R-projects

- Everything in its right place

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

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

CRF-C
HRB Clinical Research Facility Cork

HRB
Health Research Board

How is research presented?

Papers

AMERICAN SOCIETY FOR MICROBIOLOGY | 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).

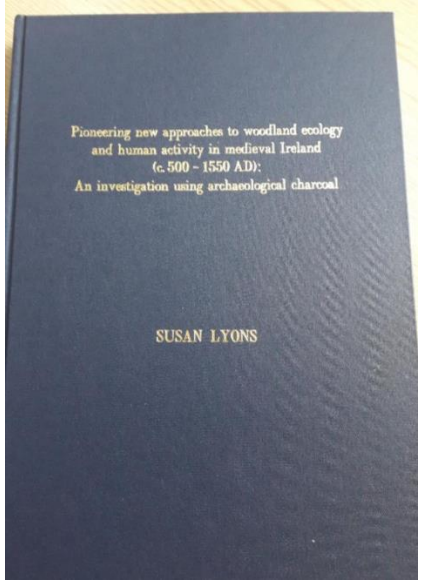
HCV is a single-stranded positive-sense RNA virus of considerable genomic heterogeneity. A recent reclassification defined the HCV global distribution into 7 genotypes and 67 subtypes, with genotypes 1 and 3 accounting for the majority of infections worldwide (6, 7). An error-prone RNA–dependent RNA 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.

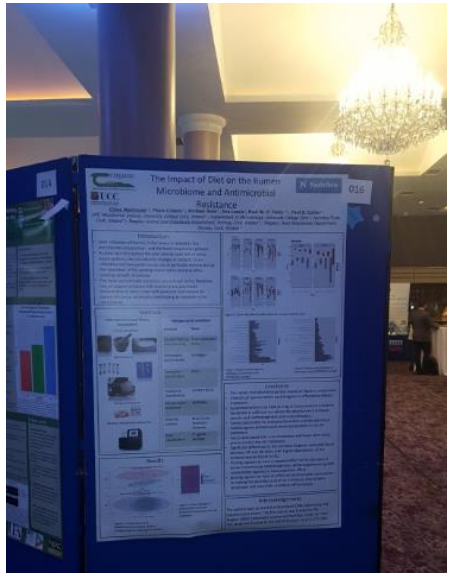
Books

Theses



Talks

Posters



But what does it really look like?

Disclaimer

Hadley Wickham
@hadleywickham

R, data, visualisation.

Houston, TX
hadley.nz

Jenny Bryan
@JennyBryan

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

David Robinson
@drob

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

New York, NY
varianceexplained.org

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

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

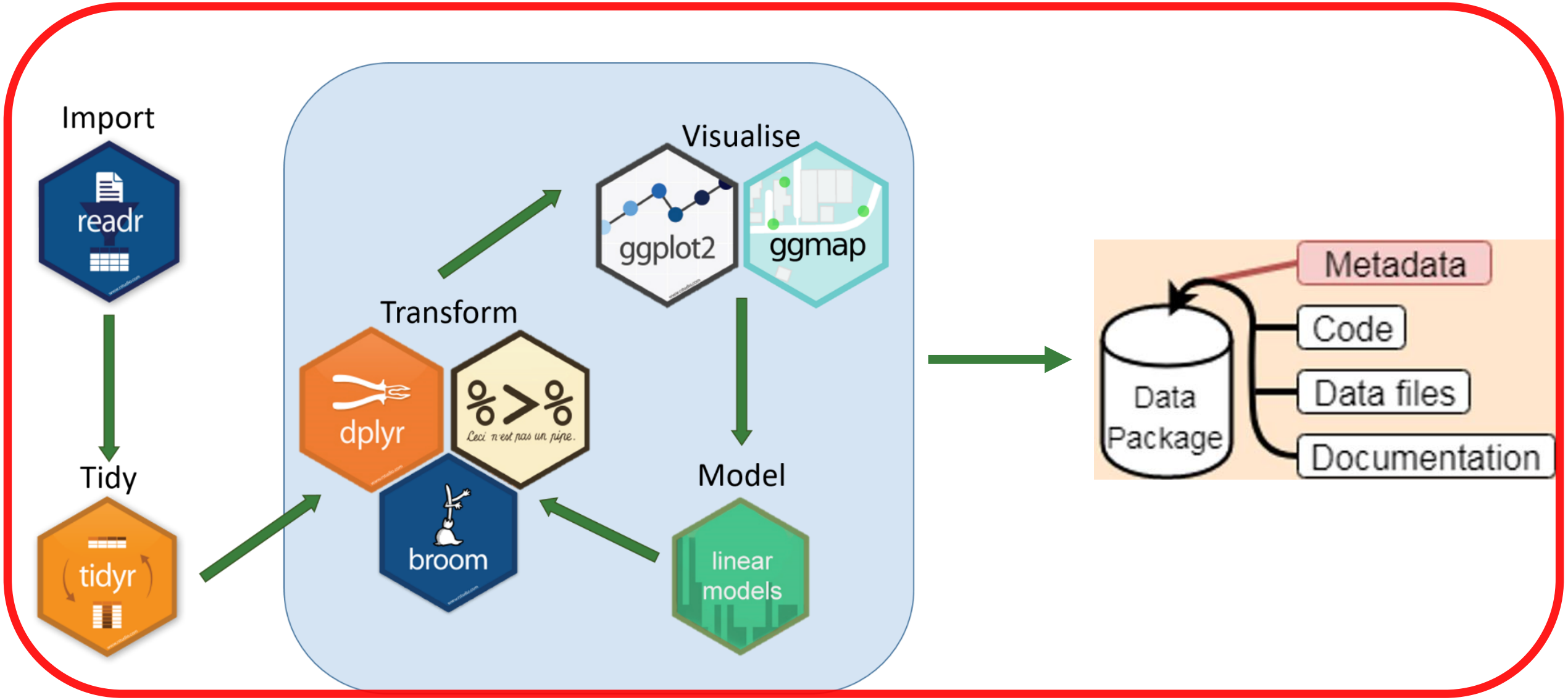
Exploring college major and income: a live data analysis in R — *October 16, 2018*
A live screencast of an exploratory data analysis from the Tidy Tuesday series. This one explores college major and income data from 538.

