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]

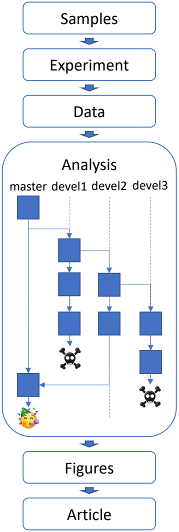
[1]Institute for Integrative Biology of the Cell (I2BC)
UMR 9198, Université Paris-Sud, CNRS, CEA
91190 - Gif-sur-Yvette, France

[2]IFB Core Cluster taskforce

June 2021

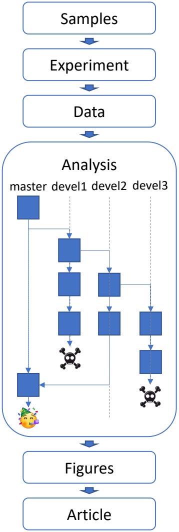
Introduction

A (not-so-uncommon) nightmare



Introduction

A (not-so-uncommon) nightmare



Samples

Experiment

Data

Analysis

master devel1 devel2 devel3

Figures

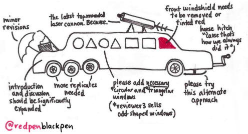
Article

Submit to a journal

Reviewer 1

👍

Reviewer 2



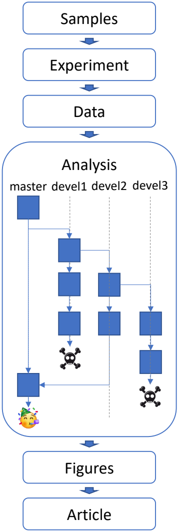
minor revisions
the latest top-mounted laser cannon. Because.
front windshield needs to be removed or tinted red
horse hitch "cause that's how we always did it"
introduction and discussion should be significantly expanded
more replicates needed
please add necessary *circular and *triangular windows (*reviewer3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

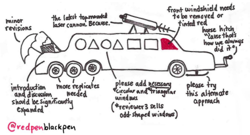
Figures

Article

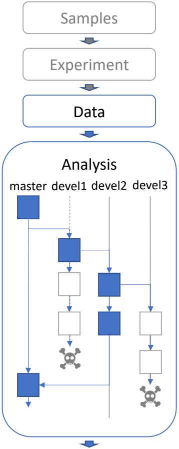
Submit to a journal

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

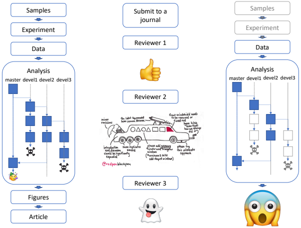
Data

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
minor revisions
the latest top-model 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 (+reviewer 3 adds odd-shaped windows)
please try this alternate approach
@redpenblackpen
Reviewer 3
Samples
Experiment
Data
Analysis
master devel1 devel2 devel3

What changed?

- Package
- Software
- Libraries
- Environment variables

- OS version
- Computer
- ..?

Different levels of encapsulation

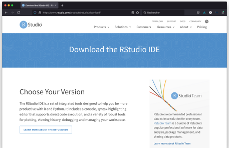
Goal : capture the system environment of applications (OS, packages, libraries,. . .) to control their execution.

- 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
RStudio
DOWNLOAD SUPPORT DOCS COMMUNITY
Products Solutions Customers Resources About Pricing

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.

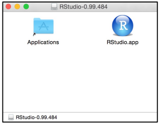
LEARN MORE ABOUT THE RSTUDIO IDE

RStudio Team

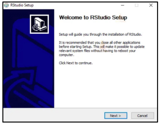
RStudio's recommended professional data science solution for every team. RStudio Team is a bundle of RStudio's popular professional software for data analysis, package management, and sharing data products.

