Workshop 1:
The tidyverse and beyond

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

CRF-C
HRB Clinical Research Facility Cork

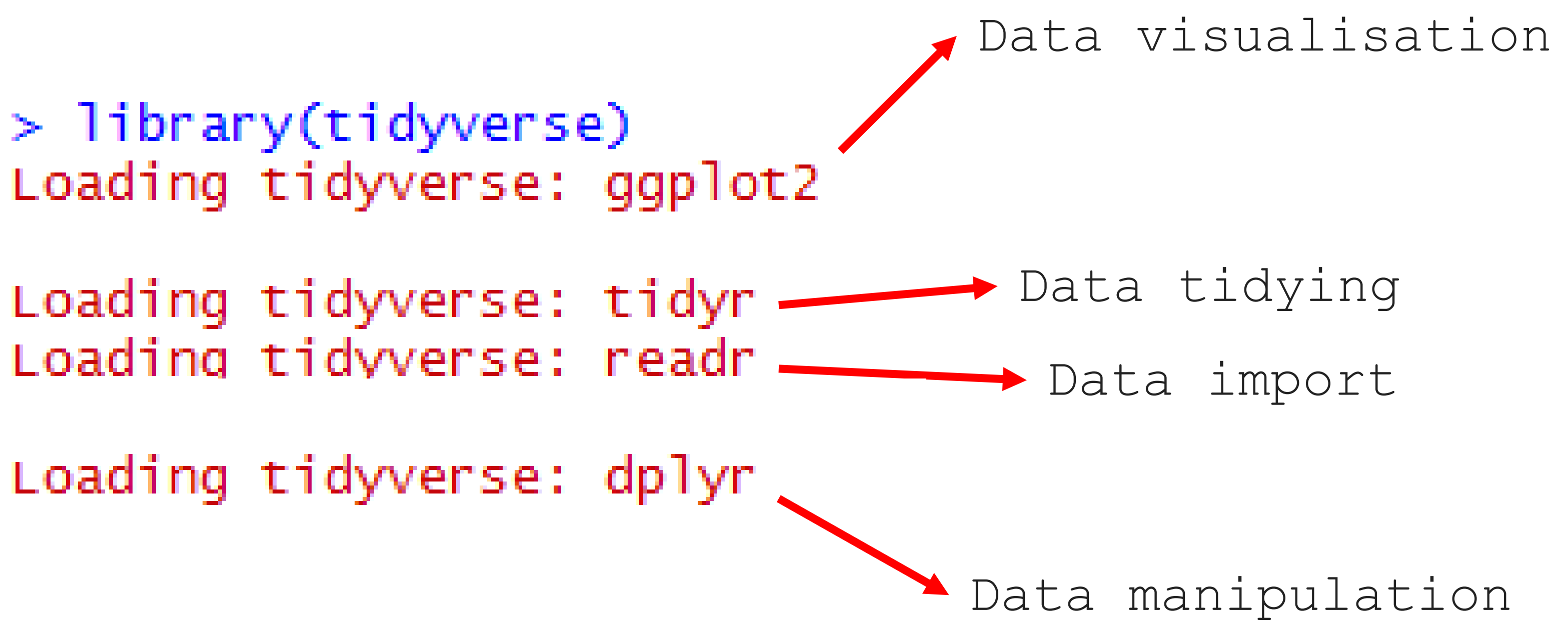
UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

HRB
Health Research Board

The tidyverse package

A bundle of ~20 individual R packages

The six main ones are loaded when the tidyverse package is called



```
> library(tidyverse)
Loading tidyverse: ggplot2

Loading tidyverse: tidyr
Loading tidyverse: readr

Loading tidyverse: dplyr
```

