

# 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

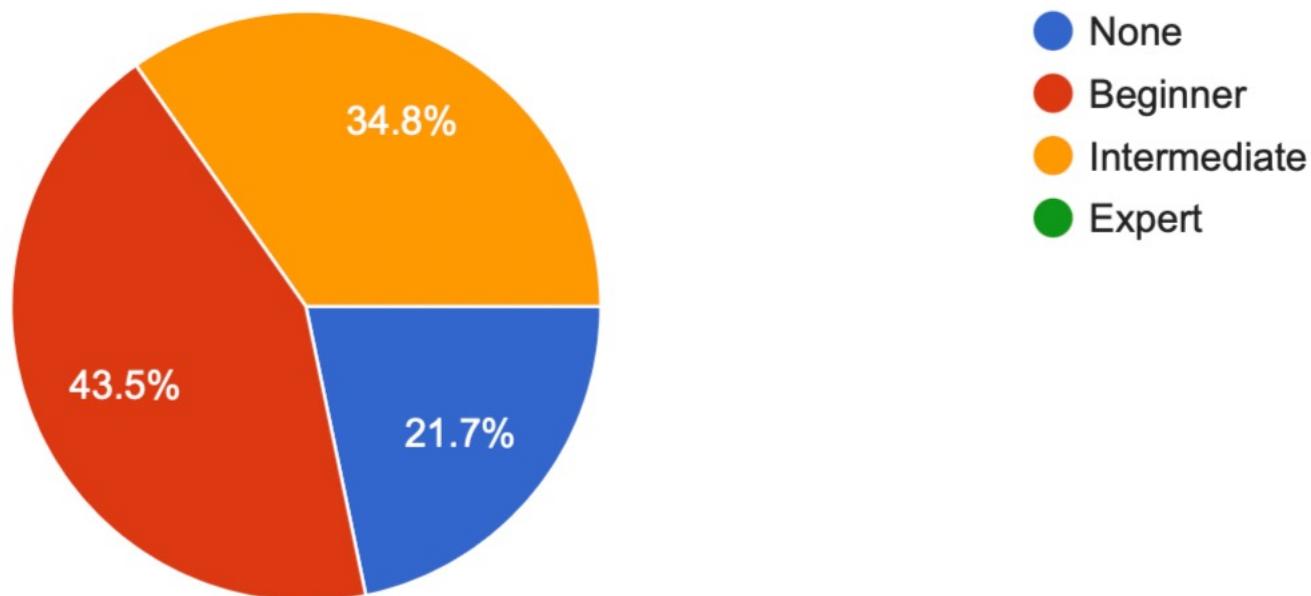
Faculty, Trivedi School of Biosciences

Ashoka University



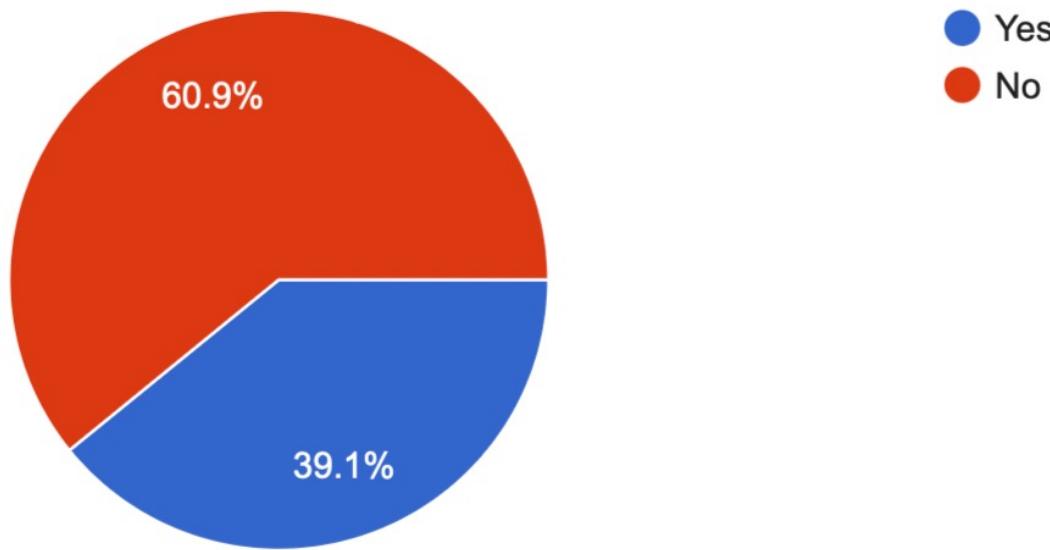
## Rate your knowledge of Biostatistics

23 responses



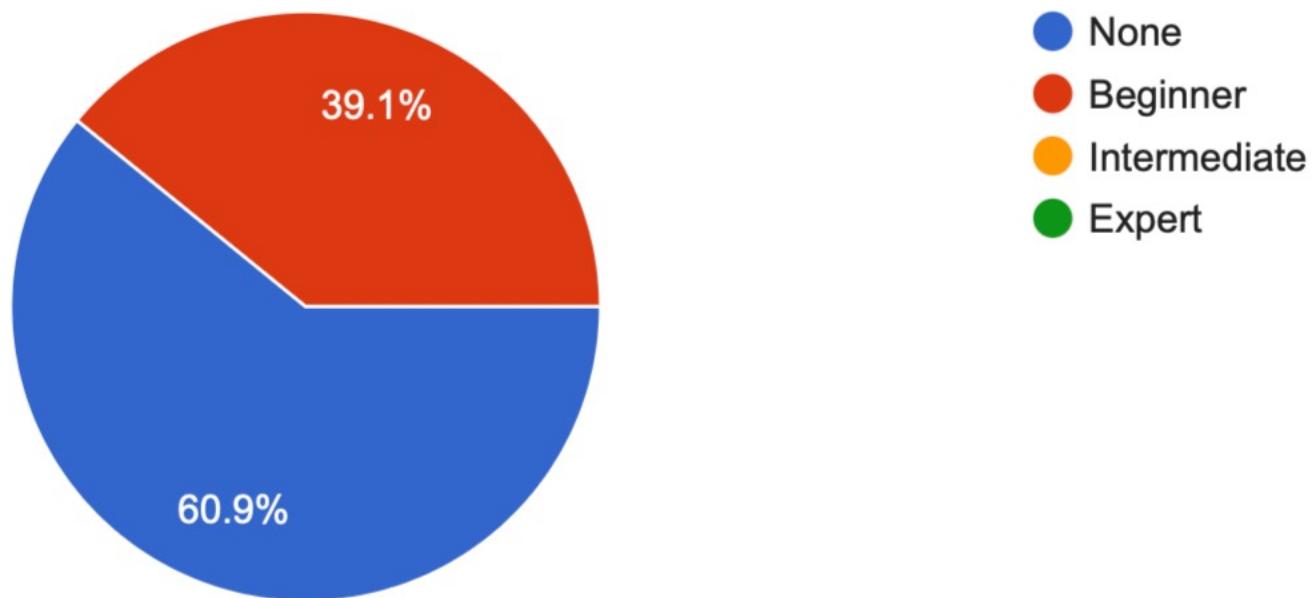
Are you aware of open source scientific computing languages?(python, R, Julia, etc.)

