Workshop IV: From MyBinder to JupyterHub

Enhancing Reproducibility in Computational Social Science

Speakers

Arnim Bleier (arnim.bleier@gesis.org) and

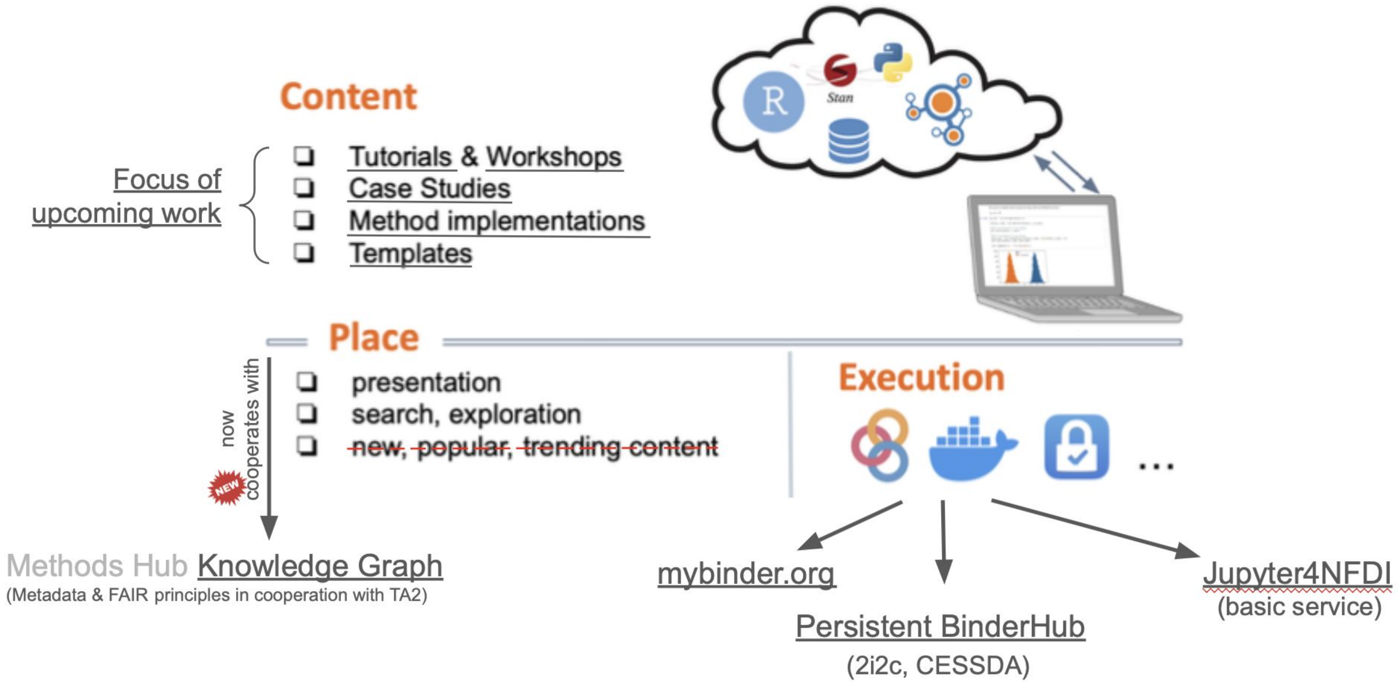
Raniere Costa da Silva (raniere.costadasilva@gesis.org)

CESSDA Conference
Split, Croatia
June 13th, 2024

Funded in parts by NFDI4DS (FKZ 460234259) and Jupyter4NFDI

Methods Hub

@gesis



Content

Focus of upcoming work

- □ Tutorials & Workshops
- □ Case Studies
- □ Method implementations
- □ Templates

Place

- □ presentation
- □ search, exploration
- □ ~~new, popular, trending content~~

NEW now cooperates with

Methods Hub Knowledge Graph

(Metadata & FAIR principles in cooperation with TA2)

Execution

...

mybinder.org

Persistent BinderHub

(2i2c, CESSDA)

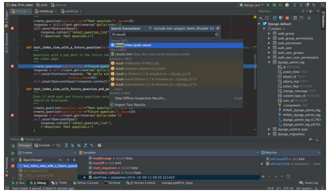
Jupyter4NFDI

(basic service)

