Hello!
I am
Rintu Kutum

Member & Instructor, **Reproducibility for Everyone (R4E) Initiative**

Faculty Fellow, Department of Computer Science

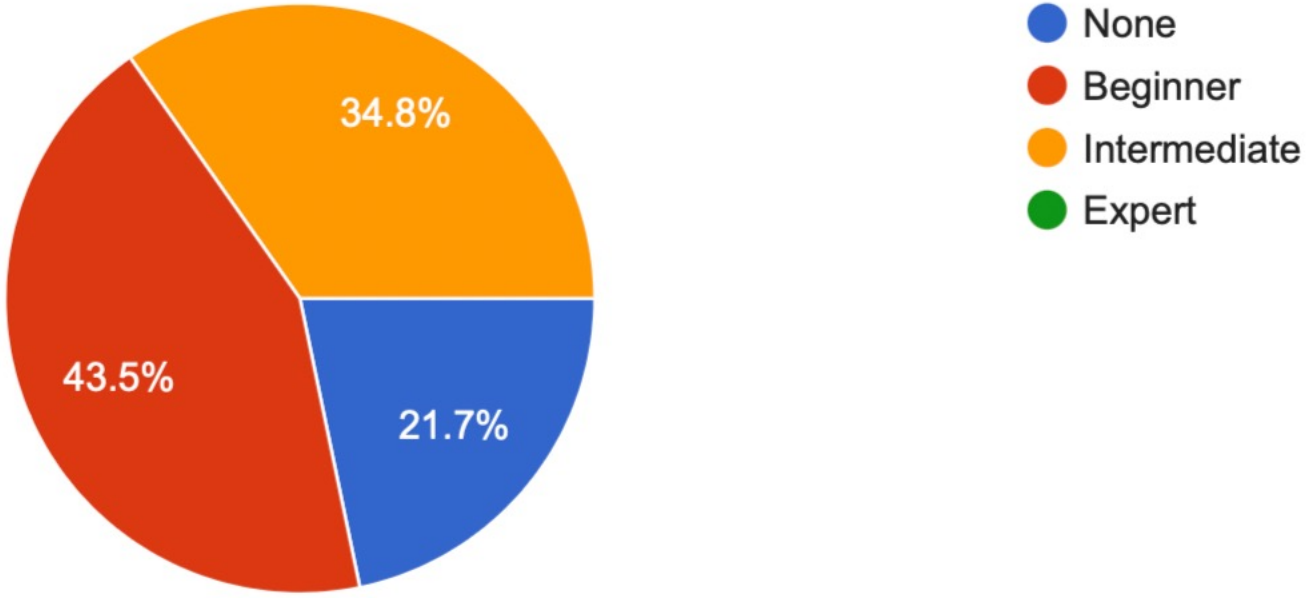
Data Scientist, Mphasis Lab for ML & Computational Thinking

Faculty, Trivedi School of Biosciences

Ashoka University

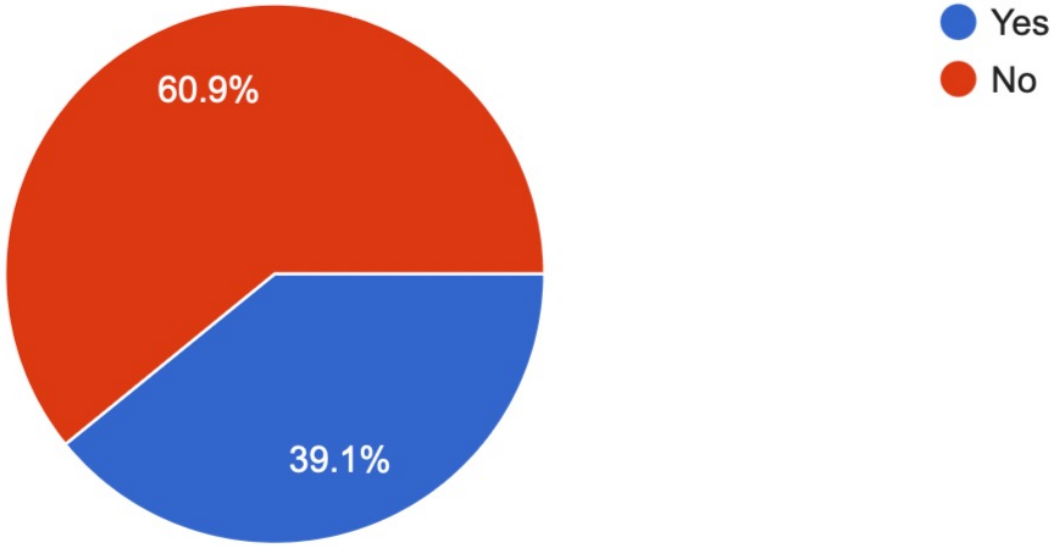
Rate your knowledge of Biostatistics

23 responses



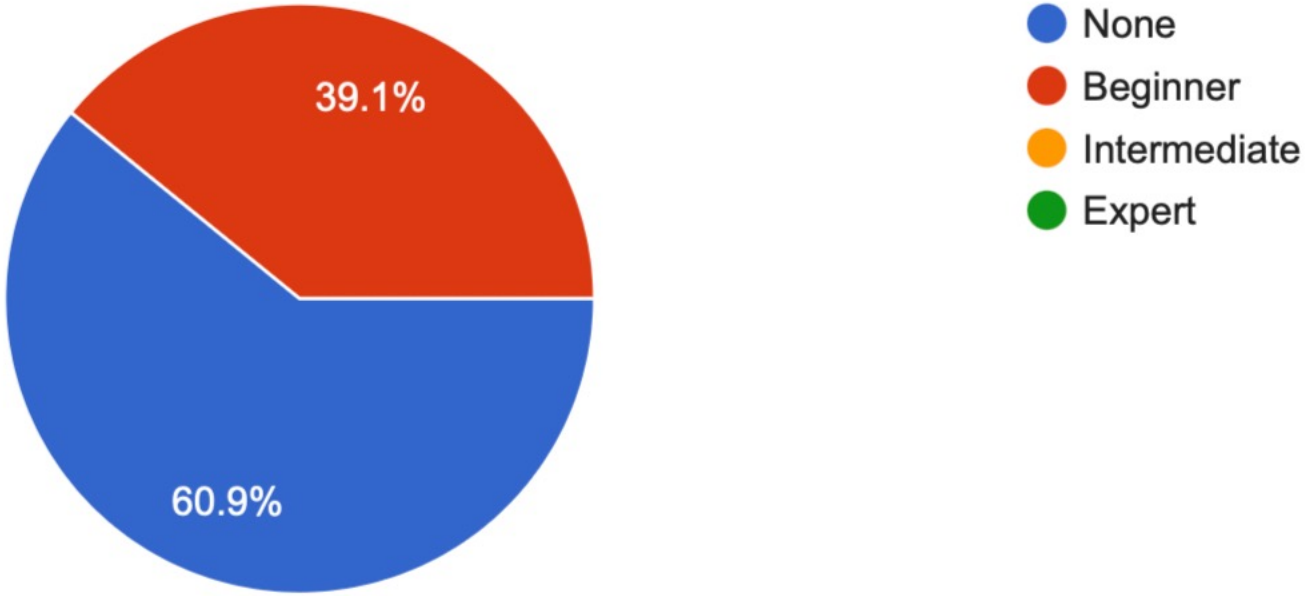
Are you aware of open source scientific computing languages?(python, R, Julia, etc.)

