New York R statistical programming meetup

Mark Hansen (on leave from UCLA at NYTimes R&D)

May 6, 2010

Unit 6:

Participatory Urban Sensing

Background on Participatory Sensing (start)

The Data-Driven Life

Horacio Salinas for The New York Times

By GARY WOLF
Published: April 26, 2010

Humans make errors. We make errors of fact and errors of judgment. We have blind spots in our field of vision and gaps in our stream of attention. Sometimes we can't even answer the simplest questions. Where was I last week at this time? How long have I had

SIGN IN TO RECOMMEND

TWITTER

COMMENTS (135)

MOST POPULAR

E-MAILED | BLOGGED | SEARCHED | VIEWED

1. Nicholas D. Kristof: New Alarm Bells About Chemicals and Cancer
2. A Sampling of Chinglish
3. Gardens That Grow on Walls

your.flowingdata

A project by FlowingData

Sign in with Twitter

Capture your life in **data**. One **tweet** at a time.

Get Started Now »

Step 1. Follow @yfd on Twitter

Step 2. Sign in to *your*.flowingdata with Twitter

Step 3. Start recording data (via direct messages) following a few simple guidelines

Making Choices

We make tiny choices every day. Those choices become habits, and those habits develop into behaviors. *your*.flowingdata helps you record these choices.

READ MORE

Collect data anywhere.

The ubiquity of Twitter allows you to record data from just about anywhere. If you can tweet, you can record data.

Interact with your data.

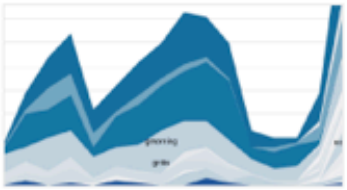
Data is meant to be played with. Use interactive data visualization and explore your data.

Customize views to your data.

All data is not created equally. Create custom visualization pages for what you're most interested in.

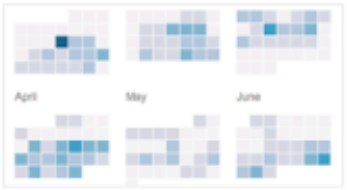
Share your findings.

Some data is meant to be private, but some is worth sharing. You decide what others can and can't see.



Understand yourself.

In the same way you can see growth from reading old entries in a diary, monitor your growth and progress through data.



Explore your data easily.

your.flowingdata was designed by a statistician, but you don't have to be one to play with your data.

Follow @yfd on Twitter and Get Started Now

your.flowingdata A project by FlowingData

DEMO USER | Start Recording Your Own Data Now

This is a DEMO. Login in with Twitter to start recording your own data.

Overview // Explore All Actions: Calendar | Treemap | Cloud | Time Series // Actions Log

Custom Pages: Click to Select a Page

Most Recent - 3 days ago

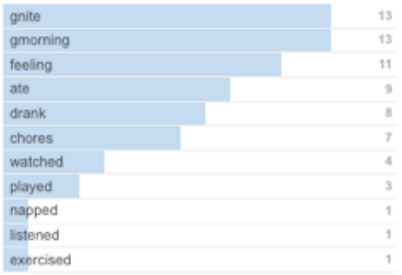
gnite

listened This American Life 4 days ago | ate turkey artichoke panini 4 days ago | feeling annoyed 4 days ago | gnite 1 week ago | drank water 1 week ago | gnite 1 week ago | drank diet dr pepper 1 week ago | feeling hopeful 1 week ago | drank diet coke 1 week ago | gmorning 1 week ago | gnite 1 week ago | feeling displeased 1 week ago | feeling sedated 1 week ago

EXPLORE ALL ACTIONS

Actions Recorded Past 30 Days

71



Actions by Time

Actions Recorded Past 24 Hours

0

Actions by Hour in the Day

Each bar represents an hour during the day, starting at midnight and ending 11:59pm.

your.flowingdata Tip

your.flowingdata pages let you create custom views into your data. Keep it private, or make it public to share with others. Create a page and start sharing now.

Most Interest

In the last 30 days, you have logged data for "gnite" the most.

Some Ideas for Stuff You Can Track

gmorning/gnite - when you go to sleep and wake up
feeling - how you're feeling right now
peed/pooped - when you go number 1 and 2
smoked - counting cigarettes smoked
watched - movies and television watched
read - books, magazines, and websites you read
traveled - places you've gone
exercised - when you get some physical activity
rode - driving, flying, biking, etc
played - making time for fun with games
ate - what you ate recently
drank - what you drank recently
glucose - keep your blood sugar down
weigh - the number on the dreaded scale

your.flowingdata A project by FlowingData

DEMO USER | Start Recording Your Own Data Now

This is a DEMO. Login in with Twitter to start recording your own data.

Overview // Explore All Actions: **Calendar** | Treemap | Cloud | Time Series // Actions Log

Custom Pages: Click to Select a Page

Previous Year | 2009 | Next Year // **This Year**

Actions Calendar /

January

February

March

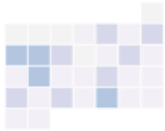
April

May

June

July

August



September

October

November

December

Saturday, July 18

drank water 10:46 p.m.
ate grapefruit sorbet 9:13 p.m.
ate pho 9:02 p.m.
watched I love you, man 7:00 p.m.
drank water 6:59 p.m.
drank diet squirt 5:31 p.m.
ate hot tamales 5:30 p.m.
ate peach cake 4:47 p.m.
watched garden state 2:31 p.m.
ate hot dog 1:47 p.m.
ate cherries 1:47 p.m.
drank diet coke 12:44 a.m.
played nba2k8 12:44 a.m.
drank water 10:27 a.m.
exercised jump rope 10:05 a.m.
drank water 9:19 a.m.
gmorning self 9:00 a.m.
gnite 1:23 a.m.
feeling tired 1:09 a.m.

