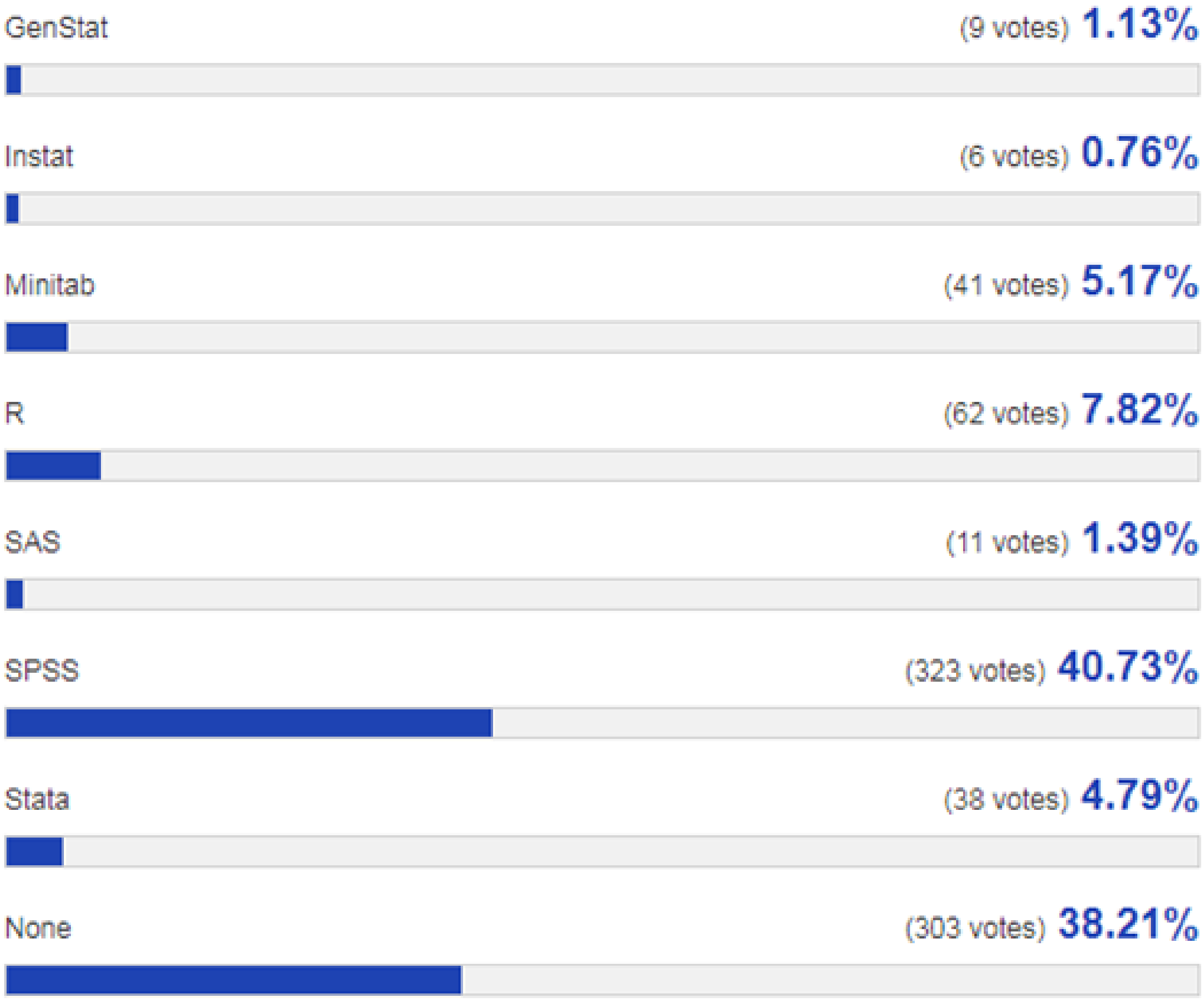
# Workshop 1:
# The tidyverse and beyond

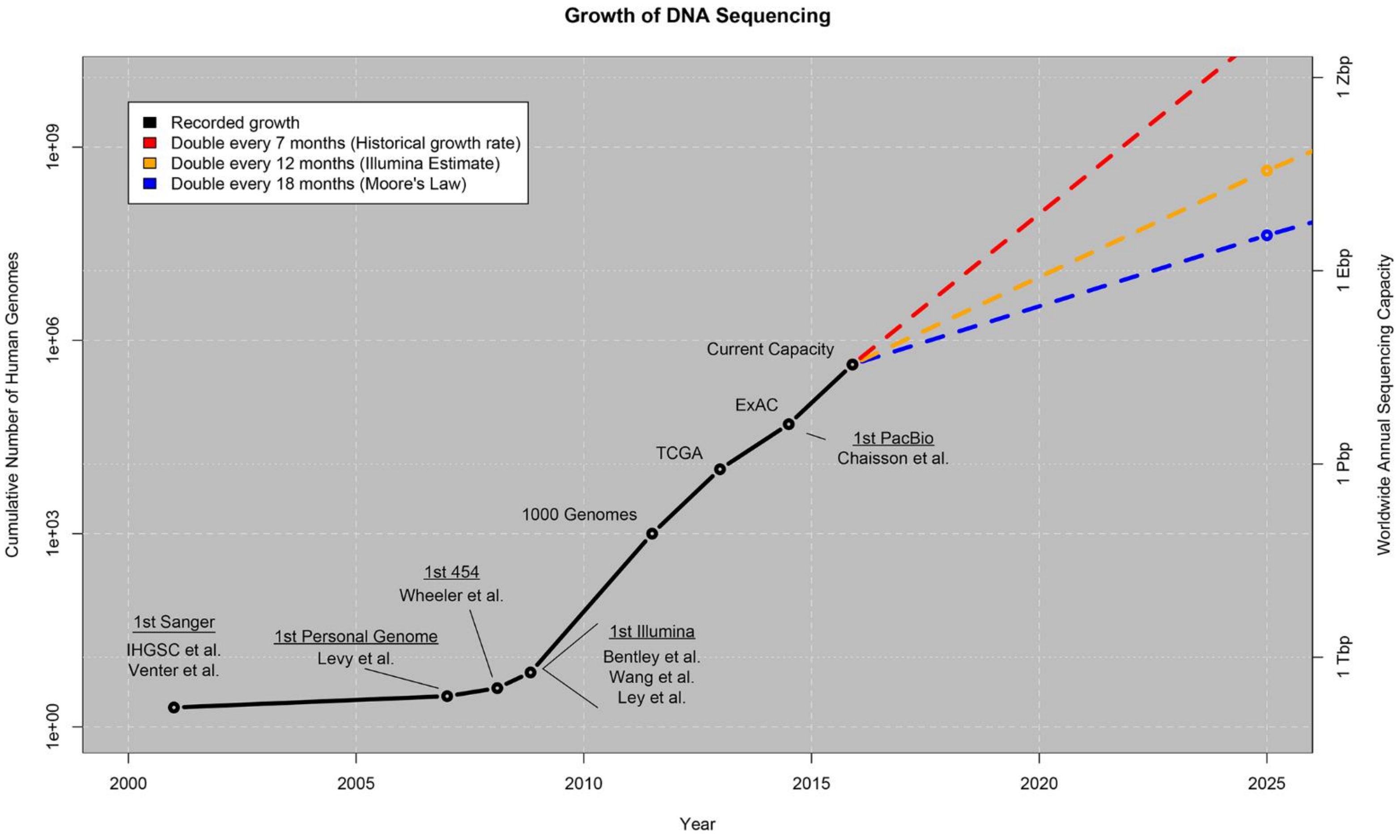
- take a parachute and jump (into the tidyverse)

