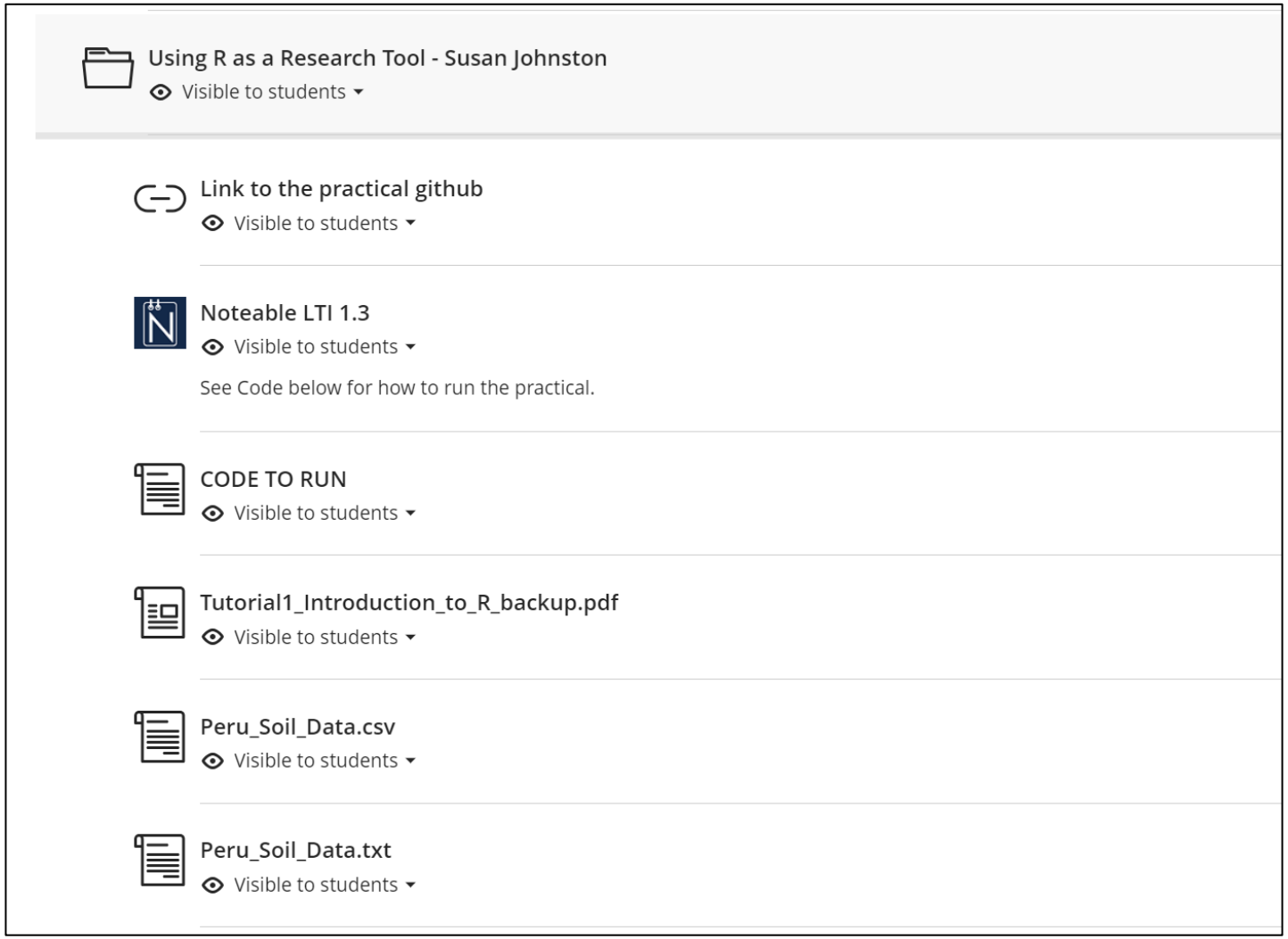
Using R as a Research Tool.

