

# Data Wrangling in R

<http://sisbid.github.io/Module1/>

# Preliminaries

# Course Info

Course name	Data Wrangling in R
Instructors	Andrew Jaffe and John Muschelli
Course website	<a href="http://sisbid.github.io/Module1/">http://sisbid.github.io/Module1/</a>
Goals	Teach you how to get and clean data
Pre-reqs	Hopefully some R programming

How many people feel  
about data wrangling



# How we feel about data wrangling



# About us

(John)

# Welcome to the Leek group

Welcome to the Leek group in the [Data Science Lab](#) and the [Department of Biostatistics](#) at the Johns Hopkins Bloomberg School of Health. We are a [group](#) of researchers, educators, and data scientists using data to solve [problems](#) in molecular biology, human health, meta-research, education, and anything else we think could be useful for the world. We produce [data tools and code](#) that you can use for your projects as well. We teach [online open classes](#) so you can learn how to use data too. If you think any of this sounds cool consider [joining us](#) in working to make the world a better place. If you just want to keep up with everything we are working on, follow Jeff on Twitter <https://twitter.com/jtleek>.

# about



## Links

[Blog](#)

[HopStat](#)

## Social Media

[Twitter](#)

[GitHub](#)

[YouTube Channel](#)

## Classes

[Introduction to R](#)

[Faculty Office Hour](#)

## Short Courses

[Neurohacking](#)

[Imaging in R \(aka Neurohacking 2.0\)](#)

[Building R Packages](#)

# A HopStat and Jump Away

*Trying to at least Doggie Paddle through the Sea of Data, Contributor to <http://bmorebiostat.com>*

Home

Why are you here

Search ...

Search

## Recent Posts

[The 3 ‘Times’ of a Project](#)

[Some Thoughts as a Junior Faculty  
\(at JHSPH\)](#)

[The way people use AI is ruining  
Reproducible Science Again](#)  
[R projects may make large files](#)

## The 3 ‘Times’ of a Project

Edit

Posted on May 15, 2020 by [strictlystat](#)

During a conversation with [Sean Kross](#) about projects, particularly data science projects, I tried to explain how things can go right and wrong with a project. I was explaining things with respect to being the data scientist on academic projects, but I think these issues are cross-cutting so figured I'd post them here.

I thought back to when projects did not go well or someone was left frustrated or angry during or at the end of the interaction. To me, the issues usually come down to the 3 “time”s of a project: time, [timeline](#), and [timeliness](#).

# About us

(Andrew)

# Andrew Jaffe

- Lead Investigator at the Lieber Institute for Brain Development
- Associate Professor at Johns Hopkins University (Mental Health, Biostats, Psychiatry, and Human Genetics)
- Run academic data science team
- My research focuses on molecular correlations of psychiatric brain disorders like schizophrenia, bipolar disorder, and major depression

# Overview

The Jaffe Lab is led by Andrew E Jaffe.

The lab is associated with the [Lieber Institute for Brain Development](#) and the Departments of [Mental Health](#) and [Biostatistics](#) at Johns Hopkins Bloomberg School of Public Health.

We are also part of the [Center for Computational Biology](#) at Johns Hopkins University.

# Research Interests

We are a computational biology and genomics lab within the Lieber Institute for Brain Development (LIBD). We are interested in better understanding and characterizing genomics signatures in the human brain, including DNA methylation and gene expression.

# Contact

Email:

@andrewejaffe  
<http://www.aejaffe.com>

# Why this class



# rmarkdown

---

```
title: "My awesome website"
```

```
output:
```

```
  html_document:
```

```
    toc: true
```

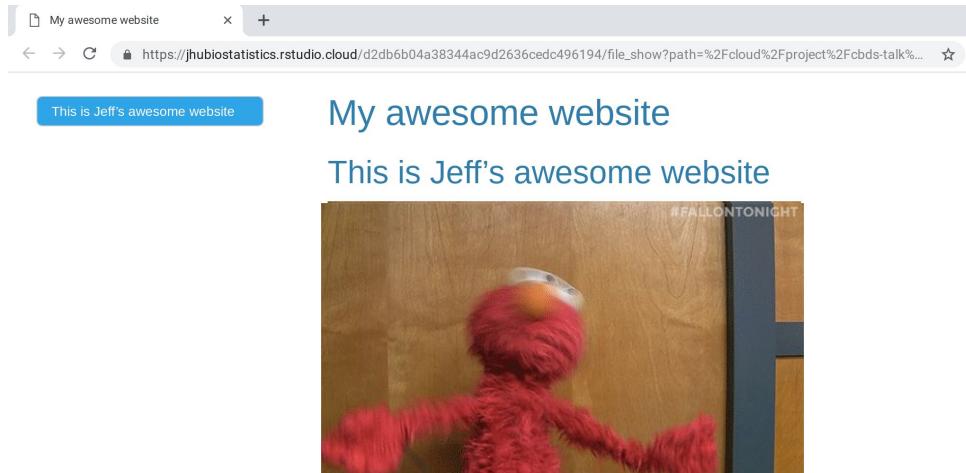
```
    toc_float: true
```

```
    theme: cerulean
```

---

```
# This is Jeff's awesome website
```

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



# dbplyr

```
library(bigrquery)
set_service_token("file.json"))

con <- dbConnect(
  bigrquery(),
  project = "project_name",
  dataset = "dataset_name"
)

unique_elements = con %>%
 tbl("dataset1") %>%
  count()
```

```
unique_elments
Running job 'job_id.US'...
Complete
Billed: 32.51 MB
Downloading 10 rows in 1 pages.
# Source:   lazy query [?? x 2]
# Database: BigQueryConnection
```

	n
	<int>
1	3700675

# httr

```
library(httr)
library(dplyr)

username = 'janeeeverydaydoe'

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

8 commits 1 branch 0 releases 1 contributor

Branch: master New pull request Create new file Upload files Find file Clone or download

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

README.md

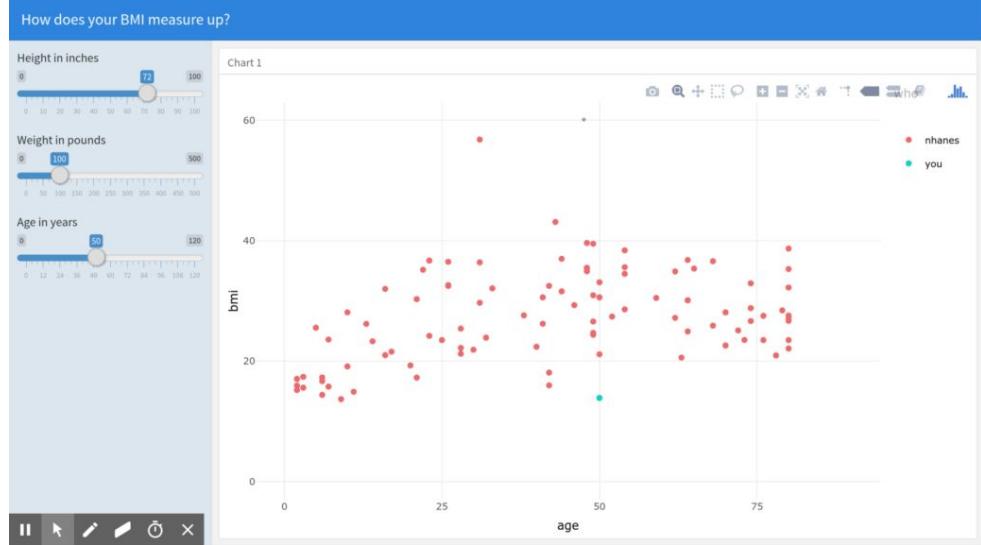
first\_project

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"
```

# 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()  
)  
})  
```
```



But also...