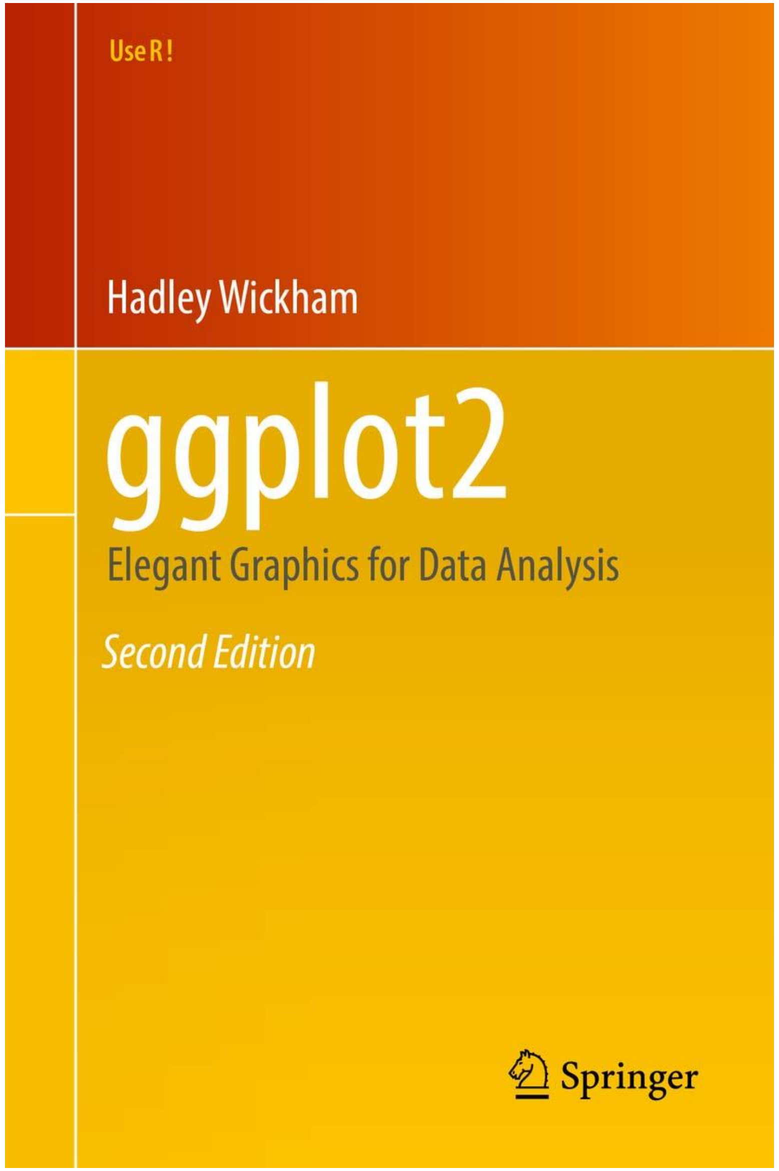
Data visualisation

Data tidying

Data import

Data manipulation

You could write a book on that!!



Use R!

