R-eproducible workflows

1-day workshop
Morning lecture

RStudio tidyverse

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

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

CRF-C
HRB Clinical Research Facility Cork

HRB
Health
Research
Board

Meetup

Cork (Ireland) R-Users Group

Plan for the day

Morning: Reproducible research through R/RStudio

10 am: Lecture - Project orientated workflows

10.30 am: Coffee break – Discussion

10.45 am: 2 × 30 minute tutorials

- Joined up thinking when writing R code
- R-projects as means to organise your research

11.45 pm: Lecture - Introduction to the tidyverse

12.30 pm: Lunch break

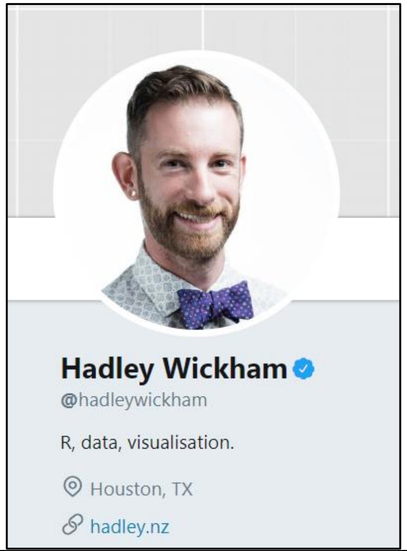
Afternoon: A crash course in the tidyverse

1.15 pm: 3 × 45 minute hands-on tidyverse tutorials including;

- Differences between the tidyverse and base R code
- Example scripts and problem sheets exploring R packages
- Useful add on packages

3.30 pm: Closing remarks, questions

Disclaimer 1



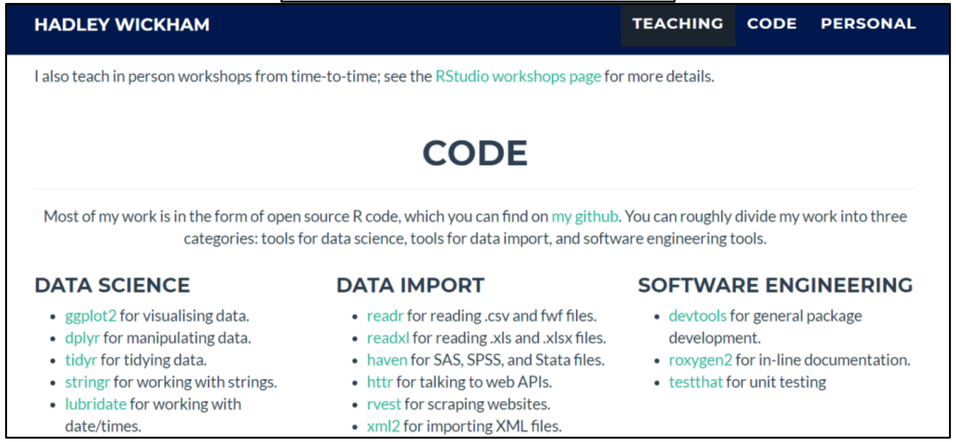
Hadley Wickham

@hadleywickham

R, data, visualisation.

Houston, TX

hadley.nz



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

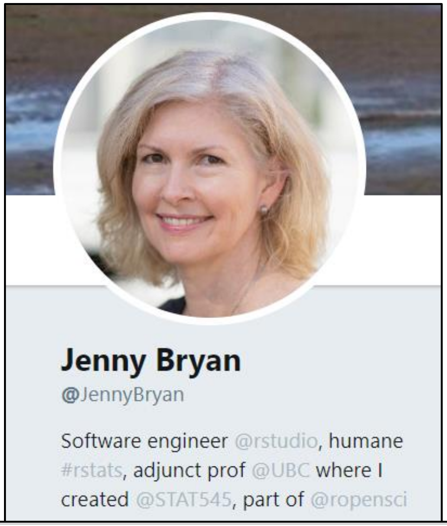
- ggplot2 for visualising data.
- dplyr for manipulating data.
- tidyr for tidying data.
- stringr for working with strings.
- lubridate for working with date/times.

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

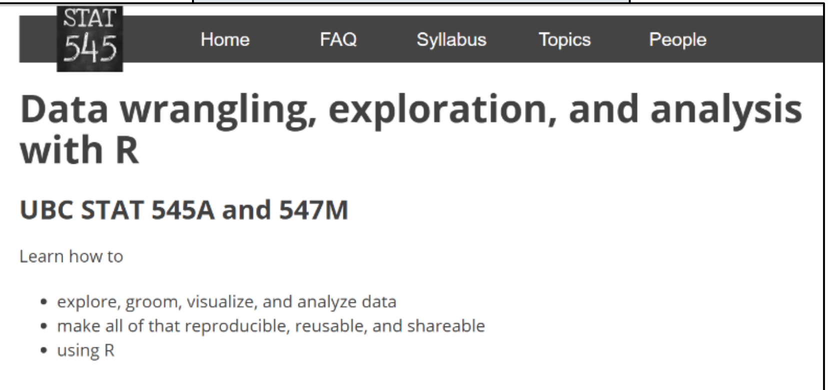
- devtools for general package development.
- roxygen2 for in-line documentation.
- testthat for unit testing



