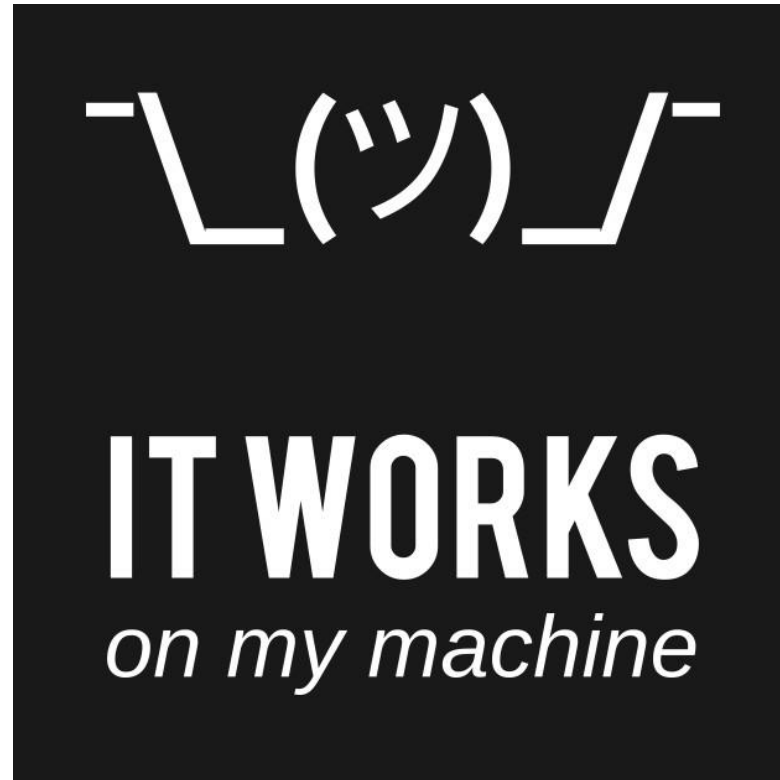


Docker Containers for Reproducibility



Kristin Chen

Agenda

- Why do we need containers?
- How can containers solve those problems?
- What is a Docker container?
 - How does Docker containerize?
- Demo: running r analysis in a container
- Reference

Why do we need containers?

- Scenario 1: rerun *my_analysis_12042018_KC.rmd* but get an error: *"Error in xx: could not find function "xx"*



- You then figure out the function *xx* is deprecated

Why do we need containers?

- Scenario 1: rerun *my_analysis_12042018_KC.rmd* but get an error: “*Error in xx: could not find function “xx”*”



- You then figure out the function *xx* is deprecated

- Scenario 2:



- You send *analysis_12042019_KC.rmd* to your teammate, along with the *private library* stored in the project directory created by *packrat* to manage package dependencies



- A few minutes later, your teammate pins you on Team and says he encountered an error: “*package ‘X’ is not available (for R version 3.3.0)*”



- You two then figure out it's because the R version in your PC is 3.6.0, which is compatible for package X; however, your teammate has R version 3.3.0 installed in his PC

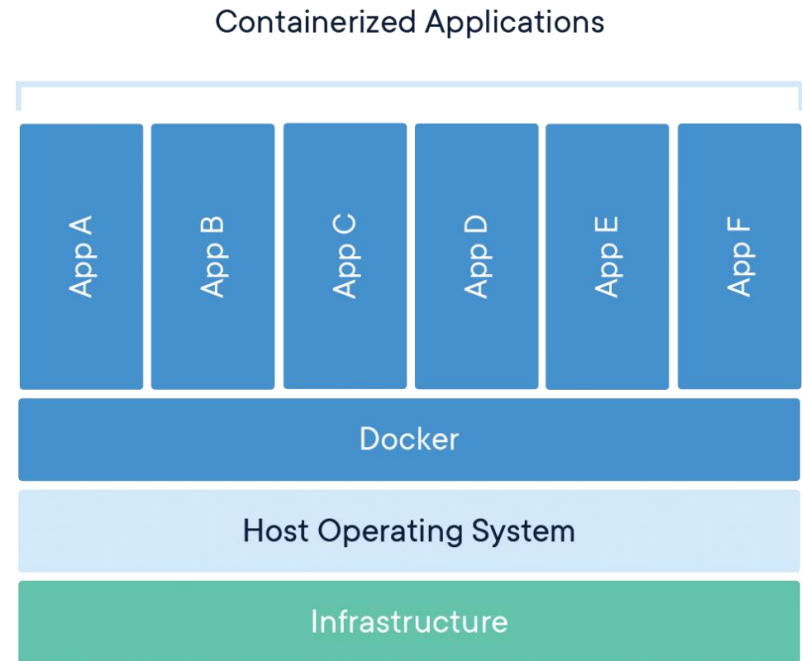
Why do we need containers?