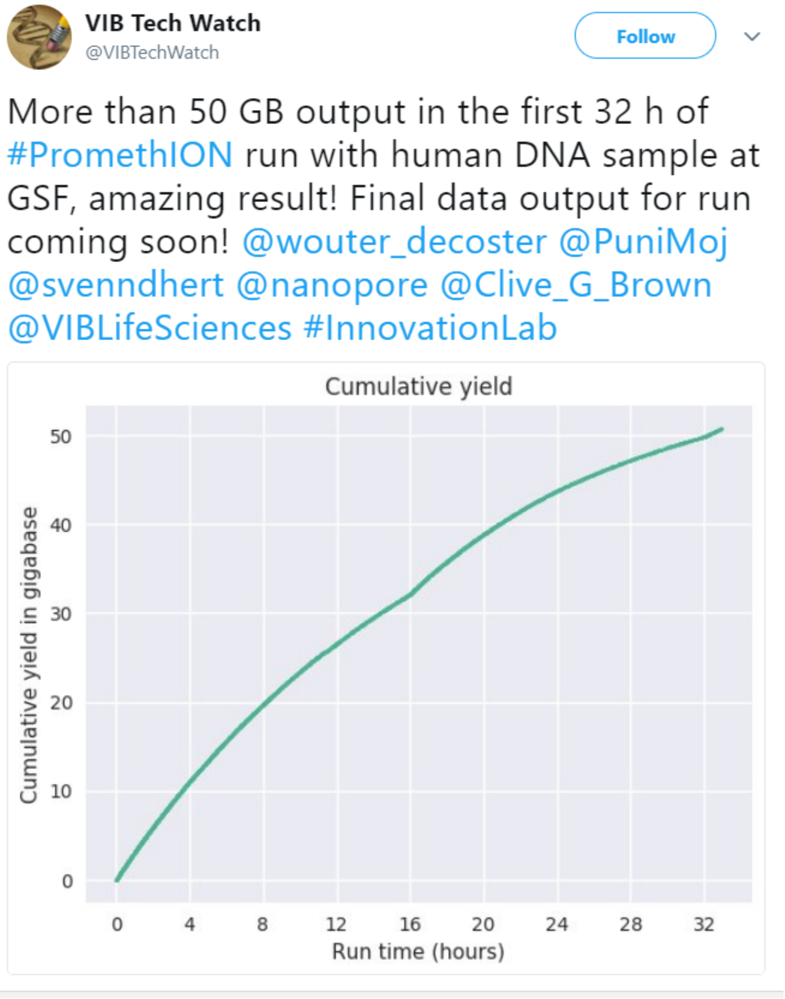
Brendan Palmer,
Statistics & Data Analysis Unit,
Clinical Research Facility - Cork

# Snapshot of data analysis tools used in UCC



# The explosion of data in the life sciences





VIB Tech Watch
@VIBTechWatch

More than 50 GB output in the first 32 h of #PromethION run with human DNA sample at GSF, amazing result! Final data output for run coming soon! @wouter_decoster @PuniMoj @svenndhert @nanopore @Clive_G_Brown @VIBLifeSciences #InnovationLab

Stevens et. al., 2015, Plos Biology

# R is worth learning

It’s free +

accessible +

reproducible +

widely used +

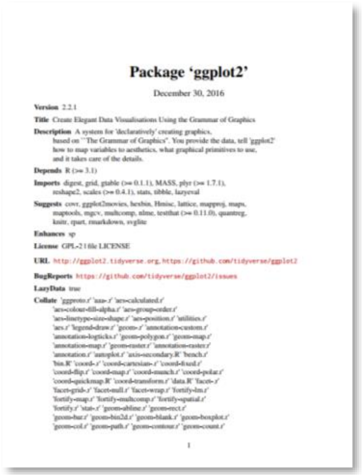
broadly applicable +

works across platforms +

and …………

# …help comes in many forms

## Vignettes



### Package 'ggplot2'

December 30, 2016

**Version** 2.2.1

**Title** Create Elegant Data Visualisations Using the Grammar of Graphics

**Description** A system for 'declaratively' creating graphics,
based on ``The Grammar of Graphics''. You provide the data, tell 'ggplot2'
how to map variables to aesthetics, what graphical primitives to use,
and it takes care of the details.

