SISBD Module 1

Accessing Biomedical Data

Jeff Leek (@jtleek)

Raphael Gottardo (@raphg)

Preliminaries

Raphael Gottardo
@raphg

Follow

Register for Summer Institute in Statistics for Big Data ow.ly/KNjBw with @rdpeng @hadleywickham @raphg and the handsome @jtleek

1:52 PM - 25 Mar 2015

11 21

Class name:	Accessing Biomedical Big Data
Instructors:	Raphael Gottardo , Jeff Leek
TAs:	Jean Morrison, Brian Williamson
Course Website:	http://sisbid.github.io/Module1
Goal:	Teach you how to get and clean data
Pre-reqs:	Hopefully some R programming
Where to get slides:	https://github.com/SISBID/Module1

Motivating Example

An exciting result

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 oncogenic pathway deregulation to identify new therapeutic strateg

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

of gene expression profiles that can predict response to

Stunning problems

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

Timeline of events

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.

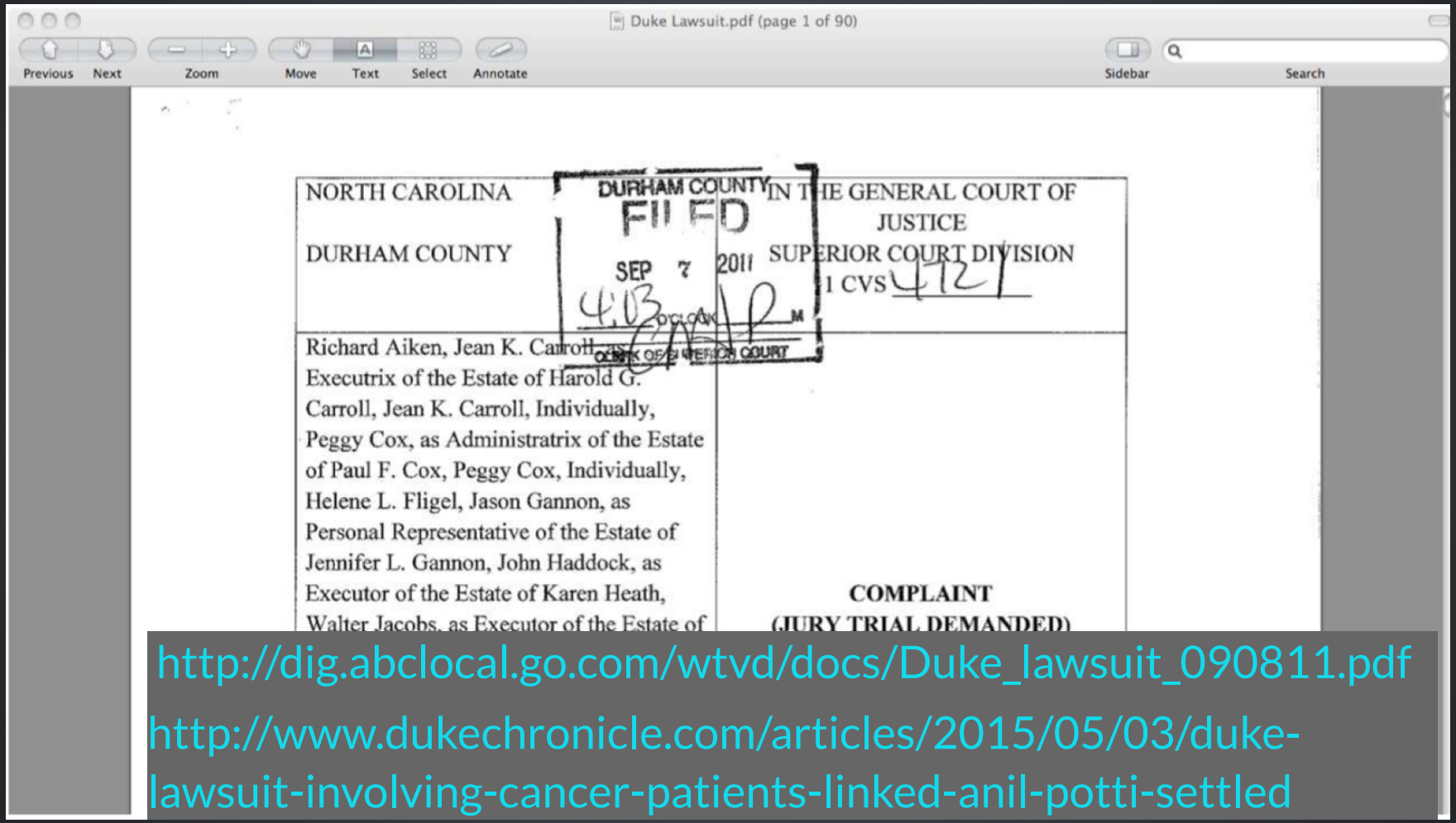
NOVEMBER 2010 *JCO* paper is retracted. Duke closes the trials permanently. Potti resigns.

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

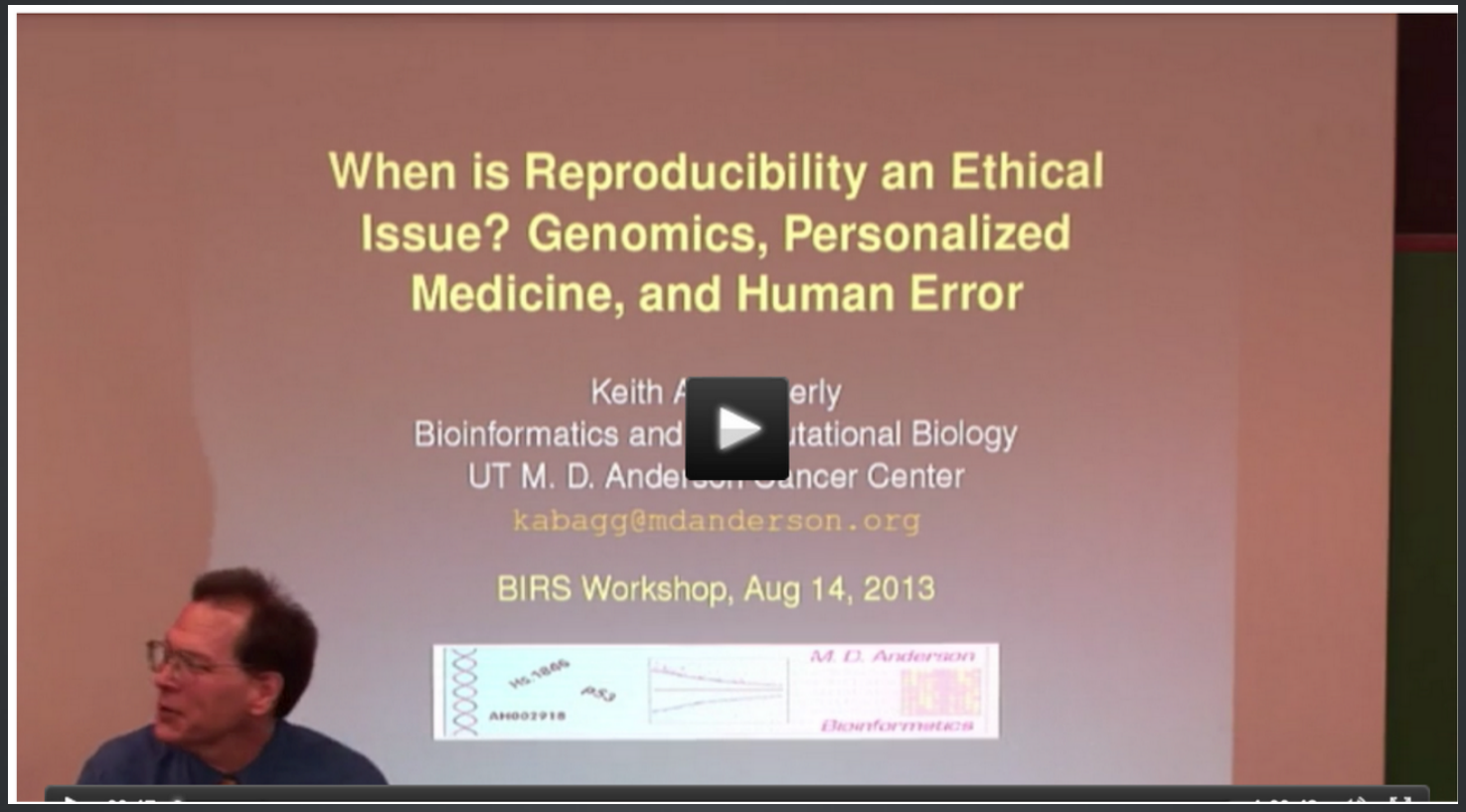
Major fallout



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

An interesting talk



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

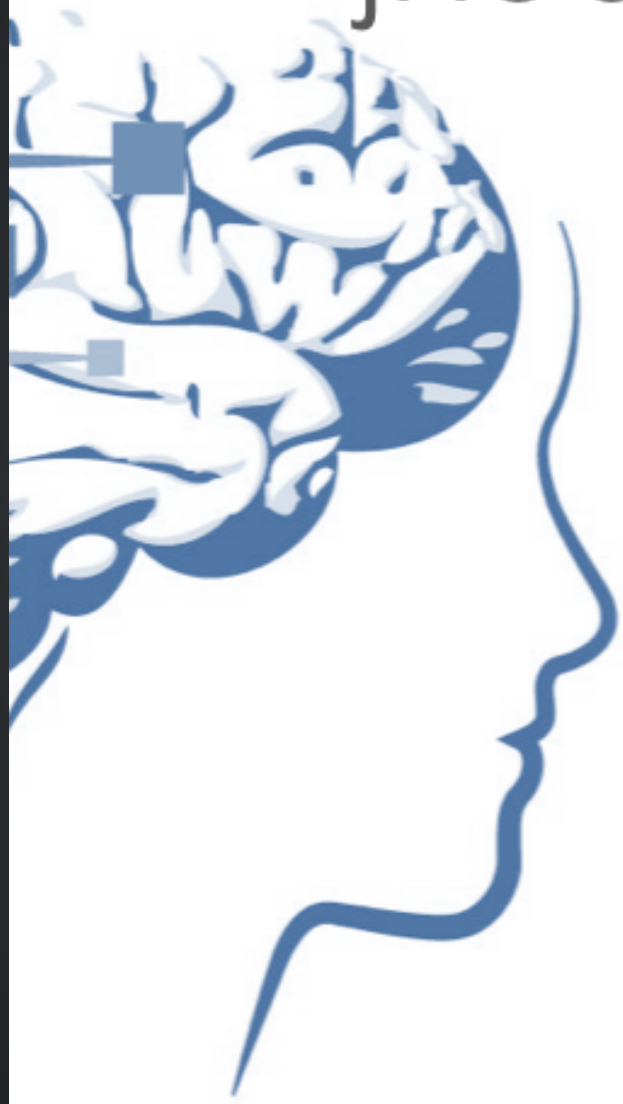
About me


How you feel about statistics


How I feel about statistics

@simplystats

jhudatascience.org



jtleek

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

- **Home**
- Alumni
- Books
- Data
- Jobs
- Papers
- People
- Software
- Talks
- Teaching

www.jtleek.com

Hi I'm Jeff

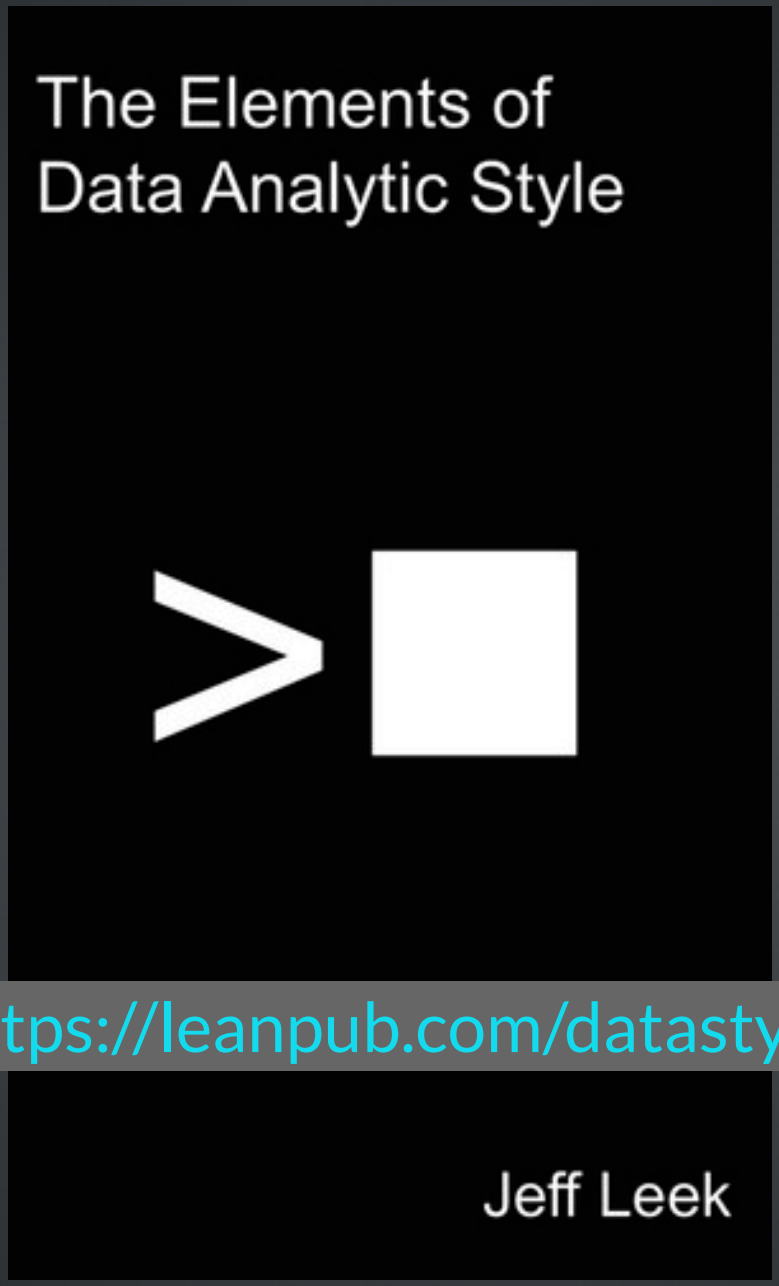
I work on figuring out how to go from raw data from next generation sequencing machines to results, turning public genomic data into clinically useful tools, and understanding how people use data analysis in real life.

I do statistical research, write data analysis software, curate and create data sets, write a blog about statistics, teach people here at Hopkins, teach a lot of people online, and work with amazing students who go do awesome things. If you want to, come do stuff with me

If you want to keep up with everything we are working on, follow me on Twitter @jtleek. The best way to contact me is my gmail account (I do not check my JHU email at all), or you can call me at my office **410-955-1166** (fair warning I have answered that phone ~3 times total since 2009), send me a fax **410-955-0958** (for real, fax is still a thing?!), or if you still use the pony express you could send me a letter at:

Johns Hopkins University
Bloomberg School of Public Health
Office E3624
615 North Wolfe Street
Baltimore, MD 21205-2179


$\frac{\hat{Y} - \mu}{\sigma_Y / \sqrt{N}}$
"Genomic data science"



https://leanpub.com/datastyle

coursera

Catalog | Search catalog

Institutions | Jeff Leek

Edit Course Description | Edit Session Descriptions | Edit Session Materials

JOHNS HOPKINS
BLOOMBERG SCHOOL of PUBLIC HEALTH

Getting and Cleaning Data

Part of the Data Science Specialization »

Learn how to gather, clean, and manage data from a variety of sources. This is the third course in the Johns Hopkins Data Science Specialization.

About the Course

Before you can work with data you have to get some. This course will cover the basic ways that data can be obtained. The course will cover obtaining data from the web, from APIs, from databases and from colleagues in various formats. It will also cover the basics of data cleaning and how to make data "tidy". Tidy data dramatically speed downstream data analysis tasks. The course will also cover the components of a complete data set including raw data, processing instructions, codebooks, and processed data. The course will cover the basics needed for collecting, cleaning, and sharing data.

Sessions

July 6, 2015 - August 1, 2015

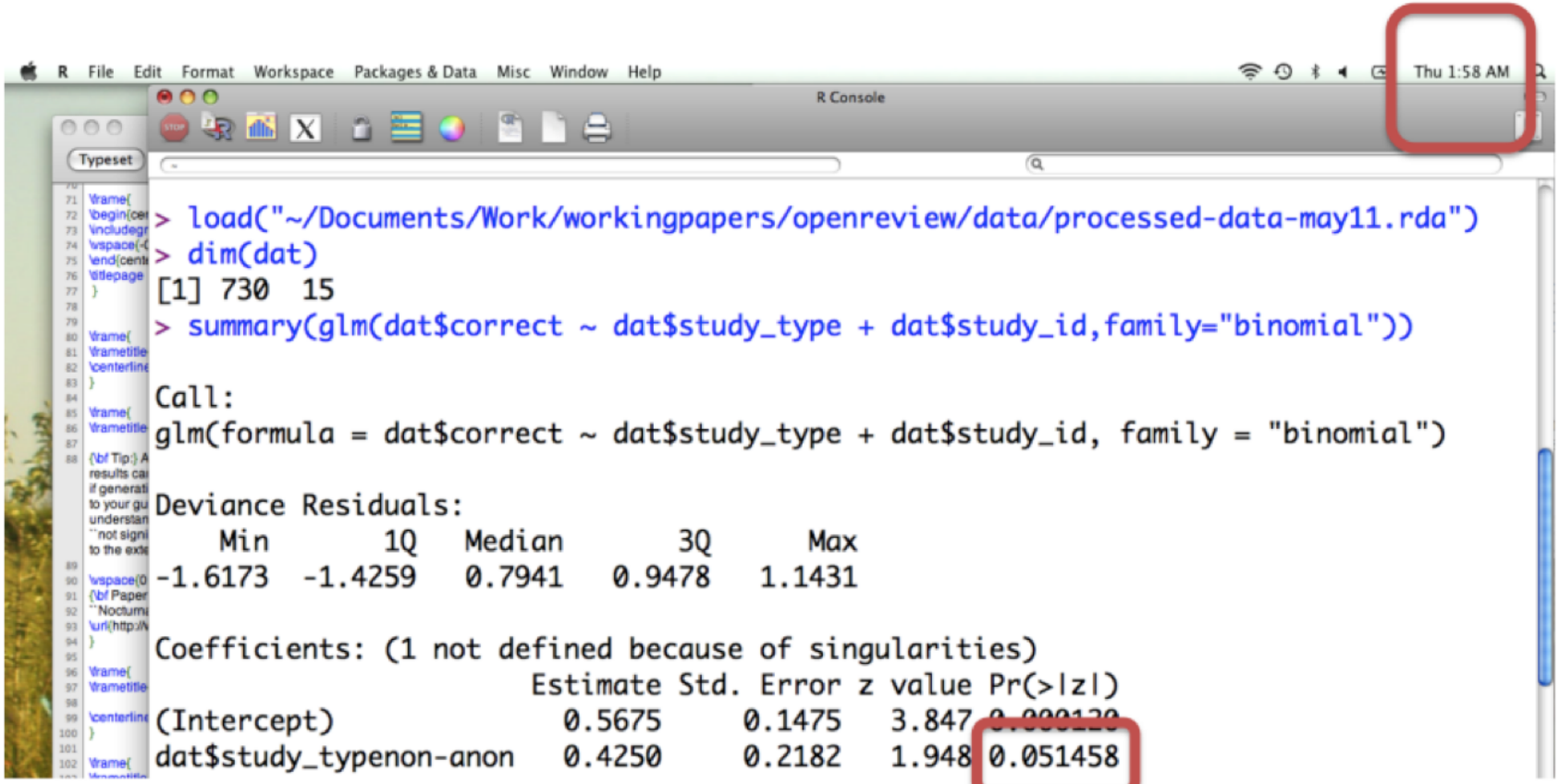
Join Course

Eligible for

Data Science Specialization

This class

What is the temptation?



Why this class?

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.

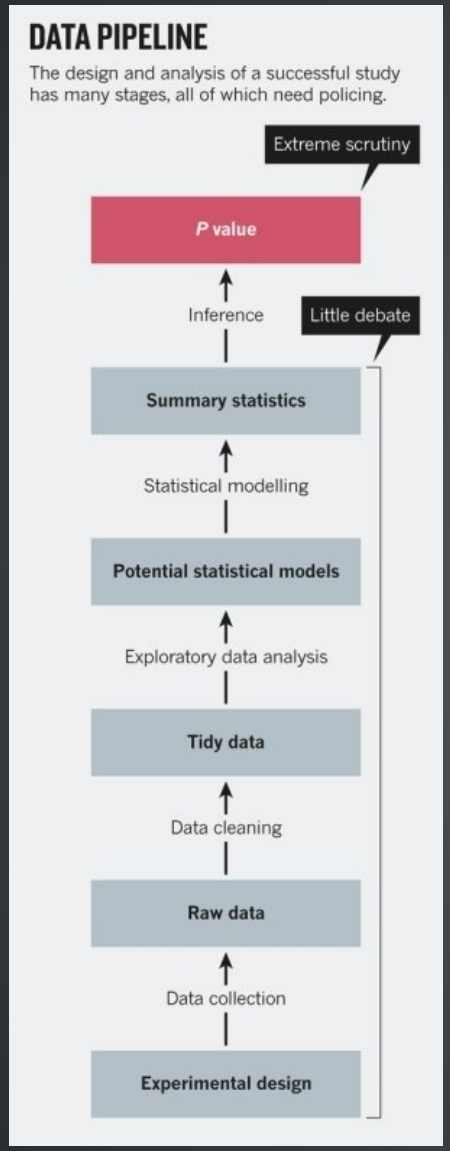
Results

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.

Why this class?

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

http://pubs.acs.org/doi/abs/10.1021/om4000067



Why this class?

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

Leek and Peng (2015) Nature

Researcher degrees of freedom

General Article

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

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

> 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

MOR

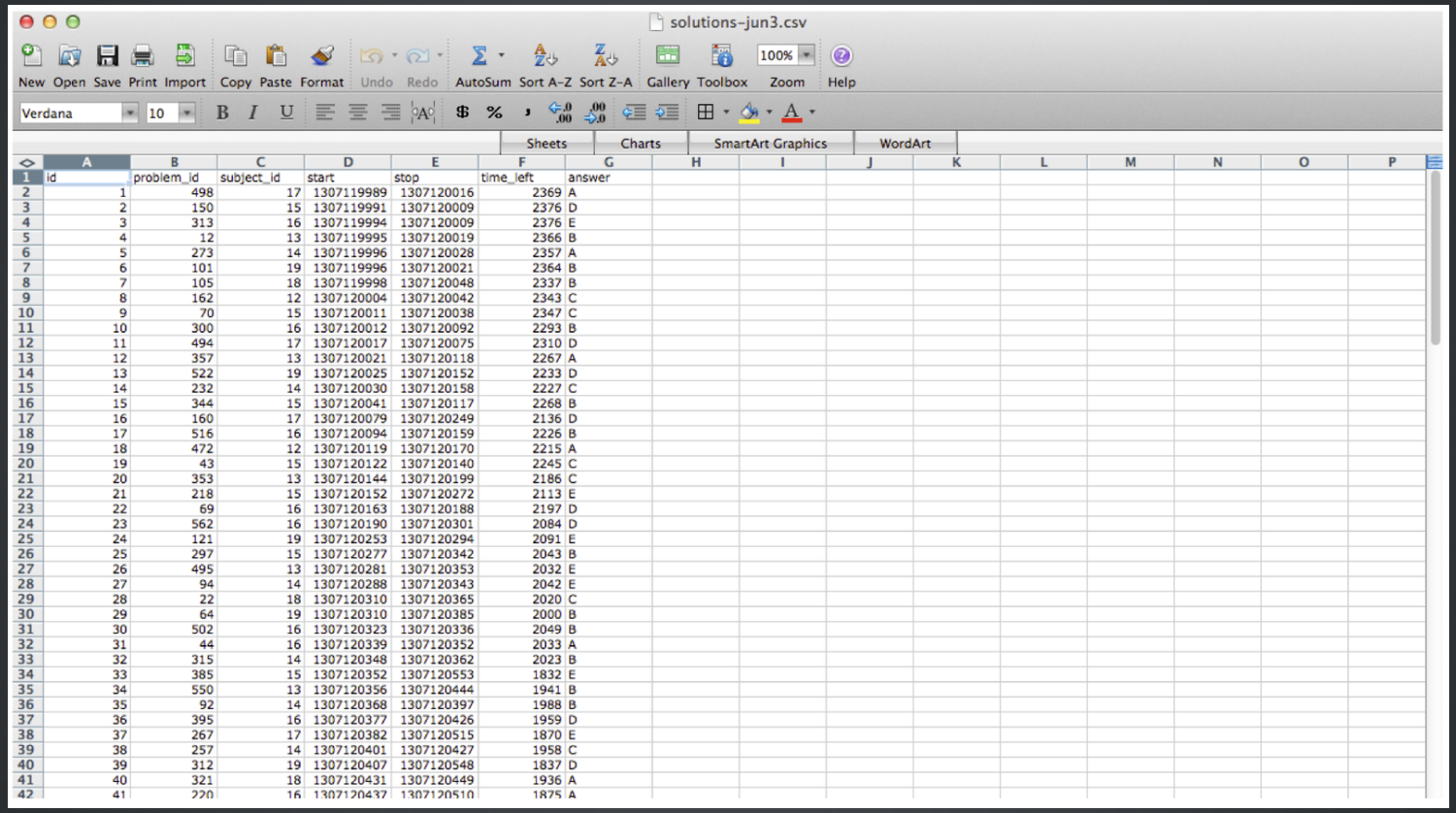
TECHNOLOGY

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

By STEVE LOHR AUG. 17, 2014

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

What you wished data looked like



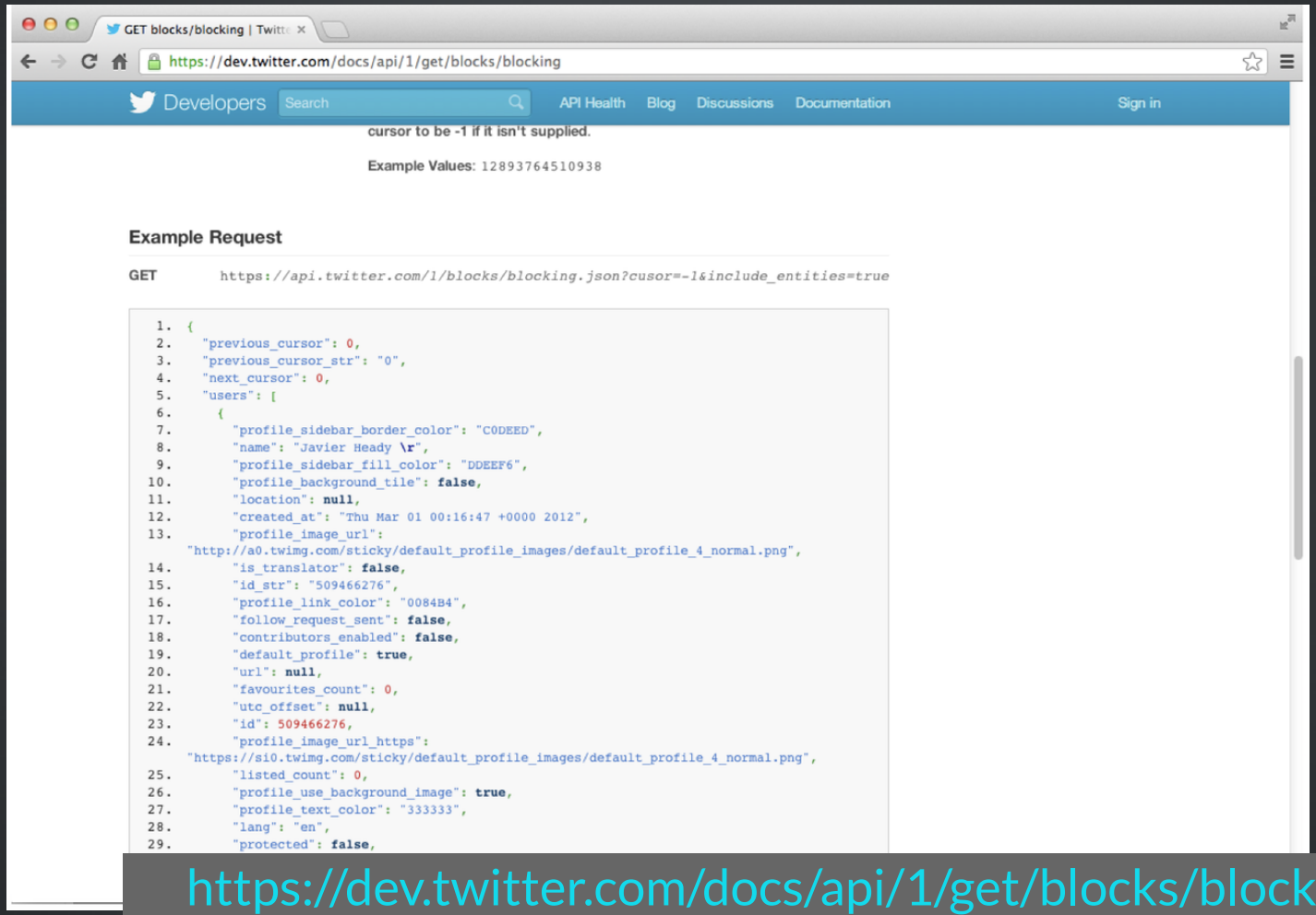
	id	problem_id	subject_id	start	stop	time_left	answer
1	1	498	17	1307119989	1307120016	2369	A
2	2	150	15	1307119991	1307120009	2376	D
3	3	313	16	1307119994	1307120009	2376	E
4	4	12	13	1307119995	1307120019	2366	B
5	5	273	14	1307119996	1307120028	2357	A
6	6	101	19	1307119996	1307120021	2364	B
7	7	105	18	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

What it actually looks like

```
@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
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_[]\ZTDUZZUSOPX]]POP\GS\WSHHD
@HWI-EAS121:4:100:1783:455#0/1
GGGTAATTCAGGGACAATGTAATGGCTGCACAAAAAAATACATCTTTCATGTTCCATTGCACCATTGACAAATACATATT
+HWI-EAS121:4:100:1783:455#0/1
abb_babbabaabbbbbbbbbbbbbbba\`b`\abbbabbbbabbbbbbaabbbbb`bb`ab_O_bab_Q_bbabaa_a
```