# Genomic signatures to guide the use of chemotherapeutics

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

**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

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

Keith A. Baggerly

Bioinformatics and Computational Biology

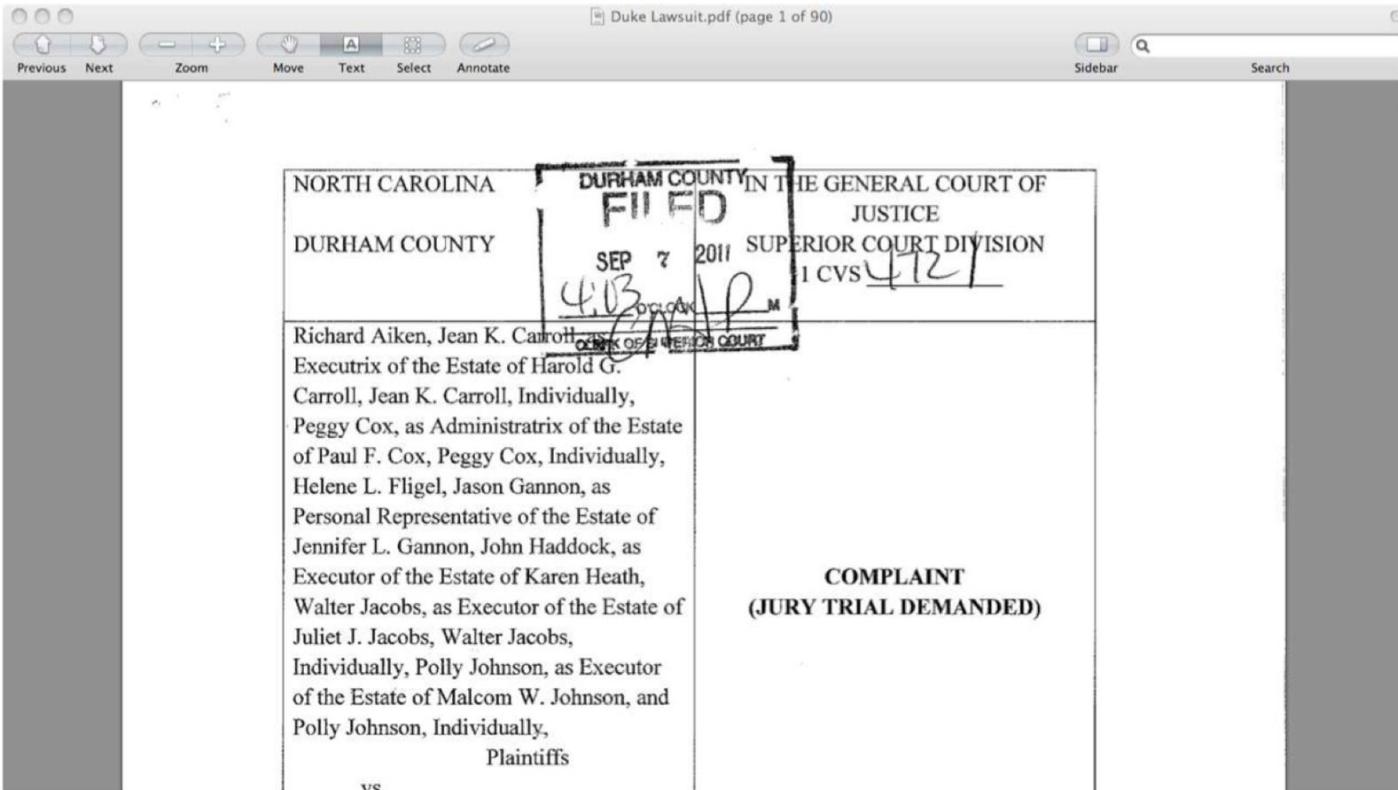
UT M. D. Anderson Cancer Center

[kabagg@mdanderson.org](mailto:kabagg@mdanderson.org)



BIRS Workshop, Aug 14, 2013





Doesn't seem that important....

Thu 1:58 AM

```
> 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)

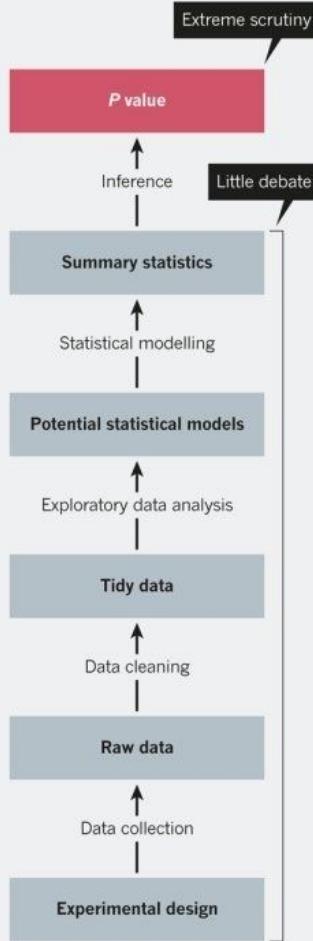
|                                       | Estimate | Std. Error | z value | Pr(> z ) |
|---------------------------------------|----------|------------|---------|----------|
| (Intercept)                           | 0.5675   | 0.1475     | 3.847   | 0.000122 |
| dat\$study_type <non-anon></non-anon> | 0.4250   | 0.2182     | 1.948   | 0.051458 |

A man in a blue suit and red tie, holding a briefcase, stands in a landscape with mountains and a city.

**ON THE ONE  
HAND...**

## DATA PIPELINE

The design and analysis of a successful study has many stages, all of which need policing.



- 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

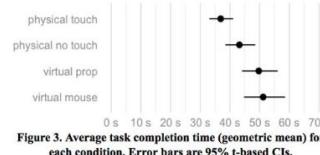
# The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time\*

Andrew Gelman<sup>†</sup> and Eric Loken<sup>‡</sup>

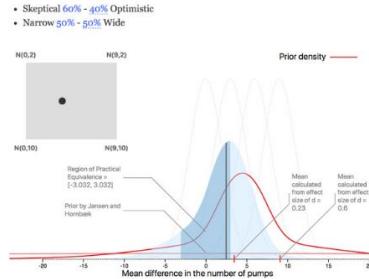
14 Nov 2013

*“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

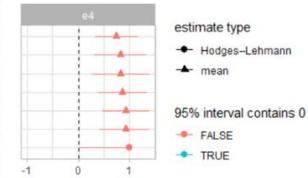


We focus our analysis on task completion times, reported in Figures 3 and 4. Dots indicate sample means, while error bars are 95% confidence intervals computed on [log-transformed data](#) [6], using the [t-distribution](#) method. Strictly speaking, all we can assert about each interval is that it comes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



| $r = 0.1$        | $r = 0.3$        | $r = 0.5$        | $r = 0.7$        | $r = 0.9$        | Overall          |
|------------------|------------------|------------------|------------------|------------------|------------------|
| pcp-neg          | pcp-neg          | scatterplot-pos  | scatterplot-neg  | scatterplot-neg  | scatterplot-pos  |
| scatterplot-pos  | scatterplot-pos  | pop-neg          | scatterplot-pos  | pcp-neg          | pcp-neg          |
| scatterplot-neg  | scatterplot-neg  | scatterplot-neg  | pop-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  | stackedarea-neg  | stackedarea-neg  | stackedarea-neg  |
| ordered line-neg | ordered line-neg | ordered line-neg | ordered line-neg | stackedarea-neg  | ordered line-neg |
| stackedline-neg  | stackedline-neg  | stackedline-neg  | stackedline-neg  | stackedarea-neg  | stackedline-neg  |
| pop-pos          | pop-pos          | pop-pos          | pop-pos          | radar-pos        | pop-pos          |
| radar-pos        | radar-pos        | radar-pos        | radar-pos        | pop-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.

And so we data wrangle

Herein lies the dirty secret about most data scientists' work -- it's more data munging than deep learning. The best minds of my generation are deleting commas from log files, and that makes me sad. A Ph.D. is a terrible thing to waste.



SEARCH

TECHNOLOGY | For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights



TECHNOLOGY

# For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

By STEVE LOHR AUG. 17, 2014



amazon web services | intel CLOUD INSIGHTS

Why Novartis is Looking Beyond On-Premises... [READ >](#)

Case Study: Cloud Supercomputing from AWS Powers... [READ >](#)

**Get Started with AWS**

**CREATE A FREE ACCOUNT >**

# What you wished data looked like

# What it actually looks like

<http://healthdesignchallenge.com/>

```
@HWI-EAS121:4:100:1783:550#0/1
CGTTACGAGATCGAACAGCGGTTCAGCAGGAATGCCGAGACGGATCTCGTATGCCGTCTGCCTGACAAGACAGGGG
+HWI-EAS121:4:100:1783:550#0/1
aaaaaa`b_aa`aa`YaX]aZ`aZM^Z]YRa]YSG[ [ZREQLHESDHNDDHNMEEDDM PENITKFLFEEDDDHEJQM EDDD
@HWI-EAS121:4:100:1783:1611#0/1
GGGTGGGCATTTCCACTCGCAGTATGGGTTGCCGCACGACAGGCAGCGGT CAGCCTGCGCTTGGCCTGGCCTTCGGAAA
+HWI-EAS121:4:100:1783:1611#0/1
a``^\\_`_``^a``a``^a_``^]a_]`a_____`_``^]X]_]XTV_\\_]NX_XVX]_]_TTTG[VTHPN]VFDZ
@HWI-EAS121:4:100:1783:322#0/1
CGTTTATTTTTGAATATGTCTTATCTAACGGTTATTTAGATGTTGGTCTTATTCTAACGGTCATATATTTCTA
+HWI-EAS121:4:100:1783:322#0/1
abaa``aaaaabbbaabbbbbbb`bbbb_bbbbbbbb`bbbaV``_a``]``aT]a__V\\]_``]a``]a_abbaV_
@HWI-EAS121:4:100:1783:1394#0/1
GGGTCTTATTGGTCTGGT GATCCCCATATTCTCCGGTTGTGGTTAACCGATCATCGCGCATTACTCCGGCTGC
+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
```

# What it actually looks like

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

The screenshot shows a web browser window with the Twitter Developers API documentation. The URL in the address bar is <https://dev.twitter.com/docs/api/1/get/blocks/blocking>. The page content includes a note about cursor values, example values, and an example request with a JSON response.

cursor to be -1 if it isn't supplied.  
Example Values: 12893764510938

**Example Request**

GET [https://api.twitter.com/1/blocks/blocking.json?cursor=-1&include\\_entities=true](https://api.twitter.com/1/blocks/blocking.json?cursor=-1&include_entities=true)

