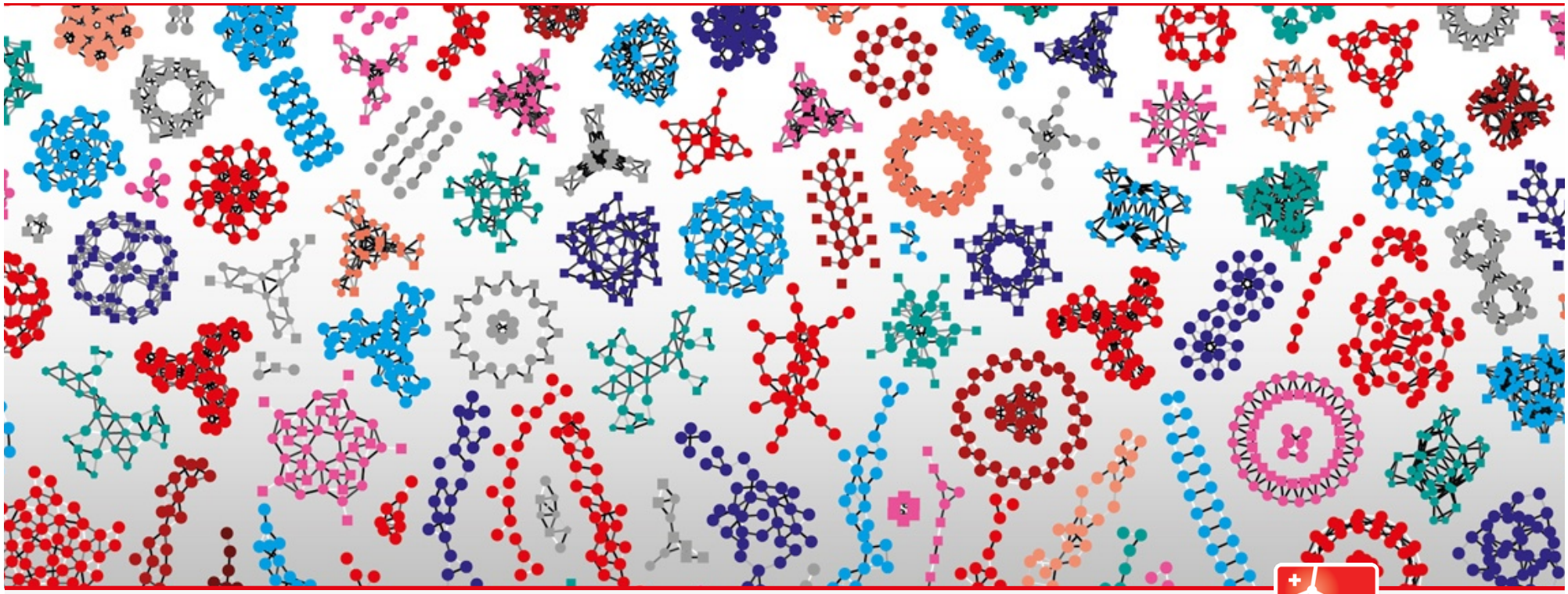


SIB
Swiss Institute of Bioinformatics

Data science: What it takes to reach reproducibility

Walid Gharib, UZH- 09.18

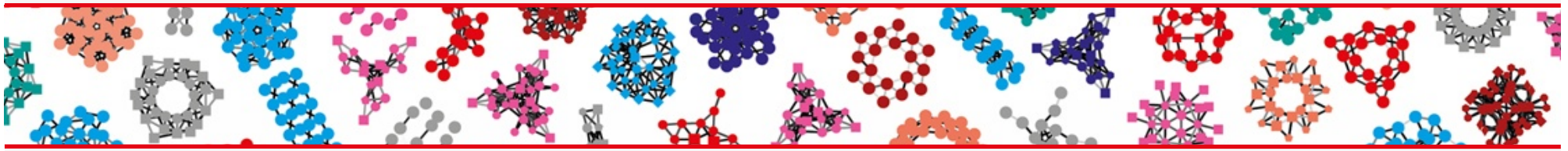
elixir SWITZERLAND

UNIVERSITÄT BERN

www.sib.swiss

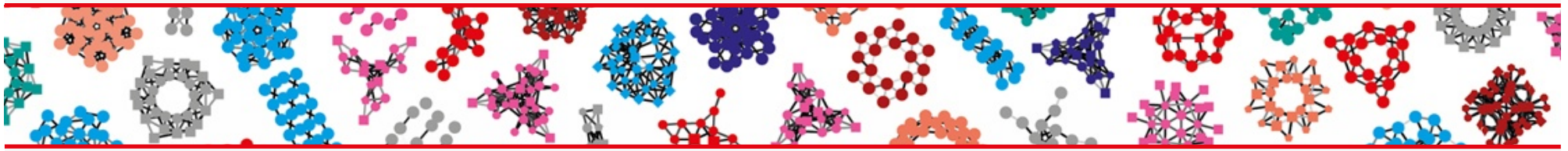
Content

- Part I: Definition
  - Redundancies
  - Reproducibility crisis
- Part II: Reproducibility in data science
  - Data availability
  - Code reproducibility
- PartIII: Best practices
  - 6 key points to reach data science reproducibility
  - Live example using containerisation
- PartIV: Training
  - A word on the role of training in reproducibility



PART I

Definition



PART I

Definitions

Reproducibility

Replicability is not reproducibility: nor is it good science (C Drummond – 2009)

- **Replicability**: is the ability of another person to produce the same results using the same tools and the same data
- **Reproducibility** involves more experimental variation

Reproducibility

- According to U.S. National Science Foundation (NSF) subcommittee on replicability in science:

> "**reproducibility** refers to the ability of a researcher to **duplicate the results of a prior study** using the **same materials** as were used by the original investigator. That is, a second researcher might use the **same raw data** to build the same analysis files and implement the same statistical analysis in **an attempt to yield the same results**.... Reproducibility is a minimum necessary condition for a finding to be believable and informative."

Replicability

- According to U.S. National Science Foundation (NSF) subcommittee on replicability in science:

> "**Replicability is** the ability of a researcher **to duplicate** the results of a prior study if the **same procedures** are followed but **new data** are collected."

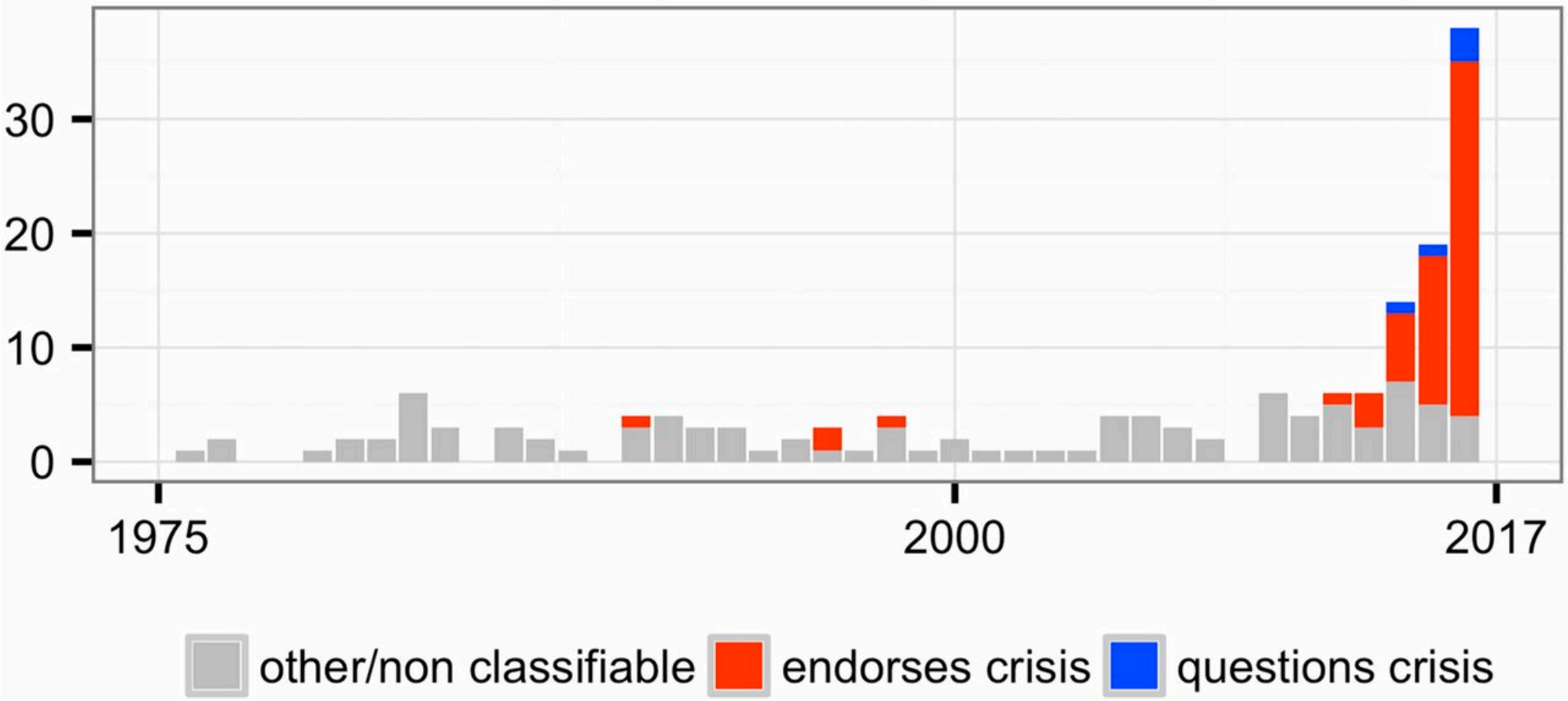
Reproducibility crisis?

Opinion: Is science really facing a reproducibility crisis, and do we need it to?

