Towards **Table** Representation Learning for End-to-End Data Management and Analysis

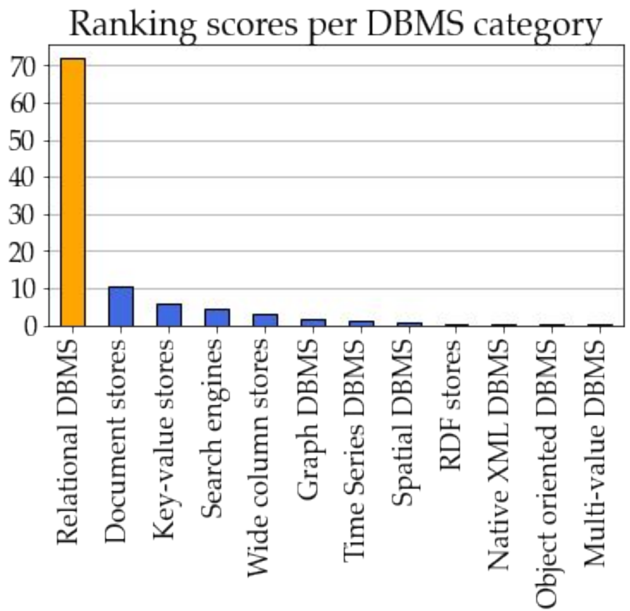
Hasso Plattner Institute, Potsdam
06/03/2023

Madelon Hulsebos

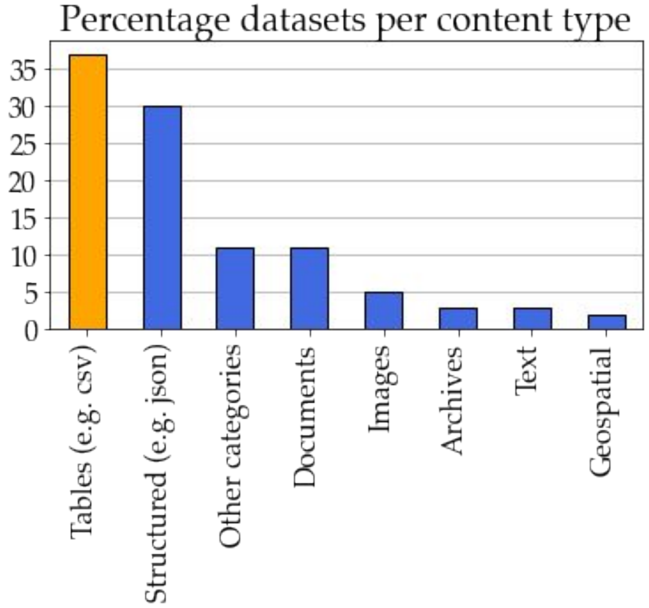
INDE lab

Tables are **everywhere**

Databases, (web) pages, documents, spreadsheets…



From DB-Engines Ranking by Category (Jan ‘23)

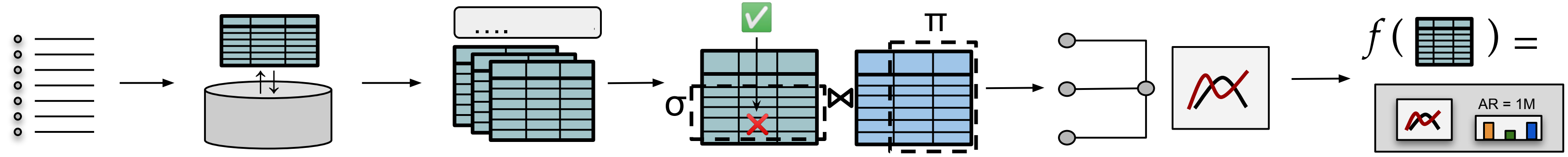


From Google Dataset Search by the Numbers (Benjelloun et al., ‘20).

Tables are **driving many analysis pipelines**

End-to-end pipelines involve tons of applications.

EXTRACT	STORE	RETRIEVE	PROCESS	ANALYSE	USE



EXTRACT
- parsing
- integration
- validation

STORE
- storage
- query processing
- cataloging

RETRIEVE
- exploration
- querying
- monitoring

PROCESS
- transformation
- augmentation
- cleaning

ANALYSE
- aggregation
- statistics
- visualization

USE
- insight extraction
- machine learning
- dashboarding

As w/ images and text: can we learn table representations to fuel these pipelines?