Total Recorded

1,345

Actions (17)

ate	429
feeling	293
drank	185
gmorning	154
gnite	121
watched	62
chores	33
played	21
exercised	15
cooked	8
listened	7
napped	4
smoked	4
read	3
headache	3

SHOW ALL

Recent Actions

gnite 3 days ago
listened This American Life 4 days ago
ate turkey artichoke panini 4 days ago
feeling annoyed 4 days ago
gnite 1 week ago
drank water 1 week ago
gnite 1 week ago
drank diet dr pepper 1 week ago
feeling hopeful 1 week ago
drank diet coke 1 week ago
gmorning 1 week ago
gnite 1 week ago
feeling displeased 1 week ago
feeling sedated 1 week ago
played nba2k8 1 week ago

your.flowingdata A project by FlowingData

DEMO USER | Start Recording Your Own Data Now

This is a DEMO. Login in with Twitter to start recording your own data.

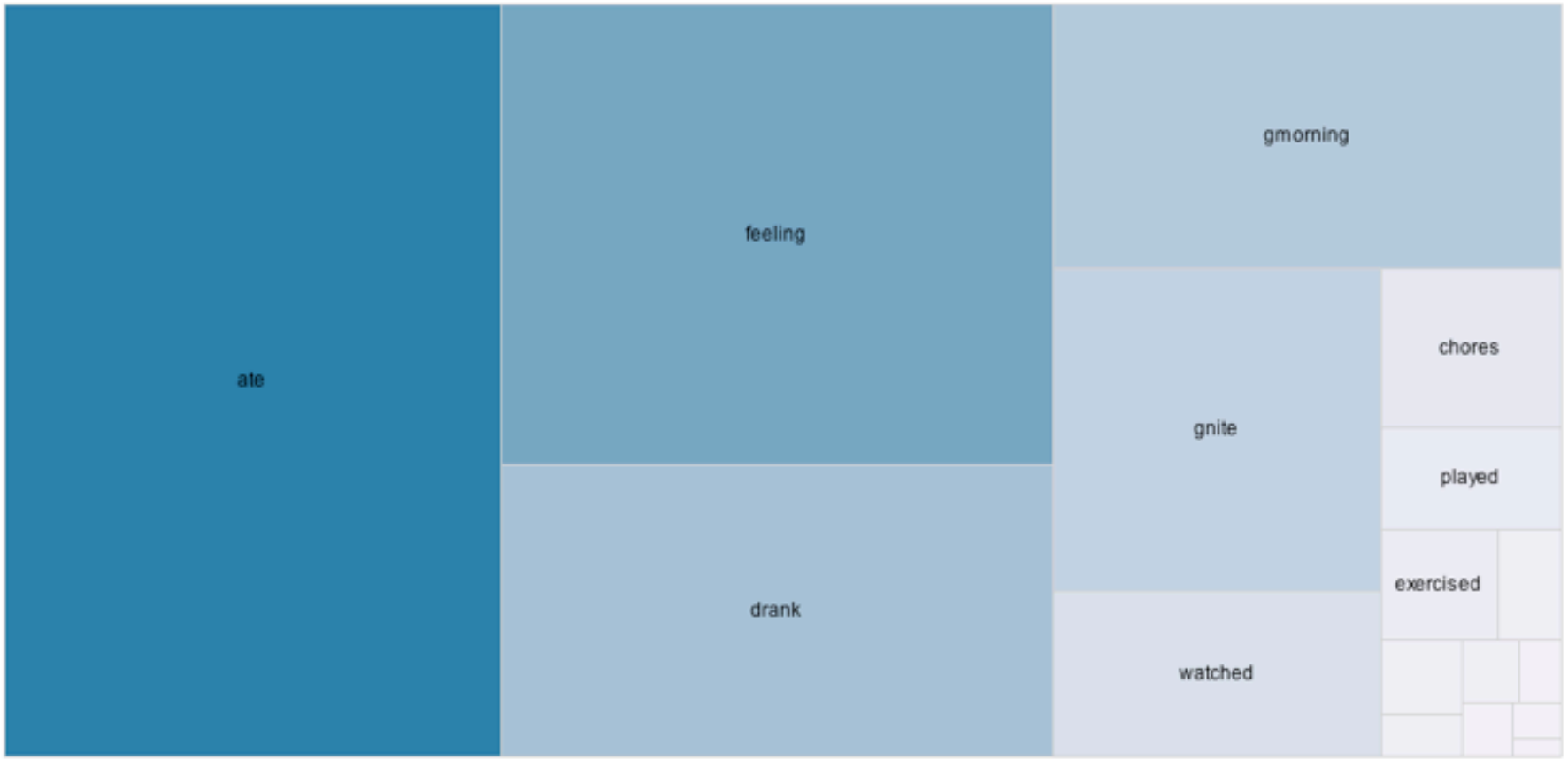
Overview // Explore All Actions: Calendar | **Treemap** | Cloud | Time Series // Actions Log

Custom Pages: Click to Select a Page

Previous Year | 2009 | Next Year // **This Year** | This Month | Today

Type a query in the search box below or click on the graph to interact.

