Large Scale Data Integration Project

Prof. Dr. Ziawasch Abedjan

Muaid Mughrabi, Binger Chen

Data Integration and Data preparation Group (D2IP)

LSDIPRO Team

- Muaid Mughrabi
  - muaid.mughrabi@tu-berlin.de
  - Seminar Organization and Topic Supervision

- Dr. Binger Chen
  - chen@tu-berlin.de
  - Topic supervision

- Prof. Dr. Ziawasch Abedjan
  - abedjan@tu-berlin.de
  - Here to enjoy

Project Goal and Grading Criteria

Soft Goals:

1. Reproduce a given paper
2. Create a competitive submission for the BTW Data Science challenge

We look at:

Code

Presentation

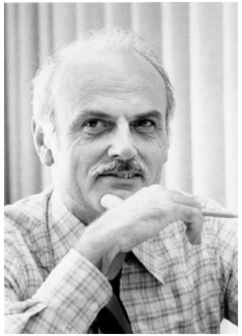
Individual contribution sheet (1A4 page)

Hard Goals (relevant for grading):

1. Clean and documented code repository
2. Systematic evaluation of generated solutions
3. Scaling solutions to the size of underlying datasets
4. Clearly describing, visualizing and presenting the results in a presentation

Database Research in the 70s-90s

- The Coddfather spoke:



Managing Data shall be Science!

Memory hierarchy
Portability
Optimization
User concurrency
Distribution
Usability
....

Jim Gray

Donald Chamberlin

Make it happen!

