UNIVERSITÄT DES SAARLANDES

Reproducibility for Everyone

Susann Auer

Slide deck

Handout: https://tinyurl.com/yckra8cb

Poster: https://doi.org/10.5281/zenodo.3641296

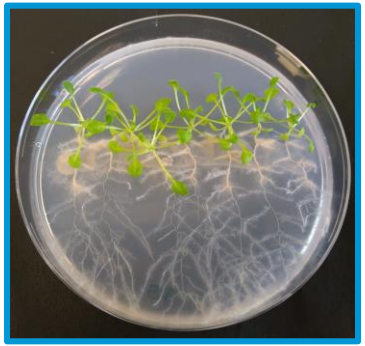
Feedback: https://forms.gle/kihnTyHEehrXCWWB7

@repro4everyone
hello@repro4everyone.org
https://www.repro4everyone.org

Replicate
Reproduce
Reuse

Hello!

I am Susann Auer



Plant Pathologist & Lecturer

Technische Universität Dresden, Germany

@SusannAuer

Replicate
Reproduce
Reuse

Read more about our initiative:

eLife FEATURE ARTICLE

SCIENCE FORUM

A community-led initiative for training in reproducible research

Abstract Open and reproducible research practices increase the reusability and impact of scientific research. The reproducibility of research results is influenced by many factors, most of which can be addressed by improved education and training. Here we describe how workshops developed by the Reproducibility for Everyone (R4E) initiative can be customized to provide researchers at all career stages and across most disciplines with education and training in reproducible research practices. The R4E initiative, which is led by volunteers, has reached more than 3000 researchers worldwide to date, and all workshop materials, including accompanying resources, are available under a CC-BY 4.0 license at https://www.repro4everyone.org/.

SUSANN AUER†, NELE A HAELTERMANN†, TRACEY L WEISSBERGER, JEFFREY C ERLICH, DAMAR SUSILARADEYA, MAGDALENA JULKOWSKA, MAŁGORZATA ANNA GAZDA, BENJAMIN SCHWESSINGER*, NAFISA M JADAVJI* AND REPRODUCIBILITY FOR EVERYONE TEAM

https://osf.io/dxw67/

https://elifesciences.org/articles/64719

Replicate
Reproduce
Reuse

Participants agree to follow the R4E Community Participation Guidelines

repro4everyone.org/pages/guidelines

DO

- ✓ Be respectful
- ✓ Give everyone a chance to contribute
- ✓ Use inclusive language
- ✓ Appreciate and accommodate differences
- ✓ Lead by example

DON'T

- ✗ Repeatedly interrupt or disrupt others
- ✗ Use sexual language or imagery
- ✗ Give unwelcome attention
- ✗ Bully, discriminate, or harass
- ✗ Make fun of personal appearance or choices

Have a concern?

- ➔ Report by telling the instructor.
- ➔ Report by emailing coc@repro4everyone.org or anonymously at https://forms.gle/UaxjwEYWVNoCDwJs5

Violations may result in removing of a participant.

Replicate
Reproduce
Reuse

When did you graduate from university?

https://cuckoo.team/R4E

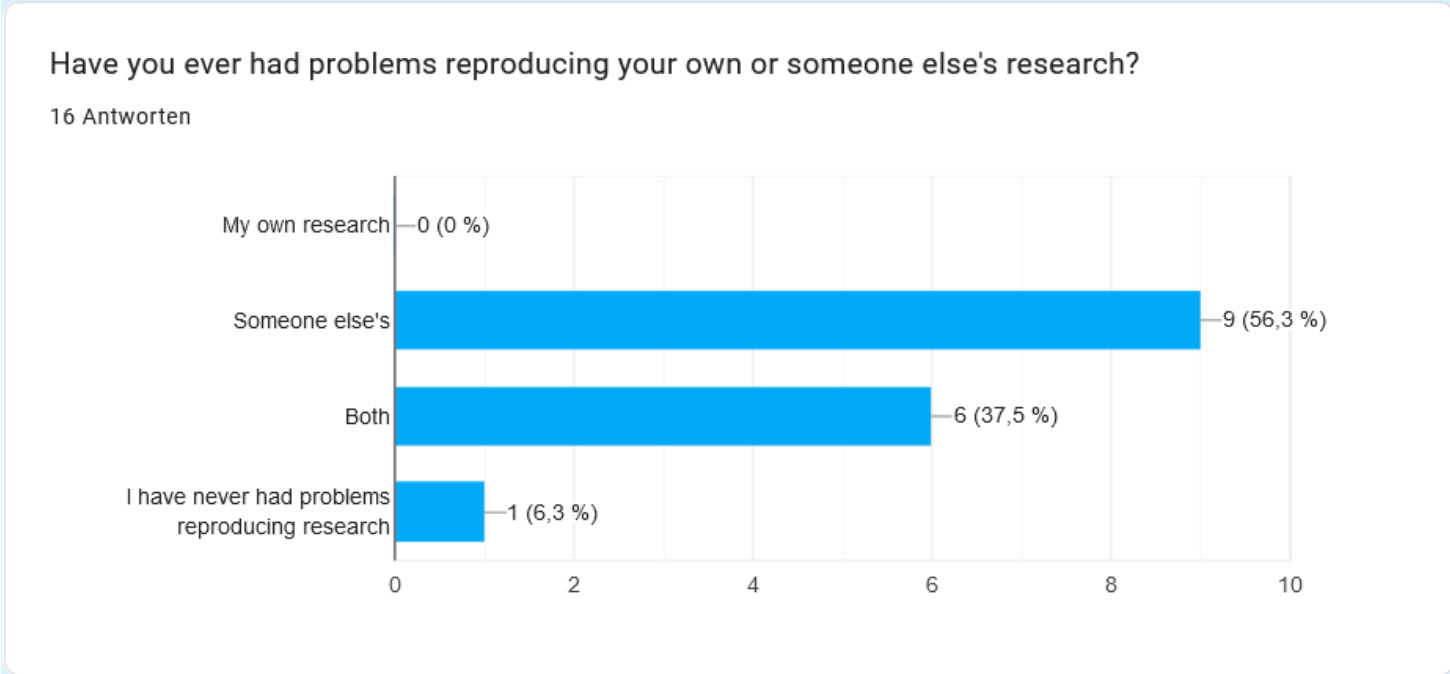
Slide deck: https://bit.ly/botanik22

Replicate
Reproduce
Reuse

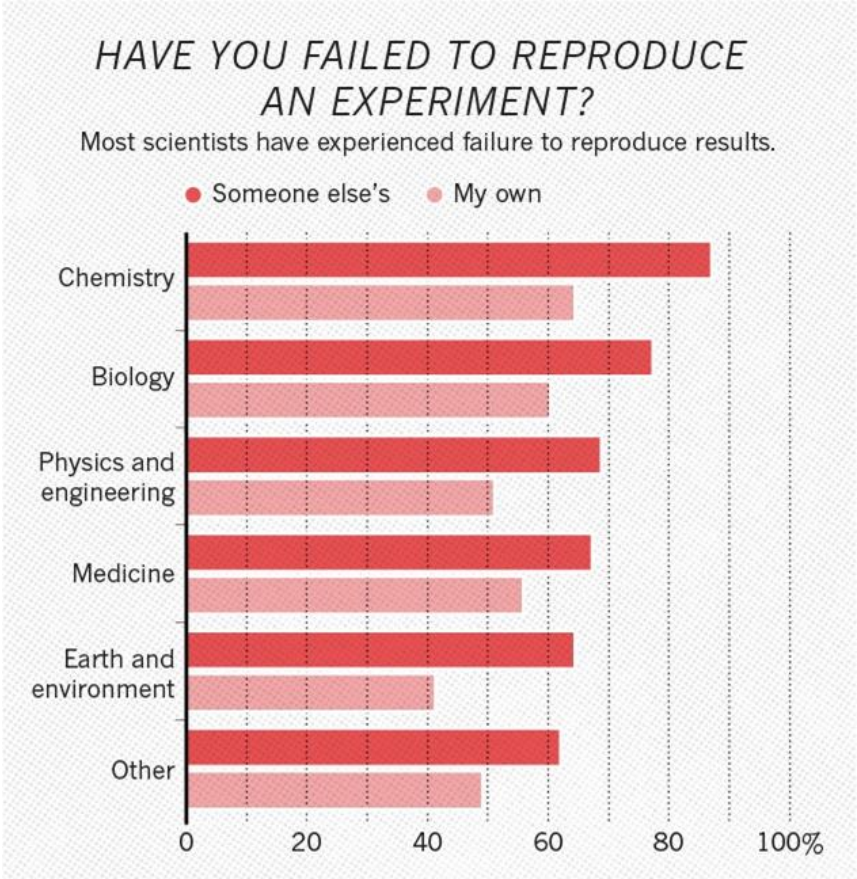
Introduction

Why does reproducibility matter to you?

Have you ever had problems reproducing your own or someone else's research?



Replicate
Reproduce
Reuse



1,500 scientists lift the lid on reproducibility, Monya Baker, *Nature*, 2016

Replicate
Reproduce
Reuse

Estimate of over $1 billion USD lost a year on irreproducible studies in US alone.

Freedman et al., 2015
(doi.org/10.1371/journal.pbio.1002165)

Mistakes in peer-reviewed papers are easy to find but hard to fix

Reproducibility: A tragedy of errors

David B. Allison, Andrew W. Brown, Brandon J. George & Kathryn A. Kaiser

https://www.nature.com/news/reproducibility-a-tragedy-of-errors-1.19264

Learning objectives

1) become familiar with the ‘Reproducibility’ framework
2) learn about ‘Reproducibility’ tools
3) be able to **implement one change** in your work routine starting tomorrow

today could be the starting point for a lifelong journey…
we are lifelong learners (:

Replicate
Reproduce
Reuse

We will look at...

- What does reproducibility mean?
- What are the different modes of reproducibility?
- Is reproducibility all that matters?
- Reproducibility tool shed:
  - Organization
  - Documentation
  - Analysis
  - Dissemination

Replicate
Reproduce
Reuse

What does reproducibility mean?

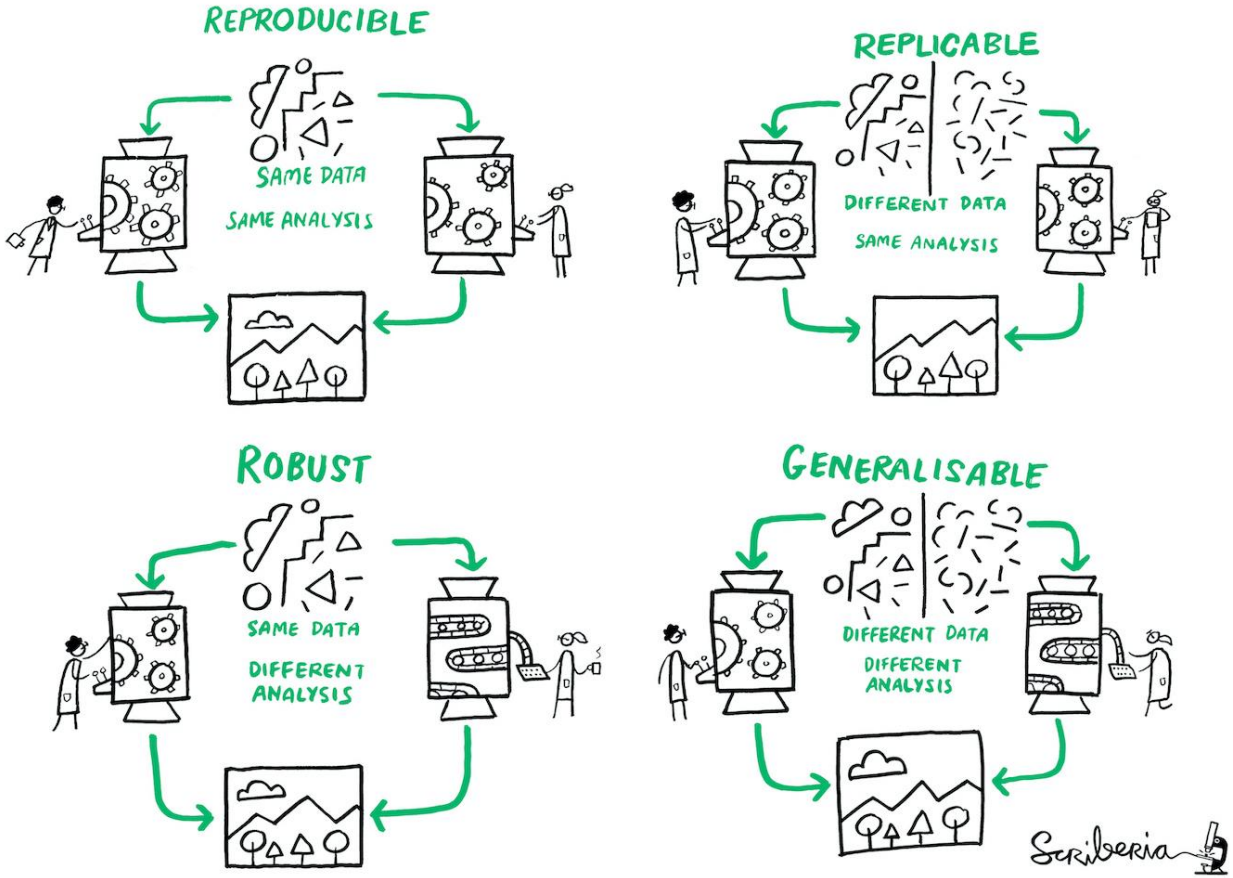
What does reproducibility mean?

Reproducible research: Authors provide all the necessary data and the computer codes to run the analysis again, re-creating the results.

Replication: A study that arrives at the same scientific findings as another study, collecting new data and completing new analyses.

Barba, 2018 (https://arxiv.org/abs/1802.03311)
Schloss, 2018 (10.1128/mBio.00525-18)

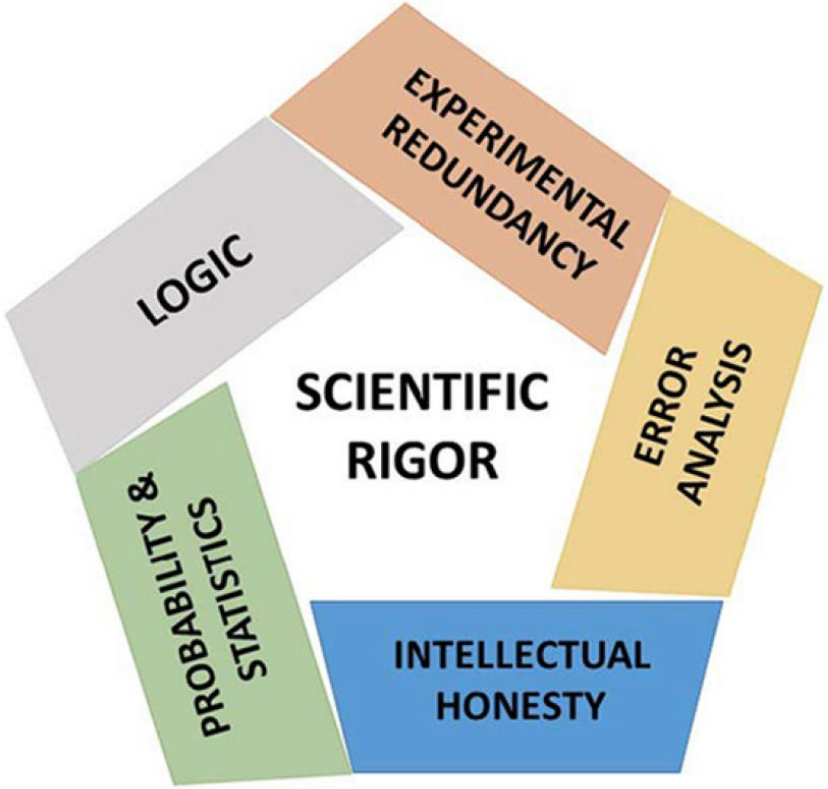
Replicate
Reproduce
Reuse



The Turing Way project illustration by Scriberia. DOI: [10.5281/zenodo.3332807](https://doi.org/10.5281/zenodo.3332807).

Replicate
Reproduce
Reuse

Is reproducibility all that matters?

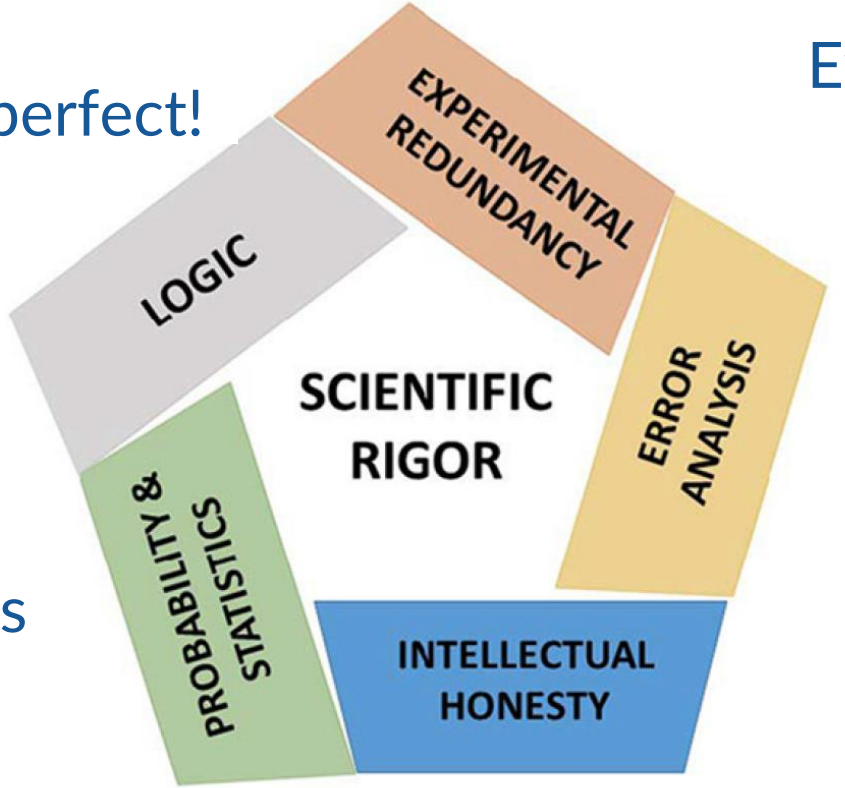


Casadevall and Fang, 2016 (10.1128/mBio.01902-16)

Replicate
Reproduce
Reuse

No one is perfect!

Every little bit helps!