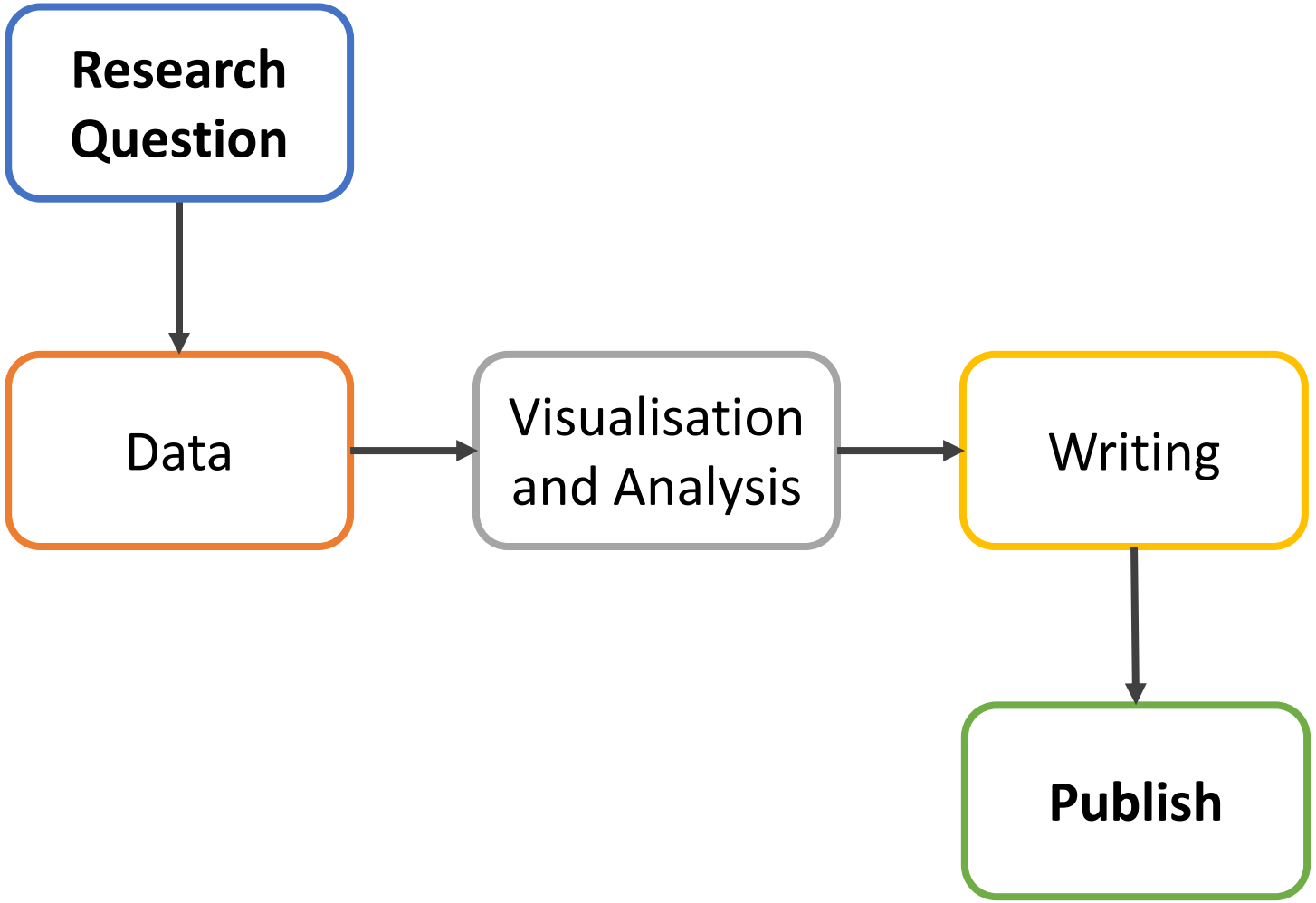
NERC E4 DTP Training

Dr Susan Johnston, Institute of Ecology and Evolution

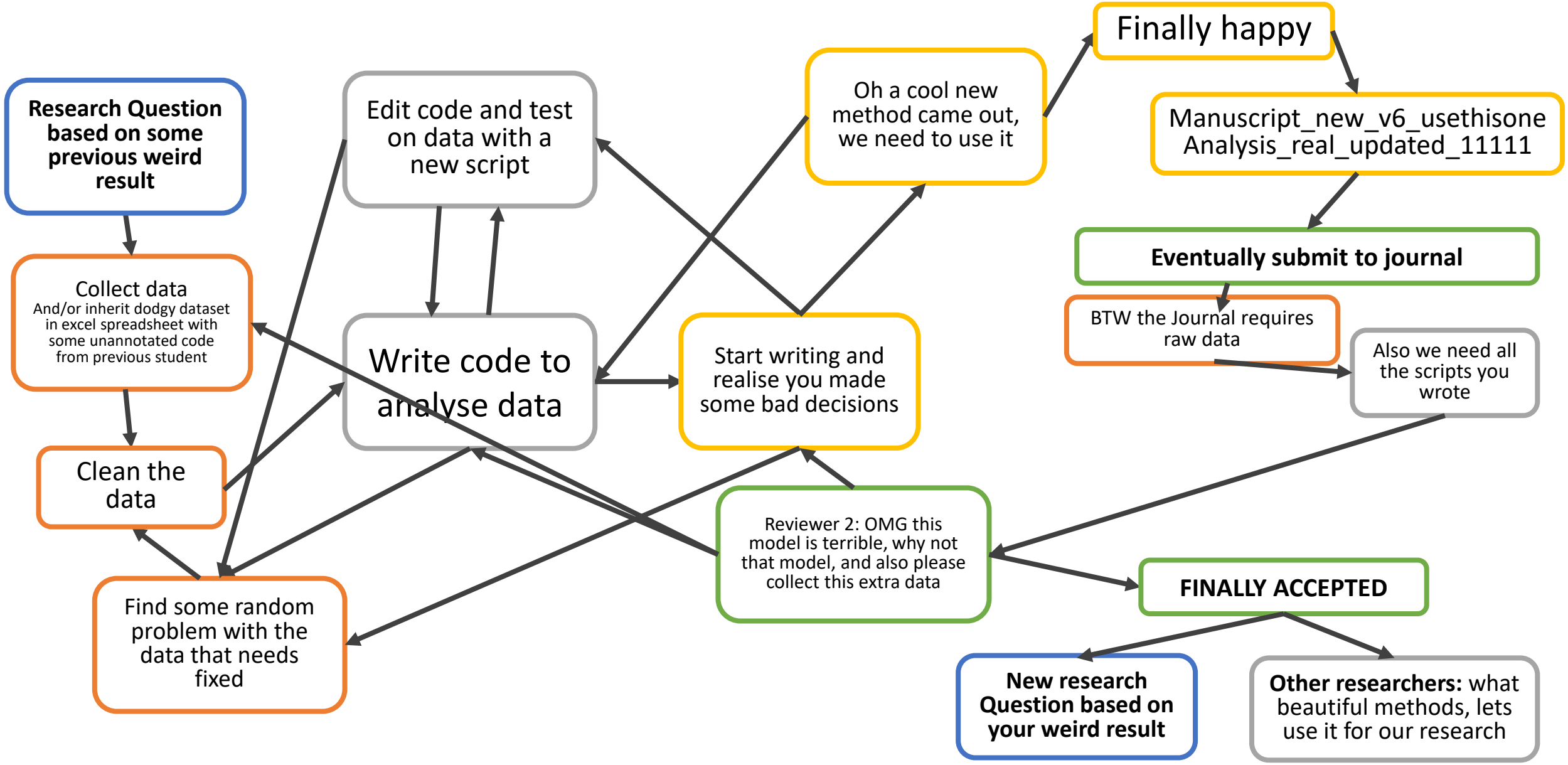


github.com/susjoh/E4StatsTutorials

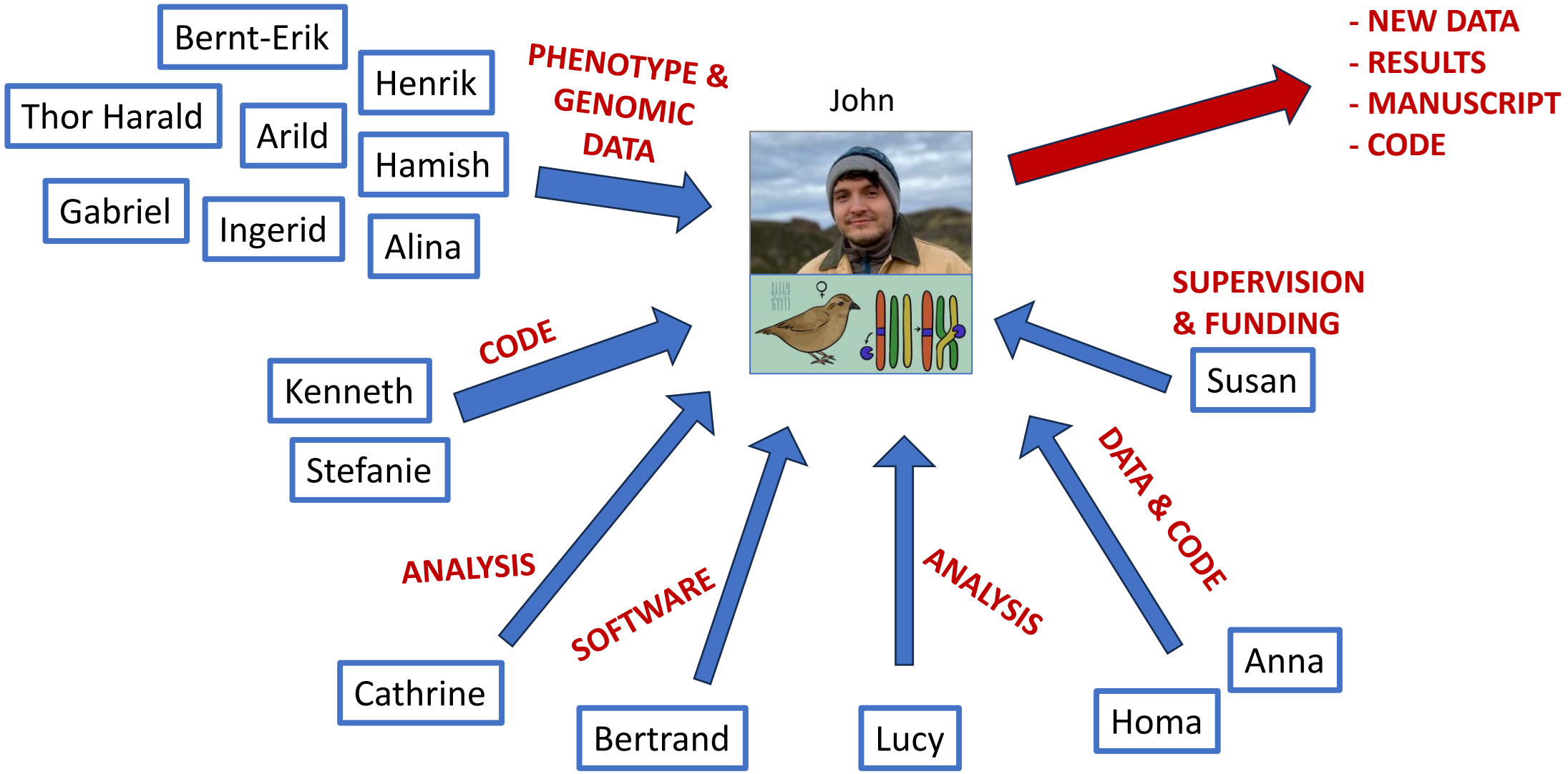
Life of a research project - expectation.



Life of a research project - reality.



Science is not a solo endeavour!



We need to be good scientists and good collaborators…

Organised data and workflows.

Reproducible analyses.

Code availability.

Version control.

What is reproducible research?

“Reproducibility is the ability of an entire experiment or study to be reproduced, either by the researcher or by someone else working independently, [and] is one of the main principles of the scientific method.“

-Wikipedia

In the lab…

8/27/08

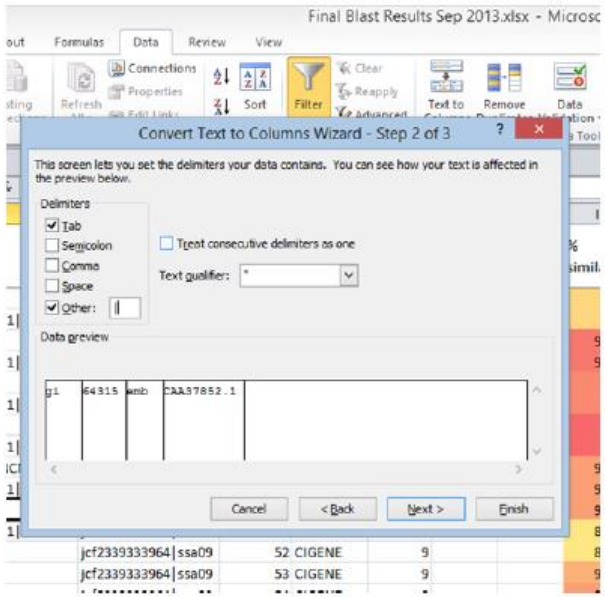
OPERON-LIKE ORGANIZATION OF THE GAL GENES

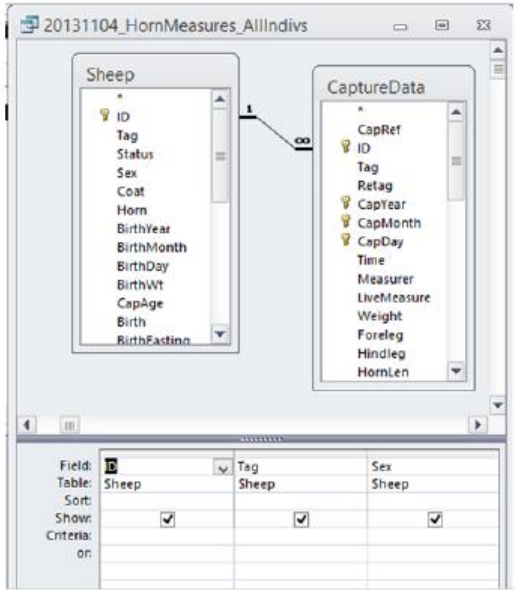
Although eukaryotes lack true operons, there are examples of operon-like gene clusters. Three examples are the galactose utilization genes in S. cerevisiae (GAL1, GAL10, GAL7), the allantoin degradation genes in S.c. (DAL1, DAL2, DAL3, DAL4, DAL7, DCG1), and the thaliandiol synthesis genes in Arabidopsis (THAS, THAH, THAD):

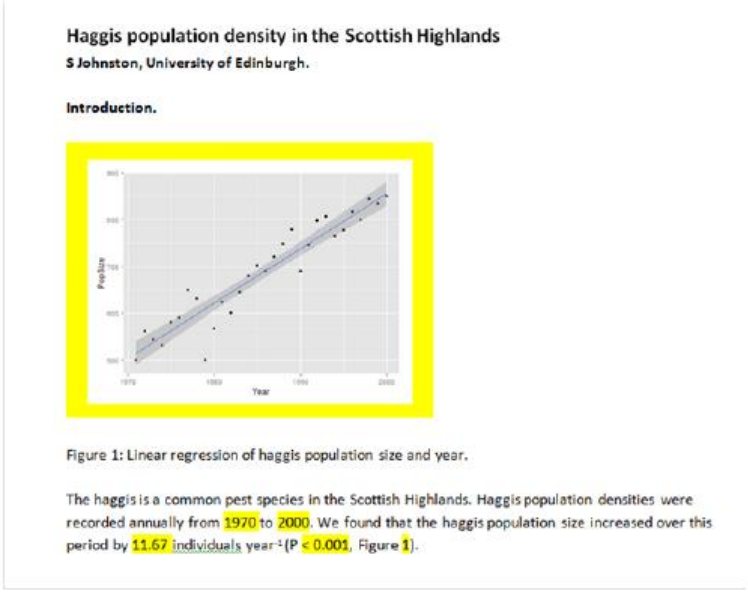
Two explanations have been given to account for this organization: genetic ~~clustering~~ linkage and metabolic channeling.

The genetic linkage hypothesis seems to be favored in the literature. It is interesting to note, however, that all three pathways above have intermediates that are toxic to the organism (in red). Here I want to test the hypothesis that the operon-like organization allows for better co-regulation of the genes and helps ~~channel~~ maintain flux through the pathway thus prevent the accumulation of the toxic intermediate ⟹

Many of us are clicking, copying and pasting...

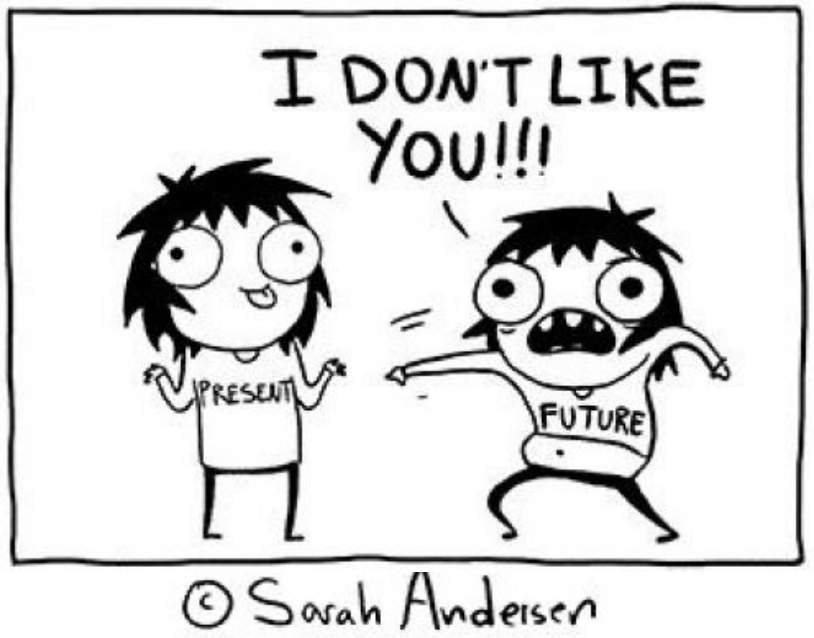






- Can you repeat all of this again...
- ...and would you get the same results every time?

