SuperPlotsOfData—a web app for the transparent display and quantitative comparison of continuous data from different conditions

Joachim Goedhart*
Swammerdam Institute for Life Sciences, Section of Molecular Cytology, van Leeuwenhoek Centre for Advanced Microscopy, University of Amsterdam, NL-1090 GE Amsterdam, The Netherlands

ABSTRACT **Plots and charts are graphical tools that make data intelligible and digestible by humans. But the oversimplification of data by only plotting the statistical summaries conflicts with the transparent communication of results. Therefore, plotting of all data are generally encouraged and this can be achieved by using a dotplot for discrete conditions. Dotplots, however, often fail to communicate whether the data are from different technical or biological replicates. The superplot has been proposed by Lord and colleagues (Lord *et al.*, 2020) to improve the communication of experimental design and results. To simplify the plotting of data from discrete conditions as a superplot, the SuperPlotsOfData web app was generated. The tool offers easy and open access to state-of-the-art data visualization. In addition, it incorporates recent innovations in data visualization and analysis, including raincloud plots and estimation statistics. The free, open source webtool can be accessed at: https://huygens.science.uva.nl/SuperPlotsOfData/.**

Monitoring Editor
Thomas Pollard
Yale University

Received: Sep 10, 2020
Revised: Jan 7, 2021
Accepted: Jan 13, 2021

INTRODUCTION

Graphs have a key role in the communication of experimental results in lab meetings, presentations, and manuscripts. Over the years, efforts have been made to increase the transparency in reporting of results and this has led to the recommendation to plot all data, rather than just a summary (Drummond and Vowler, 2011; Wilcox and Rousselet, 2018; Weissgerber *et al.*, 2019). For the display of continuous data from discrete conditions this means that instead of, or on top of, a bar that summarizes the data, the dot is the geometry of choice as it enables the display of all data. Dotplots can be generated by several popular commercial software packages, but there are also free, open source solutions (Spitzer *et al.*, 2014; Mauri *et al.*, 2017; Weissgerber *et al.*, 2017; Postma and Goedhart, 2019).

We have previously created PlotsOfData to facilitate visualization of continuous data from different (discrete) conditions (Postma and Goedhart, 2019). Under the hood, PlotsOfData uses the statistical computing software R and several packages, including ggplot2 for state-of-the-art data visualization (Wickham, 2011). The software is operated by graphical user interface in a web browser. As a result, PlotsOfData is a universally accessible open source tool that delivers high quality plots, without the need for coding skills and with a minimal learning curve.

In addition to the transparent display of the data, information about the experimental design is necessary to interpret the results. For instance, it should be clear whether data are paired, how many technical and biological replicates are plotted, and how "n" is defined for different conditions (Naegle *et al.*, 2015; Lazic *et al.*, 2018). The "superplot" has been proposed by Lord and colleagues as a way to make this clear in a plot (Lord *et al.*, 2020). A superplot identifies different replicates by color and/or shape and uses the average of each (biological) replicate as input for the comparison of conditions. An additional benefit of the explicit identification of biological replicates, instead of using the aggregated technical replicates, is that realistic *p* values are obtained.

Other approaches that aim to improve data visualization and analysis have recently been proposed, that is, raincloud plots (Allen *et al.*, 2018) and the use of estimation statistics

This article was published online ahead of print in MBoC in Press (http://www.molbiolcell.org/cgi/doi/10.1091/mbc.E20-09-0583) on January 21, 2021.
Data availability: All data and code are available in public repositories (GitHub and Zenodo) as referenced in the manuscript.
Competing interests: The authors declare no competing interests.
ORCID: 0000-0002-0630-3825; Twitter: @joachimgoedhart.
*Address correspondence to: Joachim Goedhart (j.goedhart@uva.nl).
Abbreviations used: CI, Confidence Interval; CSV, comma separated values; URL, uniform resource locator.

(Cumming, 2014; Claridge-Chang and Assam, 2016; Ho *et al.*, 2019). The benefit of an open source tool (such as PlotsOfData) that is based on a powerful statistical computing language with excellent graphics is that these innovations are readily incorporated. In contrast, implementing these new ideas in commercial software requires multistep workarounds (Lord *et al.*, 2020).

