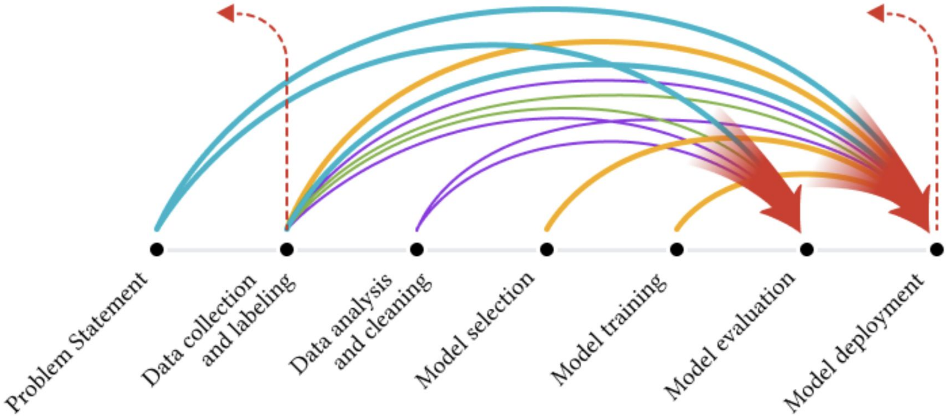
# What are we asking from **Tabular Data**?

**Madelon Hulsebos**
6 December, Copenhagen
AI for Tabular Data workshop @ EurIPS 2025

As a data scientist,
in the “real world”,
I realized 3 things…

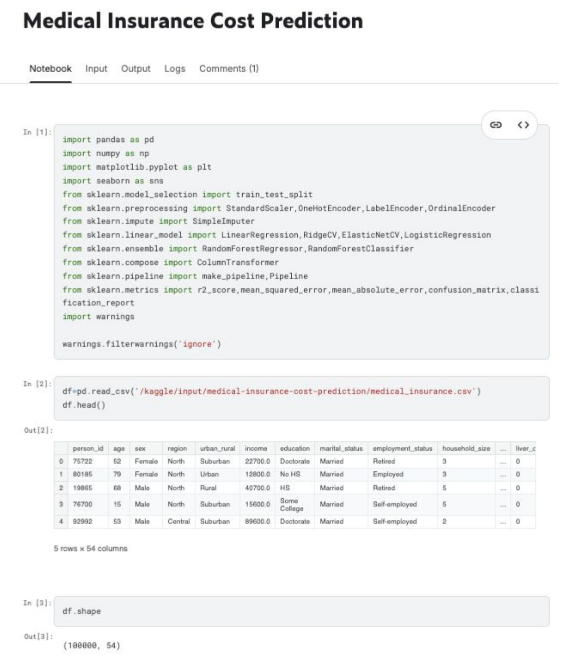
# **Realization 1:** most of my work was data work



Data science = "80% data work, 20% model work"

"Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI, Sambasivan et al., 2021

# **Realization 2**: everyone is doing the same….



Medical Insurance Cost Prediction

```
In [1]:
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler,OneHotEncoder,LabelEncoder,OrdinalEncoder
from sklearn.impute import SimpleImputer
from sklearn.linear_model import LinearRegression,RidgeCV,ElasticNetCV,LogisticRegression
from sklearn.ensemble import RandomForestRegressor,RandomForestClassifier
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import make_pipeline,Pipeline
from sklearn.metrics import r2_score,mean_squared_error,mean_absolute_error,confusion_matrix,classi
fication_report
import warnings

warnings.filterwarnings('ignore')
```

```
In [2]:
df=pd.read_csv('/kaggle/input/medical-insurance-cost-prediction/medical_insurance.csv')
df.head()
```

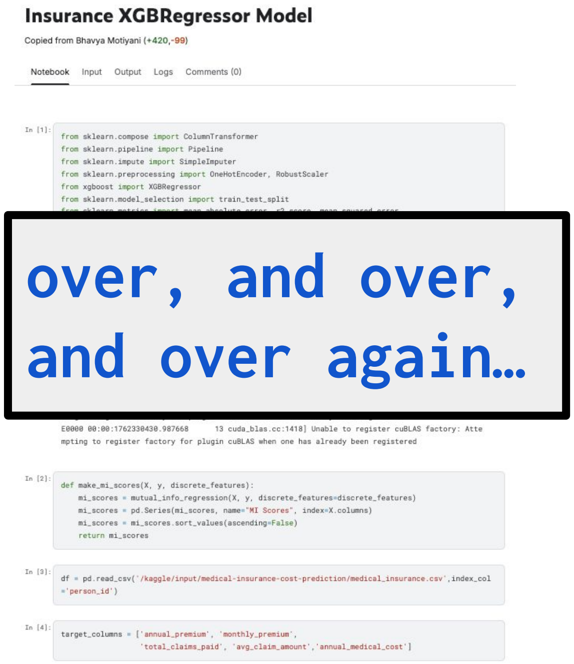
Out[2]:

| | person_id | age | sex | region | urban_rural | income | education | marital_status | employment_status | household_size | ... | liver_c |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 75722 | 52 | Female | North | Suburban | 22700.0 | Doctorate | Married | Retired | 3 | ... | 0 |
| 1 | 80185 | 79 | Female | North | Urban | 12800.0 | No HS | Married | Employed | 3 | ... | 0 |
| 2 | 19865 | 68 | Male | North | Rural | 40700.0 | HS | Married | Retired | 5 | ... | 0 |
| 3 | 76700 | 15 | Male | North | Suburban | 15600.0 | Some College | Married | Self-employed | 5 | ... | 0 |
| 4 | 92992 | 53 | Male | Central | Suburban | 89600.0 | Doctorate | Married | Self-employed | 2 | ... | 0 |

5 rows × 54 columns

```
In [3]:
df.shape
```

```
Out[3]:
(100000, 54)
```



Insurance XGBRegressor Model

Copied from Bhavya Motiyani (+420,-99)

```
In [1]:
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import OneHotEncoder, RobustScaler
from xgboost import XGBRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_absolute_error, r2_score, mean_squared_error
```

