# The Rough Road to Real-Life Reproducible Research

Christophe Pouzat, Arnaud Legrand, Konrad Hinsen

3 septembre 2018



M4-S0 : The Rough Road to Real-Life Reproducible Research

M4-S1 : Data Hell

M4-S2 : Software Hell

M4-S3 : Numerics Hell

M4-S4 : Conclusion

### Where are we?

### M4-S0 : The Rough Road to Real-Life Reproducible Research

M4-S1 : Data Hell

M4-S2 : Software Hell

M4-S3 : Numerics Hell

M4-S4 : Conclusion

### Reproducible Research Hell



WWW. PHDCOMICS. COM

Module 4. The Rough Road to Real-Life Reproducible Research

- 1. Data Hell
- 2. Software Hell
- 3. Numerics Hell
- 4. Conclusion

### Where are we?

M4-S0 : The Rough Road to Real-Life Reproducible Research

M4-S1 : Data Hell

M4-S2 : Software Hell

M4-S3 : Numerics Hell

M4-S4 : Conclusion

When we start to work on real data, we typically have to deal with two problems :

- the data are of diverse nature
- the data occupy a lot of memory

### Non-homogeneous data

- The influenza-like illness data from module can easily be presented as a table (2 dimensional object)
- Often the table form must be abandoned because
  - the columns don't all have the same length
  - the data can be a time series and a set of images, etc.

# Big data

- Text formats are not always appropriate for numbers
- Choice of a binary format because
  - Numbers occupy less memory
  - Numbers in text format must be converted to binary anyway for computation

# Text format features we wish to keep : metadata

- Text permits storing the data and all the rest...
- $\blacktriangleright$   $\Rightarrow$  add information about the data :
  - provenance
  - recording date
  - source
  - etc.
- This information about the data is what is called metadata
- ► They are vital for doing reproducible research

### Text format features we wish to keep : endianness

- Text format is universal
- Binary formats depend on hardware architecture and operating system
- The four-bit sequence 1010 can be read as
  - ▶ 1x1 + 0x2 + 1x4 + 0x8 = 5, which is little-endianness
  - ▶ 1x8 + 0x4 + 1x2 + 0x1 = 10, which is big-endianness
- A binary storage for reproducible research much specify endianness

# Binary formats for composite data allow storing metadata

Wanted : binary formats for

- working with big datasets of diverse nature
- storing metadata along with the data
- having endianness fixed once and for all

## 'FITS' and 'HDF5'

- The Flexible Image Transport System ('FITS'), developed in 1981 and still regularly updated
- The Hierarchical Data Format ('HDF'), developed at the National Center for Supercomputing Applications, is at its fifth version, 'HDF5'

### 'FITS'

- 'FITS' introduced and updates by the astrophysics community
- Format sufficiently general for use in different contexts

### The anatomy of a 'FITS' file

- One or mode segments : Header/Data Units (HDUs)
- ► A HDU is made up of :
  - ▶ a header (Header Unit) followed optionally by
  - the data (Data Unit)
- $\blacktriangleright \text{ Header} = \text{key-value pair} \rightarrow \text{metadata}$
- Data stored as binary tables (one to 999 dimensions) or as tables (text or binary)

# Manipulation of 'FITS' files

- The developers of the format offer a 'C' library and associated programs that are easy to use
- 'PyFITS' for Python users
- 'FITSio' for R users

### 'HDF5'

- Hierarchical organization, resembles a filesystem tree
- Structuring element : a group (similar to a directory) contains one of more datasets
- Groups can be nested
- No structure imposed on metadata
- No structure imposed on data they can be text

### Manipulation of 'HDF5' files

- More flexible format ⇒ the 'C' library is more complex thatn its 'FITS' equivalent
- The library is distribued with 'HDFView', a powerful tool for exploring and visualizing data
- 'h5py' is a very complete 'Python' interface
- Three 'R' packages : 'h5', 'hdf5r' et 'rhdf5'