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>

# Thank You



The nonprofit plasmid repository



@repro4everyone  
[hello@repro4everyone.org](mailto:hello@repro4everyone.org)  
[www.repro4everyone.org](http://www.repro4everyone.org)

Chan  
Zuckerberg  
Initiative 

Dorothy Bishop



Benjamin Schwessinger  
@schwessinger

# If you want to read more about our initiative:



The image shows a thumbnail of an eLife journal article. At the top left is the eLife logo, which consists of a circular icon with colored dots (blue, green, red) followed by the word "eLife". To the right of the logo is the text "FEATURE ARTICLE". Further to the right are two icons: an open padlock symbol and a circular "CC" symbol. Below these elements is the text "SCIENCE FORUM". The main title of the article is "A community-led initiative for training in reproducible research", displayed in large, bold, dark blue capital letters.

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>

## **Introduction to reproducibility**

[Read More](#)

## **Data management**

[Read More](#)

## **Electronic lab notebooks**

[Read More](#)

## **Protocol sharing**

[Read More](#)

## **Reagent sharing**

[Read More](#)

## **Bioinformatics**

[Read More](#)

## **Code and data sharing**

## **Data visualization**

## **Publishing**

[Read More](#)

<https://www.repro4everyone.org/resources>

# Open Source Tools

for computational reproducibility

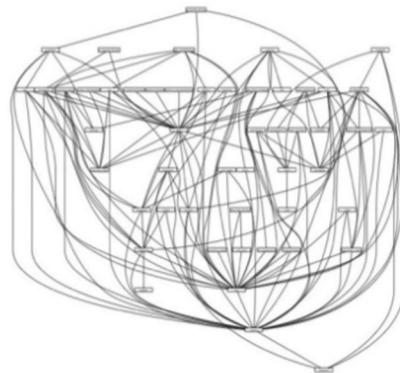
Two days hands-on workshop on Open Source Scientific  
Computing for Environmental Health Sciences

**ICMR-NIREH, Bhopal**  
27 June 2022



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

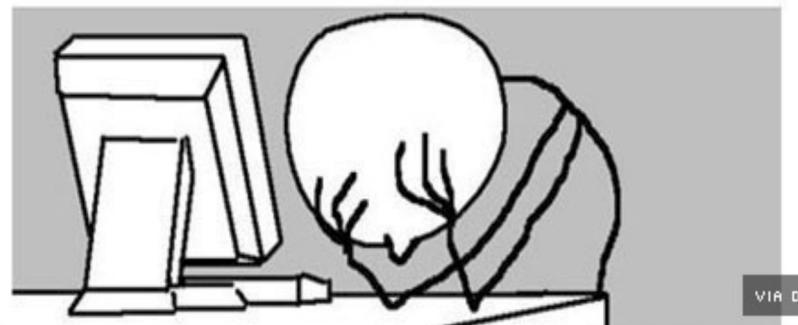
Dependency hell (files require other files)



<https://carpentries.org>

14

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/>



<https://www.rstudio.com>



# Jupyter notebooks - Example

File Edit View Run Kernel Tabs Settings Help

Files + notebooks  
Name Last Modified

- Data.ipynb an hour ago
- Fasta.ipynb a day ago
- Julia.ipynb a day ago
- Lorenz.ipynb seconds ago
- R.ipynb a day ago
- iris.csv a day ago
- lightning.json 9 days ago
- lorenz.py 3 minutes ago

Running

Commands

Cell Tools

Tabs

Lorenz.ipynb Terminal 1 Console 1 Data.ipynb README.md Python 3

In this Notebook we explore the Lorenz system of differential equations:

$$\begin{aligned}\dot{x} &= \sigma(y - x) \\ \dot{y} &= \rho x - y - xz \\ \dot{z} &= -\beta z + xy\end{aligned}$$

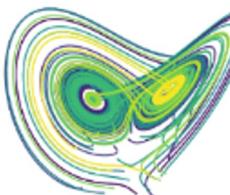
Let's call the function once to view the solutions. For this set of parameters, we see the trajectories swirling around two points, called attractors.

In [4]: `from lorenz import solve_lorenz  
t, x_t = solve_lorenz(N=10)`

Output View

lorenz.py

sigma 10.00  
beta 2.67  
rho 28.00



```
def solve_lorenz(N=10, max_time=4.0, sigma=10.0, beta=8./3, rho=28.0):
    """Plot a solution to the Lorenz differential equations."""
    fig = plt.figure()
    ax = fig.add_axes([0, 0, 1, 1], projection='3d')
    ax.axis('off')

    # prepare the axes limits
    ax.set_xlim(-25, 25)
    ax.set_ylim(-35, 35)
    ax.set_zlim(5, 55)

    def lorenz_deriv(x_y_z, t0, sigma=sigma, beta=beta, rho=rho):
        """Compute the time-derivative of a Lorenz system."""
        x, y, z = x_y_z
        return [sigma * (y - x), x * (rho - z) - y, x * y - beta * z]

    # Choose random starting points, uniformly distributed from -15 to 15
    np.random.seed(1)
    x0 = -15 + 30 * np.random(N, 3)
```



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

[git-scm.com/doc](http://git-scm.com/doc)

# 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

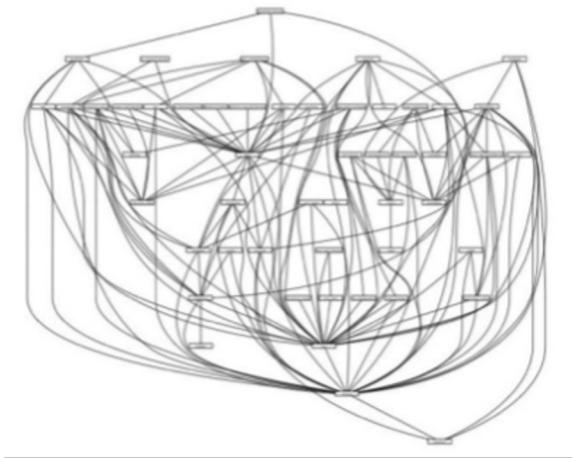


Git-based repository with a social media component.