Pat Sellinger

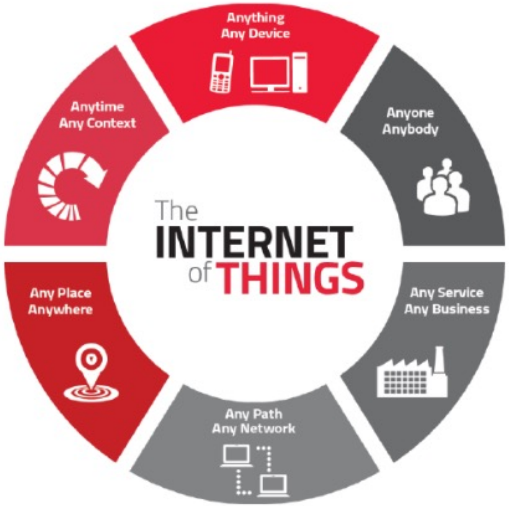
Michael Stonebraker

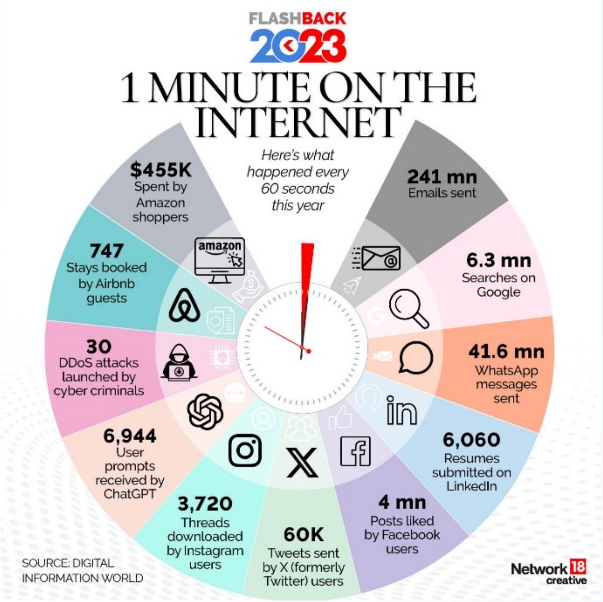
McKinsey&Company

2011

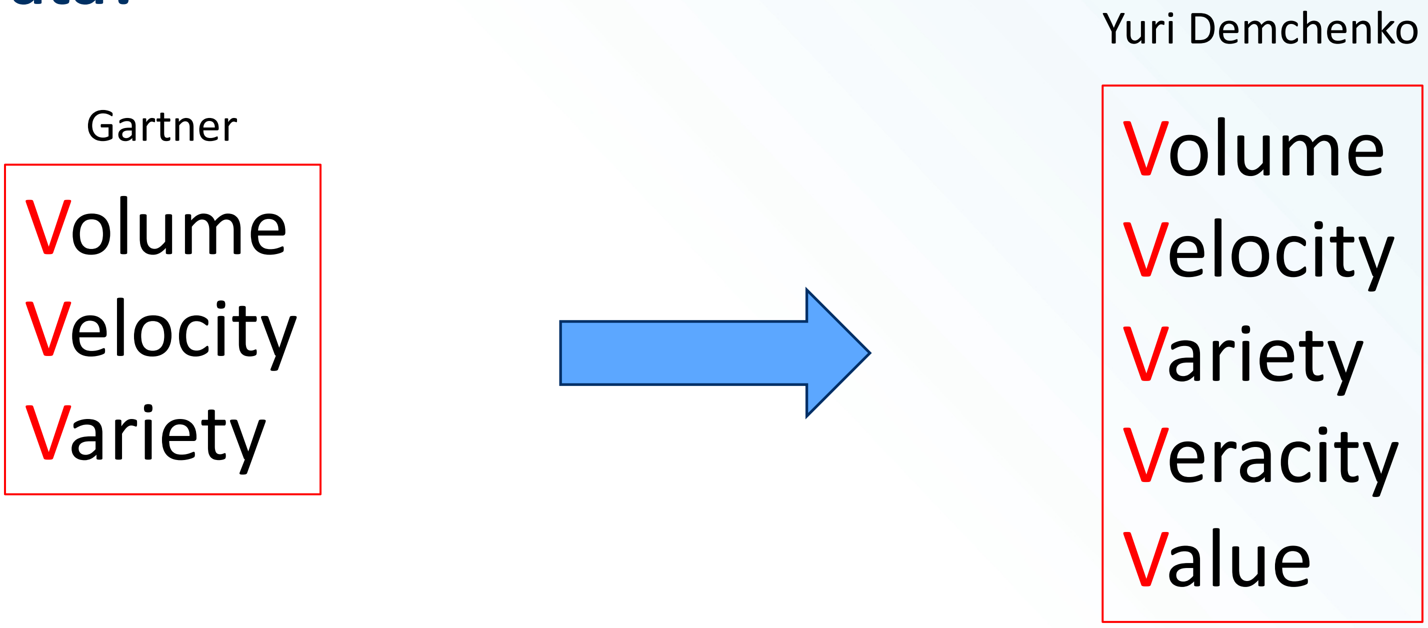
Big data: The next frontier for innovation, competition, and productivity

Where do we find Big Data ?





Big Data?

