Data Wrangling in R

http://sisbid.github.io/Data-Wrangling/

Course Info

Course name	Data Wrangling in R
Instructors	Carrie Wright and Ava Hoffman
Course website	http://sisbid.github.io/Data-Wrangling/
Goals	Teach you how to get and clean data
Pre-reqs	Hopefully some R programming

Slack is helpful for:

- Getting help
- Talking with peers
- Accessing recordings

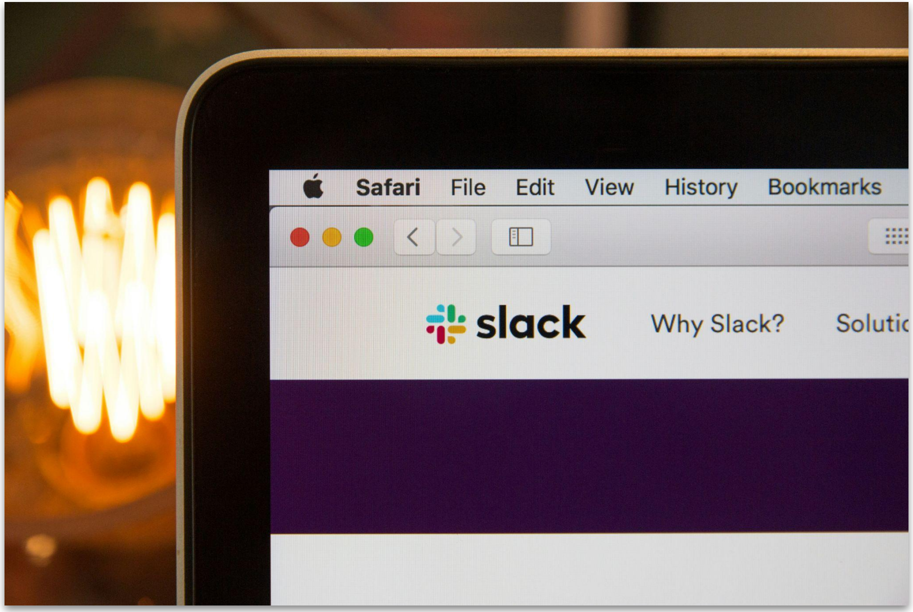


Photo by Stephen Phillips - Hostreviews.co.uk on Unsplash

What we will cover in this course:

1) Why data wrangling is important?
2) Importing data (and outporting)
3) Subsetting data
4) Summarizing data
5) Cleaning data
6) Reshaping data
7) Data merging and joining
8) Functional programming (efficiently applying functions)
9) Working with factors, strings, dates
10) Version control - Git/GitHub

But first, some jargon!

Packages

A bundle or "package" of code (and or possibly data) that can be loaded together for easy repeated use or for sharing with others.

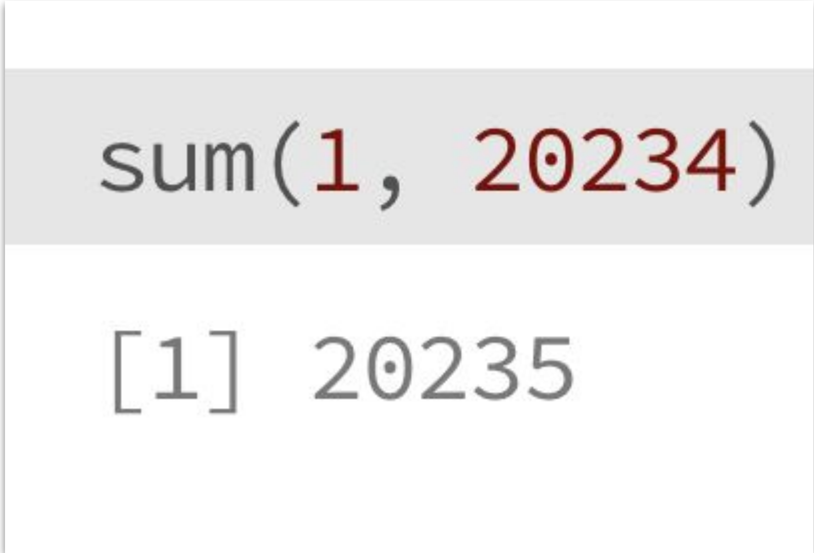
Packages are analogous to a software application like Microsoft Word on your computer.

Function

A piece of code that allows you to do something in R. Packages often contain functions.

You can think of a function as verb in R.

A function might help you add numbers together, create a plot, or organize your data. More on that soon!

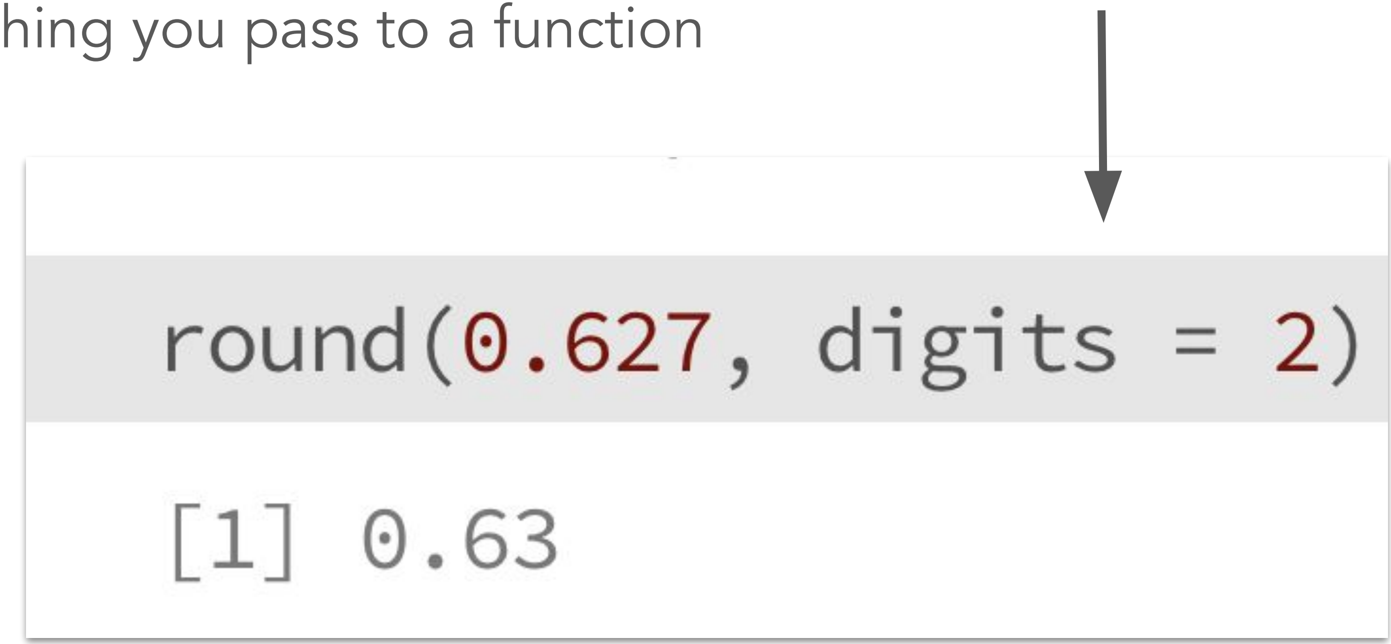


```
sum(1, 20234)
```

```
[1] 20235
```

Argument

Something you pass to a function



```
round(0.627, digits = 2)
```

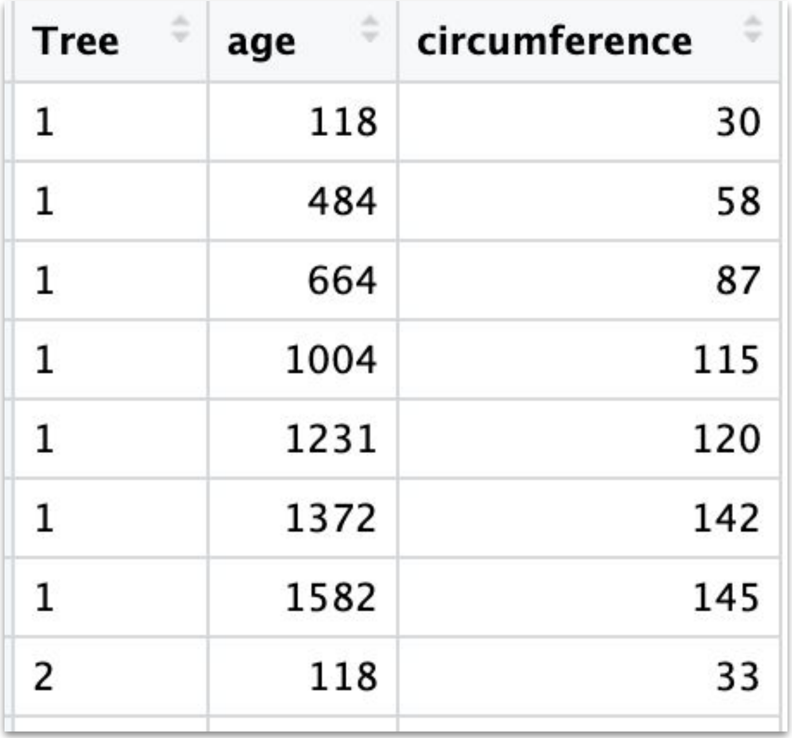
```
[1] 0.63
```

An Object

Something that can be worked with or on in R - can be lots of different things!

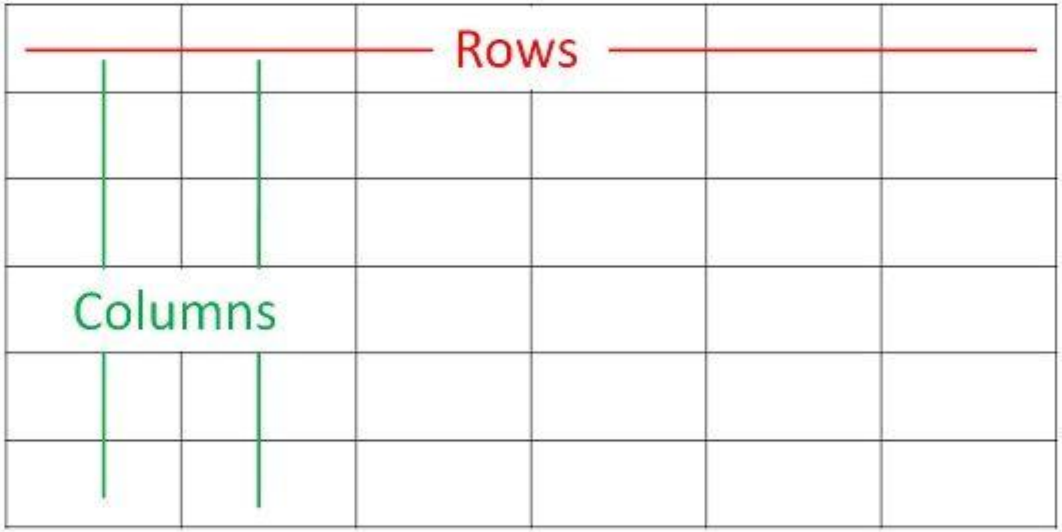
You can think of an objects as a noun in R.

An object might be a data table, a plot, a function or more!



Tree	age	circumference
1	118	30
1	484	58
1	664	87
1	1004	115
1	1231	120
1	1372	142
1	1582	145
2	118	33

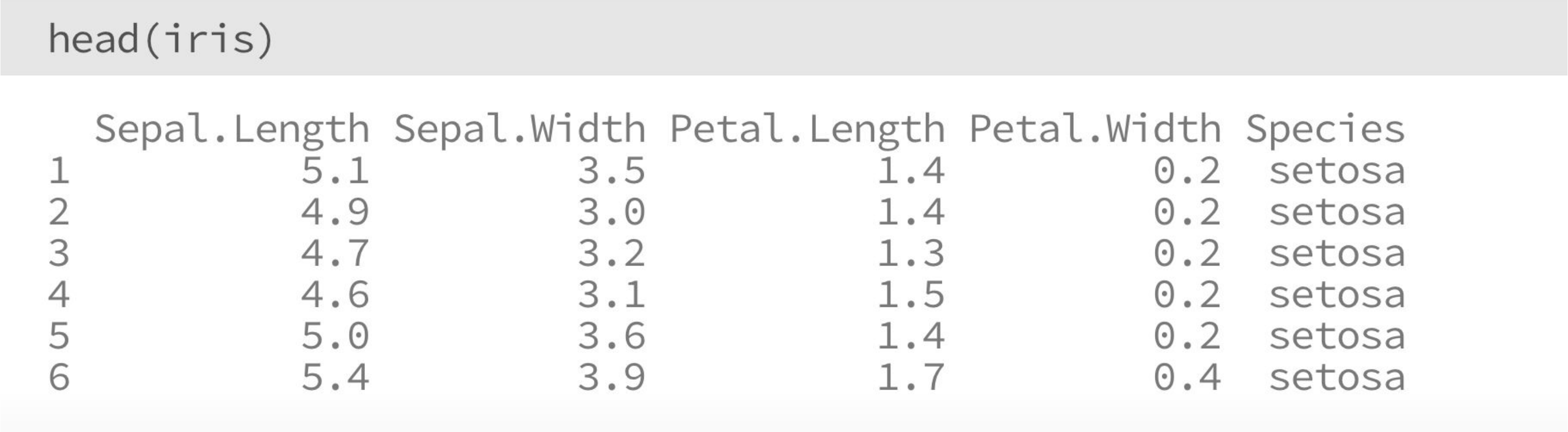
Dataframes/DataTables/Spreadsheets



Dataframes/Data tables

Rows = samples - individuals, locations, houses, viruses etc.