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.

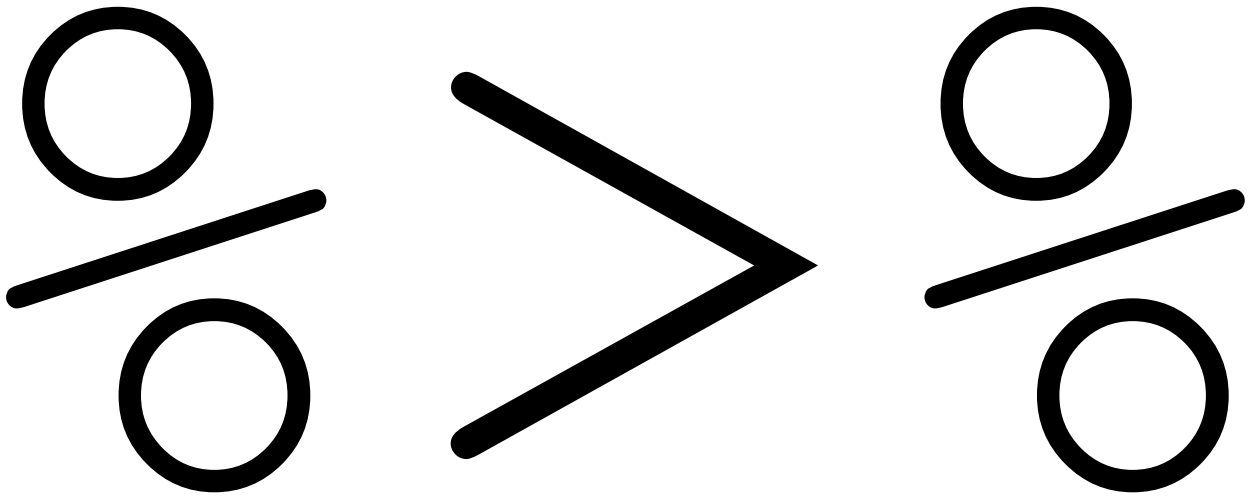
Scientific debt — *May 10, 2018*
Introducing an analogy to 'technical debt' for data scientists.

Putting the pieces together

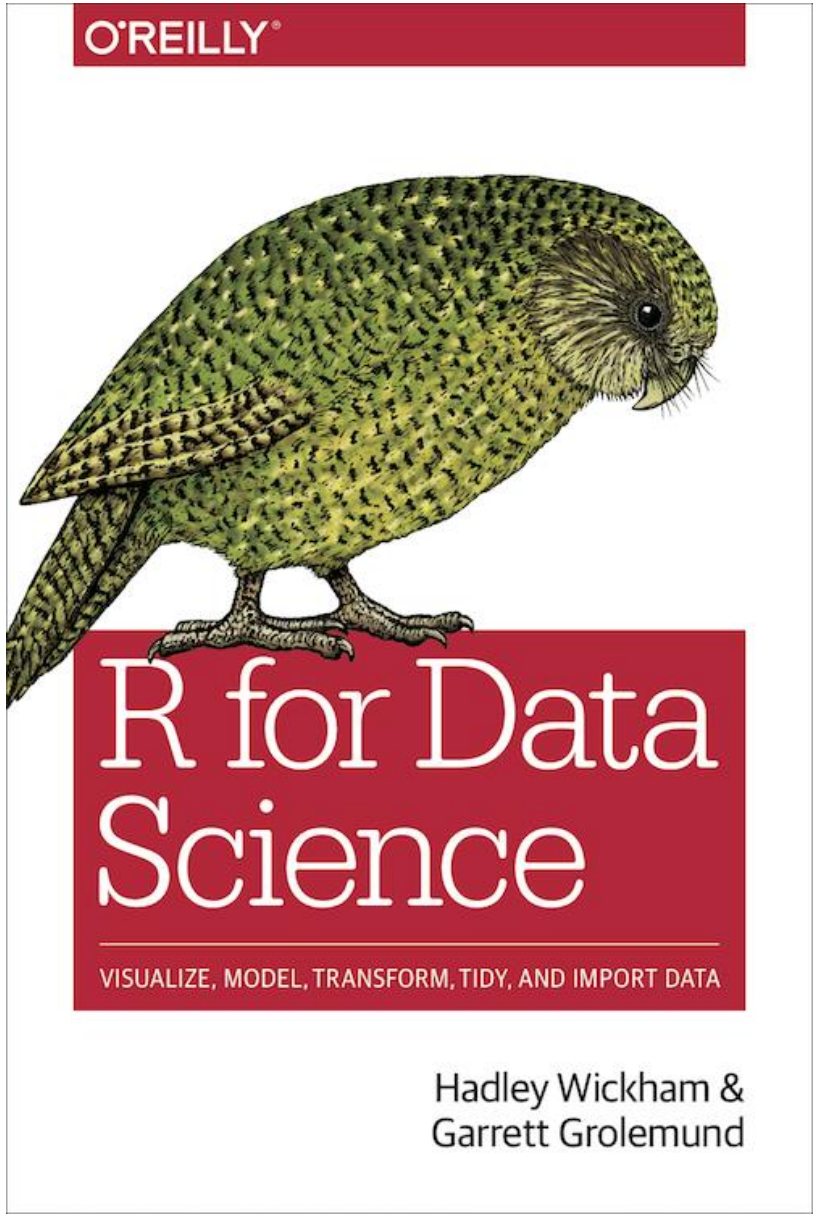
- Data analysis in a tidyverse nutshell



Putting the pieces together



You could write a book on that!!