23 responses



Level of understanding about R ?

23 responses



Who are we?

- **Community of scientists in academia and industry who want to improve the reproducibility of science**
  - Members all over the world!
  - Members help in developing new slides, hosting workshops, applying for funding, etc.
- **Developed this open access workshop, which you can use yourself for a journal club / seminar at your institute!**
- **Want to join us?**
  Reach out to the workshop organizers or through our website:
  https://www.repro4everyone.org/join

https://repro4everyone.org/

Thank You

NEW ENGLAND BioLabs® Inc.

Chan Zuckerberg Initiative

addgene
The nonprofit plasmid repository

protocols.io

Dorothy Bishop

Benjamin Schwessinger
@schwessinger

@repro4everyone
hello@repro4everyone.org
www.repro4everyone.org

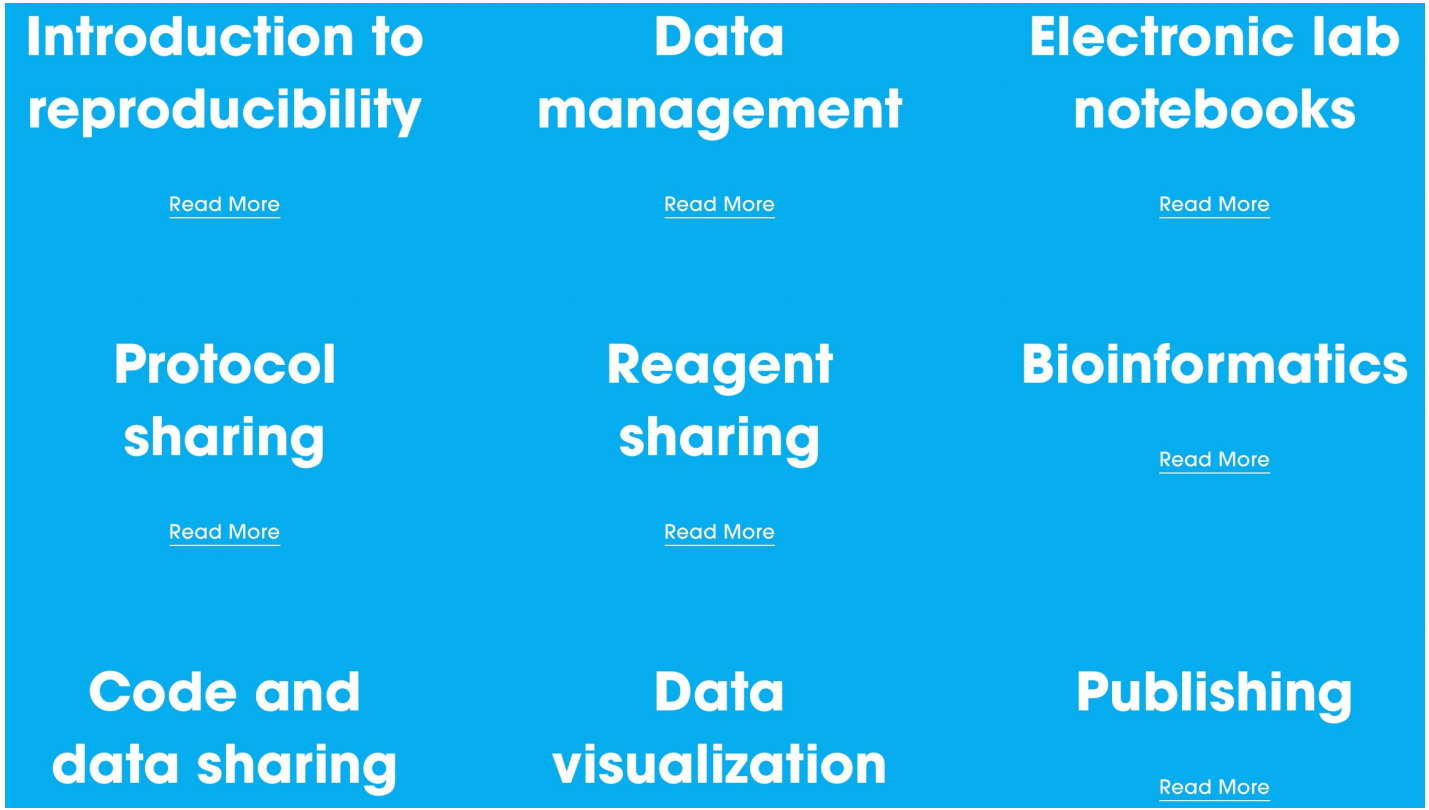
If you want to read more about our initiative:

Auer S et al., eLife 2021; https://elifesciences.org/articles/64719

Free, reusable reproducibility resources.

Explore R4E modules. Run an R4E workshop using the R4E Workshop Guide. Watch past R4E workshop recordings.

https://www.repro4everyone.org/resources



https://www.repro4everyone.org/resources

Open Source Tools

for computational reproducibility

Two days hands-on workshop on Open Source Scientific Computing for Environmental Health Sciences

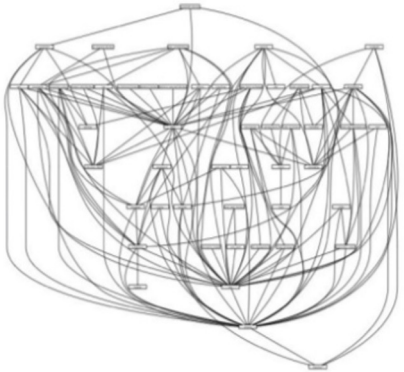
ICMR-NIREH, Bhopal

27 June 2022

Replicate
Reproduce
Reuse

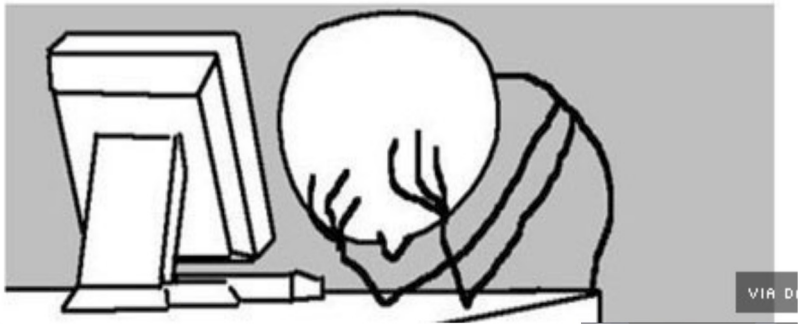
What version of the program, data, etc… did I use? Why won't it work?

Dependency hell (files require other files)



https://carpentries.org

Why did I do this?



