Computational Reproducibility

DACH Epidemiologietagung
1. - 3. September 2021, Bern

Prof. Dr. Reinhard Furrer

Department of Mathematics
Department of Computational Science

Center for Reproducible Science

@ReinhardFurrer

Content

- General comments about dangers of reproduciblity
- Pre-registration, reporting guidelines and other helpers
- Common tools of computational reproducibility
- Discussion

Reproducibility vs Replicability

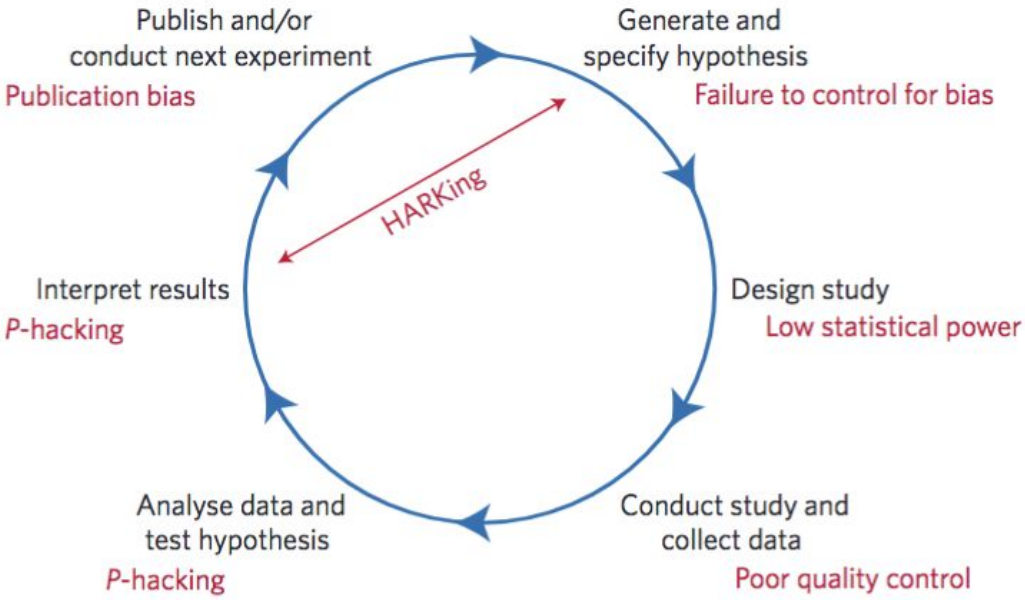
Reproducibility refers to the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator.

This requires, at minimum, the sharing of data sets, relevant metadata, analytical code, and related software.

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

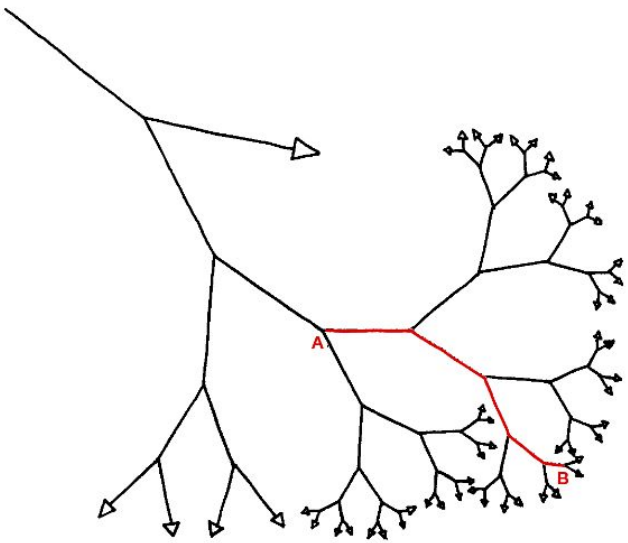
Goodman et al., 2016, *Sci Transl Med*

The scientific method and threats to reproducibility



Munafò et al., *Nat. Hum Behav*, 2017

Garden of forking paths



- Exclude some cases?
- Transform variables?
- Test different statistical models?
- Add or remove covariates?

