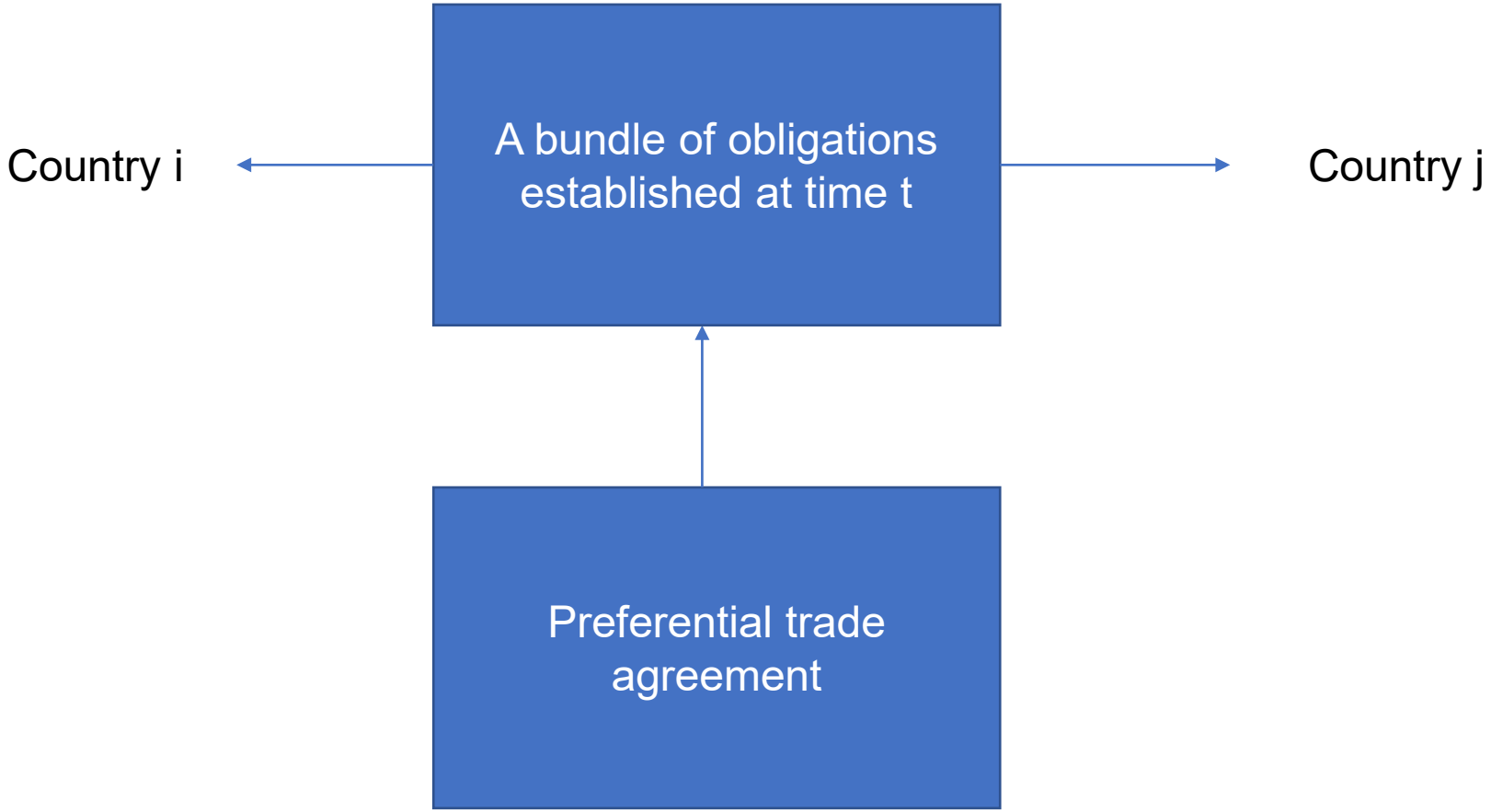
PTA obligation discontinuity design

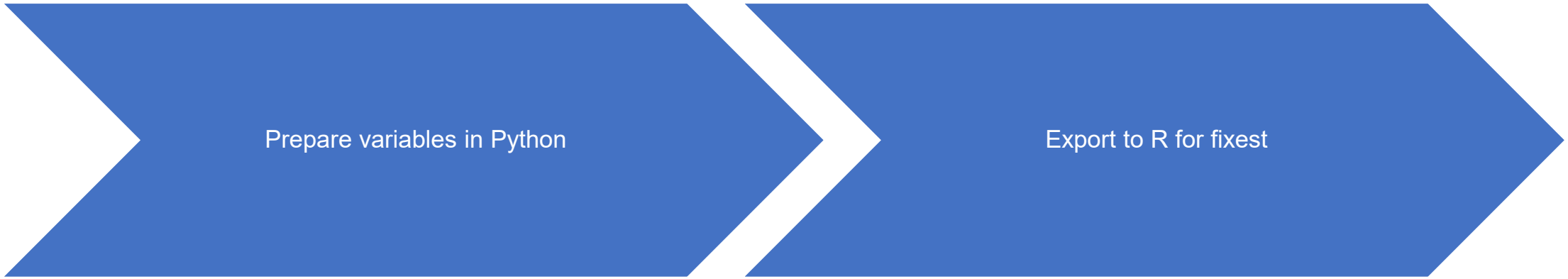
Documentation


Country i
A bundle of obligations established at time t
Country j
Preferential trade agreement

General preprocessing pipeline

Script (for inference):	*distilbert supervised text classification (after manual wrangling)* standardizeChapters_infer.py	*distilbert supervised text classification* isSubstantive_infer.py	*spaCy custom NER model* rOb_infer.py	*distilbert supervised text classification* isMandatory_infer.py	*bertopic topic classification > flan-t5-base text2text topic generation* modelTopics.py
Script (for training):	standardizeChapters_train.py	isSubstantive_train.py	rOb_train.py (cannot just run python script, has some command line steps, specified in the .py file)	isMandatory_train.py	based on pre-trained Google flan-t5-base; you can fine-tune to make better manual labels
How to do more labeling?	chapter_reassignment.csv chapter_dict.csv	isSubstantive_traindata.csv	More involved: label in label-studio, output json as input to rOb_train.py	isMandatory_train.csv	this step is new because what we previously did was cluster negotiator-assigned article titles and manually reclassified

Analysis pipeline



Script:

variablePrep.py

R_hdfe.R

analyze_exportdf.csv

is the input file you can immediately put into R fixest

Where are files?

- Scripts are in scripts_new
- Base excel/csv files which can be used for training or as temporary vessels between each inference step are in other_source
- Transformer models which can be used for preprocessing inferences are in trfModels
- Note that the latter two have to be downloaded separately because they are lage files