

R-reproducible 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?

Theses

Papers



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

Brendan A. O'Farrior,¹ Daniel Schmidt-Martin,² Zoya Dimitrova,² Pavel Skums,³ Orla Crosbie,⁴ Elizabeth Kenny-Walsh,⁵ Liam J. Fanning⁶

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

ABSTRACT
Hypervariable region 1 (HVR1) of hepatitis C virus (HCV) comprises the first 27 N-terminal amino acid residues of E2. It is classically recognized as the major antigenic target of the humoral immune response. HVR1 is also known to undergo rapid sequence evolution during chronic 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 variants 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 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 predominant dominance of a single variant within a narrow sequence space.

IMPORTANCE
Hepatitis C virus 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 variants 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 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 predominant dominance of a single variant within a narrow sequence space.

Hepatitis C virus (HCV) infection is a global health burden 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 variants 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 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 predominant dominance of a single variant within a narrow sequence space.

HCV is a single-stranded positive-sense RNA virus of considerable genomic heterogeneity. A recent reclassification defines the major genotypes 1a and 1b and 6 subtypes with genotypes 1 and 5 accounting for the majority of infections worldwide (6, 7). An error-prone RNA-dependent RNA polymerase, together with an inherent capacity for de novo hypervariable recombination, is responsible for much of this variability. These mutations are located within the envelope glycoprotein E2. The greatest heterogeneity has been identified at the 27-amino-acid HVR1 (residues 456 to 482) located at the C-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 hypothesis that a protein with reduced conformational plasticity in part accounts for high-density glycoprotein enhancement by scavenger receptor class B type I (SR-BI) interactions and is itself targeted by neutralizing antibodies (nAb) (10).

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 (18). Subsequent HVR1 evolution is driven primarily by strong selective pressure 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 10 December 2015
Editorial decision received 14 January 2016
Editorial decision received 14 January 2016
Editor: M. S. Diamond
Associate Editor: L. J. Fanning
Editorial Reviewer: L. J. Fanning, H. Hammami, G. L. Katz
B.A.O. and D.S.M. contributed equally to this article.
Copyright © 2016, American Society for Microbiology. All Rights Reserved.

Books



Talks

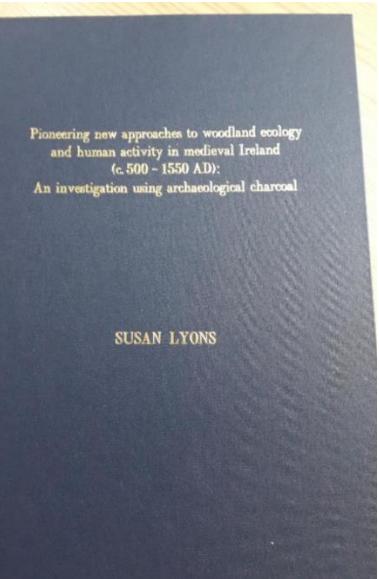


3318 J. Virol. 2016, Volume 90, Number 7

April 2016

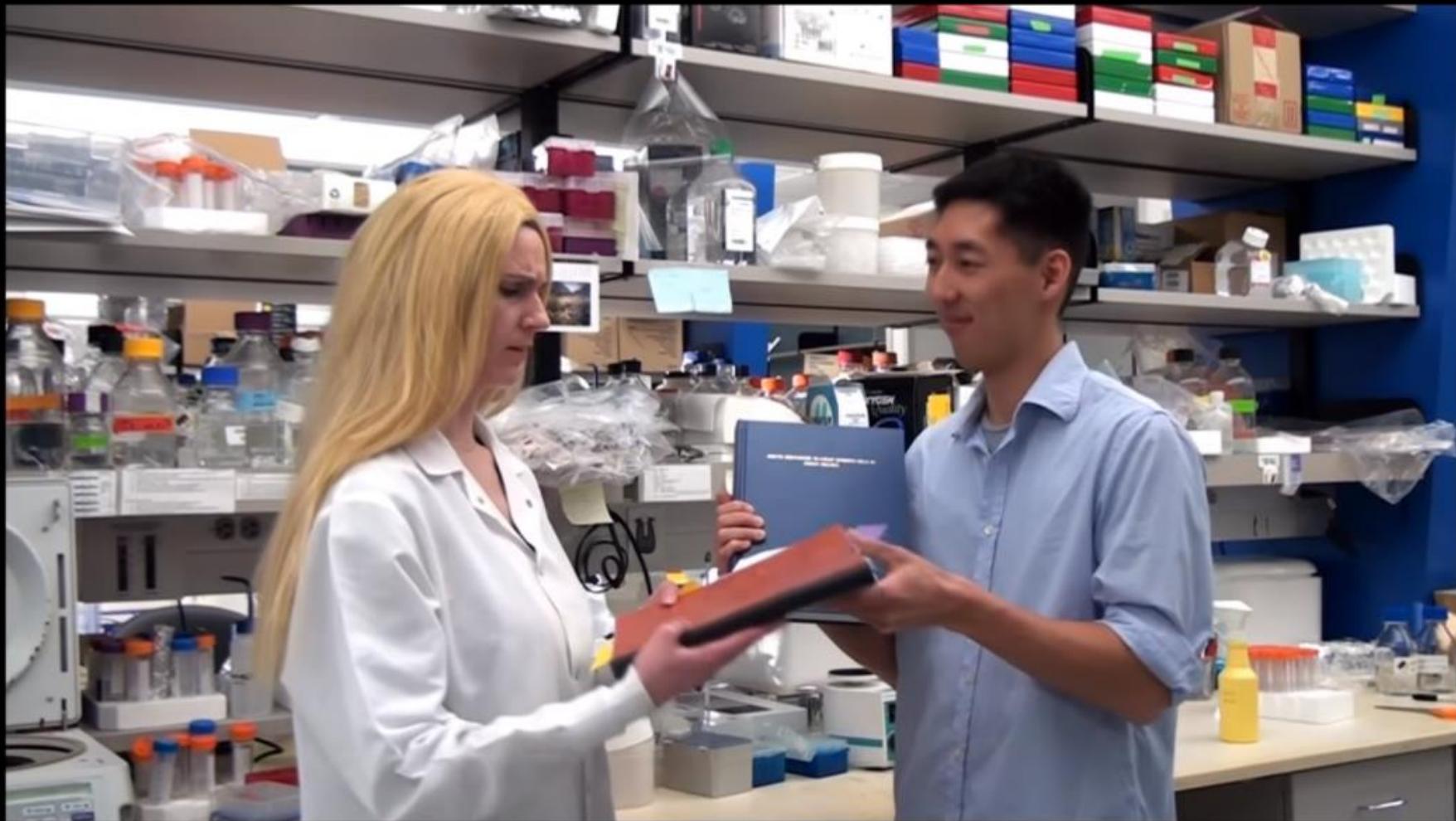
Journal of Virology

Posters



But what does it look like under the bonnet?