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

# How did I install all these different software packages?

Dependency hell

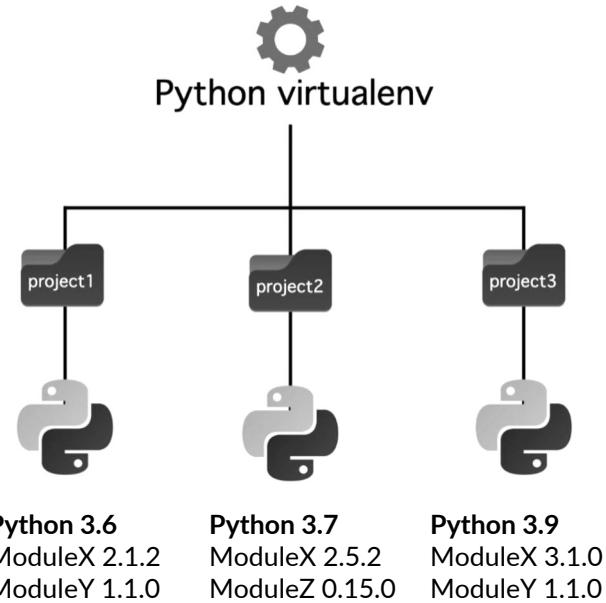


Version conflict



# 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

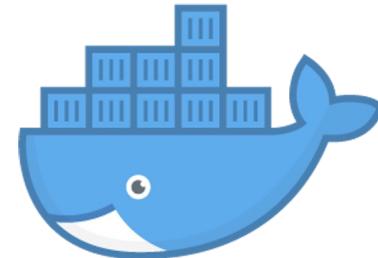


<https://conda.io/>

<https://bioconda.github.io/>

# Make your environment portable with containers

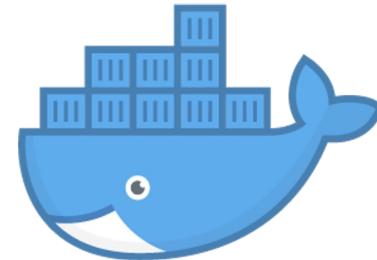
- Everything needed to run your analysis is packed up into an “image”
- Images are self-contained with all code, programs, environment, Dockerfile included
- No subsequent installation required
- Spin an image into a container using Docker or Singularity - it is like sharing your computer



<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-binder2-0>

# 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) regenerated**

# 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/ste\\_ps2rr/pages/licenses.html](http://kbroman.org/ste_ps2rr/pages/licenses.html)) and Open Source Initiative  
(<https://opensource.org/licenses>)

# Mandated vs. Disciplinary Repositories

The screenshot shows the re3data.org search interface. A search bar at the top contains the text "plant sciences". Below it is a search button with a magnifying glass icon and the word "Search". To the right of the search button is a link "Toggle short help". The main area displays a search result for "GabiPD". The result includes the title "GabiPD", a subtitle "GABI Primary Database", and a detailed description: "GABI, acronym for "Genomanalyse im biologischen System Pflanze", is the name of a large collaborative network of different plant genomic research projects. Plant data from different 'omics' fronts representing more than 10 different model or crop species are integrated in GabiPD." Below this, there are sections for "Subject(s)" (Plant Genetics, Plant Sciences, Biology, Life Sciences) and "Content type(s)" (Scientific and statistical data formats, Raw data, Structured graphics, Databases). A "Country" section indicates "Germany". At the bottom of the result card are several small icons. The page footer shows a navigation bar with links "← Previous", "1", "2", "3", "4", "5", "6", "7", "8", "Next →", and a "Sort by" dropdown menu. The overall page header also features the "re3data.org" logo.

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**



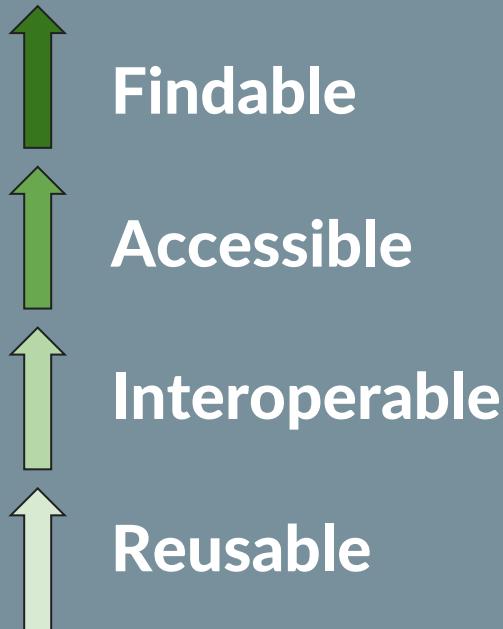
**zenodo**

# Zenodo meets GitHub

The screenshot shows the Zenodo GitHub integration interface. At the top, there's a navigation bar with 'Search', 'Upload', 'Communities', and a user dropdown. Below it, the main header says 'zenodo' and has a 'GitHub' button highlighted in blue. The left sidebar contains 'Settings' with options like 'Profile', 'Change password', 'Security', 'Linked accounts', 'Applications', 'Shared links', and 'GitHub' (which is selected). The main content area is titled 'GitHub Repositories' and includes a 'Get started' section with three steps: 1. Flip the switch (with an 'ON' button), 2. Create a release, and 3. Get the badge (with a DOI example: DOI: 10.5281/zenodo.8475). Below this, there's a list of 'Enabled Repositories' with three items: 'BatoolIMM>An-Open-Science-Approach-to-Machine-Learning' (DOI: 10.5281/zenodo.4662094), 'BatoolIMM>Collaborating-on-Open-Data-Science-Projects' (DOI: 10.5281/zenodo.4662095), and 'BatoolIMM>Open-Education-Week-2021' (DOI: 10.5281/zenodo.4765096).

