Reproducible Research in R

A Talk on How to Do the Same Thing More Than Once

Aaron Peikert

Max Planck Institute for Human Development

Edit

Reproducibility

Edit

Reproducibility

same thing in, same thing out

Edit

As a student:

Conceptually

SEXY

Technologically

SEXY

Edit

At the begin of my PhD:

Conceptually

SEXY

Technologically

SEXY

Edit

At the end of my PhD:

Conceptually

SEXY

Technologically

SEXY

Edit

Now:

Conceptually SEXY

Technologically SEXY

Edit

Conceptual

Edit

Conceptual

Edit

Conceptual

Edit

What is the purpose of reproducibility?

Edit

Same data → Same Black box → Same results

Edit

Same data → Same Black box → Same results

Reproducible?

Edit

Statistical Models
fit
data

Edit

Statistical Models

overfit

data

Edit

By how much?

Edit

By how much?

or

Reproducibility and Overfit

Edit

By how much?

- $R^2_{\text{adj.}}$
- C_p
- AIC
- Cross Validation

Edit

$$R^2_{\text{adj.}} = R^2 - (1 - R^2)\frac{p}{n - p - 1}$$

$$C_p = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sigma_e^2} - n - 2p$$

Edit

Reproducibility?

Edit

Reproducibility?

$$\mathrm{df}(\hat{y}) = \frac{\sum_{i=1}^{n} \mathrm{Cov}(\hat{y}_i, y_i)}{\sigma_e^2}$$

Edit

Reproducibility?

$$\mathrm{df}(\hat{y}) = \frac{\sum_{i=1}^{n} \mathrm{Cov}(\hat{y}_i, y_i)}{\sigma_e^2}$$

This covariance requires:

- formal derivation
- repeated computation on same or pertubated data

Edit

This covariance requires:

- formal derivation
- repeated computation on same or pertubated data

Reproducibility.

Edit

$$R^2_{\text{adj.}} = R^2 - (1 - R^2)\frac{n - df}{df - 1}$$

$$C_p = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sigma_e^2} - 3n + 2df$$

Edit

Information Criteria

Edit

Require the Hessian around the solution for their “overfit” correction.

Computed via:

- byproduct of optimization
- via finite differences

Edit

Require the Hessian around the solution for their “overfit” correction.

Computed via:

Reproducibility

Edit

Cross Validation

is

Reproduction

on subsamples.

Edit

Extended Definition

First, computational reproducibility must ensure that the same data lead to the same results.

Edit

Extended Definition

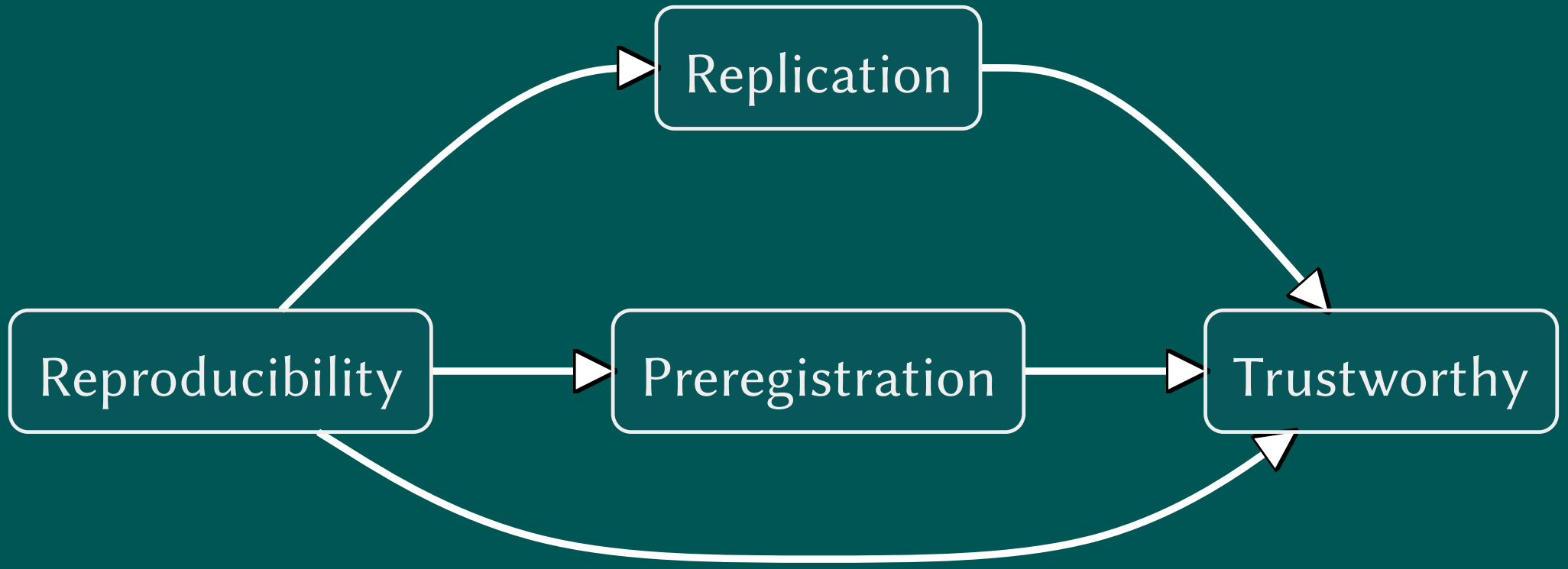
First, computational reproducibility must ensure that the same data lead to the same results.

