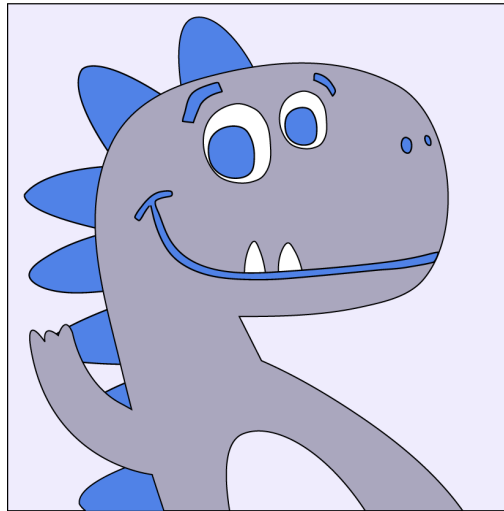


TIRA for IR: Theory and Hands-On Tutorial

Towards Shared Tasks in IR Courses



Augsburg IR Lab, 6th June, 2024

Maik Fröbe

University of Jena

@webis_de

www.webis.de

TIRA for IR: Theory and Hands-On Tutorial

Motivation

Michael Granitzer

Leiter OpenWebSearch.eu



"I want to
choose my
search engine
like my daily
newspaper"



TIRA for IR: Theory and Hands-On Tutorial

Motivation



Michael Granitzer
Leiter OpenWebSearch.eu

"I want to choose my search engine like my daily newspaper"

 open search foundation

Open Search Foundation

- ❑ Joint EU project
- ❑ Open Web Index to foster competition
- ❑ Shared tasks and data challenges planned

TIRA for IR: Theory and Hands-On Tutorial

Motivation



Michael Granitzer
Leiter OpenWebSearch.eu

"I want to choose my search engine like my daily newspaper"

 open search foundation

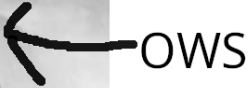
Open Search Foundation

- ❑ Joint EU project
- ❑ Open Web Index to foster competition
- ❑ **Shared tasks** and data challenges planned