**Depends** R (>= 3.1)

**Imports** digest, grid, gtable (>= 0.1.1), MASS, plyr (>= 1.7.1),
reshape2, scales (>= 0.4.1), stats, tibble, lazyeval

**Suggests** covr, ggplot2movies, hexbin, Hmisc, lattice, mapproj, maps,
maptools, mgcv, multcomp, nlme, testthat (>= 0.11.0), quantreg,
knitr, rpart, rmarkdown, svglite

**Enhances** sp

**License** GPL-2 | file LICENSE

**URL** http://ggplot2.tidyverse.org, https://github.com/tidyverse/ggplot2

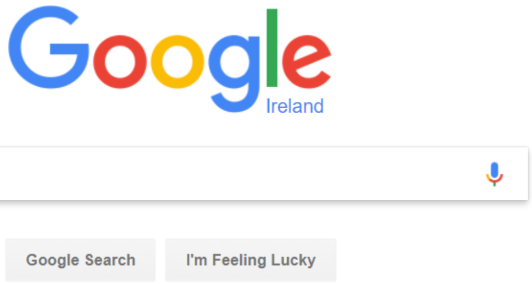
**BugReports** https://github.com/tidyverse/ggplot2/issues

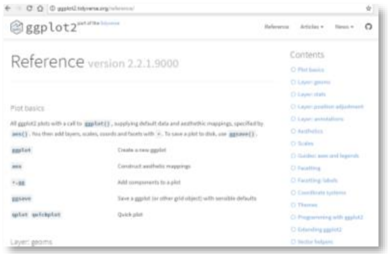
**LazyData** true

**Collate** 'ggproto.r' 'aaa-.r' 'aes-calculated.r'
'aes-colour-fill-alpha.r' 'aes-group-order.r'
'aes-linetype-size-shape.r' 'aes-position.r' 'utilities.r'
'aes.r' 'legend-draw.r' 'geom-.r' 'annotation-custom.r'
'annotation-logticks.r' 'geom-polygon.r' 'geom-map.r'
'annotation-map.r' 'geom-raster.r' 'annotation-raster.r'
'annotation.r' 'autoplot.r' 'axis-secondary.R' 'bench.r'
'bin.R' 'coord-.r' 'coord-cartesian-.r' 'coord-fixed.r'
'coord-flip.r' 'coord-map.r' 'coord-munch.r' 'coord-polar.r'
'coord-quickmap.R' 'coord-transform.r' 'data.R' 'facet-.r'
'facet-grid-.r' 'facet-null.r' 'facet-wrap.r' 'fortify-lm.r'
'fortify-map.r' 'fortify-multcomp.r' 'fortify-spatial.r'
'fortify.r' 'stat-.r' 'geom-abline.r' 'geom-rect.r'
'geom-bar.r' 'geom-bin2d.r' 'geom-blank.r' 'geom-boxplot.r'
'geom-col.r' 'geom-path.r' 'geom-contour.r' 'geom-count.r'

1

## Webpages




ggplot2.tidyverse.org/reference/
ggplot2 part of the tidyverse
Reference
Articles
News

Reference version 2.2.1.9000

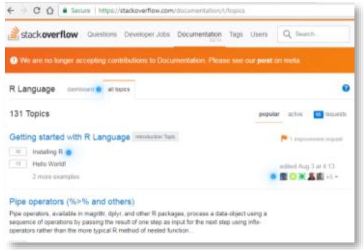
Plot basics

All ggplot2 plots with a call to `ggplot()`, supplying default data and aesthethic mappings, specified by `aes()`. You then add layers, scales, coords and facets with `+`. To save a plot to disk, use `ggsave()`.

| | |
|---|---|
| `ggplot` | Create a new ggplot |
| `aes` | Construct aesthetic mappings |
| `+.gg` | Add components to a plot |
| `ggsave` | Save a ggplot (or other grid object) with sensible defaults |
| `qplot` `quickplot` | Quick plot |

Layer: geoms

Contents
- Plot basics
- Layer: geoms
- Layer: stats
- Layer: position adjustment
- Layer: annotations
- Aesthetics
- Scales
- Guides: axes and legends
- Facetting
- Facetting: labels
- Coordinate systems
- Themes
- Programming with ggplot2
- Extending ggplot2
- Vector helpers



We are no longer accepting contributions to Documentation. Please see our **post** on meta.

R Language

131 Topics

Getting started with R Language Introduction Topic

1 improvement request

Installing R

Hello World!

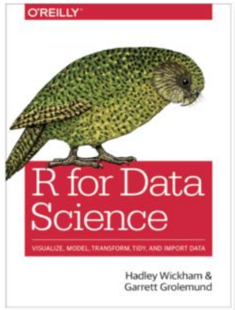
2 more examples

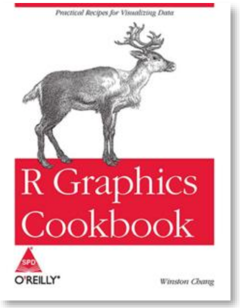
edited Aug 3 at 4:13

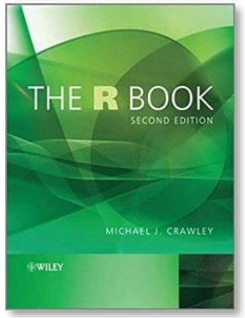
Pipe operators (%>% and others)

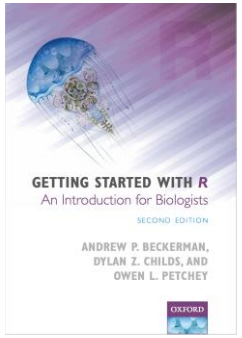
Pipe operators, available in magrittr, dplyr, and other R packages, process a data-object using a sequence of operations by passing the result of one step as input for the next step using infix-operators rather than the more typical R method of nested function…

## eBooks









## Twitter

**GettingStartedWithR**
@GSwithR

Getting Started with R (Beckerman, Childs & Petchey) is a book designed for getting started with R. Collecting & making beginner R tweets. r4all.org

Sheffield, UK & Zurich, CH
r4all.org

**R Tips**
@R_Programming

Great #rstats Tips, Offering #datascience course at goo.gl/is9vyV and Tuts on #datamining #analytics rstatistics.net. Tweets by @selva86

r-statistics.co

**Hadley Wickham**
@hadleywickham

R, data, visualisation.

Houston, TX
hadley.nz

**Jenny Bryan**
@JennyBryan

@rstudio, humane #rstats, statistics, @ropensci, teach @STAT545, on leave from @UBC, math.ubc.ca/~jbryan/JennyC...

**David Robinson**
@drob

Data Scientist at @StackOverflow, #rstats fan/evangelist

New York, NY
varianceexplained.org

**Data Scientists IRL**
@DataSci_Ireland Follows you

Promoting the Data Science professions in Ireland.

