

Big Data Wrangling in R

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

Preliminaries

Course Info

Course name	Accessing Biomedical Data
Instructors	Andrew Jaffe and Jeff Leek
TAs	Jeremy Roth
Course website	<u>http://sisbid.github.io/Module1/</u>
Goals	Teach you how to get and clean data
Pre-reqs	Hopefully some R programming

Course Info

Course name	Big Data Wrangling in R
Instructors	Andrew Jaffe and Jeff Leek
TAs	Jeremy Roth
Course website	<u>http://sisbid.github.io/Module1/</u>
Goals	Teach you how to get and clean data
Pre-reqs	Hopefully some R programming

Course Info

Course name	Not borking data processing :)
Instructors	Andrew Jaffe and Jeff Leek
TAs	Jeremy Roth
Course website	<u>http://sisbid.github.io/Module1/</u>
Goals	Teach you how to get and clean data
Pre-reqs	Hopefully some R programming

About us



How many people feel about statistics

Abstract

Formula display: **MathJax** [?](#)

Background

Many groups, including our own, have proposed the use of DNA methylation profiles as biomarkers for various disease states. While much research has been done identifying DNA methylation signatures in cancer vs. normal etc., we still lack sufficient knowledge of the role that differential methylation plays during normal cellular differentiation and tissue specification. We also need thorough, genome level studies to determine the meaning of methylation of individual CpG dinucleotides in terms of gene expression.

Results

In this study, we have used (insert statistical method here) to compile unique DNA methylation signatures from normal human heart, lung, and kidney using the Illumina Infinium 27 K methylation arrays and compared those to gene expression by RNA sequencing. We have identified unique signatures of global DNA methylation for human heart, kidney and liver, and showed that DNA methylation data can be used to correctly classify various tissues. It indicates that DNA methylation reflects tissue specificity and may play an important role in tissue differentiation. The integrative analysis of methylation and RNA-Seq data showed that gene methylation and its transcriptional levels were comprehensively correlated. The location of methylation markers in terms of distance to transcription start site and CpG island showed no effects on the regulation of gene expression by DNA methylation in normal tissues.

ORGANOMETALLICS



[Home](#) | [Browse the Journal](#) | [Articles ASAP](#) | [Current Issue](#) | [Multimedia](#) | [Submission & Review](#)

Article

[Prev.](#)

Synthesis, Structure, and Catalytic Studies of Palladium and Platinum Bis-Sulfoxide Complexes

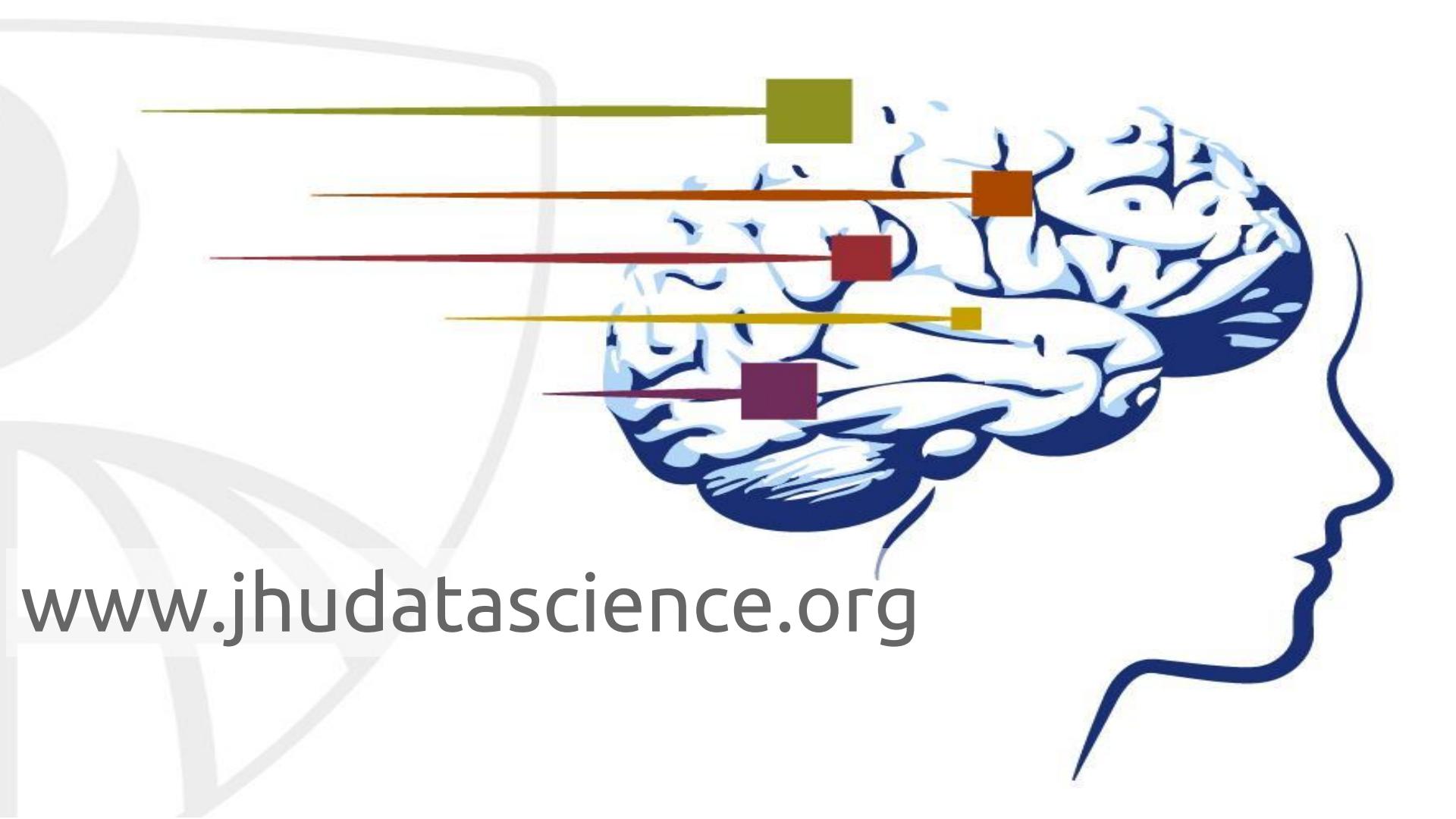
Emma, please insert NMR data here! where are they? and for this compound, just make up an elemental analysis...



How we* feel about statistics



@simplystats
<http://simplystatistics.org>



www.jhudatascience.org



What are your
colleagues reading in

Biostatistics?

READ THIS JOURNAL



[View Current Issue \(Volume 16 Issue 3 July 2015\)](#)

[Advance Access](#)

[Browse the Archive](#)

Among the important scientific developments of the 20th century is the explosive growth in statistical reasoning and methods for application to studies of human health. Examples include developments in likelihood methods for inference, epidemiologic statistics, clinical trials, survival analysis, and statistical genetics. Substantive problems in public health and biomedical research have fueled the development of statistical methods, which in turn have improved our ability to draw valid inferences from data. The objective of *Biostatistics* is to advance statistical science and its application to problems of human health and disease, with the ultimate goal of advancing the public's health.

The ten highest cited articles from Biometrika and Biostatistics from 2013 are free to

read online. [Read now.](#)

LATEST ARTICLES

[Double robust and efficient estimation of a prognostic model for events in the presence of dependent censoring](#)
Schnitzer, M. E., Lok, J. J., Bosch, R. J.

THE JOURNAL

[About this journal](#)

[Rights & Permissions](#)

[Dispatch date of the next issue](#)

[We are mobile – find out more](#)

[Journals Career Network](#)

Impact factor: 2.649

5-Yr impact factor: 2.853

Co-Editors

Geert Molenberghs

Anastasios Tsiatis

[View full editorial board](#)

jtleek

Find me online @jtleek,
@simplystats, Simply
Statistics, and Github.

[Home](#)

[Alumni](#)

[Books](#)

[Data](#)

[Jobs](#)

[Papers](#)

[People](#)

[Software](#)

[Talks](#)

[Teaching](#)

@jtleek

<http://www.jtleek.com>

I defined data science



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item

Article Talk

Read Edit View history Search

Create account Log in

Data science

From Wikipedia, the free encyclopedia



This article's tone or style may not reflect the encyclopedic tone used on Wikipedia. See Wikipedia's guide to writing better articles for suggestions. (February 2014)

Data Science is the extraction of knowledge from large volumes of **data** that are structured or unstructured,^{[1][2]} which is a continuation of the field **data mining** and **predictive analytics**, also known as **knowledge discovery and data mining** (KDD). "Unstructured data" can include emails, videos, photos, social media, and other user-generated content. Data science often requires sorting through a great amount of information and writing algorithms to extract insights from this data.

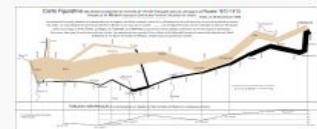


This is me

Contents [hide]

- 1 Overview
- 2 History
- 3 Domain specific interests
- 4 Criticism
- 5 Research areas
 - 5.1 Security Data Science
 - 5.2 Clinical data science
 - 5.3 Genomic data science
 - 5.4 Agriculture
 - 5.5 Retail
- 6 Further reading

Part of a series on Statistics
Data visualization



- Major dimensions [hide]
Exploratory data analysis • Information design
Interactive data visualization
Descriptive statistics • Inferential statistics
Statistical graphics • Plot
Data analysis • Infographic
Data science
- Thought leaders [hide]
John W. Tukey • Edward Tufte
- Information graphic types [hide]
Line chart • Bar chart

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.

