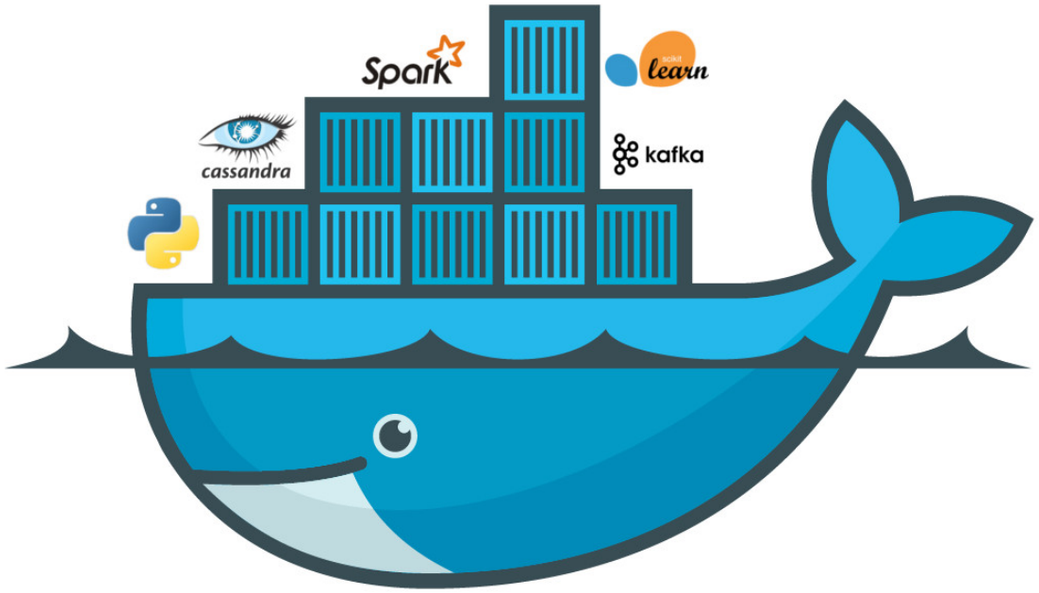
ISP 2017-2018

Reproducible Research via Docker

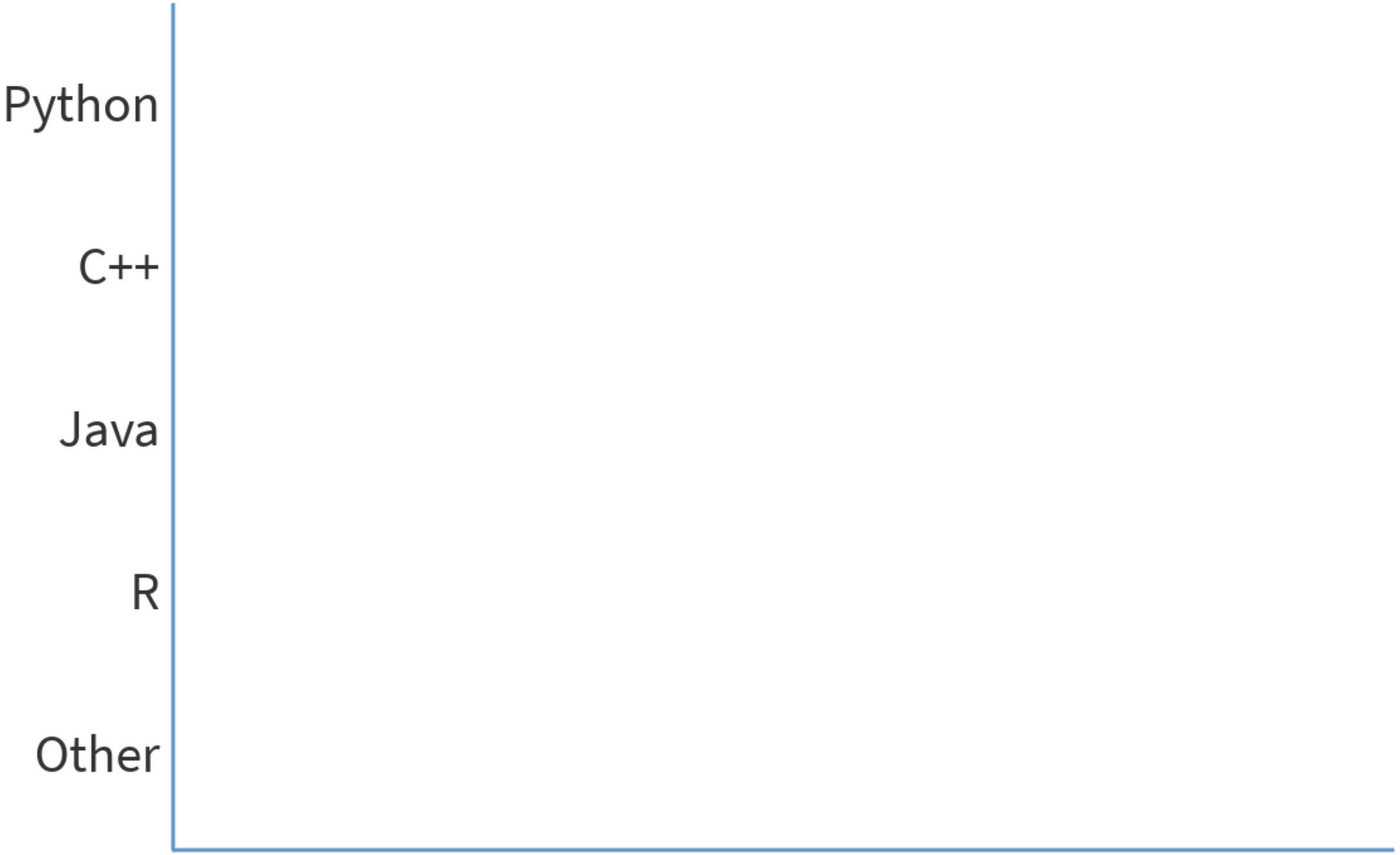


Artyom Nikitin (artem.nikitin@skolkovotech.ru),
Marina Munkhoeva (marina.munkhoeva@skolkovotech.ru)

Small Poll

pollev.com/artyomnikiti713

What is your main programming language for the research?



Have you ever tried to compile/run someone's code?

Yes

No

Was it painful?

Pretty much always

Sometimes

Never

Research Reproducibility

U.S. National Science Foundation:

"*Reproducibility* refers to the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator …"

Research Reproducibility

ACM - Artifact Review and Badging Policy (2016)

