MONASH University

MONASH
BIOINFORMATICS
PLATFORM

Introducing R

Intro to R Workshop
Trainers – Dr Paul Harrison & Adele Barugahare
Helpers – Dr Sarah Williams & Dr Nick Wong

@MonashBioinfo

Bioinformatics.platform@monash.edu

http://bioinformatics.erc.monash.edu

GROUP OF EIGHT AUSTRALIA

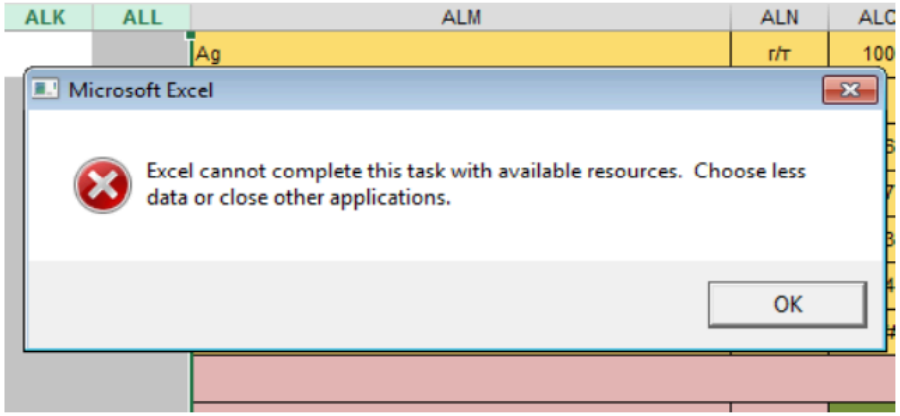
Today's Objective(s) – What we hope you will get out of today

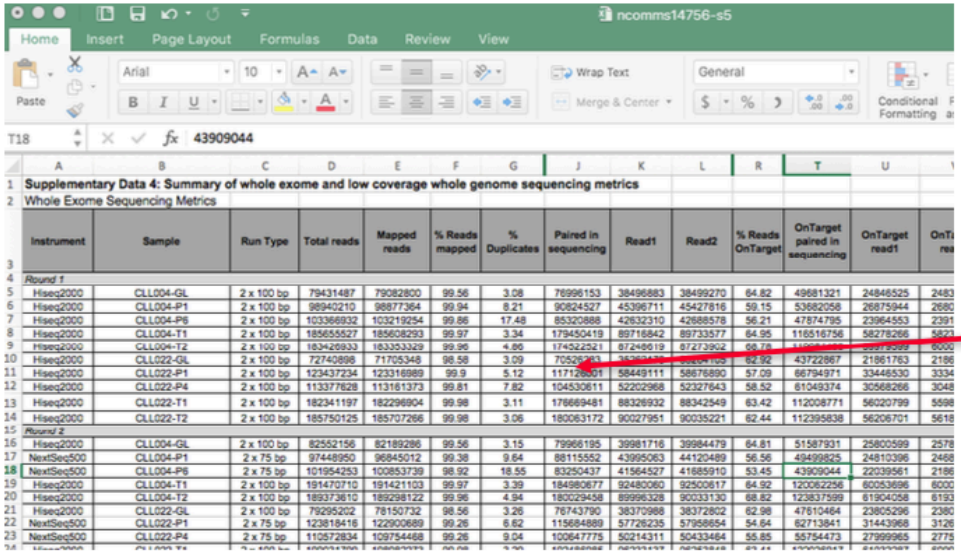
- Introduction to R
- The basics of R – data objects
- Data plotting
- How to find more information for a specific task
- The start of the journey, a network of R learners/users around you.
- Entry into more advanced and specialized uses of R for your research.

Why R? (as opposed to Excel?)

- It's free, legally free. Active development.
- Reproducible analysis.
  - Document what you have done with your data in code.
  - Come back to it in days, months, years and you should know what was done.
- Low risk of inadvertent data loss/mutation.
  - By design, R requires you to load your data in, what you do with the data is then written in code (R language).
- R can handle really large datasets.
  - Excel is limited by 1,048,576 rows and 16,384 columns.
- Collaborative.
  - Share your data and analysis.

