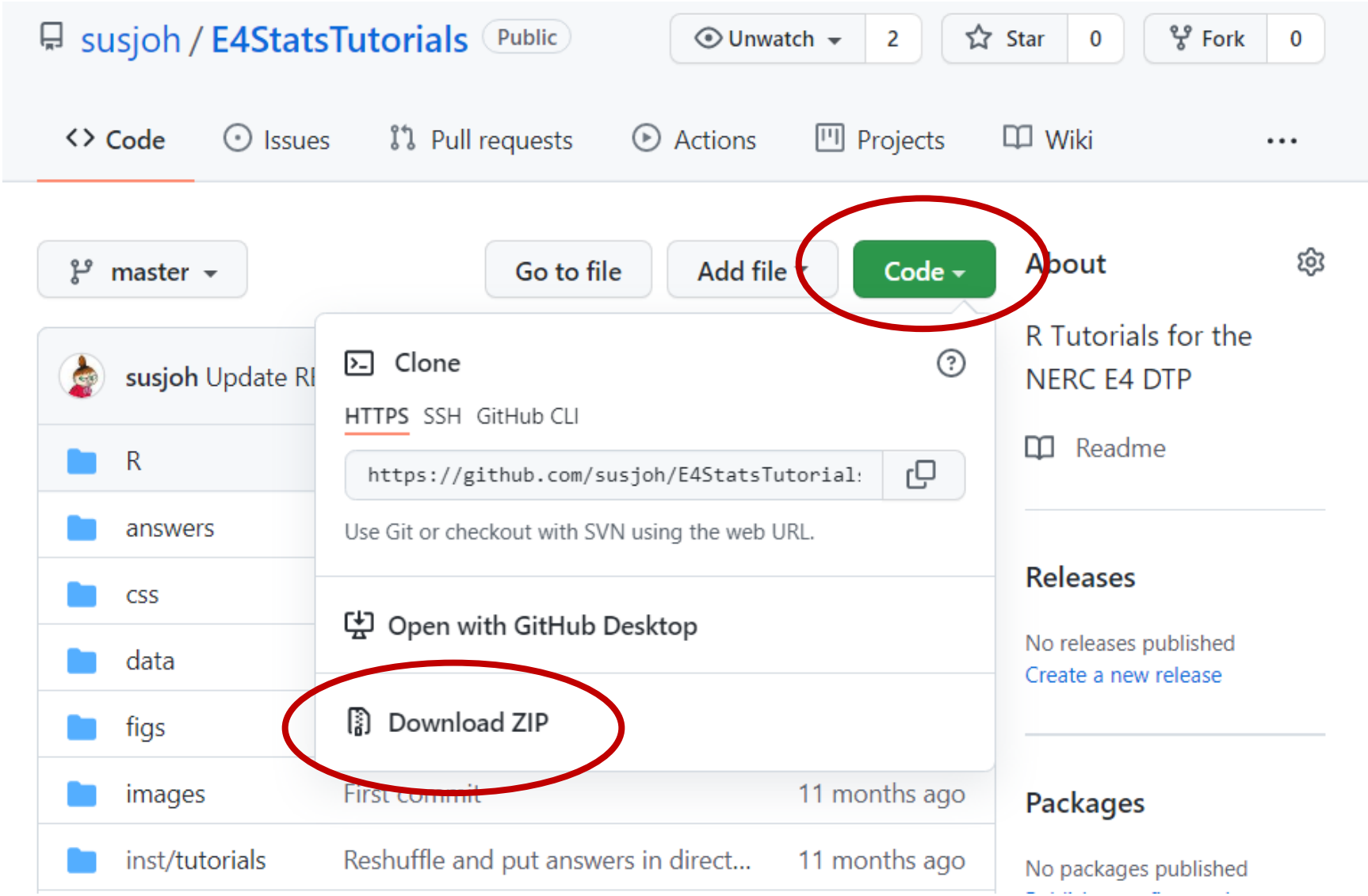
Using R as a Research Tool.

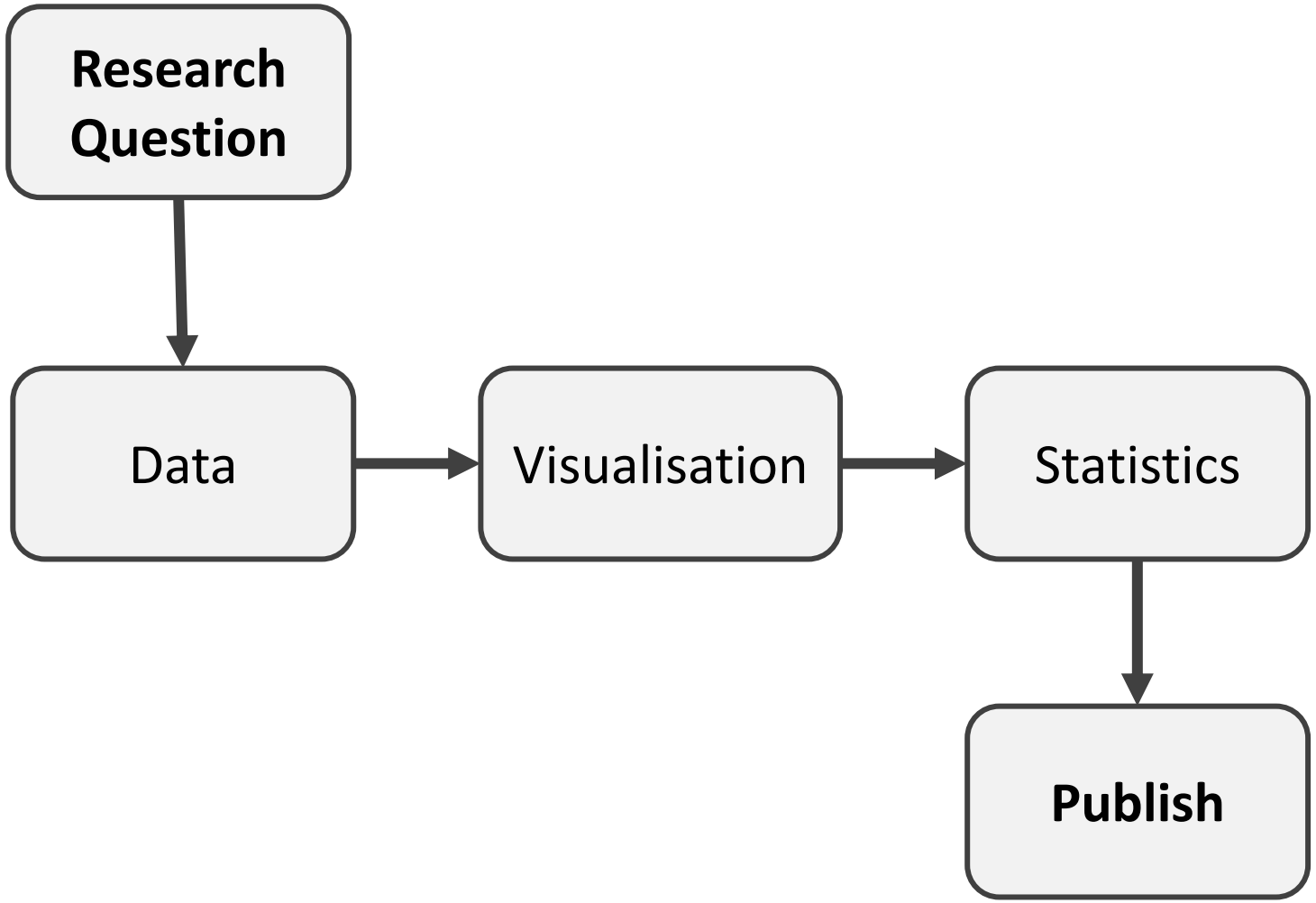
NERC E4 DTP Training

Dr Susan Johnston, Institute of Evolutionary Biology

github.com/susjoh/E4StatsTutorials



Using R as a Research Tool: Overview



What is  ?