R for Data Science webpage

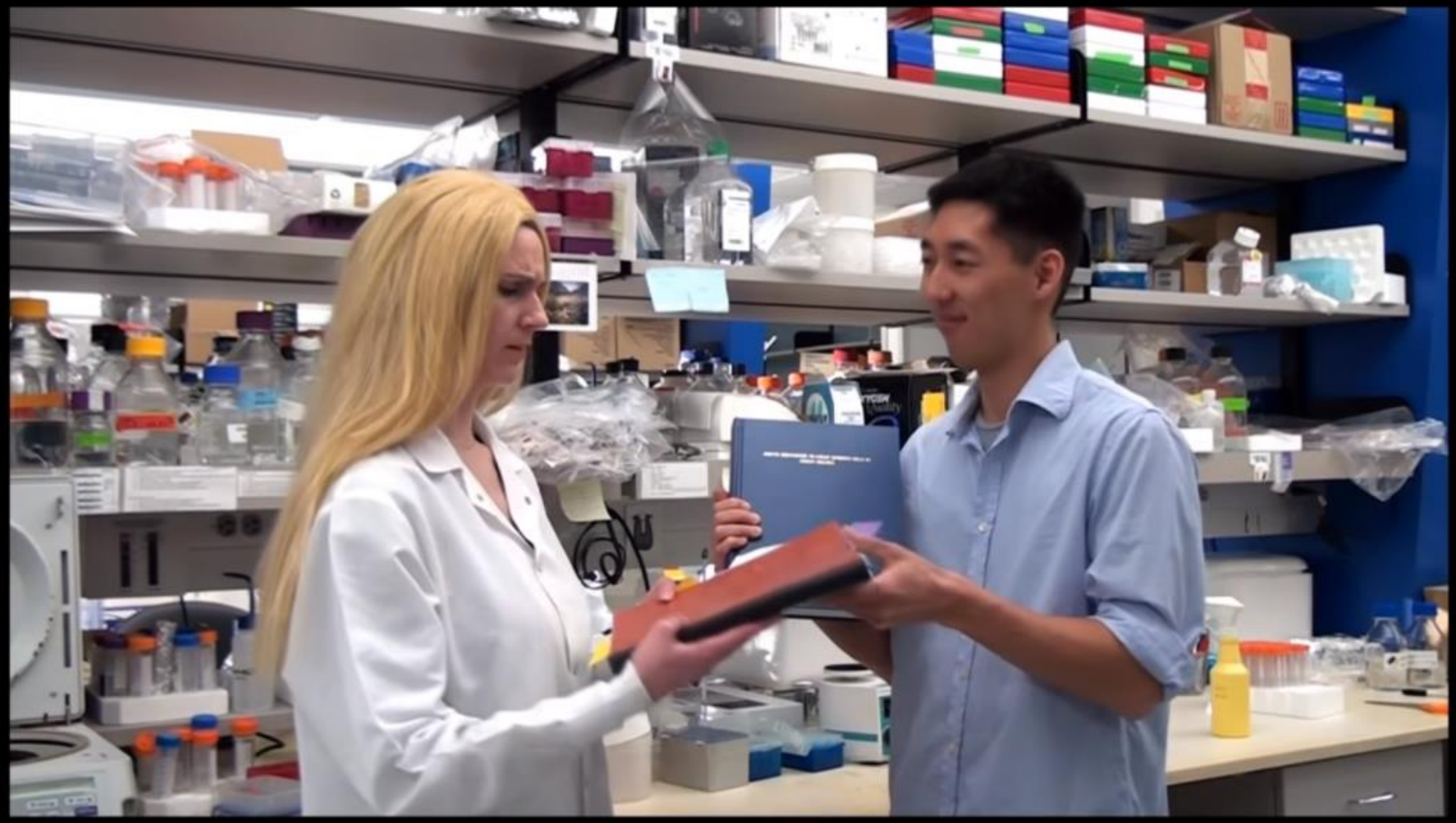
Putting the pieces together

A: Define a project structure

B: Set a naming convention

C: Use scripted workflows

D: Reproducible research



You were defending, one foot out the door

Link to the video on YouTube

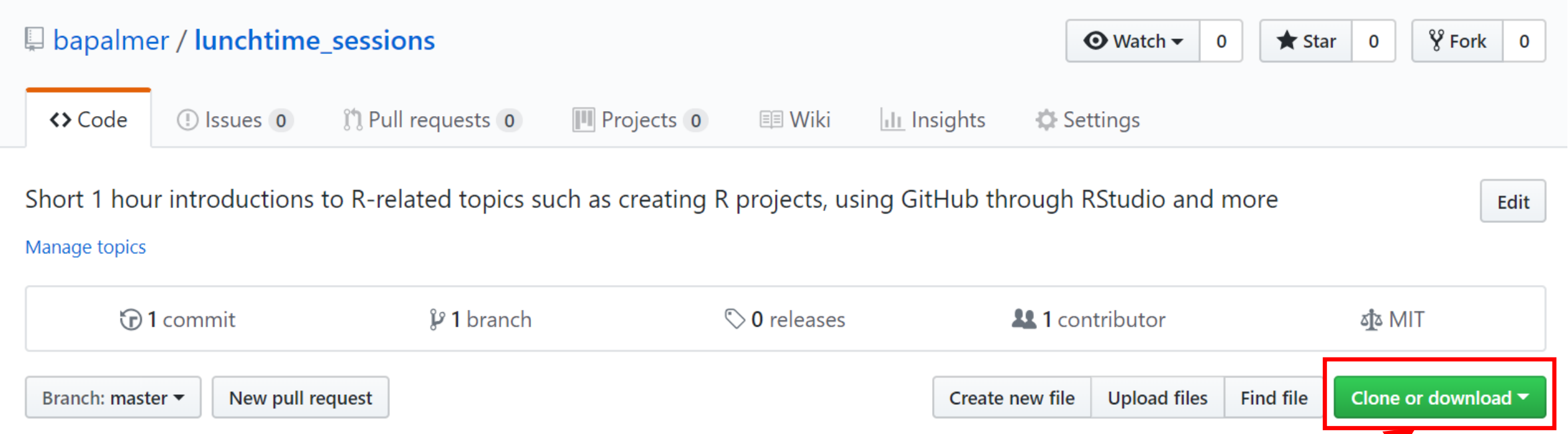
I got your project and its problems galore

Link to the video on YouTube

Link to the video on YouTube