Consequence:

- Essentially a **multiple testing problem**
- Inflated **false positive** rate

Gelman and Loken, 2013, unpublished

Publication bias

- Significant results have a higher likelihood of getting published, non-significant results remain often unpublished

- Publication bias is a major concern in research. For example, in medical research the bias can mislead clinical practice and harm patients

(Pre)- Registration

Time-stamped read-only **research protocol** created before the study containing as a minimum:

- Hypotheses
- Description of population, inclusion/exclusion criteria, sample size
- Data collection procedure or database used
- General design
- Variables (primary vs. secondary, explanatory vs. dependent variables, raw vs manipulated variables)
- Specify exactly how the key confirmatory analyses will be conducted under all probable scenarios

Why should you (pre)-register your study?

- Add credibility to your research
- Set a time-stamped record of your ideas
- Lets you think more deeply about your research and planning
- Helps you remember your exact a-priori hypotheses
- Can save you a lot of time
- Documents your research and your career
- Allows your study to be included in meta research projects

Registries (non-exhaustive list)

NIH U.S. National Library of Medicine
ClinicalTrials.gov

Registry for clinical trials since 1997
https://clinicaltrials.gov/

World Health Organization

International Clinical Trials Registry Platform (ICTRP)
https://www.who.int/ictrp/network/primary/

OSF

Research management tool (not just for preregistration), embargo possible for up to 4 years
https://osf.io/

ASPREDICTED

Protocols can be private forever
https://aspredicted.org/

PRECLINICALTRIALS.EU
International register of preclinical trial protocols

Comprehensive listing of preclinical animal study protocols
https://preclinicaltrials.eu/

NIHR | National Institute for Health Research

PROSPERO International prospective register of systematic reviews
https://www.crd.york.ac.uk/prospero/

“Challenges” of (pre)-registration

- Changes regarding data collection (e.g., less observations)
- Violation of statistical assumptions, e.g. distribution of variables
- Possible variables unknown until we actually get the data
- Many experiments, large datasets, i.e. discovery science
- Can somebody scoope me?

“Challenges” of (pre)-registration

- Changes regarding data collection (e.g., less observations)
  Solution: document changes to procedures and data acquisition such that it is possible to assess their impact.
- Violation of statistical assumptions, e.g. distribution of variables
  Solution: pre-register decision tree
- Possible variables unknown until we actually get the data
  Solution: look at meta-data or subset of full data for the analysis plan
- Many experiments, large datasets, i.e. discovery science
  Solution: initial experiment exploration, follow-up experiment for confirmation
- Can somebody scoope me?
  Solution: set embargo, registration is time-stamped

Level of detail necessary to (pre)-register a study

Find a middle ground between:

Providing not enough information

- Only time-stamped abstract of the study
- Incomplete information about how exactly data are collected and analyzed
- Failure to provide sample size, exclusion criteria etc.

Providing too much information

- Large section on theoretical background
- Entire paper

(Pre)-registration should be complete and concise in order to make the concept effective

Reporting guidelines

- Simple, structured tool to use while writing manuscripts.
- Providing a minimum list of information needed to ensure a manuscript can be, for example:
  - understood by a reader
  - replicated by a researcher
  - used by a doctor to make a clinical decision
  - included in a systematic review
- Guiding authors in reporting a specific type of research
- Developed using explicit methodology

