

Data Wrangling in R

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

Course Info

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

What we will cover in this course:

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

The Tidvyerse

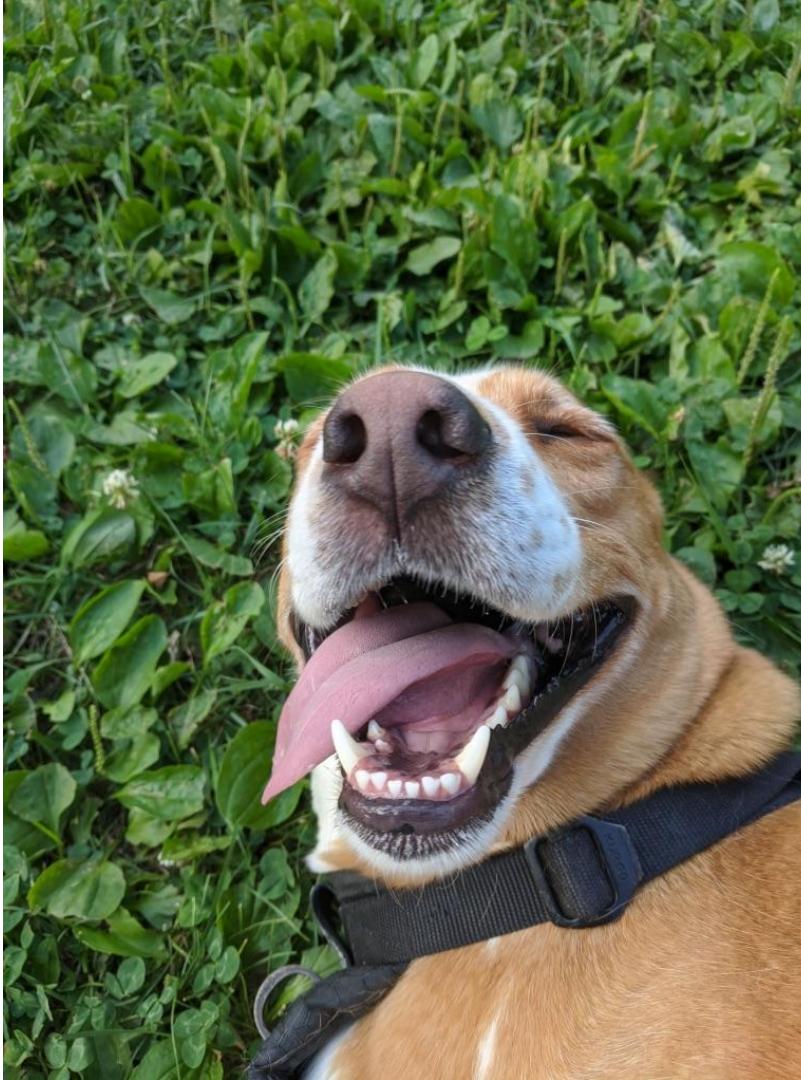
Newer packages designed for data science that make R code more intuitive.



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 \(DaSL\)](#).

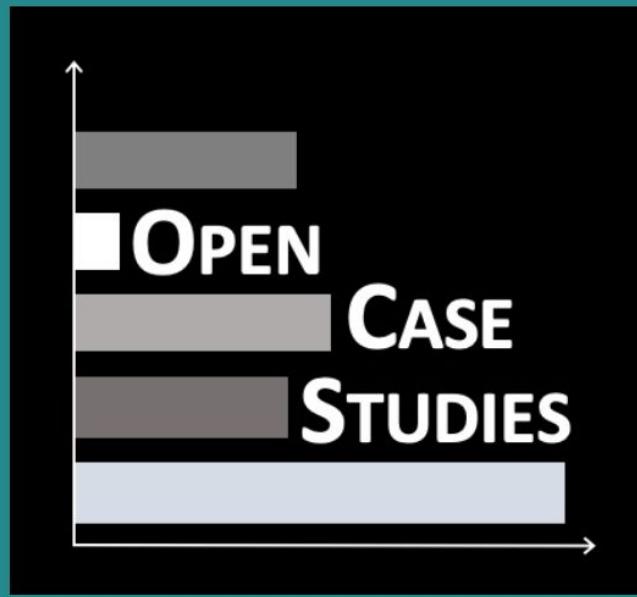
Prior to joining the JHSPH, Dr. Wright was a Postdoctoral Fellow at the [Lieber Institute for Brain Development \(LIBD\)](#), where her research focused on uncovering genetic mechanisms in psychiatric disease (with a particular emphasis on non-coding RNA) through the utilization of data science tools. 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](#).



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.





ABOUT

RESEARCH

RESOURCES

FUN STUFF

RESUME / CV

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:
```

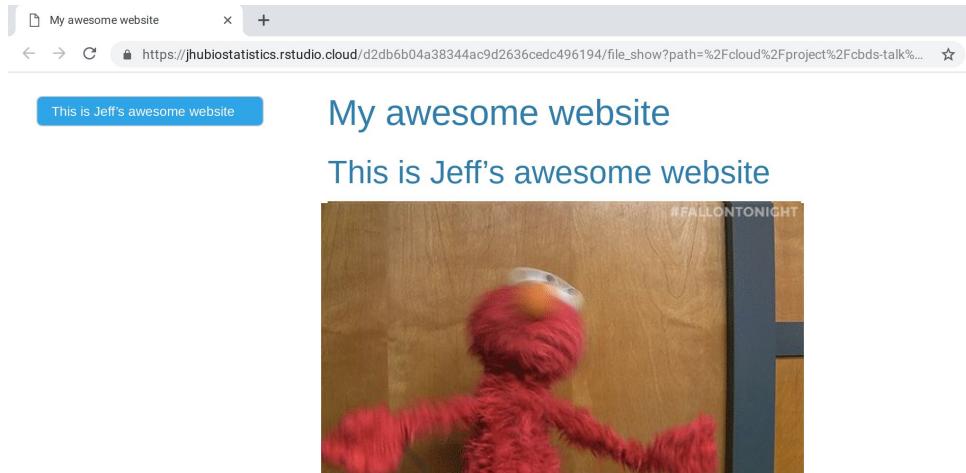
```
    toc: true
```

