CSP

New Orleans, Louisiana
February 14–16 2019

Reproducible Research

Best Practices and Tools for Support

Dr. Je’Anna Abbott, Professor
Jamey Johnston, Sr. Data Scientist/Engineer

Code is available at this GitHub Repo - https://github.com/STATCowboy/Reproducible-Research

Please silence cell phones

Dr. Je'Anna Abbott

Spec's Charitable Foundation Professor

in /jeannaabbott

@STATWonderWoman

University of Houston

Education

Texas A&M - MS in Analytics

U. of Houston - PhD

U. of Houston – JD

U. of Chicago – MBA

Jamey Johnston

Sr. Data Scientist

in /jameyj

@STATCowboy

Project Coach Texas A&M Analytics

Education

Texas A&M - MS in Analytics

LSU - BS in Spatial Analysis

Semi-Pro Photographer

http://jamey.photos

Blog

http://STATCowboy.com

Code

https://github.com/STATCowboy/Reproducible-Research

Agenda

- Why is Reproducible Research Necessary?
- Methodology Proposed to Assist in Reproducible Research
- What is git?
- What is GitHub?
- What is Docker?
- What is Docker Hub?
- Docker Image
- Dockerfile
- Docker Container
- Demos

Source: https://www.github.com

Source: https://www.docker.com/

Why is Reproducible Research Necessary?

- Recent studies - inability to reproduce or replicate previously published research
- Leading some to question the credibility of scientific research
- Only 11-25% of selected biomedical findings were independently replicable Economist (2013)
- Researchers in several disciplines lack training in proper procedures
- Others engage in questionable research practices

Methodology Proposed to Assist in Reproducible Research

git, GitHub and Docker

- GitHub to track changes in code
- Docker image hosted on Docker Hub
- Allows others to download and run the image as a container
- Contains the exact code, data, and software versions/packages used for the research

git and GitHub

GitHub

What is git?

- Git is a Version Control System (VCS)
- Used to track changes, versioning, in computer files/code in repositories
- Assist in the coordination of joint program work by multiple collaborators
- Branching and merging
- It was developed by Linus Torvalds (creator of Linux kernel)
- https://git-scm.com/

git and GitHub

What is GitHub?

- GitHub is web-based service for hosting VCSs based on git
- Both private repositories and free accounts (public) are available on GitHub
- http://www.github.com

Docker

What is Docker?

- Docker is platform develop, deploy, and run applications with containers
- Generally Linux based but can be Windows based
- Docker containers have the advantage of being:
  - Flexible
  - Lightweight
  - Interchangeable
  - Portable
  - Scalable
  - Stackable
- http://www.docker.com

Docker

What is Docker Hub?

- Docker Hub is a cloud-based service to store and download and test Docker images
- Build Docker images of your research code and data
- Publish as a Docker image on Docker Hub
- Allow others to download, test, and validate your research

Docker

Docker Image

- Docker container images contain everything needed to run a piece of software
- Contain code, runtime, system tools, system libraries, and settings
- They are lightweight, stand-alone, and executable
- Use an image to start containers

Docker

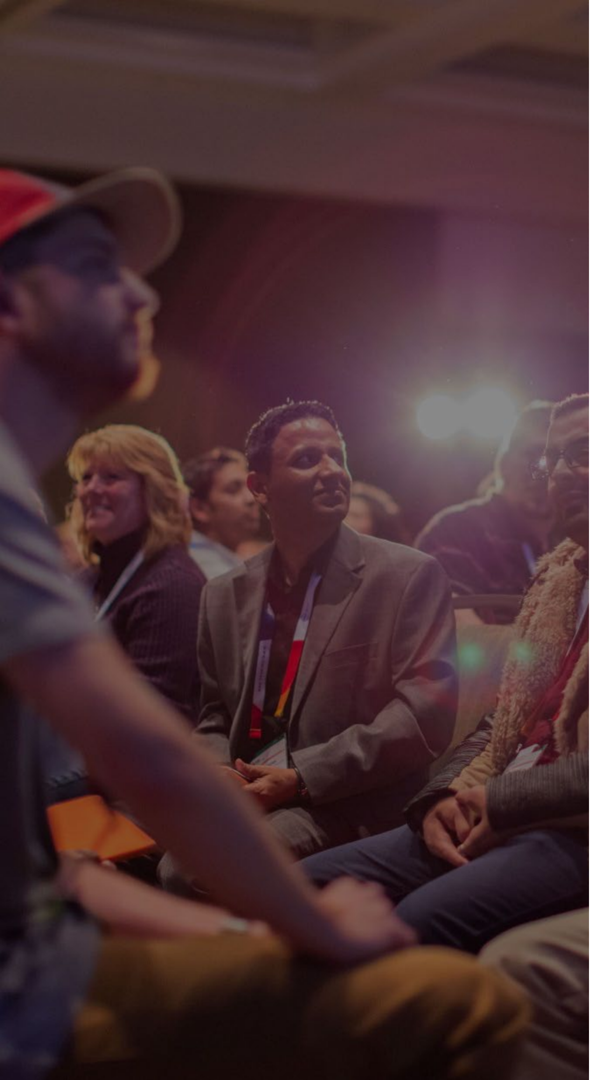
Dockerfile

- Dockerfile is a filed used to build Docker images
- It contains the instructions for Docker to build the images
- Which can later be instantiated as a Docker container
- Essentially it is the source code for a Docker image

Docker

Docker Container

- Is a running instance of a Docker image
- They run the actual application



Thank You

Jamey Johnston

@STATCowboy ✉ jameyj@tamu.edu

Dr. Je’Anna Abbott

@STATWonderWoman ✉ jabbott@uh.edu