SISBID: Accessing Biomedical Big Data

Jeff Leek
@jtleek

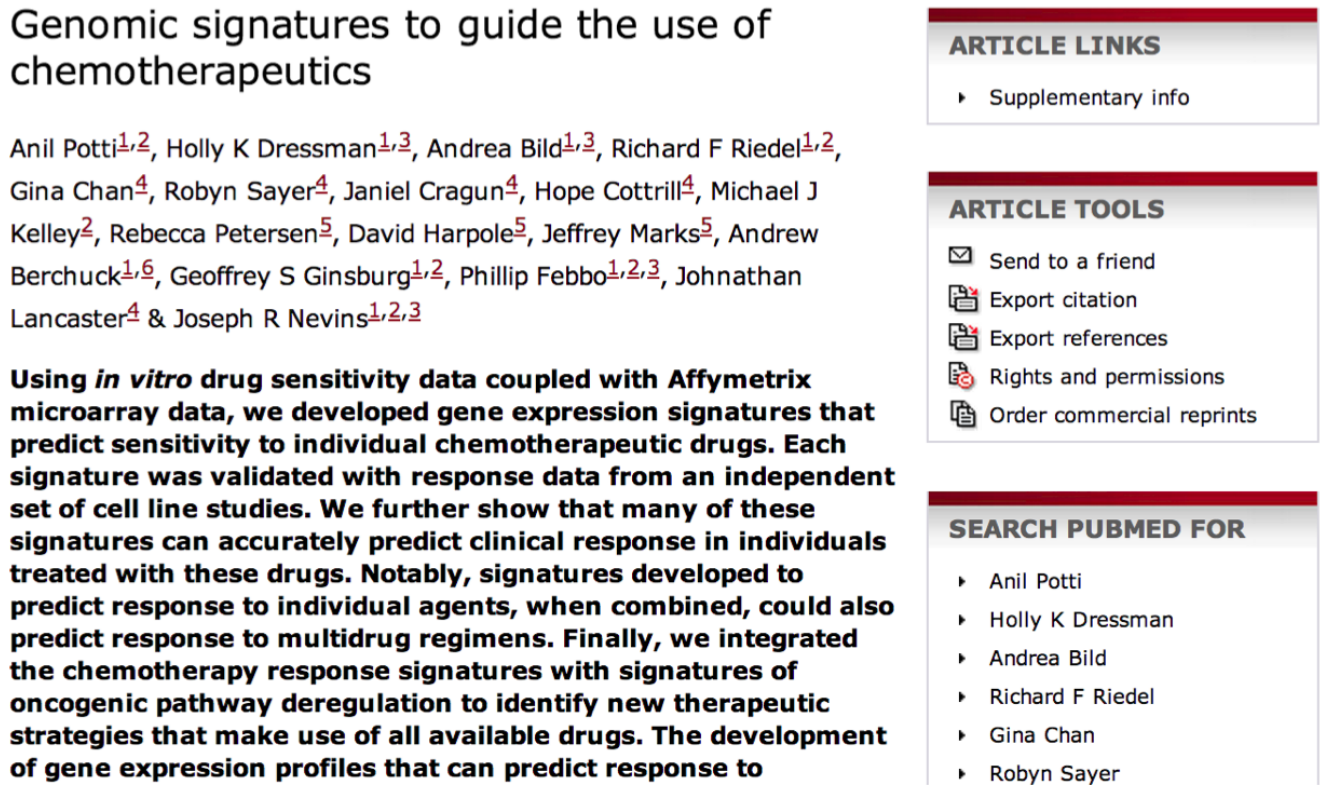
Preliminaries

About this course

- **Class name**: Accessing Biomedical Big Data
- **Instructors**: Jeff Leek, Andrew Jaffe
- **TAs**: TBD
- **Course website**: http://sisbid.github.io/Module1
- **Goal** : Teach you how to get and clean data
- **Pre-reqs**: Some basic knowledge of R
- **Where to get the 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 strategies that make use of all available drugs. The development of gene expression profiles that can predict response to

ARTICLE LINKS

- Supplementary info

ARTICLE TOOLS

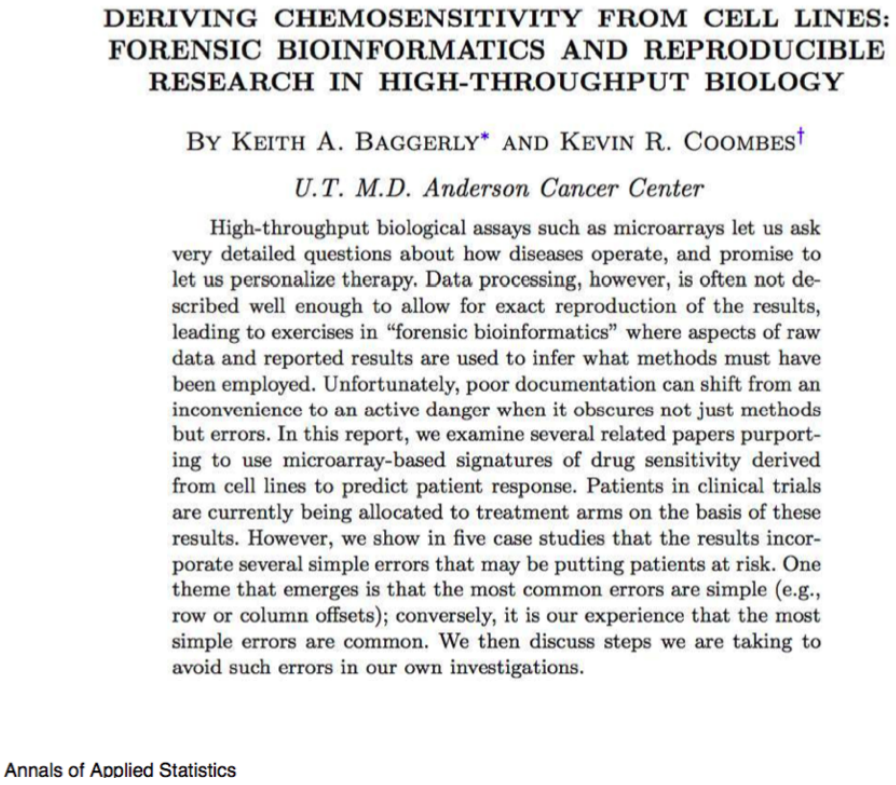
- Send to a friend
- Export citation
- Export references
- Rights and permissions
- Order commercial reprints

SEARCH PUBMED FOR

- Anil Potti
- Holly K Dressman
- Andrea Bild
- Richard F Riedel
- Gina Chan
- Robyn Sayer

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

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.

Annals of Applied Statistics

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

Timeline of events

Box: How events unfolded

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.

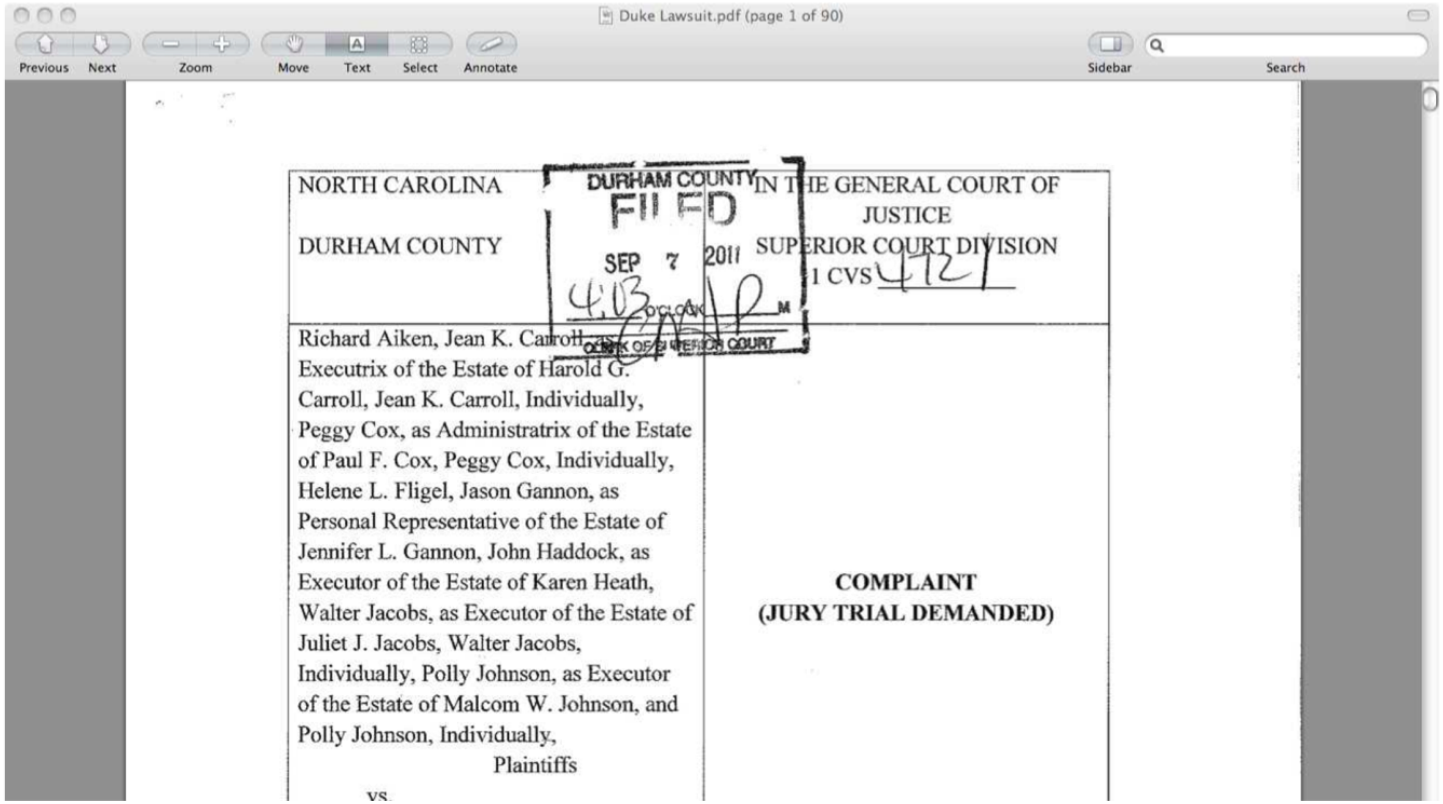
NOVEMBER 2010 *JCO* paper is retracted. Duke closes the trials permanently. Potti resigns.