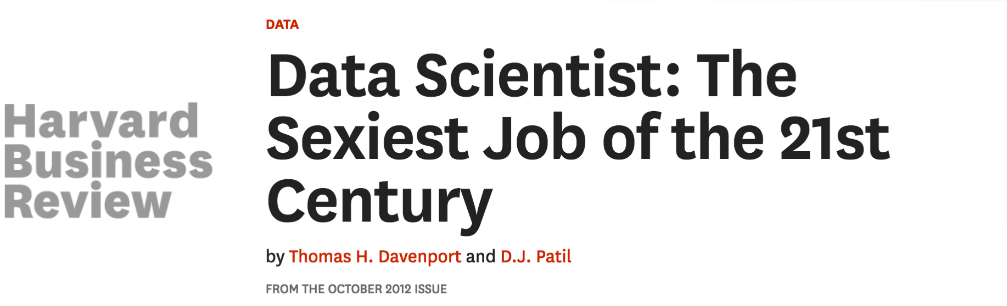


Gartner

- Volume
- Velocity
- Variety

Yuri Demchenko

- Volume
- Velocity
- Variety
- Veracity
- Value



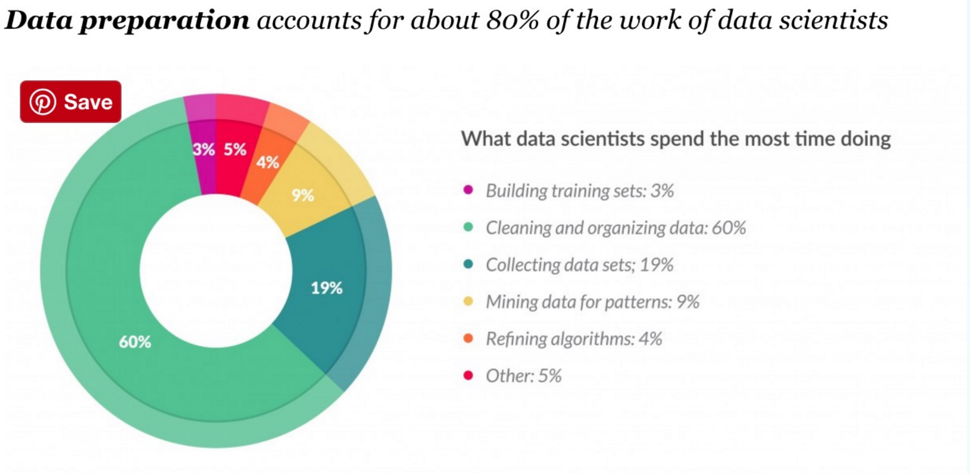
Harvard Business Review

DATA

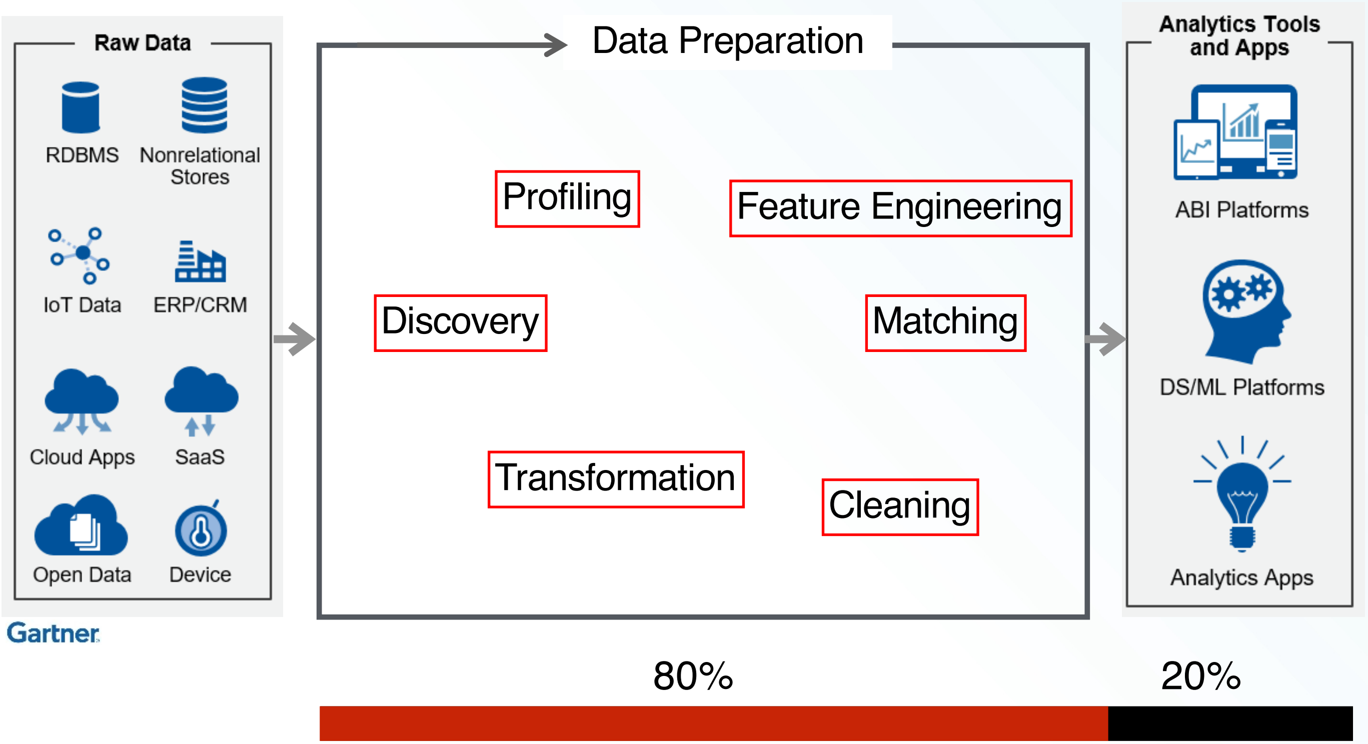
Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE



