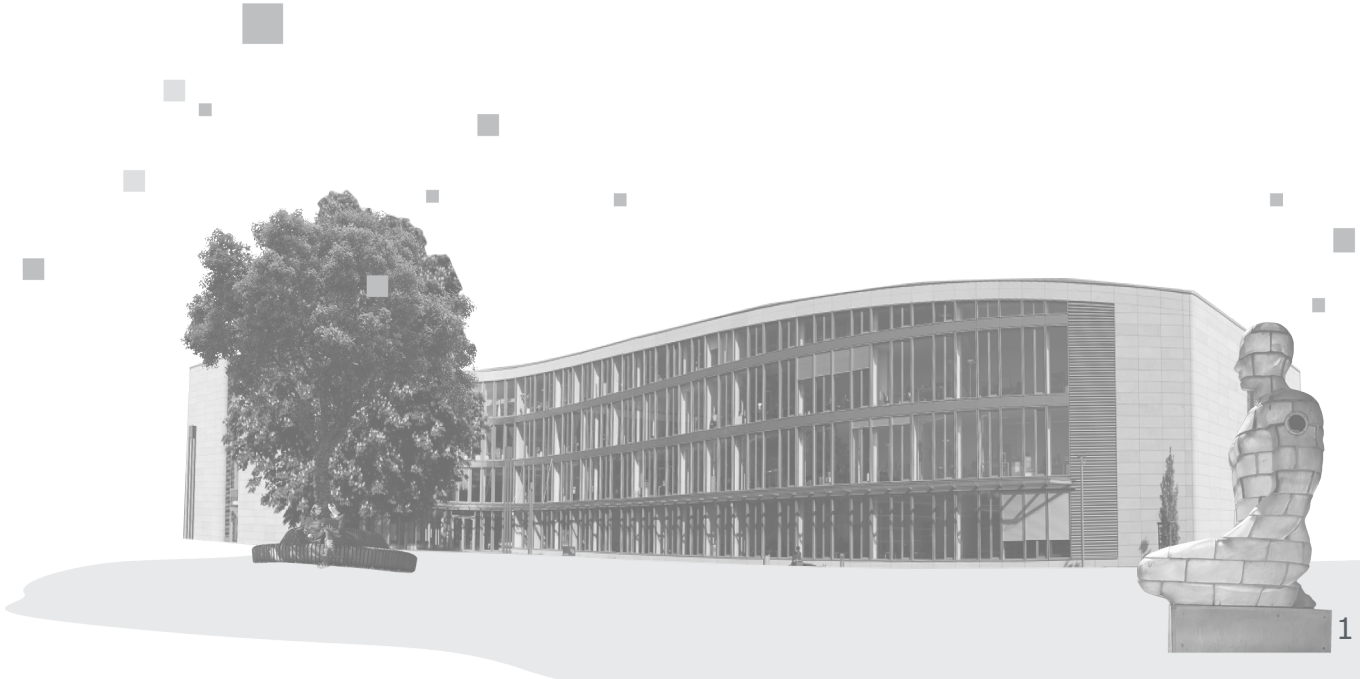
KI Service Zentrum

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Containerize your intelligence: A hands-on workshop on deploying AI models with Docker

SPONSORED BY THE
Federal Ministry of Education and Research

Felix Boelter and Johanna Reiml



Design IT.
Create Knowledge.

www.hpi.de

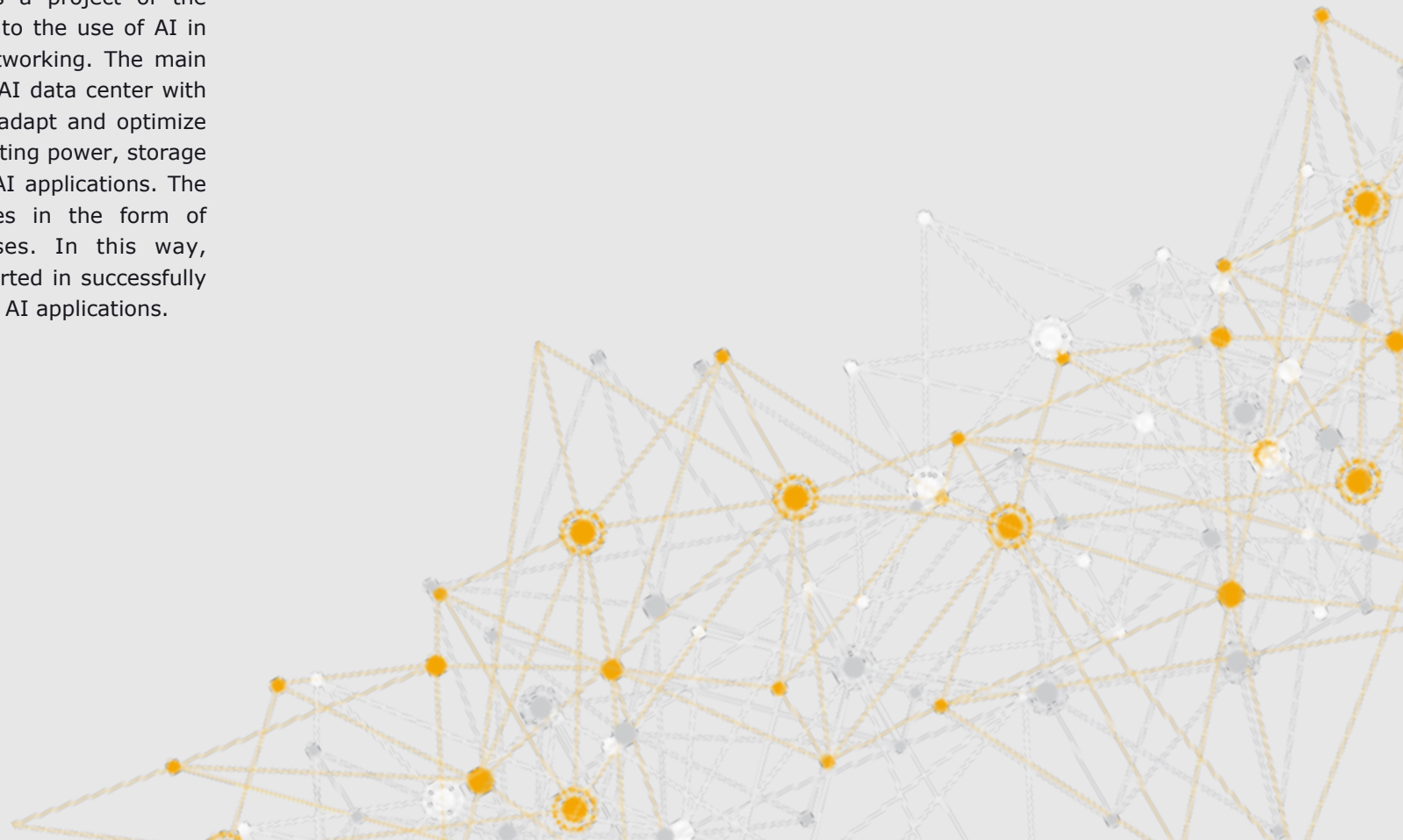
KISZ-BB

KI Service Zentrum
by Hasso-Plattner-Institut

HPI

SPONSORED BY THE
Federal Ministry of Education and Research

The AI Service Center Berlin-Brandenburg (KISZ-BB) is a project of the Hasso-Plattner-Institute with the aim of lowering barriers to the use of AI in business and society through knowledge transfer and networking. The main research areas are operational research to investigate an AI data center with heterogeneous hardware and methodological research to adapt and optimize AI models. The KISZ-BB provides resources such as computing power, storage space, data and models for the development and use of AI applications. The KISZ-BB also offers educational and consulting services in the form of workshops, individual consultations and online courses. In this way, companies, start-ups and non-profit institutions are supported in successfully mastering the next steps towards the professionalization of AI applications.