Jenny Bryan

@JennyBryan

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

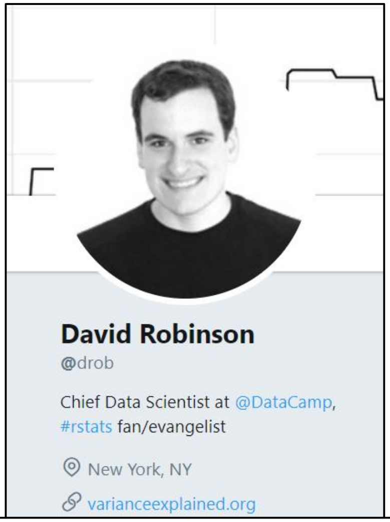


Data wrangling, exploration, and analysis with R

UBC STAT 545A and 547M

Learn how to

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



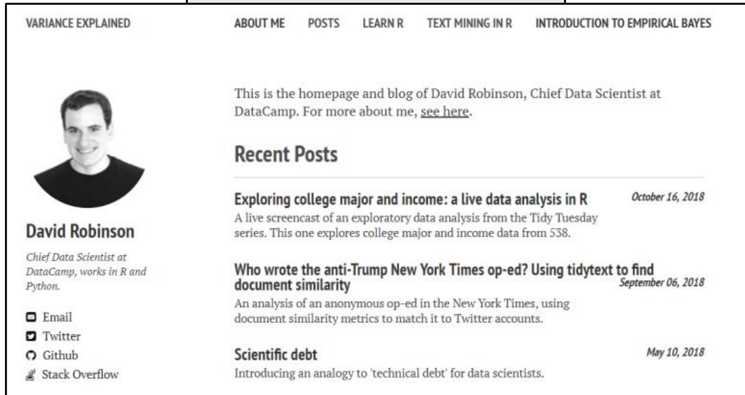
David Robinson

@drob

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

New York, NY

varianceexplained.org



David Robinson

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

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*

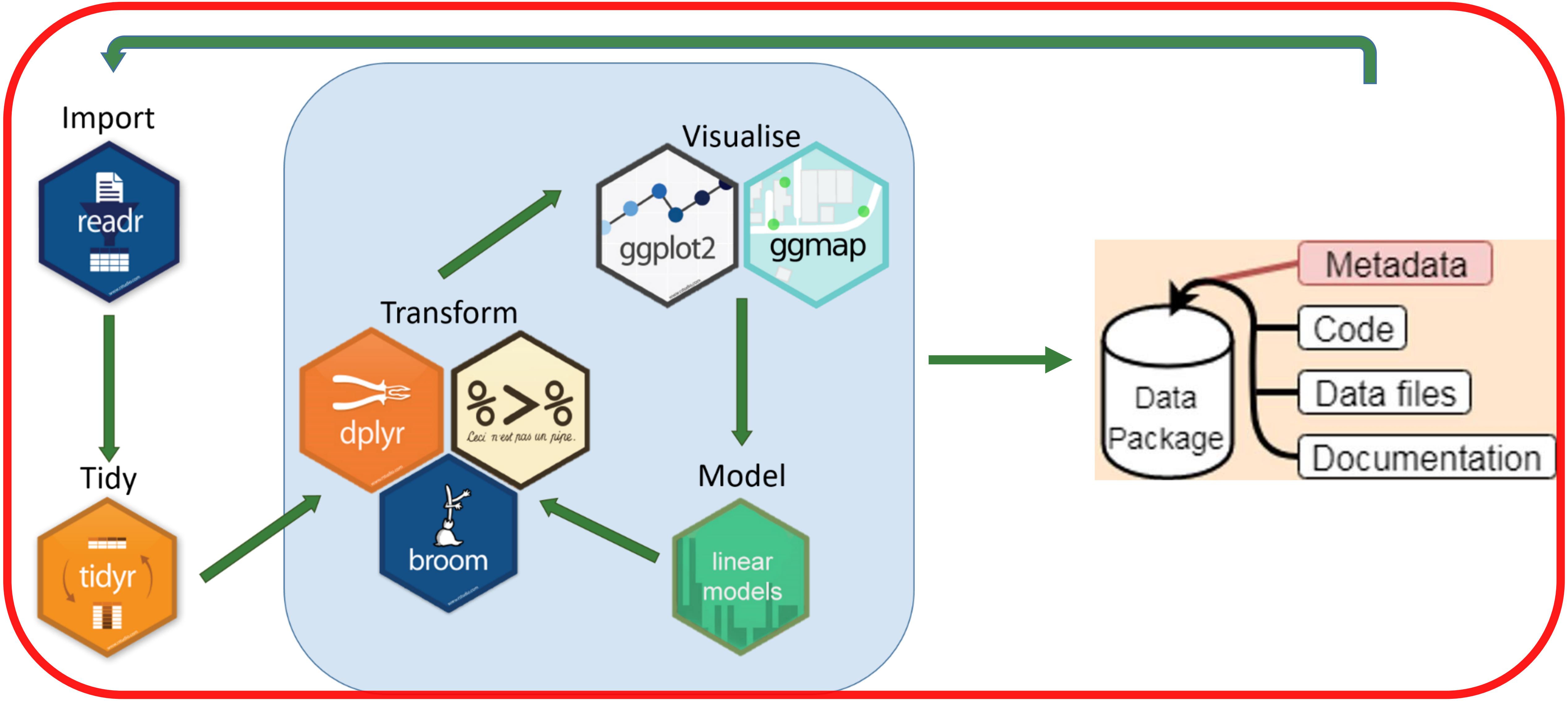
An analysis of an anonymous op-ed in the New York Times, using document similarity metrics to match it to Twitter accounts.

Scientific debt *May 10, 2018*

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

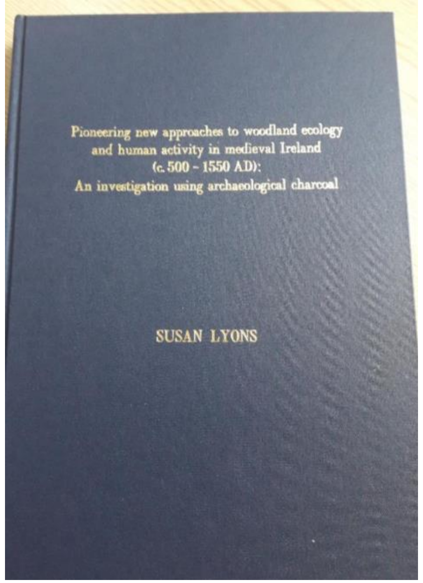
Putting the pieces together

- Data analysis in a tidyverse nutshell



How is research presented?

Theses


Pioneering new approaches to woodland ecology
and human activity in medieval Ireland
(c. 500 - 1550 AD):
An investigation using archaeological charcoal

SUSAN LYONS

Papers

AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Virology

CrossMark

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.

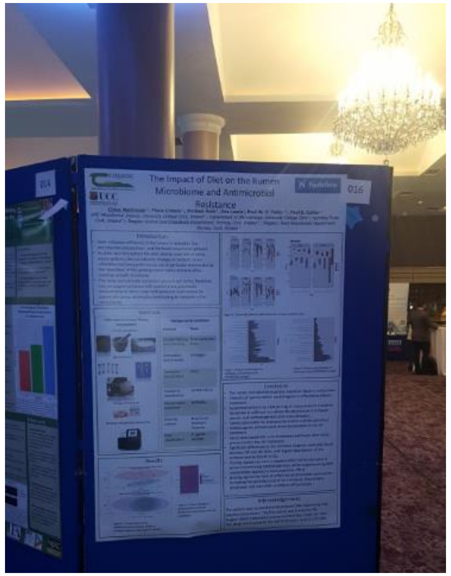
Copyright © 2016, American Society for Microbiology. All Rights Reserved.