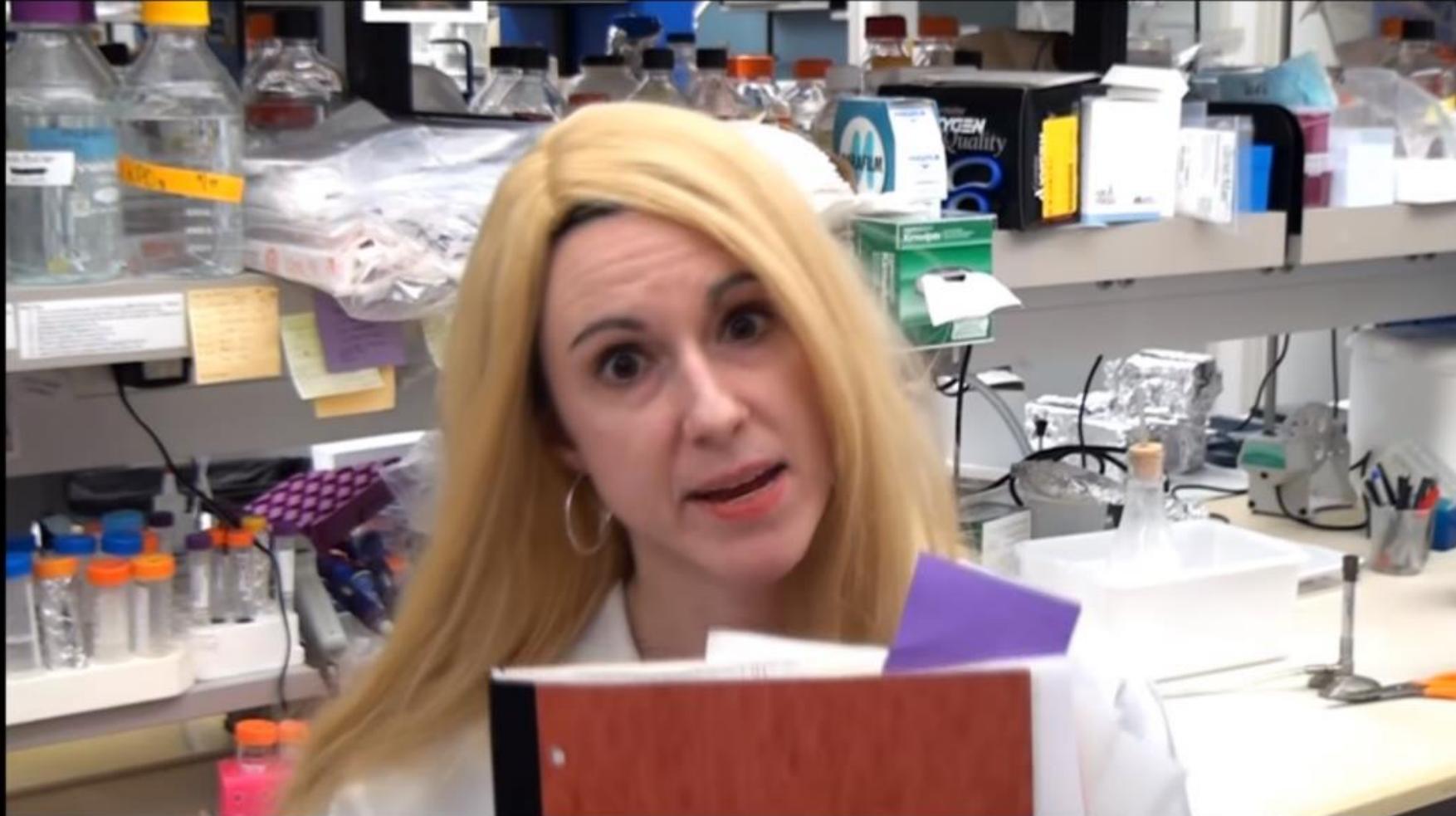




You were defending, one foot out the door



I got your project and its problems galore



I hate my life,

DON'T WORRY,
YOU DON'T HAVE
TO START YOUR
CODE FROM
SCRATCH.

YOU CAN REUSE THE
SOFTWARE THAT THE
PREVIOUS PERSON
ON THE PROJECT
WROTE SEVERAL
YEARS AGO.

ARE THERE
INSTRUCTIONS FOR
HOW TO USE IT?

I DOUBT IT.
IS THE CODE
COMMENTED?

NOT LIKELY.

WHERE ARE
THE FILES?

WHO KNOWS.

THIS IS GOING
TO BE PAINFUL,
ISN'T IT?

JUST A
SCRATCH.

THIS PERSON IS likely to be YOU BTW!!

Disclaimer



Jenny Bryan

@JennyBryan

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

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



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.
- [tidyverse](#) 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

VARIANCE EXPLAINED



David Robinson

@drob

Chief Data Scientist at [@DataCamp](#),
#rstats fan/evangelist

⌚ New York, NY

🔗 varianceexplained.org

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

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



David Robinson

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

- ✉ Email
- ⌚ Twitter
- ⌚ Github
- 🌐 Stack Overflow

Recent Posts

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

October 16, 2018

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

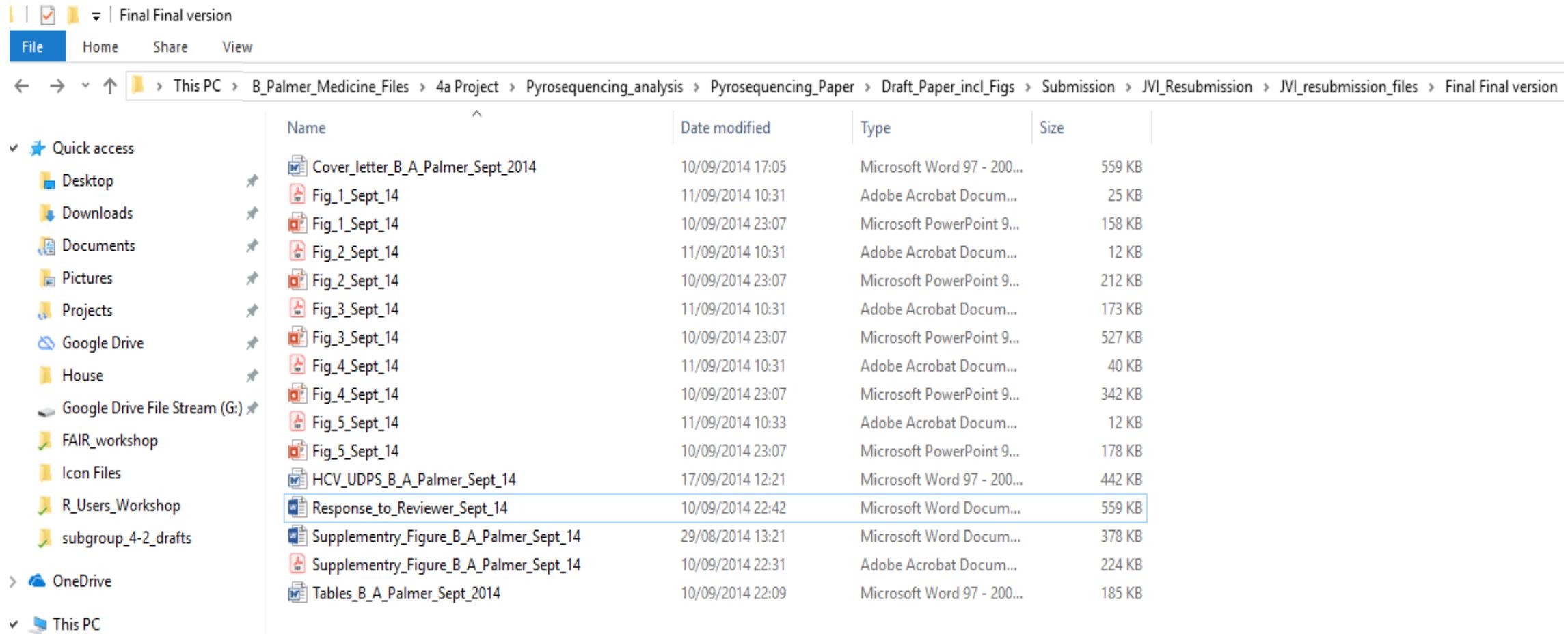
September 06, 2018

Scientific debt
Introducing an analogy to 'technical debt' for data scientists.