### Git (hub, lab, ...) : not well suited for data storage



### Conclusions

- Real data  $\Rightarrow$  size and structure problems
- $\blacktriangleright \text{ Read data are complex} \Rightarrow \text{metadata}$
- 'FITS' and 'HDF5' = practical solutions
- In terms of complexity and flexibility : 'FITS' < 'HDF5'</p>
- ► Archiving platforms ⇒ persistent storage accessible for everyone

### Where are we?

M4-S0 : The Rough Road to Real-Life Reproducible Research

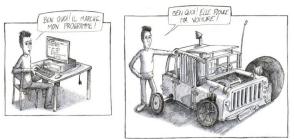
M4-S1 : Data Hell

M4-S2 : Software Hell

M4-S3 : Numerics Hell

M4-S4 : Conclusion

# Scaling up



WWW. 4VC\_DAMAS.FR



- A real spaghetti bowl
  - No global view
  - Interaction between multiple languages = danger

	Ddit.	View	c 1	nsert		Dell 1	Kerr	nel .	Woga	61	Help	Hide 0	Code			
+	× @	e		٠	н	Run		2 1	landowr			40 H	Ide show i	ode	B Exp	ort to HT
		le:	Mano			-	e aria	nal								
	in 10	surgicities inform														
		Manual sectors and a sector and														
		Ren de calone				stars (1)	(Here)								-	
		ndeator					netro									
		10							A vordra de		and the second				-	
		-							5.6. sorder							
		1	10.04										4) 10 (0) (0) (0) (0) (0)		04411	
		100	9.10	1.01				-					The on-the pro-			
		(40,700) (40,700)							Circle PARES				data wa hogi		-	
	A 181												Accord (at 1) or			
				-046.734				- 15.0	raws-cp							
	6410		(west)	-		but he	d ber	Internet	e (ne see ) ee	-	a lana ina	dan me				
						1004.5	47411		1.0	12.5	10	Farm				
							20014	-	3.5	4.0	14	Farm				
		E.				1874.3		8	43	4.0	14	Fana				
		1			10.71	10.00	79614		0	0.5	Ľ.	Farm				
		1			10.71	10.00	79614		0	0.5	Ľ.	Farm				
		1			10.71	10.00	79614		0	0.5	Ľ.	Farm				
					10.71	10.00	79614		43	0.5	Ľ.	Farm				
		-11 - 2222			10.71	10.00	79614		0	0.5	Ľ.	Farm				
			A STATE	100 H	10.71	10.000	79614		0	0.5	Ľ.	Farm				
			A DE LE COLOR DE L	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	10.71	10.000	79614		0	0.5	Ľ.	Farm				
			AN APPLY AND A APP	1	10.71	10.000	79614		0	0.5	Ľ.	Farm				
			AN OF		10.71	10.000	79614		0	0.5	Ľ.	Farm				
			AN OF THE OWNER OWNE		10.71	10.000	79614		0	0.5	Ľ.	Farm				
			AN OF THE OWNER OWNE		10.71	10.000	79614		0	0.5	Ľ.	Farm				
			A DI LI	· · · · · · · · · · · · · · · · · · ·	10.71	10.000	79614		0	0.5	Ľ.	Farm				
							79854			0.0	12	Press To control of the second				
				1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1		ana a	79854			0.0	12	Press To control of the second				
	a (10)		Annual An		aghi hear		Part of the second seco			0.0	12	Press To control of the second				
	34 (24)-		Annual An		aghi hear		Part of the second seco			0.0	12	Press To control of the second				
	34 (24)-		Annual An		aghi hear		Part of the second seco			0.0	12	Press To control of the second				
	34 (24)-		Annual An		aghi hear		Part of the second seco			0.0	12	Press To control of the second				
	34 (24)-		Annual An		aghi hear		Part of the second seco			0.0	12	Press To control of the second				
	34 (24)-		Annual An		aghi hear		Part of the second seco			0.0	12	Press To control of the second				
	34 (24)-		Annual An		aghi hear		Part of the second seco			0.0	12	Press To control of the second				

- A real spaghetti bowl
  - No global view
  - Interaction between multiple languages = danger





