Docker

for reproducible and portable (Data) Science

Luca Verginer

01/03/2017

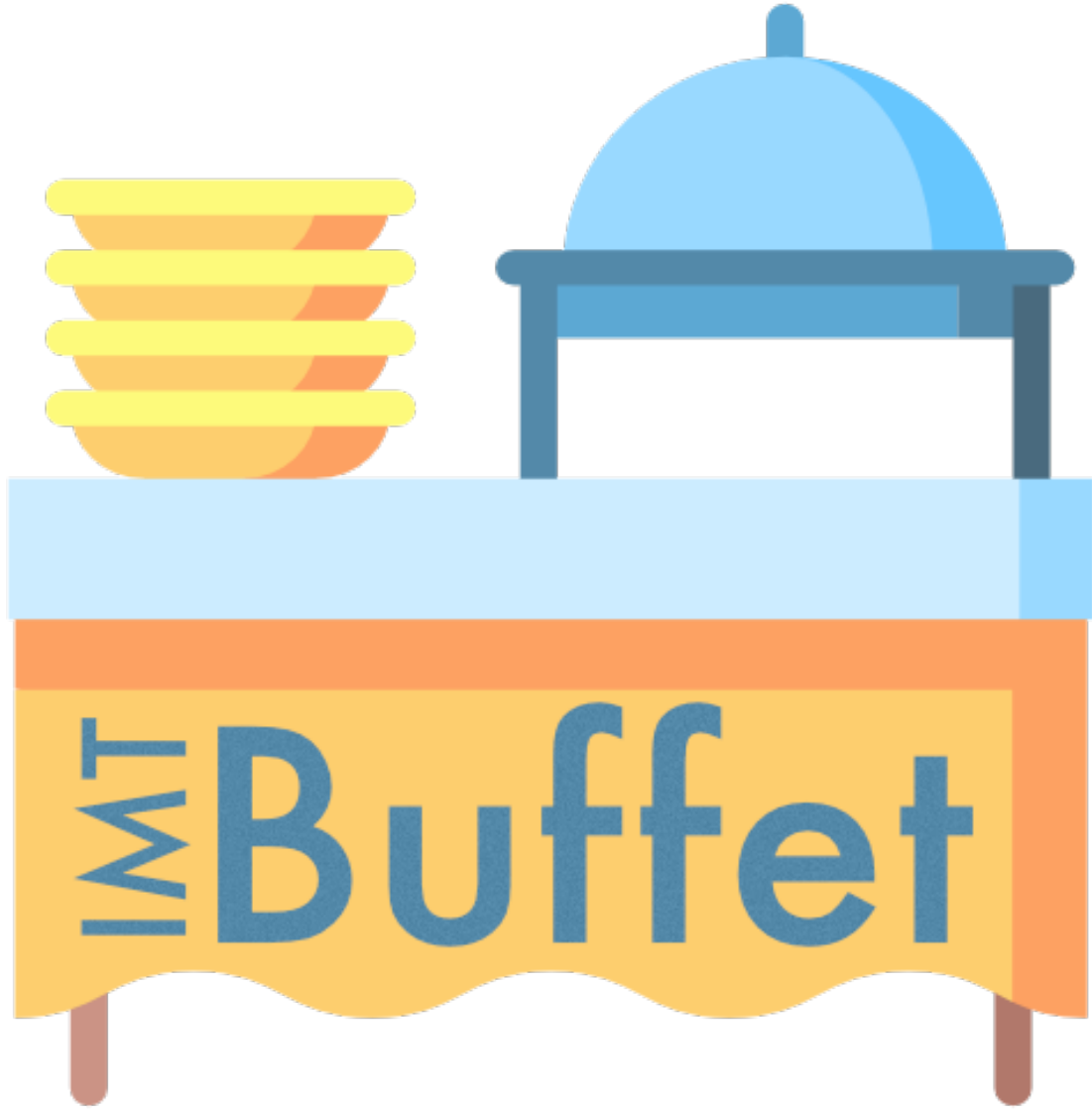

Figure 1:

- IMT Reading Group

Figure 2:

Docker containers wrap a piece of software in a complete filesystem that contains everything needed to run: code, runtime, system tools, system libraries – anything that can be installed on a server. This guarantees that the software will always run the same, regardless of its environment.

Overview

- What is Docker?
- Why would you want to use it?
- “Nice! But how do I use it?!”
- Basic Architecture
- Demo

What is it?

