Developing AI-Driven Data Catalogs: Challenges and Opportunities

AnHai Doan

University of Wisconsin-Madison

Joint work with Ting Cai, Minh Pham, Amanpreet Saini, Stephen Sheen, Mark Tervo (UW-Madison), Gaurav Pathak (Informatica), Goetz Graefe (Google), Nan Tang, Mourad Ouzzani (QCRI)

Motivation

- **Organizations increasingly have many datasets**
  - Relational databases, noSQL databases, tables, files, images, emails, documents, …
- **Most projects use only a few datasets**
- **But finding them in a sea of datasets is often very difficult**

- **To solve this, organizations use data catalogs**

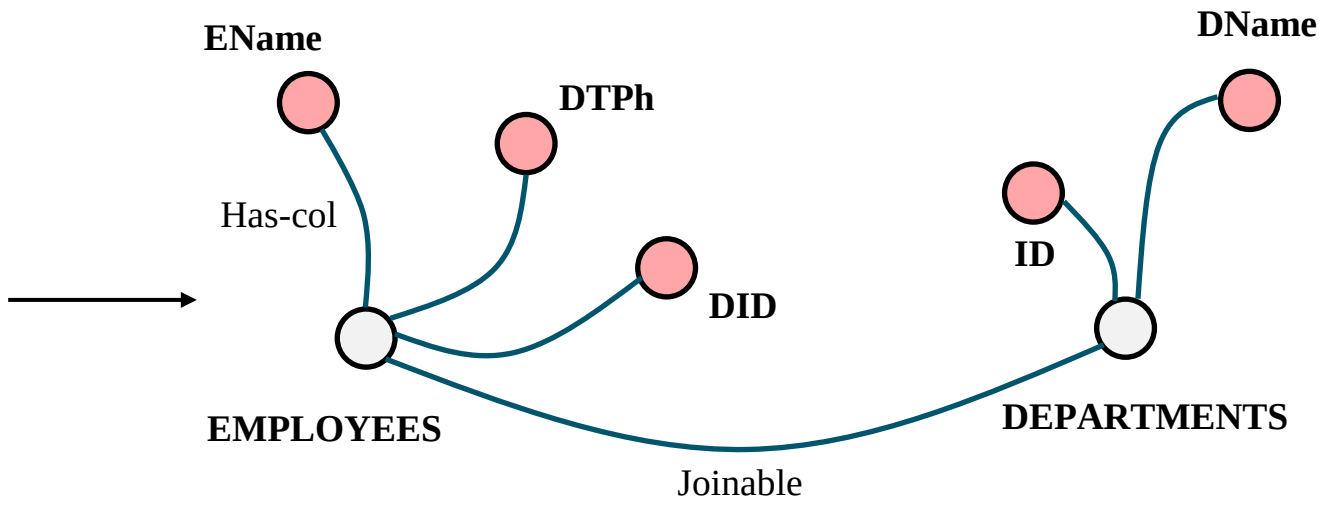
Example 1: Companies

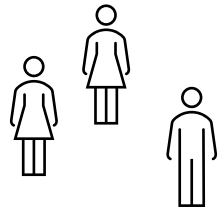
EMPLOYEES

EName	DTPh	DID
Dave Smith	4399	d_1
Jane Miller	5603	d_2
Mike Davis	2862	d_1

DEPARTMENTS

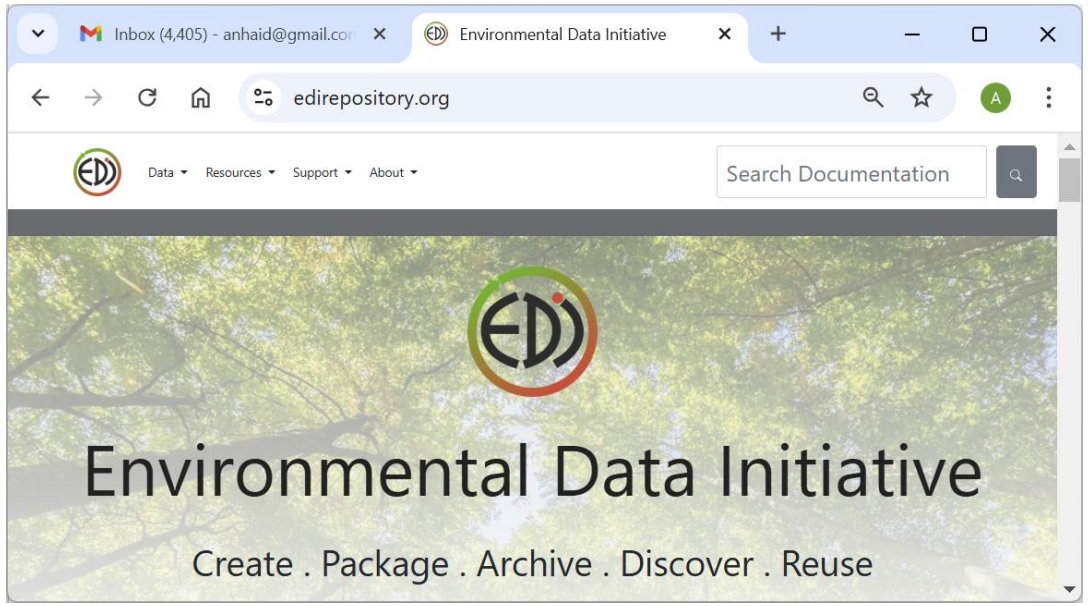
ID	DName
d_1	Sales
d_2	Legal

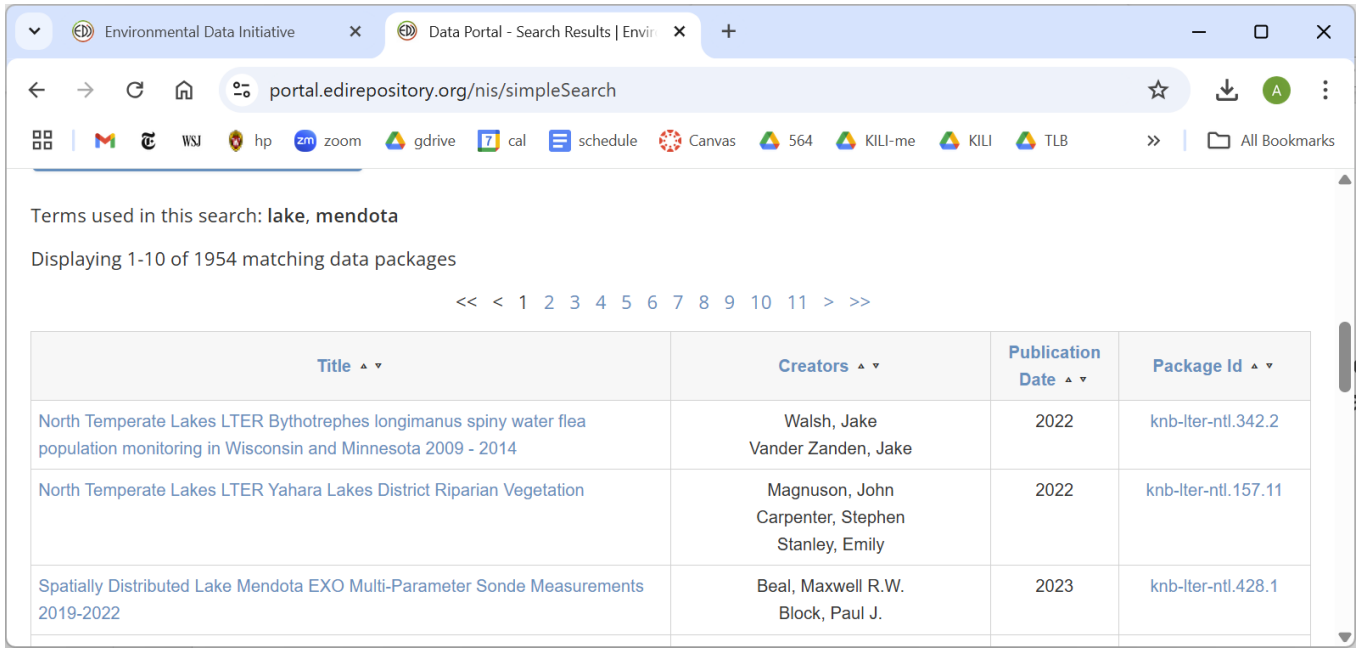




Keyword search
NL querying
Browsing

Example 2: Domain Sciences



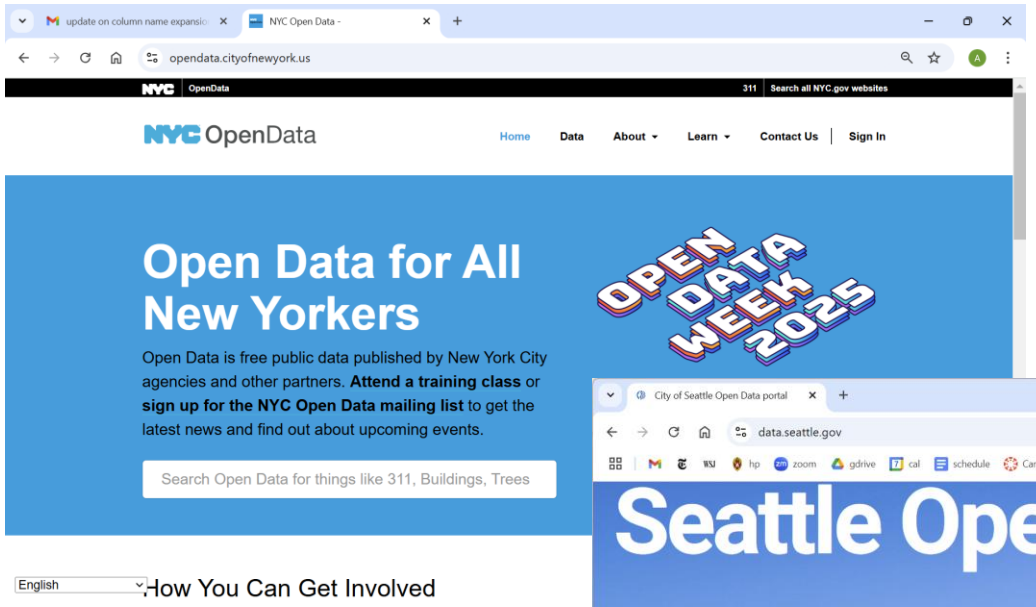


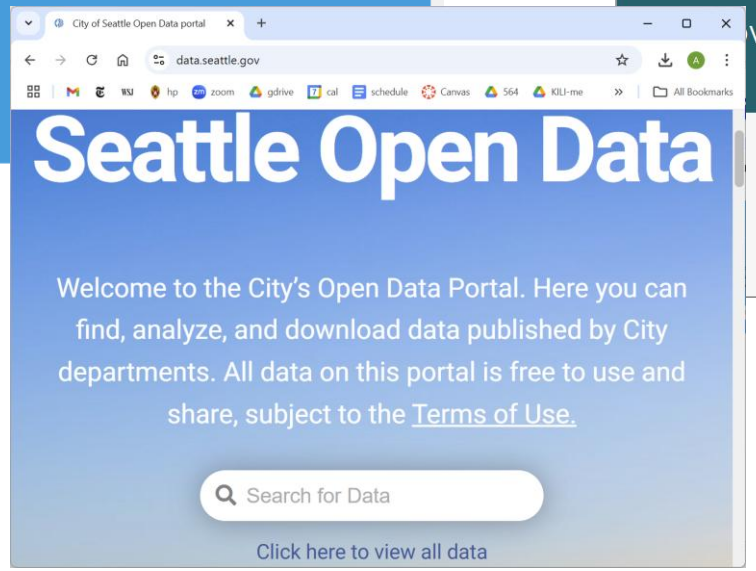
Since 2007
87K data packages, 18K tables
60TB storage, 10K downloads / week

