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

Literate programming

Introduction

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.
— Donald E. Knuth, Literate Programming, 1984

Introduction

What is literate programming ?

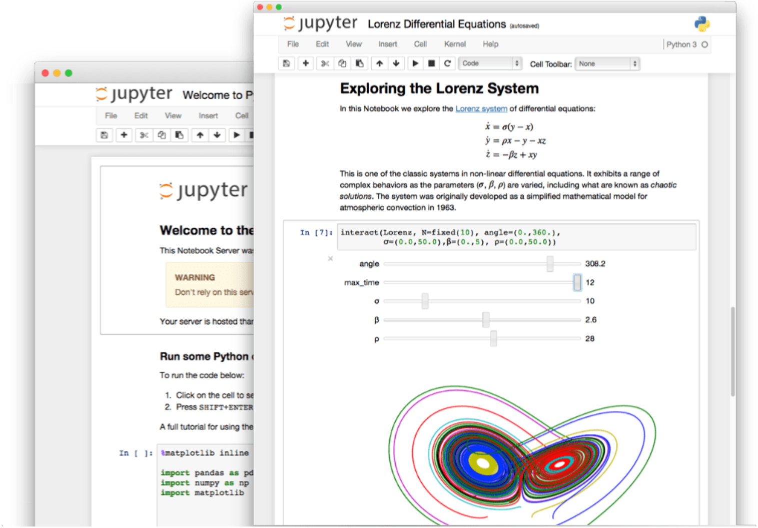
Definition

"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." Donald Knuth, 1984.

Wikipedia, 18/08/2020
`https://en.wikipedia.org/wiki/Literate_programming#Workflow`

Introduction

What does it look like ?



jupyter

Welcome to the

This Notebook Server wa

WARNING

Don't rely on this serv

Your server is hosted tha

Run some Python

To run the code below:

1. Click on the cell to se
2. Press SHIFT+ENTER

A full tutorial for using the

```
In [ ]: %matplotlib inline

        import pandas as pd
        import numpy as np
        import matplotlib
```

Exploring the Lorenz System

In this Notebook we explore the Lorenz system of differential equations:

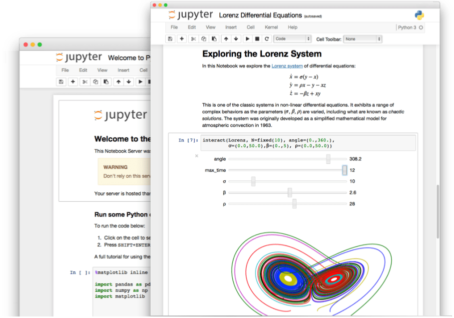
$$\dot{x} = \sigma(y - x)$$
$$\dot{y} = \rho x - y - xz$$
$$\dot{z} = -\beta z + xy$$

This is one of the classic systems in non-linear differential equations. It exhibits a range of complex behaviors as the parameters (σ, β, ρ) are varied, including what are known as *chaotic solutions*. The system was originally developed as a simplified mathematical model for atmospheric convection in 1963.

```
In [7]: interact(Lorenz, N=fixed(10), angle=(0.,360.),
                 σ=(0.0,50.0),β=(0.,5), ρ=(0.0,50.0))
```

angle 308.2
max_time 12
σ 10
β 2.6
ρ 28

Introduction



Welcome to the

This Notebook Server wa

Don't rely on this ser

Your server is hosted tha

Run some Python

To run the code below:

1. Click on the cell to se
2. Press SHIFT+ENTER

A full tutorial for using the

```
In [ ]: %matplotlib inline

        import pandas as pd
        import numpy as np
        import matplotlib
```

Exploring the Lorenz System

In this Notebook we explore the Lorenz system of differential equations:

$$\dot{x} = \sigma(y - x)$$
$$\dot{y} = \rho x - y - xz$$
$$\dot{z} = -\beta z + xy$$

This is one of the classic systems in non-linear differential equations. It exhibits a range of complex behaviors as the parameters (σ, β, ρ) are varied, including what are known as *chaotic solutions*. The system was originally developed as a simplified mathematical model for atmospheric convection in 1963.

```
In [7]: interact(Lorenz, N=fixed(10), angle=(0.,360.),
                 σ=(0.0,50.0),β=(0.,5), ρ=(0.0,50.0))
```

Interactive programming interface allowing to combine both natural and computer languages.