```
1. {
2.   "previous_cursor": 0,
3.   "previous_cursor_str": "0",
4.   "next_cursor": 0,
5.   "users": [
6.     {
7.       "profile_sidebar_border_color": "C0DEED",
8.       "name": "Javier Heady \ud83d\udcbb",
9.       "profile_sidebar_fill_color": "DDEEF6",
10.      "profile_background_tile": false,
11.      "location": null,
12.      "created_at": "Thu Mar 01 00:16:47 +0000 2012",
13.      "profile_image_url":
14.        "http://a0.twimg.com/sticky/default_profile_images/default_profile_4_normal.png",
15.        "is_translator": false,
16.        "id_str": "509466276",
17.        "profile_link_color": "0084B4",
18.        "follow_request_sent": false,
19.        "contributors_enabled": false,
20.        "default_profile": true,
21.        "url": null,
22.        "favourites_count": 0,
```

# What it actually looks like

## ALLERGIES

Last Updated: 01 Dec 2011 @ 0851

Allergy Name: TRIMETHOPRIM  
Location: DAYT29

Date Entered: 09 Mar 2011  
Reaction:

Allergy Type: DRUG

A Drug Class: ANTI-INFECTIVES, OTHER

Observed/Historical: HISTORICAL

Comments: The reaction to this allergy was MILD (NO SQUELAE)

Allergy Name: TRAMADOL

Location: DAYT29

## MEDICATION HISTORY

Last Updated: 11 Apr 2011 @ 1737

Medication: AMLODIPIINE BESYLATE 10MG TAB

Instructions: TAKE ONE TABLET BY MOUTH TAKE ON GRAPEFRUIT JUICE--

Status: Active

Refills Remaining: 3

Last Filled On: 28 Aug 2010

Initially Ordered On: 13 Aug 2010

Quantity: 45

Days Supply: 90

Pharmacy: DAYTON

Prescription Number: 2718953



**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](https://github.com/jennybc/2016-06_spreadsheets/blob/master/2016-06_useR-stanford.pdf))

| A  | B                                                                                | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U |
|----|----------------------------------------------------------------------------------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1  |  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 2  |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 3  |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 4  |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 5  |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 6  |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 7  |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 8  |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 9  |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 10 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 11 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 12 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 13 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 14 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 15 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 16 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 17 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 18 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 19 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 20 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 21 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 22 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 23 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 24 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 25 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 26 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 27 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 28 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 29 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 30 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 31 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 32 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 33 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 34 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 35 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 36 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 37 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 38 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 39 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 40 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 41 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 42 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 43 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 44 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 45 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 46 |                                                                                  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |

Slide from Jenny Bryan

([https://github.com/jennybc/2016-06\\_spreadsheets/blob/master/2016-06\\_useR-stanford.pdf](https://github.com/jennybc/2016-06_spreadsheets/blob/master/2016-06_useR-stanford.pdf))



Desiree Narango

@DLNarango

Follow



Today's updates on #otherpeoplesdata:



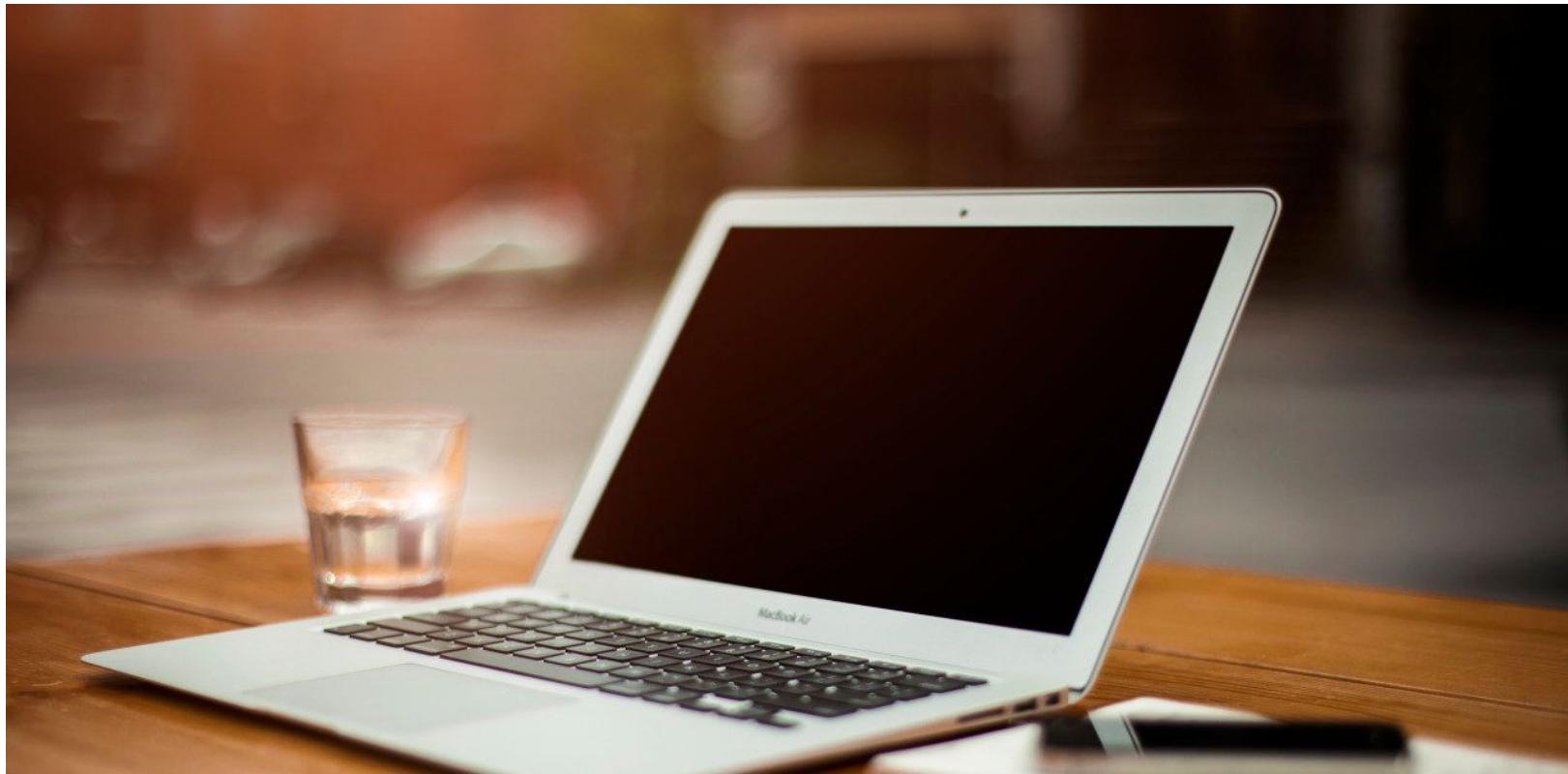
8:56 AM - 22 Oct 2018

---

1 Like

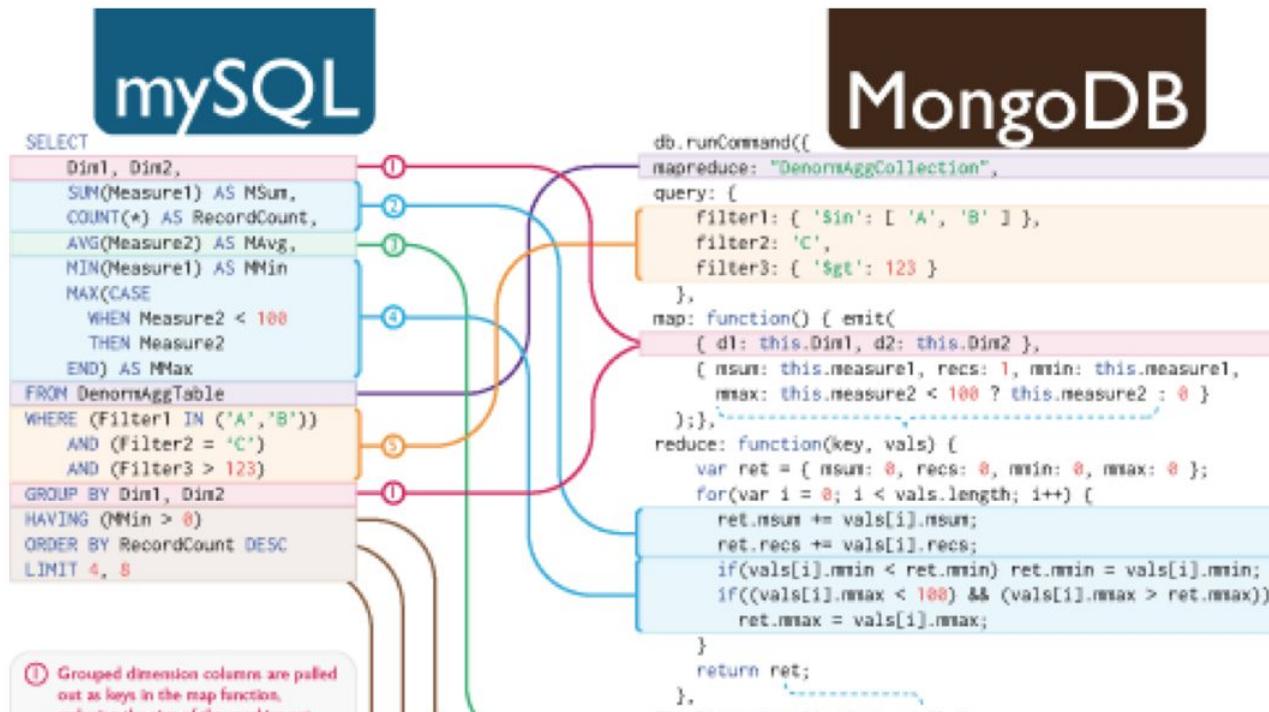


# Where you wish data was



# Where data actually is

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



# Where data actually is

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

The screenshot shows a web browser window with the URL <https://dev.twitter.com/docs/api/1/get/blocks/blocking> in the address bar. The page is titled "GET blocks/blocking | Twitter API". The main content area displays the API documentation for the GET blocks/blocking endpoint, including example values and an example request.

**Example Values:** 12893764510938

**Example Request**

**GET** [https://api.twitter.com/1/blocks/blocking.json?cursor=-1&include\\_entities=true](https://api.twitter.com/1/blocks/blocking.json?cursor=-1&include_entities=true)