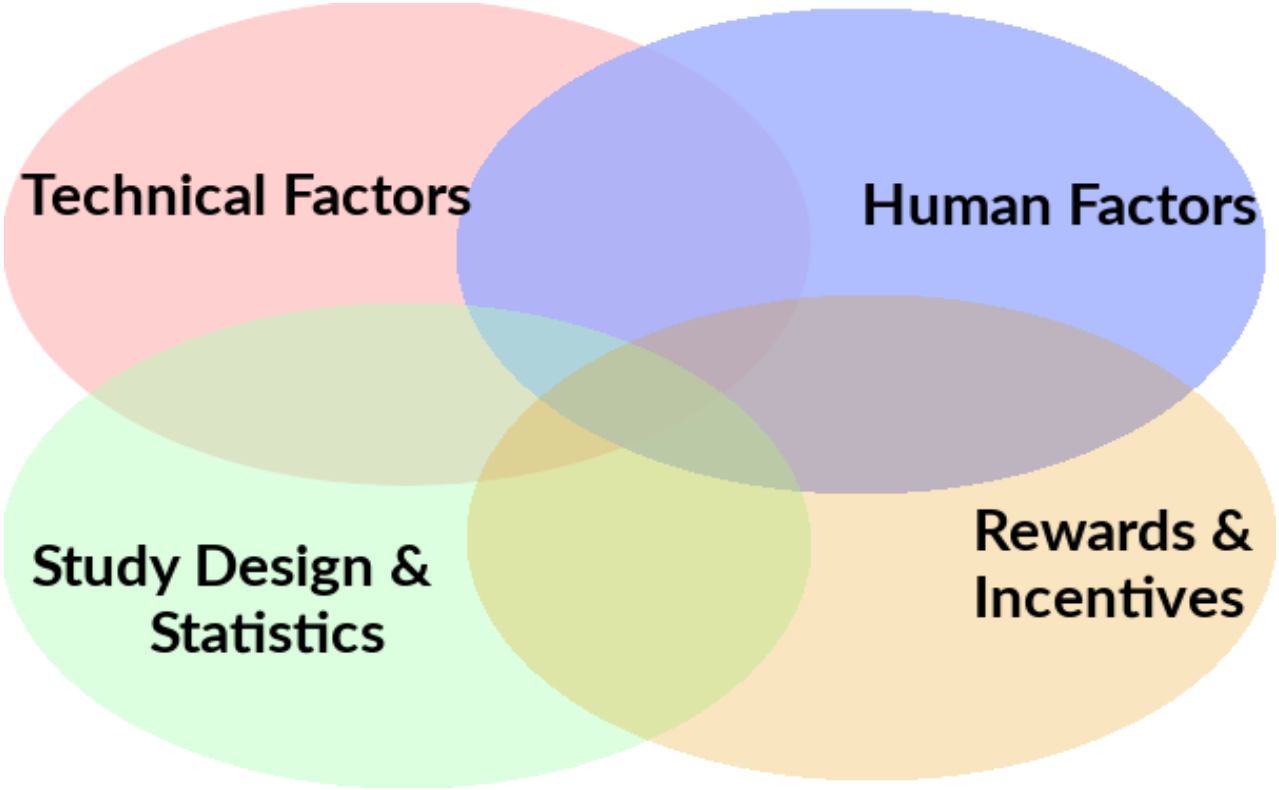
Everyone starts somewhere!

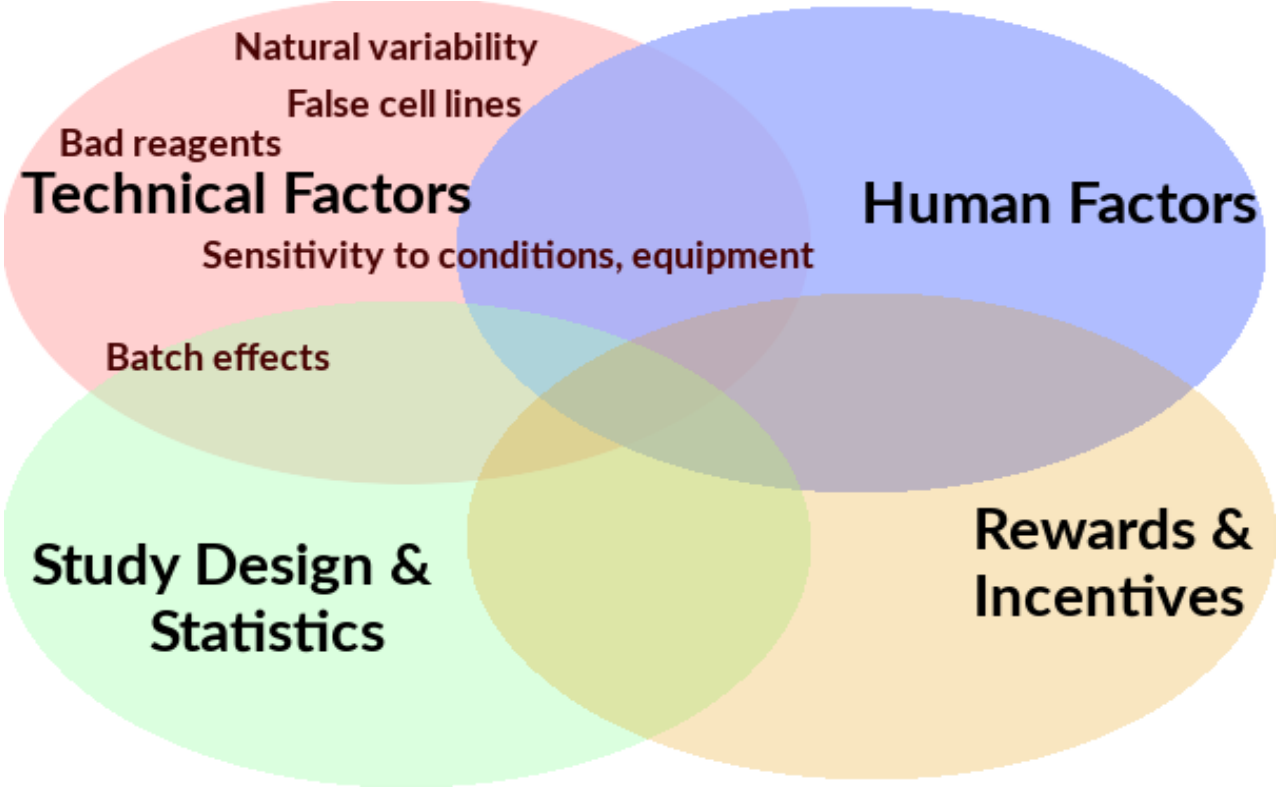
Transparent and open science!

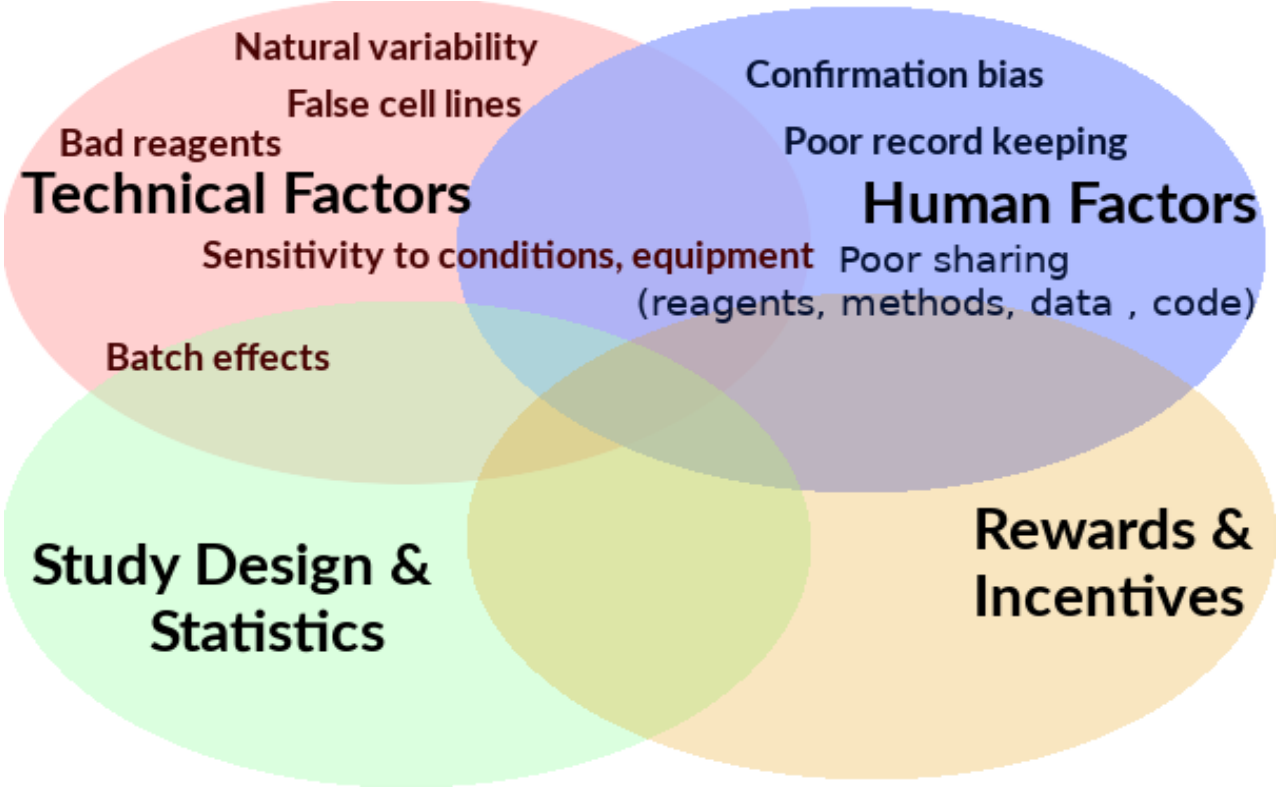
Casadevall and Fang, 2016 (10.1128/mBio.01902-16)

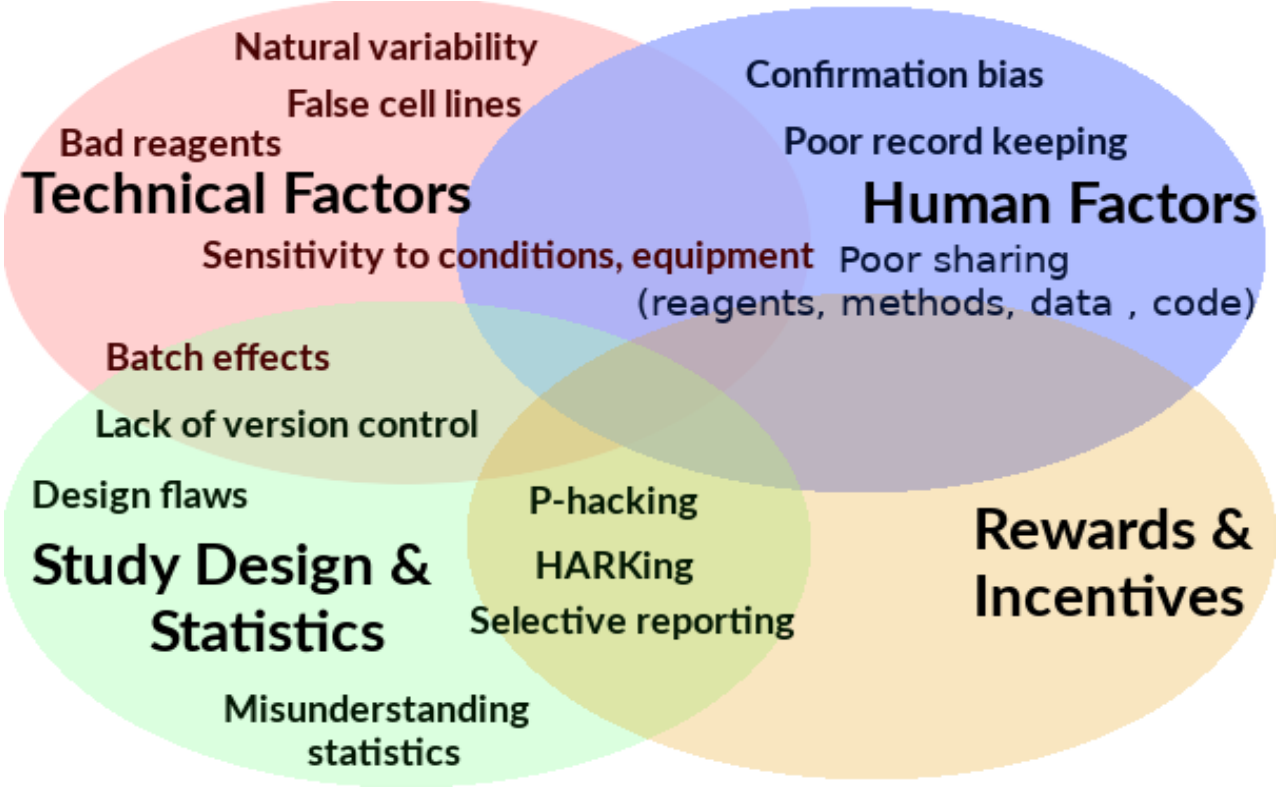
Replicate
Reproduce
Reuse

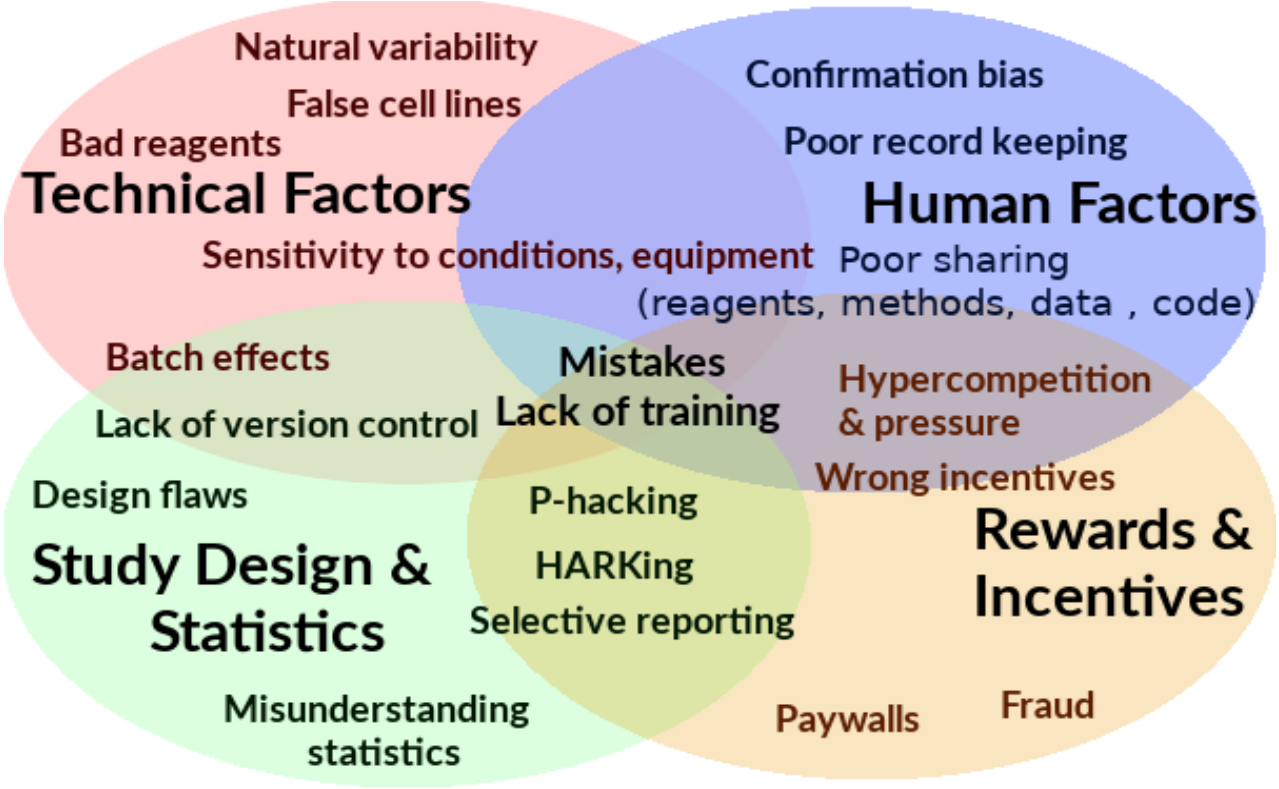
Factors decreasing reproducibility



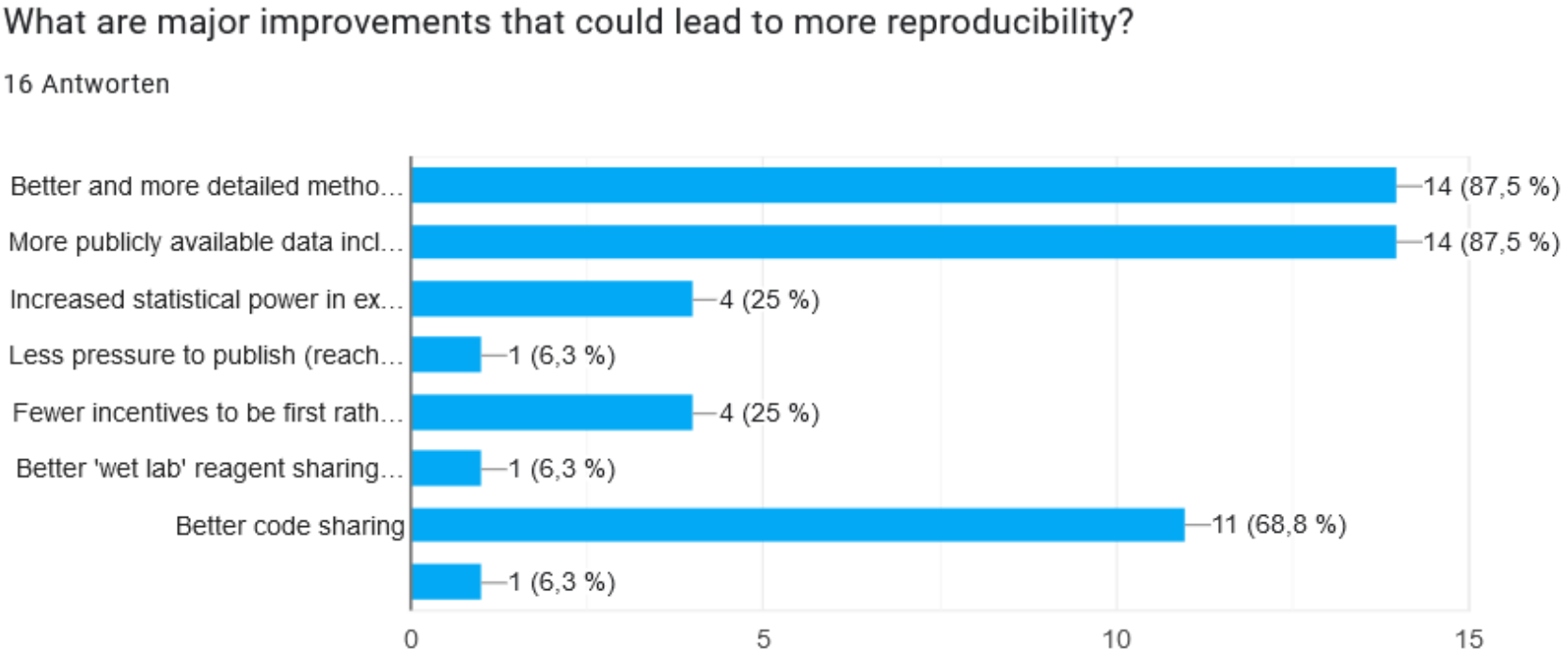

Natural variability
False cell lines
Bad reagents
Technical Factors
Sensitivity to conditions, equipment
Batch effects
Human Factors
Study Design & Statistics
Rewards & Incentives


Natural variability
False cell lines
Bad reagents
Technical Factors
Sensitivity to conditions, equipment
Batch effects
Confirmation bias
Poor record keeping
Human Factors
Poor sharing
(reagents, methods, data , code)
Study Design &
Statistics
Rewards &
Incentives


Natural variability
False cell lines
Bad reagents
Technical Factors
Sensitivity to conditions, equipment
Confirmation bias
Poor record keeping
Human Factors
Poor sharing
(reagents, methods, data , code)
Batch effects
Lack of version control
Design flaws
Study Design & Statistics
P-hacking
HARKing
Selective reporting
Misunderstanding statistics
Rewards & Incentives


Natural variability
False cell lines
Bad reagents
Technical Factors
Sensitivity to conditions, equipment
Confirmation bias
Poor record keeping
Human Factors
Poor sharing
(reagents, methods, data , code)
Batch effects
Lack of version control
Mistakes
Lack of training
Hypercompetition
& pressure
Wrong incentives
Design flaws
Study Design &
Statistics
P-hacking
HARKing
Selective reporting
Rewards &
Incentives
Misunderstanding
statistics
Paywalls
Fraud

Factors improving reproducibility



Replicate
Reproduce
Reuse

“Respondents said that the most important element that would enable the better reproducibility of published research would be that authors describe methods and analyses in detail.”