***Data preparation** accounts for about 80% of the work of data scientists*



Raw Data

RDBMS

Nonrelational Stores

IoT Data

ERP/CRM

Cloud Apps

SaaS

Open Data

Device

Gartner

Data Preparation

Profiling

Feature Engineering

Discovery

Matching

Transformation

Cleaning

Analytics Tools and Apps

ABI Platforms

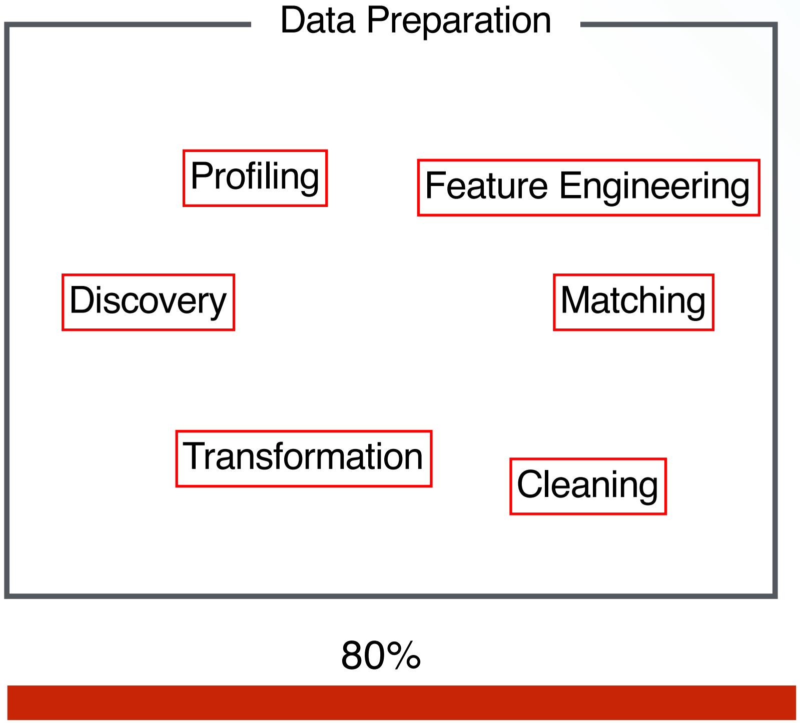
DS/ML Platforms

Analytics Apps

80%

20%

Focus of our research



Data Preparation

Profiling

Feature Engineering

Discovery

Matching

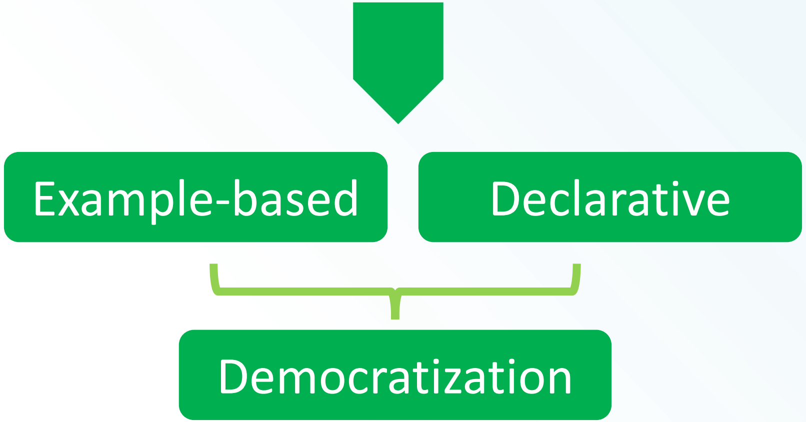
Transformation

Cleaning

80%

Our goal:

- Hide complexity
- Empower domain experts
- Make routines task-aware



Example-based

Declarative

Democratization

Topics

1.BTW 2025 Data Science Challenge

Integrating various sources for improving forecasting

2.Transformation Discovery

Using Web resources to find transformation functions via input output examples

3.Table Question Answering

Use the concept of chain of thoughts on an LLM to generate functions that retrieve the answer for a question in natural languages

4.AutoTables:

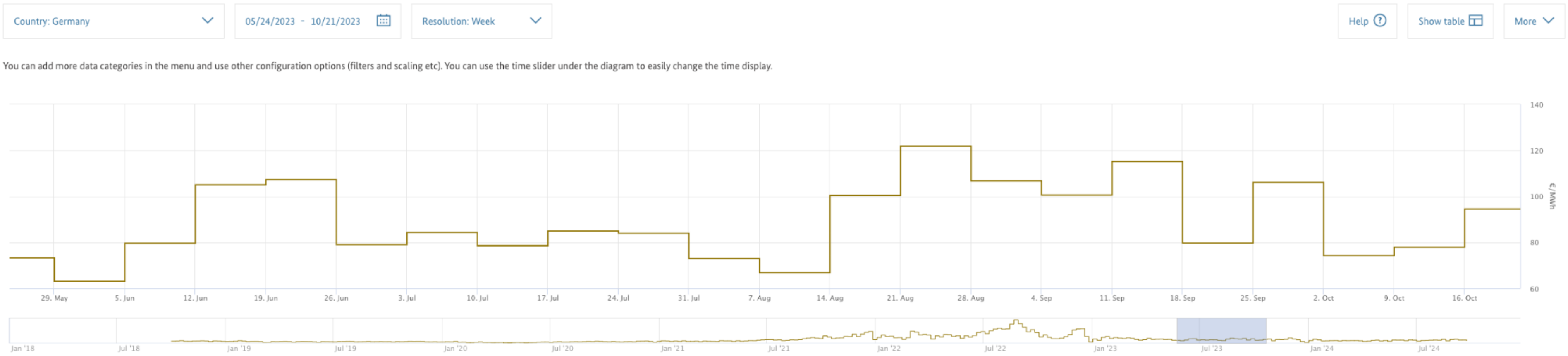
Synthesizing tables without example

5.Finding related tables

Given an input table and a corpus of tables, find tables that enrich your table https://btw2025.gi.de/data-science-challenge

BTW 2025

- Supervisor: Muaid Mughrabi
- Goal: Predict hourly day-ahead energy prices for Germany on February 18, 2025
- Problem: Not enough data and features available for accurate prediction