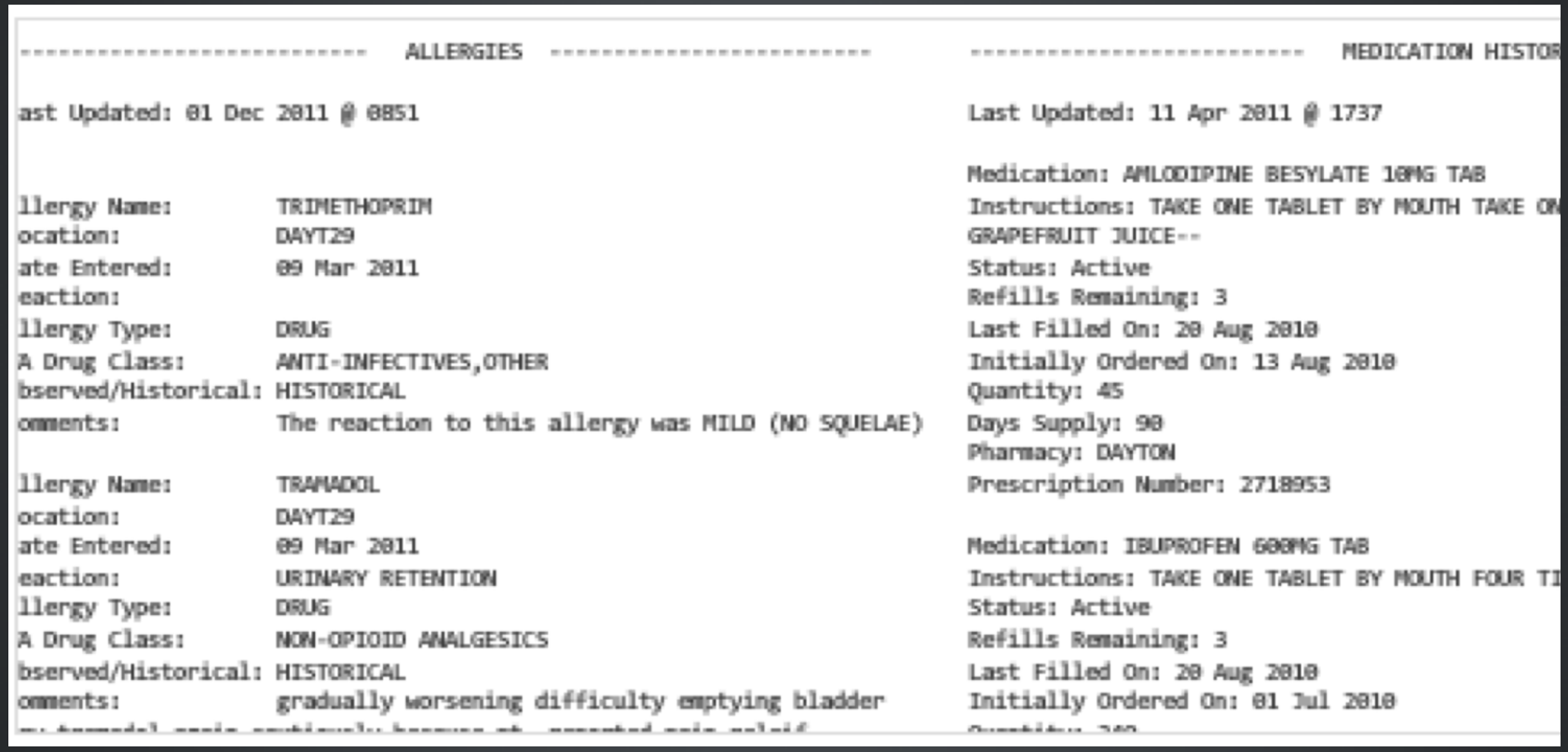
http://brianknaus.com/software/srtoolbox/s_4_1_sequence80.txt
```

What it actually looks like



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

What it actually looks like



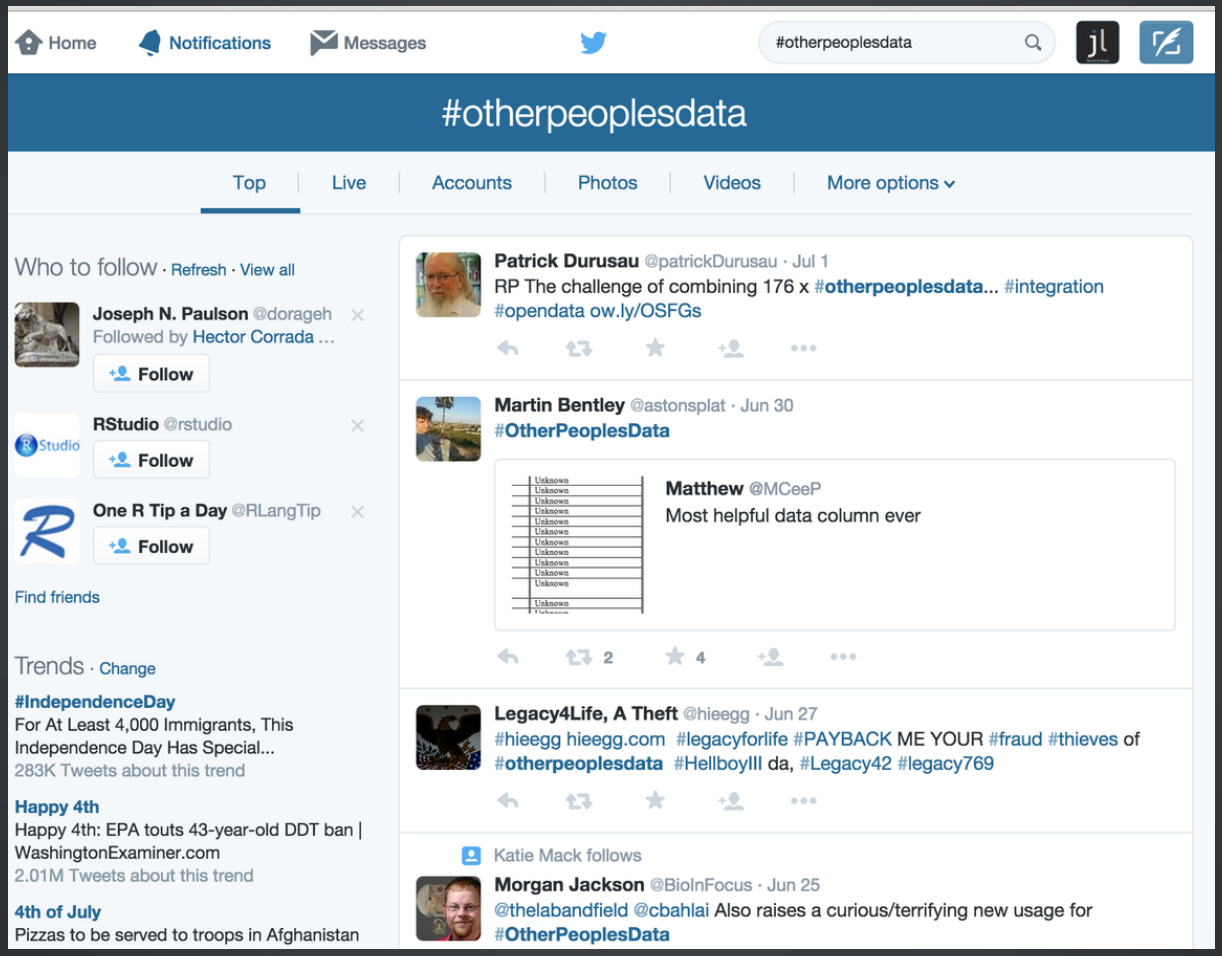
```
-------------------------   ALLERGIES   -------------------------      -------------------------   MEDICATION HISTOR

ast Updated: 01 Dec 2011 @ 0851                                       Last Updated: 11 Apr 2011 @ 1737

                                                                      Medication: AMLODIPINE BESYLATE 10MG TAB
llergy Name:          TRIMETHOPRIM                                    Instructions: TAKE ONE TABLET BY MOUTH TAKE ON
ocation:              DAYT29                                          GRAPEFRUIT JUICE--
ate Entered:          09 Mar 2011                                     Status: Active
eaction:                                                              Refills Remaining: 3
llergy Type:          DRUG                                            Last Filled On: 20 Aug 2010
A Drug Class:         ANTI-INFECTIVES,OTHER                           Initially Ordered On: 13 Aug 2010
bserved/Historical:   HISTORICAL                                      Quantity: 45
omments:              The reaction to this allergy was MILD (NO SQUELAE)  Days Supply: 90
                                                                      Pharmacy: DAYTON
llergy Name:          TRAMADOL                                        Prescription Number: 2718953
ocation:              DAYT29
ate Entered:          09 Mar 2011                                     Medication: IBUPROFEN 600MG TAB
eaction:              URINARY RETENTION                               Instructions: TAKE ONE TABLET BY MOUTH FOUR TI
llergy Type:          DRUG                                            Status: Active
A Drug Class:         NON-OPIOID ANALGESICS                           Refills Remaining: 3
bserved/Historical:   HISTORICAL                                      Last Filled On: 20 Aug 2010
omments:              gradually worsening difficulty emptying bladder  Initially Ordered On: 01 Jul 2010
```

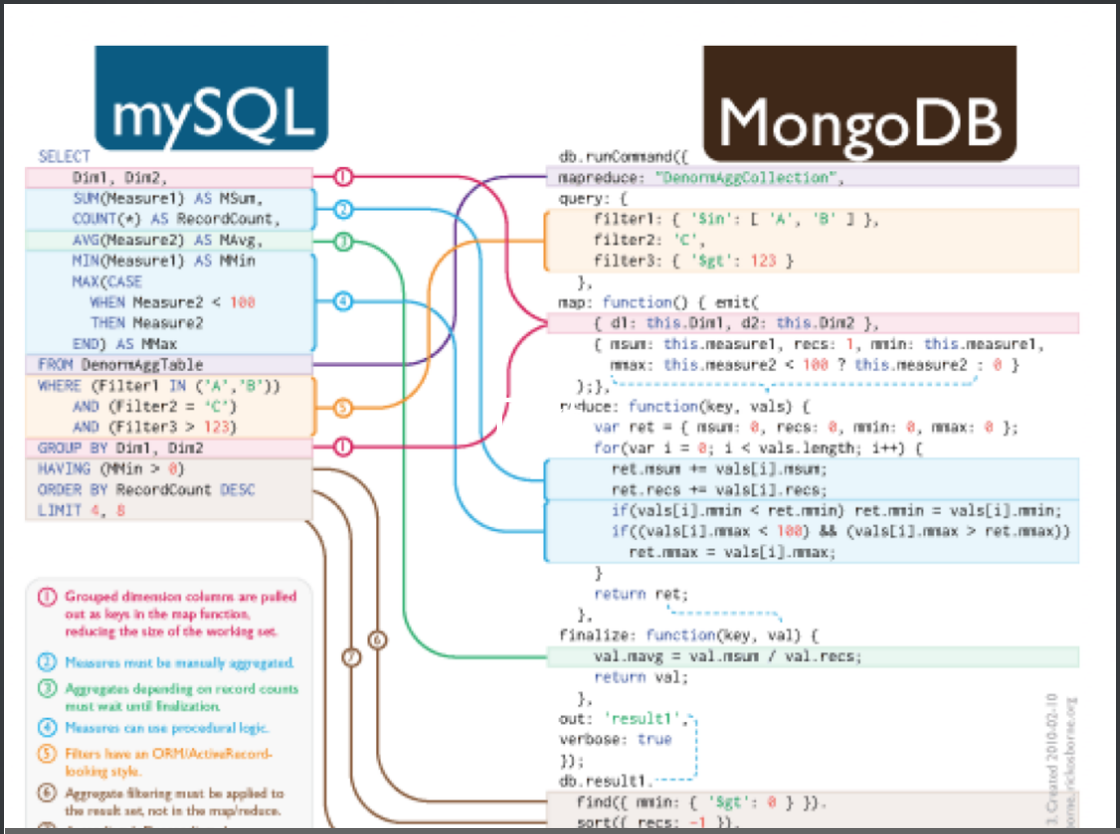
http://blue-button.github.com/challenge/

#otherpeoplesdata



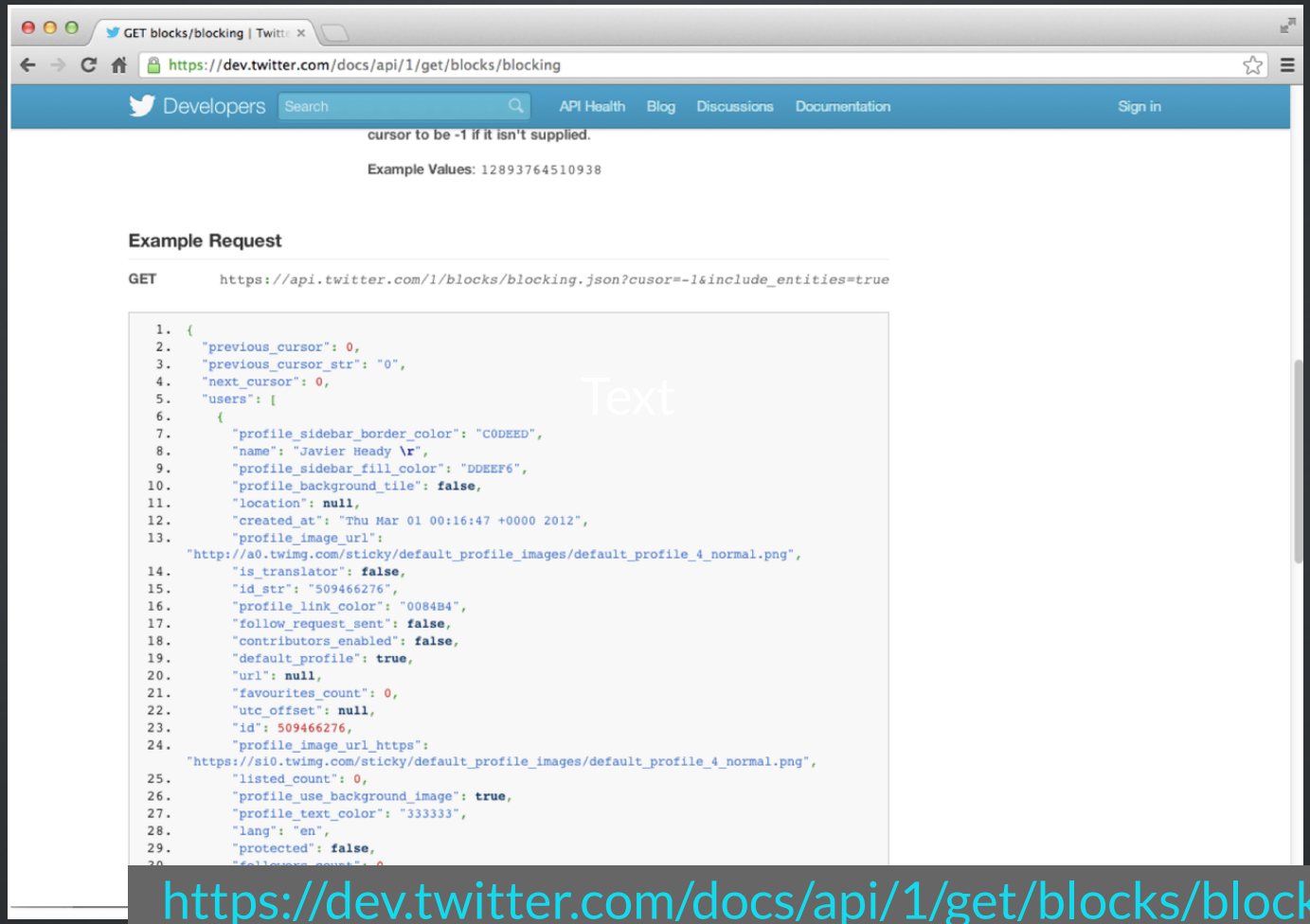

Where you wish data was

Where is data?



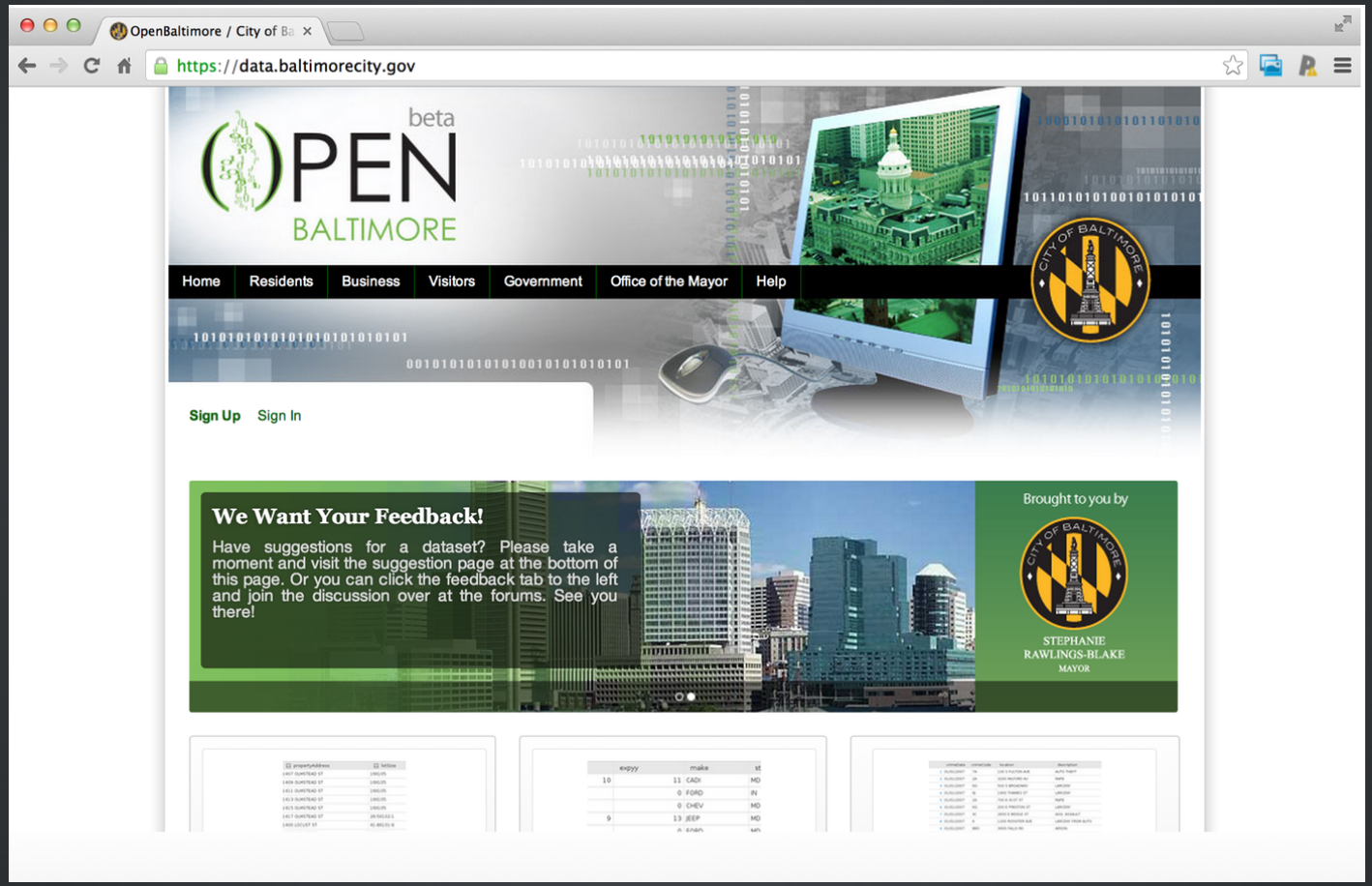
http://rickosborne.org/blog/2010/02/infographic-migrating-from-sql-to-mapreduce-with-mongodb/

Where is data?



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

Where is data?



https://data.baltimorecity.gov/

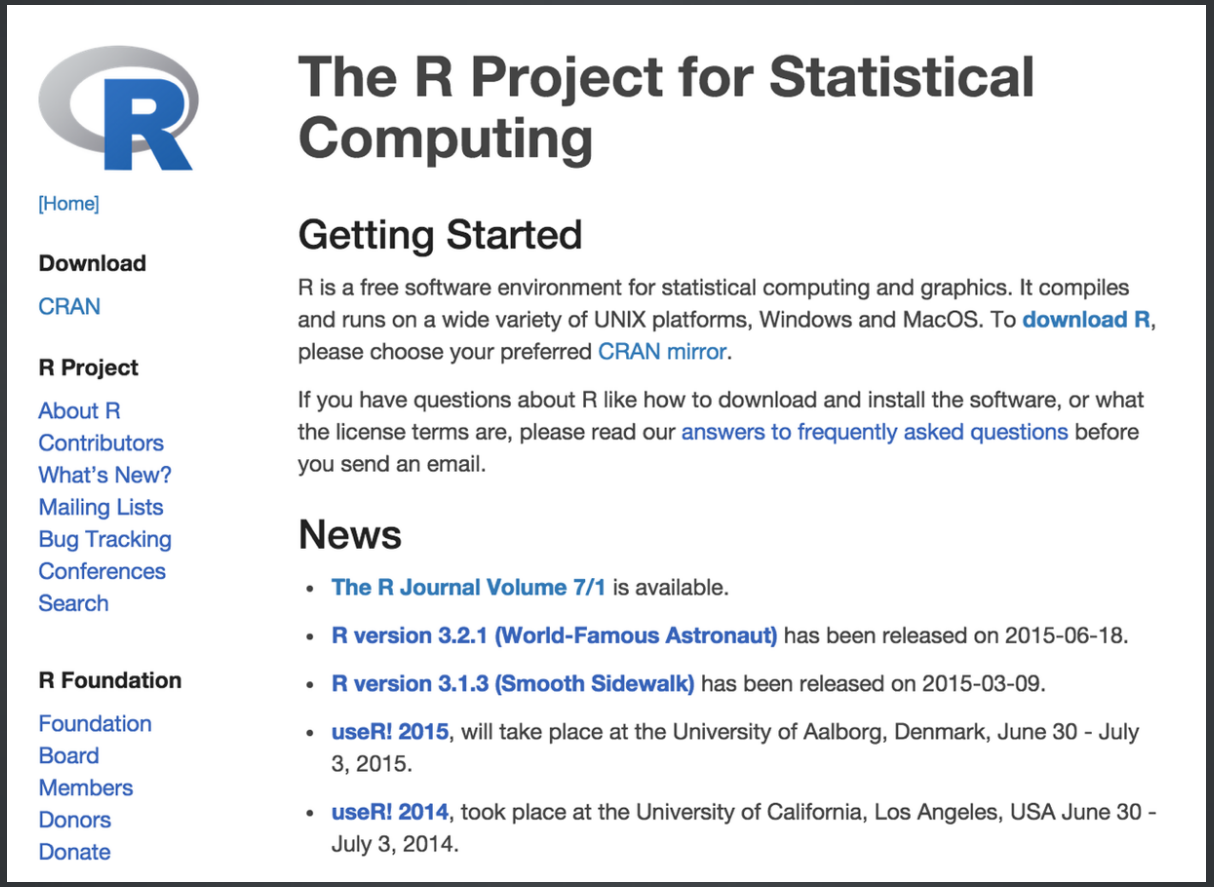

GET ALL THE DATA!!!
memegenerator.net

Data Source Brainstorming

http://bit.ly/1FWssE2

Tools

The workhorse



[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.

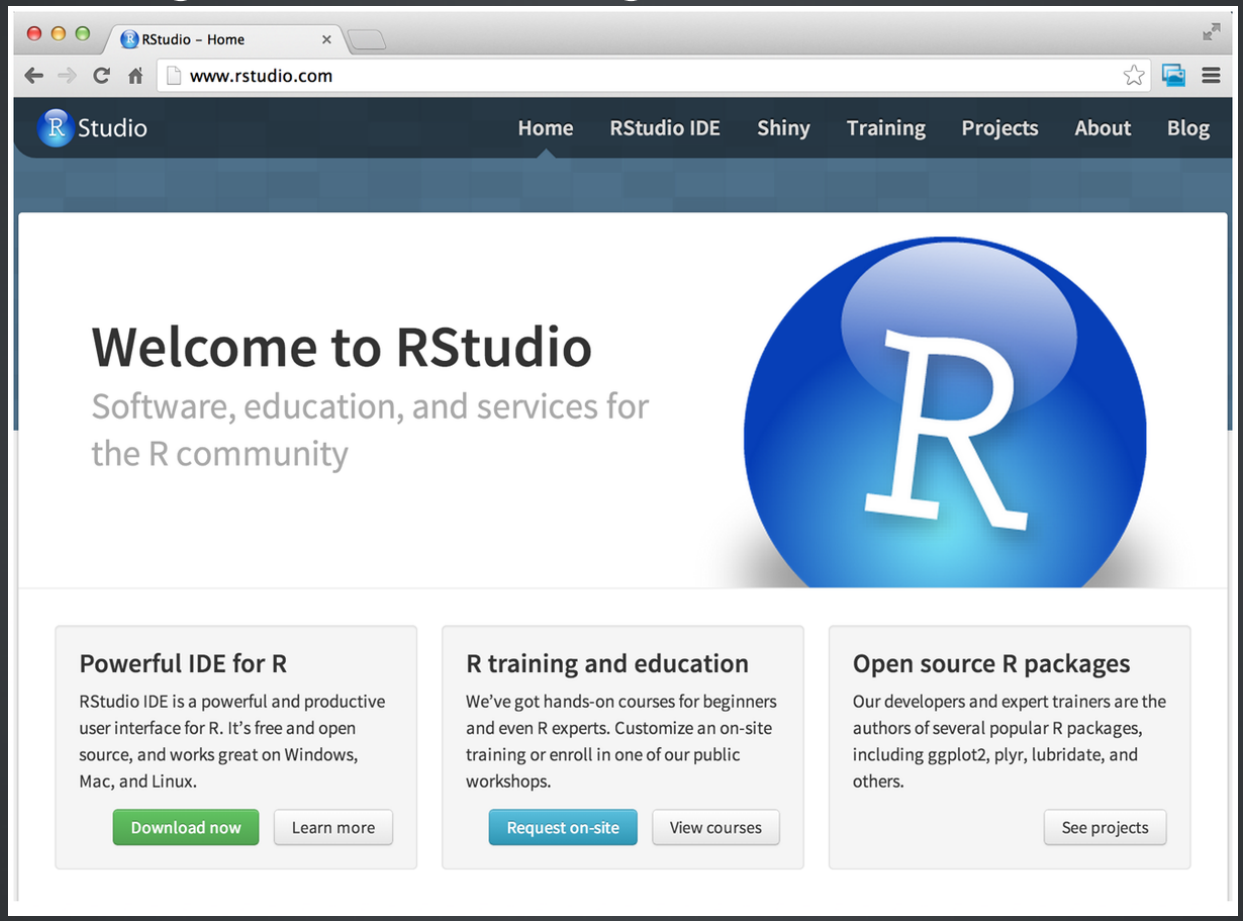
If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

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.

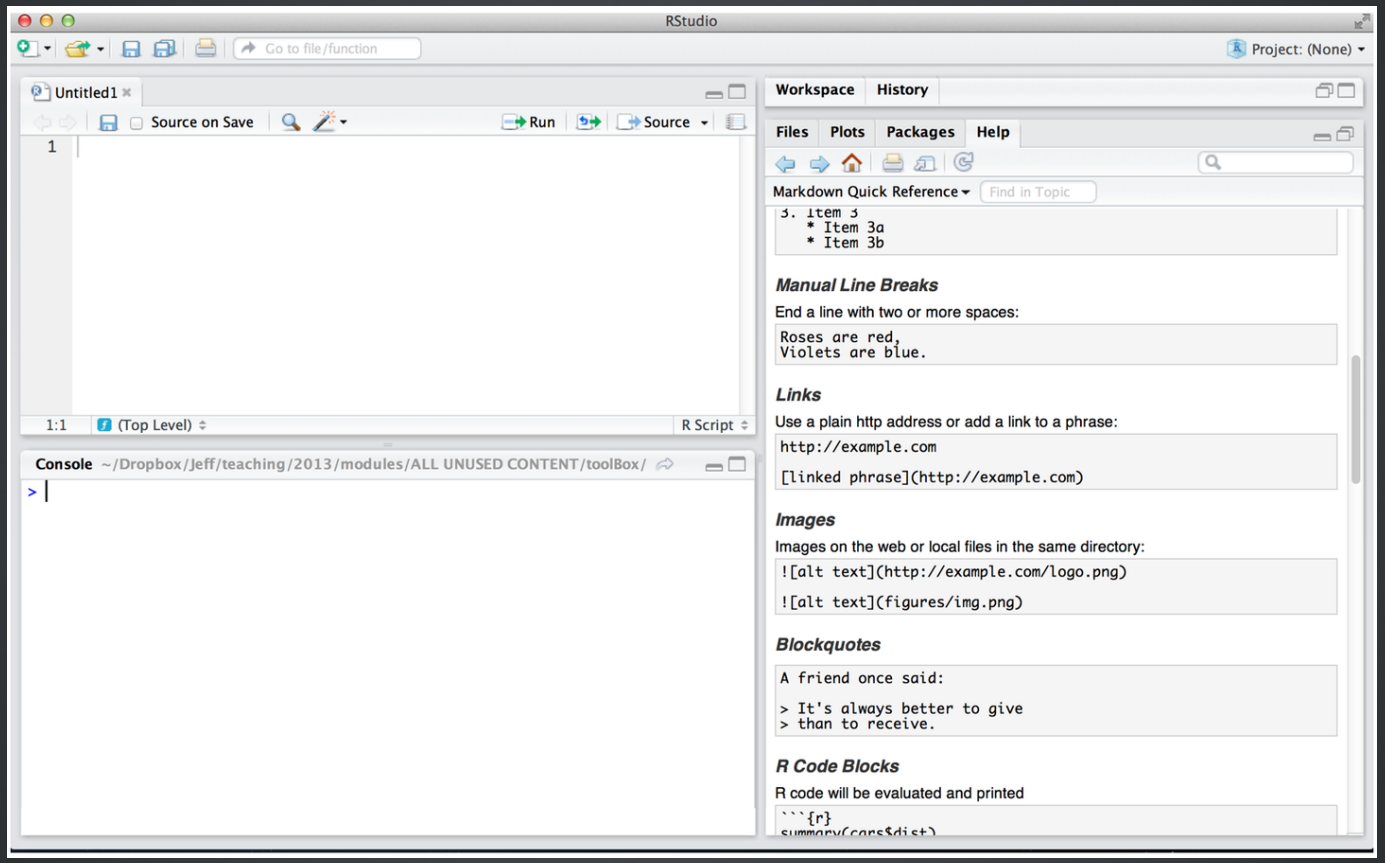
http://www.r-project.org/

A good coding environment



http://www.rstudio.com/

Rstudio Interface



http://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

Rstudio Tour

http://bit.ly/1RdftK1

R packages

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

Available CRAN Packages By Name

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

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

http://cran.rstudio.com/

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

Home » BiocViews

All Packages

Bioconductor version 3.1 (Release)

Autocomplete biocViews search:

- Software (1024)
  - AssayDomain (345)
  - BiologicalQuestion (313)
  - Infrastructure (211)
  - ResearchField (225)
  - StatisticalMethod (293)
  - Technology (645)
  - WorkflowStep (525)
- AnnotationData (883)
- ExperimentData (241)

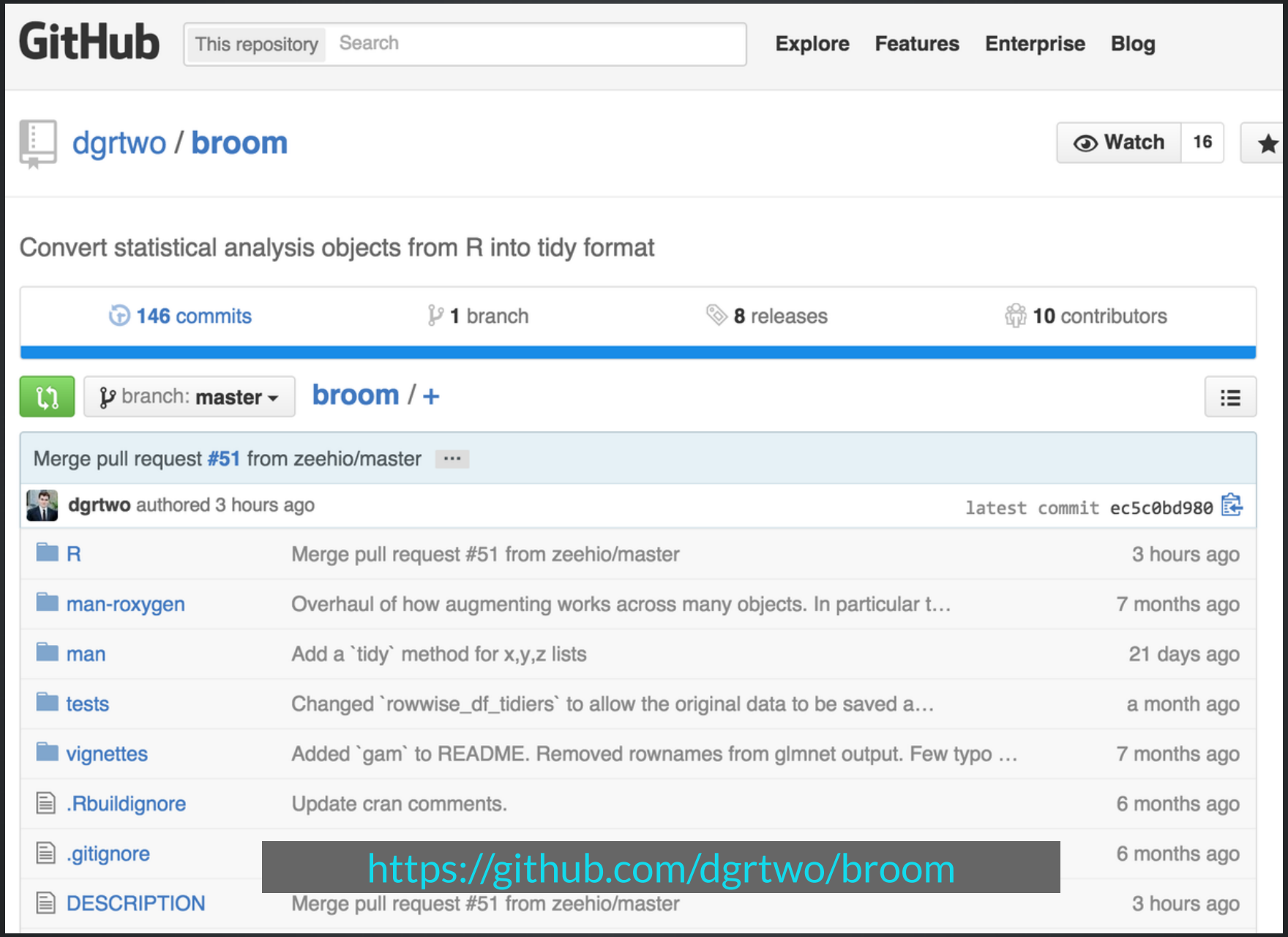
Packages found under Software:

Show All entries Search table:

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

http://bioconductor.org/

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


GitHub
This repository
Search
Explore
Features
Enterprise
Blog
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
R
Merge pull request #51 from zeehio/master
3 hours ago
man-roxygen
Overhaul of how augmenting works across many objects. In particular t…
7 months ago
man
Add a `tidy` method for x,y,z lists
21 days ago
tests
Changed `rowwise_df_tidiers` to allow the original data to be saved a…
a month ago
vignettes
Added `gam` to README. Removed rownames from glmnet output. Few typo …
7 months ago
.Rbuildignore
Update cran comments.
6 months ago
.gitignore
6 months ago
DESCRIPTION
Merge pull request #51 from zeehio/master
3 hours ago
https://github.com/dgrtwo/broom