The screenshot shows a Zenodo record for a presentation. The header says 'zenodo' and has a 'Presentation' button. The main title is 'The Adoption of Open Science in The Middle East' by 'Batool Almarzouq' on May 15, 2021. It features an 'Open Access' button and a green 'Edit' button. The record includes a 'Communities' section for 'Open Science Community Saudi Arabia', showing 304 views and 165 downloads. A thumbnail image shows a presentation slide with the title 'The Adoption of Open Science in The Middle East' and 'Open Science Community in Saudi Arabia'. The slide features a cartoon illustration of people working together around a tree labeled 'OUR COMMUNITY'. Below the slide, there's a file list for 'The-Adoption-of-Open-Science-in-The-Middle-East.pdf' (8.4 MB) with download and preview buttons. The publication date is May 15, 2021, and the DOI is DOI: 10.5281/zenodo.4765109.

# FAIR Data



Get organized! Be happy!



The Turing Way project illustration by Scriberia. Original version

on Zenodo. <http://doi.org/10.5281/zenodo.3695300>.

# Data Visualization & Analysis

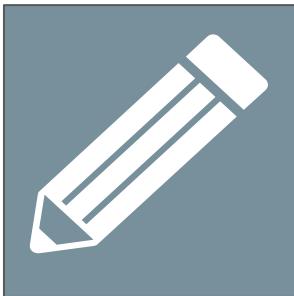


# Data presentation is the foundation of our collective scientific knowledge...

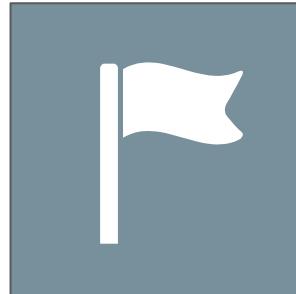


Figures are especially important. They often show data for key findings.

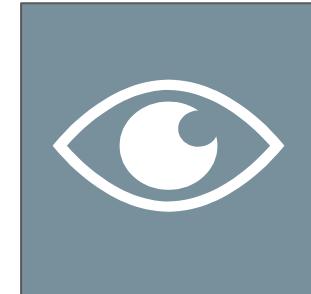
# Effective figures should...



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)

# The usual way and its flaws...



Google Sheets

A screenshot of an Excel spreadsheet. The formula bar at the top shows 'fx'. Below it is a table with three columns labeled A, B, and C. Row 1 contains values in cells A1 and B1. Row 2 is empty. Rows 3 through 10 are also empty. The table has a light gray background and thin black borders between cells.

## 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**

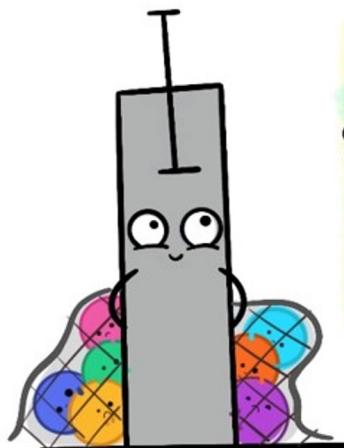
Weissgerber et al. 2017

[10.1074/jbc.RA117.00014](https://doi.org/10.1074/jbc.RA117.00014)