- Environment for statistical computing and graphics.
- Interactive programming language.
- 16,454 packages on CRAN
- **Free and open-source** multi-platform software.

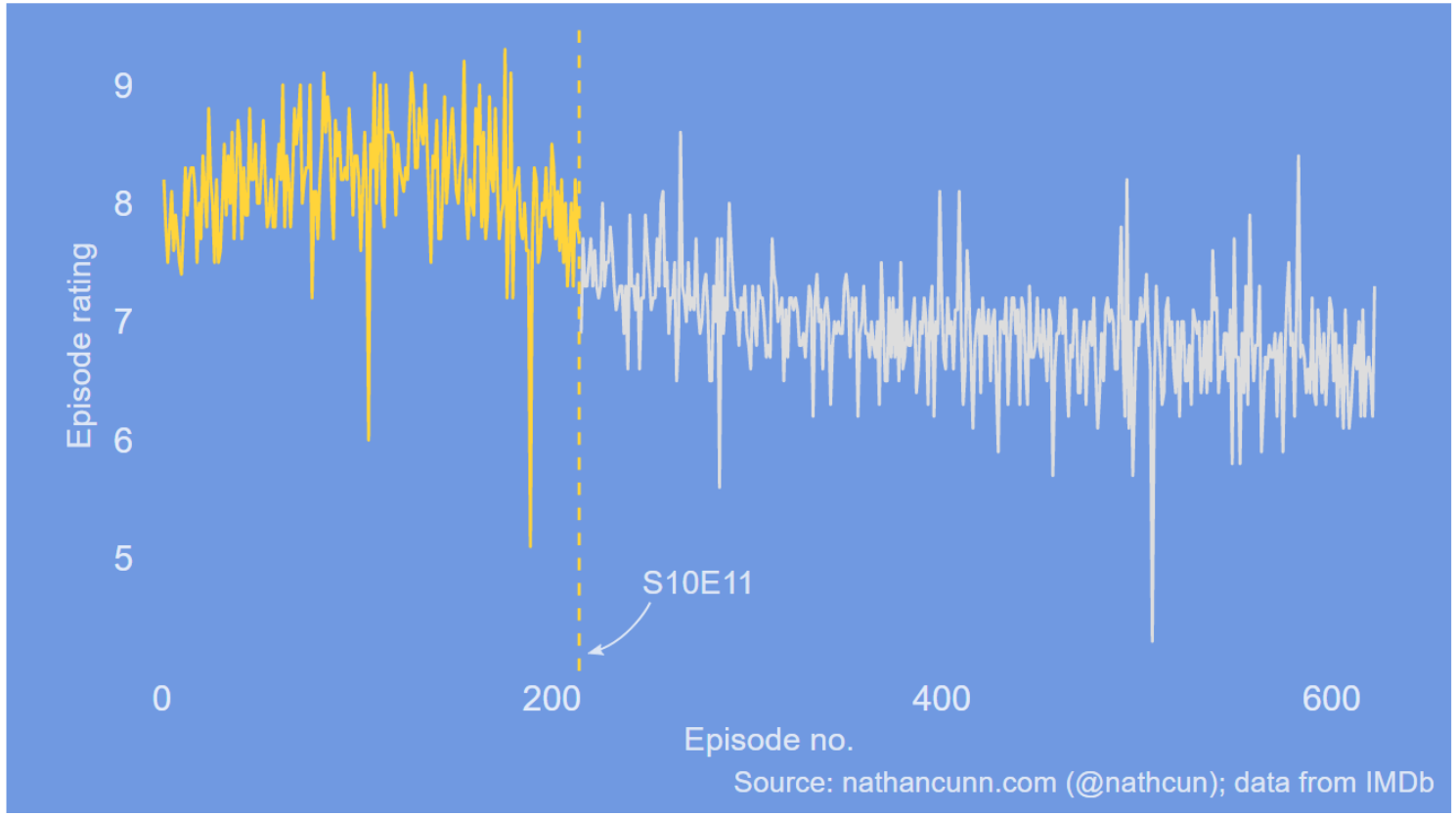
Why use  ?

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

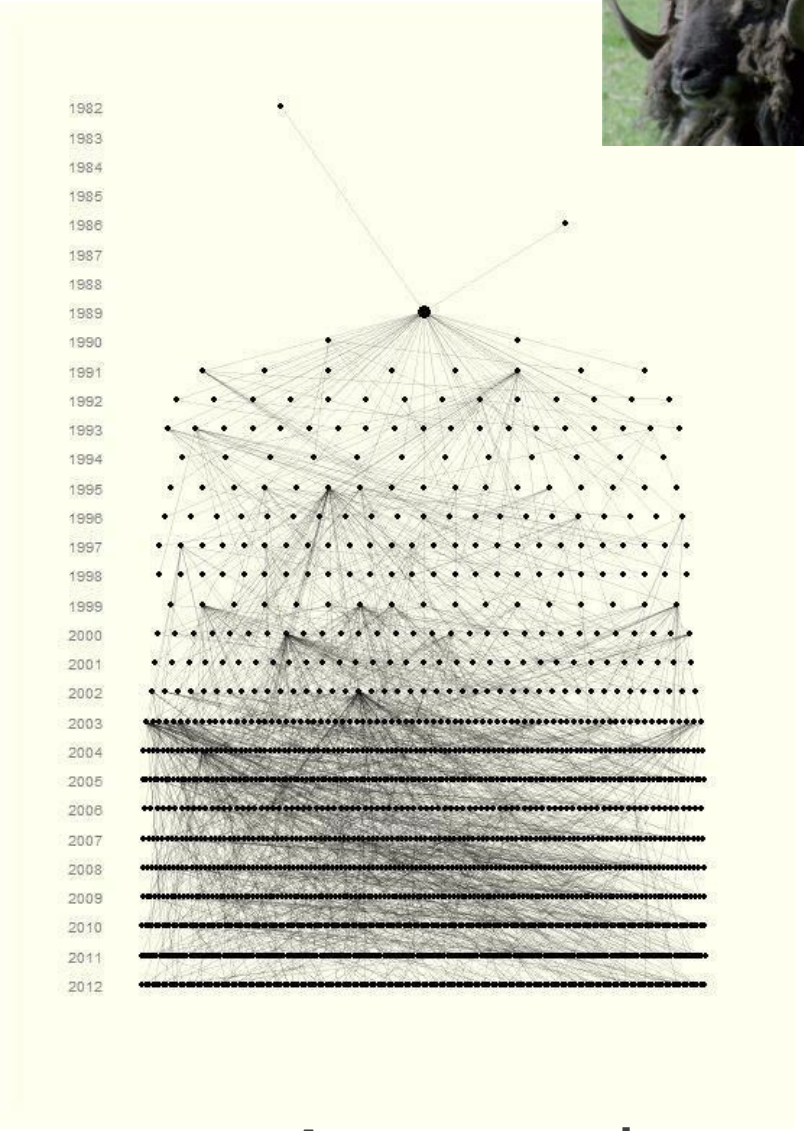
- Statistics.