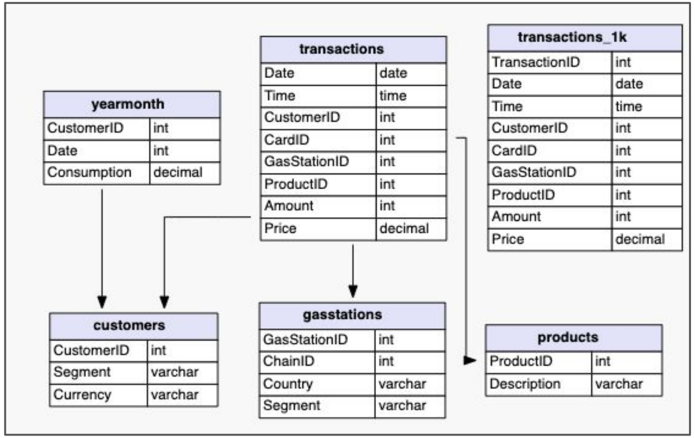
Tables are **rich and challenging**

Content: measurements, messy, heterogeneous dtypes.

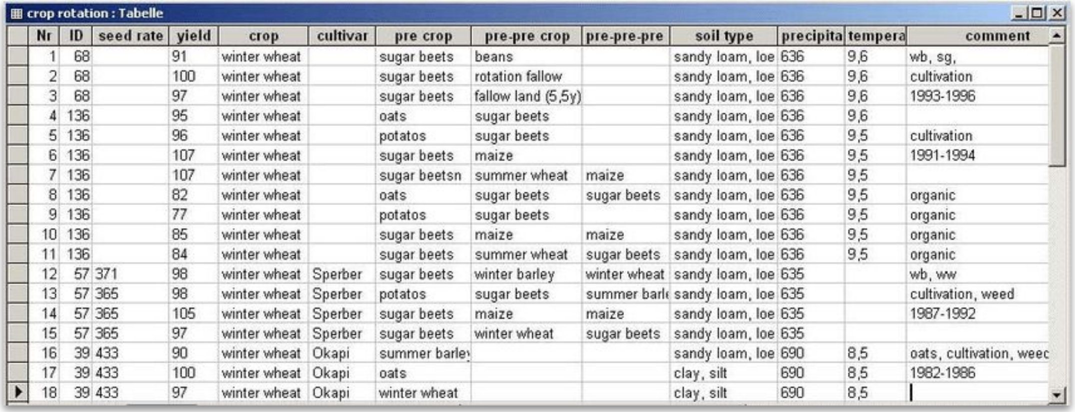
Structure: columns, rows, cells, headers, hierarchical.

Context: relations, constraints, metadata.

Usage: analyses, ml models, visualizations.



CTU Prague Relational Learning Repository



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 barle			sandy loam, loe	690	8,5	oats, cultivation, weec
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	

First table when searching "crop data"

	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%	23.1
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	68.4%	66.4%	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%

Published by a Tableau user

Today: **learning over tables**

How to **represent** a table?

How to **understand** them?

How to **adapt** table models?

How to find good **data**?

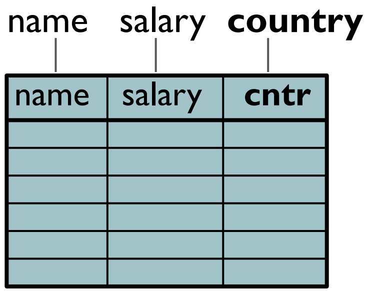
And beyond…

Images, videos, text...

Tables

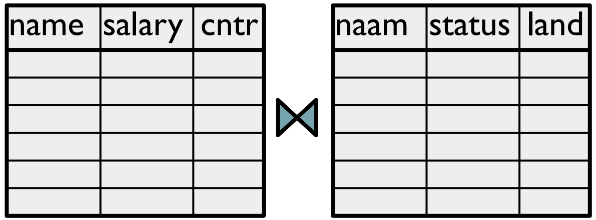
Column type detection: **why?**

Essential **understanding** of a table comes **through its columns**.

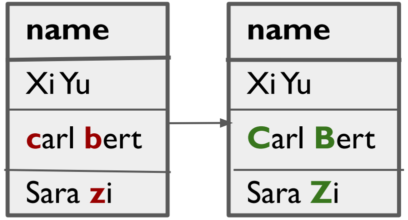


Looks easy, but….
- Undescriptive header?
- Messy and heterogeneous values?
- Unknown types?

