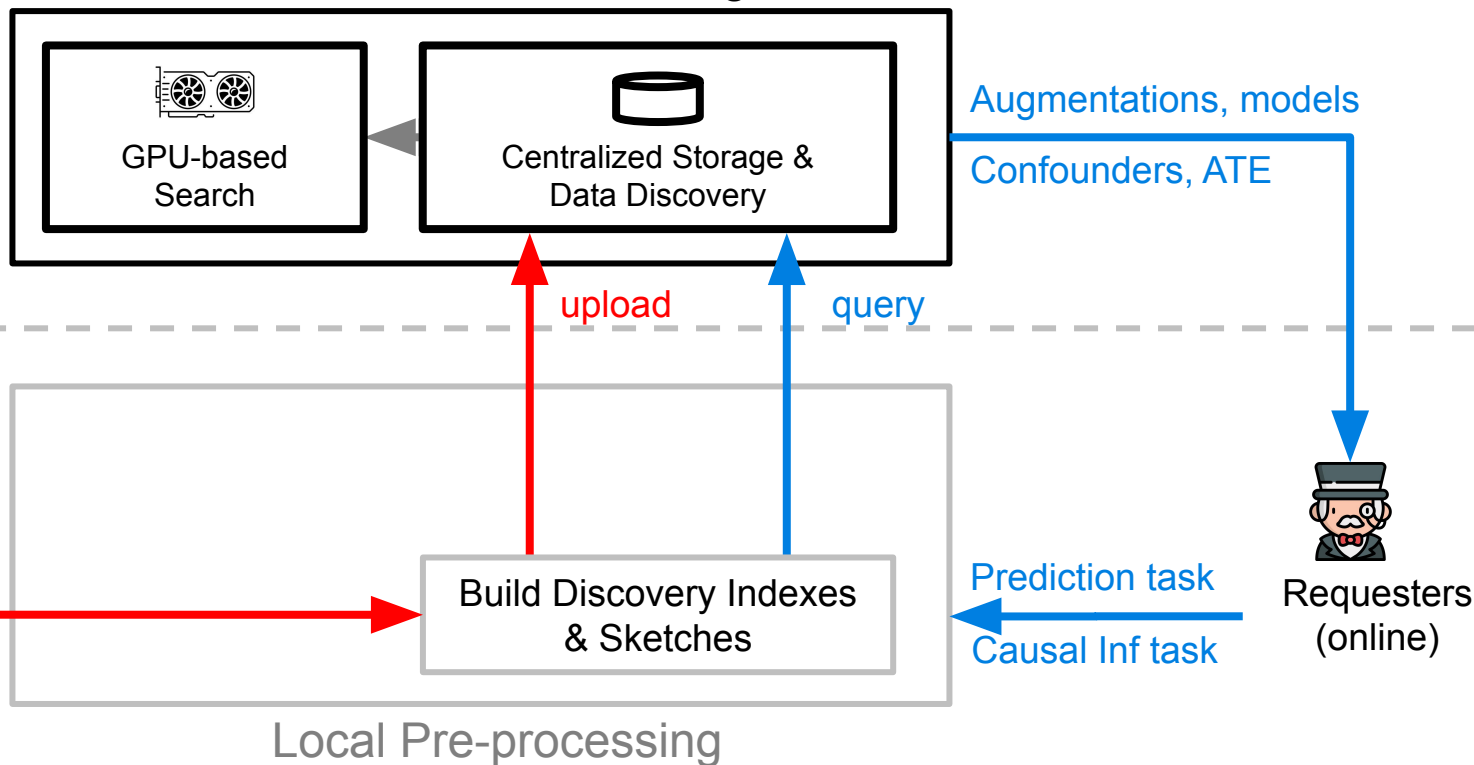
DataEx

Cloud Dataset Search Engine

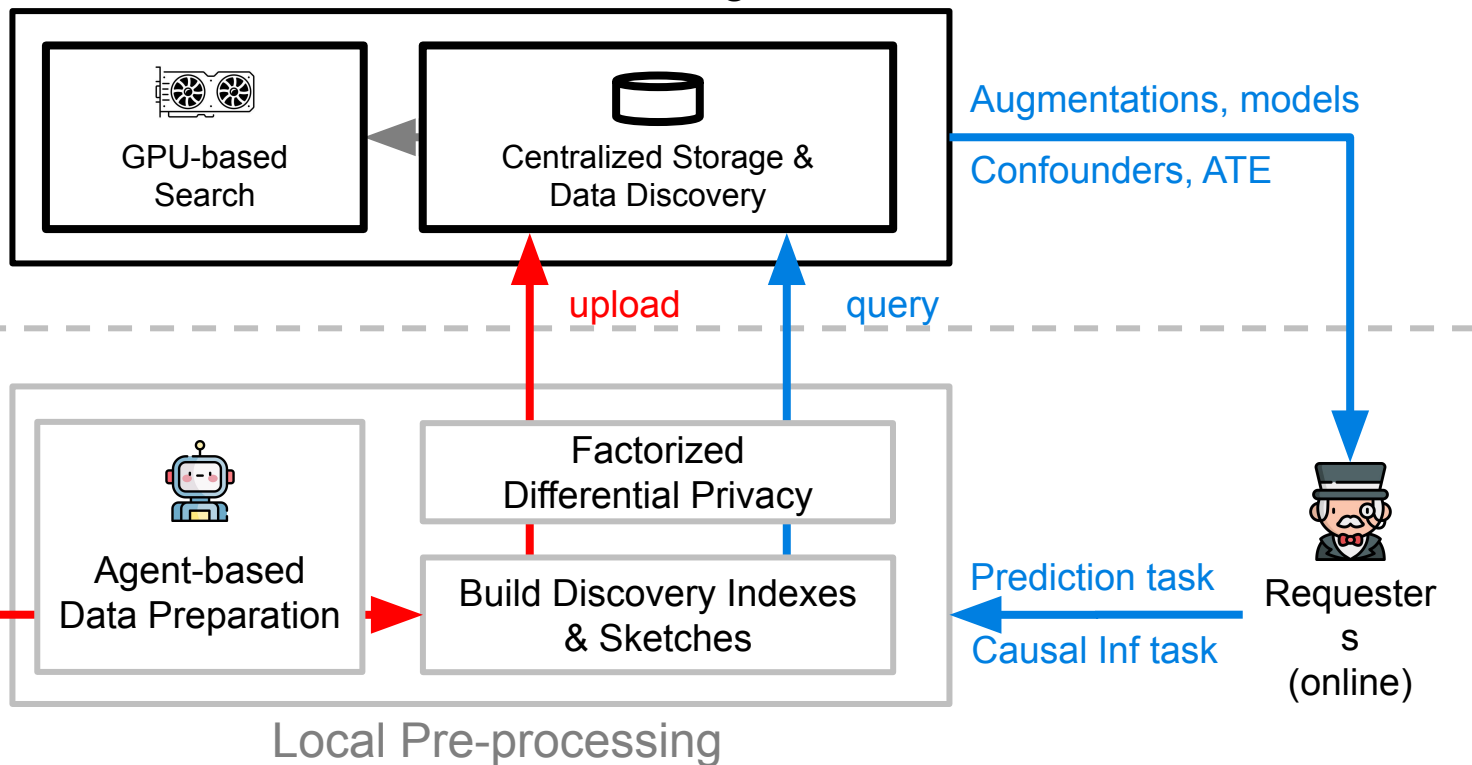


DataEx

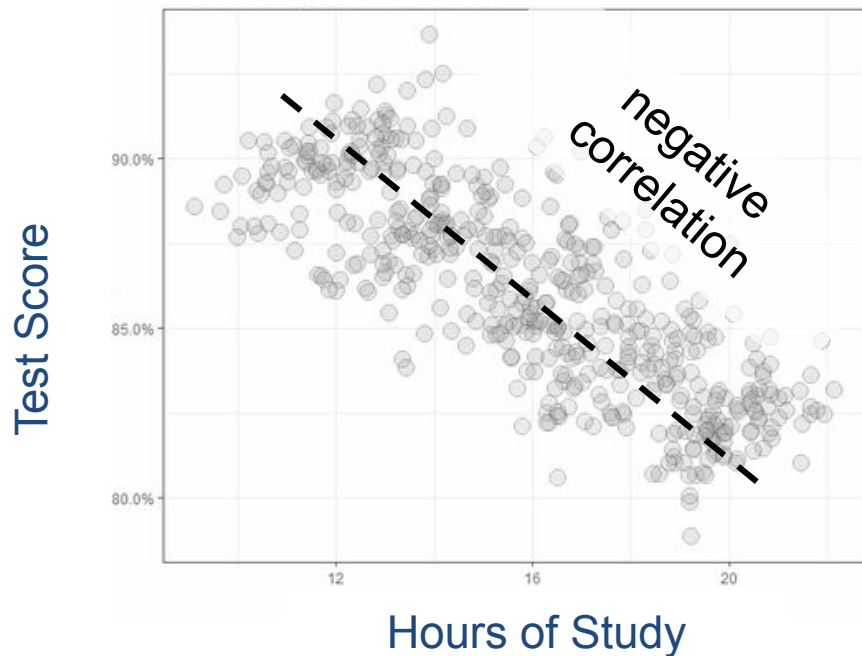
Cloud Dataset Search Engine



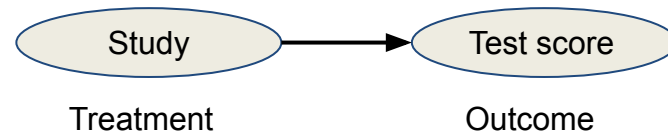
Cloud Dataset Search Engine



Confounders in Causal Analysis

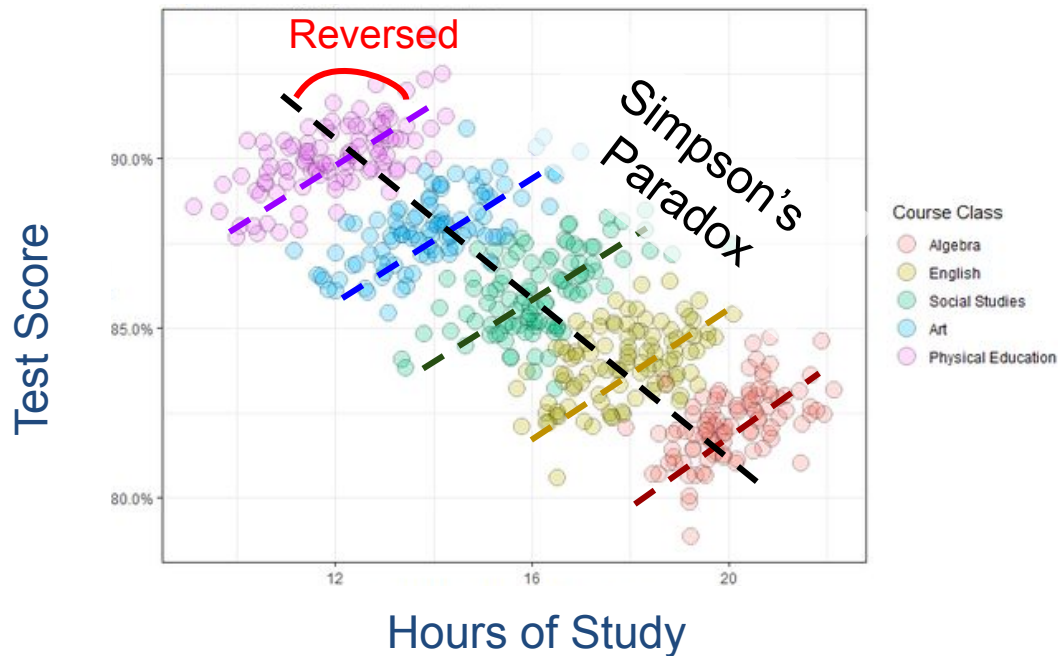


Causal Diagram

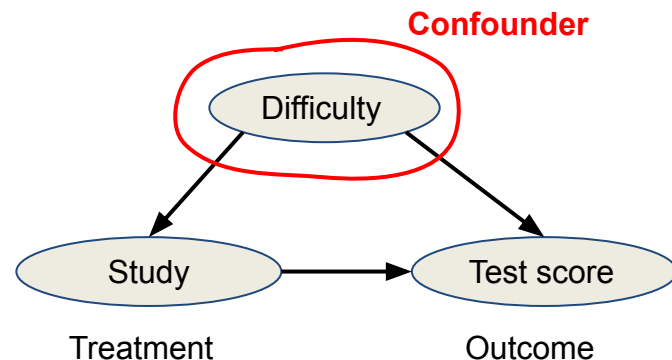


Studying causes poor grades?
Study $\rightarrow$ TestScores

Confounders in Causal Analysis



Causal Diagram



Why does the trend reverse?
Study → Test Scores

Confounders in Causal Analysis

