# Die Wahl der Mittel – Jupyter-Notebooks als Forschungsinfrastruktur

Kerstin Jung[1], Pascal Hein[2], André Blessing[1], Jan Hess[3], Volodymyr Kushnarenko[4]

[1] Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
[2] Institut für Literaturwissenschaft, Universität Stuttgart
[3] Deutsches Literaturarchiv Marbach
[4] Höchstleistungsrechenzentrum Stuttgart (HLRS), Universität Stuttgart
https://www.sdc4lit.de

## Ausgangslage im Projekt

- Literatur im Netz als Untersuchungsgegenstand (z. B. literarische Blogs)
  - teilweise große Datenmengen, z. B. umfasst der Blog-Crawl des Techniktagebuchs (https://techniktagebuch.tumblr.com) aus der Sammlung des DLA im WARC-Format (Standard zur Archivierung von Webinhalten) ca. 9 GB und mehr als 7000 Blogbeiträge
- unterschiedliche Analyseinteressen
  - basierend auf extrahiertem Text: Topics, Named Entities, …
  - basierend auf Linkstruktur: warc2graph https://github.com/dla-marbach/warc2graph
- heterogene Nutzendengruppe (Autor*innen, Forschende, Schüler*innen)

unterschiedliche technische Vorkenntnisse

## Verfügbarmachung ausführbarer Jupyter-Notebooks?

- Betrieb eines eigenen Hubs, Voraussetzungen: https://jupyter.org/hub
  - Hardware
  - Administration (Nutzungsverwaltung, Monitoring von Speicher und Rechenkapazitäten)
  - Wartungskapazitäten für Updates auf Maschine, Hub und Paketen, Abhängigkeiten in den Notebooks
  - Sicherheitskonzept, da es sich bei Jupyter-Notebooks um ausführbaren Quellcode handelt

- Externe Notebook-Services wie z. B. Colaboratory (Google) https://colab.research.google.com/
  - große Gestaltungsmöglichkeiten für die Nutzenden, z. B. durch dauerhaft installierbare Pakete
  - oft problematische Speicherung aller Daten und teilweise unvorhersehbarer Umgang mit den eigenen Diensten
  - kein geeigneter Standard für die Forschung

- Nutzung von Notebooks innerhalb bestehender integrierter Entwicklungsumgebungen
- Download einer Konfiguration als Docker-Container
- Bereitstellung von Notebooks in Repositorien oder Versionskontrollsystemen für Download und Weiterentwicklung

➢ geeignet für Nutzende der entsprechenden Infrastrukturen bzw. mit technischen Vorkenntnissen.

- Forschungsgetriebene Ansätze und Umgebungen, z. B.
  - GESIS Notebooks, bringt zugängliche Notebooks aus Repositorien mittels Binder in eine Ausführungsumgebung https://notebooks.gesis.org/binder/; https://mybinder.org/
  - DHVLab und DH2go erlauben die Ausführung von Notebooks https://dhvlab.gwi.uni-muenchen.de/; https://dh2go.ilw.uni-stuttgart.de/

