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

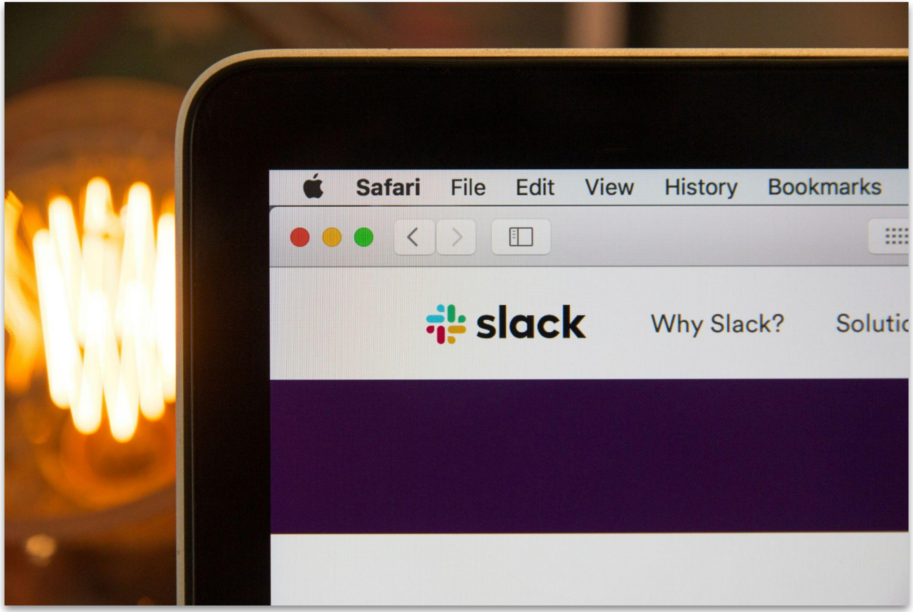


Photo by Stephen Phillips - Hostreviews.co.uk on Unsplash

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.

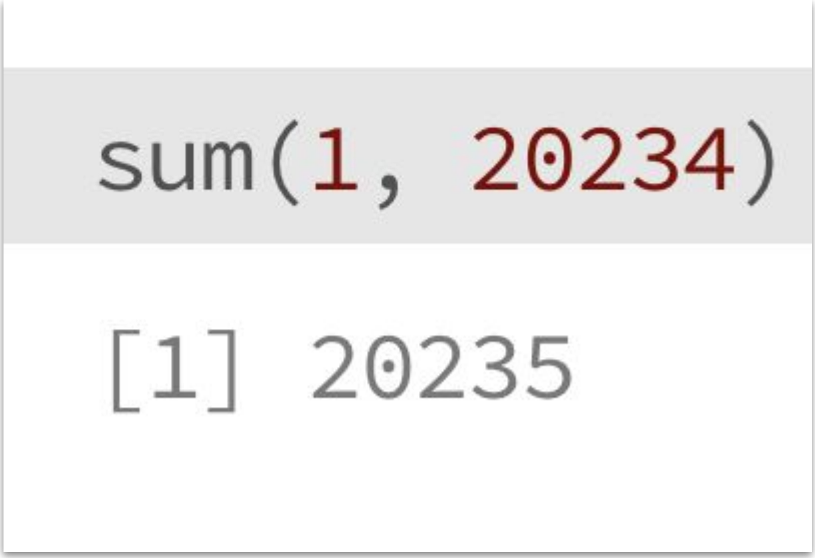
Packages are analogous to a software application like Microsoft Word on your computer.

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.

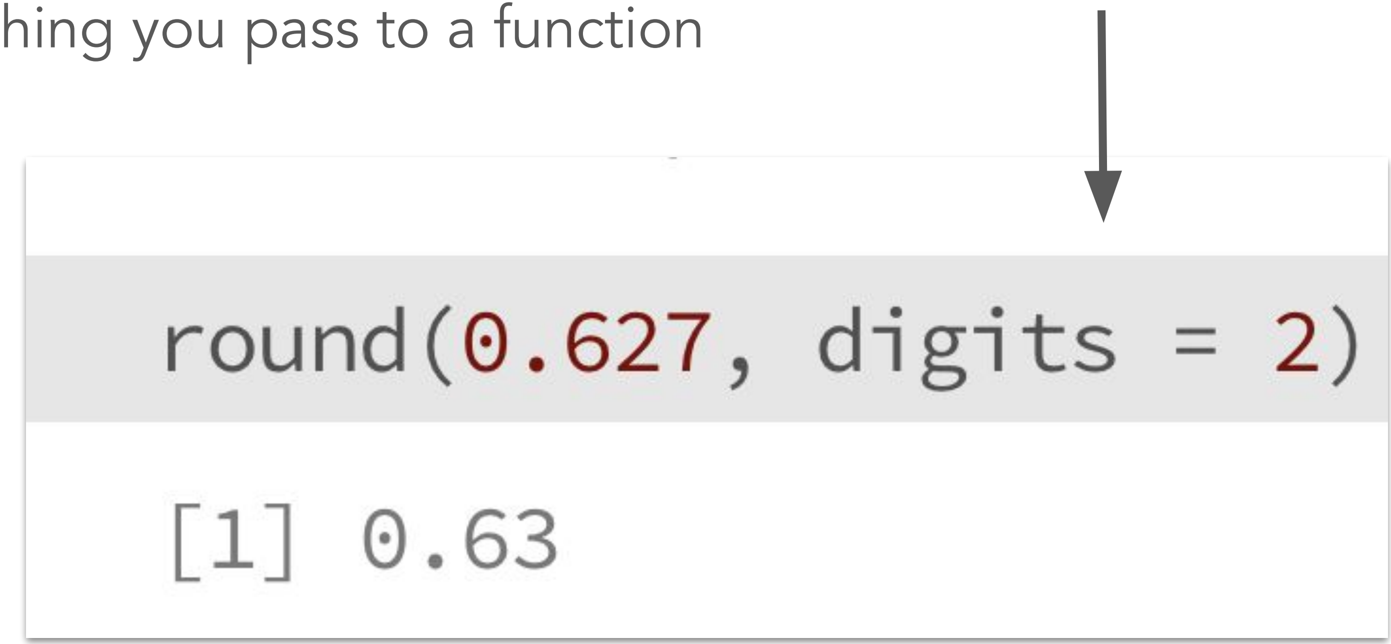
A function might help you add numbers together, create a plot, or organize your data. More on that soon!



```
sum(1, 20234)

[1] 20235
```

Argument

Something you pass to a function



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

```
[1] 0.63
```

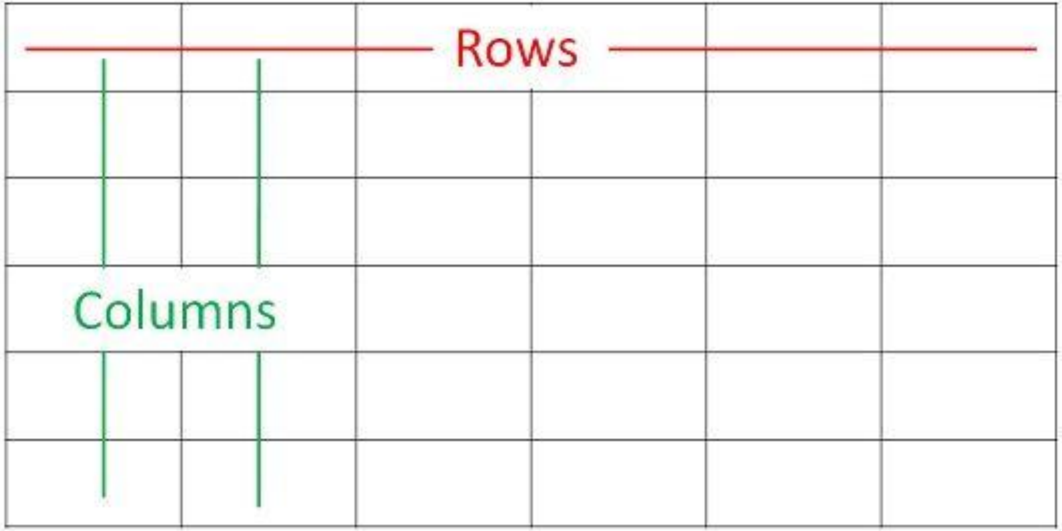
An Object

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!

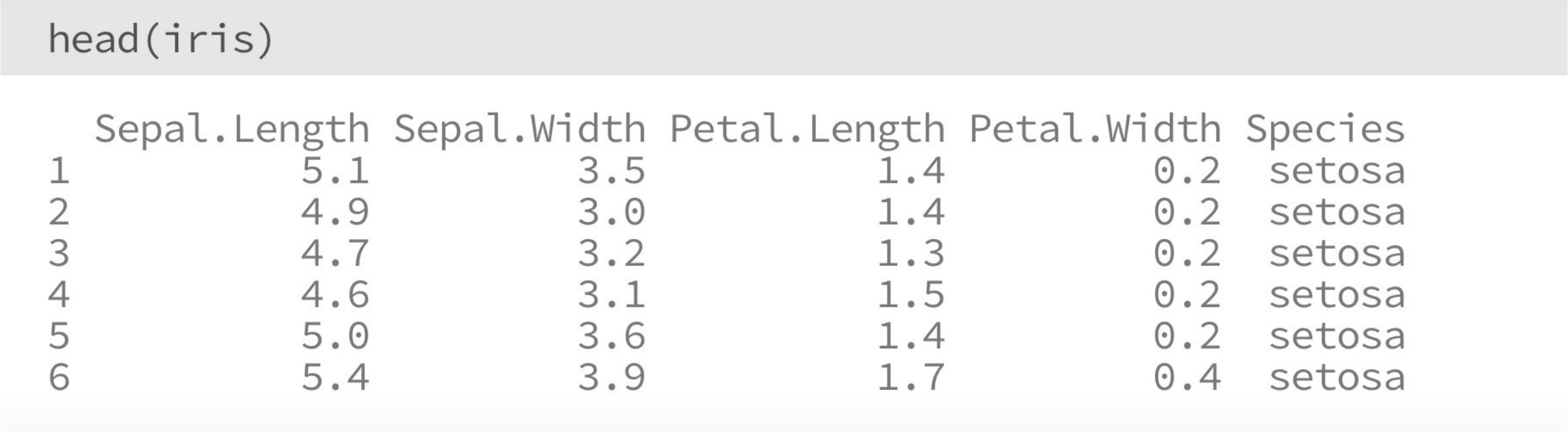
Dataframes/DataTables/Spreadsheets



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:

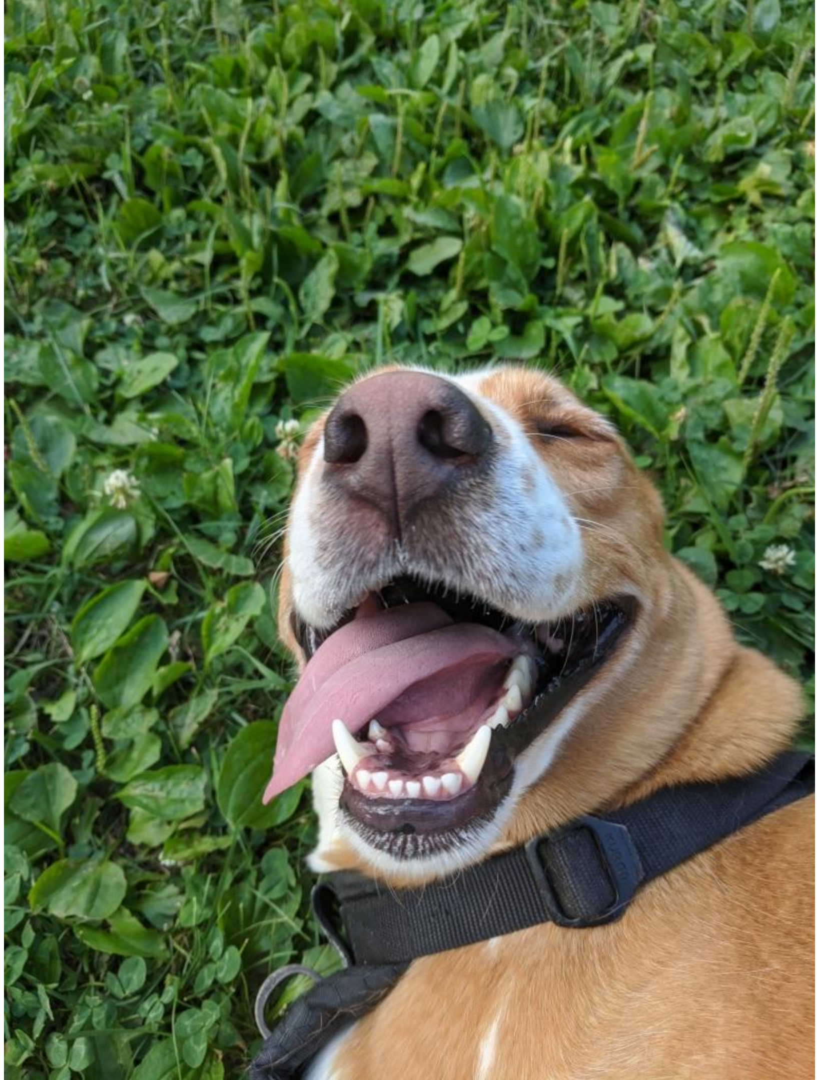
https://link.springer.com/content/pdf/bbm%3A978-1-4419-1318-0%2F1.pdf

The Tidvyerse

Newer packages designed for data science that make R code more intuitive.

How many people feel about data wrangling

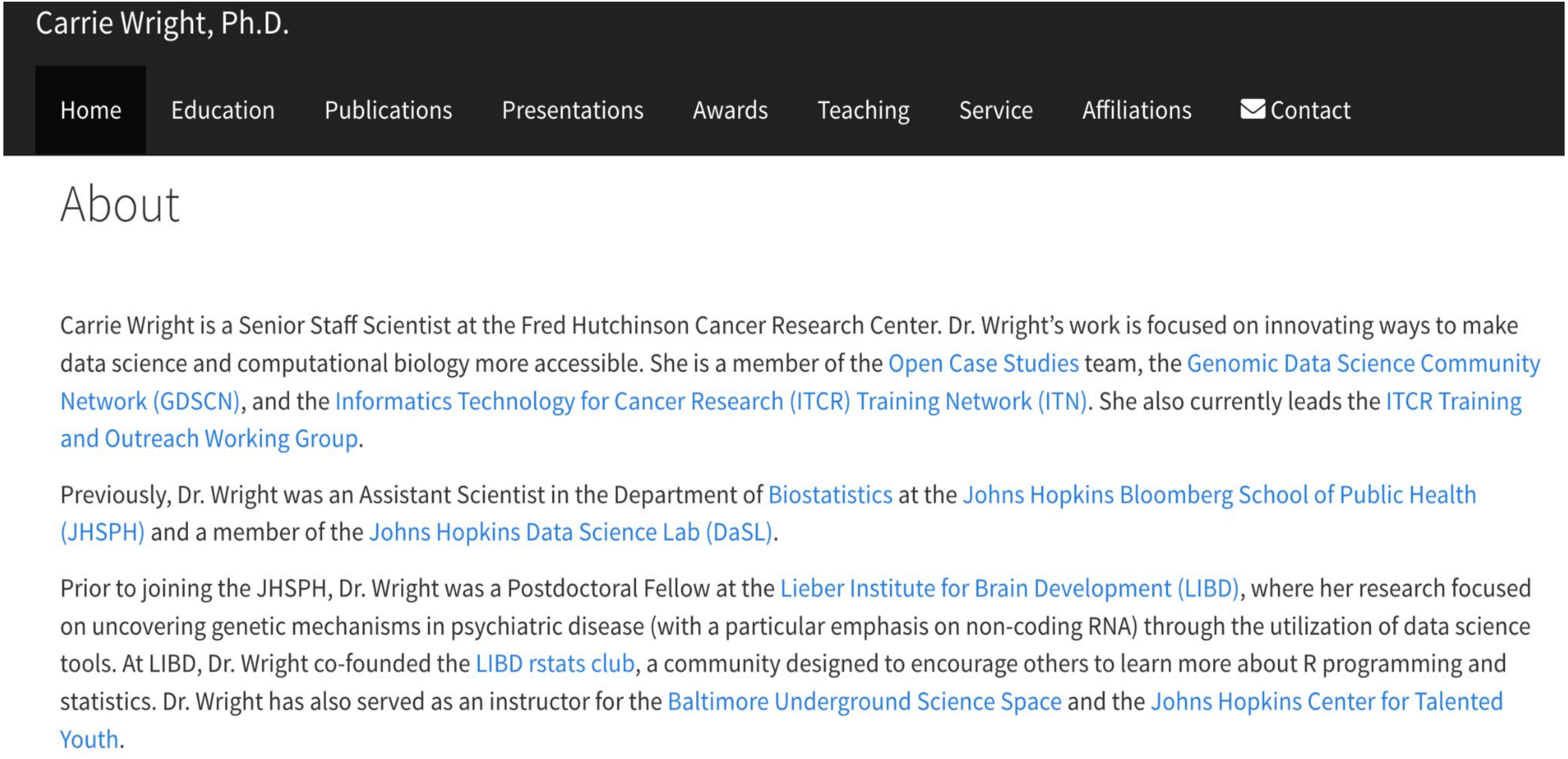
How we feel about data wrangling



Carrie

About us

Ava



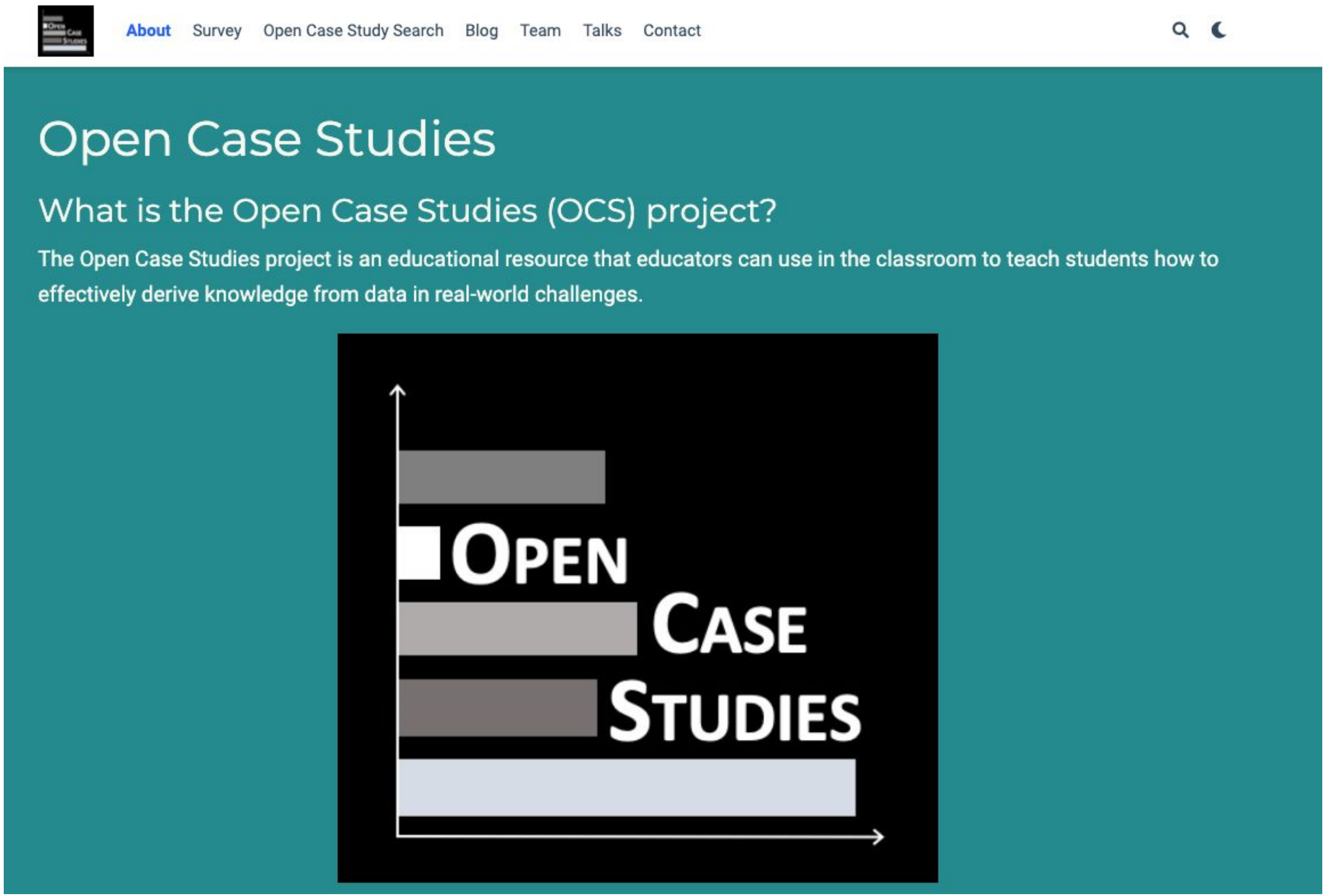
About

Carrie Wright is a Senior Staff Scientist at the Fred Hutchinson Cancer Research Center. Dr. Wright's work is focused on innovating ways to make data science and computational biology more accessible. 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 leads the ITCR Training and Outreach Working 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/

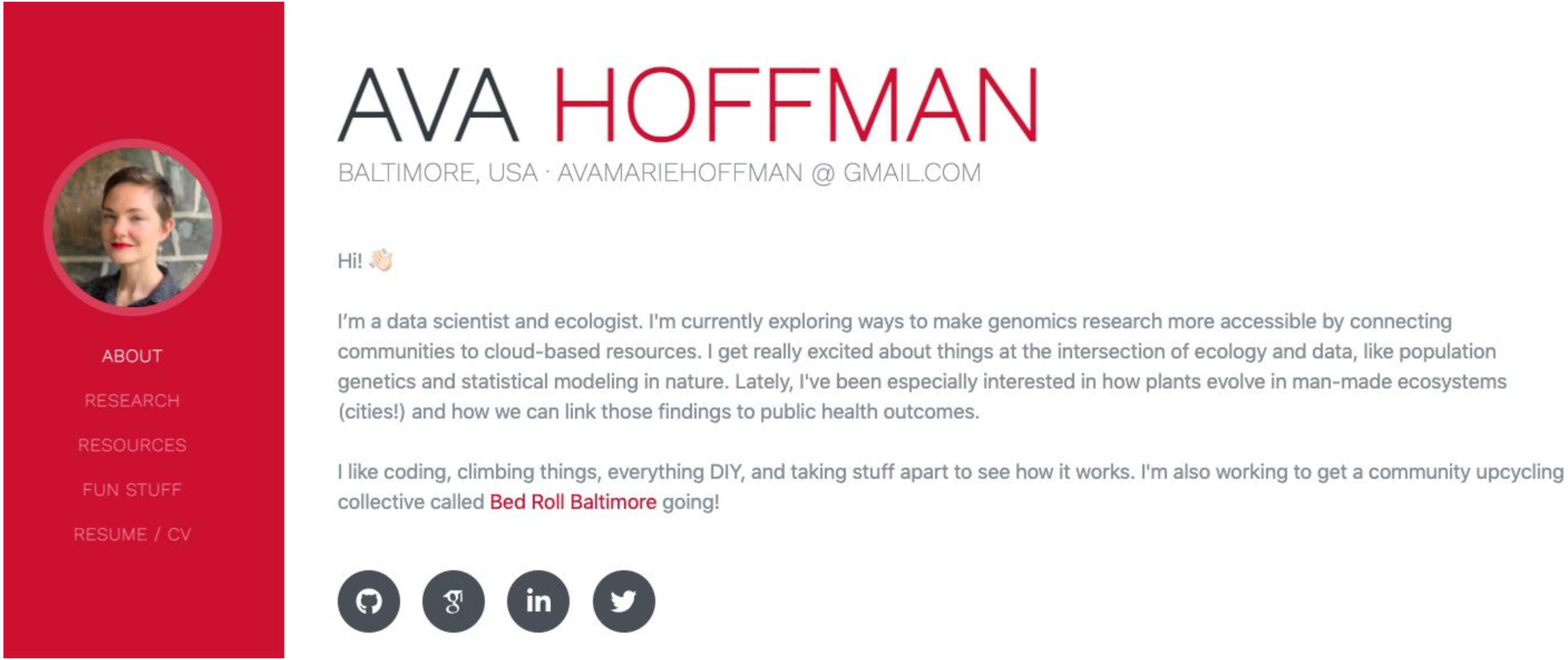


Open Case Studies

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.

