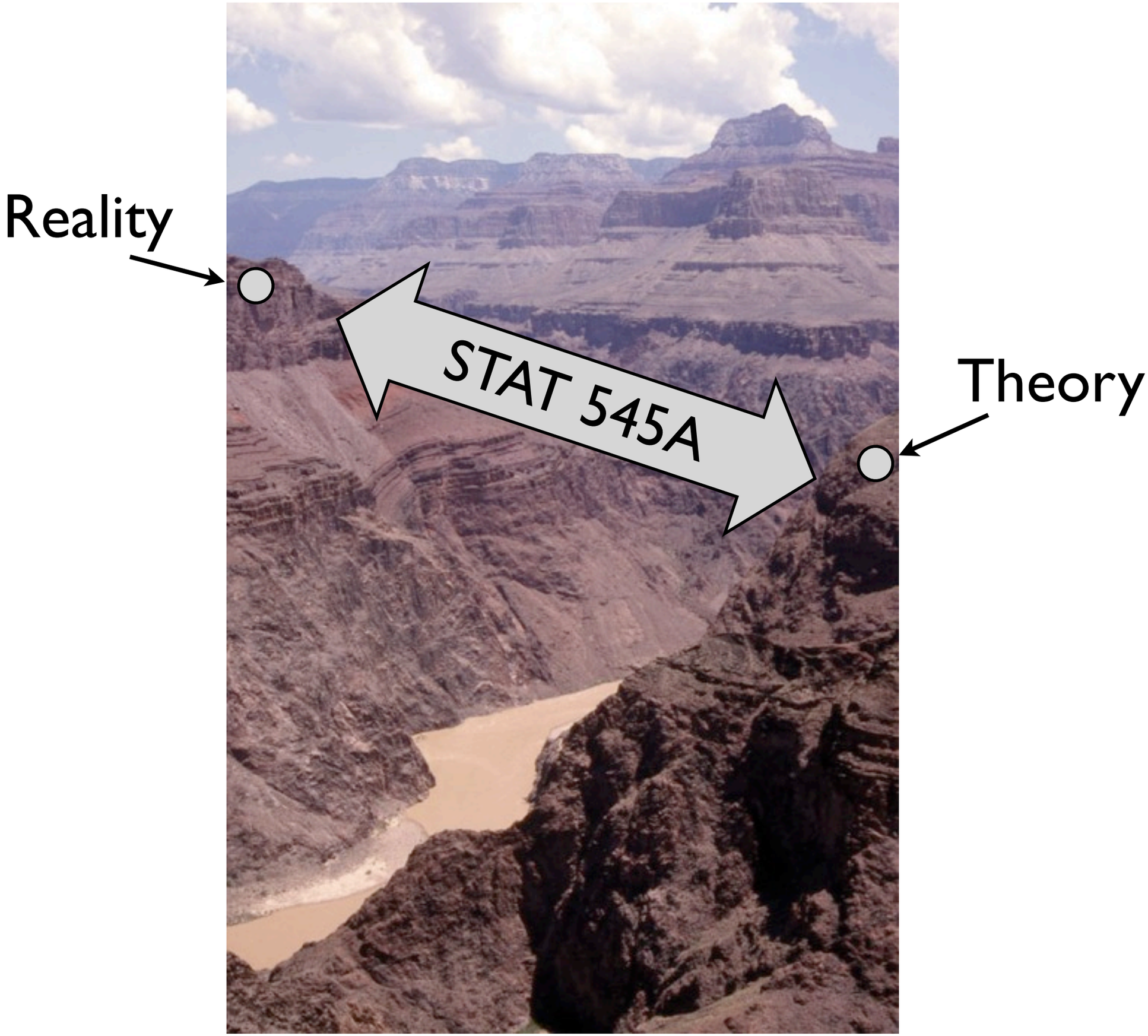
STAT 545A
Class meeting #1
Wednesday, September 4, 2013

Dr. Jennifer (Jenny) Bryan

Department of Statistics and Michael Smith Laboratories


Reality
STAT 545A
Theory

Two inter-related goals

- Foster your development of a personal philosophy for data analysis, esp. exploratory and descriptive analysis.
- Help you assemble a modern toolchain and workflows for data analysis.

My hope:

You'll leave this course with (at least the beginnings of) a confident, deliberate attitude about how to approach data analysis and the practical skills to put your attitude into action.

data analysis is awesome

make pictures obsessively

get serious about process, packaging and presentation

---

course stuff
software stuff
transition to “lab” mode

why data analysis?

"If a tree falls in a forest and no one is around to hear it, does it make a sound?"

http://pinker.wjh.harvard.edu/photos/Santa_Barbara_95-96/pages/fallen%20tree%20in%20Sequoia%20Forest.htm

Importance of data analysis and presentation for our discipline

"If a tree falls in a forest and no one is around to hear it, does it make a sound?"

"If a wonderful statistical method exists and no one uses it ... does it really exist? Is it accurate to call it 'wonderful'?

"If an important statistical result exists and no one truly grasps it ... does it really exist? Is it fair to call it 'important'?

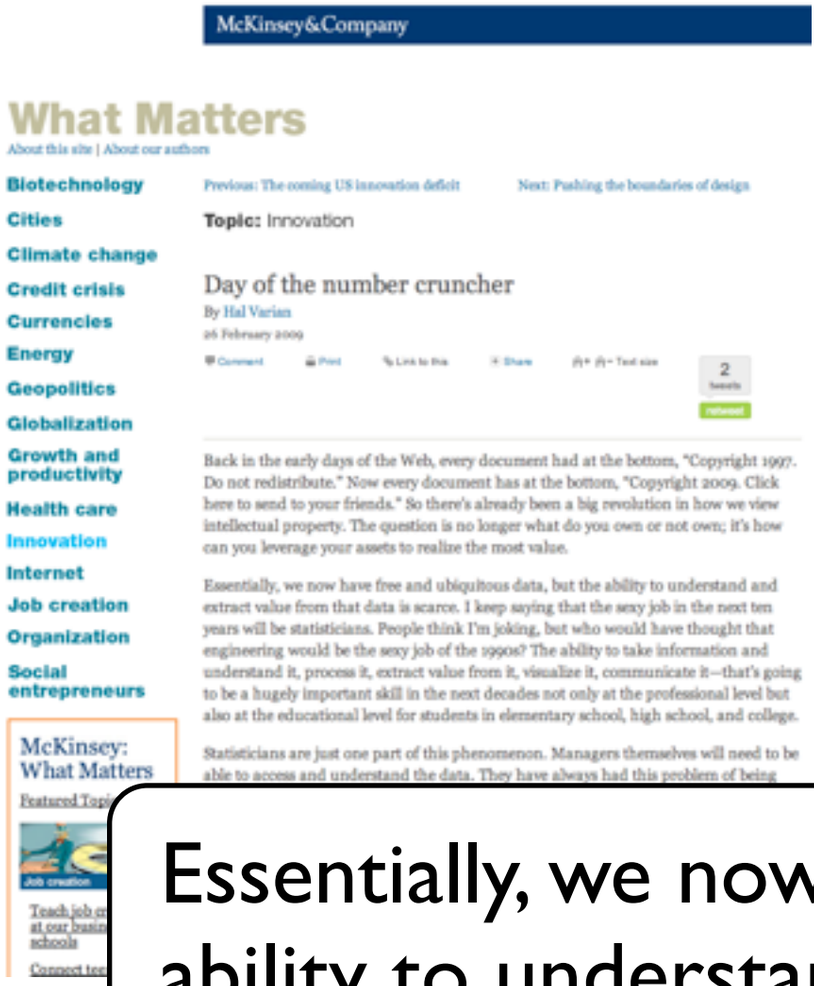
Importance of data analysis for our discipline

My claim:

Thoughtful, reproducible, well-presented data analyses present a tremendous opportunity for statistics to impact scientific research.

Maybe it's not necessary for every individual to excel in applied statistics, but it's vital that some people do. You could be one of them!

Importance of data analysis for your employability and quality of life ....


McKinsey&Company

What Matters

About this site | About our authors

Biotechnology
Cities
Climate change
Credit crisis
Currencies
Energy
Geopolitics
Globalization
Growth and productivity
Health care
Innovation
Internet
Job creation
Organization
Social entrepreneurs

Previous: The coming US innovation deficit Next: Pushing the boundaries of design

Topic: Innovation

Day of the number cruncher

By Hal Varian

26 February 2009

Comment Print Link to this Share Text size 2 tweets retweet

Back in the early days of the Web, every document had at the bottom, "Copyright 1997. Do not redistribute." Now every document has at the bottom, "Copyright 2009. Click here to send to your friends." So there's already been a big revolution in how we view intellectual property. The question is no longer what do you own or not own; it's how can you leverage your assets to realize the most value.

Essentially, we now have free and ubiquitous data, but the ability to understand and extract value from that data is scarce. I keep saying that the sexy job in the next ten years will be statisticians. People think I'm joking, but who would have thought that engineering would be the sexy job of the 1990s? The ability to take information and understand it, process it, extract value from it, visualize it, communicate it—that's going to be a hugely important skill in the next decades as well.

Statisticians are just one part of this phenomenon. Managers themselves will need to be able to access and understand the data. They have always had this problem of being

McKinsey: What Matters

Featured Topi

Job creation

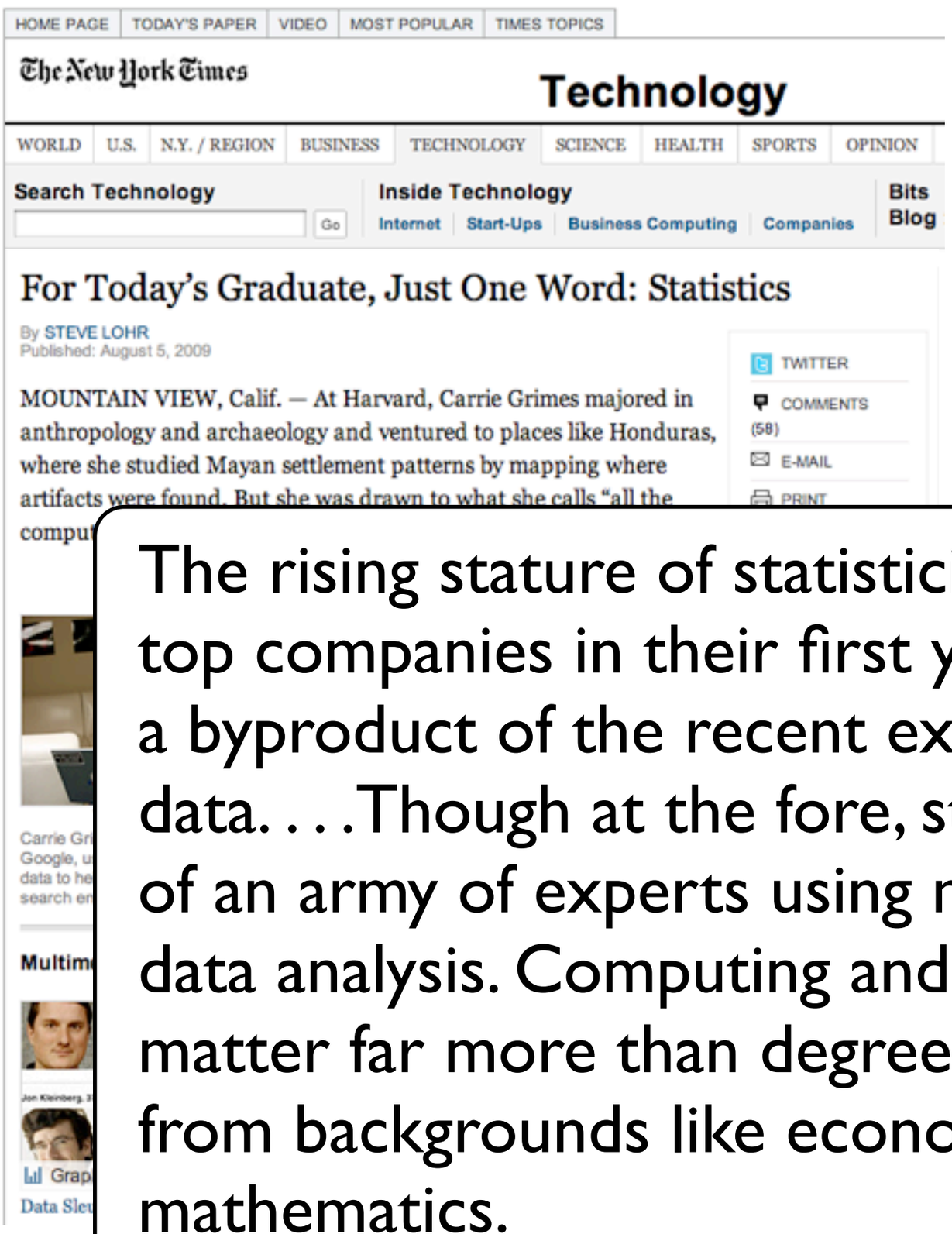
Teach job cr... at our busin... schools

Connect tod...

Hal Varian
Chief Economist, Google

Essentially, we now have free and ubiquitous data, but the ability to understand and extract value from that data is scarce. I keep saying that the sexy job in the next ten years will be statisticians. People think I'm joking, but who would have thought that engineering would be the sexy job of the 1990s? The ability to take information and understand it, process it, extract value from it, visualize it, communicate it—that's going to be a hugely important skill in the next decades

http://whatmatters.mckinseydigital.com/innovation/day-of-the-number-cruncher


HOME PAGE | TODAY'S PAPER | VIDEO | MOST POPULAR | TIMES TOPICS

The New York Times

Technology

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

Search Technology | Go

Inside Technology: Internet | Start-Ups | Business Computing | Companies

Bits Blog

For Today's Graduate, Just One Word: Statistics

By STEVE LOHR
Published: August 5, 2009

MOUNTAIN VIEW, Calif. — At Harvard, Carrie Grimes majored in anthropology and archaeology and ventured to places like Honduras, where she studied Mayan settlement patterns by mapping where artifacts were found. But she was drawn to what she calls "all the

TWITTER
COMMENTS (58)
E-MAIL
PRINT

Carrie Gri
Google, u
data to he
search en

Multim

Jon Kleinberg, 3

Grap

Data Sleu

"For Today's Graduate, Just One Word: Statistics" by Steve Lohr, New York Times, August 5, 2009.

The rising stature of statisticians, who can earn $125,000 at top companies in their first year after getting a doctorate, is a byproduct of the recent explosion of digital data. . . . Though at the fore, statisticians are only a small part of an army of experts using modern statistical techniques for data analysis. Computing and numerical skills, experts say, matter far more than degrees. So the new data sleuths come from backgrounds like economics, computer science and mathematics.

http://www.nytimes.com/2009/08/06/technology/06stats.html

What is data science?

Analysis: The future belongs to the companies and people that turn data into products.

by Mike Loukides | @mikeloukides | Comments: 52 | 2 June 2010

We've all heard it: according to Hal Varian, statistics is the next sexy job. Five years ago, in What is Web 2.0, Tim O'Reilly said that "data is the next Intel Inside." But what does that statement mean? Why do we suddenly care about statistics and about data?

In this post, I examine the many sides of data science -- the technologies, the companies and the unique skill sets.

What is data science?

The web is full of "data-driven apps." Almost any e-commerce application is a data-driven application. There's a database behind a web front end, and middleware that talks to a number of other databases and data services (credit card processing companies, banks, and so on). But merely using data isn't really what we mean by "data science." A data application acquires its value from the data itself, and creates more data

Report sections

http://radar.oreilly.com/2010/06/what-is-data-science.html

"Statistics: Your chance for happiness (or misery)" by Xiao-Li Meng, The Harvard Undergraduate Research Journal, Volume 2 Issue 1 | Spring 2009.

Statistics: Your chance for happiness (or misery)

by ADMIN on JANUARY 30, 2011 · LEAVE A COMMENT

By Professor Xiao-Li Meng

Whipple V.N. Jones Professor of Statistics and Department Chair

Professor Meng and his "Happy Team" on the opening day of Stat 105
Cassandra Wolos, Kari Lock, Xiao-Li Meng, Yves Chretien, and Paul Edlefsen

1. "I keep saying the sexy job in the next ten years will be statisticians."

Hal Varian, Google's chief economist, recently was interviewed by McKinsey Quarterly, and was quoted (see www.mckinseyquarterly.com/Strategy/Innovation/):

"I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s? The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it."

SAN FRANCISCO BAY AREA | APRIL 8, 2010

New Hiring Formula Values Math Pros

Region's Employers Seek Statistical Experts Over Computer-Science Generalists

BY JESSICA E. VASCELLARO

Being a math geek has never been cooler, at least in Silicon Valley.

As Bay Area technology companies ramp up hiring out of the recession, they are in hot pursuit of a particular kind of employee: those with experience in statistics and other data-manipulation techniques.