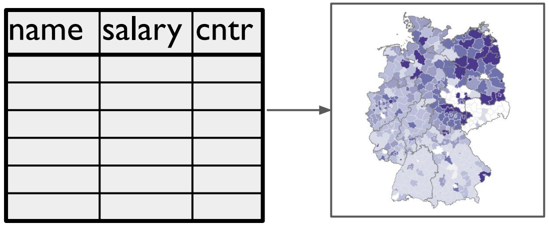
As in other type systems, semantic column types dictate operations to perform on them.



Join tables on “name” and “country” columns



Capitalize “name” columns



Plot “country” data

Column type detection: **how?**

Matching header or values by ① matching column values, ② aggregating to types.

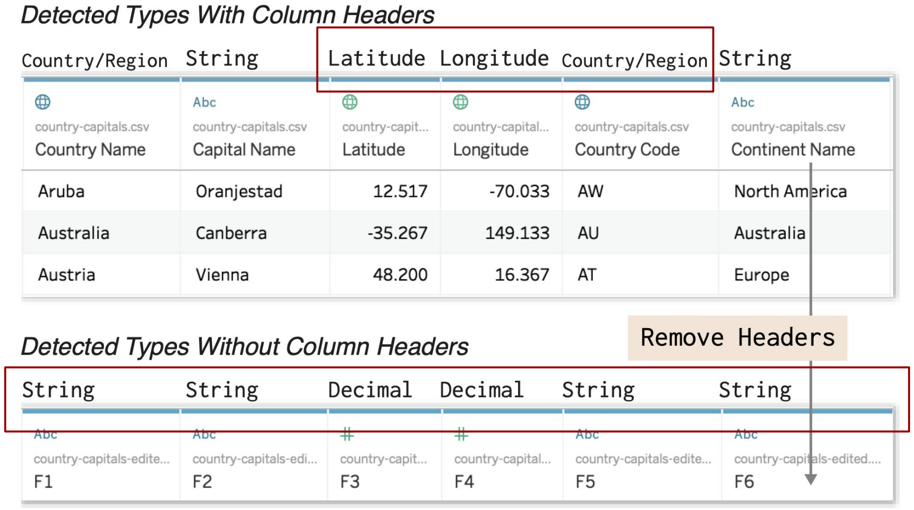
In commercial systems (e.g. Tableau):

- Preset regular expressions.
- Preset type:values dictionary.

SOTA:

- Ontology-based [1].
- Extracted rules from GitHub [2].

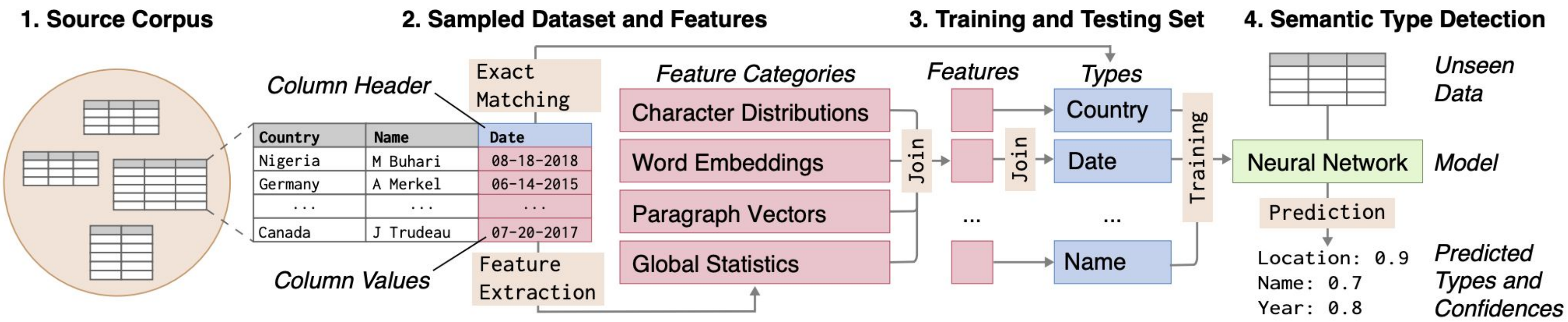
What if we remove column names?

