Reproducibility: an old friend, the laboratory notebook

Better reproducibility with documented code

P. Marin, M. Hiriart, P. Ruiz & N. Goué
aubi@uca.fr

Université Clermont Auvergne, AuBi, Mésocentre

21 mars 2023

. This work is derived from the IFB and I2BC team members

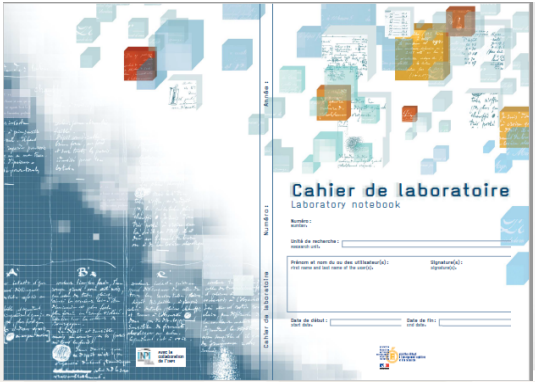
Sommaire

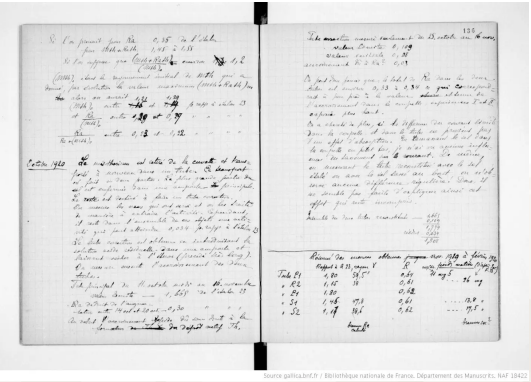
1. The laboratory notebook
   - The aim
2. Notebook in bioinformatic
3. Practicial training
   - Build your own documentation

Paper version

Laboratory notebook allow to :

- Day-to-day recording each step in a process, experiments...





Source gallica.bnf.fr / Bibliothèque nationale de France, Département des Manuscrits, NAF 18422

Paper version

Laboratory notebook allow to :

- Day-to-day recording each step in a process, experiments...
- Report on the progress, and scientific experimentations from the idea to final conclusions

Paper version

Laboratory notebook allow to :

- Day-to-day recording each step in a process, experiments...
- Report on the progress, and scientific experimentations from the idea to final conclusions
- Keep track of knowledge in a lab

Paper version

Laboratory notebook allow to :

- Day-to-day recording each step in a process, experiments...
- Report on the progress, and scientific experimentations from the idea to final conclusions
- Keep track of knowledge in a lab
- Useful drafting a patent

Paper version

Laboratory notebook allow to :

- Day-to-day recording each step in a process, experiments...
- Report on the progress, and scientific experimentations from the idea to final conclusions
- Keep track of knowledge in a lab
- Useful drafting a patent
- Proof of anteriority

Paper version

Paper version

This is a legal tool :

Paper version

This is a legal tool :

- Page numbered in each notebook

Paper version

This is a legal tool :

- Page numbered in each notebook
- Cover page with the owner of the results

Paper version

This is a legal tool :

- Page numbered in each notebook
- Cover page with the owner of the results
- Each page contain a part to date, to sign for at least to people

Paper version

This is a legal tool :

- Page numbered in each notebook
- Cover page with the owner of the results
- Each page contain a part to date, to sign for at least to people

At each research level :

Paper version

This is a legal tool :

- Page numbered in each notebook
- Cover page with the owner of the results
- Each page contain a part to date, to sign for at least to people

At each research level :

- Researchers

Paper version

This is a legal tool :

- Page numbered in each notebook
- Cover page with the owner of the results
- Each page contain a part to date, to sign for at least to people

At each research level :

- Researchers
- Engineers

Paper version

This is a legal tool :

- Page numbered in each notebook
- Cover page with the owner of the results
- Each page contain a part to date, to sign for at least to people

At each research level :

- Researchers
- Engineers
- Technicians

Paper version

This is a legal tool :

- Page numbered in each notebook
- Cover page with the owner of the results
- Each page contain a part to date, to sign for at least to people

At each research level :

- Researchers
- Engineers
- Technicians
- Students...

End what's happen for bioinformatic ?