Search:



About | Help | Forum | A project by FlowingData

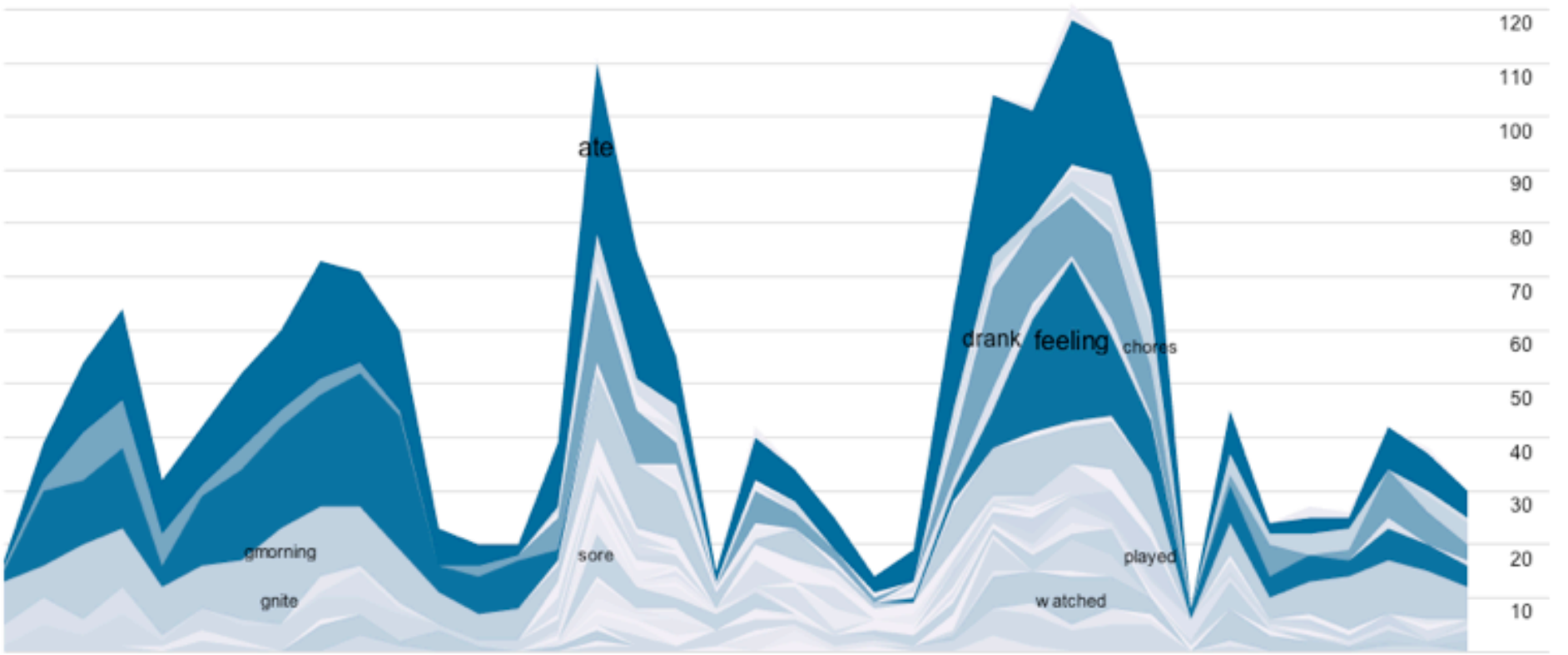
your.flowingdata

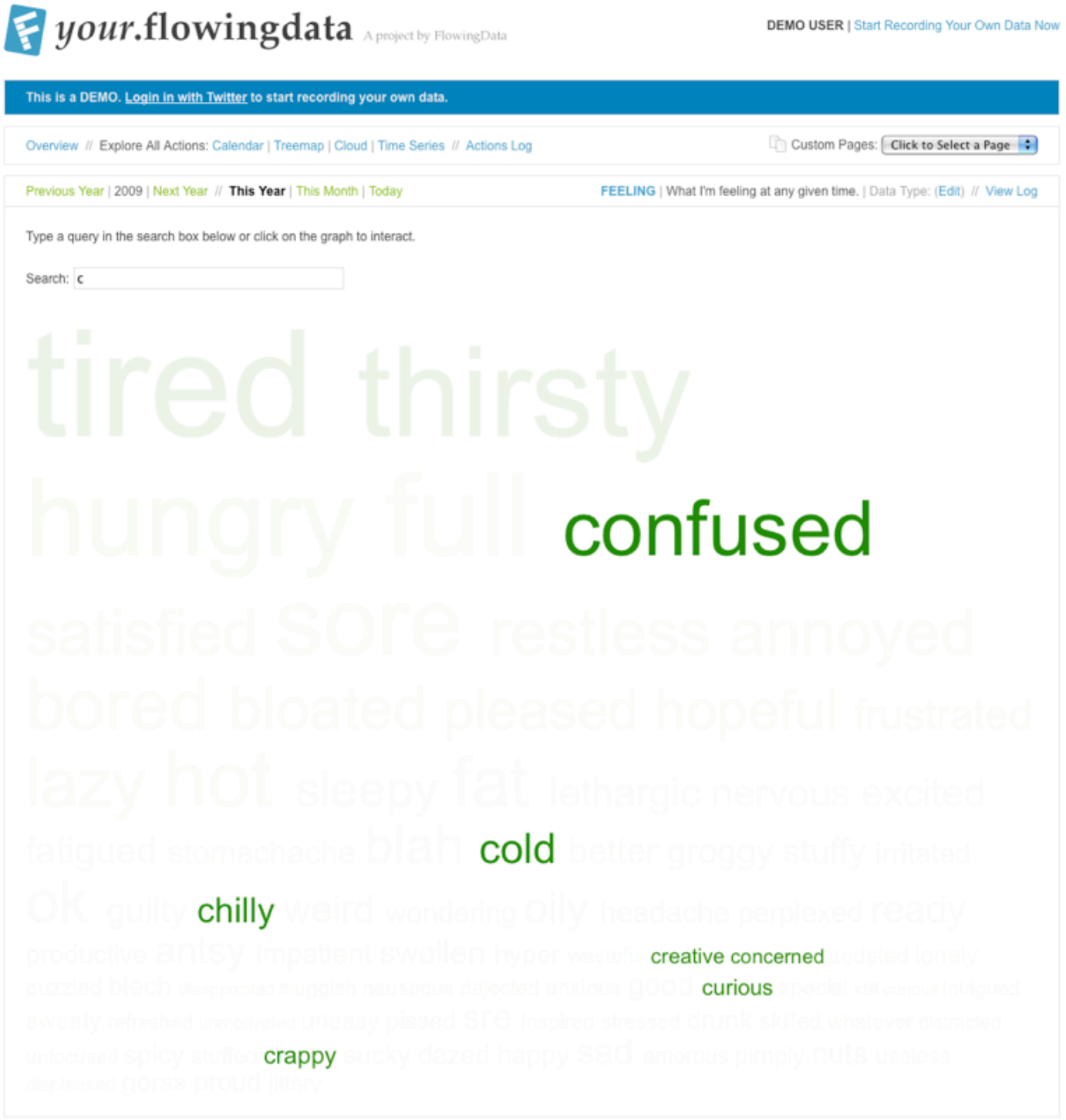
A project by FlowingData

Type a query in the search box below or click on the graph to interact.

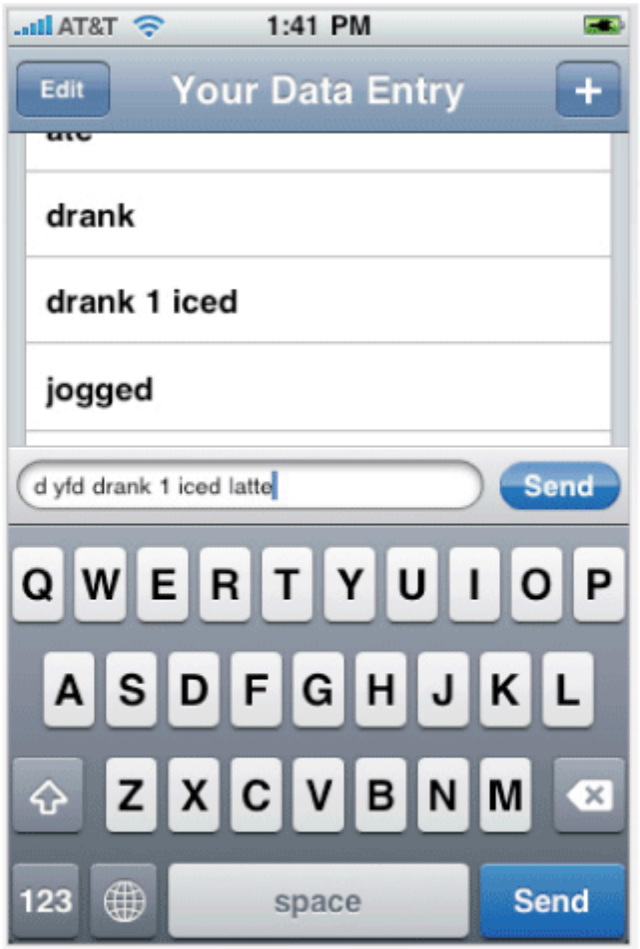
Search:

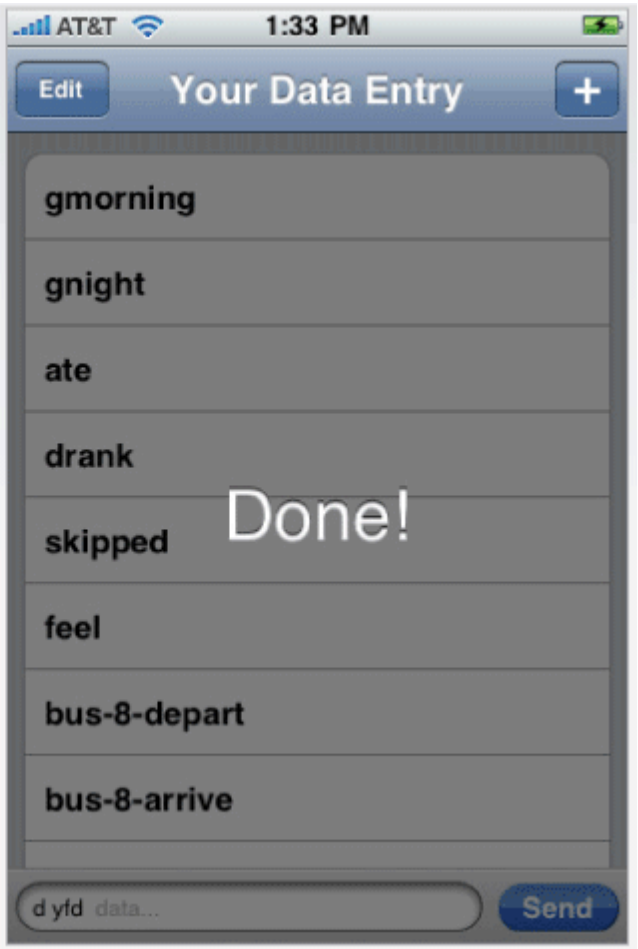
Graph by: day | week | month




your.flowingdata
A project by FlowingData
DEMO USER | Start Recording Your Own Data Now
This is a DEMO. Login in with Twitter to start recording your own data.
Overview // Explore All Actions: Calendar | Treemap | Cloud | Time Series // Actions Log
Custom Pages: Click to Select a Page
Previous Year | 2009 | Next Year // This Year | This Month | Today
FEELING | What I'm feeling at any given time. | Data Type: (Edit) // View Log
Type a query in the search box below or click on the graph to interact.
Search: c
tired thirsty
hungry full confused
satisfied sore restless annoyed
bored bloated pleased hopeful frustrated
lazy hot sleepy fat lethargic nervous excited
fatigued stomachache blah cold better groggy stuffy irritated
ok guilty chilly weird wondering oily headache perplexed ready
productive antsy impatient swollen hyper creative concerned lonely
puzzled blech curious
sweaty refreshed uneasy pissed sore inspired stressed drunk skilled
crappy sad
About | Help | Forum | A project by FlowingData


iPod
1:12 PM
Edit
Your Data Entry
+
gmorning
gnight
ate
drank
weigh
d yfd data...
Send


