FAIR_bioinfo : Open Science and FAIR principles in a bioinformatics project

How to make a bioinformatics project more reproducible

C. Hernandez[1] T. Denecker[2] J. Sellier[2] G. Le Corguillé[2]
C. Toffano-Nioche[1]

[1]Institute for Integrative Biology of the Cell (I2BC)
UMR 9198, Université Paris-Sud, CNRS, CEA
91190 - Gif-sur-Yvette, France

[2]IFB Core Cluster taskforce

June 2021

Table of contents

1 Introduction to reproducibility

2 History management

3 Control your development environment
- Introduction
- Encapsulation
- The Conda stack
- What is Docker?
- Practical session

4 Workflow

5 Tracability with Notebook

6 IFB resources

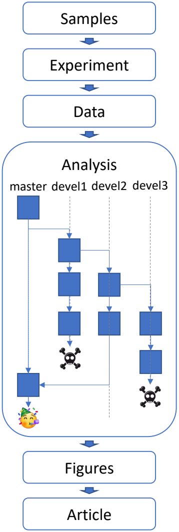
7 Sharing and dissemination

8 Conclusion

9 3rd Day

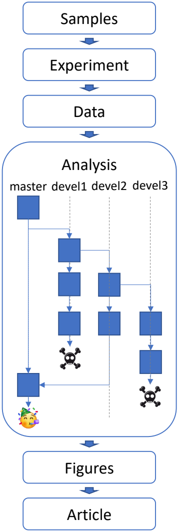
Introduction

A (not-so-uncommon) nightmare



Introduction

A (not-so-uncommon) nightmare



Samples

Experiment

Data

Analysis

master devel1 devel2 devel3

Figures

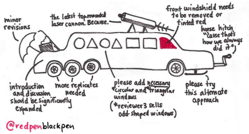
Article

Submit to a journal

Reviewer 1

👍

Reviewer 2



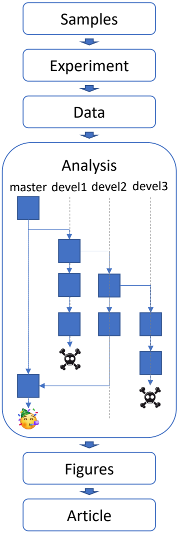
minor revisions
the latest top-mounted laser cannon. Because.
front windshield needs to be removed or tinted red
horse hitch "cause that's how we always did it"
introduction and discussion should be significantly expanded
more replicates needed
please add necessary *circular and *triangular windows
(*reviewer3 sells odd-shaped windows)
please try this alternate approach
@redpenblackpen

Reviewer 3

👻

Introduction

A (not-so-uncommon) nightmare



Samples

Experiment

Data

Analysis

master devel1 devel2 devel3

☠ ☠ 🥳

Figures

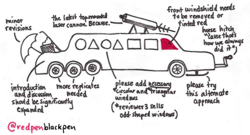
Article

Submit to a journal

Reviewer 1

👍

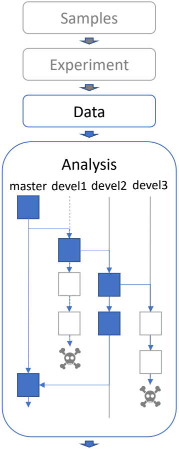
Reviewer 2



minor revisions
the latest top-mounted laser cannon. Because.
front windshield needs to be removed or tinted red
horse hitch "cause that's how we always did it"
introduction and discussion should be significantly expanded
more replicates needed
please add necessary *circular and *triangular windows
(*reviewer3 sells odd-shaped windows)
please try this alternate approach
@redpenblackpen

Reviewer 3

👻



Samples

Experiment

Data

Analysis

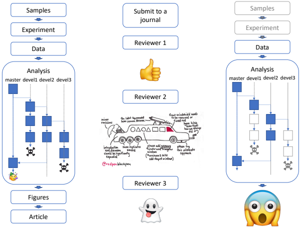
master devel1 devel2 devel3

☠ ☠

😱

Introduction

A (not-so-uncommon) nightmare



Samples

Experiment

Data

Analysis

master devel1 devel2 devel3

Figures

Article

Submit to a journal

Reviewer 1

Reviewer 2

Reviewer 3

Samples

Experiment

Data

Analysis

master devel1 devel2 devel3

What changed?