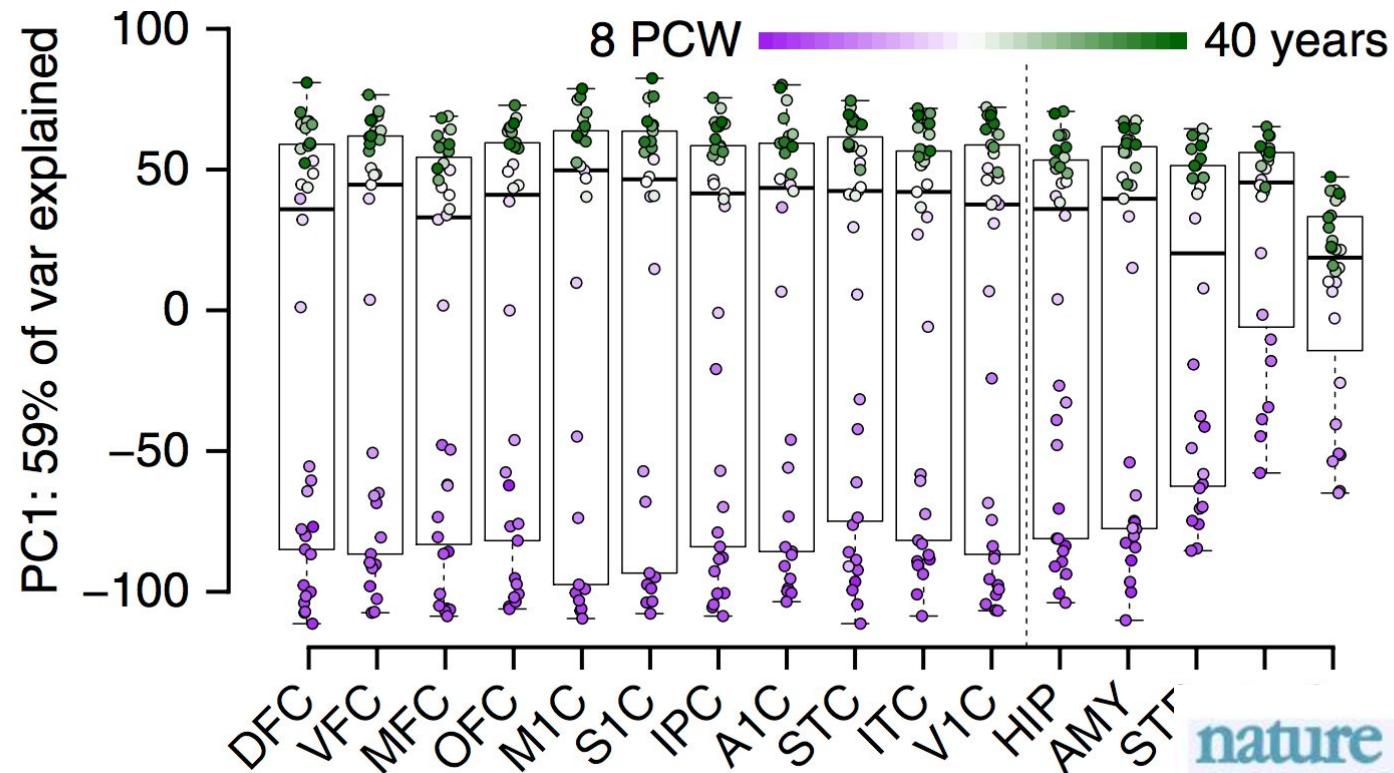
Contact

Email:

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

Developmental regulation of human cortex transcription and its clinical relevance at single base resolution

Andrew E Jaffe¹⁻³, Jooheon Shin¹, Leonardo Collado-Torres^{1,2}, Jeffrey T Leek^{2,4}, Ran Tao¹, Chao Li¹, Yuan Gao¹,
Yankai Jia¹, Brady J Maher^{1,5,6}, Thomas M Hyde^{1,5-8}, Joel E Kleinman^{1,9} & Daniel R Weinberger^{1,4-7,9}



Now you

Introduce yourself to your neighbor

Today

Morning

Background

R/Rstudio

Version Control

Reproducible Research

Afternoon

Data I/O

Intro to Bioconductor

About us

<https://goo.gl/3jq2nE>

Skill (self) Assessment

<https://goo.gl/TX5Dey>

Motivation

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

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.

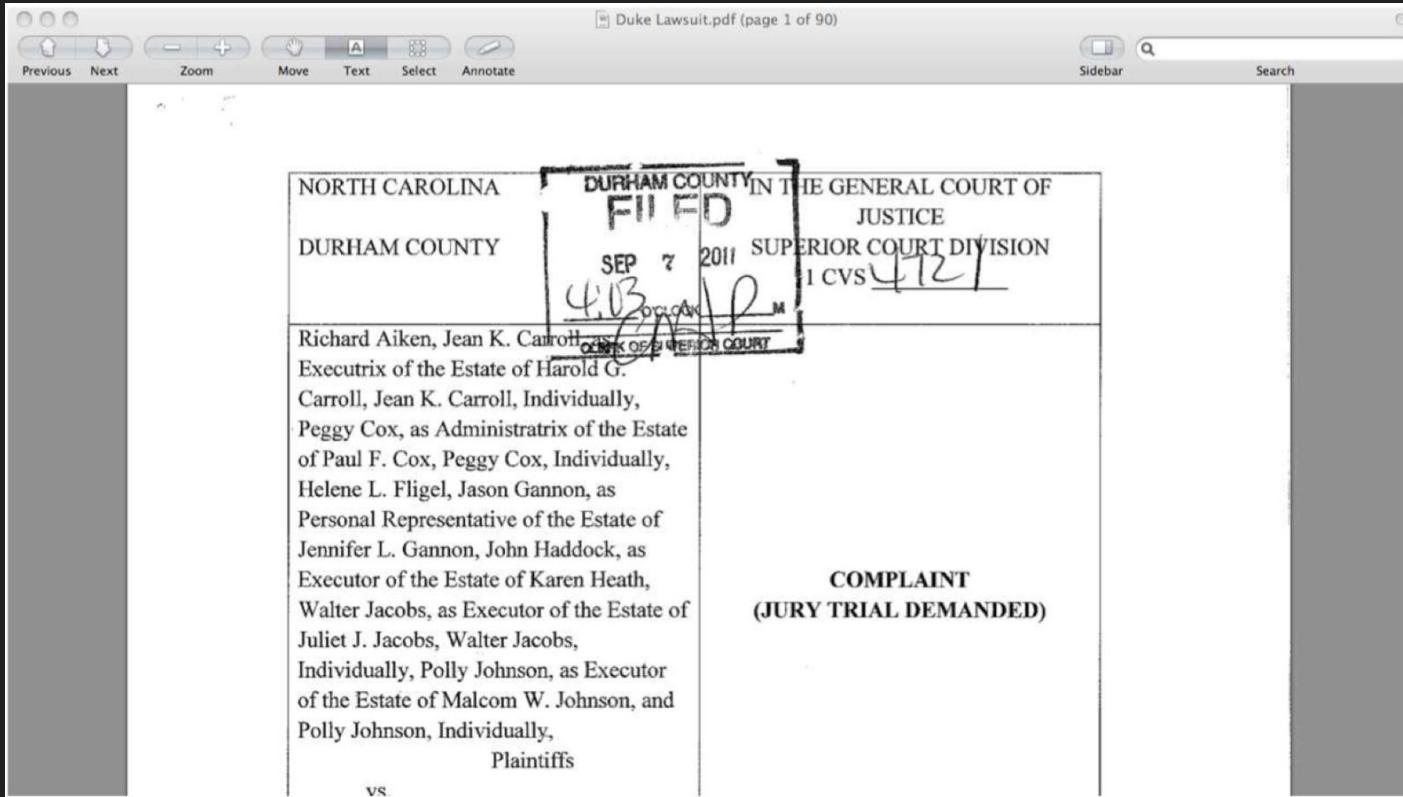
JANUARY 2010 Duke restarts clinical trials.

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.



http://dig.abclocal.go.com/wtvd/docs/Duke_lawsuit_090811.pdf

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

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>

It is not the critic who counts

"It is not the critic who counts: not the man who points out how the strong man stumbles or where the doer of deeds could have done better. The credit belongs to the man who is actually in the arena, whose face is marred by dust and sweat and blood, who strives valiantly, who errs and comes up short again and again, because there is no effort without error or shortcoming, but who knows the great enthusiasms, the great devotions, who spends himself for a worthy cause; who, at the best, knows, in the end, the triumph of high achievement, and who, at the worst, if he fails, at least he fails while daring greatly, so that his place shall never be with those cold and timid souls who knew neither victory nor defeat."



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 Myer, Mathematics Educator](#)



Why this
class?

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:

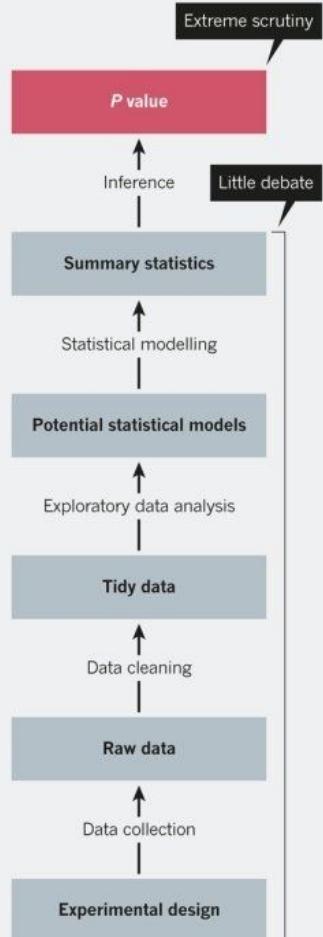
Min	1Q	Median	3Q	Max
-1.6173	-1.4259	0.7941	0.9478	1.1431

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

DATA PIPELINE

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



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

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

Joseph P. Simmons¹, Leif D. Nelson², and Uri Simonsohn¹

¹The Wharton School, University of Pennsylvania, and ²Haas School of Business, University of California, Berkeley

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

<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*

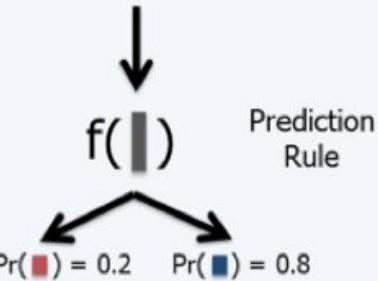
Andrew Gelman[†] and Eric Loken[‡]