In one file:

- Explanations
- Code
- Results
- Graphs and plots

Introduction

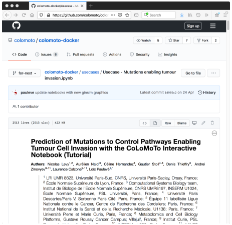
Why using literate programming frameworks ?

Use cases:

- Day to day analyses
- Analysis reports
- Writing scientific articles

Example of an article entirely written using a notebook

File (on a repository)



colomoto / colomoto-docker

colomoto-docker / usecases / Usecase - Mutations enabling tumour invasion.ipynb

pauleve update notebooks with new ginsim graphics Latest commit 14985c2 on 24 Apr

1 contributor

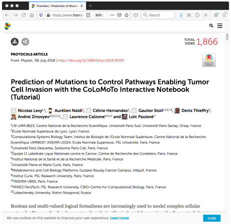
2513 lines (2513 sloc) | 422 KB

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

Authors: Nicolas Levy[1,2], Aurélien Naldi[3], Céline Hernandez[3], Gautier Stoll[4-8], Denis Thieffry[3], Andrei Zinovyev[9-11], Laurence Calzone[9-11], Loïc Paulevé[1,*]

> [1] LRI UMR 8623, Université Paris-Sud, CNRS, Université Paris-Saclay, Orsay, France; [2] École Normale Supérieure de Lyon, France; [3] Computational Systems Biology team, Institut de Biologie de l'Ecole Normale Supérieure, CNRS UMR8197, INSERM U1024, École Normale Supérieure, PSL Université, Paris, France; [4] Université Paris Descartes/Paris V, 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, U1138; Paris, France; [7] Université Pierre et Marie Curie, Paris, France; [8] Metabolomics and Cell Biology Platforms, Gustave Roussy Cancer Campus; Villejuif, France; [9] Institut Curie, PSL

Published article



TOTAL VIEWS 1,866

PROTOCOLS ARTICLE

Front. Physiol., 06 July 2018 | https://doi.org/10.3389/fphys.2018.00787

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

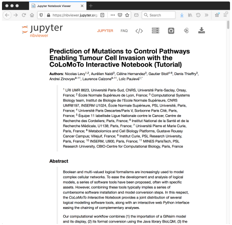
Nicolas Levy[1,2], **Aurélien Naldi**[3], **Céline Hernandez**[3], **Gautier Stoll**[4,5,6,7,8], **Denis Thieffry**[3], **Andrei Zinovyev**[9,10,11,12], **Laurence Calzone**[9,10,11] and **Loïc Paulevé**[1,*]

[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
[7] Université Pierre et Marie Curie, Paris, France
[8] Metabolomics and Cell Biology Platforms, Gustave Roussy Cancer Campus, Villejuif, France
[9] Institut Curie, PSL Research University, Paris, France
[10] INSERM U900, Paris, France
[11] MINES ParisTech, PSL Research University, CBIO-Centre for Computational Biology, Paris, France
[12] Lobachevsky University, Nizhni Novgorod, Russia

Boolean and multi-valued logical formalisms are increasingly used to model complex cellular

We use cookies on this website to improve your user experience. Learn more

Executable file



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

Authors: Nicolas Levy[1,2], Aurélien Naldi[3], Céline Hernandez[3], Gautier Stoll[4-8], Denis Thieffry[3], Andrei Zinovyev[9-11], Laurence Calzone[9-11], Loïc Paulevé[1,*]

> [1] LRI UMR 8623, Université Paris-Sud, CNRS, Université Paris-Saclay, Orsay, France; [2] École Normale Supérieure de Lyon, France; [3] Computational Systems Biology team, Institut de Biologie de l'Ecole Normale Supérieure, CNRS UMR8197, INSERM U1024, École Normale Supérieure, PSL Université, Paris, France; [4] Université Paris Descartes/Paris V, 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, U1138; Paris, France; [7] Université Pierre et Marie Curie, Paris, France; [8] Metabolomics and Cell Biology Platforms, Gustave Roussy Cancer Campus; Villejuif, France; [9] Institut Curie, PSL Research University, Paris, France; [10] INSERM, U900, Paris, France; [11] MINES ParisTech, PSL Research University, CBIO-Centre for Computational Biology, Paris, France