TIRA for IR: Theory and Hands-On Tutorial

Best Case

Your Search Engine



TIRA for IR: Theory and Hands-On Tutorial

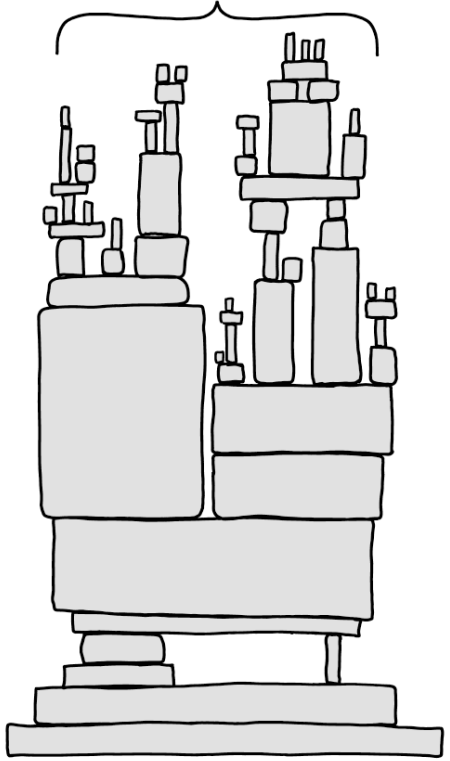
Best Case

Your Search Engine



Worst Case

Your Search Engine



TIRA for IR: Theory and Hands-On Tutorial

Best Case

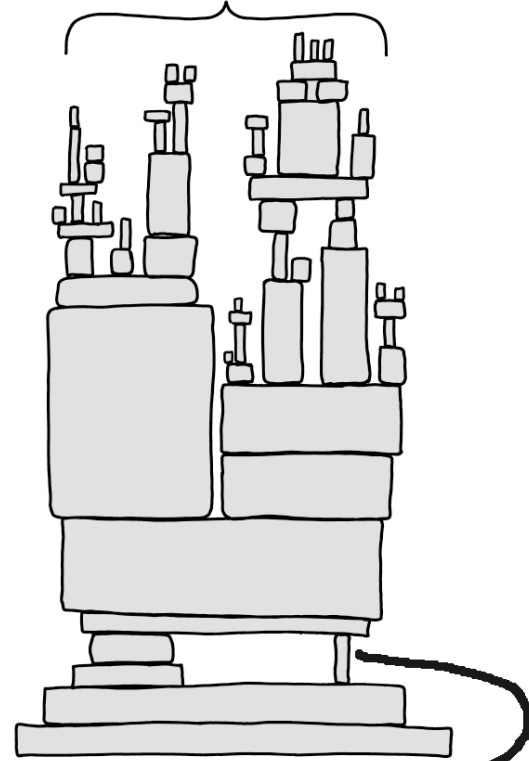
Worst Case

Your Search Engine

Your Search Engine



OWS



Potential problems:

[Fuhr'21]

- ❑ Problem 1: Internal validity
- ❑ Problem 2: External validity

TIRA for IR: Theory and Hands-On Tutorial

Best Case

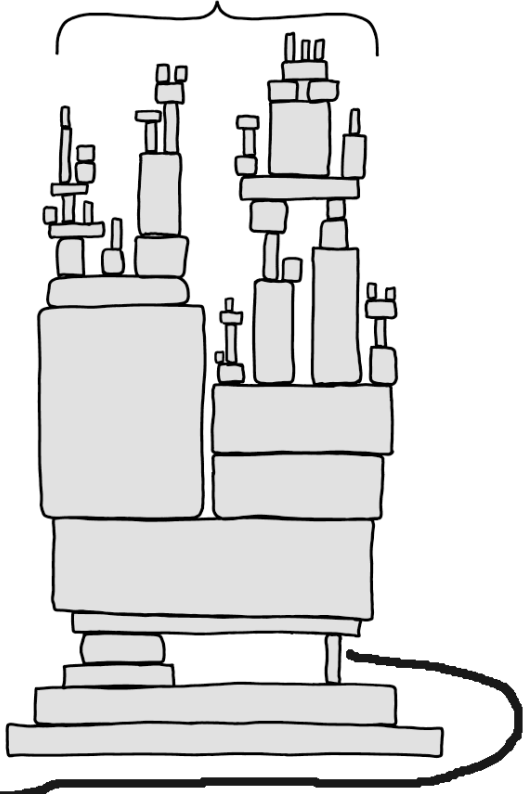
Worst Case

Your Search Engine

Your Search Engine



OWS



Potential problems:

[Fuhr'21]

- ❑ Problem 1: Internal validity
- ❑ Problem 2: External validity
- ❑ Problem 3: Blinded experimentation with LLMs

TIRA for IR: Theory and Hands-On Tutorial

Problem 1: Internal Validity [Fuhr'21]

Goal

The hypothesis is supported by the data.

TIRA for IR: Theory and Hands-On Tutorial

Problem 1: Internal Validity [Fuhr'21]

Goal

The hypothesis is supported by the data.

Possible problems

- ❑ Wrong baseline
[Armstrong'09,Lin'18]
- ❑ Formulate hypothesis after experiments
[Fuhr'21]

TIRA for IR: Theory and Hands-On Tutorial

Problem 1: Internal Validity [Fuhr'21]

Goal

The hypothesis is supported by the data.

Possible problems

- ❑ Wrong baseline
[Armstrong'09,Lin'18]
- ❑ Formulate hypothesis after experiments
[Fuhr'21]

Possible solutions

- ❑ Centralized leaderboards
 - E.g., Run uploads to EvaluateIR
[Armstrong'09]
- ❑ Task-specific leaderboards
 - E.g., MS MARCO, MIRACL
[Lin'22,Zhang'22]

TIRA for IR: Theory and Hands-On Tutorial

Problem 1: Internal Validity [Fuhr'21]

Goal

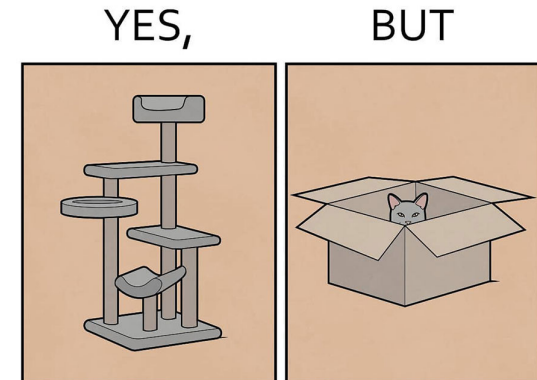
The hypothesis is supported by the data.

