

**MAKE YOUR LIFE
(AND ANALYSIS)
EASIER WITH
CONTAINERS**

AUDIENCE

- Are you a biologist?
- Have you heard of Docker?
- Not sure where to start?

YOU'VE COME TO THE RIGHT PLACE!

ME

- Software Engineer
- Build software infrastructure for researchers
- Help researchers to use computational tools
- Was a 'container skeptic'

CYVERSE

Helps researchers:

1. Learn about, and
2. Productively use

New tech like containers

ANALYSIS IS GETTING COMPLEX

- Multiple software packages (R, Python, etc.)
- With specific versions
- Have to work together
- On different platforms

THE PAIN

- Hard to install one-by-one
- Wasted effort and time
- Fragile, hard-to-reproduce analyses

HELP! MAKE IT STOP!

How we we make it easy to install & use things consistently?

CONTAINERS! *

New packages & apps are increasingly available as containers (BioContainers, etc.)

Speaker notes

- "BioContainers is an open source and community-driven framework which provides system-agnostic executable environments for bioinformatics software. BioContainers framework allows software to be installed and executed under an isolated and controllable environment."
- There will be a webinar specifically on BioContainers in the near future

CONCEPTS & TERMS

Speaker notes

- These are broad strokes

IMAGE

A self-contained, read-only 'snapshot' of your applications and packages, with all their dependencies

DOCKERFILE (OR SINGULARITY RECIPE)

Executable instructions (script) for:

- Creating an image
- Specifying the 'entry point' for the container

CONTAINER

- A 'running image'

Speaker notes

- The entry point is executed
- From Matt Rich's Singularity tutorial: The running container will have exactly the environment defined in the image.

DOCKER

- A server (sometimes called a daemon): A program that runs in the background, and handles life cycle of images and containers
- A command-line client: You use it to tell the server what to do

Download from: <https://www.docker.com/>

Speaker notes

The reason they made a separate server and client is so that you can have the server program running on a different machine from the the client

SINGULARITY

A way to run containers on HPC

Find out more: <https://www.sylabs.io/singularity/>

Speaker notes

- Because of computer security reasons HPC folks usually don't allow Docker
- It is easy to create Singularity images from Docker images
- With Singularity there is no separate server and client

WHAT ABOUT MY DATA?

Do not put your data in the image!

- Local data: 'Mount' it into a container when you start it
- Remote data: Pull into the container once it's running (e.g. CyVerse Data Store, S3, etc.)

Speaker notes

- "Bind mounts" make the host's filesystem accessible inside the container.

COMPUTE RESOURCES

I need more!

Talk to us. There are a few options, and it depends on what you need.

SHARING CONTAINERS

Image registries

Speaker notes

- Singularity Hub and Docker Hub

USING CONTAINERS

Speaker notes

- These are pre-recorded

DEMO: COMMAND LINE APP

```
mkdir -p ~/blast
cd ~/blast
docker pull biocontainers/blast:v2.2.31_cv2
docker run biocontainers/blast:v2.2.31_cv2 blastp -help
wget http://www.uniprot.org/uniprot/P04156.fasta
curl -O ftp://ftp.ncbi.nih.gov/\
refseq/D_rerio/mRNA_Prot/zebrafish.1.protein.faa.gz
gunzip zebrafish.1.protein.faa.gz
docker run -v $PWD:/data/ biocontainers/blast:v2.2.31_cv2 \
    makeblastdb -in zebrafish.1.protein.faa -dbtype prot
docker run -v $PWD:/data/ biocontainers/blast:v2.2.31_cv2 \
    blastp -query P04156.fasta -db zebrafish.1.protein.faa \
    -out results.txt
cat results.txt
```

Speaker notes

- See BioContainer Example: <https://biocontainers.pro/docs/101/running-example/>
- The above commands assumes that you have Docker installed

DEMO: WEB APP (JUPYTER)

```
cd ~  
git clone https://github.com/plyte/blastn-jupyter-docker.git  
cd blastn-jupyter-docker/  
docker build --tag blastn-jupyter-docker:local .  
docker run -p 8888:8888 blastn-jupyter-docker:local
```

Speaker notes

- See example: <https://github.com/plyte/blastn-jupyter-docker>
- Explain local IP & port

CYVERSE SUPPORT FOR CONTAINERS

1. Command line (Atmosphere)
2. Interactive apps (VICE)
3. HPC (XSEDE & OSG)

Speaker notes

- On Atmosphere run: `ezd` or `ezs`
- First will install Docker, the second Singularity

SUMMARY

- Package your analysis pipeline in a single container
- Everyone in your lab can have a consistent environment

NEXT TIME

- How to build containers
- Running on different platforms
- Science applications

LINKS & REFERENCES

- [Docker](#)
- [Singularity](#)
- [Play with Docker Classroom](#)
- [Katacode - Learn Docker](#)
- [CyVerse Container Camp materials](#)
- [Reproducible research with containers](#)
- [Upendra's Cybercarpentry workshop notes](#)
- [Tyson Swetnam's Container Camp Presentation](#)
- [Matthew Rich's Singularity workshop](#)
- [BioContainers](#)

THANKS!

- Nirav Merchant
- Upendra Devisetty
- Tyson Swetnam
- Blake Joyce
- Eric Lyons
- Ariella Gladstein
- Tina Lee
- Shelley Littin



CyVerse is supported by the National Science Foundation under Grants No. DBI-0735191, DBI-1265383 and DBI-1743442