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.

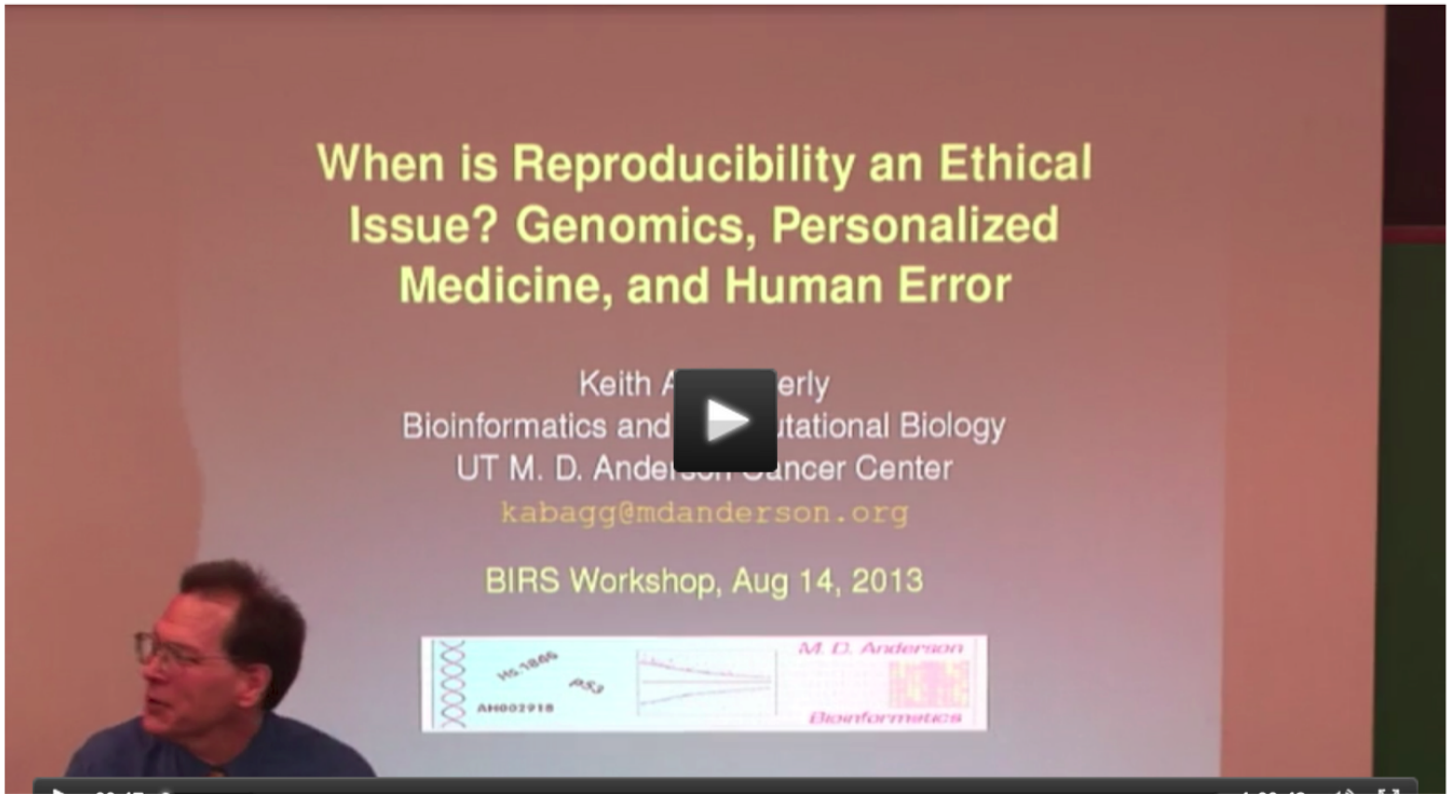
http://www.nature.com/news/2011/110111/full/469139a/box/1.html

Major fallout



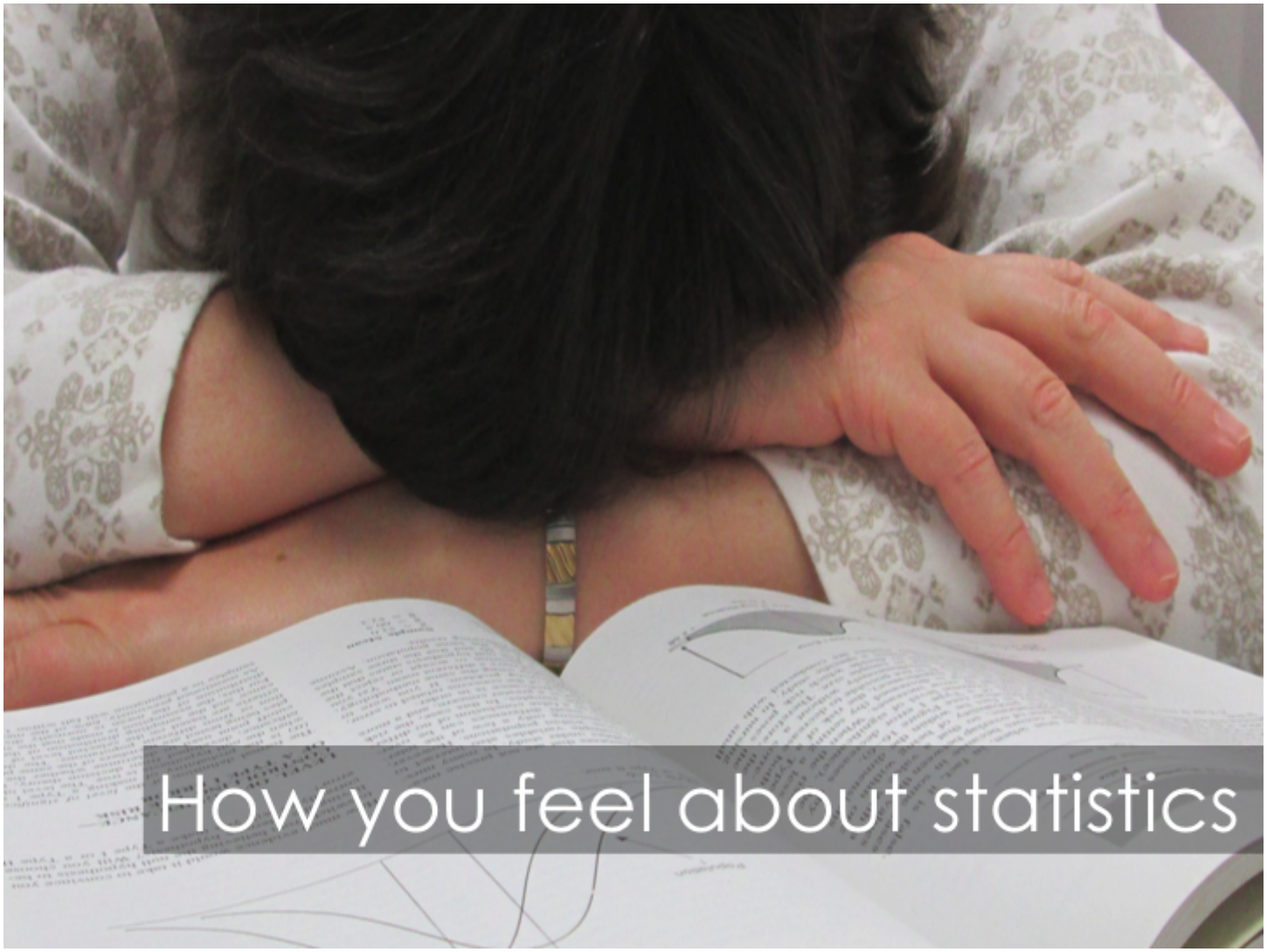
http://dig.abclocal.go.com/wtvd/docs/Duke_lawsuit_090811.pdf
http://www.dukechronicle.com/articles/2015/05/03/duke-lawsuit-involving-cancer-patients-linked-anil-potti-settled

