Teaching a New Dog Old Tricks

Reusing Data Management Principles in the Age of LLMs

Tova Milo
Tel Aviv University

A Brief (personal) history

1991 Object Oriented Databases

1994 Semi structured Data

1998 Data Integration

1999 XML and Web services

2006 Data-centered Workflows/Business processes

2011 Crowd sourcing

2018 Data exploration + ML

2020 Data disposal

…

The AI Era

From sports, to health care, to the way we drive our cars, or choose how to invest our money, AI is entering every aspect of our lives…

This is particularly true for the way we interact with data!

An eye opener (for me)

Can Foundation Models Wrangle Your Data?

A. Narayan, I. Chami, L, J. Orr, C. Ré, **VLDB 2022**

Have LLMs Made These Challenges Disappear?

Query languages

Semi-structured data

Data integration

Data cleaning

Crowdsourcing

Explanations

…

Just ask an LLM?

So, do we still need these old ideas?

Perhaps more than ever

LLM-Based Data Management still faces many challenges

- Hallucinations
- Non-grounded explanations
- Ignoring constraints and inconsistencies
- No guarantees of correctness, completeness, reproducibility

Old principles can help address new AI problems - often in new roles

The Rest of This Talk

1. 2 Examples
   (Exploratory Data Analysis)

2. The general picture
   (Toolbox for LLMs and related technology)

3. The future
   (agentic AI)

Context

Exploratory data analysis (EDA):

The process of examining & investigating a given dataset

EDA agent

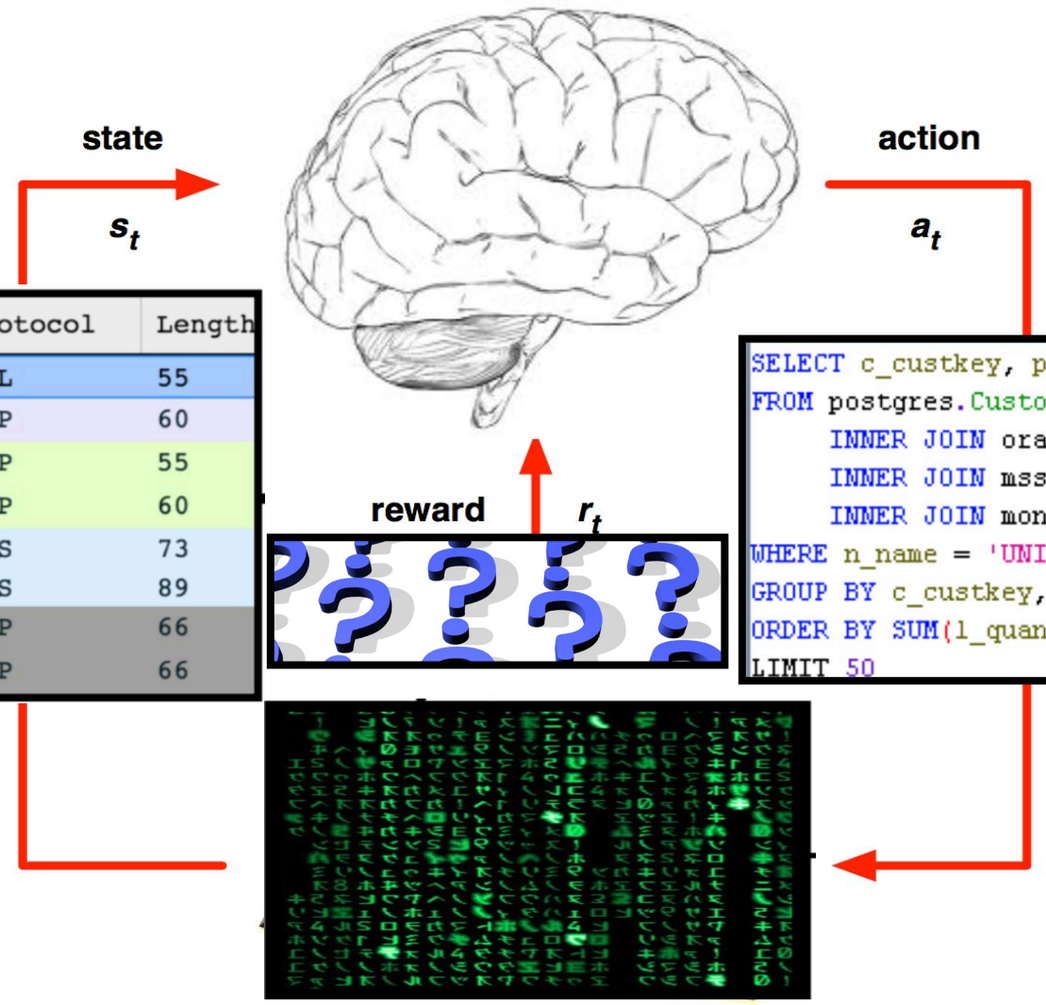
Can we teach a machine to generate a coherent, meaningful sequence of exploratory queries?

[SIGMOD'20, SIGMOD'23, VLDB'25, EDBT'25]

RL-EDA Settings

Utilizing the RL paradigm for EDA:

1. Agent observes a **dataset/result** set
2. Agent formulates a **query**
3. Agent receives a **reward**
4. Agent learns a policy that **maximizes expected reward**



Source	Destination	Protocol	Length
172.16.254.128	216.58.208.206	SSL	55
216.58.208.206	172.16.254.128	TCP	60
172.16.254.128	216.58.208.226	TCP	55
216.58.208.226	172.16.254.128	TCP	60
172.16.254.128	8.8.8.8	DNS	73
8.8.8.8	172.16.254.128	DNS	89
172.16.254.128	216.58.208.227	TCP	66
172.16.254.128	216.58.208.227	TCP	66

2 examples

Query languages for semi-structured data [SIGMOD'23, EDBT'25]

Crowdsourcing [SIGMOD'25, EDBT'26]