Electronic version

Electronic Laboratory Notebooks (ELN)
Modern LN since 2009 (C.U.R.I.E. Network)

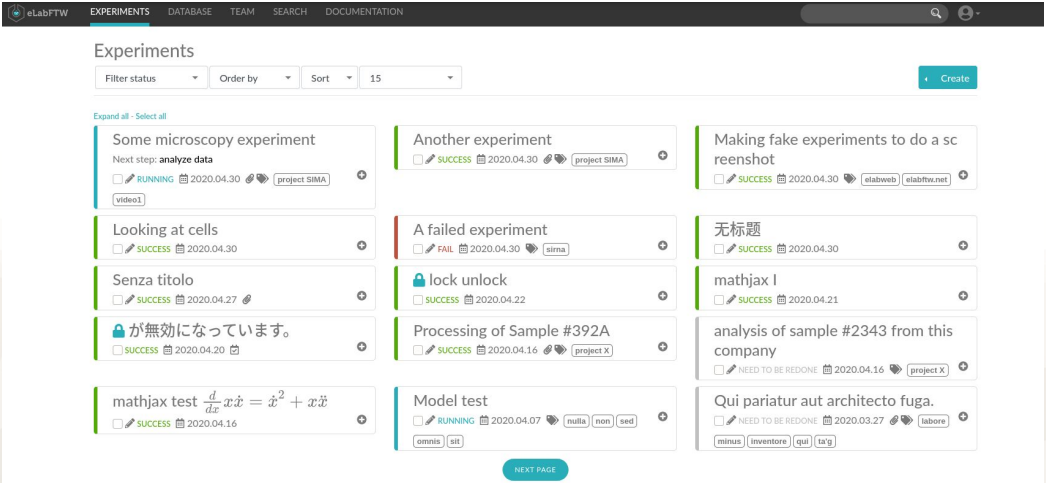
Electronic version

Electronic Laboratory Notebooks (ELN)

Modern LN since 2009 (C.U.R.I.E. Network) 

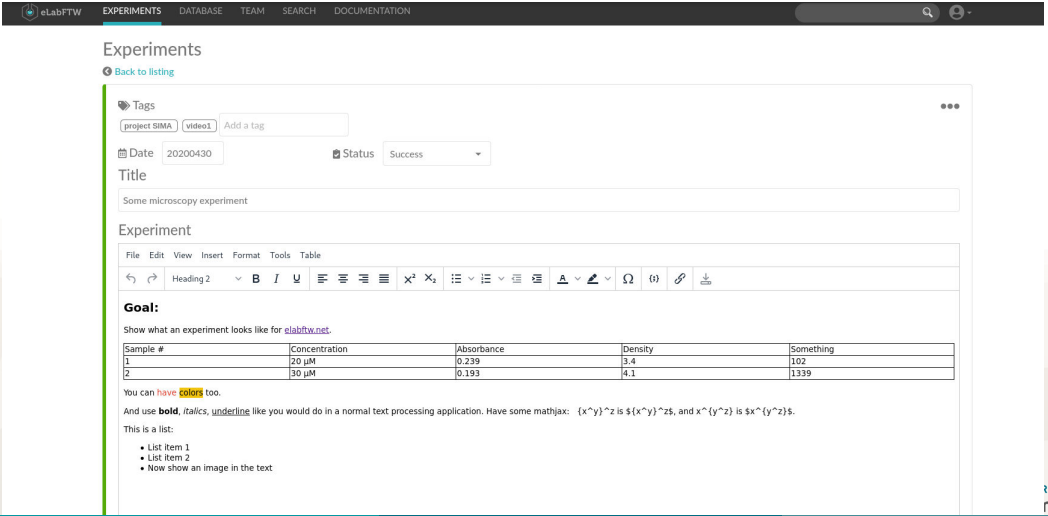
Electronic version

Electronic Laboratory Notebooks (ELN)
Modern LN since 2009 (C.U.R.I.E. Network)



Electronic version

Electronic Laboratory Notebooks (ELN)
Modern LN since 2009 (C.U.R.I.E. Network)



Electronic version

Electronic Laboratory Notebooks (ELN)
Modern LN since 2009 (C.U.R.I.E. Network)

- dematerialised
- archivable
- sharable
- secure

Electronic version

Electronic Laboratory Notebooks (ELN)
Modern LN since 2009 (C.U.R.I.E. Network)

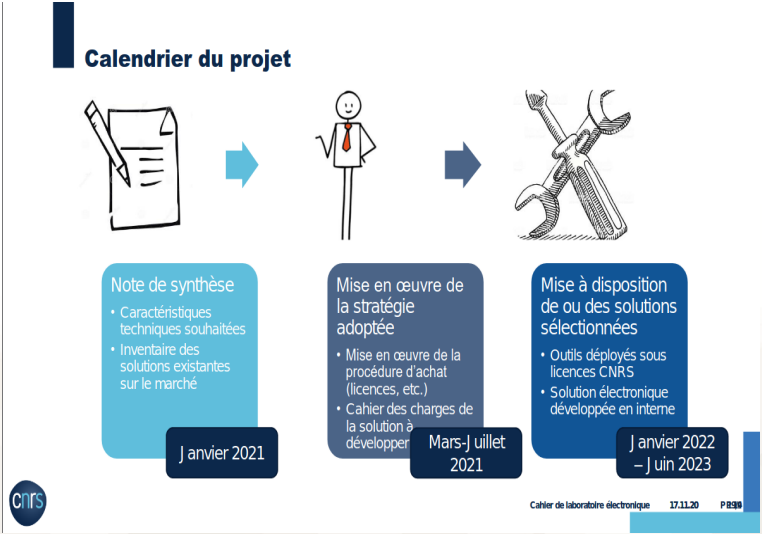
- dematerialised
- archivable
- sharable
- secure