Samota and Davey, 2018
(https://doi.org/10.1101/581033)

Replicate
Reproduce
Reuse

Where is your greatest potential for growth?

Replicate
Reproduce
Reuse

- Everyone starts somewhere!
- Every little bit helps!
- No one is perfect!

- Everyone starts somewhere!
- Every little bit helps!
- No one is perfect!

- **Adopting some of these best practices isn't just good for other scientists…**
- **It's good for you & will save you time in the long term!**

Data Management

Are you familiar with these feelings?

I cannot find this file!

What did I call it again??

Where is my file??

Was it the wild type picture or the mutant one?

What version was it??

Where is my RAW!!! data???

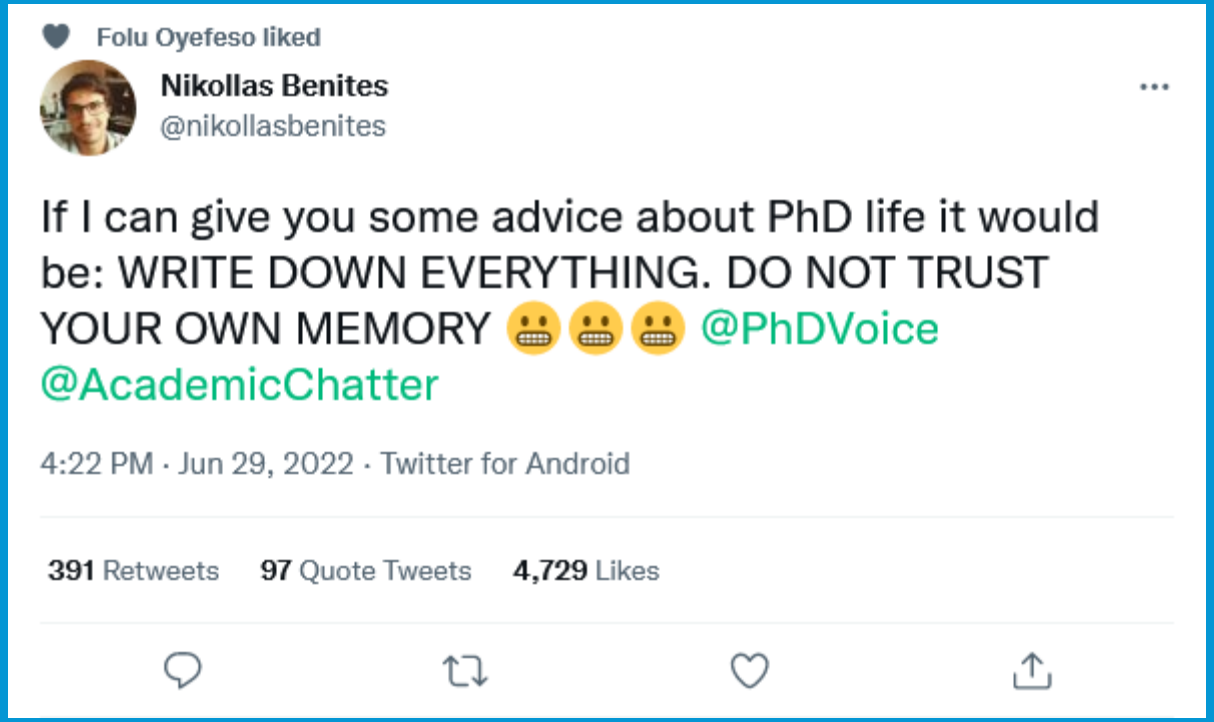
Replicate
Reproduce
Reuse

Make a plan!
Be happy!

DMPTool
Build your Data Management Plan

No need to become a science superhero overnight!

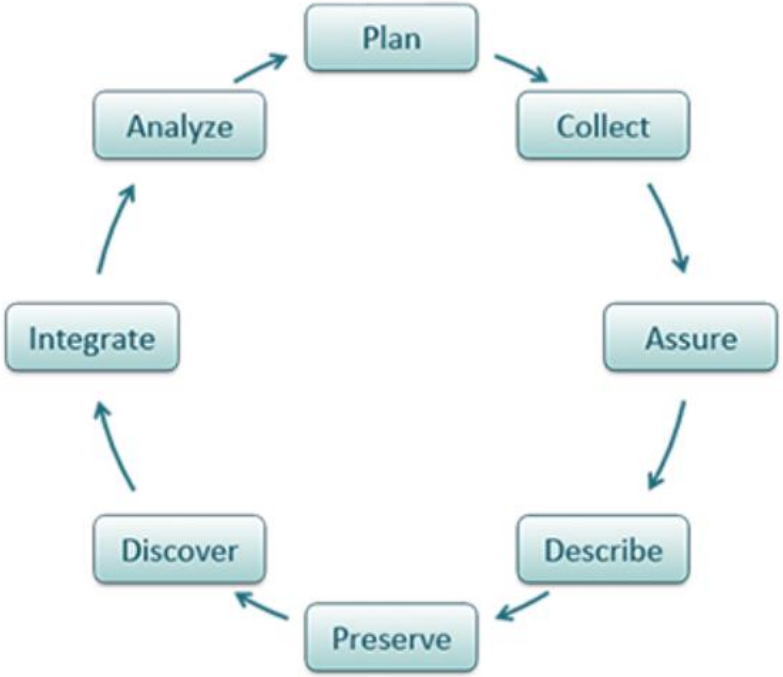
Replicate
Reproduce
Reuse



Replicate
Reproduce
Reuse

Think about…

- **What** data will be produced as a part of the project?
- **How** each type of data will be organized, documented, standardized, stored, protected, shared and archived?
- **Who** will take responsibility for carrying out the activities listed above, and
- **When** these activities will take place over the course of the project (and beyond)?
- **Metadata**



guides.lib.purdue.edu/c.php?g=353013&p=2378292, www.dataone.org/best-practices

Replicate
Reproduce
Reuse

Project directory structure

Project_1

- methods
- raw_data
- analysis
- scripts
- manuscript
- readme and/or ELN link

- **Develop an informative directory structure**
- **Keep research materials together**

Inspired by 'Bioinformatic data skills' by Vincent Buffalo

Replicate
Reproduce
Reuse

Project directory structure

Project_1

- methods
- raw_data
  - readme
- analysis
  - analysis_method_1
    - 2017
    - 2018
  - analysis_method_2
- scripts
- manuscript
  - text
    - version_1
- readme and/or ELN link

Specific content in each category for Project #1

Raw data,
Data analysis, and
Manuscript

Inspired by 'Bioinformatic data skills'
by Vincent Buffalo

Replicate
Reproduce
Reuse

Project directory structure

- **Project_1**
  - **methods**
  - **raw_data**
    - readme
  - **analysis**
    - analysis_method_1
      - 2017
      - 2018
    - Analysis_method_2
  - **scripts**
  - **manuscript**
    - text
      - Version_1
  - **readme and/or ELN link**

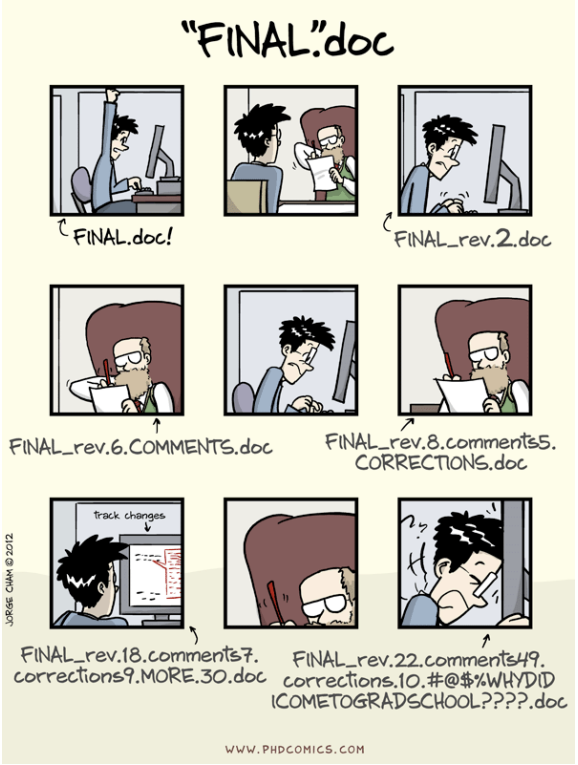
- Always keep **raw** data!
- Always backup data

(3x and synchronized: 3 unique locations - cloud, server, personal drive)

Inspired by 'Bioinformatic data skills'
by Vincent Buffalo

File naming conventions

- What did you call the last file you generated?
- Did you have rules?
- Where can you (and other people) find it 6 months from now?



http://guides.lib.purdue.edu/c.php?g=353013&p=2378292
http://kbroman.org/dataorg/

Replicate
Reproduce
Reuse

File naming conventions

The rules don't matter; that you have rules matters

- Include date in yyyy-mm-dd format
- Use meaningful abbreviations
- Have group identifiers
- Document your decisions
- Be consistent
- Use version numbers

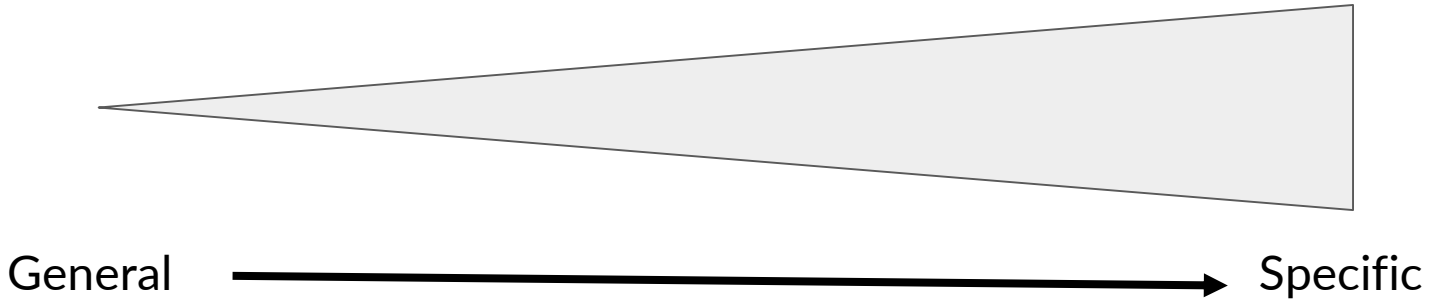
guides.lib.purdue.edu/c.php?g=353013&p=2378292, kbroman.org/dataorg/

File naming conventions

- Example

  20130825_DOEProject_Ex1Test1_Data_Gonzalez_v3-03.xlsx

  Date **Project** **Experiment** **Type** **ID** **Version**



http://guides.lib.purdue.edu/c.php?g=353013&p=2378292

Replicate
Reproduce
Reuse

What should Ben do?

Ben is really excited to join a new team that is performing a screen of plant hormones on root architecture. However,

- The previous Postdoc started a new job and refuses to respond to his emails.
- The technician on the project was only involved in the data acquisition steps.
- Unfortunately, the lab notebook went missing in a recent move to a new floor.
    - The methods section in a previous paper reads like this -

Materials and Methods
Plants were grown on appropriate media and roots photographed. Images were analyzed using WinRhizo (Arsenault, J-L., et al. 1995) and data presented as graphs.

(1) Identify the problem(s)

(2) Suggest solutions that could have prevented these problems in the first place.

https://cuckoo.team/R4E

Replicate
Reproduce
Reuse

Electronic Notebooks

Paper Lab-notebooks - in use since the 15th Century!

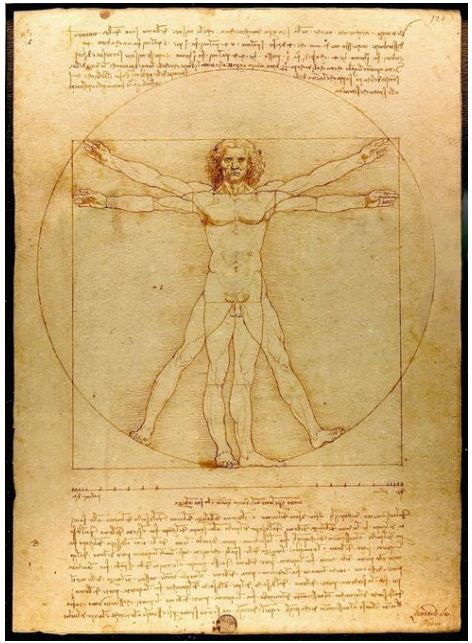
Good record keeping is important for:

- Dissemination of ideas, findings,
- Legally binding records that protects intellectual property

Not searchable!

Can be easily damaged, misplaced
Not easy to back up

Hard to share with collaborators



Leonardo da Vinci's notebook, Codex Arundel c. 1458-1518
British Library

Replicate
Reproduce
Reuse

Why should you use an Electronic Lab Notebook?

Search

Easily searchable both in online and offline searches

Export

Data can be exported as PDF (must backup regularly)

Share

Easily shareable with collaborators and broader community

Embed

Can embed high res images, protocols, & more

Access

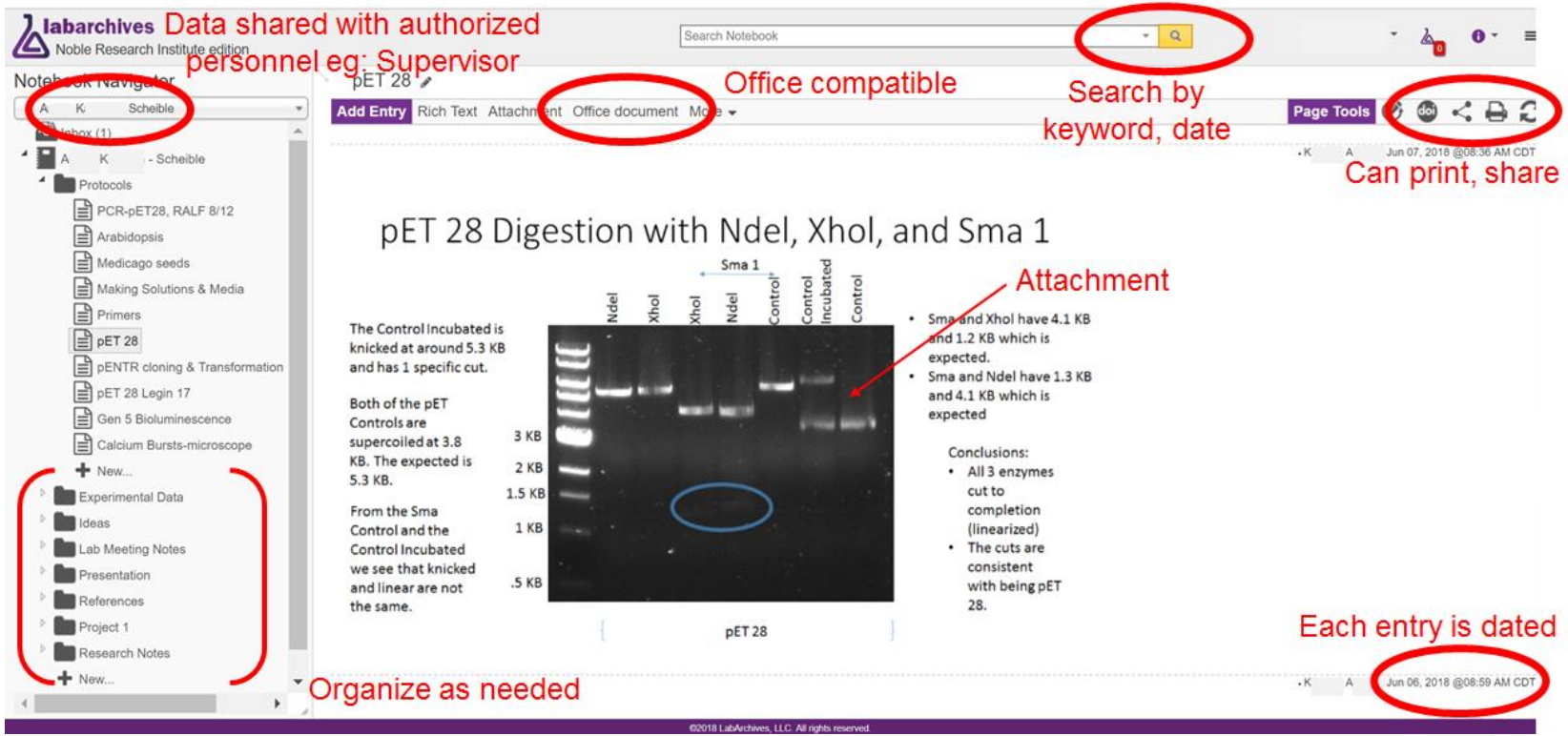
Easily accessible globally (depend on connectivity)

Mobile

Use the mobile app to quickly upload images

Replicate
Reproduce
Reuse

Basic features of an Electronic Lab Notebook



Replicate
Reproduce
Reuse

Cost considerations - Available Products

- Paid for - Bio-Itech, LabArchives, LabGuru
- Paid (with free version) - SciNote, Benchling
- Open source - Open wet ware, ELOG
- Free - Open Science Framework (OSF.io), LocalWiki

