Reproducibility from a Mostly Selfish Point of View

Noam Ross

EHA Science Meeting

2016-01-26

Reproducibility: Can scientific results be re-created from data and source material?

- Can the tables and figures be made again?
- Does the analysis actually do what you think it does?
- Is it clear *why* it was done?
- Can the analysis be used on new data?
- Can analysis be extended do other things?

- Not the same thing as, but related to: *replicability, computational, openness, etc.*

Selfish reasons for reproducibility

- Quality Control and Risk Management
- Productivity
- Collaboration
- Impact
- Funding Mandates

Quality Control and Risk Management

Is The Reinhart-Rogoff Result Based on a Simple Spreadsheet Error?

By Matthew Yglesias

2k 0

Update: Reinhart and Rogoff have responded.

So this is huge. Or, rather, it won't matter even a tiny little bit but it ought to be a big deal anyway. You've probably heard that countries with a high debt:GDP ratio suffer from slow economic growth. The specific number 90 percent has been invoked frequently. That's all thanks to a study conducted by Carmen Reinhardt and Kenneth Rogoff for their book *This Time It's Different*. But the results have been difficult for other researchers to replicate. Now three scholars at the University of Massachusetts have done so in "**Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff**" and they find that the Reinhart/Rogoff result is based on opportunistic exclusion of Commonwealth data in the late-1940s, a debatable premise about how to weight the data, and most of all a sloppy Excel coding error.

http://www.slate.com/blogs/moneybox/2013/04/16/reinhart_rogoff_coding_error_austerity_policies_founded_on_bad_coding.html

Quality Control and Risk Management

Sciencexpress

Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent

M. Gallego Llorente,[1*†] E. R. Jones,[2*†] A. Eriksson,[1,3] V. Siska,[1] K. W. Arthur,[4] J. W. Arthur,[4] M. C. Curtis,[5,6] J. T. Stock,[7] M. Coltorti,[8] P. Pieruccini,[8] S. Stretton,[9] F. Brock,[10,11] T. Higham,[10] Y. Park,[12] M. Hofreiter,[13,14] D. G. Bradley,[2] J. Bhak,[15] R. Pinhasi,[16*] A. Manica[1*]

dx.doi.org/10.1126/science.aad2879

The New York Times

Science

SUBSCRIBE | LOG IN

Scientists Recover First Genome of Ancient Human From Africa

"The most astonishing thing is there's quite a lot of backflow in all modern African populations," Dr. Pinhasi said. He and his colleagues estimate that 7 percent of the genomes of the Yoruba people of Nigeria are of Eurasian origin. In the genomes of Mbuti pygmies who live in the rain forest in the Democratic Republic of Congo, 6 percent of the DNA comes from Eurasians.

www.nytimes.com/2015/10/09/science/scientists-sequence-first-ancient-human-genome-from-africa.html

Quality Control and Risk Management

There Was No Vast Migration of Eurasians Into Africa

RAZIB KHAN • JANUARY 25, 2016 • 700 WORDS • 9 COMMENTS • REPLY

The results presented in the Report "Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent" were affected by a bioinformatics error. A script necessary to convert the input produced by samtools v0.1.19 to be compatible with PLINK was not run when merging the ancient genome, Mota, with the contemporary populations SNP panel, leading to homozygote positions to the human reference genome being dropped as missing data (the analysis of admixture with Neanderthals and Denisovans was not affected). When those positions were included, 255,922 SNP out of 256,540 from the contemporary reference panel could be called in Mota. The conclusion of a large migration into East Africa from Western Eurasia, and more precisely from a source genetically close to the early Neolithic farmers, is not affected. However, the geographic extent of the genetic impact of this migration was overestimated: the Western Eurasian backflow mostly affected East Africa and only a few Sub-Saharan populations; the Yoruba and Mbuti do not show higher levels of Western Eurasian ancestry compared to Mota.