Columns = variables - aspects or features measured, color, count, type etc.



```
head(iris)
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

Need more help?

R jargon:

https://link.springer.com/content/pdf/bbm%3A978-1-4419-1318-0%2F1.pdf

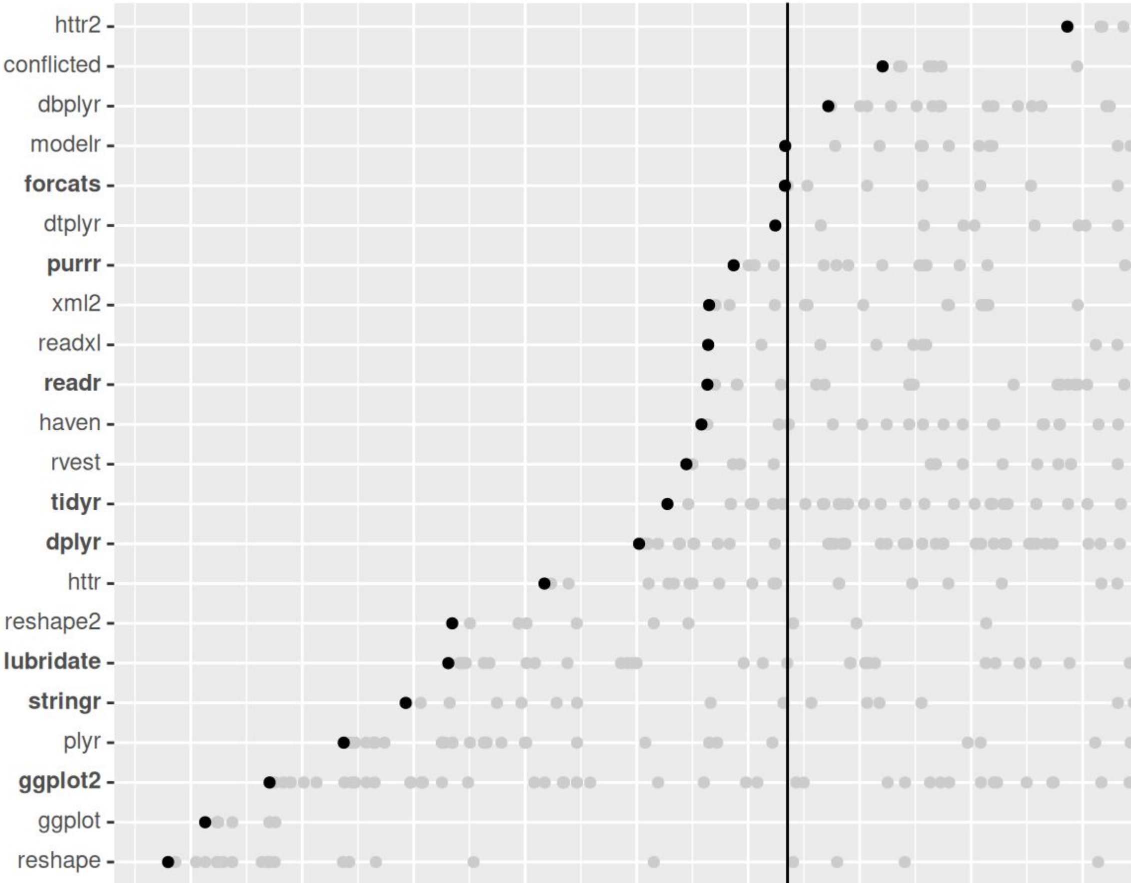
The Tidyverse

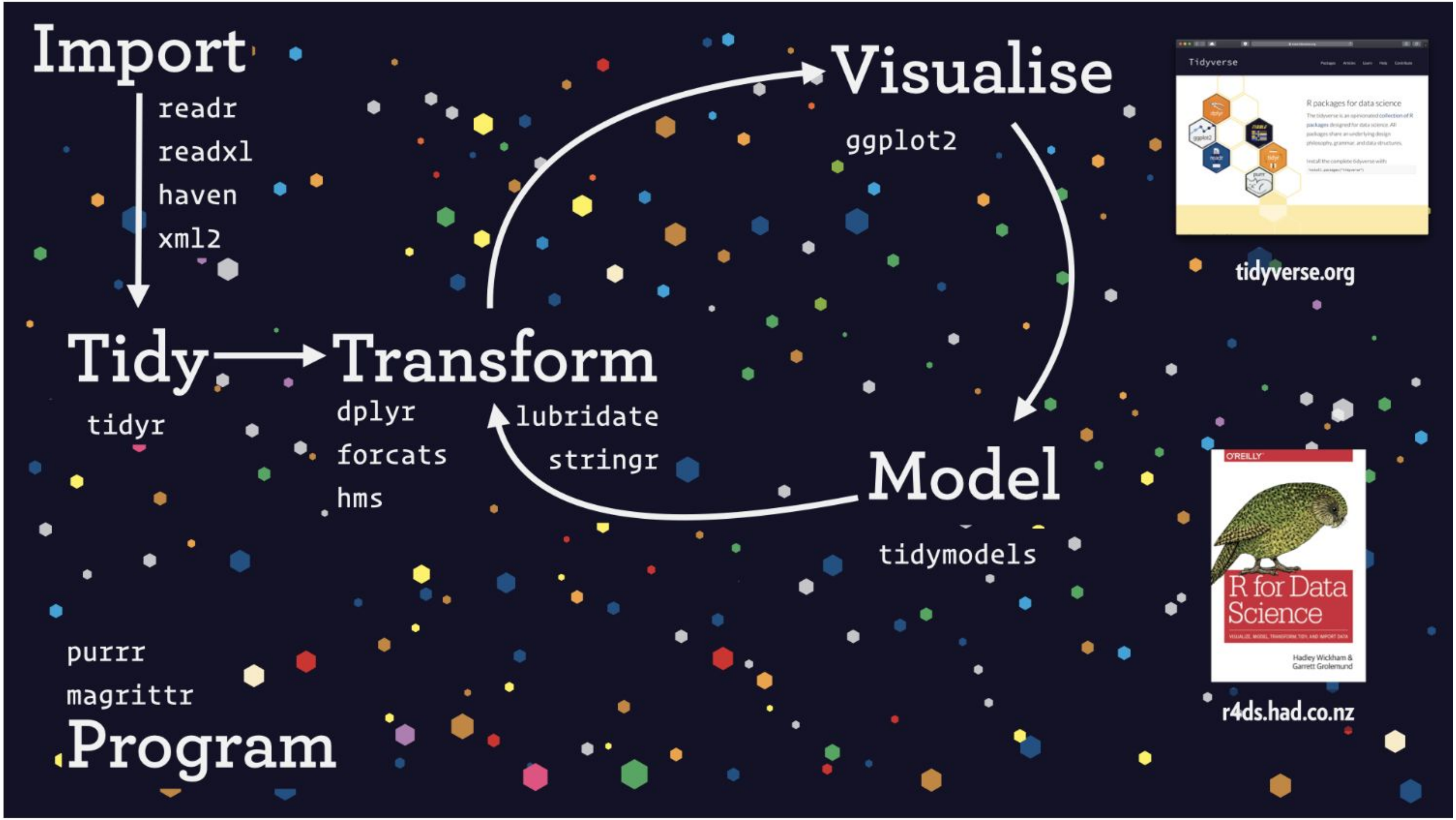
Packages designed for data science that make analysis more intuitive.

Super powerful for wrangling and data viz.

History of the tidyverse

"The tidyverse is just one way to get the job done and I don't think it's wrong to use other tools (indeed, it's usually not possible to do an analysis using only the tidyverse)."





Import

readr
readxl
haven
xml2

Tidy

tidyr

Transform

dplyr
forcats
hms

lubridate
stringr

Visualise

ggplot2

Model

tidymodels

Program

purrr
magrittr

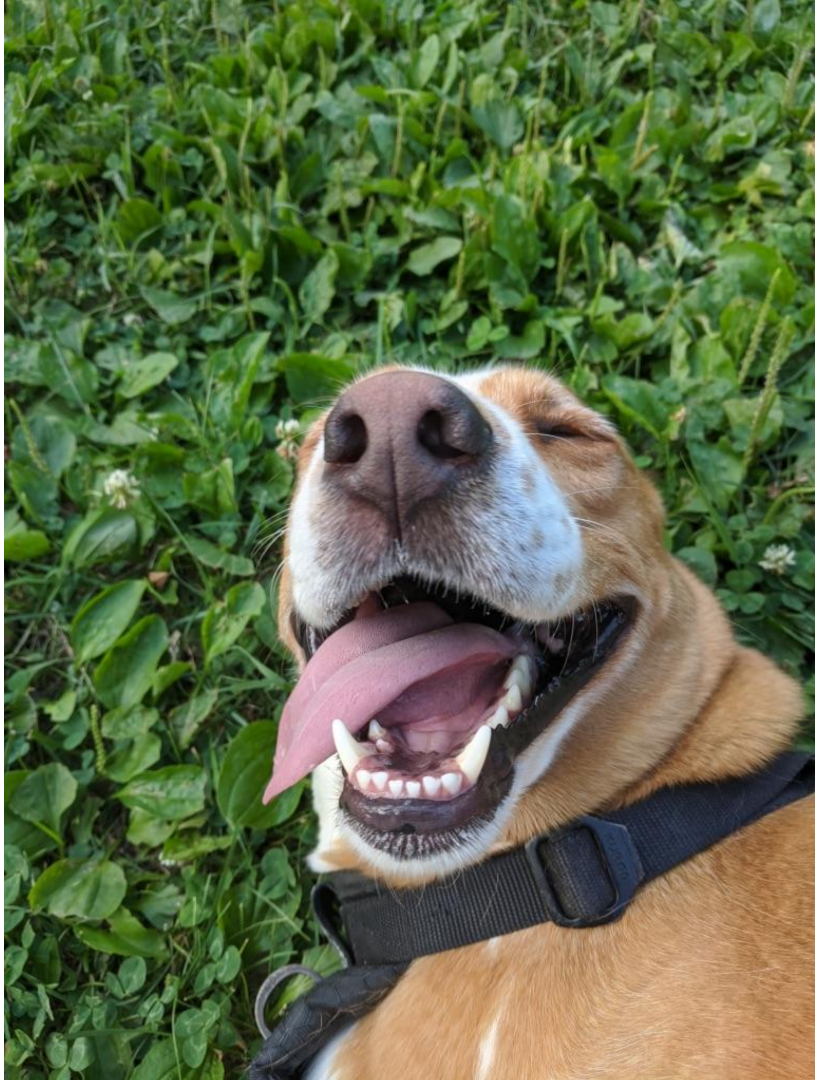
tidyverse.org

r4ds.had.co.nz

https://hadley.github.io/25-tidyverse-history/

How many people feel about data wrangling

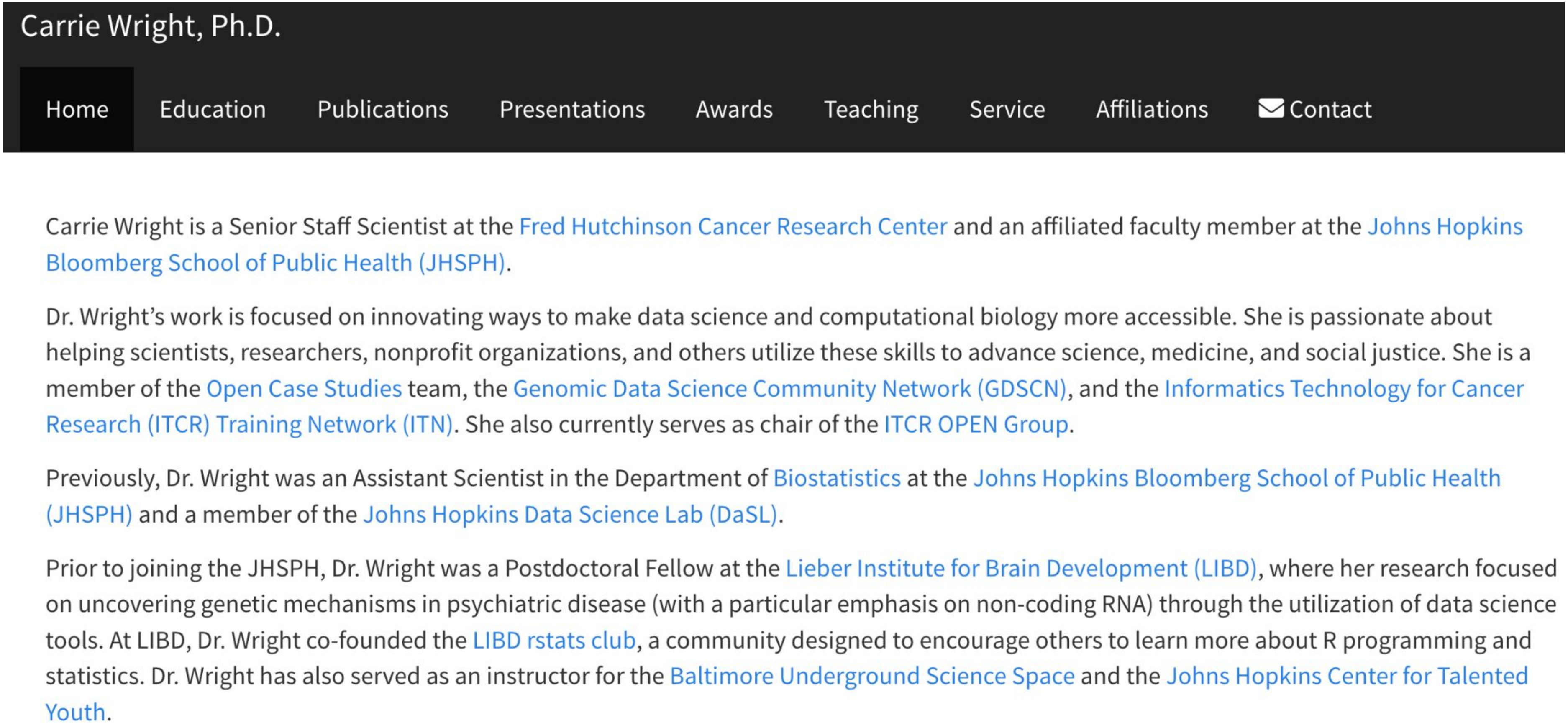
How we feel about data wrangling



About us

Carrie

Ava



Carrie Wright, Ph.D.

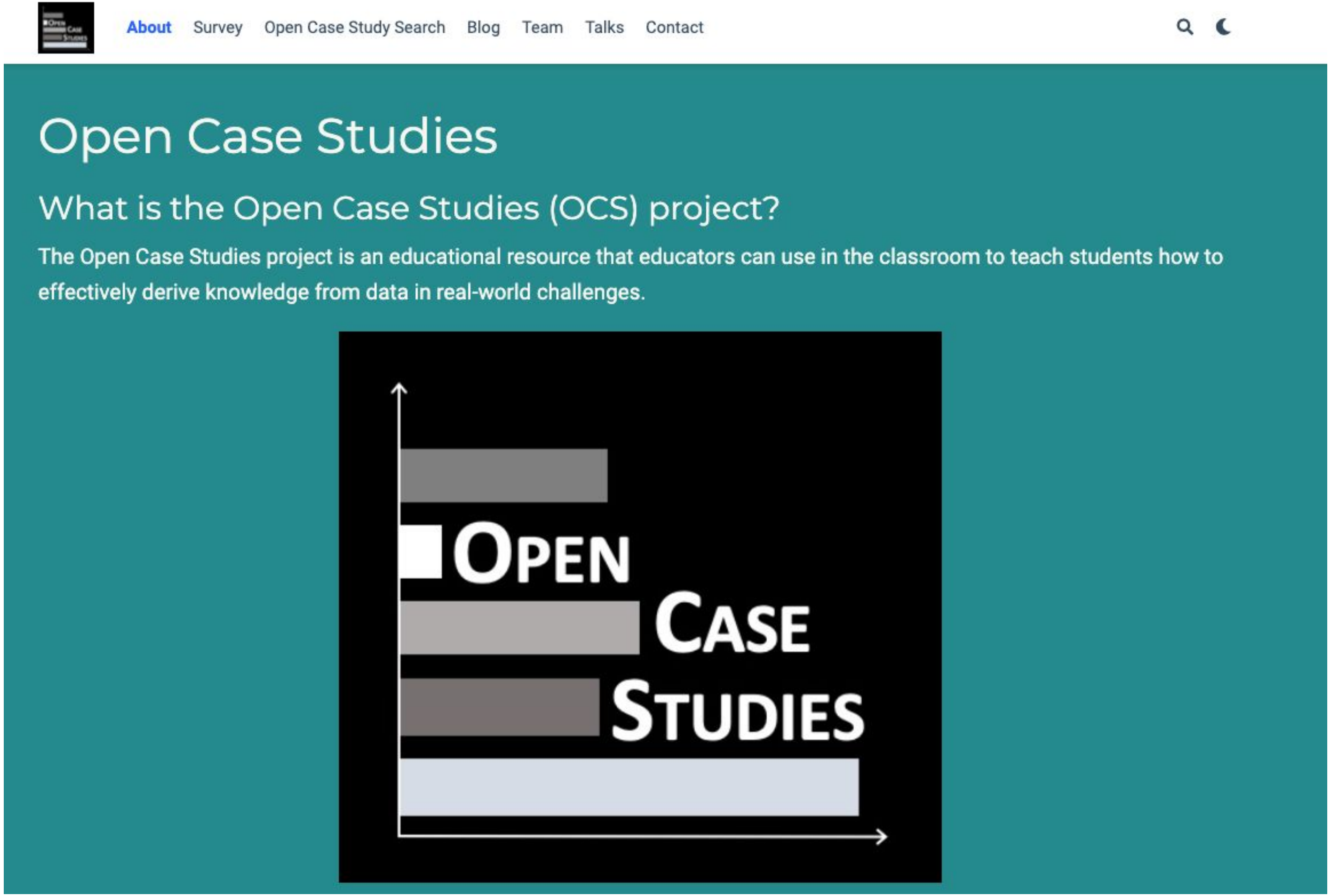
Carrie Wright is a Senior Staff Scientist at the Fred Hutchinson Cancer Research Center and an affiliated faculty member at the Johns Hopkins Bloomberg School of Public Health (JHSPH).

Dr. Wright's work is focused on innovating ways to make data science and computational biology more accessible. She is passionate about helping scientists, researchers, nonprofit organizations, and others utilize these skills to advance science, medicine, and social justice. She is a member of the Open Case Studies team, the Genomic Data Science Community Network (GDSCN), and the Informatics Technology for Cancer Research (ITCR) Training Network (ITN). She also currently serves as chair of the ITCR OPEN Group.

Previously, Dr. Wright was an Assistant Scientist in the Department of Biostatistics at the Johns Hopkins Bloomberg School of Public Health (JHSPH) and a member of the Johns Hopkins Data Science Lab (DaSL).

Prior to joining the JHSPH, Dr. Wright was a Postdoctoral Fellow at the Lieber Institute for Brain Development (LIBD), where her research focused on uncovering genetic mechanisms in psychiatric disease (with a particular emphasis on non-coding RNA) through the utilization of data science tools. At LIBD, Dr. Wright co-founded the LIBD rstats club, a community designed to encourage others to learn more about R programming and statistics. Dr. Wright has also served as an instructor for the Baltimore Underground Science Space and the Johns Hopkins Center for Talented Youth.

https://carriewright11.github.io/

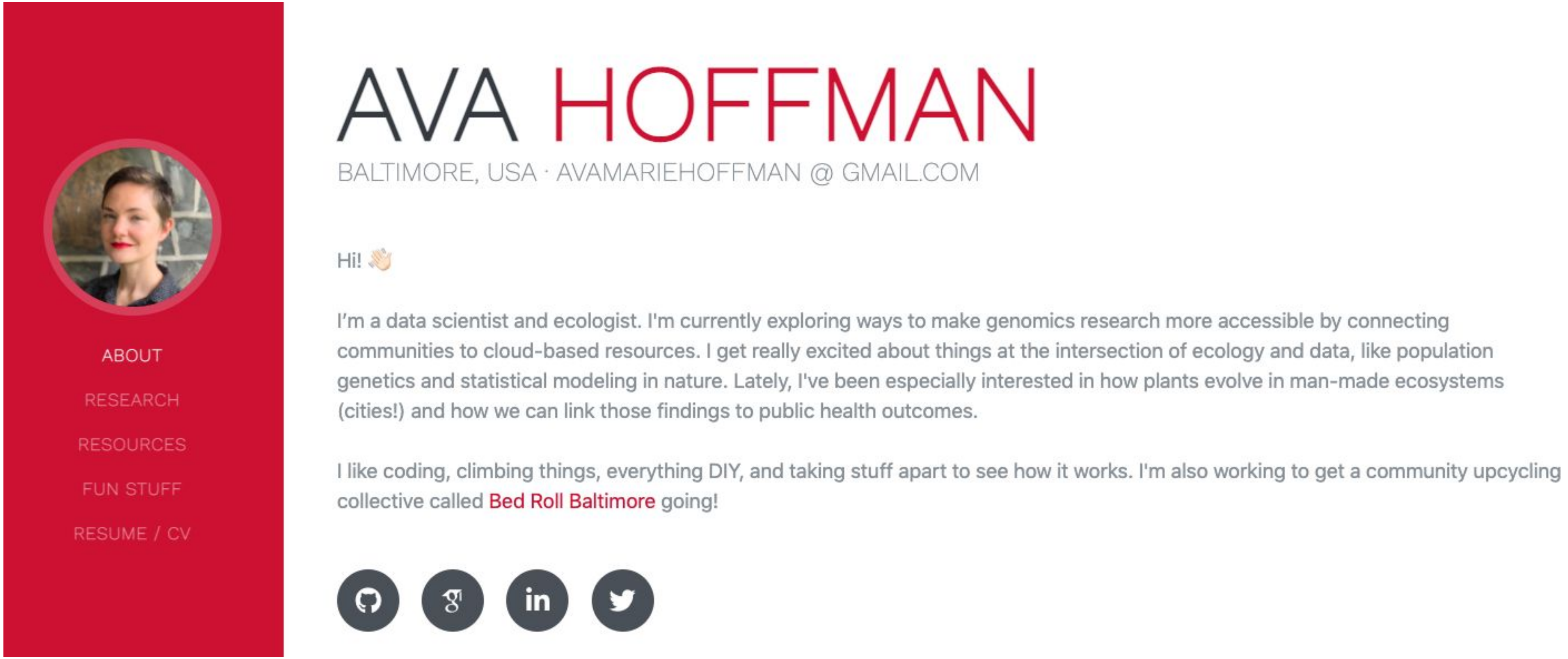


Open Case Studies

What is the Open Case Studies (OCS) project?

The Open Case Studies project is an educational resource that educators can use in the classroom to teach students how to effectively derive knowledge from data in real-world challenges.

https://www.opencasestudies.org/



AVA HOFFMAN

BALTIMORE, USA · AVAMARIEHOFFMAN @ GMAIL.COM

Hi! 👋🏻

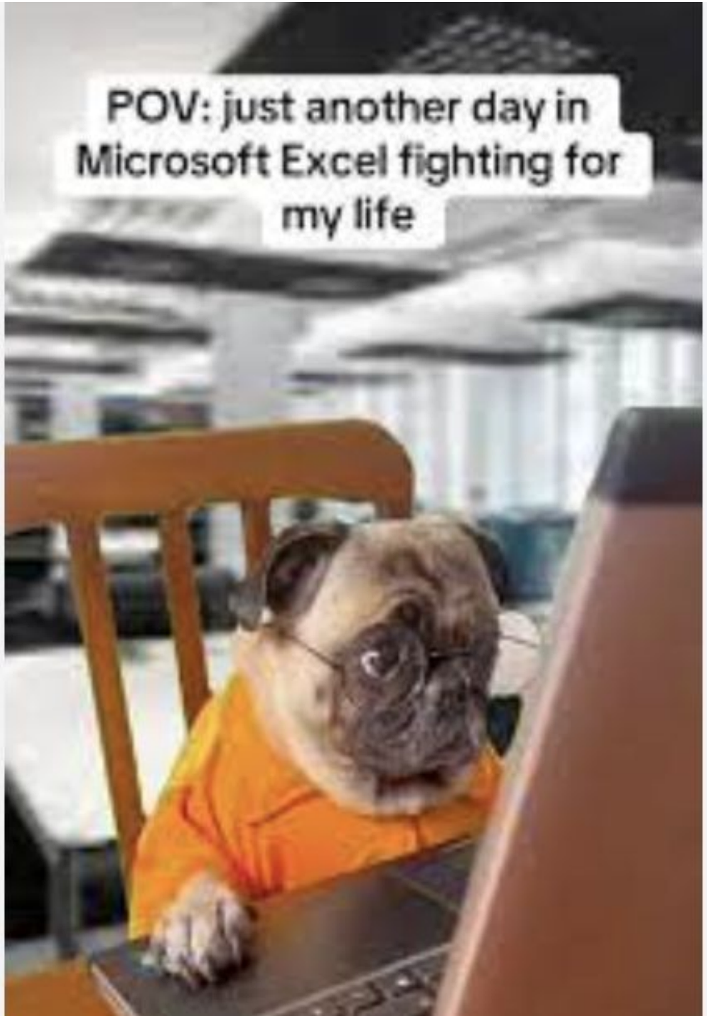
I'm a data scientist and ecologist. I'm currently exploring ways to make genomics research more accessible by connecting communities to cloud-based resources. I get really excited about things at the intersection of ecology and data, like population genetics and statistical modeling in nature. Lately, I've been especially interested in how plants evolve in man-made ecosystems (cities!) and how we can link those findings to public health outcomes.

I like coding, climbing things, everything DIY, and taking stuff apart to see how it works. I'm also working to get a community upcycling collective called Bed Roll Baltimore going!

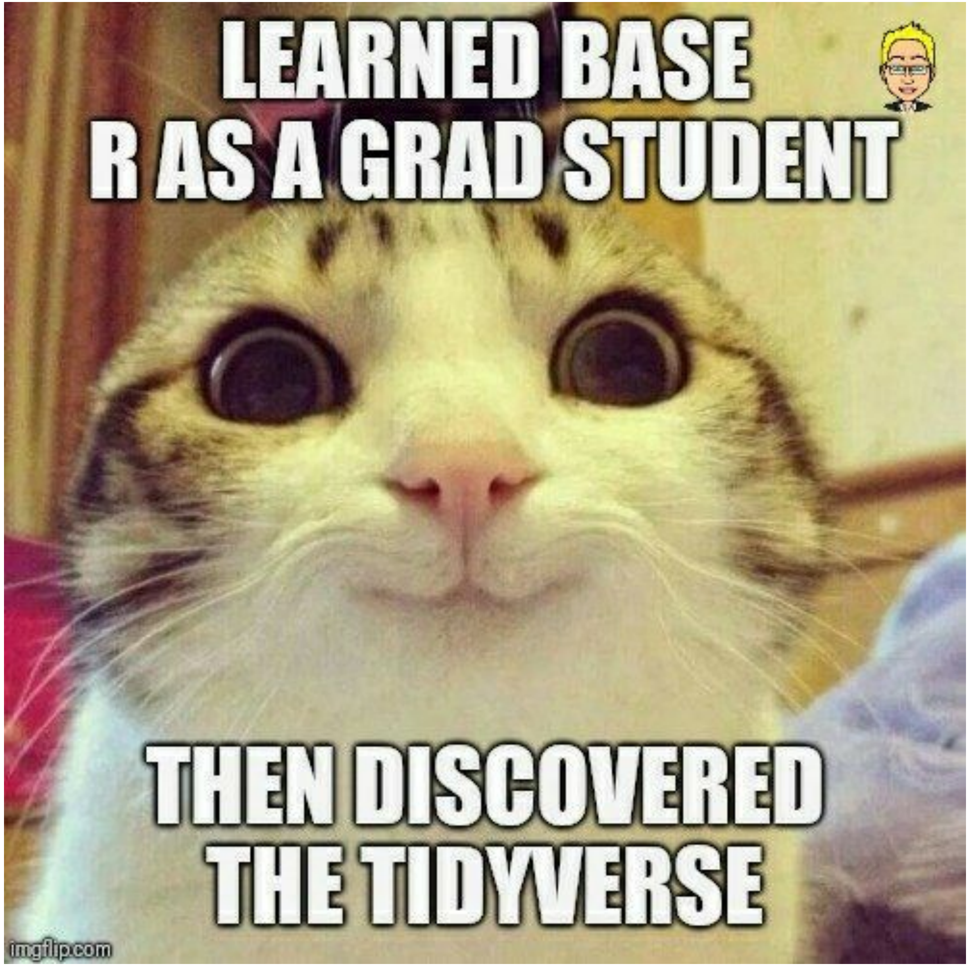
https://www.avahoffman.com/

Why this class

link



POV: just another day in Microsoft Excel fighting for my life



https://images.app.goo.gl/TDUz7aBxEHx6wonh8

But also…

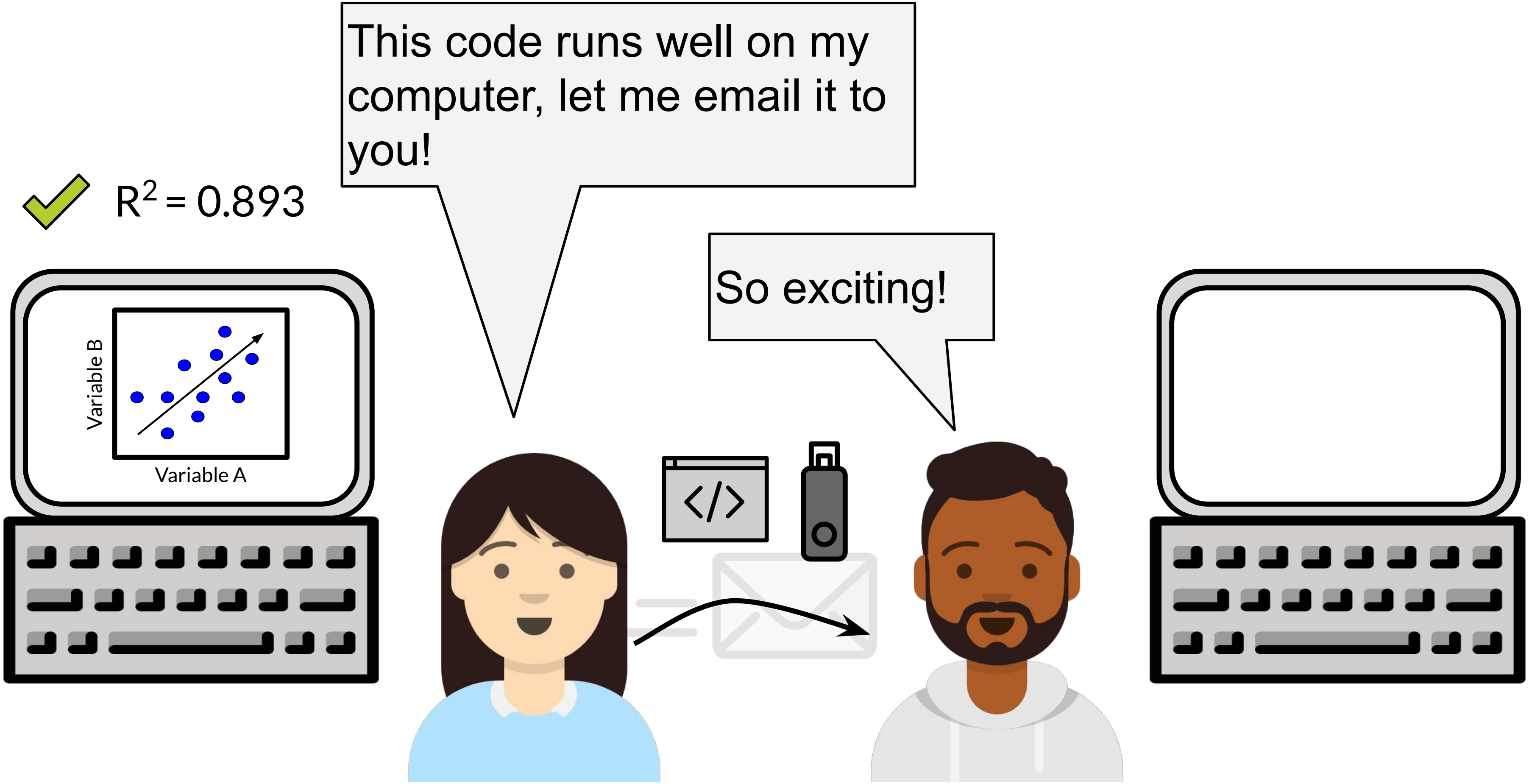


Image created by Candace Savonen using Avataars.

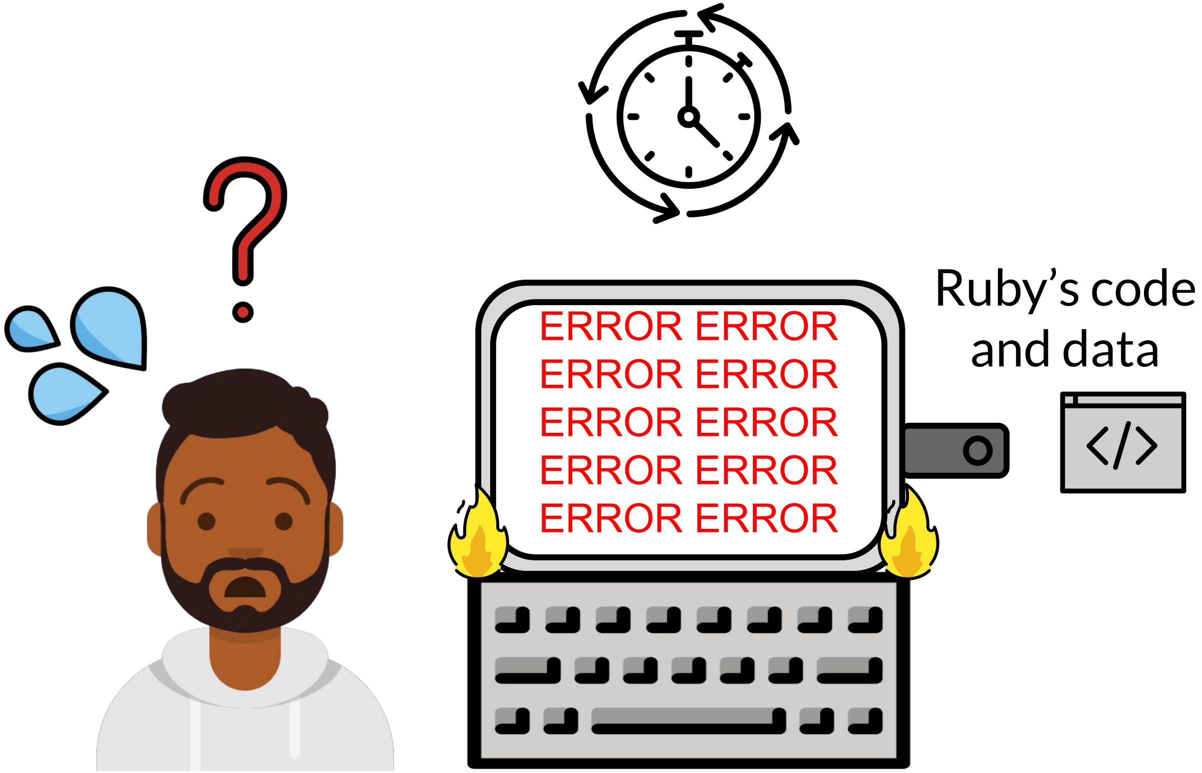
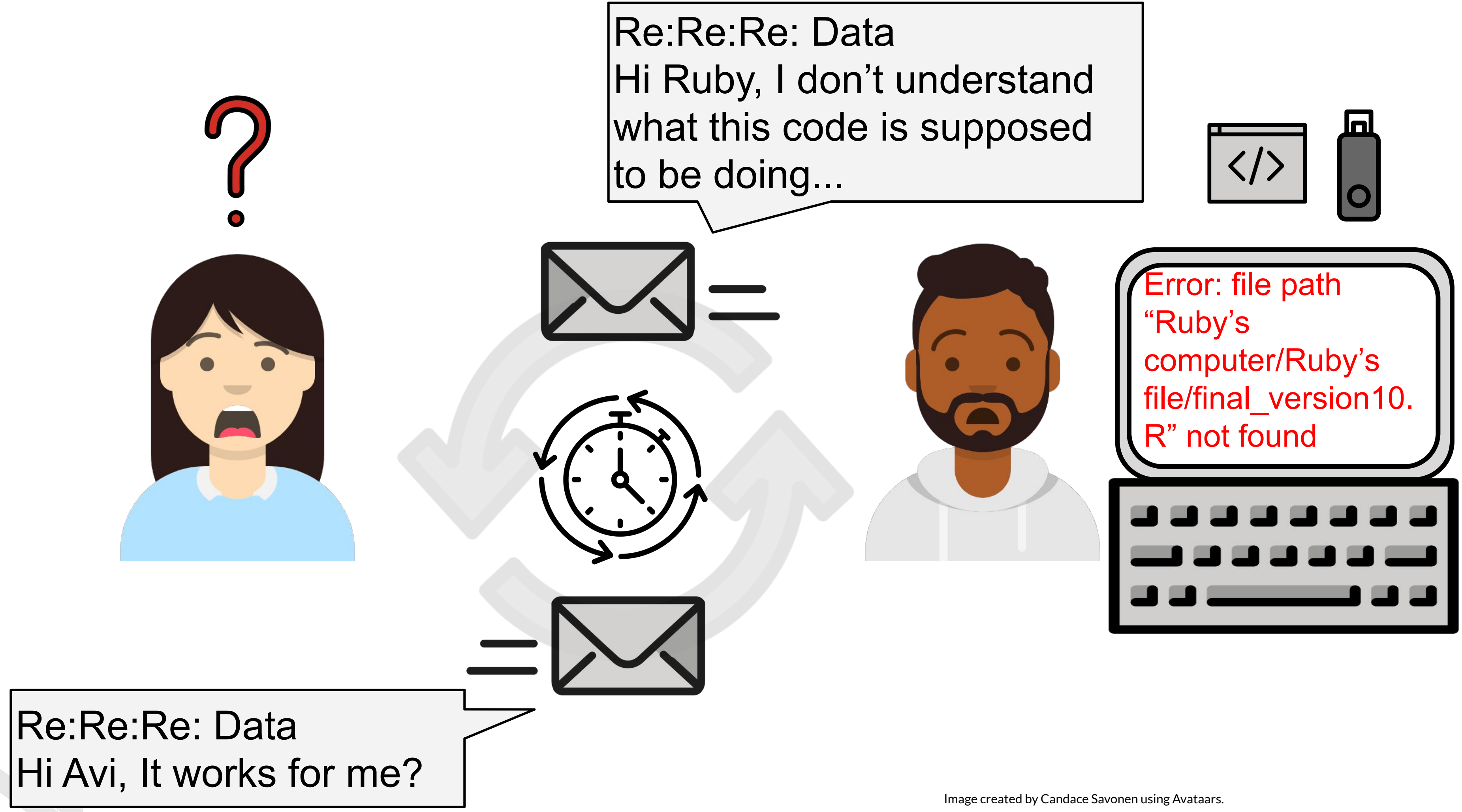


Image created by Candace Savonen using Avataars.

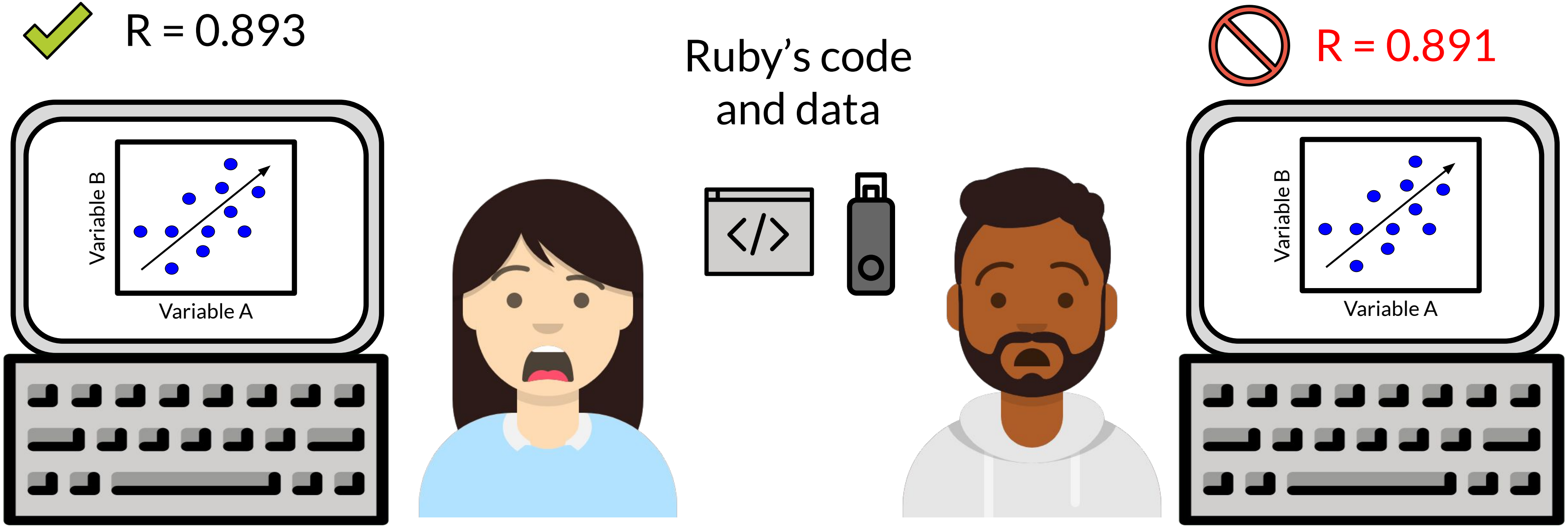


Re:Re:Re: Data
Hi Ruby, I don't understand what this code is supposed to be doing...

Error: file path "Ruby's computer/Ruby's file/final_version10.R" not found

Re:Re:Re: Data
Hi Avi, It works for me?

Image created by Candace Savonen using Avataars.

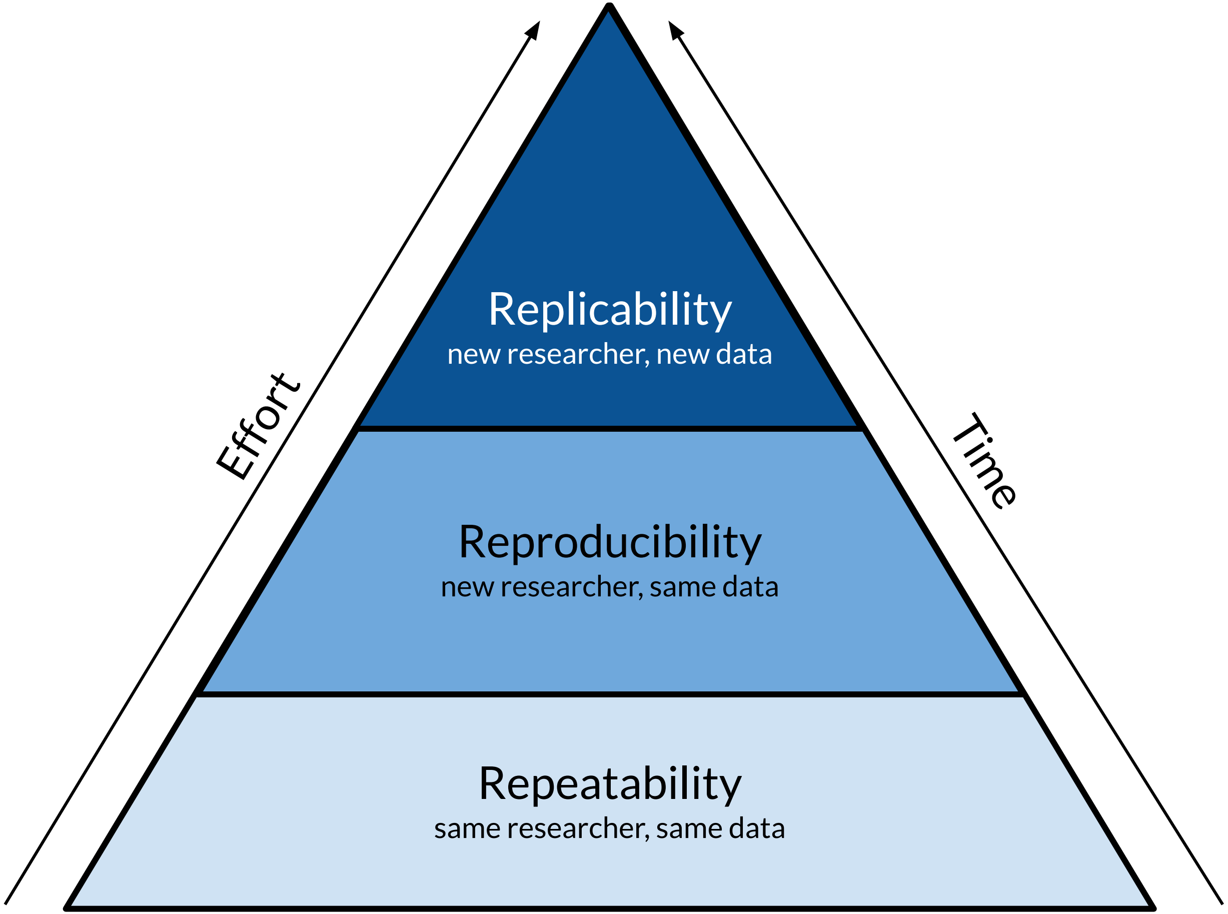


✓ R = 0.893

Ruby’s code and data

🛇 R = 0.891

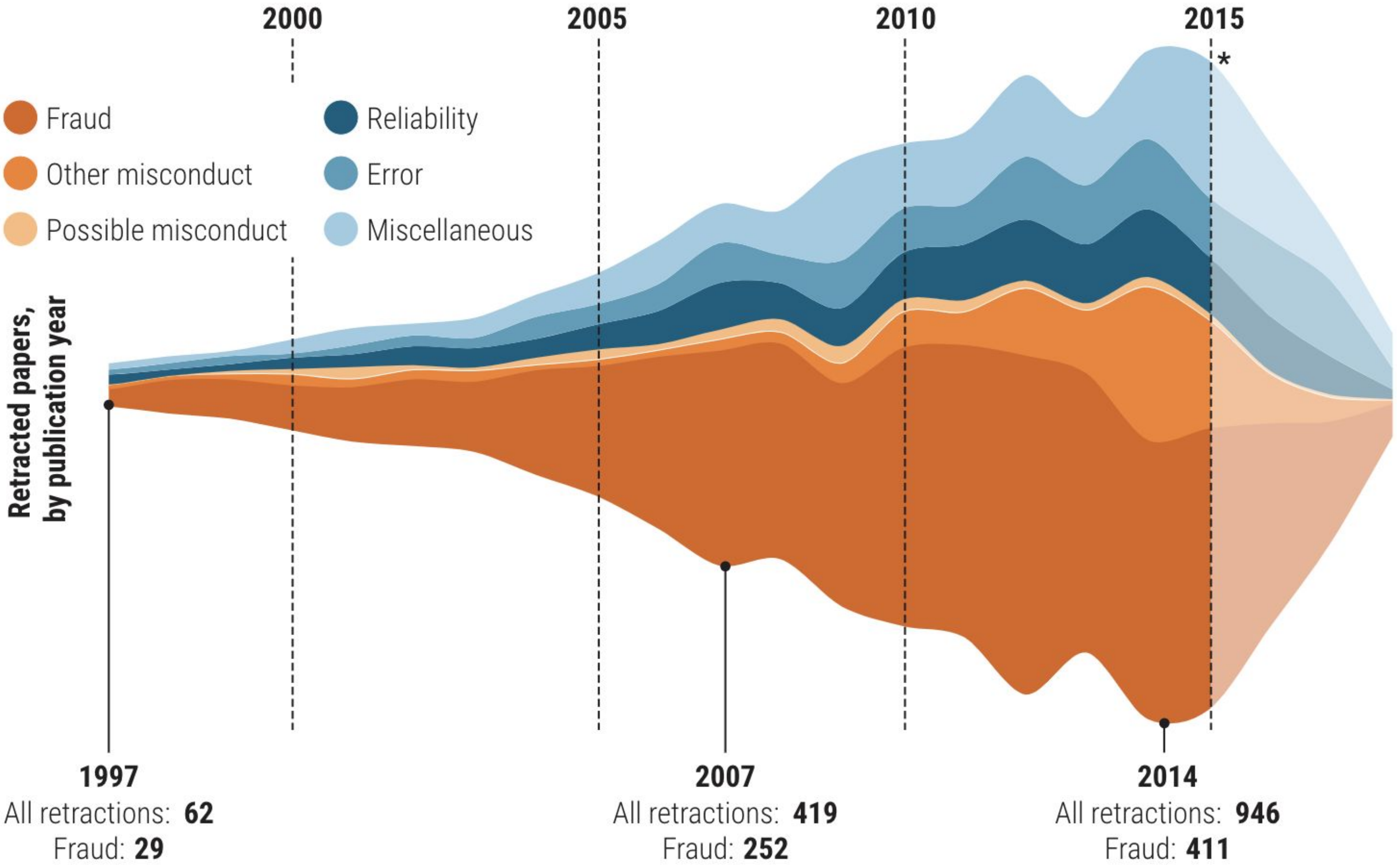
Image created by Candace Savonen using Avataars.



Based off of a figure from Essawy et al, 2020 https://doi.org/10.1016/j.envsoft.2020.104753

Data wrangling using more error-prone ways can lead to less transparent science

Transparent Data Wrangling is Required for Transparent Data Science



https://www.science.org/content/article/what-massive-database-retracted-papers-reveals-about-science-publishing-s-death-penalty

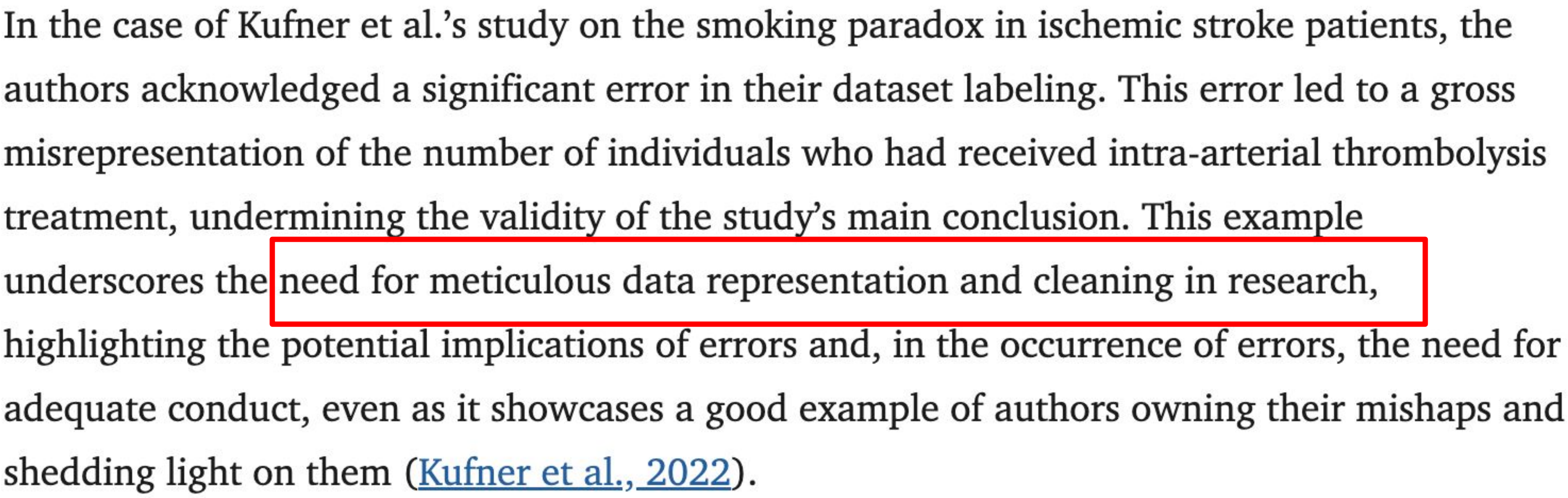
Dataset Error:

Data Errors are One of the Leading Causes of Retractions

https://pmc.ncbi.nlm.nih.gov/articles/PMC10485848/

Dataset Error:

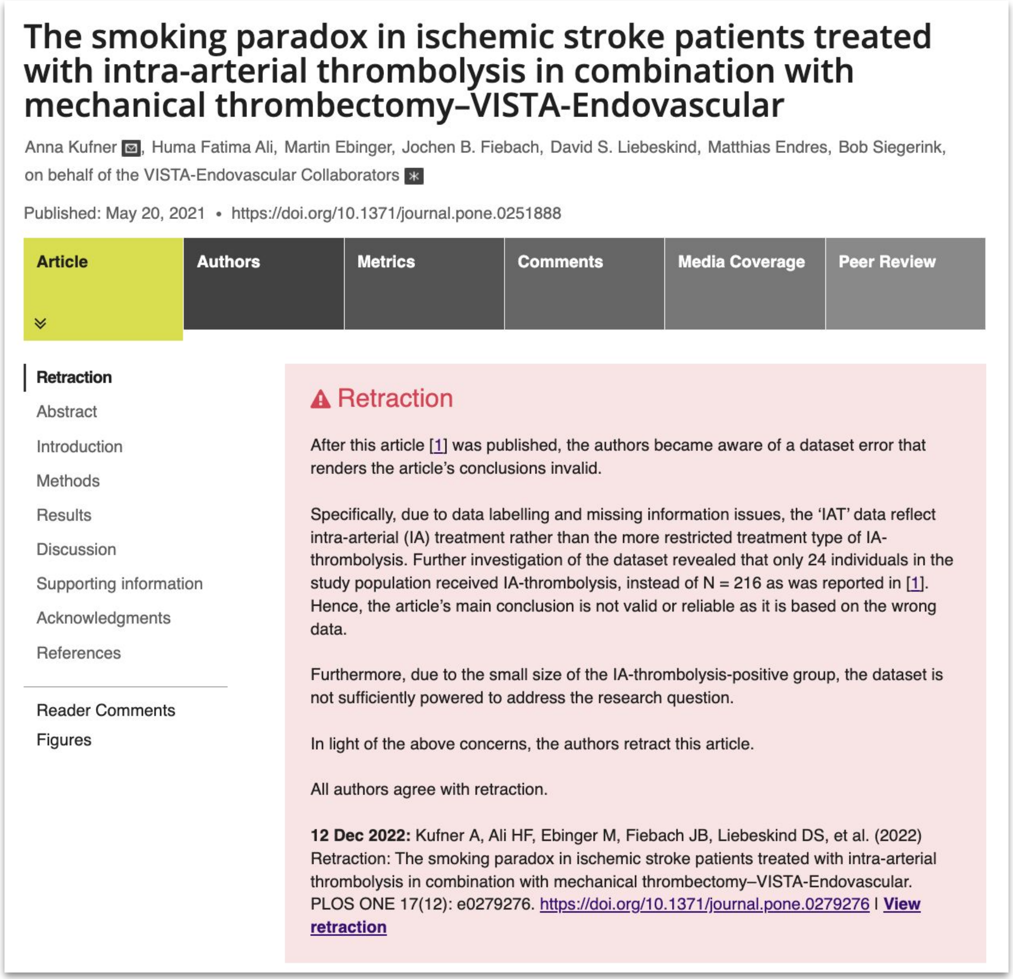
Data Errors are One of the Leading Causes of Retractions



In the case of Kufner et al.'s study on the smoking paradox in ischemic stroke patients, the authors acknowledged a significant error in their dataset labeling. This error led to a gross misrepresentation of the number of individuals who had received intra-arterial thrombolysis treatment, undermining the validity of the study's main conclusion. This example underscores the need for meticulous data representation and cleaning in research, highlighting the potential implications of errors and, in the occurrence of errors, the need for adequate conduct, even as it showcases a good example of authors owning their mishaps and shedding light on them (Kufner et al., 2022).

https://pmc.ncbi.nlm.nih.gov/articles/PMC10485848/

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0251888



The smoking paradox in ischemic stroke patients treated with intra-arterial thrombolysis in combination with mechanical thrombectomy–VISTA-Endovascular

Anna Kufner, Huma Fatima Ali, Martin Ebinger, Jochen B. Fiebach, David S. Liebeskind, Matthias Endres, Bob Siegerink, on behalf of the VISTA-Endovascular Collaborators

Published: May 20, 2021 • https://doi.org/10.1371/journal.pone.0251888

⚠ Retraction

After this article [1] was published, the authors became aware of a dataset error that renders the article's conclusions invalid.

Specifically, due to data labelling and missing information issues, the 'IAT' data reflect intra-arterial (IA) treatment rather than the more restricted treatment type of IA-thrombolysis. Further investigation of the dataset revealed that only 24 individuals in the study population received IA-thrombolysis, instead of N = 216 as was reported in [1]. Hence, the article's main conclusion is not valid or reliable as it is based on the wrong data.

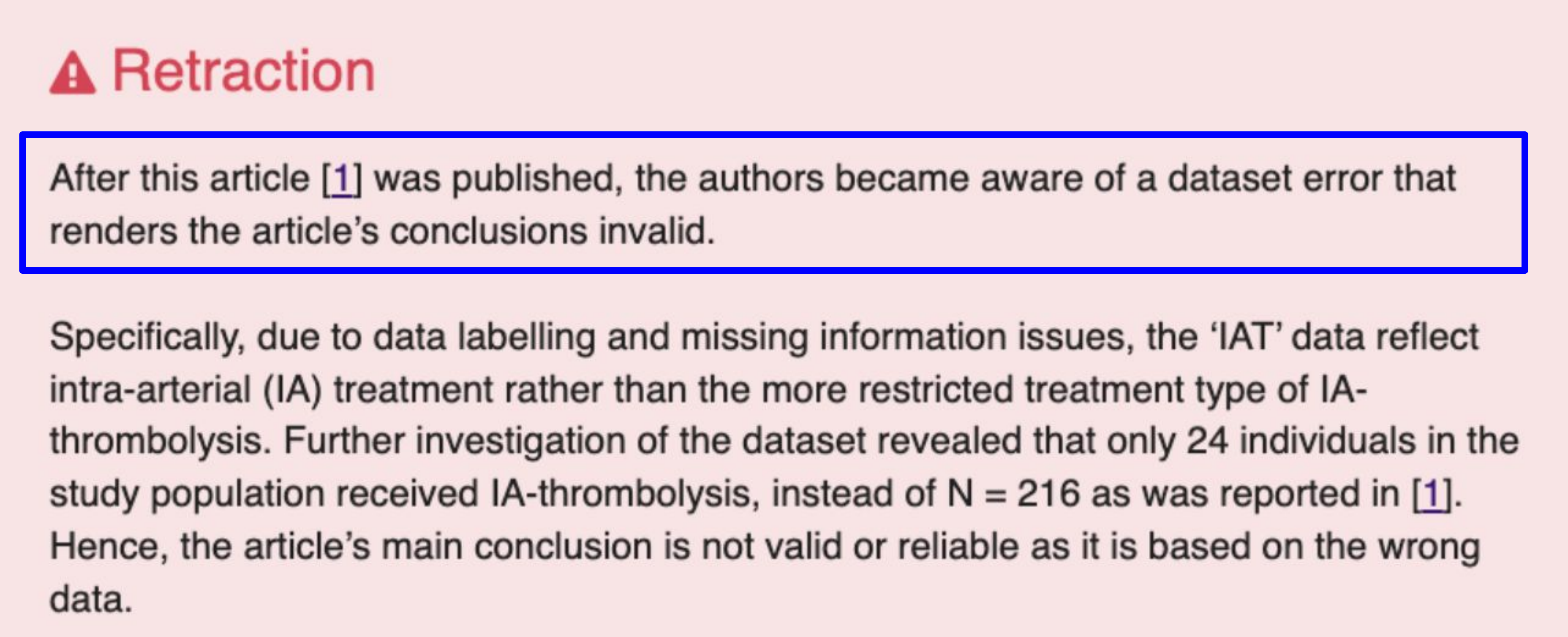
Furthermore, due to the small size of the IA-thrombolysis-positive group, the dataset is not sufficiently powered to address the research question.

In light of the above concerns, the authors retract this article.

All authors agree with retraction.

12 Dec 2022: Kufner A, Ali HF, Ebinger M, Fiebach JB, Liebeskind DS, et al. (2022) Retraction: The smoking paradox in ischemic stroke patients treated with intra-arterial thrombolysis in combination with mechanical thrombectomy–VISTA-Endovascular. PLOS ONE 17(12): e0279276. https://doi.org/10.1371/journal.pone.0279276 | **View retraction**

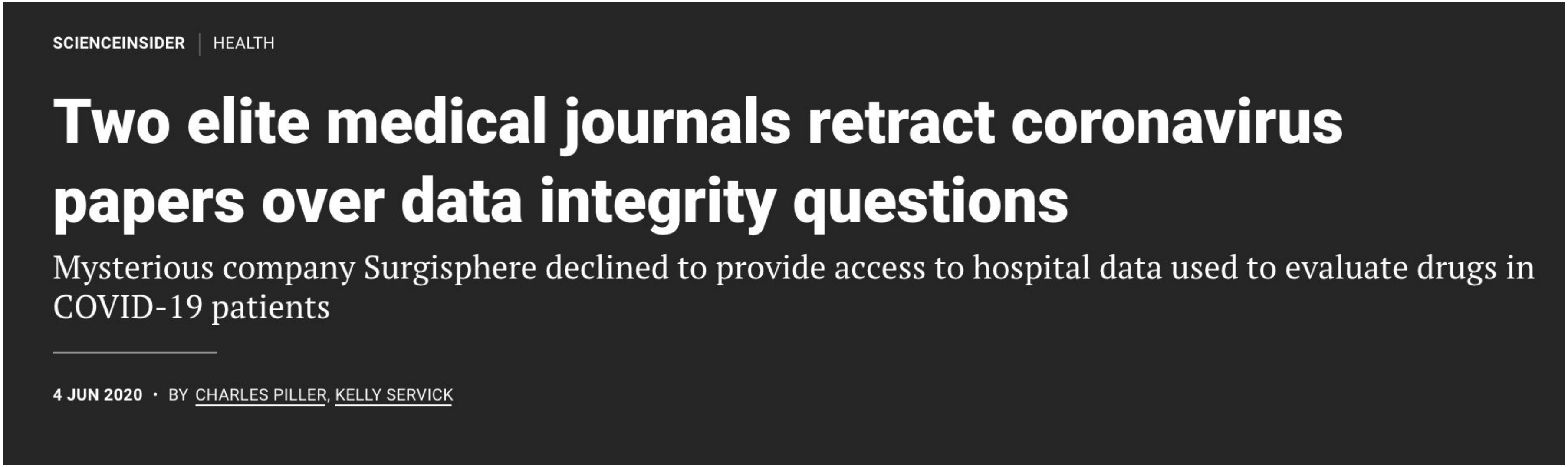
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0251888



⚠ Retraction

After this article [1] was published, the authors became aware of a dataset error that renders the article's conclusions invalid.

Specifically, due to data labelling and missing information issues, the 'IAT' data reflect intra-arterial (IA) treatment rather than the more restricted treatment type of IA-thrombolysis. Further investigation of the dataset revealed that only 24 individuals in the study population received IA-thrombolysis, instead of N = 216 as was reported in [1]. Hence, the article's main conclusion is not valid or reliable as it is based on the wrong data.



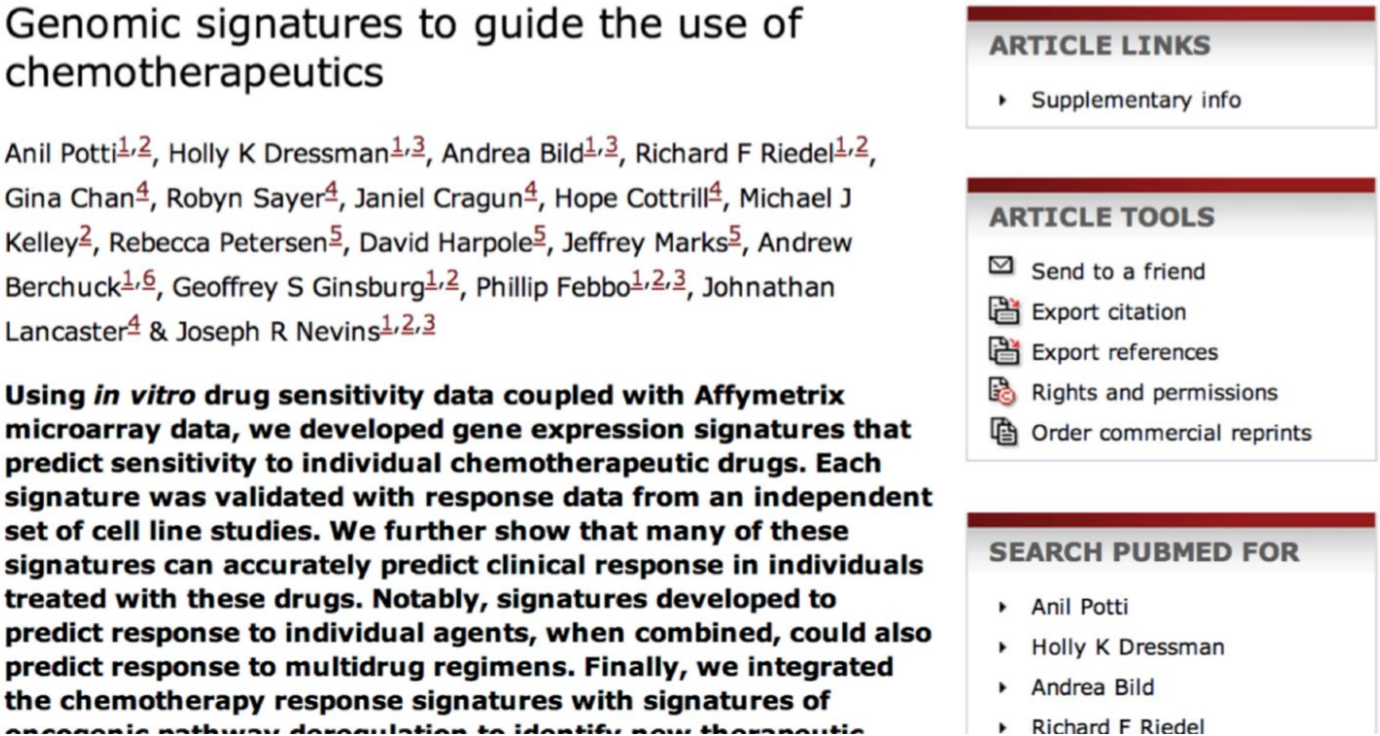
SCIENCEINSIDER | HEALTH

Two elite medical journals retract coronavirus papers over data integrity questions

Mysterious company Surgisphere declined to provide access to hospital data used to evaluate drugs in COVID-19 patients

4 JUN 2020 • BY CHARLES PILLER, KELLY SERVICK

Older example - but of great consequence



Genomic signatures to guide the use of chemotherapeutics

Anil Potti[1,2], Holly K Dressman[1,3], Andrea Bild[1,3], Richard F Riedel[1,2], Gina Chan[4], Robyn Sayer[4], Janiel Cragun[4], Hope Cottrill[4], Michael J Kelley[2], Rebecca Petersen[5], David Harpole[5], Jeffrey Marks[5], Andrew Berchuck[1,6], Geoffrey S Ginsburg[1,2], Phillip Febbo[1,2,3], Johnathan Lancaster[4] & Joseph R Nevins[1,2,3]

Using *in vitro* drug sensitivity data coupled with Affymetrix microarray data, we developed gene expression signatures that predict sensitivity to individual chemotherapeutic drugs. Each signature was validated with response data from an independent set of cell line studies. We further show that many of these signatures can accurately predict clinical response in individuals treated with these drugs. Notably, signatures developed to predict response to individual agents, when combined, could also predict response to multidrug regimens. Finally, we integrated the chemotherapy response signatures with signatures of oncogenic pathway deregulation to identify new therapeutic

ARTICLE LINKS

- Supplementary info

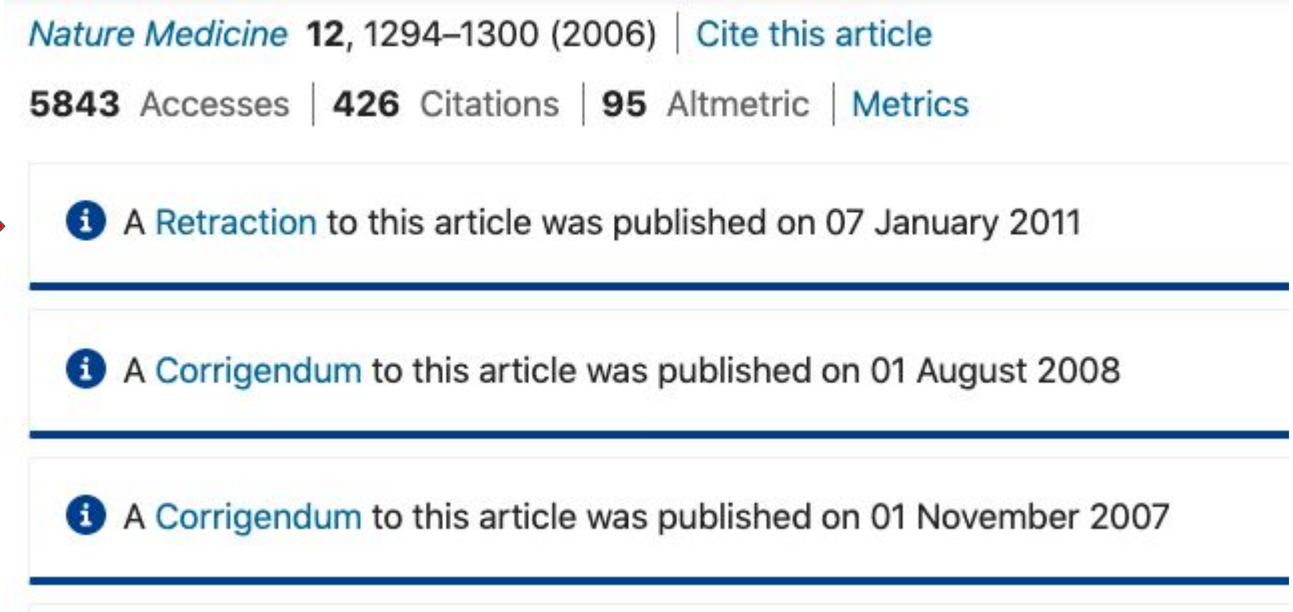
ARTICLE TOOLS

- Send to a friend
- Export citation
- Export references
- Rights and permissions
- Order commercial reprints

SEARCH PUBMED FOR

- Anil Potti
- Holly K Dressman
- Andrea Bild
- Richard F Riedel

https://doi.org/10.1038/nm1491



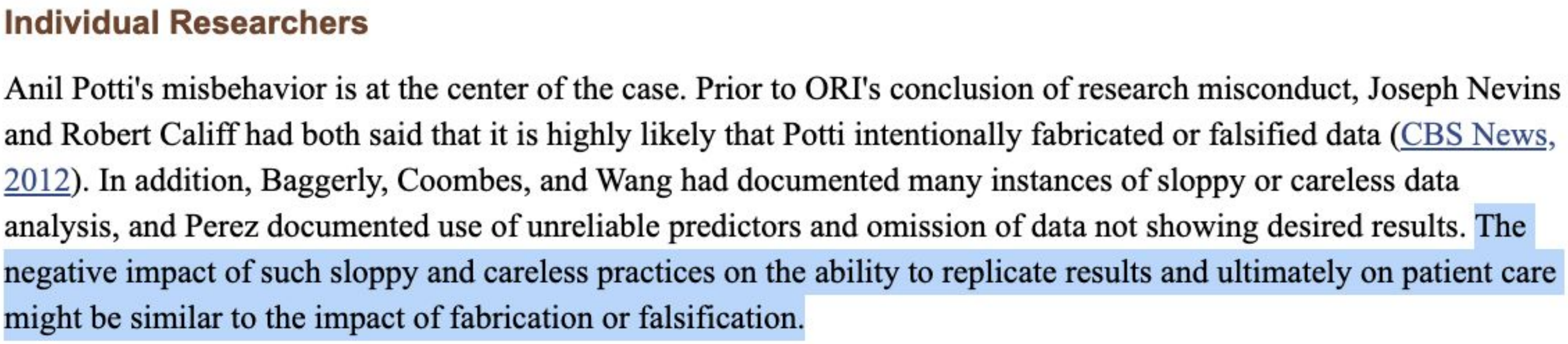
Nature Medicine **12**, 1294–1300 (2006) | Cite this article

5843 Accesses | **426** Citations | **95** Altmetric | Metrics

ℹ A Retraction to this article was published on 07 January 2011

ℹ A Corrigendum to this article was published on 01 August 2008

ℹ A Corrigendum to this article was published on 01 November 2007



Individual Researchers

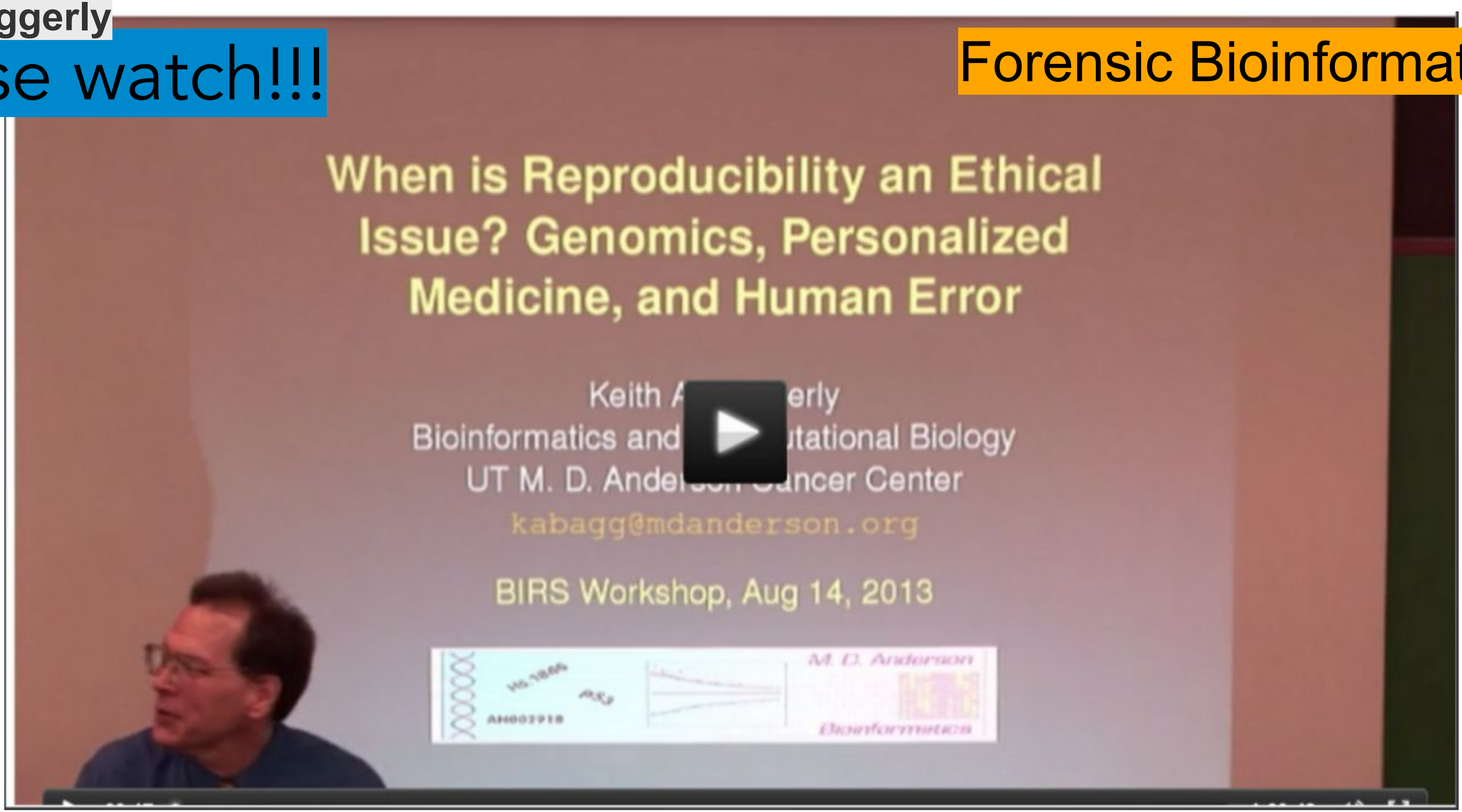
Anil Potti's misbehavior is at the center of the case. Prior to ORI's conclusion of research misconduct, Joseph Nevins and Robert Califf had both said that it is highly likely that Potti intentionally fabricated or falsified data (CBS News, 2012). In addition, Baggerly, Coombes, and Wang had documented many instances of sloppy or careless data analysis, and Perez documented use of unreliable predictors and omission of data not showing desired results. The negative impact of such sloppy and careless practices on the ability to replicate results and ultimately on patient care might be similar to the impact of fabrication or falsification.

https://www.ncbi.nlm.nih.gov/books/NBK475955/

Keith Baggerly

Please watch!!!

Forensic Bioinformatics



When is Reproducibility an Ethical Issue? Genomics, Personalized Medicine, and Human Error

Keith A erly

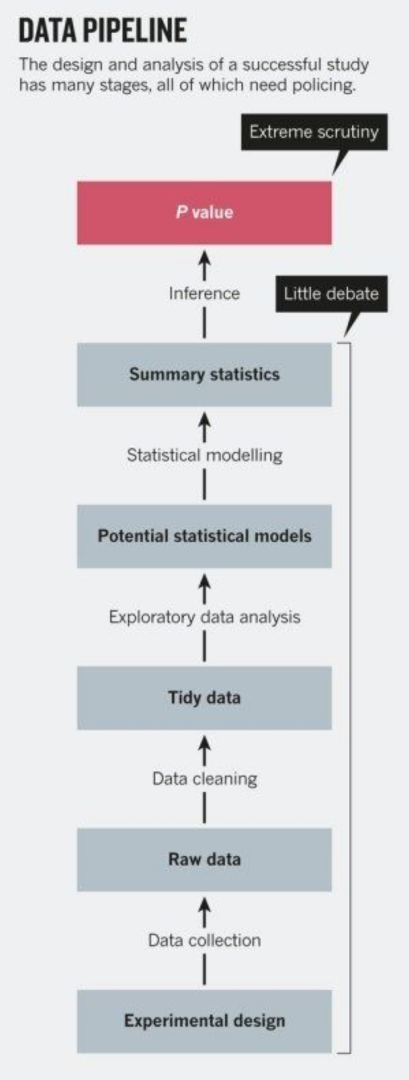
Bioinformatics and utational Biology

UT M. D. Ande ancer Center

kabagg@mdanderson.org

BIRS Workshop, Aug 14, 2013

http://www.birs.ca/events/2013/5-day-workshops/13w5083/videos/watch/201308141121-Baggerly.mp4



- Most of the attention is on the last step
- This course is about all the steps that come before
- They are *critical* for getting things rights

Leek and Peng Nature 2015

Reproducibility/Transparency is Important

Edited by

Roger J Davis et al.

Reproducibility Project: Cancer Biology

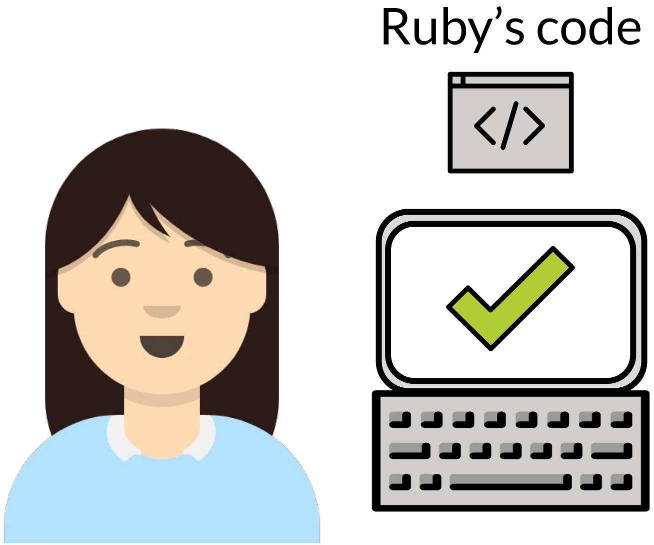
Collection • Dec 10, 2014

“We report the challenges confronted during a large-scale effort to replicate findings in cancer biology, and describe how improving **transparency and sharing** can make it easier to assess rigor and replicability and, therefore, to **increase research efficiency.**”

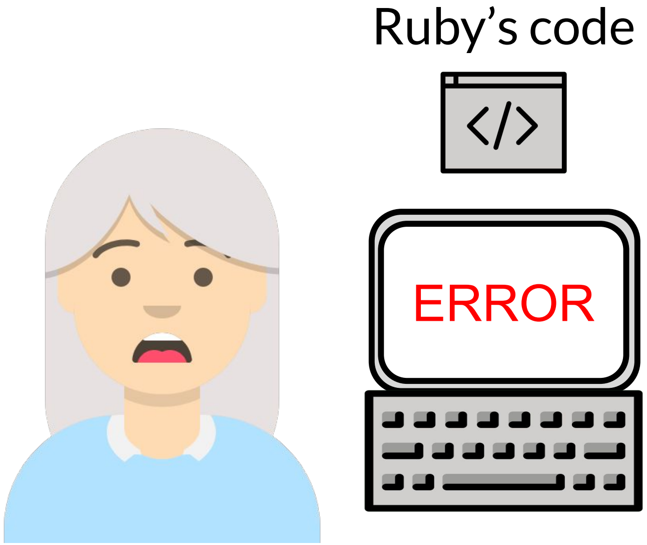
https://elifesciences.org/articles/67995

Reproducibility saves everyone time and effort!

Now Ruby



Future Ruby



It saves your future self time and effort!

Image created by Candace Savonen using Avataars.

Your closest collaborator is you six months ago, but you don’t reply to emails

- Karl Broman

http://kbroman.org/Tools4RR/assets/lectures/06_org_eda.pdf

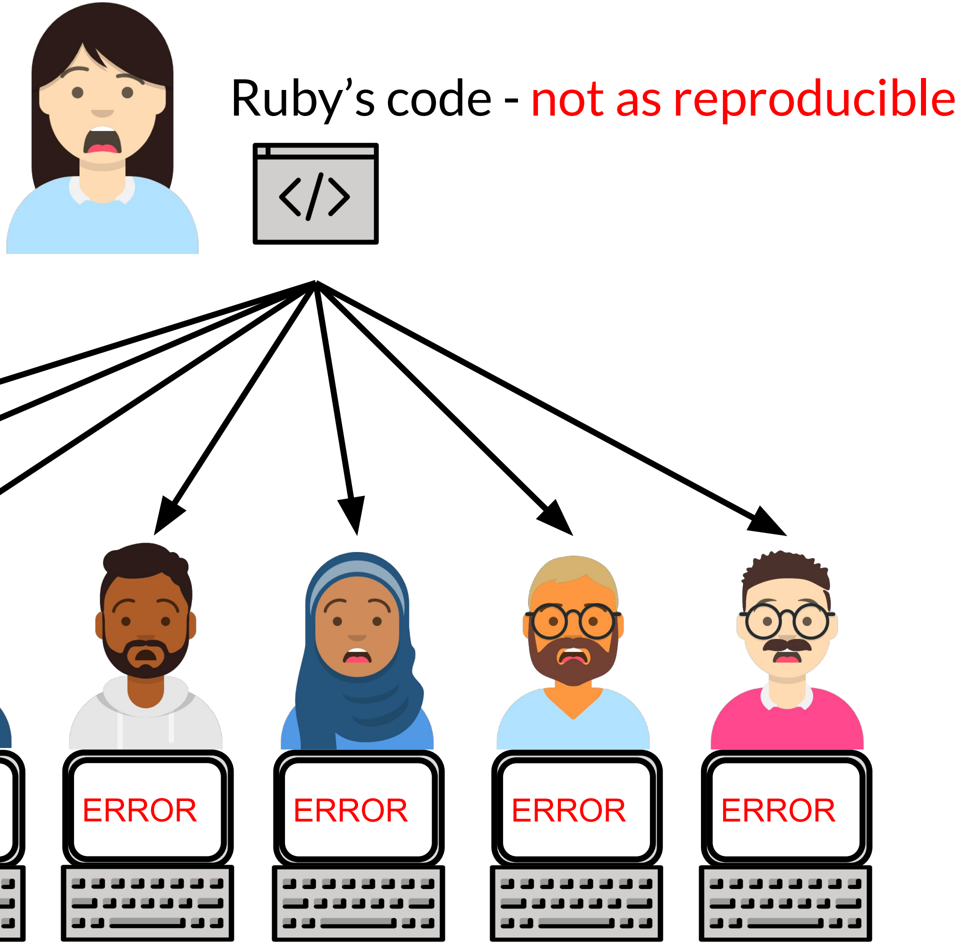


Image created by Candace Savonen using Avataars.

It saves the time & effort of others!

Ruby's code - more reproducible

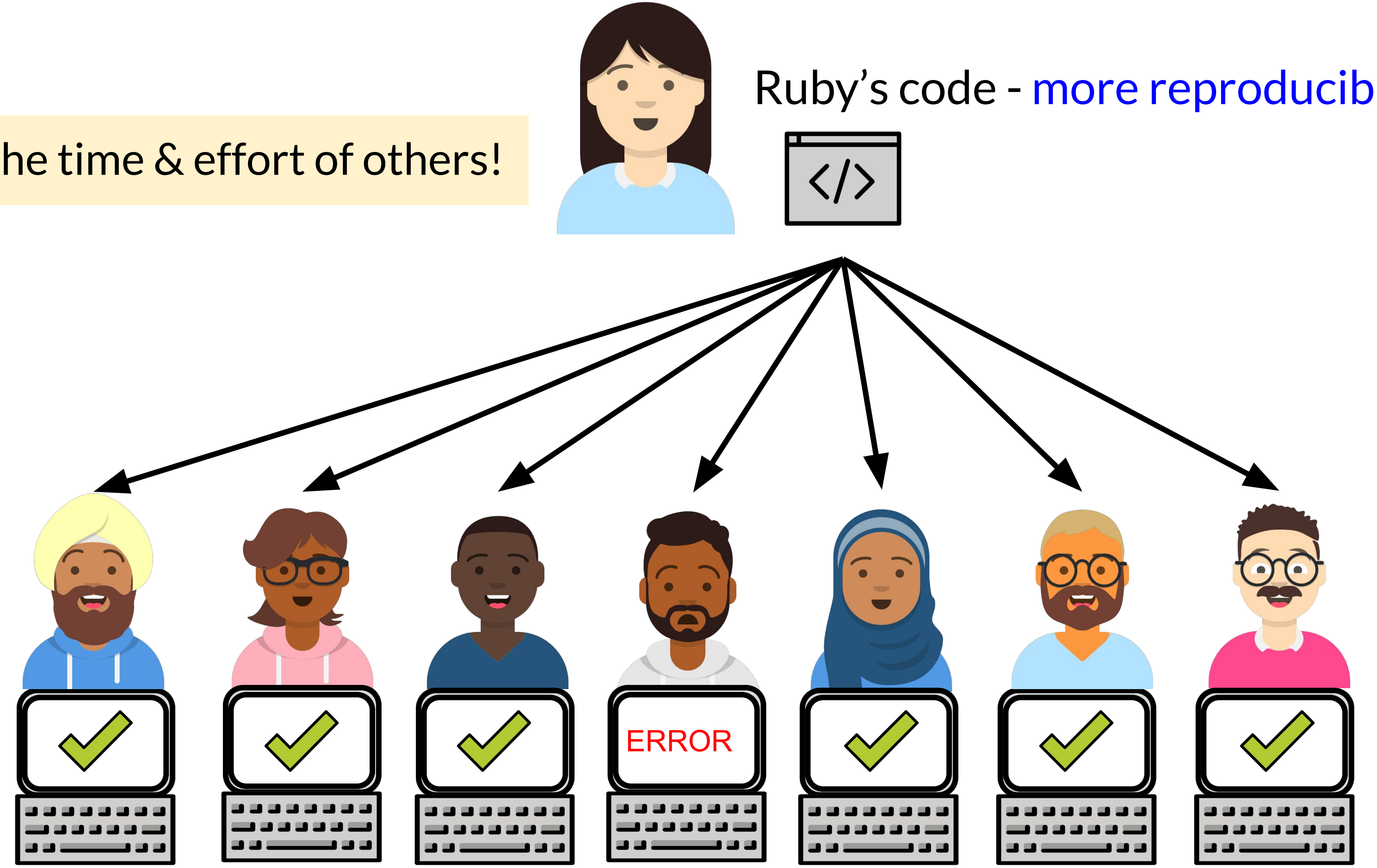
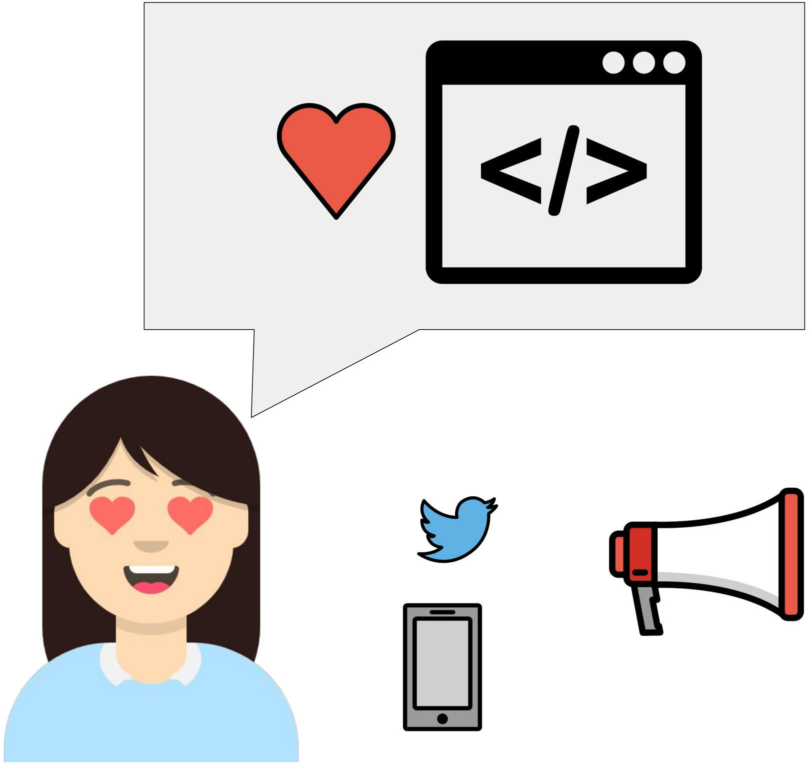


Image created by Candace Savonen using Avataars.

A useable, well-documented analysis is more likely to be used and disseminated!



It improves trust and reuse!

⚠️ Reproducibility != Correctness ⚠️

Reproducibility ~ Consistency

You could be consistently wrong in the same way....

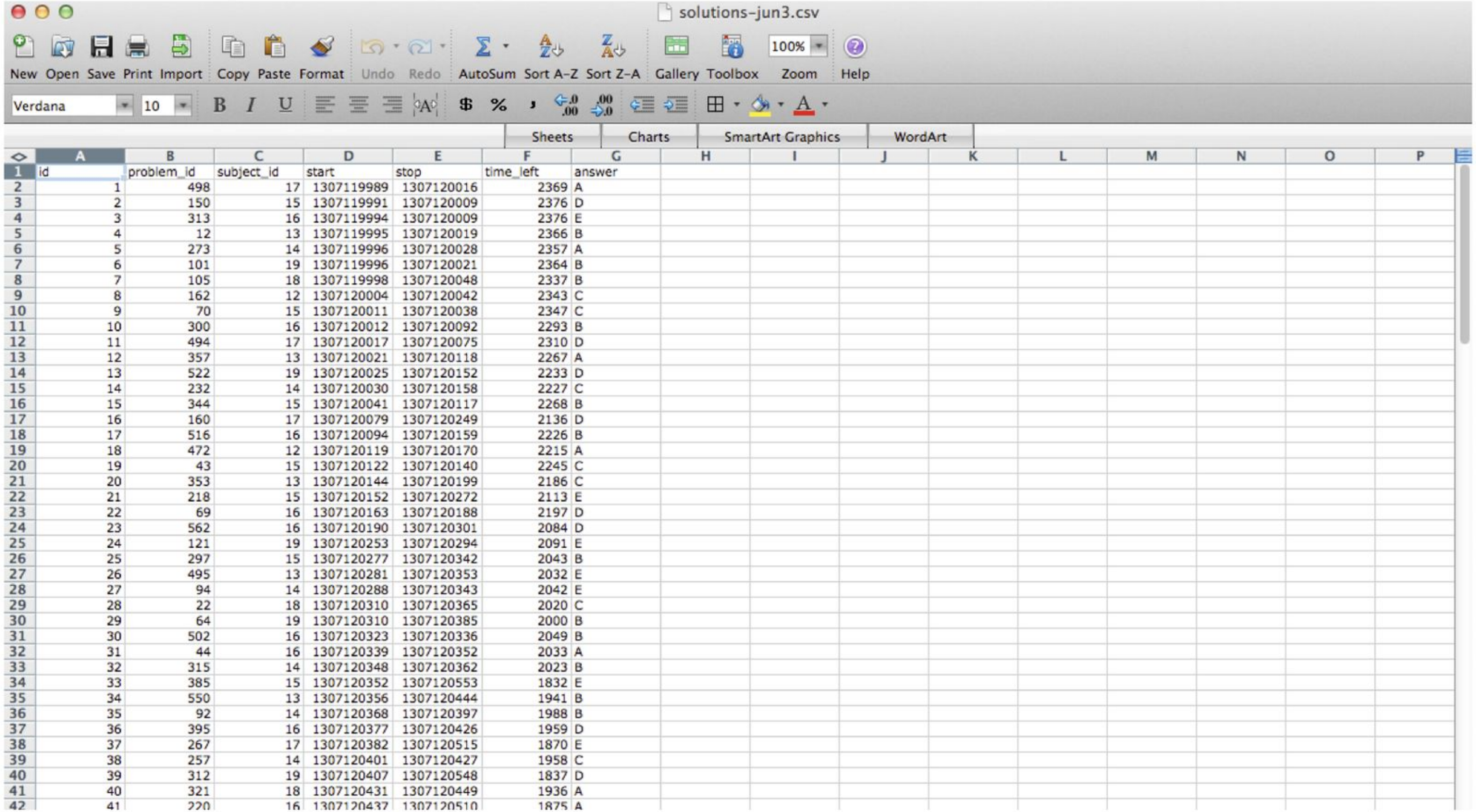
However, being consistent and transparent is a necessary step for doing trustworthy science.

It makes it easier for you and others to determine if your work was correct.

Reproducible work is typically easier to update!

Reviewer 2:
Please redo your full analysis but change this one detail.

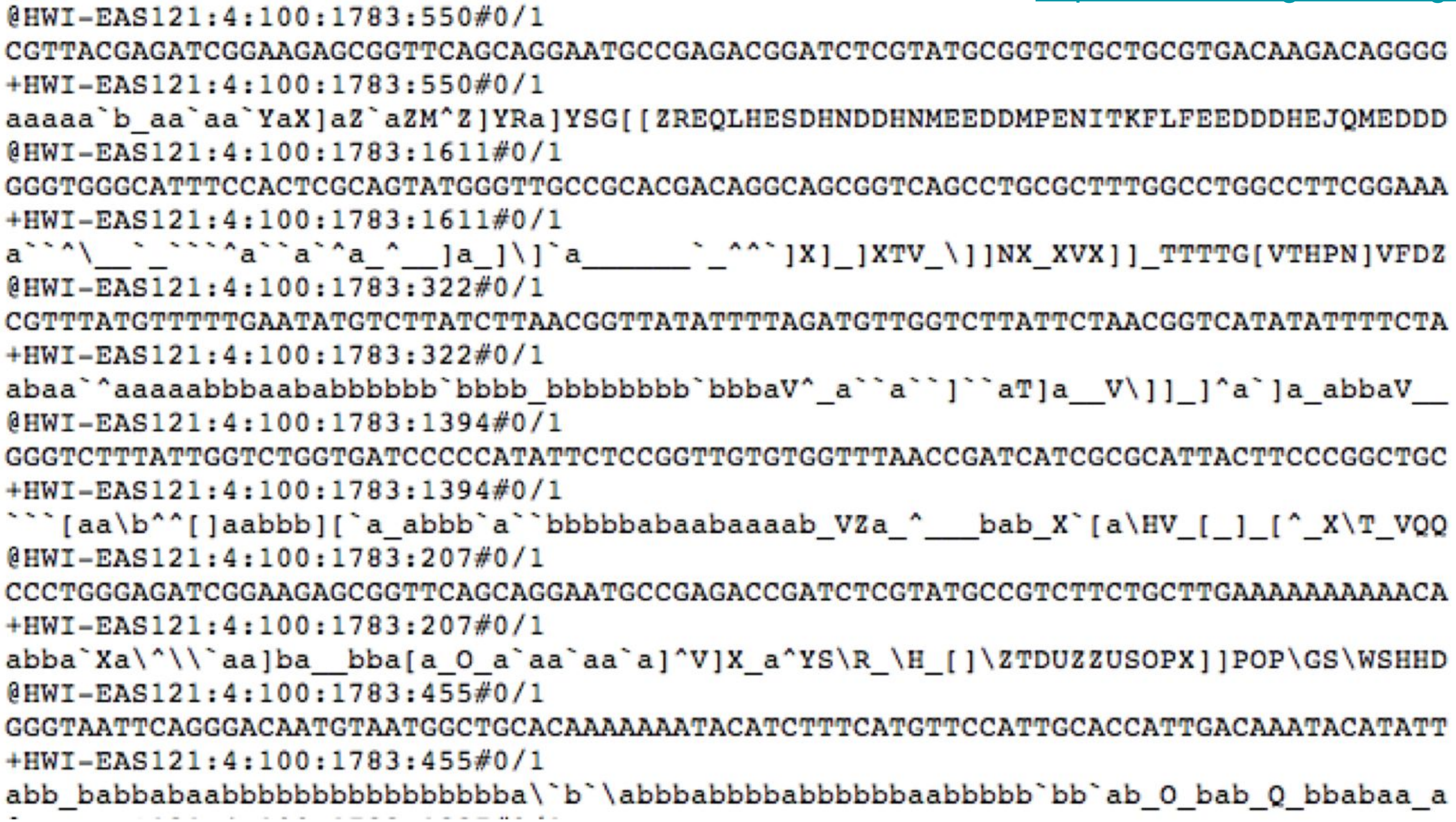
What you wished data looked like



	A	B	C	D	E	F	G
1	id	problem_id	subject_id	start	stop	time_left	answer
2	1	498	17	1307119989	1307120016	2369	A
3	2	150	15	1307119991	1307120009	2376	D
4	3	313	16	1307119994	1307120009	2376	E
5	4	12	13	1307119995	1307120019	2366	B
6	5	273	14	1307119996	1307120028	2357	A
7	6	101	19	1307119996	1307120021	2364	B
8	7	105	18	1307119998	1307120048	2337	B
9	8	162	12	1307120004	1307120042	2343	C
10	9	70	15	1307120011	1307120038	2347	C
11	10	300	16	1307120012	1307120092	2293	B
12	11	494	17	1307120017	1307120075	2310	D
13	12	357	13	1307120021	1307120118	2267	A
14	13	522	19	1307120025	1307120152	2233	D
15	14	232	14	1307120030	1307120158	2227	C
16	15	344	15	1307120041	1307120117	2268	B
17	16	160	17	1307120079	1307120249	2136	D
18	17	516	16	1307120094	1307120159	2226	B
19	18	472	12	1307120119	1307120170	2215	A
20	19	43	15	1307120122	1307120140	2245	C
21	20	353	13	1307120144	1307120199	2186	C
22	21	218	15	1307120152	1307120272	2113	E
23	22	69	16	1307120163	1307120188	2197	D
24	23	562	16	1307120190	1307120301	2084	D
25	24	121	19	1307120253	1307120294	2091	E
26	25	297	15	1307120277	1307120342	2043	B
27	26	495	13	1307120281	1307120353	2032	E
28	27	94	14	1307120288	1307120343	2042	E
29	28	22	18	1307120310	1307120365	2020	C
30	29	64	19	1307120310	1307120385	2000	B
31	30	502	16	1307120323	1307120336	2049	B
32	31	44	16	1307120339	1307120352	2033	A
33	32	315	14	1307120348	1307120362	2023	B
34	33	385	15	1307120352	1307120553	1832	E
35	34	550	13	1307120356	1307120444	1941	B
36	35	92	14	1307120368	1307120397	1988	B
37	36	395	16	1307120377	1307120426	1959	D
38	37	267	17	1307120382	1307120515	1870	E
39	38	257	14	1307120401	1307120427	1958	C
40	39	312	19	1307120407	1307120548	1837	D
41	40	321	18	1307120431	1307120449	1936	A
42	41	220	16	1307120437	1307120510	1875	A

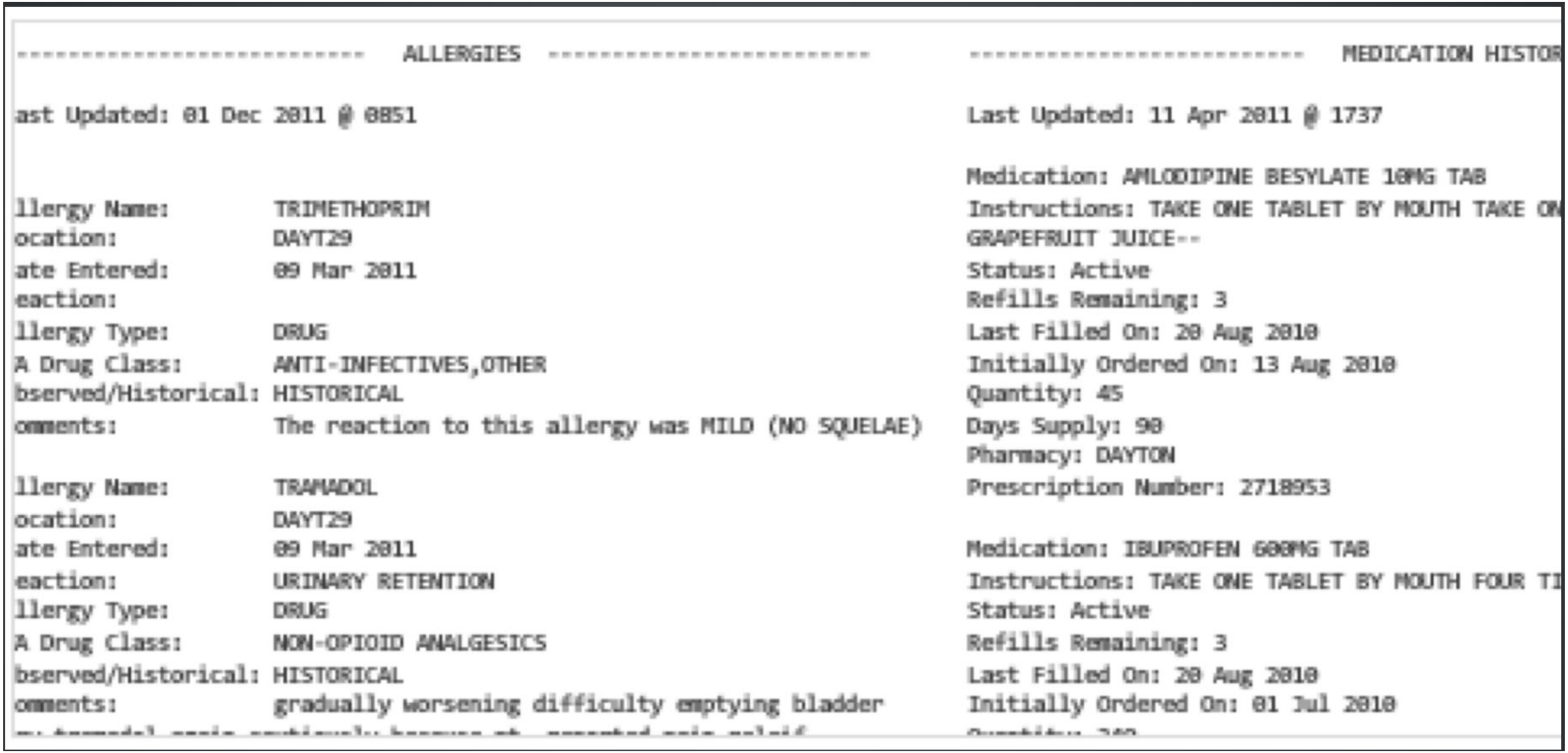
What it actually looks like

http://healthdesignchallenge.com/

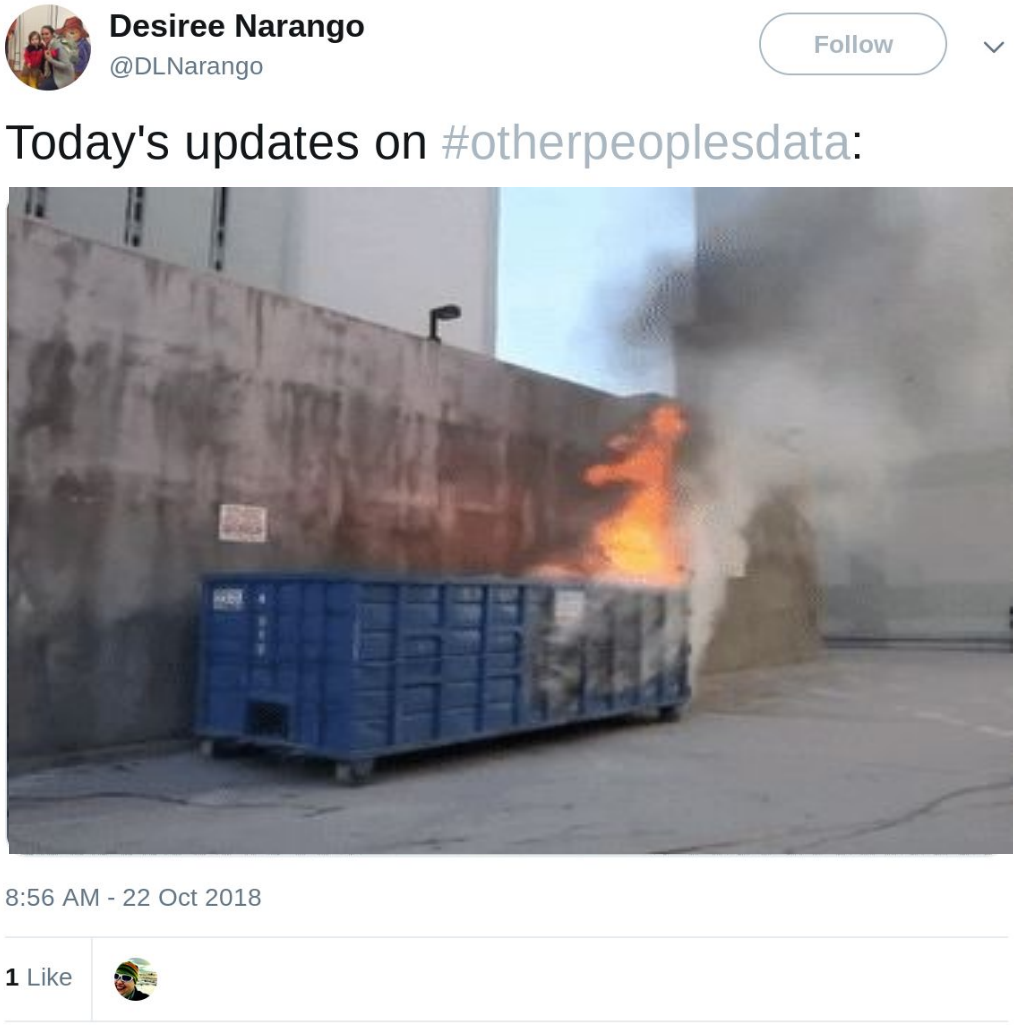