- Repeatability (same team, same setup)
- Replicability (different team, same setup)
- Reproducibility (different team, different setup)

First IEEE Workshop on The Future of Research Curation and Research Reproducibility (2016)

ACM SIGCOMM Reproducibility Workshop (2017)

The ACM Task Force on Data, Software, and Reproducibility in Publication (2017)

Research Reproducibility

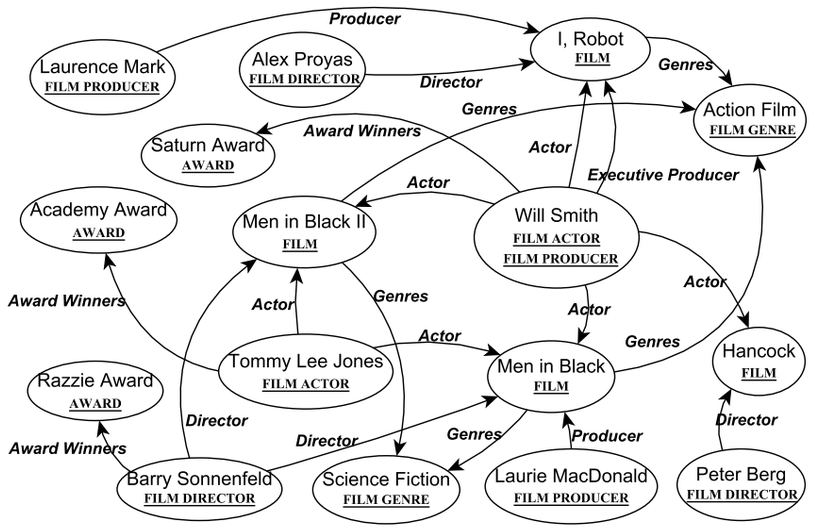
ACM SIGMOD (Management of Data) conference

Generating Preview Tables for Entity Graphs

[1]Ning Yan [2]Sona Hasani [2]Abolfazl Asudeh [2]Chengkai Li
[1]Huawei U.S. R&D Center [2]The University of Texas at Arlington
ning.yan.uta@gmail.com {sona.hasani,ab.asudeh}@mavs.uta.edu cli@uta.edu

