- Problem to Solve:
  - Maintaining multiple data science software stacks for my hobby work: machine learning, deep learning, computer vision.
  - Work on both my personal laptop (MacBook) and Cloud (AWS).

- Work inspired by these talks:
  - Docker for Data Scientists
  - Capital One Analytic Garage on Docker

  and Kaggle's Docker offerings
  - Dockerhub-Kaggle

Data Science Software Stack Docker Prototype

Docker images providing the following data science software stacks for peronsal use:

- Anaconda Python with Jupyter Notebook
- Rstudio Server
- Apache Spark (Stand-alone)
- Tensorflow (cpu and gpu versions)
- h2o
- xgboost
- lightgbm

See wiki for additional information.

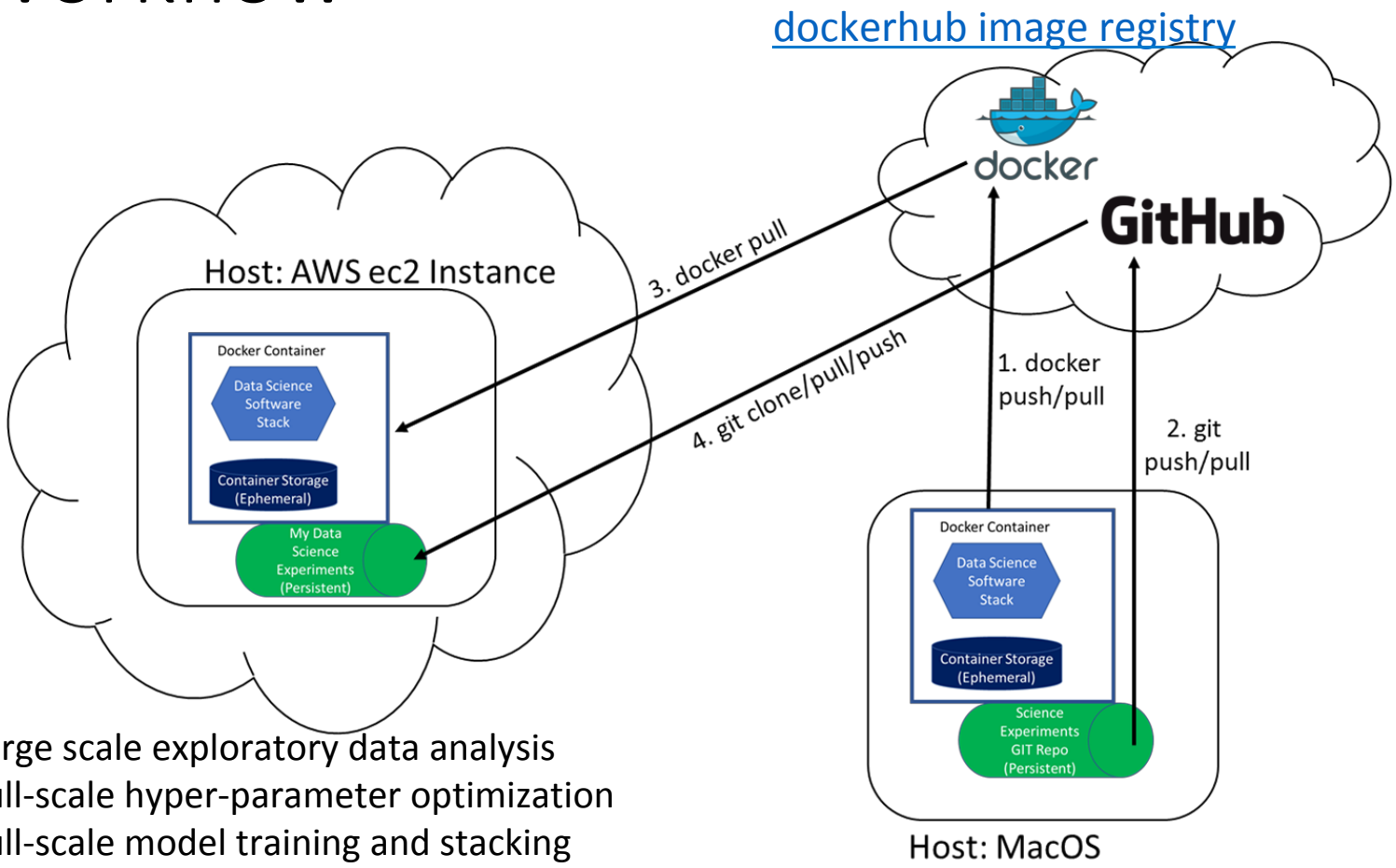
Built docker images can be found at dockerhub.com

System Requirements:

- MacOS
- Docker for Mac 18.06.1 ce (requires kubernetes enabled)
- Chrome Browser

Github Repo containing the work.

Workflow



- Large scale exploratory data analysis
- Full-scale hyper-parameter optimization
- Full-scale model training and stacking

- Exploratory data analysis
- Develop and test ML stacking pipeline with limited data
- Develop and test hyper-parameter optimization workflow with limited data

Steps

- MacOS:
  - Set up docker image definition
  - Define run-time configuration for container
  - Build docker image
  - Push image to docker hub
  - Do data science work
  - Push data science work to github
- AWS:
  - Start AWS ec2 instance and configure for use
  - Pull image from docker hub
  - Pull data science work from github
  - Continue data science work on AWS

Experiences

- It works
- Considerations
  - Some images take a while to build
  - Due to size some images take a while to download from dockerhub
  - Lots of command-line interface interactions
  - On AWS default location for docker images are on root volume, which is most likely ephemeral, necessitating subsequent downloads when instantiating new instances