~~~
@HWI-EAS121:4:100:1783:550#0/1
CGTTACGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACGGATCTCGTATGCGGTCTGCTGCGTGACAAGACAGGGG
+HWI-EAS121:4:100:1783:550#0/1
aaaaa`b_aa`aa`YaX]aZ`aZM^Z]YRa]YSG[[ZREQLHESDHNDDHNMEEDDMPENITKFLFEEDDDHEJQMEDDD
@HWI-EAS121:4:100:1783:1611#0/1
GGGTGGGCATTTCCACTCGCAGTATGGGTTGCCGCACGACAGGCAGCGGTCAGCCTGCGCTTTGGCCTGGCCTTCGGAAA
+HWI-EAS121:4:100:1783:1611#0/1
a``^\__`_````^a``a`^a_^__]a_]\]`a_____`_^^`]X]_]XTV_\]]NX_XVX]]_TTTTG[VTHPN]VFDZ
@HWI-EAS121:4:100:1783:322#0/1
CGTTTATGTTTTTGAATATGTCTTATCTTAACGGTTATATTTTAGATGTTGGTCTTATTCTAACGGTCATATATTTTCTA
+HWI-EAS121:4:100:1783:322#0/1
abaa`^aaaaabbbaababbbbbb`bbbb_bbbbbbbb`bbbaV^_a``a``]``aT]a__V\]]_]^a`]a_abbaV__
@HWI-EAS121:4:100:1783:1394#0/1
GGGTCTTTATTGGTCTGGTGATCCCCCATATTCTCCGGTTGTGTGGTTTAACCGATCATCGCGCATTACTTCCCGGCTGC
+HWI-EAS121:4:100:1783:1394#0/1
```[aa\b^^[]aabbb][`a_abbb`a``bbbbbabaabaaaab_VZa_^___bab_X`[a\HV_[_]_[^_X\T_VQQ
@HWI-EAS121:4:100:1783:207#0/1
CCCTGGGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAACA
+HWI-EAS121:4:100:1783:207#0/1
abba`Xa\^\\`aa]ba__bba[a_O_a`aa`aa`a]^V]X_a^YS\R_\H_[]\ZTDUZZUSOPX]]POP\GS\WSHHD
@HWI-EAS121:4:100:1783:455#0/1
GGGTAATTCAGGGACAATGTAATGGCTGCACAAAAAAATACATCTTTCATGTTCCATTGCACCATTGACAAATACATATT
+HWI-EAS121:4:100:1783:455#0/1
abb_babbabaabbbbbbbbbbbbbbbba\`b`\abbbabbbbabbbbbbaabbbbb`bb`ab_O_bab_Q_bbabaa_a
~~~