```
    toc_float: true
```

```
    theme: cerulean
```

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

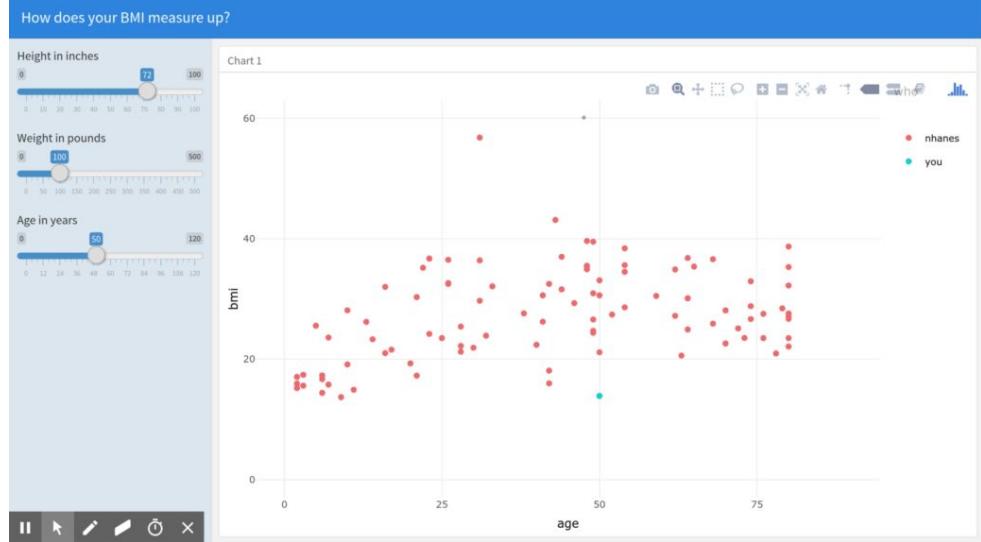
```
![] (https://media.giphy.com/media/d  
rXGoW1iudhKw/giphy.gif)
```



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()
)
})
```
```



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"
```

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 Chan⁴, Robyn Sayer⁴, Janel Cragun⁴, Hope Cottrill⁴, Michael J Kelley², Rebecca Petersen⁵, David Harpole⁵, Jeffrey Marks⁵, Andrew Berchuck^{1,6}, Geoffrey S Ginsburg^{1,2}, Phillip Febbo^{1,2,3}, Johnathan Lancaster⁴ & 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

Please watch!!!

Forensic Bioinformatics

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

Keith A. Baggerly

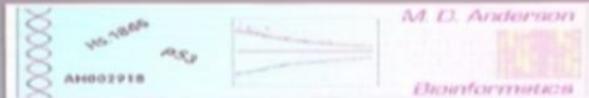
Bioinformatics and Computational Biology

UT M. D. Anderson Cancer Center

kabagg@mdanderson.org



BIRS Workshop, Aug 14, 2013



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

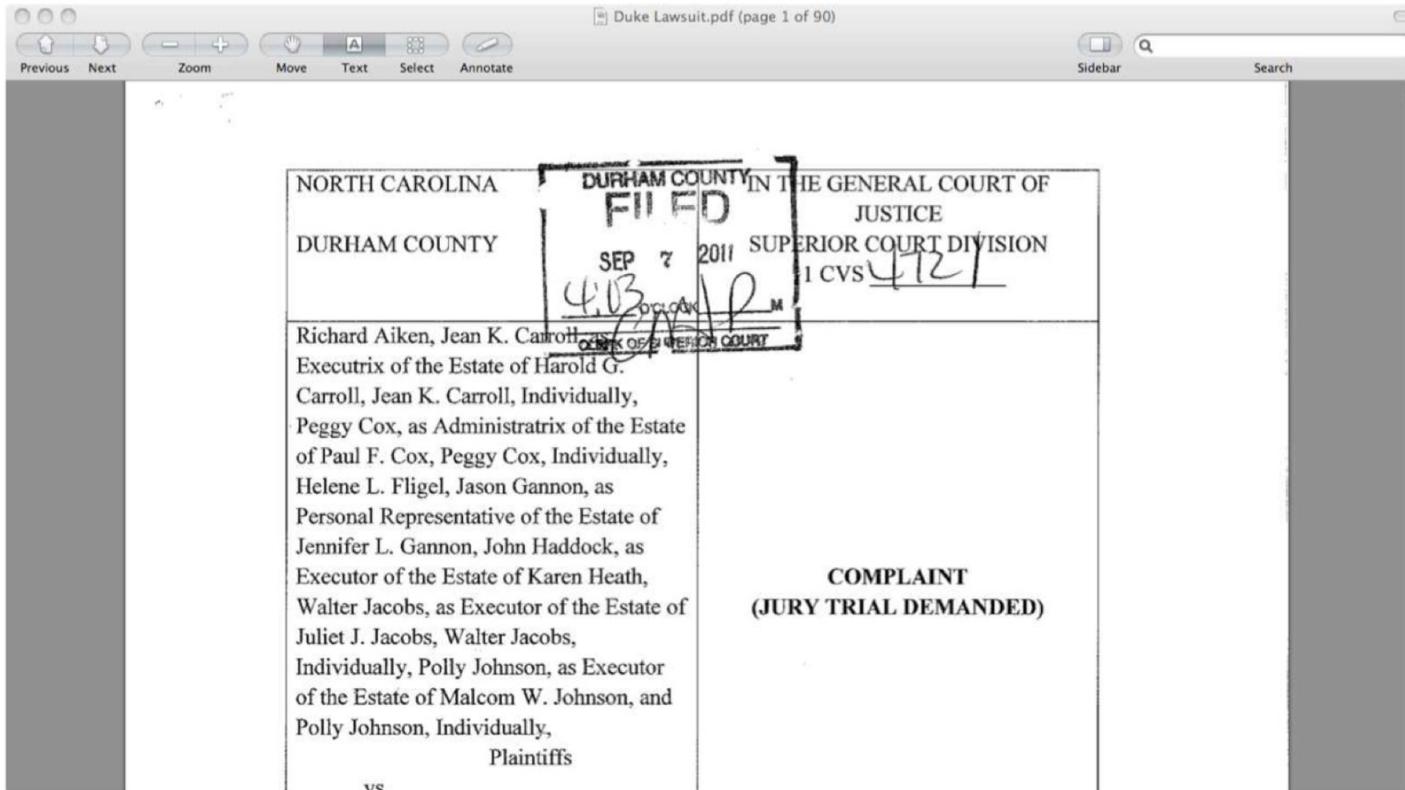
Nature Medicine **12**, 1294–1300 (2006) | [Cite this article](#)

5843 Accesses | **426** Citations | **95** Altmetric | [Metrics](#)

 **i** A [Retraction](#) to this article was published on 07 January 2011

i A [Corrigendum](#) to this article was published on 01 August 2008

i A [Corrigendum](#) to this article was published on 01 November 2007



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)