Ireland
facebook.com/DataScientists...

**Darren L Dahly**
@statsepi Follows you

Principal Statistician, Epidemiologist, Sr Lecturer | @HRBIreland Clinical Research Facility @CRF_CORK | Cork #Rstats Users Group meetup.com/Cork-Ireland-R...

Cork, Ireland
darrendahly.github.io

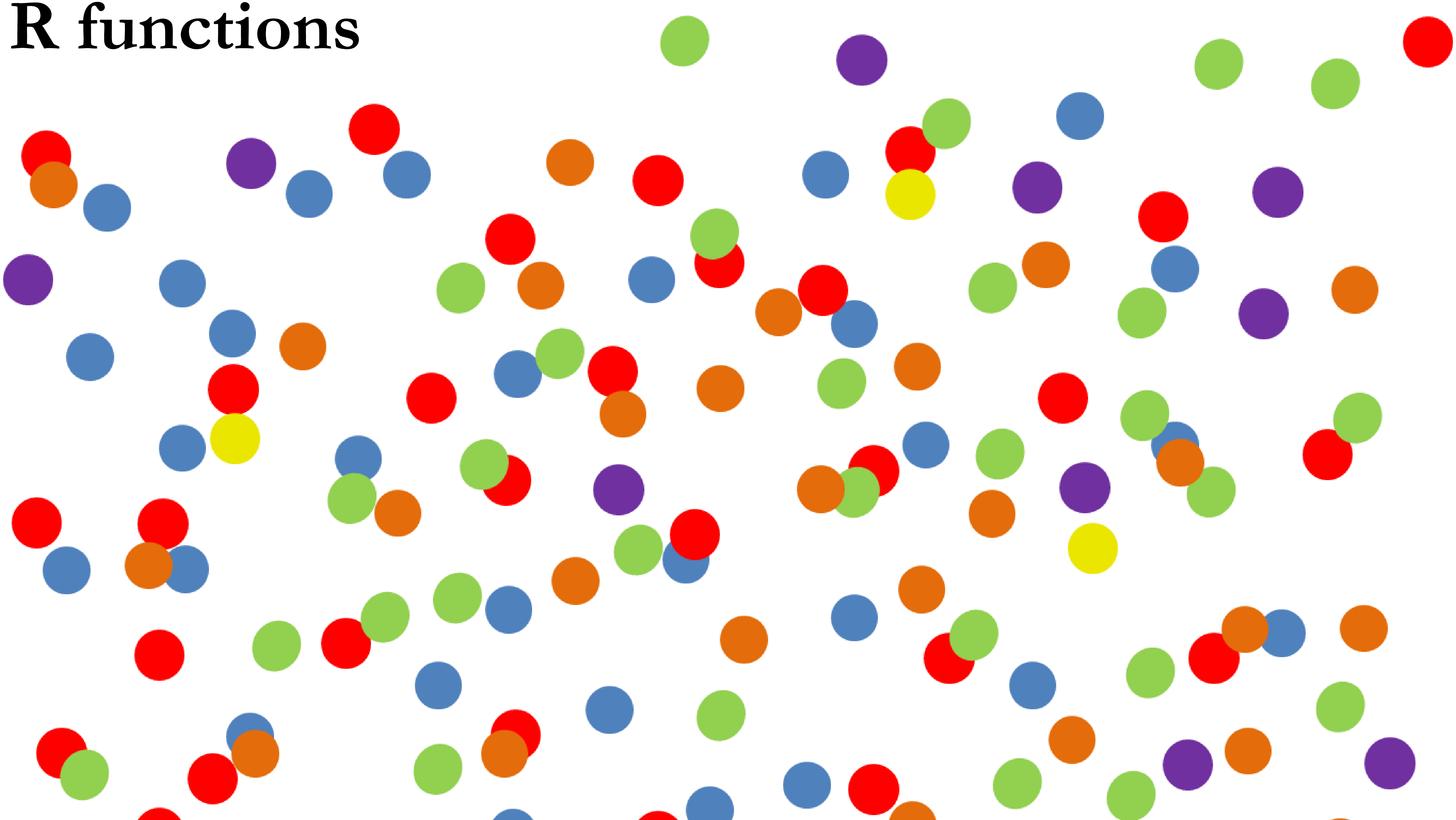
**Data Science Renee**
@BecomingDataSci

Creator/Host of Becoming a Data Scientist Podcast // @DataSciGuide @DataSciLearning @NewDataSciJobs // Personal acct: @paix120 // Data Scientist at @HelioCampus

