IEEE
PACIFIC VIS 2023
SEOUL, KOREA

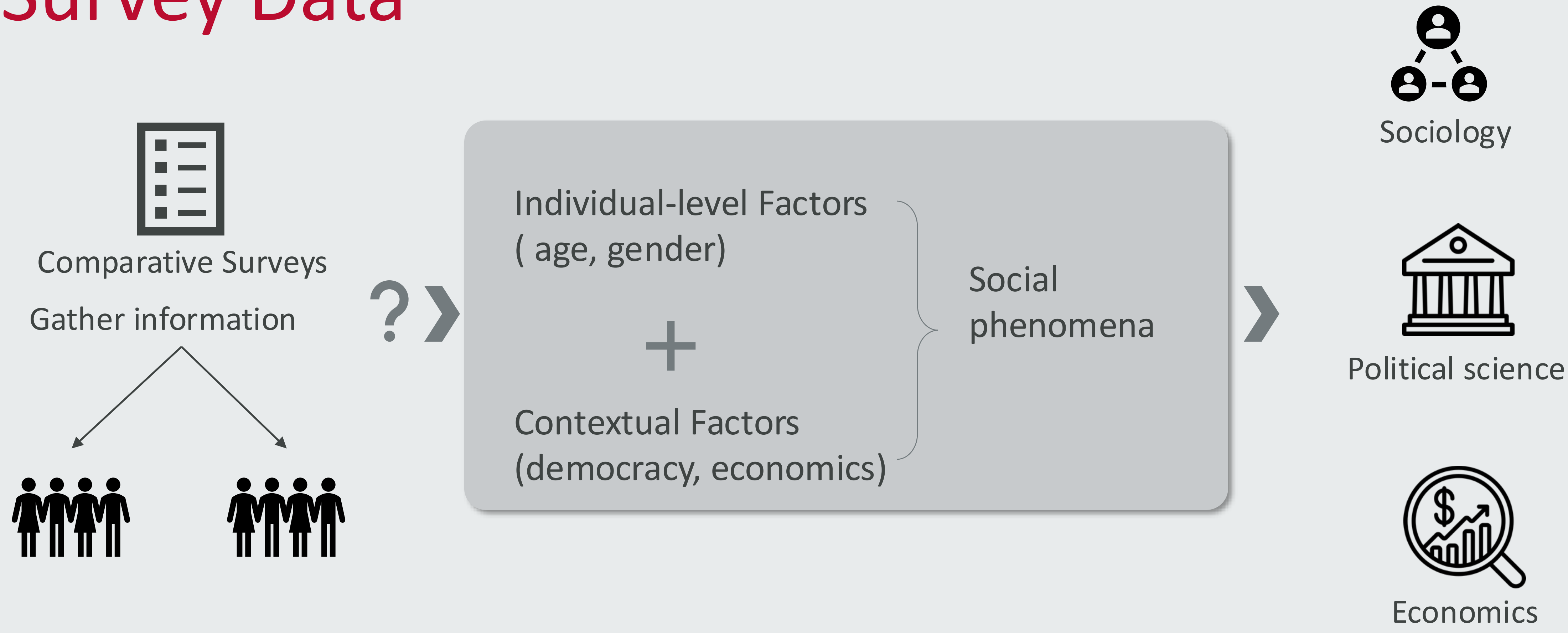
*SDR*Querier: A Visual Querying Framework for Cross-National Survey Data Recycling

Yamei Tu, Olga Li, Junpeng Wang, Han-Wei Shen, Przemek Powałko, Irina Tomescu-Dubrow, Kazimierz M. Slomczynski, Spyros Blanas,J. Craig Jenkins

THE OHIO STATE UNIVERSITY

IFiS PAN
Institute of Philosophy and Sociology, Polish Academy of Sciences (IFiS PAN)

Survey Data



Comparative Surveys

Gather information

?

Individual-level Factors (age, gender)

+

Contextual Factors (democracy, economics)

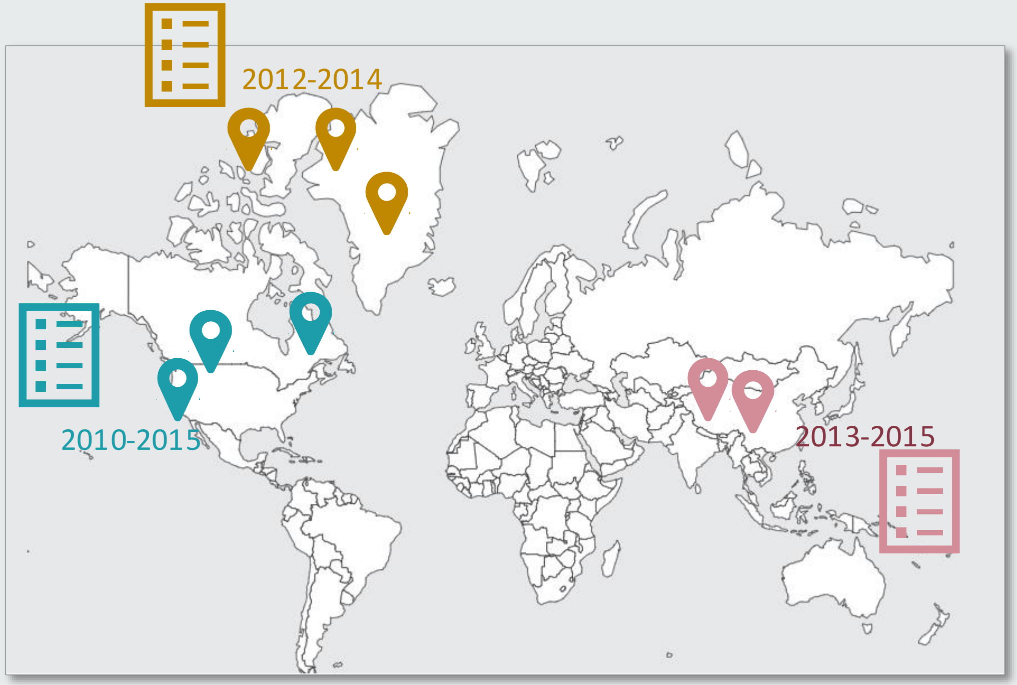
Social phenomena

Sociology

Political science

Economics

Survey Data Recycling



22 International Surveys

rescale

recode

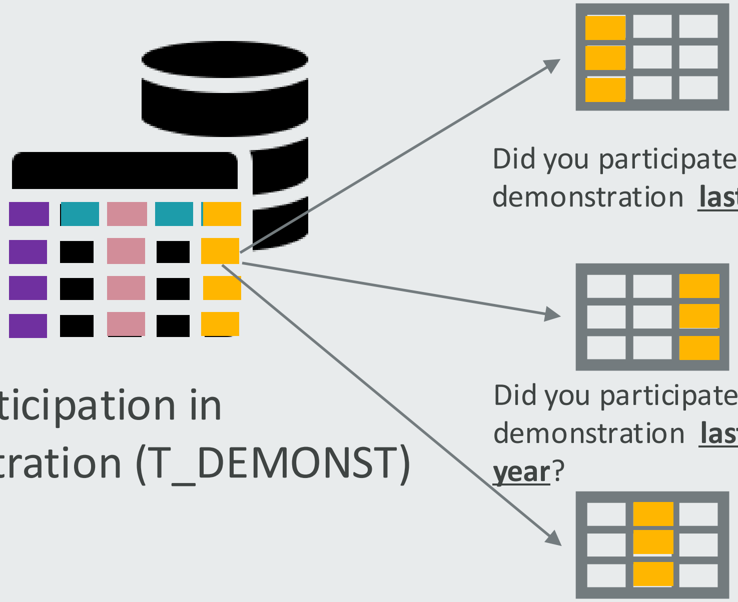
transform

Harmonized Dataset

Harmonized Data

Each row indicates one respondent’s responses to all questions.

Each column is called a **variable.**



Did you participate in demonstration **last year**?

Did you participate in demonstration **last five year**?

Did you participate in demonstration **ever**?

e.g., participation in demonstration (T_DEMONST)

Target Variable: represents a specific question from the original questionnaire.

Source Variable: original variables taken from different surveys for harmonization

Harmonization Controls: inter-survey methodological variability in the formulation of the source questions.

Quality Controls: biases and errors that stem from differences in the quality of the source survey data.

Utilizing Harmonized Dataset

Study "gender differences in political participation”?

Data Understanding

To decide whether there is relevant variables to scientists’ research interests?

Data Exploration

To decide whether there is sufficient data related to target variables.

Data Evaluation

Run the statistical models on the valid data to verify and generate insights.

Visual Querying Framework: SDRQuerier

Query-by-Question

Whether there are related variables?

- BERT-based model for variable recommendation from user question.
- Visualizing harmonized dataset structure.

Query-by-Condition

Whether variables have sufficient data to use?

Temporal Availability Profiler to display multi-faceted information

Query-by-Relation

Whether expected patterns exist in the data?