Kanza, 10.1186/s13321-017-0221-3

One size does not fit all

Parameters to consider

Available lab notebooks

ELN Features Matrix

Features	Specifications															
	Benchling	Biovia	Confluence	Docollab	ECL	ELOG	Evernote	Exemplar	Findings	Hivebench	IDBS	LabArchives	LabCollector	LabWare	LabVantage	Lab
Interactivity																
Intuitive Interface Design	✓	No response received	★	★	No response received	★	No response received	No response received	★	★	★	✓	No response received	★	★	★
Auto Metadata Harvest	★	No response received	✗	✓	No response received	✗	No response received	No response received	✗	✓	★	✓	No response received	✗	✓	✓
Search functions can search across file formats and beyond typos	★	★	★	★	No response received	★	★	★	★	★	★	★	No response received	★	★	★
Ability to manipulate files and images	★	No response received	★	★	No response received	★	No response received	★	★	★	★	★	No response received	★	★	★
Support for multiple open windows	✓	★	✓	✓	No response received	✓	✓	★	★	✓	✓	✓	No response received	✓	★	★
Ability to link out	✗	No response received	★	✓	✓	✓	✓	✓	✓	✓	✓	✓	No response received	✓	✓	✓
Support for Researcher Documentation																
Hyperlink support	✓	No response received	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	No response received	✓	✓	✓
Metadata Creation Prompts	✓	No response received	✗	✓	No response received	✓	✗	No response received	✗	✓	✓	✓	No response received	✗	✓	✓
Rights Management (licensing)	★	No response received	★	✓	No response received	★	No response received	No response received	✓	✓	★	✓	✓	✓	✓	✓
Protocol Integration	✓	★	✓	✓	No response received	✓	★	★	✓	✓	✓	✓	★	✓	★	✓
Adaptability to Lab workflows																
Accounts/Permissions Levels	✓	No response received	★	✓	✓	✓	✓	★	★	✓	✓	✓	✓	✓	★	✗
Internal Data Sharing	✓	★	★	✓	No response received	✓	✓	No response received	✓	✓	✓	✓	✓	✓	✓	★
Adaptable to a Variety of Workflows	★	No response received	★	★	No response received	★	No response received	★	★	★	★	★	★	★	★	★
Compatibility with authoring tools	✓	No response received	★	✓	No response received	✗	No response received	No response received	✗	★	★	✓	No response received	✗	★	✗
Windows Compatible	✓	No response received	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓
Macintosh Compatible	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	★
Linux Compatible	✓	✗	✓	✓	No response received	✓	No response received	✓	✗	✓	✓	✓	No response received	✓	✗	★
Android Compatible	✓	✓	✓	✓	No response received	✓	✓	✓	✗	✗	✓	✓	✓	✓	✓	✗
iOS Compatible	✓	✓	✓	✓	No response received	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
Storage																
Cloud Storage	✓	No response received	✗	✓	No response received	✓	No response received	No response received	✓	✓	✓	✓	✓	✓	✓	✗
Local Storage	✗	No response received	✓	✗	No response received	✓	No response received	No response received	✓	✗	✓	✓	No response received	✓	✓	✓
Hybrid (cloud/local) Storage	✗	No response received	✗	✗	No response received	✗	No response received	No response received	✓	✓	✗	✓	No response received	✗	✗	✗
Versioning	★	★	★	★	No response received	★	No response received	★	★	★	★	★	★	★	★	★
File Redundancy	★	No response received	★	★	No response received	★	No response received	No response received	★	★	★	★	No response received	★	★	★
Creates stable URLs or persistent identifiers for entries	✓	No response received	✓	✓	No response received	✓	No response received	No response received	★	✓	✓	✓	No response received	✓	✓	✓
Can unregistered users access the data found at persistent links?	✓	No response received	✓	✗	No response received	✗	No response received	No response received	★	★	✗	✗	No response received	✗	✗	✗
Storage Capacity - Users	★	No response received	★	★	No response received	★	★	No response received	★	★	★	★	No response received	★	✓	★
Storage Capacity - Max File Size	★	No response received	★	★	No response received	★	No response received	No response received	★	★	★	★	No response received	★	✓	★

https://datamanagement.hms.harvard.edu/electronic-lab-notebooks

Replicate
Reproduce
Reuse

General tips on electronic record keeping

- Back-up data regularly
- Maintain a physical observation notebook in parallel
- Mobile apps provide added portability
- If using free ELNs, check privacy and data ownership policies

Organizing and Sharing Protocols

“

Has anyone ever educated you on the best practices on scientific record keeping?

Replicate
Reproduce
Reuse

Being able to find optimized protocols for your organism / tissue of choice can save you months!

https://medium.com/@tpoi

Replicate
Reproduce
Reuse

Lab member

I create them myself

Literature

Where do you find your protocols?

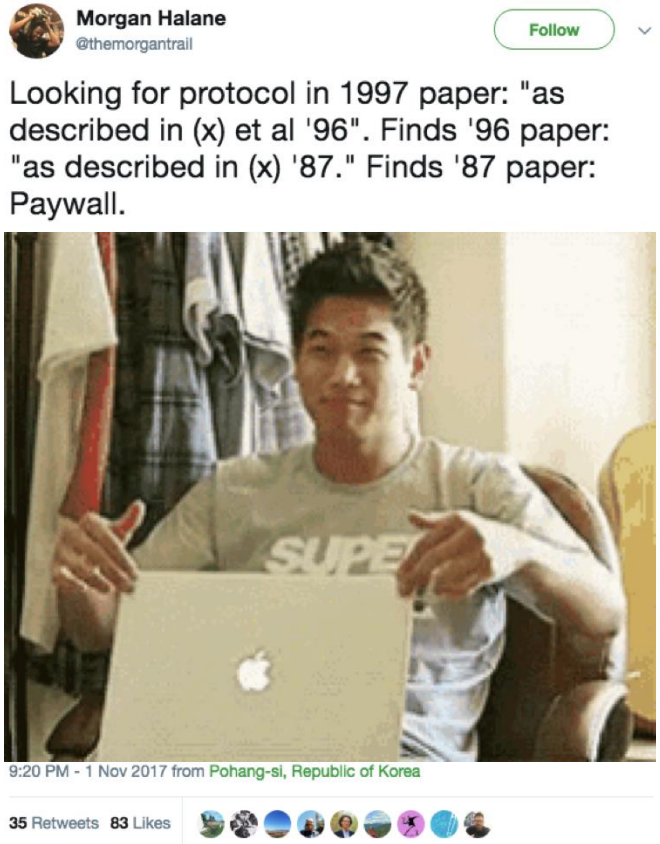
Repositories

Researchgate

Contacting an expert in the field

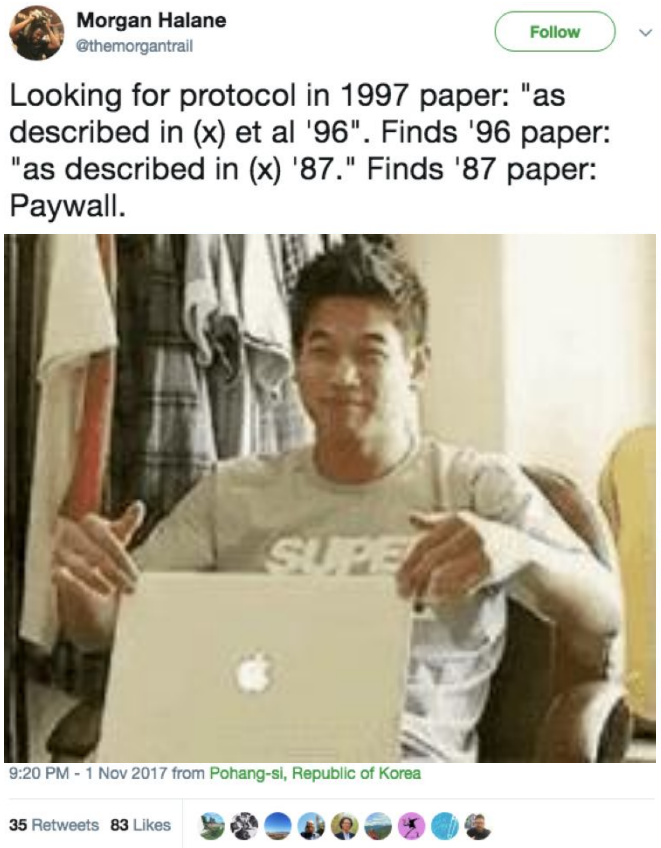
Replicate
Reproduce
Reuse

Description Unavailable



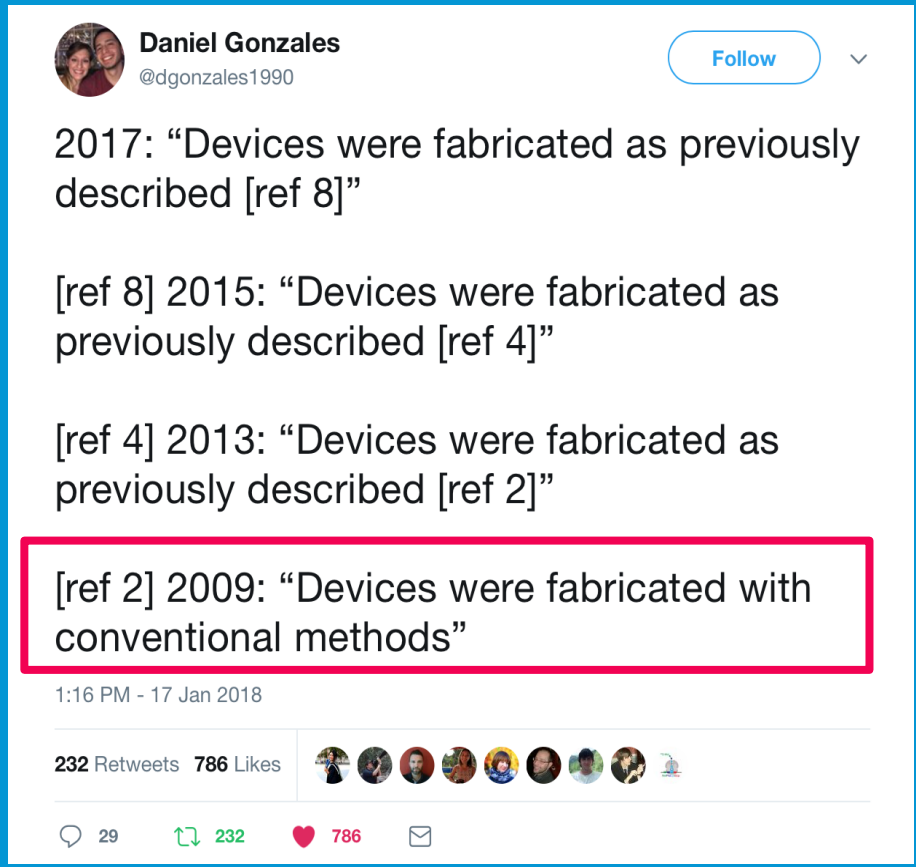
Replicate
Reproduce
Reuse

Description Unavailable



Replicate
Reproduce
Reuse

Description Ambiguous



Timothée Poisot Follow
Ecologist. Not that kind of doctor.
Sep 8, 2015 · 2 min read

Step 2—do the rest of the [redacted] analysis

How to draw an owl

1.

1. Draw some circles

2.

2. Draw the rest of the [redacted] owl

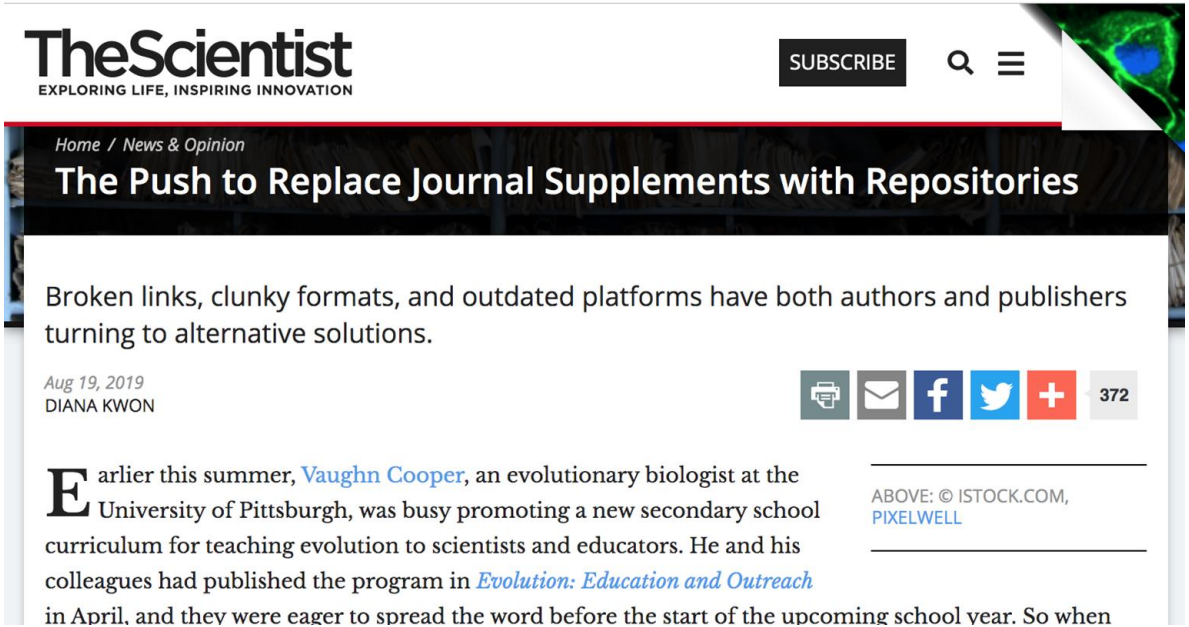
So when starting a new research project, one can feel like one is trying to draw an owl using the above tutorial. This is because we tend to learn about methods by reading papers, and the Methods section of any given paper is often, to put it mildly, *terse*. To pursue the *How to draw an owl* analogy, a Methods section could read

We draw the owl on 60.2 gsm white paper of the A4 dimension (210mm by 297mm), using 3H and 6B graphite pencils (Derwent, Cumbria, UK). We did so by looking at owls, and drawing what we saw on paper. This protocol yielded one drawn owl.

https://medium.com/@tpoi

Replicate
Reproduce
Reuse

Use repositories, not supplemental files



TheScientist
EXPLORING LIFE, INSPIRING INNOVATION

SUBSCRIBE

Home / News & Opinion

The Push to Replace Journal Supplements with Repositories

Broken links, clunky formats, and outdated platforms have both authors and publishers turning to alternative solutions.

Aug 19, 2019
DIANA KWON

372

Earlier this summer, Vaughn Cooper, an evolutionary biologist at the University of Pittsburgh, was busy promoting a new secondary school curriculum for teaching evolution to scientists and educators. He and his colleagues had published the program in *Evolution: Education and Outreach* in April, and they were eager to spread the word before the start of the upcoming school year. So when

ABOVE: © ISTOCK.COM, PIXELWELL

Replicate
Reproduce
Reuse

Protocol drawing exercise

Work status

busy	pink
done	green

1. Use a pen and a piece of paper.
2. Imagine a typical computer. Everyone knows how a computer looks like. Draw the computer, don´t share it.

5 mins

3. Write detailed instructions on another piece of paper.

5-10 mins

4. Share your instructions with us.
5. Draw a new computer according to the instructions you picked.

5 mins

6. Share the instructions and your drawings with us.
7. Be amazed by wonders (:

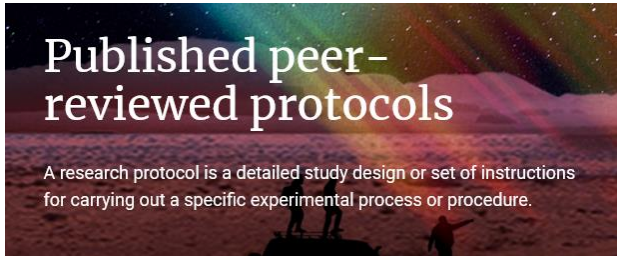
Detail & Digitize Protocols

- Think of protocol as brief, modular, self-contained scientific publication
- Include 3-4 sentence abstract that puts methodology in context
- Include as much detail as possible
  - duration/time per step
  - reagent amount
  - vendor name
  - catalogue number
  - secret sauce
- Expected result
- Safety information
- Software package
- Chronology of steps
- Notes, recipes, tips and tricks
- Use protocols.io, Google Docs, and/or ELN (not paper or MS Word) - **need versioning**

https://www.protocols.io/view/how-to-make-your-protocol-more-reproducible-discov-g7vbzn6

https://www.aje.com/en/arc/how-to-write-an-easily-reproducible-protocol/

Protocol repositories



https://plos.org/protocols/

https://www.protocols.io

Operational procedures, instructions/manuals, computational workflows, safety checklists, biology, chemistry, clinical trials…

https://www.nature.com/nprot/protocolexchange

Resource and reagent sharing

Have you ever shared resources like reagents or other physical resources?

Have you ever used other people's resources? What was it?

Problems with wet-lab reagent availability

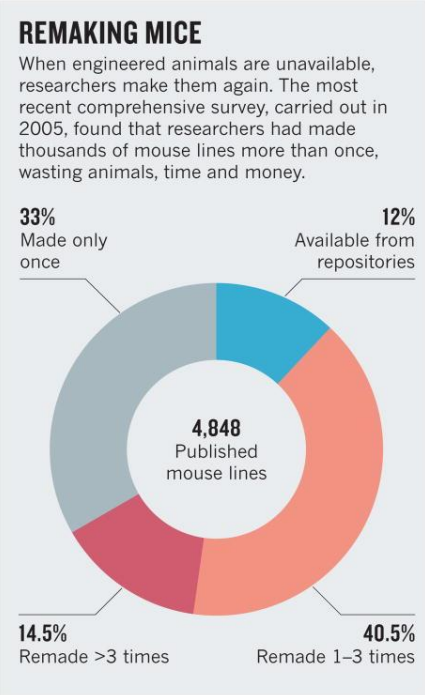
Scientist makes and publishes a reagent **“available upon request”**

Scientist leaves the lab and stores published reagent in freezer

Other scientists request the reagent, but no one else remembers where it is stored!

Replicate
Reproduce
Reuse

Problems with wet-lab reagent availability



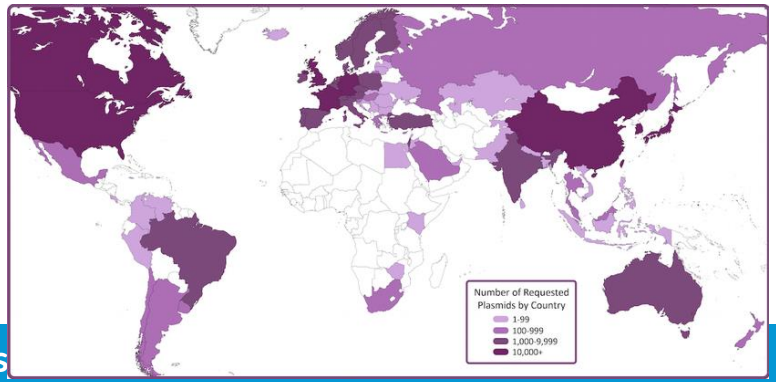
- In 2005, NIH Study found that half of the mouse lines had been remade at least twice
- Only 12% were available from repositories
- It can take years and cost $$$ for researchers to make a mouse strain.