3318 jvi.asm.org Journal of Virology April 2016 Volume 90 Number 7

Books

Implementation Guide and Toolkit
for National Clinical Guidelines

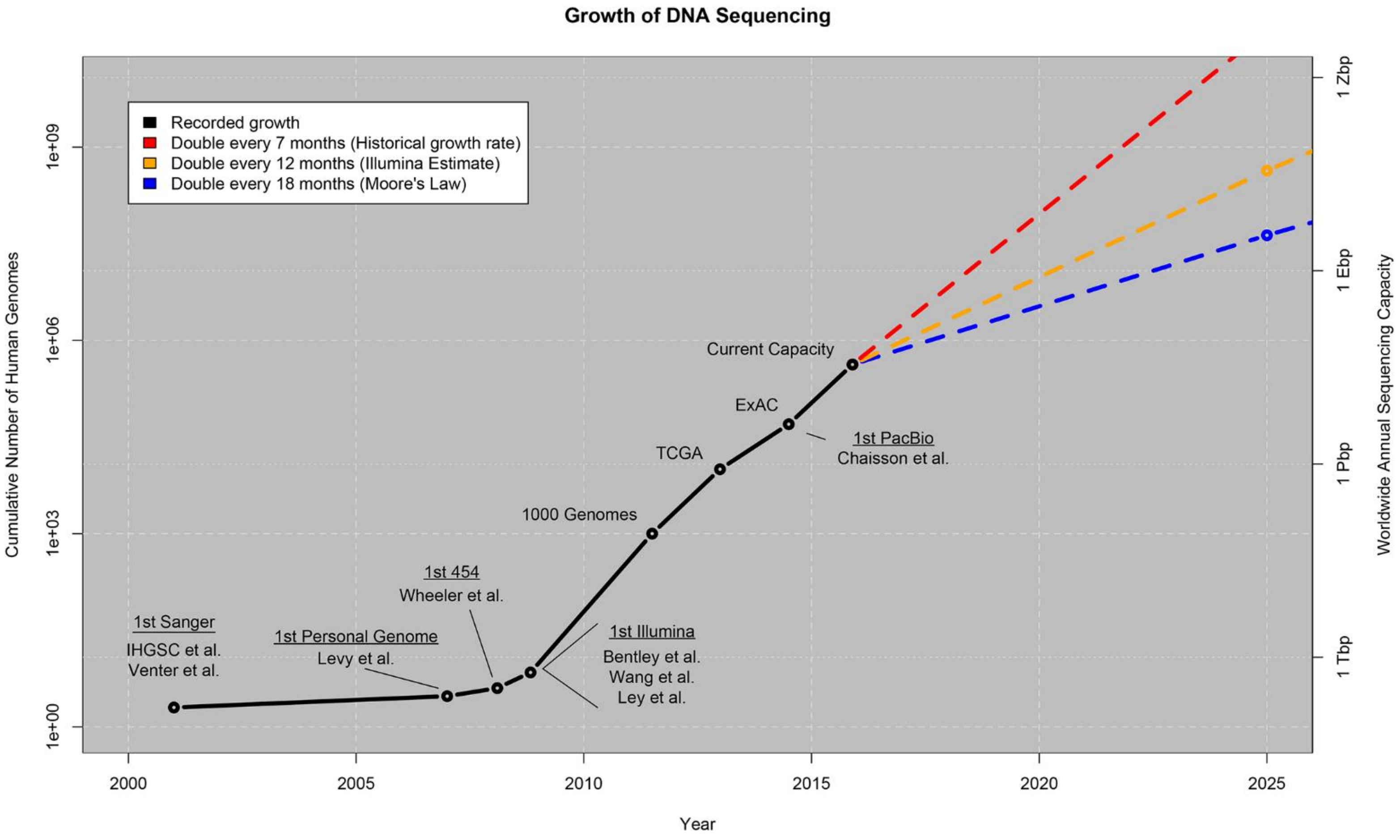
Posters

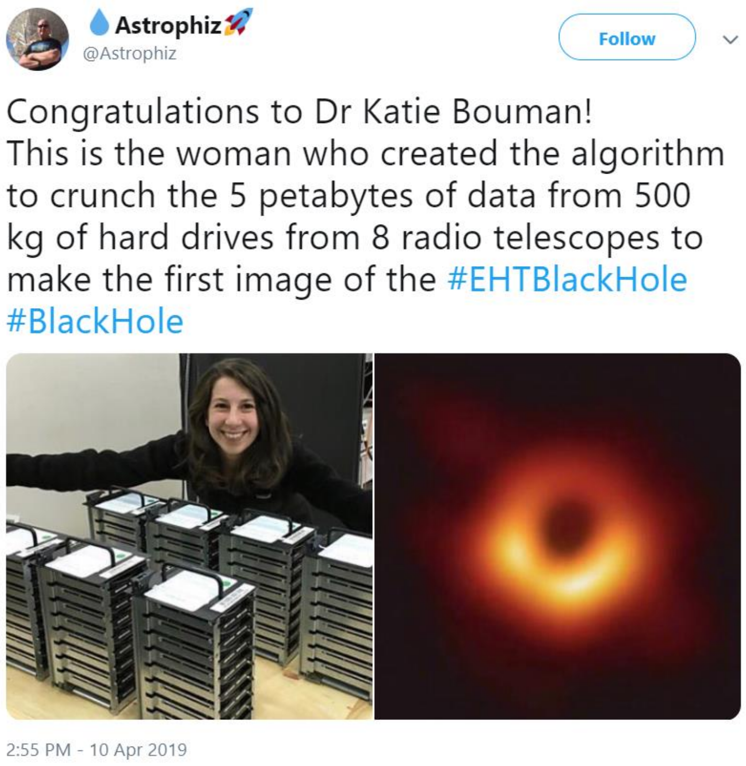


Talks

But what does it look like under the bonnet?

The explosion of data


Growth of DNA Sequencing



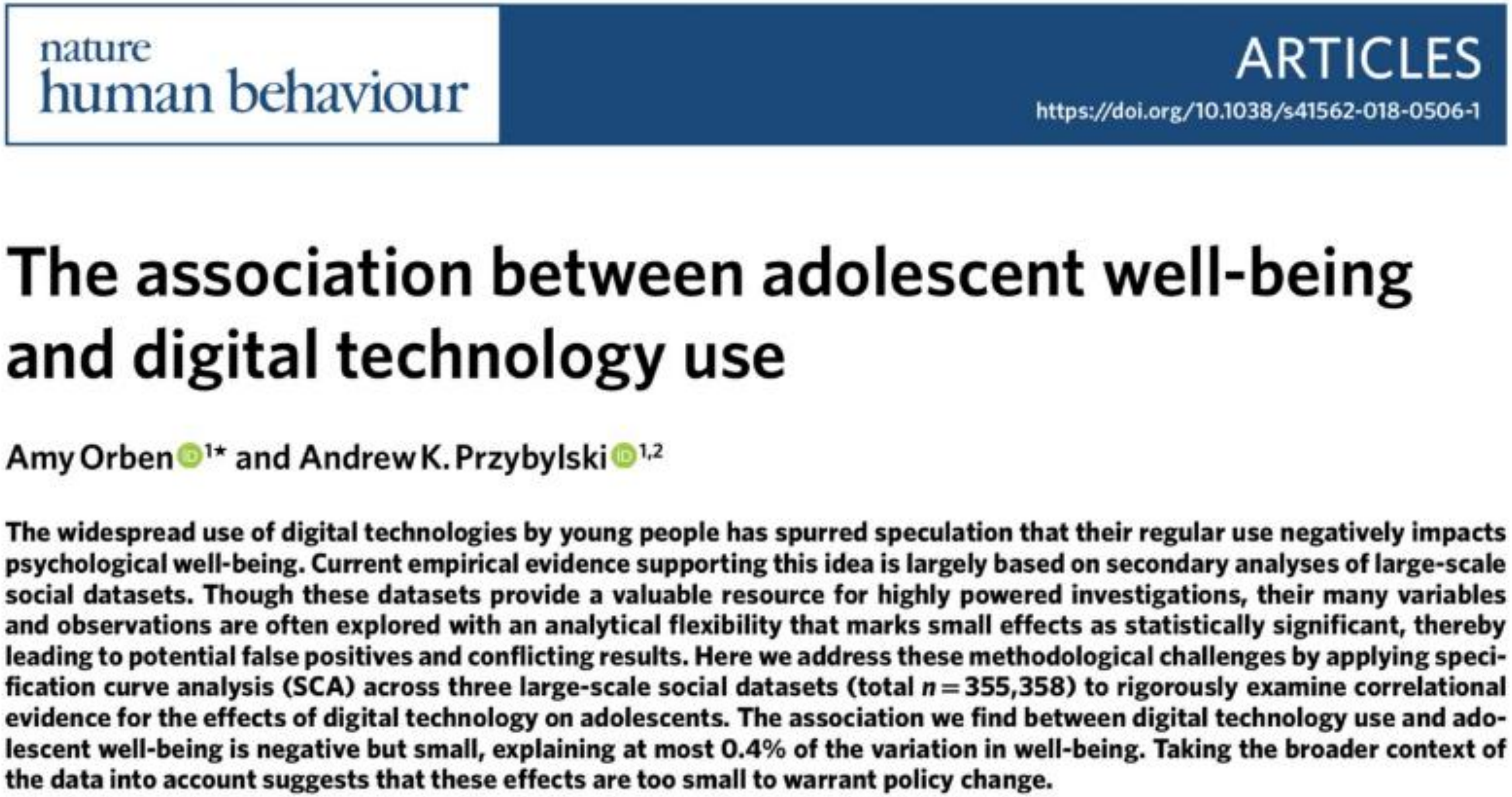
💧Astrophiz🚀
@Astrophiz

Congratulations to Dr Katie Bouman!
This is the woman who created the algorithm to crunch the 5 petabytes of data from 500 kg of hard drives from 8 radio telescopes to make the first image of the #EHTBlackHole #BlackHole

2:55 PM - 10 Apr 2019

Stevens et. al., 2015, Plos Biology

Large quantities of data ≠ high quality of science



nature
human behaviour

ARTICLES

https://doi.org/10.1038/s41562-018-0506-1

The association between adolescent well-being and digital technology use

Amy Orben[1*] and Andrew K. Przybylski[1,2]

The widespread use of digital technologies by young people has spurred speculation that their regular use negatively impacts psychological well-being. Current empirical evidence supporting this idea is largely based on secondary analyses of large-scale social datasets. Though these datasets provide a valuable resource for highly powered investigations, their many variables and observations are often explored with an analytical flexibility that marks small effects as statistically significant, thereby leading to potential false positives and conflicting results. Here we address these methodological challenges by applying specification curve analysis (SCA) across three large-scale social datasets (total $n = 355{,}358$) to rigorously examine correlational evidence for the effects of digital technology on adolescents. The association we find between digital technology use and adolescent well-being is negative but small, explaining at most 0.4% of the variation in well-being. Taking the broader context of the data into account suggests that these effects are too small to warrant policy change.