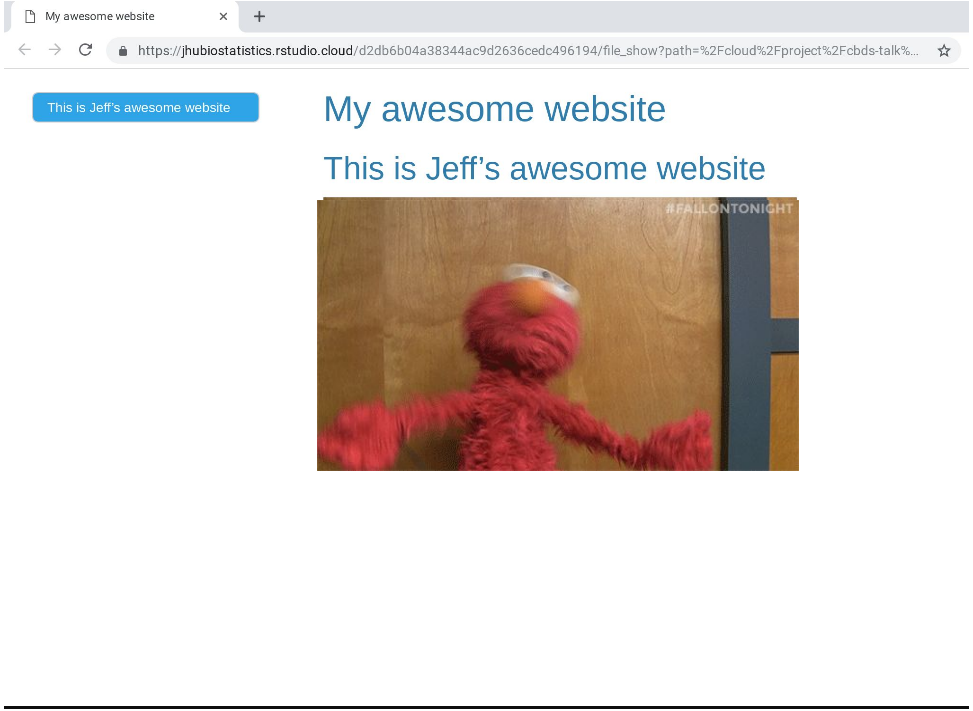
I like coding, climbing things, everything DIY, and taking stuff apart to see how it works. I'm also working to get a community upcycling collective called Bed Roll Baltimore going!

https://www.avahoffman.com/

Why this class

rmarkdown

```
---
title: "My awesome website"
output:
    html_document:
        toc: true
        toc_float: true
        theme: cerulean
---
# This is Jeff's awesome website

![](https://media.giphy.com/media/d
rXGoW1iudhKw/giphy.gif)
```


My awesome website
https://jhubiostatistics.rstudio.cloud/d2db6b04a38344ac9d2636cedc496194/file_show?path=%2Fcloud%2Fproject%2Fcbds-talk%...
This is Jeff's awesome website
#FALLONTONIGHT

My awesome website

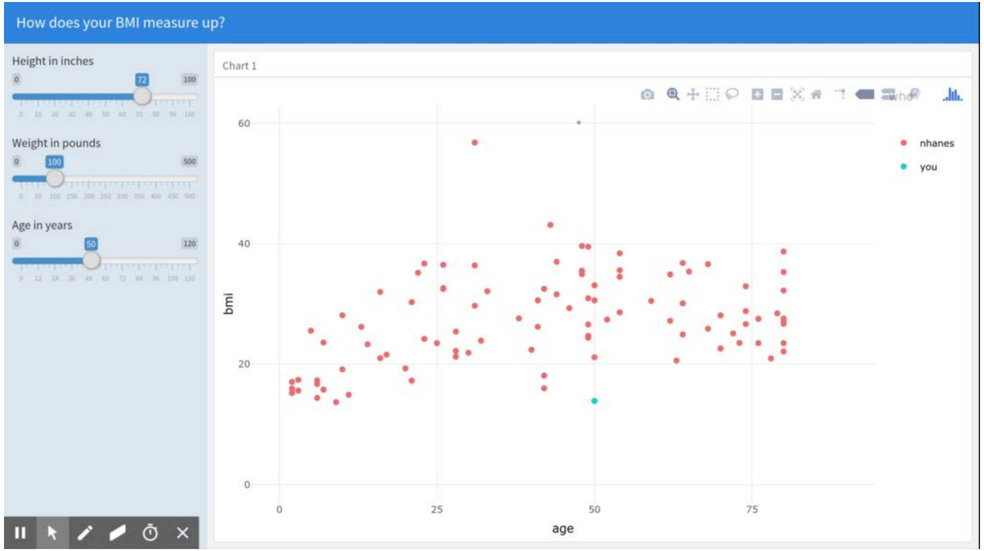
This is Jeff's awesome website

flexdashboard

~~~
---
title: "How does your BMI measure up?"
output: flexdashboard::flex_dashboard
runtime: shiny
---

Inputs {.sidebar}
-------------------------------------

```{r}
library(flexdashboard); library(NHANES); library(plotly);library(dplyr)
sliderInput("height", "Height in inches",0,100,72)
sliderInput("weight", "Weight in pounds",0,500,100)
sliderInput("age", "Age in years",0,120,50)

~~~

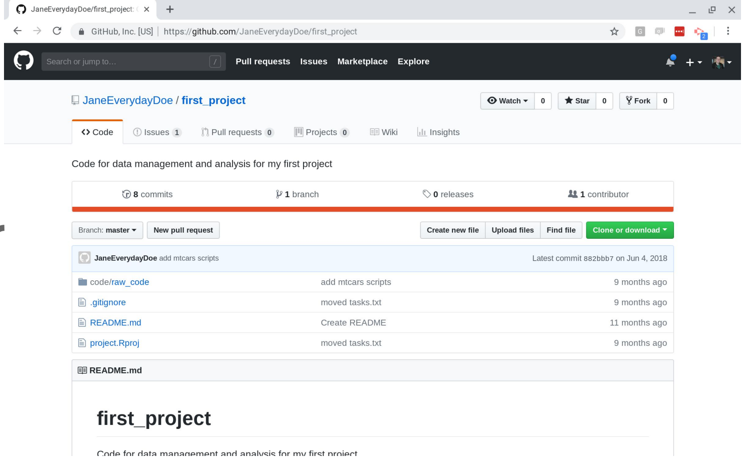
Column

### Chart 1

```{r}
nhanes = sample_n(NHANES,100)
renderPlotly({
  df = data.frame(bmi = c(nhanes$BMI,input$weight*0.45/(input$height*0.025)^2),
                  age = c(nhanes$Age,input$age),
                  who = c(rep("nhanes",100),"you"))
  ggplotly(ggplot(df) +
             geom_point(aes(x=age,y=bmi,color=who)) +
             scale_x_continuous(limits=c(0,90)) +
             scale_y_continuous(limits=c(0,60)) +
             theme_minimal()
           )
})
```
```



How does your BMI measure up?

Height in inches: 72

Weight in pounds: 100

Age in years: 50
```

# httr

```
library(httr)
library(dplyr)

username = 'janeeverydaydoe'

url_git = 'https://api.github.com/'


api_response =
GET(url = paste0(url_git, 'users/',
username, '/repos'))

content(api_response)[[1]]
```



JaneEverydayDoe / first_project

Code for data management and analysis for my first project

JaneEverydayDoe add mtcars scripts — Latest commit 882bbb7 on Jun 4, 2018

code/raw_code	add mtcars scripts	9 months ago
.gitignore	moved tasks.txt	9 months ago
README.md	Create README	11 months ago
project.Rproj	moved tasks.txt	9 months ago

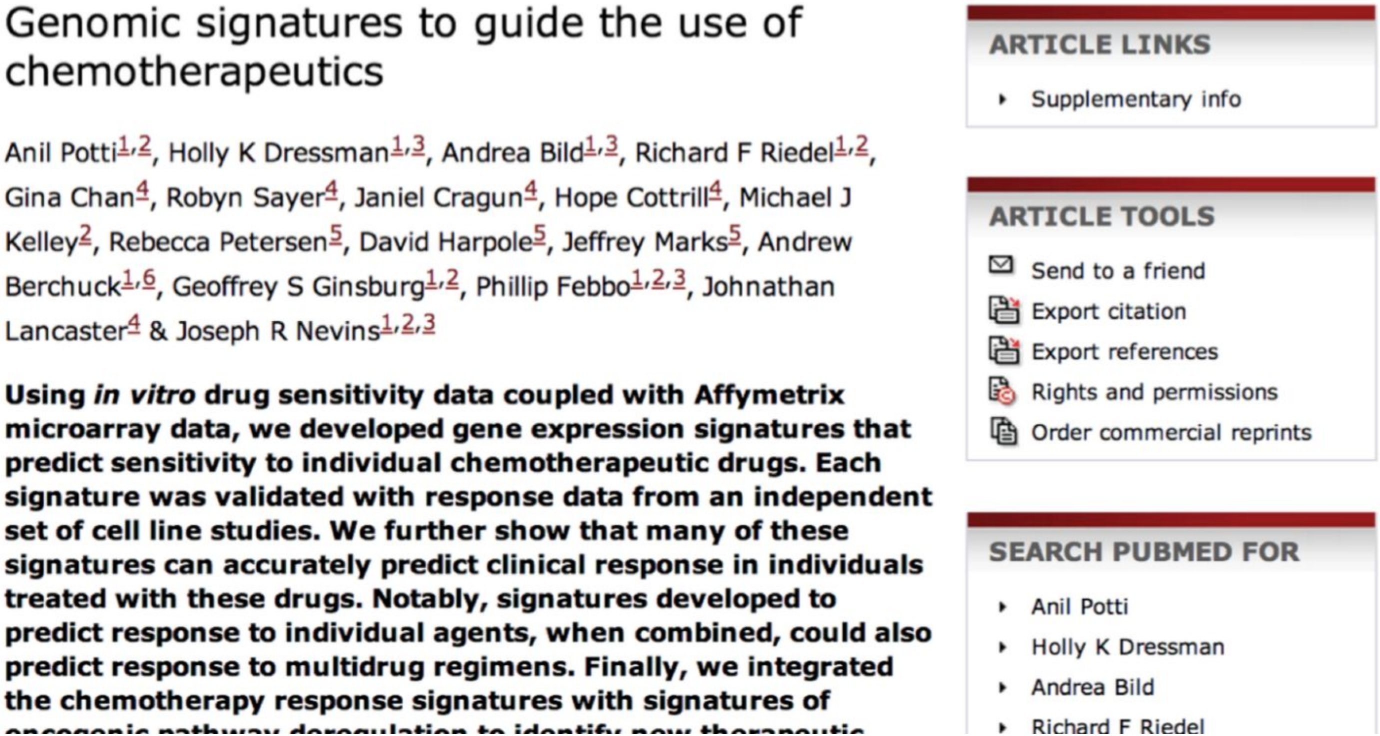
first_project

Code for data management and analysis for my first project

```
$id
[1] 130377298
$node_id
[1] "MDEwOlJlcG9zaXRvcnkxMzAzNzcyOTg="
$name
[1] "first_project"
$full_name
[1] "JaneEverydayDoe/first_project"
$owner$gravatar_id
[1] ""
$owner$url
[1] "https://api.github.com/users/JaneEverydayDoe"
```

...

But also…



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

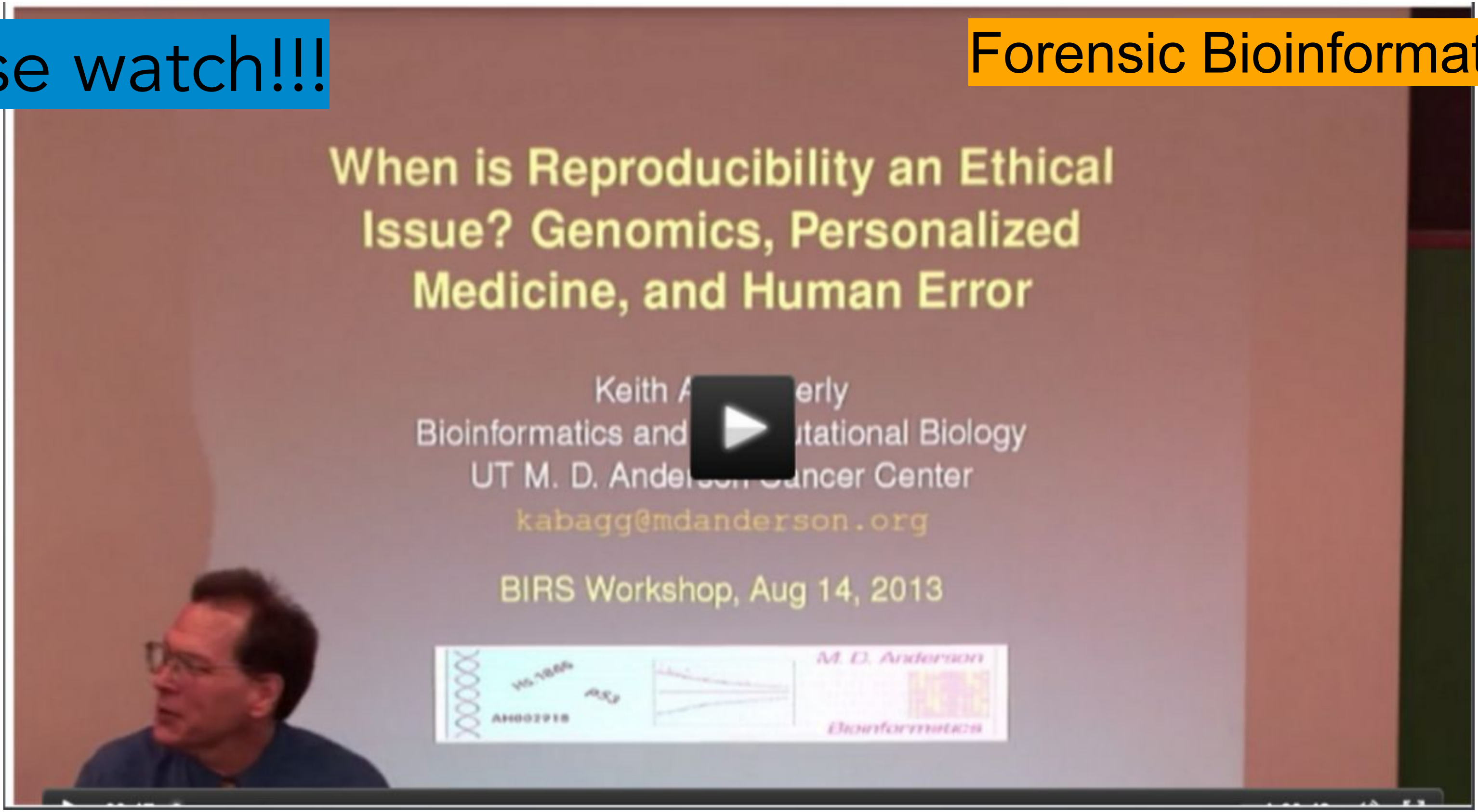
SEARCH PUBMED FOR

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

https://doi.org/10.1038/nm1491

Please watch!!!

Forensic Bioinformatics

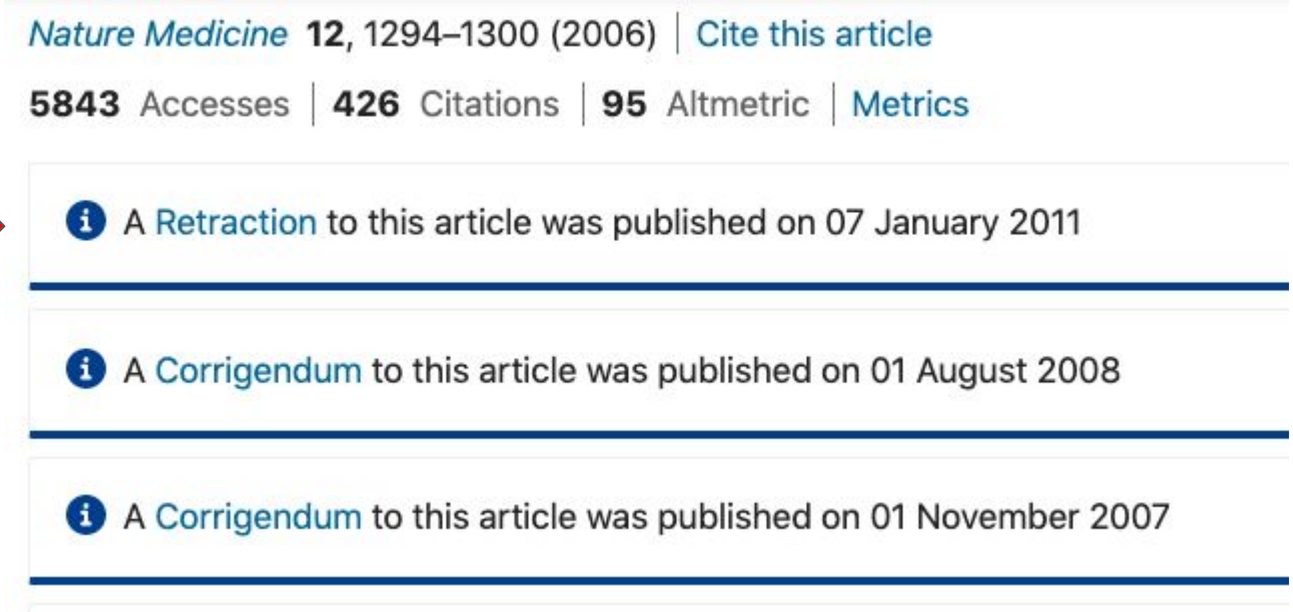


When is Reproducibility an Ethical Issue? Genomics, Personalized Medicine, and Human Error

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



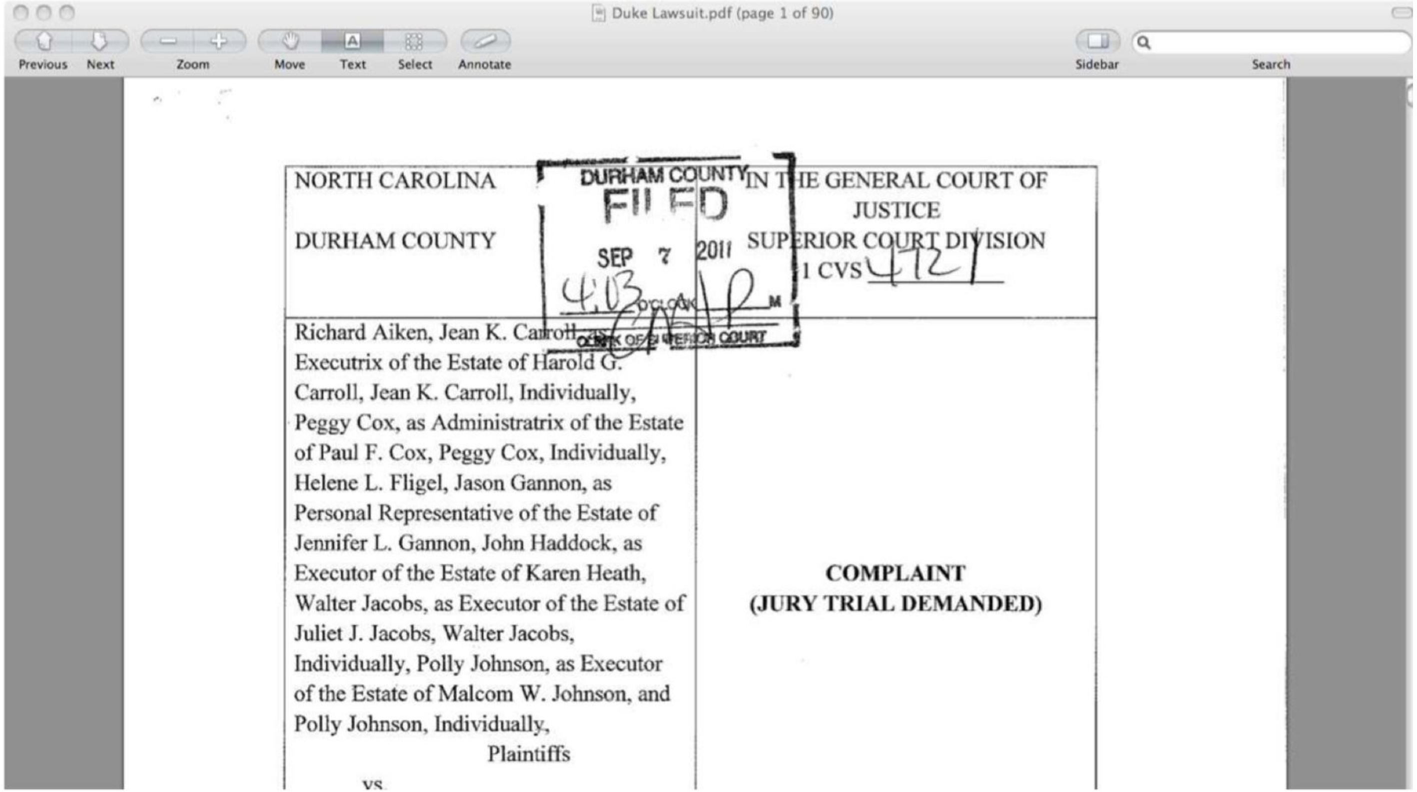
Nature Medicine **12**, 1294–1300 (2006) | Cite this article

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

A Retraction to this article was published on 07 January 2011

A Corrigendum to this article was published on 01 August 2008

A Corrigendum to this article was published on 01 November 2007

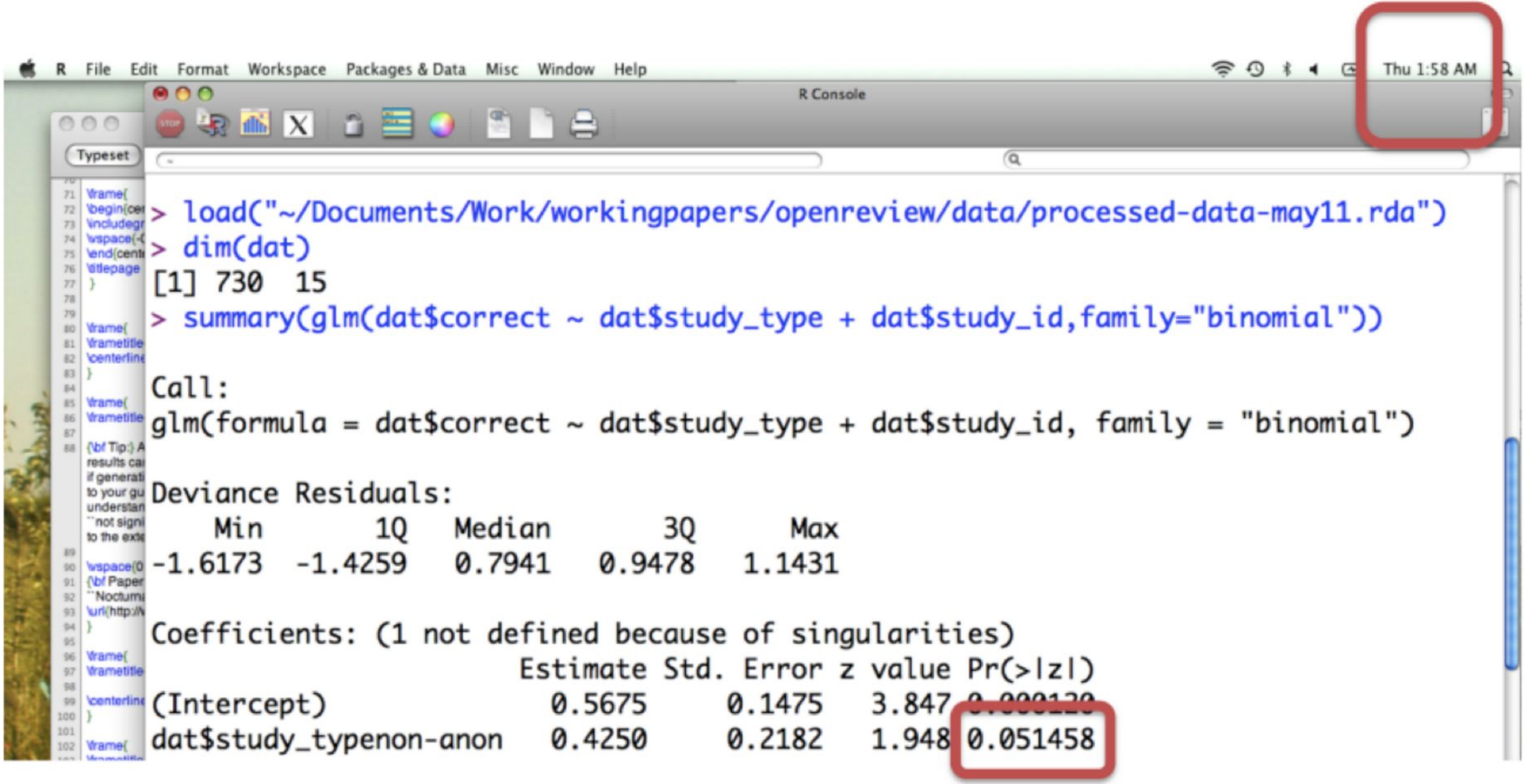


NORTH CAROLINA DURHAM COUNTY	IN THE GENERAL COURT OF JUSTICE SUPERIOR COURT DIVISION 1 CVS 4721
Richard Aiken, Jean K. Carroll, as Executrix of the Estate of Harold G. Carroll, Jean K. Carroll, Individually, Peggy Cox, as Administratrix of the Estate of Paul F. Cox, Peggy Cox, Individually, Helene L. Fligel, Jason Gannon, as Personal Representative of the Estate of Jennifer L. Gannon, John Haddock, as Executor of the Estate of Karen Heath, Walter Jacobs, as Executor of the Estate of Juliet J. Jacobs, Walter Jacobs, Individually, Polly Johnson, as Executor of the Estate of Malcom W. Johnson, and Polly Johnson, Individually, Plaintiffs vs.	**COMPLAINT** **(JURY TRIAL DEMANDED)**

DURHAM COUNTY
FILED
SEP 7 2011
4:03 O'CLOCK P M
CLERK OF SUPERIOR COURT

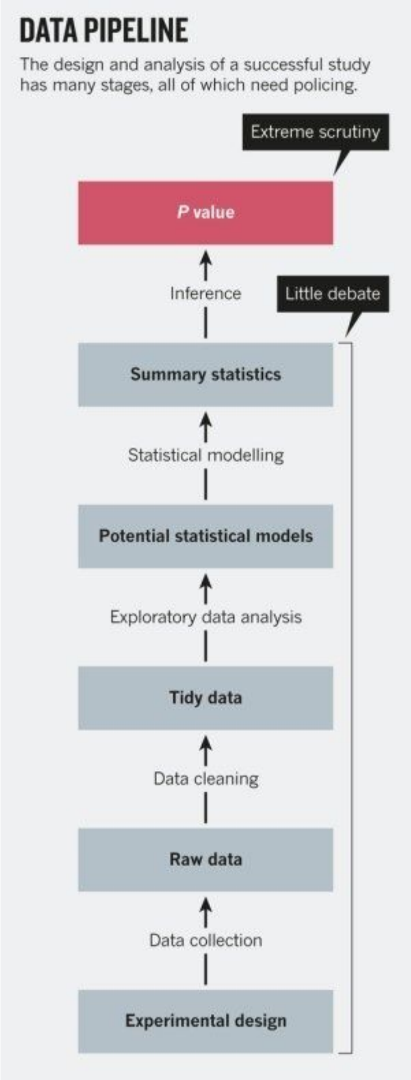
https://www.dukechronicle.com/article/2015/05/duke-lawsuit-involving-cancer-patients-linked-anil-potti-settled

Doesn’t seem that important….



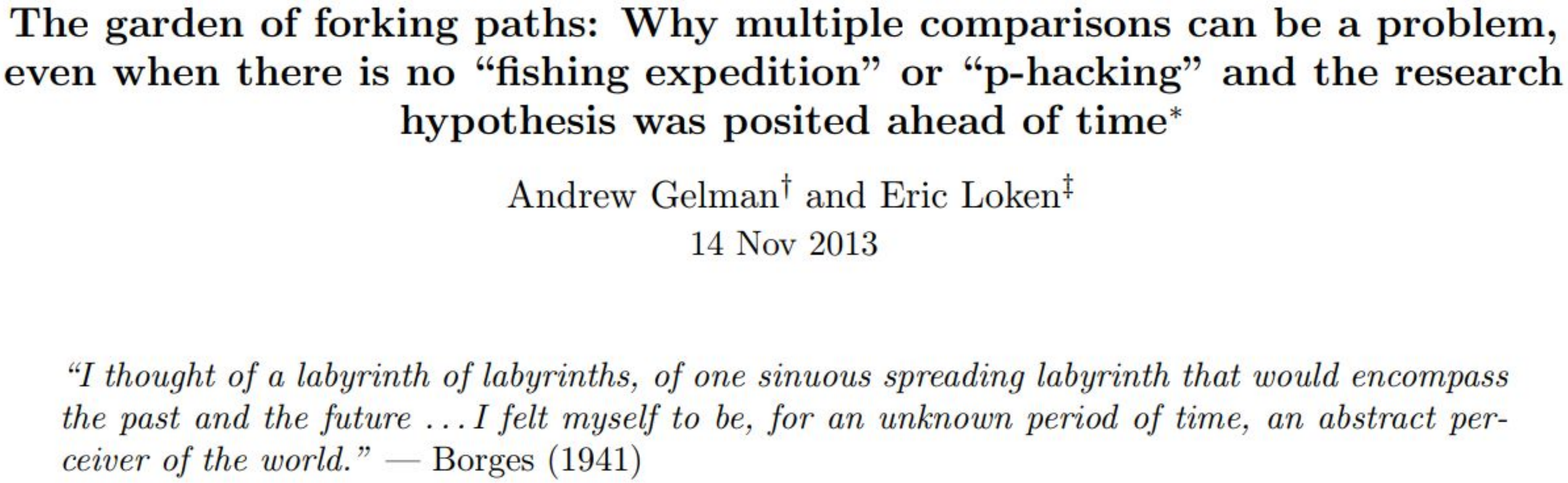
```
> load("~/Documents/Work/workingpapers/openreview/data/processed-data-may11.rda")
> dim(dat)
[1] 730  15
> summary(glm(dat$correct ~ dat$study_type + dat$study_id,family="binomial"))