What it actually looks like



```
-------------------------  ALLERGIES  -----------------------      ------------------------  MEDICATION HISTOR

ast Updated: 01 Dec 2011 @ 0851                                     Last Updated: 11 Apr 2011 @ 1737

                                                                    Medication: AMLODIPINE BESYLATE 10MG TAB
llergy Name:          TRIMETHOPRIM                                  Instructions: TAKE ONE TABLET BY MOUTH TAKE ON
ocation:              DAYT29                                        GRAPEFRUIT JUICE--
ate Entered:          09 Mar 2011                                   Status: Active
eaction:                                                            Refills Remaining: 3
llergy Type:          DRUG                                          Last Filled On: 20 Aug 2010
A Drug Class:         ANTI-INFECTIVES,OTHER                         Initially Ordered On: 13 Aug 2010
bserved/Historical: HISTORICAL                                      Quantity: 45
omments:              The reaction to this allergy was MILD (NO SQUELAE)  Days Supply: 90
                                                                    Pharmacy: DAYTON
llergy Name:          TRAMADOL                                      Prescription Number: 2718953
ocation:              DAYT29
ate Entered:          09 Mar 2011                                   Medication: IBUPROFEN 600MG TAB
eaction:              URINARY RETENTION                             Instructions: TAKE ONE TABLET BY MOUTH FOUR TI
llergy Type:          DRUG                                          Status: Active
A Drug Class:         NON-OPIOID ANALGESICS                         Refills Remaining: 3
bserved/Historical: HISTORICAL                                      Last Filled On: 20 Aug 2010
omments:              gradually worsening difficulty emptying bladder  Initially Ordered On: 01 Jul 2010
```



