Reproducibility in data science – An overview

Making reproducible research in DS for dummies?

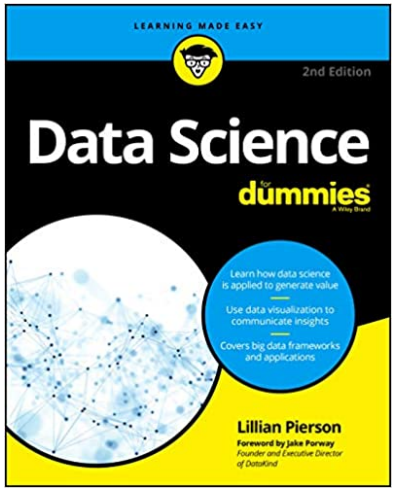
Jean-Baptiste DURAND[1]

[1]Ensimag, Laboratoire Jean Kuntzmann and Inria, Statify (Grenoble)

Grenoble, April 9, 2021

Acknowledgments:

- Anne-Marie Dols
- Arnaud Legrand
- Arnaud Seigneurin
- Hans Rocha IJzerman
- Simon Barthelmé



Reminder: Statement of the problem

Randall, D. and Welser, C.

The Irreproducibility Crisis of Modern Science: Causes, Consequences, and the Road to Reform.

National Association of Scholars. Princeton, NJ, USA (2018).

> "A 2012 study, for example, aimed at reproducing the results of 53 landmark studies in hematology and oncology, but succeeded in replicating only six (11 percent) of those studies."

Some possible achievements in reproducible DS

What is 1) *common* to computer science / others vs.
2) *specific* to data science?

Randomness: intrinsic variability making exact replication of experiments sometimes impossible.

Possible ambitions / challenges:

- To replicate my own analyses
- To allow other researchers to replicate my analyses
- To allow other researchers to replicate my experiments and qualitative conclusions
- Conclusions and perspectives.
  Where does my responsibility lie?

Replicability

Research data: position of CNRS

[...] CNRS just implemented a plan for “Research Data”. What are its aims?

Alain Schuhl.: This plan and the proposed actions are related to **data**, which are destined, as stated by the European Community, to be “as open as possible, as closed as necessary”, should it be raw or processed data in any format, texts and documents, **software**, **algorithms**, **protocols**, etc.

Source: `https://www.cnrs.fr/fr/cnrsinfo/cnrs-un-plan-ambitieux-pour-des-donnees-accessibles-\et-reutilisables`

Why would I need version control?

- Principles of version control

Why would I need version control?

- Principles of version control
- Collaborative work

Why would I need version control?

- Principles of version control
- Collaborative work
- What if I work alone?

Why would I need version control?

- Principles of version control
- Collaborative work
- What if I work alone?

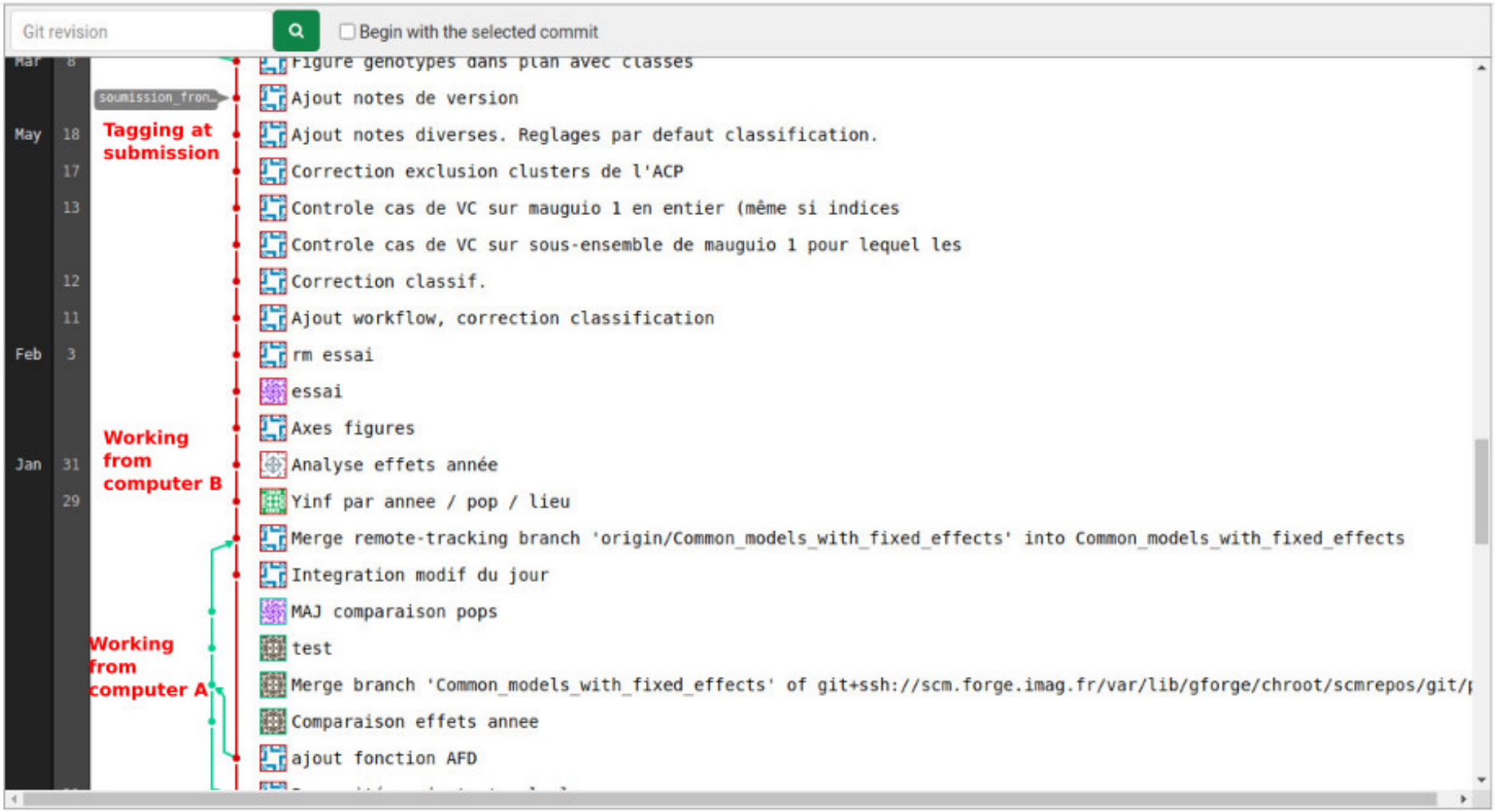
- Revisions: commenting / tagging

Why would I need version control?

- Principles of version control
- Collaborative work
- What if I work alone?

- Revisions: commenting / tagging
- Do I need to rely on external servers?

Version control with git: an example



Gitlab: more than git

- Interacting with users

- Organizing collaborative work
- Continuous integration
- Managing docker images

- ...
- `https://gricad-gitlab.univ-grenoble-alpes.fr/`

Why would I need virtualization?

- Principles of virtualization

Why would I need virtualization?

- Principles of virtualization
- Collaborative work

Why would I need virtualization?

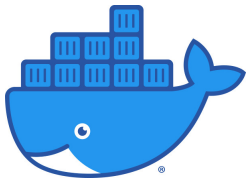
- Principles of virtualization
- Collaborative work
- What if I work alone?

Why would I need virtualization?