20.11.2023

KISZ-BB

AI Service Centers

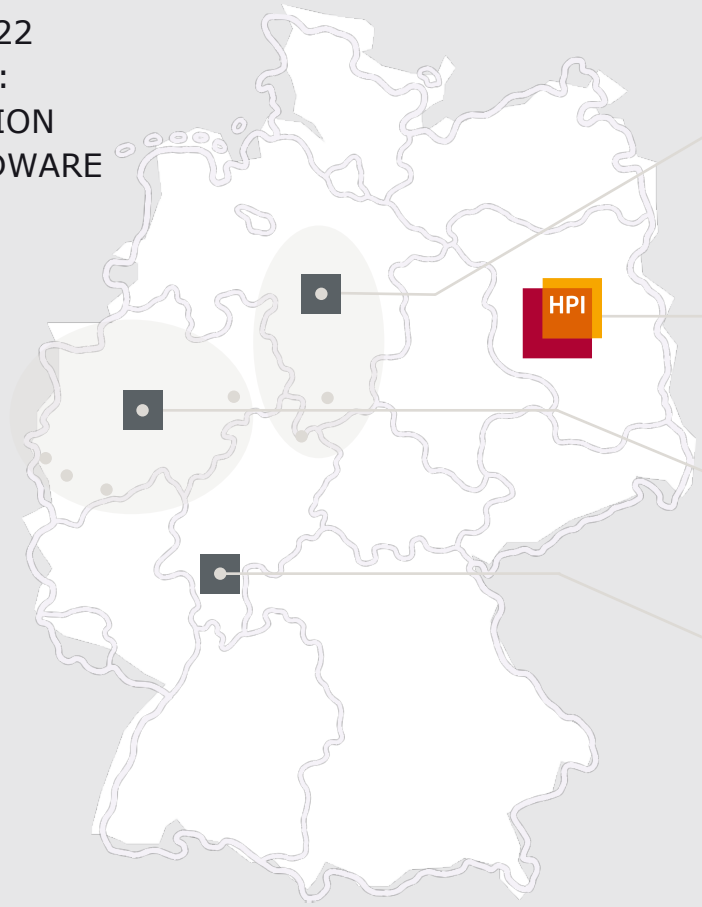
MOTIVATION, TASK & SERVICES

KI Service Zentrum
by Hasso-Plattner-Institut

HPI

SPONSORED BY THE
Federal Ministry of Education and Research

KICK-OFF: DEC 2022
FUNDING AMOUNT:
APPROX. €10 MILLION
17 PEOPLE + HARDWARE
5 SPECIALISTS



KISSKI
Hannover I Göttingen I Kassel
Sensitive and critical infrastructures
Focus: Medicine & Energy

KISZ-BB
Berlin I Brandenburg
Educational and advisory services
Use of AI in business and society

West AI
Dortmund I Bonn I Jülich I Aachen I Paderborn
Large and transferable AI models

hessian AI
Service Center
Darmstadt
Explainability, generalizability
and contextual adaptation

KI Service Zentrum by Hasso-Plattner-Institut

HPI

SPONSORED BY THE

Federal Ministry of Education and Research

Discussion

Who are you?

KI Service Zentrum by Hasso-Plattner-Institut

HPI

SPONSORED BY THE

Federal Ministry of Education and Research

Learning Goals

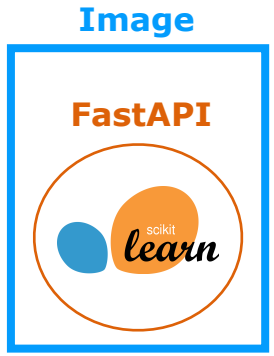
- Understand how we can use a **basic language detection model**
- Understand **APIs** and **FastAPI**
- Understand essential **Docker features**
- Be able to **deploy a basic ML application**

What we will build

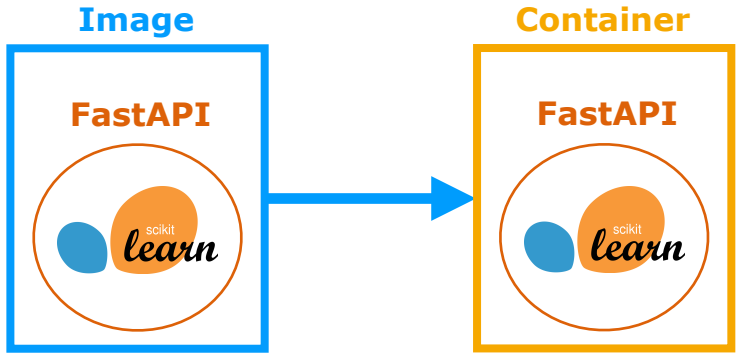
SPONSORED BY THE Federal Ministry of Education and Research

Step 1:
Create Model API

Step 2:
Create Docker Image



Step 3:
Build a Container and Deploy it



SPONSORED BY THE Federal Ministry of Education and Research

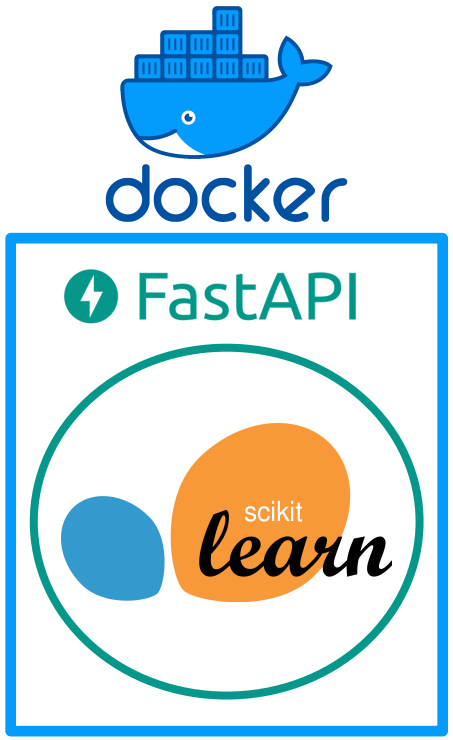
What we will build

Input:
Guten Morgen

Send request via web interface →

Output:
German

← Receive response via web interface



KI Service Zentrum by Hasso-Plattner-Institut

HPI

Hands-On: Setting up the repository

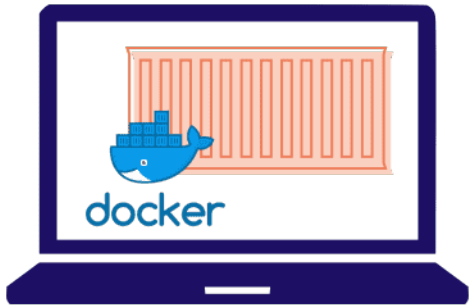
SPONSORED BY THE
Federal Ministry of Education and Research

