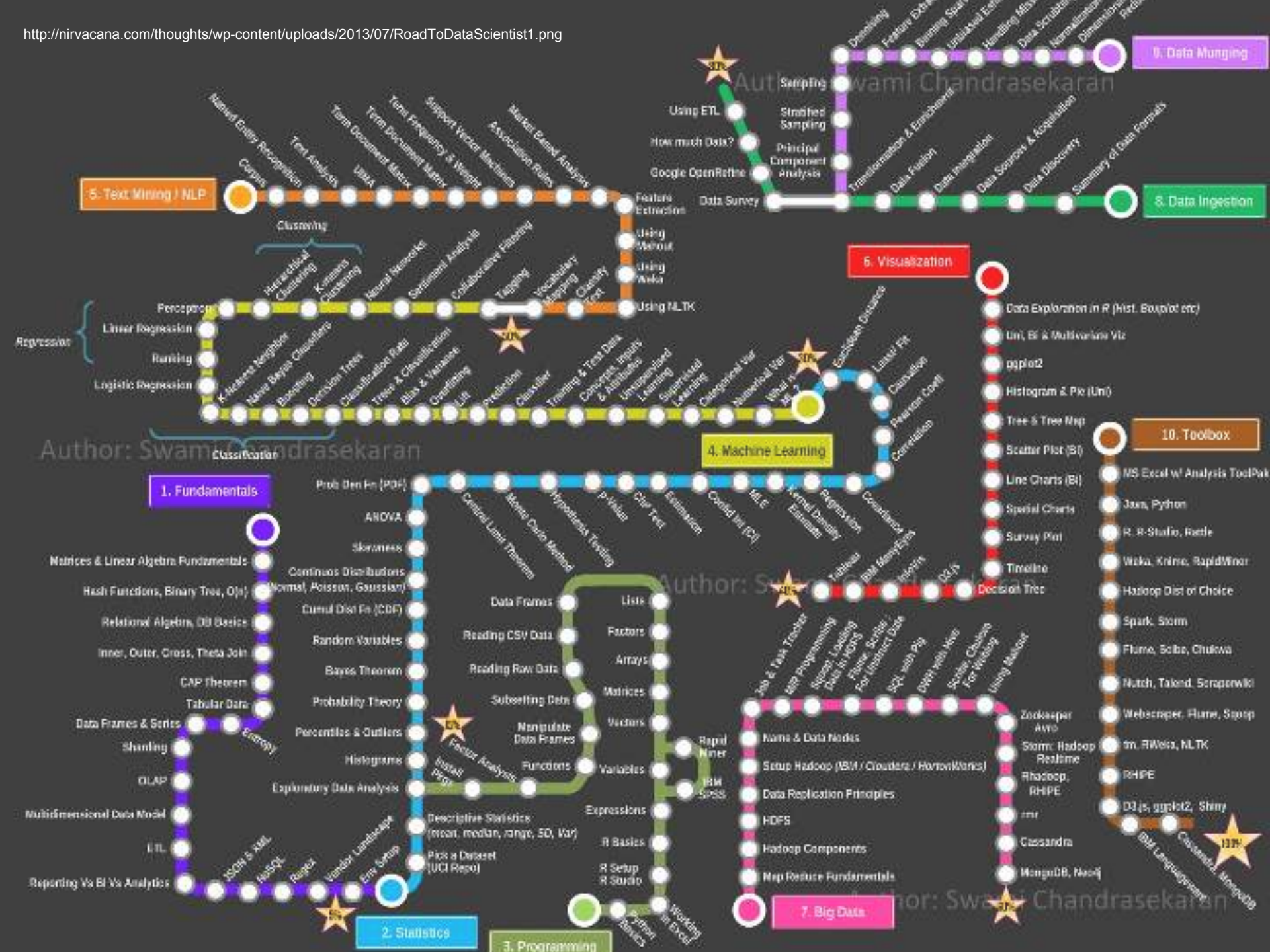
docker for Data Science

https://www.meetup.com/Data-Science-Meetup-Muenster

Daniel Nüst
Institute for Geoinformatics
University of Münster

@nordholmen | http://nördholmen.net

http://nirvacana.com/thoughts/wp-content/uploads/2013/07/RoadToDataScientist1.png



Docker for Data Science

http://blog.kaggle.com/2016/02/05/how-to-get-started-with-data-science-in-containers/

https://github.com/wiseio/datascience-docker (Hackday container - nice!)

http://www.datadan.io/containerized-data-science-and-engineering-part-2-dockerized-data-science/

http://www.slideshare.net/CalvinGiles/docker-for-data-science

https://civisanalytics.com/blog/data-science/2016/05/11/strata-2016-talk/

https://www.quora.com/What-are-use-cases-for-Docker-in-Data-Science-and-Machine-Learning

"**Isolation! Portability! Repeatability!**"

http://nirvacana.com/thoughts/wp-content/uploads/2013/07/RoadToDataScientist1.png

Agenda

What is Docker? Why?

