BENCHMARKING FOR TABULAR REPRESENTATION MODELS ON LONGITUDINAL DATASETS

Pratik Pokharel, Oliver Kennedy

Introduction

Longitudinal dataset schemas are inherently dynamic, evolving over time. Questionnaires may be added, re-moved, rephrased, or restructured to meet the evolving demands of the study at different points.

Form 1

Family Details:

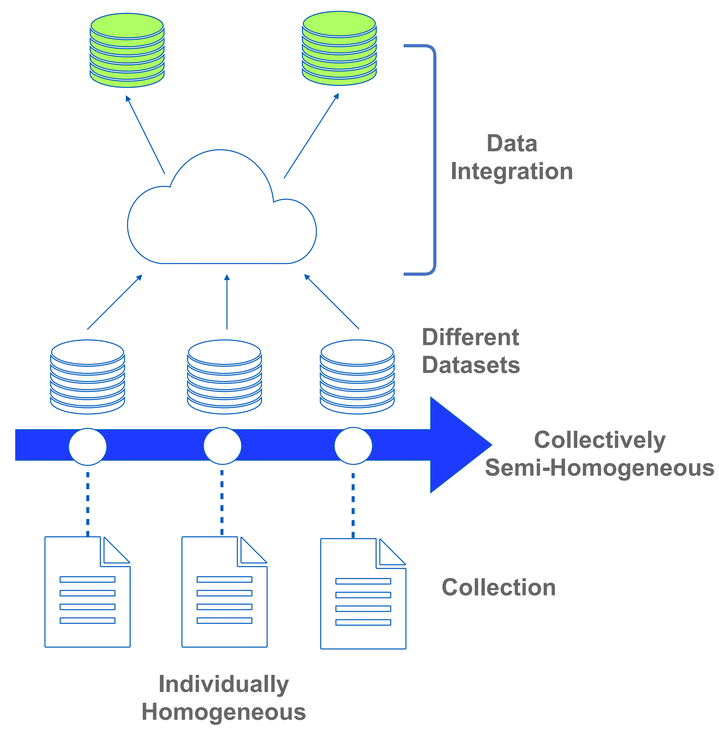
- Do you have children? **(Yes/ No)**

Form 2

Family:

- Do you have children?
- **If yes, How many?**

Need to Integrate Form Questions



Longitudinal Datasets vs Open Web Tables

Mountain	Height	Country
Everest	8848	NPL
Wetterhorn	3690	SWI
Kilimanjaro	5895	KEN

1. **Semantically Dense Attribute Identifiers**
2. **Inter Attribute Relationships**
   a. **Mountain *has* Height**
   b. **Mountain *is located in* Country**

col_01	col_02	col_03
Tom Cruise	1.70	USA
Brad Pitt	180.34	USA
Mark Viduka	187.96	AUS
	177.42	NPL

1. **Semantics from Entities- *Celebrities, Heights and Nationalities***
2. **Null Values have no semantics**

Family:

Q1. Do you have children?

Q2. How many?

Q3. What are their ages?

Q4.Do you have siblings

Q5. How many?

Semantics(Q5) = Q5 + Response to Q4

1. **Prose Questions; Ambiguous in Isolation**
2. **Frequent references to preceding questions.**