- A real spaghetti bowl
  - No global view
  - Interaction between multiple languages = danger

### Following Color Names by Web Image Sourchas Sea

i barrer Antien Effertrop

And the second s

Manual Annual Information

We as apply to the distance of the standard from the loss parts in a finance of the standard from the





statute for make on Ulbrac data

We first use a survey, here "it is if it when any distribution an order to be book any pp or words information of its sets. We will be an a single-book and any pp or source and the set of the sets of the set of the set of the subfinite of the sets of the order is the set of the set of the sets of the sets of the set of the subfinite of the sets of the

Pedital and V. Taring lands variance



### () () () (i () Poblacce

### A real spaghetti bowl

- No global view
- Interaction between multiple languages = danger

In characterization of the second at these problems of the of the second sec

Construction of the second second



- A real spaghetti bowl
  - No global view
  - Interaction between multiple languages = danger

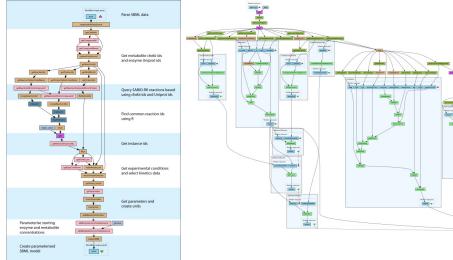
emace File Edit Options Buffers Tools Org Tbl Text Help • 🛄 😹 🗶 🛄 Save 🦐 Undo 🔛 🐘 🛤 🛤 🕰 #+TITLE: Incidence du syndrôme grippal #+LANGUAGE: fr #+OPTIONS: \*:nil num:1 toc:t #+PROPERTY: header-args :exports both \* Préface... \*\* Emacs 25... \* Préparation des données \*\* Description et téléchargement... \*\* Extraction des colonnes utiles... \*\* Vérification de la cohérence des données... \*\* Gestion des dates... \*\* Un premier apercu des données Nous passons au langage R pour avoir un résumé statistique de notre jeux de données. #+BEGIN SRC R :results output :var data=date-inc-sorted sunnary(data) #+END SRC #+RESULTS: date : Min. :1984-12-31 Min. : : 1st Ou.: 1993-02-25 1st Ou.: 5164 : Median :2001-04-16 Median : 16188 : Mean :2001-04-15 Mean : 63053 : 3rd Ou.: 2009-06-04 3rd Ou.: 49576 : Max. :2017-07-24 Max. :1001824 Regardons enfin à quoi ressemblent nos données ! #+BEGIN SRC R :results output graphics :file inc-plot.png :var data=date-inc-sorted plot(data, type="l", xlab="Date", vlab="Incidence hebdomadaire") #+END SRC #+RESULTS: U:--- analyse-syndrome-grippal-orgmode.org Top L1 Git:master (Org) eginning of buffer

- A real spaghetti bowl
  - No global view
  - Interaction between multiple languages = danger

