

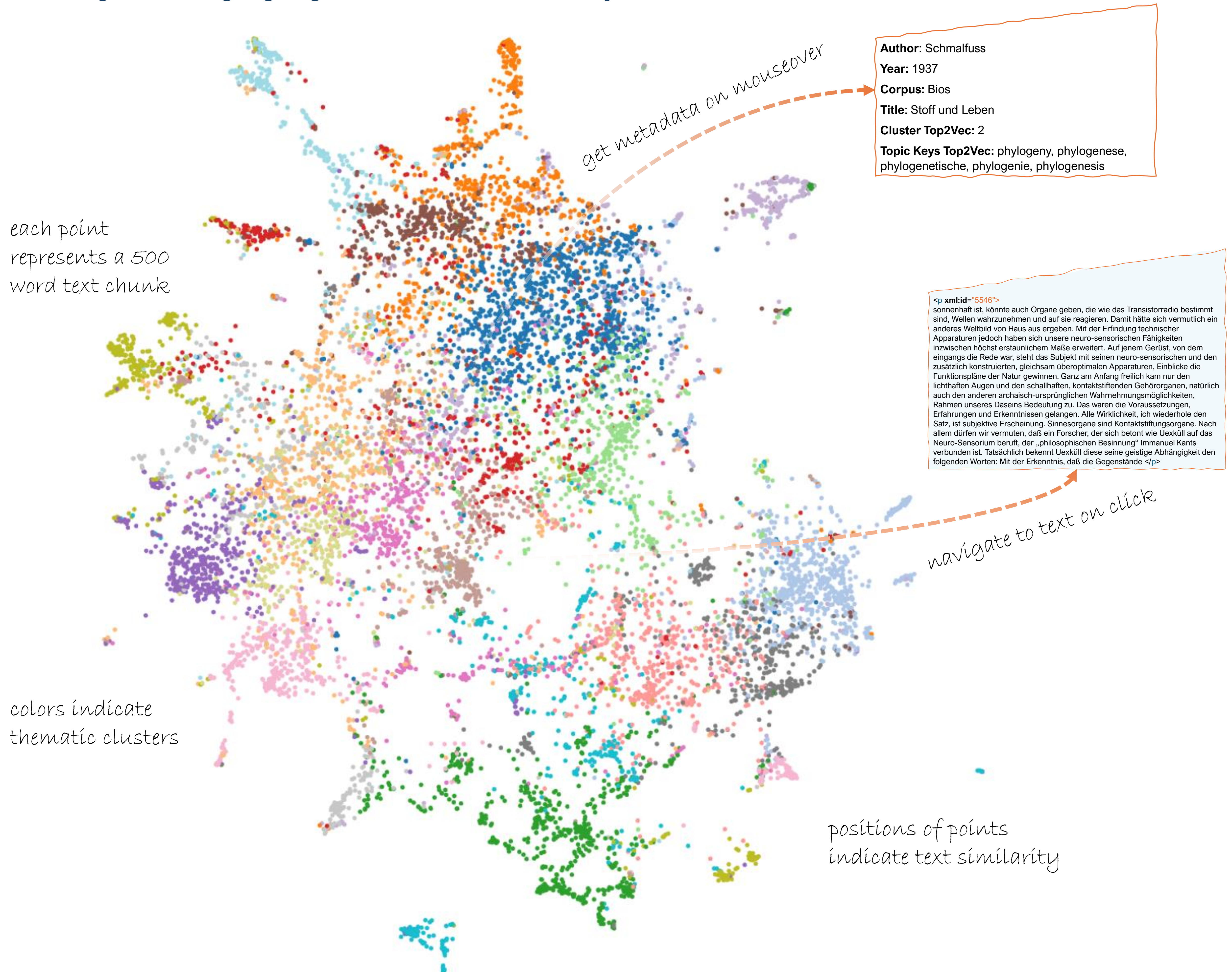
# HERMENEUTOPIC

A Workflow for Multilingual Text Corpus Exploration

Stefan Reiners-Selbach  
Faculty of Arts & Humanities

## LET'S GO BACK TO THE TEXTS...

Text mining techniques such as topic modeling and document embedding gain more and more traction in the Humanities as they enable us to engage with large amounts of text, enabling new kinds of research questions. But, for Humanities scholars there is the need to return to the texts: The Human Reader has to stay in the loop, hermeneutical approaches remain important. **HermeneuTopic** aims to leverage the results of digital analyses to structure and inform the reading process. **Multilingual, text length and language agnostic, and novice friendly.**

