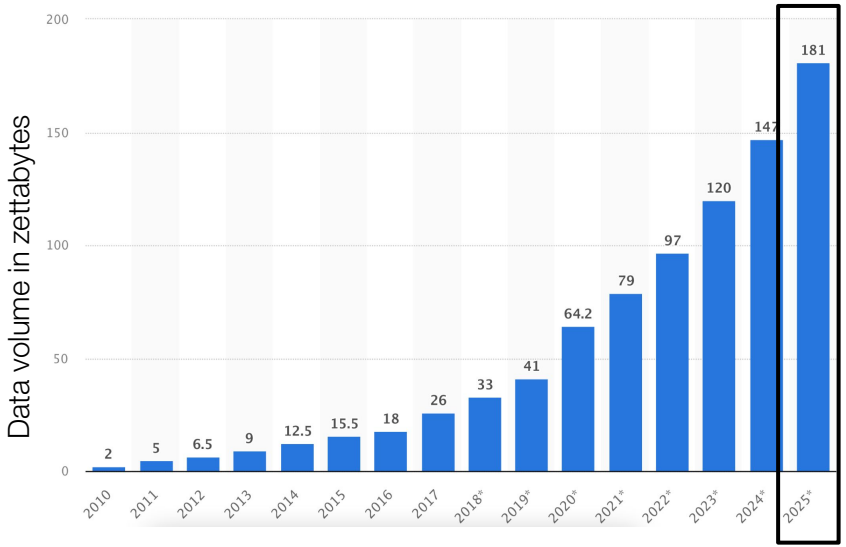
Revisiting Dataset Search

Madelon Hulsebos

In collaboration with: Wenjing Lin, Shreya Shankar, Fatma Özcan, Aditya Parameswaran

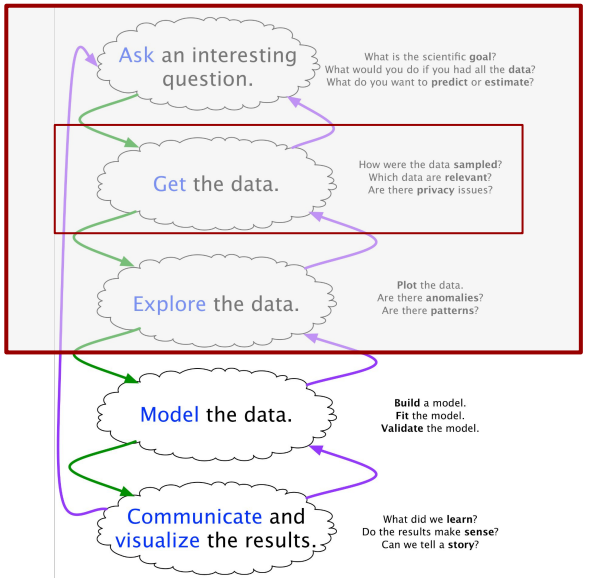
The Dataset Search Problem

Immense growth of data → desire for insight



Source: Taylor, P., Statista, 2023

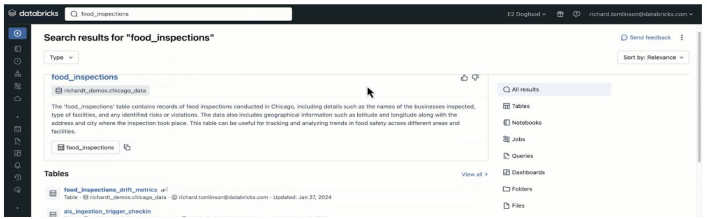
A typical DS/A workflow



Source: Blitzstein and Pfister, Harvard DS course

How Do We Get the Data?

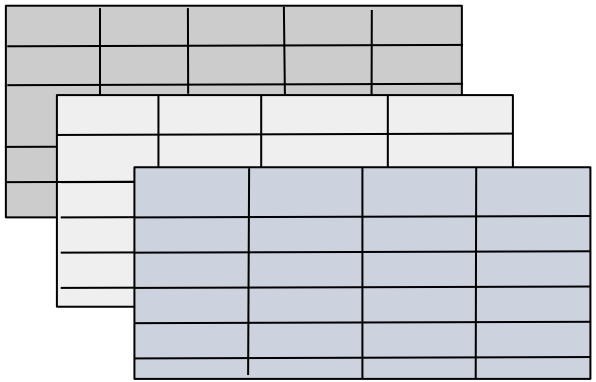
"Basic" dataset search



databricks

query "`product revenue`"

retrieved result



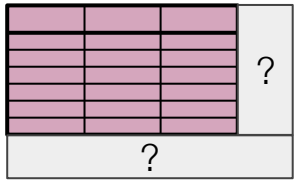
What Are We Researching?