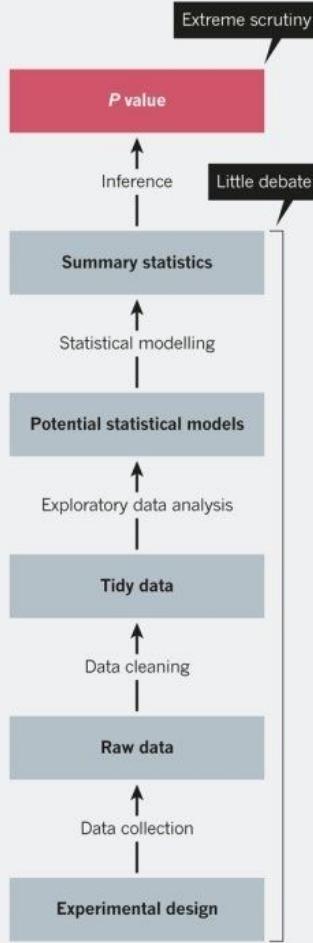
| | Estimate | Std. Error | z value | Pr(> z) |
|---------------------------------------|----------|------------|---------|----------|
| (Intercept) | 0.5675 | 0.1475 | 3.847 | 0.000122 |
| dat\$study_type <non-anon></non-anon> | 0.4250 | 0.2182 | 1.948 | 0.051458 |

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 right

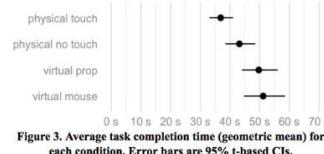
The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time*

Andrew Gelman[†] and Eric Loken[‡]

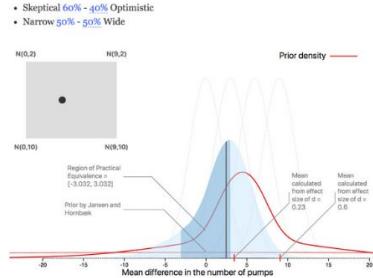
14 Nov 2013

“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



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 | pcp-neg | scatterplot-pos | pcp-neg | 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 | 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 |
| 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 |

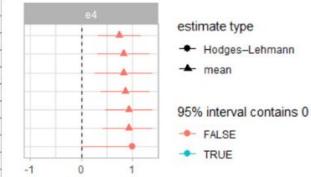


Figure 4. Perceptually-driven ranking of visualizations depending on the correlation sign (-neg / -pos), as a function of correlation value (r) and overall (right column).

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.

And so we data wrangle



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>

What it actually looks like

<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>. The page has a blue header with links for Developers, Search, API Health, Blog, Discussions, Documentation, and Sign in. Below the header, there is a note about cursor values and an example value of 12893764510938. The main content area is titled "Example Request" and shows a GET request to [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). A code block displays the JSON response structure:

```
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.        "default_profile": true,
21.        "url": null,
22.        "favourites_count": 0,
23.        "utc_offset": null,
24.        "id": 509466276,
25.        "profile_image_url_https":
26.          "https://si0.twimg.com/sticky/default_profile_images/default_profile_4_normal.png",
27.          "listed_count": 0,
28.          "profile_use_background_image": true,
29.          "profile_text_color": "333333",
30.          "lang": "en",
31.          "protected": false,
32.          "followers_count": 0,
33.          "geo_enabled": false,
34.          "description": null,
```

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
Date Entered: 09 Mar 2011
Reaction: URINARY RETENTION
Allergy Type: DRUG
A Drug Class: NON-OPIOID ANALGESICS
Observed/Historical: HISTORICAL
Comments: gradually worsening difficulty emptying bladder

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

Medication: IBUPROFEN 600MG TAB
Instructions: TAKE ONE TABLET BY MOUTH FOUR TI Status: Active
Refills Remaining: 3
Last Filled On: 28 Aug 2010
Initially Ordered On: 01 Jul 2010



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)

| 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 | | | | | | | | | | | | | | | | | | | | |

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

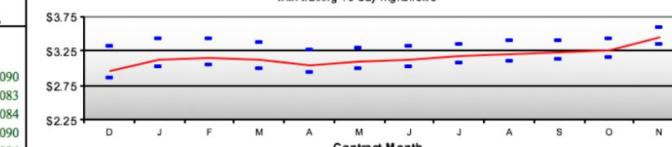
| | | |
|-----------------|----------------|----------|
| Mark Whitt | (303) 575-6473 | Denver |
| Paul Lucci | (303) 575-6474 | Denver |
| Tyrell Harrison | (303) 575-6478 | Denver |
| Dave Fuller | (503) 464-3732 | Portland |

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 |

Forward NYMEX Strip
with trailing 10-day highs/lows



IF NWPL Rocky Mountains

| | Fixed Price | | Basis |
|-------|-------------|---------|---------|
| BID | OFFER | BID | OFFER |
| 1.890 | 1.910 | | |
| 2.060 | 2.080 | | |
| 2.395 | 2.415 | (0.565) | (0.545) |
| 2.594 | 2.614 | (0.494) | (0.474) |
| 2.581 | 2.601 | (0.585) | (0.565) |
| 3.356 | 3.376 | (0.295) | (0.275) |
| 2.634 | 2.654 | (0.530) | (0.510) |

IF CIG Rocky Mountains

| | Fixed Price | | Basis |
|-------|-------------|---------|---------|
| BID | OFFER | BID | OFFER |
| 1.940 | 1.960 | | |
| 1.960 | 1.980 | | |
| 2.345 | 2.365 | (0.615) | (0.595) |
| 2.548 | 2.568 | (0.540) | (0.520) |
| 2.471 | 2.491 | (0.695) | (0.675) |
| 3.311 | 3.331 | (0.340) | (0.320) |
| 2.551 | 2.571 | (0.614) | (0.594) |

IF EL Paso Permian

| | Fixed Price | | Basis |
|-------|-------------|---------|---------|
| BID | OFFER | BID | OFFER |
| 2.375 | 2.395 | | |
| 2.420 | 2.440 | | |
| 2.700 | 2.720 | (0.260) | (0.240) |
| 2.855 | 2.875 | (0.233) | (0.213) |
| 3.009 | 3.029 | (0.158) | (0.138) |
| 3.499 | 3.519 | (0.153) | (0.133) |
| 2.982 | 3.002 | (0.182) | (0.162) |

IF NWPL Canadian Border (Sumas)

| | Fixed Price | | Basis |
|-------|-------------|---------|---------|
| BID | OFFER | BID | OFFER |
| 2.480 | 2.500 | | |
| 2.460 | 2.480 | | |
| 2.800 | 2.820 | (0.160) | (0.140) |
| 2.892 | 2.912 | (0.196) | (0.176) |
| 2.796 | 2.816 | (0.370) | (0.350) |
| 3.706 | 3.726 | 0.055 | 0.075 |
| 2.880 | 2.900 | (0.285) | (0.265) |

