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

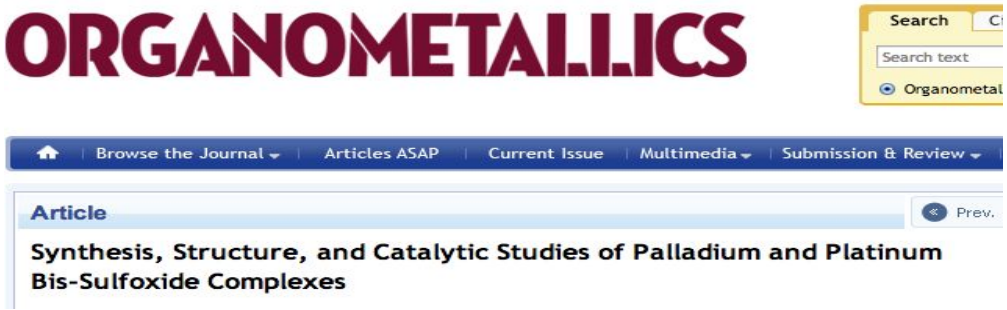
Background

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 arraysand 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.

http://bit.ly/OgW3xv



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

Drinkel et al. *Oganometalics* 2013

How we* feel about statistics

@simplystats
http://simplystatistics.org

www.jhudatascience.org

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.**

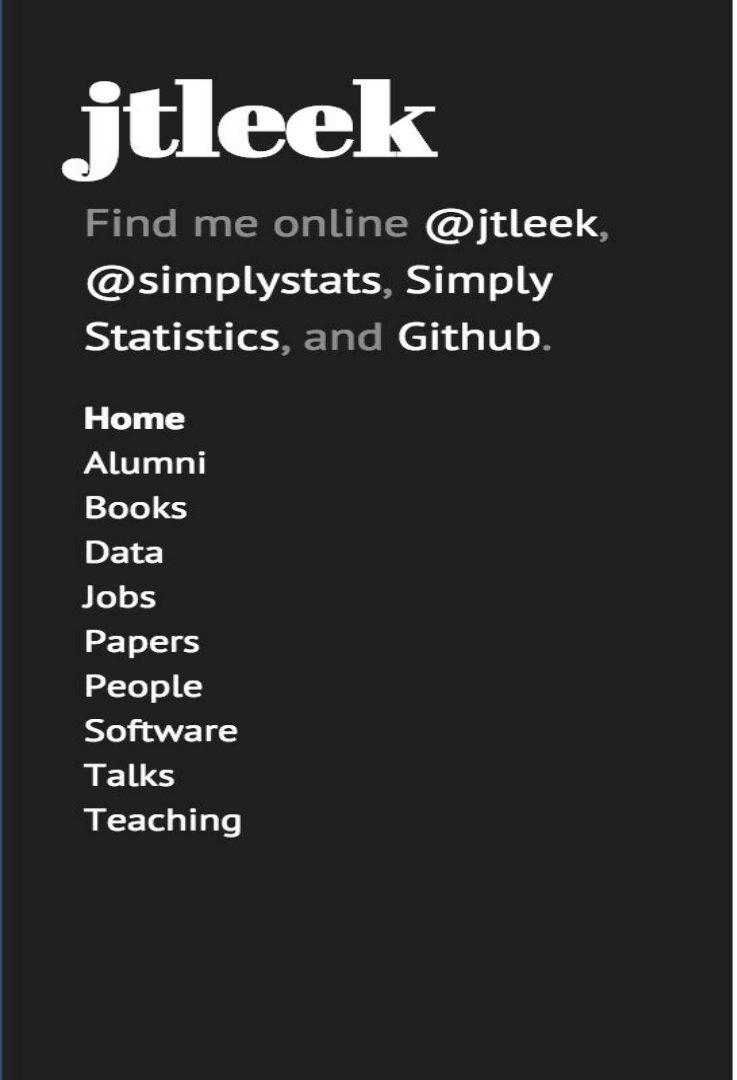
LATEST ARTICLES

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

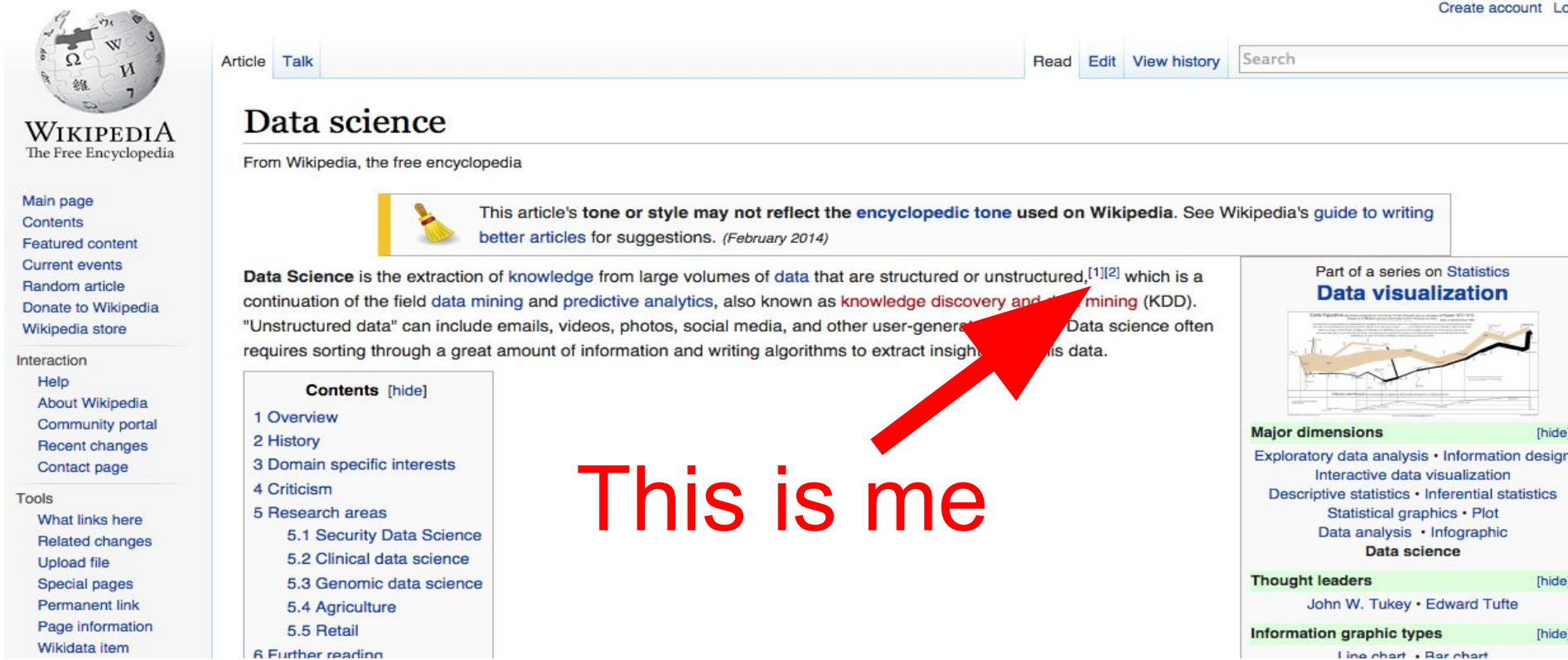
Impact factor: 2.649
5-Yr impact factor: 2.853



@jtleek

http://www.jtleek.com

I defined data science



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.

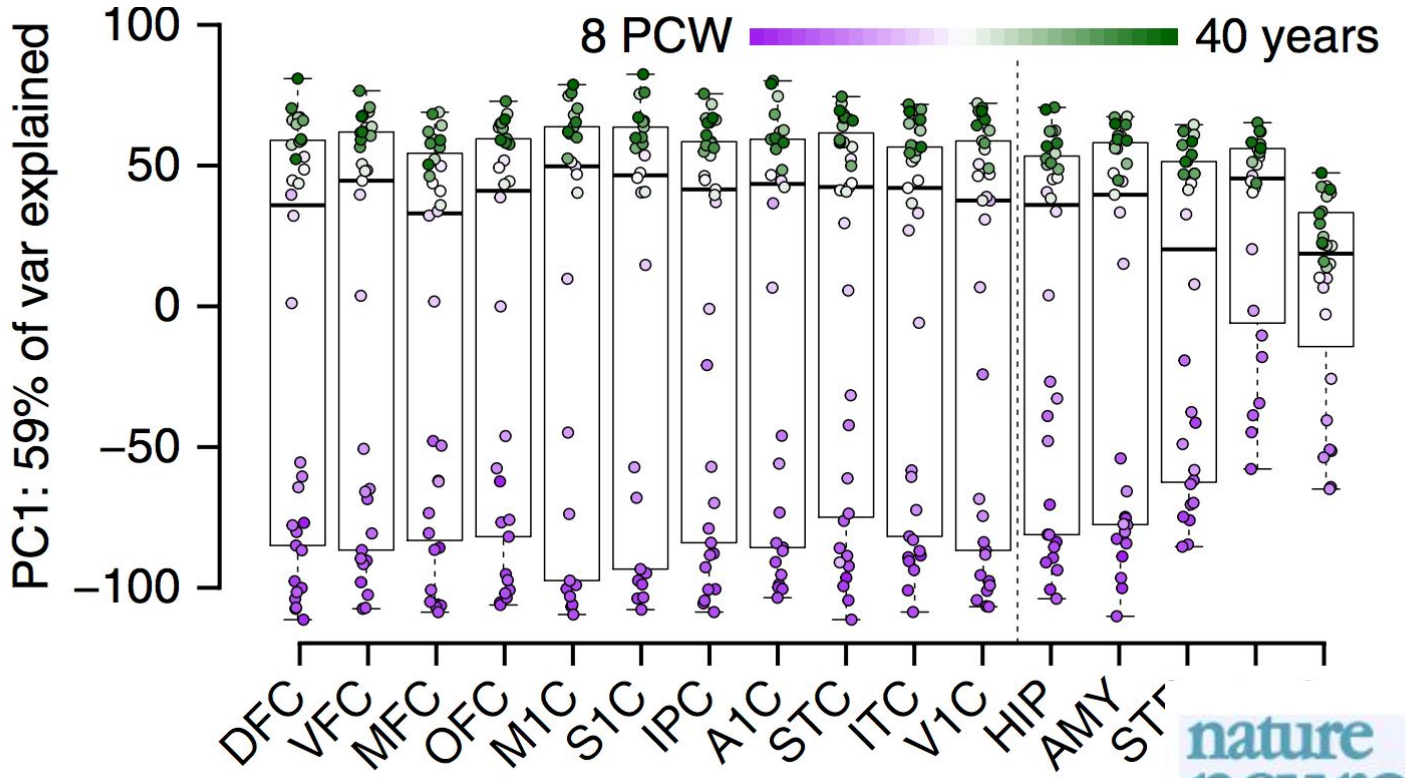
Contact

@andrewejaffe
http://www.aejaffe.com

Email:

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

Andrew E Jaffe[1–3], Jooheon Shin[1], Leonardo Collado-Torres[1,2], Jeffrey T Leek[2,4], Ran Tao[1], Chao Li[1], Yuan Gao[1], Yankai Jia[1], Brady J Maher[1,5,6], Thomas M Hyde[1,5–8], Joel E Kleinman[1,9] & Daniel R Weinberger[1,4–7,9]



nature neuroscience

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 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

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.

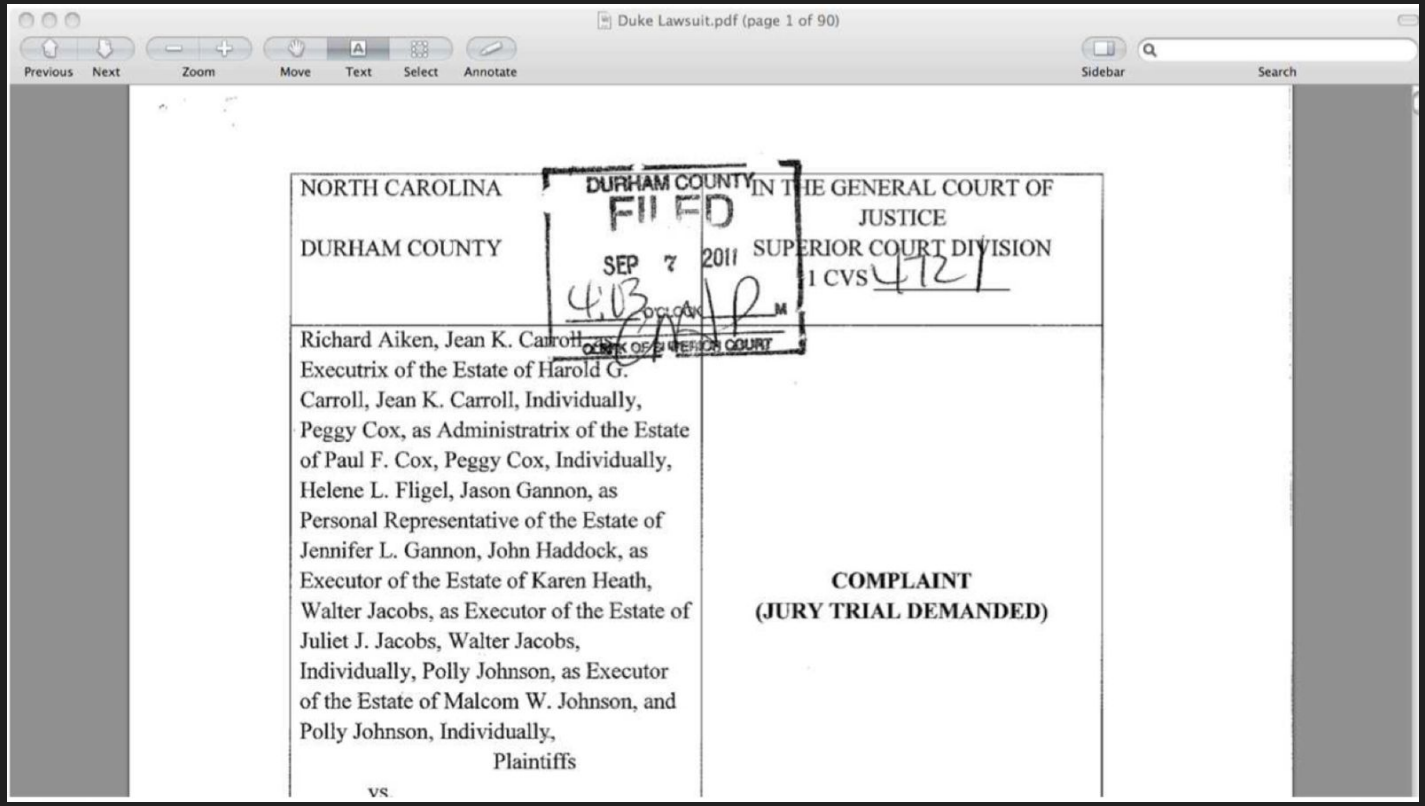
JULY 2010 *The Cancer Letter* reveals that Potti made false claims about his CV. Trials are suspended and an investigation begins. Harold Varmus, director of the National Cancer Institute in Bethesda, Maryland, asks the Institute of Medicine to review Duke's trials.

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.

