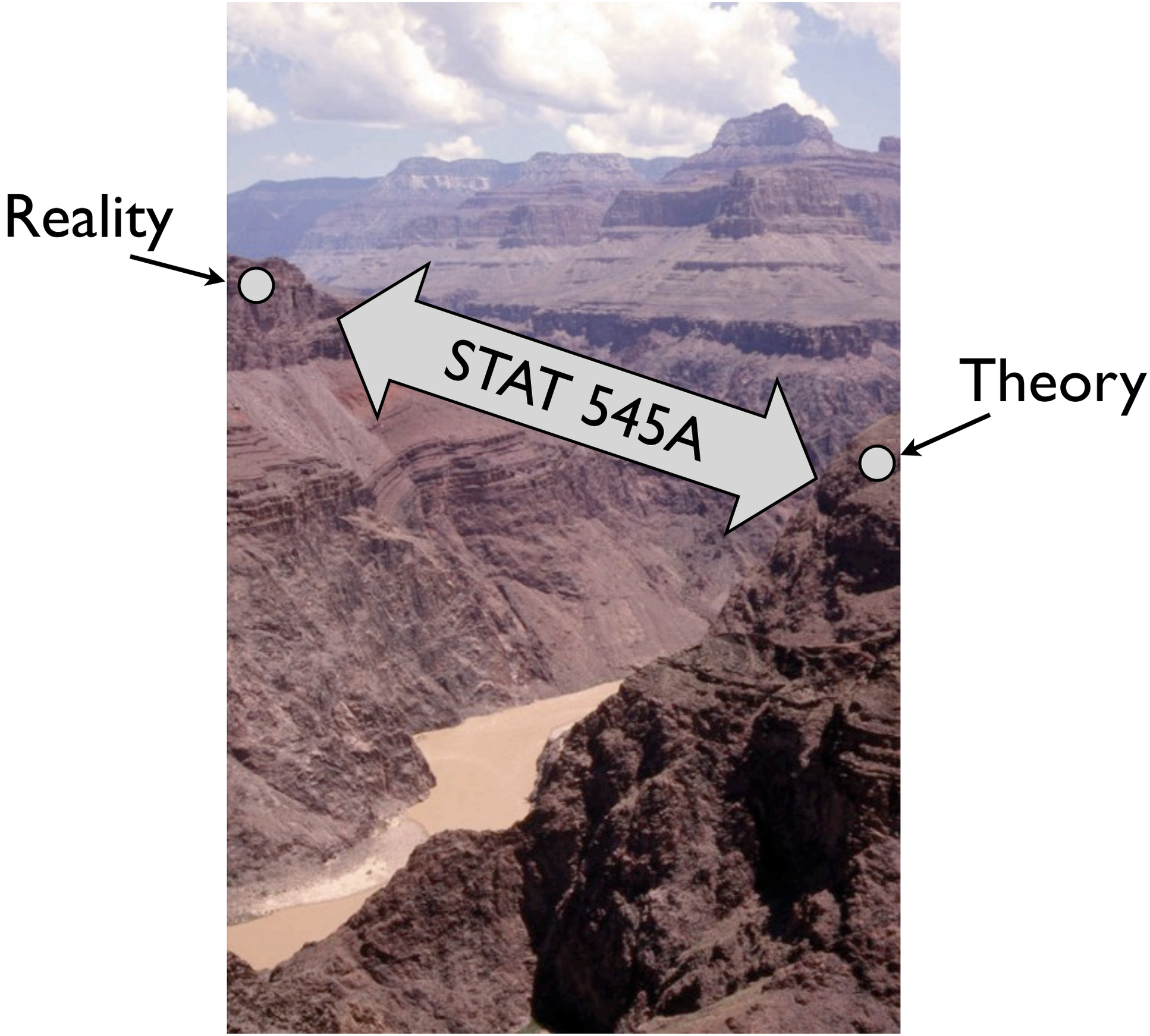
STAT 545A
Class meeting #1
Wednesday, September 5, 2012

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 on data analysis, especially exploratory analysis.
- Strengthen your data analysis skills.

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.

Importance of data analysis for our discipline

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

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

Importance of data analysis 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'?

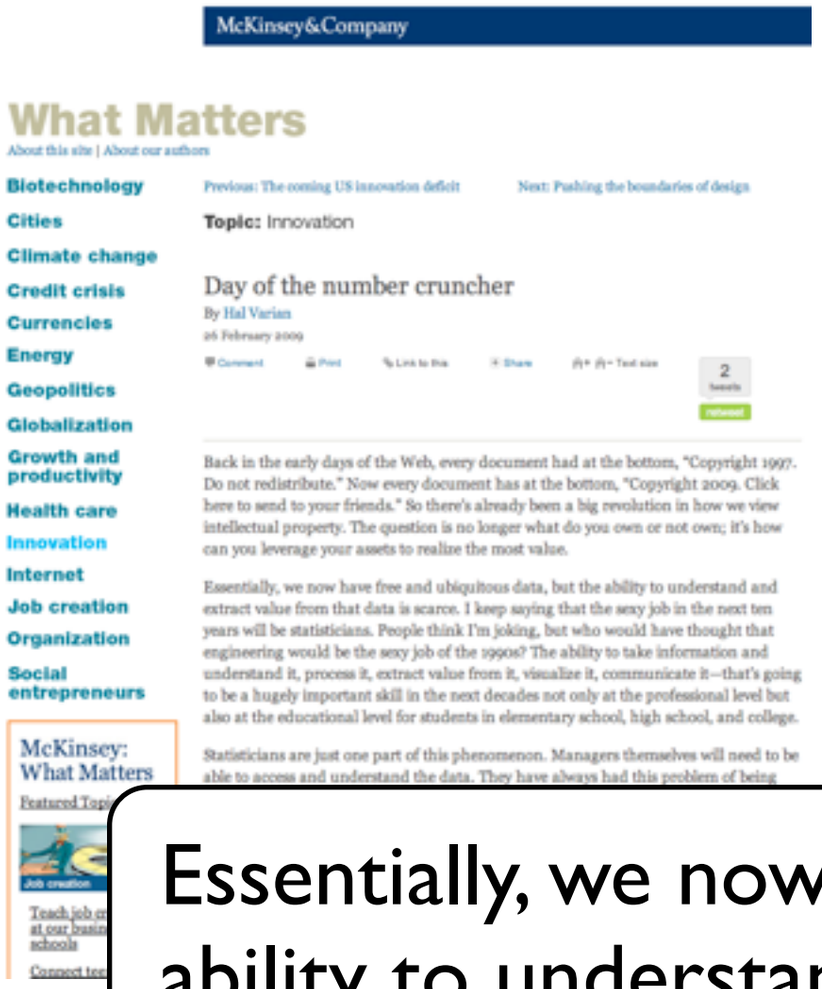
Importance of data analysis for our discipline

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 not only at the professional level but also at the educational level for students in elementary school, high school, and college.

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

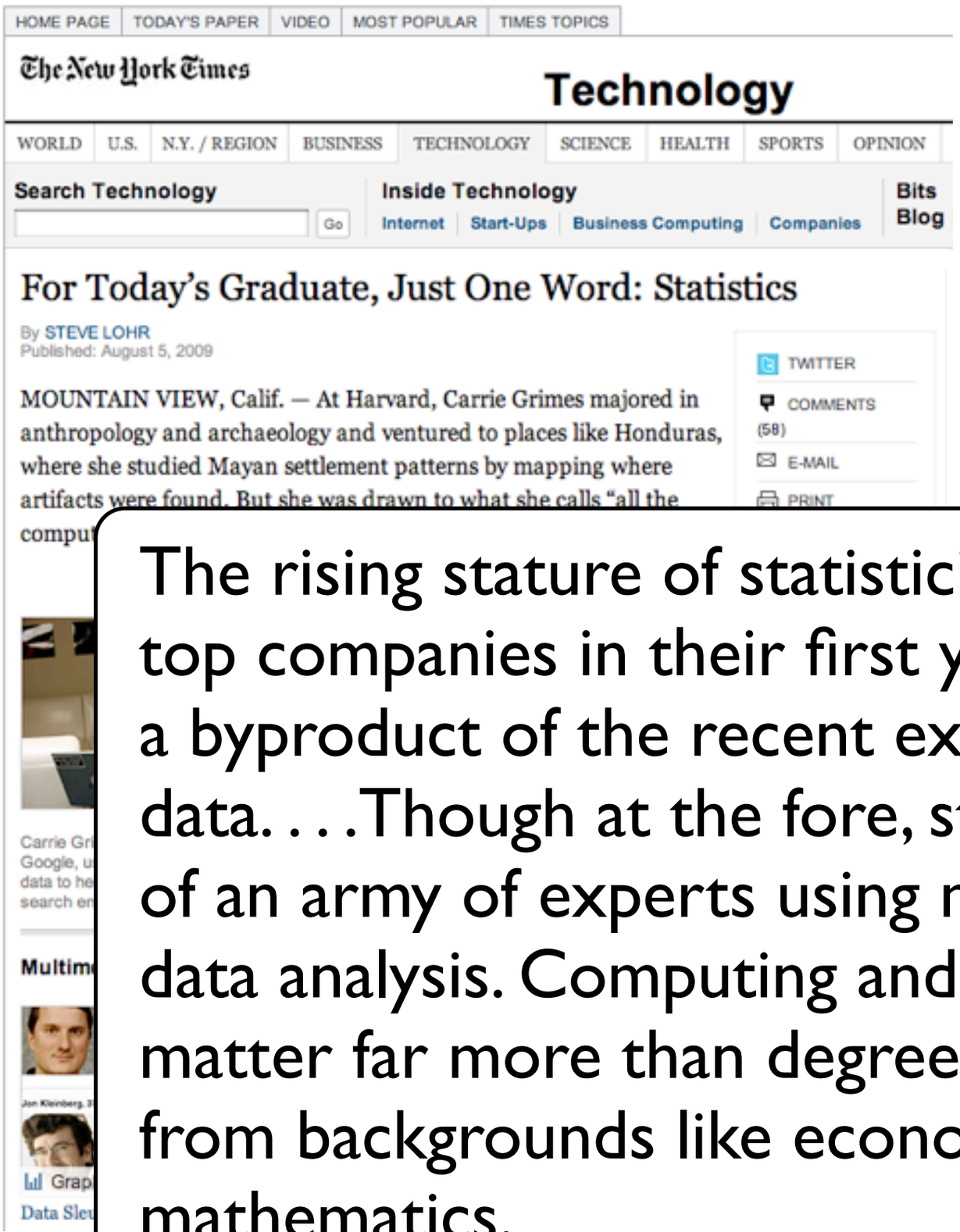
Teach job cr... at our busin... schools

Connect tea...

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 | 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

TWITTER
COMMENTS (58)
E-MAIL
PRINT

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

Carrie Gri
Google, u
data to he
search en

Multim

Jon Kleinberg,

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 ...