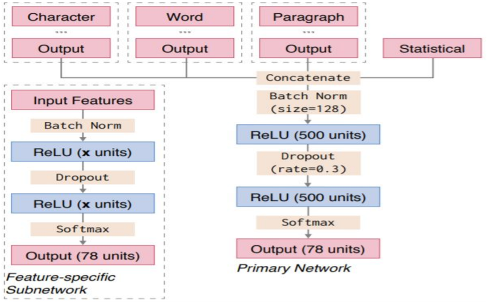


[1] Recovering semantics of tables on the web. Petros et al, 2011
[2] Synthesizing type-detection logic for rich semantic data types using open-source code, Cong Yan and Yeye He, 2018.

Column type detection: **Sherlock**

Scale, robustness, accuracy?





mit media lab

MIT CSAIL

Published at KDD 2019

Can **Sherlock** detect types?

Evaluated on >600K columns from Web tables.

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)	–

Current usage:

- Adopted in industry: health tech and fashion (e.g. data integration).
- People contributed bugfixes, speedups.
- Was extended to SATO (w context).
- Research benchmarks (competitive!).

Paper, model, data and code: https://sherlock.media.mit.edu

In the wake of **Sherlock**

Pre-trained models for table understanding: large-scale training without ground-truth labels.

Industry feedback Sherlock: nice but **data mismatch**, cannot add **custom types.**

① How to transfer to new data domains?

② How to detect new types?

What **data** do we need?

① How to transfer to new data domains? → Why asked?

Tables needed:

- Large to facilitate learning → WebTables [3] ✅
- Table semantics (e.g. col types) → WebTables ✅
- DB-like table content and structure (semantics, dtypes, size) ❓
- Coverage to generalize across domains ❓

President	Party	Term as President	Vice-President
1. George Washington (1732-1799)	None, Federalist	1789-1797	John Adams
2. John Adams (1735-1826)	Federalist	1797-1801	Thomas Jefferson
3. Thomas Jefferson (1743-1826)	Democratic-Republican	1801-1809	Aaron Burr, George Clinton
4. James Madison (1751-1836)	Democratic-Republican	1809-1817	George Clinton, Elbridge Gerry
5. James Monroe (1758-1831)	Democratic-Republican	1817-1825	Daniel Tompkins
6. John Quincy Adams (1767-1848)	Democratic-Republican	1825-1829	John Calhoun
7. Andrew Jackson (1767-1845)	Democrat	1829-1837	John Calhoun, Martin van Buren
8. Martin van Buren (1782-1862)	Democrat	1837-1841	Richard Johnson
9. William H. Harrison (1773-1841)	Whig	1841	John Tyler
10. John Tyler (1790-1862)	Whig	1841-1845	
11. James K. Polk (1795-1849)	Democrat	1845-1849	George Dallas
12. Zachary Taylor (1784-1850)	Whig	1849-1850	Millard Fillmore
13. Millard Fillmore (1800-1874)	Whig	1850-1853	
14. Franklin Pierce (1804-1869)	Democrat	1853-1857	William King
15. James Buchanan (1791-1868)	Democrat	1857-1861	John Breckinridge

Table from a Web page.