Why R? (as opposed to Excel?)





Typo and saved it

Ziemann *et al. Genome Biology* (2016) 17:177
DOI 10.1186/s13059-016-1044-7

Genome Biology

COMMENT Open Access

Gene name errors are widespread in the scientific literature

Mark Ziemann[1], Yotam Eren[1,2] and Assam El-Osta[1,3*]

Abstract

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

Keywords: Microsoft Excel, Gene symbol, Supplementary data

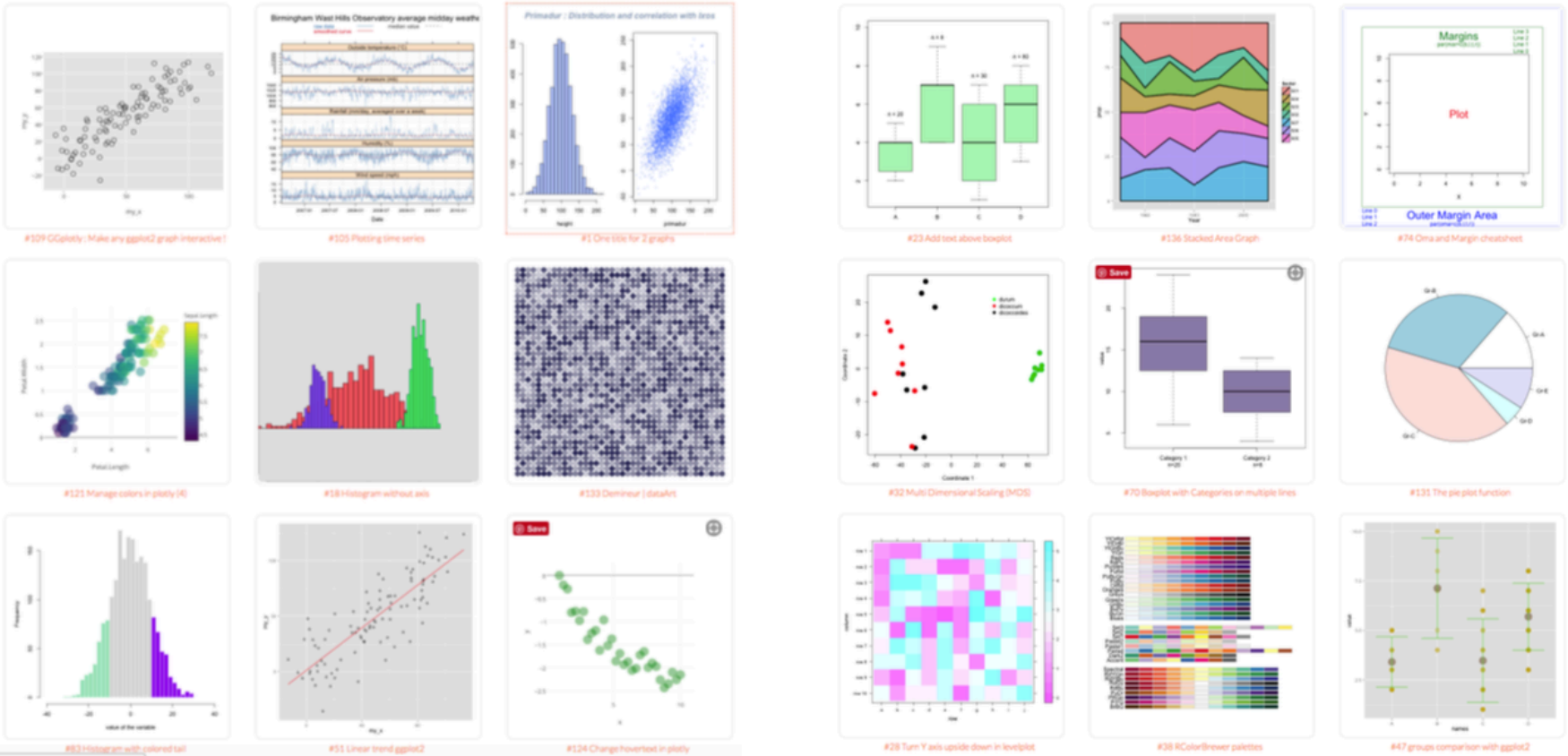
Abbreviations: GEO, Gene Expression Omnibus; JIF, journal impact factor

frequently reused. Our aim here is to raise awareness of the problem.