- Open your terminal
- Navigate to a directory for storing the workshop code using the `cd` command
- **Clone the repository:**
  - `git clone https://github.com/KISZ-BB/kisz-mlops-docker.git`
- **Navigate to the repository using:**
  - `cd kisz-mlops-docker`
- Check the contents with `ls` (Linux/Mac) or `dir` (Windows)
- Look at the `README.md` and the `handout.pdf`

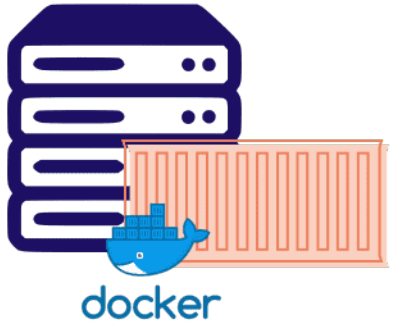
Part 1: Introduction to Docker

Background Talk: What is Docker?

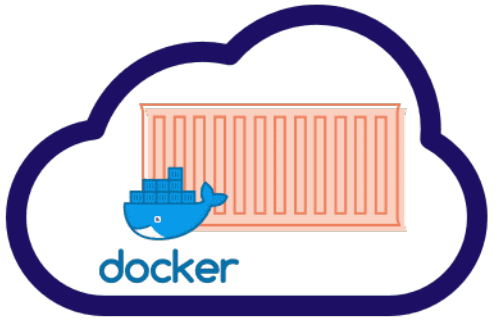
SPONSORED BY THE Federal Ministry of Education and Research



Works well on my laptop

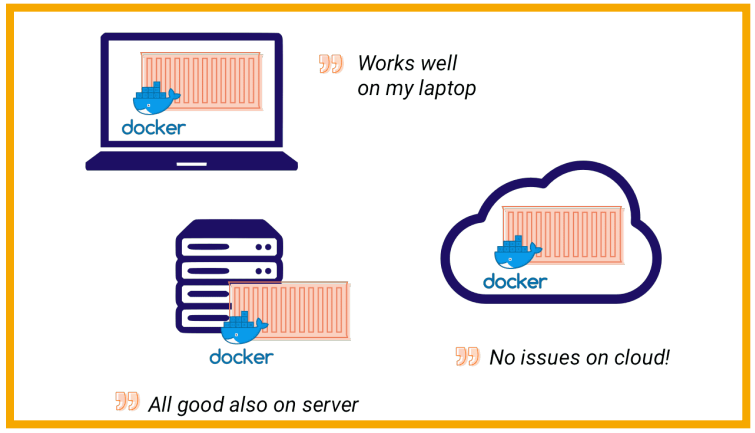


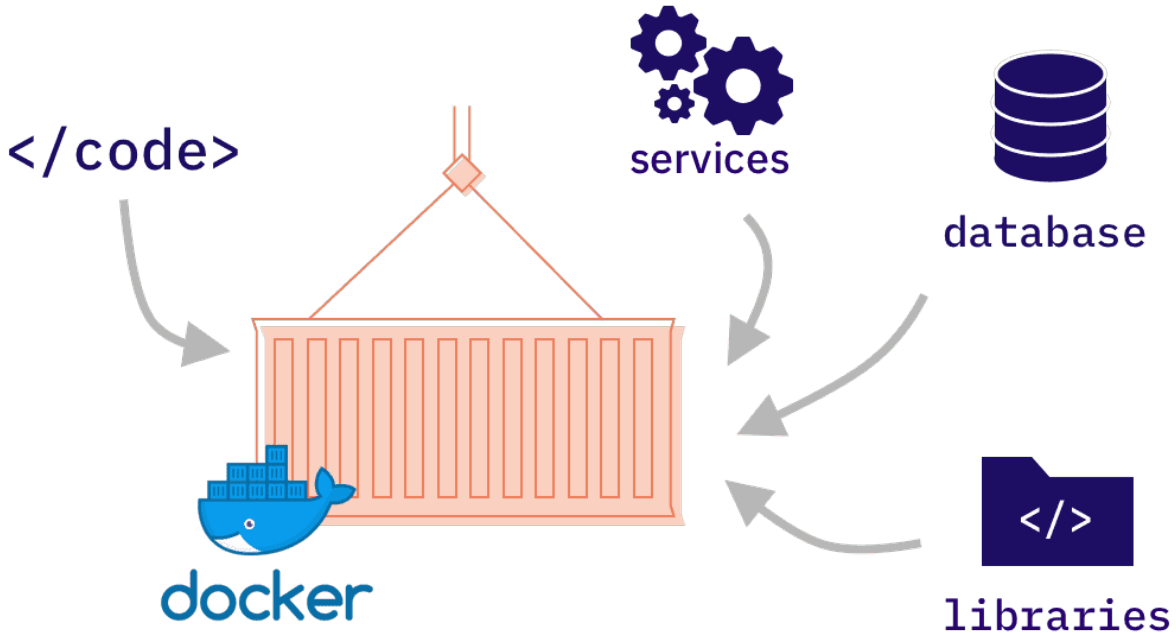
All good also on server



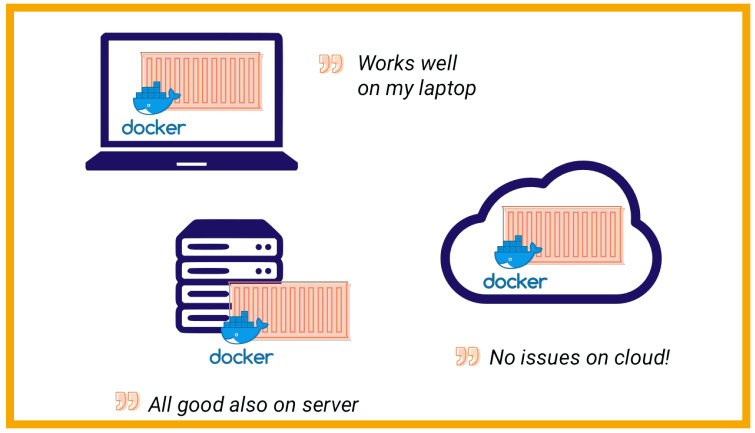
No issues on cloud!

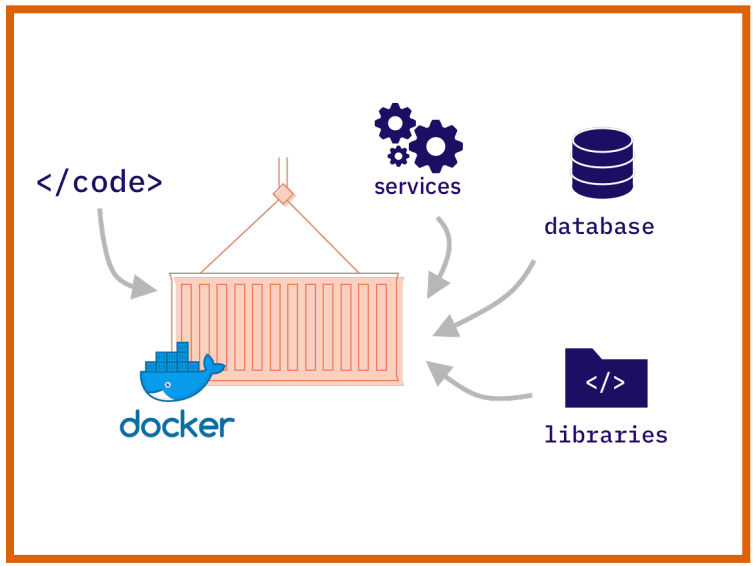
Background Talk: What is Docker?

