

1 **Studentsourcing - aggregating and re-using data from a**  
2 **practical cell biology course**

3 **Joachim Goedhart<sup>1</sup>**

4 <sup>1</sup>Molecular Cytology, SILS - University of Amsterdam, The Netherlands,

---

Corresponding author: Joachim Goedhart, [j.godhart@uva.nl](mailto:j.godhart@uva.nl)

## Abstract

Data that is generated by students during a course is often lost as there is no centrally organized collection and storage of the data. The loss of data prevents its re-use. To provide access to these data, I present an approach that I call studentsourcing. It collects, aggregates and re-uses data that is generated by students in a practical course on cell biology. The course runs annually and I have recorded the data that was generated by >100 students over 3 years. Two use cases illustrate how the data can be aggregated and re-used either for the scientific record or for teaching. As the data is obtained by different students, in different groups, over different years, it is an excellent opportunity to discuss experimental design and modern data visualization methods such as the superplot. The first use case demonstrates how the data can be presented as an online, interactive dashboard, providing real-time data of the measurements. The second use case shows how central data storage provides a unique opportunity to get precise quantitative data due to the large sample size. Both use cases illustrate how data can be effectively aggregated and re-used.

## Plain Language Summary

Data acquired by students has value and in this work we present ways to collect and re-use that data.

## Introduction

Teaching practical skills in a lab course is a crucial part of education in biology, biomedical science, and life sciences (Hofstein & Lunetta, 2003; Reid & Shah, 2007). In these lab courses data is generated, reported and interpreted, much like *real* experimental lab work. However, students use their data just for their own lab report and the data is not centrally stored or aggregated. As a consequence, most of the data that is gathered in a lab course is lost. Yet, these data are potentially useful. Especially for larger course, an impressive amount of data under well-controlled conditions can be generated. Therefore, by collecting and aggregating the data of multiple students over multiple years, one can easily gather a large dataset with high numbers of independent observations (Lazic et al., 2018).

Microscopy is an essential tool in cell biology. The use of microscopes to observe cells and organisms has changed from a qualitative, descriptive approach, into a quantitative method (Renz, 2013; Senft et al., 2023; Wait et al., 2020; Waters, 2009). The development of digital cameras and image analysis software has catalyzed this transition (Carpenter, 2007). Therefore, experiments that use microscopes are often followed by bioimage analysis to extract quantitative information from the data. To teach these skills, we combine a basic course on microscopy in a course on cell biology with teaching image processing and analysis in ImageJ/FIJI (Schneider et al., 2012). In a typical year, over one hundred students are enrolled in this course and therefore, a substantial amount of data is generated in the course.

I decided to collect the data that was generated by the students in the lab course over several years and store the measurement results in a central location. The data by itself can be valuable for the scientific community as precise estimates with good statistics can be obtained. Moreover, the data are a starting point to discuss data visualization, experimental design and how experimental design affects the statistics and interpretation of data. Here, I report the methods to collect, process and visualize the data. The data re-use is demonstrated in two use cases.

## Methods

For full reproducibility, this document is written using Quarto (Posit, <https://quarto.org/>), and the source code of the manuscript and the notebooks, and the data are available in a repository: <https://github.com/JoachimGoedhart/MS-StudentSourcing>. A version rendered as HTML is available and it provides

56 easy access to the notebooks as well: [https://joachimgoedhart.github.io/](https://joachimgoedhart.github.io/MS-StudentSourcing/)  
57 [MS-StudentSourcing/](https://joachimgoedhart.github.io/MS-StudentSourcing/)

## 58 Use case 1

### 59 *Sample preparation and measurements*

60 A buccal swab is used to harvest cheek cells by scraping ~5 times over the inside of  
61 the cheek. The tip of the sample collector is dipped into an eppendorf tube with 40  
62  $\mu\text{l}$  PBS, and the cells are transferred to an object slide by touching the slide with the  
63 tip. Next, 10  $\mu\text{l}$  of 0.1% methyleneblue solution is added and the sample is enclosed  
64 by a square coverslip (22 x 22 mm). The sample is used immediately to observe  
65 the cells with a Leica microscope, equipped with a Lumenera camera. A 20x or 40x  
66 objective is used to observe and image the cells. A separate image of a micrometer  
67 (Electron Microscopy Sciences 6804208, Stage Micrometer S8, Horizontal Scale, 1  
68 mm Length) is acquired at the same magnification. The images are processed in  
69 ImageJ/FIJI and the dimensions of the images are calibrated with the micrometer  
70 image (Figure 1). The line tool is used to measure the diameter of the cells (the  
71 longest axis).