ABSTRACT

Users are tapping into massive, heterogeneous entity graphs for many applications. It is challenging to select entity graphs for a particular need, given abundant datasets from many sources and the oftentimes scarce information for them. We propose methods to produce preview tables for compact presentation of important entity types and relationships in entity graphs. The preview tables assist users in attaining a quick and rough preview of the data. They can be shown in a limited display space for a user to browse and explore, before she decides to spend time and resources to fetch and investigate the complete dataset. We formulate several optimization problems that look for previews with the highest scores according to intuitive goodness measures, under various constraints on preview size and distance between preview tables. The opti-



Research Reproducibility

Minimum Requirements

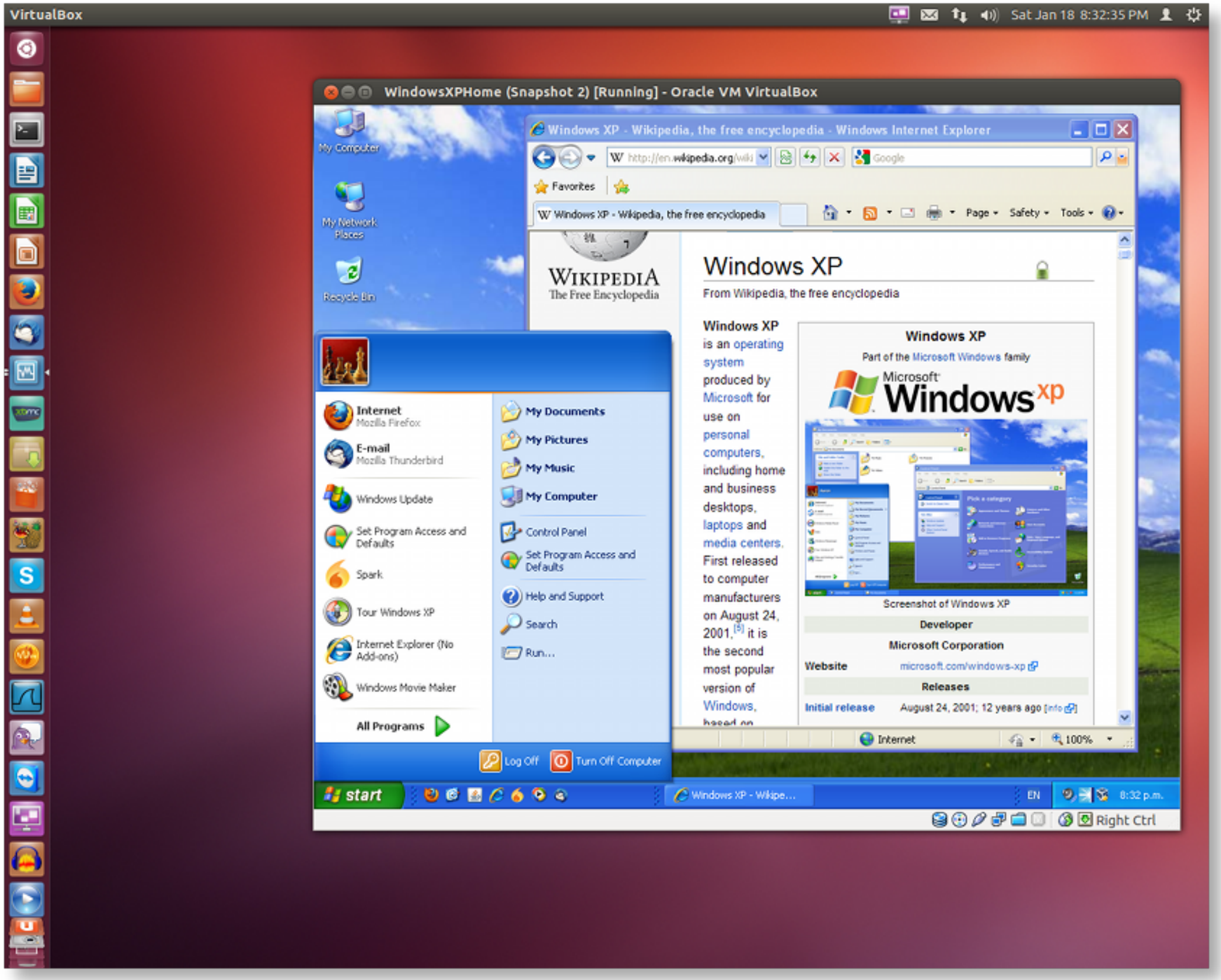
- Available source code
- Thorough documentation
- Comments
- Hardware used
- Software / libraries used
- Step-by-step installation / run instructions