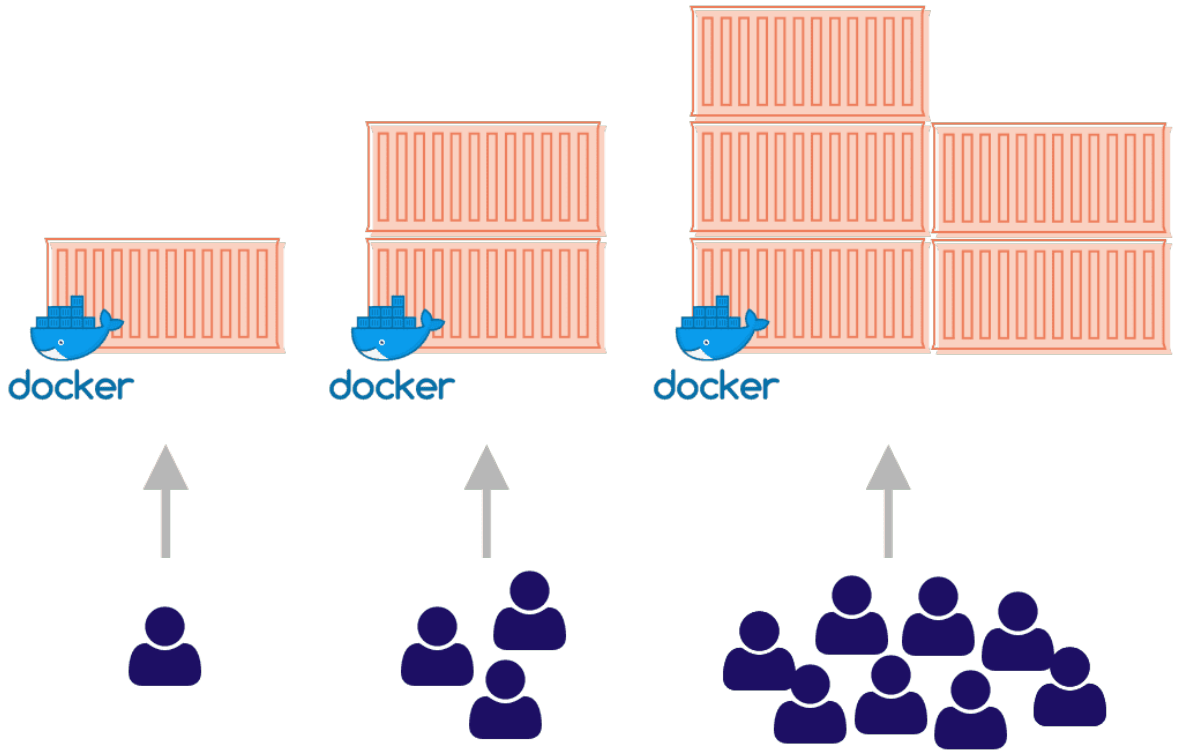




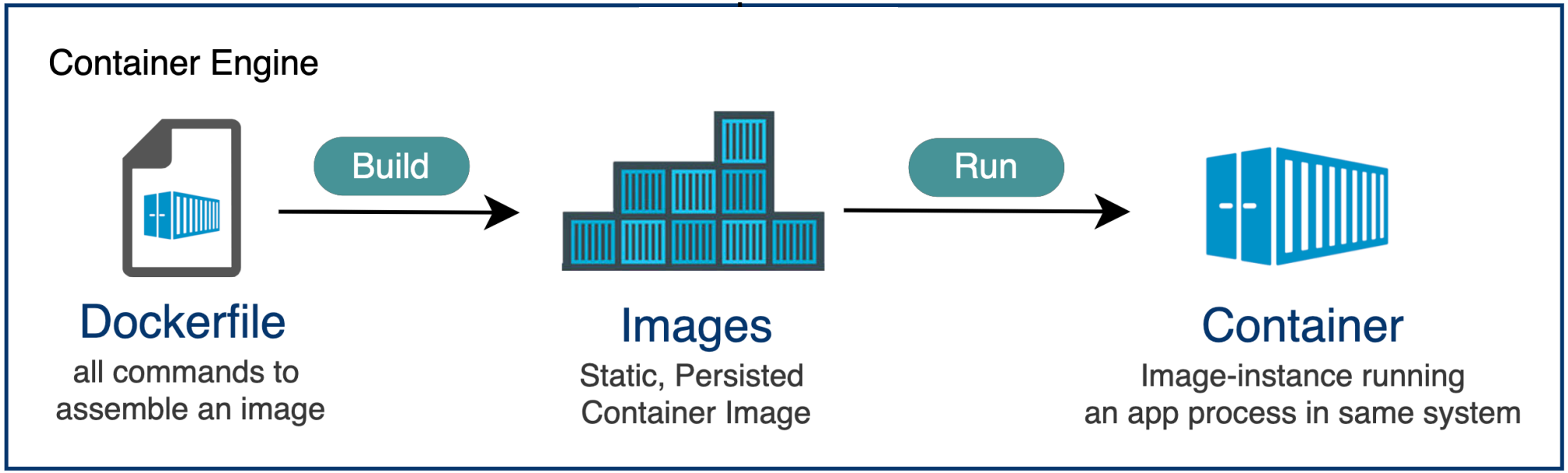
Background Talk: What is Docker?