What can it be used for?

Live Demo

Crossed fingers created by Michael A. Salter from Noun Project (https://thenounproject.com/richard.zeid/)

Why containerization?
Why Docker?

Motivation

Pets vs. Cattle

Service Model

- Pets are given names like pussinboots.cern.ch
- They are unique, lovingly hand raised and cared for
- When they get ill, you nurse them back to health

- Cattle are given numbers like vm0042.cern.ch
- They are almost identical to other cattle
- When they get ill, you get another one

- Future application architectures should use Cattle but Pets with strong configuration management are viable and still needed

http://www.slideshare.net/gmccance/cern-data-centre-evolution

Motivations for Docker in mainstream IT

https://www.docker.com/use-cases

CI/CD

Enable developers to develop and test applications more quickly and within any environment

DEVOPS

Break down barriers between Dev and Ops teams to improve the app development process

BIG DATA

Empower your enterprise to leverage big data analytics

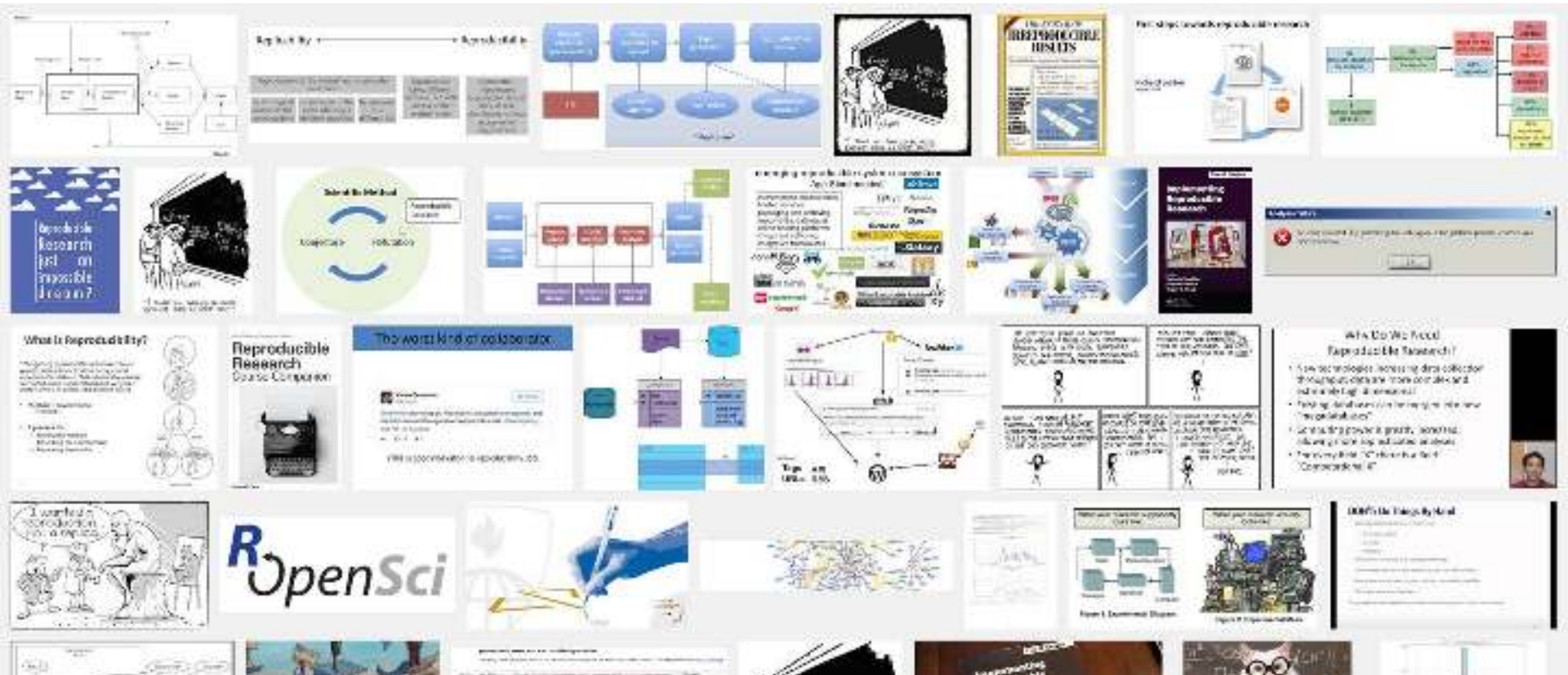
INFRASTRUCTURE OPTIMIZATION

Decrease infrastructure costs while increasing its efficiency

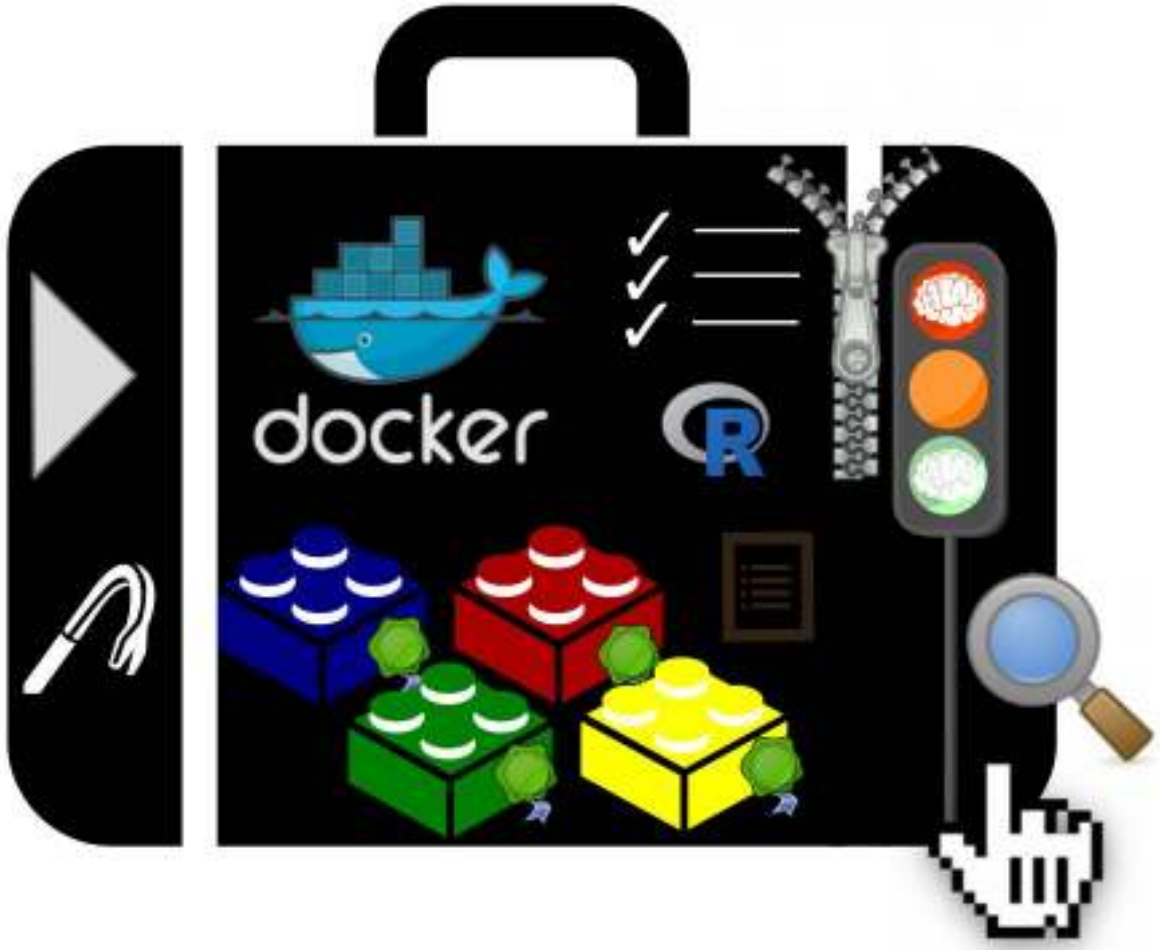
Motivation for Reproducible Research

Reproducibility is at the ♥ of

Science



Executable Research Compendium



o2r opening reproducible research

http://o2r.info

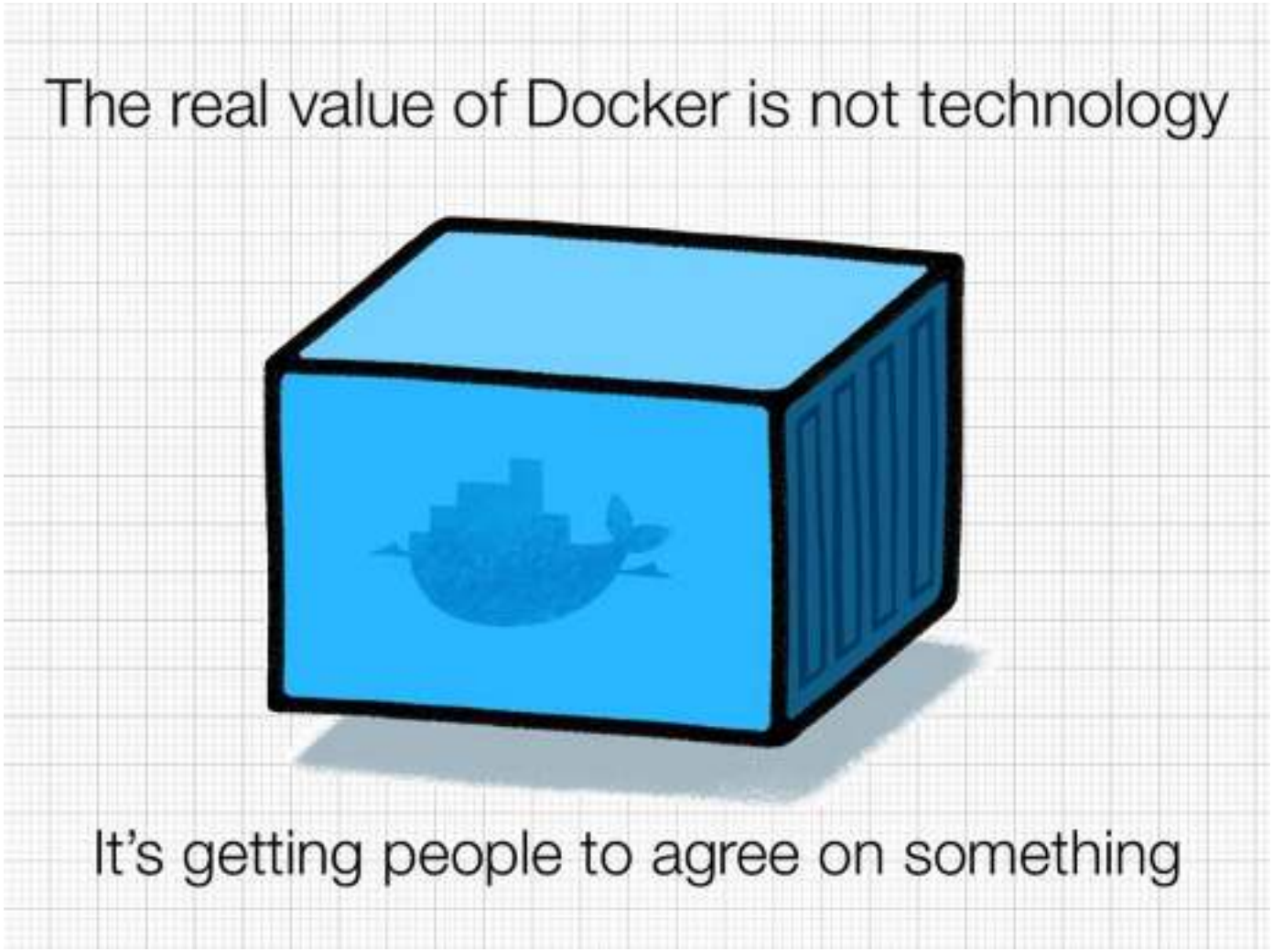
Motivation for Data Scientists - Proof by quote

> *"There was only one problem — all of my work was done in my local machine in R. People appreciate my efforts but they don't know how to consume my model because it was not "productionized" and the infrastructure cannot talk to my local model.* ***Hard lesson learned!****"*
> *-Robert Chang, data scientist at Twitter*
>
> *"Data engineers are often frustrated that data scientists produce inefficient and poorly written code, have little consideration for the maintenance cost of productionizing ideas, demand unrealistic features that skew implementation effort for little gain... The list goes on, but you get the point". -Jeff Magnusson, director of data platform Stitchfix*

But don't worry! There is a better way: **Dockerize your data science applications for ease of deployment, portability, and integration within your infrastructure**.

Home/Posts
About
Contact
Talks/Schedule
twitter
github
linkedin

via Daniel Whitenack,
http://www.datadan.io/containerized-data-science-and-engineering-part-2-dockerized-data-science/