May 10, 2018

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

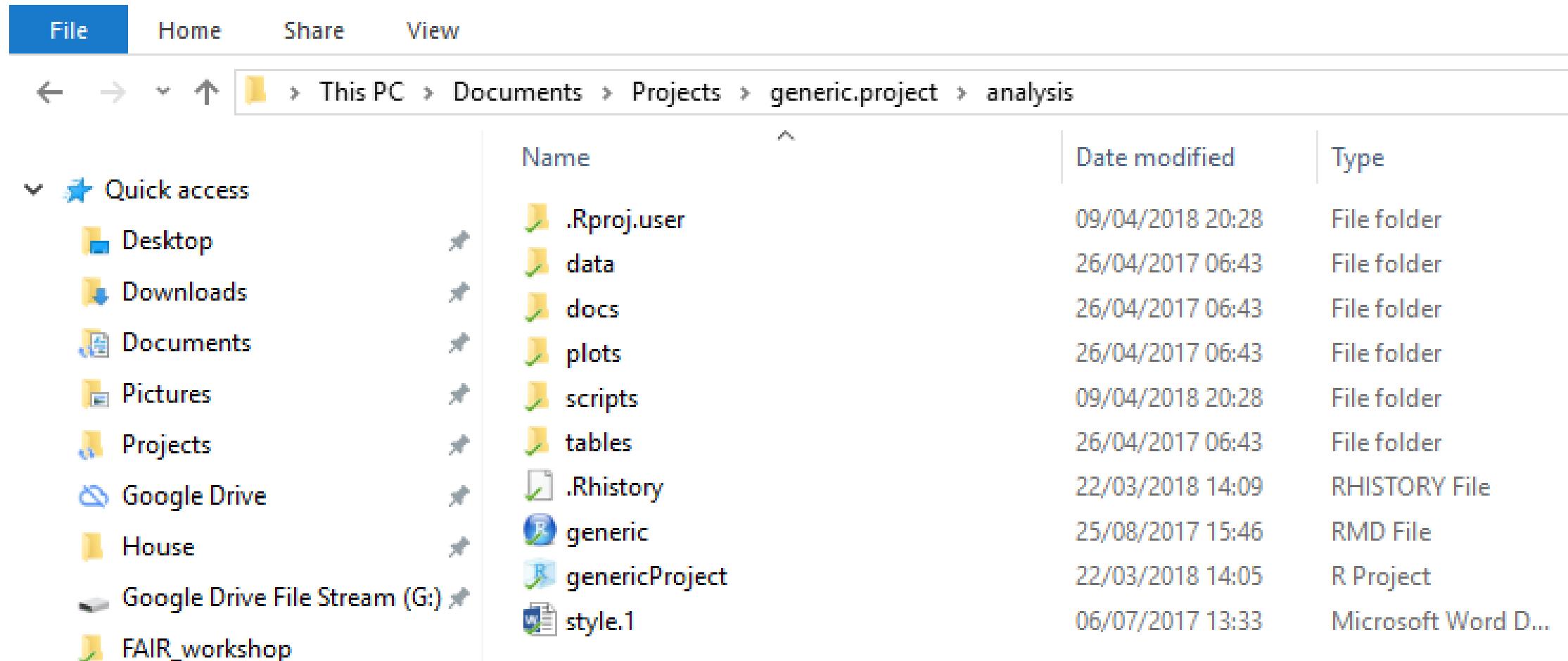
- Help your future-self



	Name	Date modified	Type	Size
Quick access	Cover_letter_B_A_Palmer_Sept_2014	10/09/2014 17:05	Microsoft Word 97 - 200...	559 KB
	Fig_1_Sept_14	11/09/2014 10:31	Adobe Acrobat Docum...	25 KB
	Fig_1_Sept_14	10/09/2014 23:07	Microsoft PowerPoint 9...	158 KB
	Fig_2_Sept_14	11/09/2014 10:31	Adobe Acrobat Docum...	12 KB
	Fig_2_Sept_14	10/09/2014 23:07	Microsoft PowerPoint 9...	212 KB
	Fig_3_Sept_14	11/09/2014 10:31	Adobe Acrobat Docum...	173 KB
	Fig_3_Sept_14	10/09/2014 23:07	Microsoft PowerPoint 9...	527 KB
	Fig_4_Sept_14	11/09/2014 10:31	Adobe Acrobat Docum...	40 KB
	Fig_4_Sept_14	10/09/2014 23:07	Microsoft PowerPoint 9...	342 KB
	Fig_5_Sept_14	11/09/2014 10:33	Adobe Acrobat Docum...	12 KB
	Fig_5_Sept_14	10/09/2014 23:07	Microsoft PowerPoint 9...	178 KB
	HCV_UDPS_B_A_Palmer_Sept_14	17/09/2014 12:21	Microsoft Word 97 - 200...	442 KB
	Response_to_Reviewer_Sept_14	10/09/2014 22:42	Microsoft Word Docum...	559 KB
	Supplementary_Figure_B_A_Palmer_Sept_14	29/08/2014 13:21	Microsoft Word Docum...	378 KB
	Supplementary_Figure_B_A_Palmer_Sept_14	10/09/2014 22:31	Adobe Acrobat Docum...	224 KB
	Tables_B_A_Palmer_Sept_2014	10/09/2014 22:09	Microsoft Word 97 - 200...	185 KB

Define a generic project structure

- STEP 1: Give your research projects a shared structure



The screenshot shows a Windows File Explorer window with the following details:

- File Explorer ribbon:** File, Home, Share, View.
- Address bar:** This PC > Documents > Projects > generic.project > analysis
- Left sidebar (Quick access):** Desktop, Downloads, Documents, Pictures, Projects, Google Drive, House, Google Drive File Stream (G:), FAIR_workshop.
- Right pane (File list):**

Name	Date modified	Type
.Rproj.user	09/04/2018 20:28	File folder
data	26/04/2017 06:43	File folder
docs	26/04/2017 06:43	File folder
plots	26/04/2017 06:43	File folder
scripts	09/04/2018 20:28	File folder
tables	26/04/2017 06:43	File folder
.Rhistory	22/03/2018 14:09	RHISTORY File
generic	25/08/2017 15:46	RMD File
genericProject	22/03/2018 14:05	R Project
style.1	06/07/2017 13:33	Microsoft Word D...

Give your files informative names

- STEP 1: Give your research projects a shared structure

The screenshot shows a Windows File Explorer window. The navigation bar at the top includes 'File', 'Home', 'Share', and 'View' tabs. The address bar shows the path: 'This PC > Documents > Projects > 18.04.27-WP3_Feeding_Trial > analysis > data'. On the left, there's a 'Quick access' sidebar with links to 'Desktop', 'Downloads', 'Documents', 'Pictures', and 'Projects'. The main pane displays a list of files in the 'data' folder. The columns are 'Name' (sorted by date modified) and 'Date modified'. The files listed are:

Name	Date modified
raw_files_password_protected	27/09/2018 11:54
master_database	12/06/2018 12:22
nutritics_food_level	09/05/2018 14:25
nutritics_grouped	09/05/2018 14:27
nutritics_grouped_reduced	14/05/2018 01:14

Everything in its right place

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

NO

Name
All unique 4a amino acid Sequences (B-N).fas
All unique 4a amino acid Sequences (B-N).meg
All_AA_haplotypes.meg
All_AA_haplotypes_with_clonal_sequences.meg
BS100_AA_with_clones
BS100_AA_with_clones.nwk
BS1000_AA_pyro&clones
BS1000_AA_pyro&clones.nwk
BS1000_AA_pyro_only
BS1000_AA_pyro_only.nwk
BS1000_Uncalibrated_Clonal_AA
BS1000_Uncalibrated_Clonal_AA.nwk
BS1000_Uncalibrated_Pyro_AA
BS1000_Uncalibrated_Pyro_AA.nwk
pic