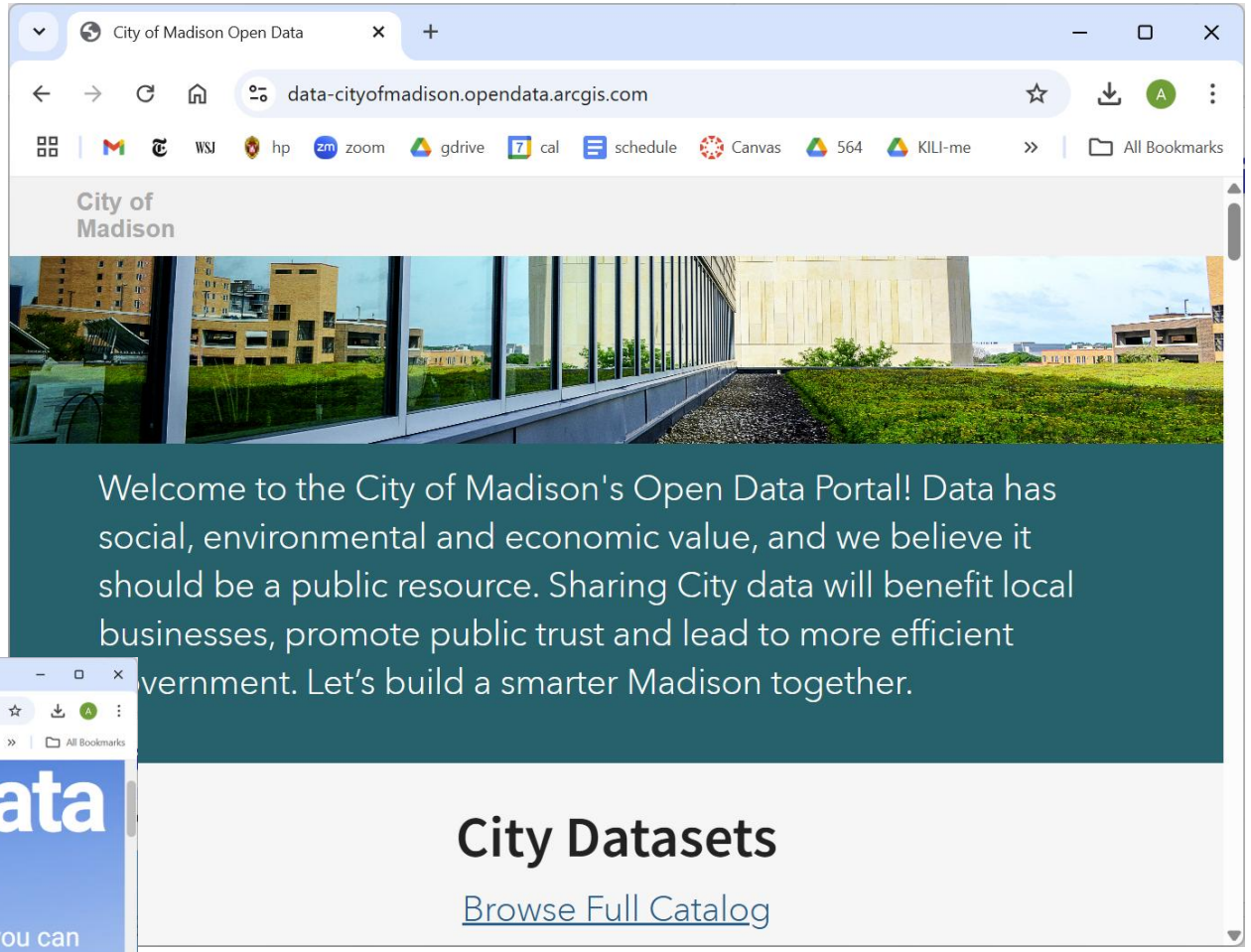
- **Other lakes in environmental science**
  - Dryad, CUAHSI, Zenodo, Figshare, BCO-DMO, Artic Data Center, IDigBio

Example 3: Government Agencies

Data is a valuable asset
Taxpayers pay for it
So they should have access







Current State of the Art

- **Critical enabler for DS and AI projects**
- **Lot of research**
  - Focus on a few problems, make many assumptions, evaluated on small datasets
  - Mainly produce papers
- **Little effort in building systems & working with customers**

- **We don't really know how good current research effort is**
- **We don't have good open-source catalog systems**
- **We cannot help "small fish" customers**
  - Domain sciences, government agencies, small business, citizen scientists

The SmartCat Project @ Wisconsin

- **Help “small fish” build catalogs quickly / advance research**
- **Build SmartCat, an open-source catalog system**
  - Focus on tables for now
- **Work with customers**
  - The EDI environmental science data lake team
- **Use GenAI**
- **Key findings so far**
  - Can do a lot to help small fish
  - GenAI very promising but
    - Less accurate on enterprise/domain science data
    - Sometimes doesn’t work
    - Must be combined with other technologies (e.g., Big Data scaling, RDBMS, curation)

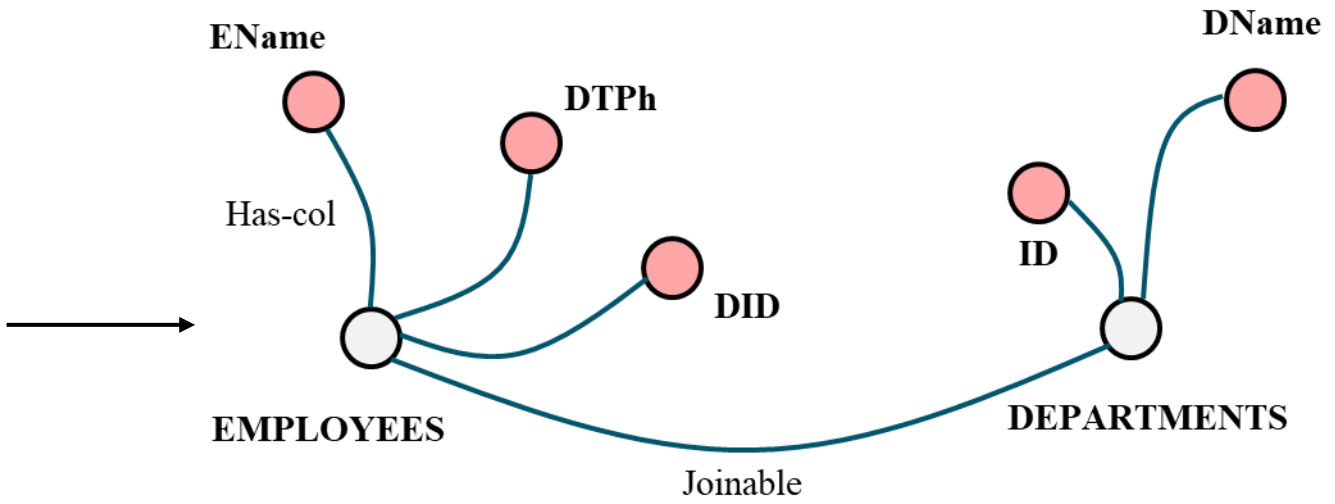
Five Key Challenges

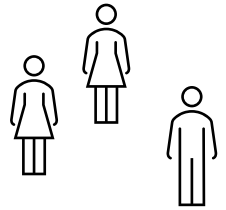
EMPLOYEES

EName	DTPh	DID
Dave Smith	4399	d_1
Jane Miller	5603	d_2
Mike Davis	2862	d_1

DEPARTMENTS

ID	DName
d_1	Sales
d_2	Legal





Keyword search
NL querying
Browsing

- **Inferring metadata**
- **Ways to find datasets**
- **Curation**
- **Scaling**
- **Handling changes**

Inferring Metadata

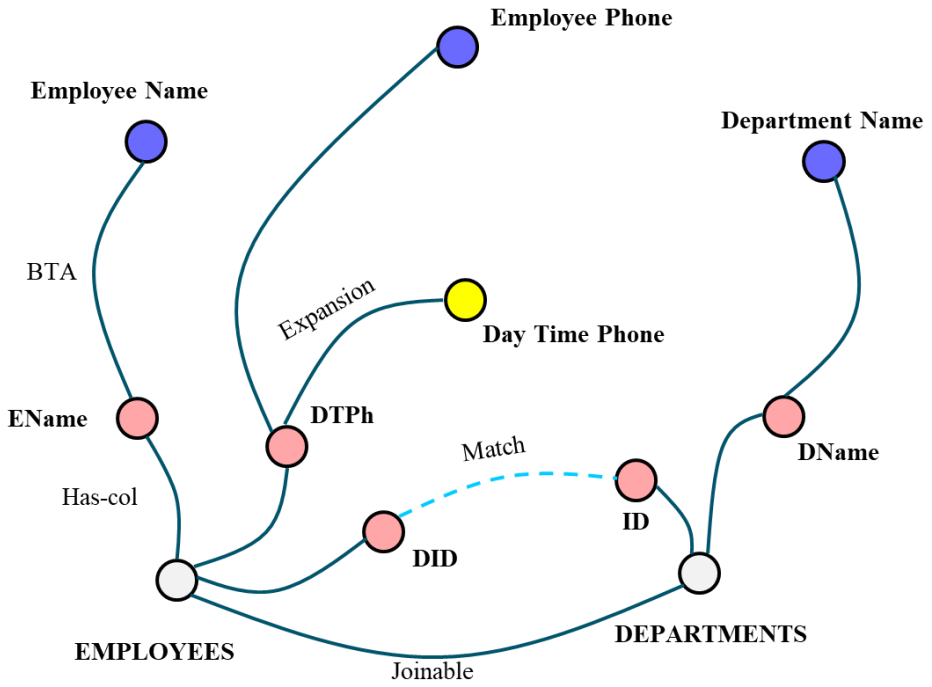
- Table/column name expansion
- Table/column description
- Tags
- Column annotation
- Business term association
- Key discovery
- Schema matching
- Joinable/unionable relationships
- Inferred lineage
- Related tables
- …