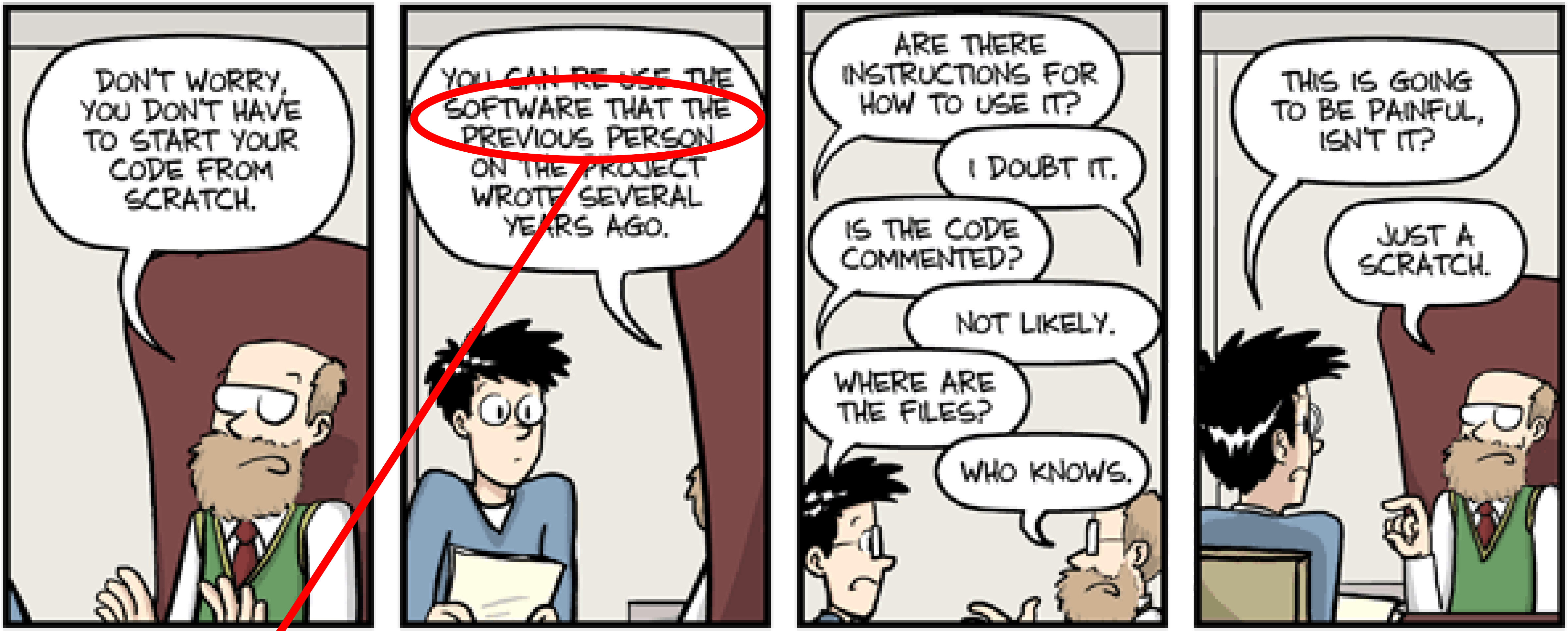
Orben & Przybylski (2019), Nature Human Behaviour

You were defending, one foot out the door

[Zheng Lab - Bad Project (Lady Gaga parody)]

This session

- **Project structure**
  - **Naming conventions**
    - **Scripted workflows**
      - R Markdown
        - **Reproducible research**

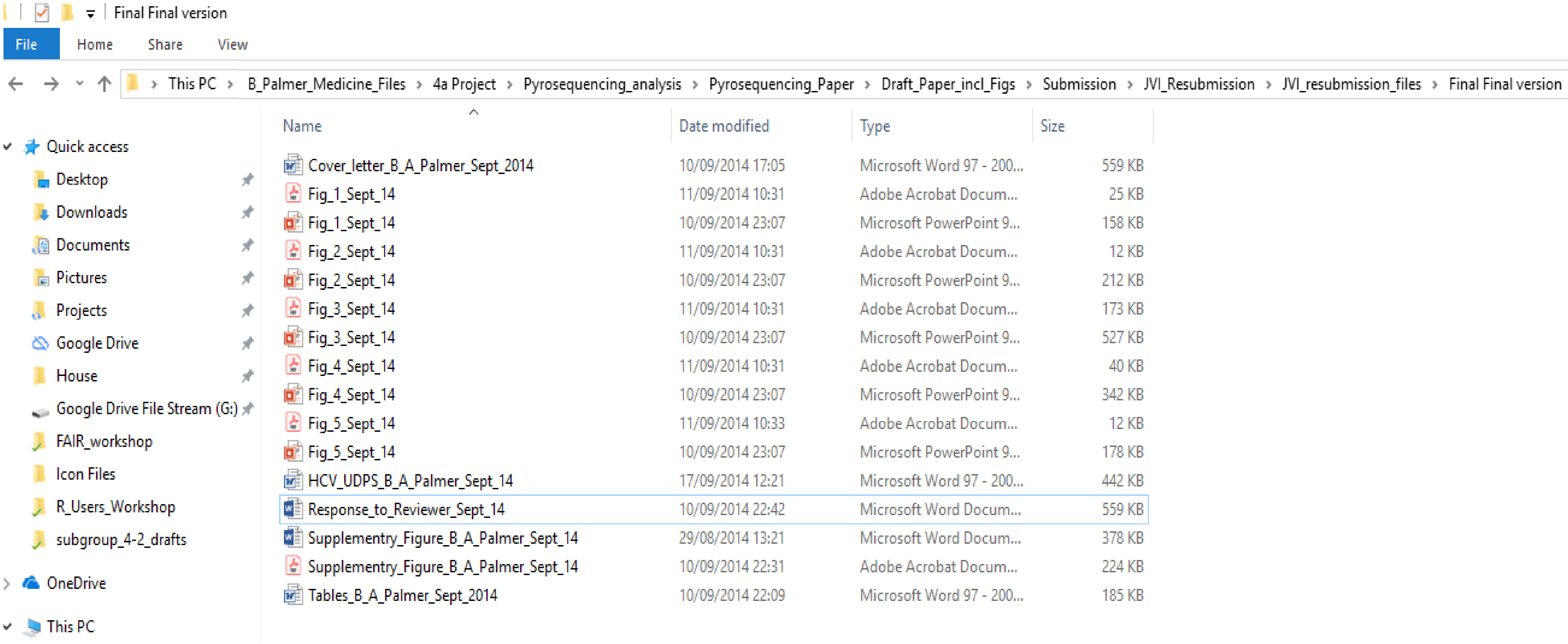

DON'T WORRY, YOU DON'T HAVE TO START YOUR CODE FROM SCRATCH.
YOU CAN RE-USE THE SOFTWARE THAT THE PREVIOUS PERSON ON THE PROJECT WROTE SEVERAL YEARS AGO.
ARE THERE INSTRUCTIONS FOR HOW TO USE IT?
I DOUBT IT.
IS THE CODE COMMENTED?
NOT LIKELY.
WHERE ARE THE FILES?
WHO KNOWS.
THIS IS GOING TO BE PAINFUL, ISN'T IT?
JUST A SCRATCH.

THIS PERSON IS likely to be YOU BTW!!