Rather than looking for just plain-vanilla computer scientists, who typically don't have as deep a study of math and statistics, companies from Facebook Inc. to online advertising company AdMob Inc. say they need more workers with stronger backgrounds in statistics and a related field called machine learning, which involves writing algorithms that get smarter over ...

http://online.wsj.com/article_email/SB10001424052702304871704575160553254798886-lMyQjAxMTAwMDAwODEwNDgyWj.html

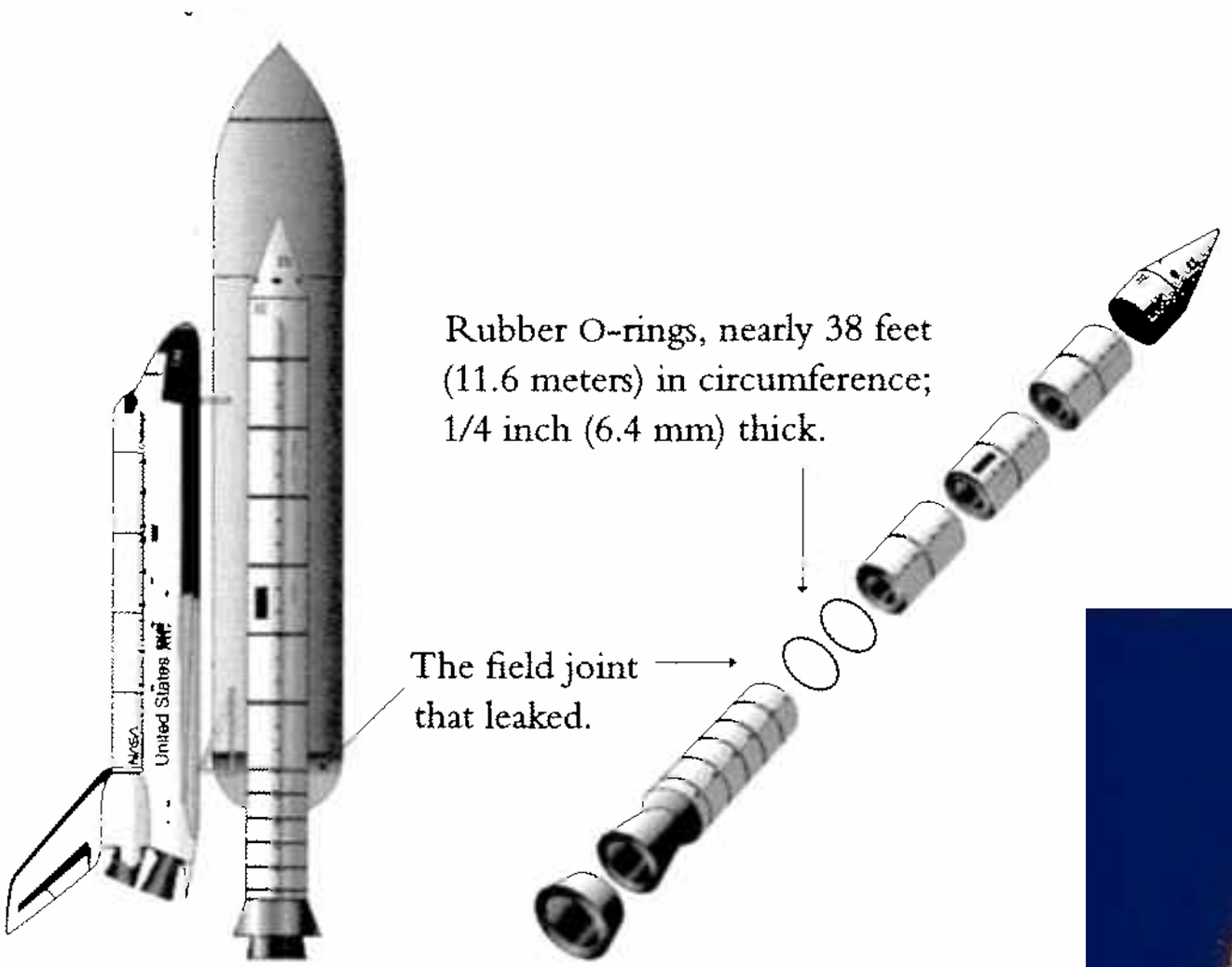
1. **Build your communication skills.** Unless you live in a plastic bubble, you are going to need to work with other people. You will be given tasks by other people, collaborate with other people to achieve those tasks, and ultimately have to report the results of your work to other people. You need to be able to speak clearly and concisely, listen carefully, write well (and quickly), and give informative and interesting presentations. Contrary to popular belief *the person who learns to do these things well will advance farther than someone who has better technical capabilities but poor communication skills!* Management usually can't tell the difference between a good statistician and a great one, but they can see immediately who communicates their results well and who does so poorly. Unfortunately, most university environments stress working alone and in isolation, completely the opposite of what life will be like on the other side of graduation.

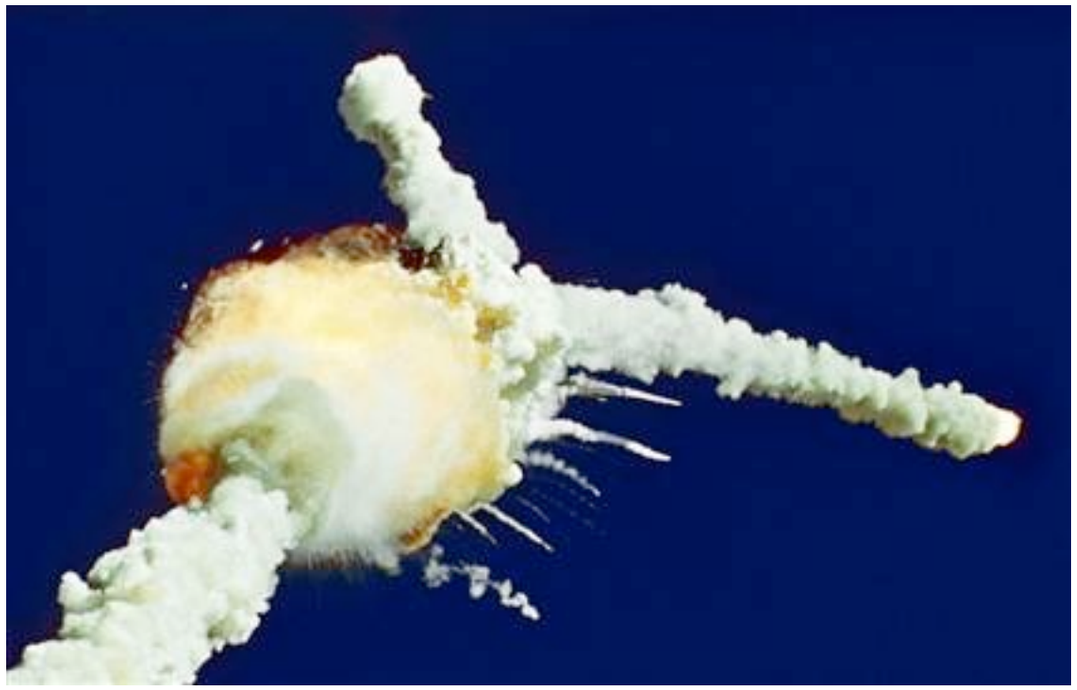
   You need to take actions to ensure that your communication skills are sharp. These actions can include: (i) Taking a writing class, especially one that stresses technical writing, which has a completely different flavour from essay writing; (ii) Taking a class in verbal communication, and in particular one that covers the fine art of making and delivering presentations; (iii) Taking business courses, especially those in business communication and organizational structure and behaviour, so that you can better understand your audience and learn to arrange your communications accordingly; (iv) Seeking out courses that expressly advertise group project work and/or presentations, even (*especially!*) if these things scare you. All of our speakers indicate that anything that you can do to practice your communication skills will have a positive effect on your employability and advancement.

Excerpt from "Real Advice from Real People" by Tom Loughin, Statistical Society of Canada Liaison, Vol. 22 No. 4 (November 2008).

“A picture is worth a thousand words”

1986 Challenger space shuttle disaster
Favorite example of Edward Tufte





http://msnbcmedia1.msn.com/j/msnbc/Components/Photos/050709/050609_columbia_hmed_6p.hmedium.jpg

TEMPERATURE CONCERN ON SRM JOINTS

27 JAN 1986

HISTORY OF O-RING DAMAGE ON SRM FIELD JOINTS

	SRM No.	Cross Sectional View: Erosion Depth (in.)	Cross Sectional View: Perimeter Affected (deg)	Cross Sectional View: Nominal Dia. (in.)	Top View: Length Of Max Erosion (in.)	Top View: Total Heat Affected Length (in.)	Clocking Location (deg)
61A LH Center Field**	22A	None	None	0.280	None	None	36°--66°
61A LH ~~CENTER~~ FIELD**	22A	NONE	NONE	0.280	NONE	NONE	338°-18°
51C LH Forward Field**	15A	0.010	154.0	0.280	4.25	5.25	163
51C RH Center Field (prim)***	15B	0.038	130.0	0.280	12.50	58.75	354
51C RH Center Field (sec)***	15B	None	45.0	0.280	None	29.50	354
41D RH Forward Field	13B	0.028	110.0	0.280	3.00	None	275
41C LH Aft Field*	11A	None	None	0.280	None	None	--
41B LH Forward Field	10A	0.040	217.0	0.280	3.00	14.50	351
STS-2 RH Aft Field	2B	0.053	116.0	0.280	--	--	90

*Hot gas path detected in putty. Indication of heat on O-ring, but no damage.
**Soot behind primary O-ring.
***Soot behind primary O-ring, heat affected secondary O-ring.

Clocking location of leak check port - 0 deg.

OTHER SRM-15 FIELD JOINTS HAD NO BLOWHOLES IN PUTTY AND NO SOOT NEAR OR BEYOND THE PRIMARY O-RING.

SRM-22 FORWARD FIELD JOINT HAD PUTTY PATH TO PRIMARY O-RING, BUT NO O-RING EROSION AND NO SOOT BLOWBY. OTHER SRM-22 FIELD JOINTS HAD NO BLOWHOLES IN PUTTY.

BLOW BY HISTORY

SRM-15 WORST BLOW-BY
- 2 CASE JOINTS (80°), (110°) ARC
- MUCH WORSE VISUALLY THAN SRM-22

SRM 22 BLOW-BY
- 2 CASE JOINTS (30-40°)

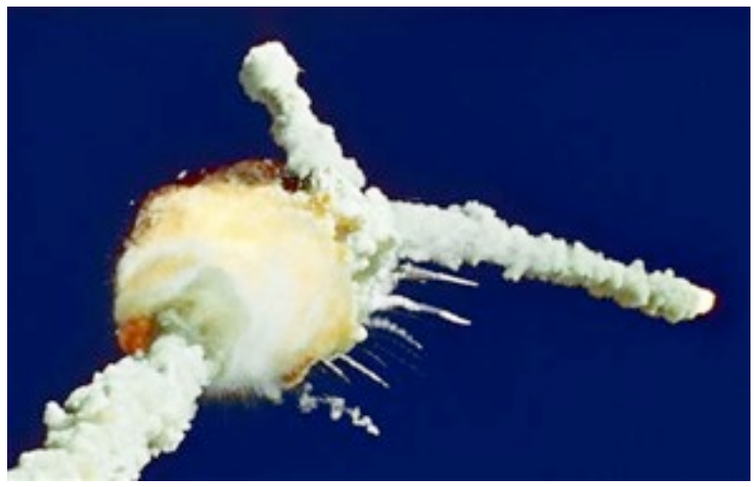
SRM-13A, 15, 16A, 18, 23A 24A
- NOZZLE BLOW-BY

HISTORY OF O-RING TEMPERATURES (DEGREES-F)

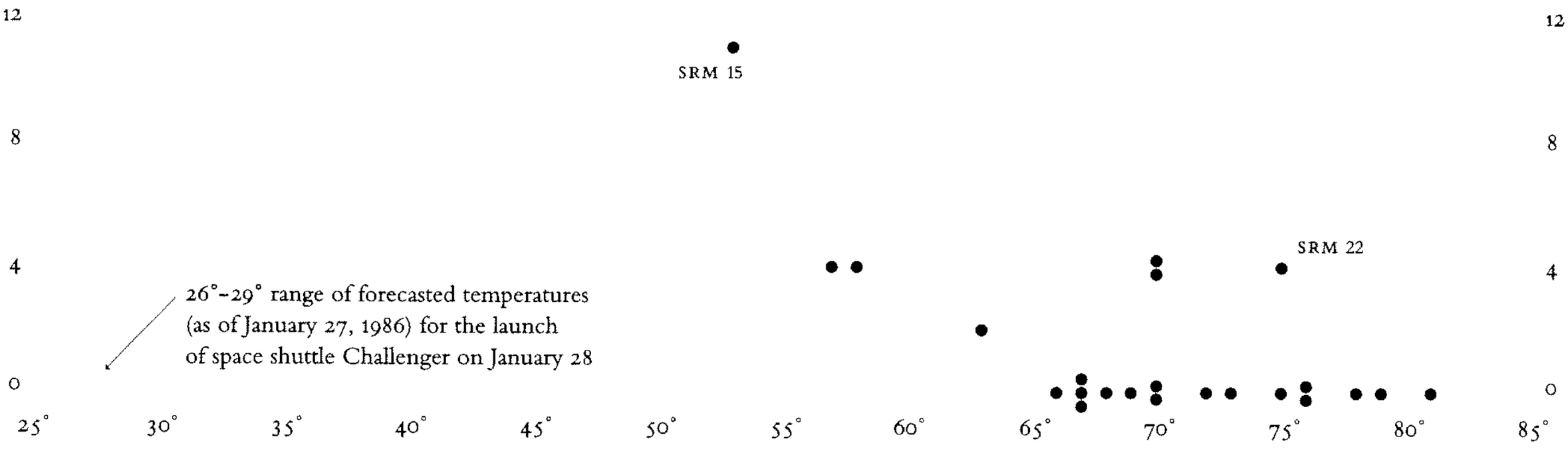
MOTOR	MBT	AMB	O-RING	WIND
DM-1	68	36	47	10 MPH
DM-2	76	45	52	10 MPH
QM-3	72.5	40	48	10 MPH
QM-4	76	48	51	10 MPH
SRM-15	52	64	53	10 MPH
SRM-22	77	78	75	10 MPH
SRM-25	55	26	29 27	10 MPH 25 MPH

MOTOR	O-RING
DM-1	47
DM-2	52
QM-3	48
QM-4	51
SRM-15	53
SRM-22	75
SRM-25	29 27

“A picture is worth a thousand words”



O-ring damage
index, each launch



Temperature (°F) of field joints at time of launch

“A picture is worth a thousand words”

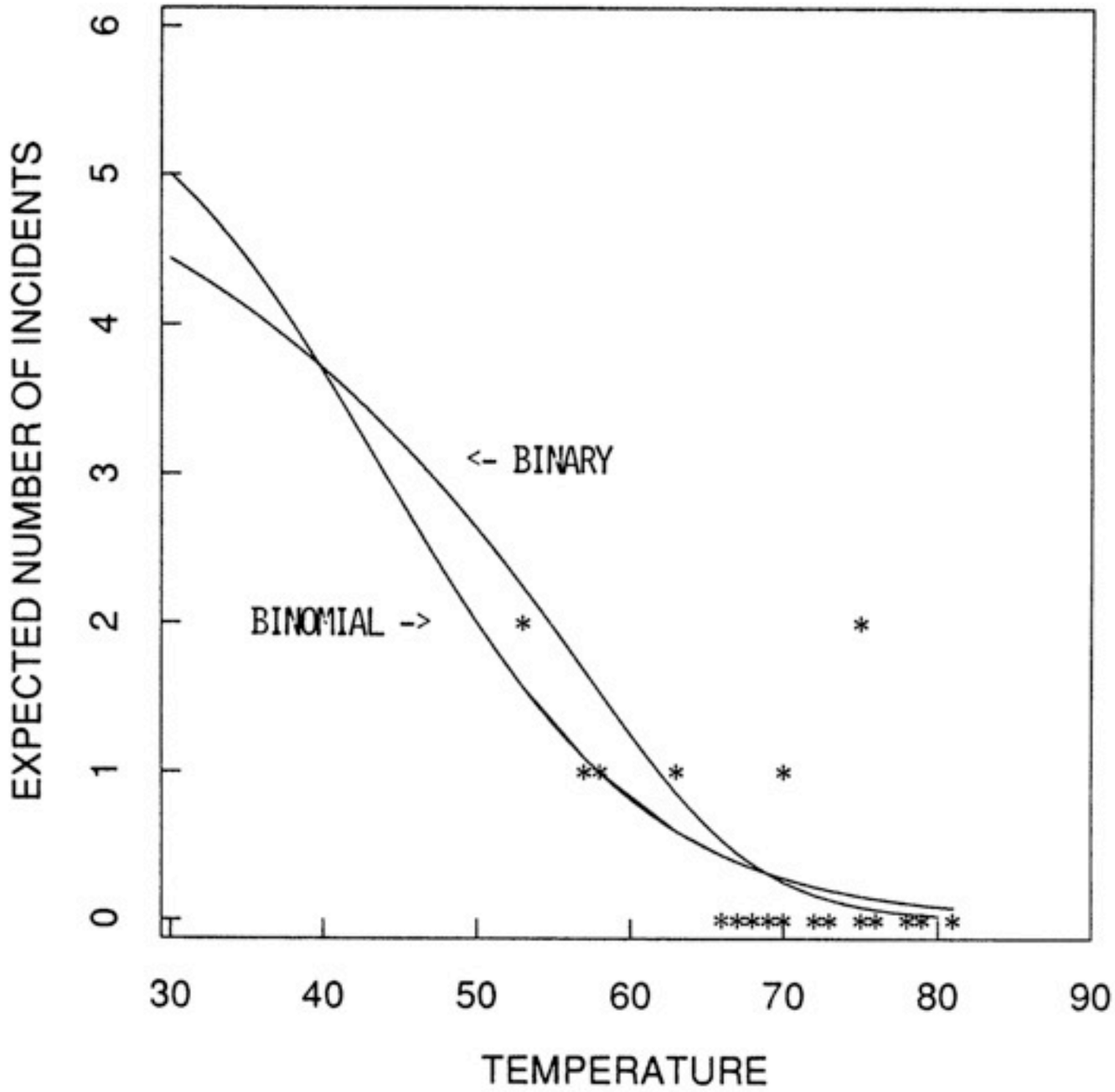
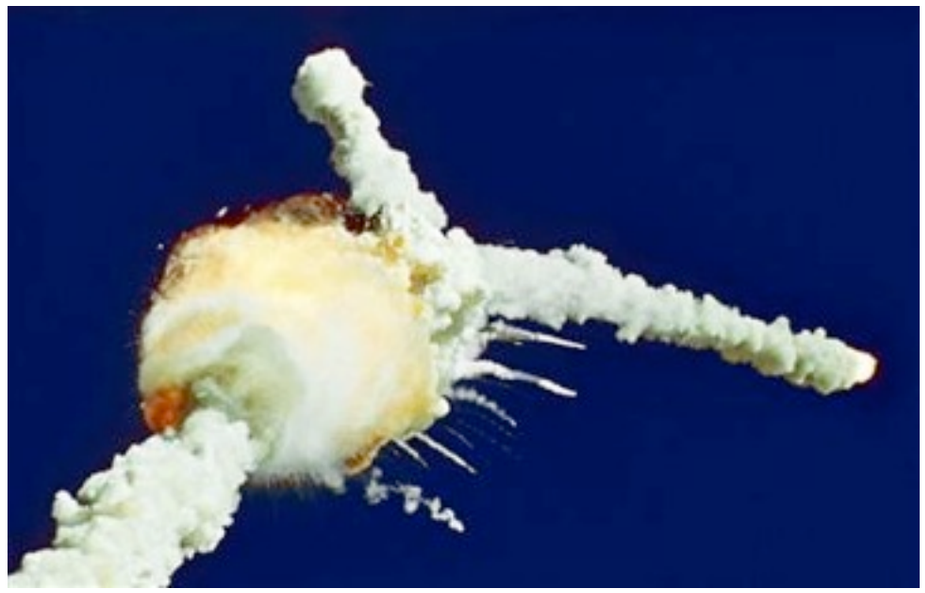


Figure 4. O-Ring Thermal-Distress Data: Field-Joint Primary O-Rings, Binomial-Logit Model, and Binary-Logit Model.



Siddhartha R. Dalal; Edward B. Fowlkes; Bruce Hoadley. Risk Analysis of the Space Shuttle: Pre-Challenger Prediction of Failure. JASA, Vol. 84, No. 408 (Dec., 1989), pp. 945-957. Access via JSTOR.

Edward Tufte
http://www.edwardtufte.com

BOOK:
Visual Explanations: Images and Quantities, Evidence and Narrative

Ch. 5 deals with the Challenger disaster
That chapter is available for $7 as a downloadable booklet:
http://www.edwardtufte.com/tufte/books_textb

“A picture is worth a thousand words”

Always, always, always plot the data.

Replace (or complement) ‘typical’ tables of data or statistical results with figures that are more compelling and accessible.

Whenever possible, generate figures that overlay / juxtapose observed data and analytical results, e.g. the ‘fit’.

“A picture is worth a thousand words”

Why?

