Data Wrangling in R

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

Preliminaries

Course Info

Course name	Data Wrangling in R
Instructors	Andrew Jaffe and Jeff Leek
Course website	http://sisbid.github.io/Module1/
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


SURF
SAND

About us

(Jeff)

Leek group | Home | Projects | Data and code | Members | Books | Talks | Blog | Join us

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.

The Johns Hopkins Data Science Lab

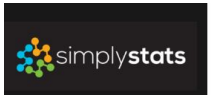
The Johns Hopkins Data Science Lab (DaSL) is a group based in the Johns Hopkins Bloomberg School of Public Health whose mission is to enhance data science thinking everywhere and make data science accessible to the entire world. Data science is a fundamental way of thinking in many areas of science, business, and government. We believe all people should be able to develop literacy, fluency and skill in data science so they can make sense of the data they encounter in their personal and professional lives. We recognize data science as a fundamentally human activity and focus our activities on helping people build data analyses for people.

Our goal is to

- Teach people how to design, collect, interpret, and interact with data
- Build a supportive environment for the people at Johns Hopkins who creatively use data to answer questions
- Provide leadership on how people doing data science should be supported at Johns Hopkins and in academia, industry, and government
- Build resources and products that help people learn and do data science
- Conduct research into the theory and practice of data science

We have previously built massive online open courses in data science that have enrolled more than 8 million people around the world, published best selling books, widely-subscribed blogs, developed podcasts on data science, statistics, and academia, and have developed a software platform for interactive learning of statistics in R. We make our impact by combining cutting edge research in machine learning, artificial intelligence and statistics with a deep understanding of applications and an eye toward the human behavioral component of data analysis.

DaSL News



2011 - 2017. All rights reserved.

Built with blogdown and Hugo.
Theme Blackburn.

Simply Statistics

A statistics blog by Rafa Irizarry, Roger Peng, and Jeff Leek

Research quality data and research quality databases

Jeff Leek 2019/05/29

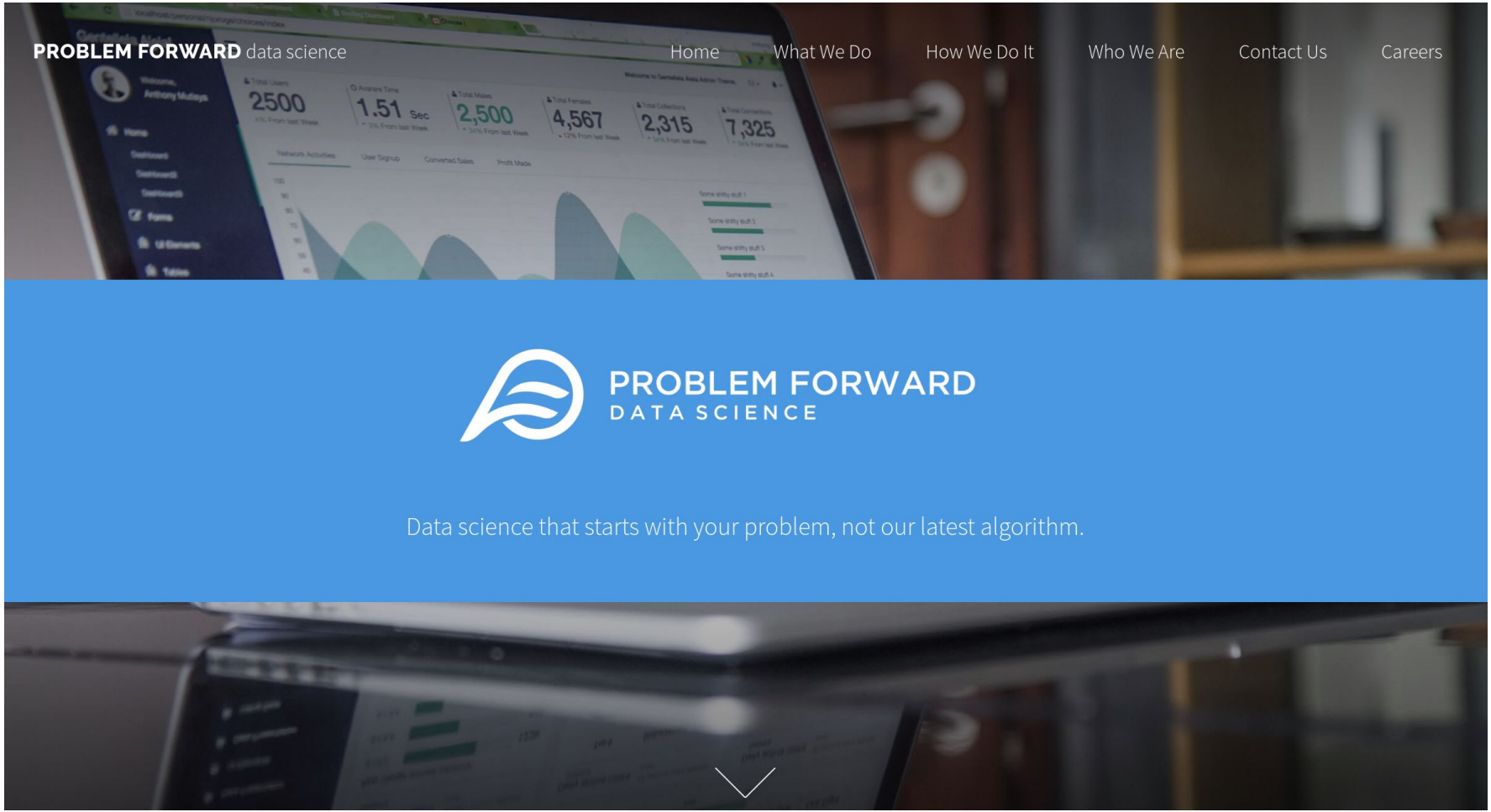
When you are doing data science, you are doing research. You want to use data to answer a question, identify a new pattern, improve a current product, or come up with a new product. The common factor underlying each of these tasks is that you want to use the data to answer a question that you haven't answered before. The most effective process we have come up for getting those answers is the scientific research process.

Read more »

I co-founded a company! Meet Problem Forward Data Science

Jeff Leek 2019/05/20

I have some exciting news about something I've been working on for the last year or so. I started a company! It's called Problem Forward data science. I'm pumped about this new startup for a lot of reasons. My co-founder is one of my families closest friends, Jamie McGovern, who has more than 2 decades of experience in the consulting world and who I've known for 15 years. We are creating a cool new model of "data scientist as a service" (more on that below) We have a problem forward, not solution backward approach to data science that


PROBLEM FORWARD data science
Home
What We Do
How We Do It
Who We Are
Contact Us
Careers
PROBLEM FORWARD
DATA SCIENCE
Data science that starts with your problem, not our latest algorithm.

About us

(Andrew)

Andrew Jaffe

- Lead Investigator at the Lieber Institute for Brain Development
- Assistant 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

Now you

Introduce yourself to your neighbor

About us

https://bit.ly/2xLMh7l

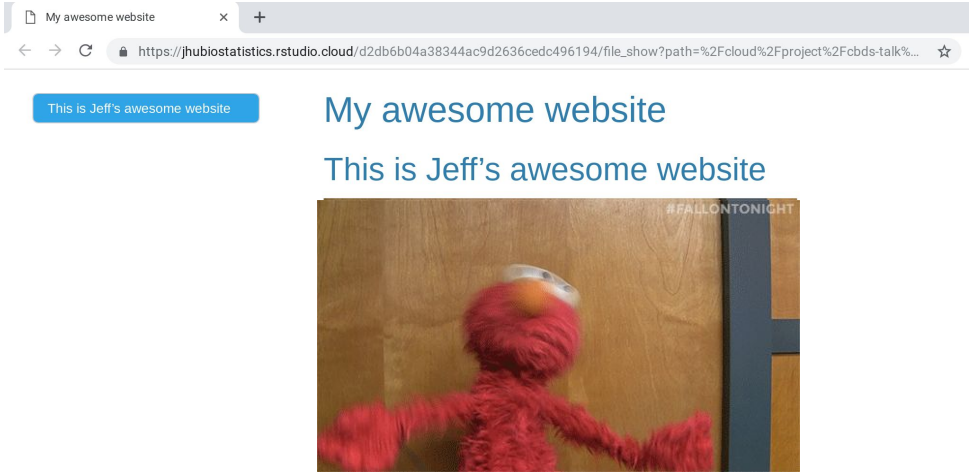
Skills self assessment

https://bit.ly/2YUy4B0

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/drXGoW1iudhKw/giphy.gif)
```



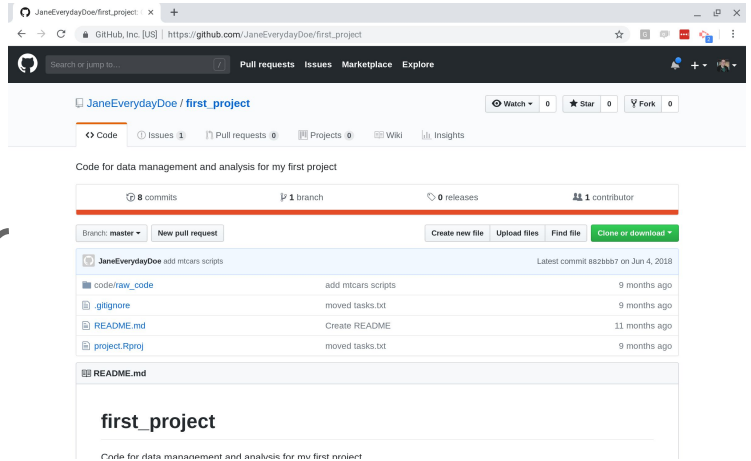
dbplyr

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

con <- dbConnect(
  bigquery(),
  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 = 'janeeverydaydoe'

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


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

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

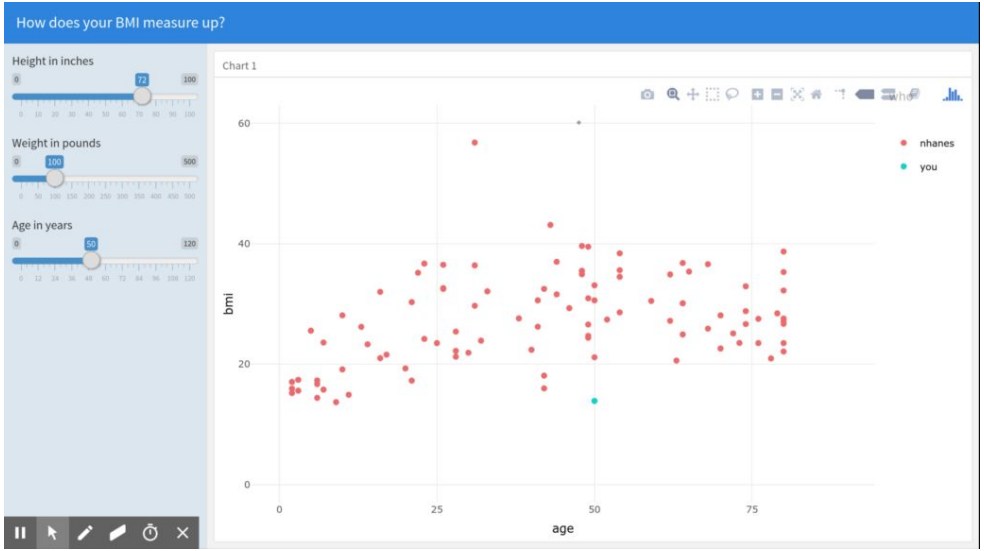