User Query

Treatment Outcome

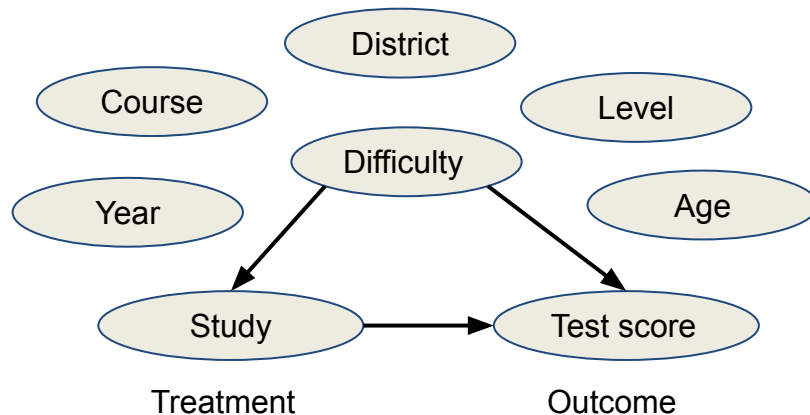
ID	Study	Score
1	15 hr	75
2	10 hr	90
3	20 hr	85

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

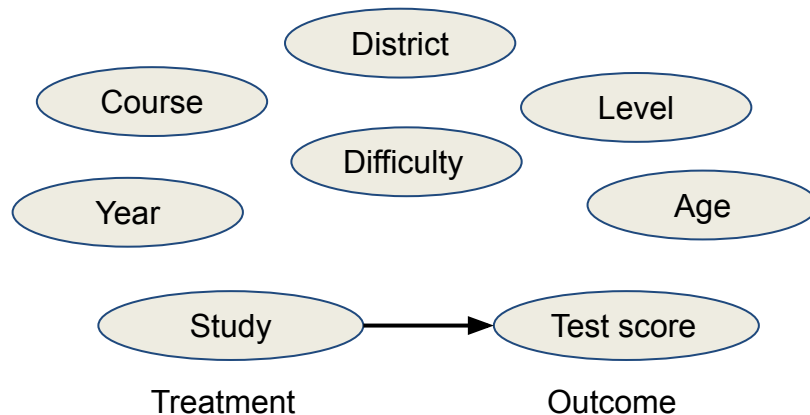
ID	Study	Score
1	15 hr	75
2	10 hr	90
3	20 hr	85

Data Repository

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

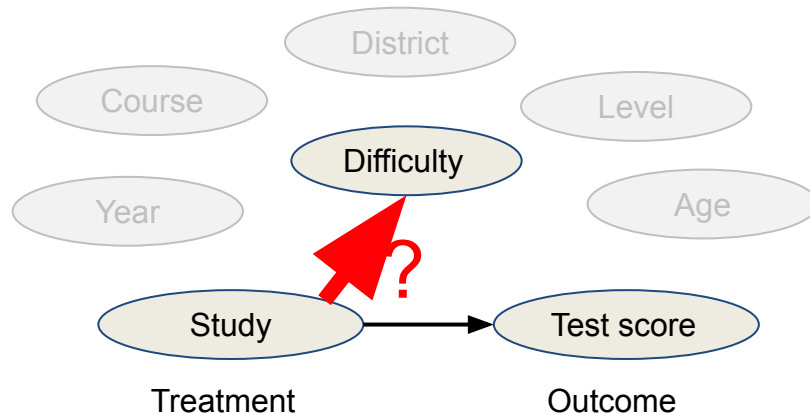
ID	District
1	1
2	2
3	3

...



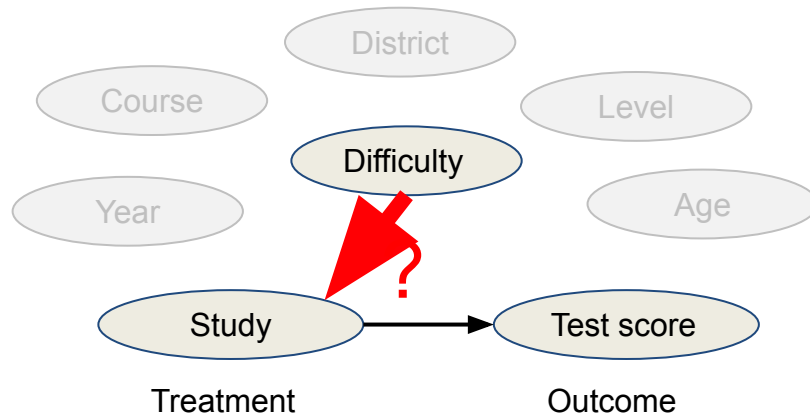
Confounders in Causal Analysis

Background: Bivariate Causal Discovery (BCD) estimates $\rightarrow$ or $\leftarrow$ edges from data