Possible problems

- ❑ Wrong baseline [Armstrong'09,Lin'18]
- ❑ Formulate hypothesis after experiments [Fuhr'21]

Possible solutions

- ❑ Centralized leaderboards
 - E.g., Run uploads to EvaluateIR [Armstrong'09]
- ❑ Task-specific leaderboards
 - E.g., MS MARCO, MIRACL [Lin'22,Zhang'22]



TIRA for IR: Theory and Hands-On Tutorial

Problem 1: Internal Validity [Fuhr'21]

Goal

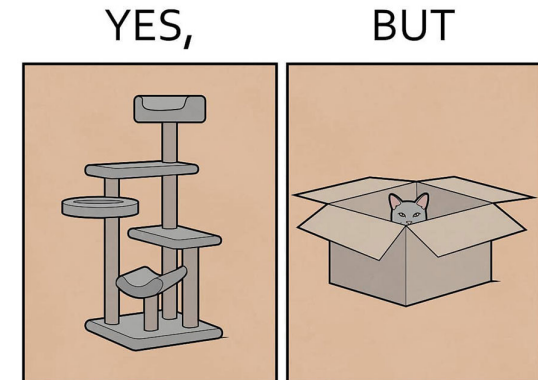
The hypothesis is supported by the data.

Possible problems

- ❑ Wrong baseline [Armstrong'09,Lin'18]
- ❑ Formulate hypothesis after experiments [Fuhr'21]

Possible solutions

- ❑ Centralized leaderboards
 - E.g., Run uploads to EvaluateIR [Armstrong'09]
- ❑ Task-specific leaderboards
 - E.g., MS MARCO, MIRACL [Lin'22,Zhang'22]



“EvaluateIR never gained traction, and a number of similar efforts following it have also floundered” [Lin'18]

TIRA for IR: Theory and Hands-On Tutorial

Problem 2: External Validity [Fuhr'21]

Goal

Repeating an experiment on similar data yields similar observations.

TIRA for IR: Theory and Hands-On Tutorial

Problem 2: External Validity [Fuhr'21]

Goal

Repeating an experiment on similar data yields similar observations.

Possible problems

- ❑ Non-reproducible results

TIRA for IR: Theory and Hands-On Tutorial

Problem 2: External Validity [Fuhr'21]

Goal

Repeating an experiment on similar data yields similar observations.

Possible problems

- ❑ Non-reproducible results

Possible Solutions

- ❑ TREC Open Runs
[Voorhees'16]
- ❑ Reproducibility initiatives
 - OSIRRC: Archive artifacts
[Arguello'15, Clancy'19]
 - CENTRE: Reimplementation
[Ferro'19, Sakai'19]
- ❑ Platforms + documentation
 - CodaLab, EvalAI, PRIMAD, STELLA, TIRA
- ❑ Meta evaluations: BEIR
[Thakur'21]

TIRA for IR: Theory and Hands-On Tutorial

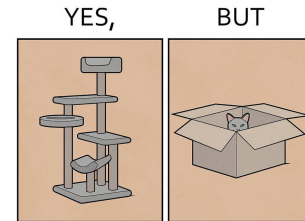
Problem 2: External Validity [Fuhr'21]

Goal

Repeating an experiment on similar data yields similar observations.

Possible problems

- ❑ Non-reproducible results



Possible Solutions

- ❑ TREC Open Runs [Voorhees'16]
- ❑ Reproducibility initiatives
 - OSIRRC: Archive artifacts [Arguello'15, Clancy'19]
 - CENTRE: Reimplementation [Ferro'19, Sakai'19]
- ❑ Platforms + documentation
 - CodaLab, EvalAI, PRIMAD, STELLA, TIRA
- ❑ Meta evaluations: BEIR [Thakur'21]
- ❑ 19 of 69 runs (Problems: 11)
- ❑ 2015: 8 systems archived
2019: 1 system fully reproducible [Lin'19]
- ❑ Limited adoption of jig + CIFF [Clancy'19]
- ❑ Additional effort
- ❑ Evaluations on subsets
- ❑ Often sparse judgments

TIRA for IR: Theory and Hands-On Tutorial

Problem 3: Blinded Experimentation with LLMs



Percy Liang

@percyliang



I worry about language models being trained on test sets. Recently, we emailed support@openai.com to opt out of having our (test) data be used to improve models. This isn't enough though: others running evals could still inadvertently contribute those test sets to training.