a series and the base based of the series of the	Harat 2000 to here have been		[ ] Sections And Thirs Sect Transcere Section And Stream And Transcere	(HE)10.710
a second se		And and a second		Contraction Spart Lines of the International Social States
B. Mariada and Part 1	a set all locates les los locates los locates/los	BUILDER - Contraction Contraction	and the second s	in the second se
and of the state o	A STATE CONTRACT OF A CONTRACT	A - Manager and an interaction of the Association of the	100 miles	100 Arrs 100
<ul> <li>See The Second Control of Second Co</li></ul>	person have been been preserved by the	San	N In Inc.	
Comparison of the second	The first function of the first and information of the second sec		and the set of the set	
			44 - 10000 45 20 20 20 20 20 20 20 20 20 20 20 20 20	in the materia
5 22 23 23	a a war block, many shire there have	Long R. Martin and M. Kong, Phys. Rev. B 100, 15 (1990) Rev. and Martine and Martine and Society an	6 (and a second se	6.4 an enclose and the second seco
1 101 101 10	First and all all all all all all all all all al		10	have a start of the start of the start of the start
<ul> <li>CALLERING MARKET</li> <li>CALLERING</li> </ul>		The Property of Arms of the second se	to that a my heart when a sit in the set of	water and the later of the second second
				party on and much only a start of the start
B. Samuel Kitserstein Annuel State (1998) 1 (2019)		And the first function of the second se	a state and a state of the state	The Party State of the State of
2. Second least	No. 10. 10. Annual and a statement of the statement		and the solution in the solution of the soluti	
Second Seco	A set of the total the set	• As it is realized with a simple fits the same fit was and it uses if it is not an analyzed on a simple fit was and an analyzed with a same interaction of the same and a simple same is the same fits the same same is a simple same is a simple same is the same fits a billing of the same same is a simple same is the same same is a billing of the same same is a simple same is a simple same is a simple same same is a simple same same same is a simple same same is a simple same same same same same same same sam		
5 751	Planet and it an	The Real Process of the American Street Process of the Process of		E: 2.022
<ul> <li>A start start of a start start</li> <li>A start start start start start</li> </ul>	had been and the had been and			And Address Terrar
The state of the s	and any set of the barrier water and the set	And have been as the field of the best for a characteristic of the barry of the bar		CONTRACTOR OF A DESCRIPTION OF A DESCRIP
the second concerning from the second s		[20] K. Mark, K. Mark, K. Mark, R. Mark, S. Mark, S. Markel, M. Mark, "Phys. Rev. Lett. B44, 81 (1997) 11 (1997).	Total cite Jone	NAME & ADDRESS OF TAXABLE AND ADDRESS OF
a set of the	Start and a second where a second start and a secon	test and the second second second second		( he al he he
a and a second		(and a spectrum)	Review range a	M 27 161 2
B. PETA Distance, Salar Marke Tor Inflation Brind B. Alf in Consulty, Kind on Marke Tor Inflat Brind	The area of the constrainty of the second	A DATE OF THE OWNER.	h = // a hat help handal	a a han a
		and the second sec		a to transmo
a server of the server of the basely to be based	which a first that is a first the relative provider. Here should not a provide reaction	A REAL POINT OF A REAL POINT AND A REAL POINT AND A REAL POINT AND A	10 1000	P. In Johnson II.
Marke article colours have been been to been at the second of the second second at the second second second at the second s	Harmonic also film had a generalization for the francessory with the data to the second	THE REAL PROPERTY AND AND ADDRESS OF A DREAM AND ADDRESS OF ADDRES		and or Publication of the
	A. M. M. M. Andrewski, and the state of the state of the last of the state of the state.	W. Start is because in the second se second second sec	ALL A LAND AND ALL AND A	
Contraction of the second second		Contraction of the providence of the providence of the second sec	Notice and assessed to realize the contract of the second	00 00 000000 00 00 00 000000 00 00 00 00
	t is minimum closedenations at it. http://	Chrester rest time-		A Date of the second se
World and an Art of Phone And Analytics and the other And a setting	Terrer and the standard in the	No water during prime of industry was indus- ted on calors to by Memorical Perception, and instance for water, and a restriction of the col-	<ul> <li>M. Organization, "Inter-Tables," Inter-Tables, Inter-Tables, 2011.</li> </ul>	and the second s
There makes a	An item for the	1 m	the base base had	
Reading to the second s	maximum and and share the second seco	Longer and an extent may	an in Joseff Ham	A 10 Parter 10
a be- mine real	management of the second secon	A di te carra contra sulla resta data del con que contra a conte entre a partecente en la desentata constato de salura effer pero		ALL DEL TO