Slide from Jenny Bryan

(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

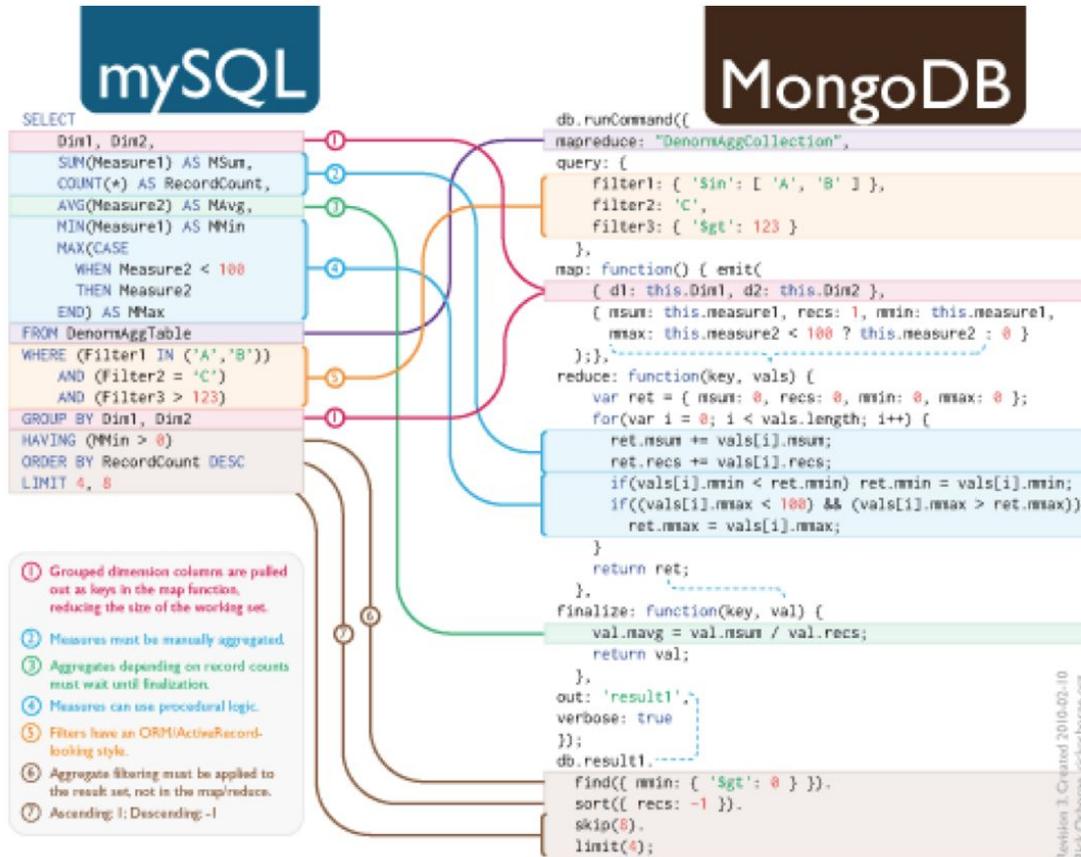


Where you wish data was



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](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](https://dev.twitter.com/docs/api/1/get(blocks/blocking). The page title is "GET blocks/blocking | Twitter". The main content area displays the API endpoint details, including the example value "12893764510938". Below this, there is a section titled "Example Request" which shows a GET request to "https://api.twitter.com/1/blocks/blocking.json?cursor=-1&include_entities=true". A large code block below the request shows a JSON response object with 32 numbered properties, such as "previous_cursor", "name", "profile_sidebar_border_color", and "id".

```
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 🇪🇸",
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,
23.        "utc_offset": null,
24.        "id": 509466276,
25.        "profile_image_url_https":
26.          "https://si0.twimg.com/sticky/default_profile_images/default_profile_4_normal.png",
27.          "listed_count": 0,
28.          "profile_use_background_image": true,
29.          "profile_text_color": "333333",
30.          "lang": "en",
31.          "protected": false,
32.          "followers_count": 0,
33.          "geo_enabled": false,
34.          "description": null,
```

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

<https://twitter.com/hadleywickham>

Abstract

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

A tidy data set

| | A | B | C | D | E | F | G | 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 data with
Linking variables across tables

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)

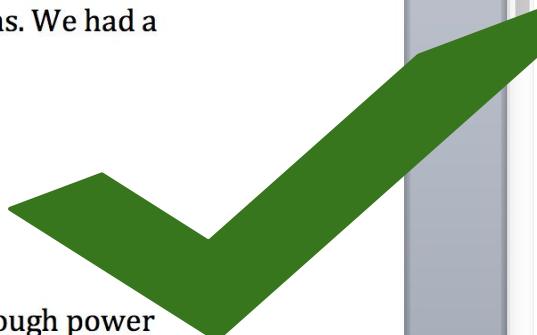
G. Diagnosis: DSM/ICD-10 diagnosis

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 A² Aa A^{BD} 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|