TIRA for IR: Theory and Hands-On Tutorial

Problem 3: Blinded Experimentation with LLMs

Touche 2020 Task #1 Topic Descriptions / Narratives

From: <ANONYMIZED>@openai.com

To: touche@webis.de

Hey!

Is there a list of all the topic descriptions / narratives for task #1 available (like in Table #1's example in the paper), and / or any other information that shines light on how the human evaluation scores were made?

Great work on the dataset!

Best,

--

<ANONYMIZED>

Member of the Technical Staff

OpenAI | www.openai.com

TIRA for IR: Theory and Hands-On Tutorial

Problem 3: Blinded Experimentation with LLMs

Touche 2020 Task #1 Topic Descriptions / Narratives

From: <ANONYMIZED>@openai.com

To: touche@webis.de

Dataset	GPT-4 (Random Exemplars)	GPT-4 (Curated Exemplars)
MedQA US 5-option	78.63	78.24
MedQA US 4-option	81.38	82.33
MedMCQA	72.36	71.36
PubMedQA	74.40	74.00

Table 5: Random few-shot exemplar selection vs. expert curation.

6.2 Memorization

GPT-4’s strong performance on benchmark datasets raises the possibility that the system is leveraging *memorization* or *leakage* effects, which can arise when benchmark data is included in a model’s training set. Given that LLMs are trained on internet-scale datasets, benchmark data may inadvertently appear

OpenAI | www.openai.com

TIRA for IR: Theory and Hands-On Tutorial

Problem 3: Blinded Experimentation with LLMs

Touche 2020 Task #1 Topic Descriptions
 From: <ANONYMIZED>@openai.com
 To: touche@webis.de


Dataset
MedQA US 5-option
MedQA US 4-option
MedMCQA
PubMedQA

Table 5: Random f

6.2 Memorization

GPT-4's strong performance on bench
memorization or leakage effects, which
 set. Given that LLMs are trained on i

OpenAI | www.openai.com



Horace He
@cHHillee

...

I suspect GPT-4's performance is influenced by data contamination, at least on Codeforces.

Of the easiest problems on Codeforces, it solved 10/10 pre-2021 problems and 0/10 recent problems.

[Tweet übersetzen](#)

g's Race	implementation, math	🚩	★	greedy, implementation	🚩	★	
nd Chocolate	implementation, math	🚩	★	Cat?	implementation, strings	🚩	★
triangle!	brute force, geometry, math	🚩	★	Actions	data structures, greedy, implementation, math	🚩	★
	greedy, implementation, math	🚩	★	Interview Problem	brute force, implementation, strings	🚩	★
umbers	brute force	🚩	★	vers	brute force, implementation, strings	🚩	★
ine Line	implementation	🚩	★	nd Suffix Array	strings	🚩	★
r or Stairs?	implementation	🚩	★	ther Promotion	greedy, math	🚩	★
Loves 3 I	math	🚩	★	iForces	greedy, sortings	🚩	★
s	implementation, math	🚩	★	i and Append	implementation, two pointers	🚩	★
	greedy, implementation, sortings	🚩	★	ig Directions	geometry, implementation	🚩	★

TIRA for IR: Theory and Hands-On Tutorial

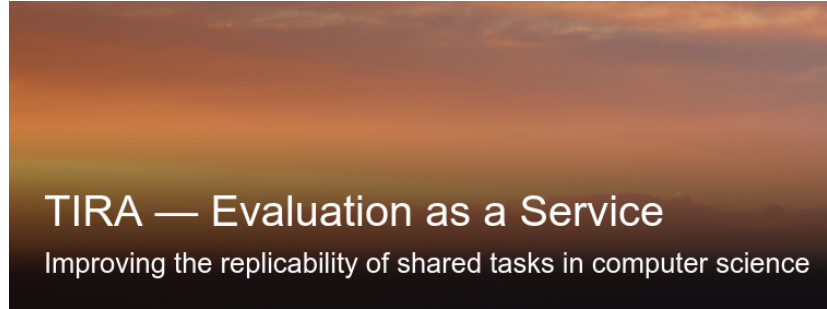
How Do We Try to Address those Problems in The Open Web Search Project?