EMPLOYEES

EName	DTPh	DID
Dave Smith	4399	d_1
Jane Miller	5603	d_2
Mike Davis	2862	d_1

DEPARTMENTS

ID	DName
d_1	Sales
d_2	Legal



Table/Column Name Expansion

1997lgextnuts_csv (['Date', 'Site', 'Community', 'Core', 'Horizon', 'Code', 'Core location', 'CAN#', 'CANWT', 'WETWT', 'DRYWT', 'KCLWT', 'HCLWT', '1N HCLWT', 'MOISTURE', 'corwetwt', 'BulkDens', 'Dry_Wet', 'length', 'NH4 um/l', 'NO3 um/l', 'PO4 um/l', '1NPO4 um/l', 'NH4blank', 'NO3blank', 'PO4blank', '1NPO4blk', 'NH4 ug_g', 'NO3 ug_g', 'PO4 ug_g', '1NPO4 ug_g', 'NH4 g/m2', 'NO3 g/m2', 'PO4 g/m2', '1NPO4 g/m2', 'pH', 'N:P', 'COMMENTS'])

1997 Long Term Ecological Experiment Nutrient Study CSV (['Date', 'Site', 'Community', 'Horizon', 'Code', 'Core Location', 'Canister Number', 'Canister Weight', 'Wet Weight', 'Dry Weight', 'Potassium Chloride Weight', 'Hydrochloric Acid Weight', 'One Normal Hydrochloric Acid Weight', 'Moisture', 'Corrected Wet Weight', 'Bulk Density', 'Dry Wet', 'Length', 'Ammonium Micromoles per Liter', 'Nitrate Micromoles per Liter', 'Phosphate Micromoles per Liter', '1 Normal Phosphate Micromoles per Liter', 'Ammonium Blank', 'Nitrate Blank', 'Phosphate Blank', '1 Normal Phosphate Blank', 'Ammonium Micrograms per Gram', 'Nitrate Micrograms per Gram', 'Phosphate Micrograms per Gram', '1NPO4 Micrograms per Gram', 'Ammonium Grams per Square Meter', 'Nitrate Grams per Square Meter', 'Phosphate Grams per Square Meter', '1NPO4 Grams per Square Meter', 'Potential of Hydrogen', 'Nitrogen Phosphorus', 'Comments'])

- **Fundamental problem, yet very little attention so far**

The NameGuess Solution [EMNLP-23]

```
[developer]: You are a helpful assistant, answer the question from the user
and reply in the same format.
[user]: As abbreviations of column names from a table, c_name | pCd | dt
stand for
[assistant]: Customer Name | Product Code | Date
[user]: As abbreviations of column names from a table, SId | DbVer | Date |
ChangeDate stand for
```

- **69.3% EM accuracy with GPT4 (vs 43.4% EM accuracy with human)**
- **Limitations**
  - Uses just one public synthetic dataset (895 tables, 9218 columns)
  - Do not exploit table names
  - Very basic prompt

Our Solution

- **Evaluated on more datasets (enterprise, domain science)**
- **Exploits table names**
- **Uses rules & chains-of-thought reasoning**
  - Forces LLM to tokenize then expand each token

```
[developer] You are a helpful assistant, answer the question from the user and reply in the same format.

[user] Your task is to expand abbreviated column names into full-form phrases.
First reason step by step and then return your final answer.


Follow the guidelines below when you expand:
1. Expand all abbreviations in the column names.
2. Expand chemical symbols and units of measure to their full names.
3. Do not expand or mutate numbers.
4. Do not add extra words or explanations.
5. Maintain the original order of tokens in the expansion.
```

[Question] As abbreviations of column names from a table named Prchs_info, c_name | pCd | dt stand for? Think step by step

[Answer]
Reasoning
1. **Identify the table name context:**
   - The table is named **"Prchs_info"**, which likely stands for **"Purchase Information"**.
   - This suggests that the column names are related to purchase details.

2. **Break down each column name into tokens and infer meanings:**
   - **c_name** → `c`, `name`
     - `c` is commonly used for **Customer** in business contexts.
     - `name` clearly refers to **Name**.
     - Together, `c_name` most likely means **Customer Name**.

   - **pCd** → `p`, `Cd`
     - `p` is frequently used for **Product** in sales or inventory tables.
     - `Cd` is a common abbreviation for **Code**.
     - Together, `pCd` most likely means **Product Code**.

Final Answer
Customer Name | Product Code | Date

...

[Question] As abbreviations of column names from a table named PAF_SUBK, SUBK_FFATA_RPT_REQ_IND | SUBK_FULLY_EXCTD_DT | SUBK_FED_CNTRCT_IND stand for
[Answer]

Our Solution

Dataset	Domain	# Tables	# Columns	NameGuess Solution	Our Solution
NameGuess	Public	895	9218	81.5	81.87
Finance	Enterprise	23	443	73.81	87.81
University	Enterprise	122	1563	63.17	70.76
EDI	Domain Science	251	3830	65.3	85.9
AdventureWork 1	Enterprise	101	826	72.64	92.37
AdventureWork 2	Enterprise	101	826	74.82	94.31

- **Enterprise / domain science data is more difficult**

Generating Table Descriptions and Tags

STaxRate (STRGUID, STRTaxRate, STRProvinceId, STRId, STRTaxType, STRModDate, STRName)

• Description: The Sales Tax Rate table contains detailed information about various sales tax rates applicable across different provinces. Each entry is uniquely identified by a `Sales Tax Rate Globally Unique Identifier` and includes the `Sales Tax Rate Tax Rate`, which specifies the percentage of tax applied. The table also records the `Sales Tax Rate Province Identifier` to indicate the region where the tax rate is applicable, and the `Sales Tax Rate Tax Type` to categorize the type of tax. Additionally, it includes the `Sales Tax Rate Modified Date` to track changes over time and the `Sales Tax Rate Name` for easy reference.

• Tags: tax name, province, tax rate, sales tax, tax type, tax categorization, regional tax

1997lgextnuts_csv (['Date', 'Site', 'Community', 'Core', 'Horizon', 'Code', 'Core location', 'CAN#', 'CANWT', 'WETWT', 'DRYWT', 'KCLWT', 'HCLWT', '1N HCLWT', 'MOISTURE', 'corwetwt', 'BulkDens', 'Dry_Wet', 'length', 'NH4 um/l', 'NO3 um/l', 'PO4 um/l', '1NPO4 um/l', 'NH4blank', 'NO3blank', 'PO4blank', '1NPO4blk', 'NH4 ug_g', 'NO3 ug_g', 'PO4 ug_g', '1NPO4 ug_g', 'NH4 g/m2', 'NO3 g/m2', 'PO4 g/m2', '1NPO4 g/m2', 'pH', 'N:P', 'COMMENTS'])

1997 Long Term Ecological Experiment Nutrient Study CSV (['Date', 'Site', 'Community', 'Core', 'Horizon', 'Code', 'Core Location', 'Canister Number', 'Canister Weight', 'Wet Weight', 'Dry Weight', 'Potassium Chloride Weight', 'Hydrochloric Acid Weight', 'One Normal Hydrochloric Acid Weight', 'Moisture', 'Corrected Wet Weight', 'Bulk Density', 'Dry Wet', 'Length', 'Ammonium Micromoles per Liter', 'Nitrate Micromoles per Liter', 'Phosphate Micromoles per Liter', '1 Normal Phosphate Micromoles per Liter', 'Ammonium Blank', 'Nitrate Blank', 'Phosphate Blank', '1 Normal Phosphate Blank', 'Ammonium Micrograms per Gram', 'Nitrate Micrograms per Gram', 'Phosphate Micrograms per Gram', '1NPO4 Micrograms per Gram', 'Ammonium Grams per Square Meter', 'Nitrate Grams per Square Meter', 'Phosphate Grams per Square Meter', '1NPO4 Grams per Square Meter', 'Potential of Hydrogen', 'Nitrogen Phosphorus', 'Comments'])

• Description: The **1997 Long Term Ecological Experiment Nutrient Study CSV** table provides comprehensive data on nutrient levels and soil characteristics from various ecological sites. Key columns include *Date*, *Site*, *Community*, and *Core*, which identify the sampling details, while measurements such as *Ammonium Micromoles per Liter*, *Nitrate Micromoles per Liter*, and *Phosphate Micromoles per Liter* offer insights into nutrient concentrations. This table supports ecological research by enabling analysis of nutrient dynamics and soil properties across different environments.

• Tags: ['soil dynamics', 'ecological data 1997', 'environmental research', 'site sampling', 'nutrient analysis', 'soil characteristics', 'nutrient concentrations', 'ecological study']

Inferring Metadata

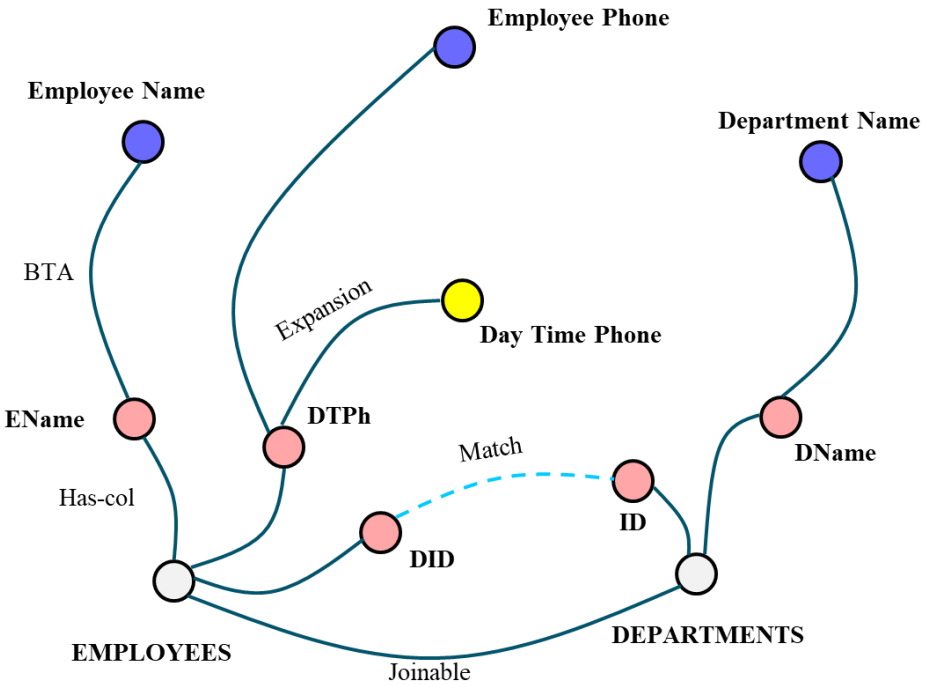
- Table/column name expansion
- Table/column description
- Tags
- Column annotation
- Business term association
- Key discovery
- Schema matching
- Joinable/unionable relationships
- Inferred lineage
- Related tables
- …

