Big Data, organisation and analysis

Hands on to build data access middleware

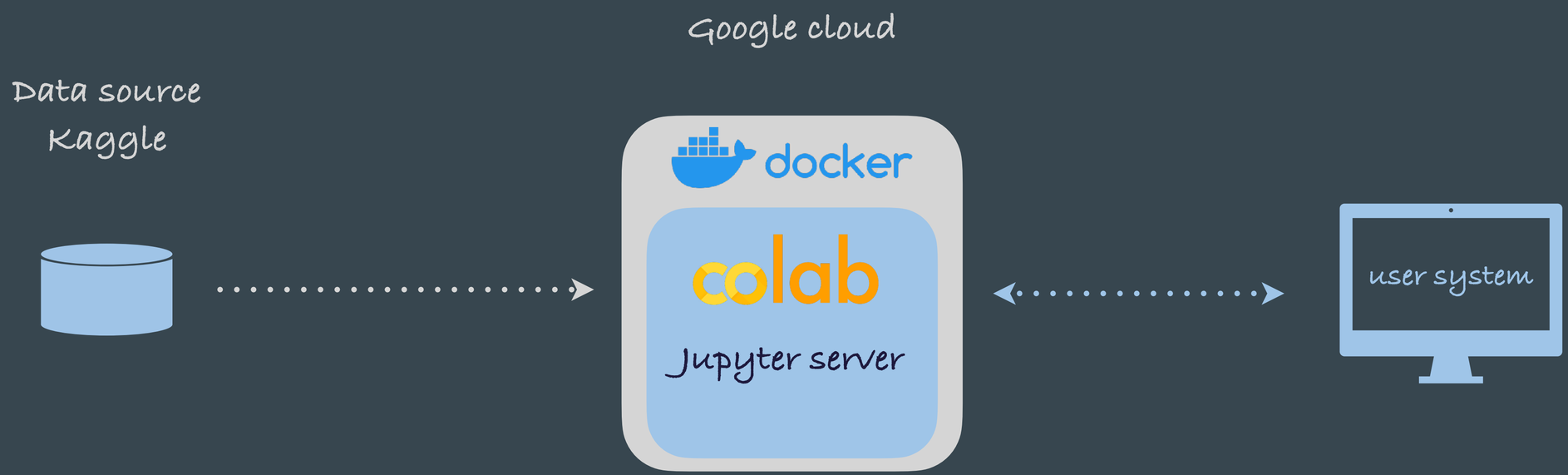
Steffen M. Noe, Spring 2024

The topics for today

- We will use Google Colab
- We will create data access middleware
-

Example 1

Use temperature data from Kaggle



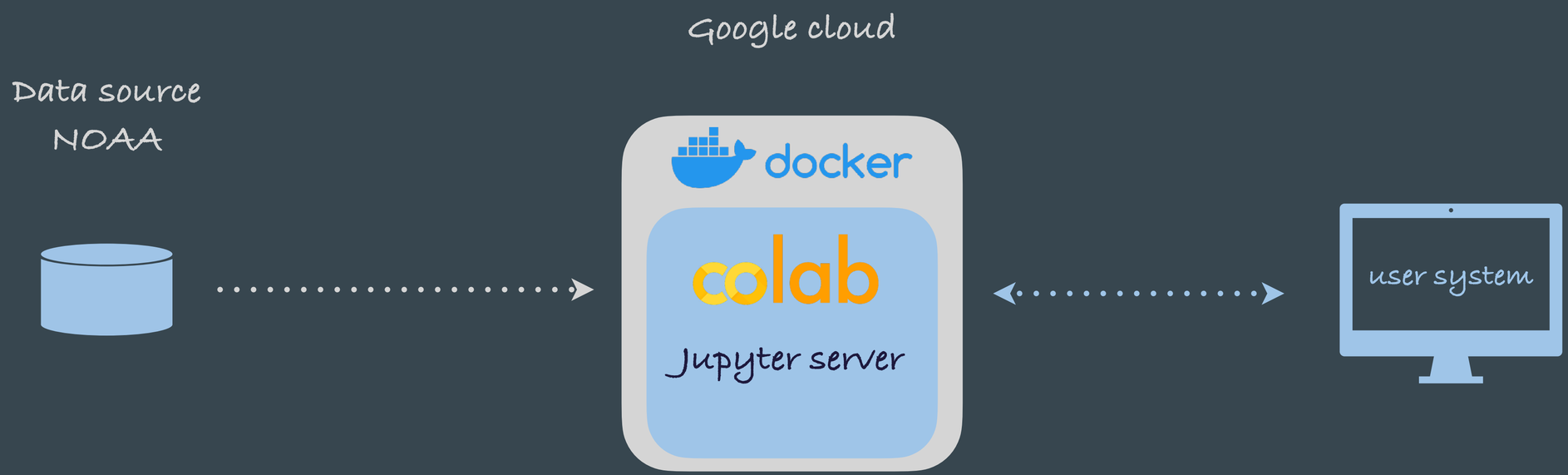
Example 1

Use Temperature data from Kaggle

- Sign up to the Kaggle system
- Make yourself a token to download data
- Upload the token to Colab
- download the city temperatures (https://www.kaggle.com/datasets/sudalairajkumar/daily-temperature-of-major-cities) via the opendata package
- Load the data into a DataFrame
- Choose a city you like
- Create a datetime column and set it as index to have a timeseries
- Visualise the timeseries

Example 2

Use CO2 data from NOAA Mauna Loa Observatory



Example 2

Use CO2 data from NOAA Mauna Loa observatory

- Check up the website to understand the data structure (https://gml.noaa.gov/ccgg/trends/data.html)
- We can read with pandas the .csv data directly (simplest solution)
- Use pandas.read_csv() to do so!
- Create a datetime column and set it as index to have again a time series (Note: it's not toatally straightforward because of your dataset!)
- Process the data to have only the monthly averages and the deseasonalised smoothed time series
- Visualise them in one plot

Example 2 alternative reading

Use CO2 data from NOAA Mauna Loa observatory

- Check up the website to understand the data structure (https://gml.noaa.gov/ccgg/trends/data.html)
- We can read any textual data via the requests package (need decoding)
- Use requests.get() to read the data
- Use io.StringIO() with decode() to load the data into a DataFrame
- Create a datetime column and set it as index to have again a timeseries (Note: it's not toatally straightforward because of your dataset!)
- Process the data to have only the monthly averages and the deseasonalised smoothed time series
- Visualise them in one plot