Lloyd et al. Nature 2015 10.1038/522151a

Functions of reagent repositories

They:

- Verify reagents
- Curate reagents
- Facilitate and track shipping
- Protect IP



Process is easier if you:

- Record how a reagent was created
- Provide associated publications
- Provide associated protocols

(All of these are facilitated by other tools discussed in this workshop)

Benefits of depositing reagents

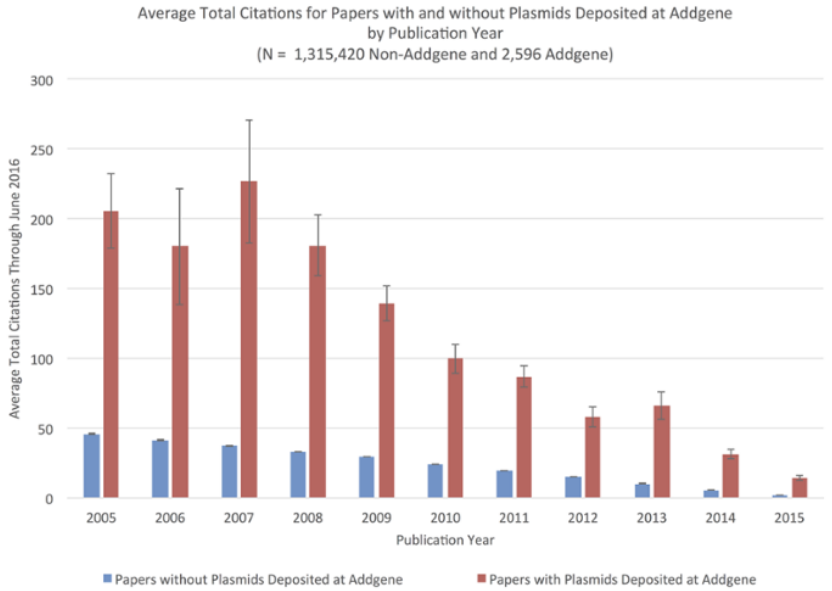
DIRECT

- Archiving
- Reducing time spent sending out reagents
- Authentication/Quality Control

INDIRECT

- Creation of educational content
- Direct promotion
- Analysis of reagent distribution

Figure 1



https://blog.addgene.org/addgene-depositors-get-more-citations

Replicate
Reproduce
Reuse

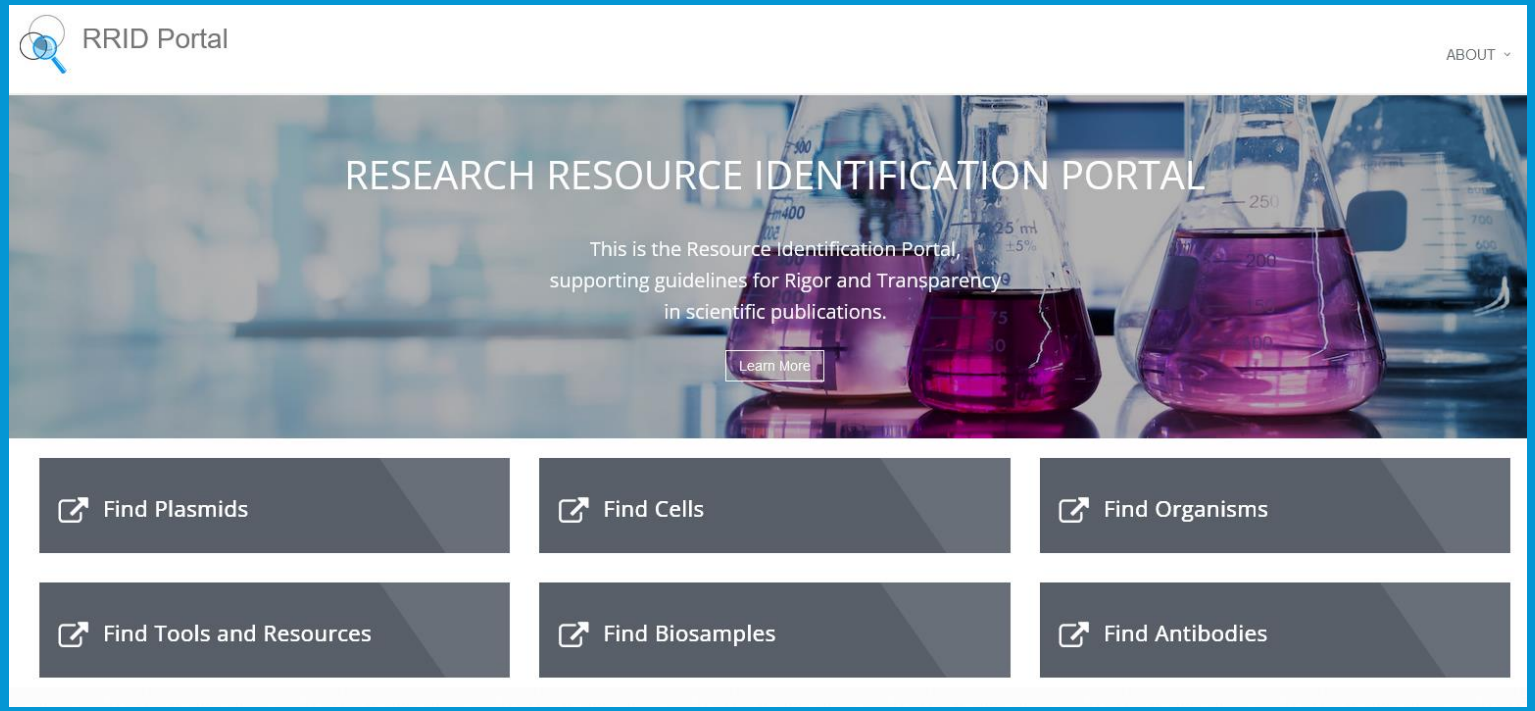
How to improve resource sharing & reporting

- Record how a resource was created
- Authenticate!
- Provide associated publications & protocols
- When naming resources you make, use descriptive and standardized naming conventions
- Record and include the reagent's catalog # or RRID in your lab notebook and manuscript(s)
- Deposit reagents & resources you make with a repository!

https://scicrunch.org/resources

Replicate
Reproduce
Reuse

Register an RRID for your resource



https://scicrunch.org/resources

Replicate
Reproduce
Reuse

Examples of reagent/material repositories

- Chemistry repositories:
  - https://www.nfdi4chem.de/de/chemie-repositorien/
- Open Science repository chemistry
  - http://www.open-science-repository.com/journal-of-chemistry.html
- EMBL European Bioinformatics Institute
  - https://www.ebi.ac.uk/
- Material Sciences
  - NIST Materials Data repository
  - NoMad repository
  - Materials Cloud
- Language Science
  - Linguistik-Repository (GER)
  - LingBuzz
  - Semanticsarchive.net

Do a quick internet search for a repository in your field of interest. How many do you find?

Data and code sharing

„We drown in data but cannot find them.“

NFDI - @NFDI@nfdi.social
@NFDI_de

"Wir ertrinken in Daten, können sie aber nicht finden."
Direktor @ysurevetter erzählt im Interview mit der @sz über #NFDI und die "Veredelung von #Forschungsdaten.

Translate Tweet

sueddeutsche.de
Wie Wissenschaftler Daten in Zukunft teilen wollen
Die Wissenschaft erzeugt massenhaft Daten. Ein weitgehend ungenutzter Schatz, den eine Nationale Forschungsdateninfrastruktur nun heben will

5:52 PM · Jul 10, 2023 · **4,224** Views

40 Retweets **5** Quotes **60** Likes **2** Bookmarks

Make your research data accessible and findable!

Replicate
Reproduce
Reuse

What to share?

- Data & code necessary to **validate findings** & **reproduce results**
- Data & code that might be **valuable** to other researchers/policy makers
- Data & code which **cannot be (easily) re-generated**

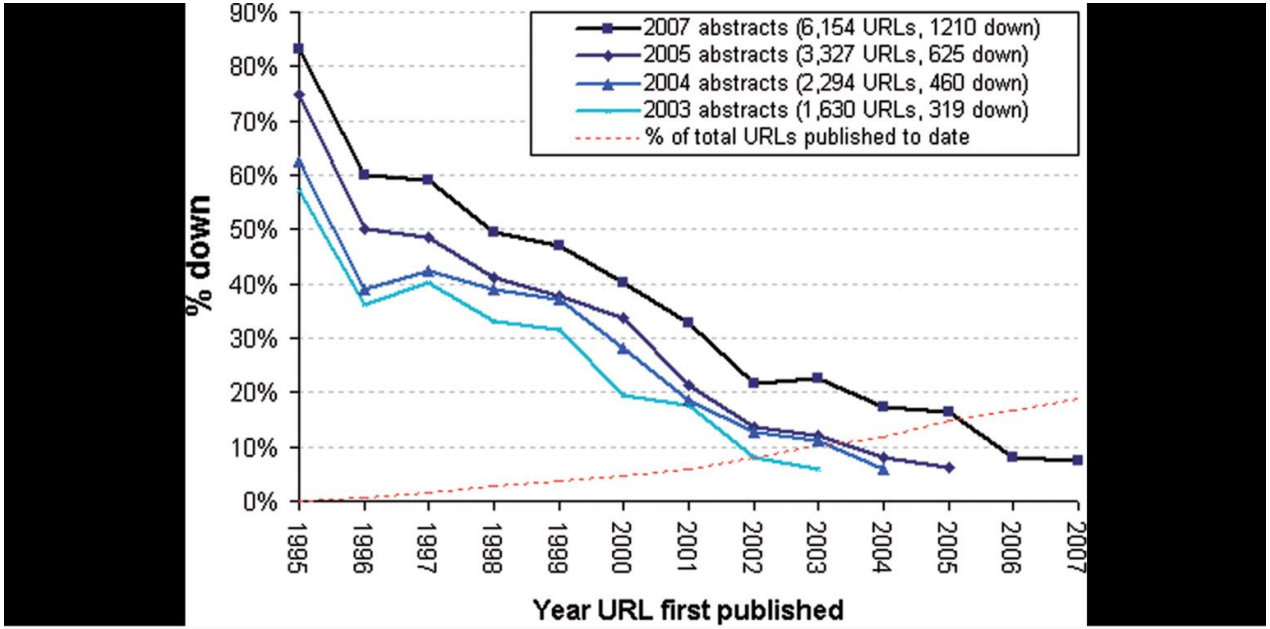
Why share?

- Funder or publisher mandates
- Citation benefits (Piwowar 2013, https://doi.org/10.7717/peerj.175)
- Preserve long-term access to data

How to share?

- Choose open, persistent, and non-proprietary **file formats**
- Create and share **documentation** to enable reuse
- Include **data citations** of source data
- Create rich **metadata**

Data sharing – don´t use websites



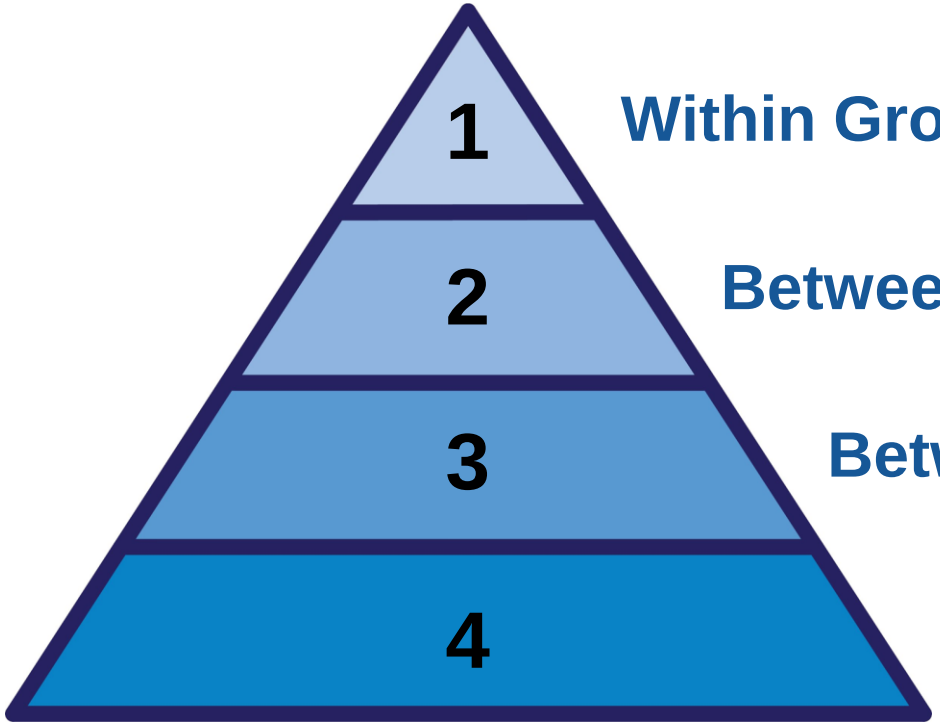
Time-dependent decay of URLs published in MEDLINE abstracts. Surveys taken in 2004, 2005 and 2007 are compared to the original 2003 survey. The number of URLs published per year is displayed as a percentage of all URLs published (e.g. the 1162 unique URLs published in 2007 represent 19% of all URLs published to date).

Jonathan D. Wren; **URL decay in MEDLINE—a 4-year follow-up study**, 2008 *Bioinformatics*

https://doi.org/10.1093/bioinformatics/btn127

Replicate
Reproduce
Reuse

Tiers of Data Sharing



Within Group

-Data Management
-Scientific Record Keeping

Between Groups

-'Wet Lab' protocol Sharing
-'Wet Lab' Reagent Sharing

Between Organizations

-Publishing
-Data Visualization

Within the community

https://repro4everyone.org/

Replicate
Reproduce
Reuse

Data Repository

Use a data repository not your website. They provide:

- Persistent identifiers for your data (like DOI)
  - Unique & citable
  - Prevents “link rot”
- Persistent access
- Preservation
- Backup
- Management of access
- Versioning
- Licensing

Data License

- Consider **Creative Commons licenses** for data & text
- Either CC-0 or CC-BY
- Guidance on data licenses available through Digital Curation Center: http://www.dcc.ac.uk/resources/how-guides-license-research-data

Code License

- Consider open source license such as MIT, BSD, or Apache license
- Guidance on software licenses by Karl Broman (http://kbroman.org/steps2rr/pages/licenses.html) and Open Source Initiative (https://opensource.org/licenses)

General Purpose Repositories

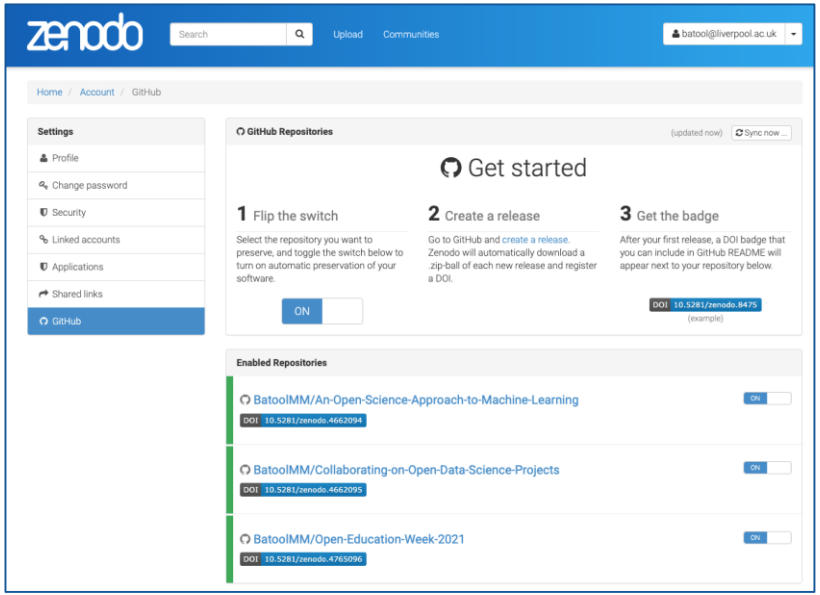
In addition to a specified data repository, you can make a deposit to a general purpose repository:

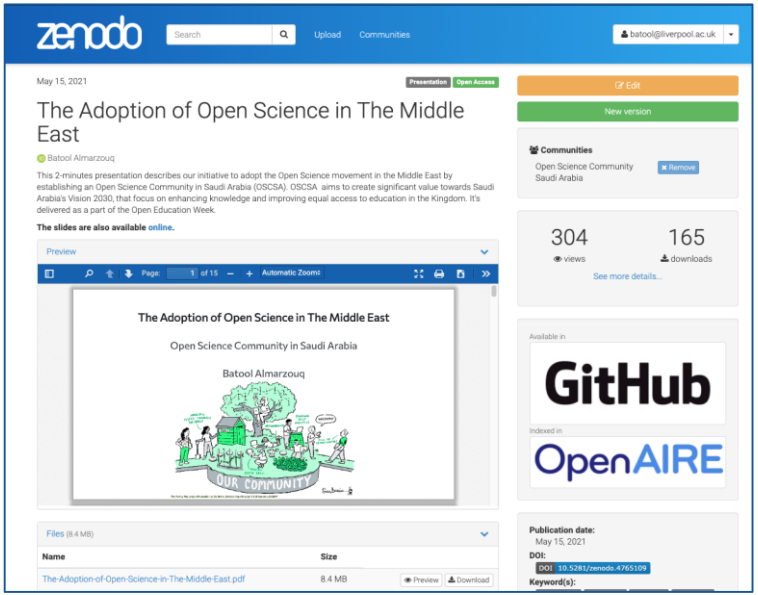
- DataDryad http://datadryad.org/ (curated digital repository; free to access, $ to publish datasets)
- Figshare https://figshare.com/
- Zenodo https://zenodo.org/
- Open Science Framework https://osf.io/
- Find data repositories: https://www.re3data.org/

re3data.org DRYAD figshare zenodo

Replicate Reproduce Reuse

Zenodo meets GitHub





Replicate
Reproduce
Reuse

Why share?

- Sharing increases access and speeds up science.
- Improves your science and record keeping.
- Improves reproducibility and increases citations.
- Increases your loveliness.

Replicate
Reproduce
Reuse

FAIR Data

- Findable
- Accessible
- Interoperable
- Reusable

Get organized! Be happy!

The Turing Way project illustration by Scriberia. Original version on Zenodo. http://doi.org/10.5281/zenodo.3695300.

Replicate
Reproduce
Reuse

WORLD VIEW | 05 September 2022

Without appropriate metadata, data-sharing mandates are pointless

Funders and investigators must demand appropriate metadata standards to take data from foul to FAIR.

Mark A. Musen

https://www.nature.com/articles/d41586-022-02820-7

Replicate
Reproduce
Reuse

Bioinformatics Tools

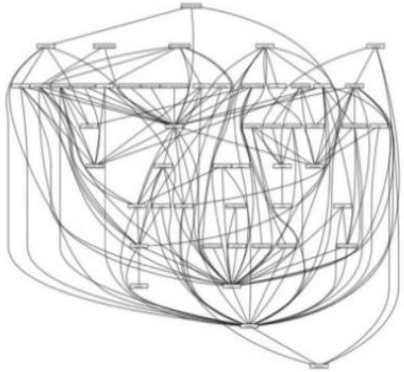
for computational reproducibility

Replicate
Reproduce
Reuse

What version of the program, data, etc… did I use? Why won't it work?

Dependency hell

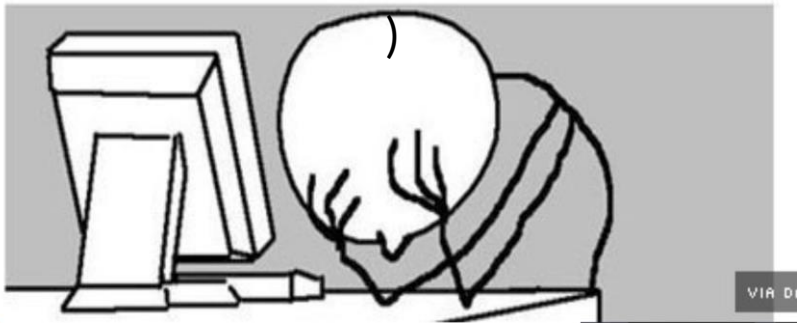
(files require other files)



https://carpentries.org

Why did I do this?

