A Hitchhikers Guide to Reproducible Research

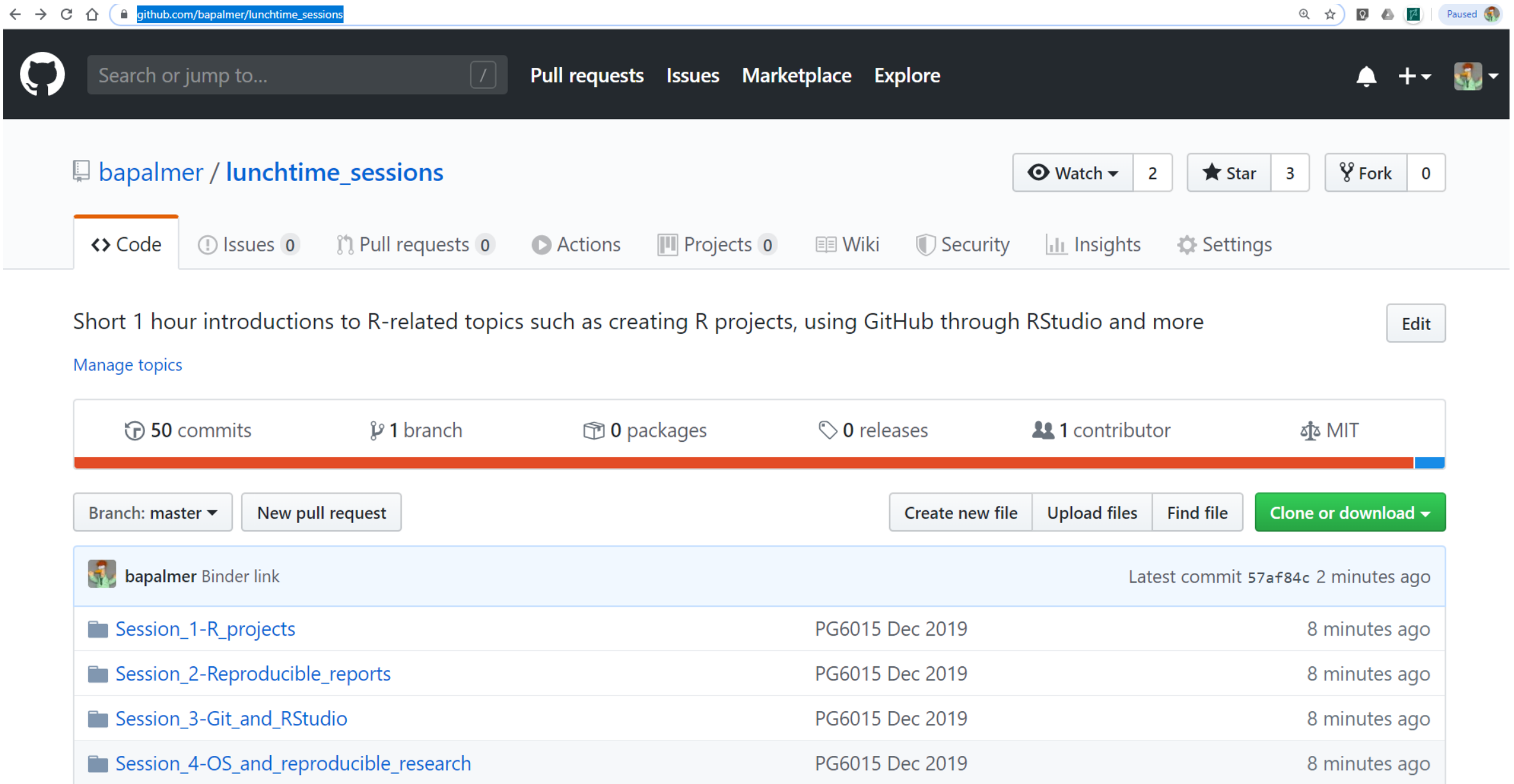
Thursday 20th February 2020

Brendan Palmer,
Clinical Research Facility - Cork &
School of Public Health
@B_A_Palmer

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh
School of
Public Health

CRF-C
HRB Clinical Research Facility Cork

Talk materials



Can we believe what we see in the literature?

AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

PERSPECTIVE

The Prevalence of Inappropriate Image Duplication in Biomedical Research Publications

Elisabeth M. Bik,[a] Arturo Casadevall,[b,c] Ferric C. Fang[d]

Department of Medicine, Division of Infectious Diseases, Stanford School of Medicine, Stanford, California, USA[a]; Department of Molecular Microbiology and Immunology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA[b]; Department of Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland, USA[c]; Departments of Laboratory Medicine and Microbiology, University of Washington School of Medicine, Seattle, Washington, USA[d]

ABSTRACT **Inaccurate data in scientific papers can result from honest error or intentional falsification. This study attempted to determine the percentage of published papers that contain inappropriate image duplication, a specific type of inaccurate data. The images from a total of 20,621 papers published in 40 scientific journals from 1995 to 2014 were visually screened. Overall, 3.8% of published papers contained problematic figures, with at least half exhibiting features suggestive of deliberate manipulation. The prevalence of papers with problematic images has risen markedly during the past decade. Additional papers written by authors of papers with problematic images had an increased likelihood of containing problematic images as well. As this analysis focused only on one type of data, it is likely that the actual prevalence of inaccurate data in the published literature is higher. The marked variation in the frequency of problematic images among journals suggests that journal practices, such as prepublication image screening, influence the quality of the scientific literature.**

Can we believe what we read in the literature?

OPEN ACCESS

ESSAY

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • https://doi.org/10.1371/journal.pmed.0020124

And publication bias is very real

Rink Hoekstra
@RinkHoekstra

Follow

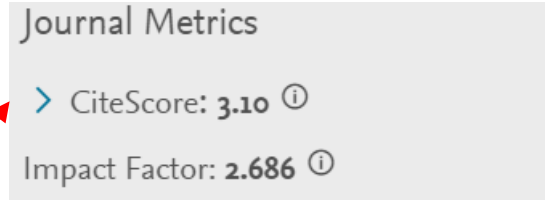
Elsevier editor Spada acknowledging that null results are not even considered for Addictive Behaviors, seemingly not realizing how problematic that is. Offering a lower prestige alternative journal doesn't make that right.

Professor M. M. Spada said:

"Articles that may not traditionally be considered by Addictive Behaviors, including negative/null data papers, studies using smaller samples and cross-sectional designs, replication studies, cross-cultural research, and case reports will be welcome by its sister journal Addictive Behaviors Reports."

Editor-in-Chief
Professor M. M. Spada
London South Bank University

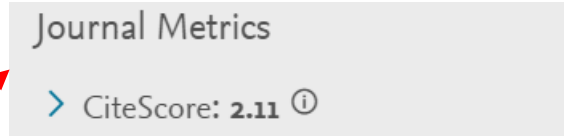
7:55 AM - 19 Mar 2019



Journal Metrics

CiteScore: **3.10**

Impact Factor: **2.686**



Journal Metrics

CiteScore: **2.11**

...and this is where we put the non-significant results.

someecards
user card

p-values do not define a study

nature International weekly journal of science

Search Go

Advanced search

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive > Volume 520 > Issue 7549 > Comment > Article