Hadley Wickham

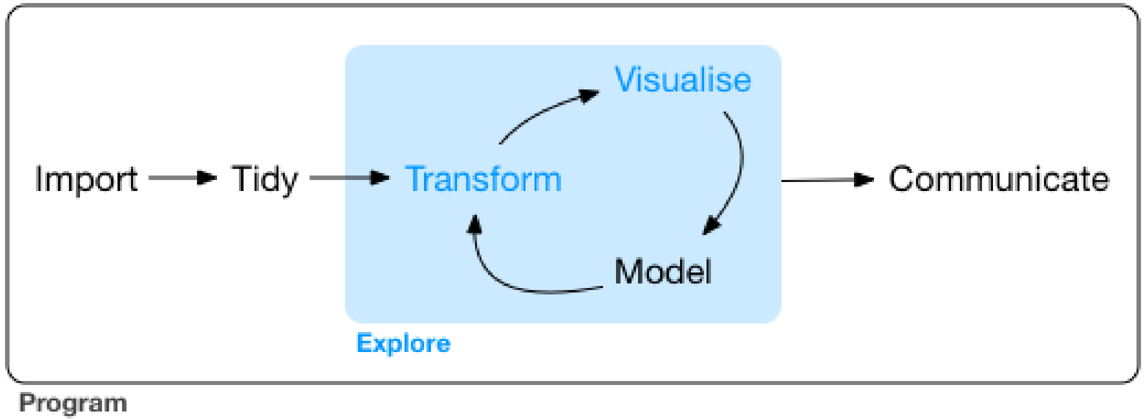
ggplot2

Elegant Graphics for Data Analysis

Second Edition

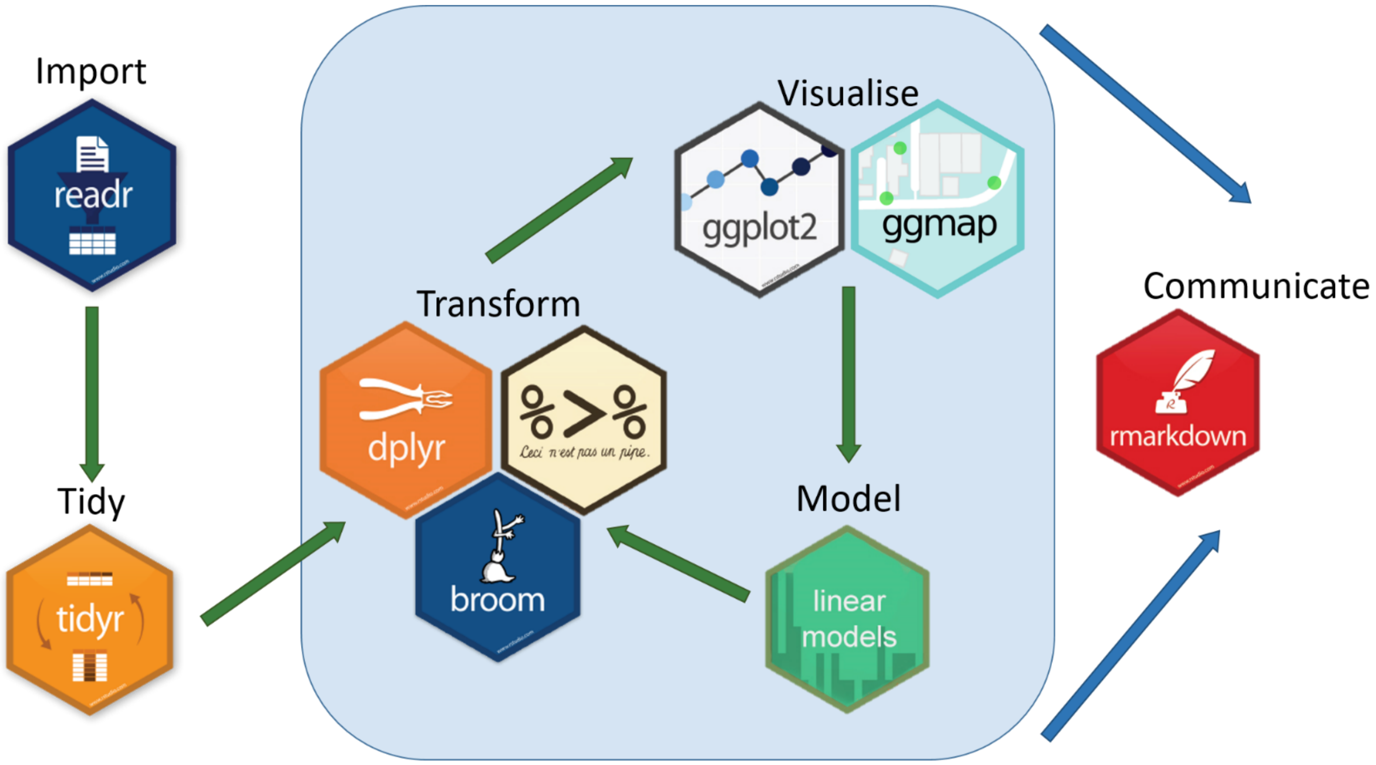
Springer

