

TETYS: towards the next-generation open-source Web topic explorer

Anna Bernasconi, Francesco Invernici, and Stefano Ceri

Contact: anna.bernasconi@polimi.it

Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy

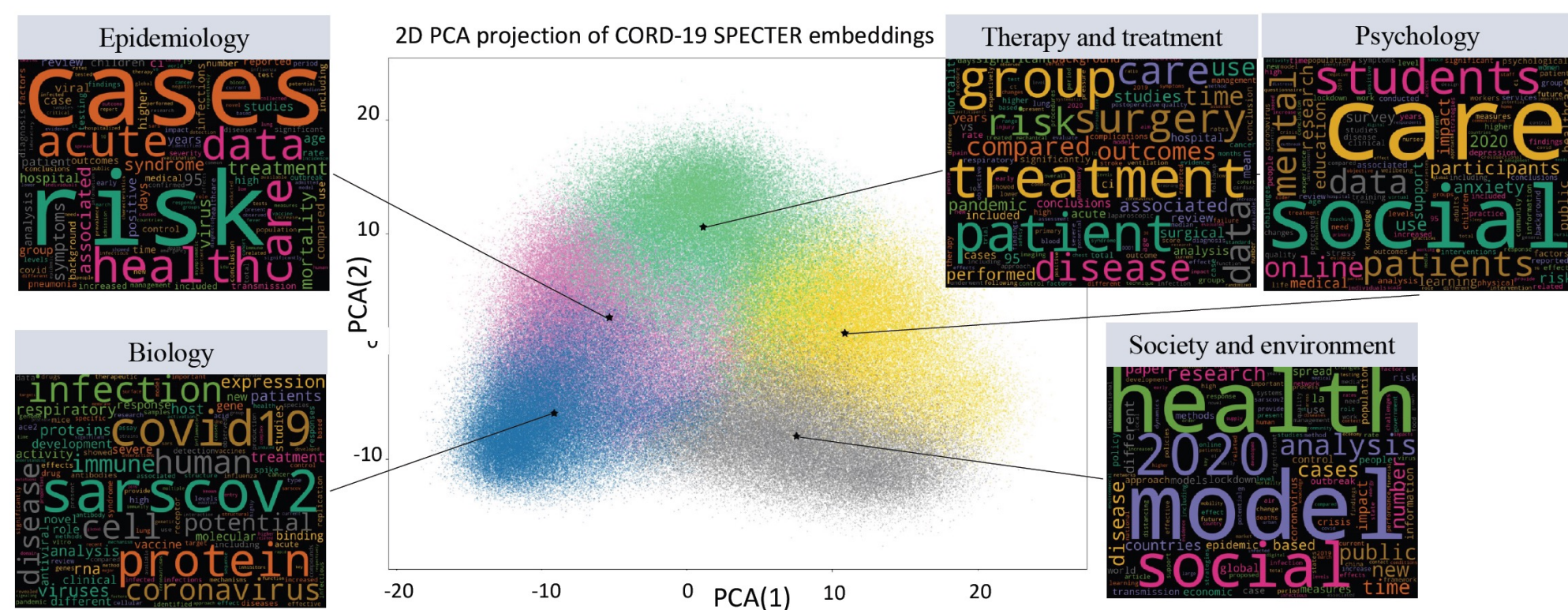
Web users who search specialized content typically lack knowledge of the precise topology of the dataset.

→ We propose a *next-generation open-source Web topic explorer*, an architecture with:

- 1) a **pipeline** for ingesting huge data corpora, extracting highly relevant topics
- 2) an interactive **dashboard**, supporting topic visualization, exploration of temporal series, and statistical testing

Our prototype, CORToViz, explores the CORD-19 dataset (COVID-19 research abstracts). Other domains will be explored in TETYS (e.g., climate change).