- find bizarre data and results when it is least embarrassing and painful
- facilitate comparisons and reveal trends

Recommended reference: Gelman A, Pasarica C, Dodhia R. “Let's Practice What We Preach: Turning Tables into Graphs”. The American Statistician, Volume 56, Number 2, 1 May 2002 , pp. 121-130(10). via JSTOR

Statistical Computing and Graphics

Let’s Practice What We Preach: Turning Tables into Graphs

Andrew GELMAN, Cristian PASARICA, and Rahul DODHIA

Statisticians recommend graphical displays but often use tables to present their own research results. Could graphs do better? We study the question by going through the tables in a recent issue of the *Journal of the American Statistical Association*. We show how it is possible to improve the presentations using graphs that actually take up less space than the original tables. We find a particularly effective tool to be multiple repeated line plots,

plays. Our advice follows well-known principles of data display (see, e.g., Tufte 1983; Cleveland 1985) but applied to the presentation or research results as well as raw data.

2. DISPLAYING NUMERICAL RESULTS

Statistical research requires the display of many different kinds of numerical results, including raw numbers, data reductions, inferences, and—for research in theory and

"All models are wrong, some models are useful."

Box, G.E.P., Robustness in the strategy of scientific model building, in Robustness in Statistics, R.L. Launer and G.N. Wilkinson, Editors. 1979, Academic Press: New York.

Entia non sunt multiplicanda praeter necessitatem

The principle, known as Occam's Razor, that says: when there are two competing theories or explanations -- both compatible with observed data, known facts -- the simpler one is better.

Implication for statistical analysis: if two models are equally wrong-but-compatible-with-data, the simpler one is more useful!

http://blog.okcupid.com/index.php/dont-be-ugly-by-accident/

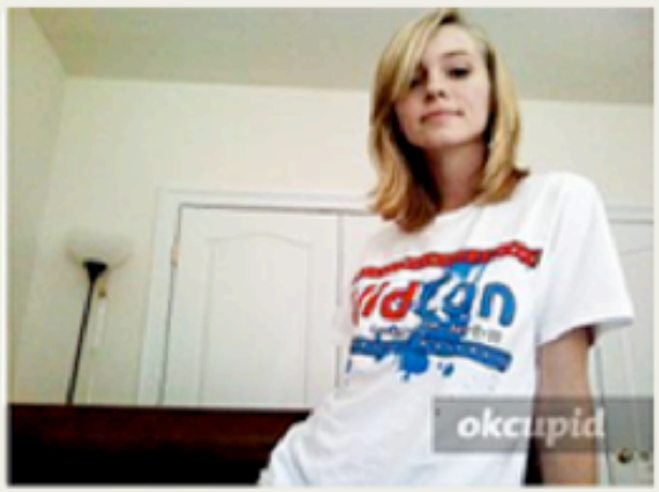
oktrends

Dating Research from OkCupid

Our experiment:

1. We collected 552,000 example user pictures.
2. We paired them up and asked people to make snap judgments, like so:

Who would you rather go on a date with?



Her

Her

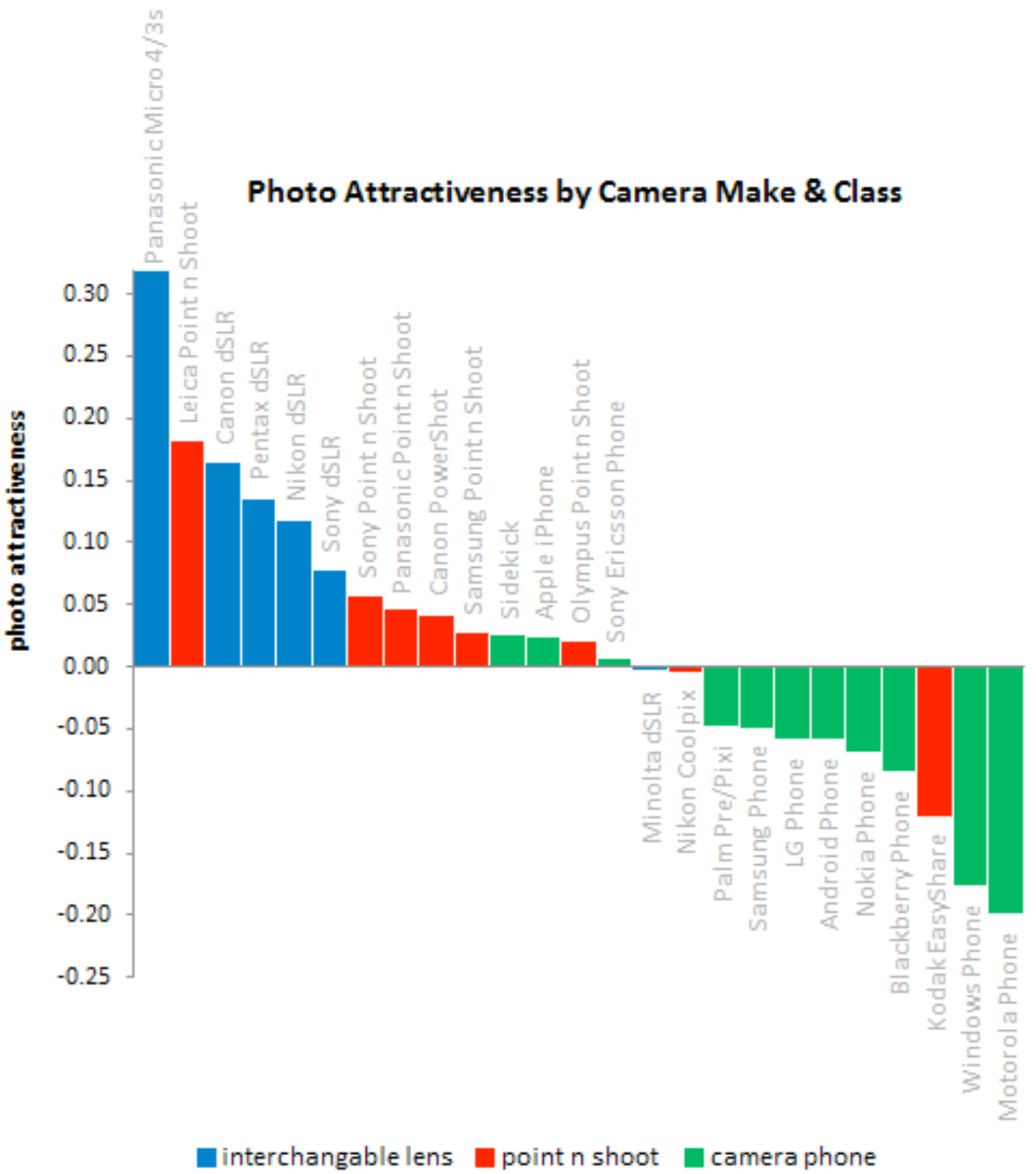
3. We collated these millions of judgments with the time of day each picture was taken, what the shutter speed was, and so on. Almost all modern cameras embed this stuff in a special header, called *EXIF data*.
4. We made graphs.

About OkTrends

OkTrends is original research and insights from OkCupid, the best dating site on earth. We've compiled our observations and statistics from hundreds of millions of OkCupid user interactions, all to explore the data side of the online dating world.

1. Panasonic > Canon > Nikon.

The type and brand of camera you use has a huge effect on how good you look in your pictures. This is a plot of the most popular makes:



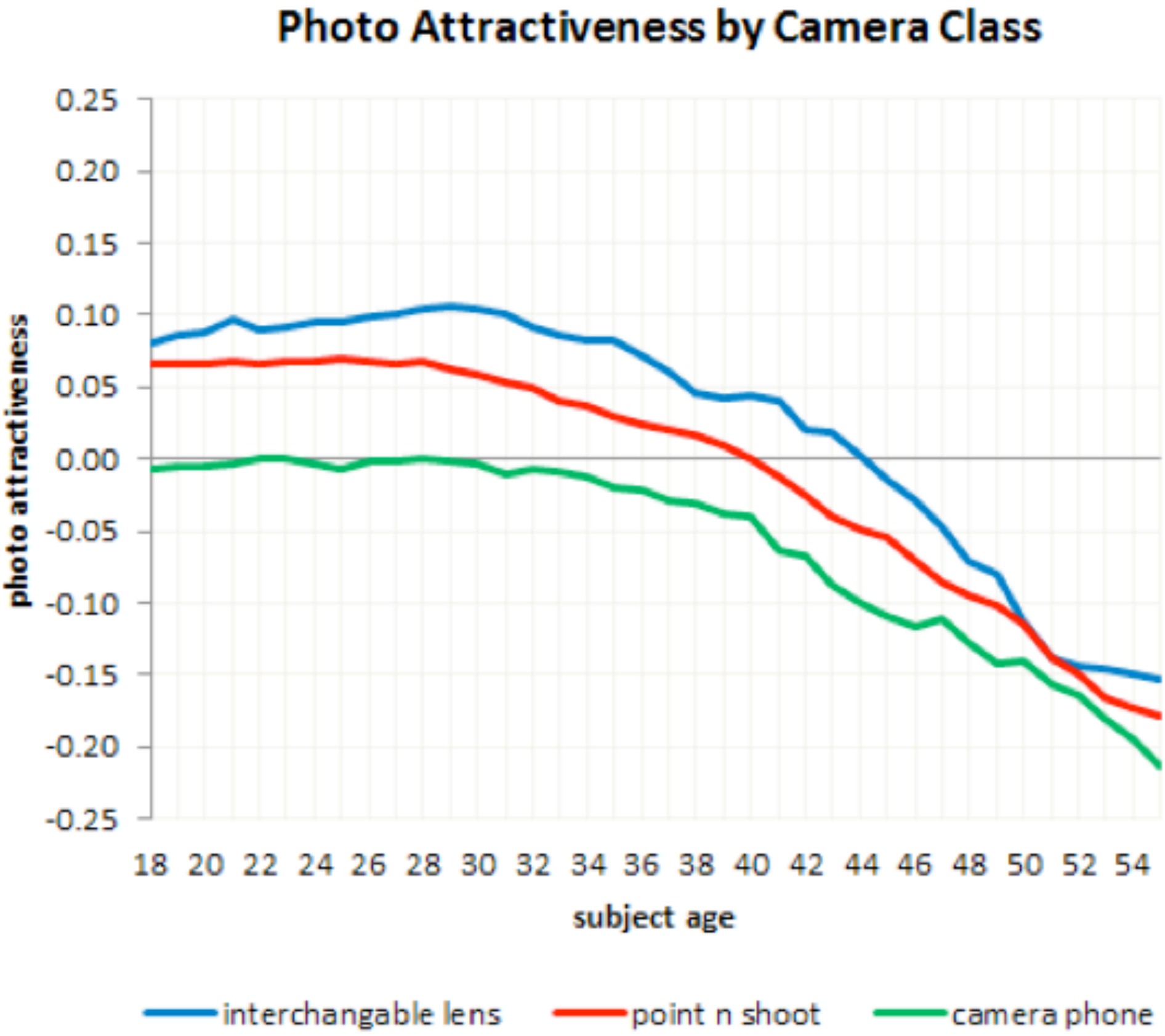
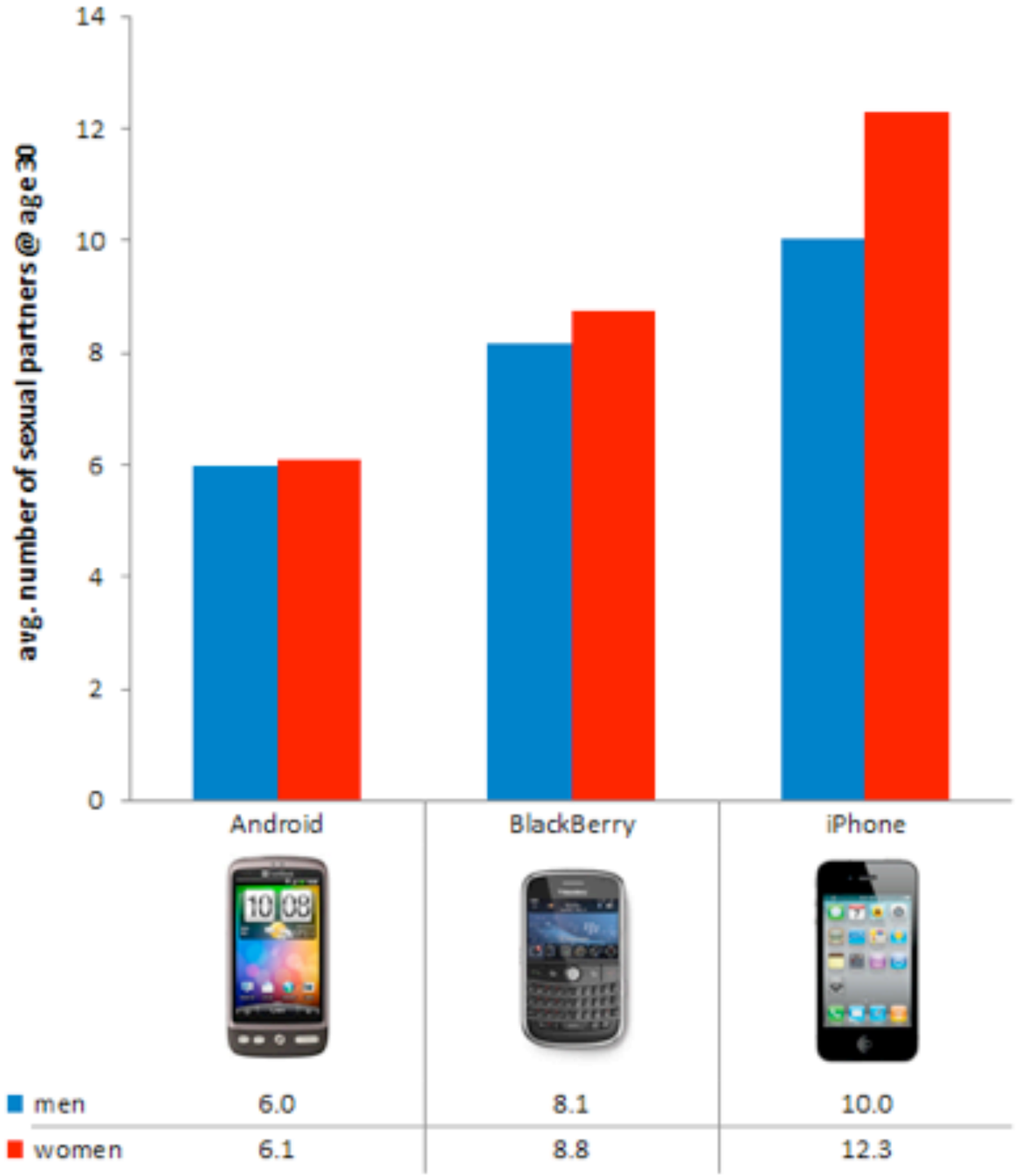
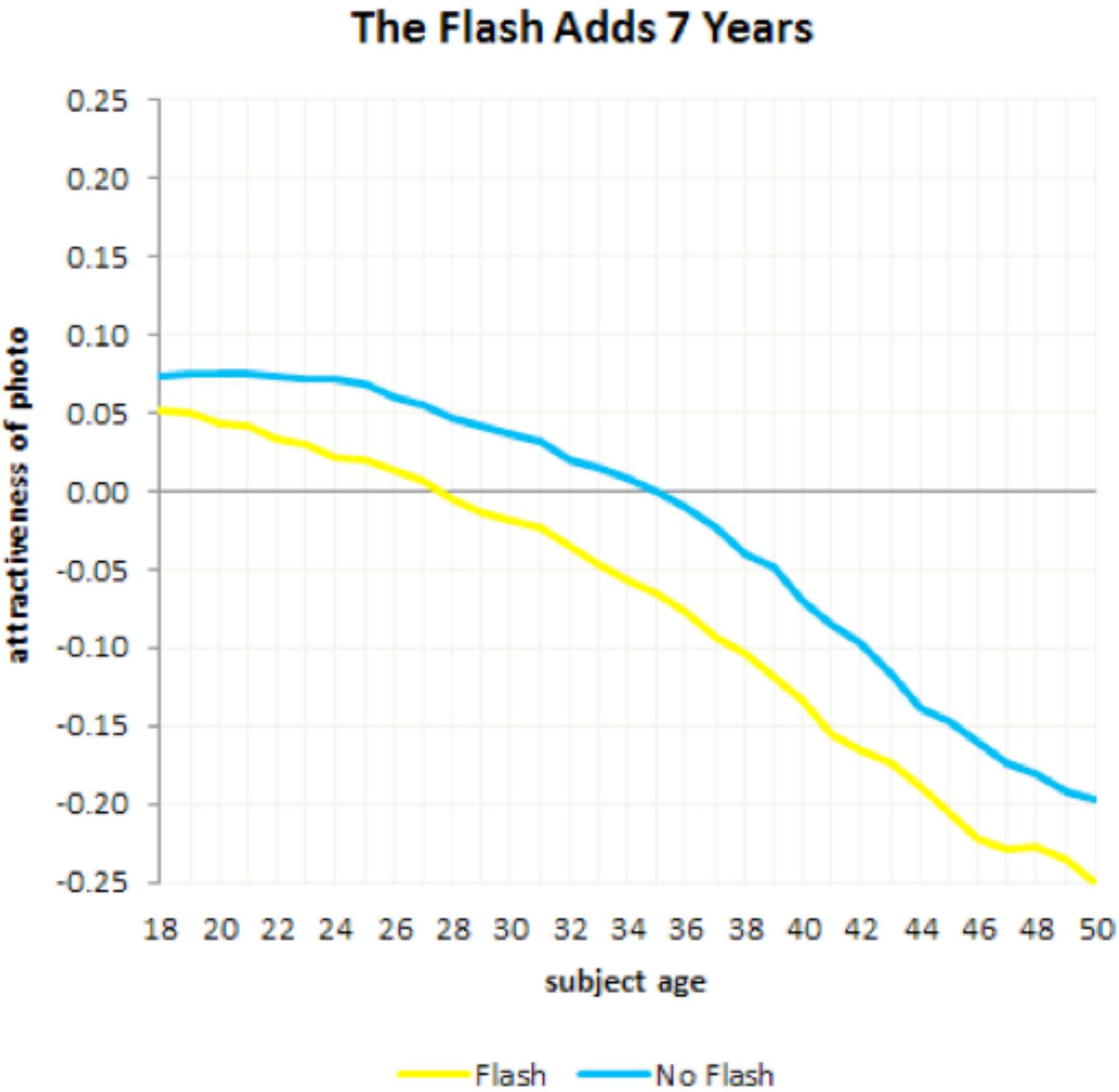