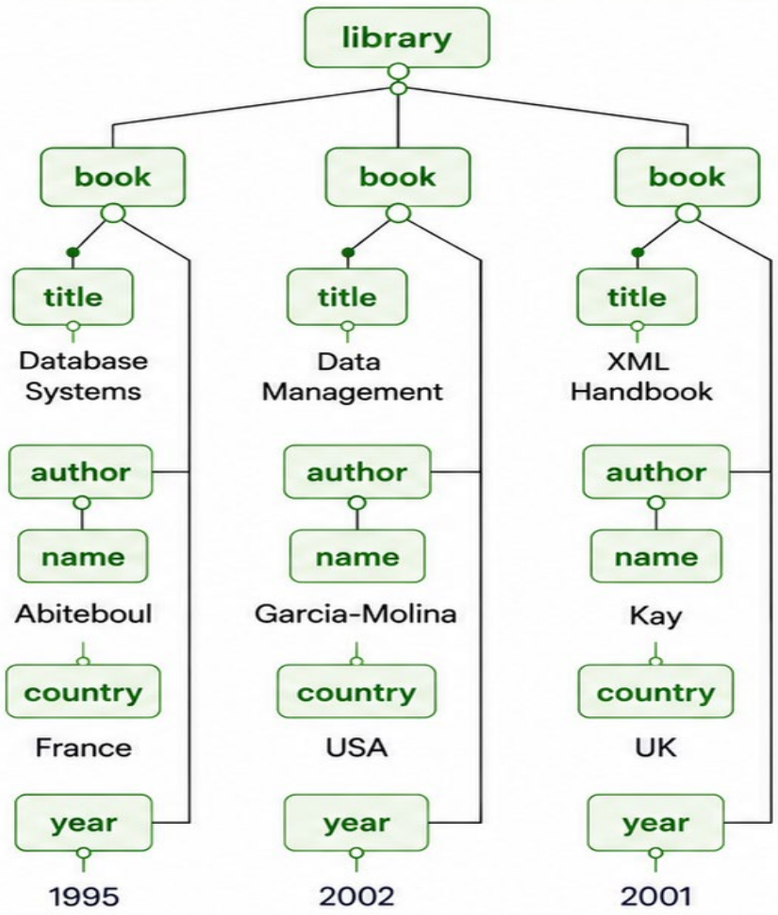
Semi-structured data

Very hot topic between 1995-2005

1. Semi-Structured Data (XML)

```
<library>
  <book id="b1">
    <title>Database Systems</title>
    <author id="a1">
      <name>Abiteboul</name>
      <country>France</country>
    </author>
    <year>1995</year>
  </book>
  <book id="b2">
    <title>Data Management</title>
    <author id="a2">
      <name>Garcia-Molina</name>
      <country>USA</country>
    </author>
    <year>2002</year>
  </book>
  <book id="b3">
    <title>XML Handbook</title>
    <author id="a3">
      <name>Kay</name>
      <country>UK</country>
    </author>
    <year>2001</year>
  </book>
</library>
```

2. Modeling as a Tree

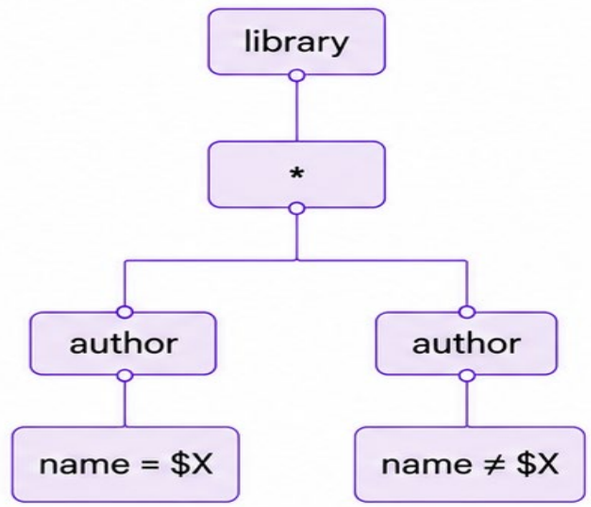


3. Example of a Tree Pattern Query

Query Goal:

Find all publications with two different authors.

Tree Pattern Query



Back to Data Exploration

Explore a dataset D in light of a goal g

Find a country with different viewing habits than the rest of the world

type	title	country	...	dur.	genre	rating
Movie	Bareilly Ki..	India	...	110	Intl.	TV-14
TV	Imposters	USA	...	2	Comedies	TV-MA
Movie	Pulp Fiction	USA	...	154	Classic	R
TV	Thomas & Fri..	UK	...	2	Kids	TV-Y
...	...	...	...	...	...	...