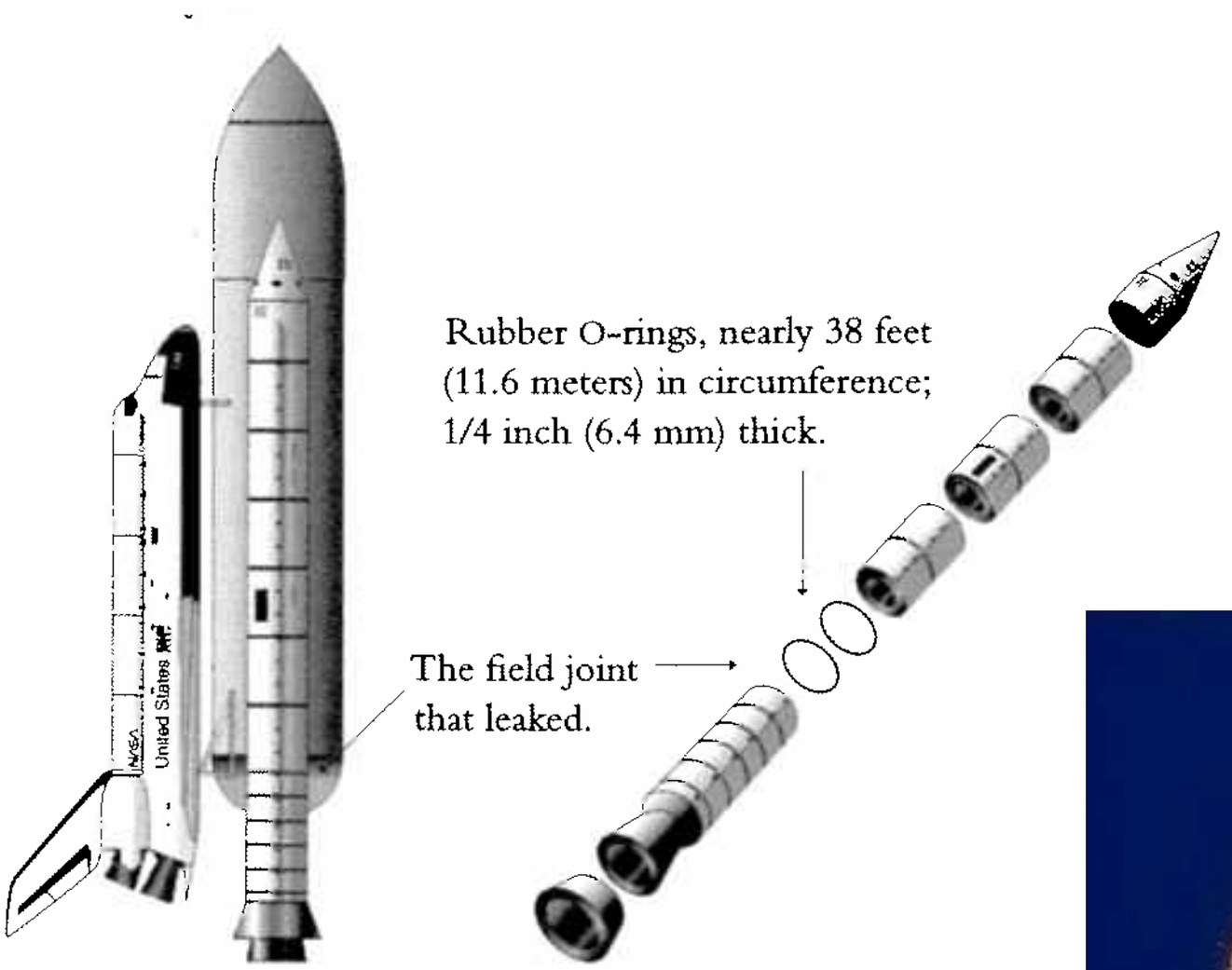
1. **Build your communication skills.** Unless you live in a plastic bubble, you are going to need to work with other people. You will be given tasks by other people, collaborate with other people to achieve those tasks, and ultimately have to report the results of your work to other people. You need 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.

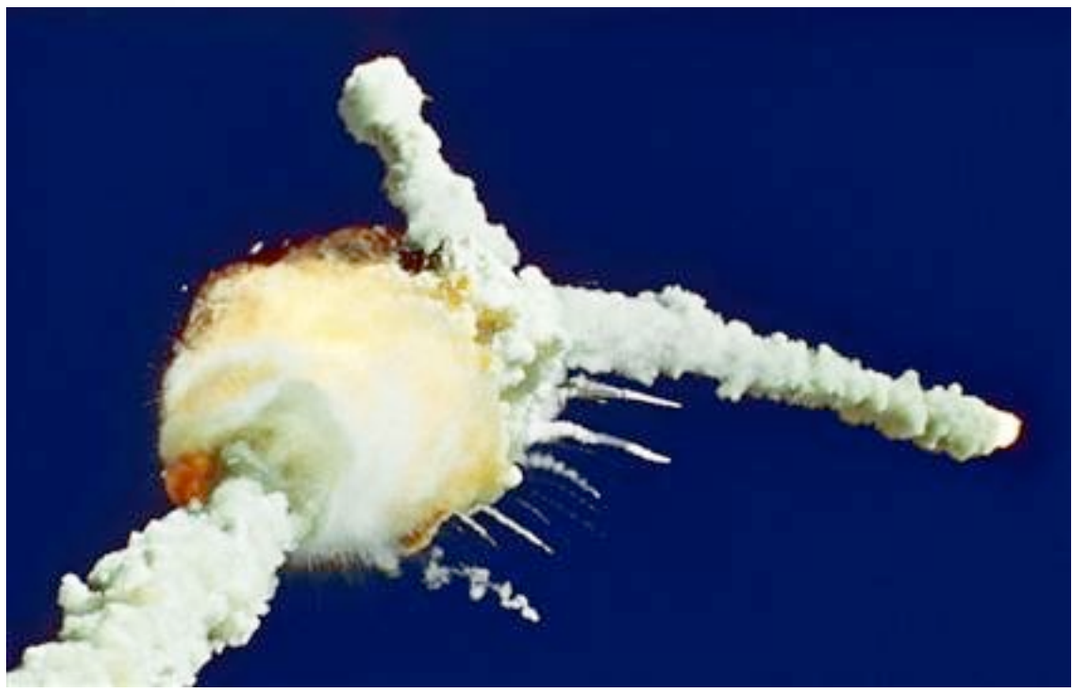
You need to take actions to ensure that your communication skills are sharp. These actions can include: (i) Taking a writing class, especially one that stresses technical writing, which has a completely different flavour from essay writing; (ii) Taking a class in verbal communication, and in particular 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°)

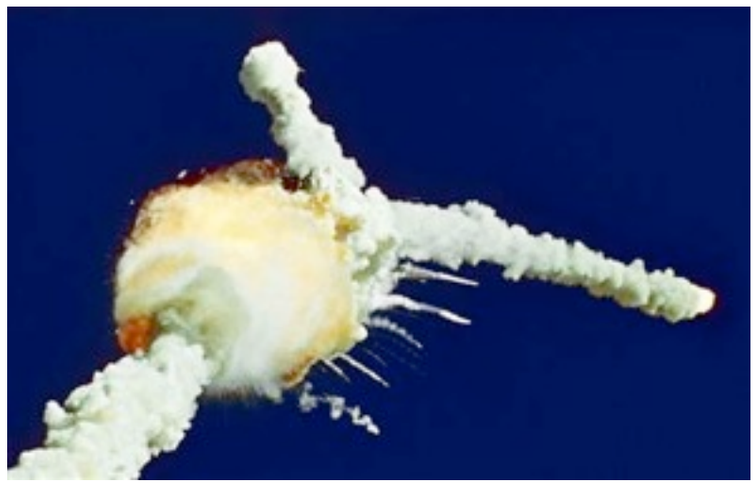
SRM-13A, 15, 16A, 18, 23A 24A
- NOZZLE BLOW-BY

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”

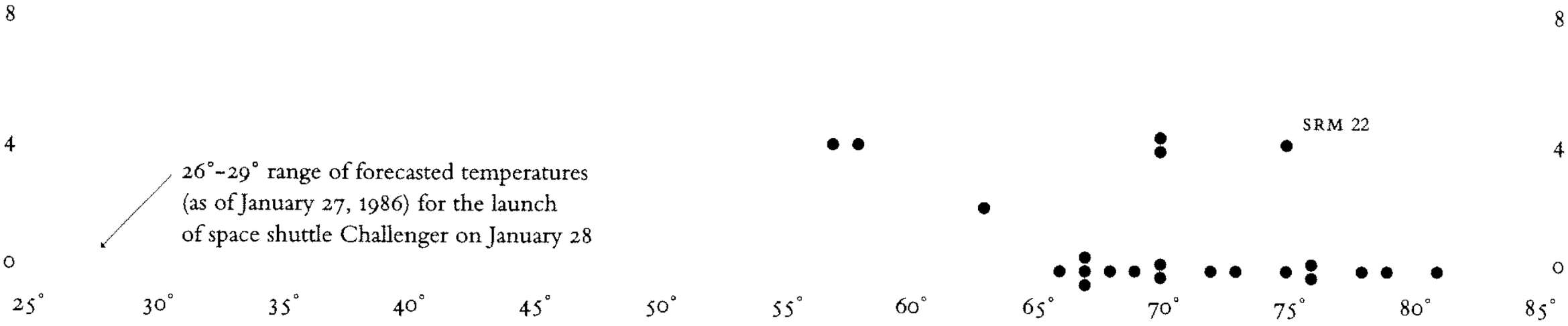


O-ring damage
index, each launch

12

SRM 15

12



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

“A picture is worth a thousand words”