Second, computational reproducibility must make the inductive process repeatable on similar data.

Edit

Edit

Summary



Edit

Technological

Edit

Technological

> For reproducibility, it really needs to be reproducible and checkable by a stranger with little time or energy to spare, because even the author will soon enough be that stranger.

— Gwern Branwen

Edit

Now you!

Try to add your name to the next slide. Every time I show this presentation your name will be there.

Edit

Awesome People:

- Aaron Peikert

Edit

Four Problems with Reproducibility

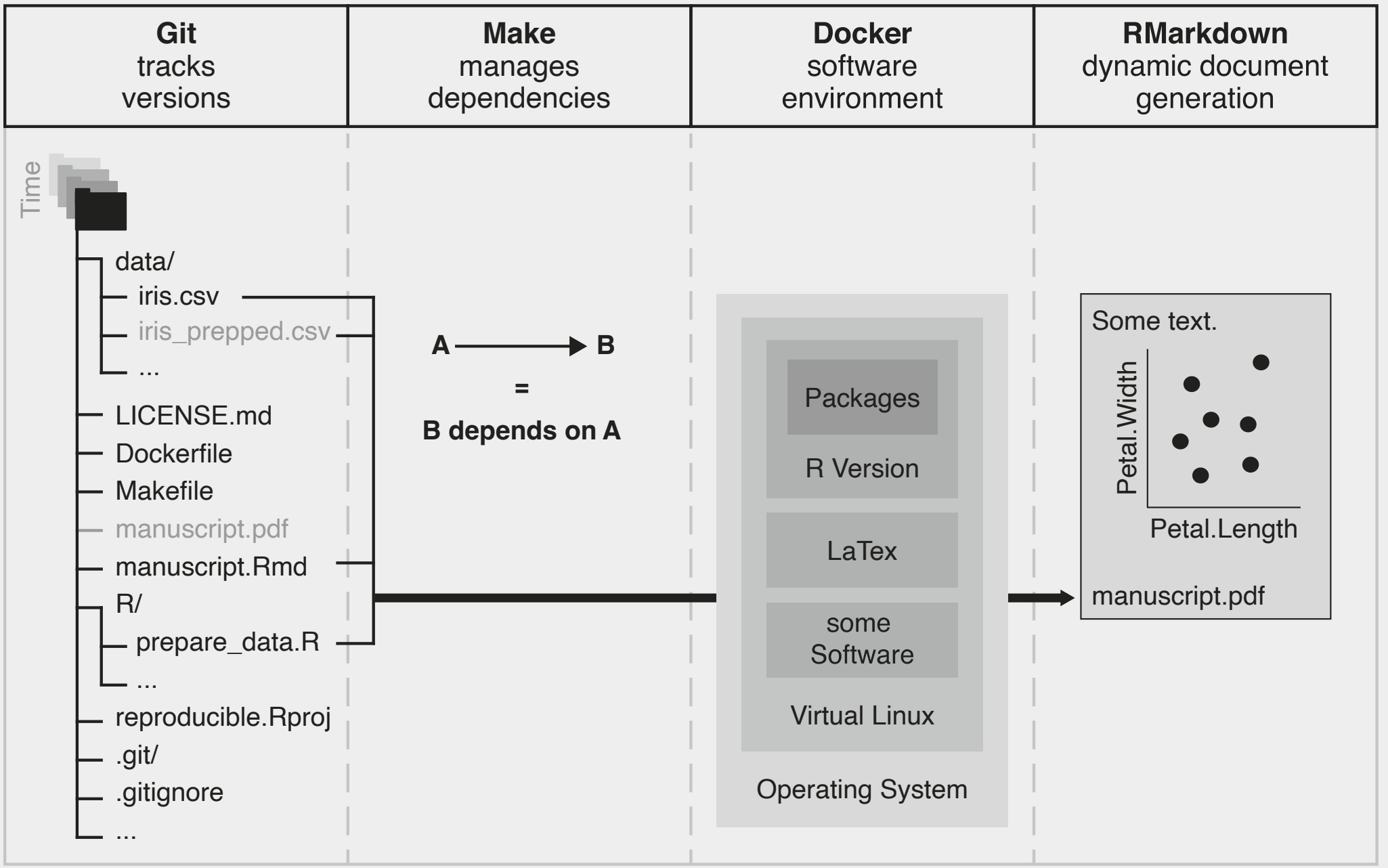
1. versioning
2. copy&paste errors
3. software dependencies
4. linking everything together

Edit

Four Solutions for Reproducibility

1. version control
2. dynamic documents
3. software management
4. workflow orchestration