```
devtools::install_github("broom")
```

Average Trustworthiness

>

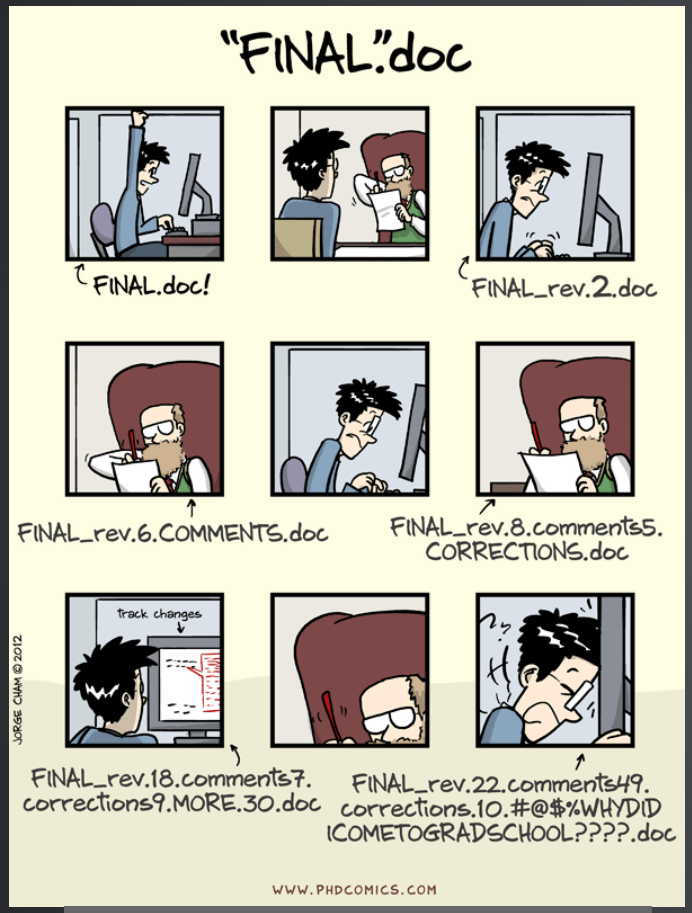
>

Installing Packages

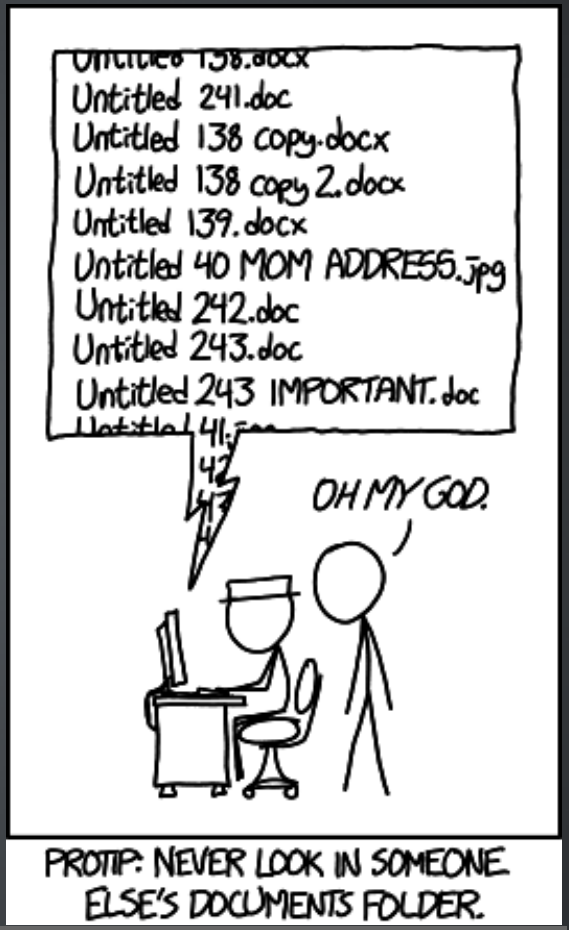
http://bit.ly/1CdylSm

Version Control

Just. No.



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

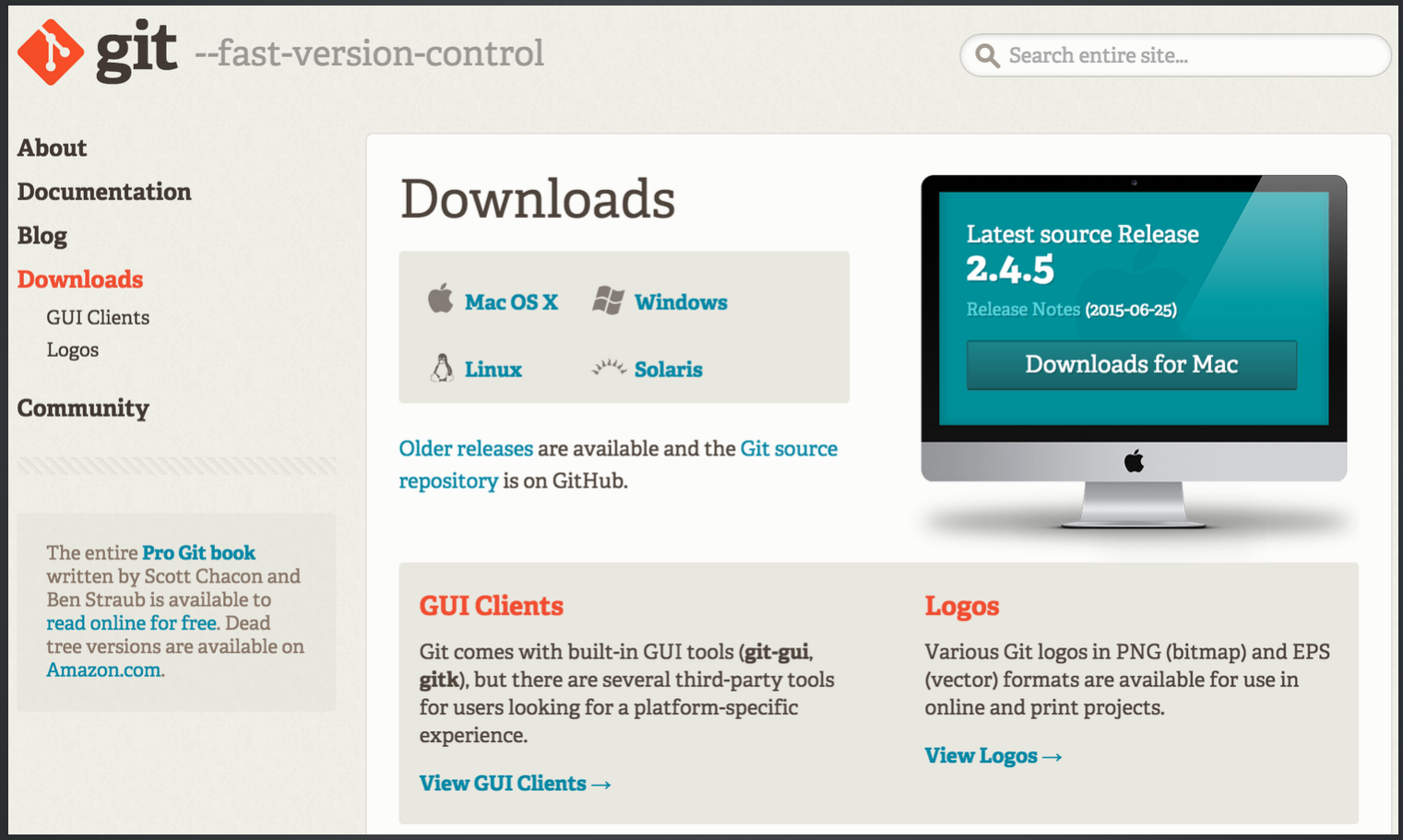


https://xkcd.com/1459/

Unfortunate truth

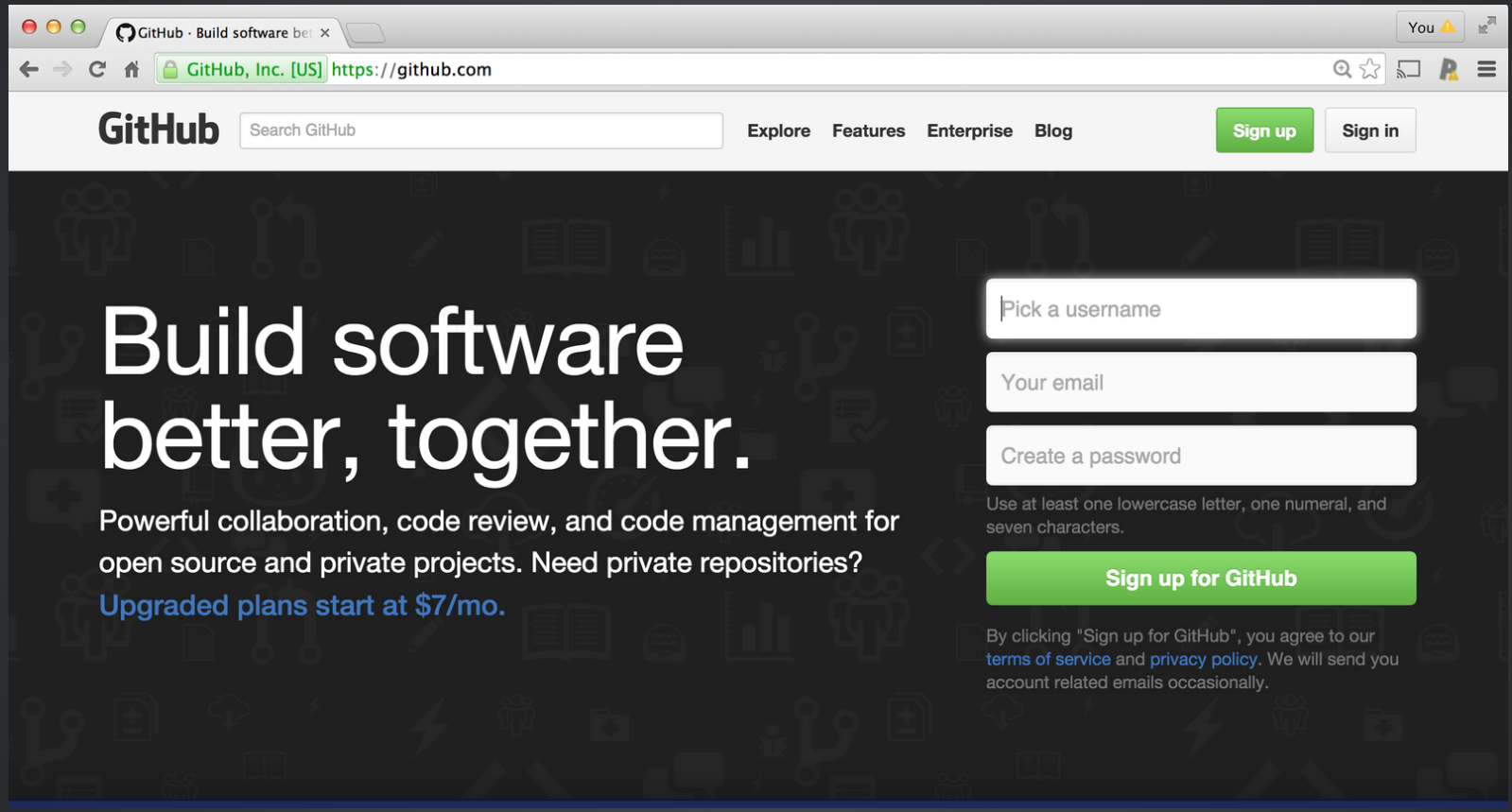
https://twitter.com/mtrc/status/617075570761965568

Git



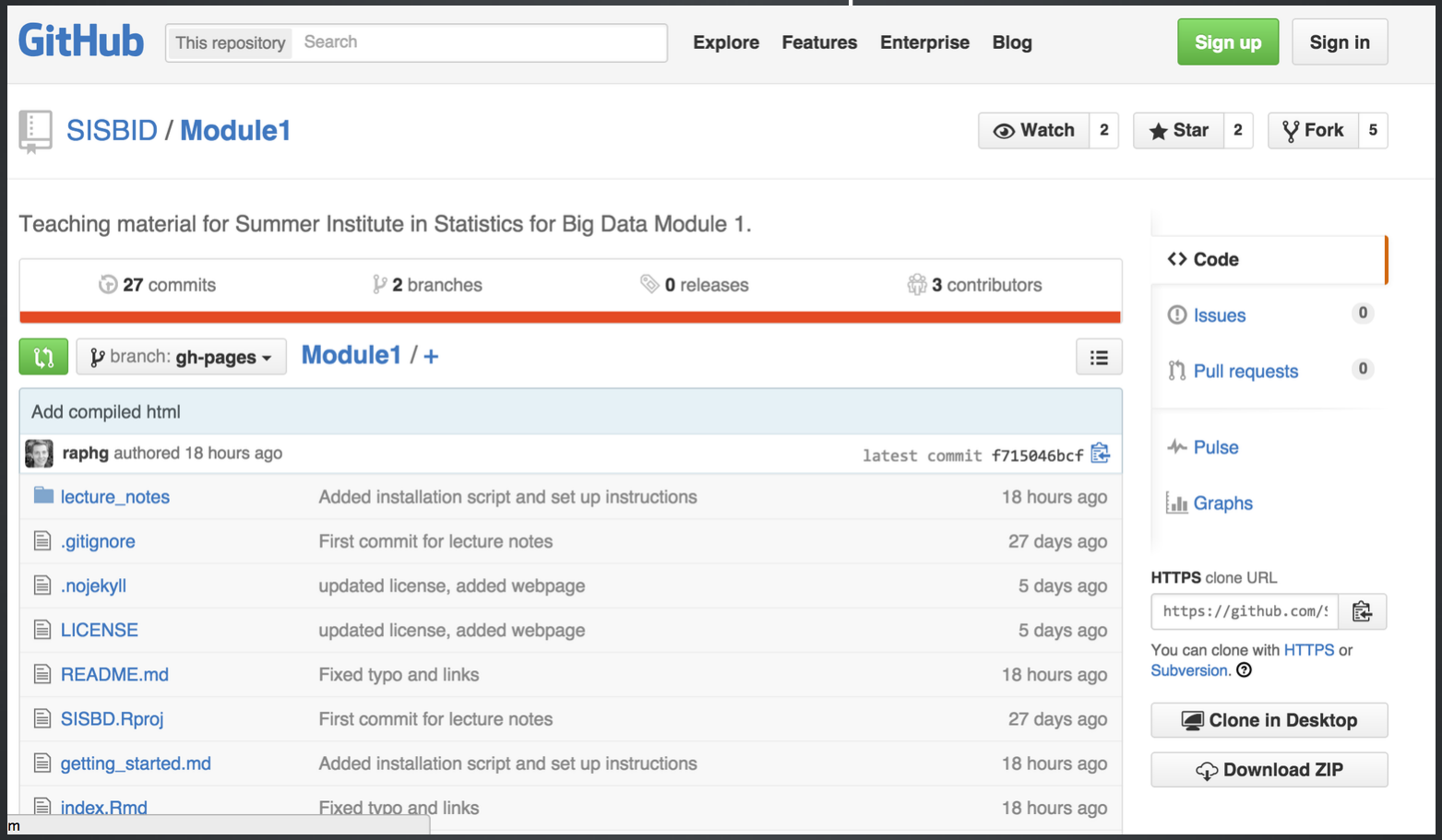
http://git-scm.com/downloads

Github



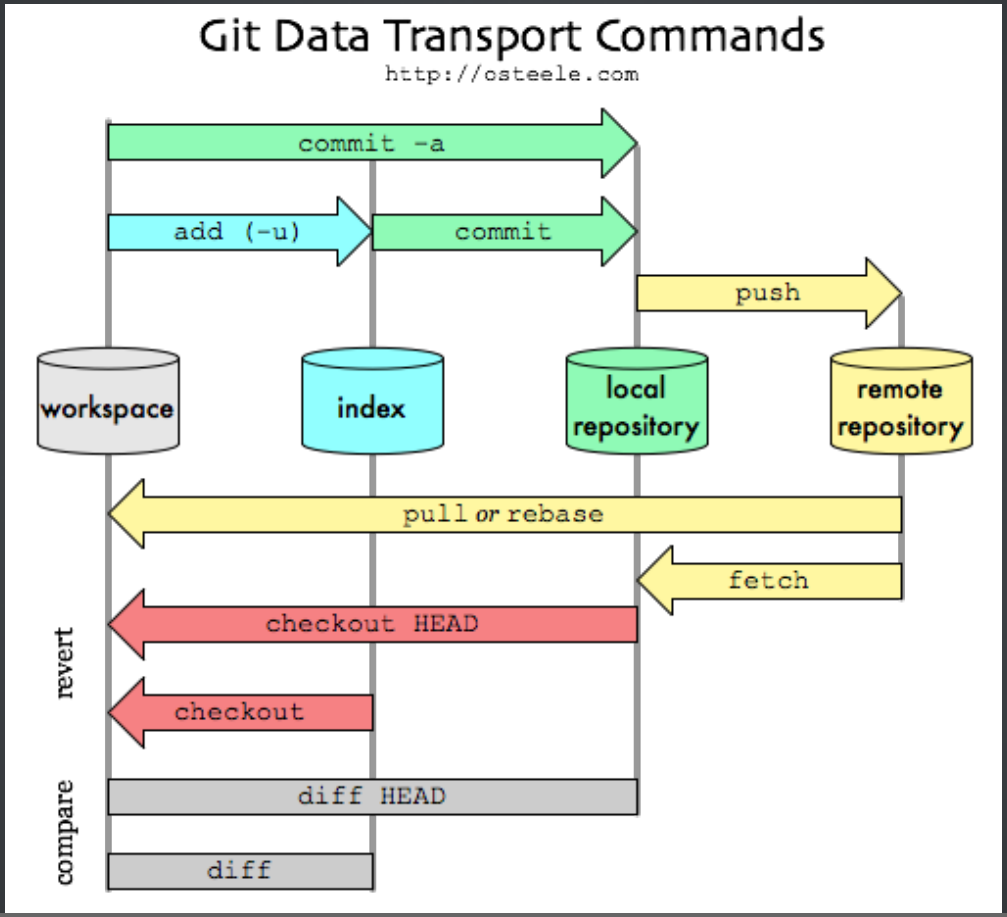
https://github.com/

Github repo



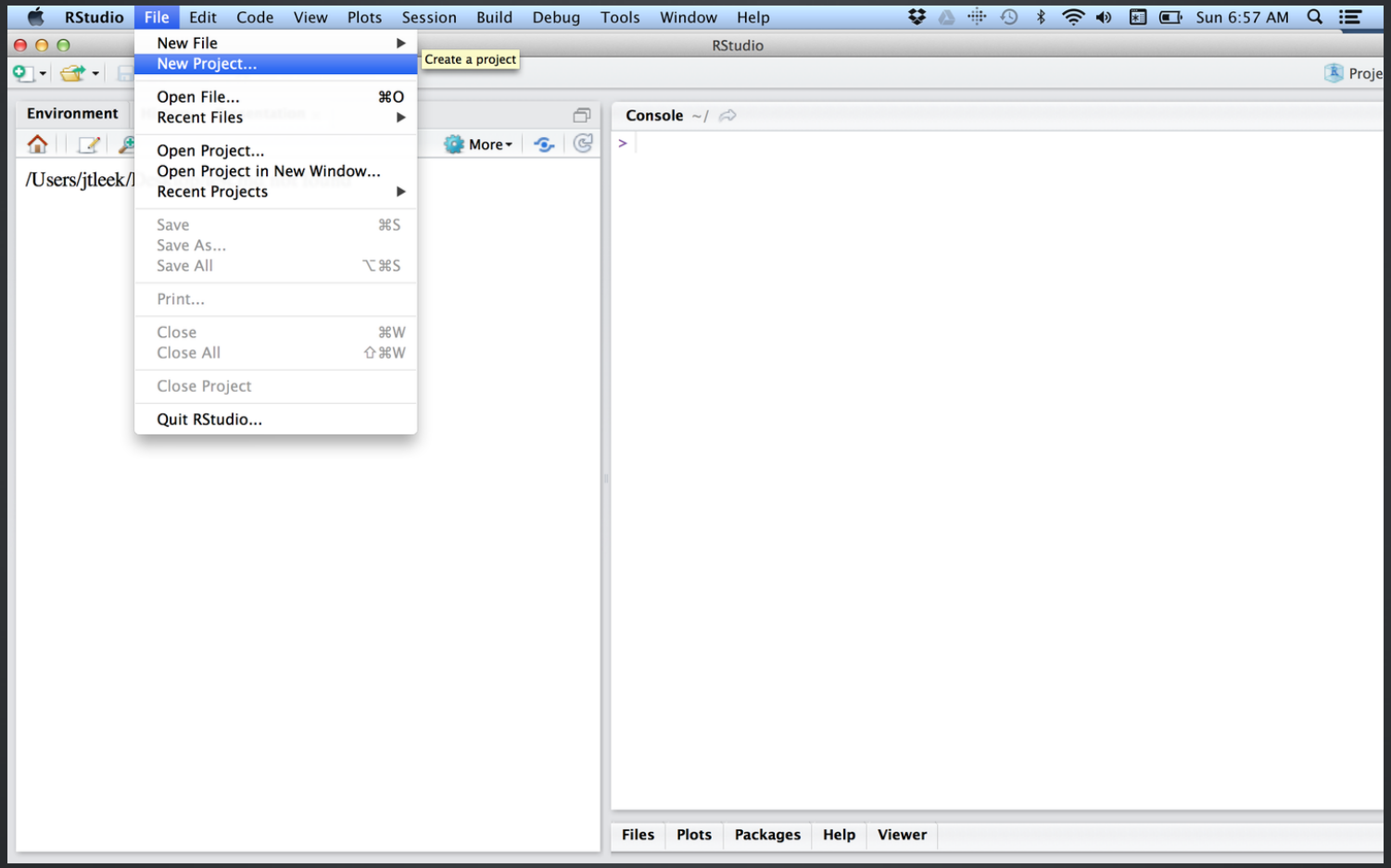
https://github.com/jtleek/datasharing

Basic Scheme

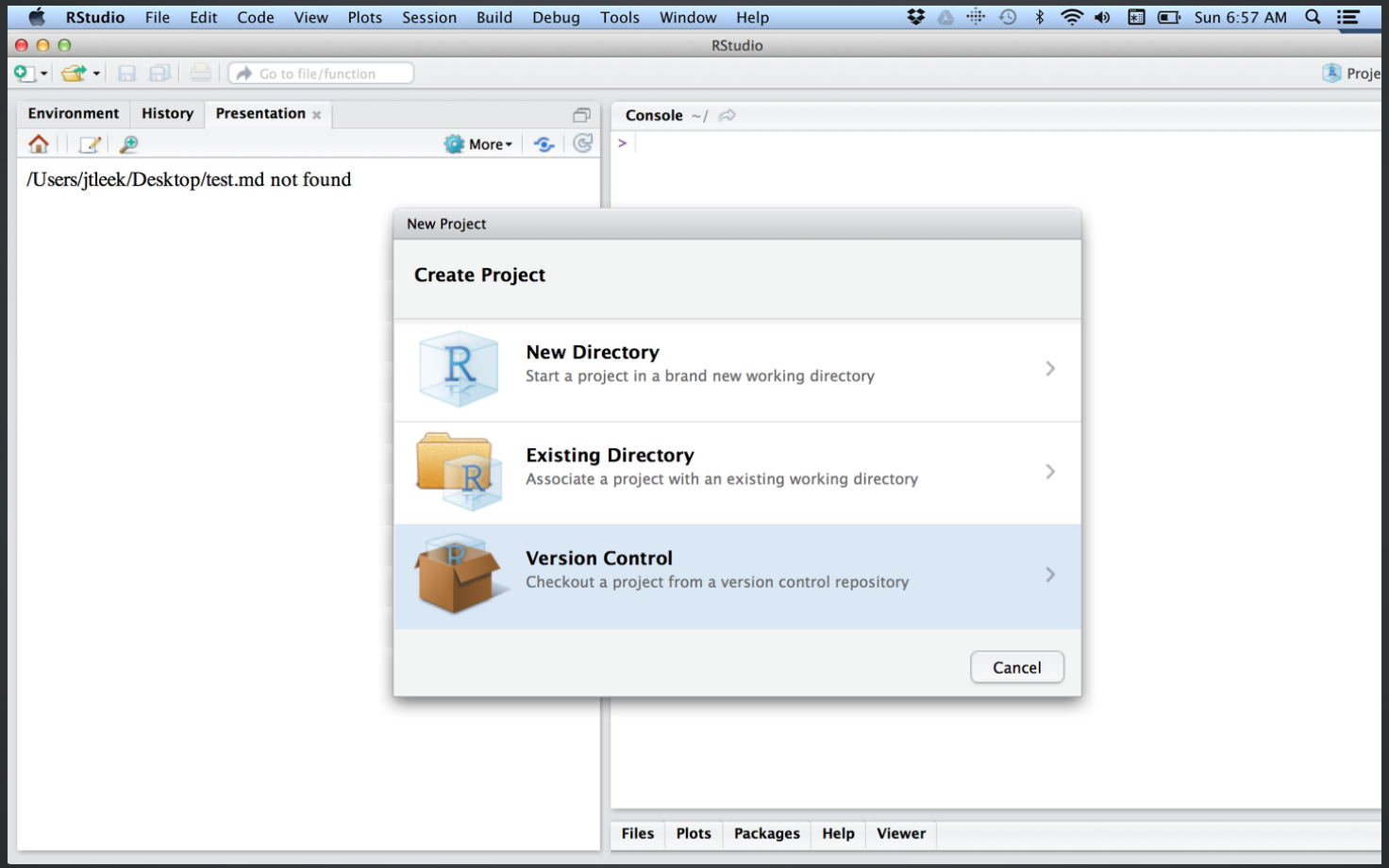


http://gitready.com/beginner/2009/01/21/pushing-and-pulling.html

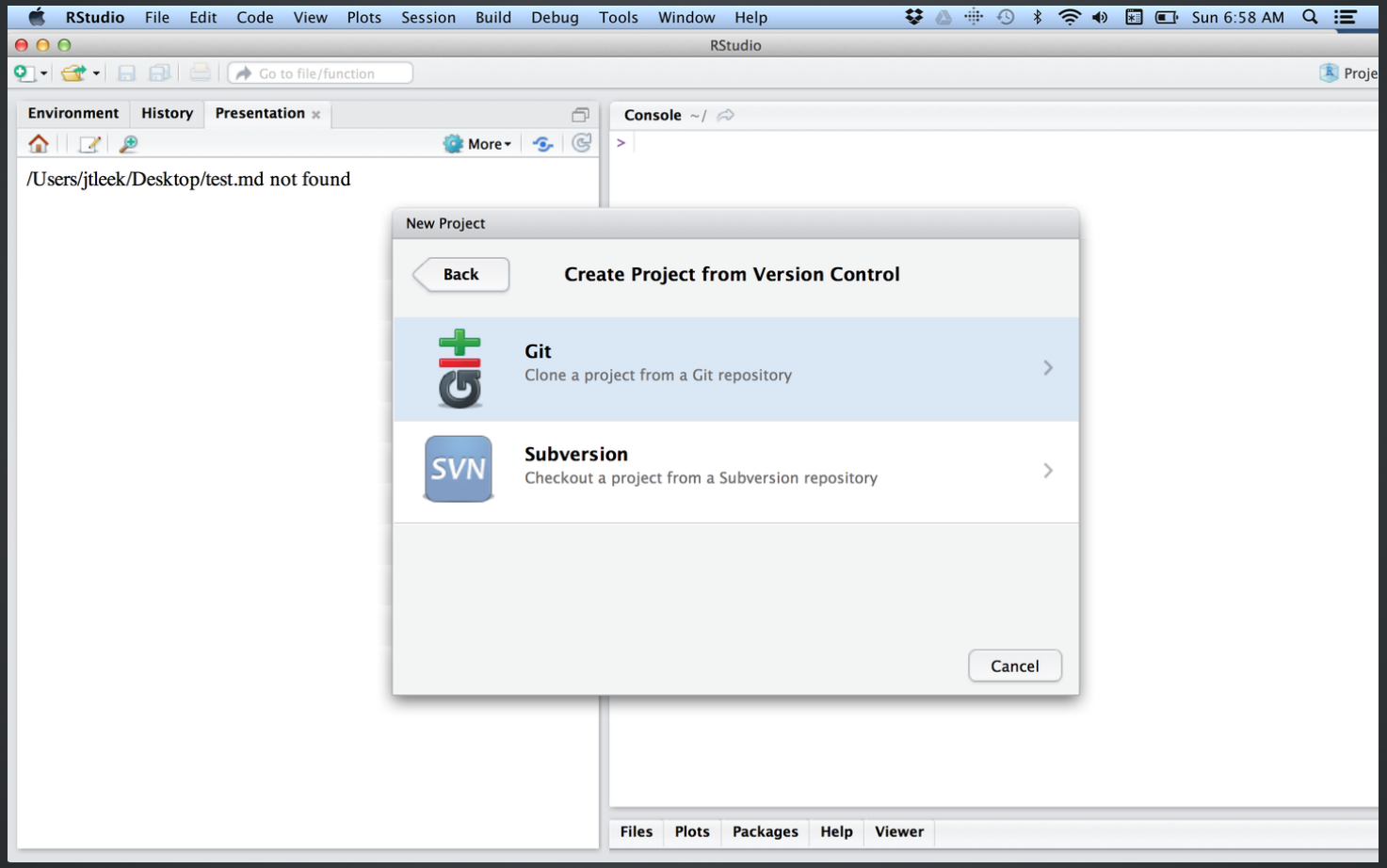
Git/Github in Rstudio



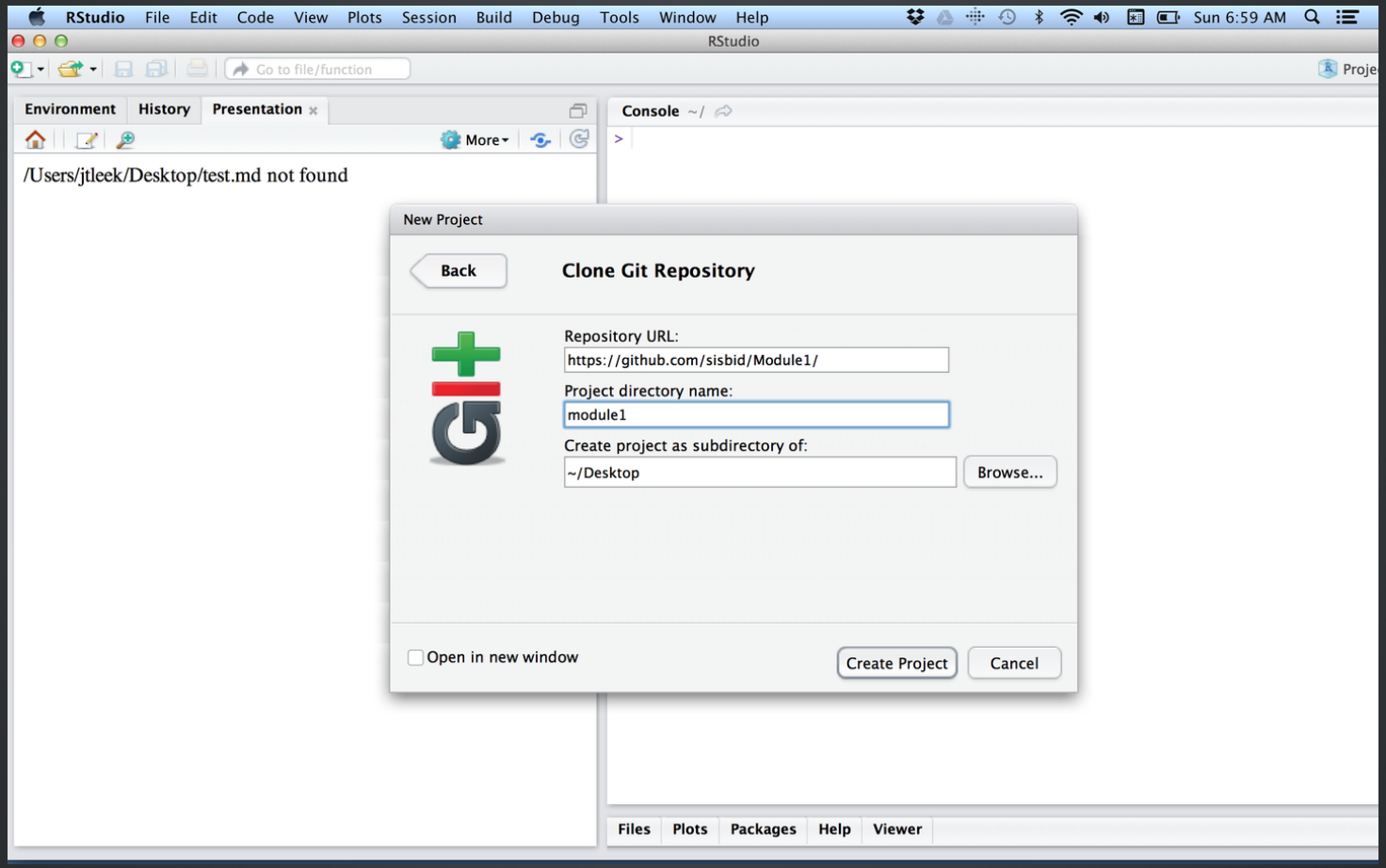
Git/Github in Rstudio



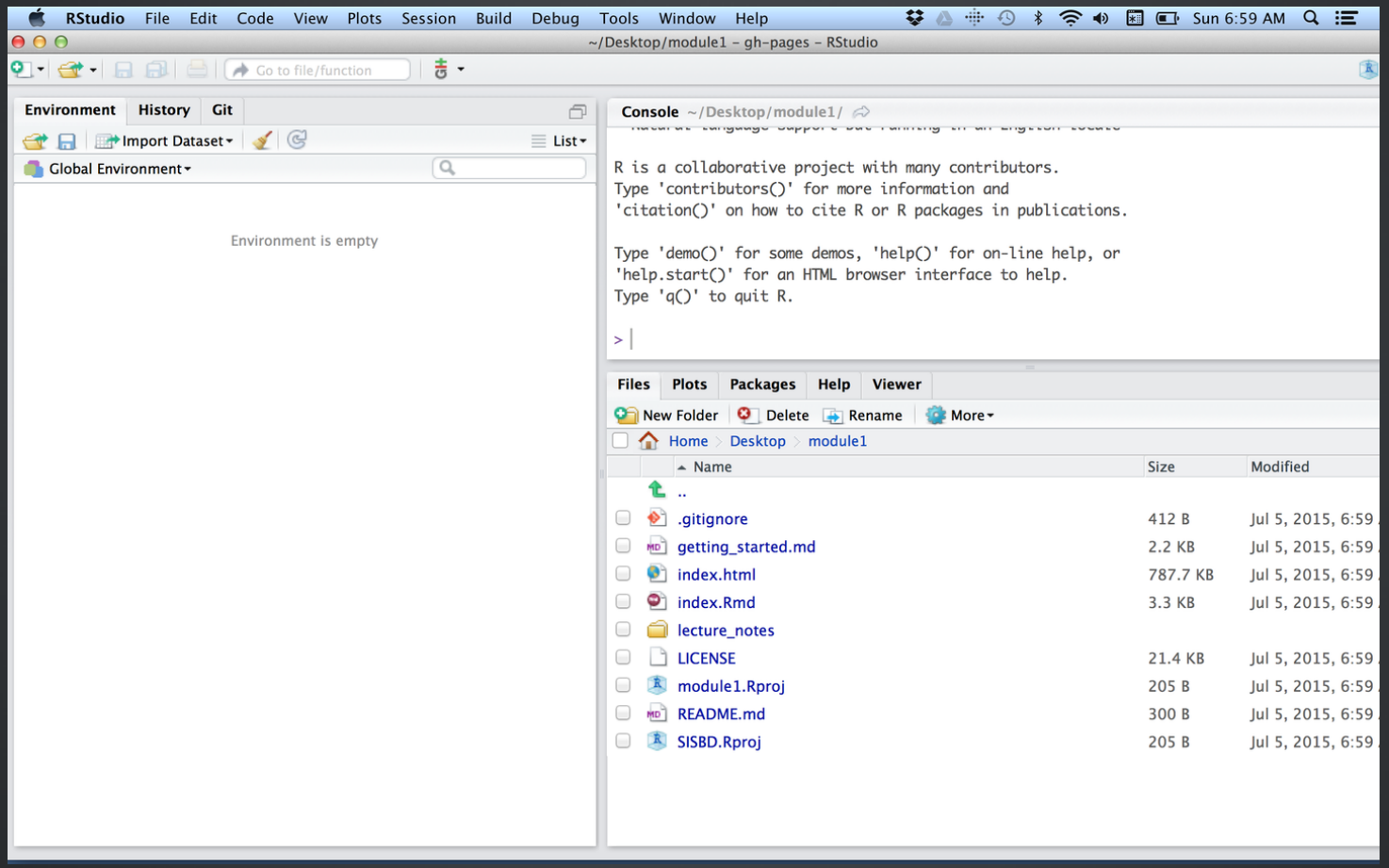
Git/Github in Rstudio



Git/Github in Rstudio



Git/Github in Rstudio



Installing Git/Github

http://r-pkgs.had.co.nz/git.html#git-init

Git/Github Practice

http://bit.ly/1H0qcwB

Reproducible Research

Why reproducible research?

> Your closest collaborator is you six months ago, but you don't reply to emails.

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

Why reproducible research?

> Could you just re-run all that code with the [latest/different/best] parameters?
>
> - every collaborator

Why reproducible research?

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

Why reproducible research?

REPORT BRIEF MARCH 2012

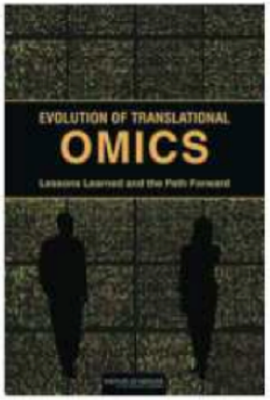
INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES

Advising the nation • Improving health

For more information visit www.iom.edu/translationalomics

Evolution of Translational Omics

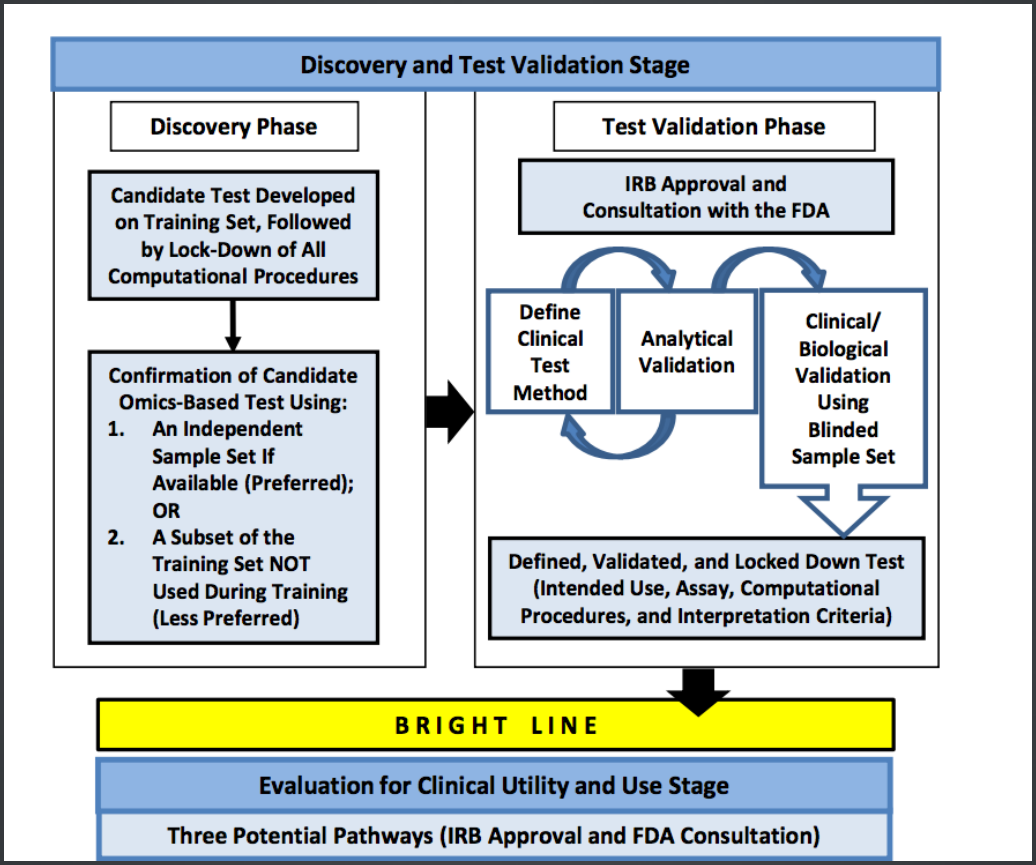
Lessons Learned and the Path Forward



Sequencing the human genome opened a new era in biomedical science. Researchers have begun to untangle the complex roles of biology and genetics in specific diseases, and now better understand why particular therapies do or do not work in individual patients. New technologies have made it feasible to measure an enormous number of molecules within a tissue or cell; for example, genomics investigates thousands of DNA sequences, and proteomics examines large numbers of proteins. Collectively, these technologies

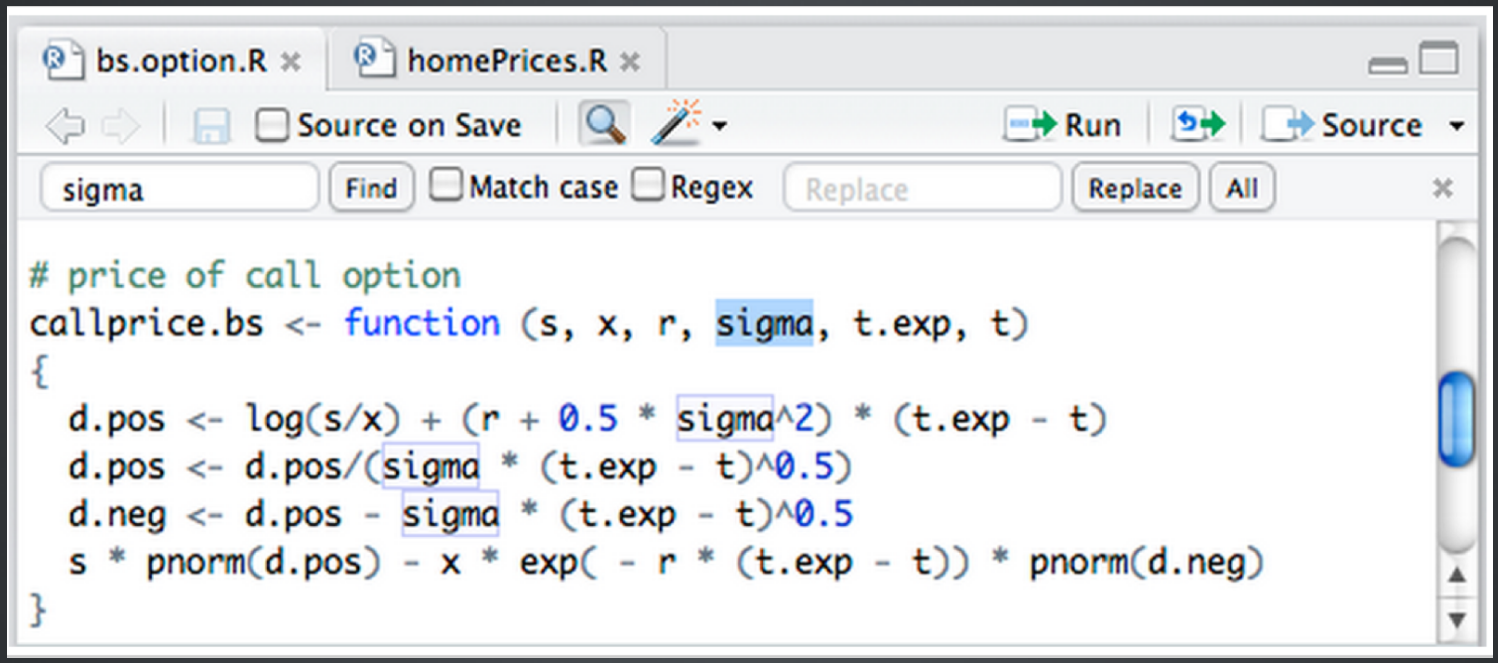
http://www.iom.edu/Reports/2012/Evolution-of-Translational-Omics.aspx

The bright line



http://www.iom.edu/Reports/2012/Evolution-of-Translational-Omics.aspx

R scripts



http://www.rstudio.com/ide/docs/using/source

Markdown ⟶ HTML

foo.md ⟶ foo.html

easy to write
(and read!)

easy to publish
easy to read in
browser

https://speakerdeck.com/jennybc/new-tools-and-workflows-for-data-analysis

Markdown → HTML