Learn more about RStudio Team

MacOS



Windows



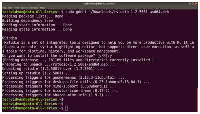
Welcome to RStudio Setup

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.

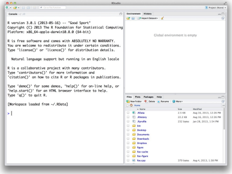
Click Next to continue.

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 (3.13.3-1ubuntu1) ...
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:~$ 
```

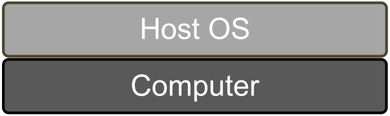
Use Rstudio



```
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]

> 
```

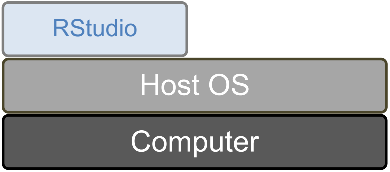
Global environment is empty

Encapsulation



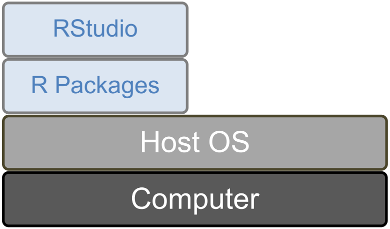
We started with a computer using a specific OS...

Encapsulation



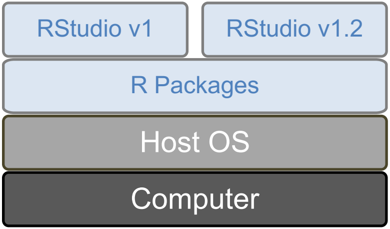
We started with a computer using a specific OS...
And inside this environment, we installed a new application.

Encapsulation



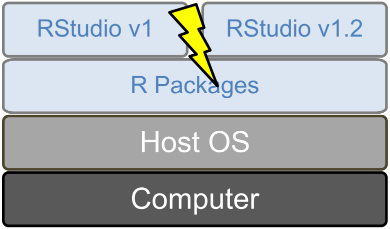
We started with a computer using a specific OS...
And inside this environment, we installed a new application.
Applications rely on dependencies, e.g. external libraries.

Encapsulation



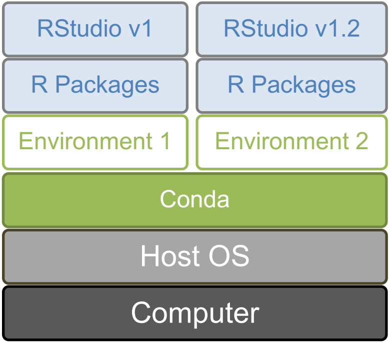
Usually dependencies of different applications don't interfere.
But what if we want to test the latest version of our favourite tool?
There might be conflicts. . .

Encapsulation



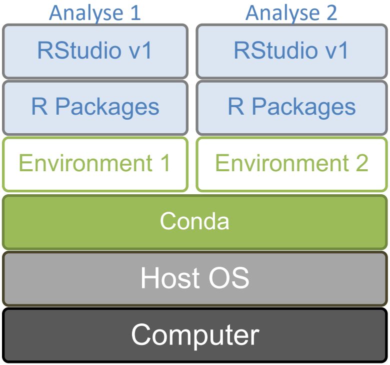