- Package
- Software
- Libraries
- Environment variables

- OS version
- Computer
- ..?

Different levels of encapsulation

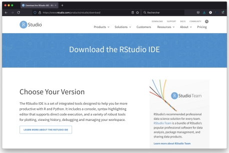
Goal : capture the system environment of applications (OS, packages, libraries,. . .) to control their execution.

- Hardware virtualisation (virtual machines)
- OS virtualisation (images and containers) docker
- Environment management CONDA

Encapsulation

Let’s say we want to install RStudio…

Install Rstudio ?


https://www.rstudio.com/products/rstudio/download/
Rechercher
Studio
DOWNLOAD SUPPORT DOCS COMMUNITY
Products Solutions Customers Resources About Pricing
Studio Team
LEARN MORE ABOUT THE RSTUDIO IDE
Learn more about RStudio Team

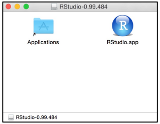
Download the RStudio IDE

Choose Your Version

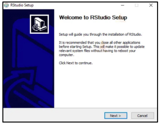
The RStudio IDE is a set of integrated tools designed to help you be more productive with R and Python. It includes a console, syntax-highlighting editor that supports direct code execution, and a variety of robust tools for plotting, viewing history, debugging and managing your workspace.

RStudio’s recommended professional data science solution for every team. RStudio Team is a bundle of RStudio’s popular professional software for data analysis, package management, and sharing data products.

MacOS



Windows



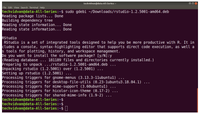
Welcome to RStudio Setup

Setup will guide you through the installation of RStudio.

It is recommended that you close all other applications before starting Setup. This will make it possible to update relevant system files without having to reboot your computer.

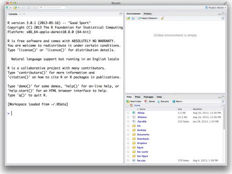
Click Next to continue.

Unix-based


techvidvan@data-All-Series: ~
File Edit View Search Terminal Help

```
techvidvan@data-All-Series:~$ sudo gdebi ~/Downloads/rstudio-1.2.5001-amd64.deb
Reading package lists... Done
Building dependency tree
Reading state information... Done
Reading state information... Done

RStudio
 RStudio is a set of integrated tools designed to help you be more productive with R. It in
cludes a console, syntax-highlighting editor that supports direct code execution, as well a
s tools for plotting, history, and workspace management.
Do you want to install the software package? [y/N]:y
(Reading database ... 181189 files and directories currently installed.)
Preparing to unpack .../rstudio-1.2.5001-amd64.deb ...
Unpacking rstudio (1.2.5001) over (1.2.5001) ...
Setting up rstudio (1.2.5001) ...
Processing triggers for gnome-menus (3.13.3-11ubuntu1) ...
Processing triggers for desktop-file-utils (0.23-1ubuntu3.18.04.1) ...
Processing triggers for mime-support (3.60ubuntu1) ...
Processing triggers for hicolor-icon-theme (0.17-2) ...
Processing triggers for shared-mime-info (1.9-2) ...
techvidvan@data-All-Series:~$
techvidvan@data-All-Series:~$
techvidvan@data-All-Series:~$ 
```

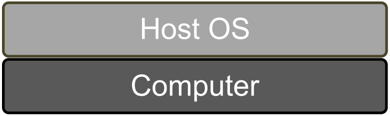
Use Rstudio



```
R version 3.0.1 (2013-05-16) -- "Good Sport"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin10.8.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

> |
```