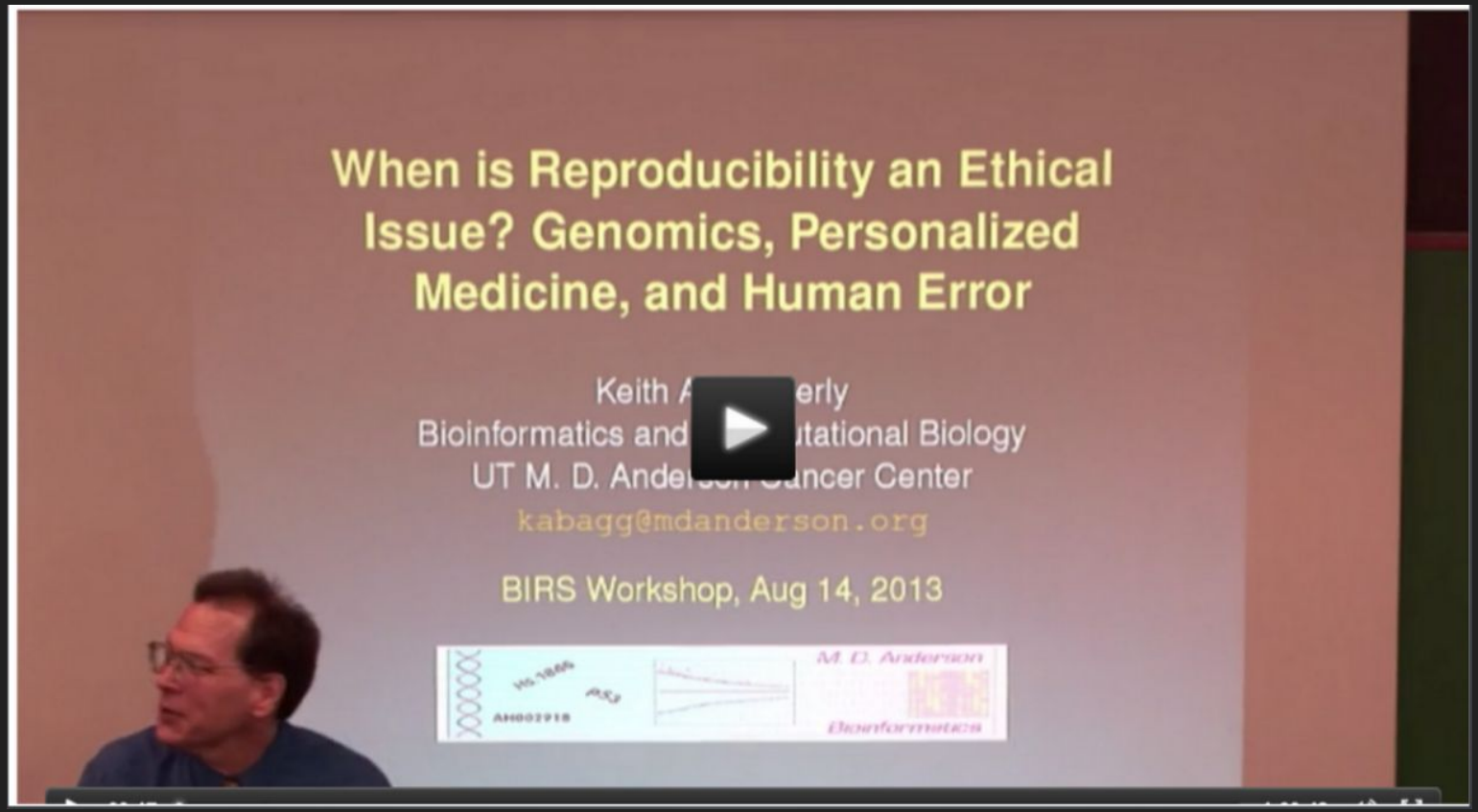
JANUARY 2011 *Nature Medicine* paper is retracted.

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



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

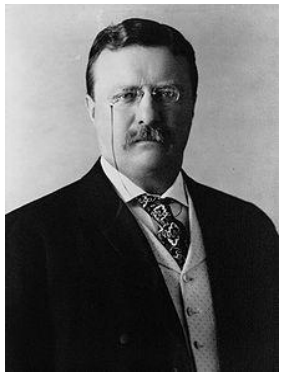
http://www.dukechronicle.com/articles/2015/05/03/duke-lawsuit-involving-cancer-patients-linked-anil-potti-settled



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

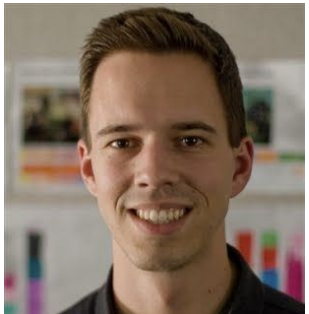
It is not the critic who counts

"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."

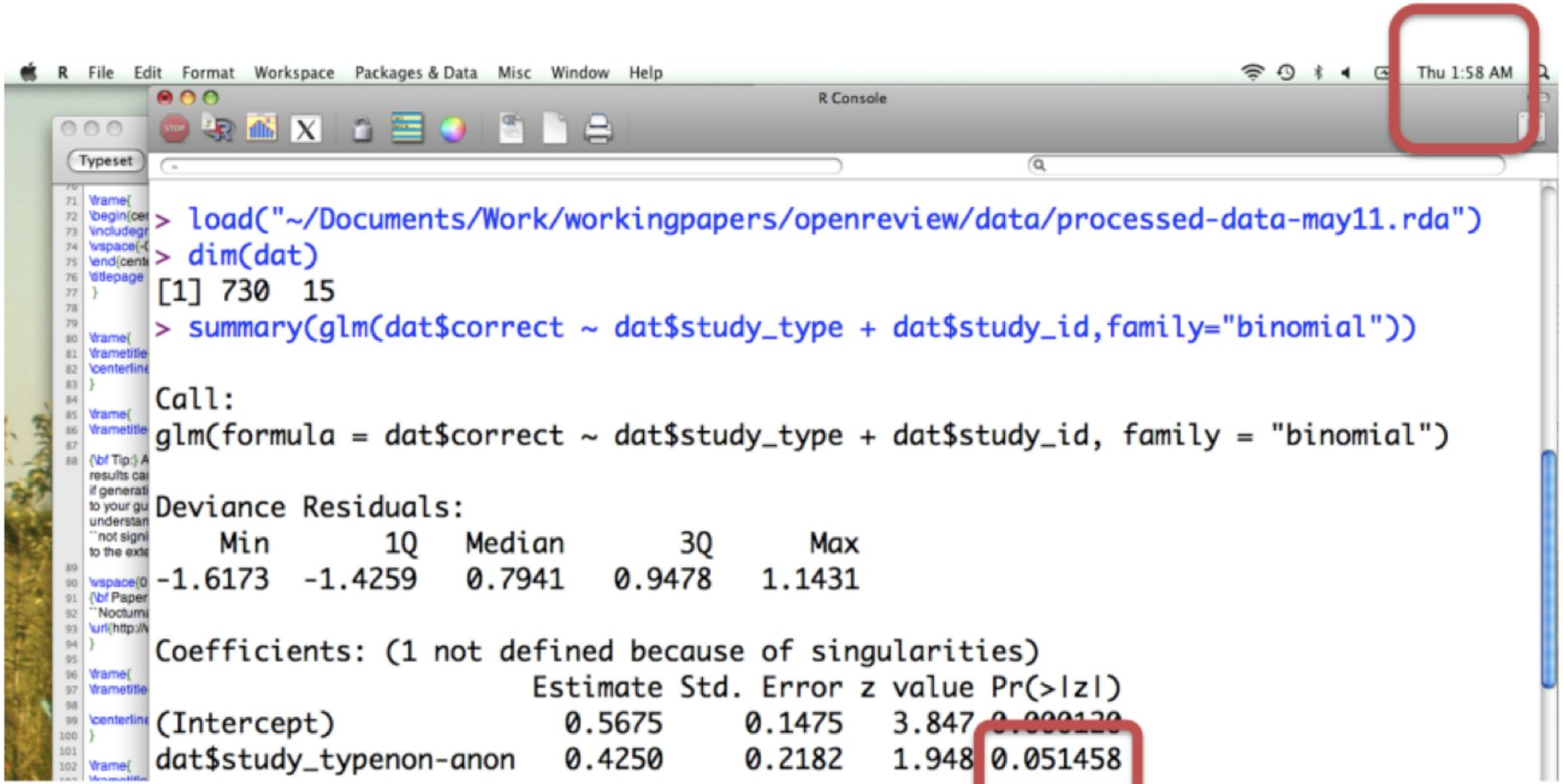


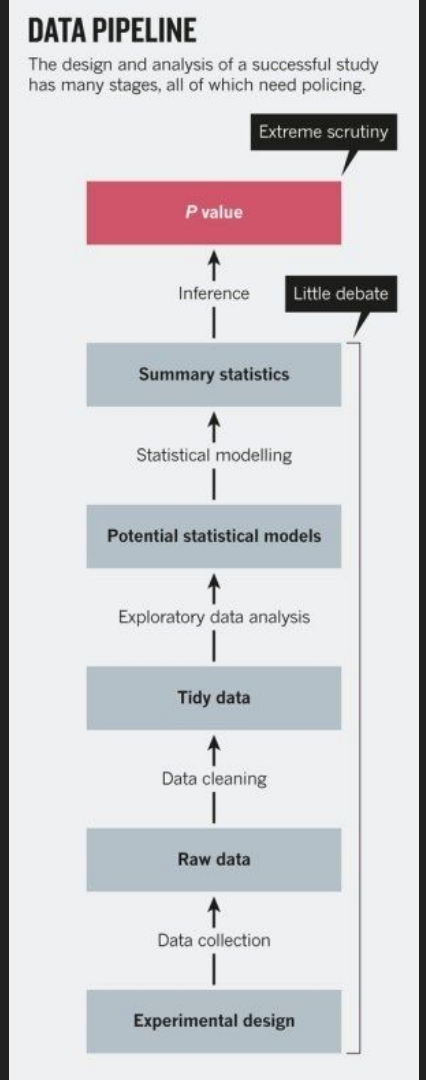
Ask yourselves, what problem have you solved, ever, that was worth solving, where you knew knew all of the given information in advance? Where you didn’t have a surplus of information and have to filter it out, or you didn’t have insufficient information and have to go find some?

Dan Myer, Mathematics Educator



Why this class?


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



- 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

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

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

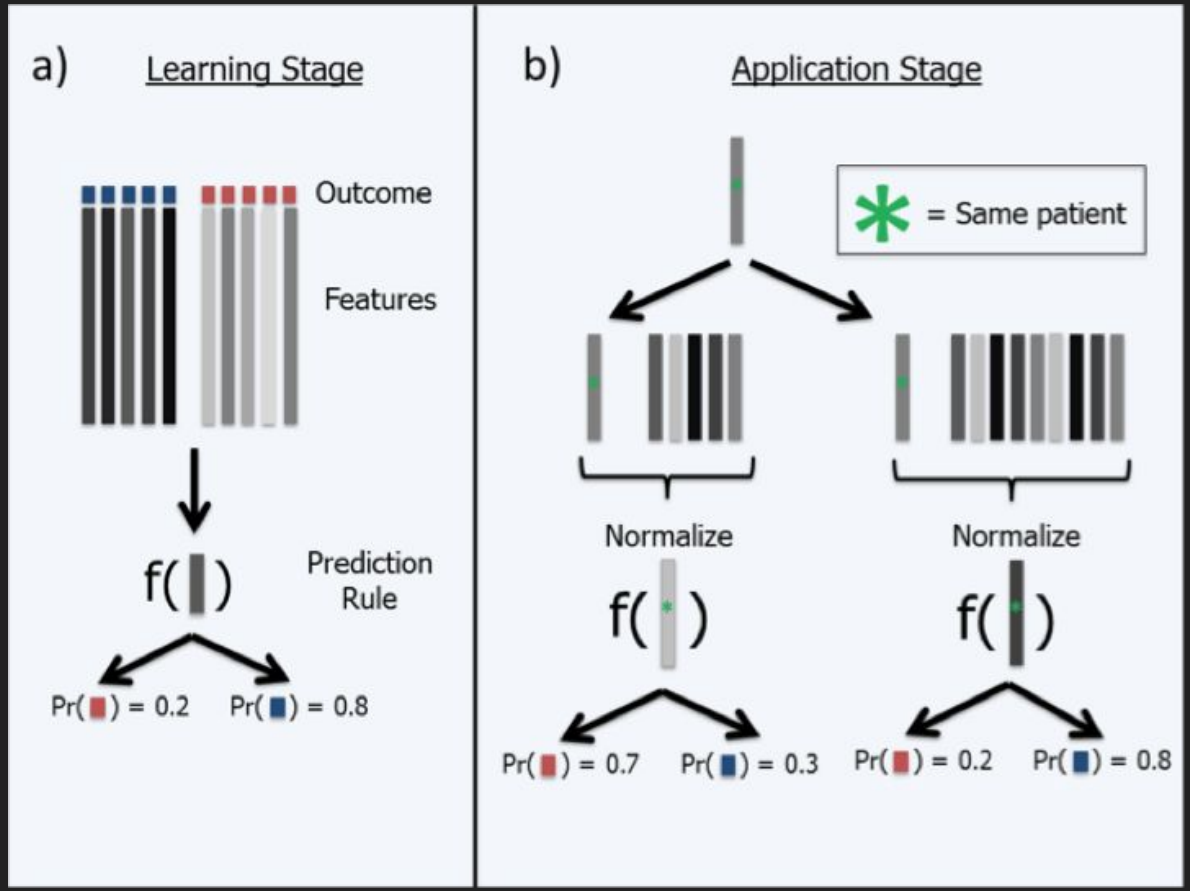
The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*