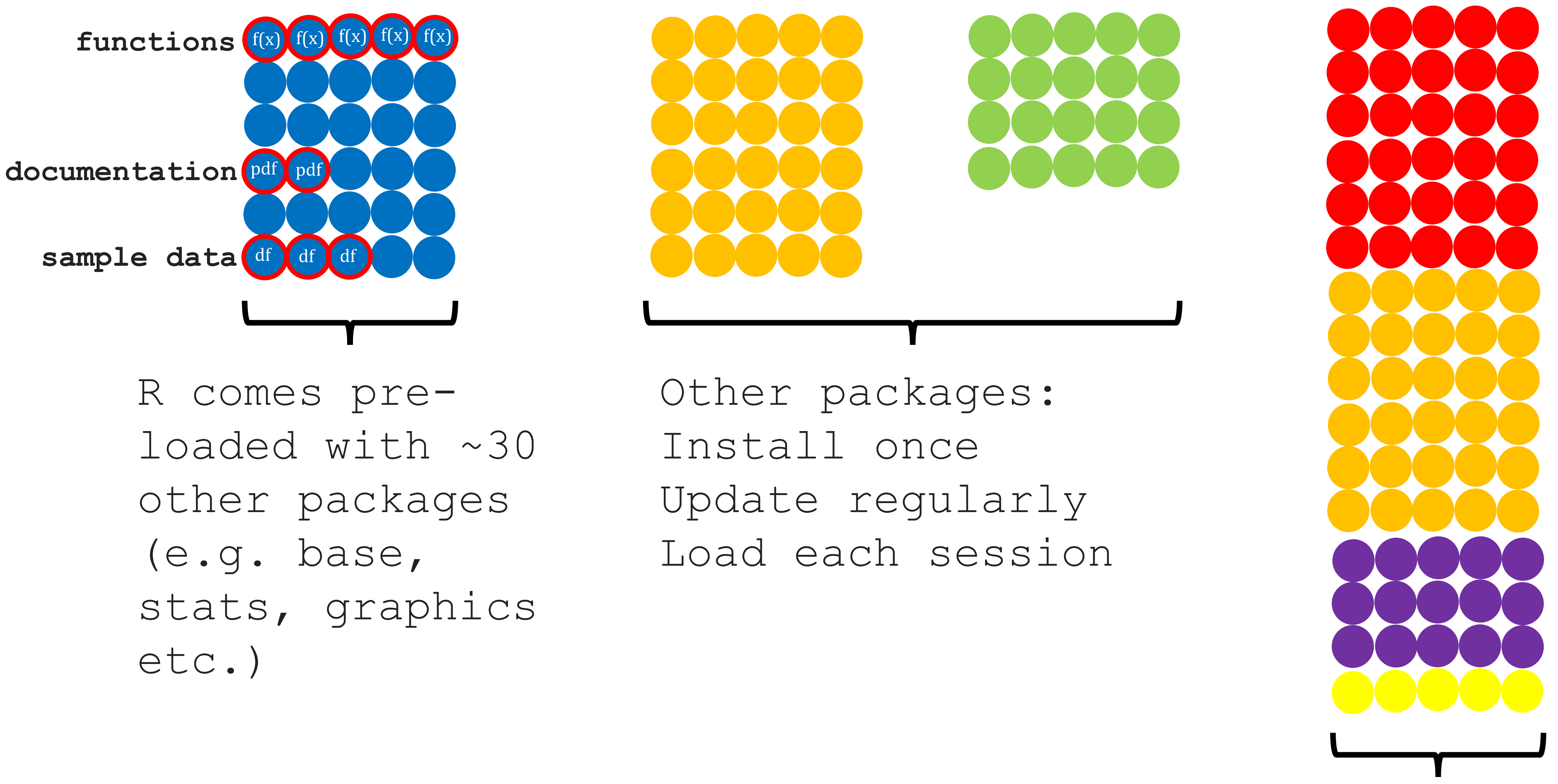
# Cheatsheets

The Caret Package
caret Package
Base R
Data Transformation with dplyr :: CHEAT SHEET
R Markdown :: CHEAT SHEET
R Markdown Reference Guide
Data Science in Spark with Sparklyr :: CHEAT SHEET
R Syntax Comparison :: CHEAT SHEET
Dollar sign syntax
Formula syntax
Tidyverse syntax
The variety of R syntaxes give you many ways to "say" the same thing
Data Import :: CHEAT SHEET
RStudio IDE
Data Visualization with ggplot2 :: CHEAT SHEET
Shiny :: CHEAT SHEET
Advanced R

# R functions



# R packages



R comes pre-loaded with ~30 other packages (e.g. base, stats, graphics etc.)

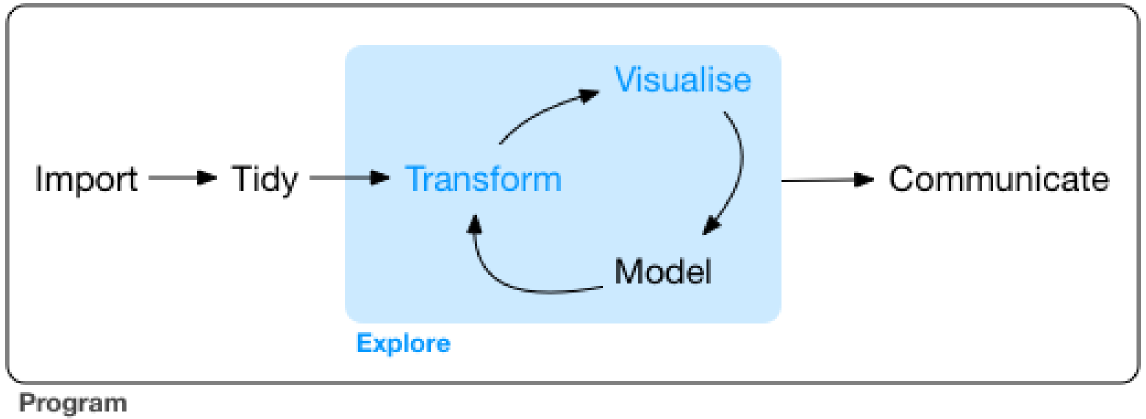
Other packages:
Install once
Update regularly
Load each session

