Containerization for Reproducible Bioinformatics Research

Lessons from the NCI Cloud Resources and Hackathons

Steve Tsang, NCI-CBIIT
tsang@mail.nih.gov
Prepared for NCBI hackathon Sept 10-12, 2018

Recording - https://youtu.be/gQ2PcxyzW7s GitHub - https://github.com/stevetsa/nlmreproducibility
Slides - https://docs.google.com/presentation/d/16f8_z6tlULBc9nXLghCp04TGyVlZzKzazUNbdprROhs/edit

Reproducibility and Containerization

PERSPECTIVE

Reproducible Research in Computational Science

Roger D. Peng

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore MD 21205, USA.

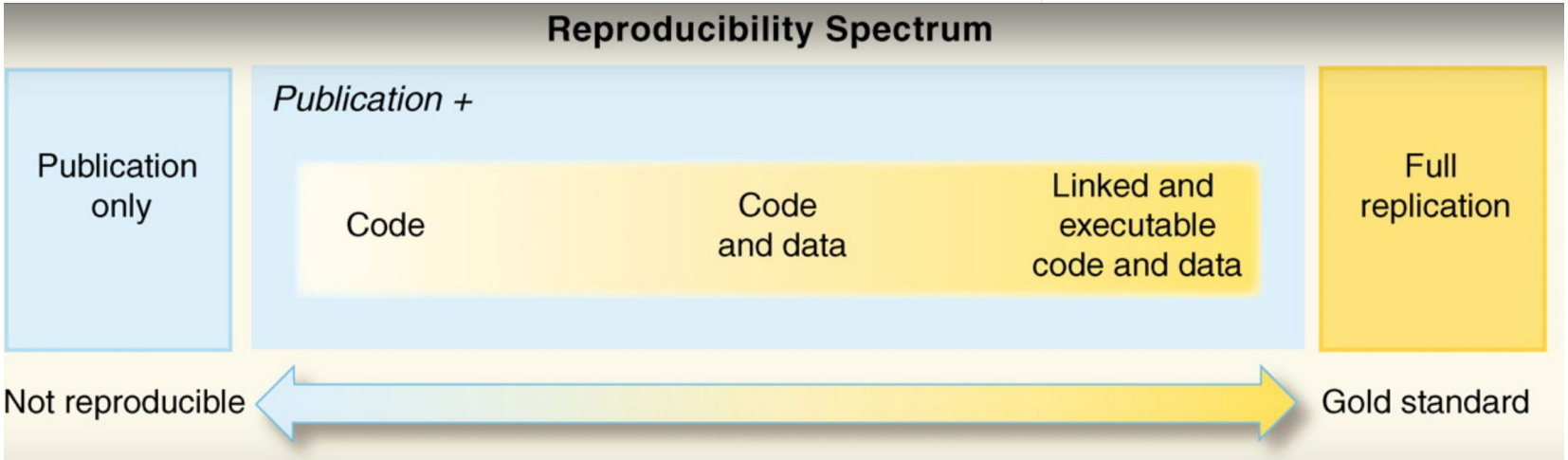
To whom correspondence should be addressed. E-mail: rpeng@jhsph.edu

– Hide authors and affiliations

Science 02 Dec 2011:
Vol. 334, Issue 6060, pp. 1226-1227
DOI: 10.1126/science.1213847