Here, SuperPlotsOfData is presented, which builds on the PlotsOfData web tool and uses the same philosophy of providing easy and open access to state-of-the-art data visualization. In addition, it enables the identification of replicates as a superplot and it incorporates recent innovations, including raincloud plots and estimation statistics.

Availability, code, and issue reporting

The SuperPlotsOfData app is available at: https://huygens.science.uva.nl/SuperPlotsOfData.

The code was written using R (https://www.r-project.org) and Rstudio (https://www.rstudio.com). To run the app, several freely available packages are required: shiny, ggplot2, magrittr, dplyr, readr, tidyr, ggbeeswarm, readxl, DT, broom and RCurl. This version of the manuscript is connected to version 1.0.3 of the web app (https://github.com/JoachimGoedhart/SuperPlotsOfData/releases/tag/v1.0.3), which is archived at zenodo, doi: https://doi.org/10.5281/zenodo.4423341.

Up-to-date code and new release will be made available on Github, together with information on running the app locally: https://github.com/JoachimGoedhart/SuperPlotsOfData/.

The Github page of SuperPlotsOfData is the preferred way to communicate issues and request features (https://github.com/JoachimGoedhart/SuperPlotsOfData/issues). Alternatively, the users can contact the developers by email or Twitter. Contact information is found on the "About" page of the app.

Data input and data format

The default data structure is the tidy format (Wickham, 2014) and an example is shown in Table 1. The minimum input consists of a column with conditions and a column with measured variables. A third column that identifies the replicates is recommend to take full advantage of the application.

Data are, however, often organized and stored in spreadsheets. The web app supports this type of "wide" data, when it is structured according to Table 2. Users are asked to define the number of rows that are used to define the conditions and replicates. For a superplot, typically two rows are necessary, one for conditions and another for specifying the replicates. Users can identify the information of each of the rows with a label. After upload, the data will be converted to the tidy format.

Condition	Replicate	Value
Control	Replicate1	7.1
Control	Replicate1	6.7
Control	Replicate1	6.3
Drug	Replicate1	9.9
Drug	Replicate1	9.5
Drug	Replicate1	8.5
Control	Replicate2	5.1
Control	Replicate2	4.5
Control	Replicate2	4.4
Drug	Replicate2	8.7
Drug	Replicate2	9.1
Drug	Replicate2	8.5

TABLE 1: Synthetic data in the tidy format, where each row is an observation and all measured values are in one column.

Control	Drug	Control	Drug
Replicate1	Replicate1	Replicate2	Replicate2
7.1	9.9	5.1	8.7
6.7	9.5	4.5	9.1
6.3	8.5	4.4	8.5

TABLE 2: Data, identical to the data in Table 1, in a spreadsheet format. The first two rows specify the condition and replicate.

The data can be supplied as a CSV file, XLS(X) file, by copy-paste or through a URL (CSV files only). One example dataset is available to demonstrate the data structure and for testing the app.

After data upload, the user selects the columns that hold the information on the Conditions, Measurements, and (optional) Replicates. When no replicates are selected, the measurements are grouped per condition.

Data visualization

When the data are composed of different replicates, these are indicated for each condition with a different color and/or a different symbol as suggested before (Galbraith *et al.*, 2010; Weissgerber *et al.*, 2017; Lord *et al.*, 2020). The mean or median of each of the replicates is indicated with a larger dot (Figure 1). Lines are drawn between the means or medians from the same sample when the data are qualified by the user as "paired." Pairing of the data will affect the statistics for the quantitative comparison of conditions, as will be explained below.

Statistics

Technical replicates. A median or mean can be selected as the summary statistic for each replicate and this is shown in the graphs as a larger dot. Several other statistics for individual replicates are available under the Data Summary tab and include n, standard deviation standard error of the mean and the 95% confidence interval (CI). The table (Figure 2) with the statistics for the individual replicates lists the *p* value from a Shapiro–Wilk test which can be used to evaluate whether the data of the replicates are normally