```
$id
[1] 130377298
$node_id
[1] "MDEwOlJlcG9zaXRvcnkxMzAzNzcyOTg="
$name
[1] "first_project"
$full_name
[1] "JaneEverydayDoe/first_project"
$owner$gravatar_id
[1] ""
$owner$url
[1] "https://api.github.com/users/JaneEverydayDoe"
```

...

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

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

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

www.nature.com/nm/journal/v12/n11/full/nm1491.html

DERIVING CHEMOSENSITIVITY FROM CELL LINES: FORENSIC BIOINFORMATICS AND REPRODUCIBLE RESEARCH IN HIGH-THROUGHPUT BIOLOGY

By Keith A. Baggerly* and Kevin R. Coombes†

U.T. M.D. Anderson Cancer Center

High-throughput biological assays such as microarrays let us ask very detailed questions about how diseases operate, and promise to let us personalize therapy. Data processing, however, is often not described well enough to allow for exact reproduction of the results, leading to exercises in "forensic bioinformatics" where aspects of raw data and reported results are used to infer what methods must have been employed. Unfortunately, poor documentation can shift from an inconvenience to an active danger when it obscures not just methods but errors. In this report, we examine several related papers purporting to use microarray-based signatures of drug sensitivity derived from cell lines to predict patient response. Patients in clinical trials are currently being allocated to treatment arms on the basis of these results. However, we show in five case studies that the results incorporate several simple errors that may be putting patients at risk. One theme that emerges is that the most common errors are simple (e.g., row or column offsets); conversely, it is our experience that the most simple errors are common. We then discuss steps we are taking to avoid such errors in our own investigations.

https://projecteuclid.org/euclid.aoas/1267453942

From the article:
Cancer trial errors revealed

2006 Anil Potti, a cancer geneticist at Duke University in Durham, North Carolina, and others file patent applications on the idea of using gene-expression data to predict sensitivity to cancer drugs. Potti is first author on a paper in *Nature Medicine*[1].

2007 Potti is last author on a paper in the *Journal of Clinical Oncology* (*JCO*)[2]. Duke begins three clinical trials to test Potti's predictors in patients with breast or lung cancer.

SEPTEMBER 2009 Keith Baggerly and Kevin Coombes, statisticians at the University of Texas M. D. Anderson Cancer Centre in Houston, publish a paper in *Annals of Applied Statistics*[3] stating that they could not replicate Potti's claims. Duke suspends the trials and asks a review panel to investigate.

NOVEMBER 2009 Potti places data underlying the *JCO* paper online. Baggerly writes to Sally Kornbluth, Duke vice-dean for research, and Michael Cuffe, Duke vice-president for medical affairs, to point out differences from raw data.

DECEMBER 2009 An unredacted copy of the report by Duke's review panel, later obtained by *Nature*, shows that the panel replicated Potti's claims using his data, but were unaware that those data contained discrepancies.

JANUARY 2010 Duke restarts clinical trials.

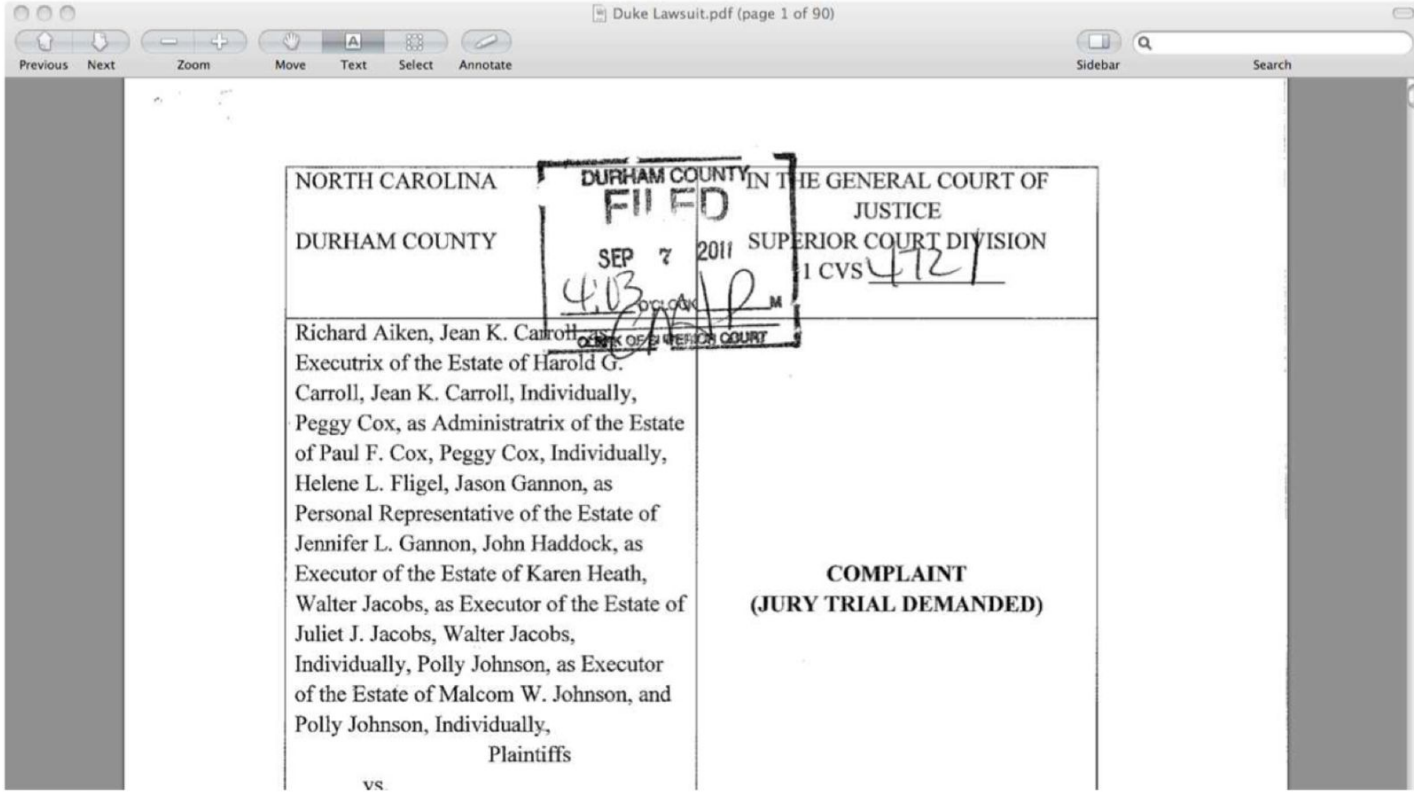
JULY 2010 *The Cancer Letter* reveals that Potti made false claims about his CV. Trials are suspended and an investigation begins. Harold Varmus, director of the National Cancer Institute in Bethesda, Maryland, asks the Institute of Medicine to review Duke's trials.

NOVEMBER 2010 *JCO* paper is retracted. Duke closes the trials permanently. Potti resigns.

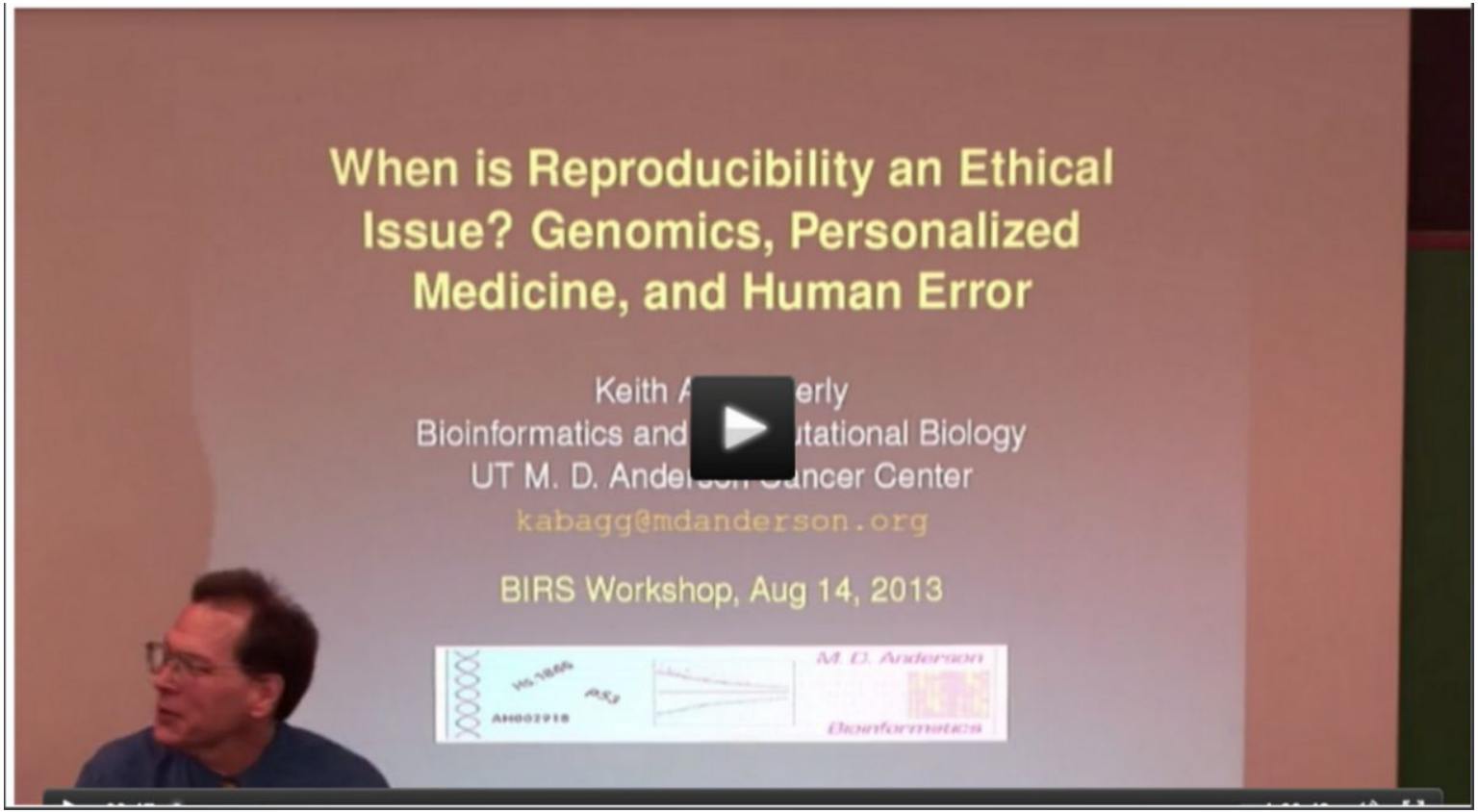
DECEMBER 2010 Institute of Medicine study begins, but will now focus more generally on criteria for genomics predictor.

JANUARY 2011 *Nature Medicine* paper is retracted.

www.nature.com/news/2011/110111/full/469139a/box/1.html



http://www.dukechronicle.com/articles/2015/05/03/duke-lawsuit-involving-cancer-patients-linked-anil-potti-settled



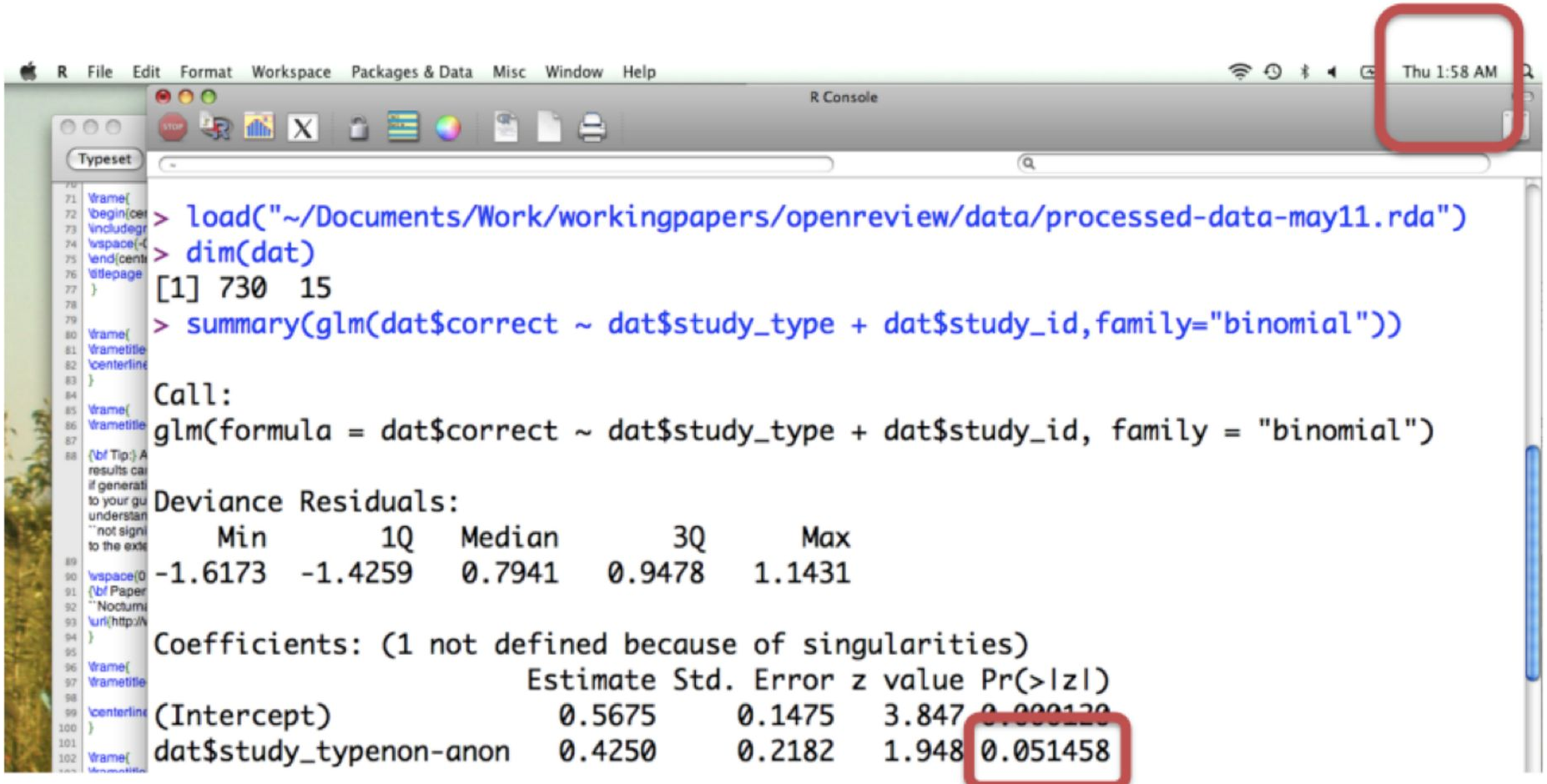
http://www.birs.ca/events/2013/5-day-workshops/13w5083/videos/watch/201308141121-Baggerly.mp4

http://www.ted.com/talks/dan_meyer_math_curriculum_makeover

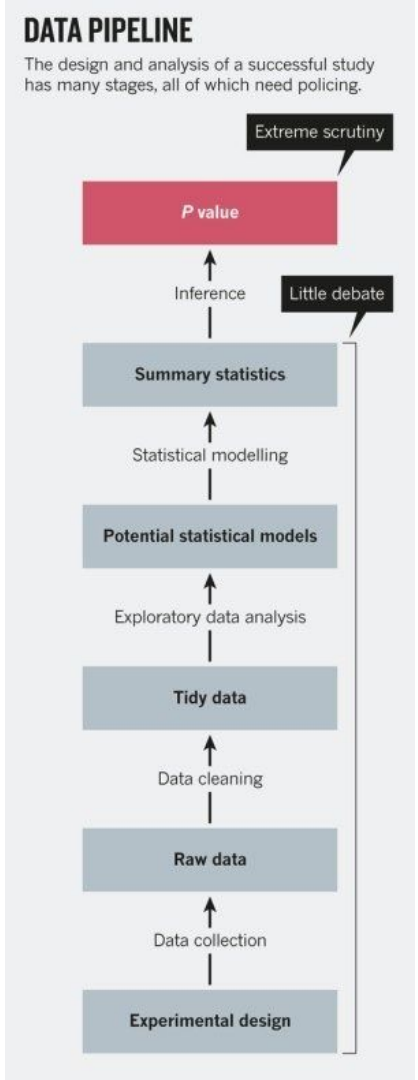
“Ask yourselves, what problem have you solved, ever, that was worth solving, where you knew knew all of the given information in advance? Where you didn’t have a surplus of information and have to filter it out, or you didn’t have insufficient information and have to go find some?

-Dan Meyer”

Doesn’t seem that important….


R File Edit Format Workspace Packages & Data Misc Window Help
Thu 1:58 AM
R Console
Typeset
> 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
dat$study_typenon-anon 0.4250 0.2182 1.948 0.051458


ON THE ONE
HAND...



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

Leek and Peng Nature 2015

JAMA Network
JAMA
Journals
Enter Search

This Issue Citations 1,078

PDF More Cite Permissions

Original Contribution FREE

March 8, 2006

Fine Particulate Air Pollution and Hospital Admission for Cardiovascular and Respiratory Diseases

Francesca Dominici, PhD; Roger D. Peng, PhD; Michelle L. Bell, PhD; et al

Article Information

Eight outcomes were considered based on the ICD-9 codes for 5 cardiovascular outcomes (heart failure [428], heart rhythm disturbances [426-427], cerebrovascular events [430-438], ischemic heart disease [410-414, 429], peripheral vascular disease [440-448]), 2 respiratory outcomes (chronic obstructive pulmonary disease [COPD; 490-492], respiratory tract infections [464-466, 480-487]), and hospitalizations caused by injuries and other external causes (800-849). The county-wide daily hospitalization rates for each outcome for 1999-2002 appear in Table 1.

The study population includes 11.5 million Medicare enrollees residing an average of 5.9 miles from a PM2.5 monitor. The analysis was restricted to the 204 US counties with populations larger than 200 000. Of these 204 counties, 90 had daily PM2.5 data across the study period and the remaining counties had PM2.5 data collected once every 3 days for at least 1 full year. The locations of the 204 counties appear in Figure 1. The counties were clustered into 7 geographic regions by applying the K-means clustering algorithm to longitude and latitude for the counties.10,11

aps
ASSOCIATION FOR
PSYCHOLOGICAL SCIENCE

General Article

False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Psychological Science
XX(X) 1–8
© The Author(s) 2011
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797611417632
http://pss.sagepub.com
SAGE

Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]
[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berkeley

http://pss.sagepub.com/content/22/11/1359.abstract

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

Andrew Gelman† and Eric Loken‡

14 Nov 2013

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

Explorable Multiverse Analyses

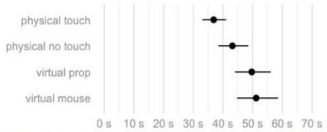
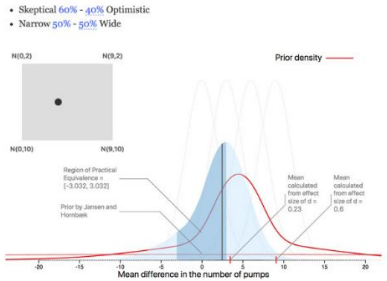


Figure 3. Average task completion time (geometric mean) for each condition. Error bars are 95% t-based CIs.

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



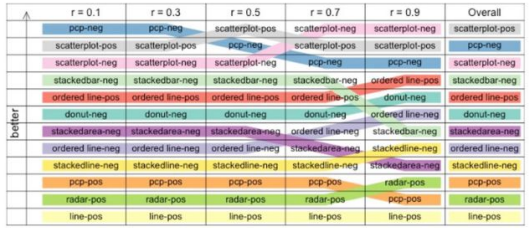
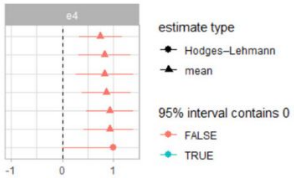


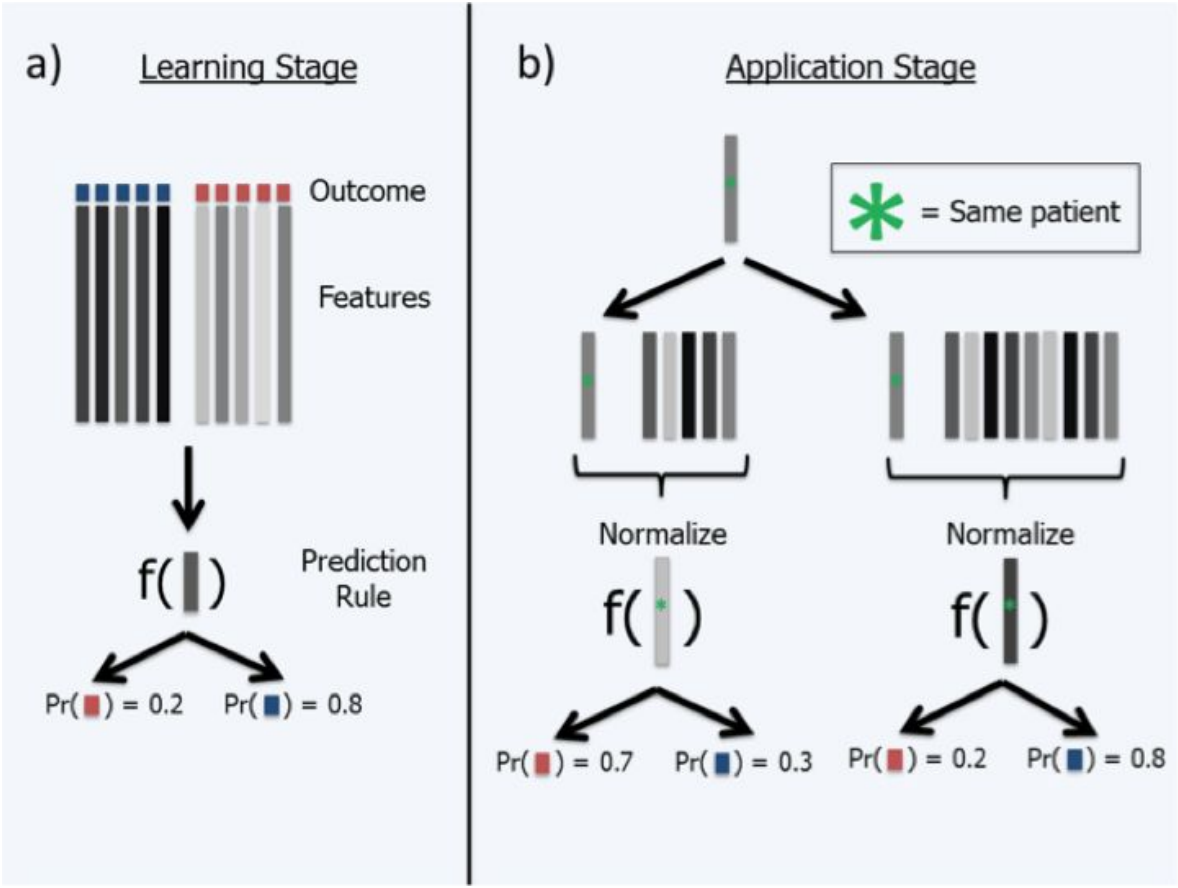
Figure 4. Perceptually-driven ranking of visualizations depending on the correlation sign (-neg / -pos), as a function of correlation value (r) and overall (right column).



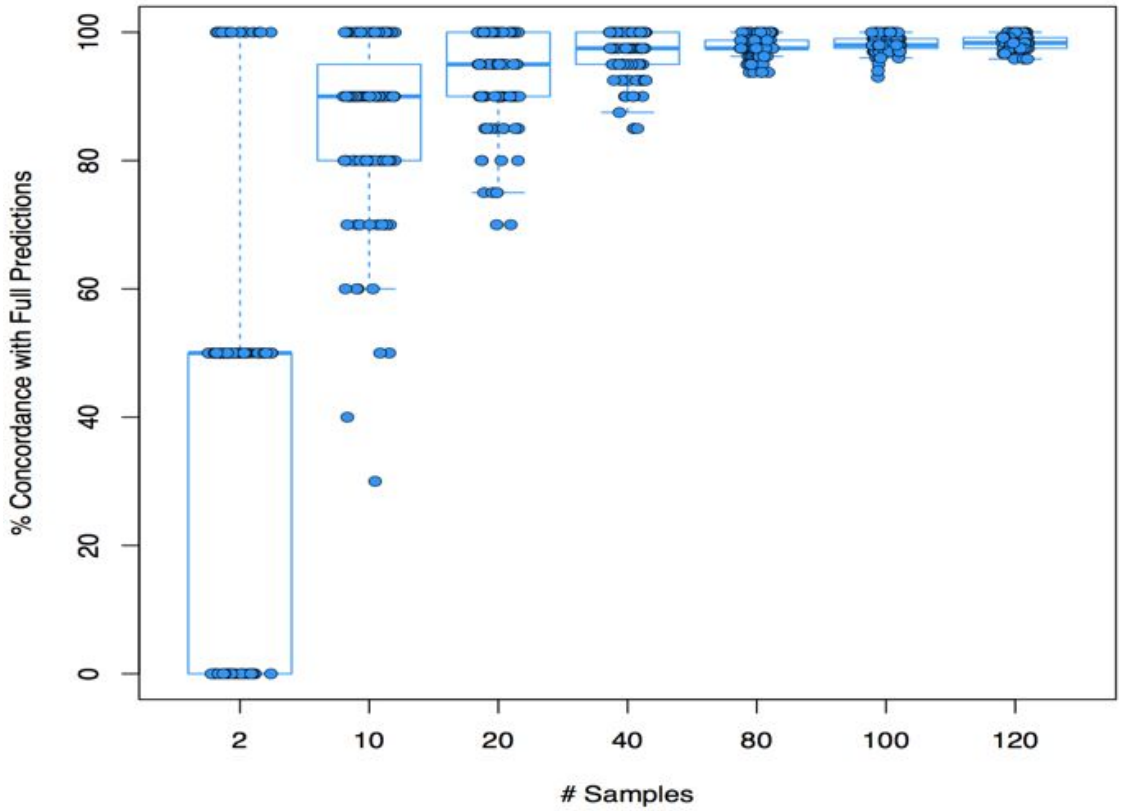
Pierre Dragicevic (Inria), Yvonne Jansen (CNRS - Sorbonne Université), Abhraneel Sarma (University of Michigan)
Matthew Kay (University of Michigan), Fanny Chevalier (University of Toronto)

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

https://explorablemultiverse.github.io/examples/frequentist/



http://www.ncbi.nlm.nih.gov/pubmed/25788628



http://www.ncbi.nlm.nih.gov/pubmed/25788628

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.

http://adage.com/article/digitalnext/dear-madison-avenue-set-data-scientists-free/298676/

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 ›

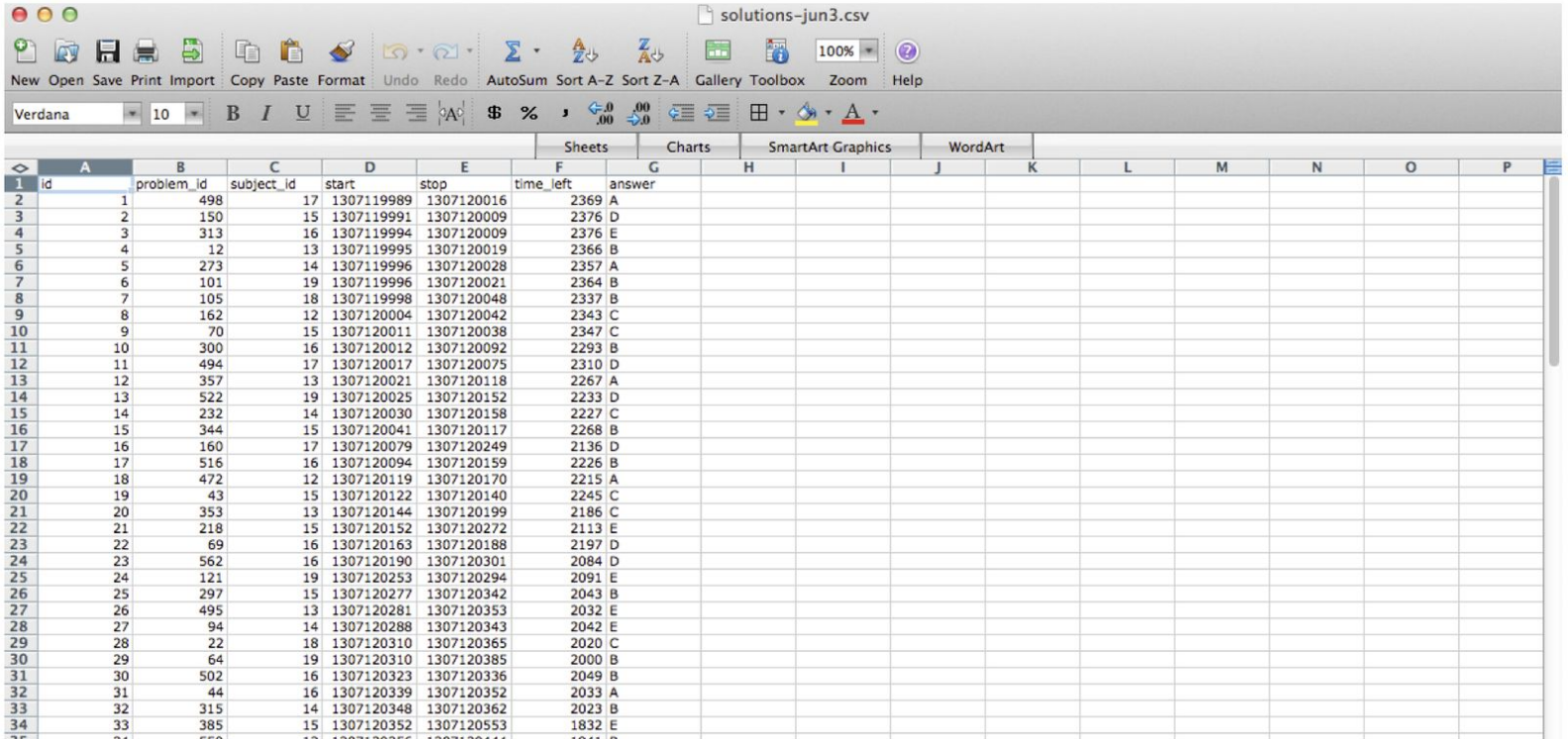
Case Study: Cloud Supercomputing from AWS Powers... READ ›

Get Started with AWS

CREATE A FREE ACCOUNT ›

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

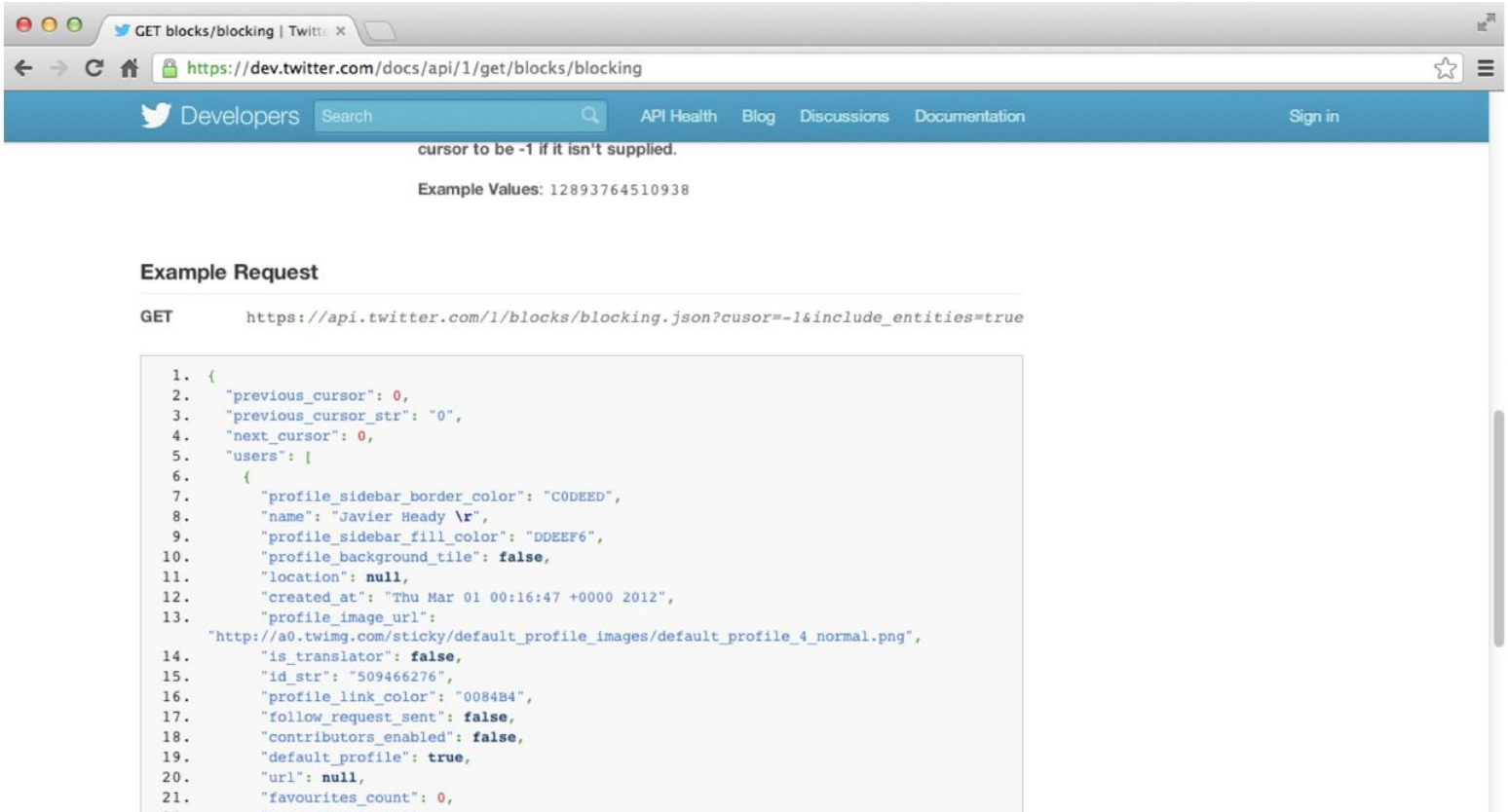
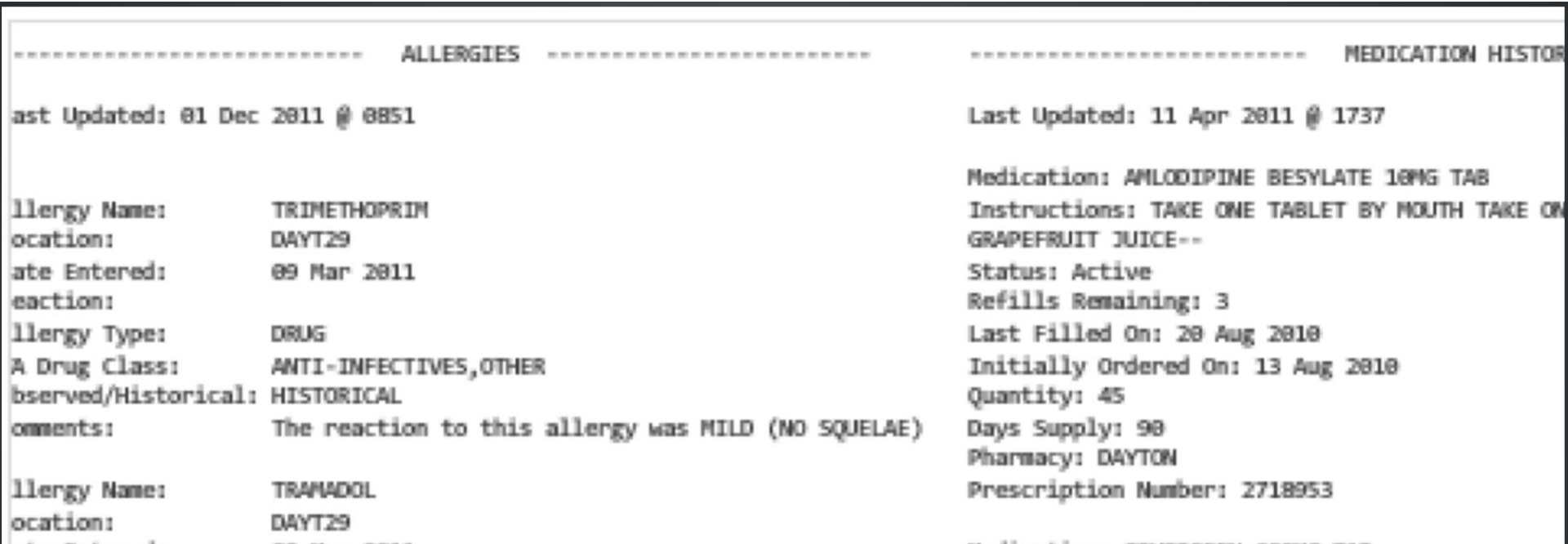
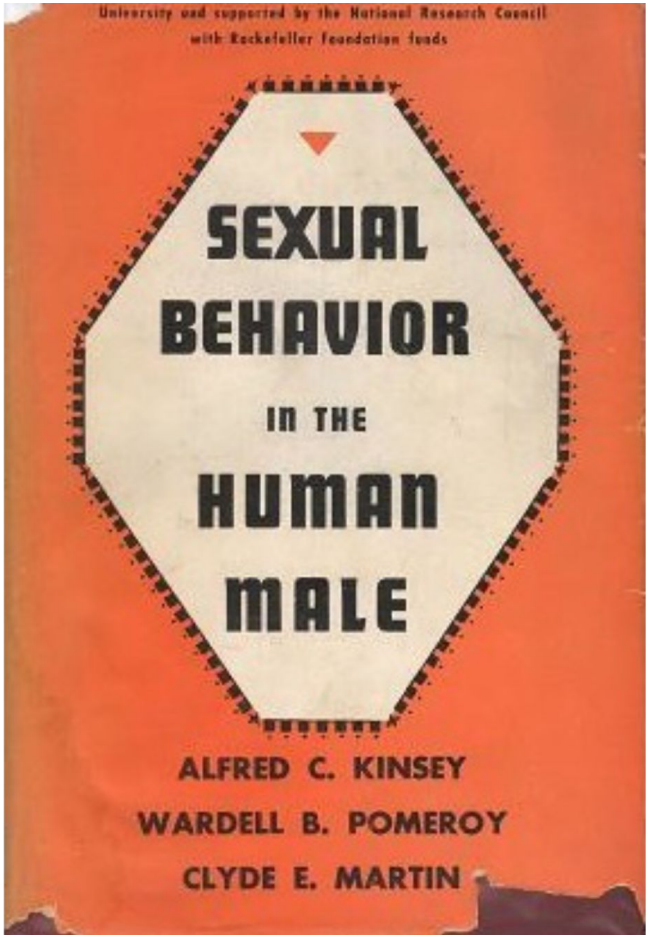
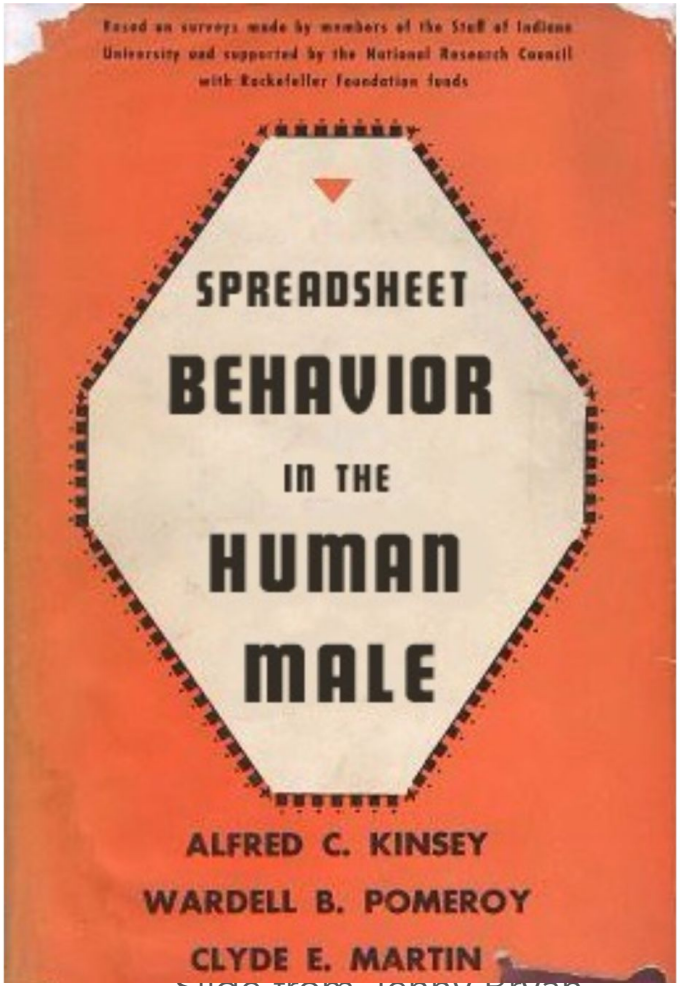
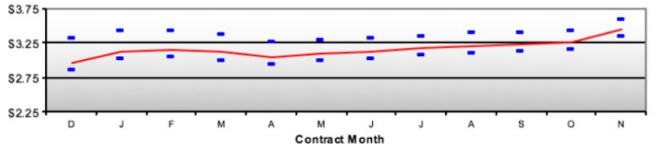
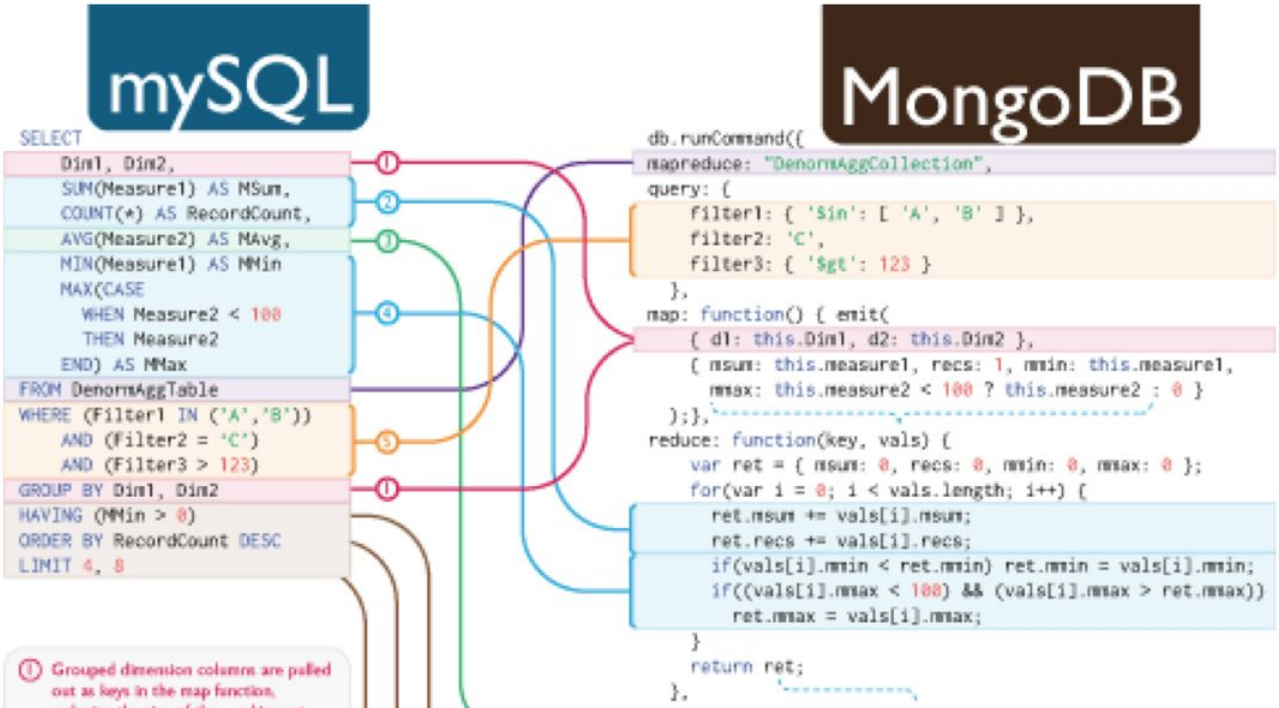
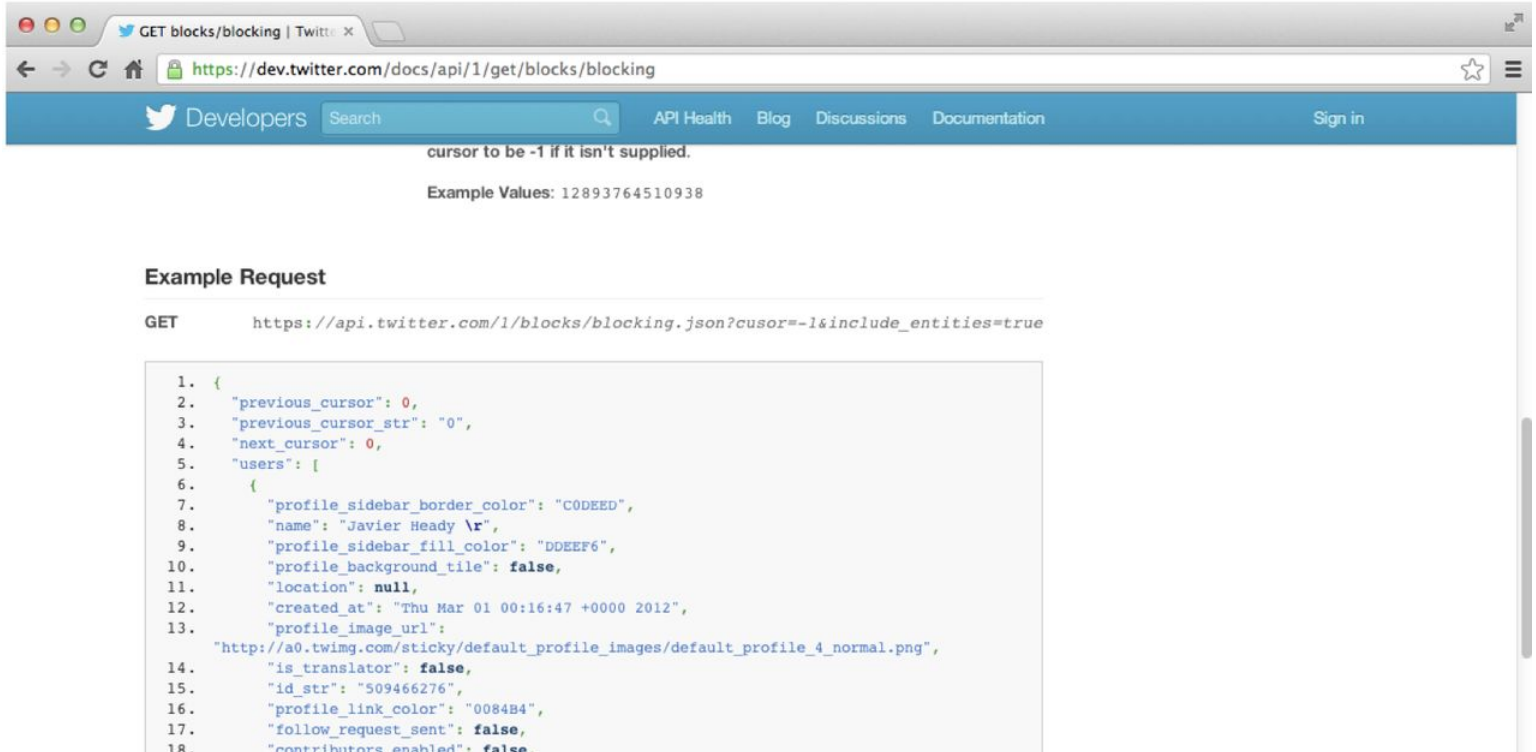
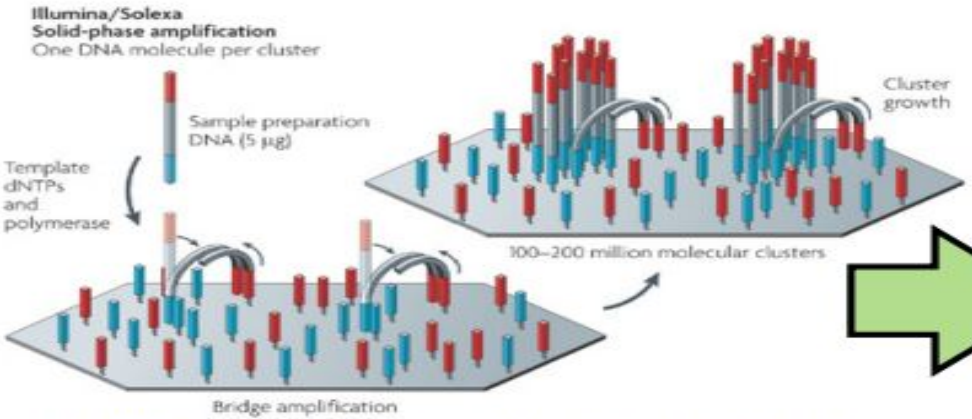
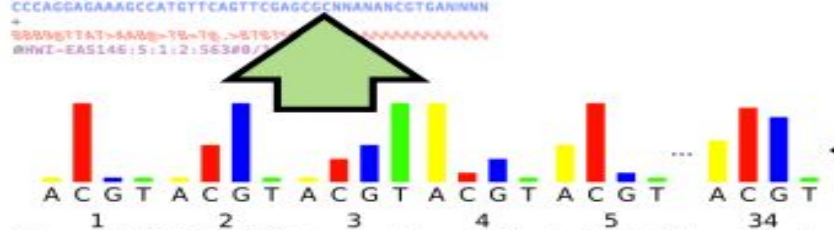
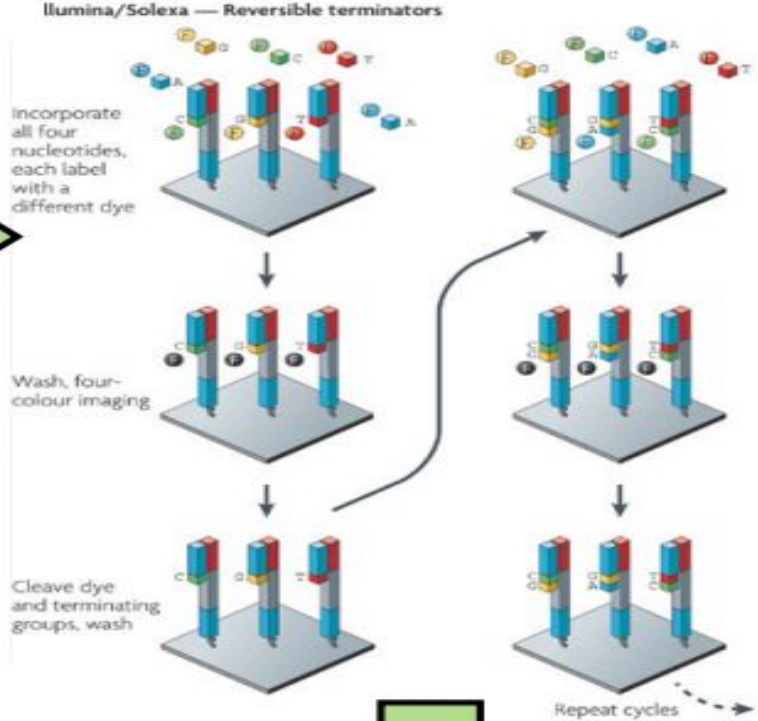
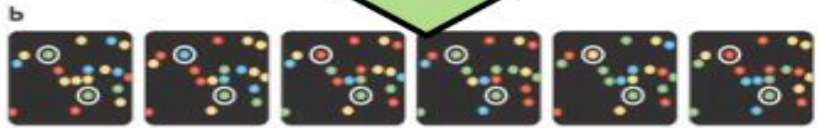
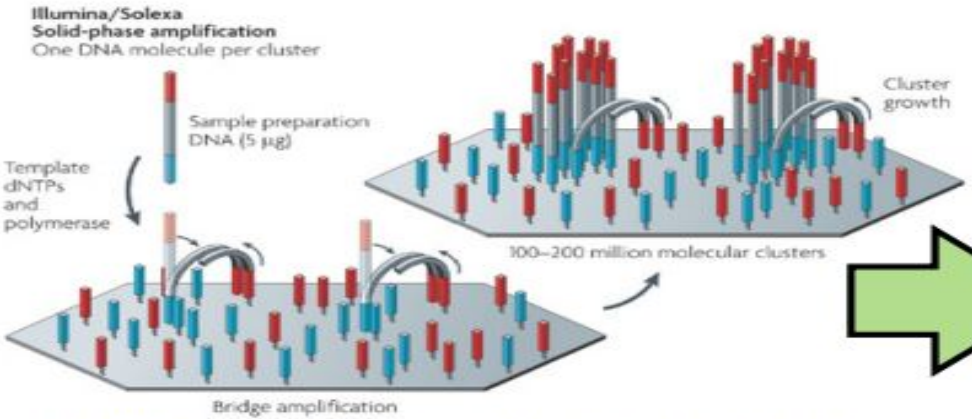
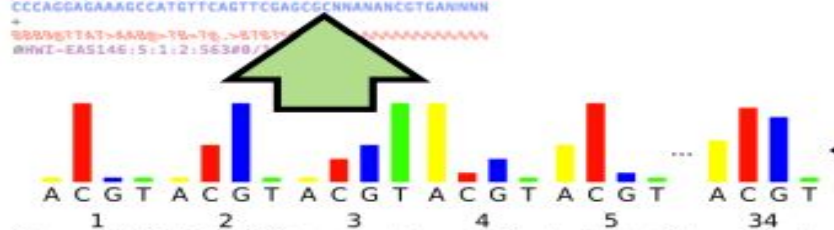
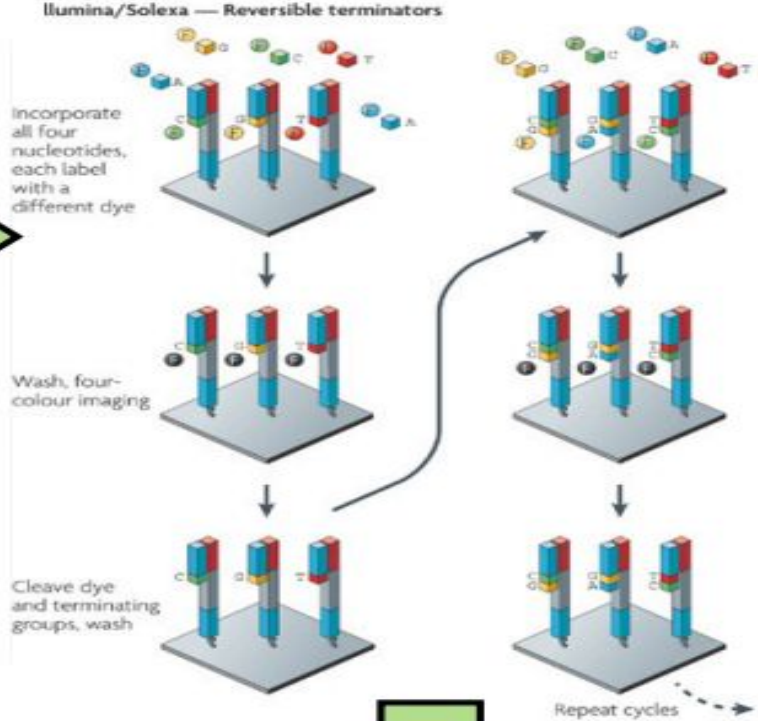
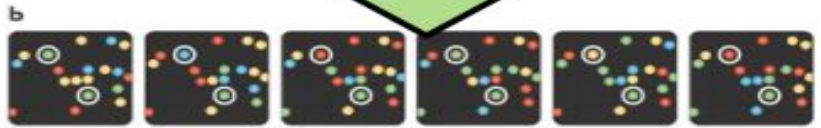
What you wished data looked like



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

What it actually looks like

http://healthdesignchallenge.com/

