



# UTOPIA: Automatic Pivot Table Assistant

Whanhee Cho, Anna Fariha



## Messy spreadsheet data causes unusable pivot tables

Movie Title	Genre	Gross
2001: A Space Odyssey	Technology Fiction	100 M
Interstellar	Science Fiction, drama, Adventure	500 M
The Matrix	action, sci-fi	200 M

### Traditional pivot table tools struggle with messy data

Messy spreadsheet data is common [1]

Data **parsing** requires additional expertise

Resolving **synonymous variants** requires manual effort

Users don't prefer **switching environments**

🤔 Which movie genre achieves the most gross? Let's find out this through a Microsoft Excel pivot table with Genre and Gross.



Genre	Sum of Gross
Technology Fiction	100 M
Science Fiction, drama, Adventure	500 M
action, sci-fi	200 M

Multiple Values

No one wants to know the total gross of movies of **science fiction, drama, adventure** genre!  
- **Multiple values** in the Genre cause this problem.



Genre	Sum of Gross
Technology Fiction	100 M
Science Fiction	500 M
Sci-Fi	200 M
...	...

Synonymous Variants

I split the data with Microsoft Power Query. But, some values share **the same semantic meaning**.



Genre	Sum of Gross
Action	900 M
Sci-Fi	800 M
Drama	920 M
Adventure	600 M

I finally got the desired result, but I had to find and replace variants about **240 times**. It took me **3 hours**!

## UTOPIA provides ideal pivot tables without data preprocessing

The screenshot shows the UTOPIA interface. On the left, there's a 'Data Overview' table with columns: Title, Year, Genre, Gross, Director, Casts. Below it are 'Select Attributes' and 'Pivot Table Fields' sections. The 'Pivot Table' section shows a table with 'Sum of Gross' as the row label and 'Year' (2010, 2011, 2012) as the column labels. The 'Genre' column is expanded to show sub-groups like 'crime', 'action', 'superhero action', etc., with their respective gross values.

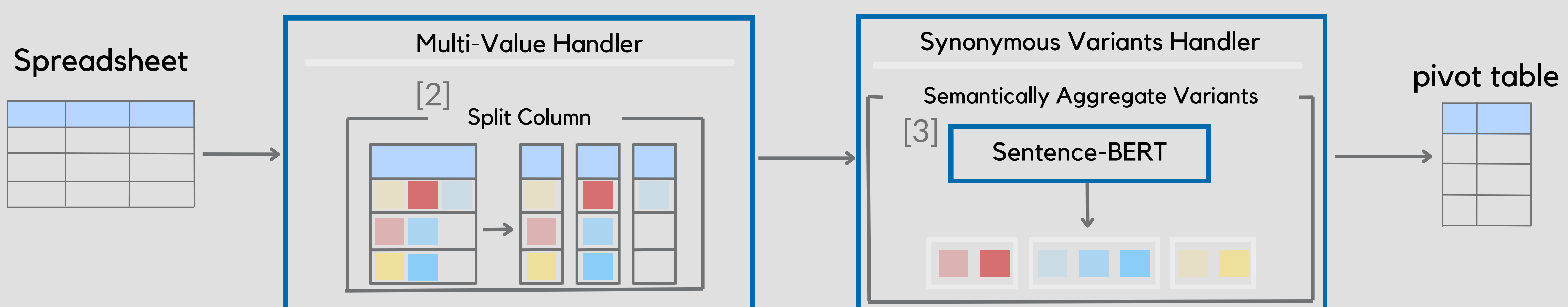
### Synonymous variants are grouped

UTOPIA **semantically aggregates** synonymous variants. Expand or collapse rows to view groups.

### Each cell contains a single value

UTOPIA **parses** multi-valued attributes even in the **absence of delimiters**.

## UTOPIA uses SOTA parsing technology and embeddings



## Future Work: Recommend k-budgeted pivot tables



What are the interesting pivot tables?

[1] P Li, Y He, C Yan, Y Wang, and S Chaudhuri. 2023. Auto-Tables: Synthesizing Multi-Step Transformations to Relationalize Tables without Using Examples. PVLDB

[2] M Raza and S Gulwani. 2017. Automated Data Extraction Using Predictive Program Synthesis. AAAI

[3] T Gao, X Yao, and D Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. EMNLP