Yes

Documents > Projects > 18.04.27-WP3_Feeding_Trial > analysis > scripts		
	Name	Date modified
...	01_data_import_and_tidying_master_file	02/10/2018 18:51
...	02_data_import_and_tidying_nutritics_grouped	19/10/2018 19:47
...	03_figures	17/10/2018 16:40
...	04_tables	22/05/2018 12:26
...	05_study_overview	19/10/2018 23:06
...	functions	13/05/2018 23:13

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_marshall-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

Outline a file naming convention

Chronological order:

```
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H01.csv  
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H02.csv  
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H03.csv  
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_platefile.csv  
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A01.csv  
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A02.csv  
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A03.csv  
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A04.csv
```

Logical order:

```
01_marshall-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

- benefits of using R projects for data analysis tasks

Documents > R_Users_Workshop > T_L_workshop > stats-teaching > project-structure

Name	Date modified	Type
data	03/11/2018 21:08	File folder
docs	04/11/2018 10:19	File folder
example_project	03/11/2018 13:51	File folder
lecture	03/11/2018 13:53	File folder
plots	03/11/2018 15:10	File folder
scripts	04/11/2018 10:13	File folder
LICENSE	03/11/2018 14:47	File
project-structure	04/11/2018 10:00	R Project
README	04/11/2018 10:23	MD File



- Open the script 03_project_clean_data.R
- Open the script 04_project_project_data.R

Don't Do What Donny Don't Does!!



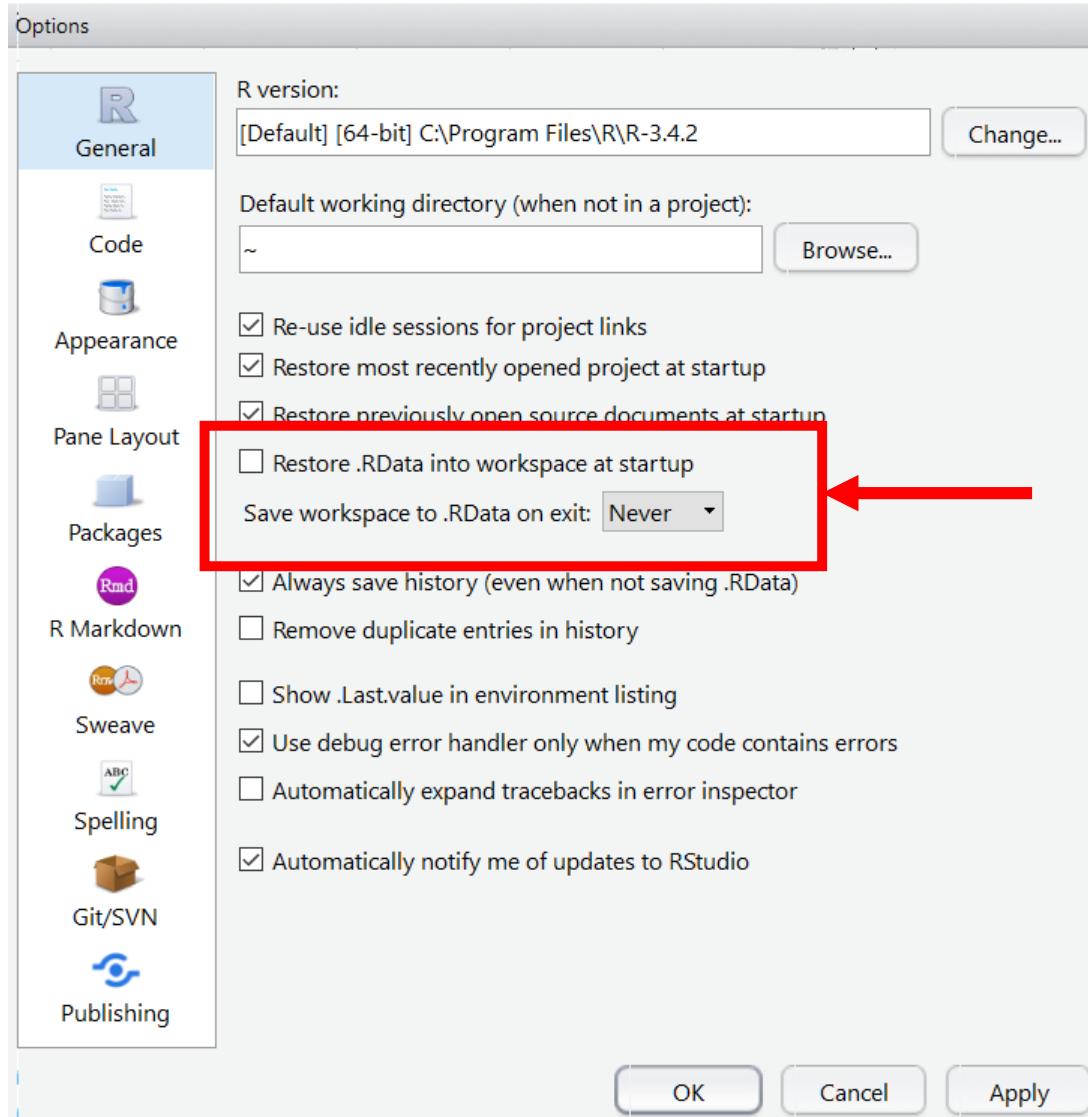
Donny Don't:

- Start your script with...
`setwd()`

Donny Don't:

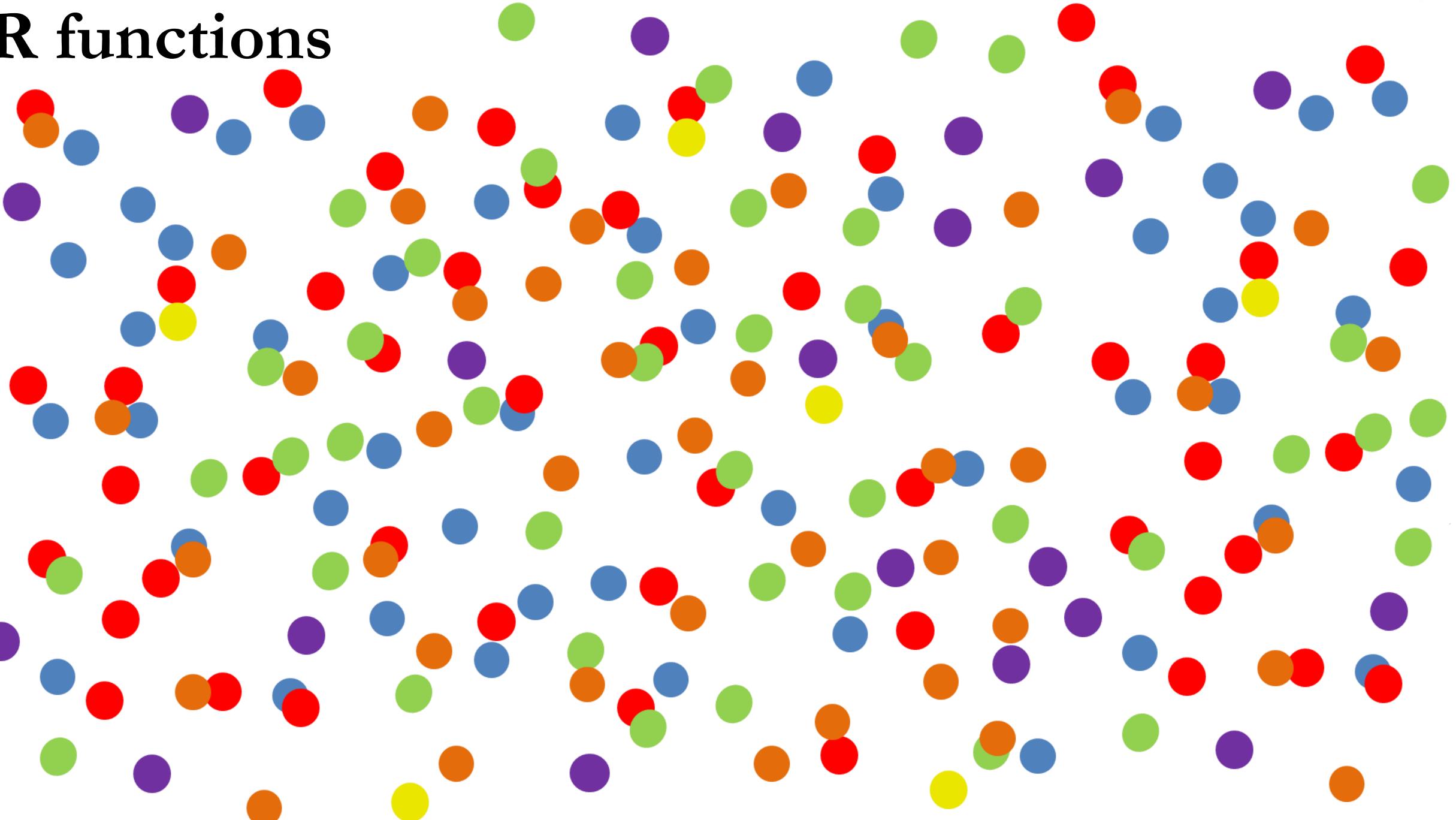
- Start your script with...
`rm(list = ls())`

Other points to note

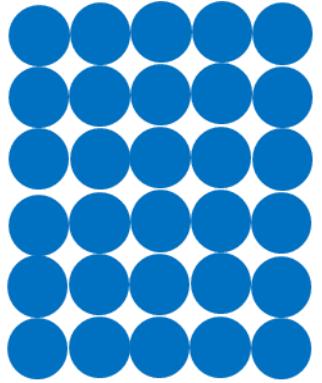


- You might consider your environment as “real”
- If you continue to use R, it is better for you to consider your R scripts as “real”, as these should recreate the environment
- You may suffer short term pain
- This will prevent long term agony

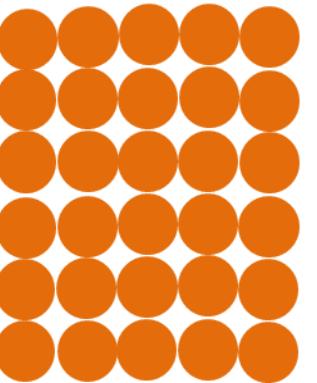
R functions



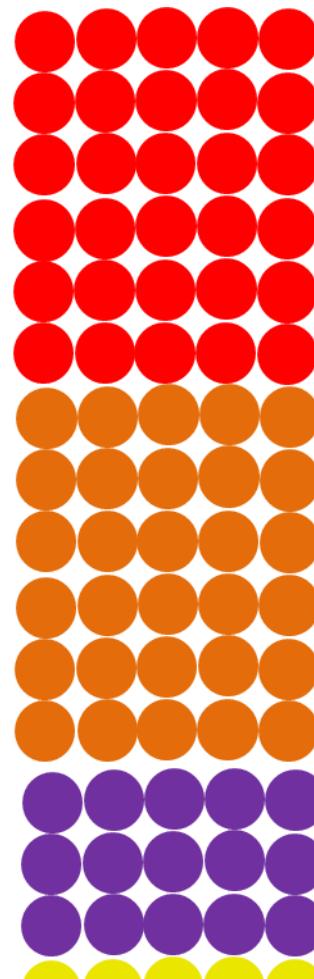
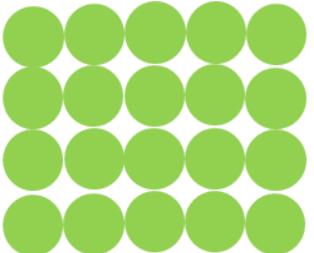
R packages



Base R:
Comes
pre-
loaded



Other packages:
Install once
Update regularly
Load each session



core
tidyverse

What is the tidyverse?

Tidyverse

Packages Articles Learn Help Contribute

R packages for data science

The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

Install the complete tidyverse with:

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

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

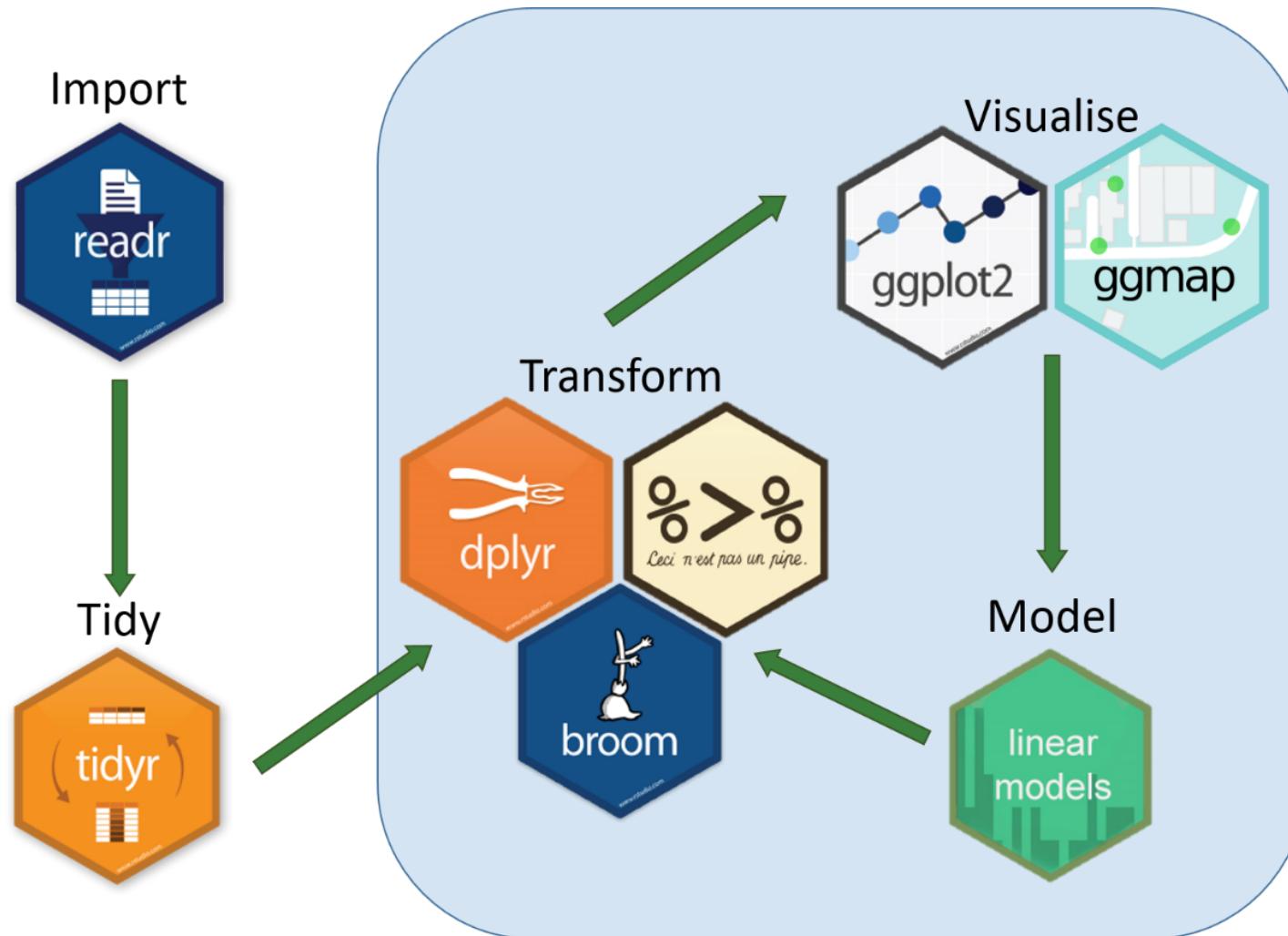
What is the tidyverse?