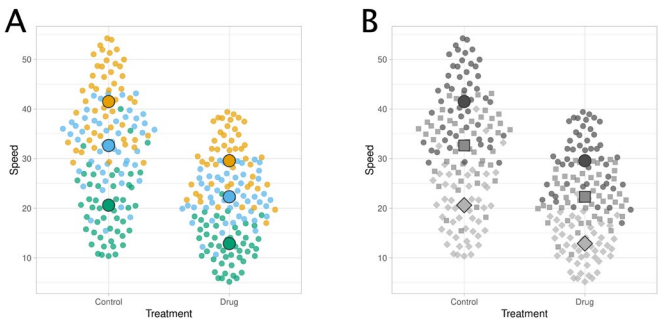

FIGURE 1: Output of the application based on example data. (A) By default, the (biological) replicates are identified by unique, colorblind friendly colors and the mean of each replicate is indicated with a larger dot. (B) An alternative presentation of the same data that uses different symbols and gray values to identify replicates.

Table 1: Statistics for individual replicates

Copy CSV Excel PDF Search:

	Condition	Replica	n	mean	sd	sem	95%CI_lo	95%CI_hi	median	p(Shapiro-Wilk)
1	Control	1	50	41.49	7.48	1.07	39.34	43.64	41.7	0.034
2	Control	2	50	32.64	6.88	0.98	30.66	34.61	34.37	0.061
3	Control	3	50	20.62	6.84	0.98	18.66	22.59	20.29	0.099
4	Drug	1	50	29.57	6.09	0.87	27.83	31.32	30.05	0.084
5	Drug	2	50	22.32	4.68	0.67	20.98	23.66	21.92	0.263
6	Drug	3	50	12.9	4.4	0.63	11.63	14.16	12.55	0.497

Showing 1 to 6 of 6 entries Previous 1 Next

Table 2: Statistics for conditions

Copy CSV Excel PDF Search:

	Condition	n	mean	sd	sem	95%CI_lo	95%CI_hi
1	Control	3	31.58	10.48	7.41	-0.29	63.45
2	Drug	3	21.6	8.36	5.91	-3.83	47.03

Showing 1 to 2 of 2 entries Previous 1 Next

Table 3: Statistics for differences between conditions

Copy CSV Excel PDF Search:

	Condition	difference	95%CI_lo	95%CI_hi	p.value
1	Control vs Drug	-9.99	-31.89	11.92	0.2695074

Showing 1 to 1 of 1 entries Previous 1 Next

FIGURE 2: A screenshot of statistics from the example data that are calculated by SuperPlotsOfData. The statistics are available under the Data Summary tab. Each of the tables can be downloaded in a number of formats.

distributed. A high p value provides evidence for a normal distribution. When the majority of the replicates shows a deviation from normality (threshold for $p < 0.05$), the user is notified and advised to use the median as a summary statistic.

Biological replicates. The statistics for each condition are presented in a second table that is available under the Data Summary tab. The number of (biological) replicates, the mean, standard deviation, and standard error of the mean and 95% CI are displayed.

Comparing conditions. The aim of an experiment with different conditions is often to detect a difference between those conditions. One way to do this is in a plot is by comparing the 95% CI. SuperPlotsOfData has an option to display the mean and 95% CI to enable inference by eye (Cumming and Finch, 2005; Cumming, 2009). Alternatively, the mean can be displayed together with the SD to summarize the spread in the data that is measured for a condition.

The predominant statistical methods for the quantitative comparison of data are a null-hypothesis significance test (NHST). A low p value provides evidence for a statistical difference between conditions. SuperPlotsOfData offers an ordinary t test for a difference between the means of the of the individual replicates. Depending on the experimental design, the data can be paired or connected. To highlight a paired relation, lines can be added to connect the mean or median values of the replicates. This will affect the result of the statistical analysis and a notification is displayed by the app. A paired t test is done when the means are connected with lines and an unpaired t test with correction for unequal variances (also known as Welch's t test) is done when the means are not connected. The table with statistics can be displayed under the plot and is also available under the Data Summary tab.

Instead of testing for statistical significance, it is often more interesting and biologically relevant to answer the question: how large is the difference? (Cumming, 2014). The calculated difference between conditions and its 95% CI is also known as the effect size and this type of analysis is termed estimation statistics (Claridge-Chang and Assam, 2016). In SuperPlotsOfData there is an option to display the difference between a reference condition, which can be selected by the user, and the other conditions.

