

R: The tidyverse and beyond

1-day workshop
Morning lecture



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

Plan for the day

Morning: Reproducible research through R/RStudio

10.45 am: Lecture - Project orientated workflows

11.15 am: Coffee break - Discussion

11.30 am: 2 × 30 minute tutorials

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

12.30 pm: Lecture - Introduction to the tidyverse

1.15 pm: Lunch break

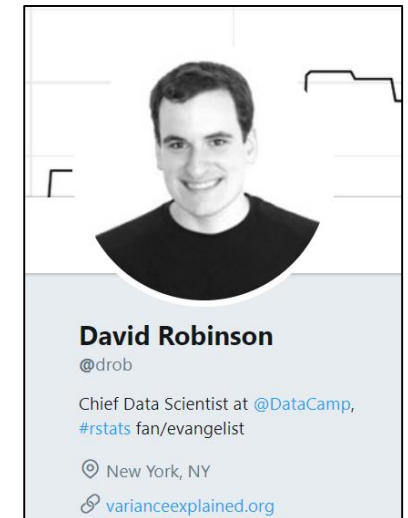
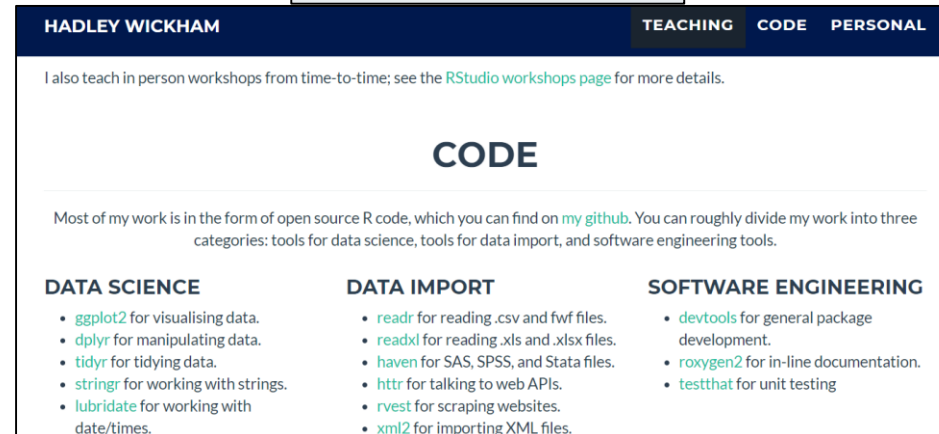
Afternoon: A crash course in the tidyverse

1.45 pm: 3 × 45 minute hands-on tidyverse tutorials including example scripts and problem sheets exploring