Country: Germany

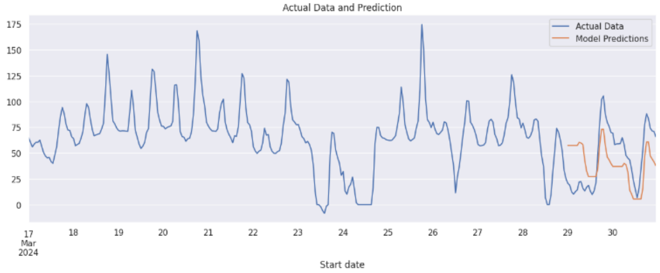
05/24/2023 - 10/21/2023

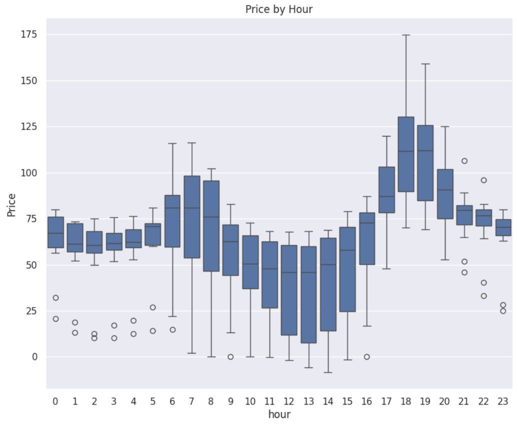
Resolution: Week

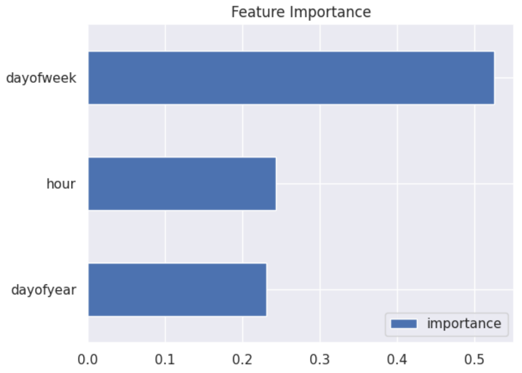
You can add more data categories in the menu and use other configuration options (filters and scaling etc). You can use the time slider under the diagram to easily change the time display.

BTW 2025

- Data collection
- Data curation
- Prediction
- Explanation and Visualization
- Be Creative!







BTW 2025: Tasks

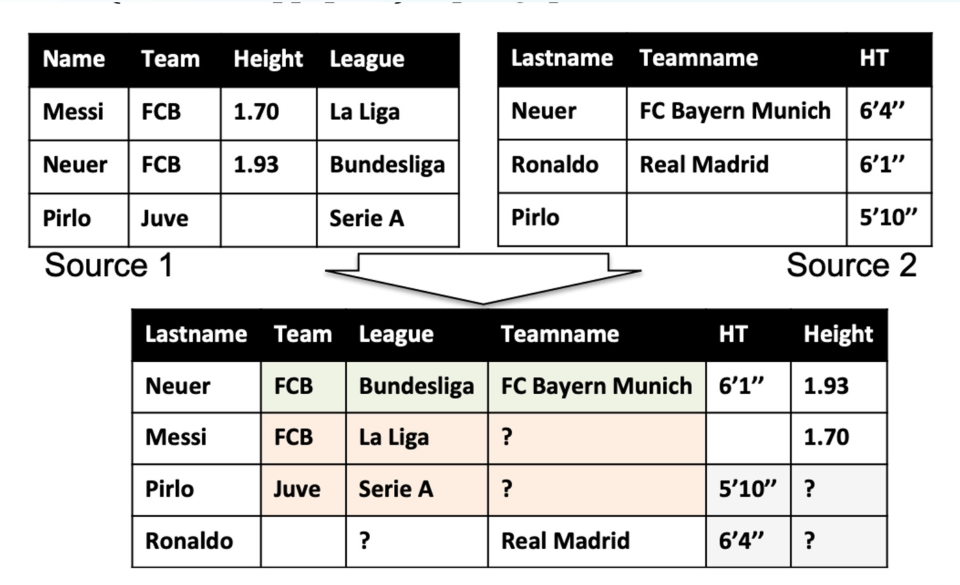
- Gathering Domain Knowledge & Data Sources
  - Data crawling from DMARD.DE
  - Weather information?
  - Special events and how to account for them: pandemic, war, news, etc.
- Data Preparation
  - Data cleaning, removing anomalies if needed, investigating variable relationships, and creating statistical summaries.
  - Training, Testing, and Validation splits (without harming temporal information)
- Visualization
  - Plot trends and forecasts, sanity check model performance visually, and explain your findings
- Predictive Modelling:
  - Develop a forecasting model, compare it against a reasonable baseline, and iterate on it.

https://cs.uwaterloo.ca/~ilyas/papers/AbedjanICDE16.pdf

Transformation Discovery with DataXFormer

Supervisor: Muaid Mughrabi
Problem:

- Tables often have missing values/ Representations need to be unified
- Using external resources, such as tables, functions, catalogues, one can fill in the gaps
- DataXFormer swifts through relational tables finding tables that contain examples of one-to-one and many-to-one mappings



Name	Team	Height	League
Messi	FCB	1.70	La Liga
Neuer	FCB	1.93	Bundesliga
Pirlo	Juve		Serie A