```
> 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", "tidyverse", "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

	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 polyadenylation* 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 <i>Saccharomyces</i>)* 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

```

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

Console Terminal

> 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.

> |

Line by line

workshop_1_project — 8_weeks_Oct-Dec_17

Environment History Connections

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

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

```

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

```

20:1 Section 1: Data import, tidying and transformation

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 = "\\|\\|\\|")
> |

```

Line by line

workshop_1_project — 8_weeks_Oct-Dec_17

Environment History Connections

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

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

```

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

```

27:1 Section 1: Data import, tidying and transformation

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)
>

```

Line by line

Environment History Connections

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

```

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

```

Console Terminal

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

```

> .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)
>

```

Line by line

Environment History Connections

Global Environment

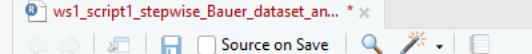
Name	Type	Length	Size	Value
gathered_gene_df	tbl_df	6	9.8 MB	199332 obs. of 6 variables
mutated_gene_df	tbl_df	44	3.5 MB	5537 obs. of 44 variables
raw_gene_df	tbl_df	40	3.3 MB	5537 obs. of 40 variables
selected_gene_df	tbl_df	40	2.4 MB	5537 obs. of 40 variables
separated_gene...	tbl_df	44	3.6 MB	5537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



```

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

```

Console Terminal x

```

~/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)
>

```

Line by line

Environment History Connections

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

```

12 raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")
13 separated_gene_df <- separate(raw_gene_df, NAME,
14                               c("name", "BP", "MF", "systematic_name",
15                                 "number"),
16                               sep = "\\|\\|\\|")
17 mutated_gene_df <- mutate_at(separated_gene_df,
18                               vars(name:systematic_name),
19                               funs(trimws)
20 )
21
22 selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
23
24 gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
25
26 nearly_there_df <- separate(gathered_gene_df, sample,
27                               c("nutrient", "rate"), sep = 1, convert = TRUE)
28
29 nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
30                       S = "Sulfate", N = "Ammonia", U = "Uracil")
31
32 cleaned_genes_df <- mutate(nearly_there_df,
33                               nutrient = plyr::revalue(nutrient, nutrient_names)
34                               ) %>%
35   filter(!is.na(expression), systematic_name != "")
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
279
280
281
282
283
284
285
286
287
288
289
289
290
291
292
293
294
295
296
297
298
299
299
300
301
302
303
304
305
306
307
308
309
309
310
311
312
313
314
315
316
317
318
319
319
320
321
322
323
324
325
326
327
328
329
329
330
331
332
333
334
335
336
337
338
339
339
340
341
342
343
344
345
346
347
348
349
349
350
351
351
352
353
354
355
356
357
358
359
359
360
361
362
363
364
365
366
367
368
369
369
370
371
372
373
374
375
376
377
378
379
379
380
381
382
383
384
385
386
387
388
389
389
390
391
392
393
394
395
396
397
398
399
399
400
401
402
403
404
405
406
407
408
409
409
410
411
412
413
414
415
416
417
418
419
419
420
421
422
423
424
425
426
427
428
429
429
430
431
432
433
434
435
436
437
438
439
439
440
441
442
443
444
445
446
447
448
449
449
450
451
452
453
454
455
456
457
458
459
459
460
461
462
463
464
465
466
467
468
469
469
470
471
472
473
474
475
476
477
478
479
479
480
481
482
483
484
485
486
487
488
489
489
490
491
492
493
494
495
496
497
498
499
499
500
501
502
503
504
505
506
507
508
509
509
510
511
512
513
514
515
516
517
518
519
519
520
521
522
523
524
525
526
527
528
529
529
530
531
532
533
534
535
536
537
538
539
539
540
541
542
543
544
545
546
547
548
549
549
550
551
552
553
554
555
556
557
557
558
559
559
560
561
562
563
564
565
566
567
568
569
569
570
571
572
573
574
575
576
577
578
579
579
580
581
582
583
584
585
586
587
588
589
589
590
591
592
593
594
595
596
597
598
599
599
600

```

Line by line

Environment History Connections

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

ws1_script1_stepwise_Bauer_dataset_analysis.R

```

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 < Section 1: Data import, tidying and transformation
42
43
44:1

```

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

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_df	tbl_df	44	3.6 MB	5537 obs. of 44 variables

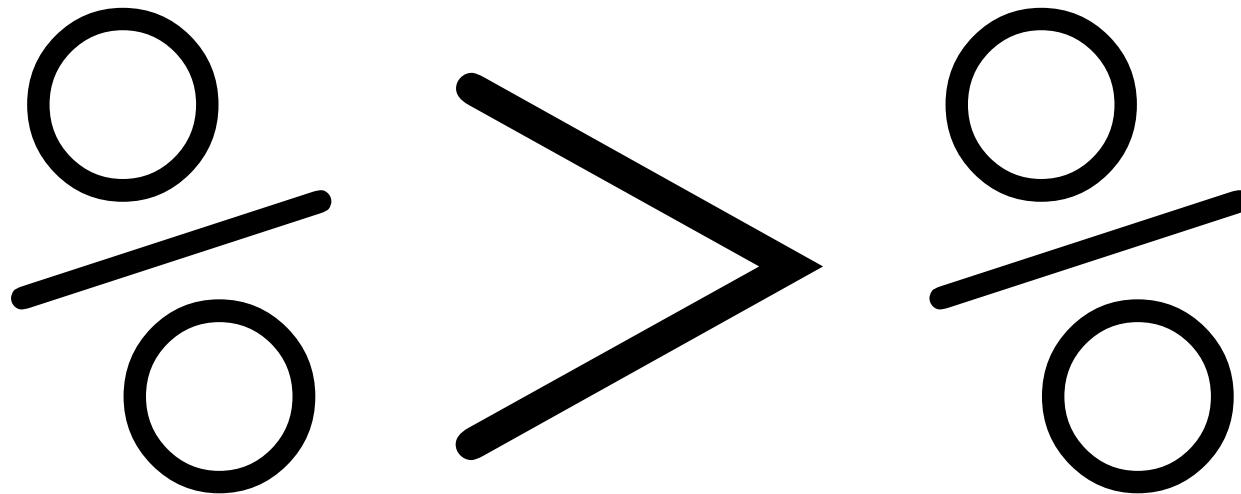
Files Plots Packages Help Viewer

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,           ← First argument is no longer the data  
    c("name", "BP", "MF", "systematic_name", "number"),  
    sep = "\\|\\|"  
  )
```

Here the input data is outside the function

```

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

```

9:18 (Top Level) ▾

Console Terminal

~/R_Users_Workshop/8_weeks_Oct-Dec_17/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.