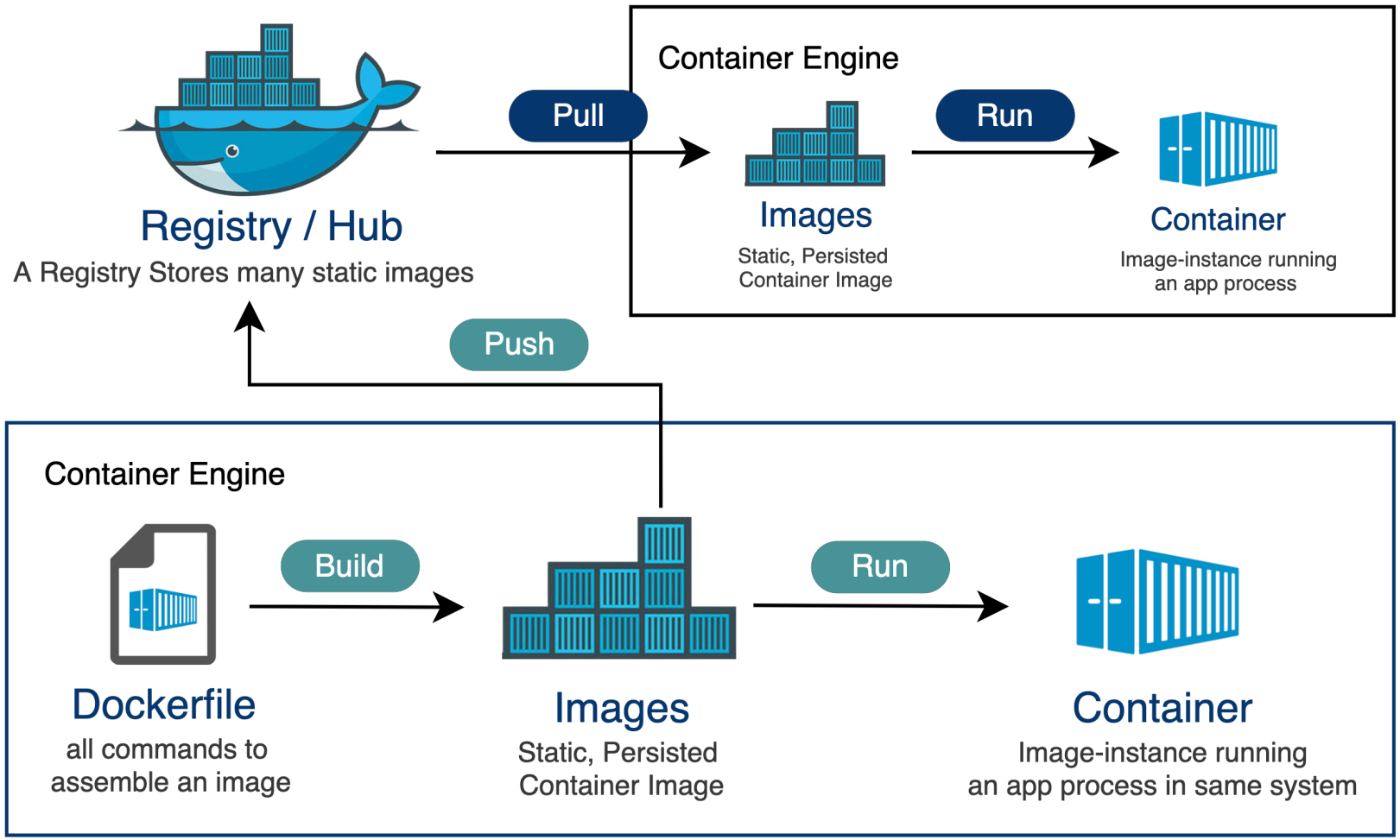




Background Talk: How does it work?



Docker Hub and Registries



Hands-On: Hello-World toy example

SPONSORED BY THE
Federal Ministry of Education and Research

Goal: Pulling an image, creating a container and removing the container.

Find a detailed description of the steps in the GitHub repository under "Hands-On 1: Hello-World toy example".

Discussion

SPONSORED BY THE
Federal Ministry of Education and Research

How was your experience?


Break

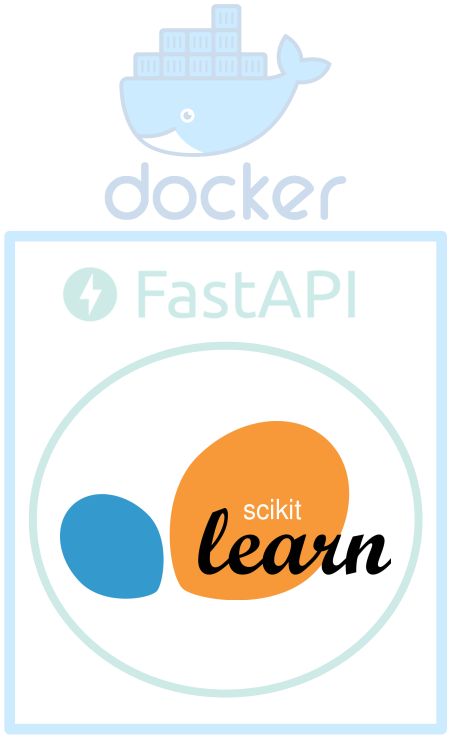
Part 2: Creating the ML Model

What we will build

SPONSORED BY THE
Federal Ministry of Education and Research

Input:
Guten Morgen

Send request via web interface



Output:
German

Receive response via web interface

KI Service Zentrum by Hasso-Plattner-Institut
HPI
SPONSORED BY THE
Federal Ministry of Education and Research

Using the Model

Language Detection Pipeline

SPONSORED BY THE Federal Ministry of Education and Research

Input

A text for which we want to determine the language.

Machine learning model

Output

The best matching language.

Discussion: What is Jupyter?

SPONSORED BY THE
Federal Ministry of Education and Research

KI Service Zentrum by Hasso-Plattner-Institut
HPI
SPONSORED BY THE
Federal Ministry of Education and Research

Training & Testing the Model

Hands-On: Running Jupyter with Docker

Goal: Running Jupyter Notebook in a Docker container

Find a detailed description of the steps in the GitHub repository under "Hands-On 2: Model Training with Jupyter "

Discussion

SPONSORED BY THE
Federal Ministry of Education and Research

How was your experience?