```
Title (header 1, actually)
==================================

This is a Markdown document.

## Medium header (header 2, actually)

It's easy to do *italics* or __make things bold__.

> All models are wrong, but some are useful. An
approximate answer to the right problem is worth a
good deal more than an exact answer to an
approximate problem. Absolute certainty is a
privilege of uneducated minds-and fanatics. It is,
for scientific folk, an unattainable ideal. What
you do every day matters more than what you do
once in a while. We cannot expect anyone to know
anything we didn't teach them ourselves.
Enthusiasm is a form of social courage.

Code block below. Just affects formatting here but
we'll get to R Markdown for the real fun soon!

```
x <- 3 * 4
```

I can haz equations. Inline equations, such as ...
the average is computed as $\frac{1}{n} \sum_{i=1}
^{n} x_{i}$. Or display equations like this:

$$
\begin{equation*}
|x|=
\begin{cases} x & \text{if $x≥0$,} \\\\
-x &\text{if $x\le 0$.}
\end{cases}
\end{equation*}
```

Title (header 1, actually)

This is a Markdown document.

Medium header (header 2, actually)

It's easy to do *italics* or **make things bold**.

> All models are wrong, but some are useful. An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem. Absolute certainty is a privilege of uneducated minds-and fanatics. It is, for scientific folk, an unattainable ideal. What you do every day matters more than what you do once in a while. We cannot expect anyone to know anything we didn't teach them ourselves. Enthusiasm is a form of social courage.

Code block below. Just affects formatting here but we'll get to R Markdown for the real fun soon!

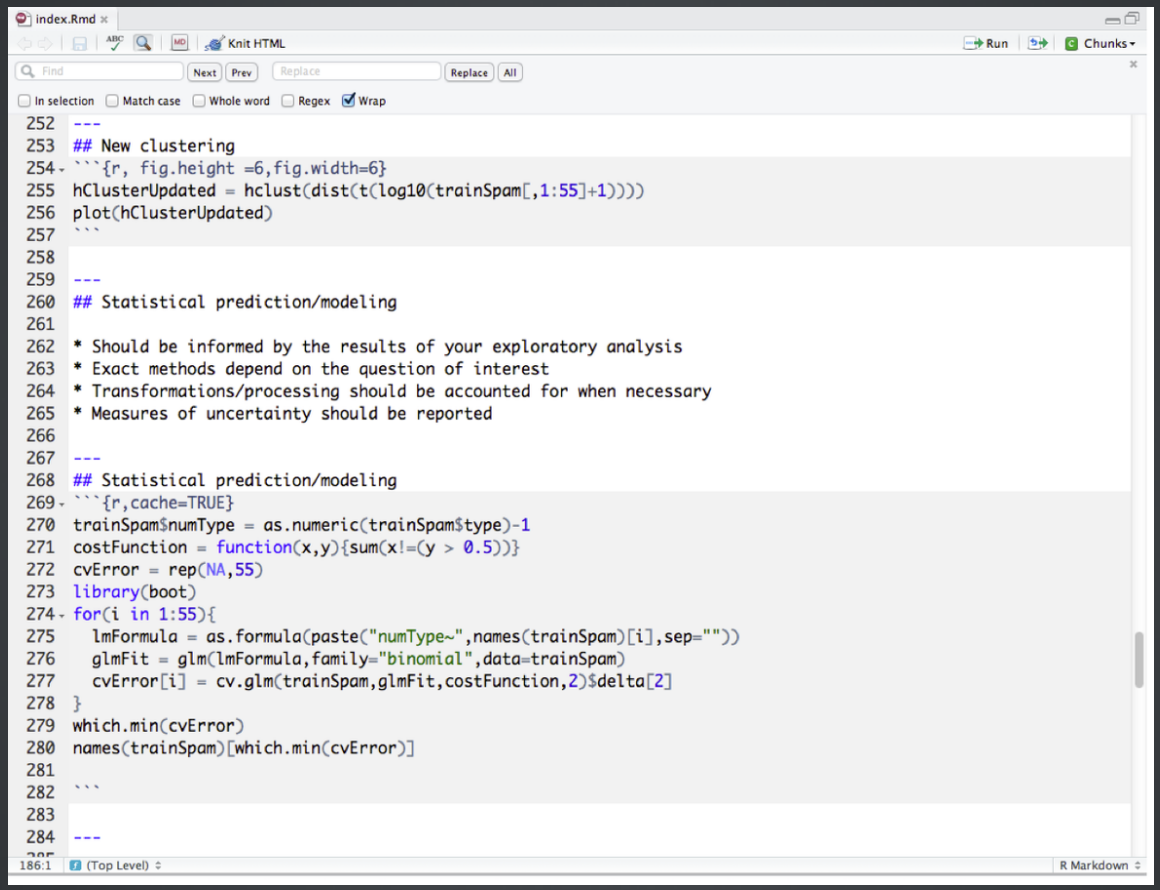
```
x <- 3 * 4
```

I can haz equations. Inline equations, such as ... the average is computed as $\frac{1}{n} \sum_{i=1}^{n} x_i$. Or display equations like this:

$$|x| = \begin{cases} x & \text{if } x \geq 0, \\ -x & \text{if } x < 0. \end{cases}$$

https://speakerdeck.com/jennybc/new-tools-and-workflows-for-data-analysis

R markdown files



http://www.rstudio.com/ide/docs/authoring/using_markdown

R Markdown → Markdown

````
R Markdown rocks
==================================

This is an R Markdown document.

```{r}
x <- rnorm(1000)
head(x)
```

See how the R code gets executed and a
representation thereof appears in the document?
`knitr` gives you control over how to represent all
conceivable types of output. In case you care, then
average of the `r length(x)` random normal variates
we just generated is `r round(mean(x), 3)`. Those
numbers are NOT hard-wired but are computed on-the-
fly. As is this figure. No more copy-paste ... copy-
paste ... oops forgot to copy-paste.

```{r}
plot(density(x))
```

Note that all the previously demonstrated math
typesetting still works. You don't have to choose
between having math cred and being web-friendly!

Inline equations, such as ... the average is
computed as $\frac{1}{n} \sum_{i=1}^{n} x_{i}$. Or
display equations like this:

$$
\begin{equation*}
|x|=
\begin{cases} x & \text{if $x≥0$,} \\\\
-x &\text{if $x\le 0$.}
\end{cases}
````

````
R Markdown rocks
==================================

This is an R Markdown document.

```r
x <- rnorm(1000)
head(x)
```

```
## [1] -1.3007  0.7715  0.5585 -1.2854  1.1973
2.4157
```

See how the R code gets executed and a
representation thereof appears in the document?
`knitr` gives you control over how to represent all
conceivable types of output. In case you care, then
average of the 1000 random normal variates we just
generated is -0.081. Those numbers are NOT hard-
wired but are computed on-the-fly. As is this
figure. No more copy-paste ... copy-paste ... oops
forgot to copy-paste.

```r
plot(density(x))
```

![plot of chunk unnamed-chunk-2](figure/unnamed-
chunk-2.png)

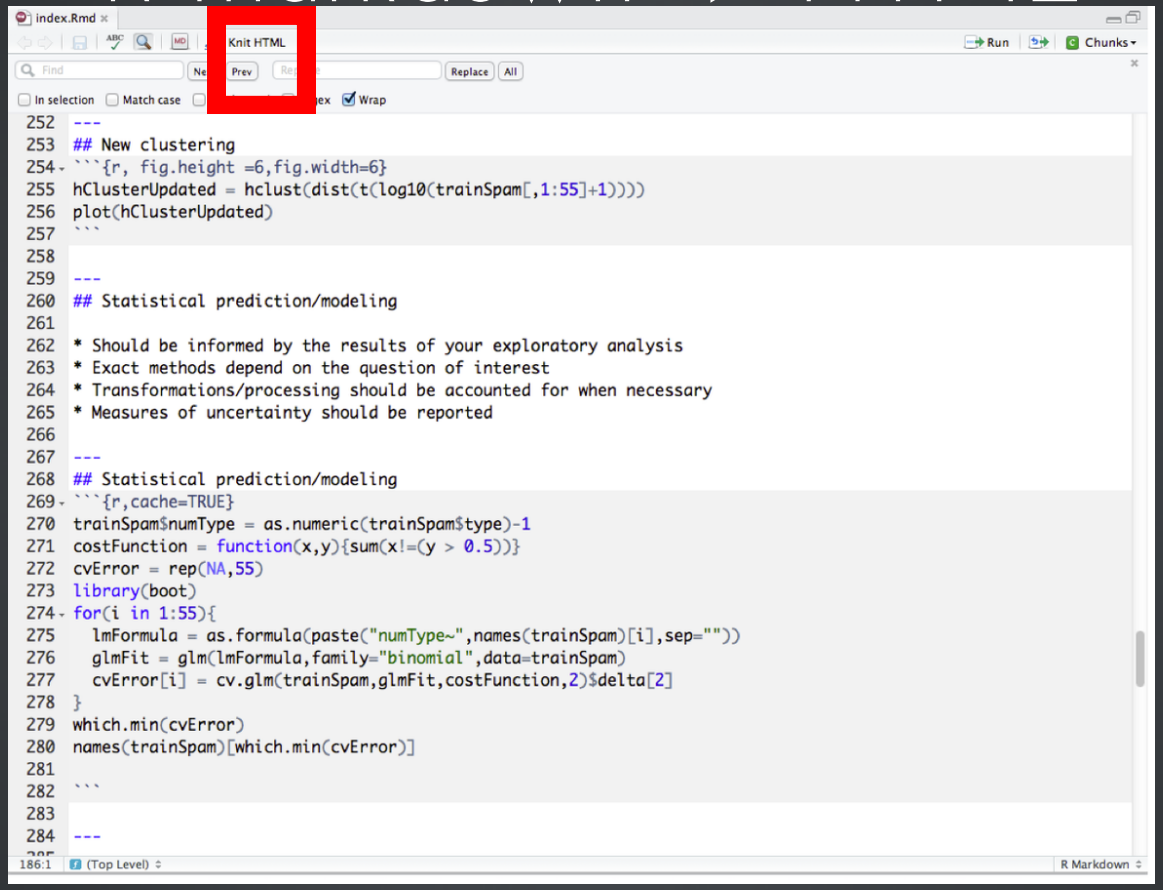
...
````

https://speakerdeck.com/jennybc/new-tools-and-workflows-for-data-analysis

```
library(rmarkdown
render("foo.Rmd")
```

https://speakerdeck.com/jennybc/new-tools-and-workflows-for-data-analysis

R markdown -> HTML



http://www.rstudio.com/ide/docs/authoring/using_markdown

R markdown Lab

http://bit.ly/1LRk3ds

Downloading Data

Organizing yourself

```
Project/
    raw/
    processed/
    code/
    README.md
```

Relative path (do this)

```
setwd("../data")
setwd("./files")

setwd("..\tmp")
```

Absolute path (not this)

```
setwd("/Users/jtleek/data")
setwd("~/Desktop/files/data")

setwd("C:\\Users\\Andrew\\Downloads")
```

Finding and creating files

```
if(!file.exists("data")){
  dir.create("data")
}
```

Listing files

```
list.files("data")
[1] "chicago.rds" "geo"
```

Downloading files

```
fileUrl <- "https://data.baltimorecity.gov/api/views/dz54-2aru/rows.csv?accessType=DOWNLOA
```

```
download.file(fileUrl,
    destfile="./data/cameras.csv",
    method="curl")
list.files("./data")
dateDownloaded <- date()
dateDownloaded
```

Maybe easier (on Macs)

```
fileUrl <- "https://data.baltimorecity.gov/api/views/dz54-2aru/rows.csv?accessType=DOWNLOA
```

