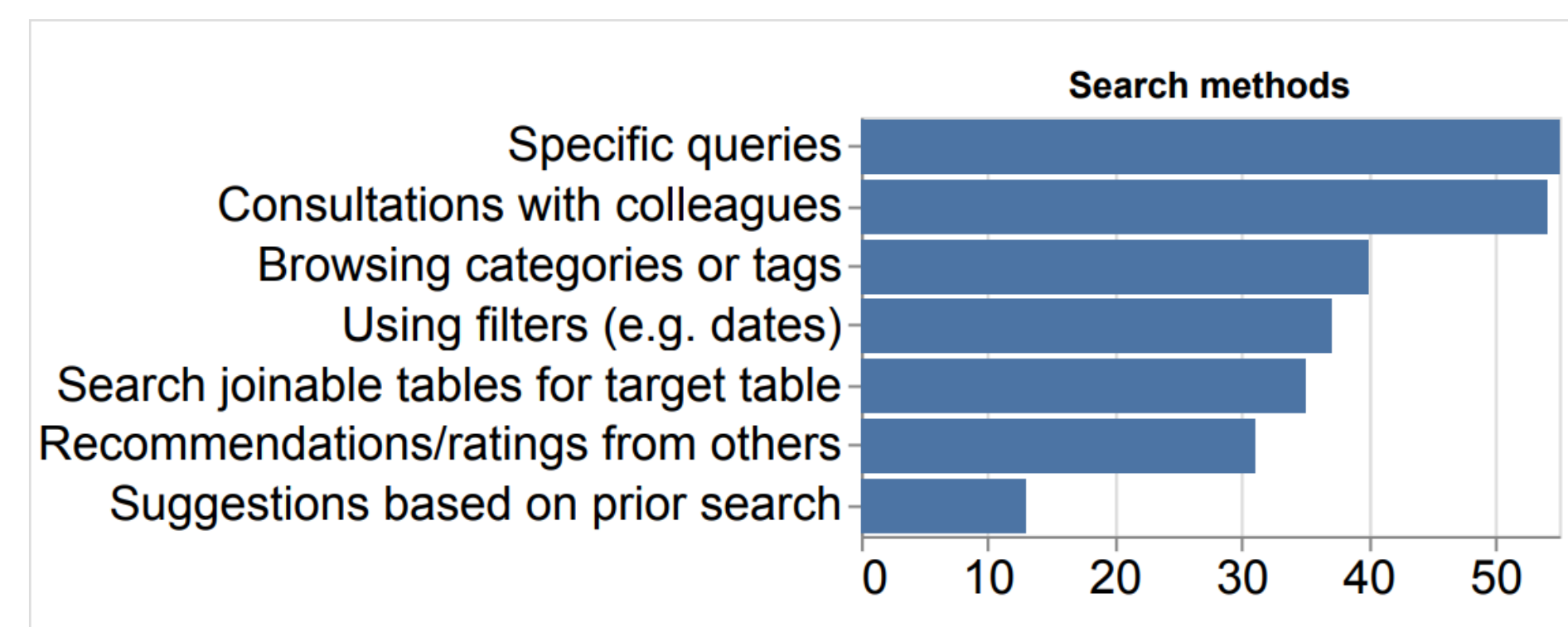


Background

Problem

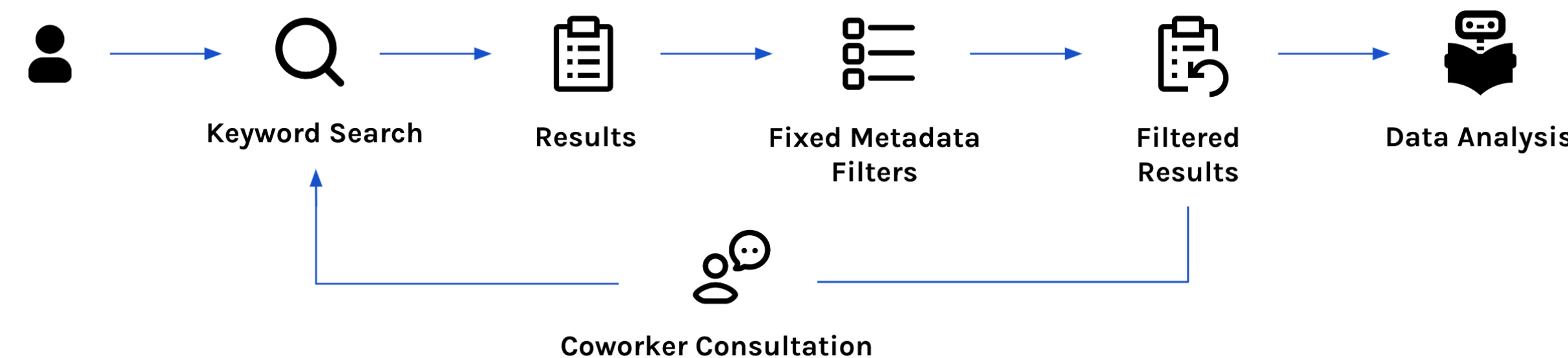
Why and How We Search: Insights from Survey (79%) To find the right dataset for analysis



Top search methods:

1. Specific queries
2. Consultations with coworkers
3. Filters/categories

Figure 1: The standard industry dataset search workflow is rigid and offers little support for improving queries, often leading to significant time spent consulting with colleagues.