Andrew Gelman† and Eric Loken‡

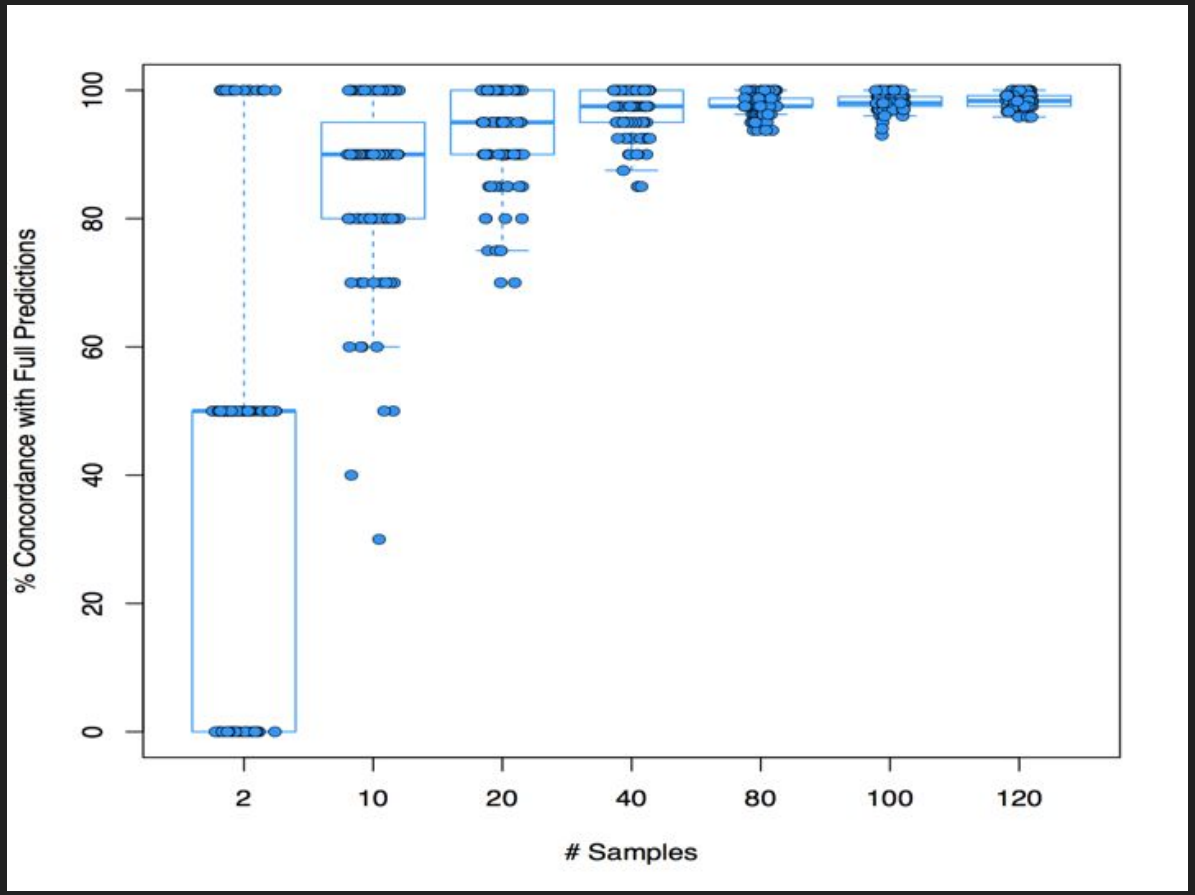
14 Nov 2013

"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)

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



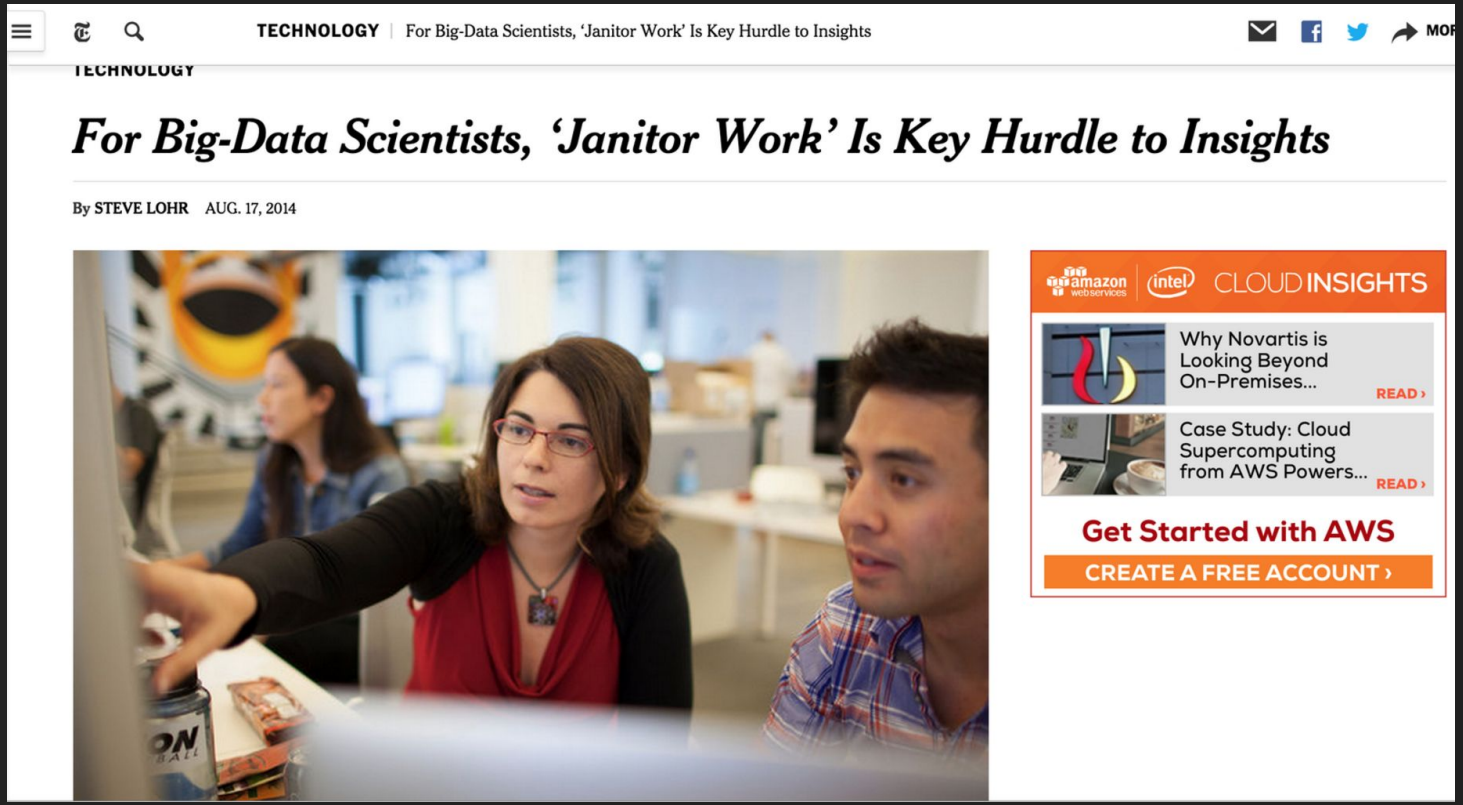
http://www.ncbi.nlm.nih.gov/pubmed/25788628



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

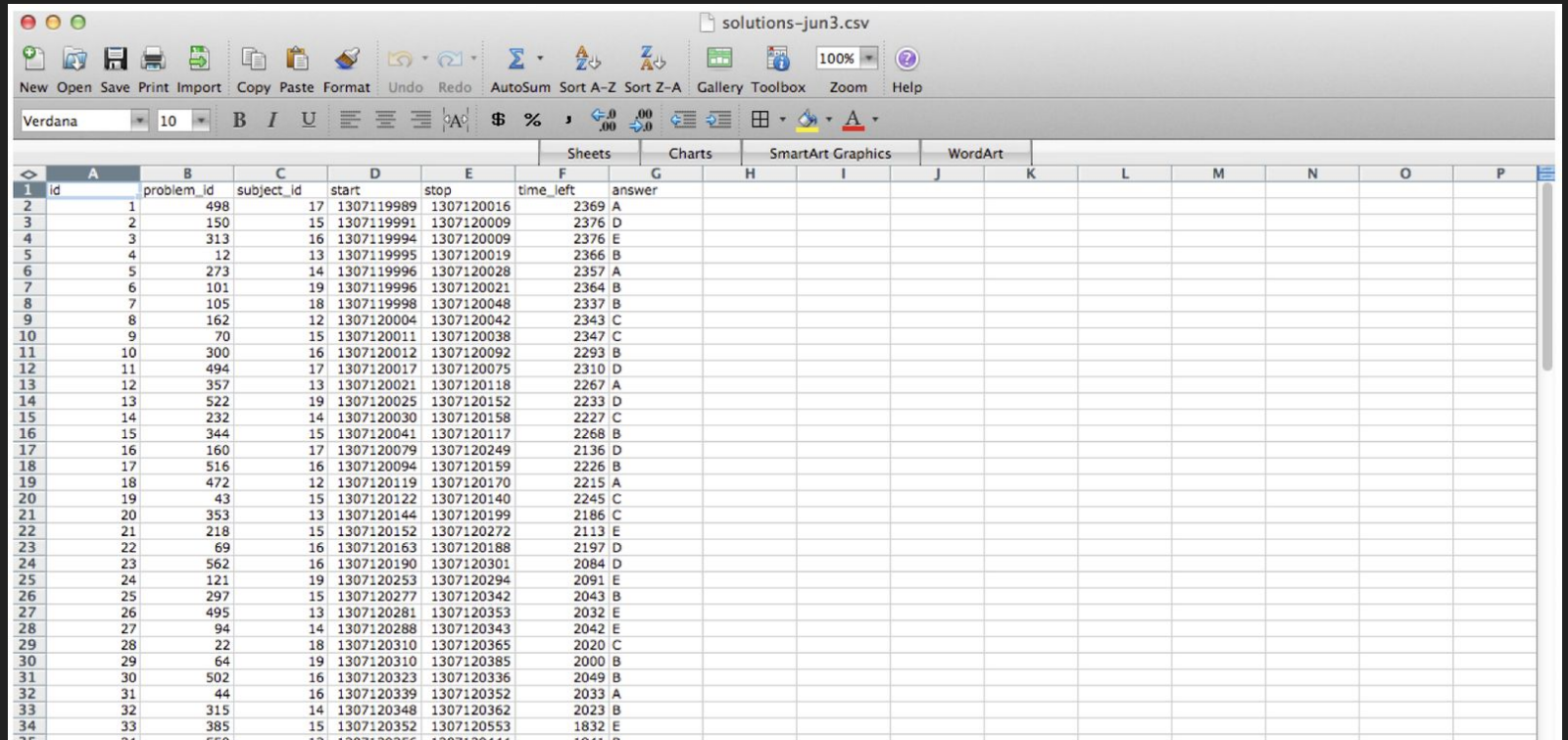
Herein lies the dirty secret about most data scientists' work -- it's more data munging than deep learning. The best minds of my generation are deleting commas from log files, and that makes me sad. A Ph.D. is a terrible thing to waste.

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



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



solutions-jun3.csv

New Open Save Print Import Copy Paste Format Undo Redo AutoSum Sort A-Z Sort Z-A Gallery Toolbox 100% Zoom Help