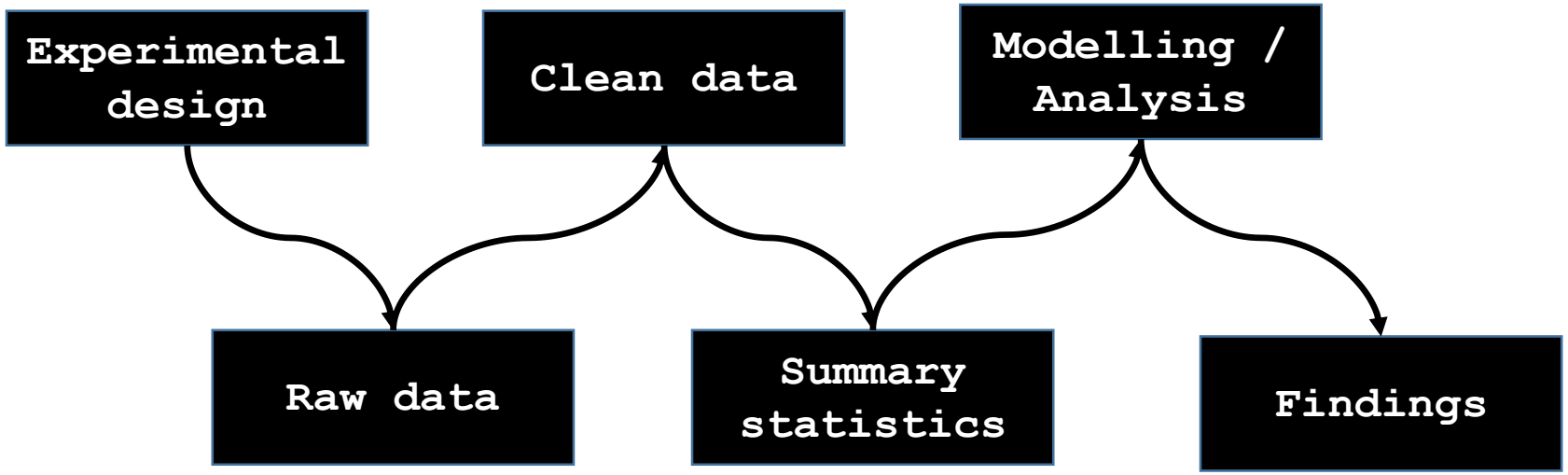
NATURE | COMMENT

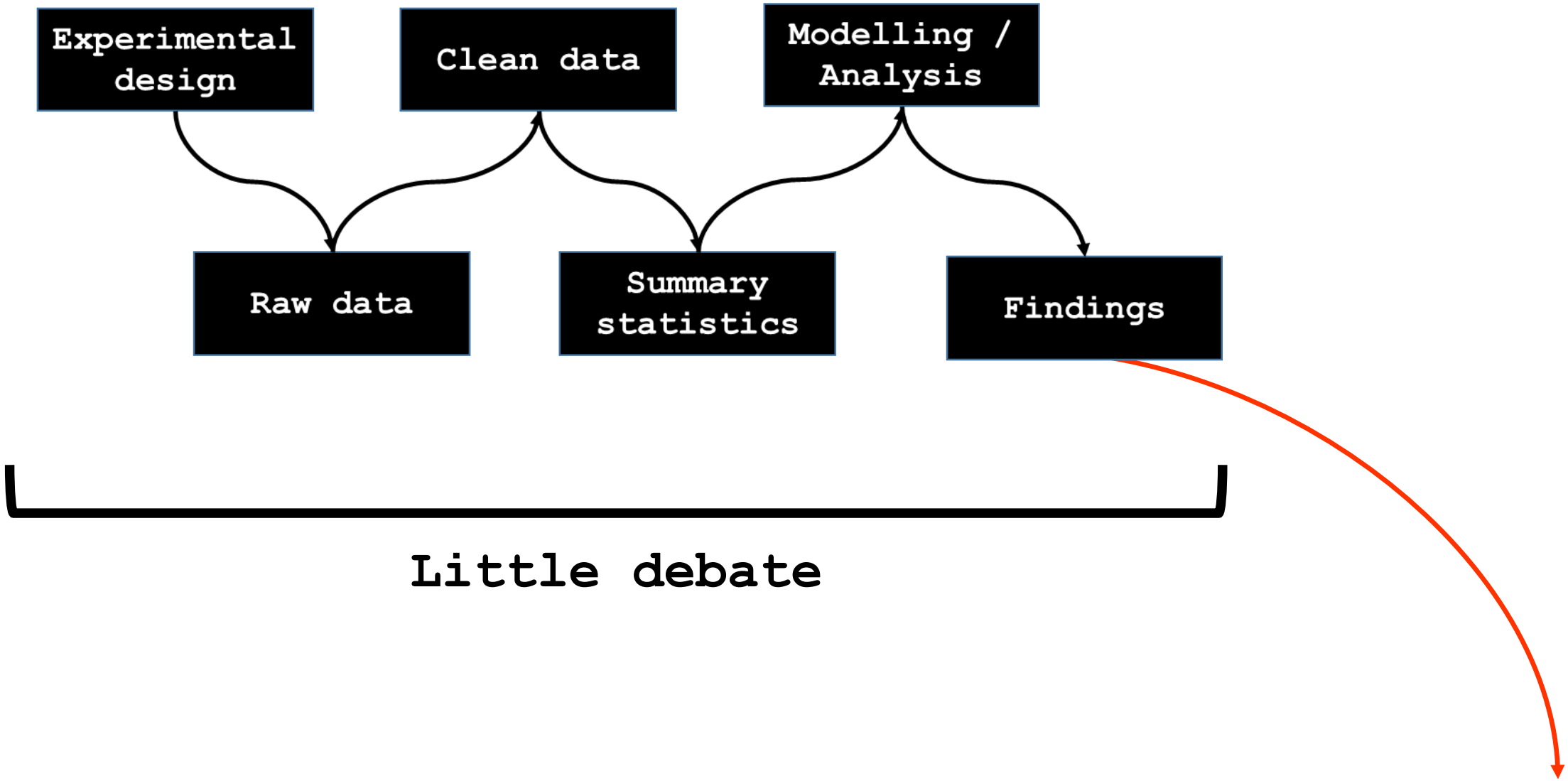
E-alert RSS Facebook Twitter

Statistics: *P* values are just the tip of the iceberg

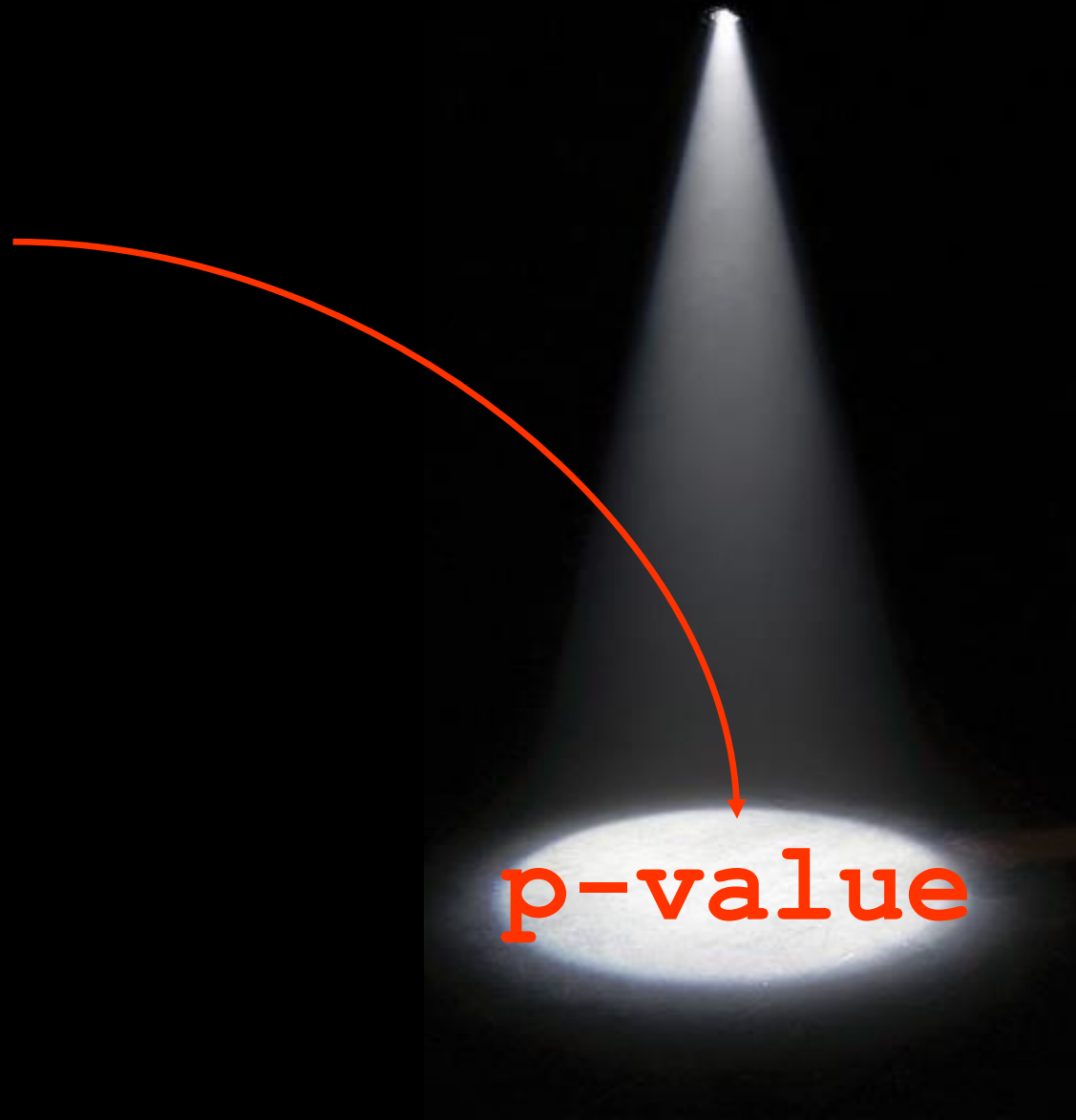
Jeffrey T. Leek & Roger D. Peng

28 April 2015




Experimental design
Clean data
Modelling / Analysis
Raw data
Summary statistics
Findings
Little debate

Extreme scrutiny



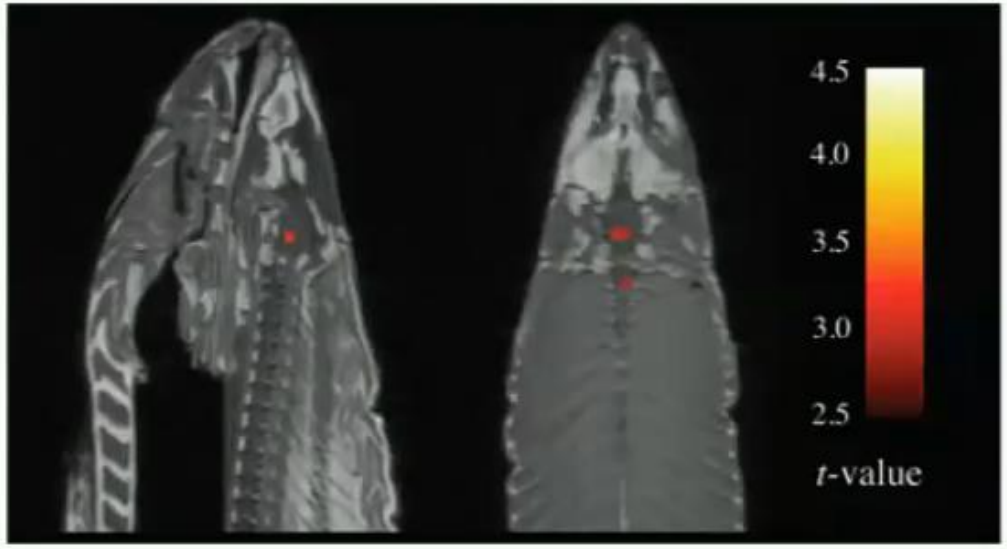
Ig Noble research - 2012 Neuroscience winner

Journal of Serendipitous and Unexpected Results

Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction

Craig M. Bennett [1*], Abigail A. Baird [2], Michael B. Miller [1] and George L. Wolford [3]

[1]Department of Psychology, University of California at Santa Barbara, Santa Barbara, CA 93106
[2]Department of Psychology, Blodgett Hall, Vassar College, Poughkeepsie, NY 12604
[3]Department of Psychological and Brain Sciences, Moore Hall, Dartmouth College, Hanover, NH 03755



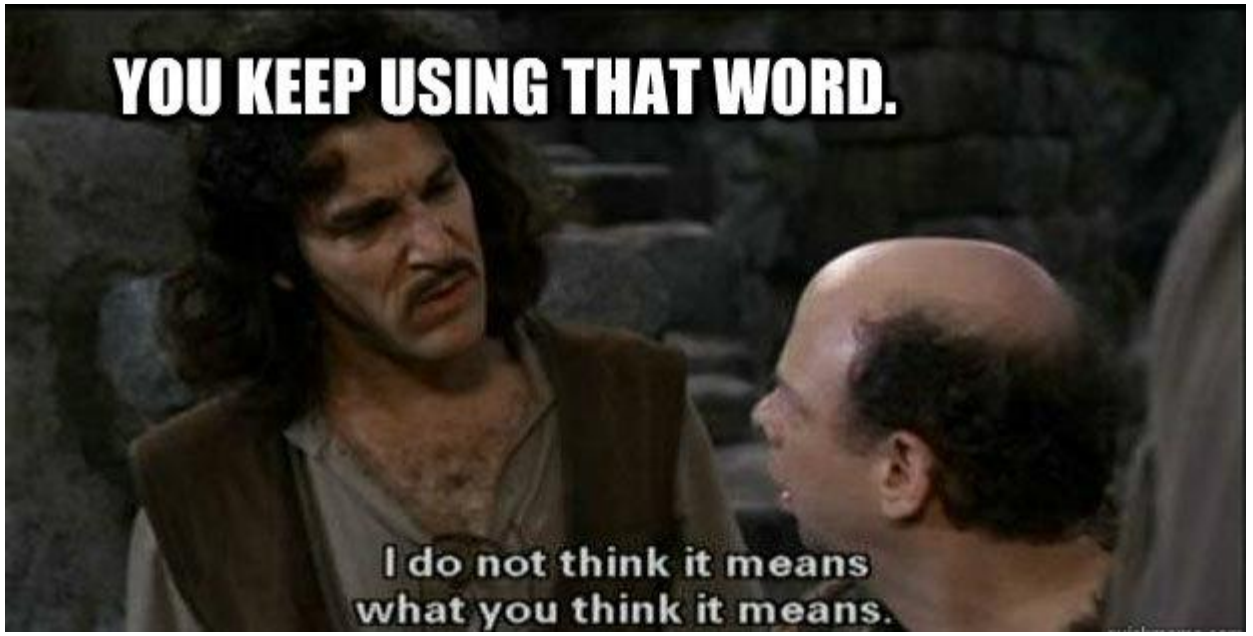
One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon measured approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning. It is not known if the salmon was male or female, but given the post-mortem state of the subject this was not thought to be a critical variable.

The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence, either socially inclusive or socially exclusive. The salmon was asked to determine which emotion the individual in the photo must have been experiencing.

Several active voxels were observed in a cluster located within the salmon's brain cavity (see Fig. 1). The size of this cluster was 81 mm^3 with a cluster-level significance of $p = 0.001$.

Either we have stumbled onto a rather amazing discovery in terms of post-mortem ichthyological cognition, or there is something a bit off with regard to our uncorrected statistical approach.

Significant



Taylor & Francis
Taylor & Francis Group

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: http://amstat.tandfonline.com/loi/utas20

The ASA's Statement on *p*-Values: Context, Process, and Purpose

Ronald L. Wasserstein & Nicole A. Lazar

Eur J Epidemiol (2016) 31:337–350
DOI 10.1007/s10654-016-0149-3

CrossMark

ESSAY

Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations

Sander Greenland[1] · Stephen J. Senn[2] · Kenneth J. Rothman[3] · John B. Carlin[4] · Charles Poole[5] · Steven N. Goodman[6] · Douglas G. Altman[7]

Received: 9 April 2016 / Accepted: 9 April 2016 / Published online: 21 May 2016

Abstract Misinterpretation and abuse of statistical tests, confidence intervals, and statistical power have been decried for decades, yet remain rampant. A key problem is that there are no interpretations of these concepts that are at once simple, intuitive, correct, and foolproof. Instead, correct use and interpretation of these statistics requires an attention to detail which seems to tax the patience of working scientists. This high cognitive demand has led to an epidemic of shortcut definitions and interpretations that are simply wrong, sometimes disastrously so—and yet these misinterpretations dominate much of the scientific literature. In light of this problem, we provide definitions and a discussion of basic statistics that are more general and critical than typically found in traditional introductory expositions. Our goal is to provide a resource for instructors, researchers, and consumers of statistics whose knowledge of statistical theory and technique may be limited but who wish to avoid and spot misinterpretations. We emphasize how violation of often unstated analysis protocols (such as selecting analyses for presentation based on the *P* values they produce) can lead to small *P* values even if the declared test hypothesis is correct, and can lead to large *P* values even if that hypothesis is incorrect. We then provide an explanatory list of 25 misinterpretations of *P* values, confidence intervals, and power. We conclude with guidelines for improving statistical interpretation and reporting.

Editor's note This article has been published online as supplementary material with an article of Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process and purpose. The American Statistician 2016.

Albert Hofman, Editor-in-Chief EJE.

✉ Sander Greenland
lesdomes@ucla.edu

Stephen J. Senn
stephen.senn@lih.lu

John B. Carlin
john.carlin@mcri.edu.au

Charles Poole
cpoole@unc.edu

Steven N. Goodman
steve.goodman@stanford.edu

Douglas G. Altman
doug.altman@csm.ox.ac.uk

[1] Department of Epidemiology and Department of Statistics, University of California, Los Angeles, CA, USA

[2] Competence Center for Methodology and Statistics, Luxembourg Institute of Health, Strassen, Luxembourg

[3] RTI Health Solutions, Research Triangle Institute, Research Triangle Park, NC, USA

[4] Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, School of Population Health, University of Melbourne, Melbourne, VIC, Australia

[5] Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC, USA

[6] Meta-Research Innovation Center, Departments of Medicine and of Health Research and Policy, Stanford University School of Medicine, Stanford, CA, USA

[7] Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK

What is your research hypothesis?

Article

Social Psychological and
Personality Science
2017, Vol. 8(8) 875-881

DOI: 10.1177/1948550617693058
journals.sagepub.com/home/spp

Too True to be Bad: When Sets of Studies With Significant and Nonsignificant Findings Are Probably True

Daniël Lakens[1] and Alexander J. Etz[2]

Consider the following example:

- Starting out, H_0 and H_1 are equally likely
  - α is controlled for 0.05
  - The study has 80% power
  - What is the most likely outcome?

  A – True positive, B – True negative,
  C – False Positive, D – False negative

Result

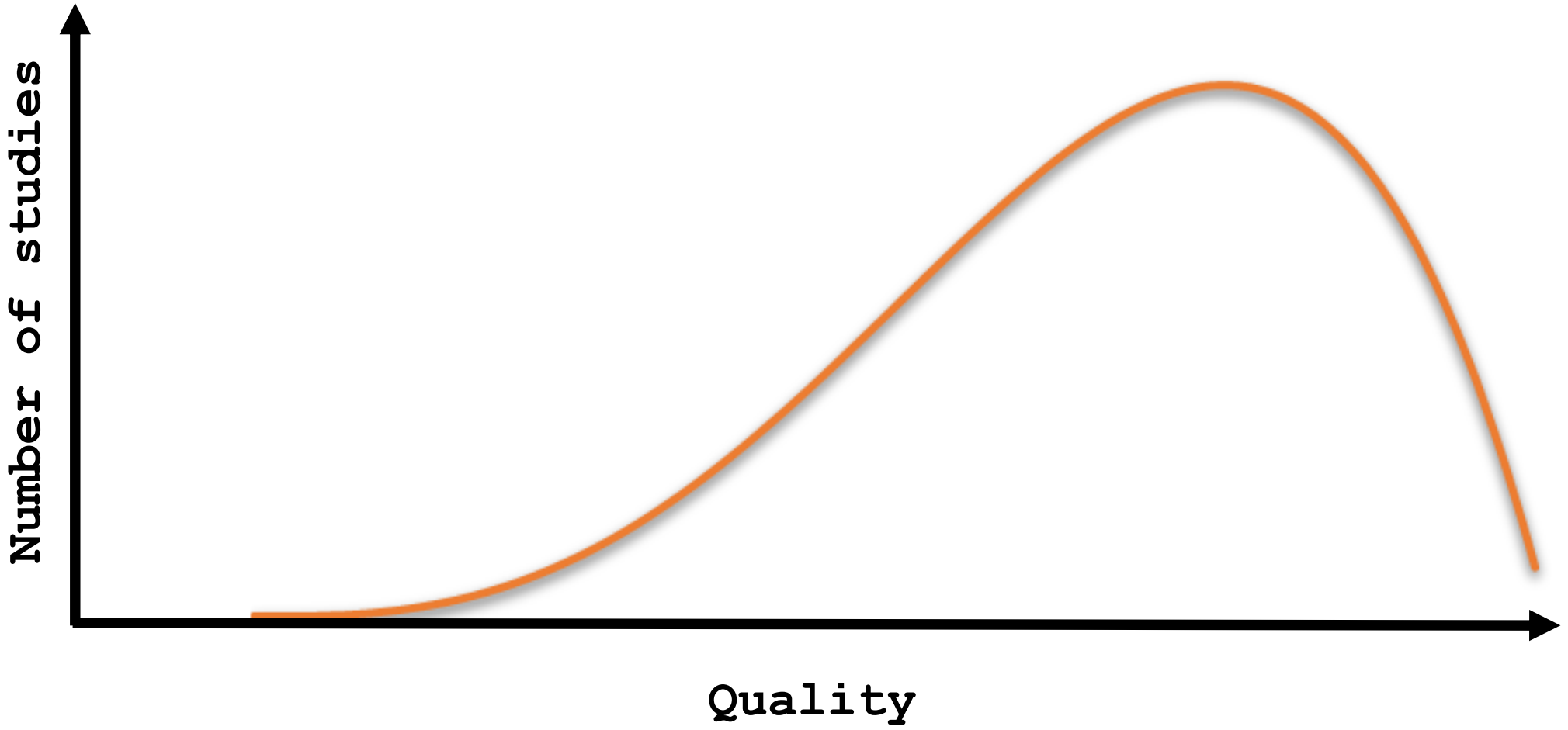
- What is the most likely outcome?