Desiree Narango
@DLNarango

Today's updates on #otherpeoplesdata:

8:56 AM - 22 Oct 2018

1 Like

And so we data wrangle

Raw & processed data

http://en.wikipedia.org/wiki/Data

“Data are values of qualitative or quantitative variables, belonging to a set of items.”

http://en.wikipedia.org/wiki/Data

> “Data are values of qualitative or quantitative variables, belonging to a set of items.”

Set of items: Sometimes called the population; the set of objects you are interested in

http://en.wikipedia.org/wiki/Data

> “Data are values of qualitative or quantitative variables, belonging to a set of items.”

Variables: A measurement or characteristic of an item

http://en.wikipedia.org/wiki/Data

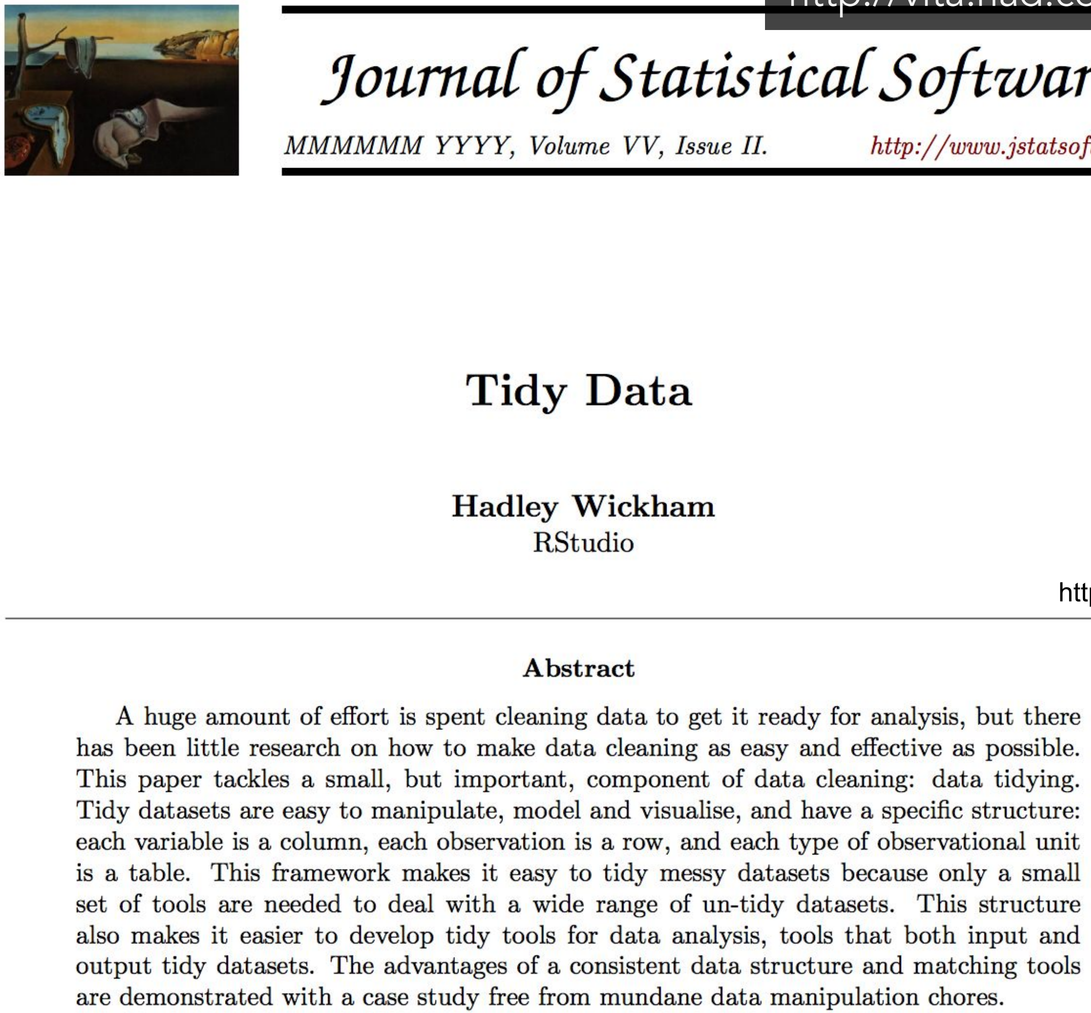
> “Data are values of qualitative or quantitative variables, belonging to a set of items.”

Qualitative: Country of origin, sex, treatment
Quantitative: Height, weight, blood pressure

Data sharing

1. The raw data.
2. A tidy data set
3. A code book describing each variable and its values in the tidy data set.
4. An explicit and exact recipe you used to go from 1 -> 2,3

http://vita.had.co.nz/papers/tidy-data.pdf



Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

Tidy Data

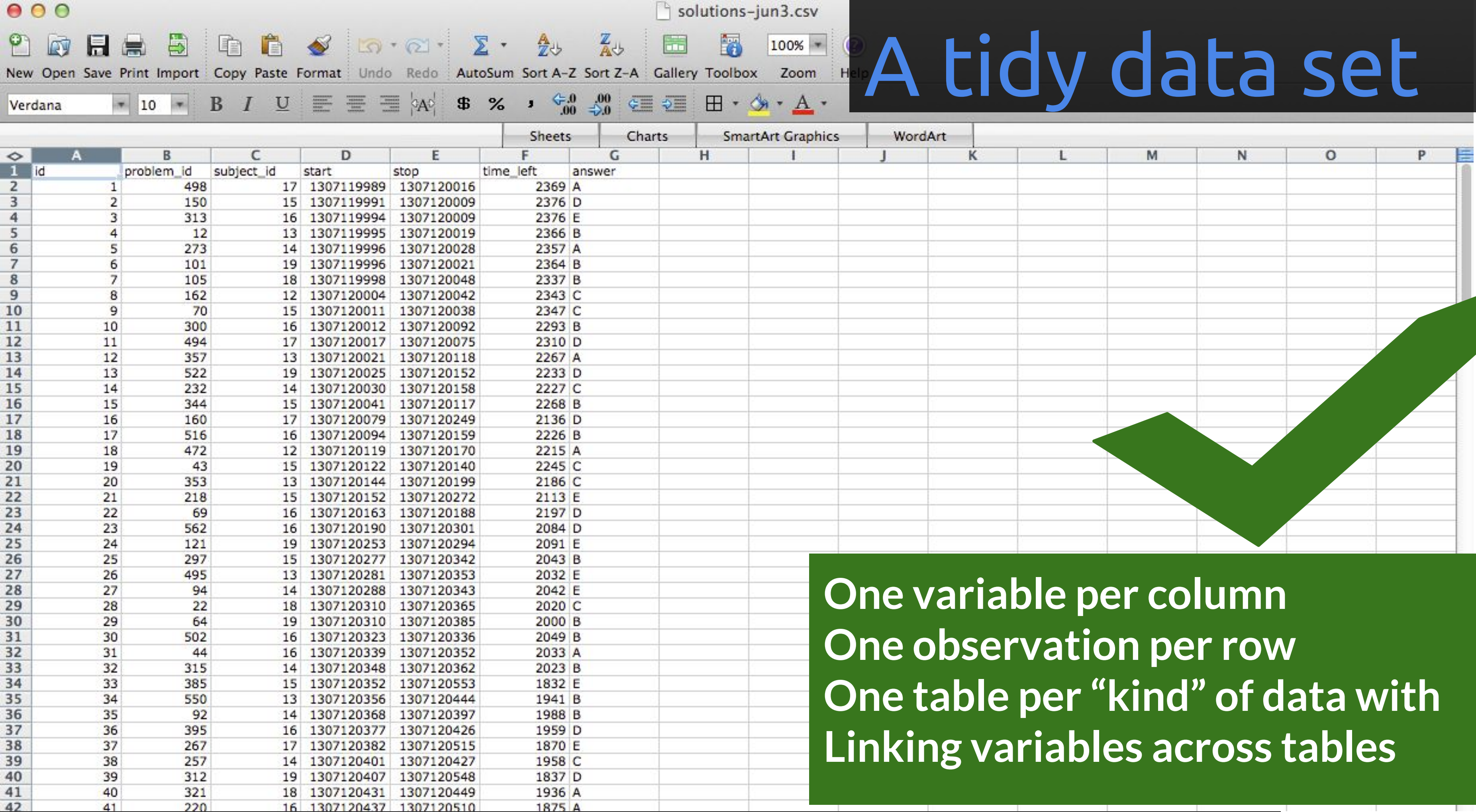
Hadley Wickham
RStudio

https://twitter.com/hadleywickham

Abstract

A huge amount of effort is spent cleaning data to get it ready for analysis, but there has been little research on how to make data cleaning as easy and effective as possible. This paper tackles a small, but important, component of data cleaning: data tidying. Tidy datasets are easy to manipulate, model and visualise, and have a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table. This framework makes it easy to tidy messy datasets because only a small set of tools are needed to deal with a wide range of un-tidy datasets. This structure also makes it easier to develop tidy tools for data analysis, tools that both input and output tidy datasets. The advantages of a consistent data structure and matching tools are demonstrated with a case study free from mundane data manipulation chores.

A tidy data set

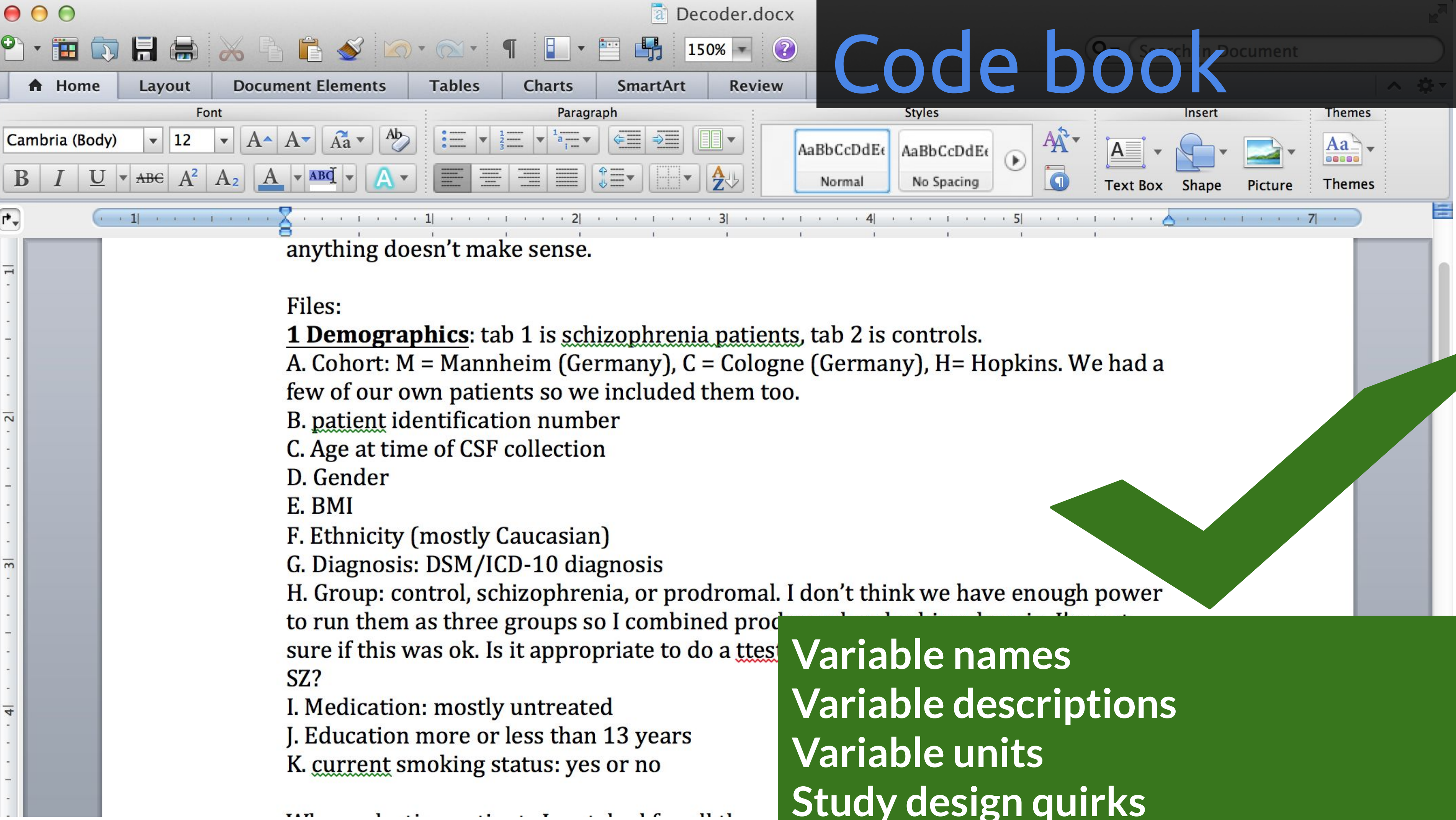


	A	B	C	D	E	F	G
1	id	problem_id	subject_id	start	stop	time_left	answer
2	1	498	17	1307119989	1307120016	2369	A
3	2	150	15	1307119991	1307120009	2376	D
4	3	313	16	1307119994	1307120009	2376	E
5	4	12	13	1307119995	1307120019	2366	B
6	5	273	14	1307119996	1307120028	2357	A
7	6	101	19	1307119996	1307120021	2364	B
8	7	105	18	1307119998	1307120048	2337	B
9	8	162	12	1307120004	1307120042	2343	C
10	9	70	15	1307120011	1307120038	2347	C
11	10	300	16	1307120012	1307120092	2293	B
12	11	494	17	1307120017	1307120075	2310	D
13	12	357	13	1307120021	1307120118	2267	A
14	13	522	19	1307120025	1307120152	2233	D
15	14	232	14	1307120030	1307120158	2227	C
16	15	344	15	1307120041	1307120117	2268	B
17	16	160	17	1307120079	1307120249	2136	D
18	17	516	16	1307120094	1307120159	2226	B
19	18	472	12	1307120119	1307120170	2215	A
20	19	43	15	1307120122	1307120140	2245	C
21	20	353	13	1307120144	1307120199	2186	C
22	21	218	15	1307120152	1307120272	2113	E
23	22	69	16	1307120163	1307120188	2197	D
24	23	562	16	1307120190	1307120301	2084	D
25	24	121	19	1307120253	1307120294	2091	E
26	25	297	15	1307120277	1307120342	2043	B
27	26	495	13	1307120281	1307120353	2032	E
28	27	94	14	1307120288	1307120343	2042	E
29	28	22	18	1307120310	1307120365	2020	C
30	29	64	19	1307120310	1307120385	2000	B
31	30	502	16	1307120323	1307120336	2049	B
32	31	44	16	1307120339	1307120352	2033	A
33	32	315	14	1307120348	1307120362	2023	B
34	33	385	15	1307120352	1307120553	1832	E
35	34	550	13	1307120356	1307120444	1941	B
36	35	92	14	1307120368	1307120397	1988	B
37	36	395	16	1307120377	1307120426	1959	D
38	37	267	17	1307120382	1307120515	1870	E
39	38	257	14	1307120401	1307120427	1958	C
40	39	312	19	1307120407	1307120548	1837	D
41	40	321	18	1307120431	1307120449	1936	A
42	41	220	16	1307120437	1307120510	1875	A

One variable per column
One observation per row
One table per “kind” of data with
Linking variables across tables

Reference: http://brianknaus.com/software/srtoolbox/s_4_1_sequence80.txt

Code book



anything doesn't make sense.