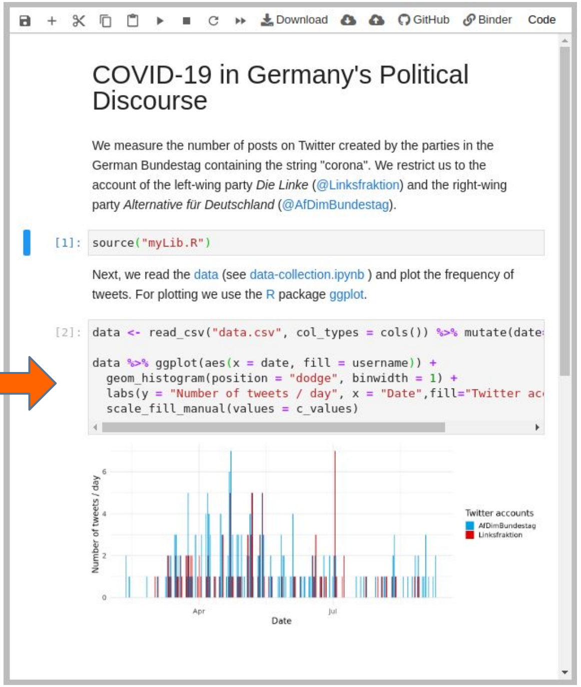
The **Methods Hub extends** and **builds upon Notebooks**. The components of GESIS Notebooks (**execution**, **place**, and **pontent**) will become part of the Methods Hub.

doi.org/10.5281/zenodo.7362188

What are Notebooks: Literate Programming



Source code



COVID-19 in Germany's Political Discourse

We measure the number of posts on Twitter created by the parties in the German Bundestag containing the string "corona". We restrict us to the account of the left-wing party *Die Linke* (@Linksfraktion) and the right-wing party *Alternative für Deutschland* (@AfDimBundestag).

```
[1]: source("myLib.R")
```

Next, we read the data (see data-collection.ipynb) and plot the frequency of tweets. For plotting we use the R package ggplot.

```
[2]: data <- read_csv("data.csv", col_types = cols()) %>% mutate(date

     data %>% ggplot(aes(x = date, fill = username)) +
       geom_histogram(position = "dodge", binwidth = 1) +
       labs(y = "Number of tweets / day", x = "Date",fill="Twitter ac
       scale_fill_manual(values = c_values)
```

Natural language

Examples:

- ❑ Jupyter
- ❑ Quarto
- ❑ Pluto.jl
- ❑ ...

Try Jupyter Notebooks (exercise)

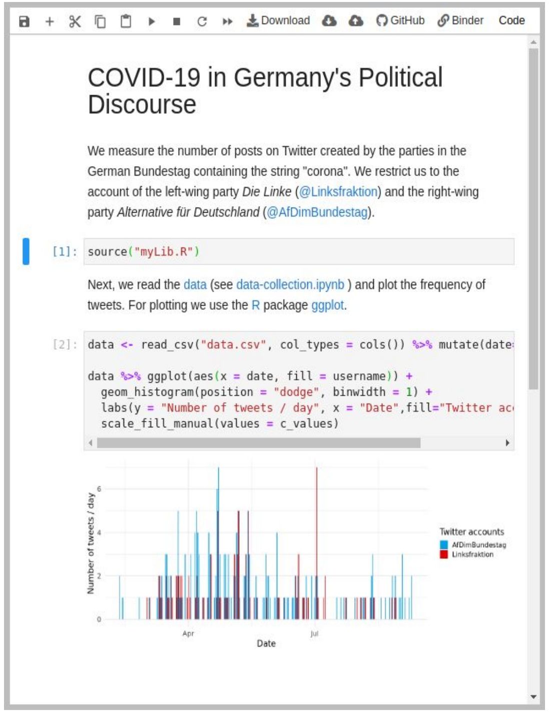
https://mybinder.org/v2/gh/arnim/RStan-Binder/master or

https://**t.ly/iTPTt**

Computation

Cloud:

❑ potentially large Data
❑ standardized environment
❑ 1-Click reproducibility



COVID-19 in Germany's Political Discourse

We measure the number of posts on Twitter created by the parties in the German Bundestag containing the string "corona". We restrict us to the account of the left-wing party *Die Linke* (@Linksfraktion) and the right-wing party *Alternative für Deutschland* (@AfDimBundestag).

```
[1]: source("myLib.R")
```

Next, we read the data (see data-collection.ipynb) and plot the frequency of tweets. For plotting we use the R package ggplot.

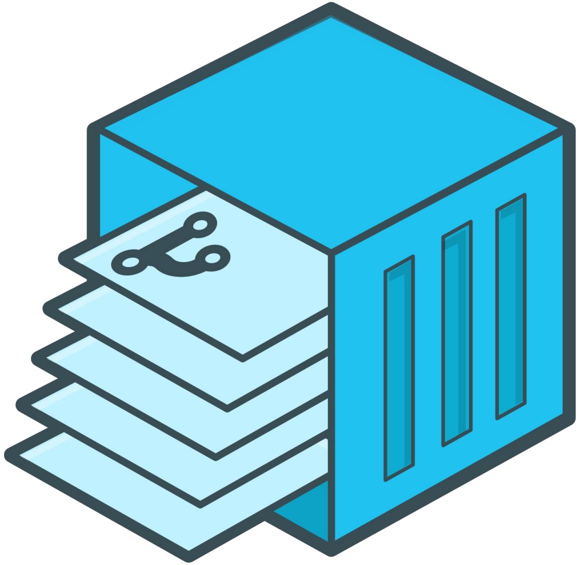
```
[2]: data <- read_csv("data.csv", col_types = cols()) %>% mutate(date

     data %>% ggplot(aes(x = date, fill = username)) +
       geom_histogram(position = "dodge", binwidth = 1) +
       labs(y = "Number of tweets / day", x = "Date",fill="Twitter ac
       scale_fill_manual(values = c_values)
```