Presenting relational patterns.

Visual Querying Framework: SDRQuerier

Query-by-Question

Whether there are related variables?

- BERT-based model for variable recommendation from user question.
- Visualizing harmonized dataset structure.

Query-by-Condition

Whether variables have sufficient data to use?

Temporal Availability Profiler to display multi-faceted information

Query-by-Relation

Whether expected patterns exist in the data?

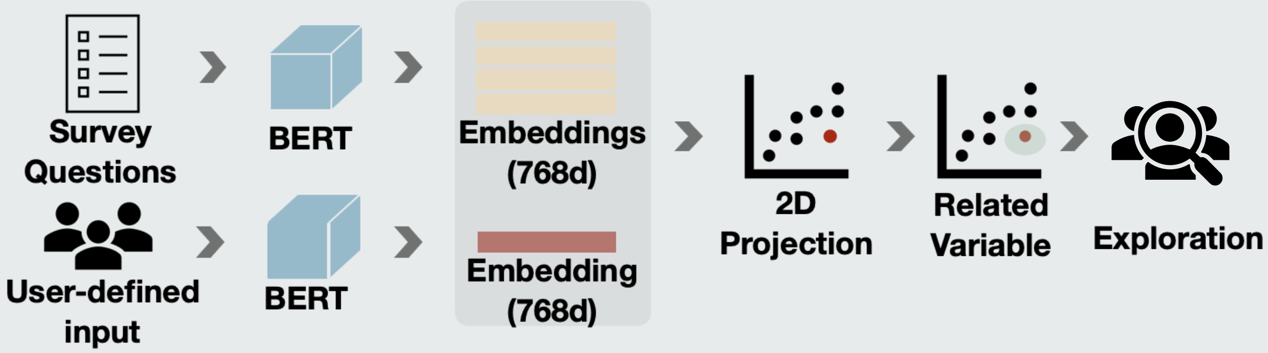
Presenting relational patterns.

Difficult to identify relevant variables?

- It can be difficult to identify the theoretical concept and their concrete meanings from the abbreviated names.
- It is hard for scientists to find related control variables efficiently due to the complex structure.
- Survey questions can provide good contextual meanings for each target variable.
- e.g., T_DEMONST (a target variable) can be described differently in the original questionnaires:
  - *authorized demonstrations in democratic countries*
  - *unauthorized activities in non-democratic countries*

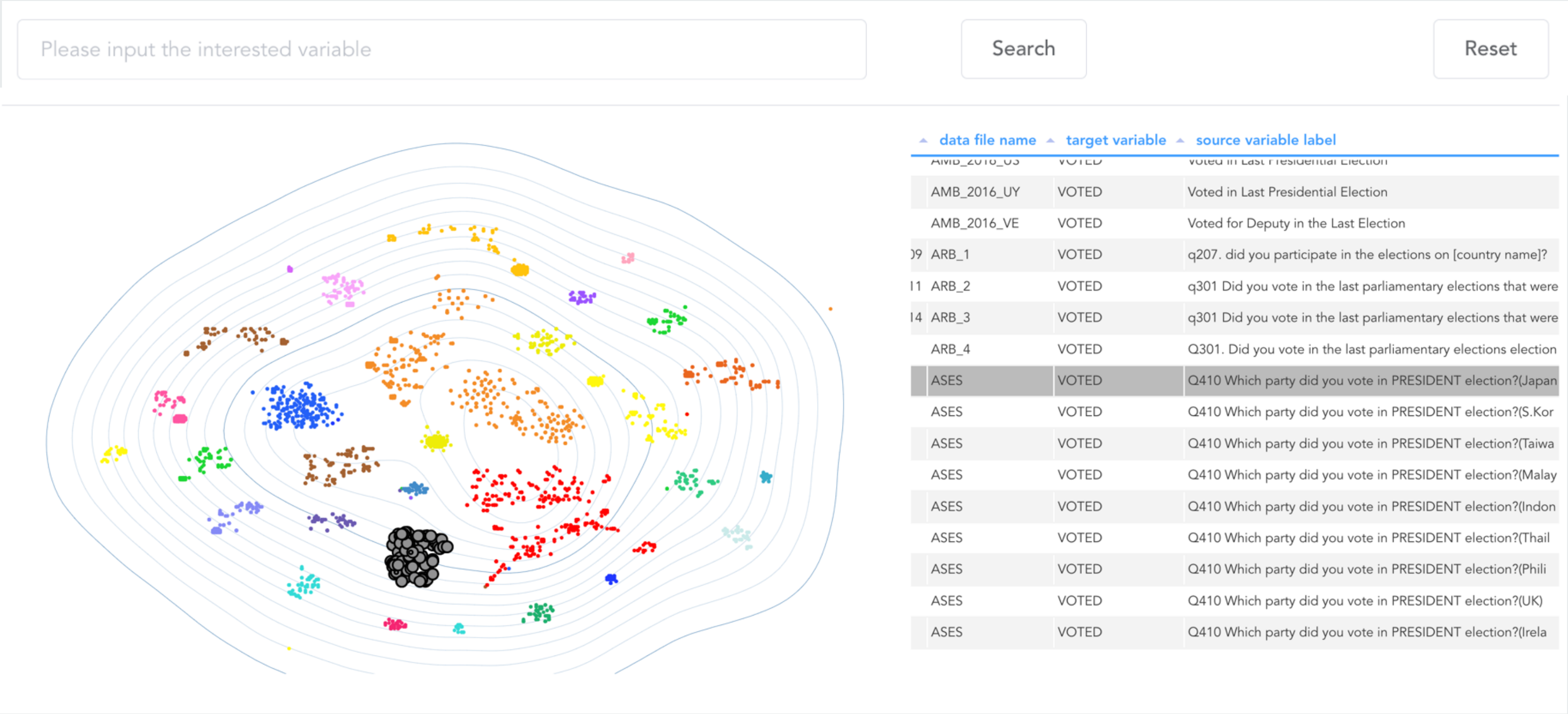
BERT-based Model

- Developed BERT-based classification model trained on survey questions for predicting target variable.
- Model enables two types of variable recommendation:
  - Hard: predict relevant target variables directly from user input.
  - Soft: generate embeddings of user input and compare to survey question embeddings for more flexible explorations.



Soft recommendation

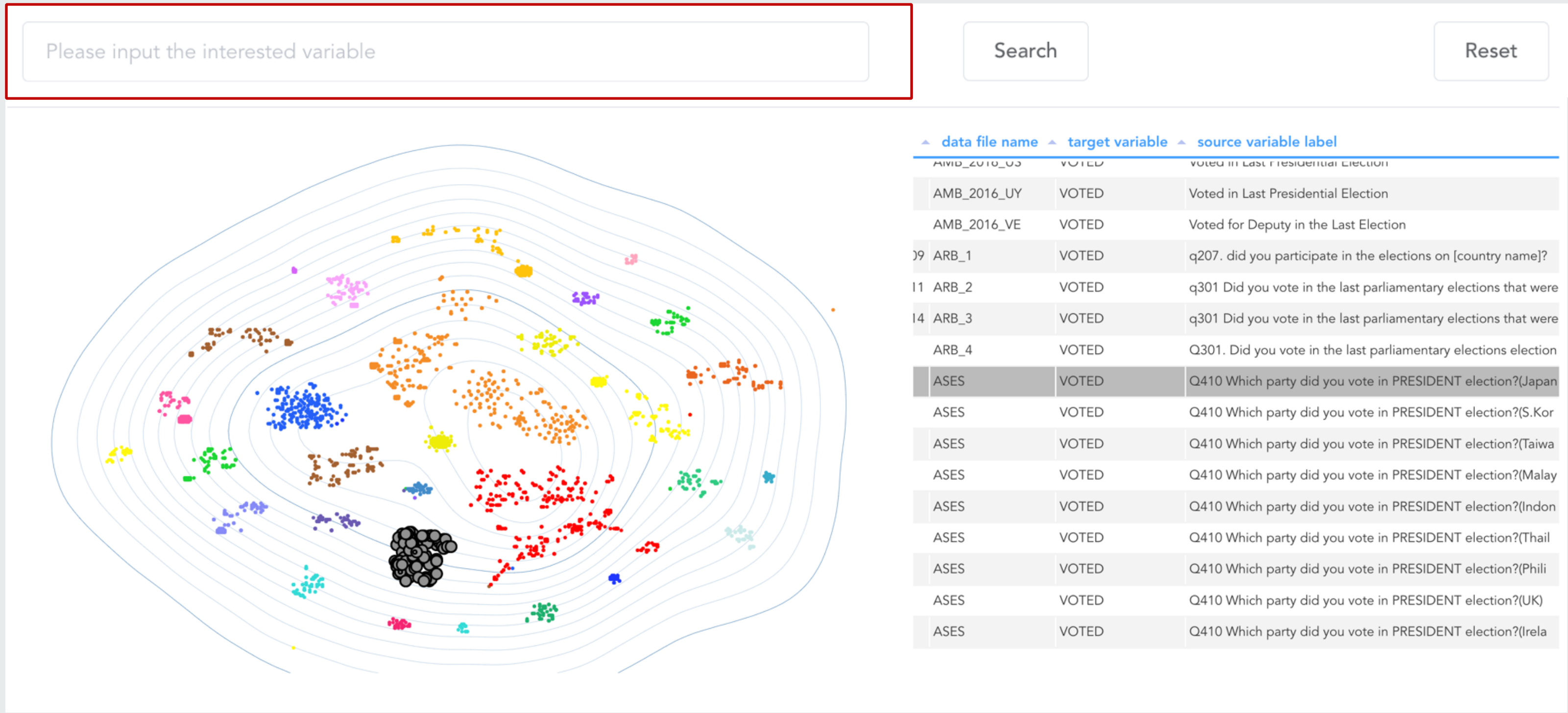
NLP Model + Visualization + Interactions



