Madelon Hulsebos

Berkeley UNIVERSITY OF CALIFORNIA

CWI

Table Representation Learning

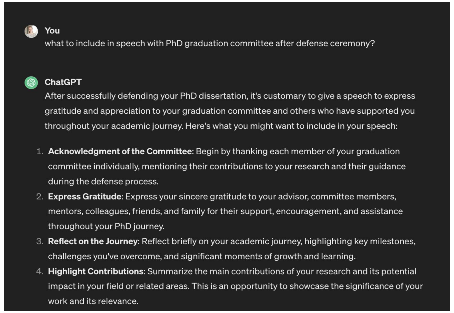
UW Seattle

May 24th, 2024

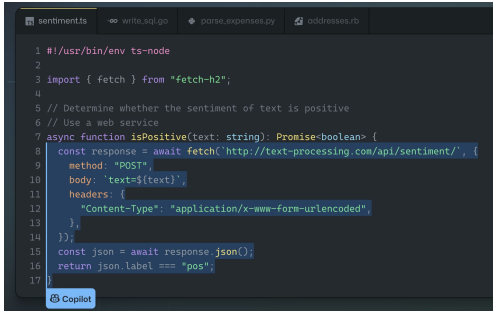
The Impressive Capabilities of Transformers

Transformers leveraged for applications over images, text, code:

Generates images of dogs



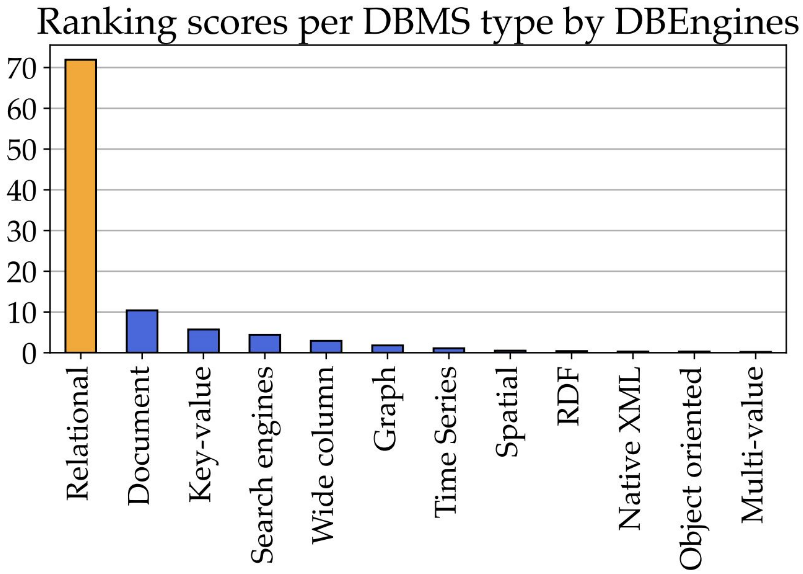
Helps writing graduation speech

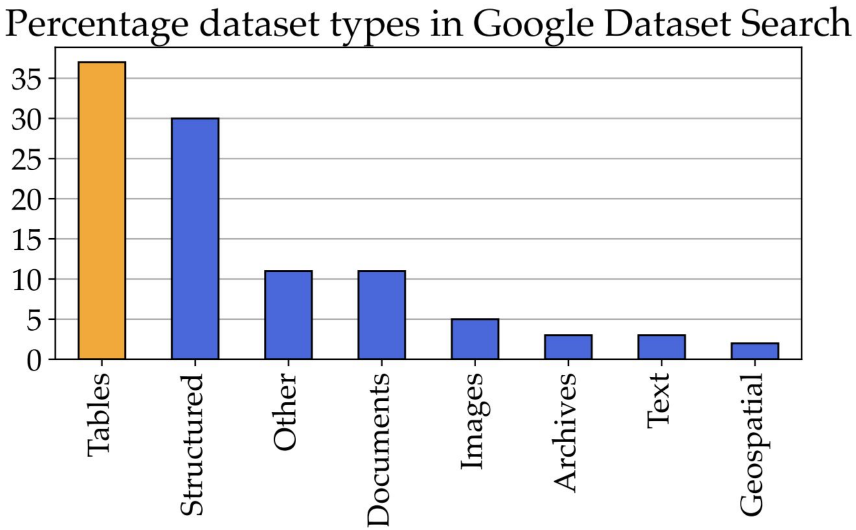


Completes code

What about **tables**?

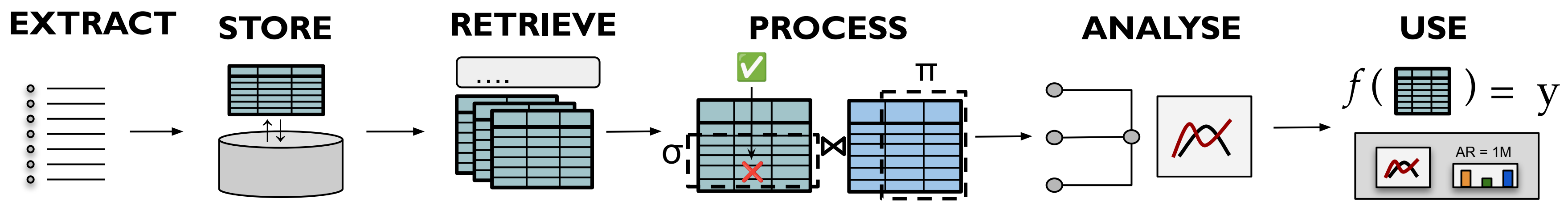
Tables Dominate the Data Landscape





Application Potential of Table Representations

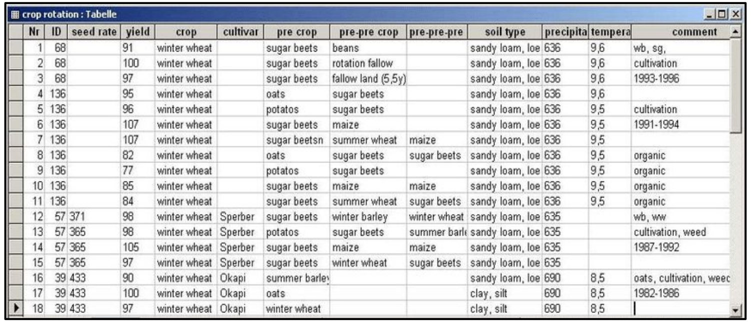
Many tasks in **high-value use-cases** operate over tables, e.g. *data analytics*.