Scenarios that benefit from reproducibility



- The first researcher who will need to reproduce results is likely to be **YOU.**
- New data becomes available.
- You return to a project after a period of time.
- You give the project to a new student/collaborator.
- A reviewer wants you to change something.
- Other researchers want to used your data/methods.
- You found an error, but not sure how it happened.

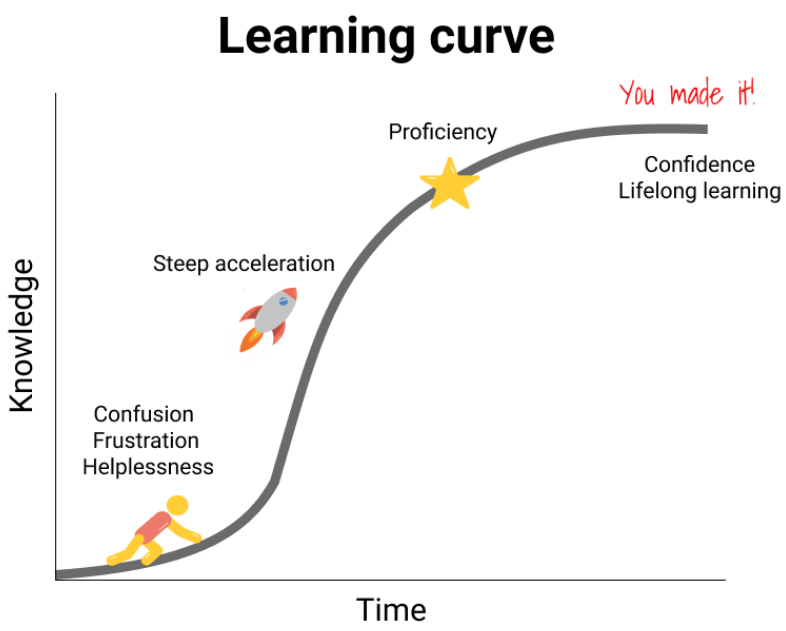
Coding is Key!

"This is R. There is no if. Only how."
-- Simon `Yoda' Blomberg, R-help (April 2005)

- Automates analyses.
- Provides a record of all data manipulation analysis.
- Can be adapted and rerun.
- You should always use code whenever possible.

```r
#
# Car speed analysis
# SEJ, Sep 2023
#

data(cars)

# Calculate the mean speed and distance

mean(cars$speed)
mean(cars$dist)

# Linear regression:

car_regression <- lm(speed ~ dist, data = cars)
car_regression                  # very basic results
summary(car_regression)         # more detailed results

# Plot the regression

plot(speed ~ dist, data = cars)
abline(car_regression)

```

Learning to code is about the ✨ process ✨



It will take time
(but not much time!)

You will (should) screw up
(a lot) and learn many lessons.

Code can be improved: it's
fine to get the job done.

Anyone can code.

Ada Lovelace, 1840

FEATURE

The Secret History of Women in Coding

Computer programming once had much better gender balance than it does today. What went wrong?

https://www.nytimes.com/2019/02/13/magazine/women-coding-computer-programming.html

Mary Jackson at NASA, 1977

Rear Admiral Grace Hopper, 1960

What is R?

- Environment for statistical computing and graphics.
- Interactive programming language.
- 20,406 packages on CRAN.
- **Free and open-source** multi-platform software.

What is RStudio®?

- Integrated development environment (IDE) for running R.
- Integrates other tools to aid reproducibility.
- R Projects allow portability.
- **Free** multi-platform software.

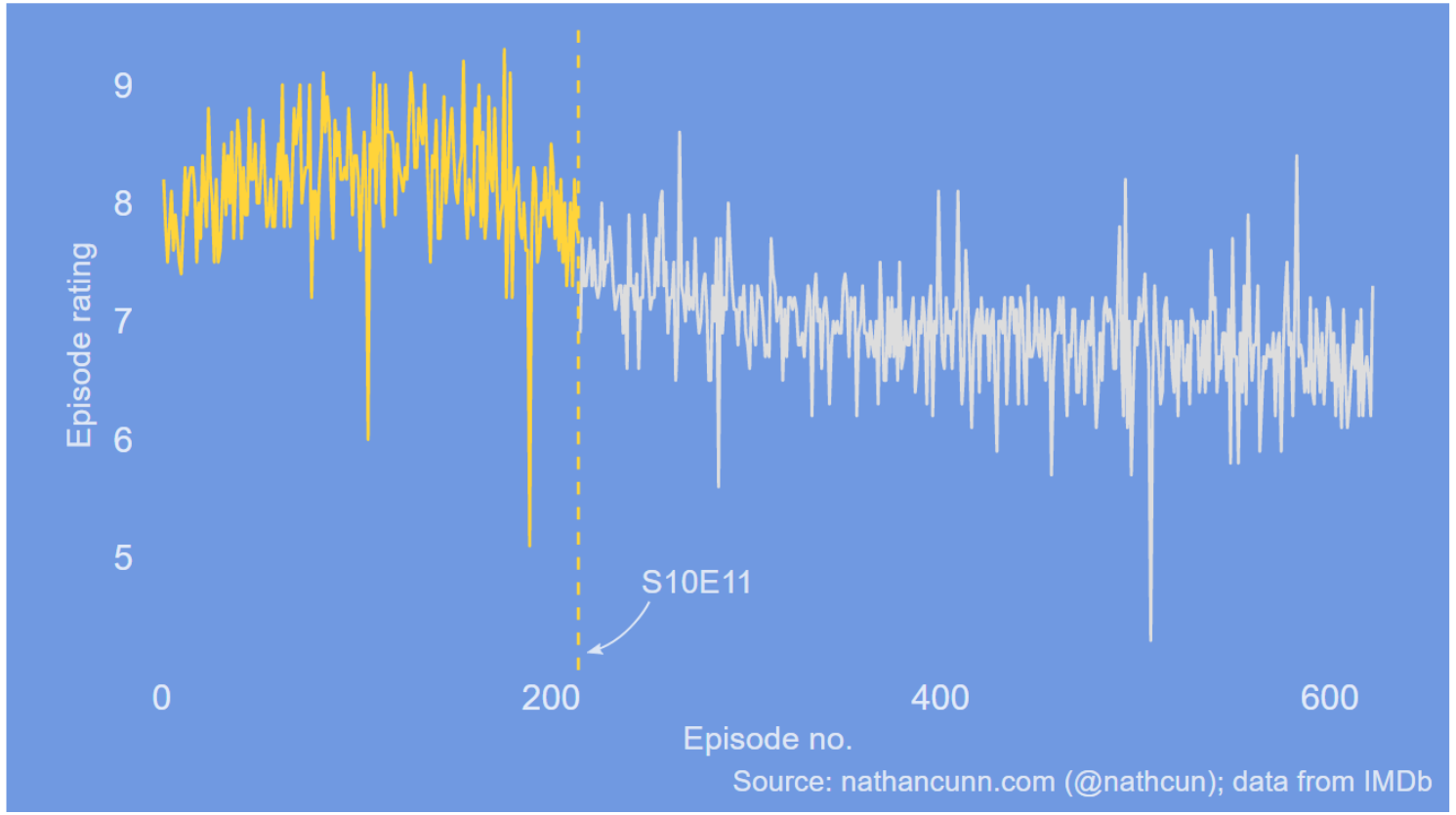
Why use R?

"This is R. There is no if. Only how."
-- Simon `Yoda' Blomberg, R-help (April 2005)

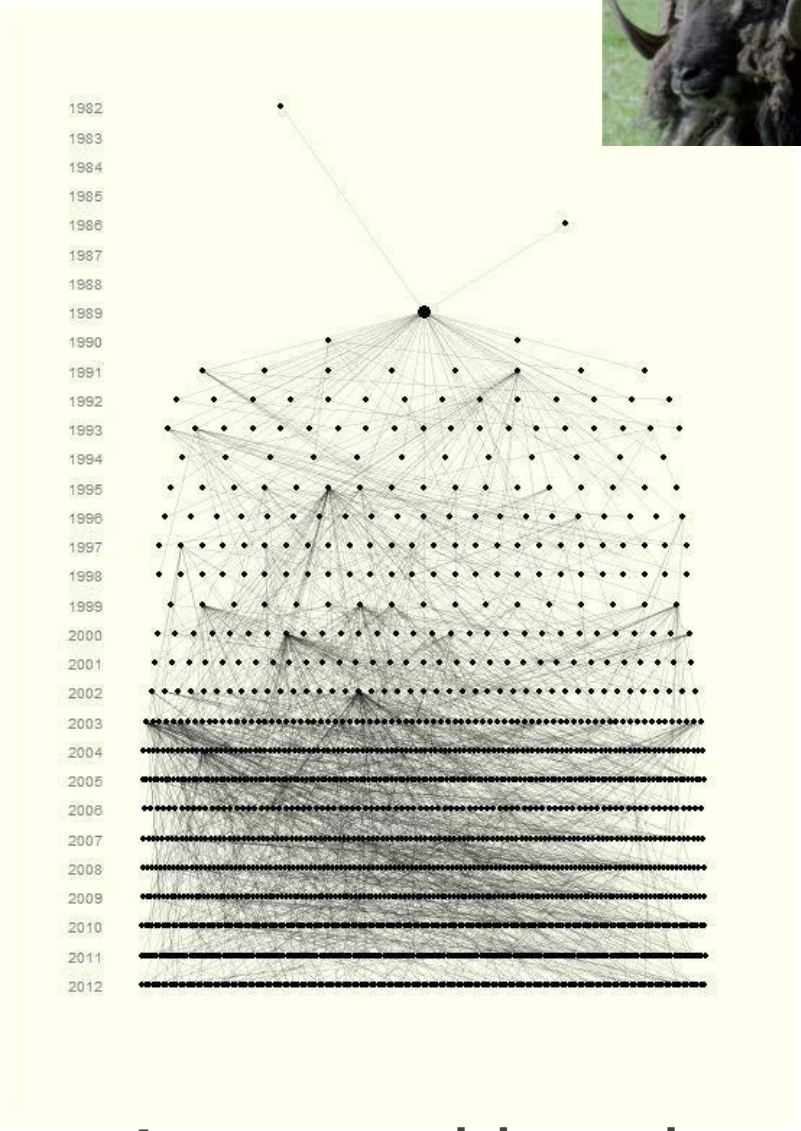
- Statistics.