```
1. {
2.   "previous_cursor": 0,
3.   "previous_cursor_str": "0",
4.   "next_cursor": 0,
5.   "users": [
6.     {
7.       "profile_sidebar_border_color": "CODEED",
8.       "name": "Javier Heady \r",
9.       "profile_sidebar_fill_color": "DDEEF6",
10.      "profile_background_tile": false,
11.      "location": null,
12.      "created_at": "Thu Mar 01 00:16:47 +0000 2012",
13.      "profile_image_url":
14.        "http://a0.twimg.com/sticky/default_profile_images/default_profile_4_normal.png",
15.      "is_translator": false,
16.      "id_str": "509466276",
17.      "profile_link_color": "0084B4",
18.      "follow_request_sent": false,
19.      "contributors_enabled": false
20.    }
21.  ]
22.}
```

# Raw & processed data

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

“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

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

**Variables:** A measurement or characteristic of an item

“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



# *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.

# A tidy data set

|    | A  | B          | C          | D          | E          | F         | G      | H | I | J | K | L | M | N | O | P |
|----|----|------------|------------|------------|------------|-----------|--------|---|---|---|---|---|---|---|---|---|
| 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 variable  
Linking indicators for columns

Decoder.docx

# 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 prodromal and schizophrenia. Not sure if this was ok. Is it appropriate to do a ttest between SZ and C?

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

```
33 library(sva)
34 library(affy)
35 library(RColorBrewer)
36 library(corrplot)
37 library(limma)
38 trop = RSkittleBrewer('tropical')
39 ...
40
41
42 ## Load the data
43
44 You will need to download the GEUVADIS ballgown object from this site: https://github.com/ctazee/ballgown\_code
45
46
47 ```{r loaddata, dependson="load"}
48 load("fpkm.rda")
49 pd = ballgown::pData(fpkm)
50 pd$dirname = as.character(pd$dirname)
51 ss = function(x, pattern, slot=1,...) sapply(strsplit(
52 pd$IndividualID = ss(pd$dirname, "_", 1)
53 tfpkm = expr(fpkm)$trans
54 ...
55
56 ## Subset to non-duplicates
57
58 You will need the GEUVADIS quality control information and population information available from these
1:1  (Top Level) 
```



R/Python Code  
Input raw data -> output tidy  
No parameters

recipe.docx

Home Layout Document Elements Tables Charts SmartArt Review

Cambria (Body) 15 A A Aa Ab B I U ABC A<sup>2</sup> Aa A<sup>BD</sup> Aa

Font Paragraph Styles Insert Themes

AaBbCcDdEe AaBbCcDdEe AaBbCcDdEe Normal No Spacing Heading 1 AA Text Box Shape Picture Themes

1 2 3 4 5 6 7

1| 2|

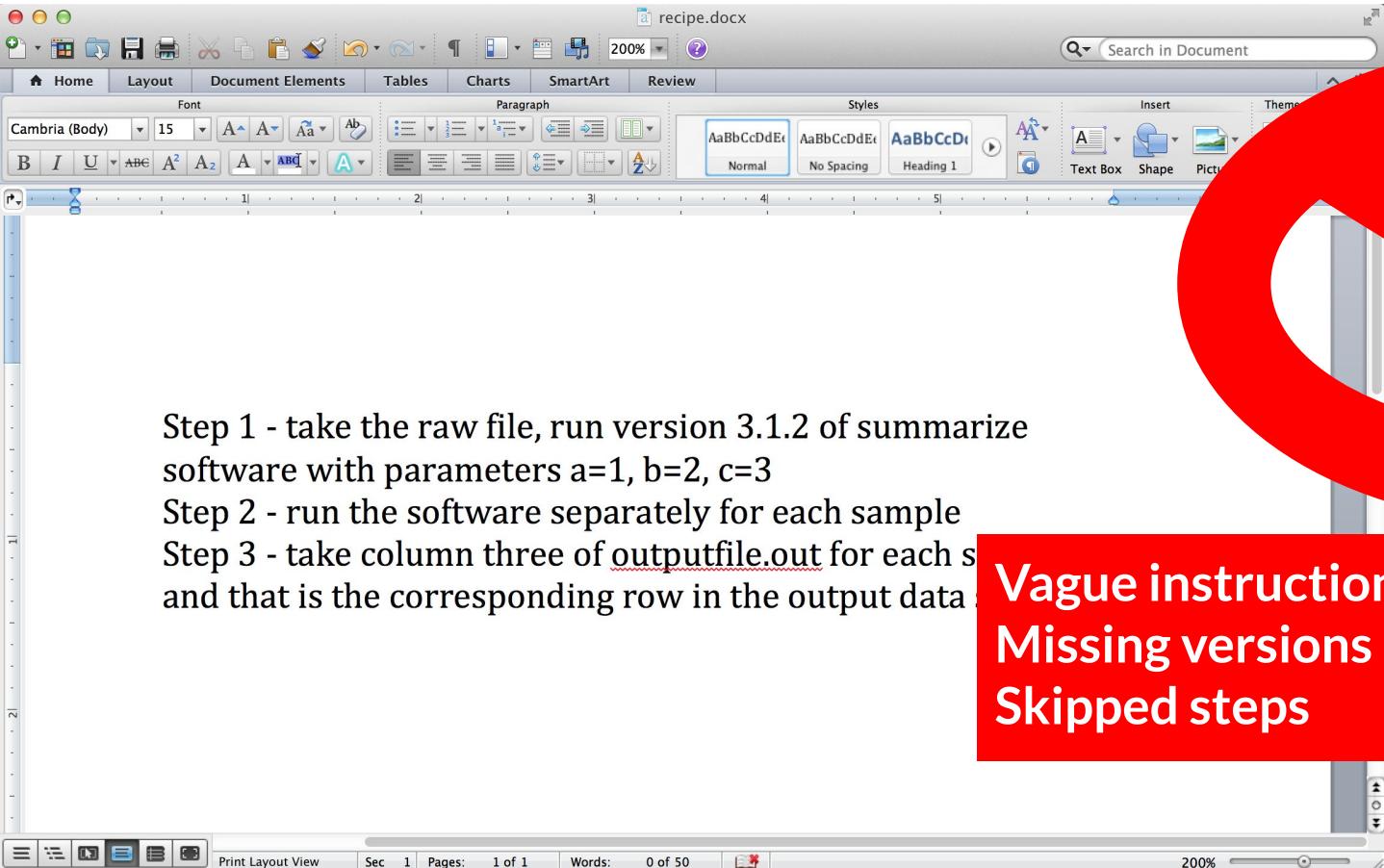
Print Layout View Sec 1 Pages: 1 of 1 Words: 0 of 50 200%

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 sample and that is the corresponding row in the output data

Explicit instructions  
Versions of software  
Parameters included



recipe.docx

Home Layout Document Elements Tables Charts SmartArt Review

Font Paragraph Styles Insert Themes

Cambria (Body) 15 A A Aa Ab B I U ABC A<sup>2</sup> Aa A ABD A Aa

Normal No Spacing Heading 1

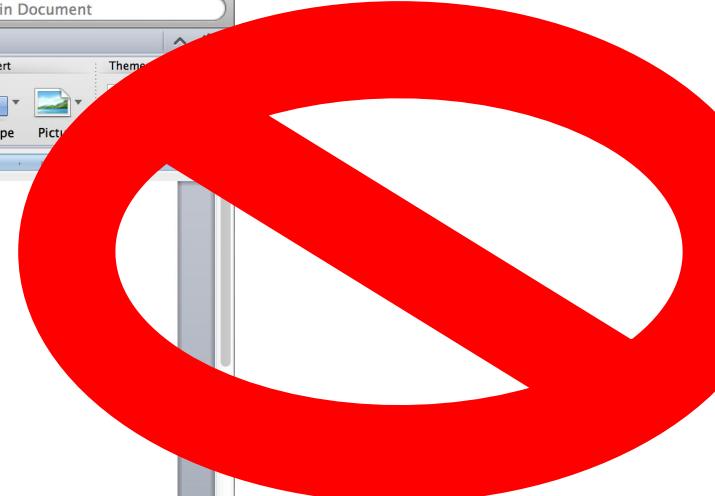
Text Box Shape Picture

1 2 3 4 5

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<br>(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.                                                                                                               |

# 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

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 |
| 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

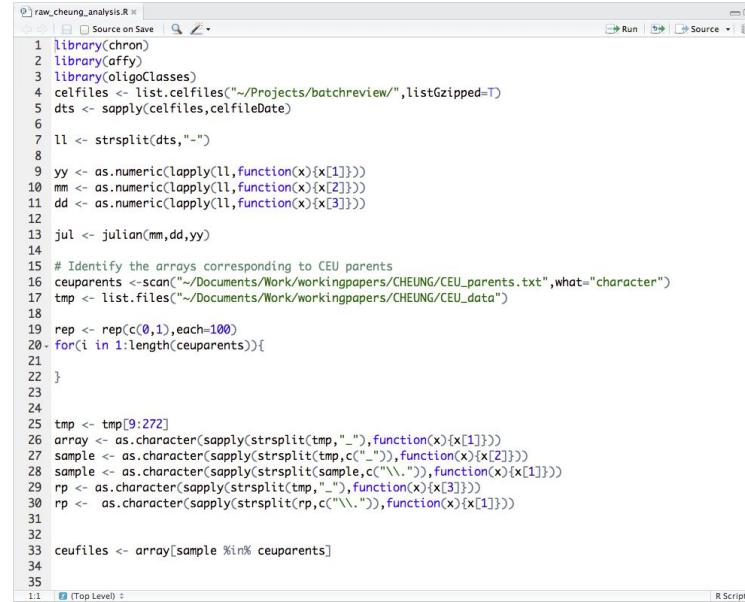
| ALLERGIES                        |                                                    | MEDICATION HISTORY               |                                                                      |
|----------------------------------|----------------------------------------------------|----------------------------------|----------------------------------------------------------------------|
| Last Updated: 01 Dec 2011 @ 0851 |                                                    | Last Updated: 11 Apr 2011 @ 1737 |                                                                      |
| Allergy Name:                    | TRIMETHOPRIM                                       | Medication:                      | AMLODIPINE BESYLATE 10MG TAB                                         |
| Location:                        | DAYT29                                             | Instructions:                    | TAKE ONE TABLET BY MOUTH TAKE ONE-HALF TABLET FOR GRAPEFRUIT JUICE-- |
| Date Entered:                    | 09 Mar 2011                                        | Status:                          | Active                                                               |
| Action:                          |                                                    | Refills Remaining:               | 3                                                                    |
| Allergy Type:                    | DRUG                                               | Last Filled On:                  | 28 Aug 2010                                                          |
| A Drug Class:                    | ANTI-INFECTIVES, OTHER                             | Initially Ordered On:            | 13 Aug 2010                                                          |
| Observed/Historical:             | HISTORICAL                                         | Quantity:                        | 45                                                                   |
| Comments:                        | The reaction to this allergy was MILD (NO SQUELAE) | Days Supply:                     | 90                                                                   |
| Allergy Name:                    | TRAMADOL                                           | Pharmacy:                        | DAYTON                                                               |
| Location:                        | DAYT29                                             | Prescription Number:             | 2718953                                                              |
| Date Entered:                    | 09 Mar 2011                                        | Medication:                      | IBUPROFEN 600MG TAB                                                  |
| Action:                          | URINARY RETENTION                                  | Instructions:                    | TAKE ONE TABLET BY MOUTH FOUR TIMES A DAY WITH FOOD                  |
| Allergy Type:                    | DRUG                                               | Status:                          | Active                                                               |
| A Drug Class:                    | NON-OPIOID ANALGESICS                              | Refills Remaining:               | 3                                                                    |
| Observed/Historical:             | HISTORICAL                                         | Last Filled On:                  | 28 Aug 2010                                                          |
| Comments:                        | gradually worsening difficulty emptying bladder    | Initially Ordered On:            | 01 Jul 2010                                                          |

# Processed data

|    | A  | B          | C          | D          | E          | F         | G | H | I | J | K | L | M | N | O | P |
|----|----|------------|------------|------------|------------|-----------|---|---|---|---|---|---|---|---|---|---|
| 1  | Id | problem_id | subject_id | start      | stop       | time_left |   |   |   |   |   |   |   |   |   |   |
| 2  | 1  | 498        | 17         | 1307119989 | 1307120016 | 2369      | A |   |   |   |   |   |   |   |   |   |
| 3  | 2  | 150        | 15         | 1307119991 | 1307120009 | 2376      | D |   |   |   |   |   |   |   |   |   |
| 4  | 3  | 313        | 16         | 1307119992 | 1307120010 | 2376      | E |   |   |   |   |   |   |   |   |   |
| 5  | 4  | 32         | 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        | 17         | 1307120004 | 1307120038 | 2343      | C |   |   |   |   |   |   |   |   |   |
| 10 | 9  | 70         | 15         | 1307120007 | 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 | 1307120186 | 2227      | C |   |   |   |   |   |   |   |   |   |
| 16 | 15 | 344        | 15         | 1307120031 | 1307120117 | 2208      | 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         | 1307120150 | 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         | 1307120283 | 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         | 1307120352 | 1307120362 | 2025      | 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 | 1307120513 | 1870      | E |   |   |   |   |   |   |   |   |   |
| 39 | 38 | 257        | 14         | 1307120401 | 1307120527 | 1935      | 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 |   |   |   |   |   |   |   |   |   |

- 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

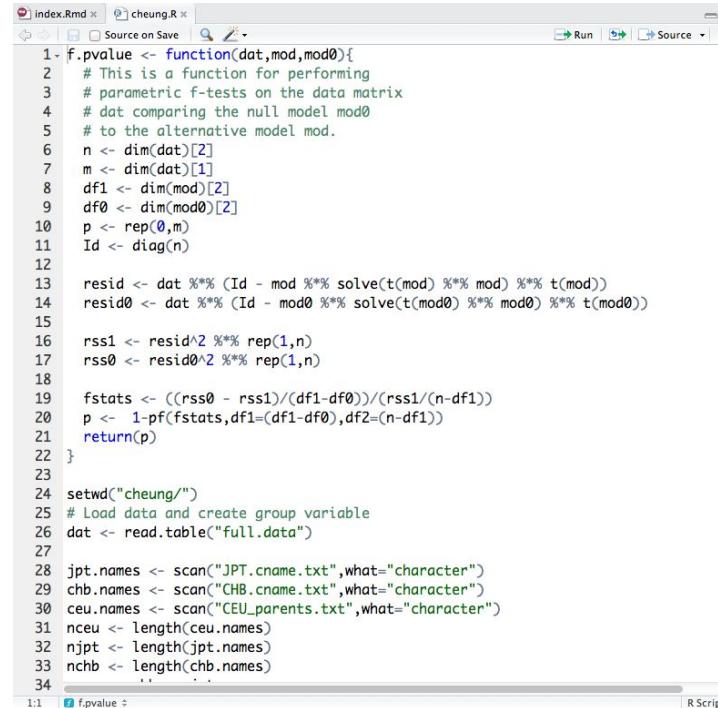