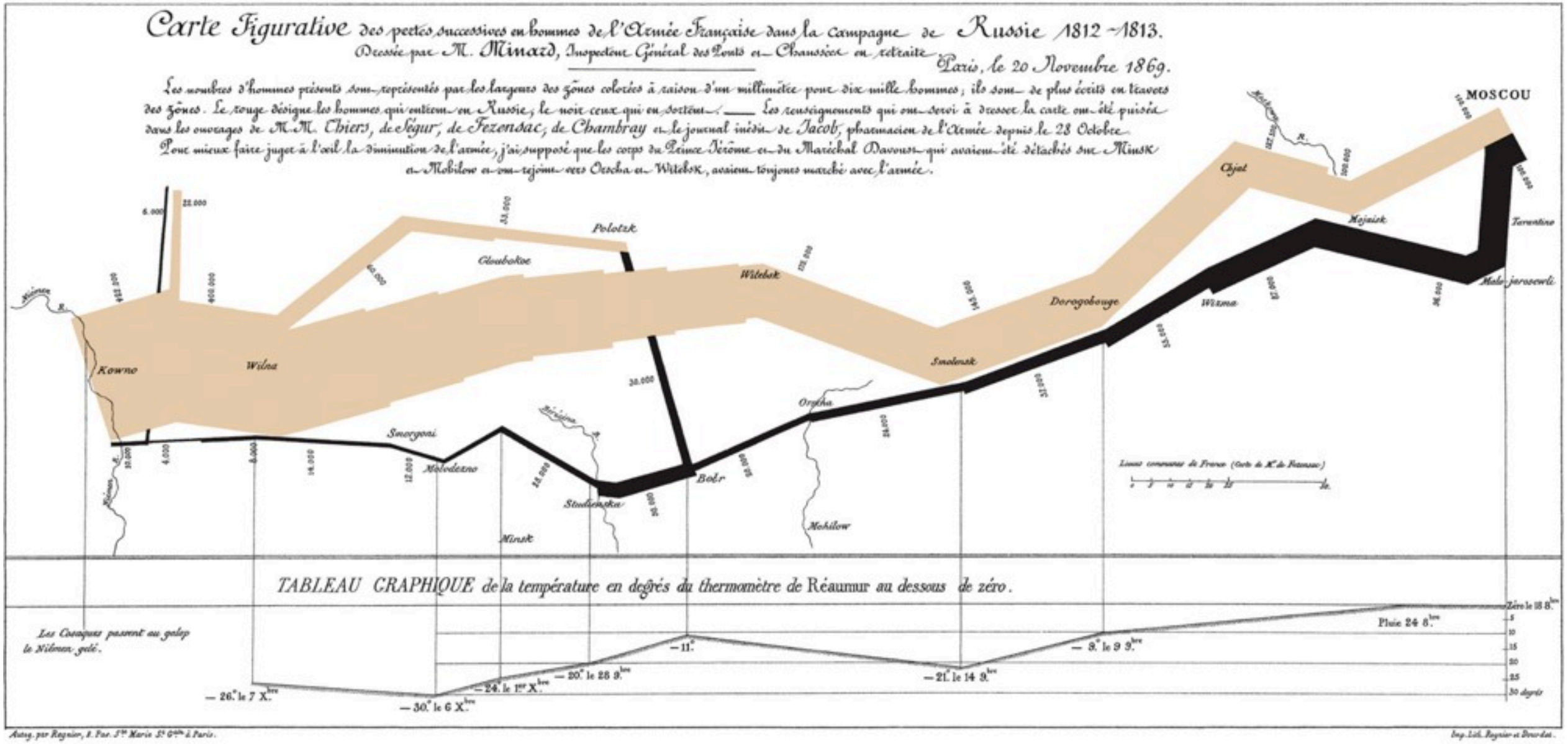
Photo Attractiveness by Camera Class
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.20
-0.25
photo attractiveness
18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54
subject age
interchangable lens
point n shoot
camera phone

Sexual Activity by Smart Phone Brand




The Flash Adds 7 Years
attractiveness of photo
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.20
-0.25
18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50
subject age
Flash
No Flash

www.gapminder.org (via YouTube)

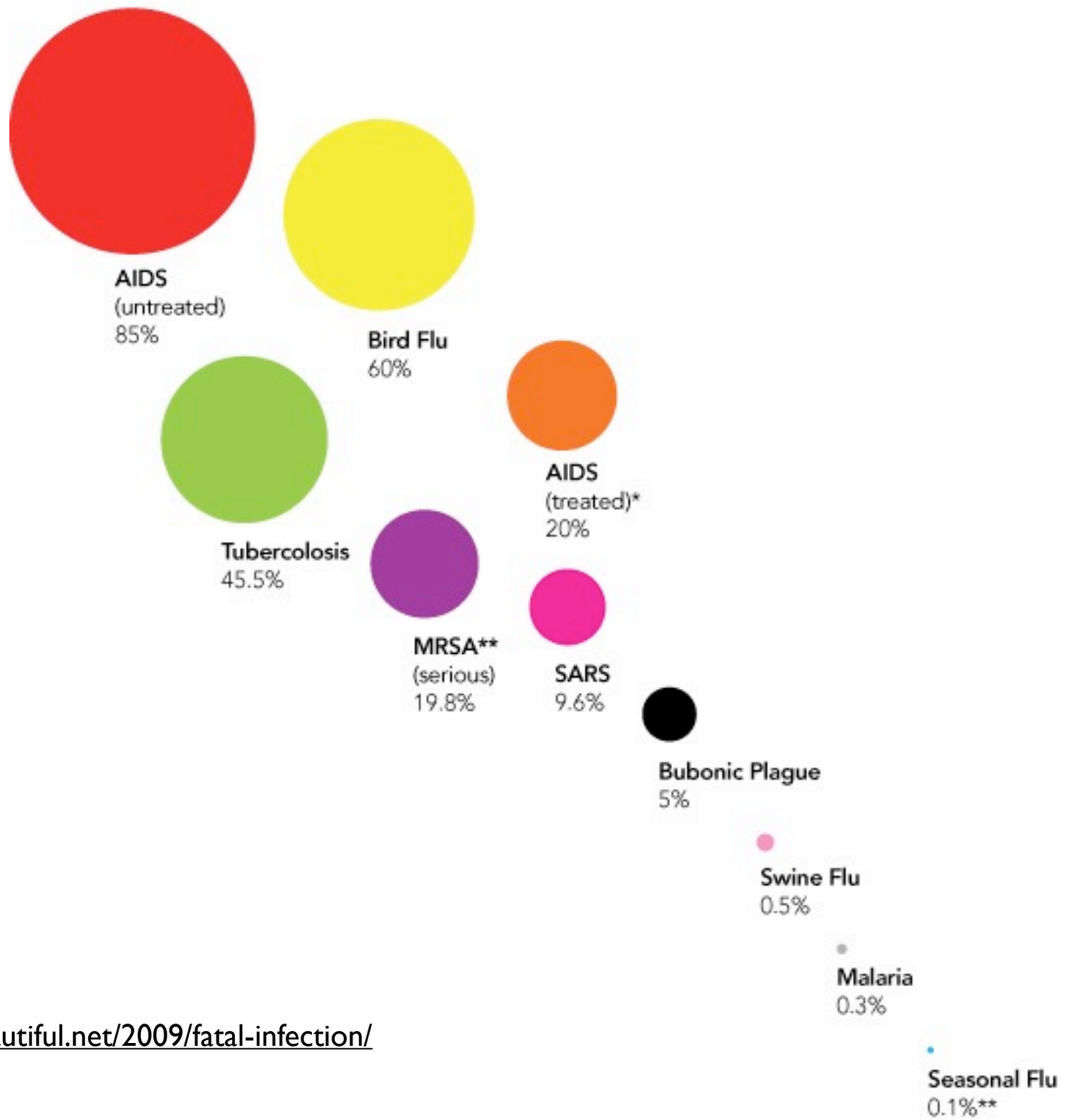


Famous chart/map by Charles Joseph Minard, much beloved by Tufte, depicting Napoleonic army during Russian campaign of 1812

http://upload.wikimedia.org/wikipedia/commons/2/29/Minard.png

Disease Case Fatality Rates

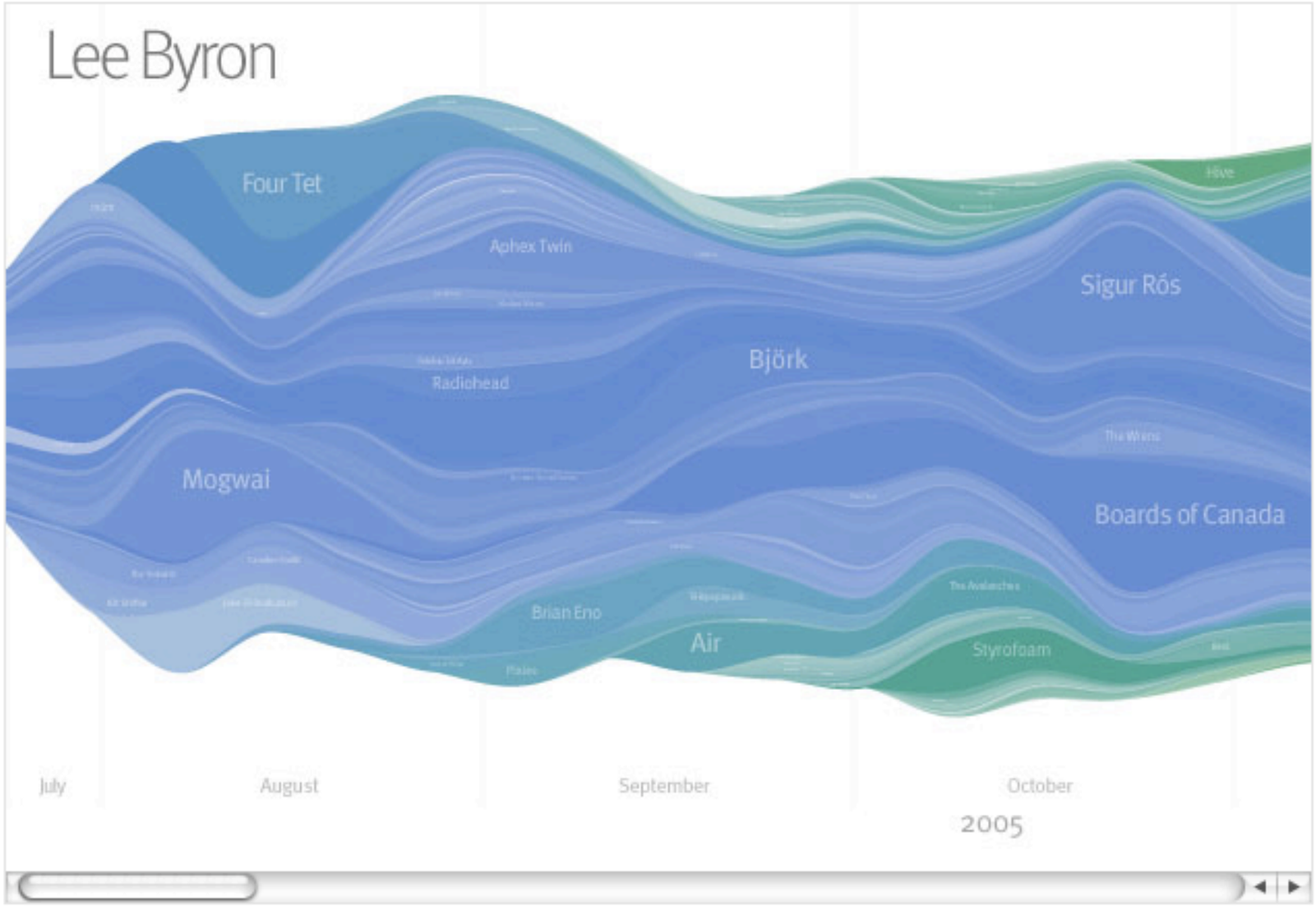
Average % of infected who die



http://www.informationisbeautiful.net/2009/fatal-infection/

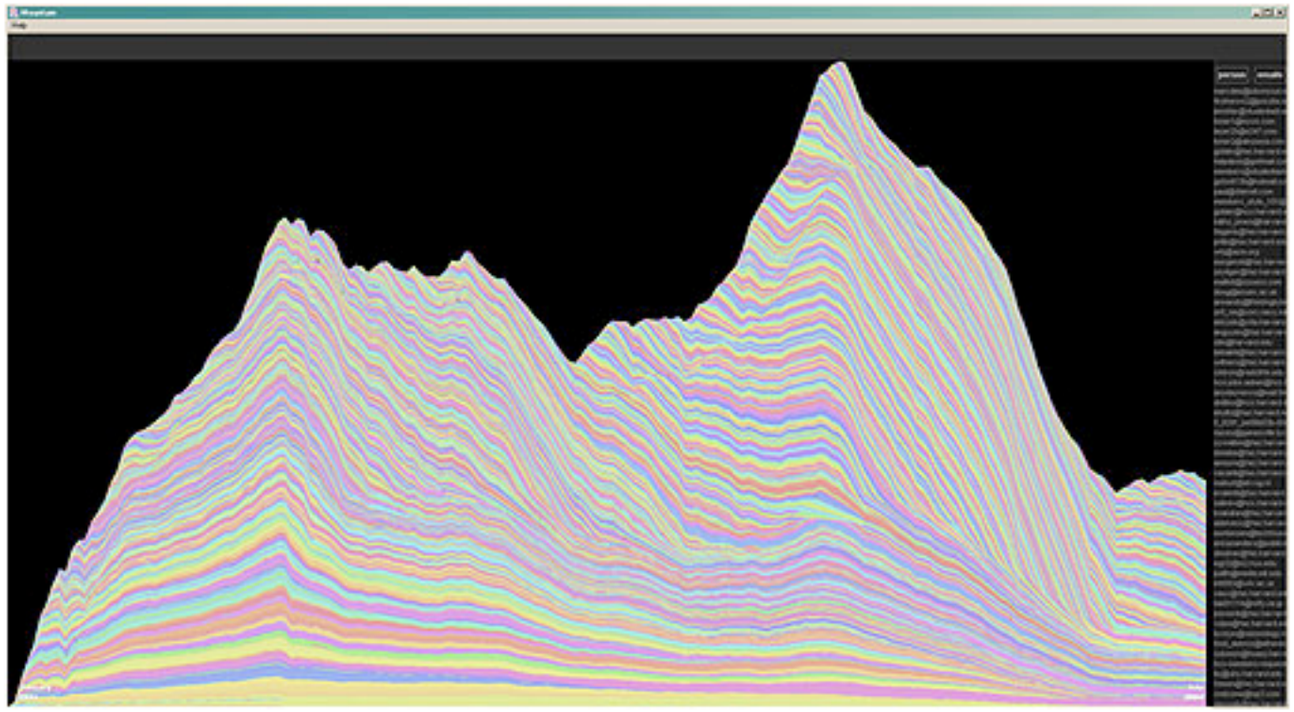
source: Worldwide figures from World Health Organisation, CDC, Guardian
*estimated from secondary sources // ** based on US figures

David McCandless // v1.2 // Sep 09
InformationIsBeautiful.net



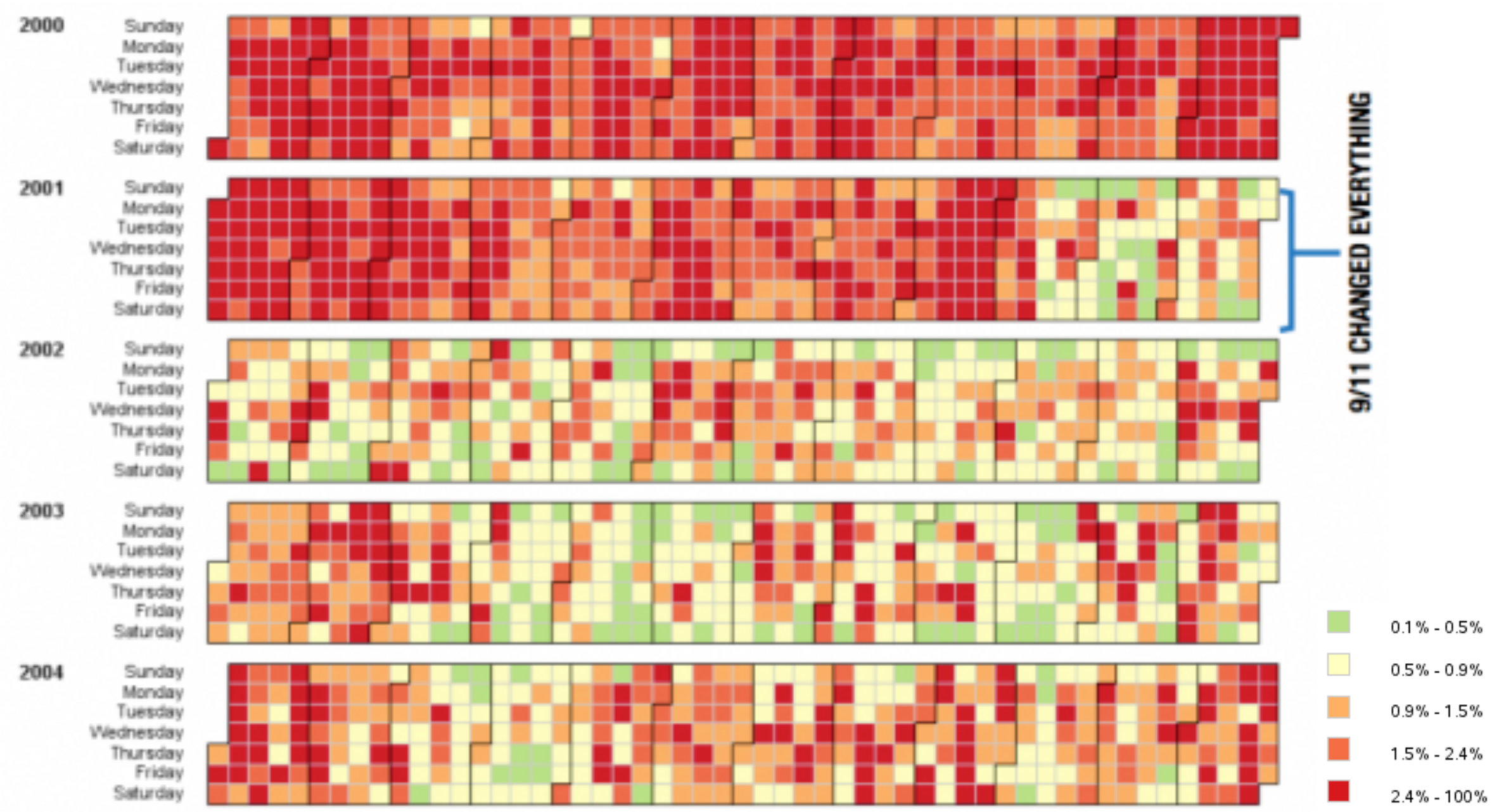
http://www.leebyron.com/what/lastfm/

http://alumni.media.mit.edu/~fviegas/projects/mountain/



In the mountain above, the owner of the email archive has graduated from one school and moved to a new university for his graduate studies. This is the reason why we see two distinct peaks; the mountain on the right represents the surge of new contacts this person has made in the new school.