But less and less adapted to recent evolutions of our work
We need an electronic tool for individual traceability

Electronic version

Le rapport du groupe de travail « Cahier de laboratoire électronique » (ELN) présente une vision partagée sur la définition, le cadrage, les usages et le périmètre fonctionnel de l'ELN, qui doit pouvoir s'intégrer dans les environnements informatiques et institutionnels existants. Il émet un ensemble de recommandations sur les critères de choix d'un outil et intègre une liste comparative d'outils existants.

Electronic version



Literate programming

What is literate programming ?

1. Donald E. Knuth, Literate Programming, 1984
. https://en.wikipedia.org/wiki/Literate_programming#Workflow

Literate programming

What is literate programming ?

"Let us change our traditional attitude to the construction of programs :
Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on explaining to humans what we want the computer to do."[1]

1. Donald E. Knuth, Literate Programming, 1984
. `https://en.wikipedia.org/wiki/Literate_programming#Workflow`

Literate programming

What is literate programming ?

"Let us change our traditional attitude to the construction of programs :
Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on explaining to humans what we want the computer to do."[1]

"Literate programming is a programming paradigm introduced by Donald Knuth in which a computer program is given an explanation of its logic in a natural language, such as English, interspersed with snippets of macros and traditional source code, from which compilable source code can be generated."

1. Donald E. Knuth, Literate Programming, 1984
. https://en.wikipedia.org/wiki/Literate_programming#Workflow

Literate programming

What does it look like ?

Literate programming

hello.ipynb | Untitled1.ipynb | hello.ipynb | Untitled.ipynb

Markdown | My Python 3.8

```
# This is an example title
I can add more infos about my code to explain to other users and aslo to remember what my
code is doing

> - Cell with text to give informations
> - Cell with code to view what's happen
> - plot are also present in the output
> - You can export the raw file in jupyter format but also in pdf, word, tex...
```

Subtitle for configuration import

[1]:
```
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
```

Subtitle for my first plot

[2]:
```
def gaussian(x, a, b, c):
    return a * np.exp(-b * (x-c)**2)

def noisy_gaussian():
    # gaussian array y in interval -5 <= x <= 5
    nx = 100
    x = np.linspace(-5.0, 5.0, nx)
    y = gaussian(x, a=2.0, b=0.5, c=1.5)
    noise = np.random.normal(0.0, 0.2, nx)
    y += noise
    return x, y

def fit(x, y, n):
    pfit = np.polyfit(x, y, n)
    yfit = np.polyval(pfit, x)
    return yfit

def plot(x, y, yfit):
    plt.plot(x, y, "r", label="Data")
```

This is an example title

I can add more infos about my code to explain to other users and aslo to rememb
is doing

- Cell with text to give informations
- Cell with code to view what's happen
- plot are also present in the output
- You can export the raw file in jupyter format but also in pdf, word, tex

Subtitle for configuration import

In [1]:
```
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
```

Subtitle for my first plot

In [2]:
```
def gaussian(x, a, b, c):
    return a * np.exp(-b * (x-c)**2)

def noisy_gaussian():
    # gaussian array y in interval -5 <= x <= 5
    nx = 100
    x = np.linspace(-5.0, 5.0, nx)
    y = gaussian(x, a=2.0, b=0.5, c=1.5)
    noise = np.random.normal(0.0, 0.2, nx)
    y += noise
    return x, y

def fit(x, y, n):
    pfit = np.polyfit(x, y, n)
    yfit = np.polyval(pfit, x)
    return yfit
```

Literate programming

What does it look like ?
Interactive programming interface allowing to combine both natural and computer languages

Literate programming

What does it look like ?
Interactive programming interface allowing to combine both natural and computer languages
In one file

- Explanation
- Code
- Results
- Graphs and plots

Literate programming

Why using literate programming frameworks ?

- Labbook

Literate programming

Why using literate programming frameworks ?

- Labbook
- Day-to-day analysis

Literate programming

Why using literate programming frameworks ?

- Labbook
- Day-to-day analysis
- Make automatic reports

Literate programming

Why using literate programming frameworks ?