(missing documentation)



'Bioinformatic data skills' Vincent Buffalo

Replicate
Reproduce
Reuse

Document your analysis with literate programming

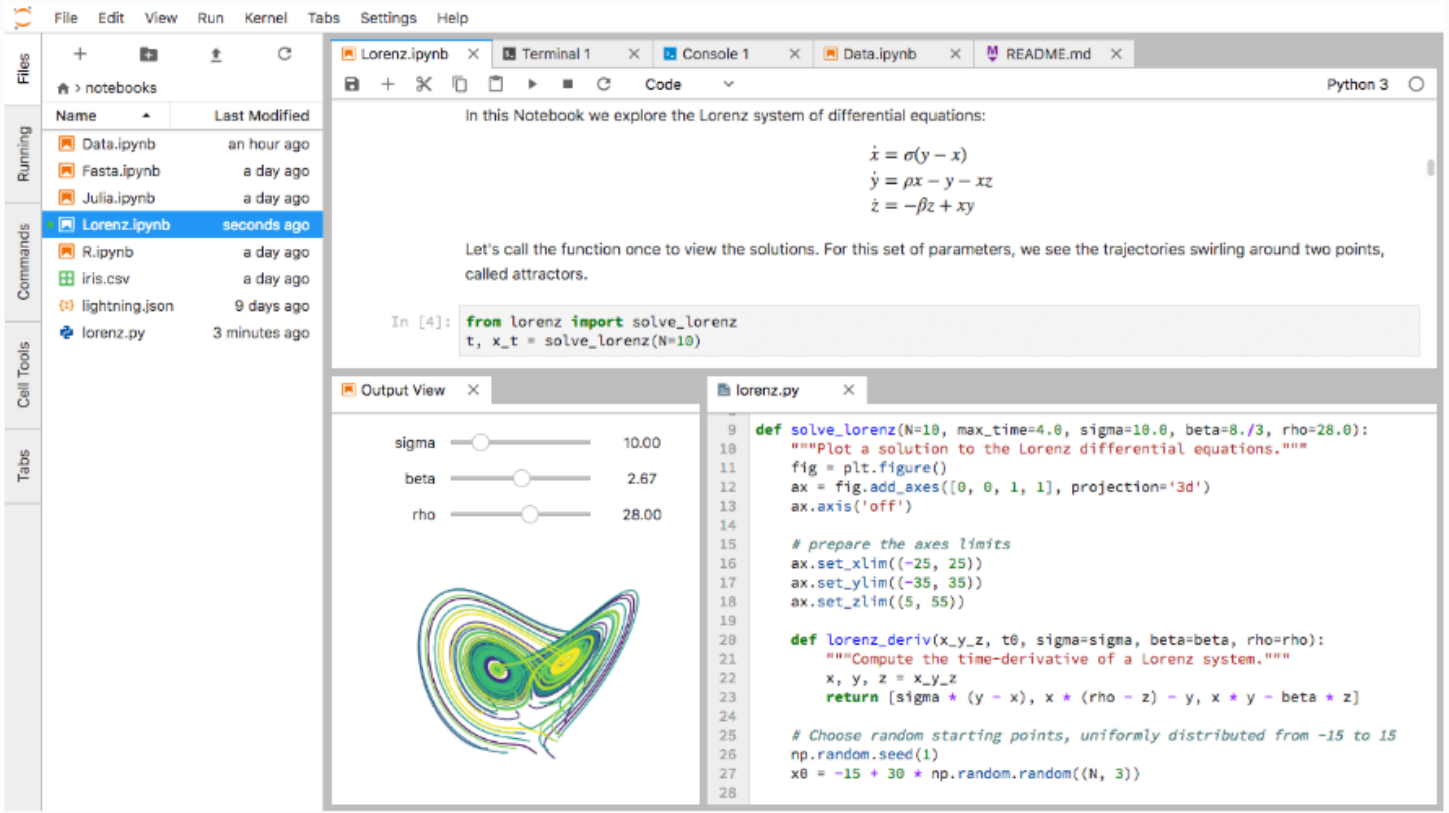
- Documentation of your analysis narrative and the analysis code together in one executable document
- What you did and why you did it
- Interactive data exploration
- Easily shared
- Best start: Jupyter Notebooks or RMarkdown with KnitR

https://jupyter.org/

RStudio®

https://www.rstudio.com

Jupyter Notebooks - Example



RMarkdown

PS1_solutions.Rmd

PS1_solutions.Rmd ×

Users > jerlich > Box Sync > CMN2017 > PS1_solutions.Rmd

```
1  ---
2  title: "Problem Set 1 Solutions"
3  author: "Prof. Jeffrey Erlich"
4  date: "9/19/2017"
5  output: html_document
6  ---
7  ```{r setup, include=FALSE}
8  knitr::opts_chunk$set(echo = TRUE)
9  options(scipen = 1, digits = 2)
10 require(cowplot)
11
12 ```
13
14 You can see the code that generated this page [here](https://int.erlichlab.org/
   teaching/cmn2017/PS1_solutions.Rmd).
15
16 ## Question 1
17
18 We setup the baseline cell parameters first.
19
20 ```{r q1-setup}
21 R1 = 400e6 # Ohm
22 C = 20e-12 # F
23 I = 0.1e-9 # A
24 tau1 = R1 * C # s
25 ```
26
27 We assume some resistance of the cell $R_1 = `r R1/1e6` \ M\Omega$. Also
   capacitance, $C= `r C/1e-12` \ pF$, which makes the time-constant $\tau_1=R_1\cdot
   C= `r R1*C*1e3` \ ms$. For now, we will not simulate the dynamical system, just use
   the known solution.
28 For a current step, we get a rising voltage, $V(t) = IR(1-e^{-t/\tau})$.
29 At the end of the step we get a falling voltage $V(t) = V_{step}(e^{-t/\tau})$,
   where $V_{step}$ is the voltage at the end of the current injection.
30 Note, we write $V_{step}$ because if the current step was not long enough to reach
   steady state, then $V_{step} \neq IR$.
31
32 Next, we will setup the parameters for the injection.
33 ```{r q1}
34 tax = seq(-25, 220)/1000 # setup the time axes: -25 to 220 ms
35 endtime = 0.1 # The duration of the injection
36 step = tax>0 & tax<endtime # the current injection will go from 0 to endtime seconds
37 stepdown = tax>=endtime  #
38 V1 = tax*0     # Initialize V1 to all zeros.
```

0 0 Server not selected Ln 23, Col 15 Spaces: 2 UTF-8 LF R Markdown [off]

Problem Set 1 Solutions

Prof. Jeffrey Erlich

9/19/2017

You can see the code that generated this page here.

Question 1

We setup the baseline cell parameters first.

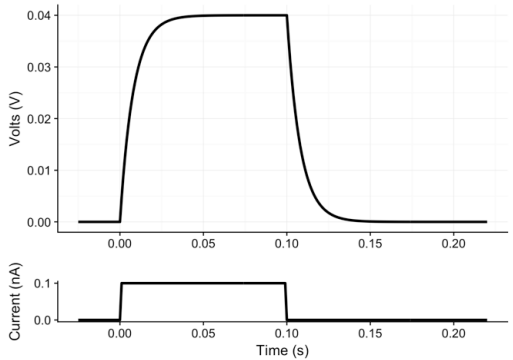
```
R1 = 400e6 # Ohm
C = 20e-12 # F
I = 0.1e-9 # A
tau1 = R1 * C # s
```

We assume some resistance of the cell $R_1 = 400\ M\Omega$. Also capacitance, $C = 20\ pF$, which makes the time-constant $\tau_1 = R_1 \cdot C = 8\ ms$. For now, we will not simulate the dynamical system, just use the known solution.
For a current step, we get a rising voltage, $V(t) = IR(1 - e^{-t/\tau})$.
At the end of the step we get a falling voltage $V(t) = V_{step}e^{-t/\tau}$, where V_{step} is the voltage at the end of the current injection.
Note, we write V_{step} because if the current step was not long enough to reach steady state, then $V_{step} \neq IR$.

Next, we will setup the parameters for the injection.

```
tax = seq(-25, 220)/1000 # setup the time axes: -25 to 220 ms
endtime = 0.1 # The duration of the injection
step = tax>0 & tax<endtime # the current injection will go from 0 to endtime seconds
stepdown = tax>=endtime  #
V1 = tax*0     # Initialize V1 to all zeros.
V1[step] = I*R1*(1 - exp(-tax[step]/tau1))
# During the step we use the formula for rising voltage

V1[stepdown] = max(V1)*(exp(-(tax[stepdown]-endtime)/tau1))
# After the step we use the formula for decaying voltage
```

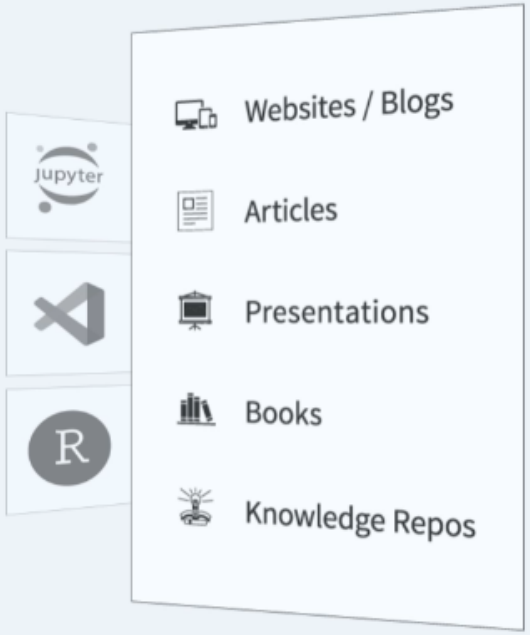


Quarto – for output creation

An open-source scientific and technical publishing system

- Author using Jupyter notebooks or with plain text markdown in your favorite editor.
- Create dynamic content with Python, R, Julia, and Observable.
- Publish reproducible, production quality articles, presentations, websites, blogs, and books in HTML, PDF, MS Word, ePub, and more.
- Share knowledge and insights organization-wide by publishing to Posit Connect, Confluence, or other publishing systems.
- Write using Pandoc markdown, including equations, citations, crossrefs, figure panels, callouts, advanced layout, and more.

Analyze. Share. Reproduce. You have a story to tell with data—tell it with Quarto.



https://quarto.org/

Replicate
Reproduce
Reuse

Quarto example

cell-options.ipynb

Code | Python 3 (ipykernel)

Palmer Penguins

```
---
author: Norah Jones
format:
  html:
    code-tools: true
    code-fold: true
---
```

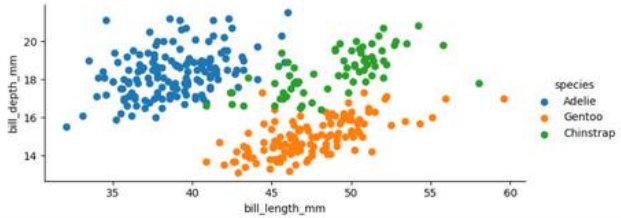
[3]:
```
#| echo: false
import pandas as pd
df = pd.read_csv("palmer-penguins.csv")
```

Exploring the Data

See @fig-bill-sizes for an exploration of bill sizes by species.

[6]:
```
#| label: fig-bill-sizes
#| fig-cap: Bill Sizes by Species
import matplotlib.pyplot as plt
import seaborn as sns
g = sns.FacetGrid(df, hue="species", height=3, aspect=3.5/1.5)
g.map(plt.scatter, "bill_length_mm", "bill_depth_mm").add_legend()
```

[6]:
```
<seaborn.axisgrid.FacetGrid at 0x2946720e0>
```



Palmer Penguins

Code

Show All Code
Hide All Code

AUTHOR
Norah Jones

PUBLISHED
March 12, 2023

Exploring the Data

See Figure 1 for an exploration of bill sizes by species.

▼ Code

```
import matplotlib.pyplot as plt
import seaborn as sns
g = sns.FacetGrid(df, hue="species", height=3, aspect=3.5/2)
g.map(plt.scatter, "bill_length_mm", "bill_depth_mm").add_legend()
```

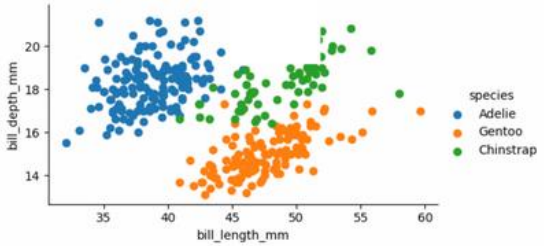


Figure 1: Bill Sizes by Species

Replicate
Reproduce
Reuse

Notebooks

- Help to keep track of analyses
- Interactive coding and data exploration
- Embedded visualization
- Automatic completion and easy access to library documentation
- Mix of code and documentation.
- Acts as a "live" version of a manuscript.
- Can embed code results in formatted text
- Easily shared (https://nbviewer.jupyter.org)
- Widgets available (customization)
- Compatible with >40 programming languages
- Can run remotely on server

https://jupyter.org/

https://www.rstudio.com

Document changes

Version-control system that uses *forks* and *pull requests*.

git-scm.com/doc

With version control

Git

- Records changes (what, when, who)
- Documents version history
- Illustrates changes between versions (diffs)

Github

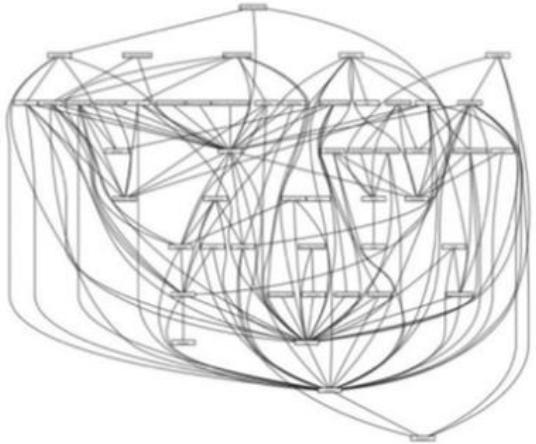
- Lets you share code easily
- Lets you collaborate on your code more easily

Git-based repository with a social media component.

http://smutch.github.io/VersionControlTutorial//

How did I install all these different software packages???

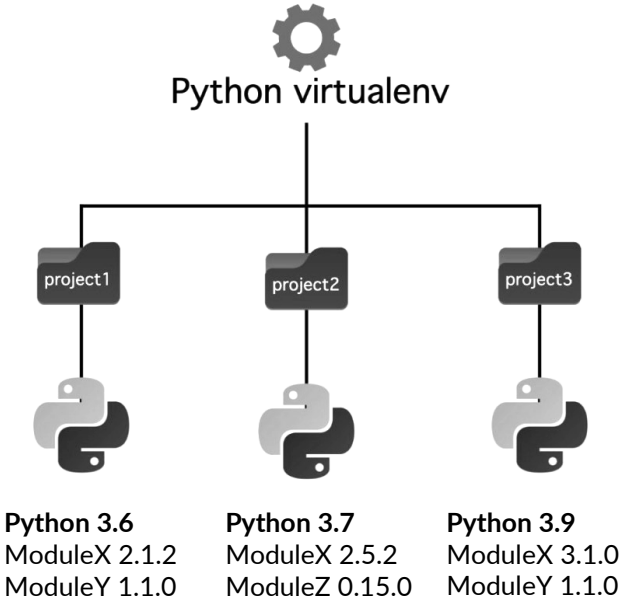
Dependency hell



Version conflict

Document your computational environment with package, dependency & environment managers

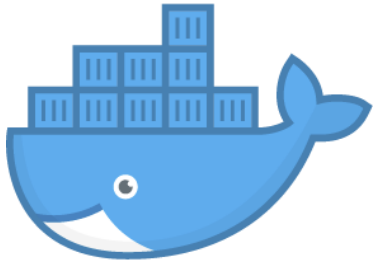
- Python virtual environments
- Dependency managers
- Handles installs & dependencies
- Document your environment (requirements.txt)
- In RStudio, use CRAN and RProjects
- Document your packages (install.R)



https://conda.io/ https://bioconda.github.io/

Make your environment portable with containers

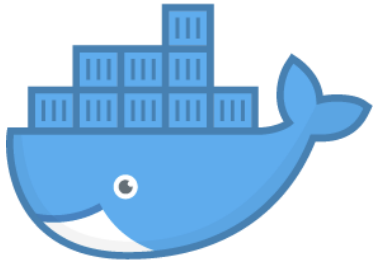
- Everything needed to run your analysis is packed up into an “image”
- Images are self-contained with all code, programs, environment, Dockerfile included
- No subsequent installation required
- Spin an image into a container using Docker or Singularity - it is like sharing your computer



https://docs.docker.com

Make your environment portable with containers

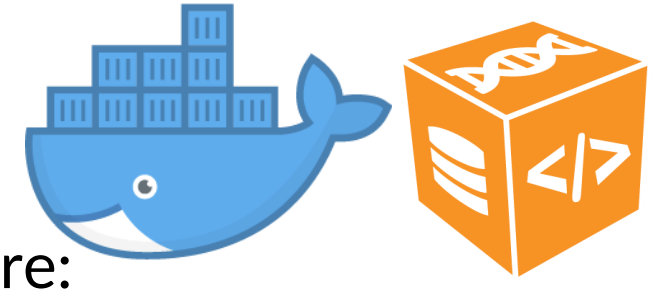
- Binder uses the environment documentation file from your Github repo to automatically build a shareable Docker image
- Runs in the cloud
- Executable



https://tinyurl.com/jupyter-binder2-0

https://tinyurl.com/eLife-binder2-

Containers



Docker runs images as containers that are:

- No subsequent installation required
- Isolated
- Portable including dissemination

  But...

  can use up a lot of system resources, networking can be a pain.

https://docs.docker.com

Have you ever used these kind of tools? What advice would you give to a novice?

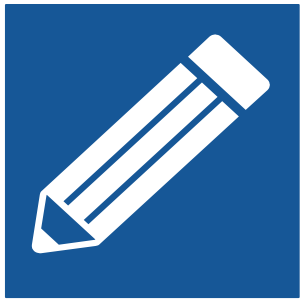
@repro4everyone
#repro4everyone

Data Visualization & Analysis

Data presentation is the foundation of our collective scientific knowledge...

Figures are especially important. They often show data for key findings.

Effective figures should...