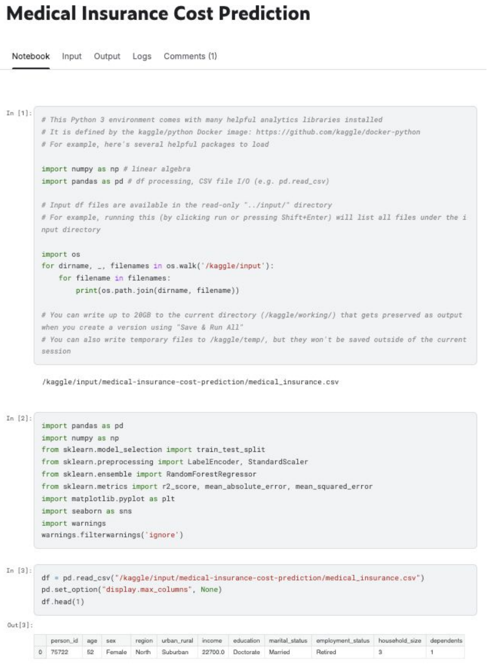
```
E0000 00:00:1762330430.987668      13 cuda_blas.cc:1418] Unable to register cuBLAS factory: Atte
mpting to register factory for plugin cuBLAS when one has already been registered
```

```
In [2]:
def make_mi_scores(X, y, discrete_features):
    mi_scores = mutual_info_regression(X, y, discrete_features=discrete_features)
    mi_scores = pd.Series(mi_scores, name="MI Scores", index=X.columns)
    mi_scores = mi_scores.sort_values(ascending=False)
    return mi_scores
```

```
In [3]:
df = pd.read_csv('/kaggle/input/medical-insurance-cost-prediction/medical_insurance.csv',index_col
='person_id')
```

```
In [4]:
target_columns = ['annual_premium', 'monthly_premium',
                  'total_claims_paid', 'avg_claim_amount','annual_medical_cost']
```

**over, and over,**
**and over again…**



Medical Insurance Cost Prediction

```
In [1]:
# This Python 3 environment comes with many helpful analytics libraries installed
# It is defined by the kaggle/python Docker image: https://github.com/kaggle/docker-python
# For example, here's several helpful packages to load

import numpy as np # linear algebra
import pandas as pd # df processing, CSV file I/O (e.g. pd.read_csv)

# Input df files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list all files under the i
nput directory

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# You can write up to 20GB to the current directory (/kaggle/working/) that gets preserved as output
when you create a version using "Save & Run All"
# You can also write temporary files to /kaggle/temp/, but they won't be saved outside of the current
session
```

```
/kaggle/input/medical-insurance-cost-prediction/medical_insurance.csv
```

```
In [2]:
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import r2_score, mean_absolute_error, mean_squared_error
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

```
In [3]:
df = pd.read_csv("/kaggle/input/medical-insurance-cost-prediction/medical_insurance.csv")
pd.set_option("display.max_columns", None)
df.head(1)
```

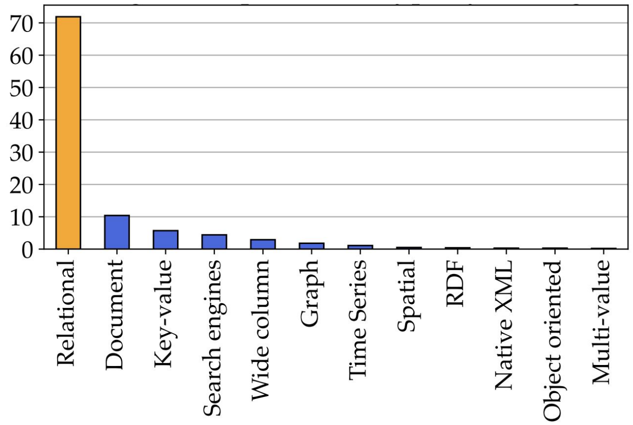
Out[3]:

| | person_id | age | sex | region | urban_rural | income | education | marital_status | employment_status | household_size | dependents |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 75722 | 52 | Female | North | Suburban | 22700.0 | Doctorate | Married | Retired | 3 | 1 |

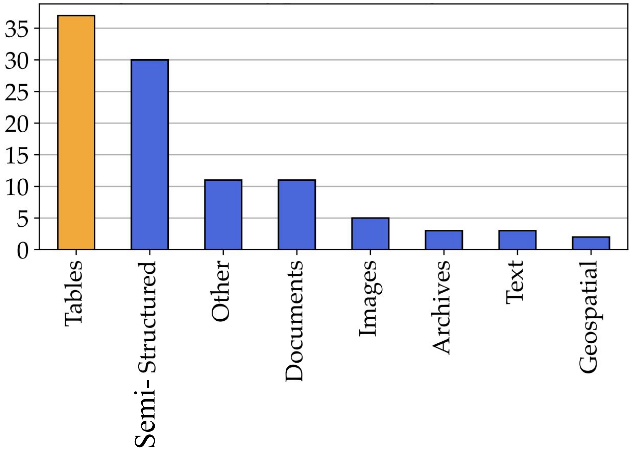
There must be latent patterns to *learn* (data, code, etc).

# **Realization 3:** tables prevail in the org data landscape

Popularity of database systems

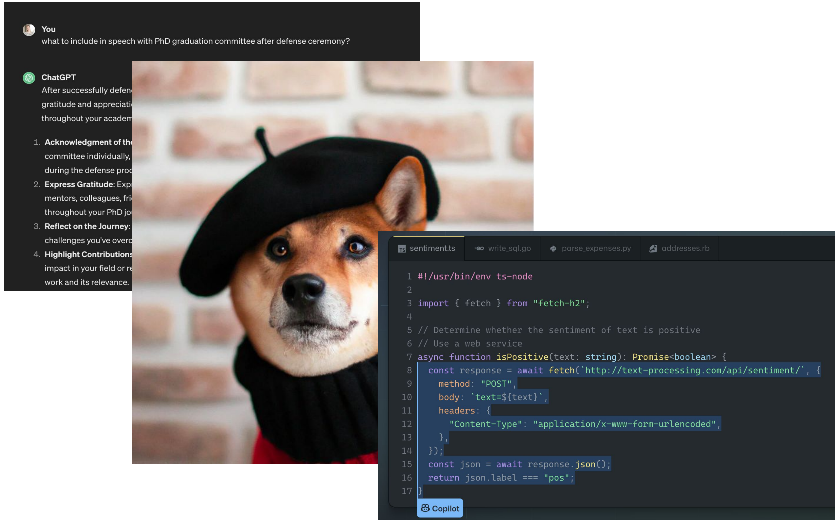


Distribution Google Dataset Search



For a reason: **tables serve high-value decisions**

# Surprisingly, “tables” ignored as modality in neural AI



You
what to include in speech with PhD graduation committee after defense ceremony?

ChatGPT
After successfully defen
gratitude and appreciati
throughout your academ

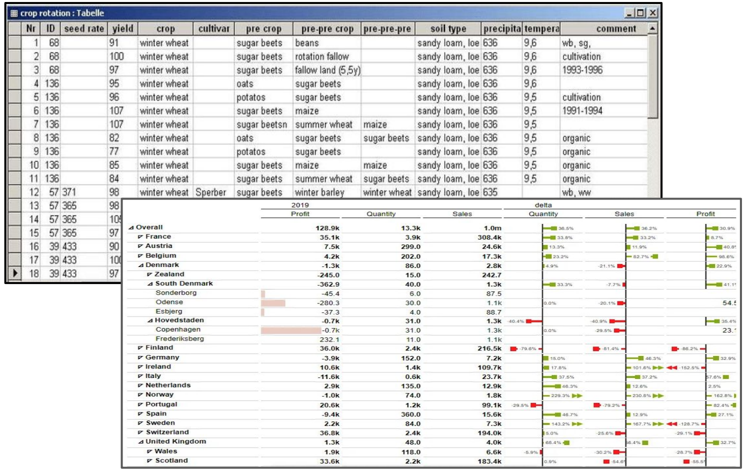
1. **Acknowledgment of th**
committee individually,
during the defense proc
2. **Express Gratitude**: Exp
mentors, colleagues, fri
throughout your PhD jo
3. **Reflect on the Journey**:
challenges you've overc
4. **Highlight Contributions**
impact in your field or re
work and its relevance.