Call:
glm(formula = dat$correct ~ dat$study_type + dat$study_id, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6173  -1.4259   0.7941   0.9478   1.1431

Coefficients: (1 not defined because of singularities)
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)               0.5675     0.1475   3.847 0.000120
dat$study_typenon-anon    0.4250     0.2182   1.948 0.051458
```

ON THE ONE
HAND...



- 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



The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time*

Andrew Gelman[†] and Eric Loken[‡]

14 Nov 2013

> *“I thought of a labyrinth of labyrinths, of one sinuous spreading labyrinth that would encompass the past and the future . . . I felt myself to be, for an unknown period of time, an abstract perceiver of the world.”* — Borges (1941)

Explorable Multiverse Analyses

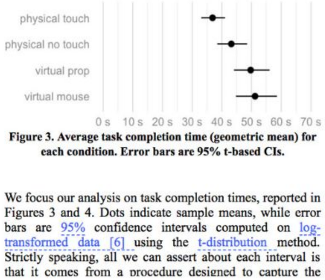
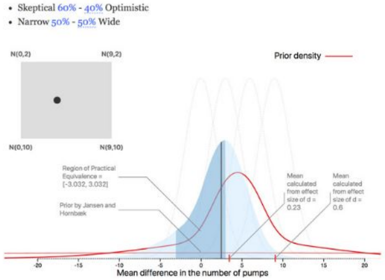
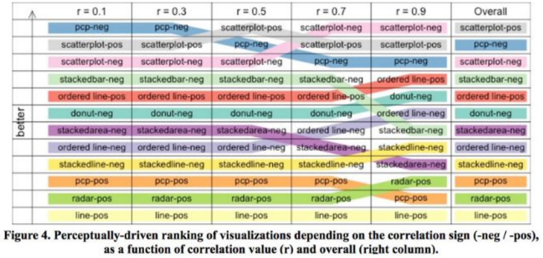


Figure 3. Average task completion time (geometric mean) for each condition. Error bars are 95% t-based CIs.

We focus our analysis on task completion times, reported in Figures 3 and 4. Dots indicate sample means, while error bars are 95% confidence intervals computed on log-transformed data [6] using the t-distribution method. Strictly speaking, all we can assert about each interval is that it comes from a procedure designed to capture the

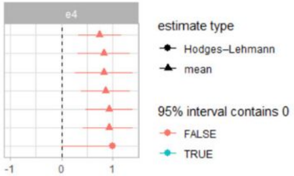


- Skeptical 60% - 40% Optimistic
- Narrow 50% - 50% Wide



r = 0.1	r = 0.3	r = 0.5	r = 0.7	r = 0.9	Overall
pcp-neg	pcp-neg	scatterplot-pos	scatterplot-neg	scatterplot-neg	scatterplot-pos
scatterplot-pos	scatterplot-pos	pcp-neg	scatterplot-pos	scatterplot-pos	pcp-neg
scatterplot-neg	scatterplot-neg	scatterplot-neg	pcp-neg	pcp-neg	scatterplot-neg
stackedbar-neg	stackedbar-neg	stackedbar-neg	stackedbar-neg	ordered line-pos	stackedbar-neg
ordered line-pos	ordered line-pos	ordered line-pos	ordered line-pos	donut-neg	ordered line-pos
donut-neg	donut-neg	donut-neg	donut-neg	ordered line-neg	donut-neg
stackedarea-neg	stackedarea-neg	stackedarea-neg	ordered line-neg	stackedbar-neg	stackedarea-neg
ordered line-neg	ordered line-neg	ordered line-neg	stackedarea-neg	stackedline-neg	ordered line-neg
stackedline-neg	stackedline-neg	stackedline-neg	stackedline-neg	stackedarea-neg	stackedline-neg
pcp-pos	pcp-pos	pcp-pos	pcp-pos	radar-pos	pcp-pos
radar-pos	radar-pos	radar-pos	radar-pos	pcp-pos	radar-pos
line-pos	line-pos	line-pos	line-pos	line-pos	line-pos

Figure 4. Perceptually-driven ranking of visualizations depending on the correlation sign (-neg / -pos), as a function of correlation value (r) and overall (right column).



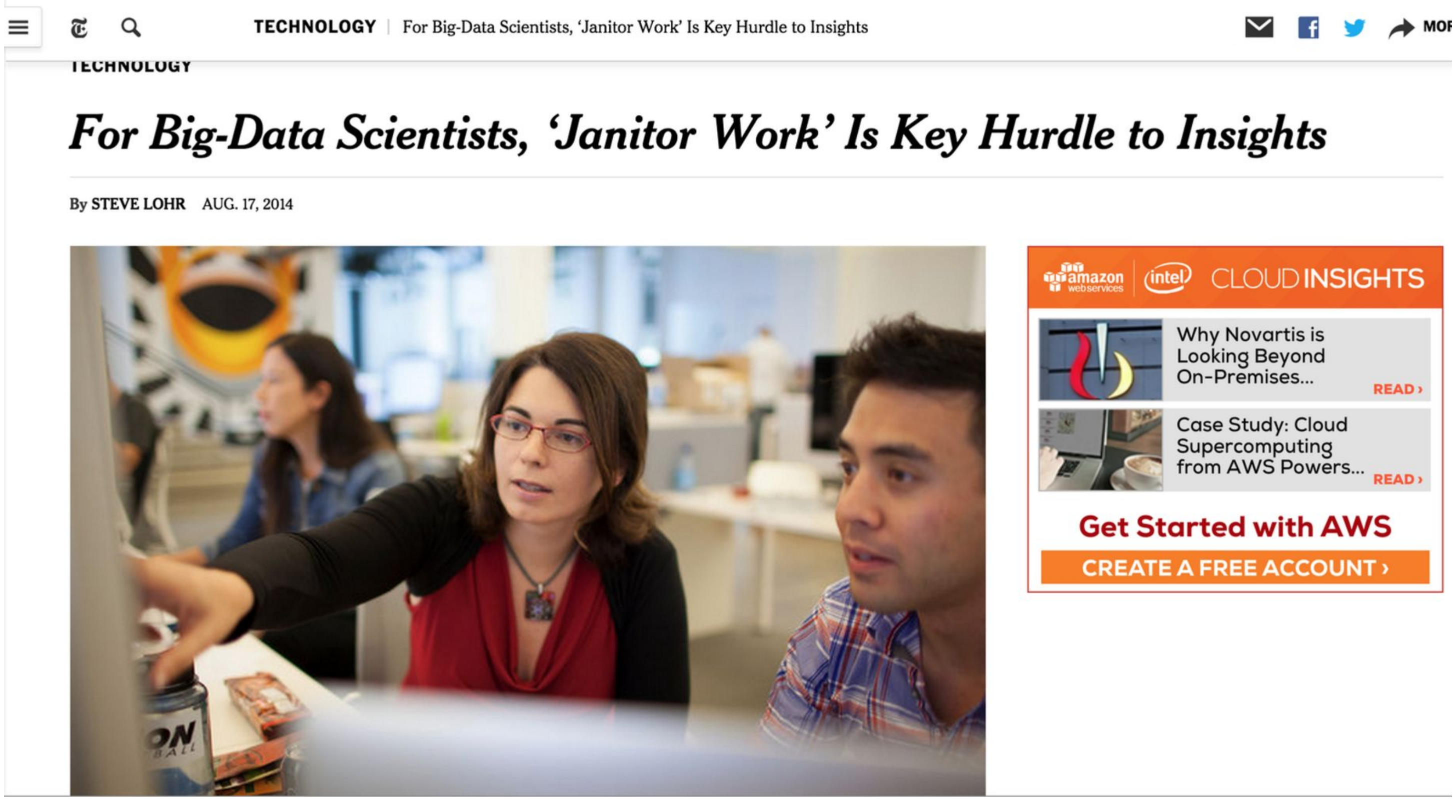
Pierre Dragicevic (Inria), Yvonne Jansen (CNRS - Sorbonne Université), Abhraneel Sarma (University of Michigan)

Matthew Kay (University of Michigan), Fanny Chevalier (University of Toronto)

With **explorable multiverse analysis reports**, readers of research papers can explore alternative analysis options by interacting with the paper itself. This new approach to statistical reporting draws from two recent ideas: multiverse analysis, a philosophy of statistical reporting where paper authors report the outcomes of many different statistical analyses in order to show how fragile or robust their findings are; and explorable explanations, narratives that can be read as normal explanations but where the reader can also become active by dynamically changing some elements of the explanation.

https://explorablemultiverse.github.io/

And so we data wrangle



TECHNOLOGY | For Big-Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights
MOF

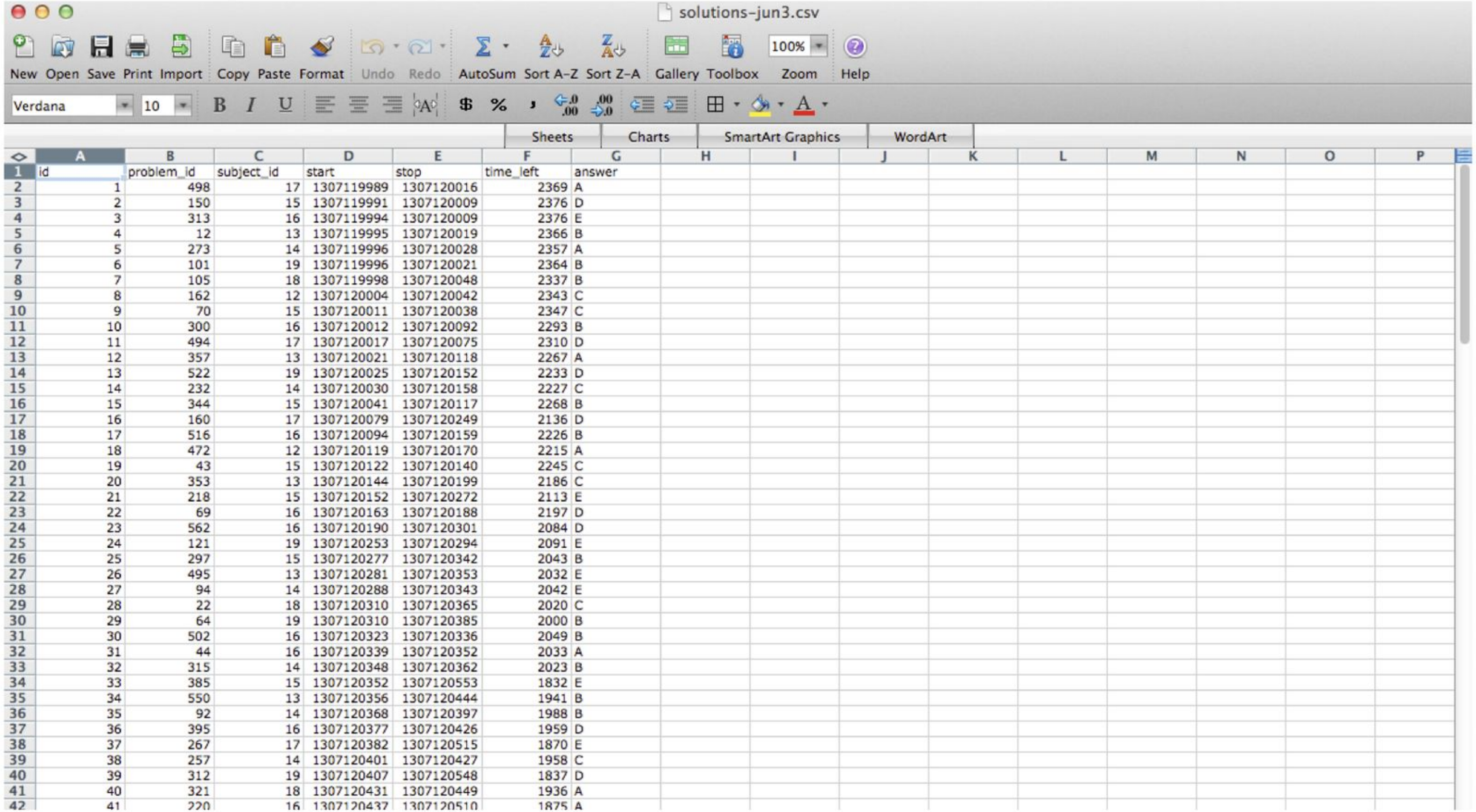
TECHNOLOGY

For Big-Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights

By STEVE LOHR AUG. 17, 2014

http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html?_r=0

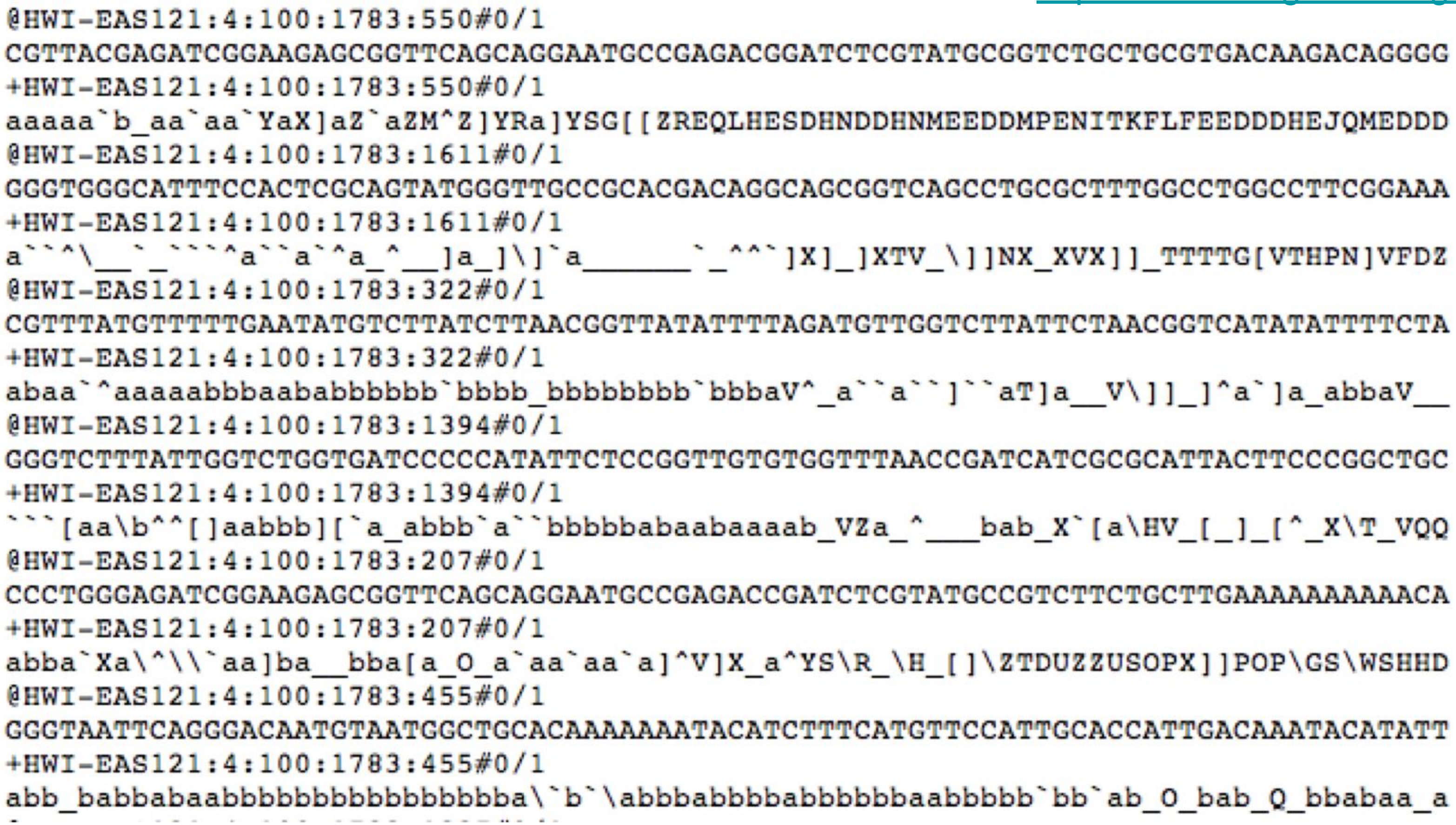
What you wished data looked like



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

What it actually looks like

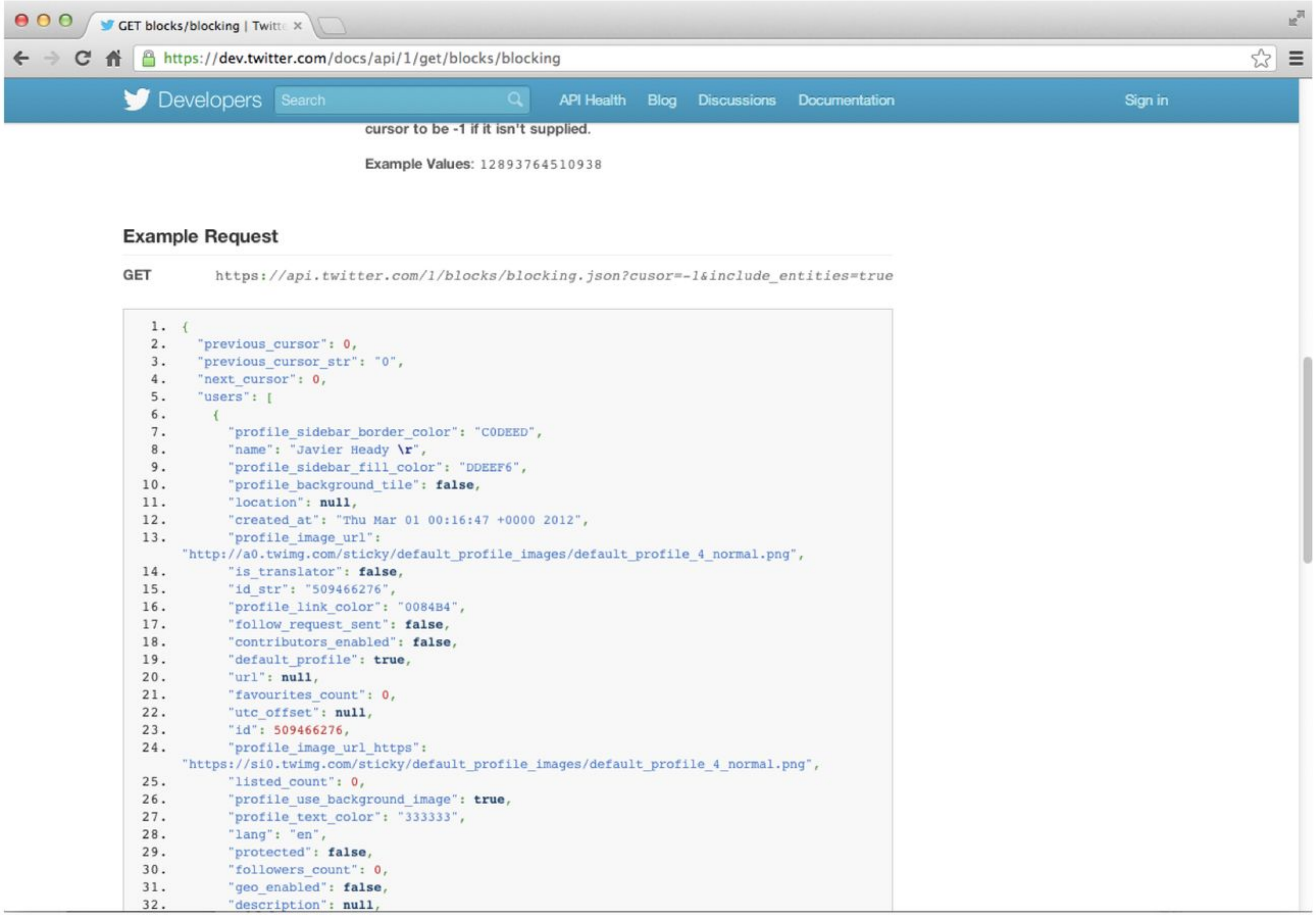
http://healthdesignchallenge.com/