Daniele Fanelli

PNAS March 12, 2018. 201708272; published ahead of print March 12, 2018. https://doi.org/10.1073/pnas.1708272114

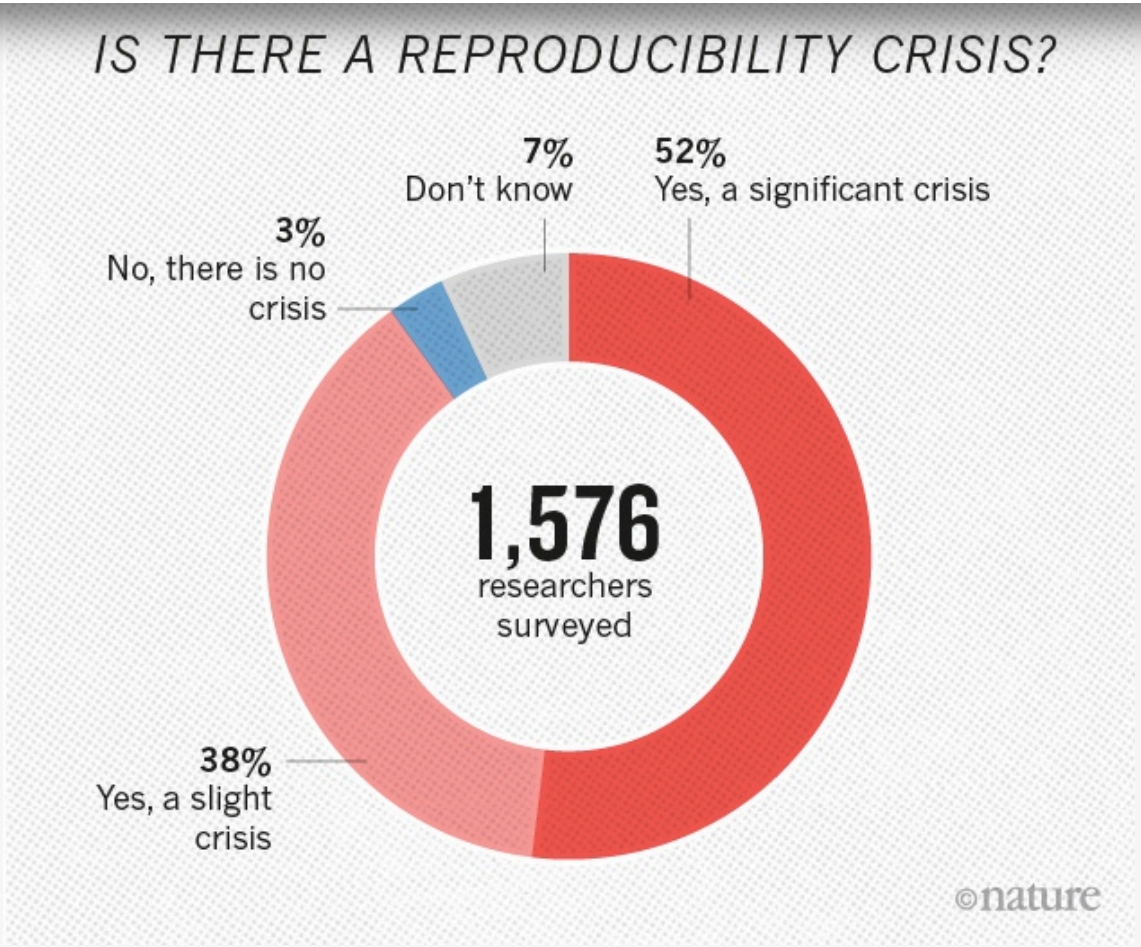
Frequency of Crisis Narrative in Web of Science Records



Reproducibility crisis?

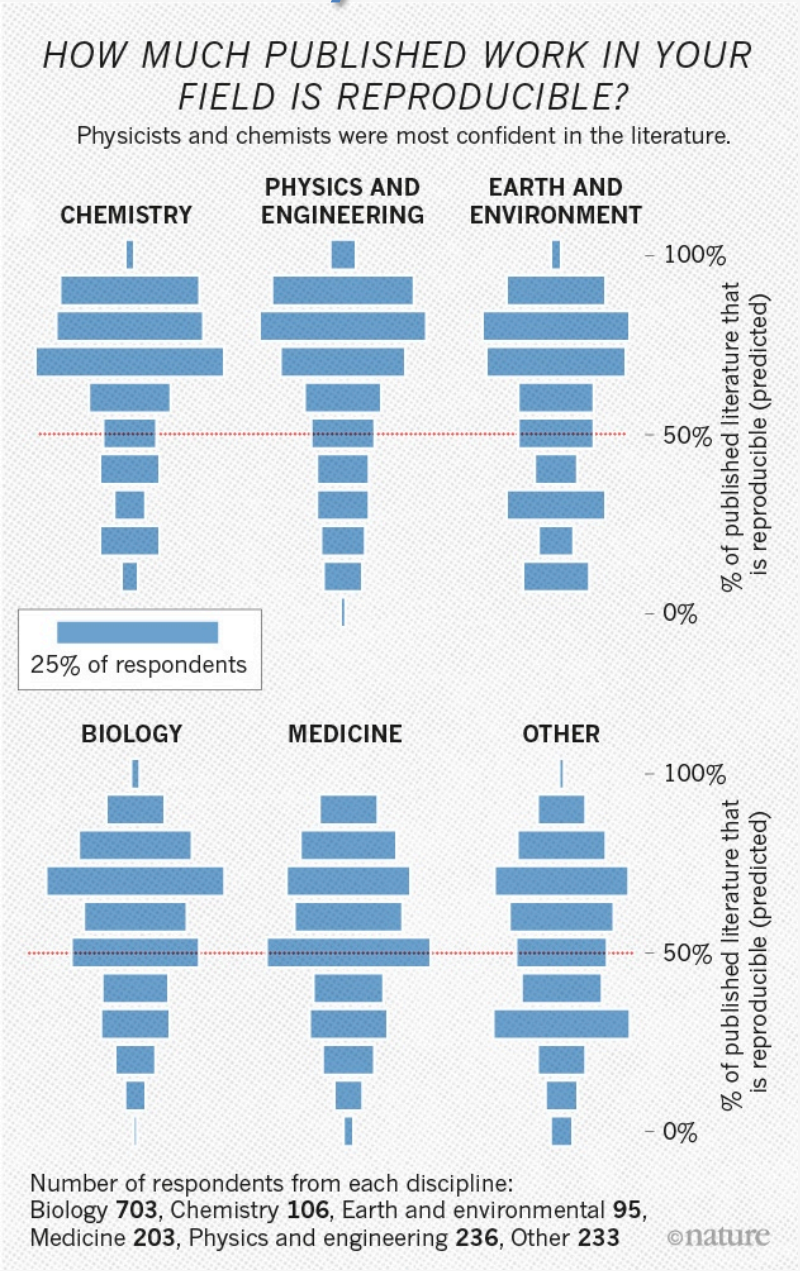
NATURE | NEWS FEATURE

1,500 scientists lift the lid on reproducibility



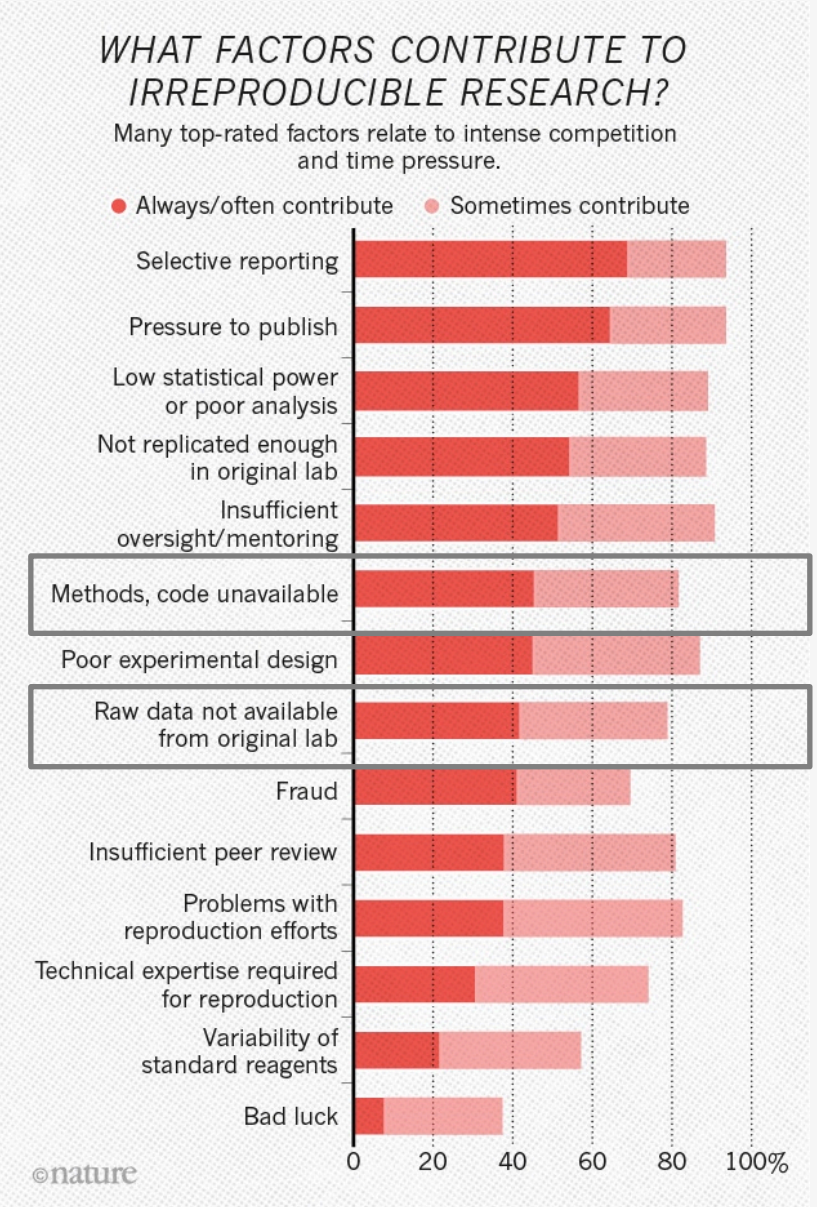
Baker M (2016) Is there a reproducibility crisis? Nature 533:452–454

Reproducibility crisis?