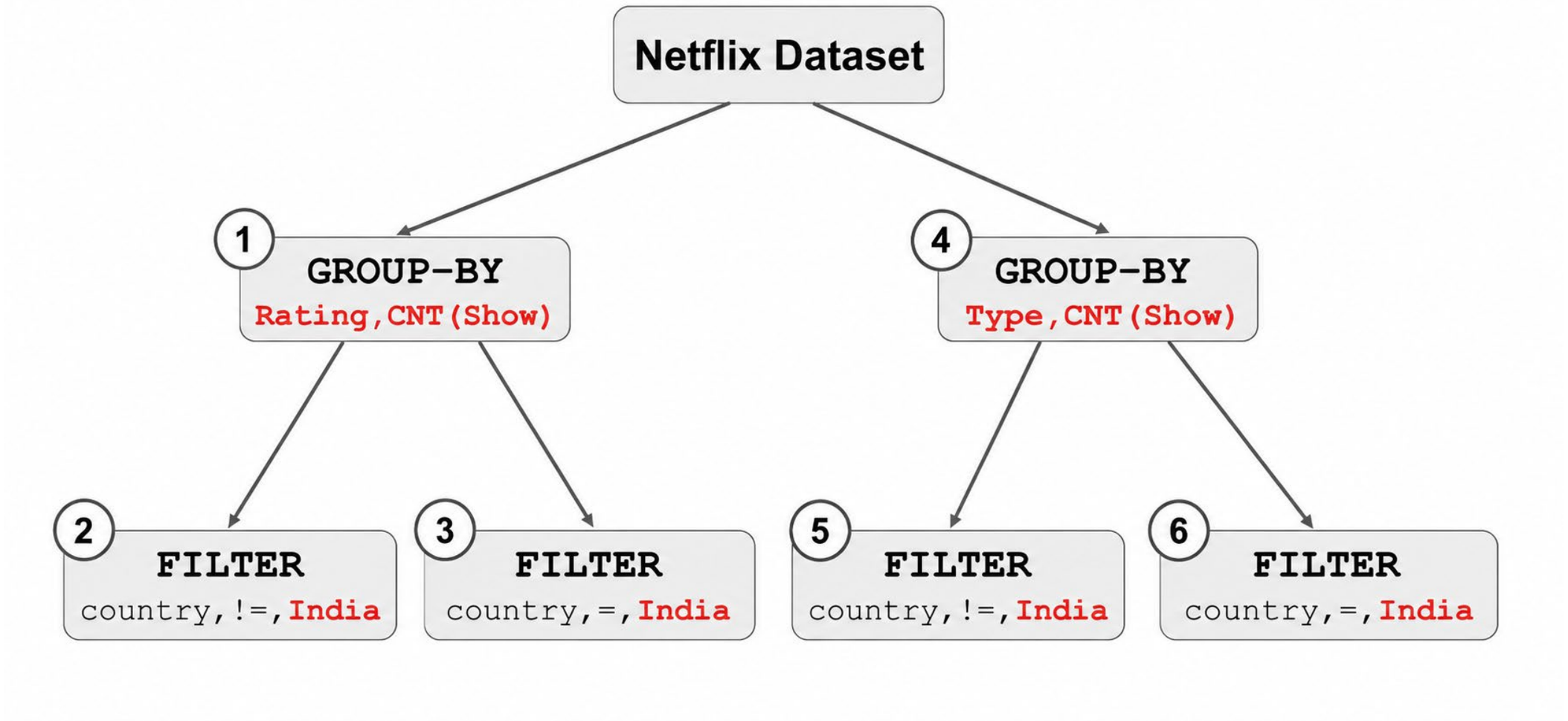
Netflix TV-shows and Movies Dataset

Analysis Results

India (1046 titles) stands out in terms of **rating** and **type**:

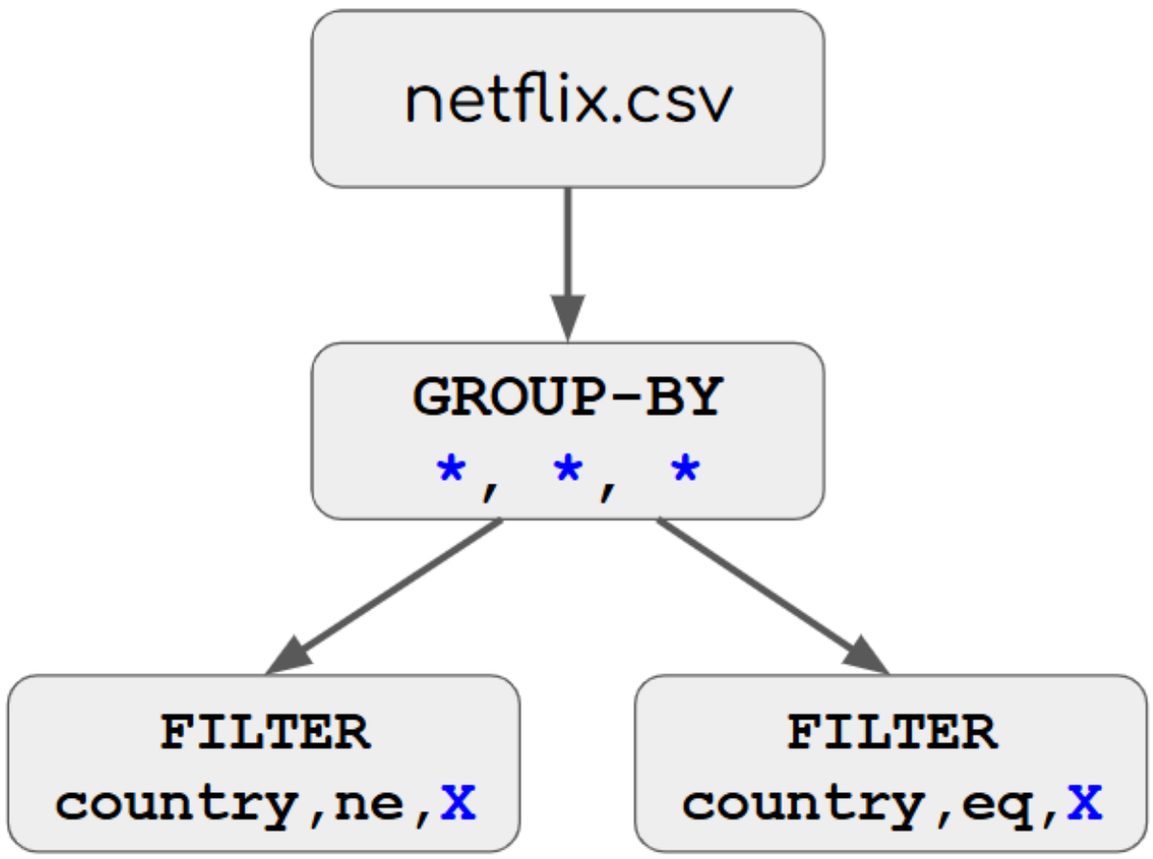
- **Content Rating:**
  - In **India**, 56% of titles are rated TV-14.
  - Rest of the world averages only 20% for TV-14 rating.
- **Content Type:**
  - In **India**, 93% of titles are movies.
  - Globally, movies constitute 66%, with the remaining 34% being TV shows.

Exploration as a Tree



Goal as a Tree Query

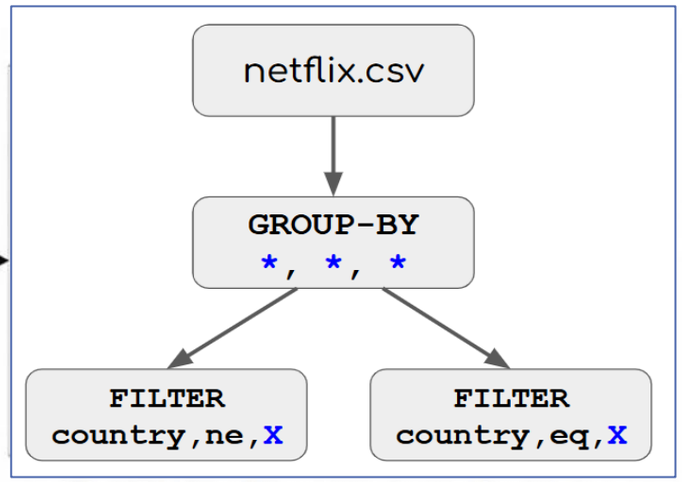
Find countries with different viewing habits than the rest of the world