Data visualisation

e.g. http://www.r-graph-gallery.com/portfolio/ggplot2-package/



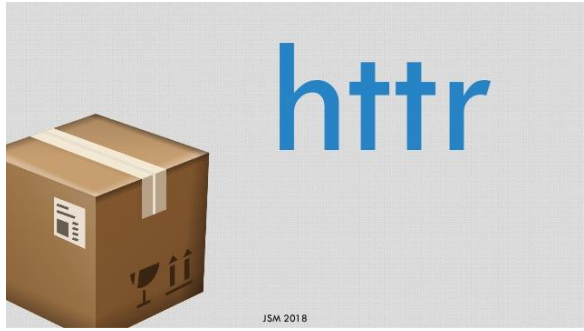
When did the golden age of The Simpsons end?

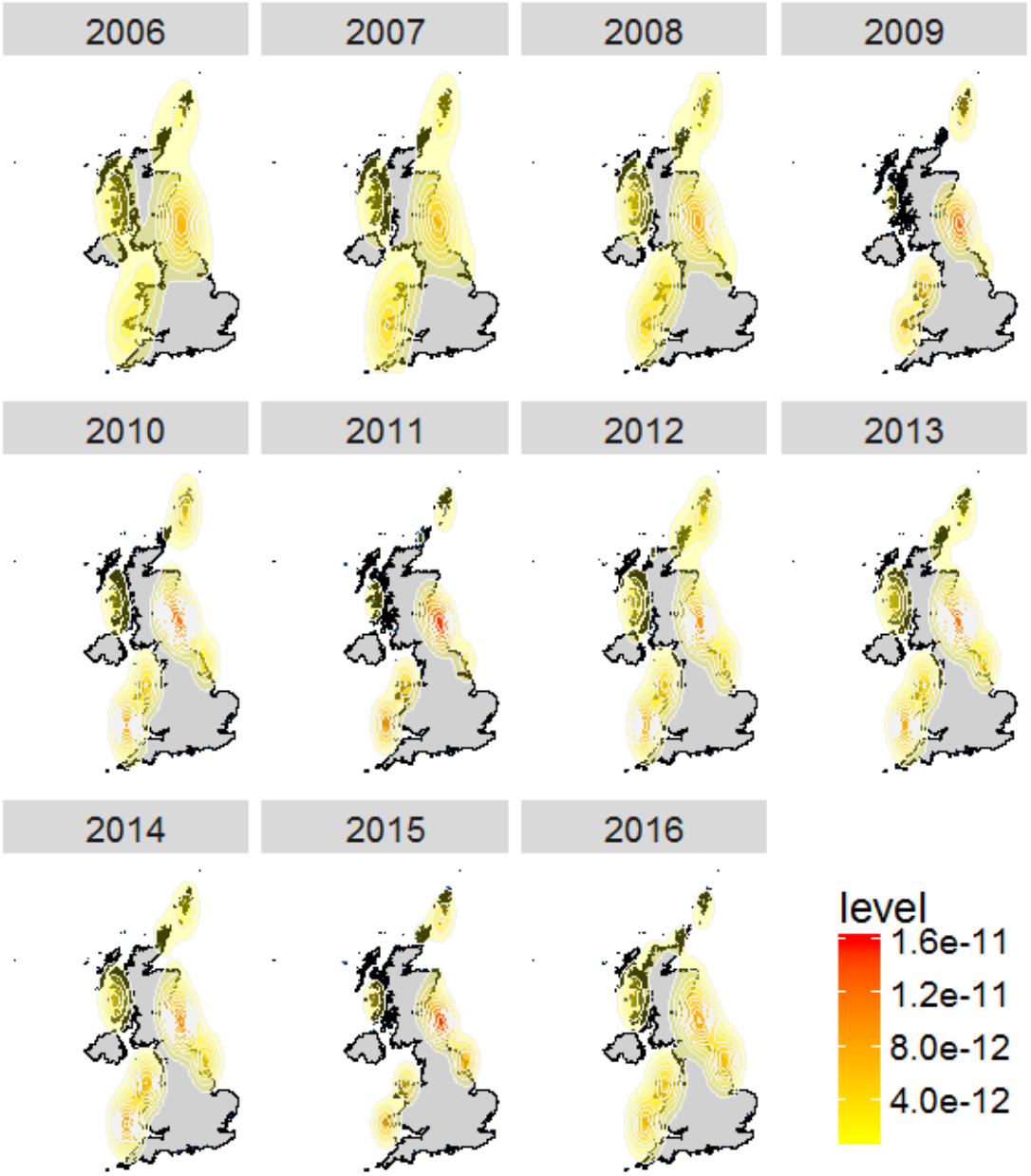


Ancestors and descendants of a single Soay sheep called Snowball.

UK distribution of Atlantic Puffins

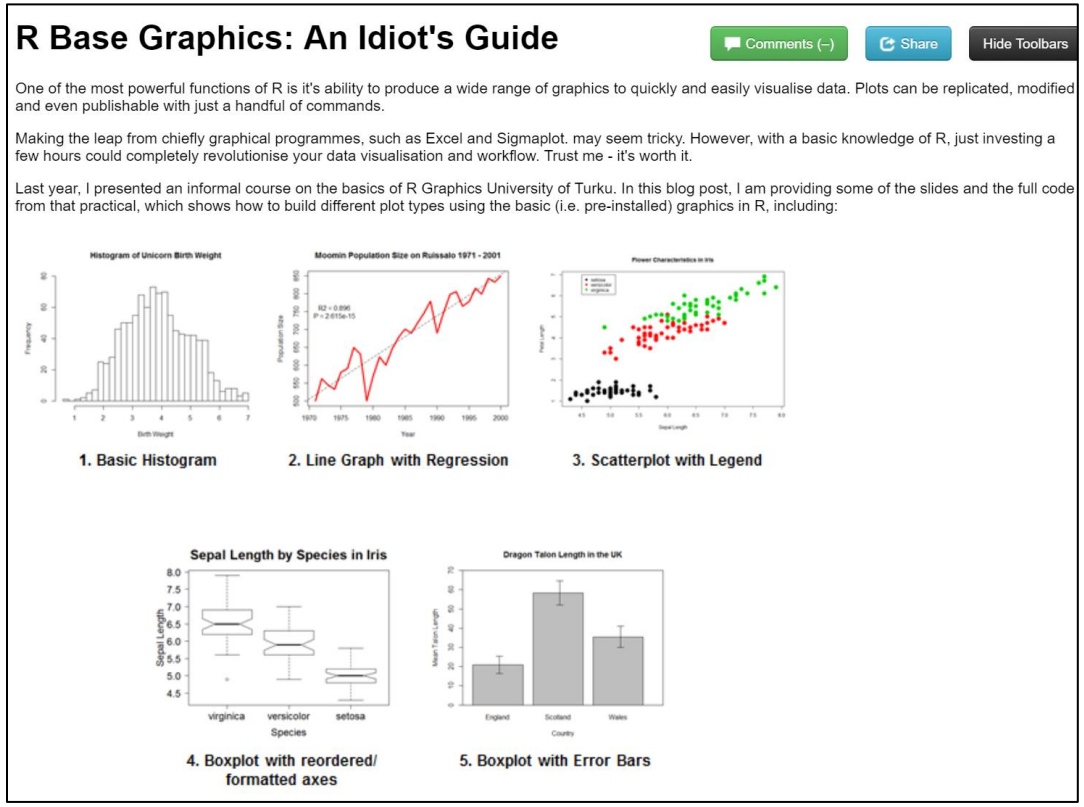
Access data from the Global Biodiversity Information Facility And Flickr directly through R





Team Shrub in School of Geosciences:
https://ourcodingclub.github.io/tutorials/seecc_1/index.html

Report writing



R Base Graphics: An Idiot's Guide

Comments (–) Share Hide Toolbars

One of the most powerful functions of R is it's ability to produce a wide range of graphics to quickly and easily visualise data. Plots can be replicated, modified and even publishable with just a handful of commands.

Making the leap from chiefly graphical programmes, such as Excel and Sigmaplot. may seem tricky. However, with a basic knowledge of R, just investing a few hours could completely revolutionise your data visualisation and workflow. Trust me - it's worth it.

Last year, I presented an informal course on the basics of R Graphics University of Turku. In this blog post, I am providing some of the slides and the full code from that practical, which shows how to build different plot types using the basic (i.e. pre-installed) graphics in R, including:

Using R as a Research Tool.

Dr Susan Johnston: Susan.Johnston@ed.ac.uk

Demonstrators: Gergana Dalaskova, John Godlee.
Hat-Tips to Kyle Dexter, The Coding Club and R4all.

November 6, 2017

1 Introduction

1.1 What is R?

R began its life in New Zealand in 1993 as a language and environment for statistical computing and graphics. It is an interpreted programming language, meaning that rather than pointing and clicking, the user types in commands. It is **free** and works across all platforms.

1.2 Why use R?

`knitr` to HTML

LaTeX and R Sweave

https://scotland.shinyapps.io/babynames/

Interactive applications (shiny)

National Records of Scotland

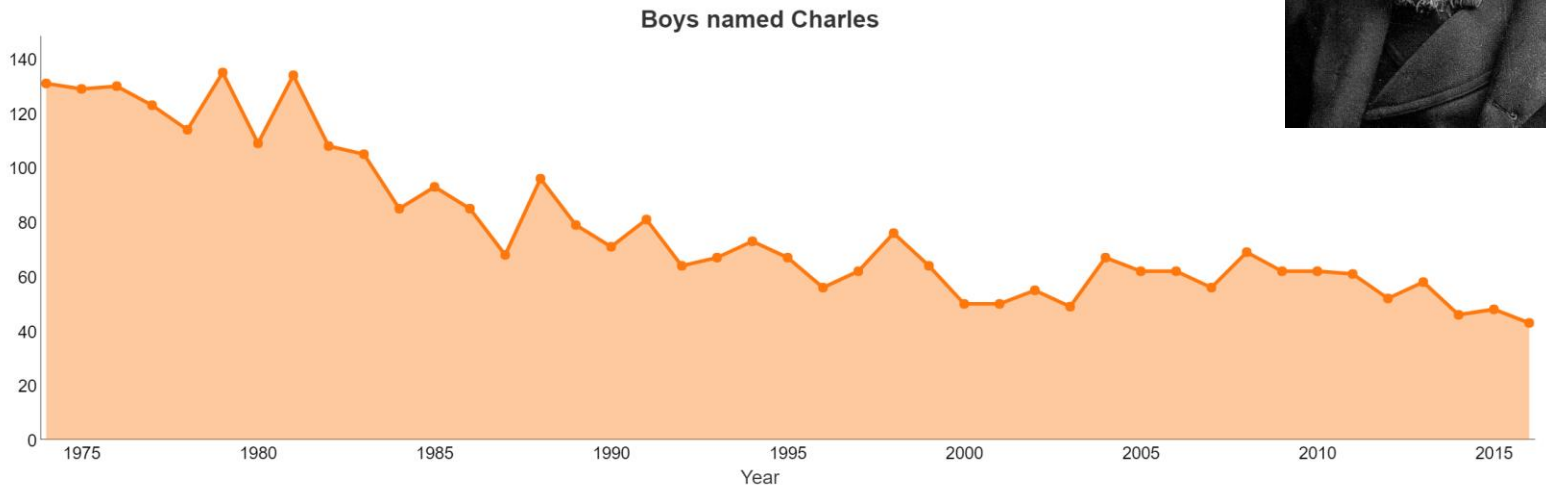
Baby names trends in Scotland since 1974

Enter a **name,** select the **gender** and click on **'Apply'** to see how a name's popularity has changed over the years.
App might be slow at busy times. Please be patient.