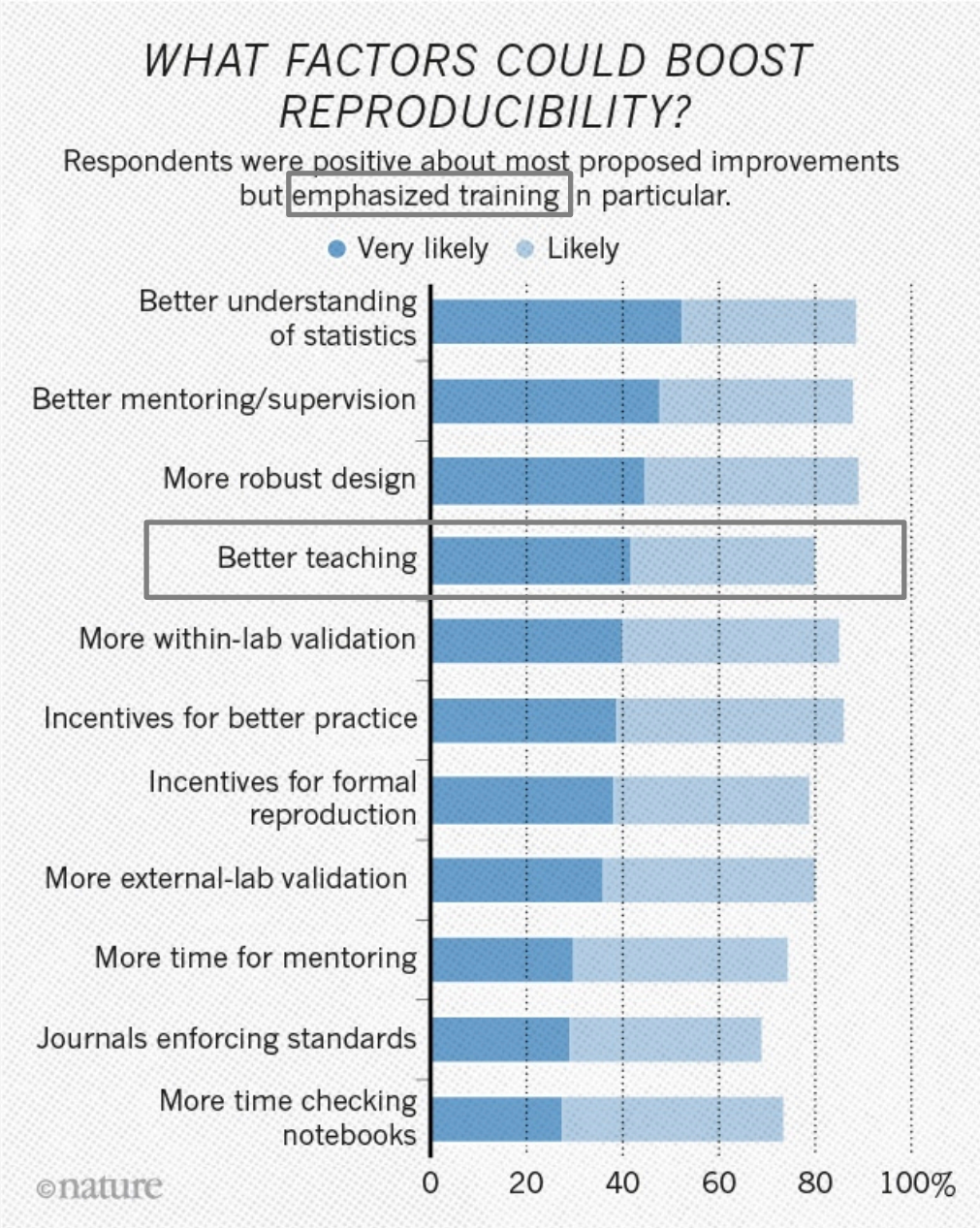
Baker M (2016) Is there a reproducibility crisis? Nature 533:452–454

Reproducibility crisis?



Baker M (2016) Is there a reproducibility crisis? Nature 533:452–454

Reproducibility crisis?



Baker M (2016) Is there a reproducibility crisis? Nature 533:452–454

Reproducibility crisis?

How Common Are Fabricated, False, Biased, and Irreproducible Findings?

Scientific misconduct and questionable research practices (QRP) occur at frequencies that, while nonnegligible, are relatively small and therefore unlikely to have a major impact on the literature. In anonymous surveys, on average 1–2% of scientists admit to having fabricated or falsified data at least once (2). Much higher percentages admit to other QRP, such as dropping data points based on a gut feeling or failing to publish a contradictory result. However, the percentage of scientific literature that is actually affected by these

Fanelli D (2018) Is science really facing a reproducibility crisis, and do we need it to? Proc Natl Acad Sci USA 115:2628–2631

Reproducibility crisis?

How Common Are Fabricated, False, Biased, and Irreproducible Findings?

The occurrence of questionable or flawed research and publication practices may be revealed by a high rate of false-positives and "*P*-hacked" (8) results. However, while these issues do appear to be more common than outright scientific misconduct, their impact on the reliability of the literature appears to be contained. Analyses based on the distribution of *P* values reported in the medical literature, for example, suggested a false-discovery rate of only 14% (9). A similar but broader analysis concluded that *P*-hacking was

Fanelli D (2018) Is science really facing a reproducibility crisis, and do we need it to? Proc Natl Acad Sci USA 115:2628–2631

Reproducibility crisis?

How Common Are Fabricated, False, Biased, and Irreproducible Findings?

that, of 18 studies, at least 11 had been successfully replicated (22). The largest reproducibility initiative to date suggested that in psychological science, reproducibility was below 50% (23). This latter estimate, however, is likely to be too pessimistic for at

Fanelli D (2018) Is science really facing a reproducibility crisis, and do we need it to? Proc Natl Acad Sci USA 115:2628–2631

Reproducibility crisis?

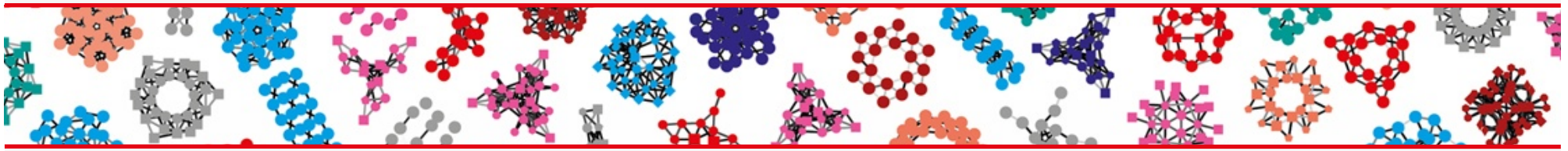
Do We Need the "Science in Crisis" Narrative to Promote Better Science?

The new "science is in crisis" narrative is not only empirically unsupported, but also quite obviously counterproductive. Instead of inspiring younger generations to do more and better science, it might foster in them cynicism and indifference. Instead of inviting greater respect for and investment in research, it risks discrediting the value of evidence and feeding antiscientific agendas.

Fanelli D (2018) Is science really facing a reproducibility crisis, and do we need it to? Proc Natl Acad Sci USA 115:2628–2631

Reproducibility & Replicability

- Deserves a standardized definition
- Reproducibility doesn't imply correctness
- Reproducibility is associated with transparency
- Replicability seeks the trueness of a claim
- Is there a "Crisis"?



PART II

Reproducibility in data science

Reproducibility in data science

1. Raw data availability
   - Metadata for large dataset and computationally intensive intermediate results

2. Code reproducibility
   - Code/scripts/software/versioning
   - Computing environment
     - Third party software installations
     - System dependencies

1.Raw data availability

- 36 articles in APA* journals out of 141 articles sent their datasets within 6 months

Wicherts et al. 2006

- Follow up analysis, reported publications with statistical inconsistencies were unlikely to share data highlighting the importance of mandatory data archiving