Abstract

> Boolean and multi-valued logical formalisms are increasingly used to model complex cellular networks. To ease the development and analysis of logical models, a series of software tools have been proposed, often with specific assets. However, combining these tools typically implies a series of cumbersome software installation and model conversion steps. In this respect, the *CoLoMoTo Interactive Notebook* provides a joint distribution of several logical modelling software tools, along with an interactive web Python interface easing the chaining of complementary analyses.
>
> Our computational workflow combines (1) the importation of a GINsim model and its display, (2) its format conversion using the Java library BioLQM, (3) the

Literate programming

This session :

- Markdown
- Rmarkdown / RStudio
- Jupyter

Markup and markdown

Definition

A markup language uses tags to define elements within a document.

Three different types and usage :

- Presentational (used by traditional word-processing systems)
  - Markup is invisible
- Procedural, provides instructions to process the text (e.g. TeX, PostScript)
  - Markup is visible and can be directly manipulated by the author.
- Descriptive, to label documents parts (e.g. LaTeX, HTML, XML...)
  - Emphasizes the document structure.

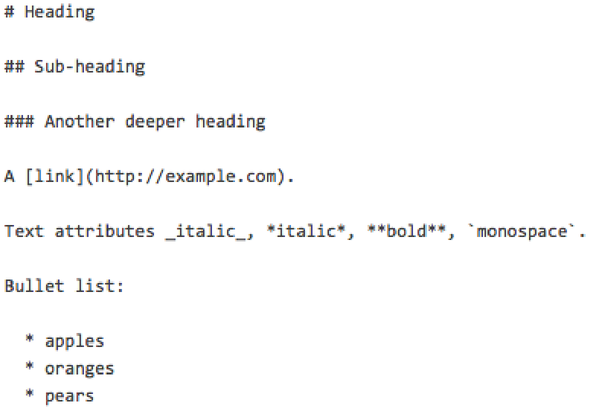
Markdown language

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 its raw form

Markdown language

You've probably see it already on GitHub (README), Wikipedia...



```
# Heading

## Sub-heading

### Another deeper heading

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

Text attributes _italic_, *italic*, **bold**, `monospace`.

Bullet list:

  * apples
  * oranges
  * pears
```

Github guide :
urlhttps://guides.github.com/features/mastering-markdown/

Literate programming

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.

The challengers

No need to hide, there are currently two main frameworks used in bioinformatics:
RMarkdown and Jupyter

RMarkdown

RMarkdown

RMarkdown

At the beginning, there was nothing.

RMarkdown

At the beginning, there was nothing.

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

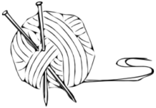
RMarkdown

At the beginning, there was nothing.

Then came Sweave.
Leisch, Friedrich (2002). "Sweave, Part I: Mixing R and LaTeX: A short introduction to the Sweave file format and corresponding R functions"
And people saw that the path would be long...

RMarkdown

knitR (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"
`https://yihui.org/knitr/`