Data visualisation

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



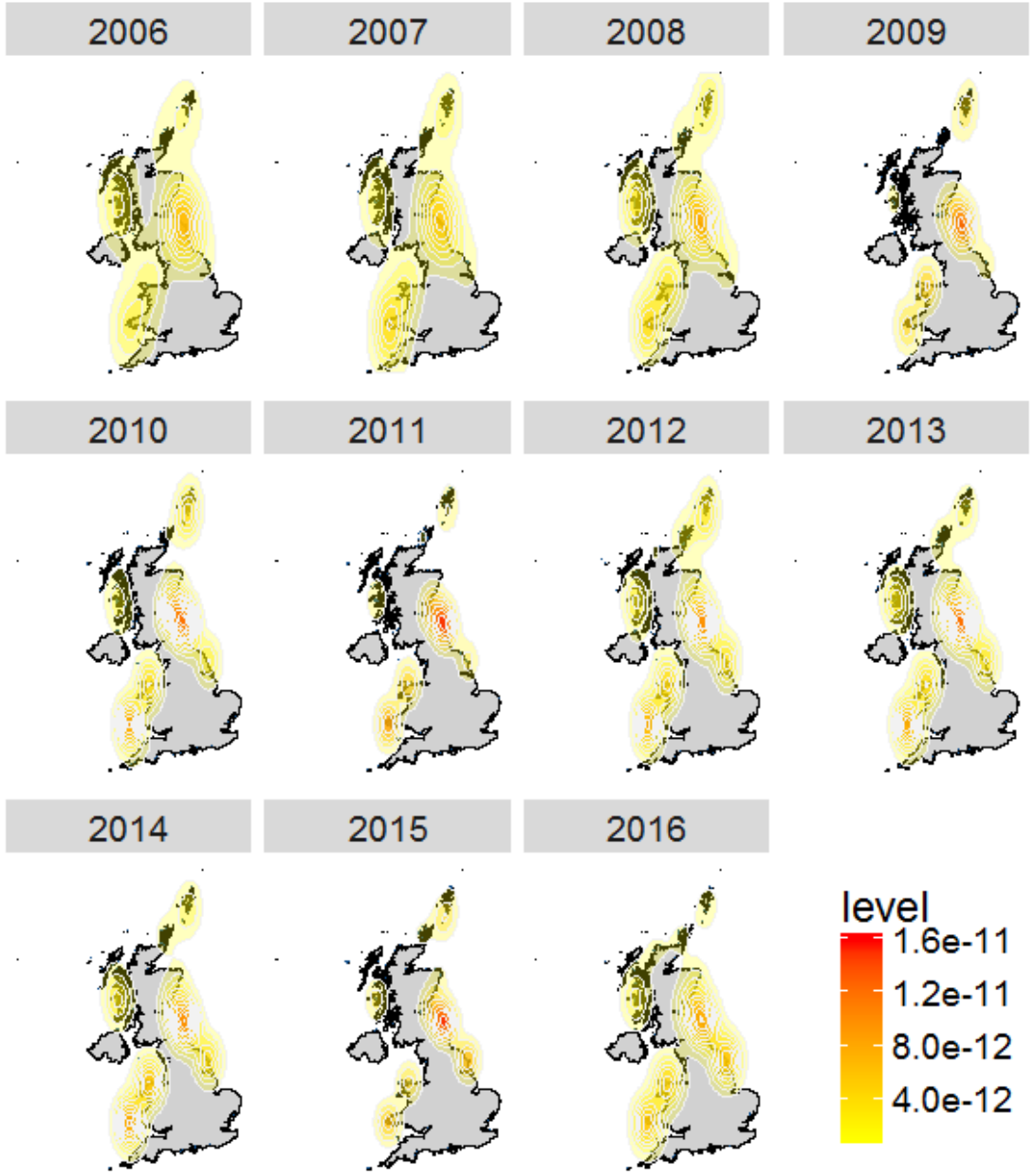
When did the golden age of The Simpsons end?



Ancestors and descendants of a single Soay sheep called Snowball.