Jorge Cham | www.phdcomics.com

Still haven’t found what I’m looking for

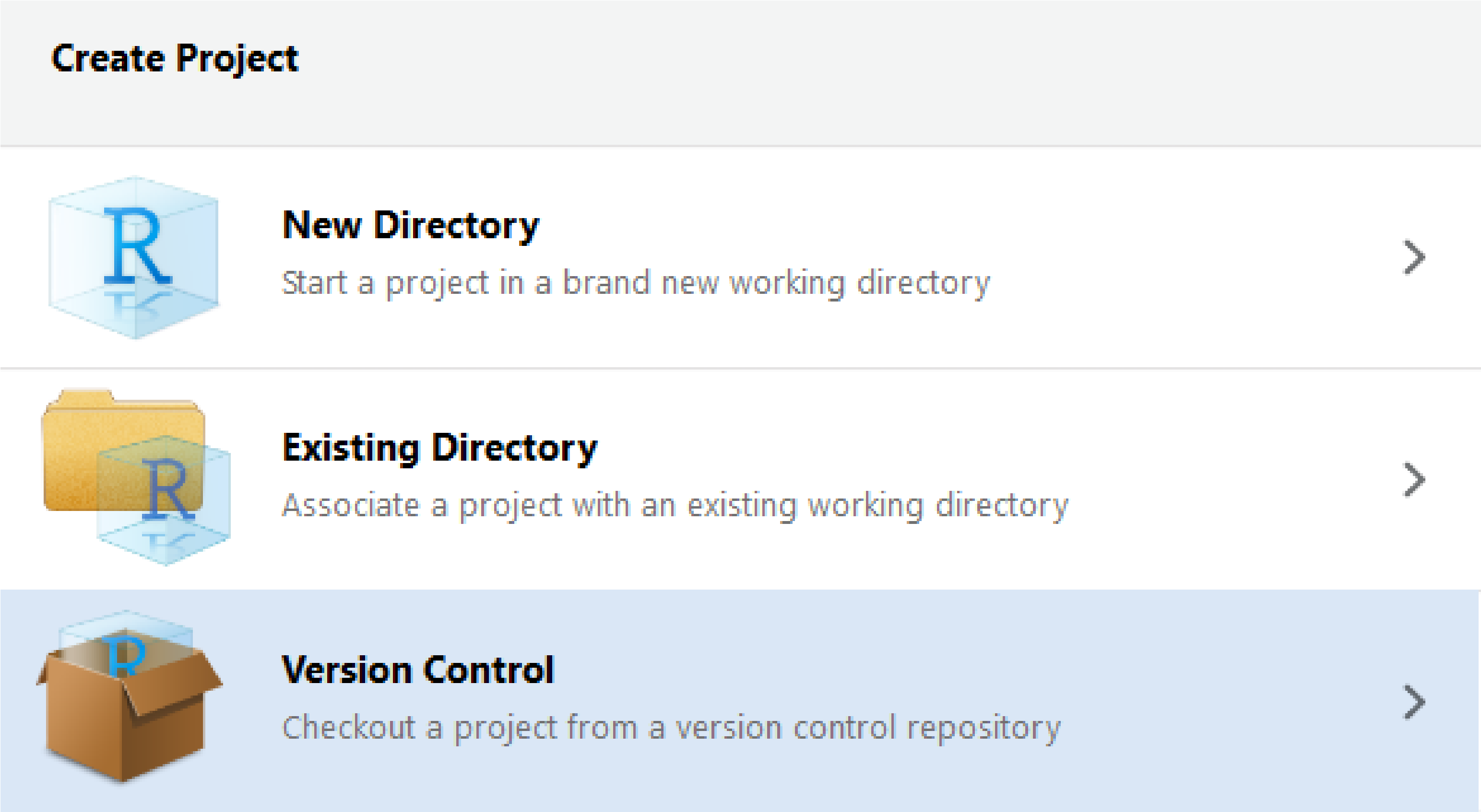
- **Help your future-self**



This PC › B_Palmer_Medicine_Files › 4a Project › Pyrosequencing_analysis › Pyrosequencing_Paper › Draft_Paper_incl_Figs › Submission › JVI_Resubmission › JVI_resubmission_files › Final Final version

Name	Date modified	Type	Size
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
Supplementry_Figure_B_A_Palmer_Sept_14	29/08/2014 13:21	Microsoft Word Docum...	378 KB
Supplementry_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

R-projects



Create Project

New Directory

Start a project in a brand new working directory

Existing Directory

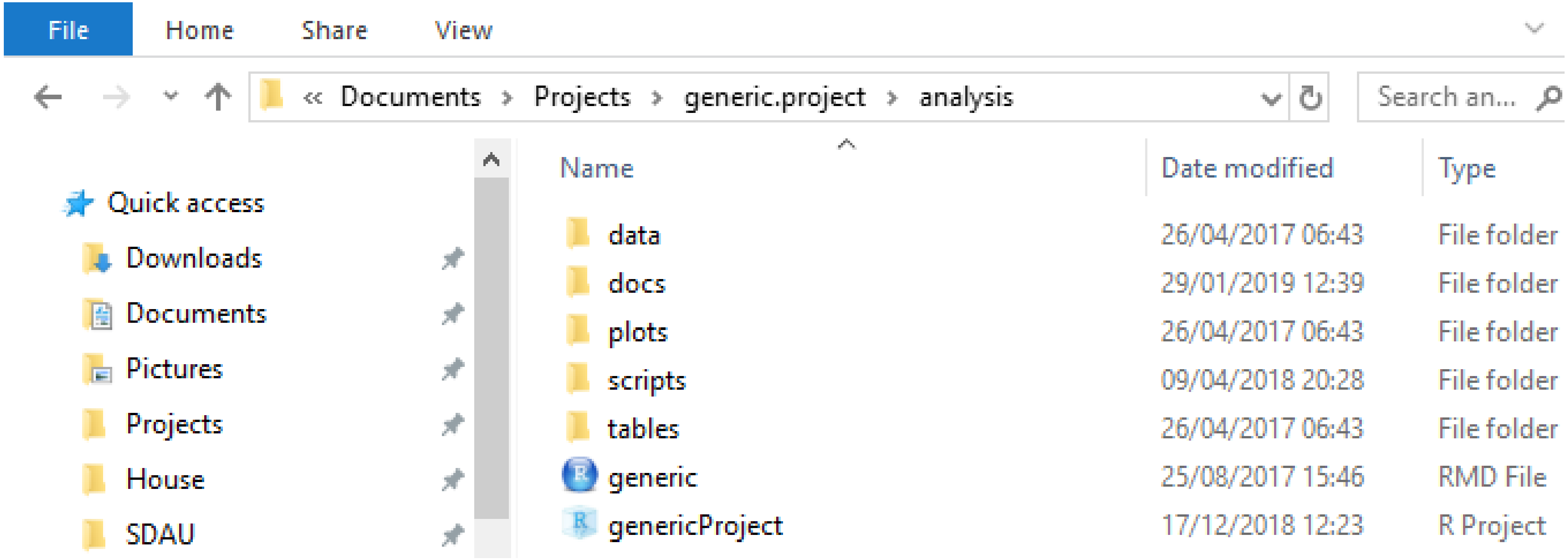
Associate a project with an existing working directory

Version Control

Checkout a project from a version control repository

Define a generic project structure

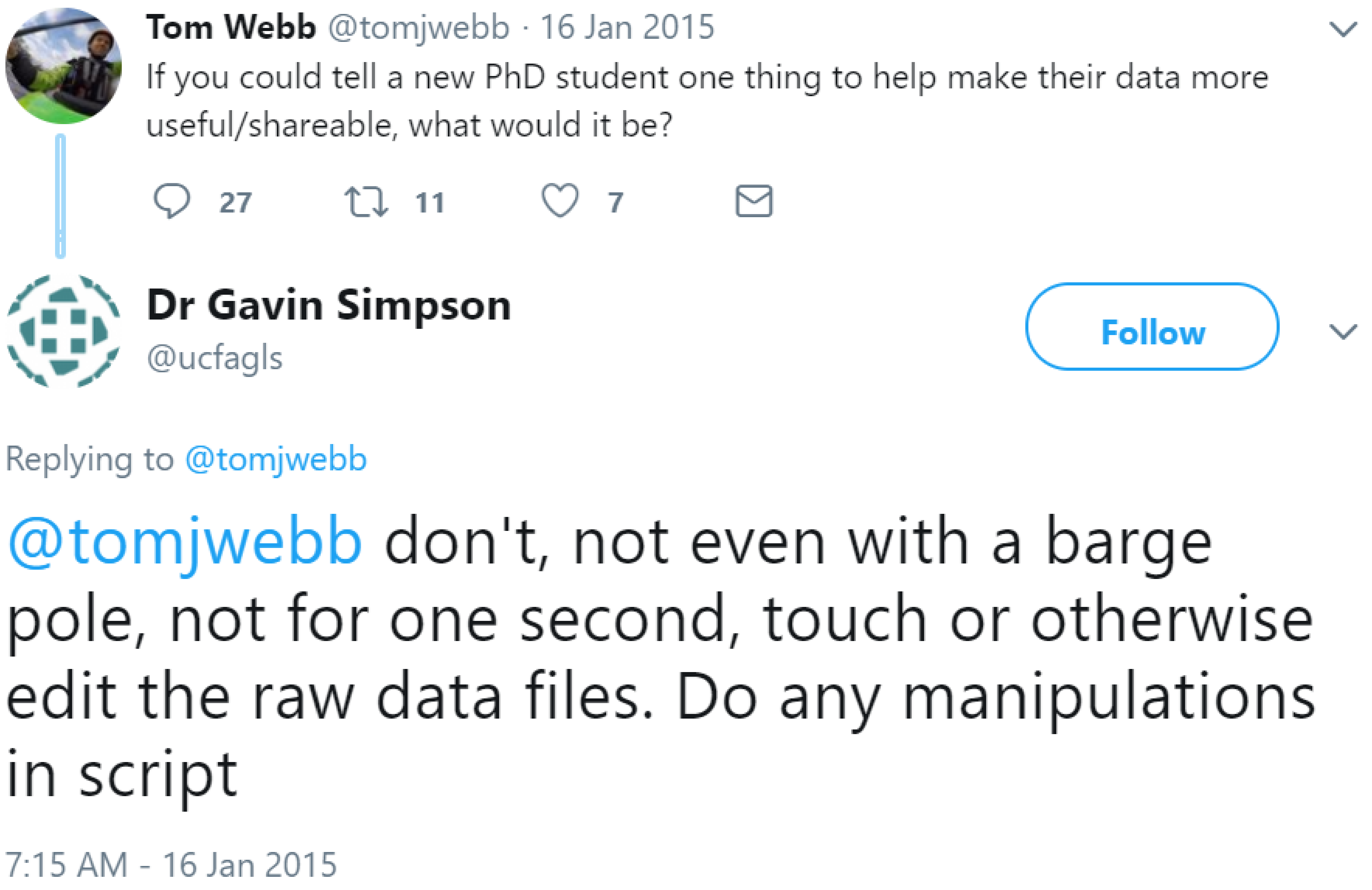
- STEP 1: Give your research projects a shared structure



« Documents › Projects › generic.project › analysis

Name	Date modified	Type
data	26/04/2017 06:43	File folder
docs	29/01/2019 12:39	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
generic	25/08/2017 15:46	RMD File
genericProject	17/12/2018 12:23	R Project

Work from the raw data ALWAYS!!