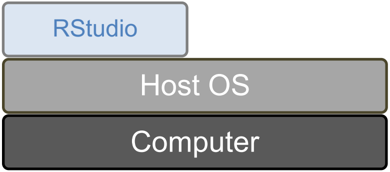
Name	Size	Modified
.RData	2.5 KB	Aug 16, 2013, 12:26 PM
.Rhistory	22.2 KB	Aug 16, 2013, 12:26 PM
.Rprofile	232 bytes	Jan 28, 2013, 1:54 PM
bin		
Desktop		
Documents		
Downloads		
Dropbox		
figure		
foo-cache		
foo-figure		
foo.cpp	379 bytes	Aug 4, 2013, 1:38 PM

Encapsulation



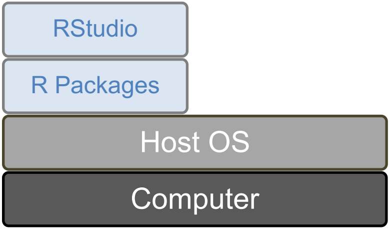
We started with a computer using a specific OS...

Encapsulation



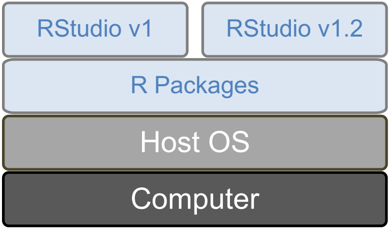
We started with a computer using a specific OS...
And inside this environment, we installed a new application.

Encapsulation



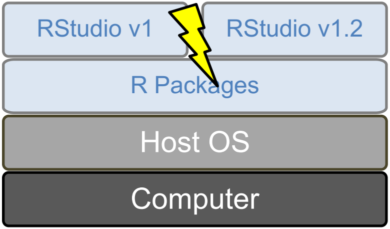
We started with a computer using a specific OS...
And inside this environment, we installed a new application.
Applications rely on dependencies, e.g. external libraries.

Encapsulation



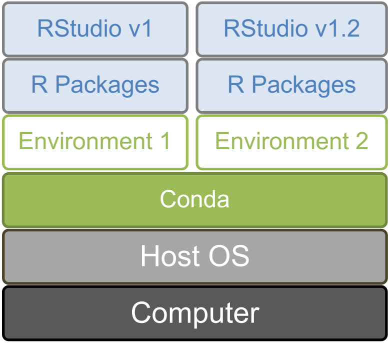
Usually dependencies of different applications don't interfere.
But what if we want to test the latest version of our favourite tool?
There might be conflicts. . .

Encapsulation



Usually dependencies of different applications don't interfere.
But what if we want to test the latest version of our favourite tool?
There might be conflicts. . .

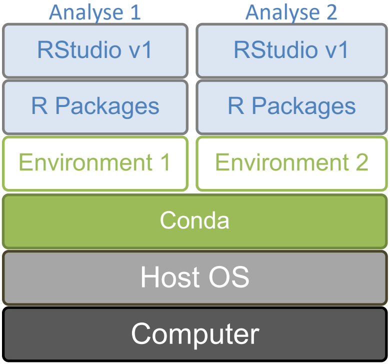
Encapsulation : managing environments



Idea : create separated environments for each application.

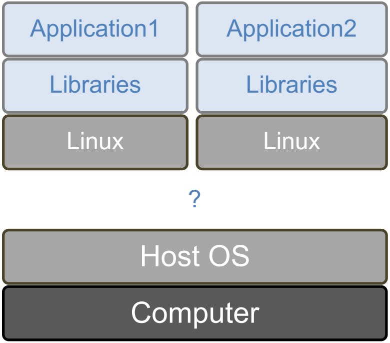
CONDA

Encapsulation : managing environments



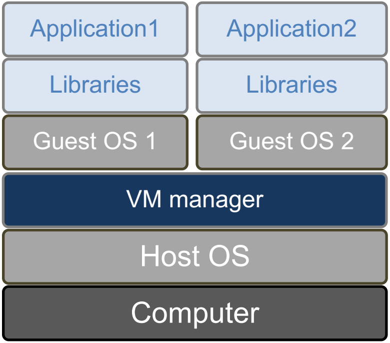
Idea : create separated environments for each application.
More versatile: create a new environment per analysis.