~~~
@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_babbabaabbbbbbbbbbbbbbba\`b`\abbbabbbbabbbbbbaabbbbb`bb`ab_O_bab_Q_bbabaa_a
~~~

What it actually looks like

https://dev.twitter.com/docs/api/1/get/blocks/blocking


GET blocks/blocking | Twitte ×
https://dev.twitter.com/docs/api/1/get/blocks/blocking
Developers
Search
API Health
Blog
Discussions
Documentation
Sign in

cursor to be -1 if it isn't supplied.

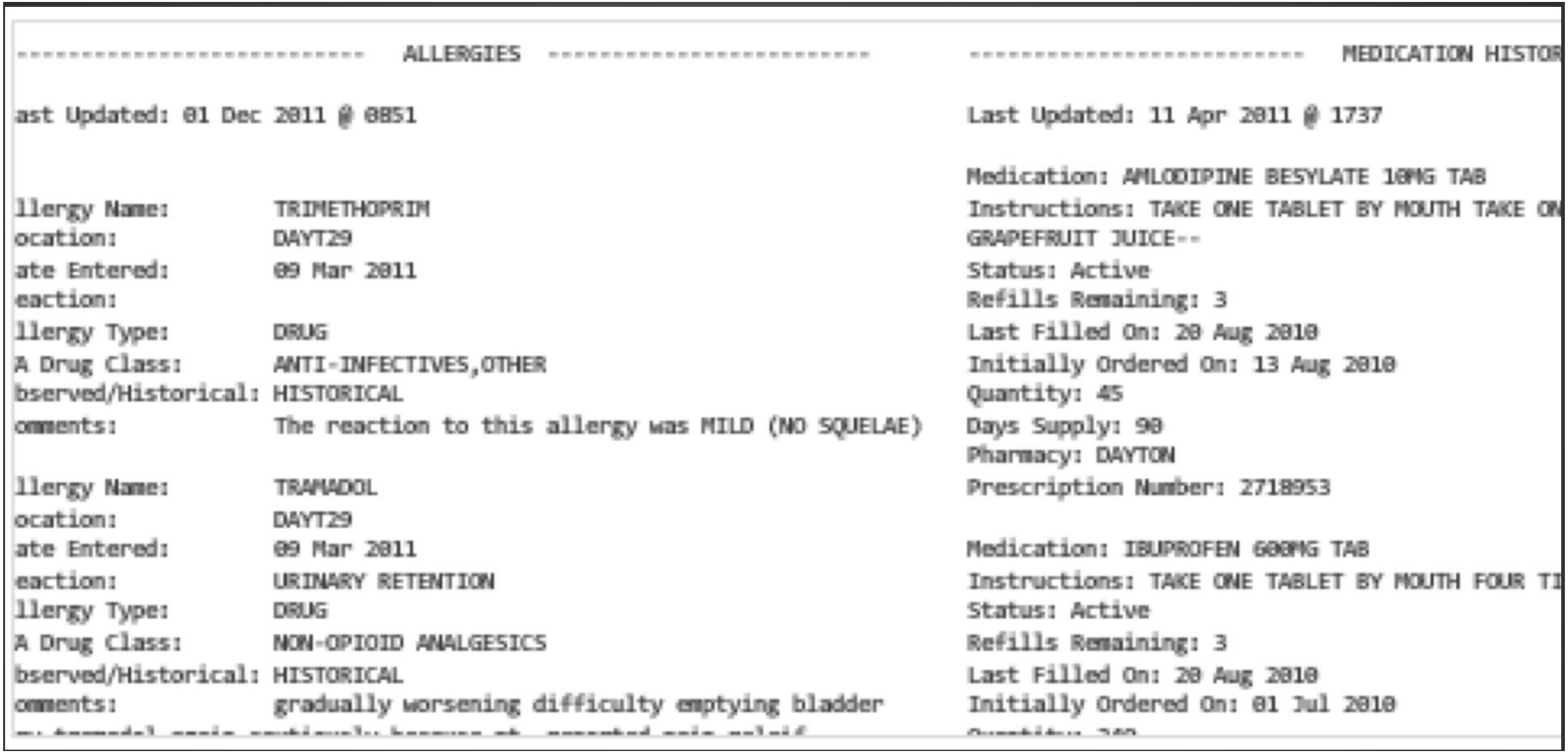
Example Values: 12893764510938

Example Request

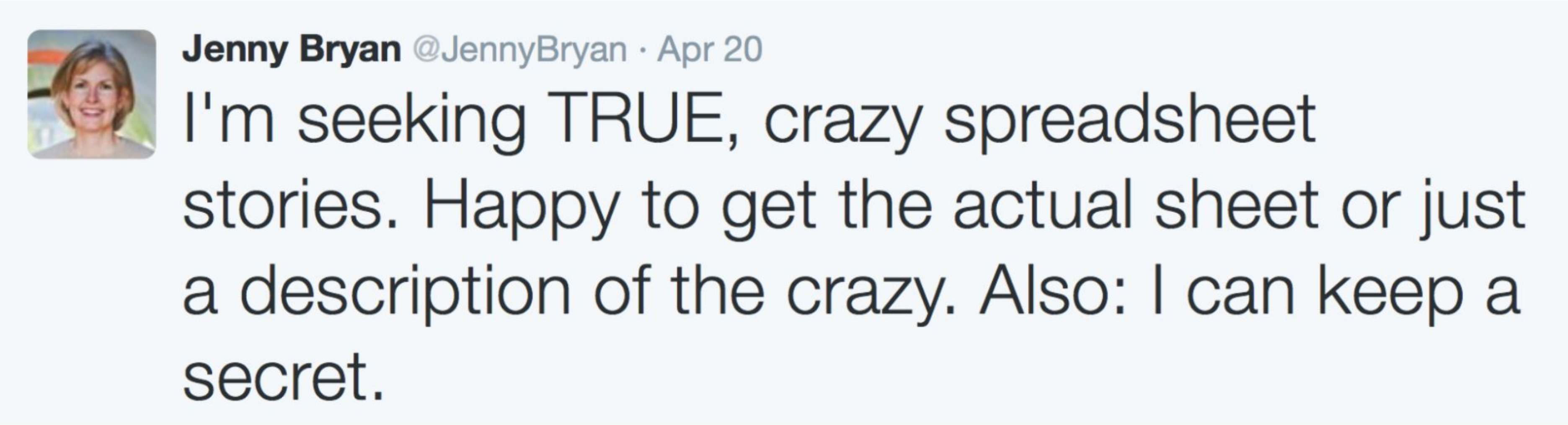
GET https://api.twitter.com/1/blocks/blocking.json?cusor=-1&include_entities=true

```
{
  "previous_cursor": 0,
  "previous_cursor_str": "0",
  "next_cursor": 0,
  "users": [
    {
      "profile_sidebar_border_color": "C0DEED",
      "name": "Javier Heady \r",
      "profile_sidebar_fill_color": "DDEEF6",
      "profile_background_tile": false,
      "location": null,
      "created_at": "Thu Mar 01 00:16:47 +0000 2012",
      "profile_image_url":
"http://a0.twimg.com/sticky/default_profile_images/default_profile_4_normal.png",
      "is_translator": false,
      "id_str": "509466276",
      "profile_link_color": "0084B4",
      "follow_request_sent": false,
      "contributors_enabled": false,
      "default_profile": true,
      "url": null,
      "favourites_count": 0,
      "utc_offset": null,
      "id": 509466276,
      "profile_image_url_https":
"https://si0.twimg.com/sticky/default_profile_images/default_profile_4_normal.png",
      "listed_count": 0,
      "profile_use_background_image": true,
      "profile_text_color": "333333",
      "lang": "en",
      "protected": false,
      "followers_count": 0,
      "geo_enabled": false,
      "description": null,
```

What it actually looks like



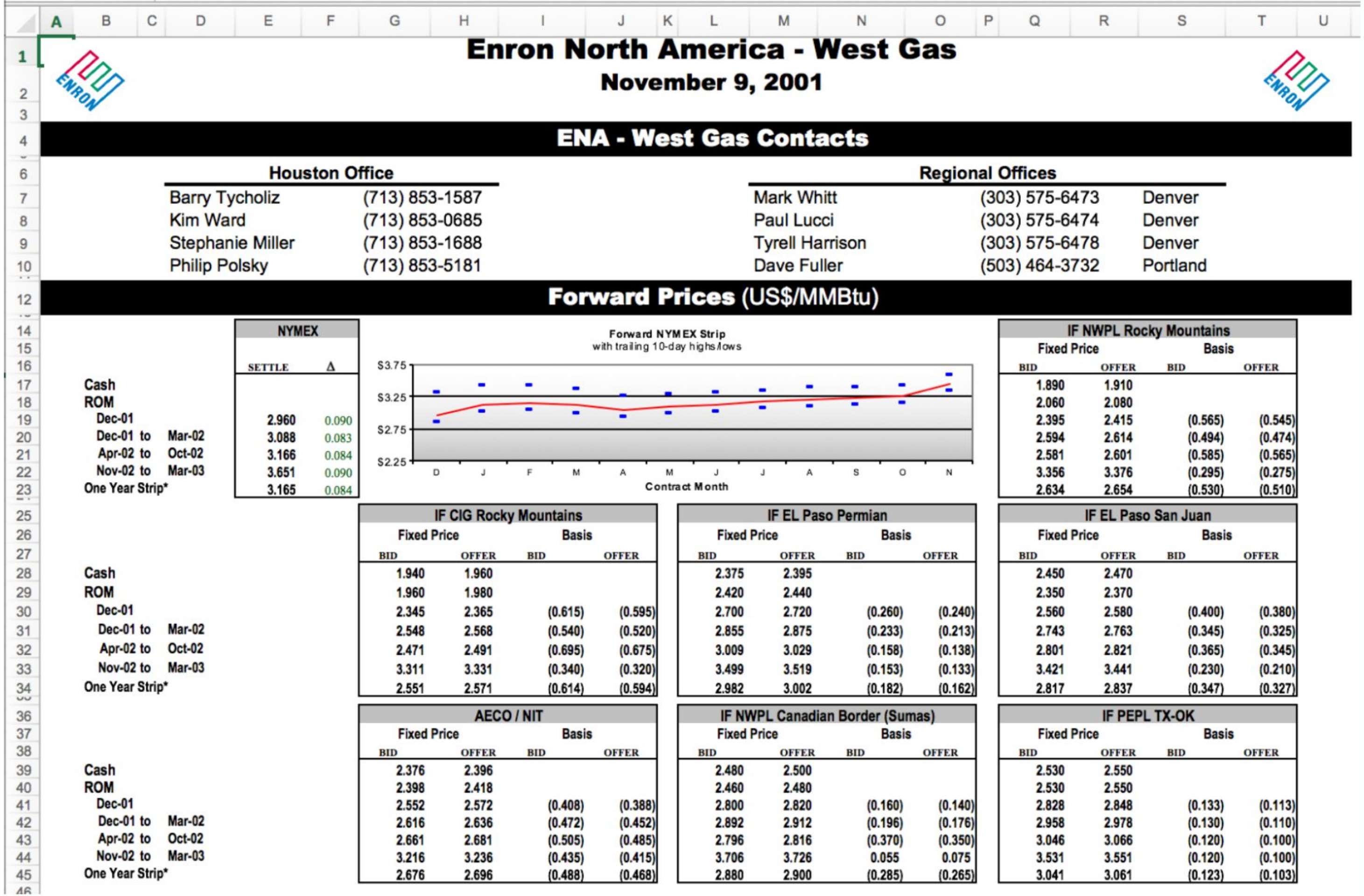
```
-------------------------   ALLERGIES  -----------------------        -------------------------   MEDICATION HISTOR

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
rx tramadol again cautiously because pt. reported pain relief       Quantity: 240
```



Jenny Bryan @JennyBryan · Apr 20

I'm seeking TRUE, crazy spreadsheet stories. Happy to get the actual sheet or just a description of the crazy. Also: I can keep a secret.

Slide from Jenny Bryan
(https://github.com/jennybc/2016-06_spreadsheets/blob/master/2016-06_useR-stanford.pdf)



Enron North America - West Gas

November 9, 2001

ENA - West Gas Contacts

Houston Office	
Barry Tycholiz	(713) 853-1587
Kim Ward	(713) 853-0685
Stephanie Miller	(713) 853-1688
Philip Polsky	(713) 853-5181

Regional Offices		
Mark Whitt	(303) 575-6473	Denver
Paul Lucci	(303) 575-6474	Denver
Tyrell Harrison	(303) 575-6478	Denver
Dave Fuller	(503) 464-3732	Portland

Forward Prices (US$/MMBtu)

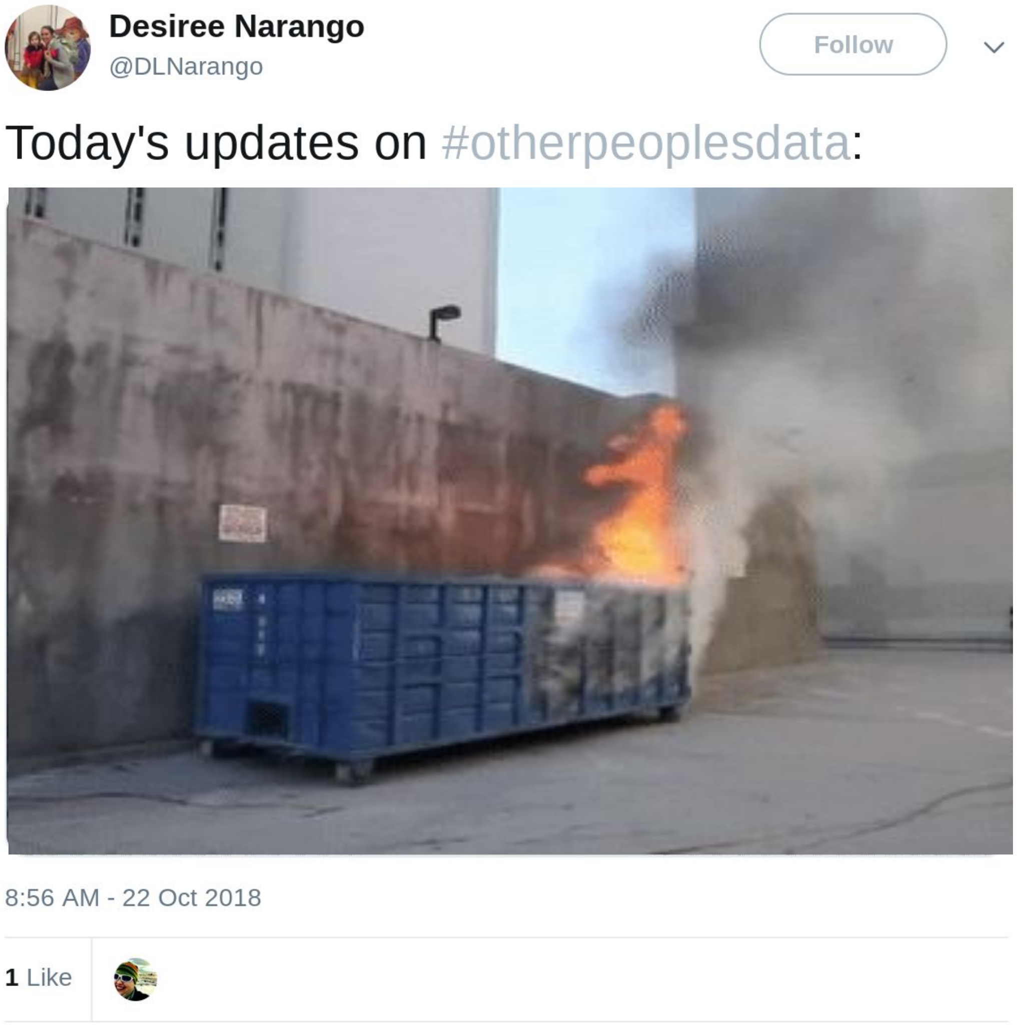
	NYMEX	
	SETTLE	Δ
Cash		
ROM		
Dec-01	2.960	0.090
Dec-01 to Mar-02	3.088	0.083
Apr-02 to Oct-02	3.166	0.084
Nov-02 to Mar-03	3.651	0.090
One Year Strip*	3.165	0.084

	IF NWPL Rocky Mountains			
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	1.890	1.910		
ROM	2.060	2.080		
Dec-01	2.395	2.415	(0.565)	(0.545)
Dec-01 to Mar-02	2.594	2.614	(0.494)	(0.474)
Apr-02 to Oct-02	2.581	2.601	(0.585)	(0.565)
Nov-02 to Mar-03	3.356	3.376	(0.295)	(0.275)
One Year Strip*	2.634	2.654	(0.530)	(0.510)

	IF CIG Rocky Mountains				IF EL Paso Permian				IF EL Paso San Juan			
	Fixed Price		Basis		Fixed Price		Basis		Fixed Price		Basis	
	BID	OFFER	BID	OFFER	BID	OFFER	BID	OFFER	BID	OFFER	BID	OFFER
Cash	1.940	1.960			2.375	2.395			2.450	2.470		
ROM	1.960	1.980			2.420	2.440			2.350	2.370		
Dec-01	2.345	2.365	(0.615)	(0.595)	2.700	2.720	(0.260)	(0.240)	2.560	2.580	(0.400)	(0.380)
Dec-01 to Mar-02	2.548	2.568	(0.540)	(0.520)	2.855	2.875	(0.233)	(0.213)	2.743	2.763	(0.345)	(0.325)
Apr-02 to Oct-02	2.471	2.491	(0.695)	(0.675)	3.009	3.029	(0.158)	(0.138)	2.801	2.821	(0.365)	(0.345)
Nov-02 to Mar-03	3.311	3.331	(0.340)	(0.320)	3.499	3.519	(0.153)	(0.133)	3.421	3.441	(0.230)	(0.210)
One Year Strip*	2.551	2.571	(0.614)	(0.594)	2.982	3.002	(0.182)	(0.162)	2.817	2.837	(0.347)	(0.327)

	AECO / NIT				IF NWPL Canadian Border (Sumas)				IF PEPL TX-OK			
	Fixed Price		Basis		Fixed Price		Basis		Fixed Price		Basis	
	BID	OFFER	BID	OFFER	BID	OFFER	BID	OFFER	BID	OFFER	BID	OFFER
Cash	2.376	2.396			2.480	2.500			2.530	2.550		
ROM	2.398	2.418			2.460	2.480			2.530	2.550		
Dec-01	2.552	2.572	(0.408)	(0.388)	2.800	2.820	(0.160)	(0.140)	2.828	2.848	(0.133)	(0.113)
Dec-01 to Mar-02	2.616	2.636	(0.472)	(0.452)	2.892	2.912	(0.196)	(0.176)	2.958	2.978	(0.130)	(0.110)
Apr-02 to Oct-02	2.661	2.681	(0.505)	(0.485)	2.796	2.816	(0.370)	(0.350)	3.046	3.066	(0.120)	(0.100)
Nov-02 to Mar-03	3.216	3.236	(0.435)	(0.415)	3.706	3.726	0.055	0.075	3.531	3.551	(0.120)	(0.100)
One Year Strip*	2.676	2.696	(0.488)	(0.468)	2.880	2.900	(0.285)	(0.265)	3.041	3.061	(0.123)	(0.103)

Slide from Jenny Bryan
(https://github.com/jennybc/2016-06_spreadsheets/blob/master/2016-06_useR-stanford.pdf)



Desiree Narango
@DLNarango

Today's updates on #otherpeoplesdata:

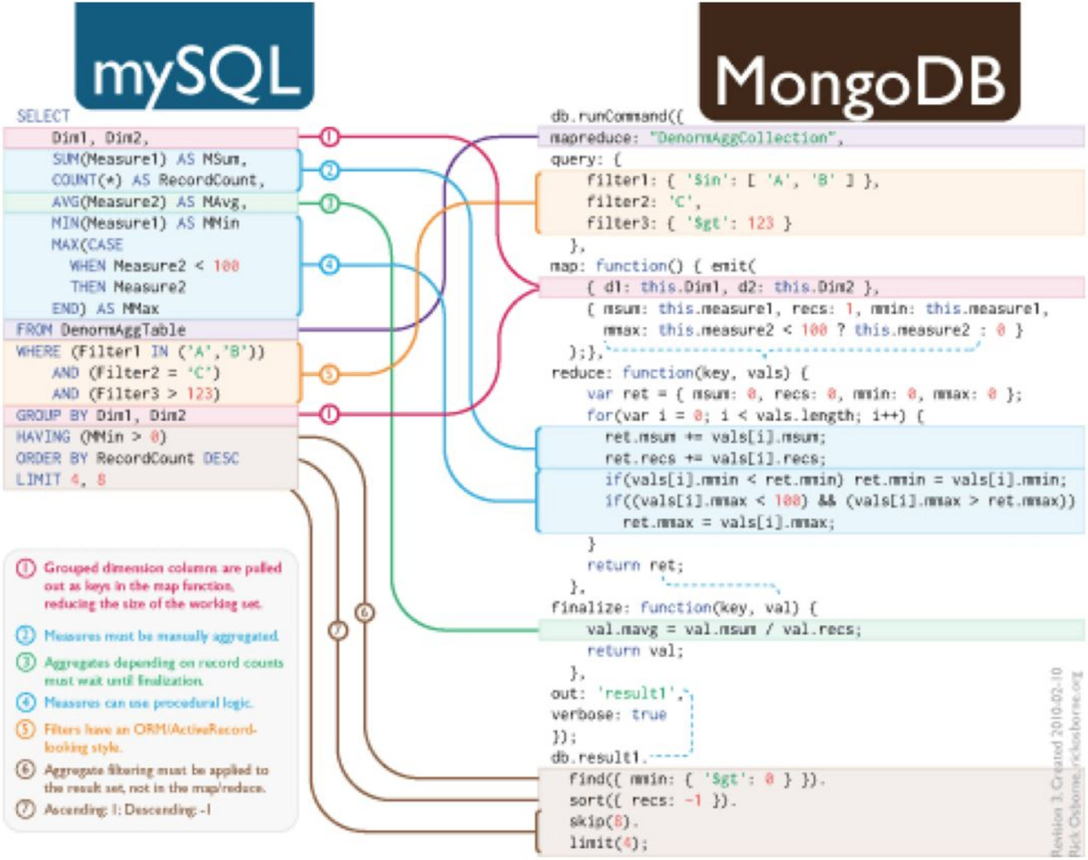
8:56 AM - 22 Oct 2018

1 Like

Where you wish data was

Where data actually is

https://rickosborne.org/blog/2010/02/infographic-migrating-from-sql-to-mapreduce-with-mongodb/



```
SELECT
    Dim1, Dim2,
    SUM(Measure1) AS MSum,
    COUNT(*) AS RecordCount,
    AVG(Measure2) AS MAvg,
    MIN(Measure1) AS MMin
    MAX(CASE
      WHEN Measure2 < 100
      THEN Measure2
    END) AS MMax
FROM DenormAggTable
WHERE (Filter1 IN ('A','B'))
    AND (Filter2 = 'C')
    AND (Filter3 > 123)
GROUP BY Dim1, Dim2
HAVING (MMin > 0)
ORDER BY RecordCount DESC
LIMIT 4, 8
```

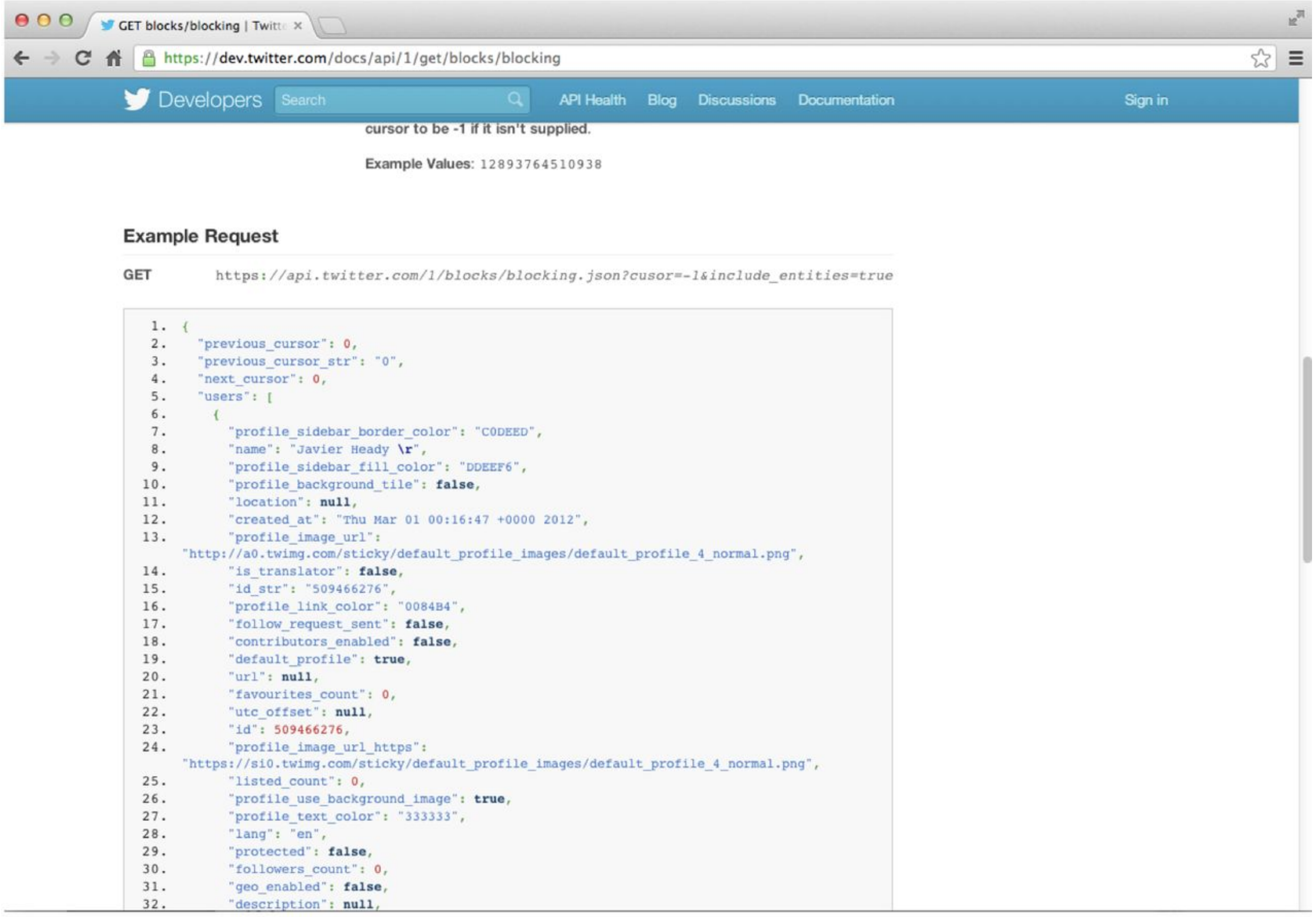
```
db.runCommand({
mapreduce: "DenormAggCollection",
query: {
    filter1: { '$in': [ 'A', 'B' ] },
    filter2: 'C',
    filter3: { '$gt': 123 }
  },
map: function() { emit(
    { d1: this.Dim1, d2: this.Dim2 },
    { msum: this.measure1, recs: 1, mmin: this.measure1,
      mmax: this.measure2 < 100 ? this.measure2 : 0 }
  );},
reduce: function(key, vals) {
    var ret = { msum: 0, recs: 0, mmin: 0, mmax: 0 };
    for(var i = 0; i < vals.length; i++) {
      ret.msum += vals[i].msum;
      ret.recs += vals[i].recs;
      if(vals[i].mmin < ret.mmin) ret.mmin = vals[i].mmin;
      if((vals[i].mmax < 100) && (vals[i].mmax > ret.mmax))
        ret.mmax = vals[i].mmax;
    }
    return ret;
  },
finalize: function(key, val) {
    val.mavg = val.msum / val.recs;
    return val;
  },
out: 'result1',
verbose: true
});
db.result1.
  find({ mmin: { '$gt': 0 } }).
  sort({ recs: -1 }).
  skip(8).
  limit(4);
```

1. Grouped dimension columns are pulled out as keys in the map function, reducing the size of the working set.
2. Measures must be manually aggregated.
3. Aggregates depending on record counts must wait until finalization.
4. Measures can use procedural logic.
5. Filters have an ORM/ActiveRecord-looking style.
6. Aggregate filtering must be applied to the result set, not in the map/reduce.
7. Ascending: 1; Descending: -1

Revision 3, Created 2010-02-10
Rick Osborne, rickosborne.org

Where data actually is

https://dev.twitter.com/docs/api/1/get/blocks/blocking


GET blocks/blocking | Twitt
https://dev.twitter.com/docs/api/1/get/blocks/blocking
Developers
Search
API Health
Blog
Discussions
Documentation
Sign in

cursor to be -1 if it isn't supplied.

Example Values: 12893764510938

Example Request

GET https://api.twitter.com/1/blocks/blocking.json?cusor=-1&include_entities=true