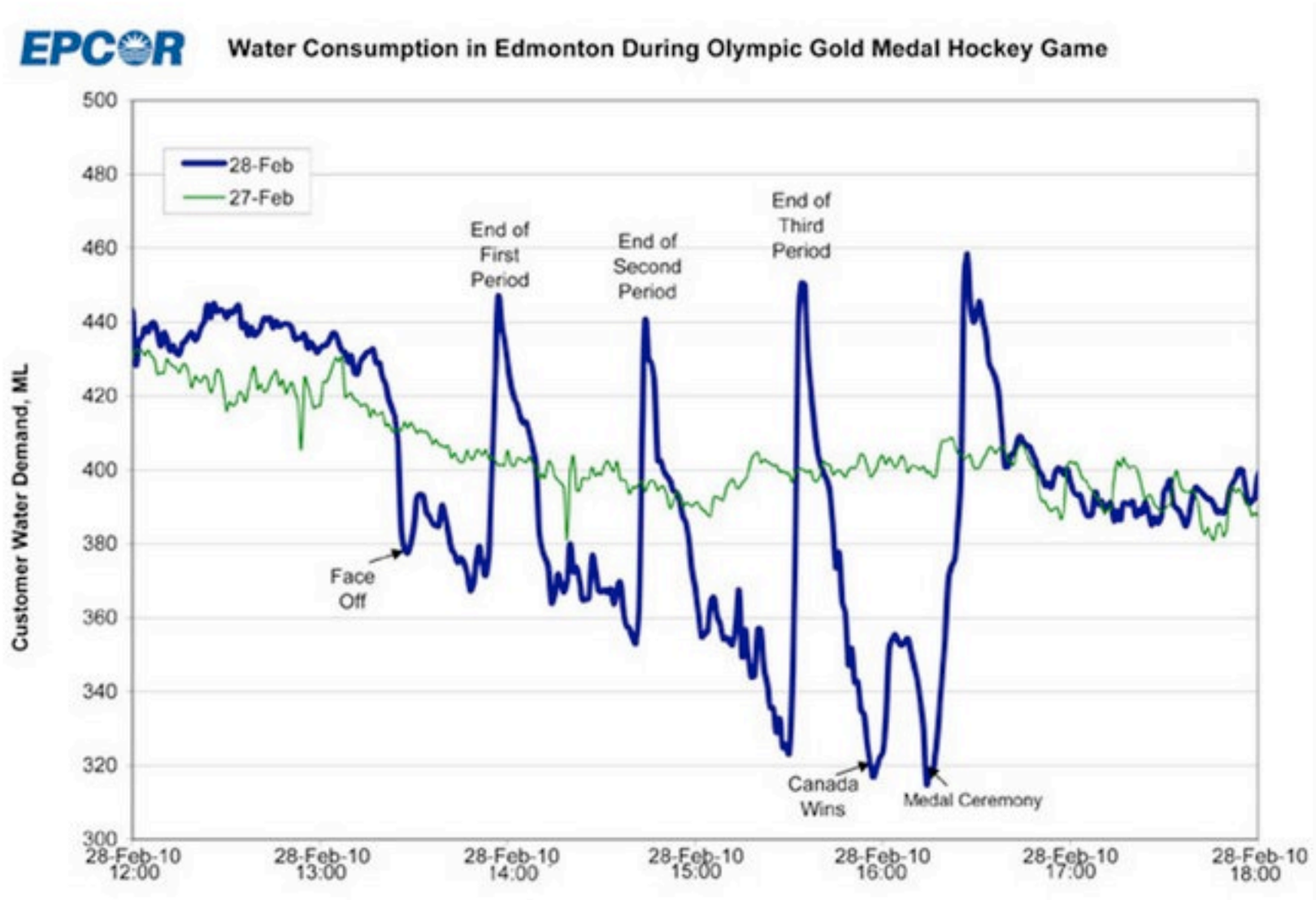
This is a view of flight cancellations. The more red a rectangle is, the higher the percentage of cancellations.



http://flowingdata.com/2009/09/10/3-in-depth-views-of-flight-delays-and-cancellations/

What If Everybody in Canada Flushed At Once?

http://www.patspapers.com/blog/item/what_if_everybody_flushed_at_once_Edmonton_water_gold_medal_hockey_game/



Good leads for inspiration ....
(visualization, simple stats)

http://chartsnthings.tumblr.com
http://blog.blprnt.com/
http://addictedtor.free.fr/graphiques/
http://flowingdata.com/
http://www.informationisbeautiful.net/
http://www-958.ibm.com/software/data/cognos/manyeyes/
http://blog.okcupid.com/
http://chartporn.org/
http://junkcharts.typepad.com/junk_charts/

weak links in the chain: process, packaging and presentation

project organization / literate programming /
reproducible research

collaboration / open science

The Trifecta of Vexing Issues in Scientific Statistical Computing

version control / back up / archive

project organization / literate programming /
reproducible research

Sweave
knitr

collaboration / open science

github
Rforge
sourceforge

The Trifecta of Vexing Issues in Scientific Statistical Computing

git
subversion
mercurial

version control / back up / archive

project organization / literate programming /
reproducible research

Sweave
knitr

What the cool kids seem to be doing

collaboration / open science

github
Rforge
sourceforge

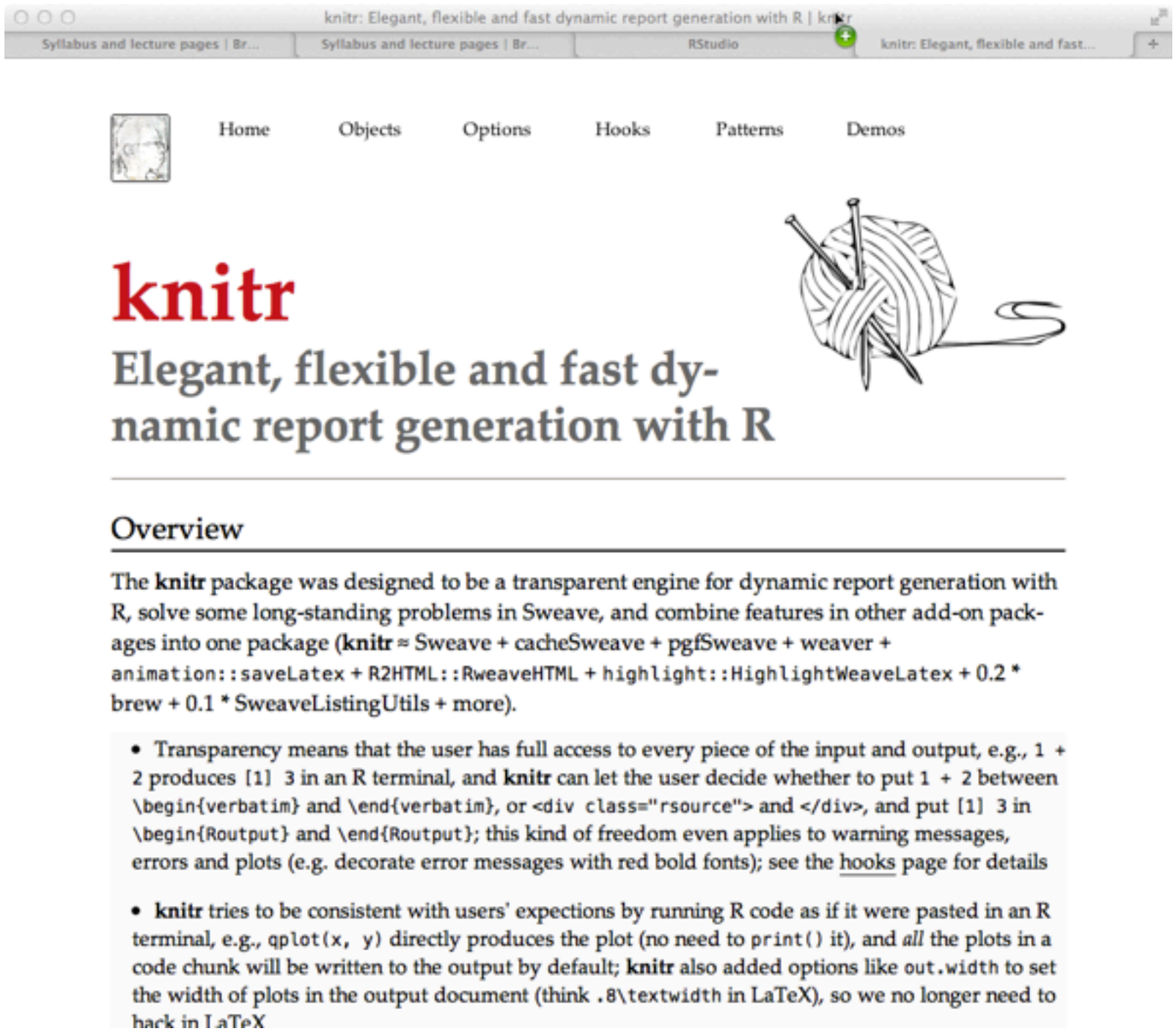
The Trifecta of Vexing Issues in Scientific Statistical Computing

git
subversion
mercurial

version control / back up / archive

project organization / literate programming / reproducible research

knitr



Getting Started with R Markdown, knitr, and Rstudio 0.96

Syllabus and lecture pages | Br... | Syllabus and lecture pages | Br... | RStudio | Getting Started with R Markdo... | +

Jeromy Anglim's Blog: Psychology and Statistics

THURSDAY, MAY 17, 2012

Getting Started with R Markdown, knitr, and Rstudio 0.96

This post examines the features of R Markdown using knitr in Rstudio 0.96. This combination of tools provides an exciting improvement in usability for reproducible analysis. Specifically, this post (1) discusses getting started with R Markdown and `knitr` in Rstudio 0.96; (2) provides a basic example of producing console output and plots using R Markdown; (3) highlights several code chunk options such as caching and controlling how input and output is displayed; (4) demonstrates use of standard Markdown notation as well as the extended features of formulas and tables; and (5) discusses the implications of R Markdown. This post was produced with R Markdown. The source code is available here as a gist. The post may be most useful if the source code and displayed post are viewed side by side. In some instances, I include a copy of the R Markdown in the displayed HTML, but most of the time I assume you are reading the source and post side by side.

Getting started

To work with R Markdown, if necessary:

- Install R
- Install the lastest version of RStudio (at time of posting, this is 0.96)
- Install the latest version of the `knitr` package:

```
install.packages("knitr")
```

To run the basic working example that produced this blog post:

- Open R Studio, and go to File - New - R Markdown
- If necessary install `ggplot2` and `lattice` packages:

```
install.packages("ggplot2"); install.packages("lattice")
```

- Paste in the contents of the gist (which contains the R Markdown file

I'm an academic bridging I/O psychology and statistics. My blog contains 100+ posts focused on data analysis in the social sciences. If you're new, check out the Site Map. If you love R, check out the 40+ posts on R. If you want to follow the blog, see the RSS and email subscription options.

- Overview of Blog Content
- Academic Publications
- Teaching Resources
- Google+ Profile: JeromyAnglim
- Twitter @JeromyAnglim
- GitHub @JeromyAnglim
- Follow on

Search This Blog

Search

FeedBurner FeedCount

901 readers

Subscribe via RSS

Posts

Comments

Categories

ability academia APAStyle Article Deconstruction Australia basic analyses bayesian Beamer BibTeX binary variable blogging book review bootstrapping calculus causation CFA cluster analysis computers correlation data mining data sharing descriptive statistics design difference scores discriminant function analysis dyads Eclipse Endnote Excel experiments factor analysis focus groups formatting fun GEE general advice ggplot2 I/O Psych Inquisit internet interviews introduction JabRef

http://jeromyanglim.blogspot.ca/2012/05/getting-started-with-r-markdown-knitr.html

http://rstudio.org/docs/authoring/using_markdown

Syllabus and lecture pages | Br... Syllabus and lecture pages | Br... RStudio RStudio

Home Screenshots Download Docs Support Development Blog

Using R Markdown with RStudio

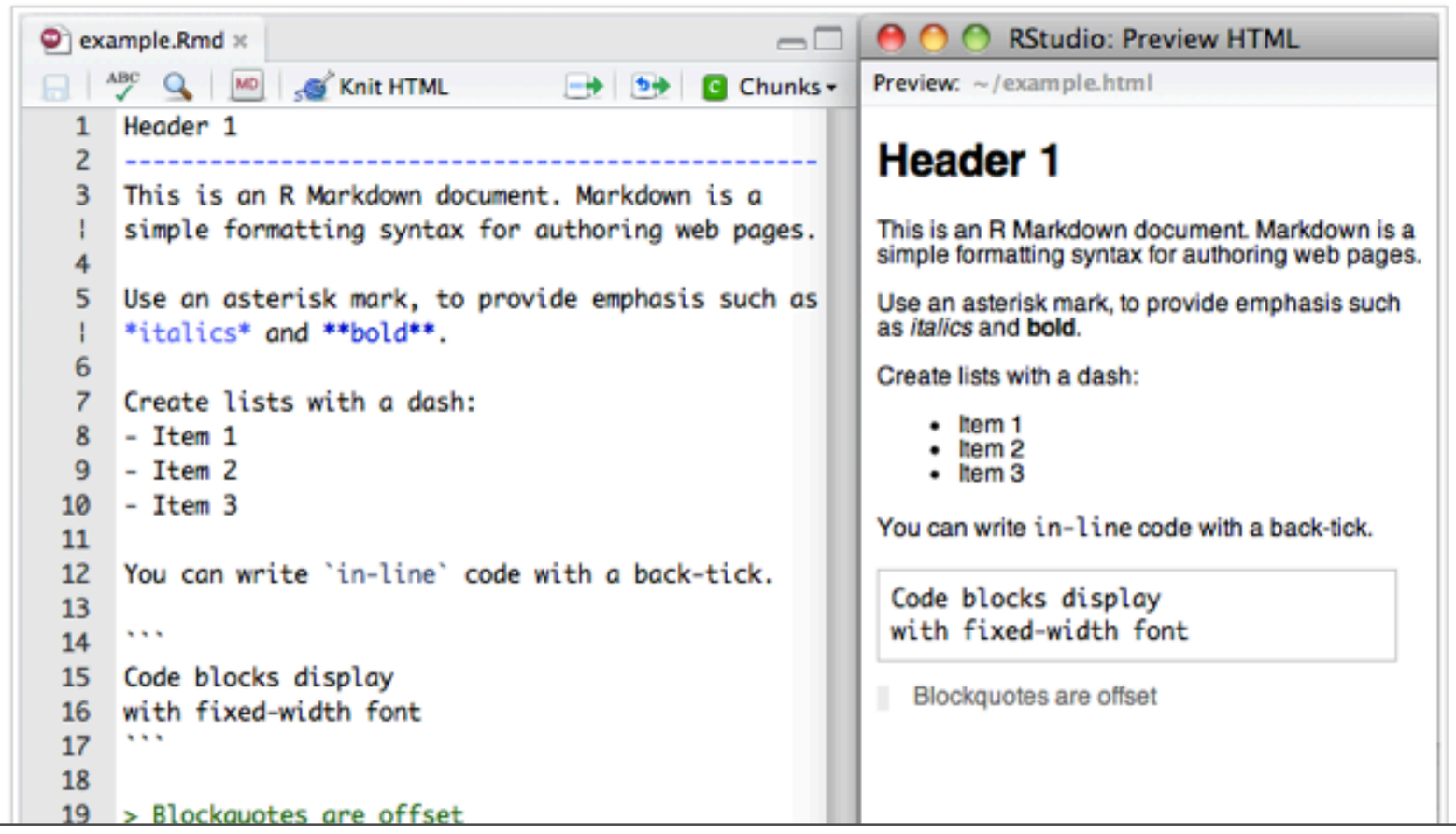
R Markdown enables easy authoring of reproducible web reports from R. It offers:

- Easy creation of web reports from R that can be automatically regenerated whenever underlying code or data changes.
- A highly accessible syntax (markdown) which lower the barriers to entry for reproducible research.
- Output of a standalone HTML file (with images embedded directly in the file) that is easy to share using email, Dropbox, or by deploying to a web server.
- Support for publishing dynamic and interactive web content.

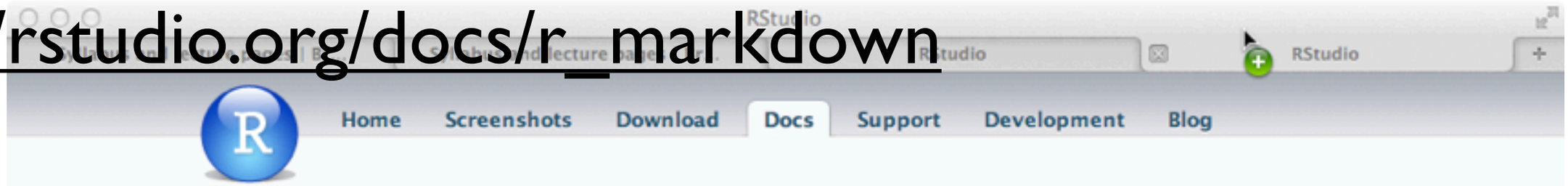
This article includes an overview of how to use R Markdown within RStudio. For more specific details on syntax and implementation, see the R Markdown article.

Markdown Basics

Markdown is a simple markup language designed to make authoring web content easy for everyone. Rather than writing HTML and CSS code, Markdown enables the use of a syntax much more like plain-text email. For example the file on the left shows basic Markdown and the resulting output as an HTML file on the right:



http://rstudio.org/docs/r_markdown



R Markdown

Overview

R Markdown is a format that enables easy authoring of reproducible web reports from R. It combines the core syntax of Markdown (an easy-to-write plain text format for web content) with embedded R code chunks that are run so their output can be included in the final document.

The concept of R Markdown is similar to Sweave (a system built-in to R for combining R with LaTeX). In Sweave, Rnw files are "weaved" into TeX files that are then compiled into PDFs. In R Markdown, Rmd files are similarly "weaved" into plain markdown (.md) files that are then compiled into HTML documents.

The R Markdown syntax is described in detail below. At a high-level it is a combination of the following:

- The core Markdown syntax
- The ability to embed R code and the results of its execution
- Additional extensions provided by the Sundown library
- Support for LaTeX and MathML equations
- Bundling of images within generated HTML files

Implementation

The implementation of R Markdown is provided by two packages:

- **knitr** — Weaves Rmd files into plain markdown (.md) files
- **markdown** — Converts markdown files into HTML documents

For example, to run the R code chunks inside an Rmd file and then convert the resulting markdown file into HTML you would execute the following:

```
library(knitr)
library(markdown)
knit("Foo.Rmd")
markdownToHTML("Foo.md", "Foo.html")
```

Note that this can also be accomplished in one step by calling `knitr::knit2html`. However, the fact that converting from Rmd to HTML is broken into two steps allows for the use of alternate markdown rendering programs (e.g. Pandoc).

RStudio also implements support for rendering R Markdown files into HTML using the **Knit HTML** commands. This produces a result equivalent to the above code.