```typescript
#!/usr/bin/env ts-node

import { fetch } from "fetch-h2";

// Determine whether the sentiment of text is positive
// Use a web service
async function isPositive(text: string): Promise<boolean> {
  const response = await fetch(`http://text-processing.com/api/sentiment/`, {
    method: "POST",
    body: `text=${text}`,
    headers: {
      "Content-Type": "application/x-www-form-urlencoded",
    },
  });
  const json = await response.json();
  return json.label === "pos";
}
```

**Text**, **Images**, **Code**…



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 | | | | | | | | | |
| 14 | 57 | 365 | 105 | | | | | | | | | |
| 15 | 57 | 365 | 97 | | | | | | | | | |
| 16 | 39 | 433 | 90 | | | | | | | | | |
| 17 | 39 | 433 | 100 | | | | | | | | | |
| 18 | 39 | 433 | 97 | | | | | | | | | |

| | 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 | 68.4% | 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 |

**Tables**?!

# Did a PhD on table semantics, but had a larger vision.

**The Table Representation Learning workshop @ NeurIPS 2022 was born.**

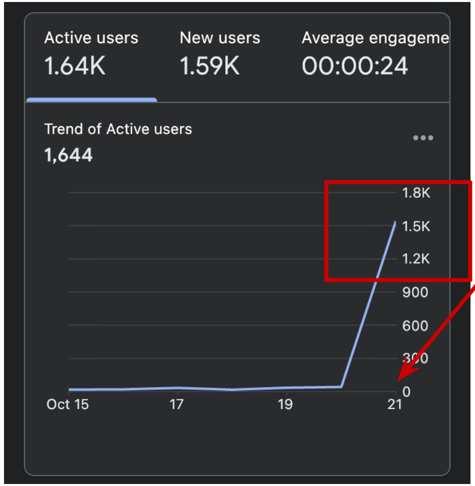
We received what we expected:

- Tabular QA / text-to-SQL,
- Synthetic data generation,
- Data preparation, etc.

And… neural models for predictive tabular ML.

# Anecdotes on neural models for tabular predictive ML

- Received neural predictive ML papers, rejected from main ML confs(?)
- Loud “pro-XGBoost” camp vs. small “pro-neural models” camp.
- Betted on pre-trained neural models for tables in ‘18: let’s empower vision!
- And, 1 paper intro’d a pre-trained cross-table model: TabPFN. It went viral:



Frank’s TabPFN tweet

Congrats team TabPFN, TabICL, ConTexTab, etc for pushing through and heading leaderboards!

# Tables weren't really cool in AI

But something happened in a tiny room in New Orleans at TRL @ NeurIPS 2022.

Great vibes, a wildly diverse community, trying to connect the dots.

**Tables were Back.**

# Fast-forward to 2025.

**Tabular AI is the “new hot topic” (quote CV researcher)**

**Tabular AI is Europe-led** 🇪🇺 (but let’s diversify).

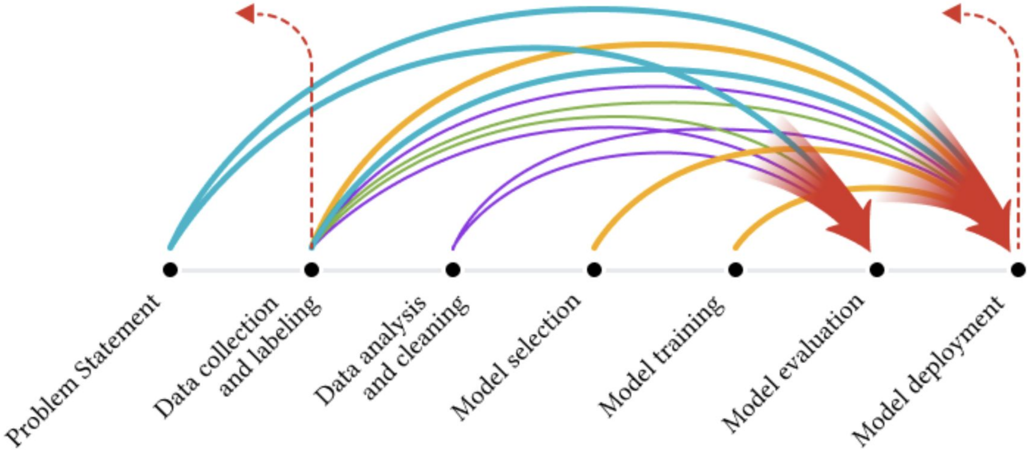
From an anonymous peer:

```
> is there no tab workshop this year at main conf bc you all took to eurips?
> either way wish I could be at the workshop :)
```

**And we’re only just beginning.**

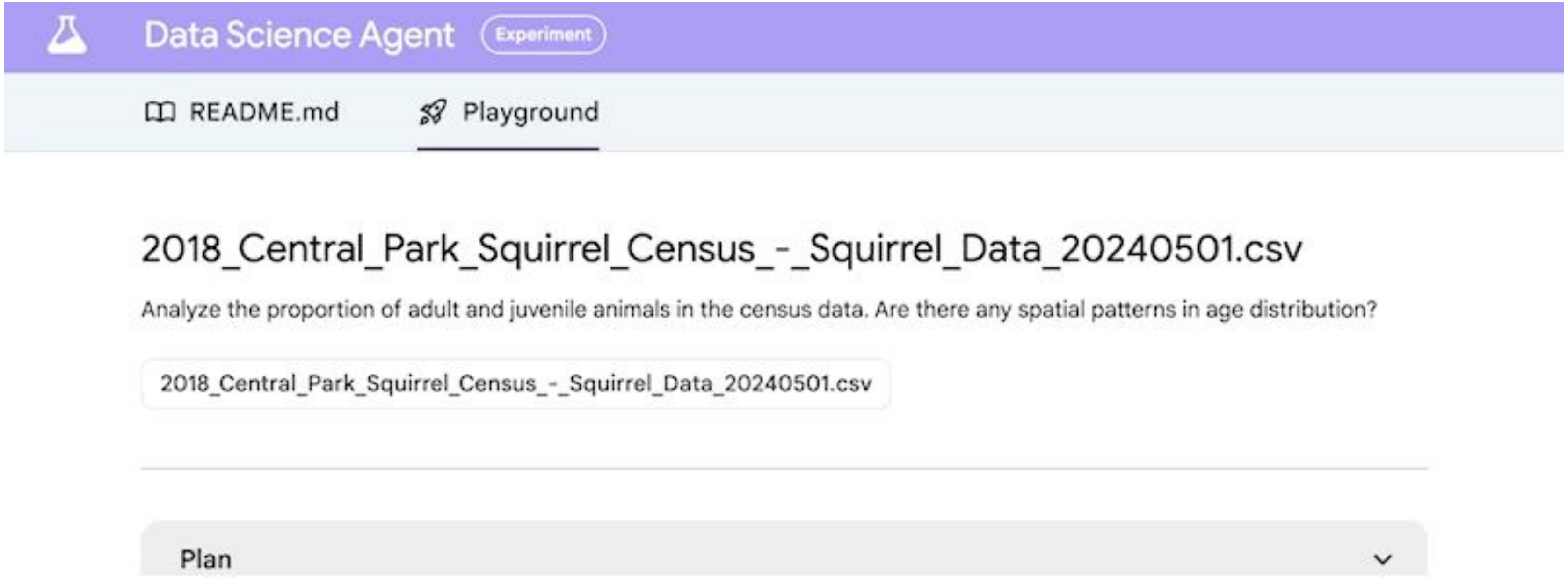
So, what *are we asking* from tabular data?

# Tabular pipelines are multi-faceted



*Replace “model” with “tool/dashboard” and it’s BI.

# LLM-enthusiasts make believe that DS is solved…

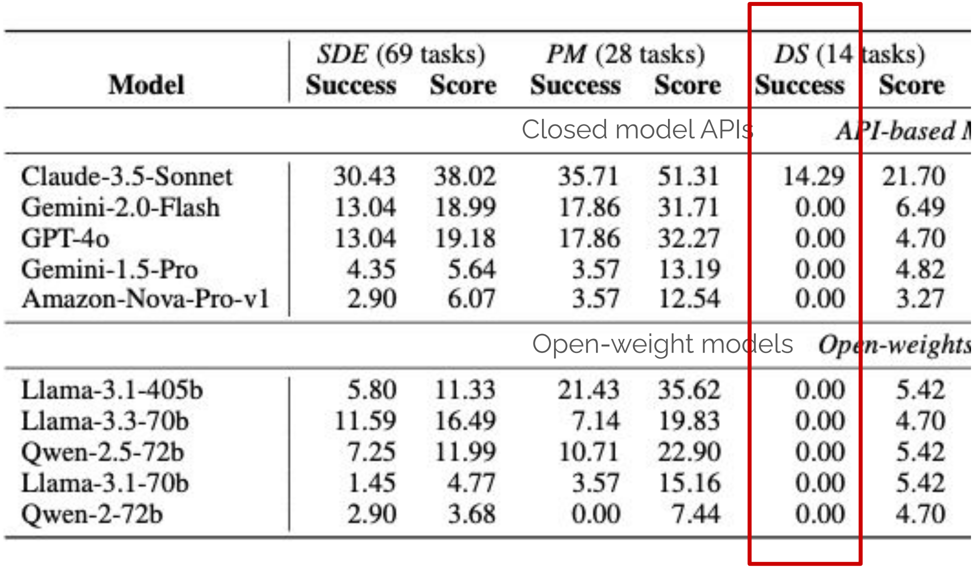


2018_Central_Park_Squirrel_Census_-_Squirrel_Data_20240501.csv

Analyze the proportion of adult and juvenile animals in the census data. Are there any spatial patterns in age distribution?

2018_Central_Park_Squirrel_Census_-_Squirrel_Data_20240501.csv

# It’s not



| Model | *SDE* (69 tasks) | | *PM* (28 tasks) | | *DS* (14 tasks) | |
|---|---|---|---|---|---|---|
| | **Success** | **Score** | **Success** | **Score** | **Success** | **Score** |
| | | | Closed model APIs | | | ***API-based M*** |
| Claude-3.5-Sonnet | 30.43 | 38.02 | 35.71 | 51.31 | 14.29 | 21.70 |
| Gemini-2.0-Flash | 13.04 | 18.99 | 17.86 | 31.71 | 0.00 | 6.49 |
| GPT-4o | 13.04 | 19.18 | 17.86 | 32.27 | 0.00 | 4.70 |
| Gemini-1.5-Pro | 4.35 | 5.64 | 3.57 | 13.19 | 0.00 | 4.82 |
| Amazon-Nova-Pro-v1 | 2.90 | 6.07 | 3.57 | 12.54 | 0.00 | 3.27 |
| | | | Open-weight models | | | ***Open-weights*** |
| Llama-3.1-405b | 5.80 | 11.33 | 21.43 | 35.62 | 0.00 | 5.42 |
| Llama-3.3-70b | 11.59 | 16.49 | 7.14 | 19.83 | 0.00 | 4.70 |
| Qwen-2.5-72b | 7.25 | 11.99 | 10.71 | 22.90 | 0.00 | 5.42 |
| Llama-3.1-70b | 1.45 | 4.77 | 3.57 | 15.16 | 0.00 | 5.42 |
| Qwen-2-72b | 2.90 | 3.68 | 0.00 | 7.44 | 0.00 | 4.70 |

(TheAgentCompany: Benchmarking LLM Agents on Consequential Real World Tasks, Xu et al., 2024)

# It all starts with retrieving the right data

# What table to use for my task?

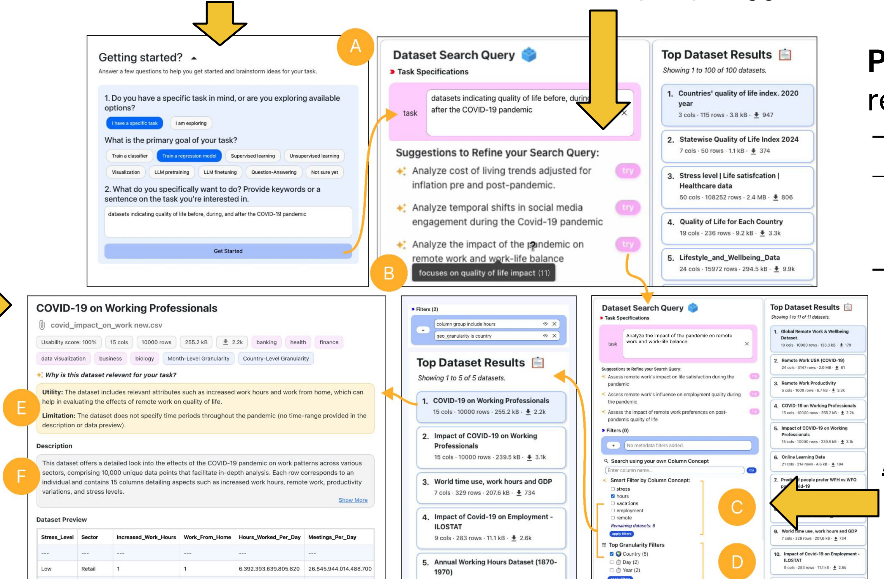
**In 2025**, we get “AGI” but *it still takes weeks to find the right dataset.*

*Task-driven* query

*Result-driven* query suggestions

*Task-driven* dataset relevance

*Data-driven* column suggestions



Getting started?

Answer a few questions to help you get started and brainstorm ideas for your task.

1. Do you have a specific task in mind, or are you exploring available options?

[✓ I have a specific task] I am exploring

What is the primary goal of your task?

Train a classifier [✓ Train a regression model] Supervised learning Unsupervised learning Visualization LLM pretraining LLM finetuning Question-Answering Not sure yet

2. What do you specifically want to do? Provide keywords or a sentence on the task you're interested in.

datasets indicating quality of life before, during, and after the COVID-19 pandemic

Get Started

Dataset Search Query

Task Specifications

task: datasets indicating quality of life before, during, and after the COVID-19 pandemic

Suggestions to Refine your Search Query:

- Analyze cost of living trends adjusted for inflation pre and post-pandemic. try