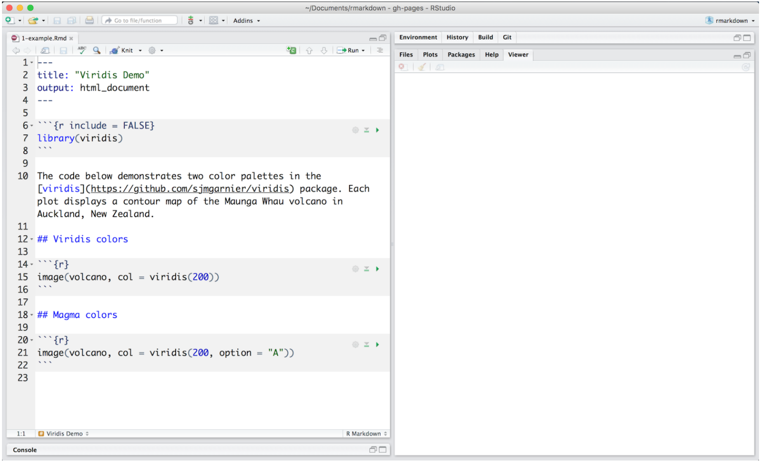
RMarkdown

RMarkdown

"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

RMarkdown

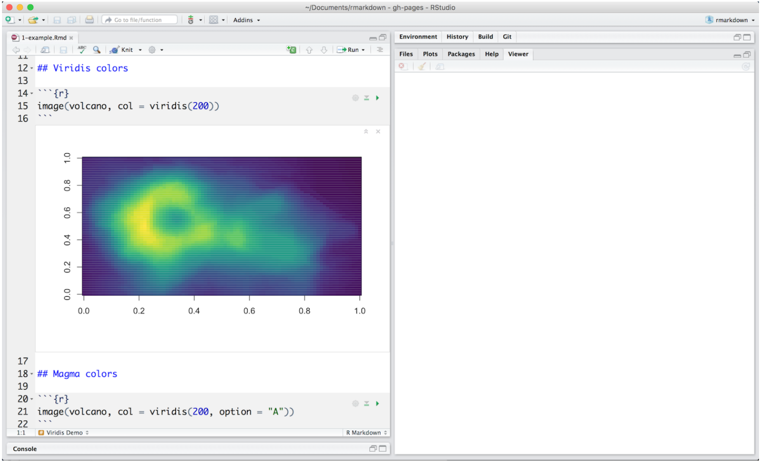


~~~
---
title: "Viridis Demo"
output: html_document
---

```{r include = FALSE}
library(viridis)
```

The code below demonstrates two color palettes in the
[viridis](https://github.com/sjmgarnier/viridis) package. Each
plot displays a contour map of the Maunga Whau volcano in
Auckland, New Zealand.

## Viridis colors

```{r}
image(volcano, col = viridis(200))
```

## Magma colors

```{r}
image(volcano, col = viridis(200, option = "A"))
```
~~~

Integrated into RStudio, IDE for R.

RMarkdown

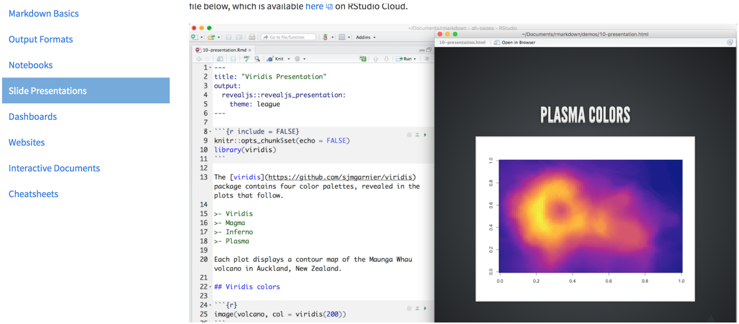
R Notebooks



~~~
## Viridis colors

```{r}
image(volcano, col = viridis(200))
```

## Magma colors

```{r}
image(volcano, col = viridis(200, option = "A"))
```
~~~

RMarkdown

R Notebooks and more...



file below, which is available here on RStudio Cloud.

~~~
---
title: "Viridis Presentation"
output:
  revealjs::revealjs_presentation:
    theme: league
---

```{r include = FALSE}
knitr::opts_chunk$set(echo = FALSE)
library(viridis)
```


The [viridis](https://github.com/sjmgarnier/viridis) package contains four color palettes, revealed in the plots that follow.

>- Viridis
>- Magma
>- Inferno
>- Plasma

Each plot displays a contour map of the Maunga Whau volcano in Auckland, New Zealand.

## Viridis colors

