

1 **Reproducible, flexible and high throughput data extraction from primary**  
2 **literature: The metaDigitise R package**

3 Joel L. Pick<sup>1,\*</sup>, Shinichi Nakagawa<sup>1</sup>, Daniel W.A. Noble<sup>1</sup>

4 <sup>1</sup> Ecology and Evolution Research Centre, School of Biological, Earth and  
5 Environmental Sciences, University of New South Wales, Kensington, NSW 2052,  
6 Sydney, AUSTRALIA

7 \*Corresponding Author: joel.l.pick@gmail.com

## 8 Abstract

9 1. Research synthesis requires data extraction from primary studies with effect sizes for  
10 meta-analyses being calculated from summary statistics. However, exact values of such  
11 statistics are commonly hidden in figures.

12 2. Extracting summary statistics from figures can be a slow process that is not easily  
13 reproducible. Additionally, current software lacks an ability to incorporate important  
14 meta-data (e.g., sample sizes, treatment / variable names) about experiments and is not  
15 integrated with other software to streamline analysis pipelines.

16 3. Here we present the R package **metaDigitise** which extracts descriptive statistics  
17 such as means, standard deviations and correlations from the four plot types: 1)  
18 mean/error plots (e.g. bar graphs with standard errors), 2) box plots, 3) scatter plots  
19 and 4) histograms. **metaDigitise** is user-friendly and easy to learn as it interactively  
20 guides the user through the data extraction process. Notably, it enables large-scale  
21 extraction by automatically loading image files, letting the user stop processing, edit  
22 and add to the resulting data frame at any point.

23 4. Digitised data can be easily re-plotted and checked, facilitating reproducible data  
24 extraction from plots with little inter-observer bias. We hope that by making the  
25 process of figure extraction more flexible and easy to conduct it will improve the  
26 transparency and quality of meta-analyses in the future.

27 Keywords: meta-analysis, comparative analysis, data extraction, R, reproducibility,  
28 figures, images, summary statistics

# 1 Introduction

In many different contexts, researchers make use of data presented in primary literature. Most notably, this includes meta-analysis, which is becoming increasingly common in many research fields. Meta-analysis uses effect size estimates and their sampling variance, taken from many studies, to understand whether particular effects are common across studies and to explain variation among these effects (Glass, 1976; Koricheva, Gurevitch & Mengersen, 2013; Nakagawa et al., 2017). Meta-analysis relies on descriptive statistics (e.g. means, standard deviations (SD), sample sizes, correlation coefficients) extracted from primary literature that have been reported in the text or tables of research papers. Descriptive statistics are also, however, frequently presented in figures and so need to be manually extracted using digitising programs. While inferential statistics (e.g.,  $t$ - and  $F$ -statistics) are often presented along side descriptive statistics, and can be used to derive effect sizes, descriptive statistics are much more appropriate to use because sources of non-independence in experimental designs can be dealt with more easily (Noble et al., 2017).

Although there are several tools that data extraction from figures (e.g. **DataThief** (Tummers, 2006), **GraphClick** (Arizona-Software, 2008), **WebPlotDigitizer** (Rohatgi, 2017)), these tools do not cater to needs of meta-analysis for three main reasons. First, they typically only provide the user with calibrated  $x,y$  coordinates from imported figures, and do not differentiate between common plot types that are used to present data. Consequently a large amount of downstream data manipulation is required, that is different across plots types. For example, data are frequently presented in mean/error plots (Figure 1A), from which the user wants a mean and SD for each group presented. From  $x,y$  coordinates, users must manually discern between mean and error coordinates and assign points to groups. Error then needs to be calculated as the deviation from the mean, and then transformed to SD, according to the type of error presented. Second, digitising programs do not allow the integration of metadata at the time of data

56 extraction, such as experimental group or variable names, and sample sizes. This makes  
57 the downstream calculations laborious, as information has to be added later using  
58 different software. Finally, existing programs do not import sets of images for the user  
59 to systematically work through. Instead they require the user to manually import  
60 images one by one, and export data into individual files, that need to be imported and  
61 edited using different software.