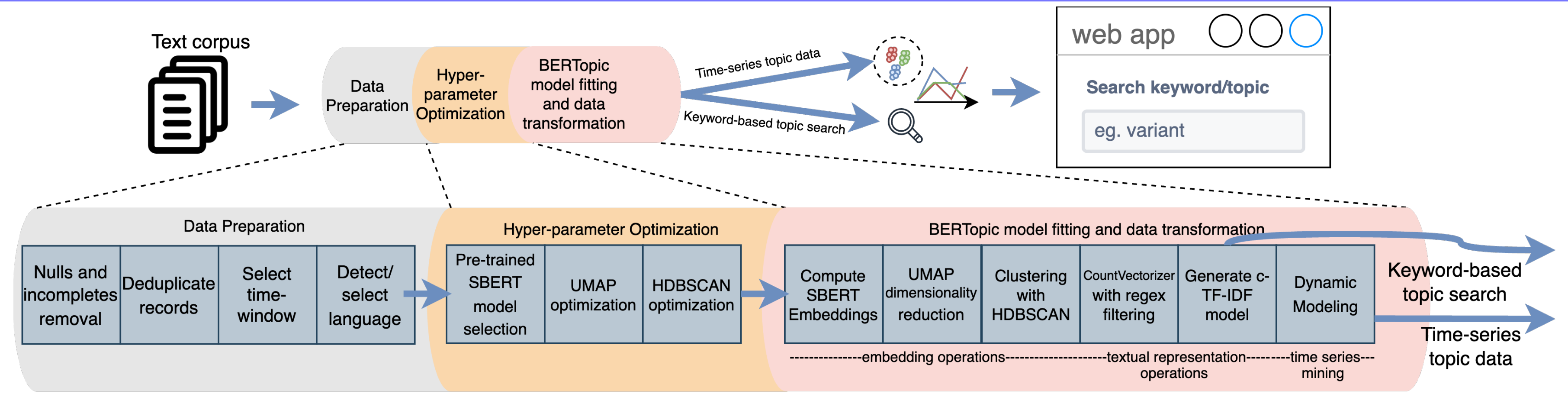
Topic modeling



- Based on up-to-date technologies (including LLMs) and self-optimizing models (see BERTopic framework [1]);
- Unsupervised topic modeling via clustering of articles along the orthogonal dimensions of the document embeddings (latent representations generated by language models) – computed with SPECTER [2];
- Visualization of topics through wordclouds (generated via *gensim* Python package)
- Temporal topic modeling via time series mining and statistical testing.