R projects

- Here's one I made earlier……
https://github.com/bapalmer/lunchtime_sessions



- Open the R project folder and explore it

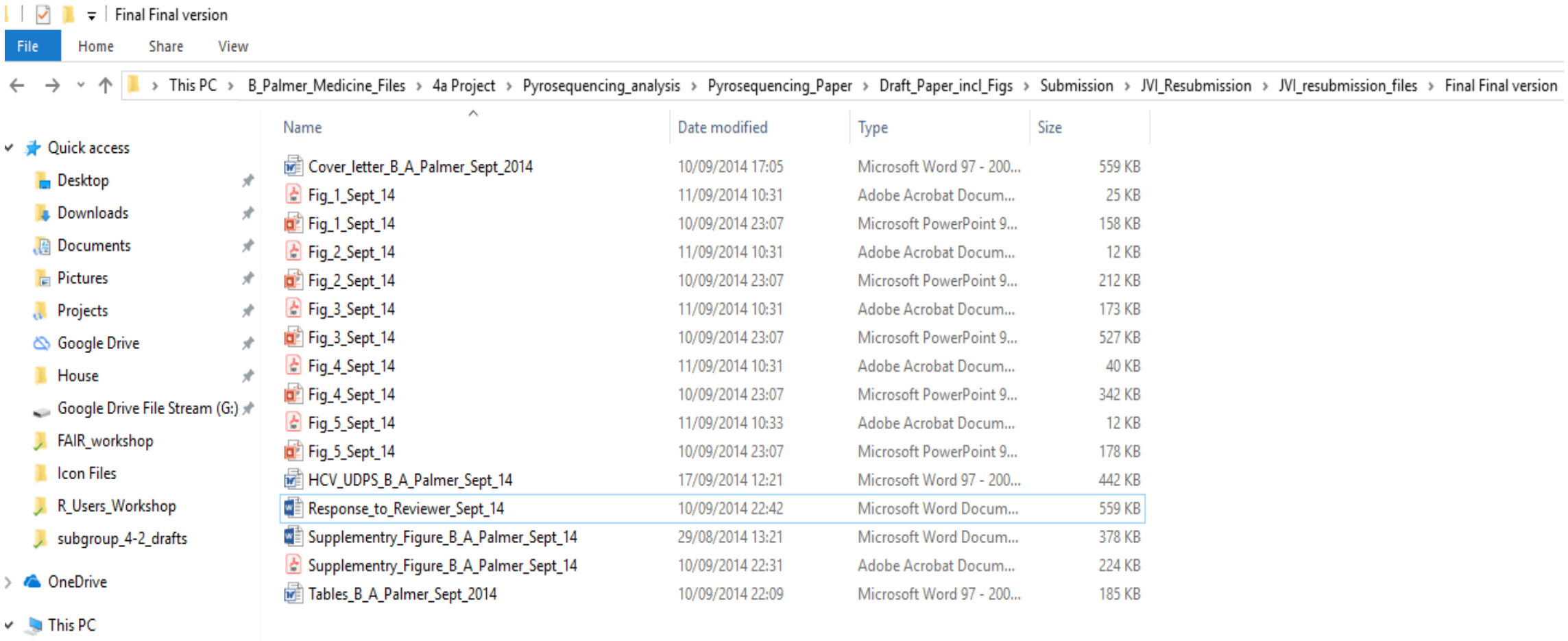
R projects

Documents › R_Users_Workshop › lunchtime_sessions-master › R-projects

Name	Date modified
data	14/11/2018 23:03
docs	15/11/2018 00:19
figs	14/11/2018 22:53
scripts	14/11/2018 23:48
tables	15/11/2018 00:19
R-projects	15/11/2018 00:27