Files:

1 Demographics: tab 1 is schizophrenia patients, tab 2 is controls.

A. Cohort: M = Mannheim (Germany), C = Cologne (Germany), H= Hopkins. We had a few of our own patients so we included them too.

B. patient identification number

C. Age at time of CSF collection

D. Gender

E. BMI

F. Ethnicity (mostly Caucasian)

G. Diagnosis: DSM/ICD-10 diagnosis

H. Group: control, schizophrenia, or prodromal. I don't think we have enough power to run them as three groups so I combined prod

sure if this was ok. Is it appropriate to do a ttest

SZ?

I. Medication: mostly untreated

J. Education more or less than 13 years

K. current smoking status: yes or no

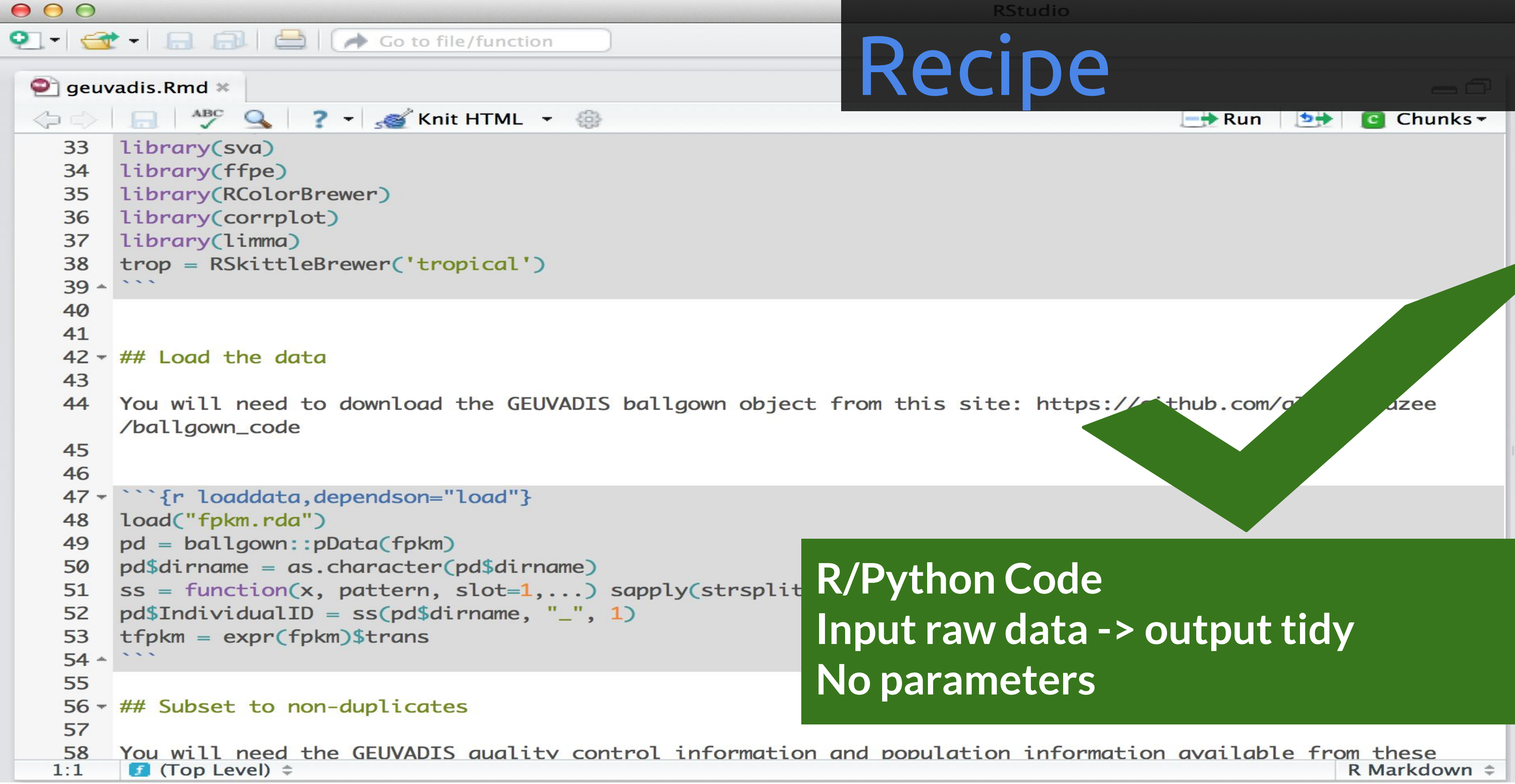
Variable names

Variable descriptions

Variable units

Study design quirks

Recipe

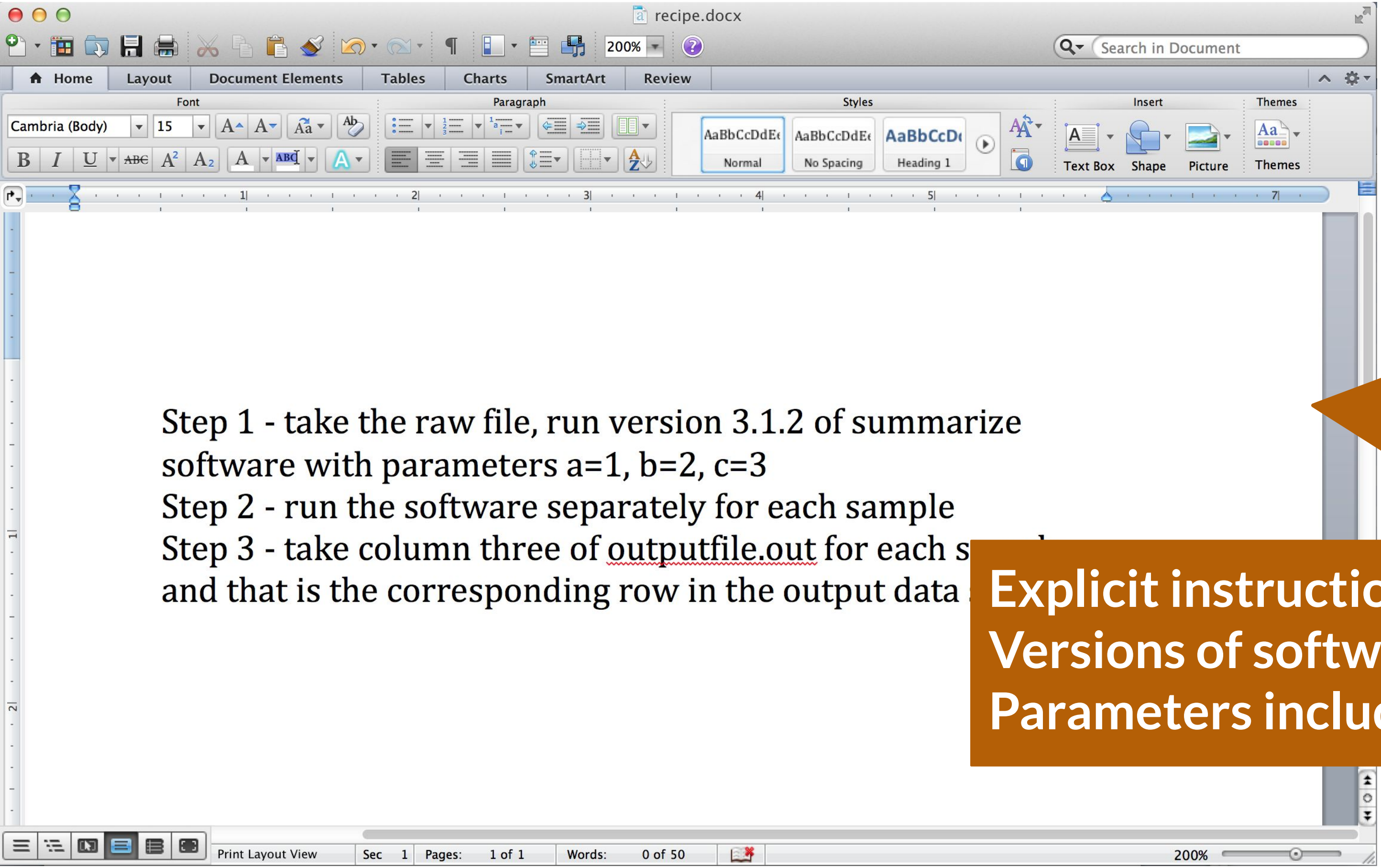


~~~
library(sva)
library(ffpe)
library(RColorBrewer)
library(corrplot)
library(limma)
trop = RSkittleBrewer('tropical')
```‌``


## Load the data

You will need to download the GEUVADIS ballgown object from this site: https://[illegible]thub.com/a[illegible]dzee
/ballgown_code


```‌``{r loaddata,dependson="load"}
load("fpkm.rda")
pd = ballgown::pData(fpkm)
pd$dirname = as.character(pd$dirname)
ss = function(x, pattern, slot=1,...) sapply(strsplit
pd$IndividualID = ss(pd$dirname, "_", 1)
tfpkm = expr(fpkm)$trans
```‌``