TETYS pipeline

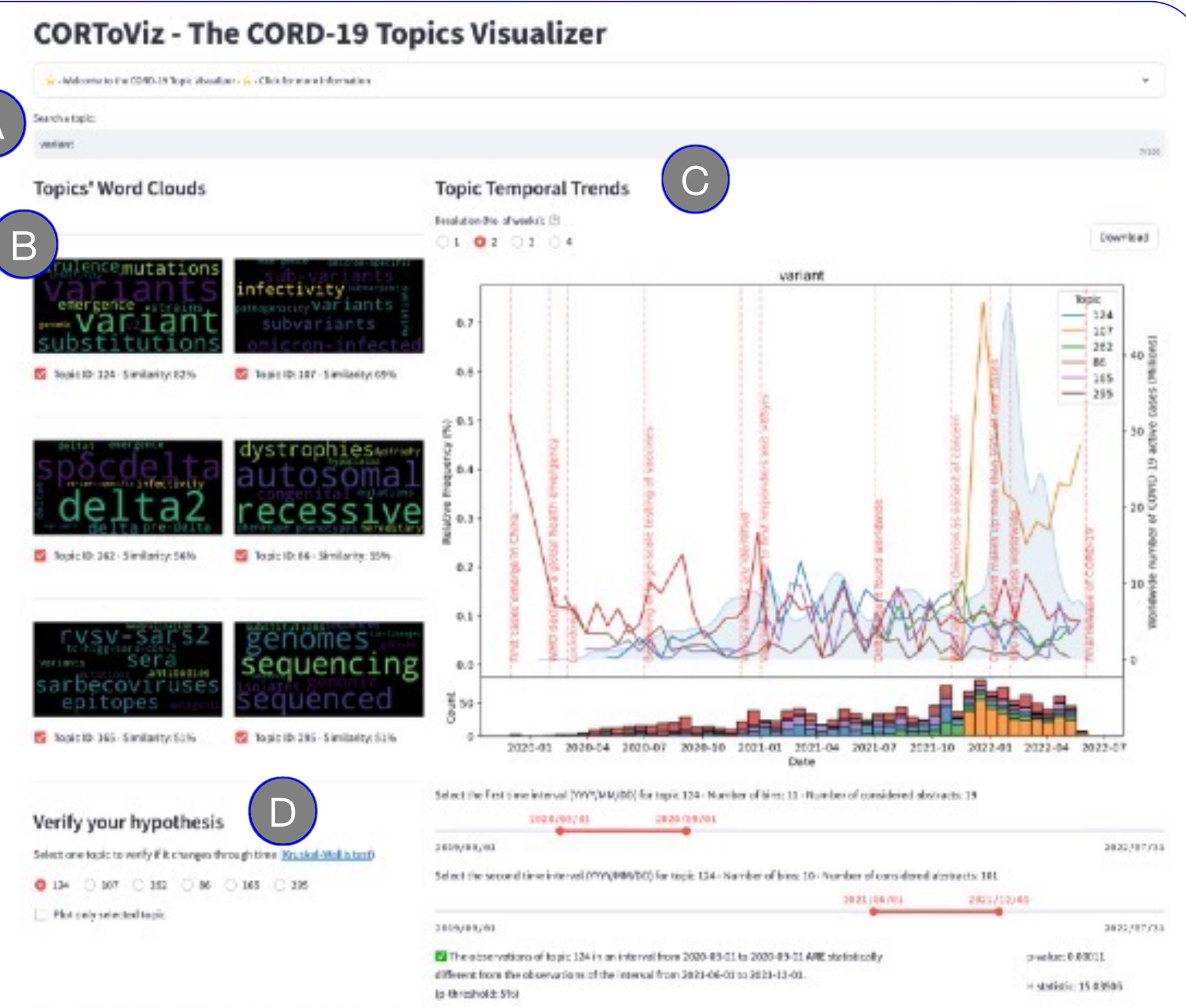
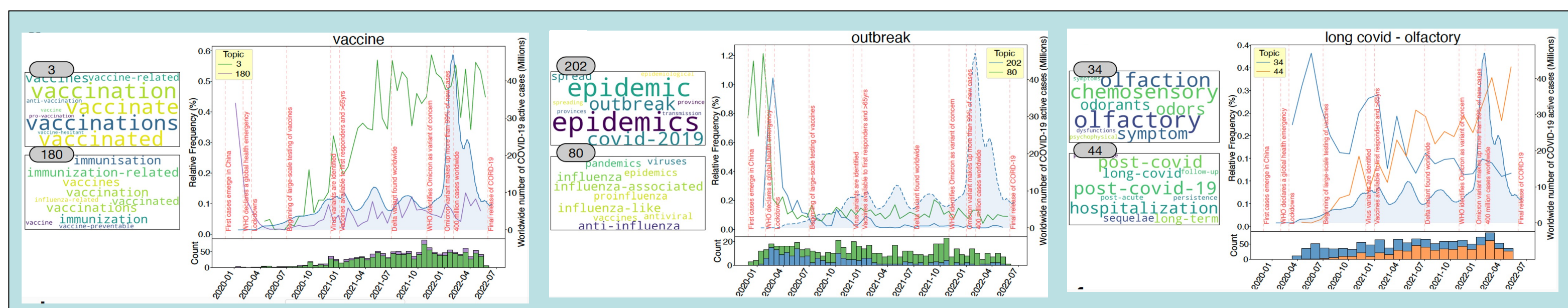
- Data preparation
- Hyper-parameter optimization
- Fitting of the topic model
- Data transformation, for enabling
 - Time series visualization
 - Keyword search functionalities



CORToViz dashboard (<http://gmql.eu/cortoviz>) [3]



- Keyword-based search bar
- Six top-ranked topics, explained through their word clouds
- Line plot of the intensities (i.e., the relative frequencies of appearance) of the two selected topics – with configurable bin resolution (1-4 weeks).
- Panel showing statistical testing (Kruskal-Wallis test for the difference between groups)