Dataset search for data enrichment

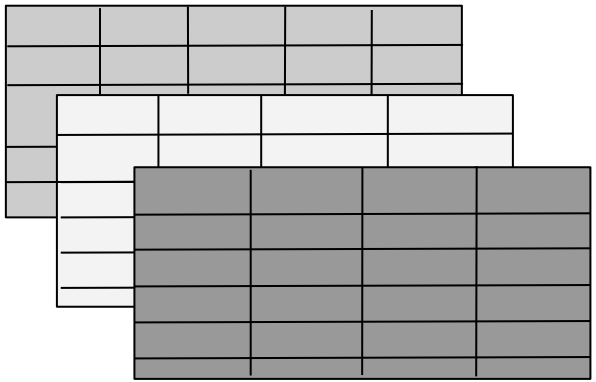
Method	Task	Rep. Learning	ANN Index
Octopus [18]	KS	✗	✗
G.D.S. [2]	KS	✗	✗
Aurum [13]	KS	✗	LSH
LSH-Ensemble [3]	Join	✗	LSH
Juneau [4]	Join	✗	✗
JOSIE [5]	Join	✗	✗
MATE [6]	Join	✗	XASH
DeepJoin [7]	Join	✔	HNSW
D^3L [14]	Union, Join	❙	LSH
Starmie [8]	Union, Join	✔	LSH, HNSW
TUS [9]	Union	❙	LSH
SANTOS [10]	Union	✗	✗
TURL [12]	TU	✔	✗
Sherlock [11]	TU	✔	✗
SATO [19]	TU	✔	✗

Source: Taha et al., ICSC, 2024

query

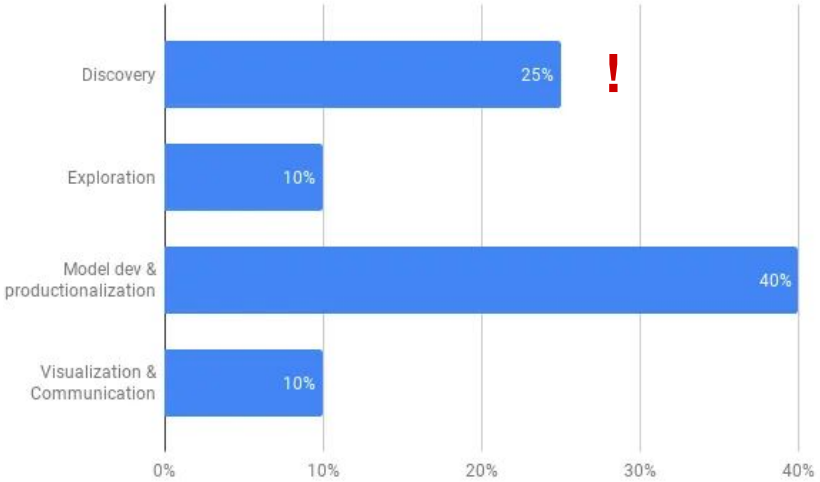


retrieved result



Is “Basic” Dataset Search A Solved Problem?

Finding the right dataset for data analytics tasks is still **a time-consuming process**



Source: Grover, M., Lyft Engineering, 2019

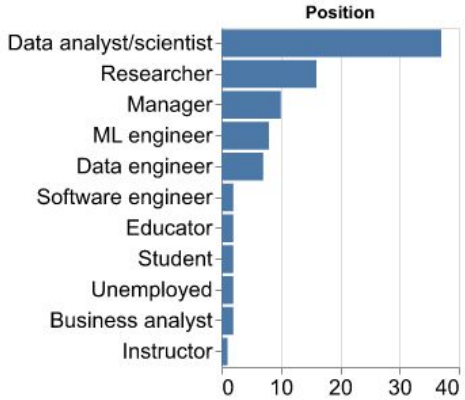
Outline for today

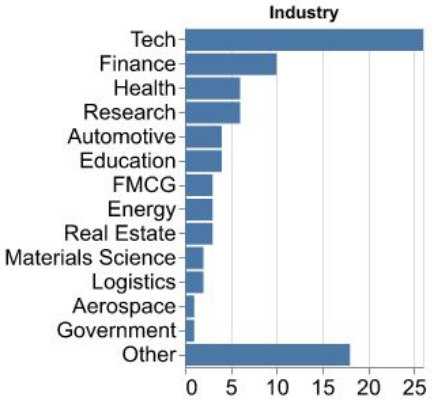
1. Insights from practice
2. Proposal for next-gen dataset search systems

Insights from practice

We asked ~~ourselves~~: **why is dataset search *still* so hard in practice**?