```
raw_cheung_analysis.R < Source on Save Run Source
1 library(chron)
2 library(affy)
3 library(oligoClasses)
4 celfiles <- list.celfiles("~/Projects/batchreview/",listGzipped=T)
5 dts <- sapply(celfiles,celfileDate)
6
7 ll <- strsplit(dts,"-")
8
9 yy <- as.numeric(lapply(ll,function(x){x[1]}))
10 mm <- as.numeric(lapply(ll,function(x){x[2]}))
11 dd <- as.numeric(lapply(ll,function(x){x[3]}))
12
13 jul <- julian(mm,dd,yy)
14
15 # Identify the arrays corresponding to CEU parents
16 ceuparents <- scan("~/Documents/Work/workingpapers/CHEUNG/CEU_parents.txt",what="character")
17 tmp <- list.files("~/Documents/Work/workingpapers/CHEUNG/CEU_data")
18
19 rep <- rep(c(0,1),each=100)
20 for(i in 1:length(ceuparents)){
21
22 }
23
24
25 tmp <- tmp[9:272]
26 array <- as.character(sapply(strsplit(tmp,"_"),function(x){x[1]}))
27 sample <- as.character(sapply(strsplit(tmp,c("_")),function(x){x[2]}))
28 sample <- as.character(sapply(strsplit(sample,c("\\.")),function(x){x[1]}))
29 rp <- as.character(sapply(strsplit(tmp,"."),function(x){x[3]}))
30 rp <- as.character(sapply(strsplit(rp,c("\\\\.")),function(x){x[1]}))
31
32
33 ceufiles <- array[sample %in% ceuparents]
34
35
```

R Script

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

# Final scripts



The screenshot shows the RStudio interface with two tabs open: "index.Rmd" and "cheung.R". The "cheung.R" tab is active, displaying an R script. The script contains several commented sections and functions. It includes code for performing f-tests, setting working directories, reading tables, and scanning character vectors. The code is well-organized with comments explaining its purpose and flow.