We thank Pontus Skoglund and David Reich for letting us know about this problem.

-- "If something like this happened to me I'd probably literally throw up."

www.unz.com/gnxp/there-was-no-vast-migration-of-eurasians-into-africa/

Productivity

Karl -- this is very interesting, however you used an old version of the data (n=143 rather than n=226).

I'm really sorry you did all that work on the incomplete dataset.

Bruce

Collaboration

Gaps in reproducibility are gaps in communication

- With each other across the organization
- With outside partners, now and future
- With those who re-use or extend our research
- With ourselves, six months from now

Impact

- More products, more audiences
- Papers
- Reports
- Blog posts
- Tutorials
- Datasets
- Tools/interactives

rambaut.github.io/MERS-Tools/cases2.html

MERS-CoV spatial, temporal and epidemiological information

Submitted by Andrew Rambaut on Tue, 2013-06-18 17:29

Distribution of cases

Temporal distribution of cases by country, age, gender and clinical outcome. Drag to select time ranges or age ranges, click on bars to include/exclude. Select the 'primary' bar in 'By cluster' to filter out those known to be secondary cases. Country is by probable location of infection.

> *This list of cases does not include the 113 additional laboratory confirmed cases reported to the WHO on 3 June 2014. These date from May 2013 and include 34 deaths.*

Notes: The live data files for these graphics are now hosted on my GitHub repository: MERS-Cases. Please feel free to fork and update these files. Push changes to update the page. The list of cases has been gathered from various sources including WHO bulletins and media reports. It contains some cases that were not laborary confirmed (but are extremely likely). There will be inaccuracies and omissions and it should be considered illustrative of the current situation. Thanks to Paul Wikramaratna for extensive work on this list to keep it as accurate and comprehensive as possible.

Follow @epidemicks on Twitter to see edit by edit updates to the data file.

Visualisations coded using Leaflet, CrossFilter, D3.js, dc.js and a few other bits and pieces.