- readr, tidyr
- dplyr
- ggplot2

4 pm: Closing remarks, questions

Disclaimer



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

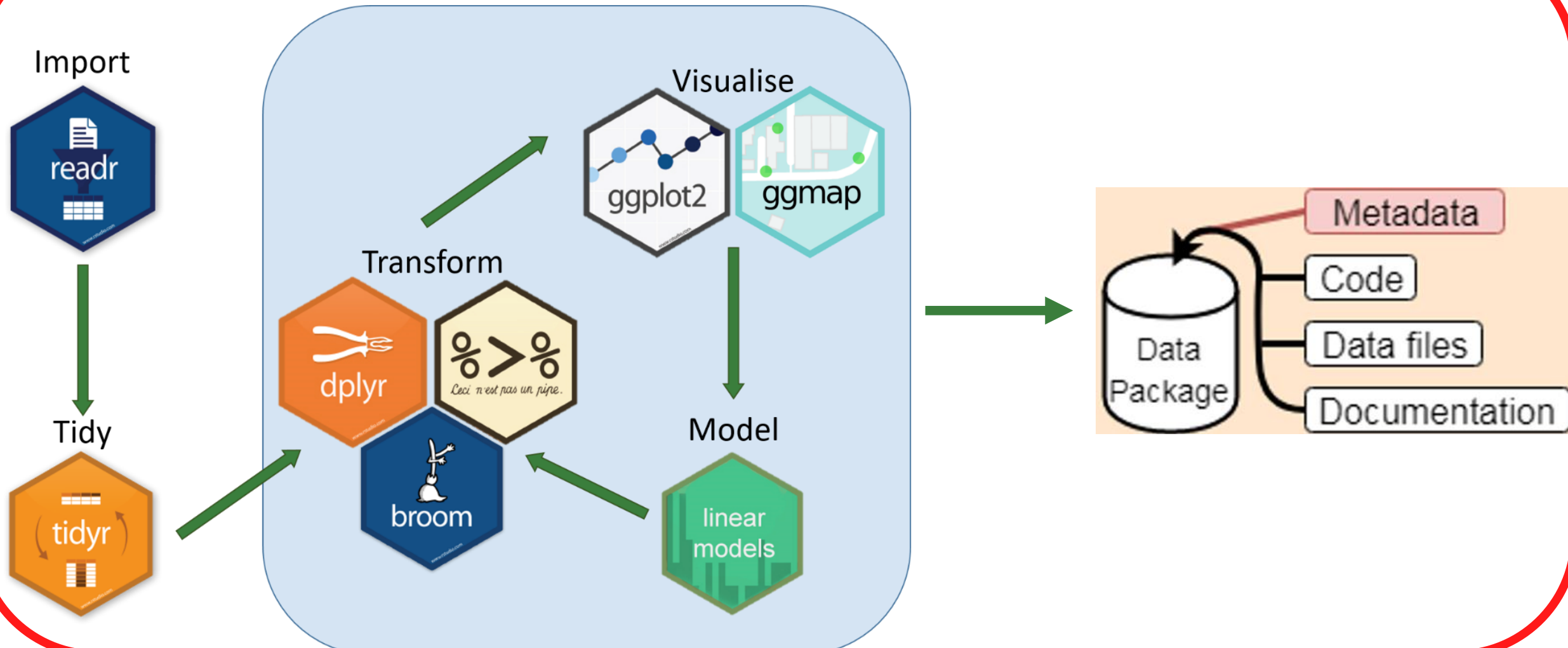
Exploring college major and income: a live data analysis in R October 16, 2018
A live screencast of an exploratory data analysis from the Tidy Tuesday series. This one explores college major and income data from 538.

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.

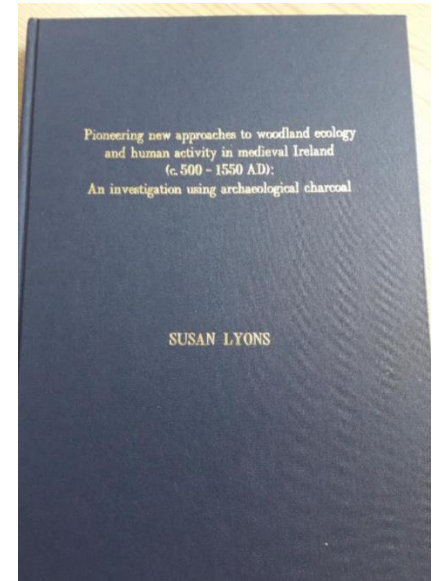
Putting the pieces together

- Data analysis in a tidyverse nutshell



How is research presented?

Theses



Books



Papers

Journal of Virology

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

Brendan A. Palmer,* Daniel Schmidt-Martin,* Zoya Dimitrova,* Pavel Skums,* Orla Crosbie,* Elizabeth Kenny-Walsh,* Liam J. Fanning*
Molecular Biology Diagnostic & Research Laboratory, Department of Medicine, University College Cork, Ireland†; Division of Viral Hepatitis, Centers for Disease Control and Prevention, Atlanta, Georgia, USA‡; Department of Hepatology, 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 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-naïve, 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.