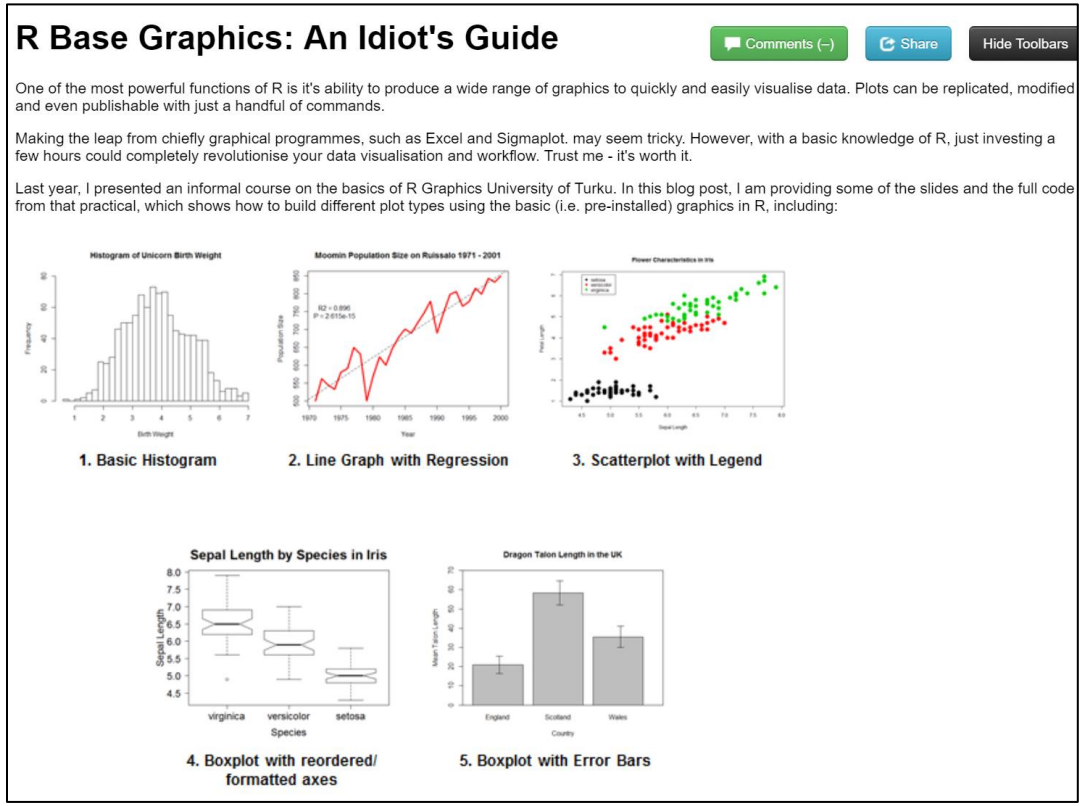
UK distribution of Atlantic Puffins

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

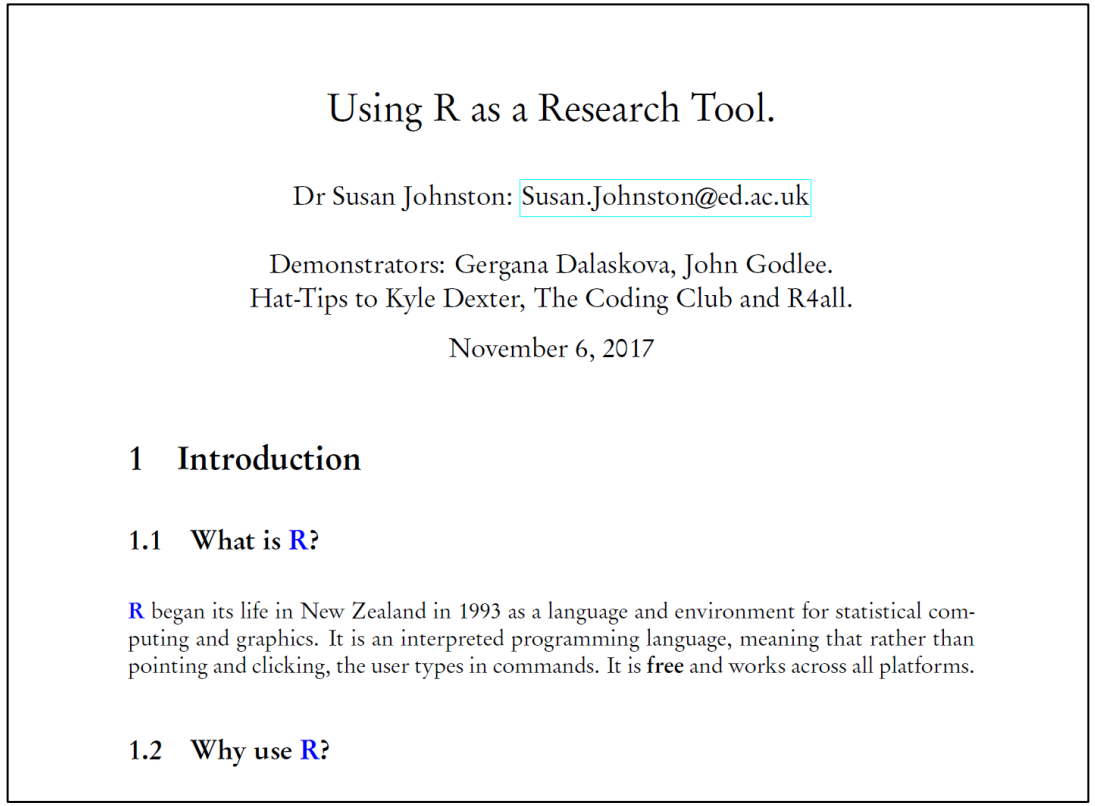
Comments (–) Share Hide Toolbars

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:

`knitr` to HTML



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?

LaTeX and R Sweave

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

Interactive applications (shiny)

National Records of Scotland

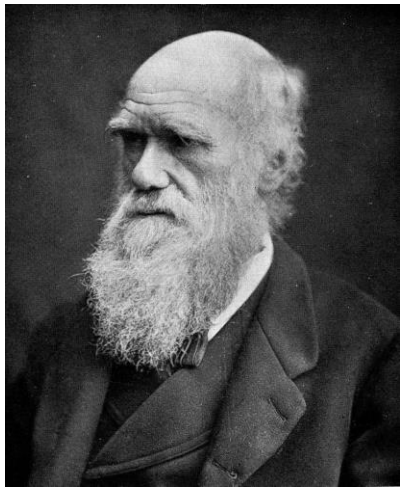
Baby names trends in Scotland since 1974

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.

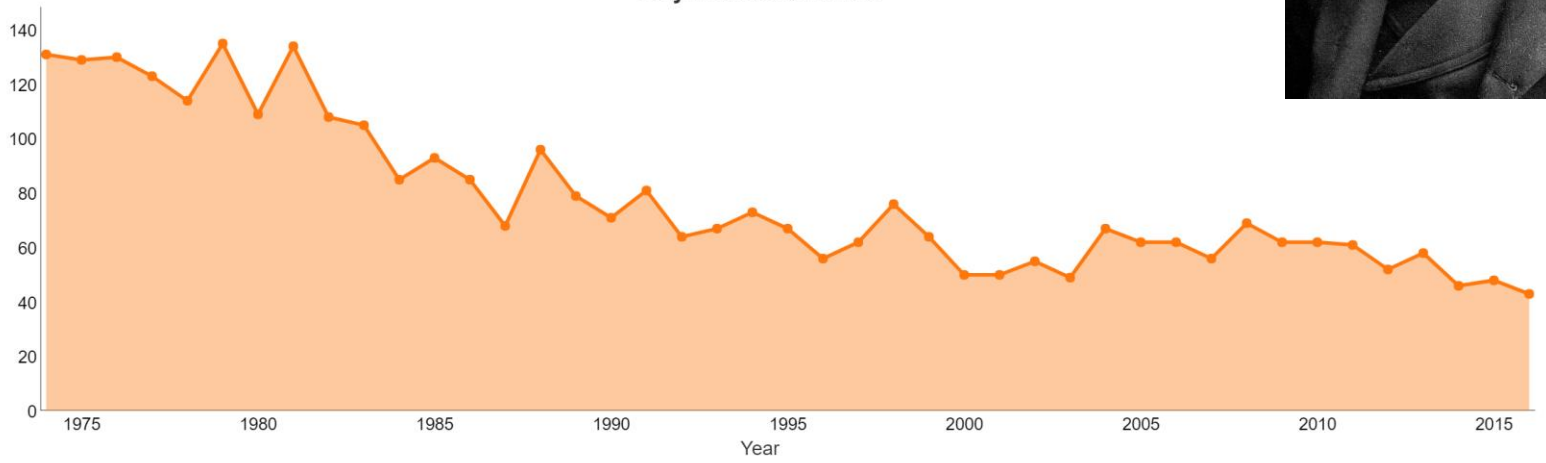
Name Charles

Gender Male

Apply



Boys named Charles

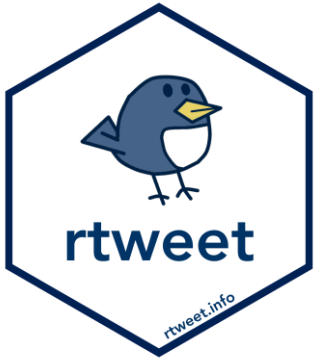


How to use this app
Hover over years to highlight individual values
Click and drag to zoom
Double-click to zoom out

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.