Science

Vol 334, Issue 6060
02 December 2011



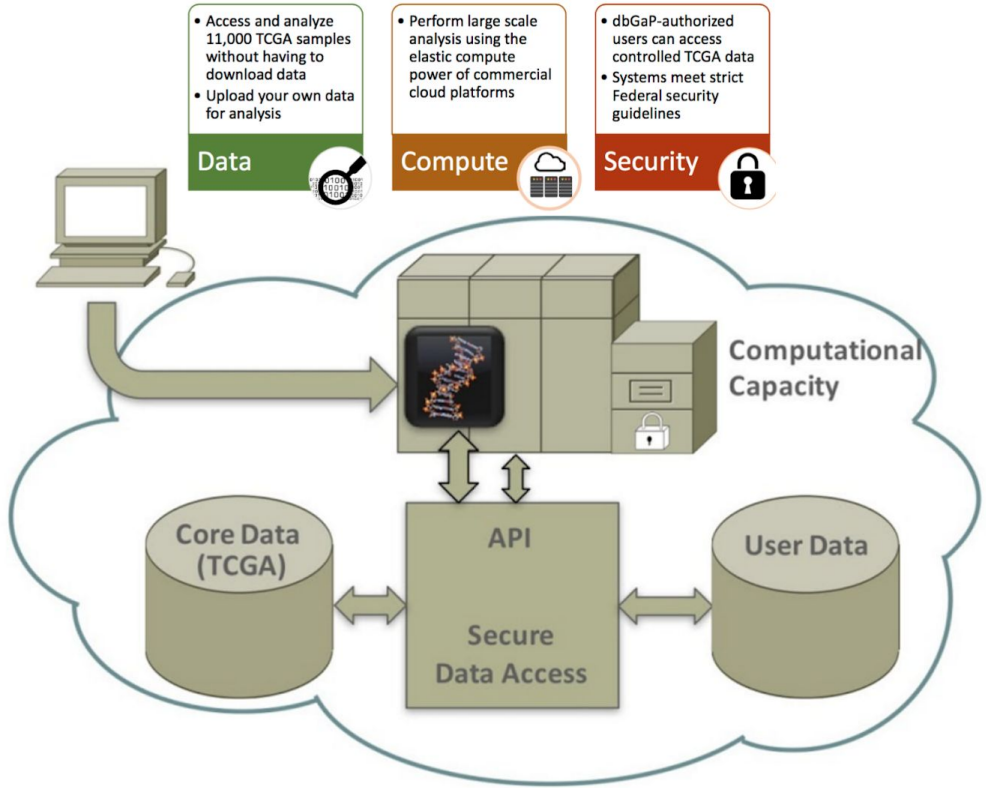
Reproducibility is HARD...

NIH Data Science Learning and Testing Hackathon

February 23, 2018!

The NIH will host a Learning and Testing Data Science hackathon on February 23rd, 2018 on the main campus in Bethesda, MD. Learners will test alpha and beta code that have been generated in full, collaborative development hackathons for a wide range of scientific problems, including general bioinformatics and genomic analyses in addition to text, image, and sequence processing. This event is for researchers who are in the early stages of their data science journey, including students and postdocs. Other non-scientific developers, mathematicians, or librarians in a similar educational place are also welcome! Learning in this event will be primarily

Cloud Resources Concept: Co-located Compute & Data



docker

COMMON WORKFLOW LANGUAGE

{wdl}

Google Cloud Platform

amazon web services

Democratize access to NCI-generated genomic & related data
Provide cost-effective computational capacity for the cancer research community

F.A.I.R. Analysis?

AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

PERSPECTIVE

Identifying and Overcoming Threats to Reproducibility, Replicability, Robustness, and Generalizability in Microbiome Research

Patrick D. Schloss[a]

[a]Department of Microbiology and Immunology, University of Michigan, Ann Arbor, Michigan, USA