Name Charles

Gender Male

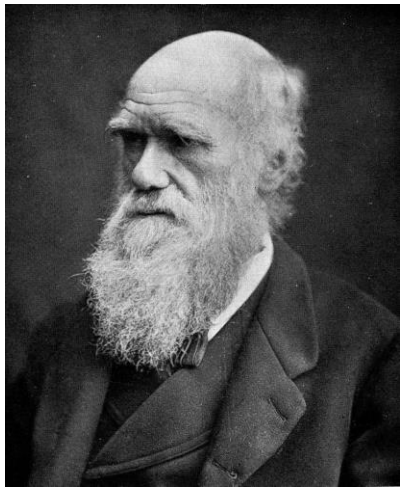
Apply



How to use this app
Hover over years to highlight individual values
Click and drag to zoom
Double-click to zoom out

Data: Baby names, Scotland, 1974-2016 (xlsx)
Data: Baby names, Scotland, 1974-2016 (csv)
Publications: Baby names, Scotland, 1974-2016

National Records of Scotland
© Crown Copyright 2017 - Copyright conditions
Follow us on Twitter - @NatRecordsScot
See more Infographics & Visualisations



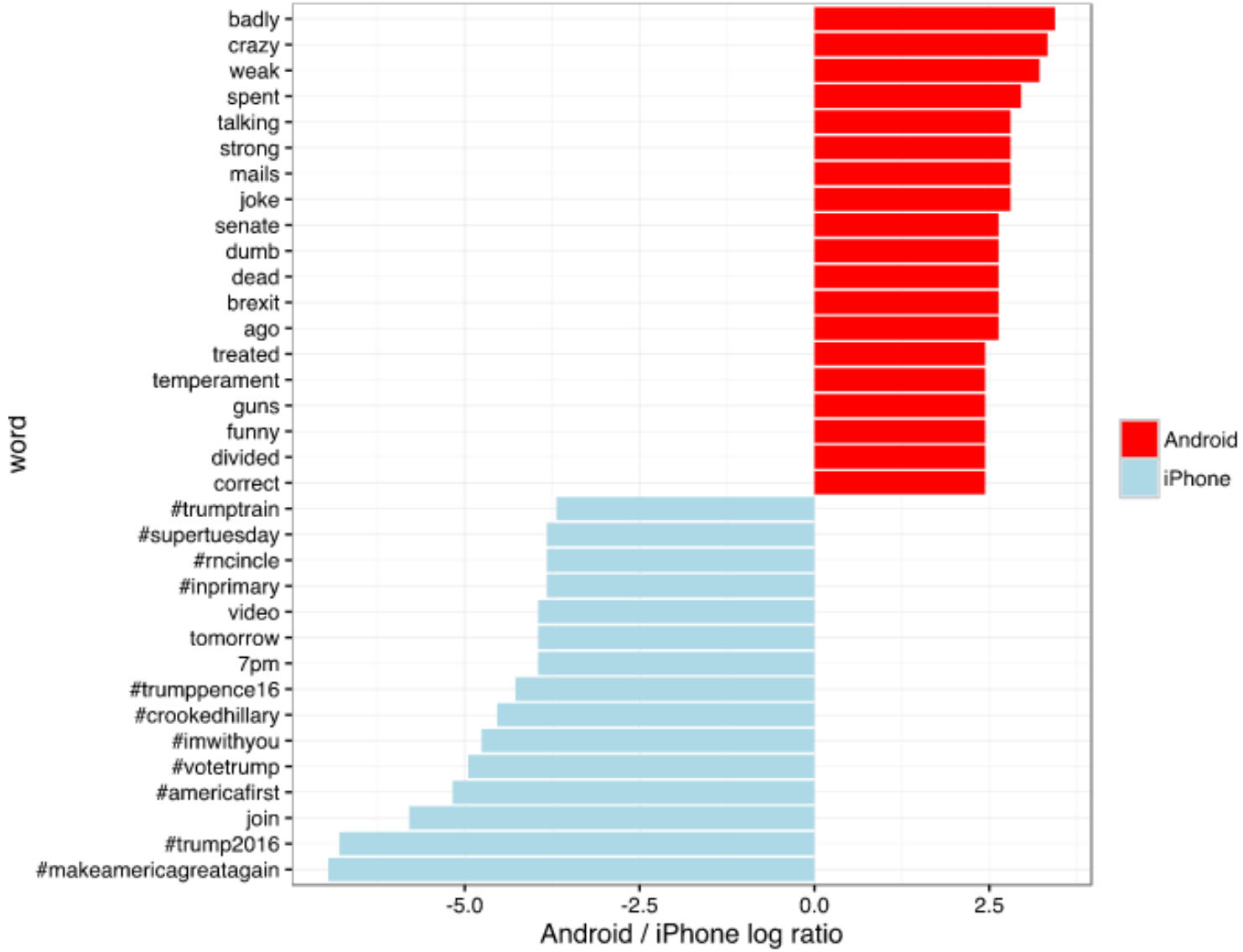
Analytics e.g.

Todd Vaziri
@tvaziri

Every non-hyperbolic tweet is from iPhone (his staff).

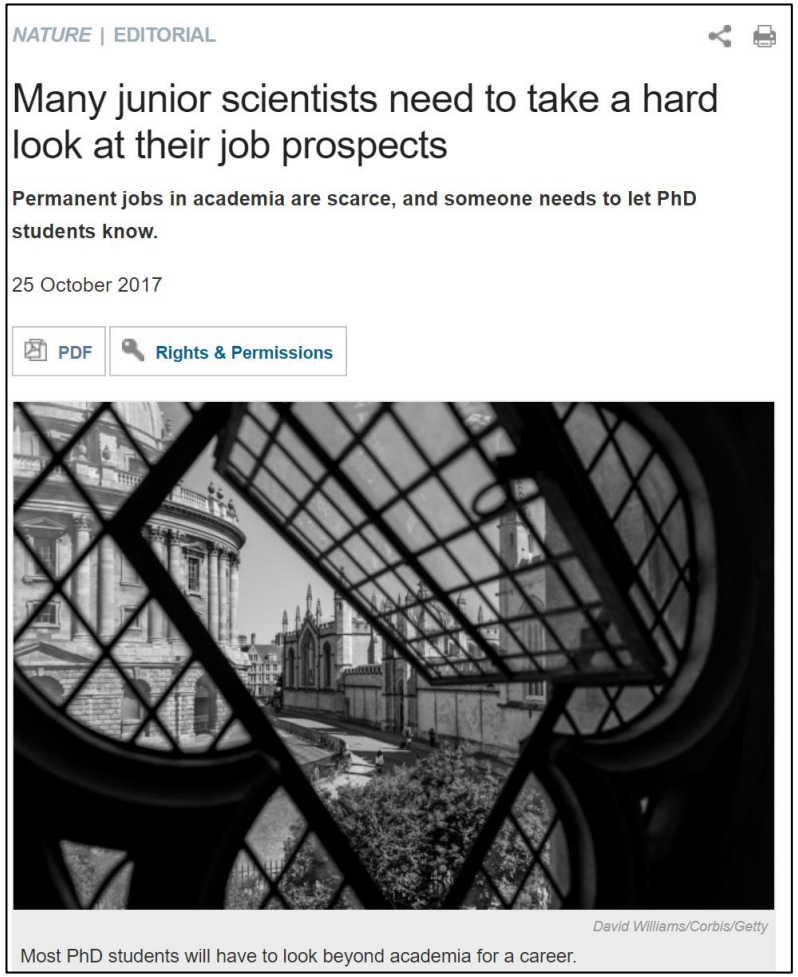
Every hyperbolic tweet is from Android (from him).

http://varianceexplained.org/r/trump-tweets/



Programmers are in demand.

- Transferable skill which makes you competitive for postdocs and academic positions.
- Similar to Python and easy path to other languages.
- Research companies, Facebook, Google, Twitter, AirBnB.
- Scotland R jobs at Scottish Government, Met Office, JP Morgan, RBS & other Banks, Rockstar North, University of Edinburgh, Energy Companies, start-ups, etc.



NATURE | EDITORIAL

Many junior scientists need to take a hard look at their job prospects

Permanent jobs in academia are scarce, and someone needs to let PhD students know.

25 October 2017

PDF Rights & Permissions

David Williams/Corbis/Getty

Most PhD students will have to look beyond academia for a career.

https://www.nature.com/news/many-junior-scientists-need-to-take-a-hard-look-at-their-job-prospects-1.22879

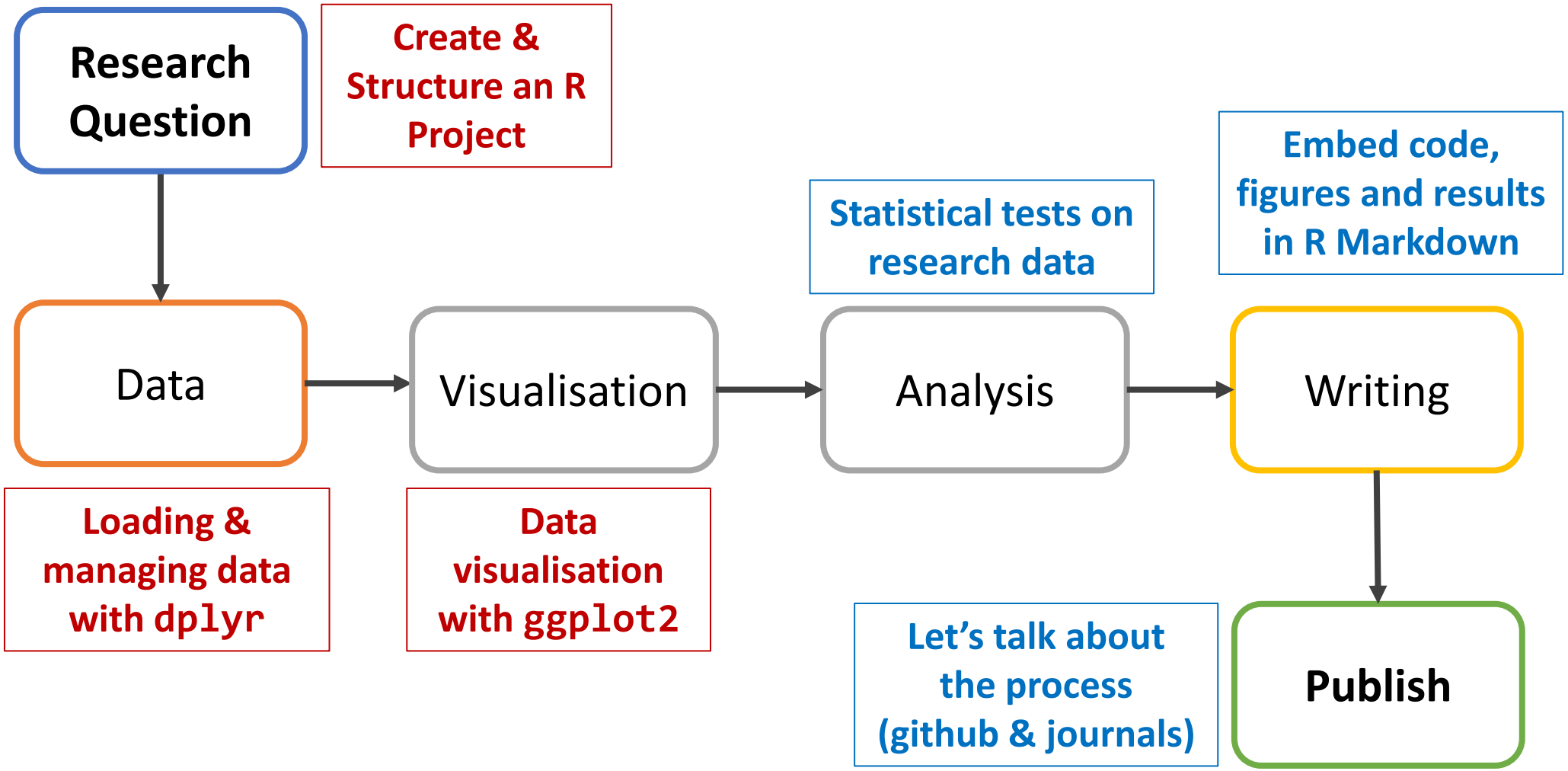
Best method is implemented in X.