End-to-end Architecture

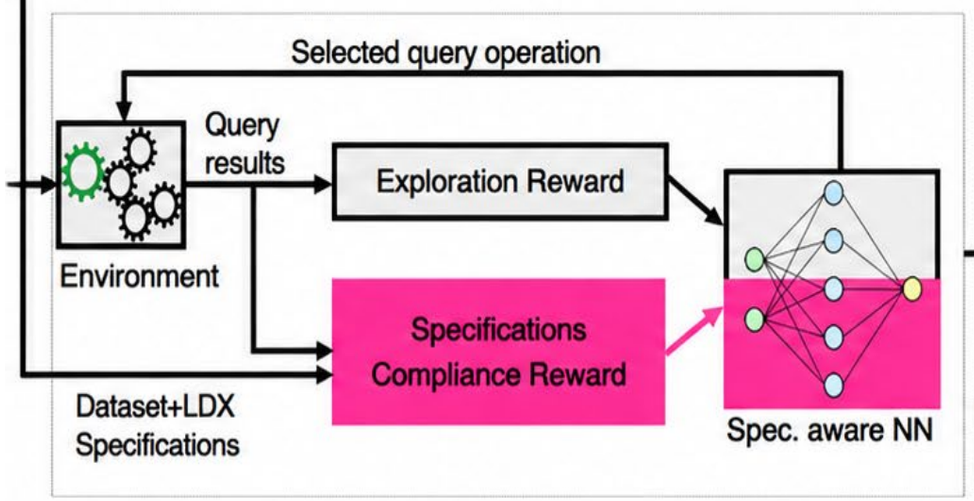
Find a country with different viewing habits than the rest of the world

type	title	country	...	dur.	genre	rating
Movie	Bareilly Ki..	India	...	110	Intl.	TV-14
TV	Imposters	USA	...	2	Comedies	TV-MA
Movie	Pulp Fiction	USA	...	154	Classic	R
TV	Thomas & Fri..	UK	...	2	Kids	TV-Y
...	...	...	...	...	...	...

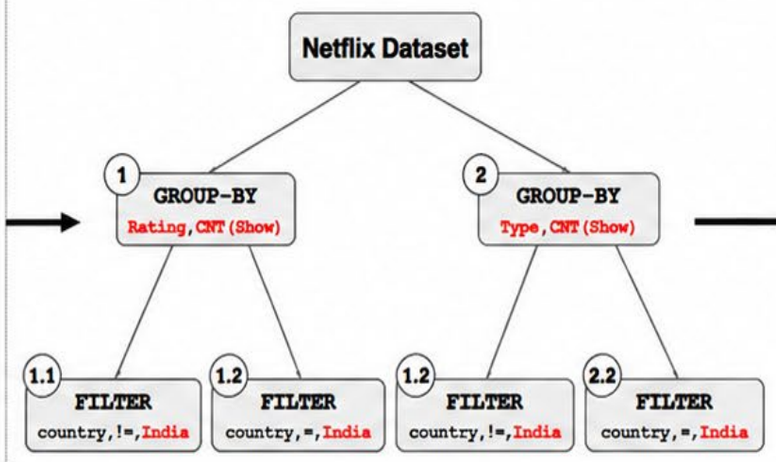


(a) Input: Dataset and Analysis Goal

(b) LDX Exploration Specifications



Constrained Deep Reinforcement Learning (CDRL) ADE Engine



(d) Executable Exploration Tree

1. Group by *rating*, count(*show*);
Filter by *country* != 'India'

[1.1]:

rating	COUNT
TV-MA (mature audience)	2623 (38%)
TV-14 (14+ children)	1398 (20%)
...	...
TV-G (all ages)	274 (4%)

1.2 Group by *rating*, count(*show*);
Filter by *country* = 'India'

[1.2]:

rating	COUNT
TV-14 (14+ chidren)	533 (56%)
TV-MA (mature audience)	240 (25%)
...	...
TV-G (all ages)	9 (1%)

2.1 Group by *type*, count(*show*);
Filter by *country* != 'India'

[2.1]:

type	COUNT
Movie	4494 (66%)
TV Show	2337 (34%)

2.2 Group by *type*, count(*show*);
Filter by *country* = 'India'

[2.2]:

type	COUNT
Movie	883 (92%)
TV Show	73 (8%)

(e) Final Exploration Notebook (snippet)

