R-eproducible workflows

Half-day workshop

(after 2.5 days of intense R training!!)

Brendan Palmer, University College Cork
Adam Kane, University College Dublin
Enrico Pirotta, Washington State University

Cork (Ireland) R-Users Group

https://www.meetup.com/Cork-Ireland-R-Users-Group/

Putting the pieces together

- Project structure
  - Naming conventions
    - Scripted workflows
      - Reproducible research

How is research presented?

Papers

Journal of Virology

Network Analysis of the Chronic Hepatitis C Virome Defines Hypervariable Region 1 Evolutionary Phenotypes in the Context of Humoral Immune Responses

Brendan A. Palmer,[a] Daniel Schmidt-Martin,[a] Zoya Dimitrova,[b] Pavel Skums,[b] Orla Crosbie,[c] Elizabeth Kenny-Walsh,[c] Liam J. Fanning[a]

Molecular Virology Diagnostic & Research Laboratory, Department of Medicine, University College Cork, Cork, Ireland[a]; Division of Viral Hepatitis, Centers for Disease Control and Prevention, Atlanta, Georgia, USA[b]; Department of Hepatology, Cork University Hospital, Cork, Ireland[c]

ABSTRACT

Hypervariable region 1 (HVR1) of hepatitis C virus (HCV) comprises the first 27 N-terminal amino acid residues of E2. It is classically seen as the most heterogeneous region of the HCV genome. In this study, we assessed HVR1 evolution by using ultradeep pyrosequencing for a cohort of treatment-naive, chronically infected patients over a short, 16-week period. Organization of the sequence set into connected components that represented single nucleotide substitution events revealed a network dominated by highly connected, centrally positioned master sequences. HVR1 phenotypes were observed to be under strong purifying (stationary) and strong positive (antigenic drift) selection pressures, which were coincident with advancing patient age and cirrhosis of the liver. It followed that stationary viromes were dominated by a single HVR1 variant surrounded by minor variants comprised from conservative single amino acid substitution events. We present evidence to suggest that neutralization antibody efficacy was diminished for stationary-virome HVR1 variants. Our results identify the HVR1 network structure during chronic infection as the preferential dominance of a single variant within a narrow sequence space.

IMPORTANCE

HCV infection is often asymptomatic, and chronic infection is generally well established in advance of initial diagnosis and subsequent treatment. HVR1 can undergo rapid sequence evolution during acute infection, and the variant pool is typically seen to diverge away from ancestral sequences as infection progresses from the acute to the chronic phase. In this report, we describe HVR1 viromes in chronically infected patients that are defined by a dominant epitope located centrally within a narrow variant pool. Our findings suggest that weakened humoral immune activity, as a consequence of persistent chronic infection, allows for the acquisition and maintenance of host-specific adaptive mutations at HVR1 that reflect virus fitness.

Hepatitis C virus (HCV) infection is a global health issue and is recognized as a major etiological agent of liver-related diseases (1). It has been estimated that the current prevalence of HCV represents approximately 2% of the global adult (15 years of age and older) population (2). Following transmission, HCV infection may remain asymptomatic for decades, resulting in the majority of infections initially passing undetected (3). It is estimated that up to 4 million Americans are living with the virus, the majority of whom became infected prior to the isolation and identification of the virus (4, 5). Consequently, the U.S. Centers for Disease Control and Prevention now recommend that Americans born from 1945 to 1965 be screened for the presence of the virus notwithstanding the presence of clinical symptoms (3, 5).

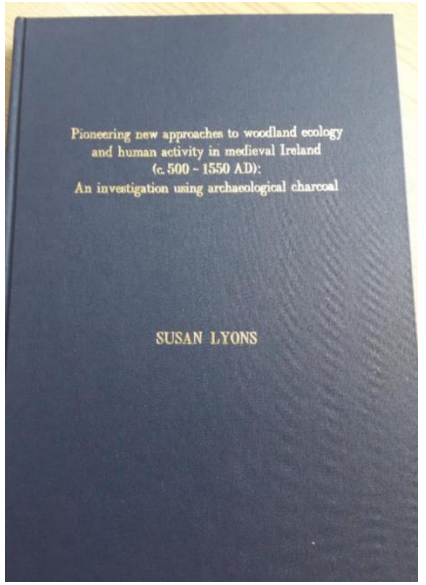
HCV is a single-stranded positive-sense RNA virus of considerable genomic heterogeneity. A recent reclassification defined the HCV global distribution into 7 genotypes and 67 subtypes, with genotypes 1 and 3 accounting for the majority of infections worldwide (6, 7). An error-prone RNA-dependent RNA polymerase, together with an inherent tolerance of defined hypervariable regions (HVR), accounts for much of this variability. Three HVRs are located within the envelope glycoprotein E2. The greatest heterogeneity has been identified at the 27-amino-acid HVR1 (residues 384 to 410 of the H77 reference strain), located at the amino-terminal end of the E2 glycoprotein (8). Recent studies indicated that the central region of E2 (residues 456 to 656) is globular and surprisingly compact, whereas the first 80 amino acids (including HVR1) lack this structural rigidity (9). This observation is consistent with a region that is proposed to shield conserved neutralizing epitopes and to participate in high-density lipoprotein enhancement of infection via scavenger receptor class B type I (SRBI) interactions and is itself targeted by neutralizing antibodies (nAb) (10–16).

Mutational flexibility at HVR1 was characterized soon after the initial identification of HCV (8, 17). Rapid mutational change of HVR1 has been documented over weeks during the acute phase of infection, where HVR1 evolution is governed predominantly by strong selective pressures, with fixation of beneficial mutations (11, 18, 19). Reports examining samples collected over years to decades have documented the emergence of convergent HVR1

Received 25 November 2015 Accepted 22 December 2015
Accepted manuscript posted online 30 December 2015
Citation Palmer BA, Schmidt-Martin D, Dimitrova Z, Skums P, Crosbie O, Kenny-Walsh E, Fanning LJ. 2016. Network analysis of the chronic hepatitis C virome defines hypervariable region 1 evolutionary phenotypes in the context of humoral immune responses. J Virol 90:3318–3329. doi:10.1128/JVI.02995-15.
Editor M. S. Diamond
Address correspondence to Liam J. Fanning, l.fanning@ucc.ie.
B.A.P. and D.S.-M. contributed equally to this article.

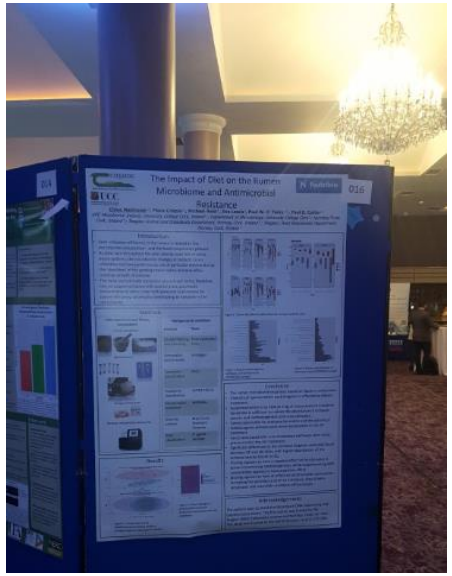
Books

Theses



Talks

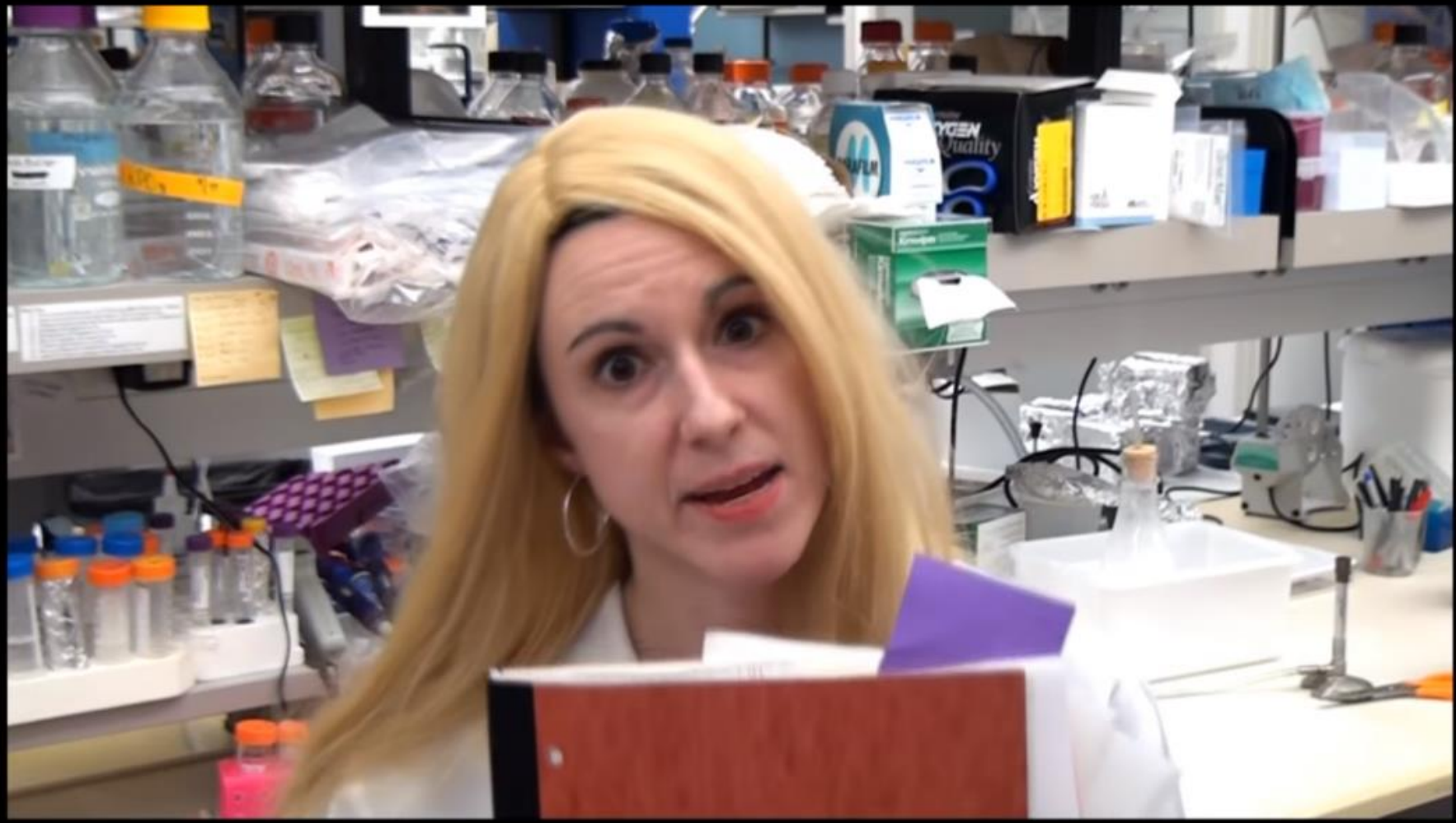
Posters



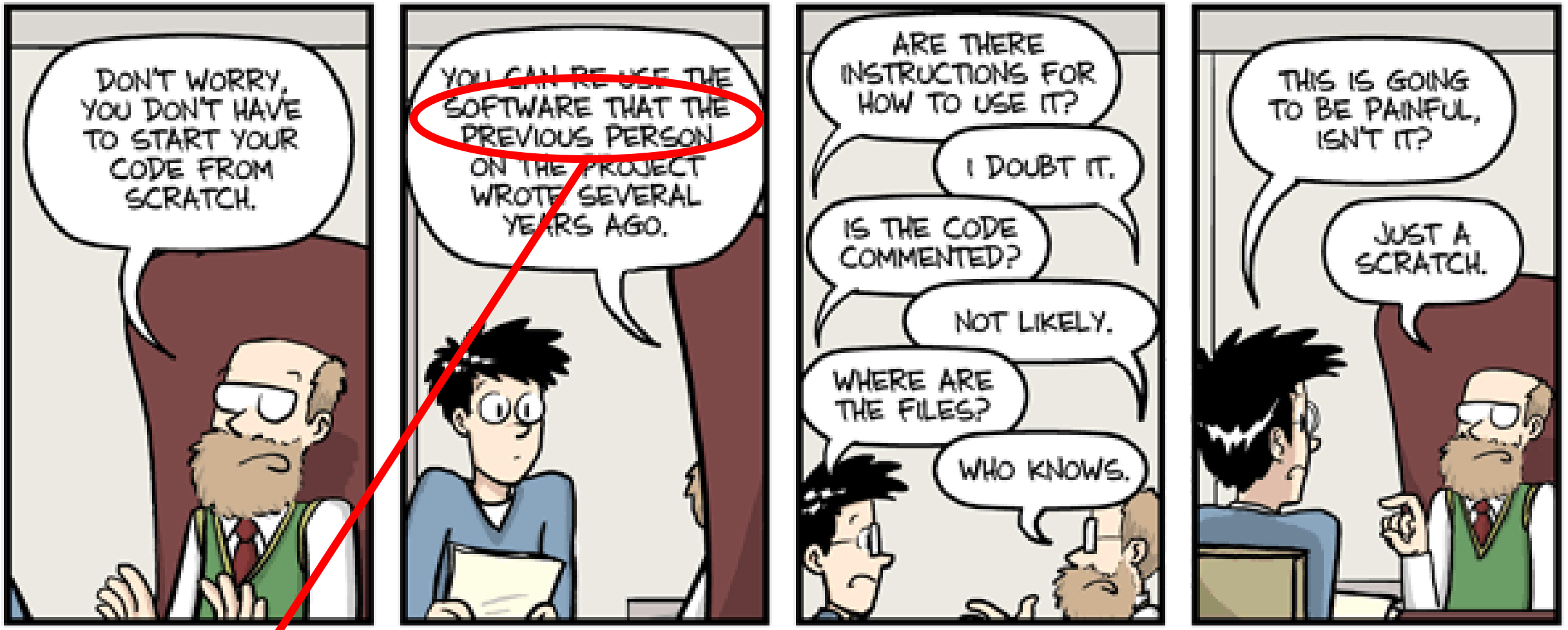
But what does it look like under the bonnet?

www.youtube.com/watch?v=Fl4L4M8m4d0

www.youtube.com/watch?v=Fl4L4M8m4d0



www.youtube.com/watch?v=Fl4L4M8m4d0



THIS PERSON IS likely to be YOU BTW!!

