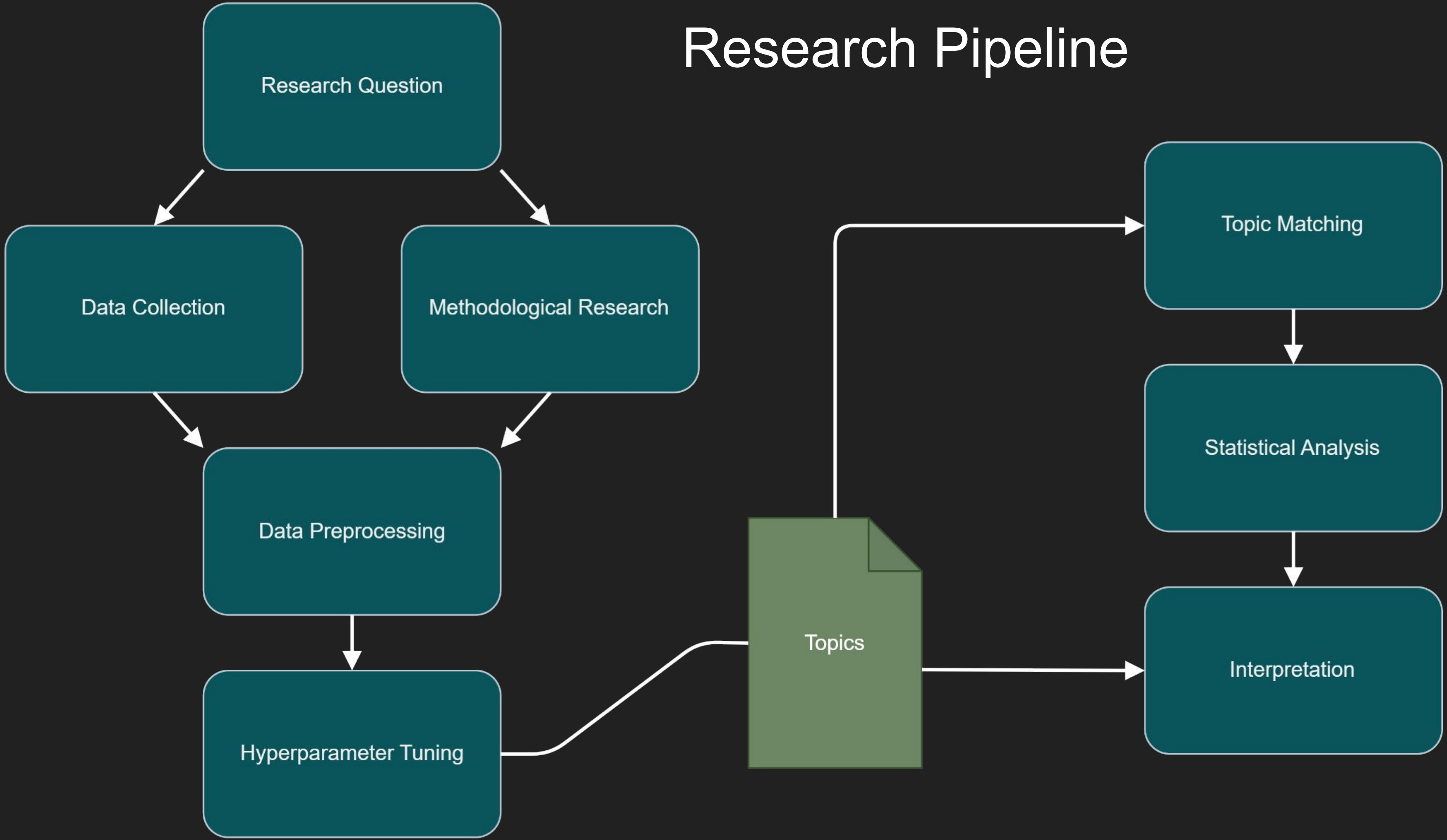
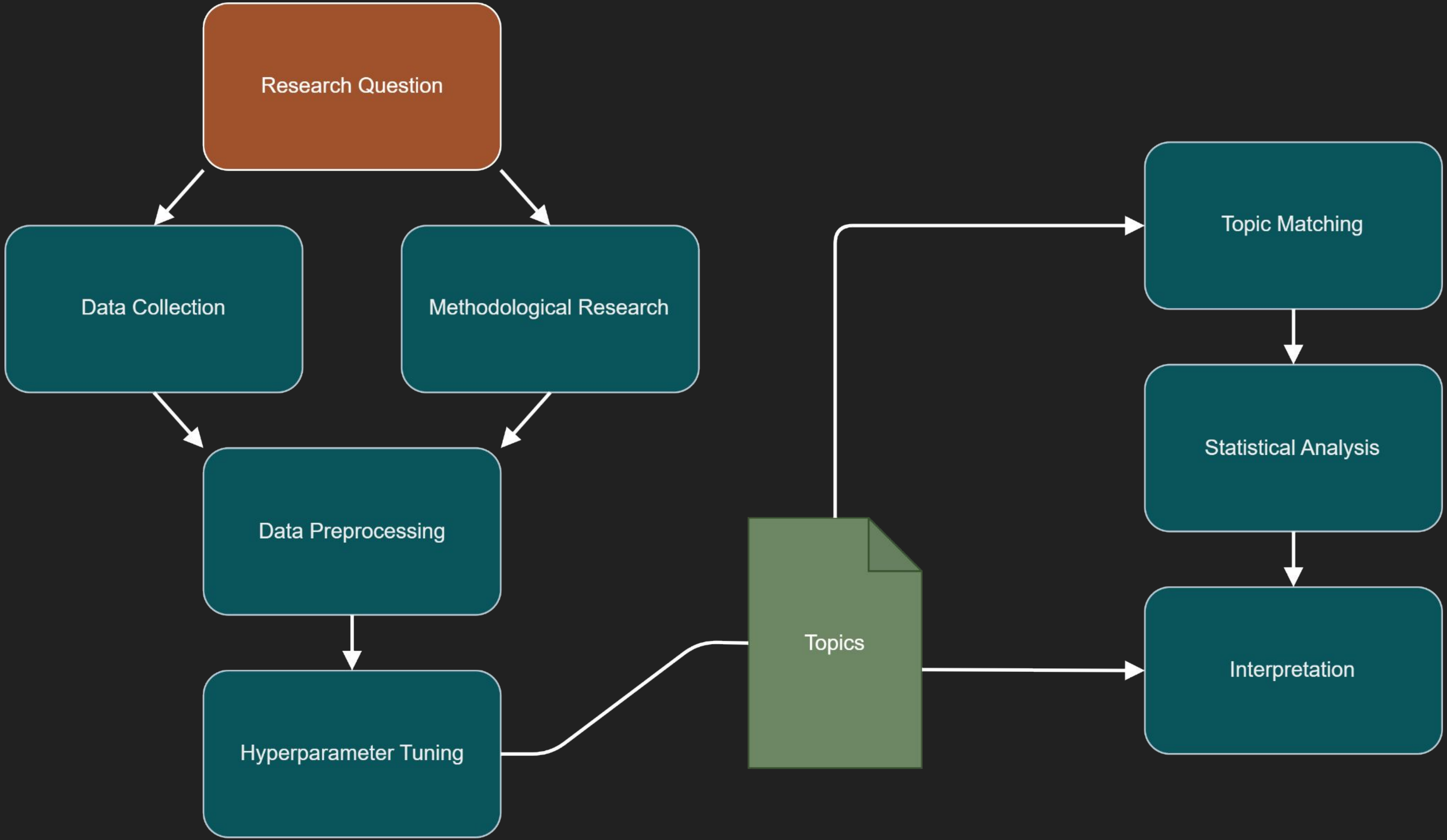
Political debates in Parliament and on Twitter