Research Reproducibility

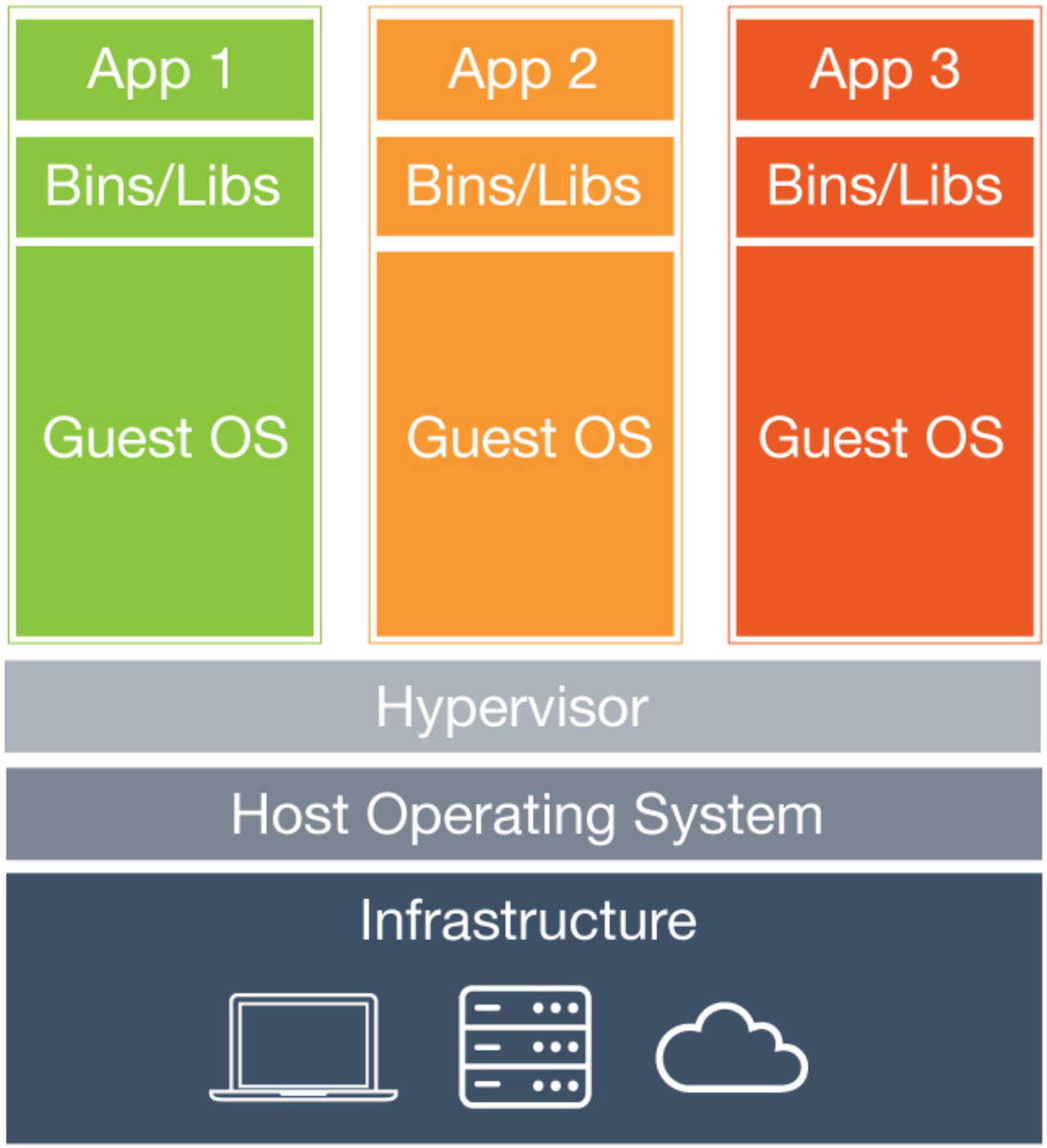
~~Minimum~~ Requirements

- Available source code
- Thorough documentation
  - Comments
  - Hardware used
  - ~~Software / libraries used~~
  - ~~Step-by-step installation / run instructions~~
  - **ONE-click build / run**