The state of the s	mandaria in the same wanted as the last of same at a	THE EXCLUSION OF THE R. P. LEWIS CO., LANSING MICH.	In Distance In In Distance State	and an and an an
			In Property of the Industry Strategy	B2 (2 2
Works, N. 199, States for addition from the state of the	and and a second part of the sec	a cancel activate the proof activate and a set of the second set of the second	We Record over 10 10 Descent Provide	AB         AT         AT         AT           AB         AT         AT         AT         AT           AB         AT         AT         AT         AT
ELECTRICAL PLATERS	The same	Contra among a particular and a second		March Coloradore 14
Name of a factor field of the second		Production of the latter for an international system of the last function in the	International Contraction of the Parameters of t	And in Amount in the
The second	the second	President Constraint teaching on a president and the president sector of the president	in Antonia State of the Antonia State of the	and Participation of the
THE R. LEWIS CO., LANSING MICH.	1. State Control and Control of Control o	Distance and a service of the servic		All in Annual States and particular
a second state a first in a second read of a first state of a second state of a s	the case of the second second second	[5] Shing Re (Sec) and a strength of a strength of particular devices of the strength of th		
the party of the party of the party of the party of the state of the party of the p	with the bring with the same and	<sup>1</sup> Second to be the second	and course in an and and from the second sec	an an instant in the
STERNING AF Proven		[15] S. Andrew M. M. Markett, and M. M. Markett, and an experimental second structures of the state of the local structure of a structure of the structure o	an Pa San in Manager Annale	and the second s
	3. The Electricity		and the fact of the financial states	
The second party have a first and and and and and and the second party second and	11 95-05		(a) See Sec. 1. (a) 1. (b) Sec. 1. (b) Sec. 1. (b) Sec. 1. (c)	B2 02 2
Fig Bridding Benfelder and Water and	which handles have	A CONTRACT OF THE RECEIPTING OF AN ADDRESS OF A CONTRACT O		
And the second s	and states and an other states and and	the second of the second s	Marchine Annual Inc. 10 Marchine Annual Marchine Marchine Annual Inc. 10 Marchine Annual Inc.	10 00 (1.00) (1.00) 00 00 10 00 (1.00)
manda. Annual an and an annual a particular a	· · · persite mentioners · · ·	A DESCRIPTION OF A DESC		
a la de la sector a la sector de la sector de las sector a sector a			W Schultz in 10 Stands States	Mar Internet A
10.5.5	a contraction contraction for			
and I taken taked and all hand	<ul> <li>Bernell A sun</li> <li>Constants</li> <li>Bernell A sun</li> </ul>	CONTRACT CONTRACTOR OF A DESCRIPTION OF A DESCRIPANTE A DESCRIPANTE A DESCRIPANTE A DESCRIPTION OF A DESCRIP	en funtumen in in futures hauten	10 11 111 11 11 10 12 11001 100 11
Players by his herebere have been			and the state for the fact that	
Playing his by house houses hereing	<ul> <li>and any first start of the start base from the start of t</li></ul>	(a) an addition of the descent of the second state of the secon	and the second s	and an Annual An
Frank ton be bester biging the		The second		
Participa Anno an Anno anno anno anno anno	<ul> <li>A second a frame conception to the loss of the loss o</li></ul>	An instruction of the production and the second sec	and Texa and the second	
		a con-		and the function of the second
	All Million Control Contr	Charles and a second second	and the section of th	
		(a) and a final particular of particular structures.		and Access in the second secon
Participant and the Second Street Street	Westminister merelinant for the second	118		
P spread pressure by propriets further holes for pressure	Not the state of t	1.10 The second seco	an particulation in a second	and the second s
a particular Sussession (1) Data Sana Sana Sana Sana Sana Sana Sana S	To consider a fill allow the advector and the second to the second			in in paratical in
France Property is bother from Property		In the case of the second system of the second system in the second system of the second s	and an and a second sec	E and C and an
	Receipt for any effective composition on any proceeding of the second proceeding of the second secon	Carbon and a second s		A NAME AND THE PARTY DESCRIPTION OF THE
	Version of an one-of-one for instance and a constraint of the second state of the second state of the		and the spectrum of the law of the state of	the same local particular to the last same
Property property 1 Australia View Name	And a reason was not a first the set of the	B. S. Martin, C. R. Martin, R. S. Shinaka, and C. Shinaka, K. Matta, M. A. Shinaka, and C. S.		and a later of the second seco
Playing provident by Social Instally State	states and a second sec		andras in in product	- increasing to operation of a straight