62 Data extraction from figures is therefore an incredibly time-consuming process as  
63 existing software does not provide an optimized research pipeline to facilitate data  
64 extraction and editing. Furthermore, although meta-analysis is an important tool in  
65 consolidating the data from multiple studies, many of the processes involved in data  
66 extraction are opaque and difficult to reproduce, making extending studies problematic.  
67 Having a tool that facilitates reproducibility in meta-analyses will increase transparency  
68 and aid in resolving the reproducibility crises seen in many fields (Peng, Dominici &  
69 Zeger, 2006; Peng, 2011; Parker et al., 2016).

70 Here, we present an interactive R package, **metaDigitise** (available at  
71 <https://github.com/daniel1noble/metaDigitise>), which is designed for large scale,  
72 reproducible data extraction from figures, specifically catering to the the needs of  
73 meta-analysts. To this end, we provide tools to extract data from common plot types  
74 (mean/error plots, box plots, scatter plots and histograms, see Figure 1). **metaDigitise**  
75 operates within the R environment making data extraction, analysis and export more  
76 streamlined. The necessary calculations are carried out on processed data immediately  
77 after extraction so that comparable summary statistics can be obtained quickly.  
78 Summary data from multiple figures is returned into a single data frame which can be  
79 can easily exported or use in downstream analysis within R. Completed digitisations are  
80 automatically saved for each figure, meaning users can redraw their digitisations on  
81 figures, make corrections and access calibration and proceeded data. This makes  
82 sharing figure digitisation and reproducing the work of others simple and easy, and

83 allows meta-analyses to be updated more efficiently.

## 84 2 Directory Structure and Reproducibility

85 The **metaDigitise** package was created with the idea that users would have multiple  
86 images to extract from and therefore operates in the same way whether the user has one  
87 or multiple images. There is one main function, **metaDigitise()**, which interactively  
88 takes the user through the process of extracting data from figures. **metaDigitise()**  
89 works on a directory containing images of figures copied from primary literature, in  
90 .png, .jpg, .tiff, .pdf format, specified to **metaDigitise()** through the **dir** argument.  
91 The user should think carefully about their directory structure early on in their project.  
92 Although different directory structures may be used, we would recommend having all  
93 files for one project in a single directory with an informative and unambiguous naming  
94 scheme for images to help identify the paper and figure that data come from (e.g.  
95 paper\_figure\_trait.png).

96 **metaDigitise()** recognizes all the images in a directory and automatically imports  
97 them one by one, allowing the user to extract the relevant information about a figure as  
98 they go. Having all figures in one directory therefore expedites digitisation by  
99 preventing users from having to constantly change directories and / or open new  
100 images. The data from each completed image is automatically saved as a **metaDigitise**  
101 object in a separate .RDS file to a **caldat** directory that is created within the parent  
102 directory when first executing **metaDigitise()**. These files enable re-plotting and  
103 editing of images at a later point (see below). When run, **metaDigitise()** also  
104 identifies the images within a directory that have been previously digitised and only  
105 imports new images to process. The data of all images is then automatically integrated  
106 into the final output. This means that all figures do not need to be extracted at one  
107 time and new figures can be added to the directory as the project develops.

108 This directory structure allows the complete digitisation process to be reproduced at a  
109 later stage, shared with collaborators and presented as supplementary materials for a  
110 publication. As long as all the images and the caldat directory are still in one directory,  
111 `metaDigitise()` will be able to reproduce all figure extractions, regardless of the  
112 computer it is run on. For an analysis to be updated, new figures can simply be added  
113 to the directory and `metaDigitise()` run to incorporate the new data.

### 114 3 Image Processing

115 Running `metaDigitise()` presents the user with three options; ‘Process new images’,  
116 ‘Import existing data’ or ‘Edit existing data’. Selecting ‘Process New Images’ starts the  
117 digitisation process on images within the directory that have not previously been  
118 digitised; the other functions are discussed below.