Objective

Build a **flexible** interface that allow users to **independently & iteratively** search for datasets, providing **proactive guidance** to help build effective queries tailored to their specific tasks.

System Design

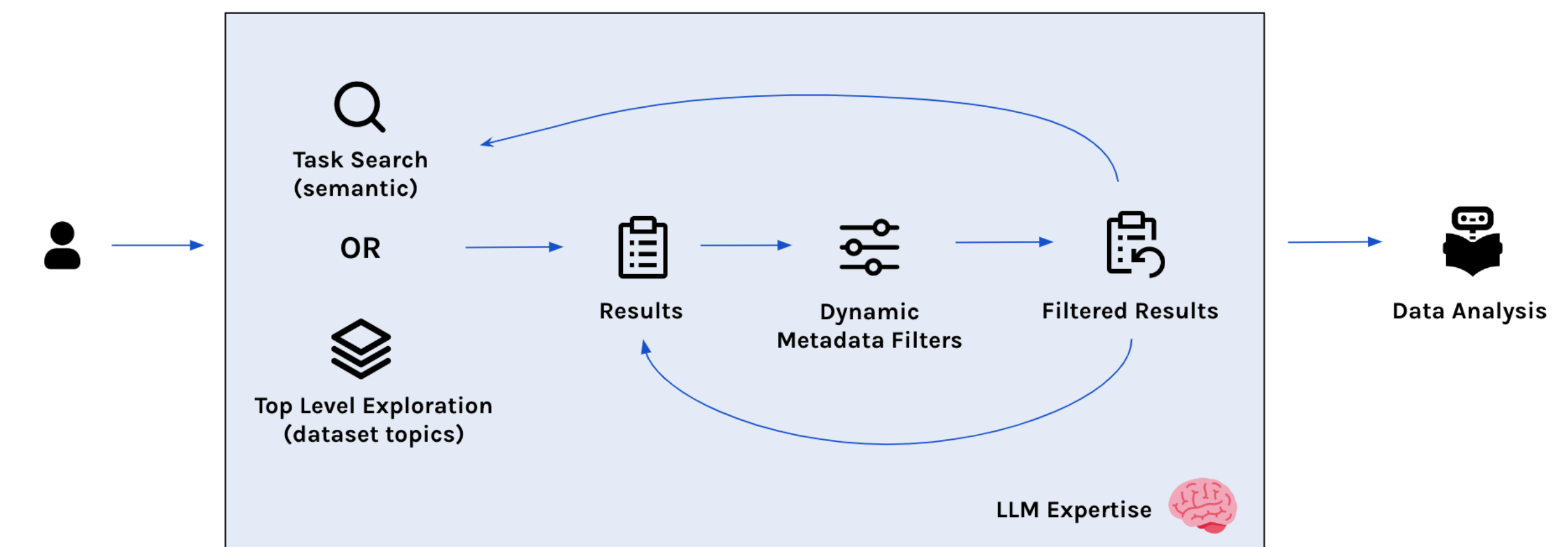


Figure 2: Framework of the proposed interface. Task/topic and metadata searches are performed through a natural language interface. The entire search process is supported by an LLM, which provides query suggestions, dynamically generates task and metadata filters, and answers clarification questions.

Design Principles

- D1 LLM Elicitation through Proactive Guidance**
Purpose: Prompt users to share more information about their needs, which will be reflected in the query blocks & search interface.
- D2 Dynamic Query Decomposition**
Purpose: Allow users to see how the LLM is dynamically updating and refining the search space, providing transparency into the search process.
- D3 Allowing Users to Compare Datasets Efficiently**
Purpose: Facilitate high-level exploration of datasets by organizing them into topics and enable users to delve into metadata details of individual datasets as they iteratively build and refine their queries.

Interface

Top Dataset Results

Showing 10 out of 355

1. Hourly Energy Consumption
Bob Evans • Updated 8 days ago
Usability 100% • 6MB
2000 records • 40 columns

2. World Energy Consumption
Lucy Evans • Updated 10 days ago
Usability 98% • 8MB
5320 records • 23 columns

3. International Energy Statistics
Hose Painter • Updated 2 months ago
Usability 89% • 18MB
8900 records • 34 columns

4. Appliances Energy Prediction
Amy Smith • Updated 15 days ago
Usability 88% • 7MB
500 records • 20 columns

5. Energy Efficiency Dataset
Andrew Fong • Updated 3 weeks ago
Usability 88% • 8MB
550 records • 14 columns

6. Hourly energy demand generation
Anna Ye • Updated 2 months ago
Usability 79% • 11MB
354 records • 24 columns

7. Renewable Energy
Matt Chang • Updated 3 days ago
Usability 78% • 2MB
75 records • 3 columns

8. Nuclear Energy Datasets
Ken Roth • Updated 8 days ago

About Dataset

energy hourly range environment

Description

PJM Interconnection LLC (PJM) is a regional transmission organization (RTO) in the United States. It is part of the Eastern Interconnection grid operating an electric transmission system serving all or parts of Delaware, Illinois, Indiana, Kentucky, Maryland, Michigan, New Jersey, North Carolina, Ohio, Pennsylvania, Tennessee, Virginia, West Virginia, and the District of Columbia.