Break

Part 3: Creating the ML API

SPONSORED BY THE Federal Ministry of Education and Research

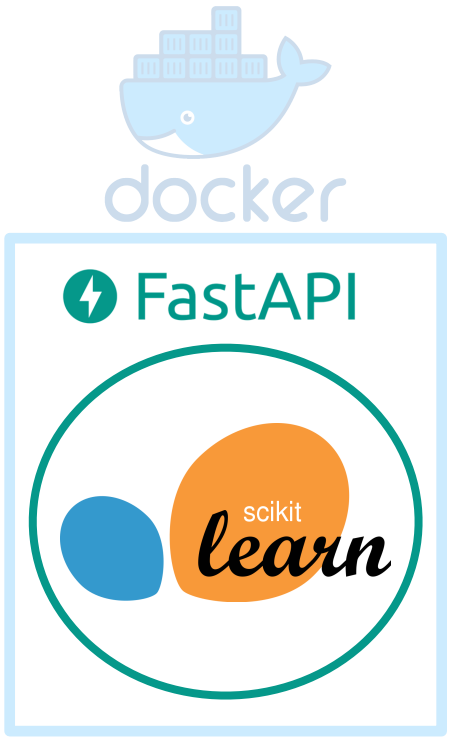
What we will build

Input:
Guten Morgen

Send request via web interface →

Output:
German

← Receive response via web interface

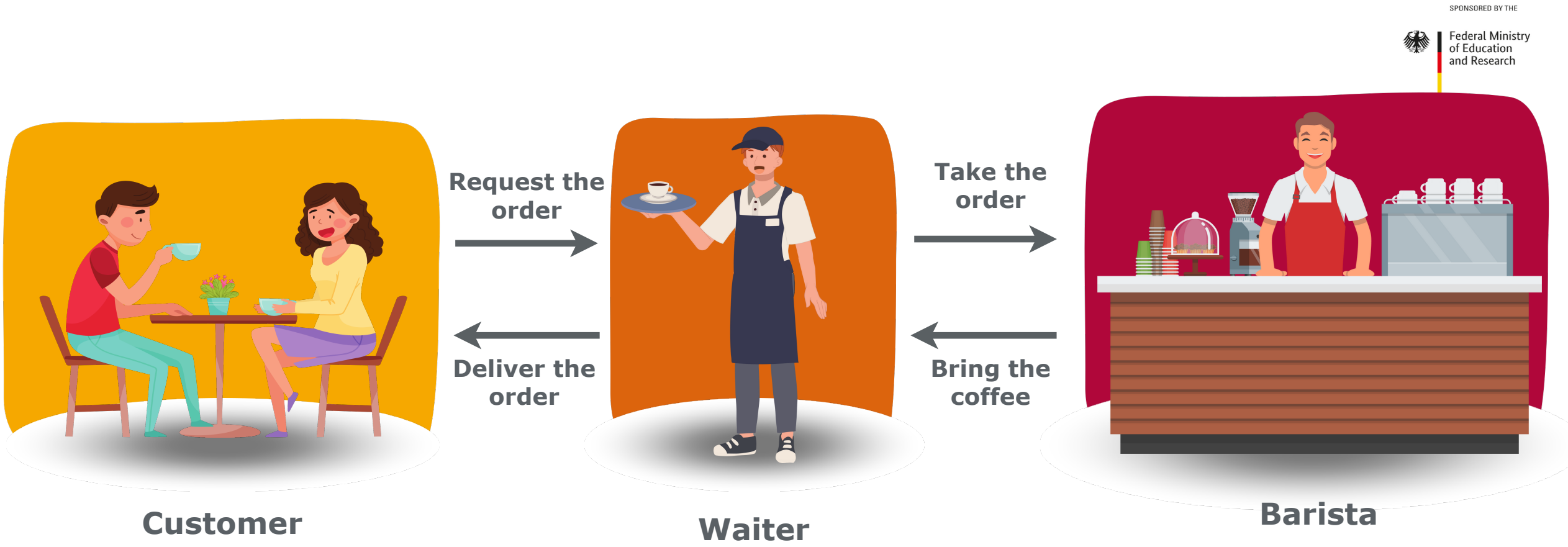


Introduction to APIs

SPONSORED BY THE
Federal Ministry of Education and Research

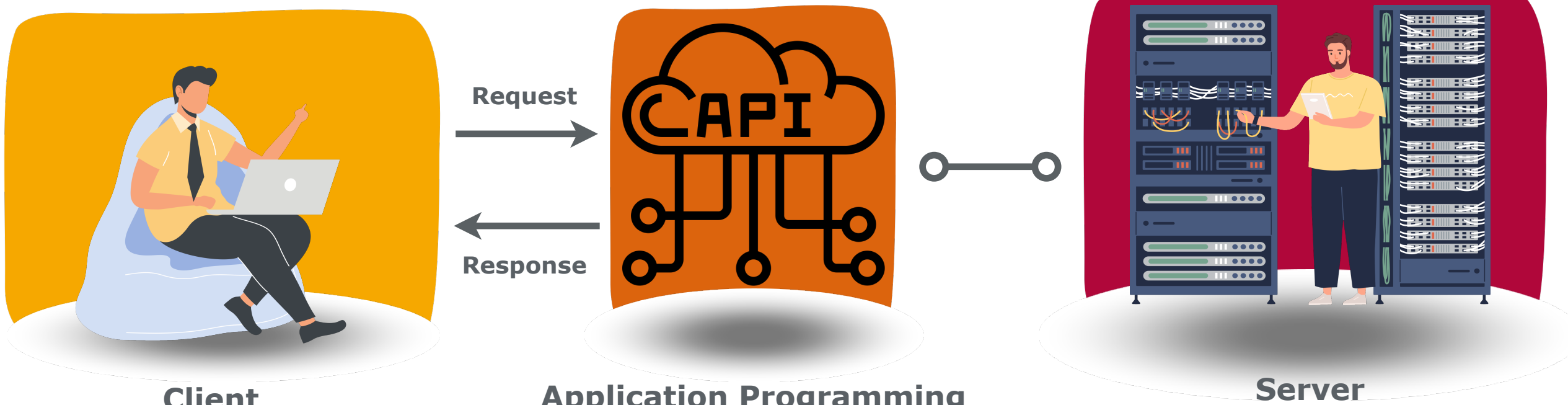
Introduction to APIs

SPONSORED BY THE
Federal Ministry of Education and Research



Introduction to APIs

SPONSORED BY THE
Federal Ministry of Education and Research

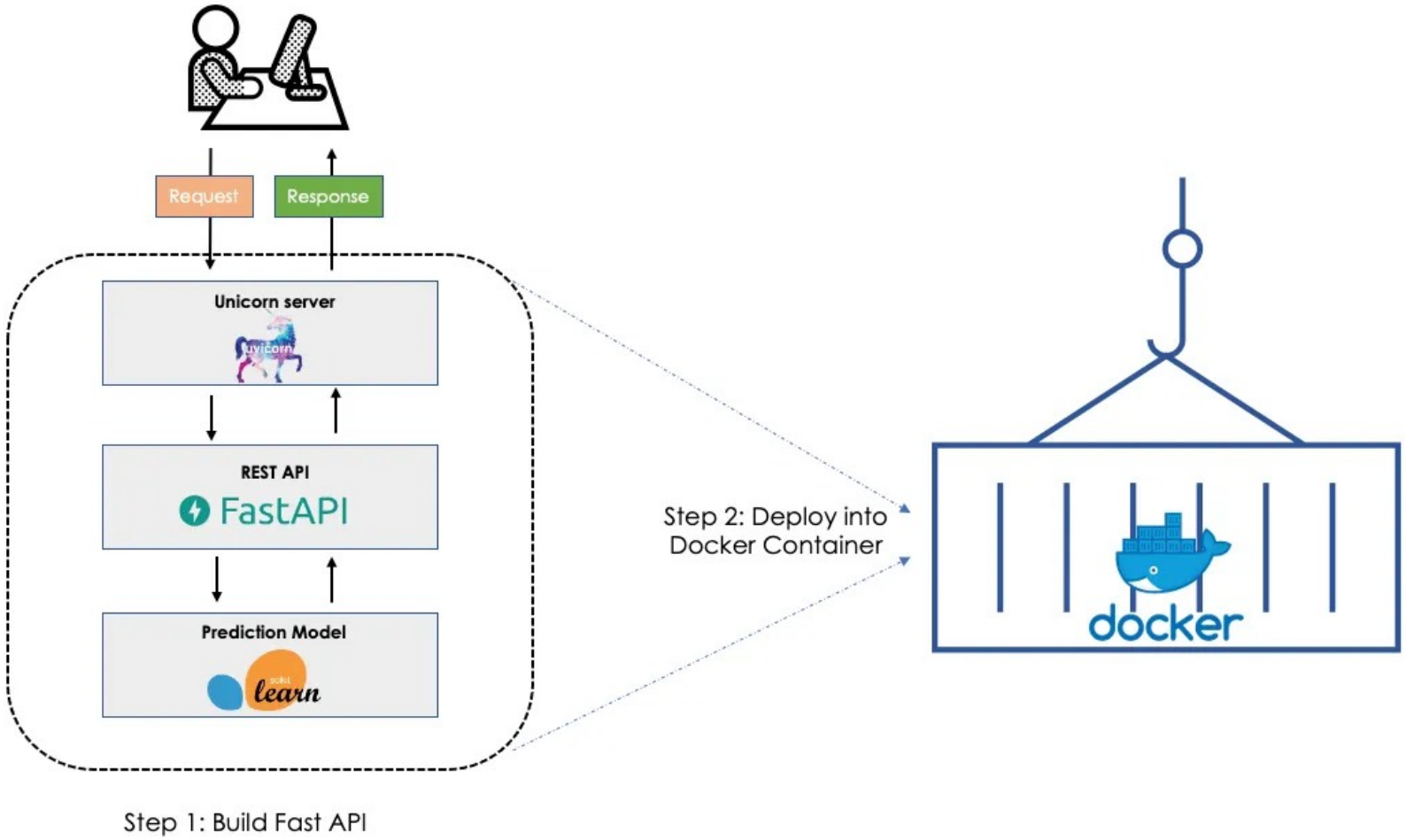


SPONSORED BY THE
Federal Ministry of Education and Research

API calls with FastAPI

Introduction to Web APIs for Machine Learning

SPONSORED BY THE Federal Ministry of Education and Research



KI Service Zentrum by Hasso-Plattner-Institut

HPI

SPONSORED BY THE

Federal Ministry of Education and Research

Simplest FastAPI file

```python
from fastapi import FastAPI

app = FastAPI()


@app.get("/")
async def root():
    return {"message": "Hello World"}
```

Hands-On: Our FastAPI file

SPONSORED BY THE Federal Ministry

```
model_api = FastAPI()

class Result(BaseModel):
    language : str

def get_detection_model():
    full_detection_model = joblib.load("model/multinomial_language_detector.joblib")
    return full_detection_model

# TODO: Create a post request with the path "/predict" and the Result response_model.
@model_api.post(. . . .)
async def predict(input_text, full_detection_model = Depends(get_detection_model)):
    detection_model, cv = full_detection_model
    vectorized = cv.transform([input_text])
    language_prediction = detection_model.predict(vectorized)[0]
    return Result(language = language_prediction)
model_api.mount("/", StaticFiles(directory="app/static", html=True), name="static")
```

Discussion

SPONSORED BY THE
Federal Ministry of Education and Research

How was your experience?


Break

Part 4: Containerizing and Deploying the ML Model API

What we will build

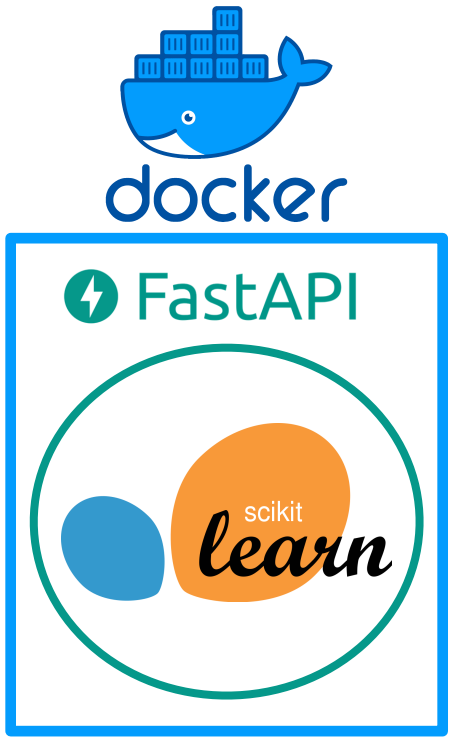
SPONSORED BY THE Federal Ministry of Education and Research

Input:
Guten Morgen

Send request via web interface →

Output:
German

← Receive response via web interface



The Dockerfile: Important Commands

SPONSORED BY THE Federal Ministry of Education and Research

FROM
Specifies a base image
`FROM python:3.11`

WORKDIR
Sets the working directory
`WORKDIR /app`

COPY