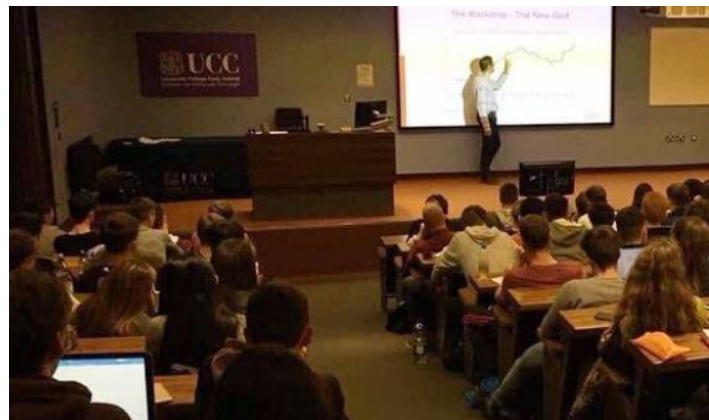
Hepatitis C virus (HCV) infection is a global health issue and is recognized as a major etiological agent of liver-related diseases (1). It has been estimated that the current prevalence of HCV represents approximately 2% of the global adult (15 years of age and older) population (2). Following transmission, HCV infection may remain asymptomatic for decades, resulting in the majority of infections initially passing undetected (3). It is estimated that up to 1 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 (SRB1) 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 21 November 2015; Accepted 22 December 2015
Accepted manuscript posted online 10 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. 90:3218–3228. doi:10.1128/JVI.02090-15
Editor: M. S. Diamond
Address correspondence to Liam J. Fanning, lfanning@ucc.ie.
BA-P and ZS-M contributed equally to this article.
Copyright © 2016, American Society for Microbiology. All Rights Reserved.