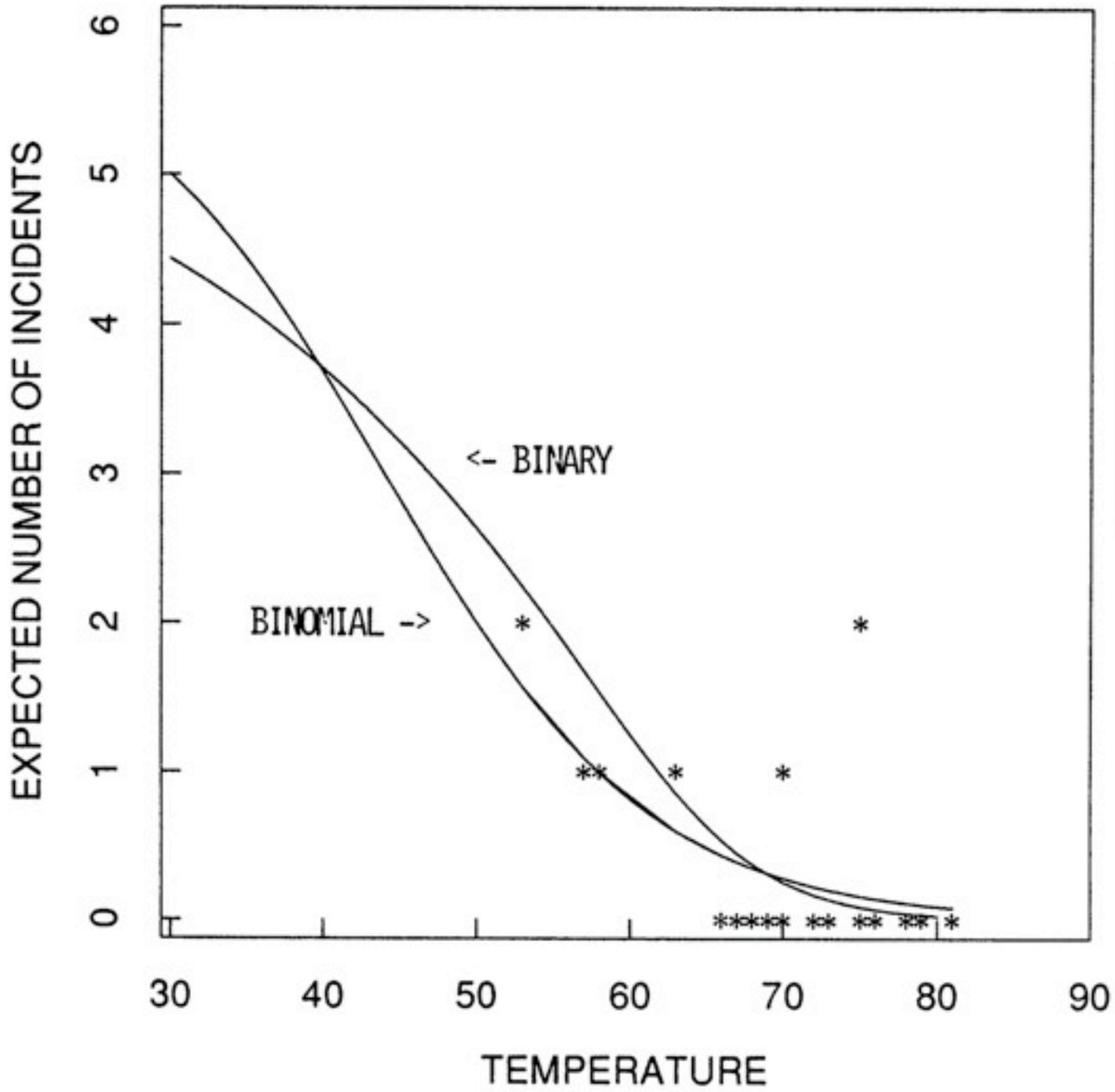
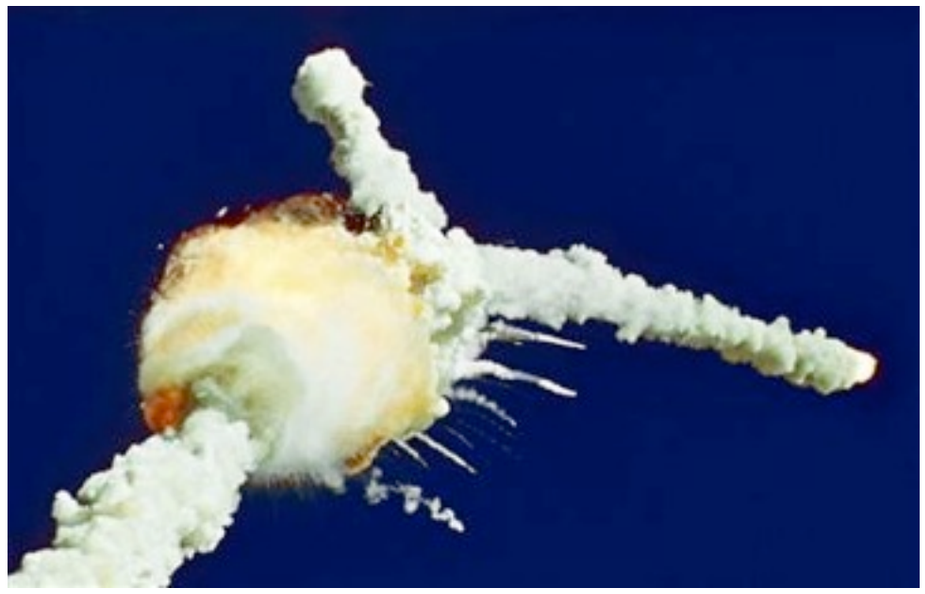


Figure 4. O-Ring Thermal-Distress Data: Field-Joint Primary O-Rings, Binomial-Logit Model, and Binary-Logit Model.



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!

oktrends

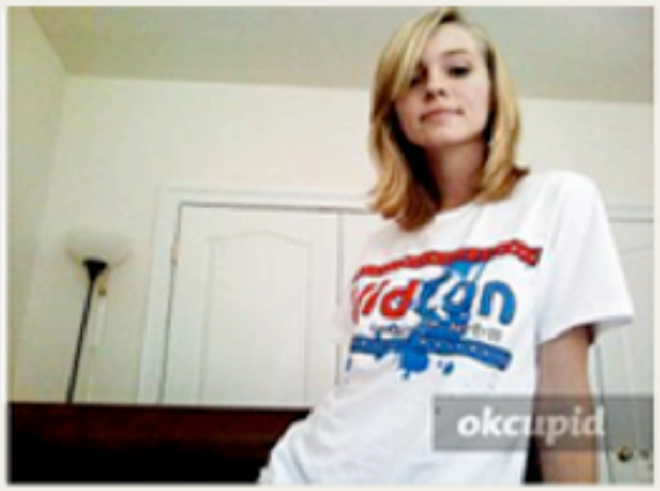
Dating Research from OkCupid

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

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

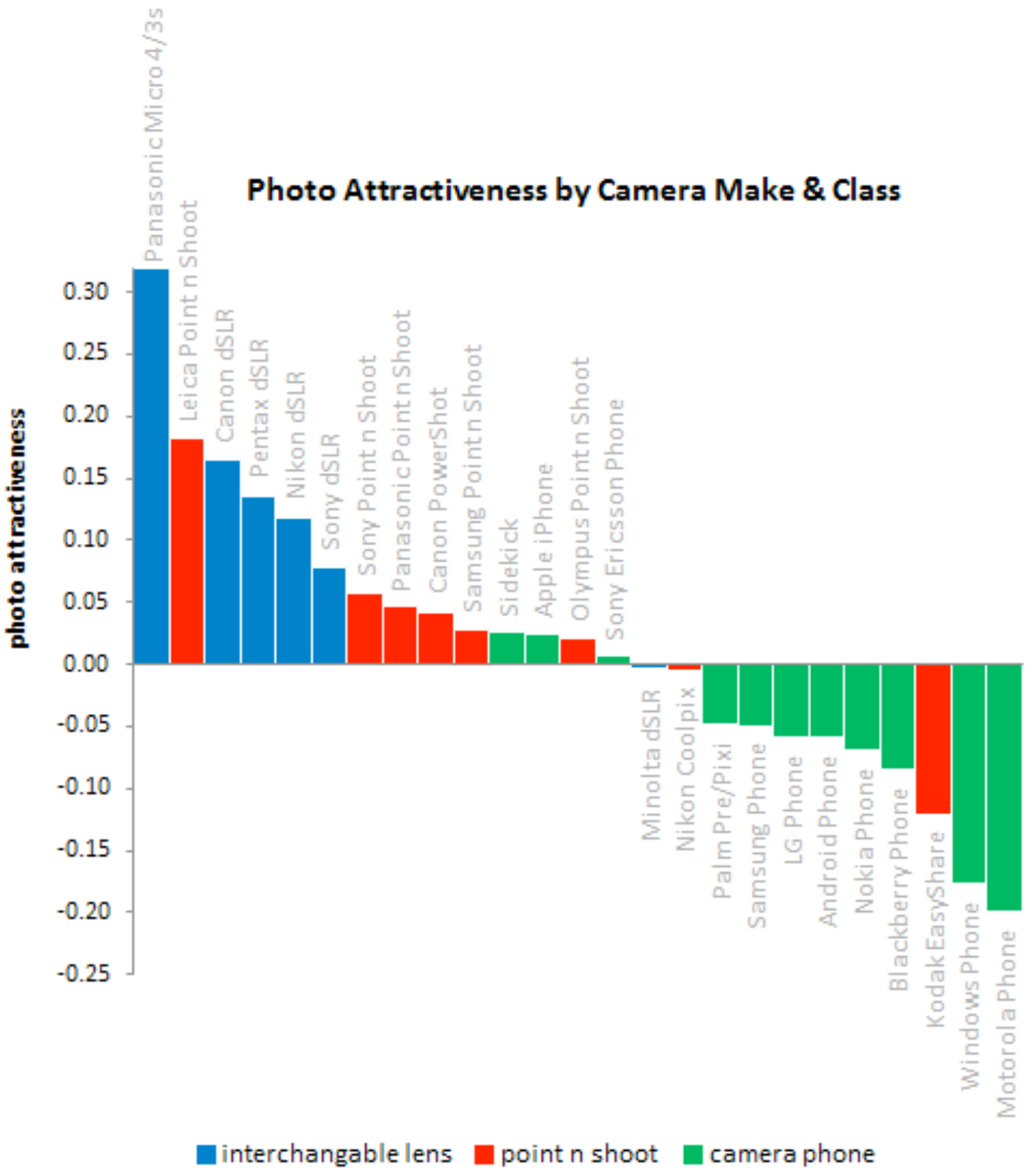
3. We collated these millions of judgments with the time of day each picture was taken, what the shutter speed was, and so on. Almost all modern cameras embed this stuff in a special header, called *EXIF data*.
4. We made graphs.