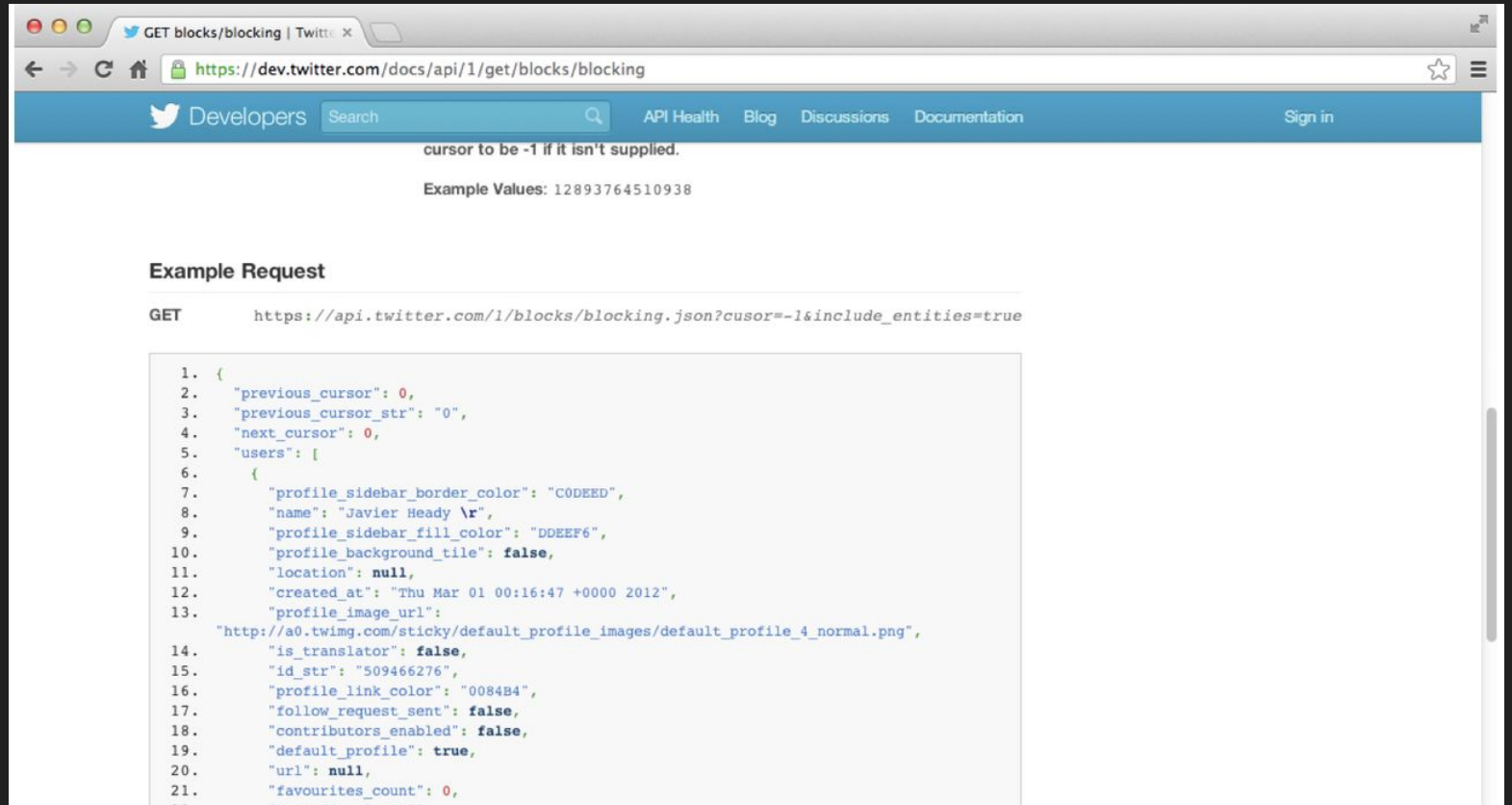
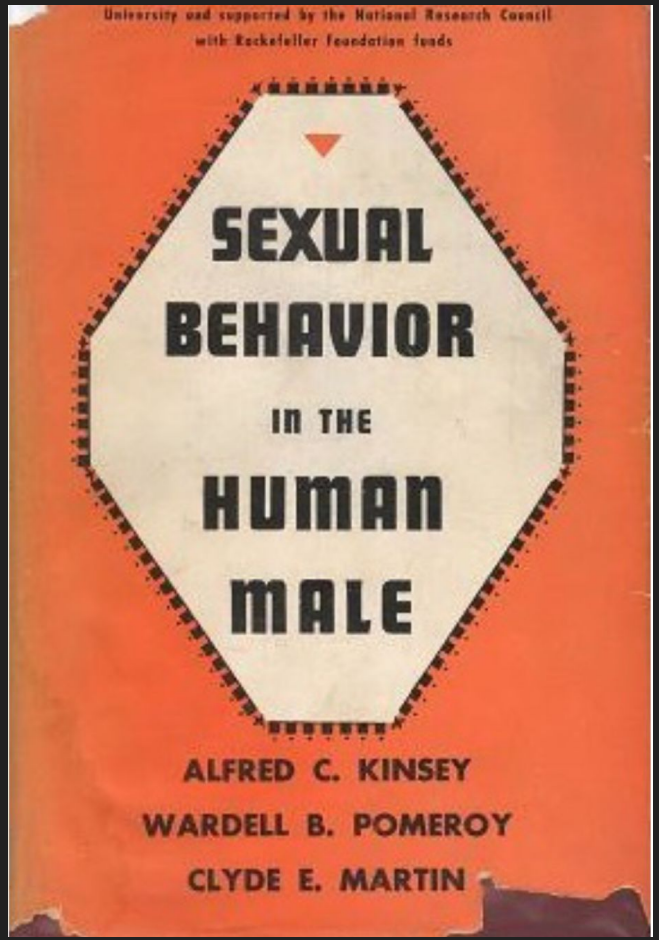
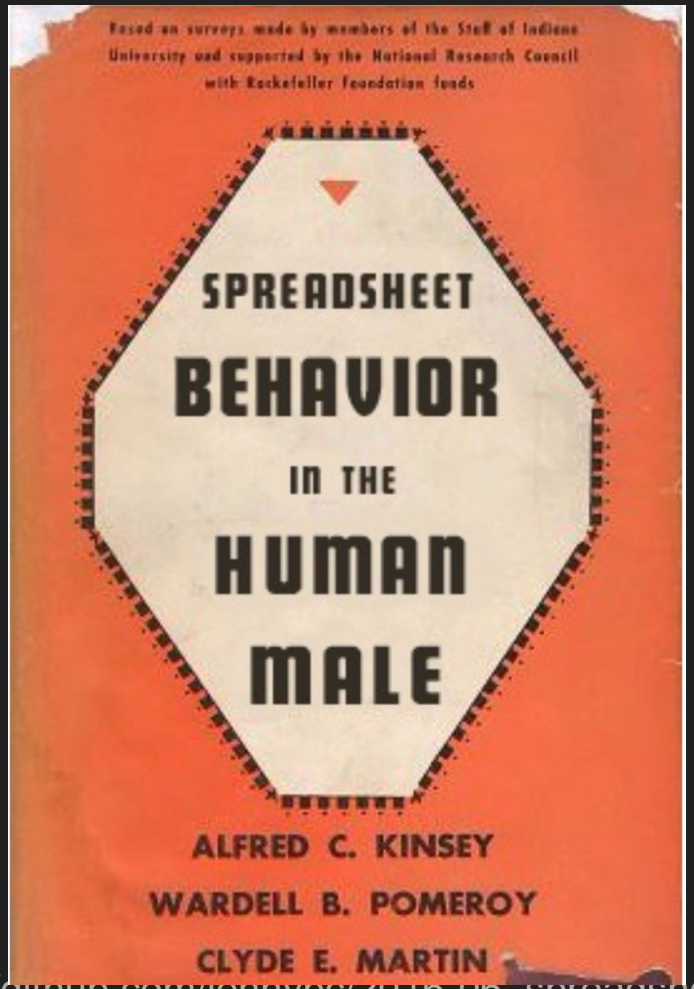
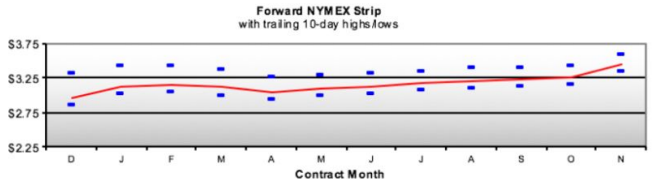
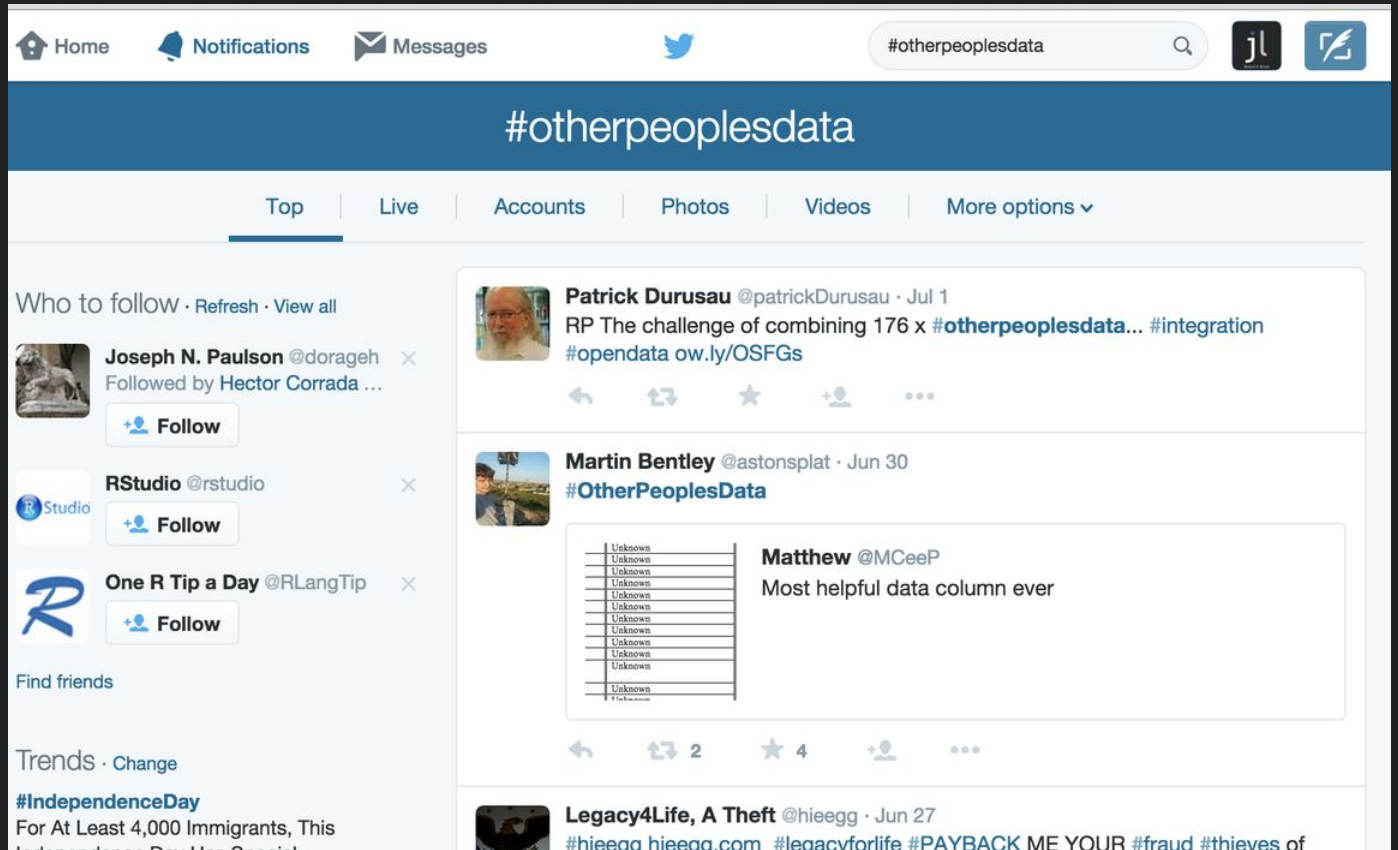
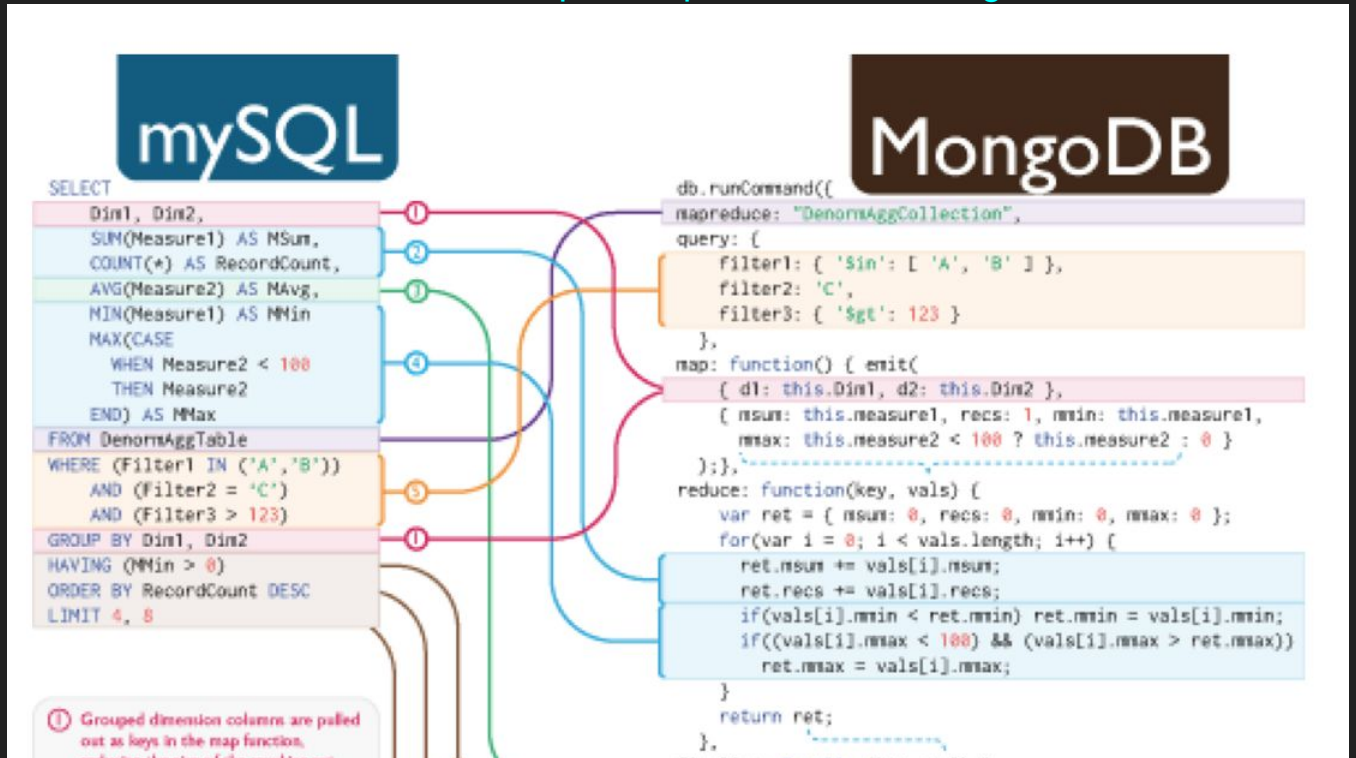
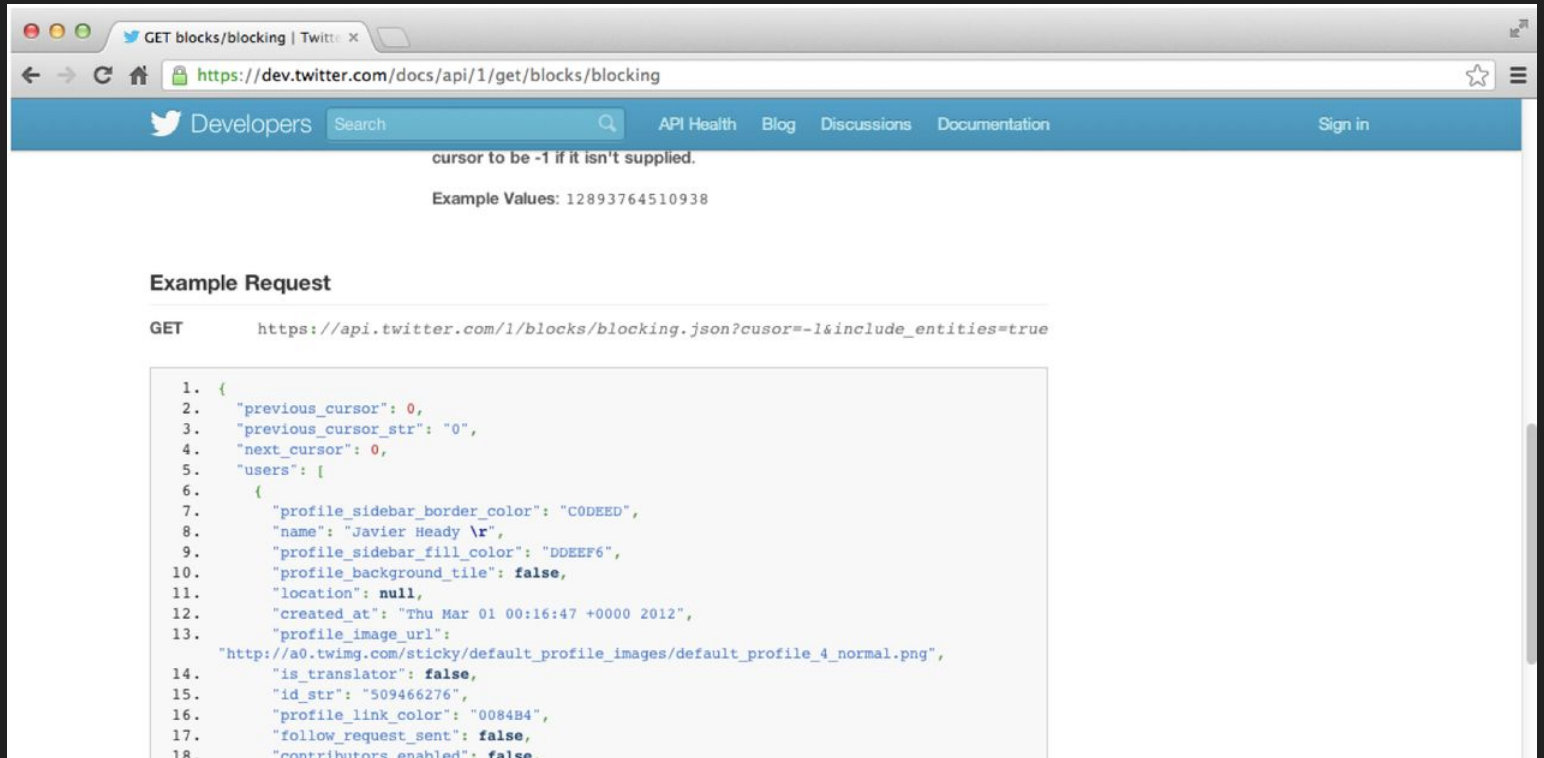
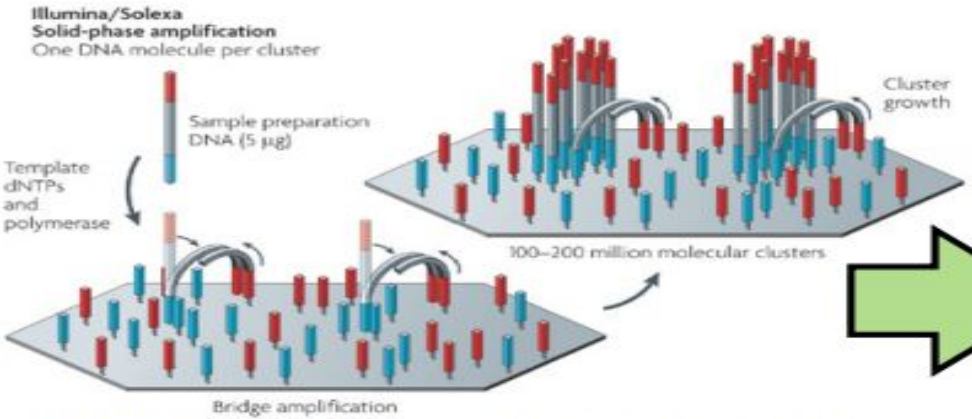
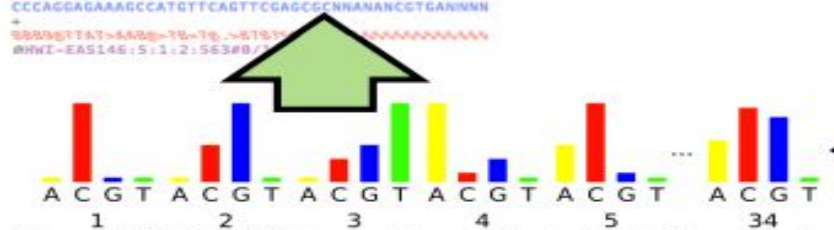
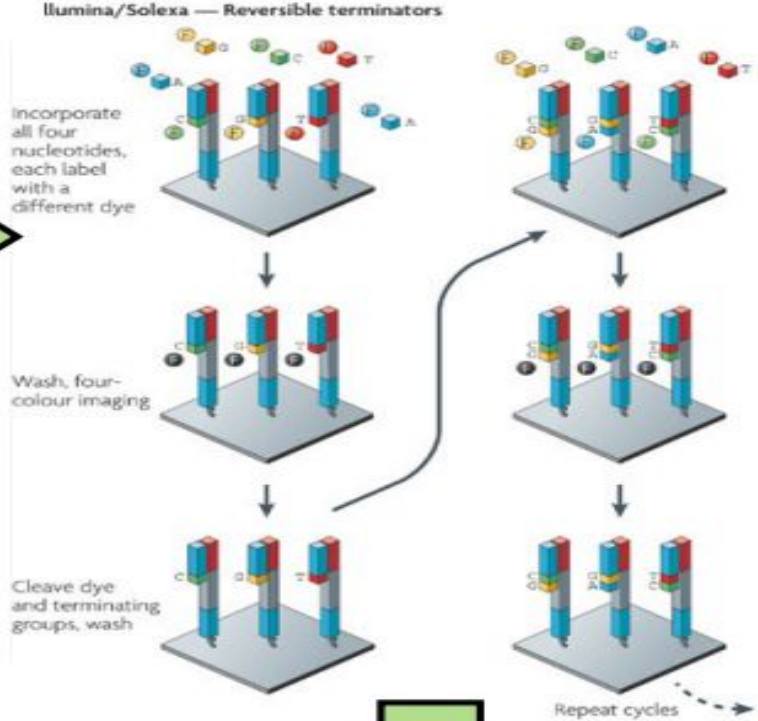
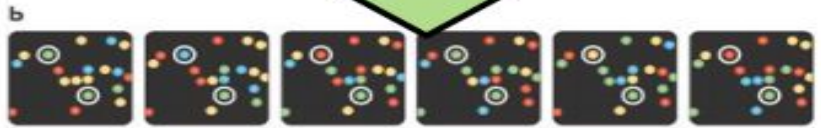
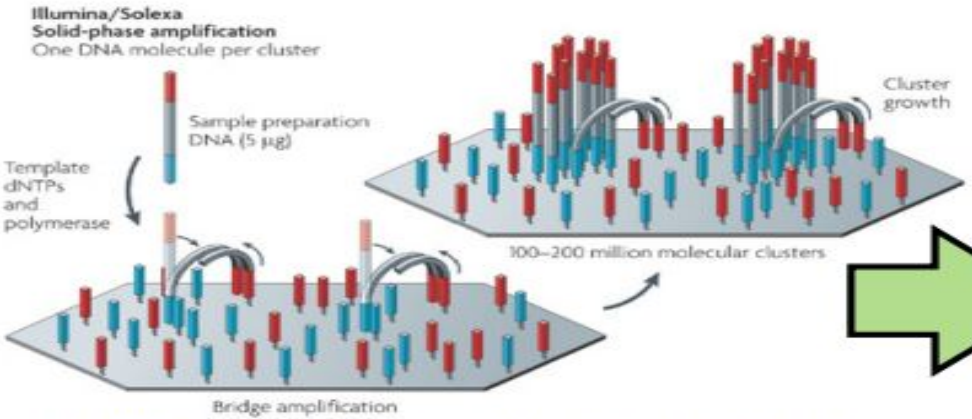
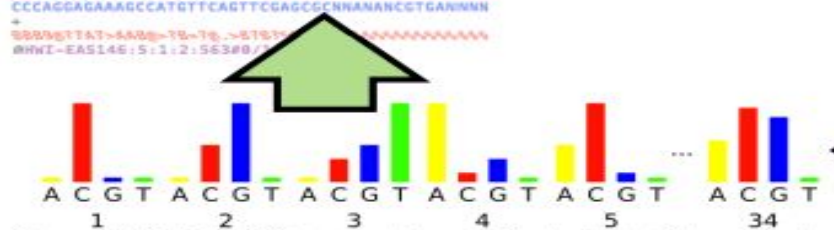
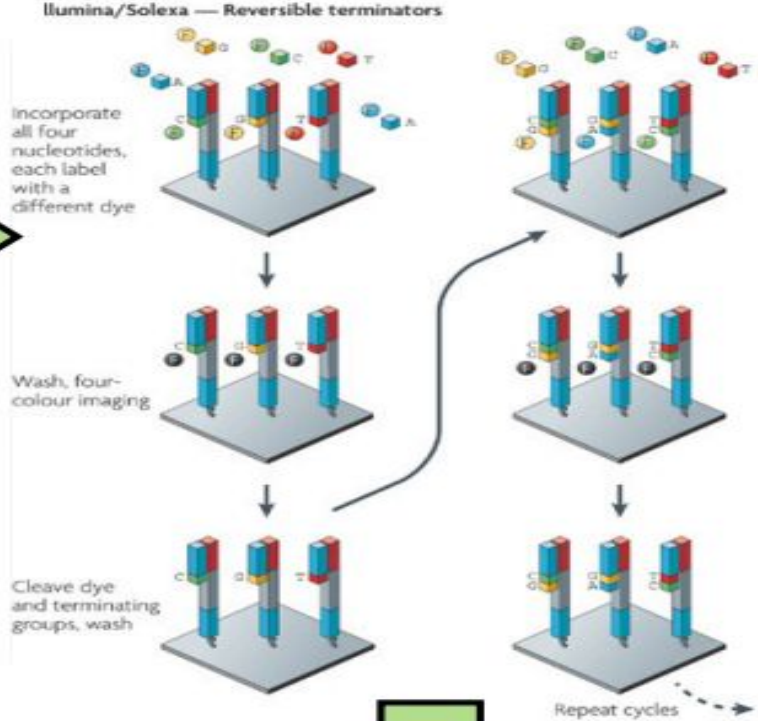
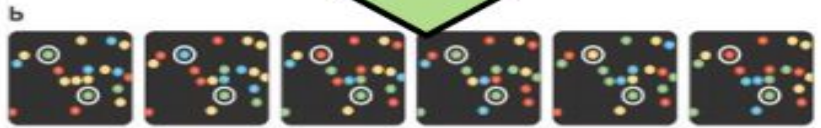
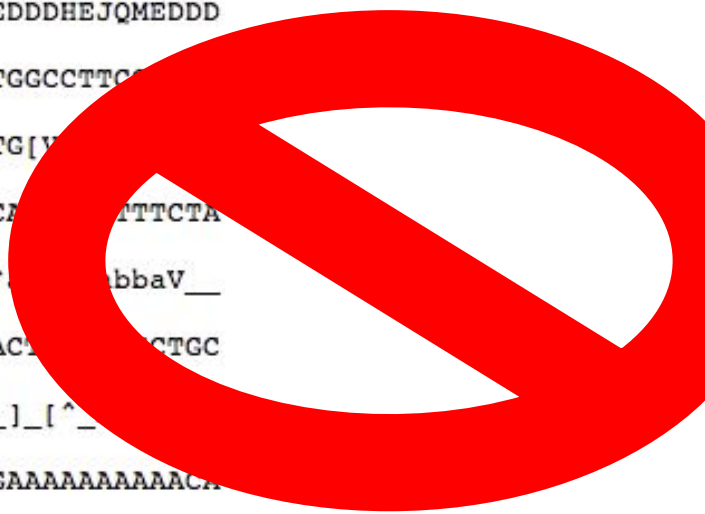
Verdana 10