Copies files to image
`COPY ./ /app`

RUN
Executes terminal commands
`RUN pip install -r requirements.txt`

CMD
Initial container command. One per Dockerfile.
`CMD ["uvicorn", "app.api:model_api", "--host", "0.0.0.0"]`

Hands-On: Building a Docker Image

SPONSORED BY THE
Federal Ministry of Education and Research

Build: `docker build -t mlops/language-detection:v1 .`

Start: `docker run -p 8000:8000 mlops/language-detection:v1`

Dockerfile

FROM Python Base Image

WORKDIR in the container set to **`/app`**

COPY the **`requirements.txt`** file into the container

RUN pip install for packages specified in **`requirements.txt`**

COPY the rest of the application into the container

CMD for starting the uvicorn server through the terminal command:
`uvicorn app.location:api_name --host 0.0.0.0`

Discussion

SPONSORED BY THE
Federal Ministry of Education and Research

What tedious aspects of the build process did you observe?

SPONSORED BY THE Federal Ministry of Education and Research

What we noticed

- Slow build times due to big Docker images
- Running the container in attached mode is not suitable for production
- Using individual `build` and `run` commands is not useful in a production setting

Running the Container with Docker Compose

Build: `docker build -t mlops/language-detection:v1 .`
Start: `docker run -p 8080:8000 mlops/language-detection:v1`

Simpler:

docker compose up --build

```yaml
version: "3"

services:
  api:
    image: mlops/language-detection:v1
    build:
      context: .
    ports:
      - "8000:8000"
    # volumes:
    #   - .:/app
```

KI Service Zentrum by Hasso-Plattner-Institut

HPI

Live-Demo: Deploying the Container with Docker Compose

SPONSORED BY THE
Federal Ministry of Education and Research

1. **Problem: Container shuts down when terminal is closed**
2. Run the container in detached mode: `docker compose up -d`

3. **Problem: Container does not restart when the machine restarts**
4. Add `restart: always` to the `compose.yaml` file

5. **Problem: The application has errors and I need to see the logs**
6. Check that the container is running by running `docker ps`
7. Check the logs of the specific container using: `docker logs YOUR_CONTAINER_ID`

8. Close the container using: `docker compose down`

Discussion

SPONSORED BY THE
Federal Ministry of Education and Research

How can we make the image smaller?

KI Service Zentrum by Hasso-Plattner-Institut

HPI

Competition: Who can create the smallest image size?