```
@HWI-EAS121:4:100:1783:550#0/1
CGTTACGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACGGATCTCGTATGCGGTCTGCTGCGTGACAAGACAGGGG
+HWI-EAS121:4:100:1783:550#0/1
aaaaa`b_aa`aa`YaX]aZ`aZM^Z]YRa]YSG[[ZREQLHESDHNDDHNMEEDDMPENITKFLFEEDDDHEJQMEDDD
@HWI-EAS121:4:100:1783:1611#0/1
GGGTGGGCATTTCCACTCGCAGTATGGGTTGCCGCACGACAGGCAGCGGTCAGCCTGCGCTTTGGCCTGGCCTTCGGAAA
+HWI-EAS121:4:100:1783:1611#0/1
a``^\__`_```^a``a`^a_^__]a_]\]`a_____`_^^`]X]_]XTV_\]]NX_XVX]]_TTTTG[VTHPN]VFDZ
@HWI-EAS121:4:100:1783:322#0/1
CGTTTATGTTTTTGAATATGTCTTATCTTAACGGTTATATTTTAGATGTTGGTCTTATTCTAACGGTCATATATTTTCTA
+HWI-EAS121:4:100:1783:322#0/1
abaa`^aaaaabbbaabababbbbbb`bbbb_bbbbbbbb`bbbaV^_a``a``]``aT]a__V\]]_]^a`]a_abbaV__
@HWI-EAS121:4:100:1783:1394#0/1
GGGTCTTTATTGGTCTGGTGATCCCCCATATTCTCCGGTTGTGTGGTTTAACCGATCATCGCGCATTACTTCCCGGCTGC
+HWI-EAS121:4:100:1783:1394#0/1
```[aa\b^^[]aabbb][`a_abbb`a``bbbbbabaabaaaab_VZa_^___bab_X`[a\HV_[_]_[^_X\T_VQQ
@HWI-EAS121:4:100:1783:207#0/1
```
```

What it actually looks like

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



What it actually looks like



---------- ALLERGIES ----------

ast Updated: 01 Dec 2011 @ 0851

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

llergy Name: TRAMADOL
ocation: DAYT29

---------- MEDICATION HISTOR

Last Updated: 11 Apr 2011 @ 1737

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

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)



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



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

Enron North America - West Gas

November 9, 2001

ENA - West Gas Contacts

Houston Office		Regional Offices		
Barry Tycholiz	(713) 853-1587	Mark Whitt	(303) 575-6473	Denver
Kim Ward	(713) 853-0685	Paul Lucci	(303) 575-6474	Denver
Stephanie Miller	(713) 853-1688	Tyrell Harrison	(303) 575-6478	Denver
Philip Polsky	(713) 853-5181	Dave Fuller	(503) 464-3732	Portland

Forward Prices (US$/MMBtu)

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



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

	IF CIG Rocky Mountains Fixed Price BID	Fixed Price OFFER	Basis BID	Basis OFFER
Cash	1.940	1.960		
ROM	1.960	1.980		
Dec-01	2.345	2.365	(0.615)	(0.595)
Dec-01 to Mar-02	2.548	2.568	(0.540)	(0.520)
Apr-02 to Oct-02	2.471	2.491	(0.695)	(0.675)
Nov-02 to Mar-03	3.311	3.331	(0.340)	(0.320)
One Year Strip*	2.551	2.571	(0.614)	(0.594)

	IF EL Paso Permian Fixed Price BID	Fixed Price OFFER	Basis BID	Basis OFFER
Cash	2.375	2.395		
ROM	2.420	2.440		
Dec-01	2.700	2.720	(0.260)	(0.240)
Dec-01 to Mar-02	2.855	2.875	(0.233)	(0.213)
Apr-02 to Oct-02	3.009	3.029	(0.158)	(0.138)
Nov-02 to Mar-03	3.499	3.519	(0.153)	(0.133)
One Year Strip*	2.982	3.002	(0.182)	(0.162)

	IF EL Paso San Juan Fixed Price BID	Fixed Price OFFER	Basis BID	Basis OFFER
Cash	2.450	2.470		
ROM	2.350	2.370		
Dec-01	2.560	2.580	(0.400)	(0.380)
Dec-01 to Mar-02	2.743	2.763	(0.345)	(0.325)
Apr-02 to Oct-02	2.801	2.821	(0.365)	(0.345)
Nov-02 to Mar-03	3.421	3.441	(0.230)	(0.210)
One Year Strip*	2.817	2.837	(0.347)	(0.327)

	AECO / NIT Fixed Price BID	Fixed Price OFFER	Basis BID	Basis OFFER
Cash	2.376	2.396		
ROM	2.398	2.418		
Dec-01	2.552	2.572	(0.408)	(0.388)
Dec-01 to Mar-02	2.616	2.636	(0.472)	(0.452)
Apr-02 to Oct-02	2.661	2.681	(0.505)	(0.485)
Nov-02 to Mar-03	3.216	3.236	(0.435)	(0.415)
One Year Strip*	2.676	2.696	(0.488)	(0.468)

	IF NWPL Canadian Border (Sumas) Fixed Price BID	Fixed Price OFFER	Basis BID	Basis OFFER
Cash	2.480	2.500		
ROM	2.460	2.480		
Dec-01	2.800	2.820	(0.160)	(0.140)
Dec-01 to Mar-02	2.892	2.912	(0.196)	(0.176)
Apr-02 to Oct-02	2.796	2.816	(0.370)	(0.350)
Nov-02 to Mar-03	3.706	3.726	0.055	0.075
One Year Strip*	2.880	2.900	(0.285)	(0.265)

	IF PEPL TX-OK Fixed Price BID	Fixed Price OFFER	Basis BID	Basis OFFER
Cash	2.530	2.550		
ROM	2.530	2.550		
Dec-01	2.828	2.848	(0.133)	(0.113)
Dec-01 to Mar-02	2.958	2.978	(0.130)	(0.110)
Apr-02 to Oct-02	3.046	3.066	(0.120)	(0.100)
Nov-02 to Mar-03	3.531	3.551	(0.120)	(0.100)
One Year Strip*	3.041	3.061	(0.123)	(0.103)

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

Desiree Narango
@DLNarango

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



Where data actually is

https://data.baltimorecity.gov/

Data brainstorming

https://bit.ly/2LXhVap

Raw & processed data

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

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

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

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

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

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

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

Variables: A measurement or characteristic of an item

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

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

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



Source: Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010

name
sequence
quality scores

x 100s of millions



Source: Whiteford et al. Swift: primary data analysis for the Illumina Solexa sequencing platform. Bioinformatics. 2009





Source: Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010

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



Source: Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010

name
sequence
quality scores

x 100s of millions



Source: Whiteford et al. Swift: primary data analysis for the Illumina Solexa sequencing platform. Bioinformatics. 2009





Source: Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010

Raw data