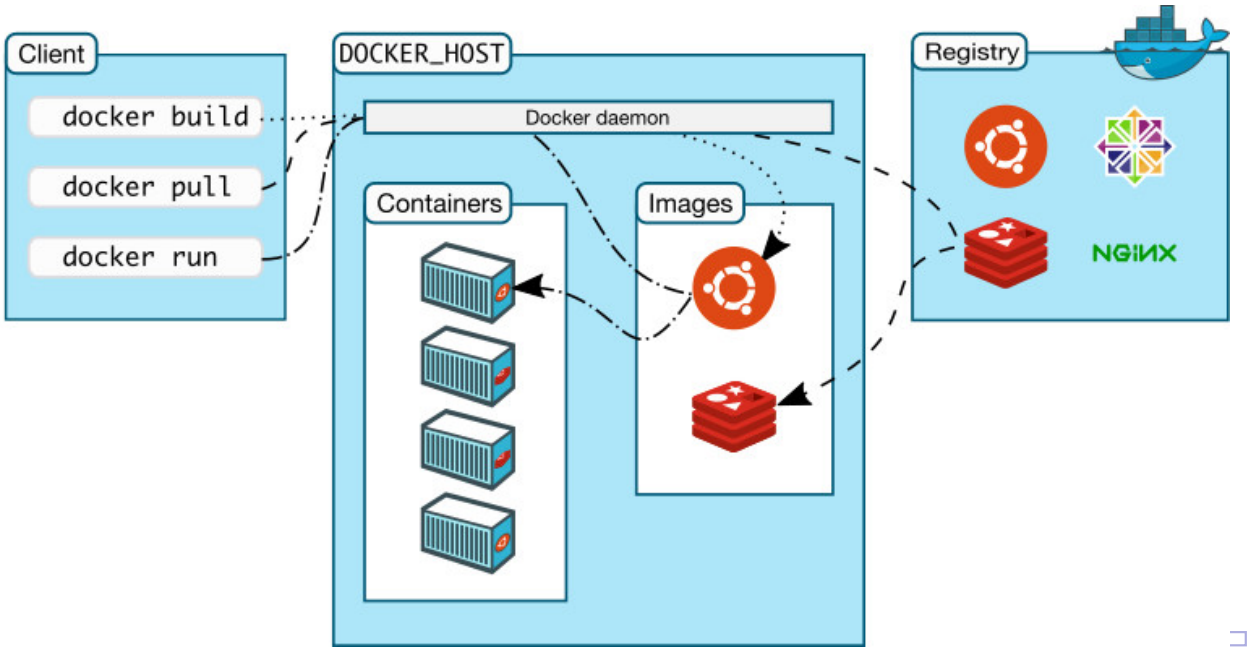
- Principles of virtualization
- Collaborative work
- What if I work alone?

- Virtualization and MS Windows

Docker



- Light virtual system (not fully-furnished with drivers)
- Runs on “any” host
- Hosts / Images / Containers

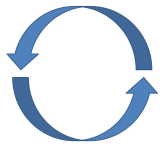


Docker files (recipes)

```
# Download base image ubuntu 18.04 N.B. Prefer stable debian versions
FROM ubuntu:18.04

# Metadata
LABEL maintainer="Jean-Baptiste Durand <jean-baptiste.durand@inria.fr>"
LABEL version="1.0"

# Environment variables
ENV LD_LIBRARY_PATH="/usr/local/cuda/lib64/:/usr/local/nvidia/lib:/usr/local/nvidia/lib64"

# Update the image & install some tools
RUN apt update && apt install -y gedit

RUN apt install -y python3-pip

RUN python3.7 -m pip install --upgrade pip && \
        python3.7 -m pip install jupyter==1.0.0

[...] # Skipping repetitive stuff

RUN mkdir /root/r_analysis &&\
         cd /root/r_analysis &&\
         echo 'install.packages("sp", repos="'$R_CONTRIBS'")' >> r_install.txt &&\
         Rscript r_install.txt &&\
         rm -Rf /root/r_analysis

# Switch to new user
USER $user

COPY ./notebooks/ /home/stat/devlp/bnp_mrf/notebooks

# Change working directory
WORKDIR /home/stat/devlp/bnp_mrf/notebooks/
```

Layers / registry servers

- Images made of layers
- Layers can be shared
- By default containers cannot see hosts
- Running a container: the shell
- Running a notebook server



- Pulling images on registry servers
- Cost of the whole thing
- MS Windows images

Singularity

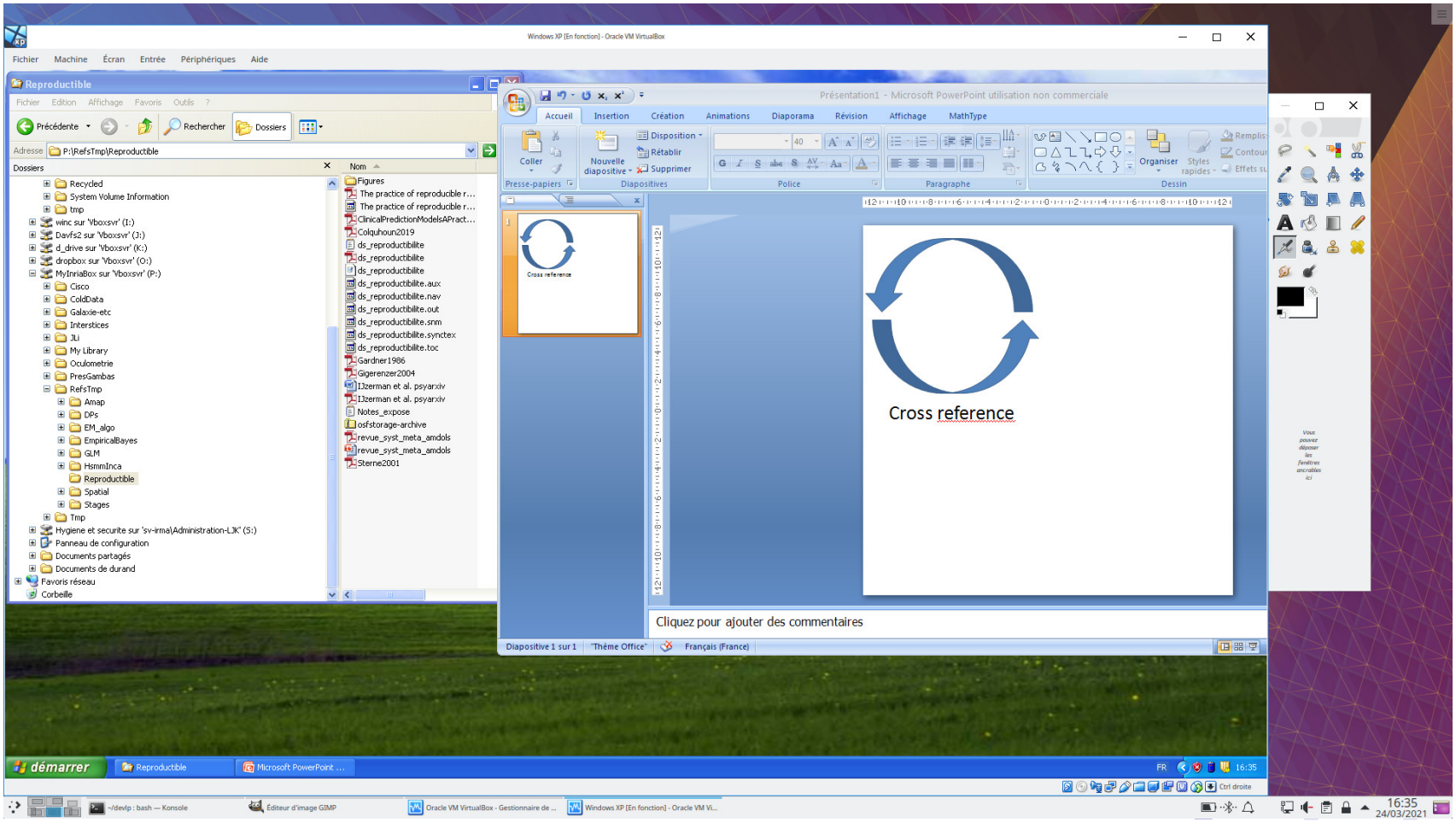
- Pretty much the same as Docker 
- Pretty much the opposite regarding isolation from host
- No persistent container, no layers, no image registry servers?