- Think of `Docker` as a *Ultra Light* Virtual Machine (VM) .
- It is a tool to create linux images with **“Known Good State”**.
- Offers *free* infrastrucure to store, distribute and retrieve images from anywhere on the web.

The 2 central concepts

Image and **Container**

- Image:
  - Large files containing a frozen Linux instance
  - i.e. an Snapshot of a machine you care about.
- Container:
  - A running machine created from an Image

- Image <-> Container
- OOP: Class <-> Class instance
- OS: binary <-> Process (i.e. something with a `PID`)

What is it good for?

Reproducibility

The 4 stages of code isoltion:

1. Separate folder for each project
2. Using Github or other VCS
3. python virtual environments (i.e. manipulate `$PATH`)
4. Virtual Machine

Problems with the above solutions

- Could impact an other project
- No guarantee it will work on different machine
- . . . or tomorrow on the same
- Slow and huge
- Difficult to move around

Docker to the rescue!

- A machine is completely defined by its `Dockerfile`

```
# set base image
FROM ubuntu:14.04

# install what you need
RUN apt-get update && apt-get install -y package-bar
RUN pip install numpy scipy

# default command to execute when starting
CMD bash
```

- Allows to deterministically recreate a machine from scratch
- Anyone can reproduce your “research” if you publish it’s Dockerimage and Raw Data

Portability

- If a image works as expected on your machine it is **guaranteed** to work on an other
- Can be easily shared with collaborators (-> all work with the same development environment)
- If your computer breaks, you start working immmediately using any other Machine
  - Assuming you backup your data.

Scalability

- Start exploring on your local machine
- You do not have enough RAM/CPU ...
- Deploy to an Amazon Server or your department's Servers in a few minutes.

Specific Applications

- Example 1: Run a complex server set.
  - MySQL
  - Jupyter

- Example 2: Graph analysis.
  - Neo4j
  - RStudio
  - `graph_tool` notoriously difficult to install parallel version.