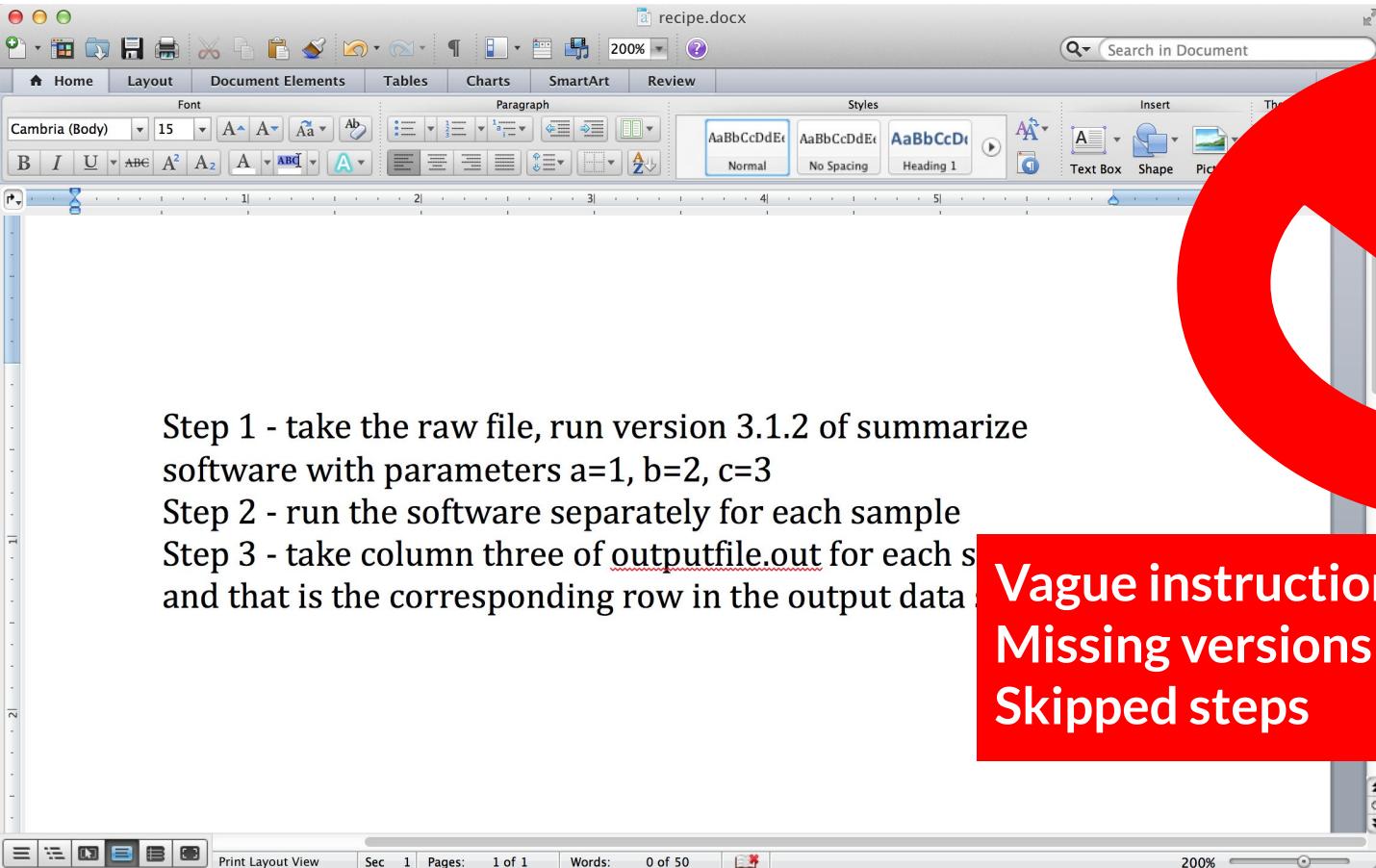
Print Layout View Sec 1 Pages: 1 of 1 Words: 0 of 50 200%

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

200%

Search in Document

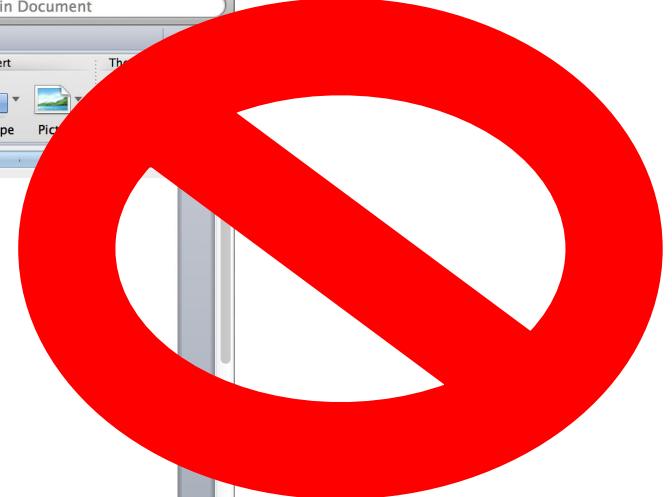
Home Layout Document Elements Tables Charts SmartArt Review

Font Paragraph Styles Insert

Cambria (Body) 15 A A Aa Ab B I U ABC A² Aa A ABD A AaBbCcDdEe AaBbCcDdEe AaBbCcDdEe Normal No Spacing Heading 1 AA 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
(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

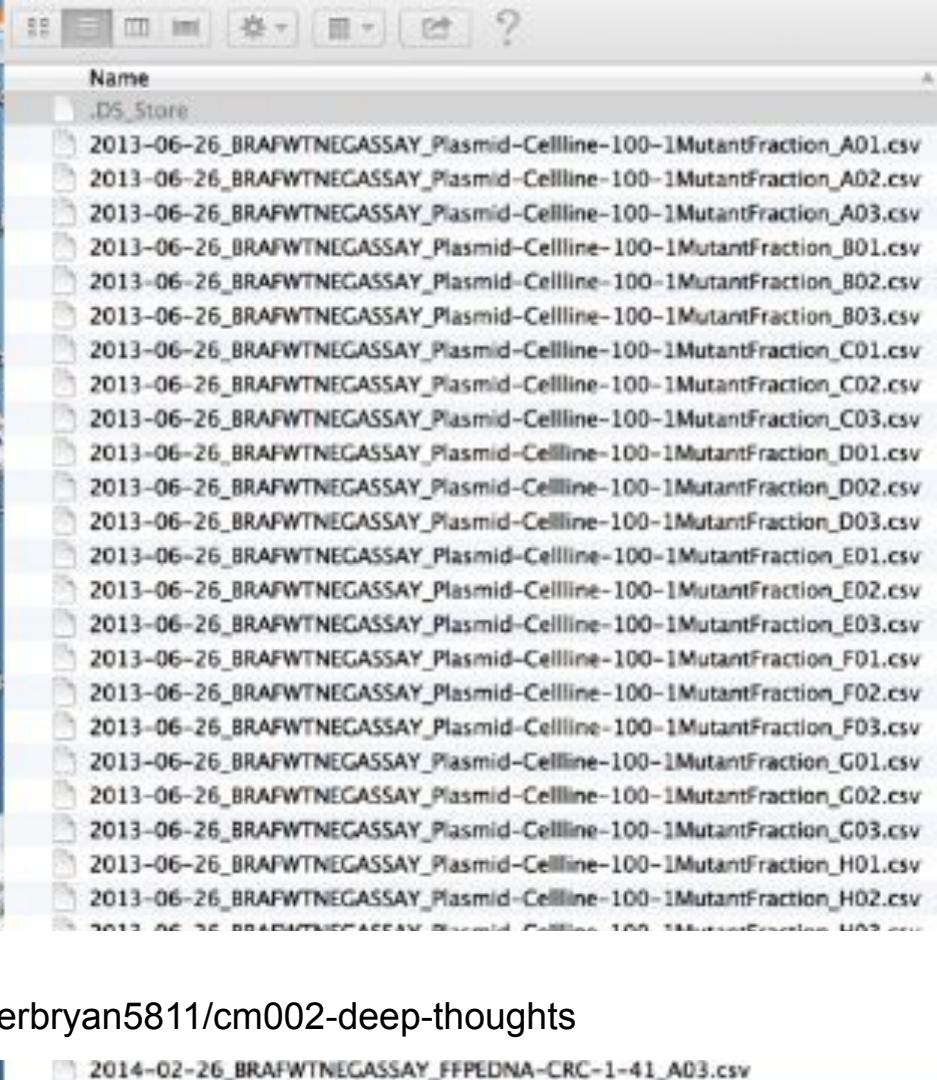
1. Be consistent
2. Choose good names for things
3. Write dates as YYYY-MM-DD
4. No empty cells
5. Put just one thing in a cell
6. Don't use font color or highlighting as data
7. Save the data as plain text files