```{r}
image(volcano, col = viridis(200))
~~~

Jupyter

Jupyter

Jupyter

A bit of history...

- 2011 : IPython (interactive Python shell) with notebook functionalities
- 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 can it do?

Jupyter

What can it do?
Everything (excepted coffee)

Jupyter

But what is it exactly ?

Jupyter

But what is it exactly ?
Web-based interactive computational environment.

Jupyter

But what is it exactly ?
Web-based interactive computational environment.

- Web-based : client/server

Jupyter

But what is it exactly ?
Web-based interactive computational environment.

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

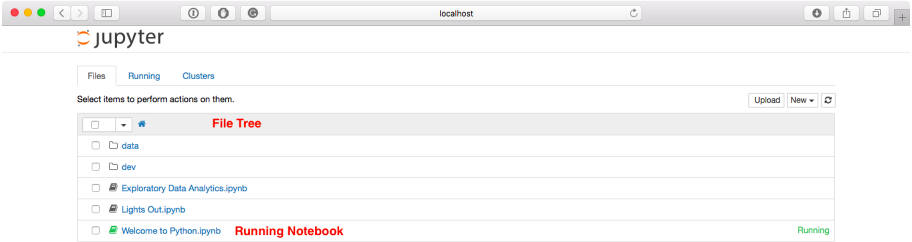
Jupyter

But what is it exactly ?
Web-based interactive computational environment.

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

Jupyter

Dashboard



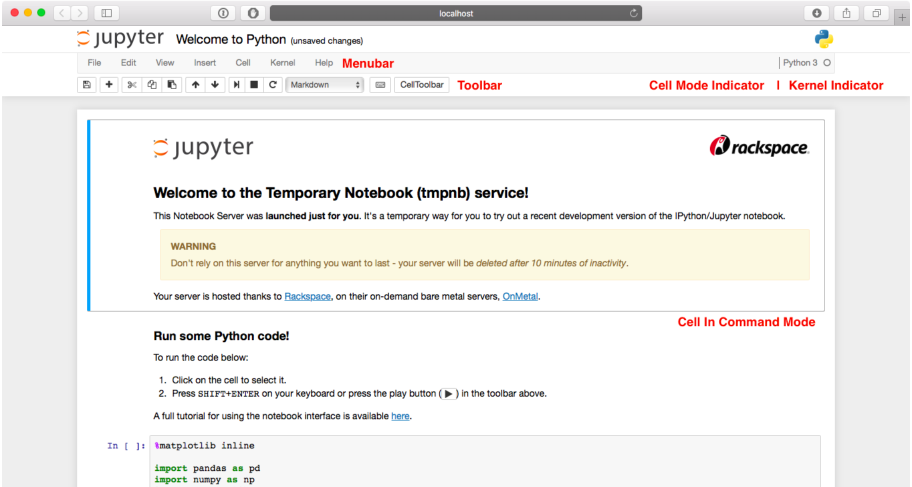
Select items to perform actions on them.

File Tree

- ☐ data
- ☐ dev
- ☐ Exploratory Data Analytics.ipynb
- ☐ Lights Out.ipynb
- ☐ Welcome to Python.ipynb **Running Notebook** Running

Jupyter

Notebook editor



jupyter

Welcome to the Temporary Notebook (tmpnb) service!

This Notebook Server was **launched just for you**. It's a temporary way for you to try out a recent development version of the IPython/Jupyter notebook.

WARNING

Don't rely on this server for anything you want to last - your server will be *deleted after 10 minutes of inactivity*.

Your server is hosted thanks to Rackspace, on their on-demand bare metal servers, OnMetal.

Run some Python code!

To run the code below:

1. Click on the cell to select it.
2. Press `SHIFT+ENTER` on your keyboard or press the play button (▶) in the toolbar above.

A full tutorial for using the notebook interface is available here.