Talks



Posters

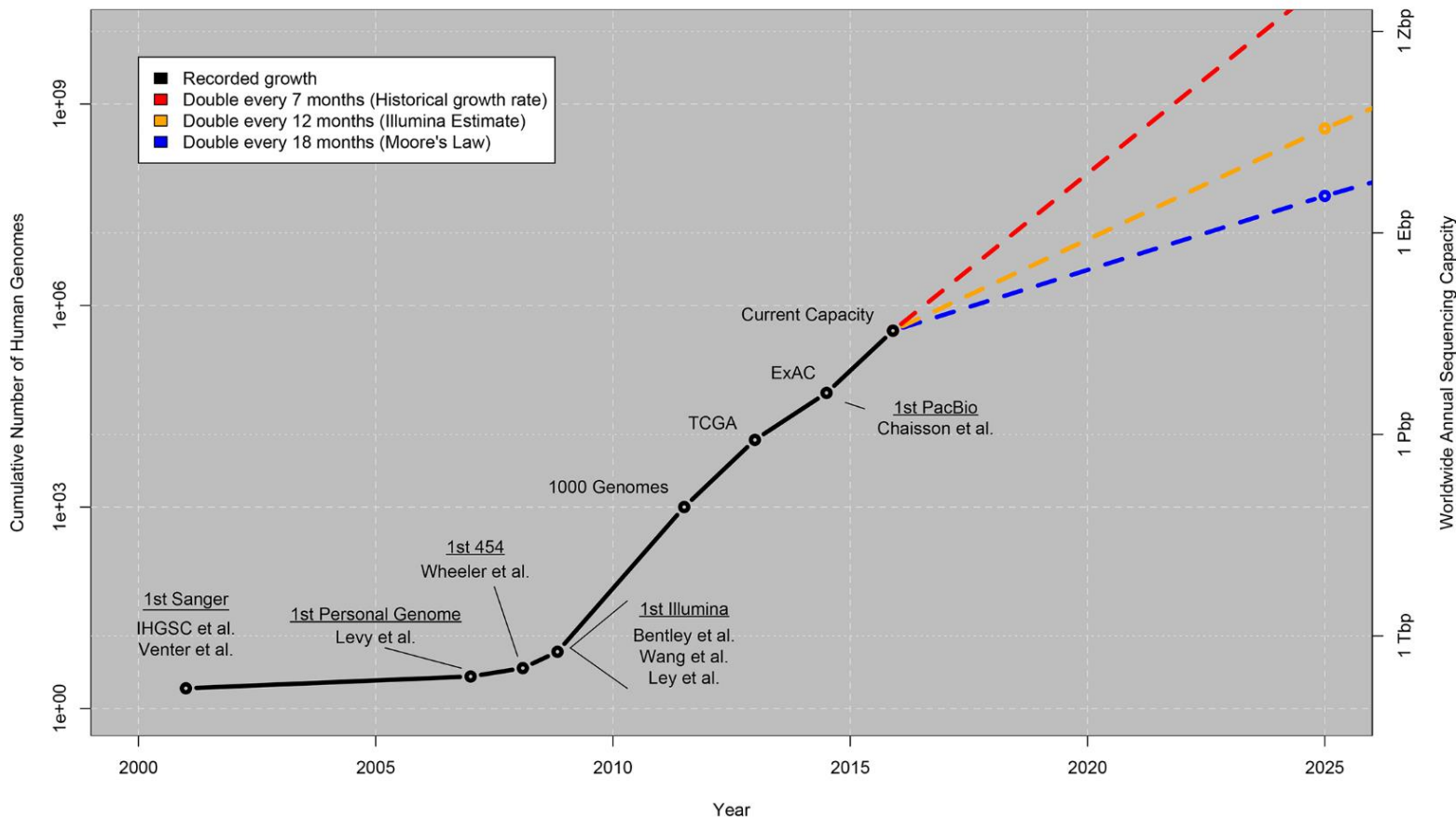


But what does it look like under the bonnet?



The explosion of data in the life sciences

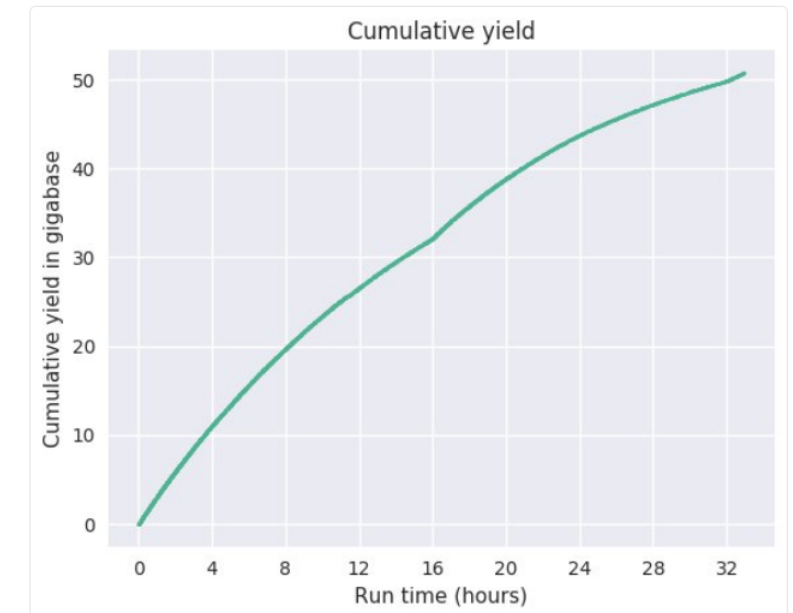
Growth of DNA Sequencing



VIB Tech Watch
@VIBTechWatch

Follow

More than 50 GB output in the first 32 h of [#PromethION](#) run with human DNA sample at GSF, amazing result! Final data output for run coming soon! [@wouter_decoster](#) [@PuniMoj](#) [@svenndhert](#) [@nanopore](#) [@Clive_G_Brown](#) [@VIBLifeSciences](#) [#InnovationLab](#)