TIRA for IR: Theory and Hands-On Tutorial

Evaluation and Prototyping with TIRA

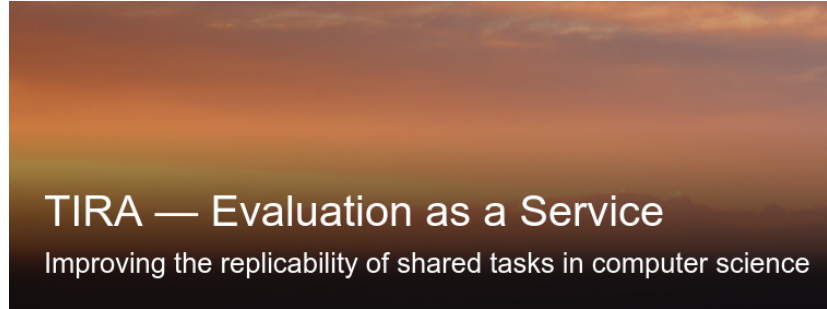
`http://tira.io`



TIRA for IR: Theory and Hands-On Tutorial

Evaluation and Prototyping with TIRA

`http://tira.io`



Evolution of TIRA

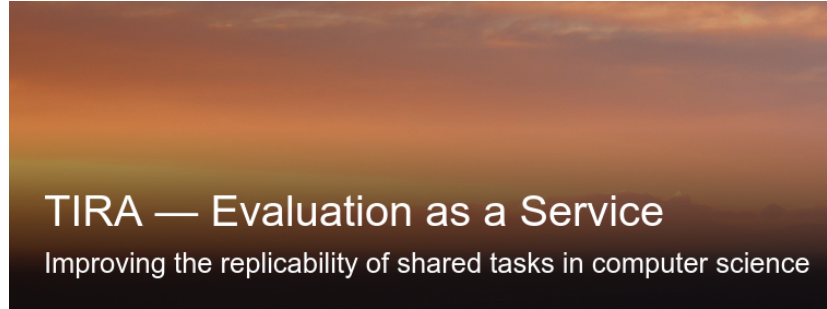
[Gollub'12, Potthast'19, Fröbe'23]

- ❑ 2005–2011: Pipelines, eval. run submissions, manual software submissions
- ❑ 2012–2022: Software submissions with virtual machines
- ❑ 2023–today: Immutable software submissions with Docker + Git CI/CD

TIRA for IR: Theory and Hands-On Tutorial

Evaluation and Prototyping with TIRA

`http://tira.io`



Evolution of TIRA

[Gollub'12, Potthast'19, Fröbe'23]

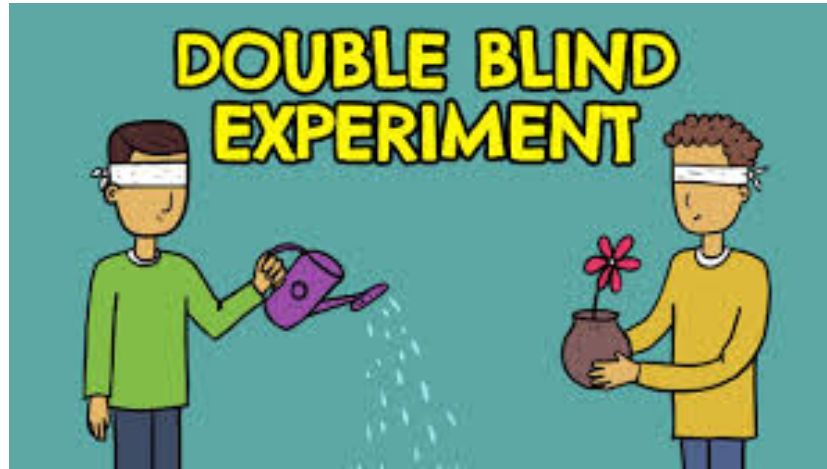
- ❑ 2005–2011: Pipelines, eval. run submissions, manual software submissions
- ❑ 2012–2022: Software submissions with virtual machines
- ❑ 2023–today: Immutable software submissions with Docker + Git CI/CD