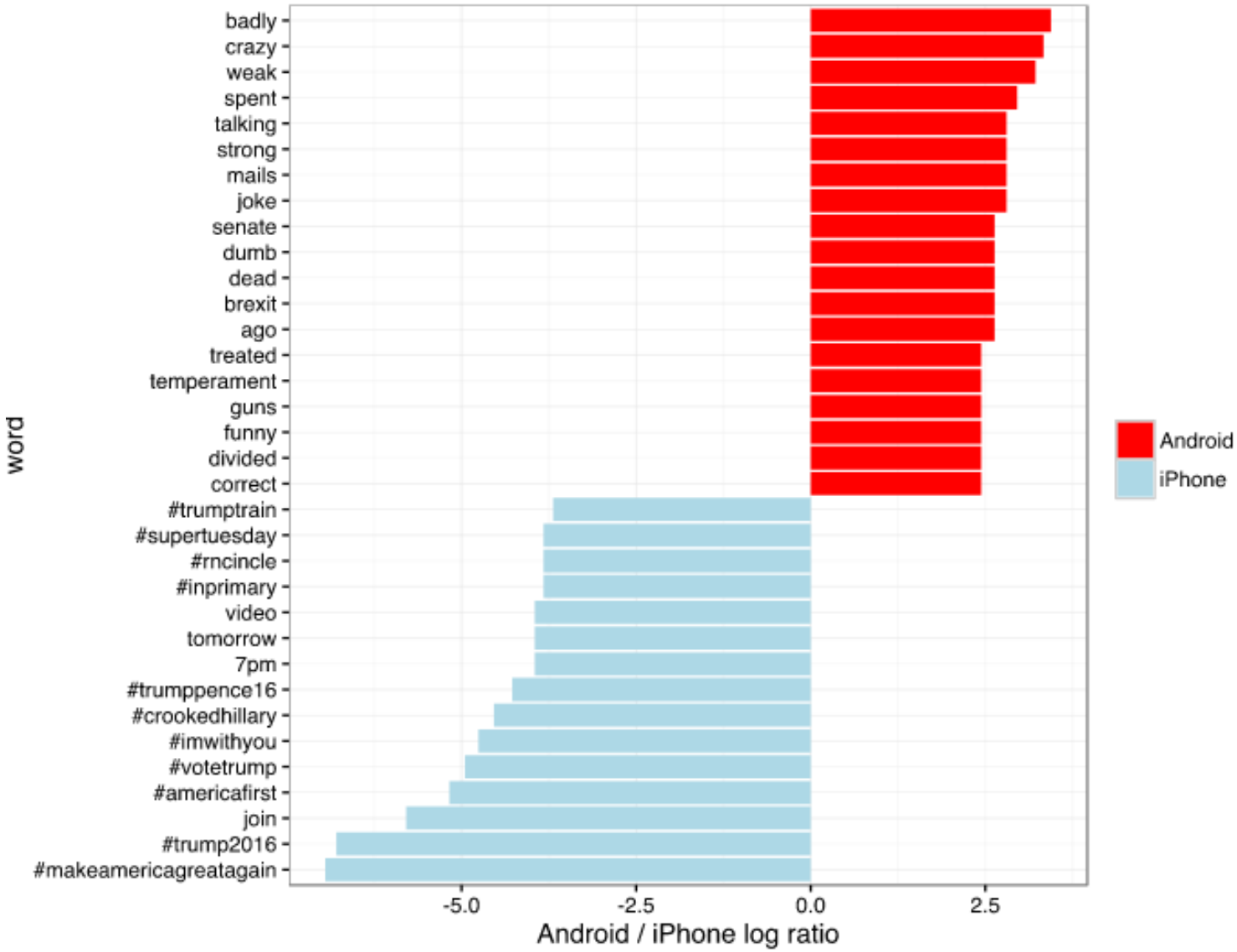


Todd Vaziri ✓
@tvaziri

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

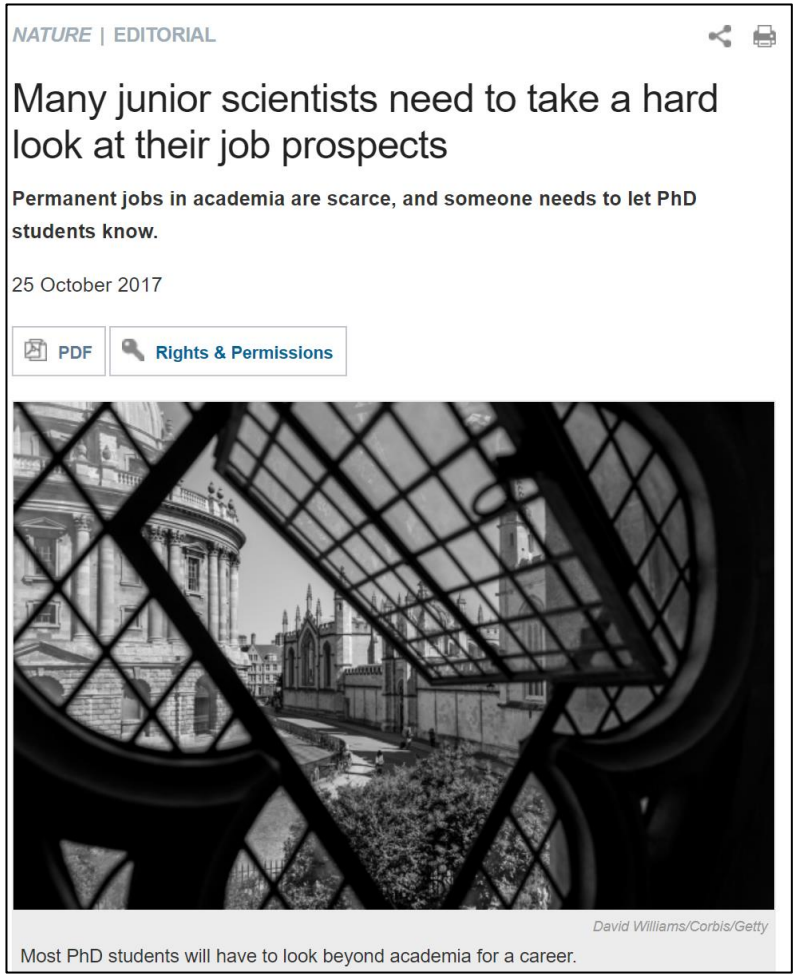
Every hyperbolic tweet is from Android (from him).

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.
- Edinburgh R jobs at Scottish Government, RBS, Tesco & Sainsburys Bank, Rockstar North, DataLab, 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

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

facilitates 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

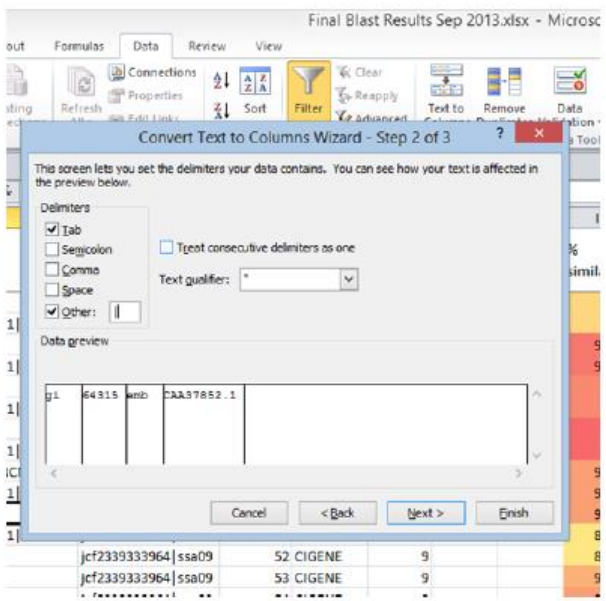
Although eukaryotes lack true operons, there are examples of operon-like gene clusters. Three examples are the galactose utilization genes in S. cerevisiae (GAL1, GAL10, GAL7), the allantoin degradation genes in S.c. (DAL1, DAL2, DAL3, DAL4, DAL7, DCG1), and the thalrendiol synthesis genes in Arabadopsis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

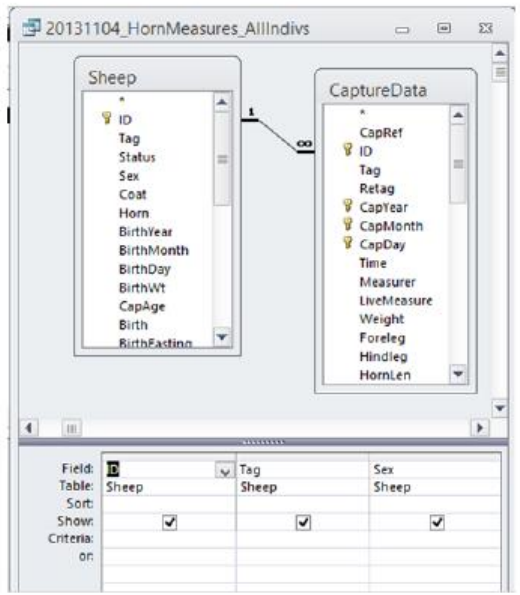
Two explanations have been given to account for this organization: genetic ~~clustering~~ linkage and metabolic channeling.