We downloaded and screened supplementary files from 18 journals published between 2005 and 2015 using a suite of shell scripts. Excel files (.xls and.xlsx suffixes) were converted to tabular separated files (tsv) with ssconvert (v1.12.9). Each sheet within the Excel file was converted to a separate tsv file. Each column of data in the tsv file was screened for the presence of gene symbols. If the first 20 rows of a column contained five or more gene symbols, then it was suspected to be a list of gene symbols, and then a regular expression (regex) search of the entire column was applied to identify gene symbol errors. Official gene symbols from Ensembl version 82, accessed November 2015, were obtained for

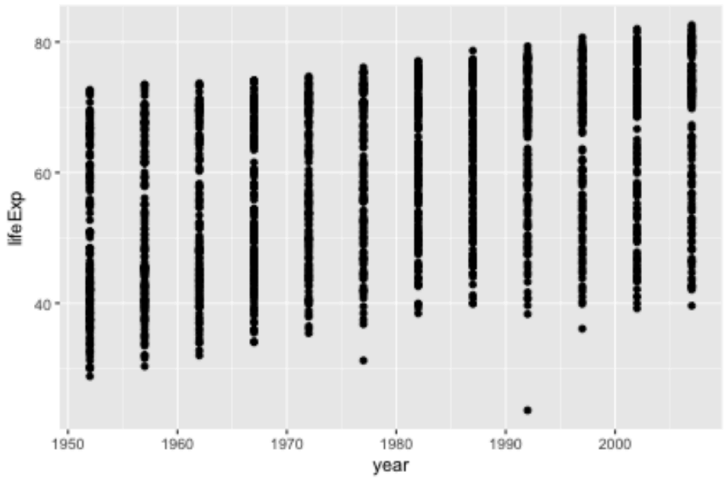
Why R? (as opposed to Excel?)

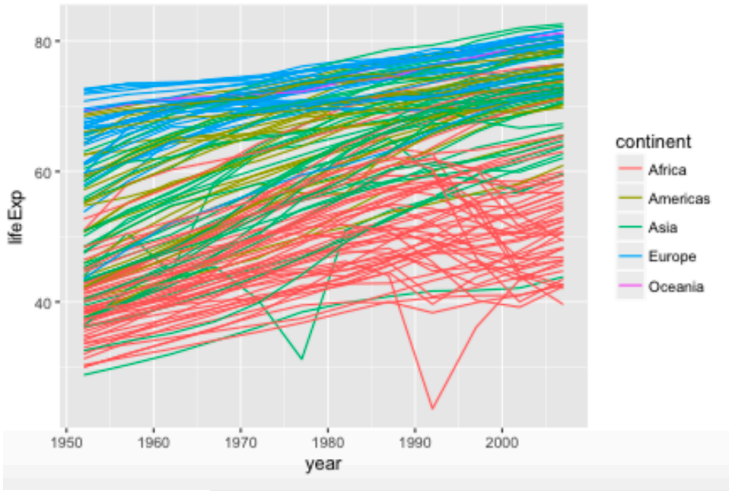
- Graphs and plotting.
- Can you plot a box and whisker plot in excel?