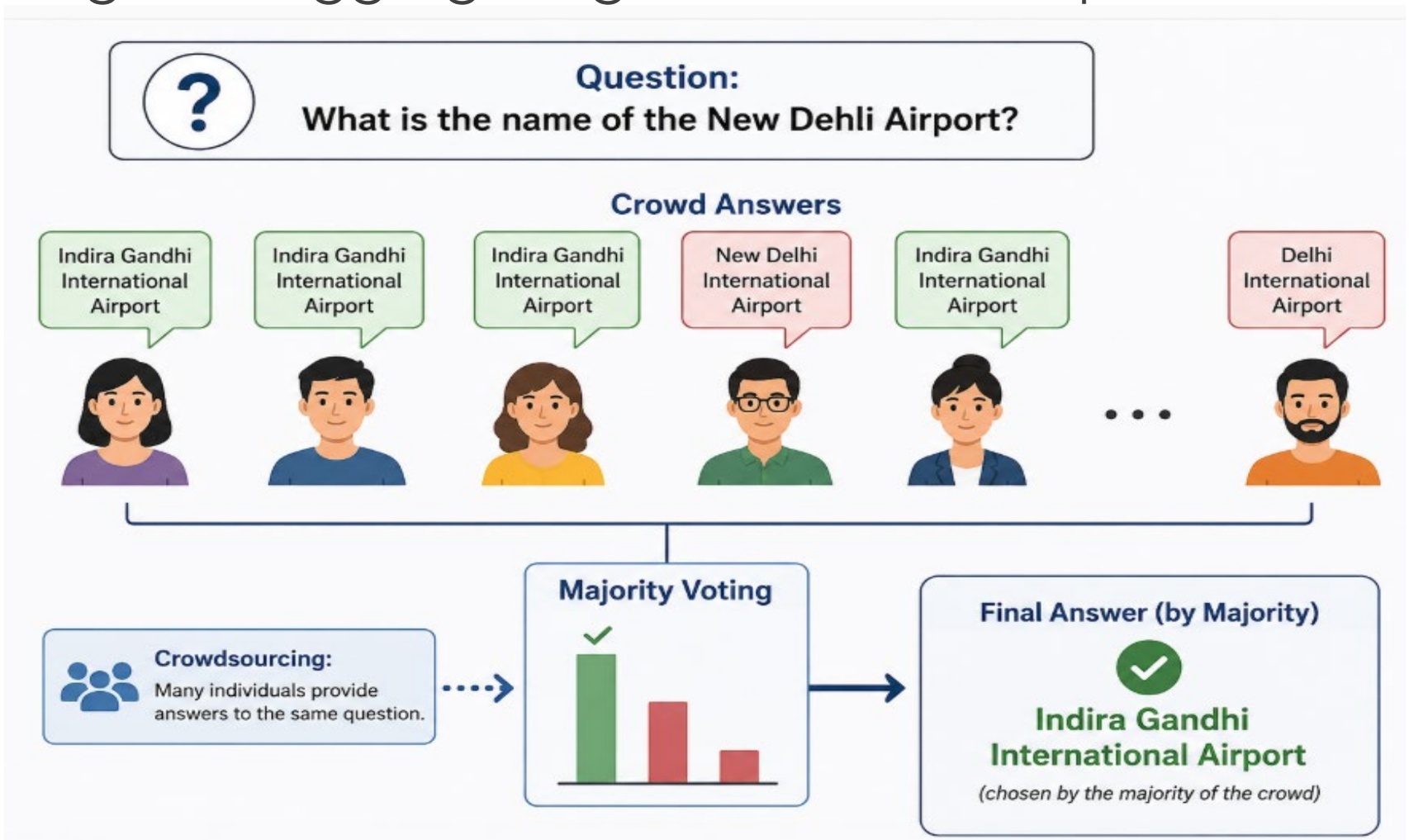
Key message

Formal specifications can constrain and guide AI exploration

Crowdsourcing algos

Very hot topic between 2010-2018

Collecting and aggregating data from multiple users/experts



Back to Data Analysis

Data gathering via LLM API

LLMs have different costs, accuracy, some are domain specific (e.g. has their own expertise)

Just like the crowd…

Data Gathering

1. Incomplete Table (with Missing Data)

ProductID	ProductName	Category	Price (USD)	Rating	ReleaseYear
P001	Wireless Earbuds	Electronics	49.99	4.2	2022
P002	Yoga Mat	Sports	29.99	?	?
P003	?	Kitchen	15.99	4.0	?
P004	Smart Watch	Electronics	?	4.5	2021
?	Bluetooth Speaker	Electronics	39.99	?	2023
P006	Coffee Maker	?	?	4.1	?

Missing Information

- Missing values (e.g., Price, Rating, ReleaseYear)
- Missing entries (rows)
- Missing attributes (columns)
- Inconsistent / incomplete knowledge

2. Multiple LLMs as "Crowd"

Each LLM proposes completions for the missing parts.

LLM A (e.g., GPT-4o)
Suggests values for missing cells, rows, and columns

LLM B (e.g., Claude 3.5)
Provides alternative completions

LLM C (e.g., domain specific)
Adds plausible rows and attributes

...

Aggregation (Majority / Confidence Voting)

Combine outputs using:

- Majority voting / Agreement
- Confidence scores
- Consistency & constraints

3. Augmented (More Complete) Table

ProductID	ProductName	Category	Price (USD)	Rating	ReleaseYear
P001	Wireless Earbuds	Electronics	49.99	4.2	2022
P002	Yoga Mat	Sports	29.99	4.3	2021
P003	Blender	Kitchen	15.99	4.0	2020
P004	Smart Watch	Electronics	89.99	4.5	2021
P005	Bluetooth Speaker	Electronics	39.99	4.1	2023
P006	Coffee Maker	Kitchen	59.99	4.1	2022

What just happened?

Query languages Crowdsourcing

RL/LLM-based EDA system

Old database ideas used in a completely new setting

Not old solutions — reusable principles

The Rest of This Talk

1. 2 Examples
   (Exploratory Data Analysis)

2. **The general picture**
   (Toolbox for LLMs and related technology)

3. The future
   (agentic AI)

LLMs Changed the Interface

Classical systems	**AI systems**
Explicit queries	Natural language
Operators	Prompts
Schemas	Embeddings
Workflows	Agents