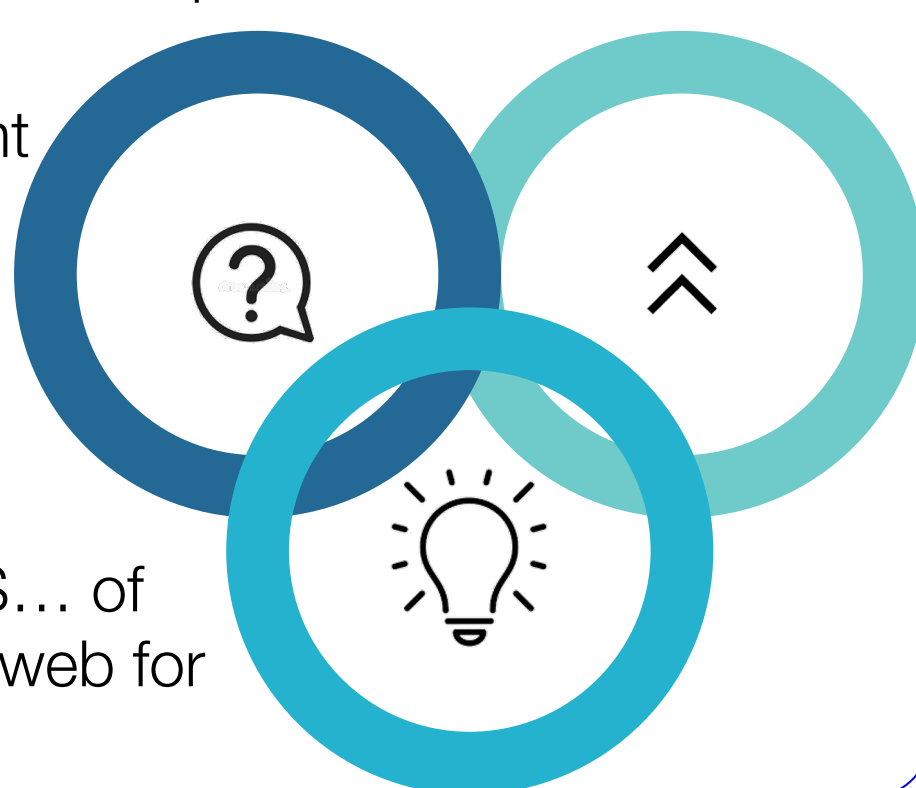


Excellence

INNOVATION... bringing a one-click solution to apply topic modeling on virtually any possible document corpus, at the service of Web searchers

ADDRESSING NEEDS... of users who search the web for specialized content

AMBITION... to make research information more accessible to common Web users who access technical / specialized contributions



Impact

SCIENTIFIC IMPACT... contribution to fostering the diffusion of knowledge through the Open Science paradigm

ADDED VALUE TO PLATFORMS... increasing the engagement of users and prospectively bringing more interest toward them



COMMUNICATION STRATEGY IMPACT... in multiple fields

Next steps

- Build second prototype to investigate United Nations' SDG-related research [4]
- Generalize the pipeline for any corpus of Web textual documents
- Add a *statistical-toolbox* with possibility to further analyse topics time series
- Transform the web-application user experience and validate it with solid user studies

Bibliography

- [1] Grootendorst, M., 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794.
- [2] Cohan, A., et al. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In Proc. 58th ACL.
- [3] Invernici, F., Bernasconi, A. and Ceri, S., 2024. Exploring the evolution of research topics during the COVID-19 pandemic. Expert Systems with Applications, 252, p.124028.
- [4] United Nations' Sustainable Development Goals (SDG) - <https://sdgs.un.org/goals>

Team



Anna Bernasconi
Assistant Professor
Bioinformatics, Databases,
and Data Science



Francesco Invernici
1st year PhD Candidate
Knowledge Management,
Databases, and NLP



Stefano Ceri
Full Professor
Databases and Web Technologies

Acknowledgement & Funding



Funded by the
European Union

This research is supported by the TETYS project, a beneficiary of the NGI Search 2nd Open Call (Number of the Sub-grant Agreement SEARCH_OC2_18), funded by the European Union. The NGI Search project has received funding from the European Union's Horizon Europe research and innovation programme under the grant agreement 101069364 and it is framed under the Next Generation Internet Initiative.