- Labbook
- Day-to-day analysis
- Make automatic reports
- Write scientific article

Literate programming

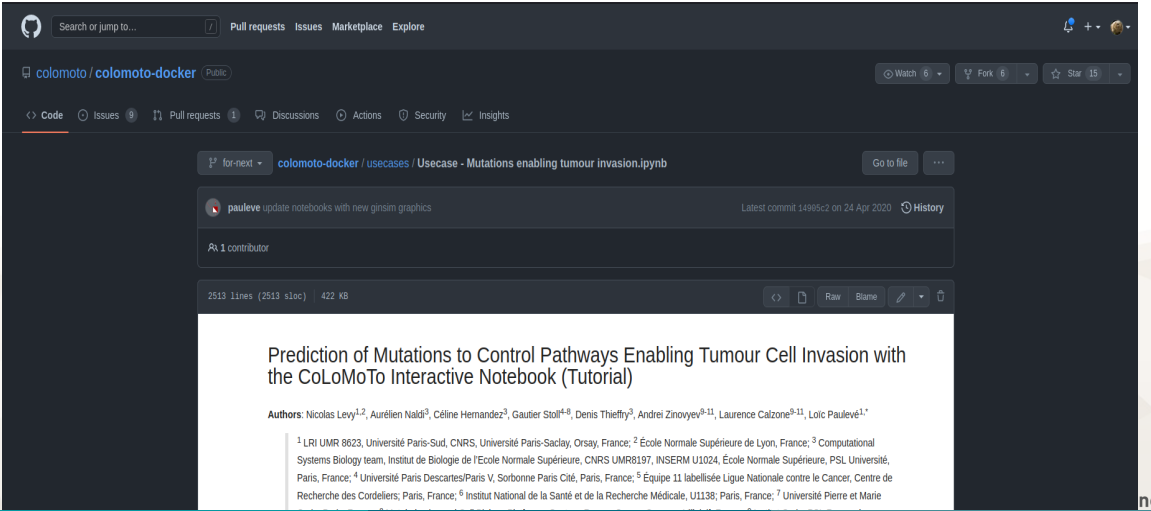
example

Example of an article written using a notebook

. https://www.frontiersin.org/articles/10.3389/fphys.2018.00787/full

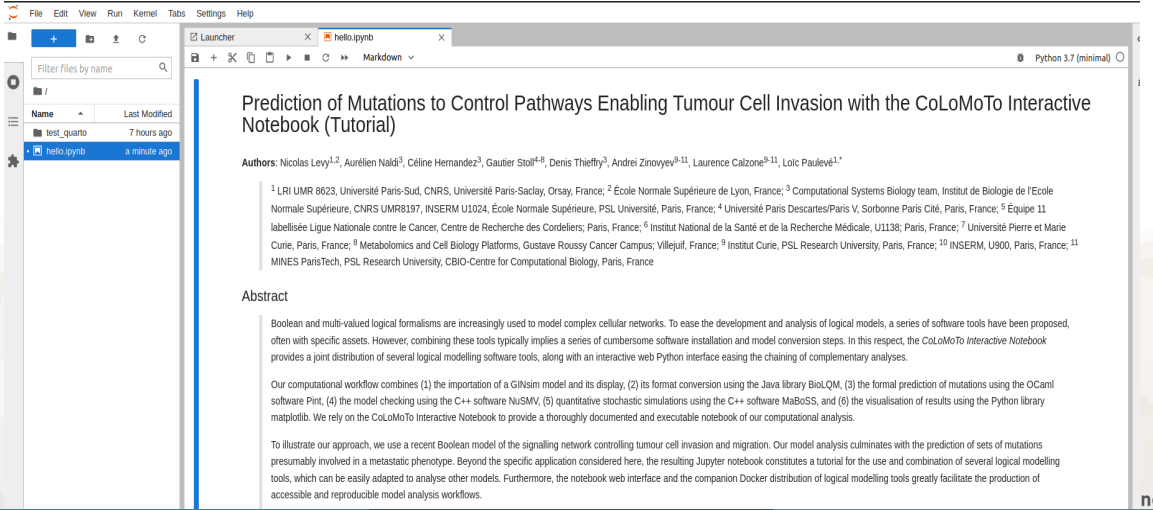
Literate programming

example



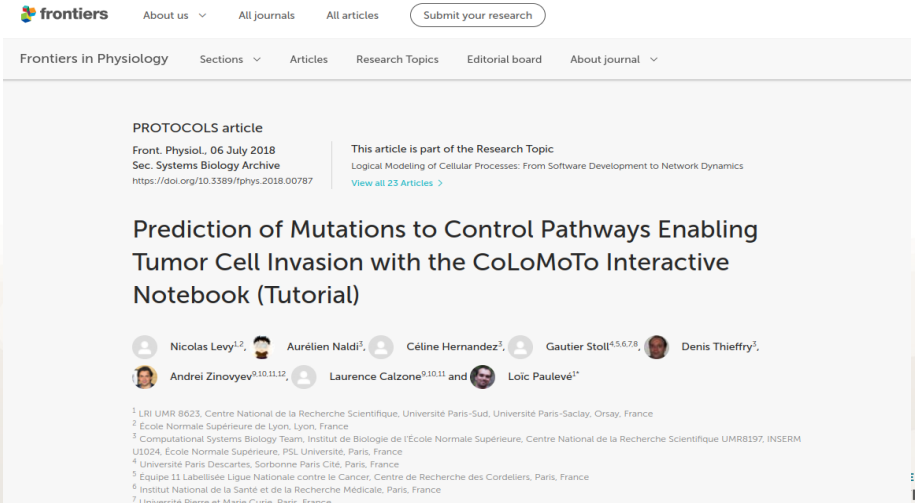
Literate programming

example



Literate programming

example



frontiers About us All journals All articles Submit your research

Frontiers in Physiology Sections Articles Research Topics Editorial board About journal