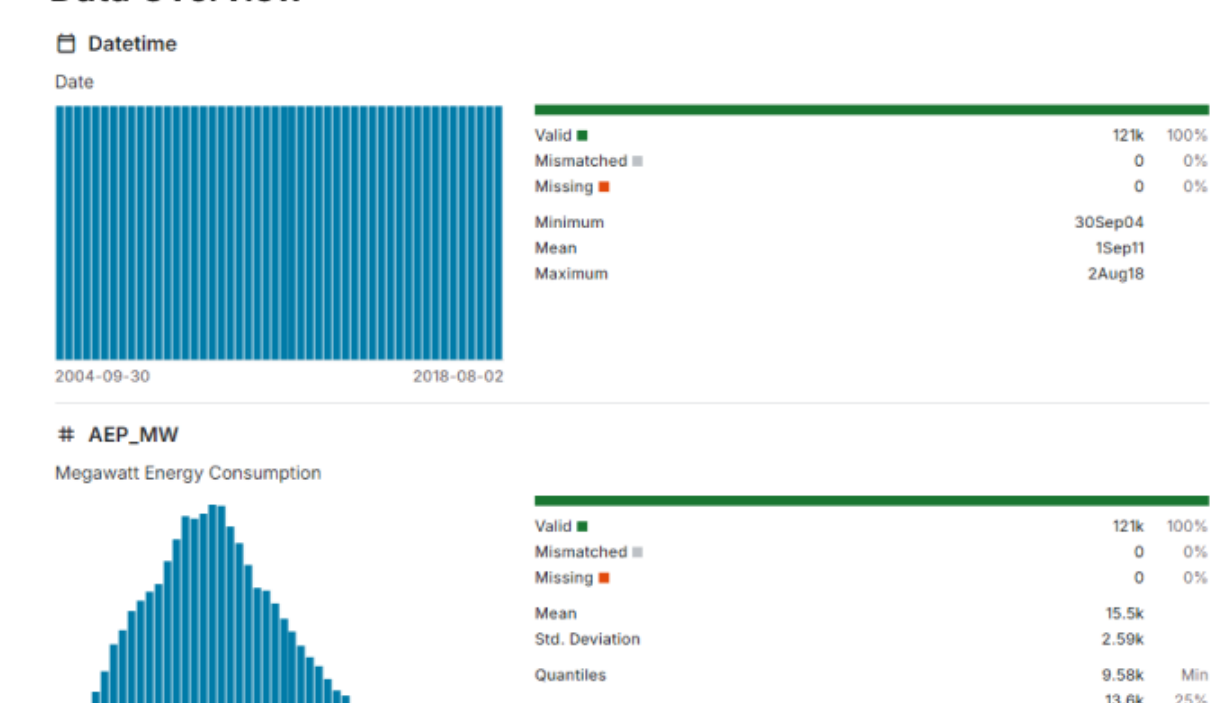
Previous queries

'Build a machine learning model to classify the type of appliance generating energy'
'Evaluate the relationship between energy consumption by seasons'

Example Rows

Date	# AEP_MW
2004-12-31 01:00:00	13478.0
2004-12-31 02:00:00	12865.0

Data Overview



Query Blocks

View edit history

task classification model for renewable energy trends

metadata (3)

usability > 70%

rows > 5000

columns > 20

SYSTEM: Hi, I'm chatGPT! Please start your dataset search with a task!

USER: I need a dataset to train a classification model for renewable energy trends.

SYSTEM UPDATED TASK BLOCK

SYSTEM:

Would you like to focus on solar, wind, or other types of renewable energy?

Are you interested in global trends or specific regions?

USER: I am interested in global solar energy trends datasets with good usability scores.

SYSTEM CREATED FILTER BLOCK

USER MANUALLY ADDED FILTER BLOCK (2)

Q

Next Steps

1. Providing "better" proactive guidance
 - How to inform search space for remaining metadata attributes?
 - How to convey which proposed change will result in the greatest amount of disparity?
2. Creating a baseline interface
 - Features to include: natural language queries, fixed filters, displaying search results
3. Evaluation against baseline interface with user studies

Work Referenced

Madelon Hulsebos, Wenjing Lin, Shreya Shankar, and Aditya Parameswaran. 2024. It Took Longer than I was Expecting: Why is Dataset Search Still so Hard? In Proceedings of the 2024 Workshop on Human-In-the-Loop Data Analytics (HILDA 24). Association for Computing Machinery, New York, NY, USA, 1–4. <https://doi.org/10.1145/3665939.3665959>