Wicherts et al. 2011

*American **Psychological** Association

1.Raw data availability

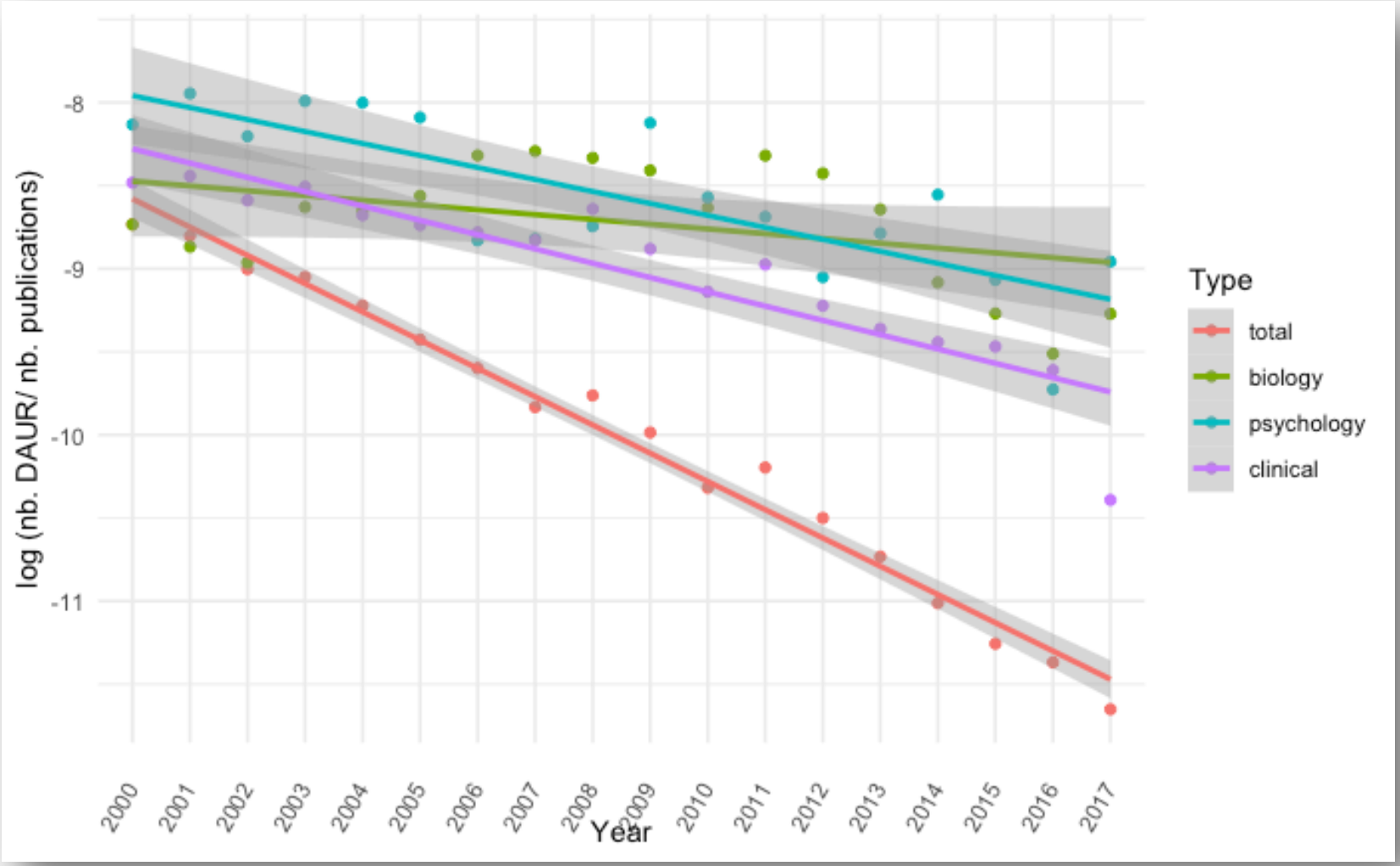
"data available upon request"

- 200 emails to corresponding authors in Buisness & economics related fields;
  - 128 replied;
  - 88 delivered the data with the next 30 days

(Un) available upon request, M Krawczyk, E Reuben - Accountability in research, 2012 - Taylor & Francis

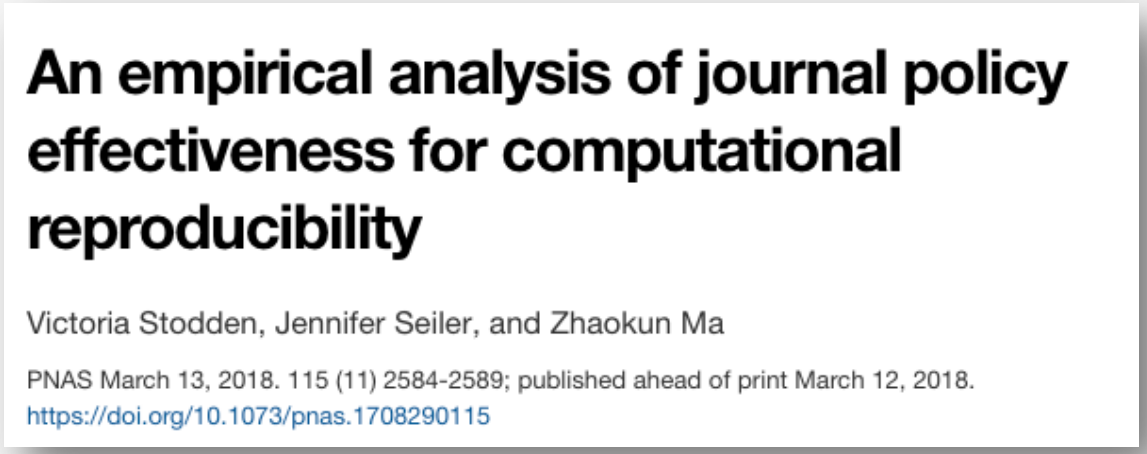
1.Raw data availability

Nb. occurrence of "Data available upon request" 2000-2017
Nb. DAUR/Nb. publications 2000-2017



https://app.dimensions.ai/browse/publication/for?redirect_path=/discover/publication

1.Raw data availability



An empirical analysis of journal policy effectiveness for computational reproducibility

Victoria Stodden, Jennifer Seiler, and Zhaokun Ma

PNAS March 13, 2018. 115 (11) 2584-2589; published ahead of print March 12, 2018.
https://doi.org/10.1073/pnas.1708290115

- 204 papers published in Science journal, after data policy implementation

Victoria Stodden et al. PNAS March 13, 2018

1.Raw data availability

- 204 computational papers published in Science journal, after data policy implementation [2011-2012]
  - 24 had sufficient information to locate the data without contacting the authors
  - 180 emails to corresponding authors

Victoria Stodden et al. PNAS March 13, 2018

1.Raw data availability

- 180 emails to corresponding authors

Type of response	Count	Percent, %
Did not share data or code:		
Contact another person	**20**	**11**
Asked for reasons	**20**	**11**
Refusal to share	**12**	**7**
Directed back to supplement	**6**	**3**
Unfulfilled promise to follow up	**5**	**3**
Impossible to share	**3**	**2**
Shared data and code	**65**	**36**
Email bounced	**3**	**2**
No response	**46**	**26**

36% shared "some" data and code

Victoria Stodden et al. PNAS March 13, 2018

1.Raw data availability

- Before and after the policy

Disclosure practice	2009–2010,% n(214)	2011–2012,% n(204)
Citations to data and/or code in references	25	29
Data location given in acknowledgements	**29**	**48**
Code location given in acknowledgements	4	5

Victoria Stodden et al. PNAS March 13, 2018

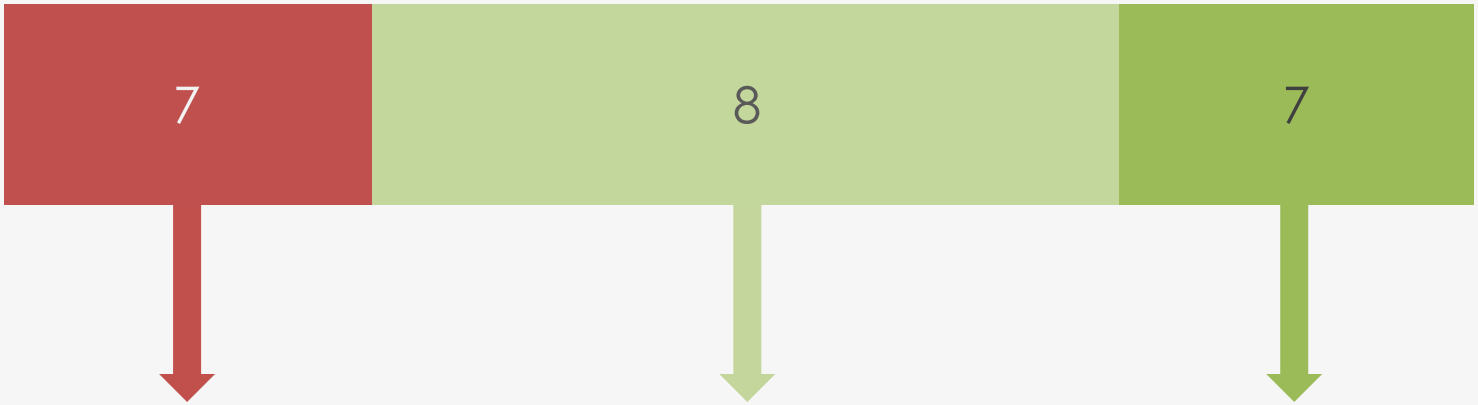
2. Code reproducibility

1. Code/scripts/software

2. Computing environment
   - Third party software installations
   - System dependencies

2. Code reproducibility

- 22 randomly chosen out of 56 "judged reproducible" articles out of 204

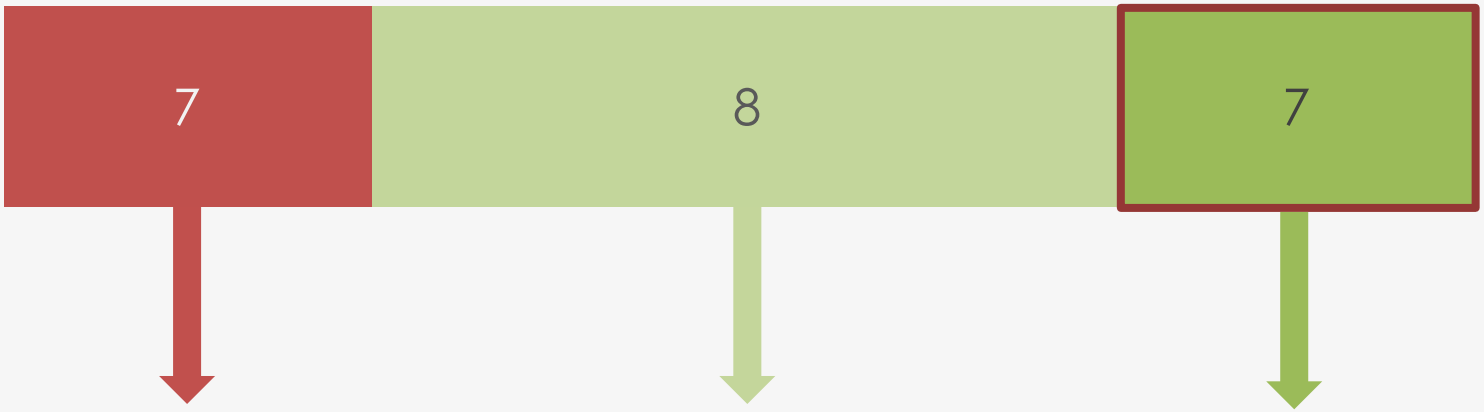


- Impossible or difficult to reproduce
  - Missing essential code/data/methods
  - inherent complexity (computationally intensive months of computations)
- Reproducible with effort
  - hand coding of data into a new format
  - Pseudo code with functions not detailed described in text
  - missing arguments required for differing architecture
  - required GPU programing abilities
- Reproducible with minor effort
  - installing a specialized library