B-E: Disclaimer

Jenny Bryan
@JennyBryan

Software engineer @rstudio, humane #rstats, adjunct prof @UBC where I created @STAT545, part of @ropensci

Hadley Wickham
@hadleywickham

R, data, visualisation.

Houston, TX

hadley.nz

HADLEY WICKHAM | TEACHING | CODE | PERSONAL

I also teach in person workshops from time-to-time; see the RStudio workshops page for more details.

CODE

Most of my work is in the form of open source R code, which you can find on my github. You can roughly divide my work into three categories: tools for data science, tools for data import, and software engineering tools.

DATA SCIENCE

- ggplot2 for visualising data.
- dplyr for manipulating data.
- tidyr for tidying data.
- stringr for working with strings.
- lubridate for working with date/times.

DATA IMPORT

- readr for reading .csv and fwf files.
- readxl for reading .xls and .xlsx files.
- haven for SAS, SPSS, and Stata files.
- httr for talking to web APIs.
- rvest for scraping websites.
- xml2 for importing XML files.

SOFTWARE ENGINEERING

- devtools for general package development.
- roxygen2 for in-line documentation.
- testthat for unit testing

David Robinson
@drob

Chief Data Scientist at @DataCamp, #rstats fan/evangelist

New York, NY

varianceexplained.org

STAT 545 | Home | FAQ | Syllabus | Topics | People

Data wrangling, exploration, and analysis with R

UBC STAT 545A and 547M

Learn how to

- explore, groom, visualize, and analyze data
- make all of that reproducible, reusable, and shareable
- using R

VARIANCE EXPLAINED | ABOUT ME | POSTS | LEARN R | TEXT MINING IN R | INTRODUCTION TO EMPIRICAL BAYES

David Robinson

Chief Data Scientist at DataCamp, works in R and Python.

Email
Twitter
Github
Stack Overflow

This is the homepage and blog of David Robinson, Chief Data Scientist at DataCamp. For more about me, see here.

Recent Posts

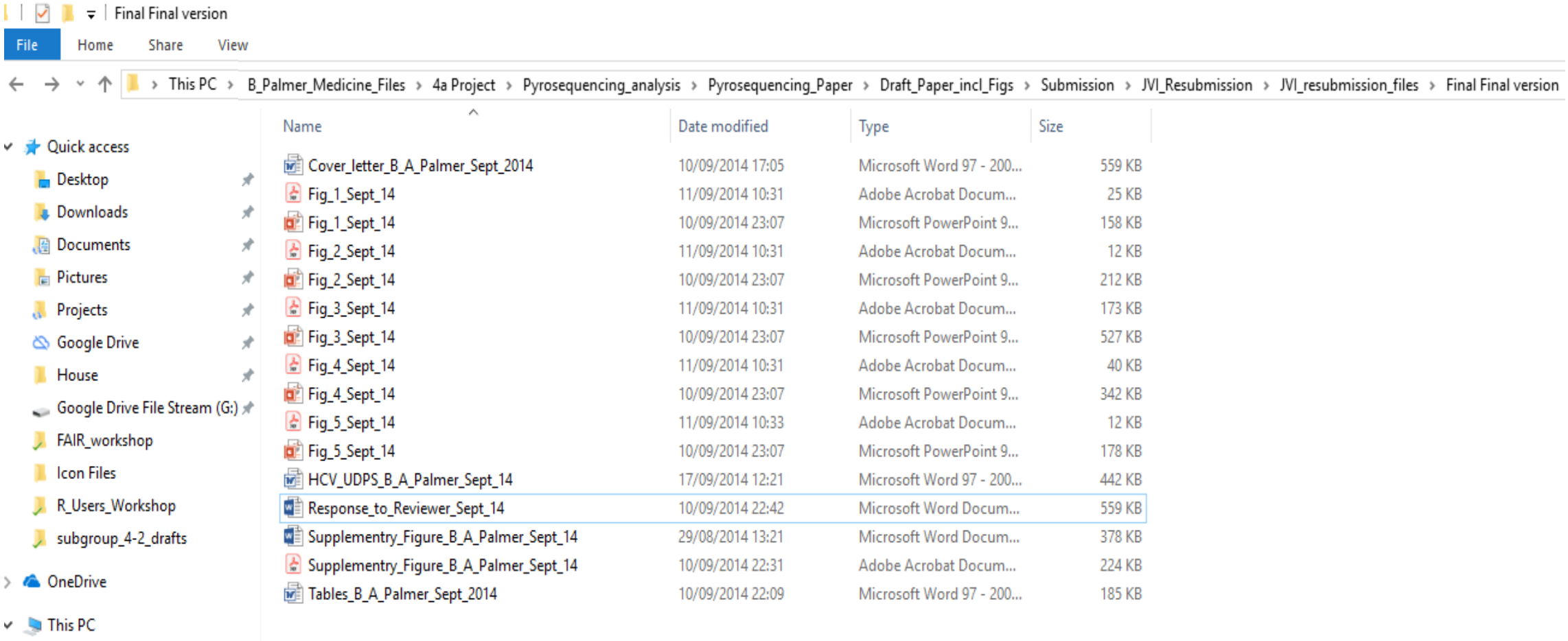
Exploring college major and income: a live data analysis in R — *October 16, 2018*
A live screencast of an exploratory data analysis from the Tidy Tuesday series. This one explores college major and income data from 538.

Who wrote the anti-Trump New York Times op-ed? Using tidytext to find document similarity — *September 06, 2018*
An analysis of an anonymous op-ed in the New York Times, using document similarity metrics to match it to Twitter accounts.

Scientific debt — *May 10, 2018*
Introducing an analogy to 'technical debt' for data scientists.

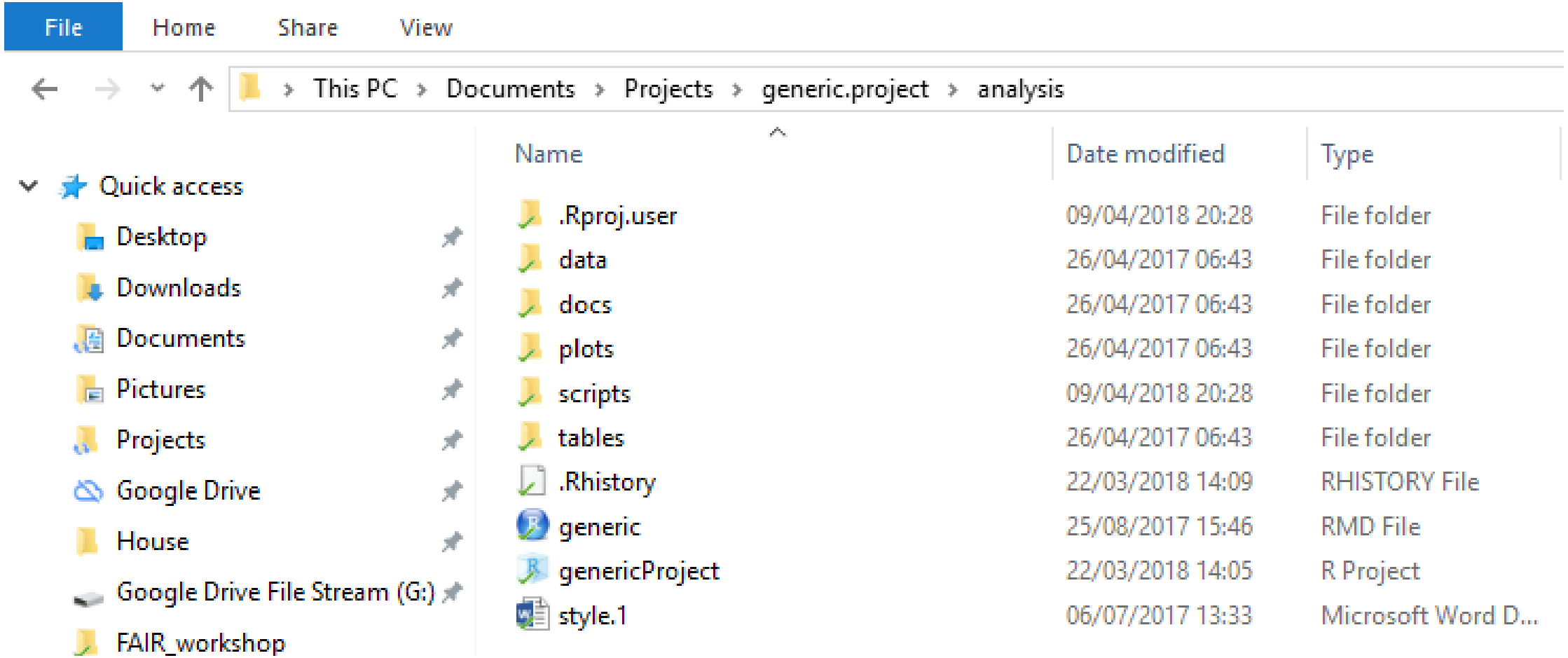
A: Still haven't found what I'm looking for

- **Help your future-self**



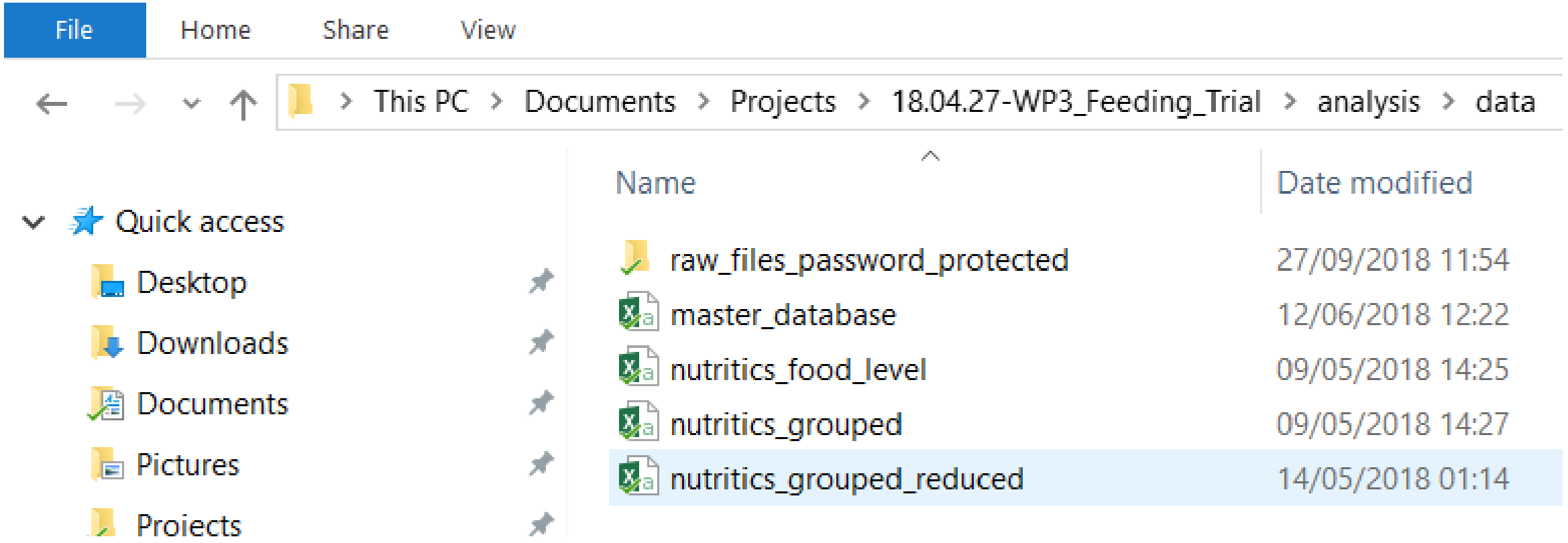
Define a generic project structure

- **STEP 1: Give your research projects a shared structure**



Give your files informative names

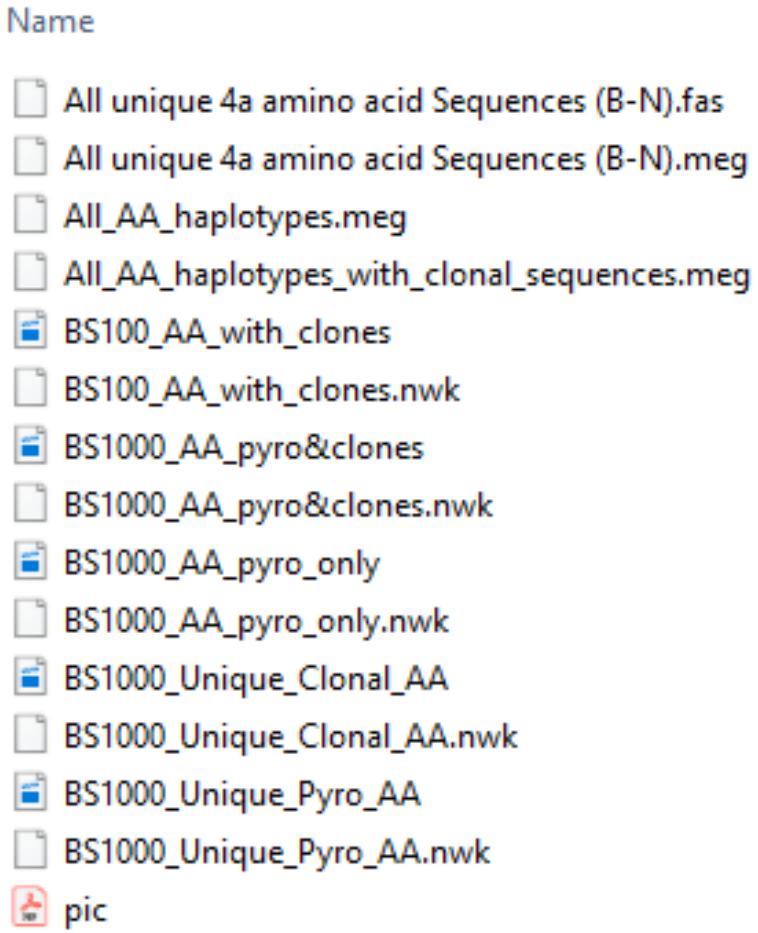
- **STEP 1: Give your research projects a shared structure**