1,286 selected out of 1,286 cases | reset all

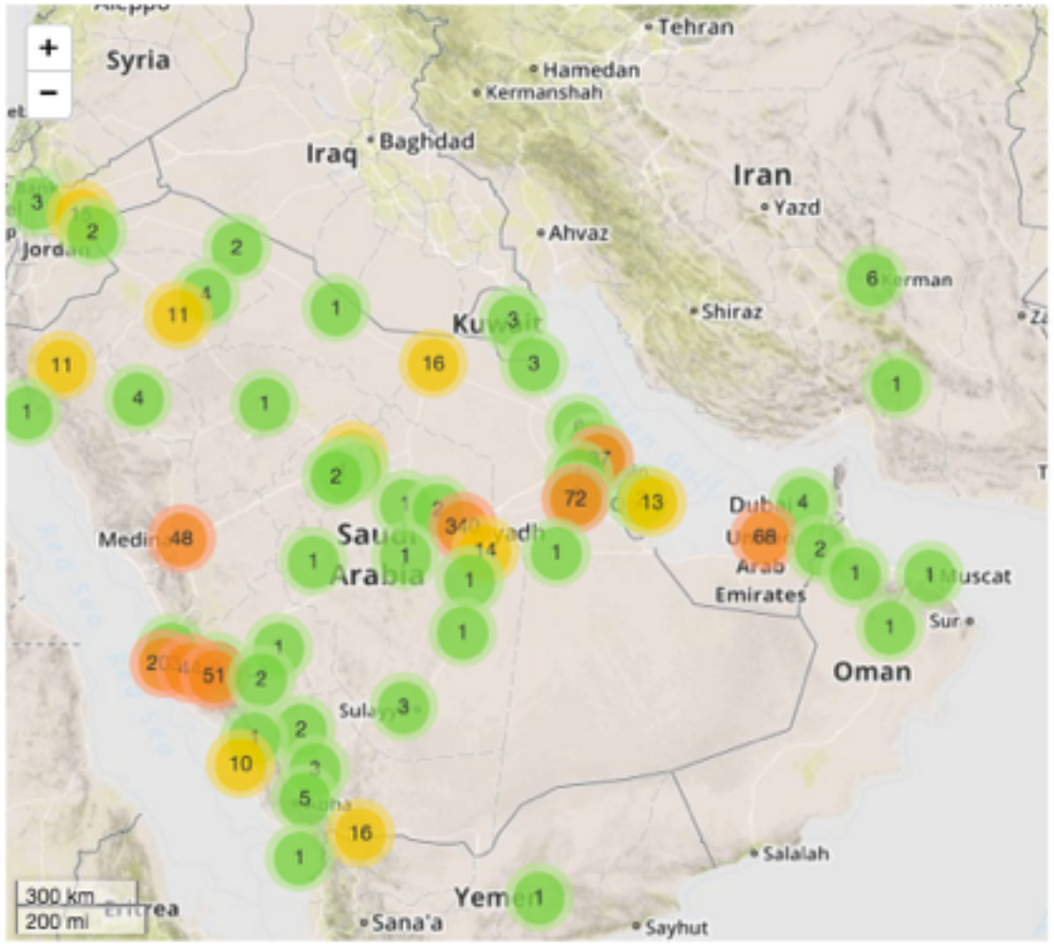
Spatial Distribution

Country	
1 France	
8 Iran	
2 Italy	
20 Jordan	
959 KSA	
3 Kuwait	
1 Lebanon	
9 Oman	
15 Qatar	
186 South Korea	
2 Tunisia	
77 UAE	
2 UK	
1 Yemen	

0 200400600800

KSA location	
1 Abqaiq	
1 Afif	
1 Al Badaya	
1 Al Hawiyah	
11 Al Jawf Region	
1 Al Makhwah	
10 Al Qunfudhah	
1 Al-Aflaj	
5 Al-Ahsa	
72 Al-Hofuf	
14 Al-Kharj	
1 Al-Quway'iyah	
2 Ar Rass	
2 Arar	
1 Asfan	
3 Asir Province	
2 Bisha	
14 Buraydah	
13 Dammam	
5 Dhahran	
1 Duba	
13 Eastern Province	
1 Ha'il	
16 Hafr Al-Batin	
1 Haradh	
1 Hotat Bani Tamim	
2 Huraymila	
203 Jeddah	
1 Jizan	
6 Jubail	
3 Khafji	
5 Khamis Mushait	
11 Khobar	
1 Khurmah	
48 Madinah	
44 Makkah	
16 Najran	
3 Qatif	
2 Qurayyat	
1 Rafha	
340 Riyadh	
4 Sakakah	
1 Shaqraa	
11 Tabuk	
51 Taif	
4 Taimah	
2 Turaba	
2 Unaizah	
3 Wadi al-Dawasir	
327 other countries	
1 unknown	

0 100 200 300



Mandates

- Funders
- Publishers

Availability of data, material and methods

An inherent principle of publication is that others should be able to replicate and build upon the authors' published claims. A condition of publication in a Nature journal is that **authors are required to make materials, data, code, and associated protocols promptly available to readers without undue qualifications**. Any restrictions on the availability of materials or information must be disclosed to the editors at the time of submission. Any restrictions must **also** be disclosed in the submitted manuscript.