- A real spaghetti bowl
  - No global view
  - Interaction between multiple languages = danger

### ... that is difficult to orchestrate



### ... that is difficult to orchestrate

### Workflows :

- Clearer high-level view
- Composition of codes and data movement made explicit
- ► Safer sharing, reusing, and execution
- Notebooks are a variant that is both impoverished and richer
- ► No simple/mature path from a notebook to a workflow Examples :
  - Galaxy, Kepler, Taverna, Pegasus, Collective Knowledge, VisTrails
  - Light-weight : dask, drake, swift, snakemake, ...
  - Hybrids : SOS-notebook, ....

### The mess of expensive computations

Long-running computations and big datasets

- JupyterHub and supercomputers : under development
- Checkpoints and caching
- Workflows permit scaling up

### Complex ecosystems

What is hiding behind the simple

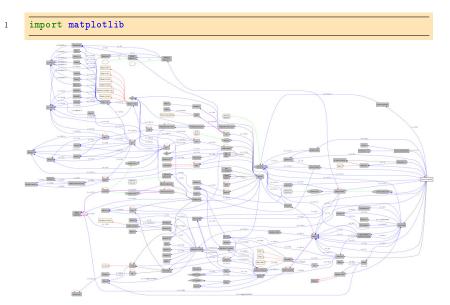
```
import matplotlib
```

1

```
Package: python3-matplotlib
Version: 2.1.1-2
Depends: python3-dateutil, python-matplotlib-data (>= 2.1.1-2),
python3-pyparsing (>= 1.5.6), python3-six (>= 1.10), python3-tz,
libjs-jquery, libjs-jquery-ui, python3-numpy (>= 1:1.13.1),
python3-numpy-abi9, python3 (<< 3.7), python3 (>= 3.6~),
python3-cycler (>= 0.10.0), python3:any (>= 3.3.2-2~), libc6 (>=
2.14), libfreetype6 (>= 2.2.1), libgcc1 (>= 1:3.0), libpng16-16 (>=
1.6.2-1), libstdc++6 (>= 5.2), zliblg (>= 1:1.1.4)
```

### Complex ecosystems

### What is hiding behind the simple



### Complex ecosystems

No standard :

1

 $\mathbf{2}$ 

3

4

5

6

- Linux (apt, rpm, yum), MacOS X (brew, McPorts, Fink), Windows (?)
- Neither for installation nor for retrieving the information...

```
import sys
                                                library(ggplot2)
 print(sys.version)
                                                sessionInfo()
                                          2
 import matplotlib
 print(matplotlib.__version__)
                                               R version 3.4.3 (2017-11-30)
 import pandas as pd
                                               Platform: x86_64-pc-linux-gnu (64-bit)
                                               Running under: Debian GNU/Linux buster/sid
 print(pd.__version__)
                                               Matrix products: default
                                               BLAS: /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.7.1
3.6.3 (default, Oct 3 2017, 21:16:13)
                                               LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.7.1
[GCC 7.2.0]
2.1.1
                                               locale:
0.20.3
                                                [1] LC CTYPE=fr FR.UTF-8
                                                                             LC NUMERIC=C
                                                [3] LC TIME=fr FR.UTF-8
                                                                             LC COLLATE=fr FR.UTF-8
                                                [5] LC MONETARY=fr FR.UTF-8
                                                                             LC MESSAGES=fr FR.UTF-8
                                               other attached packages:
                                               [1] ggplot2_2.2.1
                                               loaded via a namespace (and not attached):
                                                [1] colorspace_1.3-2 scales_0.5.0
                                                                                    compiler_3.4.3
                                                                                                    laz
                                                [5] plvr_1.8.4 pillar_1.1.0
                                                                                    gtable_0.2.0
                                                                                                    tib
                                                [9] Rcpp_0.12.15 grid_3.4.3
                                                                                    rlang_0.1.6
                                                                                                    mun
```

# Controlling one's environment

### A controlled environment :

 Work in a virtual machine (heavy) or a Docker container (light)