Large quantities of data \neq high quality of science

nature
human behaviour

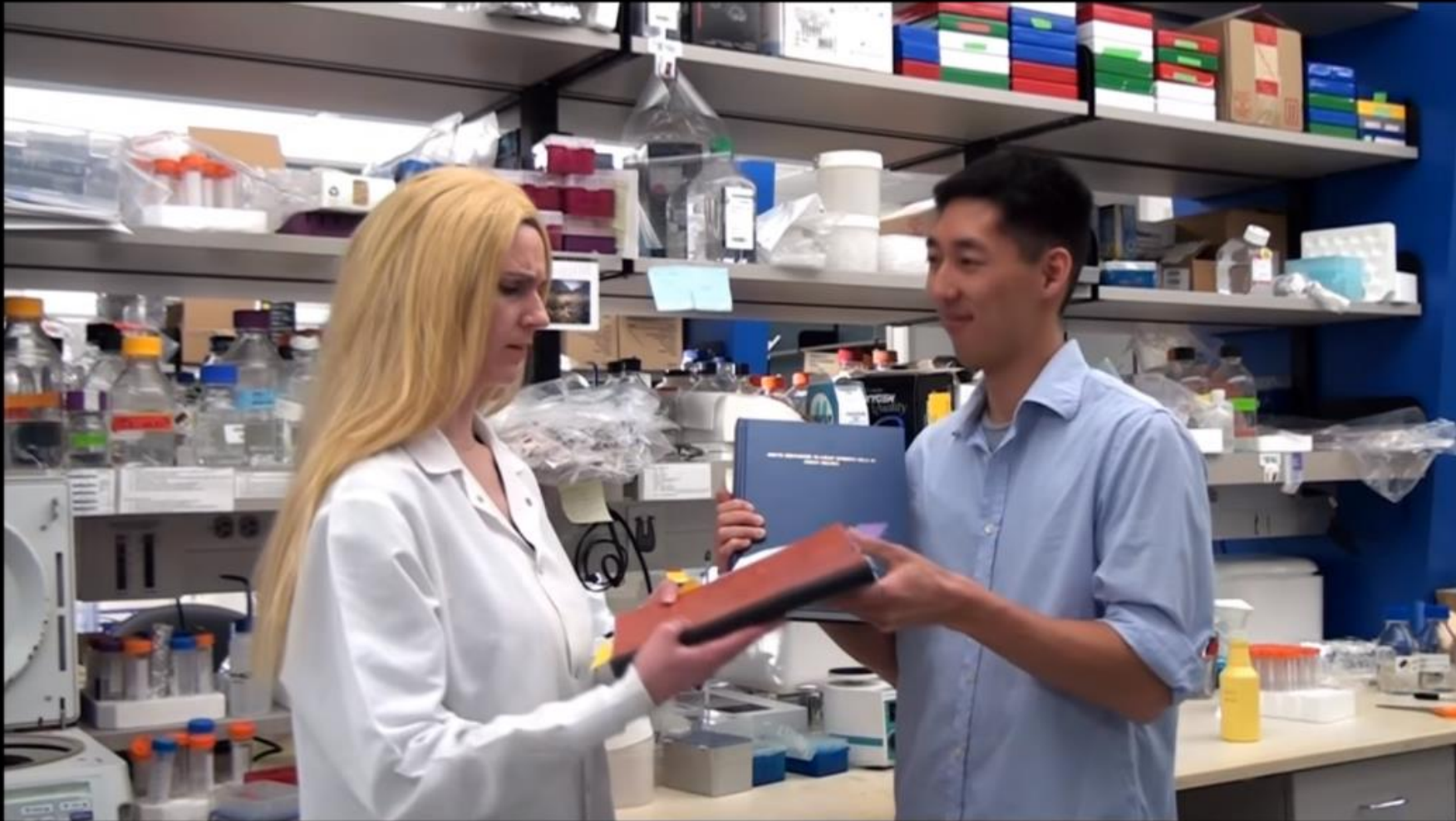
ARTICLES

<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.



You were defending, one foot out the door



I got your project and its problems galore



I hate my life,

This session

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



THIS PERSON IS likely to be YOU BTW!!

Still haven't found what I'm looking for

- Help your future-self

Final Final version

File Home Share View

← → ↕ ↑ 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
Quick access				
Desktop				
Downloads				
Documents				
Pictures				
Projects				
Google Drive				
House				
Google Drive File Stream (G:)				
FAIR_workshop				
Icon Files				
R_Users_Workshop				
subgroup_4-2_drafts				
OneDrive				
This PC				
	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