### 72 *Data collection*

73 The data of the measurements is collected through a Google Form, an example of  
74 which is shown in Figure 1. By submitting the form, the students give permission for  
75 the anonymous use of the data. The data that is recorded by the form is the group  
76 (A/B/C/D), the size measurements of the cheek cells and the size measurements of  
77 the nucleus. The data is aggregated in a Google Sheet which has four columns with  
78 data on Timestamp, Group, size of cells, size of nuclei. When correctly uploaded,  
79 the two columns with the size data have comma separated values of 10 measure-  
80 ments.

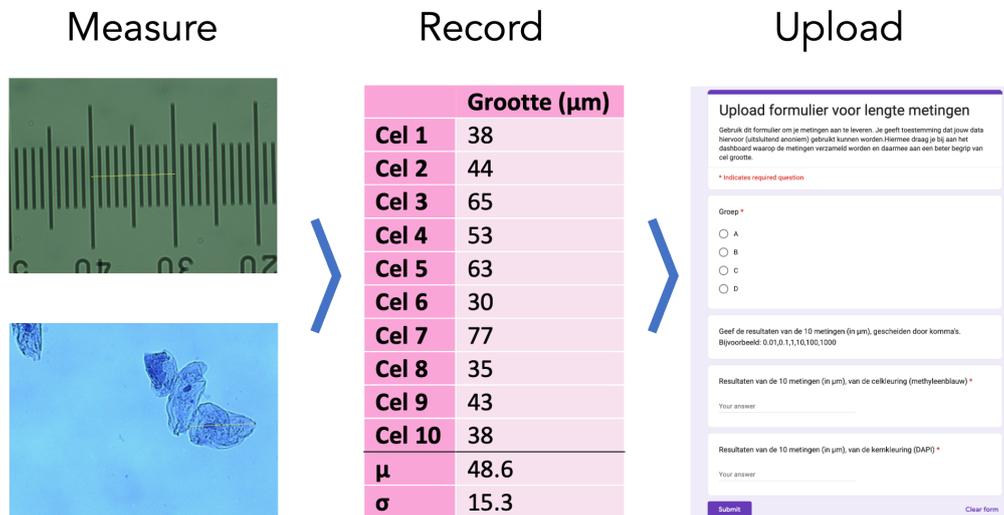


Figure 1: Overview of steps taken by the students to acquire and upload their data. The size of cells is measured by microscopy, the data is processed and recorded in their lab journal and finally, the data are uploaded by a Google form

### 81 *Data processing*

The data that is in the Google Sheet can be downloaded and read into R as a CSV. All subsequent processing and data visualisation and presentation in dashboard style is done in R. The code is available on Github: <https://github.com/JoachimGoedhart/CellSizeR>. The cleaning of the data consists of removing empty cells, changing the column names, listing all individual measurements in a single row, forcing the data into a ‘numeric’ type and filtering for sensible values (anything outside the generous range of 0-1000 will be removed). A detailed protocol that explains the processing is available as protocol 10 (Goedhart, 2022).

#### *Data visualisation*

A dashboard is composed in R Markdown with the {flexdashboard} package. The code is available here: <https://github.com/JoachimGoedhart/CellSizeR> and the live dashboard is available online: <https://amsterdamstudygroup.shinyapps.io/CellSizeR/>