Sheets Charts SmartArt Graphics WordArt

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

IF CIG Rocky Mountains	Fixed Price		Basis	
	BID	OFFER	BID	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		Basis	
	BID	OFFER	BID	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		Basis	
	BID	OFFER	BID	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		Basis	
	BID	OFFER	BID	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		Basis	
	BID	OFFER	BID	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		Basis	
	BID	OFFER	BID	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)

#otherpeoplesdata



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/


GET ALL THE
DATA!!!

Data brainstorming

https://goo.gl/r26muZ

Raw and processed data

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

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/*

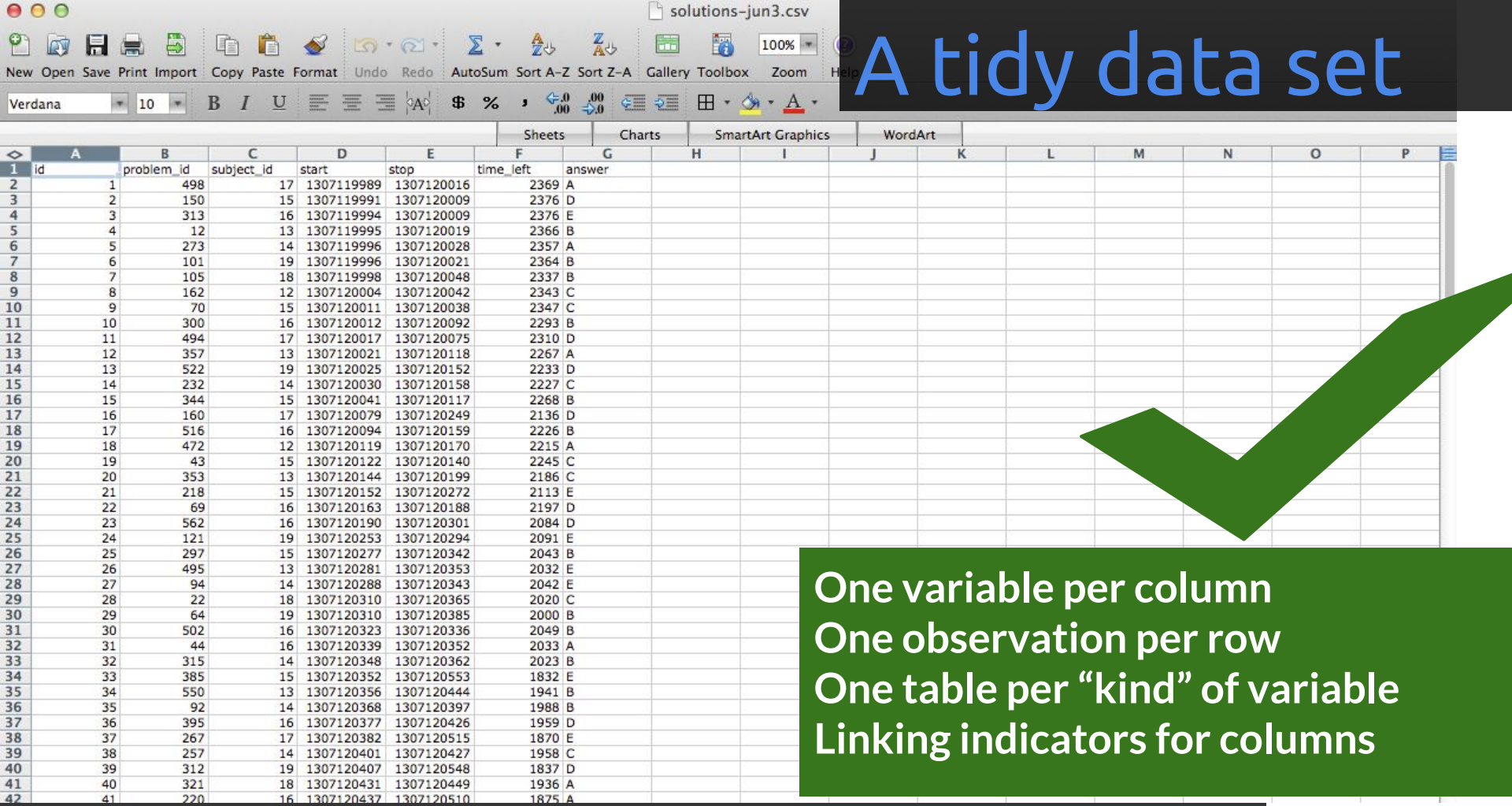
Tidy Data

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

Code book

Decoder.docx

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)