```
{
  "previous_cursor": 0,
  "previous_cursor_str": "0",
  "next_cursor": 0,
  "users": [
    {
      "profile_sidebar_border_color": "C0DEED",
      "name": "Javier Heady \r",
      "profile_sidebar_fill_color": "DDEEF6",
      "profile_background_tile": false,
      "location": null,
      "created_at": "Thu Mar 01 00:16:47 +0000 2012",
      "profile_image_url":
"http://a0.twimg.com/sticky/default_profile_images/default_profile_4_normal.png",
      "is_translator": false,
      "id_str": "509466276",
      "profile_link_color": "0084B4",
      "follow_request_sent": false,
      "contributors_enabled": false,
      "default_profile": true,
      "url": null,
      "favourites_count": 0,
      "utc_offset": null,
      "id": 509466276,
      "profile_image_url_https":
"https://si0.twimg.com/sticky/default_profile_images/default_profile_4_normal.png",
      "listed_count": 0,
      "profile_use_background_image": true,
      "profile_text_color": "333333",
      "lang": "en",
      "protected": false,
      "followers_count": 0,
      "geo_enabled": false,
      "description": null,
```

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

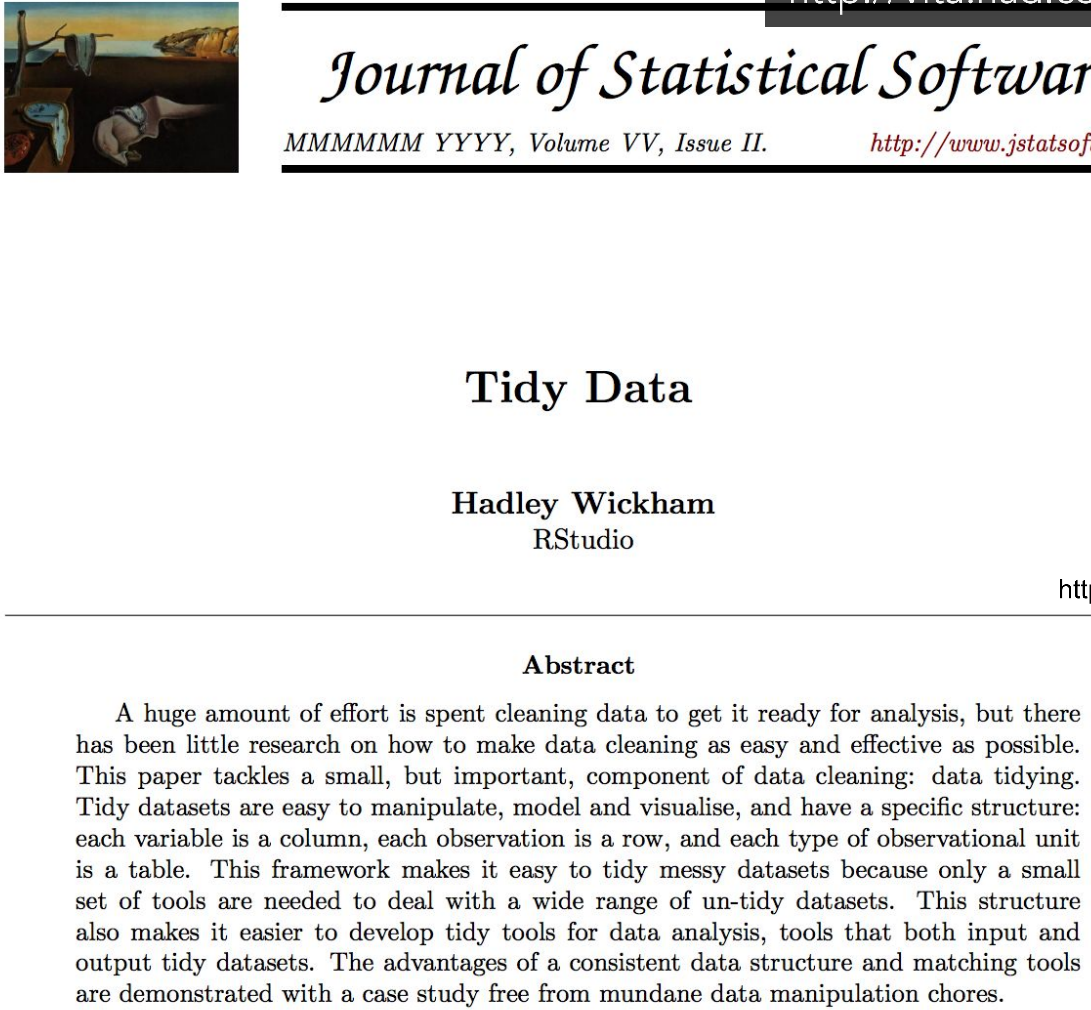
> “Data are values of qualitative or quantitative variables, belonging to a set of items.”

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

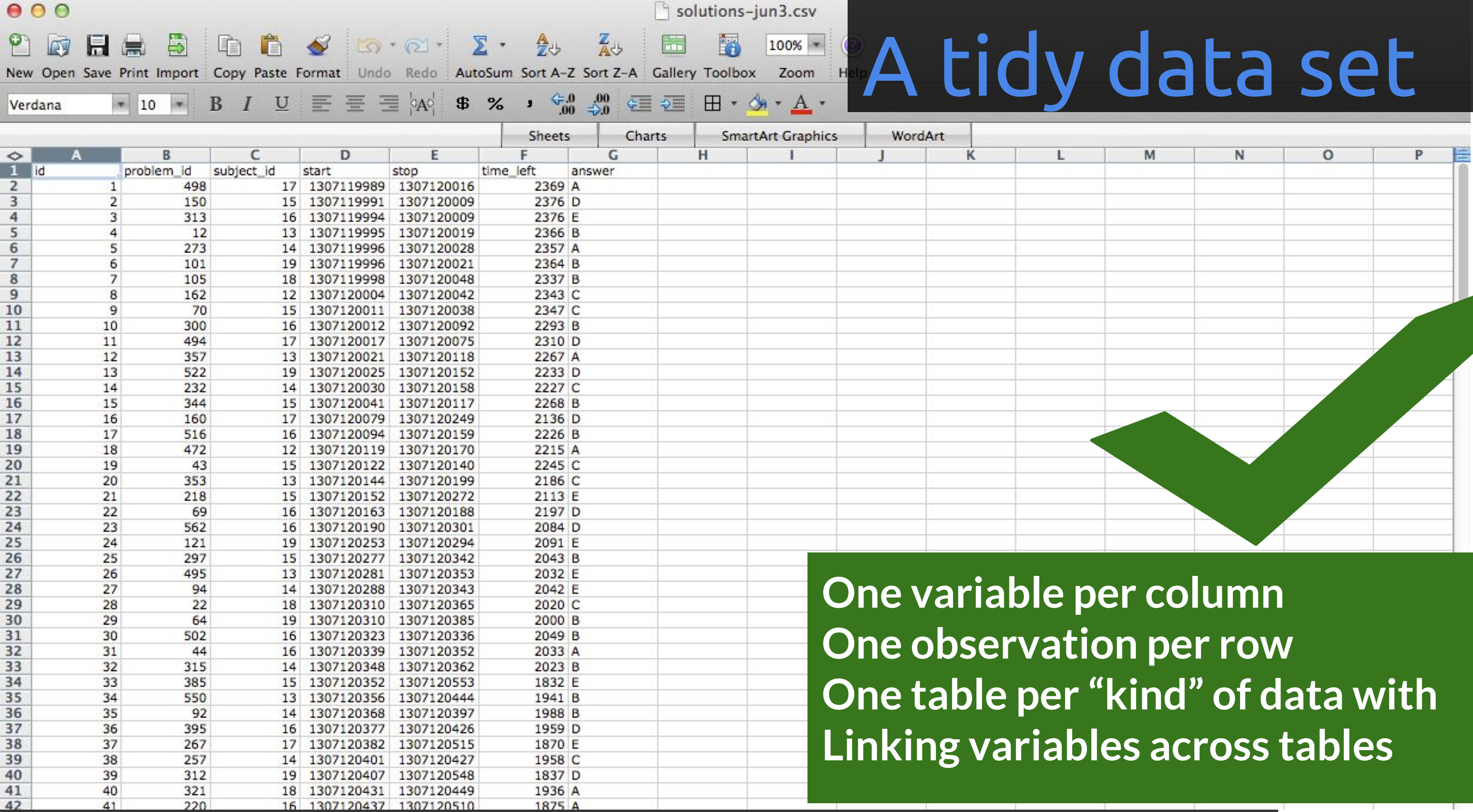
Hadley Wickham
RStudio

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.

https://twitter.com/hadleywickham

A tidy data set

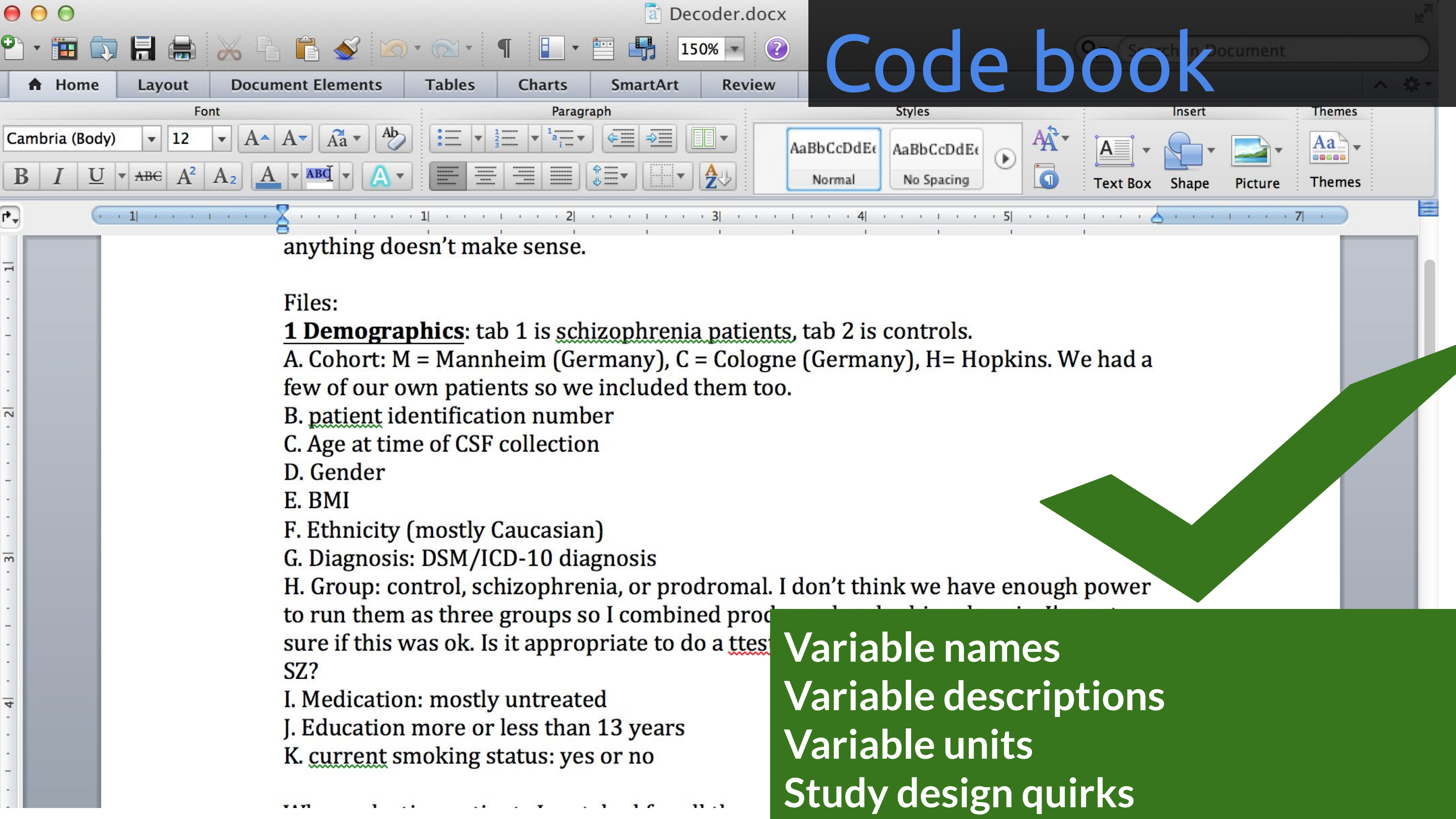


	A	B	C	D	E	F	G
1	id	problem_id	subject_id	start	stop	time_left	answer
2	1	498	17	1307119989	1307120016	2369	A
3	2	150	15	1307119991	1307120009	2376	D
4	3	313	16	1307119994	1307120009	2376	E
5	4	12	13	1307119995	1307120019	2366	B
6	5	273	14	1307119996	1307120028	2357	A
7	6	101	19	1307119996	1307120021	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

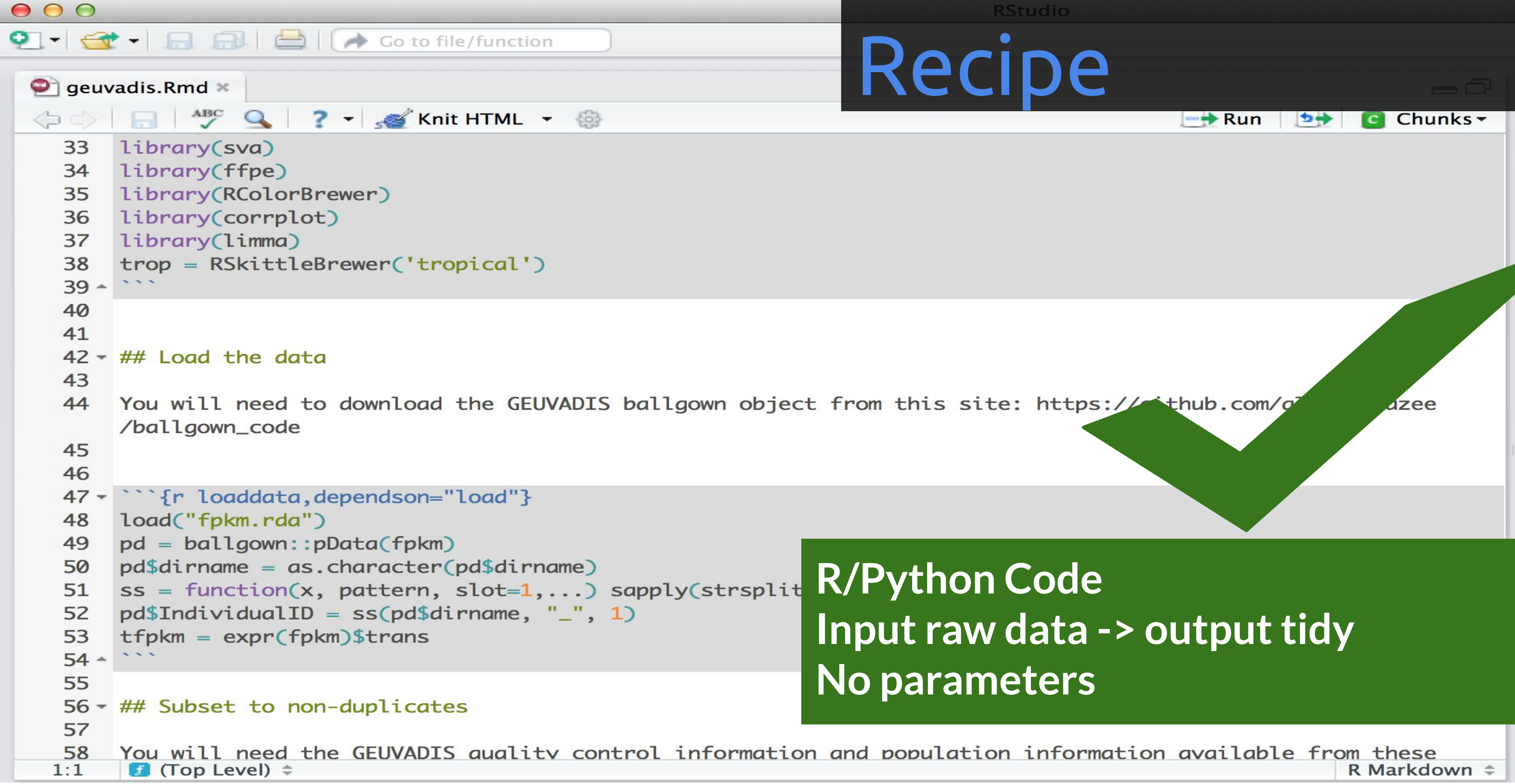
Variable names

Variable descriptions

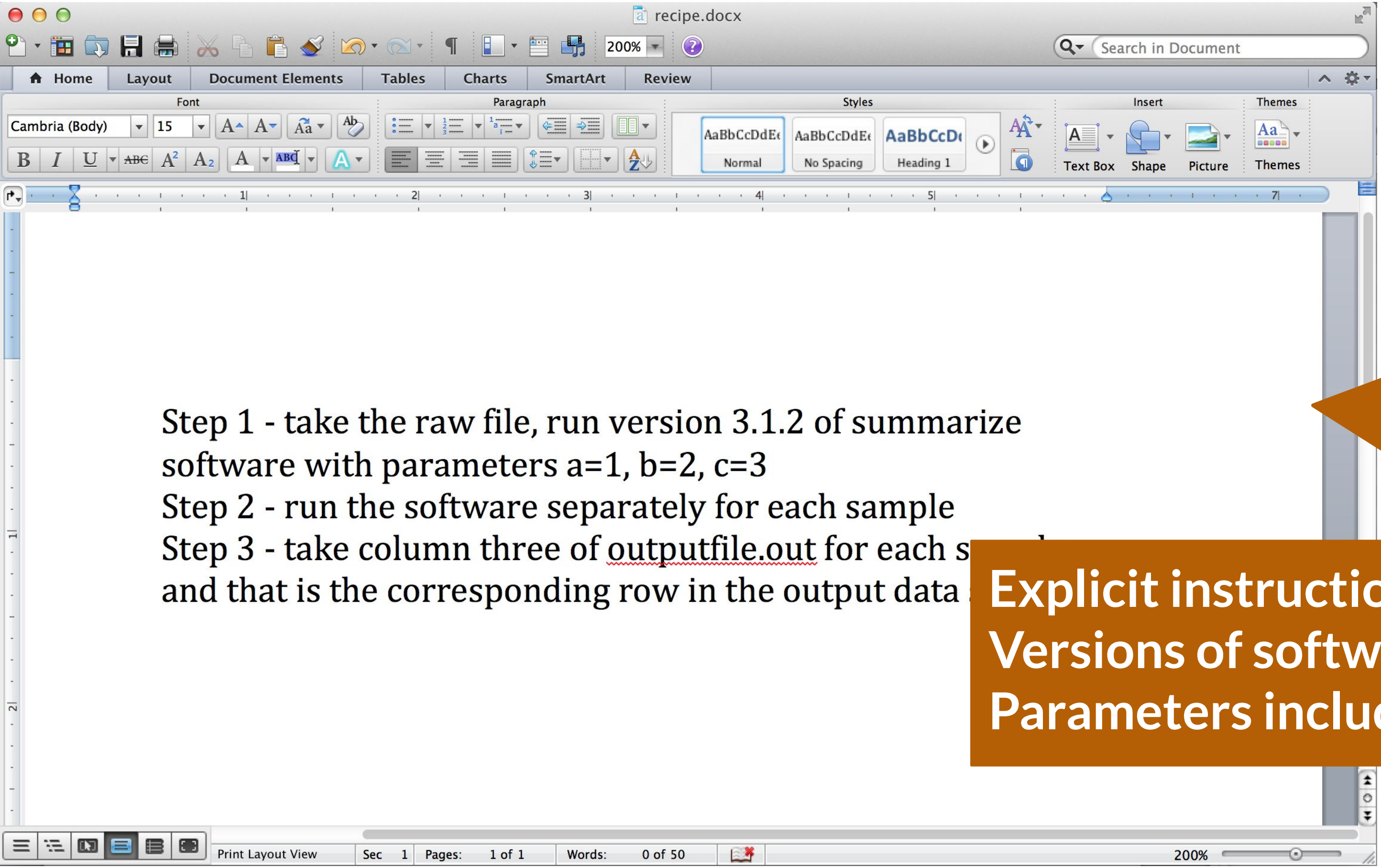
Variable units

Study design quirks

Recipe

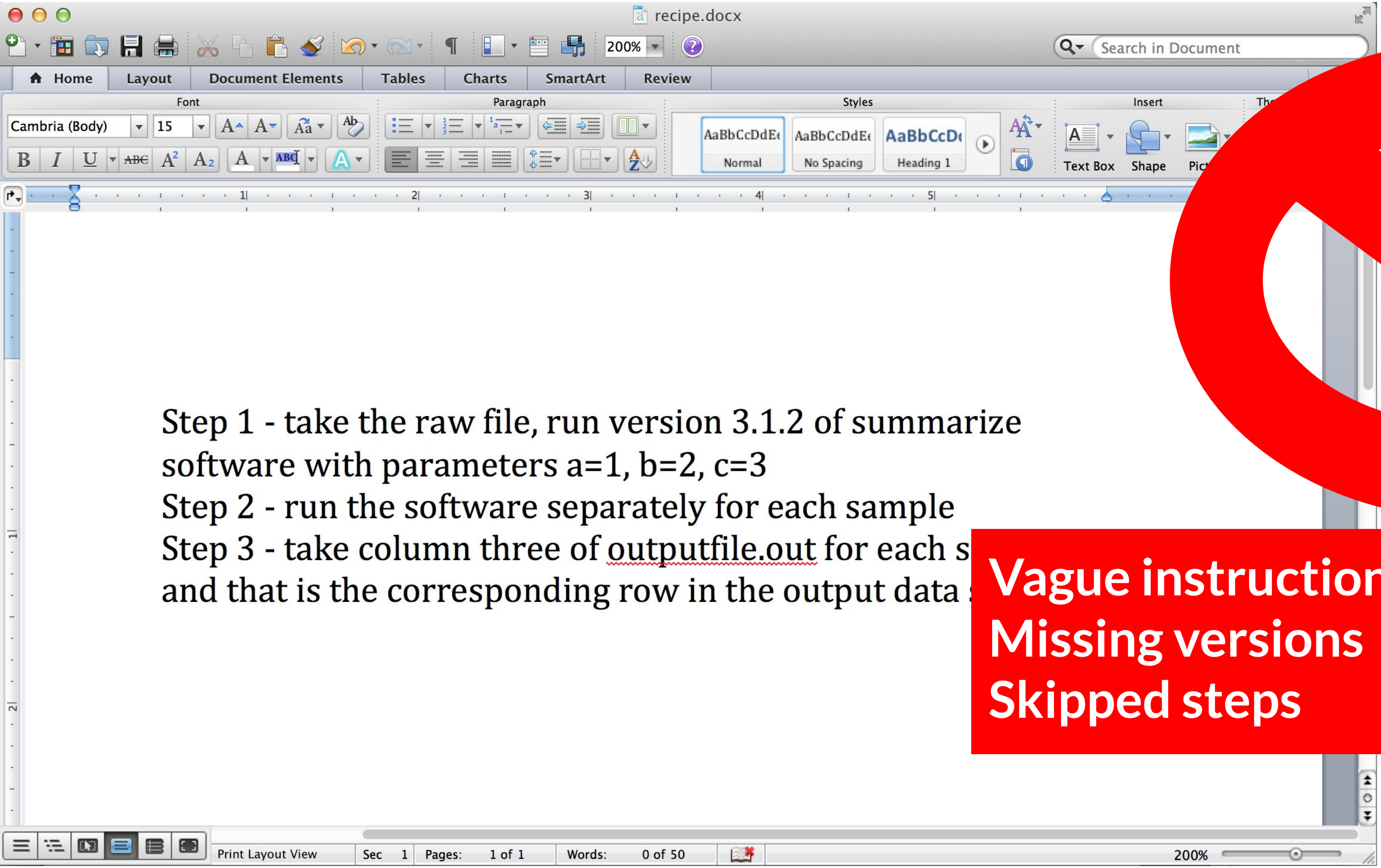