A great talk about this event



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

About me



how you feel

how i feel

@simplystats

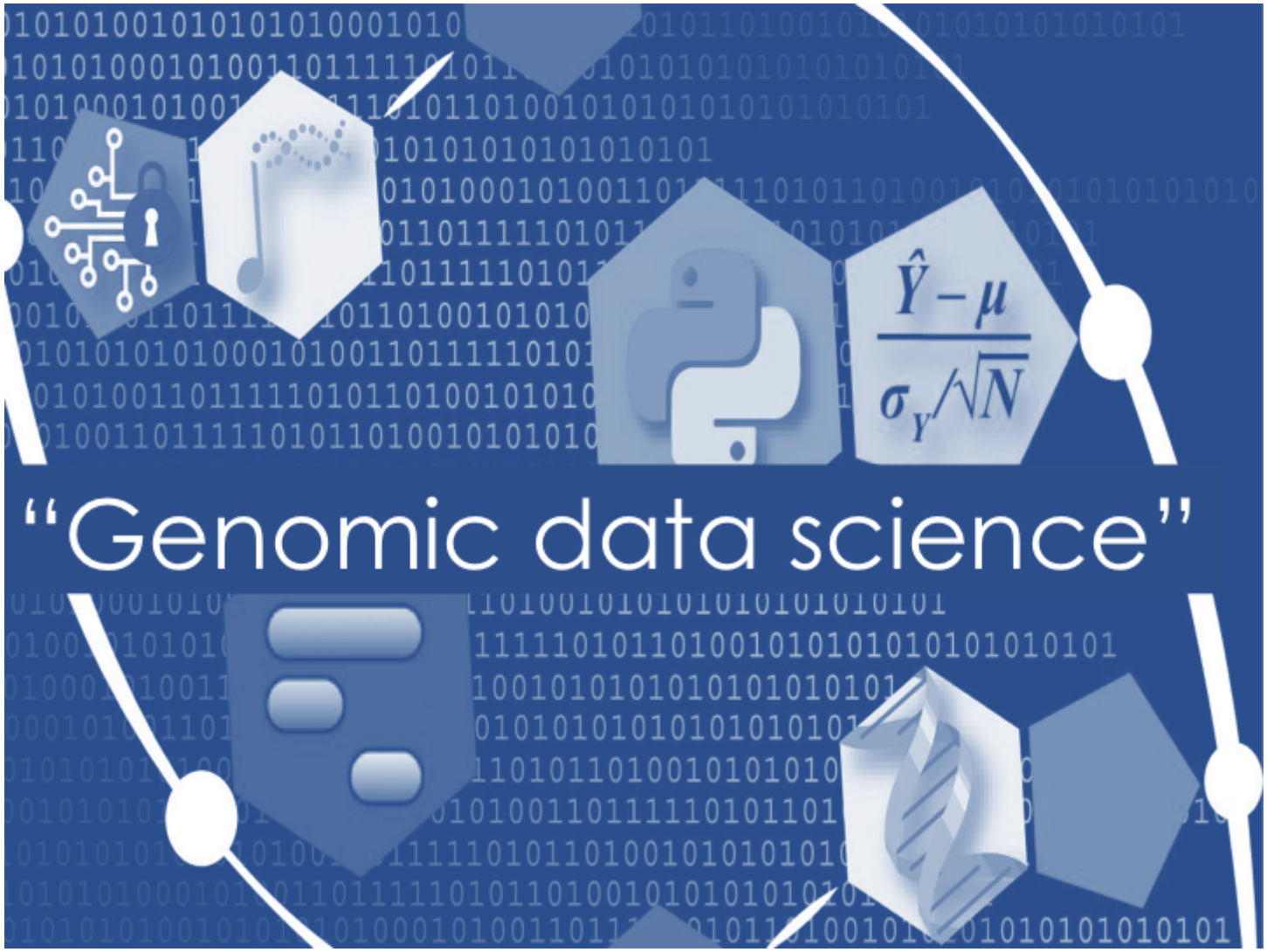
simplystats

http://simplystatistics.org/

jhudatascience.org

dss

https://www.coursera.org/specialization/jhudatascience/1



gdss

https://www.coursera.org/specialization/genomics/41

jtleek

http://www.jtleek.com

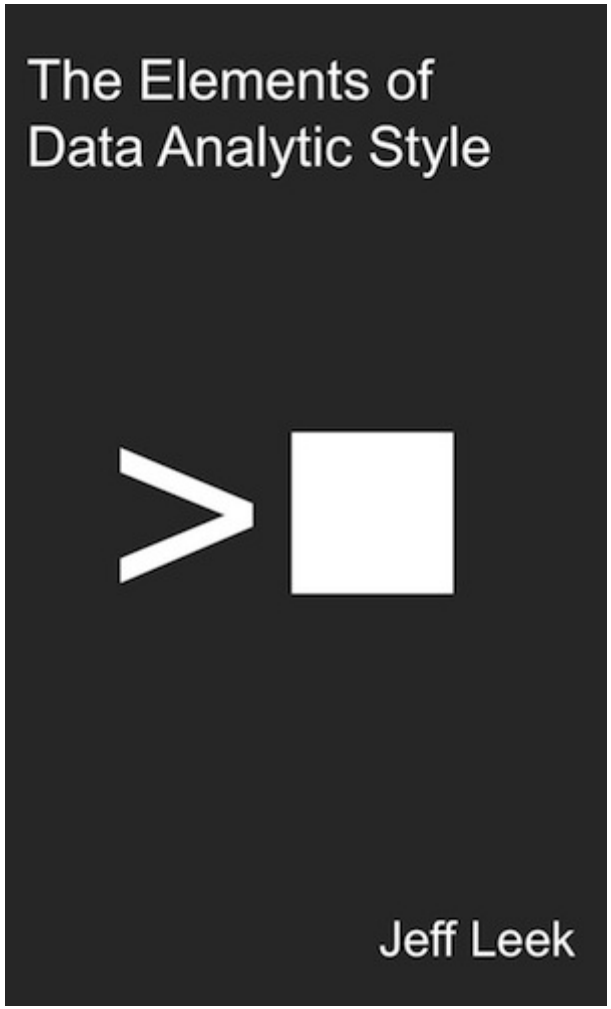
Hi I'm Jeff

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



edas

https://leanpub.com/datastyle

Session Motivation

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

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

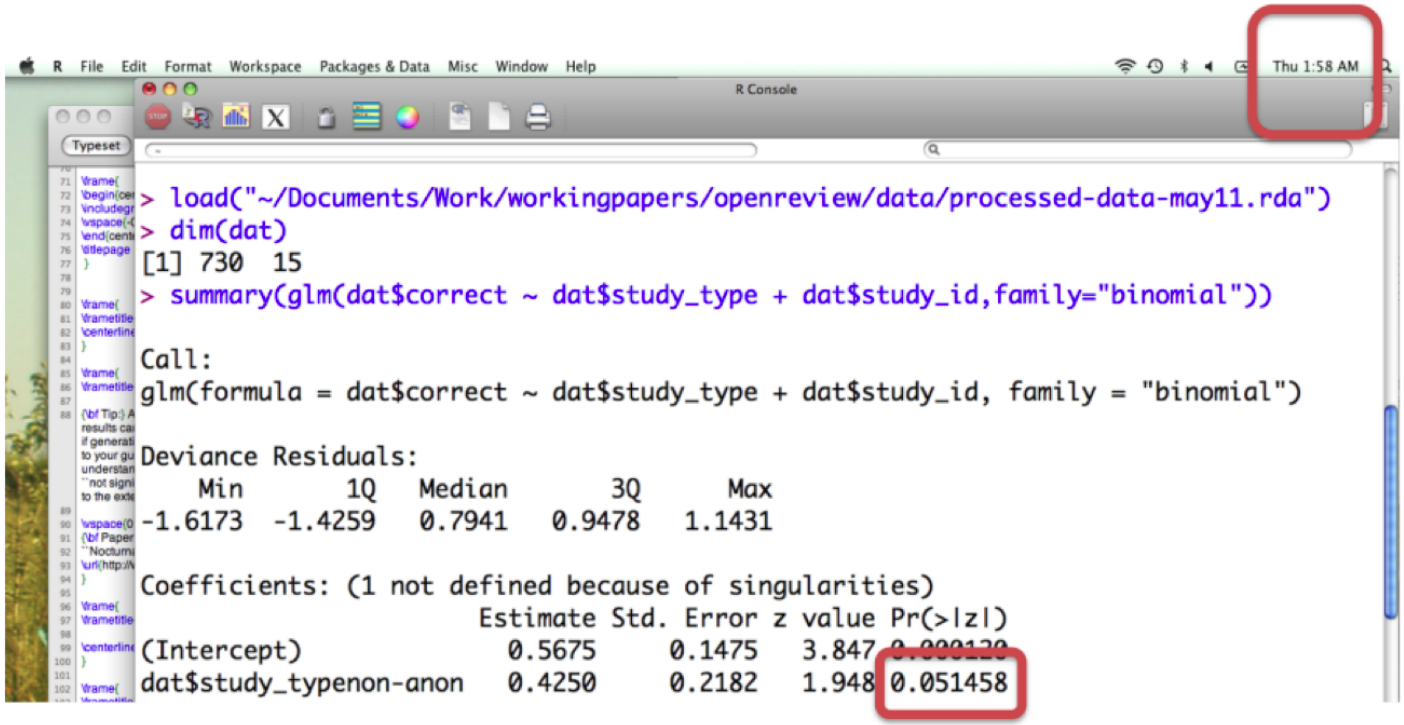
Psychological Science
XX(X) 1–8

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

What is the temptation?



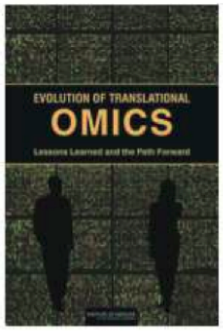
This is now codified for clinical genomics

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

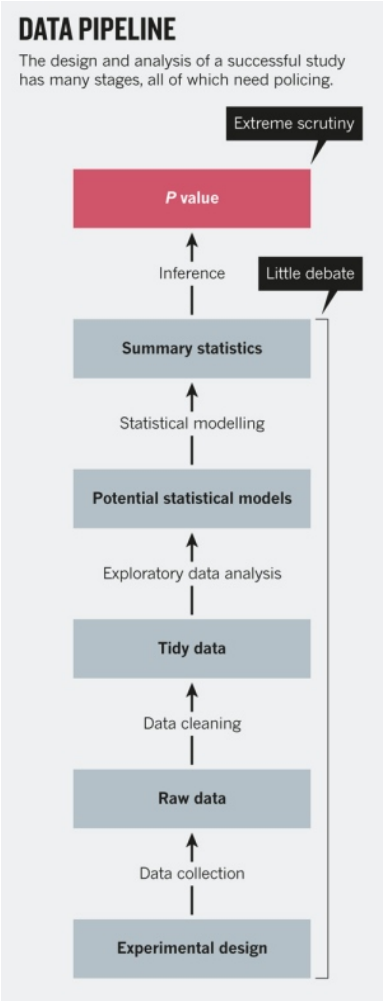


Sequencing the human genome opened a new era in biomedical science. Researchers have begun to untangle the complex roles of biology and genetics in specific diseases, and now better understand why particular therapies do or do not work in individual patients. New technologies have made it feasible to 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 are referred to as *omics*.