Source

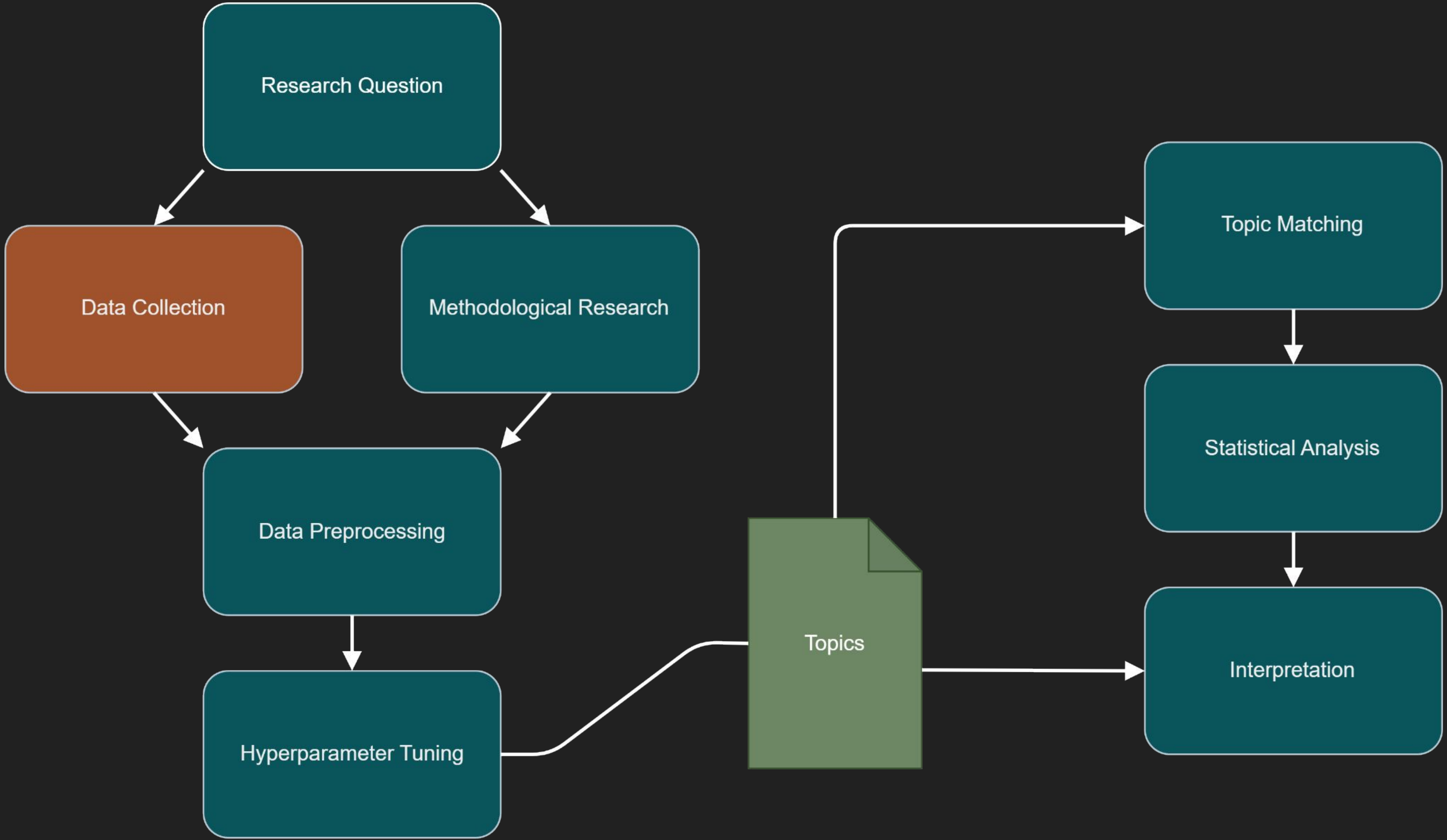
Research Pipeline




Research Question
Data Collection
Methodological Research
Data Preprocessing
Hyperparameter Tuning
Topics
Topic Matching
Statistical Analysis
Interpretation

Research Question

Do the debated topics on Twitter by German politicians differ from the topics debated in the German parliament?

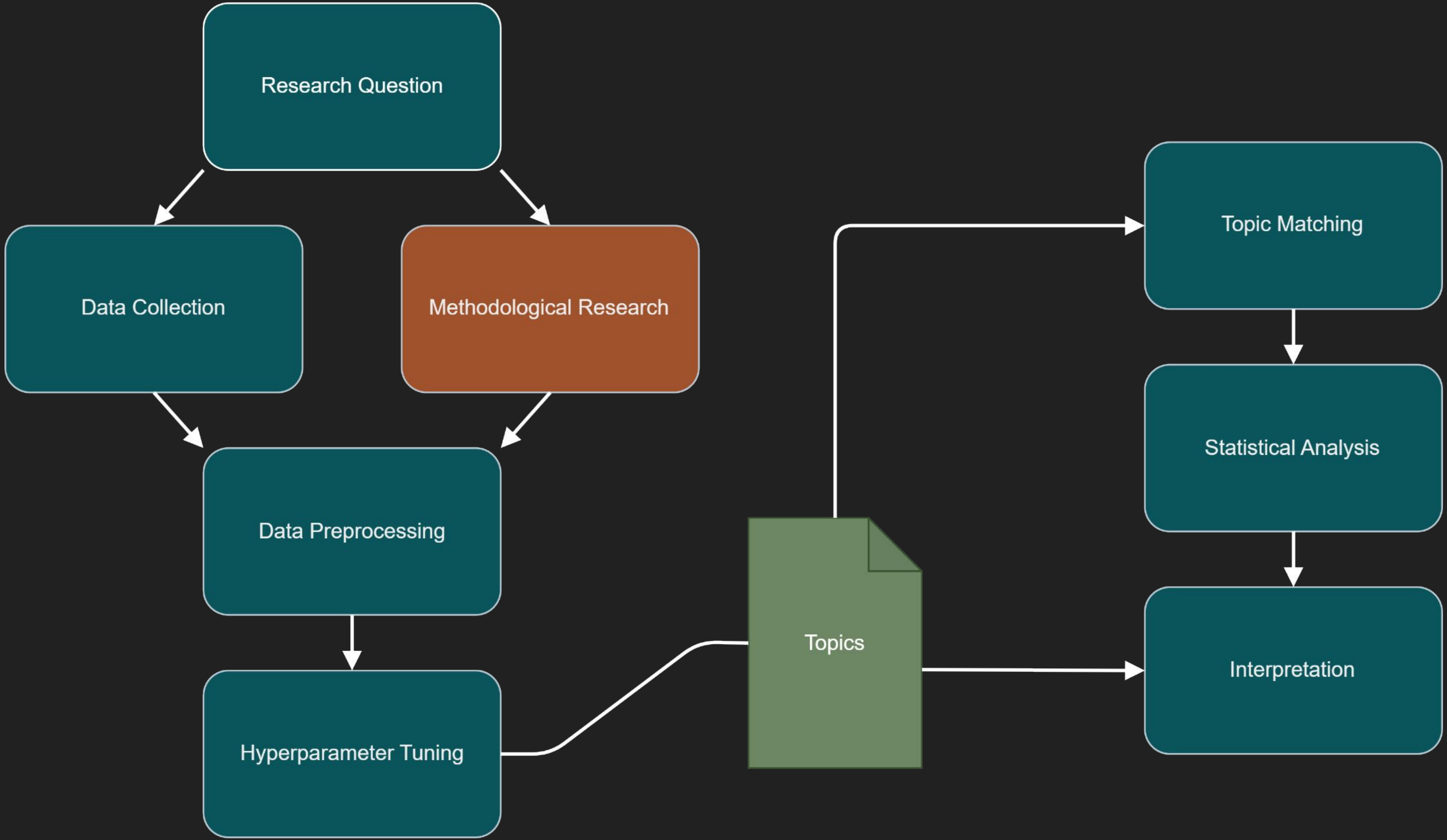

Research Question
Data Collection
Methodological Research
Data Preprocessing
Hyperparameter Tuning
Topics
Topic Matching
Statistical Analysis
Interpretation

Data Collection

- Data from 24.10.2017 to 26.10.2021 (19th legislative period)
- Plenary Protocols of German Parliament (Open Data Portal)
  - Used available parser
- Tweets of German MPs (from Lasser et al., 2022)

Thank you, Jana!

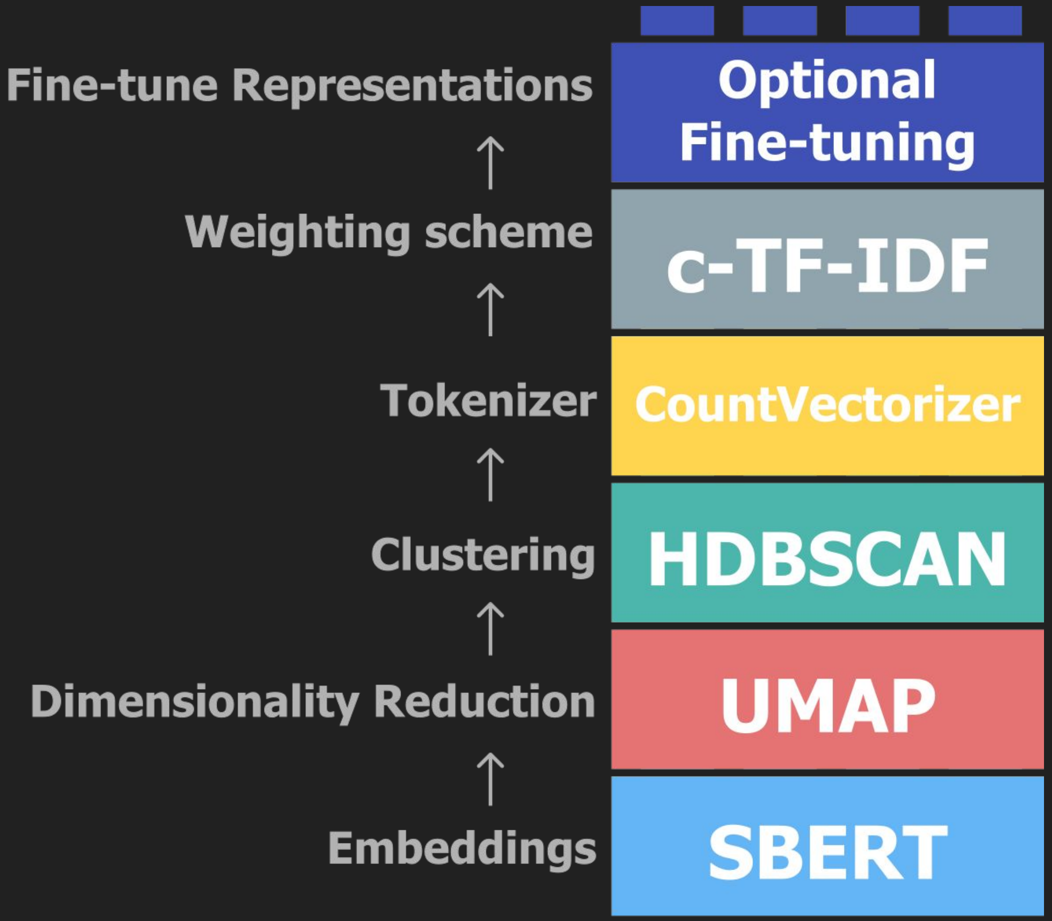
https://github.com/pournaki/bundestag-parser https://www.bundestag.de/services/opendata