Virtualization

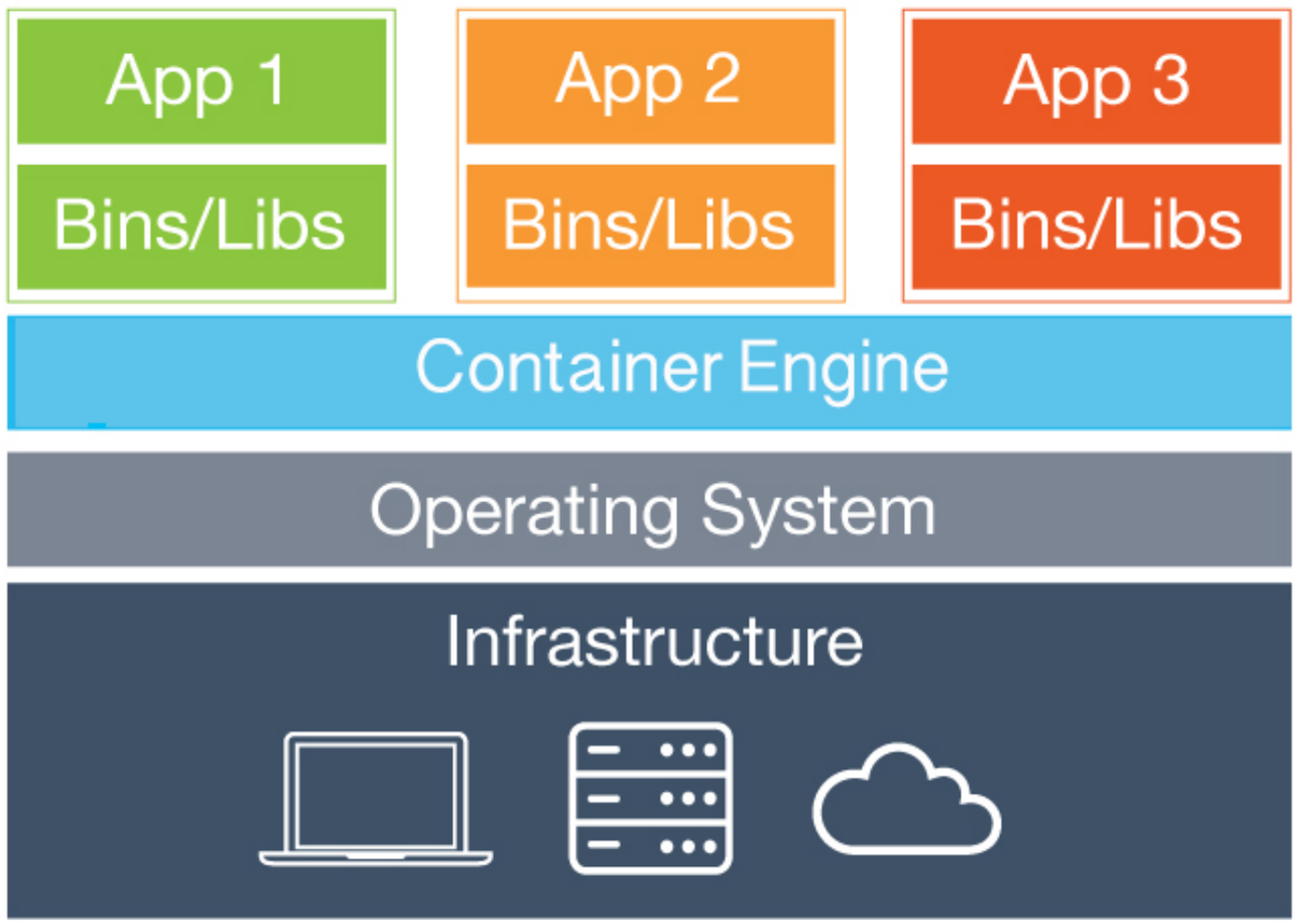


Virtualization: Hypervisor



Hypervisor-based Virtualization

Virtualization: Containers



Container virtualization

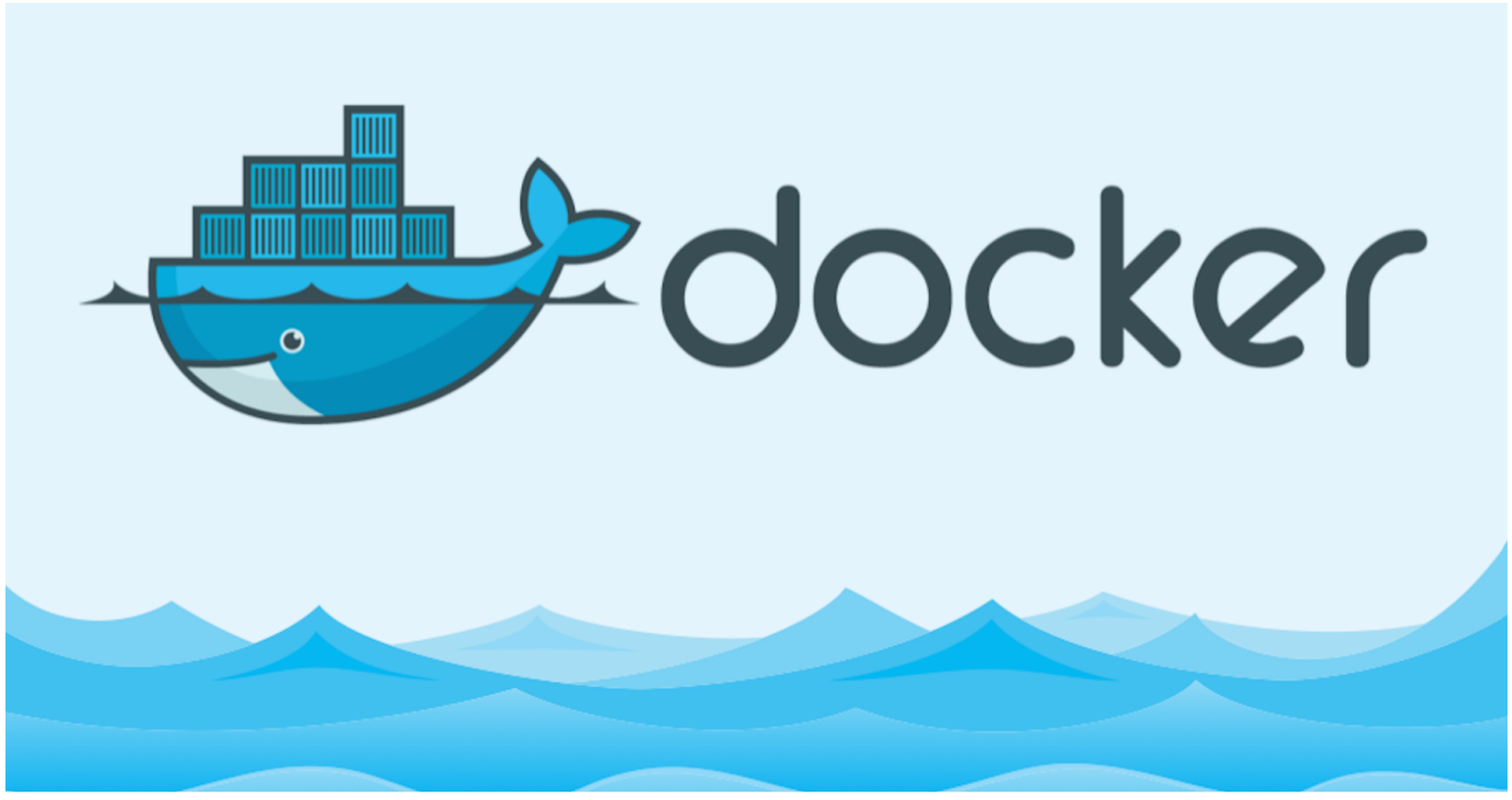
Virtualization: Containers

Hypervisor-based:

- Any OS on any OS
- Complete isolation of guest instances
- Worse performance

Container-based:

- Share kernel with host
- Worse security
- Almost native performance



- Convenient
- Efficient
- Open-source (split recently to CE and EE)
- Huge community

Docker

What you need to know…

1. *Docker Image* - basis of a container.

    - OS distribution-specific code.
    - Your libraries, source code, data, etc.

2. *Docker Container.*

3. *Docker Engine* - creates, ships and runs docker containers on physical / virtual host.