• parsing • integration • validation	• storage • query processing • cataloging	• exploration • querying • monitoring	• transformation • augmentation • cleaning	• aggregation • statistics • visualization	• insight extraction • machine learning • dashboarding

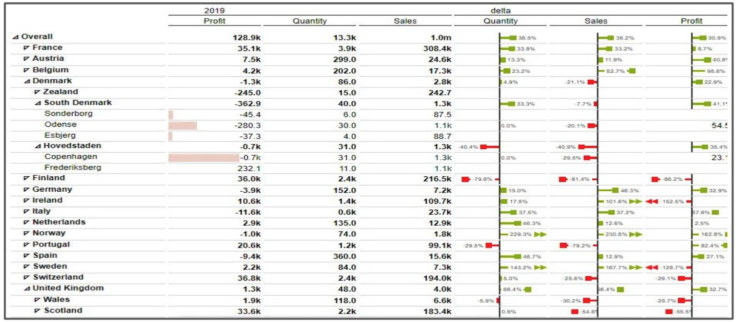
Rich and challenging!

Tables come with diversity in structure, dimensions, content, and semantics…



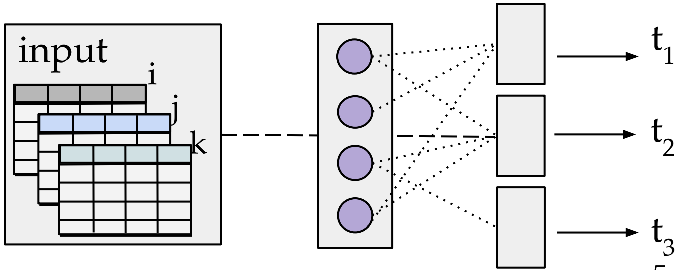
crop rotation : Tabelle

Nr	ID	seed rate	yield	crop	cultivar	pre crop	pre-pre crop	pre-pre-pre	soil type	precipita	tempera	comment
1	68		91	winter wheat		sugar beets	beans		sandy loam, loe	636	9,6	wb, sg,
2	68		100	winter wheat		sugar beets	rotation fallow		sandy loam, loe	636	9,6	cultivation
3	68		97	winter wheat		sugar beets	fallow land (5,5y)		sandy loam, loe	636	9,6	1993-1996
4	136		95	winter wheat		oats	sugar beets		sandy loam, loe	636	9,6	
5	136		96	winter wheat		potatos	sugar beets		sandy loam, loe	636	9,5	cultivation
6	136		107	winter wheat		sugar beets	maize		sandy loam, loe	636	9,5	1991-1994
7	136		107	winter wheat		sugar beetsn	summer wheat	maize	sandy loam, loe	636	9,5	
8	136		82	winter wheat		oats	sugar beets	sugar beets	sandy loam, loe	636	9,5	organic
9	136		77	winter wheat		potatos	sugar beets		sandy loam, loe	636	9,5	organic
10	136		85	winter wheat		sugar beets	maize	maize	sandy loam, loe	636	9,5	organic
11	136		84	winter wheat		sugar beets	summer wheat	sugar beets	sandy loam, loe	636	9,5	organic
12	57	371	98	winter wheat	Sperber	sugar beets	winter barley	winter wheat	sandy loam, loe	635		wb, ww
13	57	365	98	winter wheat	Sperber	potatos	sugar beets	summer barl	sandy loam, loe	635		cultivation, weed
14	57	365	105	winter wheat	Sperber	sugar beets	maize	maize	sandy loam, loe	635		1987-1992
15	57	365	97	winter wheat	Sperber	sugar beets	winter wheat	sugar beets	sandy loam, loe	635		
16	39	433	90	winter wheat	Okapi	summer barley			sandy loam, loe	690	8,5	oats, cultivation, weed
17	39	433	100	winter wheat	Okapi	oats			clay, silt	690	8,5	1982-1986
18	39	433	97	winter wheat	Okapi	winter wheat			clay, silt	690	8,5	



	2019			delta		
	Profit	Quantity	Sales	Quantity	Sales	Profit
Overall	128.9k	13.3k	1.0m	36.5%	36.2%	30.9%
France	35.1k	3.9k	308.4k	33.8%	33.2%	8.7%
Austria	7.5k	299.0	24.6k	13.3%	11.9%	40.8
Belgium	4.2k	202.0	17.3k	23.2%	82.7%	98.6%
Denmark	-1.3k	86.0	2.8k	4.9%	-21.1%	22.9%
Zealand	-245.0	15.0	242.7			
South Denmark	-362.9	40.0	1.3k	33.3%	-7.7%	41.1
Sonderborg	-45.4	6.0	87.5			
Odense	-280.3	30.0	1.1k	0.0%	-20.1%	54.5
Esbjerg	-37.3	4.0	88.7			
Hovedstaden	-0.7k	31.0	1.3k	-40.4%	-40.9%	35.4%
Copenhagen	-0.7k	31.0	1.3k	0.0%	-29.5%	23.
Frederiksberg	232.1	11.0	1.1k			
Finland	36.0k	2.4k	216.5k	-79.6%	-81.4%	-86.2%
Germany	-3.9k	152.0	7.2k	15.0%	46.3%	32.9%
Ireland	10.6k	1.4k	109.7k	17.8%	101.6%	-152.5%
Italy	-11.6k	0.6k	23.7k	37.5%	37.2%	57.6%
Netherlands	2.9k	135.0	12.9k	46.3%	12.6%	2.5%
Norway	-1.0k	74.0	1.8k	229.3%	230.8%	162.8%
Portugal	20.6k	1.2k	99.1k	-29.5%	-79.2%	82.4%
Spain	-9.4k	360.0	15.6k	46.7%	12.9%	27.1%
Sweden	2.2k	84.0	7.3k	143.2%	167.7%	-128.7%
Switzerland	36.8k	2.4k	194.0k	5.0%	-25.6%	-29.1%
United Kingdom	1.3k	48.0	4.0k	66.4%	56.4%	32.7%
Wales	1.9k	118.0	6.6k	-5.9%	-30.2%	-28.7%
Scotland	33.6k	2.2k	183.4k	0.9%	-54.6	-55.5