	data file name	target variable	source variable label
	AMB_2016_US	VOTED	Voted in Last Presidential Election
	AMB_2016_UY	VOTED	Voted in Last Presidential Election
	AMB_2016_VE	VOTED	Voted for Deputy in the Last Election
09	ARB_1	VOTED	q207. did you participate in the elections on [country name]?
11	ARB_2	VOTED	q301 Did you vote in the last parliamentary elections that were
14	ARB_3	VOTED	q301 Did you vote in the last parliamentary elections that were
	ARB_4	VOTED	Q301. Did you vote in the last parliamentary elections election
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(Japan
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(S.Kor
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(Taiwa
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(Malay
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(Indon
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(Thail
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(Phili
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(UK)
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(Irela

Soft recommendation

NLP Model + Visualization + Interactions

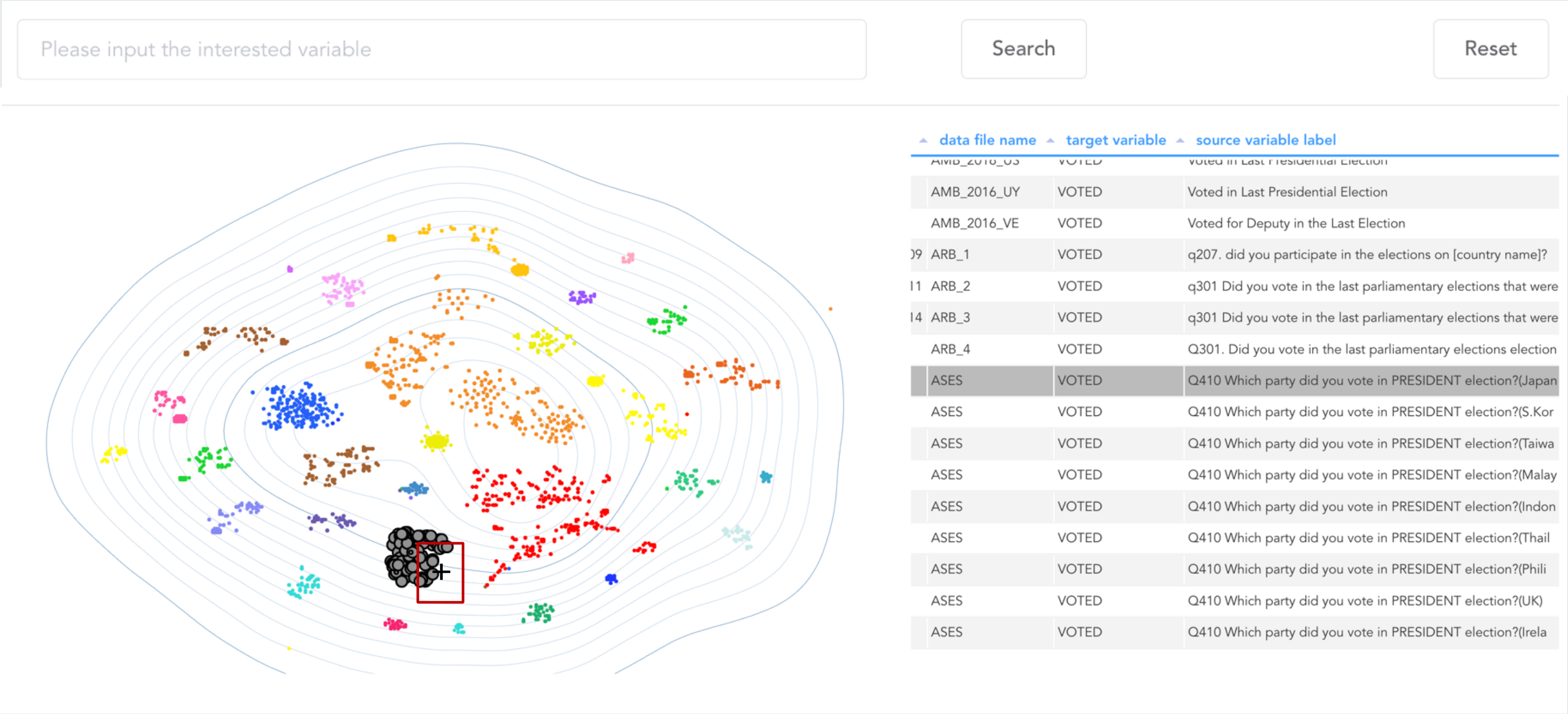


	data file name	target variable	source variable label
	AMB_2016_US	VOTED	Voted in Last Presidential Election
	AMB_2016_UY	VOTED	Voted in Last Presidential Election
	AMB_2016_VE	VOTED	Voted for Deputy in the Last Election
)9	ARB_1	VOTED	q207. did you participate in the elections on [country name]?
11	ARB_2	VOTED	q301 Did you vote in the last parliamentary elections that were
14	ARB_3	VOTED	q301 Did you vote in the last parliamentary elections that were
	ARB_4	VOTED	Q301. Did you vote in the last parliamentary elections election
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(Japan
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(S.Kor
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(Taiwa
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(Malay
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(Indon
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(Thail
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(Phili
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(UK)
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(Irela

- Input the questions /keyword /key-phrases to describe the research interests.