```
library(sva)
library(ffpe)
library(RColorBrewer)
library(corrplot)
library(limma)
trop = RSkittleBrewer('tropical')
```
```

## Load the data

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



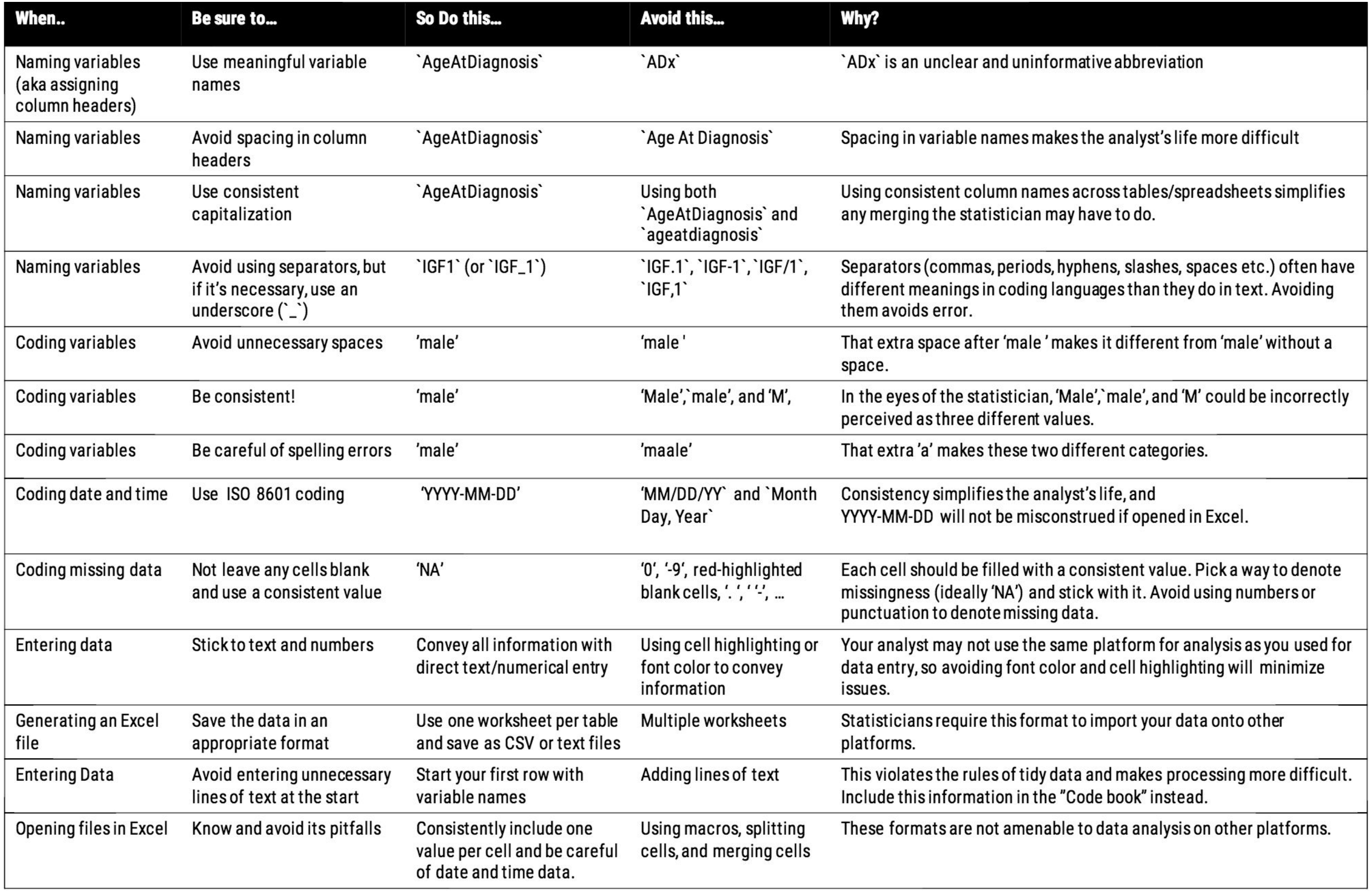
Step 1 - take the raw file, run version 3.1.2 of summarize software with parameters a=1, b=2, c=3
Step 2 - run the software separately for each sample
Step 3 - take column three of outputfile.out for each s
and that is the corresponding row in the output data

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



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

Rules for Tidy Spreadsheets

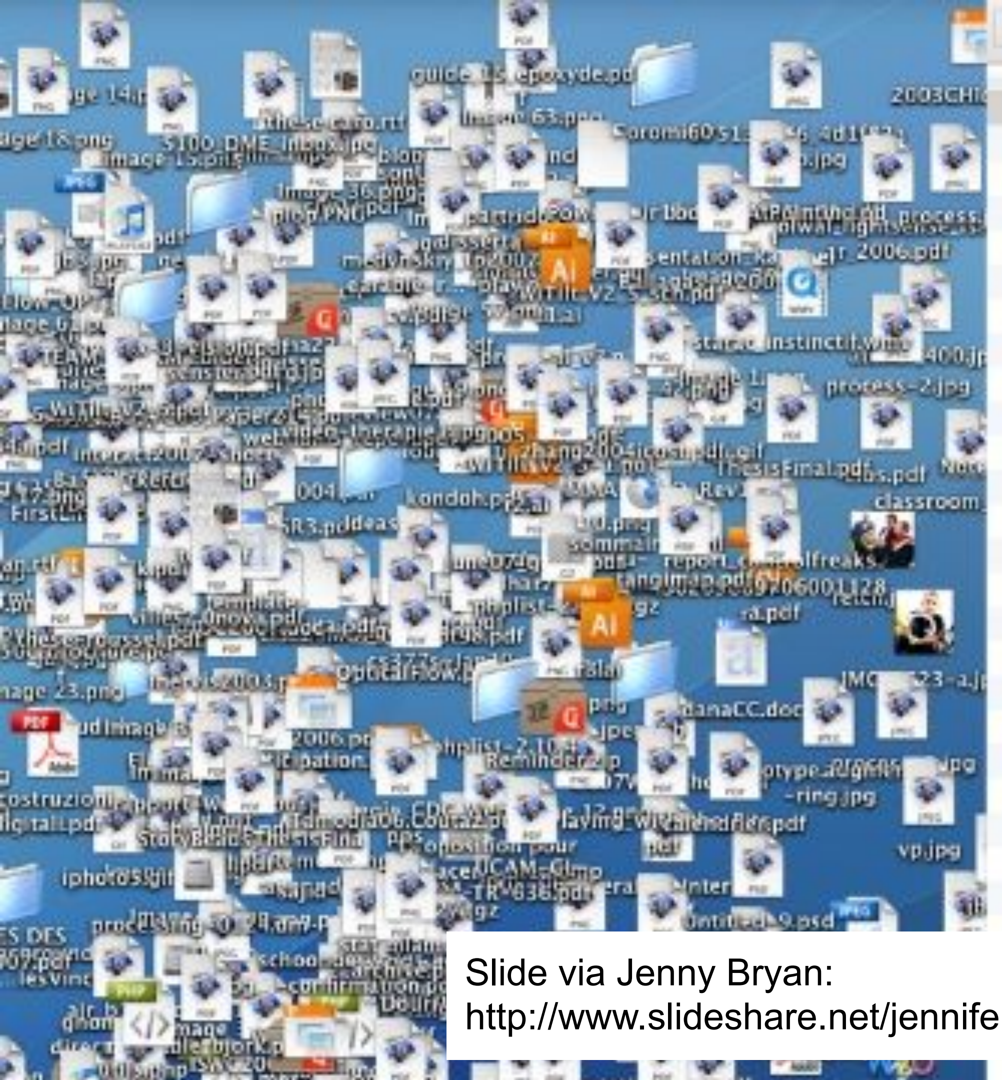
1. Be consistent
2. Choose good names for things
3. Write dates as YYYY-MM-DD
4. No empty cells
5. Put just one thing in a cell
6. Don't use font color or highlighting as data
7. Save the data as plain text files

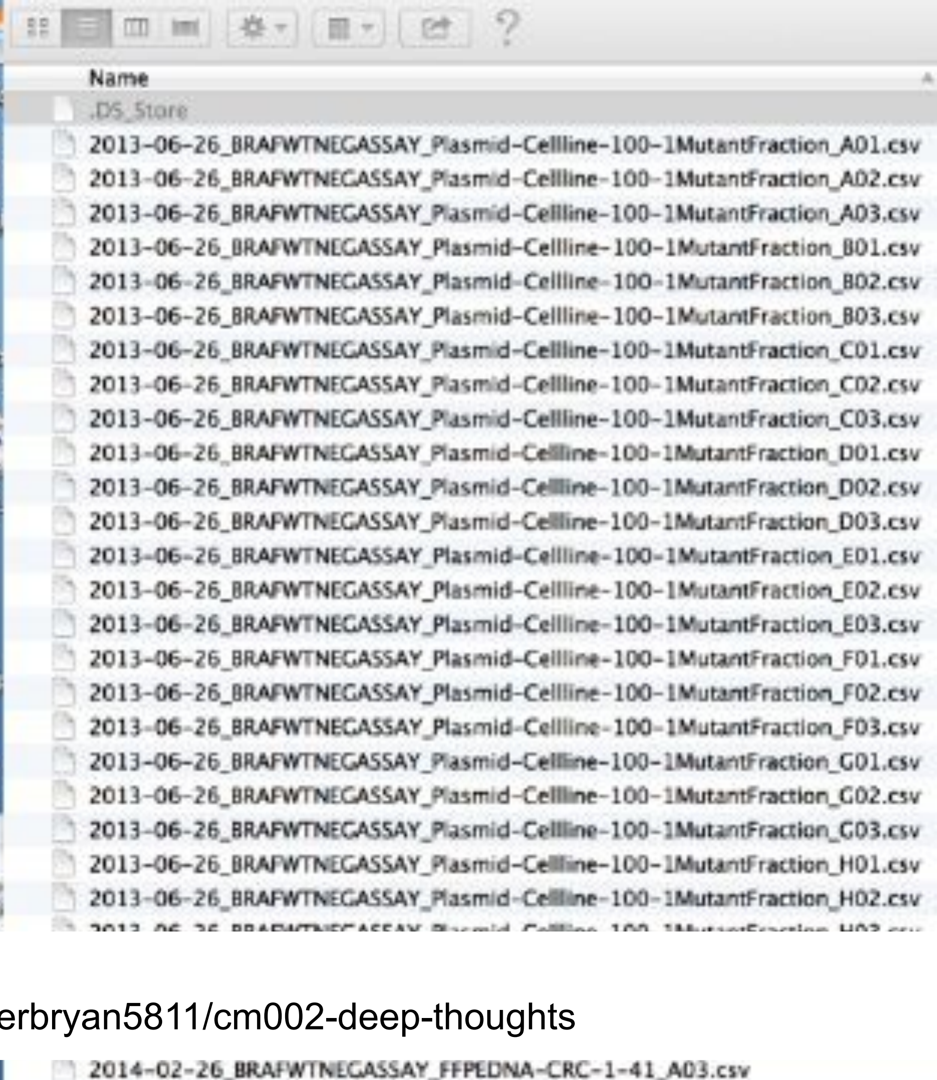
https://doi.org/10.1080/00031305.2017.1375989

Organize thyself

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

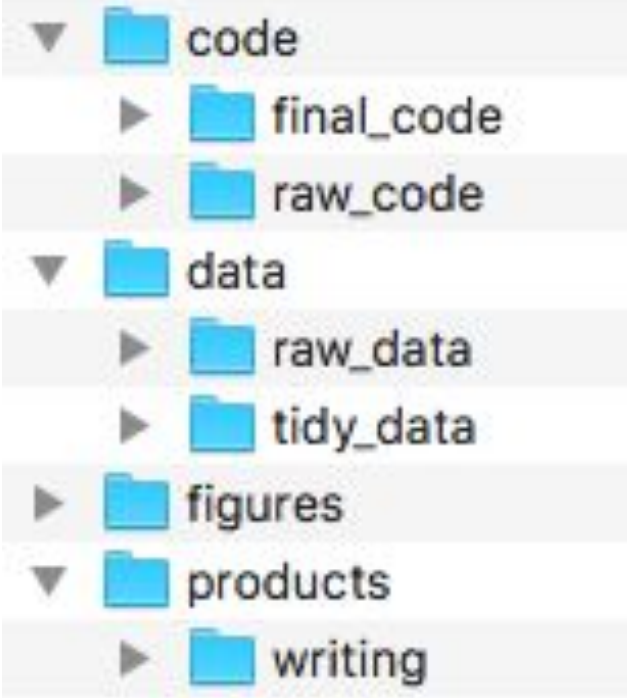
- Jenny Bryan





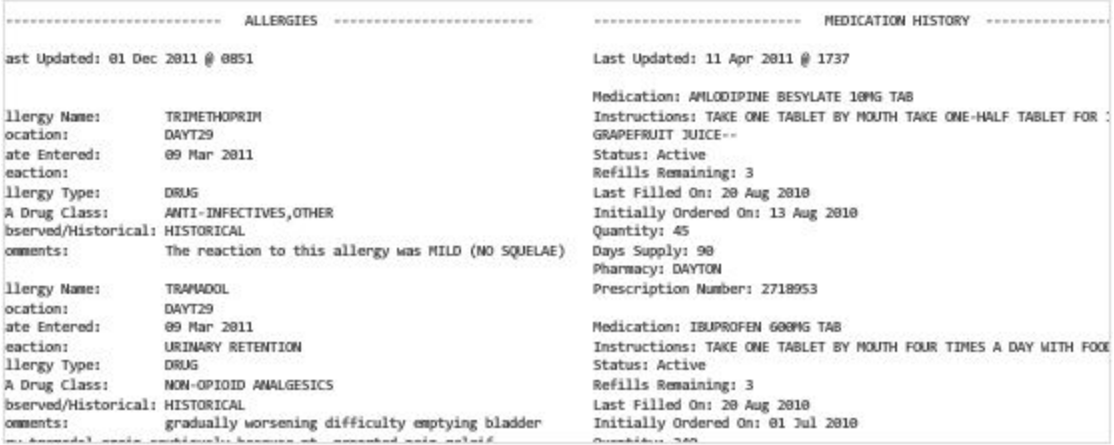
Name
.DS_Store
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_A01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_A02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_A03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_B01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_B02.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
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H03.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
```

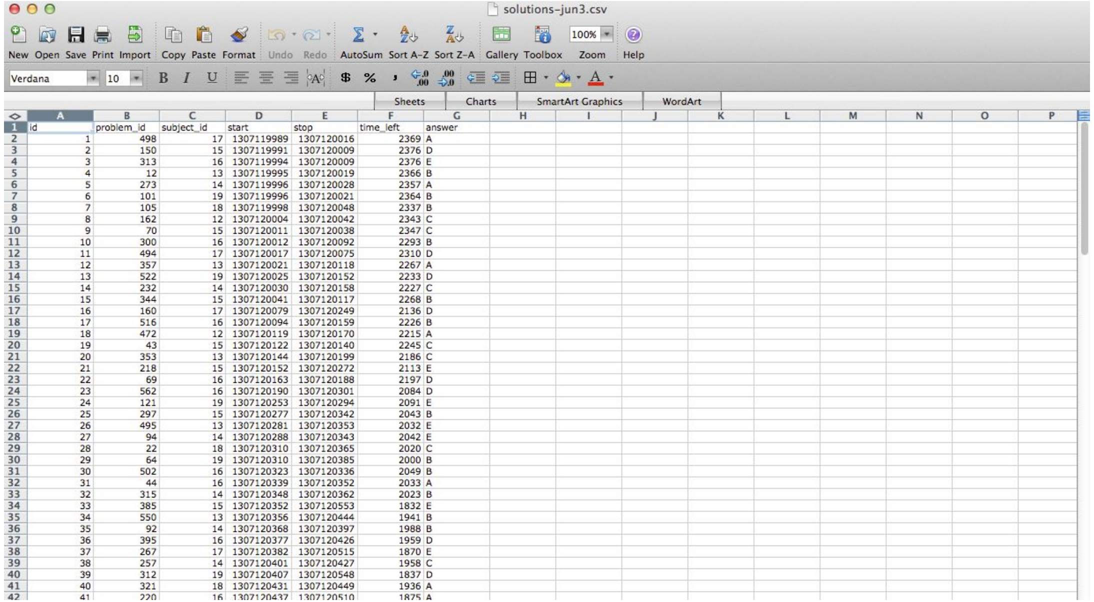
Raw data



```
---------------------------  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 FOOD
Status: Active
Refills Remaining: 3
Last Filled On: 20 Aug 2010
Initially Ordered On: 01 Jul 2010
```

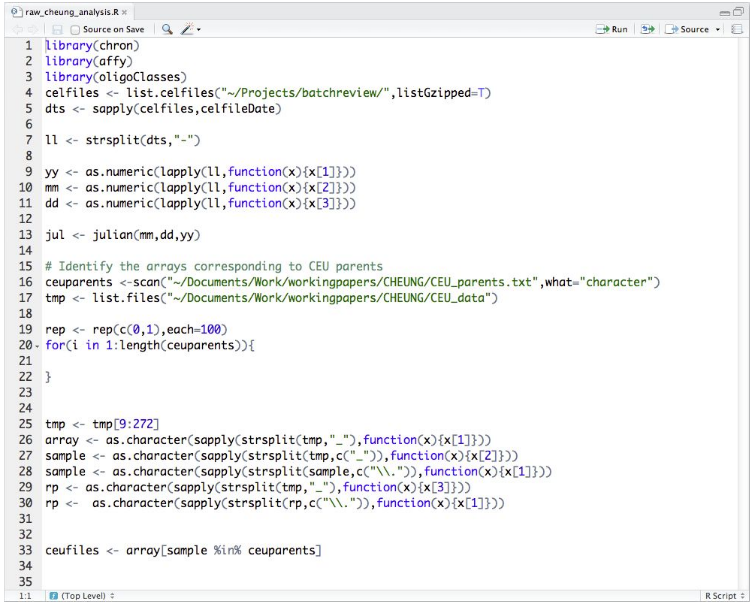
Processed data



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

- 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

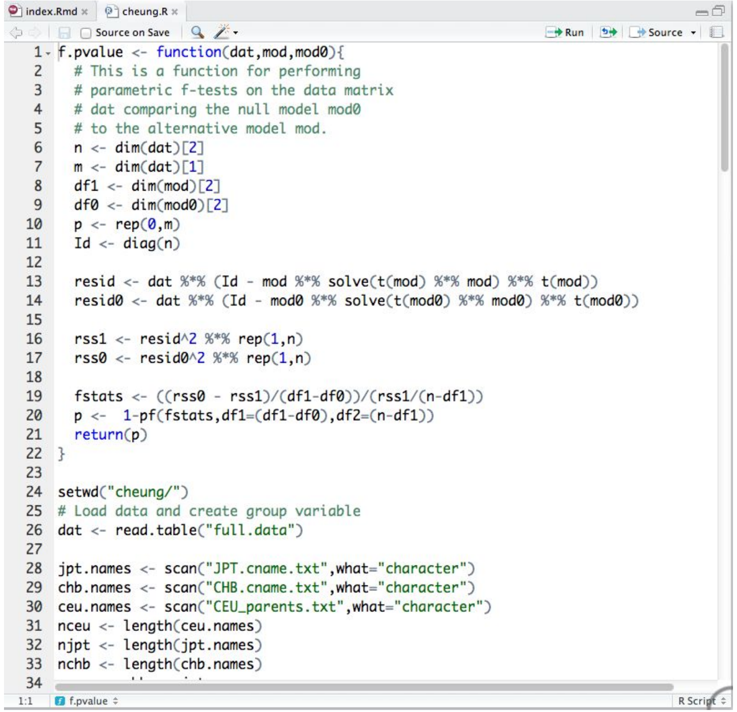
Raw scripts



```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]
```

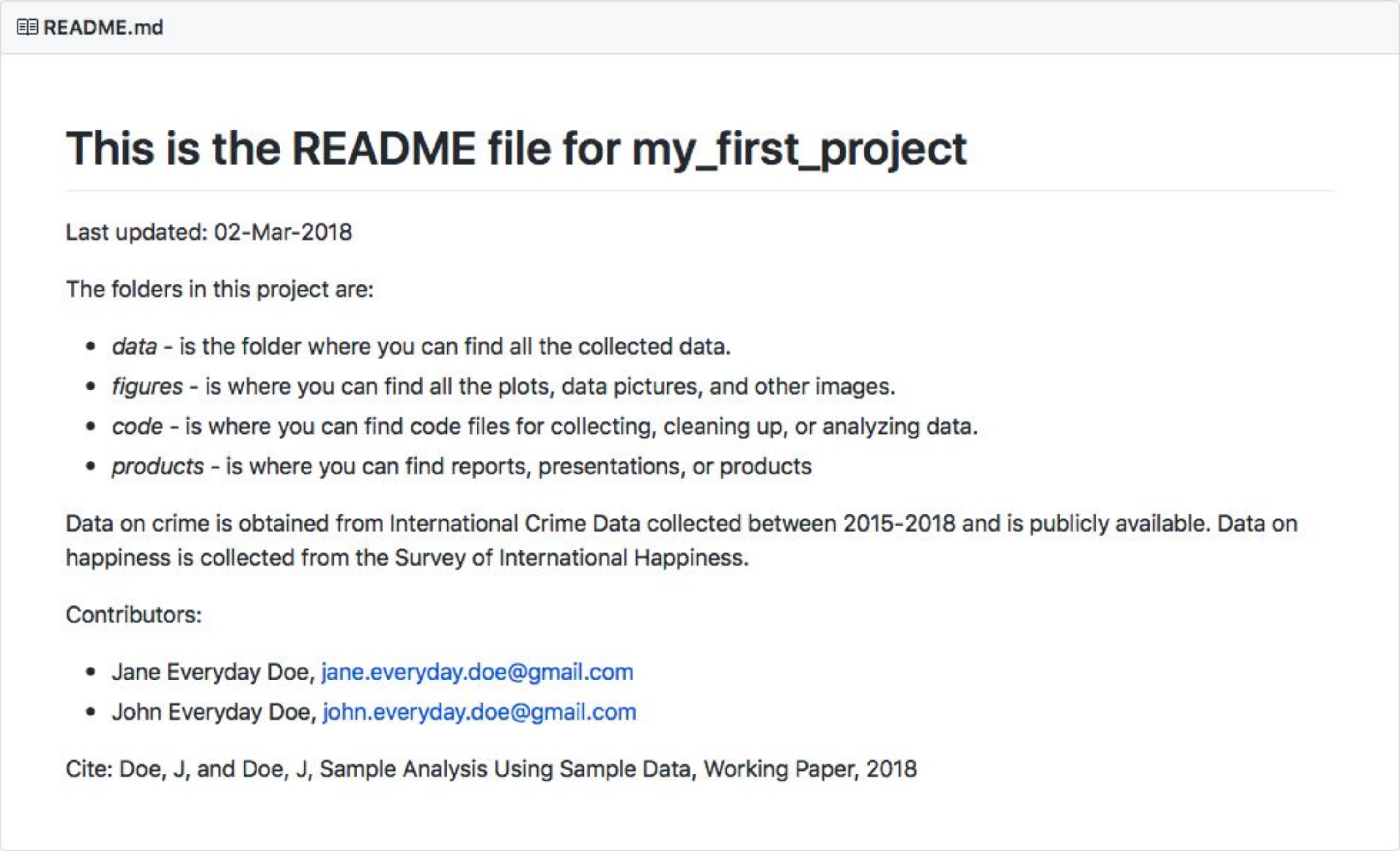
- May be less commented (but comments help you!)
- May be multiple versions
- May include analyses that are later discarded

Final scripts