119 For all plot types, `metaDigitise()` requires the user to calibrate the axes in the figure,  
120 by clicking on two known points on the axis in question, and entering the value of those  
121 points (Figure 1). `metaDigitise()` then calculates the value of any clicked points in  
122 terms of the figure axes. For mean/error and box plots, only the y-axis is calibrated  
123 (Figure 1A,B), assuming the x-axis is redundant. For scatter plots and histograms both  
124 axes are calibrated (Figure 1C,D).

125 As figures may have been copied from older, scanned publications, they may not be  
126 perfectly orientated. This makes calibration of the points in the figure problematic.  
127 `metaDigitise()` allows users to rotate the image (Figure 2A,B). Furthermore,  
128 mean/error plots, box plots and histograms, may be presented with horizontal bars.  
129 `metaDigitise()` assumes that bars are vertical, but allows the user to flip the image to  
130 make the bars are vertical (Figure 2C,D).

131 **metaDigitise** recognises four main types of plot; Mean/error plots, box plots, scatter  
132 plots and histograms (1). All plot types can be extracted in a single call of

133 `metaDigitise()` and integrated into one output. Alternatively, users can process  
134 different plot types separately, using separate directories. All four plot types are  
135 extracted slightly differently (outlined below). Upon completing all images, or quitting,  
136 either summarised or processed data is returned (specified by the user through the  
137 `summary` argument). Summarised data consists of a mean, SD and sample size, for each  
138 identified group within the plot (should multiple groups exist). In the case of scatter  
139 plots, the correlation coefficient between x and y variables within each identified group  
140 is also returned. Processed data consists of a list with slots for each of the four figure  
141 types, containing the calibrated points that the user has clicked. This may be  
142 particularly useful in the case of scatter plots.

### 143 **3.1 Mean/Error and Box Plots**

144 `metaDigitise()` handles mean/error and box plots in a very similar way. For each  
145 mean/box, the user enters group names and sample sizes. If the user does not enter a  
146 sample size at the time of data extraction (if, for example, the information is not readily  
147 available) a SD is not calculated. Sample sizes can, however, be entered at a later time  
148 (see below). For mean/error plots, the user clicks on an error bar and the mean. Error  
149 bars above or below the mean can be clicked, as sometimes one is clearer than the  
150 other. `metaDigitise()` assumes that the error bars are symmetrical. Points are  
151 displayed where the user has clicked, with the error in a different colour to the mean  
152 (Figure 1A). The user also enters the type of error used in the figure: SD, standard  
153 error (SE) or 95% confidence intervals (CI95). For box plots, the user clicks on the  
154 maximum, upper quartile, median, lower quartile and minimum. For both plot types,  
155 the user can add, edit or remove groups. Three functions, `error_to_sd()`,  
156 `rqm_to_mean()` and `rqm_to_sd()`, that convert different error types to SD, box plot data  
157 to mean and box plot data SD, respectively, are also available in the package (see  
158 supplements for further details of these conversions).

## 159 3.2 Scatter plots

160 Users can extract points from multiple groups from scatter plots. Different groups are  
161 plotted in different colours and shapes to enable them to be distinguished, with a legend  
162 at the bottom of the figure (Figure 1C). Mean, SD and sample size are calculated from  
163 the clicked points, for each group. Data points may overlap with each other making it  
164 impossible to know whether points have been missed. This may result in the sample  
165 size of digitised groups conflicting with what is reported in the paper. For example, in  
166 Figure 1C only 49 points have been clicked when the sample size is known to be 50.  
167 Hence, **metaDigitise** also provides the user with the option to input known sample sizes  
168 directly. Nonetheless, it is important to recognise the impact that overlapping points  
169 can have on summary statistics, and in particular on sampling variance.

## 170 3.3 Histograms

171 The user clicks on the top corners of each bar, which are drawn in alternating colours  
172 (Figure 1D). Bars are numbered to allow the the user to edit them. As with scatter  
173 plots, if the sample size from the extracted data does not match a known sample size,  
174 the user can enter an alternate sample size. The calculation of mean, SD and sample  
175 size from this data is shown in the supplements.