Singularity files (recipes)

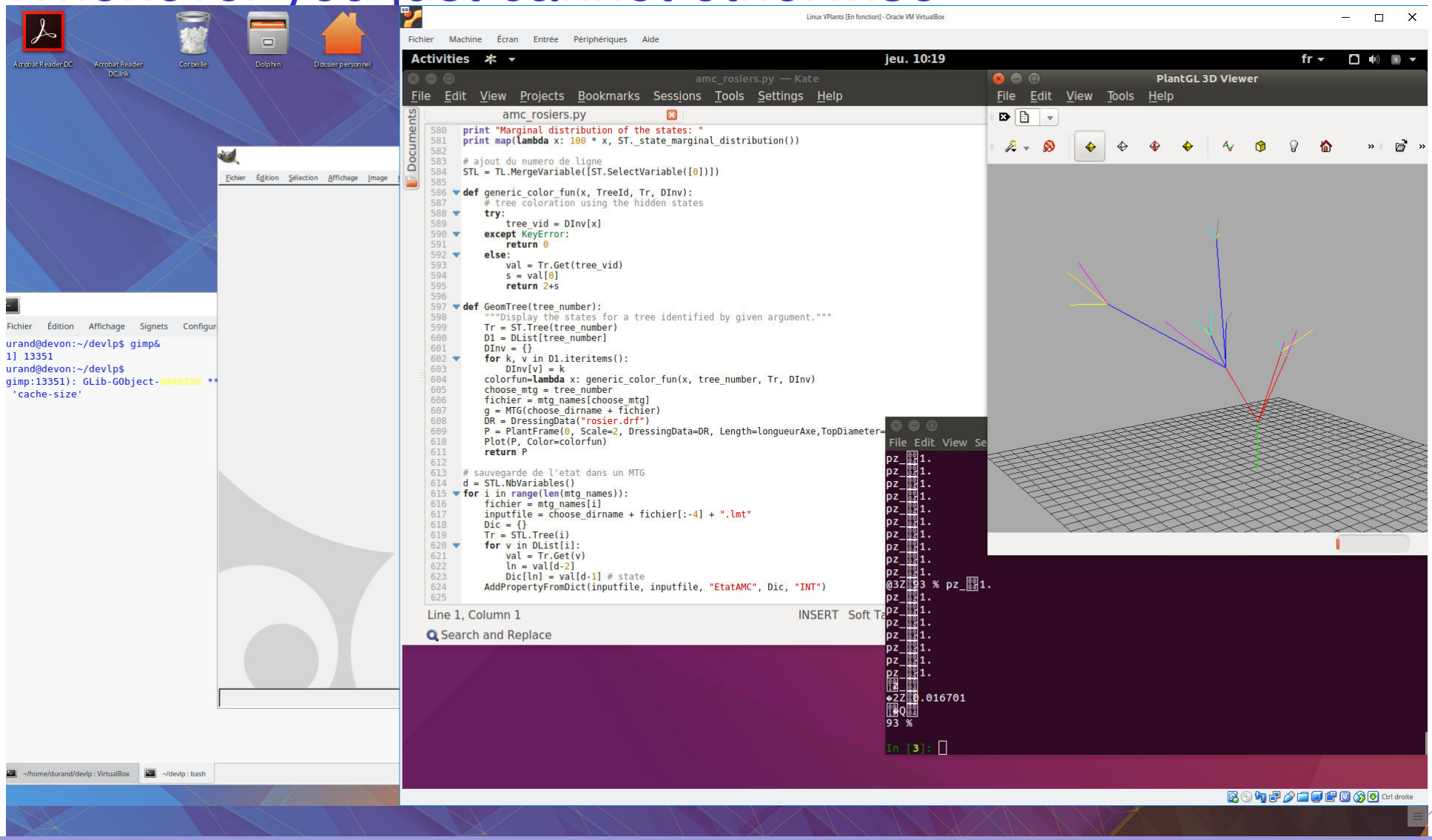
```
Bootstrap: docker
From: ubuntu:18.04

%labels
AUTHOR Jean-Baptiste Durand
VERSION="1.0"

%setup
mkdir -p ${SINGULARITY_ROOTFS}/r_analysis

%files

# Environment variables
%environment
export LANG="C.UTF-8" LC_ALL="C.UTF-8"
export LD_LIBRARY_PATH="/usr/local/cuda/lib64/:/usr/local/nvidia/lib:/usr/local/nvidia/lib64"

%post
# Update the image & install some tools
apt update && apt install -y gedit

apt install -y python3-pip

python3.7 -m pip install --upgrade pip && \

[...] # Skipping repetitive stuff


%apprun jupyter
jupyter notebook --ip 0.0.0.0 --no-browser --allow-root

%apphelp jupyter
Use "run --app jupyter -e tensorflow-1-4-1_gpu_count.simg" to run a jupyter notebook
```

Full virtualization and VirtualBox

- Presumably exports (much) more than needed
- It works...



Whenever you just cannot otherwise



Package management: conda

- Creates virtual environments
- Install packages from `https://anaconda.org`
- Limited isolation with system
- Most released packages are R- python-oriented
- Possibility of exporting your environment ($\rightarrow$ sharing)
- Limited OS interoperability
  (availability of packages, e.g., boost-python=1.60.0)
- Custom package creation/releasing more demanding than using Docker.

Package management in R: renv

Overview

The renv package helps you create **r**eproducible **env**ironments for your R projects. Use renv to make your R projects more:

- **Isolated**: Installing a new or updated package for one project **won't break your other projects**, and vice versa. That's because renv gives each project its own private package library.
- **Portable**: Easily transport your projects from one computer to another, even across different platforms. renv makes it easy to install the packages your project depends on.
- **Reproducible**: renv records the exact package versions you depend on, and ensures those **exact versions** are the ones that get installed **wherever you go**.

Environment control with Sumatra

Through specific modification of commands in shells for snapshots of

- the code that was run (through connections with git, etc.)
- parameter files and command line options
- platform on which the code was run

```
smt configure -e python -m myscript.py
smt run -i input_file -o output_file
    --version=3e6f02a --label=mytest
    --reason="Test the effect of using
              a low-pass filter"
smt comment "Doesn't make much of a difference"
smt tag "Figure 5" 20141203-093401
```

Incorporates comments:

- the reason for which the simulation/analysis was run
- a summary of the outcome of the simulation/analysis

Dedicated to projects involving (somewhat large-scale-) numerical simulations / analyses.