Syntax

The following is a detailed description of the R Markdown syntax. Note that the first two sections covering *Core Markdown* and *R Code Chunks* are always applicable. The remaining sections apply to the rendering of markdown to HTML as implemented by

yihui/knitr

Syllabus and lecture pages | Br... | Syllabus and lecture pages | Br... | RStudi | yihui/knitr

github | Search or Type a Command | Explore Gist Blog Help | jennybc

PUBLIC yihui / knitr

Code | Network

A general-purpose tool for dynamic report ge

http://yihui.name/knitr

Clone in M

branch: mas

Latest commit to t

bump version

yihui authored

knitr /

name
R
inst
man
tests
.Rbuildignore
.Rinstignore
.gitignore

hadley/ggplot2

Syllabus and lecture pages | Br... | Syllabus and lecture pages | Br... | RStudio

github | Search or Type a Command | Explore Gist Blog Help

PUBLIC hadley / ggplot2

Wa

Code | Network | Pull Requests 11 | Issues 114

hadley/plyr

GitHub, Inc. github.com/hadley/plyr

Syllabus and lecture pages | Br... | Syllabus and lecture pages | Br... | RStudio | hadley

github | Search or Type a Command | Explore Gist Blog Help | jennybc

PUBLIC hadley / plyr

Watch | Star 139

Code | Network | Pull Requests 15 | Issues 54

A R package for splitting, applying and combining large problems into simpler problems — Read more

http://plyr.had.co.nz

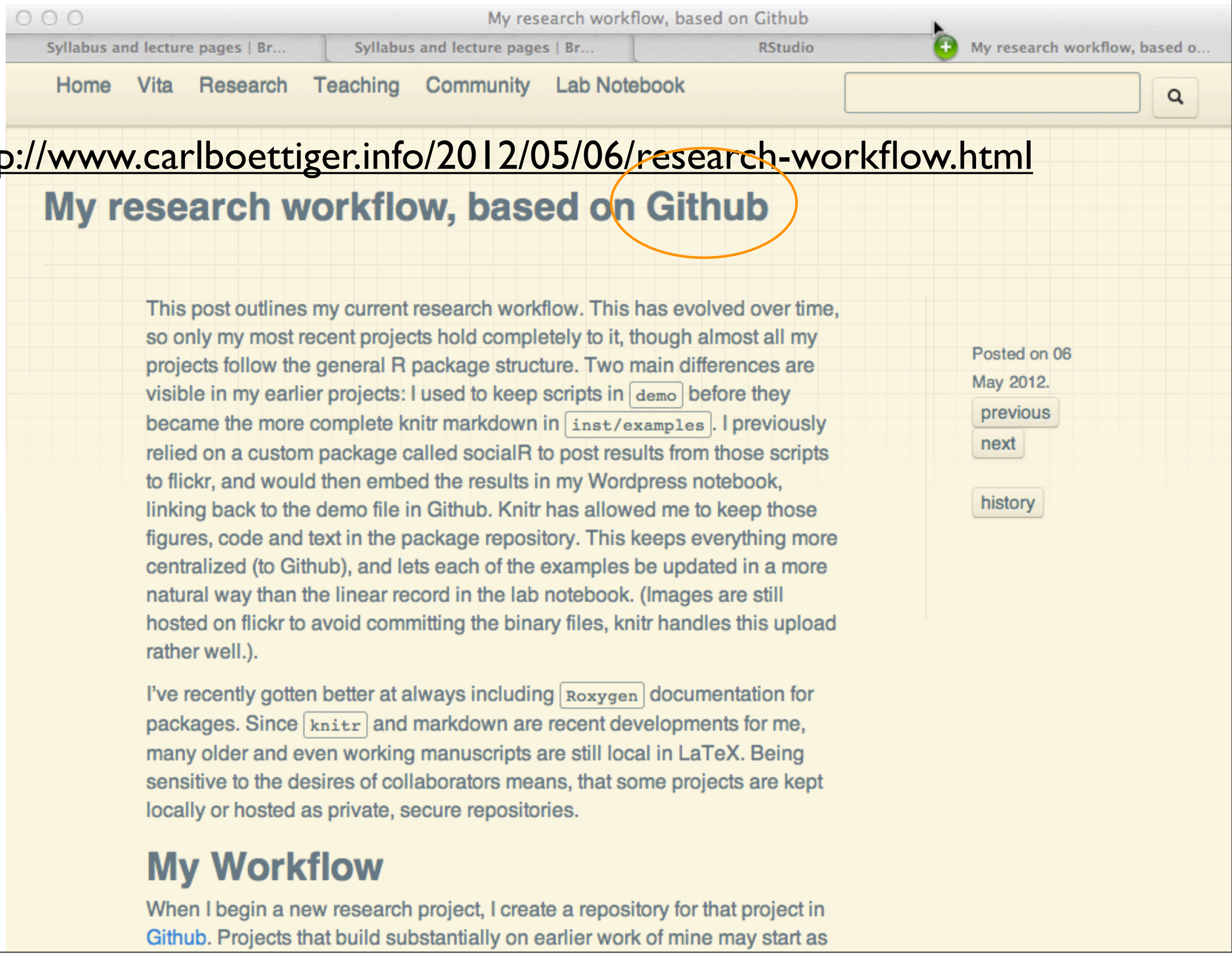
Clone in Mac | ZIP | HTTP | Git Read-Only | https://github.com/hadley/plyr.git | Read-Only acce

branch: master | Files | Commits | Branches 8 | Tags 20

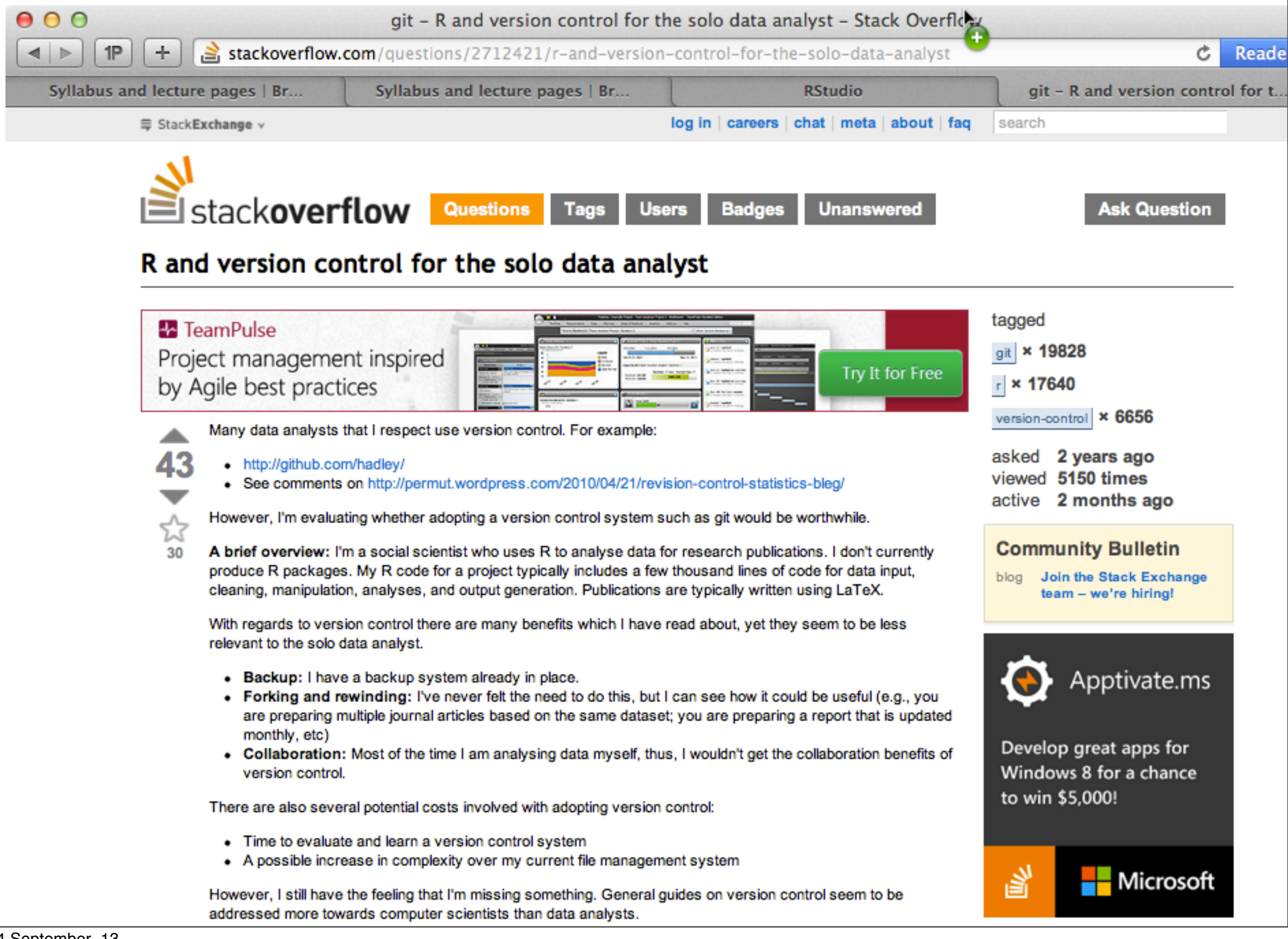
Latest commit to the master branch

Version bump

hadley authored 9 months ago com


My research workflow, based on Github
Syllabus and lecture pages | Br...
Syllabus and lecture pages | Br...
RStudio
My research workflow, based o...
Home
Vita
Research
Teaching
Community
Lab Notebook
http://www.carlboettiger.info/2012/05/06/research-workflow.html
My research workflow, based on Github
This post outlines my current research workflow. This has evolved over time, so only my most recent projects hold completely to it, though almost all my projects follow the general R package structure. Two main differences are visible in my earlier projects: I used to keep scripts in demo before they became the more complete knitr markdown in inst/examples. I previously relied on a custom package called socialR to post results from those scripts to flickr, and would then embed the results in my Wordpress notebook, linking back to the demo file in Github. Knitr has allowed me to keep those figures, code and text in the package repository. This keeps everything more centralized (to Github), and lets each of the examples be updated in a more natural way than the linear record in the lab notebook. (Images are still hosted on flickr to avoid committing the binary files, knitr handles this upload rather well.).
I've recently gotten better at always including Roxygen documentation for packages. Since knitr and markdown are recent developments for me, many older and even working manuscripts are still local in LaTeX. Being sensitive to the desires of collaborators means, that some projects are kept locally or hosted as private, secure repositories.
Posted on 06 May 2012.
previous
next
history
My Workflow
When I begin a new research project, I create a repository for that project in Github. Projects that build substantially on earlier work of mine may start as

http://stackoverflow.com/questions/2712421/r-and-version-control-for-the-solo-data-analyst



http://support.rstudio.org/help/kb/advanced/using-version-control-with-rstudio

RStudio

rstudio.org/docs/version_control/overview

Reader

Syllabus and lecture pages | Bryan Lab | Syllabus and lecture pages | Bryan Lab | RStudio

Home Screenshots Download Docs Support Development Blog

Using Version Control with RStudio

Overview

Version control is an indispensable tool for coordinating the work of teams and also has many benefits for individual work. The following StackOverflow discussions describe some of these benefits:

- Why should I use version control?
- R and version control for the solo data analyst

RStudio includes integrated support for two open source version control systems:

- Git
- Subversion

To use version control with RStudio, you should first ensure that you have installed Git and/or Subversion (details below). Next, you should become familiar with using RStudio Projects (which are required for version control features to be enabled).

Installation

Prior to using RStudio's version control features you will need to ensure that you have Git and/or Subversion installed on your system. The following describes how to do this for various platforms.

Git

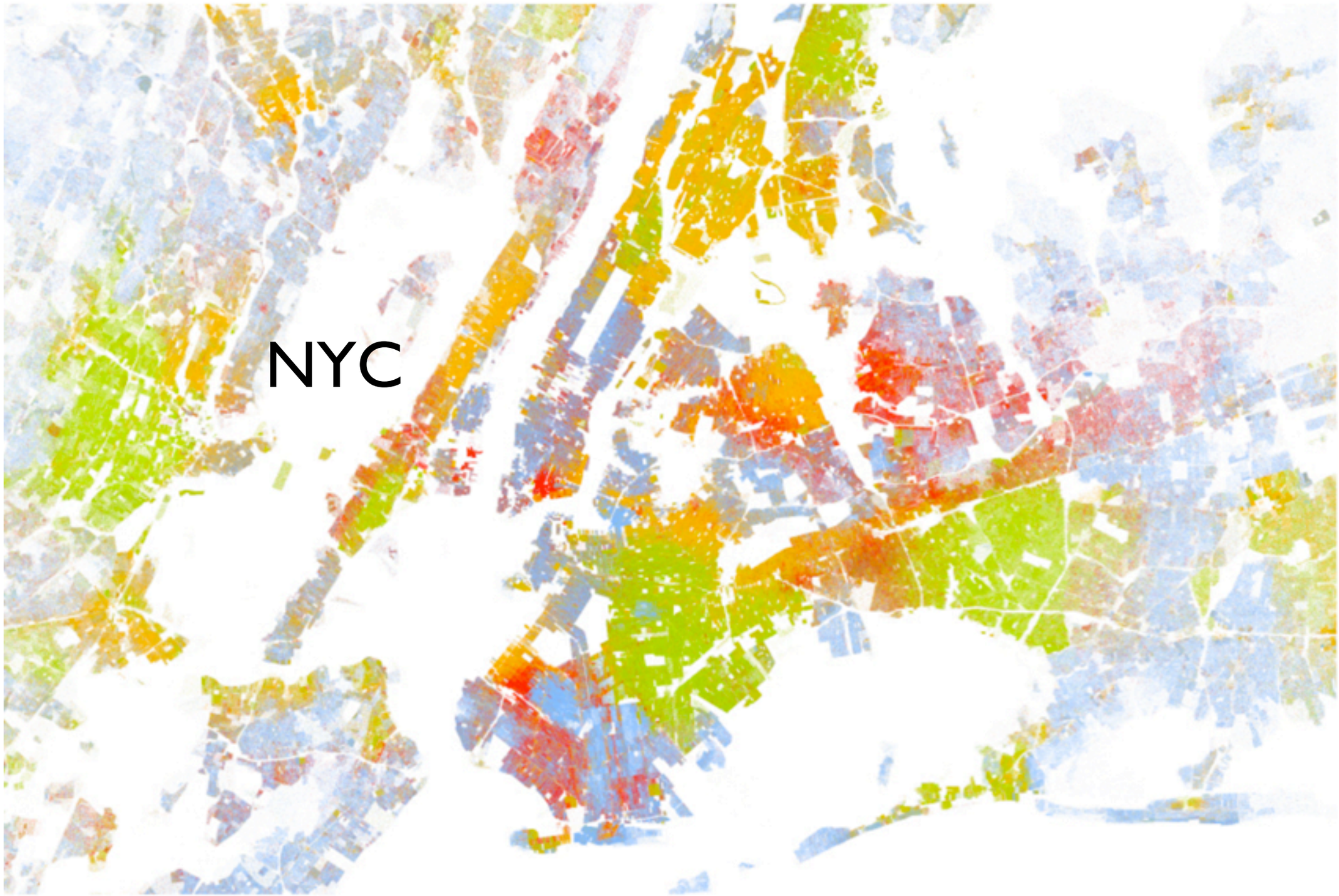
Prior to using Git with RStudio you should install it using the appropriate method for your platform:

- **Windows:** http://code.google.com/p/msysgit/
- **OSX:** http://code.google.com/p/git-osx-installer/
- **Debian/Ubuntu:** `sudo apt-get install git-core`
- **Fedora/RedHat:** `sudo yum install git-core`

An excellent resource for learning more about Git and how to use it is the Pro Git online book. Another good resource for learning about git is the Introduction to Git provided by GitHub.

The Best Map Ever Made of America's Racial Segregation

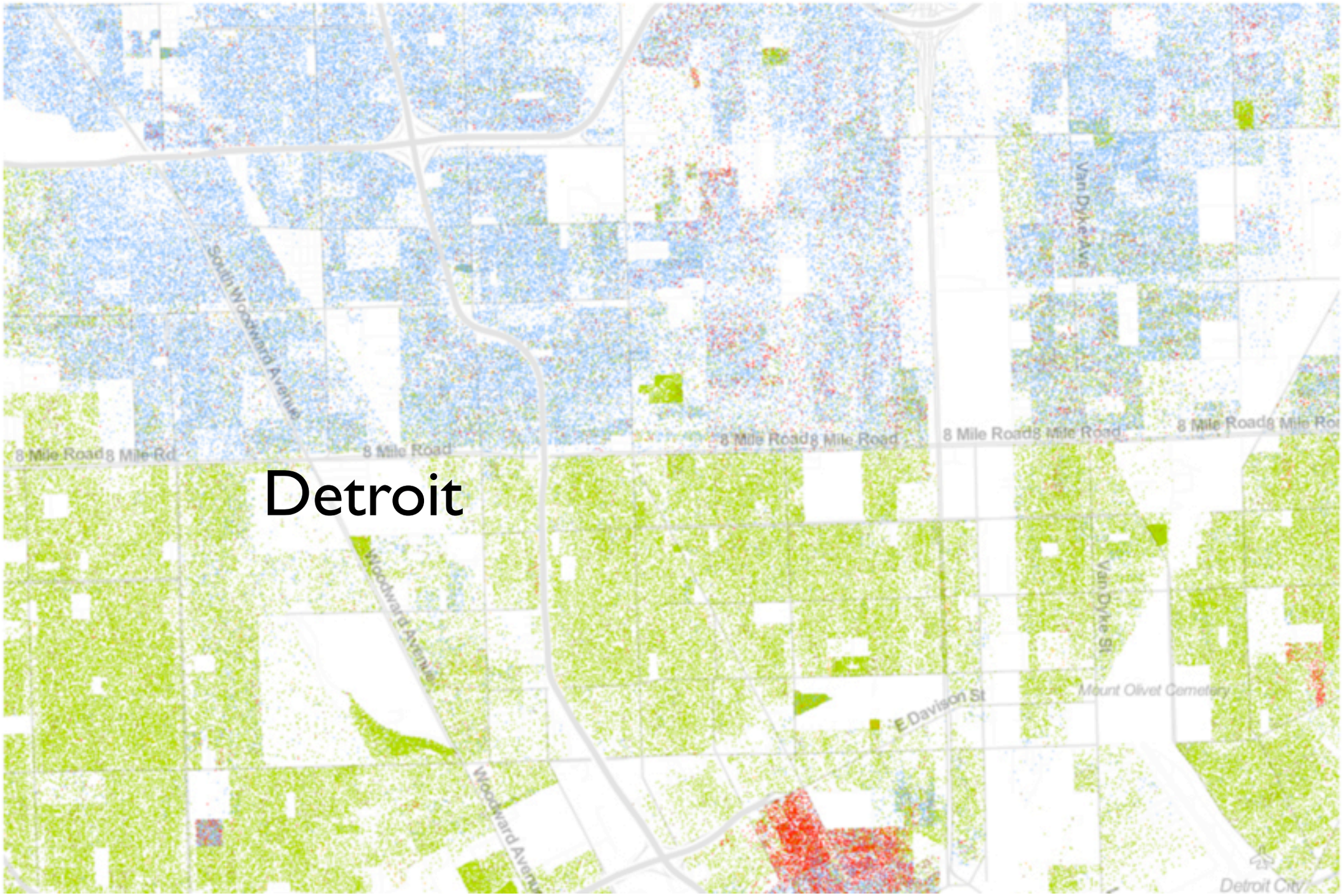
http://www.wired.com/design/2013/08/how-segregated-is-your-city-this-eye-opening-map-shows-you/?viewall=true