Victoria Stodden et al. PNAS March 13, 2018

2. Code reproducibility

- 22 randomly chosen out of 56 “judged reproducible” articles out of 204

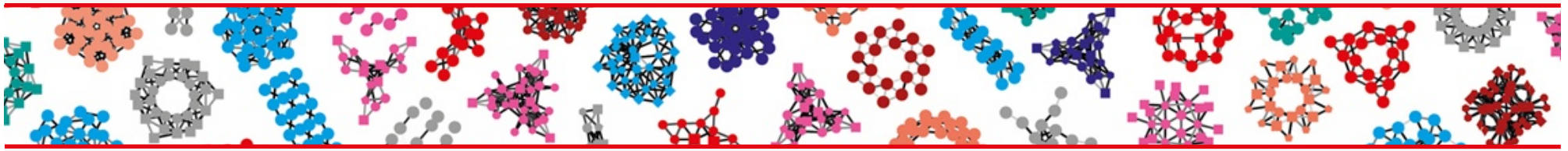


- Impossible or difficult to reproduce
  - Missing essential code/data/methods
  - inherent complexity (computationally intensive months of computations)
- Reproducible with effort
  - hand coding of data into a new format
  - Pseudo code with functions not detailed described in text
  - missing arguments required for differing architecture
  - required GPU programing abilities
- Reproducible with minor effort
  - installing a specialized library

Victoria Stodden et al. PNAS March 13, 2018

Reproducibility in DS: summary

- Data availability should be an obligation in the majority of studies
  - Deposited in scientific repositories (preferably public) e.g. similar to Sequencing read archive (for HTS data)
  - Exceptions for sensitive dataset e.g. human subject data, but still one can introduce "quasireproducibility"
- We are far from reaching the minimal standards of computational reproducibility
  - <10% of the computational studies can be reproduced with minimal effort even after journal policies changes
- Best practices in data science reproducibility should be standardized



PART III

Best practices

Best practices

1. Start by being reproducible to yourself
   - You should be able to precisely remember what you've done 6 months ago
   - Organize your harddrive or cluster project directory in folders and subfolders for each analysis
   - Separate raw data from metadata
   - Write ReadMe files to explain what is what as if you were sending the materials to collaborators (put yourself in their shoes)
   - Keep track of any external software version

Best practices

2. Code everything and avoid hand manipulations
   - Hand manipulations are easy to loose trace (personal experience)
   - The most basic reproducibility principle is to do everything via code

Best practices

3. Automate the whole process

- Combine your scripts to run one after the other (pipeline)
- Requires some skills e.g. GNU make* or Snakemake** (python)

* https://www.gnu.org/software/make/
** https://snakemake.readthedocs.io/en/stable/

Best practices

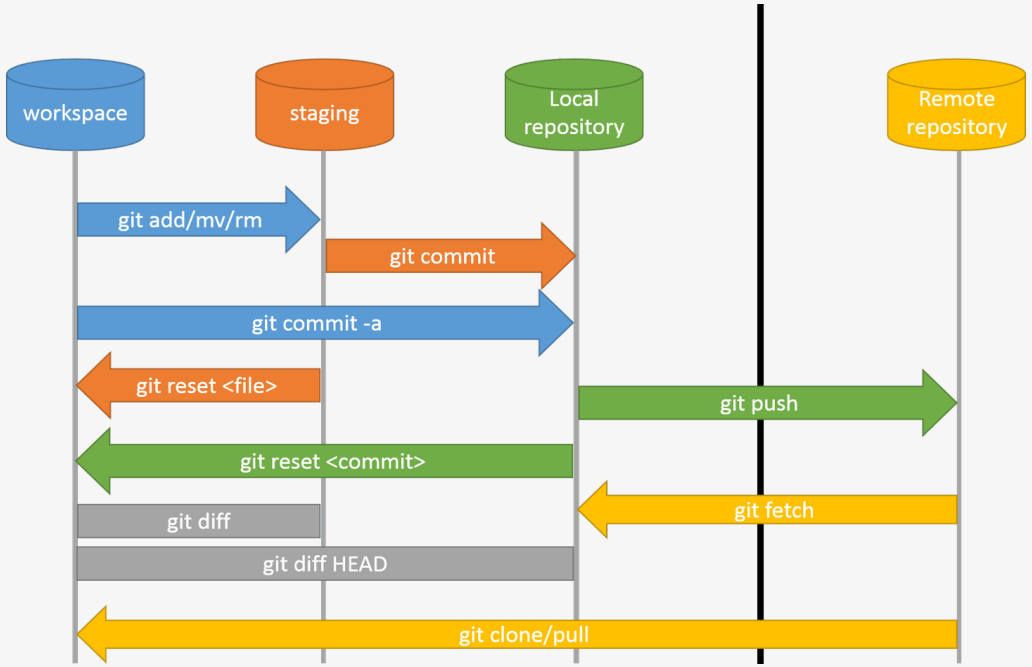
4. Generate reports from code
   - Using Markdown: lightweight markup language with plain text formatting syntax
   - I personally use R Markdown* (data analysis and teaching)
   - Python users can use python notebooks e.g. Jupyter Notebooks** (managed anaconda)

* https://rmarkdown.rstudio.com/
** http://jupyter.org/

Best practices

5. Code maintenance

- Complex repeated code should be converted to functions
- Why not creating modules or packages (R packages*) along the way
- Use version control e.g. git**



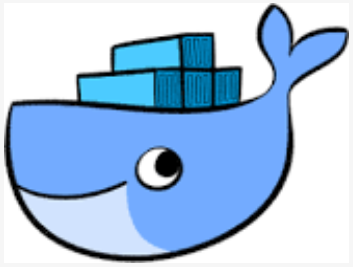
* http://kbroman.org/pkg_primer/
** https://git-scm.com/