'Bioinformatic data skills' Vincent Buffalo

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

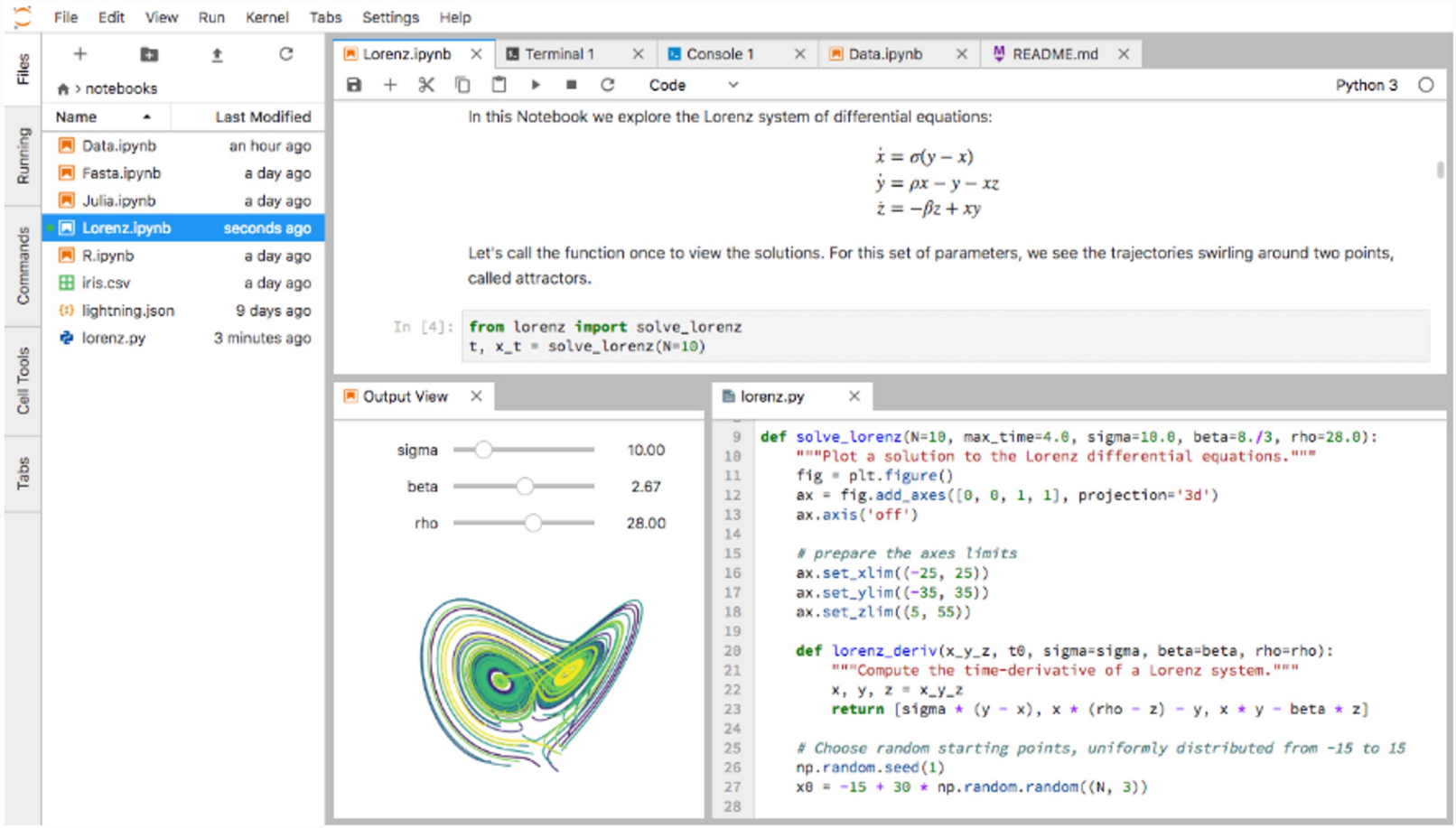
https://jupyter.org/

RStudio®

https://www.rstudio.com

Replicate
Reproduce
Reuse

Jupyter notebooks - Example



Document changes

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

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

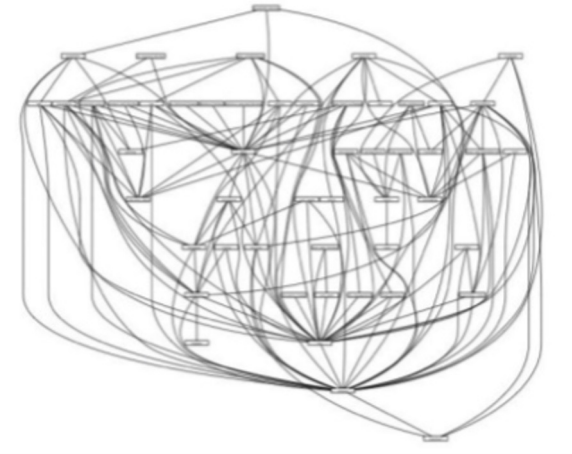
git-scm.com/doc

Git-based repository with a social media component.

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

How did I install all these different software packages?

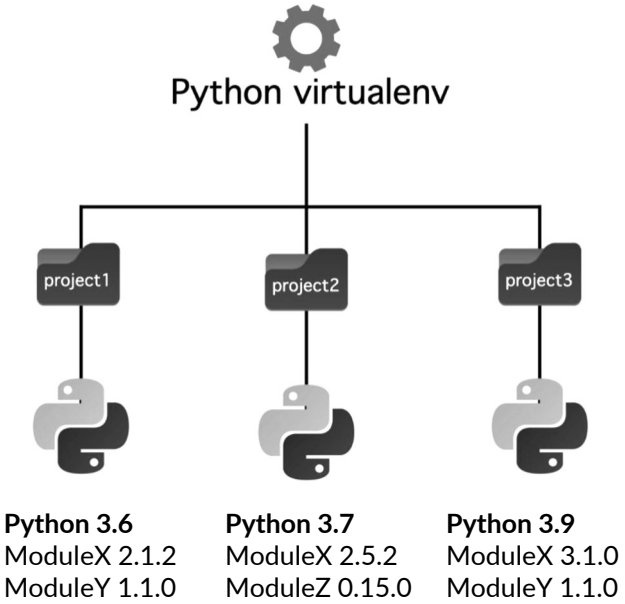
Dependency hell



Version conflict

Document your computational environment with package, dependency & environment managers

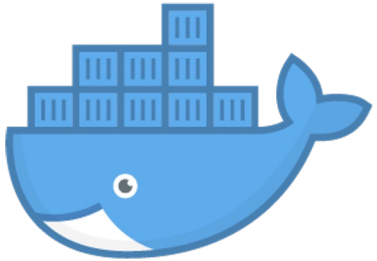
- Python virtual environments
- Dependency managers
- Handles installs & dependencies
- Document your environment (requirements.txt)
- In RStudio, use CRAN and RProjects
- Document your packages



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

Make your environment portable with containers

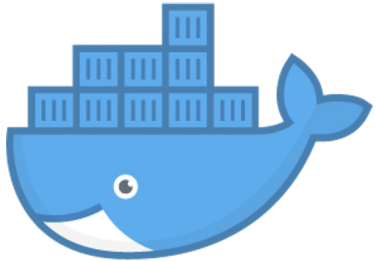
- Everything needed to run your analysis is packed up into an “image”
- Images are self-contained with all code, programs, environment, Dockerfile included
- No subsequent installation required
- Spin an image into a container using Docker or Singularity - it is like sharing your computer