```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)
```

- 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



README.md

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

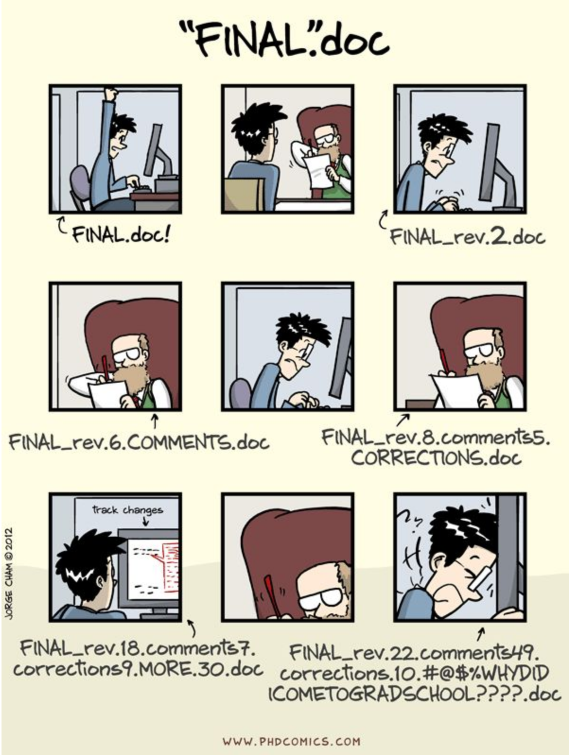
Data on crime is obtained from International Crime Data collected between 2015-2018 and is publicly available. Data on happiness is collected from the Survey of International Happiness.

Contributors:

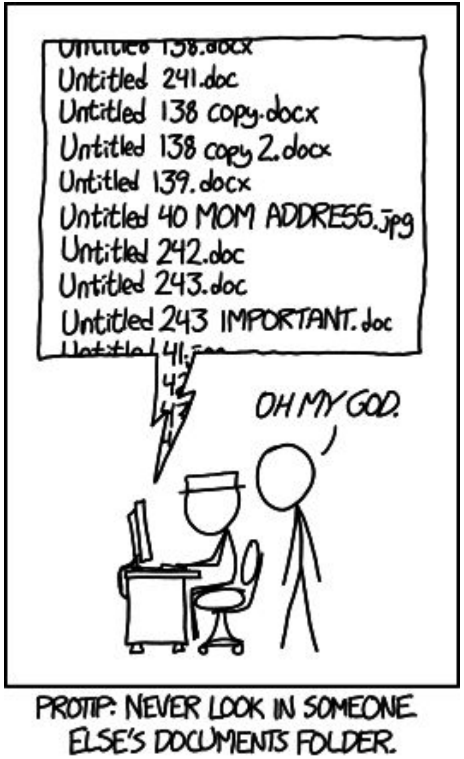
- Jane Everyday Doe, jane.everyday.doe@gmail.com
- John Everyday Doe, john.everyday.doe@gmail.com

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

Just no



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



https://xkcd.com/1459/

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

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

- Karl Broman

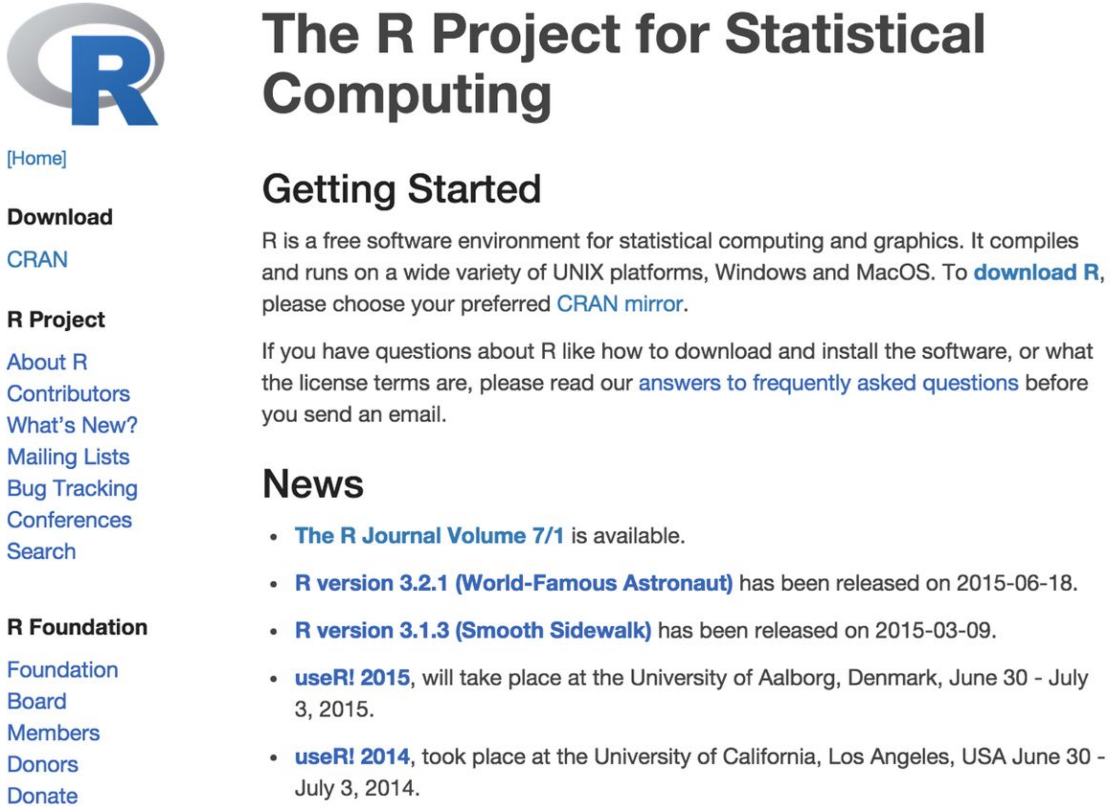
http://kbroman.org/Tools4RR/assets/lectures/06_org_eda.pdf

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



The R Project for Statistical Computing

[Home]

Download

CRAN

R Project

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

R Foundation

Foundation
Board
Members
Donors
Donate

Getting Started

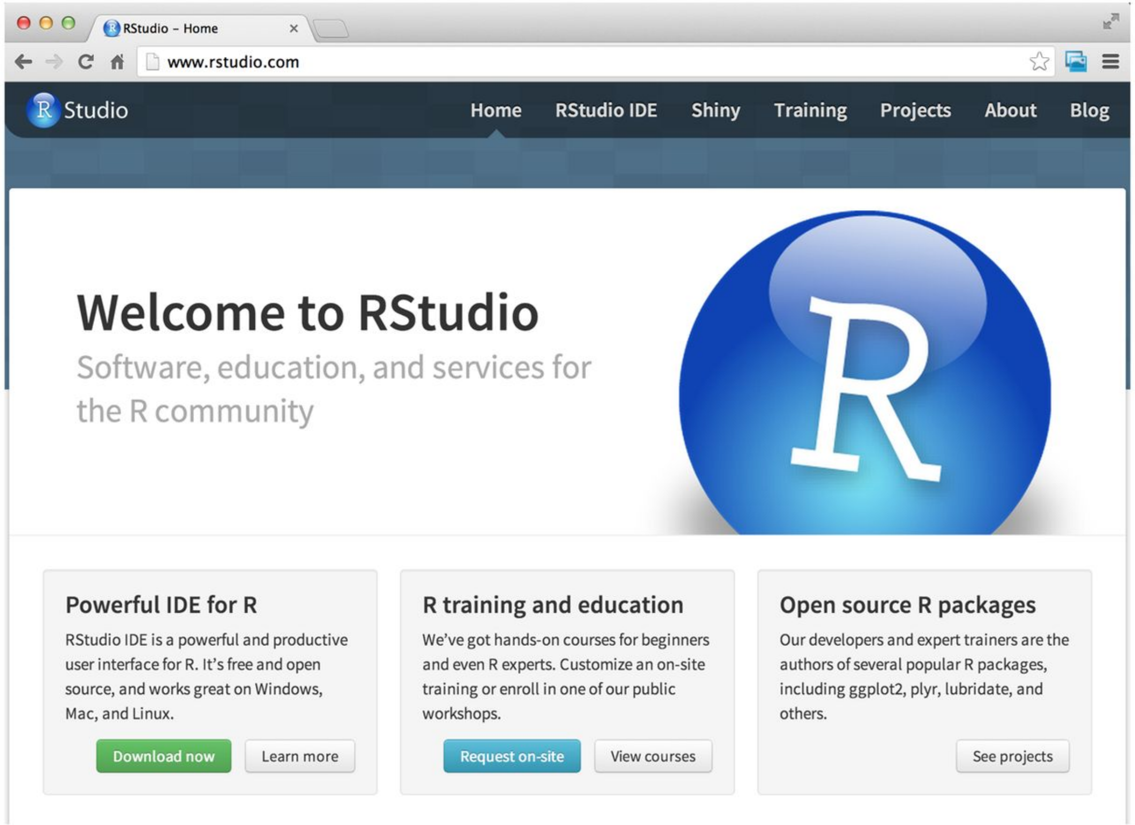
R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

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



Welcome to RStudio

Software, education, and services for the R community

Powerful IDE for R

RStudio IDE is a powerful and productive user interface for R. It's free and open source, and works great on Windows, Mac, and Linux.

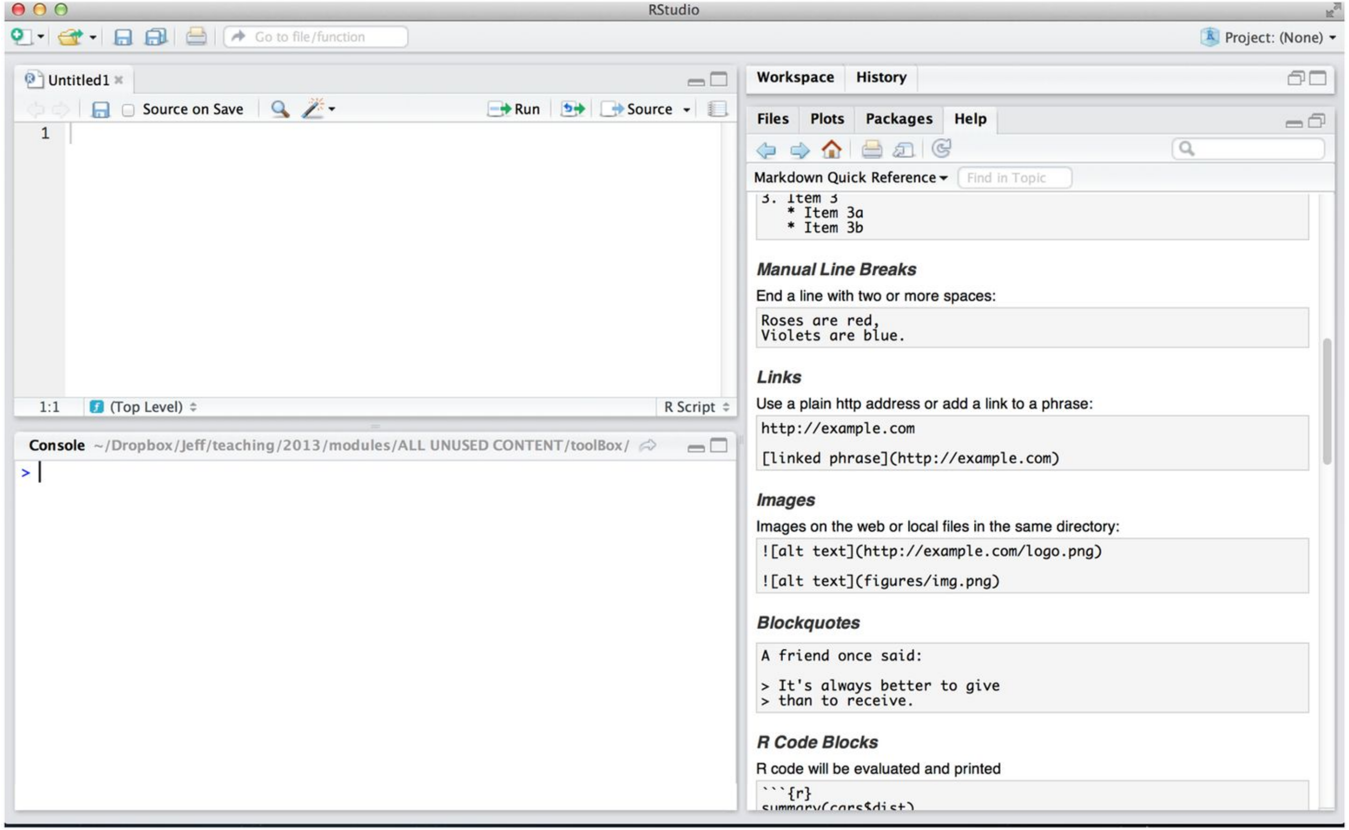
R training and education

We've got hands-on courses for beginners and even R experts. Customize an on-site training or enroll in one of our public workshops.

Open source R packages

Our developers and expert trainers are the authors of several popular R packages, including ggplot2, plyr, lubridate, and others.

https://www.rstudio.com/



```
3. Item 3
   * Item 3a
   * Item 3b
```

Manual Line Breaks

End a line with two or more spaces:

```
Roses are red,
Violets are blue.
```

Links

Use a plain http address or add a link to a phrase:

```
http://example.com

[linked phrase](http://example.com)
```

Images

Images on the web or local files in the same directory:

```
![alt text](http://example.com/logo.png)

![alt text](figures/img.png)
```

Blockquotes

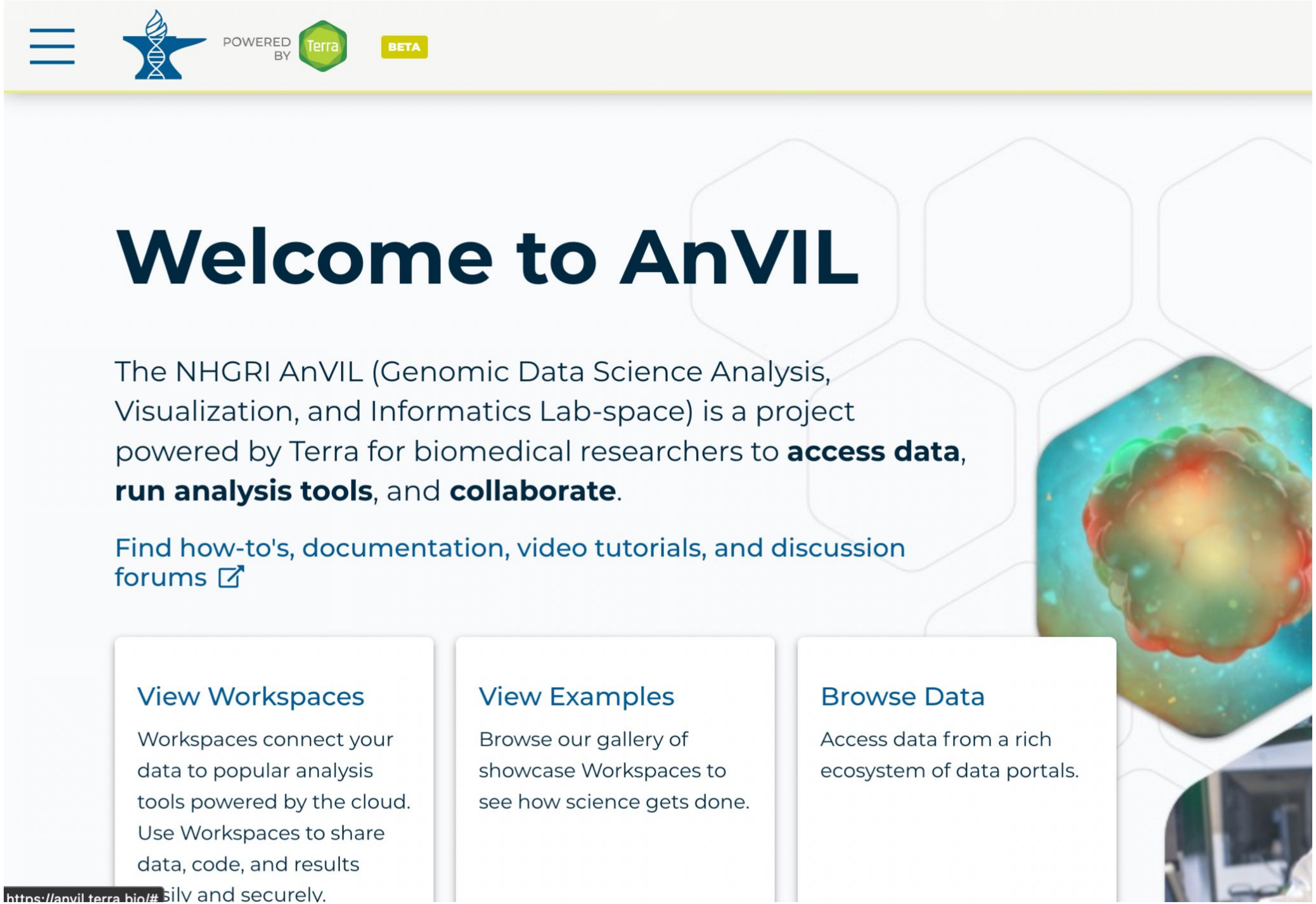
```
A friend once said:

> It's always better to give
> than to receive.
```

R Code Blocks

R code will be evaluated and printed

~~~
```{r}
summary(cars$dist)
~~~

https://www.rstudio.com/



Welcome to AnVIL

The NHGRI AnVIL (Genomic Data Science Analysis, Visualization, and Informatics Lab-space) is a project powered by Terra for biomedical researchers to **access data**, **run analysis tools**, and **collaborate**.

Find how-to's, documentation, video tutorials, and discussion forums

View Workspaces

Workspaces connect your data to popular analysis tools powered by the cloud. Use Workspaces to share data, code, and results https://anvil.terra.bio/# sily and securely.

View Examples

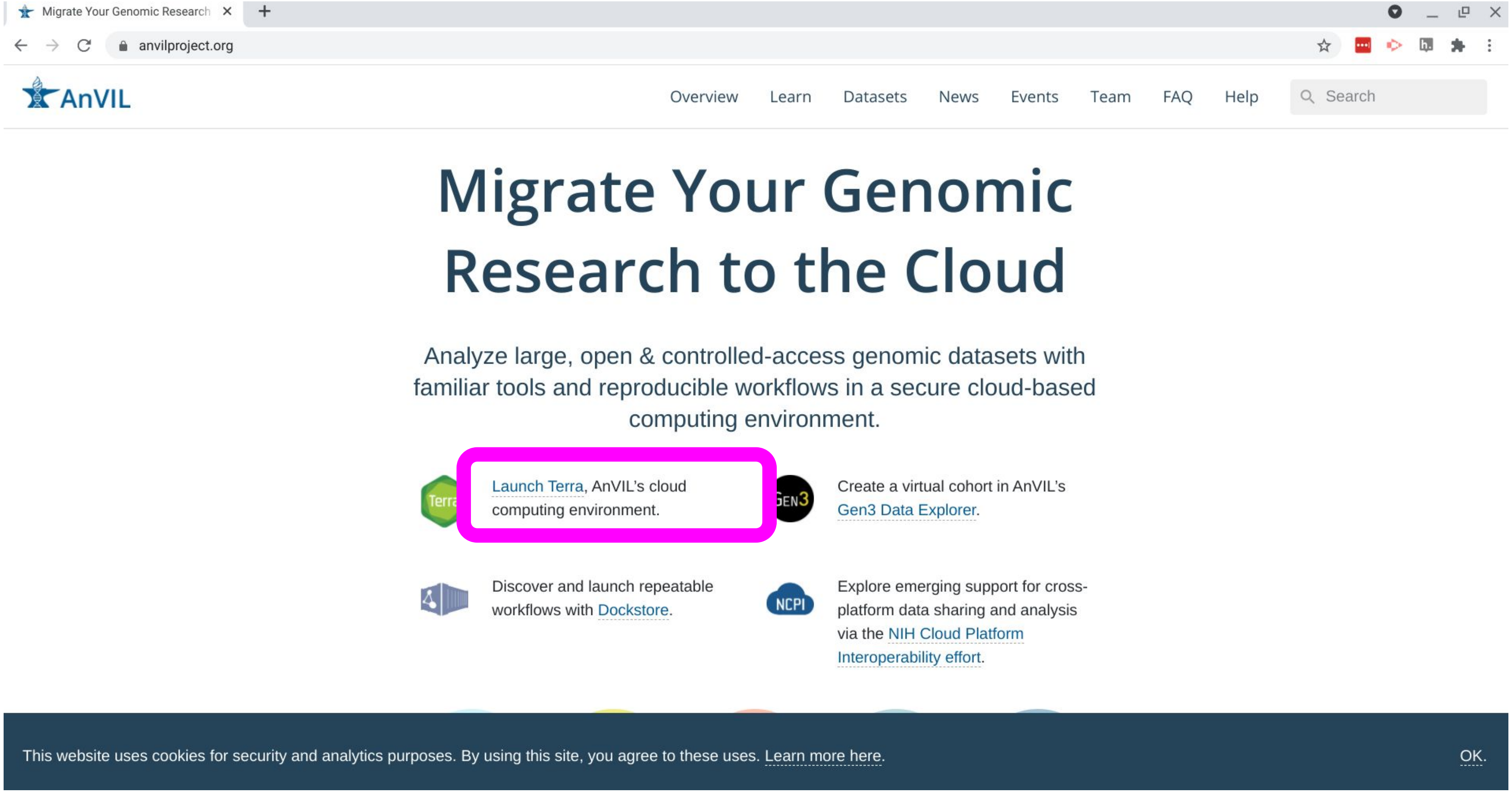
Browse our gallery of showcase Workspaces to see how science gets done.

Browse Data

Access data from a rich ecosystem of data portals.

https://anvil.terra.bio

What is AnVIL???



Migrate Your Genomic Research to the Cloud

Analyze large, open & controlled-access genomic datasets with familiar tools and reproducible workflows in a secure cloud-based computing environment.

Launch Terra, AnVIL's cloud computing environment.

Create a virtual cohort in AnVIL's Gen3 Data Explorer.

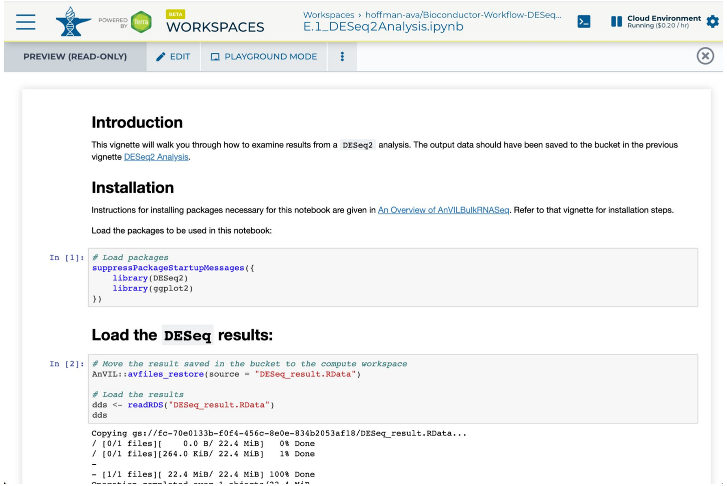
Discover and launch repeatable workflows with Dockstore.

Explore emerging support for cross-platform data sharing and analysis via the NIH Cloud Platform Interoperability effort.

This website uses cookies for security and analytics purposes. By using this site, you agree to these uses. Learn more here. OK.

https://anvilproject.org/

AnVIL is “renting computers”



Introduction

This vignette will walk you through how to examine results from a `DESeq2` analysis. The output data should have been saved to the bucket in the previous vignette DESeq2 Analysis.

Installation

Instructions for installing packages necessary for this notebook are given in An Overview of AnVILBulkRNASeq. Refer to that vignette for installation steps.

Load the packages to be used in this notebook:

```
In [1]: # Load packages
        suppressPackageStartupMessages({
            library(DESeq2)
            library(ggplot2)
        })
```

Load the `DESeq` results:

```
In [2]: # Move the result saved in the bucket to the compute workspace
        AnVIL::avfiles_restore(source = "DESeq_result.RData")

        # Load the results
        dds <- readRDS("DESeq_result.RData")
        dds
```

```
Copying gs://fc-70e0133b-f0f4-456c-8e0e-834b2053af18/DESeq_result.RData...
/ [0/1 files][    0.0 B/ 22.4 MiB]   0% Done
/ [0/1 files][264.0 KiB/ 22.4 MiB]   1% Done
-
- [1/1 files][ 22.4 MiB/ 22.4 MiB] 100% Done
Operation completed over 1 objects/22.4 MiB
```

<u>Standard Computing</u>

- You buy a laptop one time
- You get that one laptop
- You pay little per use

<u>Cloud computing</u>

- You use any web browser
- You rent the computers
- You pay per hour/gigabyte/etc.