EMPLOYEES

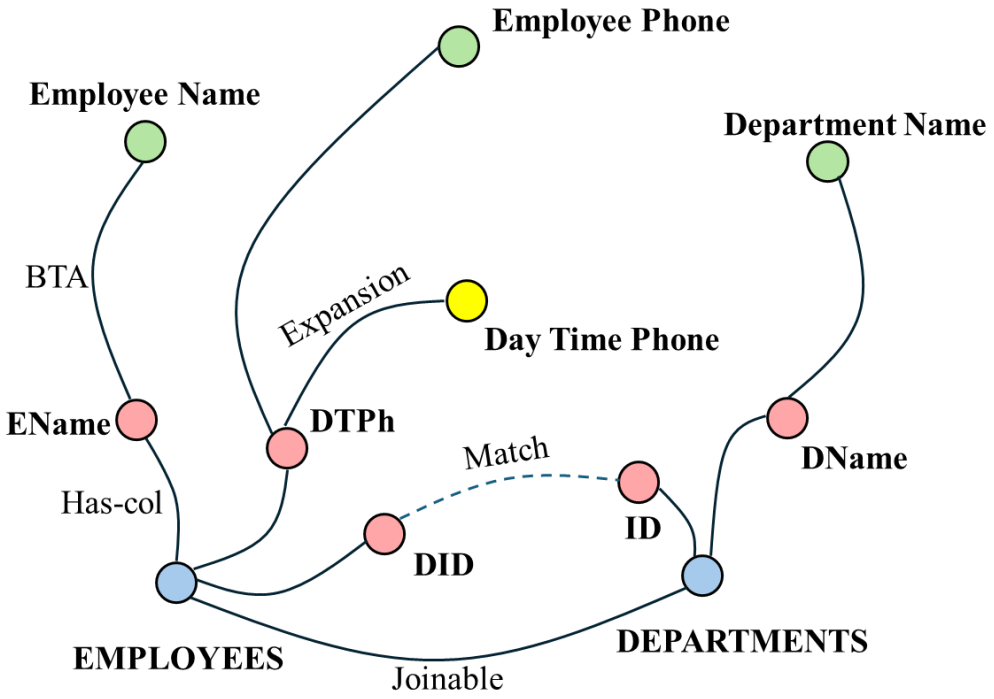
EName	DTPh	DID
Dave Smith	4399	d_1
Jane Miller	5603	d_2
Mike Davis	2862	d_1

DEPARTMENTS

ID	DName
d_1	Sales
d_2	Legal



Business Term Association (BTA)



t_1: Employee Name
t_2: Employee Phone
t_3: Department Name

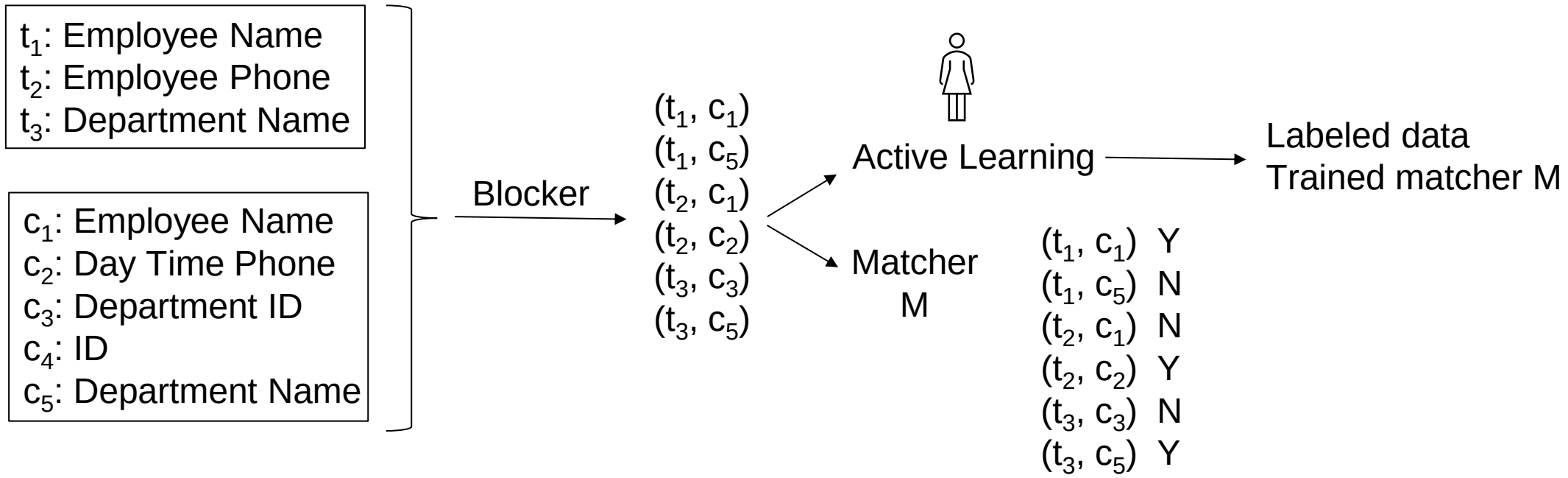
c_1: Employee Name
c_2: Day Time Phone
c_3: Department ID
c_4: ID
c_5: Department Name

Blocker → (t_1, c_1) (t_1, c_5) (t_2, c_1) (t_2, c_2) (t_3, c_3) (t_3, c_5)

Matcher → (t_1, c_1) Y, (t_1, c_5) N, (t_2, c_1) N, (t_2, c_2) Y, (t_3, c_3) N, (t_3, c_5) Y

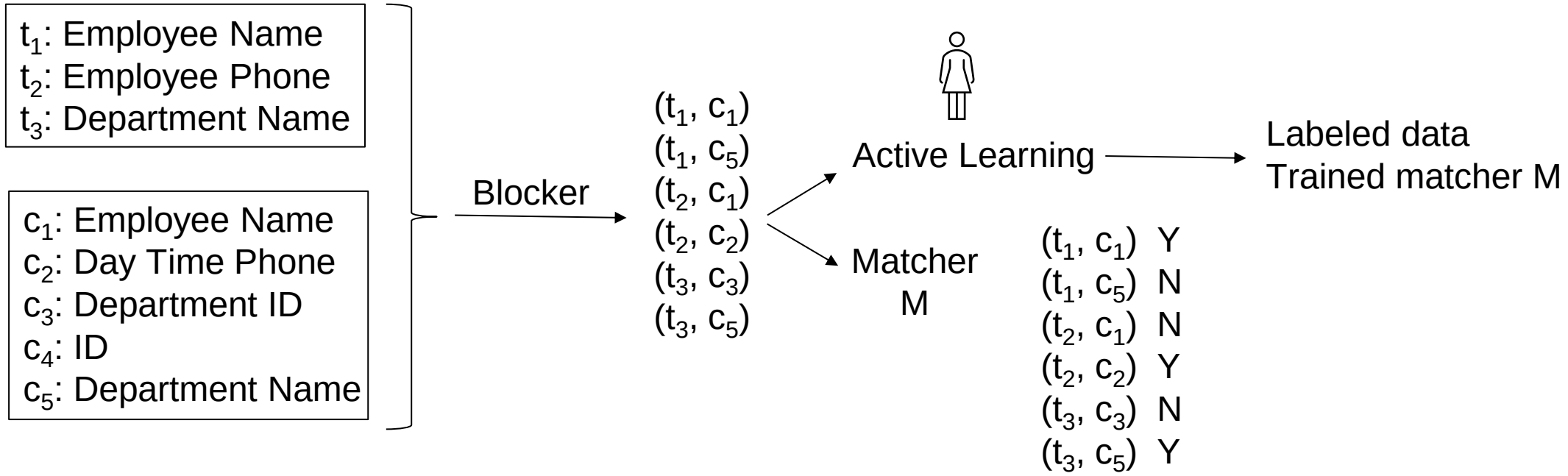
- **Zero-shot matchers**
  - Unicorn, Llama, GPT4, do not do well, too liberal
- **Need training data**
  - Use active learning

Active Learning for BTA



- **AL with random forest**
  - works well, feature engineering helps
- **AL with more complex ML models, including fine-tuning ZS matchers**
  - Can't generalize well from a small number of labeled examples
- **Use ZS matchers to label examples for AL**
  - Does not work well, too liberal

Summary for BTA



- **GenAI does not work well (for now)**
  - Too liberal in predicting matches
- **Need training data, active learning is a good way to quickly get some**
- **This training data is small, so simpler ML models work better**

Inferring Metadata

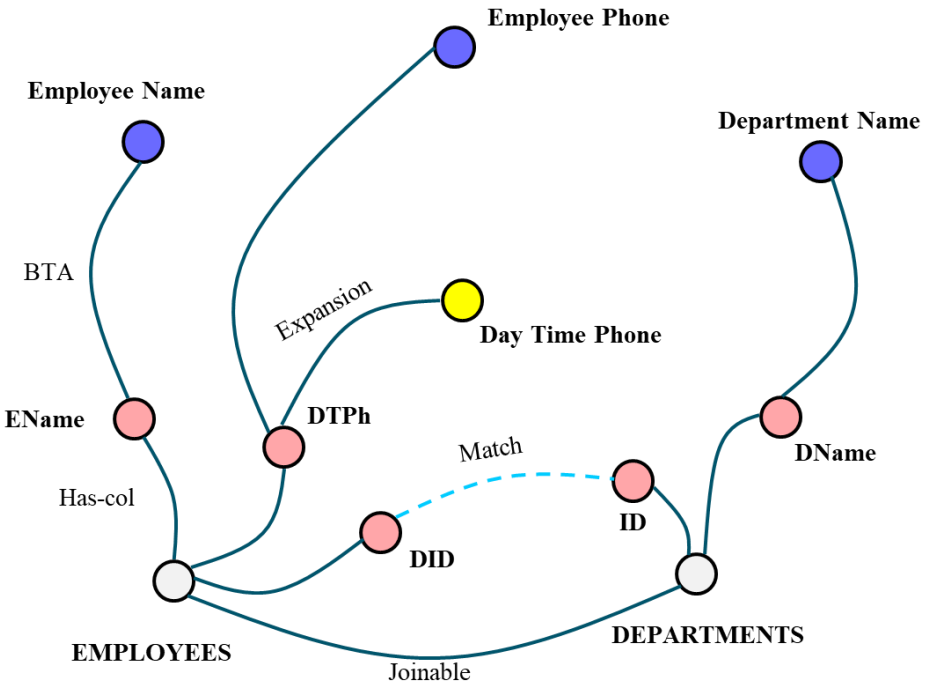
- Table/column name expansion
- Table/column description
- Tags
- Column annotation
- Business term association
- Key discovery
- Schema matching
- Joinable/unionable relationships
- Inferred lineage
- Related tables
- …