Procedure:

1. Implement approach in Docker image
2. Upload image to dedicated image registry in TIRA
3. Your approach is executed in a Kubernetes cluster via a commit

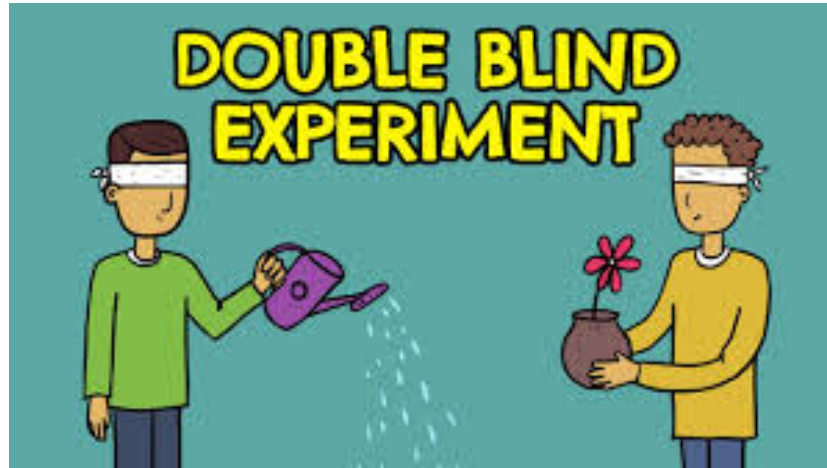
TIRA for IR: Theory and Hands-On Tutorial

Software Submissions in TIRA enable Blinded Experiments



TIRA for IR: Theory and Hands-On Tutorial

Software Submissions in TIRA enable Blinded Experiments



Blinded Experimentation

- ❑ Software executed in sandbox: No internet connection
- ❑ 2 types of datasets:

Type	Blinded	Unblinding	Feedback
Validation	Nothing	Direct	Everything
Test	Everything	Manual	✓vs ✗

Enough Preliminaries...



Enough Preliminaries...



Time to get our hands dirty :)

TIRA for IR: Theory and Hands-On Tutorial

We will use modern libraries and tools

- ❑ Docker for deployment
- ❑ `ir_datasets` for data wrangling
- ❑ PyTerrier for declarative retrieval pipelines

TIRA for IR: Theory and Hands-On Tutorial

We will use modern libraries and tools

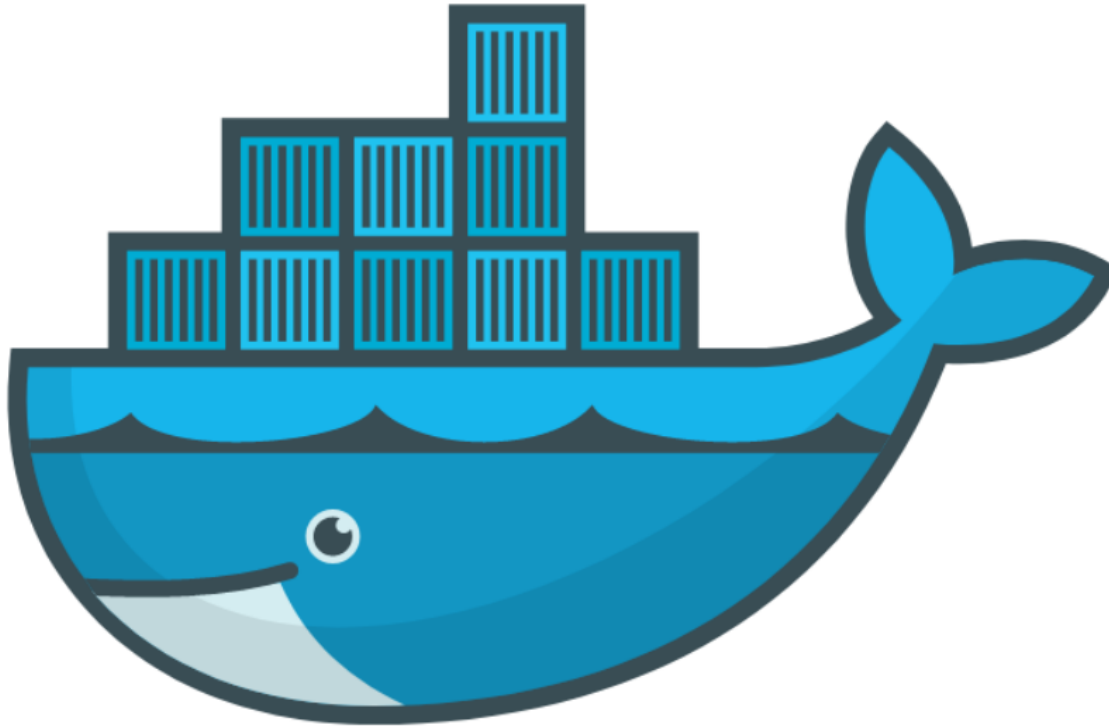
- ❑ Docker for deployment
- ❑ `ir_datasets` for data wrangling
- ❑ PyTerrier for declarative retrieval pipelines