### **Use case 2**

#### *Sample preparation and measurements*

HeLa cells are cultured according to standard procedures and seeded 1 or 2 days before the treatment on 12 mm diameter glass coverslips. HeLa cells are incubated with 10  $\mu\text{M}$  EdU for 30 minutes at 37 °C. The cells are fixed with 4% formaldehyde in PBS and permeabilised with 0.1% Triton X-100 in PBS. Click chemistry is performed with 9  $\mu\text{M}$  Cy3-azide and 2 mM  $\text{CuSO}_4$ . To start the reaction, 20 mg/ml ascorbate (final concentration) is added and the solution is used immediately to stain the cells. After 30 minutes, the cells are washed 3x with PBS and the sample is incubated with 0.1  $\mu\text{g/ml}$  DAPI for 5 minutes. Samples are mounted in Mowiol and used for observation with fluorescence microscopy. Images of at least 100 cells are acquired with the DAPI and TRITC filters sets. The nuclei in both channels are counted by hand, or in an automated way by segmentation and ‘particle analysis’ in imageJ to calculate the percentage of cells that are positive for Cy3 fluorescence, reflecting cells in the S-phase.

#### *Data collection*

The data of the measurements is collected through a Google Form. By submitting the form, the students give permission for the anonymous use of the data. The data that is recorded is the group (A/B/C/D), the percentage of cells in the S-phase for two methods, i.e. manual and using ImageJ/FIJI. The form is easy to set up and the data is collected in Google Sheets, yielding four columns; Timestamp, Group, and two columns with percentages of S-phase determined by the two methods.

#### *Data processing & visualization*

The data that is in the Google Sheet can be downloaded and read into R (R Core Team, 2022) as a CSV. All subsequent processing and data visualization is done with R and quarto. The cleaning of the data consists of removing empty cells, changing the column names, conversion to a tidy format, forcing the data into a ‘numeric’ type and filtering for sensible values (anything outside the generous range of 0-100 will be removed).

## **Results**

### **Use case 1: Comparing new results with historical data**

The aim of the experiment is to determine the average size (diameter) of a human cheek cell and nucleus. To this end, the students acquire images of their own, stained cheek cells and measure the size of the cell and its nucleus. At least 10 measurements are made and the data are uploaded with a Google form. Each sample is an independent observation as it originates from a unique human specimen. To evaluate the accuracy of their own measurements, the students can compare their

132 data with the historical data that is displayed on an online, interactive dashboard:  
 133 <https://amsterdamstudygroup.shinyapps.io/CellSizeR/>. A snapshot of the  
 134 dashboard is shown in Figure 2.

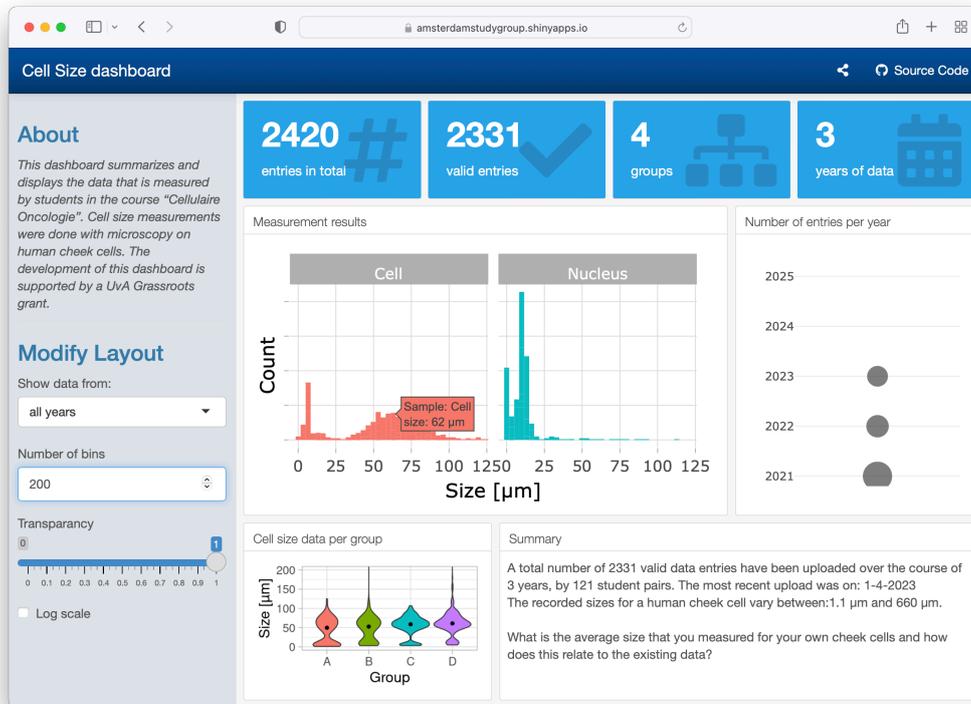


Figure 2: Screenshot of the dashboard that summarized the data on human cheek cell size measurements. The dashboard also displays the total number of entries and other metadata.