- Analyze temporal shifts in social media engagement during the Covid-19 pandemic try
- Analyze the impact of the pandemic on remote work and work-life balance try

focuses on quality of life impact (11)

Top Dataset Results

*Showing 1 to 100 of 100 datasets.*

1. Countries' quality of life index. 2020 year
   3 cols · 115 rows · 3.8 kB · 947
2. Statewise Quality of Life Index 2024
   7 cols · 50 rows · 1.1 kB · 374
3. Stress level | Life satisfcation | Healthcare data
   50 cols · 108252 rows · 2.4 MB · 806
4. Quality of Life for Each Country
   19 cols · 236 rows · 9.2 kB · 3.3k
5. Lifestyle_and_Wellbeing_Data
   24 cols · 15972 rows · 294.5 kB · 9.9k

COVID-19 on Working Professionals

covid_impact_on_work new.csv

Usability score: 100% · 15 cols · 10000 rows · 255.2 kB · 2.2k · banking · health · finance · data visualization · business · biology · Month-Level Granularity · Country-Level Granularity

***Why is this dataset relevant for your task?***

**Utility:** The dataset includes relevant attributes such as increased work hours and work from home, which can help in evaluating the effects of remote work on quality of life.

**Limitation:** The dataset does not specify time periods throughout the pandemic (no time-range provided in the description or data preview).

**Description**

This dataset offers a detailed look into the effects of the COVID-19 pandemic on work patterns across various sectors, comprising 10,000 unique data points that facilitate in-depth analysis. Each row corresponds to an individual and contains 15 columns detailing aspects such as increased work hours, remote work, productivity variations, and stress levels.

Show More

**Dataset Preview**

| Stress_Level | Sector | Increased_Work_Hours | Work_From_Home | Hours_Worked_Per_Day | Meetings_Per_Day |
|---|---|---|---|---|---|
| --- | --- | --- | --- | --- | --- |
| Low | Retail | 1 | 1 | 6.392.393.639.805.820 | 26.845.944.014.488.700 |

Filters (2)

column group include hours

geo_granularity is country

Top Dataset Results

*Showing 1 to 5 of 5 datasets.*

1. COVID-19 on Working Professionals
   15 cols · 10000 rows · 255.2 kB · 2.2k
2. Impact of COVID-19 on Working Professionals
   15 cols · 10000 rows · 239.5 kB · 3.1k
3. World time use, work hours and GDP
   7 cols · 329 rows · 207.6 kB · 734
4. Impact of Covid-19 on Employment - ILOSTAT
   9 cols · 283 rows · 11.1 kB · 2.6k
5. Annual Working Hours Dataset (1870-1970)

Dataset Search Query

Task Specifications

task: Analyze the impact of the pandemic on remote work and work-life balance

Suggestions to Refine your Search Query:

- Assess remote work's impact on life satisfaction during the pandemic try
- Assess remote work's influence on employment quality during the pandemic try
- Assess the impact of remote work preferences on post-pandemic quality of life try

Filters (0)

No metadata filters added.

Search using your own Column Concept

Enter column name... try