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



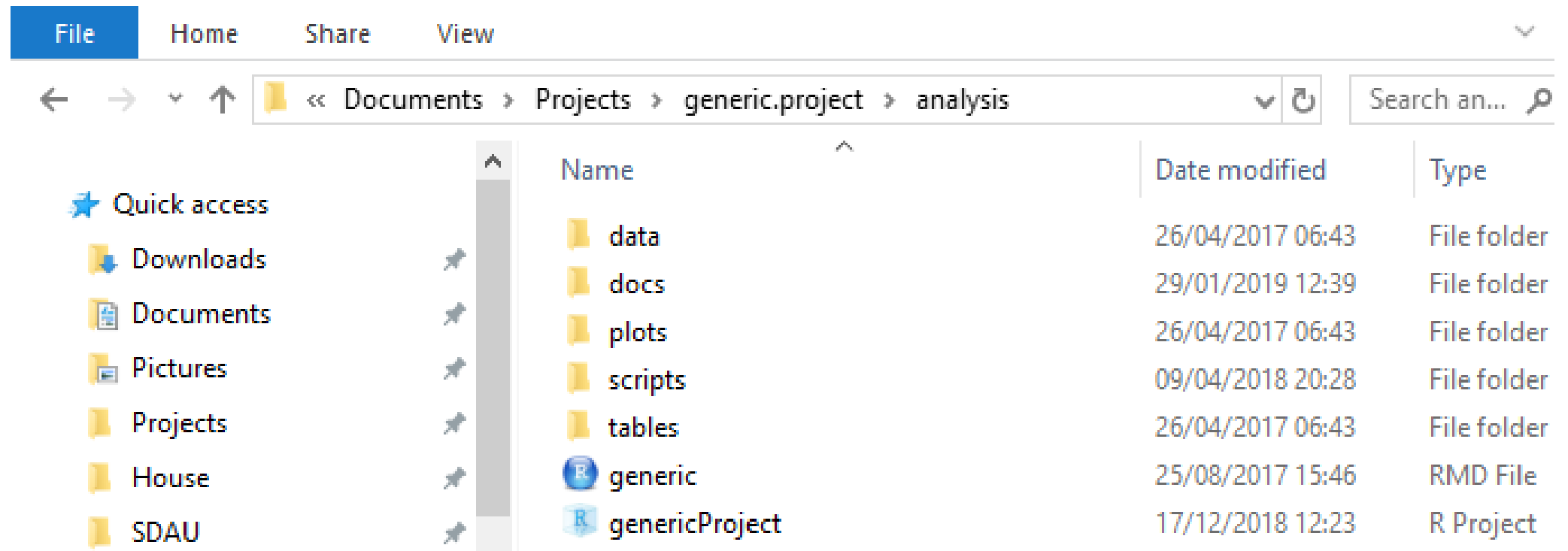
Version Control

Checkout a project from a version control repository



Define a generic project structure

- STEP 1: Give your research projects a shared structure



Give your files informative names

- STEP 1: Give your research projects a shared structure

The screenshot shows a OneDrive file explorer interface. At the top, there are tabs for 'File', 'Home', 'Share', and 'View'. Below these is a breadcrumb navigation bar showing the path: < Projects > 16-08-08_RespPCT > analysis > data. The '16-08-08_RespPCT' folder is highlighted with a red box. On the left, a sidebar shows a list of folders: Projects, House, SDAU, Google Drive File Stream, BP_half_day_ppt_lectures, and docs. The main area displays the contents of the 'data' folder, which includes a 'raw_data' folder and four Excel files. These files are highlighted with a red box: '2018-11-06_abx', '2018-11-06_monitoring', '2018-11-06_pct', and '2018-11-06_pt_info'. The files are listed with their names and the date they were modified.