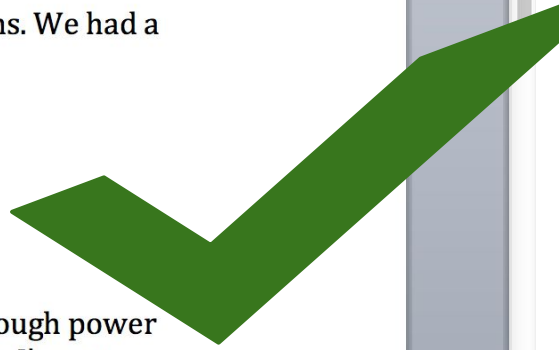
G. Diagnosis: DSM/ICD-10 diagnosis

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

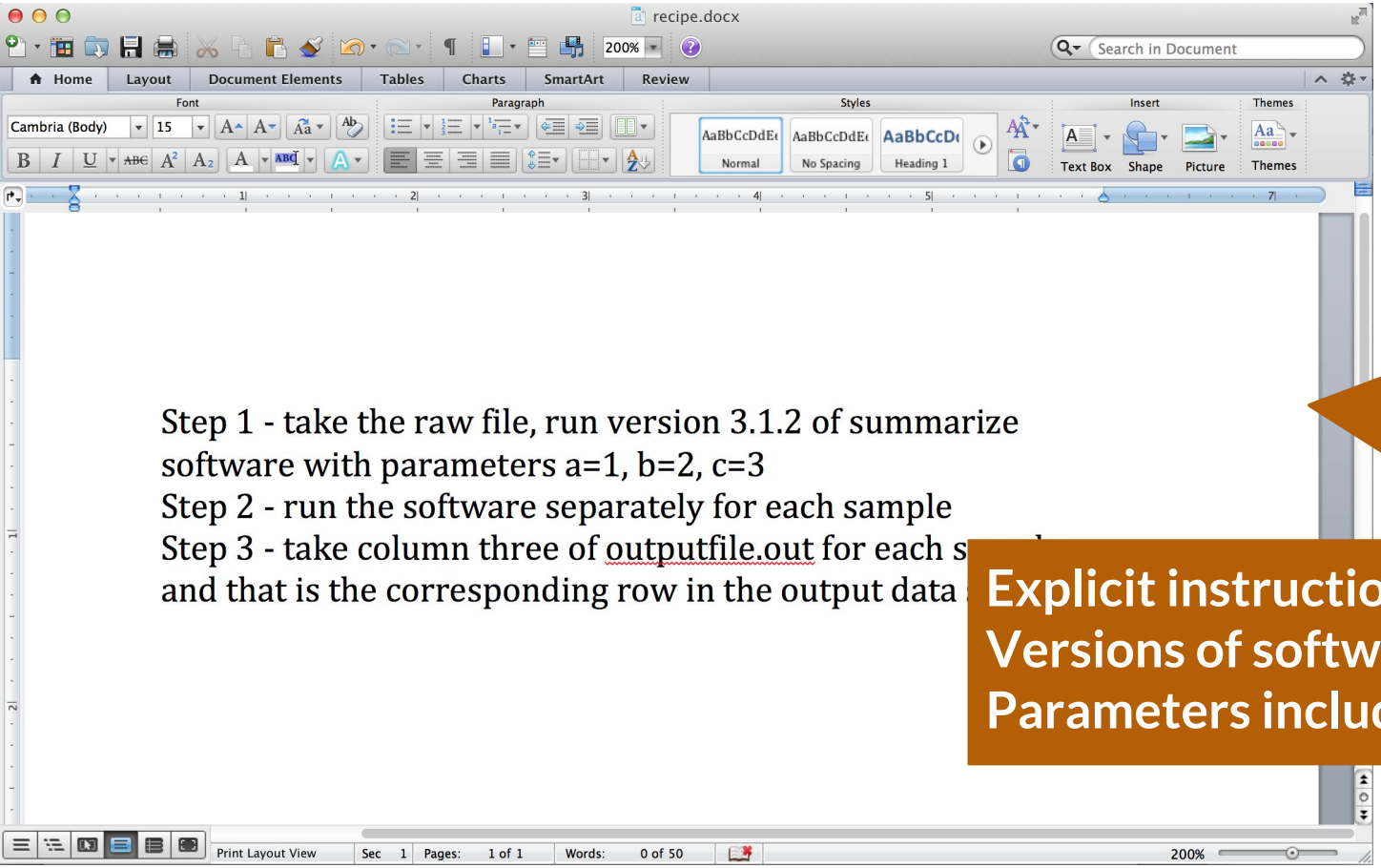
geuvadis.Rmd

Knit HTML | Run | Chunks

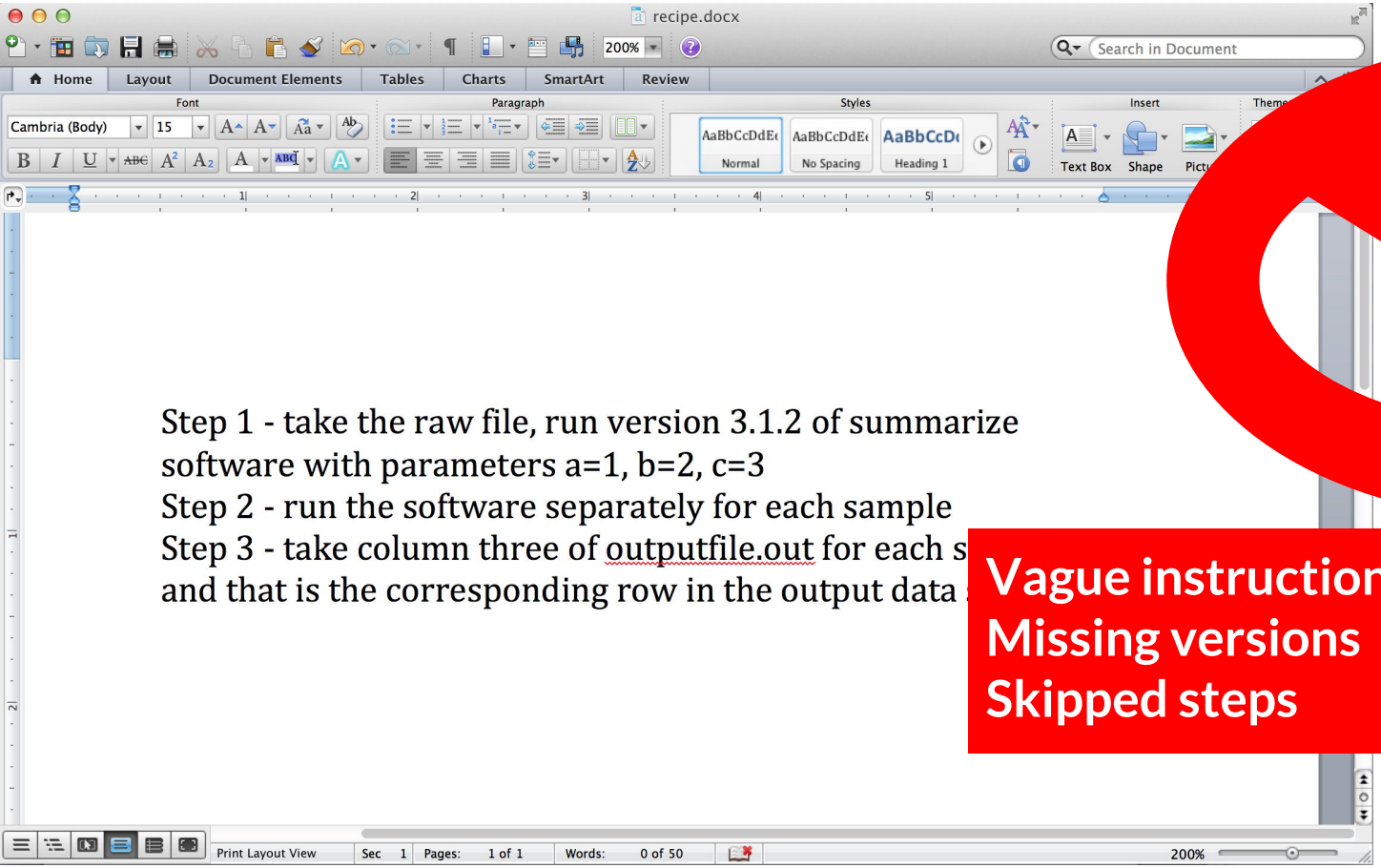
```
33  library(sva)
34  library(ffpe)
35  library(RColorBrewer)
36  library(corrplot)
37  library(limma)
38  trop = RSkittleBrewer('tropical')
39  ```
40
41
42  ## Load the data
43
44  You will need to download the GEUVADIS ballgown object from this site: https://github.com/a[illegible]uzee
    /ballgown_code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

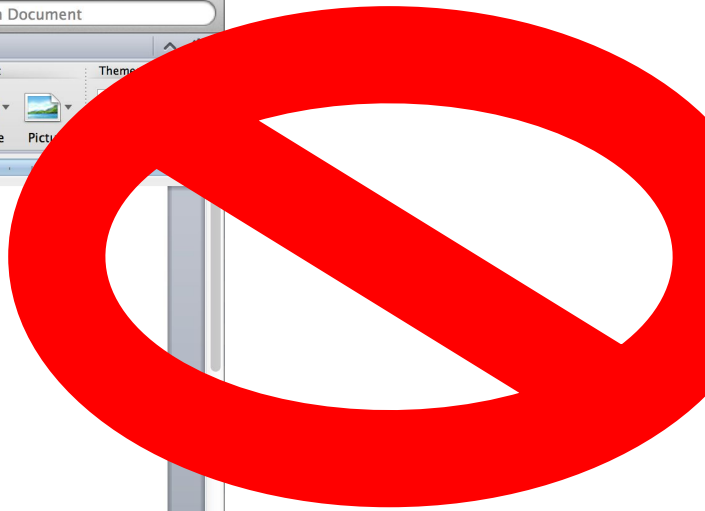
1:1 (Top Level) R Markdown

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

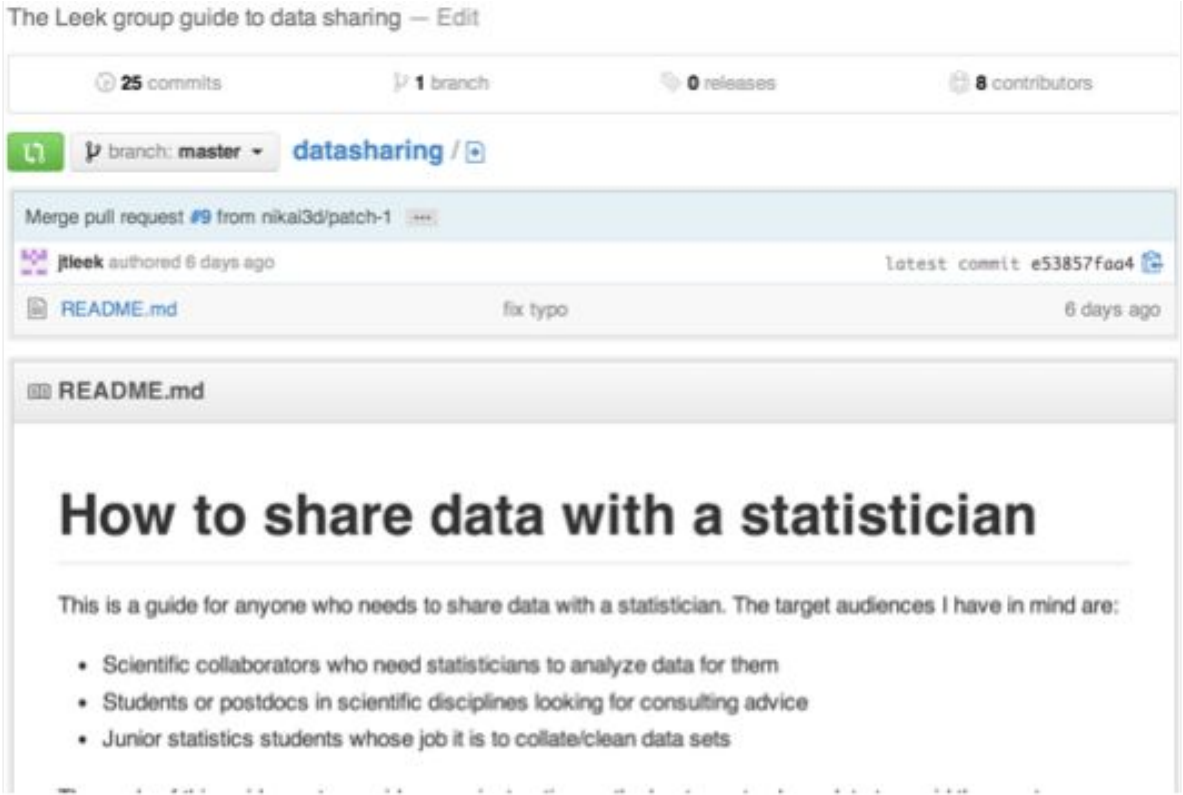


Explicit instructions
Versions of software
Parameters included





Vague instructions
Missing versions
Skipped steps



Reference: https://github.com/jtleek/datasharing

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
  - 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

R code/

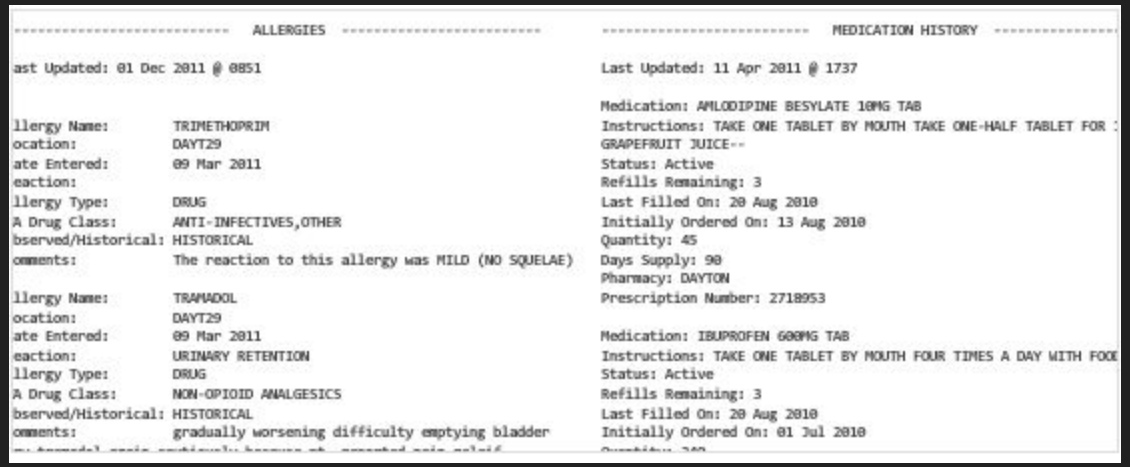
0-preprocess.R

1-explore.R

2-model.R

3-final-plots.R

Raw data

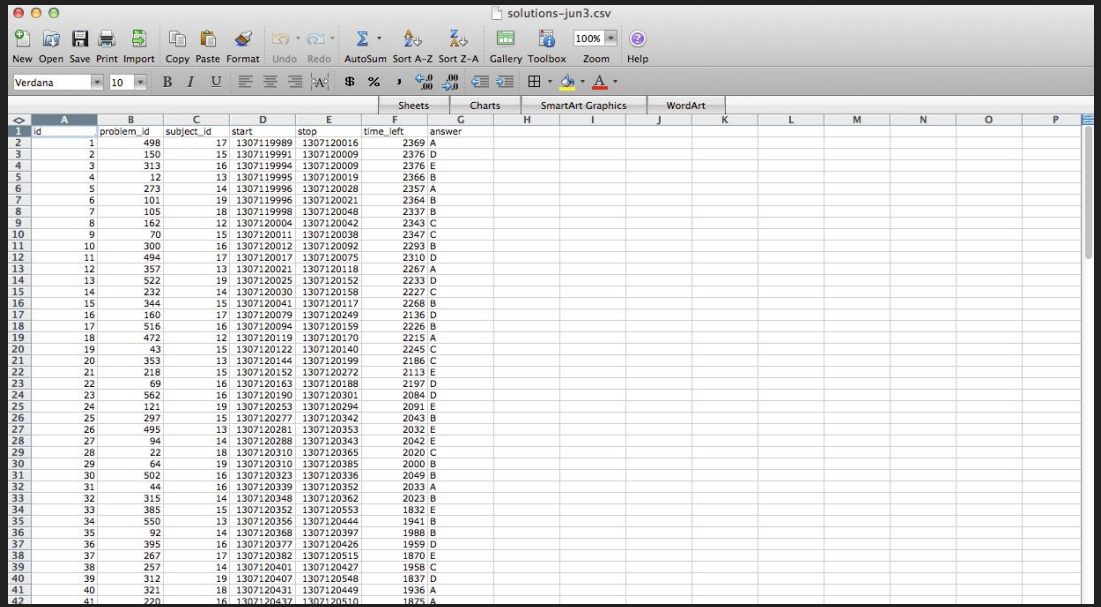


```
---------- ALLERGIES ----------

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
```

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

Processed data



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

- 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

index.Rmd × cheung.R ×

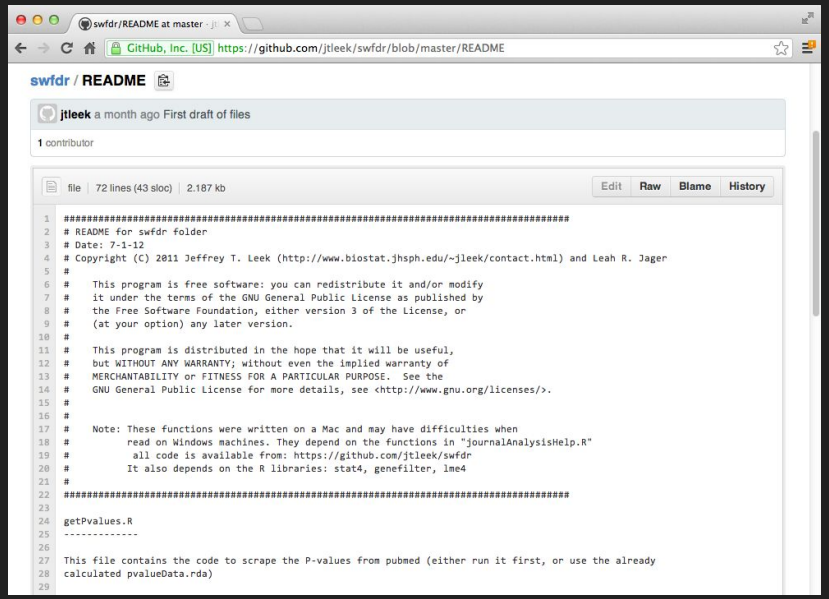
Source on Save | Run | Source