Soft recommendation

NLP Model + Visualization + Interactions

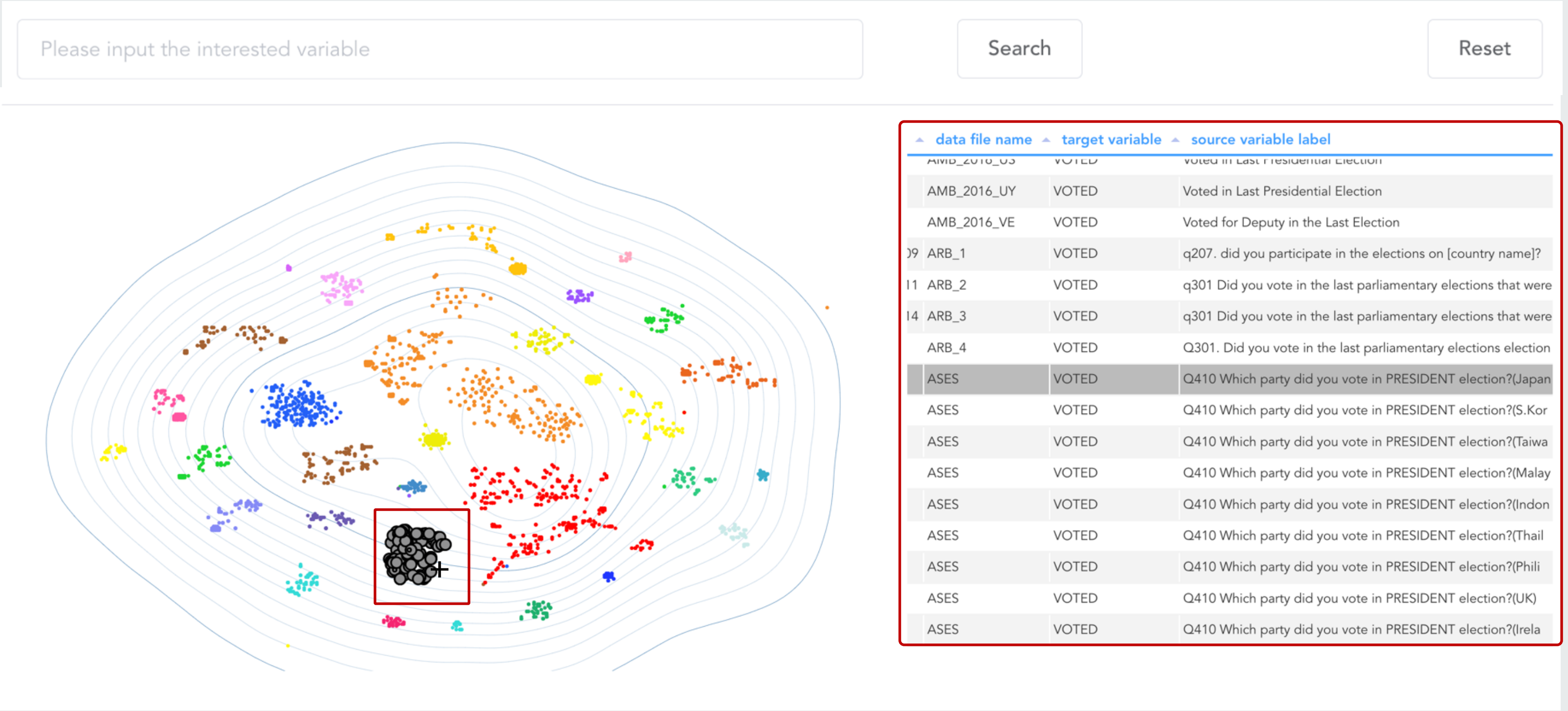


	data file name	target variable	source variable label
	AMB_2016_US	VOTED	Voted in Last Presidential Election
	AMB_2016_UY	VOTED	Voted in Last Presidential Election
	AMB_2016_VE	VOTED	Voted for Deputy in the Last Election
)9	ARB_1	VOTED	q207. did you participate in the elections on [country name]?
11	ARB_2	VOTED	q301 Did you vote in the last parliamentary elections that were
14	ARB_3	VOTED	q301 Did you vote in the last parliamentary elections that were
	ARB_4	VOTED	Q301. Did you vote in the last parliamentary elections election
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(Japan
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(S.Kor
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(Taiwa
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(Malay
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(Indon
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(Thail
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(Phili
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(UK)
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(Irela

- Input the questions /keyword /key-phrases to describe the research interests.
- Input embedding from BERT model is projected into the embedding space.

Soft recommendation

NLP Model + Visualization + Interactions

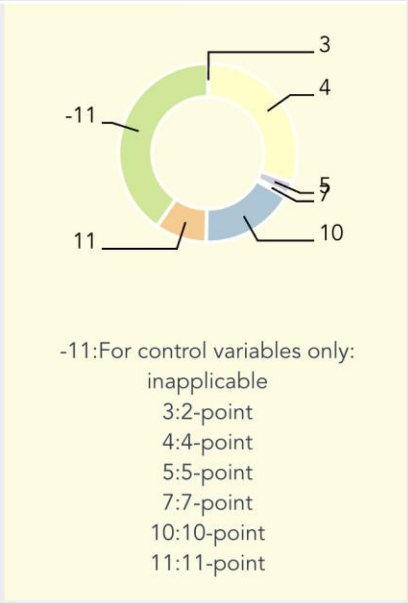


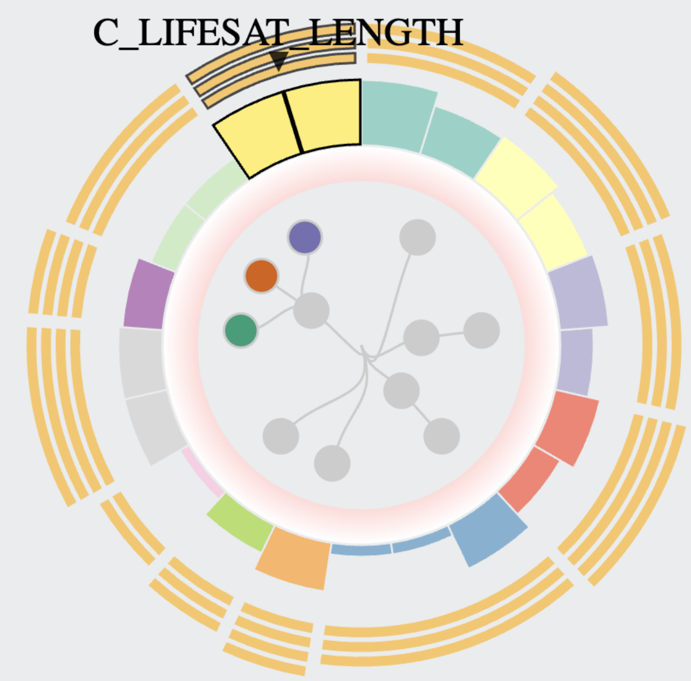
	data file name	target variable	source variable label
	AMB_2016_US	VOTED	Voted in Last Presidential Election
	AMB_2016_UY	VOTED	Voted in Last Presidential Election
	AMB_2016_VE	VOTED	Voted for Deputy in the Last Election
)9	ARB_1	VOTED	q207. did you participate in the elections on [country name]?
11	ARB_2	VOTED	q301 Did you vote in the last parliamentary elections that were
14	ARB_3	VOTED	q301 Did you vote in the last parliamentary elections that were
	ARB_4	VOTED	Q301. Did you vote in the last parliamentary elections election
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(Japan
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(S.Kor
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(Taiwa
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(Malay
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(Indon
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(Thail
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(Phili
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(UK)
	ASES	VOTED	Q410 Which party did you vote in PRESIDENT election?(Irela

- Input the questions /keyword /key-phrases to describe the research interests.
- Input embedding from BERT model is projected into the embedding space.
- Brushing related questions to understand variable's contextual meanings.

Visualizing Harmonized Data Structure

- Harmonized data's structure can be challenging for users to quickly comprehend.
- We design a Circular Graph to
  - Indicate different types of variables
  - Illustrate the relationships among variables.
- Hard recommendation will highlight the corresponding bars.



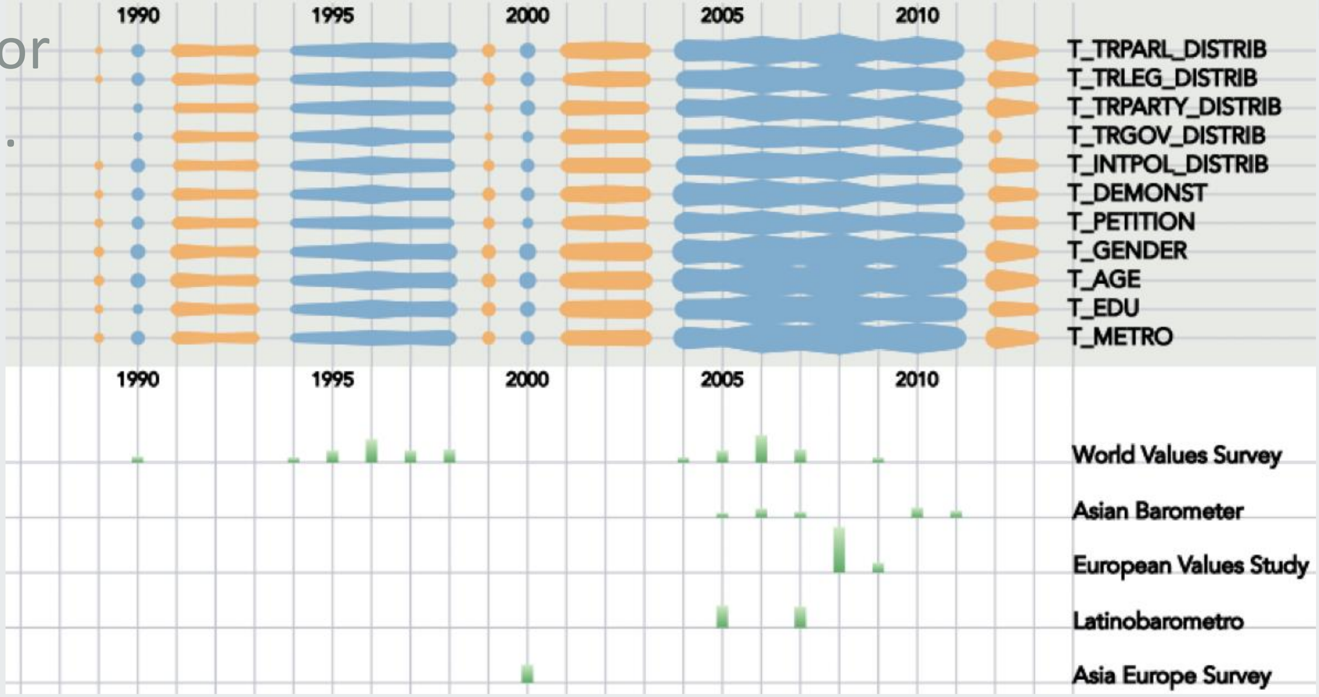


Inefficiency in Existing Data Portal

- Social scientists conduct their study based on a **set of target variables**
- Existing data portal:
  - Blindly download the full harmonized data
  - Ex-post check whether the set of target variables has sufficient data in terms of country and time coverage.
- We design a *Temporal Availability Profiler to* **effectively** and **efficiently** check data availability from multiple perspectives before downloading.