7

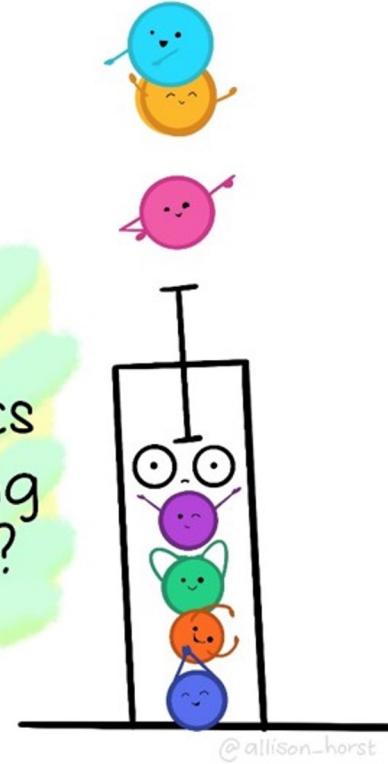


# Show your data



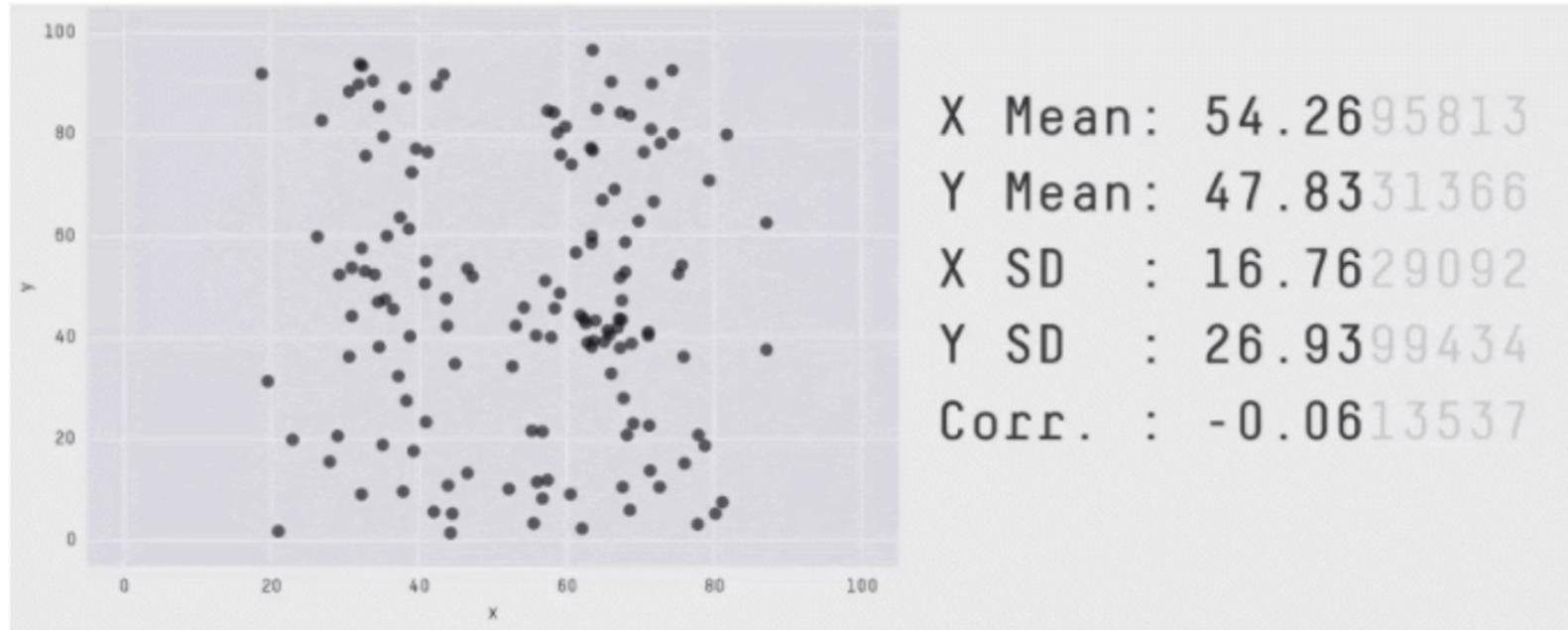
are your  
summary statistics  
hiding something  
interesting?

@AllisonHorst

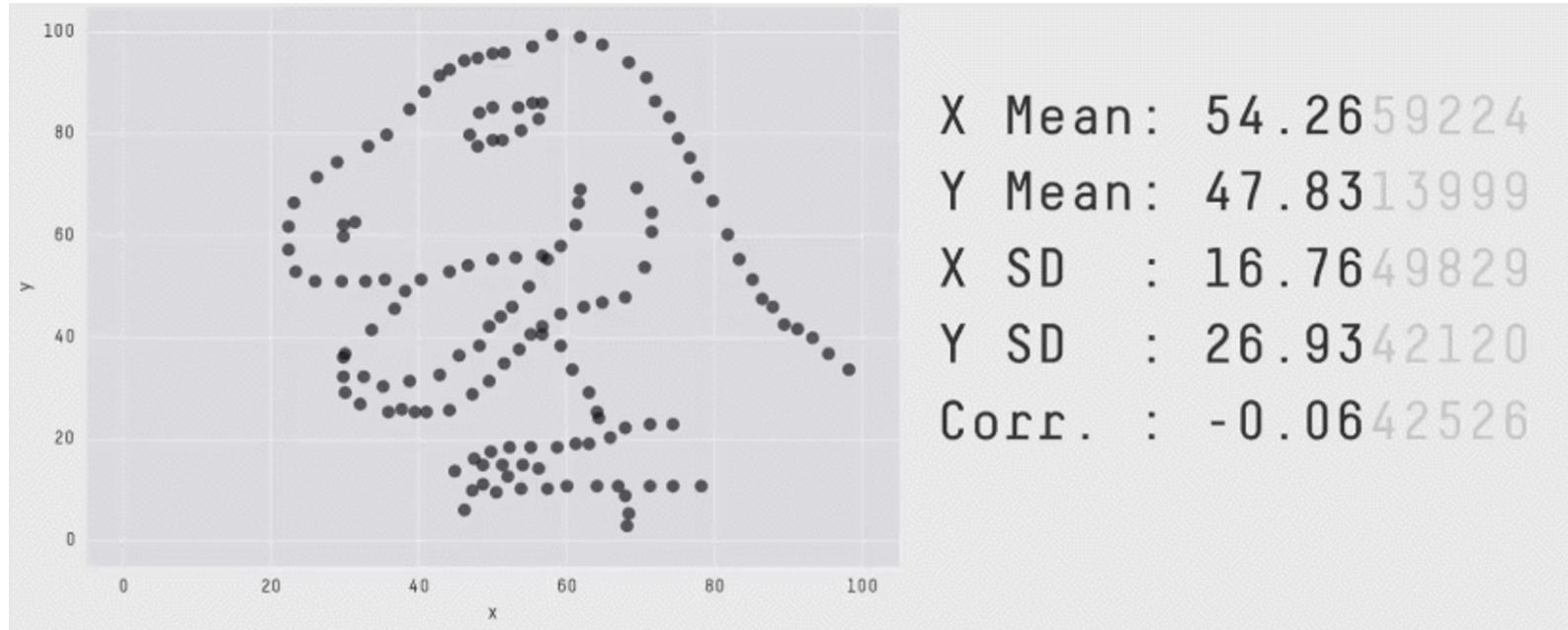


@allison\_horst

Never trust data based on the summary statistics alone!  
There really is NO alternative to showing the actual data



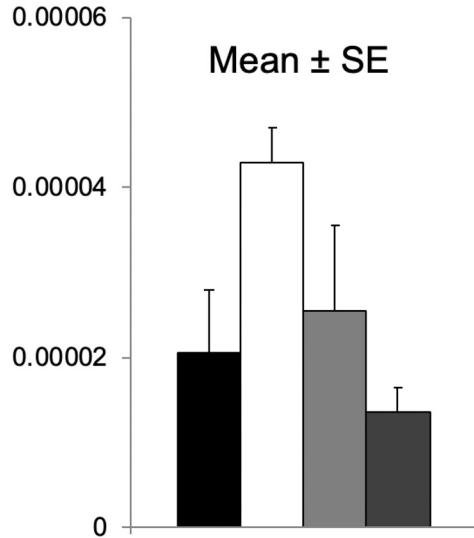
Never trust data based on the summary statistics alone!  
There really is NO alternative to showing the actual data