See Davison (2012)

Summary

- Identify your needs and those of your collaborators
- Determine demanded level of reproducibility
- Choose a set of tools accordingly
- Track precise versions of your code and environment
- In practice, prefer stable distributions / repositories (debian archive, Guix / Nix)

R / python notebooks

Pros

- Ensures consistency of a set of results
- Check reproducibility
- Documentation of analyses and choices
- Reminding necessity of commenting code

Cons

- Re-run even what is unchanged
- Linear point of view on processing
- Merging concurrent changes

CAR-simulate

March 30, 2021

[1]:
```
#!/usr/bin/python3
```

0.1 Simulate CAR models

0.1.1 Use VBAR All S 10 µM A to define the graph

Inspired by https://ecomorphisms.holobio.me/en/2019/11/27/simulating-a-spatial-conditional-autoregressive-model-car-from-a-graph-gve/ To be improved with Besag & Kooperberg (1995)?

Defining paths and variables (update to reflect your own config)

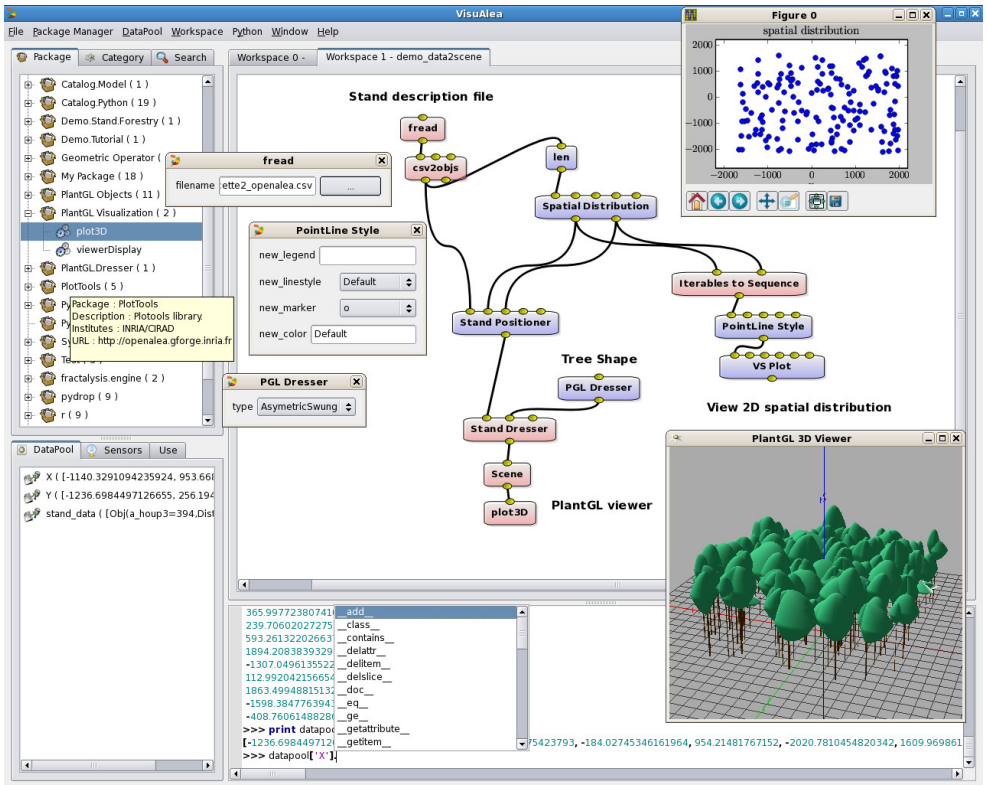
It is assumed that some directory (base_dir) contains two sub-directories: neurites and bnp-mrf

[2]:
```
# %matplotlib notebook
# allows pickle to be used in notebook
# Do not use %matplotlib notebook  if you do not intend to use pickle

import matplotlib.pyplot as plt
import numpy as np
import pickle

import os

from pathlib import Path

notebook_dir = os.getcwd() + os.sep

# path to neurite git directory

data_name = "VBAR All S 10 µM A"

data_path = notebook_dir + data_name + ".txt"
# path to data set with log rescaling (this file is created below)
data_path_log = notebook_dir + data_name +"_log_length.csv"
```

1

Workflows

- Visually and formally **sharing** the whole data processing approach
- Re-run only what is necessary by caching
- Certification of intermediate results provenance



VisuAlea, Pradal *et al.* (2015)

The R package *reproducible*

Collection of high-level, machine- and OS-independent tools for making deeply reproducible and reusable content in R.

The two workhorse functions are Cache and prepInputs; these allow for: **nested caching**, robust to environments, and objects with environments (like functions); and data retrieval and processing in continuous workflow environments. In all cases, efforts are made to make **the first and subsequent calls of functions have the same result, but vastly faster at subsequent times** by way of checksums and digesting. Several features are still under active development, including cloud storage of cached objects, allowing for sharing between users.

Panorama of replicability issues in data science

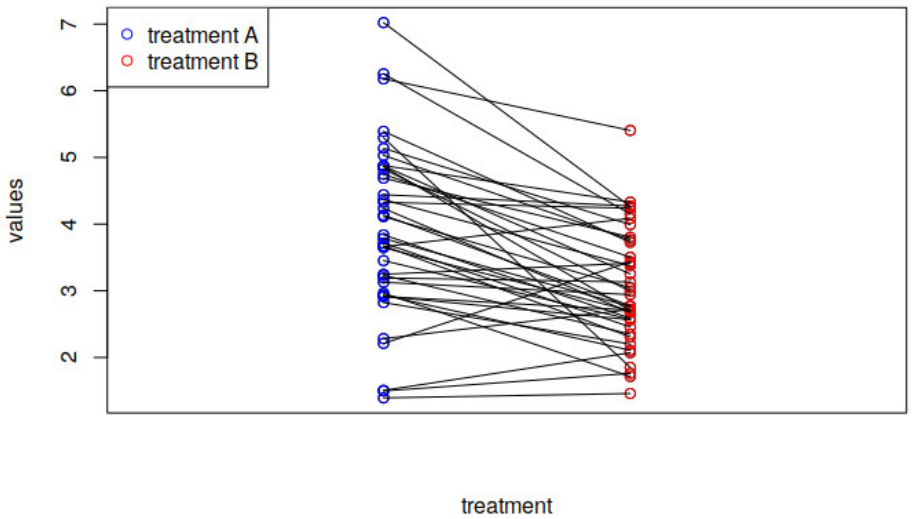
Bruns and Ioannidis (2016):

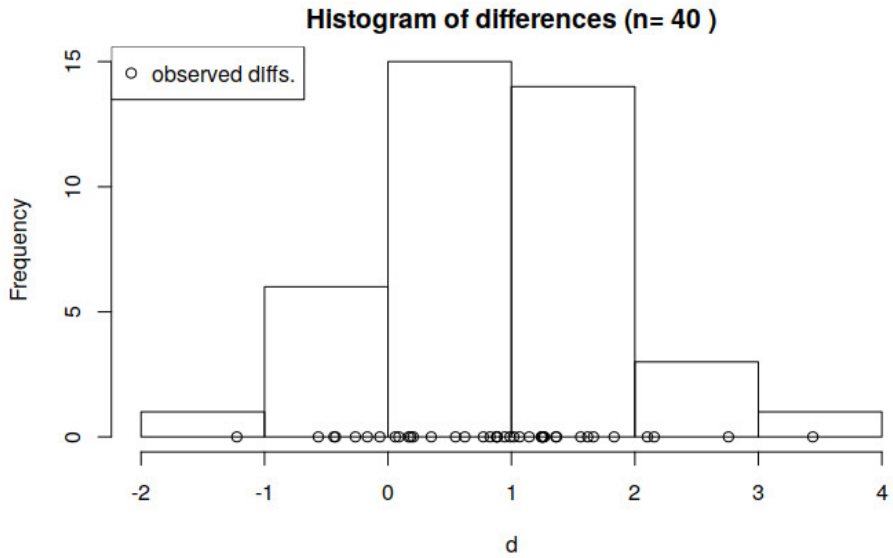
[...] “empirical surveys have documented an increased prevalence of **p-values** of 0.041–0.049 in the scientific literature over time, and the **spurious excess of statistically significant findings** in various types of both observational and experimental research that have been attributed mostly to bias.”