Edit


Git
tracks
versions
Make
manages
dependencies
Docker
software
environment
RMarkdown
dynamic document
generation
Time
data/
iris.csv
iris_prepped.csv
...
LICENSE.md
Dockerfile
Makefile
manuscript.pdf
manuscript.Rmd
R/
prepare_data.R
...
reproducible.Rproj
.git/
.gitignore
...
A
B
=
B depends on A
Packages
R Version
LaTex
some
Software
Virtual Linux
Operating System
Some text.
Petal.Width
Petal.Length
manuscript.pdf

Edit

ACRISS

Association of Car Rental Industry Systems Standards

Edit

Category	Type	Trans / Driven wheels	Fuel / air-con
E: Economy	F: SUV	A: Auto (drive unspecified)	R: Unspecified Fuel With AC
I: Intermediate	T: Convertible	B: Auto 4WD	D: Diesel With AC
S: Standard	C: 2/4 Door	D: Auto AWD	H: Hybrid With AC
F: Fullsize	D: 4-5 Door	M: Manual (drive unspecified)	E: Electric With AC
P: Premium	S: Sport		V: Petrol With AC

Edit

EDMR

Edit

The Rental Car Model

Ride it like you stole it

Edit

Four Solutions for Reproducibility

1. version control
2. dynamic documents
3. software management
4. workflow orchestration

Edit

R + RMarkdown + Docker + Make + Git

https://github.com/aaronpeikert/reproducible-research

Edit

Julia + RMarkdown + Pkg.jl + GitHub Actions + Git

https://github.com/formal-methods-mpi/pkgmanuscript/blob/main/Dockerfile#L14

Edit

Lua + Quarto + GitHub Actions + GitHub Actions + Git

https://github.com/aaronpeikert/repro-talk

Edit

Python + Quarto + Docker + GitHub Actions + Git

https://github.com/formal-methods-mpi/projects/pull/41/files

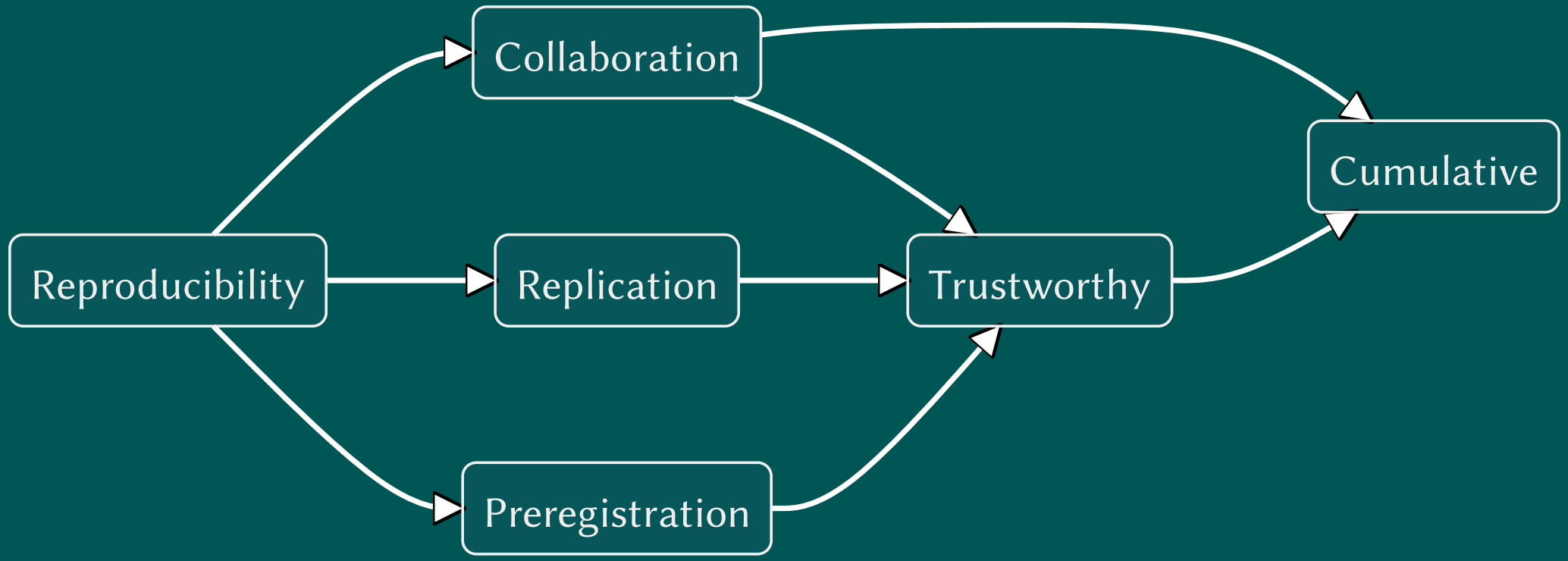
Edit

R + RMarkdown + Docker + Make + Git

https://github.com/aaronpeikert/bayes-prereg/pull/97

Edit

Summary



Edit

Edit