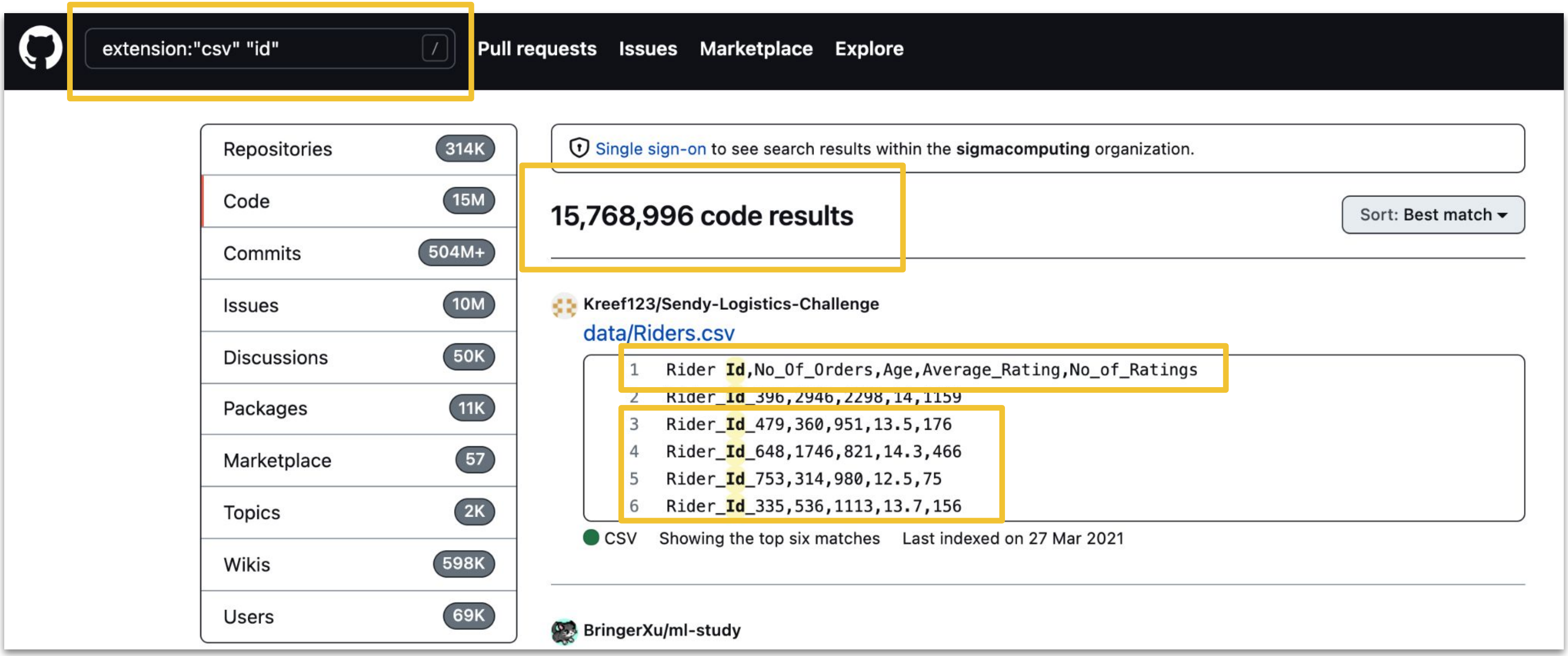
crop rotation : Tabelle

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, weec
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	

Table with crop data, first result "example database table".

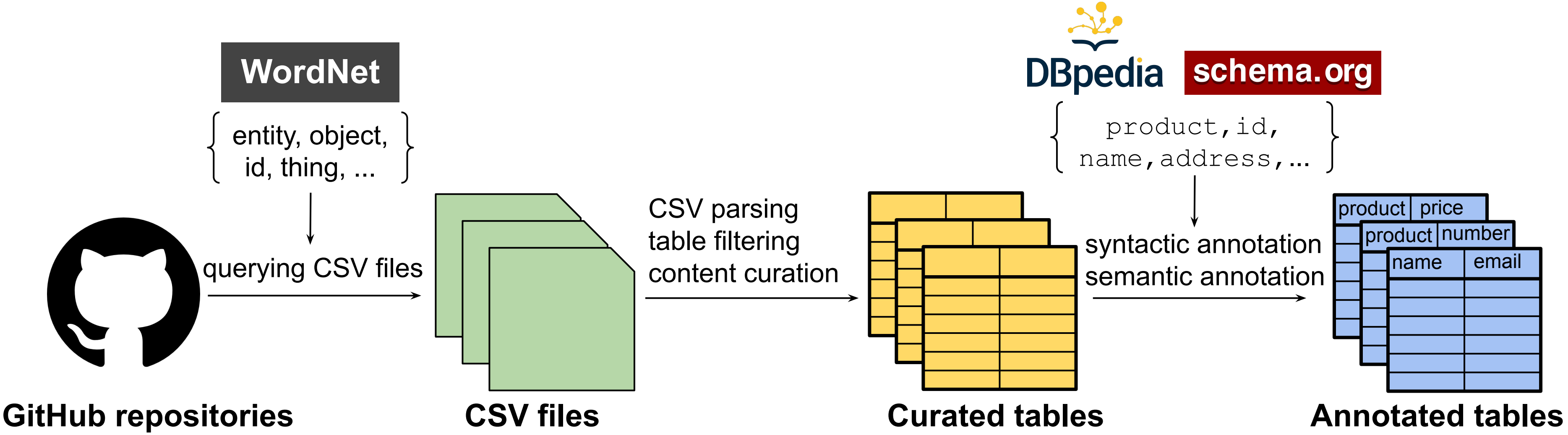
[3] WebTables: exploring the power of tables on the web, Cafarella et al., 2008

Can we use GitHub CSV files?