Smart Filter by Column Concept:

- ☐ stress
- ☑ hours
- ☐ vacations
- ☐ employment
- ☐ remote

*Remaining datasets: 8*

apply filters

Top Granularity Filters

- ☑ Country (6)
- ☐ Day (2)
- ☐ Year (2)

Top Dataset Results

*Showing 1 to 11 of 11 datasets.*

1. Global Remote Work & Wellbeing Dataset.
   10 cols · 10000 rows · 133.3 kB · 179
2. Remote Work USA (COVID-19)
   24 cols · 3147 rows · 2.0 MB · 81
3. Remote Work Productivity
   5 cols · 1000 rows · 6.7 kB · 3.3k
4. COVID-19 on Working Professionals
   15 cols · 10000 rows · 255.2 kB · 2.2k
5. Impact of COVID-19 on Working Professionals
   15 cols · 10000 rows · 239.5 kB · 3.1k
6. Online Learning Data
   21 cols · 214 rows · 4.6 kB · 184
7. Predi f people prefer WFH vs WFO p id-19
9. World time use, work hours and GDP
   7 cols · 329 rows · 207.6 kB · 734
10. Impact of Covid-19 on Employment - ILOSTAT
    9 cols · 283 rows · 11.1 kB · 2.6k

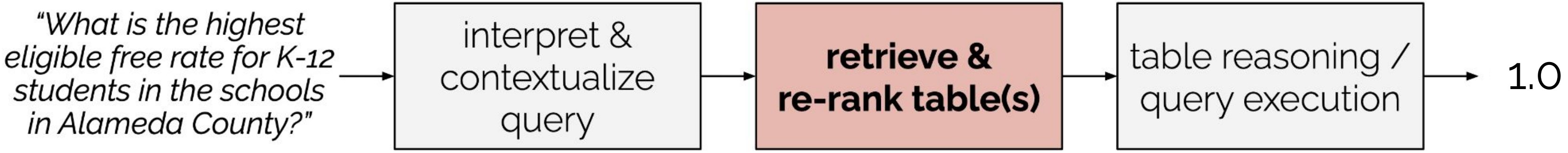
**Proactive search** makes it easier to find more relevant datasets, with higher success rate.

| Condition | Ease-of-use | Relevance | # Successes |
|---|---|---|---|
| (A) Kaggle | $\mu$=3.08; $\sigma$=0.51 | $\mu$=3.25; $\sigma$=1.05 | 7 *of* 12 |
| (B) Semantic Baseline | $\mu$=3.75; $\sigma$=0.45 | $\mu$=3.25; $\sigma$=0.86 | 6 *of* 12 |
| (C) DataScout | $\mu$=4.75; $\sigma$=0.45 | $\mu$=3.67; $\sigma$=0.78 | 10 *of* 12 |

*Rethinking Dataset Discovery with DataScout*
Lin, R., Chopra, B., Lin, W., Shankar, S., Hulsebos, M., Parameswaran, A., 2025.

UC Berkeley CWI

# How to query tables end-to-end? ~RAG over tables



*"What is the highest eligible free rate for K-12 students in the schools in Alameda County?"*

interpret & contextualize query

**retrieve & re-rank table(s)**

table reasoning / query execution

1.0

How to best retrieve tables in end-to-end tabular QA? Some findings:

- BM25 **less effective than for text**, requires highly descriptive metadata.
- Retrieval over generated summaries helps (but tricky for enterprise-y tables)
- **Row-level retrieval most effective for relational DBs**, but infeasible (many, large tables).

There's a very long road ahead.      Imagine predictive questions over data lakes?

*TARGET: benchmarking Table Retrieval for Generative Tasks*
Ji, X., Parker, G., Parameswaran, A., Hulsebos, M., 2024

*Metadata Matters in Dense Table Retrieval*
Gomm, D., Hulsebos, M., 2025

# Are We Asking the Right Questions?

In end-to-end tabular data analysis, we desire queries that perfectly express the ***insight need***: the data, the operation, and output.

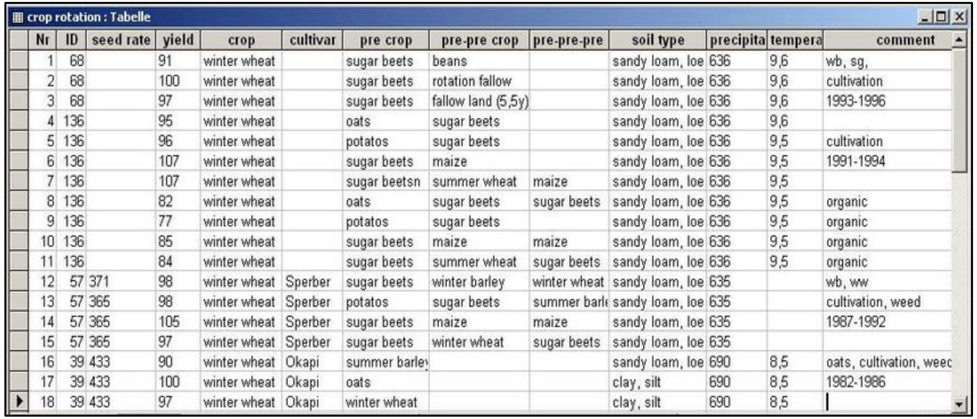
Such queries are **platonic**, we won’t see them.

But we need to understand **queries for e2d tabular data analysis** to build better systems and eval capabilities. (*join Daniel’s talk for more*)

*Are We Asking the Right Questions? On Query Ambiguity in Tabular Data Analysis*
Gomm, D., Wolff, C., Hulsebos, M., 2025