Data analysis in a nutshell



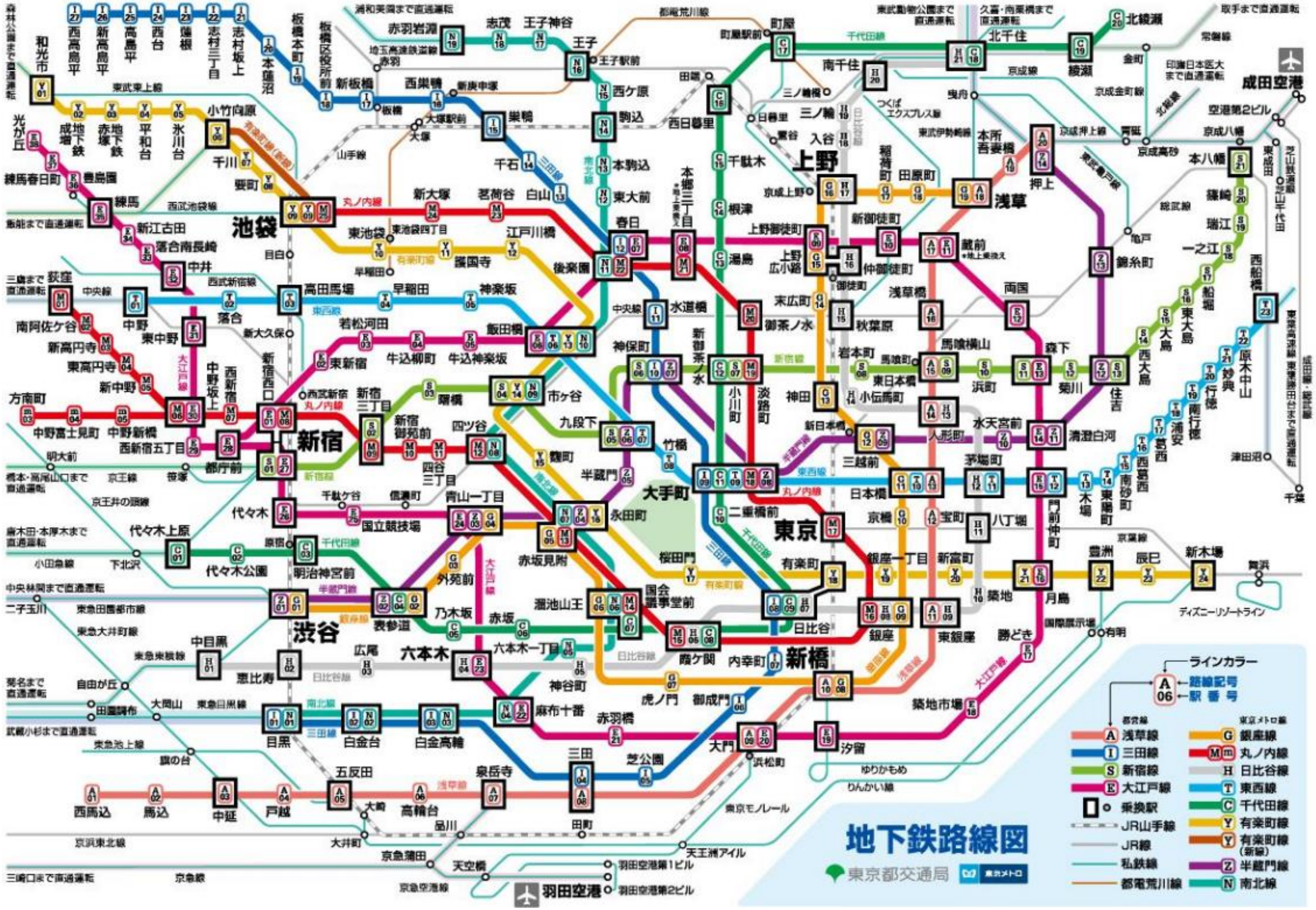
The only obstacle to this is getting the information inside your brain translated into a machine readable format