Architecture

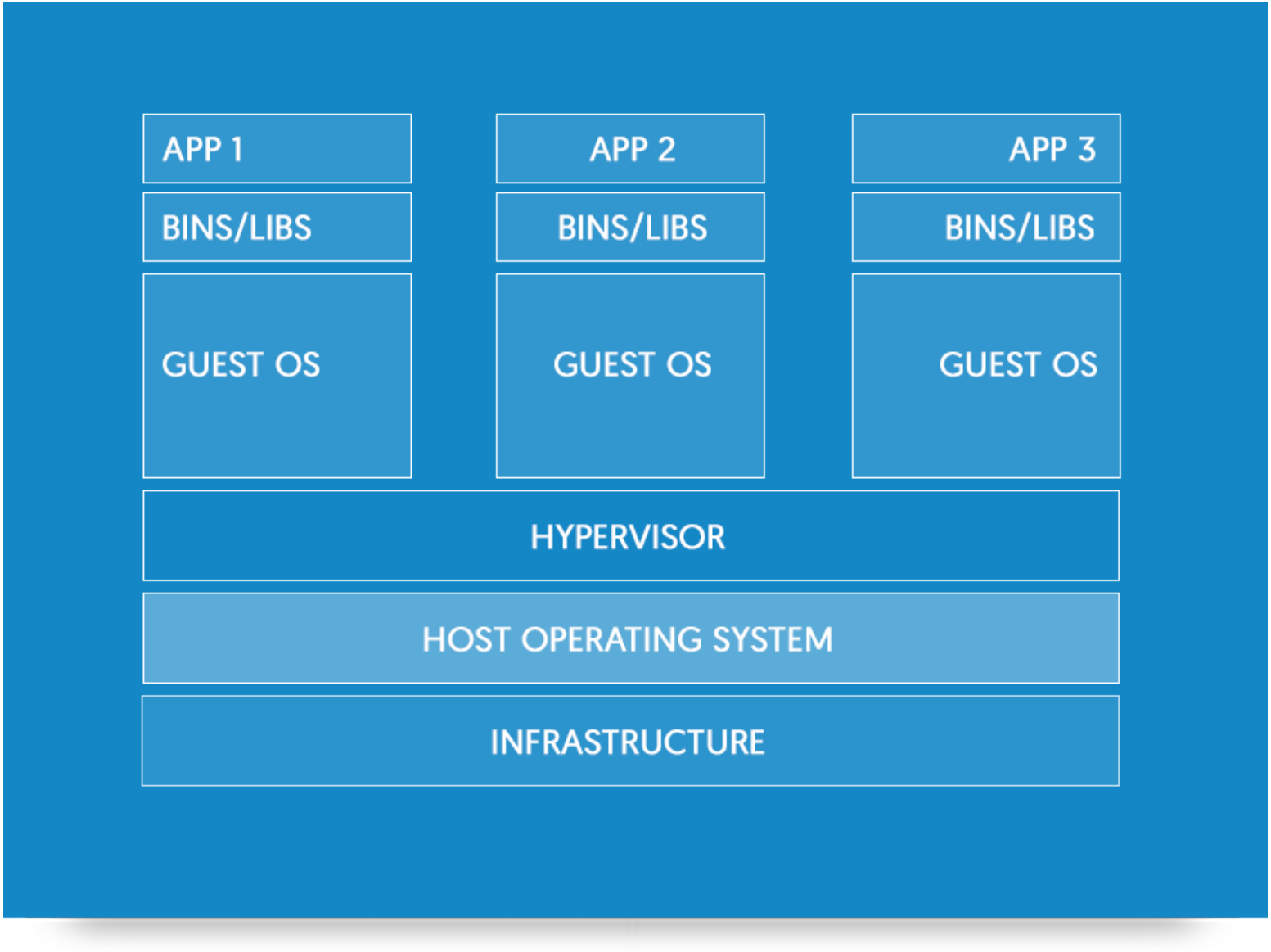

Figure 3: VM Architecture

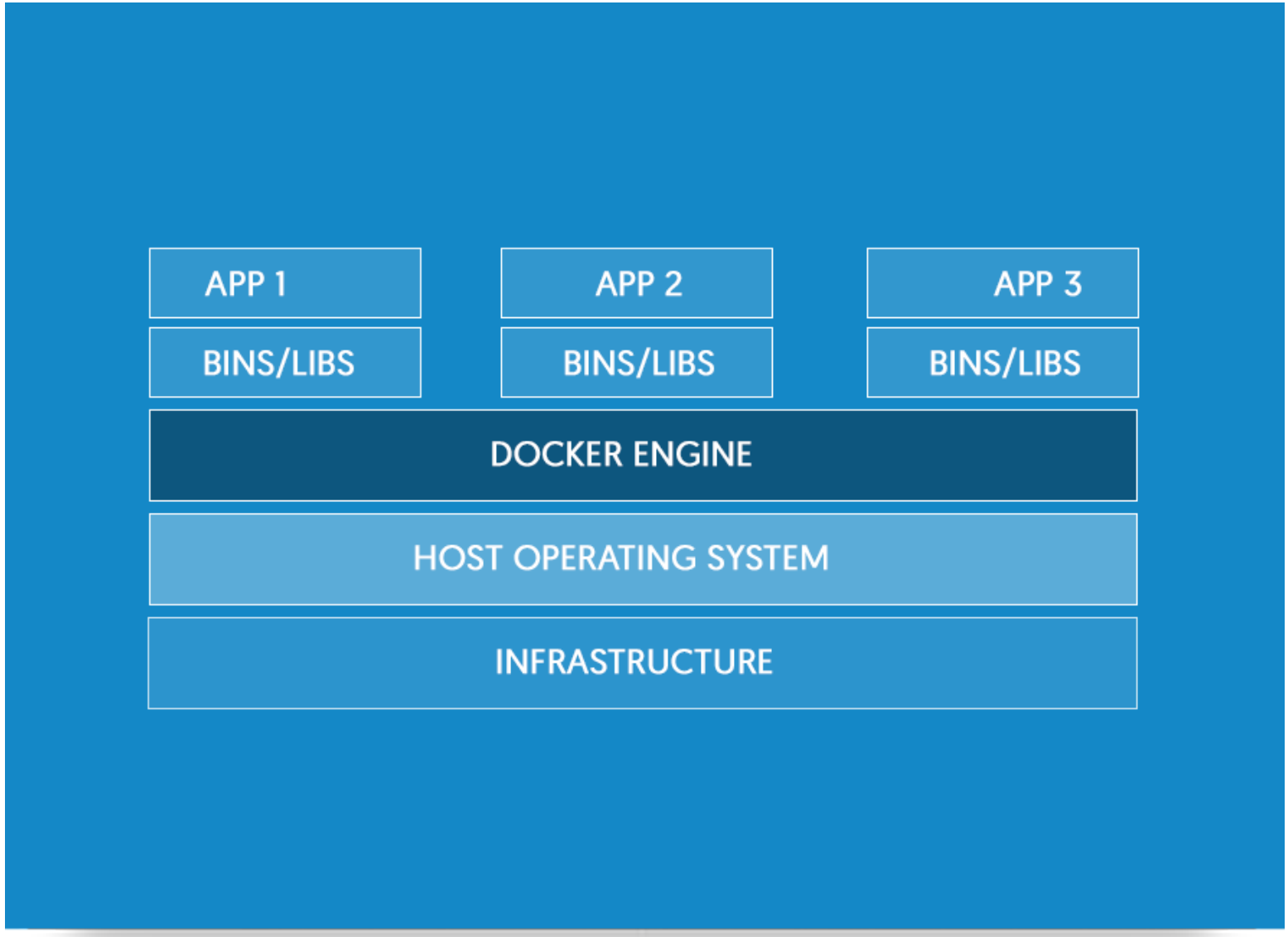

Figure 4: Docker Architecture

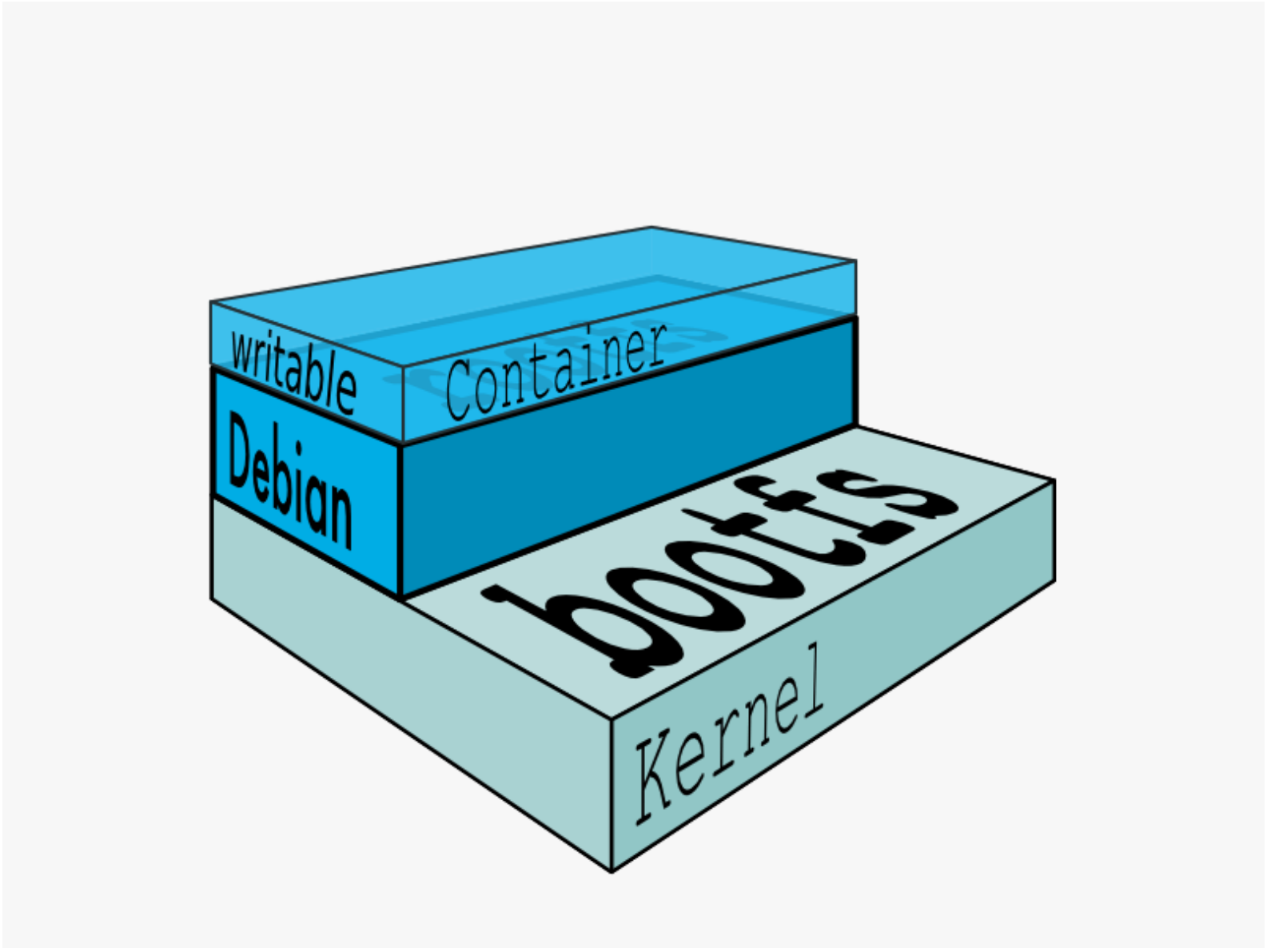

Figure 5: Layers

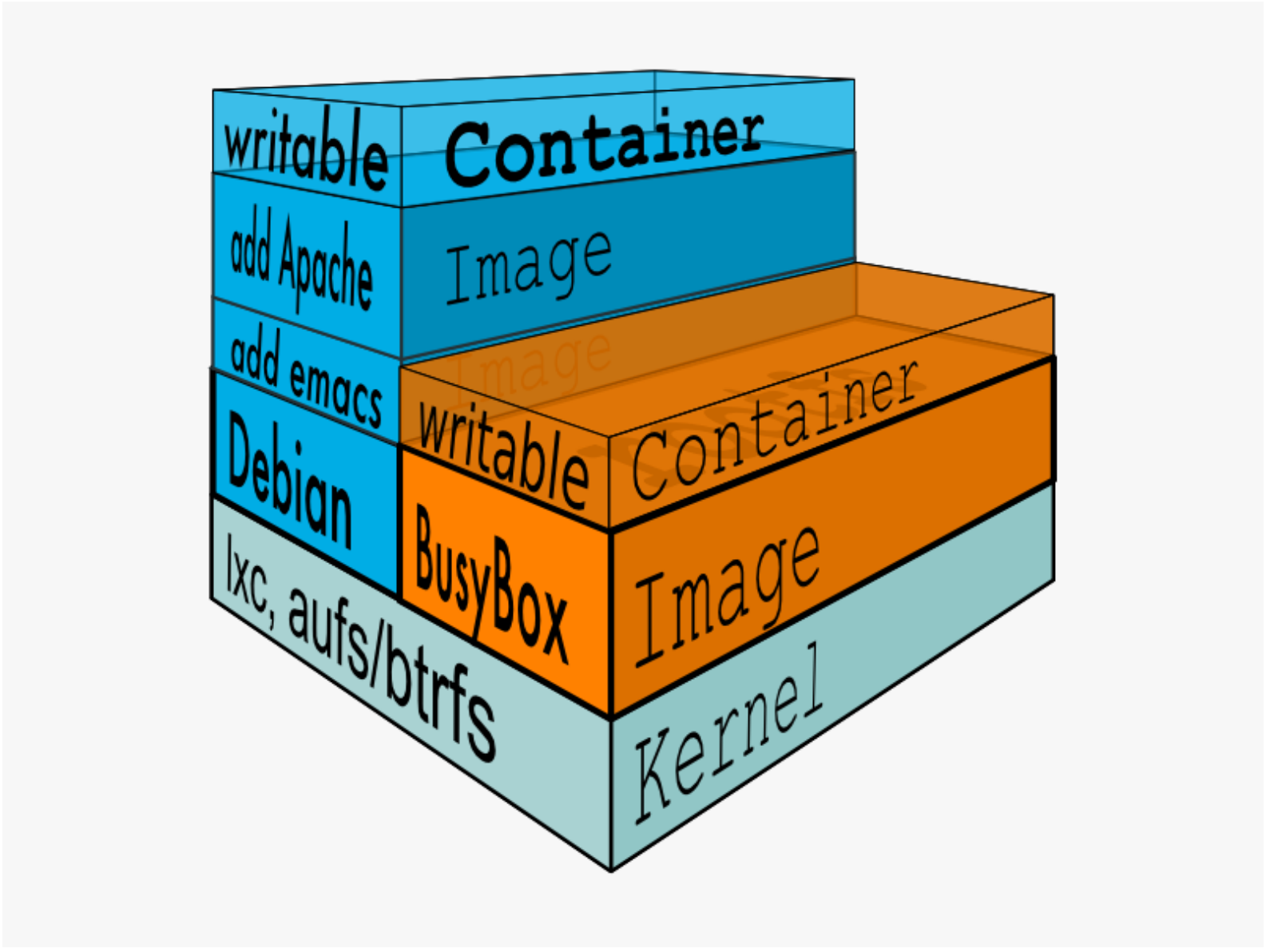

Figure 6: Layers

Demo

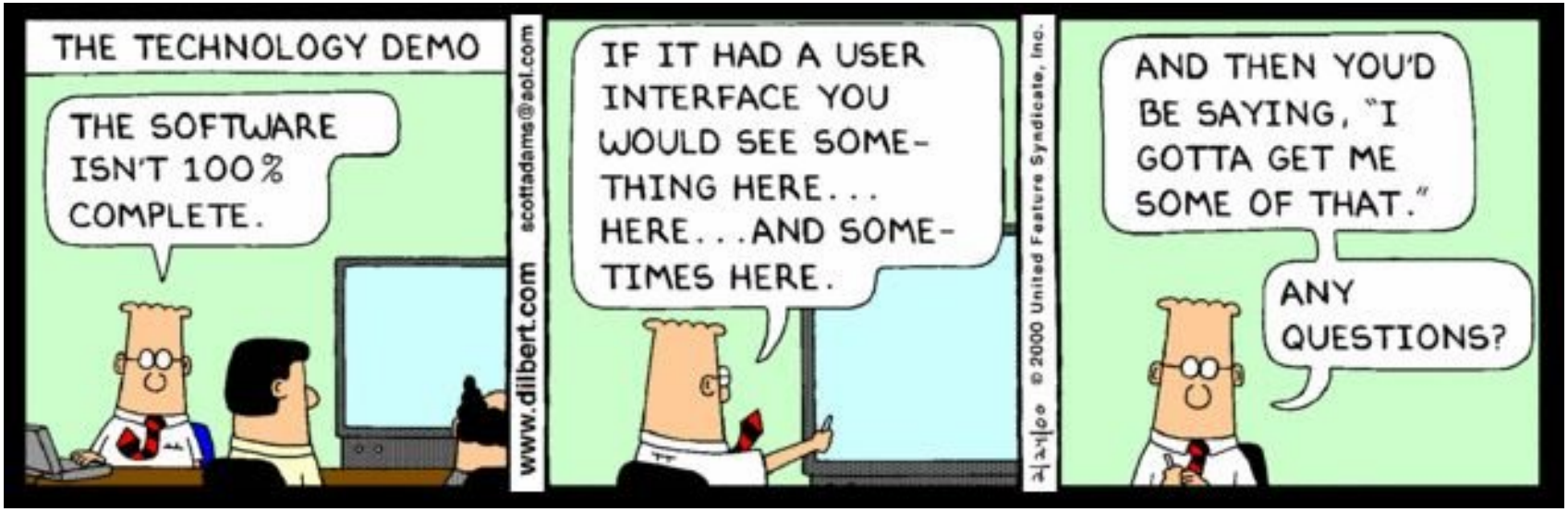
THE TECHNOLOGY DEMO
THE SOFTWARE ISN'T 100% COMPLETE.
www.dilbert.com scottadams@aol.com
IF IT HAD A USER INTERFACE YOU WOULD SEE SOMETHING HERE... HERE...AND SOMETIMES HERE.