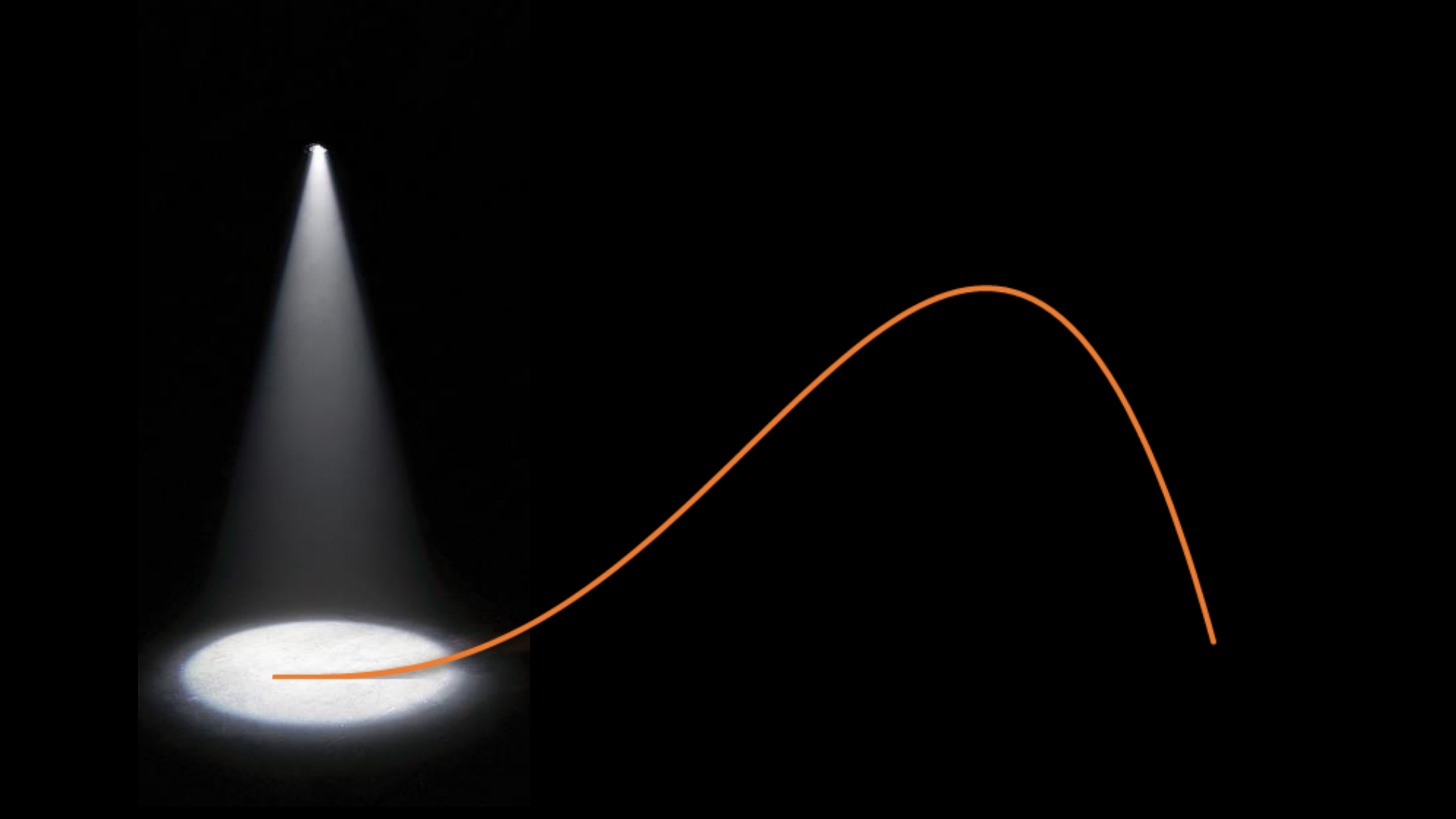
A – True positive – 40%
B – True negative – 47.5%
C – False Positive – 2.5%
D – False negative – 10%

- What is the best strategy to improve this outcome?
- Pick a better hypothesis
- If our alternate hypothesis is more likely…
- H_0 = 40% and H_1 60%

A – True positive – 48%
B – True negative – 38%
C – False Positive – 2%
D – False negative – 12%

Today



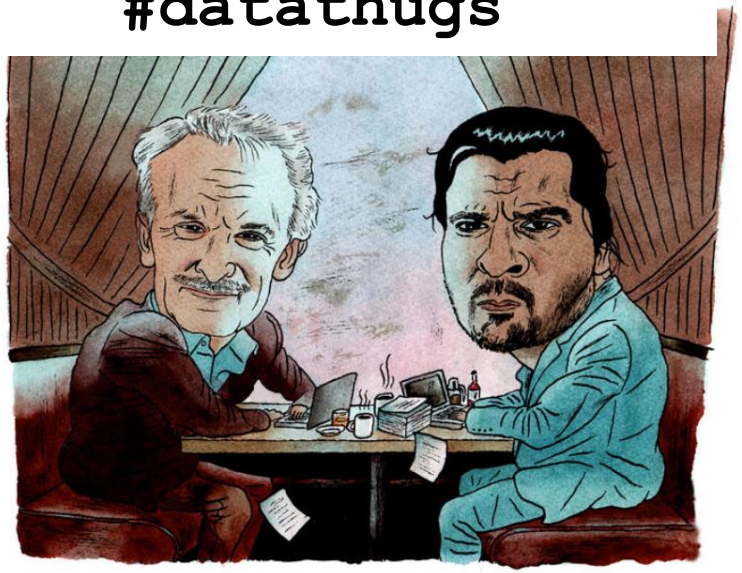


Don’t do what Donny Dont does!

“In short, peer review misses all the hard stuff, and a worrying amount of the easy stuff”

James Heathers,
Northwestern University

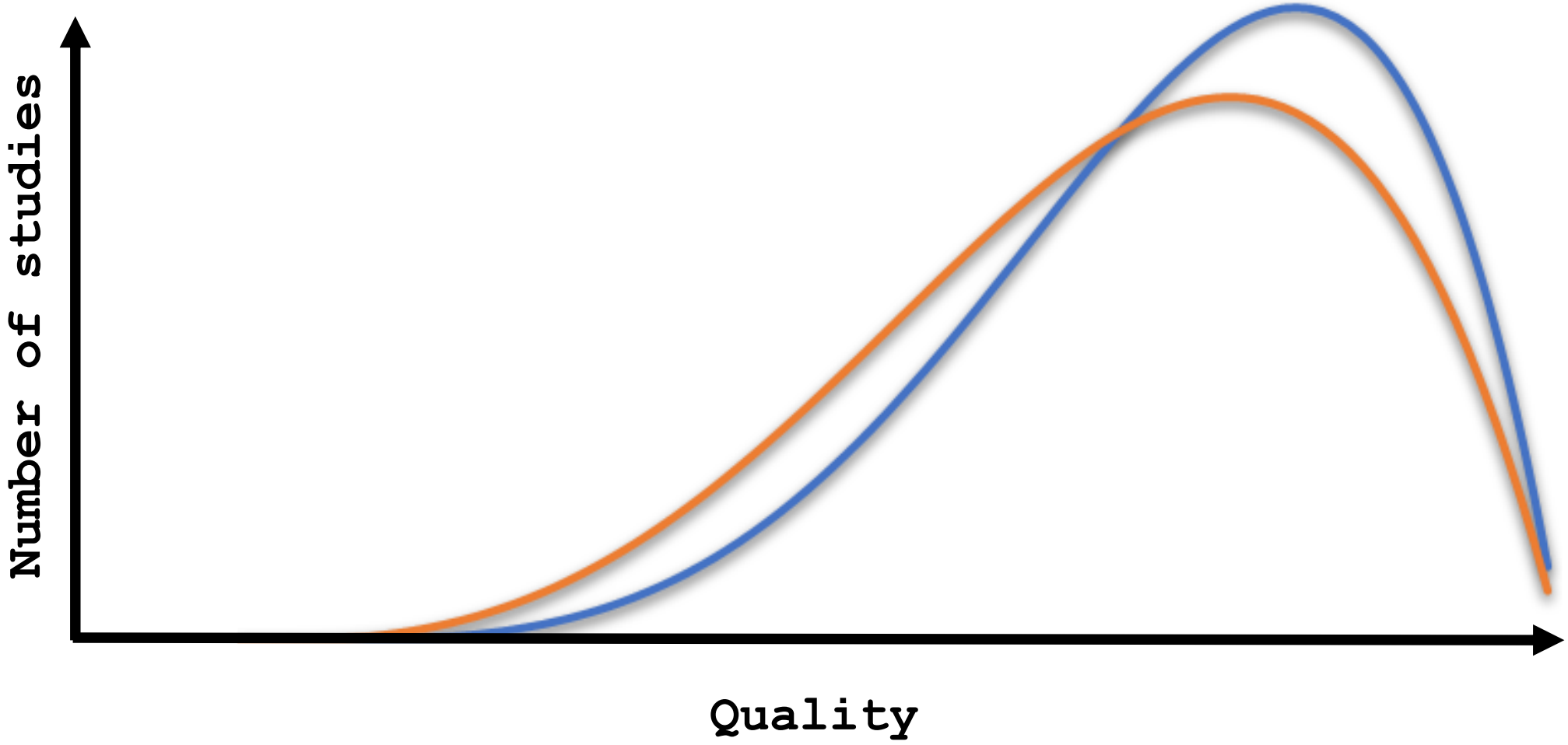
#datathugs



Brian Wansink: The grad student who never said no

“Every day we would scratch our heads, ask "Why," and come up with another way to reanalyze the data with yet another set of plausible hypotheses. Eventually we started discovering solutions”

Can we shift the distribution?



How is research presented?

Papers

Journal of Virology

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

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

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

ABSTRACT

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

IMPORTANCE

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

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

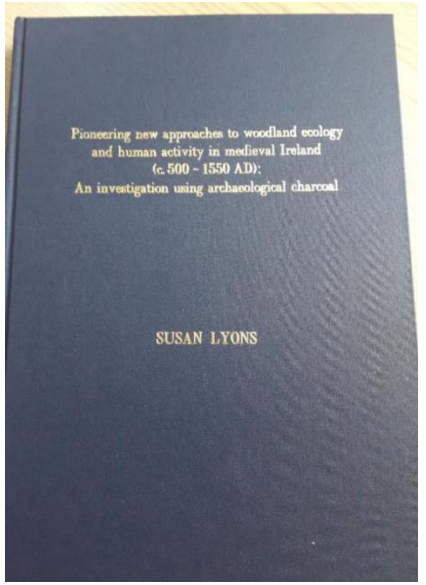
HCV is a single-stranded positive-sense RNA virus of considerable genomic heterogeneity. A recent reclassification defined the HCV global distribution into 7 genotypes and 67 subtypes, with genotypes 1 and 3 accounting for the majority of infections worldwide (6, 7). An error-prone RNA-dependent RNA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

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

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

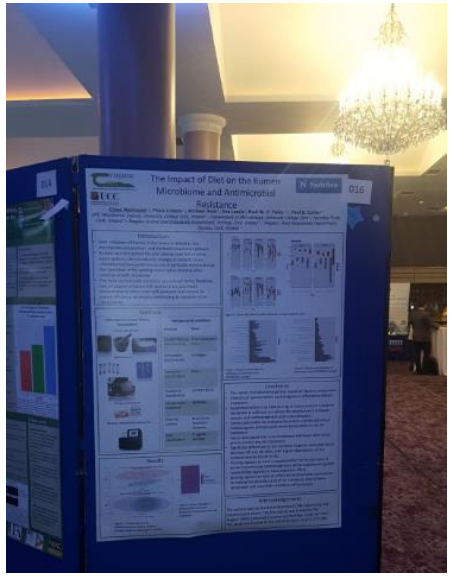
Books

Theses



Talks

Posters



But what does it really look like?

Putting the pieces together

A: Define a project structure

B: Set a naming convention

C: Use scripted workflows

D: Digital notebooks

Reproducible research

E: Version control

F: Data packaging

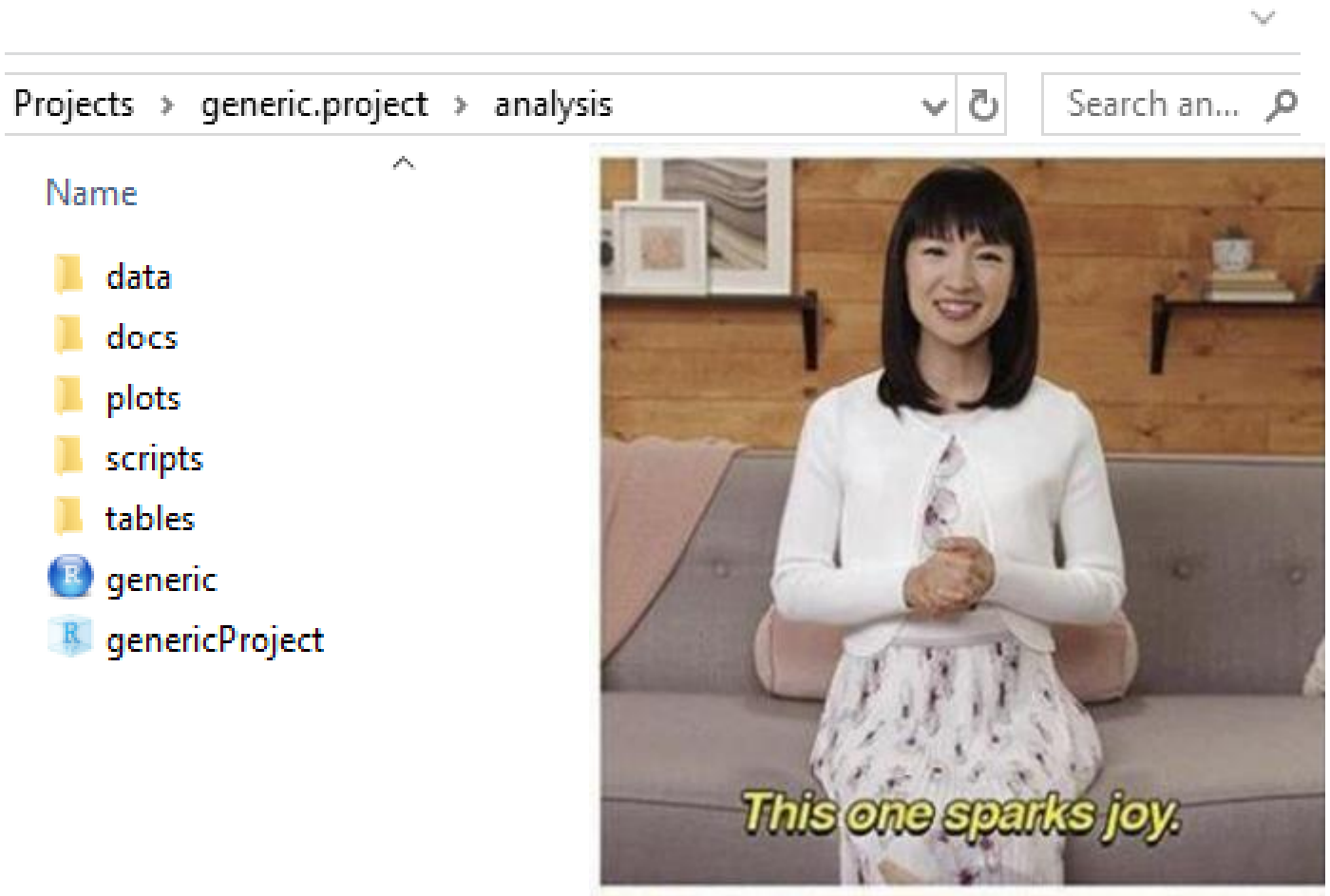
Still haven't found what I'm looking for