tidyverse

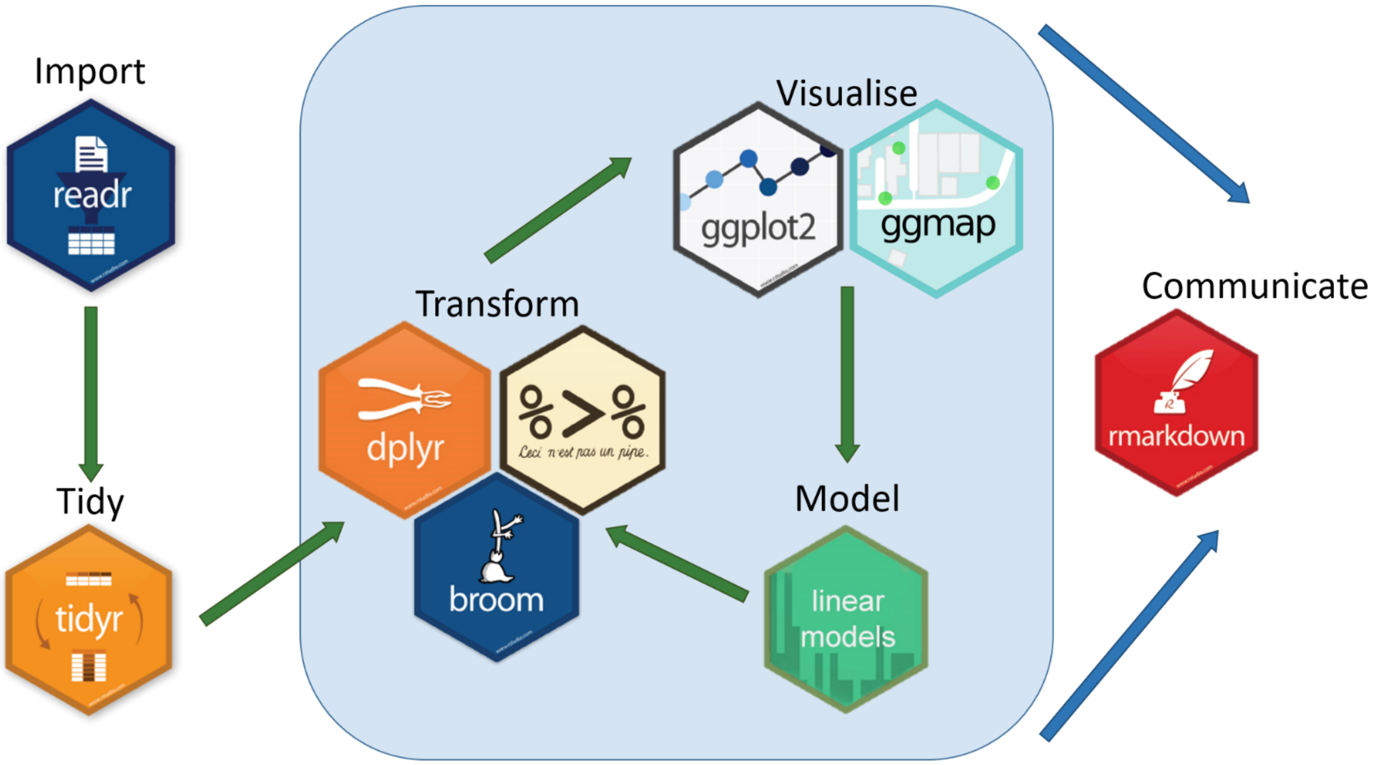
# Data analysis in a nutshell

"The good news about computers is that they do what you tell them to do. The bad news is that they do what you tell them to do." - Ted Nelson

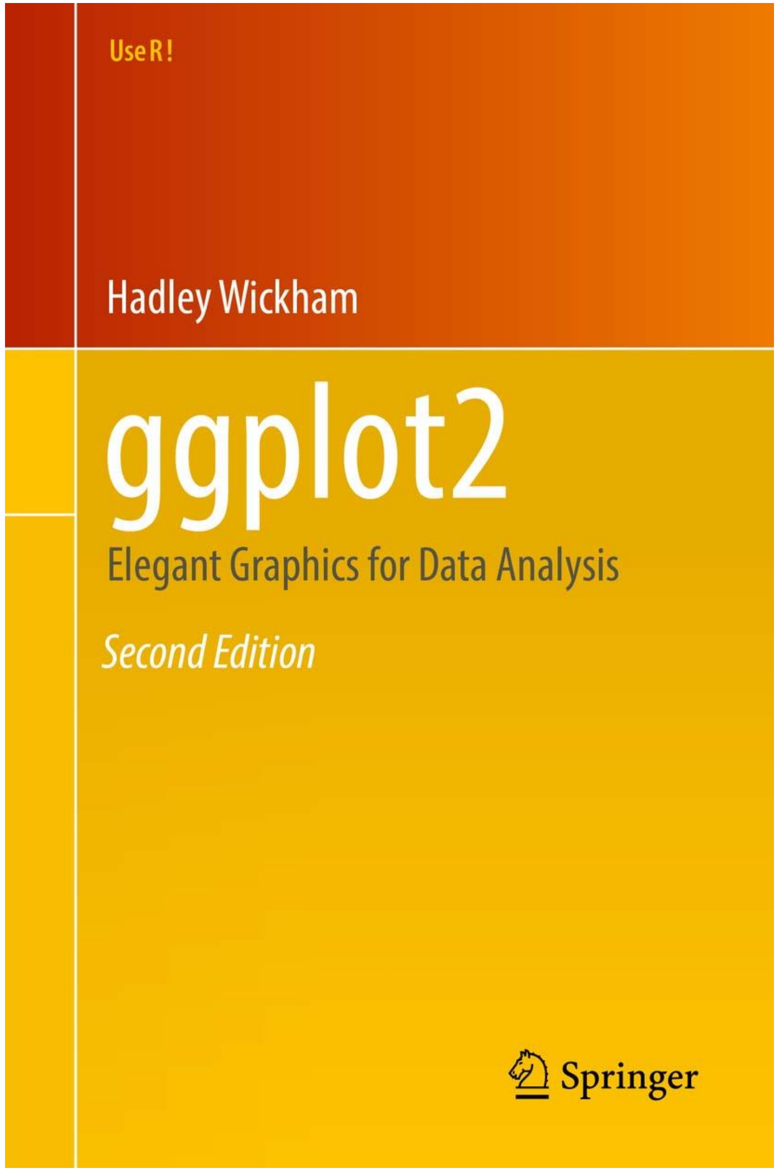
The tidyverse makes R code more human readable
- it is easier to write, run and read



# Data analysis in a tidyverse nutshell



# You could write a book on that!!



Use R!