Research Question
Data Collection
Methodological Research
Data Preprocessing
Hyperparameter Tuning
Topics
Topic Matching
Statistical Analysis
Interpretation

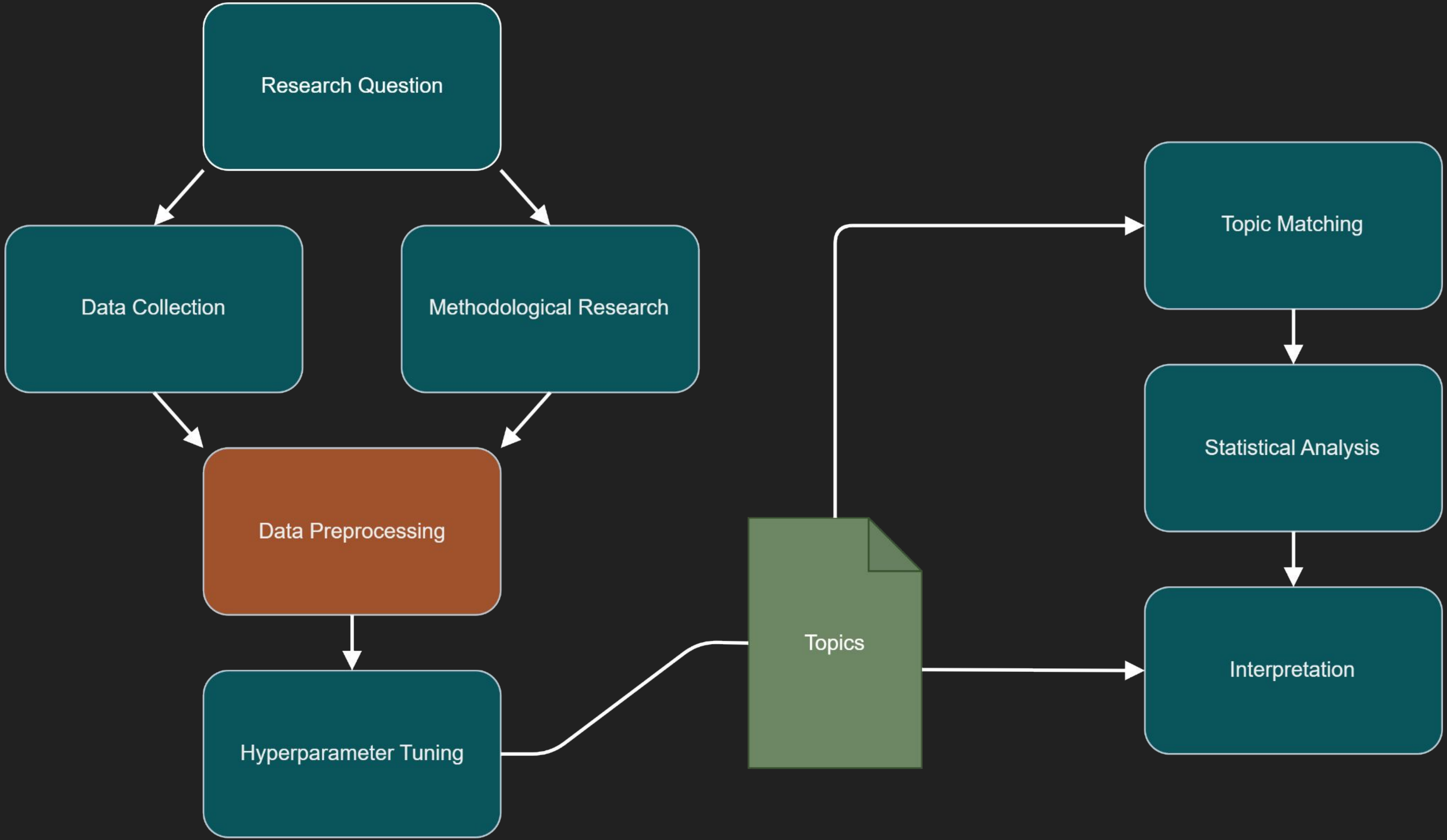
Related Work

1. Parliament data with topic modelling
    a. LDA: Curran et al. (2018)
    b. BERTopic: Contreras et al. (2022)
2. Twitter data
    a. Twitter-LDA: Zhao et al. (2011)
    b. BERTopic: Grootendorst (2022), Lasser et al. (2023)
3. Topic comparison
    a. Climate Change: Schaefer at al. (2023)