14 Nov 2013

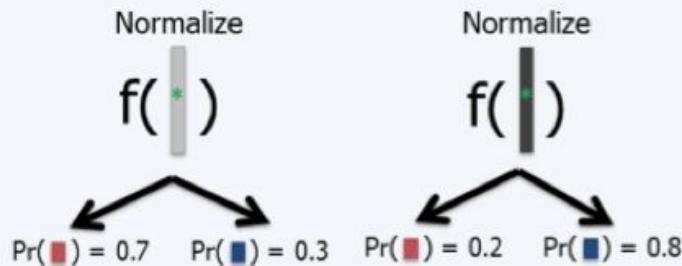
“I thought of a labyrinth of labyrinths, of one sinuous spreading labyrinth that would encompass the past and the future . . . I felt myself to be, for an unknown period of time, an abstract perceiver of the world.” — Borges (1941)

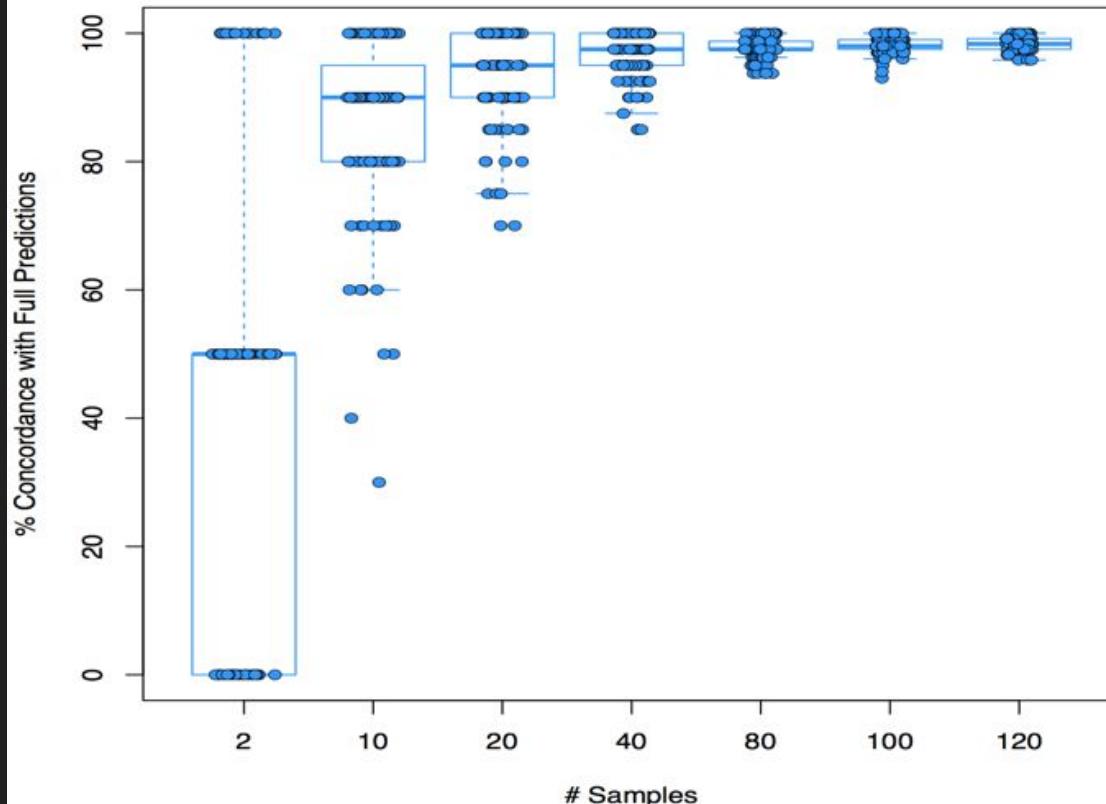
http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