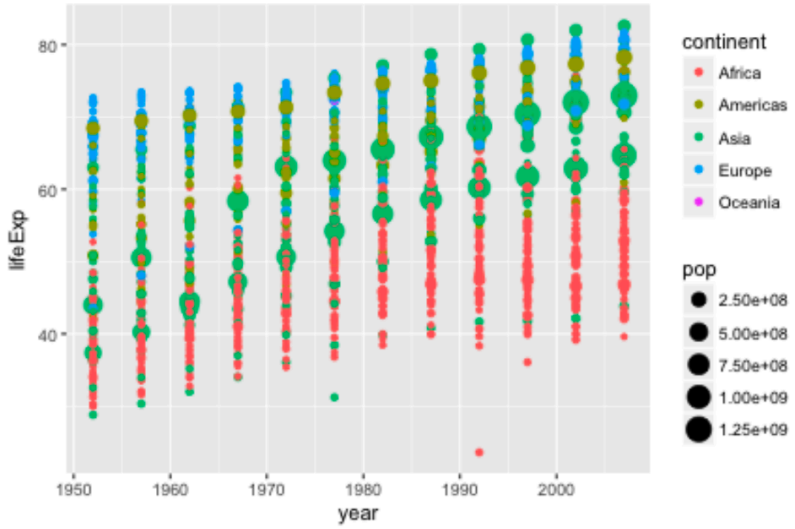


@MonashBioinfo
Bioinformatics.Platform@monash.edu
http://bioinformatics.erc.monash.edu

A better way to explore, present and interpret your data….







Learn to encode data
In a graph

How to get R?

CRAN
Mirrors
What's new?
Task Views

The Comprehensive R Archive Network

Download and Install R

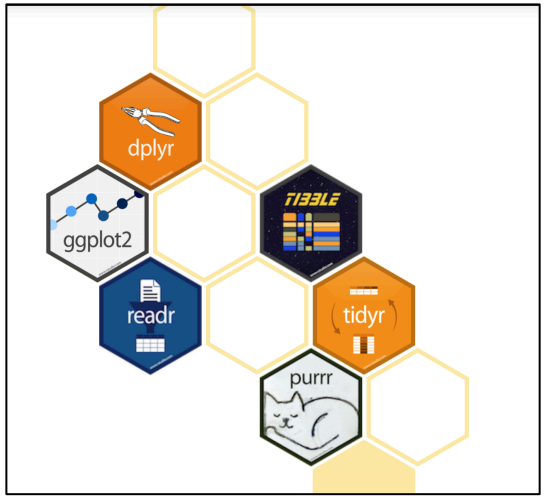
Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

https://cran.r-project.org)