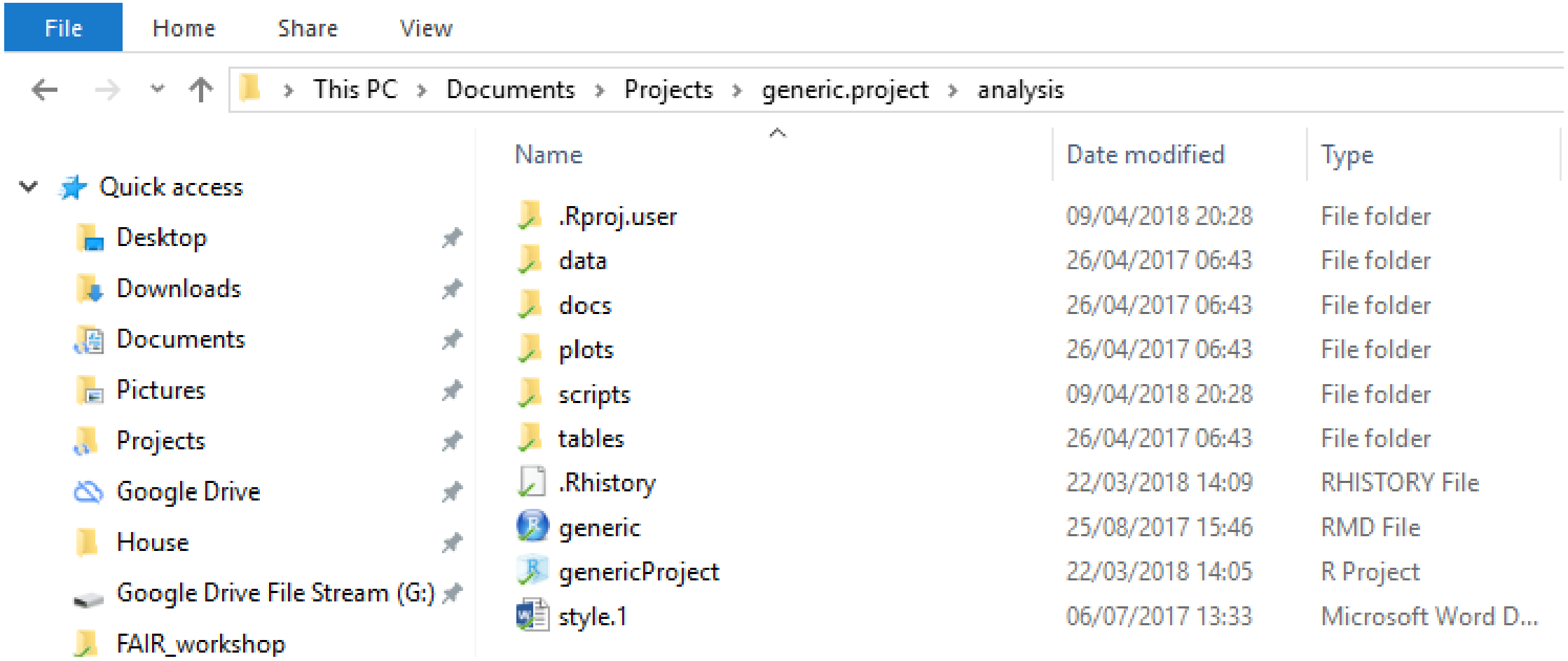
A: Still haven't found what I'm looking for

- Help your future-self



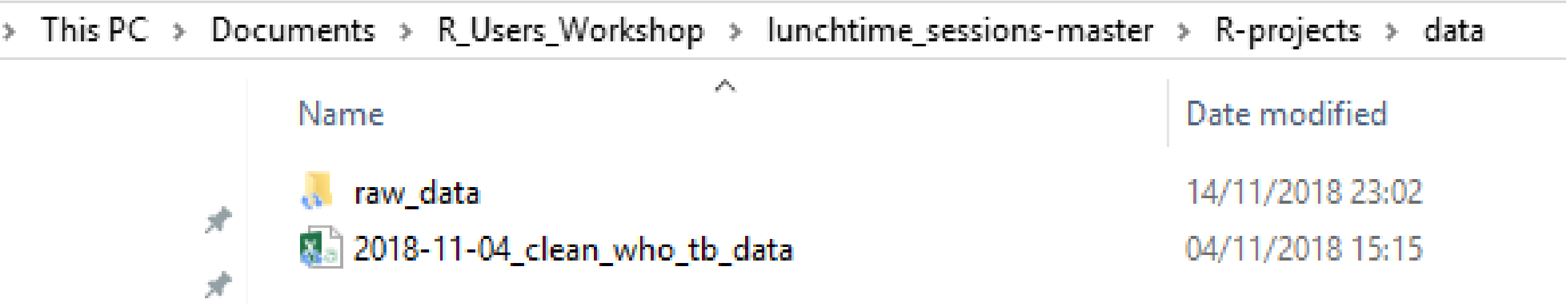
Define a generic project structure

- STEP 1: Give your research projects a shared structure



B: Give your files informative names

- STEP 2: Include metadata in the file names



Come back to what you know

- STEP 3: Make you file names machine readable, human readable and work with default ordering

NO

- Epistatic_change
- Epistatic_change_match_discovery
- Epistatic_change_match_discovery_fig_2_point_1
- Epistatic_change_v2
- epistatic_codon _change_tracking
- Epistatic_connection_network
- Heatmap_for_epistatic_syn
- Heatmap1_for_epi_site_co-change
- Heatmap2_for_epi_fdr_adjusted_p-value
- Heatmap2_for_epi_p-value

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_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
```

Image from Data Carpentry webpage

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_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
```

Image from Data Carpentry webpage

Yesterday

- Open the script 01_bad_habits.R

~/R_Users_Workshop/lunchtime_sessions/R-projects - master - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