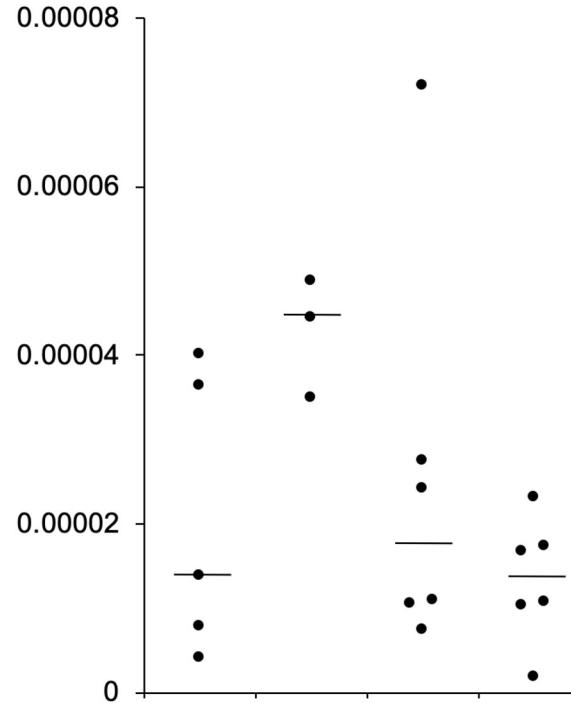
[Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing](#)