## 176 4 Importing and Editing Previously Digitised 177 data

178 **metaDigitise** is also able to re-import, edit and re-plot previously digitised figures.  
179 When running **metaDigitise()**, the user can choose to ‘Import existing data’, which  
180 returns previously digitised data, from single or all figures. Alternately, the  
181 **import\_metaDigitise()** function returns the data of previous digitisations, but



182 without user interaction, allowing easier integration into larger scripts. ‘Edit existing  
183 data’ allows the user to re-plot or edit information or digitisations that have previously  
184 be done.

## 185 4.1 Adding Sample Sizes to Previous Digitisations

186 In many cases sample sizes may not be readily available when digitising figures. This  
187 information does not need to be added at the time of digitisation. To expedite finding  
188 and adding these sample sizes at a later point, `metaDigitise()` has a specific edit  
189 option that allows users to enter previously omitted sample sizes. This first identifies  
190 missing sample sizes in the digitised output, re-plots the relevant figures and prompts  
191 the user to enter the sample sizes for the relevant groups in the figure, one by one.

## 192 5 Software Validation

193 In order to evaluate the consistency of digitisation with **metaDigitise** between users, we  
194 got 14 people to digitise the same set of 14 figures created from a simulated dataset (see  
195 supplements). We found no evidence for any inter-observer variability in digitisations  
196 for the mean (ICC = 0, 95% CI = 0 to 0.029,  $p = 1$ ), SD (ICC = 0, 95% CI = 0 to  
197 0.033,  $p = 0.5$ ) or correlation coefficient (ICC = 0.053, 95% CI = 0 to 0.296,  $p =$   
198 0.377). There was little bias between digitised and true values, on average 1.63%  
199 (mean = 0.02%, SD = 4.9%,  $r = -0.03\%$ ) and there were small absolute differences  
200 between digitised and true values, on average 2.18% (mean = 0.40%, SD = 5.81%,  $r =$   
201 0.33%) across all three summary statistics. SD estimates from digitisations are clearly  
202 most error prone. The mean absolute differences for each plot type clearly show that  
203 this effect is driven by extraction from box plots and histograms (% difference; box plot:  
204 15.805, histogram: 5.210, mean/error: 1.500, scatter plot: 0.433). SD estimation from  
205 box plot summary statistics is known to be more error prone, especially at small sample

206 sizes (Wan et al., 2014).

207 We also used simulated data to test the accuracy of digitisations with respect to known  
208 values (see supplements). **metaDigitise** was extremely accurate at matching clicked  
209 points to their true values essentially being perfectly correlated with the true simulated  
210 data for both the  $x$ -variable (Pearson’s correlation;  $r = 0.9999915$ ,  $t = 2137.4$ ,  $df = 78$ ,  
211  $p < 0.001$ ) and  $y$ -variable ( $r = 0.9999892$ ,  $t = 1897.8$ ,  $df = 78$ ,  $p < 0.001$ ) in  
212 scatterplots.

## 213 6 Limitations

214 Although **metaDigitise** is very flexible and provides functionality not seen in any other  
215 package, there are some functions that it does not perform (see Table S1). Notably  
216 **metaDigitise** lacks automated point detection. However, from our experience, manual  
217 digitising is more reliable and often equally as fast. Given the variation in image  
218 quality, calibration for automatic point detection needs to be done for each figure  
219 individually. Additionally, auto-detection often misses points which then need to be  
220 manually added. Based on tests of **metaDigitise** (see above), figures can be extracted in  
221 around 1-2 minutes, including the entry of metadata. As a result, we do not believe  
222 that current automated point detection techniques provide substantial benefits in terms  
223 of time or accuracy.