BERTopic

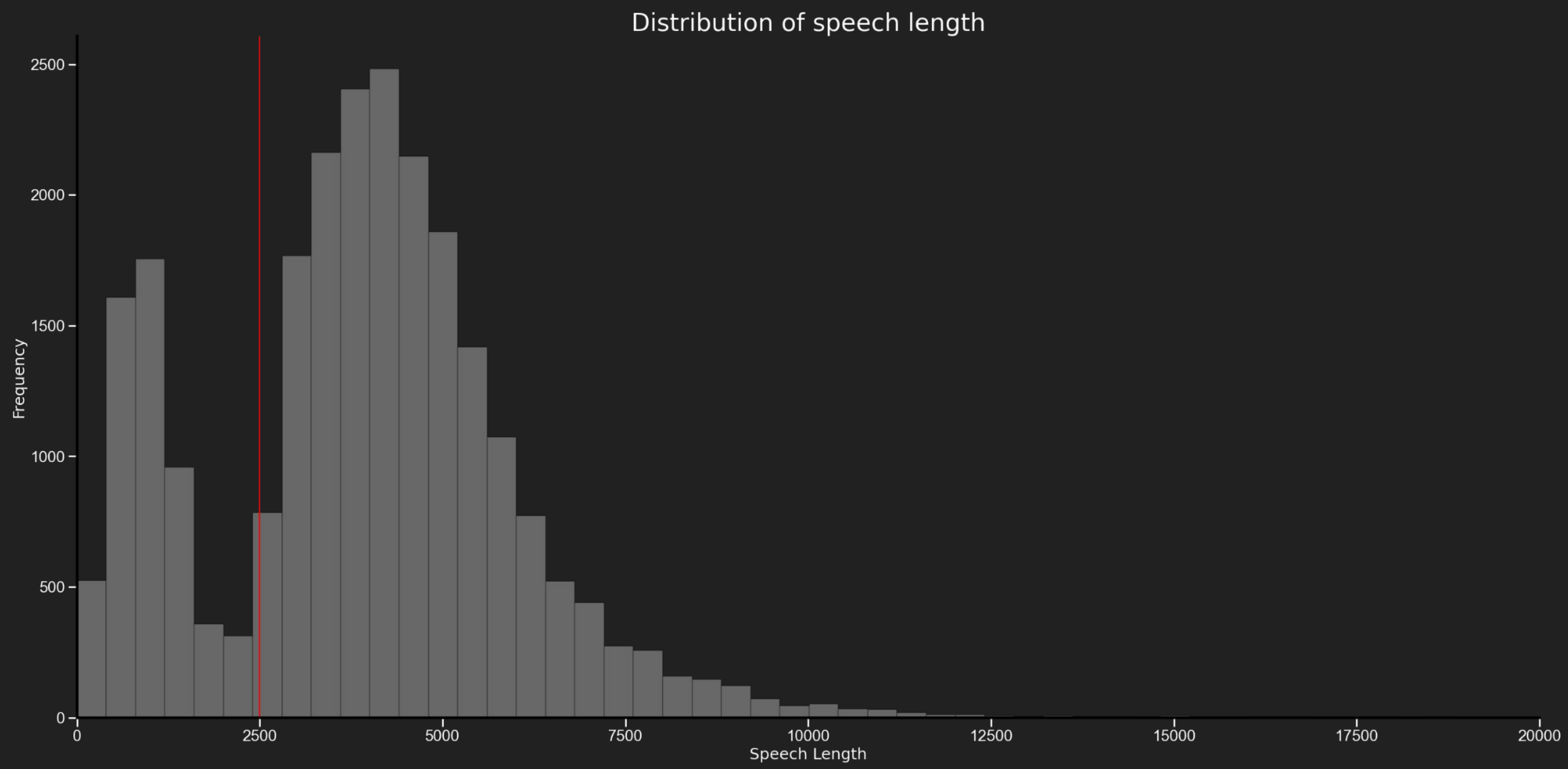


https://maartengr.github.io/BERTopic/algorithm/algorithm.html

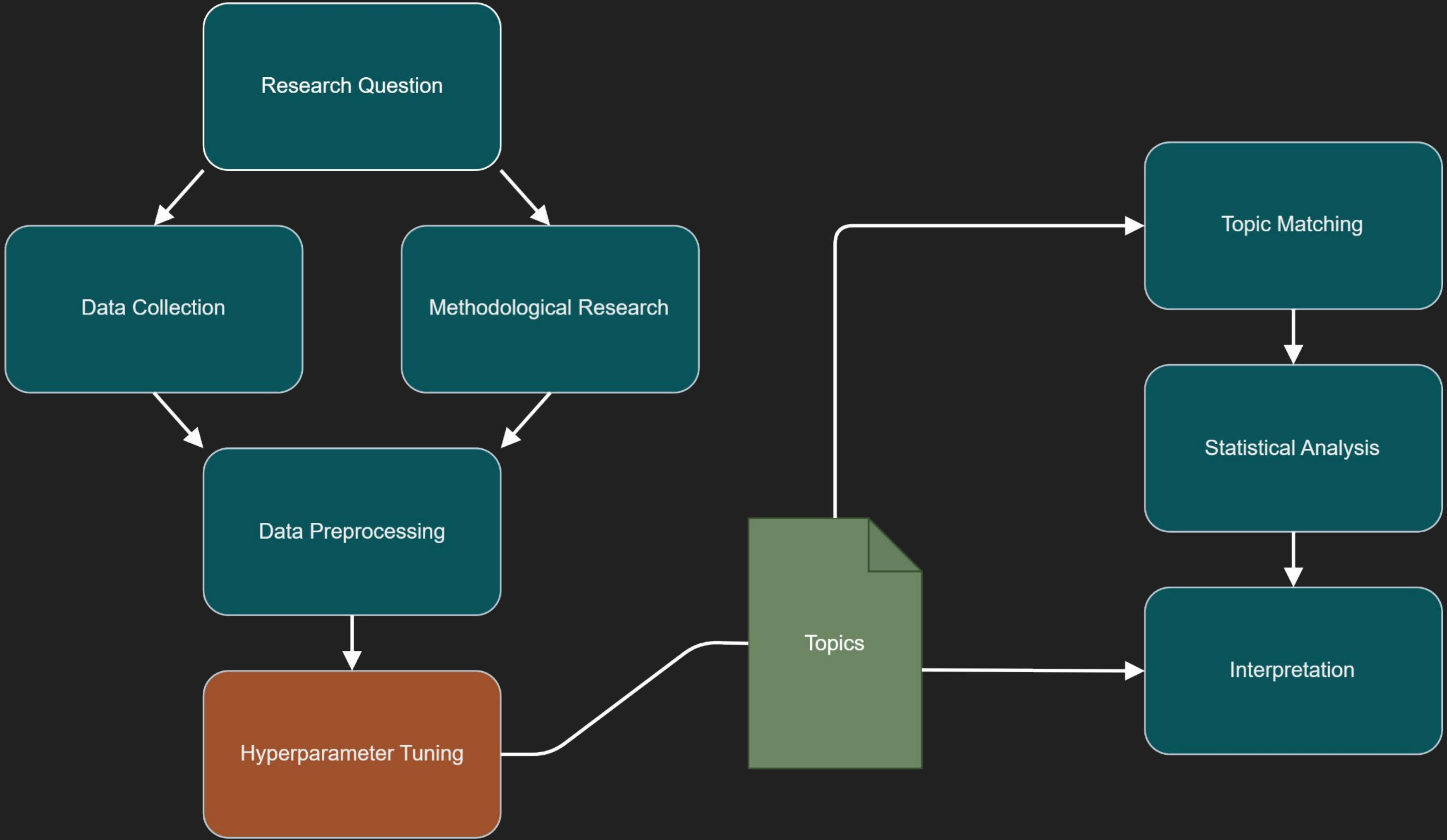

Research Question
Data Collection
Methodological Research
Data Preprocessing
Hyperparameter Tuning
Topics
Topic Matching
Statistical Analysis
Interpretation

Data Preprocessing

- Little preprocessing because LLM-embeddings deal well with stop words etc
- Some cleaning (mainly Twitter)
  - @handle → "user"
  - removing hashtags, RT, hyperlinks,
  - Filtering tweets with less than 3 words [(IF Strydom, J Grobler (2023))]
- Min length of speeches: 2 500 chars

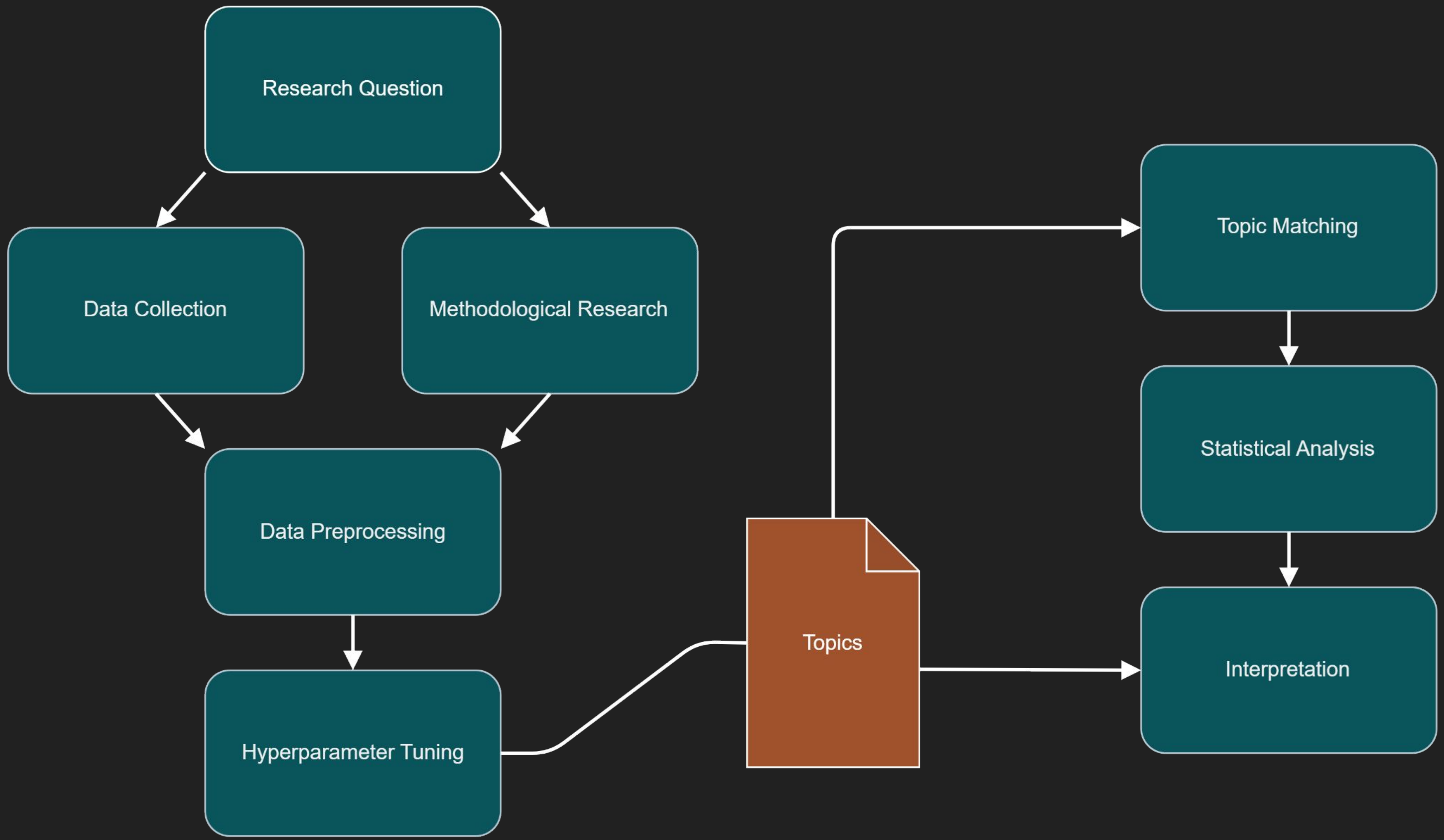


Distribution of speech length
Frequency
2500
2000
1500
1000
500
0
0
2500
5000
7500
10000
12500
15000
17500
20000
Speech Length


Research Question
Data Collection
Methodological Research
Data Preprocessing
Hyperparameter Tuning
Topics
Topic Matching
Statistical Analysis
Interpretation

Hyperparameter Tuning

- Training a model for each corpus independently
- Starting point is the default configuration
- Trying out different parameters, comparing coherence of topics
  - human evaluation
  - Tune parameters relative to document length / corpus size


Research Question
Data Collection
Methodological Research
Data Preprocessing
Hyperparameter Tuning
Topics
Topic Matching
Statistical Analysis
Interpretation