Encapsulation : hardware virtualisation



But what if we want to install a software from a different OS?

Encapsulation : hardware virtualisation

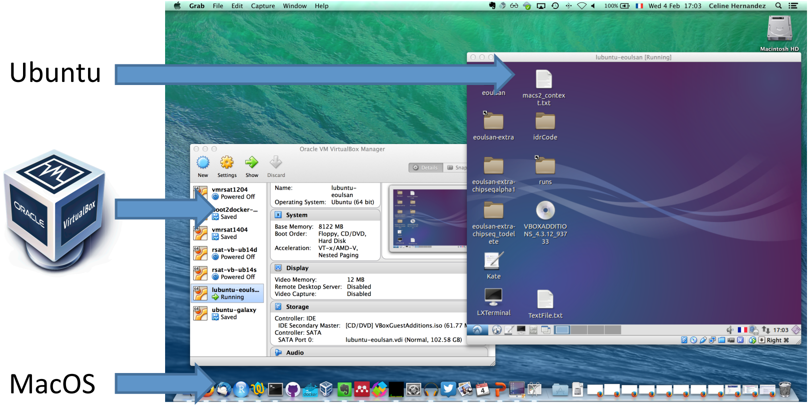


Idea: use virtual machines

Pros:

- Each application gets a completely different and independent environment
- Virtual machines can be transferred to another computer (using the same manager)

Encapsulation : hardware virtualisation



Ubuntu

MacOS

Grab File Edit Capture Window Help
100%
Wed 4 Feb 17:03 Celine Hernandez
Macintosh HD
lubuntu-eoulsan [Running]
eoulsan
macs2_context.txt
eoulsan-extra
idrCode
eoulsan-extra-chipseqalpha1
runs
eoulsan-extra-chipseq_todelete
VBOXADDITIONS_4.3.12_93733
Kate
LXTerminal
TextFile.txt
17:03
Right ⌘
Oracle VM VirtualBox Manager
New
Settings
Show
Discard
Details
Snap
Oracle
VirtualBox
docker

vmrsat1204 — Powered Off
boot2docker-... — Saved
vmrsat1404 — Saved
rsat-vb-ub14d — Powered Off
rsat-vb-ub14s — Powered Off
lubuntu-eouls... — Running
ubuntu-galaxy — Saved

```
Name:              lubuntu-
                   eoulsan
Operating System:  Ubuntu (64 bit)
```

System

```
Base Memory:   8122 MB
Boot Order:    Floppy, CD/DVD,
               Hard Disk
Acceleration:  VT-x/AMD-V,
               Nested Paging
```

Display

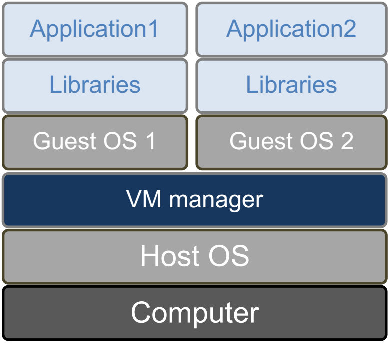
```
Video Memory:            12 MB
Remote Desktop Server:   Disabled
Video Capture:           Disabled
```

Storage

```
Controller: IDE
  IDE Secondary Master:   [CD/DVD] VBoxGuestAdditions.iso (61.77 M
Controller: SATA
  SATA Port 0:            lubuntu-eoulsan.vdi (Normal, 102.58 GB)
```

Audio

Encapsulation : hardware virtualisation



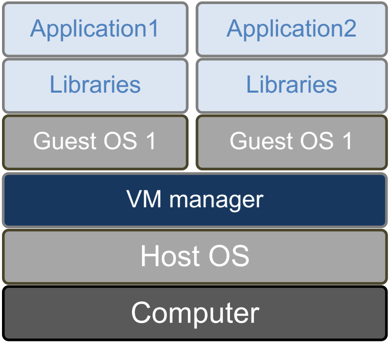
Idea: use virtual machines

Pros: transferable independent environments

Cons:

- Redundancy between VMs
- Heavy to set up
- No automation

Encapsulation : OS virtualisation



Application1 | Application2

Libraries | Libraries

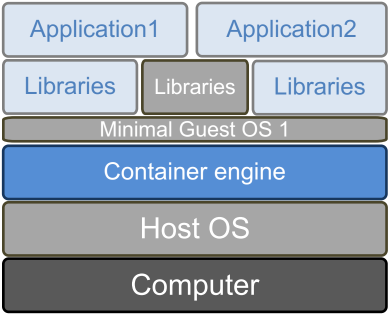
Guest OS 1 | Guest OS 1

VM manager

Host OS

Computer

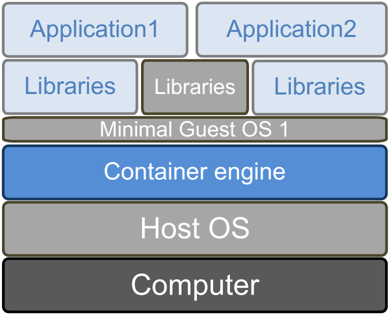
Encapsulation : OS virtualisation



Idea: "trick" applications into believing that they are in a different OS than the host's

Avoid redundancy.

Encapsulation : OS virtualisation

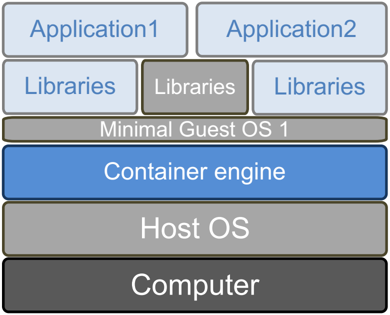


OS virtualisation vs hardware virtualisation

Pros:

- Speed
  - Installation is faster
  - No boot time
- Lightweight
  - Minimal base OS
  - Minimal libraries and application set
- Easy sharing of applications

Encapsulation : OS virtualisation

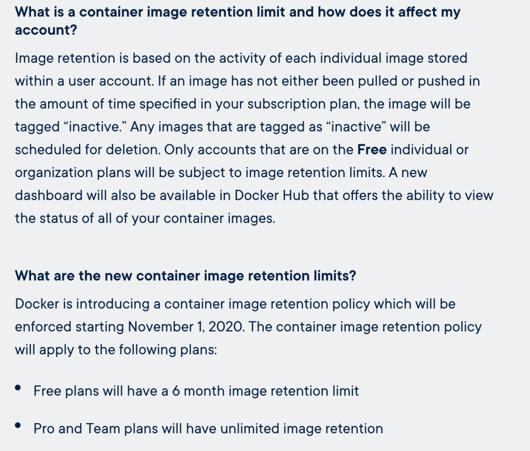


Cons:

- Singularity to use images on a cluster
- Changes of policies of the Docker company

Docker policy

Update of the Docker Image retention policy (13/08/2020)



What is a container image retention limit and how does it affect my account?

Image retention is based on the activity of each individual image stored within a user account. If an image has not either been pulled or pushed in the amount of time specified in your subscription plan, the image will be tagged "inactive." Any images that are tagged as "inactive" will be scheduled for deletion. Only accounts that are on the **Free** individual or organization plans will be subject to image retention limits. A new dashboard will also be available in Docker Hub that offers the ability to view the status of all of your container images.

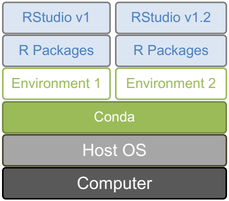
What are the new container image retention limits?