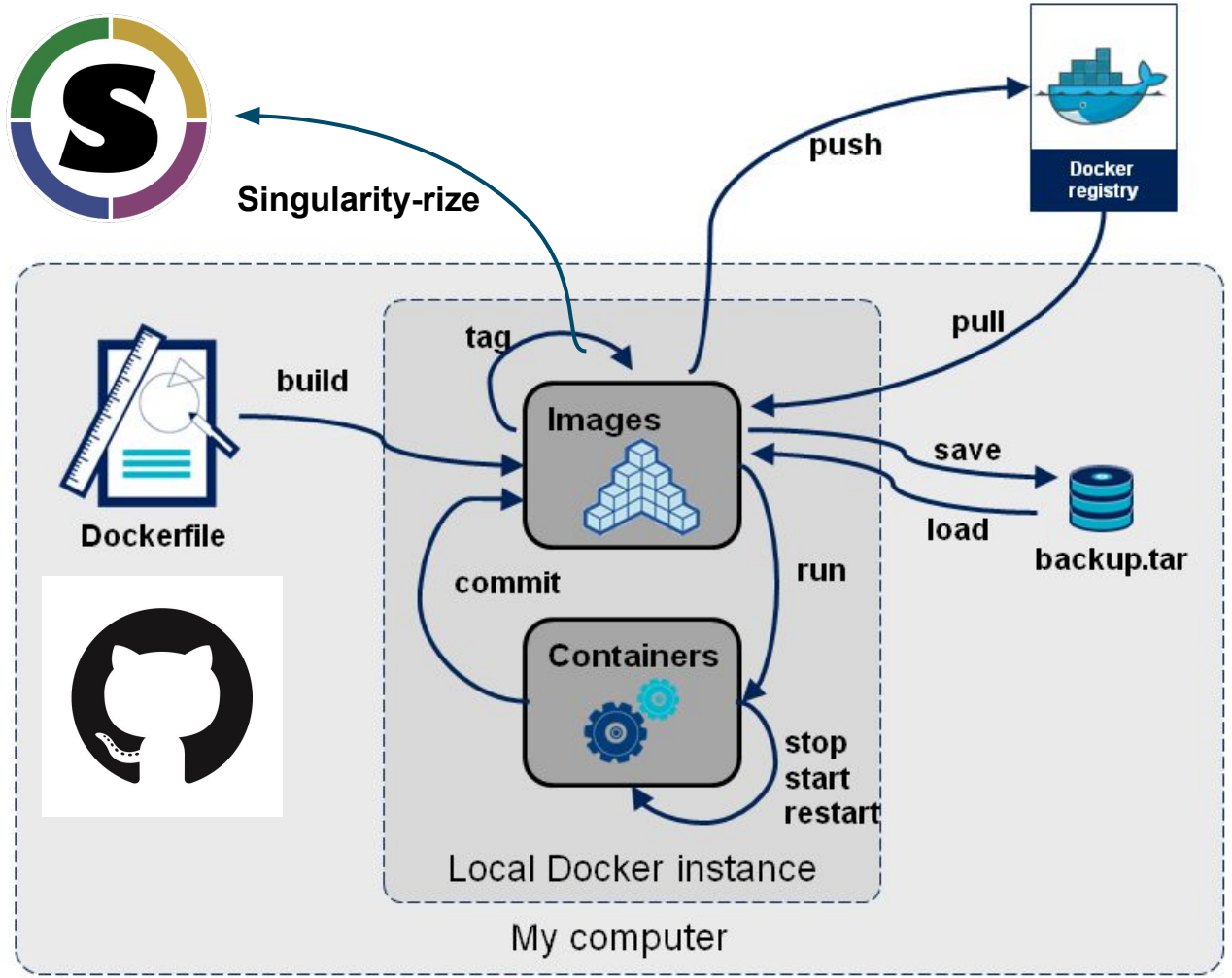
Best practices for increase the reproducibility and replicability of their work

Are Amazon Machine Images or Docker containers used to allow recreation of our work environment?

Are automated workflow tools like GNU Make and Common Workflow Language used to convert raw data into final tables, figures, and summary statistics?

Container/workflow Technologies

Container Concept



https://stackoverflow.com/questions/23735149/what-is-the-difference-between-a-docker-image-and-a-container