Immediately convey information about **study design**

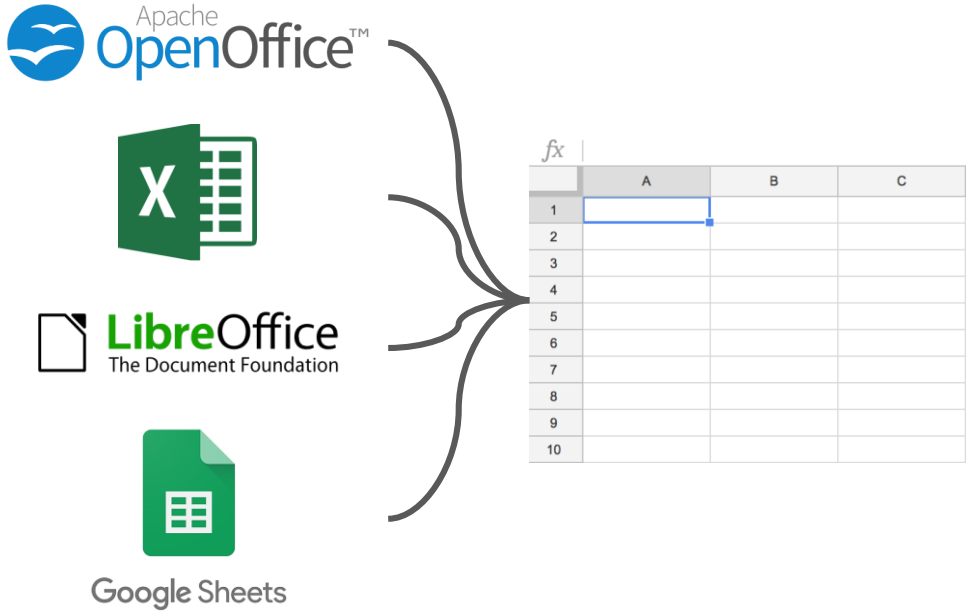
Illustrate **important findings**

Allow the reader to **critically evaluate** the data:
Show your data!

Weissgerber et al. 10.1074/jbc.RA117.000147

Replicate
Reproduce
Reuse

The usual way and its flaws...

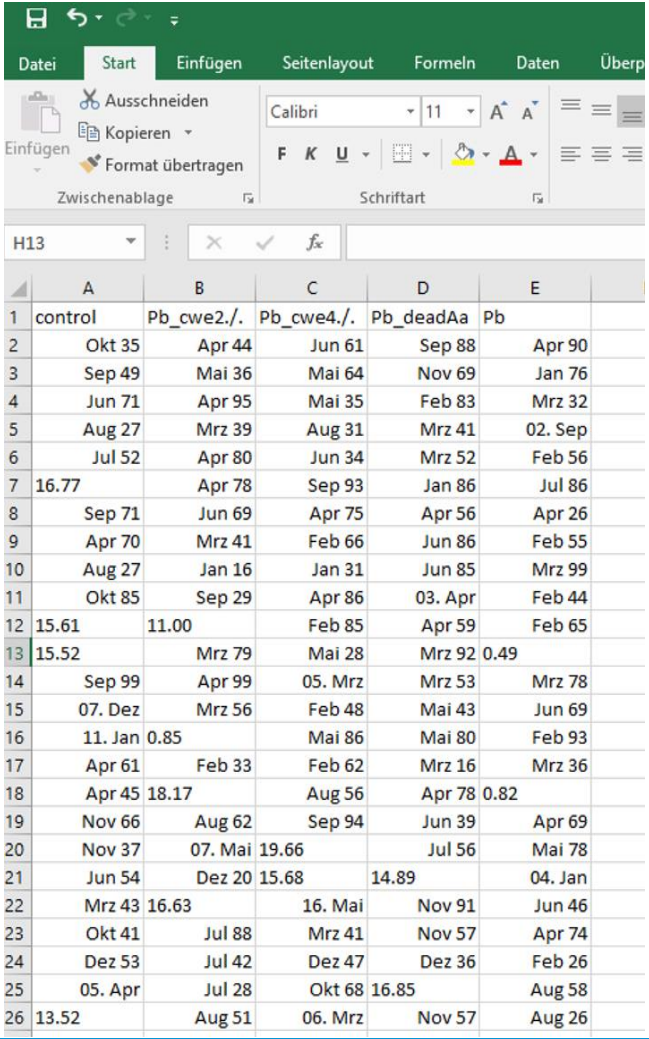


Issues:

- Reproducible Workflows?
  - Problems can be avoided by using macros or dashboards
  - However, who uses these?
- Excel renames Genes
  - Ziemann et al., 2016
  - 20% of papers in leading genomic journals contain gene list errors
- Default Plots are often Bar Charts and Line Plots

Weissgerber et al. 2017 10.1074/jbc.RA117.000147
Broman and Woo 2018 10.1080/00031305.2017.1375989

Replicate
Reproduce
Reuse



	A	B	C	D	E
1	control	Pb_cwe2./.	Pb_cwe4./.	Pb_deadAa	Pb
2	Okt 35	Apr 44	Jun 61	Sep 88	Apr 90
3	Sep 49	Mai 36	Mai 64	Nov 69	Jan 76
4	Jun 71	Apr 95	Mai 35	Feb 83	Mrz 32
5	Aug 27	Mrz 39	Aug 31	Mrz 41	02. Sep
6	Jul 52	Apr 80	Jun 34	Mrz 52	Feb 56
7	16.77	Apr 78	Sep 93	Jan 86	Jul 86
8	Sep 71	Jun 69	Apr 75	Apr 56	Apr 26
9	Apr 70	Mrz 41	Feb 66	Jun 86	Feb 55
10	Aug 27	Jan 16	Jan 31	Jun 85	Mrz 99
11	Okt 85	Sep 29	Apr 86	03. Apr	Feb 44
12	15.61	11.00	Feb 85	Apr 59	Feb 65
13	15.52	Mrz 79	Mai 28	Mrz 92	0.49
14	Sep 99	Apr 99	05. Mrz	Mrz 53	Mrz 78
15	07. Dez	Mrz 56	Feb 48	Mai 43	Jun 69
16	11. Jan	0.85	Mai 86	Mai 80	Feb 93
17	Apr 61	Feb 33	Feb 62	Mrz 16	Mrz 36
18	Apr 45	18.17	Aug 56	Apr 78	0.82
19	Nov 66	Aug 62	Sep 94	Jun 39	Apr 69
20	Nov 37	07. Mai	19.66	Jul 56	Mai 78
21	Jun 54	Dez 20	15.68	14.89	04. Jan
22	Mrz 43	16.63	16. Mai	Nov 91	Jun 46
23	Okt 41	Jul 88	Mrz 41	Nov 57	Apr 74
24	Dez 53	Jul 42	Dez 47	Dez 36	Feb 26
25	05. Apr	Jul 28	Okt 68	16.85	Aug 58
26	13.52	Aug 51	06. Mrz	Nov 57	Aug 26

Ruined raw data after Excel converted measurements into dates…

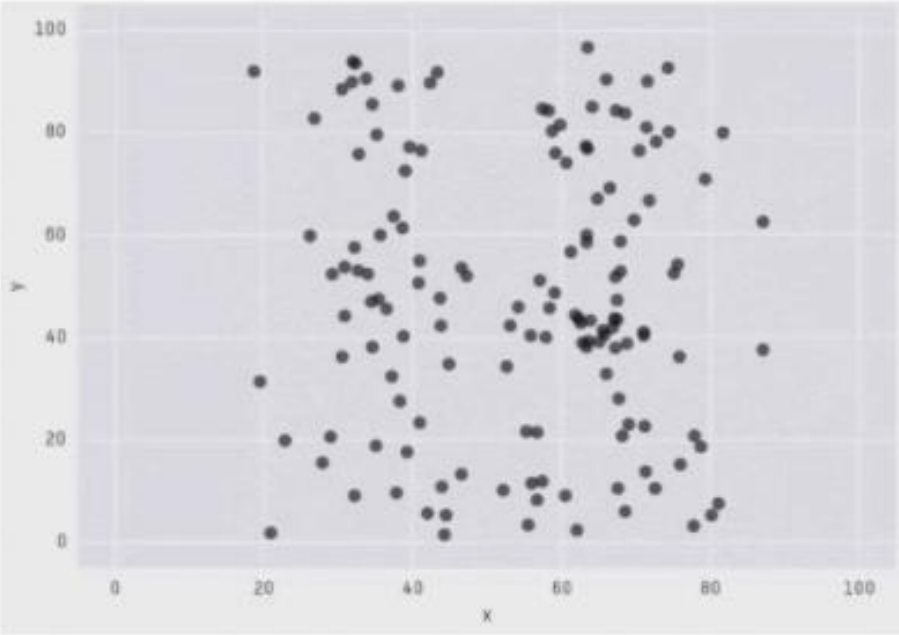
Replicate
Reproduce
Reuse

Show your data

@AllisonHorst

Never trust data based on the summary statistics alone!

There really is NO alternative to showing the actual data

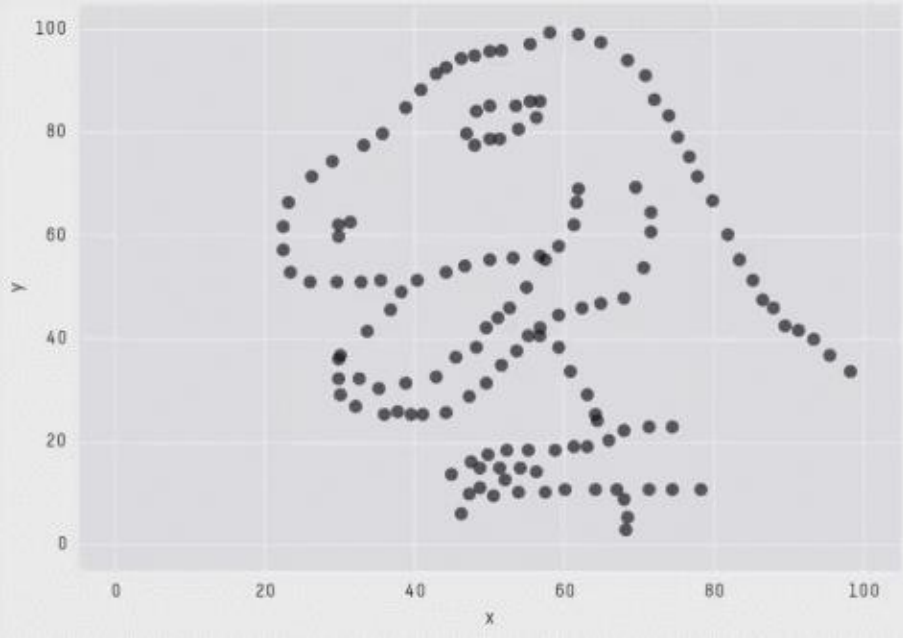


X Mean: 54.2695813
Y Mean: 47.8331366
X SD : 16.7629092
Y SD : 26.9399434
Corr. : -0.0613537

Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing

Never trust data based on the summary statistics alone!

There really is NO alternative to showing the actual data

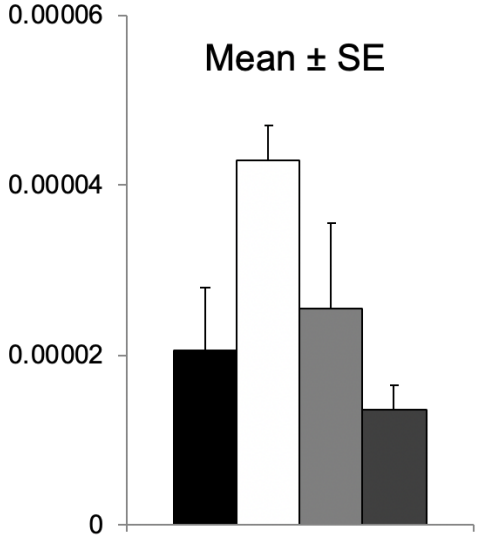


X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

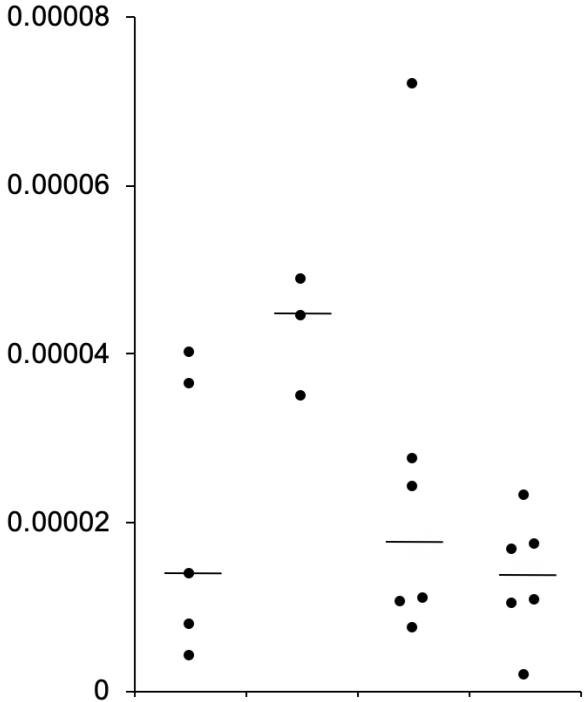
Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing

Our interpretation depends on what we see

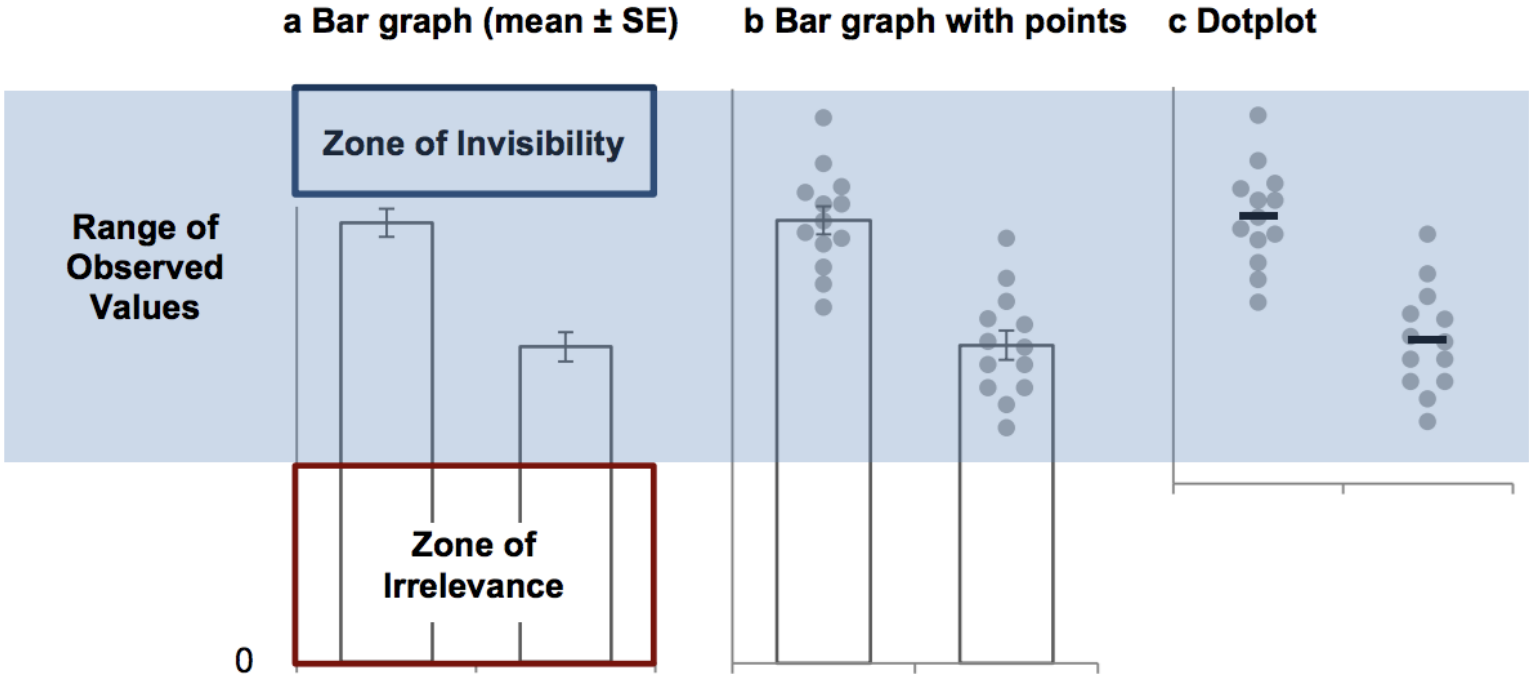
Reader is a passive observer



Reader is an active participant

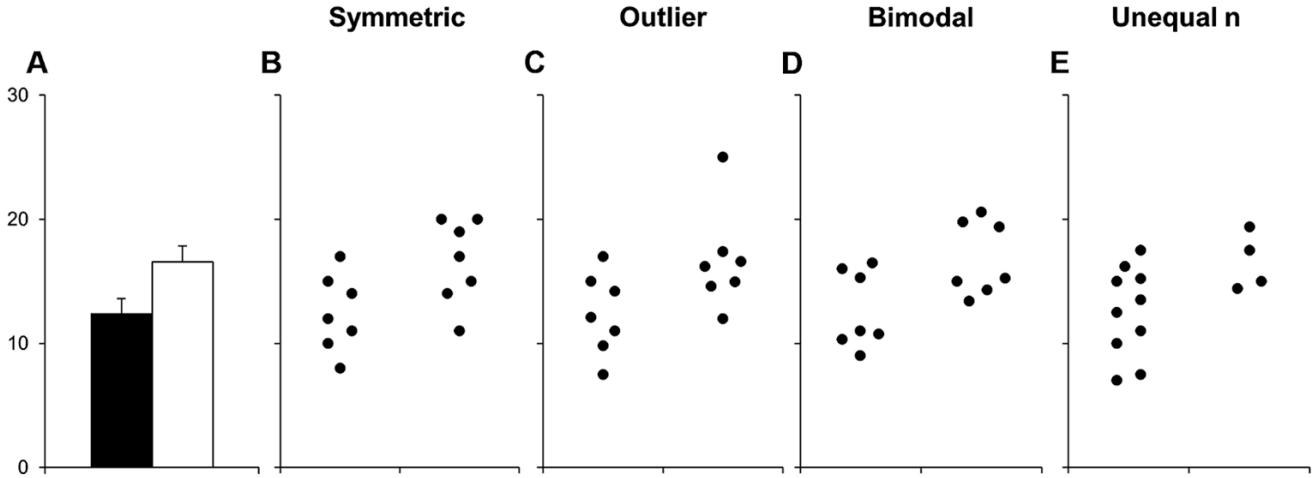


No bar graphs - even for normally distributed data



Weissgerber et al., 2017 JBC, http://www.jbc.org/content/292/50/20592.full

Avoid bar charts for continuous data

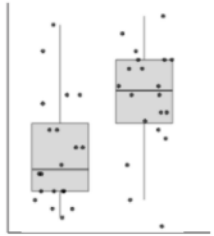


Test	p value			
T-test: Equal var.	0.035	0.050	0.026	0.063
T-test: Unequal var.	0.035	0.050	0.026	0.035
Wilcoxon	0.054	0.073	0.128	0.103

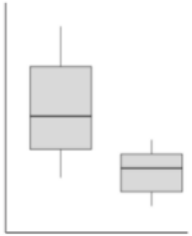
Weissgerber et al. https://doi.org/10.1371/journal.pbio.1002128

How to choose the right plot

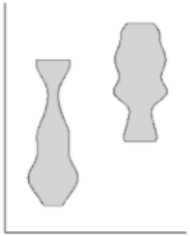
Dotplot



Boxplot with points



Boxplot



Violin plot (with or without points)

