Caring about reproducibility in scientific research

Timothée Flutre

Caring about reproducibility in scientific research

Timothée Flutre

INRA, UMR AGAP

08/12/2014

Last updated : 13/07/2016

License : CC BY-NC-SA 4.0

Abstract

Sooner or later, any research effort implies some data collection in light of a scientific theory, which in turn implies some data analysis. In practice, we enter the raw data in a computer and use a program to fit a model. But soon enough we change the model, add a preprocessing step, collect more data, share our on-going work with a student in the nearby office or with a collaborator abroad, etc. And in a realistic research project, we do this again, and again, and again... Such a process is now often called "computational science". In this context, we usually want to keep track of what we do and, even better, what our student/collaborator do. We would also like to better understand what other teams did in their recently-published paper and, why not, apply their code on our data or even combine their data with ours.

In this presentation, I will rapidly outline the current state of affairs about reproducibility in research and what it says for code and data. I will then present in some details the computational tools I am using, after years of trials and errors and inspirations from many people. In the end, everyone is welcome to bring his computer to try by himself on a minimal case study (many cables will be available for internet connection).

In these slides, terms in grey correspond to HTML links : click on them ;)

Outline

Global context

Spectrum of solutions

Outline

Global context

Spectrum of solutions

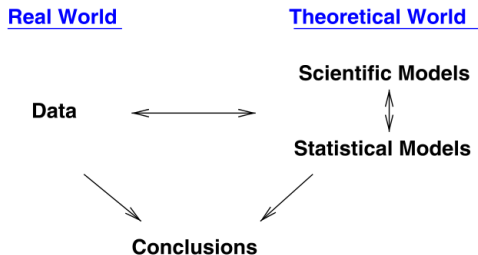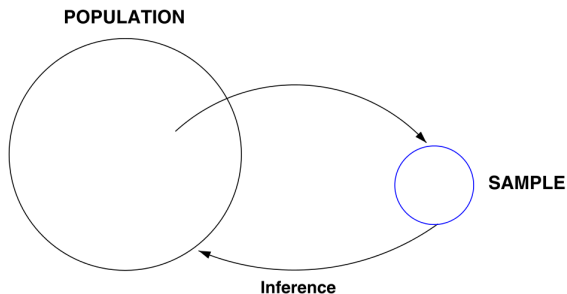# Statistical Inference : the Big Picture (Kass, 2011)

FIG. 1. *The big picture of statistical inference. Statistical procedures are abstractly defined in terms of mathematics but are used, in conjunction with scientific models and methods, to explain observable phenomena. This picture emphasizes the hypothetical link between variation in data and its description using statistical models.*

Statistical Inference : the Big Picture (Kass, 2011)

Standard conception :

Statistical Inference : the Big Picture (Kass, 2011)
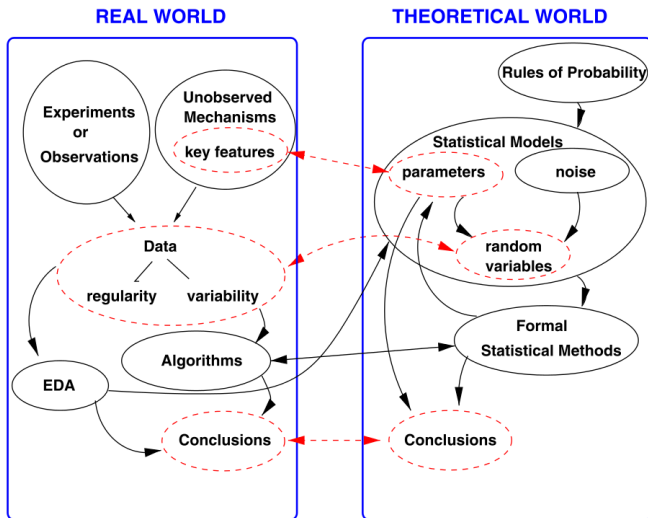
A gloomy title, but crucial statistical advice

Why Most Published Research Findings Are False

John P. A. Ioannidis

A gloomy title, but crucial statistical advice

Why Most Published Research Findings Are False

John P. A. Ioannidis

- Did you hear about this paper?

A gloomy title, but crucial statistical advice

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • DOI: 10.1371/journal.pmed.0020124

► Did you hear about this paper ?