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

Getting and cleaning data

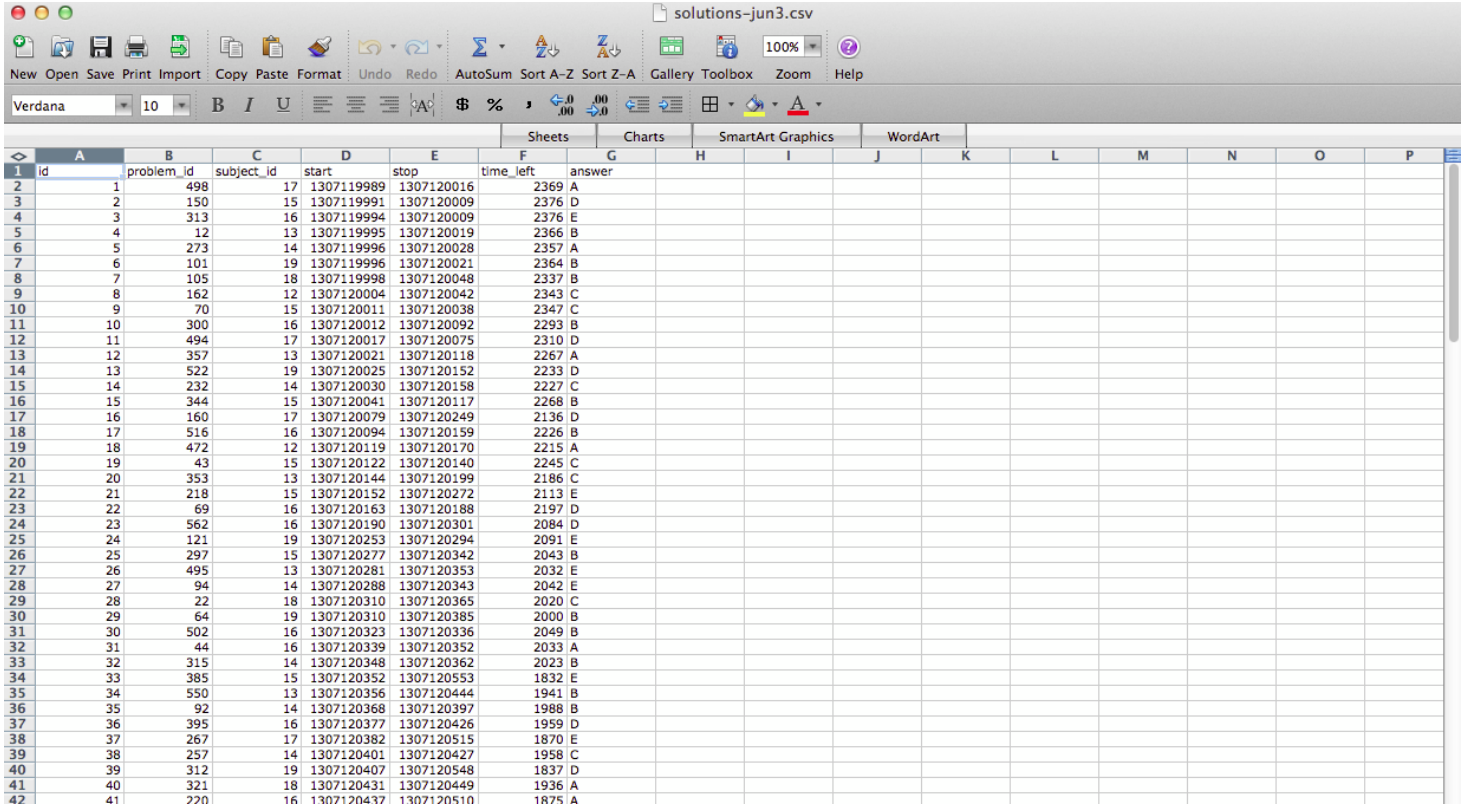
Data pipeline



- Most of the attention is on the last step

Leek and Peng (2015) Nature

What you wish data looked like



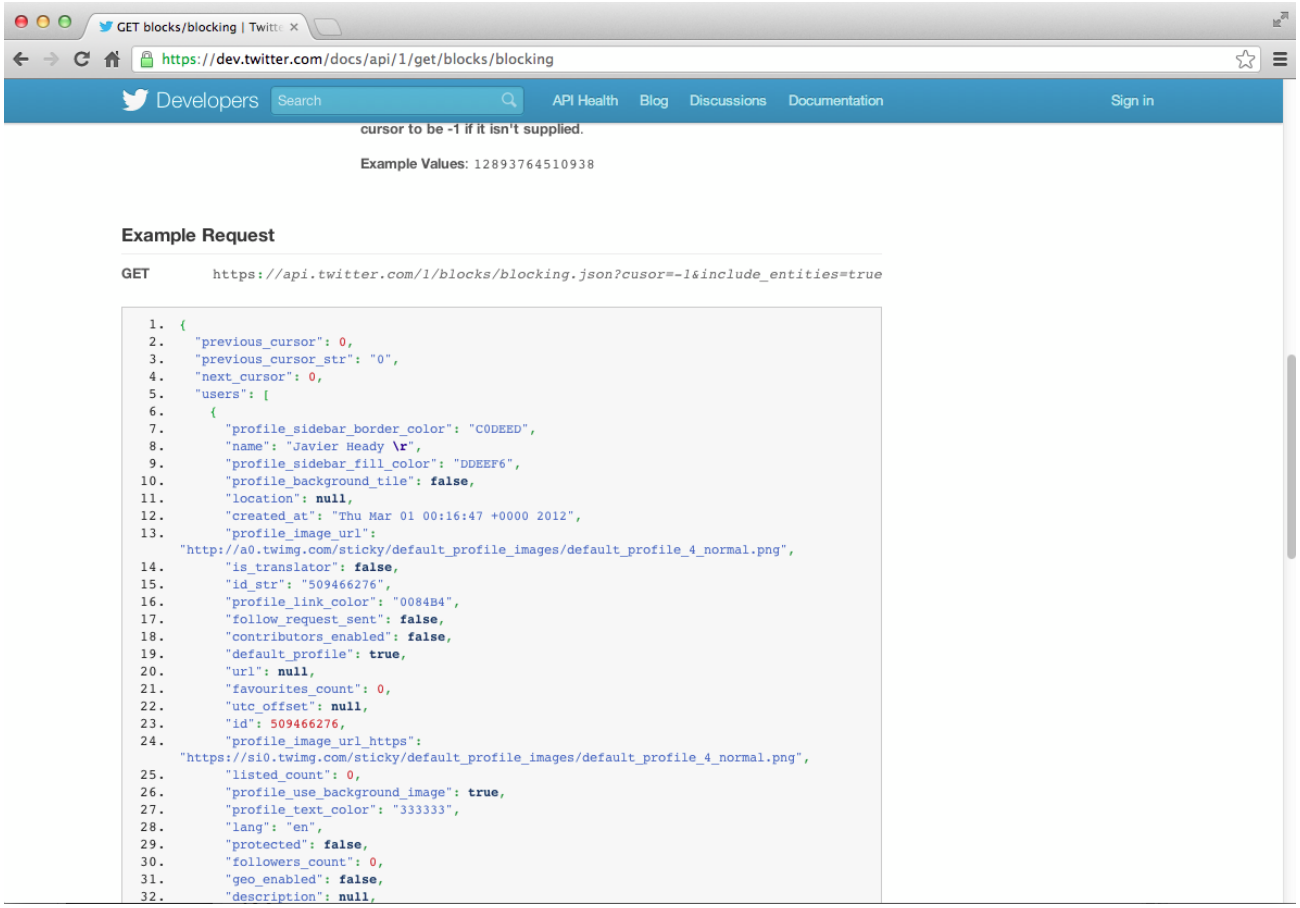
id	problem_id	subject_id	start	stop	time_left	answer
1	498	17	1307119989	1307120016	2369	A
2	150	15	1307119991	1307120009	2376	D
3	313	16	1307119994	1307120009	2376	E
4	12	13	1307119995	1307120019	2366	B
5	273	14	1307119996	1307120028	2357	A
6	101	19	1307119996	1307120021	2364	B
7	105	18	1307119998	1307120048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

What does data really look 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 does data really look like?



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

What does data really look like?

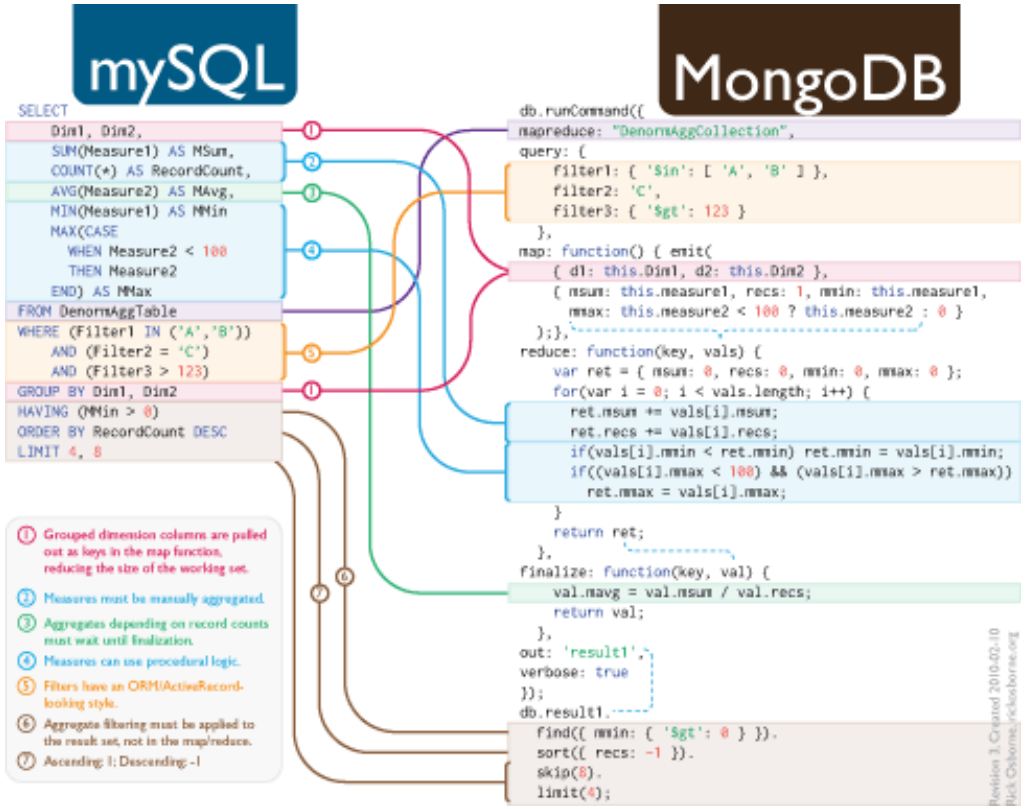
```
-------------------------   ALLERGIES   -------------------------      -------------------------   MEDICATION HISTORY   ---------

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 ONE-HALF TABLI
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 TIMES A DAY W
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
```