Goal TRL: map tables to some consistent input. Learn some representation that helps detect patterns relevant to given task(s).



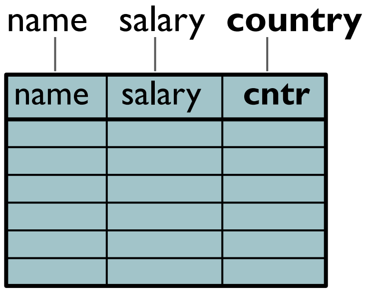
Outline for today

1. Neural Models for Table Understanding
2. Resources for Table Representation Learning
3. Retrieval systems for structured data

Neural Models for **Table Understanding**

Column type detection: **why?**

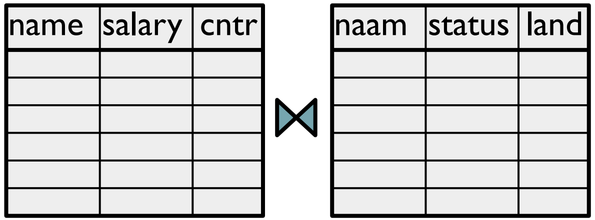
Essential understanding of a table comes through its columns.



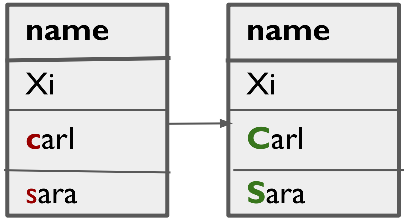
Looks easy, but….

- Undescriptive header?
- Messy and heterogeneous values?
- Unknown types?

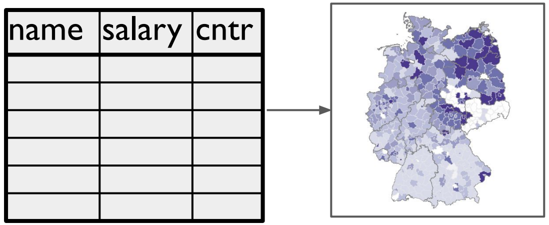
Semantic column types dictate operations to perform on them:



Join tables on “name” and “country” columns



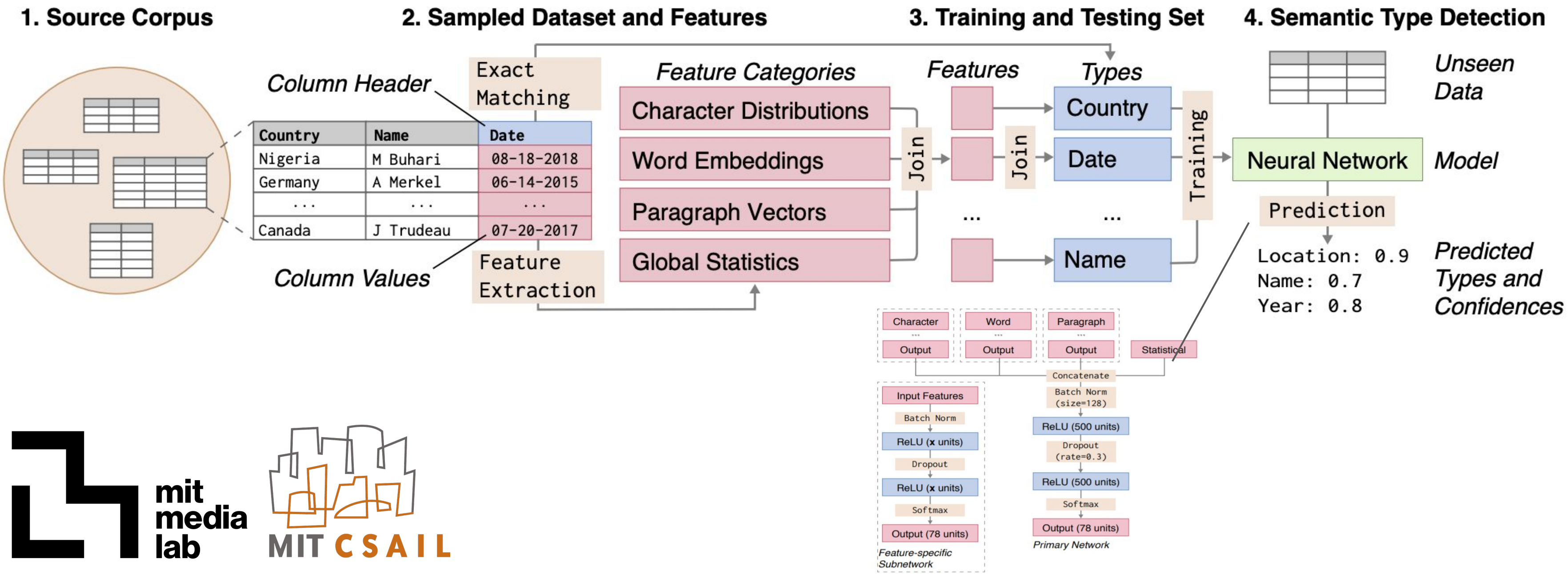
Capitalize “name” columns



Plot “country” data

Sherlock: Column Type Detection with DL

Prior: string matching (col name/values) w/ regex or dict: **robust? scale? accuracy?**



mit media lab

MIT CSAIL

Sherlock: A Deep Learning Approach to Semantic Data Type Detection. Hulsebos, Hu, Bakker, et al., KDD 2019.