PROTOCOLS article

Front. Physiol., 06 July 2018
Sec. Systems Biology Archive
https://doi.org/10.3389/fphys.2018.00787

This article is part of the Research Topic
Logical Modeling of Cellular Processes: From Software Development to Network Dynamics
View all 23 Articles >

Prediction of Mutations to Control Pathways Enabling Tumor Cell Invasion with the CoLoMoTo Interactive Notebook (Tutorial)

Nicolas Levy[1,2], Aurélien Naldi[3], Céline Hernandez[3], Gautier Stoll[4,5,6,7,8], Denis Thieffry[3], Andrei Zinovyev[9,10,11,12], Laurence Calzone[9,10,11] and Loïc Paulevé[1]*

[1] LRI UMR 8623, Centre National de la Recherche Scientifique, Université Paris-Sud, Université Paris-Saclay, Orsay, France
[2] École Normale Supérieure de Lyon, Lyon, France
[3] Computational Systems Biology Team, Institut de Biologie de l'École Normale Supérieure, Centre National de la Recherche Scientifique UMR8197, INSERM U1024, École Normale Supérieure, PSL Université, Paris, France
[4] Université Paris Descartes, Sorbonne Paris Cité, Paris, France
[5] Équipe 11 Labellisée Ligue Nationale contre le Cancer, Centre de Recherche des Cordeliers, Paris, France
[6] Institut National de la Santé et de la Recherche Médicale, Paris, France

Markup

A markup language uses tags to define elements within a document.
Three different types and usage :

Markup

A markup language uses tags to define elements within a document.
Three different types and usage :

- Presentational (used by traditional word-processing systems)

Markup

A markup language uses tags to define elements within a document.
Three different types and usage :

- Presentational (used by traditional word-processing systems)
- Procedural, provides instructions to process the text (e.g. TeX, PostScript)

Markup

A markup language uses tags to define elements within a document.
Three different types and usage :

- Presentational (used by traditional word-processing systems)
- Procedural, provides instructions to process the text (e.g. TeX, PostScript)
- Descriptive, to label documents parts (e.g. LaTeX, HTML, XML...)

Markdown

Markdown is a Lightweight markup language
Designed to be :

Markdown

Markdown is a Lightweight markup language
Designed to be :

- easy to write using any generic text editor (plain-text-formatting syntax)

Markdown

Markdown is a Lightweight markup language
Designed to be :

- easy to write using any generic text editor (plain-text-formatting syntax)
- easy to read in the raw format

Markdown

Markdown is a Lightweight markup language
Designed to be :

- easy to write using any generic text editor (plain-text-formatting syntax)
- easy to read in the raw format

```
# Heading level 1

## Heading level 2

### Heading level 3

**Bold text**

*Italic*

~~strikethrough text~~

A [link](http://www.example.com)

Item list

    - one point
        - a subpoint
    - another point
    - last point
```

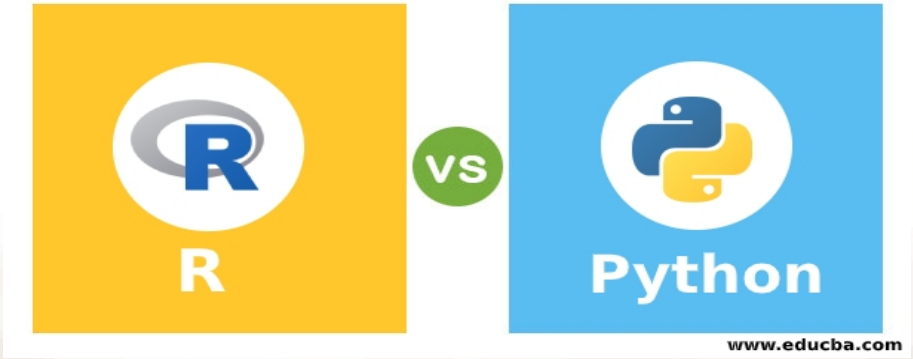
Markdown

Markdown is a Lightweight markup language
Designed to be :

- easy to write using any generic text editor (plain-text-formatting syntax)
- easy to read in the raw format

Used on Github to make the README.md
But how is this useful for literate programming ?
When you want to weave both code (to be interpreted) and formatting information, you precisely need a lightweight language for the formatting part.