224 **metaDigitise** also (currently) lacks the ability to zoom in on figures. Zooming may  
225 enable users to gain greater accuracy when clicking on points. However, from our own  
226 experience (see results above), with a reasonably sized screen accuracy is already high,  
227 and so relatively little gain is to be had from zooming in on points.

228 In contrast to some other packages **metaDigitise** does not extract lines from figures.  
229 Line extraction is not particularly useful for most meta-analyses, although we recognise  
230 that it may be useful in other fields. Should a user like to extract lines with

231 **metaDigitise**, we would recommend extracting data as a scatter plot, and clicking along  
232 the line in question. A model can then be fitted to these points (accessed by choosing to  
233 return processed rather than summary data) to estimate the parameters needed.

## 234 7 Conclusions

235 Increasing the reproducibility of figure extraction for meta-analysis and making this  
236 laborious process more streamlined, flexible and integrated with existing statistical  
237 software will go a long way in facilitating the production of high quality meta-analytic  
238 studies that can be updated in the future. We believe that **metaDigitise** will improve  
239 this research synthesis pipeline, and will hopefully become an integral package that can  
240 be added to the meta-analysts toolkit.

## 241 Acknowledgments

242 We thank the I-DEEL group and colleagues at UNSW for for testing, providing  
243 feedback and digitising including: Rose O’Dea, Fonti Kar, Malgorzata Lagisz, Julia  
244 Riley, Diego Barneche, Erin Macartney, Ivan Beltran, Gihan Samarasinghe, Dax Kellie,  
245 Jonathan Noble, Yian Noble and Alison Pick. J.L.P. was supported by a Swiss National  
246 Science Foundation Early Mobility grant (P2ZHP3\_164962), D.W.A.N. was supported  
247 by an Australian Research Council Discovery Early Career Research Award  
248 (DE150101774) and UNSW Vice Chancellors Fellowship and S.N. an Australian  
249 Research Council Future Fellowship (FT130100268).

## 250 Author Contributions

251 J.L.P. and D.W.A.N. conceived the study and J.L.P., S.N. and D.W.A.N. developed the  
252 idea. J.L.P. and D.W.A.N. developed the R-package. J.L.P. and D.W.A.N. wrote the  
253 first draft of the paper and J.L.P., S.N. and D.W.A.N. contributed substantially to  
254 subsequent revisions of the manuscript and gave final approval for publication.

## 255 References

- 256 Arizona-Software (2008) *GraphClick Software, Version 3.0*.
- 257 Glass, G. (1976) Primary, secondary, and meta-analysis research. *Educational*  
258 *Researcher*, **5**, 3–8.
- 259 Koricheva, J., Gurevitch, J. & Mengersen, K. (2013) Handbook of Meta-Analysis in  
260 Ecology and Evolution. *Princeton University Press, Princeton, New Jersey*.
- 261 Nakagawa, S., Noble, D.W., Senior, A.M. & Lagisz, M. (2017) Meta-evaluation of  
262 meta-analysis: ten appraisal questions for biologists. *BMC Biology*, **15**, 18; DOI  
263 10.1186/s12915-017-0357-7.
- 264 Noble, D.W., Lagisz, M., O’Dea, R.E. & Nakagawa, S. (2017) Nonindependence and  
265 sensitivity analyses in ecological and evolutionary meta-analyses. *Molecular Ecology*,  
266 **26**, 2410–2425.
- 267 Parker, T.H., Forstmeier, W., Koricheva, J., Fidler, F., Hadfield, J., En Chee, Y., Kelly,  
268 C.D., Gurevitch, J. & Nakagawa, S. (2016) Transparency in Ecology and Evolution:  
269 Real Problems, Real Solutions. *Trends in Ecology and Evolution*, **31**, 711–719.
- 270 Peng, R.D. (2011) Reproducible research in computational science. *Science*, **334**, 1226.
- 271 Peng, R.D., Dominici, F. & Zeger, S.L. (2006) Reproducible epidemiologic research.  
272 *American Journal of Epidemiology*, **163**, 783–789.