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:



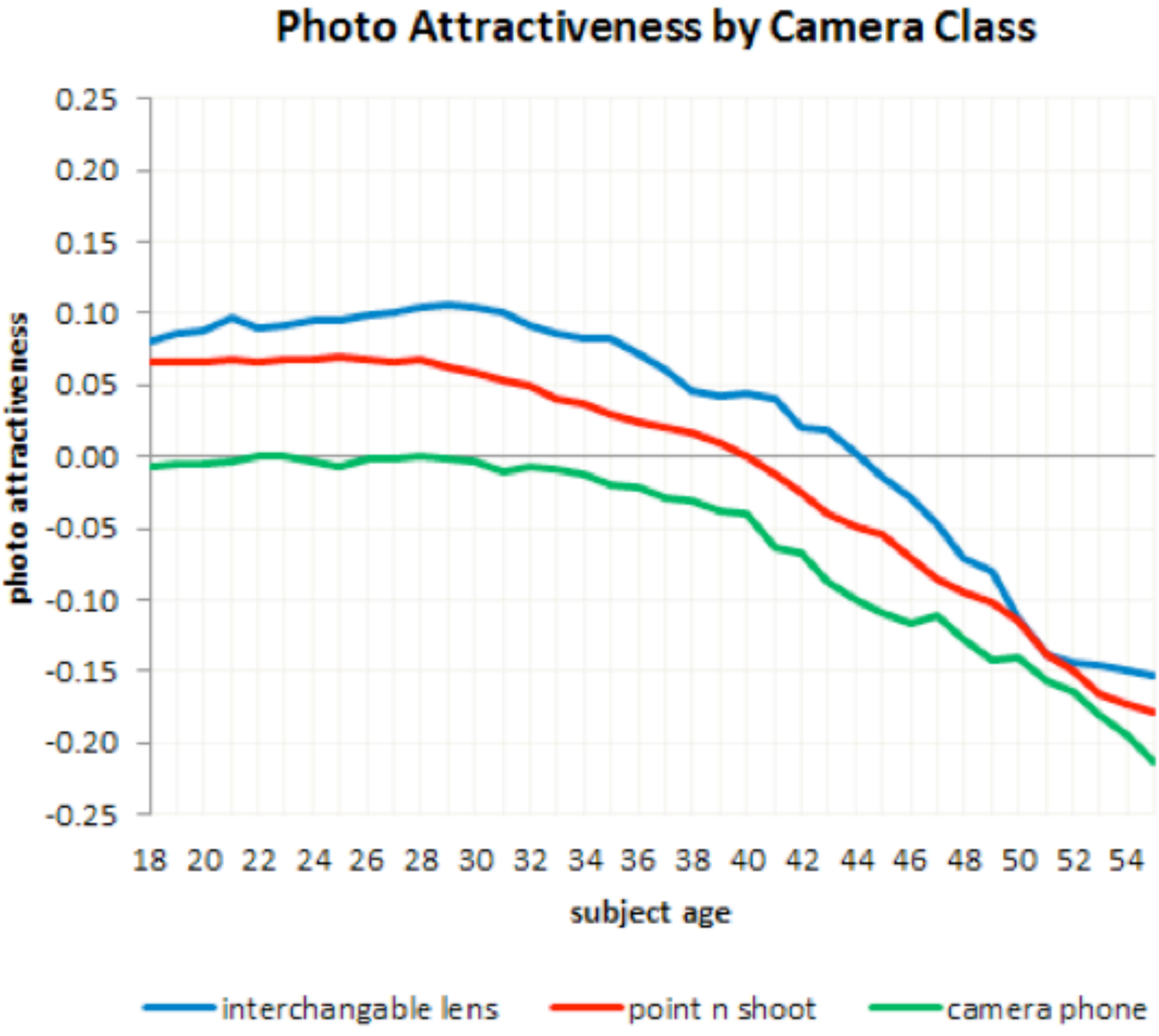
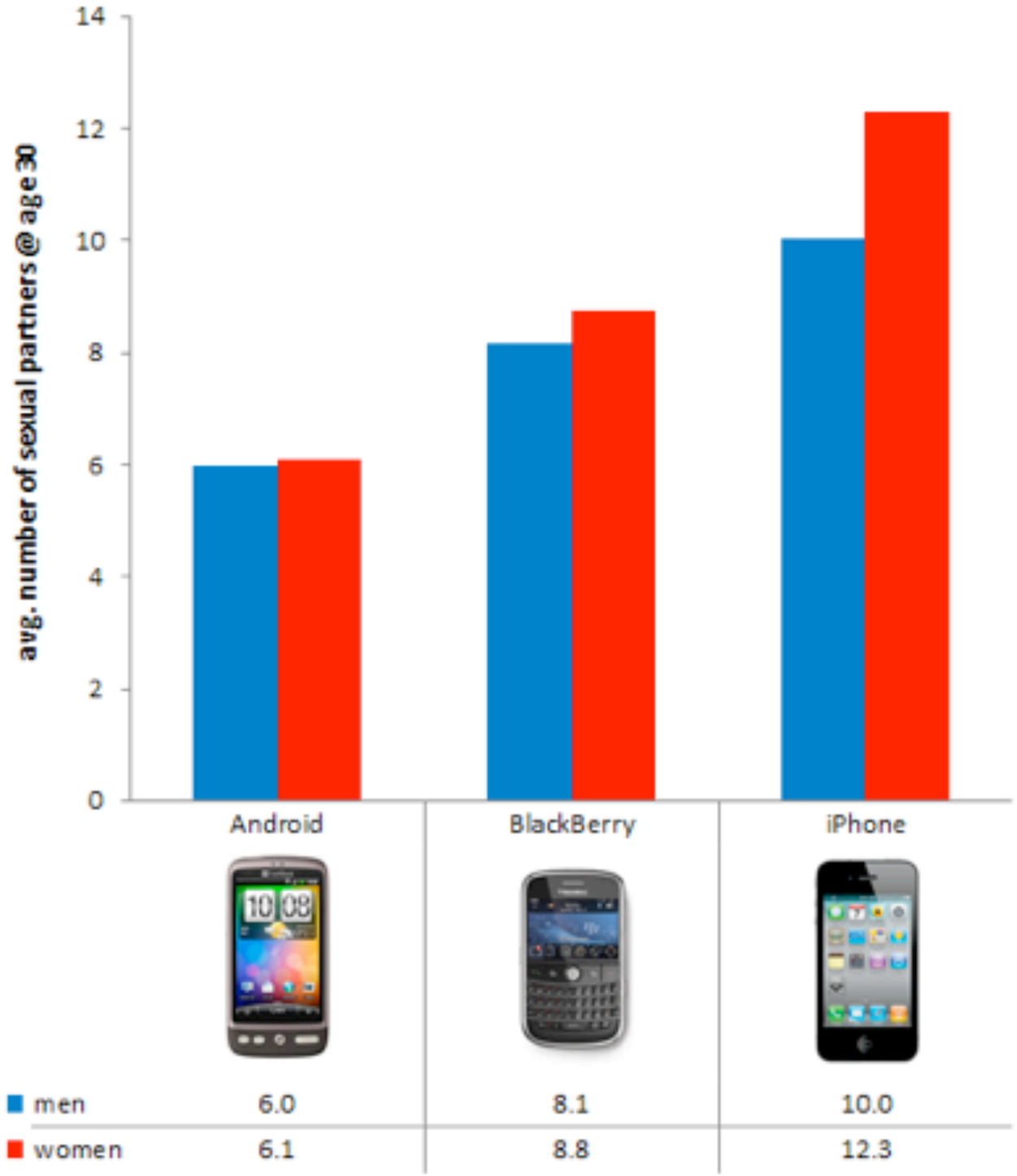
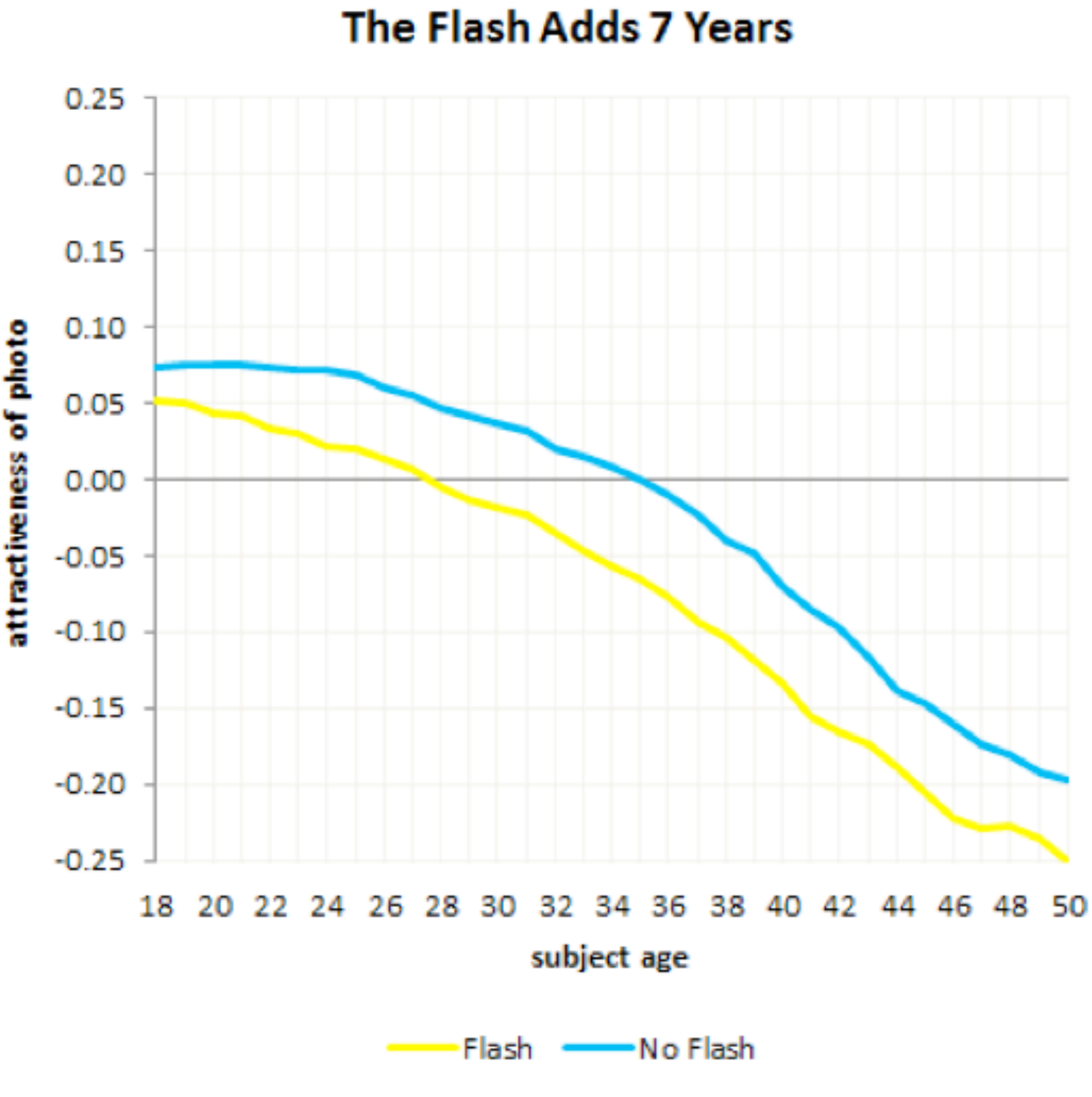


Photo Attractiveness by Camera Class
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.20
-0.25
photo attractiveness
18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54
subject age
interchangable lens
point n shoot
camera phone