- Help your future-self

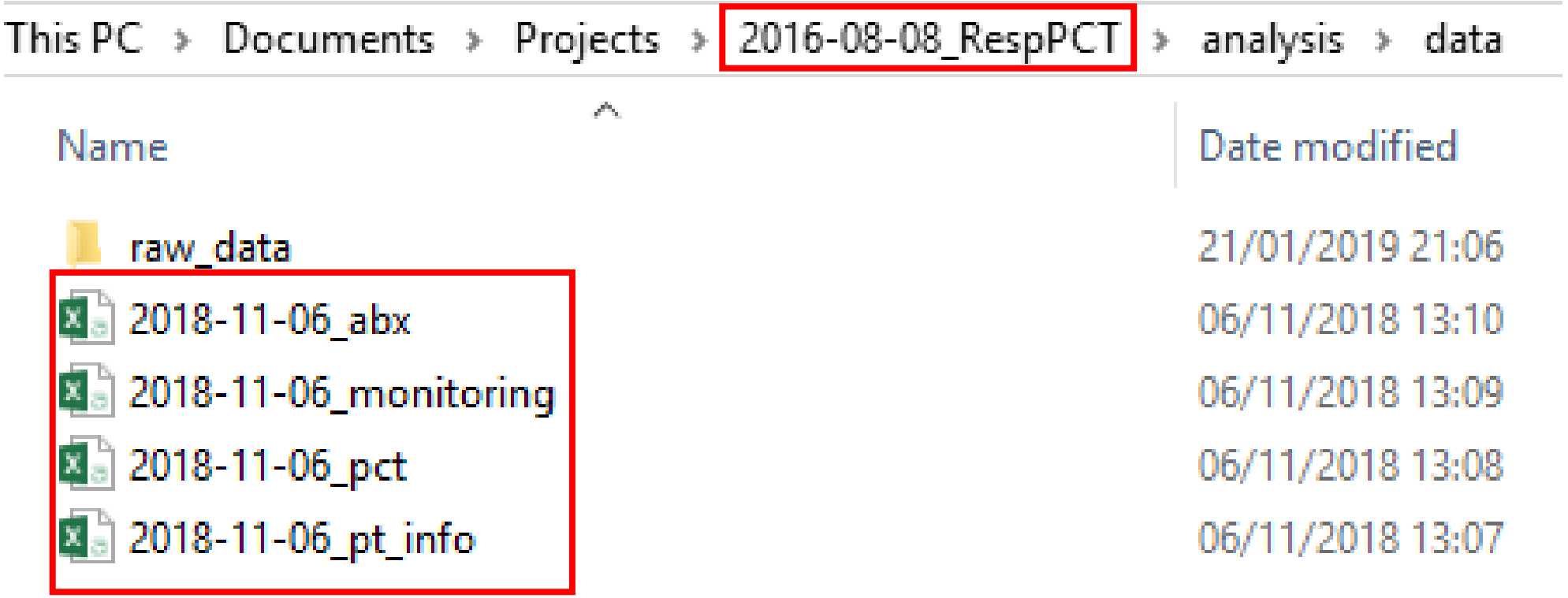
B_Palmer_Medicine_Files › 4a Project › Pyrosequencing_analysis › Pyrosequencing_Paper › Draft_Paper_incl_Figs › Submission › JVI_Resubmission › JVI_resubmission_files › Final Final version

Name	Date modified
Cover_letter_B_A_Palmer_Sept_2014	10/09/2014 17:05
Fig_1_Sept_14	11/09/2014 10:31
Fig_1_Sept_14	10/09/2014 23:07
Fig_2_Sept_14	11/09/2014 10:31
Fig_2_Sept_14	10/09/2014 23:07
Fig_3_Sept_14	11/09/2014 10:31
Fig_3_Sept_14	10/09/2014 23:07
Fig_4_Sept_14	11/09/2014 10:31
Fig_4_Sept_14	10/09/2014 23:07
Fig_5_Sept_14	11/09/2014 10:33
Fig_5_Sept_14	10/09/2014 23:07
HCV_UDPS_B_A_Palmer_Sept_14	17/09/2014 12:21
Response_to_Reviewer_Sept_14	10/09/2014 22:42
Supplementry_Figure_B_A_Palmer_Sept_14	29/08/2014 13:21
Supplementry_Figure_B_A_Palmer_Sept_14	10/09/2014 22:31
Tables_B_A_Palmer_Sept_2014	10/09/2014 22:09

A: Define a generic project structure



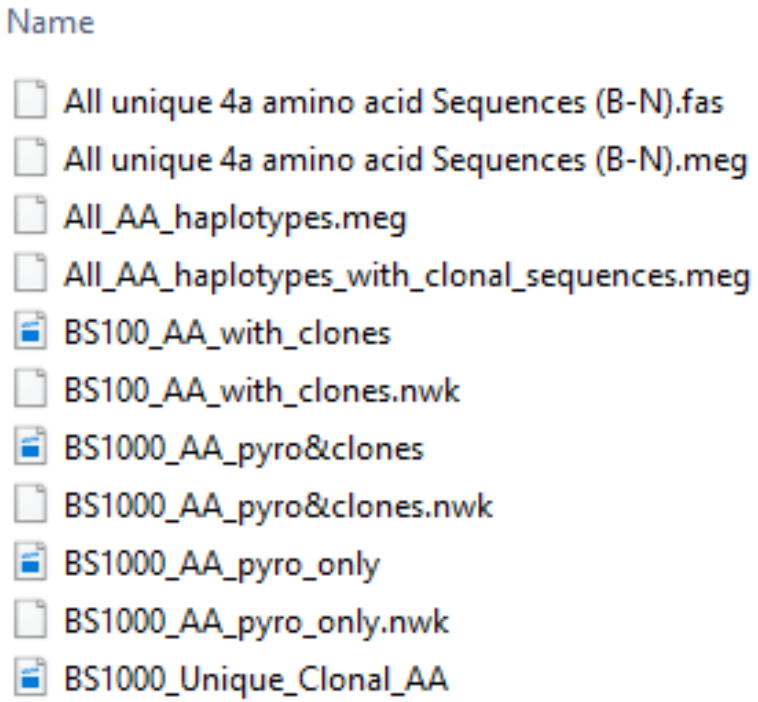
B: Give your files and folders informative names



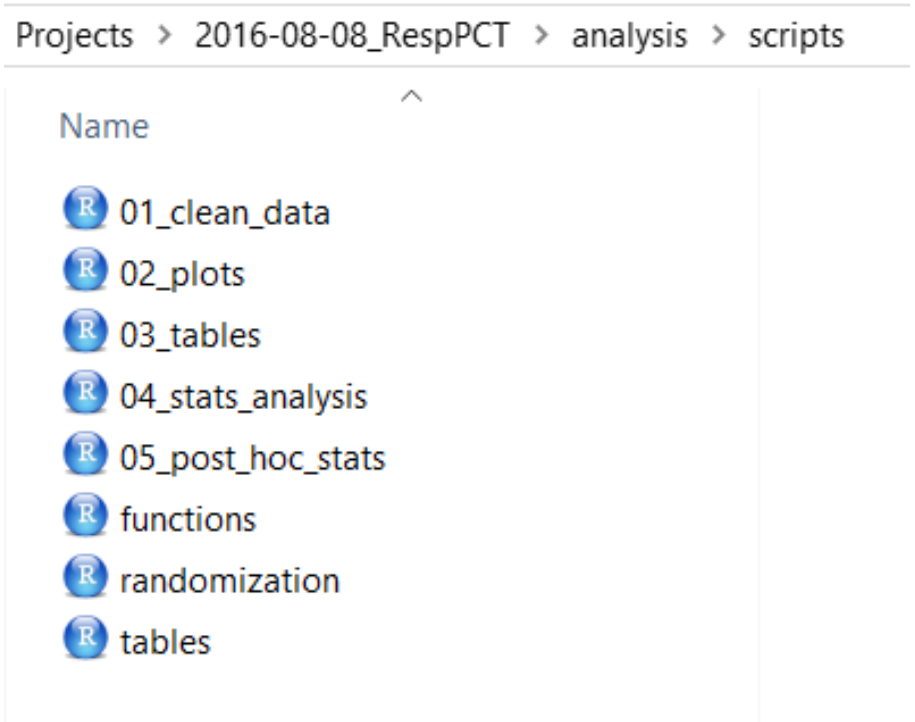
Everything in its right place

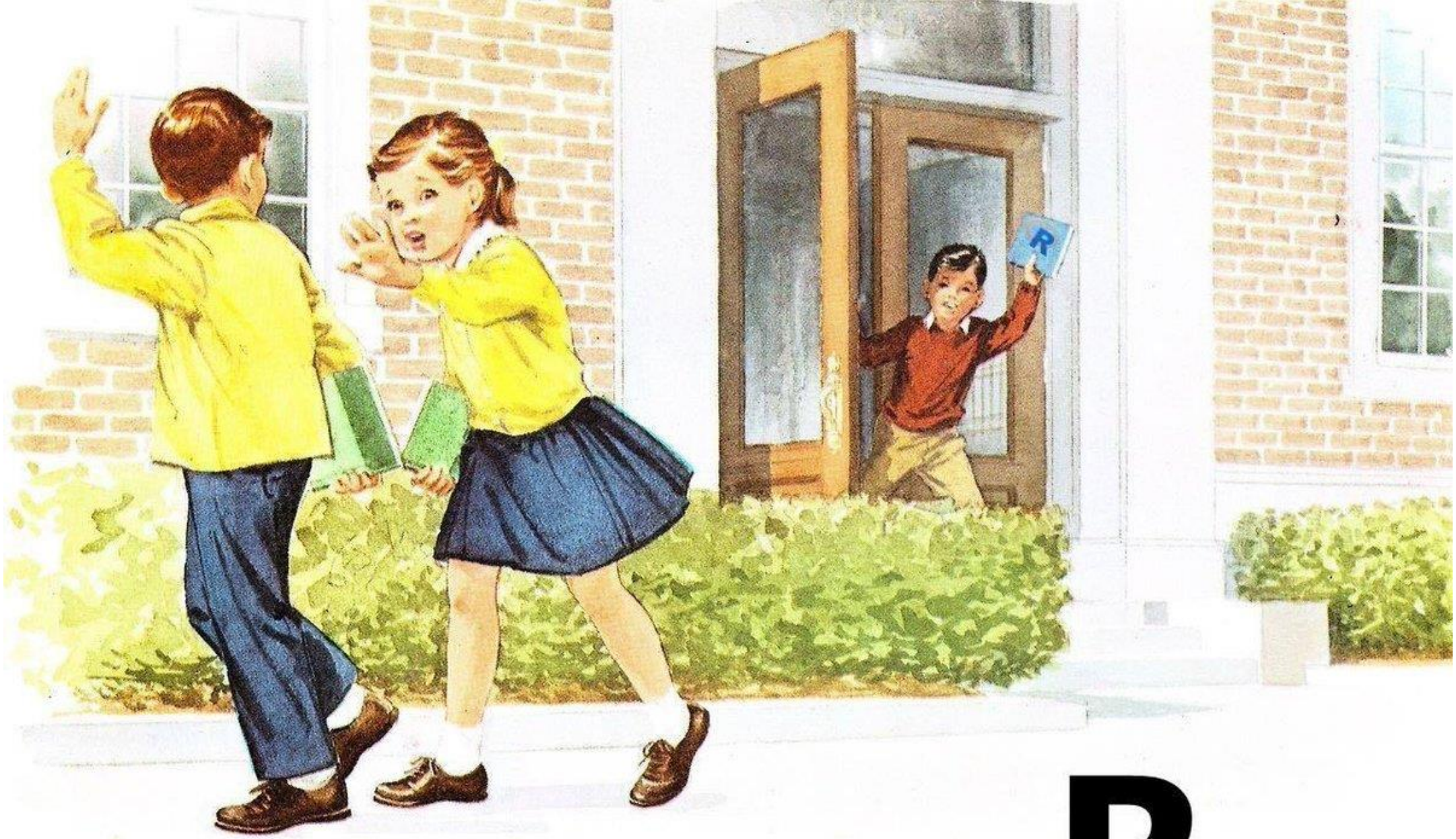
- **Make your file names:**
  1. **Machine readable**
  2. **Human readable**
  3. **Work with default ordering**

NO



Yes





Run, or he's going to tell us about **R** again!

C: Use scripted workflows

- The R scripts should also be human readable
    - Annotate the code
    - Break up the scripts into dedicated tasks
    - Interlink to other project scripts

```
# 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")
```

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?

27 11 7

Dr Gavin Simpson
@ucfagls

Follow

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

D: R Markdown

~/Open_Science/Digital_Badge/RCR - master - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

lettuce_report.Rmd*

Knit | Insert | Run

```
---
title: "This is a reproducible document"
author: "Dr. Brendan Palmer"
date: "18th June 2019"
output:
  word_document:
    fig_height: 4
    fig_width: 6
---
# This is the beginning of the project

Our initial reports might be restricted to lab meetings etc. We can use `R Markdown` to show the code we are using, so that the meetings are not just a demonstration of the results, but also an examination of the `code` used to obtain them.

## Data overview
```{r packages and setup, include = FALSE}

knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE)

# Load your packages here
library(tidyverse)
library(knitr)
```

The plot below is call from the ggplot object entitled `report_plot` created in the script `03_final_analysis.R`.