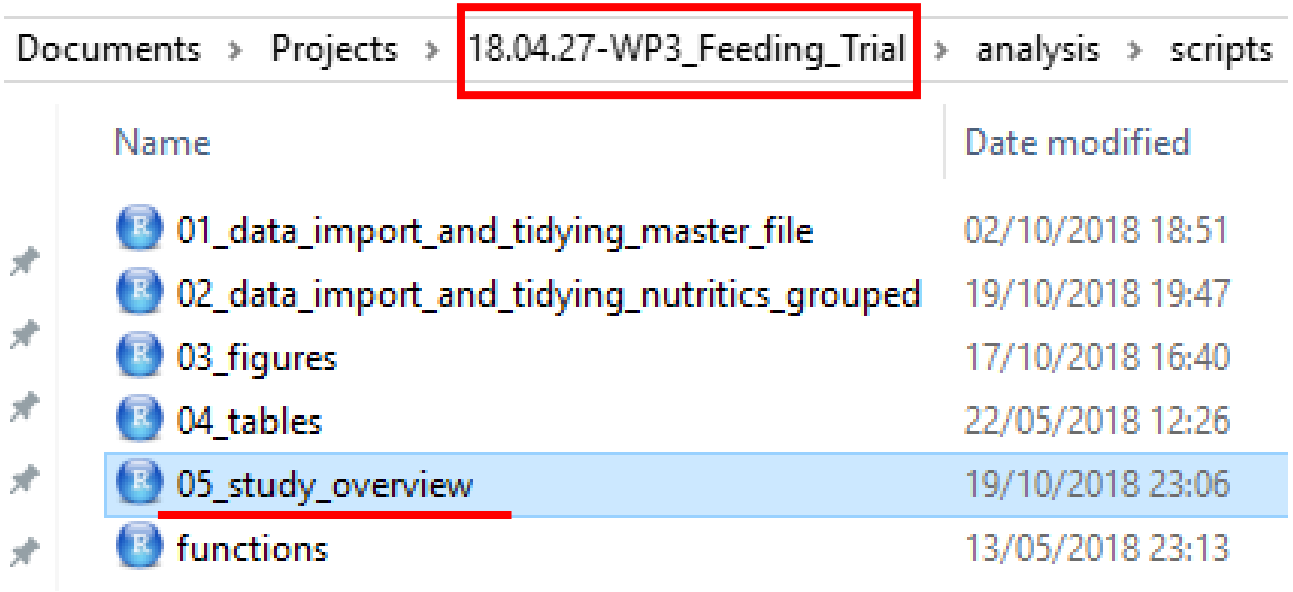
Everything in its right place

- STEP 2: Make you file names machine readable, human readable and work with default ordering

NO



Yes



Outline a file naming convention

Machine readable:

- Inherent order
- Avoid spaces
- Avoid punctuation
- Remove case-sensitivity

Human readable:

- Contains info on content
- Avoid spaces
- Avoid punctuation
- Remove case sensitivity

Metadata:

Separate with underscores ("_")

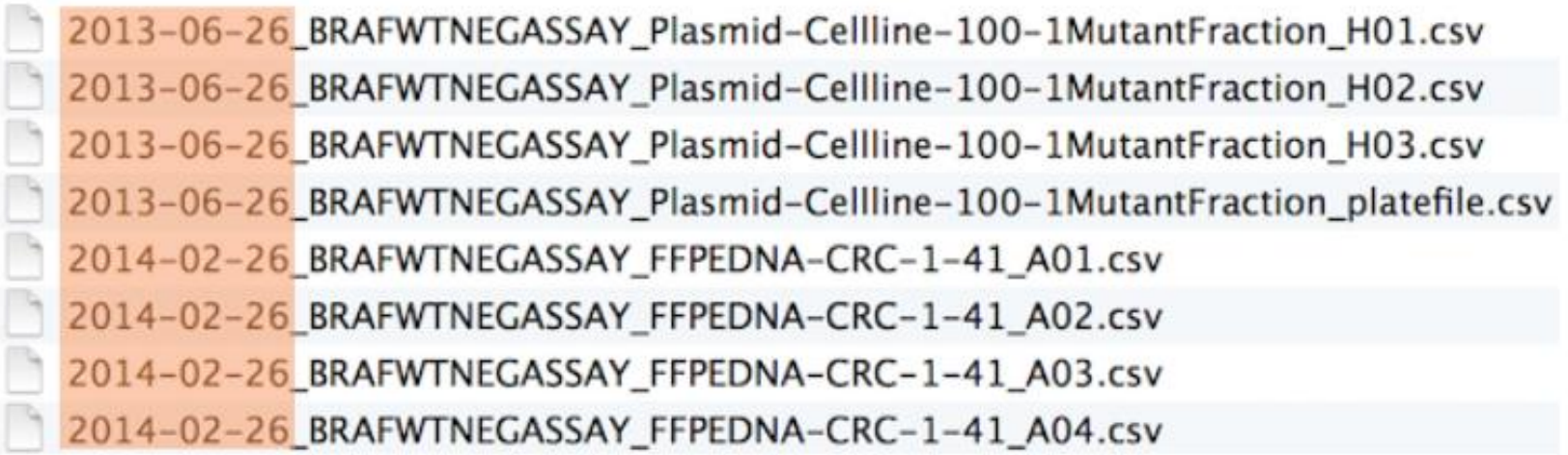
- Avoid punctuation
- Remove case-sensitivity

```
01_marshal-data.r

02_pre-dea-filtering.r

03_dea-with-limma-voom.r

04_explore-dea-results.r

90_limma-model-term-name-fiasco.r

helper01_load-counts.r

helper02_load-exp-des.r

helper03_load-focus-statinf.r

helper04_extract-and-tidy.r
```

datacarpentry.org/rr-organization1/01-file-naming/

Outline a file naming convention

Chronological order:



Logical order:

```
01_marshal-data.r
02_pre-dea-filtering.r
03_dea-with-limma-voom.r
04_explore-dea-results.r
90_limma-model-term-name-fiasco.r
helper01_load-counts.r
helper02_load-exp-des.r
helper03_load-focus-statinf.r
helper04_extract-and-tidy.r
```

Joined up thinking

- The R scripts you generate should be human readable
    - Annotate the code
    - Break up the scripts into dedicated tasks
    - Interlink with other within project scripts

```
1  # Data ----
2  # Eight tibbles returned from the 01_data_import_and_tidying_master_file.R
3  # 1. fgf23_data => FGF23 readings from study centres 01-03
4  # 2. food_level_data => Food diary entries
5  # 3. grouped_data => Dialysis and nondialysis diary entries by component
6  # 4. k_data => Serum potassium
7  # 5. master_data_clean => all the clean master file data if required
8  # 6. p_data => Serum phosphate
9  # 7. pth_data => Parathyroid hormone readings
10 # 8. pulses_nuts_data
11
12 source("scripts/01_data_import_and_tidying_master_file.R")
```

Is too much choice good or bad?

Toothpaste & Jam: The Psychology of Choice

Something as simple as buying toothpaste can be overwhelming. Do you want the anti-tartar kind or the cavity-busting option? Sensitive enamel protection or the one with whitening? Fluoride, non-fluoride? Then there's flavor: crystal mint, intense mint, fresh mint or sparkling mint – and that's *just* mint.

Inconsistent function names, inconsistent syntax

- R is a very versatile language
  - Sometimes it can be too versatile
  - Do you want to use……

```
Names or colnames
row.names or rownames
rowSums or rowsum
Sys.time, system.time
```

  - Is it written as……

```
newobject or new.Object
x = 5 or x <- 5
mapping=aes(x,y) or mapping = aes(x, y)
```

Variable selection

```
summary(starwars$name)

summary(starwars$"name")

summary(starwars["name"])

summary(starwars[,"name"])

summary(starwars[1])

summary(starwars[,1])

summary(starwars[[1]])
```

- Open the script 01_too_much.choice.R

Writing clearer code

- Annotation

- Object names
  - should use only lowercase letters, numbers, and "_"

- Spacing
  - Put a space before and after =
  - Put a space after a ,
  - Operators should be surrounded by spaces e.g. ==, <-, +

- For a more complete list visit
  - http://style.tidyverse.org/syntax.html

- Open the script 02_good_habits.R

B: Everything in its right place

- To finish up, we're going to explore the benefits of using R projects for our data analysis tasks

- Open the script 03_project_clean_data.R

- Open the script 04_project_project_data.R

Don’t Do What Donny Don’t Does!!

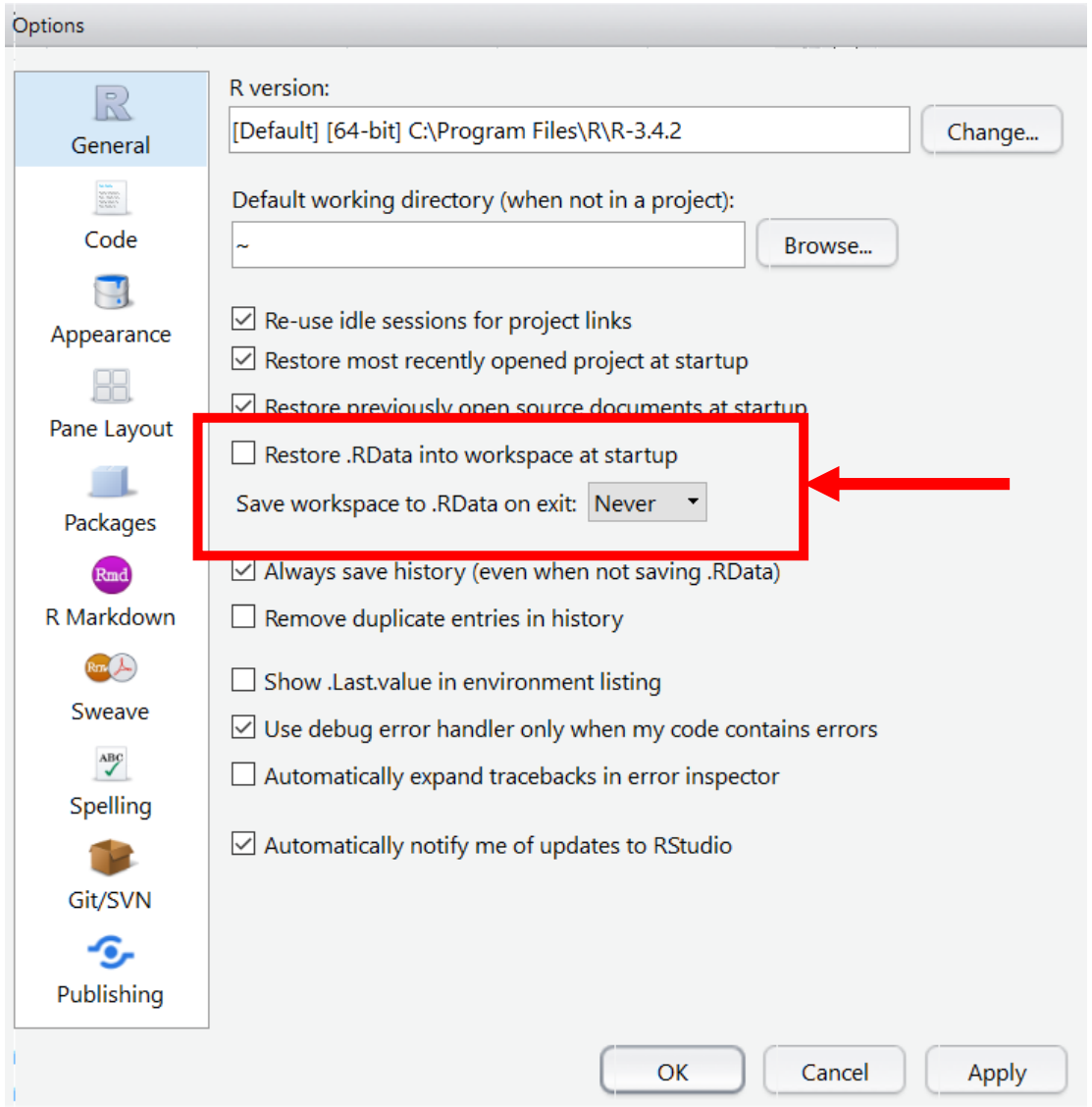
Donny Don’t:

- Start your script with…
  `setwd()`

Donny Don’t:

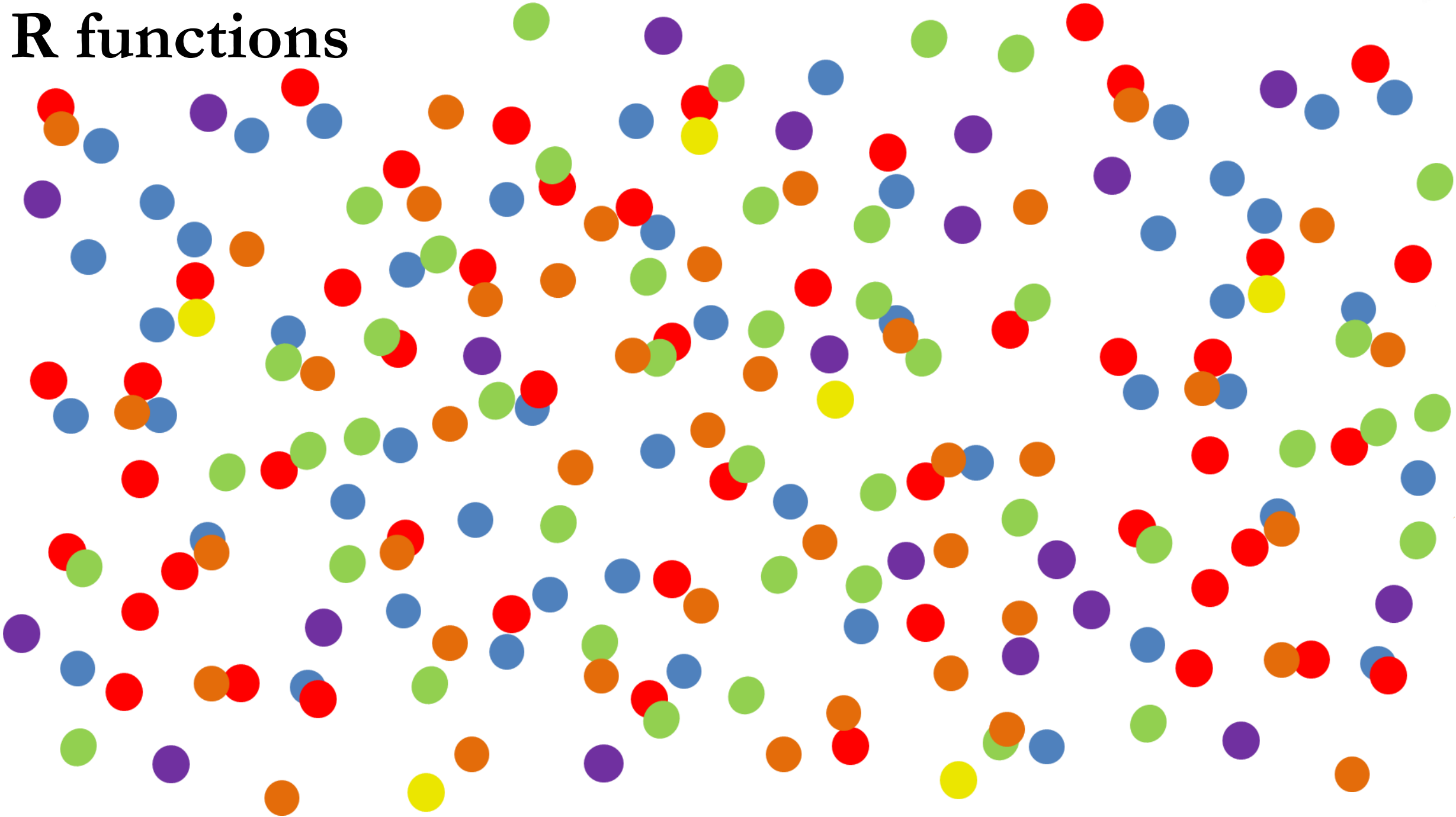
- Start your script with…
  `rm(list = ls())`

Other points to note

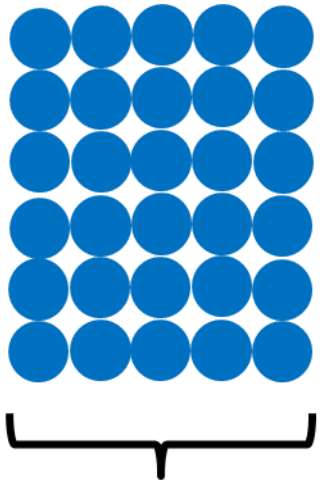


- You might consider your environment as "real"
- If you continue to use R, it is better for you to consider your R scripts as "real", as these should recreate the environment
- You may suffer short term pain
- This will prevent long term agony

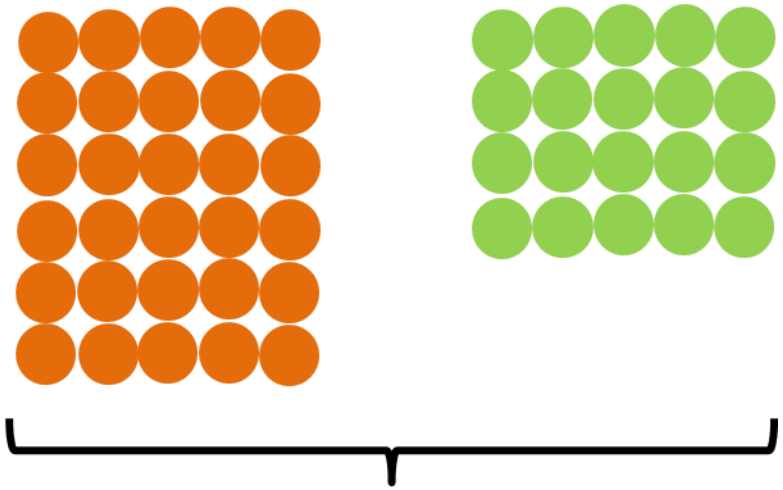
R functions



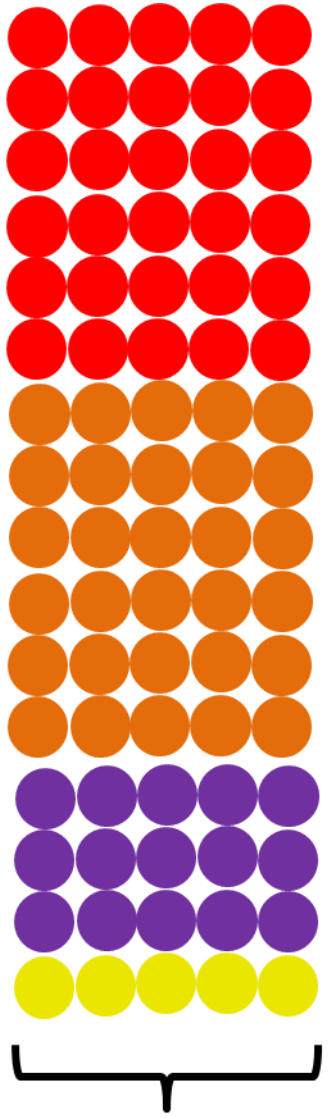
R packages



Base R:
Comes
pre-
loaded

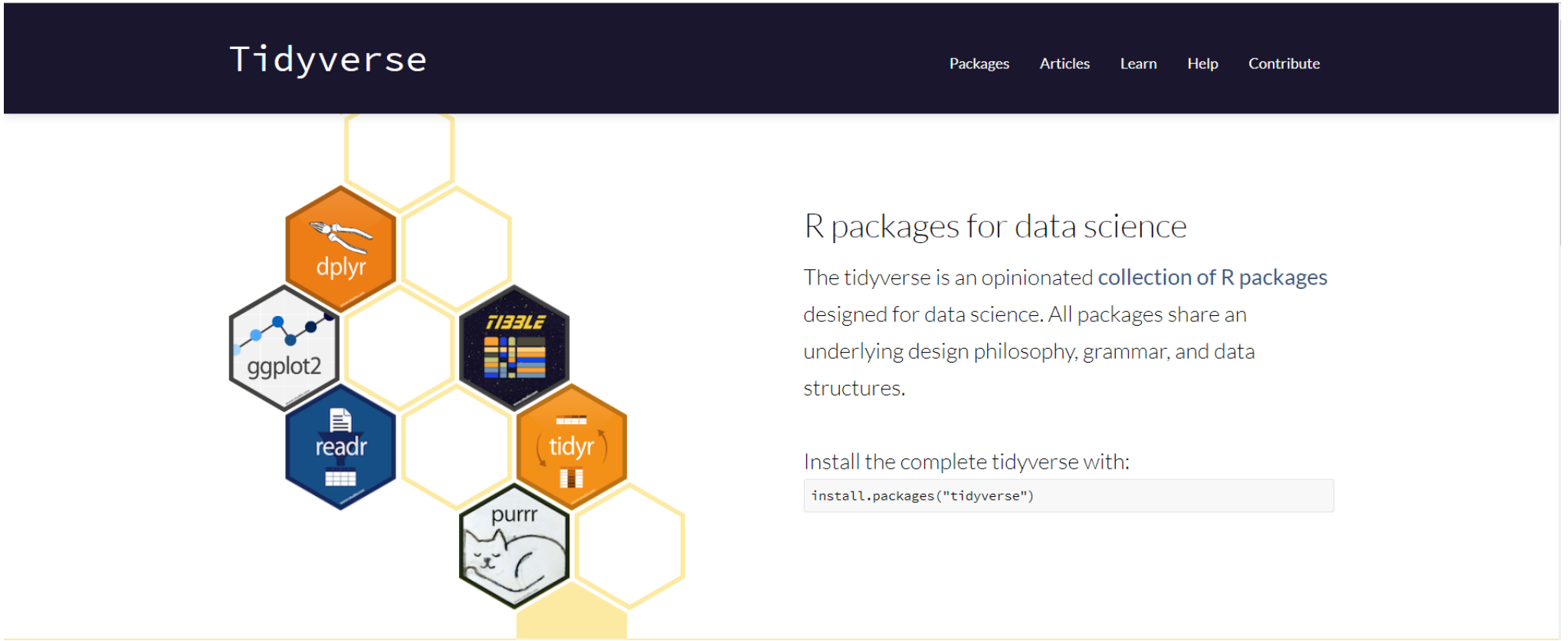


Other packages:
Install once
Update regularly
Load each session



core
tidyverse

What is the tidyverse?



- Joined up collection of packages for data analysis
    - Consistent functions
    - Uses (tidy) data
    - Supports end-to-end workflows

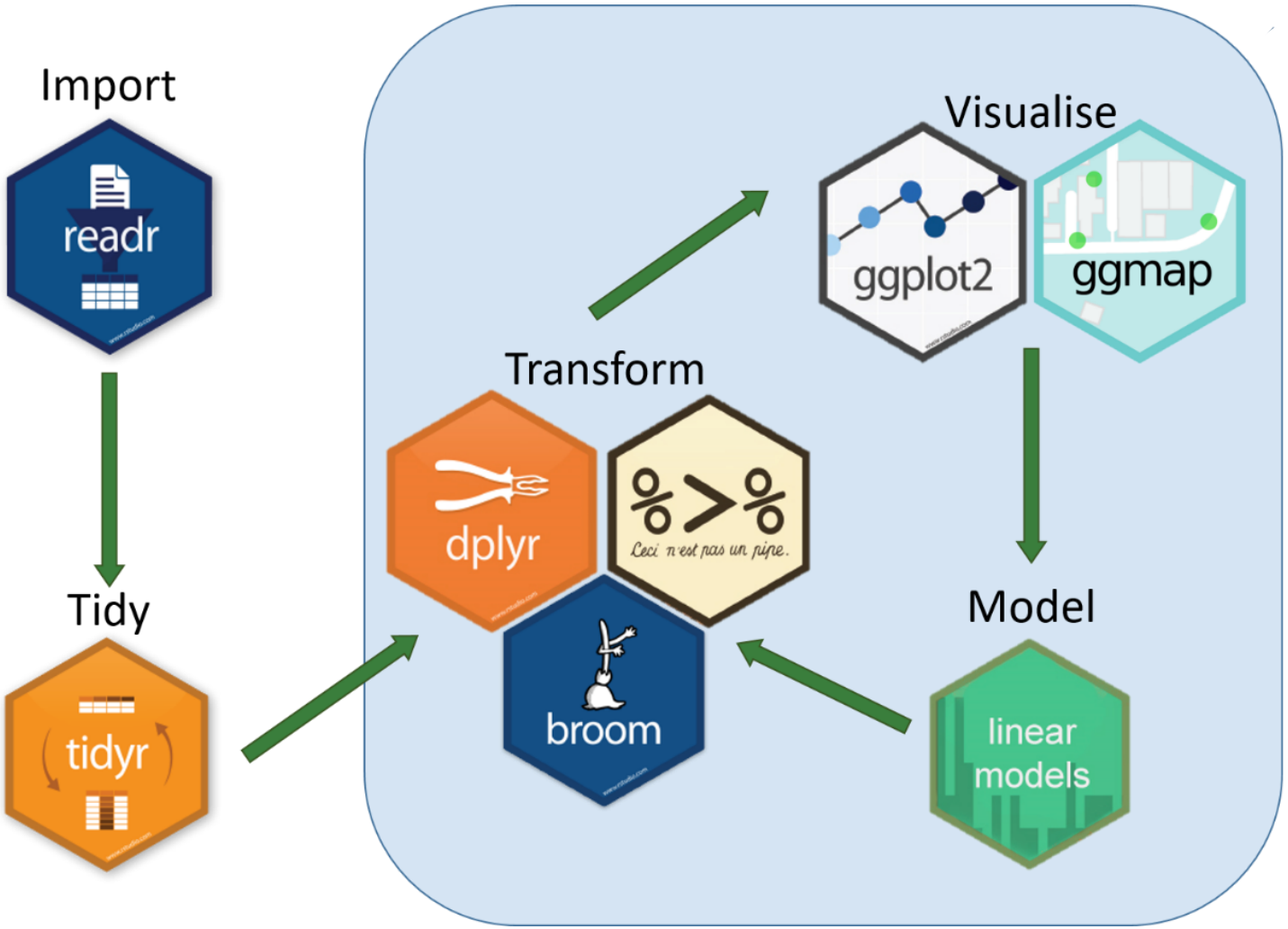
What is the tidyverse?

```
> install.packages(c("broom", "cli2", "crayon",
"dbplyr", "dplyr", "forcats", "ggplot2", "haven",
"hms", "httr", "jsonlite", "lubridate",
"magrittr", "modelr", "pillar", "purrr", "readr",
"readxl", "reprex", "rlang", "rstudioapi",
"rvest", "stringr", "tibble", "tidyr", "xml2")
```