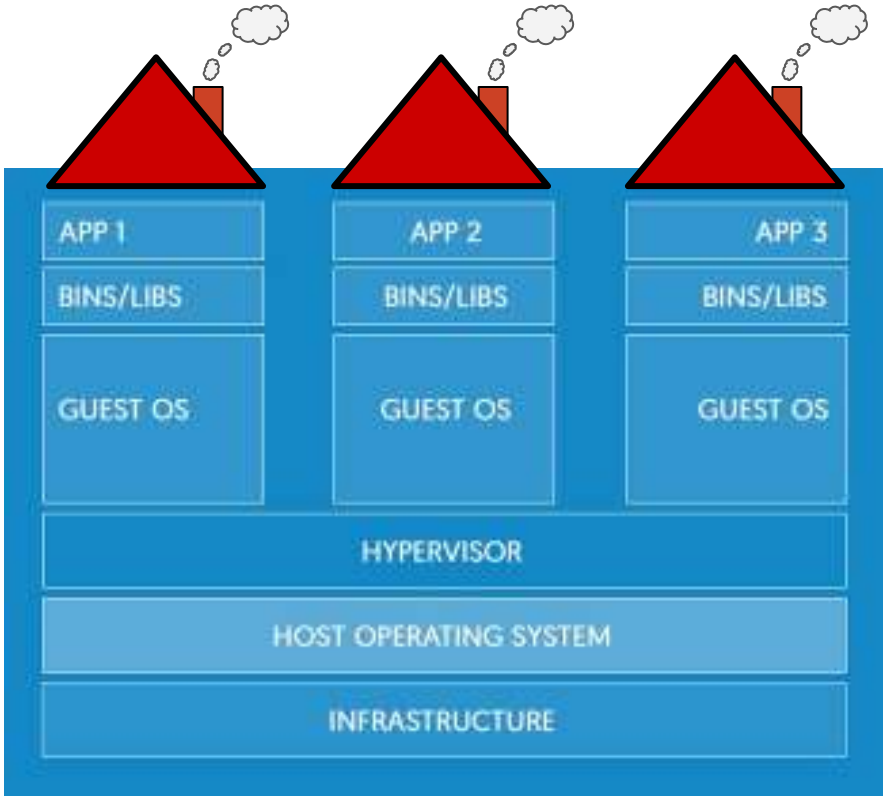


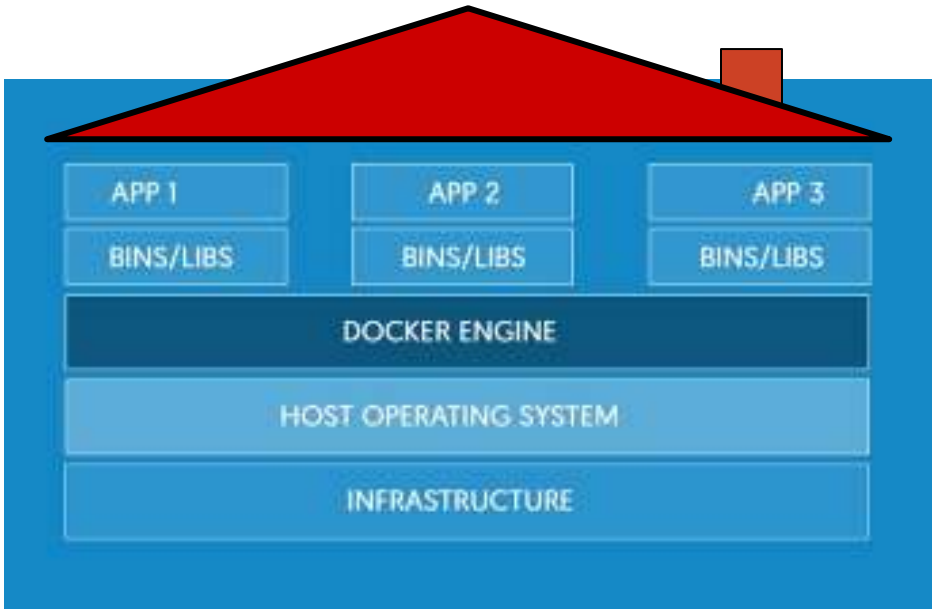
Slide by Docker inventor & Docker, Inc. CTO Solomon Hykes, DockerCon 2014

"CONTAINERIZATION"