Data analysis in a tidyverse nutshell

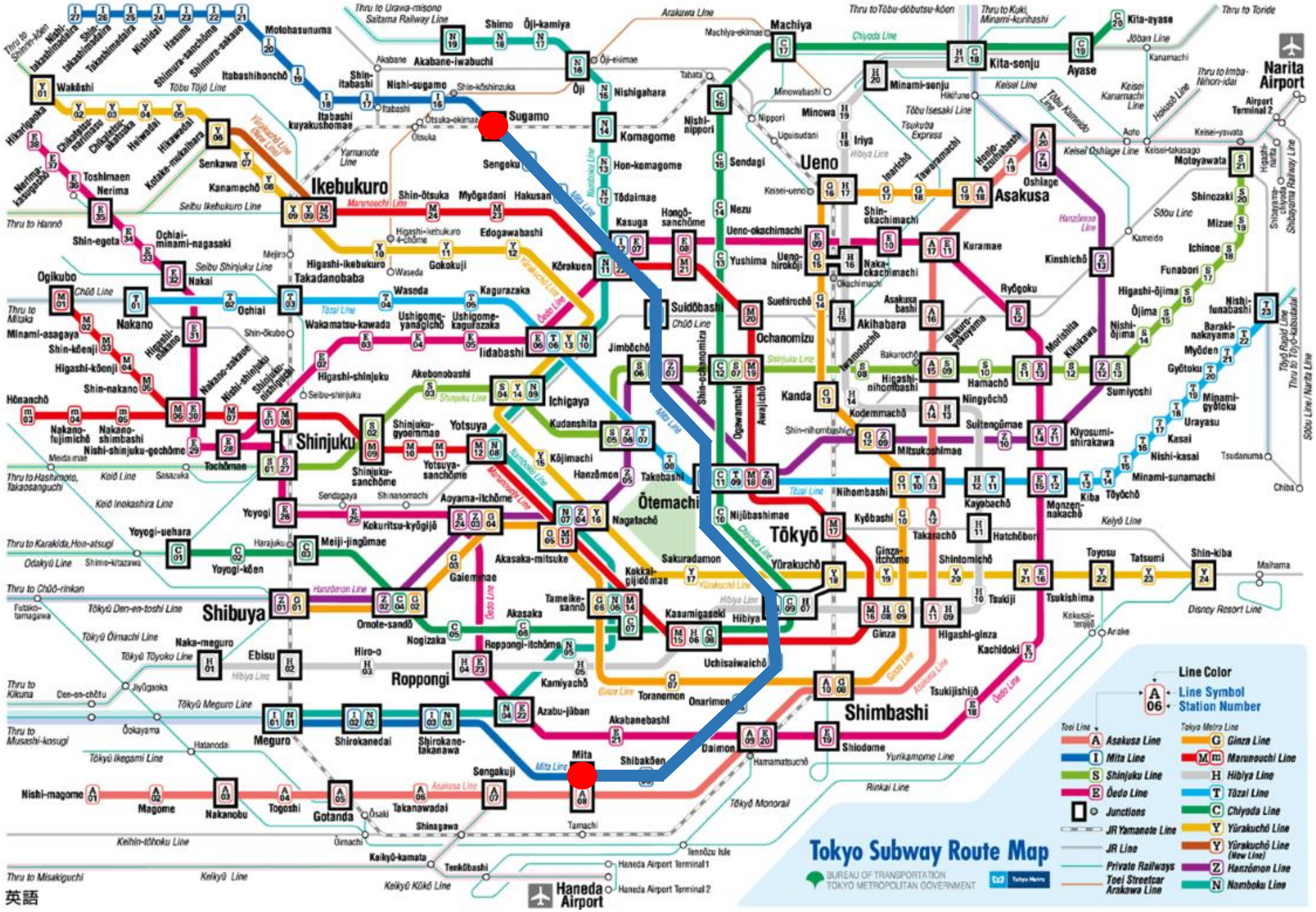


Communication with computers: C++

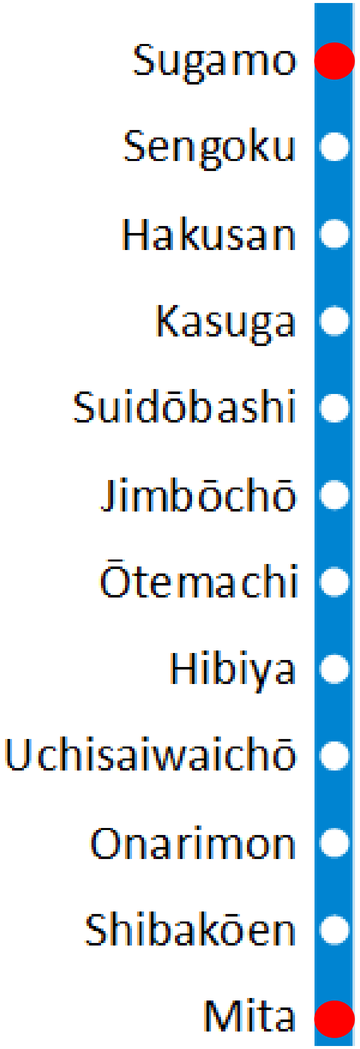
The challenge:
To find the best route from Mita to Sugamo station on the Toyko metro



Communication with computers: base R



Communication with computers: tidyversec



In general, there are many ways to achieve the same result in programming

With the tidyverse, there is one preferred way to achieve an action!