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

Putting the pieces together

- Data analysis in a tidyverse nutshell



Tidyverse works best with tidy data

- Each variable forms a column
- Each observation forms a row

Problems with Brauer et al., data…

Column headers contain values

Multiple variables 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 - letter identifies a compound
      - number is the concentration of that compound

Code structure v1

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

```
separated_gene                  - the new tibble you will create

<-                              - the assign operator

separate                        - the function you are calling on

(raw_gene,                      - the tibble 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
```

Filter

	GID	YORF	NAME	GWEIGHT	G0.05	G0.1	G0.15	G0.2	G0.25	G0.3
1	GENE1331X	A_06_P5820	SFB2 || ER to Golgi transport || molecular function unknown || YNL049C || 108...	1	-0.24	-0.13	-0.21	-0.15	-0.05	-0.05
2	GENE4924X	A_06_P5866	|| biological process unknown || molecular function unknown || YNL095C || 1...	1	0.28	0.13	-0.40	-0.48	-0.11	0.17
3	GENE4690X	A_06_P1834	QRI7 || proteolysis and peptidolysis || metalloendopeptidase activity || YDL104...	1	-0.02	-0.27	-0.27	-0.02	0.24	0.25
4	GENE1177X	A_06_P4928	CFT2 || mRNA polyadenylylation* || RNA binding || YLR115W || 1081958	1	-0.33	-0.41	-0.24	-0.03	-0.03	0.00
5	GENE511X	A_06_P5620	SSO2 || vesicle fusion* || t-SNARE activity || YMR183C || 1081214	1	0.05	0.02	0.40	0.34	-0.13	-0.14
6	GENE2133X	A_06_P5307	PSP2 || biological process unknown || molecular function unknown || YML01...	1	-0.69	-0.03	0.23	0.20	0.00	-0.27
7	GENE1002X	A_06_P6258	RIB2 || riboflavin biosynthesis || pseudouridylate synthase activity* || YOL066C...	1	-0.55	-0.30	-0.12	-0.03	-0.16	-0.11
8	GENE5478X	A_06_P7082	VMA13 || vacuolar acidification || hydrogen-transporting ATPase activity, rota...	1	-0.75	-0.12	-0.07	0.02	-0.32	-0.41
9	GENE2065X	A_06_P2554	EDC3 || deadenylylation-independent decapping || molecular function unkno...	1	-0.24	-0.22	0.14	0.06	0.00	-0.13
10	GENE2440X	A_06_P6431	VPS5 || protein retention in Golgi* || protein transporter activity || YOR069W || ...	1	-0.16	-0.38	0.05	0.14	-0.04	-0.01
11	GENE4180X	A_06_P6220	|| biological process unknown || molecular function unknown || YOL029C || 1...	1	-0.22	-0.18	0.27	0.18	0.03	-0.04
12	GENE5247X	A_06_P1410	AMN1 || negative regulation of exit from mitosis* || protein binding || YBR158...	1	0.18	0.61	1.55	1.34	0.23	-0.03
13	GENE2121X	A_06_P2983	SCW11 || cytokinesis, completion of separation || glucan 1,3-beta-glucosidas...	1	-0.67	-0.47	1.16	1.05	-0.18	-0.68
14	GENE1985X	A_06_P3720	DSE2 || cell wall organization and biogenesis* || glucan 1,3-beta-glucosidase ...	1	-0.59	-0.17	1.17	0.85	-0.12	-0.61
15	GENE4728X	A_06_P2774	COX15 || cytochrome c oxidase complex assembly* || oxidoreductase activity,...	1	-0.28	-0.81	-0.39	0.24	0.01	0.01
16	GENE3153X	A_06_P4597	SPE1 || pantothenate biosynthesis* || ornithine decarboxylase activity || YKL18...	1	-0.19	0.24	0.03	0.17	0.00	-0.01
17	GENE3704X	A_06_P5667	MTF1 || transcription from mitochondrial promoter || S-adenosylmethionine-...	1	-0.42	-0.43	-0.36	-0.12	0.05	0.24
18	GENE2141X	A_06_P3260	KSS1 || invasive growth (sensu Saccharomyces)* || MAP kinase activity || YGR...	1	-0.76	-0.32	-0.05	-0.27	-0.31	-0.01
19	GENE2978X	A_06_P3607	|| biological process unknown || molecular function unknown || YHR036W || 1...	1	-0.91	-0.43	-0.05	-0.09	-0.27	-0.45
20	GENE1203X	A_06_P5929	|| biological process unknown || molecular function unknown || YNL158W || 1...	1	-0.47	-0.43	-0.15	0.08	-0.26	-0.25

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

Line by line

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | workshop_1_project — 8_weeks_Oct-Dec_17

ws1_script1_stepwise_Bauer_dataset_an... *

Source on Save | Run | Source

```
12
13  raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")
14
15  separated_gene_df <- separate(raw_gene_df, NAME,
16                                c("name", "BP", "MF", "systematic_name",
17                                  "number"),
18                                sep = "\\|\\|")
19
20  mutated_gene_df <- mutate_at(separated_gene_df,
21                               vars(name:systematic_name),
22                               funs(trimws)
23                               )
24
25  selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
26
27  gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
28
29  nearly_there_df <- separate(gathered_gene_df, sample,
30                              c("nutrient", "rate"), sep = 1, convert = TRUE)
31
32  nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
33                      S = "Sulfate", N = "Ammonia", U = "Uracil")
34
35  cleaned_genes_df <- mutate(nearly_there_df,
36                             nutrient = plyr::revalue(nutrient, nutrient_names)
37                             ) %>%
38    filter(!is.na(expression), systematic_name != "")
```

15:1 | Section 1: Data import, tidying and transformation | R Script

Console Terminal

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project/

```
> raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")
Parsed with column specification:
cols(
  .default = col_double(),
  GID = col_character(),
  YORF = col_character(),
  NAME = col_character(),
  GWEIGHT = col_integer()
)
See spec(...) for full column specifications.
>
```

Environment History Connections

Import Dataset | Grid

Global Environment

Name	Type	Length	Size	Value
raw_gene_df	tbl_df	40	3.3 MB	5537 obs. of 40 variables

Files Plots Packages Help Viewer

New Folder | Delete | Rename | More

Home > R_Users_Workshop > 8_weeks_Oct-Dec_17 > Workshop_1 > workshop_1_project

Name	Size	Modified
..		
.RData	2.5 KB	Oct 2, 2017, 1:49 PM
.Rhistory	20.3 KB	Dec 6, 2017, 3:43 PM
Brauer2008_DataSet1.csv	1.6 MB	Sep 27, 2017, 11:32 PM
Brauer2008_DataSet1.tds	1.6 MB	Sep 28, 2017, 10:22 AM
house_completions.csv	4 KB	Sep 28, 2017, 1:35 PM
irish_population.csv	315 B	Aug 28, 2017, 4:21 PM
raw_house_completions.csv	16.2 KB	Aug 25, 2017, 3:45 PM
workshop_1.Rproj	217 B	Oct 18, 2018, 12:18 PM
ws1_script1_stepwise_Bauer_dataset_analysis.R	6.1 KB	Dec 5, 2017, 12:19 PM
ws1_script2_Bauer_dataset_analysis.R	2 KB	Dec 6, 2017, 2:33 PM
ws1_script3_house_completions.R	2.4 KB	Oct 2, 2017, 3:53 PM

Line by line

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | workshop_1_project — 8_weeks_Oct-Dec_17

ws1_script1_stepwise_Bauer_dataset_an... *

Source on Save | Run | Source

```
12
13  raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")
14
15  separated_gene_df <- separate(raw_gene_df, NAME,
16                                c("name", "BP", "MF", "systematic_name",
17                                  "number"),
18                                sep = "\\|\\|")
19
20  mutated_gene_df <- mutate_at(separated_gene_df,
21                               vars(name:systematic_name),
22                               funs(trimws)
23                               )
24
25  selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
26
27  gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
28
29  nearly_there_df <- separate(gathered_gene_df, sample,
30                              c("nutrient", "rate"), sep = 1, convert = TRUE)
31
32  nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
33                      S = "Sulfate", N = "Ammonia", U = "Uracil")
34
35  cleaned_genes_df <- mutate(nearly_there_df,
36                             nutrient = plyr::revalue(nutrient, nutrient_names)
37                             ) %>%
38    filter(!is.na(expression), systematic_name != "")
```

20:1 | Section 1: Data import, tidying and transformation | R Script

Console Terminal

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project/

```
> raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")
Parsed with column specification:
cols(
  .default = col_double(),
  GID = col_character(),
  YORF = col_character(),
  NAME = col_character(),
  GWEIGHT = col_integer()
)
See spec(...) for full column specifications.
> separated_gene_df <- separate(raw_gene_df, NAME,
+                               c("name", "BP", "MF", "systematic_name",
+                                 "number"),
+                               sep = "\\|\\|")
>
```

Environment History Connections

Import Dataset | Grid

Global Environment

Name	Type	Length	Size	Value
raw_gene_df	tbl_df	40	3.3 MB	5537 obs. of 40 variables
separated_gene...	tbl_df	44	3.6 MB	5537 obs. of 44 variables

Files Plots Packages Help Viewer

New Folder | Delete | Rename | More

Home > R_Users_Workshop > 8_weeks_Oct-Dec_17 > Workshop_1 > workshop_1_project

Name	Size	Modified
..		
.RData	2.5 KB	Oct 2, 2017, 1:49 PM
.Rhistory	20.3 KB	Dec 6, 2017, 3:43 PM
Brauer2008_DataSet1.csv	1.6 MB	Sep 27, 2017, 11:32 PM
Brauer2008_DataSet1.tds	1.6 MB	Sep 28, 2017, 10:22 AM
house_completions.csv	4 KB	Sep 28, 2017, 1:35 PM
irish_population.csv	315 B	Aug 28, 2017, 4:21 PM
raw_house_completions.csv	16.2 KB	Aug 25, 2017, 3:45 PM
workshop_1.Rproj	217 B	Oct 18, 2018, 12:18 PM
ws1_script1_stepwise_Bauer_dataset_analysis.R	6.1 KB	Dec 5, 2017, 12:19 PM
ws1_script2_Bauer_dataset_analysis.R	2 KB	Dec 6, 2017, 2:33 PM
ws1_script3_house_completions.R	2.4 KB	Oct 2, 2017, 3:53 PM

Line by line

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | workshop_1_project — 8_weeks_Oct-Dec_17

ws1_script1_stepwise_Bauer_dataset_an... *

Source on Save | Run | Source

```
12
13  raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")
14
15  separated_gene_df <- separate(raw_gene_df, NAME,
16                                c("name", "BP", "MF", "systematic_name",
17                                  "number"),
18                                sep = "\\|\\|")
19
20  mutated_gene_df <- mutate_at(separated_gene_df,
21                               vars(name:systematic_name),
22                               funs(trimws)
23                               )
24
25  selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
26
27  gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
28
29  nearly_there_df <- separate(gathered_gene_df, sample,
30                              c("nutrient", "rate"), sep = 1, convert = TRUE)
31
32  nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
33                      S = "Sulfate", N = "Ammonia", U = "Uracil")
34
35  cleaned_genes_df <- mutate(nearly_there_df,
36                             nutrient = plyr::revalue(nutrient, nutrient_names)
37                             ) %>%
38    filter(!is.na(expression), systematic_name != "")
```

27:1 | Section 1: Data import, tidying and transformation | R Script

Console Terminal

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project/

```
Parsed with column specification:
cols(
  .default = col_double(),
  GID = col_character(),
  YORF = col_character(),
  NAME = col_character(),
  GWEIGHT = col_integer()
)
See spec(...) for full column specifications.
> separated_gene_df <- separate(raw_gene_df, NAME,
+                               c("name", "BP", "MF", "systematic_name",
+                                 "number"),
+                               sep = "\\|\\|")
> mutated_gene_df <- mutate_at(separated_gene_df,
+                              vars(name:systematic_name),
+                              funs(trimws)
+                              )
> selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
>
```