01_bad_habits.R

```
20
21 sites1<-as.list(unique(RL6.7$Var1))
22 sites2<-as.list(unique(RL6.7$Var2))
23
24 sites<-as.data.frame(t(merge(sites1,sites2)))
25 colnames(sites)[1]<-"Position"
26
27 for(i in 1:nrow(sites)){
28 ans<-(sites$Position[i]<=65)
29 sites$E1[i]<-ans
30 }
31
32 # Start building network
33 RL6.7_topology<-subset(RL6.7[2:3])
34 g2<-graph.data.frame(RL6.7_topology,vertices=sites,directed=FALSE)
35 g<-simplify(g2)
36 V(g)$color<-ifelse(V(g)$E1==TRUE,"white","grey")
37 V(g)$color<-ifelse(V(g)$E1==TRUE,"white","grey")
```

Environment History Connections Git

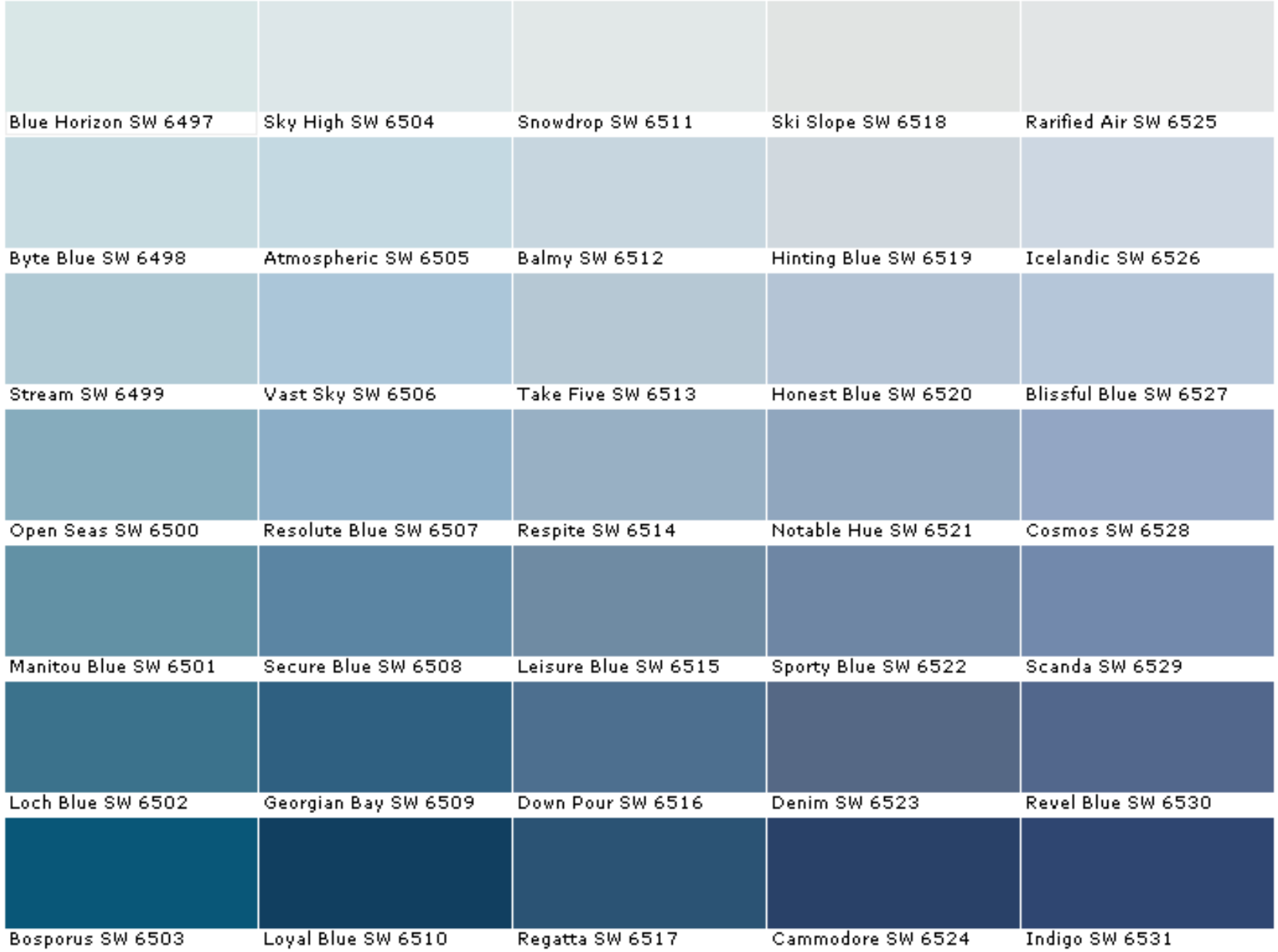
Global Environment

Name	Type	Length	Size	Value
ans	logical	1	56 B	TRUE
df	tbl_df	4	5.4 KB	39 obs. of 4 variables
g	igraph	10	2.5 KB	List of 10
g2	igraph	10	2.1 KB	List of 10
i	integer	1	56 B	4L
RL1.2	tbl_df	4	1.3 KB	3 obs. of 4 variables
RL2.3	tbl_df	4	1.3 KB	3 obs. of 4 variables
RL3.4	tbl_df	4	1.3 KB	1 obs. of 4 variables
RL4.5	tbl_df	4	1.6 KB	9 obs. of 4 variables
RL5.6	tbl_df	4	1.8 KB	20 obs. of 4 variables
RL6.7	tbl_df	4	1.3 KB	2 obs. of 4 variables
RL6.7_topolo...	tbl_df	2	1000 B	2 obs. of 2 variables
sites	data.frame	2	1.1 KB	4 obs. of 2 variables
sites1	list	2	176 B	List of 2
sites2	list	2	176 B	List of 2

Lack of annotation
Poor naming conventions
Poor readability
Spacing absent

Cluttered environment
Intermediate objects

Is too much choice good or bad?