The replicability crisis:

- p-values in question(s)
- possible safeguards
  - laboratory notebooks
  - opening data
  - pre-registration
  - meta-analyses

p-values: what are they?





- A_i (resp. B_i) observation for item i and treatment A (resp. B).
- $D_i = A_i - B_i$
- Could I believe wrongly that treatment B is more efficient than treatment A?

A touch of statistical modelling

- Sample $D_i = A_i - B_i$ with size n (n =40 here)
- Now assume D_i's are independent, normally distributed $\mathcal{N}(m, \sigma^2)$
- Unknown true mean / expectation m and variance σ^2
- Observed (variable / random) **sample** mean $\bar{D}_n = \frac{1}{n}\sum_{i=1}^{n} D_i$ and variance $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(D_i - \bar{D}_n)^2$
- B is more efficient than A (on average...) if and only if $m > 0$. Here $\bar{D}_n = 0.87$ but this is no proof of improved efficiency (also, $S_n^2 = 0.84$).

Null hypothesis significance testing (NHST)

- Two hypotheses: the null H_0 ($m \leq 0$, B less efficient) and the alternative H_1 ($m > 0$). Applies to true unknown value m, nothing random here.
- Decision at the sight of data $(n, \bar{D}_n, S_n^2)$.
- Two kinds of errors (probabilisable):
  Type I, H_0 is true but we decide H_1 (drop a fine treatment in favour of a worse one).
  Type II, H_1 is true but we decide H_0 (do not benefit from opportunity of better treatment).
- Null hypothesis significance testing controls probability of type I error. You can choose it! (say, $0 < \alpha < 0.5$)

p-values (at last)

- Null hypothesis significance testing controls probability of type I error α. You can choose it! (say, $0 < \alpha < 0.5$)
- Decide $m > 0$ if $\bar{D}_n > 0$ plus a security margin that depends on (n, S_n^2, α).

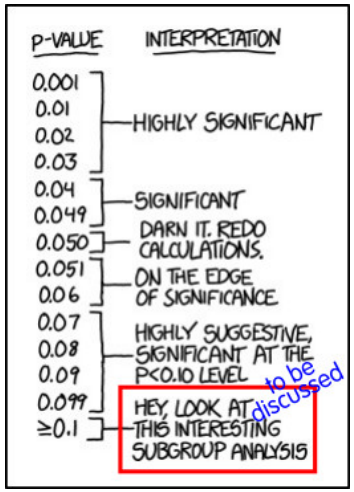
$$\text{rule: } (H_1) \quad \bar{D}_n > \ell(n, S_n^2, \alpha)$$

- Worst-case control: most unfavourable/difficult value of m (here $m = 0$).
- p-value: smallest α such that you would decide H_1 (universal statement).
- Seems reasonable but surrounded by controversies (since Fisher, 1956)
- Confidence intervals and equivalence (?) with NHST

From the Null Ritual to a ban of p-values

The Null Ritual (Gigerenzer *et al.*, 2004):

1. Set up a statistical null hypothesis of “no mean difference” or “zero correlation.” Don’t specify the predictions of your research hypothesis or of any alternative substantive hypotheses.
2. Use 0.05 as a convention for rejecting the null. If significant, accept your research hypothesis.
3. Always perform this procedure.

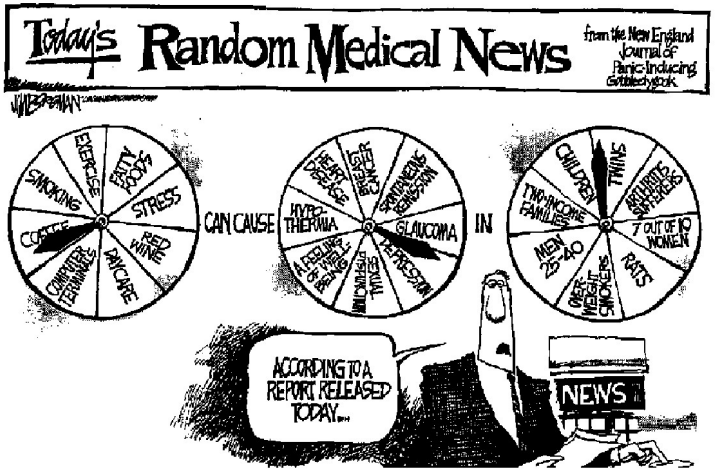


From Randall and Welser (2018)

- p-values and / or confidence intervals banned from several journals (*Basic and Applied Social Psychology*, *Political Analysis*, others).
- Did not stop the Null Ritual in some journals.

Multiple tests

- Now imagine that you have 100 candidate factors with potential effect. How many type I errors to expect?
- Corrections of level α (does it solve the problem?)
- P-hacking and HARKing (Hypothesis After Result is Known, e.g., $H_0 =$"$m < 0$")
- Effects of dimension, preprocessing... (fMRI)



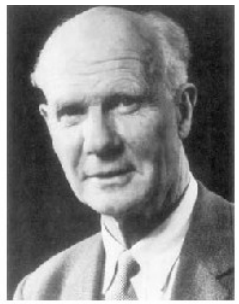
From Sterne and Smith (2001)

- Distinguish between exploratory vs. confirmatory studies.
- Think wide (how many people work in the field?)

The False Positive / Discovery Rates

- What if H_1 is true?
- What if H_0 and H_1 are both wrong?
- Assessing the frequency / probability of true H_1 among the ones you believe to be true.
- $P(H_1 \text{is true} | \text{test decided } H_1)$ and False Discovery Rate

Power

- Quantity of interest: $P(H_1 \text{ is true} | \text{test decided } H_1)$
- Related to power $\beta = P(\text{test decided } H_1 | H_1 \text{ is true})$, e.g.

$$P(\bar{D}_n > \ell(n, S_n^2, \alpha))$$

- Type II error probability $1 - \beta$
- β usually
  - increases with sample size n
  - decreases as α decreases.
  - depends on unknown true distribution / unknown parameter $m \to \beta(m)$
  - worst case really bad (consider $H_0 : m < 0$ and $H_1 : m \geq 0 \to \alpha$)

The False Discovery Rate: illustration

Power of study (proportion (%) of time we reject null hypothesis if it is false)	Percentage of "significant" results that are false positives		
	P=0.05	P=0.01	P=0.001
80% of ideas correct (null hypothesis false)			
20	5.9	1.2	0.10
50	2.4	0.5	0.05
80	1.5	0.3	0.03
50% of ideas correct (null hypothesis false)			
20	20.0	4.8	0.50
50	9.1	2.0	0.20
80	5.9	1.2	0.10
10% of ideas correct (null hypothesis false)			
20	69.2	31.0	4.30
50	47.4*	15.3	1.80
80	36.0	10.1	1.10
1% of ideas correct (null hypothesis false)			
20	96.1	83.2	33.10
50	90.8	66.4	16.50
80	86.1	55.3	11.00

*Corresponds to assumptions in table 2.