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-

Data and code sharing

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 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

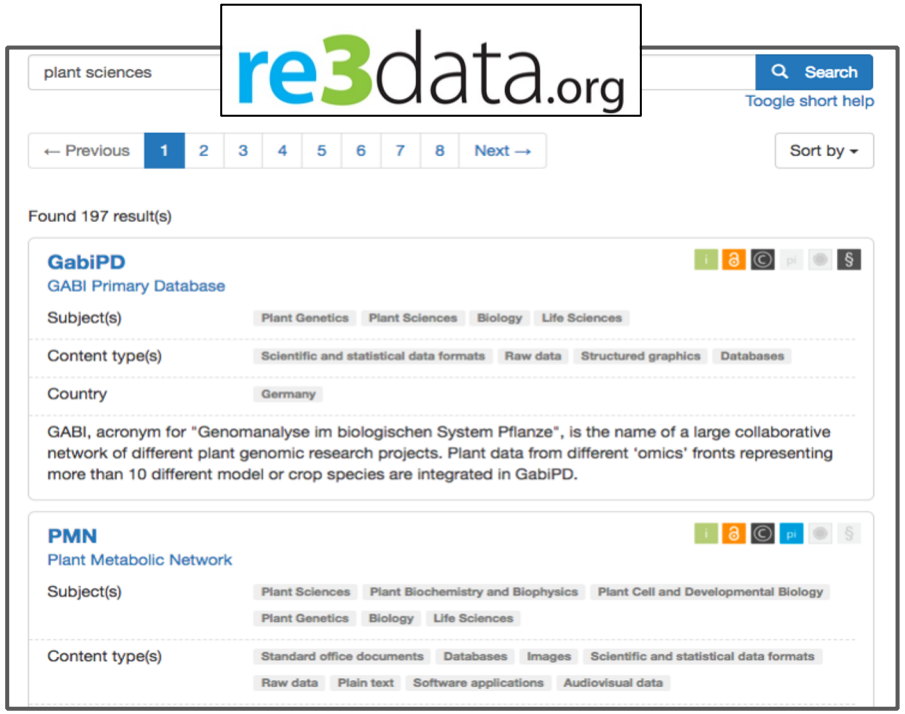
Data License

- 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)

Mandated vs. Disciplinary Repositories



Identify mandated / disciplinary repository:

- Funder specified repository
- Institutionally specified data repository
- Domain or discipline-specific data repository

Find & compare disciplinary repositories through Repository of Research Data Repositories https://www.re3data.org/

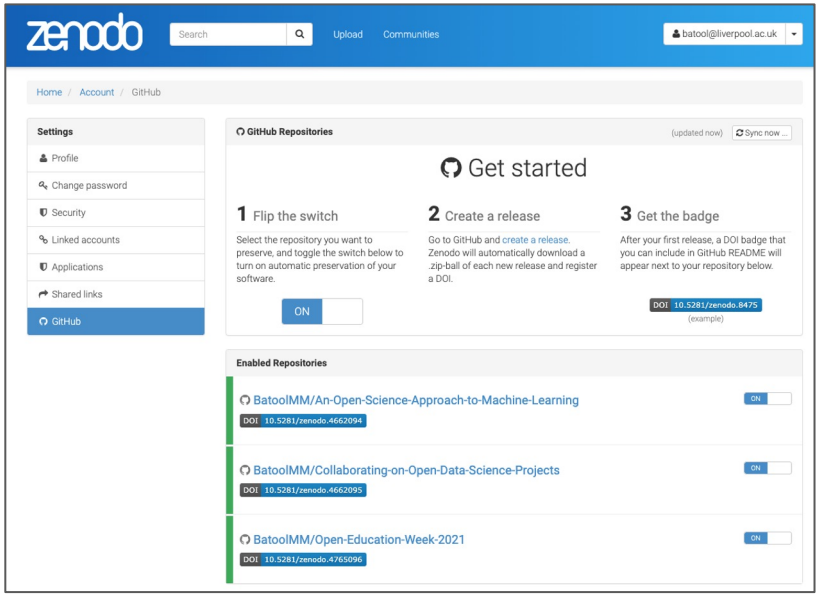
General Purpose Repositories

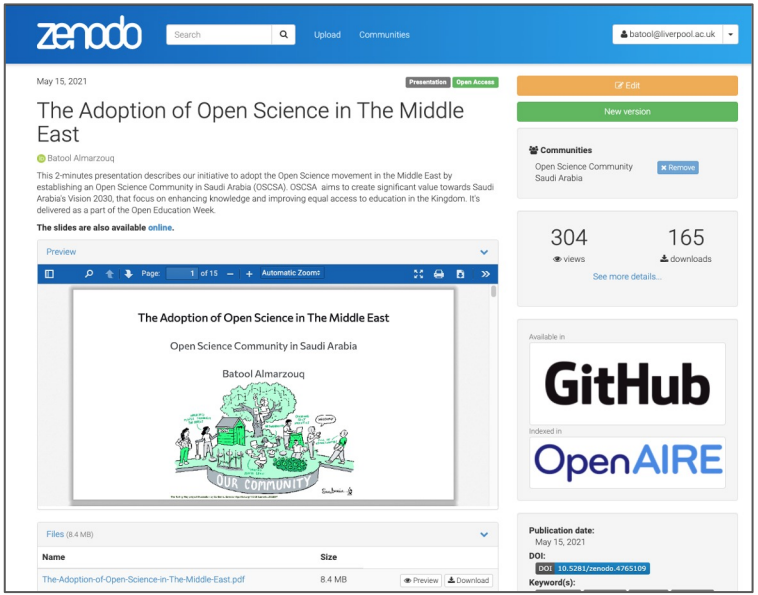
In addition to a specified data repository, you can make a deposit to a general purpose repository:

- DataDryad http://datadryad.org/ (curated digital repository; free to access, $120 to publish dataset up to 20GB)
- Figshare https://figshare.com/ (free digital repository, 5GB per file limit)
- Zenodo https://zenodo.org/ (free digital repository; 50GB per dataset limit)