After publication, readers who encounter refusal by the authors to comply with these policies should contact the chief editor of the journal. In cases where editors are unable to resolve a complaint, the journal may refer the matter to the authors' funding institution and/or publish a formal statement of correction, attached online to the publication, stating that readers have been unable to obtain necessary materials to replicate the findings.

http://www.nature.com/authors/policies/availability.html

Science

Data and Materials Availability after Publication

After publication, all data and materials necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of *Science*. All computer codes involved in the creation or analysis of data must also be available to any reader of *Science*. After

http://www.sciencemag.org/authors/science-editorial-policies

Dissemination and Sharing of Research Results

NSF Data Sharing Policy

Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing. See Award & Administration Guide (AAG) Chapter VI.D.4.

https://www.nsf.gov/bfa/dias/policy/dmp.jsp

General Principles

- Data Provenance, Metadata and Archiving
- Scripted and Documented Analysis
- Version Control
- Self Sufficiency
- Automation and Testing

Data Management and Provenance

- Track the data life cycle from raw collection through analysis
- Maintain metadata
- Use standardized data formats
- Store the data in *repositories*, private or public

Scripting and Documentation

- Recording the steps taken in data analysis via reusable *scripts*
- Create *literate* analyses that link written-up results with the original process used to produce them

Version Control

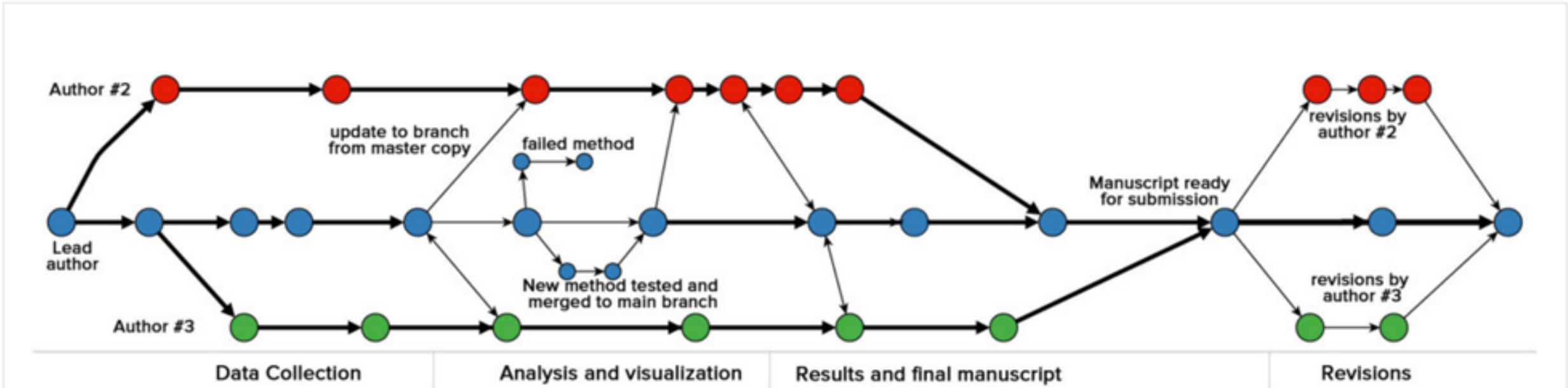
- Tracking the history and source of changes to a project