$P(H_1)$

Power $1-\beta$

P-value or level

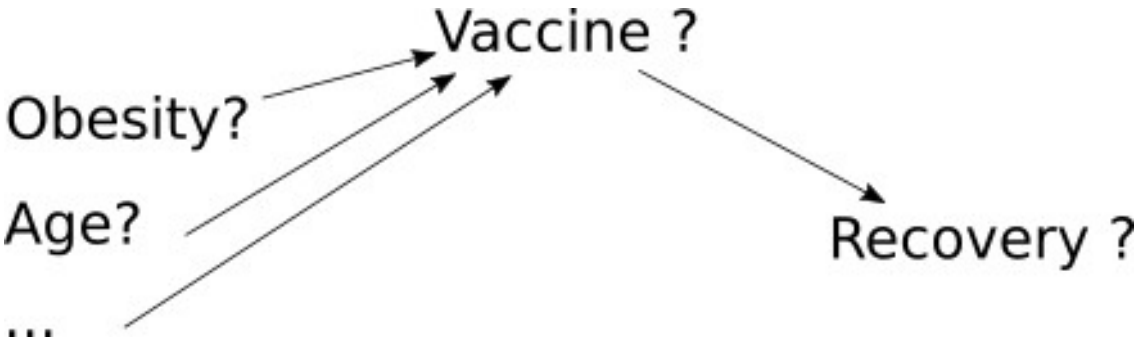
FDR

From Sterne and Smith (2001)

Breaking the Null Ritual:

- A priori formulation of hypothesis
- A priori analysis of power (target size of effect)
- Control false positive rate (e.g, Barber and Candès, 2015)

Causality and confounding



- Regression model

$$R_i = \nu_i V_i + \omega_i O_i + \alpha_i A_i + \ldots + \varepsilon_i$$

- No causal interpretation
- Adjustments (Pearl, 1998, 2000; Freeman, 2008 ; Steyerberg, 2019)
  - Known chain of measured causal factors
  - Known chain of latent causal factors
  - Ignored potential causal factors

Binary thinking

What is at stake, fundamentally? $\longrightarrow$ Binary thinking! (Born, 2019)

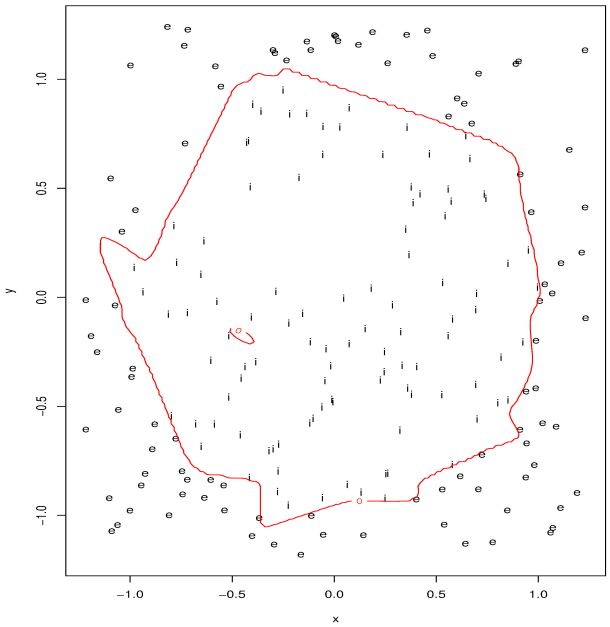
Safeguards against The Null Ritual:

- Confidence intervals (several of them): Gardner and Altman (1986)
- Investigate the alternative hypothesis / power (Colquhoun, 2019)
- Bayesian modelling (Sterne and Smith, 2001; Colquhoun, 2019)
- Replication (see also: meta-analyses)
- Model selection (same and however different ?)

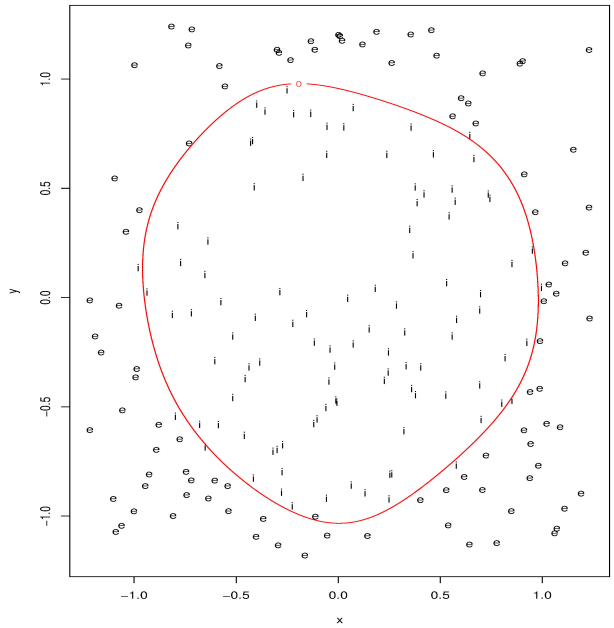
$$\min_{\theta \in \Theta} \int_{\mathcal{X}} \frac{\log P_X(x)}{\log q_\theta(x)} P_X(x) dx \approx \mathrm{IC}(\mathcal{M}(\Theta), x_1, \ldots, x_n)$$

Choose the model with best fit $\mathrm{IC}(\mathcal{M}_1(\Theta)) < \mathrm{IC}(\mathcal{M}_2(\Lambda))$.

The model selection bias



Error rate 0%
CV 10%



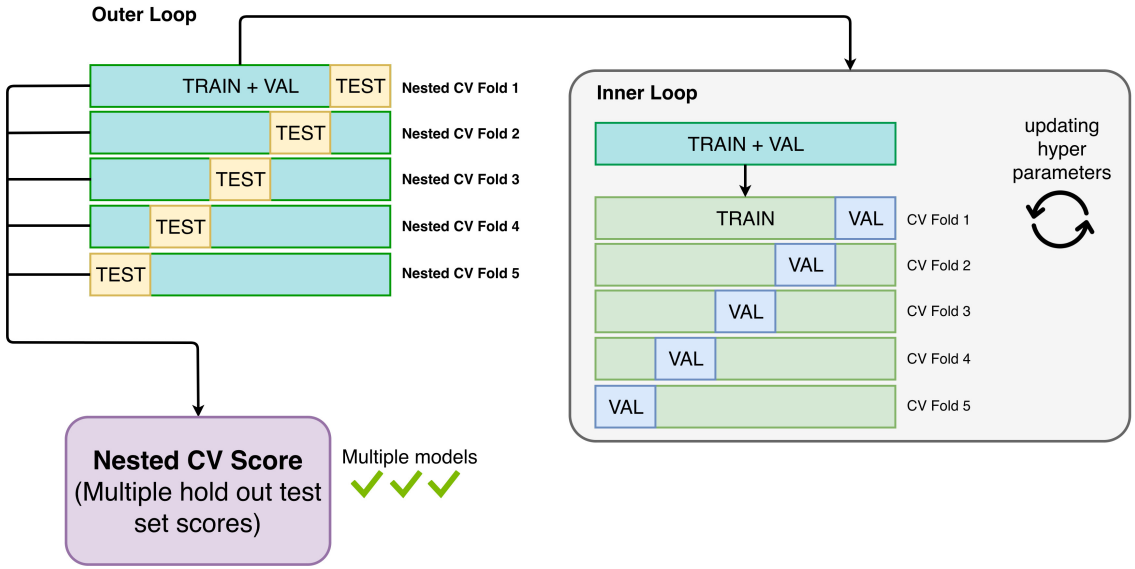
Error rate 2%
CV 3%

- Cross validation CV: leave part of observations for test vs. learning and select (regularization, etc.) parameter(s) on independent test data.
- Compare CV error rates of various classifiers (change NN architecture, SVMs, Random Forests...)
- CV error rate: expected CV on future (production) data?

