GESIS Notebooks

Online Reproducible Computational Analysis for the Social Sciences

Authors: Arnim Bleier, Kenan Erdogan, Christian Kahmann, Lisa Posch

The use of computational methods in the Social Sciences has many roots. Classic examples include the fitting of measurement-, as well as structural models or the recognition of simulation as a third way of doing science. More recently, the availability of large organic datasets has led to the emergence of Computational Social Science. All these examples have in common that potentially compute-intensive and complex analysis pipelines are an essential part of the research. Reproducibility is the cornerstone of modern science. Yet more work is needed to make reproducibility the norm for complex computational analysis pipelines.

GESIS Notebooks aims to close this gap in the Social Sciences by supporting reproducible computational analysis pipelines that enable researchers to focus on the methodological aspects of their work while increasing their visibility and following best open-science practices simultaneously. The service does not try to reinvent the wheel but reuses existing technology whenever possible. Execution, place, and content are the concepts of the service, and in the remainder, we will look at each of them.

Execution

Executability is the fundament of computational reproducibility. To provide this functionality, we build on the containerization that has revolutionized software development and IT operations in recent years. Typically referred to as DevOps, this engineering field aims to shorten development and provide continuous delivery of software. DevOps has already made its strides into machine learning under the name of MLOps; GESIS Notebooks extends these concepts to social science. Technically, docker containers are built based on standardized software package dependency descriptions. The software jupyter-repo2docker enables this building of containers for projects on GitHub, GitLab, Zenodo, Figshare, and Dataverse, ensuring a solid integration with the typical publication venues for scientific analysis code in the Social Sciences. The containers are then available for deployment to the following reproducible execution environments:

- A BinderHub instance to allow for unauthenticated short-term interactions with analysis pipelines not in need of extensive computational resources such as GPU or CPU cycles.
- A JupyterHub instance with Binder import capability to allow for authenticated work across multiple sessions with analysis pipelines in need of more computational resources and persistent storage.

Place

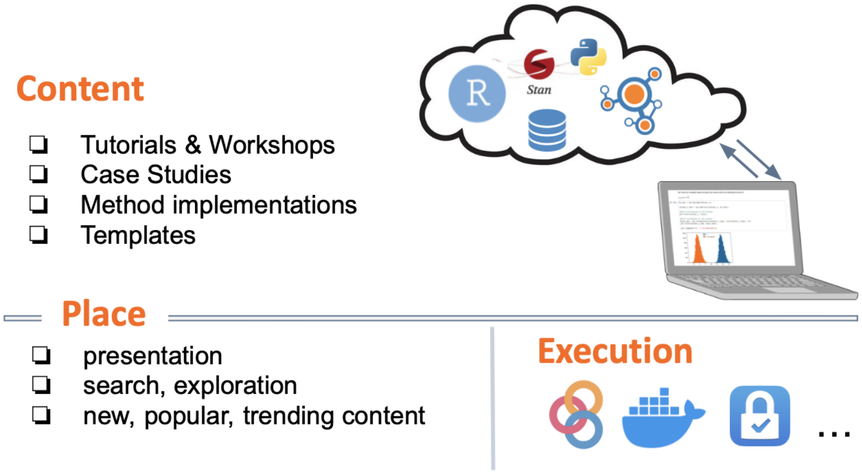
While execution provides the backend, it remains hidden behind a frontend. This place connects published analyses to computation at a click of a button in a gallery. Researchers publish their analyses on established archival platforms such as Zenodo, GitHub, and Dataverse; the gallery then provides a portal for this content tailored to the Social Sciences. In particular, this gallery allows content presentation and its search and exploration. To kick-start the content in the gallery, a hybrid approach along three categories has been taken:

- Featured analysis pipelines, selected according to a prototypical editorial acceptance process.
- Analysis pipelines created by researchers at GESIS that have not undergone any editorial selection process.
- Analysis pipelines from the MyBinder.org Events Archive may even come from researchers outside the field of the Social Sciences.

Our three-fold approach does not only provide an answer to the cold-start problem inherent to any such project. More importantly, it ensures FAIR access to selected material from the Social Sciences without unnecessarily locking interdisciplinary researchers into rigid domain-specific silos.

Content & Concluding Remarks

Place and execution are the fundamental principles in our architecture. However, they are means to enable reproducible computational analysis pipelines. Use cases demonstrate the value of the system. Examples of use cases from GESIS and beyond include the "Notebooks for Computational Sociology" toolkit (compsoc), the German Federal Election 2017 Twitter Dataset (btw17_sample_scripts) access, and LCM Binderhub.



Components of GESIS Notebooks

The project is funded by the DFG (FKZ/project number: 324867496), and carried out in cooperation with NFDI4DS (FKZ/project number: 460234259) and CESSDA.