Many specialized tools/libraries for specific purposes

Tidyverse: Data Science



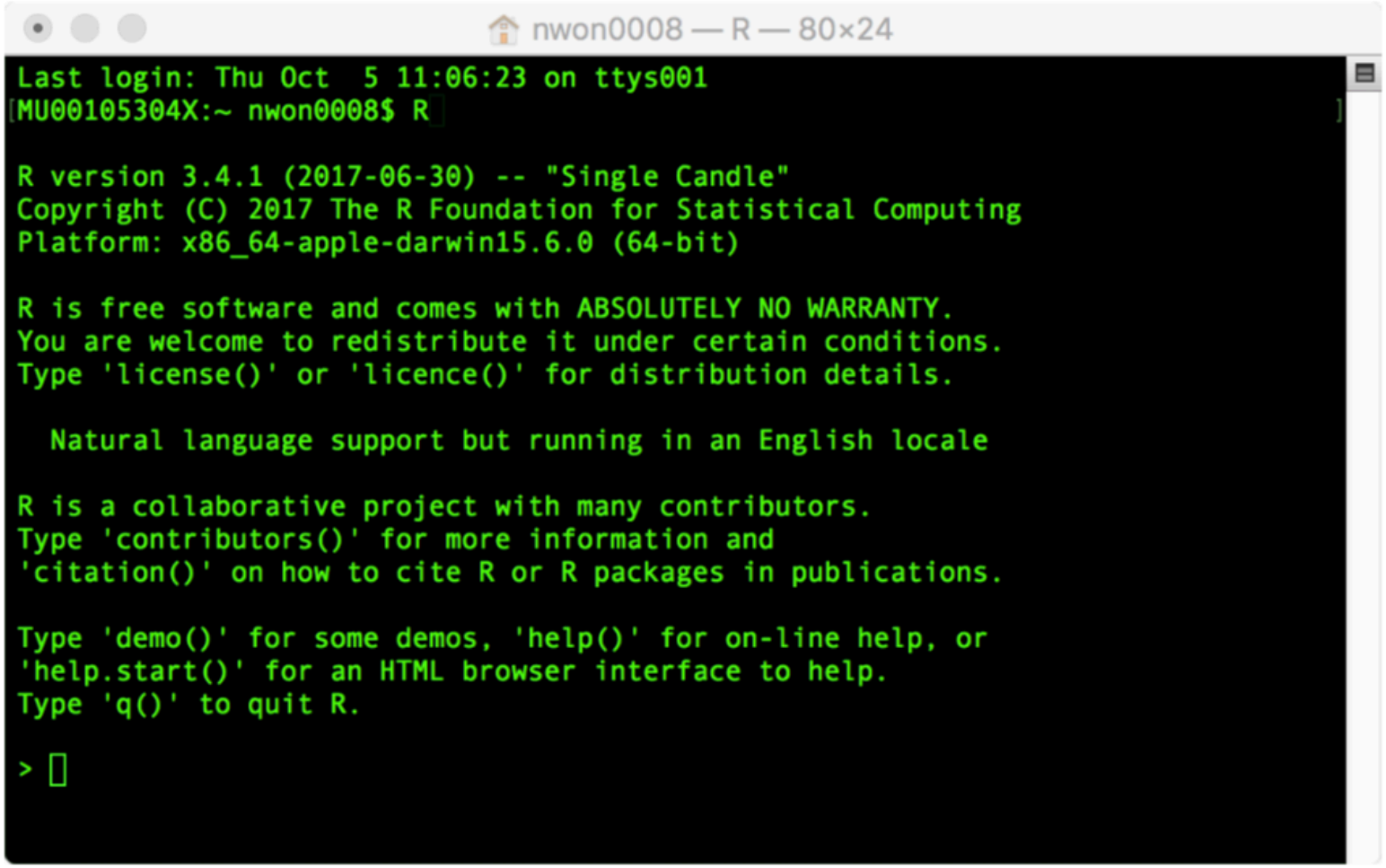
Rmarkdown: Documents

Bioconductor: Bioinformatics

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