How well does **Sherlock** detect types?

78 semantic types (name, address, etc).

Method	F_1 Score	Runtime (s)	Size (Mb)
Machine Learning			
Sherlock	0.89	0.42 (±0.01)	6.2
Decision tree	0.76	0.26 (±0.01)	59.1
Random forest	0.84	0.26 (±0.01)	760.4
Matching-based			
Dictionary	0.16	0.01 (±0.03)	0.5
Regular expression	0.04	0.01 (±0.03)	0.01
Crowdsourced Annotations			
Consensus	0.32 (±0.02)	33.74 (±0.86)	–

Examples of misclassifications.

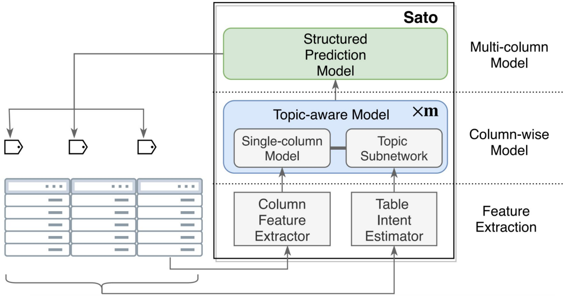
Examples	True type	Predicted type
Low Precision		
81, 13, 3, 1	Rank	Sales
316, 481, 426, 1, 223	Plays	Sales
$, $$, $$$, $$$$, $$$$$	Symbol	Sales
Low Recall		
#1, #2, #3, #4, #5, #6	Ranking	Rank
3, 6, 21, 34, 29, 36, 54	Ranking	Plays
1st, 2nd, 3rd, 4th, 5th	Ranking	Position

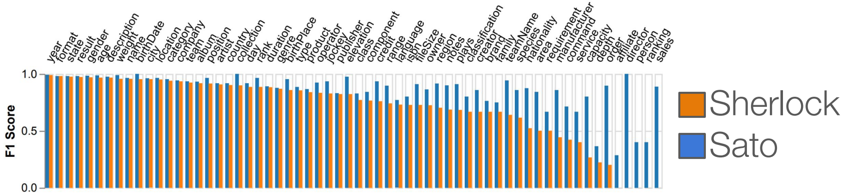
Challenges:
- Numeric data
- Non-mutually exclusive types

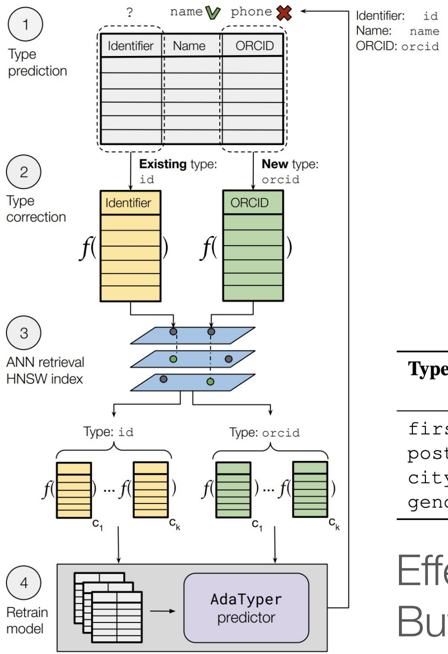
Deployed (often in healthcare), benchmarked against, and extended…

Extending Sherlock: **Contextual** and **Adaptive** Models

SATO: condition Sherlock-predicted column type on preds of neighbor columns







AdaTyper: adapt base type detection model by generating labeled samples from example columns

Type	regex	dictionary	AdaTyper		
			i=0	i=5	ΔF1
first name	0	0.508	-	**0.580**	+0.580
postal code	**0.108**	0	-	0.068	+0.068
city	0.195	0	0.296	**0.387**	+0.091
gender	0	0.098	0.340	**0.341**	+0.001

Effective w/ only few examples
But numeric/pattern? Regex..

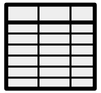
Sato: Contextual Semantic Type Detection in Tables. Zhang, Suhara, Li, Hulsebos et al, VLDB, 2020.
AdaTyper: Adaptive Semantic Column Type Detection. Hulsebos et al, arxiv, 2023.

Then came **Transformers for Tables**

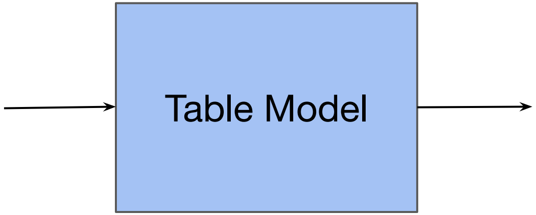
Transformers for Tables

INPUT

table



question
“...?”



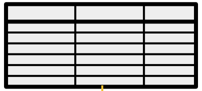
OUTPUT

embeddings
col-level: `[[1,24],[6,74],[9,10]]`
row-level: `[1,24,955,101]`

tuned (QA)
`Yes`
`5 January 2022`

Transformers for Tables

Input



“abc dz”?

[col : col1|…|colN row1 : val1|val2|…|valN row2 : …]

[abc dz?]

[token1, token2, token3, …, tokenM]

[1, 2, 3, 4, 5, 6, …, M]

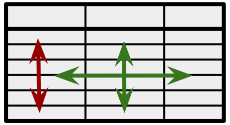
Serialize (e.g. row-wise or col-wise)

Merge tokens table & context (e.g queries)

Map tokens to “token IDs”

Model

Structured attention:



Vertical

Matrix

Pre-training tasks:

Default: recover column names or cell values.