89 data practitioners!! -> widely recruited through social media & mailing lists





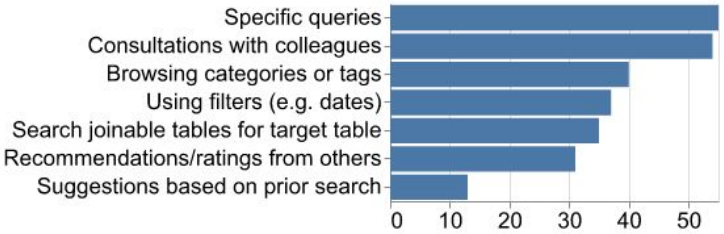
We asked:

- What and how do they search?
- What challenges do they face?
- How do they *want* to search?

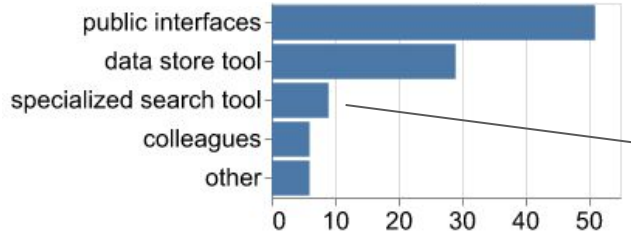
Practitioner's perspective: **what and how they search**

79% searches for **initial dataset**, 52% for **data enrichment**.

How do you search?



What tools do you use to search?



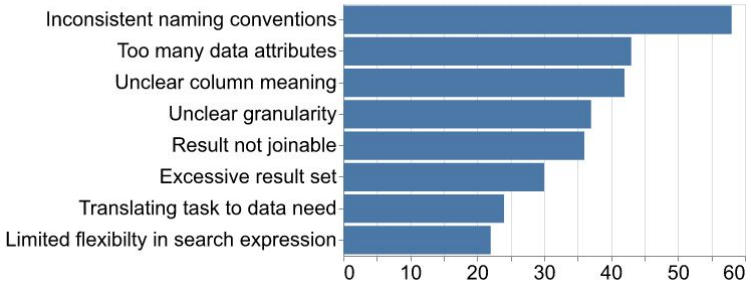
"*Identify the **problem, and the data for the problem**, … then specific keyword or tag search. Also, identify **people** who have worked on **similar problems**…*"

"*Having **so many tables**, I ask more experienced colleagues **which ones are most inherent to the analysis** I need to do. I then navigate through the categories and tags to looks for others.*"

Only 9% uses specialized tools, e.g data catalogs.

Practitioner's perspective: **key challenges**

Key challenges with existing systems?



*"The biggest challenge I've noticed is **messy variable naming** - it takes me a long time to unpack what each variable means…."*

*"**It was painful** because **almost every column had unrecognizable information** (like encrypted) it took longer than I was expecting"*

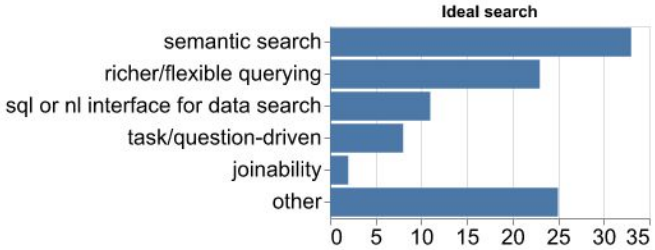
*"**Categorical level of detailing** is required, which isn't possible now."*

*"There are **too many table results** after the initial search…."*

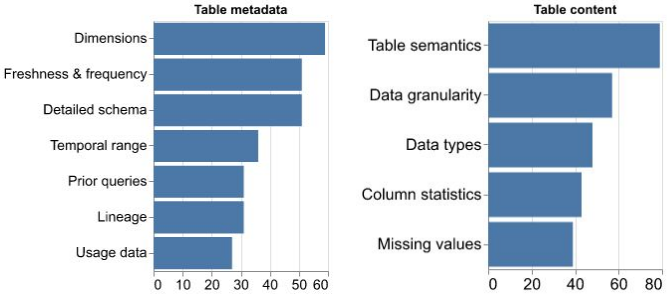
*"Not many features to search/query keywords, a lot of times **changing query still renders same data results**…"*

Practitioner's perspective: **ideal search systems**

What should search systems facilitate?