© 2000 United Feature Syndicate, Inc.
AND THEN YOU'D BE SAYING, "I GOTTA GET ME SOME OF THAT."
ANY QUESTIONS?

Figure 7:

How to install

Instructions for Mac, Windows and Linux are available
https://docs.docker.com/engine/installation/

- very easy for Linux
- easy for Mac and Windows 10
- a little work for Windows 7, 8

Test if it works

```
docker run hello-world
docker ps
docker ps -a
docker rm quirky_babbage
```

Dockerfiles light

```
docker run docker/whalesay
```

boring!

- Customize whalesay image

```
docker build -t wisewhale .
```

Jupyter and RStudio

```
docker run -d \
           --name myjupyter \
           -p 8888:8888 \
           --volume $PWD:/home/jovyan/work \
           jupyter/scipy-notebook

docker run -d \
       --name myrstudio \
       -p 8787:8787 \
       -v $PWD:/home/rstudio \
       rocker/rstudio
```

On Amazon Cloud

```
ssh -i <keyfile> ubuntu@<aws-ip-address>
# now you are on an ec2 instance
docker run -d \
       --name myjupy \
       -e GEN_CERT=yes \
       -v $PWD:/home/jovyan/work \
       verginer/scipy_graph_tool

docker logs myjupy # to get the security token
```

Or just try a new technology

For nearly every micro-service/library/rpository there is a `Dockerfile` or `docker-compose.yml` to get it running in minutes.

That's it!

Some Places to get help/containers and ideas

- Awesome Docker, a curated list of useful “Docker” stuff
-