Organize thyself

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

- Jenny Bryan





Slide via Jenny Bryan:

<http://www.slideshare.net/jenniferbryan5811/cm002-deep-thoughts>

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

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

Raw data

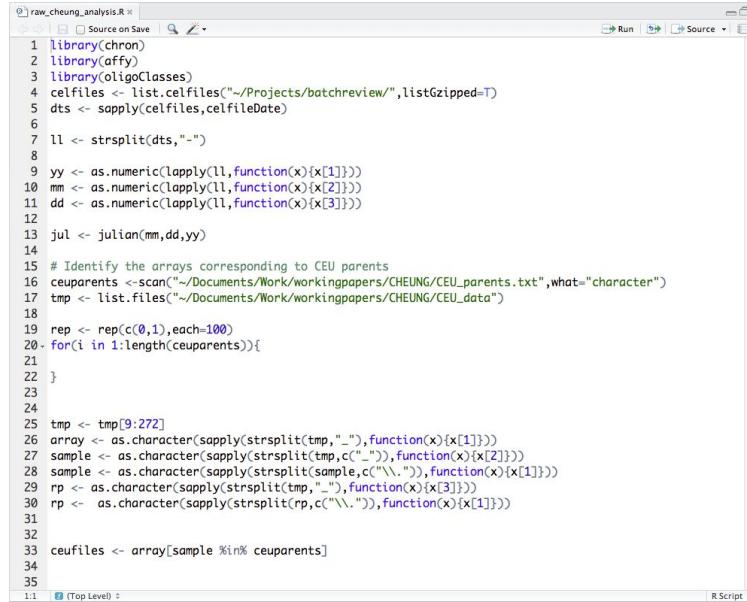
| ALLERGIES | | MEDICATION HISTORY | |
|----------------------------------|--|----------------------------------|--|
| Last Updated: 01 Dec 2011 @ 0851 | | Last Updated: 11 Apr 2011 @ 1737 | |
| Allergy Name: | TRIMETHOPRIM | Medication: | AMLODIPINE BESYLATE 10MG TAB |
| Location: | DAYT29 | Instructions: | 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 | 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 | 1307119998 | 1307120021 | 2358 | B | | | | | | | | | |
| 8 | 7 | 105 | 18 | 1307119999 | 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 | 397 | 13 | 1307120018 | 1307120119 | 2234 | 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 | 1307120104 | 1307120170 | 2215 | A | | | | | | | | | |
| 20 | 19 | 43 | 15 | 1307120122 | 1307120140 | 2245 | C | | | | | | | | | |
| 21 | 20 | 393 | 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 | 1307120310 | 2084 | D | | | | | | | | | |
| 25 | 24 | 121 | 19 | 1307120194 | 1307120394 | 2051 | 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 | 1307120311 | 1307120385 | 2050 | B | | | | | | | | | |
| 31 | 30 | 503 | 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 | 1307120367 | 1307120497 | 1958 | 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 | 270 | 16 | 1307120447 | 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



The screenshot shows an R script editor window titled "raw_cheung_analysis.R". The code is an R script for processing gene expression data. It starts by loading libraries (chron, affy, oligoClasses) and reading a list of CEL files from a specified directory. The script then splits the file names into year, month, and day components and converts them into Julian dates. It identifies arrays corresponding to CEU parents from a text file and reads data from a folder. The script then performs several steps of data manipulation, including splitting and concatenating character vectors and applying functions to specific elements of these vectors. Finally, it filters the array based on the identified CEU parents.

```
raw_cheung_analysis.R
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
```