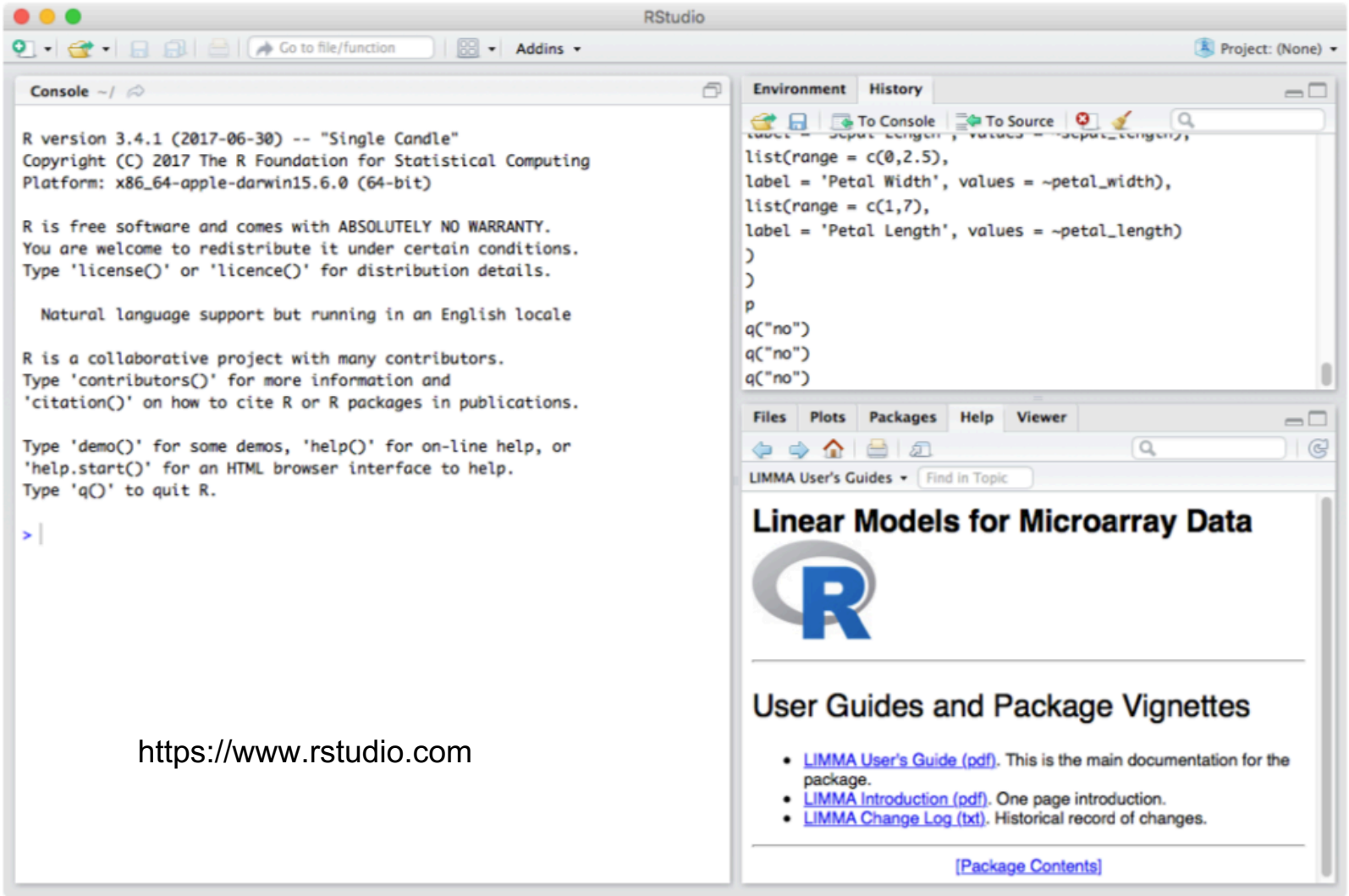
R and the interfaces to R

- R on the command line



R and the interfaces to R

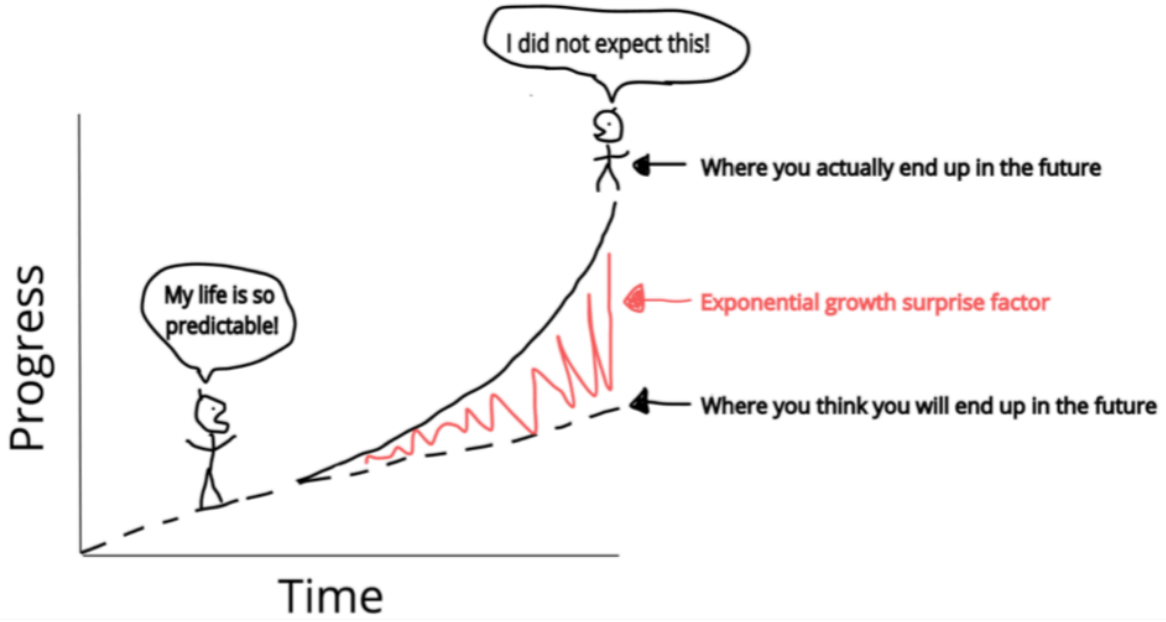
- RStudio – RStudio Server, a graphical interface to R