# Our interpretation depends on what we see

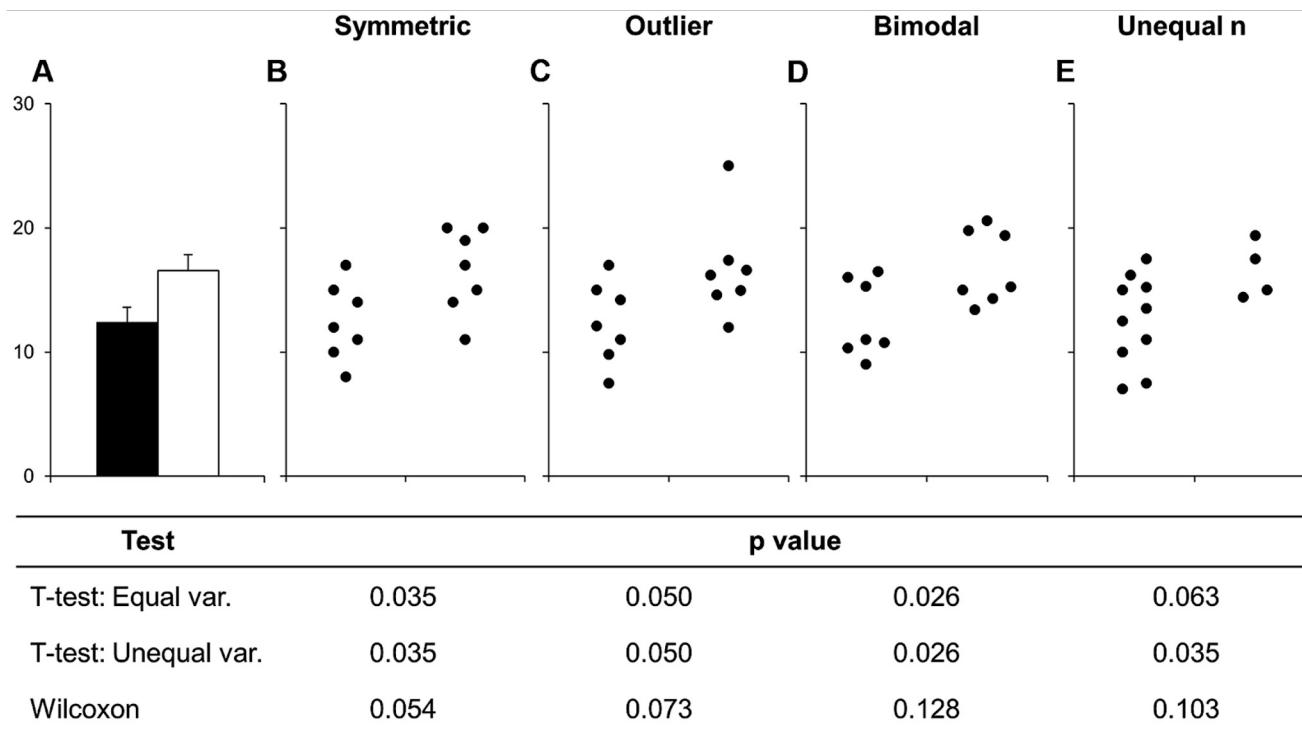
Reader is a  
passive observer



Reader is an  
active participant



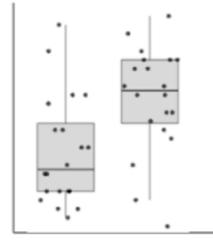
# Avoid bar charts for continuous data



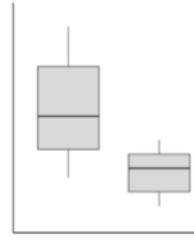
# How to choose the right plot



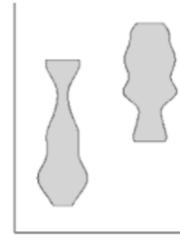
**Dotplot**



**Boxplot with points**



**Boxplot**



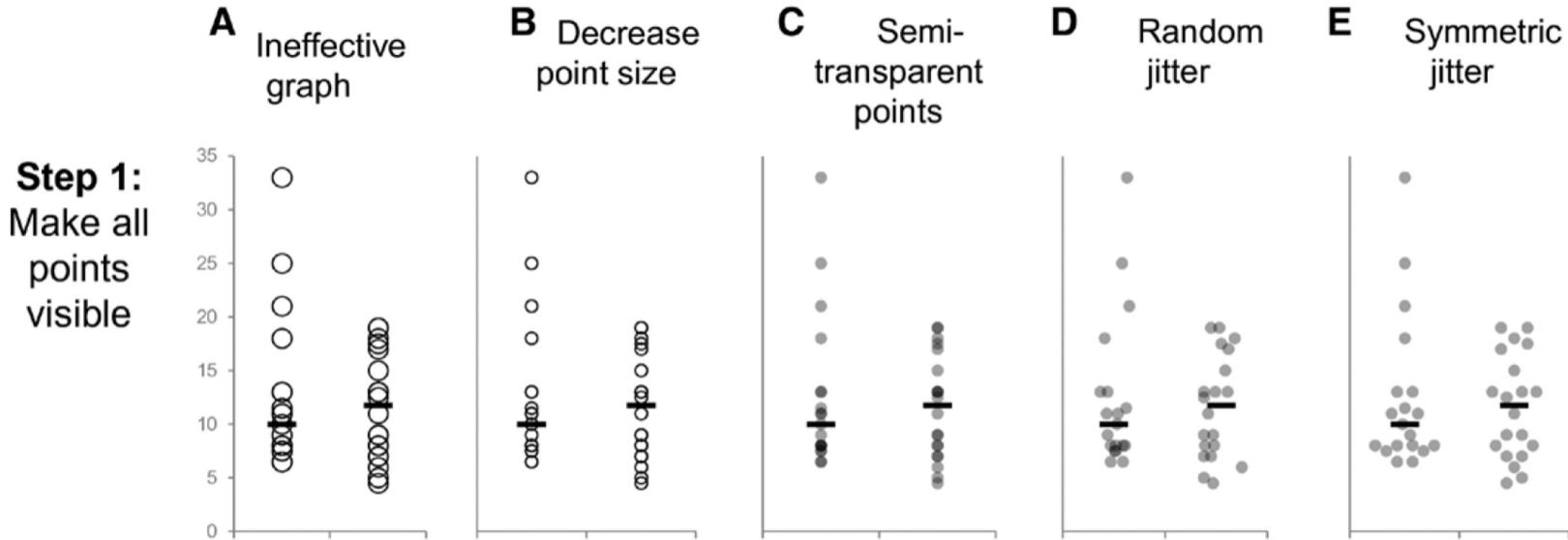
**Violin plot  
(with or  
without  
points)**