The model selection bias

Credits: Computational materials science at Berkeley Lab,

`https://hackingmaterials.lbl.gov/`



- Cross validation CV: leave part of observations for test vs. learning and select (regularization, etc.) parameter(s) on independent test data.
- Compare CV error rates of various classifiers (change NN architecture, SVMs, Random Forests...)
- CV error rate: expected CV on future (production) data?
- No $\rightarrow$ Nested cross-validation

(laboratory) notebooks

- Make your own “laboratory notebooks”
  (keeping track of your choices / ideas)
- Share them with collaborators (e.g., students)
- Access others’ notebooks (data collection, preprocessing, ...)

Opening data

- 

  The example of Hidden Semi-Markov Models to Segment Reading Phases from Eye Movements

Opening data

- 

  The example of Hidden Semi-Markov Models to Segment Reading Phases from Eye Movements

- “Data of six participants were discarded because they did not follow the rules of the experiment thoroughly or data was too noisy during the acquisition with the eye tracker.”

Opening data

- 

  The example of Hidden Semi-Markov Models to Segment Reading Phases from Eye Movements

- “Data of six participants were discarded because they did not follow the rules of the experiment thoroughly or data was too noisy during the acquisition with the eye tracker.”

- Data curation / preprocessing is part of the analyses.
  $\rightarrow$ Automatized? Reproducible?

Opening data

- 

  The example of Hidden Semi-Markov Models to Segment Reading Phases from Eye Movements

- “Data of six participants were discarded because they did not follow the rules of the experiment thoroughly or data was too noisy during the acquisition with the eye tracker.”

- Data curation / preprocessing is part of the analyses.
  $\rightarrow$ Automatized? Reproducible?

- Raw data availability / format in relation to curation.

Opening data

- 

The example of Hidden Semi-Markov Models to Segment Reading Phases from Eye Movements
- “Data of six participants were discarded because they did not follow the rules of the experiment thoroughly or data was too noisy during the acquisition with the eye tracker.”
- Data curation / preprocessing is part of the analyses.
  $\rightarrow$ Automatized? Reproducible?
- Raw data availability / format in relation to curation.
- Open data repositories
  https://cat.opidor.fr/index.php/Entrepôt_de_données

Pre-registration against p-hacking (or worse)

- Documentation of future data collection and analyses.
- Outline theoretical framing, analysis code and all materials before starting a study.
- https://osf.io/registries?view_only=
- Example: Bakker *et al.* (2020)
- Templates on https://osf.io/q29nf/

Publication bias

"Imagine that we conduct a study where we **measure as many relevant variables as possible, 10 variables**, for example. We find only **two variables statistically significant**. Then, what should we do? We could decide to write a paper highlighting these two variables (and **not reporting the other eight** at all) as if we had hypotheses about the two significant variables in the first place. Subsequently, our paper **would be published**. Alternatively, we could write a paper including all **10 variables**. When the paper is reviewed, referees might tell us that there were **no significant results** if we had 'appropriately' employed **Bonferroni corrections**, so that our study would **not be advisable for publication**."

Nakagawa, S. A farewell to Bonferroni: the problems of low statistical power and publication bias (2004)

Systematic reviews and meta-analyses to improve robustness

- Systematic review: objective, reproducible method to find answers to a certain research question, by collecting all available studies related to that question and reviewing and analyzing their results (Ahn and Kang, 2018; Dols, 2017).
- Requirements:
  - Focused research question.
  - Preregistered protocols
  - Predefined inclusion and exclusion criteria of studies

Example: The study examined **human behavioral and/or cognitive responses to temperature** (e.g., climatic, ambient, or tactile-induced conditions) or temperature primes (e.g., visual or verbal evocations of warmth or coolness); (2) human's perceived or actual temperature (e.g., via skin or core (body) measurement) was a dependent variable; (3) the study was published in English as a journal article, preprint article, working paper, dissertation, book, or thesis; (4) the study was published in 2008 and later (from IJzerman, 2021).

Systematic reviews and meta-analyses to improve robustness

- Systematic review: objective, reproducible method to find answers to a certain research question, by collecting all available studies related to that question and reviewing and analyzing their results (Ahn and Kang, 2018; Dols, 2017).
- Requirements:
  - Focused research question.
  - Preregistered protocols
  - Predefined inclusion and exclusion criteria of studies
  - Scripted literature search and study selection (data bases)

Systematic reviews and meta-analyses to improve robustness

- Systematic review: objective, reproducible method to find answers to a certain research question, by collecting all available studies related to that question and reviewing and analyzing their results (Ahn and Kang, 2018; Dols, 2017).
- Requirements:
  - Focused research question.
  - Preregistered protocols
  - Predefined inclusion and exclusion criteria of studies
  - Scripted literature search and study selection (data bases)
- Often a narrative report (as opposed to new, pooled results).

Meta-analyses

- A meta-analysis differs from a systematic review in that it uses statistical methods on estimates from two or more different studies to form a pooled estimate (Ahn and Kang, 2018).
- To improve precision in estimating effects, enhanced power in tests...
- Statistical models (study s):

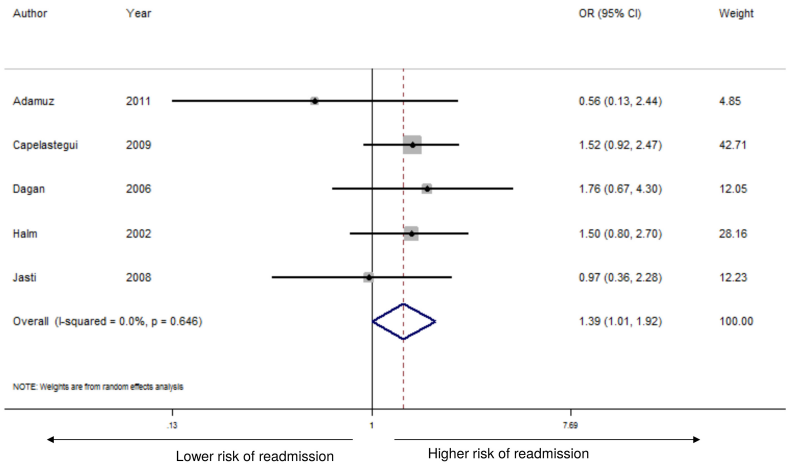
$$Y_{i,s} = \sum_k \beta_{k,s} X_{i,s} + \varepsilon_{i,s}; \qquad \varepsilon_{i,s} \sim \mathcal{N}(0, \sigma_s^2)$$

$$Y_{i,s} = \sum_k \beta_{k,s} X_{i,s} + \zeta_s + \varepsilon_{i,s}; \qquad \zeta_s \sim \mathcal{N}(0, \tau_s^2), \quad \varepsilon_{i,s} \sim \mathcal{N}(0, \sigma_s^2)$$