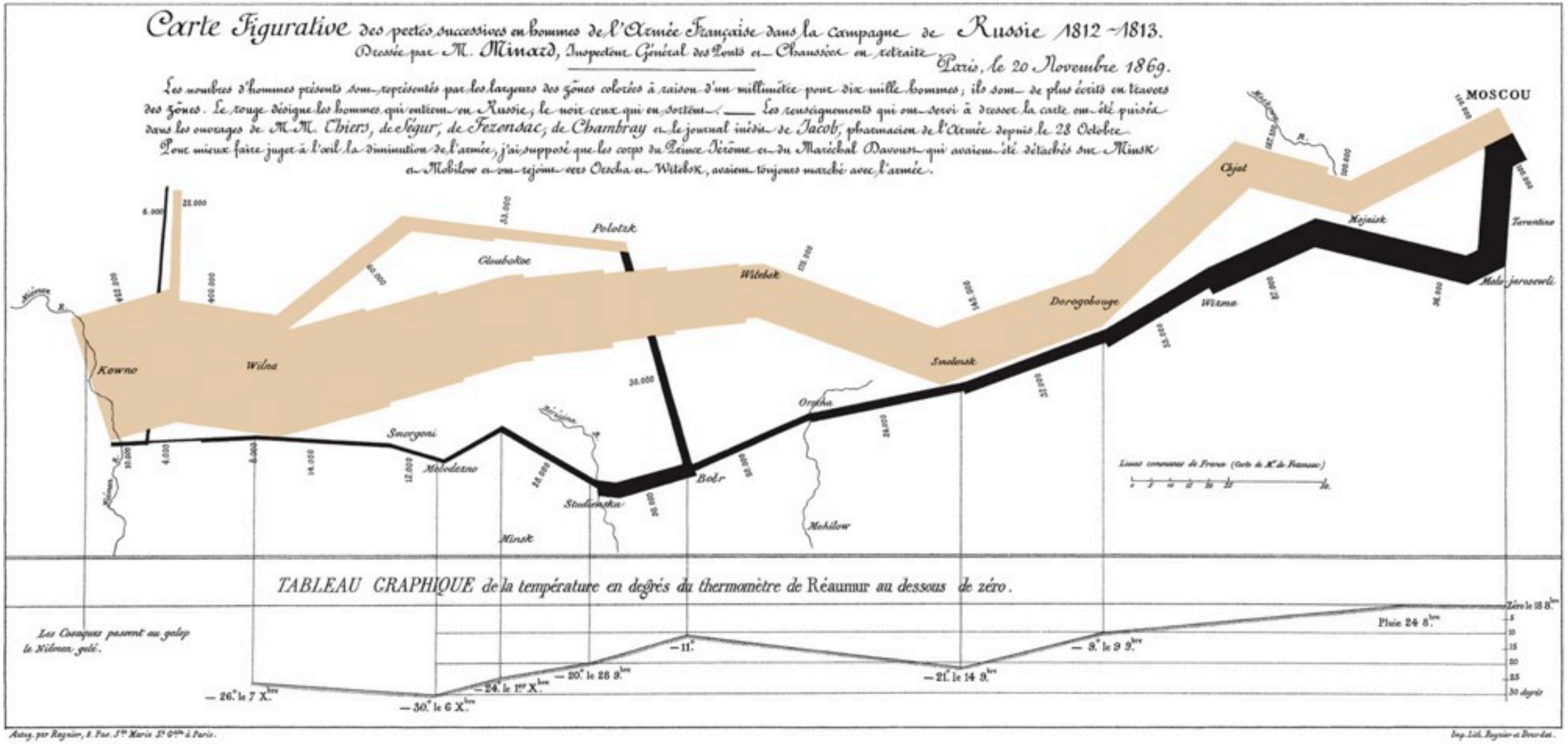
Sexual Activity by Smart Phone Brand



	Android	BlackBerry	iPhone
men	6.0	8.1	10.0
women	6.1	8.8	12.3


The Flash Adds 7 Years
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.20
-0.25
attractiveness of photo
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)

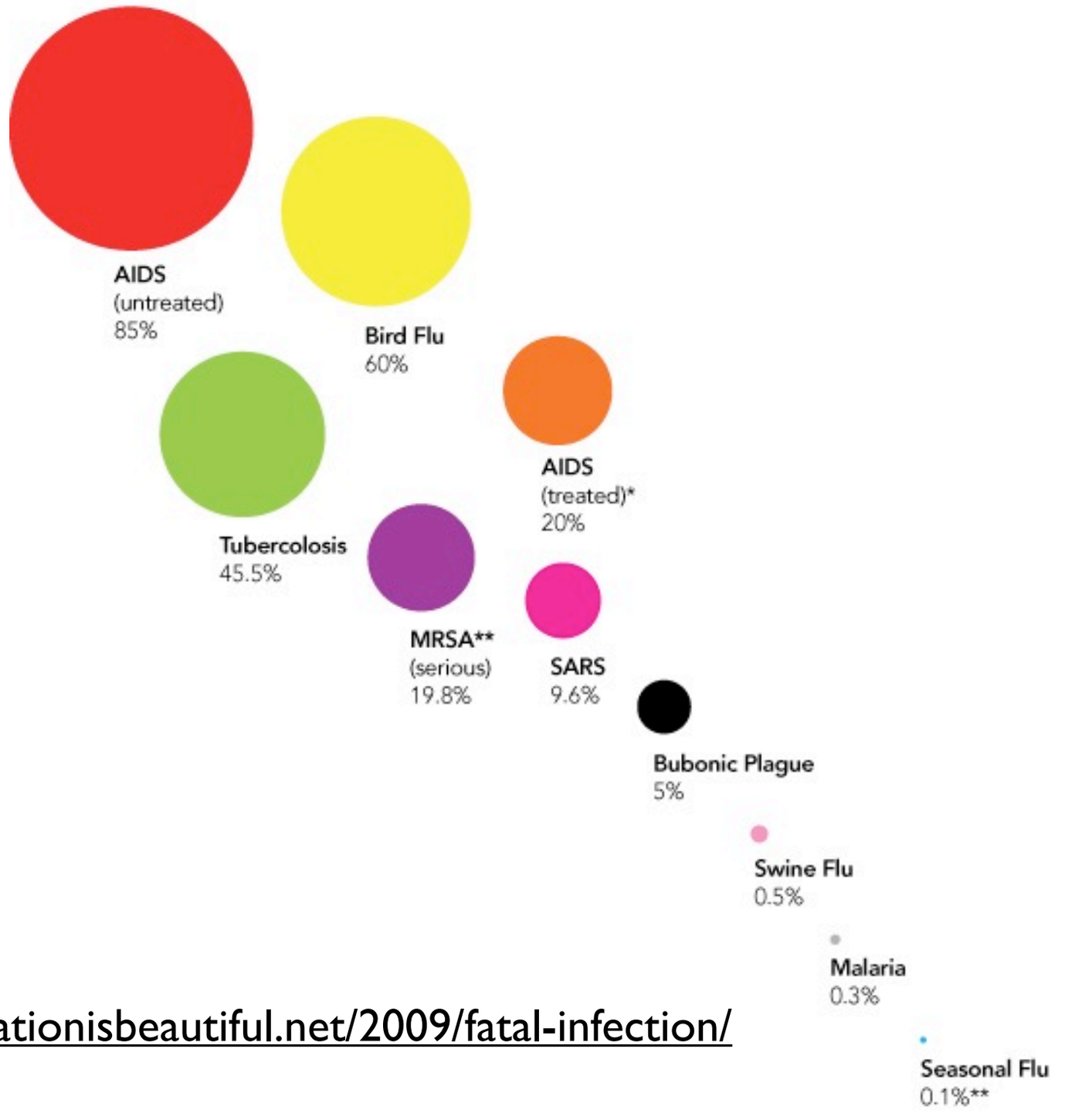


Famous chart/map by Charles Joseph Minard, much beloved by Tufte, depicting Napoleonic army during Russian campaign of 1812