Optimizing the visualization

Scientific graphs often represent data in an unpolished format with default settings that are not optimized to communicate the data in an easy-to-digest manner. In marked contrast, the field of data visualization focuses on "storytelling" and aims to convey the story that is told by the data with a clear and compelling illustration. Several of the principles of storytelling may aid the construction of graphs that are easier to understand. SuperPlotsOfData has several features that improve the plot by making the communication and interpretation of the data more effective, including 1) the option to sort data according to the measured variable, 2) the choice to rotate the graph by 90 degrees to improve the readability of the conditions, 3) effective use of colorblind friendly colors, and 4) the option to switch off the gridlines. Finally, a dark theme is available to generate plots for presentation or websites that use a dark background.

Output

The graphs can be downloaded in PDF, SVG, or PNG format. The PDF format enables editing of figures in software applications that accept vector-based graphics. The statistics can be downloaded in CSV or Excel file format.

A snapshot of the current setting can be made by the "Clone current setting" button, which returns a URL with that encodes the user-defined settings of the current session. When the data are imported from a web address, the graph can be stored and exchanged. This option enables a reproducible user interface.

In addition, the option to retrieve a hyperlink of the current setting facilitates data sharing and reuse. For details, the reader is referred to the papers that report our other apps in which this feature is implemented (Postma and Goedhart, 2019; Goedhart and Luijsterburg, 2020). The hyperlink (URL) that corresponds to the setting used in the plots shown in Figures 1 and 3 are listed in Table 3.

Limitations

A limitation of the app is the absence of a dedicated statistical analysis, which has several reasons. First, the main purpose of the app is to

Figure	URL
1A	https://huygens.science.uva.nl:/SuperPlotsOfData/?data=2
1B	https://huygens.science.uva.nl/SuperPlotsOfData/?data=2&vis=;;0.7;;;;TRUE;1;none&layout=No;;;;;;;1
3A	https://huygens.science.uva.nl:/SuperPlotsOfData/?data=1;;Treatment;Speed;Replicate&vis=quasirandom;;0.7;mean;solid;;;1;none&layout=No;;;;;;;6;;480;480&color=none&label=;;;;;;24;24;18;;&
3B	https://huygens.science.uva.nl:/SuperPlotsOfData/?data=1;;Treatment;Speed;Replicate&vis=quasirandom;;0.7;mean;solid;;;1;none&layout=Horizontal;;;TRUE;;0,60;;6;;480;480&color=none&label=;;;;;;24;24;18;;&
3C	https://huygens.science.uva.nl:/SuperPlotsOfData/?data=1;;Treatment;Speed;Replicate&vis=random;TRUE;0.7;mean;solid;;;1;none&layout=No;TRUE;TRUE;TRUE;;0,60;;6;;480;480&color=none&label=;;;;;;24;24;18;;&

Launching the webtool by using the hyperlink will reproduce the corresponding figure (make sure to copy-paste the entire URL. Clicking the hyperlink in the pdf may result in a broken link).

TABLE 3: Hyperlinks that define the plots in Figures 1 and 3.

visualize the data. Second, it is hard to implement an appropriate and fail-safe statistical analysis, since it is unpredictable what data will be supplied and analyzed (in terms of experimental design, but also the number of technical and biological replicates and the data distribution). By not implementing different analyses, a mindless, click-a-button, statistical analysis or "shopping" for the test that provides significance will be prevented. The statistical significance test should be carefully chosen to match the experimental design and data. For an overview of statistical tests and a decision tree to select the correct statistical test, the reader is referred to Pollard *et al.* (Pollard *et al.*, 2019). Finally, I believe that estimation statistics, that is, quantifying the actual difference between conditions, should be promoted and therefore an emphasis on significance testing is not desirable.

Experimental data that comprise both technical and biological replicates have a nested design. Each type of replicate contributes in a different way to the overall variance. A statistically correct comparison for these types of data uses a multilevel model (Galbraith *et al.*, 2010; Aarts *et al.*, 2014). However, this is a relatively complicated analysis. As a practical and intuitive alternative, the average of each technical replicate can be used as input for a standard (paired) *t* test. This approach is statistically valid (Galbraith *et al.*, 2010; Aarts *et al.*, 2014), but it is recommended to keep the number of measurements for technical each replicate similar (Lazic, 2010). Although averaging the technical replicates is not the best approach (Galbraith *et al.*, 2010), it is preferred over aggregating all technical replicates and ignoring the existence of biological replicates. For a detailed discussion about the superplot, the reader is referred to the original paper that proposed superplots (Lord *et al.*, 2020).