- 273 Rohatgi, A. (2017) *WebPlotDigitizer Software, Version 4.0*. Austin, Texas, USA.
- 274 Tummers, B. (2006) *DataThief Software, Version 3.0*.
- 275 Wan, X., Wang, W., Liu, J. & Tong, T. (2014) Estimating the sample mean and  
276 standard deviation from the sample size, median, range and/or interquartile range.  
277 *BMC Medical Research Methodology*, **14**, 135.

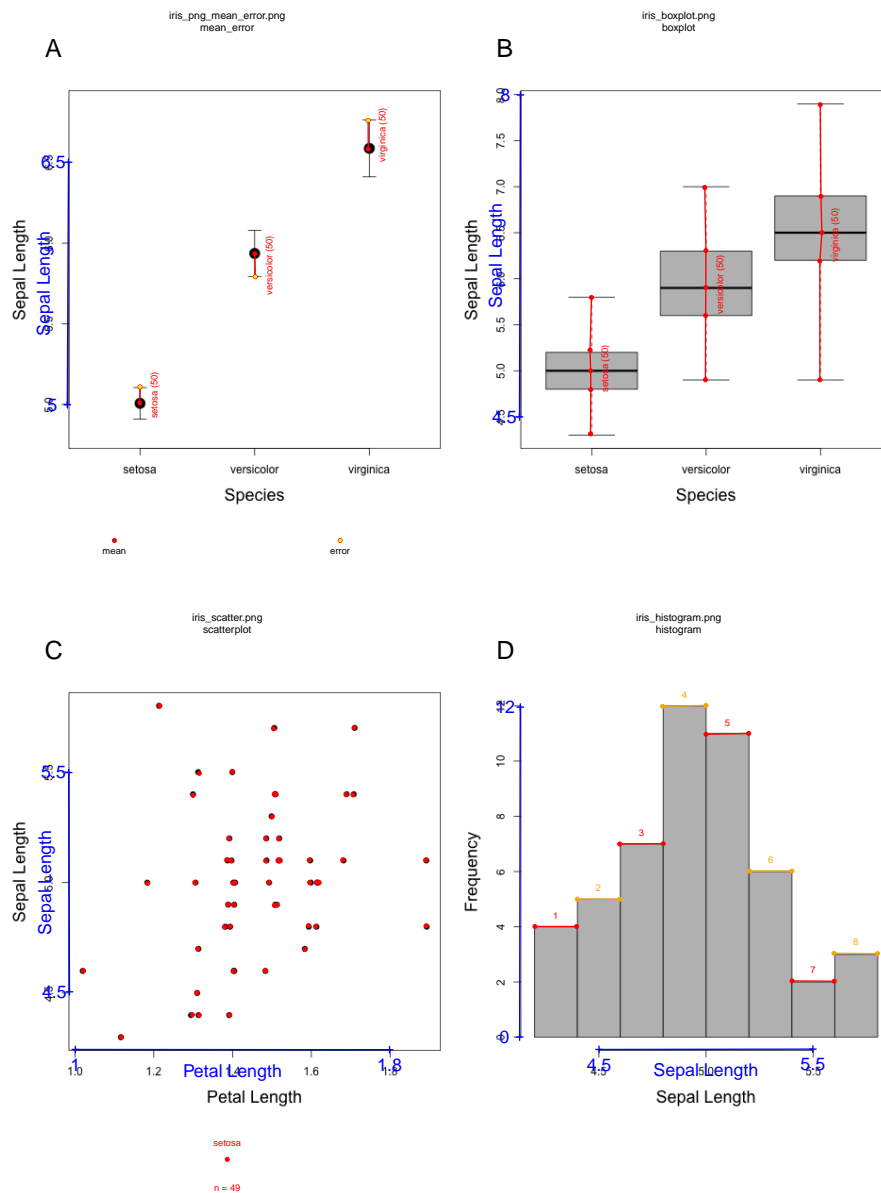


Figure 1: Four plot types that **metaDigitise** is designed to extract data from: A) mean/ error plot, B) box plot, C) scatter plot and D) histogram. Data is taken from the iris dataset in R. A and B are plotted with the whole dataset, C and D are just the data for the species *setosa*. Digitisation of the images is shown. All figures are clearly labelled at the top to remind users of the filename and plot type. This reduces errors throughout the digitisation process. Names of the variables and calibration (in blue) are plotted alongside the digitised points. In A) and B), user entered group names and sample sizes are displayed beside the relevant points. In C) the names and sample sizes for each group are shown below the figure.