This map, created by Dustin Cable at University of Virginia's Weldon Cooper Center for Public Service, is the most comprehensive representation of racial distribution in America ever made. Here: New York City. *Image: Dustin Cable* **White: blue dots; African American: green dots; Asian: red; Latino: orange; all others: brown**

The Best Map Ever Made of America's Racial Segregation

http://www.wired.com/design/2013/08/how-segregated-is-your-city-this-eye-opening-map-shows-you/?viewall=true

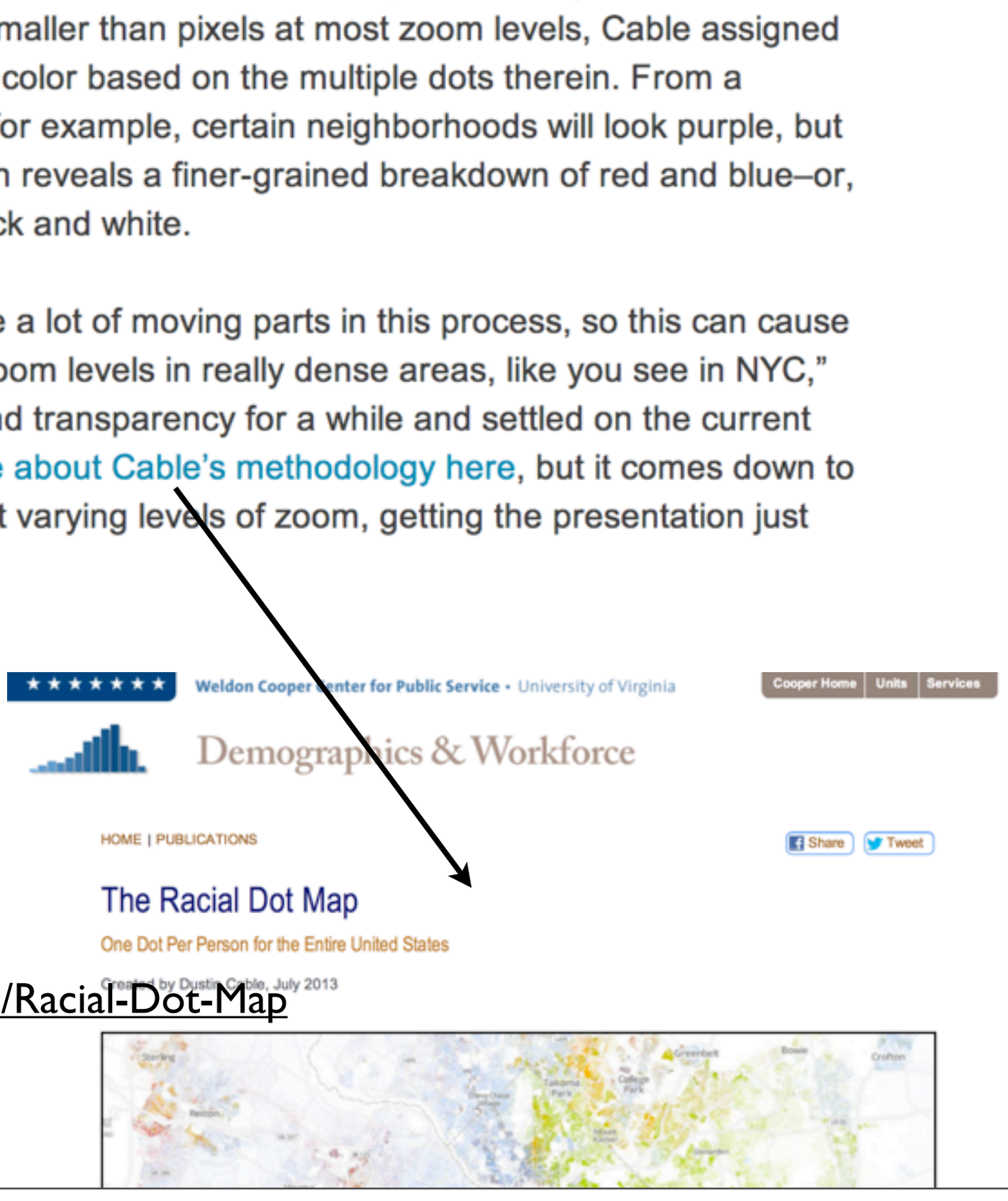


In Detroit, among the most segregated cities in America, 8 Mile Road serves as a sharp dividing line. *Image: Dustin Cable* **White: blue dots; African American: green dots; Asian: red; Latino: orange; all others: brown**

This is the most comprehensive map of race in America ever created.

White people are shown with blue dots; African-Americans with green; Asians with red; and Latinos with orange, with all other race categories from the Census represented by brown. Since the dots are smaller than pixels at most zoom levels, Cable assigned shades of color based on the multiple dots therein. From a distance, for example, certain neighborhoods will look purple, but zooming-in reveals a finer-grained breakdown of red and blue–or, really, black and white.

"There are a lot of moving parts in this process, so this can cause different shades of color to appear at different zoom levels in really dense areas, like you see in NYC," Cable explains. "I played around with dot size and transparency for a while and settled on the current scheme as being adequate." You can read more about Cable's methodology here, but it comes down to this: When you're dealing with 300 million dots at varying levels of zoom, getting the presentation just right is as much an art as a science.



http://www.coopercenter.org/demographics/Racial-Dot-Map

Weldon Cooper Center for Public Service • University of Virginia

Cooper Home | Units | Services

Demographics & Workforce

HOME | PUBLICATIONS

Share Tweet

The Racial Dot Map

One Dot Per Person for the Entire United States

Created by Dustin Cable, July 2013

Cool result is accompanied by explanation of how it was done

Methodology

Python was used to read the 50 state shapefiles (with the merged SF1 data). The GDAL and Shapely libraries were used to read the data and create the point objects. The code retrieves the population data for each census block, creates the appropriate number of geographic points randomly distributed within each census block, and outputs the point information to a database file. The resulting file has x-y coordinates for each point, a quadkey reference to the Google Maps tile system, and a categorical variable for race. The final database file has 308,745,538 observations and is about 21 GB in size. The processing time was about five hours for the entire nation.

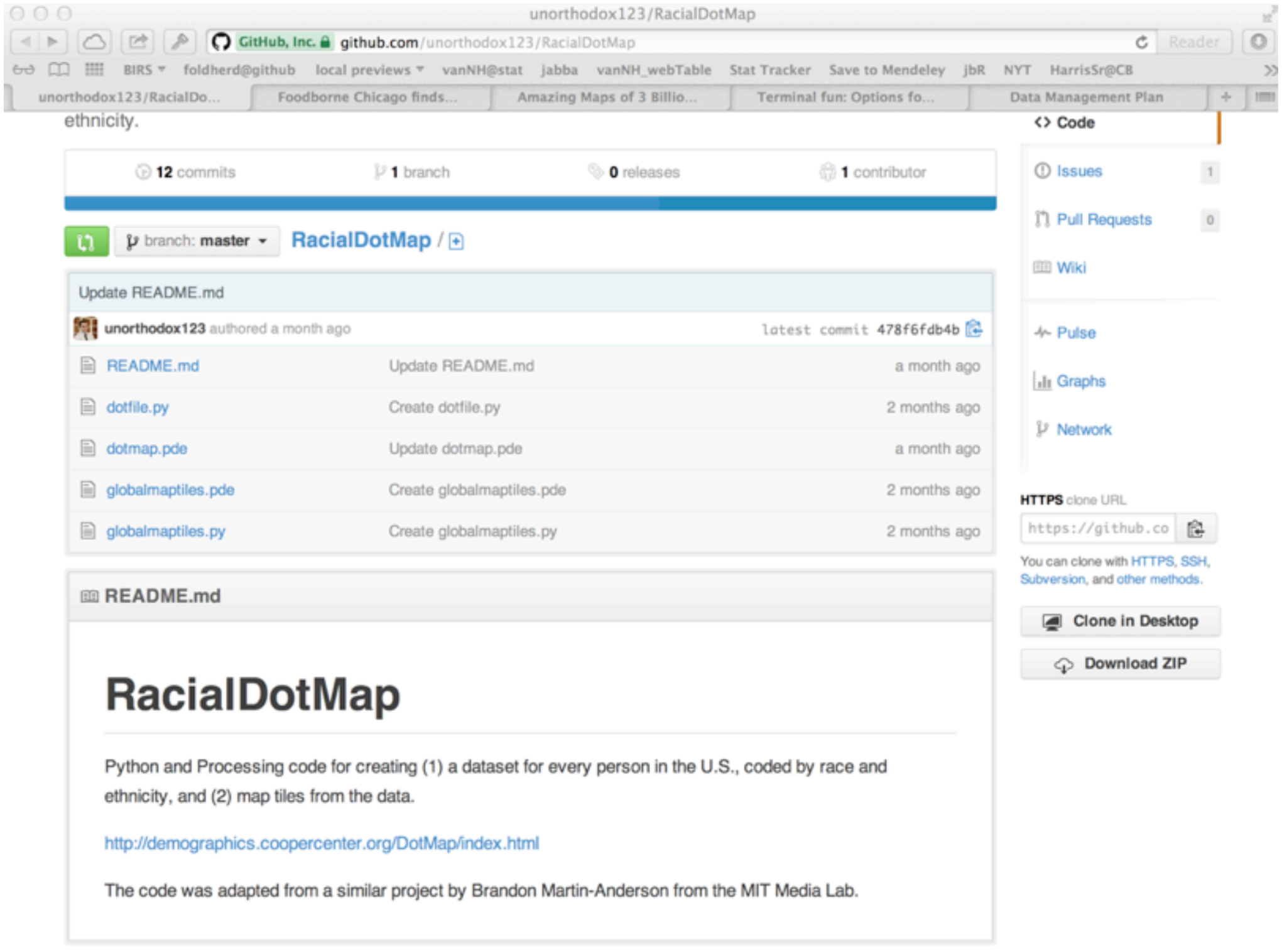
The database file was then sorted by quadkey and converted to a .csv format. SAS was able to do this within an hour without crashing.

Processing 2.0.1 for 64-bit Windows was used to create the map tiles. The Java code reads each point from the .csv file and plots a dot on a 512x512 .png map tile using the quadkey reference and x-y coordinates. The racial categorical variable is used to color-code each plotted dot. This process used the default JAVA2D renderer, but other platforms may work better using P2D. Map tiles were created for Google Maps' zoom levels 4 through 13 to make the final map. A non-color-coded map was also produced to help add more contrast for lightly populated areas. In total, the color-coded and non-color-coded maps contain 1.2 million .png files totaling about 7 GB. Producing all of the map tiles in Processing took about 16 hours for the two maps.

The Google Maps API is used to display the map tiles. Map tiles with zero population are never created using the above method. Therefore, an index was used to tell the map application whether a tile exists in order to prevent 404 errors.

The entire code is up on GitHub and was adapted from code developed by Brandon Martin-Anderson and Peter Richardson in order to account for the racial coding and errors in reading the shapefiles.

https://github.com/unorthodox123/RacialDotMap



August 16, 2013

Foodborne Chicago finds dodgy restaurants with tweets, and R

If, like me, you've ever had a sandwich from a dubious deli and then been laid up for days afterwards, you know that food poisoning is no trifling matter. In the past, local authorities would only ever learn of such public health issues if they get reported to the authorities by the victim (or the victim's doctor). But that misses the many cases of less serious illnesses that don't involve a doctor or hospital, or illnesses that simply aren't reported to the authorities.

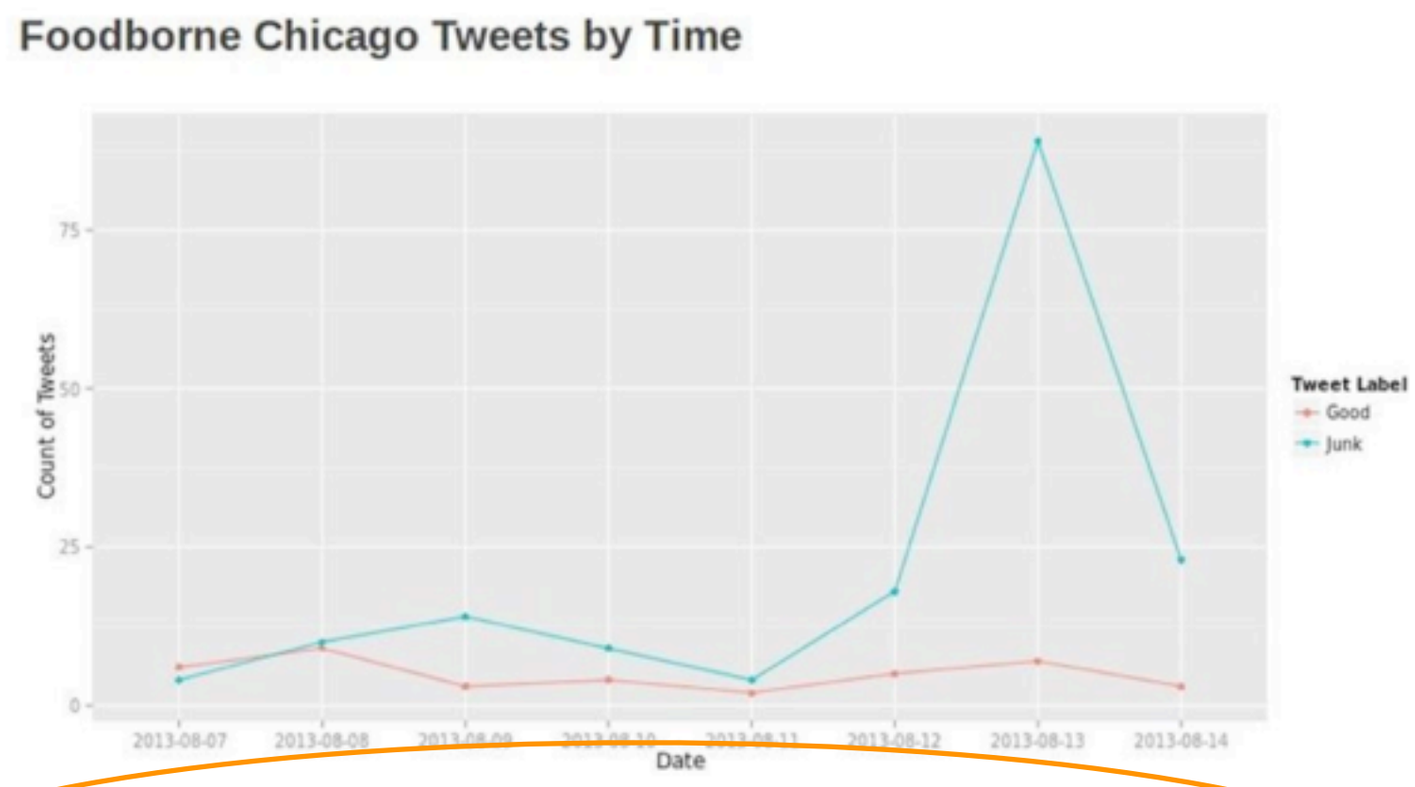
Now, the City of Chicago has found a new way of identifying sources of food poisoning: by analyzing tweets. Foodborne Chicago scans tweets posted in the Chicagoland area, responding to tweets like: *"Stomach flu/food poisoning is like eating gas station sushi without the joys of eating gas station sushi"* (but ignoring tweets like *"It's really hard to snack while watching Honey Boo Boo. It's the second best diet to food poisoning."*). If you send a such a tweet, you're likely to get a response:

http://blog.revolutionanalytics.com/2013/08/foodborne-chicago.html

The system is entirely automated, and uses real-time text analysis implemented with R language to identify those tweets that are about a specific case of food poisoning:

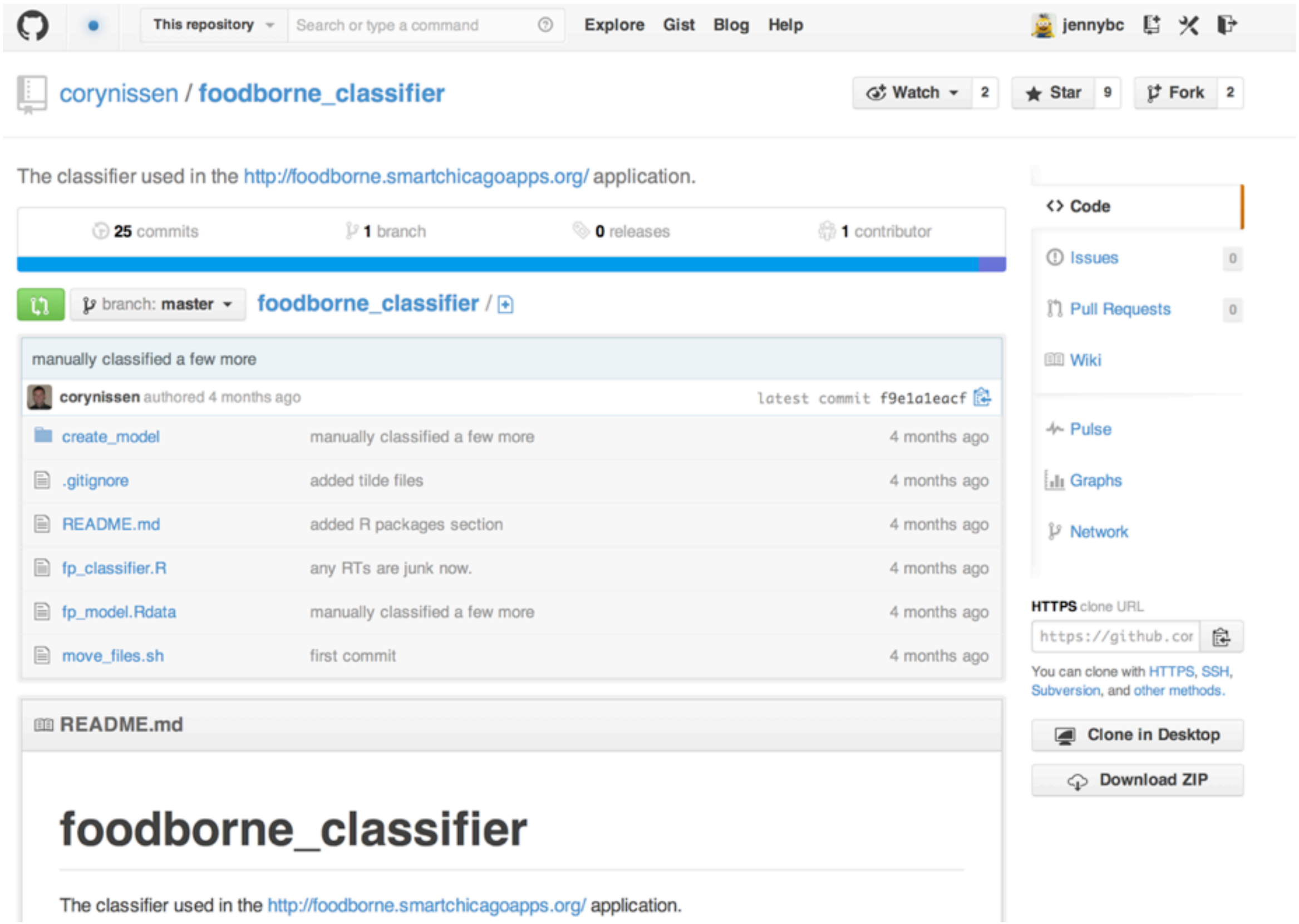
> Foodborne searches Twitter for all tweets near Chicago containing the string "food poisoning". The ingestion service consumes thousands of tweets, storing them in a large MongoDB instance. A collection of classification servers, running R, churn through the collected tweets, applying a series of filters. The tweets are classified using a model that was trained via supervised learning, which determines if the tweets are related to a food poisoning illness or not.