You can adjust the setup to your preferences:

- ❑ Docker: Higher technical expertise
- ❑ Dev-Container: Medium technical expertise needed
- ❑ Codespaces: Low technical expertise needed

Docker Tutorial

Docker Basics



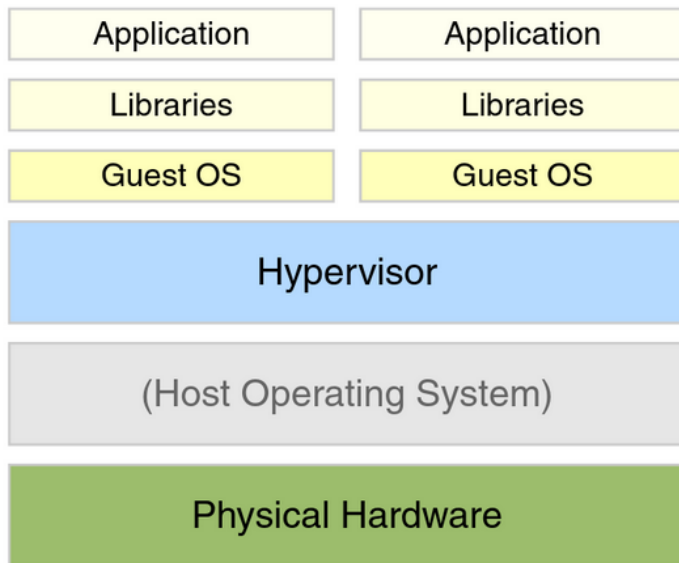
- ❑ Goal: If you can start/stop your jupyter notebook everything is fine
- ❑ <https://docs.docker.com/get-docker/>
- ❑ We will provide all required commands

Docker Tutorial

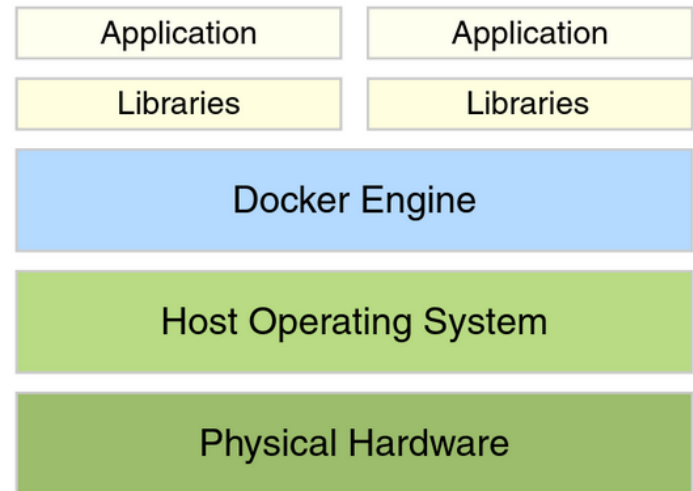
Use Cases for Docker

- ❑ Run guest systems as containers
- ❑ Shipping and running micro services as portable images
- ❑ Exploring and experimenting with new technologies
- ❑ Encapsulation mechanism to deploy applications in parallel without conflicts