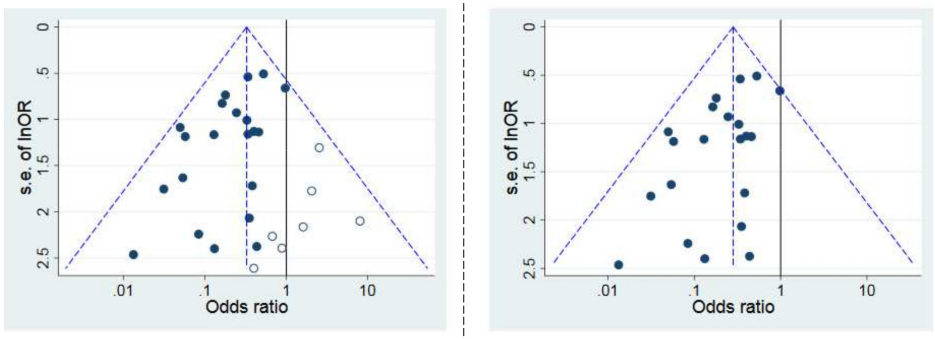
Pooled estimate $\hat{\beta}_k$ from $\hat{\beta}_{k,1}, \ldots, \hat{\beta}_{k,S}$

Meta-analyses and publication bias: funnel plots

Visual assessment of symmetries in statistics / distributions of p-values regarding precision, usually $\left(\frac{\sigma}{\sqrt{n}}\right)$.



Forest plots: confidence interval of pooled estimate



Funnel plots: estimates vs precision (illustrating potential publication bias)

Conclusions

- Combination of technical and methodological aspects.
- Each aspect is some rather huge research field.
- Assessment of required level of reproducibility / capacity of having a detailed level of documentation / code / etc.
- Statistical issues: distinguish between exploratory vs. confirmatory study.
- Identify and acknowledge weaknesses regarding reproducibility.


Ahn, E. and Kang, H.
Introduction to systematic review and meta-analysis.
Korean Journal of Anesthesiology **71**(2), 103–112 (2018).


Bakker, M., Veldkamp, C.L.S., van Assen, A.L.M., et al.
Ensuring the quality and specificity of preregistrations.
PLOS Biology **18**(12), e3000937 (2020).


Barber, R. F. and Candès, E. J.
Controlling the false discovery rate via knockoffs.
Annals of Statistics **43**(5), 2055–2085 (2015).


Born, R.T.
Banishing “Black/White Thinking”: A Trio of Teaching Tricks.
eNeuro **6**(6), ENEURO.0456-19.2019 (2019).


Bruns, S.B. and Ioannidis, J.P.A.
“*p*-Curve and *p*-Hacking in Observational Research”
PLoS ONE **11**(2), e0149144 (2016).


Colquhoun, D.
The False Positive Risk: A Proposal Concerning What to Do About p-Values
The American Statistician **73**(1), 192–201 (2019).


Davison, A.P.
Automated capture of experiment context for easier reproducibility in computational research.
Computing in Science and Engineering **14**, 48–56 (2012).


Dols, A.-M.
Revue systématique et méta-analyse.
Transparents de cours.
Méthodologie en recherche épidémiologique,
Université Grenoble Alpes (2017).


Dufour-Kowalski, S., Pradal, C., Dones, N., Barbier de Reuille, P. , Boudon, F., Chopard, J., Da Silva, D., Durand, J.-B., Theveny, F., Ferraro, P., Fournier, C., Guédon, Y., Smith, C., Stoma, S., Godin, C. and Sinoquet, H.
OpenAlea: An open-software plateform for the integration of heterogenous FSPM components.
In : *Proceedings of the Fifth International Workshop on Functional-Structural Plant Models (FSPM07)*.
Napier, New-Zealand (November 4-9 2007).


Freedman, D.A.
On regression adjustments to experimental data.
Advances in Applied Mathematics, **40**(2), 180–193 (2008).


Gardner, M.J. and Altman, D.G.
Confidence intervals rather than P values: estimation rather than hypothesis testing.
British Medical Journal, **292**(6522), 746–750 (1986).


Gigerenzer, G., Krauss, S. and Vitouch, O.
The Null Ritual. What You Always Wanted to Know About Significance Testing but Were Afraid to Ask.
In : D. Kaplan (Ed.) *The Sage handbook of quantitative methodology for the social sciences*, pp. 391–408.
Thousand Oaks, CA: Sage (2004).


IJzerman, H. Hadi, R., Coles, N. et al.
Social Thermoregulation: A Meta-Analysis
Preprint psyarxiv 10.31234/osf.io/fc6yq
https://psyarxiv.com/fc6yq/


Nakagawa, S.
A farewell to Bonferroni: the problems of low statistical power and publication bias
Behavioral Ecology, **15**(6), 1044–1045 (November 2004).


Pearl, J.
Why there is no statistical test for confounding, why many think there is, and why they are almost right.
Technical Report R-256.
Incorporated into Chapter 6 of Causality: Models, Reasoning, and Inference (January 1998).


Pearl, J.
Causality: Models, Reasoning, and Inference
Cambridge University Press. Cambridge, UK (2000).


Pradal, C., Fournier, C., Valduriez, P., Cohen-Boulakia, S.
OpenAlea: Scientific Workflows Combining Data Analysis and Simulation.
In : *Proceedings of the 27th International Conference on Scientific and Statistical Database Management (SSDBM '15).*
Association for Computing Machinery, New York, NY, USA, Article 11, 1–6 (2015).


Randall, D. and Welser, C.
The Irreproducibility Crisis of Modern Science: Causes, Consequences, and the Road to Reform.
National Association of Scholars. Princeton, NJ, USA (2018).


Sterne, J. A. and Smith, G. D. .
Sifting the evidence—what's wrong with significance tests?
Physical therapy **81**(8), 1464–1469 (2001).


Steyerberg E. W.
Clinical prediction models.
Springer International Publishing. Cham, Switzerland (2019).

Rejection region for Students' T-test

- Rejection region W: set of samples leading to decision H_1 (reject H_0)
- Reminder: H_0 is "$m \leq 0$" while H_1 is "$m > 0$", $D_i \sim \mathcal{N}(m, \sigma^2)$

$$W = \left\{ (d_1, \ldots, d_n) \in \mathbb{R}^n | \sqrt{n-1} \frac{\bar{D}_n}{S_n} > F^{-1}_{\mathcal{S}t_{n-1}}(1-\alpha) \right\}$$

where $F^{-1}_{\mathcal{S}t_{n-1}}(1-\alpha)$ is the quantile of order $1-\alpha$ of the Student distribution with parameter $n-1$ (so called degrees of freedom).

Confidence interval for parameter m

- A confidence interval with level $1-\alpha$ is a random interval that has probability $1-\alpha$ to contain m.
- α is thus the probability for the interval not to contain m (some sort of "error").

Symmetric confidence interval:

$$\left]\ \bar{D}_n - \frac{S_n}{\sqrt{n-1}} F^{-1}_{\mathcal{S}t_{n-1}}(1-\frac{\alpha}{2});\ \bar{D}_n + \frac{S_n}{\sqrt{n-1}} F^{-1}_{\mathcal{S}t_{n-1}}(1-\frac{\alpha}{2})\right[$$

where $F^{-1}_{\mathcal{S}t_{n-1}}(1-\alpha)$ is the quantile of order $1-\alpha$ of the Student distribution with parameter $n-1$ (so called degrees of freedom).

Asymmetric confidence interval associated with unilateral test H_0: "$m \leq 0$"; H_1: "$m > 0$"

$$\left]\ \bar{D}_n - \frac{S_n}{\sqrt{n-1}} F^{-1}_{\mathcal{S}t_{n-1}}(1-\alpha);\ +\infty\right[$$