Environment History Connections

Import Dataset | Grid

Global Environment

Name	Type	Length	Size	Value
mutated_gene_df	tbl_df	44	3.5 MB	5537 obs. of 44 variables
raw_gene_df	tbl_df	40	3.3 MB	5537 obs. of 40 variables
selected_gene_df	tbl_df	40	2.4 MB	5537 obs. of 40 variables
separated_gene...	tbl_df	44	3.6 MB	5537 obs. of 44 variables

Files Plots Packages Help Viewer

New Folder | Delete | Rename | More

Home > R_Users_Workshop > 8_weeks_Oct-Dec_17 > Workshop_1 > workshop_1_project

Name	Size	Modified
..		
.RData	2.5 KB	Oct 2, 2017, 1:49 PM
.Rhistory	20.3 KB	Dec 6, 2017, 3:43 PM
Brauer2008_DataSet1.csv	1.6 MB	Sep 27, 2017, 11:32 PM
Brauer2008_DataSet1.tds	1.6 MB	Sep 28, 2017, 10:22 AM
house_completions.csv	4 KB	Sep 28, 2017, 1:35 PM
irish_population.csv	315 B	Aug 28, 2017, 4:21 PM
raw_house_completions.csv	16.2 KB	Aug 25, 2017, 3:45 PM
workshop_1.Rproj	217 B	Oct 18, 2018, 12:18 PM
ws1_script1_stepwise_Bauer_dataset_analysis.R	6.1 KB	Dec 5, 2017, 12:19 PM
ws1_script2_Bauer_dataset_analysis.R	2 KB	Dec 6, 2017, 2:33 PM
ws1_script3_house_completions.R	2.4 KB	Oct 2, 2017, 3:53 PM

Line by line

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | workshop_1_project — 8_weeks_Oct-Dec_17

ws1_script1_stepwise_Bauer_dataset_an... *

Source on Save | Run | Source

```
12
13  raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")
14
15  separated_gene_df <- separate(raw_gene_df, NAME,
16                                c("name", "BP", "MF", "systematic_name",
17                                  "number"),
18                                sep = "\\|\\|")
19
20  mutated_gene_df <- mutate_at(separated_gene_df,
21                               vars(name:systematic_name),
22                               funs(trimws)
23                               )
24
25  selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
26
27  gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
28
29  nearly_there_df <- separate(gathered_gene_df, sample,
30                              c("nutrient", "rate"), sep = 1, convert = TRUE)
31
32  nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
33                      S = "Sulfate", N = "Ammonia", U = "Uracil")
34
35  cleaned_genes_df <- mutate(nearly_there_df,
36                             nutrient = plyr::revalue(nutrient, nutrient_names)
37                             ) %>%
38    filter(!is.na(expression), systematic_name != "")
```

29:1 | Section 1: Data import, tidying and transformation | R Script

Console Terminal

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project/

```
cols(
  .default = col_double(),
  GID = col_character(),
  YORF = col_character(),
  NAME = col_character(),
  GWEIGHT = col_integer()
)
See spec(...) for full column specifications.
> separated_gene_df <- separate(raw_gene_df, NAME,
+                               c("name", "BP", "MF", "systematic_name",
+                                 "number"),
+                               sep = "\\|\\|")
> mutated_gene_df <- mutate_at(separated_gene_df,
+                              vars(name:systematic_name),
+                              funs(trimws)
+                              )
> selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
> gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
>
```

Environment History Connections

Import Dataset | Grid

Global Environment

Name	Type	Length	Size	Value
gathered_gene_df	tbl_df	6	9.8 MB	199332 obs. of 6 variables
mutated_gene_df	tbl_df	44	3.5 MB	5537 obs. of 44 variables
raw_gene_df	tbl_df	40	3.3 MB	5537 obs. of 40 variables
selected_gene_df	tbl_df	40	2.4 MB	5537 obs. of 40 variables
separated_gene...	tbl_df	44	3.6 MB	5537 obs. of 44 variables

Files Plots Packages Help Viewer

New Folder | Delete | Rename | More

Home > R_Users_Workshop > 8_weeks_Oct-Dec_17 > Workshop_1 > workshop_1_project

Name	Size	Modified
..		
.RData	2.5 KB	Oct 2, 2017, 1:49 PM
.Rhistory	20.3 KB	Dec 6, 2017, 3:43 PM
Brauer2008_DataSet1.csv	1.6 MB	Sep 27, 2017, 11:32 PM
Brauer2008_DataSet1.tds	1.6 MB	Sep 28, 2017, 10:22 AM
house_completions.csv	4 KB	Sep 28, 2017, 1:35 PM
irish_population.csv	315 B	Aug 28, 2017, 4:21 PM
raw_house_completions.csv	16.2 KB	Aug 25, 2017, 3:45 PM
workshop_1.Rproj	217 B	Oct 18, 2018, 12:18 PM
ws1_script1_stepwise_Bauer_dataset_analysis.R	6.1 KB	Dec 5, 2017, 12:19 PM
ws1_script2_Bauer_dataset_analysis.R	2 KB	Dec 6, 2017, 2:33 PM
ws1_script3_house_completions.R	2.4 KB	Oct 2, 2017, 3:53 PM

Line by line

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | workshop_1_project — 8_weeks_Oct-Dec_17

ws1_script1_stepwise_Bauer_dataset_an... *

Source on Save | Run | Source

```
12
13  raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")
14
15  separated_gene_df <- separate(raw_gene_df, NAME,
16                                c("name", "BP", "MF", "systematic_name",
17                                  "number"),
18                                sep = "\\|\\|")
19
20  mutated_gene_df <- mutate_at(separated_gene_df,
21                               vars(name:systematic_name),
22                               funs(trimws)
23                               )
24
25  selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
26
27  gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
28
29  nearly_there_df <- separate(gathered_gene_df, sample,
30                              c("nutrient", "rate"), sep = 1, convert = TRUE)
31
32  nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
33                      S = "Sulfate", N = "Ammonia", U = "Uracil")
34
35  cleaned_genes_df <- mutate(nearly_there_df,
36                             nutrient = plyr::revalue(nutrient, nutrient_names)
37                             ) %>%
38    filter(!is.na(expression), systematic_name != "")
```

32:1 | Section 1: Data import, tidying and transformation | R Script

Console Terminal

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project/

```
  GID = col_character(),
  YORF = col_character(),
  NAME = col_character(),
  GWEIGHT = col_integer()
)
See spec(...) for full column specifications.
> separated_gene_df <- separate(raw_gene_df, NAME,
+                               c("name", "BP", "MF", "systematic_name",
+                                 "number"),
+                               sep = "\\|\\|")
> mutated_gene_df <- mutate_at(separated_gene_df,
+                              vars(name:systematic_name),
+                              funs(trimws)
+                              )
> selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
> gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
> nearly_there_df <- separate(gathered_gene_df, sample,
+                             c("nutrient", "rate"), sep = 1, convert = TRUE)
> |
```

Environment History Connections

Import Dataset | Grid

Global Environment

Name	Type	Length	Size	Value
gathered_gene_df	tbl_df	6	9.8 MB	199332 obs. of 6 variables
mutated_gene_df	tbl_df	44	3.5 MB	5537 obs. of 44 variables
nearly_there_df	tbl_df	7	11.3 MB	199332 obs. of 7 variables
raw_gene_df	tbl_df	40	3.3 MB	5537 obs. of 40 variables
selected_gene_df	tbl_df	40	2.4 MB	5537 obs. of 40 variables
separated_gene...	tbl_df	44	3.6 MB	5537 obs. of 44 variables

Files Plots Packages Help Viewer

New Folder | Delete | Rename | More

Home > R_Users_Workshop > 8_weeks_Oct-Dec_17 > Workshop_1 > workshop_1_project

Name	Size	Modified
..		
.RData	2.5 KB	Oct 2, 2017, 1:49 PM
.Rhistory	20.3 KB	Dec 6, 2017, 3:43 PM
Brauer2008_DataSet1.csv	1.6 MB	Sep 27, 2017, 11:32 PM
Brauer2008_DataSet1.tds	1.6 MB	Sep 28, 2017, 10:22 AM
house_completions.csv	4 KB	Sep 28, 2017, 1:35 PM
irish_population.csv	315 B	Aug 28, 2017, 4:21 PM
raw_house_completions.csv	16.2 KB	Aug 25, 2017, 3:45 PM
workshop_1.Rproj	217 B	Oct 18, 2018, 12:18 PM
ws1_script1_stepwise_Bauer_dataset_analysis.R	6.1 KB	Dec 5, 2017, 12:19 PM
ws1_script2_Bauer_dataset_analysis.R	2 KB	Dec 6, 2017, 2:33 PM
ws1_script3_house_completions.R	2.4 KB	Oct 2, 2017, 3:53 PM

Line by line

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | workshop_1_project — 8_weeks_Oct-Dec_17

ws1_script1_stepwise_Bauer_dataset_an... *

Source on Save | Run | Source

```
12
13  raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")
14
15  separated_gene_df <- separate(raw_gene_df, NAME,
16                                c("name", "BP", "MF", "systematic_name",
17                                  "number"),
18                                sep = "\\|\\|")
19
20  mutated_gene_df <- mutate_at(separated_gene_df,
21                               vars(name:systematic_name),
22                               funs(trimws)
23                               )
24
25  selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
26
27  gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
28
29  nearly_there_df <- separate(gathered_gene_df, sample,
30                              c("nutrient", "rate"), sep = 1, convert = TRUE)
31
32  nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
33                      S = "Sulfate", N = "Ammonia", U = "Uracil")
34
35  cleaned_genes_df <- mutate(nearly_there_df,
36                             nutrient = plyr::revalue(nutrient, nutrient_names)
37                             ) %>%
38    filter(!is.na(expression), systematic_name != "")
```

35:1 | Section 1: Data import, tidying and transformation | R Script

Console Terminal

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project/

```
  NAME = col_character(),
  GWEIGHT = col_integer()
)
See spec(...) for full column specifications.
> separated_gene_df <- separate(raw_gene_df, NAME,
+                               c("name", "BP", "MF", "systematic_name",
+                                 "number"),
+                               sep = "\\|\\|")
> mutated_gene_df <- mutate_at(separated_gene_df,
+                              vars(name:systematic_name),
+                              funs(trimws)
+                              )
> selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
> gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
> nearly_there_df <- separate(gathered_gene_df, sample,
+                             c("nutrient", "rate"), sep = 1, convert = TRUE)
> nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
+                     S = "Sulfate", N = "Ammonia", U = "Uracil")
> |
```

Environment | History | Connections

Import Dataset | Grid

Global Environment

Name	Type	Length	Size	Value
gathered_gene_df	tbl_df	6	9.8 MB	199332 obs. of 6 variables
mutated_gene_df	tbl_df	44	3.5 MB	5537 obs. of 44 variables
nearly_there_df	tbl_df	7	11.3 MB	199332 obs. of 7 variables
nutrient_names	character	6	984 B	Named chr [1:6] "Glucose" "...
raw_gene_df	tbl_df	40	3.3 MB	5537 obs. of 40 variables
selected_gene_df	tbl_df	40	2.4 MB	5537 obs. of 40 variables
separated_gene...	tbl_df	44	3.6 MB	5537 obs. of 44 variables

Files | Plots | Packages | Help | Viewer

New Folder | Delete | Rename | More

Home > R_Users_Workshop > 8_weeks_Oct-Dec_17 > Workshop_1 > workshop_1_project

Name	Size	Modified
..		
.RData	2.5 KB	Oct 2, 2017, 1:49 PM
.Rhistory	20.3 KB	Dec 6, 2017, 3:43 PM
Brauer2008_DataSet1.csv	1.6 MB	Sep 27, 2017, 11:32 PM
Brauer2008_DataSet1.tds	1.6 MB	Sep 28, 2017, 10:22 AM
house_completions.csv	4 KB	Sep 28, 2017, 1:35 PM
irish_population.csv	315 B	Aug 28, 2017, 4:21 PM
raw_house_completions.csv	16.2 KB	Aug 25, 2017, 3:45 PM
workshop_1.Rproj	217 B	Oct 18, 2018, 12:18 PM
ws1_script1_stepwise_Bauer_dataset_analysis.R	6.1 KB	Dec 5, 2017, 12:19 PM
ws1_script2_Bauer_dataset_analysis.R	2 KB	Dec 6, 2017, 2:33 PM
ws1_script3_house_completions.R	2.4 KB	Oct 2, 2017, 3:53 PM

Line by line

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | workshop_1_project — 8_weeks_Oct-Dec_17

ws1_script1_stepwise_Bauer_dataset_an... *

Source on Save | Run | Source

```
15 separated_gene_df <- separate(raw_gene_df, NAME,
16                               c("name", "BP", "MF", "systematic_name",
17                                 "number"),
18                               sep = "\\|\\|")
19
20 mutated_gene_df <- mutate_at(separated_gene_df,
21                              vars(name:systematic_name),
22                              funs(trimws)
23                              )
24
25 selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
26
27 gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
28
29 nearly_there_df <- separate(gathered_gene_df, sample,
30                              c("nutrient", "rate"), sep = 1, convert = TRUE)
31
32 nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
33                     S = "Sulfate", N = "Ammonia", U = "Uracil")
34
35 cleaned_genes_df <- mutate(nearly_there_df,
36                            nutrient = plyr::revalue(nutrient, nutrient_names)
37                            ) %>%
38   filter(!is.na(expression), systematic_name != "")
39
40
41
```