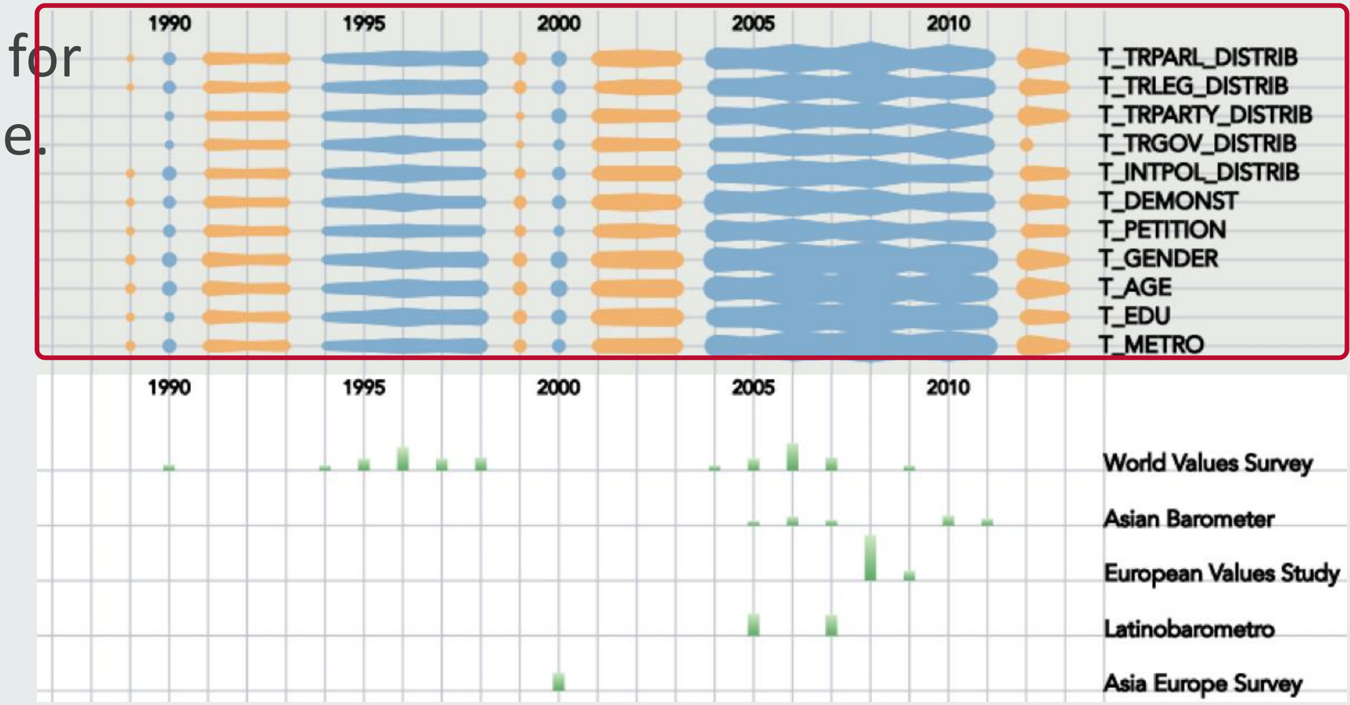
Temporal Availability Profiler

- It reveals the availability of the harmonized data at multiple levels
- Separate Availability View:
  - the number of valid samples for **each** target variable over time.
  - decide among multiple alternative variables.
- Joint Availability View:
  - available samples for **all** selected target variables.
  - precise pool of valid samples that can be used.



Temporal Availability Profiler

- It reveals the availability of the harmonized data at multiple levels
- Separate Availability View:
  - the number of valid samples for **each** target variable over time.
  - decide among multiple alternative variables.
- Joint Availability View:
  - available samples for **all** selected target variables.
  - precise pool of valid samples that can be used.

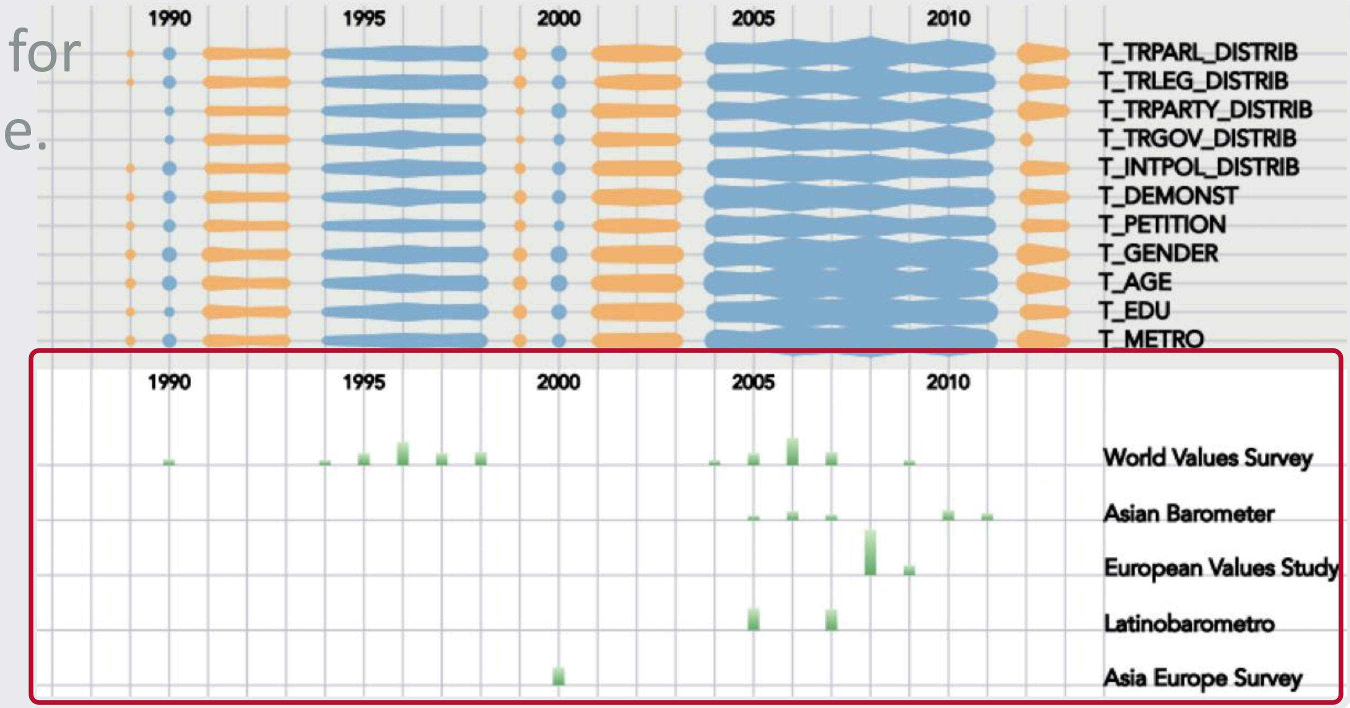


THE OHIO STATE UNIVERSITY
COLLEGE OF ENGINEERING

INSTYTUT FILOZOFII I SOCJOLOGII
POLSKIEJ AKADEMII NAUK

Temporal Availability Profiler

- It reveals the availability of the harmonized data at multiple levels
- Separate Availability View:
  - the number of valid samples for **each** target variable over time.
  - decide among multiple alternative variables.
- Joint Availability View:
  - available samples for **all** selected target variables.
  - precise pool of valid samples that can be used.



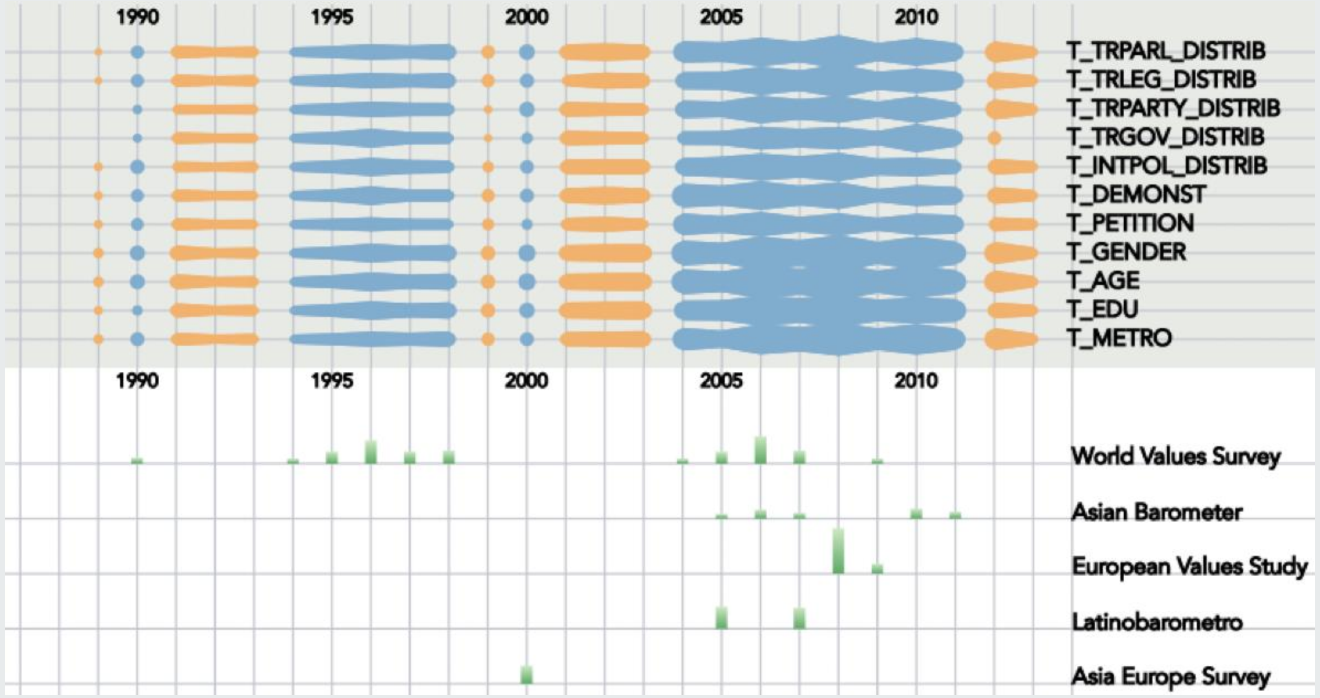
Temporal Availability Profiler

Three cases:

1. Each target variable has data, and there are also jointly available samples.

2. At least one target variable does not have data, resulting in the lack of jointly available data.

3. Each target variable has data, but there is no overlap among them.



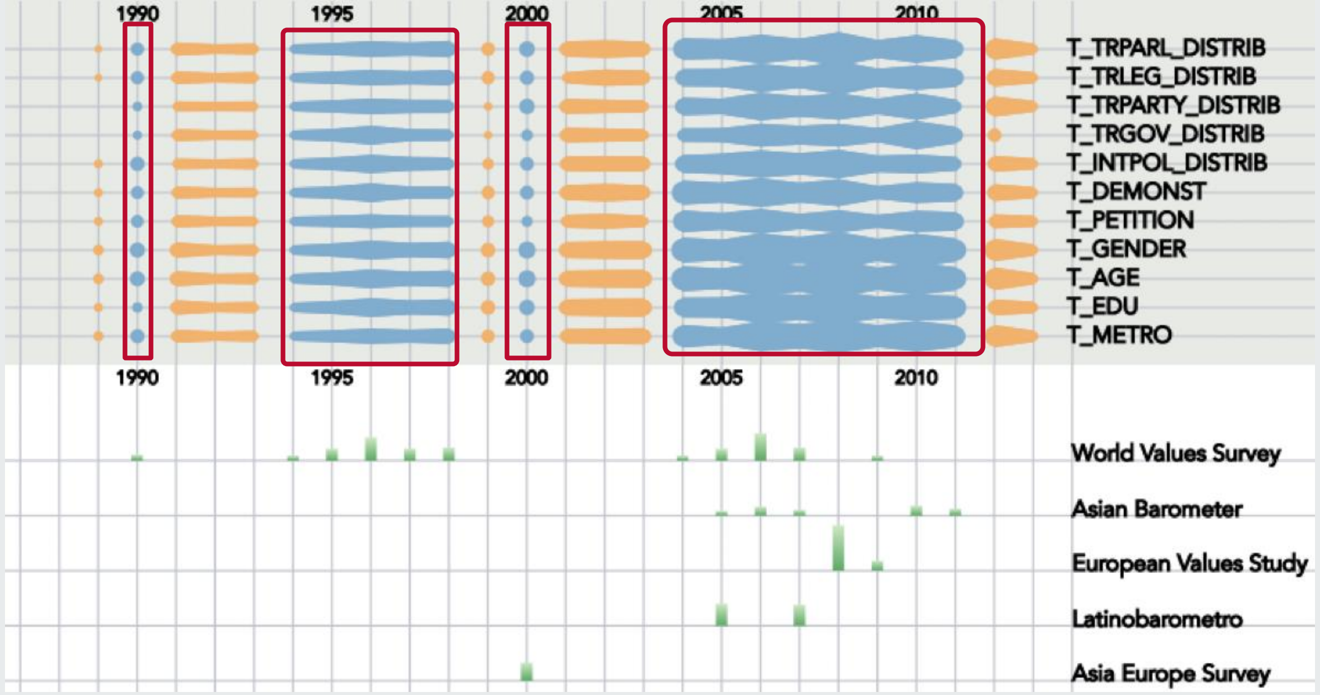
Temporal Availability Profiler

Three cases:

1. Each target variable has data, and there are also jointly available samples.

2. At least one target variable does not have data, resulting in the lack of jointly available data.

3. Each target variable has data, but there is no overlap among them.



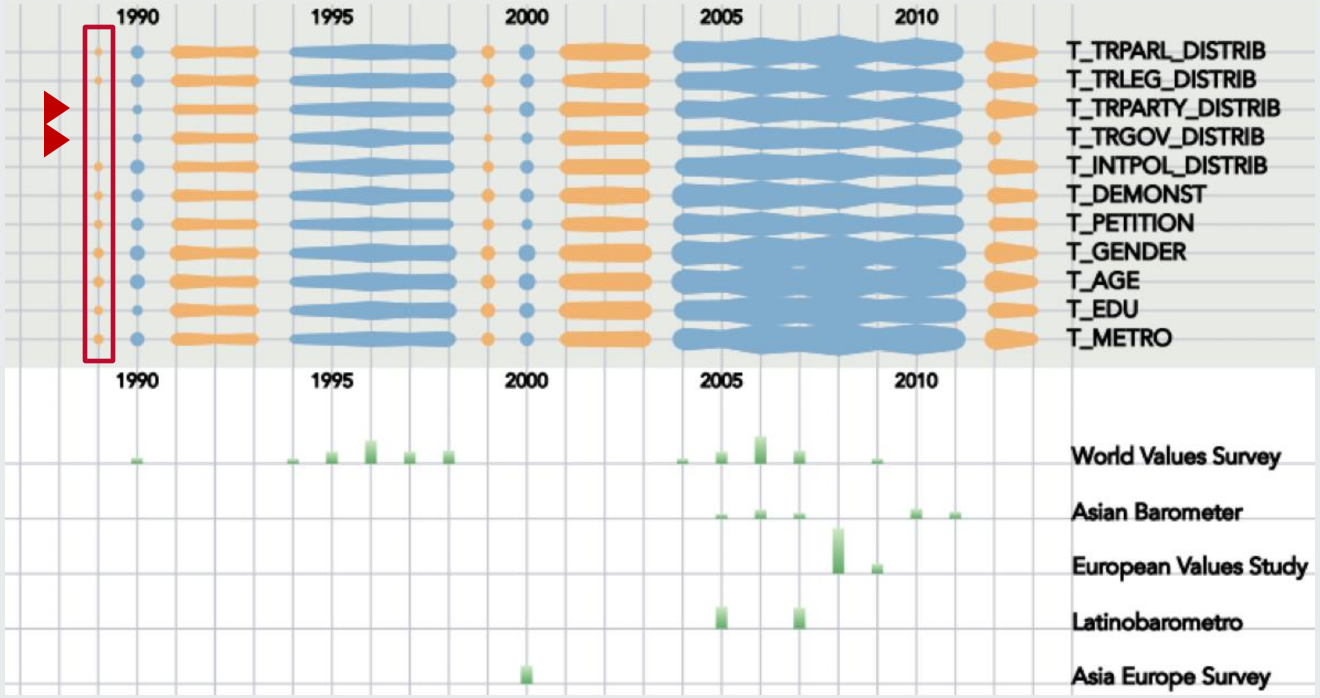
Temporal Availability Profiler

Three cases:

1. Each target variable has data, and there are also jointly available samples.

2. At least one target variable does not have data, resulting in the lack of jointly available data.

3. Each target variable has data, but there is no overlap among them.



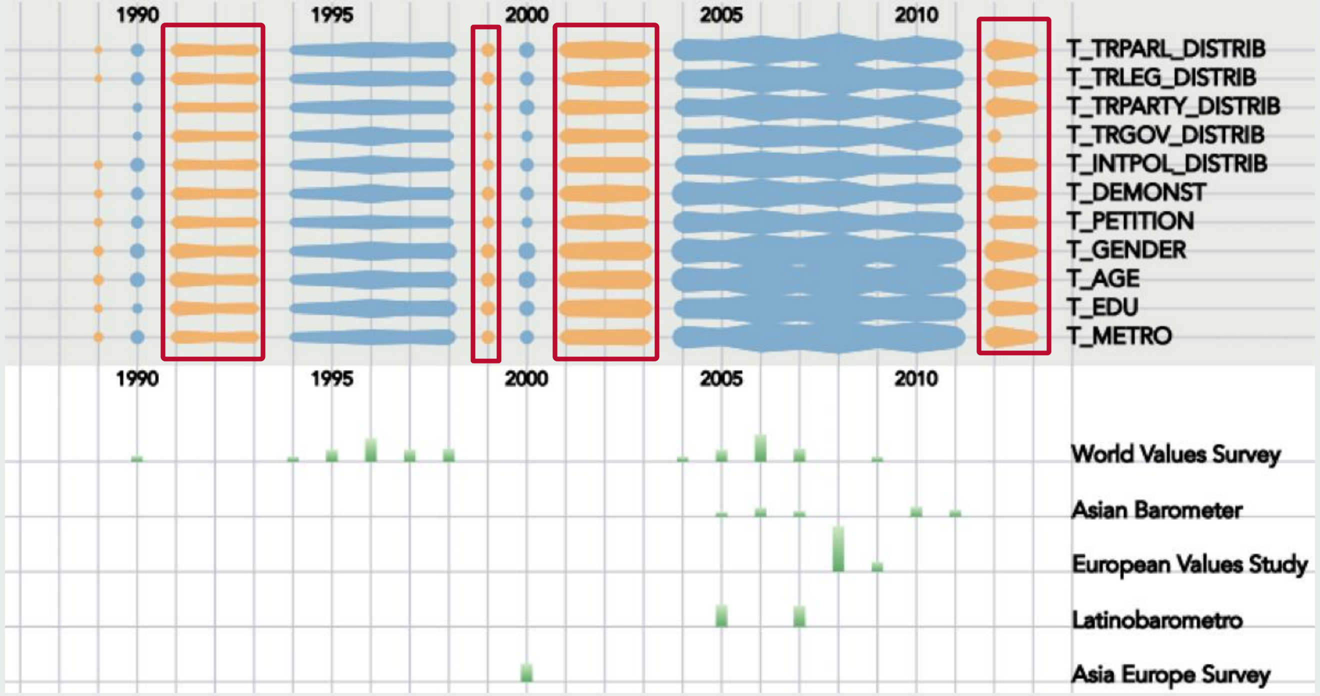
Temporal Availability Profiler

Three cases:

1. Each target variable has data, and there are also jointly available samples.

2. At least one target variable does not have data, resulting in the lack of jointly available data.

3. Each target variable has data, but there is no overlap among them.

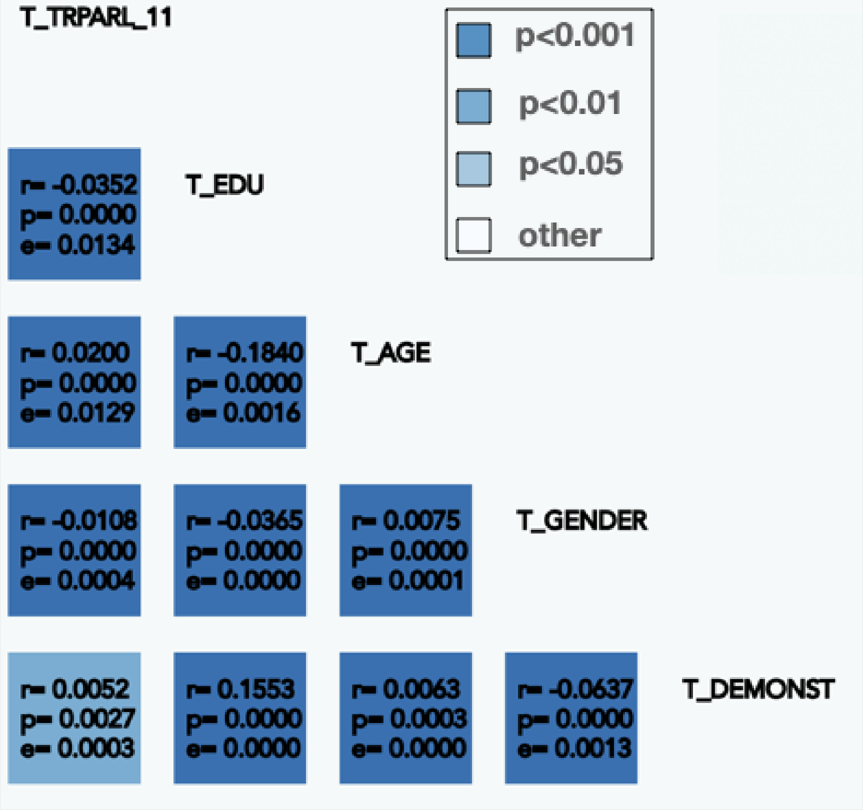


Query-by-Relation

- Social Scientists proposes ***theoretically derived hypotheses***, which are tested through ***statistical models*** using appropriate data.
- QBR is designed to query the hidden patterns from data for model verification and improvements.
- It answers two questions:
  - *what are the relationships between the selected target variables?*
  - *what are the potentially related variables?*

Correlation Matrix

- Showing pairwise correlations for user-selected target variables.
- Helps users to determine what to keep for their regression analysis.
- We compute:
  - Pearson correlation coefficient
  - p-values,
  - standard errors.



T_TRPARL_11				

r= -0.0352
p= 0.0000
e= 0.0134
	T_EDU			

r= 0.0200
p= 0.0000
e= 0.0129
	r= -0.1840
	p= 0.0000
	e= 0.0016
		T_AGE		

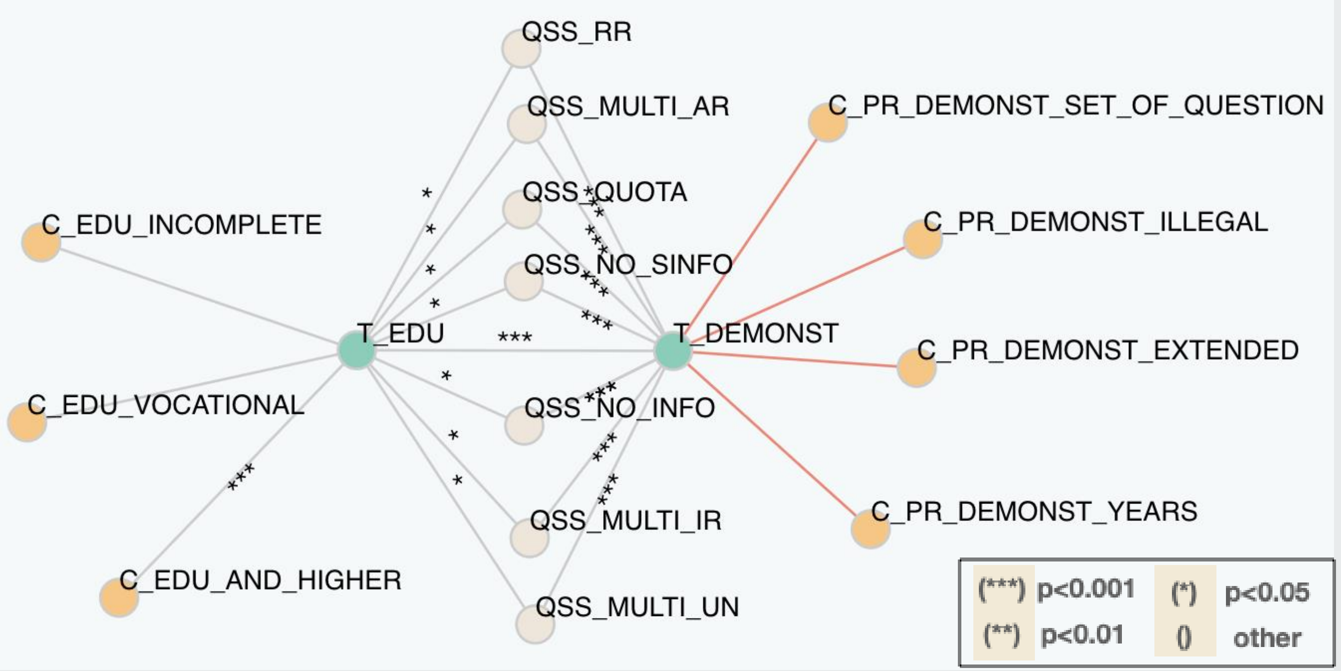
r= -0.0108
p= 0.0000
e= 0.0004
	r= -0.0365
	p= 0.0000
	e= 0.0000
		r= 0.0075
		p= 0.0000
		e= 0.0001
			T_GENDER	

r= 0.0052
p= 0.0027
e= 0.0003
	r= 0.1553
	p= 0.0000
	e= 0.0000
		r= 0.0063
		p= 0.0003
		e= 0.0000
			r= -0.0637
			p= 0.0000
			e= 0.0013
				T_DEMONST

🟦	p<0.001
🟦	p<0.01
🟦	p<0.05
□	other

Network Visualization

- Query complex relations given one pair of target variables.
- Both types of methodological variables can affect the relationship between substantive variables
  - Control variable
  - Quality variable