a) Learning Stage



b) Application Stage





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

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

What it actually looks like

<http://healthdesignchallenge.com/>

```
@HWI-EAS121:4:100:1783:550#0/1
CGTTACGAGATCGAACAGCGGTTCAGCAGGAATGCCGAGACGGATCTCGTATGCCGTCTGCGTGACAAGACAGGGG
+HWI-EAS121:4:100:1783:550#0/1
aaaaaa`b_aa`aa`YaX]aZ`aZM^Z]YRa]YSG[ [ZREQLHESDHNDDHNMEEDMPENITKFLFEEDDDHEJQMEDDD
@HWI-EAS121:4:100:1783:1611#0/1
GGGTGGGCATTTCCACTCGCAGTATGGGTTGCCGCACGACAGGCAGCGGTCAAGCCTGCGCTTGGCCTGGCCTTCGGAAA
+HWI-EAS121:4:100:1783:1611#0/1
a``^\\_`_``^`a``^a_`^_]a_]`a_____`_``^`]X]_]XTV_\\_]NX_XVX]_]_TTTG[VTHPN]VFDZ
@HWI-EAS121:4:100:1783:322#0/1
CGTTTATTTTTGAATATGTCTTATCTAACGGTTATTTAGATGTTGGTCTTATTCTAACGGTCATATATTTCTA
+HWI-EAS121:4:100:1783:322#0/1
abaa``aaaaabbbaabbbbbbb`bbbb_bbbbbbbb`bbbaV`_a``]``aT]a__V\\]]`a`]a_abbaV_
@HWI-EAS121:4:100:1783:1394#0/1
GGGTCTTATTGGTCTGGTATCCCCCATATTCTCCGGTTGTGGTTAACCGATCATCGCGCATTACTCCGGCTGC
+HWI-EAS121:4:100:1783:1394#0/1
```[aa\b``[ ]aabbb][`a_abbb`a``bbbbbabaaaab_VZa_``bab_X`[a\HV[_]_[`_X\T_VQQ
@HWI-EAS121:4:100:1783:207#0/1
```

# What it actually looks like

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

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

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

**Example Request**

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

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

# What it actually looks like

## ALLERGIES

Last Updated: 01 Dec 2011 @ 0851

Allergy Name: TRIMETHOPRIM  
Location: DAYT29  
Date Entered: 09 Mar 2011  
Reaction:

Allergy Type: DRUG  
A Drug Class: ANTI-INFECTIVES, OTHER  
Observed/Historical: HISTORICAL  
Comments: The reaction to this allergy was MILD (NO SQUELAE)

Allergy Name: TRAMADOL  
Location: DAYT29  
Reaction:

## MEDICATION HISTORY

Last Updated: 11 Apr 2011 @ 1737

Medication: AMLODIPIINE BESYLATE 10MG TAB  
Instructions: TAKE ONE TABLET BY MOUTH TAKE ON GRAPEFRUIT JUICE--  
Status: Active  
Refills Remaining: 3  
Last Filled On: 28 Aug 2010  
Initially Ordered On: 13 Aug 2010  
Quantity: 45  
Days Supply: 90  
Pharmacy: DAYTON  
Prescription Number: 2718953



**Jenny Bryan** @JennyBryan · Apr 20

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

University and supported by the National Research Council  
with Rockefeller Foundation funds

SEXUAL  
BEHAVIOR

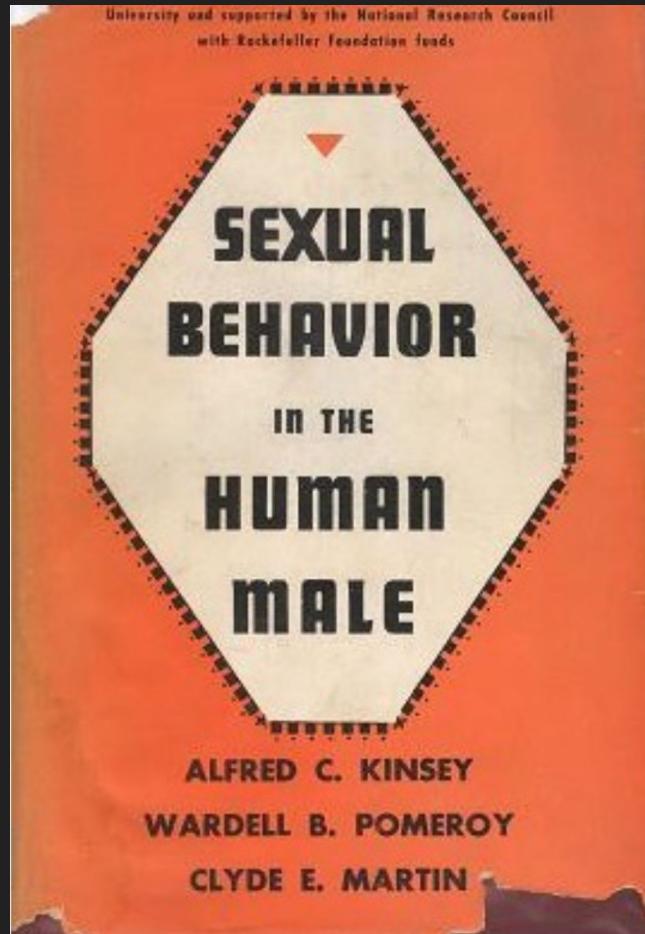
IN THE

HUMAN  
MALE

ALFRED C. KINSEY

WARDELL B. POMEROY

CLYDE E. MARTIN



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

Based on surveys made by members of the Staff of Indiana University and supported by the National Research Council with Rockefeller Foundation funds

SPREADSHEET  
**BEHAVIOR**  
IN THE  
**HUMAN**  
**MALE**

ALFRED C. KINSEY

WARDELL B. POMEROY

CLYDE E. MARTIN

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

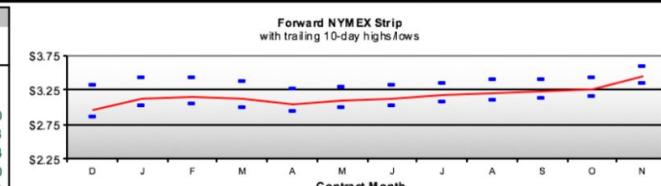
## ENA - West Gas Contacts

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
2.960	0.090
3.088	0.083
3.166	0.084
3.651	0.090
3.165	0.084



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)

AECO / NIT			
Fixed Price		Basis	
BID	OFFER	BID	OFFER
2.376	2.396		
2.398	2.418		
2.552	2.572	(0.408)	(0.388)
2.616	2.636	(0.472)	(0.452)
2.661	2.681	(0.505)	(0.485)
3.216	3.236	(0.435)	(0.415)
2.676	2.696	(0.488)	(0.468)

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)

IF PEPL TX-OK			
Fixed Price		Basis	
BID	OFFER	BID	OFFER
2.530	2.550		
2.530	2.550		
2.828	2.848	(0.133)	(0.113)
2.958	2.978	(0.130)	(0.110)
3.046	3.066	(0.120)	(0.100)
3.531	3.551	(0.120)	(0.100)
3.041	3.061	(0.123)	(0.103)

# #otherpeoplesdata

Home Notifications Messages  #otherpeoplesdata   

## #otherpeoplesdata

Top | Live | Accounts | Photos | Videos | More options ▾

Who to follow · Refresh · View all

 **Joseph N. Paulson** @dorageh Followed by Hector Corrada ... 

 **RStudio** @rstudio 

 **One R Tip a Day** @RLangTip 

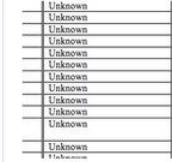
Find friends

Trends · Change

#IndependenceDay  
For At Least 4,000 Immigrants, This Independence Day Has Special

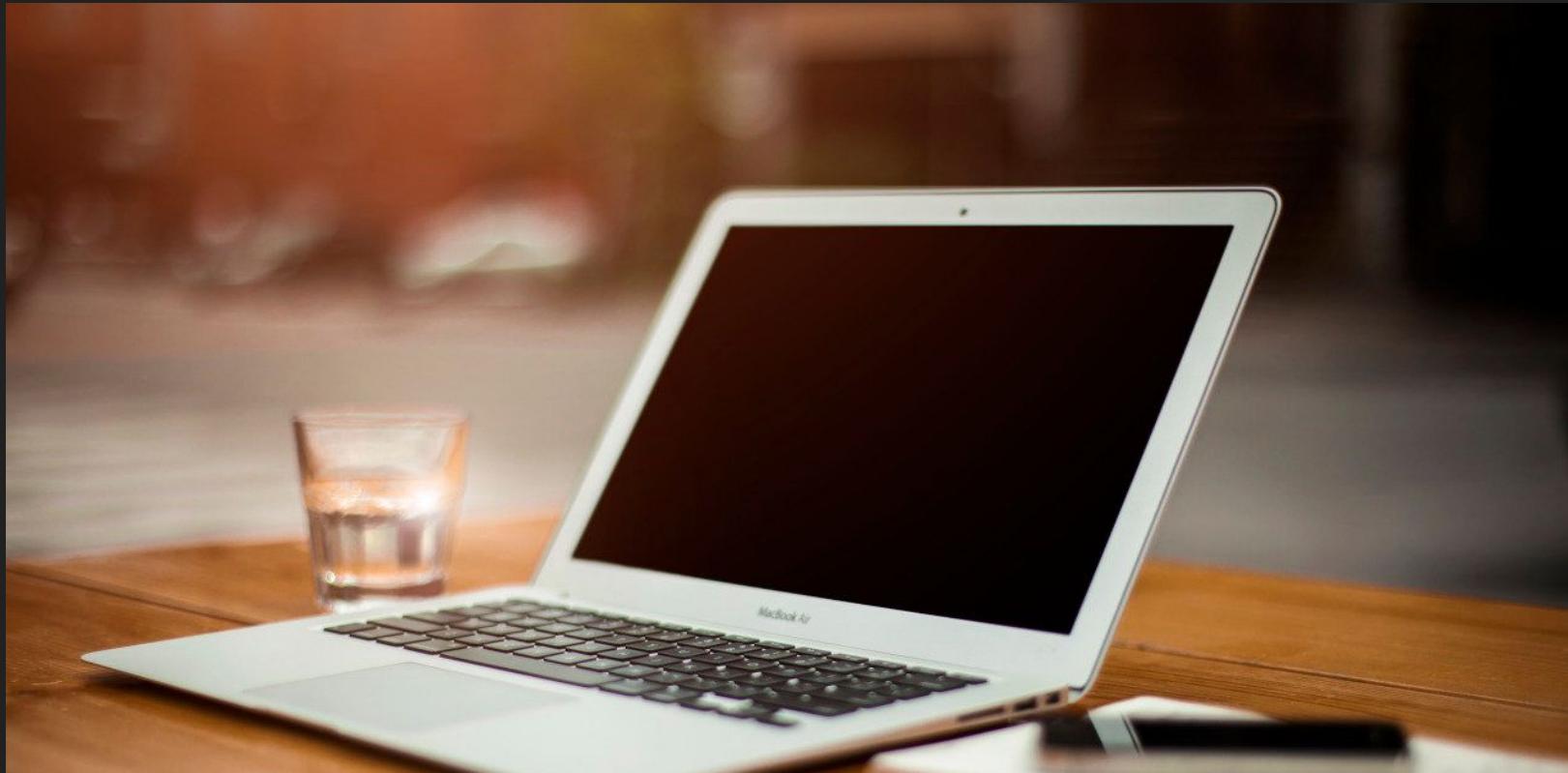
 **Patrick Durusau** @patrickDurusau · Jul 1 RP The challenge of combining 176 x #otherpeoplesdata... #integration #opendata ow.ly/OSFGs 

 **Martin Bentley** @astonsplat · Jun 30 #OtherPeoplesData 

 **Matthew** @MCeeP · Jun 27 Most helpful data column ever 

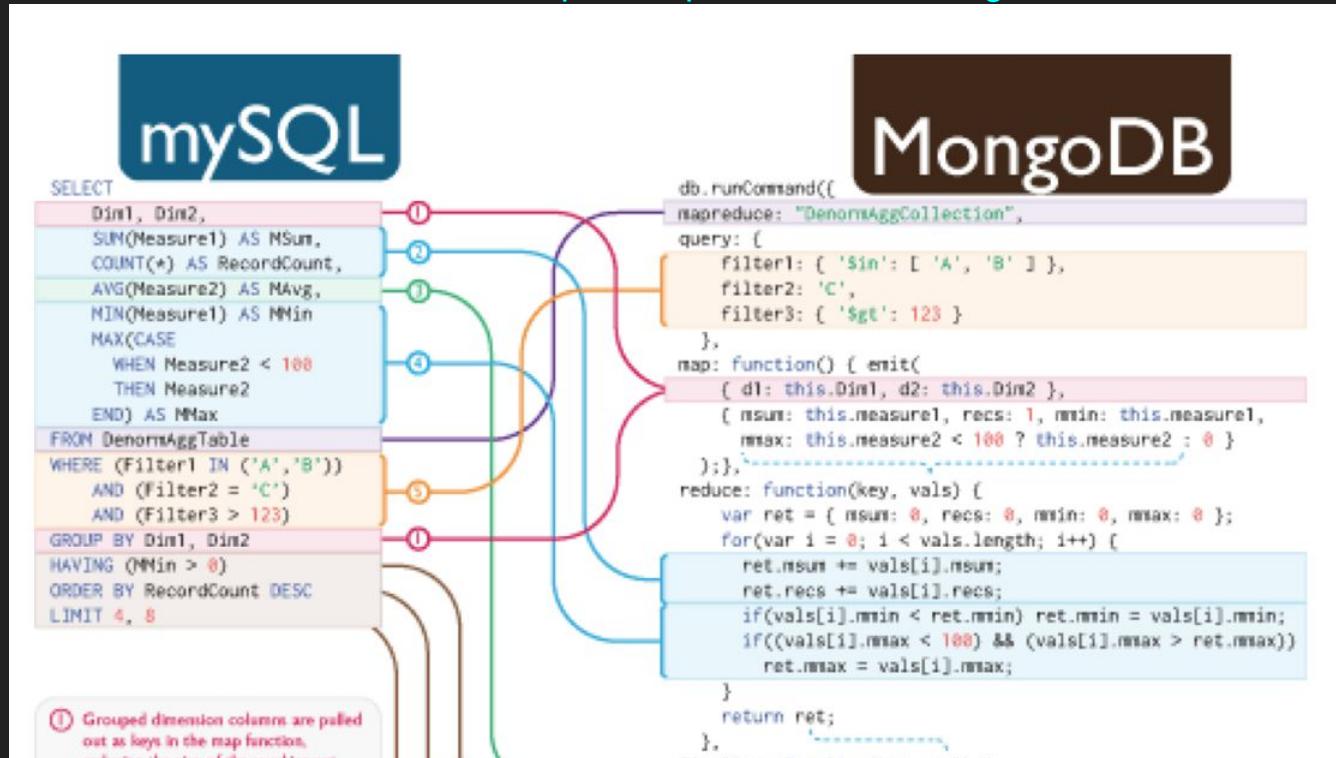
 **Legacy4Life, A Theft** @hieegg · Jun 27 #hieegg hieegg.com #legacyforlife #PAYBACK ME YOUR #fraud #thieves of

# 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 URL [https://dev.twitter.com/docs/api/1/get\(blocks/blocking](https://dev.twitter.com/docs/api/1/get(blocks/blocking) in the address bar. The page header includes the Twitter logo, a search bar, and links for API Health, Blog, Discussions, Documentation, and Sign in. Below the header, there is a note about the cursor parameter and an example value. A section titled "Example Request" shows a GET request to `https://api.twitter.com/1/blocks/blocking.json?cursor=-1&include_entities=true`. A code block below the request shows a JSON response with 18 numbered fields, including user profile information like name, profile picture URL, and background tile.

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

**Example Request**

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