Personal Computer:

❑ only small data
❑ every environment different
❑ time consuming to set up

Build Docker Images from a Git Repository



jupyter-repo2docker is a tool for building and running Docker images from source code repositories.

https://repo2docker.readthedocs.io

(Some) supported
Environment Configuration Files

python™

requirements.txt

```
numpy==1.13.1
matplotlib==2.0.2
seaborn==0.8.1
```

or

environment.yaml

```
name: example-environment
Channels:
 - conda-forge
dependencies:
 - python
 - numpy
```

install.R

```
install.packages("tidyverse", repos =
"https://cloud.r-project.org/",
dependencies=TRUE)
```

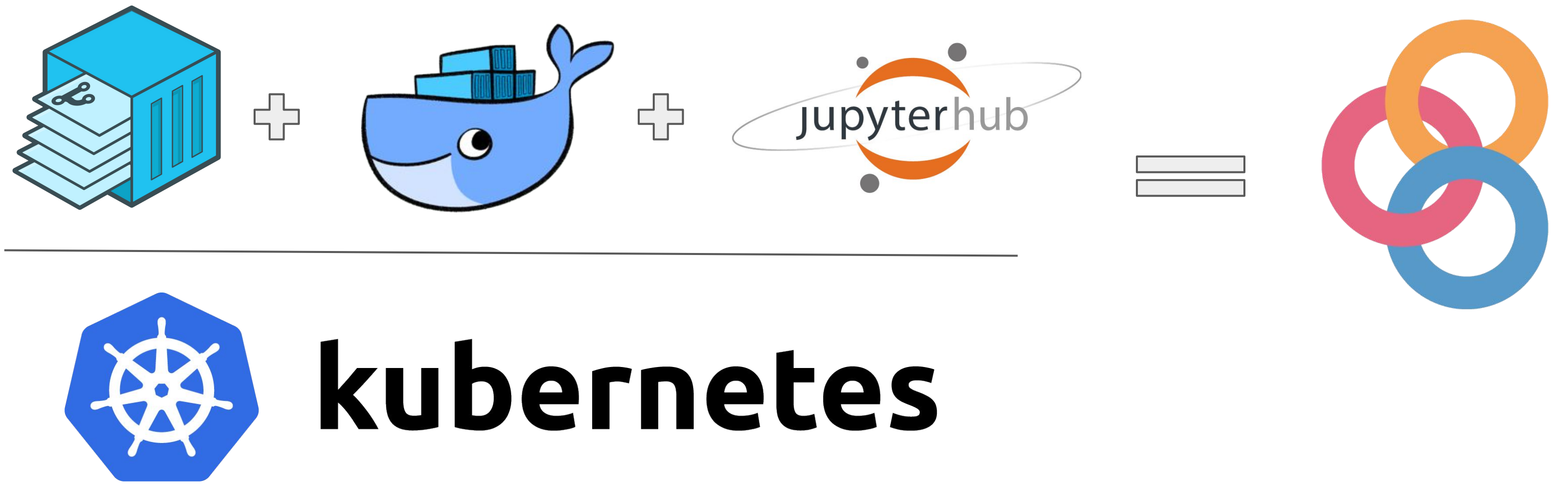
+

runtime.txt

```
r-2018-07-27
```

See https://repo2docker.readthedocs.io/en/latest/config_files.html

What is BinderHub?



https://github.com/jupyterhub/binderhub

Integrating BinderHub with JupyterHub

“binder-ready” is the de-facto standard for instant 1-click reproducible computational analysis.

JupyterHub is the de-facto standard for **long-lasting**, **persistent**, and **scalable** access to server based computational resources.

Work in cooperation with NFDI4DS, 2I2C, and CESSDA

→ https://2i2c.org/blog/2024/jupyterhub-binderhub-gesis/

Special thanks to the BinderHub Community

https://github.com/jupyterhub/binderhub/graphs/contributors

and many more who aren't in the GitHub history.

Special thanks to **Tim Head** & **The Turing Way**

for pioneering and sharing training resources

https://build-a-binder.github.io/

https://github.com/alan-turing-institute/the-turing-way/tree/main/workshops

Binderizing your repository (live demo)

Requirements to follow along

1. A laptop
2. A GitHub or GitLab account

Step 1

https://t.ly/BXSAs

How to **binderize your repository**?

Documentation of the repo2docker Configuration Files
https://repo2docker.readthedocs.io/en/latest/config_files.html

Discourse Jupyter https://discourse.jupyter.org/

Binder Examples https://github.com/binder-examples
https://github.com/binder-examples/r

Working with Jupyter & R Markdown = Jupytext
https://jupytext.readthedocs.io/en/latest/

Our WS demo repository => https://github.com/rgaiacs/2024-06-cessda-workshop-mybinder