At first glance these worlds look completely different…

But not the Problems

Retrieval

Explanations

Validation

Optimization

Orchestration

These problems now reappear inside AI systems

And our toolbox is extremely useful for them

A toolbox for AI systems

Query languages

Constraints

Provenance

Graph data

Optimization

Explanations

Crowdsourcing

…

LLM systems

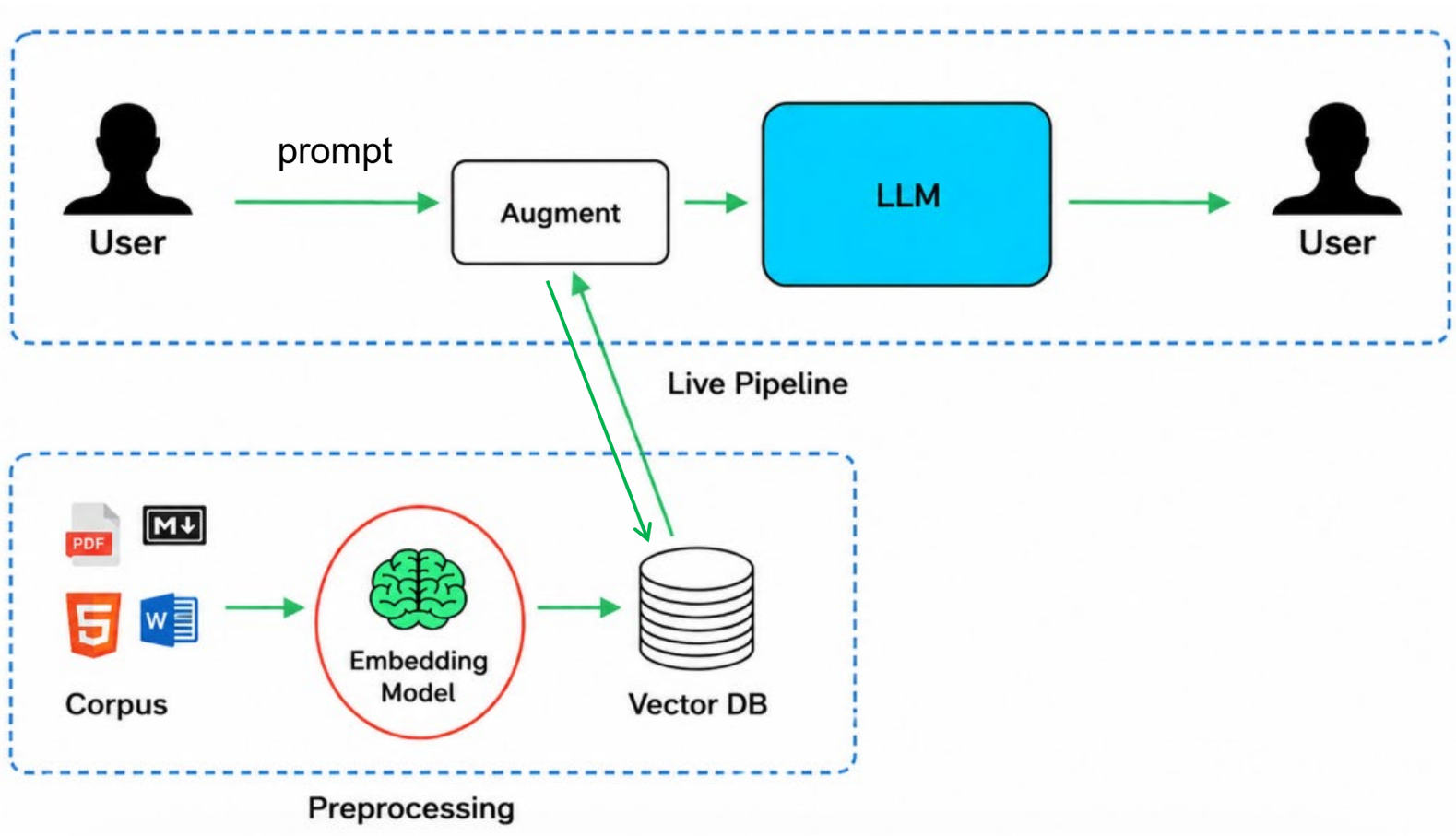
RAG

Agents

The challenge is no longer how humans use these tools – but how AI systems can

Retrieval

RAG (Retrieval Augmented Generation) was a breakthrough because it allowed LLMs to use external knowledge



But also exposed the limitations of embedding-based retrieval

What embeddings miss

Examples:

- missing multi-hop relation
- inconsistent retrieved evidence
- semantically similar but irrelevant chunks

Similarity alone is often insufficient for reliable reasoning

The Return of Structure

GraphRAG / Knowledge-Graph-Enhanced retrieval:

Vector retrieval + graph traversal

- graph edges
- path reasoning
- structured retrieval

We have seen this before

Semi-structured data, XML, Graph queries,…

and now, GraphRAG, retrieval graphs, agent memory graphs

The consumer of structure changed

Research opportunities

Reasoning

- Query planning for RAG
- Constraint-aware retrieval validation

Accountability

- Provenance for generated answers
- Privacy

Systems

- Optimization
- Updates

…

Explanations

Language models are excellent storytellers

In data analysis, explanations should not just sound plausible.

They should be tied to observable patterns in the data, supported by evidence, and open to inspection

[VLDB'23, SIGMOD'26]

Grounded Explanations

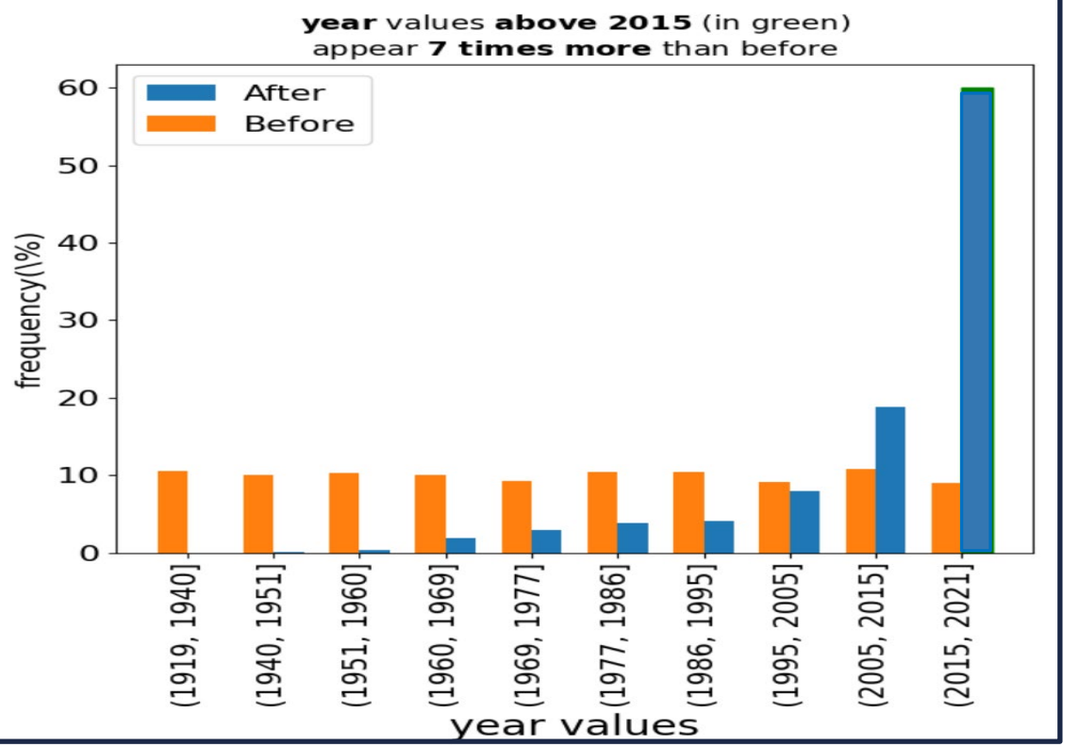
New songs are more popular on Spotify

LLM Explanation

"New songs are more popular because listeners tend to prefer fresh releases and recommendation algorithms promote recent content."