## Subset to non-duplicates

You will need the GEUVADIS quality control information and population information available from these
~~~

R/Python Code

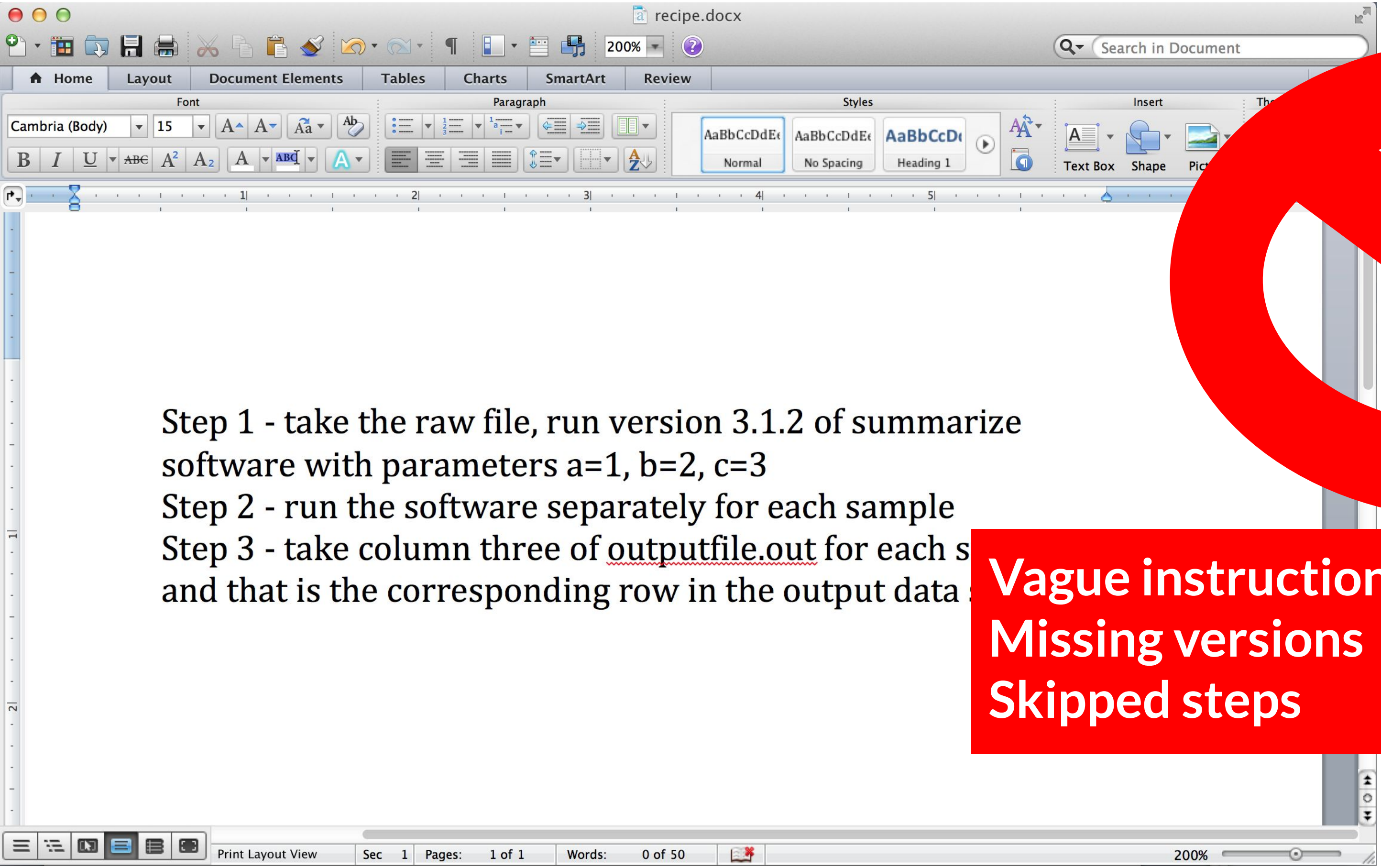
Input raw data -> output tidy

No parameters



Step 1 - take the raw file, run version 3.1.2 of summarize software with parameters a=1, b=2, c=3
Step 2 - run the software separately for each sample
Step 3 - take column three of outputfile.out for each s
and that is the corresponding row in the output data

Explicit instructions
Versions of software
Parameters included



Step 1 - take the raw file, run version 3.1.2 of summarize software with parameters a=1, b=2, c=3
Step 2 - run the software separately for each sample
Step 3 - take column three of outputfile.out for each s
and that is the corresponding row in the output data

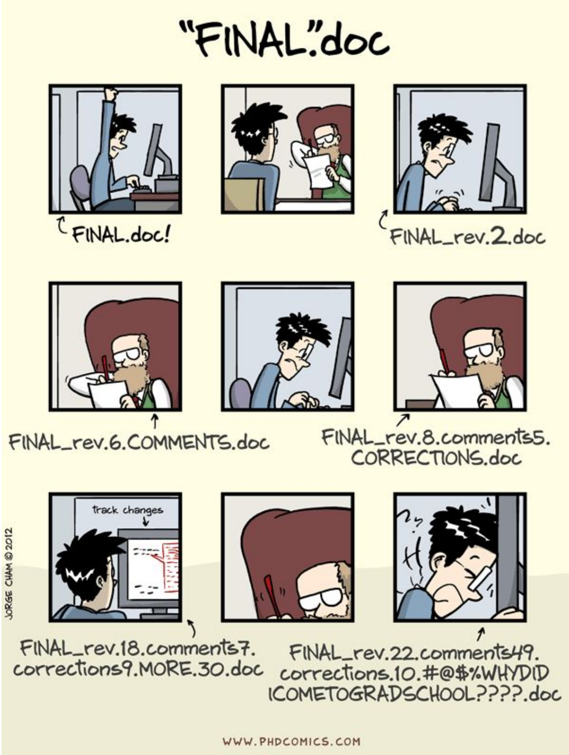
Vague instructions
Missing versions
Skipped steps

Rules for Tidy Spreadsheets

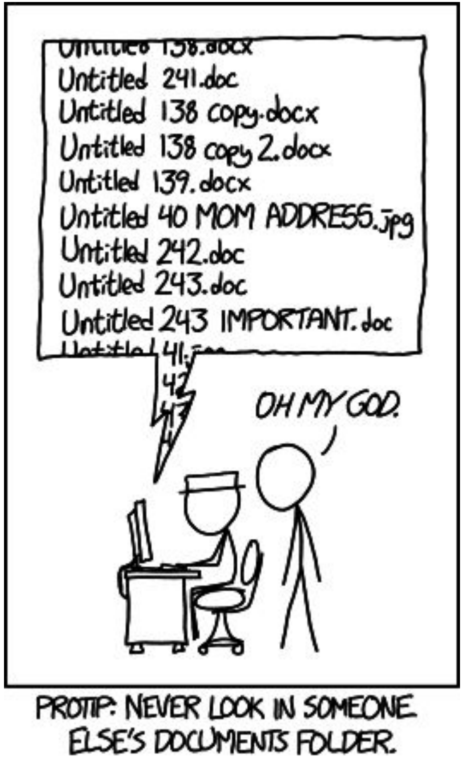
1. Be consistent
2. Choose good names for things
3. Write dates as YYYY-MM-DD
4. No empty cells
5. Put just one thing in a cell
6. Don't use font color or highlighting as data
7. Save the data as plain text files

https://doi.org/10.1080/00031305.2017.1375989

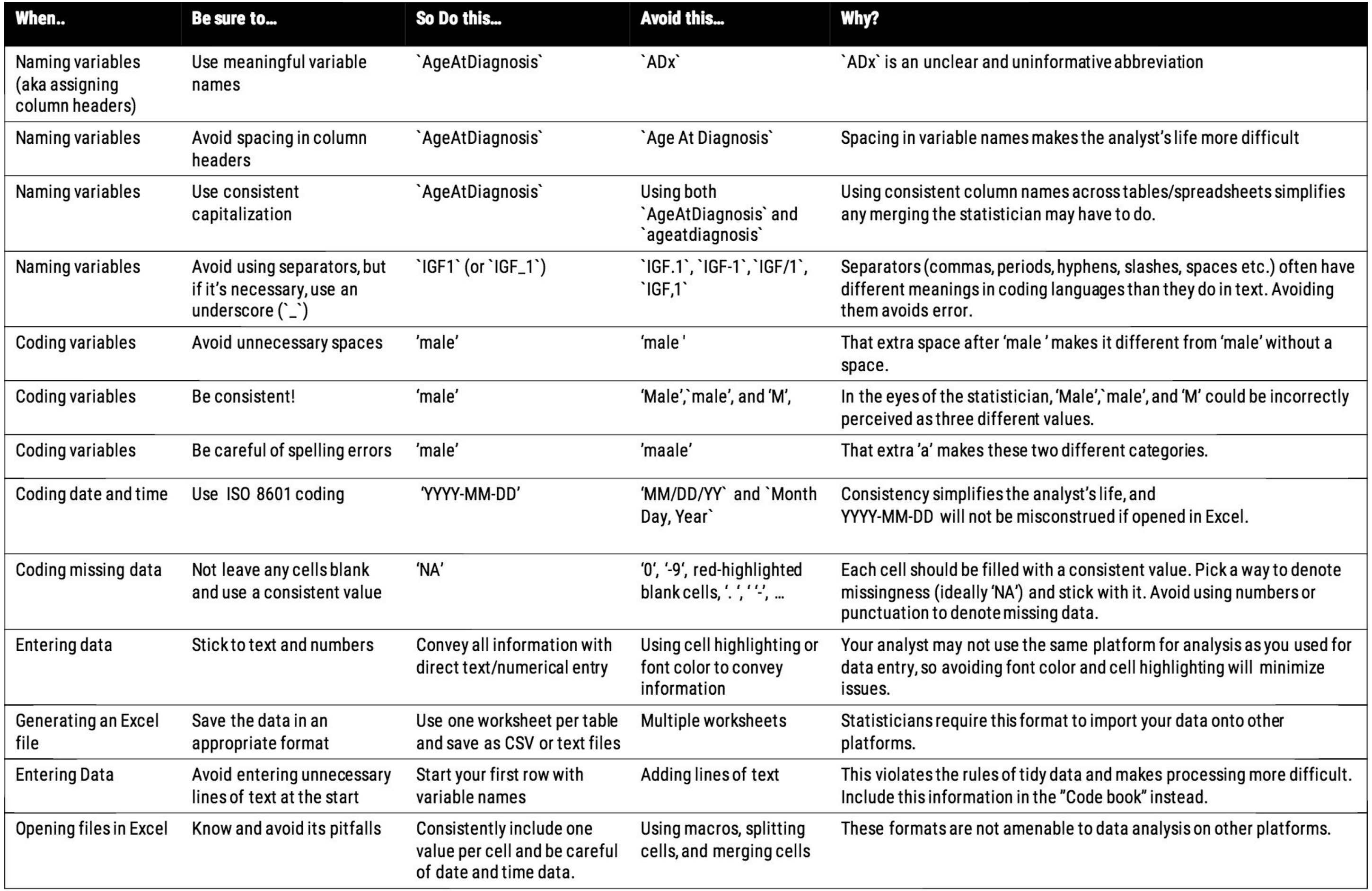
Just no



http://www.phdcomics.com/comics/archive.php?comicid=1531



https://xkcd.com/1459/



When..	Be sure to...	So Do this...	Avoid this...	Why?
Naming variables (aka assigning column headers)	Use meaningful variable names	`AgeAtDiagnosis`	`ADx`	`ADx` is an unclear and uninformative abbreviation
Naming variables	Avoid spacing in column headers	`AgeAtDiagnosis`	`Age At Diagnosis`	Spacing in variable names makes the analyst's life more difficult
Naming variables	Use consistent capitalization	`AgeAtDiagnosis`	Using both `AgeAtDiagnosis` and `ageatdiagnosis`	Using consistent column names across tables/spreadsheets simplifies any merging the statistician may have to do.
Naming variables	Avoid using separators, but if it's necessary, use an underscore (`_`)	`IGF1` (or `IGF_1`)	`IGF.1`, `IGF-1`, `IGF/1`, `IGF,1`	Separators (commas, periods, hyphens, slashes, spaces etc.) often have different meanings in coding languages than they do in text. Avoiding them avoids error.
Coding variables	Avoid unnecessary spaces	'male'	'male '	That extra space after 'male ' makes it different from 'male' without a space.
Coding variables	Be consistent!	'male'	'Male','male', and 'M',	In the eyes of the statistician, 'Male','male', and 'M' could be incorrectly perceived as three different values.
Coding variables	Be careful of spelling errors	'male'	'maale'	That extra 'a' makes these two different categories.
Coding date and time	Use ISO 8601 coding	'YYYY-MM-DD'	'MM/DD/YY' and `Month Day, Year`	Consistency simplifies the analyst's life, and YYYY-MM-DD will not be misconstrued if opened in Excel.
Coding missing data	Not leave any cells blank and use a consistent value	'NA'	'0', '-9', red-highlighted blank cells, '. ', ' '-', ...	Each cell should be filled with a consistent value. Pick a way to denote missingness (ideally 'NA') and stick with it. Avoid using numbers or punctuation to denote missing data.
Entering data	Stick to text and numbers	Convey all information with direct text/numerical entry	Using cell highlighting or font color to convey information	Your analyst may not use the same platform for analysis as you used for data entry, so avoiding font color and cell highlighting will minimize issues.
Generating an Excel file	Save the data in an appropriate format	Use one worksheet per table and save as CSV or text files	Multiple worksheets	Statisticians require this format to import your data onto other platforms.
Entering Data	Avoid entering unnecessary lines of text at the start	Start your first row with variable names	Adding lines of text	This violates the rules of tidy data and makes processing more difficult. Include this information in the "Code book" instead.
Opening files in Excel	Know and avoid its pitfalls	Consistently include one value per cell and be careful of date and time data.	Using macros, splitting cells, and merging cells	These formats are not amenable to data analysis on other platforms.

Ellis SE, Leek JT. (2017) How to share data for collaboration. *PeerJ Preprints* 5:e3139v5https://doi.org/10.7287/peerj.preprints.3139v5

Key principles of file naming for data science projects:

- Machine readable
- Human readable
- Be nicely ordered

Source: Jenny Bryan

Bad Naming	Good Naming
2013 my report.md	2013_my_report.md
malik's_report.md	maliks_report.md
01_zoë_report.md	01_zoe_report.md
AdamHooverReport.md	adam-hoover-report.md
executivereportpepsiv1.md	executive_report_pepsi_v1.md

2018_jan_sales_cust001_prod001.md

2017_mar_sales_cust001_prod001.md

2016_may_sales_cust001_prod008.md

2017_jan_sales_cust120_prod007.md

2015_oct_sales_cust034_prod001.md

2015_oct_sales_cust034_prod002.md

Year	Month	Type	Customer ID	Product ID
2018	jan	sales	001	001
2017	mar	sales	001	001
2016	may	sales	001	008
2017	jan	sales	120	007
2015	oct	sales	034	001
2015	oct	sales	034	002

Which one is better?

analysis.R

or

2017-exploratory_analysis_crime.R?

Which one is better?

05-21-2017-analysis-cust001.R
or
2017-05-21-analysis-cust001.R?

Structure of a filename

processed_pvalue_data_from_pubmed_oct24.rda

What did I do to this data

processed_pvalue_data_from_pubmed_oct24.rda

What kind of data is this?

processed_pvalue_data_from_pubmed_oct24.rda

Where did it come from?

processed_pvalue_data_from_pubmed_oct24.rda

When did I get it?

processed_pvalue_data_from_pubmed_oct24.rda

Underscores/slashes not dots/whitespace

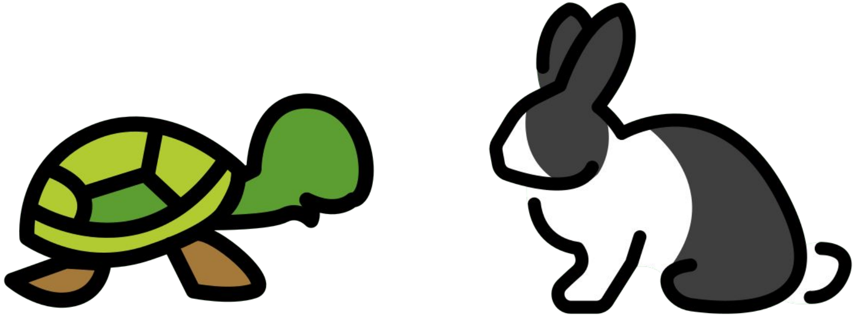
processed_pvalue_data_from_pubmed_oct24.rda

Consistency is the main rule

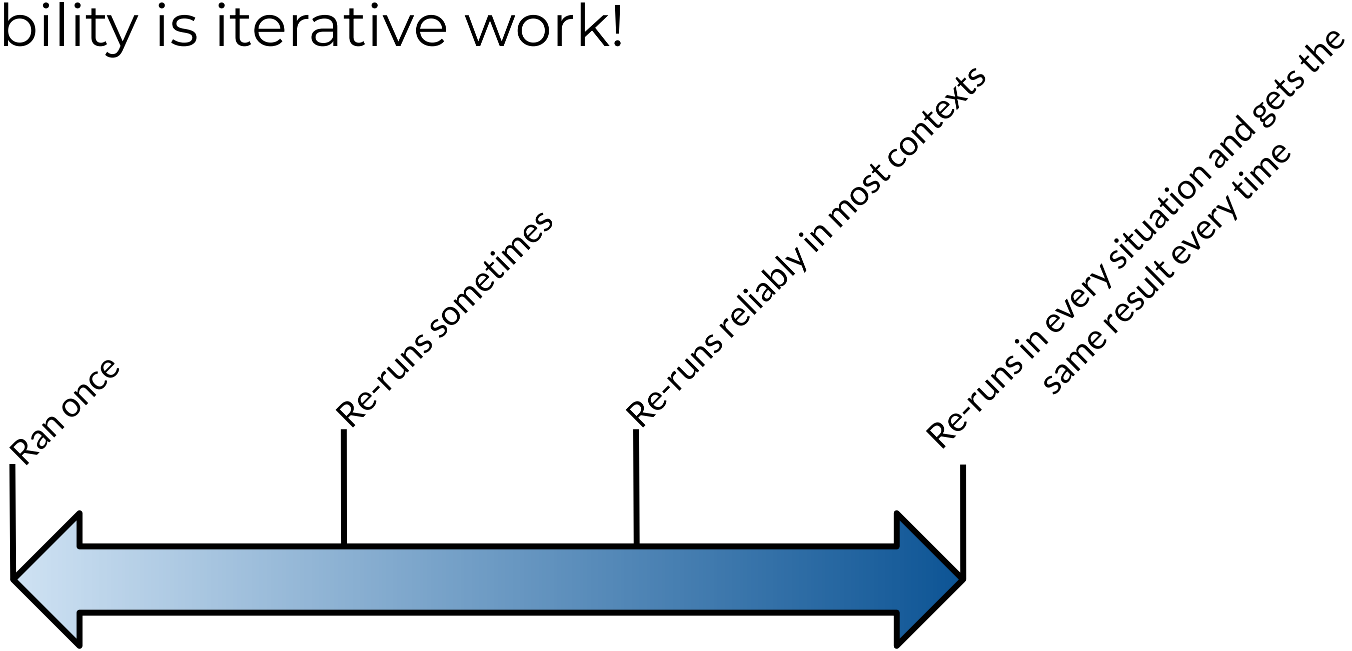
processed_pvalue_data_from_pubmed_oct24.rda
raw_pvalue_data_from_pubmed_oct24.rda

Organize thyself

Reproducibility is a tortoise’s game - it’s an incremental and slow process *but* **it has high payoffs!**



Reproducibility is iterative work!



Ran once

Re-runs sometimes

Re-runs reliably in most contexts

Re-runs in every situation and gets the same result every time

Not repeatable

Every analysis starts here

Perfectly reproducible

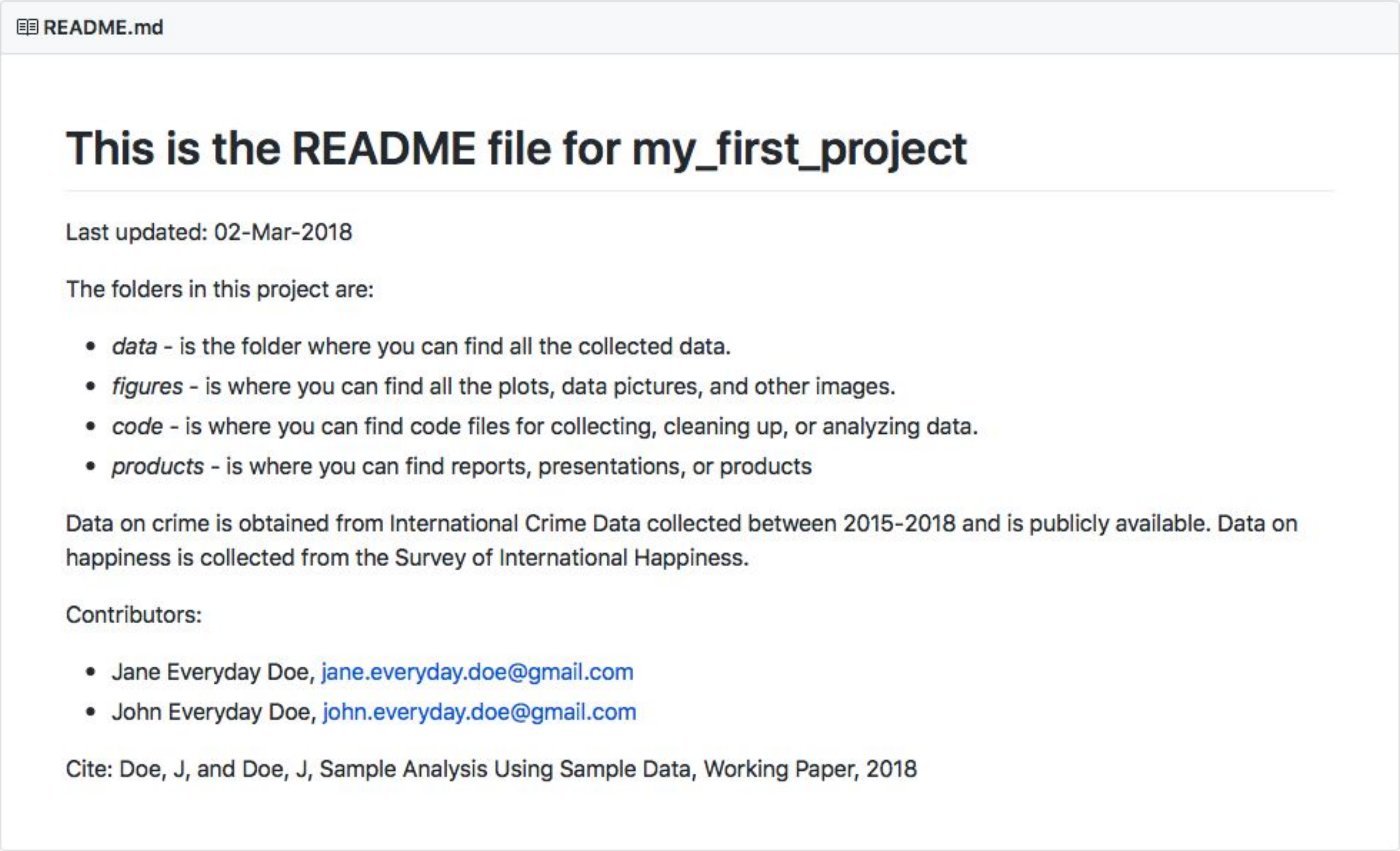
No analysis reaches here

Image created by Candace Savonen.

Documentation that every project should have!

AI can help, but check everything!

1. **READMEs**
    a. Background knowledge
    b. Usage info
    c. Software requirements to run the thing
    d. Basics on how the files are organized
2. **Code annotations:**
    a. Explain historical decisions
    b. Explain “quirks” of the code
    c. Say where more development is needed (TODO)
    d. Summarize the goals!



This is the README file for my_first_project

Last updated: 02-Mar-2018

The folders in this project are:

- *data* - is the folder where you can find all the collected data.
- *figures* - is where you can find all the plots, data pictures, and other images.
- *code* - is where you can find code files for collecting, cleaning up, or analyzing data.
- *products* - is where you can find reports, presentations, or products

Data on crime is obtained from International Crime Data collected between 2015-2018 and is publicly available. Data on happiness is collected from the Survey of International Happiness.

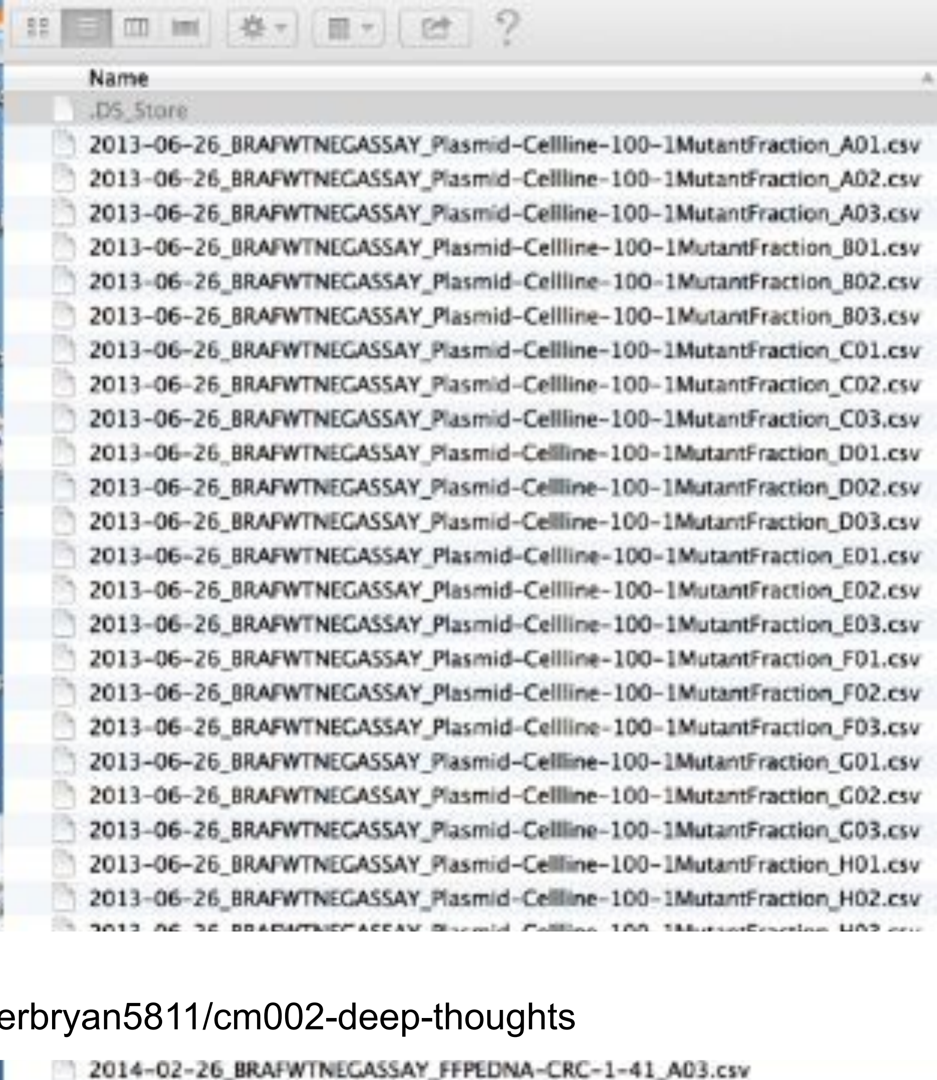
Contributors:

- Jane Everyday Doe, jane.everyday.doe@gmail.com
- John Everyday Doe, john.everyday.doe@gmail.com

Cite: Doe, J, and Doe, J, Sample Analysis Using Sample Data, Working Paper, 2018

"File organization and naming are powerful weapons against chaos."

- Jenny Bryan

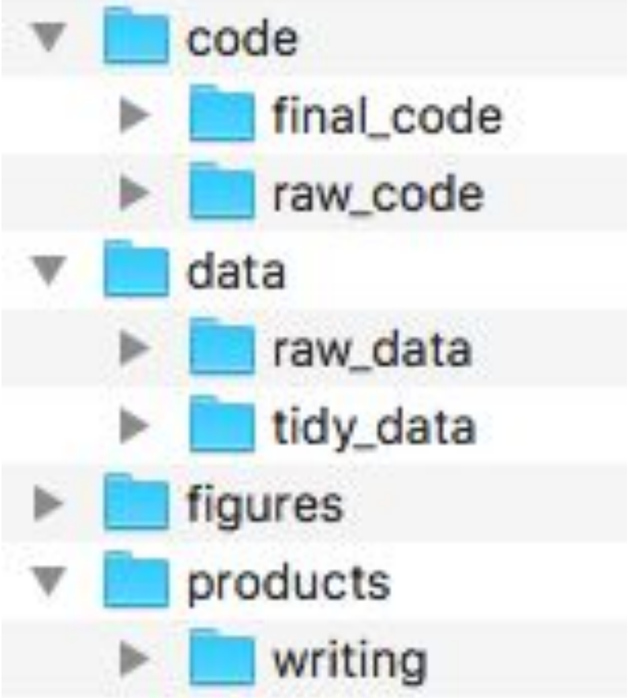

Name

.DS_Store

2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_A01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_A02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_A03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_B01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_B02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_B03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_C01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_C02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_C03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_D01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_D02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_D03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_E01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_E02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_E03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_F01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_F02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_F03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_G01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_G02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_G03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H02.csv

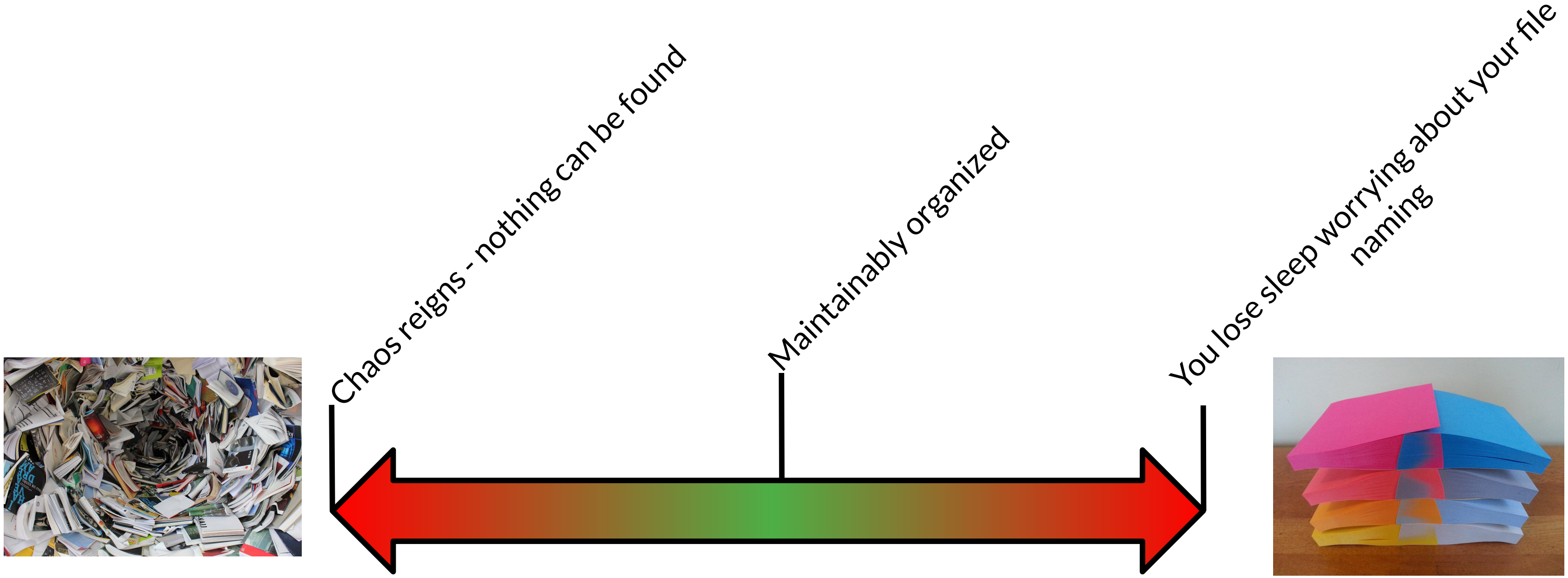
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A03.csv

Slide via Jenny Bryan:
http://www.slideshare.net/jenniferbryan5811/cm002-deep-thoughts



- code
  - final_code
  - raw_code
- data
  - raw_data
  - tidy_data
- figures
- products
  - writing

Your organizational system should work for you not the other way around!



Chaos reigns - nothing can be found

Maintainably organized

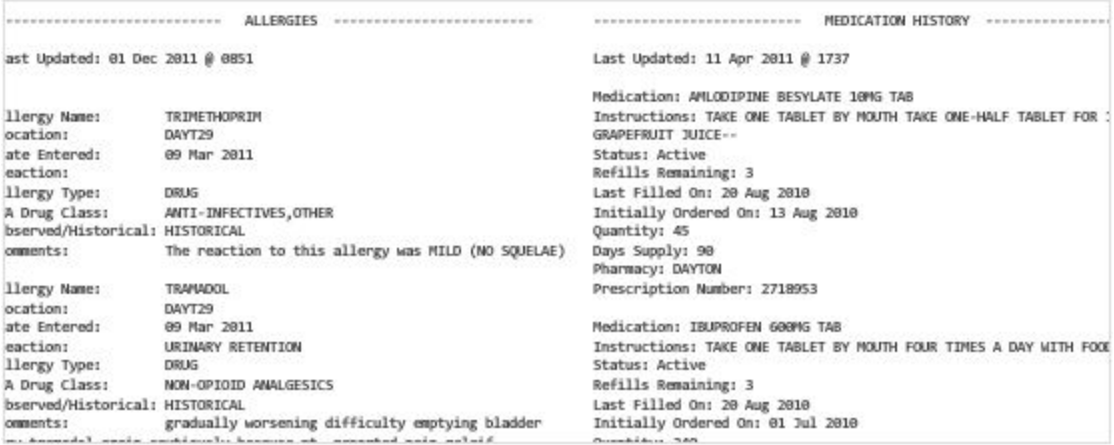
You lose sleep worrying about your file naming

Disorganized and unmanageable

Perfectly organized but maybe not maintainable

Image created by Candace Savonen.

Raw data

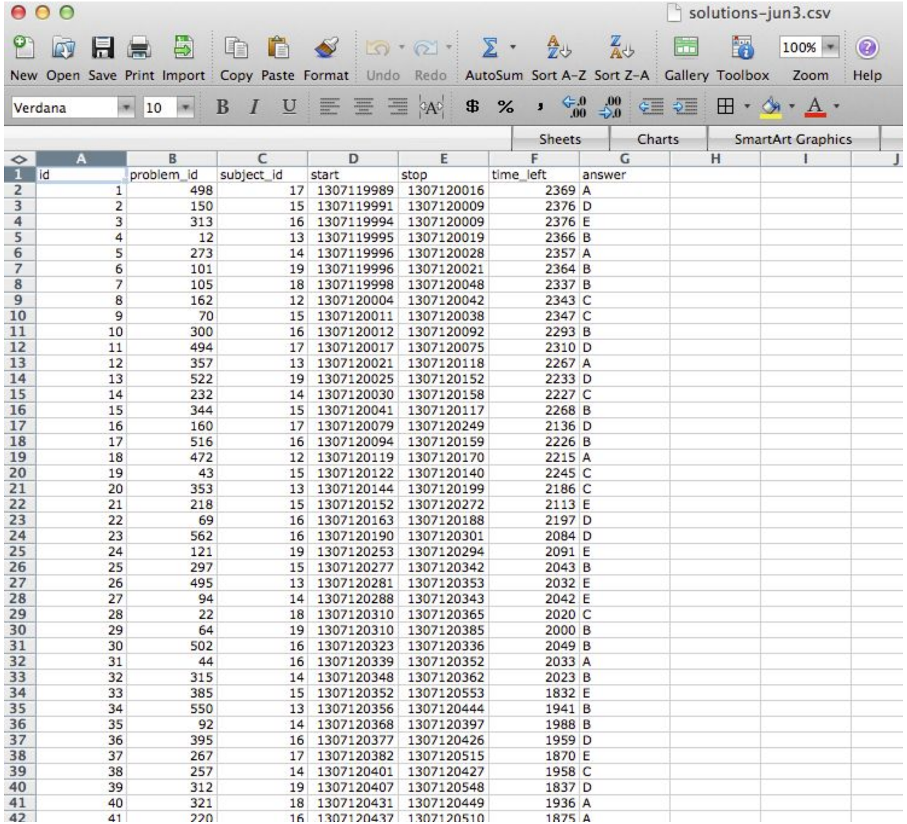


```
--------------------------   ALLERGIES   ------------------------

ast Updated: 01 Dec 2011 @ 0851


llergy Name:        TRIMETHOPRIM
ocation:            DAYT29
ate Entered:        09 Mar 2011
eaction:
llergy Type:        DRUG
A Drug Class:       ANTI-INFECTIVES,OTHER
bserved/Historical: HISTORICAL
omments:            The reaction to this allergy was MILD (NO SQUELAE)

llergy Name:        TRAMADOL
ocation:            DAYT29
ate Entered:        09 Mar 2011
eaction:            URINARY RETENTION
llergy Type:        DRUG
A Drug Class:       NON-OPIOID ANALGESICS
bserved/Historical: HISTORICAL
omments:            gradually worsening difficulty emptying bladder
```

```
--------------------------   MEDICATION HISTORY   ---------------

Last Updated: 11 Apr 2011 @ 1737

Medication: AMLODIPINE BESYLATE 10MG TAB
Instructions: TAKE ONE TABLET BY MOUTH TAKE ONE-HALF TABLET FOR 
GRAPEFRUIT JUICE--
Status: Active
Refills Remaining: 3
Last Filled On: 20 Aug 2010
Initially Ordered On: 13 Aug 2010
Quantity: 45
Days Supply: 90
Pharmacy: DAYTON
Prescription Number: 2718953

Medication: IBUPROFEN 600MG TAB
Instructions: TAKE ONE TABLET BY MOUTH FOUR TIMES A DAY WITH FOO
Status: Active
Refills Remaining: 3
Last Filled On: 20 Aug 2010
Initially Ordered On: 01 Jul 2010
Quantity: 240
```

Processed data

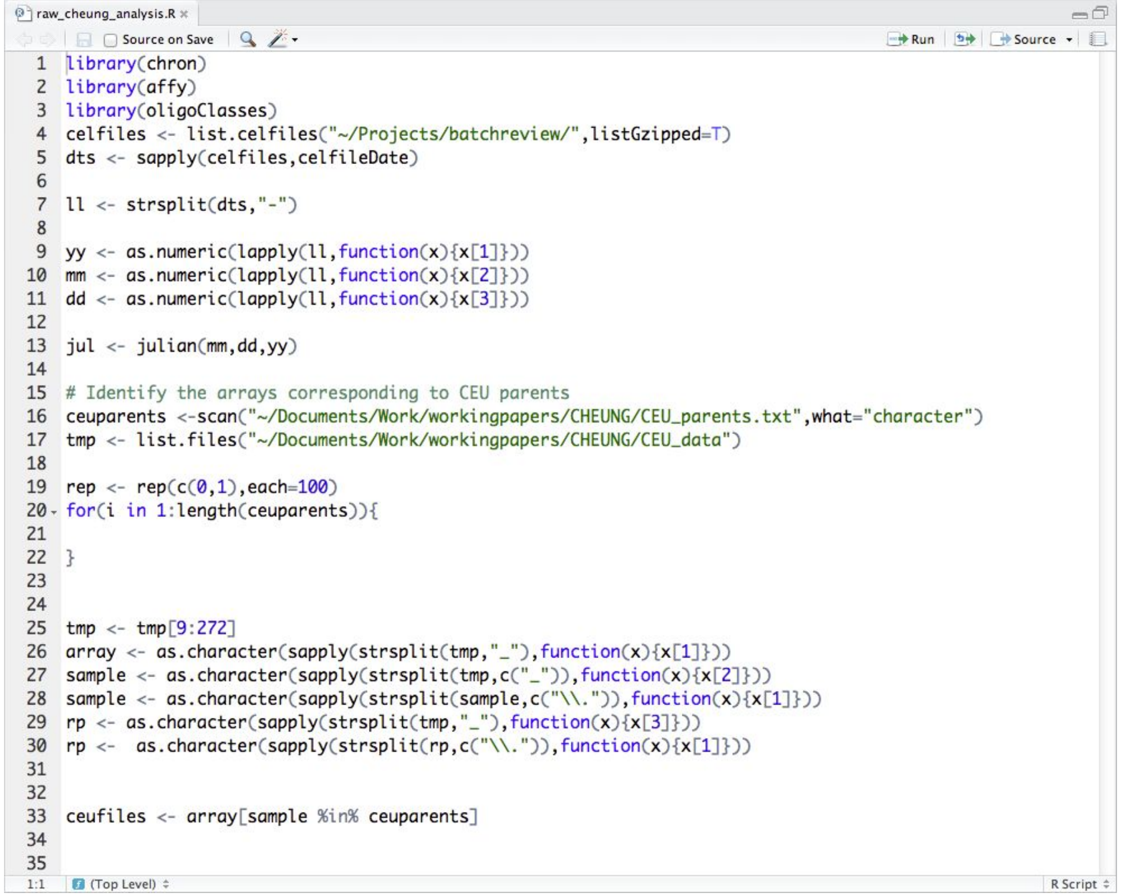
- Processed data should be named so it is easy to see which script generated the data.
- The processing script - processed data mapping should occur in the README
- Processed data should be tidy



	A	B	C	D	E	F	G	H	I
1	id	problem_id	subject_id	start	stop	time_left	answer		
2	1	498	17	1307119989	1307120016	2369	A		
3	2	150	15	1307119991	1307120009	2376	D		
4	3	313	16	1307119994	1307120009	2376	E		
5	4	12	13	1307119995	1307120019	2366	B		
6	5	273	14	1307119996	1307120028	2357	A		
7	6	101	19	1307119996	1307120021	2364	B		
8	7	105	18	1307119998	1307120048	2337	B		
9	8	162	12	1307120004	1307120042	2343	C		
10	9	70	15	1307120011	1307120038	2347	C		
11	10	300	16	1307120012	1307120092	2293	B		
12	11	494	17	1307120017	1307120075	2310	D		
13	12	357	13	1307120021	1307120118	2267	A		
14	13	522	19	1307120025	1307120152	2233	D		
15	14	232	14	1307120030	1307120158	2227	C		
16	15	344	15	1307120041	1307120117	2268	B		
17	16	160	17	1307120079	1307120249	2136	D		
18	17	516	16	1307120094	1307120159	2226	B		
19	18	472	12	1307120119	1307120170	2215	A		
20	19	43	15	1307120122	1307120140	2245	C		
21	20	353	13	1307120144	1307120199	2186	C		
22	21	218	15	1307120152	1307120272	2113	E		
23	22	69	16	1307120163	1307120188	2197	D		
24	23	562	16	1307120190	1307120301	2084	D		
25	24	121	19	1307120253	1307120294	2091	E		
26	25	297	15	1307120277	1307120342	2043	B		
27	26	495	13	1307120281	1307120353	2032	E		
28	27	94	14	1307120288	1307120343	2042	E		
29	28	22	18	1307120310	1307120365	2020	C		
30	29	64	19	1307120310	1307120385	2000	B		
31	30	502	16	1307120323	1307120336	2049	B		
32	31	44	16	1307120339	1307120352	2033	A		
33	32	315	14	1307120348	1307120362	2023	B		
34	33	385	15	1307120352	1307120553	1832	E		
35	34	550	13	1307120356	1307120444	1941	B		
36	35	92	14	1307120368	1307120397	1988	B		
37	36	395	16	1307120377	1307120426	1959	D		
38	37	267	17	1307120382	1307120515	1870	E		
39	38	257	14	1307120401	1307120427	1958	C		
40	39	312	19	1307120407	1307120548	1837	D		
41	40	321	18	1307120431	1307120449	1936	A		
42	41	220	16	1307120437	1307120510	1875	A		

Raw scripts

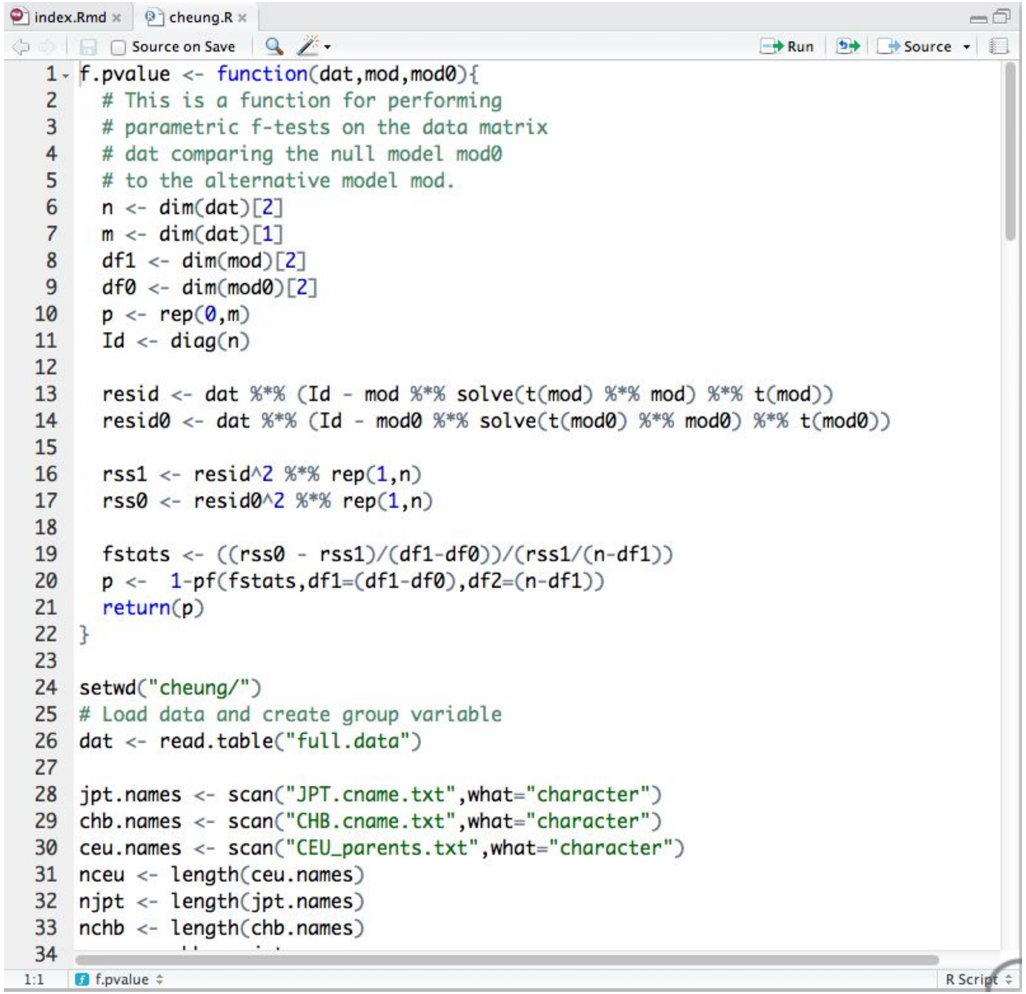
- May be less commented (but comments help you!)
- May be multiple versions
- May include analyses that are later discarded