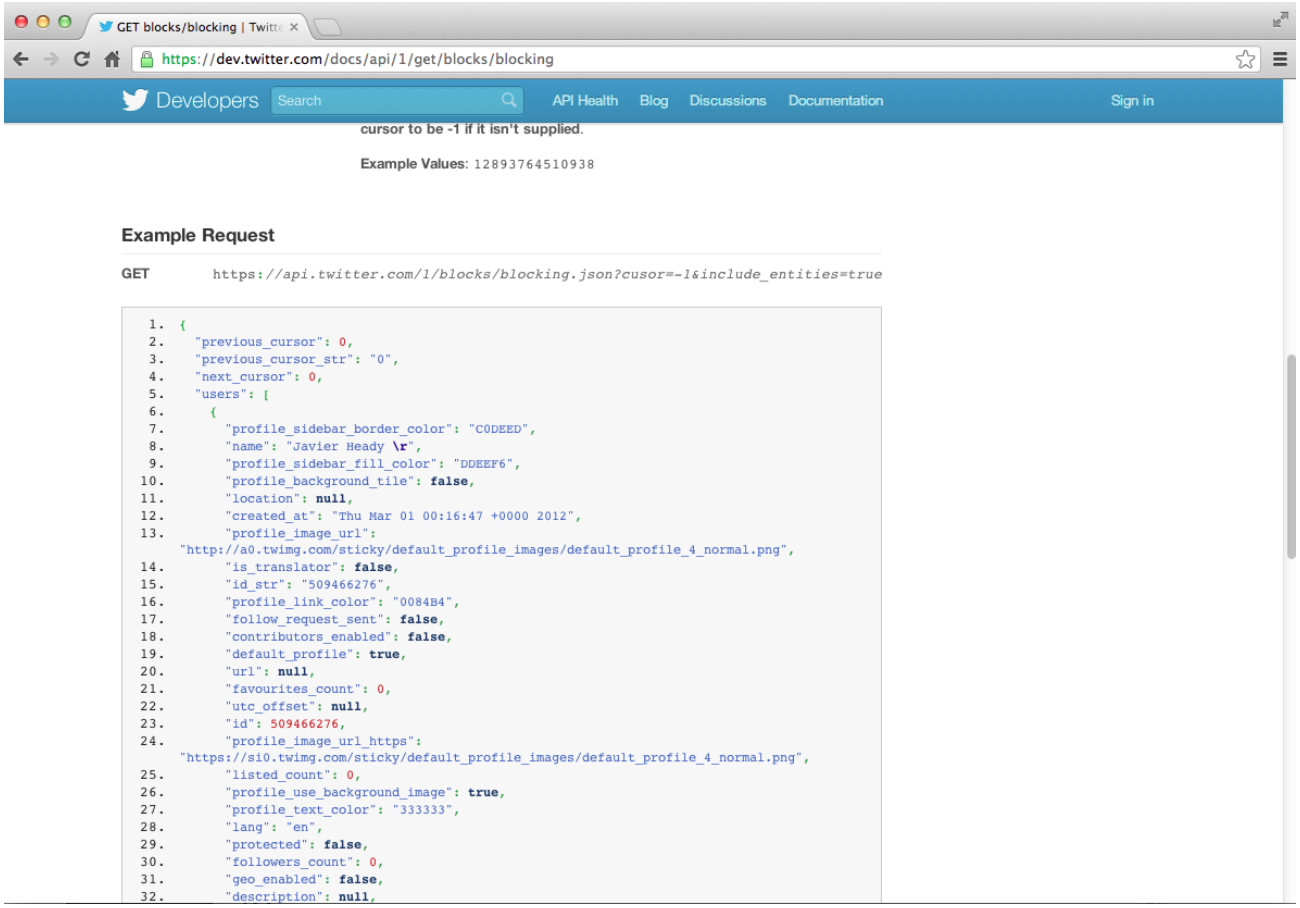
http://blue-button.github.com/challenge/

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/

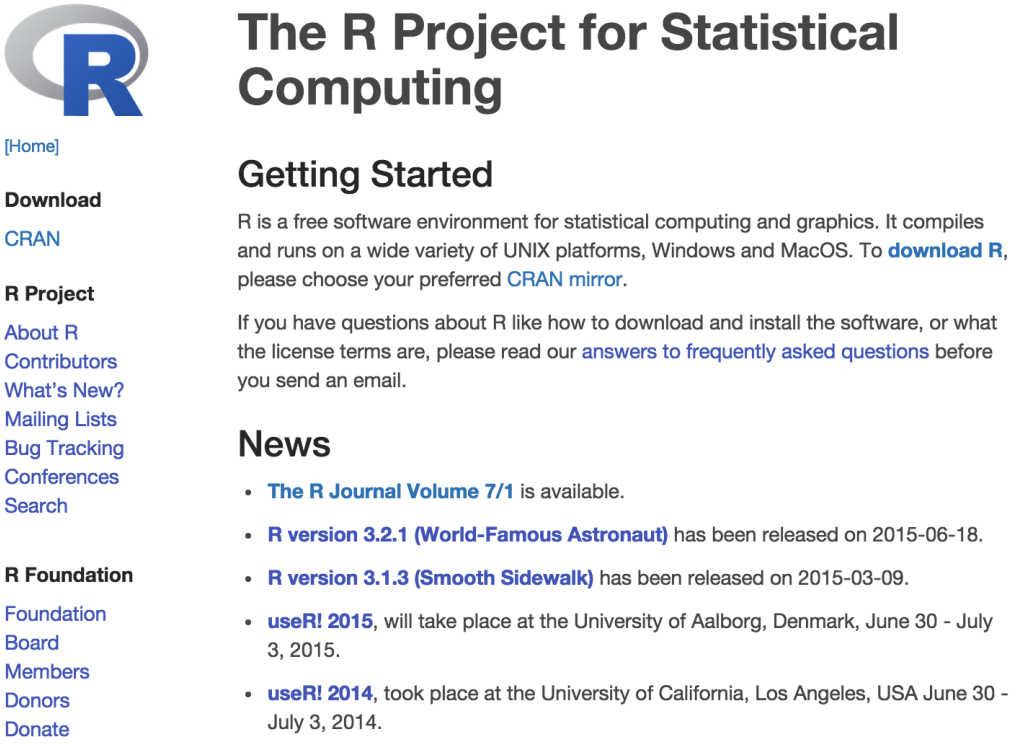
Data brainstorming exercise

The goal of this course

Raw data -> Processing script -> tidy data -> data analysis -> data communication

Tools we will use

The main workhorse of data science



[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.

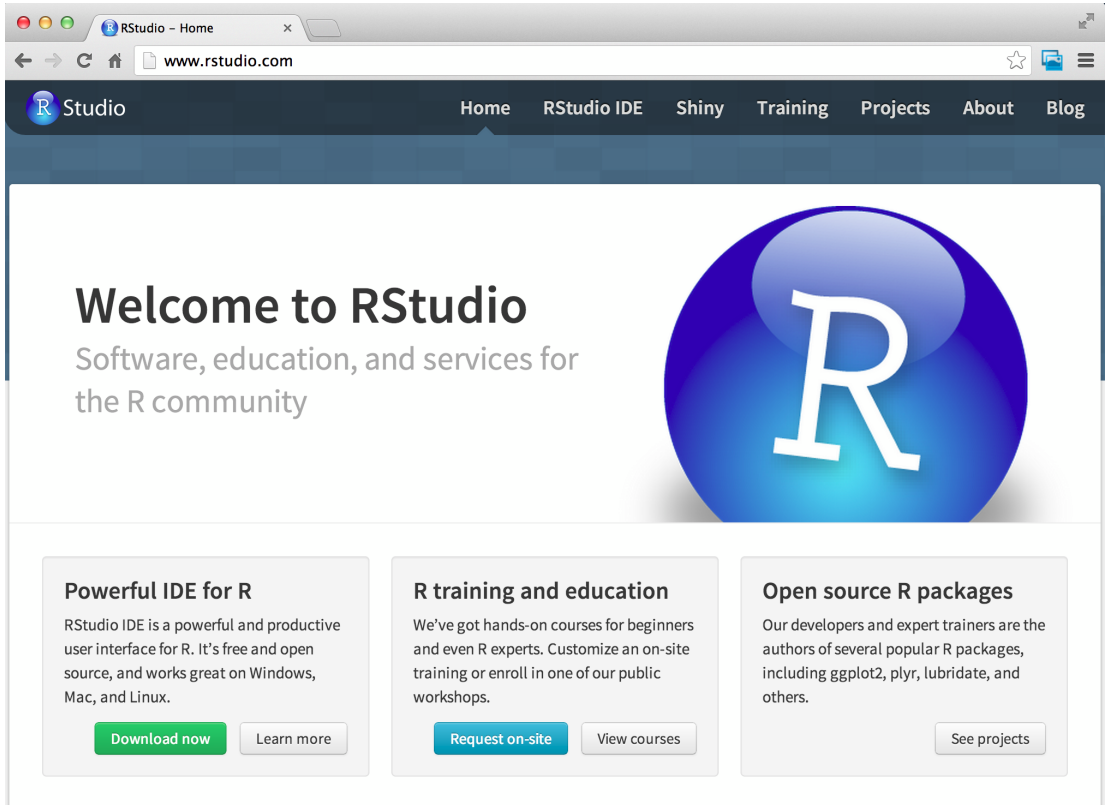
If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

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.