Docker's Layered Filesystem

$ more Dockerfile

```
FROM ubuntu:16.04
RUN apt-get install -y python3
```

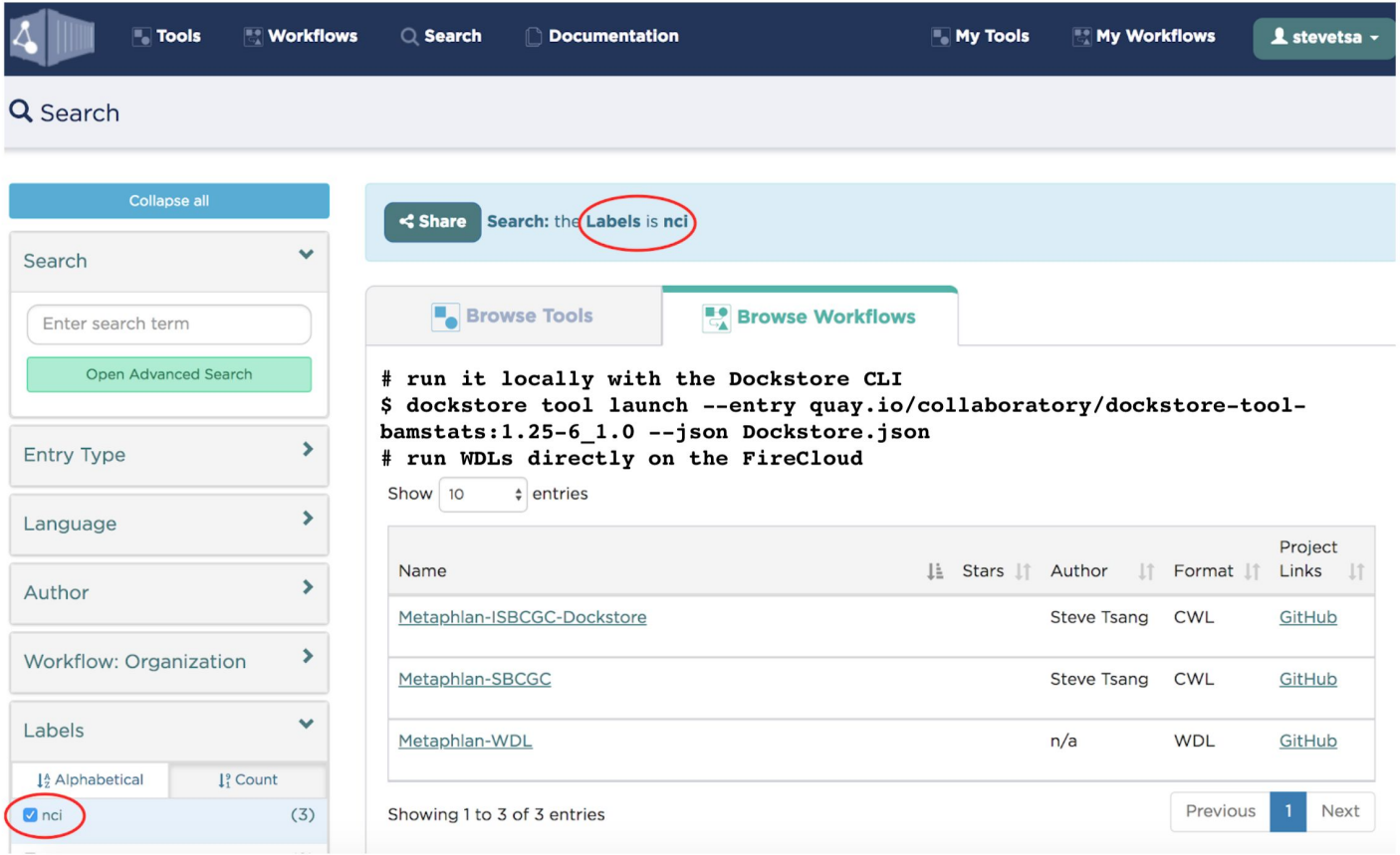
```
$ docker build -t sampleimage .
Step 1/2 : FROM ubuntu:16.04
 ---> 52b10959e8aa
Step 2/2 : RUN apt-get install -y python3
...
 ---> e0aac1c590b7
Successfully built e0aac1c590b7
Successfully tagged sampleimage:latest

$ docker images
REPOSITORY                  TAG                 IMAGE ID            CREATED                  SIZE
sampleimage                 latest              e0aac1c590b7        Less than a second ago   193MB

$ docker history sampleimage
IMAGE               CREATED                  CREATED BY                                      SIZE
COMMENT
e0aac1c590b7        Less than a second ago   /bin/sh -c apt-get install -y python3           37.3MB
52b10959e8aa        10 days ago              /bin/sh -c #(nop)  CMD ["/bin/bash"]            0B
<missing>           10 days ago              /bin/sh -c mkdir -p /run/systemd && echo '...   7B
<missing>           10 days ago              /bin/sh -c sed -i 's/^#\s*\(deb.*universe\...   2.76kB
<missing>           10 days ago              /bin/sh -c rm -rf /var/lib/apt/lists/*          0B
<missing>           10 days ago              /bin/sh -c set -xe   && echo '#!/bin/sh' >...   745B
<missing>           10 days ago              /bin/sh -c #(nop) ADD file:a83ab1826f43e88...   115MB
```