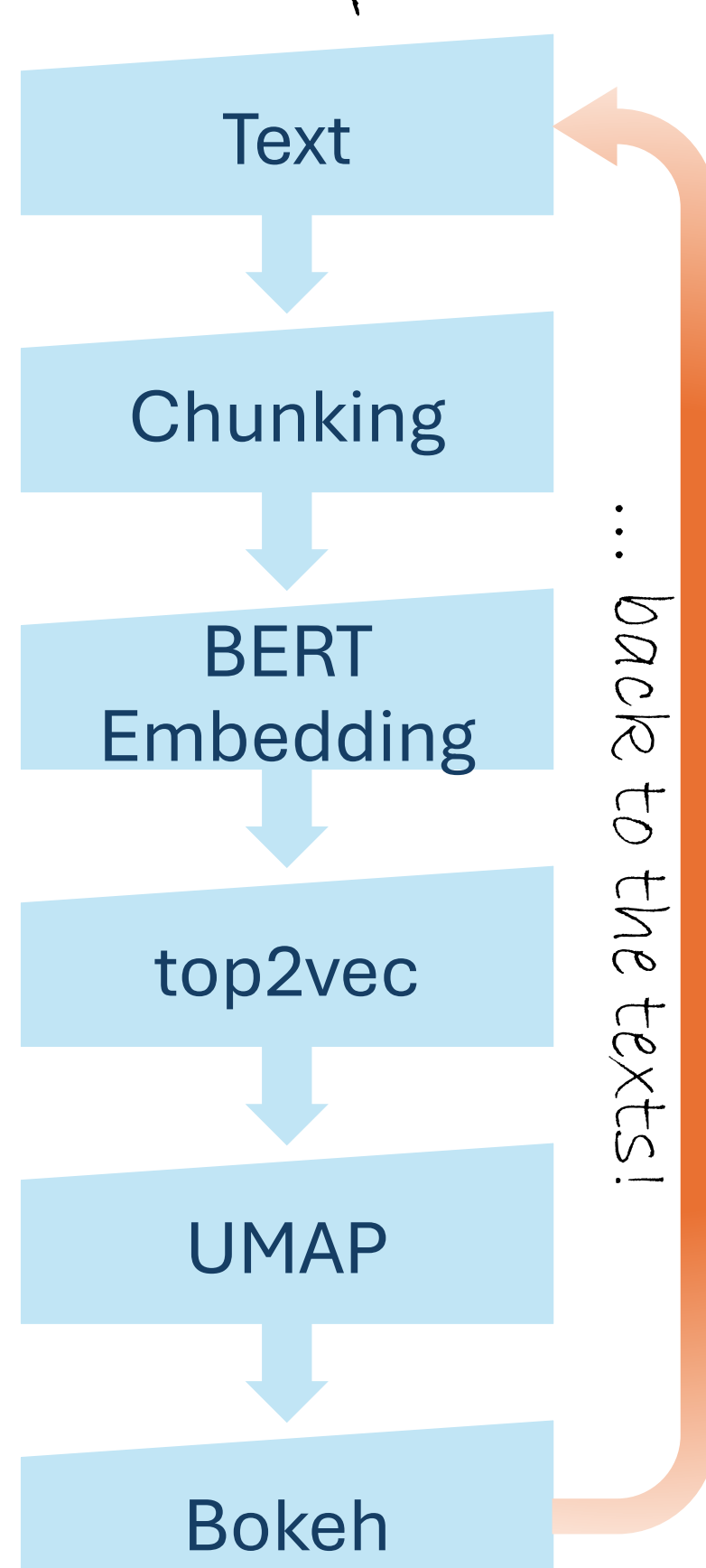


## ...AND MAKE TEXT MINING TECHNIQUES ACCESSIBLE!

Our histories, arts, and cultures are diverse – especially if we are interested in objects apart from the canon! Using text mining techniques in the history and philosophy of science for example, we can gain insights into scientific practice like never before [1], but relevant texts for a given research question may vary wildly in regard to language, text length, and format. Therefore, multilingual text mining techniques that are agnostic towards these points and highly flexible are needed [2]. To integrate these techniques in common humanities research and to enable novices to engage with their results, they can be used to help structure a more democratic reading approach: Visualizations can serve as intuitive interfaces to texts, to navigate the underlying text collections thematically and structurally, regardless of a text's canonical status today [3].

To achieve this, we first separate each text into 500 word chunks, on which we employ top2vec for embedding based topic modeling [4] without any additional pre-processing, using a multilingual BERT model [5]. UMAP is used for dimensionality reduction of the embedding [6]. The resulting two-dimensional scatter plot is then visualized interactively with Bokeh [7], linking each point that represents a text chunk with the corresponding fragment in the context of its source text and coloring according to the top2vec topic model, which can be treated as thematic clusters. This visualization serves as an interactive map of the corpus, showing its underlying thematic structure. On mouseover, users can view metadata and the topic keys to gain further insight, informing their reading; on click they can navigate to the corresponding text fragment.

## Workflow



### References

- [1] See e.g., O. M. Laan, L. Rivelli, and C. H. Pence, "Digital Literature Analysis for Empirical Philosophy of Science," *British Journal for the Philosophy of Science*, vol. 74, 2023, doi: <https://doi.org/10.1093/bjps/axz049>.
- [2] M. Nochi, "PhiloBERTa: Ein multilinguales Sprachmodell zur Beantwortung philosophischer Fragestellungen," in *DH2023: Open Humanities, Open Culture*, Q. Dombrowski, "What's a 'word': Multilingual DH and the English Default," Accessed: Jul. 19, 2023. [Online]. Available: <https://www.quindombrowski.com/blog/2020/10/15/whats-word-multilingual-dh-and-english-default/>
- [3] S. Reiners-Selbach, J. Baedke, A. Böhm, A. Fábregas Tejeda, and V. Straetmanns, "HermeneuTopic. Ein Workflow zur interaktiven mixed-methods Exploration (philosophie-)historischer Textkorpora," in *Book of Abstracts DH2024*, 2024. [Online]. Available: <https://zenodo.org/doi/10.5281/zenodo.10698453>
- [4] D. Angelov, "Top2Vec: Distributed Representations of Topics," arXiv, Aug. 19, 2020, doi: [10.48550/arXiv.2008.09470](https://doi.org/10.48550/arXiv.2008.09470).
- [5] N. Reiners and I. Gurevich, "Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation," arXiv, Oct. 05, 2020, doi: [10.48550/arXiv.2004.09813](https://doi.org/10.48550/arXiv.2004.09813).
- [6] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," arXiv, Sep. 17, 2020, doi: [10.48550/arXiv.1802.03426](https://doi.org/10.48550/arXiv.1802.03426).
- [7] Bokeh Development Team, Bokeh: Python library for interactive visualization, 2018. [Online]. Available: <https://bokeh.pydata.org/en/latest/>