Docker is introducing a container image retention policy which will be enforced starting November 1, 2020. The container image retention policy will apply to the following plans:

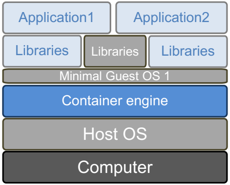
- Free plans will have a 6 month image retention limit
- Pro and Team plans will have unlimited image retention

https://www.docker.com/pricing/retentionfaq

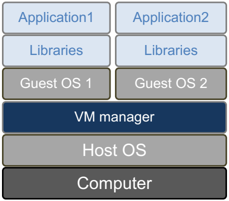
Encapsulation



RStudio v1
RStudio v1.2
R Packages
R Packages
Environment 1
Environment 2
Conda
Host OS
Computer

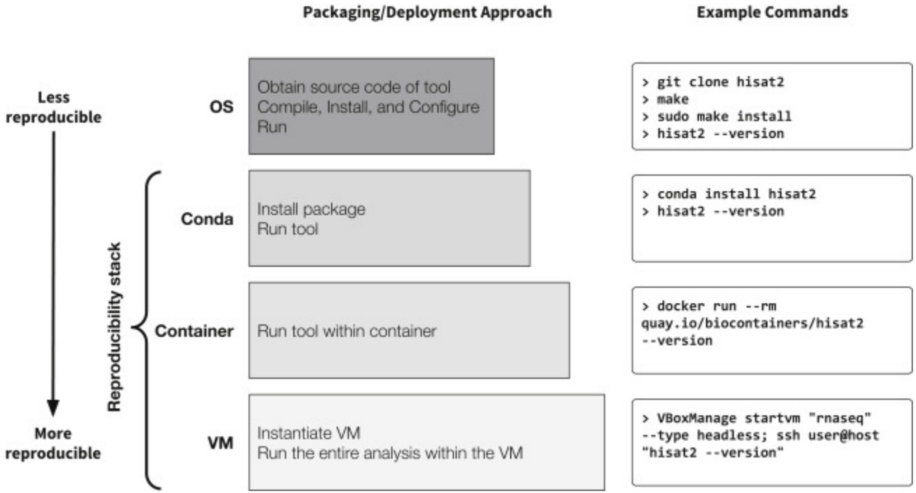


Application1
Application2
Libraries
Libraries
Libraries
Minimal Guest OS 1
Container engine
Host OS
Computer



Application1
Application2
Libraries
Libraries
Guest OS 1
Guest OS 2
VM manager
Host OS
Computer

Encapsulation and reproducibility stack



Packaging/Deployment Approach

Example Commands

Less reproducible

More reproducible

Reproducibility stack

OS

Obtain source code of tool
Compile, Install, and Configure
Run

```
> git clone hisat2
> make
> sudo make install
> hisat2 --version
```

Conda

Install package
Run tool

```
> conda install hisat2
> hisat2 --version
```

Container

Run tool within container

```
> docker run --rm
quay.io/biocontainers/hisat2
--version
```

VM

Instantiate VM
Run the entire analysis within the VM

```
> VBoxManage startvm "rnaseq"
--type headless; ssh user@host
"hisat2 --version"
```

Practical Computational Reproducibility in the Life Sciences - Björn Grüning et al (2018)

Docker

What is Docker?

Docker is not very “old”

- First commit January 2013
- First version March 2013
- Version 1.0 in June 2014

But its adoption was fast

- Officially packaged in Ubuntu since 2014 (v14.04)

What is Docker?