Source 1

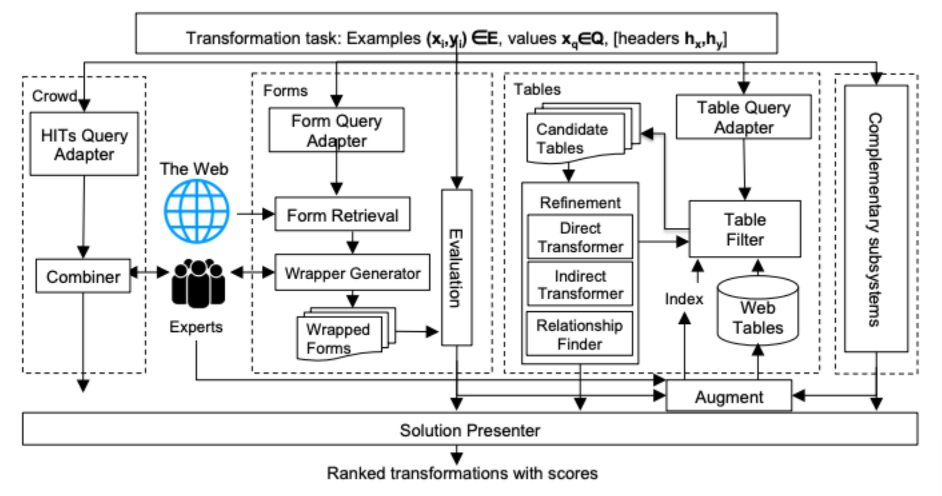
Lastname	Teamname	HT
Neuer	FC Bayern Munich	6’4”
Ronaldo	Real Madrid	6’1”
Pirlo		5’10”

Source 2

Lastname	Team	League	Teamname	HT	Height
Neuer	FCB	Bundesliga	FC Bayern Munich	6’1”	1.93
Messi	FCB	La Liga	?		1.70
Pirlo	Juve	Serie A	?	5’10”	?
Ronaldo		?	Real Madrid	6’4”	?

Task: Reproduce DataXFormer

- Implement DataXformer
  - Table retrieval
  - Transformation Discovery
  - EM Algorithm
- Evaluate your implementation
  - Correctness and Runtime
  - Dataset coverage
- Improve one aspect!
  - Partial Strings inside Cells
  - Be creative!



https://arxiv.org/pdf/2401.04398

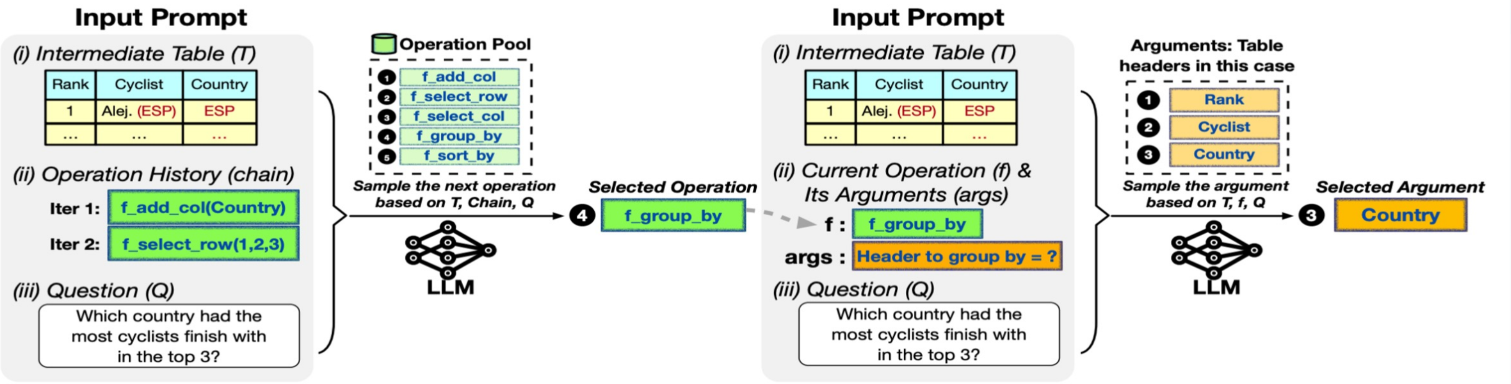
Chain of Tables

Supervisor: Muaid Mughrabi

Goal: Generating programs that extract the answers to a natural language question from a table

Problem:

- How to represent tables?
- How to prompt the LLM?
- How to keep the chain of operations?


Input Prompt
(i) Intermediate Table (T)
(ii) Operation History (chain)
Iter 1: f_add_col(Country)
Iter 2: f_select_row(1,2,3)
(iii) Question (Q)
Which country had the most cyclists finish with in the top 3?
Operation Pool
1 f_add_col
2 f_select_row
3 f_select_col
4 f_group_by
5 f_sort_by
Sample the next operation based on T, Chain, Q
LLM
Selected Operation
4 f_group_by
Input Prompt
(i) Intermediate Table (T)
(ii) Current Operation (f) & Its Arguments (args)
f : f_group_by
args : Header to group by = ?
(iii) Question (Q)
Which country had the most cyclists finish with in the top 3?
Arguments: Table headers in this case
1 Rank
2 Cyclist
3 Country
Sample the argument based on T, f, Q
LLM
Selected Argument
3 Country

Rank	Cyclist	Country
1	Alej. (ESP)	ESP
...	...	...

Chain of Tables: Tasks

- Implement Chain of Tables using small LLM
  - Table Actions implementation
  - Prompting LLM to generate plan and action args
  - Ensure that Table actions are performed correctly
- Evaluate your implementation
  - Against different LLMs
  - You need to start with smaller model with your notebooks
  - Performance and Runtime
- Improvement suggestion!
  - Instead of generating one plan, generate multiple plan candidates
    - Implement an evaluator to score each state, traverse graph based on given scores
  - Be creative!