```
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,
```

# Where data actually is

<https://data.baltimorecity.gov/>

The screenshot shows the homepage of the Open Baltimore beta website. The URL in the browser bar is <https://data.baltimorecity.gov/>. The page features a large "OPEN" logo with "BALTIMORE" below it and "beta" above the "O". The background is a collage of city images, binary code, and the City of Baltimore seal. A navigation bar at the top includes links for Home, Residents, Business, Visitors, Government, Office of the Mayor, and Help. At the bottom left, there are "Sign Up" and "Sign In" buttons. A green sidebar on the left contains the text "We Want Your Feedback!" and a message encouraging users to provide suggestions for datasets. On the right side, there is a "Brought to you by" section featuring the City of Baltimore seal.

Open Baltimore / City of Ba x

https://data.baltimorecity.gov

OPEN beta

BALTIMORE

Home Residents Business Visitors Government Office of the Mayor Help

Sign Up Sign In

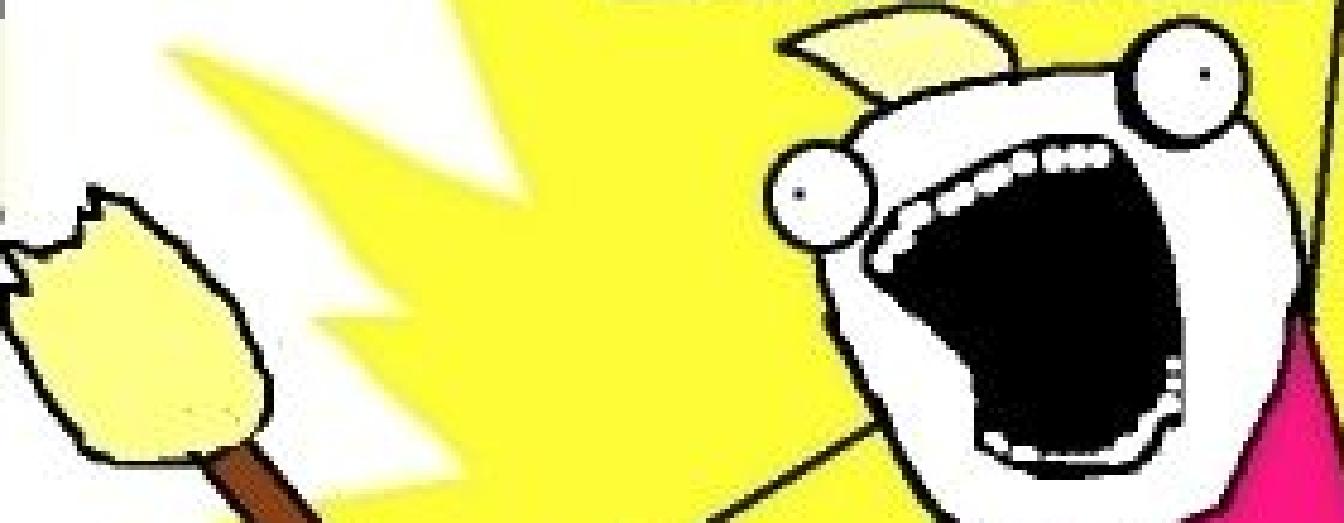
We Want Your Feedback!

Have suggestions for a dataset? Please take a moment and visit the suggestion page at the bottom of this page. Or you can click the feedback tab to the left and join the discussion over at the forums. See you there!

Brought to you by

CITY OF BALTIMORE

# GET ALL THE DATA!!!



# Data brainstorming

<https://goo.gl/r26muZ>

# Raw and processed data

Raw vs. processed, relativity of raw, data description etc.

“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

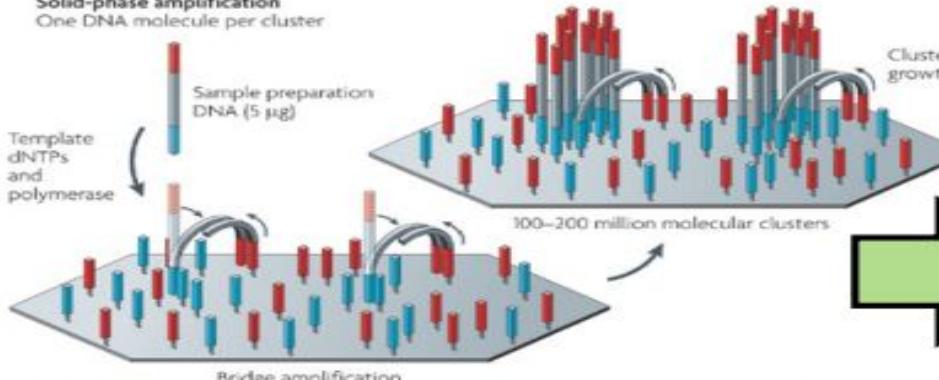
“Data are values of **qualitative** or **quantitative** variables, belonging to a set of items.”

**Qualitative:** Country of origin, sex, treatment

**Quantitative:** Height, weight, blood pressure



Illumina/Solexa  
Solid-phase amplification  
One DNA molecule per cluster



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

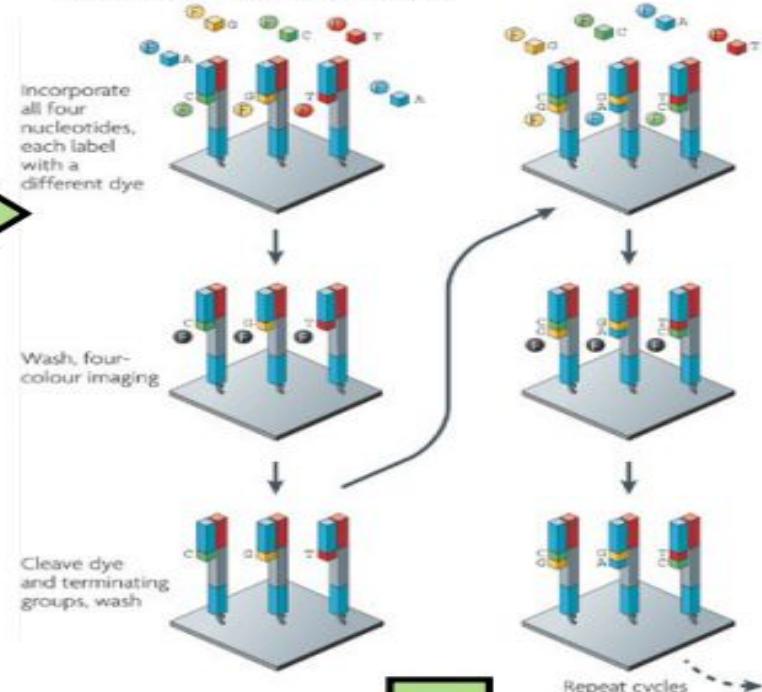
Sequence data from the Solexa platform:

@HWI-EAS146:5:1:1:961#0/1  
TCGGAGGCCAACGAGGCCGCGCGCTGNNNNNNNNNNNNNNNNN  
+  
BBBB>A7B>@BBB8AA=BA=A>>>>>>  
@HWI-EAS146:5:1:1:1595#0/1  
TCAGGAAGCAGGAAGAGACTGTGACGAGCAGNNNNNNNNNNNNNN  
+  
B9B8B<BAA<BAA<BAA<>>>>>>>>  
@HWI-EAS146:5:1:1:1048#0/1  
CTGGACTGCATCCATACCAACACTGTCNAANNNNCNNNNNNNNNN  
+  
A=B7A>>A=A>79>>747>>>>>>>>>  
@HWI-EAS146:5:1:1:1687#0/1  
CTTCTCTCAAGGCCCCAGAACAGCCAANNNNANTNTNNNN  
+  
BBCCCCCCBBCB7C8C=7>>>>=BCCB7  
@HWI-EAS146:5:1:1:1719#0/1  
CACGATCTGGTTTATTGTACCTCCGCTCHNNNNGNTNAAGNNNN  
+  
BCC7=<B=>B8B5=ABA7B6B8B84B87B>>>>>>>  
@HWI-EAS146:5:1:2:947#0/2  
CCCAGGAGAAACCATTTCAAGTCGAGCGNNANCTGANNNN  
+  
B8B5&BTAT>&AB>>B8>B8>>B7>>>>>>  
@HWI-EAS146:5:1:2:1563#0/1



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

Illumina/Solexa — Reversible terminators



b

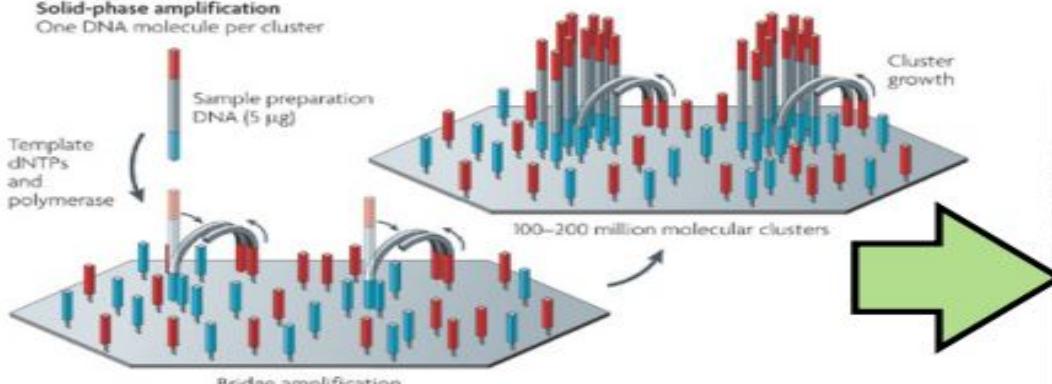


Source: Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010

# Data sharing

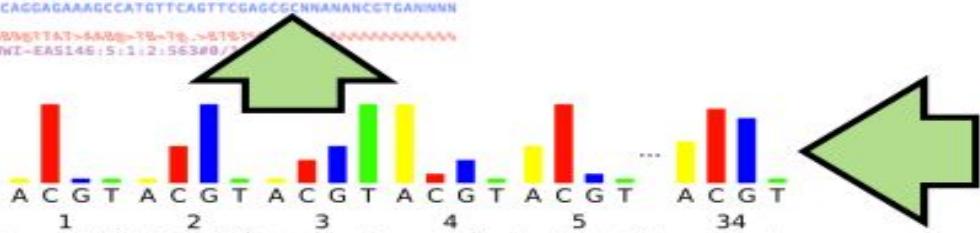
1. The raw data.
2. A tidy data set
3. A code book describing each variable  
and its values in the tidy data set.
4. An explicit and exact recipe you used  
to go from 1 -> 2,3

Illumina/Solexa  
Solid-phase amplification  
One DNA molecule per cluster



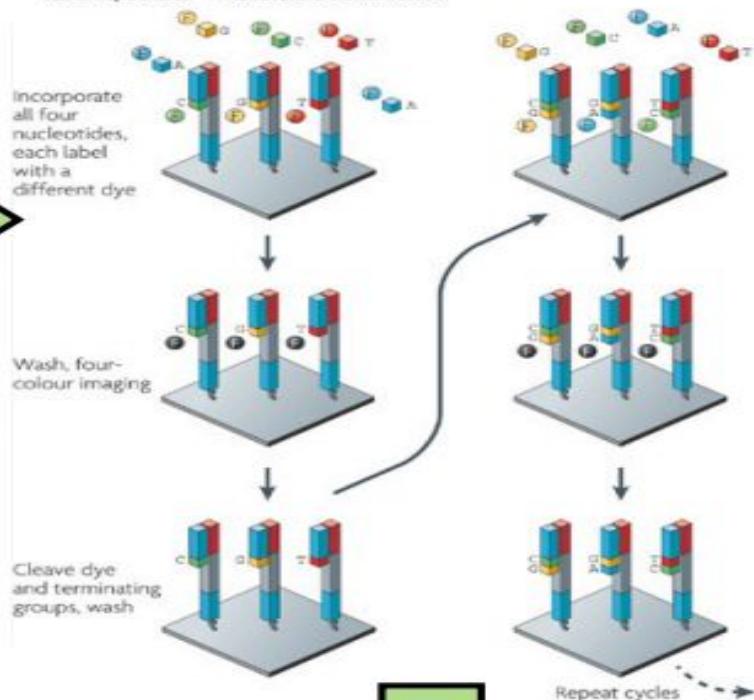
Source: Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010

Sequence data from a single lane:  
@HWI-EAS146:5:1:1:961#0/1  
TCGGAGGCCAACGAGGCCGCGCGCTGNNNNNNNNNNNNNNNNN  
+  
BBBB>A7B9;>BBB8AA=BA=A  
@HWI-EAS146:5:1:1:1595#0/1  
TCAGGAAGCAGGAAGAGACTGTGACGAGCAGNNNNNNNNNNNNNN  
+  
B9B8B<BA=<AB9=>  
@HWI-EAS146:5:1:1:1048#0/1  
CTGGACTGCATCCTACCCACACTCGTCCAANNNNCNNNNNNNNNN  
+  
A=B7A>>A=79>>747>>>>>>  
@HWI-EAS146:5:1:1:1687#0/1  
CTCTCTCAAGGCCCCAGAACAGCCAANNNNNNNNNNNNNNNNNN  
+  
BBCCCCCCBBCB7C8C=7>>>=BCBCB  
@HWI-EAS146:5:1:1:1719#0/1  
CACGATCTGGTTATTGTACCTCCGCTCHNNNNGNTNAAGNNNN  
+  
BCC7=>B=7BB5=ABA7B6BBBB4BB7B  
@HWI-EAS146:5:1:2:947#0/2  
CCCAGGAGAAACCATTTCAAGTCGAGCGNNNNANCTGANNNNN  
+  
BBB7&TAT>&AB>>7B>7B>>7B7B  
@HWI-EAS146:5:1:2:1563#0/2



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

Illumina/Solexa — Reversible terminators



b



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

# Raw data