Data-grounded Explanation

```
SELECT * FROM spotify
WHERE popularity > 65;
```



What makes an explanation Trustworthy?

Grounded in the data

Comparative

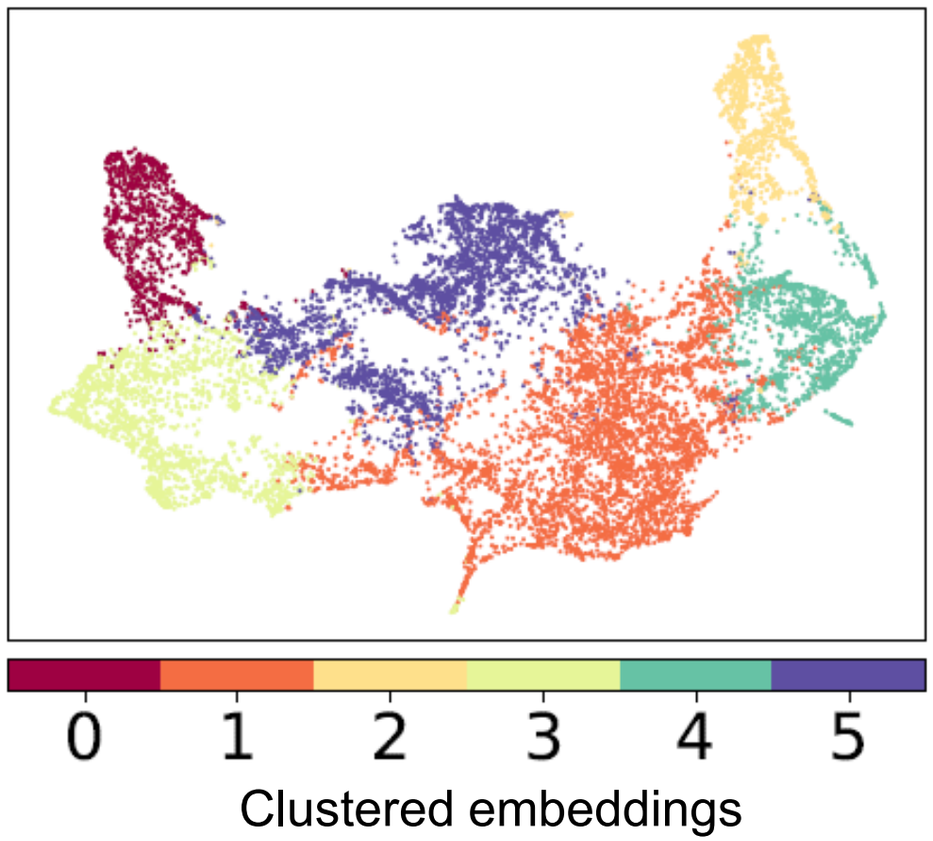
Evidence-based

Inspectable

A story may sound plausible; a grounded explanation shows the evidence.