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)
```

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

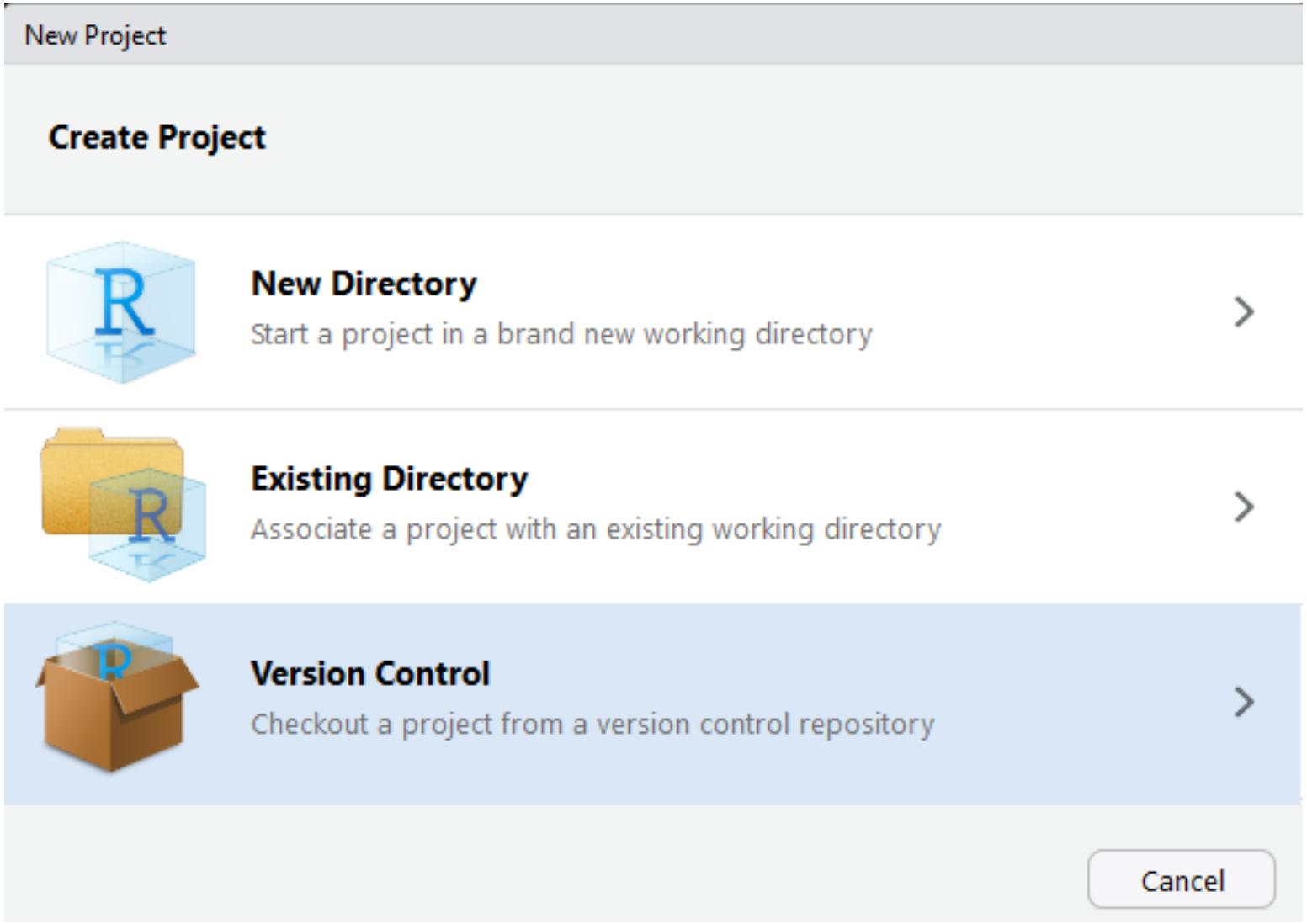
C: Joined up thinking

- 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

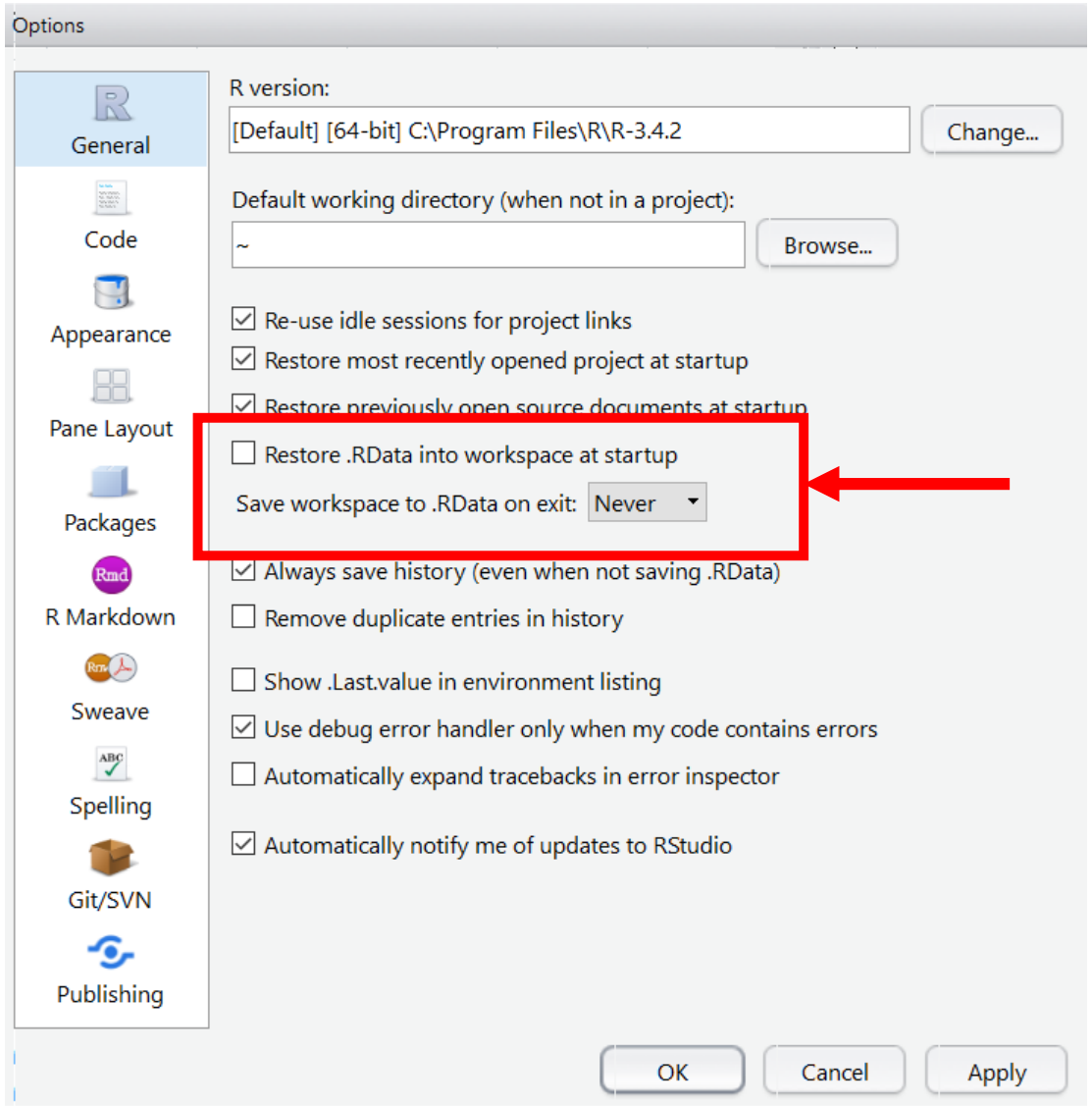
```r
# Data ----
# Eight tibbles returned from the 01_data_import_and_tidying_master_file.R
# 1. fgf23_data => FGF23 readings from study centres 01-03
# 2. food_level_data => Food diary entries
# 3. grouped_data => Dialysis and nondialysis diary entries by component
# 4. k_data => Serum potassium
# 5.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