44:1 | Section 1: Data import, tidying and transformation | R Script

Console Terminal

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project/

```
> separated_gene_df <- separate(raw_gene_df, NAME,
+                               c("name", "BP", "MF", "systematic_name",
+                                 "number"),
+                               sep = "\\|\\|")
> mutated_gene_df <- mutate_at(separated_gene_df,
+                              vars(name:systematic_name),
+                              funs(trimws)
+                              )
> selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
> gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
> nearly_there_df <- separate(gathered_gene_df, sample,
+                             c("nutrient", "rate"), sep = 1, convert = TRUE)
> nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
+                     S = "Sulfate", N = "Ammonia", U = "Uracil")
> cleaned_genes_df <- mutate(nearly_there_df,
+                            nutrient = plyr::revalue(nutrient, nutrient_names)
+                            ) %>%
+   filter(!is.na(expression), systematic_name != "")
>
```

Environment History Connections

Import Dataset | Grid

Global Environment

Name	Type	Length	Size	Value
cleaned_genes_df	tbl_df	7	11.3 MB	198430 obs. of 7 variables
gathered_gene_df	tbl_df	6	9.8 MB	199332 obs. of 6 variables
mutated_gene_df	tbl_df	44	3.5 MB	5537 obs. of 44 variables
nearly_there_df	tbl_df	7	11.3 MB	199332 obs. of 7 variables
nutrient_names	character	6	984 B	Named chr [1:6] "Glucose" "...
raw_gene_df	tbl_df	40	3.3 MB	5537 obs. of 40 variables
selected_gene_df	tbl_df	40	2.4 MB	5537 obs. of 40 variables
separated_gene...	tbl_df	44	3.6 MB	5537 obs. of 44 variables

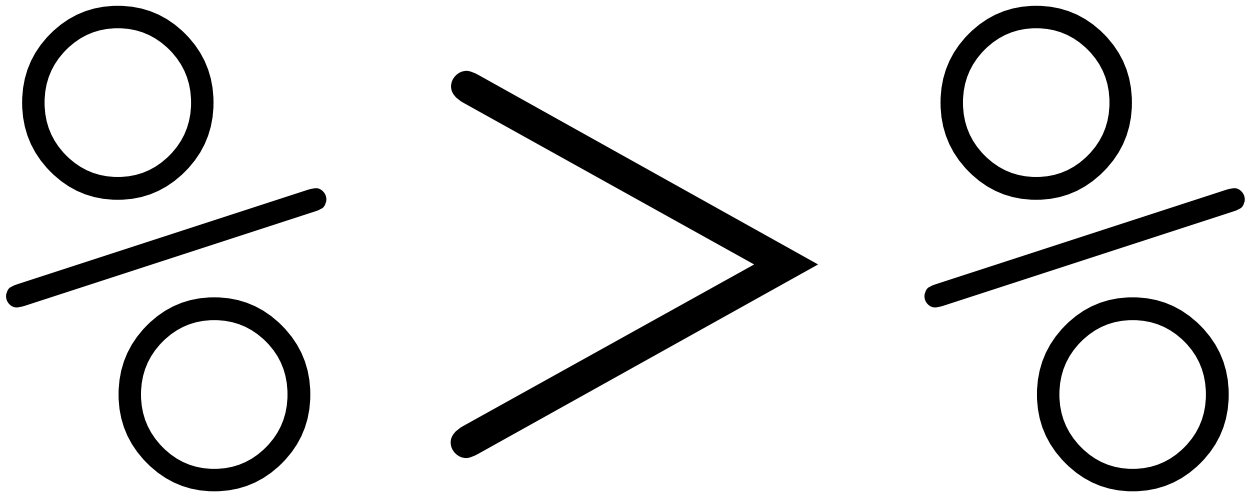
Files Plots Packages Help Viewer

New Folder | Delete | Rename | More

Home > R_Users_Workshop > 8_weeks_Oct-Dec_17 > Workshop_1 > workshop_1_project

Name	Size	Modified
..		
.RData	2.5 KB	Oct 2, 2017, 1:49 PM
.Rhistory	20.3 KB	Dec 6, 2017, 3:43 PM
Brauer2008_DataSet1.csv	1.6 MB	Sep 27, 2017, 11:32 PM
Brauer2008_DataSet1.tds	1.6 MB	Sep 28, 2017, 10:22 AM
house_completions.csv	4 KB	Sep 28, 2017, 1:35 PM
irish_population.csv	315 B	Aug 28, 2017, 4:21 PM
raw_house_completions.csv	16.2 KB	Aug 25, 2017, 3:45 PM
workshop_1.Rproj	217 B	Oct 18, 2018, 12:18 PM
ws1_script1_stepwise_Bauer_dataset_analysis.R	6.1 KB	Dec 5, 2017, 12:19 PM
ws1_script2_Bauer_dataset_analysis.R	2 KB	Dec 6, 2017, 2:33 PM
ws1_script3_house_completions.R	2.4 KB	Oct 2, 2017, 3:53 PM

Putting the pieces together



Code structure v2

```
separated_gene <- raw_gene %>%
    separate(NAME,
        c("name", "BP", "MF", "systematic_name", "number"),
        sep = "\\|\\|"
        )
```

Here the input data is outside the function

First argument is no longer the data

Piped

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | workshop_1_project — 8_weeks_Oct-Dec_17

ws1_script1_stepwise_Bauer_dataset_an... | ws1_script2_Bauer_dataset_analysis.R*

Source on Save | Run | Source

```
1  nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
2                      S = "Sulfate", N = "Ammonia", U = "Uracil"
3                      )
4
5  cleaned_genes_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t"
6                                 ) %>%
7
8    separate(NAME, c("name", "BP", "MF", "systematic_name", "number"), sep = "\\|\\|"
9             )  %>%
10
11   mutate_at(vars(name:systematic_name), funs(trimws)
12             ) %>%
13
14   select(-number, -GID, -YORF, -GWEIGHT
15          ) %>%
16
17   gather(sample, expression, G0.05:U0.3
18          ) %>%
19
20   separate(sample, c("nutrient", "rate"), sep = 1, convert = TRUE
21            ) %>%
22
23   mutate(nutrient = plyr::revalue(nutrient, nutrient_names)
24          ) %>%
25
26   filter(!is.na(expression), systematic_name != ""
27          )
```

9:18 (Top Level) R Script

Console | Terminal

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project/

```
+
+   separate(sample, c("nutrient", "rate"), sep = 1, convert = TRUE
+            ) %>%
+
+   mutate(nutrient = plyr::revalue(nutrient, nutrient_names)
+          ) %>%
+
+   filter(!is.na(expression), systematic_name != ""
+          )
Parsed with column specification:
cols(
  .default = col_double(),
  GID = col_character(),
  YORF = col_character(),
  NAME = col_character(),
  GWEIGHT = col_integer()
)
See spec(...) for full column specifications.
>
```

Environment | History | Connections

Import Dataset | Grid

Global Environment

Name	Type	Length	Size	Value
cleaned_genes_df	tbl_df	7	11.3 MB	198430 obs. of 7 variables
nutrient_names	character	6	984 B	Named chr [1:6] "Glucose" "Le...

Files | Plots | Packages | Help | Viewer

New Folder | Delete | Rename | More

Home > R_Users_Workshop > 8_weeks_Oct-Dec_17 > Workshop_1 > workshop_1_project

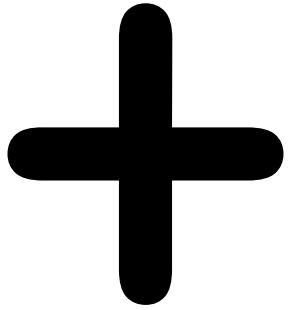
Name	Size	Modified
..		
.RData	2.5 KB	Oct 2, 2017, 1:49 PM
.Rhistory	20.3 KB	Dec 6, 2017, 3:43 PM
Brauer2008_DataSet1.csv	1.6 MB	Sep 27, 2017, 11:32 PM
Brauer2008_DataSet1.tds	1.6 MB	Sep 28, 2017, 10:22 AM
house_completions.csv	4 KB	Sep 28, 2017, 1:35 PM
irish_population.csv	315 B	Aug 28, 2017, 4:21 PM
raw_house_completions.csv	16.2 KB	Aug 25, 2017, 3:45 PM
workshop_1.Rproj	217 B	Oct 18, 2018, 12:18 PM
ws1_script1_stepwise_Bauer_dataset_analysis.R	6.1 KB	Dec 5, 2017, 12:19 PM
ws1_script2_Bauer_dataset_analysis.R	2 KB	Dec 6, 2017, 2:33 PM
ws1_script3_house_completions.R	2.4 KB	Oct 2, 2017, 3:53 PM

Piped

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

cleaned_genes_df <- read_delim("Brauer2008_DataSet1.tds", 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
           ) %>%