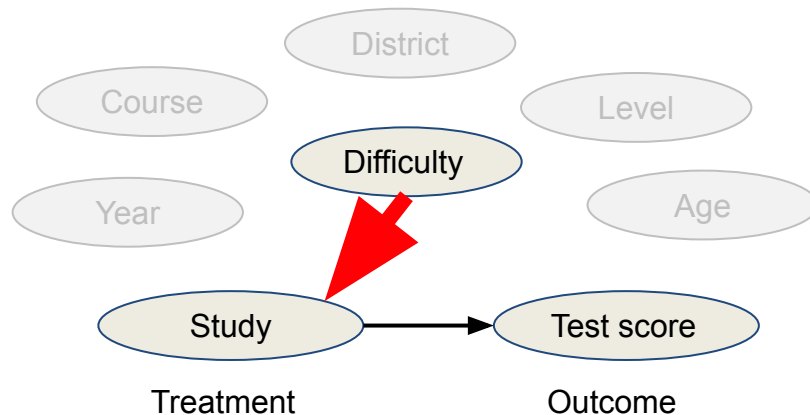
Confounders in Causal Analysis

Background: Bivariate Causal Discovery (BCD) estimates $\rightarrow$ or $\leftarrow$ edges from data



Confounders in Causal Analysis

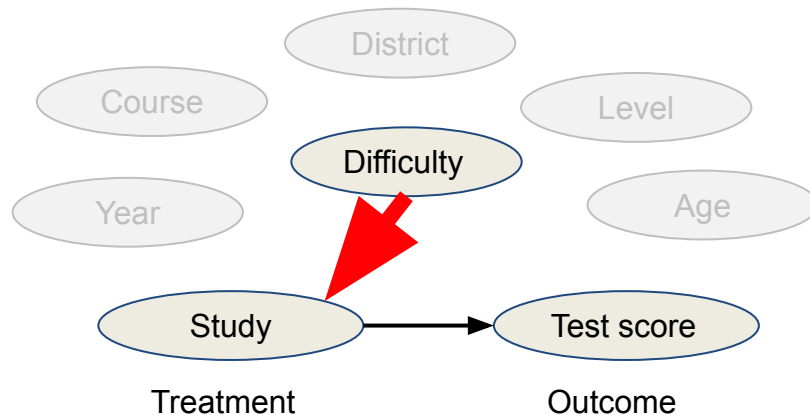
Background: Bivariate Causal Discovery (BCD) estimates $\rightarrow$ or $\leftarrow$ edges from data



Confounders in Causal Analysis

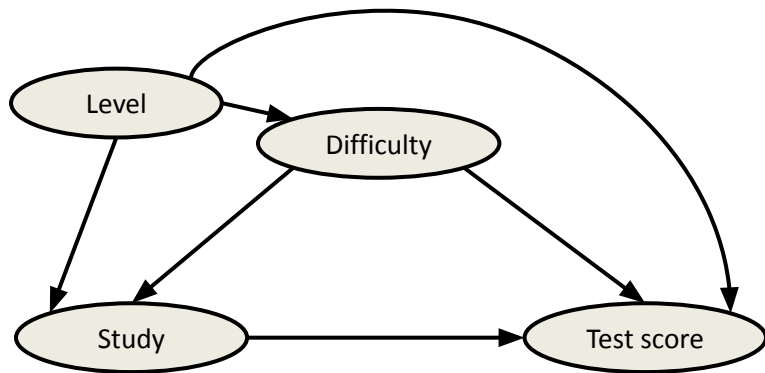
Background: Bivariate Causal Discovery (BCD) estimates $\rightarrow$ or $\leftarrow$ edges from data

Proof: existence of confounder reduces to BCD estimating “ $Ancestors \rightsquigarrow Treatment$ ”



Discovering Confounders with BCD

Building adjustment set for Study $\rightarrow$ TestScore

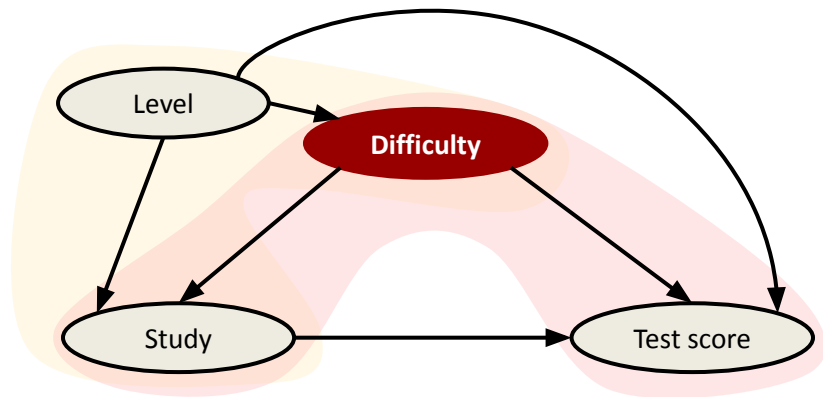


Key Observation:

treatment and outcome confounded: BCD will flag confounder $\rightarrow$ treatment.

Discovering Confounders with BCD

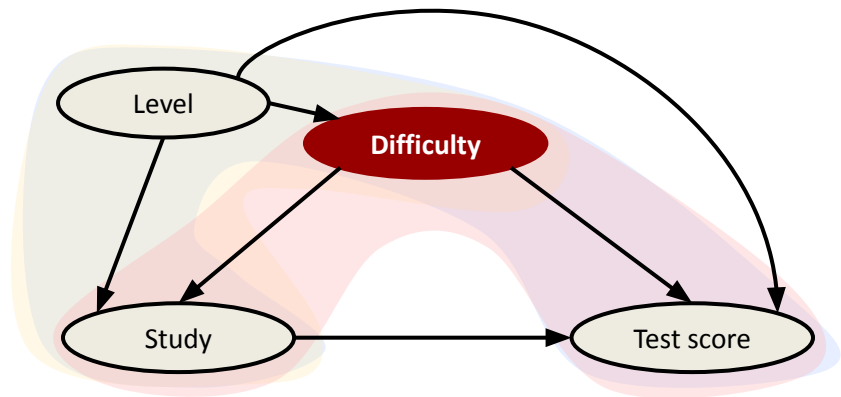
Building adjustment set for Study $\rightarrow$ TestScore



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

Discovering Confounders with BCD

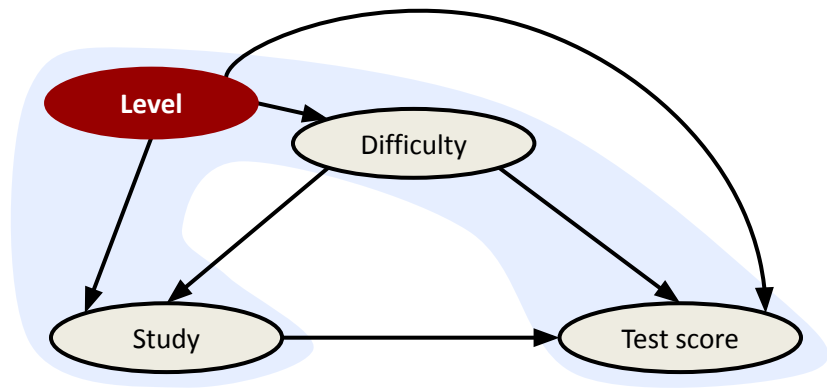
Building adjustment set for Study $\rightarrow$ TestScore



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

Discovering Confounders with BCD

Building adjustment set for Study $\rightarrow$ TestScore

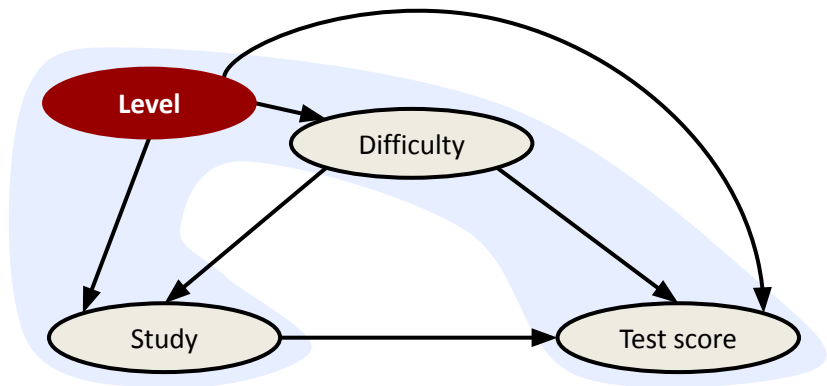


Key Insight: Level is a confounder, flagged by BCD

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

Discovering Confounders with BCD

Building adjustment set for Study $\rightarrow$ 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

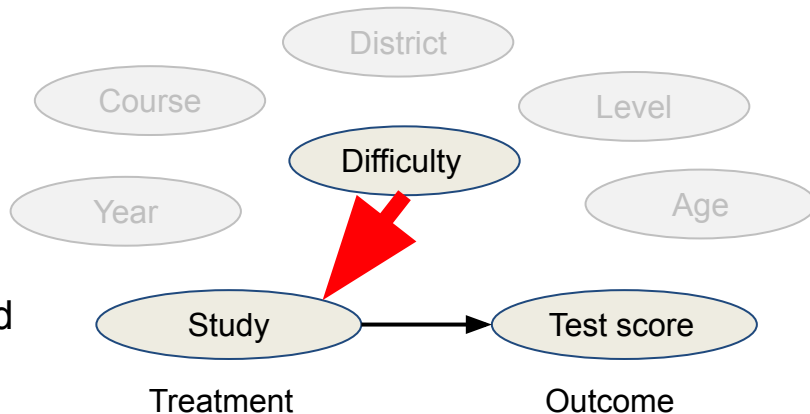
Background: Bivariate Causal Discovery (BCD) estimates $\rightarrow$ or $\leftarrow$ edges from data