It can feel a little weird

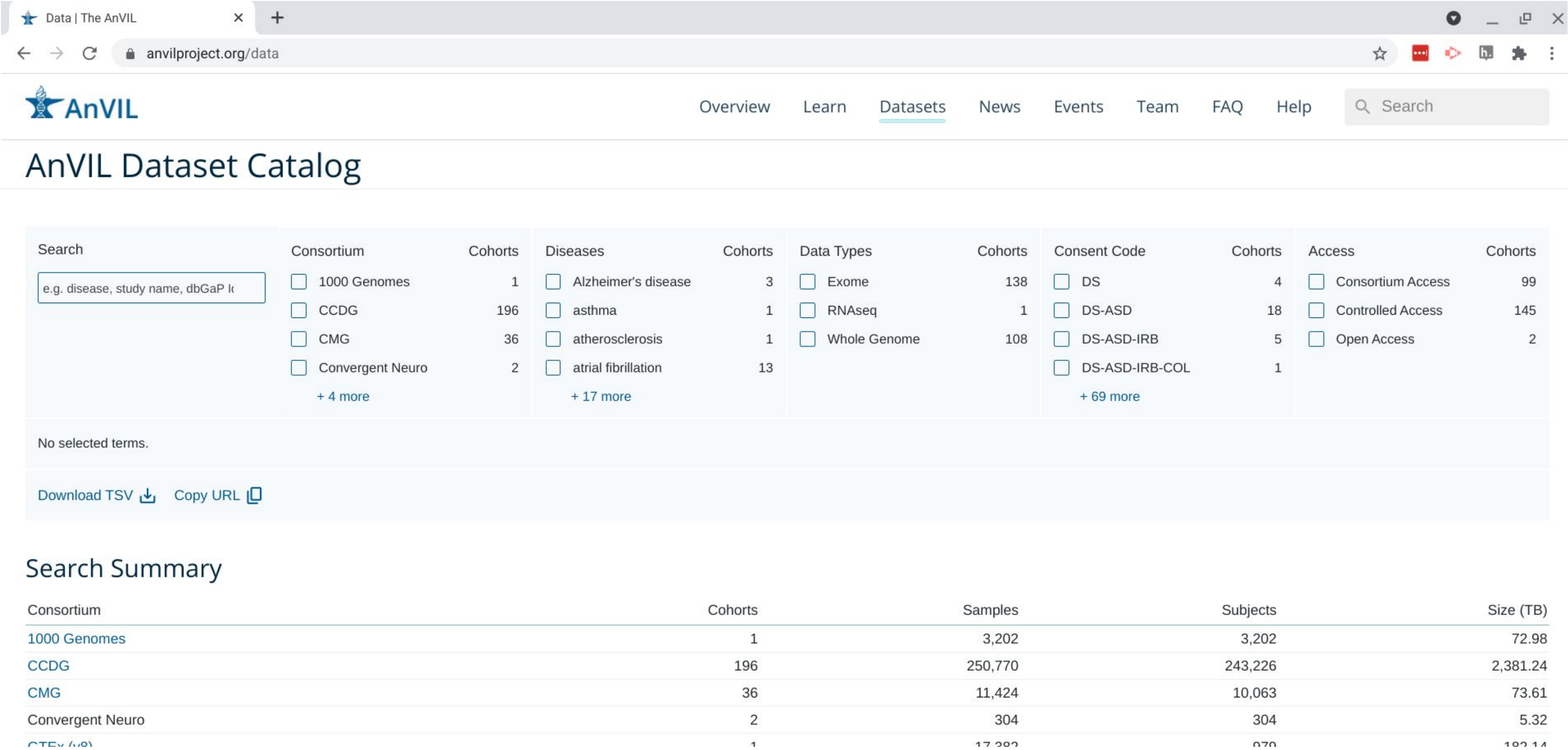
Purchased car

- You buy the car
- You fill up at a station
- You pay less per mile

ZipCar

- You don't buy the car
- You pay by the mile
- You may pay more per mile

AnVIL: Data + Sharing + Platforms



AnVIL Dataset Catalog

Search

e.g. disease, study name, dbGaP I(

Consortium	Cohorts
□ 1000 Genomes	1
□ CCDG	196
□ CMG	36
□ Convergent Neuro	2
+ 4 more	

Diseases	Cohorts
□ Alzheimer's disease	3
□ asthma	1
□ atherosclerosis	1
□ atrial fibrillation	13
+ 17 more	

Data Types	Cohorts
□ Exome	138
□ RNAseq	1
□ Whole Genome	108

Consent Code	Cohorts
□ DS	4
□ DS-ASD	18
□ DS-ASD-IRB	5
□ DS-ASD-IRB-COL	1
+ 69 more	

Access	Cohorts
□ Consortium Access	99
□ Controlled Access	145
□ Open Access	2

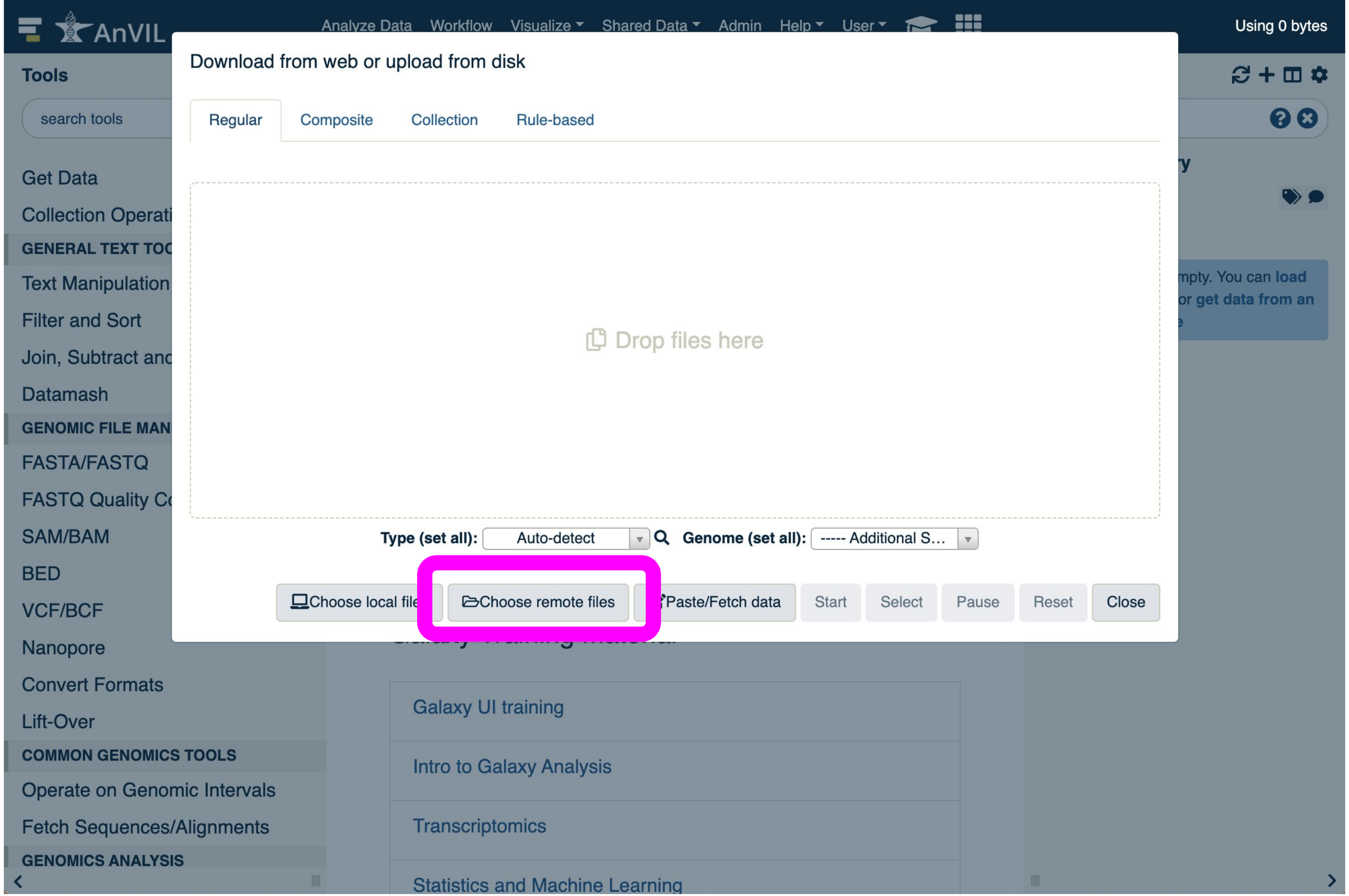
No selected terms.

Download TSV Copy URL

Search Summary

Consortium	Cohorts	Samples	Subjects	Size (TB)
1000 Genomes	1	3,202	3,202	72.98
CCDG	196	250,770	243,226	2,381.24
CMG	36	11,424	10,063	73.61
Convergent Neuro	2	304	304	5.32
GTEx (v8)	1	17,382	979	182.14

https://anvilproject.org/data



Download from web or upload from disk

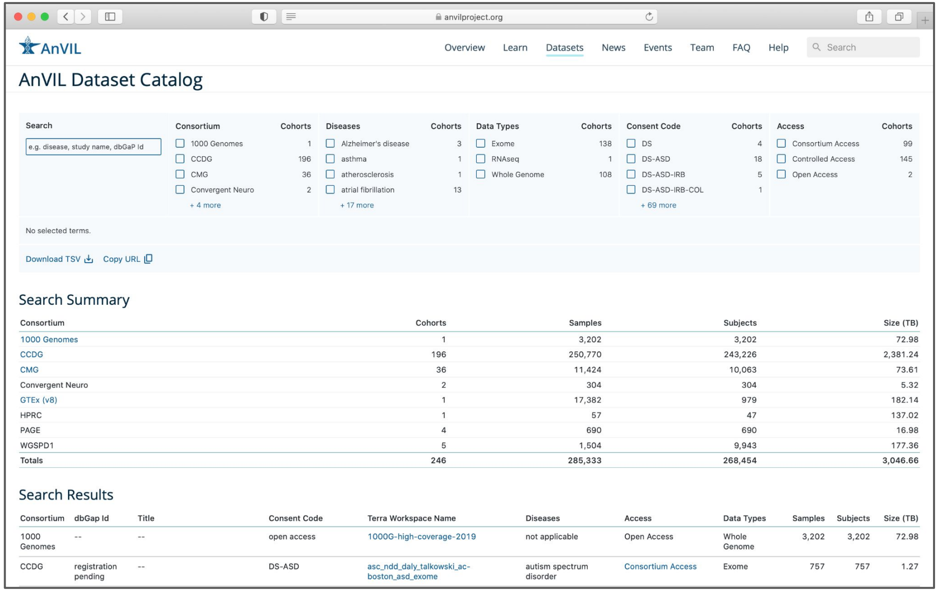
Drop files here

Type (set all): Auto-detect **Genome (set all):** ----- Additional S...

AnVIL Data Dashboard

Extensive unrestricted and protected data sets already available within AnVIL

- 246 cohorts (CCDG, CMG, GTEx,1000G, eMerge)
- 285k subjects
- 3Pb and rapidly growing
- Open access (e.g. 1000G), dbGaP authenticated (e.g. GTex) and consortium authenticated (e.g. CCDG) options available



AnVIL Dataset Catalog

Search	Consortium	Cohorts	Diseases	Cohorts	Data Types	Cohorts	Consent Code	Cohorts	Access	Cohorts
e.g. disease, study name, dbGaP Id	□ 1000 Genomes	1	□ Alzheimer's disease	3	□ Exome	138	□ DS	4	□ Consortium Access	99
	□ CCDG	196	□ asthma	1	□ RNAseq	1	□ DS-ASD	18	□ Controlled Access	145
	□ CMG	36	□ atherosclerosis	1	□ Whole Genome	108	□ DS-ASD-IRB	5	□ Open Access	2
	□ Convergent Neuro	2	□ atrial fibrillation	13			□ DS-ASD-IRB-COL	1		
	+ 4 more		+ 17 more				+ 69 more			

No selected terms.

Download TSV Copy URL

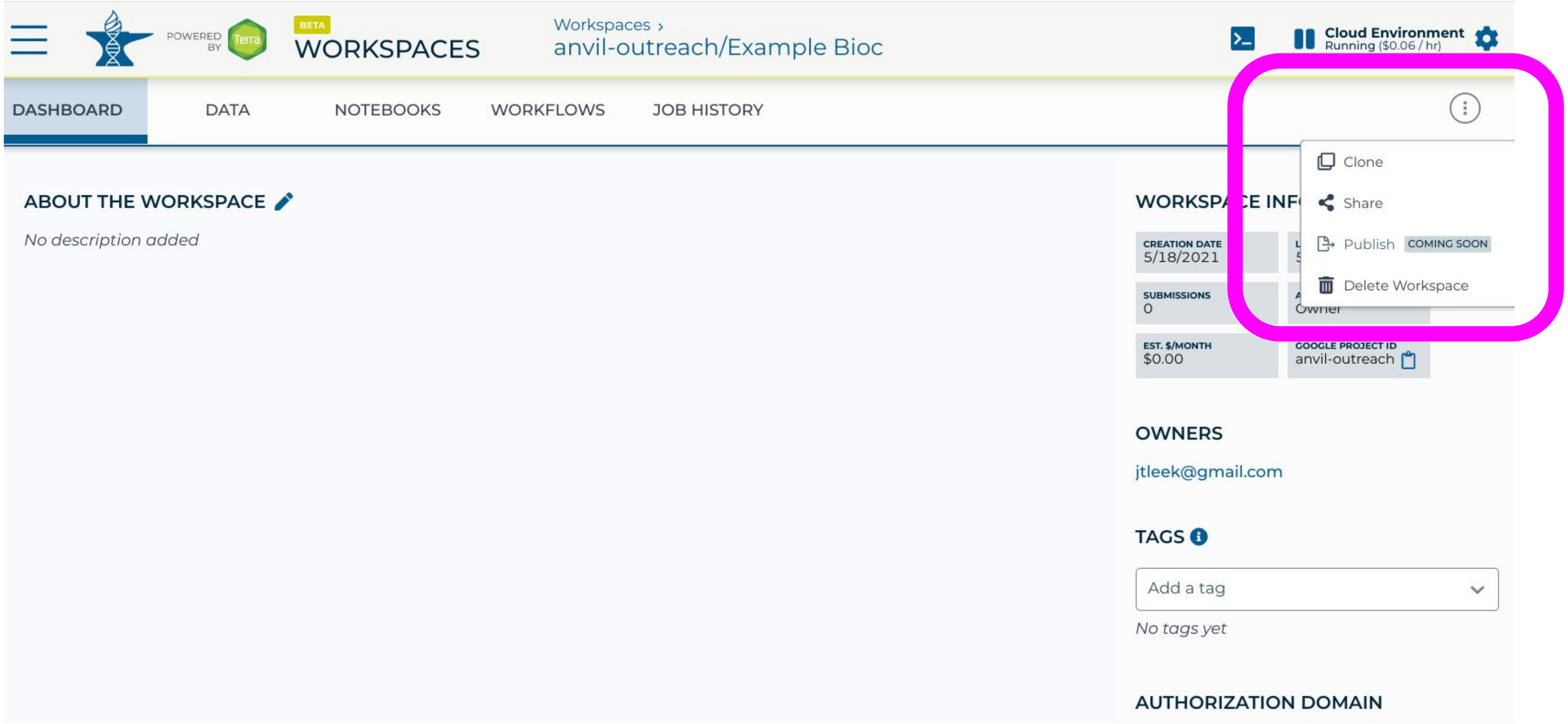
Search Summary

Consortium	Cohorts	Samples	Subjects	Size (TB)
1000 Genomes	1	3,202	3,202	72.98
CCDG	196	250,770	243,226	2,381.24
CMG	36	11,424	10,063	73.61
Convergent Neuro	2	304	304	5.32
GTEx (v8)	1	17,382	979	182.14
HPRC	1	57	47	137.02
PAGE	4	690	690	16.98
WGSPD1	5	1,504	9,943	177.36
Totals	246	285,333	268,454	3,046.66

Search Results

Consortium	dbGap Id	Title	Consent Code	Terra Workspace Name	Diseases	Access	Data Types	Samples	Subjects	Size (TB)
1000 Genomes	--	--	open access	1000G-high-coverage-2019	not applicable	Open Access	Whole Genome	3,202	3,202	72.98
CCDG	registration pending	--	DS-ASD	asc_ndd_daly_talkowski_ac-boston_asd_exome	autism spectrum disorder	Consortium Access	Exome	757	757	1.27

https://anvilproject.org/data



anvil-outreach/Example Bioc

Running ($0.06 / hr)

ABOUT THE WORKSPACE

No description added

WORKSPA CE INF

CREATION DATE 5/18/2021

SUBMISSIONS 0

Owner

EST. $/MONTH $0.00

GOOGLE PROJECT ID anvil-outreach

OWNERS

jtleek@gmail.com

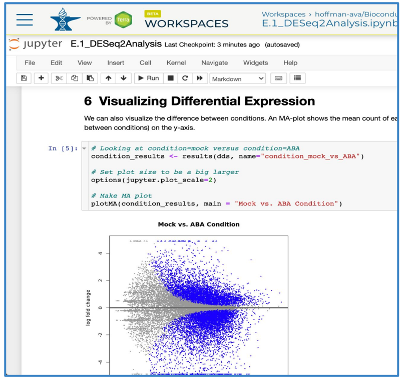
TAGS

Add a tag

No tags yet

AUTHORIZATION DOMAIN

AnVIL Analysis Platforms



6 Visualizing Differential Expression

We can also visualize the difference between conditions. An MA-plot shows the mean count of ea
between conditions) on the y-axis.

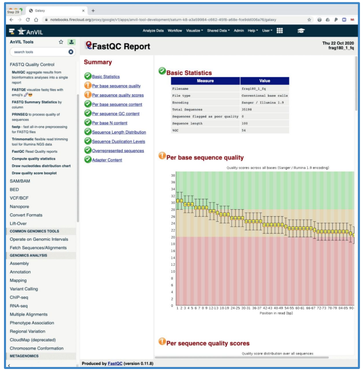
In [5]:

```
# Looking at condition=mock versus condition=ABA
condition_results <- results(dds, name="condition_mock_vs_ABA")

# Set plot size to be a big larger
options(jupyter.plot_scale=2)

# Make MA plot
plotMA(condition_results, main = "Mock vs. ABA Condition")
```

Mock vs. ABA Condition

jupyter

- + Code, text and plots in one document
- + Supports coding in Python or R
- - Least scalable, not a complete IDE


Galaxy
notebooks.firecloud.org/proxy/google/v1/apps/anvil-tool-development/saturn-k8-a3a59984-c662-45f8-a68e-fce9dd006a76/galaxy
AnVIL
Analyze Data
Workflow
Visualize
Shared Data
Admin
Help
User
AnVIL Tools
search tools
FASTQ Quality Control
SAM/BAM
BED
VCF/BCF
Nanopore
Convert Formats
Lift-Over
Quality scores across all bases (Sanger / Illumina 1.9 encoding)
Position in read (bp)

FastQC Report

Thu 22 Oct 2020
frag180_1_fq

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

Measure	Value
Filename	frag180_1_fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	35198
Sequences flagged as poor quality	0
Sequence length	100
%GC	54

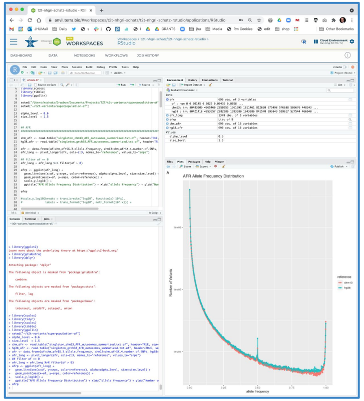
Per base sequence quality

Per sequence quality scores

Produced by FastQC (version 0.11.8)

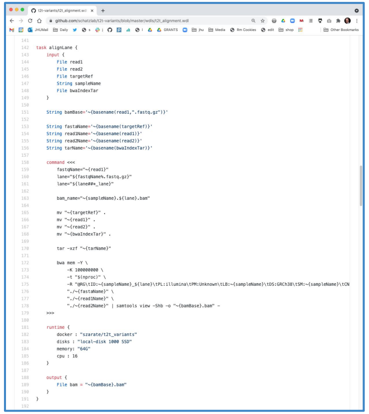
Galaxy

- + Graphical interface for thousands of tools and workflows
- + Highly accessible and reproducible
- - Tools must be preconfigured to use



RStudio

- + Feature rich IDE for programming in R
- + Rich statistics & ML and visualizations
- - Limited support for other programming languages

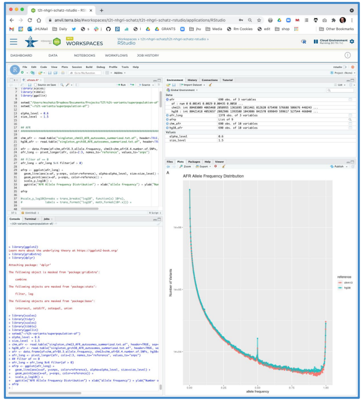

t2t-variants/t2t_alignment.wdl
github.com/schatzlab/t2t-variants/blob/master/wdls/t2t_alignment.wdl

```
task alignLane {
    input {
        File read1
        File read2
        File targetRef
        String sampleName
        File bwaIndexTar
    }

    String bamBase='~{basename(read1,".fastq.gz")}'

    String fastaName='~{basename(targetRef)}'
    String read1Name='~{basename(read1)}'
    String read2Name='~{basename(read2)}'
    String tarName='~{basename(bwaIndexTar)}'

    command <<<
        fastqName="~{read1}"
        lane="${fastqName%.fastq.gz}"
        lane="${lane##*_lane}"

        bam_name="~{sampleName}.${lane}.bam"

        mv "~{targetRef}" .
        mv "~{read1}" .
        mv "~{read2}" .
        mv "~{bwaIndexTar}" .

        tar -xzf "~{tarName}"

        bwa mem -Y \
            -K 100000000 \
            -t "$(nproc)" \
            -R "@RG\tID:~{sampleName}_${lane}\tPL:illumina\tPM:Unknown\tLB:~{sampleName}\tDS:GRCh38\tSM:~{sampleName}\tCN
            "./~{fastaName}" \
            "./~{read1Name}" \
            "./~{read2Name}" | samtools view -Shb -o "~{bamBase}.bam" -
    >>>

    runtime {
        docker : "szarate/t2t_variants"
        disks : "local-disk 1000 SSD"
        memory: "64G"
        cpu : 16
    }

    output {
        File bam = "~{bamBase}.bam"
    }
}
```

{wdl}

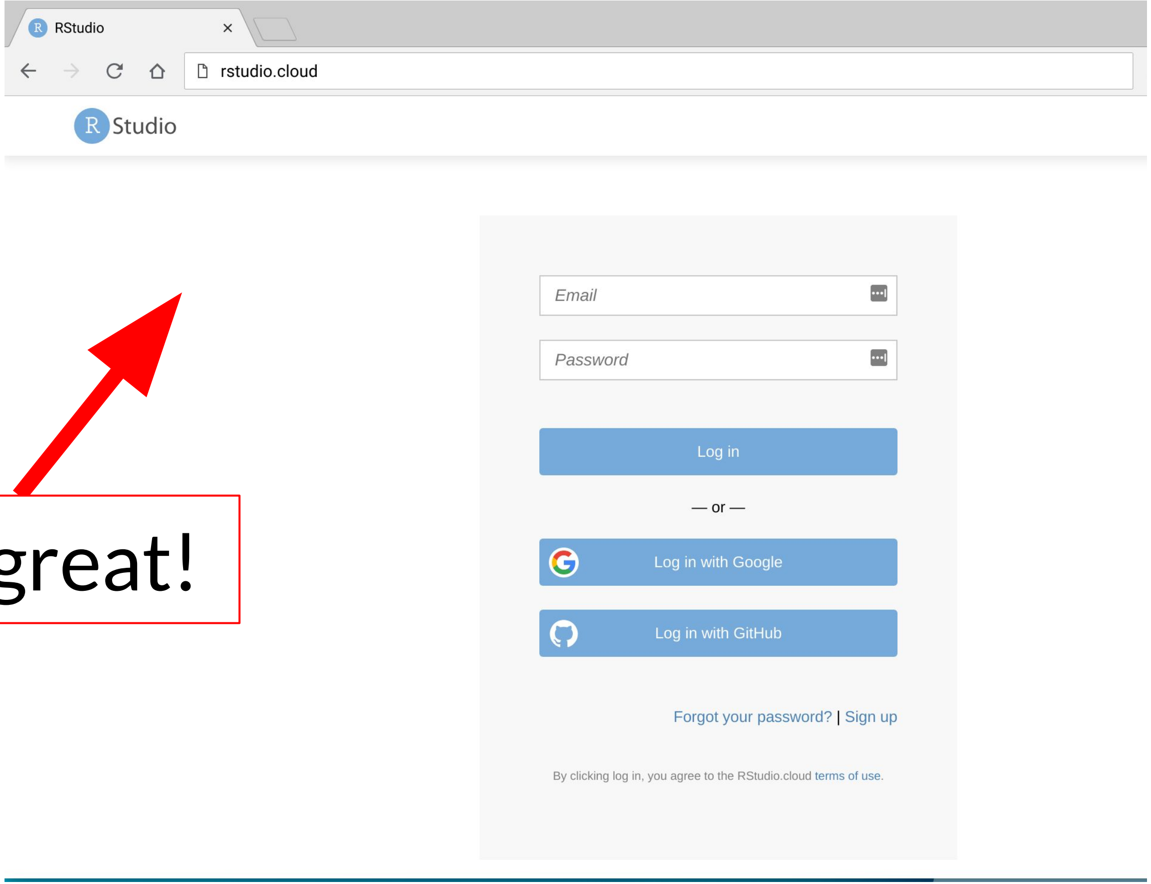
- + Extremely scalable and flexible
- - Most technically demanding
- - Unpredictable and potentially large costs

AnVIL Analysis Platforms



RStudio

+ Feature rich IDE for programming in R
+ Rich statistics & ML and visualizations
- Limited support for other programming languages



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

Also great!

https://rstudio.cloud