Cory Nissen, the data scientist who implemented the analysis behind the system, shared some of the behind-the-scenes details with me via email. He used an R package called textcat and an algorithm based on n-grams to classify the tweets. The model is trained in such a way as to bias towards sensitivity (at the 90%+ level) at the expense of specificity (50 - 60%) to better sort true food poisoning reports from "junk" tweets merely *about* food poisoning. Out of all the tweets in the Chigaco area on any given day, the system flags about 10-20 tweets a day for review, of which just a couple will typically warrant a response to the unwell citizen for followup.



The open-source R code behind the classifier is available on Github. Check out the README file for more technical details behind the implementation. You can also see how the application was presented on Fox 39 Chicago news (starting at the 2:09 mark):

https://github.com/corynissen/foodborne_classifier



http://www.wired.com/design/2013/07/design_06272013_tweetmaps/

course stuff

Culture of the class

- Teaching you to fish (vs. giving you a fish)
  - It's amazing what a determined individual can learn from documentation, small learning examples, and ... <gasp> Googling. And also stackoverflow.
- Rewarding engagement, intellectual generosity and curiosity
  - Speaking up, sharing success OR failure, showing some interest in something will earn marks.
- Zero tolerance of plagiarism
  - Generating your own approach, writing some code, and describing the process is the whole point. Process is generally more important than product.

Where marks will come from

- Small ~weekly activities; marked coarsely (think check, check minus, check plus), sometimes with peer evaluation
- Final concluding activity will be bigger, working in very loose groups organized around a few datasets
  - Think about datasets you'd like to prepare and analyze!
- Departure from past model where individual final project was almost entire course mark. I would love to decrease the back-loaded heroic efforts by students and instructor alike. Plus I think you'll learn more this way.

Excerpt from marking rubric of main assignment last year

Telling the Story	**Outstanding (A+)**	**Very good (A)**	**Adequate (A-)**	**Needs work (B)**	**Inadequate (≤ C)**
The plot "reveal the most important, most interesting aspects of the dataset" "Use the text to interpret and highlight what the figures show" "a 1-2 page guided tour through the figures"	Account is enjoyable to read and is complete but avoids unnecessary detail. Well organized -- probably with explicit use of sections. Each point / concept / figure follows logically from the previous. The figures arise as the natural support for the story and are appropriately referenced, described, and interpreted.	Close to A+, but lacking in one or two key aspects.	Overall organization, flow, integration w/ figures is adequate but there is at least one noticeable 'negative': • obvious unanswered question • major piece of information missing • creates doubt/ confusion in reader • appears to contradict itself	Substantial problems with organization, flow, completeness. Unclear how reader should transfer attention between prose and figures. Reader is forced to decode the figures -- what they show, why they are interesting / relevant, but it's possible. Requires reader to work hard, which is frustraing.	Reader can't really make sense of the work. Organization is weak or absent. Major points / concepts/ figures hard to identify. Even with considerable effort, reader can't understand the story, which is maddening.

Excerpt from marking rubric of main assignment last year

The whole package	**Outstanding (A+)**	**Very good (A)**	**Adequate (A-)**	**Needs work (B)**	**Inadequate (≤ C)**
Curiosity, skepticism, self-reflection "Give an account of the process and reflect on what was most successful, what was most disappointing."	Clear that student created different visualizations, tried different approaches. Final result comes from editing down, curating. Interesting ideas for further work or observations. Describes some lessons learned.		Modest effort to explore multiple solutions, carry out critical analysis, and identify next steps or issues. Some rather obvious or natural next steps or observations are left unmentioned, unexplored.	Student has done the bare minimum. Report barely goes beyond a basic factual description. Student let something rather simple hamper them.	Report does not contain any relevant observations, ideas for improvement, etc. Serious lack of time and/or effort is obvious.
Achievement, mastery, cleverness, creativity	Student has gone beyond what was expected and required, e.g., extraordinary effort, additional tools not addressed by this course, unusually sophisticated application of tools from course.	Tools and techniques from the course are applied very fruitfully and somewhat creatively.	Competent use of tools and techniques covered in the course. Chosen task was acceptable, but fairly conservative in ambition.	Student does not display the expected level of mastery of the tools and techniques in this course. Chosen task was too limited in scope.	Work is trivial, in scope or in implementation or both. Work demonstrates incompetence.

Excerpt from marking rubric of main assignment last year

Figures	**Outstanding (A+)**	**Very good (A)**	**Adequate (A-)**	**Needs work (B)**	**Inadequate (≤ C)**
Effective? two main goals in the course: Facilitate comparisons. Identify trends. Other worthy goals,depending on the dataset: Engage a broad audience, i.e. not just other “specialists” Demystify a complicated concept or result Illustrate a paradox or a gap between perception and reality Enable other humans to digest a very large amount of information	Graphs carefully tuned for desired purpose. Evidence that explicit effort was made to fulfill 3 or more of listed criteria. Several figure types used.	Graphs wells chosen, but have modest problems, such as inappropriate aspect ratios, poor labels, poor quality when viewed/printed. Fulfills some of the criteria, but more were within reach. More than one figure type.	Graphs fairly appropriate, but several minor problems. Fulfills only 1 or 2 of criteria. More than one figure type.	Graphs poorly chosen to support purpose. Some fundamental flaws. Seems like criteria were not explicitly considered.	Graphs do not support the purpose. Major presentation problems.

Excerpt from marking rubric of main assignment last year

Code	Outstanding (A+)	Very good (A)	Adequate (A-)	Needs work (B)	Inadequate (≤ C)
Readability Reusability (achieved through comments, informative names, transparent code, etc.)	It is extremely easy to read the code and determine what's happening, why, and how.	Close to A+, but there are a couple instances where it's hard to determine what's going on.	In broad strokes, the code is readable, but at low- to medium-level of detail, it's difficult to decipher in many places.	I have serious concerns whether the code does what it is intended to do.	Code is unreadable. I would have to run it and inspect objects and output to determine how / if it works.
micro-level: principled approach to formatting (e.g. indenting, spacing, line breaking)	Universal use of a reasonable formatting scheme -- almost certainly due to use of a smart editor.	Close to A+, but there's one or two choices that are regarded as 'bad' by the pros.	Some effort to format is detectable, but it's not uniformly applied and/or has some serious shortcomings.	Little effort to format the code.	Code formatting?
macro-level sound practices (e.g. avoiding Magic Numbers, replacing repetitive code w/ function, reference by name not number)	At every possible juncture, code uses elegant, robust practices.	Close to A+, but corners were cut here and there.	Several instances -- or perhaps general use -- of an unsound practice that will seriously impact code's robustness / reusability.	Frequent use of unsound practices, suggesting student is not aware of or trying to follow sound practices.	Code is actually broken.

Who am I?

- Associate Professor jointly appointed 50/50 in Statistics and the Michael Smith Laboratories
- Specialize in development and application of statistical methods for high-throughput, genome-scale data
- High-throughput phenotyping, formulations of cluster analysis based on graphs, R and data analysis
- PhD in 2001 in biostatistics from UC Berkeley; undergraduate in Econ/German (?!) at Yale
- Teach this course STAT 545A, STAT 540 Statistical methods for high dimensional biology and some undergrad courses (STAT 100 and 302)

Who was here in 2012

	MASC	MSC	PHD	UNCL	VISI	Sum
NA	0	0	0	1	0	1
Biochemistry & Molecular Biol	0	0	1	0	0	1
Bioinformatics	0	1	0	0	0	1
Genome Science & Technology	0	2	0	0	0	2
Kinesiology	0	0	1	0	0	1
Mathematics	0	1	0	0	0	1
Mining	1	0	0	0	0	1
Population and Public Health	0	1	0	0	0	1
Statistics	0	5	1	0	1	7
Sum	1	10	3	1	1	16

Who was here in 2011

	Master's	PHD	Sum
Forestry	1	0	1
Lib, Arch and Info Stud	0	1	1
Management Info Systems	1	0	1
Mathematics	1	0	1
Mechanical Engineering	1	1	2
Psychology	1	1	2
Resource Mgmt/ Envirn Stud	0	1	1
Statistics	14	1	15
Sum	19	5	24

Admin & communication

- Email: jenny@stat.ubc.ca
  - Please put STAT545A in subject.
  - Please be brief and consider if email is necessary, best way to handle. Speak after class? Use the Google group?
- Office is ESB 3116.
- TA is Song Cai scai@stat.ubc.ca
- Google group for the course:
  - https://groups.google.com/forum/#!forum/stat545a_2013
  - Request to join or just wait for us to invite you soon.

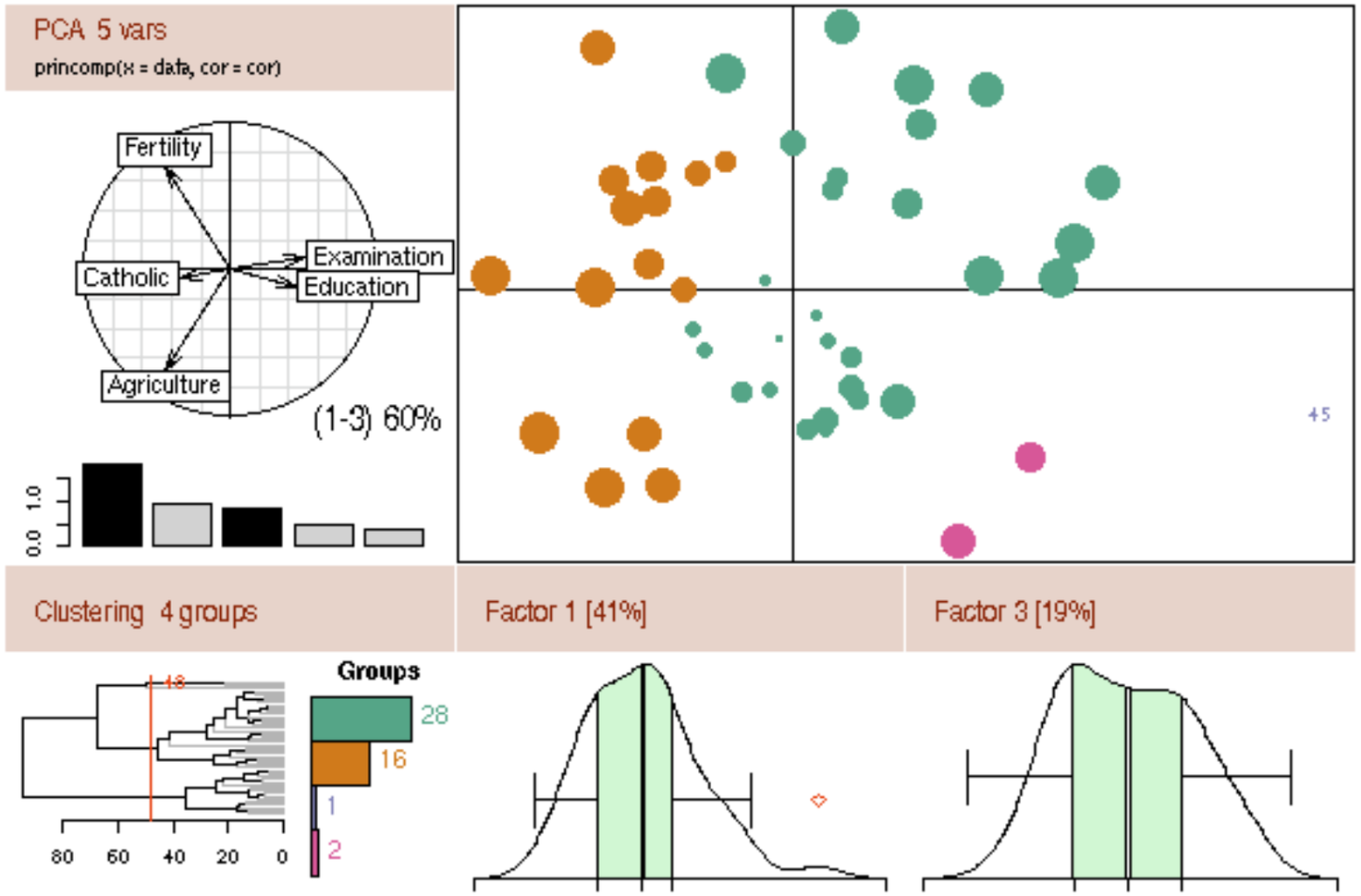
Admin & communication

- Course webpage:
  - http://www.stat.ubc.ca/~jenny/STAT545A/index.html
  - http://www.stat.ubc.ca/~jenny/STAT545A/current.html
  - The source of everything on the web will be available on github.
  - I would like to have student work and input also on this webpage or easily accessible via the webpage. How we will do that remains to be seen....
    - I suspect we'll use Rpubs.
    - Or, for the self-selecting few, maybe github.

Software

- We will use R, "a free software environment for statistical computing and graphics". The R Project.
- R is the most prevalent statistical computing environment for research in statistical methodology and is also widely used for data analysis and publication-quality graphics
- We will make heavy use of the `lattice` package for making figures; it is superior to base graphics in terms of our major goals = facilitating comparisons and revealing trends.
- Blind leading the blind: we are also going to learn `ggplot2`

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



HOME PAGE | TODAY'S PAPER | VIDEO | MOST POPULAR | TIMES TOPICS

The New York Times

Business Computing

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

Search Technology | Go

Inside Technology: Internet | Start-Ups | Business Computing | Companies

Bits Blog

Data Analysts Captivated by R's Power

Stuart Isett for The New York Times

R first appeared in 1996, when the statistics professors Robert Gentleman, left, and Ross Ihaka released the code as a free software package.

By ASHLEE VANCE
Published: January 6, 2009

TWITTER
E-MAIL
PRINT
REPRINTS
SHARE

To some people R is just the 18th letter of the alphabet. To others, it's the rating on racy movies, a measure of an attic's insulation or what pirates in movies say.

Related

Bits: R You Ready for R?

The R Project for Statistical Computing

R is also the name of a popular programming language used by a growing number of data analysts inside corporations and academia. It is becoming their lingua franca partly because data mining has entered a golden age, whether being used to set ad prices, find new drugs more quickly or fine-tune financial models. Companies as diverse as Google, Pfizer, Merck, Bank of America, the InterContinental Hotels Group and Shell use it.

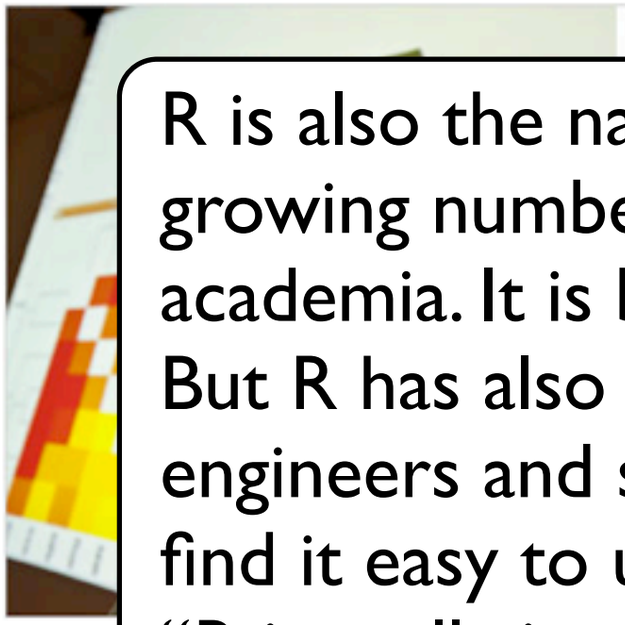
Data Analysts Captivated by R's Power in NYT January 6, 2009 by Ashlee Vance

R You Ready for R? NYT Bits blog post January 8, 2009 by Ashlee Vance

The New York Times

Business Computing

Data Analysts Captivated by R's Power



R first appe
software pa

By ASHLEE
Published:

To some
the ratin
pirates i

Related

Bits: R You Ready for R?

The R Project for Statistical Computing

growing number of data analysts inside corporations and academia. It is becoming their lingua franca partly because data mining has entered a golden age, whether being used to set ad prices, find new drugs more quickly or fine-tune financial models. Companies as diverse as Google, Pfizer, Merck, Bank of America, the InterContinental Hotels Group and Shell use it.

R is also the name of a popular programming language used by a growing number of data analysts inside corporations and academia. It is becoming their lingua franca partly ....
But R has also quickly found a following because statisticians, engineers and scientists without computer programming skills find it easy to use.
"R is really important to the point that it's hard to overvalue it," said Daryl Pregibon, a research scientist at Google, which uses the software widely. "It allows statisticians to do very intricate and complicated analyses without knowing the blood and guts of computing systems."

← La Jetee

Binary classification evaluation in R via ROCR →

Comparison of data analysis packages: R, Matlab, SciPy, Excel, SAS, SPSS, Stata

Posted on February 23, 2009

Lukas and I were trying to write a succinct comparison of the most popular packages that are typically used for data analysis. I think most people choose one based on what people around them use or what they learn in school, so I've found it hard to find comparative information. I'm posting the table here in hopes of useful comments.

Name	Advantages	Disadvantages	Open source?	Typical users
R	Library support; visualization	Steep learning curve	Yes	Finance; Statistics
Matlab	Elegant matrix support; visualization	Expensive; incomplete statistics support	No	Engineering
SciPy/NumPy/Matplotlib	Python (general-purpose programming language)	Immature	Yes	Engineering
Excel	Easy; visual; flexible	Large datasets	No	Business
SAS	Large datasets	Expensive; outdated programming language	No	Business; Government
Stata	Easy statistical analysis		No	Science
SPSS	Like Stata but more expensive and worse			

http://brenocon.com/blog/2009/02/comparison-of-data-analysis-packages-r-matlab-scipy-excel-sas-spss-stata/

Good leads for inspiration ....
(more of an R focus)

http://www.r-bloggers.com/
http://www.stattler.com/
http://learnr.wordpress.com/
http://www.drewconway.com/zia/
http://www.sigmafield.org/
http://onertipaday.blogspot.com/

Now for the hands on part ....

People variously did:

ran away!

verified they could actually log in to the lab's computers

in the lab or wherever, started working on the first two tutorial type activities posted under today on the course webpage