```
1- f.pvalue <- function(dat,mod,mod0){  
2 # This is a function for performing  
3 # parametric f-tests on the data matrix  
4 # dat comparing the null model mod0  
5 # to the alternative model mod.  
6 n <- dim(dat)[2]  
7 m <- dim(dat)[1]  
8 df1 <- dim(mod)[2]  
9 df0 <- dim(mod0)[2]  
10 p <- rep(0,m)  
11 Id <- diag(n)  
12  
13 resid <- dat %*% (Id - mod %*% solve(t(mod) %*% mod) %*% t(mod))  
resid0 <- dat %*% (Id - mod0 %*% solve(t(mod0) %*% mod0) %*% t(mod0))  
15  
16 rss1 <- resid^2 %*% rep(1,n)  
rss0 <- resid0^2 %*% rep(1,n)  
18  
19 fstats <- ((rss0 - rss1)/(df1-df0))/(rss1/(n-df1))  
20 p <- 1-pf(fstats,df1=(df1-df0),df2=(n-df1))  
21 return(p)  
22 }  
23  
24 setwd("cheung/")  
25 # Load data and create group variable  
26 dat <- read.table("full.data")  
27  
28 jpt.names <- scan("JPT.cname.txt",what="character")  
29 chb.names <- scan("CHB.cname.txt",what="character")  
30 ceu.names <- scan("CEU_parents.txt",what="character")  
31 nceu <- length(ceu.names)  
32 njpt <- length(jpt.names)  
33 nchb <- length(chb.names)  
34 }
```

- Clearly commented
  - Small comments liberally - what, when, why, how
  - Bigger commented blocks for whole sections
- Include processing details

# 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

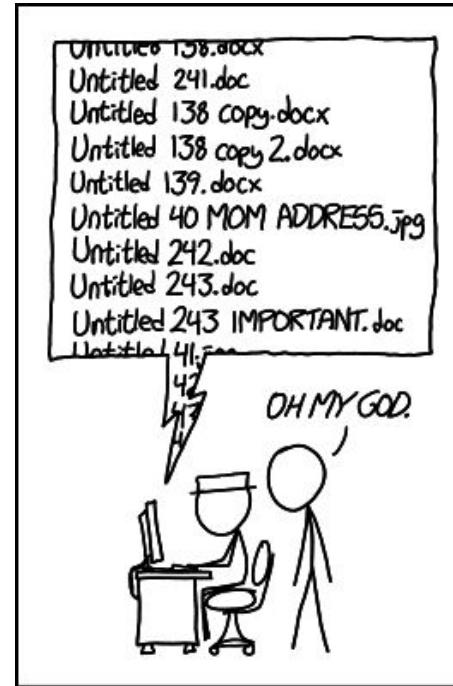
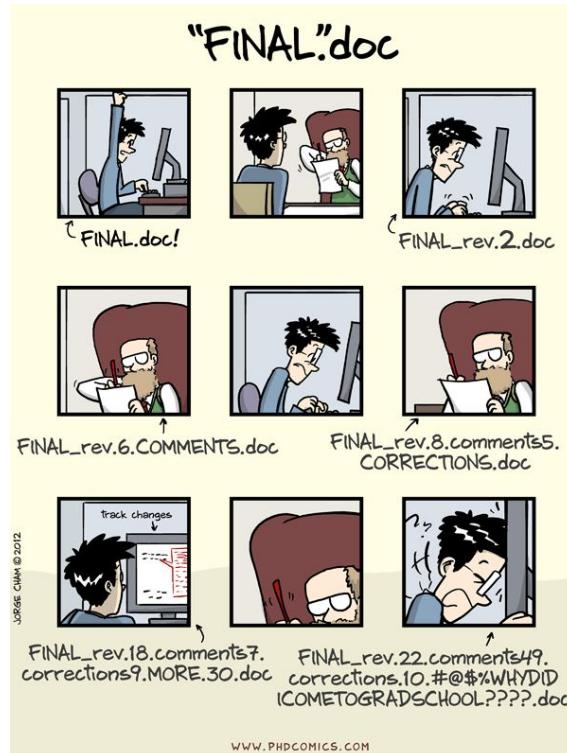
Data on crime is obtained from International Crime Data collected between 2015-2018 and is publicly available. Data on happiness is collected from the Survey of International Happiness.

Contributors:

- Jane Everyday Doe, [jane.everyday.doe@gmail.com](mailto:jane.everyday.doe@gmail.com)
- John Everyday Doe, [john.everyday.doe@gmail.com](mailto:john.everyday.doe@gmail.com)

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

# Just no



PROTIP: NEVER LOOK IN SOMEONE ELSE'S DOCUMENTS FOLDER.

<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        |
| executivereportpepsi1.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](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](http://kbroman.org/Tools4RR/assets/lectures/06_org_eda.pdf))



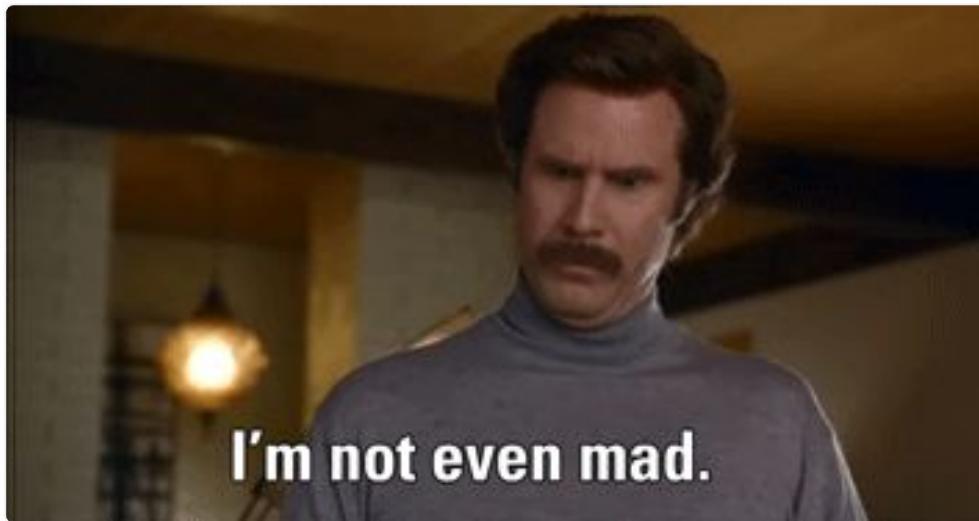
Dave Hemprich-Bennett 🦇

@hammerheadbat

Follow



\*squints at the files I was sent\*  
#otherpeoplesdata



6:01 AM - 11 Nov 2017

5 Likes



1



5



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.

RStudio - Home

www.rstudio.com

R Studio

Home RStudio IDE Shiny Training Projects About Blog

# 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.

[Download now](#) [Learn more](#)

**R training and education**

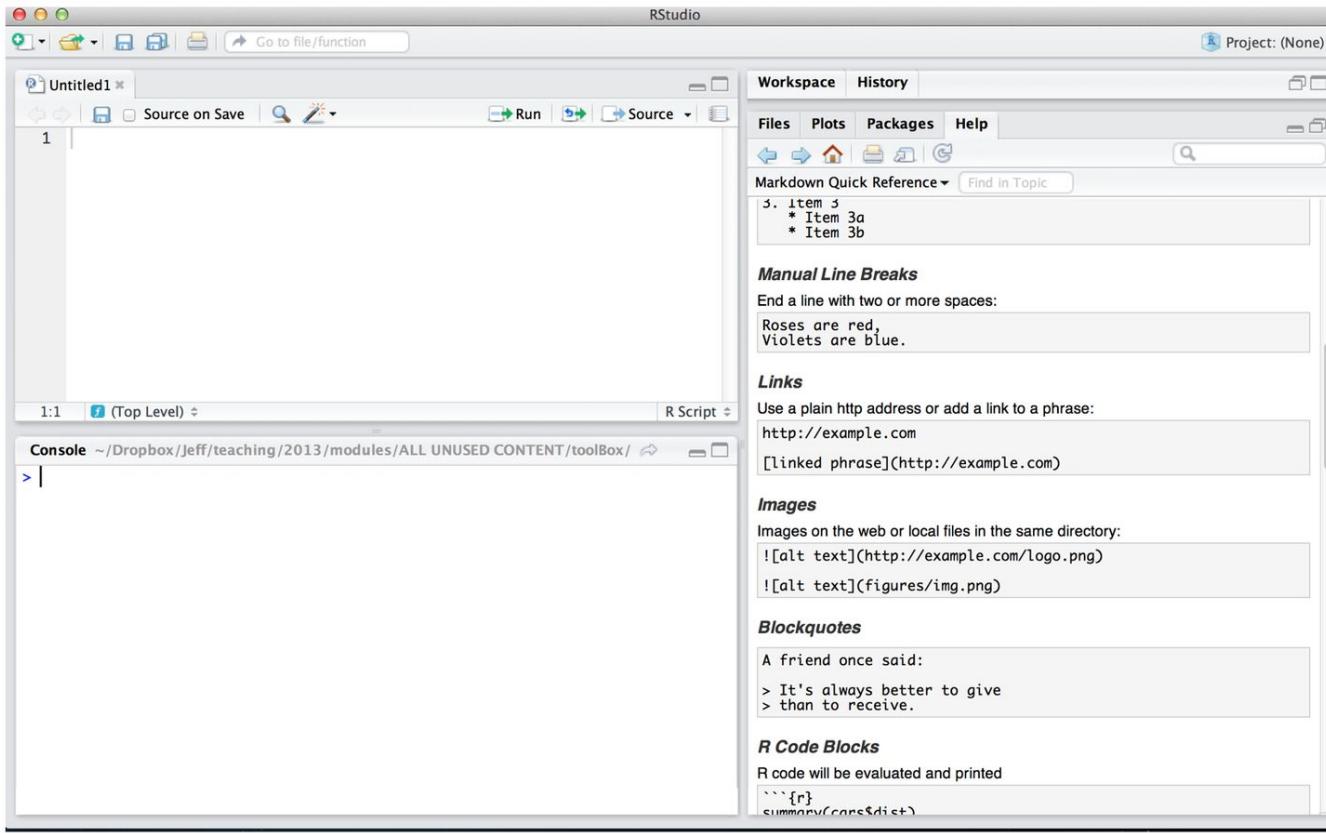
We've got hands-on courses for beginners and even R experts. Customize an on-site training or enroll in one of our public workshops.

[Request on-site](#) [View courses](#)

**Open source R packages**

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

[See projects](#)



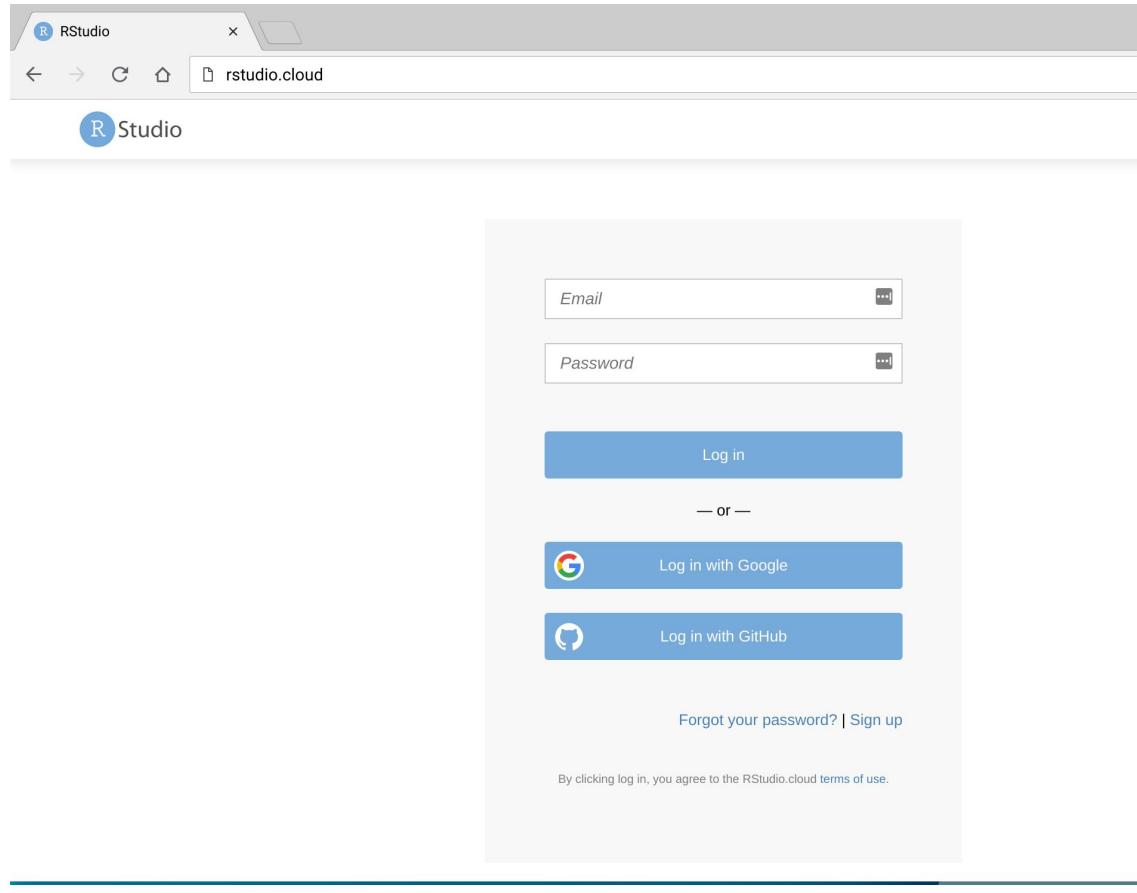
# Some useful commands

Cmd + Enter

Ctrl + Enter

Ctrl + 1

Ctrl + 2



<https://rstudio.cloud>

# R packages



[CRAN](#)  
[Mirrors](#)  
[What's new?](#)  
[Task Views](#)  
[Search](#)

[About R](#)  
[R Homepage](#)  
[The R Journal](#)

[Software](#)  
[R Sources](#)  
[R Binaries](#)  
[Packages](#)  
[Other](#)

[Documentation](#)  
[Manuals](#)  
[FAQs](#)  
[Contributed](#)

[A3](#)  
[abbyyR](#)  
[abc](#)  
[ABCAnalysis](#)  
[abc.data](#)  
[abcdeFBA](#)  
[ABCOptim](#)  
[abctools](#)  
[abd](#)  
[abf2](#)  
[abind](#)  
[abn](#)  
[abundant](#)  
[acc](#)  
[accelerometry](#)  
[AcceptanceSampling](#)  
[ACCLMA](#)  
[accrual](#)  
[accrued](#)  
[ACD](#)  
[acepack](#)

## Available CRAN Packages By Name

[A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [J](#) [K](#) [L](#) [M](#) [N](#) [O](#) [P](#) [Q](#) [R](#) [S](#) [T](#) [U](#) [V](#) [W](#) [X](#) [Y](#) [Z](#)

**A3:** Accurate, Adaptable, and Accessible Error Metrics for Predictive Models  
**Access to Abbyy** Optical Character Recognition (OCR) API  
**Tools for Approximate Bayesian Computation (ABC)**  
**Computed ABC Analysis**  
**Data Only: Tools for Approximate Bayesian Computation (ABC)**  
**ABCDE\_FBA:** A-Biologist-Can-Do-Everything of Flux Balance Analysis with this package  
**Implementation of Artificial Bee Colony (ABC) Optimization**  
**Tools for ABC Analyses**  
**The Analysis of Biological Data**  
**Load Gap-Free Axon ABF2 Files**  
**Combine Multidimensional Arrays**  
**Data Modelling with Additive Bayesian Networks**  
**Abundant regression and high-dimensional principal fitted components**  
**A Package to Processes Accelerometer Data**  
**Functions for Processing Minute-to-Minute Accelerometer Data**  
**Creation and evaluation of Acceptance Sampling Plans**  
**ACC & LMA Graph Plotting**  
**Bayesian Accrual Prediction**  
**Data Quality Visualization Tools for Partially Accruing Data**  
**Categorical data analysis with complete or missing responses**  
**ace() and avas() for selecting regression transformations**