Collaborator uses X.

Future employer uses X.

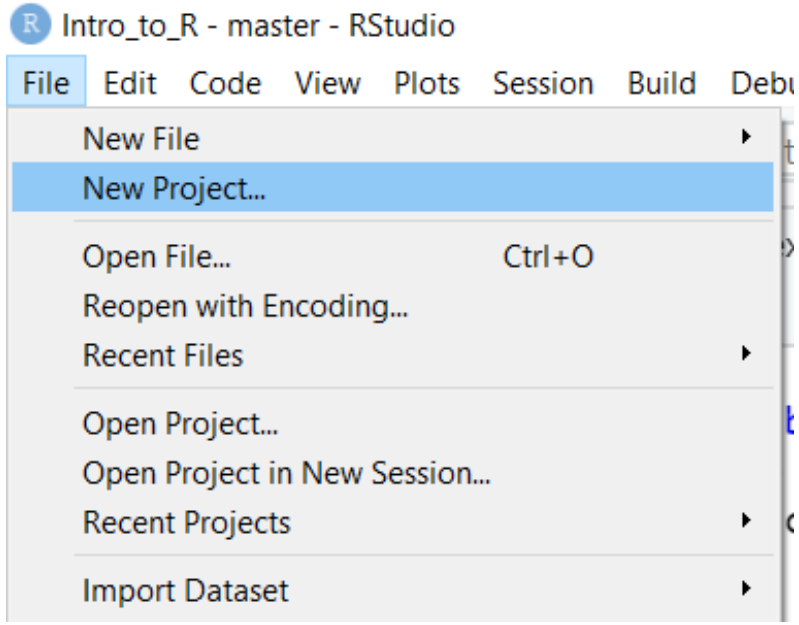
Your life could be improved by using X.

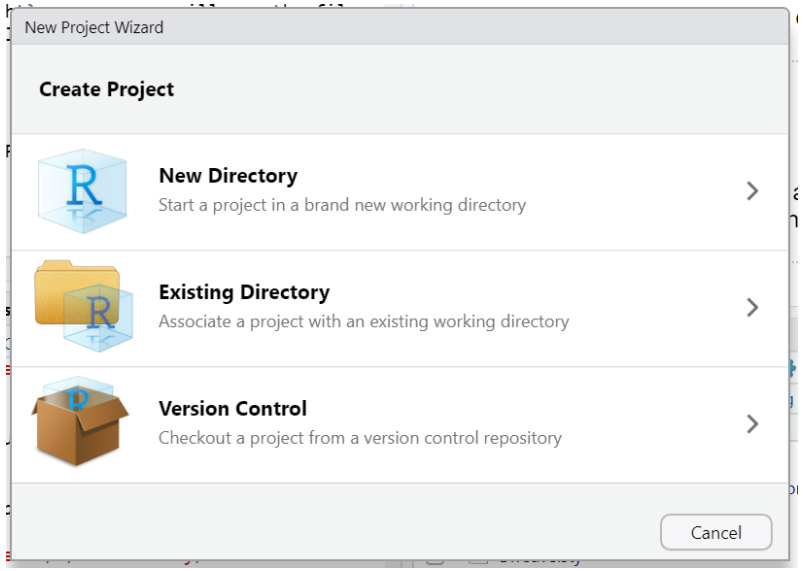
Life of a research project.



Using R Projects.

- Establishes a folder with an associated .Rproj
- One folder, one portable project.
- Saves history, profile, etc.
- Allows version control within R Studio (e.g. git)





Structuring an R Project.

https://nicercode.github.io/blog/2013-05-17-organising-my-project/

https://nicercode.github.io/blog/2013-04-05-projects/

Soft coral database backup

View Quick Look Action Dropbox

Name	Date Modified	Size	Kind
sc_database_cyclone_003.csv	03/05/2013 9:51 AM	33 KB	comm...values
function trial stacked histogram.R	30/04/2013 12:13 PM	4 KB	R Source File
plots	30/04/2013 12:13 PM	--	Folder
sc_database_areas.csv	29/04/2013 2:08 PM	49 KB	comm...values
growth_categories_log_91-92.csv	26/04/2013 2:02 PM	8 KB	comm...values
cummulative plot function.R	26/04/2013 1:55 PM	4 KB	R Source File
function for mort.freg.fusi.R	26/04/2013 11:59 AM	4 KB	R Source File
function for plotting outlines.R	22/04/2013 9:08 PM	12 KB	R Source File
outlines_area_calculations (#1) function.R	19/04/2013 5:43 PM	4 KB	R Source File
figure-fun copy.R	19/04/2013 2:05 PM	4 KB	R Source File
comparison histograms qplot 91-92.pdf	18/04/2013 11:32 AM	213 KB	Portab... (PDF)
comparison histograms qplot 89-91.pdf	18/04/2013 11:30 AM	233 KB	Portab... (PDF)
comparison histograms qplot.pdf	18/04/2013 11:22 AM	233 KB	Portab... (PDF)
merge_dataset_plots(#2).R	17/04/2013 1:56 PM	8 KB	R Source File
growth_categories_89-91.csv	17/04/2013 1:18 PM	8 KB	comm...values
growth_categories_91-92.csv	17/04/2013 1:09 PM	8 KB	comm...values
growth_categories.csv	16/04/2013 4:36 PM	8 KB	comm...values
mortality_fragmentation(#3).R	04/04/2013 4:33 PM	8 KB	R Source File
voronoid_polygons.R	04/04/2013 4:32 PM	4 KB	R Source File
summary plots and glms.docx	04/04/2013 9:41 AM	1.5 MB	Micro...ument
area_output.csv	03/04/2013 8:01 AM	16 KB	comm...values
area_script_current2.R	02/04/2013 4:17 PM	12 KB	R Source File
Appendix.docx	22/02/2013 3:17 PM	2.3 MB	Micro...ument
Chapter1 concept sheet_Marcela Diaz.pdf	22/02/2013 3:03 PM	3 MB	Portab... (PDF)
Appendix.pdf	22/02/2013 3:00 PM	2.6 MB	Portab... (PDF)
draft concept sheet_feb21.pdf	22/02/2013 2:56 PM	594 KB	Portab... (PDF)
draft concept sheet_feb21.doc	22/02/2013 2:52 PM	524 KB	Micro...ument
scale diagram_2.png	22/02/2013 2:49 PM	37 KB	Portab...image
scale diagram.png	22/02/2013 2:43 PM	41 KB	Portab...image
draft concept sheet_feb18.doc	21/02/2013 1:52 PM	90 KB	Micro...ument
Scripts from alisha's	21/02/2013 1:44 PM	--	Folder

All data, scripts and output should be kept within the same project directory (*where possible*).

Keep data here (**read only**)

Save figures here

Keep manuscript and reports here

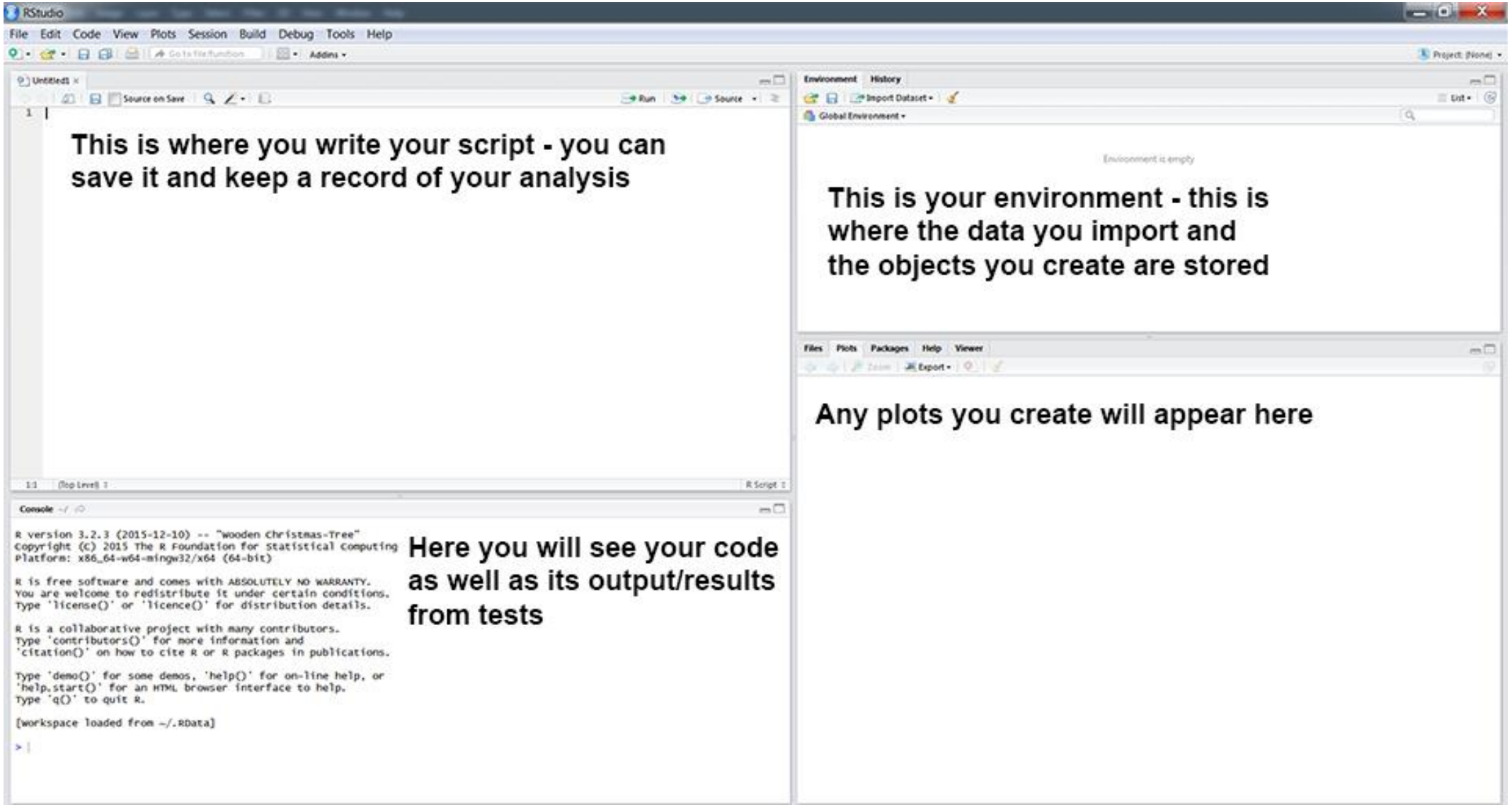
Save results here

Name	Date modified
data	29/05/2020 23:09
docs	1
figs	29/05/2020 23:07
results	20/05/2020 23:30
1_Animal_Models.R	19 01:55
2_Exploratory_GWAS.R	29/05/2020 23:14
3_Genes_in_Sig_Regions.R	08/12/2019 00:30
4_Telomere_Positions.R	31/10/2019 15:29
Soay_Telomere_GWAS.Rproj	09/06/2020 09:36