What properties to search over?



"**Topic model search results**, based on sentence similarity with the dataset description."

"Ideally I would have something **across all of the various data sources and tables** and be able to use SQL (**or a trustable NLP solution**) and pull all relevant data and metadata."

"Show me **product usage** datasets where the main fact table is **event-level usage** data with **hundreds of millions** of records and there are dimension **tables for user and account**."

"Dataset to <**solve issue of …**> with columns <1,2,3,…> on <**granularity desired**>"

Towards Next-Generation Dataset Search Systems

Desiderata for Dataset Search

Remember Bjorn's question; *do "users" know what they want? No!*

Task-driven: explicit **data needs often unknown** requiring back-and-forths w/ experts

Hybrid: search spans **multiple "views"** of a table; raw metadata + semantics

Iterative: data search queries **don't fit a search bar**; complex process

Comprehensible and diverse results: result sets **hard to digest and navigate**

Task-driven: Hypothetical Schema Embedding (HySE)

(1) Task-driven query

> What data is needed to **train a machine learning model** to **forecast demand for medicines across suppliers**?

(2) Hypothetical Schema generation

```
Instruction: generate schema needed for the given task query.
Query: {task-driven query}
```

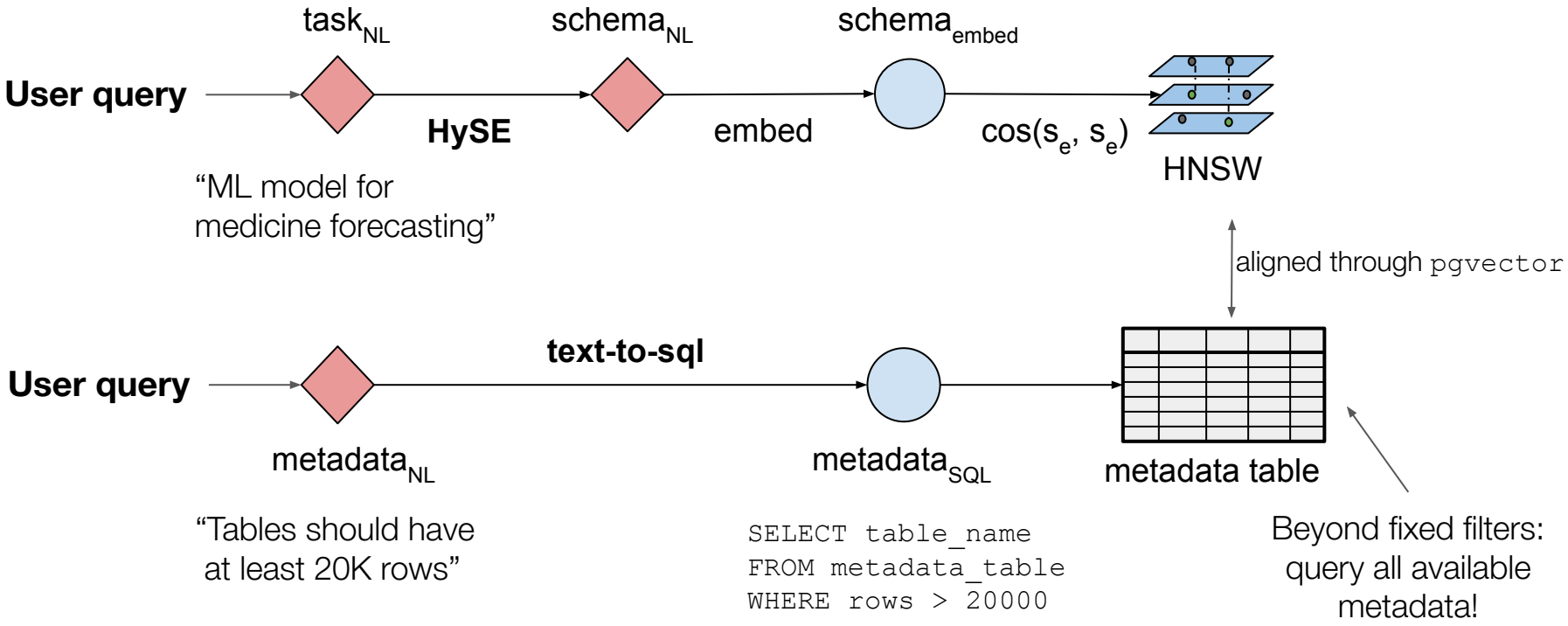
LLM output: "`hypothetical_schema`"

```
medication table: medication id, medication name, …
sales table: medication id, supplier id, date, quantity sold, …
```