  mutate(nutrient = plyr::revalue(nutrient, nutrient_names)
         ) %>%

  filter(!is.na(expression), systematic_name != ""
         )
```

The moral of the story…….

You can go from this

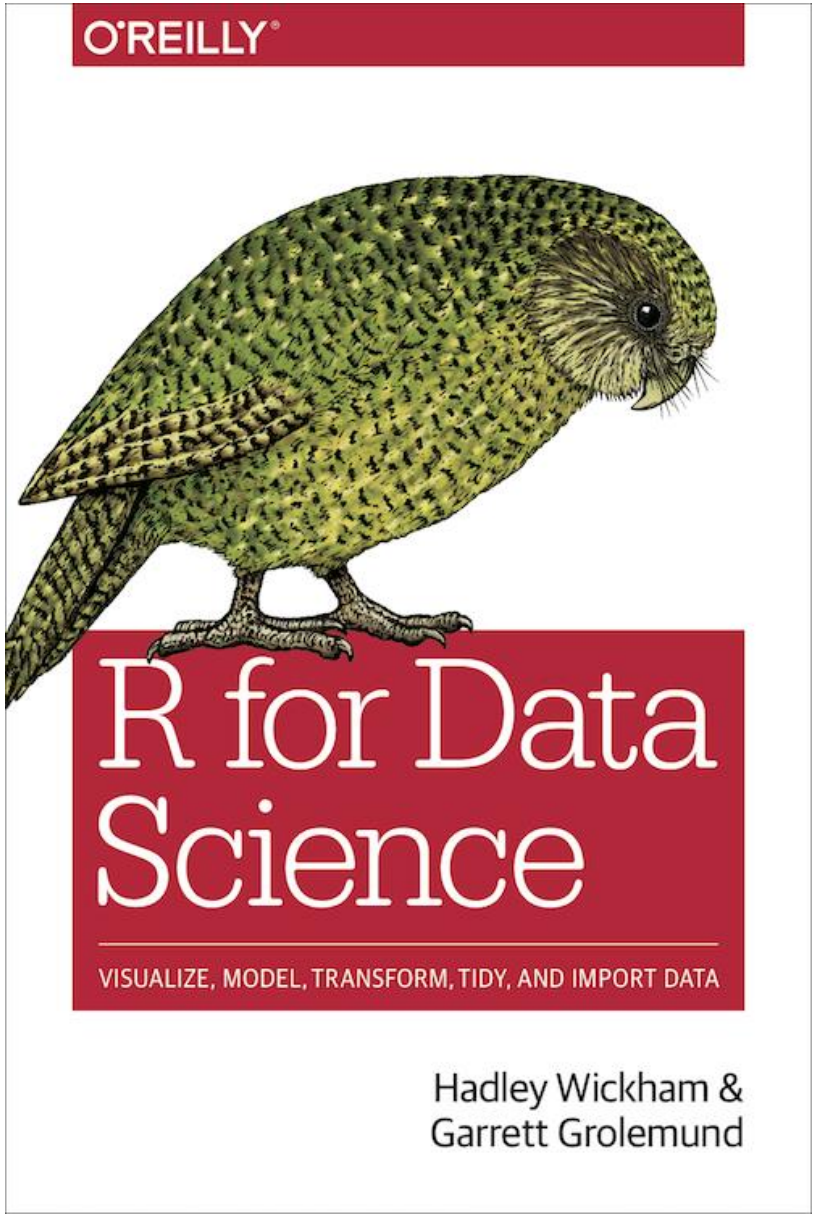


To this!!

Completely ordinary

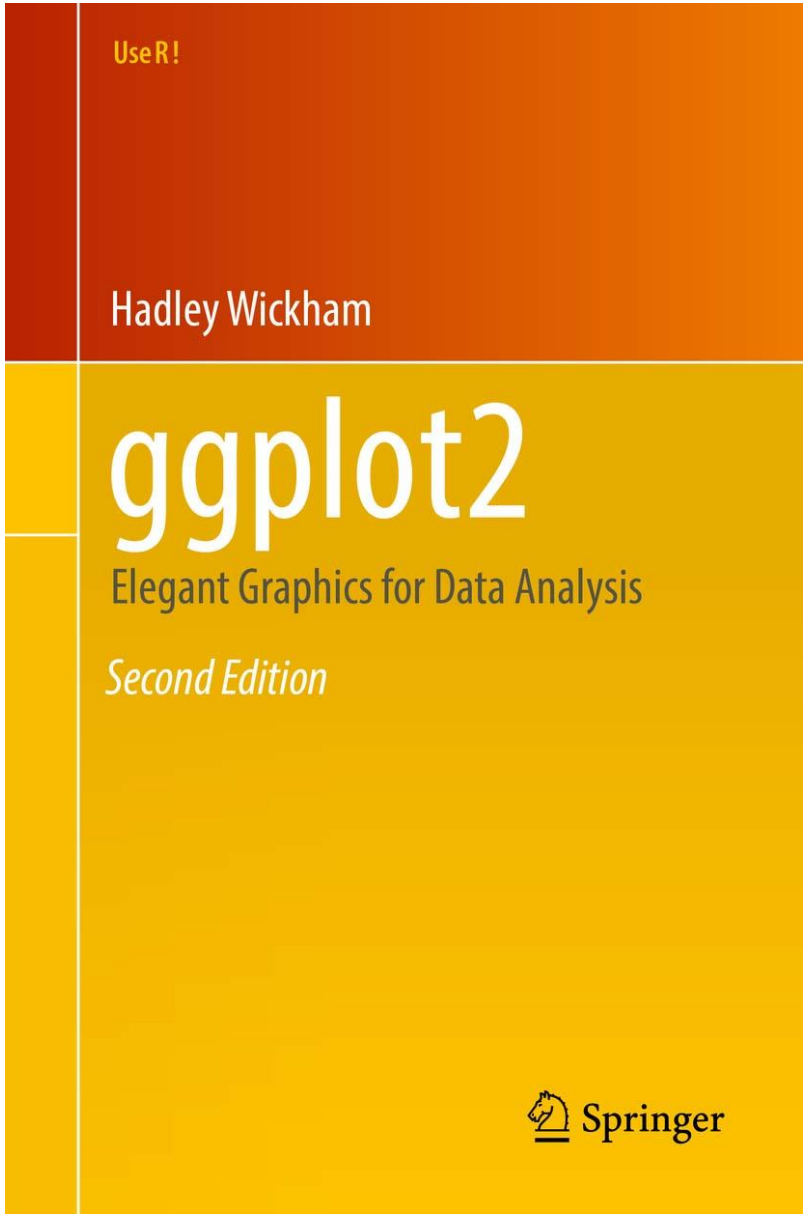
Master Builder!

You could write a book on that!!



https://r4ds.had.co.nz/

And on this!!


Use R!
Hadley Wickham
ggplot2
Elegant Graphics for Data Analysis
Second Edition
Springer

C: Tibbles

- The tidyverse equivalent of data.frames

4 main points of difference:

1. Printing in the console
2. Subsetting (The use of a placeholder (“.”)
3. Interacting with older code
4. Tibbles don’t change the input

- Open the script 05_tibbles.R

C: readr and more

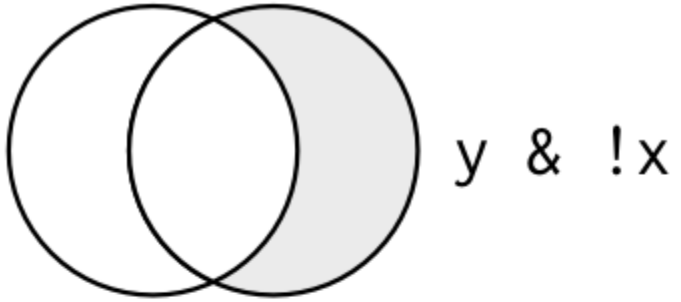
- fast way to read rectangular data (like csv, tsv)
- read_csv(): comma separated (CSV) files
- read_tsv(): tab separated files
- read_delim(): general delimited files

- readxl supports both the legacy .xls format and the modern xml-based .xlsx format
- Need to load explicitly

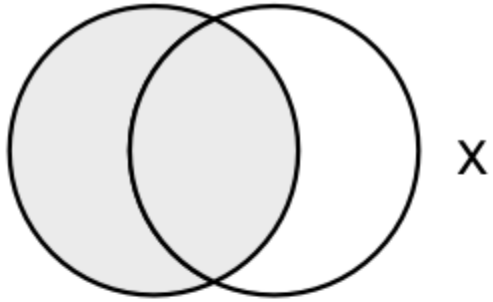
- read_sas(): SAS files
- read_sav(): SPSS files
- read_dta(): Stata files
- Also need to load explicitly

- Open the script 06_readr.R

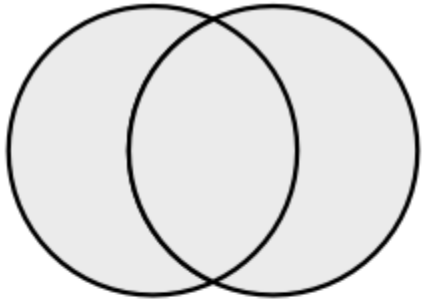
Logical operators and conditional subsetting



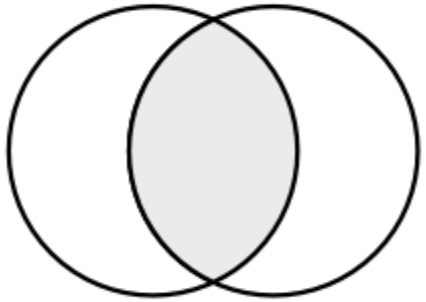
y & !x



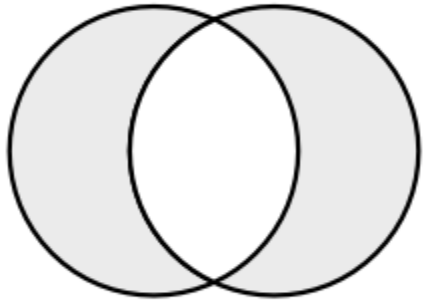
x



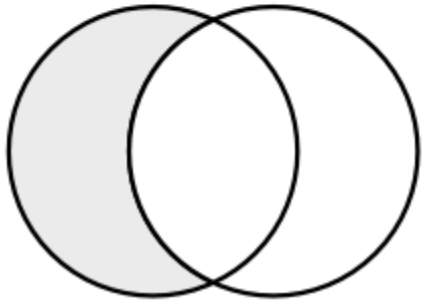
x | y



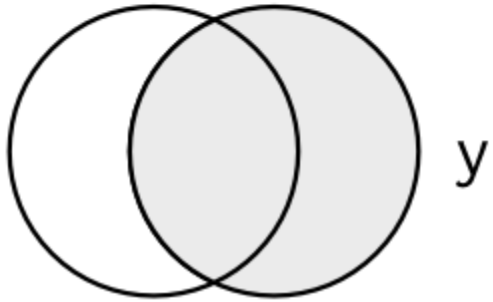
x & y



xor(x, y)



x & !y

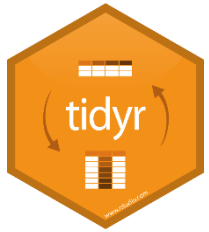


y

- & -> AND
- | -> OR (inclusive)
- ! -> NOT
- == -> EQUAL (identity)
- != -> NOT EQUAL

Image: Hadley Wickham R for Data Science Chapter 3

C: tidyr



- The goal is to create tidy data
    1. Each variable a column
    2. Each observation a row
    3. Each value is a cell

Main functions:
- gather()
- separate()

- Open the script 07_tidyr.R

Journal of Statistical Software
MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

Tidy Data

Hadley Wickham
RStudio

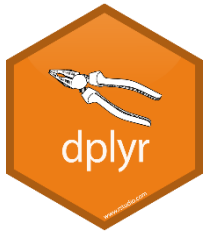
Abstract

A huge amount of effort is spent cleaning data to get it ready for analysis, but there has been little research on how to make data cleaning as easy and effective as possible. This paper tackles a small, but important, component of data cleaning: data tidying. Tidy datasets are easy to manipulate, model and visualise, and have a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table. This framework makes it easy to tidy messy datasets because only a small set of tools are needed to deal with a wide range of un-tidy datasets. This structure also makes it easier to develop tidy tools for data analysis, tools that both input and output tidy datasets. The advantages of a consistent data structure and matching tools are demonstrated with a case study free from mundane data manipulation chores.

Keywords: data cleaning, data tidying, relational databases, R.

DOI: 10.18637/jss.v059.i10

D: dplyr for data transformation



- Solves the most common data manipulation challenges

Main functions:
- select()
- filter()
- mutate()
- group_by()
- summarise()
- …… and many many more

- Open the script 08_dplyr.R

Time for some hands on application

- Open the script 09_data_cleaning_practise.R

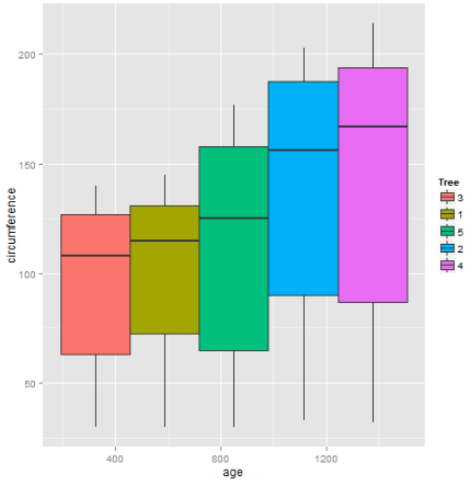
E: ggplot2

- Data visualisation based on "The Grammar of Graphics"

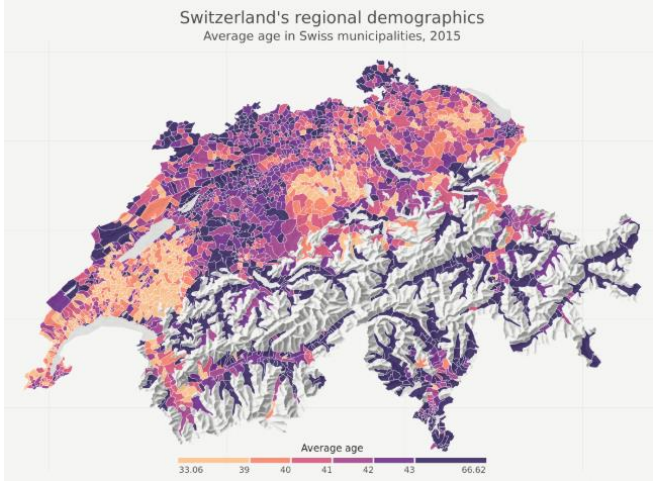
```
ggplot(data = <DATA>) +

    <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>))    +

     linear model  +

     axes formatting  +

     legend formatting  +

     title  + etc. etc.
```

C: ggplot2

- Very versatile
- Allows you to go from

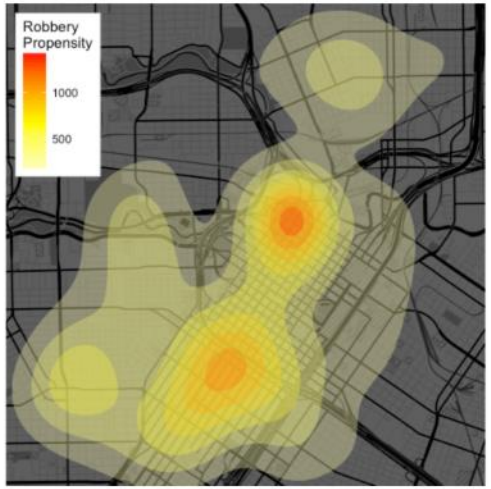


to

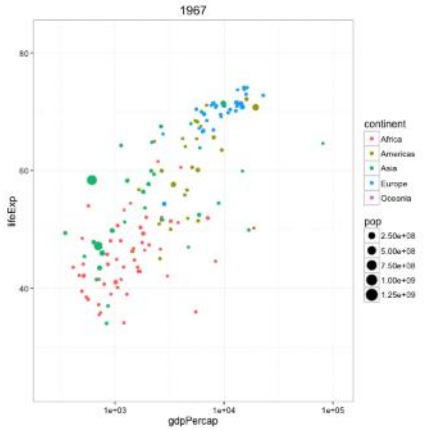


- Lots of add-on packages

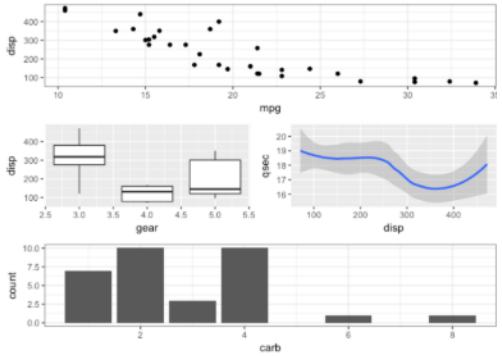
ggmap



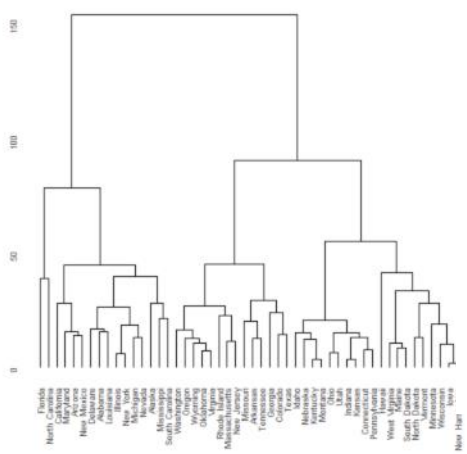
gganimate



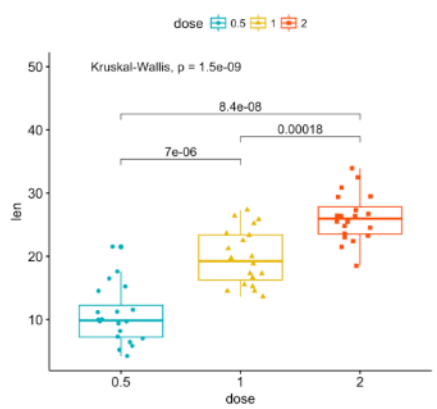
patchwork



ggdendro



ggpubr



Whistle-stop tour of ggplot2

Main features:
1. The data
2. The geoms
3. The mappings (x, y, colour, shape etc.)
4. Legends
5. Labels
6. Themes

and many many more

- Open the script 10_ggplot2.R

- Open the script 11_practise_plots.R