AT&T
1:41 PM
Edit
Your Data Entry
+
drank
drank 1 iced
jogged
d yfd drank 1 iced latte
Send
Q W E R T Y U I O P
A S D F G H J K L
Z X C V B N M
123
space
Send


AT&T
1:33 PM
Edit
Your Data Entry
+
gmorning
gnight
ate
drank
skipped
feel
bus-8-depart
bus-8-arrive
Done!
d yfd data...
Send


LG
S-CLASS USER INTERFACE

Self-experimentation as a source of new ideas: Ten examples about sleep, mood, health, and weight

Seth Roberts
Department of Psychology, University of California, Berkeley, Berkeley, CA 94720-1650.
roberts@socrates.berkeley.edu

Abstract: Little is known about how to generate plausible new scientific ideas. So it is noteworthy that 12 years of self-experimentation led to the discovery of several surprising cause-effect relationships and suggested a new theory of weight control, an unusually high rate of new ideas. The cause-effect relationships were: (1) Seeing faces in the morning on television decreased mood in the evening (>10 hrs later) and improved mood the next day (>24 hrs later), yet had no detectable effect before that (0–10 hrs later). The effect was strongest if the faces were life-sized and at a conversational distance. Travel across time zones reduced the effect for a few weeks. (2) Standing 8 hours per day reduced early awakening and made sleep more restorative, even though more standing was associated with less sleep. (3) Morning light (1 hr/day) reduced early awakening and made sleep more restorative. (4) Breakfast increased early awakening. (5) Standing and morning light together eliminated colds (upper respiratory tract infections) for more than 5 years. (6) Drinking lots of water, eating low-glycemic-index foods, and eating sushi each caused a modest weight loss. (7) Drinking unflavored fructose water caused a large weight loss that has lasted more than 1 year. While losing weight, hunger was much less than usual. Unflavored sucrose water had a similar effect. The new theory of weight control, which helped discover this effect, assumes that flavors associated with calories raise the body-fat set point. The stronger the association, the greater the increase. Between meals the set point declines. Self-experimentation lasting months or years seems to be a good way to generate plausible new ideas.

Keywords: breakfast; circadian; colds; depression; discovery; fructose; innovation; insomnia; light; obesity; sitting; standing; sugar

> Mollie: There has to be a beginning for everything, hasn't there?
> –*The Mousetrap*, Agatha Christie (1978)

1. Introduction

1.1. Missing methods

Scientists sometimes forget about idea generation. "How odd it is that anyone should not see that all observation must be for or against some view if it is to be of any service," wrote Charles Darwin to a friend (Medawar 1969, p. 11). But where did the first views come from, if not observation? According to a diagram in the excellent textbook *Statistics for Experimenters* (Box et al. 1978), the components of "data generation and data analysis in scientific investigation" (p. 4) are "deduction," "design," "new data," and so on. Scientific investigation, the diagram seems to say, begins when the scientist has a hypothesis worth testing. The book says nothing about how to obtain such a hypothesis.

It is not easy to come up with new ideas worth testing, nor is it clear how to do so. Table 1 classifies scientific methods by goal (generate ideas or test them) and time of application (before and during data collection or afterwards). The amount written about idea generation is a small fraction of the amount written about idea testing (McGuire 1997), and the amount written about what to do before and during data collection is a small fraction of the amount written about what to do afterwards – so the empty cell in Table 1, on how to collect data that generate ideas, is no surprise. Although scientific creativity has been extensively studied (e.g., Klahr & Simon 1999; Simonton 1999), this research has not yet suggested new tools or methods. Even McGuire, who listed 49 "heuristics" (p. 1) for hypothesis generation, had little to say about data gathering.

Hyman (1964) believed that "we really do not know enough about getting ideas even to speculate wisely about how to encourage fruitful research" (p. 28), but 40 years later this is not entirely true. Exploratory data analysis (Tukey 1977) helps reveal unexpected structure in data, and such discoveries often suggest new ideas worth studying. Table 1 includes only those methods useful in many areas of science, omitting methods with limited applicability

SETH ROBERTS is an Associate Professor in the Department of Psychology at the University of California at Berkeley. He is a member of the University's Center for Weight and Health. His research includes follow-up of the results reported in this article, especially the weight and mood results, and study of how things begin in another situation – the generation of new instrumental behavior by rats and pigeons.