Efficient: synthetized SQL execution.

Output

Embeddings:

token-level agg to cell-, col-, row-level.

Fine-tuning:

Predicting cells+operators, SQL, etc

Transformers Not Always SOTA…

Problem: pretrained table models poor OOD performance (e.g. col type prediction).

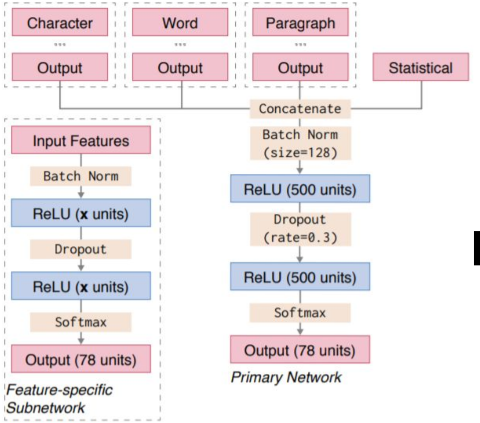
Just use GPT for col type prediction? [1]

	F_1-score	Precision	Recall	
DoDuo-VizNet∗	0.900	90.3%	89.9%	
Sherlock∗	**0.930**	**92.2%**	**93.1%**	← Specific DL model
$\boxed{\text{TaBERT} \quad 0.380 \quad 38.9\% \quad 38.3\%}$				
DoDuo-Wiki	0.815	82.6%	81.4%	
Chorus	**0.865**	**90.1%**	**86.7%**	← GPT-based

[1] CHORUS: Foundation Models for Unified Data Discovery and Exploration. Kayali, et al. VLDB, 2024.

LLMs for tables: Overkill or Unutilized Potential?

Sherlock col type pred model:

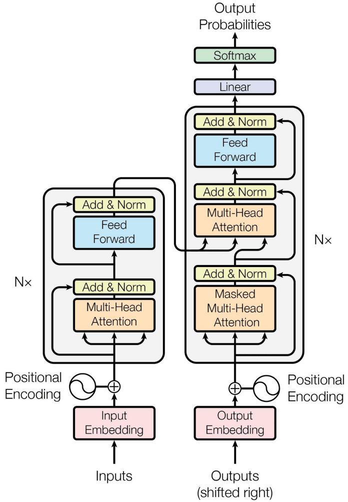


hundreds of params

vs

billions/trillions of params

Transformer architecture:



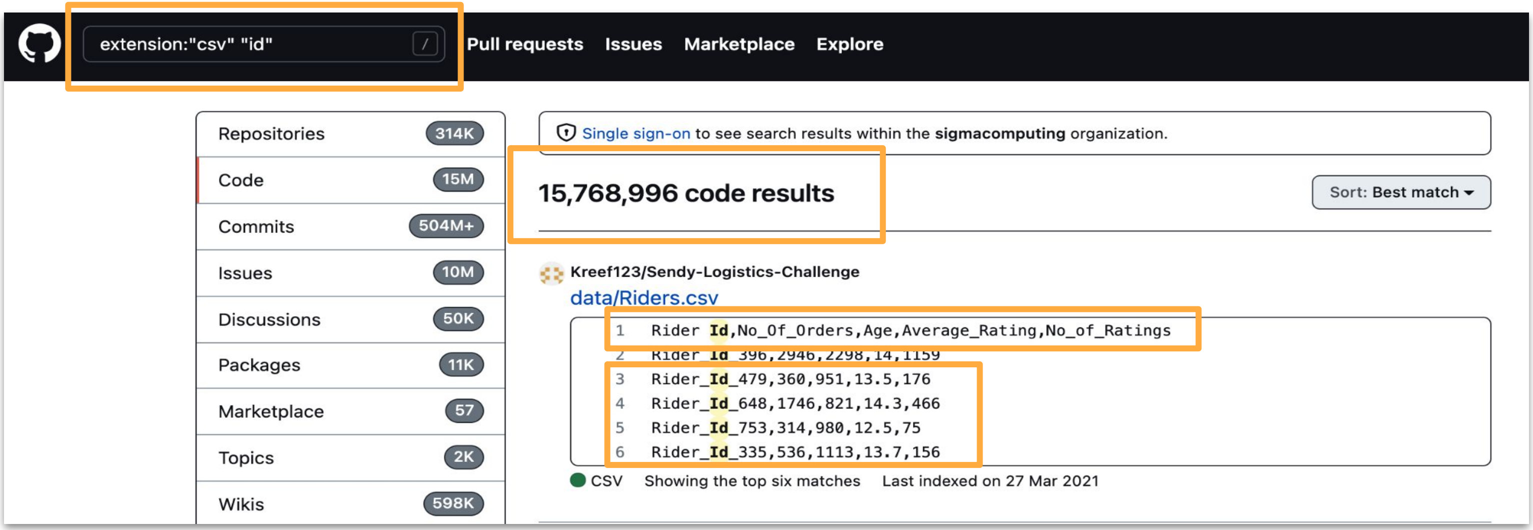
***But* LLMs are promising**… if they'd work well for tables:

How to handle messy data, large tables, full DBs, vague headers, numeric data?