Proof: existence of confounder reduces to BCD estimating “ $Ancestors \sim Treatment$ ”

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

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

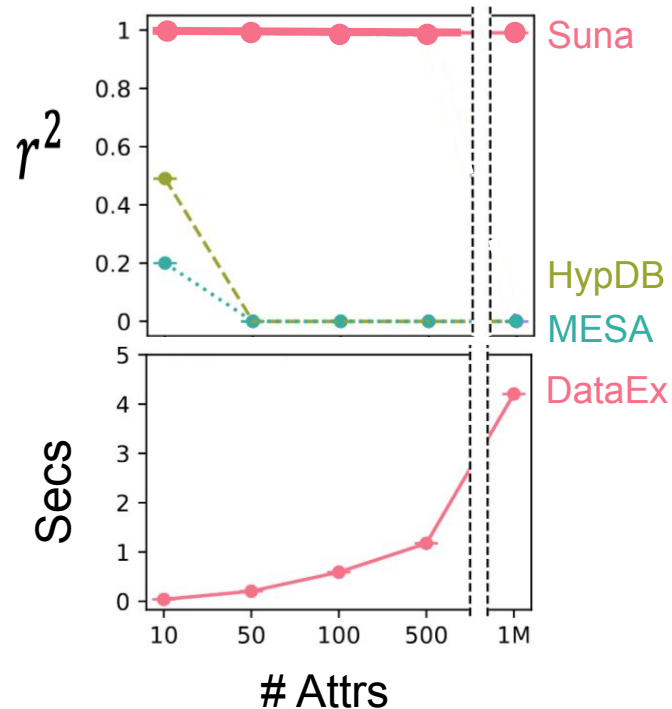


Experimental Results

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

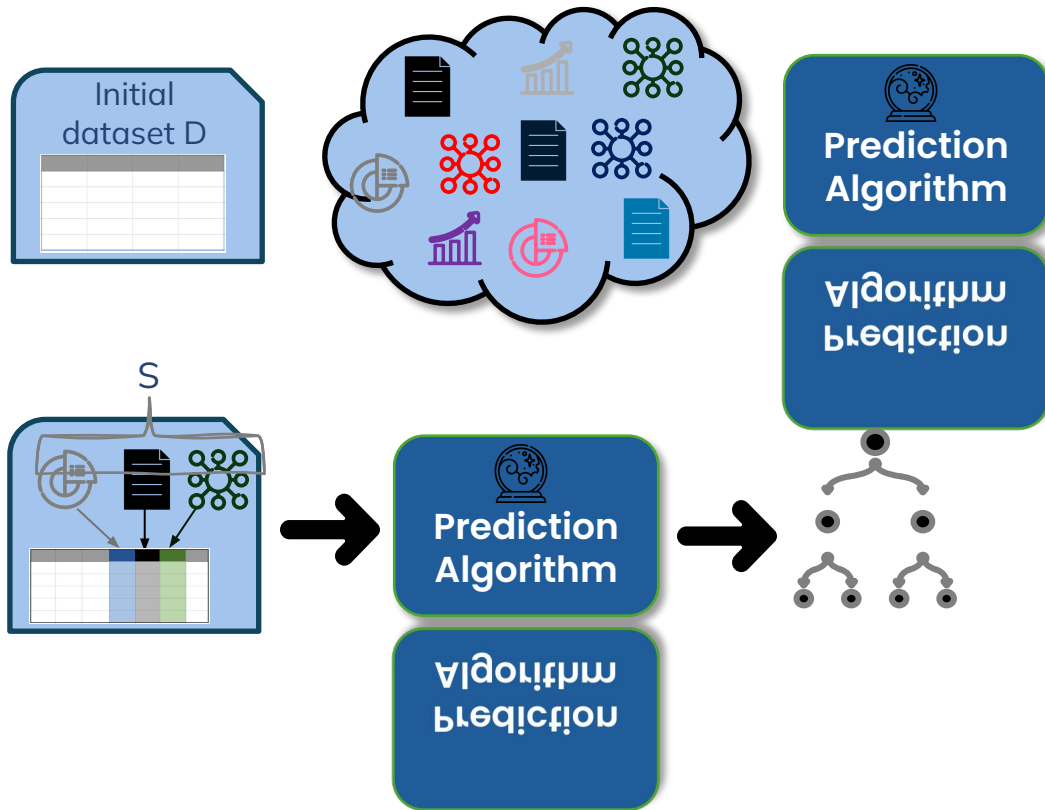
OBJECTIVE: $\max \text{utility}_t(D \bowtie S)$

CONSTRAINT:

$$S \subseteq \Gamma$$

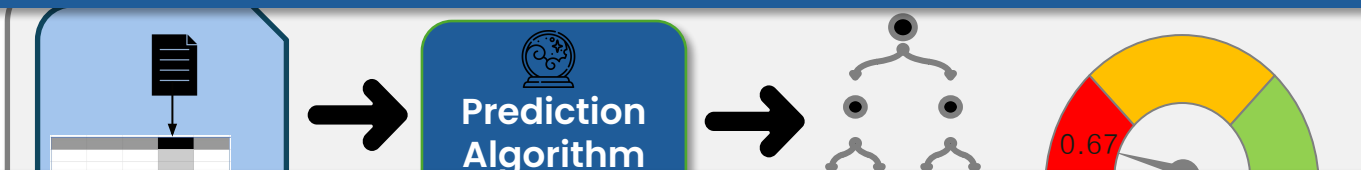
$$|S| \leq k \text{ (a constant)}$$

S is minimal

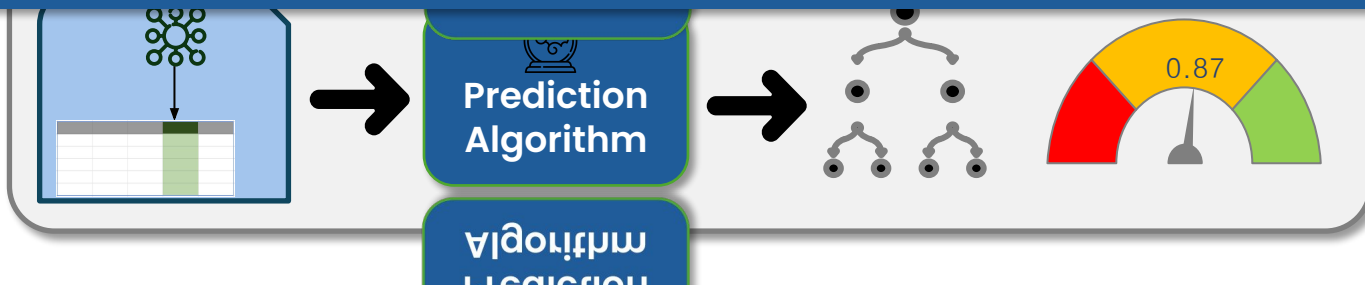


How to solve the problem?

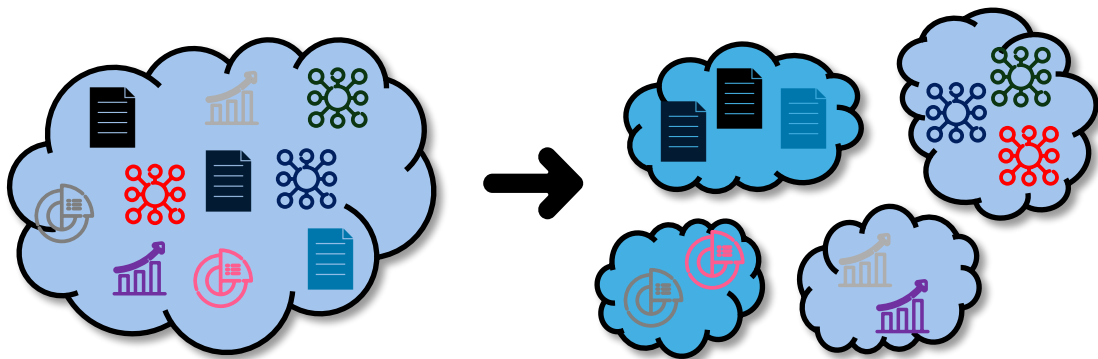
⚠ Requires n^k queries! Infeasible when n is in the order of millions



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:

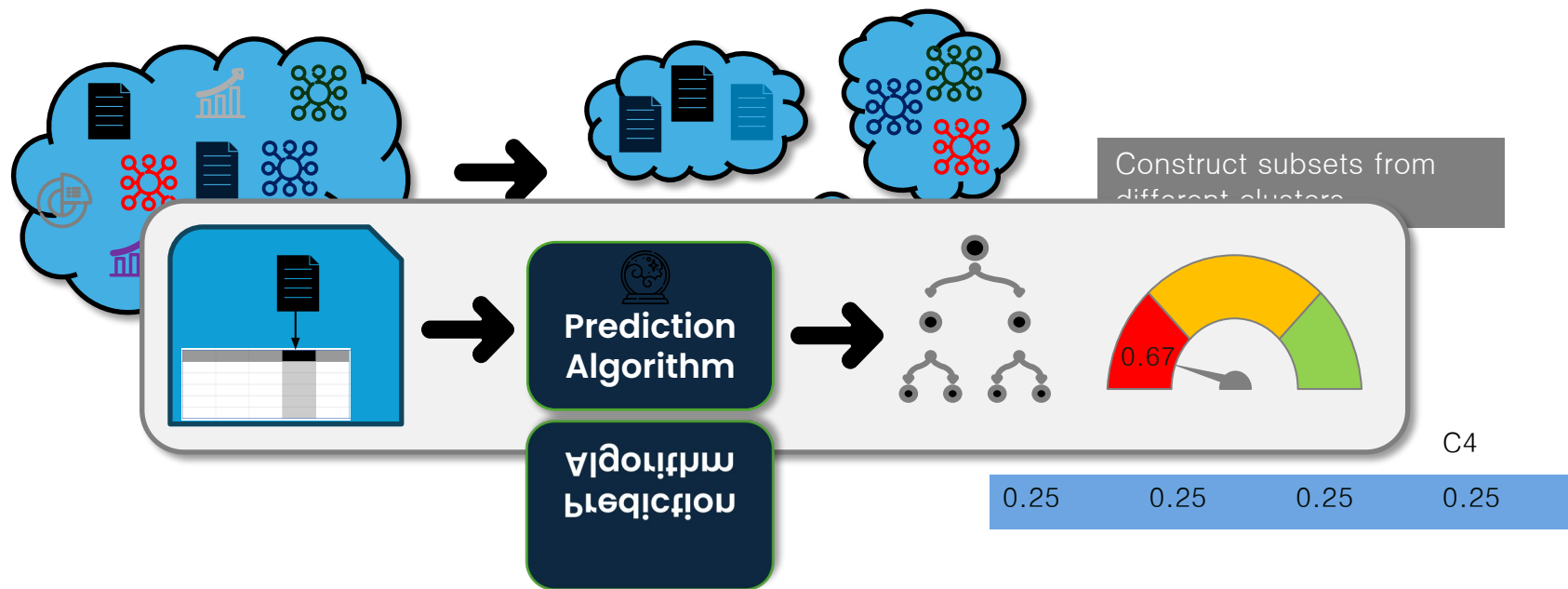
Should I sample more datasets from cluster C_i ?

OR

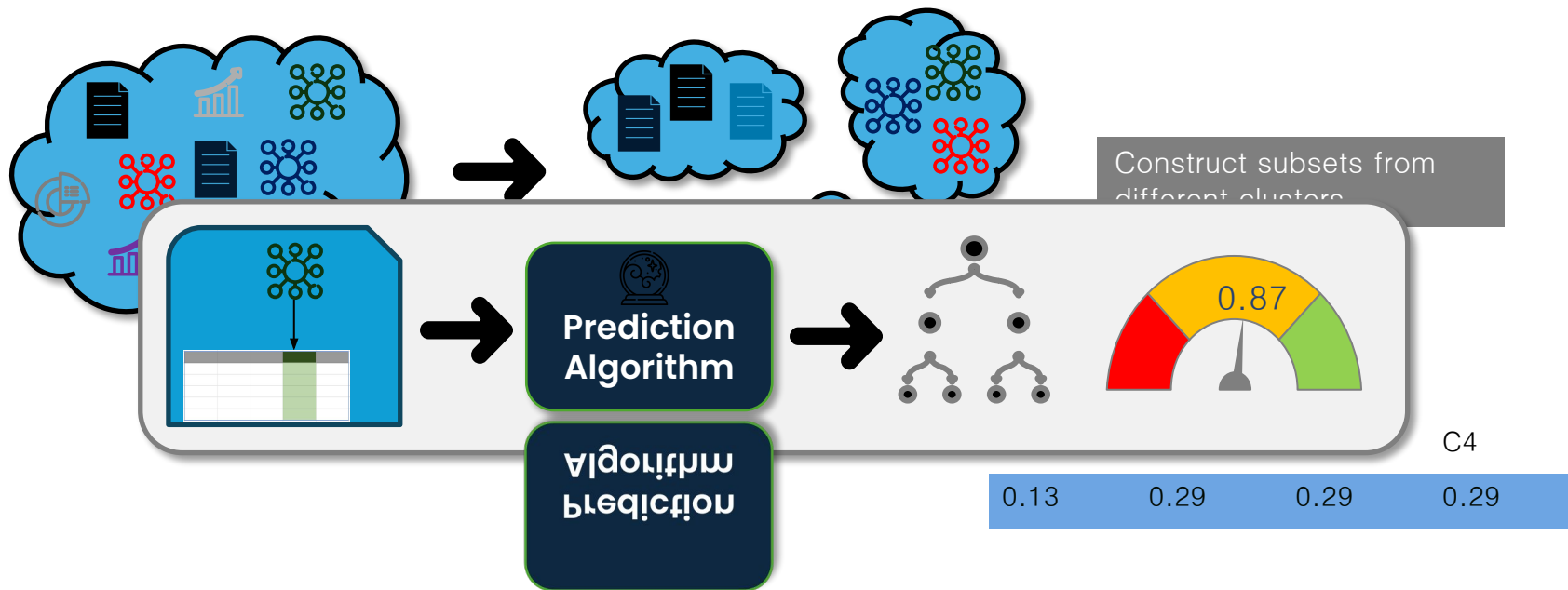
Should I explore different clusters?

SOLUTION: Bandit-based approach

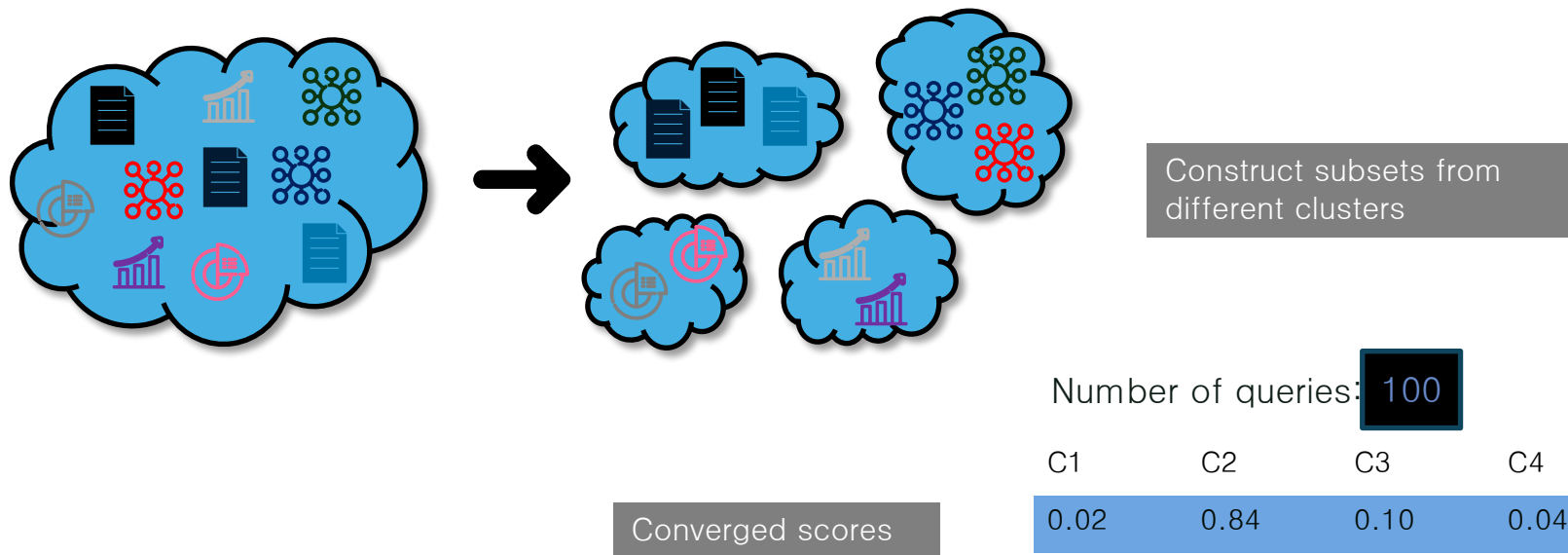
Approach 1: Diversify the Search Process