The web app enables the simultaneous visualization of multiple conditions. However, when multiple conditions are presented, their comparison may require a different test for statistical significance. For instance, testing multiple conditions may require correction for multiple testing to reduce the number of false positives. Since the primary goal of the web app is to visualize data, these analyses are currently not implemented. Even when more sophisticated statistics are used to calculate *p* values (e.g., hierarchical or nested tests), one can still use SuperPlotsOfData graphs to convey the distribution of the data.

DISCUSSION

The SuperPlotsOfData app implements some of the recent innovations in data visualization and analysis. I hope that the webtool will encourage users to adopt best practices in data presentation and analysis. These best practices include the display of individual observations, distinguishing between technical and biological replicates, and the use of estimation statistics for the quantitative comparison of conditions.

The tool democratizes modern data visualization as 1) the app is freely accessible online or it can run locally with free software, 2) it does not require any coding skills, and 3) it has a minimal learning curve. Finally, the code of the app is available which makes the analysis procedure transparent and open to modification to accommodate any future developments in analysis and visualization.

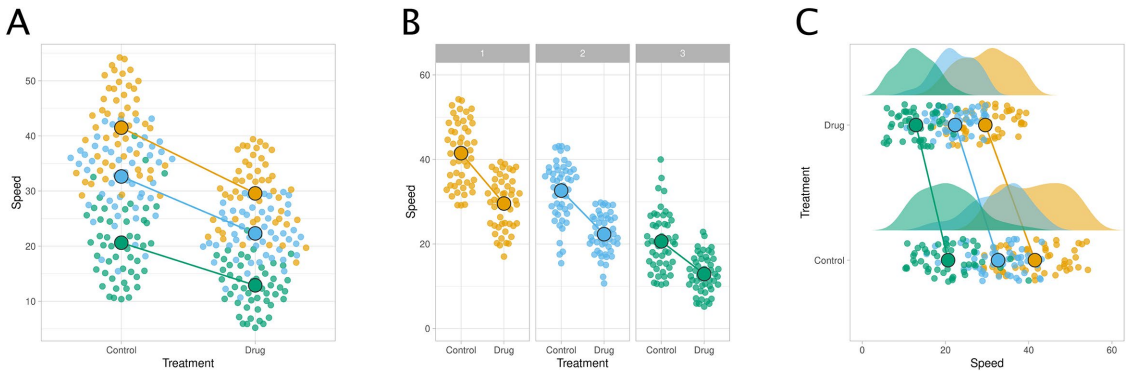

FIGURE 3: An example of the flexibility in plotting the data that is offered by SuperPlotsOfData. All plots are based on the example data and the biological replicates are paired, as indicated with the solid line. (A) Classic presentation, (B) separate presentation of each replicate, and (C) rotated plot with the data distributions on top, also known as a rain cloud plot. The URLs to recreate these figures are listed in Table 3.

ACKNOWLEDGMENTS

A good chunk of the code for SuperPlotsOfData is taken from a number of previously developed apps (https://github.com/JoachimGoedhart), PlotsOfData, VolcaNoseR, and PlotsOfDifferences, which in turn benefit from code that is shared on repositories (Github), blogs, and fora (stackoverflow). I thank Auke Folkerts (University of Amsterdam, The Netherlands) for help with the server that runs shiny and Max Grönloh (Sanquin Research, The Netherlands) for testing the app. The enthusiasm and support from colleagues on Twitter are highly appreciated and the interaction with Samuel J. Lord was particularly helpful for the implementation of the superplot.

REFERENCES

Aarts E, Verhage M, Veenvliet JV, Dolan CV, van der Sluis S (2014). A solution to dependency: using multilevel analysis to accommodate nested data. Nat Neurosci 17, 491–496. doi:10.1038/nn.3648.