135 The dashboard is interactive and users can select the data from all measurements,  
 136 or from a single year and the number of bins can be adjusted. Additionally, by hov-  
 137 ering over the plots, the values of the data can be read (as shown in Figure 2). The  
 138 dashboard also shows the data for the 4 different groups and the size distribution of  
 139 the cells by violin plots.

140 The histogram on the dashboard is the primary data that is useful for the students.  
 141 It visualizes the distribution of individual data for both the cell and the nucleus.  
 142 Since the sizes vary substantially, the data can be shown on a log-scale as well  
 143 on the dashboard (Figure 3). The main reason that the measurements differ by an  
 144 order of magnitude, is that the size measurement requires correct calibration of the  
 145 field of view with a micro ruler. When the calibration is done incorrectly, this will  
 146 affect the accuracy of the measurement, usually by a factor of 10.

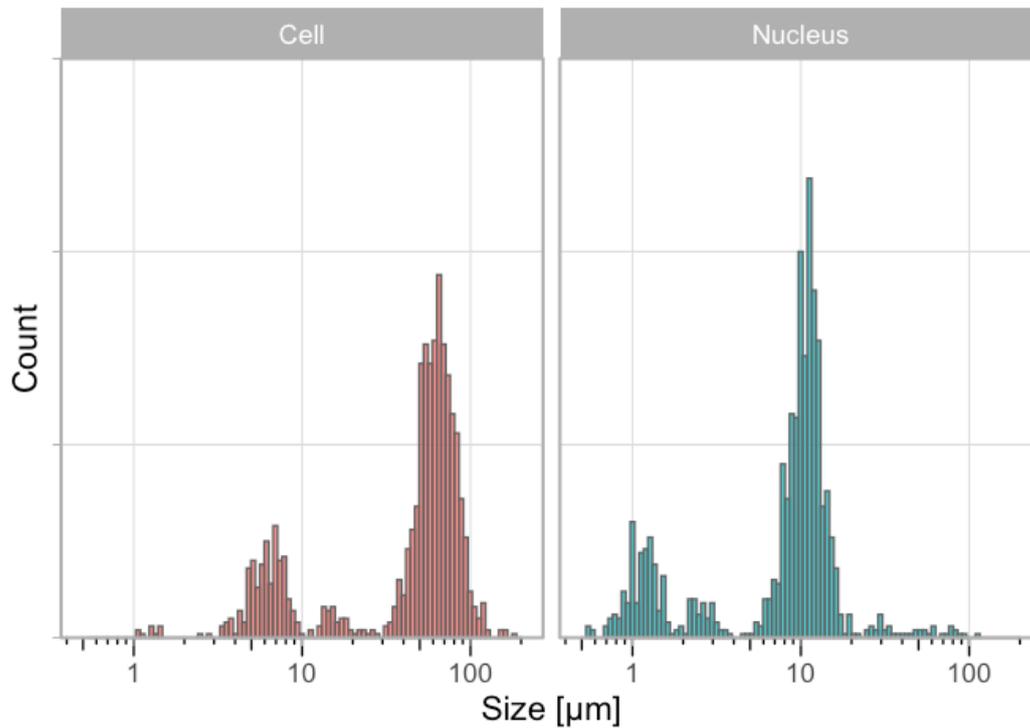


Figure 3: Distribution of the measured size of human cheek cells and their nucleus on a log scale. Data from three years. Source: [Summarizing the size of cells](#)

147 Ideally, the students see the multimodal distribution and realize that the peaks are  
 148 an order of magnitude different. Even if they don't, they will probably assume that  
 149 the majority of measurements is correct. In any case, it is possible for the students  
 150 to make a comparison and discuss their results in the context of the historical data.