Parliament

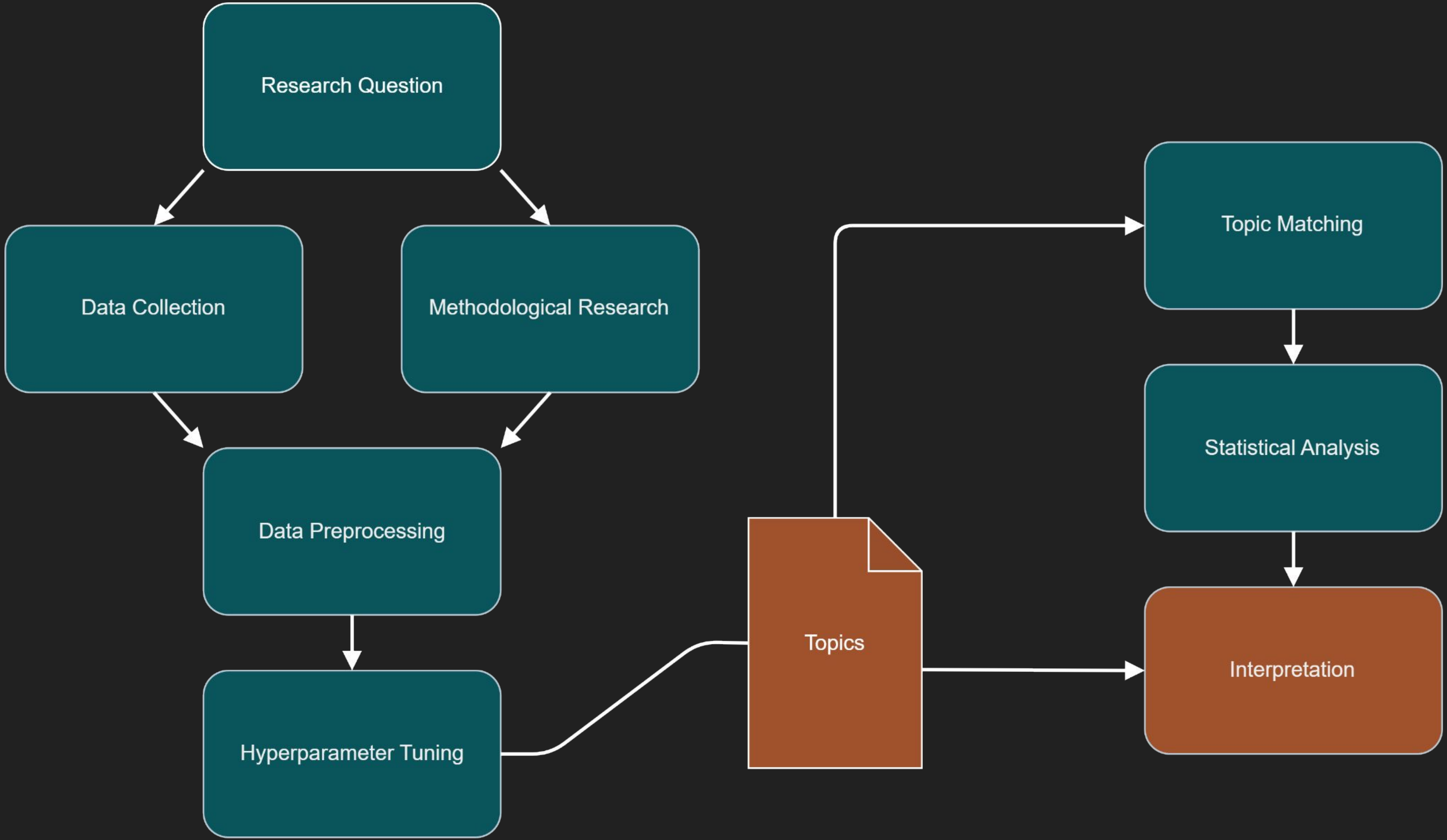
- 132 topics
- 9 889 classified (52%)

80_euro_währungszone_währungsunion_wachstumspakts
81_militäreinsatz_militarisierte_mediterranen_mittelmeerraum
82_bildungspolitischen_berufsschulen_berufsausbildung_akademikern
83_beitragszahlerinnen_ostdeutschland_altersversorgung_beitragszahler
84_aufklärungsdrohnen_drohnenkriegsführung_drohnenpiloten_kampfflugzeuge
85_transgeschlechtliche_transsexuellengesetzes_transsexuellengesetz_transgenderfrau

Twitter

- 98 topics
- 288 373 classified (58%)

65_gentechnikfreiheit_gentechnikrecht_biotech_gentechnik
66_lesenswert_lesenswerter_lesen_leseförderung
67_clubkultur_clubsterbenstoppen_clubbetreibende_clubsterben
68_organspendebereitschaft_organspendezahlen_organspendeausweis_organlebendspende
69_tierschutzgesetz_tierschutz_tierhaltungskennzeichnung_tierzahlen
70_ernährungsstrategie_nahrungsergänzungsmitteln_esseneinfachmachen_ernährungssystem


Research Question
Data Collection
Methodological Research
Data Preprocessing
Hyperparameter Tuning
Topics
Topic Matching
Statistical Analysis
Interpretation

Super Topics for analysing content

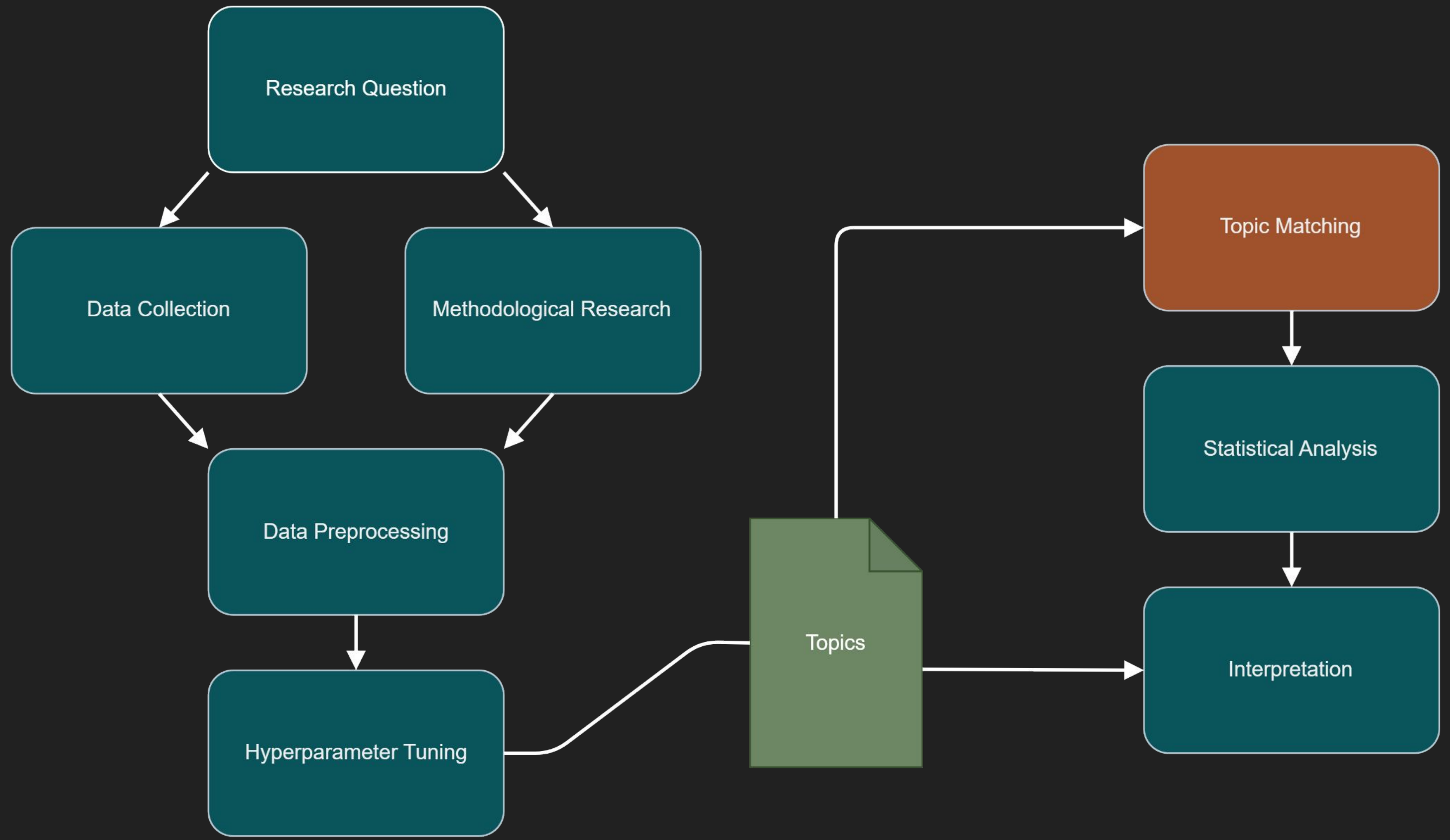
Using the ministries from the respective government and splitting their responsibilities into *Super Topics* (e.g. “Bundesministerium für Justiz und Verbraucherschutz” → Super Topics *Justice* and *Consumer Protection*)

- *Economics*
- *Energy*
- *Finance*
- *Interior and Community*
- *Foreign Affairs*
- *Justice*
- *Labour and Social Affairs*
- *Defence*
- *Food*
- *Agriculture*
- *Family Affairs, Senior Citizens, Women and Youth*
- *Health*
- *Digital Infrastructure*
- *Transport*
- *Environment, Nature Conservation and Nuclear Safety*
- *Consumer Protection*
- *Education and Research*
- *Economic Cooperation and Development*
- *Building*
- *Other*