```{r Plots from script, echo = FALSE}

source("scripts/03_final_analysis.R")

# The location of the Rmd file dictates whether the path to other files is intact
```

This is a reproducible document

Dr. Brendan Palmer

18th June 2019

This is the beginning of the project

Our initial reports might be restricted to lab meetings etc. We can use R Markdown to show the code we are using, so that the meetings are not just a demonstration of the results, but also an examination of the code used to obtain them.

Data overview

The plot below is call from the ggplot object entitled report_plot created in the script 03_final_analysis.R.

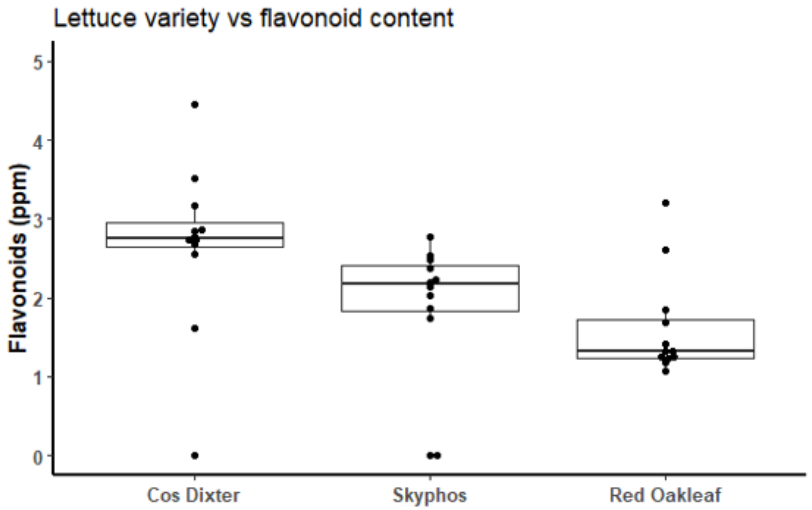
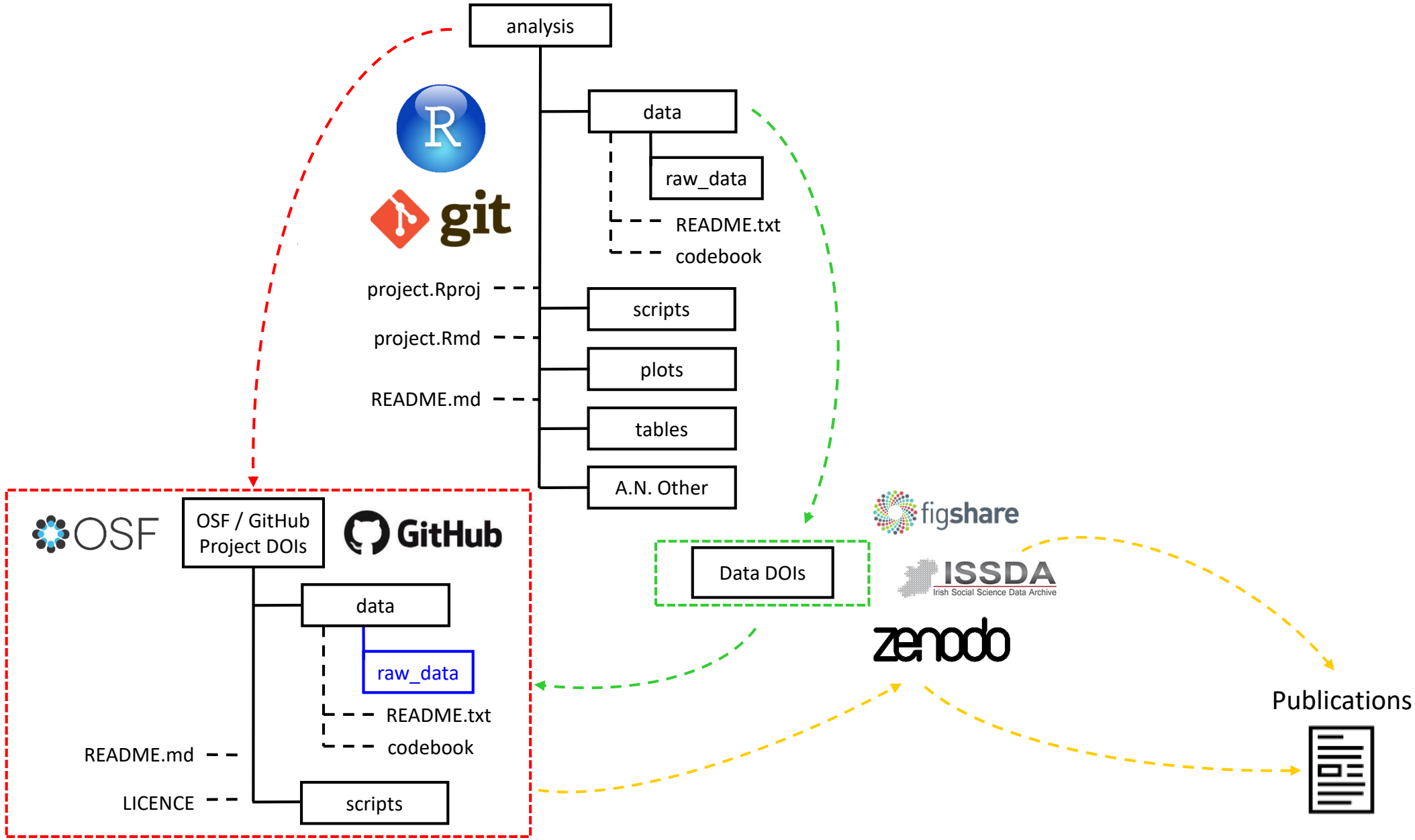


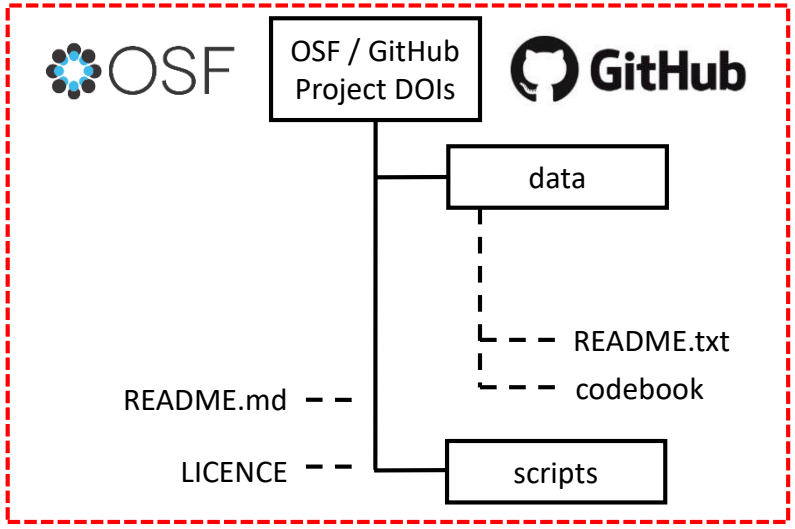
Fig. 1. Flavonoid content of three lettuce varieties under three experimental conditions.

Or we can also recreate the code within the R Markdown document as seen below.
```

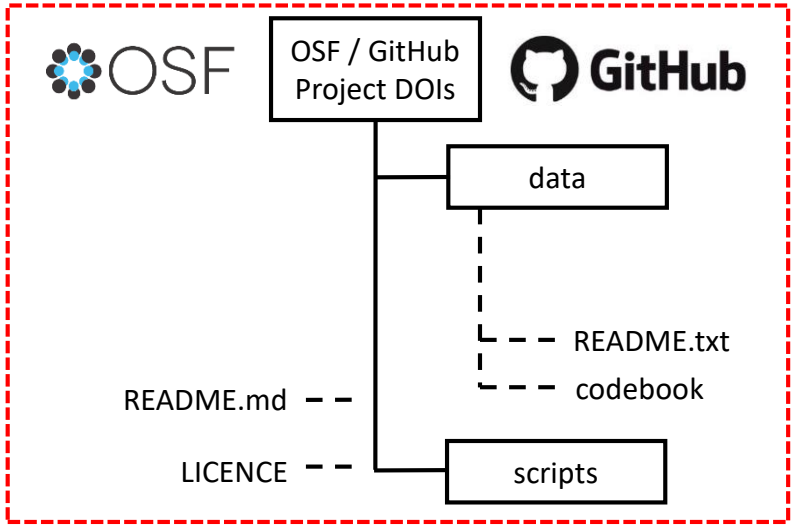
What does this allow us to do?



What does this allow us to do?



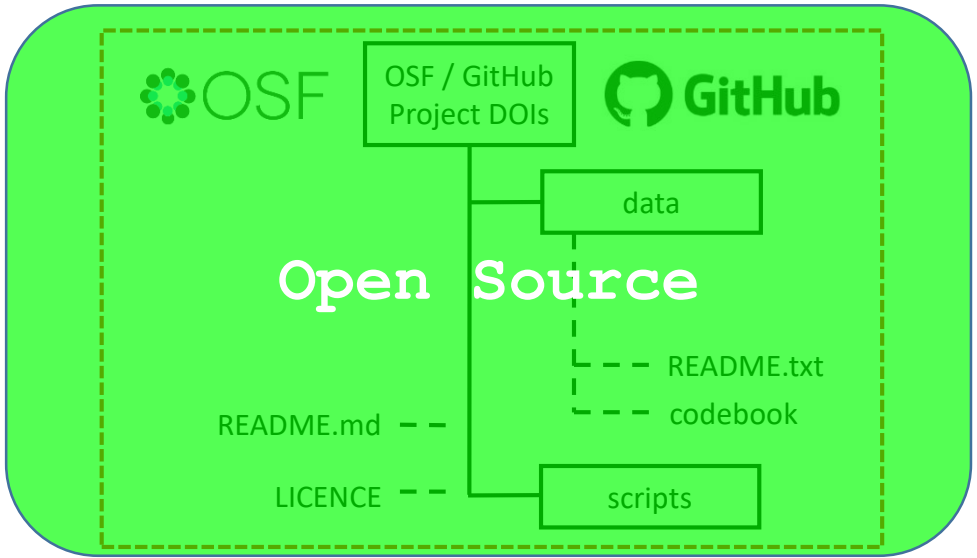
What does this allow us to do?

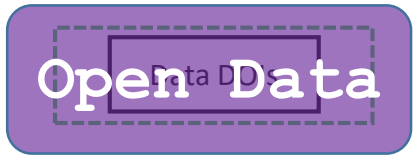


Open Data

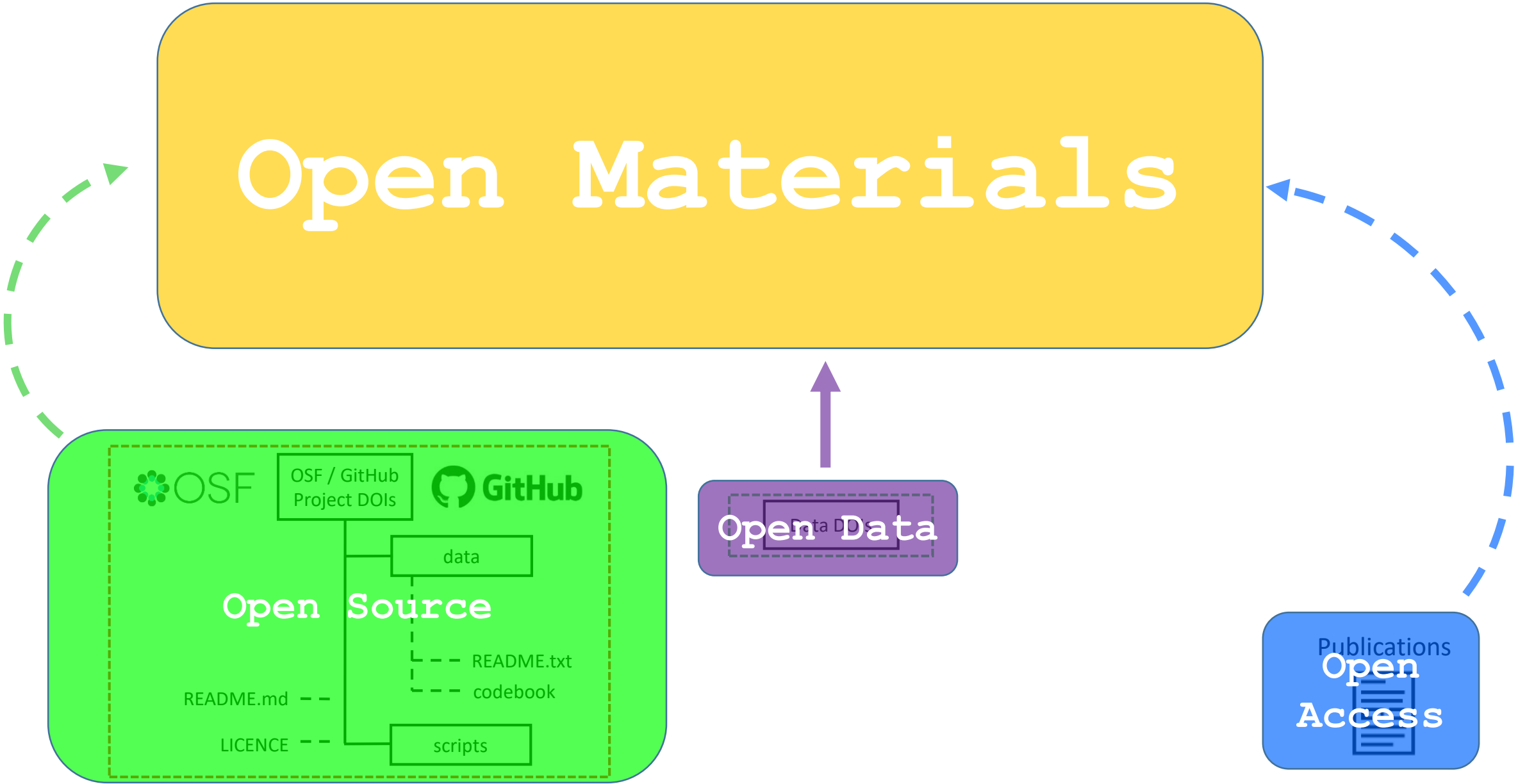
Publications
Open Access

What does this allow us to do?