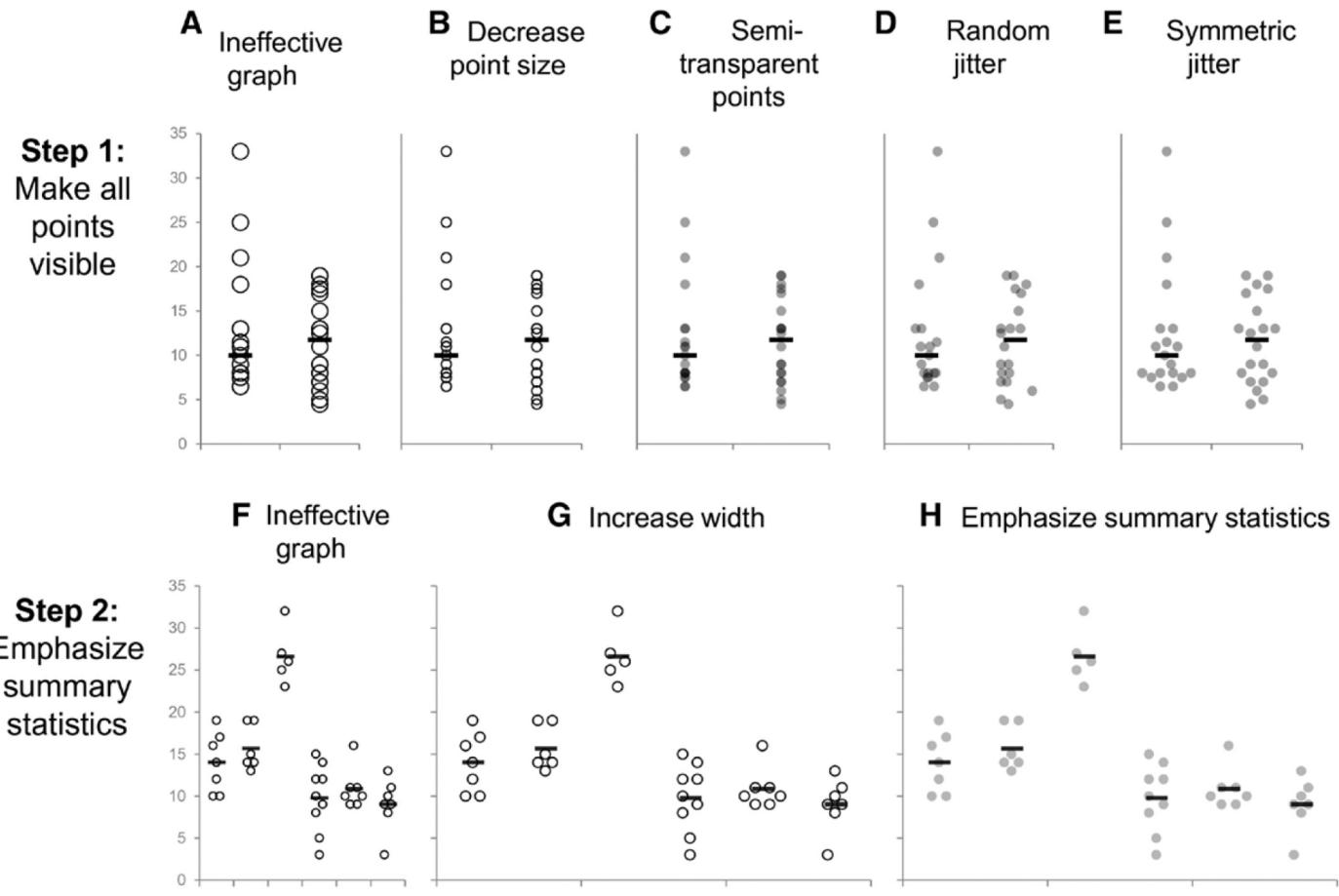
**Bar graph**

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

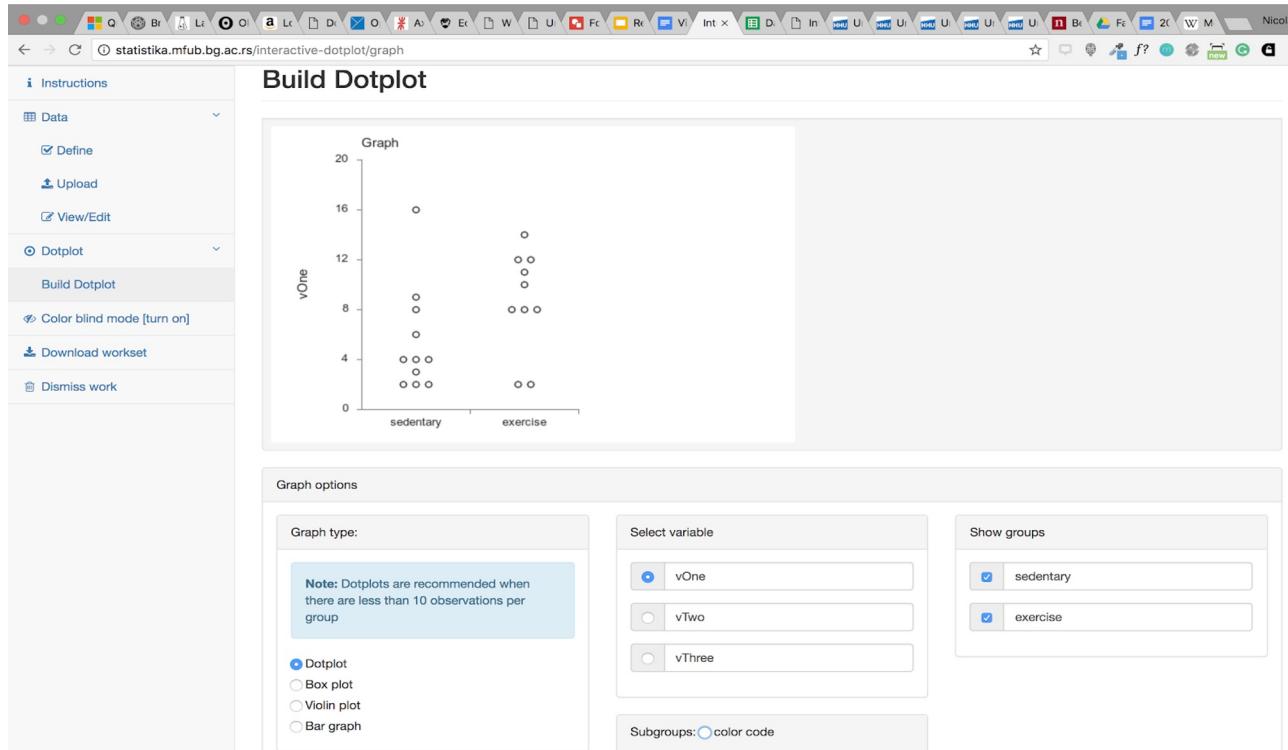
# Making effective dot plots



# Making effective dot plots



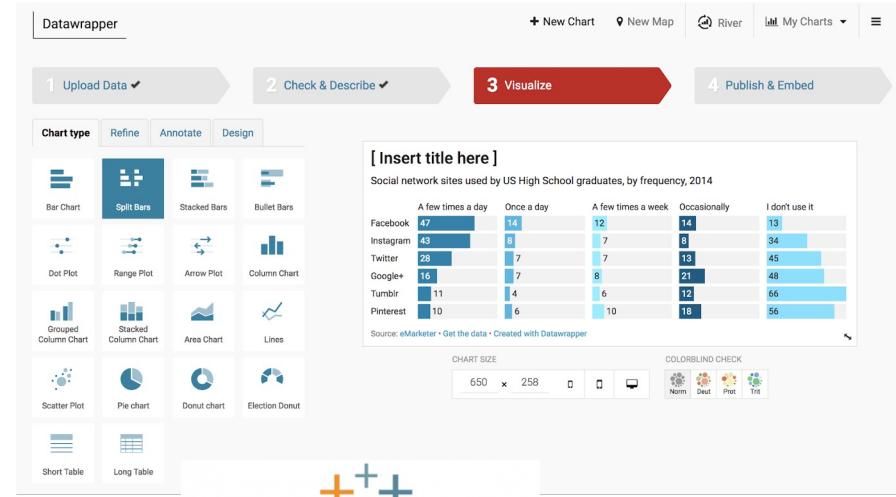
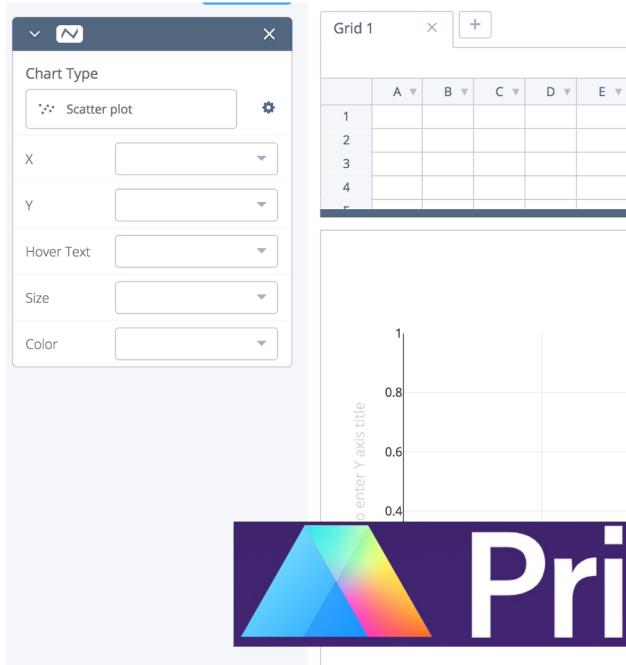
# One step further: Interactive plots



Interactive Dot Plot

Interactive Line Graph

# Some intermediate options



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

# Reproducible research practices enable you to:



Organize experiments productively



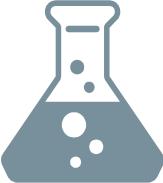
Accurately analyze results



Share results with future researchers



Share techniques



Share reagents with future researchers

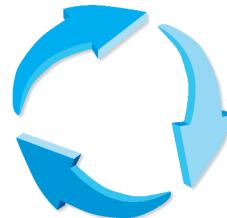


Accelerate science!

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

# Next Steps

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



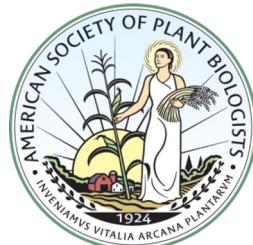
**Replicate**  
**Reproduce**  
**Reuse**

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

# Current Funders



# Past Funders



# 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)