Google docs

Dropbox

git

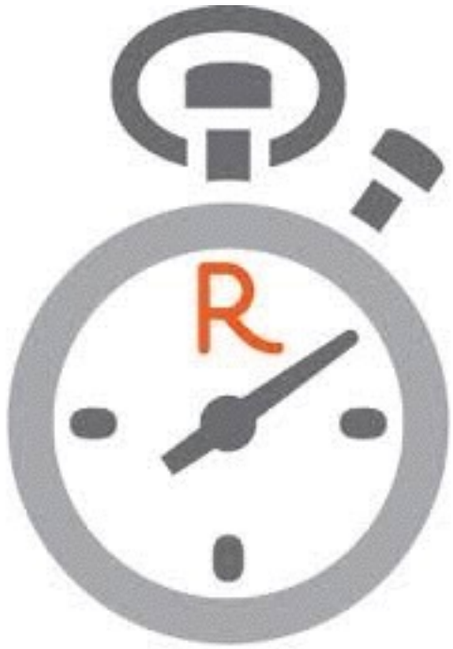
Version Control



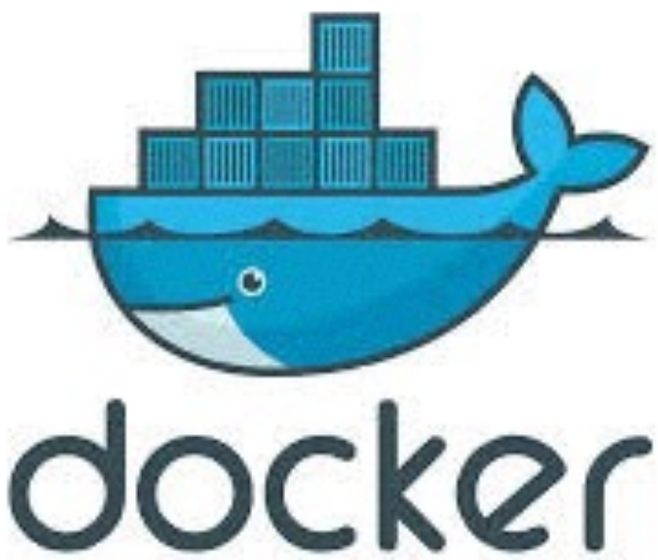
http://scfbm.biomedcentral.com/articles/10.1186/1751-0473-8-7

Self Sufficiency / Dependency Control

- Everything needed for reproducibility is packaged with the scientific output
- Documenting and storing the software and configuration used in analysis

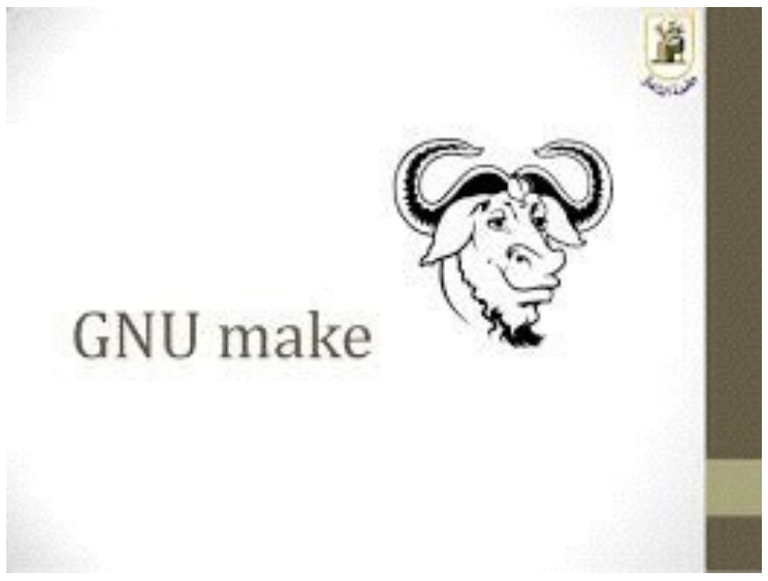


MRAN



Automation and Testing

- Link together all the steps of analysis and output creation
- Run efficiently so time- or resource-intensive steps are only repeated when needed
- Test that results are as they should be each time we make a change to data, analysis, our outputs



Jenkins

Travis CI

circleci

Karl -- this is very interesting, however you used an old version of the data (n=143 rather than n=226).

I'm really sorry you did all that work on the incomplete dataset.

Bruce

Caveats

- Learning curve can be steep
- Setting up reproducibility is not cost-free. It takes time and effort to set things up, especially at the beginning
- The tools are imperfect, and both tools and approach are largely designed by programmers, for programmers

What is most *useful* and *beneficial* to us, our collaborators and audiences?

- Data Provenance and Management
- Documentation/Literate Analysis
- Version Control
- Dependency Control
- Automation

What do we need to best reap the benefits?

- Training
- Dedicated time / personpower
- Guidelines / policies
- Paid tools / software
- Customized tools