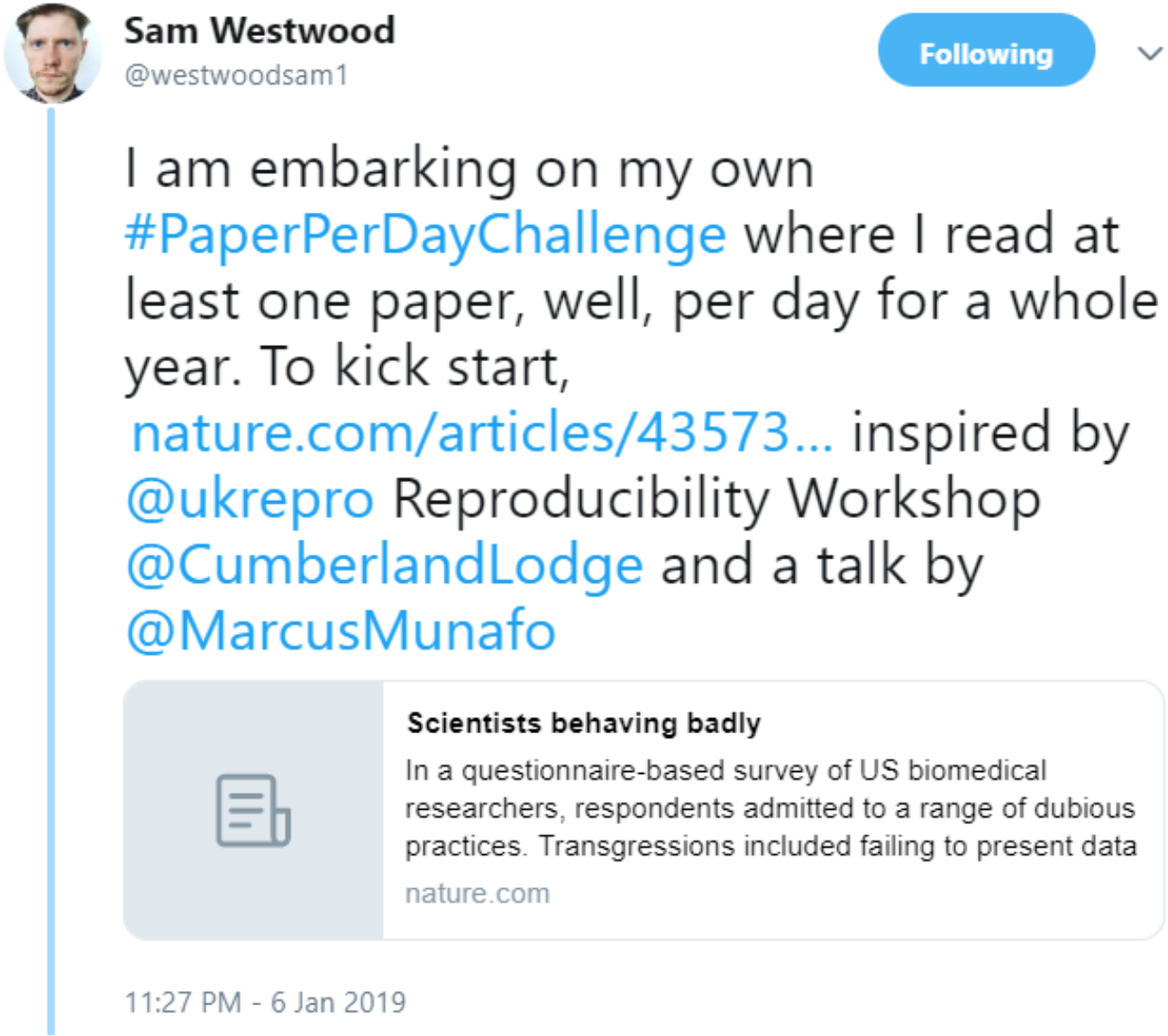


What does this allow us to do?



Install the Chrome plugin PubPeer

Further reading



Can I see your data and code?

1989

* Corresponding author.

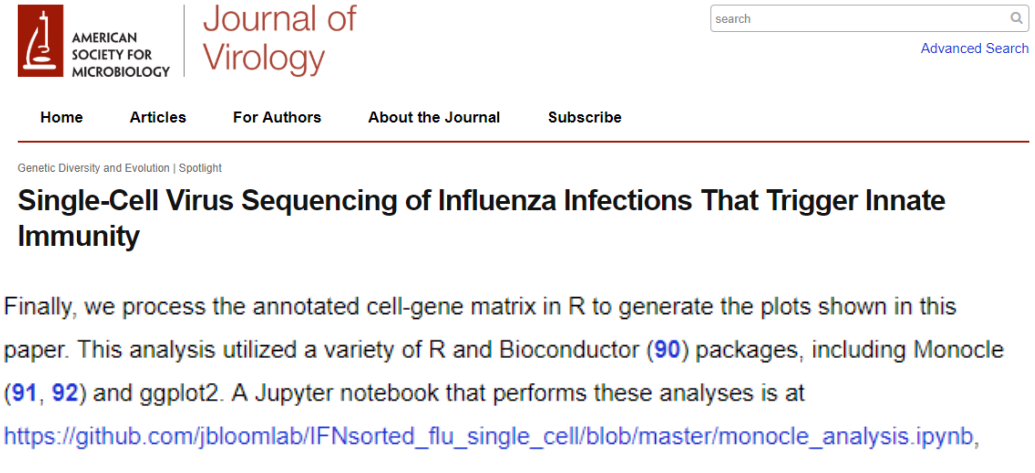
1999

* Corresponding author. Mailing address: Institute of Human Virology, 725 West Lombard St., Rm. N649, University of Maryland, Baltimore, MD 21201. Phone: (410) 706-4680. Fax: (410) 706-4694. E-mail: devico@umbi.umd.edu.

2009

* Corresponding author. Mailing address: Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 201 Althouse Laboratory, University Park, PA 16802. Phone: (814) 863-8705. Fax: (814) 865-7927. E-mail: cec9@psu.edu.

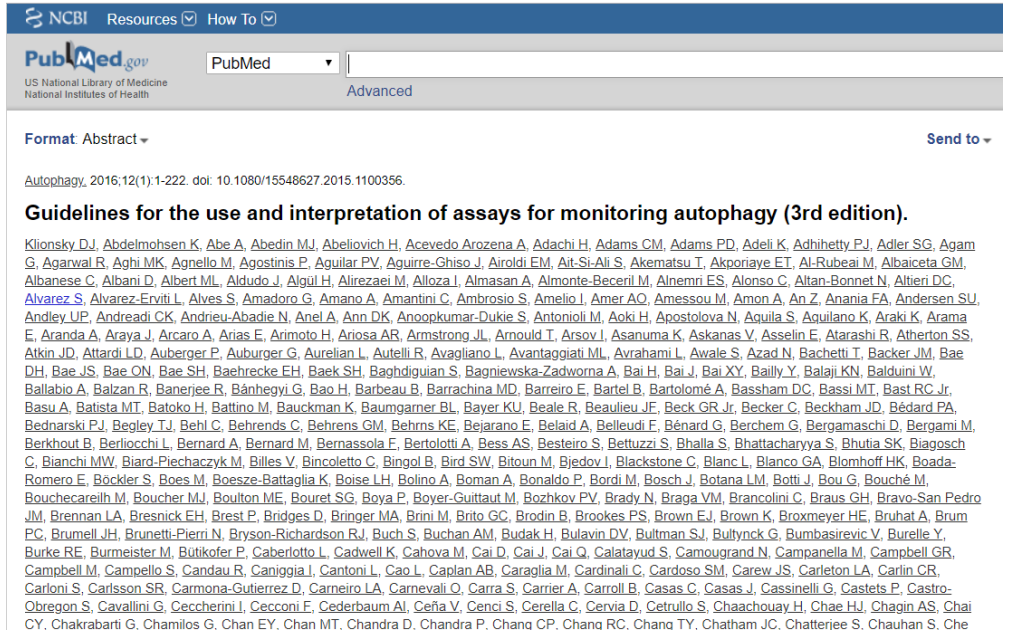
Also 2019



2019

Dr Mark Burnley
@DrMarkBurnley

"I'm the 38th author..."
"Wow, that sucks."



jbloomlab / IFNsorted_flu_single_cell

Watch

Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki | Security | Insights

Branch: master ▾ IFNsorted_flu_single_cell / monocle_analysis.ipynb

jbloom saved plot of association between co-infection / IFN

1 contributor

12.8 MB

Table of Contents

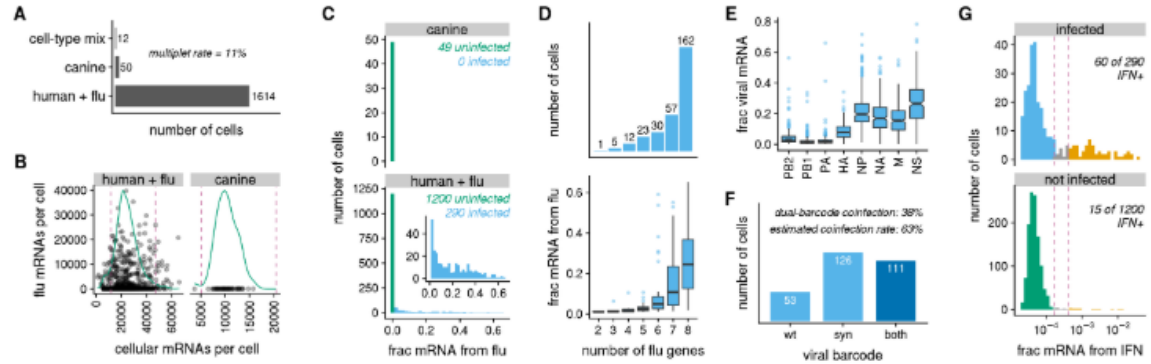
Figures for paper

We have made all of the plots above, and saved some of them to the figures directory already by using the *isfig=TRUE* argument to `saveShowPlot`. However, there are others that we want to assemble into multi-panel figures. We do that here.

First, we assemble a figure that shows the calling of cells, infected cells, and IFN+ cells:

In [101]:
```
p_cell_summary <- plot_grid(
  plot_grid(p_cellcounts, p_flu_vs_cell,
            ncol=1, rel_heights=c(1, 1.5), scale=0.9,
            labels=c("A", "B"), label_size=18, vjust=1),
  plot_grid(p_frac_flu, labels="C", scale=0.95, label_size=18, vjust=1),
  plot_grid(p_nflu_genes, labels="D", scale=0.95, label_size=18, vjust=1),
  plot_grid(p_flu_rel_expr, p_coinfect,
            scale=0.95, ncol=1,
            labels=c("E", "F"), label_size=18, vjust=1),
  plot_grid(p_ifn_dist, labels="G", scale=0.95, label_size=18, vjust=1),
  nrow=1, scale=0.95, rel_widths=c(1, 0.7, 0.6, 0.75, 0.7), align='h'
  ) +
  theme(plot.margin=unit(c(t=0, r=0, b=-0.3, l=0), "in"))

saveShowPlot(p_cell_summary, width=15.5, height=4.9, isfig=TRUE)
```

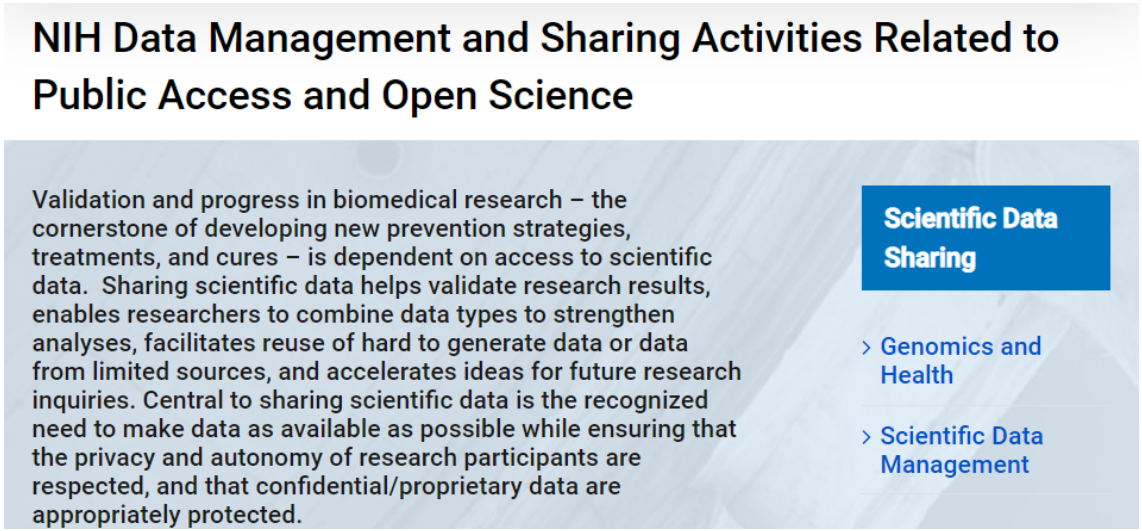


Now a supplementary figure to the one above with the single-cell transcriptomic data:

In [102]:
```
p_cell_summary_supp <- plot_grid(
  p_frac_has_gene,
  p_ifn_genes_corr,
  p_isg_dist,
  p_isg_corr,
  ncol=2,
  scale=0.9,
  rel_heights=c(0.68, 1),
  labels=c("A", "B", "C", "D"), label_size=18, vjust=2, hjust=-1
  )

saveShowPlot(p_cell_summary_supp, width=9.5, height=7.5)
```

Data management is a part of your life now!



Twitter

UK Reproducibility Network
@ukrepro

UK Reproducibility Network: a peer-led consortium to investigate factors which contribute to robust research, provide training, and disseminate best practice.

Malcolm Macleod #FBPE
@Maclomaclee Follows you

clinical neurologist, stroke trialist, and interested in improving the quality of laboratory research

Open Science MOOC
@OpenScienceMOOC Follows you

A community designed for students and researchers to help make 'Open' the default setting for the future of research. Slack: osmooc.herokuapp.com

Everywhere

Brian Nosek
@BrianNosek

Executive Director @ Center for Open Science, Professor @ University of Virginia, and co-Founder of Project Implicit

Kate Button
@ButtonKate Follows you

Academic. Psychologist. Cognitive mechanisms of depression & anxiety. Meta-science & scientific rigour. Sporadic Twitterer.

Darren L Dahly
@statsepi Follows you

Principal Statistician, Epidemiologist, Sr Lecturer | @HRBIreland Clinical Research Facility @CRF_CORK | Cork #Rstats Users Group meetup.com/Cork-Ireland-R...

Dorothy Bishop
@deevybee

Professor of developmental neuropsychology. Blog on deevybee.blogspot.com Main focus #devlangdis, see: youtube.com/radld

Elisabeth Bik
@MicrobiomDigest

Science consultant, PhD. Harbers-Bik LLC. Microbiome, research integrity & misconduct. Ex @Stanford. MicrobiomeDigest/Bik's Picks. Dutch/USA. My views.