# Tables are complex,
# they require context

# What is this table about?



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 | |

Pre crop?
Pre-pre crop?
Pre-pre-pre?

A useful signal for data viz, data prep (integration, missing val mech.), ML, etc.

*Sherlock*: (tiny!) neural model surfaces **table semantics** by column embeddings.

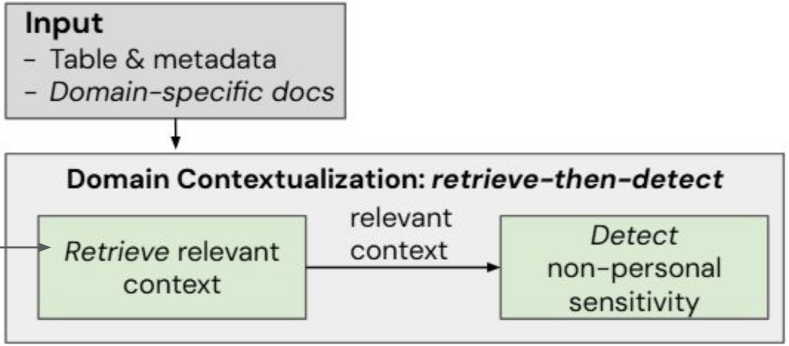
*Sherlock: A Deep Learning Approach to Semantic Data Type Detection,* Hulsebos et al., 2019.

# Is this data sensitive?

Understudied. Beyond fixed types (pii) sensitivity depends on context.

Requires ***context beyond the table***.

data governance policy documents



**Input**

- Table & metadata
- *Domain-specific docs*

**Domain Contextualization: *retrieve-then-detect***

*Retrieve* relevant context

relevant context

*Detect* non-personal sensitivity

Result:

- Better precision.
- Context-grounded (LLM) explanations.

Imagine in healthcare: contextualizing **EHR tables**, in **patient-doctor transcripts**, **medicine docs**, etc?

*Towards Contextual Sensitive Data, Telkamp and Hulsebos, 2025.*

But eventually,
we want that insight

# What is the answer to my question?

**Typical questions that drive decisions:**

- Analytical questions → *e.g.* what has happened?
- Predictive queries → *e.g.* what might happen?

Some stats from Gael's keynote at TRL @NeurIPS '24 resonated.

**Pypi #downloads last month** (updated):

- **Scikit-learn: 189,875,197**
- **Pandas: 502,255,215**

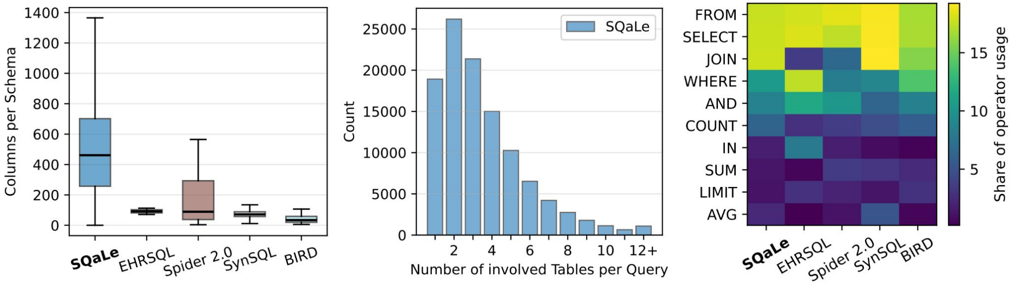
# Analytical questions: we can use SQL

Why enter the *text*-to-*SQL* game?

- Saturated (but not solved), everyone just throwing LLMs at it.
- SQL made for humans not machines: it’s too flexible, different abstraction?

Then, Cornelius (PhD) convinced me that we need **SQaLe to specialize**.

- Large text-to-SQL dataset
- 500K+ (question, sql, schema)
- Grounded in *real* schemas

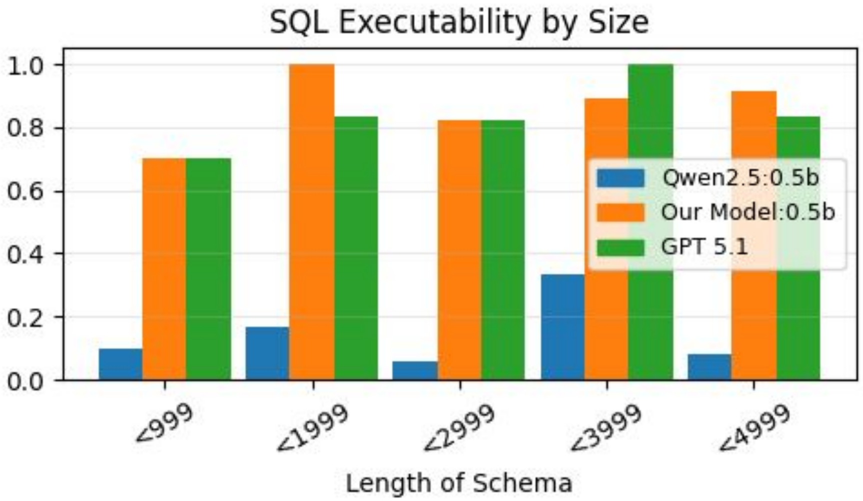


*SQaLe: A large text-to-SQL corpus grounded in real schemas, Wolff, Gomm, and Hulsebos, 2025.*

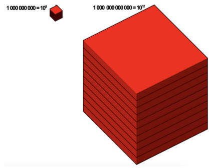
CWI

# Specialization over general-purpose

Initial Qwen model *fine-tuned* for text-to-SQL on **SQaLe**:



0.5 **billion** params vs 1.8 **trillion** params



*Find Cornelius in the poster session for more.

# Executability is not accuracy, sure.

*“For each customer in the gold or platinum loyalty tier who has placed at least two delivered orders in the last six months, return their full name, email, city and state, the total number of delivered orders, the total amount they paid based only on captured payments, their most frequently ordered product category in that period, and whether any of the products they ordered currently have low inventory in any warehouse (defined as quantity on hand minus quantity reserved less than 10), sorted by total captured payment amount from highest to lowest.”*

Realistic phrasing? Maybe not: inherited from source.
But look at that complex SQL! Promising start.