```
install.packages ("devtools")  
install.packages ("dplyr")
```

# All Packages

**Bioconductor version 3.1 (Release)**

Autocomplete biocViews search:

## Software (1024)

- ▶ AssayDomain (345)
- ▶ BiologicalQuestion (313)
- ▶ Infrastructure (211)
- ▶ ResearchField (225)
- ▶ StatisticalMethod (293)
- ▶ Technology (645)
- ▶ WorkflowStep (525)
- ▶ AnnotationData (883)
- ▶ ExperimentData (241)

**Packages found under Software:**Show [All](#) entries

Search table:

| Package                     | Maintainer                              | Title                                                                                                                            |
|-----------------------------|-----------------------------------------|----------------------------------------------------------------------------------------------------------------------------------|
| <a href="#">a4</a>          | Tobias Verbeke,<br>Willem<br>Ligtenberg | Automated Affymetrix Array Analysis Umbrella<br>Package                                                                          |
| <a href="#">a4Base</a>      | Tobias Verbeke,<br>Willem<br>Ligtenberg | Automated Affymetrix Array Analysis Base<br>Package                                                                              |
| <a href="#">a4Classif</a>   | Tobias Verbeke,<br>Willem<br>Ligtenberg | Automated Affymetrix Array Analysis<br>Classification Package                                                                    |
| <a href="#">a4Core</a>      | Tobias Verbeke,<br>Willem<br>Ligtenberg | Automated Affymetrix Array Analysis Core<br>Package                                                                              |
| <a href="#">a4Preproc</a>   | Tobias Verbeke,<br>Willem<br>Ligtenberg | Automated Affymetrix Array Analysis<br>Preprocessing Package                                                                     |
| <a href="#">a4Reporting</a> | Tobias Verbeke,<br>Willem<br>Ligtenberg | Automated Affymetrix Array Analysis Reporting<br>Package                                                                         |
| <a href="#">ABarray</a>     | Yongming<br>Andrew Sun                  | Microarray QA and statistical data analysis for<br>Applied Biosystems Genome Survey Microarray<br>(AB1700) gene expression data. |
| <a href="#">ABSSeq</a>      | Wentao Yang                             | ABSSeq: a new RNA-Seq analysis method based<br>on absolute expression differences and<br>generalized Poisson model               |
| <a href="#">aCGH</a>        | Peter Dimitrov                          | Classes and functions for Array Comparative<br>Genomic Hybridization data.                                                       |

## sva

available all platforms    downloads top 5%    posts 6 / 2 / 3 / 2  
in BioC 3.53 years    build ok    commits 1.17

### Surrogate Variable Analysis

Bioconductor version: Release (3.1)

The sva package contains functions for removing batch effects and other unwanted variation in high-throughput experiment. Specifically, the sva package contains functions for identifying and building surrogate variables for high-dimensional data sets. Surrogate variables are covariates constructed directly from high-dimensional data (like gene expression/RNA sequencing/methylation/brain imaging data) that can be used in subsequent analyses to adjust for unknown, unmodeled, or latent sources of noise. The sva package can be used to remove artifacts in three ways: (1) identifying and estimating surrogate variables for unknown sources of variation in high-throughput experiments (Leek and Storey 2007 PLoS Genetics, 2008 PNAS), (2) directly removing known batch effects using ComBat (Johnson et al. 2007 Biostatistics) and (3) removing batch effects with known control probes (Leek 2014 biorXiv). Removing batch effects and using surrogate variables in differential expression analysis have been shown to reduce dependence, stabilize error rate estimates, and improve reproducibility, see (Leek and Storey 2007 PLoS Genetics, 2008 PNAS or Leek et al. 2011 Nat. Reviews Genetics).

Author: Jeffrey T. Leek <jtleek at gmail.com>, W. Evan Johnson <wej at bu.edu>, Hilary S. Parker <hiparker at jhsph.edu>, Elana J. Fertig <ejfertig at jhmi.edu>, Andrew E. Jaffe <ajaffe at jhsph.edu>, John D. Storey <jstorey at princeton.edu>

Maintainer: Jeffrey T. Leek <jtleek at gmail.com>, John D. Storey <jstorey at princeton.edu>, W. Evan Johnson <wej at bu.edu>

## Downloads

Bioconductor workflows

### Arrays

- [High-throughput Sequencing](#)
- [Counting Reads for Differential Expression](#) (parathyroideSE vignette)
- [Annotation](#)
- [Annotating Variants](#)
- [Annotating Ranges](#)
- [Flow Cytometry and other assays](#)
- [Candidate Binding Sites for Known Transcription Factors](#)
- [Cloud-enabled cis-eQTL search and annotation](#)
- [RNA-Seq workflow: gene-level exploratory analysis and differential expression](#)
- [Changing genomic coordinate systems with rtracklayer::liftOver](#)
- [Mass spectrometry and proteomics data analysis](#)

### Mailing Lists

Post questions about Bioconductor packages to our mailing lists. Read the [posting guide](#) before posting!

- [bioconductor](#)
- [bioC-devel](#)

[Home](#) » [Bioconductor 3.1](#) » [Software Packages](#) » [sva](#)

## sva

available all platforms    downloads top 5%    posts 6 / 2 / 3 / 2  
in BioC 3.53 years    build ok    commits 1.17

### Surrogate Variable Analysis

Bioconductor version: Release (3.1)

The sva package contains functions for removing batch effects and other unwanted variation in high-throughput experiment. Specifically, the sva package contains functions for the identifying and building surrogate variables for high-dimensional data sets. Surrogate variables are covariates constructed directly from high-dimensional data (like gene expression/RNA sequencing/methylation/brain imaging data) that can be used in subsequent analyses to adjust for unknown, unmodeled, or latent sources of noise. The sva package can be used to remove artifacts in three ways: (1) identifying and estimating surrogate variables for unknown sources of variation in high-throughput experiments (Leek and Storey 2007 PLoS Genetics, 2008 PNAS), (2) directly removing known batch effects using ComBat (Johnson et al. 2007 Biostatistics) and (3) removing batch effects with known control probes (Leek 2014 biorXiv). Removing batch effects and using surrogate variables in differential expression analysis have been shown to reduce dependence, stabilize error rate estimates, and improve reproducibility, see (Leek and Storey 2007 PLoS Genetics, 2008 PNAS or Leek et al. 2011 Nat. Reviews Genetics).

Author: Jeffrey T. Leek <[jtleek@gmail.com](mailto:jtleek@gmail.com)>, W. Evan Johnson <[wej@bu.edu](mailto:wej@bu.edu)>, Hillary S. Parker <[hiparker@jhsp.h.edu](mailto:hiparker@jhsp.h.edu)>, Elana J. Fertig <[ejfertig@jhmi.edu](mailto:ejfertig@jhmi.edu)>, Andrew E. Jaffe <[ajaffe@jhsp.h.edu](mailto:ajaffe@jhsp.h.edu)>, John D. Storey <[jstorey@princeton.edu](mailto:jstorey@princeton.edu)>

Maintainer: Jeffrey T. Leek <[jtleek@gmail.com](mailto:jtleek@gmail.com)>, John D. Storey <[jstorey@princeton.edu](mailto:jstorey@princeton.edu)>, W. Evan Johnson <[wej@bu.edu](mailto:wej@bu.edu)>

## Responsiveness

- [High-throughput Sequencing](#)
- [Counting Reads for Differential Expression](#) (parathyroideSE vignette)
- [Annotation](#)
- [Annotating Variants](#)
- [Annotating Ranges](#)
- [Flow Cytometry and other assays](#)
- [Candidate Binding Sites for Known Transcription Factors](#)
- [Cloud-enabled cis-eQTL search and annotation](#)
- [RNA-Seq workflow: gene-level exploratory analysis and differential expression](#)
- [Changing genomic coordinate systems with rtracklayer::liftOver](#)
- [Mass spectrometry and proteomics data analysis](#)

### Mailing Lists

Post questions about Bioconductor packages to our mailing lists. Read the [posting guide](#) before posting!

- [bioconductor](#)
- [bioC-devel](#)

[Home](#) » [Bioconductor 3.1](#) » [Software Packages](#) » [sva](#)

## sva

available all platforms    downloads top  
in BioC 3.53 years    build ok    issues 6 / 2 / 3 / 2  
commits 1.17

### Surrogate Variable Analysis

Bioconductor version: Release (3.1)

The sva package contains functions for removing batch effects and other unwanted variation in high-throughput experiment. Specifically, the sva package contains functions for identifying and building surrogate variables for high-dimensional data sets. Surrogate variables are covariates constructed directly from high-dimensional data (like gene expression/RNA sequencing/methylation/brain imaging data) that can be used in subsequent analyses to adjust for unknown, unmodeled, or latent sources of noise. The sva package can be used to remove artifacts in three ways: (1) identifying and estimating surrogate variables for unknown sources of variation in high-throughput experiments (Leek and Storey 2007 PLoS Genetics, 2008 PNAS), (2) directly removing known batch effects using ComBat (Johnson et al. 2007 Biostatistics) and (3) removing batch effects with known control probes (Leek 2014 biorXiv). Removing batch effects and using surrogate variables in differential expression analysis have been shown to reduce dependence, stabilize error rate estimates, and improve reproducibility, see (Leek and Storey 2007 PLoS Genetics, 2008 PNAS or Leek et al. 2011 Nat. Reviews Genetics).

Author: Jeffrey T. Leek <jtleek at gmail.com>, W. Evan Johnson <wej at bu.edu>, Hilary S. Parker <hiparker at jhsph.edu>, Elana J. Fertig <ejfertig at jhmi.edu>, Andrew E. Jaffe <ajaffe at jhsph.edu>, John D. Storey <jstorey at princeton.edu>

Maintainer: Jeffrey T. Leek <jtleek at gmail.com>, John D. Storey <jstorey at princeton.edu>, W. Evan Johnson <wej at bu.edu>

# Still runs

### kflows »

non Bioconductor workflows  
de:

#### monucleotide Arrays

- [High-throughput Sequencing](#)
- [Counting Reads for Differential Expression](#) (parathyroideSE vignette)
- [Annotation](#)
- [Annotating Variants](#)
- [Annotating Ranges](#)
- [Flow Cytometry and other assays](#)
- [Candidate Binding Sites for Known Transcription Factors](#)
- [Cloud-enabled cis-eQTL search and annotation](#)
- [RNA-Seq workflow: gene-level exploratory analysis and differential expression](#)
- [Changing genomic coordinate systems with rtracklayer::liftOver](#)
- [Mass spectrometry and proteomics data analysis](#)

### Mailing Lists »

Post questions about Bioconductor packages to our mailing lists. Read the [posting guide](#) before posting!

- [bioconductor](#)
- [bioC-devel](#)

After version 3.6 of R