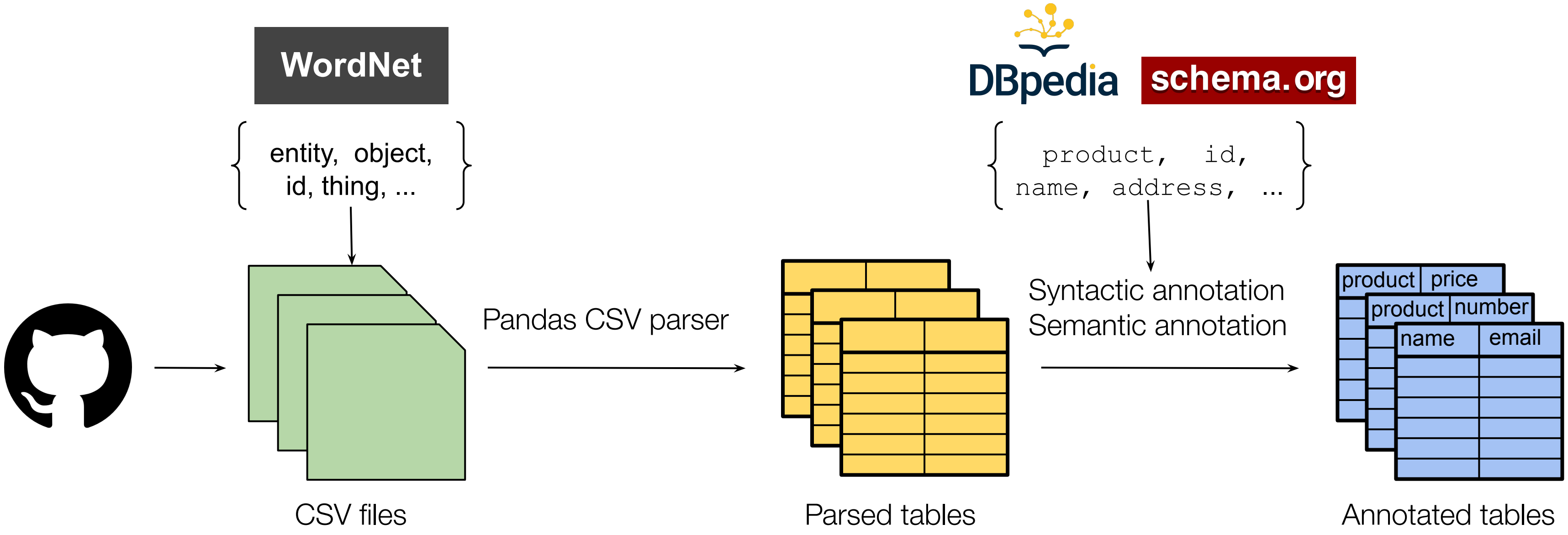
Resources for **TRL**

What Data Do We Need?

- Web/WikiTables → Web applications. Web tables ≠ DB tables…
- Data tasks on offline tables? GitHub as a data source?



GitTables



UNIVERSITY OF AMSTERDAM sigma

GitTables: A large-scale corpus of relational tables. Hulsebos, Demiralp, & Groth. SIGMOD, 2023.

Properties and Use-cases of GitTables

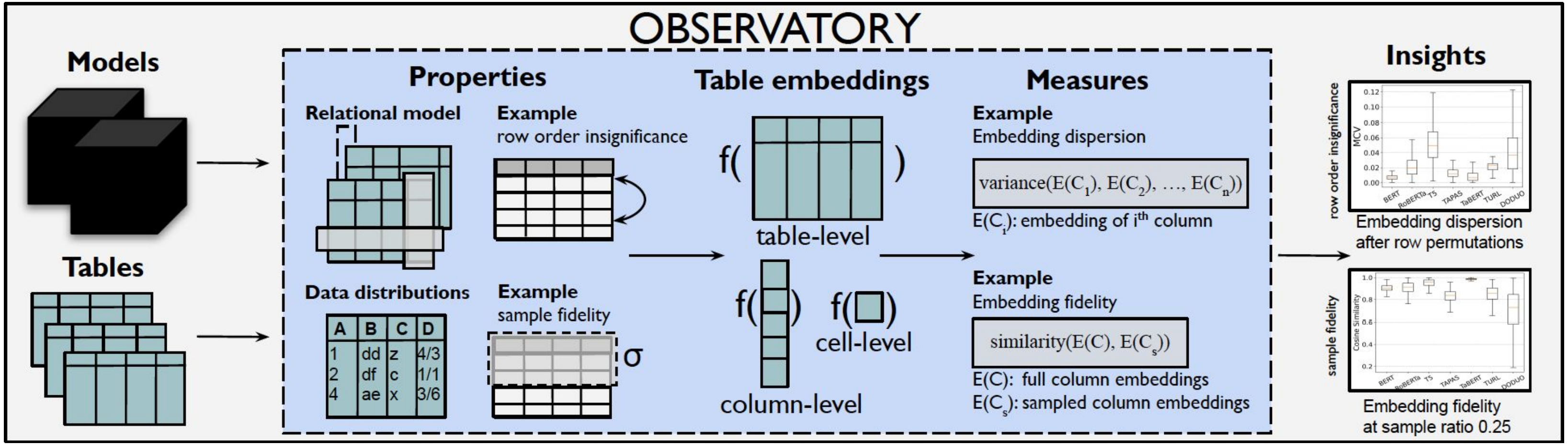
- **>1M tables** and **800K CSV** files.
- More representative: wider+taller, and IDs most common attribute.
- Usage shown for semantic **column type detection** and **schema completion**:

Header prefix	**Suggested completion**
payment_id, customer_id	**review_id**, **product_id**, **product_parent**, **product_title**, ...
id, company	**ReceivablePaymentHeader**, **ReceivablePayment**, **Status**, **Customer**, **BankEntity**, ...
id, name, location	**phone**, **email**, **uid**, **active**, **ad_organization_id**, ...

Also used for join discovery, CSV parsing, KG enhancement, etc.

Do Current **Models Capture Relational Properties**?

Neural table embeddings through the lens of Codd's relational model.



UNIVERSITY OF AMSTERDAM
UNIVERSITY OF MICHIGAN

Observatory: Characterizing Relational Table Embeddings. Cong, Hulsebos, Sun, Groth, Jagadish, VLDB, 2024.