```r
library(chron)
library(affy)
library(oligoClasses)
celfiles <- list.celfiles("~/Projects/batchreview/",listGzipped=T)
dts <- sapply(celfiles,celfileDate)

ll <- strsplit(dts,"-")

yy <- as.numeric(lapply(ll,function(x){x[1]}))
mm <- as.numeric(lapply(ll,function(x){x[2]}))
dd <- as.numeric(lapply(ll,function(x){x[3]}))

jul <- julian(mm,dd,yy)

# Identify the arrays corresponding to CEU parents
ceuparents <-scan("~/Documents/Work/workingpapers/CHEUNG/CEU_parents.txt",what="character")
tmp <- list.files("~/Documents/Work/workingpapers/CHEUNG/CEU_data")

rep <- rep(c(0,1),each=100)
for(i in 1:length(ceuparents)){

}


tmp <- tmp[9:272]
array <- as.character(sapply(strsplit(tmp,"_"),function(x){x[1]}))
sample <- as.character(sapply(strsplit(tmp,c("_")),function(x){x[2]}))
sample <- as.character(sapply(strsplit(sample,c("\\.")),function(x){x[1]}))
rp <- as.character(sapply(strsplit(tmp,"_"),function(x){x[3]}))
rp <-  as.character(sapply(strsplit(rp,c("\\.")),function(x){x[1]}))


ceufiles <- array[sample %in% ceuparents]


```

Final scripts

- Clearly commented
  - Small comments liberally - what, when, why, how
  - Bigger commented blocks for whole sections
- Include processing details
- Only analyses that appear in the final write-up



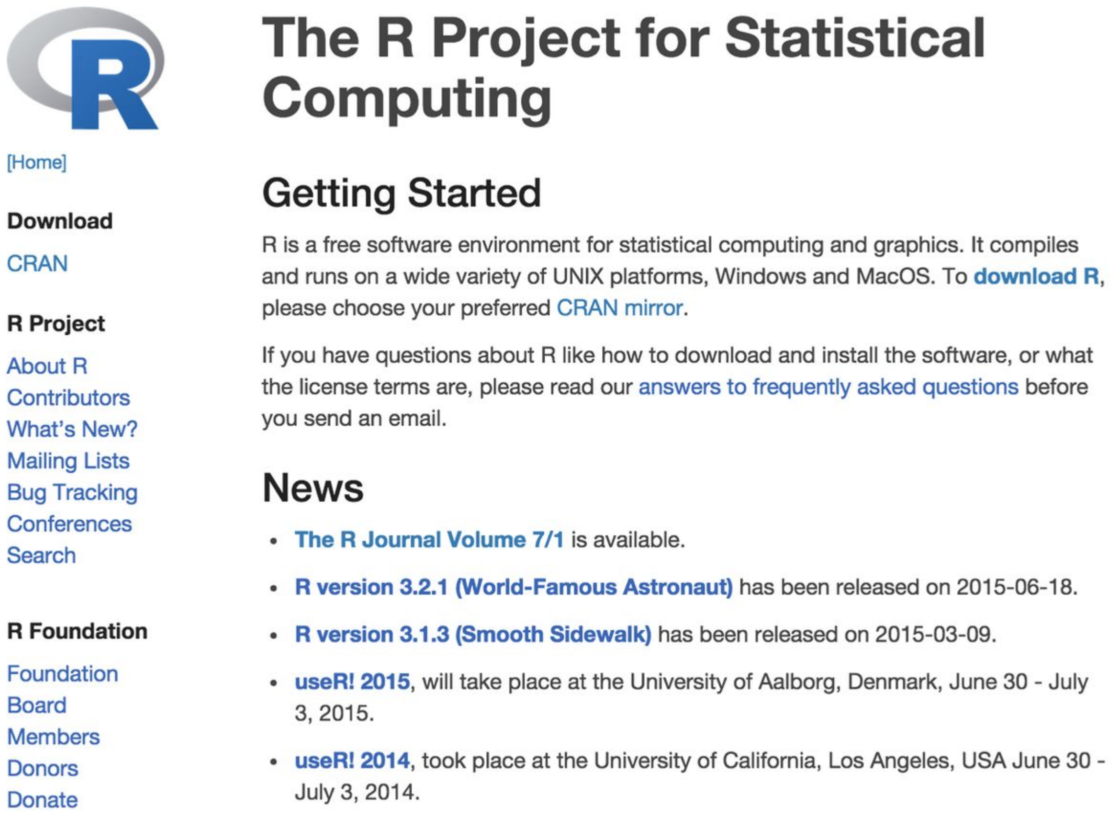
```r
f.pvalue <- function(dat,mod,mod0){
  # This is a function for performing
  # parametric f-tests on the data matrix
  # dat comparing the null model mod0
  # to the alternative model mod.
  n <- dim(dat)[2]
  m <- dim(dat)[1]
  df1 <- dim(mod)[2]
  df0 <- dim(mod0)[2]
  p <- rep(0,m)
  Id <- diag(n)

  resid <- dat %*% (Id - mod %*% solve(t(mod) %*% mod) %*% t(mod))
  resid0 <- dat %*% (Id - mod0 %*% solve(t(mod0) %*% mod0) %*% t(mod0))

  rss1 <- resid^2 %*% rep(1,n)
  rss0 <- resid0^2 %*% rep(1,n)

  fstats <- ((rss0 - rss1)/(df1-df0))/(rss1/(n-df1))
  p <-  1-pf(fstats,df1=(df1-df0),df2=(n-df1))
  return(p)
}

setwd("cheung/")
# Load data and create group variable
dat <- read.table("full.data")

jpt.names <- scan("JPT.cname.txt",what="character")
chb.names <- scan("CHB.cname.txt",what="character")
ceu.names <- scan("CEU_parents.txt",what="character")
nceu <- length(ceu.names)
njpt <- length(jpt.names)
nchb <- length(chb.names)
```

Step 1: slow down and document.
Step 2: have sympathy for your future self.
Step 3: have a system.

- Karl Broman

http://kbroman.org/Tools4RR/assets/lectures/06_org_eda.pdf

R + RStudio



[Home]

Download

CRAN

R Project

About R
Contributors
What's New?
Mailing Lists
Bug Tracking
Conferences
Search

R Foundation

Foundation
Board
Members
Donors
Donate

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

News

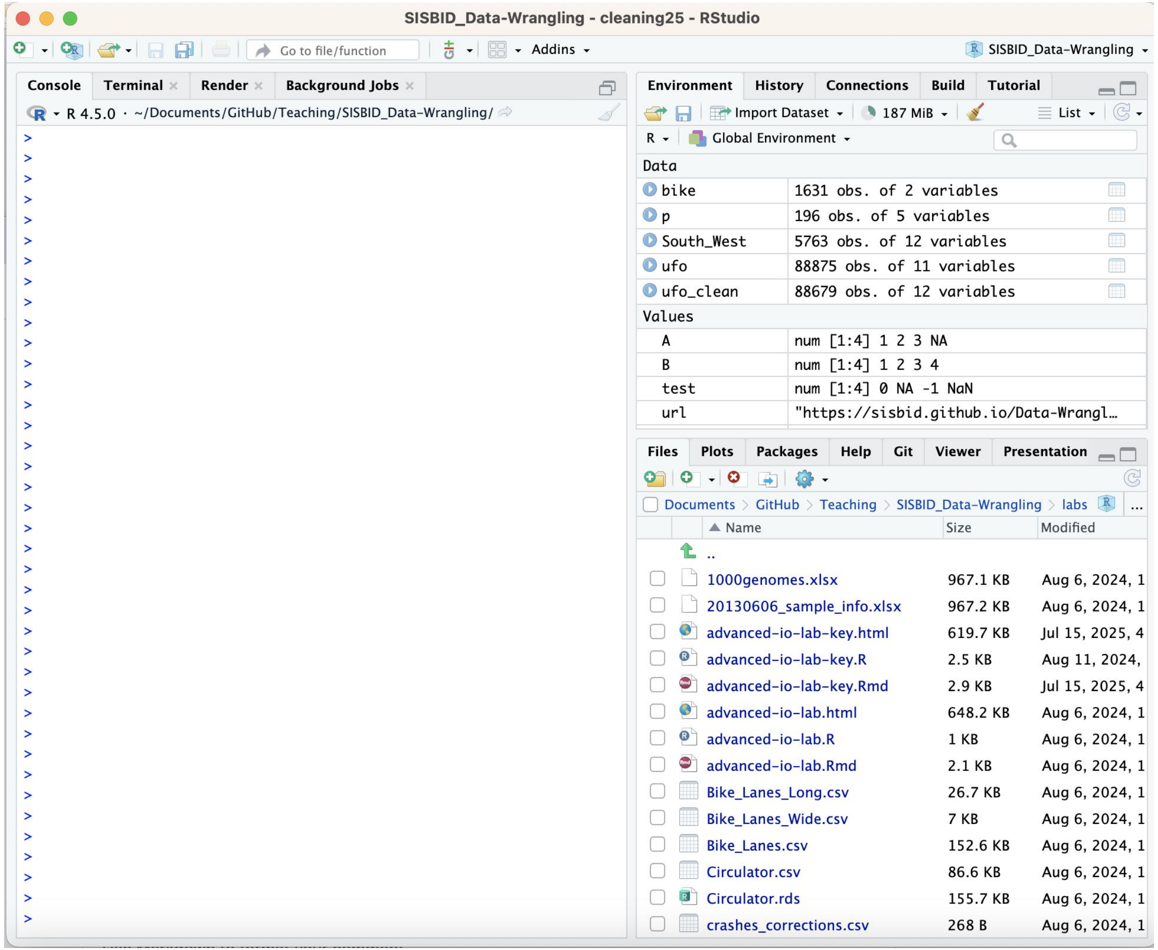
- **The R Journal Volume 7/1** is available.
- **R version 3.2.1 (World-Famous Astronaut)** has been released on 2015-06-18.
- **R version 3.1.3 (Smooth Sidewalk)** has been released on 2015-03-09.
- **useR! 2015**, will take place at the University of Aalborg, Denmark, June 30 - July 3, 2015.
- **useR! 2014**, took place at the University of California, Los Angeles, USA June 30 - July 3, 2014.

www.r-project.org

RSTUDIO IDE

The most trusted IDE for open source data science

RStudio is an integrated development environment (IDE) for R and Python. It includes a console, syntax-highlighting editor that supports direct code execution, and tools for plotting, history, debugging, and workspace management. RStudio is available in open source and commercial editions and runs on the desktop (Windows, Mac, and Linux).



Data

bike	1631 obs. of 2 variables
p	196 obs. of 5 variables
South_West	5763 obs. of 12 variables
ufo	88875 obs. of 11 variables
ufo_clean	88679 obs. of 12 variables

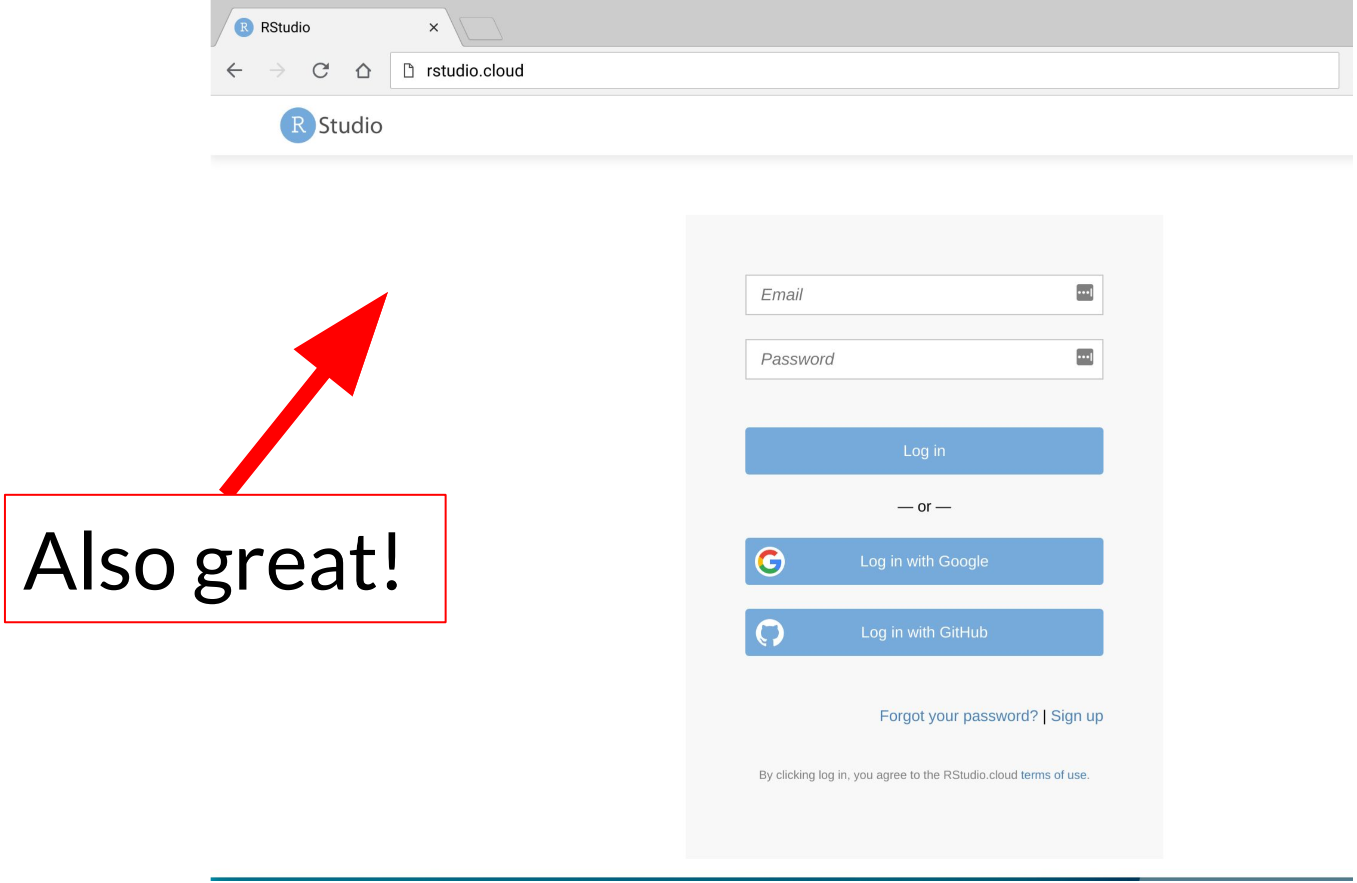
Values

A	num [1:4] 1 2 3 NA
B	num [1:4] 1 2 3 4
test	num [1:4] 0 NA -1 NaN
url	"https://sisbid.github.io/Data-Wrangl…

Documents > GitHub > Teaching > SISBID_Data-Wrangling > labs

Name	Size	Modified
..		
1000genomes.xlsx	967.1 KB	Aug 6, 2024, 1
20130606_sample_info.xlsx	967.2 KB	Aug 6, 2024, 1
advanced-io-lab-key.html	619.7 KB	Jul 15, 2025, 4
advanced-io-lab-key.R	2.5 KB	Aug 11, 2024,
advanced-io-lab-key.Rmd	2.9 KB	Jul 15, 2025, 4
advanced-io-lab.html	648.2 KB	Aug 6, 2024, 1
advanced-io-lab.R	1 KB	Aug 6, 2024, 1
advanced-io-lab.Rmd	2.1 KB	Aug 6, 2024, 1
Bike_Lanes_Long.csv	26.7 KB	Aug 6, 2024, 1
Bike_Lanes_Wide.csv	7 KB	Aug 6, 2024, 1
Bike_Lanes.csv	152.6 KB	Aug 6, 2024, 1
Circulator.csv	86.6 KB	Aug 6, 2024, 1
Circulator.rds	155.7 KB	Aug 6, 2024, 1
crashes_corrections.csv	268 B	Aug 6, 2024, 1

https://www.rstudio.com/



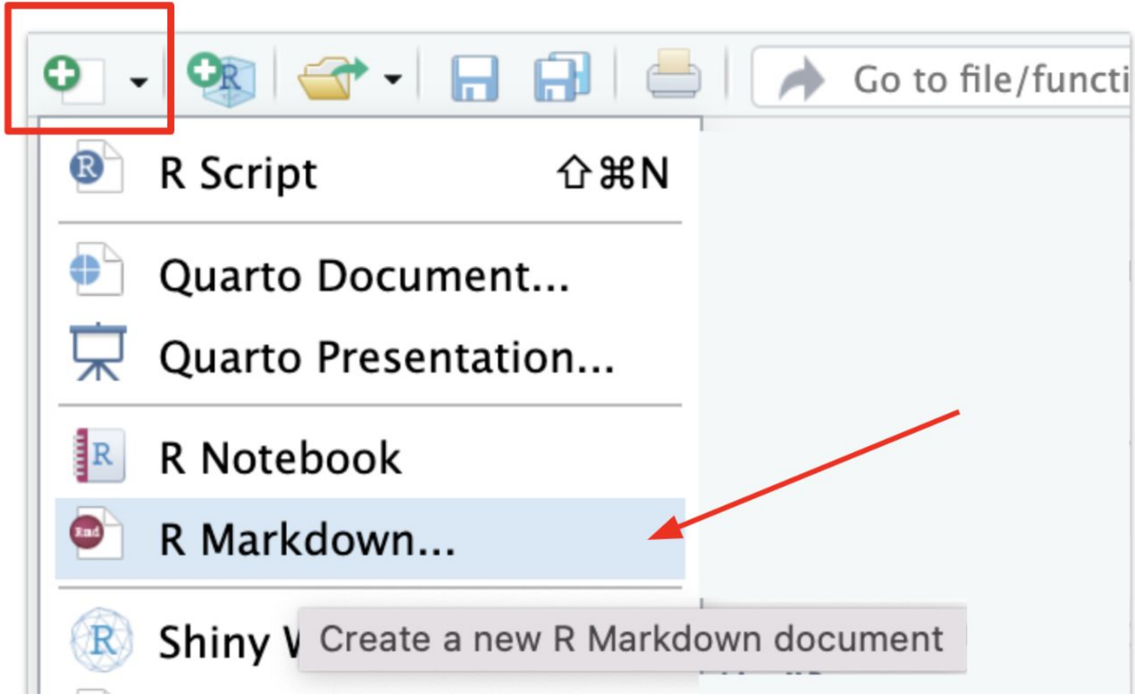
By clicking log in, you agree to the RStudio.cloud terms of use.

Also great!

https://rstudio.cloud

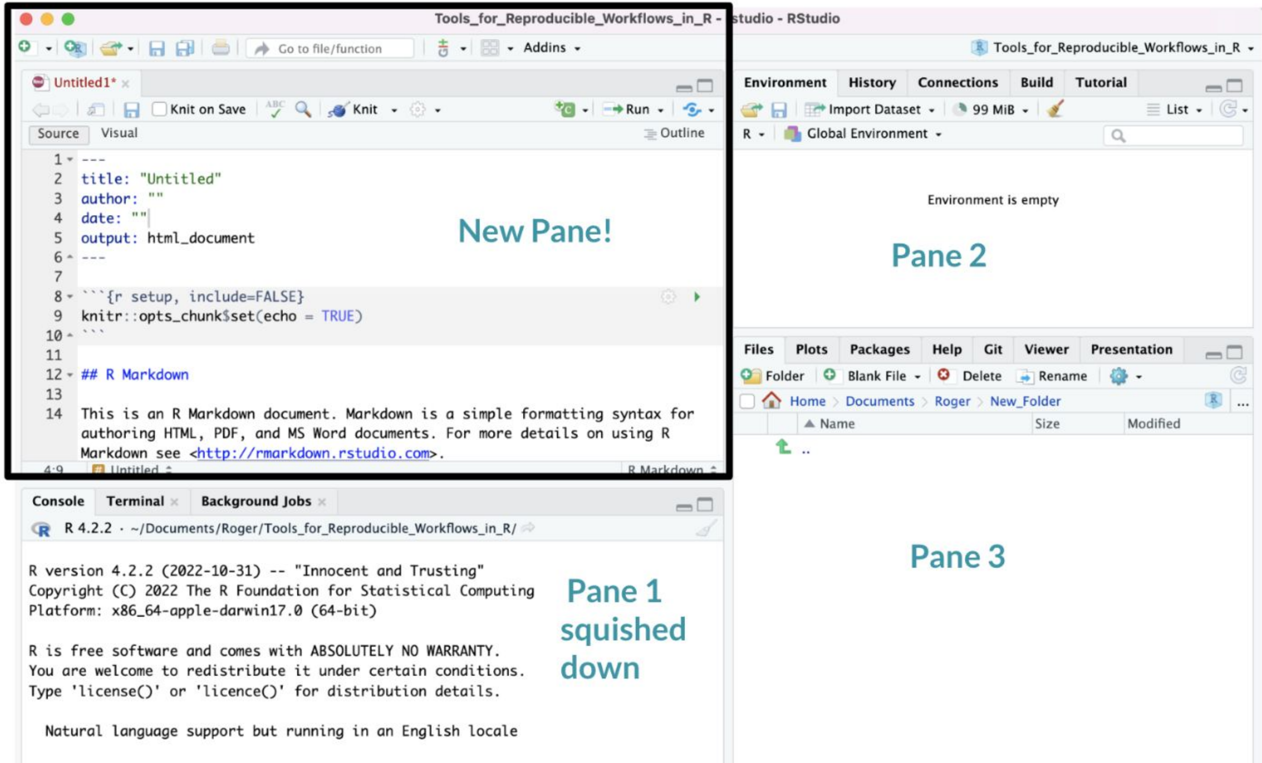
Hidden Pane

To save a copy of your code. You must open a file first - this will open a 4th pane. These files include Scripts or what are called R Markdown files.



Hidden Pane

Nice! now we have a place to save code! This is where we will mostly be working.



~~~
---
title: "Untitled"
author: ""
date: ""
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for
authoring HTML, PDF, and MS Word documents. For more details on using R
Markdown see <http://rmarkdown.rstudio.com>.
~~~

```
R version 4.2.2 (2022-10-31) -- "Innocent and Trusting"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (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
```

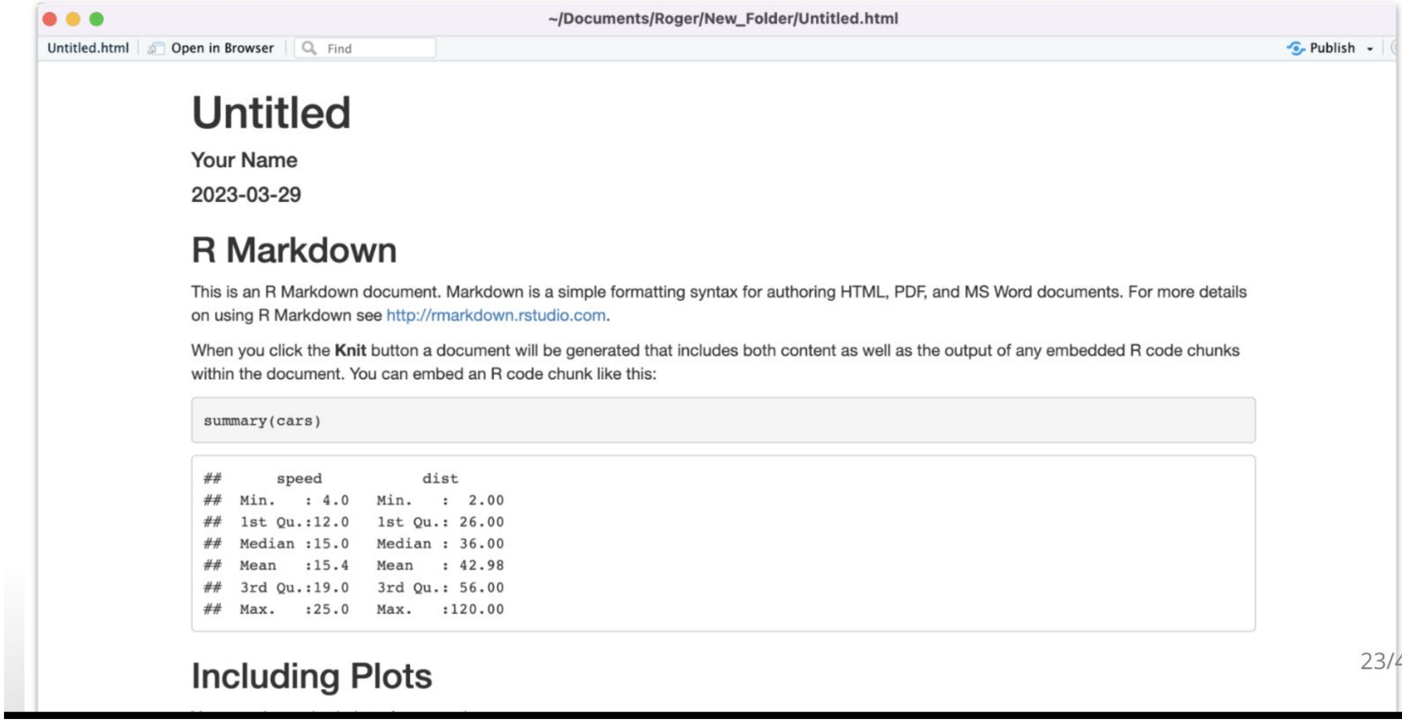
R Markdown file

R Markdown files (.Rmd) help generate reports that include your code and output.

1. Helps you describe your code
2. Allows you to check the output
3. Can create many different file types

Nice report!

This generates a nice report that you can share with others who can open in any browser.



Untitled

Your Name

2023-03-29

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed          dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00  
```

Including Plots

Summary

- Repeatable → Reproducible → Replicable
- Others know what you did well enough to use your data and code and get the same results
- Reproducibility:
  - Helps make science more efficient!
  - Helps your future self and others know what you did
    - Saving time and effort
  - Makes it easier to adjust or update your work
- Reproducible does not mean correct! But it is a good first step to help identify if your work is correct
- Organizing files, naming files consistently and informatively helps! R Markdown files do too - more about that next!