Grounding Neural Representations

Grounding should apply not only to outputs, but also to internal representations

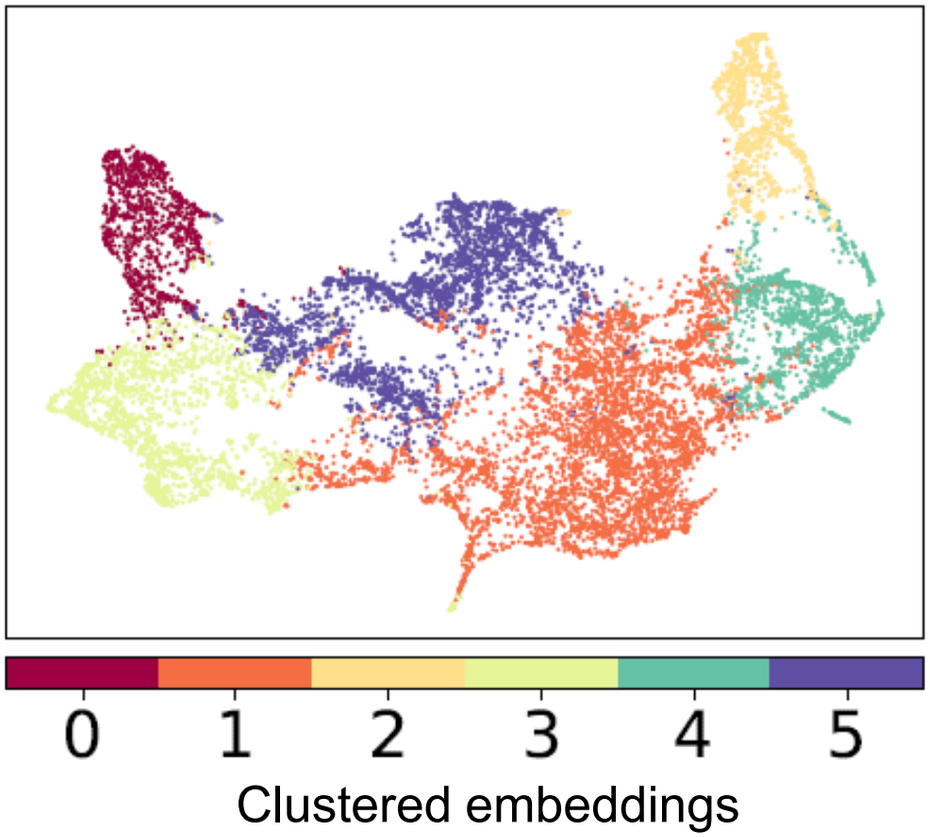


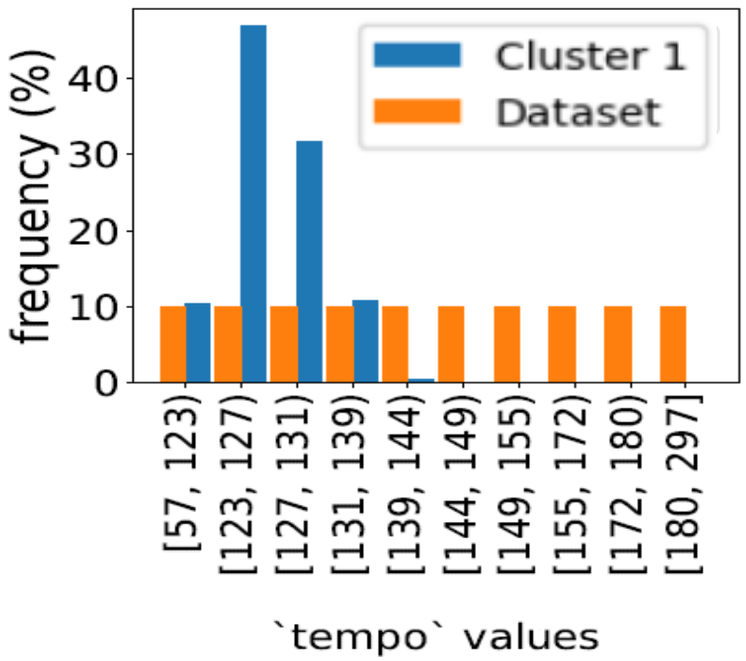
- What do representations capture?
- Which data supports them?
- How do they affect downstream decisions ?

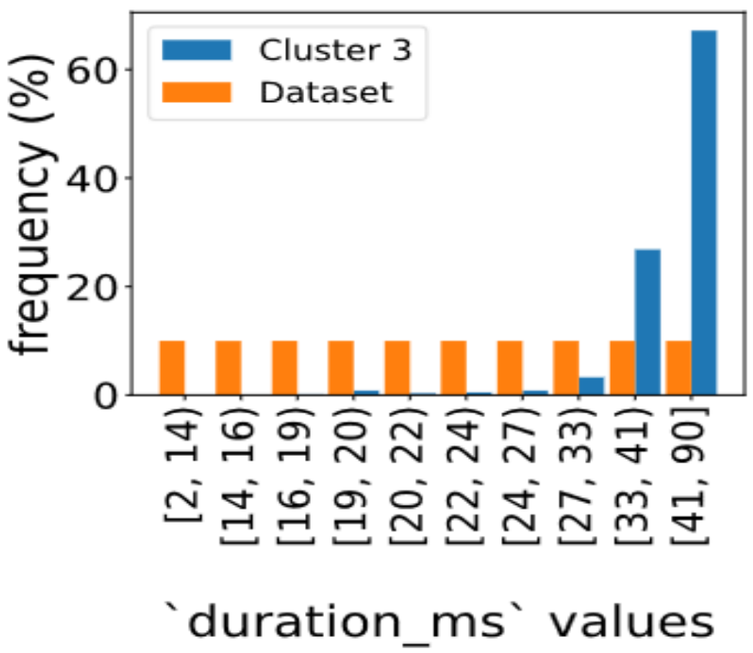
[VLDB'23, SIGMOD'25, EACL'26]

Grounding Neural Representations

Grounding should apply not only to outputs, but also to internal representations







[VLDB'23, SIGMOD'25, EACL'26]

Validation

"New songs are more popular on Spotify"
"Long movies retain viewers longer"
"Users prefer local content"
…

AI systems continuously generate claims

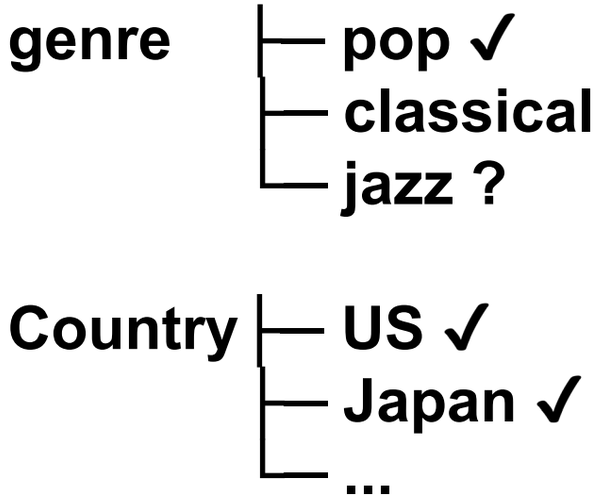
Assumptions are typically implicit

Validation becomes critical

Validation must be part of the reasoning process

Validation through Refinements and Counterexamples

New songs are more popular on Spotify



Counterexamples :
not true for classical music,
mainly true for top artists

Classical DB validation mechanisms can become part of the AI reasoning pipelines

Optimization

Classical systems

Join ordering

Caching

Materialization

Scheduling

Resource allocation

AI systems

Inference routing

KV-cache management

Batching

Retrieval planning

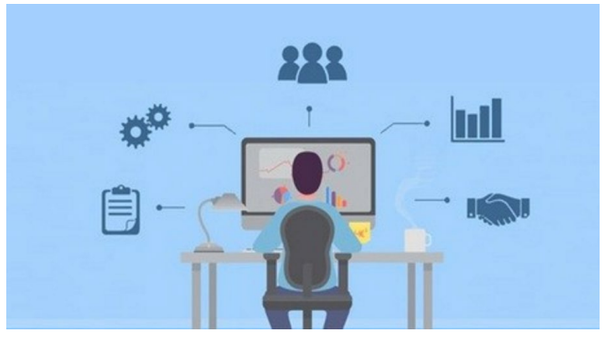
Tool/model routing

Many AI systems are becoming distributed data systems

The abstractions changed, the optimization principles did not

The Rest of This Talk

1. 2 Examples
   (Exploratory Data Analysis)



2. The general picture
   (Toolbox for LLMs and related technology)

3. **The future**
   (agentic AI)

Agentic AI

Can this be done automatically?

Can agents learn when to use classical DB tools?

Service selection: Which tool/model ?

Tool composition: In what order ?

Self-improving workflows: learn from past analyses

Autonomous validation: check claims before presenting

Time to conclude…

Teaching a new dog old tricks

We spent decades learning how to help humans reason about data.

We may now need to teach AI systems to use these ideas themselves.

Not to go back – but to move forward reliably

Thank You!

Yael Amsterdamer, Roni Copul, Susan B. Davidson, Guy Dar , Daniel Deutch, Yael Einy , Nave Frost , Amir Gilad , Aviv Hadar , Maor Juliet Lavi, Tavor Lipman, Amit Mualem, Slava Novgorodov , Kathy Razmadze, Amit Somech, Tomer Wolfson, Oz Zafar, Ron Zadicario, Gal Zeevi