```
WITH gold_platinum_customers AS
  (SELECT c.id,
          c.first_name,
          c.last_name,
          c.email,
          c.city,
          c.state
   FROM customers c
   JOIN orders o ON c.id = o.customer_id
   WHERE c.loyalty_tier IN ('gold',
                            'platinum')
     AND o.status = 'delivered'
     AND o.delivered_at >= date('now', '-6 months')
   GROUP BY c.id
   HAVING COUNT(*) >= 2),
     order_details AS
  (SELECT gpc.id,
          gpc.first_name,
          gpc.last_name,
          gpc.email,
          gpc.city,
          gpc.state,
          o.order_number,
          SUM(p.amount) AS total_paid,
          p.status AS payment_status
   FROM gold_platinum_customers gpc
   JOIN orders o ON gpc.id = o.customer_id
   LEFT JOIN payments p ON o.id = p.order_id
   AND p.status = 'captured'
   WHERE o.delivered_at >= date('now', '-6 months')
     AND o.status IN ('delivered',
                      'completed')
   GROUP BY gpc.id,
            o.id),
     product_category AS
  (SELECT od.id,
          od.first_name,
          od.last_name,
          od.email,
          od.city,
          od.state,
          p.category,
          ROW_NUMBER() OVER (PARTITION BY od.id
                             ORDER BY COUNT(*) DESC) AS rn
   FROM order_details od
   JOIN order_items oi ON od.order_number = CAST(oi.order_id AS TEXT)
   JOIN products p ON oi.product_id = p.id
   GROUP BY od.id)
SELECT gpc.first_name,
       gpc.last_name,
       gpc.email,
       gpc.city,
       gpc.state,
       COUNT(od.order_number) AS total_orders,
       SUM(od.total_paid) AS total_captured_payment_amount
FROM gold_platinum_customers gpc
JOIN order_details od ON gpc.id = od.id
AND od.payment_status = 'captured'
LEFT JOIN product_category pc ON od.id = pc.id
AND pc.rn = 1
WHERE pc.category IS NOT NULL
GROUP BY gpc.id
ORDER BY total_captured_payment_amount DESC;
```

What *should we ask* from tabular data, next?

# Generalization versus specialization?

Deriving insights from tabular data to inform decisions is an inherently multi-faceted problem, **perfect for smaller specialized models**, but still looking for the right level of abstraction, which may vary.

Eventually tabular systems are a mix and match:

- Databases and other tools for analytical queries, data prep, etc.
- Some ML models for statistical reasoning
- Some human, document, and LLM contextualization
- Some agentic capabilities
- Some fluid interfaces that connect things together

# Still, many open questions

What is a table? What is a relation? How to deal with different representations?

What data do we encounter in practice?

What queries are asked? What can and should be asked? → check Daniel’s talk

How to contextualize tabular data w/ domain knowledge from docs / human input?

How do we want to enable interaction with tabular data?

What is the ideal scope for foundation models? When to generalize vs specialize?

How do we go full-cycle from data collection to decision? What can we “agentize”?

# Where does the community stand?

Excitement:

- **Tabular AI** has gained momentum – it's recognized widely!
- Academic / VC funding is flowing.
- There's much more to **explore and exploit**.

❤ **Tabular AI community**!

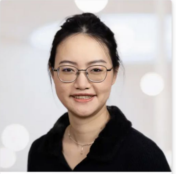
You, too? We're looking for committed members to foster the community.
*Reach out if you want to contribute, stay tuned for more initiatives :).*

# Reach out

TRL Lab

Madelon Hulsebos

Faculty



Xue (Effy) Li

Postdoctoral researcher

Daniel Gomm

PhD researcher

Cornelius Wolff

PhD researcher

Jan Henrik Bertrand

Research student (ELLIS MSc Honours Program)

Wojciech Kosiuk

Research student (ELLIS MSc Honours Program)

**madelon@cwi.nl**

**https://trl-lab.github.io/**