Total citations   Cited by 2352

2005 2006 2007 2008 2009 2010 2011 2012 2013 2014

A gloomy title, but crucial statistical advice

Why Most Published Research Findings Are False

John P. A. Ioannidis

▶ Did you hear about this paper ?

Total citations   Cited by 2352
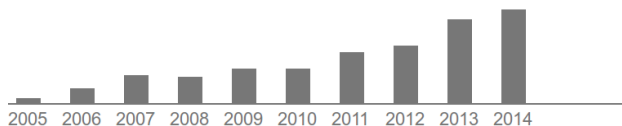
▶ Did you read it ?

Positive Predictive Value (PPV)

post-study probability that a significant research finding is true

Table 4. PPV of Research Findings for Various Combinations of Power $(1 - \beta)$, Ratio of True to Not-True Relationships *(R)*, and Bias *(u)*

$1 - \beta$	*R*	*u*	**Practical Example**	**PPV**
0.80	1:1	0.10	Adequately powered RCT with little bias and 1:1 pre-study odds	0.85
0.95	2:1	0.30	Confirmatory meta-analysis of good-quality RCTs	0.85
0.80	1:3	0.40	Meta-analysis of small inconclusive studies	0.41
0.20	1:5	0.20	Underpowered, but well-performed phase I/II RCT	0.23
0.20	1:5	0.80	Underpowered, poorly performed phase I/II RCT	0.17
0.80	1:10	0.30	Adequately powered exploratory epidemiological study	0.20
0.20	1:10	0.30	Underpowered exploratory epidemiological study	0.12
0.20	1:1,000	0.80	Discovery-oriented exploratory research with massive testing	0.0010
0.20	1:1,000	0.20	As in previous example, but with more limited bias (more standardized)	0.0015

The estimated PPVs (positive predictive values) are derived assuming $\alpha = 0.05$ for a single study.

It's not only about weak power and high bias...

DERIVING CHEMOSENSITIVITY FROM CELL LINES: FORENSIC BIOINFORMATICS AND REPRODUCIBLE RESEARCH IN HIGH-THROUGHPUT BIOLOGY

By Keith A. Baggerly[1] and Kevin R. Coombes[2]

University of Texas

High-throughput biological assays such as microarrays let us ask very detailed questions about how diseases operate, and promise to let us personalize therapy. Data processing, however, is often not described well enough to allow for exact reproduction of the results, leading to exercises in "forensic bioinformatics" where aspects of raw data and reported results are used to infer what methods must have been employed. Unfortunately, poor documentation can shift from an inconvenience to an active danger when it obscures not just methods but errors. In this report we examine several related papers purporting to use microarray-based signatures of drug sensitivity derived from cell lines to predict patient response. Patients in clinical trials are currently being allocated to treatment arms on the basis of these results. However, we show in five case studies that the results incorporate several simple errors that may be putting patients at risk. One theme that emerges is that the most common errors are simple (e.g., row or column offsets); conversely, it is our experience that the most simple errors are common. We then discuss steps we are taking to avoid such errors in our own investigations.

A hopeful title, and no less crucial advice
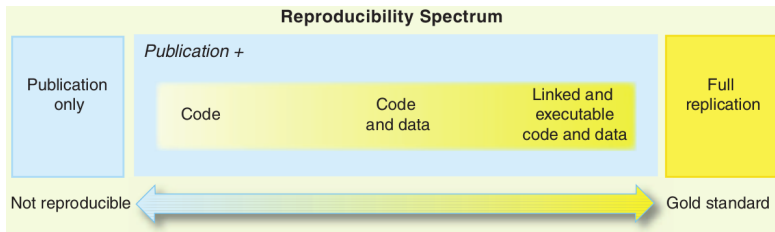
How to Make More Published Research True
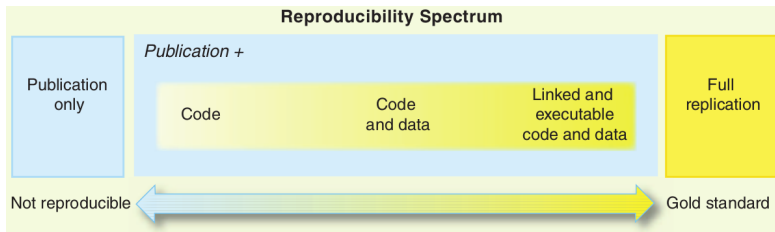
John P. A. Ioannidis ✉

Read it here !