```
@HWI-EAS121:4:100:1783:550#0/1
CGTTACGAGATCGGAAGAGCGGTTCACGCAGGAATGCCGAGACGGATCTCGTATGCCGTCTGCTCGTGACAAGACAGGGG
+HWI-EAS121:4:100:1783:550#0/1
aaaaaa`b_aa`aa`YaX]aZ`aZM^Z]YRa]YSG[[ZREQLHESDHNDDHNMEEDDM PENITKFLFEEDDDHEJQMEDDD
@HWI-EAS121:4:100:1783:1611#0/1
GGGTGGGCATTCCACTCGCAGTATGGGTTGCCGCACGGCAGCGGT CAGCCTGCGCTTGGCCTGGCCTTC
+HWI-EAS121:4:100:1783:1611#0/1
a``^`_ `` `` `` `` a``^`a``^`_a_]`]\`a____`_ ``]X]_)XTV_\])]NX_XVX]]_TTTG[Y
@HWI-EAS121:4:100:1783:322#0/1
CGTTTATGTTTGAATATGCTTATCTAACGGTTATTTAGATGTTGGTCTTATTCTAACGGTCATGTTCTA
+HWI-EAS121:4:100:1783:322#0/1
abaa``^aaaaabbbaabbbbbbb`bbbbb_bbbbbb`bbbaV^_a``^`]``^`aT]a__V\\11_1``^`bbbaV_
@HWI-EAS121:4:100:1783:1394#0/1
GGGTCTTATTGGCTGGTGATCCCCCATATTCTCCGGTTGTGGTTAACCGATCATGGGCATTAC
+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  
CCCTGGGAGATCGGAAGAGCGGTTCACGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTGCTTGGAA  
+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_[]\Z]  
@HWI-EAS121:4:100:1783:455#0/1  
GGGTAATTCAAGGGACAATGTAATGGCTGCACAAAAAAACATTTCATGTTCCAG  
+HWI-EAS121:4:100:1783:455#0/1  
abb_babbabaabbbbbbbbbbba\`b`\abbabbabbabbabbabbbaabbba
```



Processing
Computing
Summarizing
Deleting



Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II.

<http://www.jstatsoft.org/>

Tidy Data

Hadley Wickham
RStudio

Abstract

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

A tidy data set

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

One variable per column
One observation per row
One table per “kind” of variable
Linking indicators for columns

Decoder.docx

Code book

anything doesn't make sense.

Files:

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

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

B. patient identification number

C. Age at time of CSF collection

D. Gender

E. BMI

F. Ethnicity (mostly Caucasian)

G. Diagnosis: DSM/ICD-10 diagnosis

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

I. Medication: mostly untreated

J. Education more or less than 13 years

K. current smoking status: yes or no



Variable names

Variable descriptions

Variable units

Study design quirks

Recipe

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



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

recipe.docx

Home Layout Document Elements Tables Charts SmartArt Review

Cambria (Body) 15 A A Aa Ab B I U ABC 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|

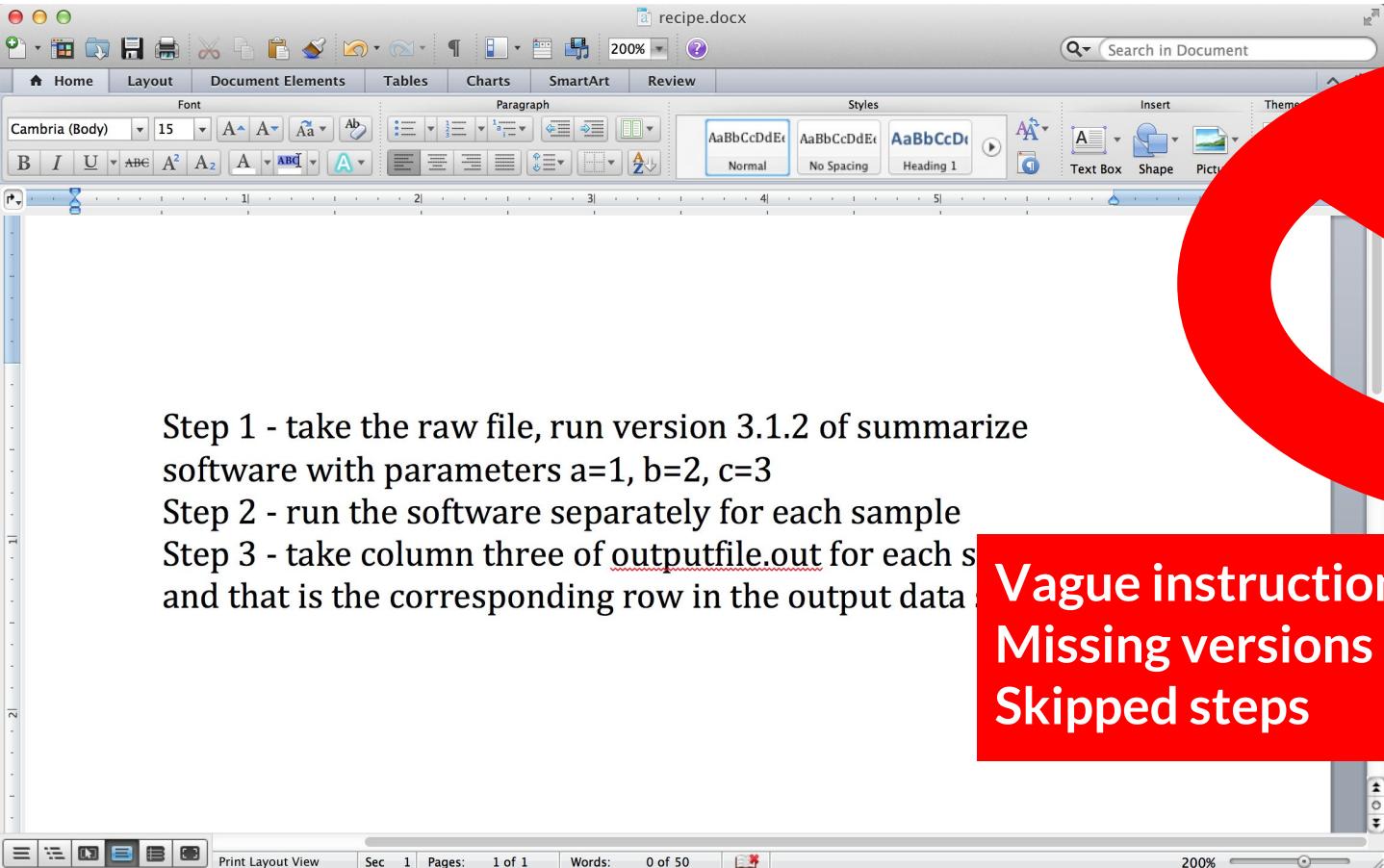
Print Layout View Sec 1 Pages: 1 of 1 Words: 0 of 50 200%

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

Cambria (Body) 15 A A Aa Ab B I U ABC A² Aa ABD A A

Font Paragraph Styles Insert Themes

AaBbCcDdEe AaBbCcDdEe AaBbCcDdEe Normal No Spacing Heading 1

Text Box Shape Picture

1 2 3 4 5

Step 1 - take the raw file, run version 3.1.2 of summarize software with parameters a=1, b=2, c=3

Step 2 - run the software separately for each sample

Step 3 - take column three of outputfile.out for each s and that is the corresponding row in the output data



Vague instructions
Missing versions
Skipped steps

The Leek group guide to data sharing — Edit

25 commits

1 branch

0 releases

8 contributors



branch: master

databsharing /

Merge pull request #9 from nikai3d/patch-1 · ...

jtleek authored 6 days ago

latest commit e53857faa4 ·

README.md

fix typo

6 days ago

README.md

How to share data with a statistician

This is a guide for anyone who needs to share data with a statistician. The target audiences I have in mind are:

- Scientific collaborators who need statisticians to analyze data for them
- Students or postdocs in scientific disciplines looking for consulting advice
- Junior statistics students whose job it is to collate/clean data sets

—

Organizing

“File organization and naming are powerful weapons against chaos.”

- Jenny Bryan

(via Karl Broman: http://kbroman.org/Tools4RR/assets/lectures/06_org_eda.pdf)



Name
.DS_Store
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_A01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_A02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_A03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_B01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_B02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_B03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_C01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_C02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_C03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_D01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_D02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_D03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_E01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_E02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_E03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_F01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_F02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_F03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_G01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_G02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_G03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H03.csv

Slide via Jenny Bryan:

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

One potential system

- Data
 - Raw data
 - Processed data
- Figures
 - Exploratory figures
 - Final figures
- R code
 - Raw scripts
 - Final scripts
 - R Markdown files (optional)
- Text
 - Readme files
 - Text of analysis

One potential system

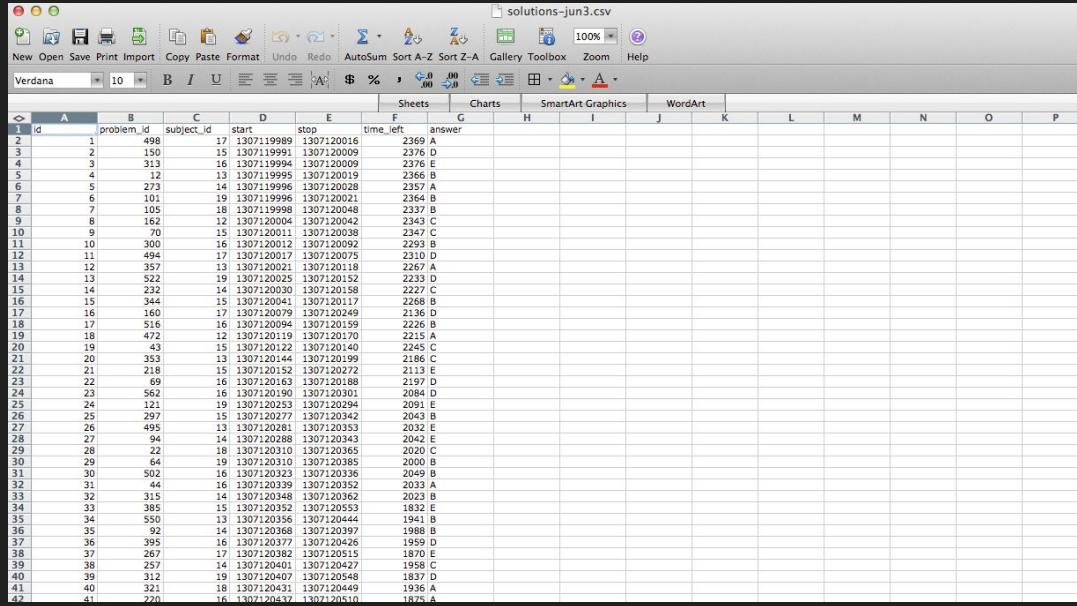
- Data R code/
0-preprocess.R
 - Raw data
 - Processed data
- Figures 1-explore.R
 - Exploratory figures
 - Final figures
- R code 2-model.R
 - Raw scripts
 - Final scripts
 - R Markdown files (optional)
- Text 3-final-plots.R
 - Readme files
 - Text of analysis

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

- Stored local copy
- If downloaded - keep date/location

Processed data

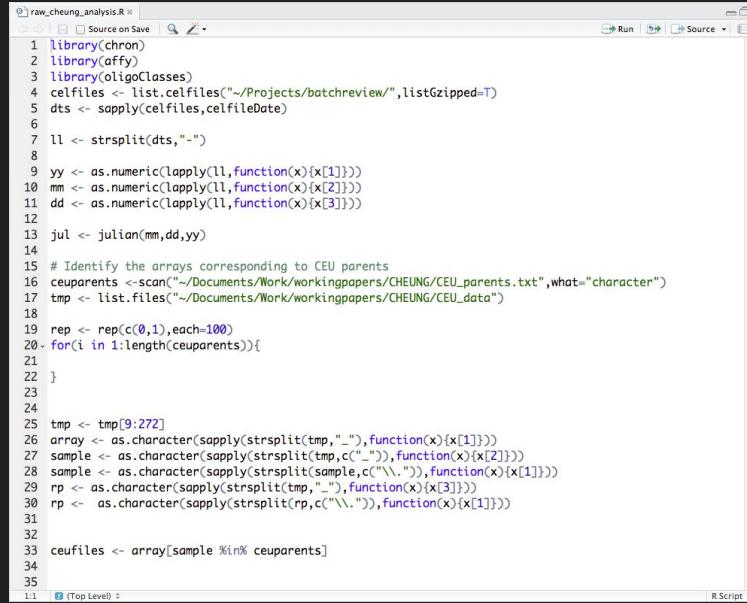


The screenshot shows a Microsoft Excel spreadsheet titled "solutions-jun3.csv". The data is organized into columns labeled A through P. Column A contains numerical IDs from 1 to 42. Columns B through P contain various data types, including numbers, dates, and letters. The first few rows of data are as follows:

	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	18	1307119991	1307120009	2376	D									
4	3	313	16	1307119992	1307120009	2375	E									
5	4	32	13	1307119995	1307120019	2366	B									
6	5	273	14	1307119996	1307120028	2357	A									
7	6	101	19	1307119996	1307120021	2364	B									
8	7	105	18	1307119998	1307120048	2337	B									
9	8	162	17	1307120004	1307120048	2343	C									
10	9	70	15	1307120014	1307120038	2347	C									
11	10	300	16	1307120012	1307120092	2293	B									
12	11	494	17	1307120017	1307120075	2310	D									
13	12	357	13	1307120021	1307120118	2267	A									
14	13	522	19	1307120025	1307120152	2233	D									
15	14	232	14	1307120030	1307120186	2227	C									
16	15	344	15	1307120031	1307120117	2208	B									
17	16	160	17	1307120079	1307120249	2136	D									
18	17	516	16	1307120094	1307120159	2226	B									
19	18	472	12	1307120119	1307120170	2215	A									
20	19	43	15	1307120122	1307120140	2245	C									
21	20	353	13	1307120144	1307120199	2186	C									
22	21	218	15	1307120151	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	15	1307120281	1307120353	2032	E									
28	27	94	14	1307120301	1307120443	2042	E									
29	28	22	18	1307120310	1307120365	2020	C									
30	29	64	19	1307120310	1307120385	2000	B									
31	30	502	16	1307120323	1307120336	2049	B									
32	31	44	16	1307120339	1307120352	2033	A									
33	32	315	14	1307120352	1307120362	2035	B									
34	33	385	15	1307120352	1307120553	1832	E									
35	34	550	13	1307120356	1307120444	1941	B									
36	35	92	18	1307120368	1307120397	1988	B									
37	36	395	16	1307120377	1307120426	1959	D									
38	37	267	17	1307120382	1307120516	1970	E									
39	38	257	14	1307120387	1307120527	1935	C									
40	39	312	19	1307120407	1307120548	1837	D									
41	40	321	18	1307120431	1307120449	1936	A									
42	41	220	16	1307120437	1307120510	1875	A									

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

Raw scripts