EMPLOYEES

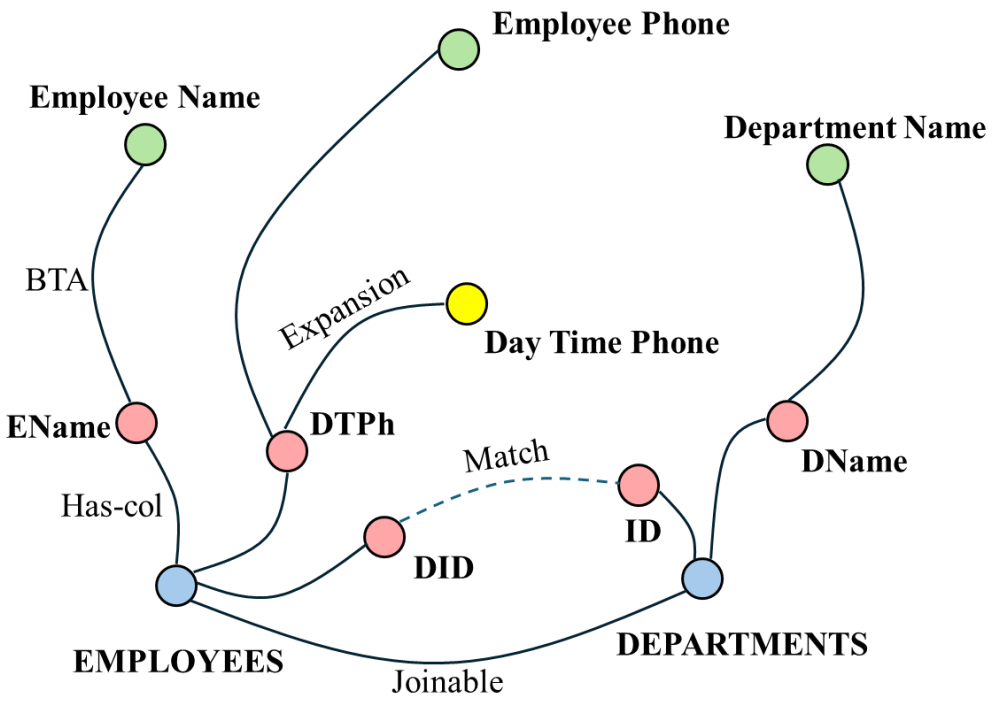
EName	DTPh	DID
Dave Smith	4399	d_1
Jane Miller	5603	d_2
Mike Davis	2862	d_1

DEPARTMENTS

ID	DName
d_1	Sales
d_2	Legal



Schema Matching for a Data Catalog



X

id	name	loc
x_1	Apple	CA
x_2	IBM	NY

Z

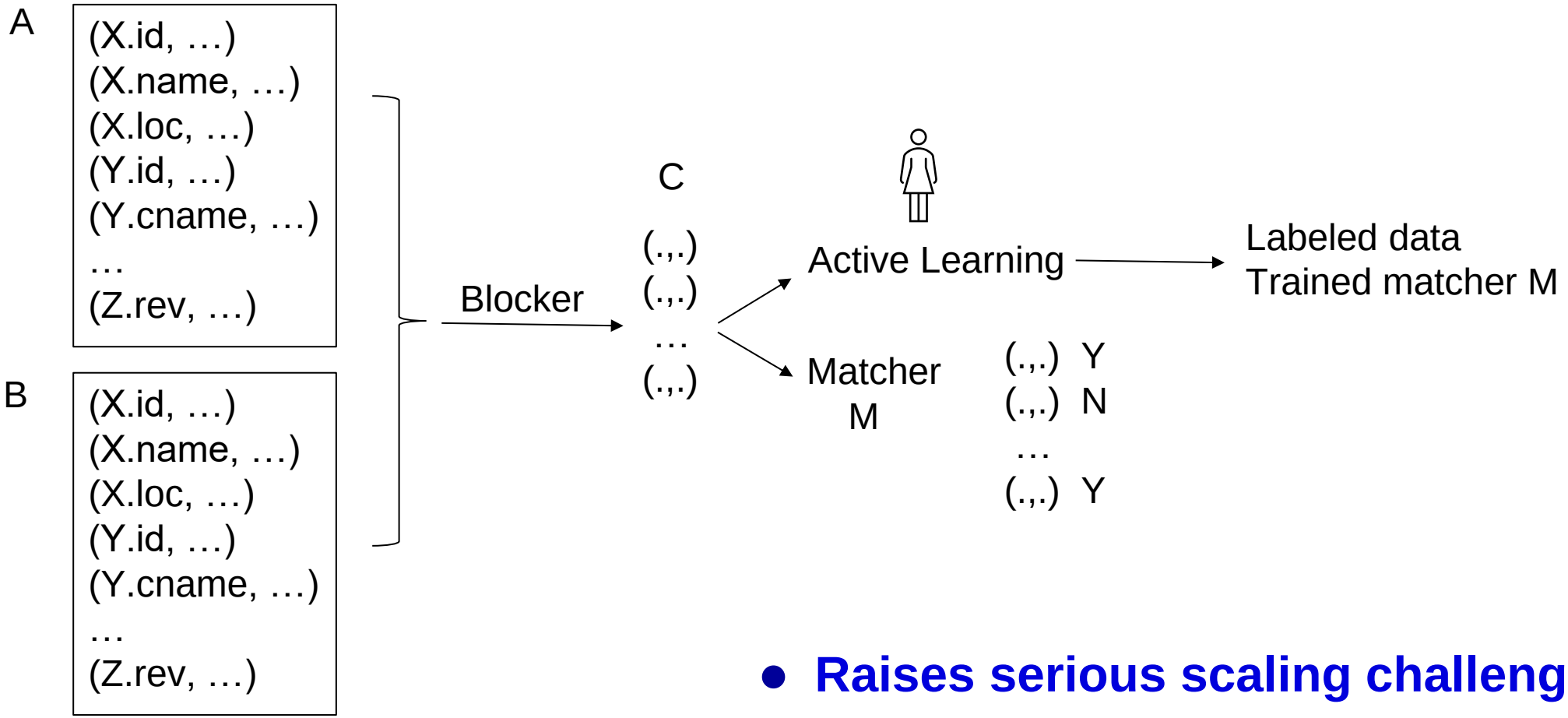
name	loc	rev
Apple Inc	CA	51
IBM Corp	NY	25

Y

id	cname	address	rev
y_1	IBM Corp	CA	25
y_2	Apple Inc	CA	51
y_3	GE	NY	351

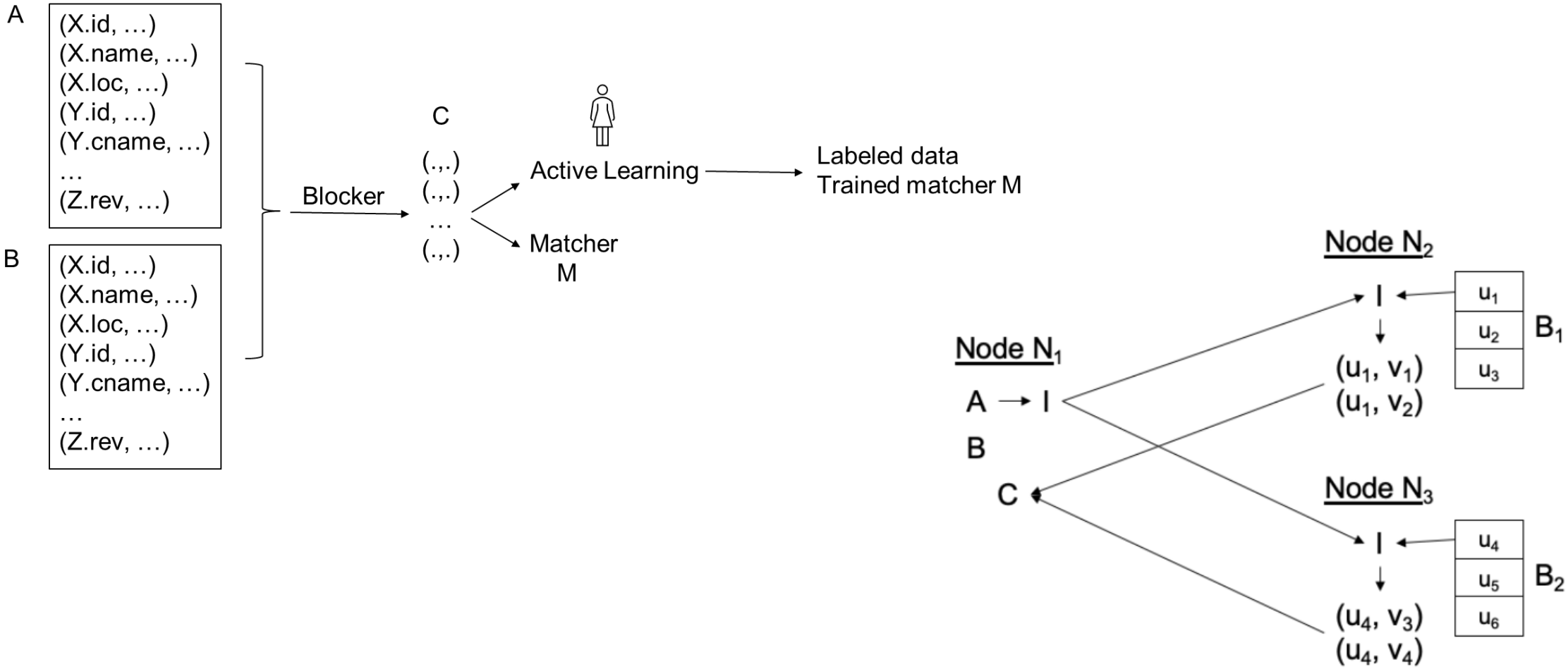
- Another fundamental problem
- Used to discover unionable/joinable tables, related tables, inferred lineage