```
install.packages("BiocManager")  
BiocManager::install(c("sva"))
```

Before version 3.6 of R

```
source("http://bioconductor.org/biocLite.R")  
biocLite("sva")
```



dgrtwo / broom

Watch

16



Convert statistical analysis objects from R into tidy format

146 commits

1 branch

8 releases

10 contributors



branch: master ▾

broom / +



Merge pull request #51 from zeehio/master ...



dgrtwo authored 3 hours ago

latest commit ec5c0bd980



Merge pull request #51 from zeehio/master

3 hours ago



Overhaul of how augmenting works across many objects. In particular t...

7 months ago



Add a `tidy` method for x,y,z lists

21 days ago



Changed `rowwise\_df\_tidiers` to allow the original data to be saved a...

a month ago



Added `gam` to README. Removed rownames from glmnet output. Few typo ...

7 months ago



Update cran comments.

6 months ago



Update cran comments.

6 months ago



Merge pull request #51 from zeehio/master

3 hours ago



jtleek / sva-devel

 Unwatch 6 Star 4 Fork 7

## Description

Short description of this repository

26 commits

1 branch

0 releases

4 contributors

# Other people like it

Code

 Issues 0 Pull requests 0

Wiki

Pulse

Graphs

Settings

HTTPS clone URL

<https://github.com/> You can clone with [HTTPS](#), [SSH](#), or [Subversion](#). 

Clone in Desktop

Commit made by the Bioconductor Git-SVN bridge.

 bioc-sync authored 27 days ago latest commit [4e9c7a2731](#)

R Made the following changes: 1) added unit tests for ComBat to check C... 2 months ago

man Made several modifications to ComBat to streamline the design matrix ... 5 months ago

src Commit made by the Bioconductor Git-SVN bridge. 7 months ago

tests Made the following changes: 1) added unit tests for ComBat to check C... 2 months ago

vignettes Made several modifications to ComBat to streamline the design matrix ... 5 months ago

.gitignore Initial commit 11 months ago

DESCRIPTION Commit made by the Bioconductor Git-SVN bridge. 27 days ago

NAMESPACE fixed documentation of sva.check 8 months ago



jtleek / sva-devel

 Unwatch 6 Star 4 Fork 7

## Description

Short description of this repository

26 commits

People have been  
working on it



branch: master

sva-devel / +

Cancel

Commit made by the Bioconductor Git-SVN bridge.

|  |                                                                          |                          |
|--|--------------------------------------------------------------------------|--------------------------|
|  | bioc-sync authored 27 days ago                                           | latest commit 4e9c7a2731 |
|  | Made the following changes: 1) added unit tests for ComBat to check C... | 2 months ago             |
|  | Made several modifications to ComBat to streamline the design matrix ... | 5 months ago             |
|  | Commit made by the Bioconductor Git-SVN bridge.                          | 7 months ago             |
|  | Made the following changes: 1) added unit tests for ComBat to check C... | 2 months ago             |
|  | Made several modifications to ComBat to streamline the design matrix ... | 5 months ago             |
|  | Initial commit                                                           | 11 months ago            |
|  | Commit made by the Bioconductor Git-SVN bridge.                          | 27 days ago              |
|  | fixed documentation of sva.check                                         | 8 months ago             |

Code

 Issues 0 Pull requests 0

Wiki

Pulse

Graphs

Settings

HTTPS clone URL

<https://github.com/> You can clone with [HTTPS](#), [SSH](#), or [Subversion](#). 

Clone in Desktop

```
install.packages("remotes")
library(remotes)
install_github("tidymodels/broom")
```

# Average trustworthiness



>



>



**github**  
SOCIAL CODING

# `rstudio.cloud` tour

<https://rstudio.cloud/spaces/77416/projects>