source("scripts/01_data_import_and_tidying_master_file.R")
```

D: Setting up an R project



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

Don’t Do What Donny Don’t Does!!

Donny Don’t:

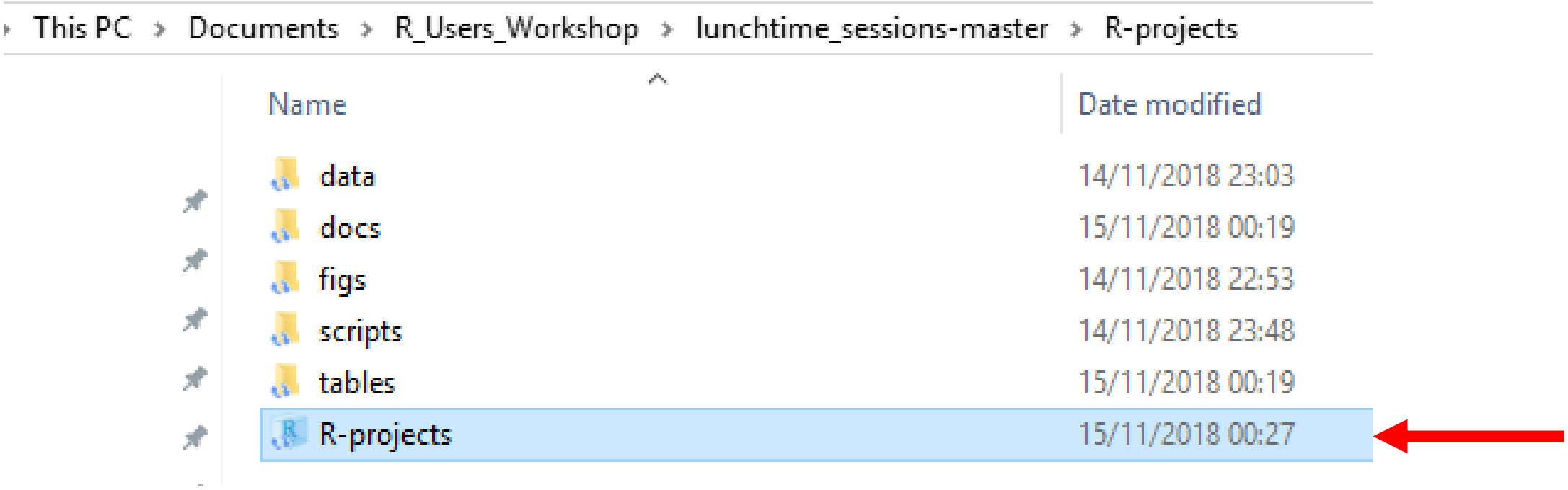
- Start your script with…
  `setwd()`

Donny Don’t:

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

Everything in its right place

- benefits of using R projects for data analysis tasks



- Open the script 03_data_cleaning.R
- Open the script 04_plots.R
- Open the script 05_tables.R
- Open the script 06_analysis.R

Meetup

Cork (Ireland) R-Users Group