> To make more published research true, practices that have improved credibility and efficiency in specific fields may be transplanted to others which would benefit from them—possibilities include the adoption of large-scale collaborative research; replication culture; registration; sharing; reproducibility practices; better statistical methods; standardization of definitions and analyses; more appropriate (usually more stringent) statistical thresholds; and improvement in study design standards, peer review, reporting and dissemination of research, and training of the scientific workforce.

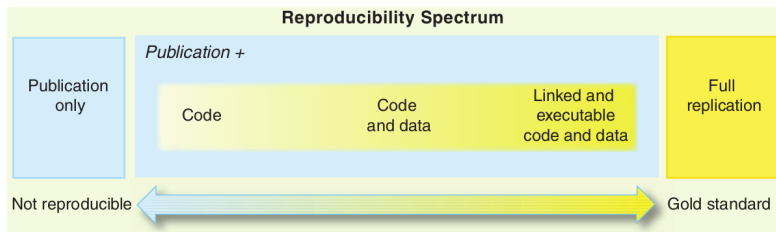# Reprod. Research in Computational Science (Peng, 2011)

Reproducibility Spectrum

Publication only

Publication +

Code | Code and data | Linked and executable code and data

Full replication

Not reproducible ← → Gold standard

Reprod. Research in Computational Science (Peng, 2011)

Reproducibility Spectrum

Publication only — Publication +: Code → Code and data → Linked and executable code and data — Full replication

Not reproducible ←→ Gold standard

▶ Are you a computational scientist ?

Reprod. Research in Computational Science (Peng, 2011)

- ▶ Are you a computational scientist ?
  - ▶ Do you use Excel ? R ? Do you plot histograms ? Do you average columns of numbers, and calculate variances ?

Reprod. Research in Computational Science (Peng, 2011)
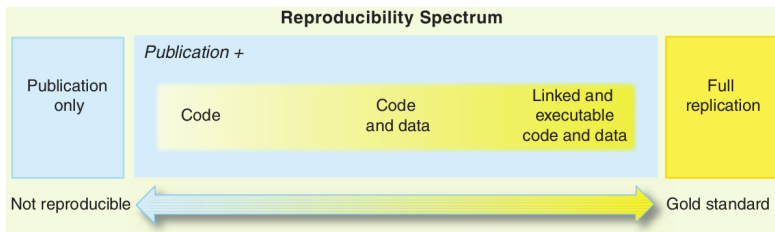
- ▶ Are you a computational scientist?
  - ▶ Do you use Excel? R? Do you plot histograms? Do you average columns of numbers, and calculate variances?

- ▶ Where are you along the spectrum?

Let us start with code...

Claerbout & Karrenbach (1992) :

An article about a computational result is advertising, not scholarship. The actual scholarship is the full software environment, code and data, that produced the result.

Let us start with code...

Claerbout & Karrenbach (1992) :

> *An article about a computational result is advertising,
> not scholarship. The actual scholarship is the full software
> environment, code and data, that produced the result.*

Buckheit & Donohoe (1995) :

> *Publishing figures or results without the complete
> software environment could be compared to a
> mathematician publishing an announcement of a
> mathematical theorem without giving the proof.*

… and now data.

Where Have All the Crop Phenotypes Gone?

Dani Zamir ✉

… and now data.

Where Have All the Crop Phenotypes Gone?

Dani Zamir ✉

Currently, virtually none of the data generated from the hundreds of phenotypic studies conducted each year are being made publically available as raw data ; thus there is little we can learn from past experience when making decisions about how to breed better crops for the future.

... and now data.

Where Have All the Crop Phenotypes Gone?

Dani Zamir ✉

Currently, virtually none of the data generated from the hundreds of phenotypic studies conducted each year are being made publically available as raw data ; thus there is little we can learn from past experience when making decisions about how to breed better crops for the future.

Zamir (Science, 2014) :

without the corresponding potentially commercially valuable phenotypic data. For example, in rice (*Oryza sativa*), which feeds roughly half the world population, 3000 variants from 89 countries were sequenced revealing 18.9 million single-nucleotide polymorphisms (SNPs) (*4*). But what good are 3000 genomes if the associated phenotypic data, and sometimes seed stocks, are kept proprietary? A wake-up call is needed for scientists, granting agencies, journal editors, and referees: What we eat are phenotypes, and seriously addressing global food security demands that, at least in the domain of crop plants, phenotypic data should be shared between scientists in the same manner as for sequences (*3*).