R notebooks *vs* Jupyter(Lab) notebook



R notebook

1. Sweave in 2002
   Leisch, Friedrich (2002). "Sweave, Part I : Mixing R and LaTeX : A short introduction to the Sweave file format and corresponding R functions"

R notebook

1. Sweave in 2002
   Leisch, Friedrich (2002). "Sweave, Part I : Mixing R and LaTeX : A short introduction to the Sweave file format and corresponding R functions"
2. knitR in 2011
   "The knitr package was designed to be a transparent engine for dynamic report generation with R, solve some long-standing problems in Sweave, and combine features in other add-on packages into one package"

RMarkdown

2012 Rmarkdown was born !

. https ://rmarkdown.rstudio.com

RMarkdown

2012 Rmarkdown was born !

. https ://rmarkdown.rstudio.com

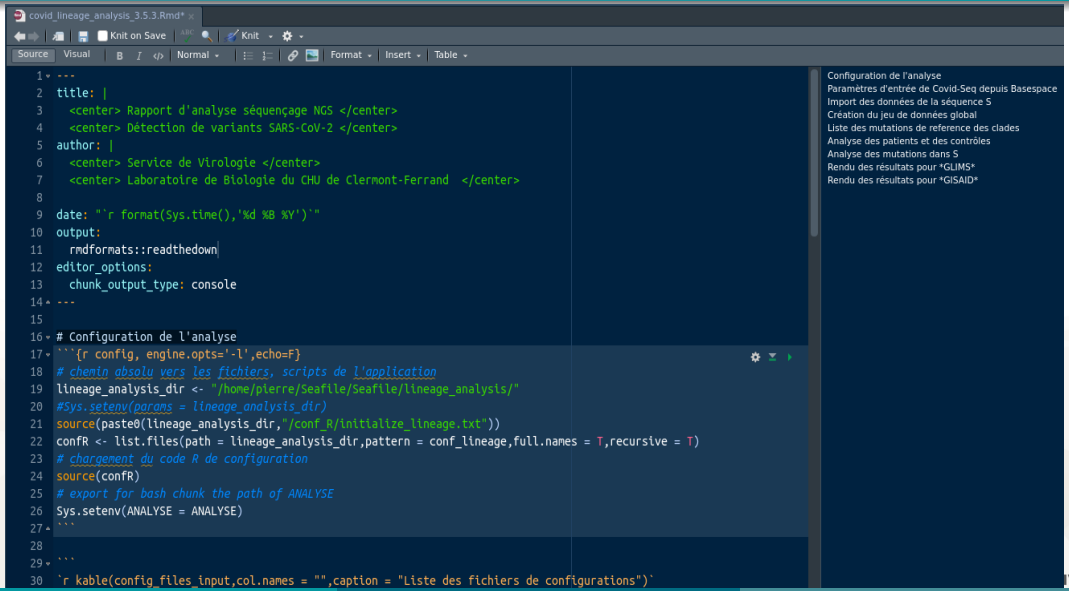
RMarkdown

2012 Rmarkdown was born !

"When you run render, R Markdown feeds the .Rmd file to knitr, which executes all of the code chunks and creates a new markdown (.md) document which includes the code and its output. The markdown file generated by knitR is then processed by pandoc which is responsible for creating the finished format."

. https ://rmarkdown.rstudio.com

RMarkdown


covid_lineage_analysis_3.5.3.Rmd*
Knit on Save
Knit
Source
Visual
Normal
Format
Insert
Table
1 ---
2 title: |
3 Rapport d'analyse séquençage NGS
4 Détection de variants SARS-CoV-2
5 author: |
6 Service de Virologie
7 Laboratoire de Biologie du CHU de Clermont-Ferrand
8
9 date: "`r format(Sys.time(),'%d %B %Y')`"
10 output:
11 rmdformats::readthedown
12 editor_options:
13 chunk_output_type: console
14 ---
15
16 # Configuration de l'analyse
17 ```{r config, engine.opts='-l',echo=F}
18 # chemin absolu vers les fichiers, scripts de l'application
19 lineage_analysis_dir <- "/home/pierre/Seafile/Seafile/lineage_analysis/"
20 #Sys.setenv(params = lineage_analysis_dir)
21 source(paste0(lineage_analysis_dir,"/conf_R/initialize_lineage.txt"))
22 confR <- list.files(path = lineage_analysis_dir,pattern = conf_lineage,full.names = T,recursive = T)
23 # chargement du code R de configuration
24 source(confR)
25 # export for bash chunk the path of ANALYSE
26 Sys.setenv(ANALYSE = ANALYSE)
27 ```
28
29 ```
30 `r kable(config_files_input,col.names = "",caption = "Liste des fichiers de configurations")`
Configuration de l'analyse
Paramètres d'entrée de Covid-Seq depuis Basespace
Import des données de la séquence S
Création du jeu de données global
Liste des mutations de reference des clades
Analyse des patients et des contrôles
Analyse des mutations dans S
Rendu des résultats pour *GLIMS*
Rendu des résultats pour *GISAID*