```
install.packages("downloader")
library(downloader)
download(fileUrl,
    destfile="./data/cameras.csv")
list.files("./data")
dateDownloaded <- date()
dateDownloaded
```

Reading CSV

Reading local files

```
cameras = read.table("./data/cameras.csv",
                     header=T,
                     sep=",")

cameras = read.csv("./data/cameras.csv")
```

For big(ger) local files

```
install.packages("readr")
library(readr)
cameras = read_csv("./data/cameras.csv")
```

Quick comparison

```
dat = matrix(rnorm(100*1000),
                   nrow=1000)

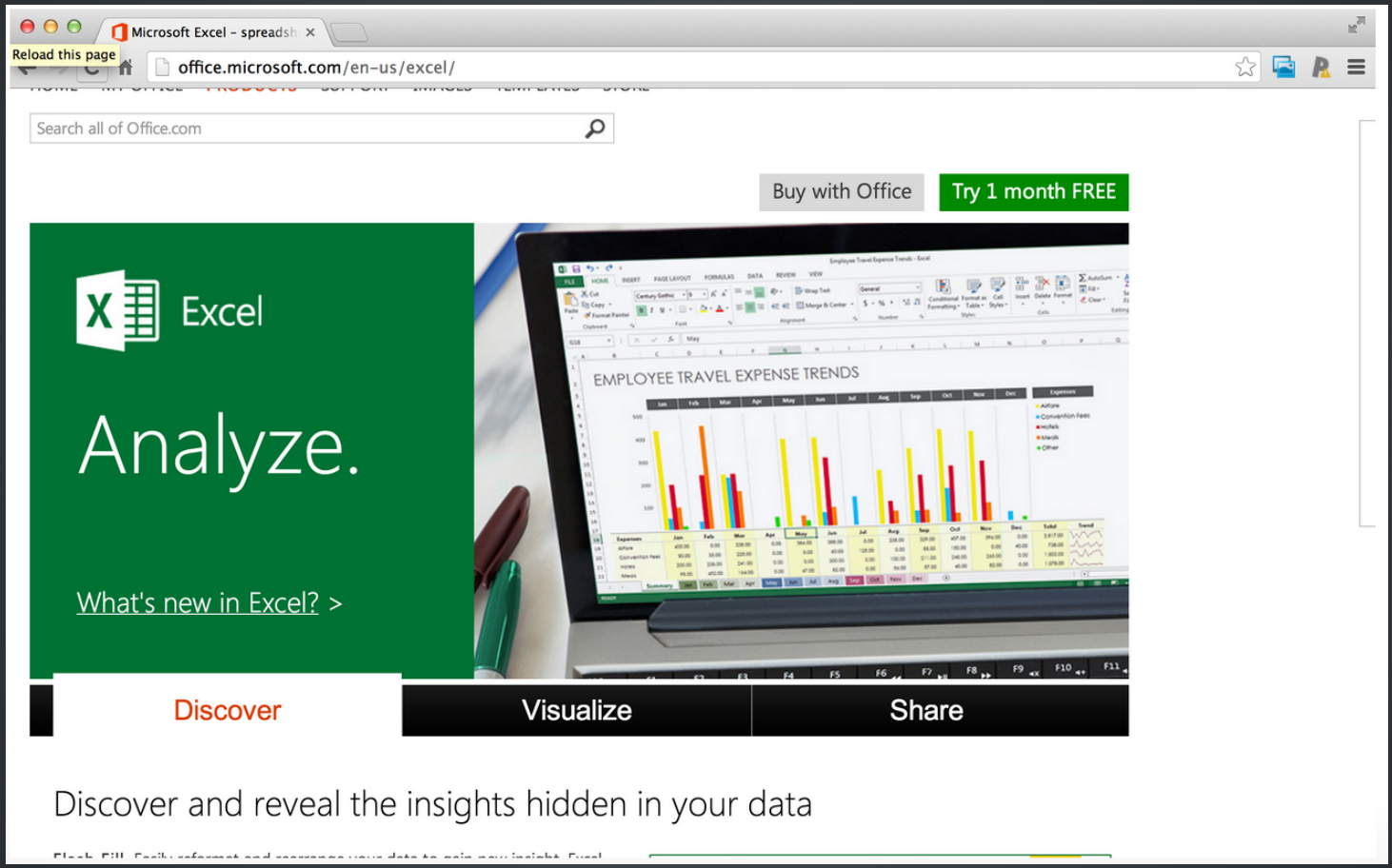
write.csv(dat,file="dat.csv")

system.time(read.csv("dat.csv"))
   user  system elapsed
  0.269   0.003   0.272

system.time(read_csv("dat.csv"))
   user  system elapsed
  0.025   0.001   0.026
```

Reading Excel

Most common file format



http://office.microsoft.com/en-us/excel/

Get an Excel File

```
fileUrl <- "https://data.baltimorecity.gov/api/views/dz54-2aru/rows.xlsx?accessType=DOWNLC
```

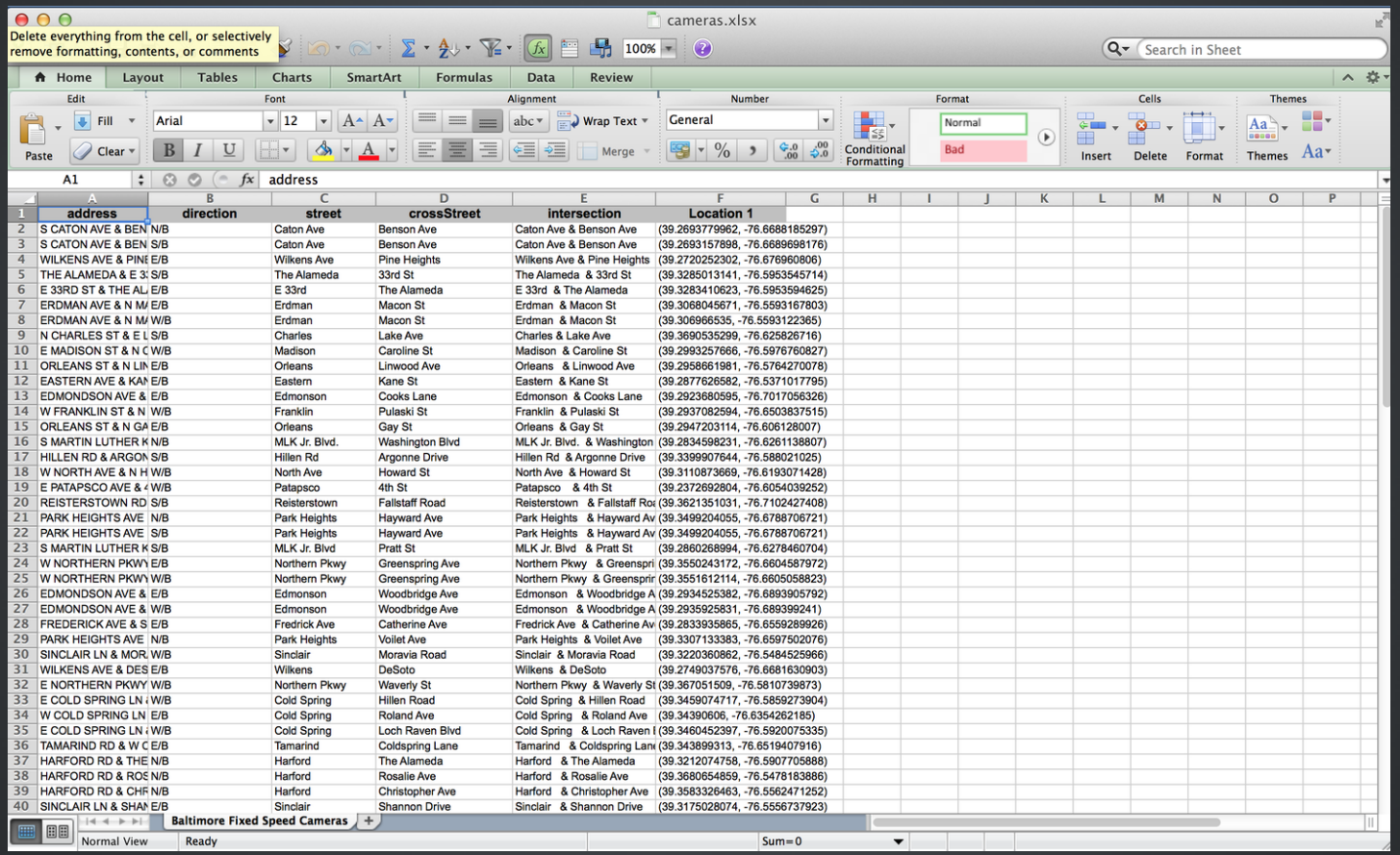
```
download(fileUrl,
    destfile="./data/cameras.xlsx")
list.files("./data")
dateDownloaded <- date()
dateDownloaded
```

Read it

```
install.packages("readxl")

cameras = read_excel("./data/cameras.xlsx",
                     sheet=1)
```

Sheets and Cells



	address	direction	street	crossStreet	intersection	Location 1
2	S CATON AVE & BEN	N/B	Caton Ave	Benson Ave	Caton Ave & Benson Ave	(39.2693779962, -76.6688185297)
3	S CATON AVE & BEN	S/B	Caton Ave	Benson Ave	Caton Ave & Benson Ave	(39.2693157898, -76.6689698176)
4	WILKENS AVE & PINE	E/B	Wilkens Ave	Pine Heights	Wilkens Ave & Pine Heights	(39.2720252302, -76.676960806)
5	THE ALAMEDA & E 3	S/B	The Alameda	33rd St	The Alameda & 33rd St	(39.3285013141, -76.5953545714)
6	E 33RD ST & THE AL	E/B	E 33rd	The Alameda	E 33rd & The Alameda	(39.3283410623, -76.5953594625)
7	ERDMAN AVE & N M	E/B	Erdman	Macon St	Erdman & Macon St	(39.3068045671, -76.5593167803)
8	ERDMAN AVE & N M	W/B	Erdman	Macon St	Erdman & Macon St	(39.306966535, -76.5593122365)
9	N CHARLES ST & E L	S/B	Charles	Lake Ave	Charles & Lake Ave	(39.3690535299, -76.625826716)
10	E MADISON ST & N C	W/B	Madison	Caroline St	Madison & Caroline St	(39.2993257666, -76.5976760827)
11	ORLEANS ST & N LIN	E/B	Orleans	Linwood Ave	Orleans & Linwood Ave	(39.2958661981, -76.5764270078)
12	EASTERN AVE & KAN	E/B	Eastern	Kane St	Eastern & Kane St	(39.2877626582, -76.5371017795)
13	EDMONDSON AVE &	E/B	Edmonson	Cooks Lane	Edmonson & Cooks Lane	(39.2923680595, -76.7017056326)
14	W FRANKLIN ST & N	W/B	Franklin	Pulaski St	Franklin & Pulaski St	(39.2937082594, -76.6503837515)
15	ORLEANS ST & N GA	E/B	Orleans	Gay St	Orleans & Gay St	(39.2947203114, -76.606128007)
16	S MARTIN LUTHER K	N/B	MLK Jr. Blvd.	Washington Blvd	MLK Jr. Blvd. & Washington	(39.2834598231, -76.6261138807)
17	HILLEN RD & ARGON	S/B	Hillen Rd	Argonne Drive	Hillen Rd & Argonne Drive	(39.3399907644, -76.588021025)
18	W NORTH AVE & N H	W/B	North Ave	Howard St	North Ave & Howard St	(39.3110873669, -76.6193071428)
19	E PATAPSCO AVE & A	W/B	Patapsco	4th St	Patapsco & 4th St	(39.2372692804, -76.6054039252)
20	REISTERSTOWN RD	S/B	Reisterstown	Fallstaff Road	Reisterstown & Fallstaff Roa	(39.3621351031, -76.7102427408)
21	PARK HEIGHTS AVE	N/B	Park Heights	Hayward Ave	Park Heights & Hayward Av	(39.3499204055, -76.6788706721)
22	PARK HEIGHTS AVE	S/B	Park Heights	Hayward Ave	Park Heights & Hayward Av	(39.3499204055, -76.6788706721)
23	S MARTIN LUTHER K	S/B	MLK Jr. Blvd	Pratt St	MLK Jr. Blvd & Pratt St	(39.2860268994, -76.6278460704)
24	W NORTHERN PKWY	E/B	Northern Pkwy	Greenspring Ave	Northern Pkwy & Greenspri	(39.3550243172, -76.6604587972)
25	W NORTHERN PKWY	W/B	Northern Pkwy	Greenspring Ave	Northern Pkwy & Greensprir	(39.3551612114, -76.6605058823)
26	EDMONDSON AVE &	E/B	Edmonson	Woodbridge Ave	Edmonson & Woodbridge A	(39.2934525382, -76.6893905792)
27	EDMONDSON AVE &	W/B	Edmonson	Woodbridge Ave	Edmonson & Woodbridge A	(39.2935925831, -76.689399241)
28	FREDERICK AVE & S	E/B	Fredrick Ave	Catherine Ave	Fredrick Ave & Catherine Av	(39.2833935865, -76.6559289926)
29	PARK HEIGHTS AVE	N/B	Park Heights	Voilet Ave	Park Heights & Voilet Ave	(39.3307133383, -76.6597502076)
30	SINCLAIR LN & MOR	W/B	Sinclair	Moravia Road	Sinclair & Moravia Road	(39.3220360862, -76.5484525966)
31	WILKENS AVE & DES	E/B	Wilkens	DeSoto	Wilkens & DeSoto	(39.2749037576, -76.6681630903)
32	E NORTHERN PKWY	W/B	Northern Pkwy	Waverly St	Northern Pkwy & Waverly St	(39.367051509, -76.5810739873)
33	E COLD SPRING LN	W/B	Cold Spring	Hillen Road	Cold Spring & Hillen Road	(39.3459074717, -76.5859273904)
34	W COLD SPRING LN	E/B	Cold Spring	Roland Ave	Cold Spring & Roland Ave	(39.34390606, -76.6354262185)
35	E COLD SPRING LN	W/B	Cold Spring	Loch Raven Blvd	Cold Spring & Loch Raven I	(39.3460452397, -76.5920075335)
36	TAMARIND RD & W C	E/B	Tamarind	Coldspring Lane	Tamarind & Coldspring Lan	(39.343899313, -76.6519407916)
37	HARFORD RD & THE	N/B	Harford	The Alameda	Harford & The Alameda	(39.3212074758, -76.5907705888)
38	HARFORD RD & ROS	N/B	Harford	Rosalie Ave	Harford & Rosalie Ave	(39.3680654859, -76.5478183886)
39	HARFORD RD & CHR	N/B	Harford	Christopher Ave	Harford & Christopher Ave	(39.3583326463, -76.5562471252)
40	SINCLAIR LN & SHAN	E/B	Sinclair	Shannon Drive	Sinclair & Shannon Drive	(39.3175028074, -76.5556737923)

If you want to read cells

```
install.packages("xlsx")
library(xlsx)
colIndex = 2:3
rowIndex = 1:4

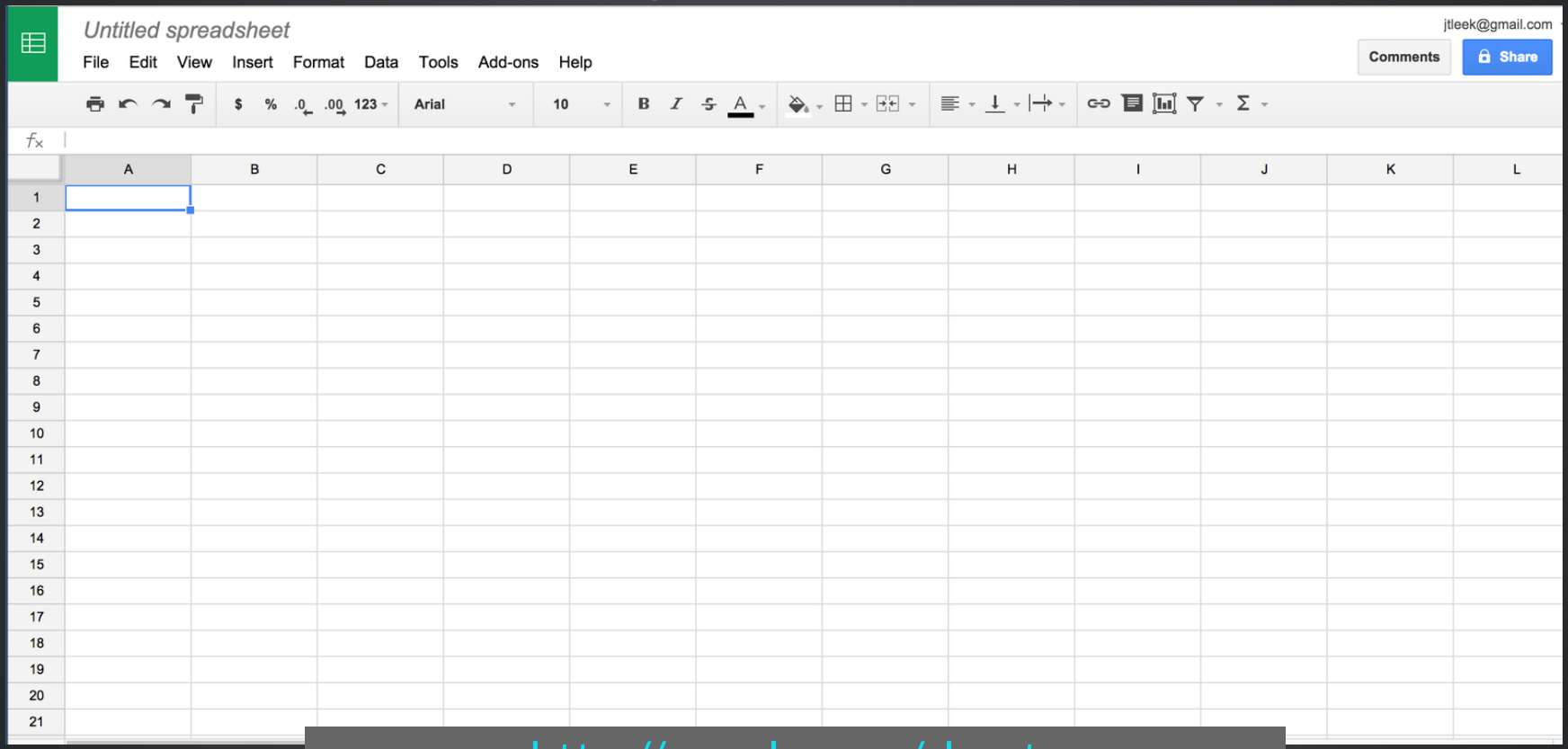
cameras = read.xlsx("./data/cameras.xlsx",
                    sheetIndex=1,
                    header=TRUE,
                    colIndex= colIndex,
                    rowIndex = rowIndex)
```

Result

```
> cameras
  direction       street
1       N/B   Caton Ave
2       S/B   Caton Ave
3       E/B Wilkens Ave
```

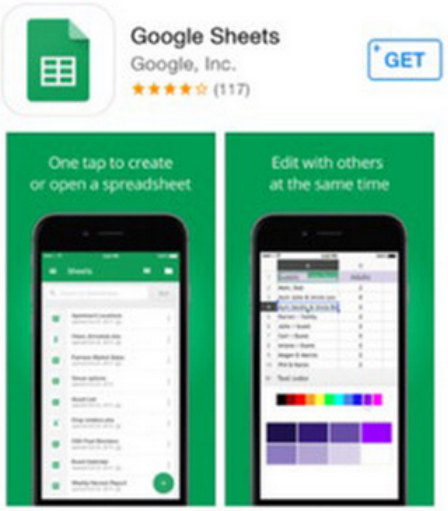
Reading Google Sheets

Google Sheets

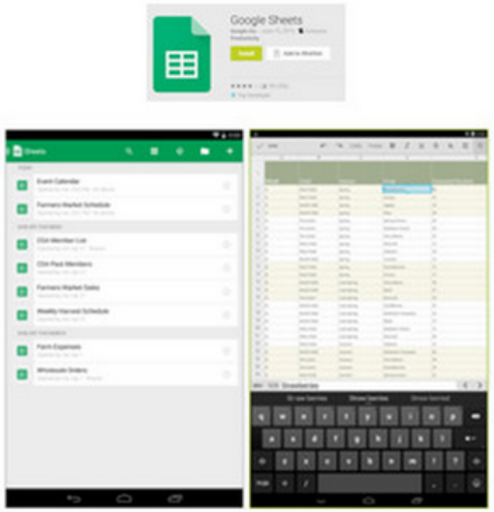


http://google.com/sheets

iOS



Android

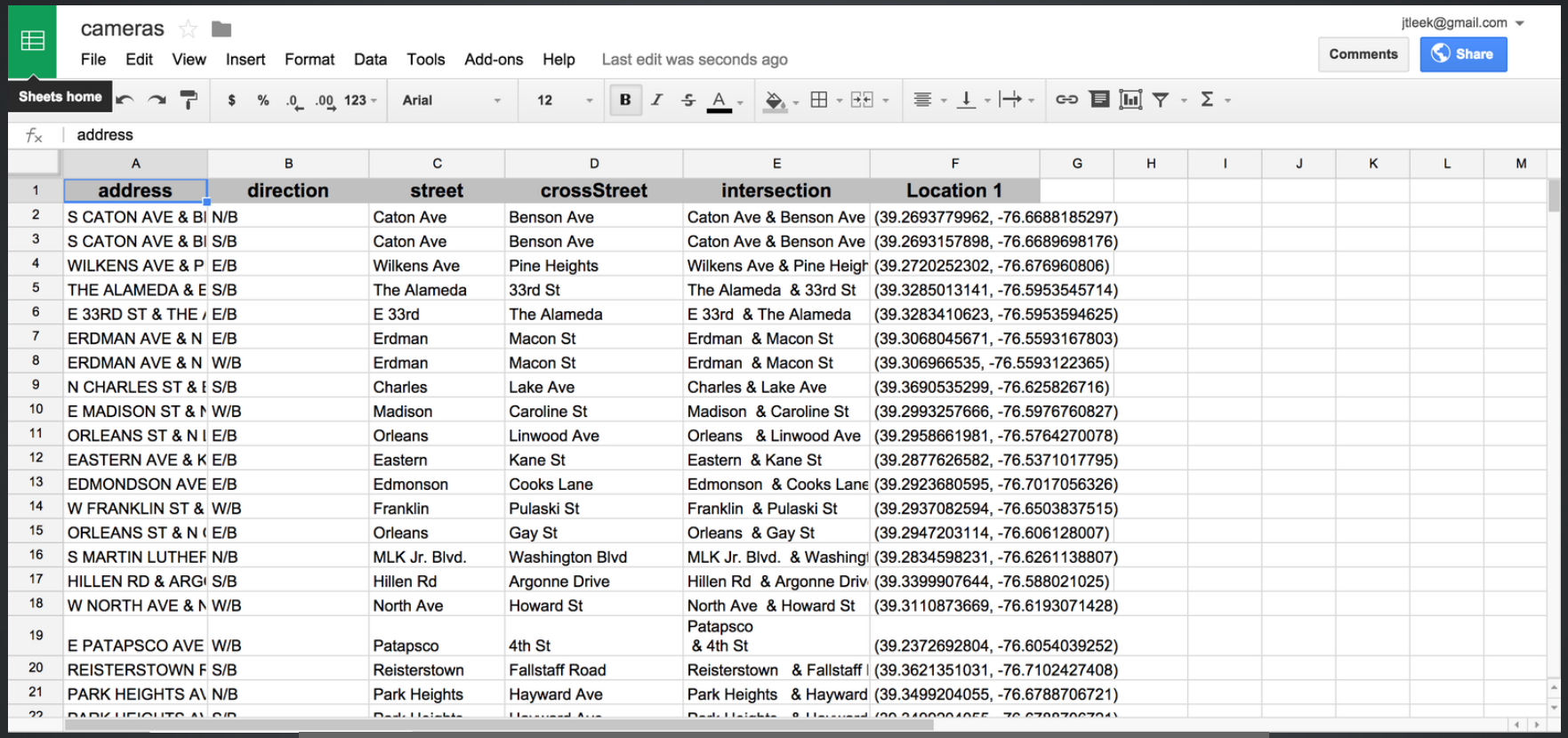


enter data from a phone
enter data w/o WiFi

Slide from: https://speakerdeck.com/jennybc/googlesheets-talk-at-user2015

googlesheets package

```
library(devtools)
install_github("jennybc/cellranger")
install_github("jennybc/googlesheets")
library(googlesheets)
?gs_read
?"cell-specification"
```



cameras

jtleek@gmail.com

File Edit View Insert Format Data Tools Add-ons Help Last edit was seconds ago

Comments Share

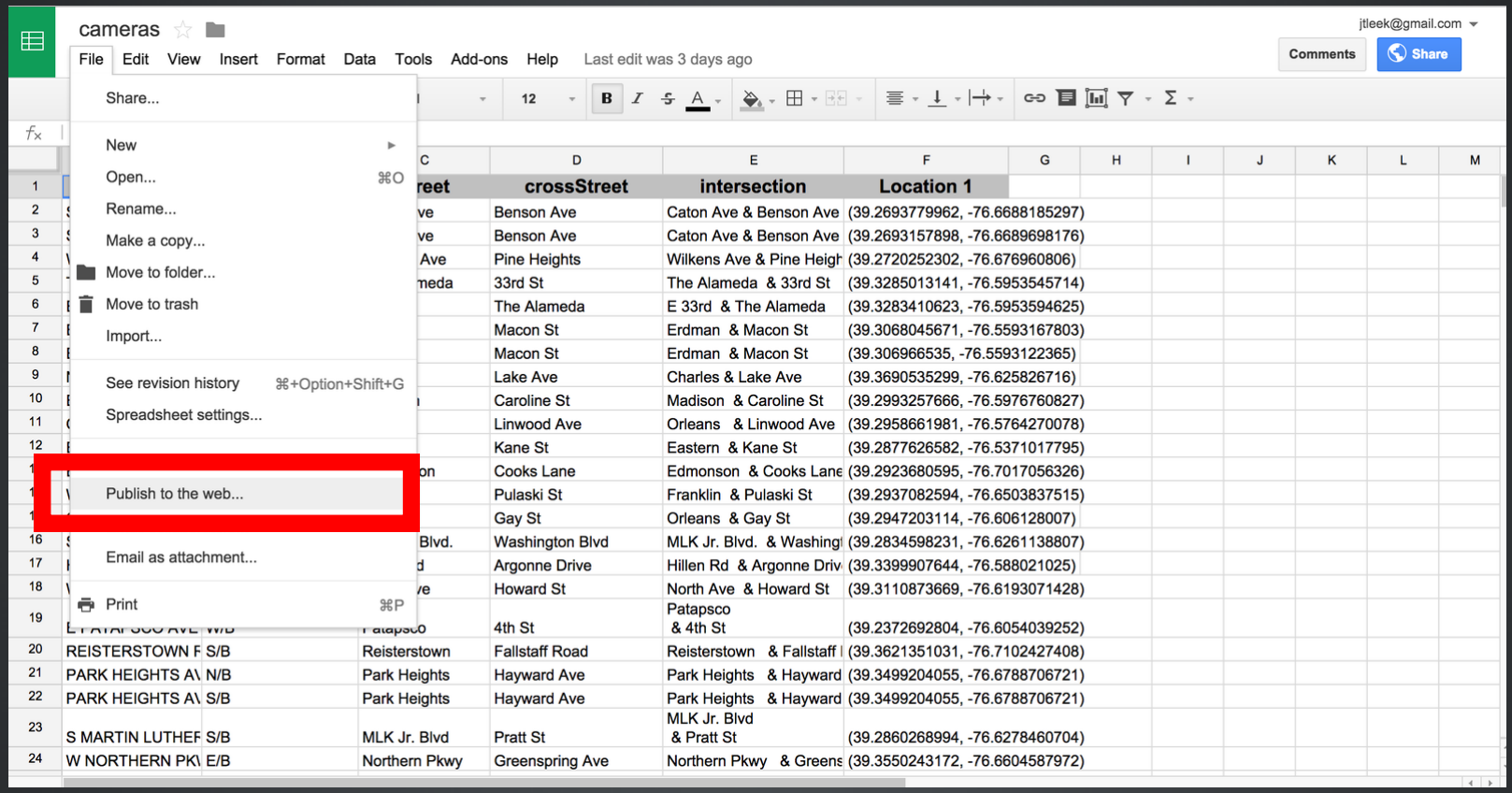
Sheets home

Arial 12

address

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	address	direction	street	crossStreet	intersection	Location 1							
2	S CATON AVE & BI	N/B	Caton Ave	Benson Ave	Caton Ave & Benson Ave	(39.2693779962, -76.6688185297)							
3	S CATON AVE & BI	S/B	Caton Ave	Benson Ave	Caton Ave & Benson Ave	(39.2693157898, -76.6689698176)							
4	WILKENS AVE & P	E/B	Wilkens Ave	Pine Heights	Wilkens Ave & Pine Heigh	(39.2720252302, -76.676960806)							
5	THE ALAMEDA & E	S/B	The Alameda	33rd St	The Alameda & 33rd St	(39.3285013141, -76.5953545714)							
6	E 33RD ST & THE /	E/B	E 33rd	The Alameda	E 33rd & The Alameda	(39.3283410623, -76.5953594625)							
7	ERDMAN AVE & N	E/B	Erdman	Macon St	Erdman & Macon St	(39.3068045671, -76.5593167803)							
8	ERDMAN AVE & N	W/B	Erdman	Macon St	Erdman & Macon St	(39.306966535, -76.5593122365)							
9	N CHARLES ST & E	S/B	Charles	Lake Ave	Charles & Lake Ave	(39.3690535299, -76.625826716)							
10	E MADISON ST & N	W/B	Madison	Caroline St	Madison & Caroline St	(39.2993257666, -76.5976760827)							
11	ORLEANS ST & N I	E/B	Orleans	Linwood Ave	Orleans & Linwood Ave	(39.2958661981, -76.5764270078)							
12	EASTERN AVE & K	E/B	Eastern	Kane St	Eastern & Kane St	(39.2877626582, -76.5371017795)							
13	EDMONDSON AVE	E/B	Edmonson	Cooks Lane	Edmonson & Cooks Lane	(39.2923680595, -76.7017056326)							
14	W FRANKLIN ST &	W/B	Franklin	Pulaski St	Franklin & Pulaski St	(39.2937082594, -76.6503837515)							
15	ORLEANS ST & N (	E/B	Orleans	Gay St	Orleans & Gay St	(39.2947203114, -76.606128007)							
16	S MARTIN LUTHEF	N/B	MLK Jr. Blvd.	Washington Blvd	MLK Jr. Blvd. & Washingt	(39.2834598231, -76.6261138807)							
17	HILLEN RD & ARG	S/B	Hillen Rd	Argonne Drive	Hillen Rd & Argonne Driv	(39.3399907644, -76.588021025)							
18	W NORTH AVE & N	W/B	North Ave	Howard St	North Ave & Howard St	(39.3110873669, -76.6193071428)							
19	E PATAPSCO AVE	W/B	Patapsco	4th St	Patapsco & 4th St	(39.2372692804, -76.6054039252)							
20	REISTERSTOWN F	S/B	Reisterstown	Fallstaff Road	Reisterstown & Fallstaff	(39.3621351031, -76.7102427408)							
21	PARK HEIGHTS AV	N/B	Park Heights	Hayward Ave	Park Heights & Hayward	(39.3499204055, -76.6788706721)							

\https://docs.google.com/spreadsheets/d/1xYzntyuetrm7mOlE7iwUtz1nS3f2uREK6y0vv_AXKN0/pubhtml



\https://docs.google.com/spreadsheets/d/1xYzntyuetrm7mOlE7iwUtz1nS3f2uREK6y0vv_AXKN0/pubhtml

Reading a Google Sheet