Application packaging using docker

Houses vs. Appartments | “binary” vs. OS

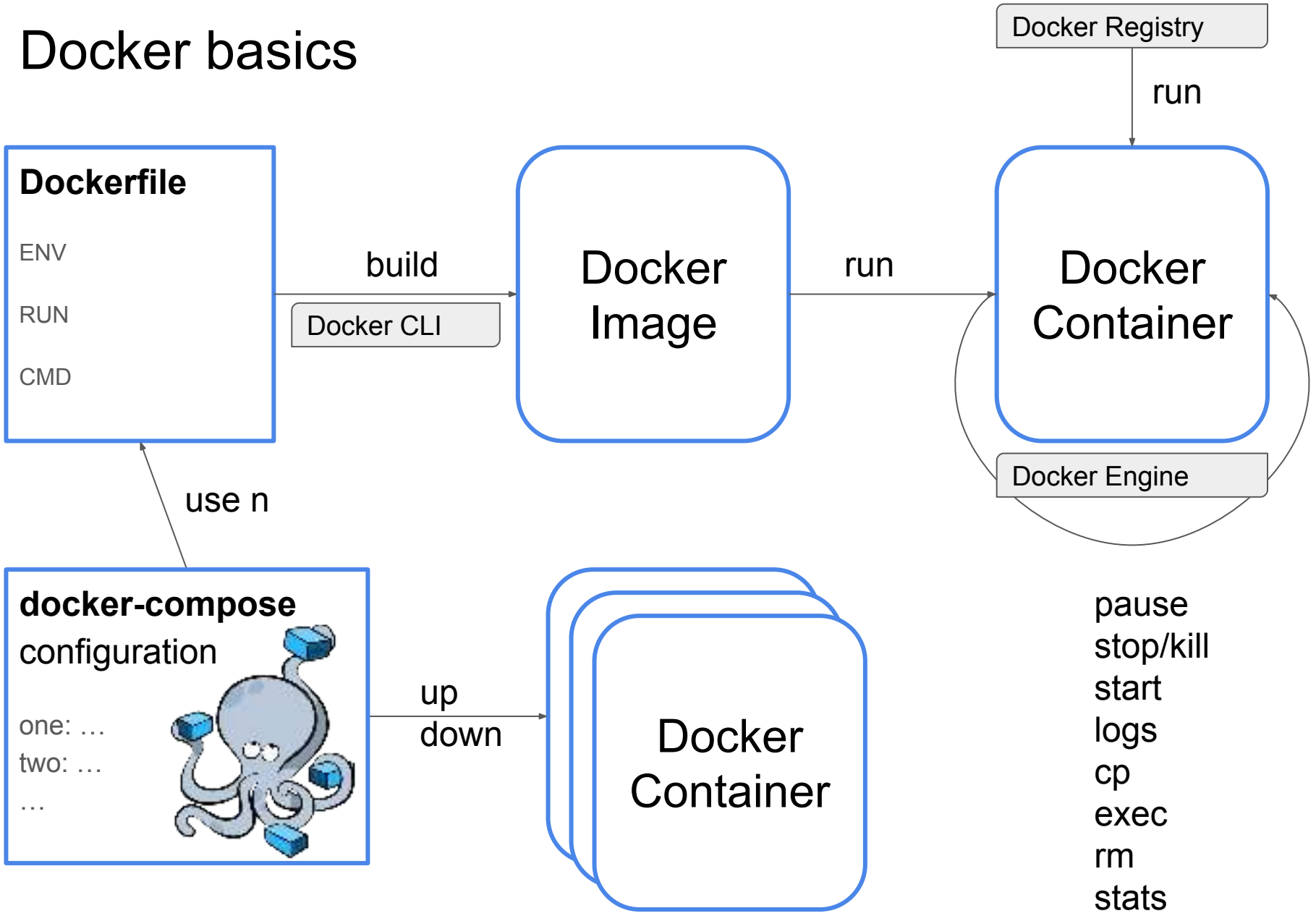




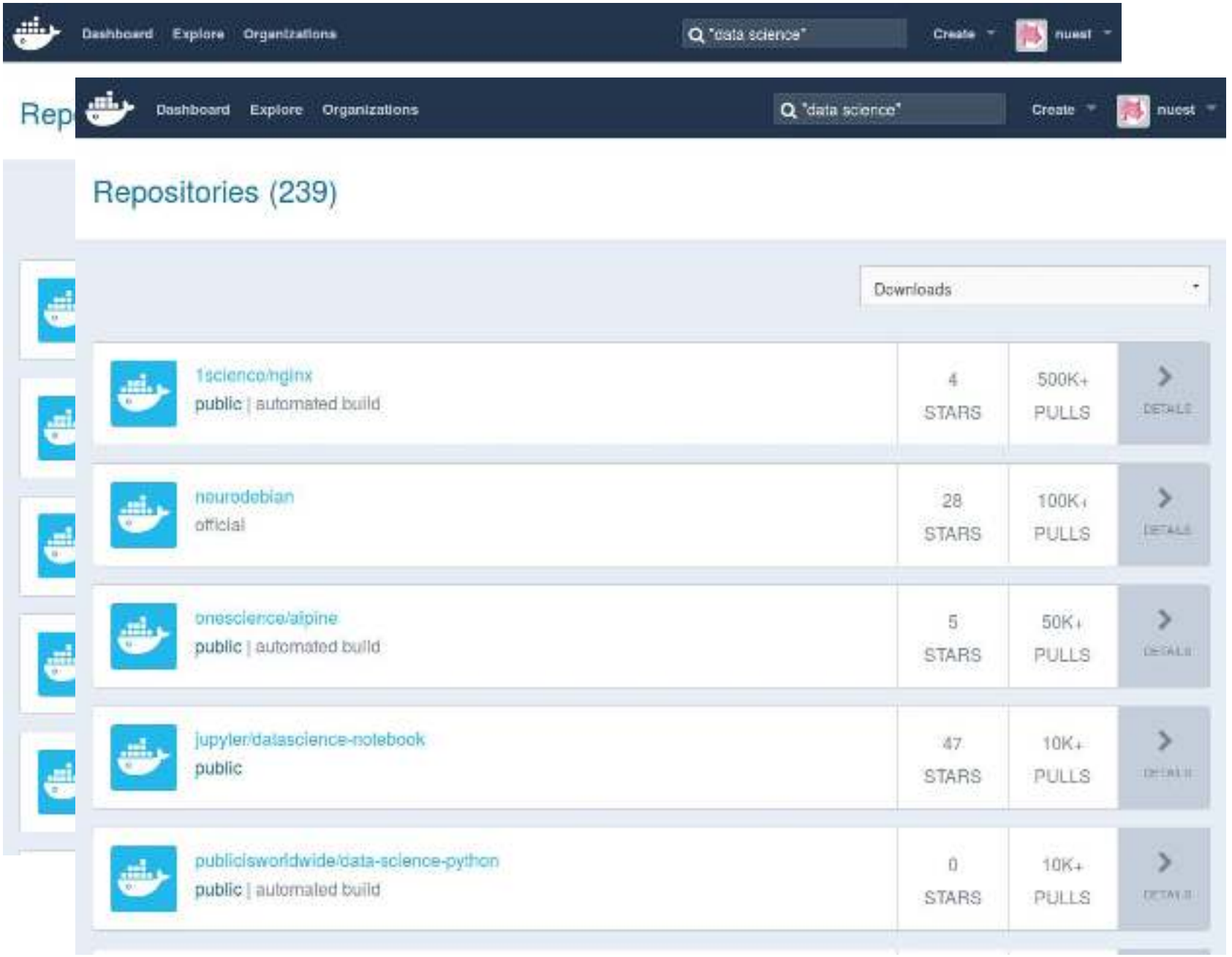
https://www.docker.com/what-docker
https://en.wikipedia.org/wiki/Operating-system-level_virtualization
https://youtu.be/ki8CZkutoxQ