Result from GitHub code search when querying for CSV files containing "id".

The birth of **GitTables**



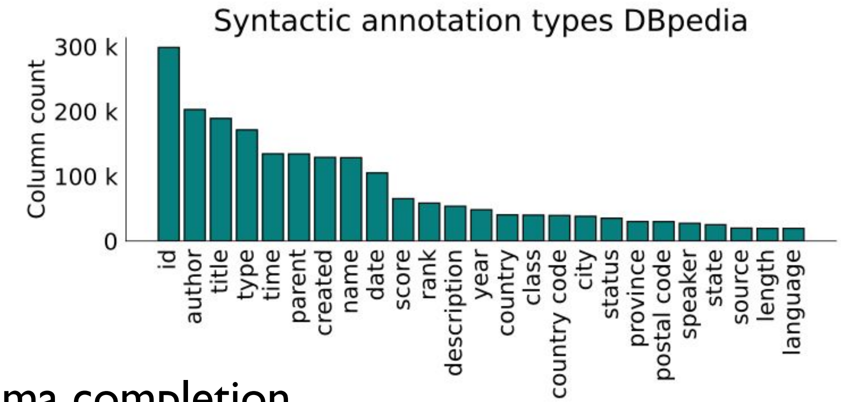
INDE lab

Published at SIGMOD 2023

What can we do with **GitTables**

We publish >1M tables, also underlying >800K CSV files.

Data type	GitTables	WDC WebTables
Numeric	57.9%	51.4%
String	41.6%	47.4%
Other	0.5%	1.2%



We show: ML for type detection and schema completion.

Other use: join discovery, schema matching, benchmarking.

General Table Representations? E.g. parsing, compression, error repair?

Paper, data and code: https://gittables.github.io

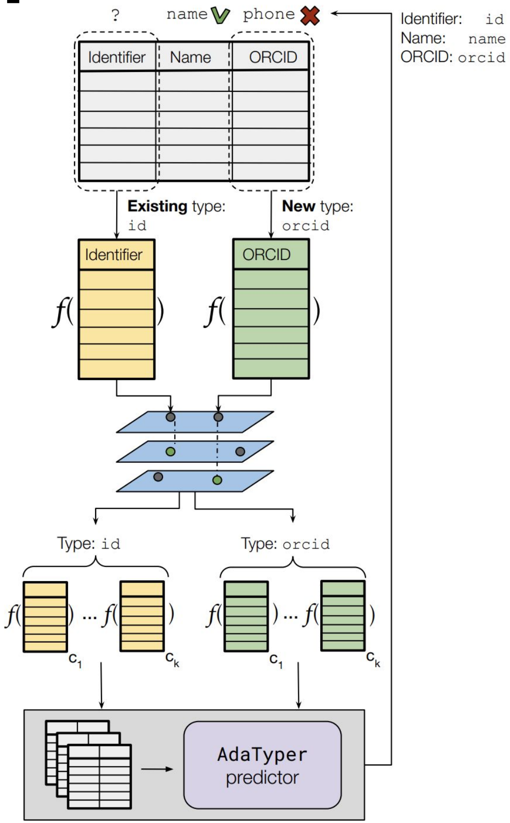
Adaptive type detection: **AdaTyper** [WIP]

② How to detect new types?

Current: by user-provided value dict or regular expression.

Interactive adaptation by example:

1. Predict initial column type.
2. User corrects with (new) type.
3. Embed example column.
4. Retrieve similar col embeddings from HNSW index [4].
5. Retrain type prediction model.