Usually dependencies of different applications don't interfere.
But what if we want to test the latest version of our favourite tool?
There might be conflicts. . .

Encapsulation : managing environments



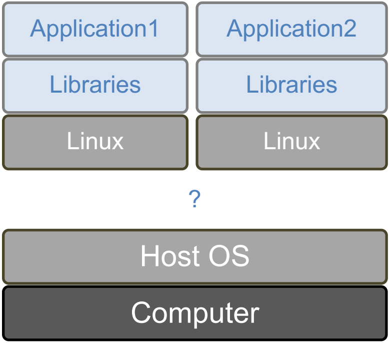
Idea : create separated environments for each application.

Encapsulation : managing environments



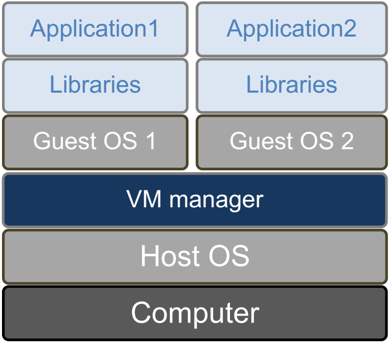
Idea : create separated environments for each application.
More versatile: create a new environment per analysis.

Encapsulation : hardware virtualisation



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

Encapsulation : hardware virtualisation

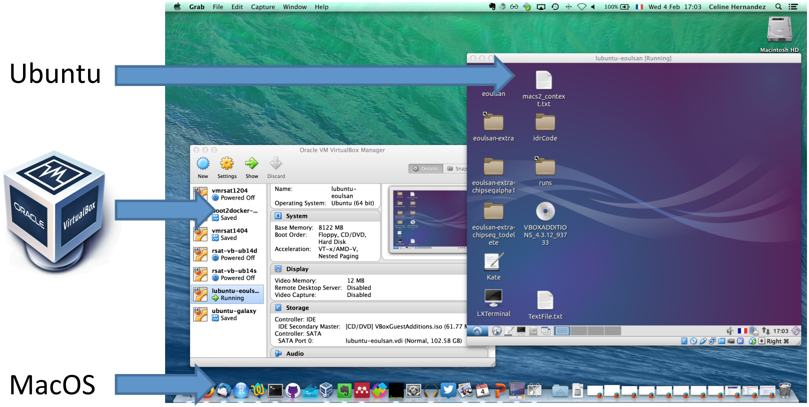


Idea: use virtual machines

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
Oracle VM VirtualBox Manager
New
Settings
Show
Discard
Details
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 ⌘

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

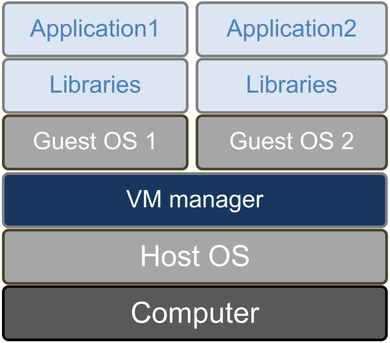
Video Memory: 12 MB
Remote Desktop Server: Disabled
Video Capture: Disabled

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



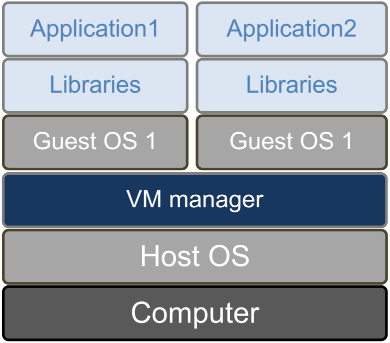
Idea: use virtual machines

Pros: transferable independent environments

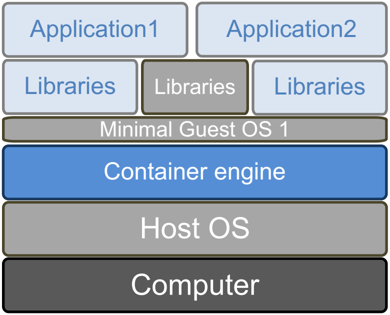
Cons:

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

Encapsulation : OS virtualisation

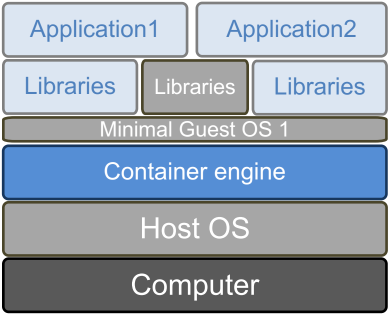