RMarkdown

Rapport d'analyse séquençage NGS
Détection de variants SARS-CoV-2

- Configuration de l'analyse
- **Paramètres d'entrée de Covid-Seq depuis Basespace**
- Import des données de la séquence S
- Création du jeu de données global
- Liste des mutations de reference des clades
- Analyse des patients et des contrôles
- Analyse des mutations dans S
- Rendu des résultats pour *GLIMS*
- Rendu des résultats pour *GISAID*

Service de Virologie
Laboratoire de Biologie du CHU de Clermont-Ferrand

Import des données de la séquence S

```
Nombre de séquences trouvées dans le fichier fasta depuis basespace : 93
Nombre de séquences trouvées dans le fichier fasta aligné via Nextclade : 93
```

Création du jeu de données global

```
Le fichier de fusion des données contient 96 entrées
```

Liste des mutations de reference des clades

Copy | CSV | Excel | PDF | Print | Show 10 entries | Search:

Tableau des mutations de références utilisées pour l'analyse des variants

scorpio_clade	lineage.ref	synonymous_mutation_ref
All	A	All
20A.EU2	B.1.160	C4543T,G5629T,C11497T,T26876C
20A/S:126A	B.1.620	C241T,C3037T,C15324T,T20049C,A23416T,T27795-,T27796-,A27797-,A28271T
20A/S:439K	B.1.258	C8047T
20A/S:98F	B.1.221	C28651T
20B/S:1122L	B.1.1.302	
20B/S:626S	B.1.1.277	
20B/S:732A	B.1.1.519	C203T,C222T,C241T,C3037T,C10954T,T19839C,C21306T,C29197T
20C/S:80Y	B.1.367	G4960T,C6070T,C7303T,C7564T,C10279T,C10525T,C10582T,C27804T

Jupyter

1. 2011 : IPython (interactive Python shell) with notebook functionalities

. https ://jupyter.org/

Jupyter

1. 2011 : IPython (interactive Python shell) with notebook functionalities
2. 2014 : Spin-off project called Project Jupyter a non-profit, open-source project maintained by a strong community

. https ://jupyter.org/

Jupyter

1. 2011 : IPython (interactive Python shell) with notebook functionalities
2. 2014 : Spin-off project called Project Jupyter a non-profit, open-source project maintained by a strong community
   - "Jupyter will always be 100% open-source software, free for all to use and released under the liberal terms of the modified BSD license"

. https ://jupyter.org/

Jupyter

1. 2011 : IPython (interactive Python shell) with notebook functionalities
2. 2014 : Spin-off project called Project Jupyter a non-profit, open-source project maintained by a strong community
   - "Jupyter will always be 100% open-source software, free for all to use and released under the liberal terms of the modified BSD license"
   - A reference to the three core programming languages supported by Jupyter (Julia, Python and R)

. https ://jupyter.org/

Jupyter

What is it exactly ?

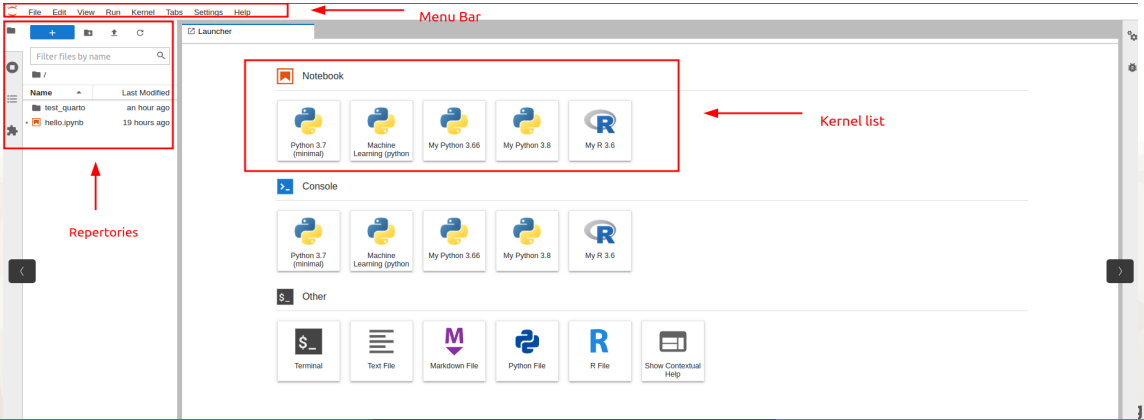
Web-based interactive computational environment

Jupyter

What is it exactly ?