Our Solution



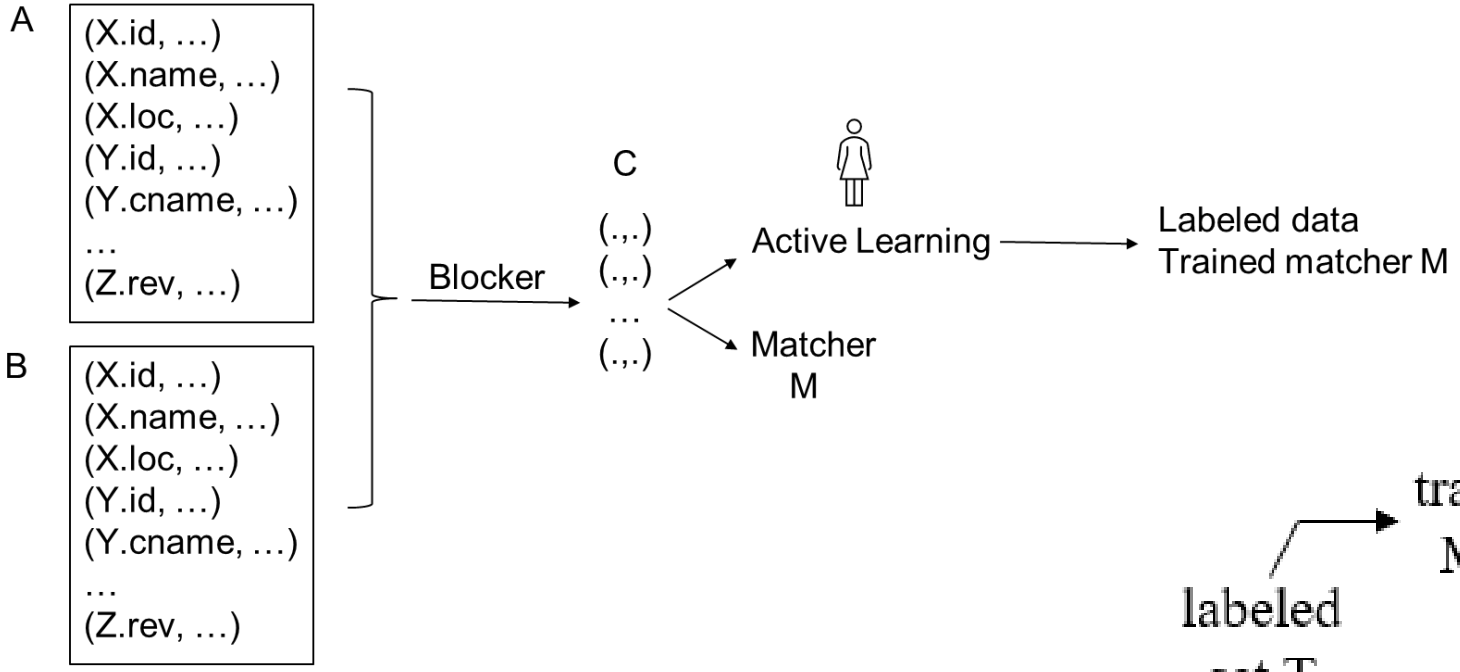
- **Raises serious scaling challenges**

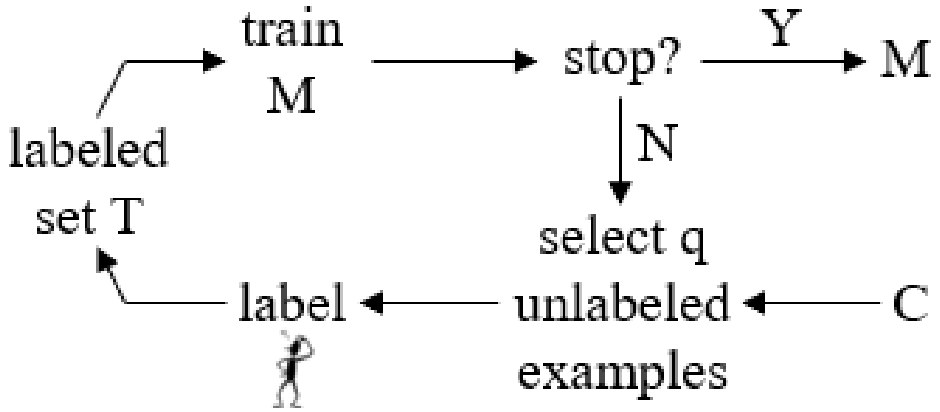
Scaling the Blocker



- **Outperforms existing blockers [VLDB-23]**

Scaling the Matcher





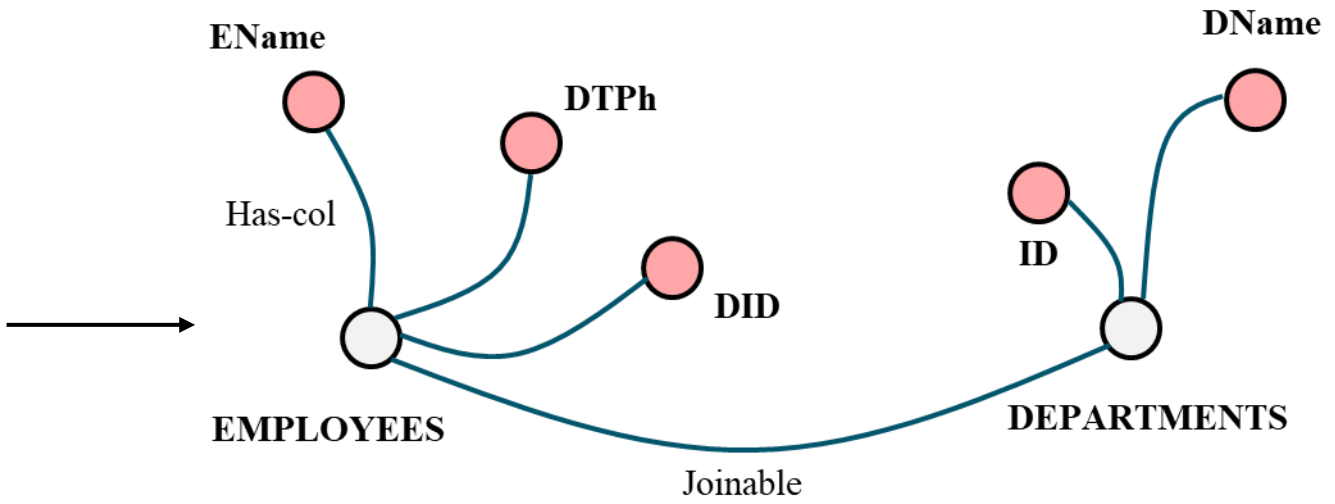
Five Key Challenges

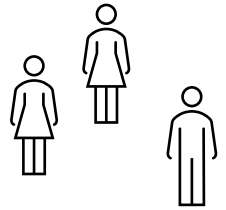
EMPLOYEES

EName	DTPh	DID
Dave Smith	4399	d_1
Jane Miller	5603	d_2
Mike Davis	2862	d_1

DEPARTMENTS

ID	DName
d_1	Sales
d_2	Legal





Keyword search
NL querying
Browsing

- **Inferring metadata**
- **Ways to find datasets**
- **Curation**
- **Scaling**
- **Handling changes**

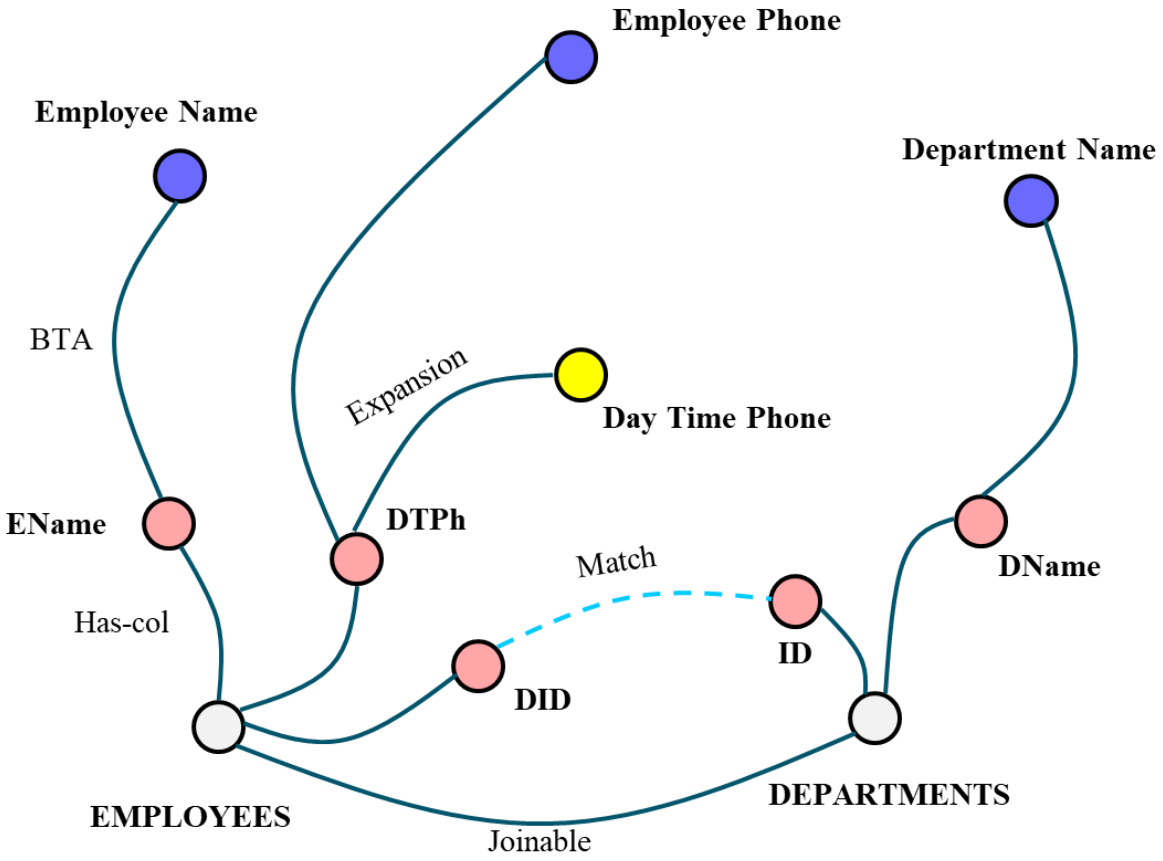
Ways to Find Datasets

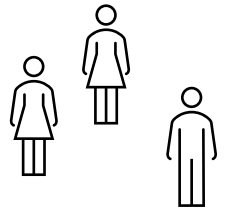
EMPLOYEES

EName	DTPh	DID
Dave Smith	4399	d_1
Jane Miller	5603	d_2
Mike Davis	2862	d_1

DEPARTMENTS

ID	DName
d_1	Sales
d_2	Legal





Keyword search
NL querying
Browsing

NL Querying

NL query Q → Process Q to create prompt P → LLM → Query S over catalog graph G → Postprocess S → Execute S over graph G

What is the number of cars with more than 4 cylinders?

```
SELECT COUNT(*)
FROM cars_data
WHERE cylinders > 4
```

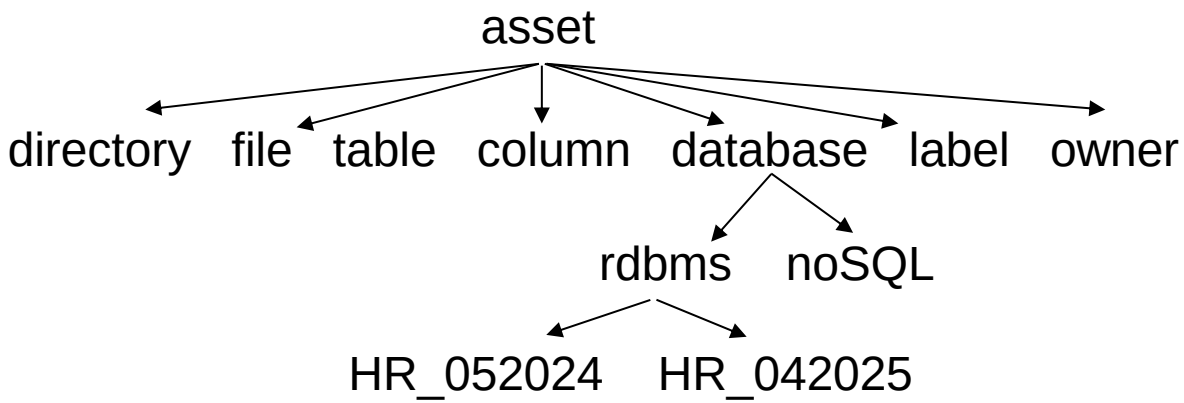
For each stadium, how many concerts are there?

```
SELECT T2.name, COUNT(*)
FROM concert AS T1 JOIN stadium AS T2
ON T1.stadium_id = T2.stadium_id
GROUP BY T1.stadium_id
```

Which countries in Europe have at least 3 car manufacturers?

```
SELECT T1.country_name
FROM countries AS T1 JOIN continents
AS T2 ON T1.continent = T2.cont_id
JOIN car_makers AS T3 ON
T1.country_id = T3.country
WHERE T2.continent = 'Europe'
GROUP BY T1.country_name
HAVING COUNT(*) >= 3
```