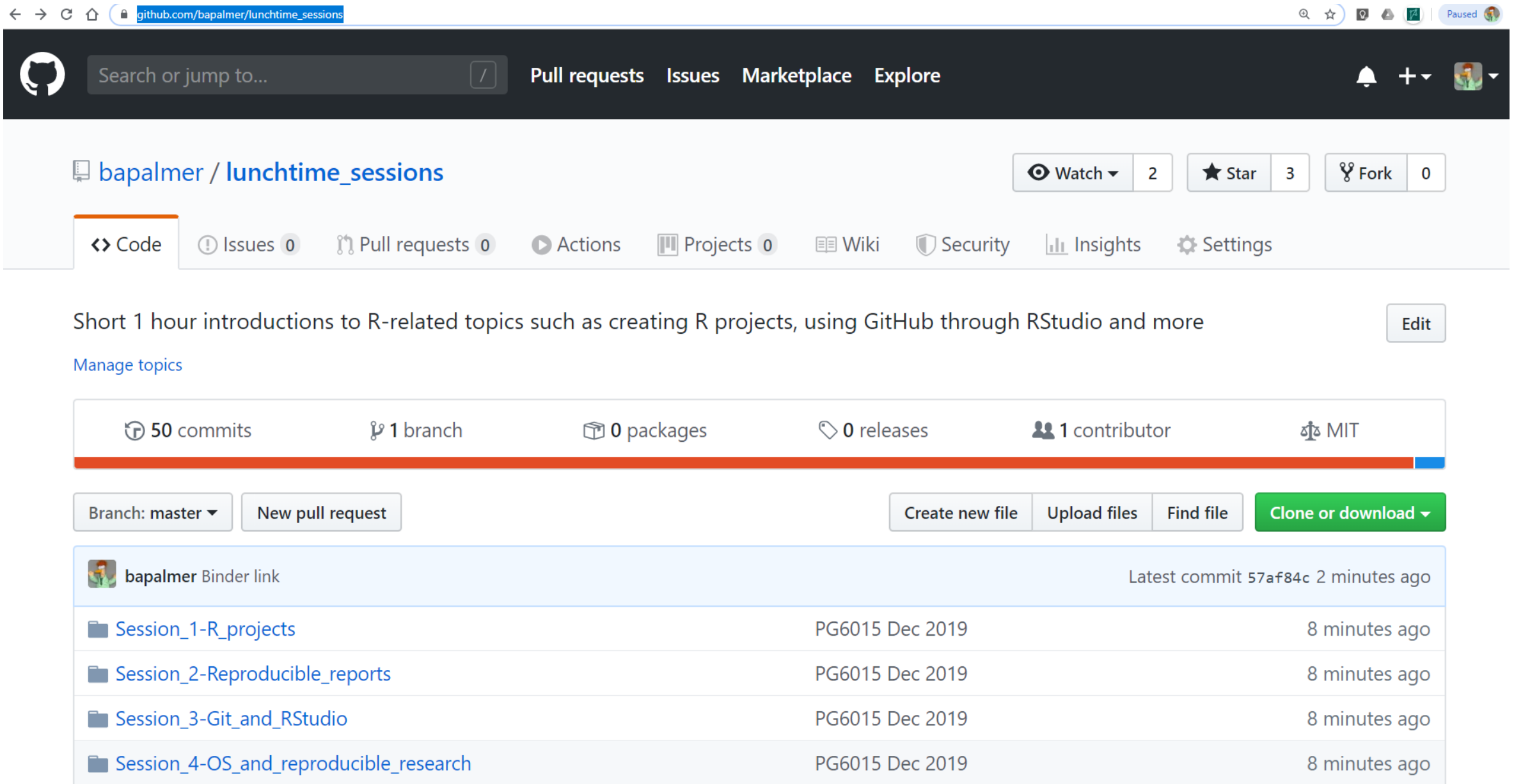
Retraction Watch
@RetractionWatch

Tracking retractions as a window into the scientific process. Sign up for our daily newsletter: eepurl.com/bNRlUn Tips? team@retractionwatch.com

Jenny Bryan
@JennyBryan

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

Lets try it out



R is for Resources

Vignettes

Package 'ggplot2'

December 30, 2016

Version 2.2.1

Title Create Elegant Data Visualisations Using the Grammar of Graphics

Description A system for 'declaratively' creating graphics, based on "The Grammar of Graphics". You provide the data, tell 'ggplot2' how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details.

Depends R (>= 3.1)

Imports digest, grid, gtable (>= 0.1.1), MASS, plyr (>= 1.7.1), reshape2, scales (>= 0.4.1), stats, tibble, lazyeval

Suggests covr, ggplot2movies, hexbin, Hmisc, lattice, mapproj, maps, maptools, mgcv, multcomp, nlme, testthat (>= 0.11.0), quantreg, knitr, rpart, rmarkdown, svglite

Enhances sp

License GPL-2 | file LICENSE

URL http://ggplot2.tidyverse.org, https://github.com/tidyverse/ggplot2

BugReports https://github.com/tidyverse/ggplot2/issues

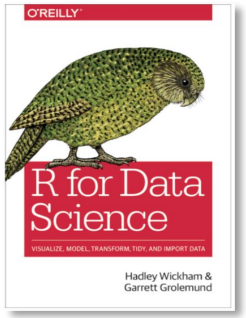
LazyData true

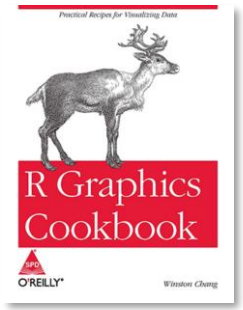
Collate 'ggproto.r' 'aaa-.r' 'aes-calculated.r' 'aes-colour-fill-alpha.r' 'aes-group-order.r' 'aes-linetype-size-shape.r' 'aes-position.r' 'utilities.r' 'aes.r' 'legend-draw.r' 'geom-.r' 'annotation-custom.r' 'annotation-logticks.r' 'geom-polygon.r' 'geom-map.r' 'annotation-map.r' 'geom-raster.r' 'annotation-raster.r' 'annotation.r' 'autolayer.r' 'axis-secondary.R' 'bench.r' 'bin.R' 'coord-.r' 'coord-cartesian-.r' 'coord-fixed.r' 'coord-flip.r' 'coord-map.r' 'coord-munch.r' 'coord-polar.r' 'coord-quickmap.R' 'coord-transform.r' 'data.R' 'facet-.r' 'facet-grid-.r' 'facet-null.r' 'facet-wrap.r' 'fortify-lm.r' 'fortify-map.r' 'fortify-multcomp.r' 'fortify-spatial.r' 'fortify.r' 'stat-.r' 'geom-abline.r' 'geom-rect.r' 'geom-bar.r' 'geom-bin2d.r' 'geom-blank.r' 'geom-boxplot.r' 'geom-col.r' 'geom-path.r' 'geom-contour.r' 'geom-count.r'

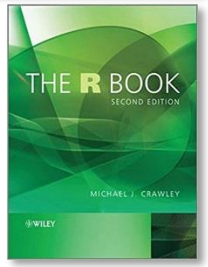
1

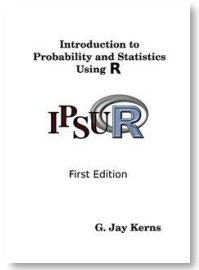
Webpages

eBooks









Cheatsheets

Twitter

Mara Averick
@dataandme

tidyverse @rstudio, hoop head, gnashgab, blatherskite, lesser ½ of @batpigandme

Massachusetts

One R Tip a Day
@RLangTip

One tip per day M-F on the R programming language #rstats. Brought to you by the R community team at Microsoft.

Hadley Wickham
@hadleywickham

R, data, visualisation.

Houston, TX

hadley.nz

David Robinson
@drob

Data Scientist at @StackOverflow, #rstats fan/evangelist

New York, NY

varianceexplained.org

Jenny Bryan
@JennyBryan

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

Vancouver, BC

jennybryan.org

Darren L Dahly
@statsepi Follows you

Principal Statistician, Epidemiologist, Sr Lecturer | @HRBIreland Clinical Research Facility @CRF_CORK | Cork #Rstats Users Group meetup.com/Cork-Ireland-R...

Cork, Ireland

darrendahly.github.io

Data Scientists IRL
@DataSci_Ireland Follows you

Promoting the Data Science professions in Ireland.

Ireland

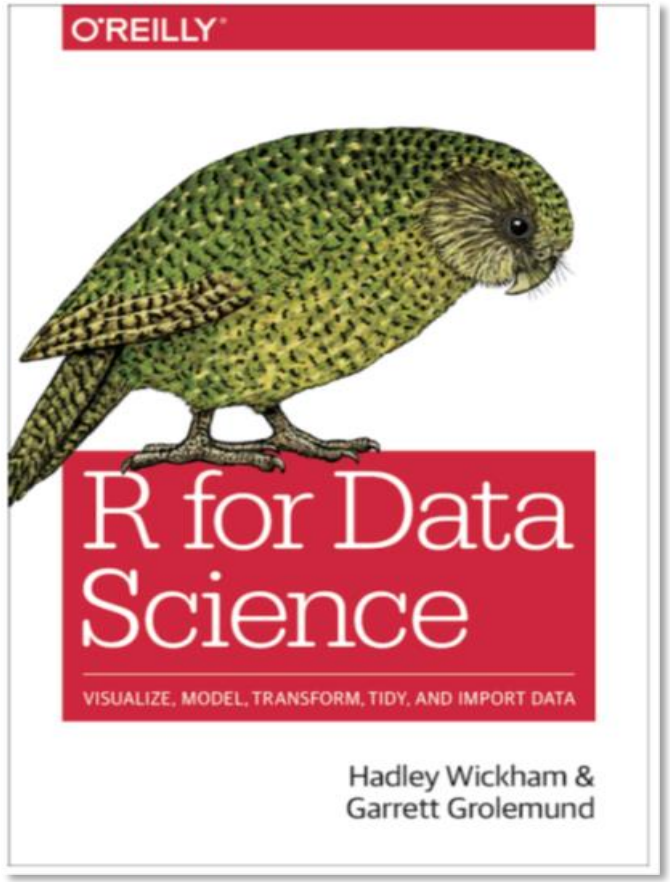
facebook.com/DataScientists...

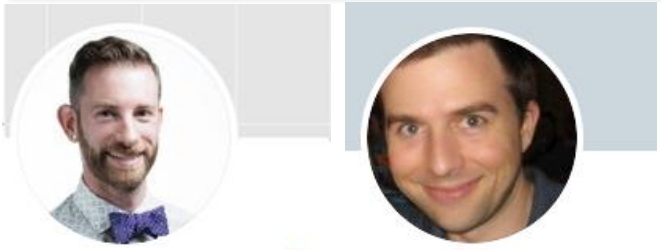
Kara Woo
@kara_woo

Research scientist at @sagebio. Data curation, visualization, #rstats, reproducibility, open science, ballet

karawoo.com

R is for Resources





Books psyTeachR

Course Books

Level 1: Grassroots

Our first-year undergraduate course covers current state of psychological science and what Open Science is as well as its importance. It also aims to make students confident and competent at using RStudio as a tool to achieve good data management skills.

Authors: Emily Nordmann, Heather Woods

Contact: Emily Nordmann

Contributors: Jack Taylor, Shannon McNee

Level 2: Practical

Our second-year undergraduate course covers data skills such as R Markdown, data wrangling with tidyverse, and data visualisation with ggplot2. It also introduces statistical concepts such as permutation tests, NHST, alpha, power, effect size, and sample size. Semester 2 focusses on correlations and the general linear model.

Authors: Phil McAleer, Helena Paterson

Contact: Phil McAleer

Level 3: Statistical Models (Coming Soon)

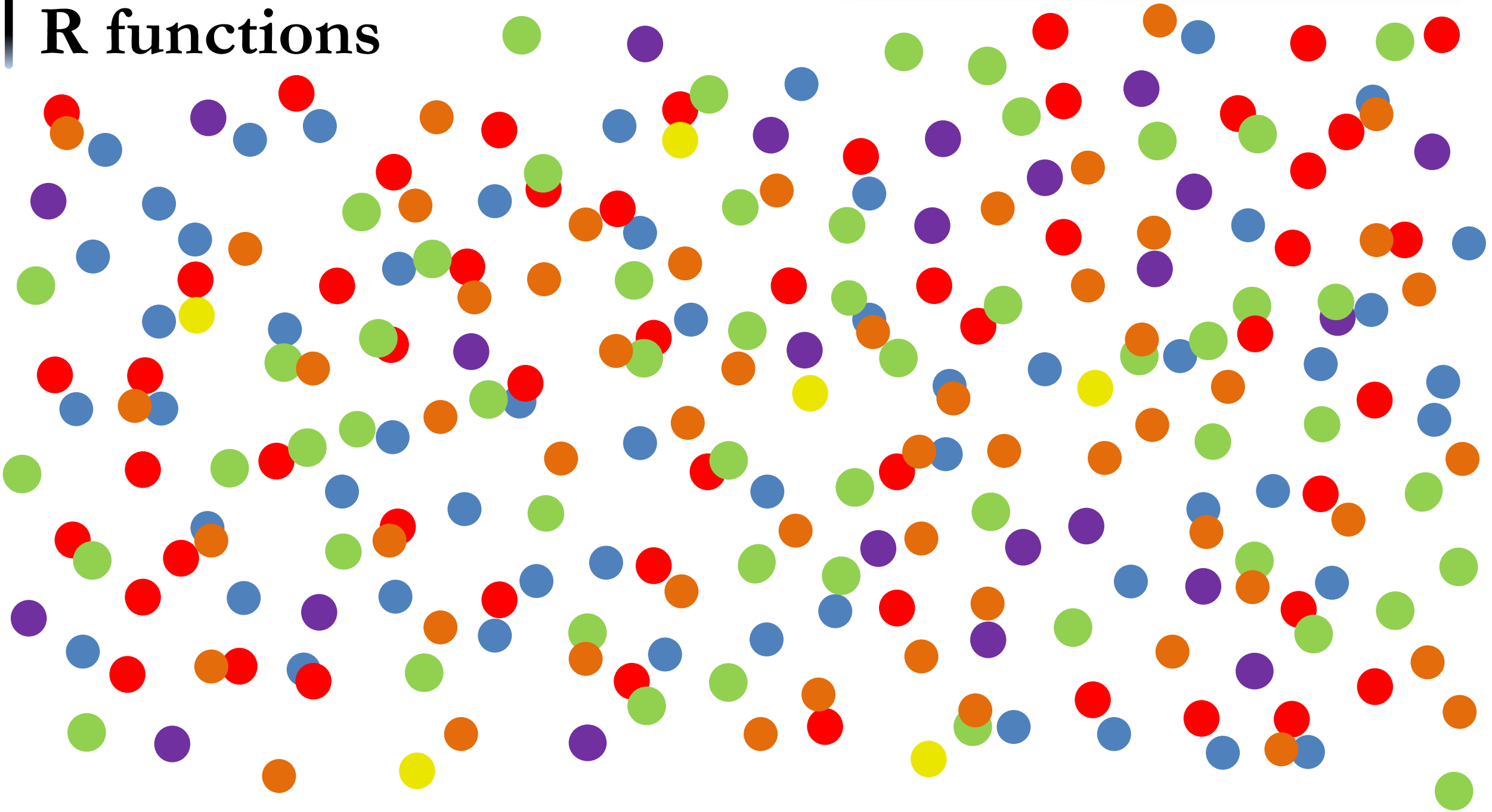
This third-year undergraduate course teaches students how to specify, estimate, and interpret statistical models corresponding to various study designs, using a General Linear Models approach.