> |

Piped

workshop_1_project — 8_weeks_Oct-Dec_17

Environment History Connections

Import Dataset

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

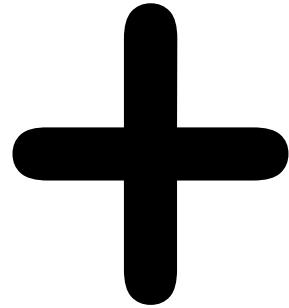
Name	Size	Modified
.RData	2.5 KB	Oct 2, 2017, 1:49 PM
.Rhistory	20.3 KB	Dec 6, 2017, 3:43 PM
Brauer2008_DataSet1.csv	1.6 MB	Sep 27, 2017, 11:32 PM
Brauer2008_DataSet1.tds	1.6 MB	Sep 28, 2017, 10:22 AM
house_completions.csv	4 KB	Sep 28, 2017, 1:35 PM
irish_population.csv	315 B	Aug 28, 2017, 4:21 PM
raw_house_completions.csv	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

```
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  select(-number, -GID, -YORF, -GWEIGHT
14         ) %>%
15
16  gather(sample, expression, G0.05:U0.3
17         ) %>%
18
19  separate(sample, c("nutrient", "rate"), sep = 1, convert = TRUE
20         ) %>%
21
22  mutate(nutrient = plyr::revalue(nutrient, nutrient_names)
23         ) %>%
24
25  filter(!is.na(expression), systematic_name != ""
26         )
```

The moral of the story.....

You can go from this

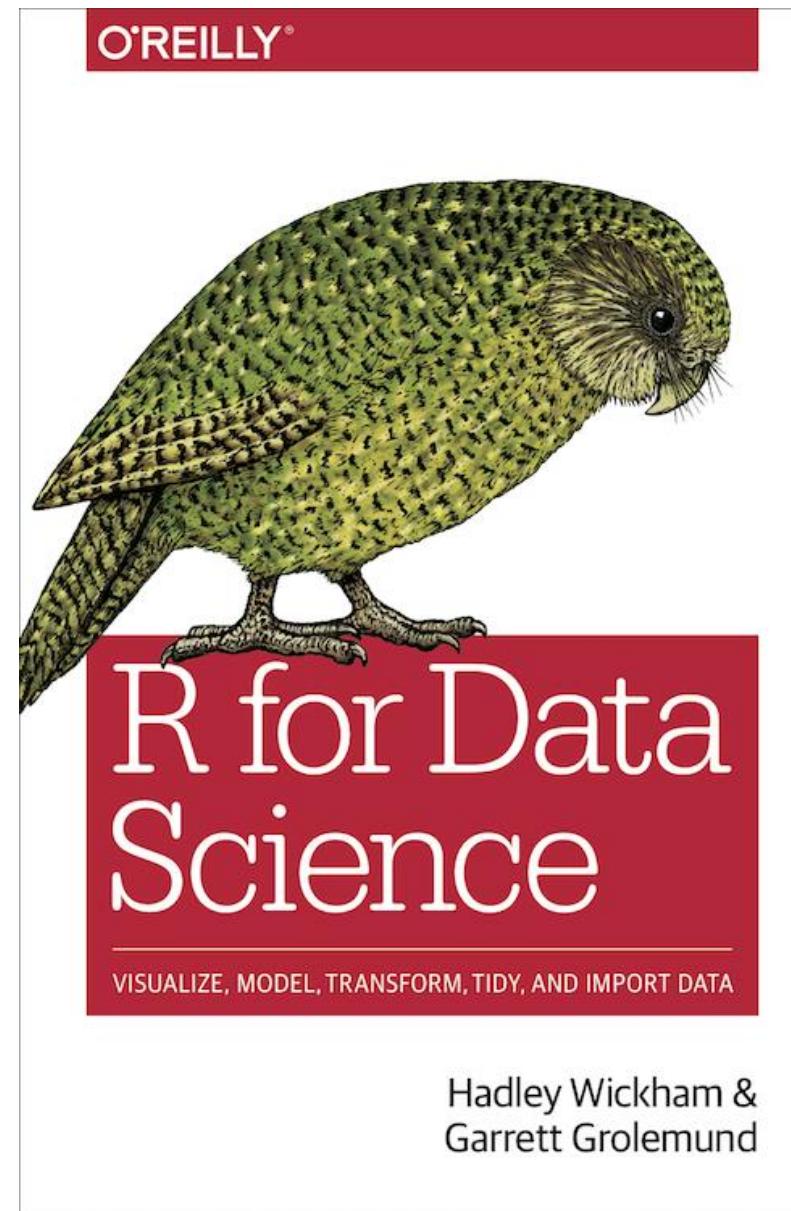


Completely ordinary

To this!!

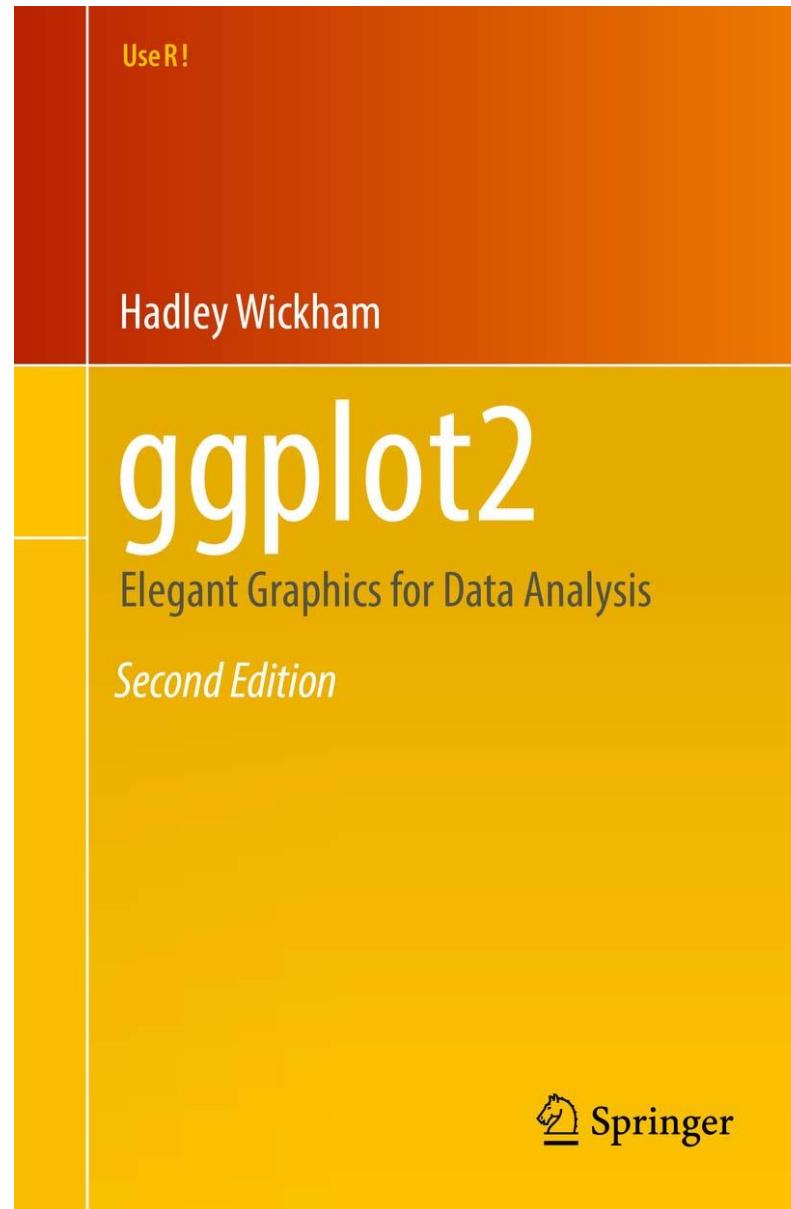
Master Builder!

You could write a book on that!!



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

And on this!!