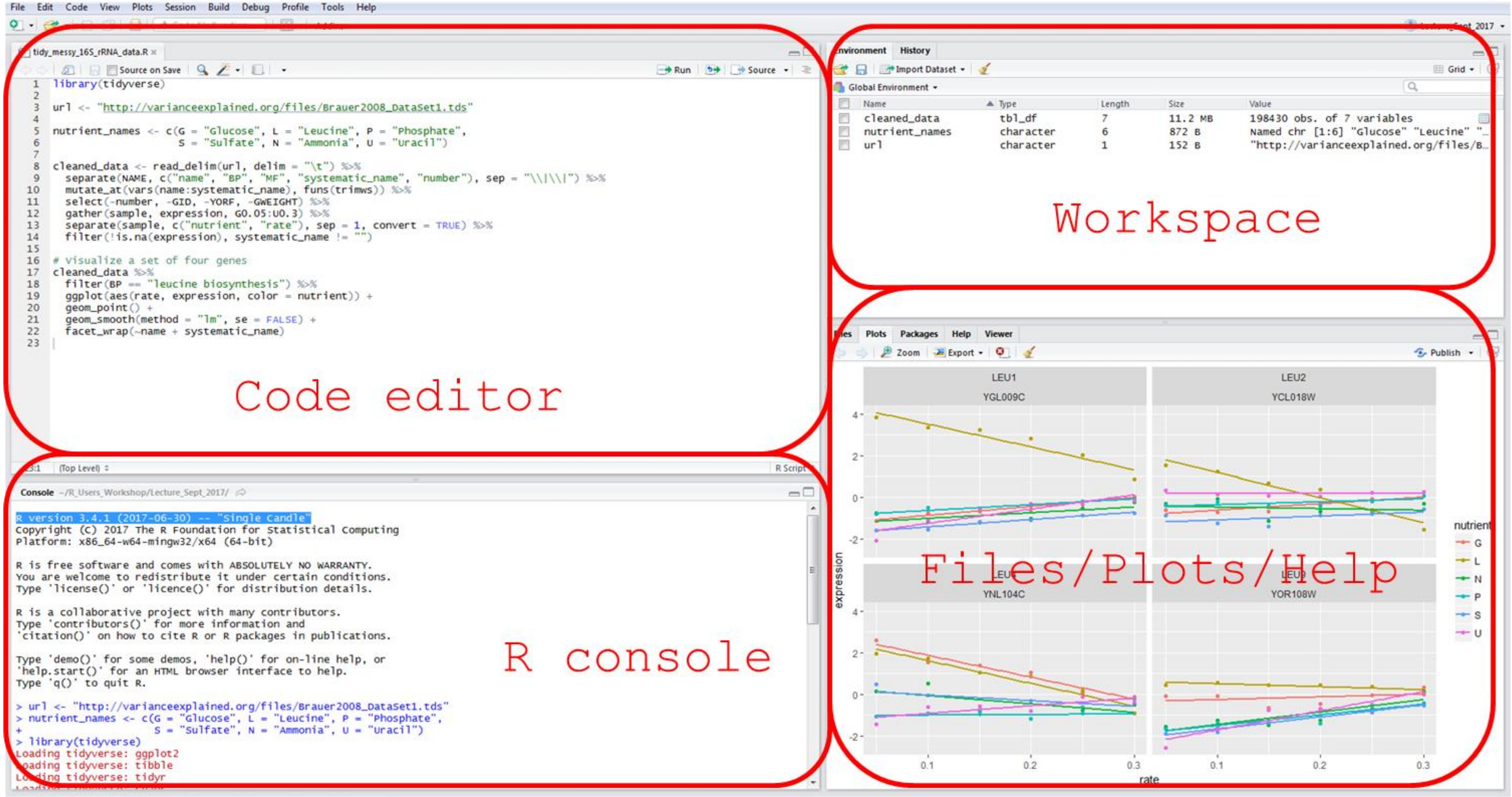
The tidyverse mantra is that each function does one thing really well