re3data.org DRYAD figshare zenodo

Zenodo meets GitHub

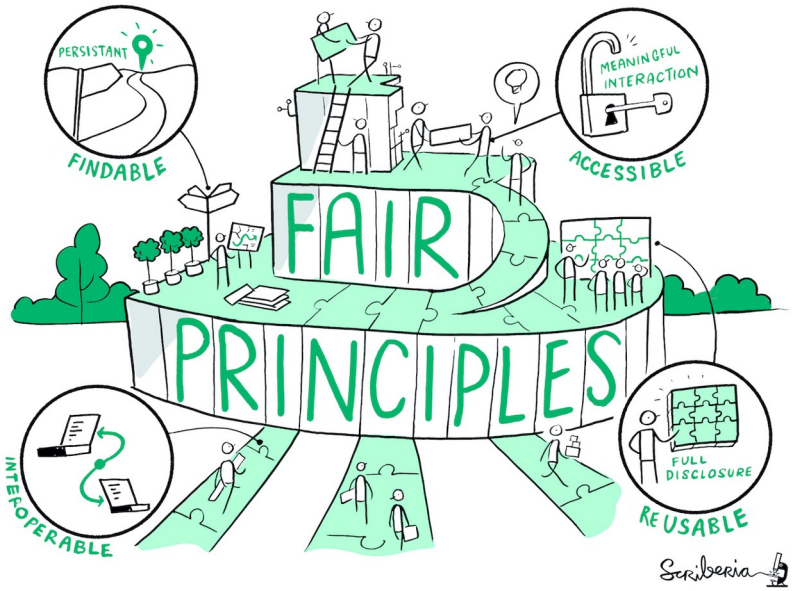




FAIR Data

- **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

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**

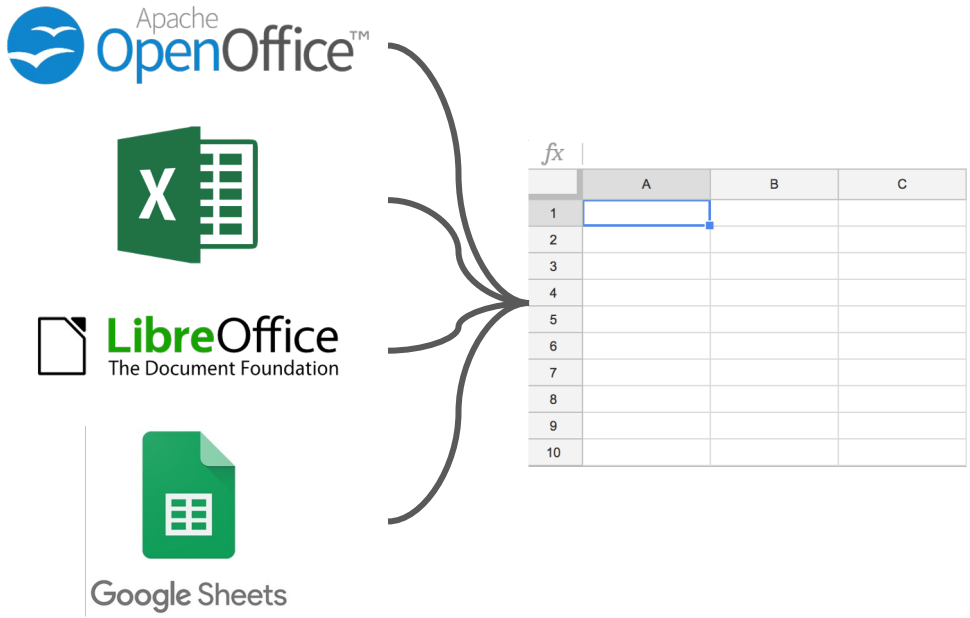
Illustrate **important findings**

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

Weissgerber et al. [10.1074/jbc.RA117.000147](https://doi.org/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

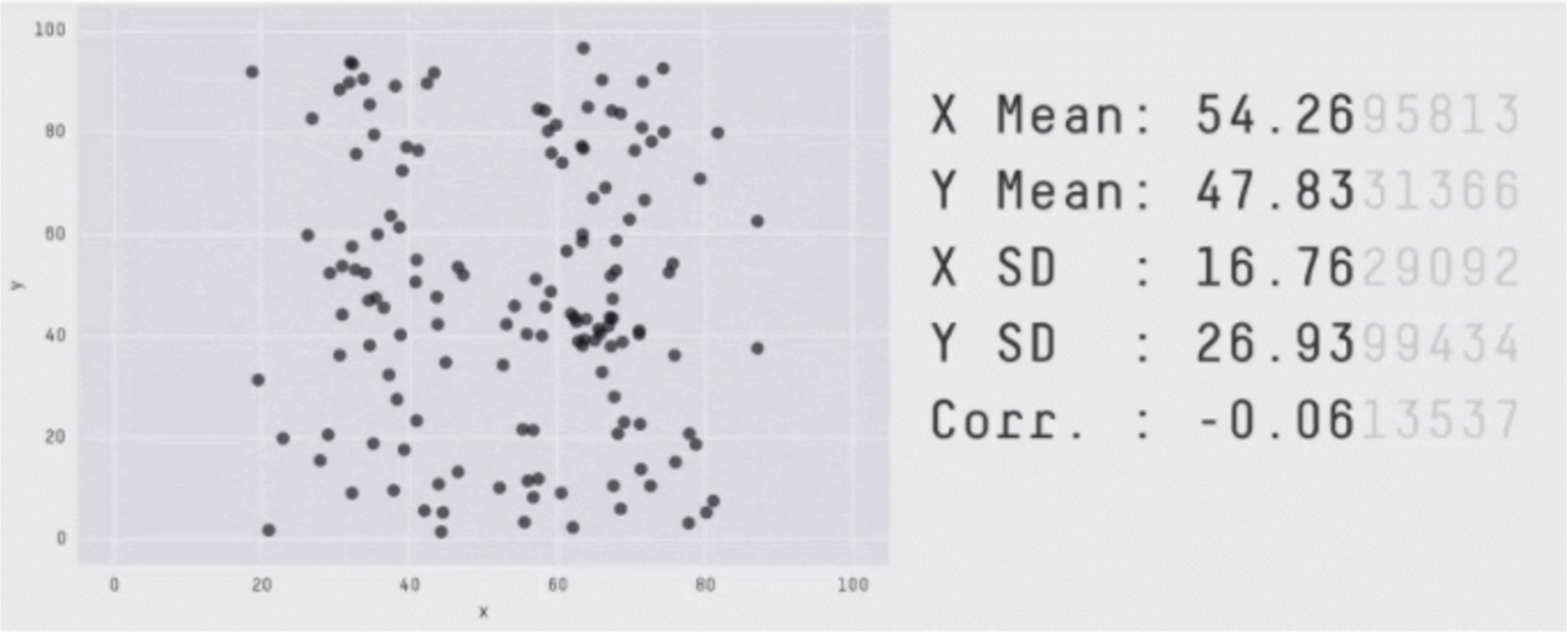
Weissgerber et al. 2017
10.1074/jbc.RA117.000147

Replicate
Reproduce
Reuse

Show your data

Never trust data based on the summary statistics alone!