### Preserve the mess

 Automatic capture of the environment

CDE, ReproZip, CARE

### Cleaning up

- Start with a clean environment
- Install only what's strictly necessary (and document it)
- Docker, Singularity, Guix, Nix,

## The test of time



## Backwards compatibility

Python and its rapidly evolving environment

```
1 python2 -c "print(10/3)"
2 python3 -c "print(10/3)"
3
```

3.3333333333333333335

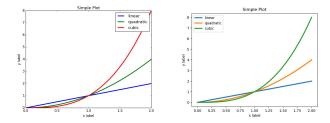
#### Backwards compatibility

Python and its rapidly evolving environment

```
1 python2 -c "print(10/3)"
2 python3 -c "print(10/3)"
```

3

3.33333333333333335



## Backwards compatibility

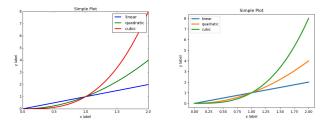
Python and its rapidly evolving environment

```
python2 -c "print(10/3)"
python3 -c "print(10/3)"
```

3.33333333333333333335

1

2



- Cortical Thickness Measurements (PLOS ONE, June 2012) : FreeSurfer : differences were found between the Mac and HP workstations and between Mac OSX 10.5 and OSX 10.6.
- Format incompatibility between orgmode 7.\*, 8.\*, 9.\*, etc.

## Rapid development tools

Rapid but also fragile and unstable :

- Correction or introduction of bugs
- It becomes necessary to check regularly if environments can still be reconstructed and work (continuous integration, regression testing)

Popper:<mark>http://falsifiable.us/</mark>

Another option :

Limit onself to what is manageable (C for example)



# Archiving

#### Source code management

Git (hub, lab, ...) : stable, open, ... durable?

Google Code, Gitorious, Code Spaces
 Software Heritage
 Archives-ouvertes fr

#### **Environment management**

- Longevity of access to Docker Hub, Nix repository, Code Ocean, ... ?
- Once an environment is frozen, what's the lifetime of a virtual machine, a Docker image, ...?

Preserve as much information as possible automatically

► Software, versions, installation procedures

#### Where are we?

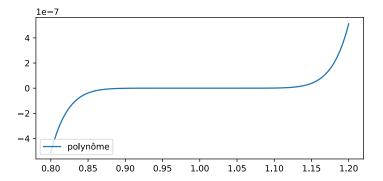
M4-S0 : The Rough Road to Real-Life Reproducible Research

M4-S1 : Data Hell

M4-S2 : Software Hell

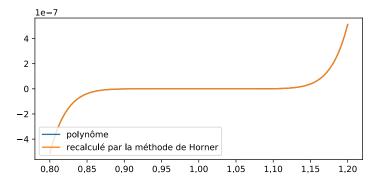
M4-S3 : Numerics Hell

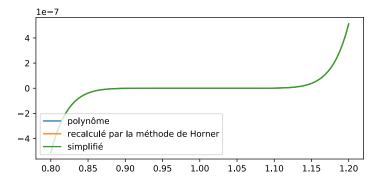
M4-S4 : Conclusion

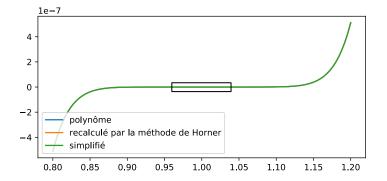


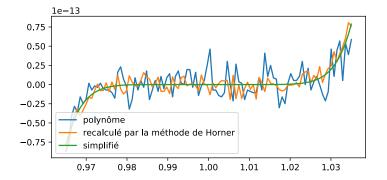
def polynome(x):
 return x\*\*9 - 9.\*x\*\*8 + 36.\*x\*\*7 - 84.\*x\*\*6 + 126.\*x\*\*5 \
 - 126.\*x\*\*4 + 84.\*x\*\*3 - 36.\*x\*\*2 + 9.\*x - 1.

1 2 3









# Rounding

- Every operation includes implicit rounding.
- ▶ a+b is actually arrondi(a+b).
- Unfortunately :
  - $\texttt{arrondi}(\texttt{arrondi}(\texttt{a}{+}\texttt{b}){+}\texttt{c}) \neq \texttt{arrondi}(\texttt{a}{+}\texttt{arrondi}(\texttt{b}{+}\texttt{c})).$
- Operation order therefore matters.