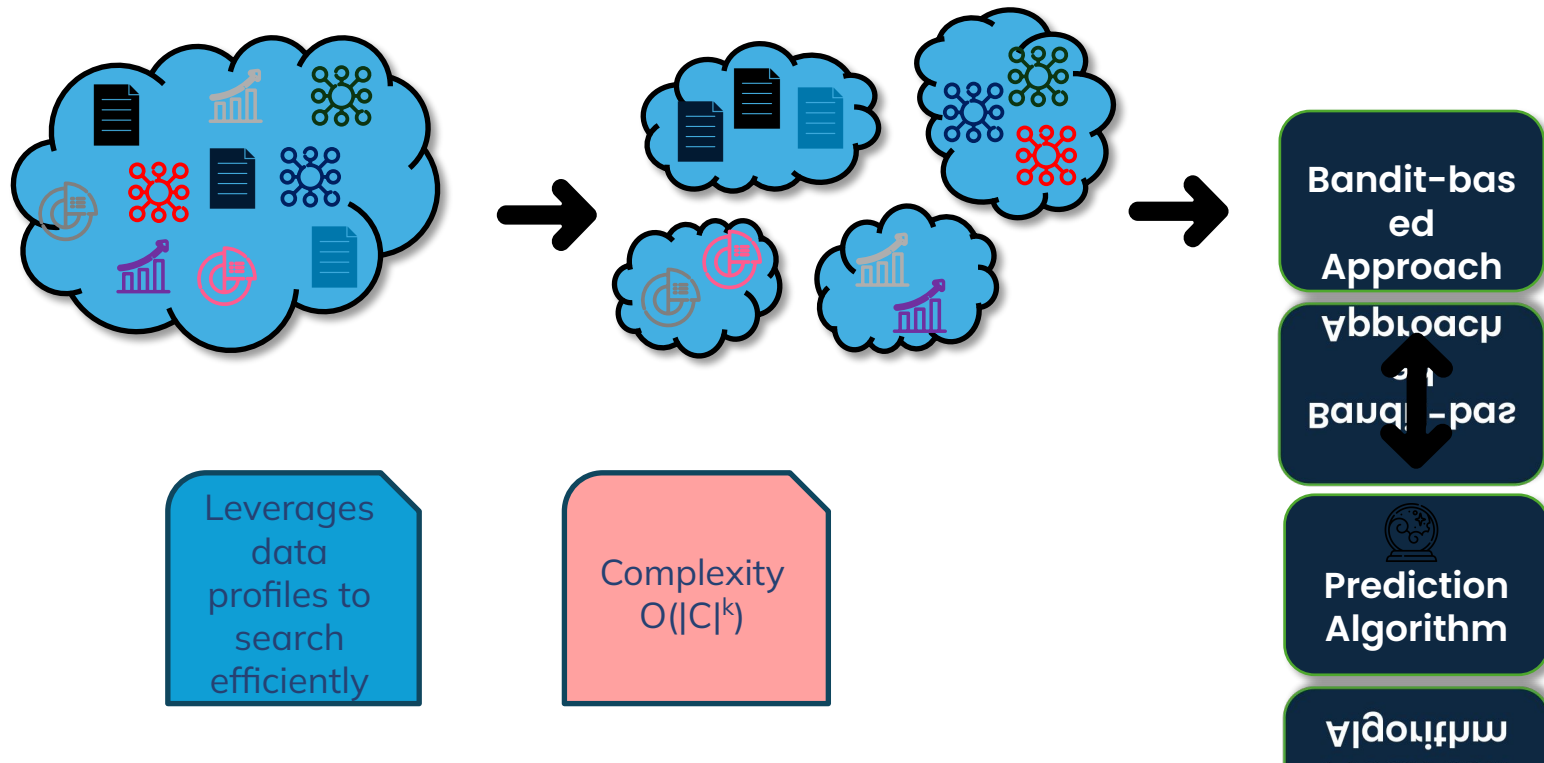
Approach 1: Diversify the Search Process



Approach 1: Diversify the Search Process



Approach 1: Diversify the Search Process



Approach 2: Leverage Monotonicity of Utility Metric

- **Monotonicity:** Easy to guarantee

$$u(D \cup T_2) \geq u(D \cup T_1) \quad \forall T_1 \subseteq T_2$$

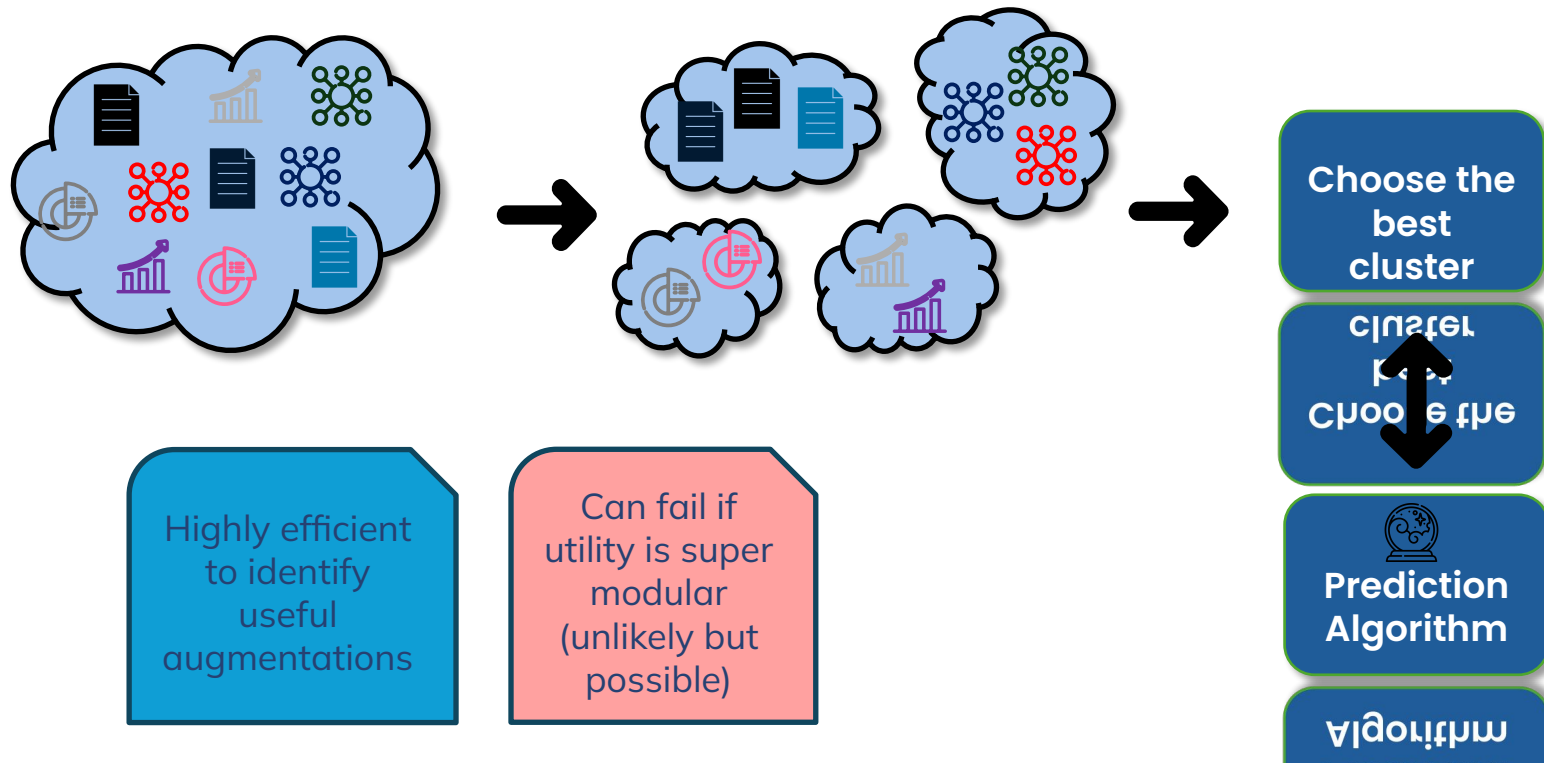
- What if the utility is **submodular** too?

- Diminishing returns property:

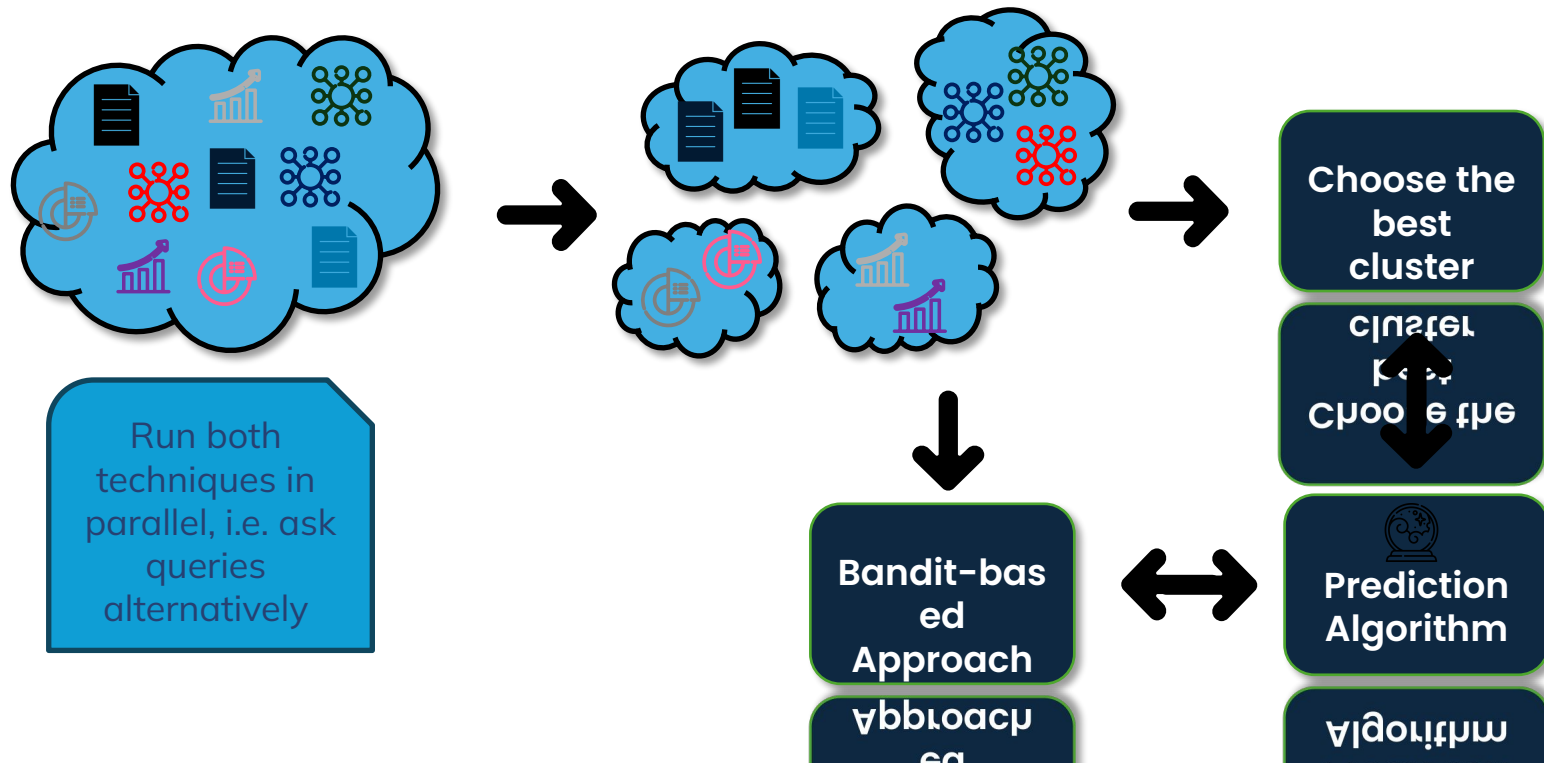
$$u(T_1 \cup \{X\}) - u(T_1) \geq u(T_2 \cup \{X\}) - u(T_2) \quad \forall T_1 \subseteq T_2$$

Solution: Greedily choose the best augmentation

Approach 2: Leverage Monotonicity of Utility Metric



Final Approach: Combining All Ideas



Privacy Challenges Are Everywhere!



Query specification
Privacy protection

Data Market/Data Discovery

Query interface
Latency, Scalability
Privacy protection

Stats

Stats

Stats

Data

Data

Data

Data acquisition
Data preparation
Privacy protection

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`

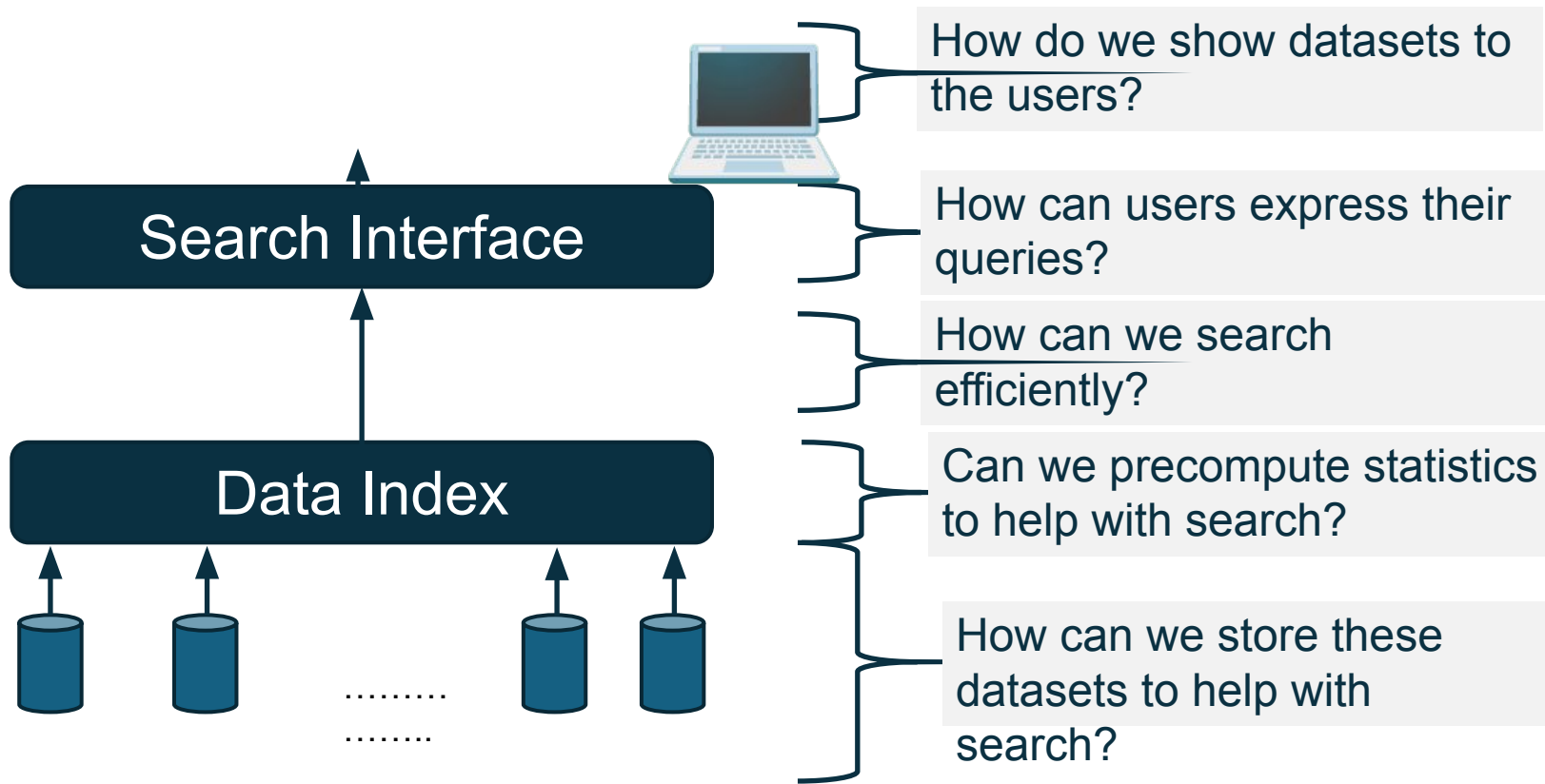
- ☐ Yes, my data must contain this attribute
- ☐ No, my data should not contain this attribute
- ☒ Does not matter

Submit

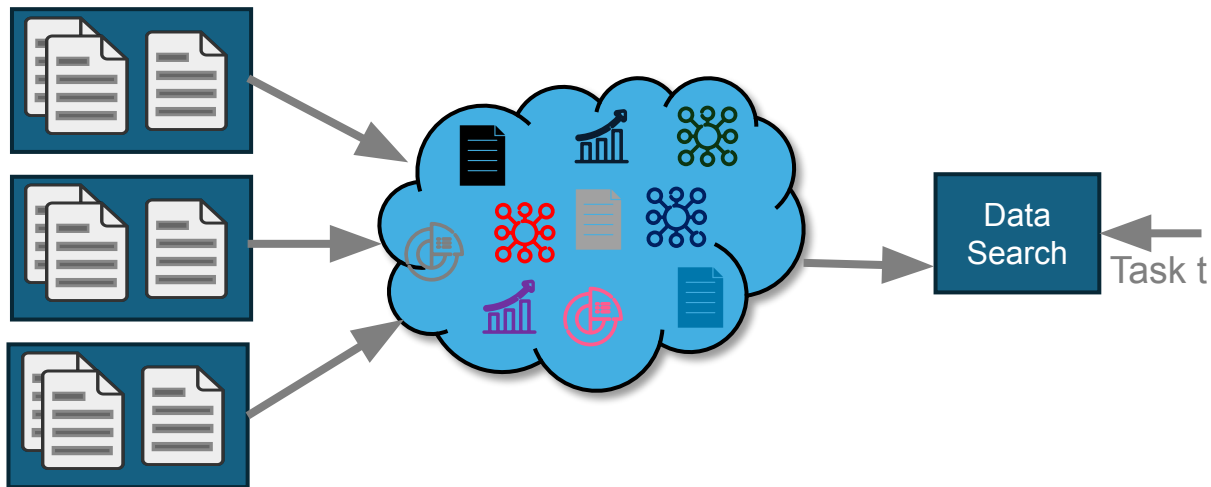
Show Shortlisted d...