- 1_Animal_Model.png
- 1_GRM_Animal_Model.png
- 1_PED_Animal_Model.png
- 1_Sex_Age.png
- 1_Sex_AgeClass.png
- 2_Distance_from_Telomere.png
- 2_Distance_to_Telomere_Binned.png
- 2_GWAS_For_Paper.png

R and the Rstudio Environment

https://ourcodingclub.github.io/tutorials/intro-to-r/



Finding help.

- In R…
  - ? searches for a specific function.
  - ?? searches for a specific string.
  - Help tab in RStudio
- Online…
  - ourcodingclub.github.io
  - Stack Overflow
  - R Cheatsheets

Tidyverse

Packages Blog Learn Help Contribute

R packages for data science

The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

Install the complete tidyverse with:

```
install.packages("tidyverse")
```

Loading data into R

Data management in R with base R & `dplyr`

- Summarise data with `summary()`
- Sort data with `arrange()`
- Select columns with `select()`
- Adding columns with `$`
- Select rows with `filter()`

filter()

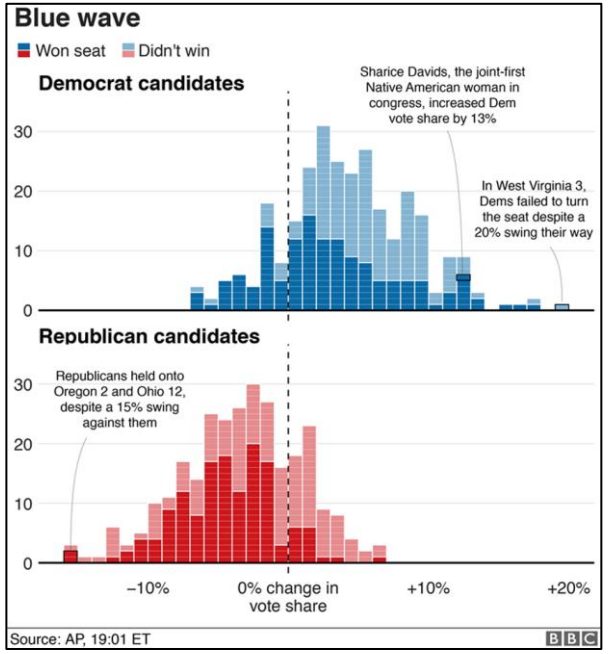
Operator	Function
$<$	less than
$>$	greater than
$=<$	less than or equal to
$=>$	greater than or equal to
$==$	equals
$!=$	does not equal
$\%in\%$	matches

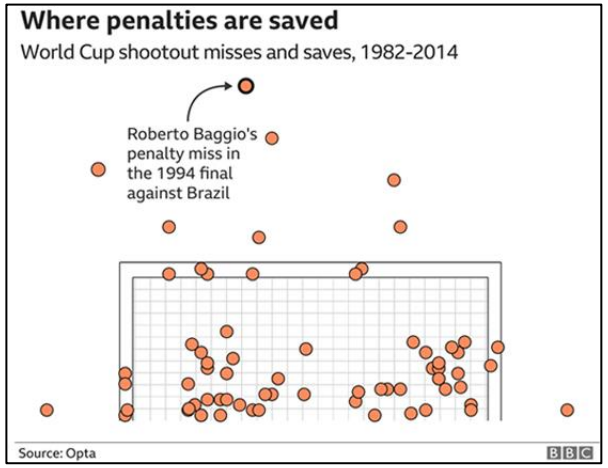
Data visualisation with

ggplot2

Data visualisation…

- Visual representation of data.
- Plots, charts, maps, infographics.
- Accessible way to identify patterns, trends and problems in your data.
- Indicates data quality and how it should be analysed.



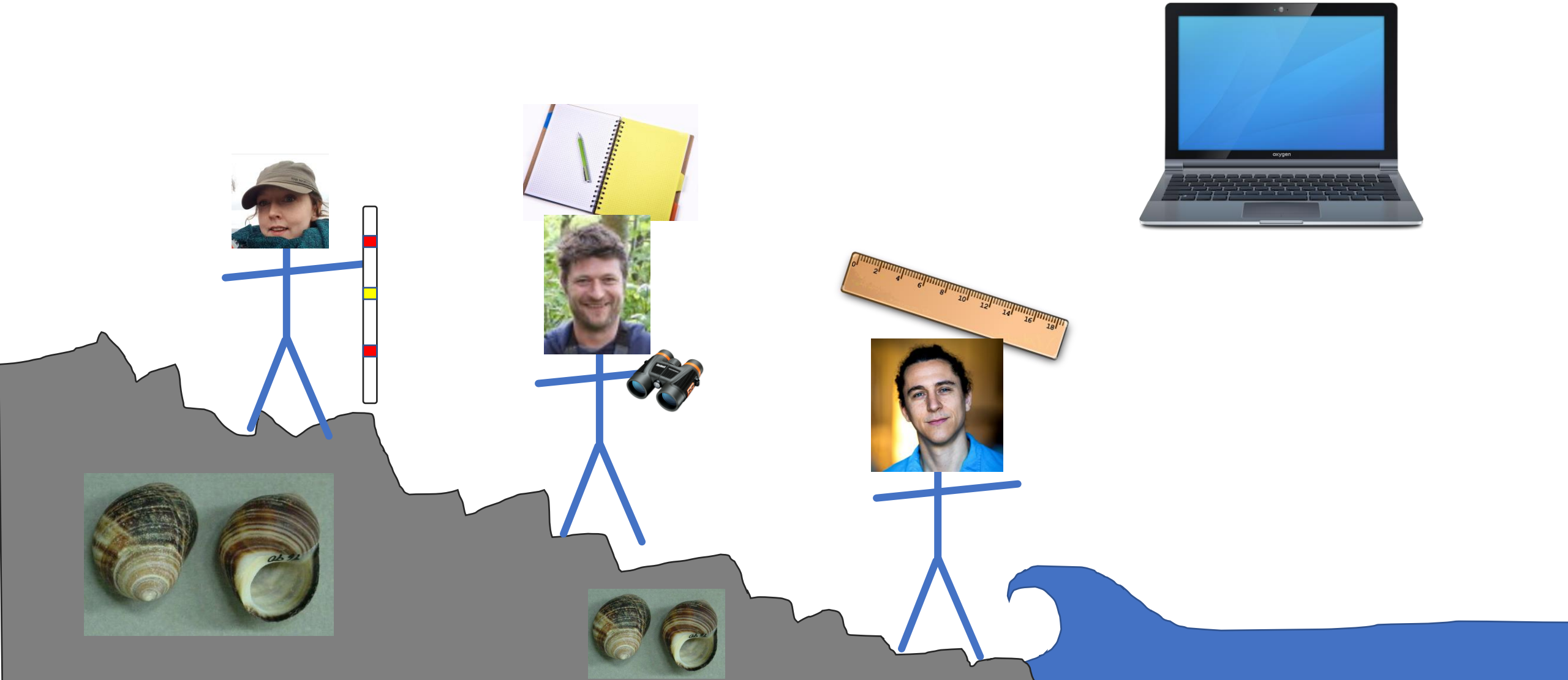


Examples from the BBC (plotted in R!)

https://bbc.github.io/rcookbook/

The first step in any analysis is to PLOT YOUR DATA!

Do *Littorina* vary in size with shore height?



x = Height on the sea shore (m)
y = Shell size (mm)

I	
x	y
10.0	8.04
8.0	6.95
13.0	7.58
9.0	8.81
11.0	8.33
14.0	9.96
6.0	7.24
4.0	4.26
12.0	10.84
7.0	4.82
5.0	5.68

x = Height on the sea shore (m)
y = Shell size (mm)

I		II	
x	y	x	y
10.0	8.04	10.0	9.14
8.0	6.95	8.0	8.14
13.0	7.58	13.0	8.74
9.0	8.81	9.0	8.77
11.0	8.33	11.0	9.26
14.0	9.96	14.0	8.10
6.0	7.24	6.0	6.13
4.0	4.26	4.0	3.10
12.0	10.84	12.0	9.13
7.0	4.82	7.0	7.26
5.0	5.68	5.0	4.74

x = Height on the sea shore (m)
y = Shell size (mm)

I		II		III	
x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46
8.0	6.95	8.0	8.14	8.0	6.77
13.0	7.58	13.0	8.74	13.0	12.74
9.0	8.81	9.0	8.77	9.0	7.11
11.0	8.33	11.0	9.26	11.0	7.81
14.0	9.96	14.0	8.10	14.0	8.84
6.0	7.24	6.0	6.13	6.0	6.08
4.0	4.26	4.0	3.10	4.0	5.39
12.0	10.84	12.0	9.13	12.0	8.15
7.0	4.82	7.0	7.26	7.0	6.42
5.0	5.68	5.0	4.74	5.0	5.73

x = Height on the sea shore (m)
y = Shell size (mm)

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

x = Height on the sea shore (m)
y = Shell size (mm)

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Anscombe's Quartet:

Regression Analysis: y1 versus x1

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	27.51	27.510	17.99	0.002
x1	1	27.51	27.510	17.99	0.002
Error	9	13.76	1.529		
Total	10	41.27			