```
@HWI-EAS121:4:100:1783:550#0/1
CGTTACGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACGGATCTCGTATGCGGTCTGCTGCGTGACAAGACAGGGG
+HWI-EAS121:4:100:1783:550#0/1
aaaaa`b_aa`aa`YaX]aZ`aZM^Z]YRa]YSG[[ZREQLHESDHNDDHNMEEDDMPENITKFLFEEDDDHEJQMEDDD
@HWI-EAS121:4:100:1783:1611#0/1
GGGTGGGCATTTCCACTCGCAGTATGGGTTGCCGCACGACAGGCAGCGGTCAGCCTGCGCTTTGGCCTGGCCTTC
+HWI-EAS121:4:100:1783:1611#0/1
a```^\__`_```^a``a`^a_^__]a_]\]`a_____`_^^`]X]_]XTV_\]]NX_XVX]]_TTTTG[V
@HWI-EAS121:4:100:1783:322#0/1
CGTTTATGTTTTTGAATATGTCTTATCTTAACGGTTATATTTTAGATGTTGGTCTTATTCTAACGGTCA      TTCTA
+HWI-EAS121:4:100:1783:322#0/1
abaa`^aaaaabbbaabababbbbbb`bbbb_bbbbbbbb`bbbaV^_a``a``]``aT]a__V\]]_]^      bbaV__
@HWI-EAS121:4:100:1783:1394#0/1
GGGTCTTTATTGGTCTGGTGATCCCCCATATTCTCCGGTTGTGTGGTTTAACCGATCATCGCGCATTAC      CTGC
+HWI-EAS121:4:100:1783:1394#0/1
```[aa\b^^[]aabbb][`a_abbb`a``bbbbbabaabaaaab_VZa_^___bab_X`[a\HV_[_]_[^_
@HWI-EAS121:4:100:1783:207#0/1
CCCTGGGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAACA
+HWI-EAS121:4:100:1783:207#0/1
abba`Xa\^\\`aa]ba__bba[a_O_a`aa`aa`a]^V]X_a^YS\R_\H_[]\ZT
@HWI-EAS121:4:100:1783:455#0/1
GGGTAATTCAGGGACAATGTAATGGCTGCACAAAAAAATACATCTTTCATGTTCCAT
+HWI-EAS121:4:100:1783:455#0/1
abb_babbabaabbbbbbbbbbbbbbba\`b`\abbbabbbbabbbbbbaabbbbb
```