Tom Webb @tomjwebb · 16 Jan 2015

If you could tell a new PhD student one thing to help make their data more useful/shareable, what would it be?

Dr Gavin Simpson
@ucfagls

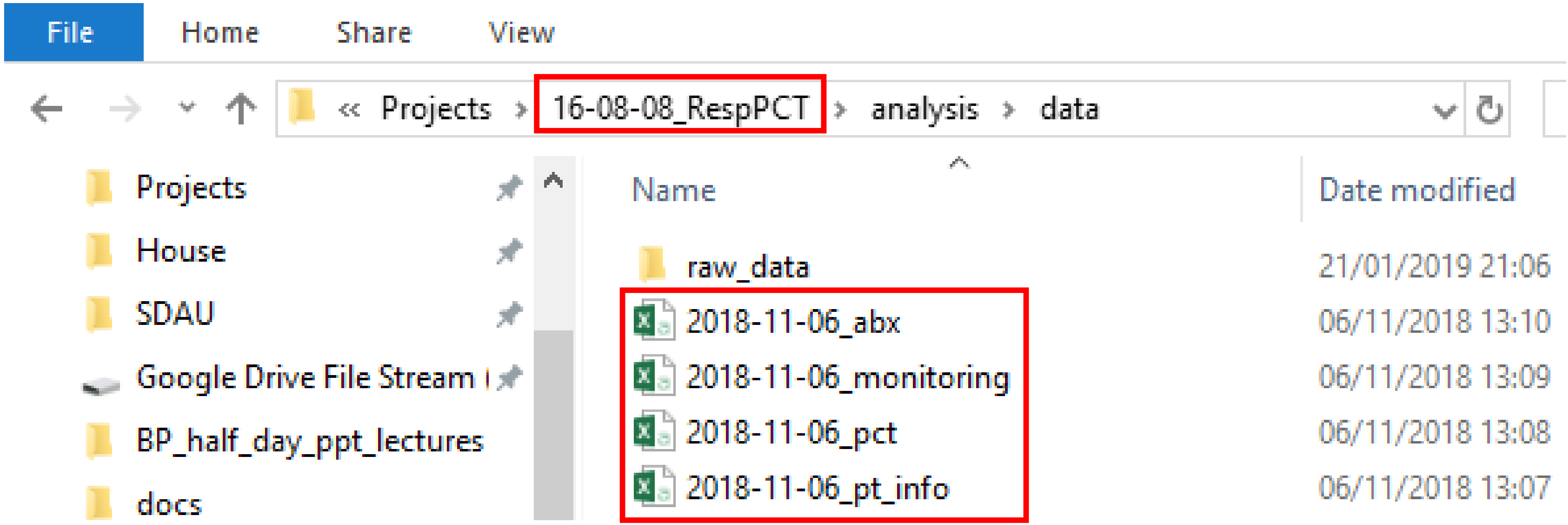
Replying to @tomjwebb

@tomjwebb don't, not even with a barge pole, not for one second, touch or otherwise edit the raw data files. Do any manipulations in script

7:15 AM - 16 Jan 2015

Give your files informative names

- STEP 1: Give your research projects a shared structure



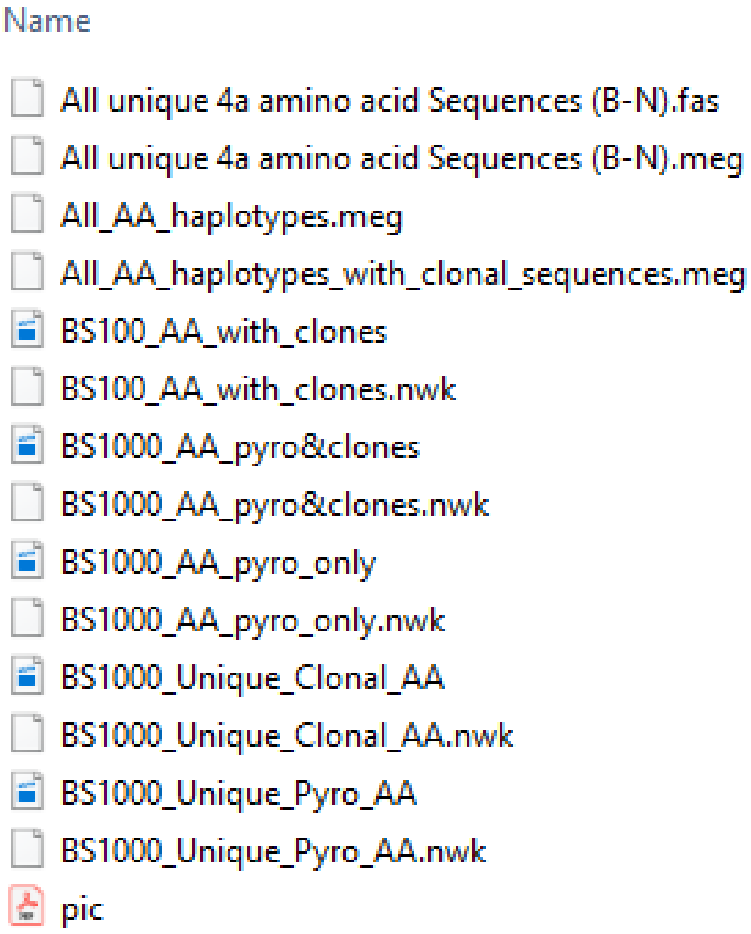
« Projects › 16-08-08_RespPCT › analysis › data

Name	Date modified
raw_data	21/01/2019 21:06
2018-11-06_abx	06/11/2018 13:10
2018-11-06_monitoring	06/11/2018 13:09
2018-11-06_pct	06/11/2018 13:08
2018-11-06_pt_info	06/11/2018 13:07

Everything in its right place

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

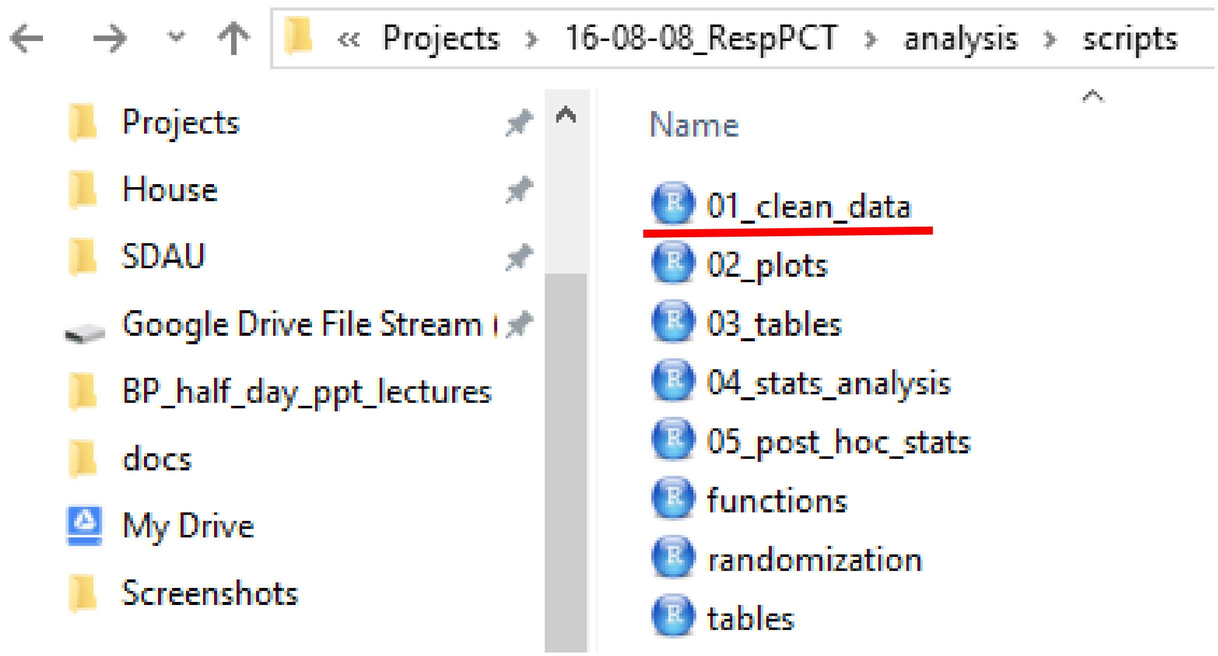
NO



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_Unique_Clonal_AA
BS1000_Unique_Clonal_AA.nwk
BS1000_Unique_Pyro_AA
BS1000_Unique_Pyro_AA.nwk
pic