kernel features
namespaces
libcontainer, LXC
cgroups
resources

Docker basics



Docker Hub


Dashboard Explore Organizations
"data science"
Create
nuest
Rep
Dashboard Explore Organizations
"data science"
Create
nuest
Repositories (239)
Downloads
1science/nginx
public | automated build
4 STARS
500K+ PULLS
neurodebian
official
28 STARS
100K+ PULLS
onescience/alpine
public | automated build
5 STARS
50K+ PULLS
jupyter/datascience-notebook
public
47 STARS
10K+ PULLS
publicisworldwide/data-science-python
public | automated build
0 STARS
10K+ PULLS

https://hub.docker.com/r/rocker/rstudio/

```
docker run --rm -it -p 8787:8787 rocker/rstudio
```

http://localhost:8787/ (rstudio/rstudio)

Great example: https://github.com/benmarwick/1989-excavation-report-Madjebebe

```
docker run --rm -it -p 8787:8787 benmarwick/mjb1989excavationpaper
```

http://localhost:8787/ (rstudio/rstudio)

Jupyter Notebook

https://github.com/jupyter/docker-stacks/tree/master/datascience-notebook

```
wget https://raw.githubusercontent.com/datascience-meetup-muenster/talks/master/meetup-01/what_are_data_scientists_and_tools.ipynb
```

```
docker run -it -p 8888:8888 -v $(pwd):/home/jovyan/work jupyter/datascience-notebook
```

http://localhost:8888/

See also https://www.dataquest.io/blog/docker-data-science/

ELK stack


```
git clone https://github.com/deviantony/docker-elk.git
cd docker-elk
# add filter to logstash/config/logstash.conf:
# filter {
#    grok {        match => { "message" => "  %{COMBINEDAPACHELOG}"}
#    }
#}

docker-compose up
```

http://localhost:5601/app/kibana
http://localhost:9200/

Example data: http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html

```
nc localhost 5000 > access_log_Aug95

docker-compose down (-v)
```

https://hub.docker.com/r/sverhoeven/cartodb/

```
docker run --rm -it -p 3000:3000 -p 8080:8080 -p 8181:8181
--name carto sverhoeven/cartodb

sudo sh -c 'echo 127.0.1.1 cartodb.localhost >> /etc/hosts'

docker run --rm -it -p 80:80 --link carto:cartodb.localhost
spawnthink/cartodb-nginx
```