http://www.equator-network.org/about-us/what-is-a-reporting-guideline/

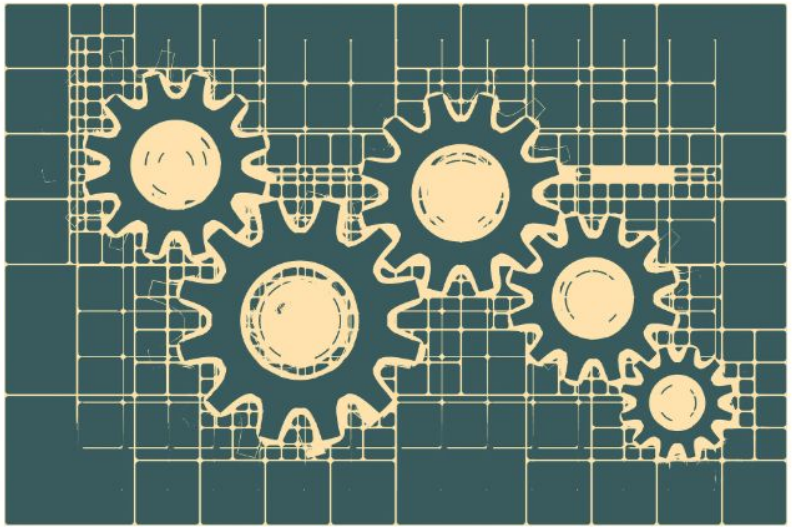
Reporting guidelines

Reporting guidelines for main study types

Randomised trials	CONSORT	Extensions
Observational studies	STROBE	Extensions
Systematic reviews	PRISMA	Extensions
Study protocols	SPIRIT	PRISMA-P
Diagnostic/prognostic studies	STARD	TRIPOD
Case reports	CARE	Extensions
Clinical practice guidelines	AGREE	RIGHT
Qualitative research	SRQR	COREQ
Animal pre-clinical studies	ARRIVE	
Quality improvement studies	SQUIRE	
Economic evaluations	CHEERS	

Towards minimal reporting standards for life scientists

01 Nov 2018 | 13:42 GMT | Posted by Rebecca Walton | Category: Nature Research



This guest blog comes from a group of journal editors and experts in reproducibility and transparent reporting, who are putting together a framework for minimal reporting standards in the life sciences.

http://blogs.nature.com/ofschemesandmemes/2018/11/01/towards-minimal-reporting-standards-life-scientists

University of Zurich[UZH]

http://www.equator-network.org/reporting-guidelines/

Data for the analysis

The following information should be kept (and passed on):

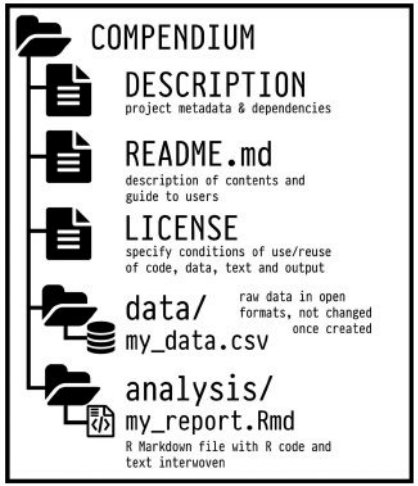
1. raw data
2. “tidy” dataset
3. code book describing each variable and its values in the tidy dataset
4. explicit and exact recipe used by the researcher to go from 1 to 2 and 3.

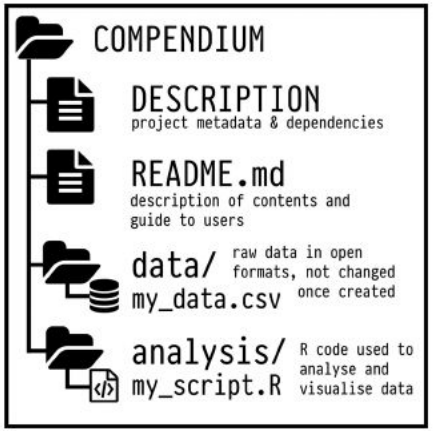
Ellis and Leek (2018) https://doi.org/10.1080/00031305.2017.1375987

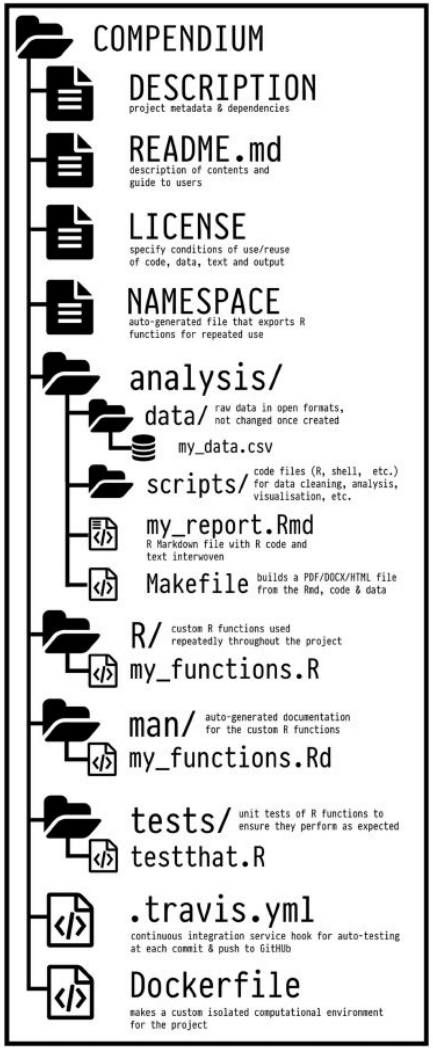
Packaging data analytical work reproducibly