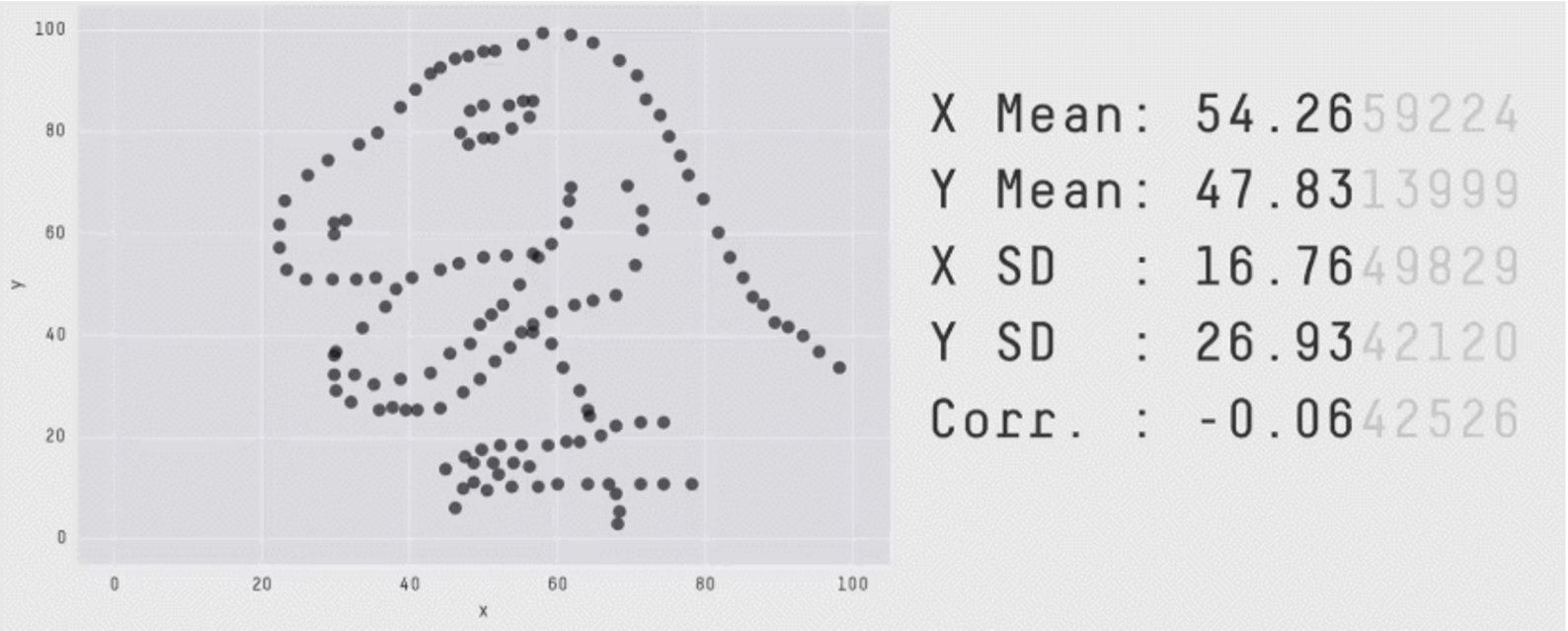
There really is NO alternative to showing the actual data



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

Never trust data based on the summary statistics alone!

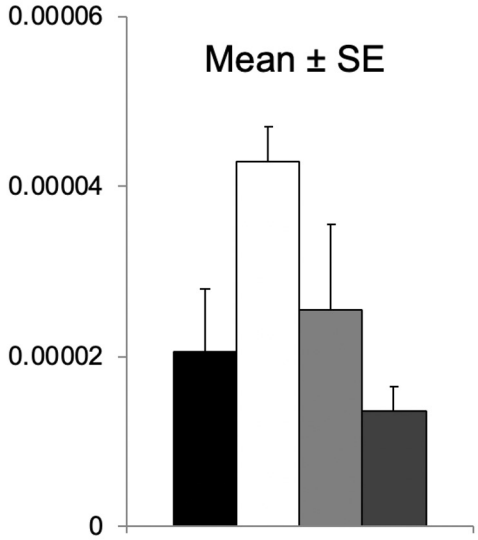
There really is NO alternative to showing the actual data



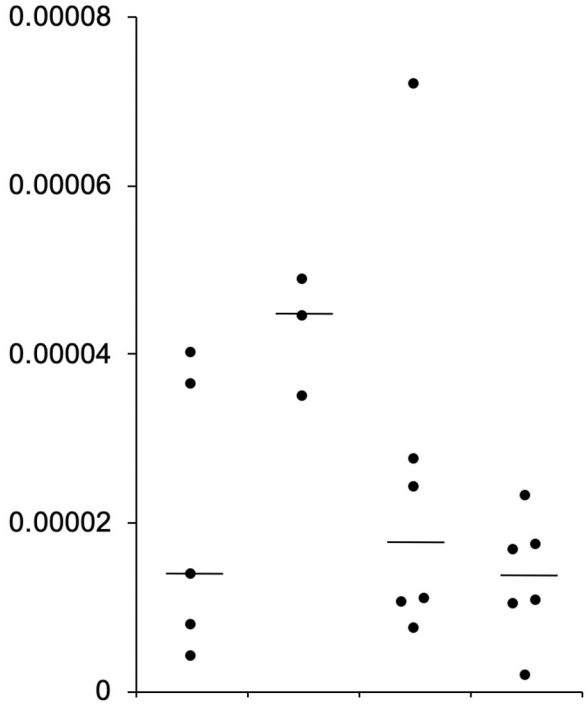
Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing

Our interpretation depends on what we see

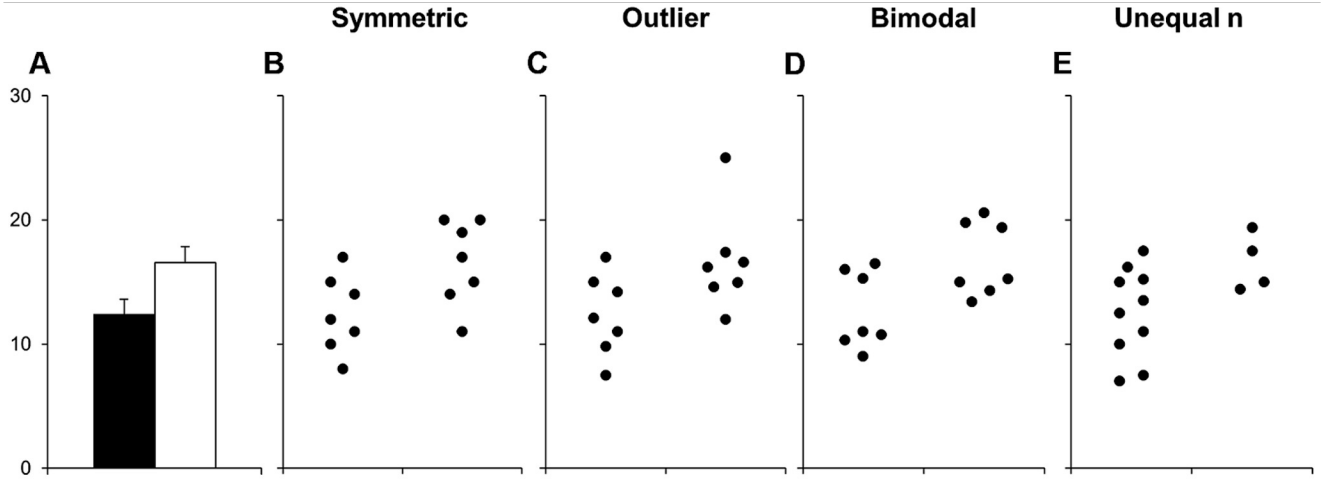
Reader is a passive observer



Reader is an active participant



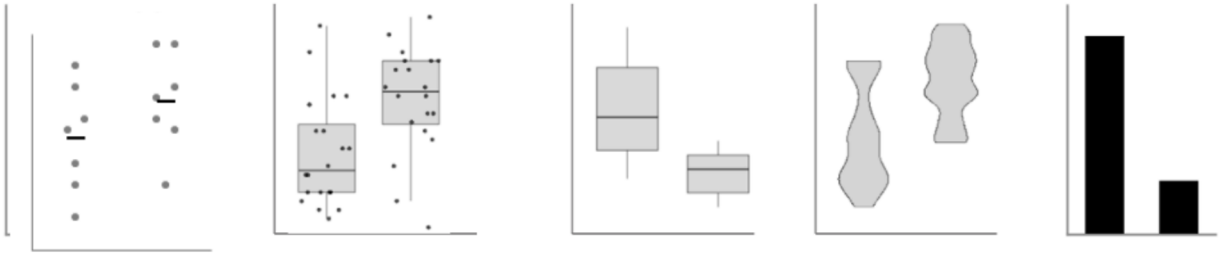
Avoid bar charts for continuous data



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

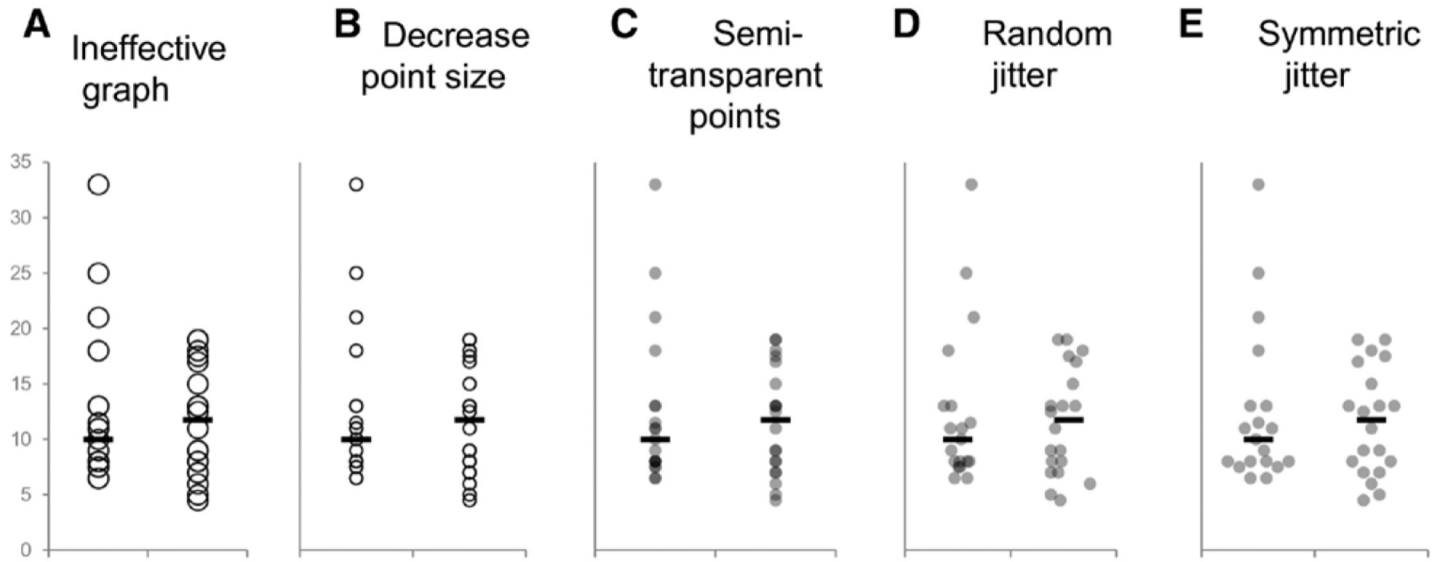


	Dotplot	Boxplot with points	Boxplot	Violin plot (with or without points)	Bar graph
Outcome variable	Continuous	Continuous	Continuous	Continuous	Counts & proportions
Sample size	Small	Medium	Large	Medium to Large	Any
Data distribution	Any	Any	Do not use for bimodal data	Any	N/A