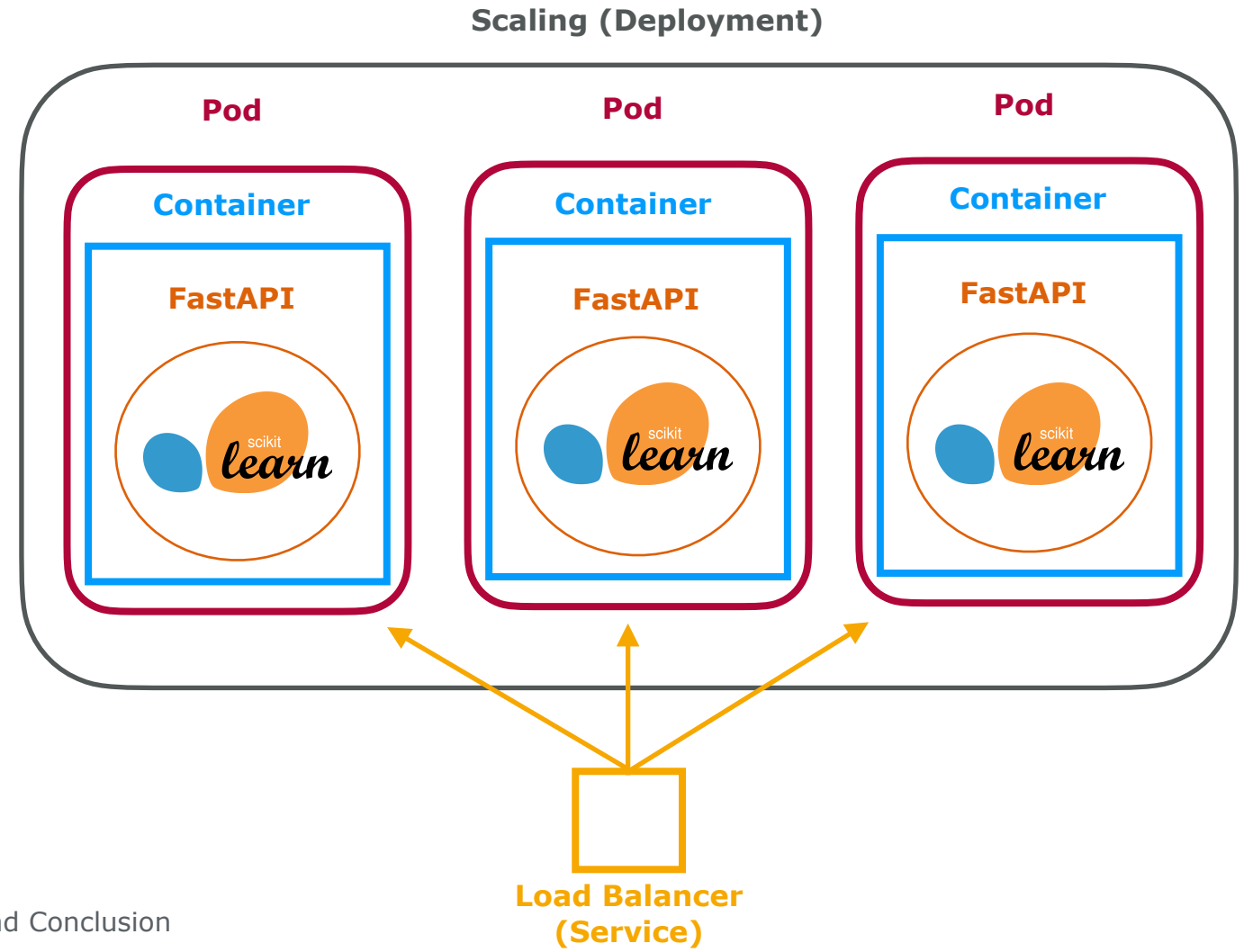
SPONSORED BY THE
Federal Ministry of Education and Research

Pair up in teams of two and work together to create the smallest possible image size.

- Use a smaller base image
- Install only what is necessary for production
- Create a `.dockerignore` file to ignore unnecessary files
- Delete any cache files that may have been created

15:00

Outlook: Scaling and Load Balancing with Kubernetes



Conclusion

- ✓ Understand how we can use a **basic language detection model**
- ✓ Understand **FastAPI** and **APIs**
- ✓ Understand essential **Docker features**
- ✓ Be able to **deploy a basic ML application**

Feedback and Q&A

forms.gle/JU6gsnDQhsY585mAA

SPONSORED BY THE
Federal Ministry of Education and Research

Docker Mount Types

SPONSORED BY THE
Federal Ministry of Education and Research

Volume Mount
Docker managed file system

Persistent Data Storage

Container Data Exchange

Performance Benefits

Bind Mount
Host file system

Real-time code changes

Access to System Files

Configuration Management

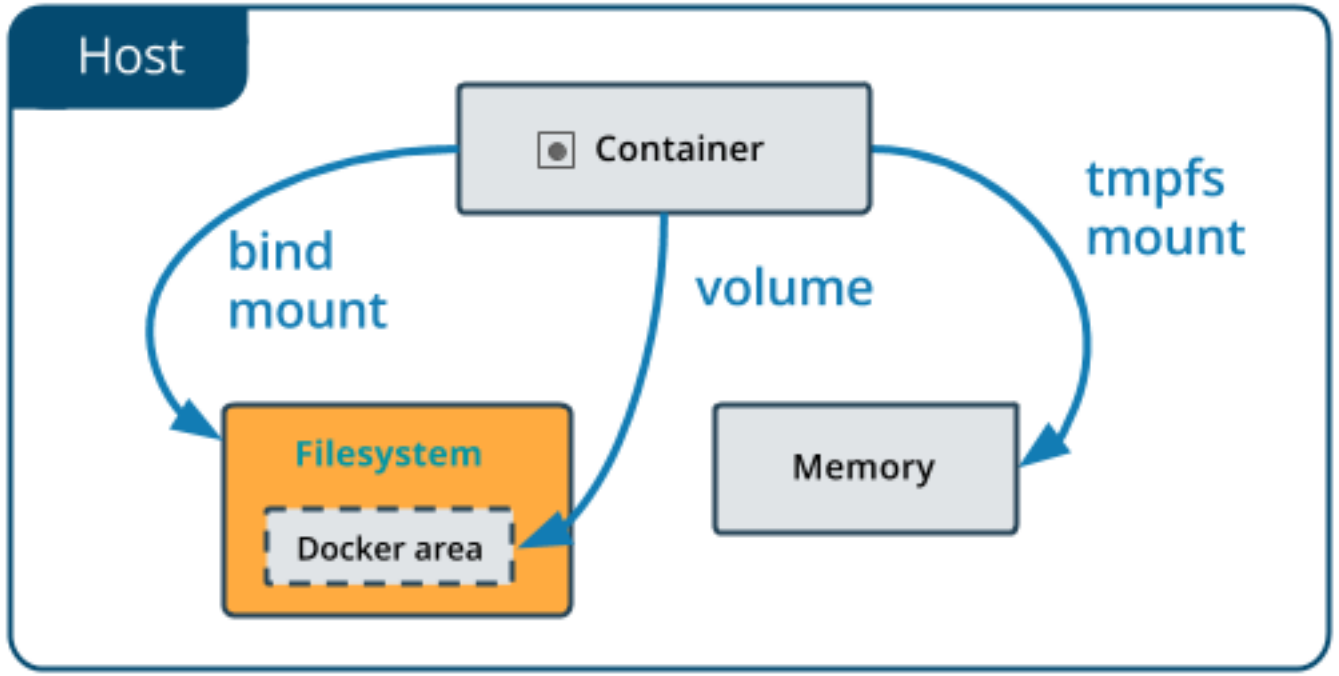
tmpfs Mount
Host memory

Non-persistent Data

Fast Data Processing

In-Memory Caching

Docker Host Interaction



Hands-On: Bind Mounts

- **Problem:** The container has no access to our files
- **Next steps:**
  - Make sure we can access the `model` directory in our Jupyter notebook container
  - Start a Jupyter notebook with access to the data pre-processing and training code