http://cartodb.localhost dev/pass1234

Deep convolutional network in Amazon Cloud

https://civisanalytics.com/blog/data-science/2016/05/11/strata-2016-talk/

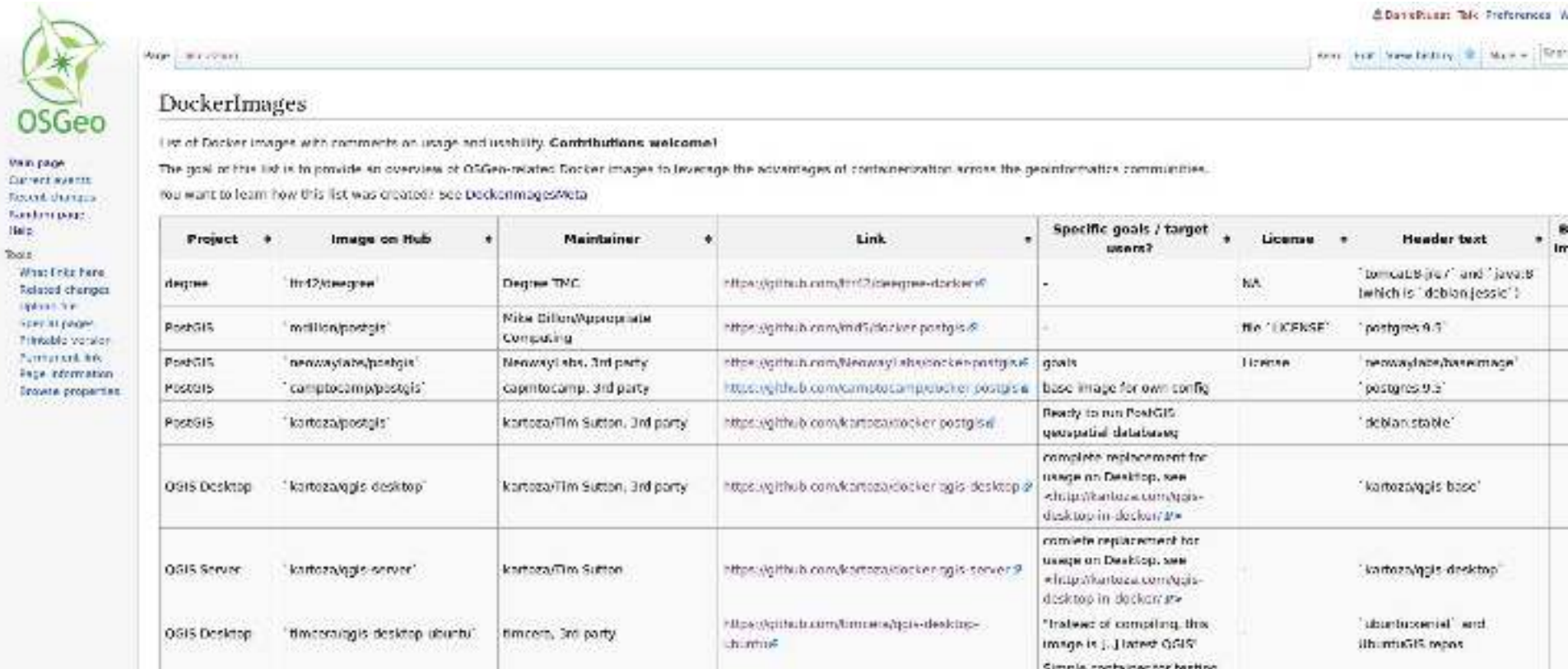
https://github.com/mdagost/pug_classifier

```
git clone https://github.com/mdagost/pug_classifier.git
docker run -d -p 8888:8888 -v /home/ubuntu/pug_classifier:/home/jovyan/work
mdagost/pug_classifier_notebook
```

Interested in “geo”? Go to OSGeo wiki +

https://wiki.osgeo.org/wiki/DockerImages
https://wiki.osgeo.org/wiki/DockerImagesMeta

http://geocontainers.org



Core arguments for Data **Scientists**

(all the Docker advantages… write once, biz ops, cloud, etc.)

Reproducibility

Project separation + don't clutter dev machine

Environment (re)creation, **documentation**

Adopt **good practices** on the way (dev cred)

Easy **collaboration**

Easy transition from testing to production

More from the Dockerverse

Docker **Machine** (provision remote host or clusters)

Docker **Cloud** (hosting of Dockerized apps)

Docker **Toolbox** (for older Mac and Windows OS)

OPEN CONTAINER INITIATIVE

Docker Universal Control Pane (custer management and monitoring UI)

Docker **Swarm mode** (container orchestration)

Docker Trusted Registry (own enterprise image storage)

kubernetes by Google (container orchestration)

Thanks for your attention!

What are your questions?

https://github.com/nuest

http://www.slideshare.net/nuest/

http://nördholmen.net

http://o2r.info daniel.nuest@wwu.de

o2r opening reproducible research

ulb Münster

ifgi Institute for Geoinformatics University of Münster

Want more Docker?
Watch Dockercon Keynote!

http://bit.ly/2cjrqQl