Dockerfile

```
####
# Kallisto - kallisto is a program for quantifying abundances of transcripts from RNA-Seq data
# Nicolas L Bray, Harold Pimentel, Páll Melsted and Lior Pachter, Near-optimal probabilistic RNA-seq quantification
# Nature Biotechnology 34, 525–527 (2016), doi:10.1038/nbt.3519
####

FROM ubuntu:17.10
#RUN rm /bin/sh
MAINTAINER Stev
RUN apt-get upd

RUN DEBIAN_FRON
 build-essential \
 gcc-multilib \
 apt-utils \
 zlib1g-dev \
 cmake \
 libhdf5-dev \
 git-all \
 autoconf \
 automake \
 libcurl4-openssl-dev

# Get latest source from releases
WORKDIR /opt
RUN git clone https://github.com/pachterlab/kallisto.git
                                                          slib

WORKDIR /opt/kallisto/ext/htslib
RUN autoconf && autoheader
WORKDIR /opt/kallisto
RUN mkdir build
WORKDIR build
RUN cmake ..
RUN make
RUN make install

COPY Dockerfile /opt/.
```

```
docker build -t stevetsa/kallisto:latest .
docker push stevetsa/kallisto:latest
docker run -v `pwd`:`pwd` -w `pwd` -i -t stevetsa/kallisto
```

Sharing Docker-based Tools



Discussions

Will the same Dockerfile always produce identical images?

Discussions

- Defining reproducibility
- Containerization allows you to run legacy software
  - Reproducibility vs “security”
- Containerization simplifies the process to run software
- Containerization provides an isolated environment for testing
- Sharing images/containers
  - Samtools in Dockerhub - defining “identical” images for tools and workflows
- Tool documentation
  Best practices (e.g. Dockerfiles) to minimize trial and error
- Training/education

NCI Containers and Workflows Interest Group

Objectives

- Initiate cross-NCI strategy to:
  - facilitate scientific computing standards, guidelines & best practices
  - share methods to promote reproducible science
  - democratize computational research and benefit the community using these methods
- Discuss approaches and possible technical solutions for describing scientific workflows and sharing containerized tools developed by NCI-funded programs

NCI Containers and Workflows Interest Group

- Approx. 130 members joined since Sept 2016
- Monthly meetings - https://goo.gl/gccfB7
  - Discussions on achieving objectives
    - Focus and direction of the interest group
  - Presentations/Lectures
    - Survey rapidly-evolving fields of container and workflow technologies and invite outside experts to inform and educate members of the Working Group
    - Use cases from Cloud Resources; Community efforts - GA4GH challenge, Dockstore, BioContainers, CWL; various scientific domains - genomics, microbe, neuroscience, imaging, etc.
  - Building a community of practice and discussing relevant topics in container and workflow technologies

Acknowledgements

National Cancer Institute - CBIIT
Tony Kerlavage
Allen Dearry
Juli Klemm
Tanja Davidsen
Durga Addepalli

NCI Cloud Resources

Sean Davis

NCBI
NIAID/RTB/GTS
UCSD

NIH NATIONAL CANCER INSTITUTE
Center for Biomedical Informatics & Information Technology