Hadley Wickham

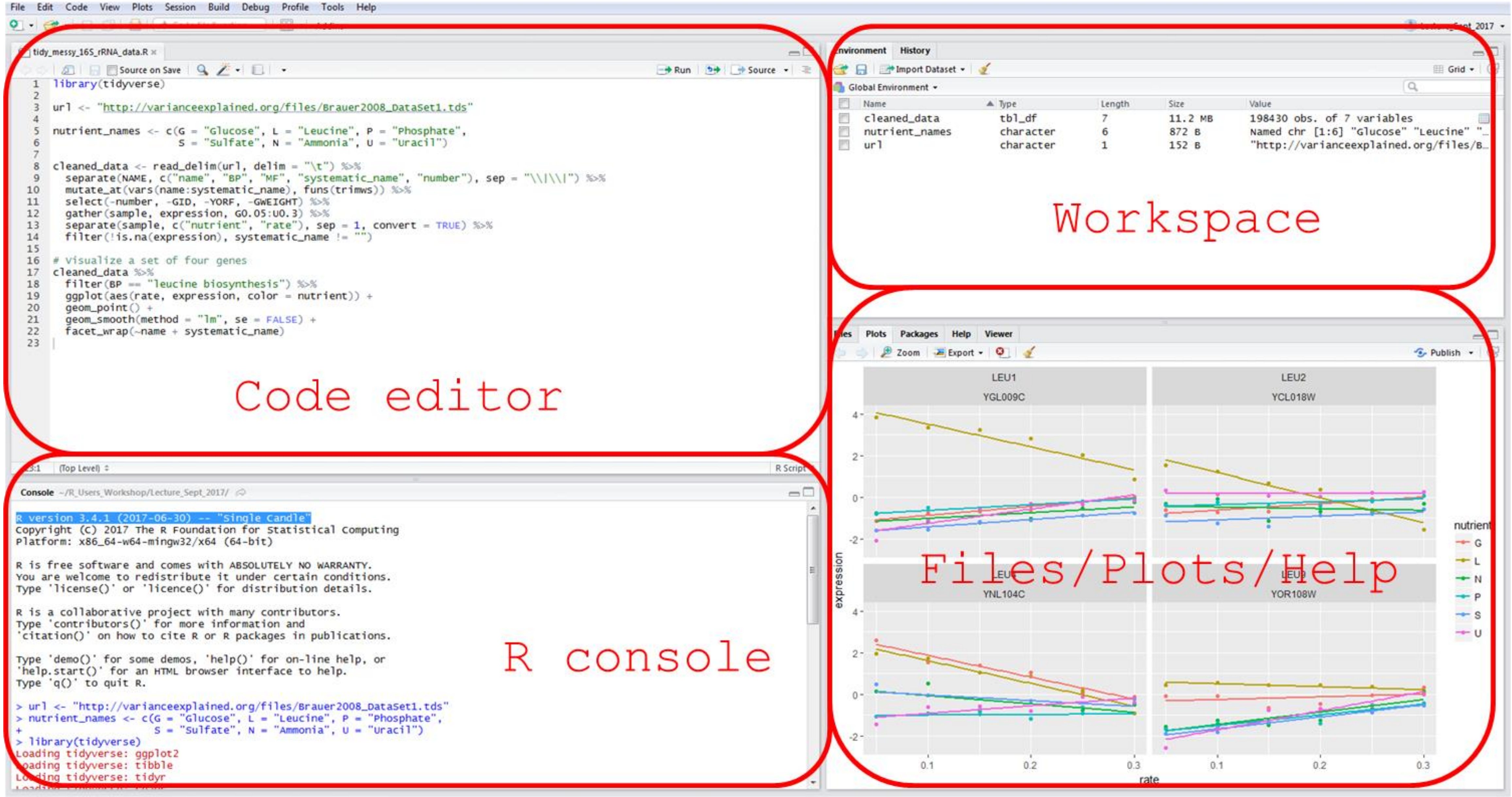
ggplot2

Elegant Graphics for Data Analysis

*Second Edition*

Springer

# RStudio



```r
library(tidyverse)

url <- "http://varianceexplained.org/files/Brauer2008_DataSet1.tds"

nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
                    S = "Sulfate", N = "Ammonia", U = "Uracil")

cleaned_data <- read_delim(url, delim = "\t") %>%
  separate(NAME, c("name", "BP", "MF", "systematic_name", "number"), sep = "\\|\\|") %>%
  mutate_at(vars(name:systematic_name), funs(trimws)) %>%
  select(-number, -GID, -YORF, -GWEIGHT) %>%
  gather(sample, expression, G0.05:U0.3) %>%
  separate(sample, c("nutrient", "rate"), sep = 1, convert = TRUE) %>%
  filter(!is.na(expression), systematic_name != "")

# visualize a set of four genes
cleaned_data %>%
  filter(BP == "leucine biosynthesis") %>%
  ggplot(aes(rate, expression, color = nutrient)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  facet_wrap(~name + systematic_name)

```

| Name | Type | Length | Size | Value |
|---|---|---|---|---|
| cleaned_data | tbl_df | 7 | 11.2 MB | 198430 obs. of 7 variables |
| nutrient_names | character | 6 | 872 B | Named chr [1:6] "Glucose" "Leucine" "… |
| url | character | 1 | 152 B | "http://varianceexplained.org/files/B… |

```
R version 3.4.1 (2017-06-30) -- "Single Candle"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> url <- "http://varianceexplained.org/files/Brauer2008_DataSet1.tds"
> nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
+                     S = "Sulfate", N = "Ammonia", U = "Uracil")
> library(tidyverse)
Loading tidyverse: ggplot2
Loading tidyverse: tibble
Loading tidyverse: tidyr
```

# Worksheet 1

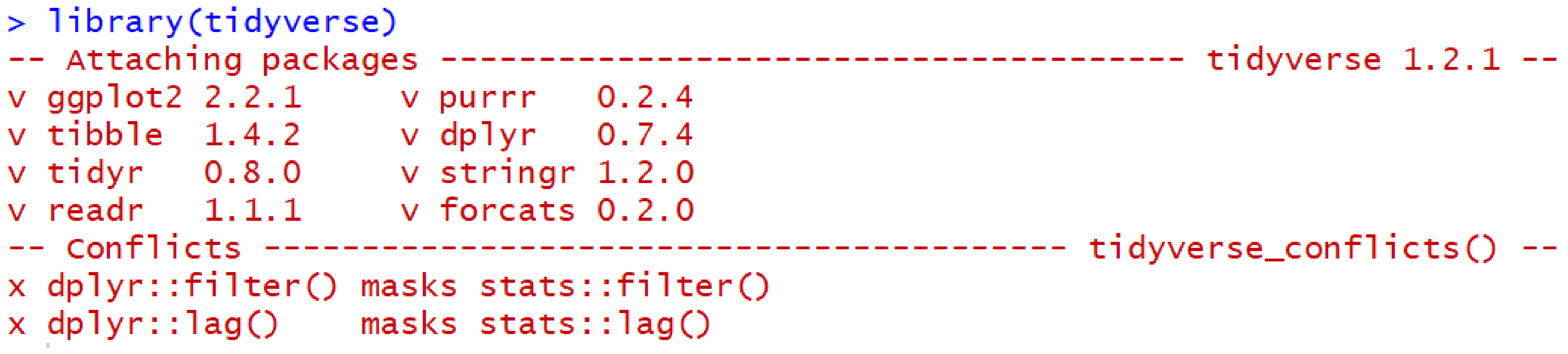
## Open ws1_script1_navigating_R_packages.R