- **Lower accuracy than on public data**
- **Lot of organization-specific synonyms**



Curation

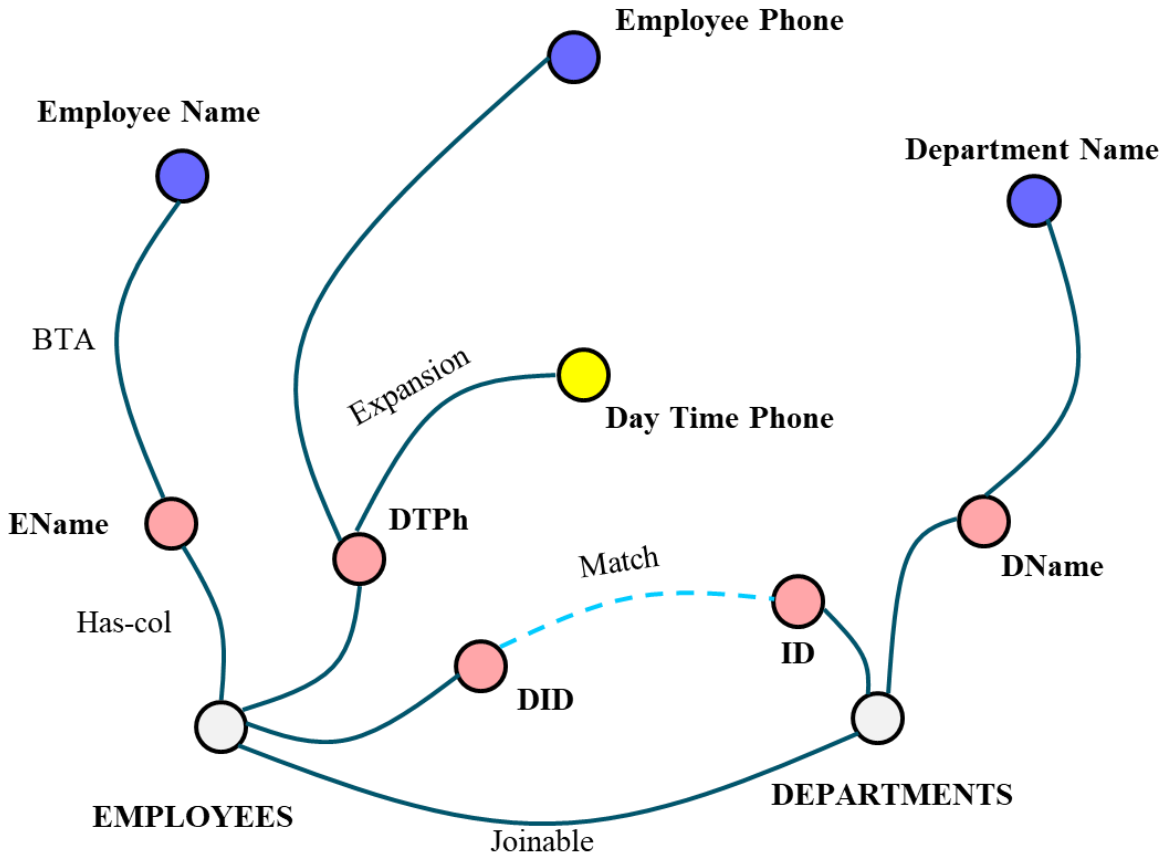
- **Bulk curation & drive-by curation**

EMPLOYEES

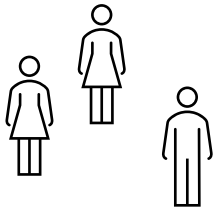
EName	DTPh	DID
Dave Smith	4399	d_1
Jane Miller	5603	d_2
Mike Davis	2862	d_1

DEPARTMENTS

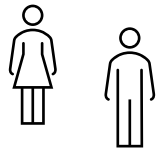
ID	DName
d_1	Sales
d_2	Legal



Users



Data stewards

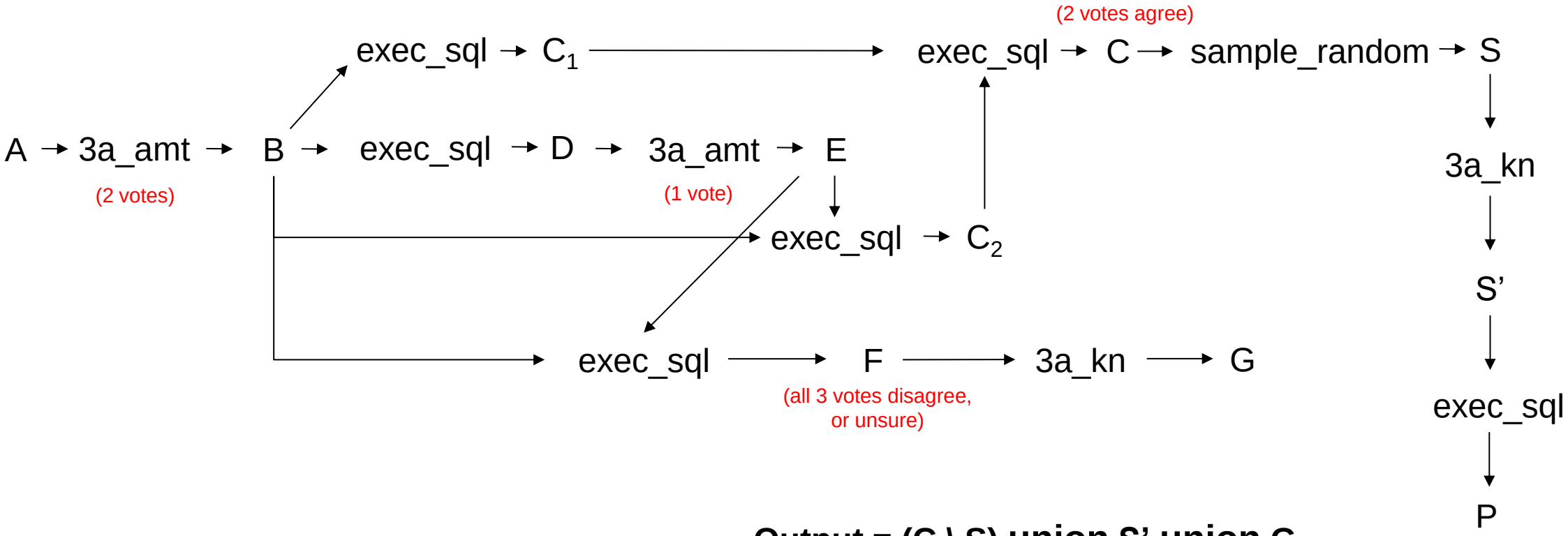


External
crowd platforms

amazon
mechanical turk

Curation Challenges

- **Often need many different curation workflows**
  - Table name expansion: 3 users agree OR 1 data steward agrees
  - Column name expansion: 2 users agree OR 1 data steward agrees
- **Workflows can be quite complex**



Output = (C \ S) union S' union G

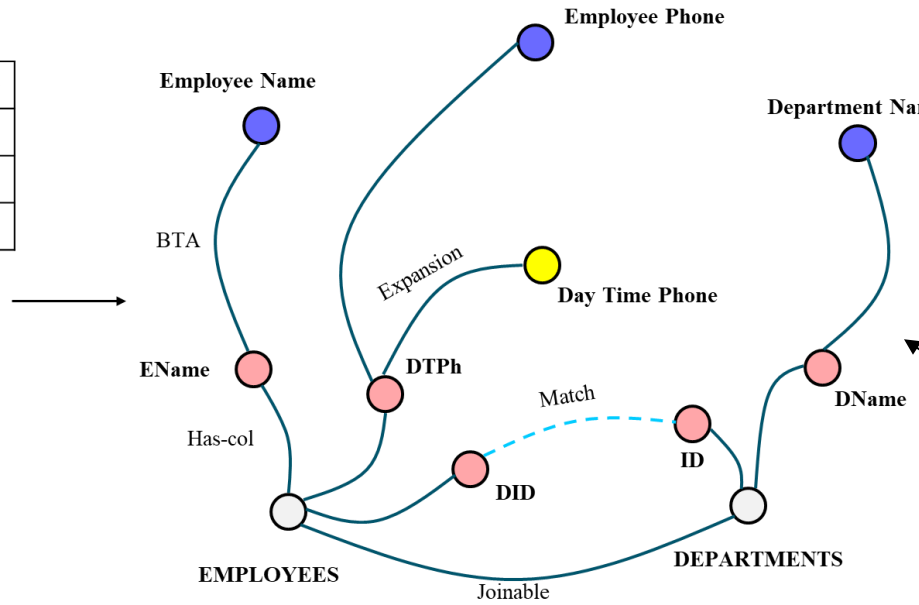
Our Solution: Cymphony

EMPLOYEES

EName	DTPh	DID
Dave Smith	4399	d_1
Jane Miller	5603	d_2
Mike Davis	2862	d_1

DEPARTMENTS

ID	DName
d_1	Sales
d_2	Legal

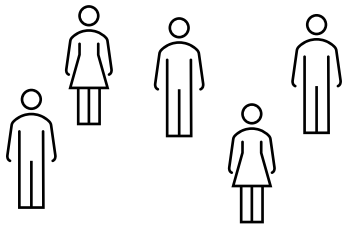



```
T1 = read_table("data1.csv");
(B1,C1) = 3a_kn(T1,"instruction1.html", k=2,n=3);
D1 = sample_random(C1,500);
D2 = exec_sql(SQL query to drop column final_label from D1 and make a new copy, D1)
(B2,C2) = 3a_kn(D2,"instruction1.html", …);
prec = exec_sql(SQL query to compute precision, D1, C2);
write_table(prec,"C1-precision.csv");
```

data1.csv → read_table → T1 → 3a_kn → C1 → sample_random → D1

exec_sql

D2 → 3a_kn → C2 → exec_sql → prec → write_table → C1-precision.csv



amazon mechanical turk

Handling Changes

Expand names

Business term association

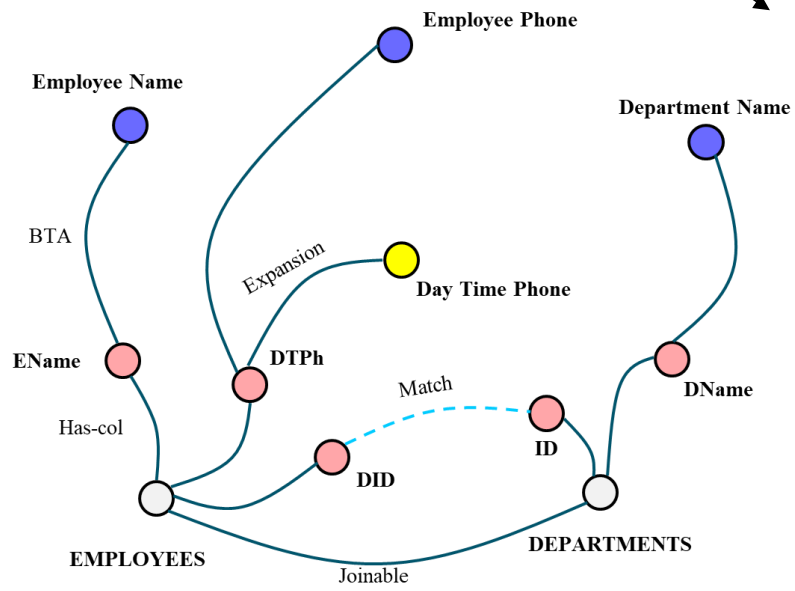
Schema matching

EMPLOYEES

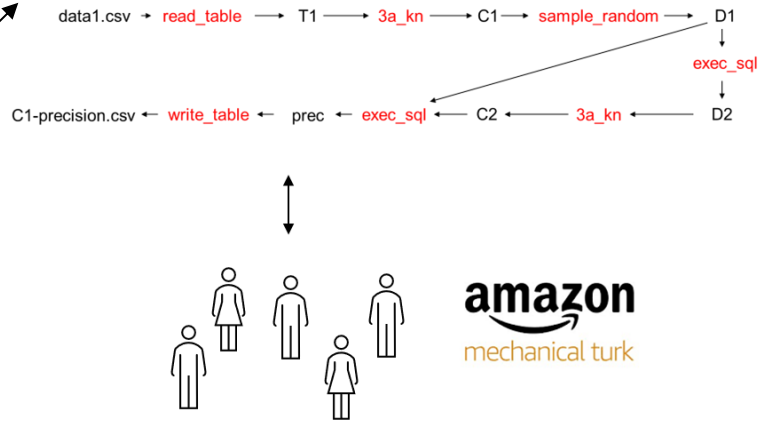
EName	DTPh	DID
Dave Smith	4399	d_1
Jane Miller	5603	d_2
Mike Davis	2862	d_1