```
sheets_url = "https://docs.google.com/spreadsheets/d/1xYzntyuetrm7mOlE7iwUtz1nS3f2uREK6y0vv_AXKN0/p
```

```
gsurl1 = gs_url(sheets_url)
dat = gs_read(gsurl1)

dat

Source: local data frame [80 x 6]

                                address direction
1          S CATON AVE & BENSON AVE         N/B
2          S CATON AVE & BENSON AVE         S/B
```

Google Sheets Lab

http://bit.ly/1Cgzjxb

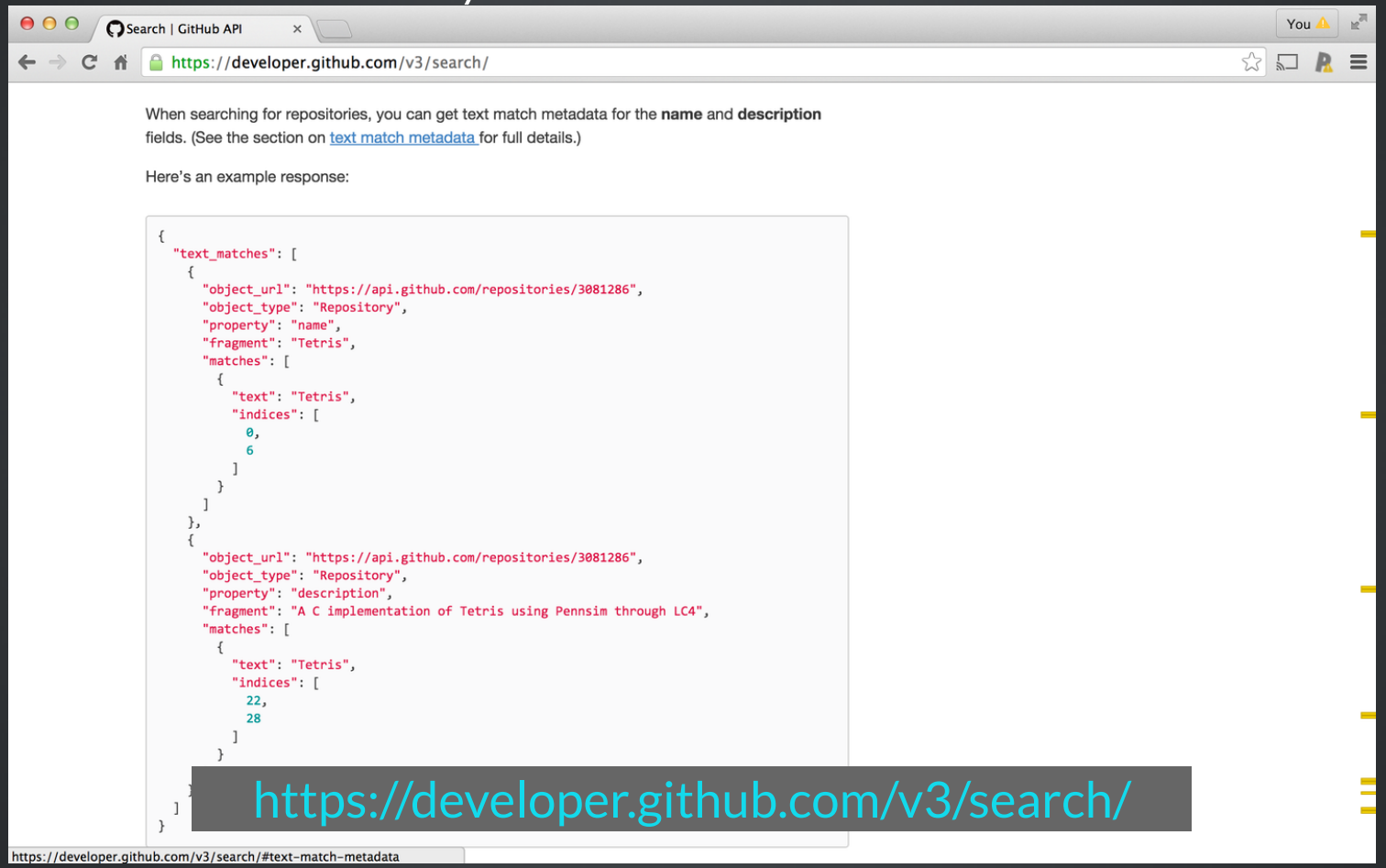
Reading Data: JSON

JSON

```
{
  "firstName": "John",
  "lastName": "Smith",
  "isAlive": true,
  "age": 25,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021-3100"
  },
  "phoneNumbers": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
    {
      "type": "office",
      "number": "646 555-4567"
    }
  ],
  "children": [],
  "spouse": null
}
```

https://en.wikipedia.org/wiki/JSON

Why JSON matters



Reading JSON

```
github_url = "https://api.github.com/users/jtleek/repos"
```

```
install.packages("jsonlite")
library(jsonlite)
jsonData <- fromJSON(github_url)
dim(jsonData)
[1] 30 67

jsonData$name
 [1] "ballgown"                "capitalIn21stCen
 [4] "dataanalysis"            "datascientist"
 [7] "datawomenontwitter"      "derfinder"
[10] "DSM"                     "EDA-Project"
```

Data frame structure from JSON

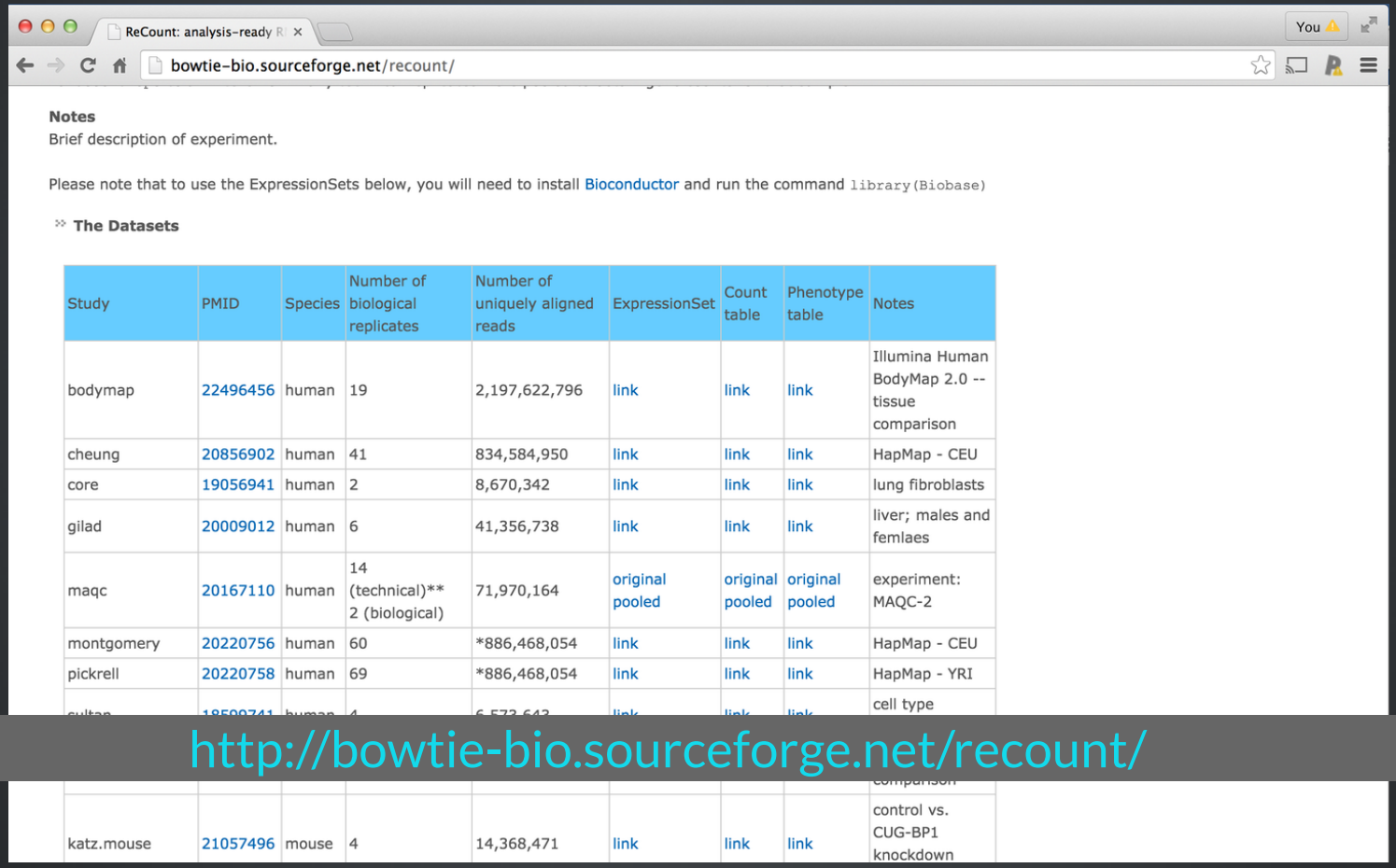
```
table(sapply(jsonData,class))
 character data.frame    integer    logical
        50          1          9          7

dim(jsonData$owner)
[1] 30 17

names(jsonData$owner)
 [1] "login"                 "id"
 [5] "url"                   "html_url"
 [9] "gists_url"             "starred_url"
```

Reading Data: HTML

This is data



Notes
Brief description of experiment.

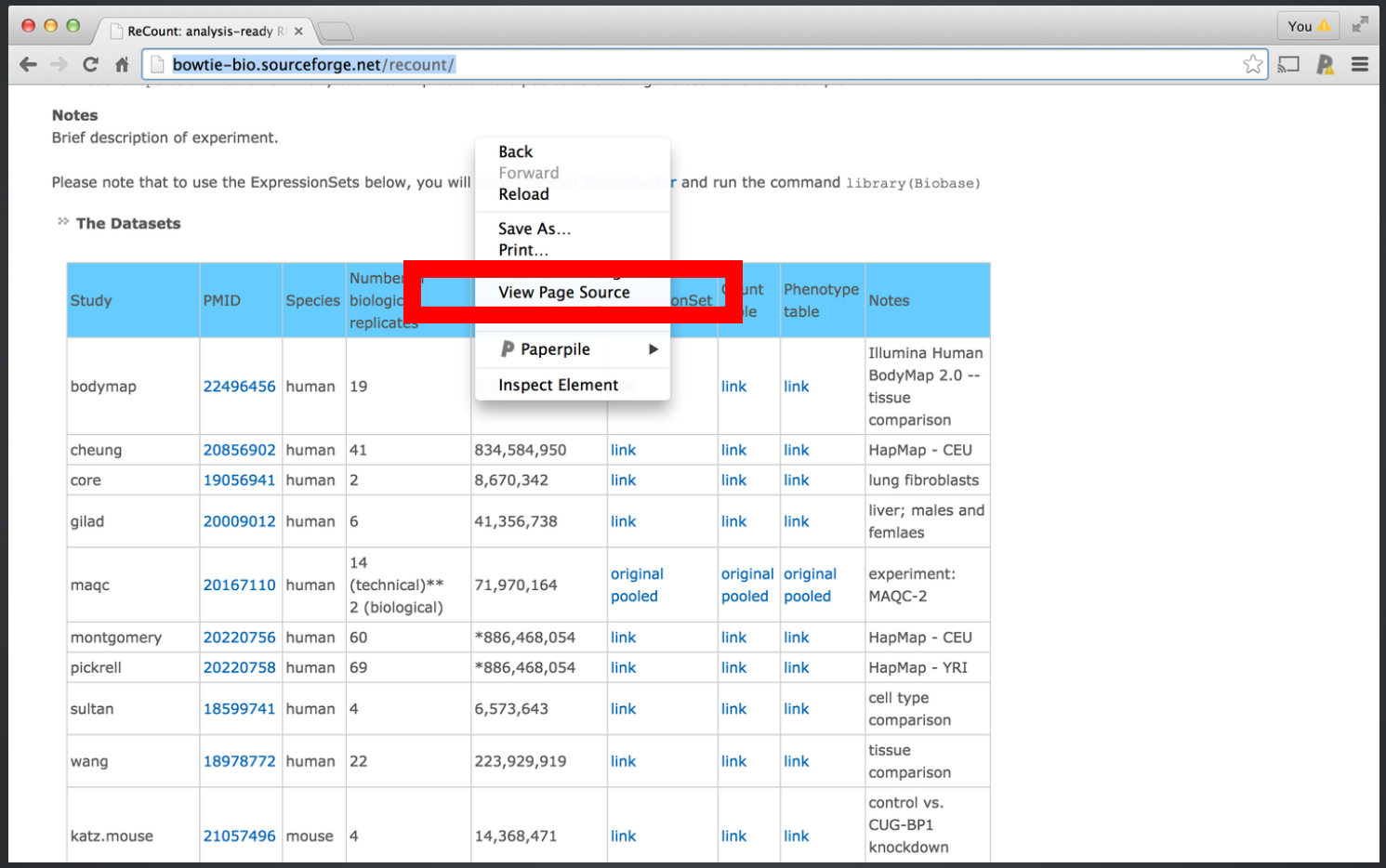
Please note that to use the ExpressionSets below, you will need to install Bioconductor and run the command `library(Biobase)`

The Datasets

Study	PMID	Species	Number of biological replicates	Number of uniquely aligned reads	ExpressionSet	Count table	Phenotype table	Notes
bodymap	22496456	human	19	2,197,622,796	link	link	link	Illumina Human BodyMap 2.0 -- tissue comparison
cheung	20856902	human	41	834,584,950	link	link	link	HapMap - CEU
core	19056941	human	2	8,670,342	link	link	link	lung fibroblasts
gilad	20009012	human	6	41,356,738	link	link	link	liver; males and femlaes
maqc	20167110	human	14 (technical)** 2 (biological)	71,970,164	original pooled	original pooled	original pooled	experiment: MAQC-2
montgomery	20220756	human	60	*886,468,054	link	link	link	HapMap - CEU
pickrell	20220758	human	69	*886,468,054	link	link	link	HapMap - YRI
sultan	18599741	human	4	6,573,643	link	link	link	cell type comparison
katz.mouse	21057496	mouse	4	14,368,471	link	link	link	control vs. CUG-BP1 knockdown

http://bowtie-bio.sourceforge.net/recount/

View the source



Study	PMID	Species	Number biological replicates		onSet	unt ble	Phenotype table	Notes
bodymap	22496456	human	19			link	link	Illumina Human BodyMap 2.0 -- tissue comparison
cheung	20856902	human	41	834,584,950	link	link	link	HapMap - CEU
core	19056941	human	2	8,670,342	link	link	link	lung fibroblasts
gilad	20009012	human	6	41,356,738	link	link	link	liver; males and femlaes
maqc	20167110	human	14 (technical)** 2 (biological)	71,970,164	original pooled	original pooled	original pooled	experiment: MAQC-2
montgomery	20220756	human	60	*886,468,054	link	link	link	HapMap - CEU
pickrell	20220758	human	69	*886,468,054	link	link	link	HapMap - YRI
sultan	18599741	human	4	6,573,643	link	link	link	cell type comparison
wang	18978772	human	22	223,929,919	link	link	link	tissue comparison
katz.mouse	21057496	mouse	4	14,368,471	link	link	link	control vs. CUG-BP1 knockdown

What the computer sees

```
Brief description of experiment.<br \><br \>

Please note that to use the ExpressionSets below, you will need to install <a href="http://www.bioconductor.org/">Bioconductor</a> and run the command
<tt>library(Biobase)</tt>

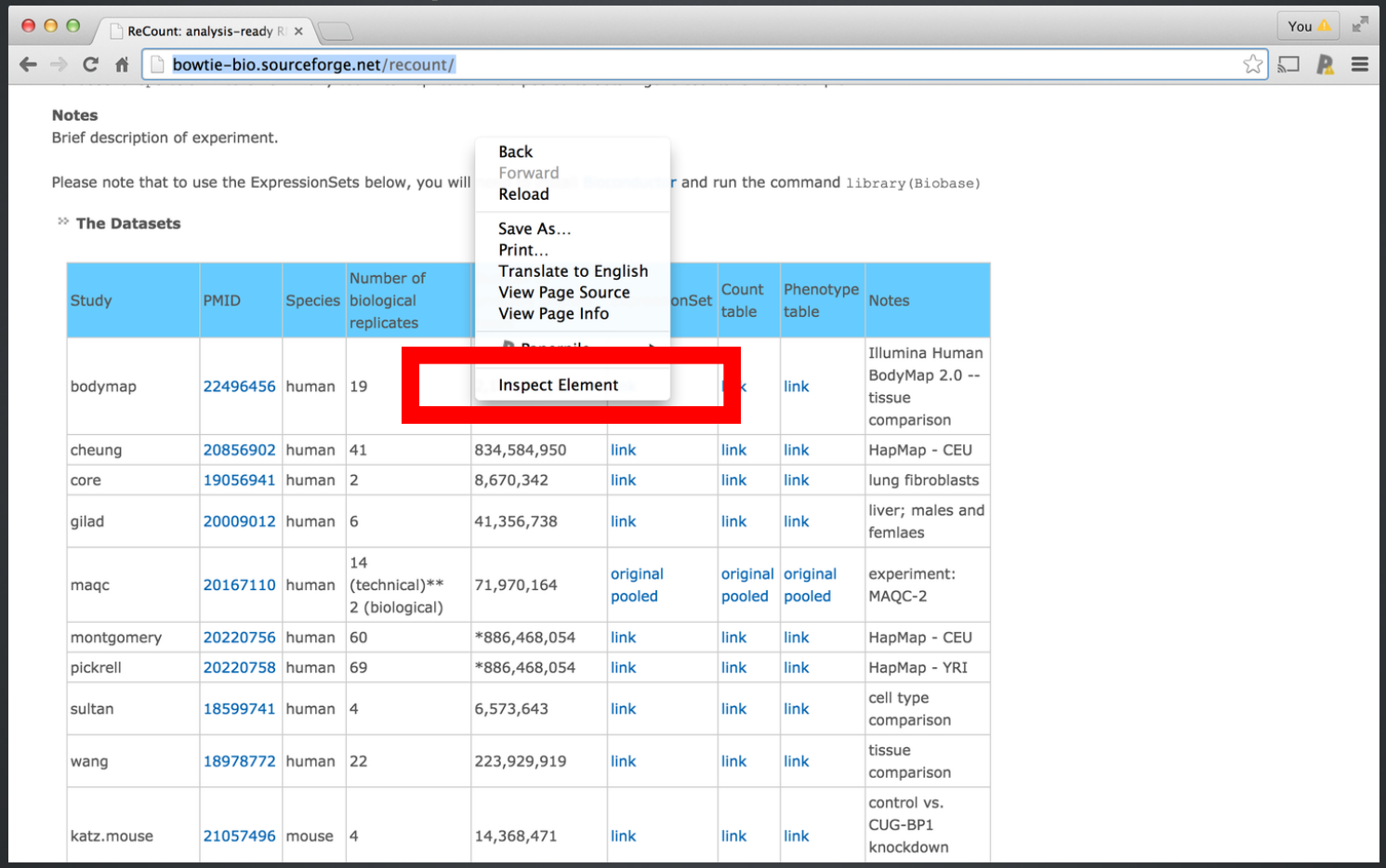
<h3>The Datasets</h3>
<div id="recounttab">
<table class="sortable"><tbody>
<tr>
<td>Study</td>
<td>PMID</td>
<td>Species</td>
<td>Number of biological replicates</td>
<td>Number of uniquely aligned reads</td>
<td>ExpressionSet</td>
<td>Count table</td>
<td>Phenotype table</td>
<td>Notes</td>
</tr>

<tr>
<td>bodymap</td>
  <td><a href="http://www.ncbi.nlm.nih.gov/pubmed/22496456">22496456</a></td>
  <td>human</td>
  <td>19</td>
  <td>2,197,622,796</td>
  <td><a href="./ExpressionSets/bodymap_eset.RData">link </a></td>
  <td><a href="./countTables/bodymap_count_table.txt">link</a></td>
  <td><a href="./phenotypeTables/bodymap_phenodata.txt">link</a></td>
  <td>Illumina Human BodyMap 2.0 -- tissue comparison</td></tr>

<tr>
<td>cheung</td>
  <td><a href="http://www.ncbi.nlm.nih.gov/pubmed?term=20856902">20856902</a></td>
  <td>human</td>
  <td>41</td>
  <td>834,584,950</td>
  <td><a href="./ExpressionSets/cheung_eset.RData">link </a></td>
  <td><a href="./countTables/cheung_count_table.txt">link</a></td>
  <td><a href="./phenotypeTables/cheung_phenodata.txt">link</a></td>
  <td>HapMap - CEU</td></tr>

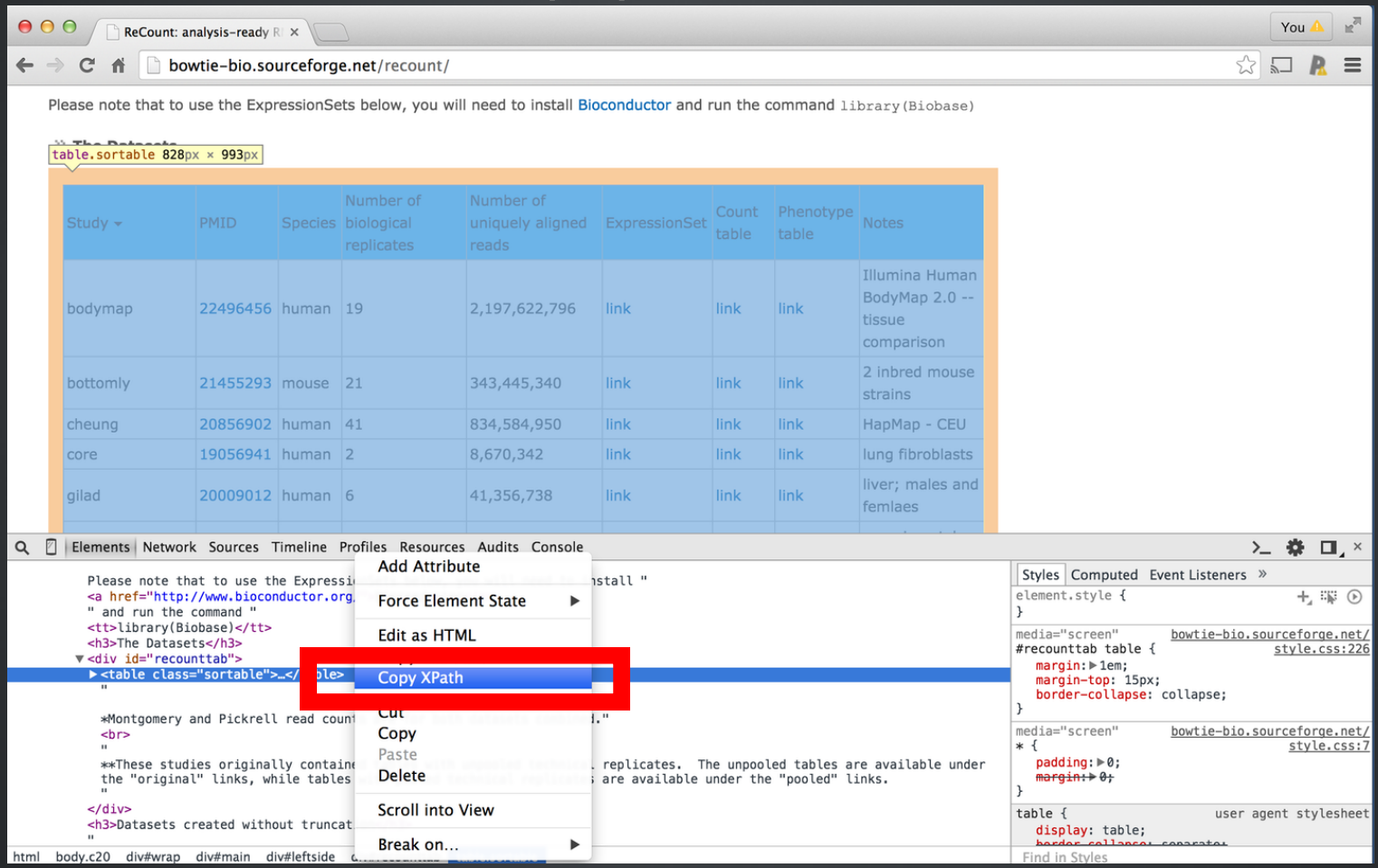
<tr><td>core</td>
  <td><a href="http://www.ncbi.nlm.nih.gov/pubmed?term=19056941">19056941</a></td>
  <td>human</td>
```

Inspect element



Study	PMID	Species	Number of biological replicates		onSet	Count table	Phenotype table	Notes
bodymap	22496456	human	19				link	Illumina Human BodyMap 2.0 -- tissue comparison
cheung	20856902	human	41	834,584,950	link	link	link	HapMap - CEU
core	19056941	human	2	8,670,342	link	link	link	lung fibroblasts
gilad	20009012	human	6	41,356,738	link	link	link	liver; males and femlaes
maqc	20167110	human	14 (technical)** 2 (biological)	71,970,164	original pooled	original pooled	original pooled	experiment: MAQC-2
montgomery	20220756	human	60	*886,468,054	link	link	link	HapMap - CEU
pickrell	20220758	human	69	*886,468,054	link	link	link	HapMap - YRI
sultan	18599741	human	4	6,573,643	link	link	link	cell type comparison
wang	18978772	human	22	223,929,919	link	link	link	tissue comparison
katz.mouse	21057496	mouse	4	14,368,471	link	link	link	control vs. CUG-BP1 knockdown

Copy XPath


ReCount: analysis-ready
bowtie-bio.sourceforge.net/recount/
Please note that to use the ExpressionSets below, you will need to install Bioconductor and run the command library(Biobase)
table.sortable 828px × 993px
Study
PMID
Species
Number of biological replicates
Number of uniquely aligned reads
ExpressionSet
Count table
Phenotype table
Notes
bodymap 22496456 human 19 2,197,622,796 link link link Illumina Human BodyMap 2.0 -- tissue comparison
bottomly 21455293 mouse 21 343,445,340 link link link 2 inbred mouse strains
cheung 20856902 human 41 834,584,950 link link link HapMap - CEU
core 19056941 human 2 8,670,342 link link link lung fibroblasts
gilad 20009012 human 6 41,356,738 link link link liver; males and femlaes
Elements Network Sources Timeline Profiles Resources Audits Console
Add Attribute
Force Element State
Edit as HTML
Copy XPath
Cut
Copy
Paste
Delete
Scroll into View
Break on...
Styles Computed Event Listeners
element.style {
media="screen" bowtie-bio.sourceforge.net/style.css:226
#recounttab table {
margin: 1em;
margin-top: 15px;
border-collapse: collapse;
media="screen" bowtie-bio.sourceforge.net/style.css:7
* {
padding: 0;
table { user agent stylesheet
display: table;
Find in Styles
html body.c20 div#wrap div#main div#leftside

rvest package

```
recount_url = "http://bowtie-bio.sourceforge.net/recount
```

```
install.packages("rvest")
library(rvest)
htmlfile = html(recount_url)

nds = html_nodes(htmlfile,
                 xpath='//*[@id="recounttab"]/table'
dat = html_table(nds)
dat = as.data.frame(dat)
```

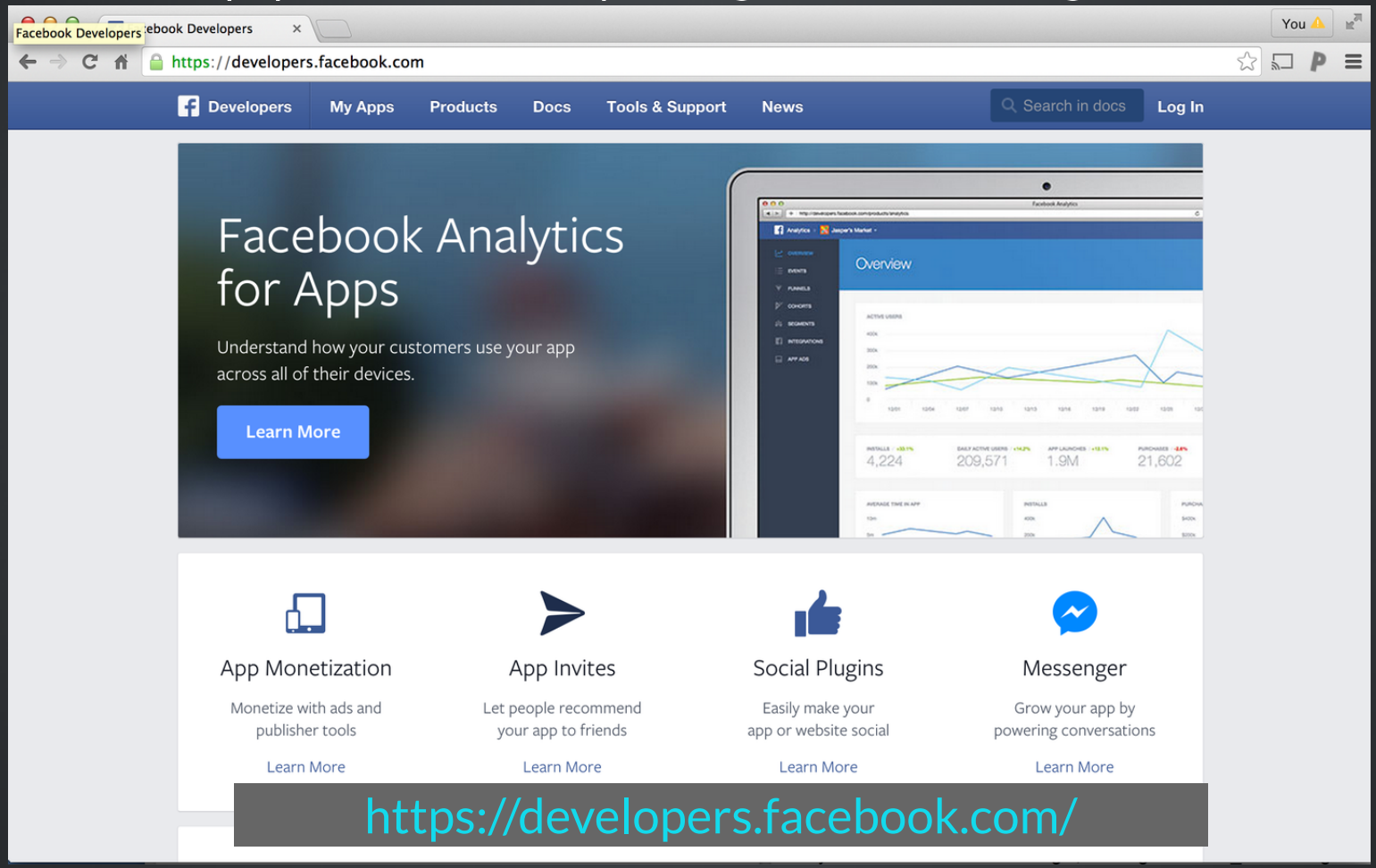
Data from the internetz!