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

Disease Case Fatality Rates

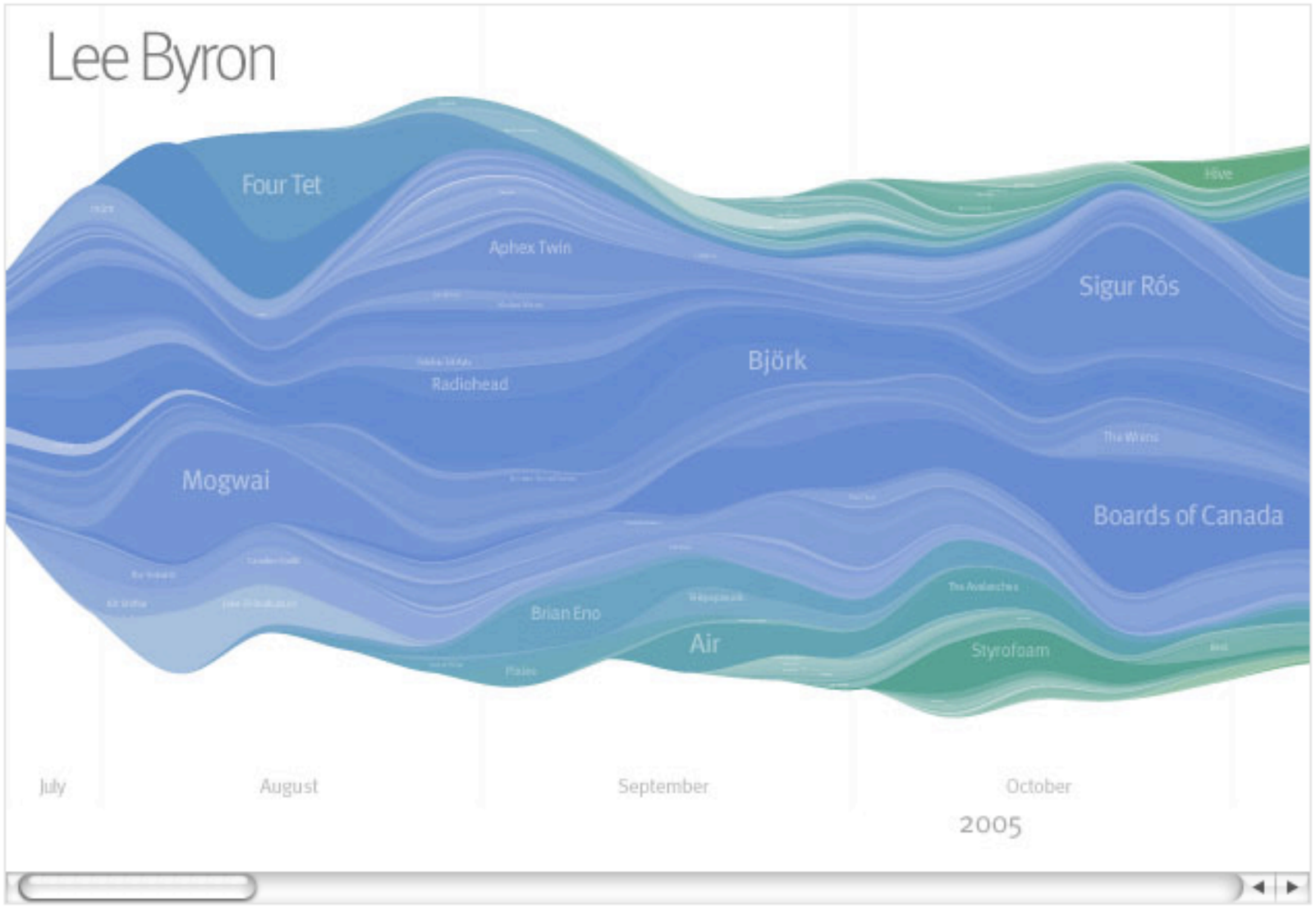
Average % of infected who die



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

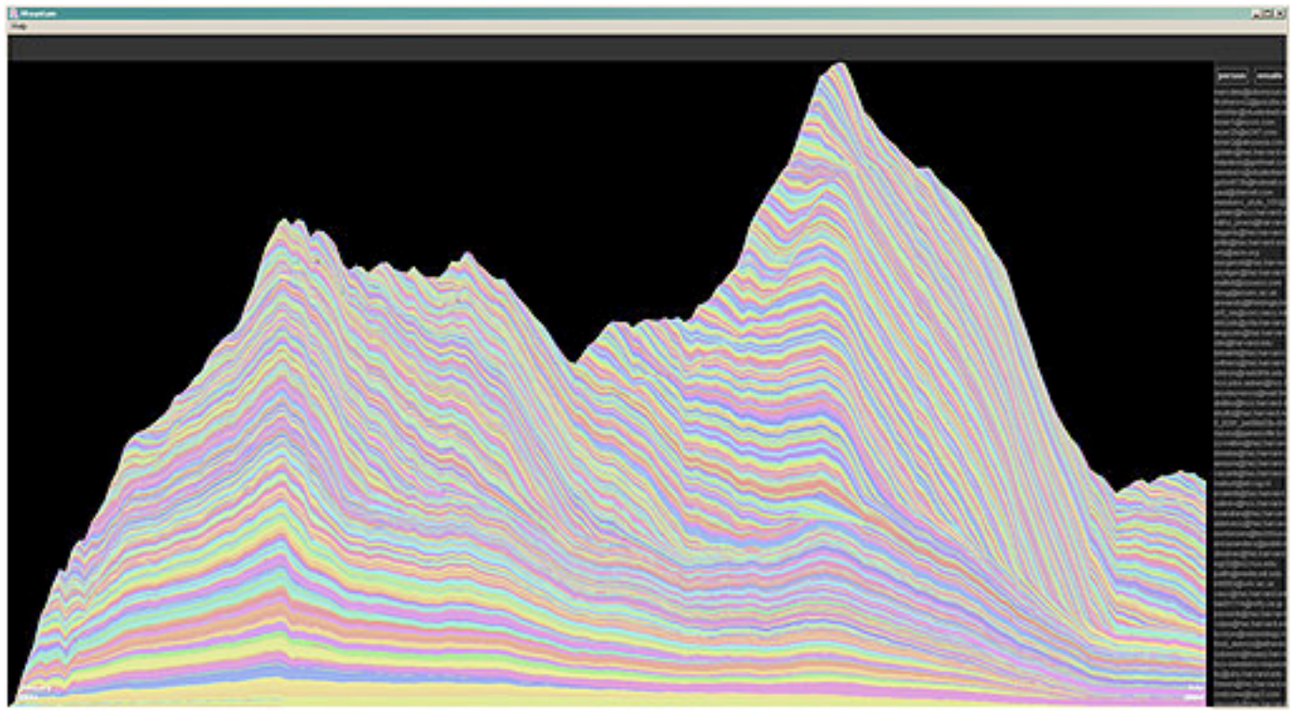
source: Worldwide figures from World Health Organisation, CDC, Guardian
*estimated from secondary sources // ** based on US figures

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



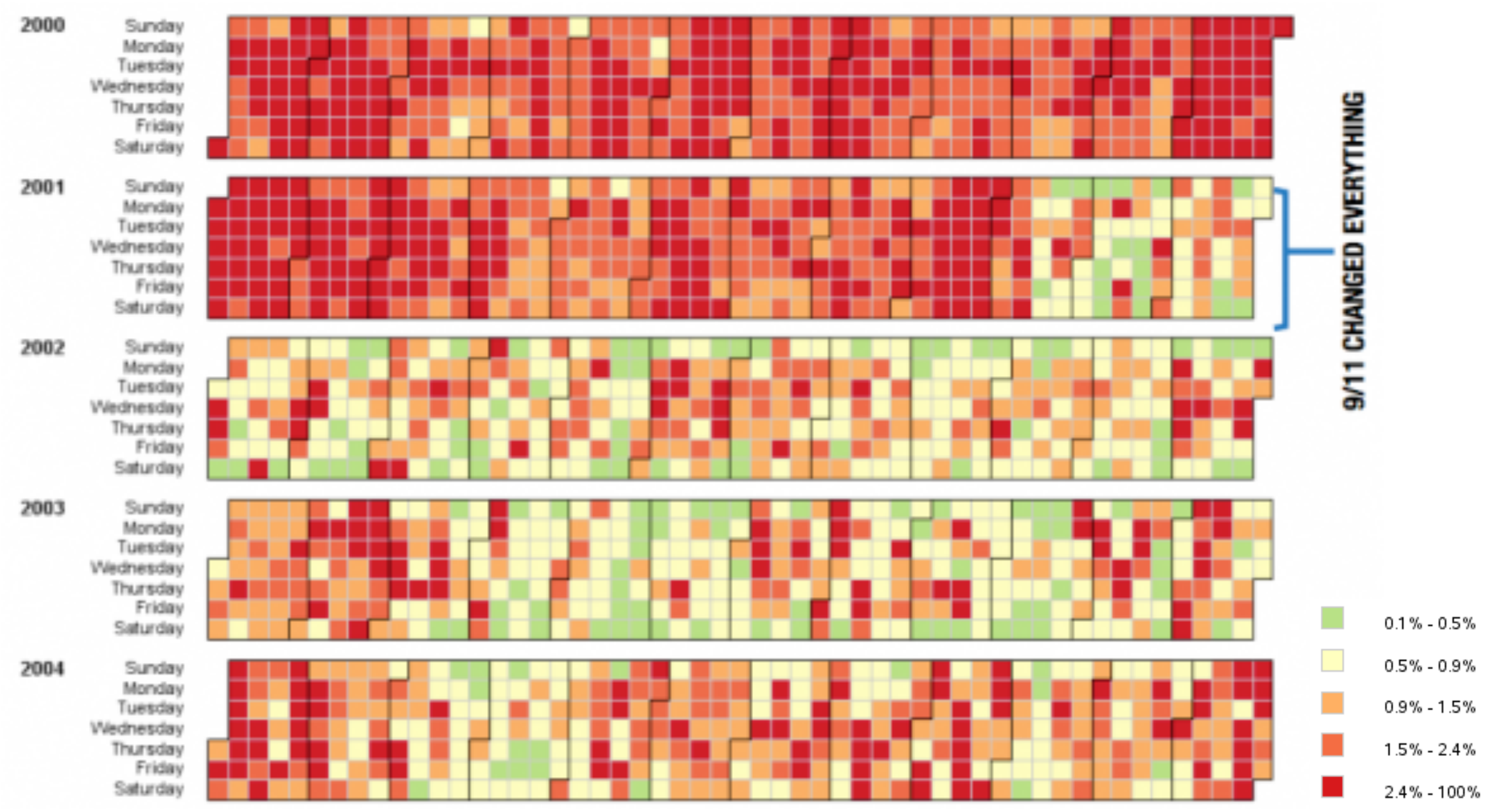
http://www.leebyron.com/what/lastfm/

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.

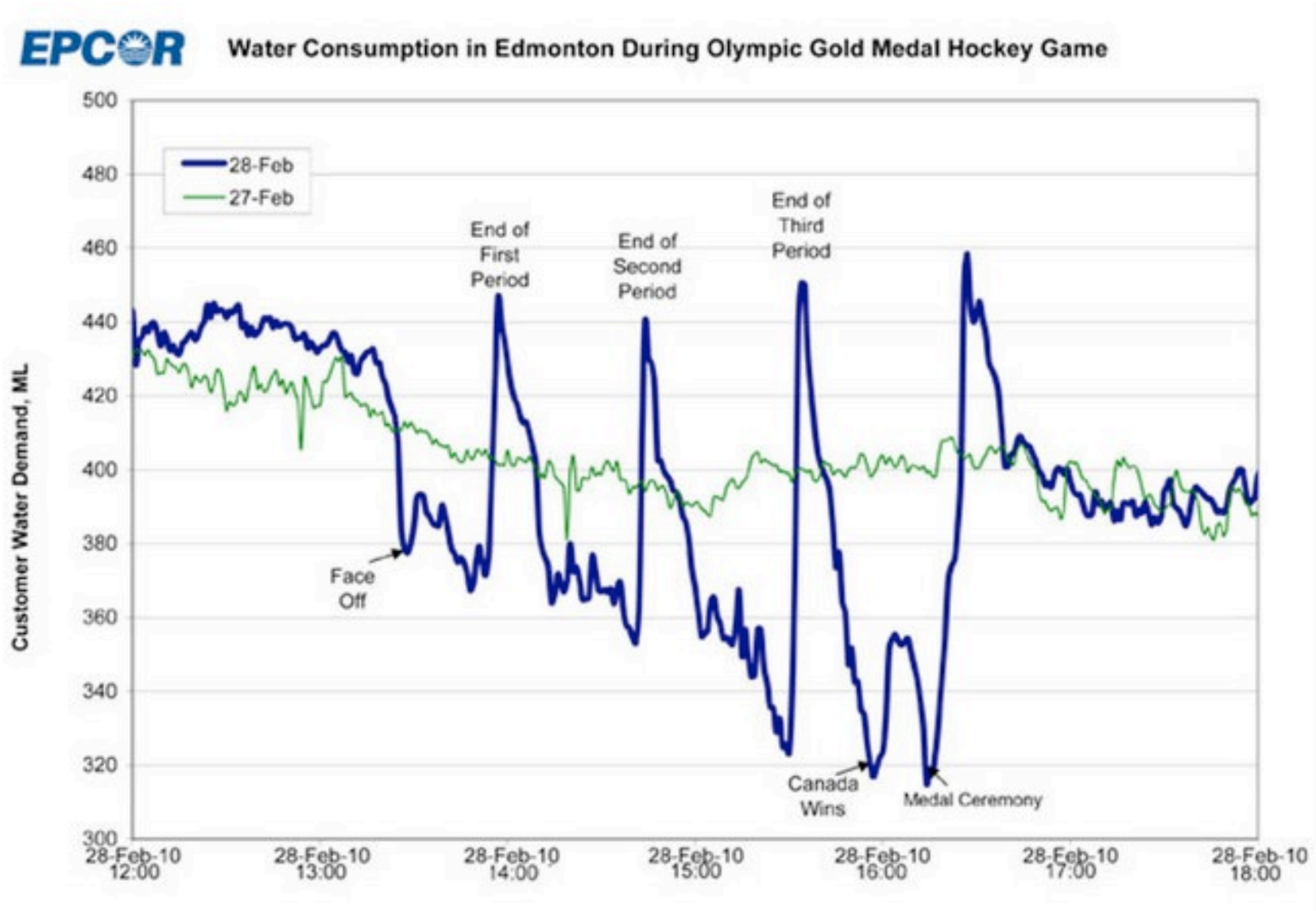
This is a view of flight cancellations. The more red a rectangle is, the higher the percentage of cancellations.