**Processing**
**Computing**
**Summarizing**
**Deleting**

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

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

Journal of Statistical Software

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

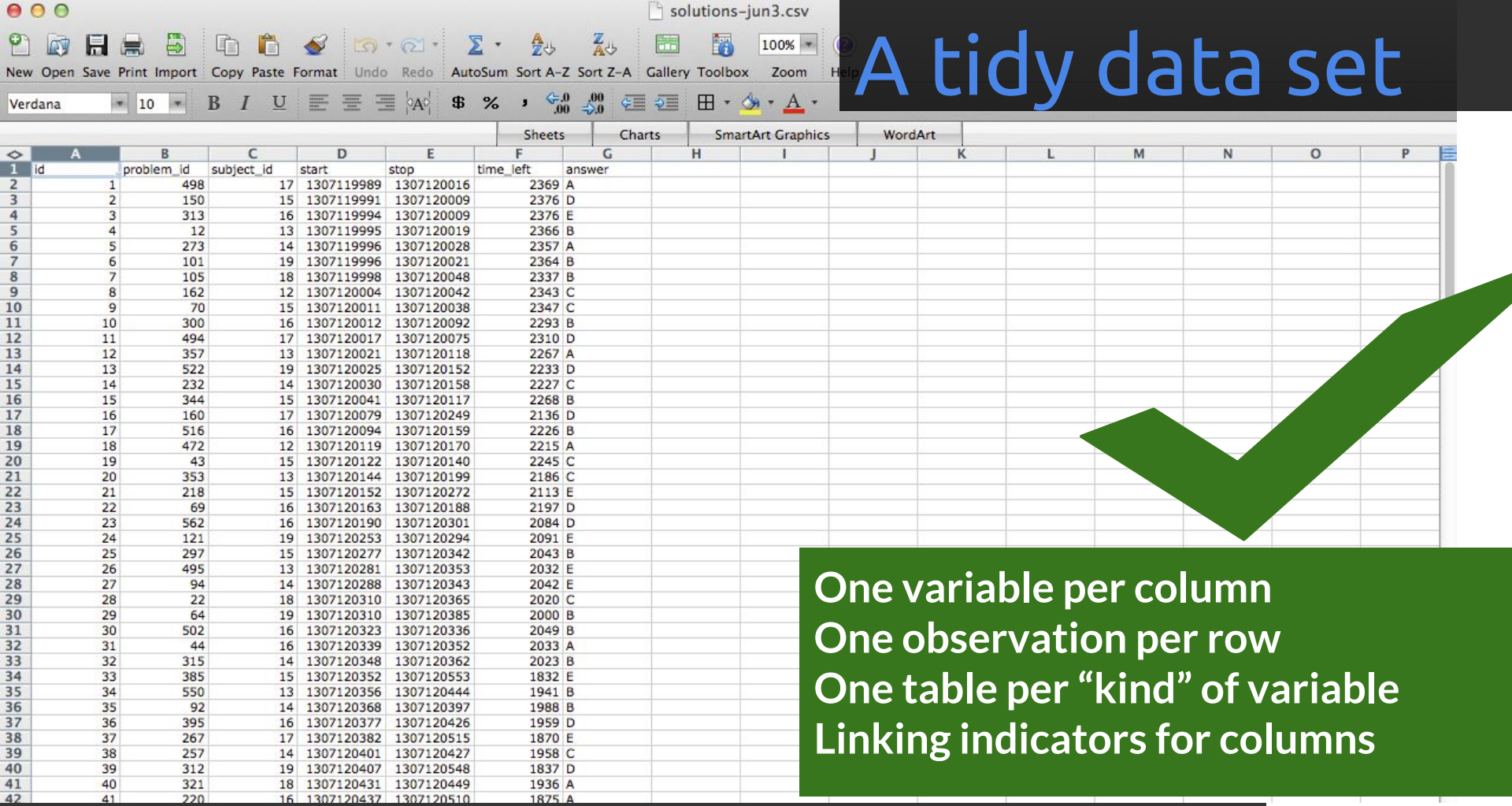
Tidy Data

Hadley Wickham
RStudio

Abstract

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

A tidy data set



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

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

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)

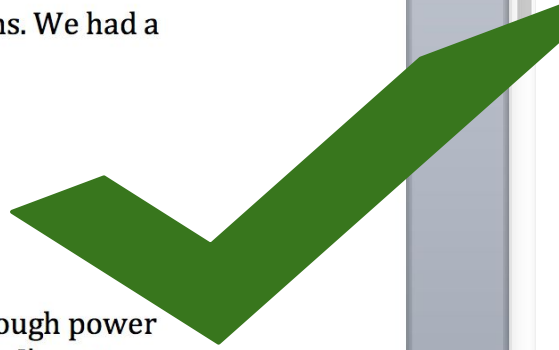
G. Diagnosis: DSM/ICD-10 diagnosis

H. Group: control, schizophrenia, or prodromal. I don't think we have enough power to run them as three groups so I combined prod
sure if this was ok. Is it appropriate to do a ttes
SZ?