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

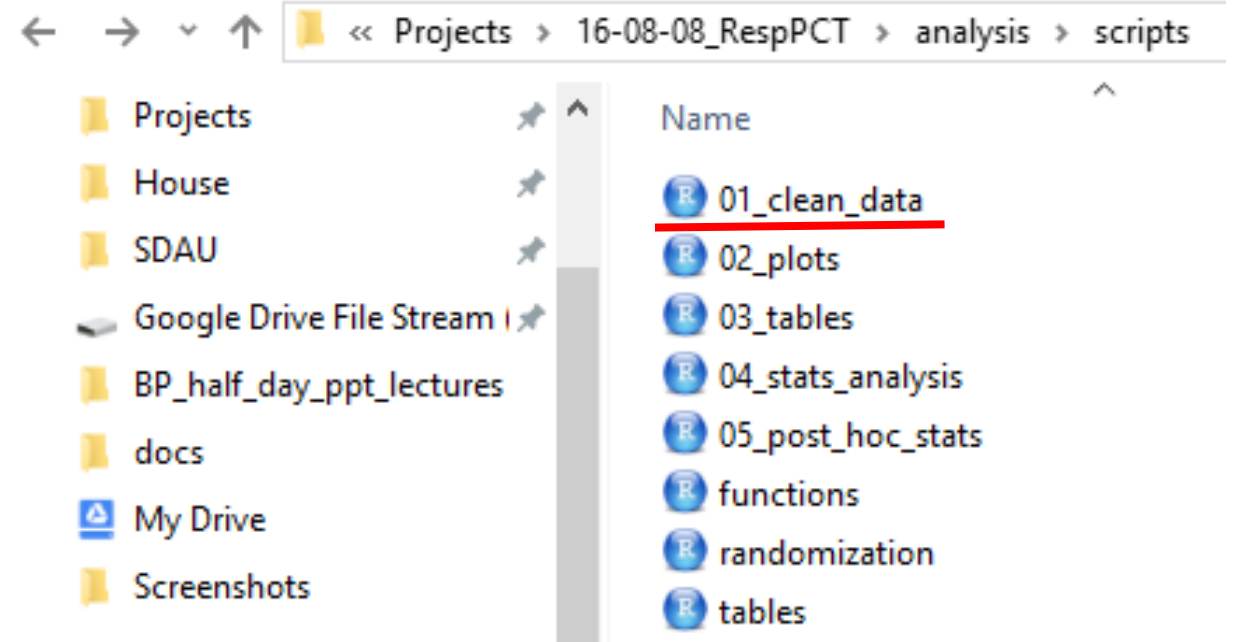
- STEP 2: Make you file names machine readable, human readable and work with default ordering

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



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

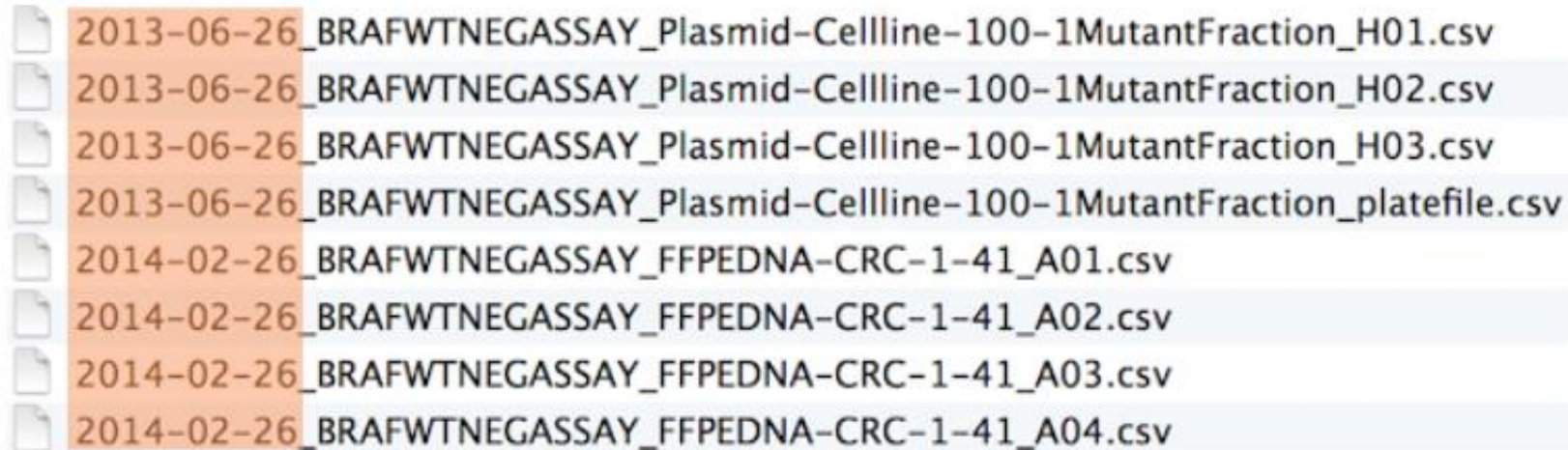
helper02_`load-exp-des`.r

helper03_`load-focus-statinf`.r

helper04_`extract-and-tidy`.r

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

```
# 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)
```
```

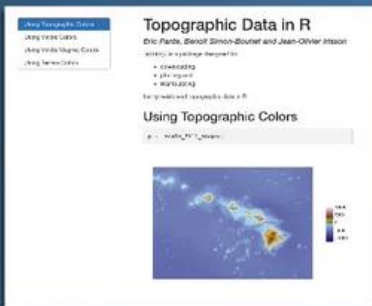
Plain text with integrated
outputs from R