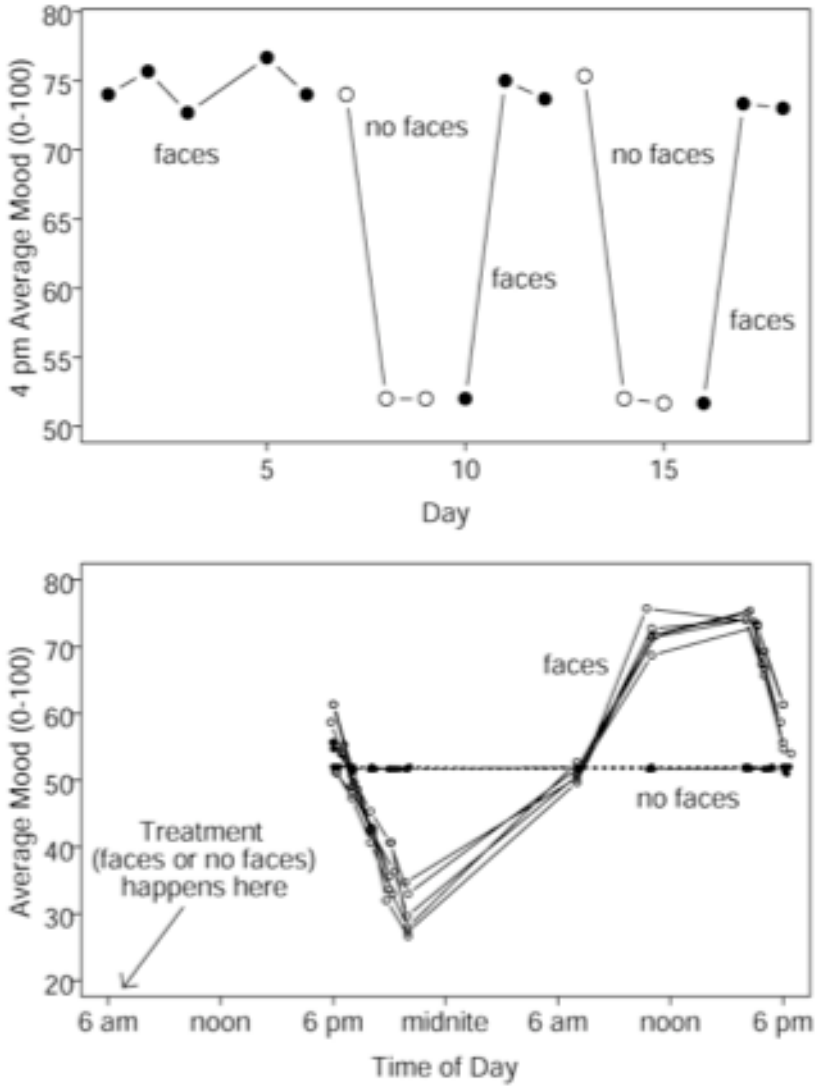


Figure 5. Mood ratings over 17 days, beginning October 13, 1999. Upper panel: Mood at 4:00 p.m. day by day. Lower panel: Time course of the effect. In both panels each point is an average of three ratings, one for each scale. The three scales measure the dimensions unhappy/happy, irritable/serene, and reluctant/eager (scale range: 5 = extremely negative, 95 = extremely positive, with 50 = neither negative nor positive). Each line is a separate series of measurements. The data start about 12 hours after the treatment because that is when the treatment began to have an effect.

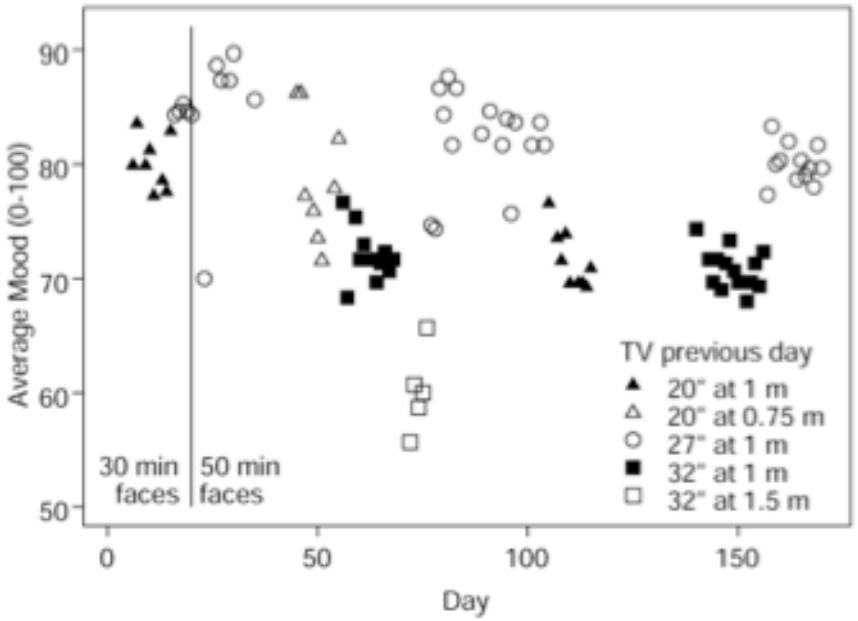


Figure 7. Effect on mood of TV size and distance. The vertical line separates results the day after viewing 30 minutes of faces from results the day after viewing 50 minutes of faces. Day 1 = November 17, 1996. Average mood = mood at 4 p.m. (average of three ratings on three scales measuring different dimensions of mood).

{ UI SKETCH }

Jane Doe | Messages (2) | Create ODL

May 1 - 31, 2009

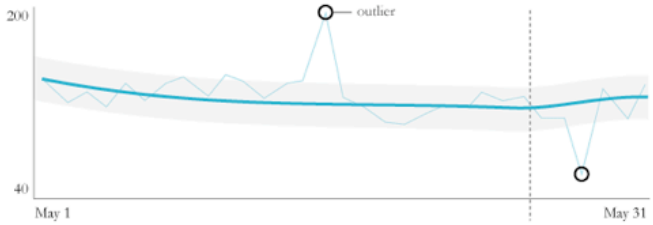
Show Current: Year **Month** Week Day select date range ▼

Calendar color-coded by hours of activity: 0 3+ No data

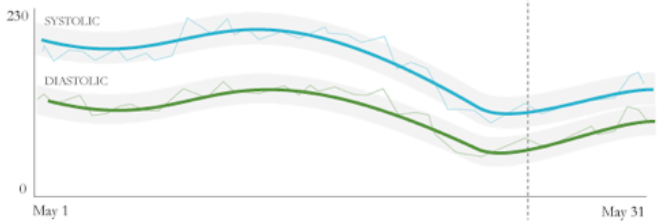
January February March April **May** June

Interventions

Glucose in mg/dL Measurements and trends for the month.



Blood Pressure Measurements and trends for the month.