A so-called **research compendium** should:

1. organize its files according to the prevailing conventions of the community
2. maintain a clear separation of data, method, output
3. should specify the computational environment that was used for s





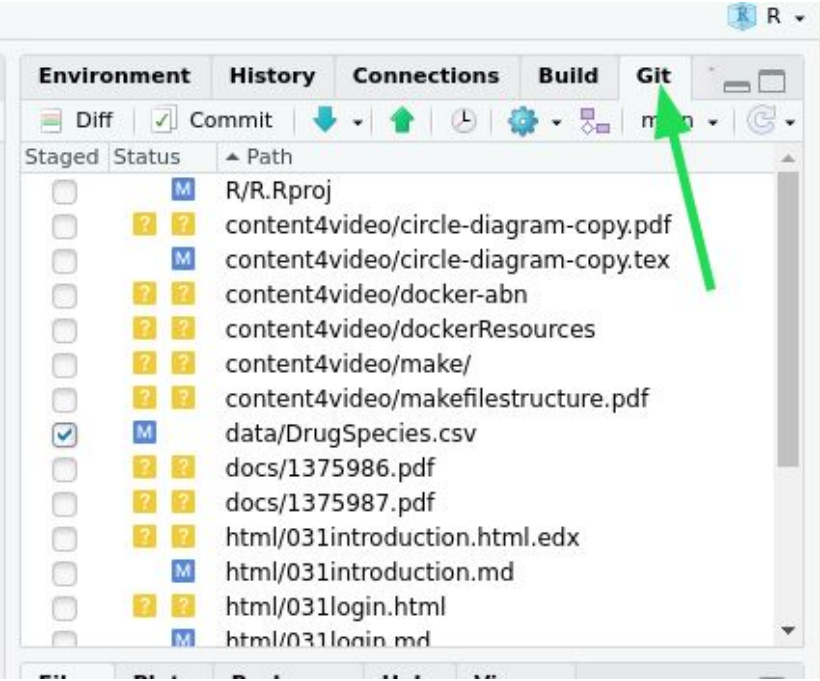


University of Zurich[UZH]

Using versioning systems

- Keep **research compendium** on a distributed version control system
- **GitLab** is a web-based tool that provides a Git repository manager
- **GitHub** is a provider of Internet hosting for software development and version control using Git.

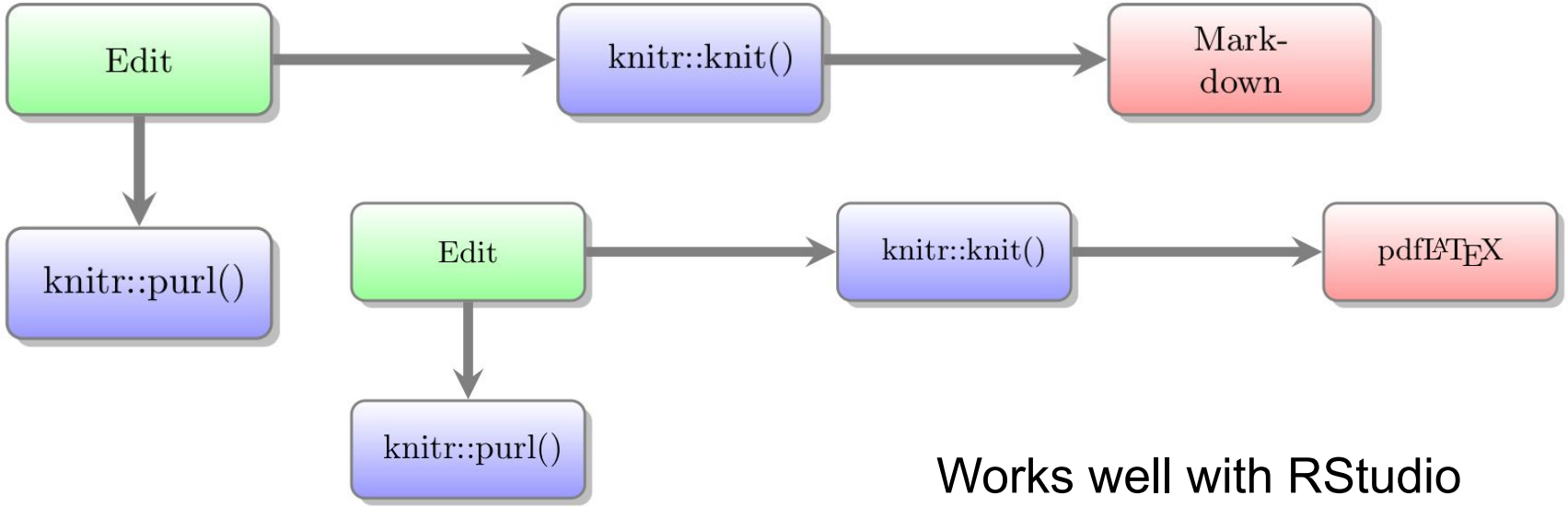
1) In the distributed approach, each developer works directly with their own local repository, and changes are shared between repositories as a separate step.
2) Overleaf can be linked to Github)