(3) Embed(`hypothetical_schema`)

(4) Retrieve source tables from vector store similar to `hypothetical_schema`

Hybrid: retrieval from multimodal index



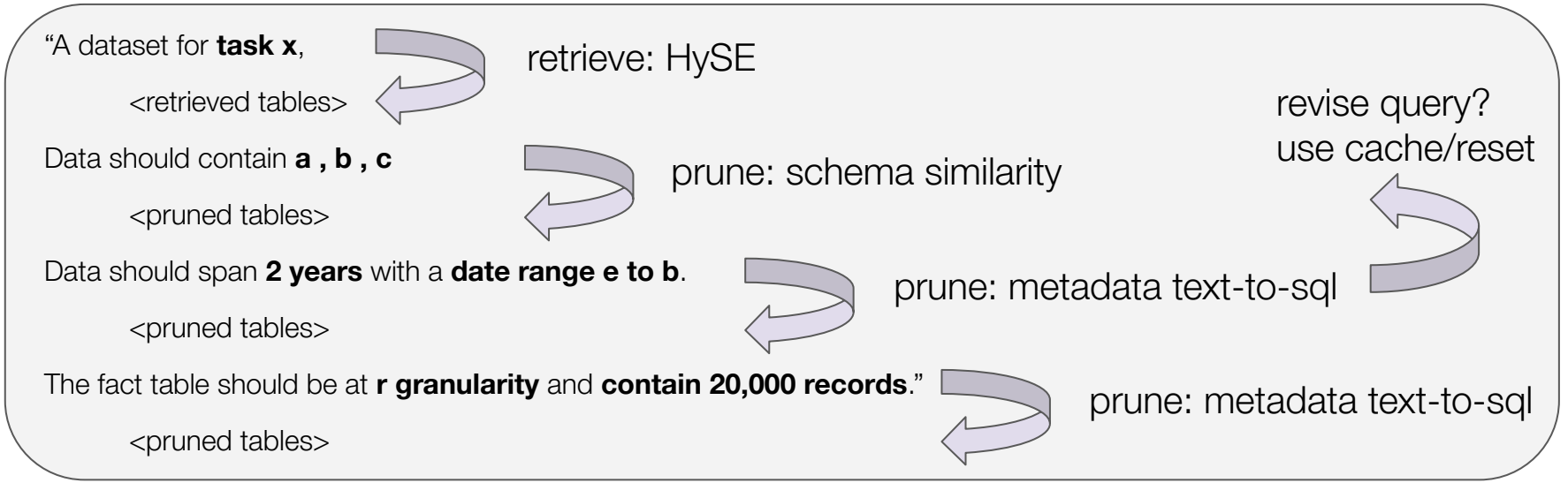
Iterative: conversational interface

Initial complex search query

A dataset to train a medicine forecast model, should contain a, b, c, and span 2 years with a date range e to b. The fact table should be at r granularity and contain 20,000 records.

AI assistance

- query interpretation & iteration
- routing through query engines
- reset or prune retrieved results



Recap

- Basic dataset search is critical to gain insight from data, *but still very hard*
- Dataset search is a complex process, we need:
  - Task-driven search
  - Hybrid search over metadata + semantics
  - Iterative interfaces
- We're well positioned to build more flexible tools with LLMs and chat interfaces!

Open questions?

- What about 4th desideratum; comprehensible results? Talk to Wenjing / Rachel
- Can we use these ideas for RAG over structured data? Talk to Carl / Rachel
- Join the poster session Wed 10AM to learn more!

Stay tuned!

madelon@berkeley.edu