I. Medication: mostly untreated

J. Education more or less than 13 years

K. current smoking status: yes or no



Variable names
Variable descriptions
Variable units
Study design quirks

RStudio

Recipe

Go to file/function

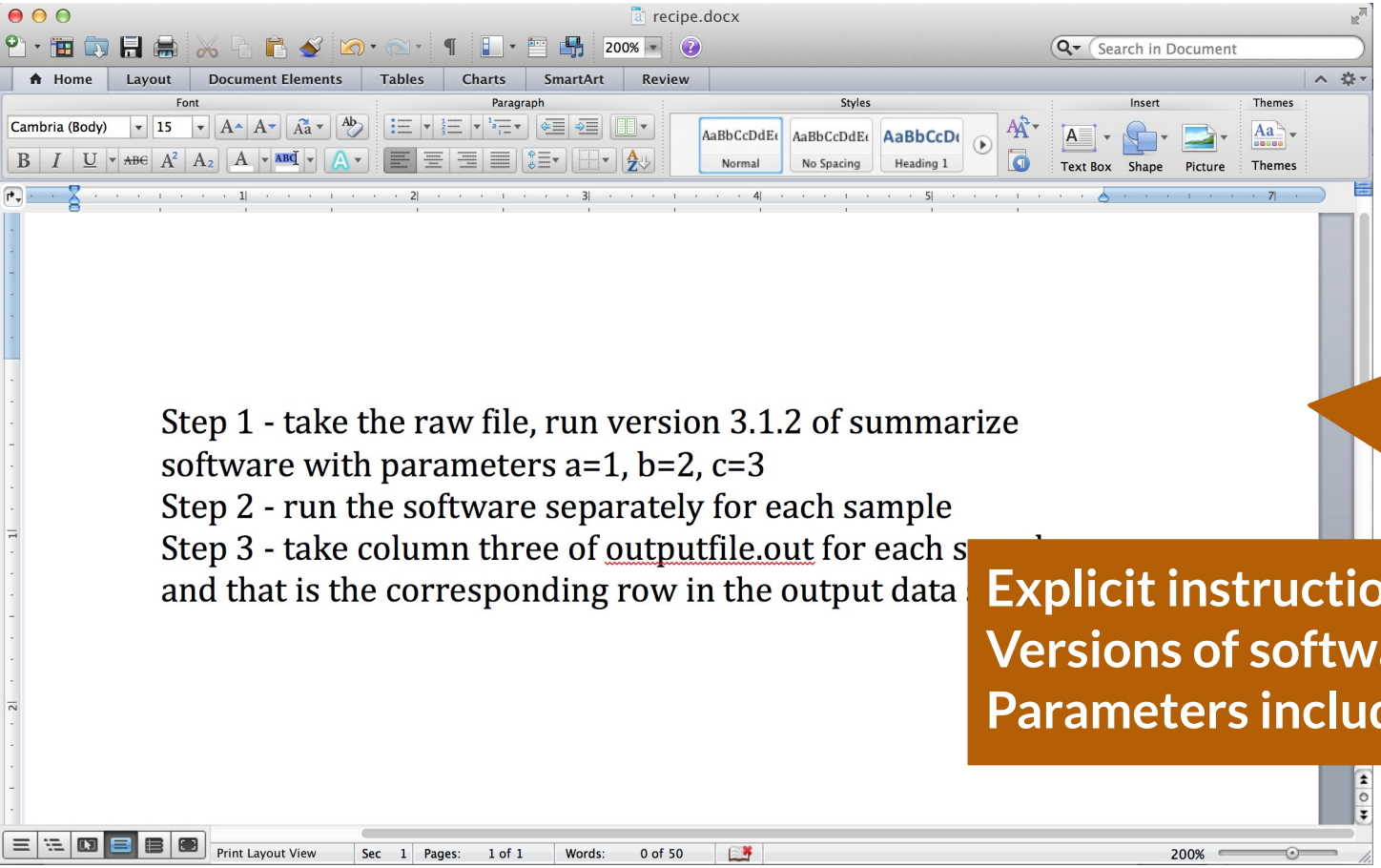
geuvadis.Rmd

Knit HTML | Run | Chunks

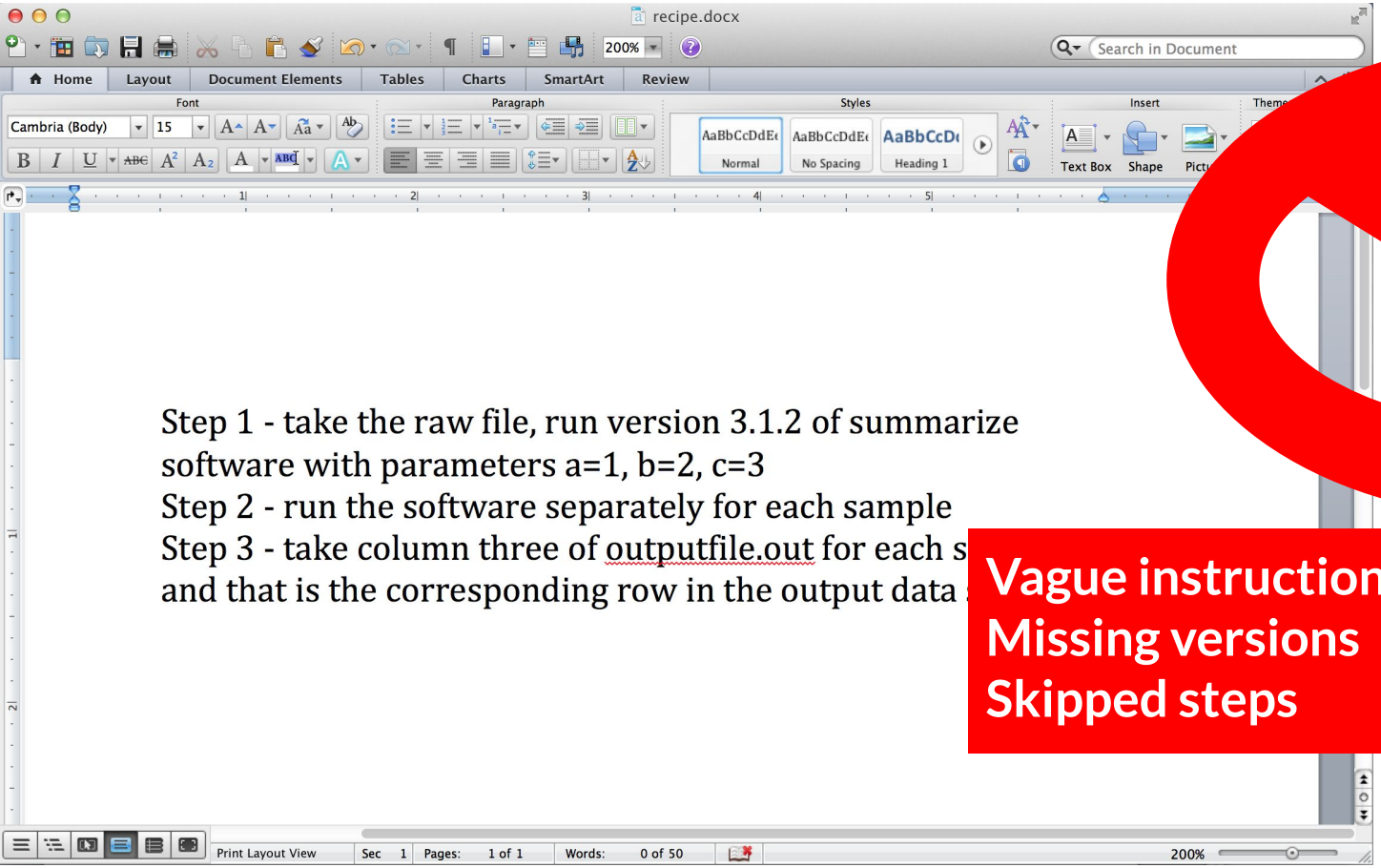
```
33  library(sva)
34  library(ffpe)
35  library(RColorBrewer)
36  library(corrplot)
37  library(limma)
38  trop = RSkittleBrewer('tropical')
39  ```
40
41
42  ## Load the data
43
44  You will need to download the GEUVADIS ballgown object from this site: https://github.com/a[illegible]uzee
    /ballgown_code
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
```

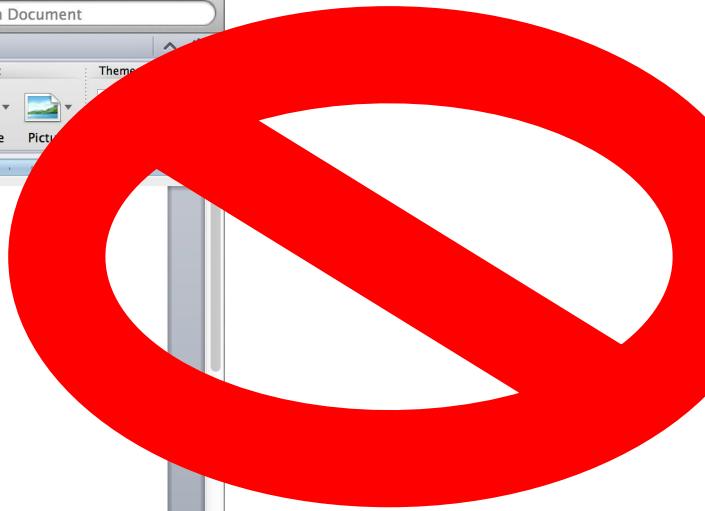
1:1 (Top Level) R Markdown

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



Explicit instructions
Versions of software
Parameters included





Vague instructions
Missing versions
Skipped steps

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

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

Rules for Tidy Spreadsheets

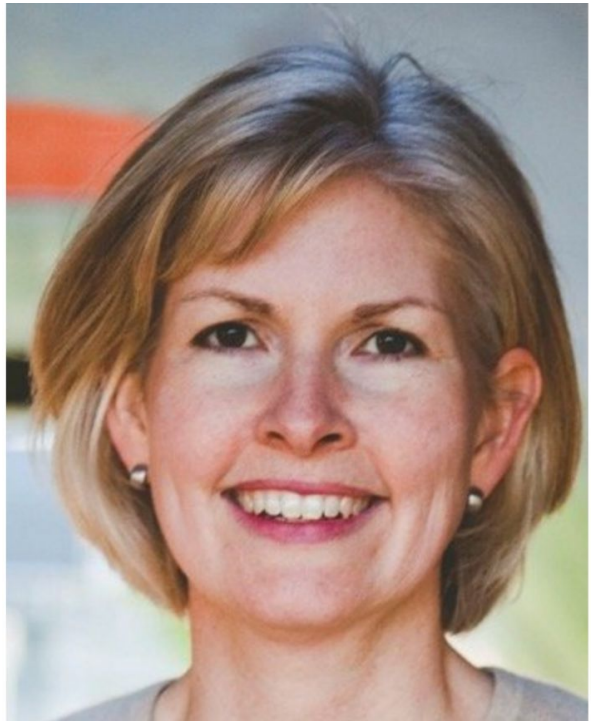
1. Be consistent
2. Choose good names for things
3. Write dates as YYYY-MM-DD
4. No empty cells
5. Put just one thing in a cell
6. Don't use font color or highlighting as data
7. Save the data as plain text files

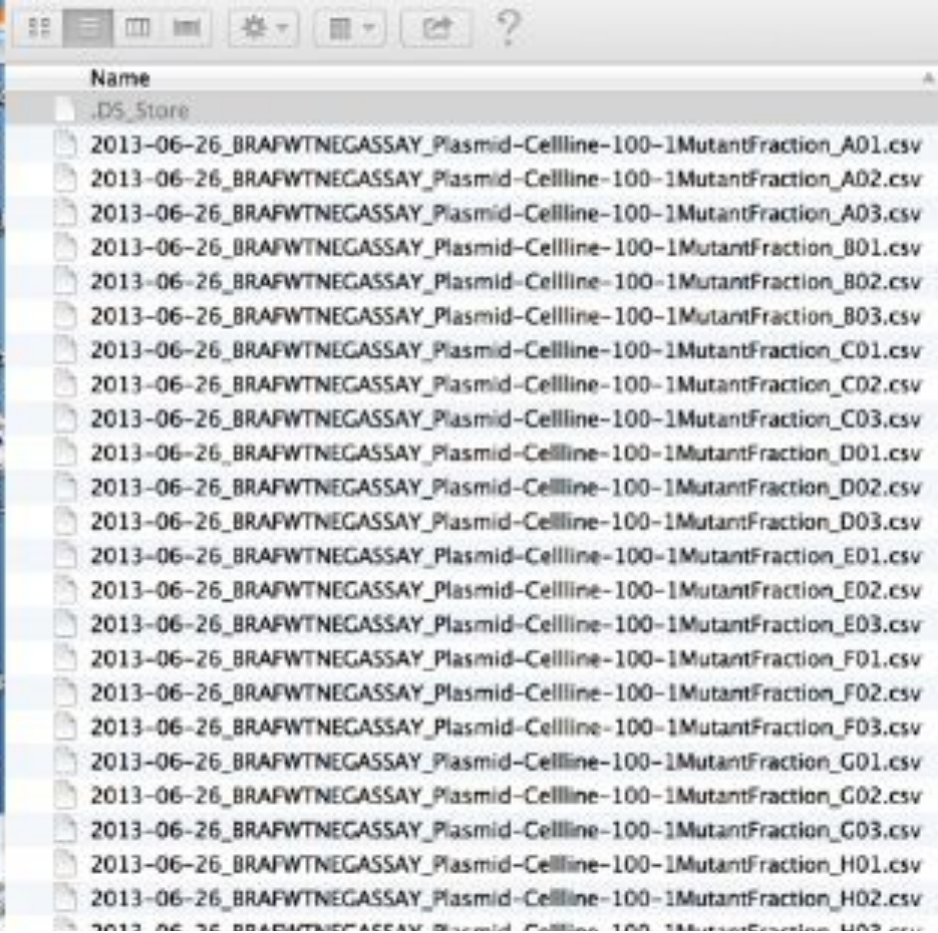
https://www.tandfonline.com/doi/full/10.1080/00031305.2017.1375989

Organize thyself

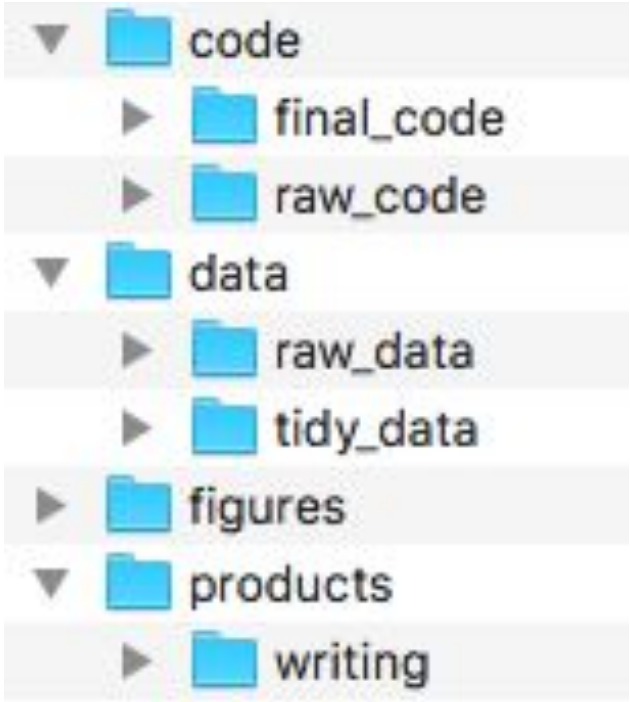
"File organization and naming are powerful weapons against chaos."