Allen M, Poggiali D, Whitaker K, Marshall TR, Kievit R (2018). Raincloud plots: a multi-platform tool for robust data visualization. PeerJ Preprints 6. e27137v1.

Claridge-Chang A, Assam PN (2016). Estimation statistics should replace significance testing. Nat Methods 13, 108–109.

Cumming G (2009). Inference by eye: Reading the overlap of independent confidence intervals. Stat Med 28, 205–220. doi:10.1002/sim.3471.

Cumming G (2014). The new statistics: Why and how. Psychol Sci 25, 7–29. doi:10.1177/0956797613504966.

Cumming G, Finch S (2005). Inference by eye: confidence intervals and how to read pictures of data. Am Psychol 60, 170–180. doi:10.1037/0003-066X.60.2.170.

Drummond GB, Vowler SL (2011). Show the data, don't conceal them. J Physiol 589, 1861–1863. doi:10.1113/jphysiol.2011.205062.

Galbraith S, Daniel JA, Vissel B (2010). A Study of Clustered Data and Approaches to Its Analysis. J Neurosci 30, 10601 LP–10608. doi:10.1523/JNEUROSCI.0362-10.2010.

Goedhart J, Luijsterburg MS (2020). VolcaNoseR – a web app for creating, exploring, labeling and sharing volcano plots. bioRxiv. 2020.05.07.082263. doi:10.1101/2020.05.07.082263.

Ho J, Tumkaya T, Aryal S, Choi H, Claridge-Chang A (2019). Moving beyond P values: data analysis with estimation graphics. Nat Methods 16, 565–566. doi:10.1038/s41592-019-0470-3.

Lazic SE (2010). The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? BMC Neurosci 11, 5. doi:10.1186/1471-2202-11-5.

Lazic SE, Clarke-Williams CJ, Munafò MR (2018). What exactly is 'N' in cell culture and animal experiments? PLOS Biol 16, e2005282.

Lord SJ, Velle KB, Mullins RD, Fritz-Laylin LK (2020). SuperPlots: Communicating reproducibility and variability in cell biology. J Cell Biol 219. doi:10.1083/jcb.202001064.

Mauri M, Elli T, Caviglia G, Uboldi G, Azzi M (2017). RAWGraphs. *In* Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter - CHItaly '17, New York: Association for Computing Machinery, 1–5.

Naegle K, Gough NR, Yaffe MB (2015). Criteria for biological reproducibility: What does " n " mean? Sci Signal 8, fs7–fs7. doi:10.1126/scisignal.aab1125.

Pollard DA, Pollard TD, Pollard KS (2019). Empowering statistical methods for cellular and molecular biologists. Mol Biol Cell 30, 1359–1368. doi:10.1091/mbc.E15-02-0076.

Postma M, Goedhart J (2019). PlotsOfData—A web app for visualizing data together with their summaries. PLOS Biol 17, e3000202. doi:10.1371/journal.pbio.3000202.

Spitzer M, Wildenhain J, Rappsilber J, Tyers M (2014). BoxPlotR: a web tool for generation of box plots. Nat Methods 11, 121–122. doi:10.1038/nmeth.2811.

Weissgerber TL, Savic M, Winham SJ, Stanisavljevic D, Garovic VD, Milic NM (2017). Data visualization, bar naked: A free tool for creating interactive graphics. J Biol Chem 292, 20592–20598. doi:10.1074/jbc.RA117.000147.

Weissgerber TL, Winham SJ, Heinzen EP, Milin-Lazovic JS, Garcia-Valencia O, Bukumiric Z, Savic MD, Garovic VD, Milic NM (2019). Reveal, don't conceal: transforming data visualization to improve transparency. Circulation 140, 1506–1518. doi:10.1161/CIRCULATIONAHA.118.037777.

Wickham H (2011). ggplot2. Wiley Interdiscip Rev Comput Stat 3, 180–185. doi:10.1002/wics.147.

Wickham H (2014). Tidy Data. J Stat Softw 59, 1–23. doi:10.18637/jss.v059.i10.

Wilcox RR, Rousselet GA (2018). A Guide to robust statistical methods in neuroscience. In: Current Protocols in Neuroscience, Hoboken, NJ: Wiley, 8.42.1-8.42.30.