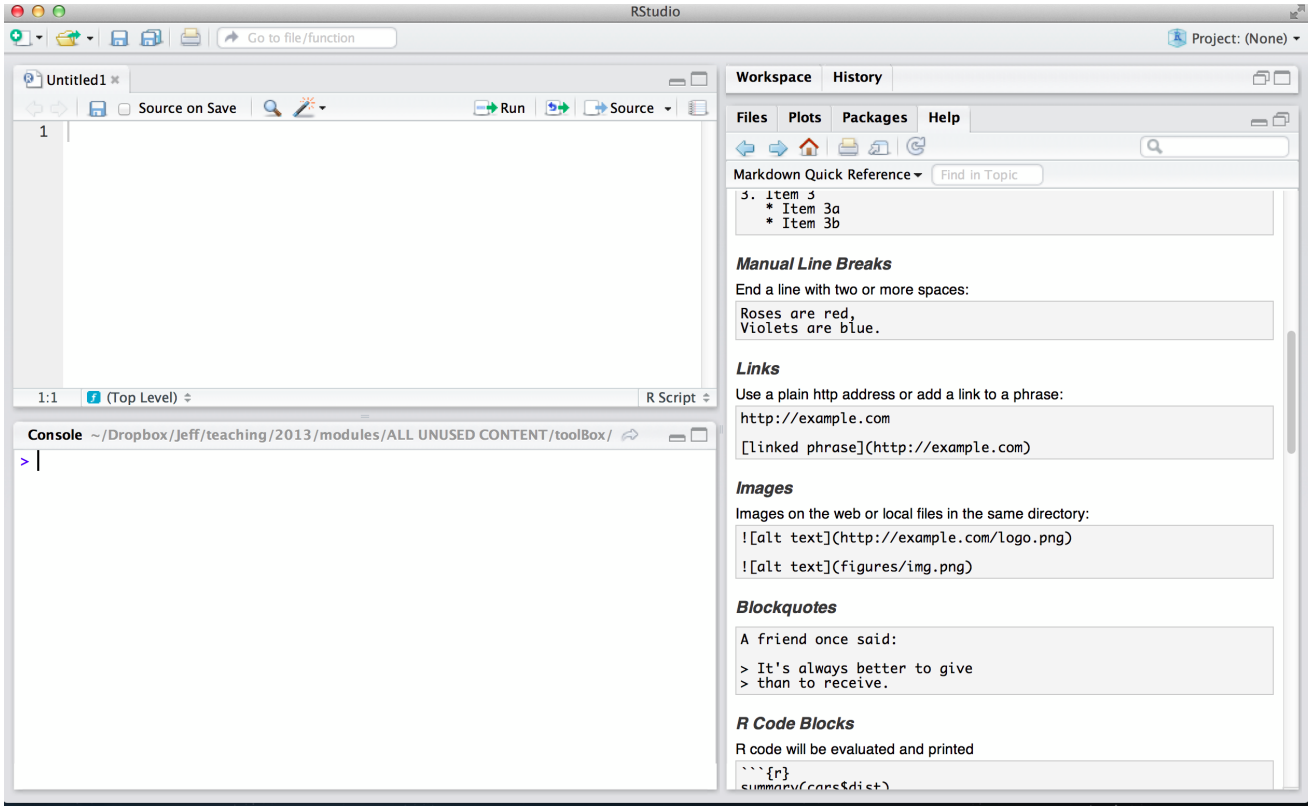
http://www.r-project.org/

Where we will work on coding



http://www.rstudio.com/

Rstudio's interface



http://www.rstudio.com/

Useful Rstudio shortcuts

- `Ctrl + Enter` (`Cmd + Enter` on OS X) in your script evaluates that line of code
    - It's like copying and pasting the code into the console for it to run.
- `Ctrl+1` takes you to the script page
- `Ctrl+2` takes you to the console
- http://www.rstudio.com/ide/docs/using/keyboard_shortcuts

Slide via: http://www.aejaffe.com/summerR_2015/

Rstudio tour/installation check-in

Replicability and reproducibility

```
Science moves forward then discoveries are replicated
and reproduced
```

Implementing Reproducible Research

Slide via: http://www.aejaffe.com/summerR_2015/

Replication

```
Replication, the practice of independently implementing
scientific experiments to validate specific findings,
 is the cornerstone of discovering scientific truth.
```

Implementing Reproducible Research

Slide via: http://www.aejaffe.com/summerR_2015

Reproducibility

```
Reproducibility can be thought of as a different
standard of validity from replication because it
forgoes independent data collection and uses the
methods and data collected by the original investigator.
```

Implementing Reproducible Research

Slide via: http://www.aejaffe.com/summerR_2015

A bit more practical

```
The sharing of analytic data and computer codes
uses to map those data into computational results
is central to any comprehensive definition of
reproducibility.
```

Implementing Reproducible Research

Slide via: http://www.aejaffe.com/summerR_2015

Why its important?

```
Except for the simplest of analyses, the computer
code used to analyze a dataset is the only record
that permits others to fully understand what a
researcher has done.
```

Implementing Reproducible Research

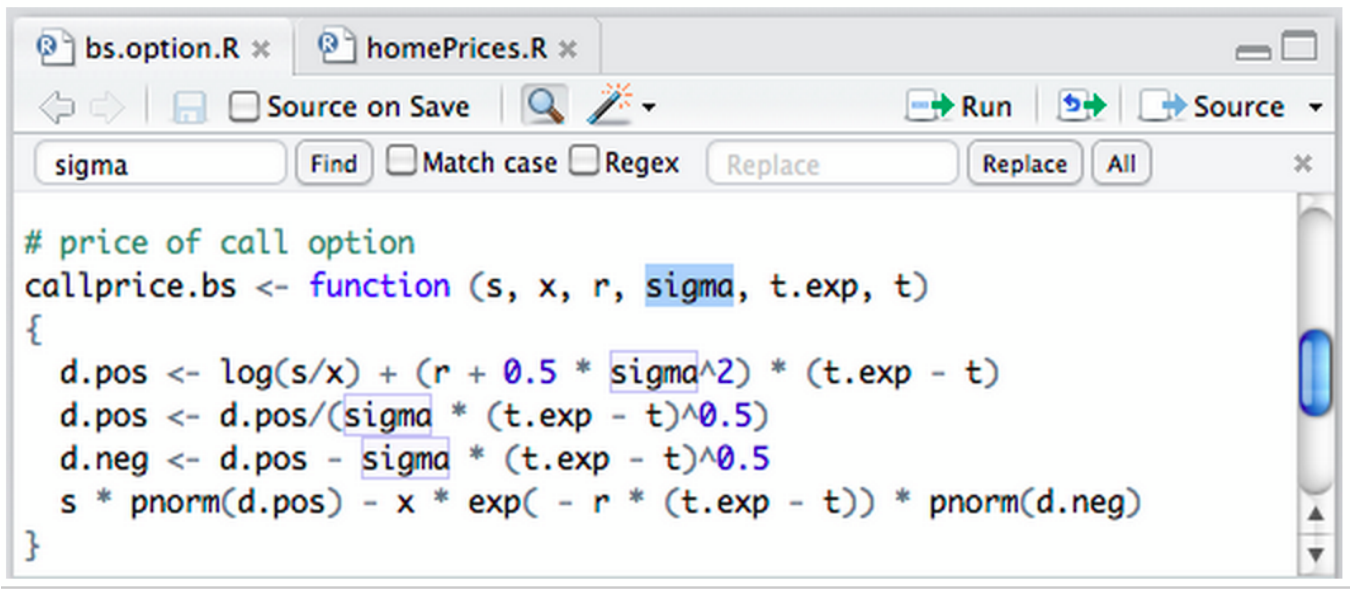
Slide via: http://www.aejaffe.com/summerR_2015

Reproducible documents

- Have you ever had your code in one file, your description of the results in another file?
- Ever made copy-paste mistakes?
- What if you were asked to change some models or revise the document?
- Was it easy to maintain?

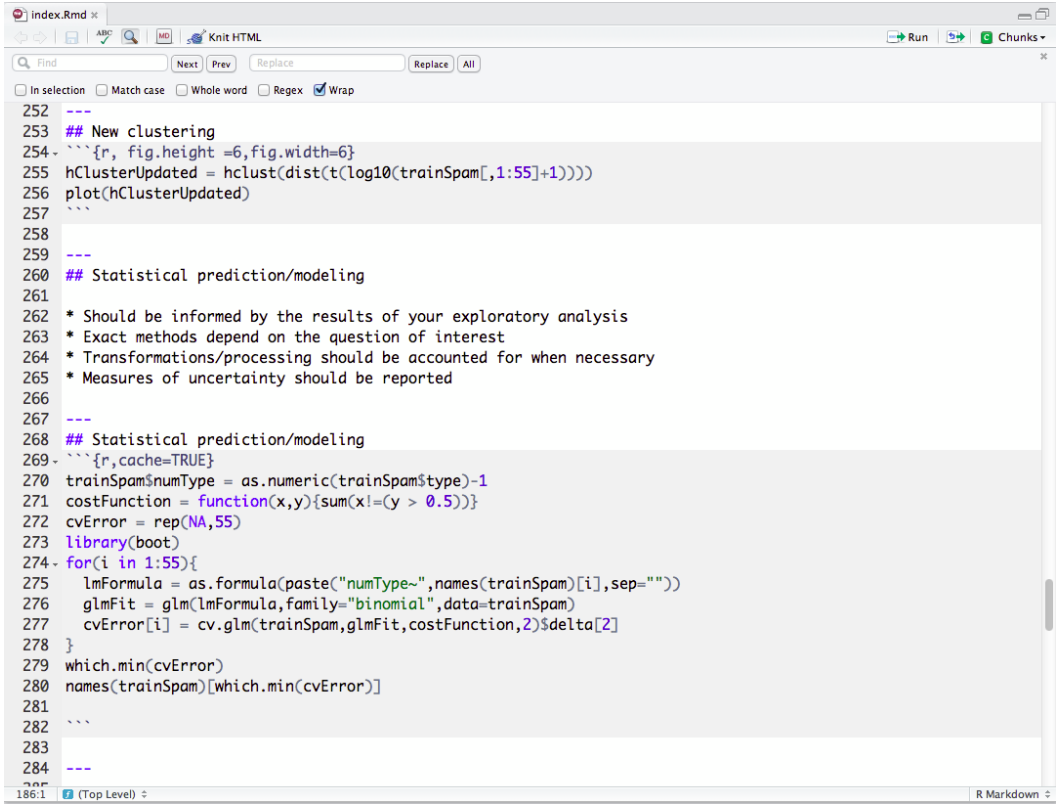
Slide via: http://www.aejaffe.com/summerR_2015