DEPARTMENTS

ID	DName
d_1	Sales
d_2	Legal




```
T1 = read_table("data1.csv");
(B1,C1) = 3a_kn(T1,"instruction1.html", k=2,n=3);
D1 = sample_random(C1,500);
D2 = exec_sql(SQL query to drop column final_label from D1 and make a new copy, D1)
(B2,C2) = 3a_kn(D2,"instruction1.html", …);
prec = exec_sql(SQL query to compute precision, D1, C2);
write_table(prec,"C1-precision.csv");
```



TF/IDF index → Keyword search
NL querying
Browsing

- **Must enable incremental execution**
- **Build on results in incremental view maintenance of RDBMS**
  - In collaboration with Goetz Graefe @ Google Madison

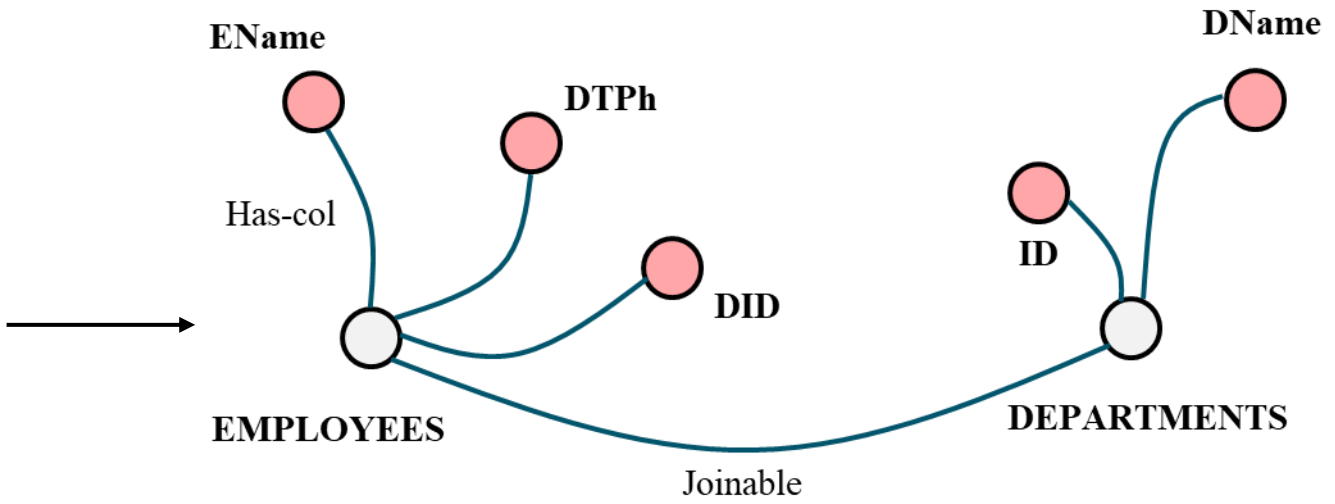
Five Key Challenges

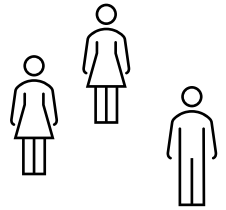
EMPLOYEES

EName	DTPh	DID
Dave Smith	4399	d_1
Jane Miller	5603	d_2
Mike Davis	2862	d_1

DEPARTMENTS

ID	DName
d_1	Sales
d_2	Legal

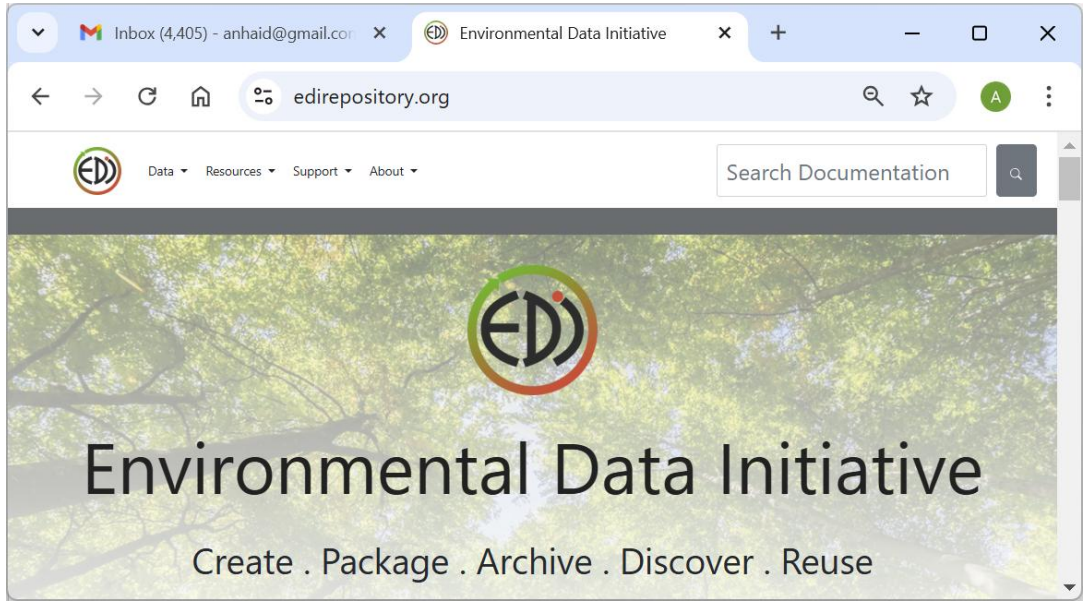




Keyword search
NL querying
Browsing

- **Inferring metadata**
- **Ways to find datasets**
- **Curation**
- **Scaling**
- **Handling changes**

Working with Customers



1997lgextnuts_csv (['Date', 'Site', 'Community', 'Core', 'Horizon', 'Code', 'Core location', 'CAN#', 'CANWT', 'WETWT', 'DRYWT', 'KCLWT', 'HCLWT', '1N HCLWT', 'MOISTURE', 'corwetwt', 'BulkDens', 'Dry_Wet', 'length', 'NH4 um/l', 'NO3 um/l', 'PO4 um/l', '1NPO4 um/l', 'NH4blank', 'NO3blank', 'PO4blank', '1NPO4blk', 'NH4 ug_g', 'NO3 ug_g', 'PO4 ug_g', '1NPO4 ug_g', 'NH4 g/m2', 'NO3 g/m2', 'PO4 g/m2', '1NPO4 g/m2', 'pH', 'N:P', 'COMMENTS'])

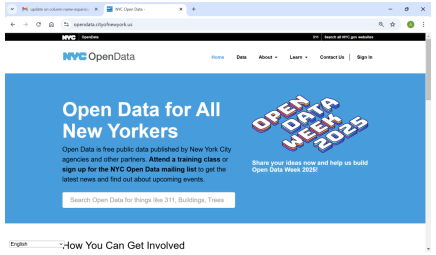
1997 Long Term Ecological Experiment Nutrient Study CSV (['Date', 'Site', 'Community', 'Core', 'Horizon', 'Code', 'Core Location', 'Canister Number', 'Canister Weight', 'Wet Weight', 'Dry Weight', 'Potassium Chloride Weight', 'Hydrochloric Acid Weight', 'One Normal Hydrochloric Acid Weight', 'Moisture', 'Corrected Wet Weight', 'Bulk Density', 'Dry Wet', 'Length', 'Ammonium Micromoles per Liter', 'Nitrate Micromoles per Liter', 'Phosphate Micromoles per Liter', '1 Normal Phosphate Micromoles per Liter', 'Ammonium Blank', 'Nitrate Blank', 'Phosphate Blank', '1 Normal Phosphate Blank', 'Ammonium Micrograms per Gram', 'Nitrate Micrograms per Gram', 'Phosphate Micrograms per Gram', '1NPO4 Micrograms per Gram', 'Ammonium Grams per Square Meter', 'Nitrate Grams per Square Meter', 'Phosphate Grams per Square Meter', '1NPO4 Grams per Square Meter', 'Potential of Hydrogen', 'Nitrogen Phosphorus', 'Comments'])

• Description: The **1997 Long Term Ecological Experiment Nutrient Study CSV** table provides comprehensive data on nutrient levels and soil characteristics from various ecological sites. Key columns include *Date*, *Site*, *Community*, and *Core*, which identify the sampling details, while measurements such as *Ammonium Micromoles per Liter*, *Nitrate Micromoles per Liter*, and *Phosphate Micromoles per Liter* offer insights into nutrient concentrations. This table supports ecological research by enabling analysis of nutrient dynamics and soil properties across different environments.

• Tags: ['soil dynamics', 'ecological data 1997', 'environmental research', 'site sampling', 'nutrient analysis', 'soil characteristics', 'nutrient concentrations', 'ecological study']

Since 2007
87K data packages, 18K tables
60TB storage, 10K downloads / week

- **Other lakes in environmental science**
  - Dryad, CUAHSI, Zenodo, Figshare, BCO-DMO, Artic Data Center, IDigBio



Conclusions

- **Building data catalogs is critical for DS & AI projects**
  - Cuts across enterprises, domain sciences, government agencies

- **Lot of research, isolated problems, mostly paper output**
- **Must build systems and work with customers**
- **SmartCat @ UW-Madison seeks to do so**
  - In collaboration with Google, Informatica, environmental sciences
  - Seeks to help small fish & advance research

- **Key findings**
  - We can do a lot to help small fish
  - GenAI very promising, but less accurate than on public data
  - Doesn’t work well for some problems
  - Must combine with other technologies (scaling, curation, RDBMS)