Virtual Machines vs Docker



Virtual Machines



Docker

Docker Tutorial

Example Docker Commands

- ❑ Visit hub.docker.com
- ❑ We use the `bash` and `webis/ir-lab-wise-2023:0.0.4` images

Docker Tutorial

Example Docker Commands

- ❑ Visit hub.docker.com
- ❑ We use the `bash` and `webis/ir-lab-wise-2023:0.0.4` images

Bash Image

```
docker run --rm -ti bash
```

- ❑ `--rm`: Remove container after completion
- ❑ `-ti`: Attach stdin and stdout
- ❑ **ToDo**: Run above comand without `-ti`. What happens?
- ❑ **ToDo**: Write text to some file, restart the container. What happens?

Docker Tutorial

Example Docker Commands

- ❑ Visit hub.docker.com
- ❑ We use the `bash` and `webis/ir-lab-wise-2023:0.0.4` images

Bash Image

```
docker run --rm -ti bash
```

- ❑ `--rm`: Remove container after completion
- ❑ `-ti`: Attach stdin and stdout
- ❑ **ToDo**: Run above command without `-ti`. What happens?
- ❑ **ToDo**: Write text to some file, restart the container. What happens?

Bash Image With Volume Mounts

```
docker run --rm -ti -v $PWD:/bla bash
```

- ❑ `-v <HOST_PATH>:<CONTAINER_PATH>`: Mount the directory `<HOST_PATH>` on the system to the directory `<CONTAINER_PATH>` within the container
- ❑ **ToDo**: Write text to some file so that it is persistent.

Docker Tutorial

Jupyter Notebook and PyTerrier Pipelines with Docker

- ❑ We have prepared a Docker image with all reasonable libraries/frameworks preinstalled

```
docker run --rm -ti -p 8888:8888 \  
  -v $PWD:/workspace/ \  
  webis/ir-lab-wise-2023:0.0.4 \  
  jupyter notebook --allow-root --ip 0.0.0.0
```

- ❑ `-p <HOST_PORT>:<CONTAINER_PORT>`: **Map port <HOST_PORT> on the system to the port <CONTAINER_PORT> within the container**
- ❑ `jupyter notebook --allow-root --ip 0.0.0.0`: **The command executed in the container. This command starts a Jupyter notebook.**
- ❑ **ToDo: Play around with Python in the notebook for a few minutes**

Docker Tutorial

Now We repeat this with Dev-Containers in VS Code

If we have time, we can see the same steps in a Dev-Container.

TIRA for IR: Theory and Hands-On Tutorial

Before We Go Into the Weekend, We Make our First Submission



TIRA for IR: Theory and Hands-On Tutorial

Before We Go Into the Weekend, We Make our First Submission



Preparation:

- ❑ Please sign up at tira.io
- ❑ Please register to the IR Lab Augsburg/Jena/Köln/Leipzig SoSe 2024
- ❑ We create a private chat for problems/questions, etc.

TIRA for IR: Theory and Hands-On Tutorial

First Submission: Step-by-Step Guide



- ❑ Step 1: Develop your System(s) on the training data (we will use a BM25 baseline in the following)
- ❑ Step 2: Connect your TIRA account to your git repository
- ❑ Step 3: Upload your Code
- ❑ Step 4: Build your Docker image via Github Actions
- ❑ Step 5: Execute your Approach in TIRA

Thats all, have fun!

- ❑ Resources/Overview at: <https://tira-io.github.io/ir-lab-bose-24/>
- ❑ Potential todos for remaining time:
 - Submit your own TF-IDF / BM25 approach?
 - Submit TinyBERT together?

Thats all, have fun!

- ❑ Resources/Overview at: <https://tira-io.github.io/ir-lab-bose-24/>
- ❑ Potential todos for remaining time:
 - Submit your own TF-IDF / BM25 approach?
 - Submit TinyBERT together?



Thats all, have fun!

- ❑ Resources/Overview at: <https://tira-io.github.io/ir-lab-bose-24/>
- ❑ Potential todos for remaining time:
 - Submit your own TF-IDF / BM25 approach?
 - Submit TinyBERT together?



Thank you!