Model Summary

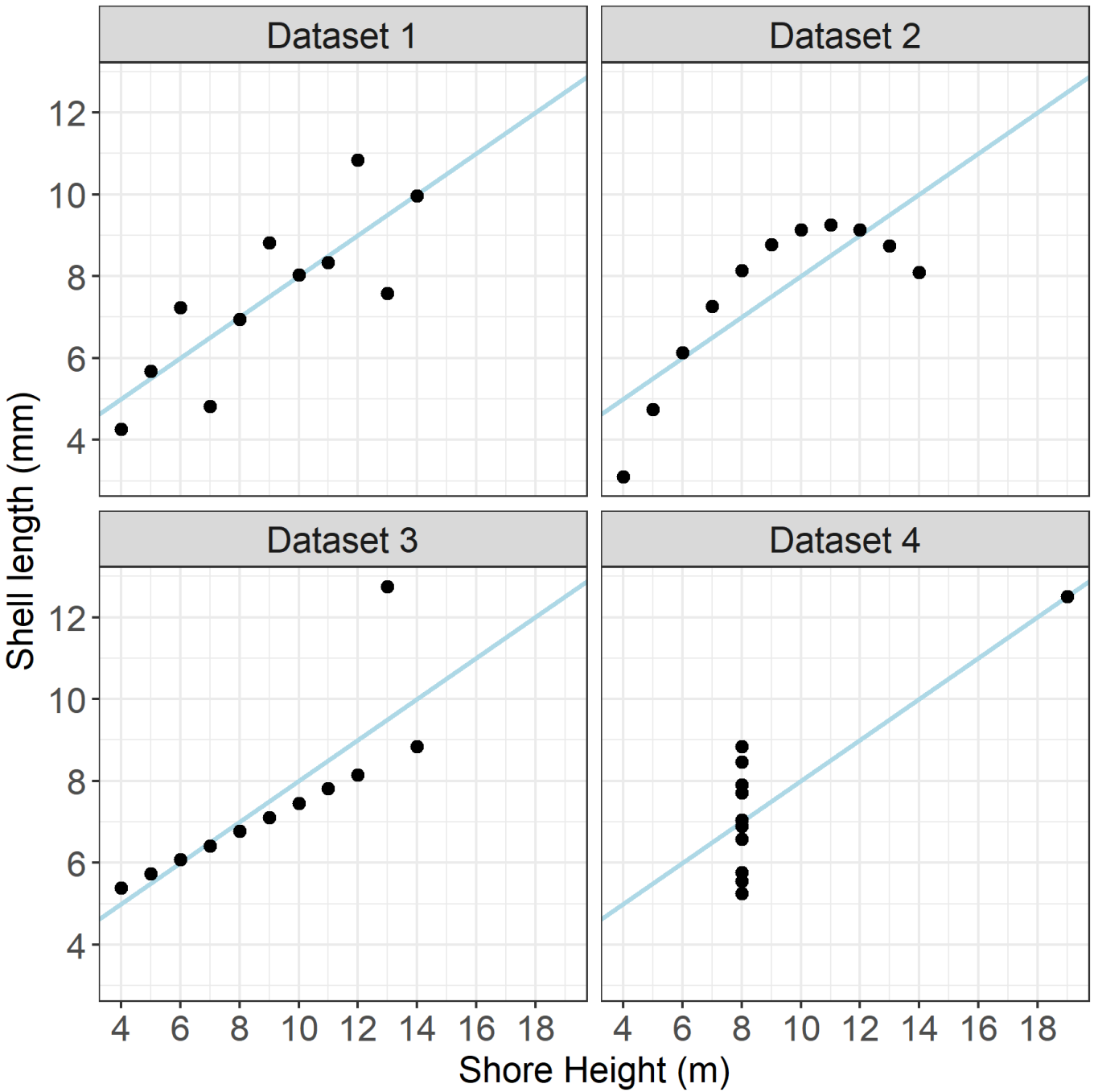
S	R-sq	R-sq(adj)	R-sq(pred)
1.23660	66.65%	62.95%	50.14%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3.00	1.12	2.67	0.026	
x1	0.500	0.118	4.24	0.002	1.00

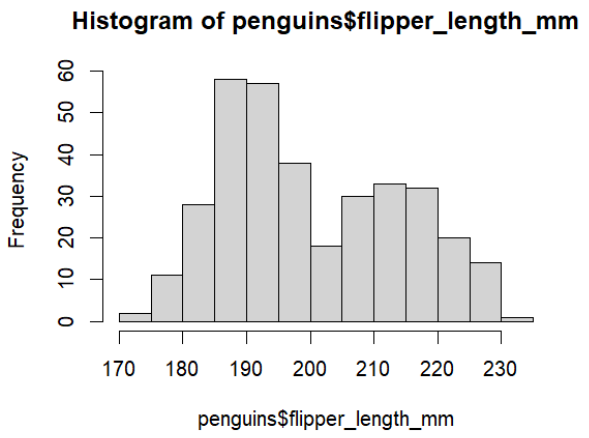
Regression Equation

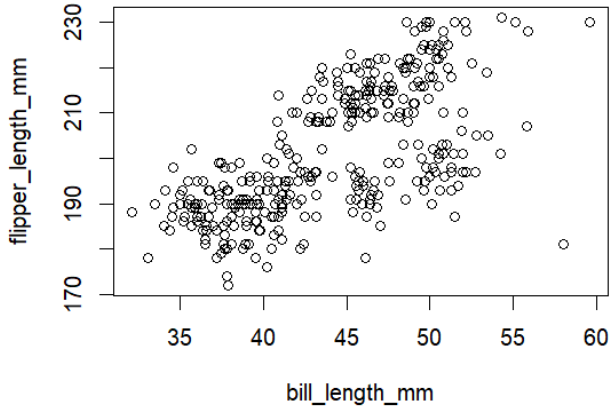
y1 = 3.00 + 0.500 x1

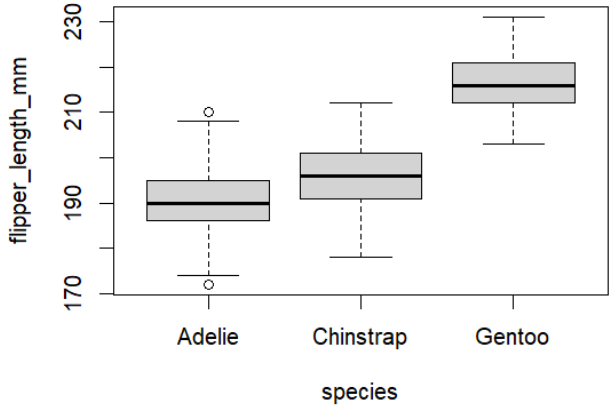

Dataset 1
Dataset 2
Dataset 3
Dataset 4
Shell length (mm)
Shore Height (m)
4
6
8
10
12
14
16
18

R has its own base graphics.

```
library(palmerpenguins)

# histogram
hist(penguins$flipper_length_mm)

# scatterplot
plot(flipper_length_mm ~ bill_length_mm,
     data = penguins)

# boxplot
boxplot(flipper_length_mm ~ species,
        data = penguins)
```







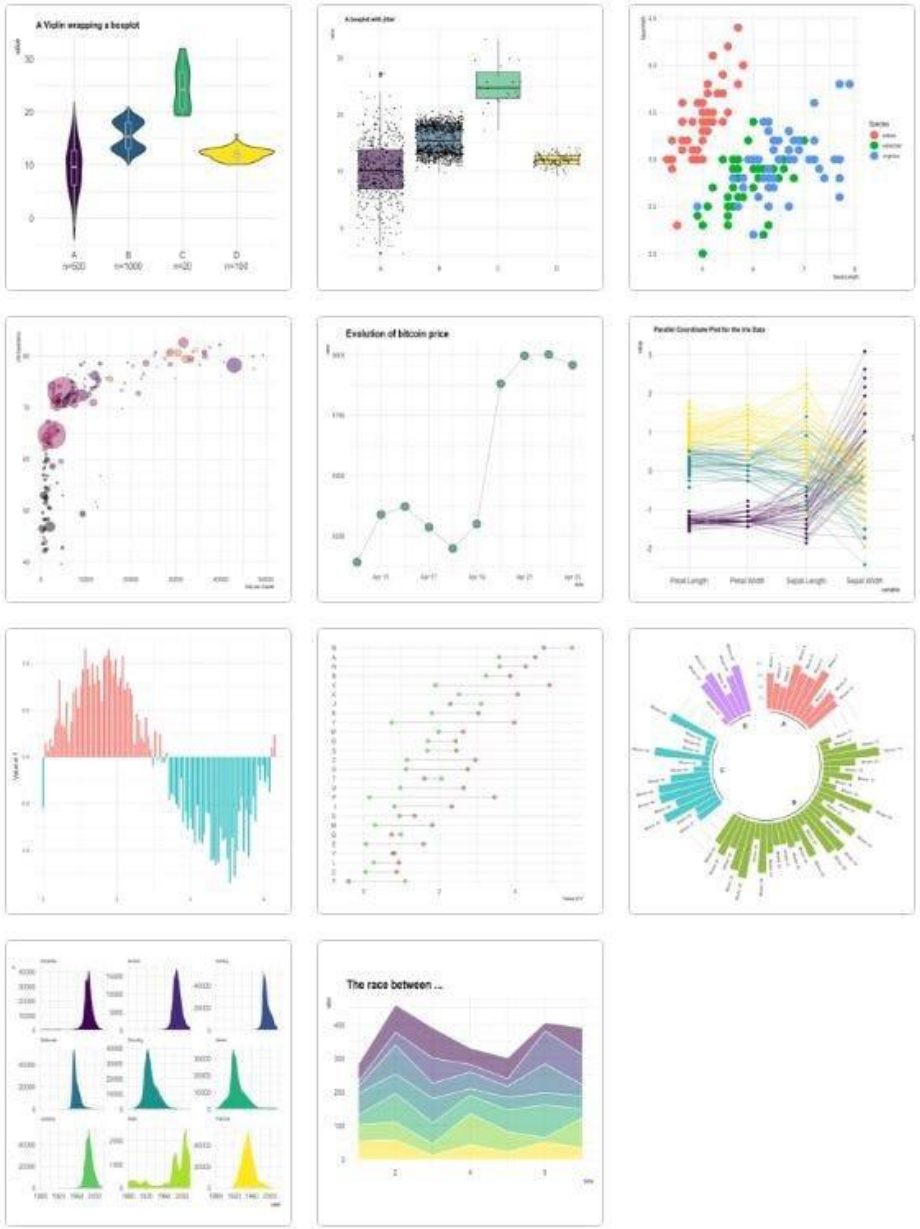
Tutorial at: http://rpubs.com/SusanEJohnston/7953

Data visualization with `ggplot2`

"You provide the data, tell ggplot2 how to map variables to ***aesthetics**** what [graph type] to use, and it takes care of the details."

**visual properties of your dataset*

https://ggplot2.tidyverse.org/index.html



Source: [R Graph Gallery]

`ggplot2` builds a graph with the following:

1. `ggplot()`
   - Data with aesthetic (visual) properties (`aes()`).

2. `geom_...()`
   - The type of plot (line, point, box-plot, etc.)

3. `stat_...()`
   - Statistical transformations (regression lines, smoothers, etc)*

4. `theme()`
   - How do you want your graph to look?

5. Other customisations
   - e.g. facets, scales, zoom, etc.