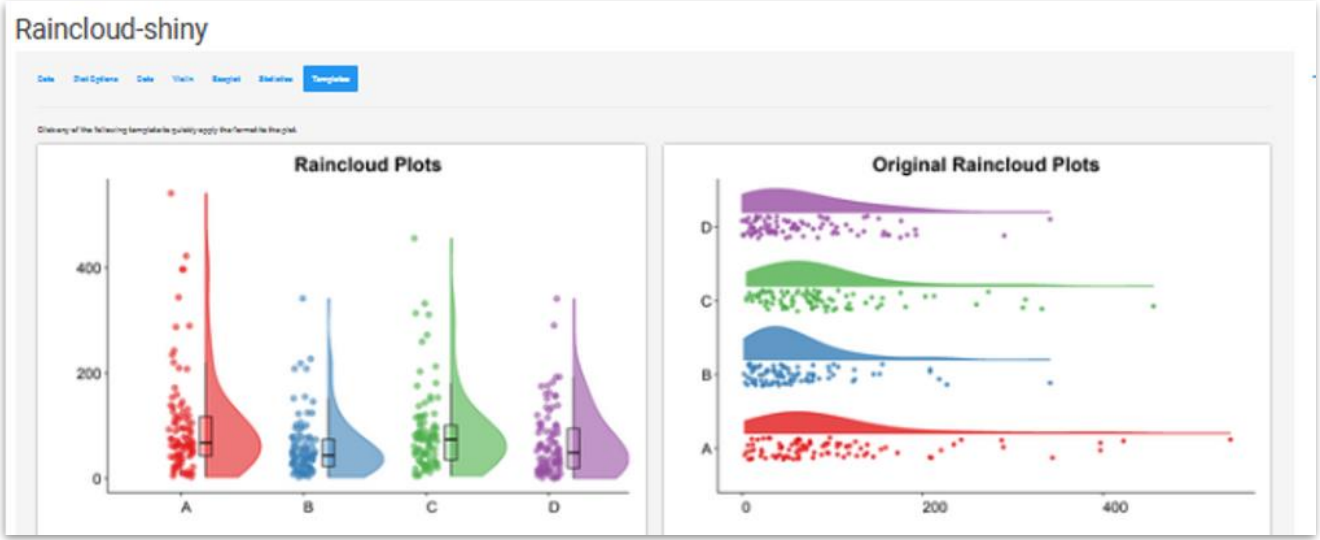
Bar graph

Add raincloud plots

Outcome variable	Continuous	Continuous	Continuous	Continuous	Counts & proportions
Sample size	Small	Medium	Large	Medium to Large	Any
Data distribution	Any	Any	Do not use for bimodal data	Any	N/A

Weissgerber et al., 2017 JBC

One step further: Interactive plots



[Interactive Dot Plot]

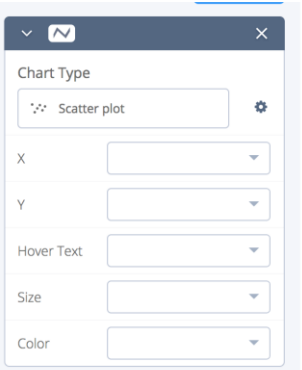
[Interactive Line Graph]

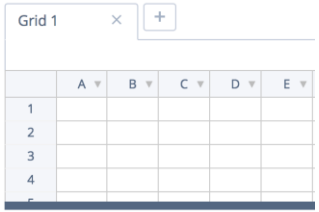
[Rain cloud plots]

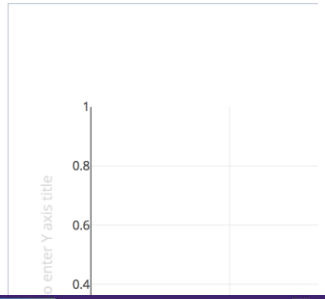
https://trolls.fandom.com/wiki/Cloud_Guy

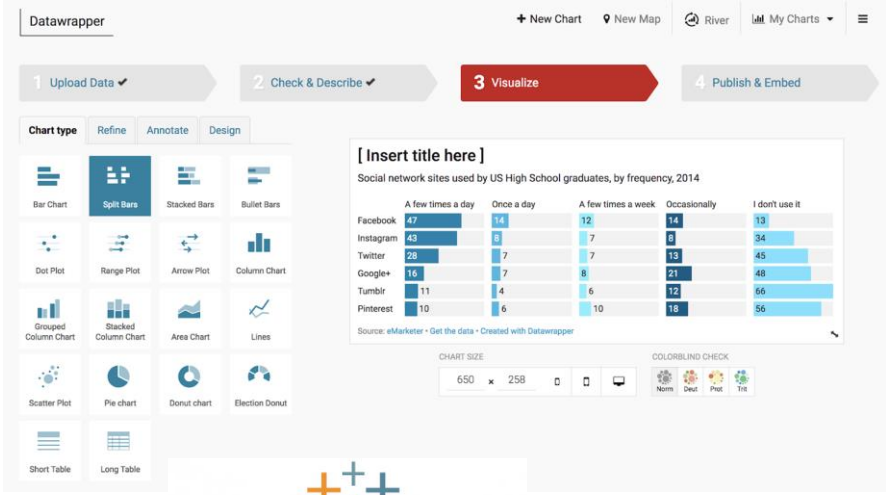
Replicate
Reproduce
Reuse

Some intermediate options









Prism 8

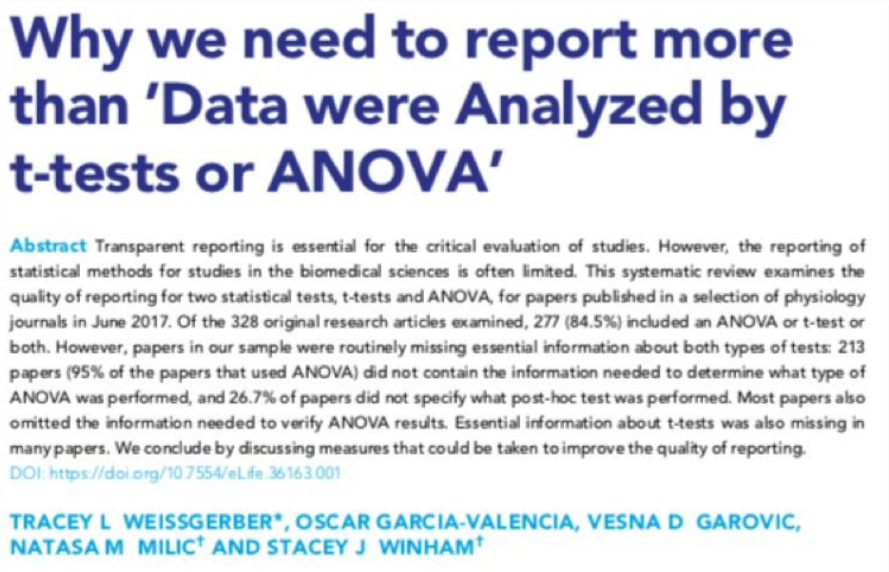
+ableau

https://plot.ly/create/#/

Replicate
Reproduce
Reuse

Statistical reporting

- Clearly report exactly what test was used & results, including exact:
  - P-values
  - Test statistics
  - Degrees of freedom
- Simple stats: include test info in figure and table legends



Why we need to report more than 'Data were Analyzed by t-tests or ANOVA'

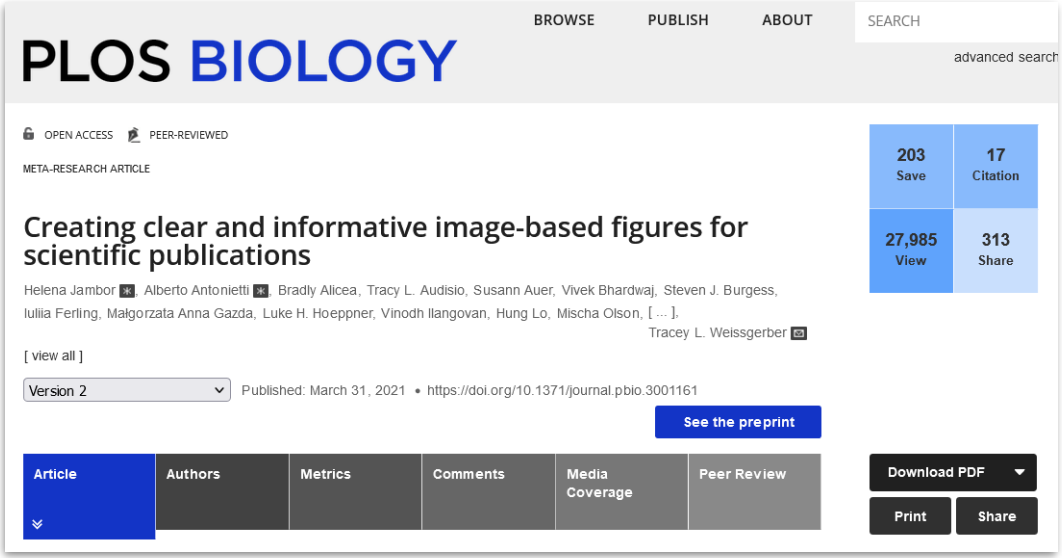
Abstract Transparent reporting is essential for the critical evaluation of studies. However, the reporting of statistical methods for studies in the biomedical sciences is often limited. This systematic review examines the quality of reporting for two statistical tests, t-tests and ANOVA, for papers published in a selection of physiology journals in June 2017. Of the 328 original research articles examined, 277 (84.5%) included an ANOVA or t-test or both. However, papers in our sample were routinely missing essential information about both types of tests: 213 papers (95% of the papers that used ANOVA) did not contain the information needed to determine what type of ANOVA was performed, and 26.7% of papers did not specify what post-hoc test was performed. Most papers also omitted the information needed to verify ANOVA results. Essential information about t-tests was also missing in many papers. We conclude by discussing measures that could be taken to improve the quality of reporting.
DOI: https://doi.org/10.7554/eLife.36163.001

TRACEY L WEISSGERBER*, OSCAR GARCIA-VALENCIA, VESNA D GAROVIC, NATASA M MILIC† AND STACEY J WINHAM†

https://elifesciences.org/articles/36163

Designing effective figures with images

Further reading on how-to for images



https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3001161

https://twitter.com/SusannAuer/status/1384412237079646210

7 step guide for figure design: https://osf.io/ycfub

Replicate
Reproduce
Reuse

Why do good figures matter?

- Often scientists, reviewers and editors examine figures first
- Search engines and journal websites allow readers to examine figures of papers
- Scientists share image-based figures on posters and social media

Design figures for a broad audience

- Think about your audience: things that are clear to you may be confusing for readers from a different field
- Your readers include scientists in your field and others, reviewers, educators, grant officers…
- **Ensure that your figures are self explanatory!**

Roland et al. 10.1002/adma.201102518

7 steps to preparing image-based figures

1. Choose a magnification & scale that fits your research question
2. Add a clearly labeled scale bar
3. Use color wisely
4. Choose a colorblind accessible color palette
5. Design your figure with a layout sketch or table
6. Annotate the figure (accessibility to broad audience)
7. Prepare legends (clear explanations)

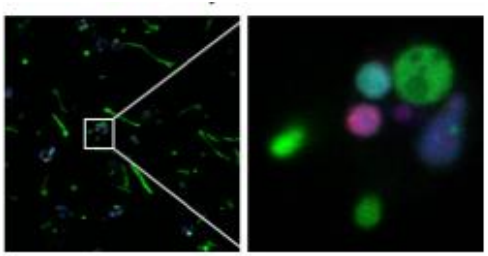
https://osf.io/ycfub

Replicate
Reproduce
Reuse

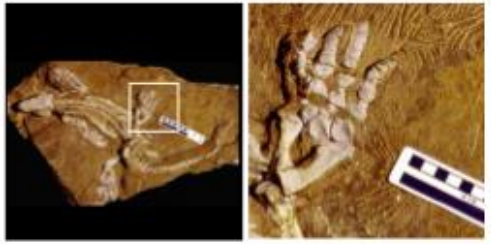
On scale and magnification: use insets to allow readers to see more than one scale

Ensure insets are

1. accurately marked
2. clearly explained



5. Inset indicated



6. Inset indicated

Replicate
Reproduce
Reuse

Scale bars convey essential information about size

1. **Every image needs a scale bar.** Differences in size (phenotype) are important for reproducibility.
2. **Scale bars & labels should be clearly visible**
3. **Annotate scale bar dimensions on the image.** Searching for dimensions in the legend is time consuming.

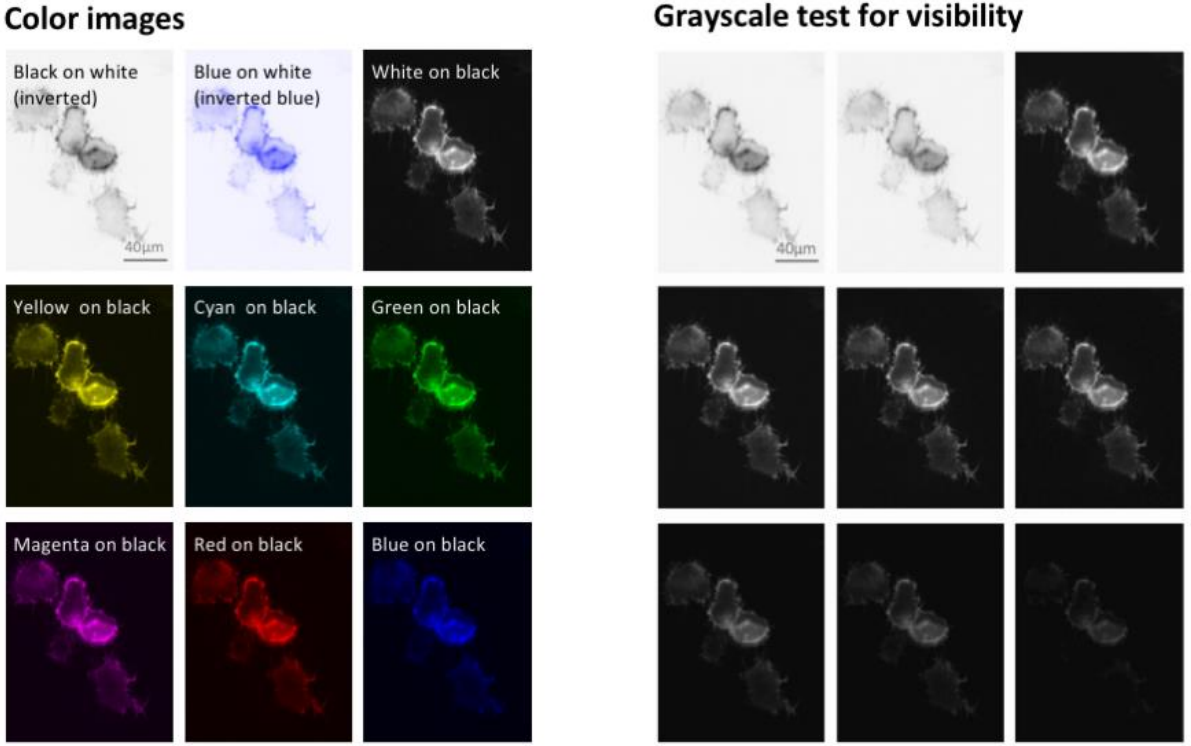
Image: Erik Streb, CC-0

Replicate
Reproduce
Reuse

Use color wisely: Visibility of colors depends on the background

Best visibility

Worst visibility



Replicate
Reproduce
Reuse

Make images colorblind-accessible

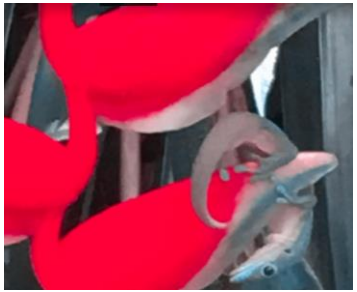
The most common form of colorblindness affects up to 8% of men and 0.5% of women of northern European ancestry.
(National Eye Institute, 2015)

It is likely that at least one of your co-authors, reviewers or editors will be colorblind. Many of your readers will also be colorblind.

Normal

Deuteranopia

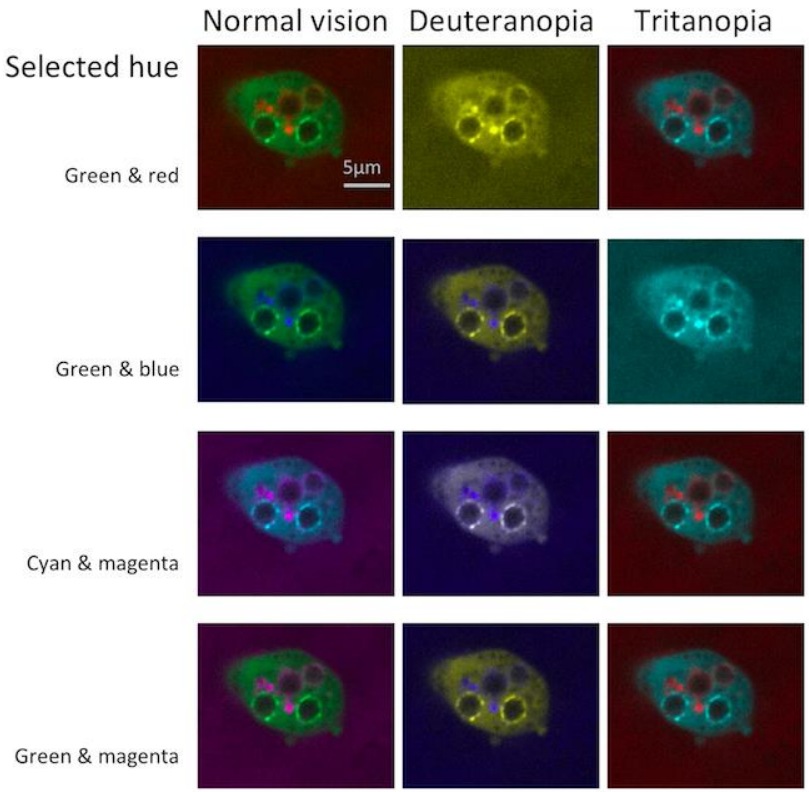
Tritanopia



Test colorblind accessibility with free tools like Color Oracle https://colororacle.org

Replicate
Reproduce
Reuse

Colorblind accessible color combinations



Colorblind accessible?

✗

✗

✓

✓

Replicate
Reproduce
Reuse

Reproducible research practices enable you to:

Organize experiments productively

Accurately analyze results

Share results with future researchers

Share techniques

Share reagents with future researchers

Accelerate science!

Tools discussed here should provide you with the framework to make you research more reproducible and will save you time and resources in the long term

Replicate
Reproduce
Reuse

Next Steps

What is one thing that you can do today to start making your research more reproducible?

Replicate
Reproduce
Reuse

@repro4everyone
https://www.repro4everyone.org
hello@repro4everyone.org

Please give feedback:

https://forms.gle/kihnTyHEehrXCWWB7

Past Funders

NEW ENGLAND BioLabs® Inc.

Chan Zuckerberg Initiative

addgene
The nonprofit plasmid repository

protocols.io

Dorothy Bishop

eLIFE

AMERICAN SOCIETY OF PLANT BIOLOGISTS
INVENIAMVS VITALIA ARCANA PLANTARVM
1924

moz://a

CODE OCEAN

License and Citation

These materials have been developed by the Reproducibility for Everyone community. Please find us at osf.io for full citation details. Slides are published under Creative Commons with Attribution license (CC-BY 4.0)

Contributors

repro4everyone Team	https://www.repro4everyone.org/pages/people/
Susann Auer	@SusannAuer
April Clyburne-Sherin	@april_cs
Sonali Roy	@SonaliRoy_
Benjamin Schwessinger	@schwessinger