Encapsulation : OS virtualisation



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

Encapsulation : OS virtualisation

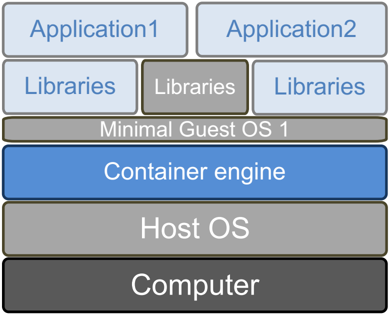


OS virtualisation vs hardware virtualisation

Pros:

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

Encapsulation : OS virtualisation

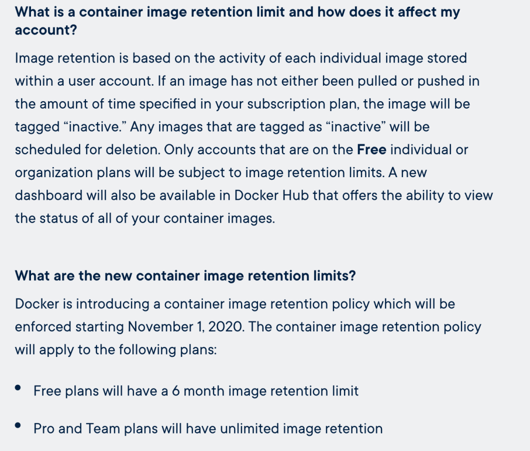


Cons:

- 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.

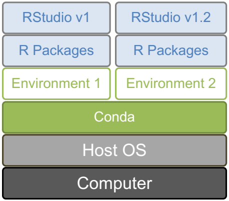
What are the new container image retention limits?

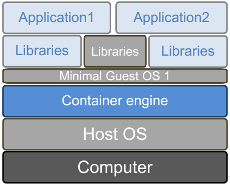
Docker is introducing a container image retention policy which will be enforced starting November 1, 2020. The container image retention policy will apply to the following plans:

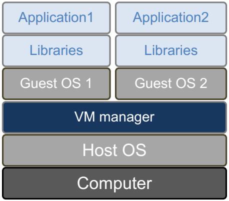
- Free plans will have a 6 month image retention limit
- Pro and Team plans will have unlimited image retention

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

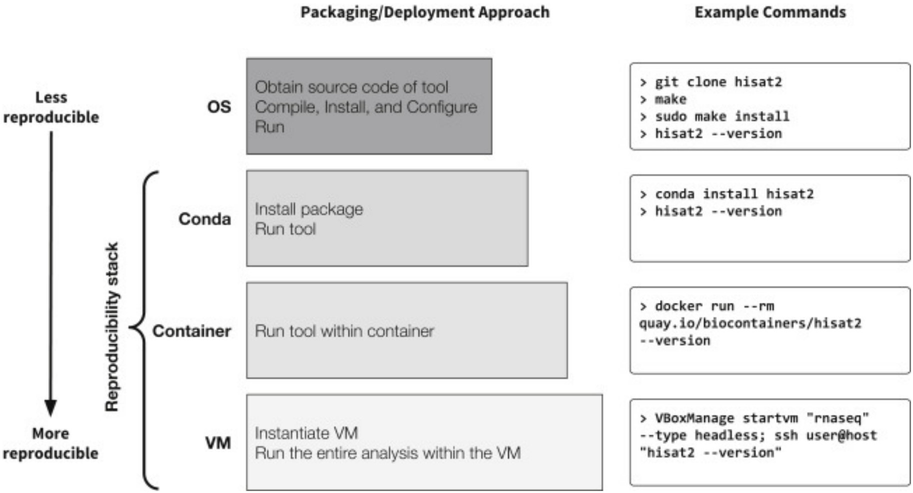
Encapsulation







Encapsulation and reproducibility stack



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

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

Install package
Run tool

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

Run tool within container

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

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)