ATTAIN

Questions

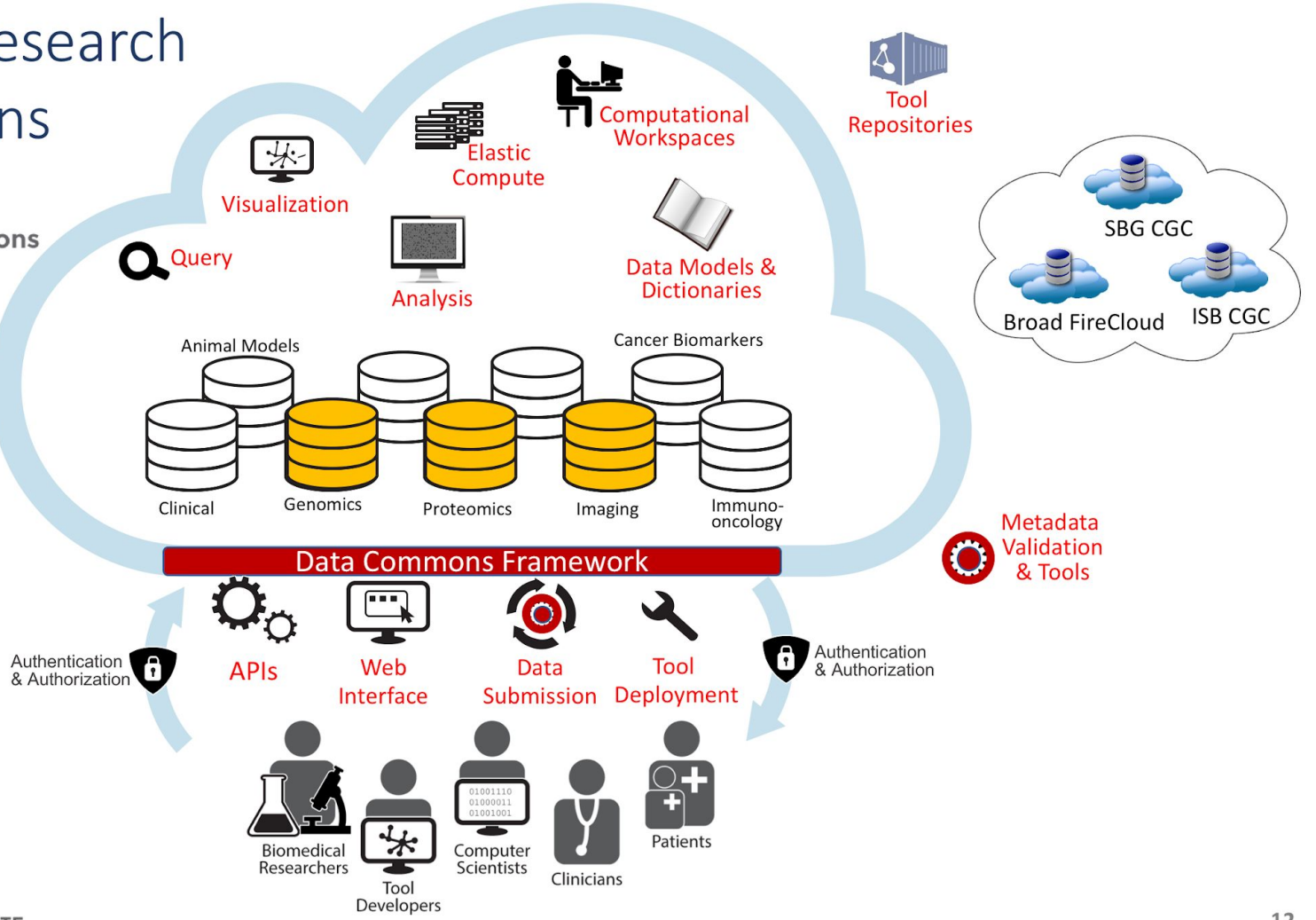

github
computing
workflows
metadata
reproducibility
environment
cgc
cancer
tools
cloud
fair
data
analysis
userfriendly
portability
open
reusability
genomics
language
interoperability
containers
workflow
proteomics
graphical
visualization
represent
meant
reproducible
publicly
facilitating
collaborators
image
research
sharing
allows
common
reusable
users
institutes
offers
consequently
built
also
form
depths
advancements
secured
minimal
given
perform
step
ability
goal
cwl
computation
described
proven
pathogens
genomic
accesscontrolled
requirements
description
resources
expanded
utility
omics
microbiome
genomes
toolset
will
alternative
commercial
currently
joined
integrated
single
architectures
interface
developed
utilize
costeffective
flexibility
interoperable
highperformance
present
softwareservice
setup
support