```
In [ ]: %matplotlib inline

        import pandas as pd
        import numpy as np
        import matplotlib
```

Jupyter

Project Jupyter

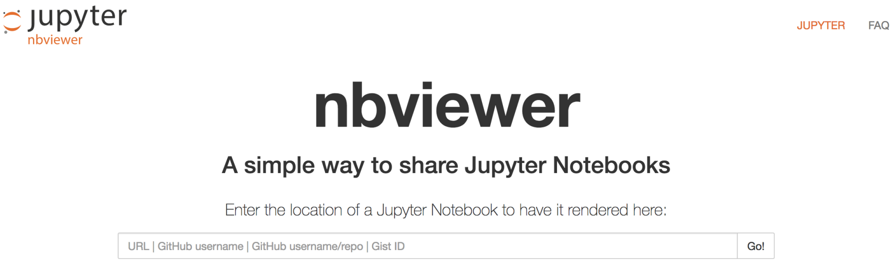
- A non-profit, open-source project maintained by a **strong** Community
- Adopted by the biggest in the Cloud industry (Google, Microsoft, Amazon...)
- And financed by the biggest (Google, Microsoft, EU Horizon 2020 program, Alfred P. Sloan Foundation...)

Inside the Python community (snakemake, conda...)

Integration with GitHub since 2015 (renderer)

Jupyter

Nbviewer : a static renderer for Jupyter notebooks



jupyter nbviewer

JUPYTER FAQ

nbviewer

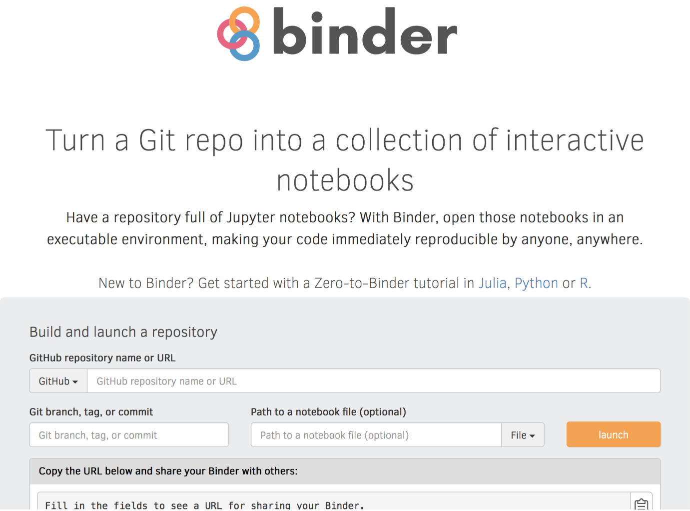
A simple way to share Jupyter Notebooks

Enter the location of a Jupyter Notebook to have it rendered here:

`https://nbviewer.jupyter.org/`

Jupyter

Jupyter + Docker = binder



binder

Turn a Git repo into a collection of interactive notebooks

Have a repository full of Jupyter notebooks? With Binder, open those notebooks in an executable environment, making your code immediately reproducible by anyone, anywhere.

New to Binder? Get started with a Zero-to-Binder tutorial in Julia, Python or R.

Build and launch a repository

GitHub repository name or URL

GitHub ▾ | GitHub repository name or URL

Git branch, tag, or commit

Git branch, tag, or commit

Path to a notebook file (optional)

Path to a notebook file (optional) | File ▾

launch

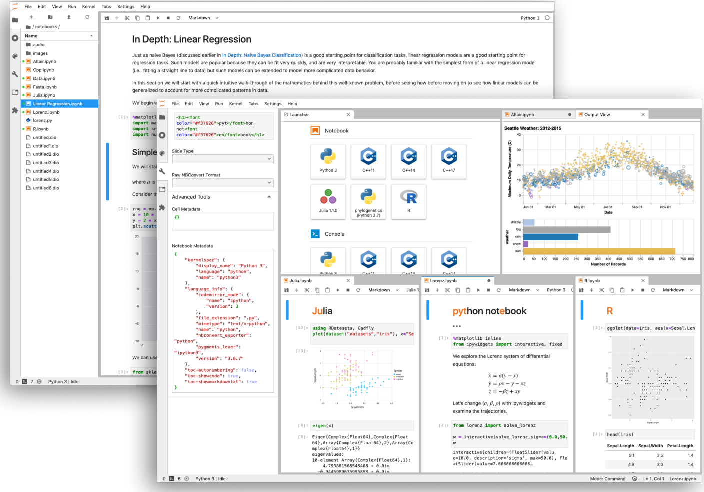
Copy the URL below and share your Binder with others:

```
Fill in the fields to see a URL for sharing your Binder.
```

`https://mybinder.org/`

Jupyter

Since June 2019 : Jupyter Lab v1.0 (now v3.0.16)



In Depth: Linear Regression

Just as naive Bayes (discussed earlier in In Depth: Naive Bayes Classification) is a good starting point for classification tasks, linear regression models are a good starting point for regression tasks. Such models are popular because they can be fit very quickly, and are very interpretable. You are probably familiar with the simplest form of a linear regression model (i.e., fitting a straight line to data) but such models can be extended to model more complicated data behavior.

In this section we will start with a quick intuitive walk-through of the mathematics behind this well-known problem, before seeing how before moving on to see how linear models can be generalized to account for more complicated patterns in data.