Primary file types - R script



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

Primary file types - R markdown document



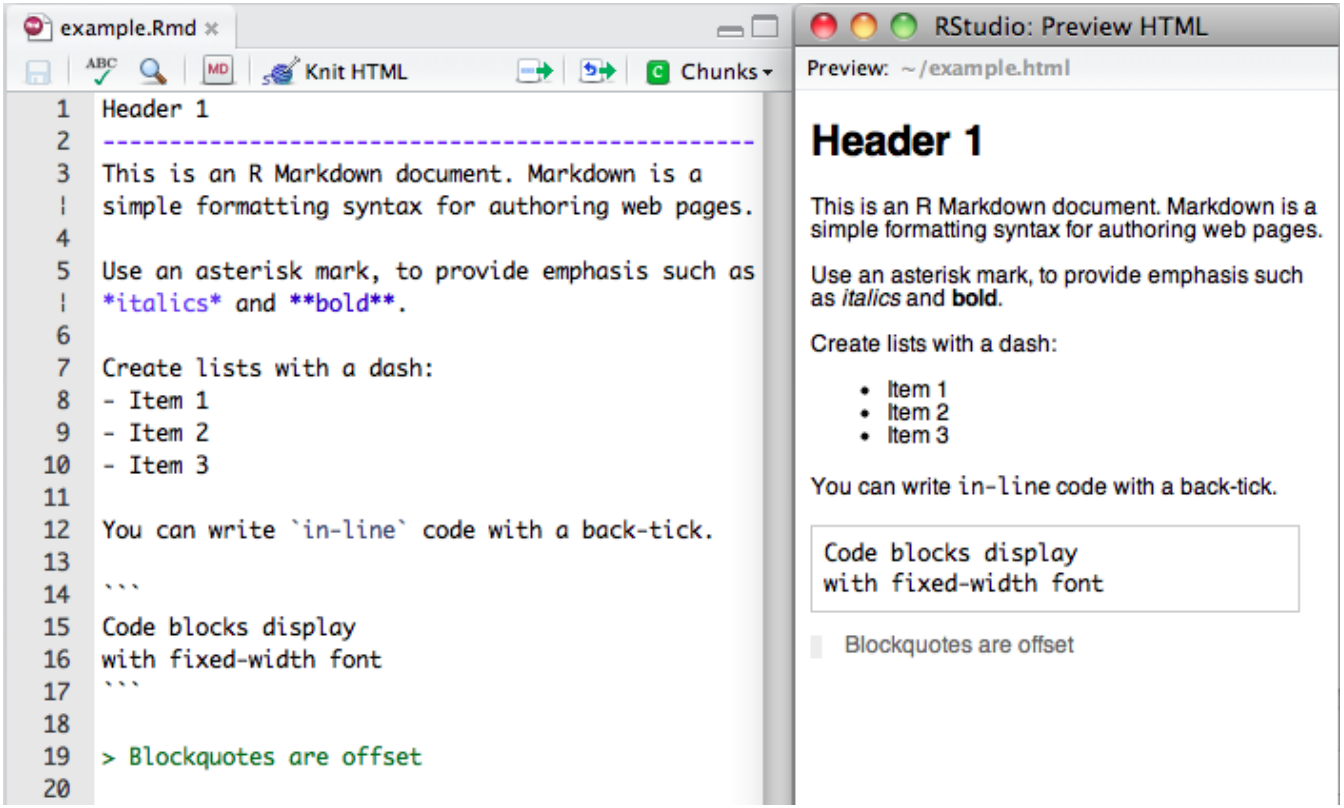
http://www.rstudio.com/ide/docs/authoring/using_markdown

Markdown basics

"Markdown is a text-to-HTML conversion tool for web writers. Markdown allows you to write using an easy-to-read, easy-to-write plain text format, then convert it to structurally valid XHTML (or HTML)."

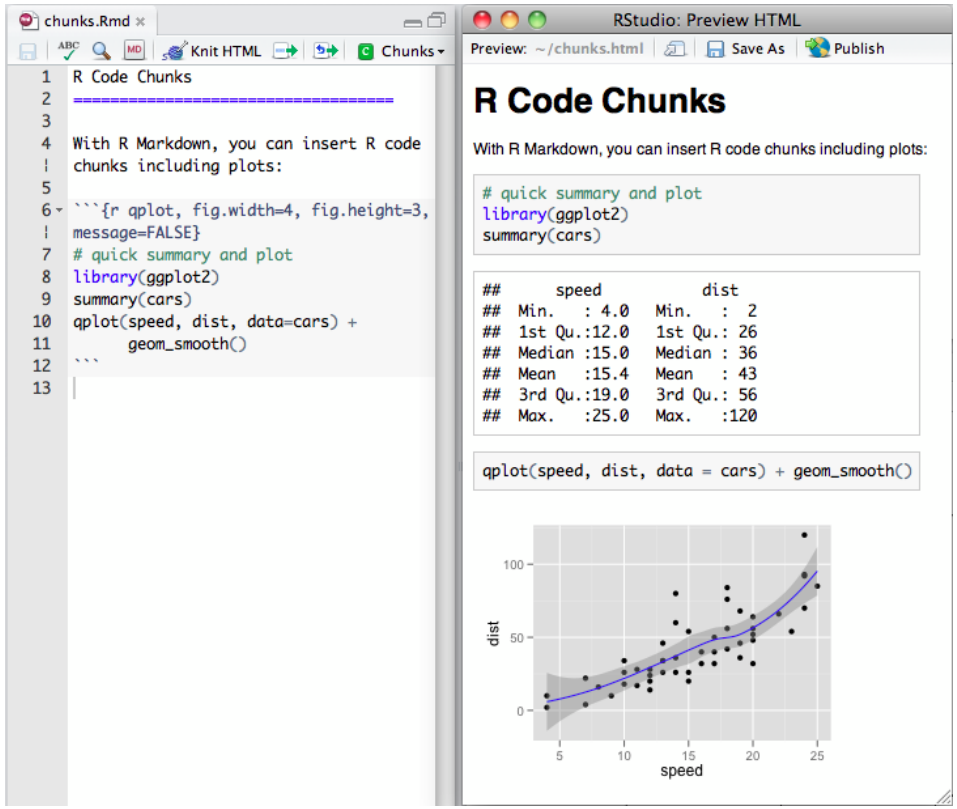
John Gruber, creator of Markdown

R markdown basics



http://rmarkdown.rstudio.com/

R code



http://rmarkdown.rstudio.com/

Equations with latex

In-line

This is an equation `$y = \beta_0 x$` in a line of text

This is an equation $y = \beta_0 x$ in a line of text

Display

This is display equation `$$y = \beta_0 x$$`

This is an equation

$$y = \beta_0 x$$

R markdown lab

Definition of data

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

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

Definition of data

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

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

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

Definition of data

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

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

Variables: A measurement or characteristic of an item.

Definition of data

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

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

Qualitative: Country of origin, sex, treatment

Quantitative: Height, weight, blood pressure

Raw data

- The original source of the data
- Often hard to use for data analyses
- Data analysis *includes* processing
- Raw data may only need to be processed once

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

Processed data

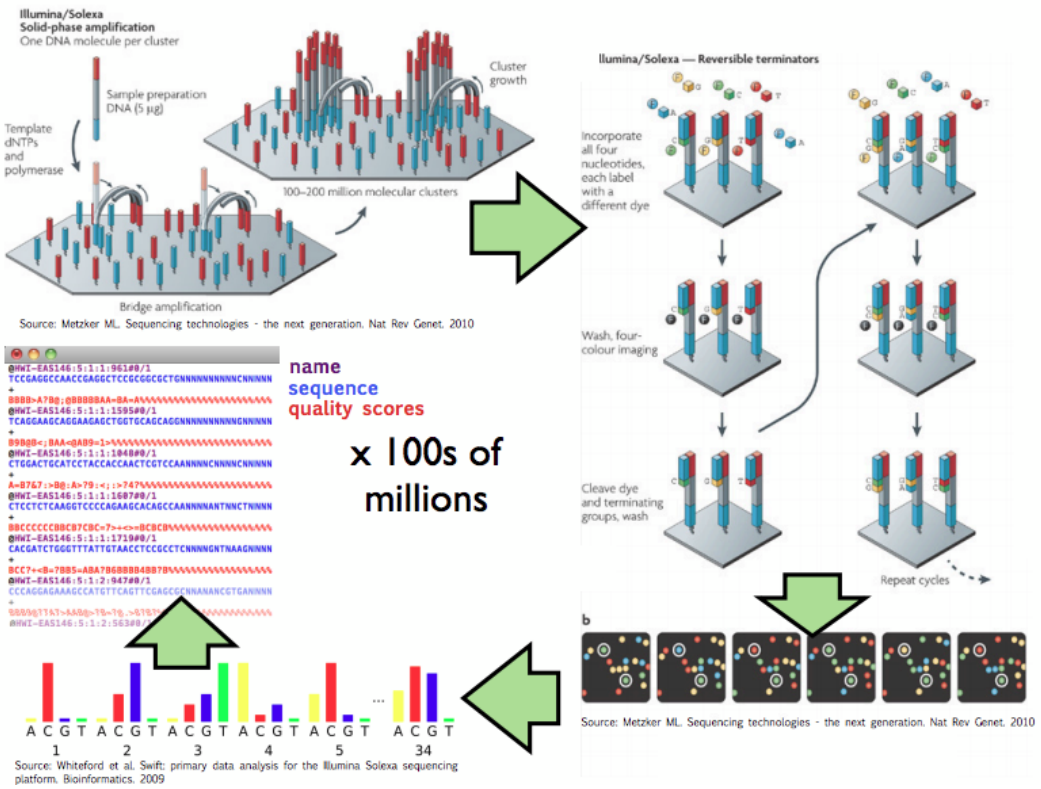
- Data that is ready for analysis
- Processing can include merging, subsetting, transforming, etc.
- There may be standards for processing
- All steps should be recorded

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

An example of a processing pipeline

http://www.illumina.com.cn/support/sequencing/sequencing_instruments/h

An example of a processing pipeline



http://www.cbcb.umd.edu/~hcorrada/CMSC858B/lectures/lect22_seqIntro/s

A common goal

A common goal

The four things you should have

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.