Best practices

6.Pack the analysis in a reproducible environment

- Containerisation

Docker



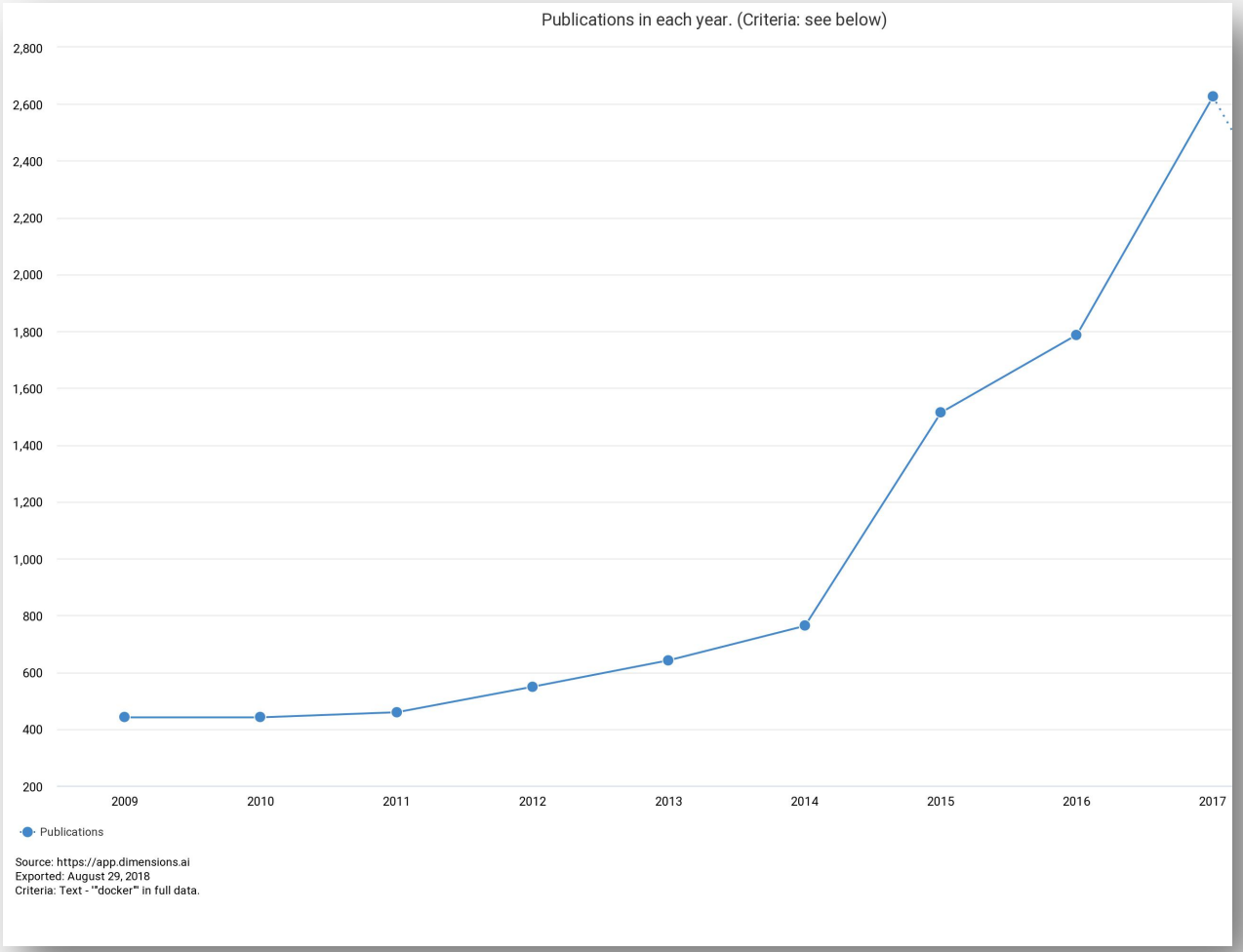
Singularity

Best practices: summary

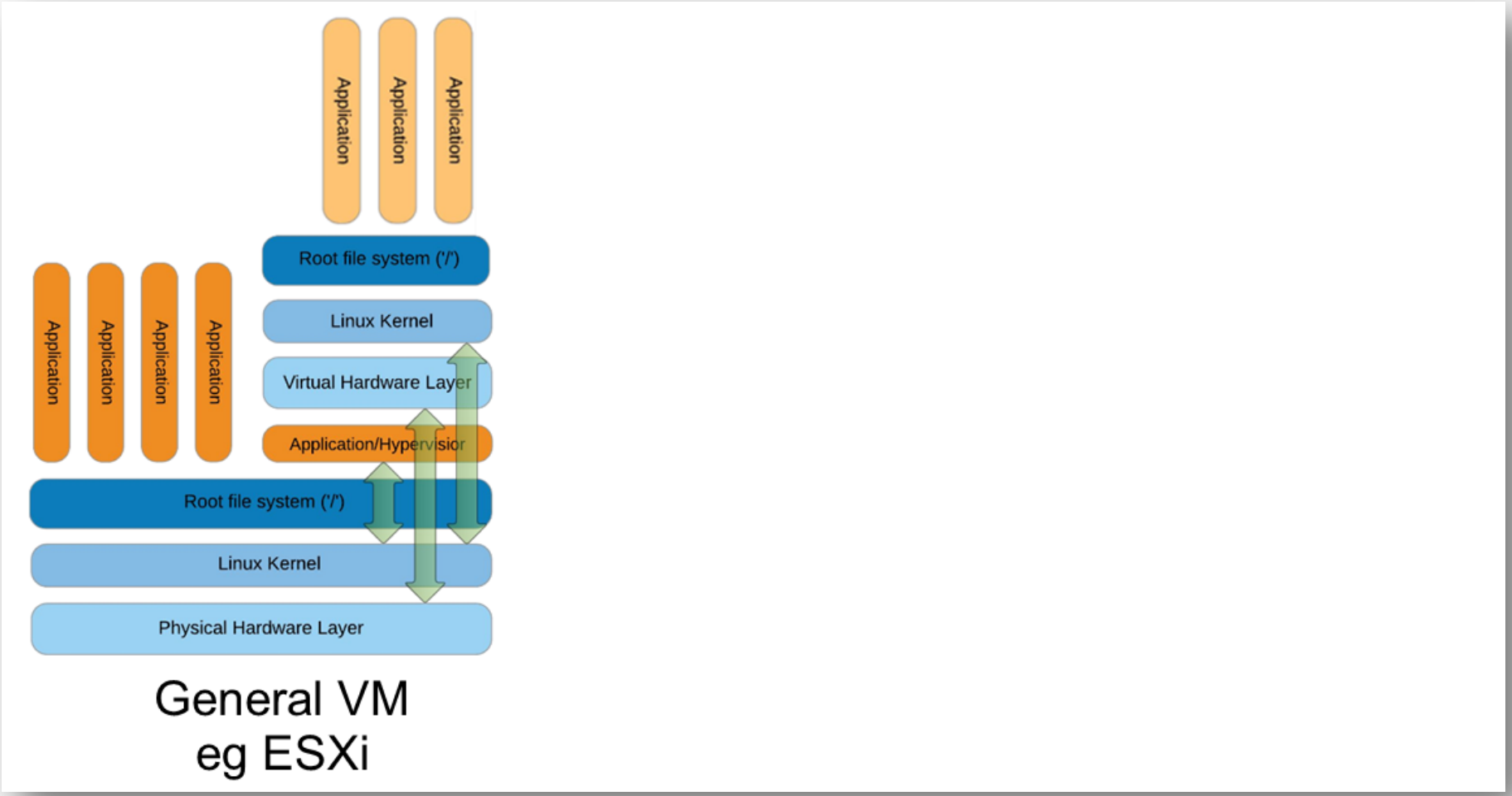
1. Start by being reproducible to yourself
2. Code everything and avoid hand manipulations
3. Automate the whole process
4. Generate reports from code
5. Code maintenance
6. Pack the analysis in a reproducible environment