Author: Dale Barr

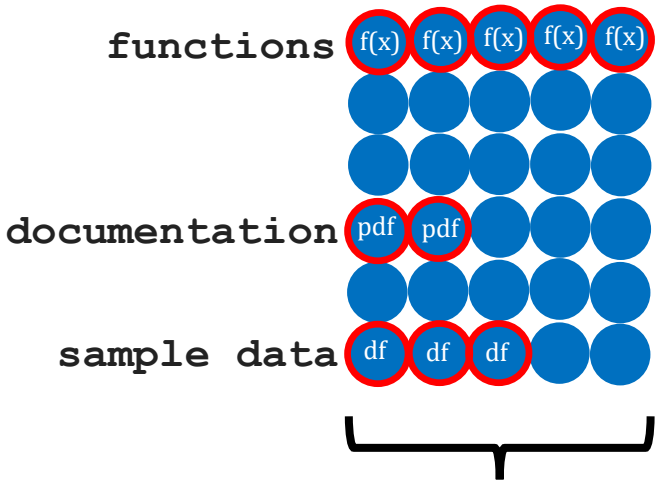
MSc Conversion

This book contains materials for students on the MSc Conversion in Psychological Studies/Science, a one-year postgraduate degree for students with a non-psychology undergraduate degree. This research methods course covers core data skills that allow you to manipulate and analyse quantitative data.

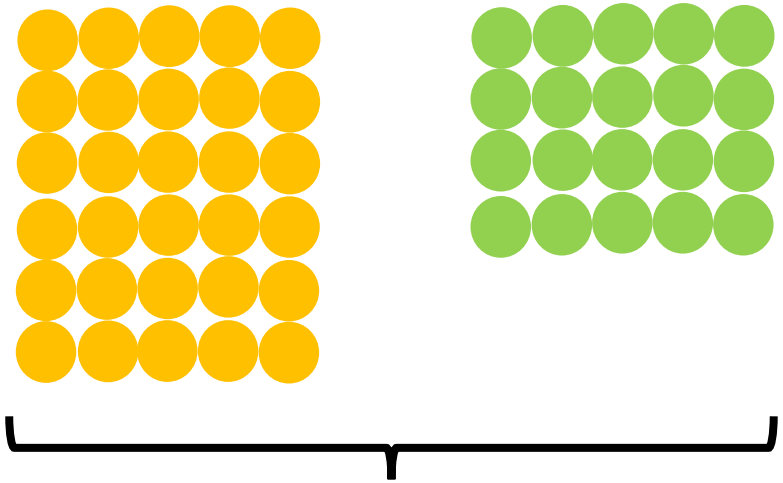
R functions



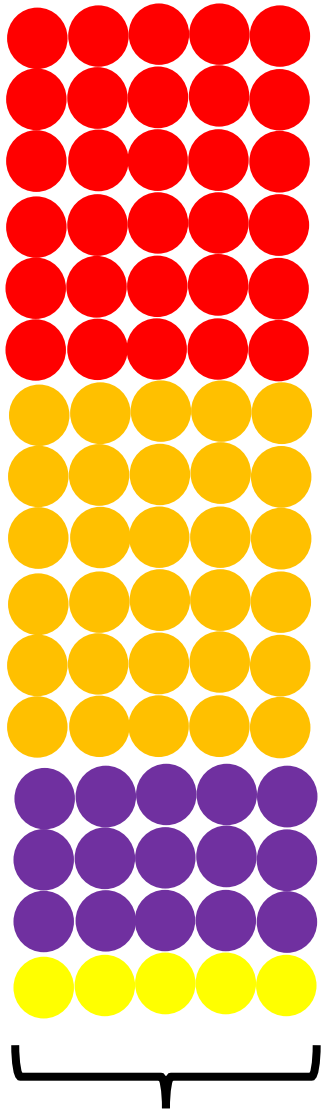
R packages



R comes pre-loaded with ~30 other packages (e.g. base, stats, graphics etc.)

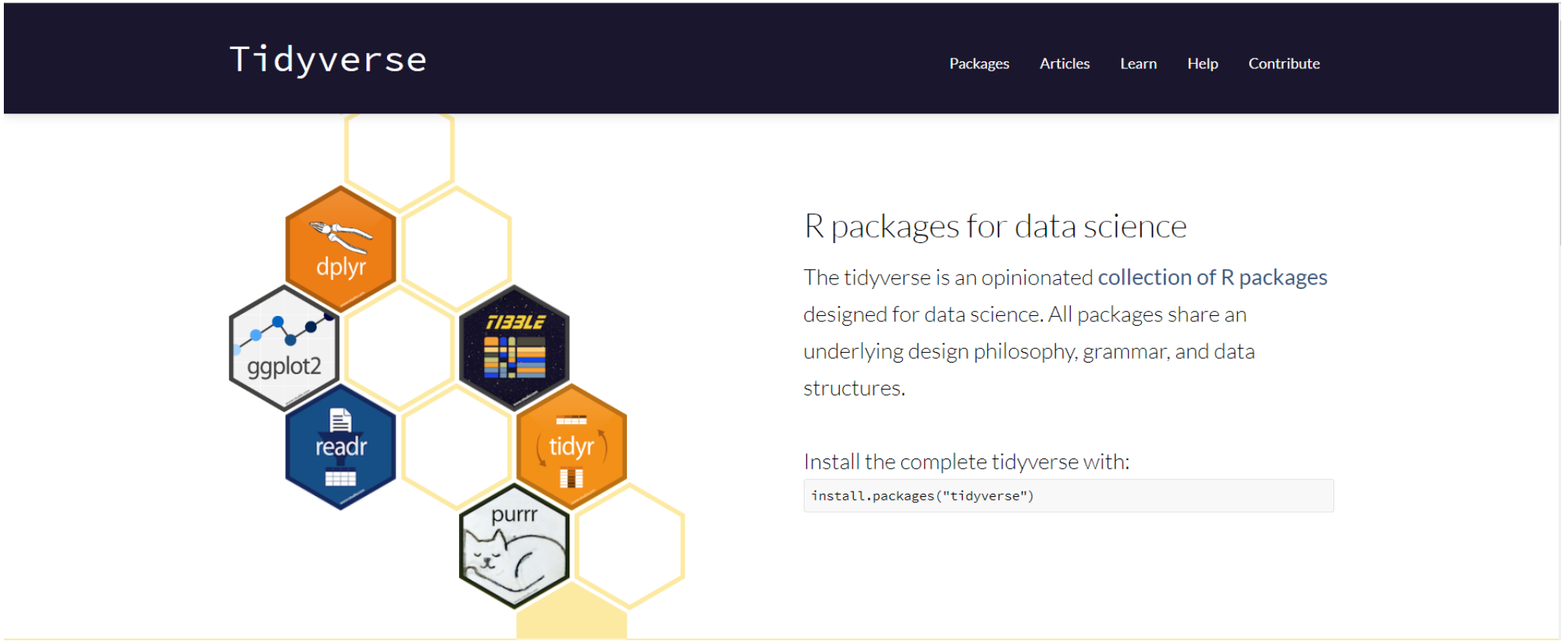


Other packages:
Install once
Update regularly
Load each session



tidyverse

What is the tidyverse?



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

R user interface versus RStudio

RGui (64-bit)

File Edit View Misc Packages Windows Help

R Console

```
R version 3.6.1 (2019-07-05) -- "Action of the Toes"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

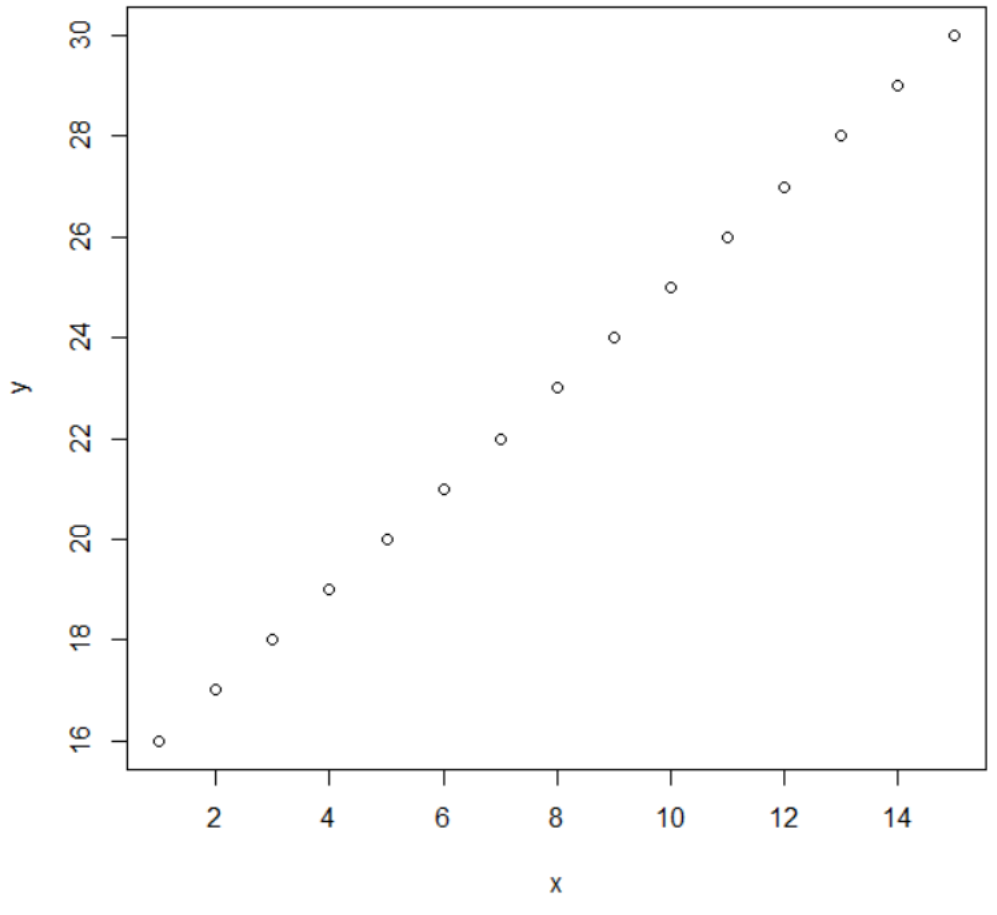
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

> print("Hello world!!")
[1] "Hello world!!"
> object <- "Hello world!!"
> object
[1] "Hello world!!"
> x <- 1:15
> y <- 16:30
> z <- mean(x + y)
> z
[1] 31
> plot(x,y)
>
```

R Graphics: Device 2 (ACTIVE)



File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Pre-workshop

test_script.R

Source on Save | Run | Source

```
31
32    gather(sample, expression, G0.05:U0.3) %>%
33
34    separate(sample, c("nutrient", "rate"), sep = 1, convert = TRUE) %>%
35
36    mutate(nutrient = plyr::revalue(nutrient, nutrient_names)) %>%
37
38    filter(!is.na(expression), systematic_name != "")
39
40  # Plot the clean data
41
42  cleaned_genes_tbl %>%
43
44    filter(BP == "leucine biosynthesis") %>%
45
46    ggplot(mapping = aes(x = rate, y = expression, color = nutrient)) +
47    geom_point() +
48    geom_smooth(method = "lm", se = FALSE) +
49    facet_wrap(~ name)
```

Code editor

49:21 | Does this work on your system? | R Script

Environment | History | Connections | Git

Import Dataset | Grid

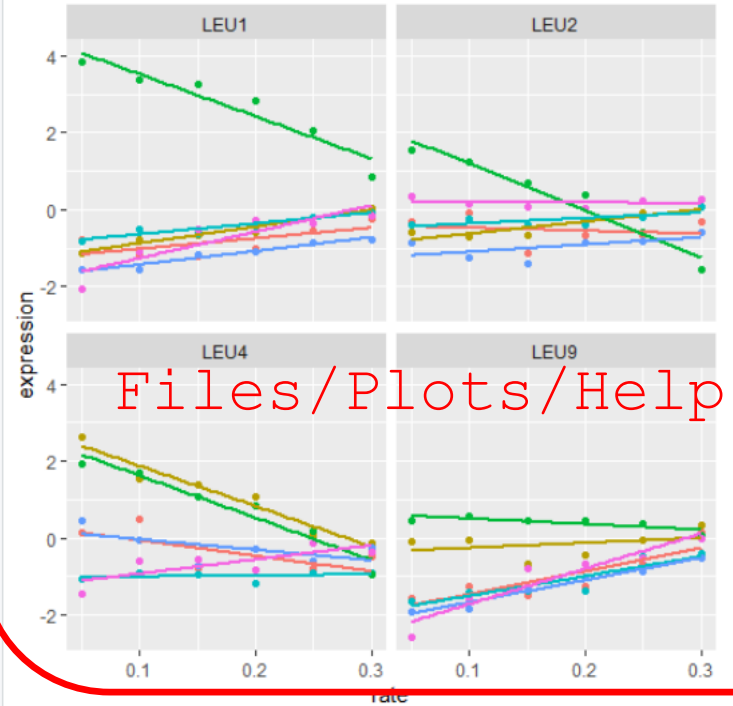
Global Environment

Name	Type	Length	Size	Value
cleaned_g...	tbl_df	7	11.3 ...	198430 obs. of 7...
nutrient_...	charac...	6	984 B	Named chr [1:6] "G...
url	charac...	1	168 B	"http://varianceex...

Environment

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish



Files/Plots/Help

Console | Terminal

~/R_Users_Workshop/PG_module/course_notes/R-A_Hitchhikers_Guide_to_Reproducible_Research/Pre-workshop/

```
  GID = col_character(),
  YORF = col_character(),
  NAME = col_character()
)

See spec(...) for full column specifications.
> cleaned_genes_tbl %>%
+
+    filter(BP == "leucine biosynthesis") %>%
+
+    ggplot(mapping = aes(x = rate, y = expression, color = nutrient)) +
+    geom_point() +
+    geom_smooth(method = "lm", se = FALSE) +
+    facet_wrap(~ name)
>
```

R console

To understand R, remember the following

- **Everything that exists is an object**
- **Everything that happens is a function**

Creating objects

For most of us, R is simply the creation of and manipulation of objects:

```
new_object <- c(1, 2, 3)
```

- the objects are then fed into functions to create amazing new objects

```
amazing_new_object <- function(new_object)
```

Broadly speaking the following is true in R:

- information

```
> data_frame  <- function(information)
> plot        <- function(data_frame)
> model       <- function(data_frame)
```

Give it a go

An example scripted, reproducible workflow

Link to packaged data, code and reports