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

Final scripts

```
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))  
14  resid0 <- dat %*% (Id - mod0 %*% solve(t(mod0) %*% mod0) %*% t(mod0))  
15  
16  rss1 <- resid^2 %*% rep(1,n)  
17  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
- Only analyses that appear in the final write-up

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:

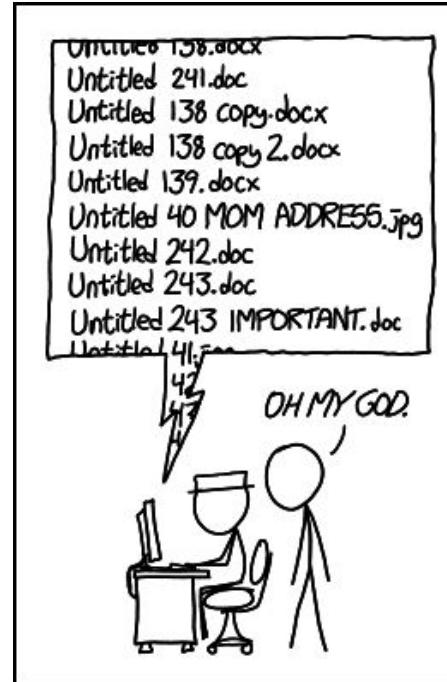
- Jane Everyday Doe, jane.everyday.doe@gmail.com
- John Everyday Doe, john.everyday.doe@gmail.com

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 |
| 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

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

- Karl Broman

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

R + RStudio



[Home]

Download

[CRAN](#)

R Project

[About R](#)

[Contributors](#)

[What's New?](#)

[Mailing Lists](#)

[Bug Tracking](#)

[Conferences](#)

[Search](#)

R Foundation

[Foundation](#)

[Board](#)

[Members](#)

[Donors](#)

[Donate](#)

The R Project for Statistical Computing

Getting Started

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

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

News

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

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

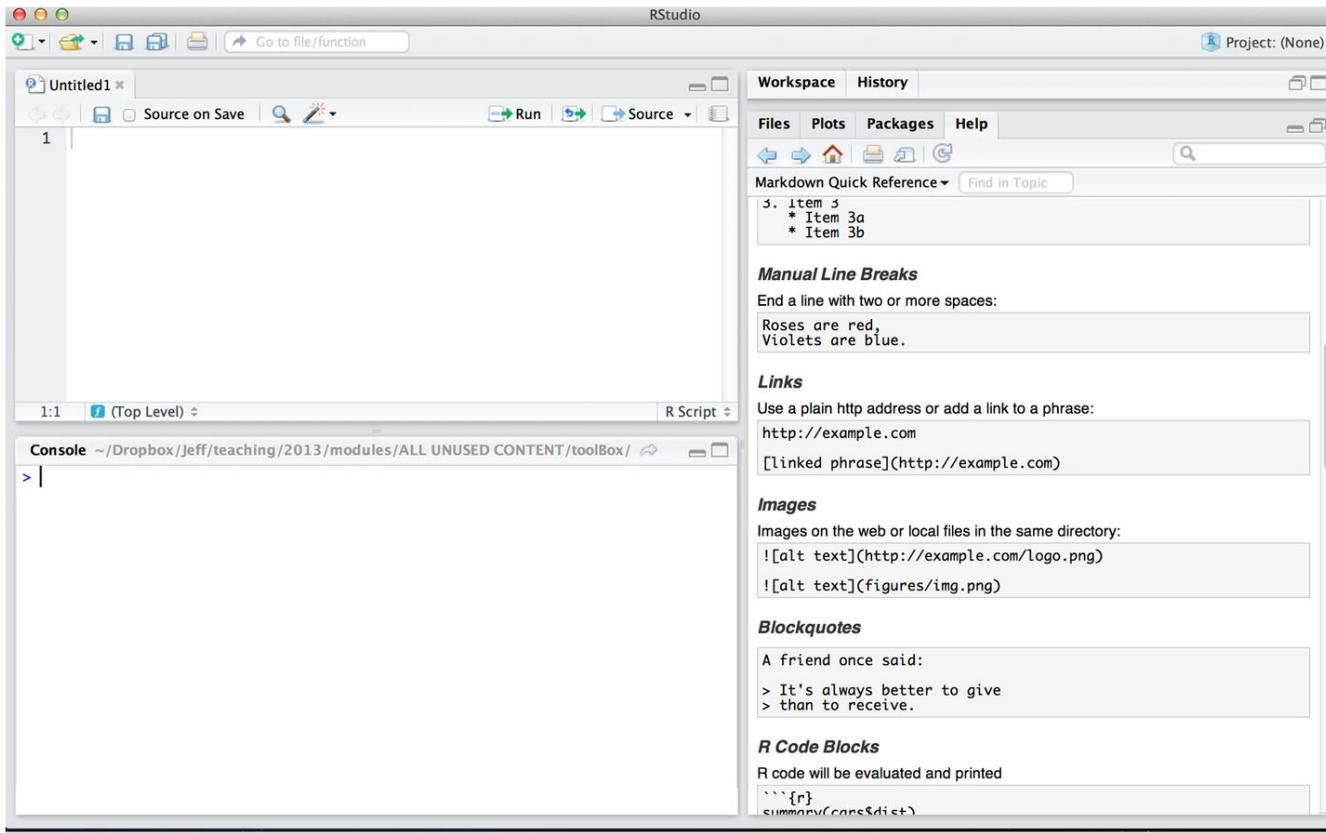
We've got hands-on courses for beginners and even R experts. Customize an on-site training or enroll in one of our public workshops.

[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](#)





POWERED
BY



BETA

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/#> silv 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”



POWERED BY WORKSPACES

Workspaces > hoffman-ava/Bioconductor-Workflow-DESeq... E.1_DESeq2Analysis.ipynb

Cloud Environment Running (\$0.20 / hr)

PREVIEW (READ-ONLY) EDIT PLAYGROUND MODE

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 AnVIL Bulk RNASeq](#). 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
```

Standard Computing

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

Cloud computing

- 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 | Consortium | Cohorts | Diseases | Cohorts | Data Types | Cohorts | Consent Code | Cohorts | Access | Cohorts |
|------------------------------------|---|---------------------|--|-------------------|--|---------|---|-------------------|--|----------------|
| e.g. disease, study name, dbGaP ID | <input type="checkbox"/> 1000 Genomes
<input type="checkbox"/> CCDG
<input type="checkbox"/> CMG
<input type="checkbox"/> Convergent Neuro
+ 4 more | 1
196
36
2 | <input type="checkbox"/> Alzheimer's disease
<input type="checkbox"/> asthma
<input type="checkbox"/> atherosclerosis
<input type="checkbox"/> atrial fibrillation
+ 17 more | 3
1
1
13 | <input type="checkbox"/> Exome
<input type="checkbox"/> RNAseq
<input type="checkbox"/> Whole Genome | 138 | <input type="checkbox"/> DS
<input type="checkbox"/> DS-ASD
<input type="checkbox"/> DS-ASD-IRB
<input type="checkbox"/> DS-ASD-IRB-COL
+ 69 more | 4
18
5
1 | <input type="checkbox"/> Consortium Access
<input type="checkbox"/> Controlled Access
<input type="checkbox"/> Open Access | 99
145
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 |
| CTEP 6.0 | 1 | 17,992 | 8,720 | 107.14 |

<https://anvilproject.org/data>

Tools

[search tools](#)[Get Data](#)[Collection Operations](#)[GENERAL TEXT TOOLS](#)[Text Manipulation](#)[Filter and Sort](#)[Join, Subtract and](#)[Datamash](#)[GENOMIC FILE MAN](#)[FASTA/FASTQ](#)[FASTQ Quality Co](#)[SAM/BAM](#)[BED](#)[VCF/BCF](#)[Nanopore](#)[Convert Formats](#)[Lift-Over](#)[COMMON GENOMICS TOOLS](#)[Operate on Genomic Intervals](#)[Fetch Sequences/Alignments](#)[GENOMICS ANALYSIS](#)

Download from web or upload from disk

Regular

Composite

Collection

Rule-based

 Drop files hereType (set all): Genome (set all): Choose local files Choose remote files Paste/Fetch data

Start

Select

Pause

Reset

Close

Galaxy UI training

Intro to Galaxy Analysis

Transcriptomics

Statistics and Machine Learning

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

The screenshot shows the AnVIL Dataset Catalog interface. At the top, there's a search bar with placeholder text "e.g. disease, study name, dbGap Id". Below the search bar are several filter categories: Consortium (1000 Genomes, CCGD, CMG, Convergent Neuro), Cohorts (1, Alzheimer's disease, 196, asthma, 36, atherosclerosis, 2, atrial fibrillation), Diseases (Alzheimer's disease, asthma, atherosclerosis, atrial fibrillation), Data Types (Exome, RNAseq, Whole Genome), Cohorts (138, DS, 1, DS-ASD, 108, DS-ASD-IRB, 1, DS-ASD-IRB-COL), Consent Code (various codes like Consortium Access, Controlled Access, Open Access), Access (Consortium Access, Controlled Access, Open Access), and Cohorts (counts for each). Below these are two tables: "Search Summary" and "Search Results". The "Search Summary" table provides a high-level overview of the dataset counts across different consortia and data types. The "Search Results" table lists specific datasets with details like dbGap ID, title, consent code, Terra workspace name, diseases, access, data type, samples, subjects, and size.