- Jenny Bryan





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

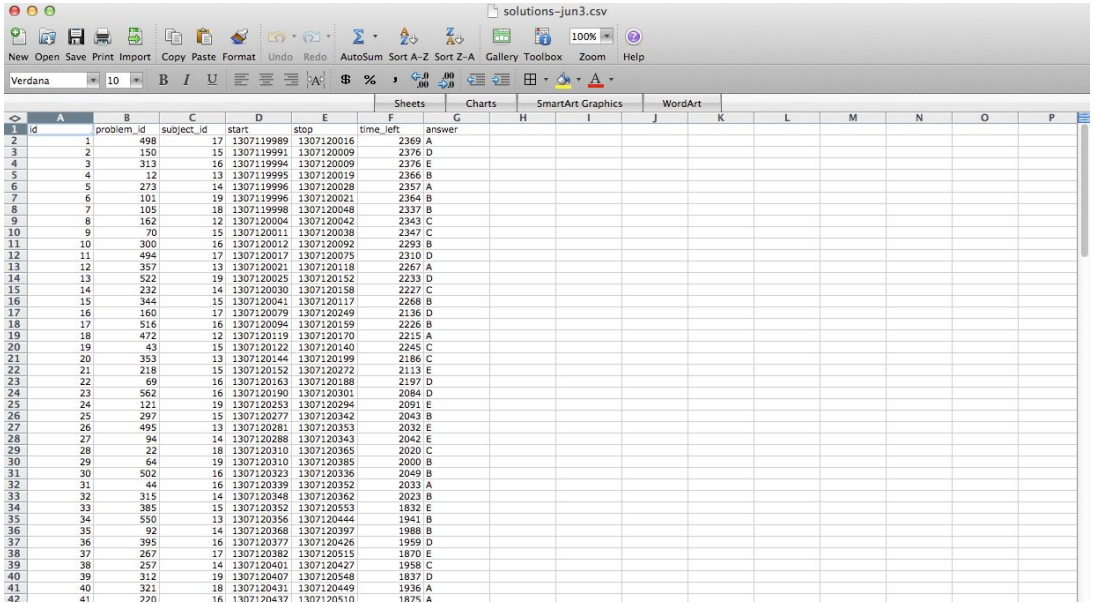

code
final_code
raw_code
data
raw_data
tidy_data
figures
products
writing

Raw data

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

ast Updated: 01 Dec 2011 @ 0851

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

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

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

Last Updated: 11 Apr 2011 @ 1737

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

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

Processed data



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

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

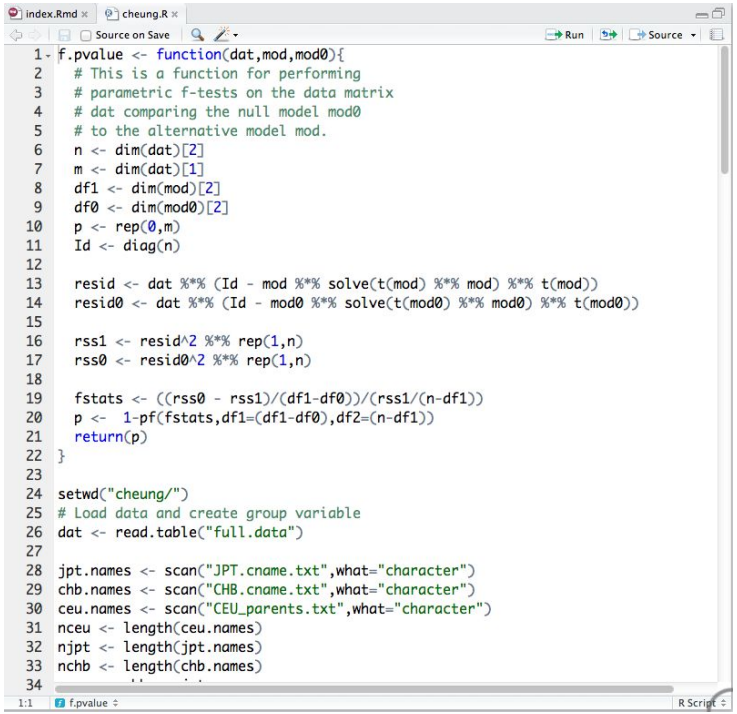
Raw scripts

raw_cheung_analysis.R

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

ll <- strsplit(dts,"-")

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

jul <- julian(mm,dd,yy)