Map Traces and stationary times with points of interest highlighted

Gym Fast Food Park

switch to full screen view map

Drug Prescribed

Patient prescribed new daily medication.

START DATE: May 26, 2009

AMOUNT: 200 mg

FREQUENCY: One pill twice per day, once in the morning and once in the evening.

PREV. AMT.: 150 mg
PREV. FREQ.: Same

NOTES:

Enter observations here.

{moves up and down w/ scrolling, meta data changes depending on what is selected on left}

How have my eating patterns in terms of place changed over time?

Monday

Tuesday

Wednesday

Thursday

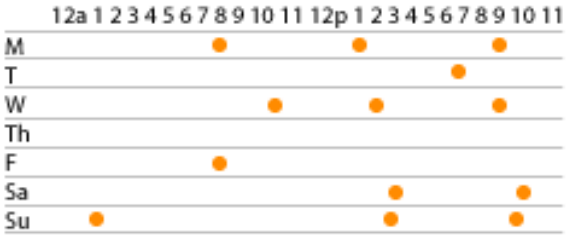
Friday

Saturday

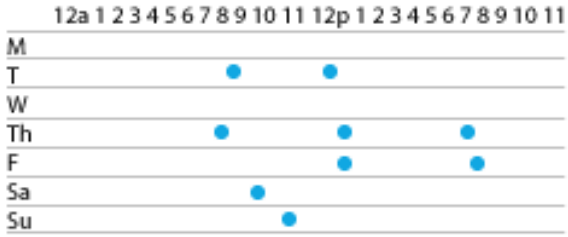
Sunday

AndWellness by CJ Cenizal

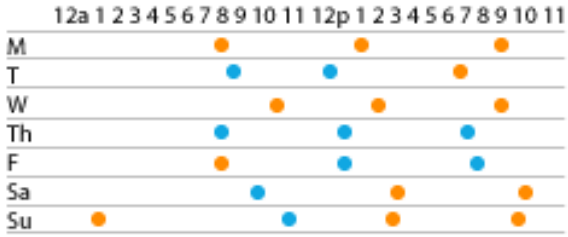
When have I eaten off plan?



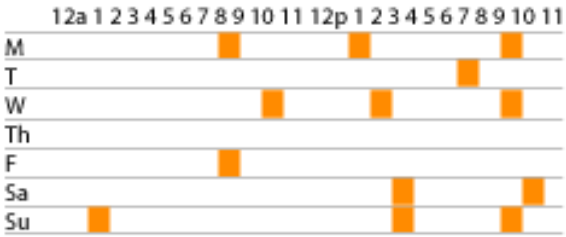
When have I eaten on plan?



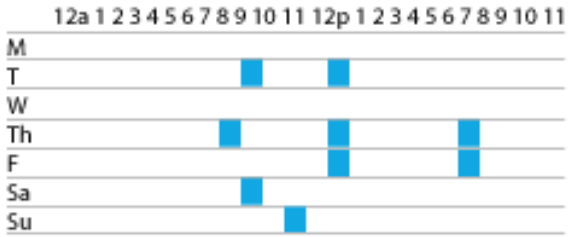
When have I eaten on and off plan?



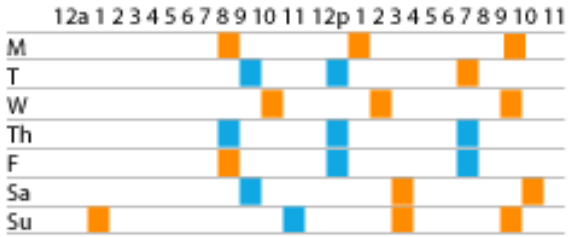
When have I eaten off plan?



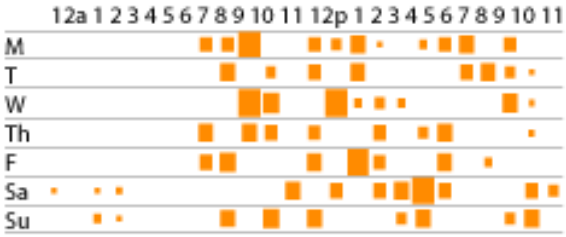
When have I eaten on plan?



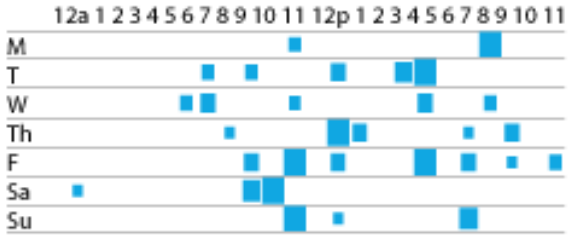
When have I eaten on and off plan?



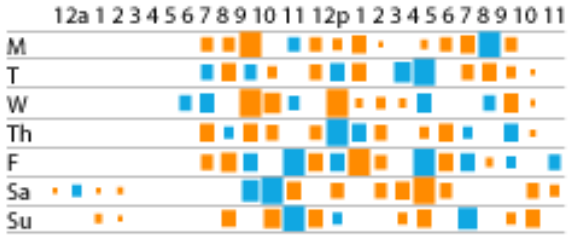
When do I tend to eat off plan?



When do I tend to eat on plan?



When do I tend to eat on/off plan?



AndWellness by CJ Cenizal

Smart phone apps may enhance epidemiological or ecological data collection

submitted by Chris Condayan on September 17, 2009
Tags: citizen, ecological, EpiCollect, epidemiology, Phones, science, Smart, technology
Source: www.plosone.org

PLoS One has published an interesting paper that considers using smart phones for scientific field data collection and suggests mobile apps could also be beneficial for recruiting 'citizen scientists' to contribute data easily to central databases through their mobile phone.