Containerisation

Usage of “Docker” in publications

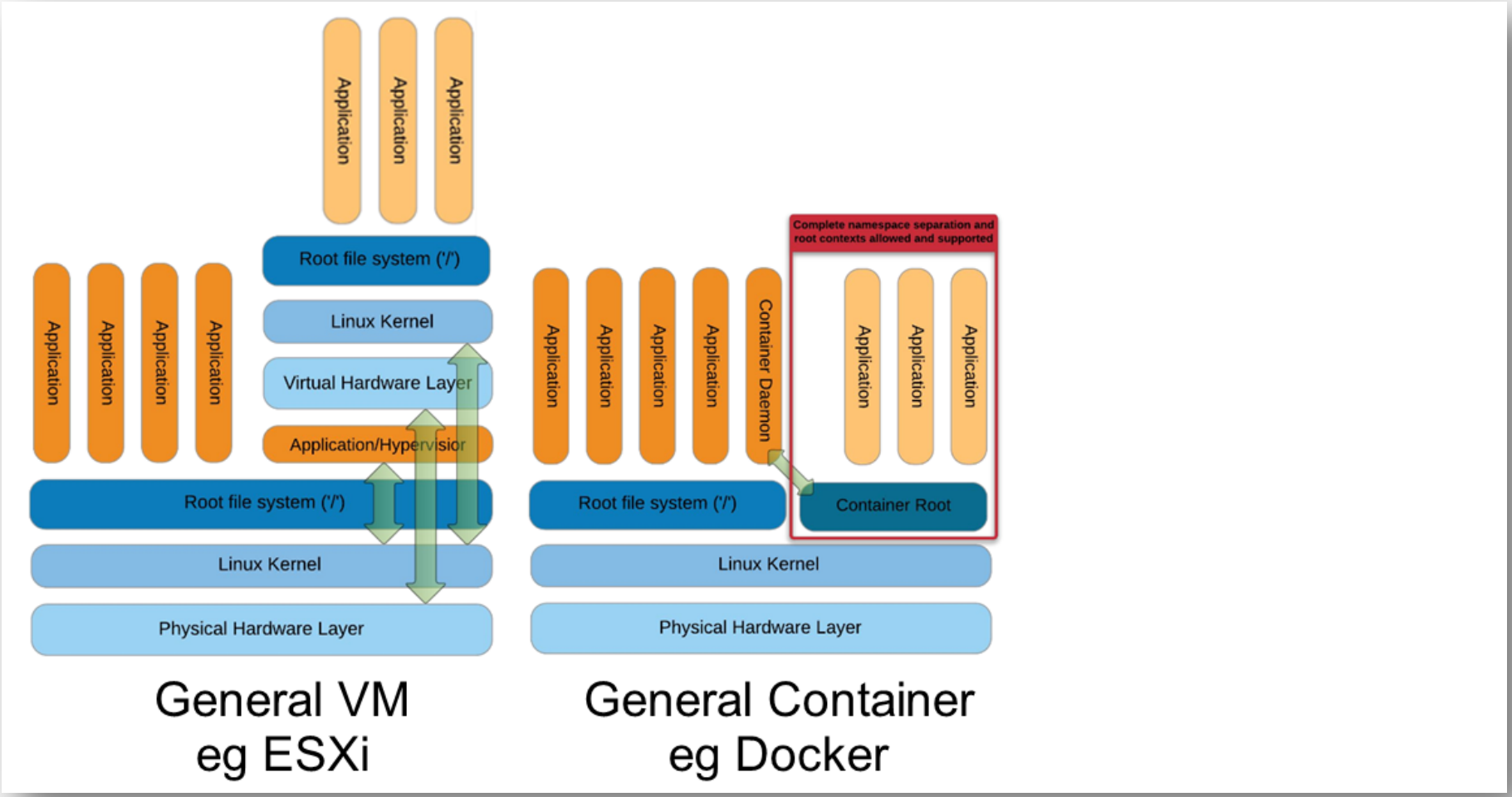


Containerisation



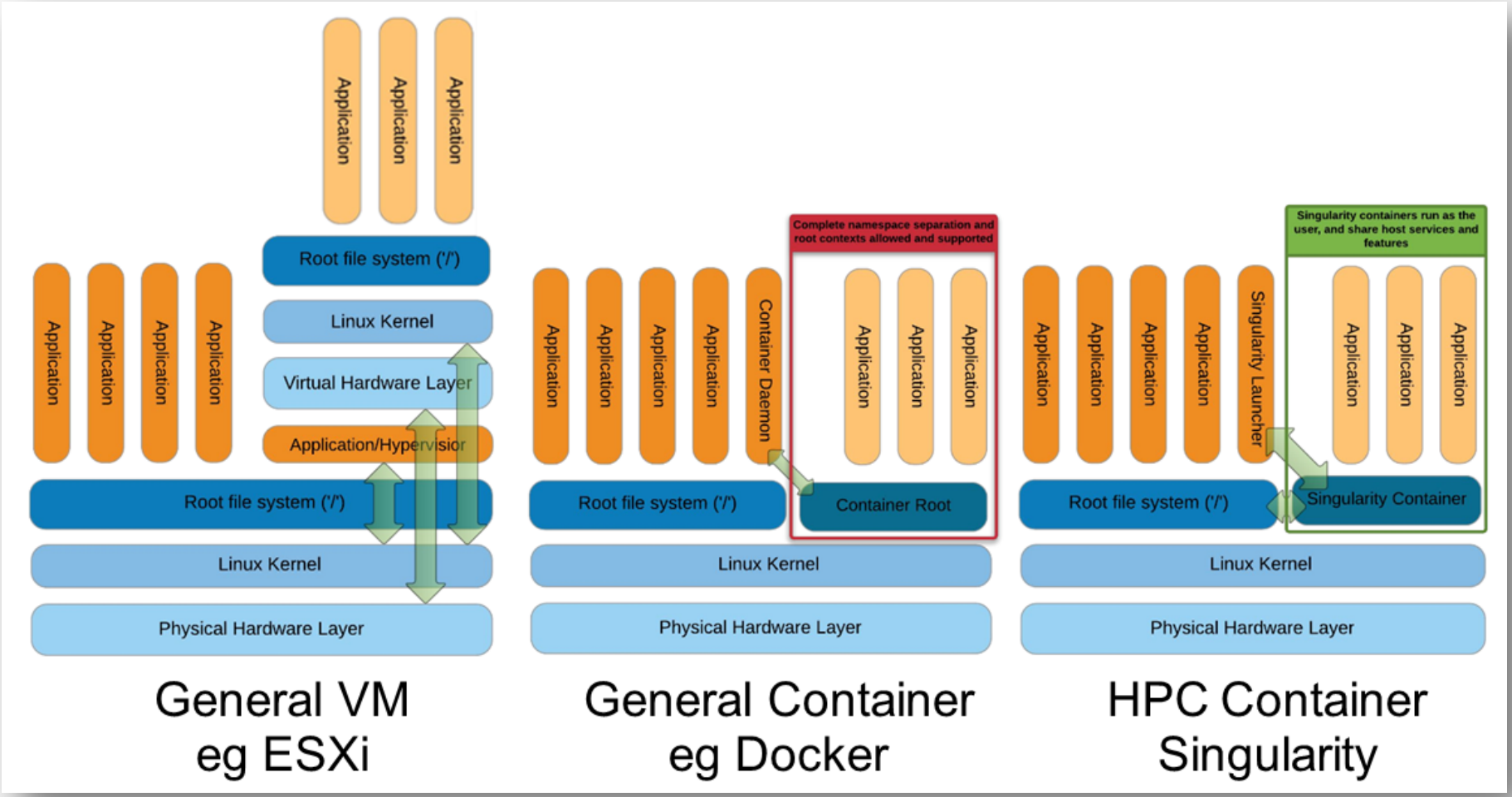
Source: Greg Kurtzer keynote at HPC Advisory Council 2017 @ Stanford

Containerisation



Source: Greg Kurtzer keynote at HPC Advisory Council 2017 @ Stanford

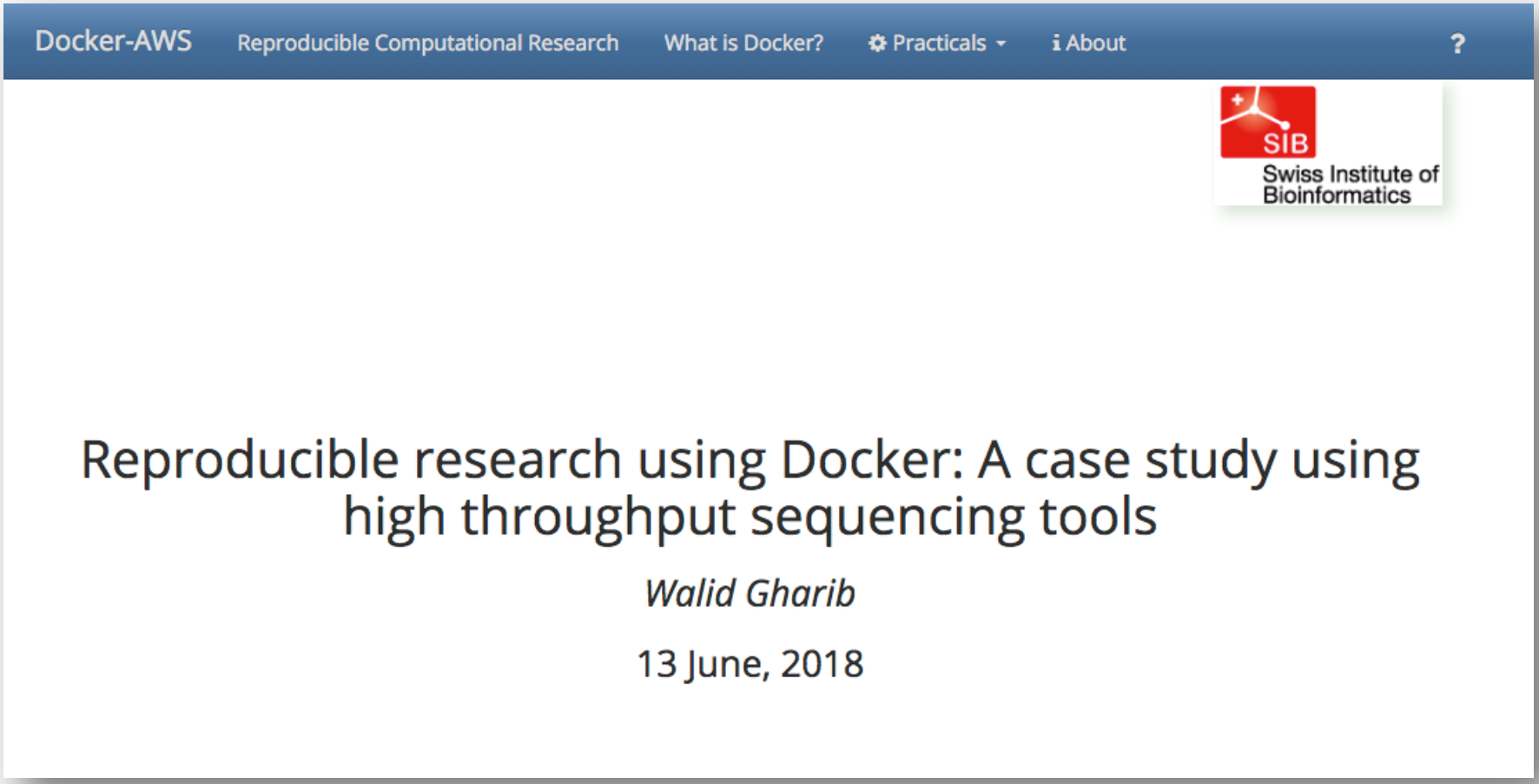
Containerisation



Source: Greg Kurtzer keynote at HPC Advisory Council 2017 @ Stanford

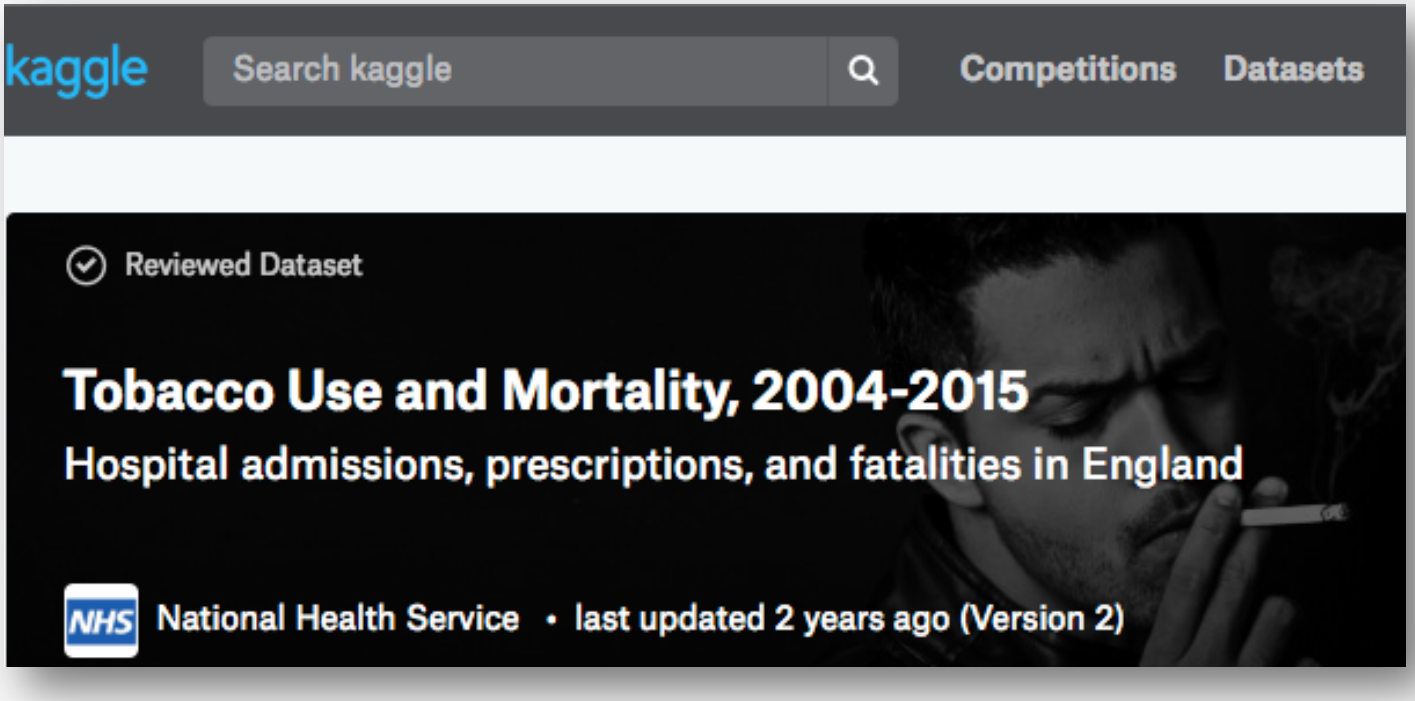
Containerisation

http://training.bioinformatics.unibe.ch/Docker2018



Now it is your turn!!