NCI Cancer Research Data Commons

NIH NATIONAL CANCER INSTITUTE
Genomic Data Commons

CPTAC
*Clinical Proteomics Tumor Analysis Consortium**

TCIA
*The Cancer Imaging Archive**



Data Contributors and Consumers

Three NCI Cloud Resources

Broad Institute

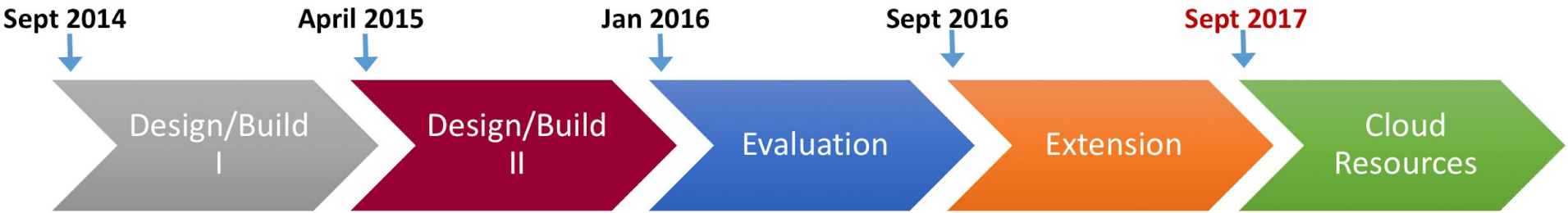
- PI: Anthony Philippakis
- Google Cloud
- Firehose in the cloud including Broad best practices workflows
- http://firecloud.org

Institute for Systems Biology

- PI: Ilya Shmulevich
- Google Cloud
- Leverage Google infrastructure; Novel query and visualization
- http://cgc.systemsbiology.net/

Seven Bridges Genomics

- PI: Brandi Davis-Dusenbery
- Amazon Web Services
- Interactive data exploration; > 30 public pipelines
- http://www.cancergenomicscloud.org



Links of interest - Courtesy of Sean Davis

- https://github.com/veggiemonk/awesome-docker
- https://dockstore.org/
- https://bioconda.github.io/ : Each package added to Bioconda also has a corresponding Docker BioContainer automatically created and uploaded to Quay.io.
- https://www.bioconductor.org/help/docker/
- https://biocontainers.pro/
- http://bioboxes.org/
- https://hub.docker.com

Additional Resources

[Bioconda](https://bioconda.github.io/)
[Bioconductor Docker Containers](https://www.bioconductor.org/help/docker/)
[BioContainers](https://biocontainers.pro/)
[Bioboxes](http://bioboxes.org/)
[NCBI Base Images](https://github.com/NCBI-Hackathons/HackathonBaseImages)

Awesome Containers
[Awesome Containers](https://github.com/tcnksm/awesome-container)
[Awesome Linux Containers](https://github.com/Friz-zy/awesome-linux-containers)
[Awesome Docker](https://github.com/veggiemonk/awesome-docker)

Additional Resources

Container Registries

[Docker Hub](http://www.dockerhub.com/);
[Quay.io](https://quay.io/repository/);
[Dockstore](https://dockstore.org/);
[Singularity Hub](https://www.singularity-hub.org);
[Google Container Registry](https://cloud.google.com/container-registry/);
[AWS Container Registry](https://aws.amazon.com/ecr/);
[Azure Container Registry](https://azure.microsoft.com/en-us/services/container-registry/);
[Seven Bridges Image Registry](https://docs.sevenbridges.com/docs/the-image-registry);
[GitLab Container Registry](https://about.gitlab.com/2016/05/23/gitlab-container-registry/);
[DGX/Nvidia Container Registry](https://www.nvidia.com/en-us/gpu-cloud/deep-learning-containers/)