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

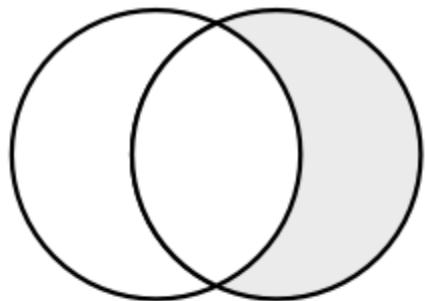


- `readxl` supports both the legacy .xls format and the modern xml-based .xlsx format
- Need to load explicitly

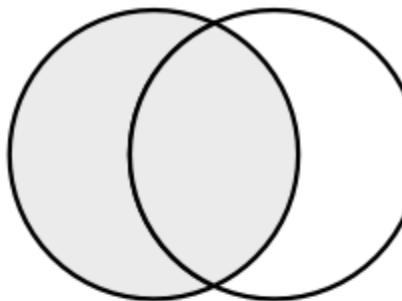


- `read_sas()`: SAS files
- `read_sav()`: SPSS files
- `read_dta()`: Stata files
- Also need to load explicitly
- Open the script 06_readr.R

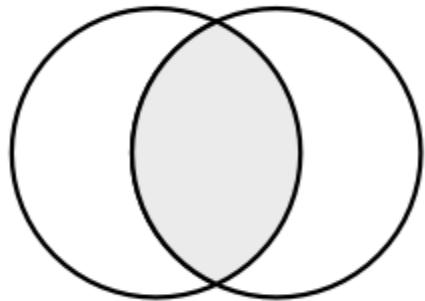
Logical operators and conditional subsetting



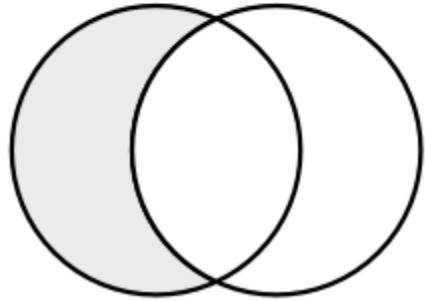
$y \& \neg x$



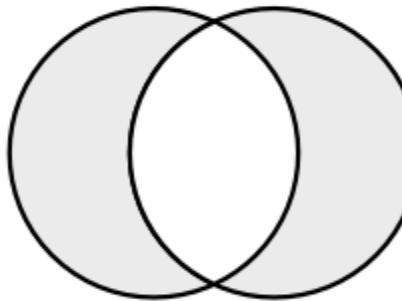
x



$x \& y$



$x \& \neg y$



$\text{xor}(x, y)$

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



C: tidyverse



- 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_tidyverse.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](https://doi.org/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 +

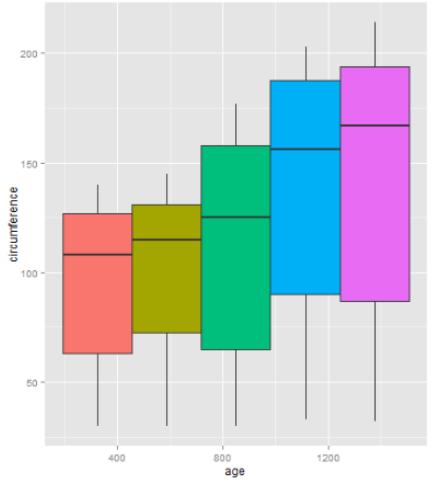
axes formatting +

legend formatting +

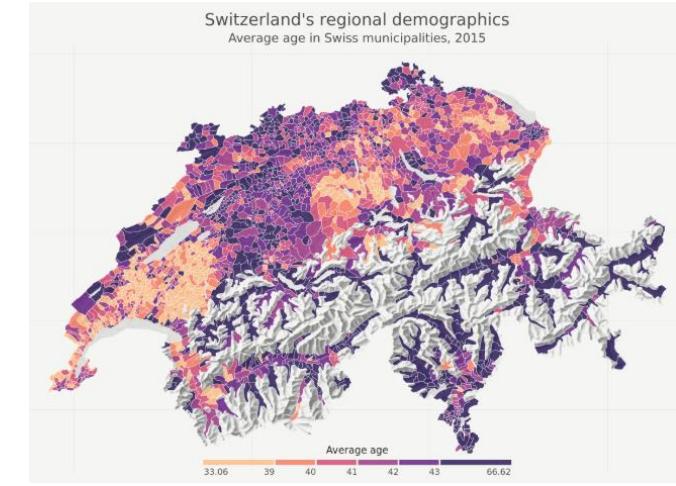
title + etc. etc.

C: ggplot2

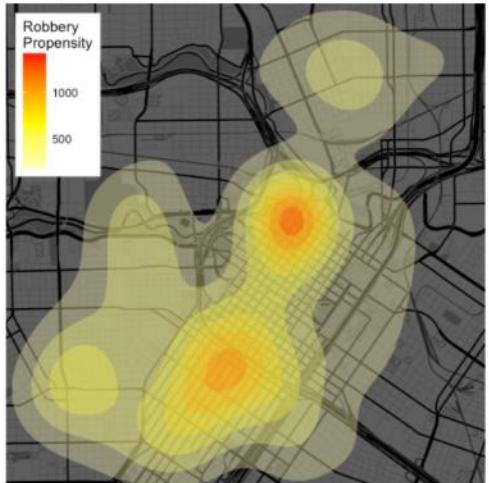
- Very versatile
- Allows you to go from
- Lots of add-on packages



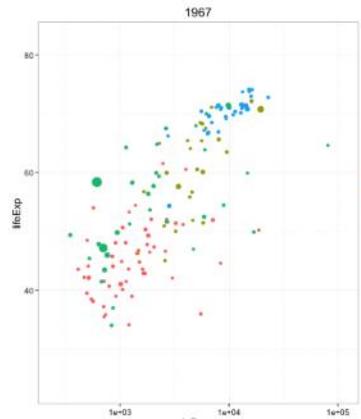
to



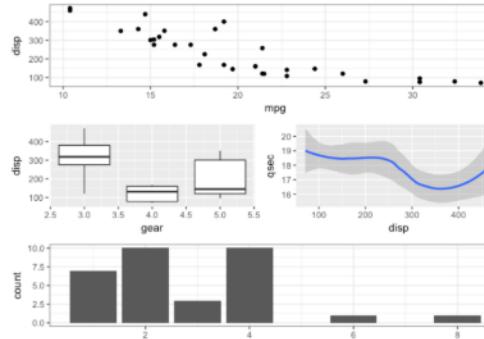
ggmap



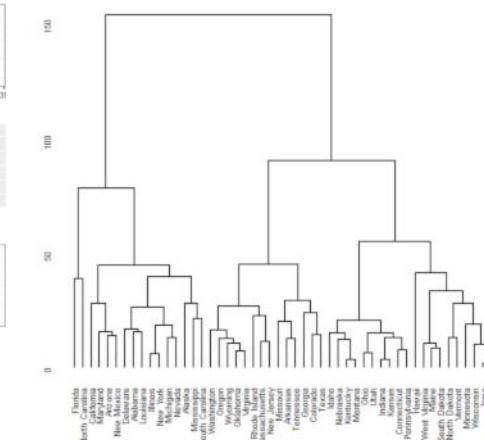
ggridge



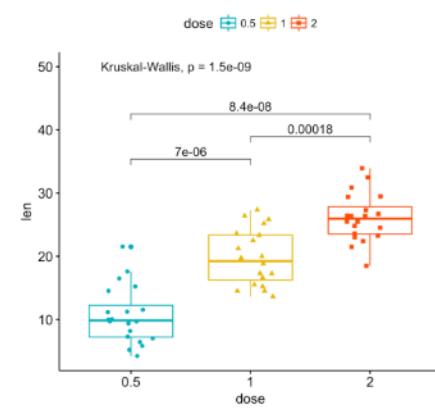
patchwork



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