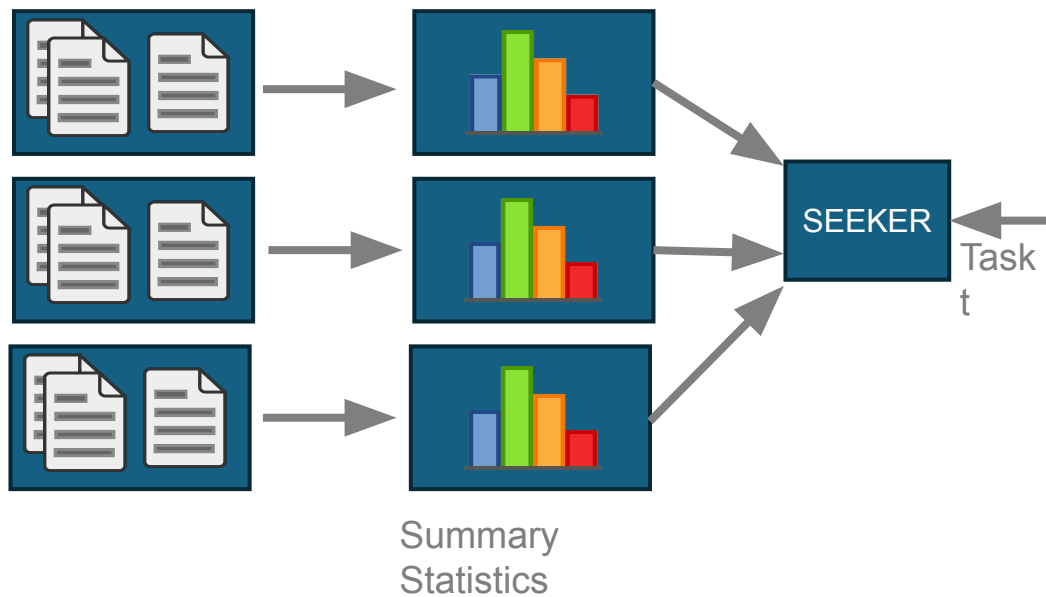
Data Discovery



Centralized Data Discovery



Federated Data Discovery



Distribution-Aware Dataset Search



lung cancer



Sign in

▼ Last updated

▼ Download format

▼ Usage rights

▼ Topic

▼ Provider

Free

Saved datasets

100+ datasets found

kaggle Lung Cancer DataSet

kaggle.com

zip

Updated Jun 1, 2023

+ more versions



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

kaggle

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

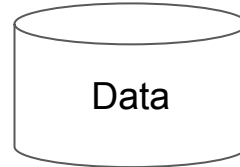
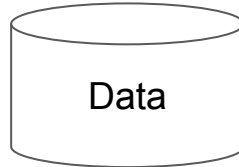
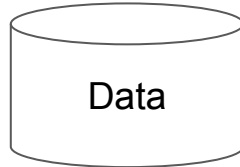
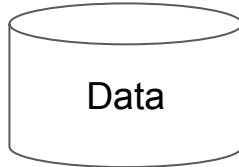
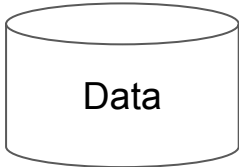
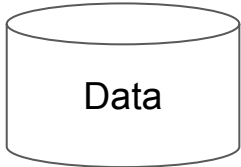
Contents



Speculating About the Future



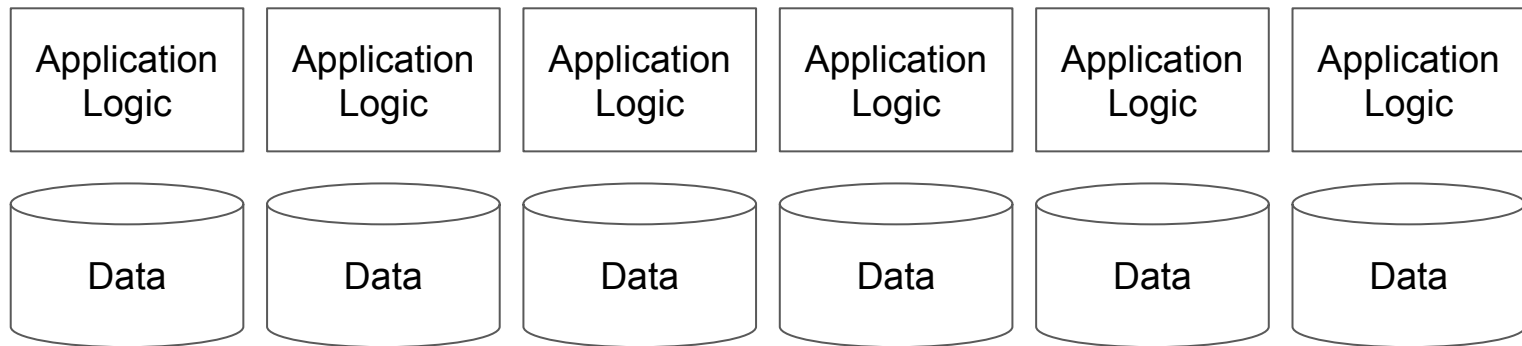
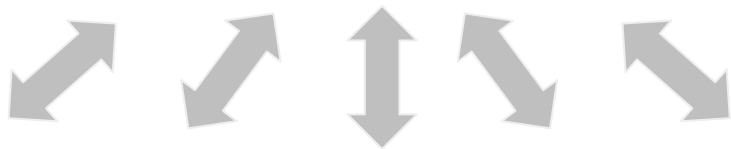
Data Market/Data Discovery





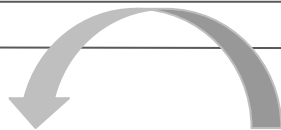
Application
Logic

Data





Data Market/Data Discovery



Application
Logic

Application
Logic

Application
Logic

Application
Logic

Application
Logic

Application
Logic



Data

Data

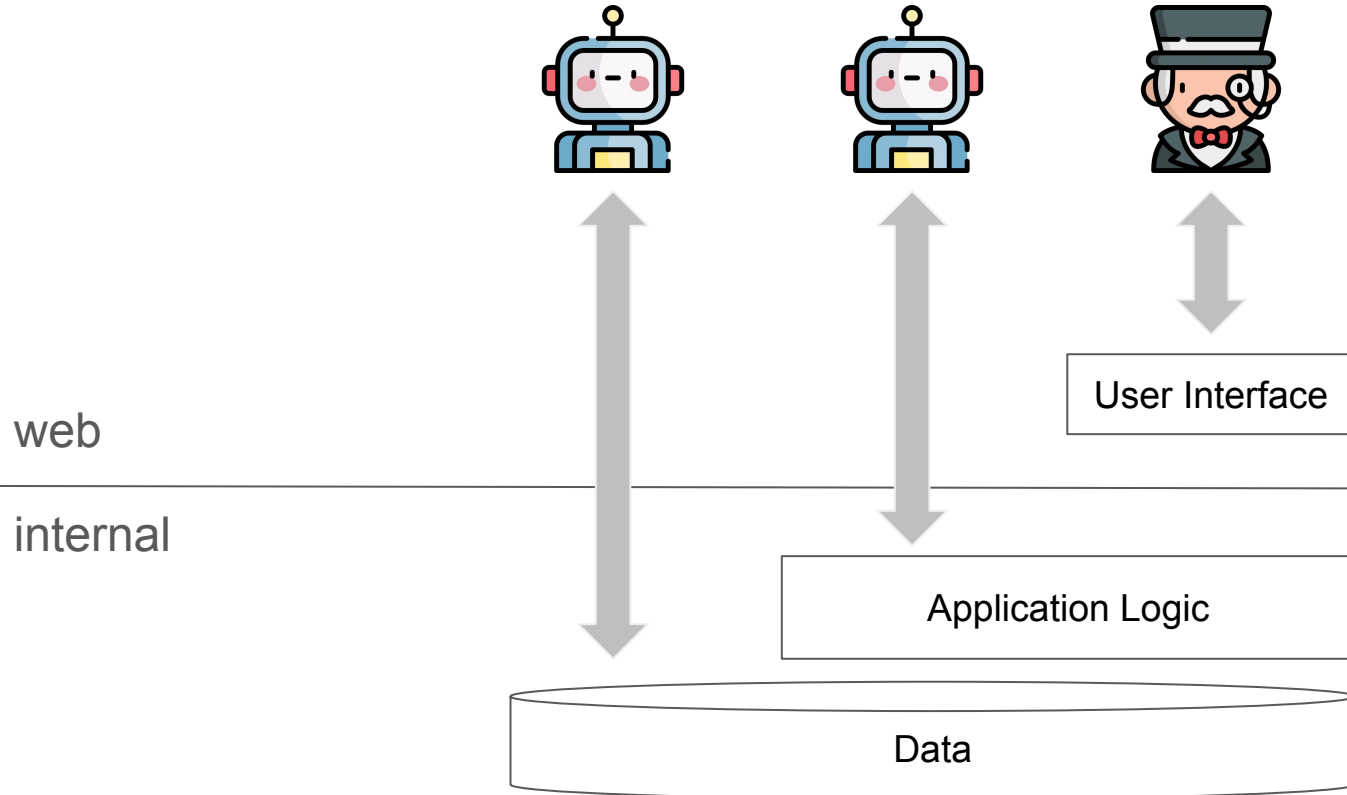
Data

Data

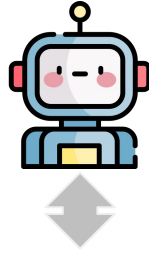
Data

Data

Will Data Markets Be More Important In the Future?



Will Data Markets Be More Important In the Future?



Data Market/Data Discovery