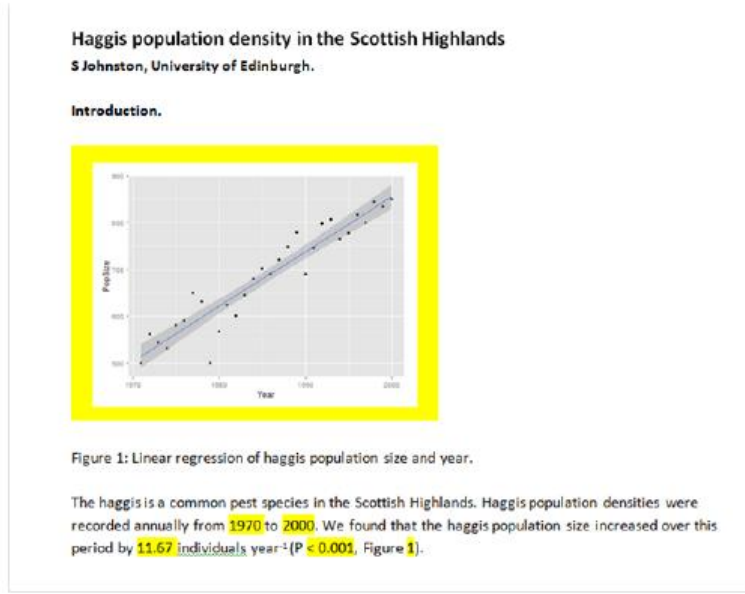
The genetic linkage hypothesis seems to be favored in the literature. It is interesting to note, however, that all three pathways above have intermediates that are toxic to the organism (in red). Here I want to test the hypothesis that the operon-like organization allows for better co-regulation of the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

1

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







Haggis population density in the Scottish Highlands

S Johnston, University of Edinburgh.

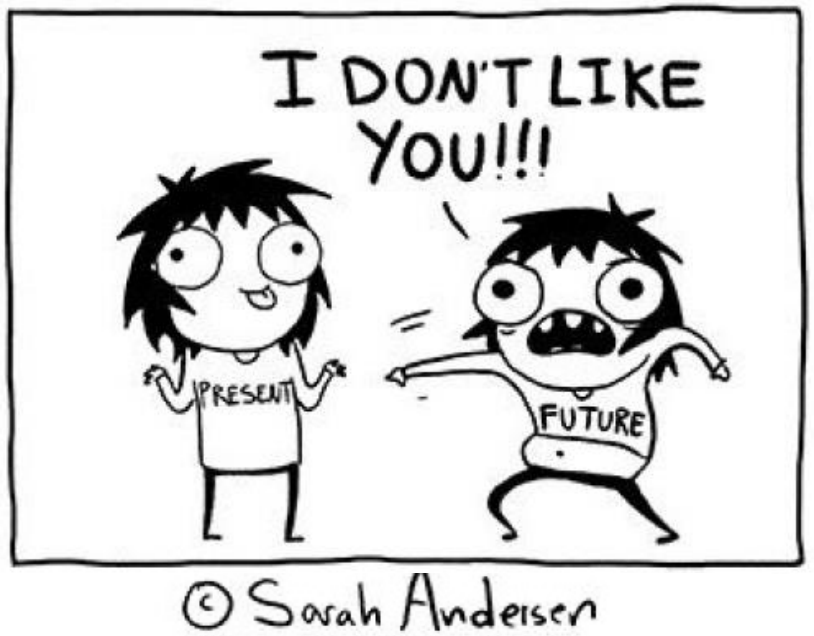
Introduction.

Figure 1: Linear regression of haggis population size and year.

The haggis is a common pest species in the Scottish Highlands. Haggis population densities were recorded annually from 1970 to 2000. We found that the haggis population size increased over this period by 11.67 individuals year^{-1} ($P < 0.001$, Figure 1).

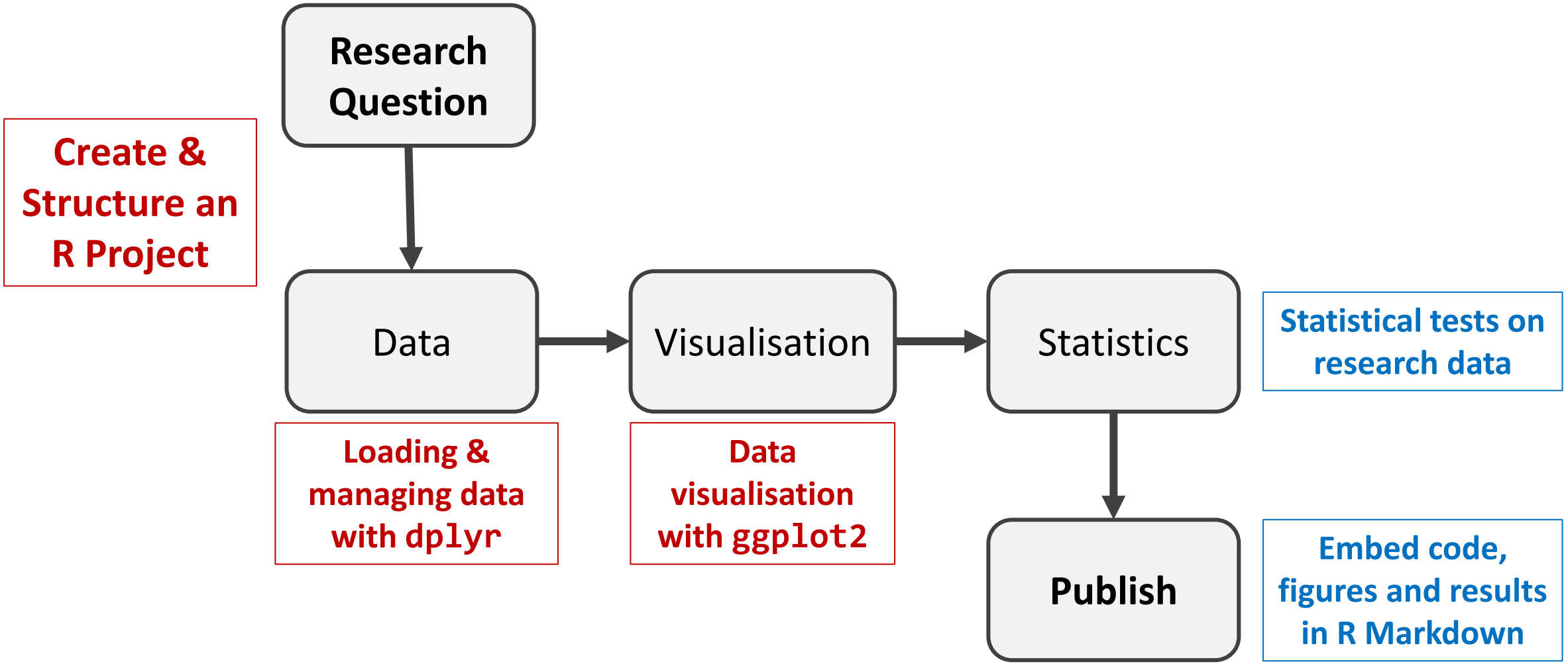
- Can you repeat all of this again...
- ...and would you get the same results every time?