```
raw_cheung_analysis.R <--> Source on Save | Run | Source | R Script | (Top Level) |
```

```
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 ceuparents <- array[sample %in% ceuparents]
34
35
```

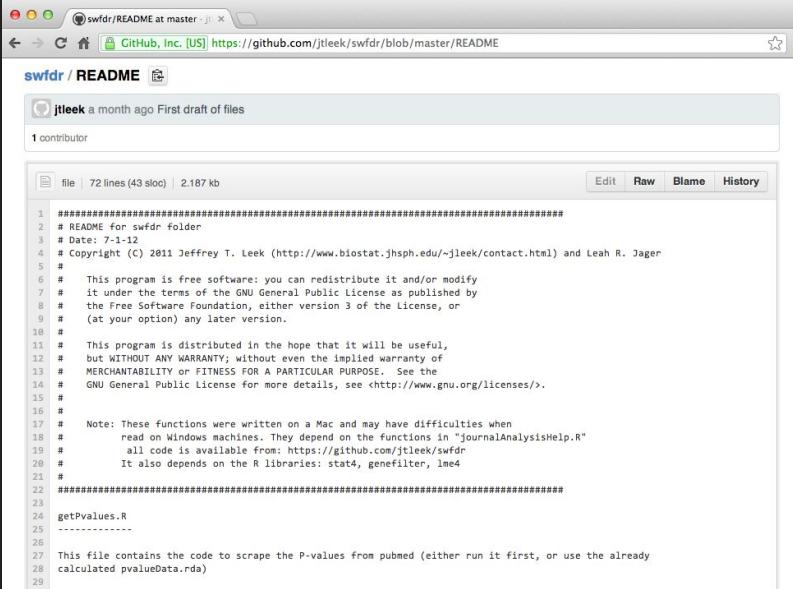
- May be less commented (but comments help you!)
- May be multiple versions
- May include analyses that are later discarded

Final scripts