Super Topics for analysing content

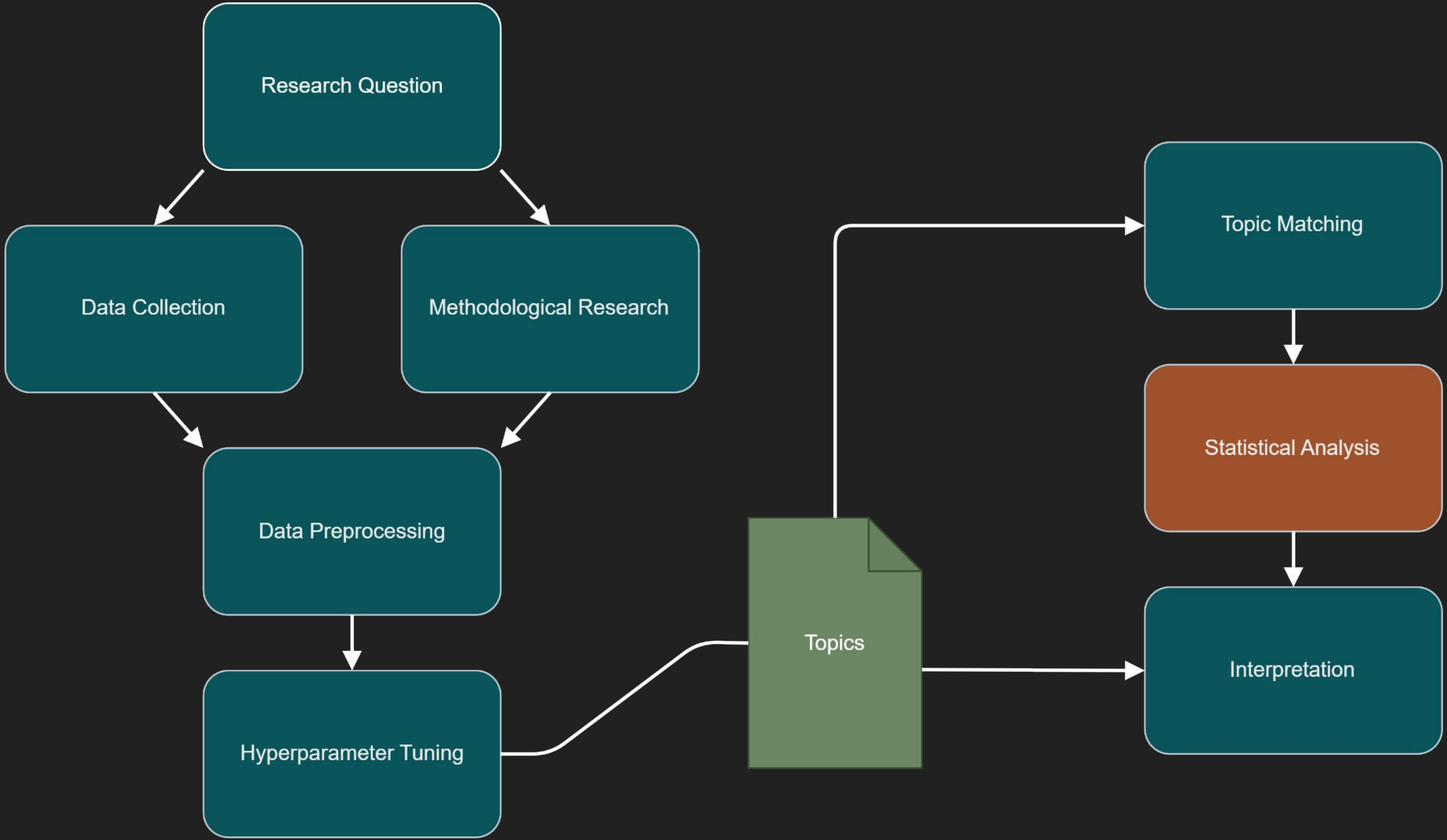
Using the ministries from the respective government and splitting their responsibilities into *Super Topics* (e.g. “Bundesministerium für Justiz und Verbraucherschutz” → Super Topics *Justice* and *Consumer Protection*)

- 44% of parliament speeches assigned to Super Topics
- 25% of Tweets assigned to Super Topics


Research Question
Data Collection
Methodological Research
Data Preprocessing
Hyperparameter Tuning
Topics
Topic Matching
Statistical Analysis
Interpretation

Matching Topics between Parliament and Twitter

- Cosine similarity between embeddings of topic representations
- Identify closest 3 candidates for matches
- Individually assign best match (majority vote)
- Interrater reliability: Fleiss' $\kappa = 0.76$
- 31% Parliament speeches matched
- 22% of Tweets matched

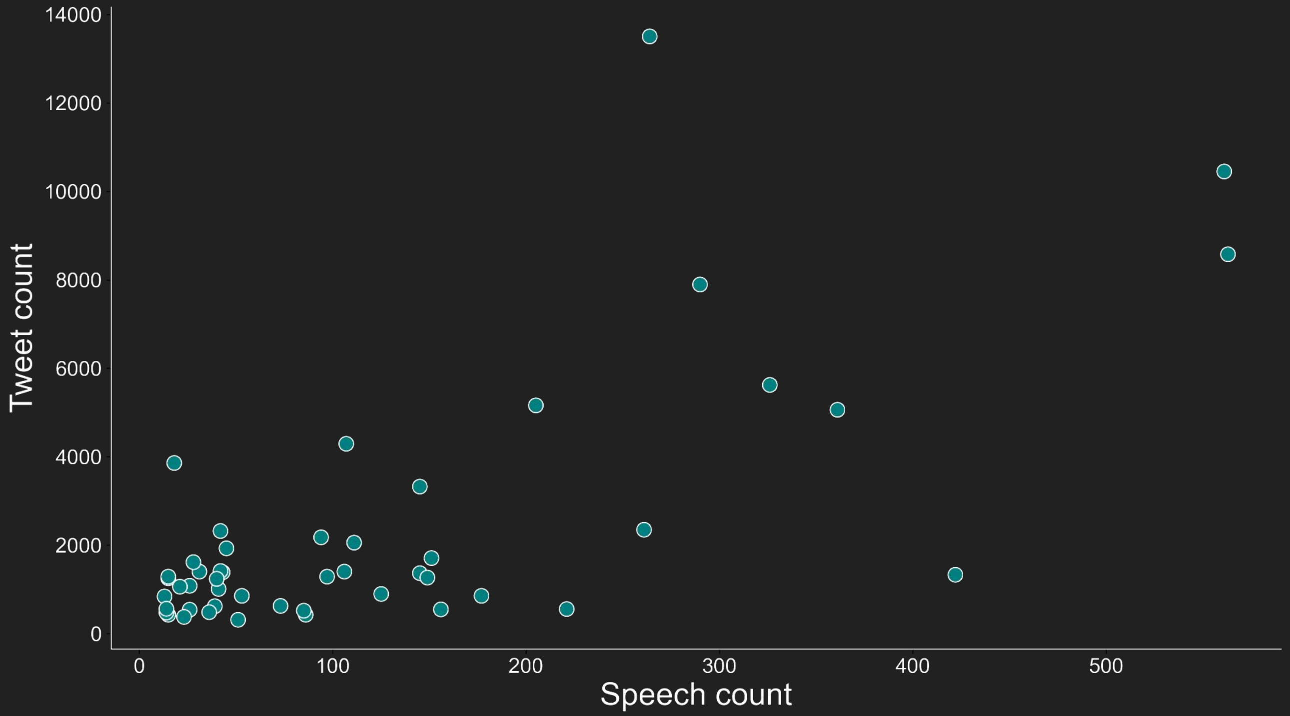

Research Question
Data Collection
Methodological Research
Data Preprocessing
Hyperparameter Tuning
Topics
Topic Matching
Statistical Analysis
Interpretation

Results

- Matched topics from parliament and Twitter
- Correlate numbers of documents in these topics
- Compute rank correlations:
  - Spearman's $\varrho = .54$, $p < .001$
  - Kendall's $\tau = .38$, $p < .001$

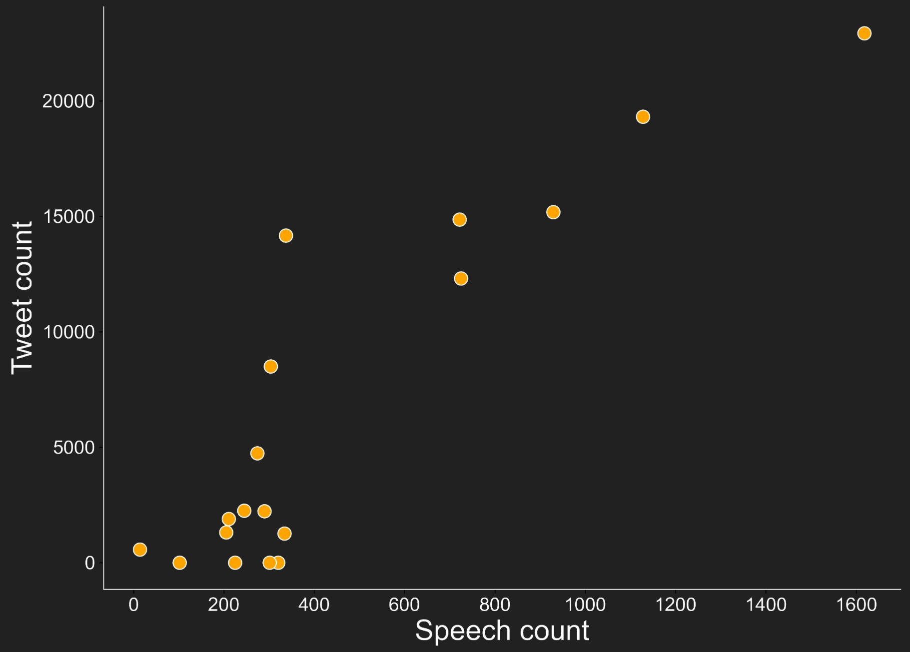
Results (2)