- challenge:
 - the computation environment where the analyses build is **different** from where the analyses execute
 - replicating prior analysis or allowing other users to reproduce your analysis in their machines become difficult
 - computation environment, in the context of analysis in R programming, is included functions in the packages, packages, and R version, etc.



How can containers solve those problems?

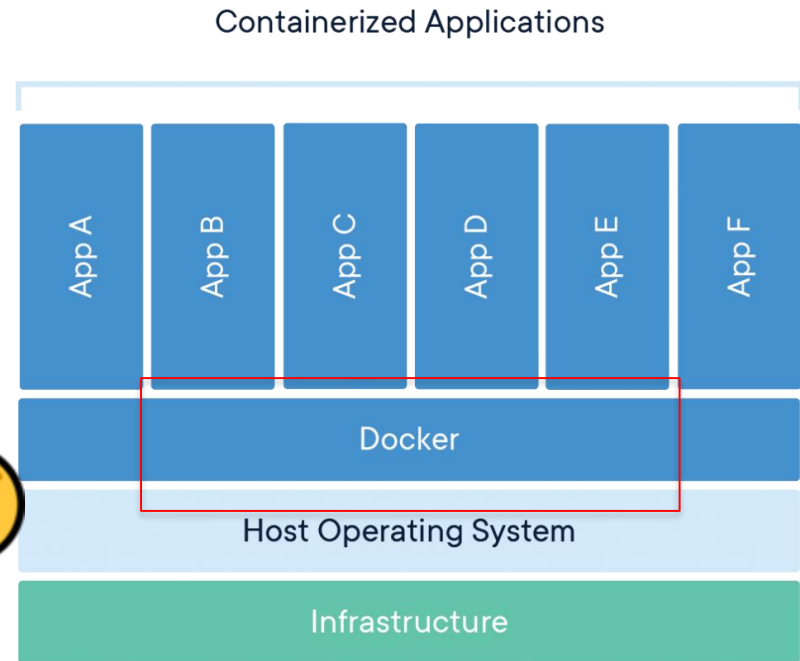
- *containerize* the computation environment as code
- containerize analyses with their computation environment in separated containers
- Run each analysis in its container to replicate same results
 - isolate from computation environment in main PC



How can containers solve those problems?



- *containerize* the computation environment as code
- containerize analyses with their computation environment in separated containers
- Run each analysis in its container to replicate same results
 - isolate from computation environment in main PC



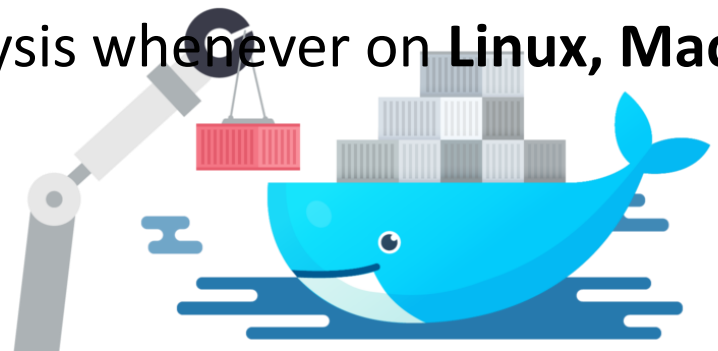
What is a Docker container?