```
<h1><font
color="#f37626">pyt</font>hon
not<font
color="#f37626">e</font>book</h1>
```

Slide Type

Raw NBConvert Format

Advanced Tools

Cell Metadata

```
{}
```

Notebook Metadata

```
{
    "kernelspec": {
        "display_name": "Python 3",
        "language": "python",
        "name": "python3"
    },
    "language_info": {
        "codemirror_mode": {
            "name": "ipython",
            "version": 3
        },
        "file_extension": ".py",
        "mimetype": "text/x-python",
        "name": "python",
        "nbconvert_exporter":
"python",
        "pygments_lexer":
"ipython3",
        "version": "3.6.7"
    },
    "toc-autonumbering": false,
    "toc-showcode": true,
    "toc-showmarkdowntxt": true
}
```

Julia

```
[10]: using RDatasets, Gadfly
      plot(dataset("datasets","iris"), x="Se
```

```
[8]: eigen(x)
```

```
[8]: Eigen{Complex{Float64},Complex{Float
     64},Array{Complex{Float64},2},Array{Co
     mplex{Float64},1}}
     eigenvalues:
     10-element Array{Complex{Float64},1}:
         4.793881566545466 + 0.0im
       -0.9445989635995898 + 0.0im
```

python notebook

```
[1]: %matplotlib inline
     from ipywidgets import interactive, fixed
```

We explore the Lorenz system of differential equations:

$$\dot{x} = \sigma(y - x)$$
$$\dot{y} = \rho x - y - xz$$
$$\dot{z} = -\beta z + xy$$

Let's change (σ, β, ρ) with ipywidgets and examine the trajectories.

```
[2]: from lorenz import solve_lorenz

     w = interactive(solve_lorenz,sigma=(0.0,50.
     w

     interactive(children=(FloatSlider(valu
     e=10.0, description='sigma', max=50.0), Flo
     atSlider(value=2.666666666666…
```

R

```
[3]: ggplot(data=iris, aes(x=Sepal.Len
```

```
[1]: head(iris)
```

Sepal.Length	Sepal.Width	Petal.Length
5.1	3.5	1.4
4.9	3.0	1.4

Conclusion ?

Who’s the best?

Conclusion ?

Who's the best?

It depends...

Conclusion ?

Who's the best?

It depends...

- R analyses? Go for RMarkdown/RStudio
- R analyses for a publication ? Consider Jupyter with an R kernel

Conclusion ?

Who's the best?

It depends...

- R analyses? Go for RMarkdown/RStudio
- R analyses for a publication ? Consider Jupyter with an R kernel
- Python analyses ? Why do you even ask...

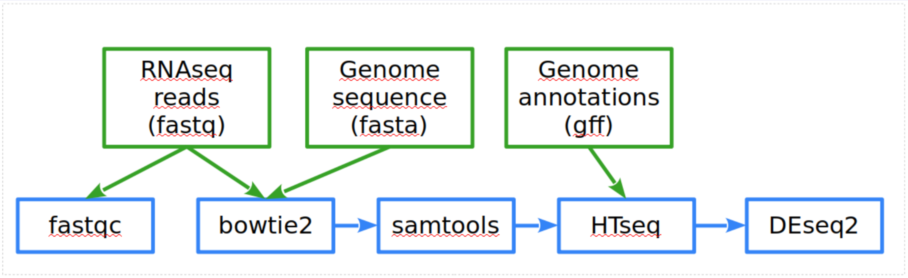
Practical session

Savoir FAIRe

- Markdown
- Learn the structure of an Rmd file
- Turn a script into a notebook
- Extend the notebook with new functionalities
- This afternoon: Jupyter with the IFB cluster

Analysis workflow

Analysis workflow



green=input, blue=tool

fastqc	control quality of the input reads
bowtie2	reads mapping on the genome sequence
samtools	mapped reads selection & formatting
HTseq	count table of mapped reads on genes (annotations)
DESeq2	statistical analysis: genes list having differential expression