```
head(dat)

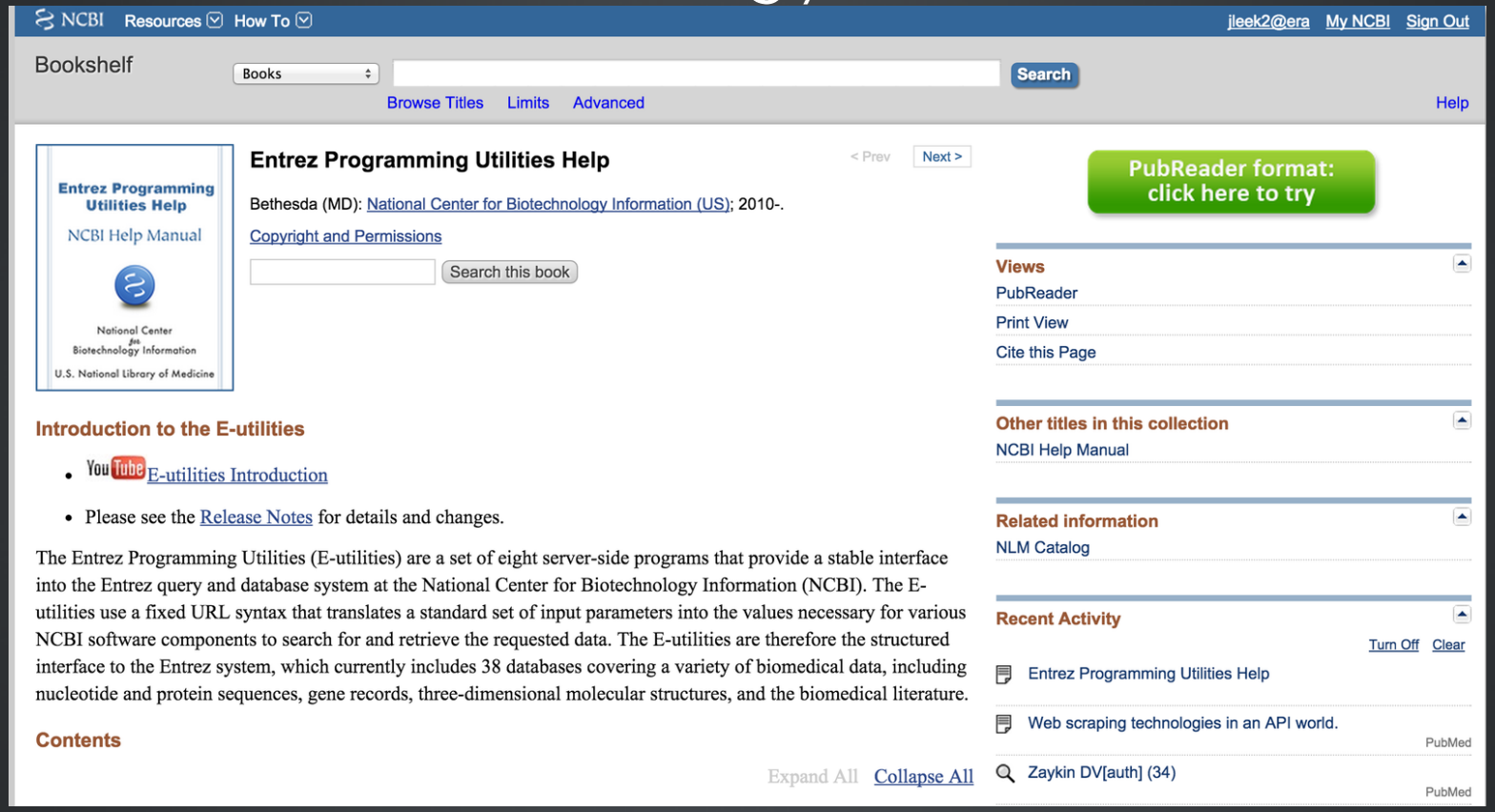
       X1        X2      X3                                      X4
1   Study      PMID Species  Number of biological replicates
2 bodymap 22496456   human                                      19
3  cheung 20856902   human                                      41
4    core 19056941   human                                       2
5   gilad 20009012   human                                       6
```

Reading Data: APIs

APIs: Application programming interface

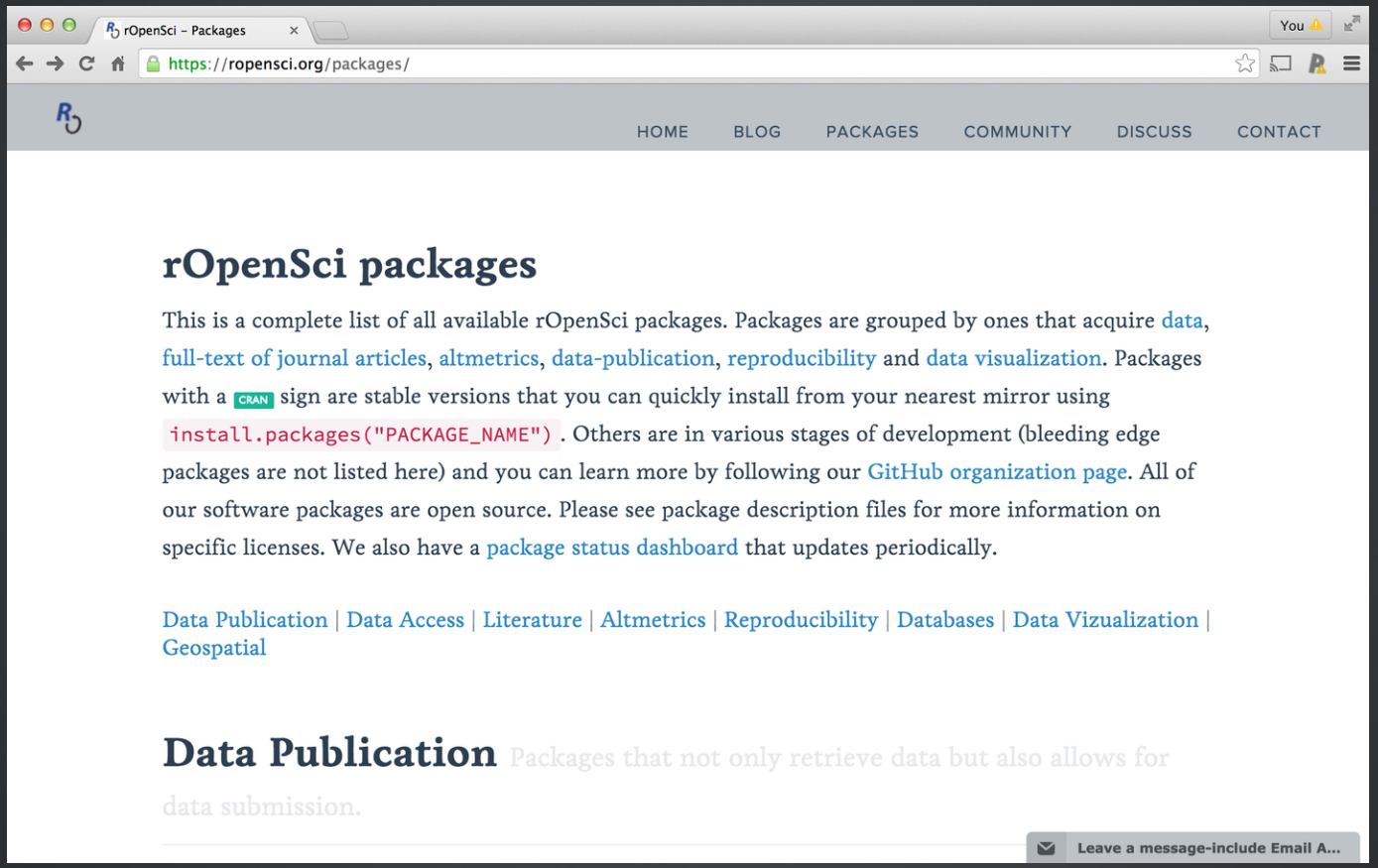


In Biology Too!



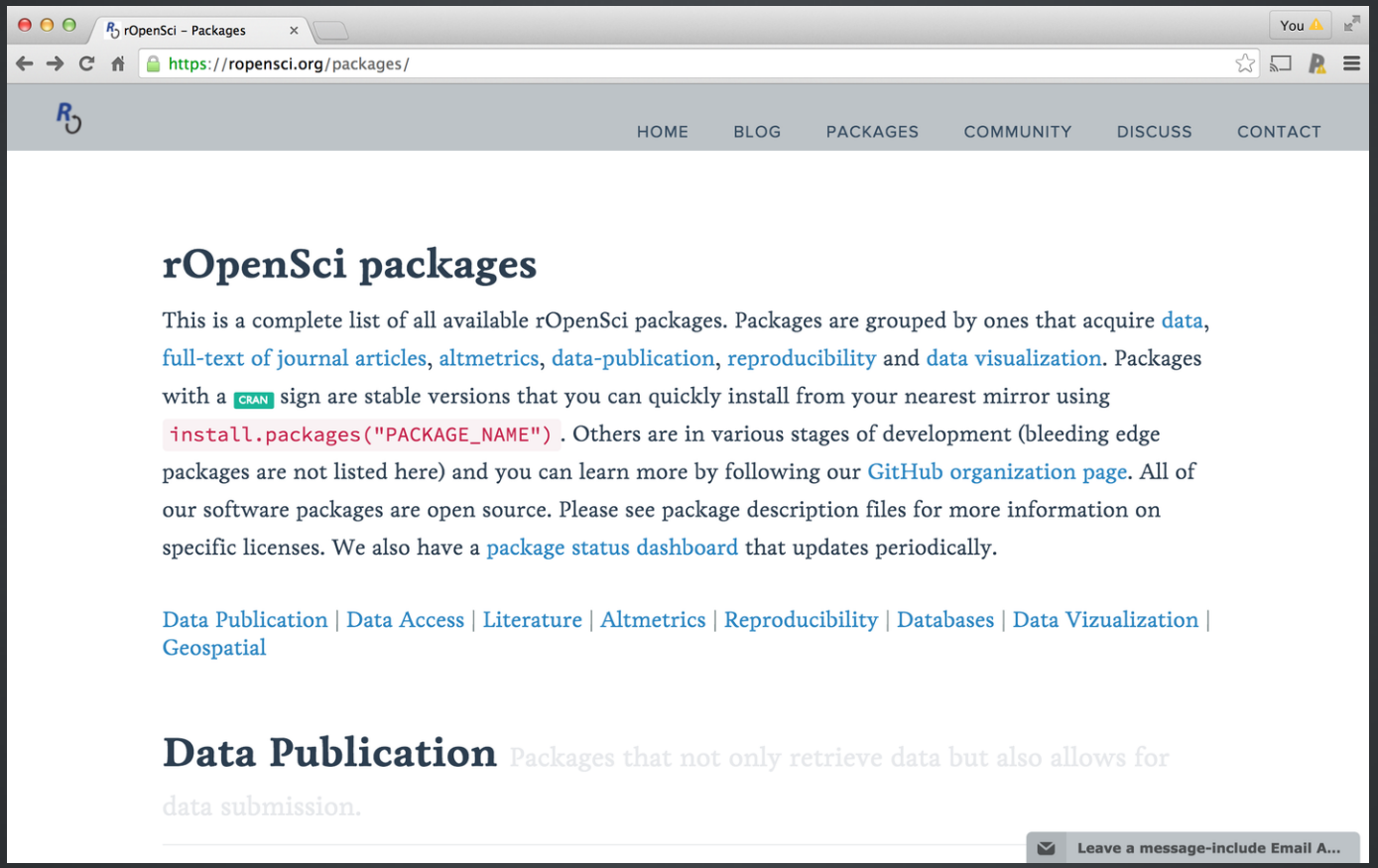
http://www.ncbi.nlm.nih.gov/books/NBK25501/

Step 0: Did someone already do this?



https://ropensci.org/

Figshare



https://ropensci.org/

Figshare API wrapper

```
install.packages("figshare")
library(figshare)
leeksearch = fs_search("Leek")

length(leeksearch)
[1] 56

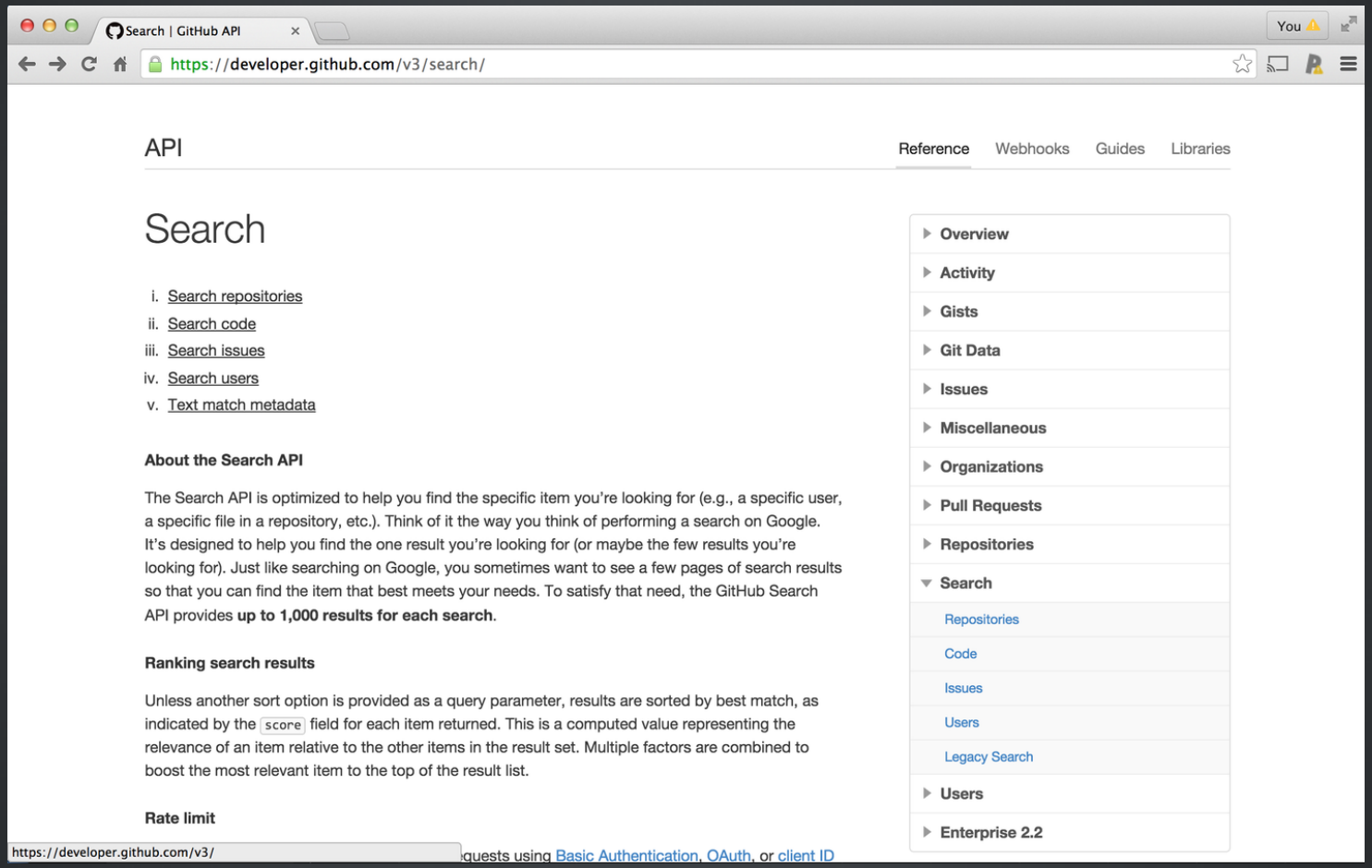
leeksearch[[1]]
$article_id
[1] 1405638

$title
[1] "GEUVADIS expressed regions coverage matrix"

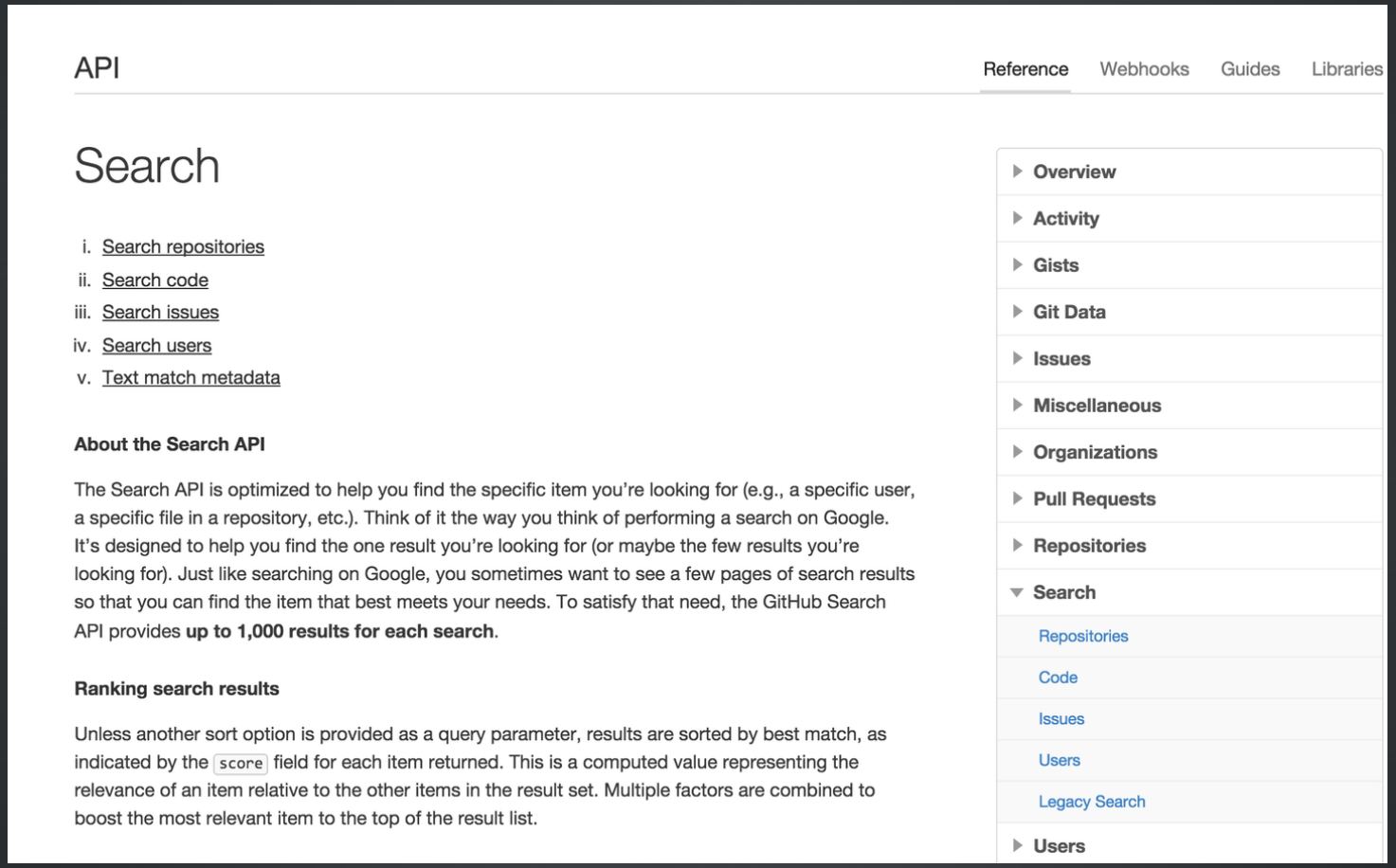
$DOI
[1] "http://dx.doi.org/10.6084/m9.figshare.1405638"