Scenarios that benefit from reproducibility



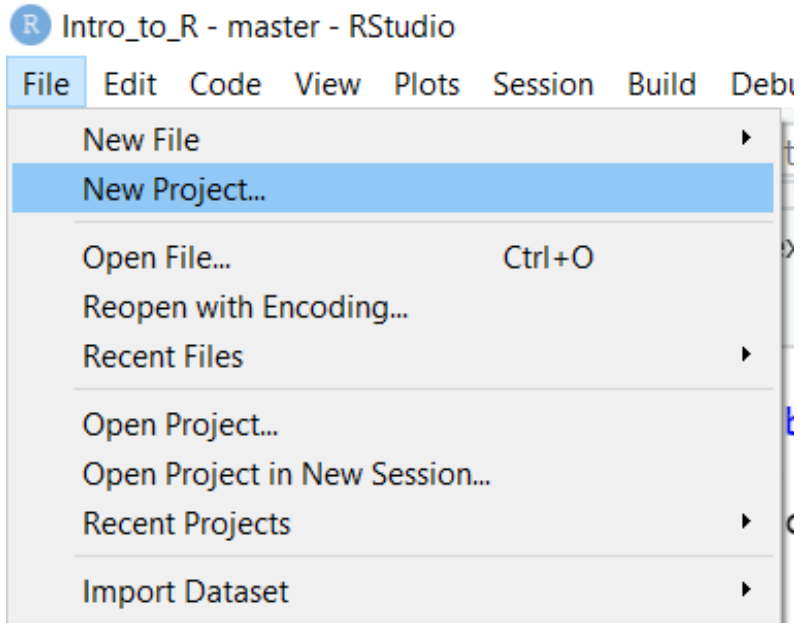
- The first researcher who will need to reproduce results is likely to be **YOU.**
- New data becomes available.
- You return to a project after a period of time.
- You give the project to a new student/collaborator.
- A reviewer wants you to change something.
- You found an error, but not sure where you went wrong.

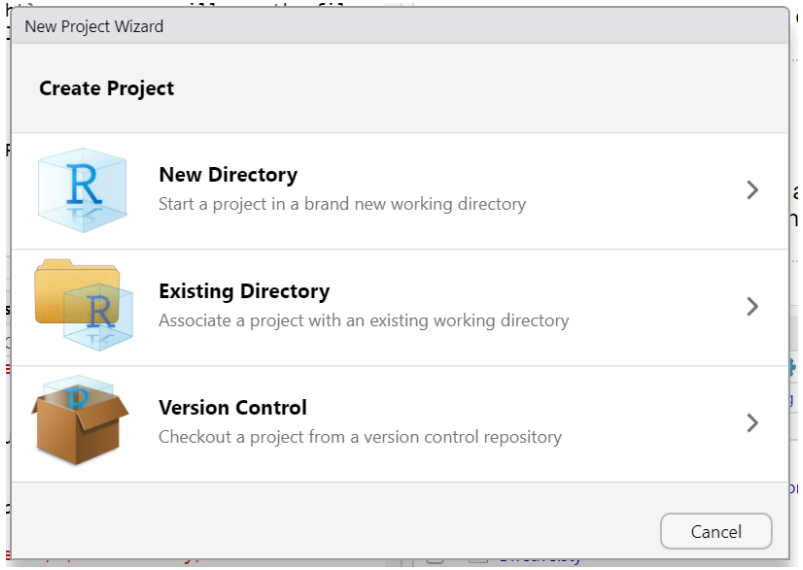
Using R as a Research Tool: Overview



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)

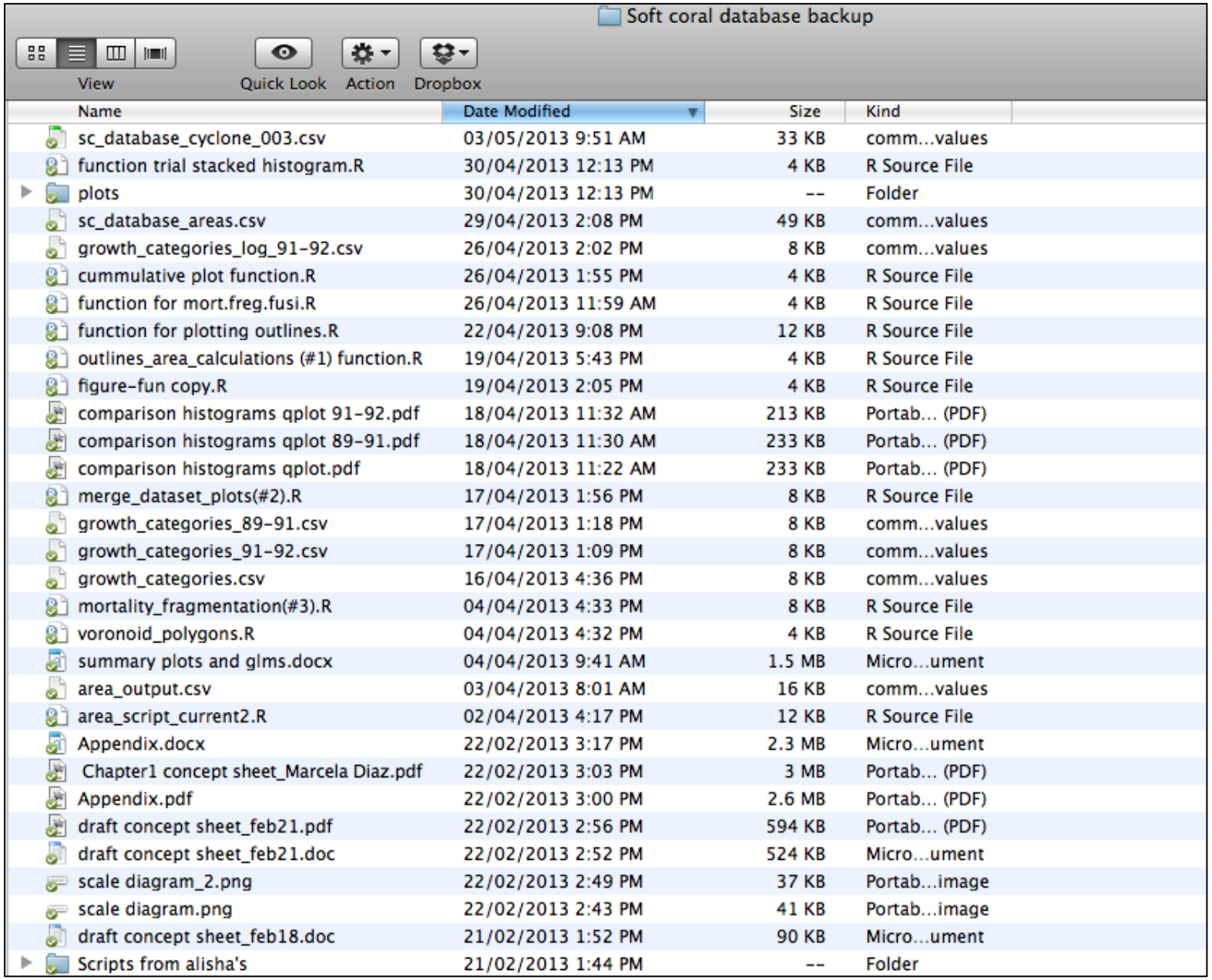




Structuring an R Project.