For a reproducible computation, operation order must be preserved !!!

#### How to explain it to my compiler?

To speed up computations, compilers may change operation order, and thus results.

Two options for computing reproducibly :

Insist on the preservation of operation order,

if the language permits it.

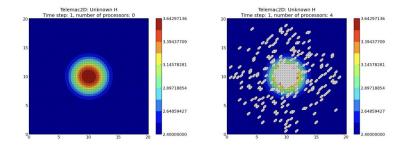
- Example : Module 'ieee\_arithmetic' in Fortran 2003
- ► Make compilation reproducible :
  - Record the precise compiler version
  - Record all compilation options

#### Parallel computation

- ▶ Goal : get results sooner → Minimize communications between processosrs → Adapt data distribution → ... hence the operation order ☺
- Consequence : results depends on the number of processors !

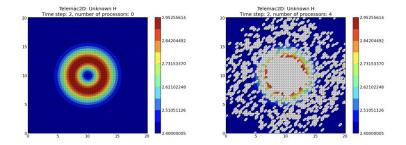
Minimizing the impact of parallelism is an active research topic.

#### Parallel computation : example



Source : Rafife Nheili, PhD. Thesis, University of Perpignan, 2016

#### Parallel computation : example



Source : Rafife Nheili, PhD. Thesis, University of Perpignan, 2016

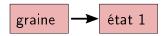
# Computing platforms

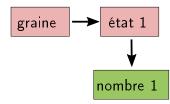
- Computing platform : hardware + infrastructure software
- Computation = platform + software + data
- ► The platform defines the interpretation of the software.
- Platform and software define the interpretation of the data.
- Other platform-defined aspects :
  - integer representation (16/32/64 bits)
  - error handling

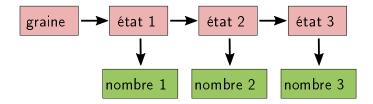
#### Random numbers

- Used to simulate stochastic processes.
- ► In practice : pseudo- random numbers.
- Series of numbers that appear to be random...
- **•** ... but are generated by a deterministic algorithm.









## Reproducibility in theory

- $\blacktriangleright$  Principle : same seed + same algorithm  $\rightarrow$  same series
- The seed is often chosen as a function of time
- It must be defined in the application code

## Reproductibility in practice

- Same seed + same algorithm → same series : not obvious with floating-point arithmetic !
- A simple trick to permit verification : test the first values of the series.

## Example : the Python language

```
1 import random
2
3 random.seed(123)
4 for i in range(5):
5 print(random.random())
```

0.0523635988509 0.0871866775226 0.40724176367 0.107700234938 0.901198877952

# Example : the Python language

1	import random
2	
3	random.seed(123)
4	assert random.random() == 0.052363598850944326
5	assert random.random() == 0.08718667752263232
6	assert random.random() == 0.4072417636703983

## Take-home message

The results of a numerical computation depend on

- the software
- the input data
- the computing platform : hardware, compilers, ...
- Platform influence is important for floating-point arithmetic. Record all parameters on which your results may depend :
  - compiler version, compilation options
  - hardware (processor type, GPU, ...)
  - number of processors
- When using a random number generator, define your own seed and verify the first elements of the series.

#### Where are we?

M4-S0 : The Rough Road to Real-Life Reproducible Research

M4-S1 : Data Hell

M4-S2 : Software Hell

M4-S3 : Numerics Hell

M4-S4 : Conclusion

## The take-home message of the MOOC

#### A major concern

- Scientific method
- Inspectability and reusability

#### Tools exist

- Computational documents and workflows, version control and archives, software environments, continuous integration, ...
- These tools evolve constantly
  - Choose those that are best adapted to your context
  - Find a compromise between modern and durable tools

Use in practice, don't get discouraged !

- Takes notes rigorously
- Make information useable and accessible
- Improve in small steps