An extreme view : what do you think ?

The NEW ENGLAND JOURNAL of MEDICINE

HOME | ARTICLES & MULTIMEDIA ⌄ | ISSUES ⌄ | SPECIALTIES & TOPICS ⌄ | FOR AUTHO

EDITORIAL

Data Sharing

Dan L. Longo, M.D., and Jeffrey M. Drazen, M.D.
N Engl J Med 2016; 374:276-277 | January 21, 2016 | DOI: 10.1056/NEJMe1516564

"research parasites" [...] people who had nothing to do with the design and execution of the study but use another group's data for their own ends, possibly stealing from the research productivity planned by the data gatherers, or even use the data to try to disprove what the original investigators had posited

Incentives, to be improved
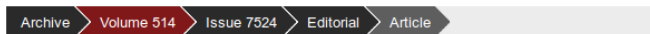
NATURE | EDITORIAL

Code share

Papers in Nature journals should make computer code accessible where possible.

29 October 2014

Publication is conditional upon the agreement of the authors to make freely available any materials and information described in their publication that may be reasonably requested by others.

Data Availability

PLOS journals require authors to make all data underlying the findings described in their manuscript fully available without restriction, with rare exception[1].

Incentives, to be improved

Archive › Volume 514 › Issue 7524 › Editorial › Article

NATURE | EDITORIAL

Code share

Papers in Nature journals should make computer code accessible where possible.

29 October 2014

Publication is conditional upon the agreement of the authors to make freely available any materials and information described in their publication that may be reasonably requested by others.

Data Availability

PLOS journals require authors to make all data underlying the findings described in their manuscript fully available without restriction, with rare exception[1].

Prod'INRA : automatic export of productions for individual assessment, including softwares, databases, biological material, etc

But that's not it!

Blocker & Meng (Bernoulli, 2013):

> *Decisions made in preprocessing constrain all later analyses and are typically irreversible. Hence, data analysis becomes a collaborative endeavor by all parties involved in data collection, preprocessing and curation, and downstream inference.*

Preprocessor's model
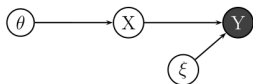
Downstream analyst's model
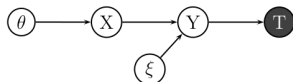
After all, why is reproducibility important ?

Which reason(s) would you choose ?

After all, why is reproducibility important?

Which reason(s) would you choose?

- ▶ reproducing work is also the first step to extending it

- ▶ we are forgetful, error-prone (or dishonest)

- ▶ helps communication with your collaborators

- ▶ bigger visibility in the community

- ▶ we are mostly funded by public money

- ▶ ...

Outline

Multi-dimensional space of "projects"

- question : mono- or inter-disciplinary ; closer to applied or to basic research
  - won't be discussed here, but obviously crucial...

Multi-dimensional space of "projects"

- ▶ question : mono- or inter-disciplinary ; closer to applied or to basic research
  - ▶ won't be discussed here, but obviously crucial...

- ▶ data : few small files or many large files ; custom or standardized formats ; databases

Multi-dimensional space of "projects"

- question : mono- or inter-disciplinary ; closer to applied or to basic research
  - won't be discussed here, but obviously crucial...

- data : few small files or many large files ; custom or standardized formats ; databases

- code : in interpreted languages (e.g. R, Python, Perl, Bash) or compiled languages (e.g. C, C++, Fortran)

Multi-dimensional space of "projects"

- **question** : mono- or inter-disciplinary ; closer to applied or to basic research
  - won't be discussed here, but obviously crucial...

- **data** : few small files or many large files ; custom or standardized formats ; databases

- **code** : in interpreted languages (e.g. R, Python, Perl, Bash) or compiled languages (e.g. C, C++, Fortran)

- **model** : informal (any word processor is enough, e.g. Writer, Word) or formal (many equations, much easier in LaTeX)

Multi-dimensional space of "projects"

- ▶ question : mono- or inter-disciplinary ; closer to applied or to basic research
  - ▶ won't be discussed here, but obviously crucial...