Example Property: Functional Dependencies

Given table T with FD: X=country → Y=continent

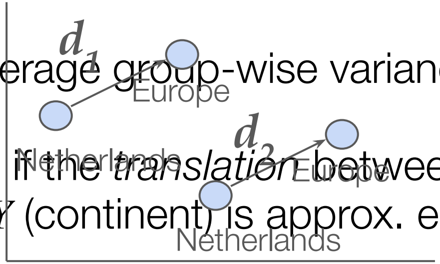
ID	name	country	continent
1	Kathryn	Netherlands	Europe
2	Oscar	Netherlands	Europe
3	Lee	Canada	North America
4	Roxanne	USA	North America
5	Fern	Netherlands	Europe
6	Raphael	USA	North America
7	Rob	USA	North America
8	Ismail	Canada	North America

We argue that:

- FD relation interpretable as *translation* between embeddings E(πX (s)) and E(πY (s))
- Model *f* preserves FD if d(E($\pi X(s)$), E($\pi Y(s)$)) = d(E($\pi X(t)$), E($\pi Y(t)$)) where *d* preserves norm+direction (L1/L2-norm).

Measure the average group-wise variance over all n "FD-groups":

$$\overline{S^2} = \frac{1}{n}\sum_{j=1}^{n} \frac{\sum_{i=1}^{m_{\mathcal{G}_j}} ||d_{ji} - \overline{d_j}||_2^2}{m_{\mathcal{G}_j} - 1}$$



$\overline{S^2}$ approaches 0 if the *translation* between group-wise FD value pairs in X (country) and Y (continent) is approx. equal. At least $\overline{S^2}$ is smaller than in non-FD value pairs.

Current Architectures Often Fall Short…

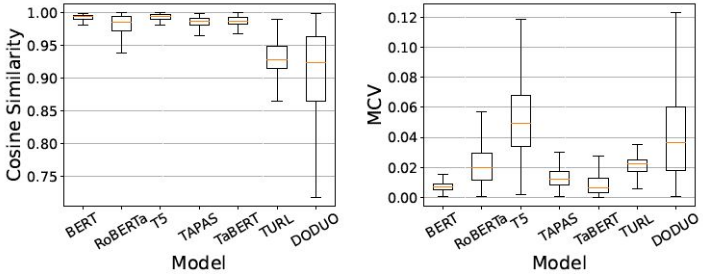
Turns out, most models do not preserve FDs!

Also more fundamental properties:

A *relation* then consists of a set of tuples, each tuple having the same set of attributes. If the domains are all simple, such a relation has a tabular representation with the following properties.

(1) There is no duplication of rows (tuples).
(2) Row order is insignificant.
(3) Column (attribute) order is insignificant.
(4) All table entries are atomic values.

Measure by avg cosine similarity of col embeddings across row permutations.



row order robustness

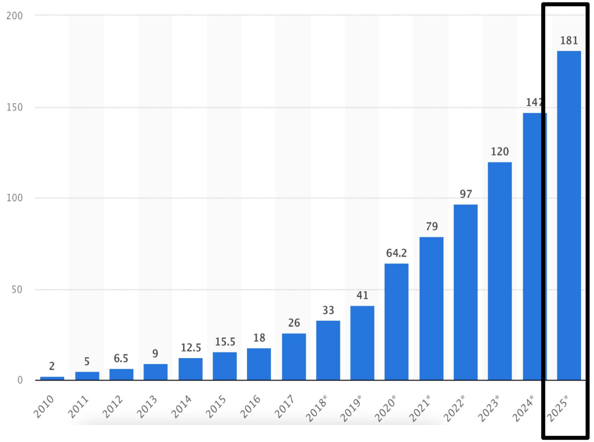
Impact on downstream tasks! **Row shuffling affects 34%** col type predictions.

[2] Extending the Database Relational Model to Capture More Meaning, Codd, 1979.

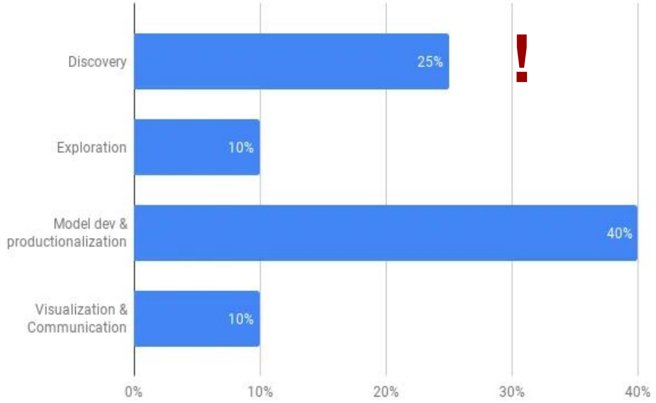
Embeddings -> retrieval systems for structured data

Use-case: dataset search for analytics/ML

Immense growth of data → desire for insights

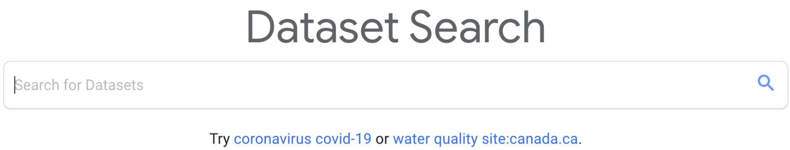


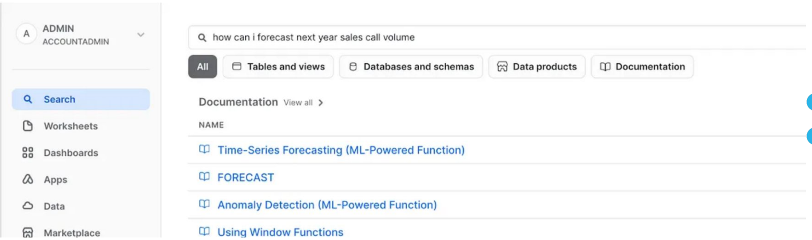
Finding the right dataset = still time-consuming