- Docker: shipping companies, which carry containers from one part of the world to another
 - no containers:
 - everything needs to be transported are individually loaded into the ship but that makes loading and unloading of goods difficult
 - with containers:
 - everything can be loaded or unloaded quickly using cranes and that helps shipping companies to transport goods more easily



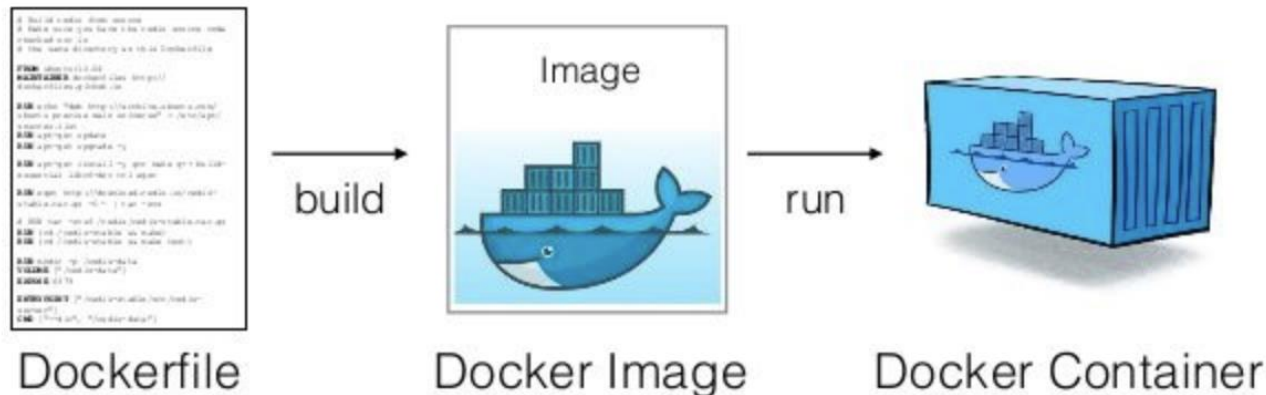
What is a Docker container?

- Docker ships containers which containerize **analyses**
 - analysis: a collection of computation pieces included packages, system dependencies, codes, reports, etc.
 - Docker creates a “**containerized**” version of the analysis, and everything needed to run the analysis is included
 - allow reproducing the analysis whenever on **Linux, Mac or Windows**



What is a Docker container? - how does Docker containerize?

- *build* an image from Dockerfile, and *spin up* a running container from that image
 - Dockerfile: a list of commands in a special text file to create an image
 - Docker image: a template of your analysis computation environment
 - Docker container: a running instance of a Docker image



Demo: running r analysis in container

- repo: <https://github.com/jiatingchen/docker-for-reproducibility>
- step:
 - `docker build Dockerfile`
 - `docker run image-nme`
 - `render my_analysis.rmd` to `my_analysis.html` in container
 - copy `my_analysis.html` to local machine
- Dockerfile:
 - rocker: <https://www.rocker-project.org/>
 - Dockerhub: Docker Hub is the world's largest public registry for versioning contain images
<https://hub.docker.com/>

Reference

<https://reproducible-analysis-workshop.readthedocs.io/en/latest/8.Intro-Docker.html>

<https://arxiv.org/pdf/1410.0846.pdf>

<https://hackernoon.com/5-free-online-courses-to-learn-docker-for-beginners-492cfc488ecb>

<https://cloudblogs.microsoft.com/opensource/2019/07/15/how-to-get-started-containers-docker-kubernetes/>

<https://nickjanetakis.com/blog/understanding-how-the-docker-daemon-and-docker-cli-work-together>

<https://ropenscilabs.github.io/r-docker-tutorial/>

<https://machinelearningmastery.com/machine-learning-in-r-step-by-step/>

<https://colinfay.me/docker-r-reproducibility/>

Thank you!