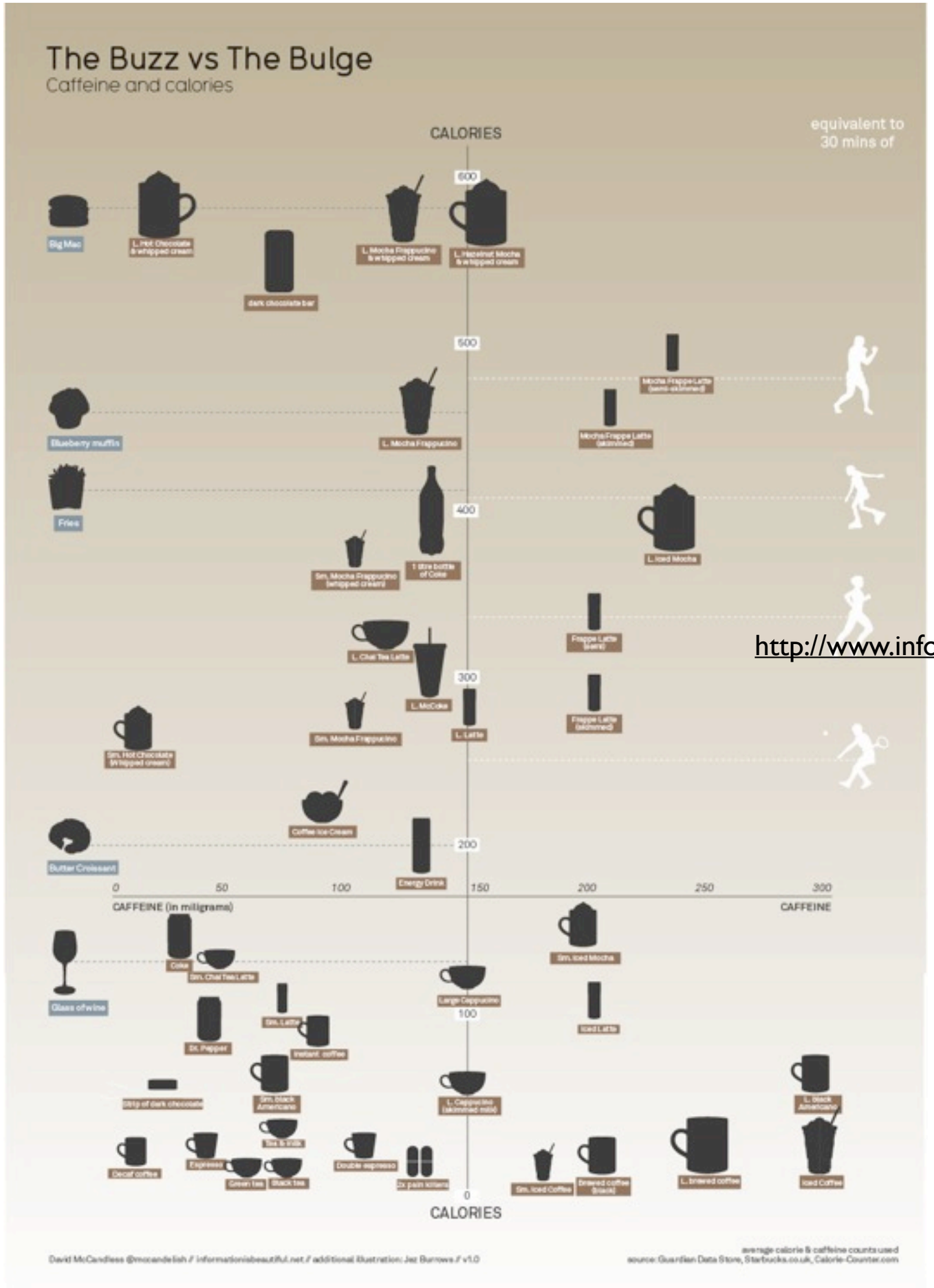


http://flowingdata.com/2009/09/10/3-in-depth-views-of-flight-delays-and-cancellations/

What If Everybody in Canada Flushed At Once?

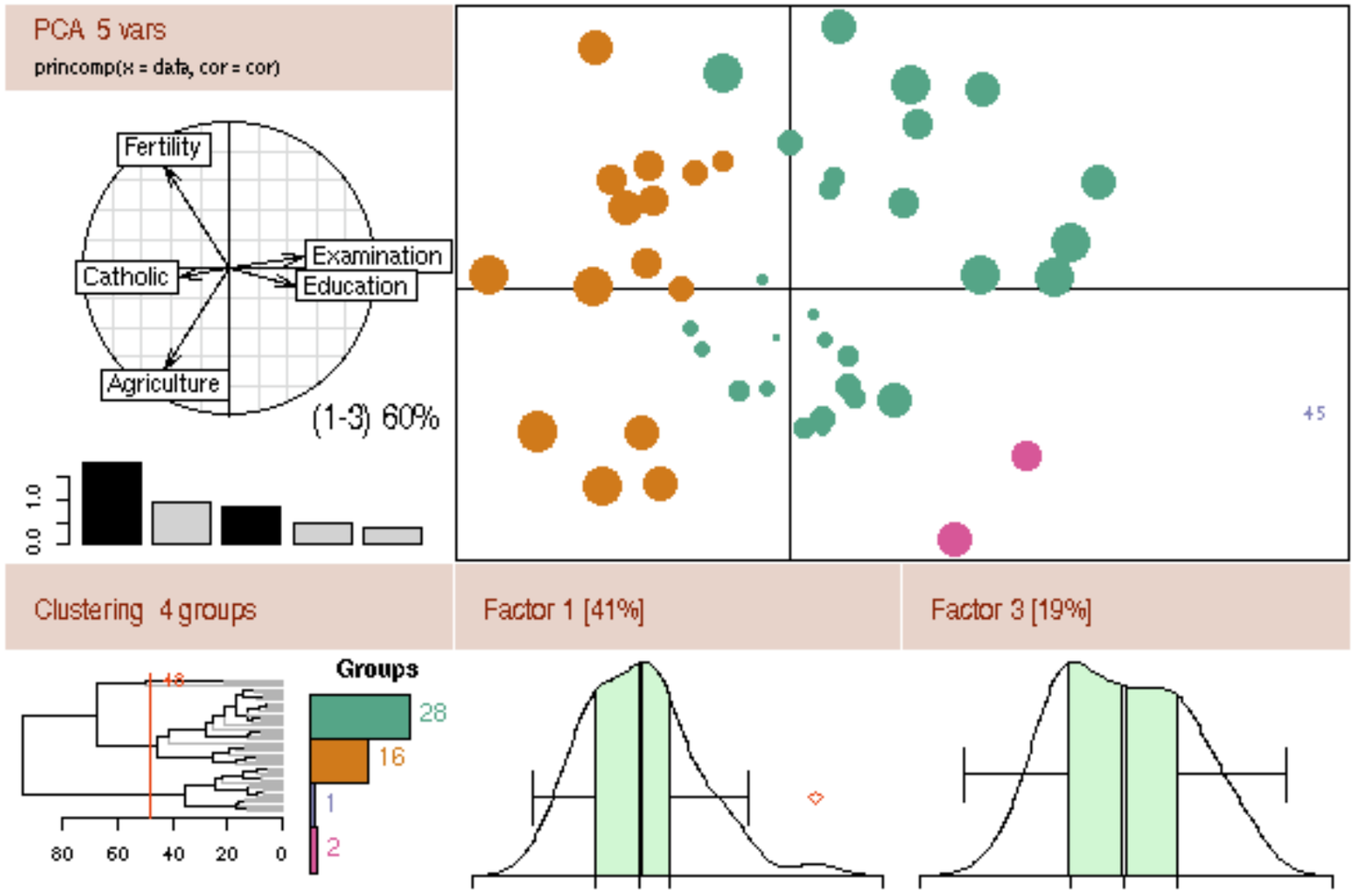
http://www.patspapers.com/blog/item/what_if_everybody_flushed_at_once_Edmonton_water_gold_medal_hockey_game/





http://www.informationisbeautiful.net/visualizations/caffeine-and-calories/

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



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/

Two inter-related goals

- Foster your development of a personal philosophy on data analysis.
- Strengthen your data analysis skills.

Summary of main philosophical points:

- Data analysis is important.
- Simple methods are preferred.
- Visual presentations of data and results are valuable.

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.*

* ggplot2 is another great add-on package for making graphics. I adopted lattice before this really existed as a viable option. The thought of starting over gives me pain, but I aspire to learn ggplot2 one day.....

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

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

The New York Times

Business Computing

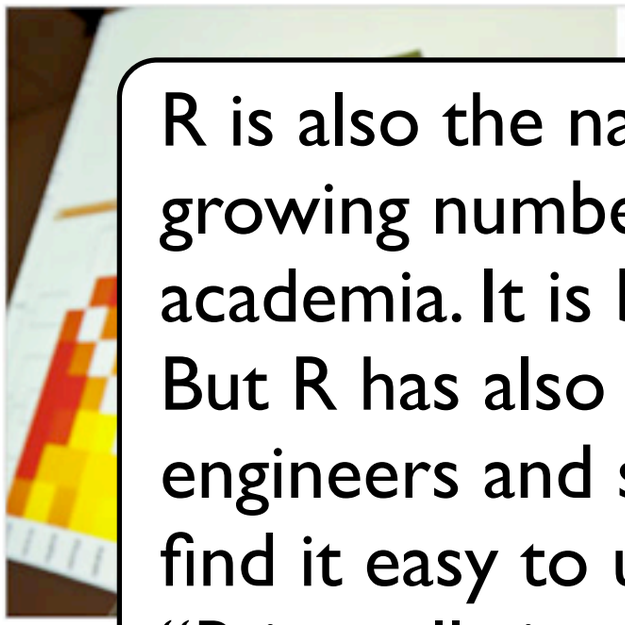
WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

Search Technology | Go

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

Bits Blog

Data Analysts Captivated by R's Power



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/

Programming / analytical practices

- R-aware text editors, such as Emacs Speaks Statistics or the hot new IDE (integrated development environment): RStudio.
- Good coding habits, e.g. naming conventions, commenting, indenting
- Organizing analytical 'project parts': data, code, numerical & graphical results
- Code management: versioning and modularity. Consider use of formal version control, such as git and github for collaboration.
- Unearthing R source code
- Managing an R installation.
- Advanced topics? Other topics?