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

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

}


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


ceufiles <- array[sample %in% ceuparents]

```

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

Final scripts



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

README.md

This is the README file for my_first_project

Last updated: 02-Mar-2018

The folders in this project are:

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

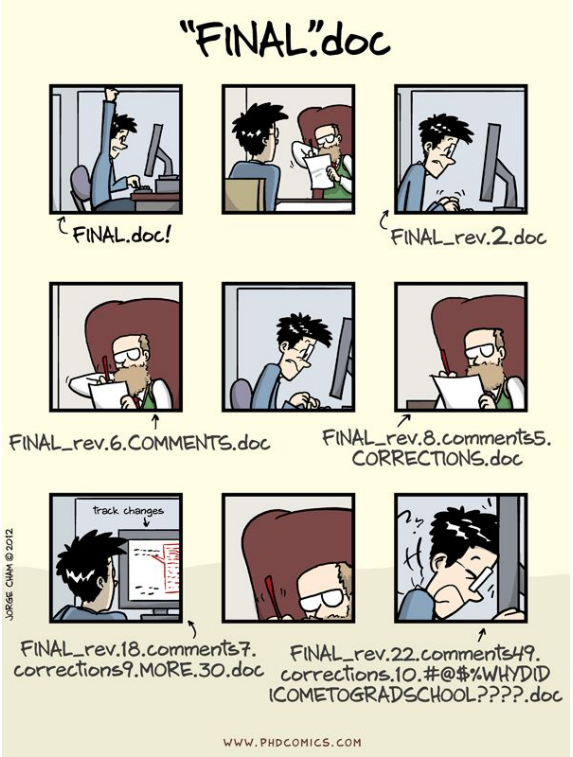
Data on crime is obtained from International Crime Data collected between 2015-2018 and is publicly available. Data on happiness is collected from the Survey of International Happiness.

Contributors:

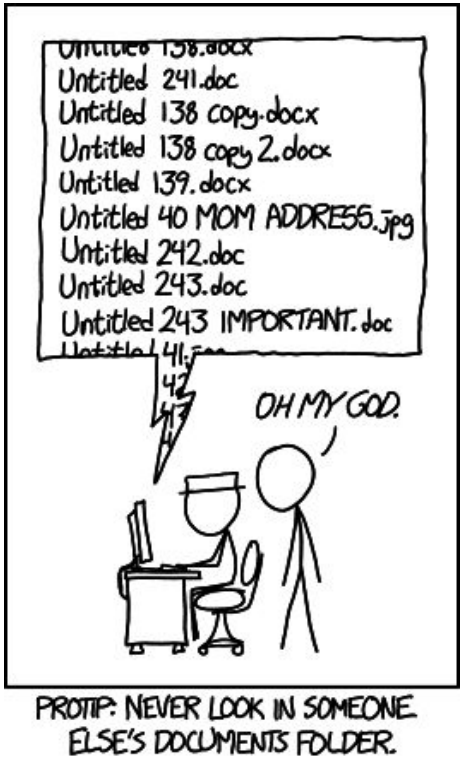
- Jane Everyday Doe, jane.everyday.doe@gmail.com
- John Everyday Doe, john.everyday.doe@gmail.com

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

Just no



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



https://xkcd.com/1459/

key principles of file naming for data science projects:

- Machine readable
- Human readable
- Be nicely ordered

Source: Jenny Bryan

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

2018_jan_sales_cust001_prod001.md
2017_mar_sales_cust001_prod001.md
2016_may_sales_cust001_prod008.md
2017_jan_sales_cust120_prod007.md
2015_oct_sales_cust034_prod001.md
2015_oct_sales_cust034_prod002.md

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

Which one is better?

analysis.R
or
2017-exploratory_analysis_crime.R?

Which one is better?

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

Structure of a filename

processed_pvalue_data_from_pubmed_oct24.rData

What did I do to this data

processed_pvalue_data_from_pubmed_oct24.rData

What kind of data is this?

processed_pvalue_data_from_pubmed_oct24.rData

Where did it come from?

processed_pvalue_data_from_pubmed_oct24.rData

When did I get it?

processed_pvalue_data_from_pubmed_oct24.rData

Underscores/slashes not dots/whitespace

processed_pvalue_data_from_pubmed_oct24.rData

Consistency is the main rule

processed_pvalue_data_from_pubmed_oct24.rData
raw_pvalue_data_from_pubmed_oct24.rData

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)

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

The R Project for Statistical Computing

[Home]

Download

CRAN

R Project

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

R Foundation

Foundation
Board
Members
Donors
Donate

Getting Started

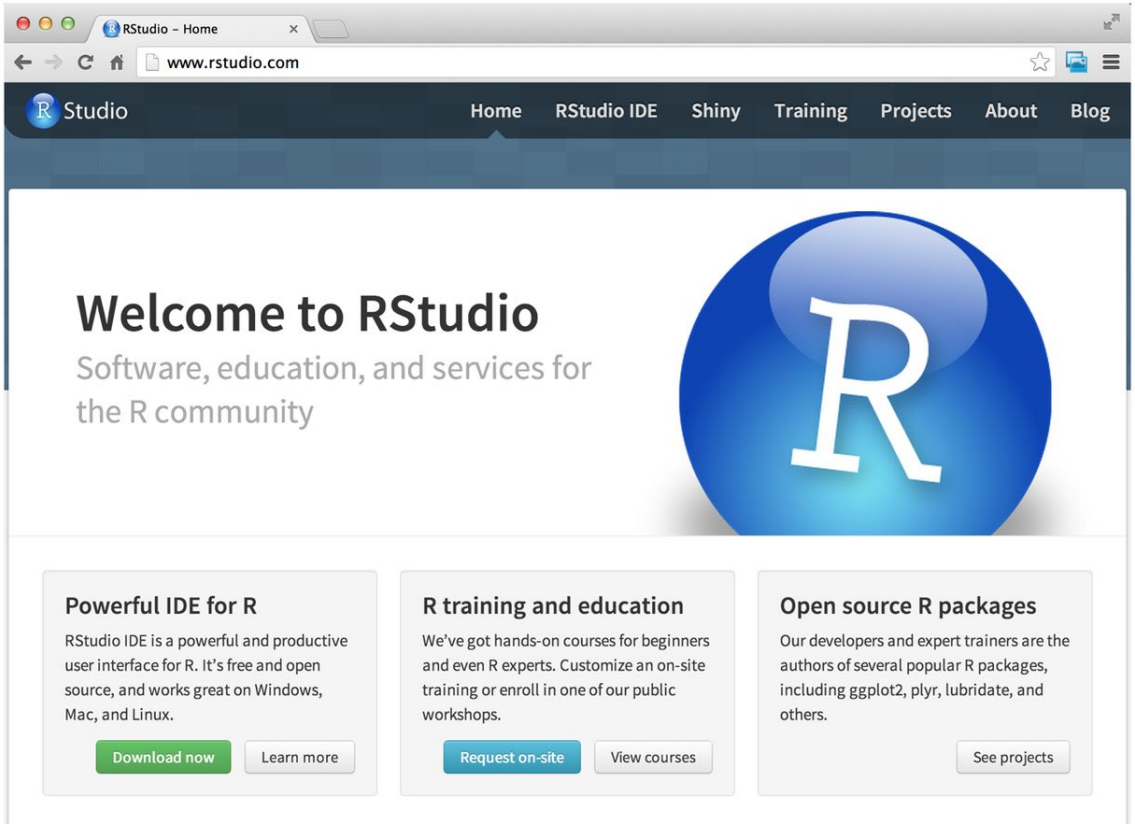
R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

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

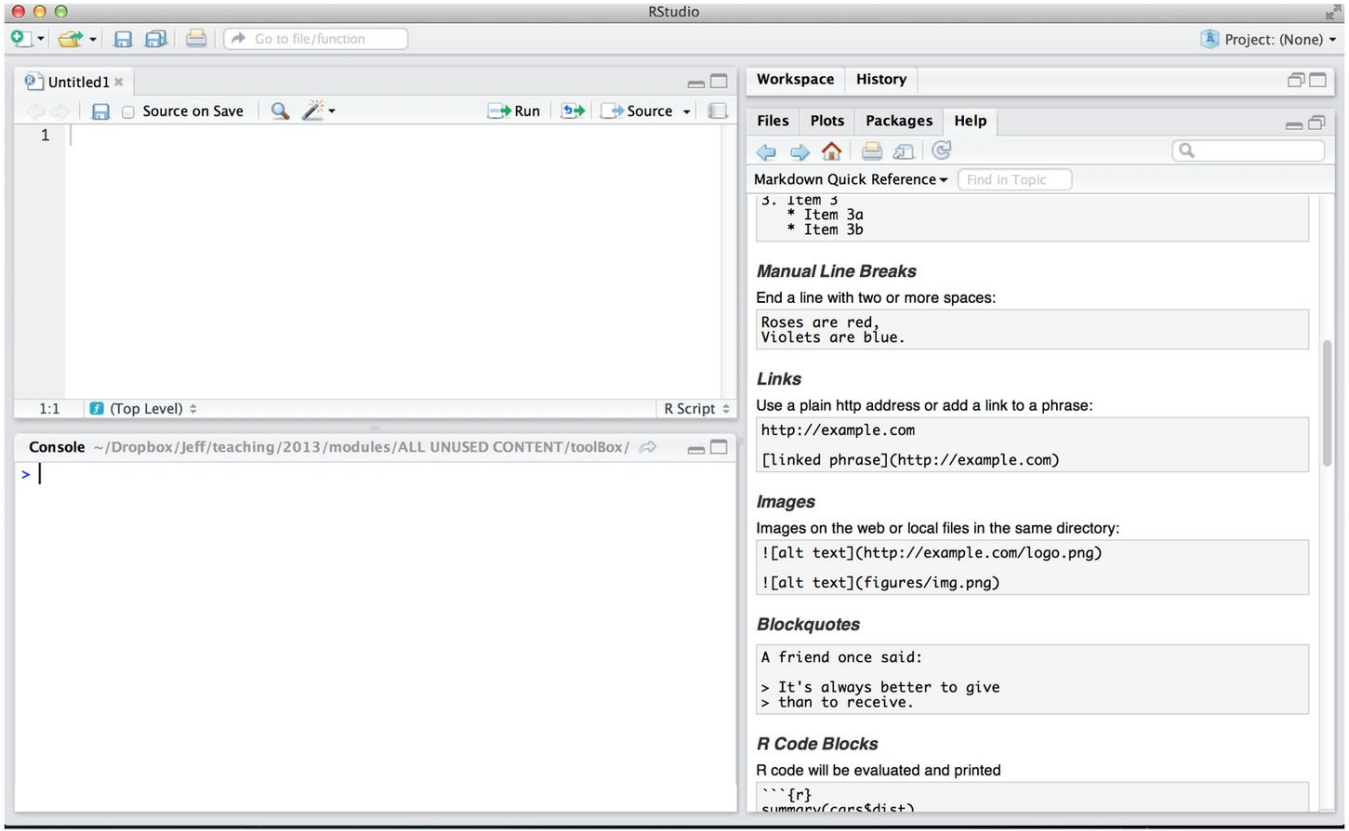
News

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

www.r-project.org



https://www.rstudio.com/



https://www.rstudio.com/

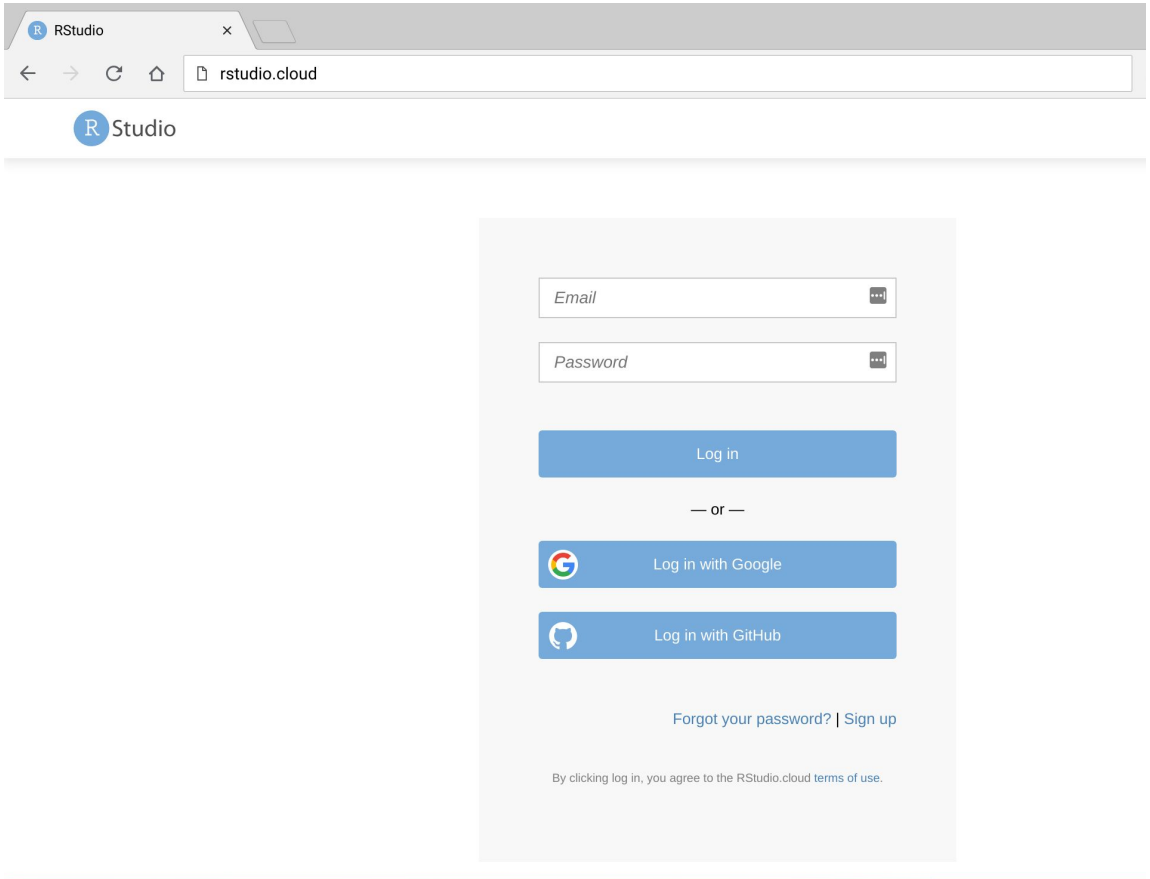
Some useful commands

Cmd + Enter

Ctrl + Enter

Ctrl + 1

Ctrl + 2



https://www.rstudio.cloud

rstudio.cloud tour

https://bit.ly/30CuRWX

Installing R Locally (For later)

http://stat545.com/block000_r-rstudio-install.html

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

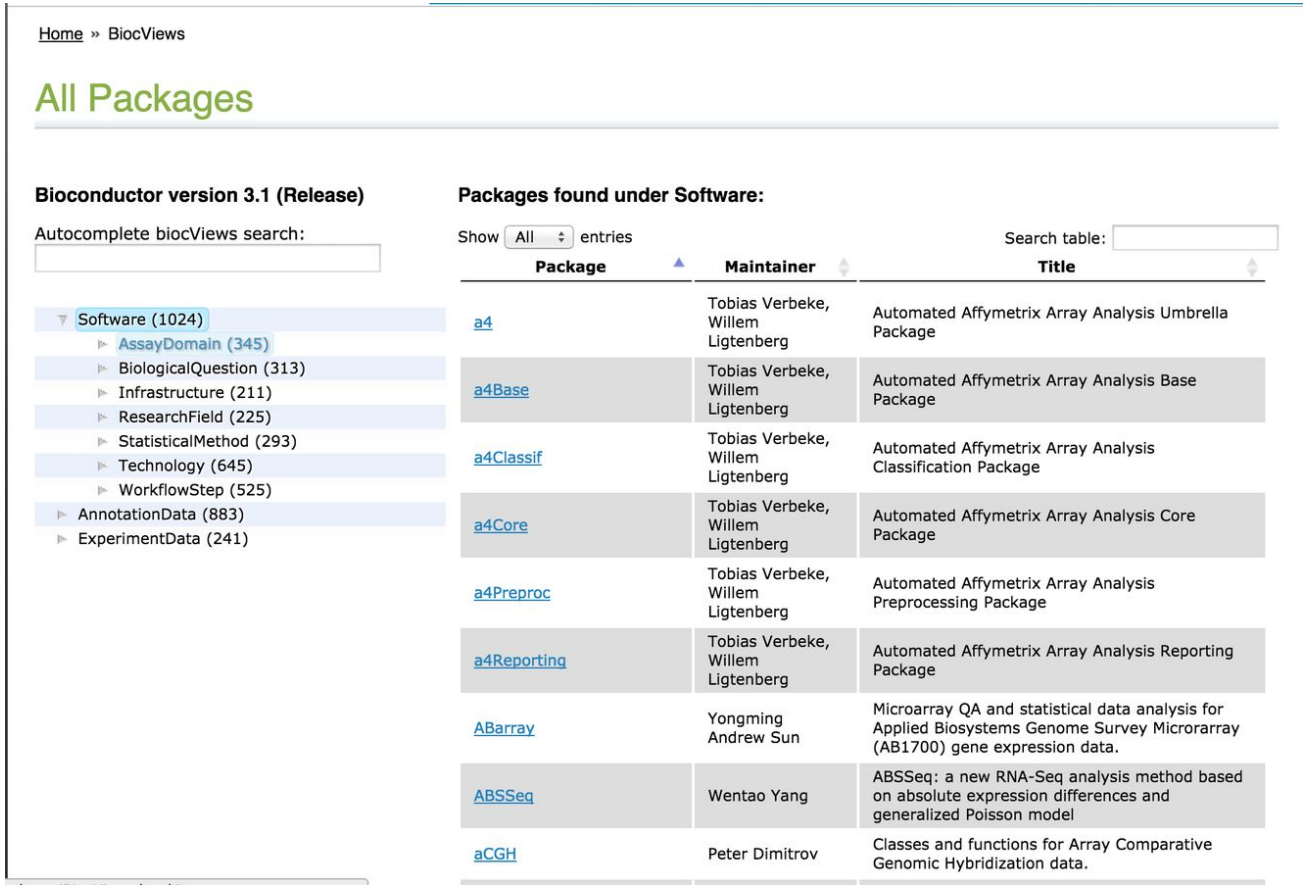
Available CRAN Packages By Name

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	A3: Accurate, Adaptable, and Accessible Error Metrics for Predictive Models
abbyyR	Access to Abbyy Optical Character Recognition (OCR) API
abc	Tools for Approximate Bayesian Computation (ABC)
ABCanalysis	Computed ABC Analysis
abc.data	Data Only: Tools for Approximate Bayesian Computation (ABC)
abcdeFBA	ABCDE_FBA: A-Biologist-Can-Do-Everything of Flux Balance Analysis with this package
ABCoptim	Implementation of Artificial Bee Colony (ABC) Optimization
abctools	Tools for ABC Analyses
abd	The Analysis of Biological Data
abf2	Load Gap-Free Axon ABF2 Files
abind	Combine Multidimensional Arrays
abn	Data Modelling with Additive Bayesian Networks
abundant	Abundant regression and high-dimensional principal fitted components
acc	A Package to Processes Accelerometer Data
accelerometry	Functions for Processing Minute-to-Minute Accelerometer Data
AcceptanceSampling	Creation and evaluation of Acceptance Sampling Plans
ACCLMA	ACC & LMA Graph Plotting
accrual	Bayesian Accrual Prediction
accrued	Data Quality Visualization Tools for Partially Accruing Data
ACD	Categorical data analysis with complete or missing responses
acepack	ace() and avas() for selecting regression transformations

https://cran.rstudio.com/

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



Home » BiocViews

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

http://bioconductor.org/

www.bioconductor.org

Downloads

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 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>

...uctor 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

www.bioconductor.org

Responsiveness

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 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>

ows

- 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

www.bioconductor.org

Still runs

sva

available all platforms | downloads top | 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 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>

kflows »

non Bioconductor workflows
de:

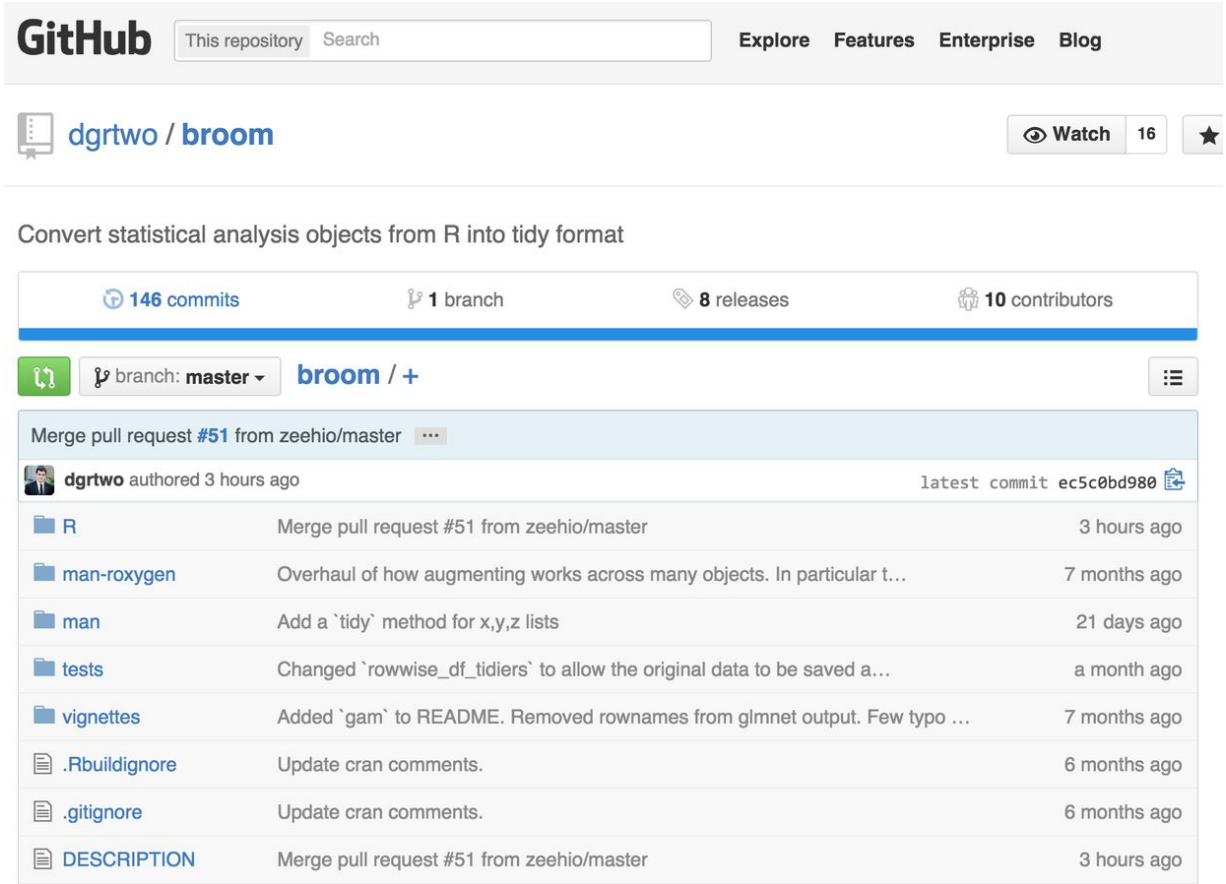
- gonucleotide Arrays
- High-throughput Sequencing
- Counting Reads for Differential Expression (parathyroideSE vignette)
- Annotation
- Annotating Variants
- Annotating Ranges
- Flow Cytometry and other assays
- Candidate Binding Sites for Known Transcription Factors
- Cloud-enabled cis-eQTL search 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

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



https://github.com/dgrtwo/broom

www.github.com

jtleek / sva-devel

Unwatch 6 Star 4 Fork 7

Description

Short description of this repository

Other people like it

26 commits 1 branch 0 releases 4 contributors

branch: master sva-devel / +

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

Code

Issues 0

Pull requests 0

Wiki

Pulse

Graphs

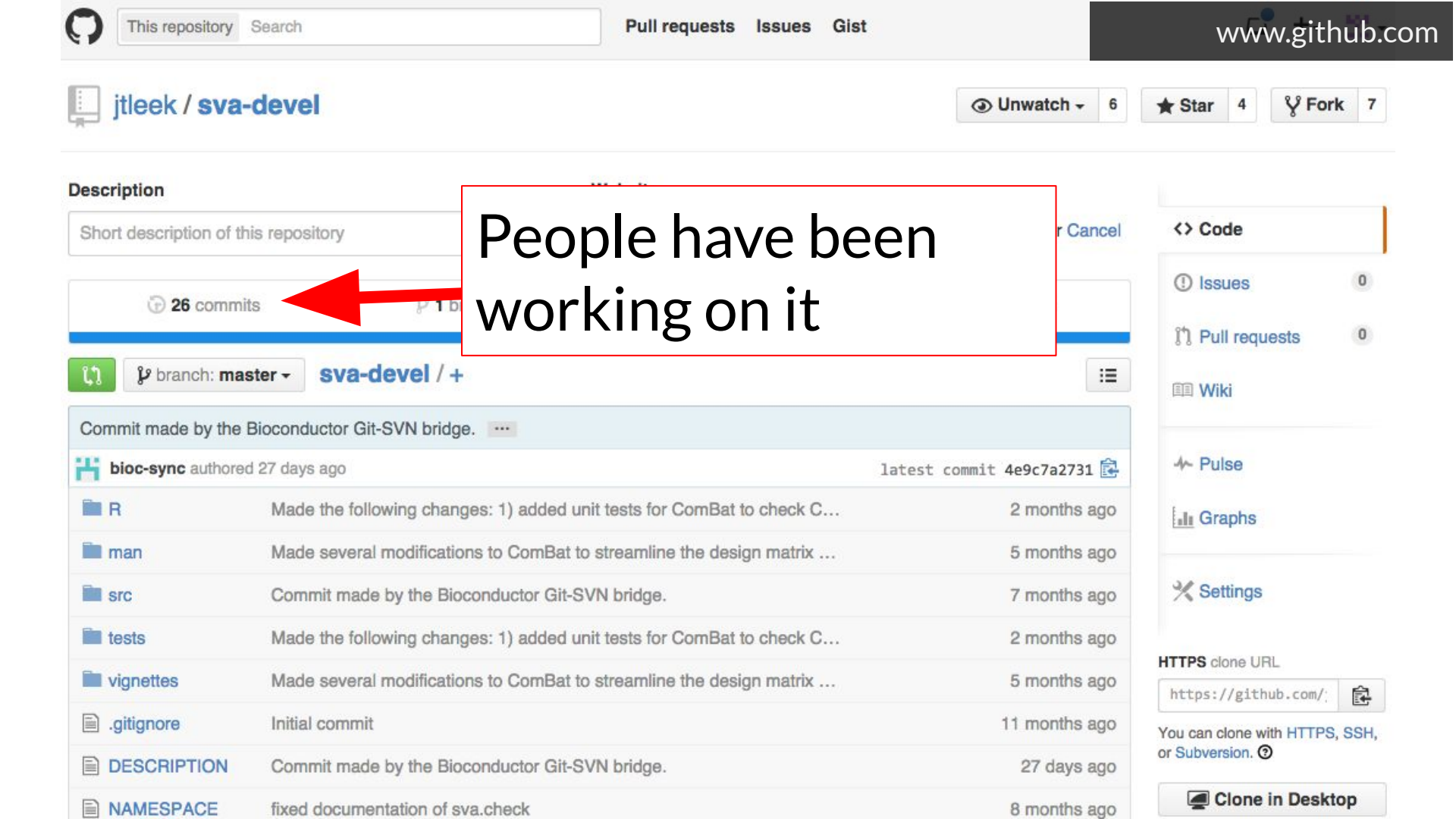
Settings

HTTPS clone URL

https://github.com/

You can clone with HTTPS, SSH, or Subversion.

Clone in Desktop


This repository
Search
Pull requests
Issues
Gist
www.github.com
jtleek / sva-devel
Unwatch 6
Star 4
Fork 7
Description
Short description of this repository
Cancel
26 commits
People have been working on it
branch: master
sva-devel / +
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
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

```
library(devtools)
install_github("dgrtwo/broom")
```

Average trustworthiness

R package installation

https://bit.ly/1CdylSm