The raw data

- The strange binary file your measurement machine spits out
- The unformatted Excel file with 10 worksheets the company you contracted with sent you
- The complicated JSON data you got from scraping the Twitter API
- The hand-entered numbers you collected looking through a microscope

You know the raw data is in the right format if you_

1. Ran no software on the data
2. Did not manipulate any of the numbers in the data
3. You did not remove any data from the data set
4. You did not summarize the data in any way

https://github.com/jtleek/datasharing

The tidy data

1. Each variable you measure should be in one column
2. Each different observation of that variable should be in a different row
3. There should be one table for each "kind" of variable
4. If you have multiple tables, they should include a column in the table that allows them to be linked

Some other important tips

- Include a row at the top of each file with variable names.
- Make variable names human readable AgeAtDiagnosis instead of AgeDx
- In general data should be saved in one file per table.

https://github.com/jtleek/datasharing

The code book

1. Information about the variables (including units!) in the data set not contained in the tidy data
2. Information about the summary choices you made
3. Information about the experimental study design you used

Some other important tips_

- A common format for this document is a Word/text file.
- There should be a section called "Study design" that has a thorough description of how you collected the data.
- There must be a section called "Code book" that describes each variable and its units.

https://github.com/jtleek/datasharing

The instruction list

- Ideally a computer script (in R :-), but I suppose Python is ok too…)
- The input for the script is the raw data
- The output is the processed, tidy data
- There are no parameters to the script

When you can't script

In some cases it will not be possible to script every step. In that case you should provide instructions like:

1. Step 1 - take the raw file, run version 3.1.2 of summarize software with parameters a=1, b=2, c=3
2. Step 2 - run the software separately for each sample
3. Step 3 - take column three of outputfile.out for each sample and that is the corresponding row in the output data set

https://github.com/jtleek/datasharing

Why is the instruction list important?

Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff

Thomas Herndon* Michael Ash Robert Pollin

April 15, 2013

http://www.colbertnation.com/the-colbert-report-videos/425748/april-23-2013/austerity-s-spreadsheet-error

dplyr

The data frame is a key data structure in statistics and in R.

- There is one observation per row
- Each column represents a variable or measure or characteristic
- Primary implementation that you will use is the default R implementation
- Other implementations, particularly relational databases systems

dplyr

- Developed by Hadley Wickham of RStudio
- An optimized and distilled version of `plyr` package (also by Hadley)
- Does not provide any "new" functionality per se, but **greatly** simplifies existing functionality in R
- Provides a "grammar" (in particular, verbs) for data manipulation
- Is **very** fast, as many key operations are coded in C++

dplyr Verbs

- `select`: return a subset of the columns of a data frame
- `filter`: extract a subset of rows from a data frame based on logical conditions
- `arrange`: reorder rows of a data frame
- `rename`: rename variables in a data frame
- `mutate`: add new variables/columns or transform existing variables
- `summarise` / `summarize`: generate summary statistics of different variables in the data frame, possibly within strata

There is also a handy `print` method that prevents you from printing a lot of data to the console.

dplyr Properties

- The first argument is a data frame.
- The subsequent arguments describe what to do with it, and you can refer to columns in the data frame directly without using the $ operator (just use the names).
- The result is a new data frame
- Data frames must be properly formatted and annotated for this to all be useful

Load the `dplyr` package

This step is important!

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

select

```
chicago <- readRDS("./data/chicago.rds")
dim(chicago)
```

```
## [1] 6940    8
```

```
head(select(chicago, 1:5))
```

```
##   city tmpd   dptp       date pm25tmean2
## 1 chic 31.5 31.500 1987-01-01         NA
## 2 chic 33.0 29.875 1987-01-02         NA
## 3 chic 33.0 27.375 1987-01-03         NA
## 4 chic 29.0 28.625 1987-01-04         NA
## 5 chic 32.0 28.875 1987-01-05         NA
## 6 chic 40.0 35.125 1987-01-06         NA
```

select

```
names(chicago)[1:3]

## [1] "city" "tmpd" "dptp"

head(select(chicago, city:dptp))

##   city tmpd   dptp
## 1 chic 31.5 31.500
## 2 chic 33.0 29.875
## 3 chic 33.0 27.375
## 4 chic 29.0 28.625
## 5 chic 32.0 28.875
## 6 chic 40.0 35.125
```

select

In dplyr you can do

```
head(select(chicago, -(city:dptp)))
```

Equivalent base R

```
i <- match("city", names(chicago))
j <- match("dptp", names(chicago))
head(chicago[, -(i:j)])
```

filter

```
chic.f <- filter(chicago, pm25tmean2 > 30)
head(select(chic.f, 1:3, pm25tmean2), 10)
```

```
##    city tmpd dptp pm25tmean2
## 1  chic   23 21.9      38.10
## 2  chic   28 25.8      33.95
## 3  chic   55 51.3      39.40
## 4  chic   59 53.7      35.40
## 5  chic   57 52.0      33.30
## 6  chic   57 56.0      32.10
## 7  chic   75 65.8      56.50
## 8  chic   61 59.0      33.80
## 9  chic   73 60.3      30.30
## 10 chic   78 67.1      41.40
```

filter

```
chic.f <- filter(chicago, pm25tmean2 > 30 & tmpd > 80)
head(select(chic.f, 1:3, pm25tmean2, tmpd), 10)
```

```
##    city tmpd dptp pm25tmean2
## 1  chic   81 71.2    39.6000
## 2  chic   81 70.4    31.5000
## 3  chic   82 72.2    32.3000
## 4  chic   84 72.9    43.7000
## 5  chic   85 72.6    38.8375
## 6  chic   84 72.6    38.2000
## 7  chic   82 67.4    33.0000
## 8  chic   82 63.5    42.5000
## 9  chic   81 70.4    33.1000
## 10 chic   82 66.2    38.8500
```

arrange

Reordering rows of a data frame (while preserving corresponding order of other columns) is normally a pain to do in R.

```
chicago <- arrange(chicago, date)
head(select(chicago, date, pm25tmean2), 3)
```

```
##         date pm25tmean2
## 1 1987-01-01         NA
## 2 1987-01-02         NA
## 3 1987-01-03         NA
```

```
tail(select(chicago, date, pm25tmean2), 3)
```

```
##            date pm25tmean2
## 6938 2005-12-29    7.45000
## 6939 2005-12-30   15.05714
## 6940 2005-12-31   15.00000
```

arrange

Columns can be arranged in descending order too.

```
chicago <- arrange(chicago, desc(date))
head(select(chicago, date, pm25tmean2), 3)
```

```
##         date pm25tmean2
## 1 2005-12-31   15.00000
## 2 2005-12-30   15.05714
## 3 2005-12-29    7.45000
```

```
tail(select(chicago, date, pm25tmean2), 3)
```

```
##            date pm25tmean2
## 6938 1987-01-03         NA
## 6939 1987-01-02         NA
## 6940 1987-01-01         NA
```

rename

Renaming a variable in a data frame in R is surprising hard to do!

```
head(chicago[, 1:5], 3)
```

```
##   city tmpd dptp       date pm25tmean2
## 1 chic   35 30.1 2005-12-31   15.00000
## 2 chic   36 31.0 2005-12-30   15.05714
## 3 chic   35 29.4 2005-12-29    7.45000
```

```
chicago <- rename(chicago, dewpoint = dptp,
                  pm25 = pm25tmean2)
head(chicago[, 1:5], 3)
```

```
##   city tmpd dewpoint       date     pm25
## 1 chic   35     30.1 2005-12-31 15.00000
## 2 chic   36     31.0 2005-12-30 15.05714
## 3 chic   35     29.4 2005-12-29  7.45000
```

mutate

```
chicago <- mutate(chicago,
                  pm25detrend=pm25-mean(pm25, na.rm=TRUE))
head(select(chicago, pm25, pm25detrend))
```

```
##       pm25 pm25detrend
## 1 15.00000   -1.230958
## 2 15.05714   -1.173815
## 3  7.45000   -8.780958
## 4 17.75000    1.519042
## 5 23.56000    7.329042
## 6  8.40000   -7.830958
```

group_by

Generating summary statistics by stratum

```
chicago <- mutate(chicago,
                  tempcat = factor(1 * (tmpd > 80),
                                   labels = c("cold", "hot")))
hotcold <- group_by(chicago, tempcat)
summarize(hotcold, pm25 = mean(pm25, na.rm = TRUE),
          o3 = max(o3tmean2),
          no2 = median(no2tmean2))
```

```
## Source: local data frame [3 x 4]
##
##   tempcat     pm25        o3      no2
##    (fctr)    (dbl)     (dbl)    (dbl)
## 1    cold 15.97807 66.587500 24.54924
## 2     hot 26.48118 62.969656 24.93870
## 3      NA 47.73750  9.416667 37.44444
```

group_by

Generating summary statistics by stratum

```
chicago <- mutate(chicago,
                  year = as.POSIXlt(date)$year + 1900)
years <- group_by(chicago, year)
summarize(years, pm25 = mean(pm25, na.rm = TRUE),
          o3 = max(o3tmean2, na.rm = TRUE),
          no2 = median(no2tmean2, na.rm = TRUE))
```

```
## Source: local data frame [19 x 4]
##
##     year     pm25       o3      no2
##    (dbl)    (dbl)    (dbl)    (dbl)
## 1   1987      NaN 62.96966 23.49369
## 2   1988      NaN 61.67708 24.52296
## 3   1989      NaN 59.72727 26.14062
## 4   1990      NaN 52.22917 22.59583
## 5   1991      NaN 63.10417 21.38194
## 6   1992      NaN 50.82870 24.78921
## 7   1993      NaN 44.30093 25.76993
## 8   1994      NaN 52.17844 28.47500
## 9   1995      NaN 66.58750 27.26042
## 10  1996      NaN 58.39583 26.38715
```

%>%

```
chicago %>% mutate(month = as.POSIXlt(date)$mon + 1)
    %>% group_by(month)
    %>% summarize(pm25 = mean(pm25, na.rm = TRUE),
          o3 = max(o3tmean2, na.rm = TRUE),
          no2 = median(no2tmean2, na.rm = TRUE))

## Source: local data frame [12 x 4]
##
##    month     pm25       o3      no2
##    (dbl)    (dbl)    (dbl)    (dbl)
## 1      1 17.76996 28.22222 25.35417
## 2      2 20.37513 37.37500 26.78034
## 3      3 17.40818 39.05000 26.76984
## 4      4 13.85879 47.94907 25.03125
## 5      5 14.07420 52.75000 24.22222
## 6      6 15.86461 66.58750 25.01140
## 7      7 16.57087 59.54167 22.38442
## 8      8 16.93380 53.96701 22.98333
## 9      9 15.91279 57.48864 24.47917
## 10    10 14.23557 47.09275 24.15217
## 11    11 15.15794 29.45833 23.56537
## 12    12 17.52221 27.70833 24.45773
```

dplyr

Once you learn the dplyr "grammar" there are a few additional benefits

- dplyr can work with other data frame "backends"
- `data.table` for large fast tables
- SQL interface for relational databases via the DBI package

Tidy data lab

Three types of genomic data

Home again

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