$description
```

Do it yourself



Read the docs



API

Reference Webhooks Guides Libraries

Search

i. Search repositories
ii. Search code
iii. Search issues
iv. Search users
v. Text match metadata

About the Search API

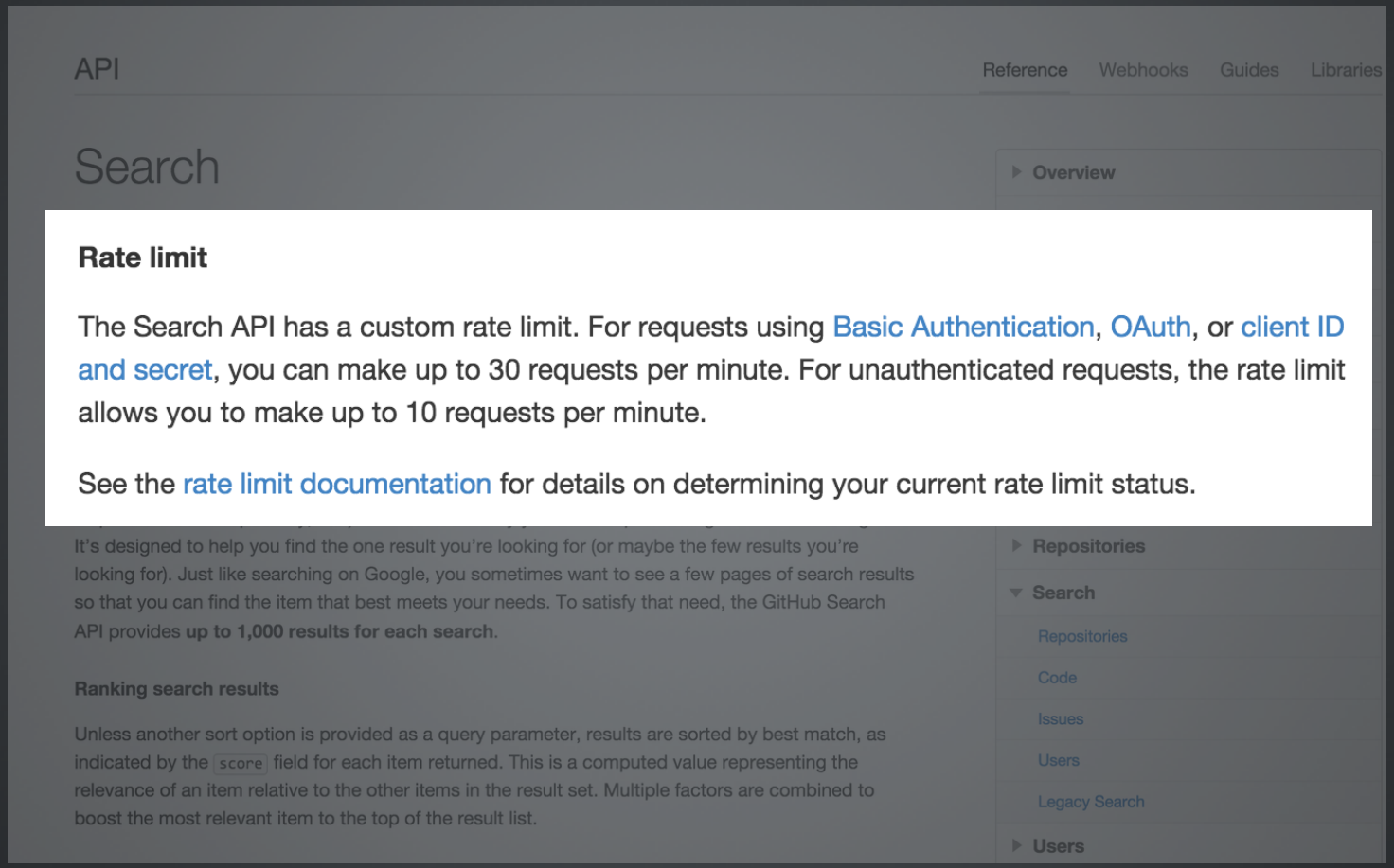
The Search API is optimized to help you find the specific item you're looking for (e.g., a specific user, a specific file in a repository, etc.). Think of it the way you think of performing a search on Google. It's designed to help you find the one result you're looking for (or maybe the few results you're looking for). Just like searching on Google, you sometimes want to see a few pages of search results so that you can find the item that best meets your needs. To satisfy that need, the GitHub Search API provides **up to 1,000 results for each search**.

Ranking search results

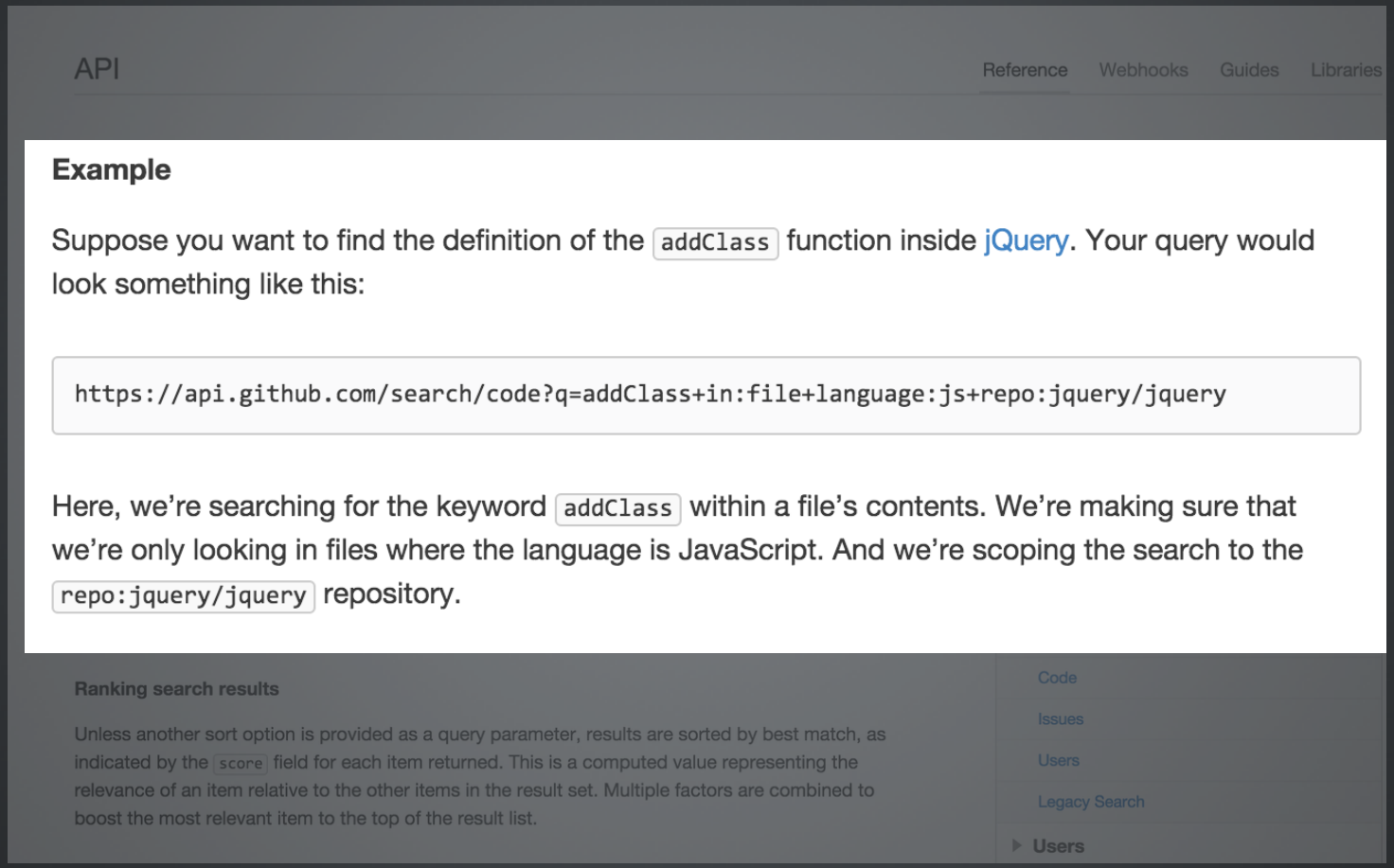
Unless another sort option is provided as a query parameter, results are sorted by best match, as indicated by the `score` field for each item returned. This is a computed value representing the relevance of an item relative to the other items in the result set. Multiple factors are combined to boost the most relevant item to the top of the result list.

- Overview
- Activity
- Gists
- Git Data
- Issues
- Miscellaneous
- Organizations
- Pull Requests
- Repositories
- Search
  - Repositories
  - Code
  - Issues
  - Users
  - Legacy Search
- Users

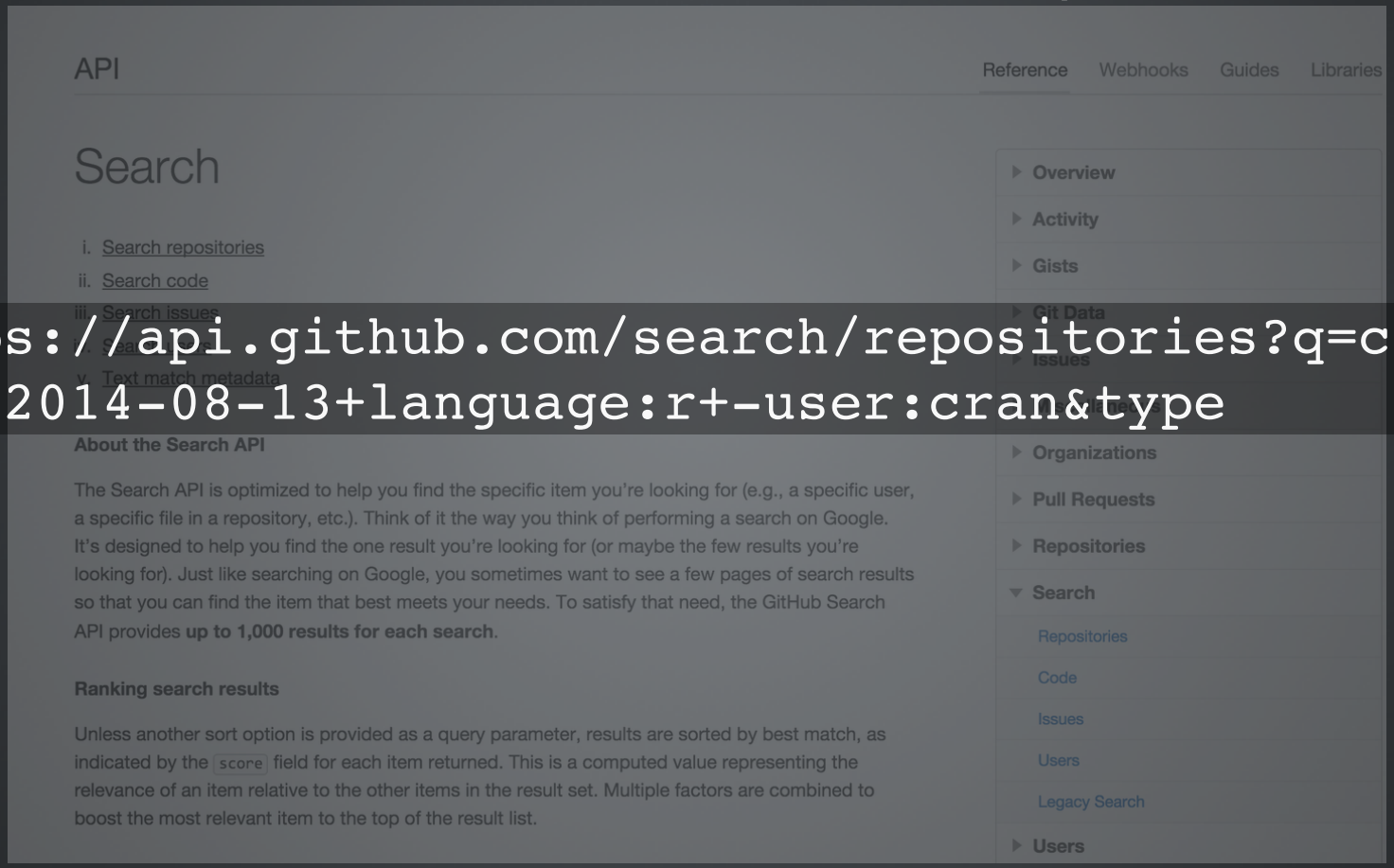
Read the docs



Read the docs

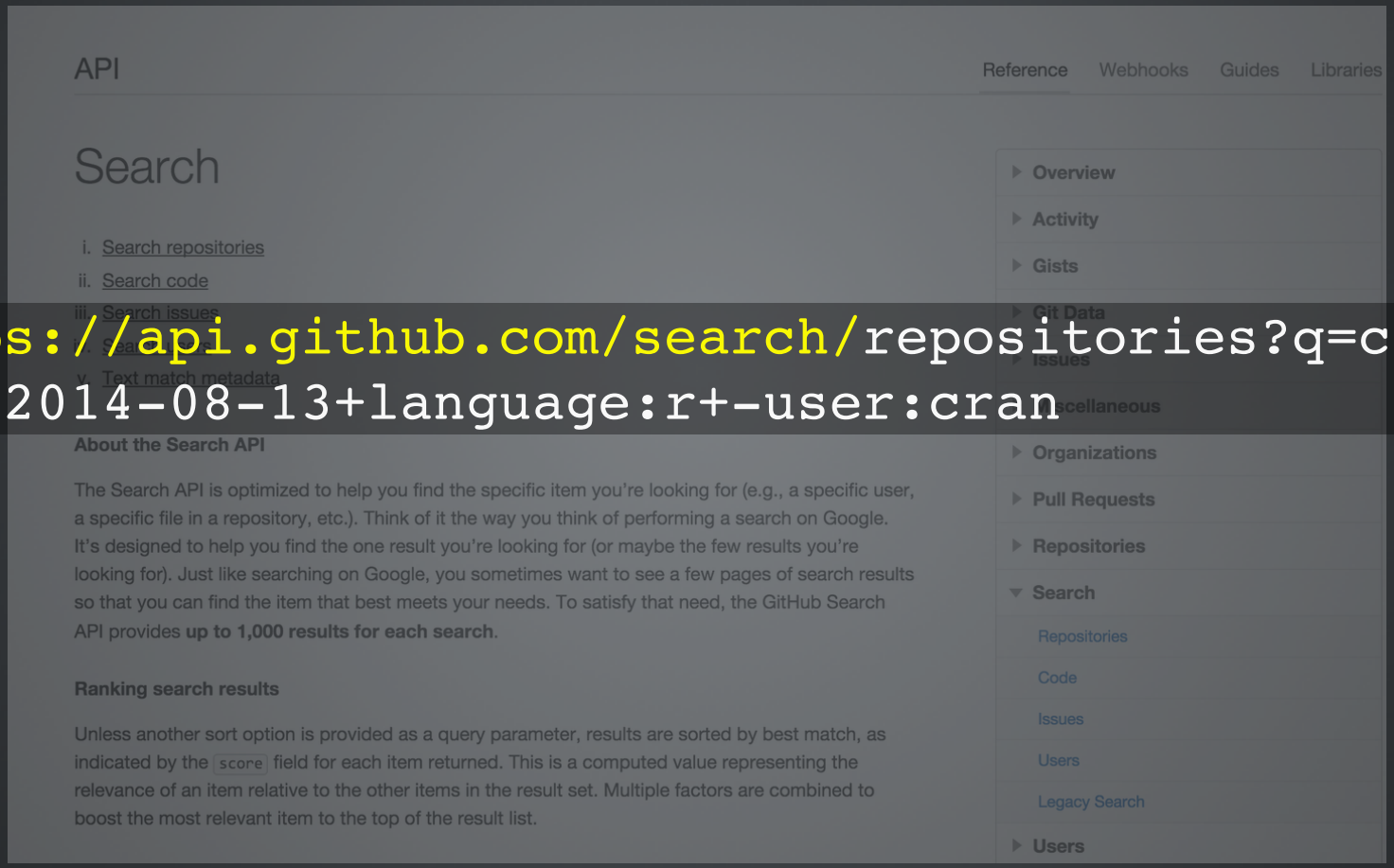


A dissected example



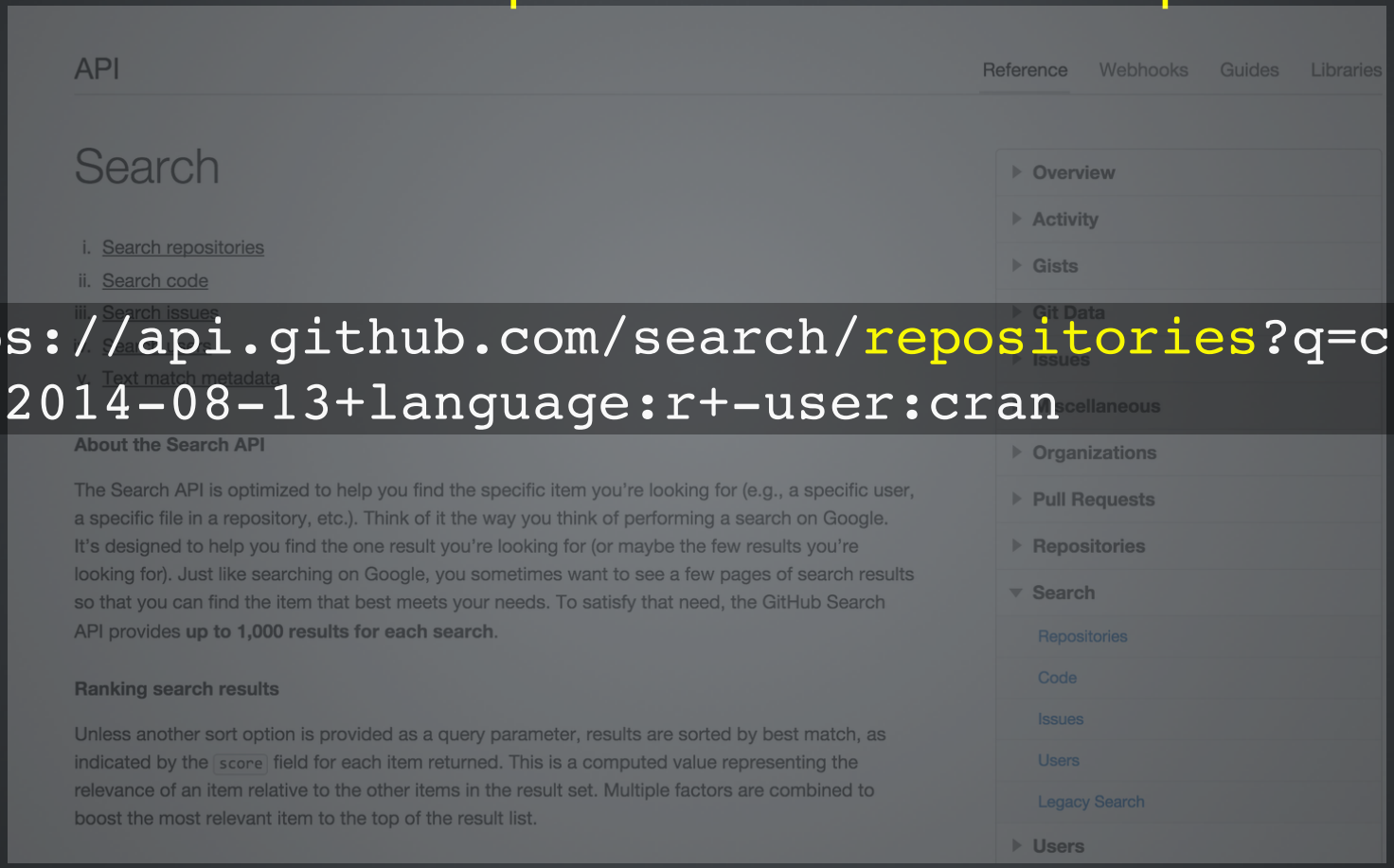
https://api.github.com/search/repositories?q=created:2014-08-13+language:r+-user:cran&type

The Base URL



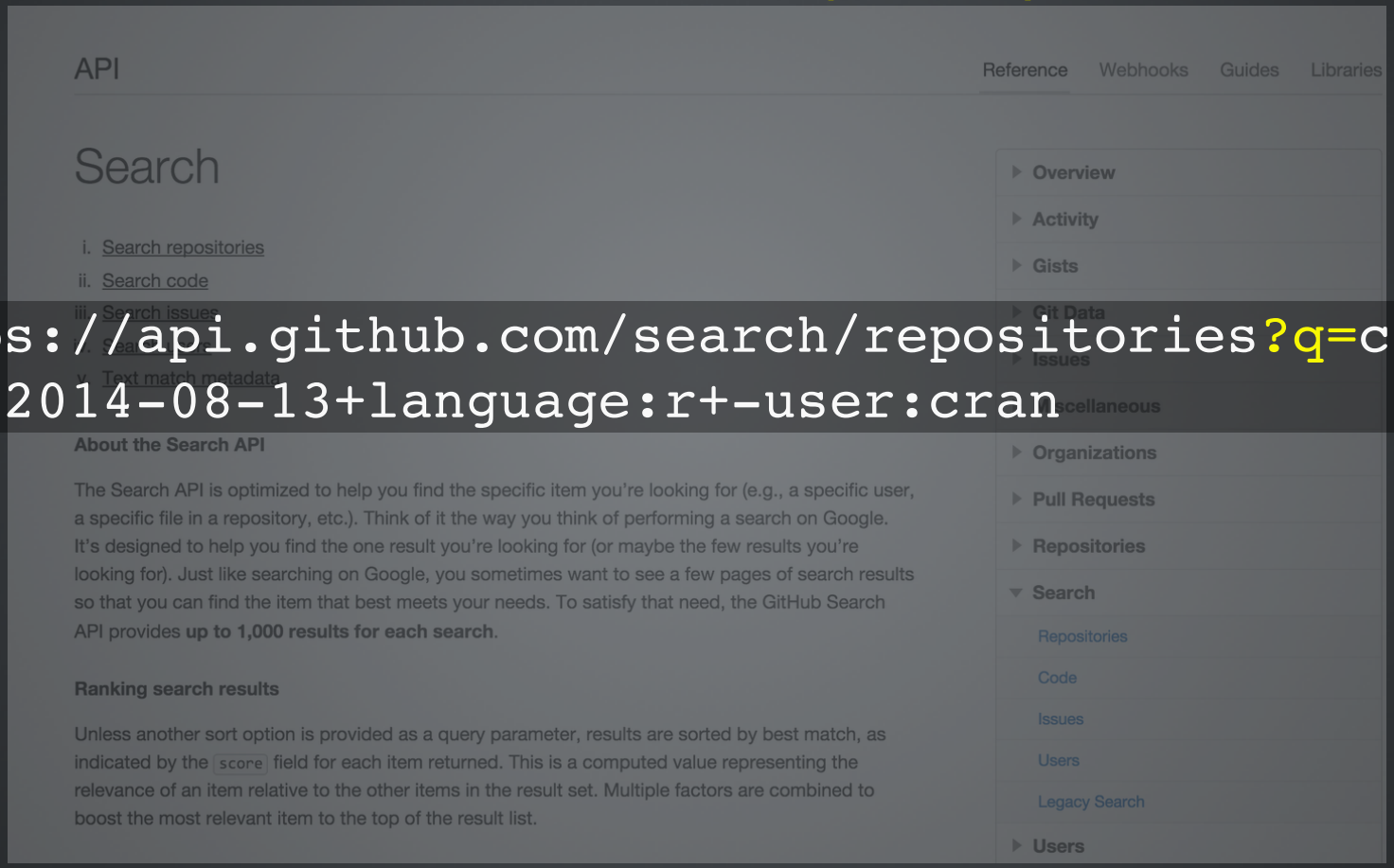
https://api.github.com/search/repositories?q=created:2014-08-13+language:r+-user:cran

Search repositories (repos)



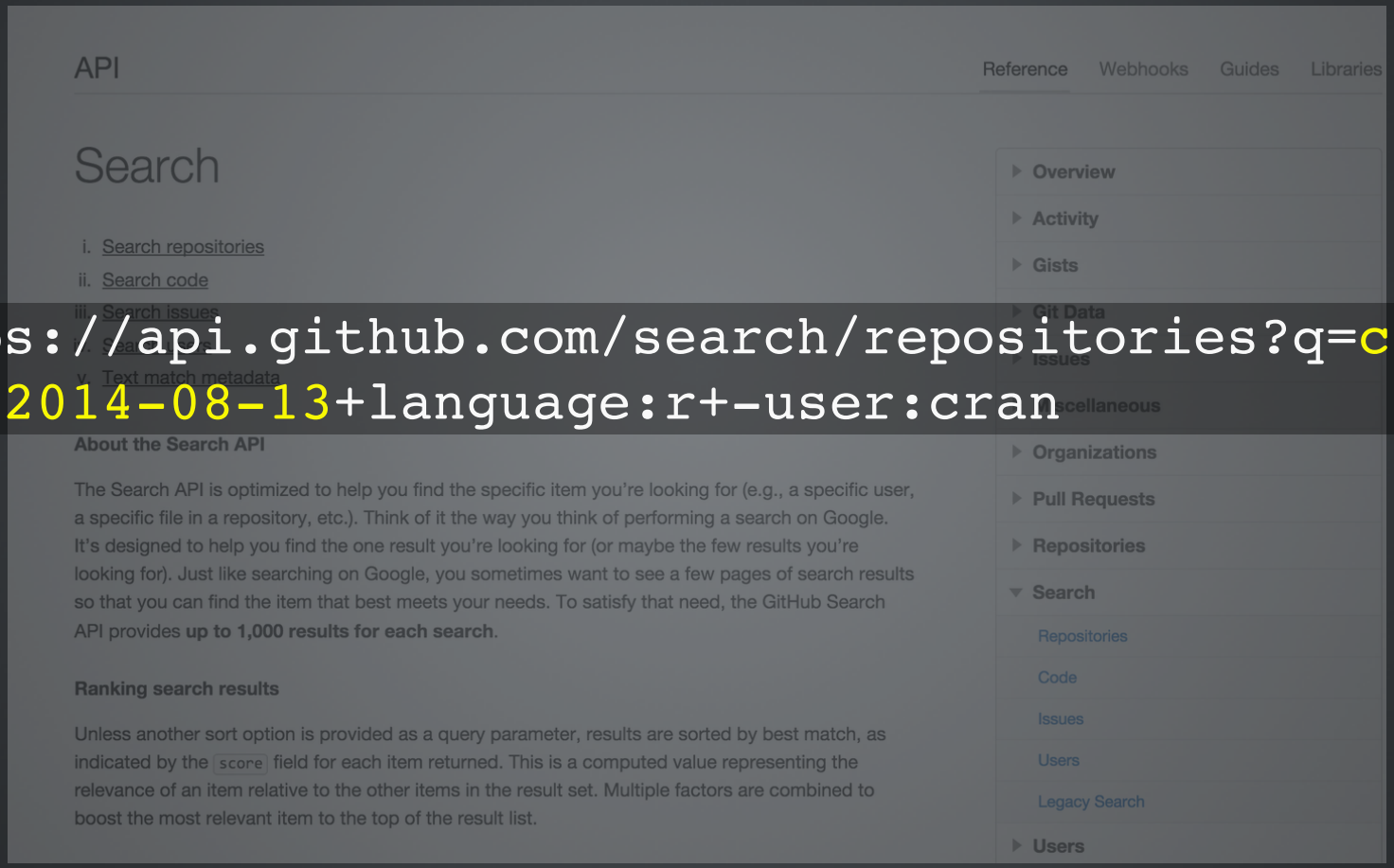
https://api.github.com/search/repositories?q=created:2014-08-13+language:r+-user:cran

Create a query



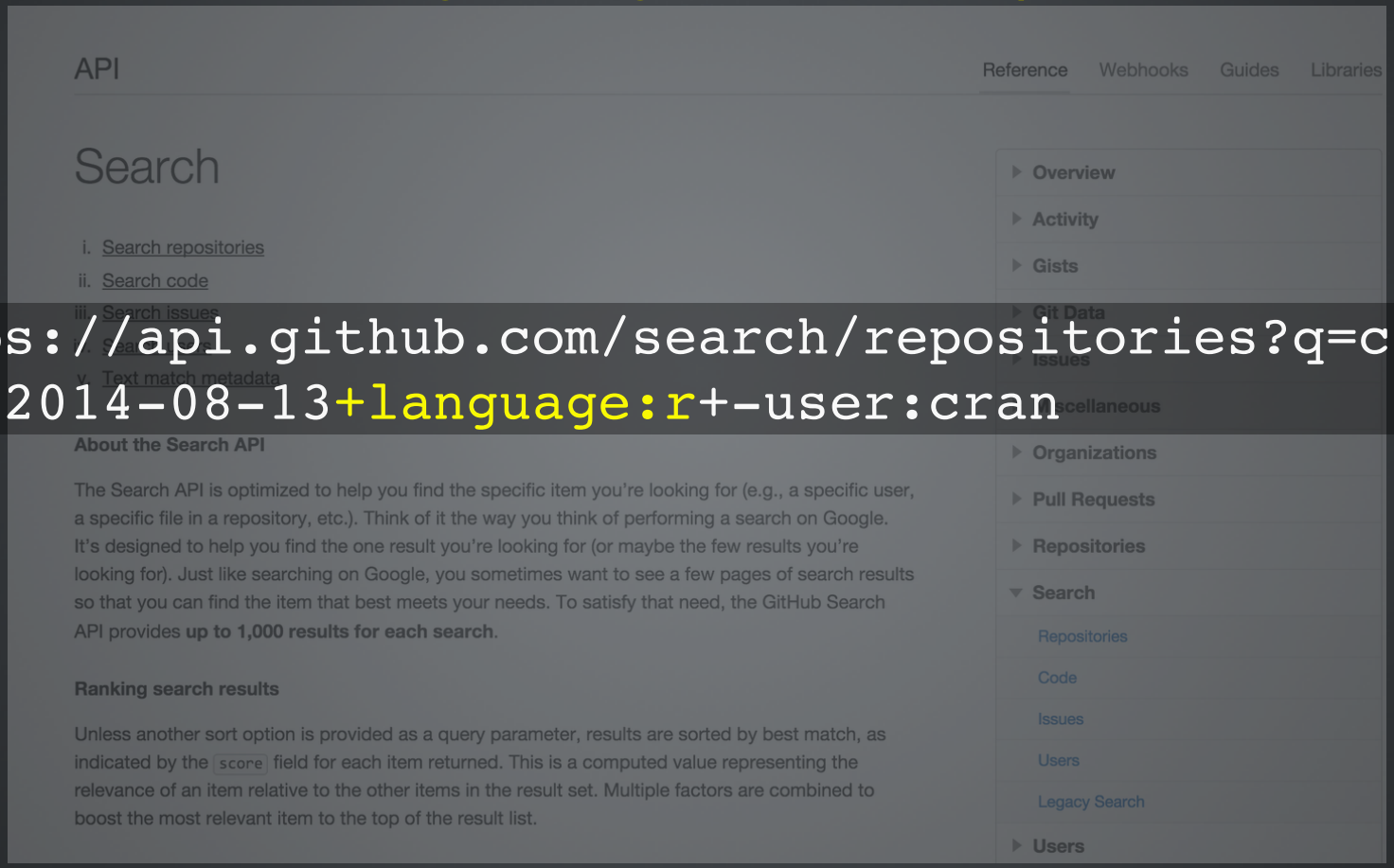
https://api.github.com/search/repositories?q=created:2014-08-13+language:r+-user:cran

Date the repo was created



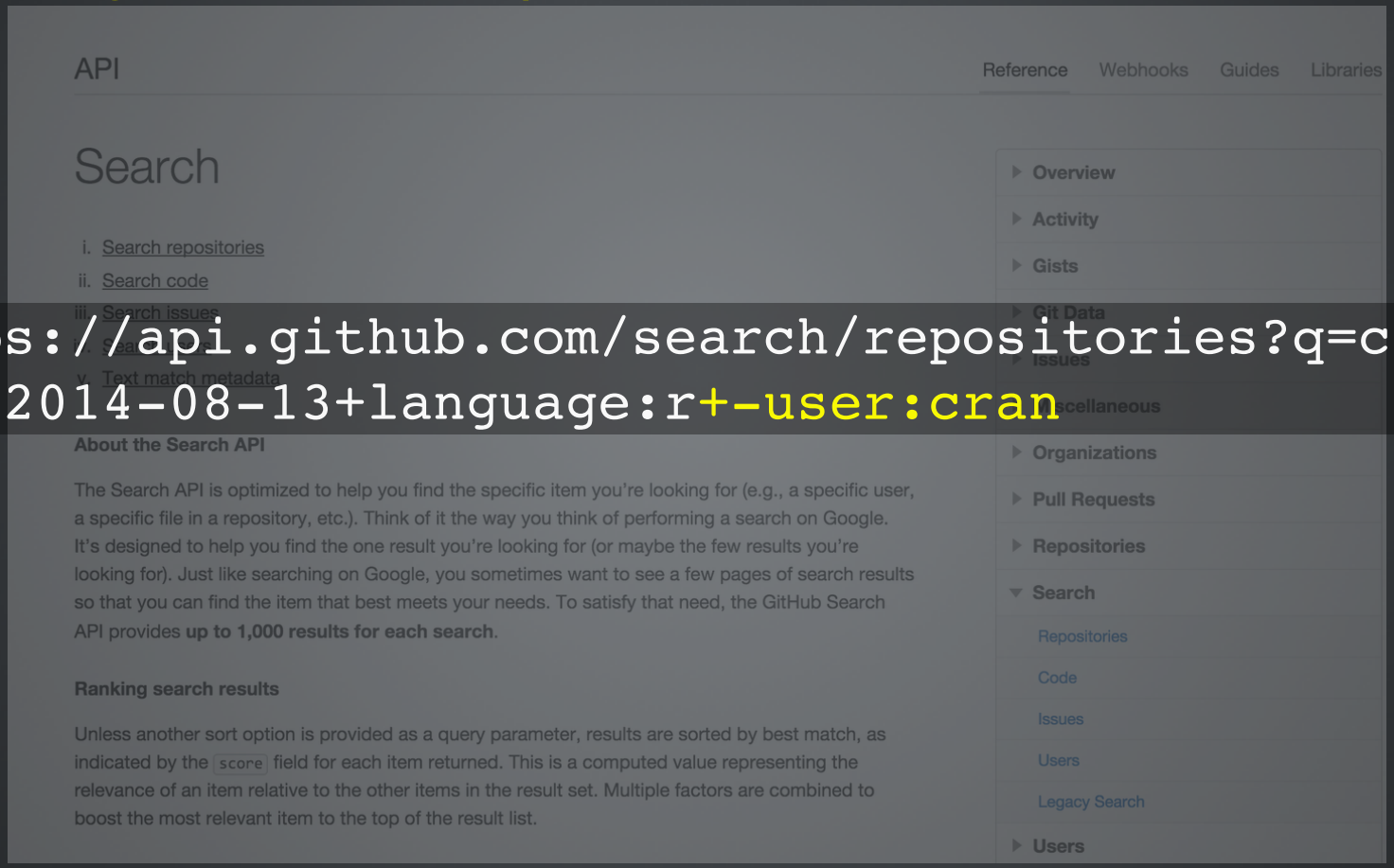
https://api.github.com/search/repositories?q=created:2014-08-13+language:r+-user:cran

Language of repo



https://api.github.com/search/repositories?q=created:2014-08-13+language:r+-user:cran

Ignore repos from "cran"




```
https://api.github.com/search/repositories?q=created:2014-08-13+language:r+-user:cran
```

httr package

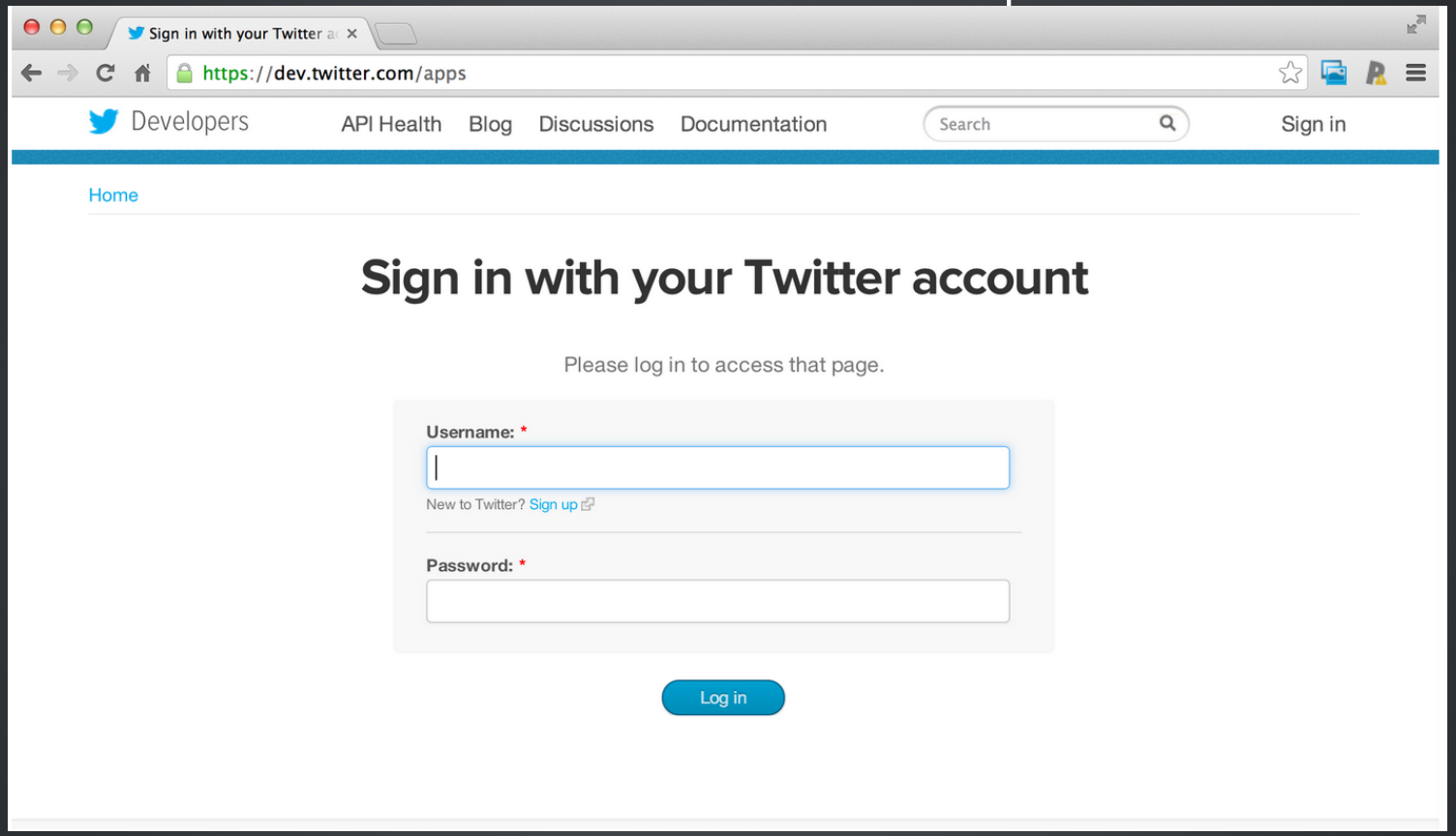
```
query_url = "https://api.github.com/search/repositories?q=created:2014-08-13+language:r+-user:cran"
```

```
install.packages("httr")
library(httr)
req = GET(query_url)
names(content(req))
[1] "total_count"        "incomplete_results"

content(req)$items[[1]]
$id
[1] 22907468

$name
[1] "computel"
```

Not all APIs are "open"



https://dev.twitter.com/apps

Not all APIs are "open"

```
myapp = oauth_app("twitter",
                  key="yourConsumerKeyHere",secret="yourConsumerSecretHere")
sig = sign_oauth1.0(myapp,
                    token = "yourTokenHere",
                     token_secret = "yourTokenSecretHere")
homeTL = GET("https://api.twitter.com/1.1/statuses/home_timeline.json", sig)
```

But you can often get cool data

```
json1 = content(homeTL)
json2 = jsonlite::fromJSON(toJSON(json1))
json2[1,1:4]

                      created_at           id              id_str
1 Mon Jan 13 05:18:04 +0000 2014 4.225984e+17 422598398940684288

1 Now that P. Norvig's regex golf IPython notebook hit Slashdot, let's see if
```

Web + API Lab

http://bit.ly/1S2oAsj