Yes



« Projects › 16-08-08_RespPCT › analysis › scripts

Projects
House
SDAU
Google Drive File Stream (
BP_half_day_ppt_lectures
docs
My Drive
Screenshots

Name

01_clean_data
02_plots
03_tables
04_stats_analysis
05_post_hoc_stats
functions
randomization
tables

Outline a file naming convention

Machine readable:

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

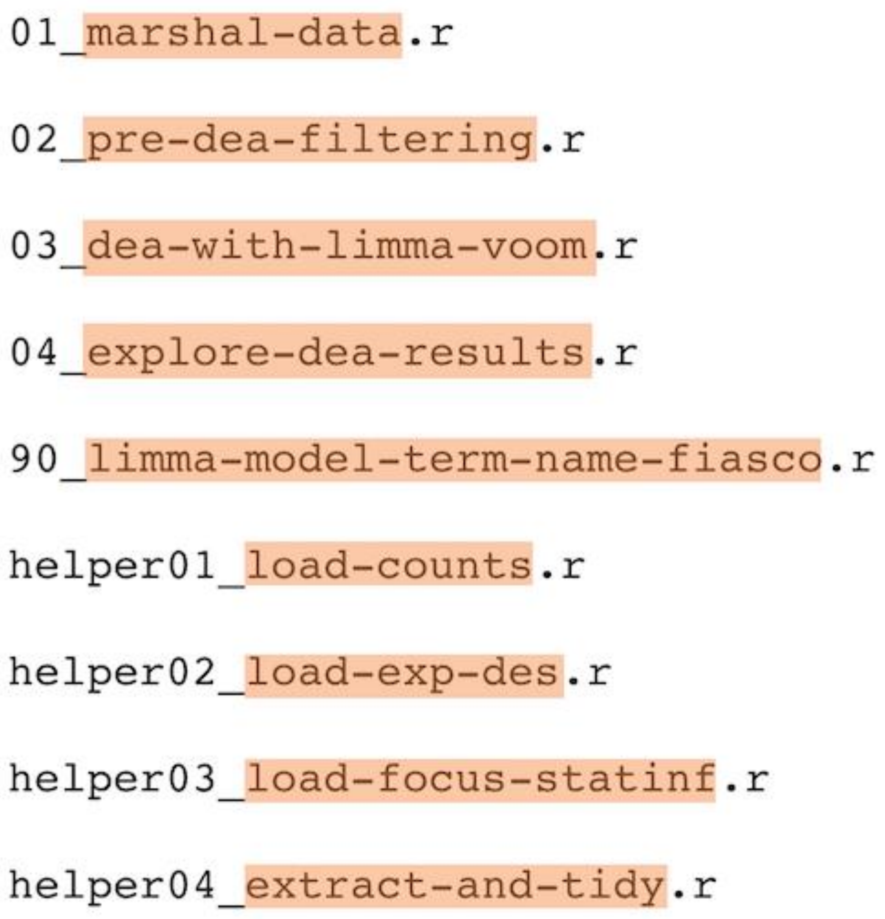
Human readable:

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

Metadata:

Separate with underscores ("_")

- Avoid punctuation
- Remove case-sensitivity

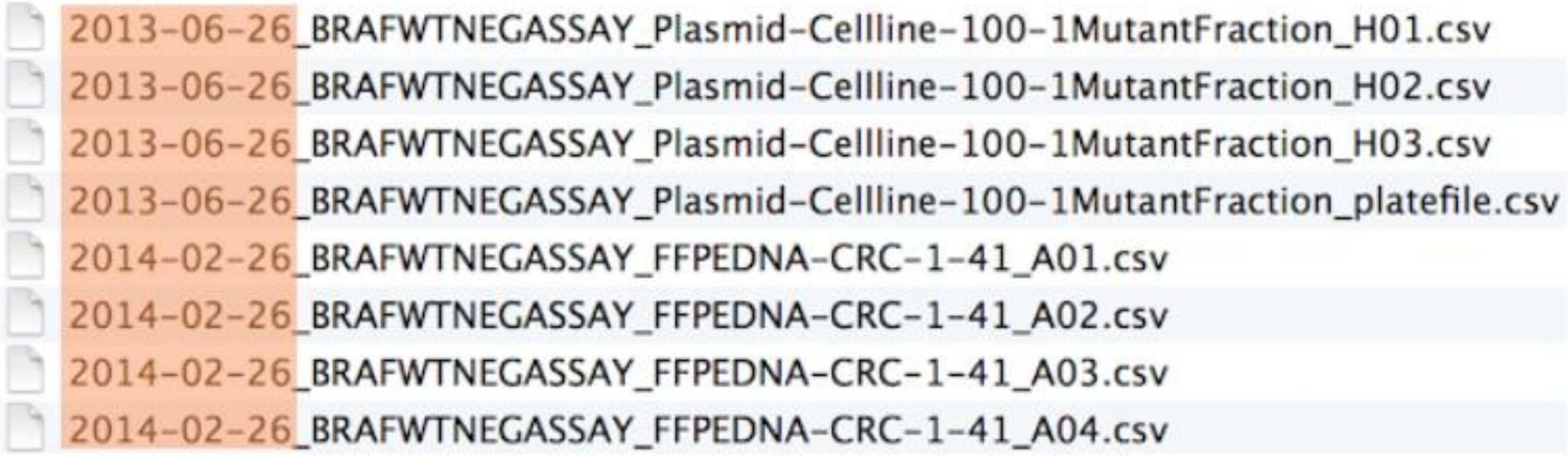


```
01_marshal-data.r
02_pre-dea-filtering.r
03_dea-with-limma-voom.r
04_explore-dea-results.r
90_limma-model-term-name-fiasco.r
helper01_load-counts.r
helper02_load-exp-des.r
helper03_load-focus-statinf.r
helper04_extract-and-tidy.r
```

[Data carpentry - File naming](#)