Peng Li, Yeye He, Cong Yan, Yue Wang, Surajit Chaudhuri: Auto-Tables: Synthesizing Multi-Step Transformations to Relationalize Tables without Using Examples. Proc. VLDB Endow. 16(11): 3391-3403 (2023)

https://www.vldb.org/pvldb/vol16/p3391-he.pdf

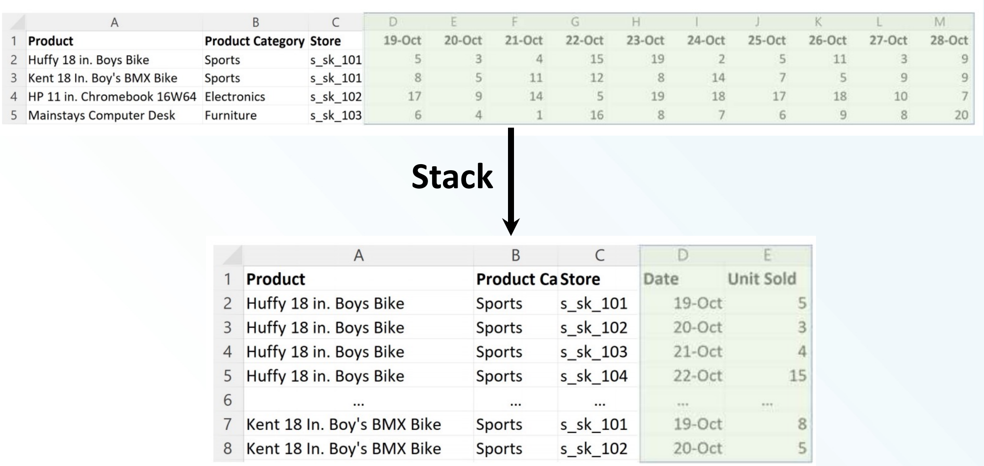
AutoTables

Supervisor: Binger Chen

Goal: Discovering the operators and transforming non-relational tables to relational tables

Transforming operators (choose 2):

- Stack: transforming homogeneous columns into rows. (Pandas API: melt)
- Transpose: transforming rows to columns and vice versa. (Pandas API: transpose)
- Wide-to-long: transforming repeating column groups into rows. (Pandas API: wide)
- Pivot: transforming repeating row groups into columns. (Pandas API: pivot)



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**Product**	**Product Category**	**Store**	19-Oct	20-Oct	21-Oct	22-Oct	23-Oct	24-Oct	25-Oct	26-Oct	27-Oct	28-Oct
2	Huffy 18 in. Boys Bike	Sports	s_sk_101	5	3	4	15	19	2	5	11	3	9
3	Kent 18 In. Boy's BMX Bike	Sports	s_sk_101	8	5	11	12	8	14	7	5	9	9
4	HP 11 in. Chromebook 16W64	Electronics	s_sk_102	17	9	14	5	19	18	17	18	10	7
5	Mainstays Computer Desk	Furniture	s_sk_103	6	4	1	16	8	7	6	9	8	20

	A	B	C	D	E
1	**Product**	**Product Ca**	**Store**	Date	Unit Sold
2	Huffy 18 in. Boys Bike	Sports	s_sk_101	19-Oct	5
3	Huffy 18 in. Boys Bike	Sports	s_sk_102	20-Oct	3
4	Huffy 18 in. Boys Bike	Sports	s_sk_103	21-Oct	4
5	Huffy 18 in. Boys Bike	Sports	s_sk_104	22-Oct	15
6	...	...	...	...	...
7	Kent 18 In. Boy's BMX Bike	Sports	s_sk_101	19-Oct	8
8	Kent 18 In. Boy's BMX Bike	Sports	s_sk_102	20-Oct	5

AutoTables: Tasks

- Generate a training dataset from relational table dataset
    - Method: leverage inverse operators
    - Relational dataset source: https://relational-data.org/
- Implement an Input-only model
    - Model input: non-relational table.
    - Model output: operator used for transformation.
    - Feature extraction based on CNN from computer vision as the pattern is visual.
- Implement Input/Output re-reranking model
    - Model input: top-k transformed tables from input-only model.
    - Model output: ranking score of these tables.
    - Similar feature extraction as Input-only model
- Scaling(optional): Implement data augmentation:
    - (1)Cropping (2)Shuffling