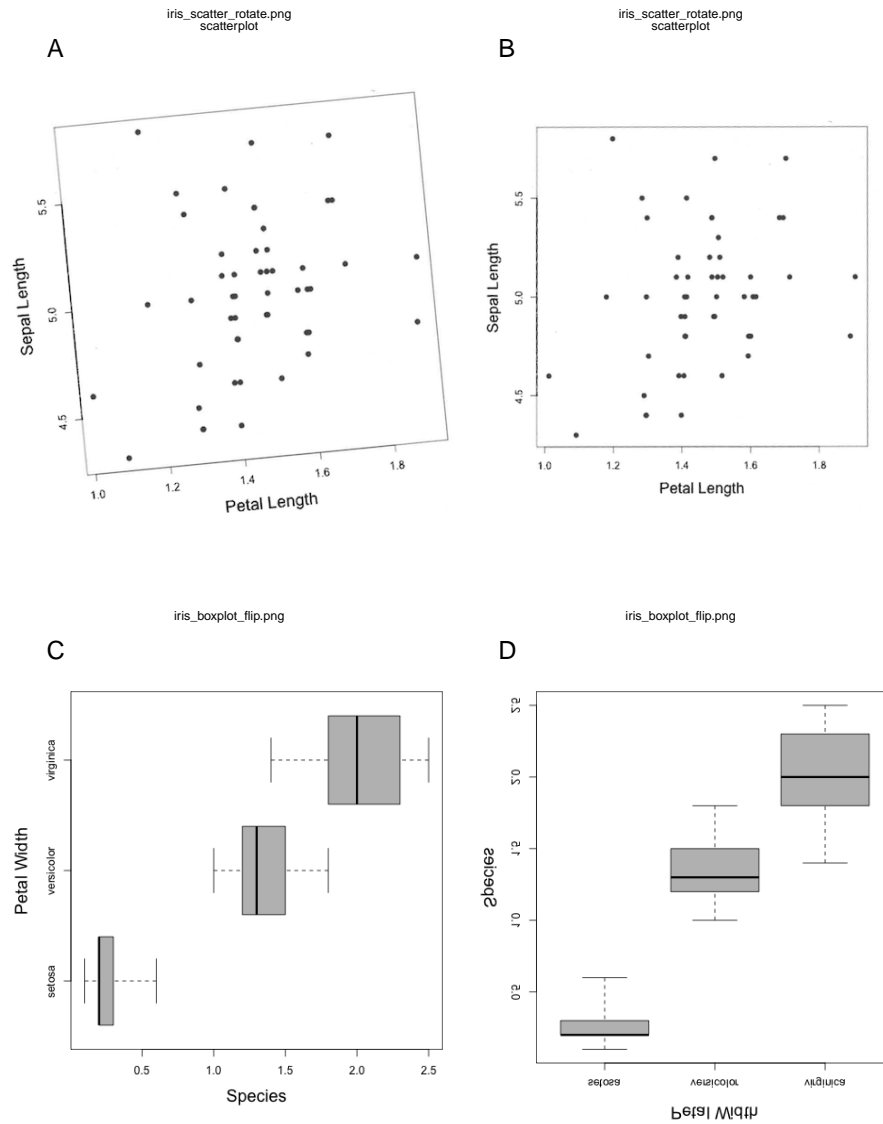


Figure 2: Figure rotation. A) and B) show how non-aligned images can be realigned through user defined rotation. C) and D) show how figures can be re-orientated so as to aid data input.