Dynamic reporting

Embed R code in your document.

Elements (figures, tables, confidence intervals) are constructed and inserted automatically.



Works well with RStudio

Dynamic reporting: example

```
28  \end{tabular}
29  \end{table}
30
31
32  \begin{rchunk}
33  \rcaption{04:ast}{Inhibition diameters}
34  <<r04:ast,fig.show="hide",purl=FALSE>>=
35  diam <- 28:40    # diameters
36  imi  <- c(0, 3, 7, 14, 32, 20, 18, 4, 1, 1, 0, 0 ,0)  #
37  mero <- c(0, 0, 0, 0,  2, 9, 33, 20, 17, 9, 6, 4, 0)
38  barplot( imi,  names.arg=paste(diam), main="Imipenem")
39  barplot( mero, names.arg=paste(diam), main="Meropenem")
40  imiDat <- rep(diam, imi)   # now a vector with the 100 c
41  c( mean(imiDat), sum( mero*diam)/100)  # means for both,
42  c( var( imiDat), sum( (imiDat-mean(imiDat))^2)/(length(i
43  c( sd( imiDat), sqrt( var( imiDat)))
44  @
45  \end{rchunk}
46
47
48  Although the data of the previous example is  rounded to
    would be reasonable to assume that the diameters are rea
    being conceptually identical to others and thus fluctuat
    common mean.
49  The following is a very simple example of a statistical
50  \begin{align}
51    Y_i = \mu + \varepsilon_i, \qquad i=1,\dots,n,
52    \label{eq02:simple}
53  \end{align}
54  where $Y_i$ are the observations, $\mu$ is an unknown diameter and $\varepsilon_i$
```

R-Code 4.1 Inhibition diameters (See Figure 4.1.)

```
diam <- 28:40    # diameters
imi  <- c(0, 3, 7, 14, 32, 20, 18, 4, 1, 1, 0, 0 ,0)  # frequencies
mero <- c(0, 0, 0, 0,  2, 9, 33, 20, 17, 9, 6, 4, 0)
barplot( imi,  names.arg=paste(diam), main="Imipenem")
barplot( mero, names.arg=paste(diam), main="Meropenem")
imiDat <- rep(diam, imi)   # now a vector with the 100 diameters
c( mean(imiDat), sum( mero*diam)/100)  # means for both, then spread
## [1] 32.40 35.12
c( var( imiDat), sum( (imiDat-mean(imiDat))^2)/(length(imiDat)-1) )
## [1] 2.2424 2.2424
c( sd( imiDat), sqrt( var( imiDat)))
## [1] 1.4975 1.4975
```