```
1:1 f.value <- 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: ncchb <- length(chb.names)  
34: ...  
R Script
```

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

README file

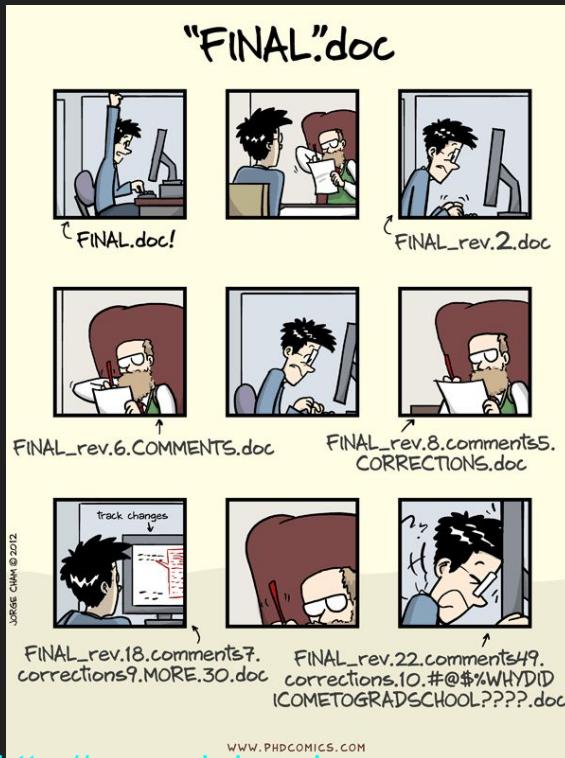


A screenshot of a web browser displaying a GitHub README file. The URL is <https://github.com/jtleek/swfdr/blob/master/README>. The page shows the file content with line numbers and various sections like license and dependencies.

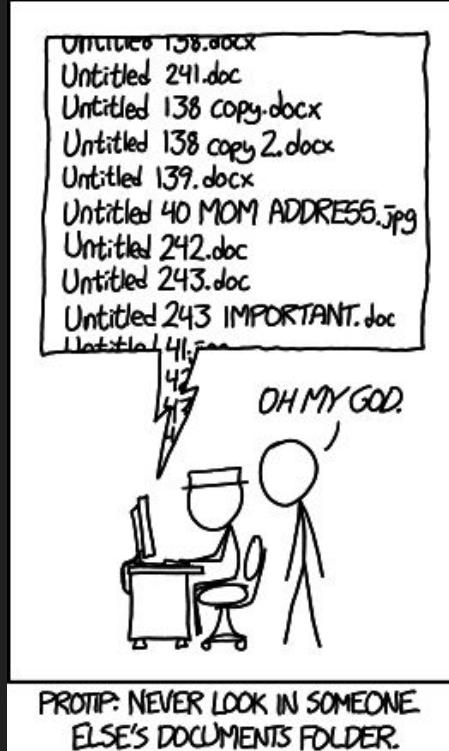
```
1 #####
2 # README for swfdr folder
3 # Date: 7-1-12
4 # Copyright (C) 2011 Jeffrey T. Leek (http://www.biostat.jhsph.edu/~jtleek/contact.html) and Leah R. Jager
5 #
6 # This program is free software: you can redistribute it and/or modify
7 # it under the terms of the GNU General Public License as published by
8 # the Free Software Foundation, either version 3 of the license, or
9 # (at your option) any later version.
10 #
11 # This program is distributed in the hope that it will be useful,
12 # but WITHOUT ANY WARRANTY; without even the implied warranty of
13 # MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the
14 # GNU General Public License for more details, see <http://www.gnu.org/licenses/>.
15 #
16 #
17 # Note: These functions were written on a Mac and may have difficulties when
18 #       read on Windows machines. They depend on the functions in "journalAnalysisHelp.R"
19 #       all code is available from: https://github.com/jtleek/swfdr
20 #       It also depends on the R libraries: stat4, genefilter, lme4
21 #
22 #####
23
24 getPValues.R
25 -----
26
27 This file contains the code to scrape the P-values from pubmed (either run it first, or use the already
28 calculated pvalueData.rda)
29
```

- Should contain step-by-step instructions for analysis
- Here is an example <https://github.com/jtleek/swfdr/blob/master/README.md>

Just no



<http://www.phdcomics.com/comics/archive.php?comicid=1521>



<https://xkcd.com/1459/>

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)

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.

A screenshot of a web browser displaying the RStudio homepage. The address bar shows 'www.rstudio.com'. The page features a large 'Welcome to RStudio' heading and a large blue circular logo with a white 'R'. Below the main heading are three sections: 'Powerful IDE for R', 'R training and education', and 'Open source R packages', each with a call-to-action button.

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.

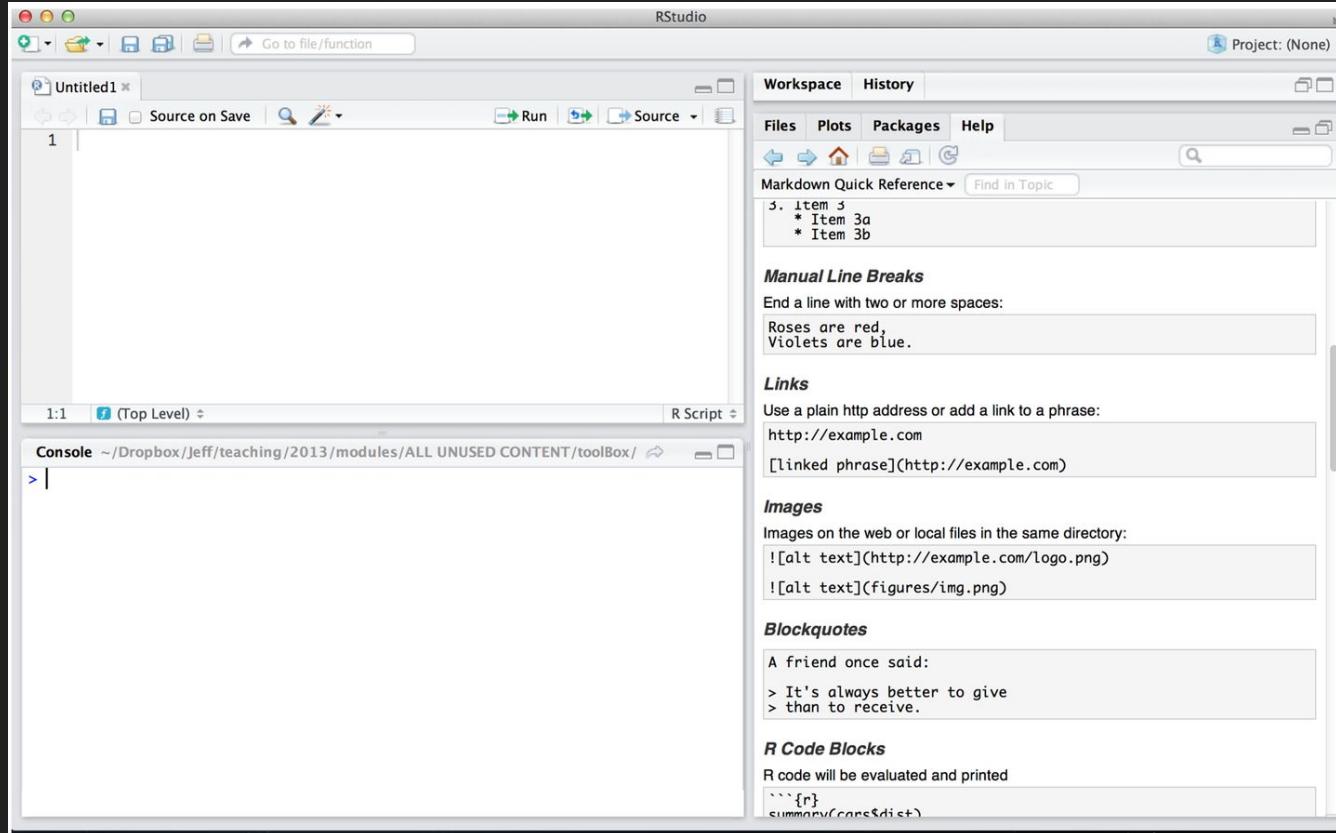
[Request on-site](#) [View courses](#)

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](#)

<https://www.rstudio.com/>



<https://www.rstudio.com/>

Some useful commands

Cmd + Enter

Evaluates line of code (Mac)

Ctrl + Enter

Evaluates line of code
(Windows)

Ctrl + 1

Switch to script page

Ctrl + 2

Switch to console

Installing Time

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

Rstudio Tour

<https://goo.gl/yAAxHC>

R packages



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

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

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

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

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

Available CRAN Packages By Name

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

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

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

All Packages

Bioconductor version 3.1 (Release)

Autocomplete biocViews search:

Software (1024)

▶ AssayDomain (345)

▶ BiologicalQuestion (313)

▶ Infrastructure (211)

▶ ResearchField (225)

▶ StatisticalMethod (293)

▶ Technology (645)

▶ WorkflowStep (525)

▶ AnnotationData (883)

▶ ExperimentData (241)

Packages found under Software:

Show All entries

Search table:

Package	Maintainer	Title
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 Microarray (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.

sva

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

Surrogate Variable Analysis

Bioconductor version: Release (3.1)

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

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

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

Downloads

Bioconductor workflows

Arrays

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

Mailing Lists

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

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

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

sva

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

Surrogate Variable Analysis

Bioconductor version: Release (3.1)

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

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

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

Responsiveness

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

Mailing Lists

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

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

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

sva

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

Surrogate Variable Analysis

Bioconductor version: Release (3.1)

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

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

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

Still runs

kflows »

non Bioconductor workflows
de:

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

Mailing Lists »

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

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

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



dgrtwo / broom

Watch

16



Convert statistical analysis objects from R into tidy format

146 commits

1 branch

8 releases

10 contributors



branch: master ▾

broom / +



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



dgrtwo authored 3 hours ago

latest commit ec5c0bd980



Merge pull request #51 from zeehio/master

3 hours ago



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

7 months ago



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

21 days ago



Changed `rowwise_df_tidiers` to allow the original data to be saved a...

a month ago



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

7 months ago



Update cran comments.

6 months ago



Update cran comments.

6 months ago



Merge pull request #51 from zeehio/master

3 hours ago

<https://github.com/dgrtwo/broom>



jtlee / 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 / +

Cancel

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

Unwatch 6Star 4Fork 7

Description

Short description of this repository

26 commits

People have been
working on it

Commit made by the Bioconductor Git-SVN bridge. ...

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

Code

Issues 0Pull 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



>



>



github
SOCIAL CODING

R package installation

<https://goo.gl/SNeI00>