#### 151 **Use case 2: Determination of the percentage of cells in S-phase**

152 The aim of the experiment is to determine the number of cells, as percentage,  
 153 that is in the S-phase. To this end, students stain cells that are treated with EdU  
 154 and they use these samples to quantify the percentage of cells in the S-phase in two  
 155 ways (manual and semi-automated). The results are uploaded via a Google Form.  
 156 The collected data can be analysed in multiple ways and here we used it to compare  
 157 the two analysis methods and, secondly, to obtain an estimate for the percentage  
 158 of S-phase cells. The data on the two analysis methods, manual and automated,  
 159 is paired and can be visualised by a dotplot in which the pairs of the data are  
 160 connected (Figure 4). The slopes of the lines vary a lot, whereas the average values per  
 161 year between the two methods is similar. This implies that there can be substan-  
 162 tial differences between the two methods, with roughly a similar number of cases  
 163 where the automated analysis over- or underestimates the percentage, relative to the  
 164 manual analysis.

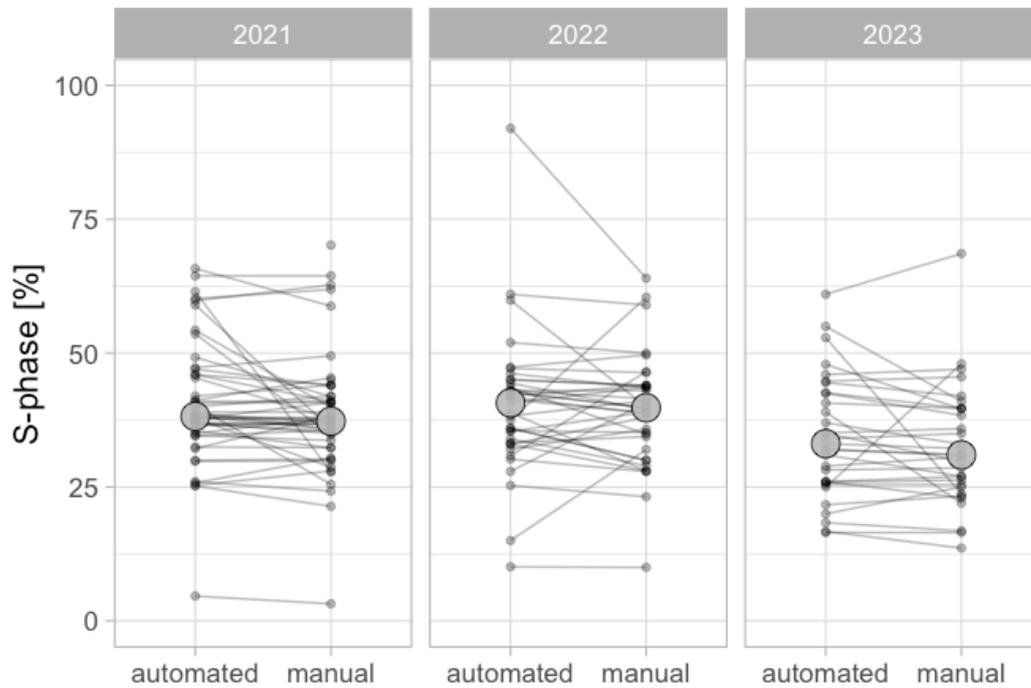


Figure 4: Comparing the manual and automated analysis. Source: [The percentage of cells in the S-phase](#)

165 There is increasing attention on effects of experimental design on data analysis and  
 166 visualization. The recently proposed superplot to distinguish biological and technical  
 167 replicates is an intuitive and straightforward way to communicate the design (Lord  
 168 et al., 2020). The data on S-phase consists of both technical and biological replicates  
 169 and is therefore ideally suited to explain the importance of correctly identifying the  
 170 independent measurements. Here, we treat the data from each group as biological  
 171 replicate, and the measurements within each group as a technical replicate. The  
 172 reason is that a group of students all stain cells that are from the same passage  
 173 number and treated at the same time and is therefore a technical replicate. On the  
 174 other hand, different groups stain different passages of cells and so we treat these  
 175 as independent observations. When the data is plotted for each individual technical  
 176 replicate (Figure 5), it can be observed that we received multiple submissions per  
 177 group, leading to a precise measurement per group. The median values range from  
 178 23% to 44%. The average value of the independent observations is 36.7% [N=12,  
 179 95%CI: 33.0%-40.3%].