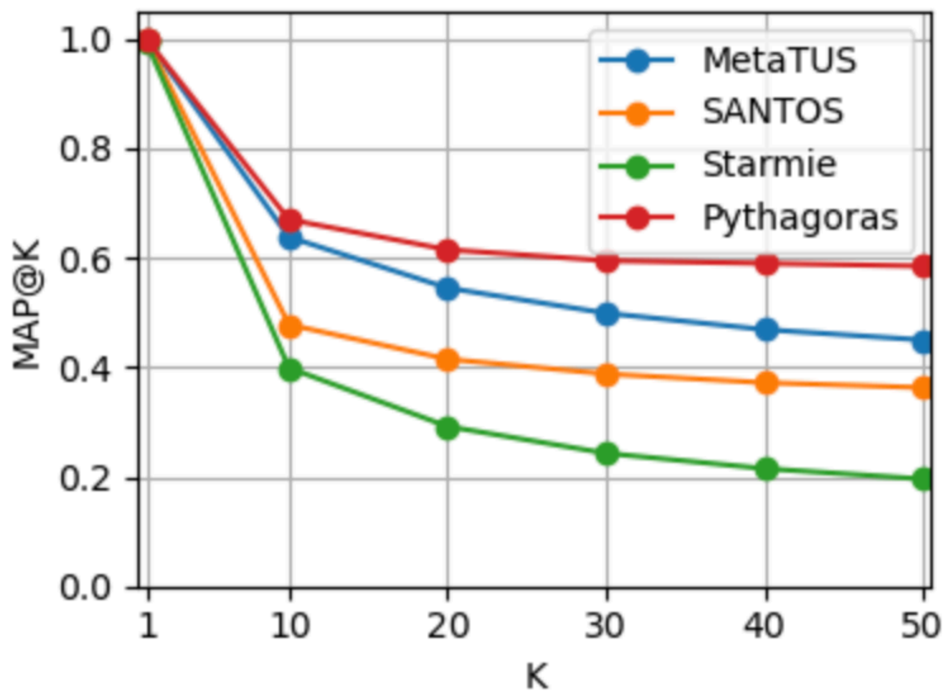
Q1. Have you ever donated to a campaign?	Q2. If yes, how much did you donate?($)
YES	100-200
NO	
DON'T KNOW	

1. **Semantics from questions, context, literals and measurements.**
2. **Null values may refer to skip semantics or conditionals**

Benchmark/Groundtruth for Longitudinal Dataset Integration

Span	72 years
Type	Labeled by SME
#Tables	2061
Size	1.32 GB
Labeled Pairs	10,320
Average Shape	2172 X 13
%Numeric Variables	15.97
%Non-Null	74.71
Column Name Overlap(10 % sample)	3.4%
Column Values Overlap (10 % non-numeric sample)	15.6%

Evaluation of SOTA Table Representation models on Table Union Search workload



Implications

Need for:

1. Longitudinal Schema aware Table-Representations.
2. Instrument-structure Modeling
3. Temporal and Conditional Modeling

References

1. Pratik Pokharel, Juseung Lee, Oliver Kennedy, Marianthi Markatou, Andrew Talal, Jeff Good, and Raktim Mukhopadhyay. Drag, drop, merge: A tool for streamlining integration of longitudinal survey instruments. HILDA, 2024.
2. Grace Fan, Jin Wang, Yuliang Li, Dan Zhang, and Ren.e Miller. 2023. Semantics-aware Dataset Discovery from Data Lakes with Contextualized Column-based Representation Learning. arXiv:2210.01922 [cs.DB] https://arxiv.org/abs/2210.01922
3. Aamod Khatiwada, Grace Fan, Roee Shraga, Zixuan Chen, Wolfgang Gatterbauer, Ren.e J. Miller, and Mirek Riedewald. 2023. SANTOS: Relationshipbased Semantic Table Union Search. Proc. ACM Manag. Data 1, 1, Article 9 (May 2023), 25 pages. doi:10.1145/3588689
4. Margherita Martorana, Tobias Kuhn, and Jacco van Ossenbruggen. 2025. Metadata-driven Table Union Search: Leveraging Semantics for Restricted Access Data Integration. arXiv:2502.20945 [cs.DB] https://arxiv.org/abs/2502.20945
5. Sven Langenecker, Christoph Sturm, Christian Schalles, and Carsten Binnig. Pythagoras: Semantic type detection of numerical data in enterprise data lakes. In Proceedings 27th International Conference on Extending Database Technology, EDBT 2024, Paestum, Italy, March 25 - March 28, pages 725–733. OpenProceedings.org, 2024.

University at Buffalo The State University of New York

Department of Computer Science and Engineering
School of Engineering and Applied Sciences
buffalo.edu