```
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:
```

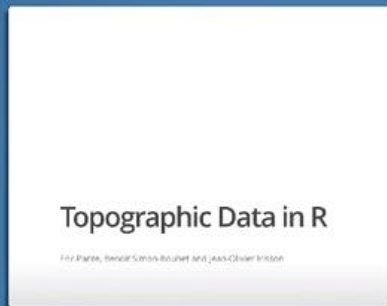
Chunks of code

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

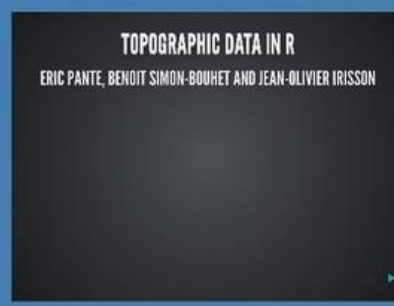
What has R Markdown ever done for us?



html



ioslides



reveal.js



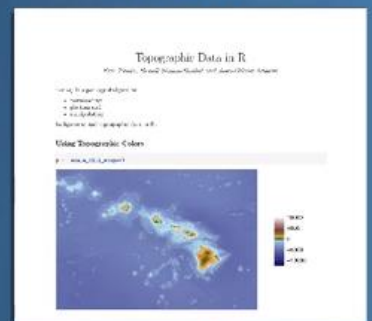
rtf



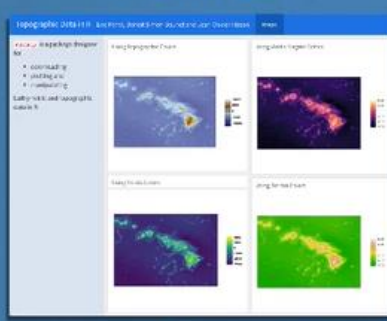
tufte handout



book



pdf



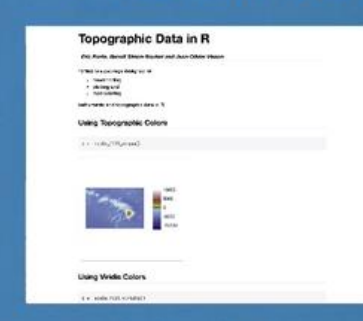
dashboard



slidy



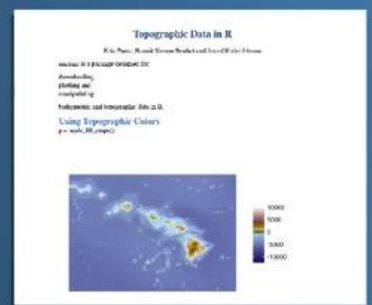
markdown



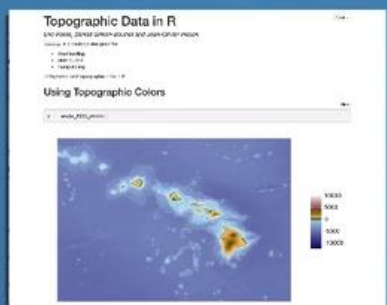
package vignette



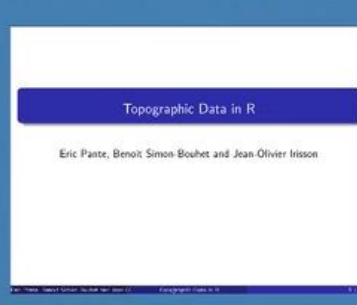
website



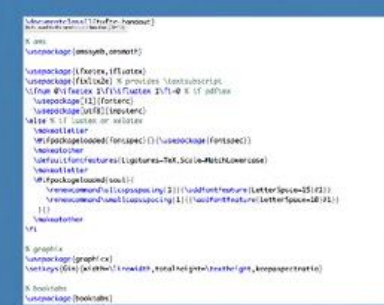
Word



notebook



beamer



latex



custom template



shiny app