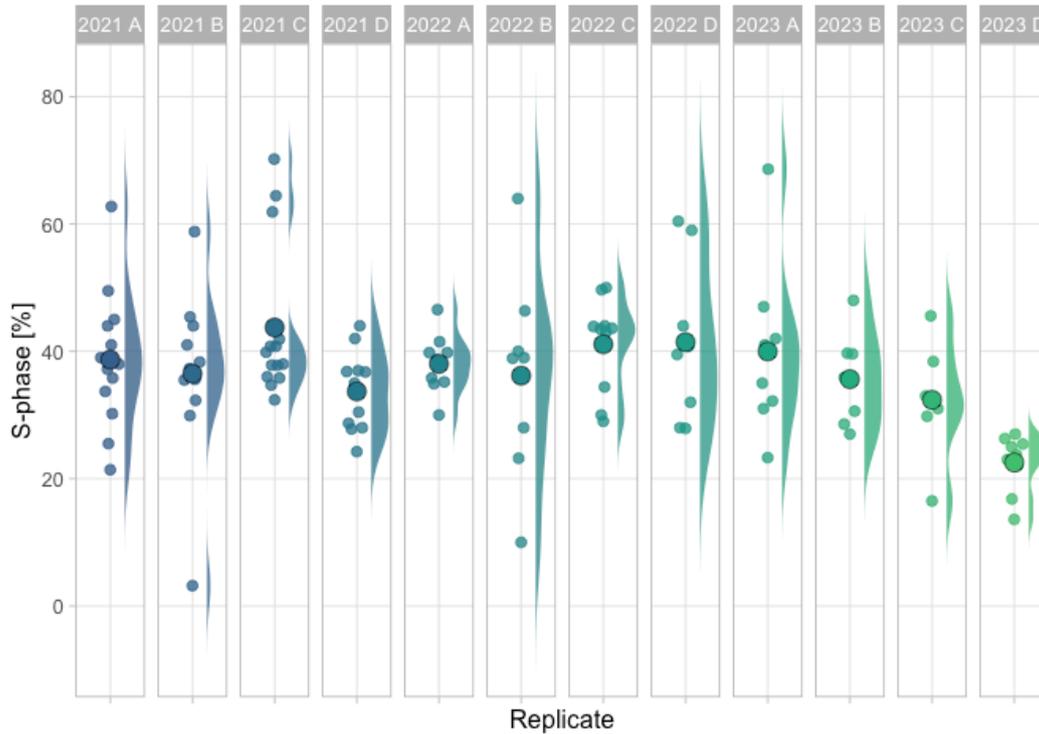


Figure 5: Data on the percentage of cells in e S-phase based on manual analysis. Each group & year defines an independent observations and is shown as dotplot and the distribution. The larger dot reflects the median value. Source: [The percentage of cells in the S-phase](#)

## 180 Discussion

181 Data that is generated in courses is often recorded by individual students or groups  
 182 of students in reports. However, it can be valuable and interesting to collect and use  
 183 these data. Here, I present a flexible and straightforward approach to collect and  
 184 display data from a large group of students and over several years. A combination of  
 185 Google Forms/Sheets for data collection and R for data processing and visualization  
 186 is used.

187 In the first use case, the data is displayed in dashboard style. The dashboard gives  
 188 a quick, interactive and complete overview of the data and it is used by students to  
 189 evaluate their results. The aggregation of the data and presentation on the dash-  
 190 board is a nice reward for the students as they see that their data has value and is  
 191 re-used by their peers.

192 In the second use case, a quarto template is used to process and visualize the data.  
 193 The use case shows how two methods can be compared and also shows that a high  
 194 number of truly independent observations can be collected.