https://www.rstudio.com

Today's agenda

- Starting out in R
- Working with tidy data frames
- Plotting with ggplot2



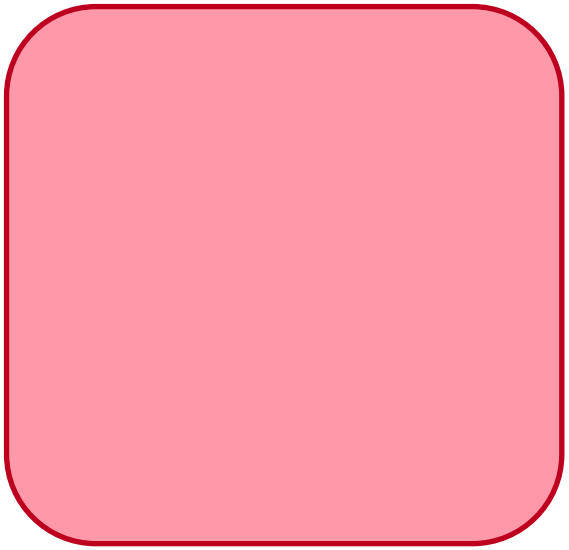
What we won't be covering but are important in the interest of time.

- Things we won't be covering but are of note:
  - –Base R
  - –Writing your own functions & loops
  - –Statistical modelling (lm, glm……).

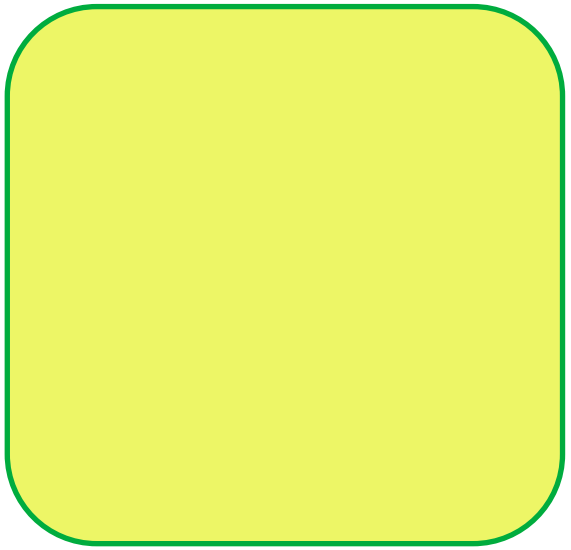
- Advanced R workshops

How today's workshop works – signalling for help

- Workshop notes
- Etherpad
- Sticky notes



I need help please



I am all good

The R workbook

Explanatory text

Under the hood, a data frame is a list of column vectors. This is why `mode` told us `gap` was a list. This means we can use `$` to retrieve columns. (Occasionally it is also useful to use `[[ ]]` to retrieve columns, for example if the column name we want is stored in a variable.)

```
head( gap$lifeExp )
```

```
## [1] 28.801 30.332 31.997 34.020 36.088 38.438
```

```
head( gap[["lifeExp"]] )
```

The code

```
## [1] 28.801 30.332 31.997 34.020 36.088 38.438
```

So to get just the lifeExp value of the third row as above, but unwrapped, we can use:

The R workbook – Challenges and tips

Challenge

What is the average gdpPercap for each continent in 2007?

Advanced: Use `weighted.mean` to calculate this correctly.

We will give you challenges and tips through the workshop to work with the example data

Tip

A data frame has column names (`colnames`), and base R data frames can also have row names (`rownames`). However the modern convention, which the Tidyverse enforces, is for a data frame to use column names but not row names. Typically a data frame contains a collection of items (rows), each having various properties (columns). If an item has an identifier such as a unique name, this would be given as just another column.

Rounding up

- This is the start of your R journey, many others are at the same stage, share your questions.
- Attend this workshop again, and others we offer
  - Advanced R
  - Specialised use-cases with R (eg: RNASeq)
- Friday help sessions at Clayton 3:00pm.

Basic concepts of R and take home messages

- R is a programming language and your code can be recorded in an R script file.

▪Variable	
▪Assignment operator	**<-**
▪Functions	**functionName()**
▪Subset	**[x , y]**
▪Run code from R script	**command/control <enter>**
▪Swap between panes	**control 1/2**