Dataset:

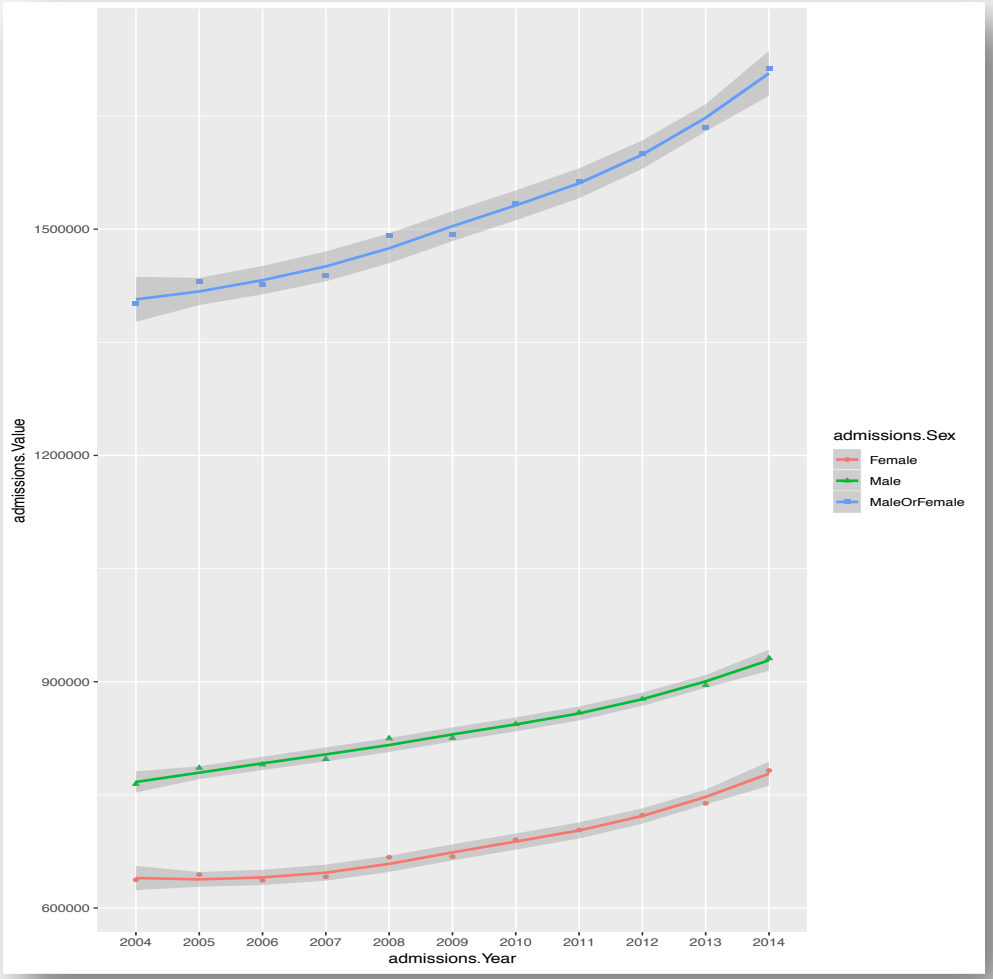


For this example, We will only use the admissions.csv dataset

https://www.kaggle.com/nhs/tobacco-use/

Now it is your turn!!

We will try to reproduce this image in the next 5-10 min using containers technology:



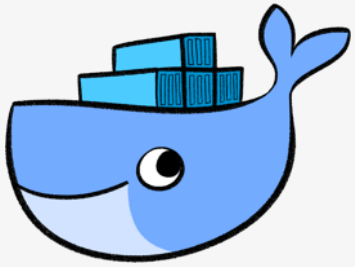
Now it is your turn!!

1. Who has a laptop (only browser is needed) with an internet connection
2. Numbered Sticky notes
3. Connect to the following URL:
   http://training.bioinformatics.unibe.ch/reprozurich
4. Locate the same number in the webpage as is on your Sticky note
5. Click on the Rstudio link corresponding to the Sticky note number
6. You will prompt for username and password: both are "training"
7. You're in ?
8. Let's do it together…

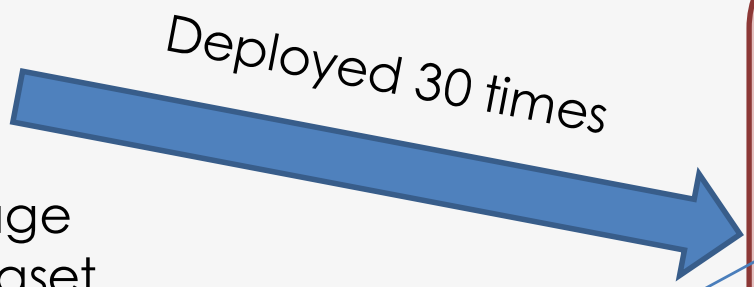
Now it is your turn!!

How many were able to reproduce the figure by executing the code?

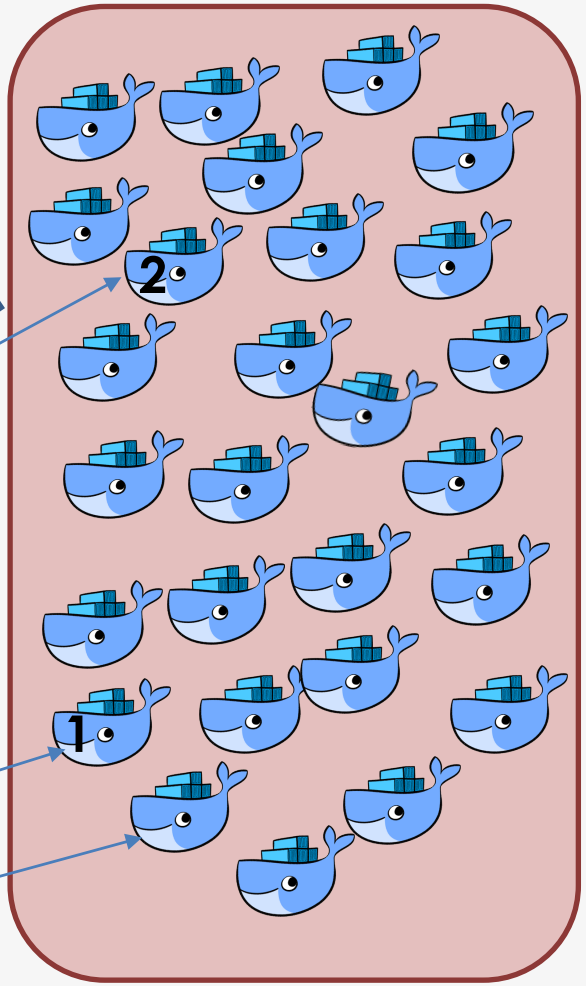
Technically

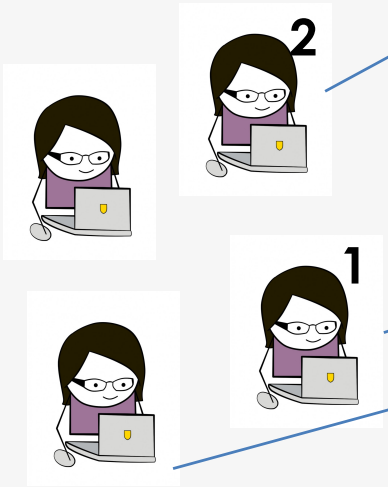


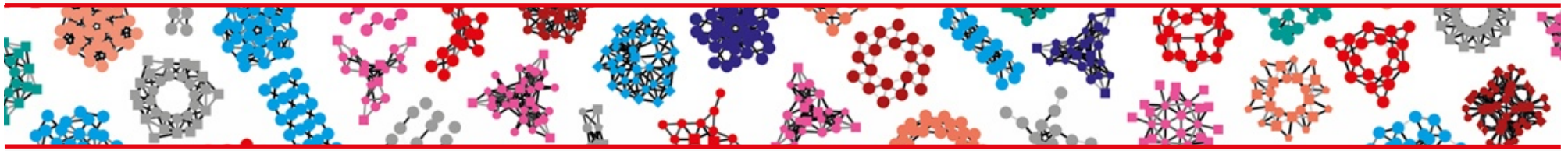
Rstudio docker image containing the dataset & code & dependencies



Cloud/server







PART IV

Training

Training

- Training is indispensable to reach reproducibility in data science
- For best practices, the learning curve isn't as steep as you think
- We are lacking courses in this area as a part of a continuous education for scientists
- The Training group at Swiss institute of Bioinformatics* (SIB) have a panel of related courses
  - Statistics for life scientists - R courses – python – Docker – Best practices in programming

https://www.sib.swiss/training

Feedbacks:

Usage of Rmarkdown & containers in genomics courses

"the informatics infrastructures that were made available to us were extremely useful and worked perfectly: the website, the etherpad, the terminal accessible through docker, etc."

"Practical, one goes out of it with tools to use"

"The docker intro was very good and extremely helpful!"

"Platform for exercices is great, website also, support for course excellent"

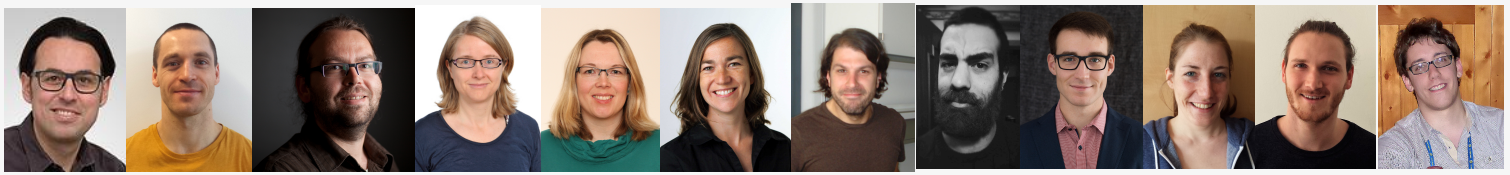
https://www.sib.swiss/training

ACKNOWLEDGEMENT

SIB training Group

Geoffrey Fucile, Patricia Palagi, Frédéric Schütz, Walid Gharib, Diana Marek and Grégoire Rossier

Interfaculty Bioinformatics Unit



Rémy Bruggmann, Kurt Wyler, Pierre Berthier, Irene Keller, Heidi Tschanz-Lischer, Simone Oberhänsli, Stephan Peischl, David Miguel Francisco Ferreira, Thomas Roder, Nicole Liechti, Lorenz Ryser, Matteo Tomasini