# Basics of R code

| Symbol | What it does | Example 1 | Example 2 |
|---|---|---|---|
| <- | Assign operator<br>Creates new objects | > x <- 5<br>> x<br>[1] 5 | > y <- "This"<br>> y<br>[1] "This" |
| c() | Helps create objects with more than one element | > v <- c(5,6,7,8)<br>> v<br>[1] 5 6 7 8 | > w <- c("This", "is", "easy! ")<br>> w<br>[1] "This" "is" "easy!" |
| # | Computer ignores what is written. Used for adding notes to code | > #print(“hello”)<br>> | > print(“hello”)<br>[1] "hello" |
| %>% | Literally translates as “then do this” | > data %>% do.something.to(data) | |
| %in% | returns a logical vector indicating if there is a match | > "x" %in% c("x", "y", "z")<br>[1] TRUE | > c("x", "y", "z") %in% "x"<br>[1]  TRUE FALSE FALSE |
| ? | Access information | > ?mean() | > ?geom_point() |

**FYI: R is case sensitive!!  Name.of.data ≠ name.of.data**

# The tidyverse package 1.2.1



```
> library(tidyverse)
-- Attaching packages -------------------------------------- tidyverse 1.2.1 --
v ggplot2 2.2.1     v purrr   0.2.4
v tibble  1.4.2     v dplyr   0.7.4
v tidyr   0.8.0     v stringr 1.2.0
v readr   1.1.1     v forcats 0.2.0
-- Conflicts ----------------------------------------- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```
ggplot2 <- data visualisation    purrr   <- functional programming
tibble  <- data frames           dplyr   <- data manipulation
readr   <- data import           stringr <- string manipulation
tidyr   <- data tidying          forcats <- categorical variables
```

# Tidy data should satisfy the following:

Each variable forms a column

Each observation forms a row

**In Brauer et al., 2008:**

- column headers are values not variable names
- multiple variable are stored in one column

e.g. 1: the column “NAME” contains information such as;

SFB2 || ER to Golgi transport || molecular function unknown || YNL049C || 1082129

- these need to be split into new columns

e.g. 2: columns G0.05 to U0.03 identify the limiting nutrient (letter) and the growth rate (number) combinations

# Try to limit “uninformative” data

“GWEIGHT” contains the same information in every cell
    - this isn’t going to add to our analysis

“GID” and “YORF” appear to be study specific IDs

“NAME” column contains a lot of information

Going back to the previous example;
SFB2 || ER to Golgi transport || molecular function unknown || YNL049C || 1082129

SFB2: Gene names, but not present in all cases
ER to Golgi transport: Biological process
molecular function unknown: Molecular function
YNL049C: Gene ID listed on public repositories
1082129: Another identifier that does not appear to be useful

# Code structure example

```
separated_gene_df <- separate(raw_gene_df, NAME, c("name", "BP", "MF", "systematic_name", "number"), sep = "\\|\\|")
```

`separated_gene_df` -the new object you will create

`<-` -the assign operator

`separate` -the function you are calling

`(raw_gene_df,` -the input data

`NAME,` -the column to be separated

`c("name", "BP", "MF", "systematic_name", "number"),` -new columns IDs

`sep = "\\|\\|")` -how the data will be split

# Worksheet 2

## Open ws1_script2_stepwise_Brauer_analysis.R

# How to plot in ggplot

Template:

```
ggplot(data = <DATA>) +
     <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>))    +
      linear model   +
      axes formatting   +
      legend formatting   +
      title    + etc. etc.
```

# Worksheet 3

## Open ws1_script3_piped_Brauer_analysis.R

# Introductory R Workshops

~~**Week 1 (13th February):**~~

~~**Take a parachute and jump (into the tidyverse)**~~

~~- tidying and visualisation of NGS data using sample R scripts~~

**Week 2 (20th February):**

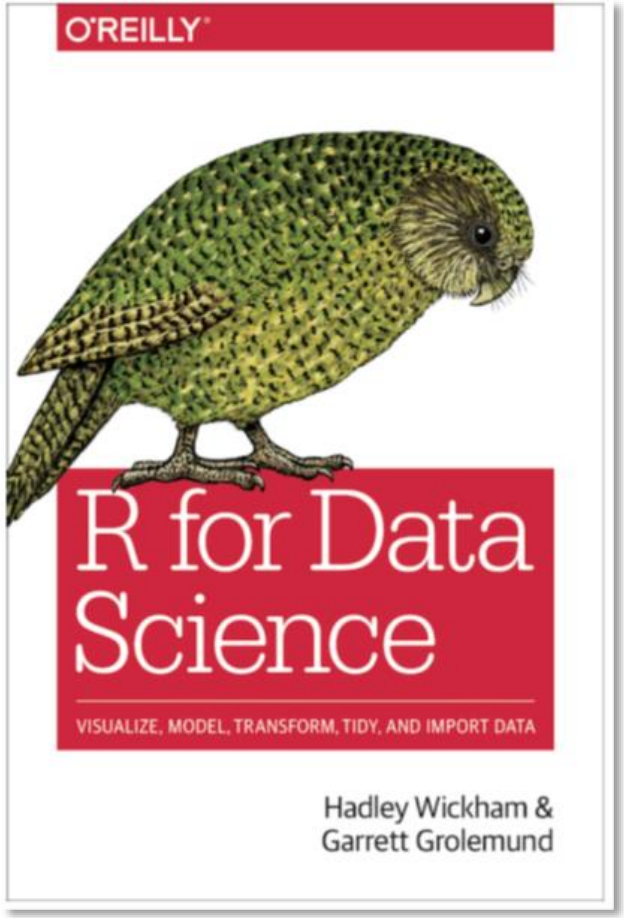
**We built this software on base R code**

- overview and structure of R syntax

**Week 3 (27th February):**

**Sending an SOS to the world**

- how to identify with errors in your code and get help



# Introductory R Workshops

**Week 4 (6th March):**
**It's the end of base R as you know it**

- introduction to the tidyverse packages tidyr and dplyr

**Week 5 (13th March):**
**Welcome to the ggungle**

- analysis and visualisation of data

**Week 6 (20th March):**
**Yesterday**

- writing clear code and making your work reproducible