| Consortium | Cohorts | Samples | Subjects | Size (TB) |
|------------------|---------|---------|----------|-----------|
| 1000 Genomes | 1 | 3,202 | 3,202 | 72.98 |
| CCGD | 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 |

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

<https://anvilproject.org/data>



BETA
WORKSPACES

Workspaces >
anvil-outreach/Example Bioc



Cloud Environment
Running (\$0.06 / hr)



DASHBOARD

DATA

NOTEBOOKS

WORKFLOWS

JOB HISTORY

ABOUT THE WORKSPACE

No description added

WORKSPACE INFO

CREATION DATE
5/18/2021

SUBMISSIONS
0

EST. \$/MONTH
\$0.00

GOOGLE PROJECT ID
anvil-outreach

Clone

Share

Publish COMING SOON

Delete Workspace

Owner

OWNERS

jtleek@gmail.com

TAGS

Add a tag

No tags yet

AUTHORIZATION DOMAIN

AnVIL Analysis Platforms

The screenshot shows a Jupyter notebook interface with the title 'E_1_DESeq2Analysis.ipynb'. The code cell contains R code for visualizing differential expression between 'Mock vs. ABA Condition'. The resulting scatter plot shows log2 fold change on the y-axis versus -log10(p-value) on the x-axis, with points colored by condition.



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

The screenshot shows the Galaxy web interface. On the left, a 'Summary' panel lists various QC metrics for a dataset. On the right, a 'Basic Statistics' panel displays a histogram of sequence quality scores.



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

The screenshot shows the RStudio interface. The code editor contains R code for an 'AFR Allele Frequency Distribution' analysis. The plot shows the frequency distribution of alleles across different samples.



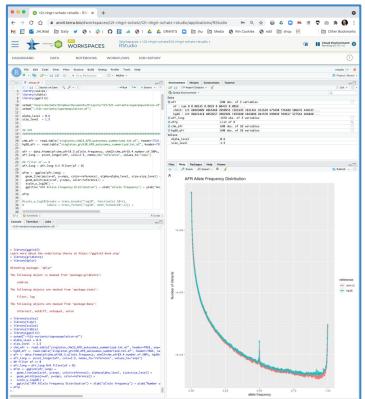
- + Feature rich IDE for programming in R
- + Rich statistics & ML and visualizations
- Limited support for other programming languages

The screenshot shows the wd1 command-line interface. It displays R code for a 'FastQC' pipeline, including commands for reading fastq files, extracting sequence names, and running FastQC. The code also includes logic for handling multiple samples and generating reports.



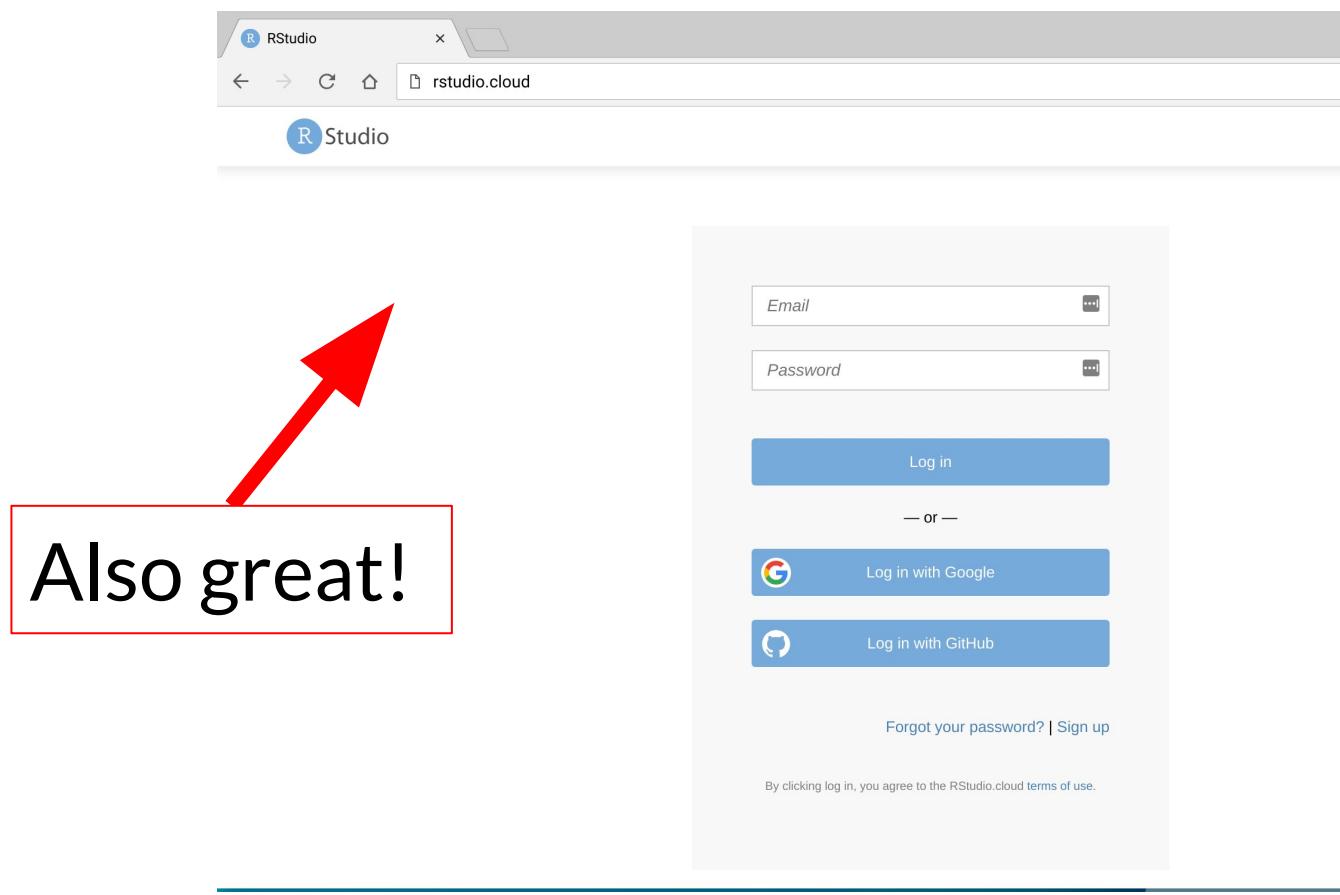
- + Extremely scalable and flexible
- Most technically demanding
- Unpredictable and potentially large costs

AnVIL Analysis Platforms



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





<https://rstudio.cloud>