Basics of R code

Symbol	What it does	Example 1	Example 2
<-	Creates objects	> x <- 5 > x [1] 5	> y <- "This" > y [1] "This"
c()	Helps create objects with more than one element	> v <- c(5,6,7,8) > v [1] 5 6 7 8	> w <- c("This", "is", "easy! ") > w [1] "This" "is" "easy!"
#	Computer ignores what is written. Used for adding notes to code	> #print("hello") >	> print("hello") [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") [1] TRUE	> c("x", "y", "z") %in% "x" [1] TRUE FALSE FALSE
?	Access information	> ?mean()	> ?geom_point()

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

Recall:

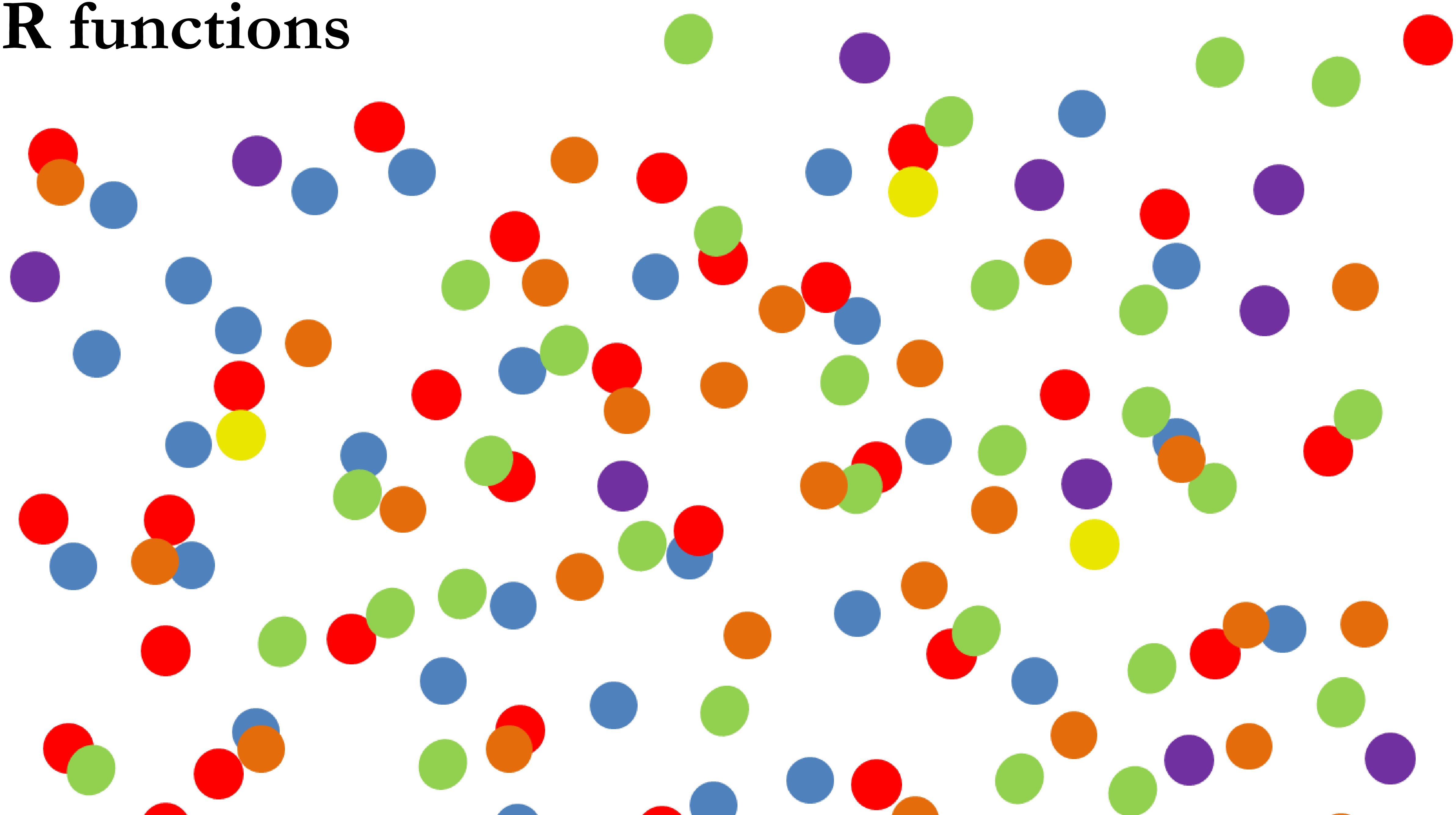


```r
library(tidyverse)

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)
```