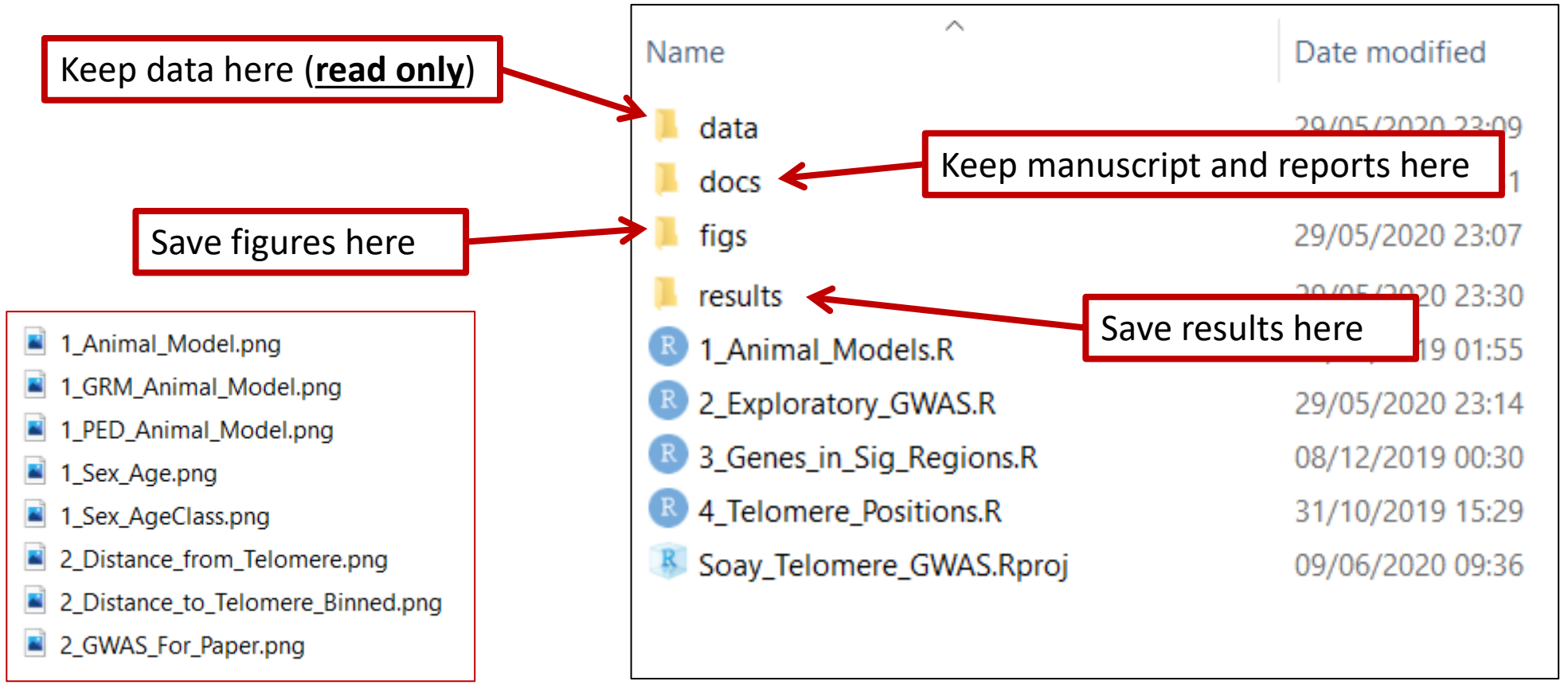
https://nicercode.github.io/blog/2013-05-17-organising-my-project/

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



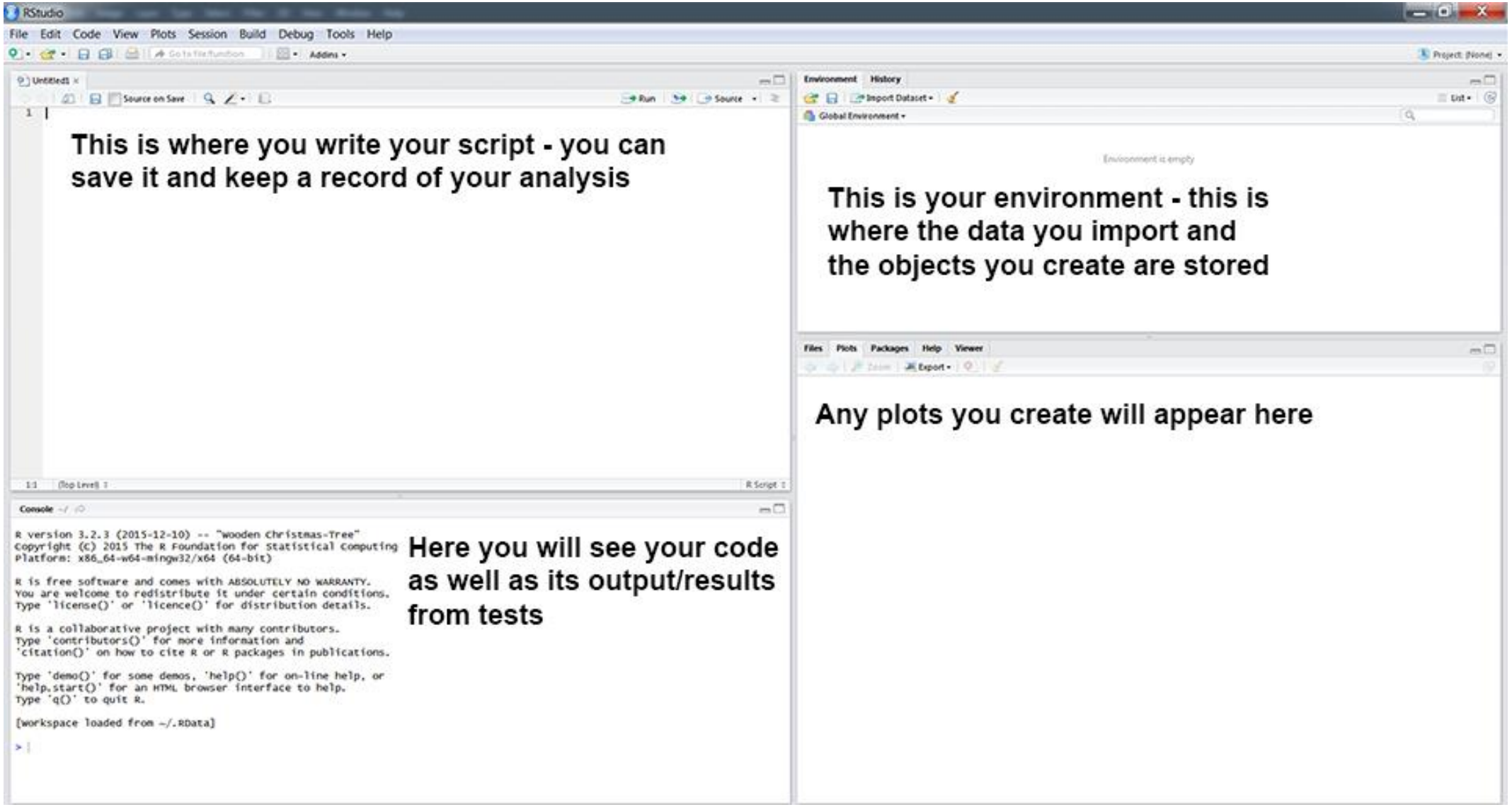
Name	Date Modified	Size	Kind
sc_database_cyclone_003.csv	03/05/2013 9:51 AM	33 KB	comm...values
function trial stacked histogram.R	30/04/2013 12:13 PM	4 KB	R Source File
plots	30/04/2013 12:13 PM	--	Folder
sc_database_areas.csv	29/04/2013 2:08 PM	49 KB	comm...values
growth_categories_log_91-92.csv	26/04/2013 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*).



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

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

http://ggplot2.tidyverse.org/reference/

Base graphics…
http://rpubs.com/SusanEJohnston/7953

`ggplot2` uses three components to construct a graph.

- Layers: `ggplot()`
  - Data with aesthetic properties (`aes()`)

- Geoms: `geom_...()`
  - Type of plot (line, scatter, box-plot, etc).

- Stats: `stat_...()`
  - Statistical transformations
  - NB. Most geoms have a default stat, so this is not always need.