195 For both use cases, the code is available and can be used as a starting point for the  
 196 processing and visualization of other datasets. The approach that is presented here  
 197 is not limited to practical courses. It can also be used to collect data from other  
 198 crowds, or in collaborative science projects.

199 Collecting and reusing the data has a number of advantageous aspects. First, a high  
 200 number of measurements increases the precision of the measurement and therefore  
 201 allows us to obtain precise numbers. Second, the historical data can be shared with

202 the students and they can interpret and discuss their results in light of the existing  
 203 data. Third, the obtained data serves as material that can be used to teach data ma-  
 204 nipulation, statistics and data visualization which is a fundamental aspect of science  
 205 (Sailem et al., 2016). The use cases described in this paper deal with these aspects.

206 The aggregation of the data inevitably leads to a discussion on experimental design,  
 207 as this is important to establish whether measurements are independent or not. This  
 208 aspect of experimental design has received attention over the last years (Aarts et al.,  
 209 2015; Eisner, 2021; Sikkel et al., 2017) and it is valuable to teach this aspect of data  
 210 analysis and visualization. Although I have not implemented this yet, I think that  
 211 having students participate in the data aggregation, creates a very practical opportu-  
 212 nity to teach experimental design and the identification of biological units (Lazic et  
 213 al., 2018). In addition, it may stimulate cooperative learning (Tanner et al., 2003).

214 In conclusion, I feel it is valuable to collect data from practical courses and here we  
 215 report one way to achieve that. I hope that serves as a starting point for others that  
 216 want to collect, store and use data from large groups of students.

### 217 **Data availability**

218 The data is available at: <https://doi.org/10.5281/zenodo.8359955>

### 219 **Code availability**

220 The code for this manuscript is available here: <https://github.com/JoachimGoedhart/MS-StudentSourcing>  
 221 and it includes the notebook that was generated to analyze  
 222 the S-phase data: [https://github.com/JoachimGoedhart/MS-StudentSourcing/](https://github.com/JoachimGoedhart/MS-StudentSourcing/tree/main/notebooks)  
 223 [tree/main/notebooks](https://github.com/JoachimGoedhart/MS-StudentSourcing/tree/main/notebooks) The code for the dashboard ‘CellSizeR’ is deposited here:  
 224 <https://github.com/JoachimGoedhart/CellSizer>

225 Versioned code with a DOI will be made available upon acceptance.

### 226 **Contributions**

227 J.G. conceived the project, acquired funding, wrote code, and wrote the manuscript.

### 228 **Competing interests**

229 The authors declare no competing interests

### 230 **Acknowledgments**

231 A [blog post by Garrick Aden-Buie](#) was very helpful in the initial phase of this  
 232 project. Many thanks to the people involved in Quarto, which was used to write  
 233 and shape this paper. Most importantly, I’d like to thank all students involved in  
 234 the course that have generously shared their data, making this project a success.

### 235 **References**

- 236 Aarts, E., Dolan, C. V., Verhage, M., & Sluis, S. van der. (2015). Multilevel analysis  
 237 quantifies variation in the experimental effect while optimizing power and pre-  
 238 venting false positives. *BMC Neuroscience*, *16*(1). [https://doi.org/10.1186/](https://doi.org/10.1186/s12868-015-0228-5)  
 239 [s12868-015-0228-5](https://doi.org/10.1186/s12868-015-0228-5)
- 240 Carpenter, A. E. (2007). Software opens the door to quantitative imaging. *Nature*  
 241 *Methods*, *4*(2), 120–121. <https://doi.org/10.1038/nmeth0207-120>
- 242 Eisner, D. A. (2021). Pseudoreplication in physiology: More means less. *Journal of*  
 243 *General Physiology*, *153*(2). <https://doi.org/10.1085/jgp.202012826>
- 244 Goedhart, J. (2022). *DataViz protocols - an introduction to data visualization proto-*  
 245 *cols for wet lab scientists*. Zenodo. <https://doi.org/10.5281/ZENODO.7257808>
- 246 Hofstein, A., & Lunetta, V. N. (2003). The laboratory in science education: Foun-  
 247 dations for the twenty-first century. *Science Education*, *88*(1), 28–54. [https://](https://doi.org/10.1002/sce.10106)  
 248 [doi.org/10.1002/sce.10106](https://doi.org/10.1002/sce.10106)