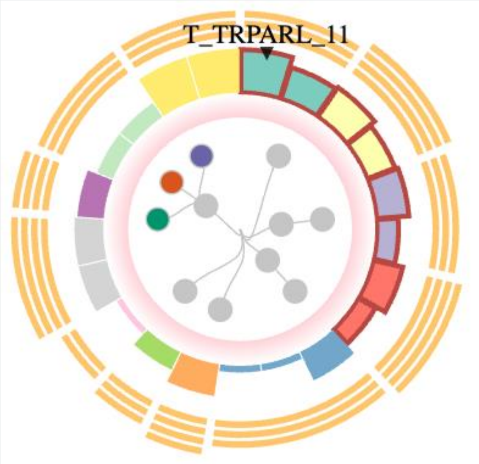


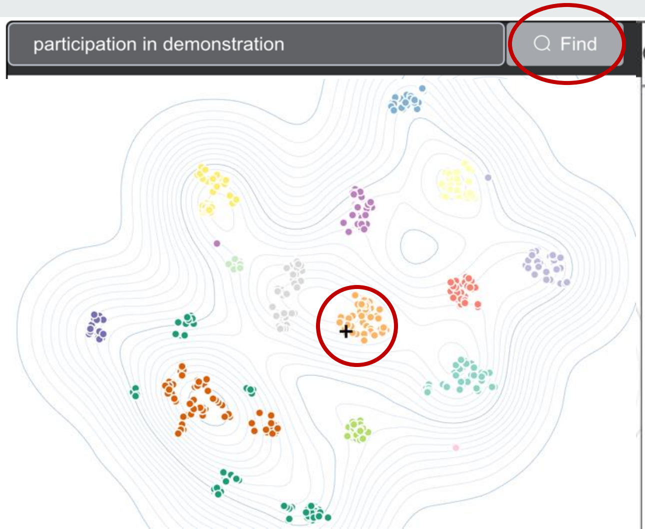
Case Study: Participation in Demonstrations worldwide

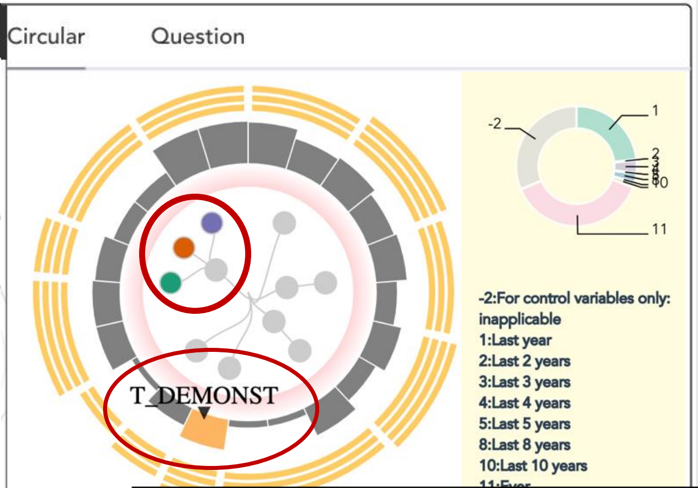
T_EDU, T_AGE, T_GENDER T_TRPARL_11

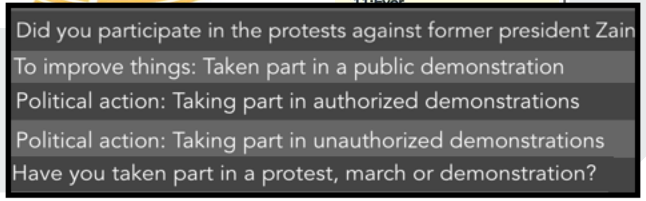
- Proposal: that resources and political attitudes have different effects on the levels of participation in demonstrations in democratic and non-democratic countries.

T_DEMONST



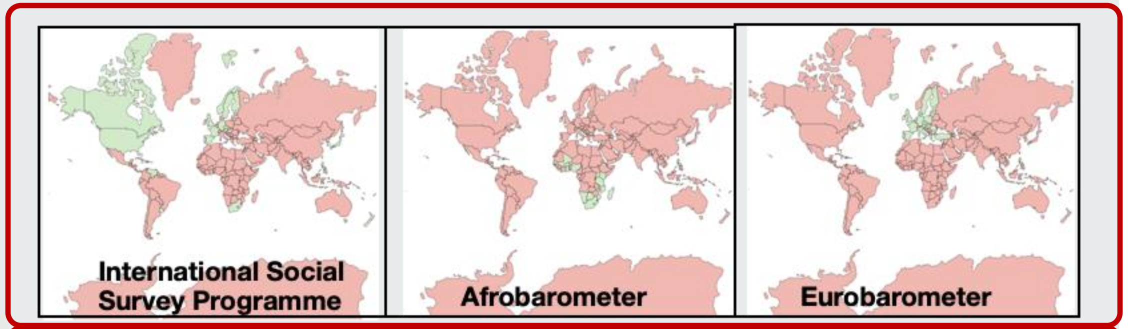


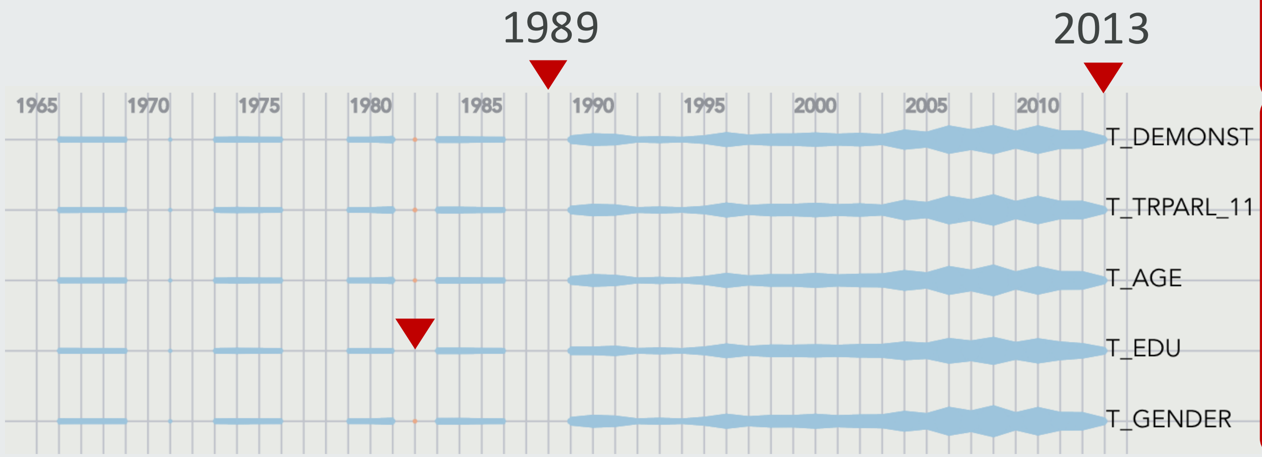


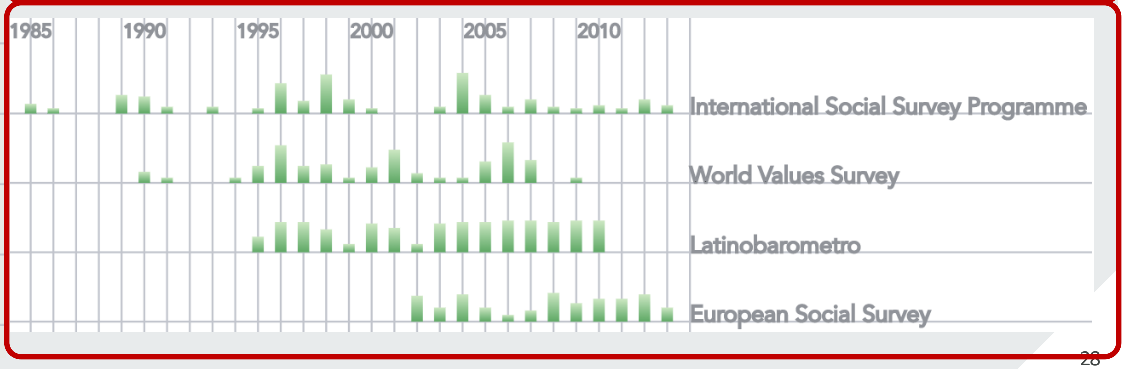

Did you participate in the protests against former president Zain
To improve things: Taken part in a public demonstration
Political action: Taking part in authorized demonstrations
Political action: Taking part in unauthorized demonstrations
Have you taken part in a protest, march or demonstration?

Case Study: Participation in Demonstrations worldwide

- High Temporal Coverage
- Harmonized data provides 22 available surveys to use
- Diverse country coverage allows comparison between democratic and non-democratic countries.







Conclusion

- We abstract the challenges in analyzing harmonization survey data
- We propose **a visual analytics system**, *SDR*Querier, to help scientists locate target variables and evaluate their theoretical models.

Acknowledgement

This work is supported by the NSF project 1738502, RIDIR: Survey Data Recycling: New Analytic Framework, Integrated Database, and Tools for Cross-national Social, Behavioral and Economic Research. The work is also partially supported by ICICLE project under grant number OAC-2112606.

Questions?