```r
f.pvalue <- function(dat,mod,mod0){
  # This is a function for performing
  # parametric f-tests on the data matrix
  # dat comparing the null model mod0
  # to the alternative model mod.
  n <- dim(dat)[2]
  m <- dim(dat)[1]
  df1 <- dim(mod)[2]
  df0 <- dim(mod0)[2]
  p <- rep(0,m)
  Id <- diag(n)

  resid <- dat %*% (Id - mod %*% solve(t(mod) %*% mod) %*% t(mod))
  resid0 <- dat %*% (Id - mod0 %*% solve(t(mod0) %*% mod0) %*% t(mod0))

  rss1 <- resid^2 %*% rep(1,n)
  rss0 <- resid0^2 %*% rep(1,n)

  fstats <- ((rss0 - rss1)/(df1-df0))/(rss1/(n-df1))
  p <-  1-pf(fstats,df1=(df1-df0),df2=(n-df1))
  return(p)
}

setwd("cheung/")
# Load data and create group variable
dat <- read.table("full.data")

jpt.names <- scan("JPT.cname.txt",what="character")
chb.names <- scan("CHB.cname.txt",what="character")
ceu.names <- scan("CEU_parents.txt",what="character")
nceu <- length(ceu.names)
njpt <- length(jpt.names)
nchb <- length(chb.names)
```

1:1 f.pvalue R Script

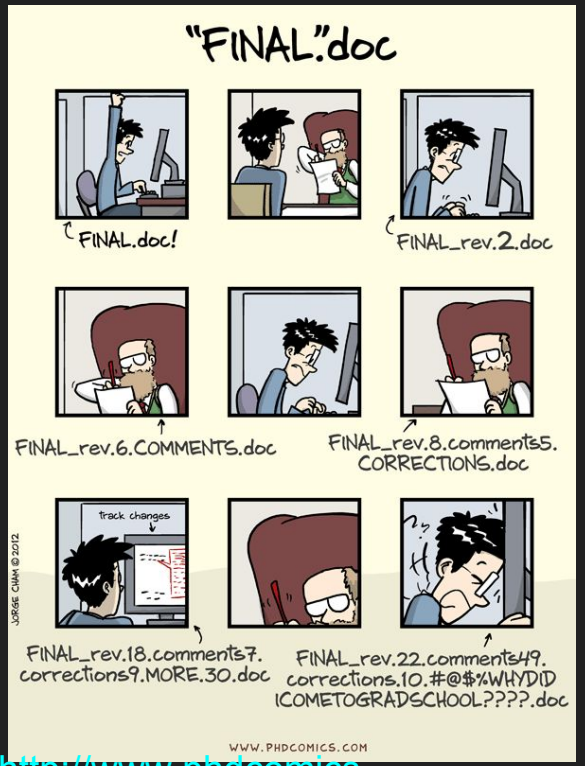
- Clearly commented
  - Small comments liberally - what, when, why, how
  - Bigger commented blocks for whole sections
- Include processing details
- Only analyses that appear in the final write-up

README file

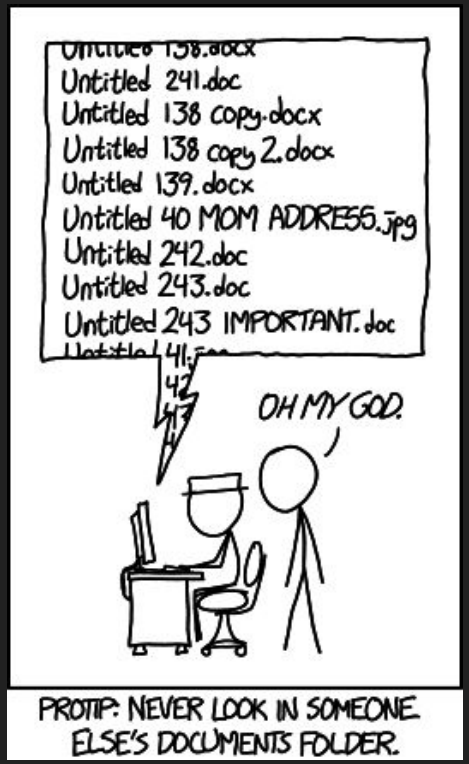


- Should contain step-by-step instructions for analysis
- Here is an example https://github.com/jtleek/swfdr/blob/master/README.md

Just no



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



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

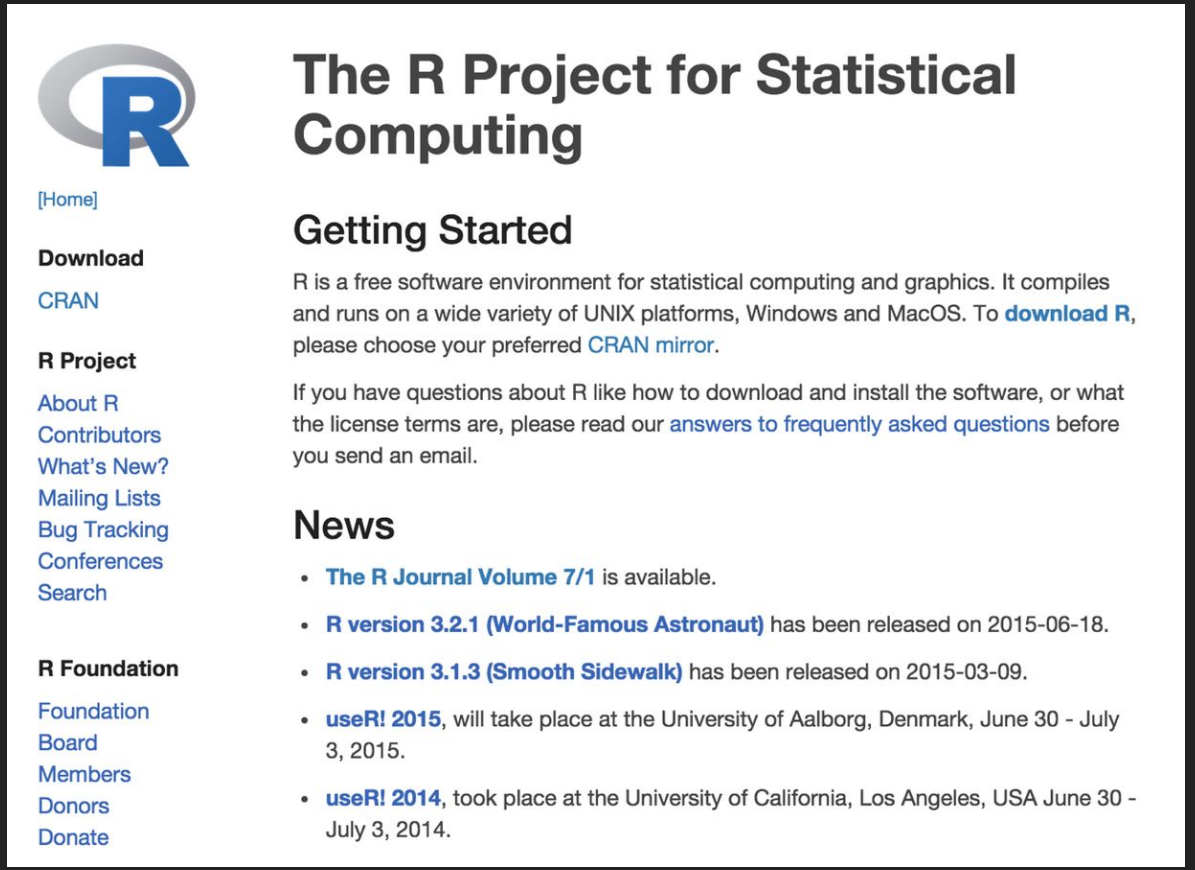
(http://kbroman.org/Tools4RR/assets/lectures/06_org_eda.pdf)

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

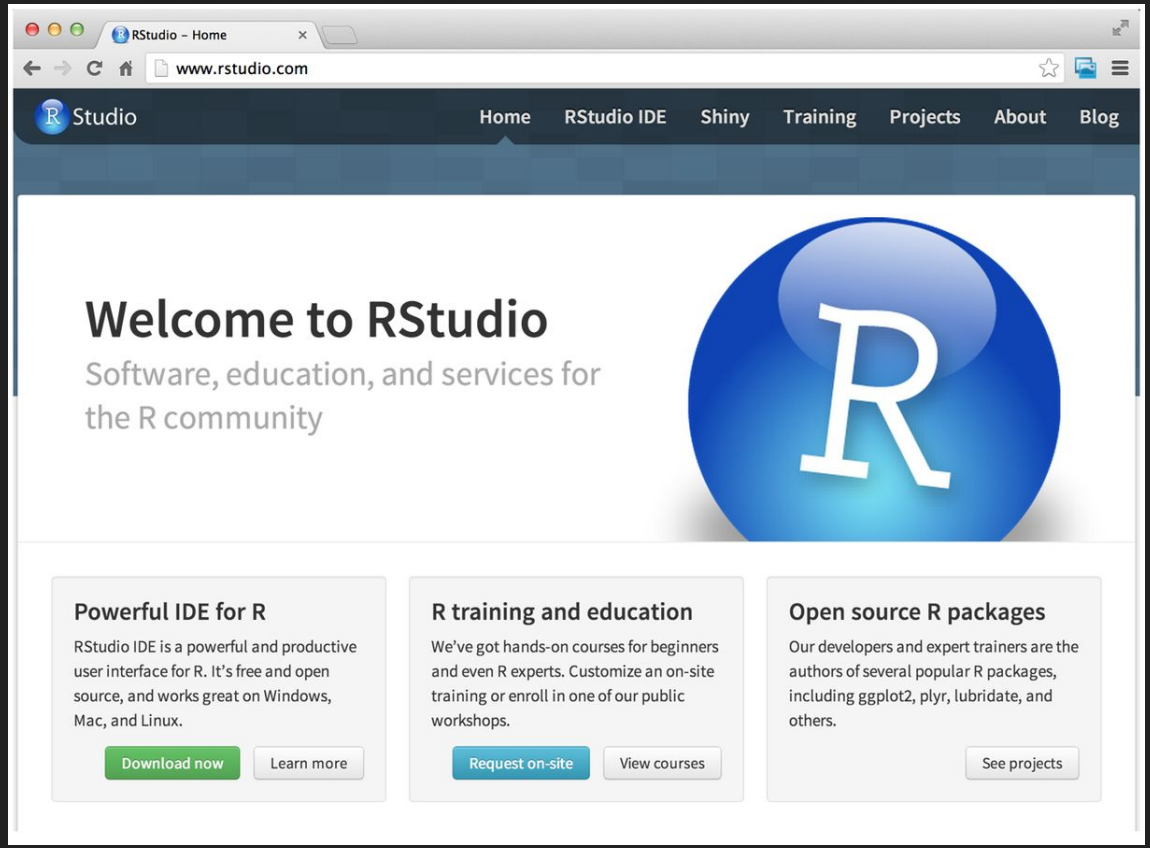
- Karl Broman

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

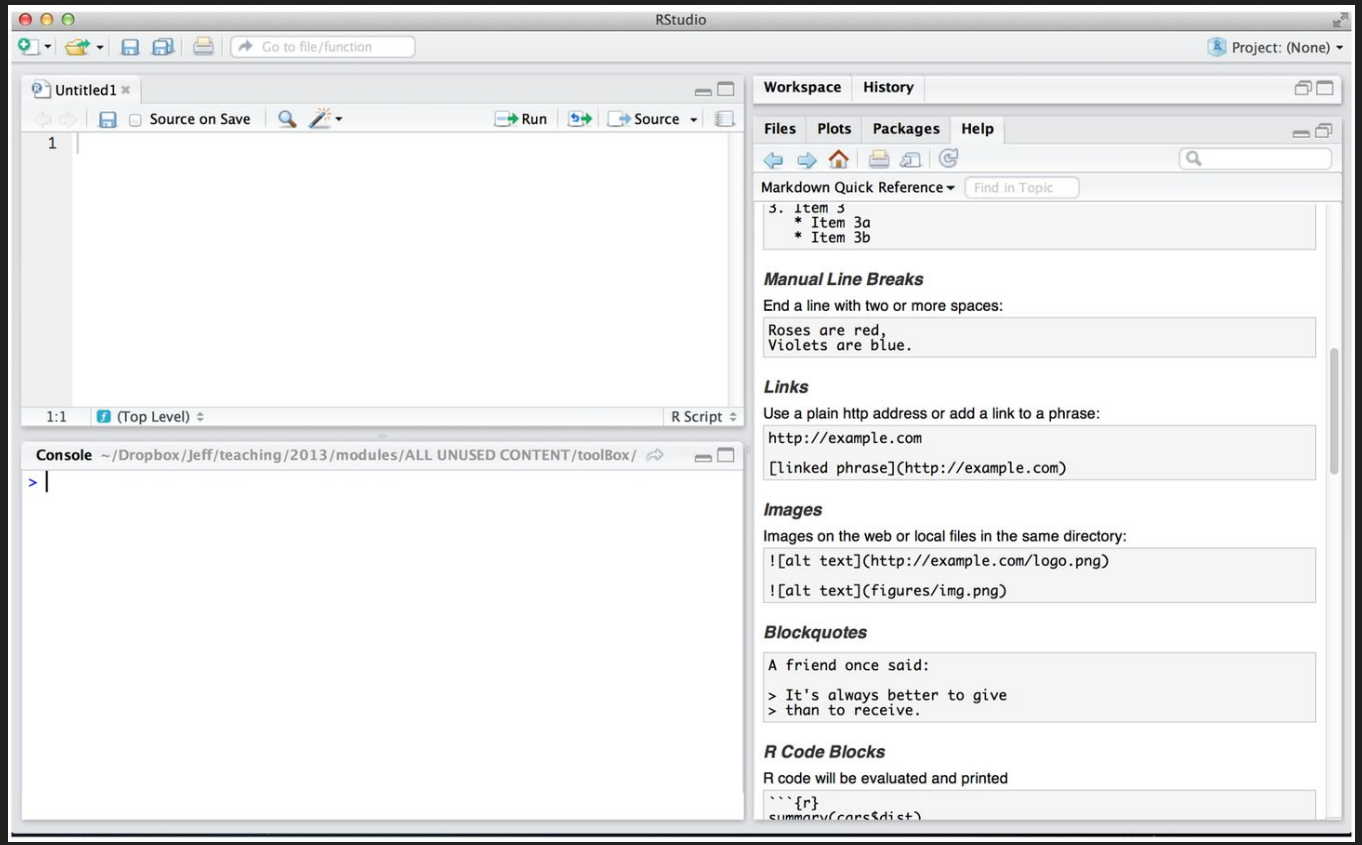
R + Rstudio



www.r-project.org



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

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

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

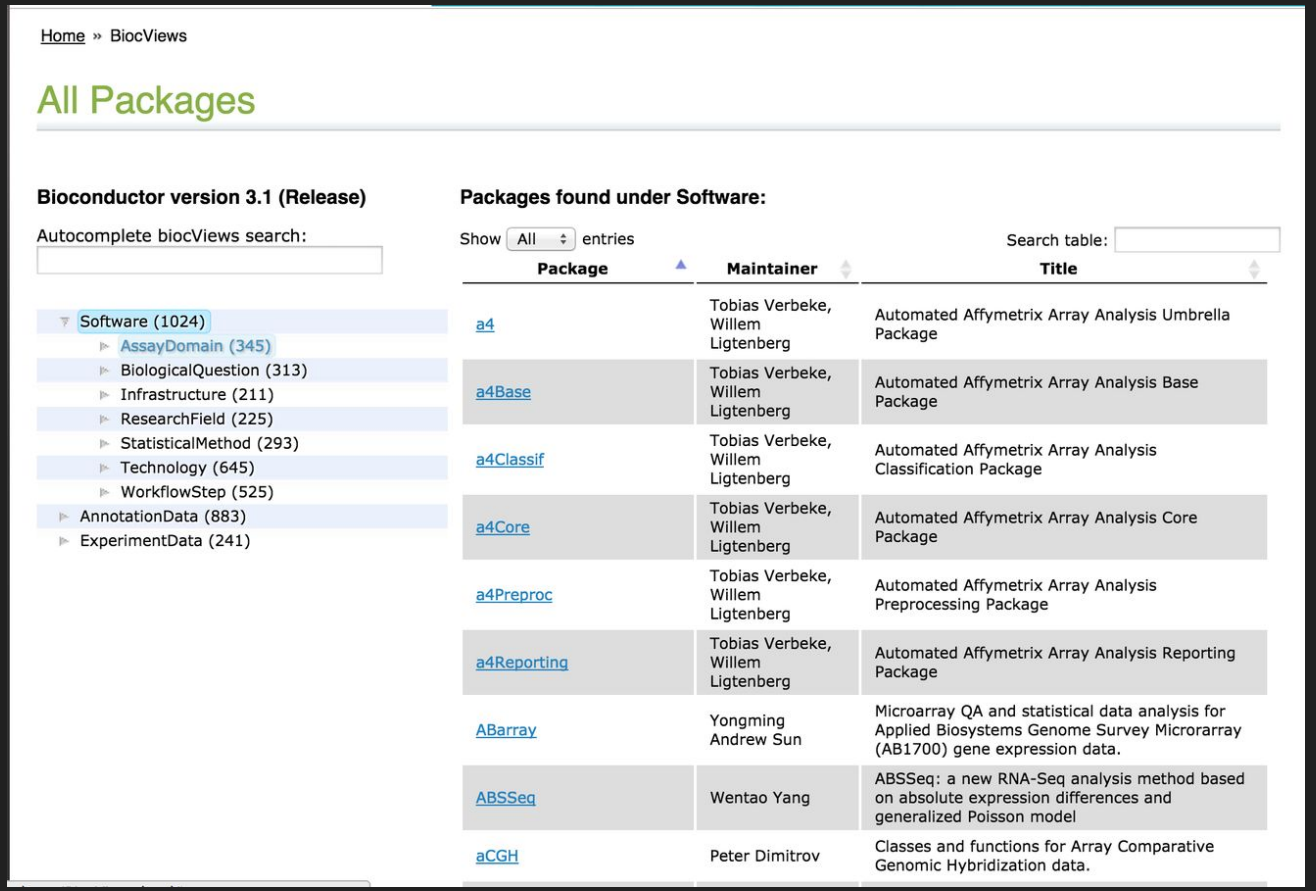
Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

A3	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”)
```