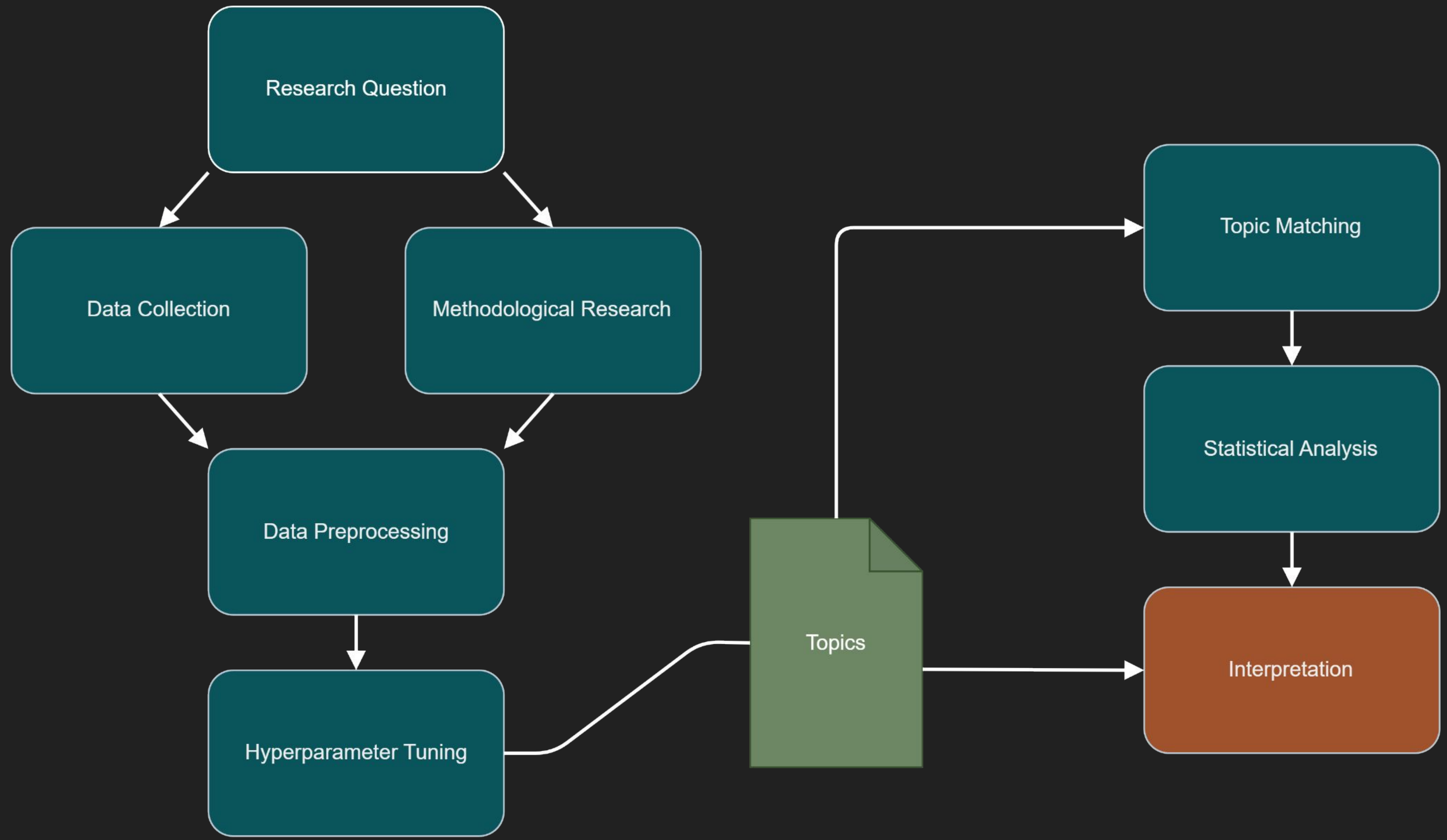
- $\varrho = .54$,
  $p < .001$
- $\tau = .38$,
  $p < .001$



Results (3)

- Same procedure applied to the data on a super topic level
- Spearman's $\varrho = .74$, $p < .001$
- Kendall's $\tau = .62$, $p < .001$



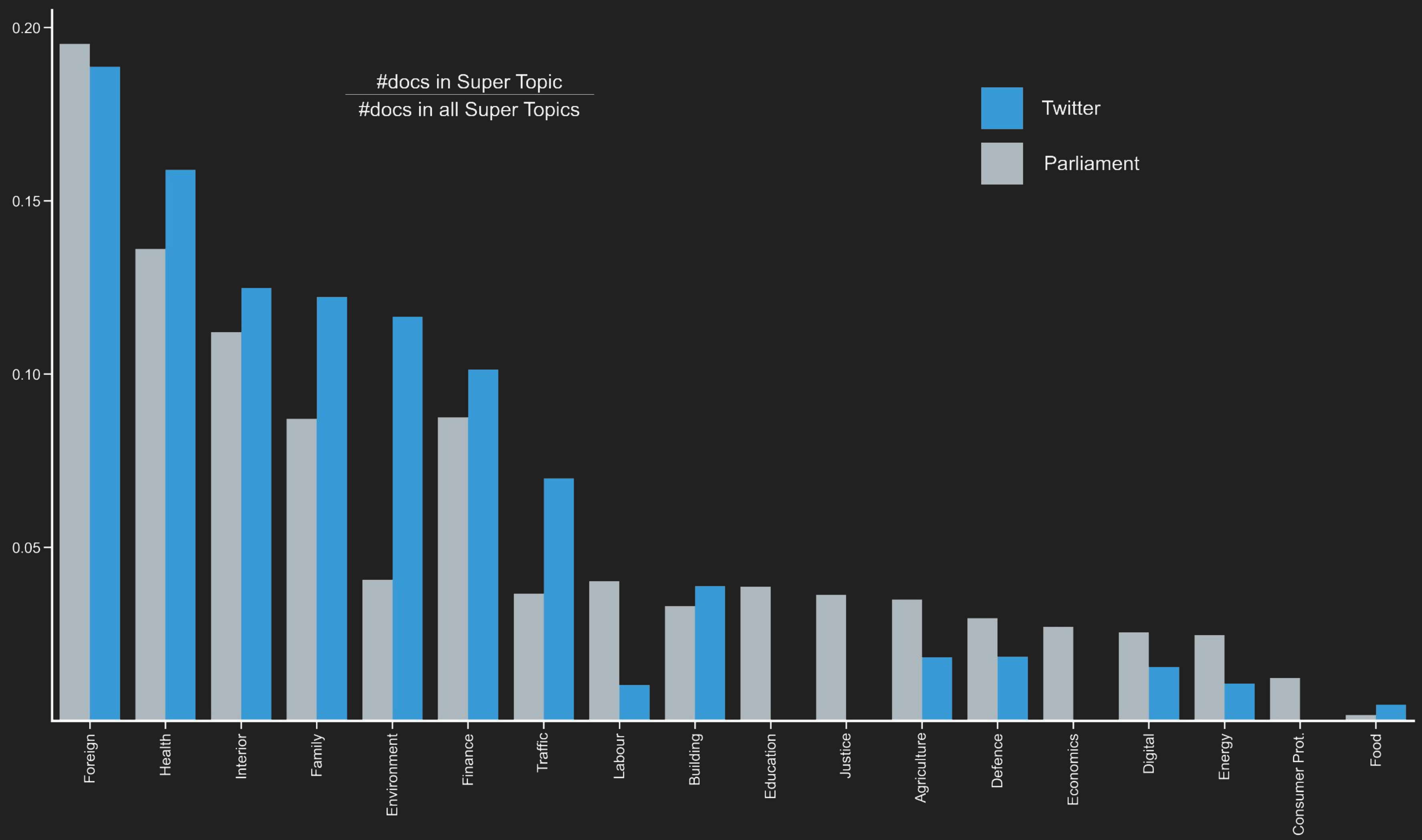

Research Question
Data Collection
Methodological Research
Data Preprocessing
Hyperparameter Tuning
Topics
Topic Matching
Statistical Analysis
Interpretation