- 249 Lazic, S. E., Clarke-Williams, C. J., & Munafò, M. R. (2018). What exactly is  
250 ‘N’ in cell culture and animal experiments? *PLOS Biology*, *16*(4), e2005282.  
251 <https://doi.org/10.1371/journal.pbio.2005282>
- 252 Lord, S. J., Velle, K. B., Mullins, R. D., & Fritz-Laylin, L. K. (2020). SuperPlots:  
253 Communicating reproducibility and variability in cell biology. *Journal of Cell*  
254 *Biology*, *219*(6). <https://doi.org/10.1083/jcb.202001064>
- 255 R Core Team. (2022). *R: A language and environment for statistical computing*.  
256 Vienna, Austria: R Foundation for Statistical Computing. Retrieved from  
257 <https://www.R-project.org/>
- 258 Reid, N., & Shah, I. (2007). The role of laboratory work in university chemistry.  
259 *Chem. Educ. Res. Pract.*, *8*(2), 172–185. <https://doi.org/10.1039/b5rp90026c>
- 260 Renz, M. (2013). Fluorescence microscopy—A historical and technical perspective.  
261 *Cytometry Part A*, *83*(9), 767–779. <https://doi.org/10.1002/cyto.a.22295>
- 262 Sailem, H. Z., Cooper, S., & Bakal, C. (2016). Visualizing quantitative microscopy  
263 data: History and challenges. *Critical Reviews in Biochemistry and Molecular*  
264 *Biology*, *51*(2), 96–101. <https://doi.org/10.3109/10409238.2016.1146222>
- 265 Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). NIH Image to ImageJ:  
266 25 years of image analysis. *Nature Methods*, *9*(7), 671–675. [https://doi.org/](https://doi.org/10.1038/nmeth.2089)  
267 [10.1038/nmeth.2089](https://doi.org/10.1038/nmeth.2089)
- 268 Senft, R. A., Diaz-Rohrer, B., Colarusso, P., Swift, L., Jamali, N., Jambor, H., et  
269 al. (2023). A biologist’s guide to planning and performing quantitative bioimag-  
270 ing experiments. *PLOS Biology*, *21*(6), e3002167. [https://doi.org/10.1371/](https://doi.org/10.1371/journal.pbio.3002167)  
271 [journal.pbio.3002167](https://doi.org/10.1371/journal.pbio.3002167)
- 272 Sikkel, M. B., Francis, D. P., Howard, J., Gordon, F., Rowlands, C., Peters, N. S.,  
273 et al. (2017). Hierarchical statistical techniques are necessary to draw reliable  
274 conclusions from analysis of isolated cardiomyocyte studies. *Cardiovascular*  
275 *Research*, *113*(14), 1743–1752. <https://doi.org/10.1093/cvr/cvx151>
- 276 Tanner, K., Chatman, L. S., & Allen, D. (2003). Approaches to Cell Biology Teach-  
277 ing: Cooperative Learning in the Science Classroom—Beyond Students Working  
278 in Groups. *Cell Biology Education*, *2*(1), 1–5. [https://doi.org/10.1187/](https://doi.org/10.1187/cbe.03-03-0010)  
279 [cbe.03-03-0010](https://doi.org/10.1187/cbe.03-03-0010)
- 280 Wait, E. C., Reiche, M. A., & Chew, T.-L. (2020). Hypothesis-driven quantitative  
281 fluorescence microscopy – the importance of reverse-thinking in experimental de-  
282 sign. *Journal of Cell Science*, *133*(21). <https://doi.org/10.1242/jcs.250027>
- 283 Waters, J. C. (2009). Accuracy and precision in quantitative fluorescence mi-  
284 croscopy. *Journal of Cell Biology*, *185*(7), 1135–1148. [https://doi.org/](https://doi.org/10.1083/jcb.200903097)  
285 [10.1083/jcb.200903097](https://doi.org/10.1083/jcb.200903097)