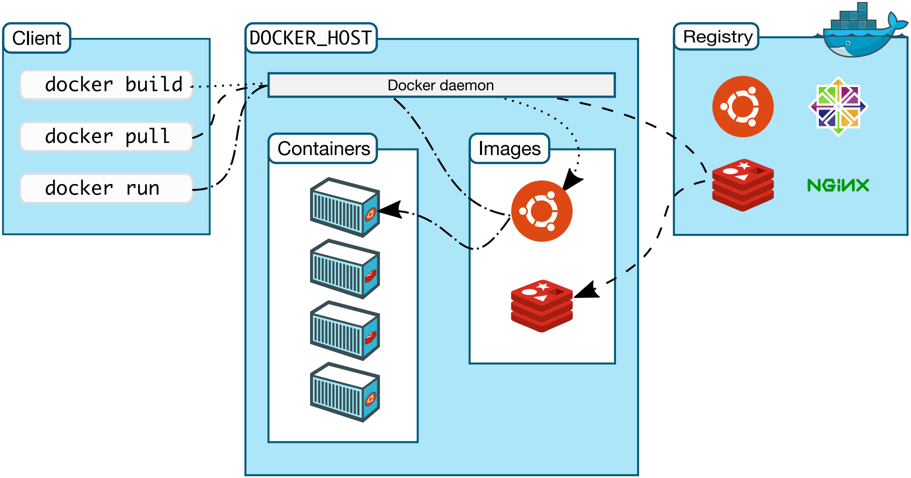
Image

- Set of libraries and functions
- Fixed. Cannot be modified
- Can be stored/shared online
- Can be automatically built

Container

- "Active image"
- Can be modified (interactive)
- Can be turned into an image
- One image, many containers

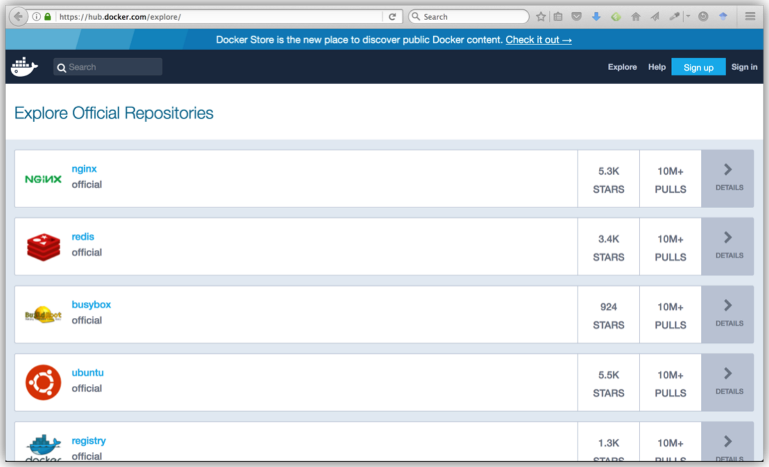
What is Docker?



(https://docs.docker.com/get-started/overview/)

What is Docker?

DockerHub



https://hub.docker.com/explore/

Docker Store is the new place to discover public Docker content. Check it out →

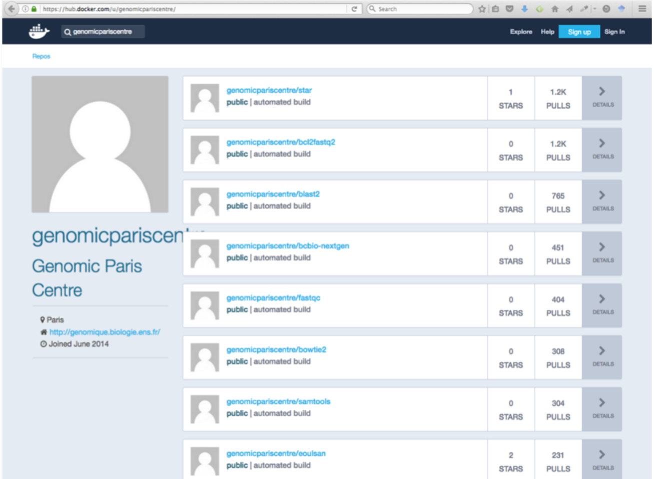
Explore Official Repositories

nginx	official	5.3K STARS	10M+ PULLS	DETAILS
redis	official	3.4K STARS	10M+ PULLS	DETAILS
busybox	official	924 STARS	10M+ PULLS	DETAILS
ubuntu	official	5.5K STARS	10M+ PULLS	DETAILS
registry	official	1.3K STARS	10M+ PULLS	DETAILS

(https://hub.docker.com/)

What is Docker?