- ▶ data : few small files or many large files ; custom or standardized formats ; databases

- ▶ code : in interpreted languages (e.g. R, Python, Perl, Bash) or compiled languages (e.g. C, C++, Fortran)

- ▶ model : informal (any word processor is enough, e.g. Writer, Word) or formal (many equations, much easier in LaTeX)

- ▶ collaborators : professional scientist or anyone else ; beginner or experienced ; "open curious" or already fully proficient

Initial steps toward reproducible research

Karl Broman's tutorial : `http://kbroman.org/steps2rr/`

1. Everything with a script
2. Organize your data and code
3. Automate the process
4. Turn scripts into reproducible reports
5. Turn repeated code into functions
6. Package functions for reuse
7. Use version control

Initial steps toward reproducible research

Karl Broman's tutorial : `http://kbroman.org/steps2rr/`

1. Everything with a script
2. Organize your data and code
3. Automate the process
4. Turn scripts into reproducible reports
5. Turn repeated code into functions
6. Package functions for reuse
7. Use version control

See also : `http://software-carpentry.org/`

A text editor, not a word processor

- ▶ word processors : LibreOffice Writer, Microsoft Word, etc
- ▶ text editors : Emacs, Vim, Notepad++, Gedit, Geany, etc

A text editor, not a word processor

- ▶ word processors : LibreOffice Writer, Microsoft Word, etc
- ▶ text editors : Emacs, Vim, Notepad++, Gedit, Geany, etc

Why knowing/mastering a good text editor is crucial ?

- ▶ essential tool of any data analysis, e.g. scripts
- ▶ light-weight markup languages convert plain text to PDF/HTML/odt/docx/etc, check out pandoc
- ▶ delivers full power of text manipulation tools, e.g. `diff` to compare, `grep` to search, `awk` to extract, ...
- ▶ allows efficient use of version control systems such as git

Project organization

Follow the spirit of Noble (PLoS Comput Biol, 2009) :

► someone unfamiliar with your project should be able to look at your computer files and understand in detail what you did and why ;

► everything you do, you will probably have to do it over again.

Project organization

Follow the spirit of Noble (PLoS Comput Biol, 2009) :

▶ someone unfamiliar with your project should be able to look at your computer files and understand in detail what you did and why ;

▶ everything you do, you will probably have to do it over again.

1. choose a short project name and create a directory
2. give brief explanations in a text file named README
3. describe sharing/modifying rights in a text file named COPYING or LICENSE
4. list authors in a text file named AUTHORS
5. create subdirectories doc/, data/, src/ and results/

Version control systems (VCS)

Karl Broman on VCS :

- ▶ not strictly necessary for *reproducibility*, but can be hugely useful for *sanity*
- ▶ requires a big initial investment in time and effort, but become a natural part of your daily workflow after a month or so
- ▶ huge *short-term* advantages in collaborative projects (keeping in sync, merging simultaneous changes)

Version control systems (VCS)

Karl Broman on VCS :

- ▶ not strictly necessary for *reproducibility*, but can be hugely useful for *sanity*
- ▶ requires a big initial investment in time and effort, but become a natural part of your daily workflow after a month or so
- ▶ huge *short-term* advantages in collaborative projects (keeping in sync, merging simultaneous changes)

My advice : use git (free software ; multi-platform ; used all over the globe by programmers, scientists, journalists, etc)

- ▶ download it from http://git-scm.com/
- ▶ *read* chapters 1 to 3 of the official book

Two examples

1. project "ligth" : a few small data files ; laptop ; classical models ; available implementations for interpreted languages
   - ok to re-run the whole analysis from time to time
   - write notebook in text file in Rmd format and use RStudio

Two examples

1. project "ligth" : a few small data files ; laptop ; classical models ; available implementations for interpreted languages
   - ok to re-run the whole analysis from time to time
   - write notebook in text file in Rmd format and use RStudio

2. project "heavy" : many large files, e.g. sequencing data ; cluster ; possibly new models and implementations
   - can't re-run the whole analysis because of intensive computations
   - write notebook in text file in org format and use Emacs or pandoc

Project "light"

1. create your project directory : `README`, `COPYING`, `AUTHORS`, `doc/`, `data/`, `src/`, `results/`
2. (re-)install the latest version of RStudio : http://www.rstudio.com/
3. visit github.com/timflutre/tuto-reproducible-research and download the Rmd file in `project_light/`
4. convert it from Rmd to HTML in RStudio by clicking on `Knit to HTML`

Project "light"

1. create your project directory : `README`, `COPYING`, `AUTHORS`, `doc/`, `data/`, `src/`, `results/`
2. (re-)install the latest version of RStudio : `http://www.rstudio.com/`
3. visit github.com/timflutre/tuto-reproducible-research and download the Rmd file in `project_light/`
4. convert it from Rmd to HTML in RStudio by clicking on `Knit to HTML`