Image: Dockerfile

FROM ubuntu:14.04

WORKDIR /hello-world

RUN sudo apt-get update && apt-get install -y python

ADD hello-world.py /hello-world/

CMD python /hello-world/hello-world.py

```
docker build -t <image_name> <directory>

docker run <image_name>
```

Image: Structure

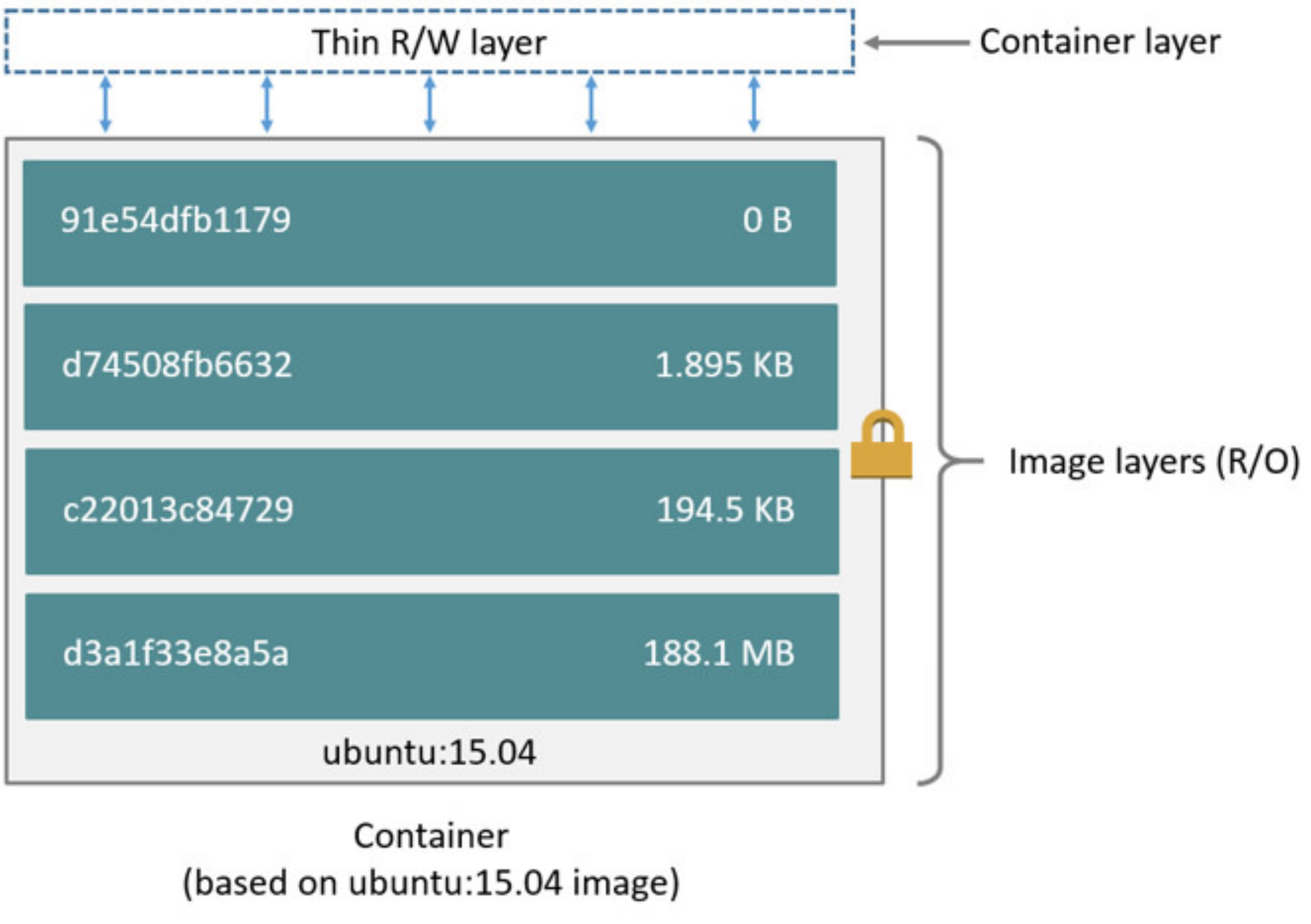


Image: Structure

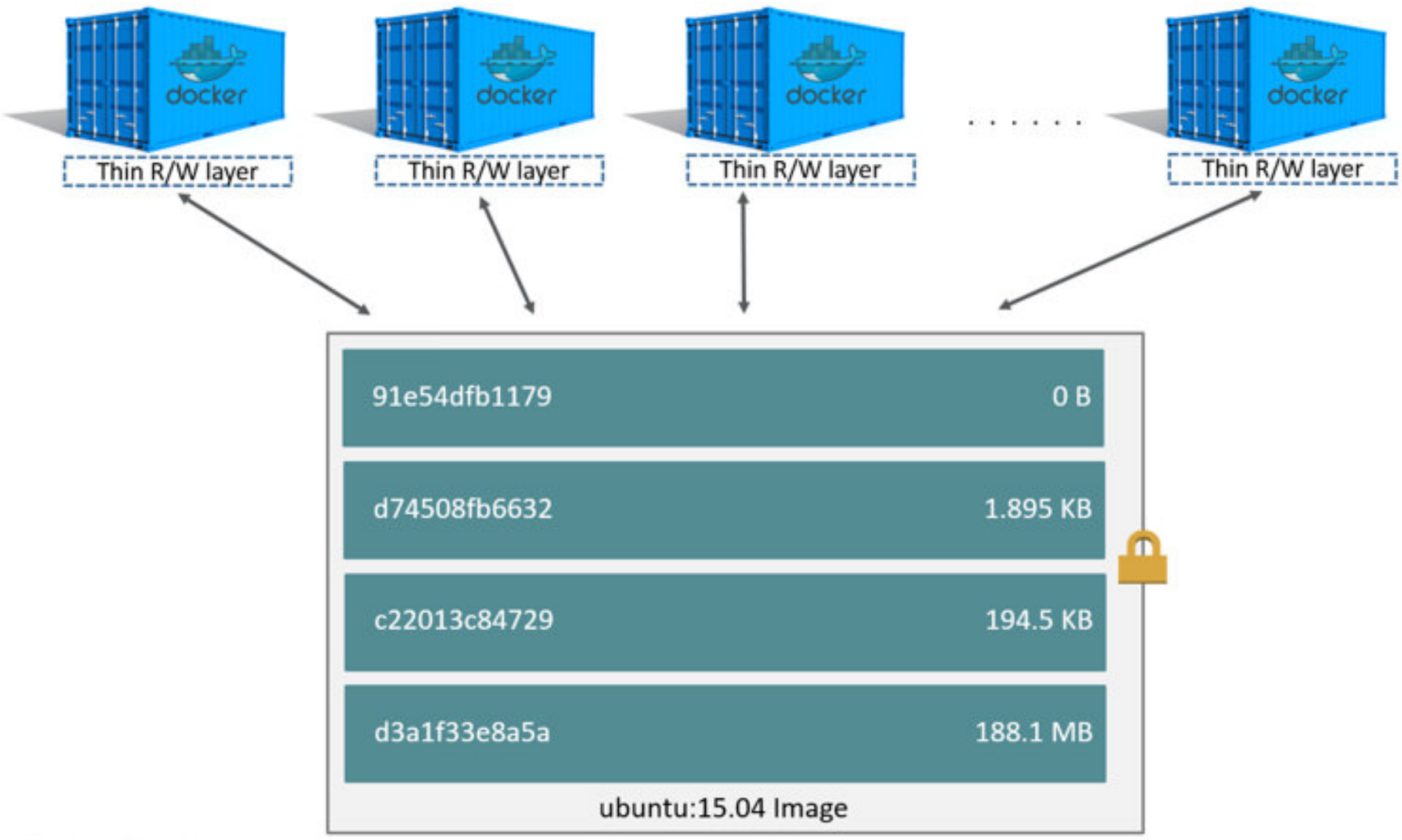


Image: Sharing

- Official Docker registry
  - hub.docker.com
    - 1 free private repo
- You can deploy your own registry
  - check docs.docker.com/registry/

Hands-on!

https://docs.docker.com/engine/installation/

Post-installation tips:

- Change default images storage directory - they **WILL** eat up your disk space

- Linux - enable non-root use:
  1. sudo groupadd docker
  2. sudo useradd -G docker $USER
  3. Logout - login

If you are pythonic…

Binder (beta) mybinder.org

- "Turn a GitHub repo into a collection of interactive notebooks"
- Upload your code to GitHub
- Put *requirements.txt*, *environment.yml* or a *Dockerfile* in the root of repo
- It will build and deploy the container, just connect to the jupyter notebook!
- Free so far, 4 GB RAM limit
- You can deploy your own Binder server!

Homework Assignment

1. Select any project you worked on, e.g.:
    - Research paper
    - Application period project
2. Pack it using Docker. Upon run:
    - Prepare datasets
    - Conduct experiments
    - Generate plots
    - Generate paper / report
3. Push image to docker repository and submit a link in canvas
4. Verify randomly assigned submission from another student