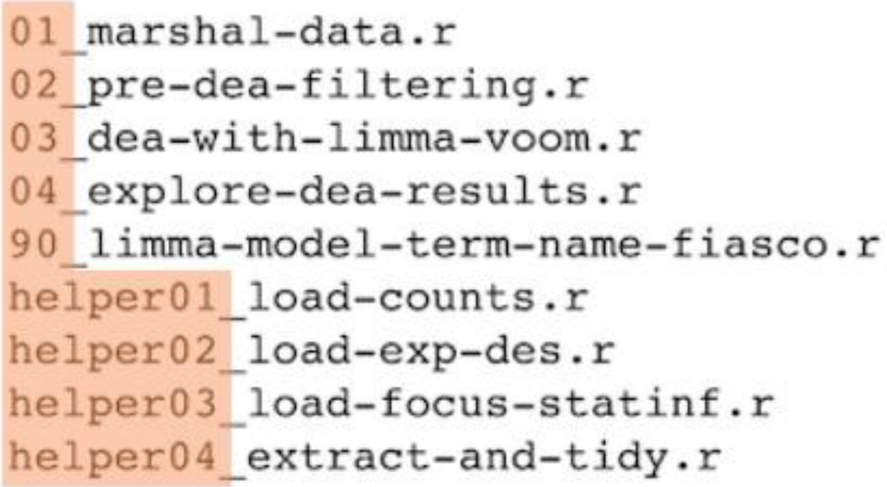
Outline a file naming convention

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
```

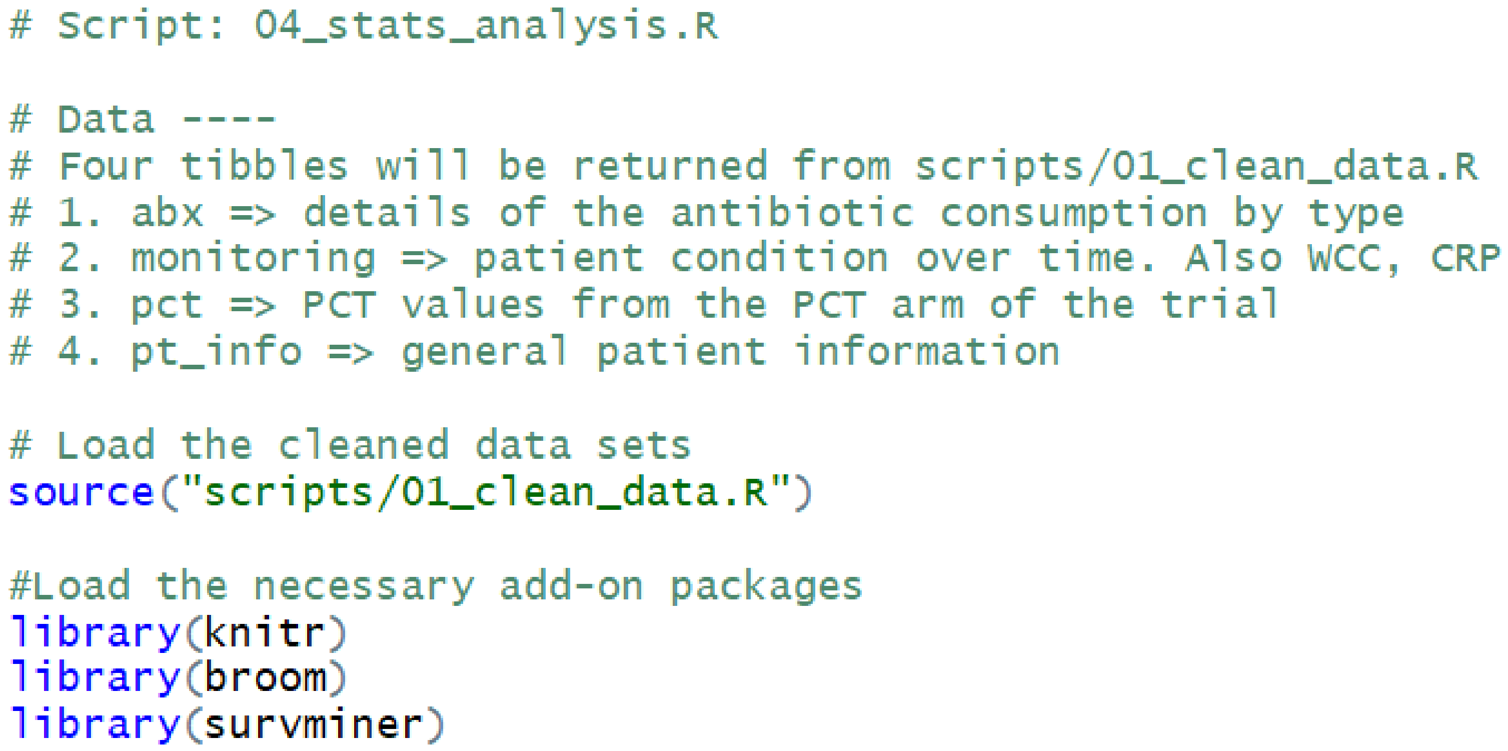
Logical order:



```
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
```

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
# Script: 04_stats_analysis.R

# Data ----
# Four tibbles will be returned from scripts/01_clean_data.R
# 1. abx => details of the antibiotic consumption by type
# 2. monitoring => patient condition over time. Also WCC, CRP
# 3. pct => PCT values from the PCT arm of the trial
# 4. pt_info => general patient information

# Load the cleaned data sets
source("scripts/01_clean_data.R")

#Load the necessary add-on packages
library(knitr)
library(broom)
library(survminer)
```

R Markdown

- R Markdown combines the code you wrote, the output produced and you own comments

- You can view it as a digital lab notebook, where you are both recording what you’re doing, and what you were thinking while you were thinking it!

- R Markdown outputs can take many forms
    - Word documents, PDFs, slideshows etc.

R Markdown

YAML header

```
---
title: "Diamond sizes"
date: 2016-08-25
output: html_document
---
```

Chunks of code

~~~
```{r setup, include = FALSE}
library(ggplot2)
library(dplyr)
smaller <- diamonds %>%
filter(carat <= 2.5)
```
~~~

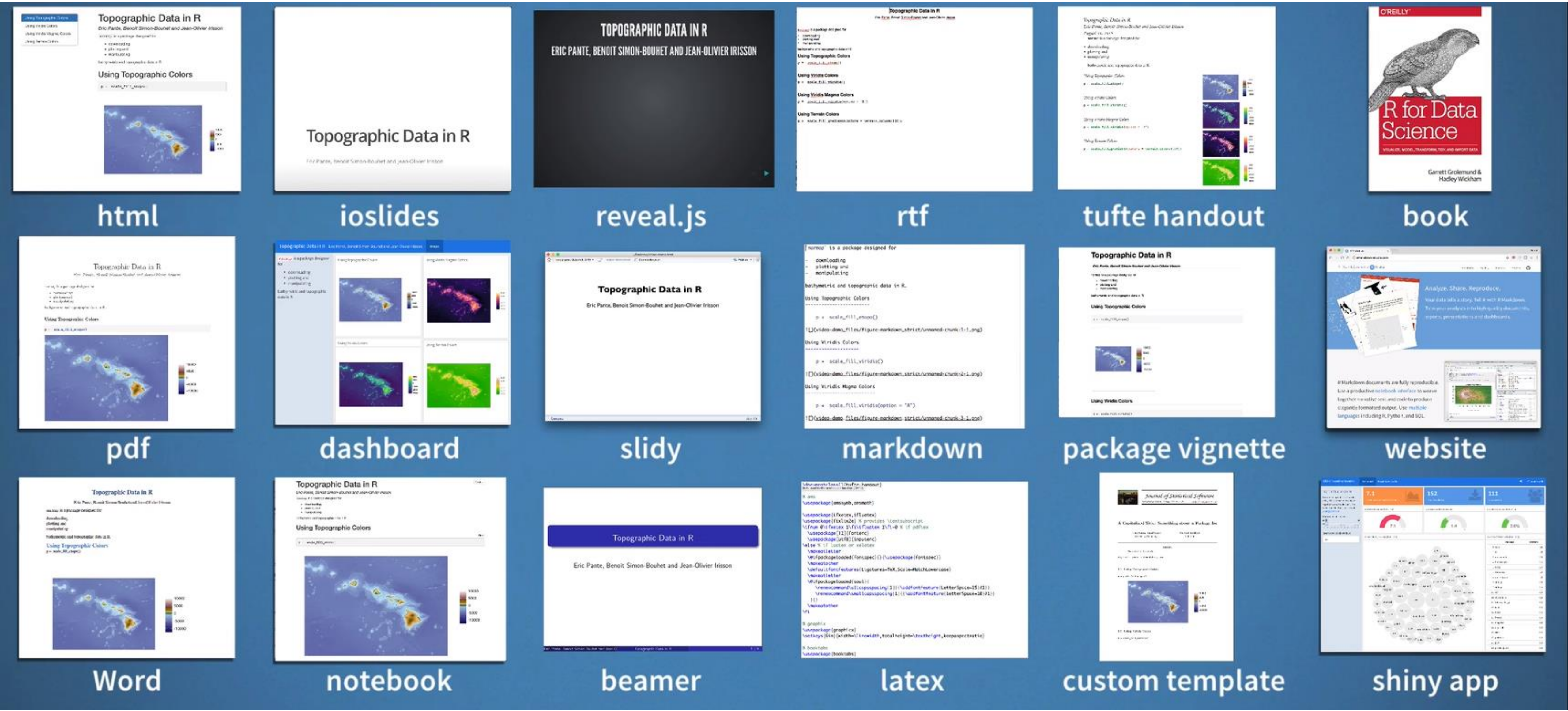
Plain text with integrated outputs from R

```
We have data about `r nrow(diamonds)`
diamonds. Only
`r nrow(diamonds) - nrow(smaller)` are
larger than
2.5 carats. The distribution of the
remainder is shown below:
```

Chunks of code

~~~
```{r, echo = FALSE}
smaller %>%
ggplot(aes(carat)) +
geom_freqpoly(binwidth = 0.01)
```
~~~

What has R Markdown ever done for us?


html
ioslides
reveal.js
rtf
tufte handout
book
pdf
dashboard
slidy
markdown
package vignette
website
Word
notebook
beamer
latex
custom template
shiny app
Topographic Data in R
Eric Pante, Benoit Simon-Bouhet and Jean-Olivier Irisson
R for Data Science
O'REILLY
Garrett Grolemund & Hadley Wickham