Play !
`https://github.com/csgillespie/statslang/tree/master/R`
`http://cran.r-project.org/web/packages/agridat/index.html`
`http://cran.r-project.org/web/packages/HistData/`

Project "heavy"

Usually, if large data, then computer cluster, and "shared space" :

- external_public/, external_private/, internal/

Project "heavy"
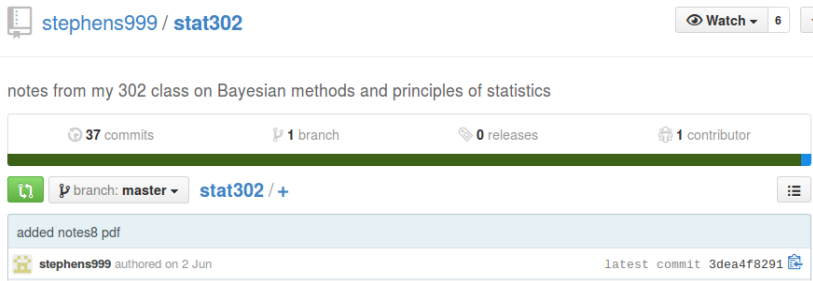
Usually, if large data, then computer cluster, and "shared space" :

- external_public/, external_private/, internal/

1. create the backbone for project_heavy in your home
2. version it with git, and push it to the server
3. clone the repository for project_heavy in the shared space
4. put the large data files data/ in the shared space, but don't version them
5. edit versioned files in your home, and commit/push
6. run intensive computations in the shared space (coordinate with other users !)
7. update (pull) repository on the shared space (need one origin-<user> per user)

Are you teaching ?

Example in statistics from Matthew Stephens : class notes versioned with git and hosted on GitHub

stephens999 / **stat302**

Watch ▾ 6

notes from my 302 class on Bayesian methods and principles of statistics

⊙ **37** commits   ⑂ **1** branch   ◎ **0** releases   👤 **1** contributor
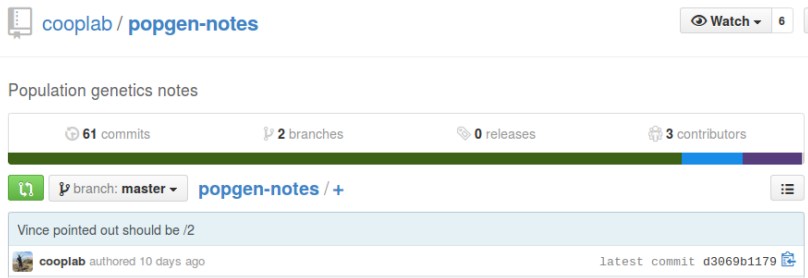
⑂ branch: **master** ▾   **stat302** / +

added notes8 pdf

stephens999 authored on 2 Jun   latest commit 3dea4f8291

Are you teaching ? Cont'd

Example in population genetics from Graham Coop : class notes
versioned with git and hosted on GitHub

Are you writing an article ?

Example in plant biology from Rubén Rellán Álvarez : article (text, figures) and software versioned with git and hosted on GitHub
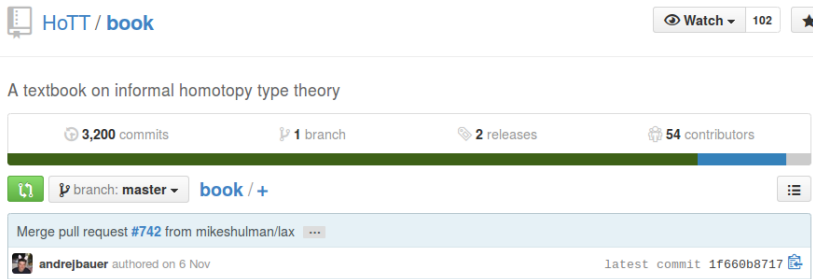
Are you writing a book ?

Example in mathematics from the "Homotopy Type Theory" group which started at Princeton (IAS) and now writes a book versioned with git and hosted on GitHub

Take-home message

Whatever the exact tools you are using, it is the spirit that is important !

Take-home message

Whatever the exact tools you are using, it is the spirit that is important !

- ► How well does your behavior favor others understanding and building on your work ?

- ► When and how will you start ?