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/

Two inter-related goals

- Foster your development of a personal philosophy on data analysis.
- Strengthen your data analysis skills.

Summary re: practical skills:

- We use R. We mostly use `lattice` for plots.
- We are disciplined and professional about writing code and organizing projects, because this makes analysis transparent, reproducible, and (at some point!) more enjoyable.

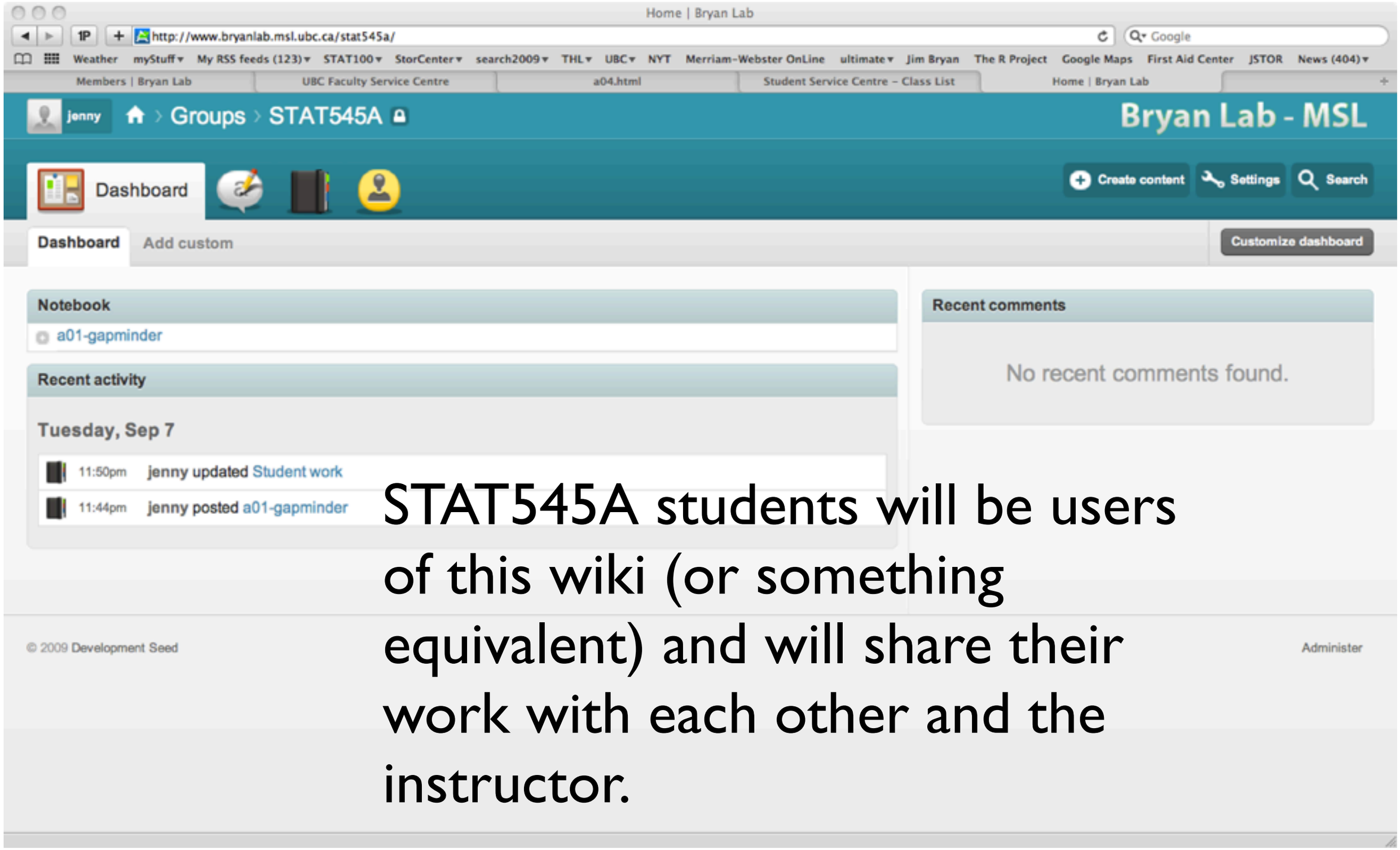
Culture of the class

- 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.
- Rewarding engagement, intellectual generosity and curiosity
  - Speaking up, sharing success OR failure, showing some interest in something will earn brownie points.
- 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.

How I plan to use our time and energy

- 1/3 = Concepts
- 1/3 = Mechanics
- 1/3 = Reflecting on our work, asking and answering questions, discussion

Probably will happen: a class collaborative webspace thingy. More on that later .....



STAT545A students will be users of this wiki (or something equivalent) and will share their work with each other and the instructor.

Where marks will come from

- Homework: 3 assignments (or more but smaller, if I can chop it more finely).
- First 2 assignments are designed to motivate (scare?) you into getting set up and started. Marking scheme is very coarse and, esp. at first, effort based!
  - Example: if completed, mark = 87
  - Example: check plus = 90, check = 85, check minus = 80
- First assignment: we all use Gapminder data.
- Second assignment: you find your own data (what you might use for final assignment).

Where marks will come from

- Final assignment is the main one. Two choices:
  - Find your own data (in assignment 2? join forces with someone with more exciting data?) and present it.
  - A more proscribed, statistical exercise (in past has been a bootstrap problem ... this option may be eliminated).
- Consists of a series of figures, accompanied by a ~3 page report, plus code.

Where marks will come from

- Course mark will mostly be that of the final assignment. Exceptionally good (or bad) performance on first two assignments, contributions during/outside of class, how easy and pleasant for others to help you, etc. can lead to slight adjustments up or down. FYI: Last year I adjusted two marks, one +1 and one +2.

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.

UBC Grading System, as applied/interpreted by Dr. Jenny Bryan for STAT 545A

	A+ **90-100**	**A** **85-89**	**A-** **80-84**	**B+/B/B-** **76-79/72-75/68-71**	**C, D, F** **0-67**
Re: graduate study in Statistics	Outstanding. You are encouraged to pursue a course of study with a strong applied statistics / statistical computing component. Your natural abilities and deliberate efforts suggest you will excel in this area.	Very good. You show potential to become a very strong applied statistician and data analyst. The deficiencies with your work are likely to improve dramatically with more experience.	Adequate. You will be able to perform the necessary applied statistics / computing tasks in your studies, but this does not seem to be one of your strengths.	Needs work. You will need to improve in this area to achieve the level of competence expected in someone with a graduate degree in statistics.	Inadequate. Graduate studies in statistics might not be right for you.
Personal viewpoint: Would I be willing to supervise such a student and fund her/him as an RA from a research grant?	I would be pleased to work with this student as an RA. I expect his/her work to be of very high quality and require very little correction, refinement, and editing. It will be pleasant and efficient to work with this student in my research projects, collaborations, and publications.	I would be willing to work with this student, if we can identify a specific suitable project. I expect his/her work will initially require a fair amount of checking and refinement, but he/she will rapidly become more independent.	I would be reluctant to work with this student as an RA. I expect his/her work to contain many errors and omissions and will require a great deal of my time and energy to make the analysis, writing, figures, and tables good enough to show to my collaborators and incorporate into publications.	I would not be willing to work with this student as an RA.	I would not be willing to work with this student as an RA.

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
- Recent focus on studies of large collection of 'knockout' organisms and statisical methods for analyzing graphs
- 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's in here

	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

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 class webspace?
- Offices (I prefer appointments):
  - ESB 3116 (main --> only office)
  - MSB 229 (being vacated)

Admin & communication

- Course webpage:
  - Collaborative web space COMING SOON?
  - Will be linked from my regular webpages:
  - http://www.stat.ubc.ca/~jenny/teach/
  - http://www.stat.ubc.ca/~jenny/teach/STAT545/
  - Stat dept --> faculty --> jenny --> teaching

An incomplete list of good books

- Venables WN, Ripley BD (2002) Modern applied statistics with S. Springer.
- Sarkar, D (2008) Lattice: Multivariate Data Visualization with R. Springer. Available via PDFs through UBC's arrangement with SpringerLink. *
- Spector, P (2008) Data Manipulation with R. Springer. Available via PDFs through UBC's arrangement with SpringerLink.
- Chambers, J.M. (2008) Software for Data Analysis: Programming with R. Springer, New York. (via SpringerLink).

* *This is a real treasure trove of statistics and statistical computing books you can get online!*

More sources & books of interest

- Cleveland, William S (1993). Visualizing Data. AT&T Bell Laboratories.
- STATSnetBASE (CRC Press)
  - A Handbook of Statistical Analyses Using R by Brian Everitt and Torsten Hothorn.
  - Handbook of Statistical Analyses using S-Plus, Second Edition by Brian Everitt.
  - R Graphics by Paul Murrell.
  - R Programming for Bioinformatics by R Gentleman.

Sept

	5
10	12
17	19
24	26

Oct

1	3
Thanksgiving	10
15	17

12 class meetings

Breakdown of methods

- univariate data (single quantitative variable) and, optionally, a categorical variable
- bivariate data (two quantitative variables) and, optionally, a categorical variable
- multivariate data (3 or more quantitative variables)
- multiway data (single quantitative variable, two or more categorical variables)

What to do now?

- YESTERDAY: Get R installed. Strongly encourage working with R via Rstudio.

- ASAP: Start using resources to get started with R. I will not provide super-basic details that are widely available in eBooks, tutorials, etc. but will focus on higher-level issues (though I am easily engaged in specific discussions of syntax & approach). A few examples:

  - http://www.r-project.org/ (see Documentation, esp. Manuals)
  - Phil Spector's R intro from his class.
  - More detailed coverage from Phil Spector.
  - Nice collection of R Resources from Cerebral Mastication.

What to do now?

- WHEN PROMPTED: Reply to my email, so I can capture your email address and invite you to join the web space, if that happens. If you don't get my email, maybe you're not registered (yet)? In that case, just send me an email (jenny@stat.ubc.ca).

- BEFORE CLASS on MONDAY SEPT 10: Complete "Assignment 1: R Gapminder challenge" and (try to) post your results in the course webspace.