Research versus Practice

Systems in **industry** vs focus in **research**





"Basic" dataset search (e.g. keyword search)

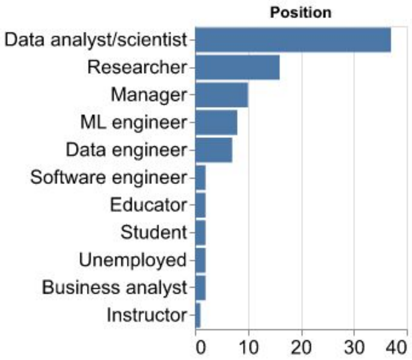
Method	Task	Rep. Learning	ANN Index
Octopus [18]	KS	✗	✗
G.D.S. [2]	KS	✗	✗
Aurum [13]	KS	✗	LSH
LSH-Ensemble [3]	Join	✗	LSH
Juneau [4]	Join	✗	✗
JOSIE [5]	Join	✗	✗
MATE [6]	Join	✗	XASH
DeepJoin [7]	Join	✓	HNSW
D^3L [14]	Union, Join	▮	LSH
Starmie [8]	Union, Join	✓	LSH, HNSW
TUS [9]	Union	▮	LSH
SANTOS [10]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

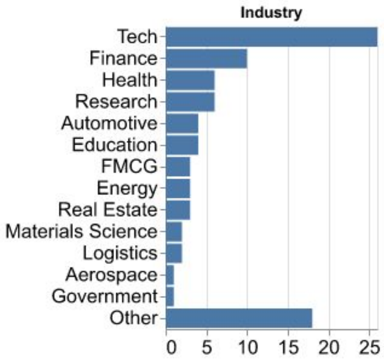
Majority research on data augmentation

Basic Dataset Search is not a Solved Problem!

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

89 data practitioners!! recruited through social media & mailing lists:





We asked:

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

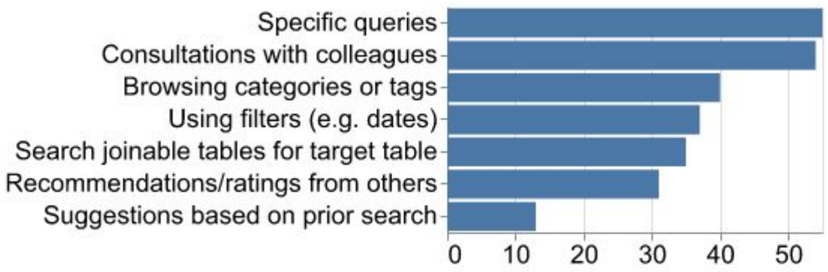
UNIVERSITY OF CALIFORNIA

“It Took Longer than I was Expecting:” Why is Dataset Search Still so Hard?. Hulsebos, et al, HILDA@SIGMOD, 2024.

Practitioner's perspective: what and how they search

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

How do you search?

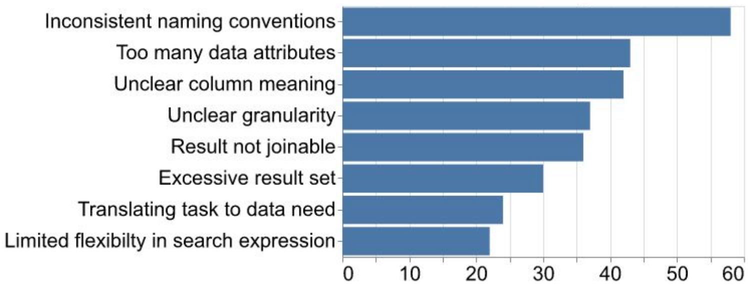


"*Identify the **problem and the data for the problem**, … then specific keyword or tag search. Also, **identify people** who have worked on similar problems…*"

"*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.*"

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…."*

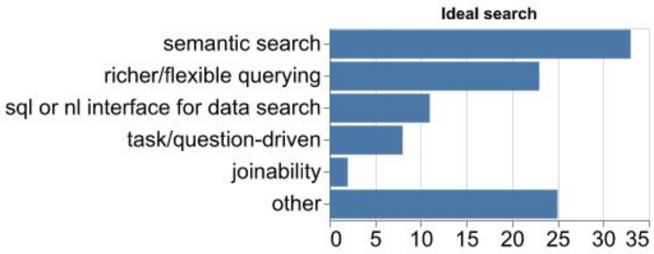
*"****Categorical level of detailing*** *is required, which isn't possible now."*

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

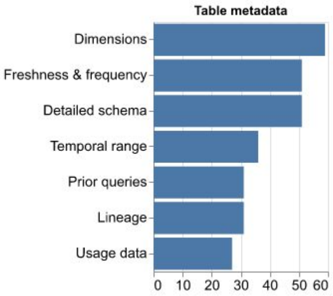
"Not many features to search/query keywords, a lot of times ***changing query still renders same data results****…"*

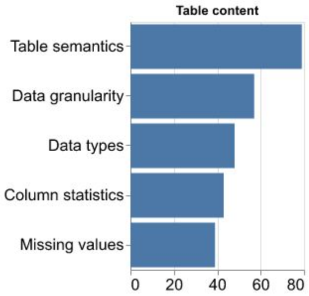
Practitioner's perspective: ideal search systems

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**>"

Desiderata for Dataset Search

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

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

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

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

TBC…

Key takeaways…

- **Tables prevalent** in the data landscape, especially enterprises (*eg* for analytics).
- Capabilities of **transformer should extend beyond** images & text -> tables & DBs.
- Pre-train & tune table models on **representative data** (*eg* GitTables).
- **Tables ≠ natural language**: specific challenges and properties (Observatory).
- Important applications of table embeddings, e.g. **retrieval systems**.

Interested?

- Reach out: madelon@berkeley.edu
- TRL papers/resources: madelonhulsebos.com/trl
- Table Representation Learning workshop @ NeurIPS 2024??

Thanks!