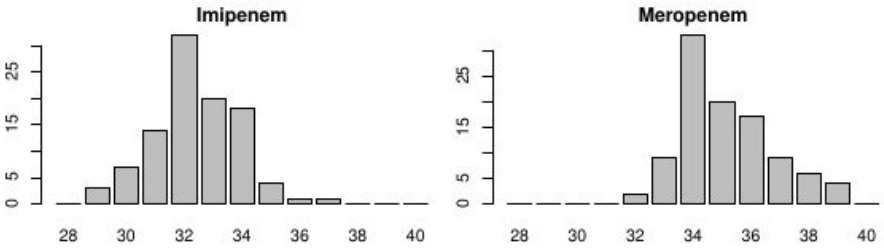
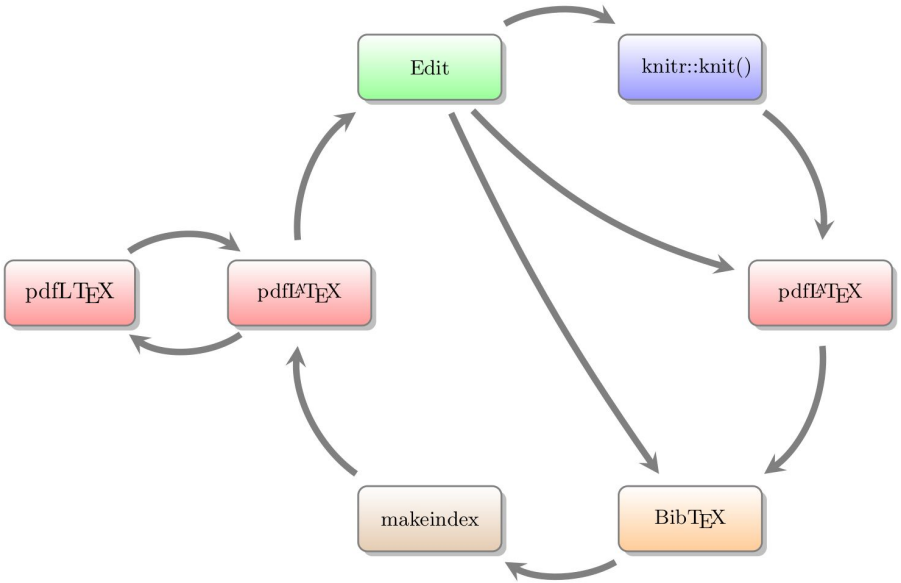


Figure 4.1: Frequencies of inhibition diameters (in mm) by E. coli and imipenem and meropenem (total 100 measurements). (See R-Code 4.1.)

Dynamic reporting

For larger documents, careful dissection into sections and tasks is advisable.

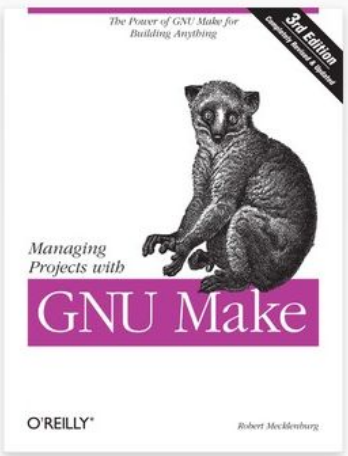


Make is a software tool that compiles and builds components according to specified dependencies.

Further resources

- knitr:
  - overview and basics: https://yihui.org/knitr/
  - summary of options: https://yihui.org/knitr/options/
  - reference card: https://cran.r-project.org/web/packages/knitr/vignettes/knitr-refcard.pdf

- make:
  - accessible description & examples: https://kbroman.org/minimal_make/
  - O'Reilly's GNU make (open!): https://www.oreilly.com/openbook/make3/book/



Containerization

Issues:

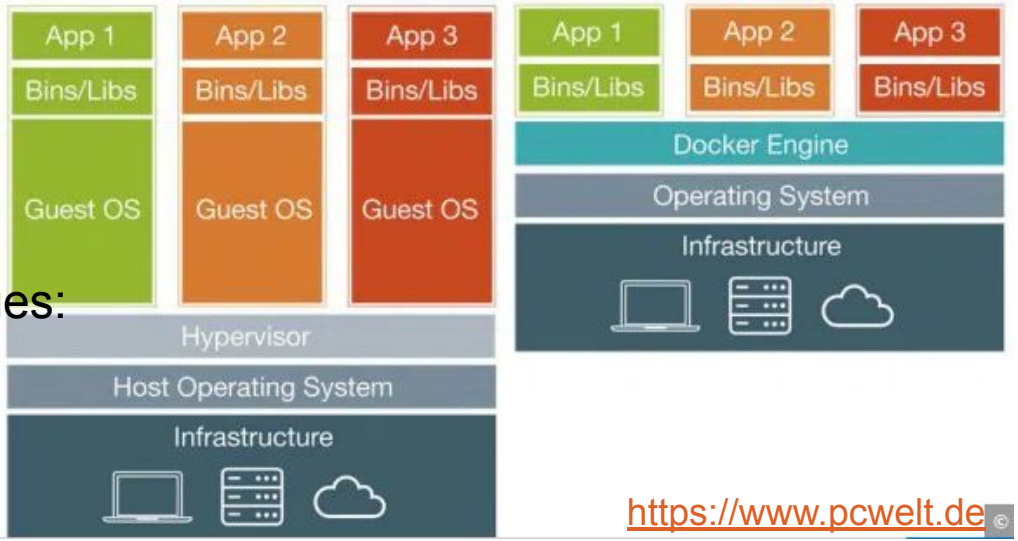
- Evolving software/version dependency
- OS specific software components
- Developing core software

Solution:

- Freeze software environment
- Distribute

Docker

- A Docker **container** is a standardized and enclosed environment that runs applications.
- A Docker **image** is used to build containers. Images are used to store and ship applications.
- A Docker **registry** is a repository for Docker images: https://hub.docker.com/



https://www.pcwelt.de

Docker for R

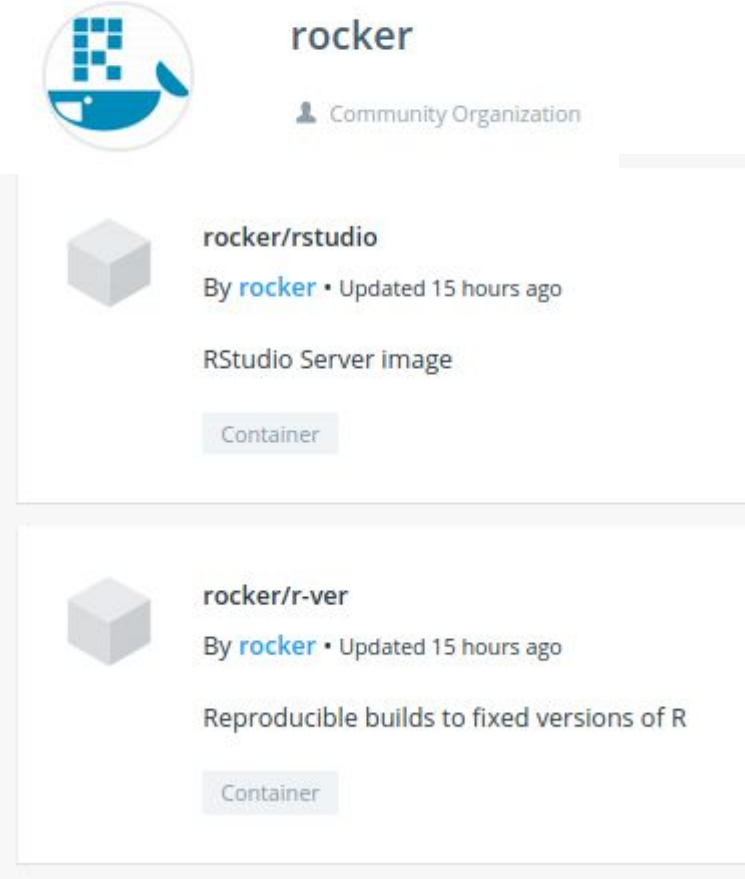
- Prebuild registries for R environments:

  https://hub.docker.com/u/rocker

  All R versions, with RStudio, tidyverse, verse, ...

- containerit: Generating Dockerfiles for reproducible research with R, Nüst & Hinz (2019) https://doi.org/10.21105/joss.01603

- Ten simple rules for writing Dockerfiles for reproducible data science, Nüst et al. (2020) https://doi.org/10.1371/journal.pcbi.1008316



Take home message

- Be aware of reproducibility pitfalls
- Pre-register your study/projects
- Use reporting guidelines
- Strive for complete reporting
- Distribute entire research compendia
- Use software tools for dynamic reporting
- Use a (distributed) version control system for the entire project