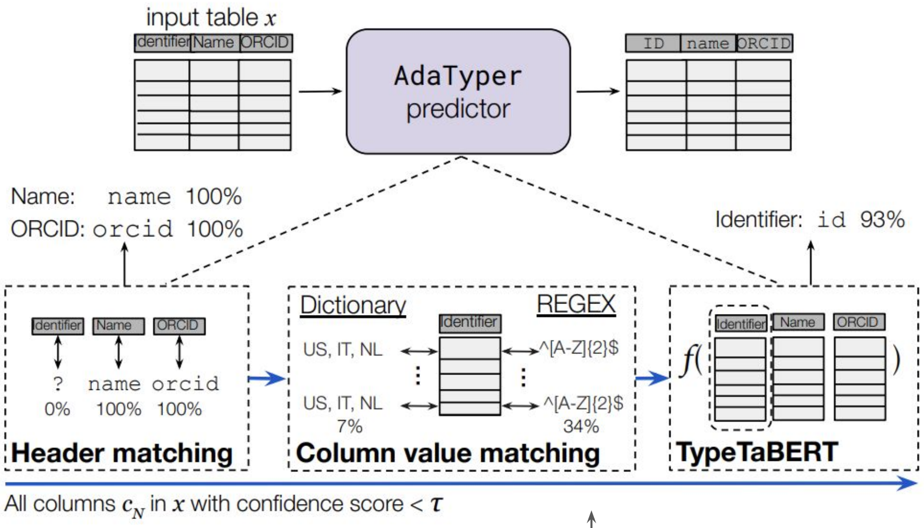


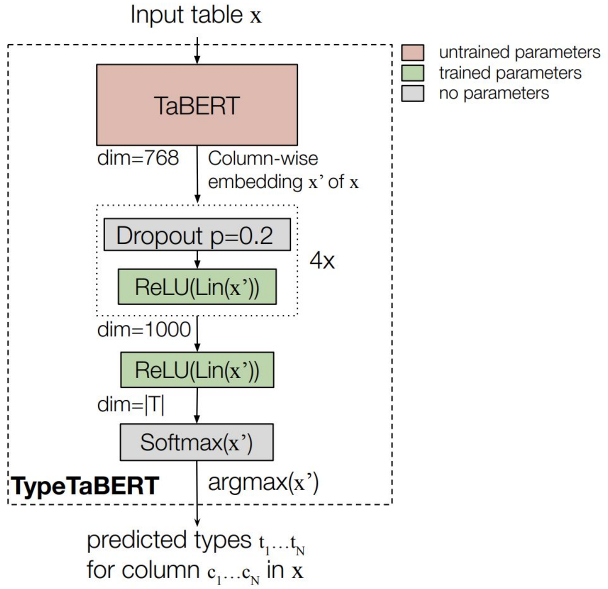
INDE lab

[4] Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs, Malkov and Yashunin, 2018

AdaTyper predictor

Hybrid type detection pipeline enabling **different adaptation methods**.





So, we can still adapt through regular expressions….

How well does **AdaTyper** adapt?

Measuring **performance after *x* examples** of new type

- Human annotated tables from Prague Relational Learning Repo (not used for training!).
- High precision.
- WIP: low recall, increase -> drop: issues w example diversity and label errors?

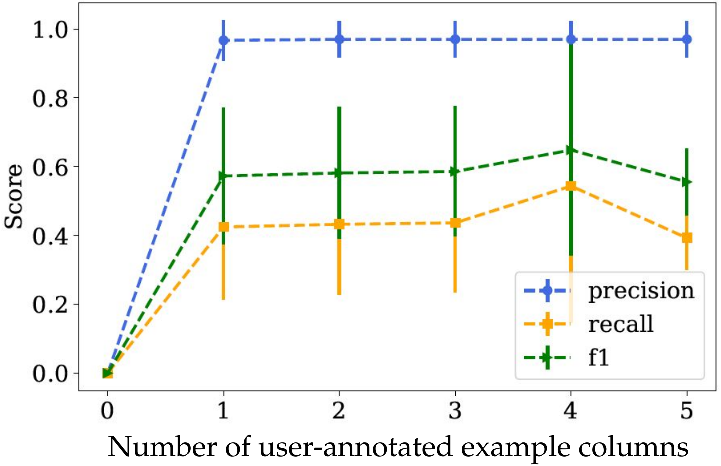


Table on the **Horizon**

1. What *do* table representations capture? Now blindly adopting models for any task.

2. What *can* table representations capture within E2E pipeline?

    - Left and right, from storage & query optimization to analysis recommendation
    - Developing new neural architectures aligned with data management tasks
    - Contextualizing tables w.r.t. downstream usage

3. Table-specific deployment challenges.

Interested?

New research area with many challenging problems and impactful applications!

Exciting community spanning different communities (e.g. NLP, DB, ML). Take part:

1. **Join**: Dedicated TRL Slack space → reach out m.hulsebos@uva.nl!
2. **Learn**: SIGMOD '23 Tutorial "Models and Practice of Neural Table Representations".
3. **Contribute**: hopefully 2nd Table Representation Learning workshop at NeurIPS '23.

Ideas for TRL applications, challenges, questions → m.hulsebos@uva.nl?