Usermade images (1/2)



genomicpariscen

Genomic Paris Centre

Paris

http://genomique.biologie.ens.fr/

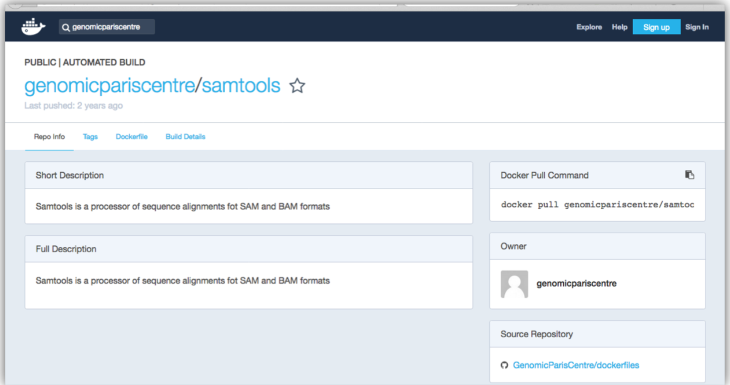
Joined June 2014

Repository	Stars	Pulls
genomicpariscentre/star public | automated build	1 STARS	1.2K PULLS
genomicpariscentre/bcl2fastq2 public | automated build	0 STARS	1.2K PULLS
genomicpariscentre/blast2 public | automated build	0 STARS	765 PULLS
genomicpariscentre/bcbio-nextgen public | automated build	0 STARS	451 PULLS
genomicpariscentre/fastqc public | automated build	0 STARS	404 PULLS
genomicpariscentre/bowtie2 public | automated build	0 STARS	308 PULLS
genomicpariscentre/samtools public | automated build	0 STARS	304 PULLS
genomicpariscentre/eoulsan public | automated build	2 STARS	231 PULLS

(urlhttps://hub.docker.com/u/genomicpariscentre/)

What is Docker?

Usermade images (2/2)
Be critical!



PUBLIC | AUTOMATED BUILD

genomicpariscentre/samtools ☆

Last pushed: 2 years ago

Short Description

Samtools is a processor of sequence alignments fot SAM and BAM formats

Full Description

Samtools is a processor of sequence alignments fot SAM and BAM formats

Docker Pull Command

```
docker pull genomicpariscentre/samtoc
```

Owner

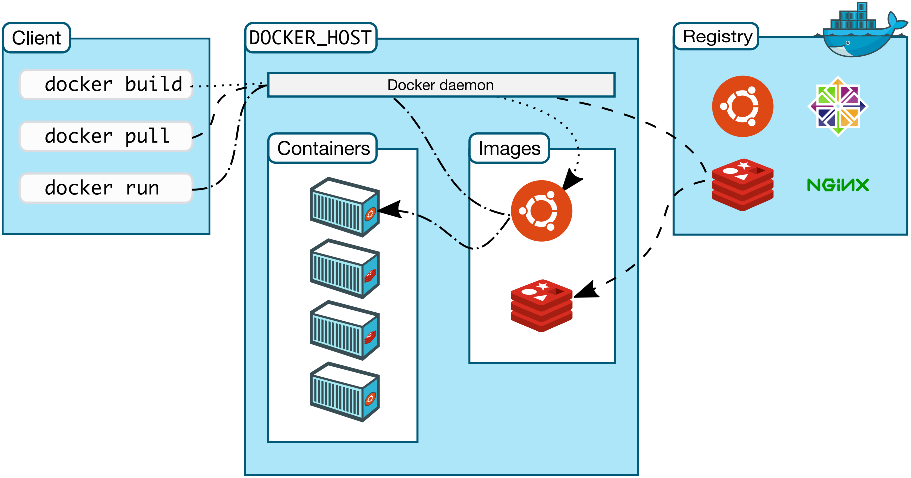
genomicpariscentre

Source Repository

GenomicParisCentre/dockerfiles

(https://hub.docker.com/r/genomicpariscentre/samtools/)

What is Docker?



(https://docs.docker.com/get-started/overview/)

What is Docker?

Other commands :

- docker images : list images available locally
- docker ps : status of containers
- docker rm : delete a container
- docker rmi : delete an image
- ...

(More details during the practical session.)