Parliament

1. Covid (5.7%)
2. Immigration (4.3%)
3. Budget and Finance (3.5%)

Twitter

1. Climate (4.7%)
2. Tax politics (3.6%)
3. Covid (3%)



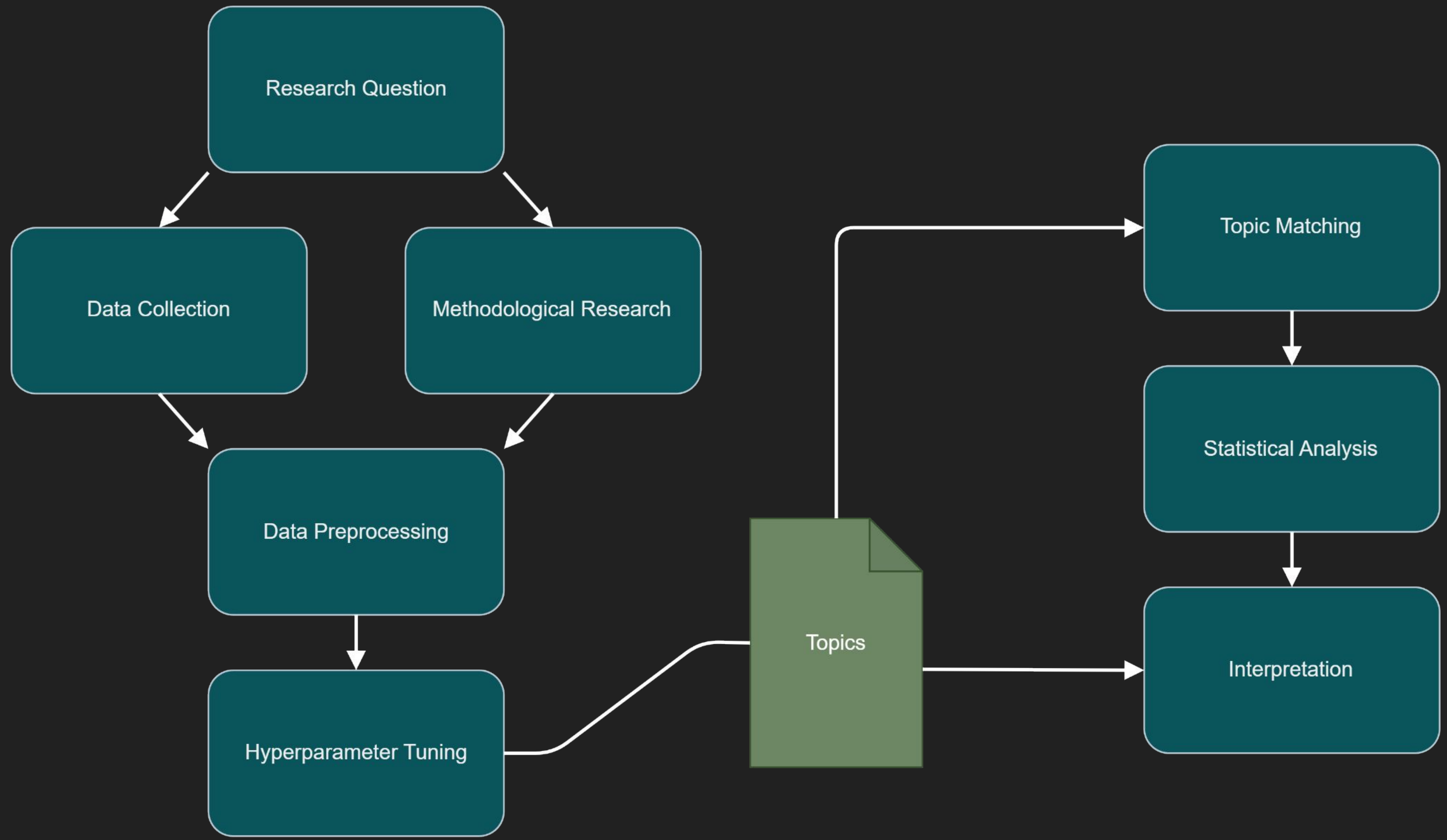
#docs in Super Topic
#docs in all Super Topics
Twitter
Parliament
0.20
0.15
0.10
0.05
Foreign
Health
Interior
Family
Environment
Finance
Traffic
Labour
Building
Education
Justice
Agriculture
Defence
Economics
Digital
Energy
Consumer Prot.
Food

Research Question

Do the debated topics on Twitter by German politicians differ from the topics debated in the German parliament?

- significant, moderate to strong correlation
- but: for some topics clear deviations

Yesn’t


Research Question
Data Collection
Methodological Research
Data Preprocessing
Hyperparameter Tuning
Topics
Topic Matching
Statistical Analysis
Interpretation

Limitations

- Corpora are vastly different
  - language
  - length of documents
  - number of documents
- Topic modelling leaves almost 50% of documents unclassified
  - and even more for Super Topics
  - and some topics do not seem coherent (especially for Tweets)

['fischereipolitik', 'maritime', 'maritimen', 'fischereibetriebe', 'überwasserschiffbau', 'unterwasserschiffbau', 'schiffbauindustrie', 'schiffbau', 'fischereiaufsicht', 'schiffsbesetzungsverordnung']

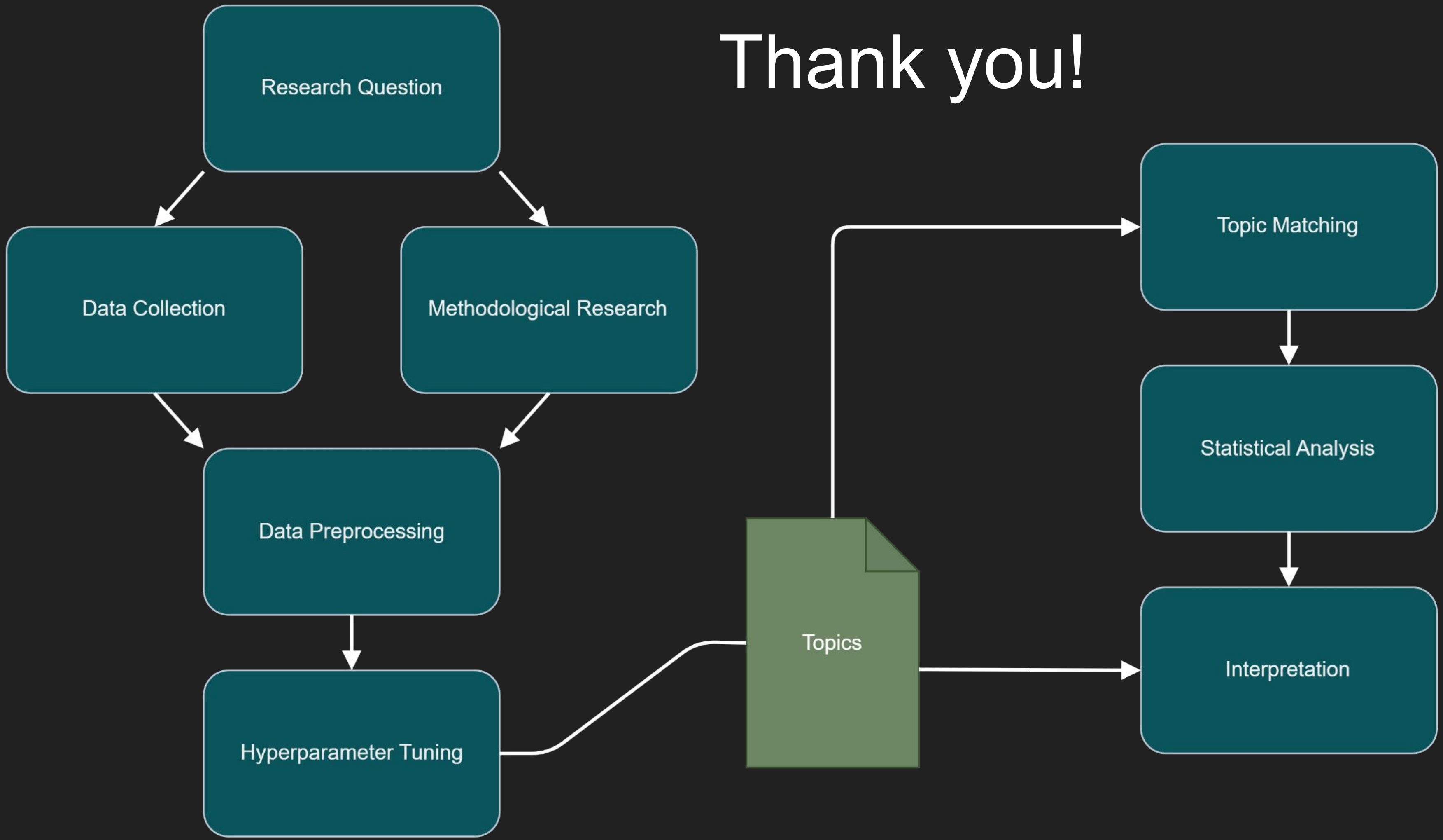
Limitations

- Corpora are vastly different
- Topic modelling leaves almost 50% of documents unclassified
- Matching handwavy, although controlled for interpersonal difference
- Assignment of topics to super topics unclear at times
    - Super topics overlap heavily (e.g. Agriculture and Environment, Nature Conservation)
    - Some topics not covered by super topic (media, journalism)

Future Research

- Does the discrepancy in Environment Topic hold now, while the Green Party is in Government?
- Fuzzy Matching of Topics to Super Topics (Overlapping Clusters)
- Temporal relation between Topics in Parliament and Topics on Twitter

Thank you!


Research Question
Data Collection
Methodological Research
Data Preprocessing
Hyperparameter Tuning
Topics
Topic Matching
Statistical Analysis
Interpretation