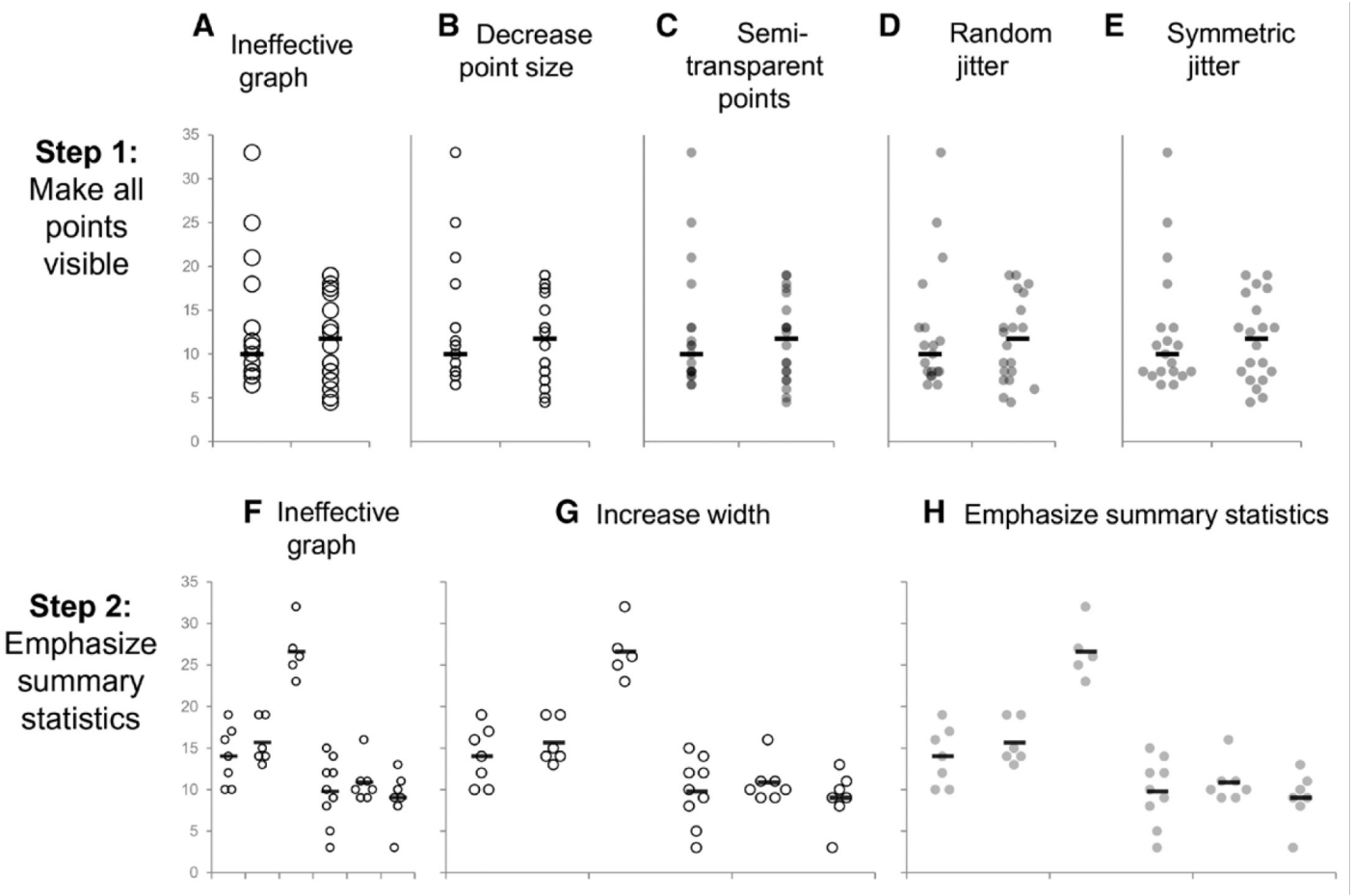
Weissgerber et al., 2017 JBC

Making effective dot plots

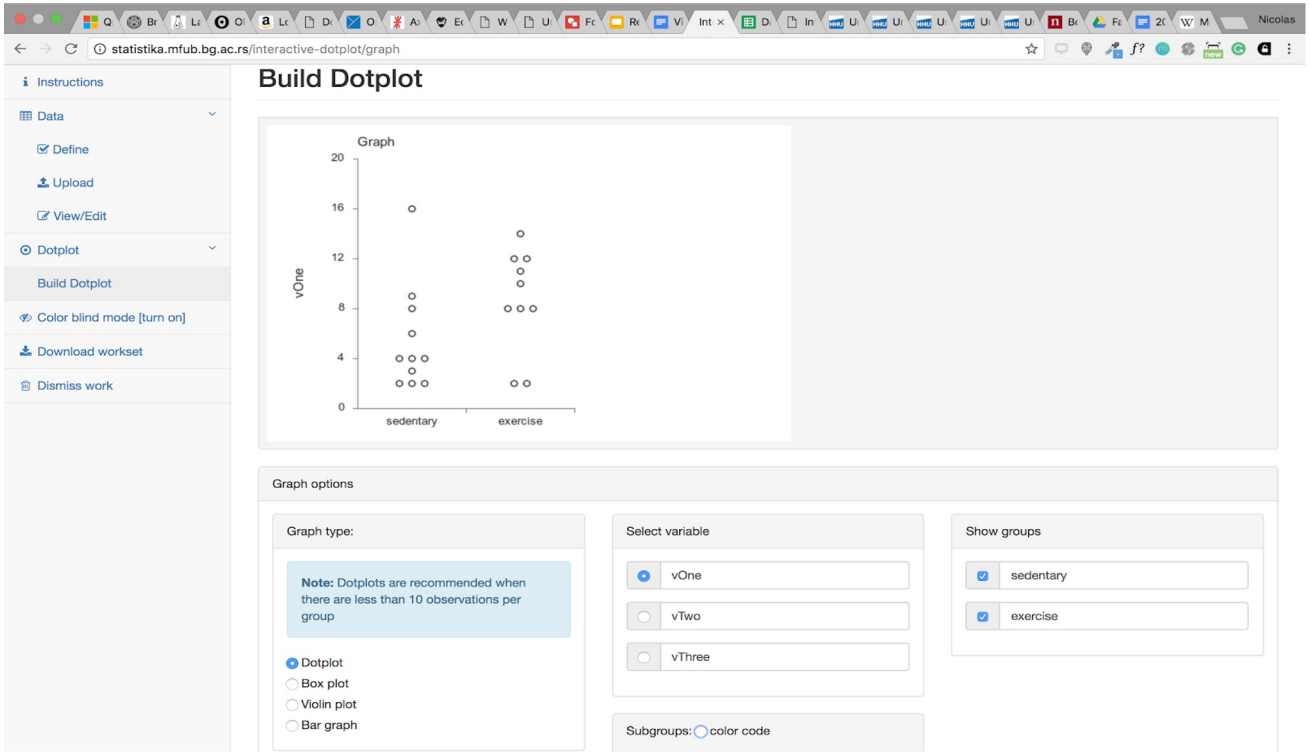
Step 1: Make all points visible



Making effective dot plots



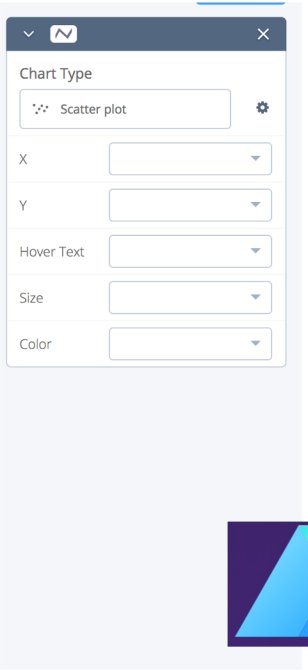
One step further: Interactive plots

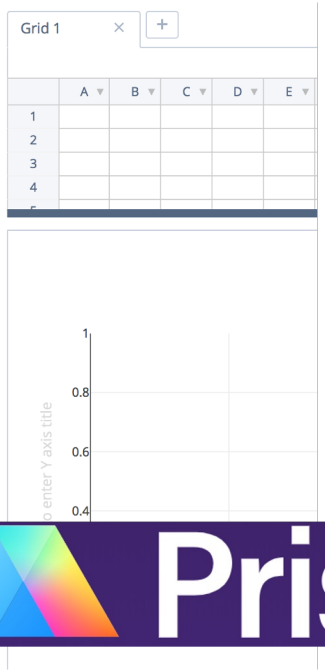


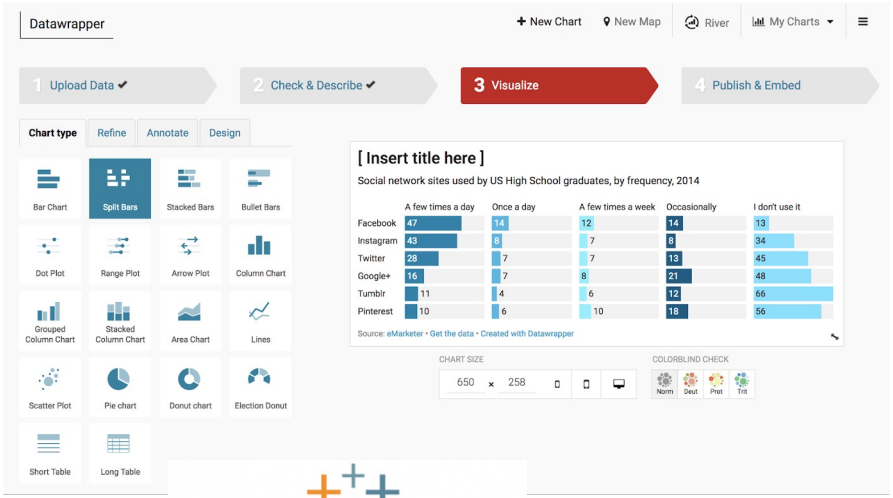
Interactive Dot Plot

Interactive Line Graph

Some intermediate options







Prism 8

+ableau

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

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

Current Funders

Chan Zuckerberg Initiative

addgene
The nonprofit plasmid repository

Past Funders

protocols.io

eLIFE

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)