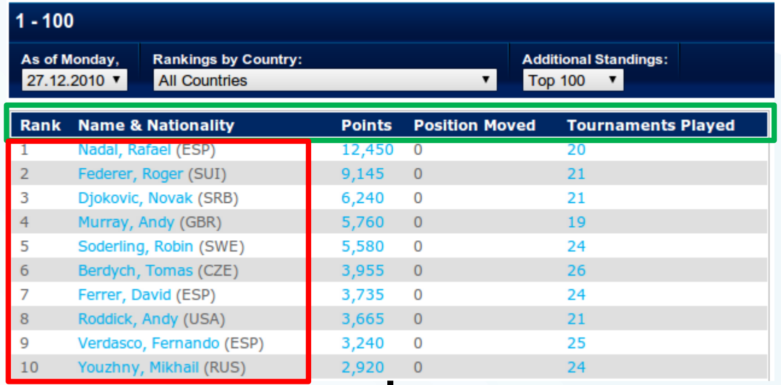
Anish Das Sarma, Lujun Fang, Nitin Gupta, Alon Y. Halevy, Hongrae Lee, Fei Wu, Reynold Xin, Cong Yu: Finding related tables. SIGMOD Conference 2012: 817-828

https://amplab.cs.berkeley.edu/wp-content/uploads/2012/06/finding-related-tables.pdf

Finding Related Tables

Supervisor: Binger Chen

Goal: Finding Related Tables based on Entity Complement
- Find Columns that contain semantically similar entities but complement each other

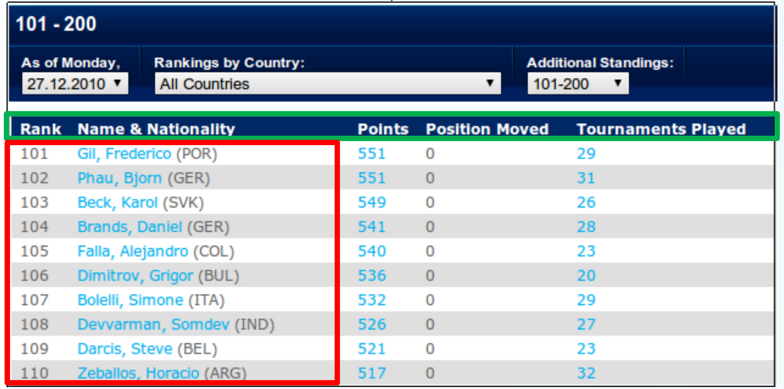


1 - 100

As of Monday, 27.12.2010 ▼ Rankings by Country: All Countries ▼ Additional Standings: Top 100 ▼

Rank	Name & Nationality	Points	Position Moved	Tournaments Played
1	Nadal, Rafael (ESP)	12,450	0	20
2	Federer, Roger (SUI)	9,145	0	21
3	Djokovic, Novak (SRB)	6,240	0	21
4	Murray, Andy (GBR)	5,760	0	19
5	Soderling, Robin (SWE)	5,580	0	24
6	Berdych, Tomas (CZE)	3,955	0	26
7	Ferrer, David (ESP)	3,735	0	24
8	Roddick, Andy (USA)	3,665	0	21
9	Verdasco, Fernando (ESP)	3,240	0	25
10	Youzhny, Mikhail (RUS)	2,920	0	24

Entity complement



101 - 200

As of Monday, 27.12.2010 ▼ Rankings by Country: All Countries ▼ Additional Standings: 101-200 ▼

Rank	Name & Nationality	Points	Position Moved	Tournaments Played
101	Gil, Frederico (POR)	551	0	29
102	Phau, Bjorn (GER)	551	0	31
103	Beck, Karol (SVK)	549	0	26
104	Brands, Daniel (GER)	541	0	28
105	Falla, Alejandro (COL)	540	0	23
106	Dimitrov, Grigor (BUL)	536	0	20
107	Bolelli, Simone (ITA)	532	0	29
108	Devvarman, Somdev (IND)	526	0	27
109	Darcis, Steve (BEL)	521	0	23
110	Zeballos, Horacio (ARG)	517	0	32

Finding Related Tables: Tasks

- Measure the relatedness of two tables:
  - entity consistency: they have entities from the same class, such as all tennis players
  - schema similarity: they have same table schema

- Leverage signals from the web source to infer entity groups
  - Knowledge bases providing signals: WebIsA, DBpedia, etc.
  - Example of signals: (Entity: Paris, Class: City)

- Leverage signals from the web source to infer entity groups

- Scaling(optional): implement all above on big dataset

Important Dates

17.10. Topic selection, group formation

24.10. Weekly meetings

Identify the main idea, the problem being solved, who/what is involved, understood paper, find datasets, get familiar with repository

31.10.Weekly meetings

Present a plan

Per person: what is going to be implemented/ tested and what was done the week before.

07.11. Weekly meetings (development)

14.11. Weekly meetings (first pipeline of the software should e ready, development)

21.11. Weekly meetings (development)

28.11. Weekly meetings (development)

05.12. Expert review

A different group will test your system

12.12. System improvement, experiments.

19.12. Weekly meetings (experiments)

09.01. Weekly meetings (experiments/Start writing the report)

16.01. Weekly meetings (experiments, visualization)

23.01. Weekly meetings (experiments, visualization)

30.01. Weekly meetings (Wrap up documentation, and presentation slides)

06.02. Final presentations.

13.02. Final presentations.

14.02. One A4 page describing individual contributions on the project.

TODOS

- Identify group members via ISIS
- Send Muaid Mughrabi until October 21st 5pm your chosen top 3 topics