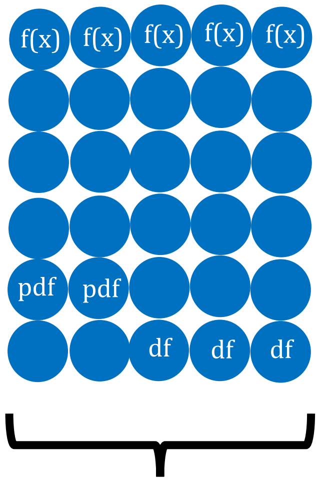
```
R version 3.4.1 (2017-06-30) -- "Single Candle"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

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
```

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…

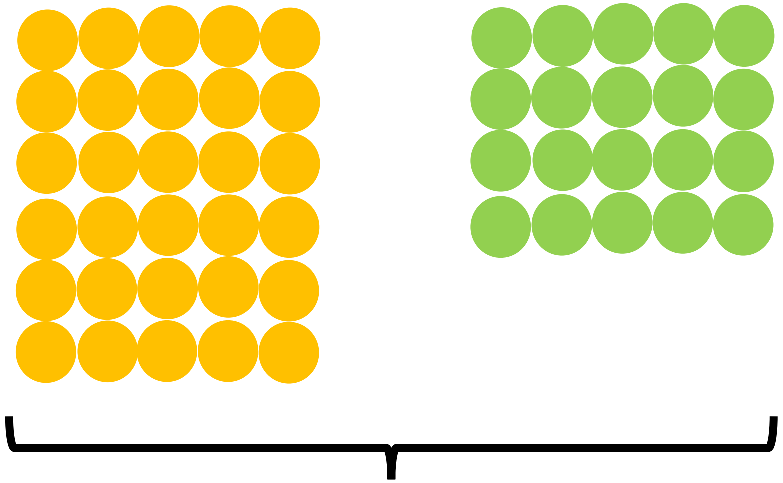
R functions



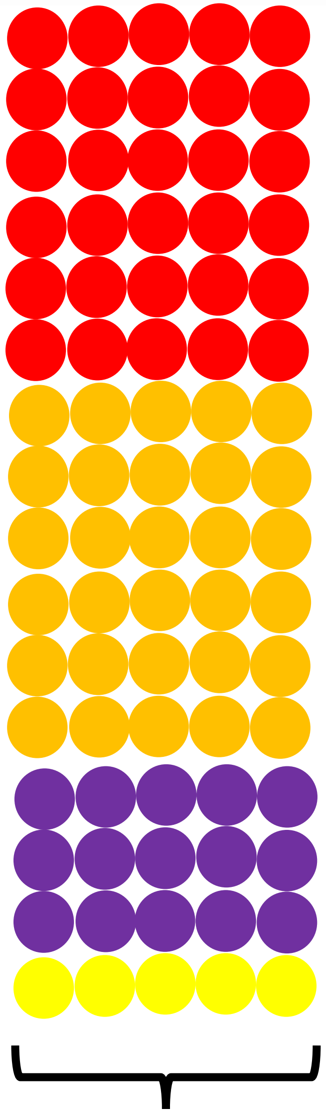
R packages



Base R:
Comes
pre-
loaded



Other packages:
Install once
Update regularly
Load each session



core
tidyverse

Worksheet 1

Part A

Tidy data should satisfy the following:

Each variable forms a column

Each observation forms a row

Bauer et al., 2008:
Column headers are values not variable names

Multiple variable are stored in one column

e.g. column "NAME" contains values such as;
SFB2 || ER to Golgi transport || molecular function unknown || YNL049C || 1082129

These need to be split up

G0.05:U0.03 letter is limiting nutrient and the number is the growth rate

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

Worksheet 1

Part B

Code structure

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

```
separated_gene_df            -the new data frame you will create

<-                           -the assign operator

separate                     -the function you are calling on

(raw_gene_df,                -the data frame to be used

NAME,                        -the column to be altered

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

sep = "\\|\\|")              -identify the separator to be used
```

Worksheet 1

Part C

How to plot in ggplot

Template:

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

Worksheet 1

Part D