http://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

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

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:

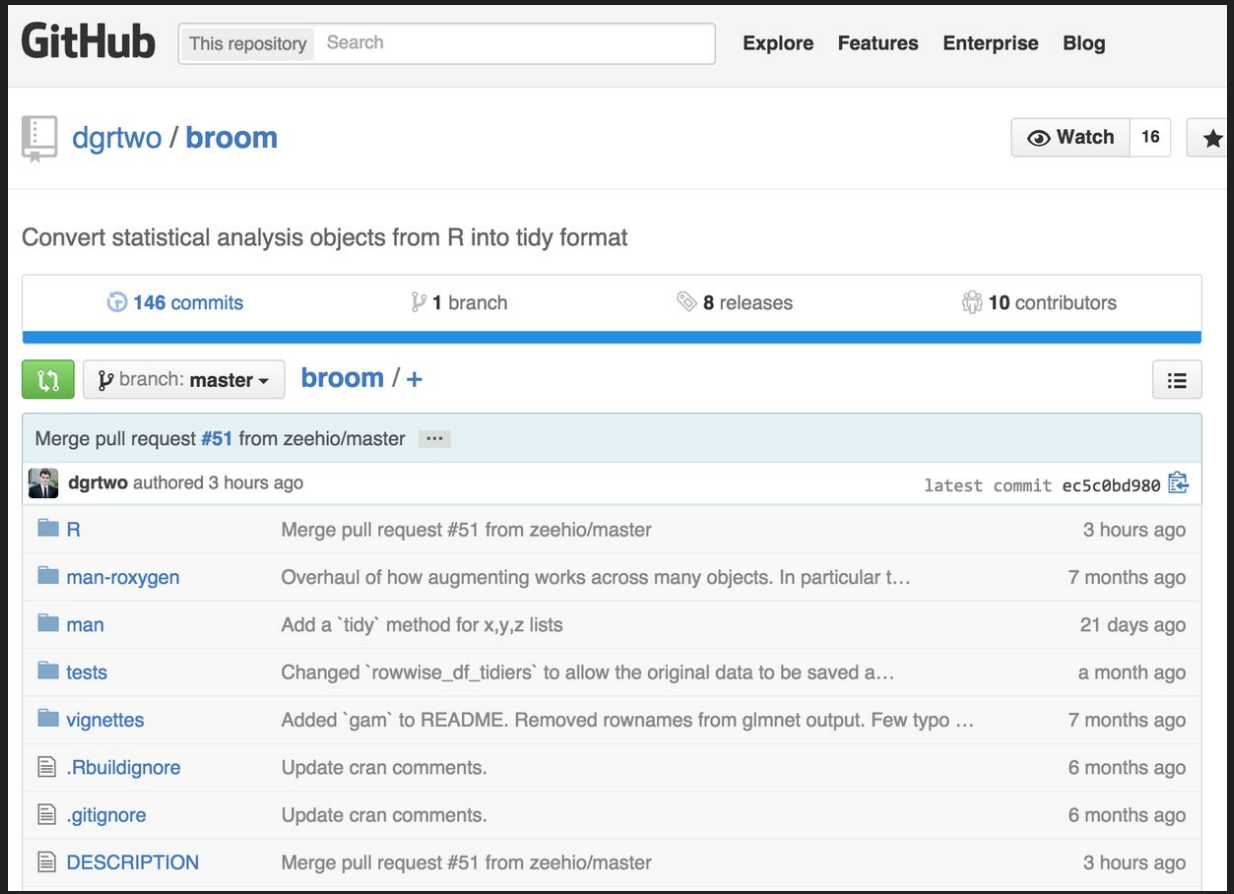
- gonucleotide Arrays
- High-throughput Sequencing
- Counting Reads for Differential Expression (parathyroideSE vignette)
- Annotation
- Annotating Variants
- Annotating Ranges
- Flow Cytometry and other assays
- Candidate Binding Sites for Known Transcription Factors
- Cloud-enabled cis-eQTL search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

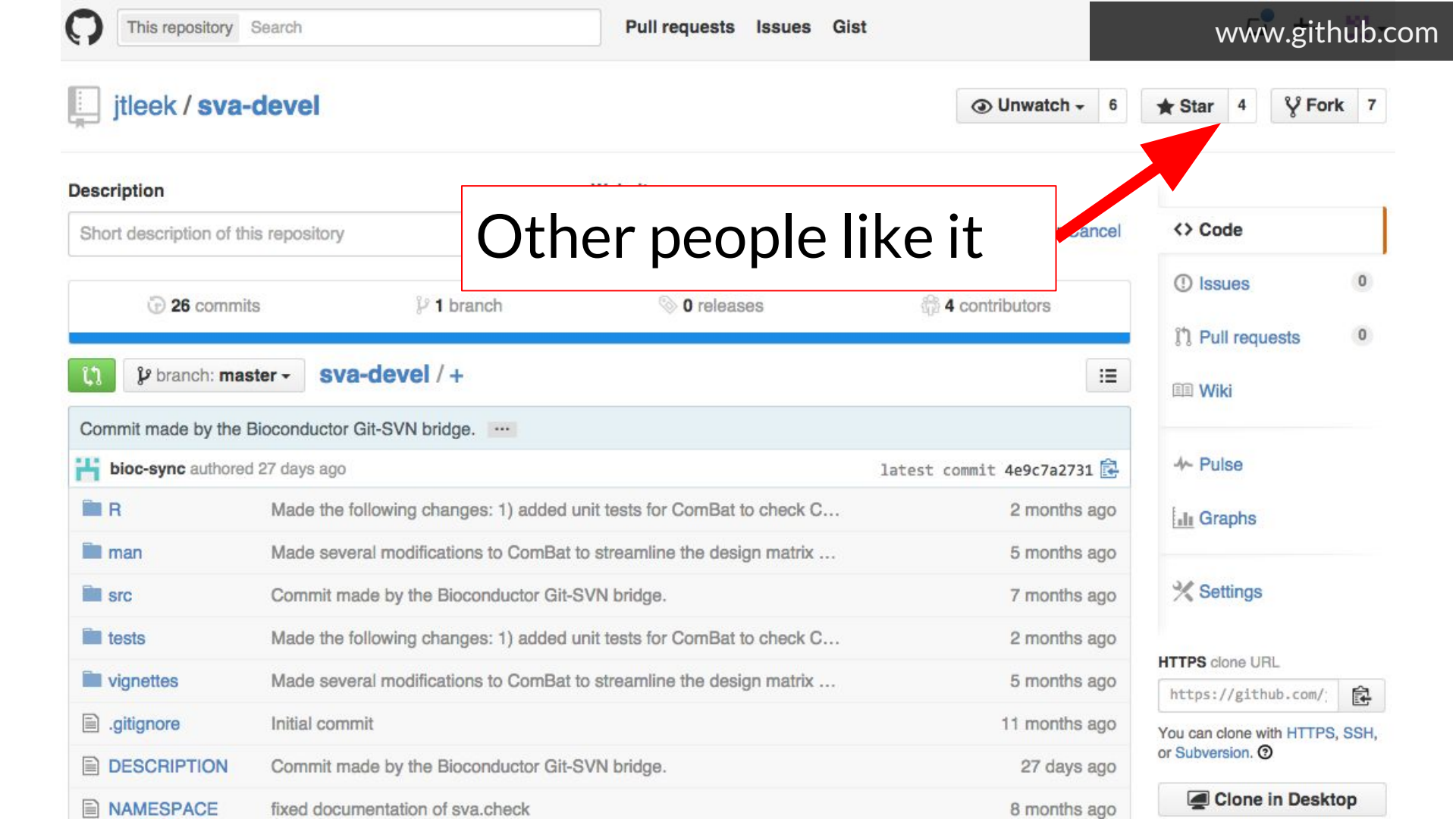
Post questions about Bioconductor packages to our mailing lists. Read the posting guide before posting!

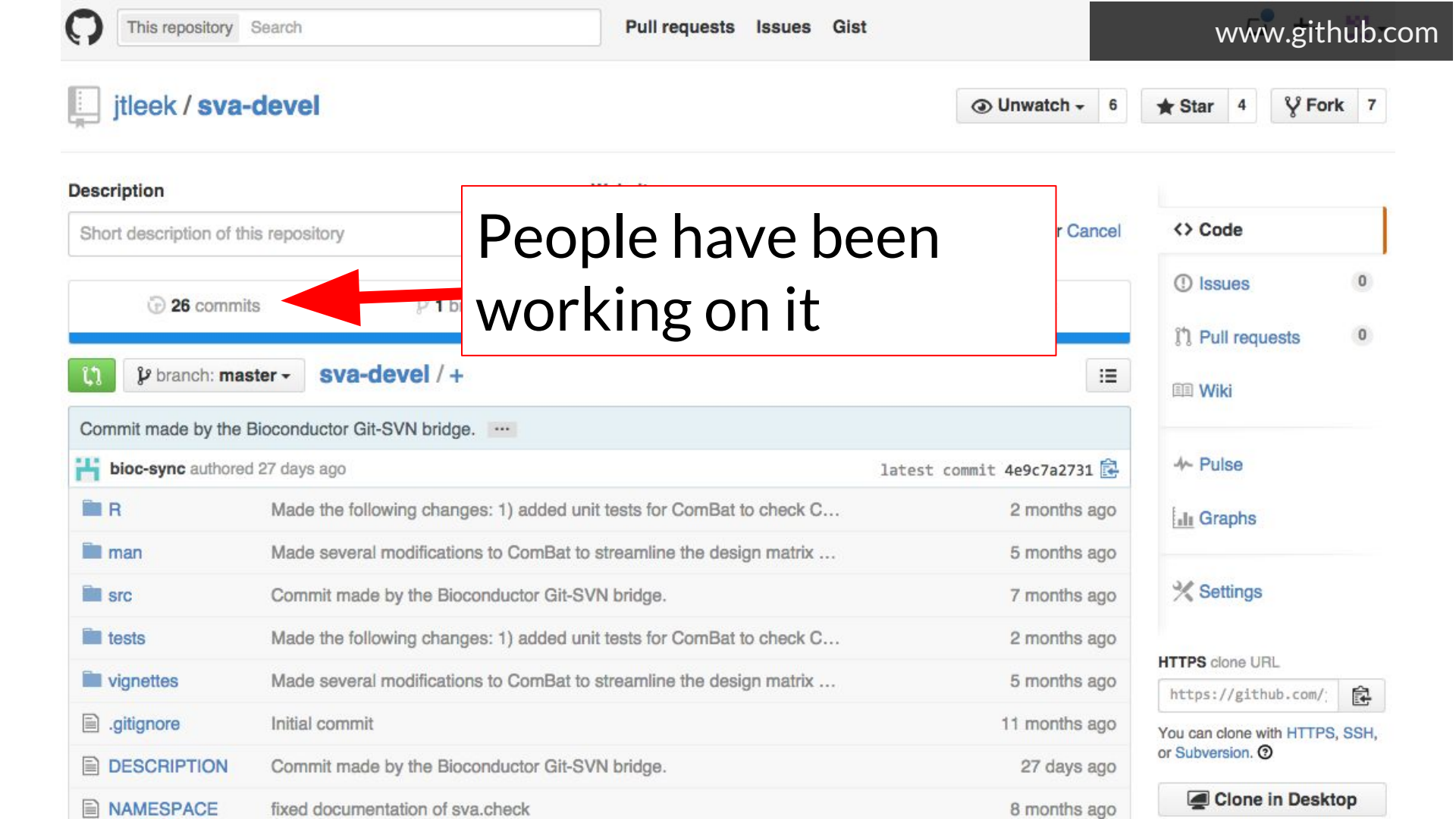
- bioconductor
- bioc-devel

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



https://github.com/dgrtwo/broom


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
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://goo.gl/SNeI00