Web-based interactive computational environment

- Web-based : client/server

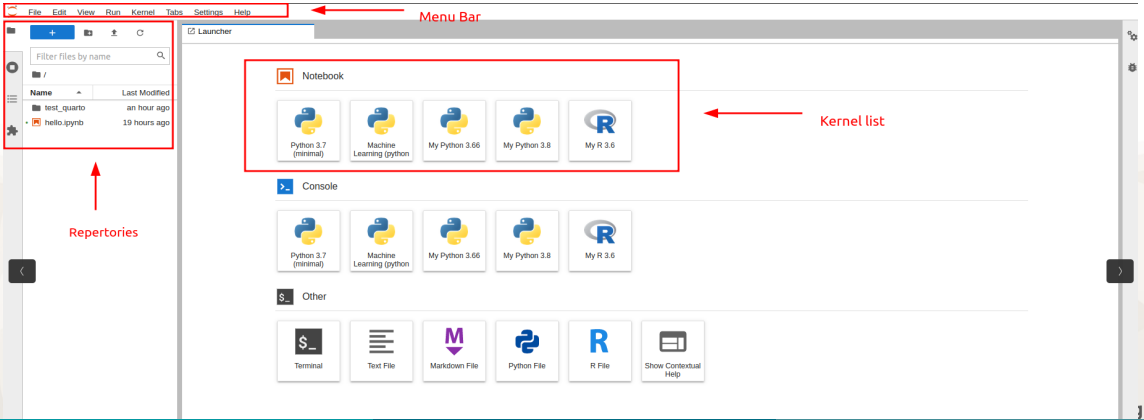


Jupyter

What is it exactly ?

Web-based interactive computational environment

- Web-based : client/server
- Interactive : notebook system

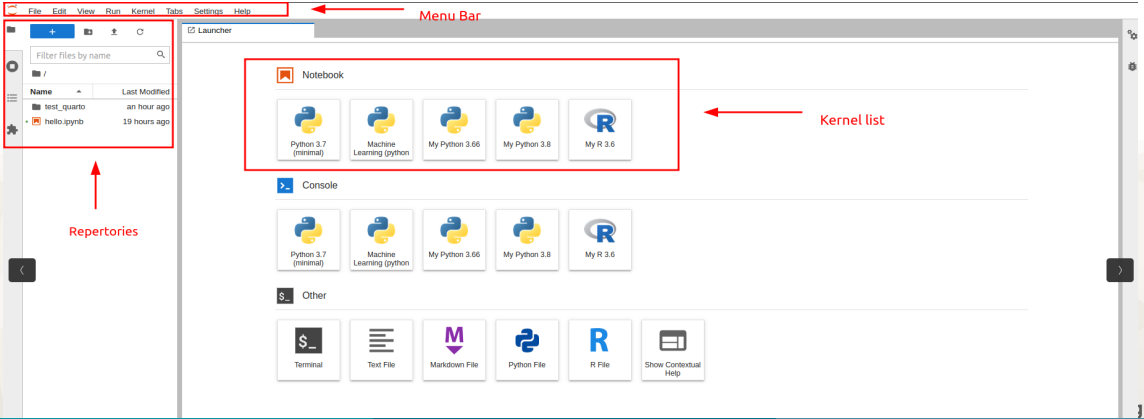


Jupyter

What is it exactly ?

Web-based interactive computational environment

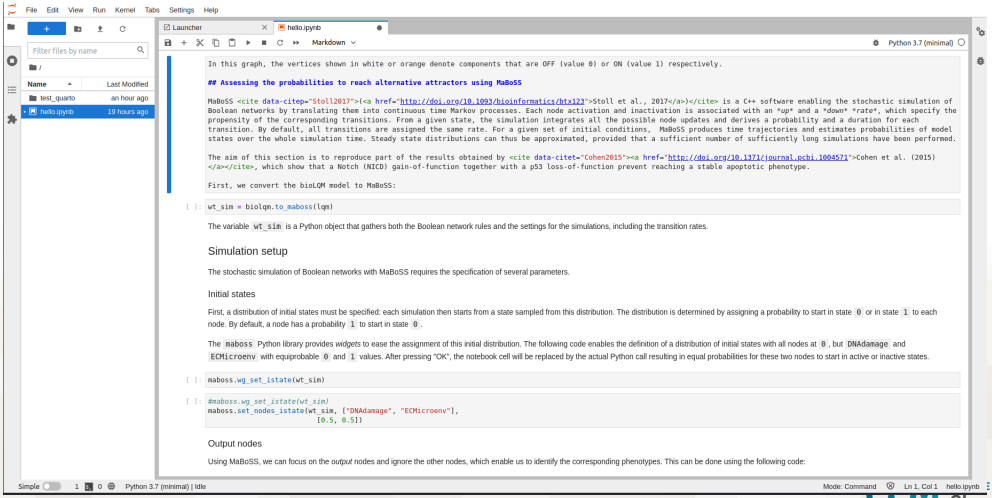
- Web-based : client/server
- Interactive : notebook system
- Computational environment : console, many kernels available...



Jupyter

What is it exactly ?

Web-based interactive computational environment



PRACTICE

- Jupyter training link
- Rmarkdown training link