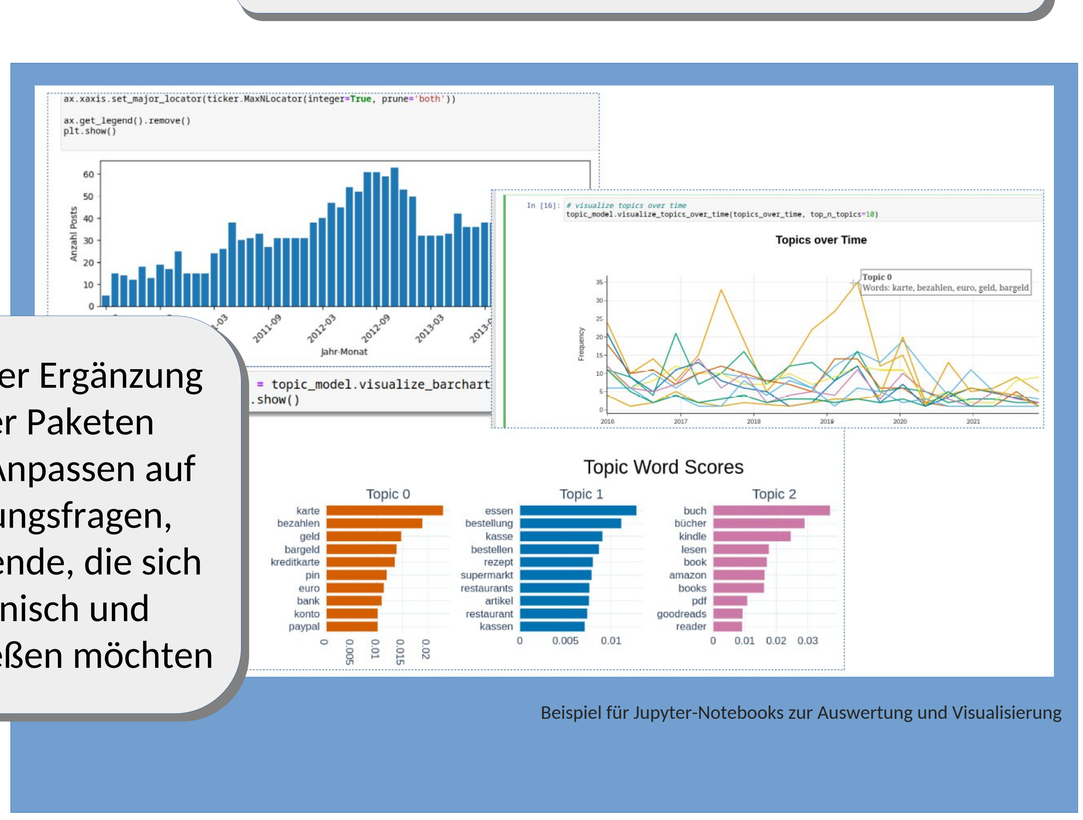
➢ Fokus auf Nutzende aus bestimmten Fachbereichen bzw. Institutionen

## Jupyter-Notebooks

https://jupyter.org

- kombinieren Dokumentation, Ausführung und Visualisierung von Programmcode, insbesondere von Python als vielgenutzter Programmiersprache in den Digital Humanities
- können ein geeigneter Zugang zu Daten und Analysemöglichkeiten für eine breite Nutzendengruppe sein
  - technische Voraussetzungen, z. B. benötigte Pakete, sind im Notebook selbst spezifiziert
  - bestehende Notebooks können fast ohne Vorkenntnisse mit Python betrieben werden und ermöglichen trotzdem an das eigene Forschungsinteresse angepasste Anfragen (z. B. durch interaktive Elemente wie Dropdown-Menüs oder Range-Sliders)
  - verändern der Fragestellung durch schrittweises Anpassen des Programmcodes möglich
- technische Hürden: Pakete ggf. in aufeinander abgestimmten Versionen erforderlich, oder nur in Abhängigkeit des Betriebssystems verfügbar, …

Vorteile kommen nur in konfigurierter Ausführungsumgebung zum Tragen




Beispiel für Jupyter-Notebooks zur Auswertung und Visualisierung

Einschränkung der Ergänzung von Code oder Paketen verhindert das Anpassen auf eigene Forschungsfragen, gerade für Nutzende, die sich das Feld technisch und inhaltlich erschließen möchten

**Infrastrukturbedarf einer sicheren und flexiblen Ausführungsumgebung für Jupyter-Notebooks**

für eine breite Forschungscommunity zur Verfügbarmachung, Dokumentation und Nachnutzung von Forschungsmethoden und -ergebnissen

deutsches literatur archiv marbach
HLRS High-Performance Computing Center | Stuttgart
Universität Stuttgart
Gefördert von: Baden-Württemberg MINISTERIUM FÜR WISSENSCHAFT, FORSCHUNG UND KUNST