Here's the abstract:

Background

Epidemiologists and ecologists often collect data in the field and, on returning to their laboratory, enter their data into a database for further analysis. The recent introduction of mobile phones that utilise the open source Android operating system, and which include (among other features) both GPS and Google Maps, provide new opportunities for developing mobile phone applications, which in conjunction with web applications, allow two-way communication between field workers and their project databases.

Methodology

Here we describe a generic framework, consisting of mobile phone software, EpiCollect, and a web application located within www.spatialepidemiology.net. Data collected by multiple field workers can be submitted by phone, together with GPS data, to a common web database and can be displayed and analysed, along with previously collected data, using Google Maps (or Google Earth). Similarly, data from the web database can be requested and displayed on the mobile phone, again using Google Maps. Data filtering options allow the display of data submitted by the individual field workers or, for example, those data within certain values of a measured variable or a time period.

URBAN SENSING
CENS / UCLA

discovery + connected participation = change

Google™ Custom Search

vision | blog | projects | technology | resources | results | contact

URBANSENSING
CENS / UCLA

URBAN SENSING
CENS / UCLA

Dashboard

prev **September 22 - September 29, 2008**

Tell us what you think! Send feeback to peir-info@cens.ucla.edu.

COLOR TRIPS BY: Carbon Impact

STATUS

Last update from your phone received **21 hours ago**.

10,195 data points uploaded this week.

16 data points are outstanding as part of your most recent incomplete trip.

EXPLORE

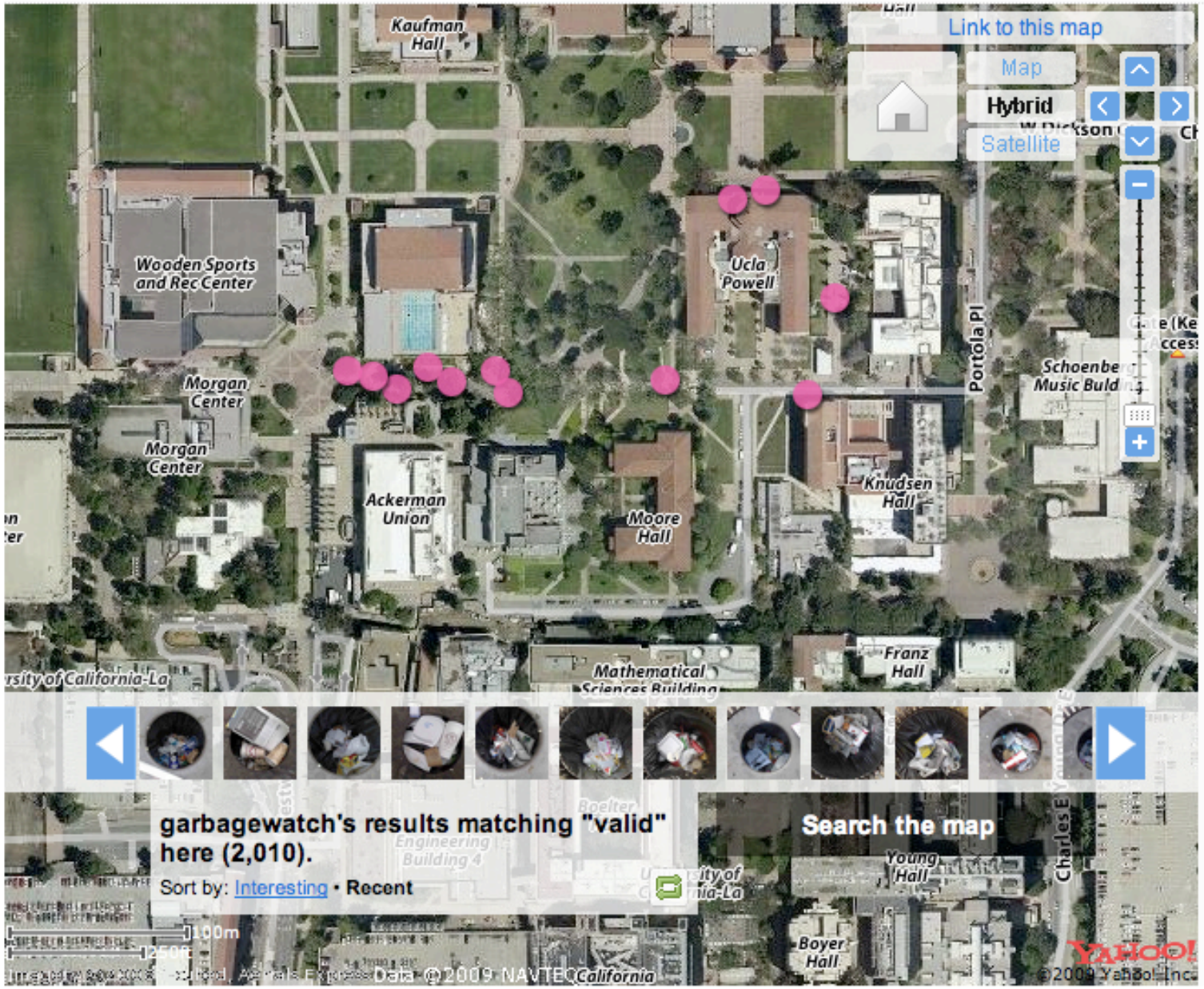
- Trip Log
- Your Network

SUPPORT

- About
- FAQ

GarbageWatch

Map | Contributions | Tag | Participate | myGW














Doritos
















WHAT'S INVASIVE! Santa Monica Mountains

Home: SAMO | Top Invasives | Data & Photos | Maps | Help / About | Login || Register

Use your Mobile Phone to Help Us Locate Invasive Plants!

Top Invasive Plants!

Santa Monica Mountains

Harding grass

Perennial Pepperweed

Poison hemlock

Spanish Broom

Terracina spurge

Yellow Starthistle

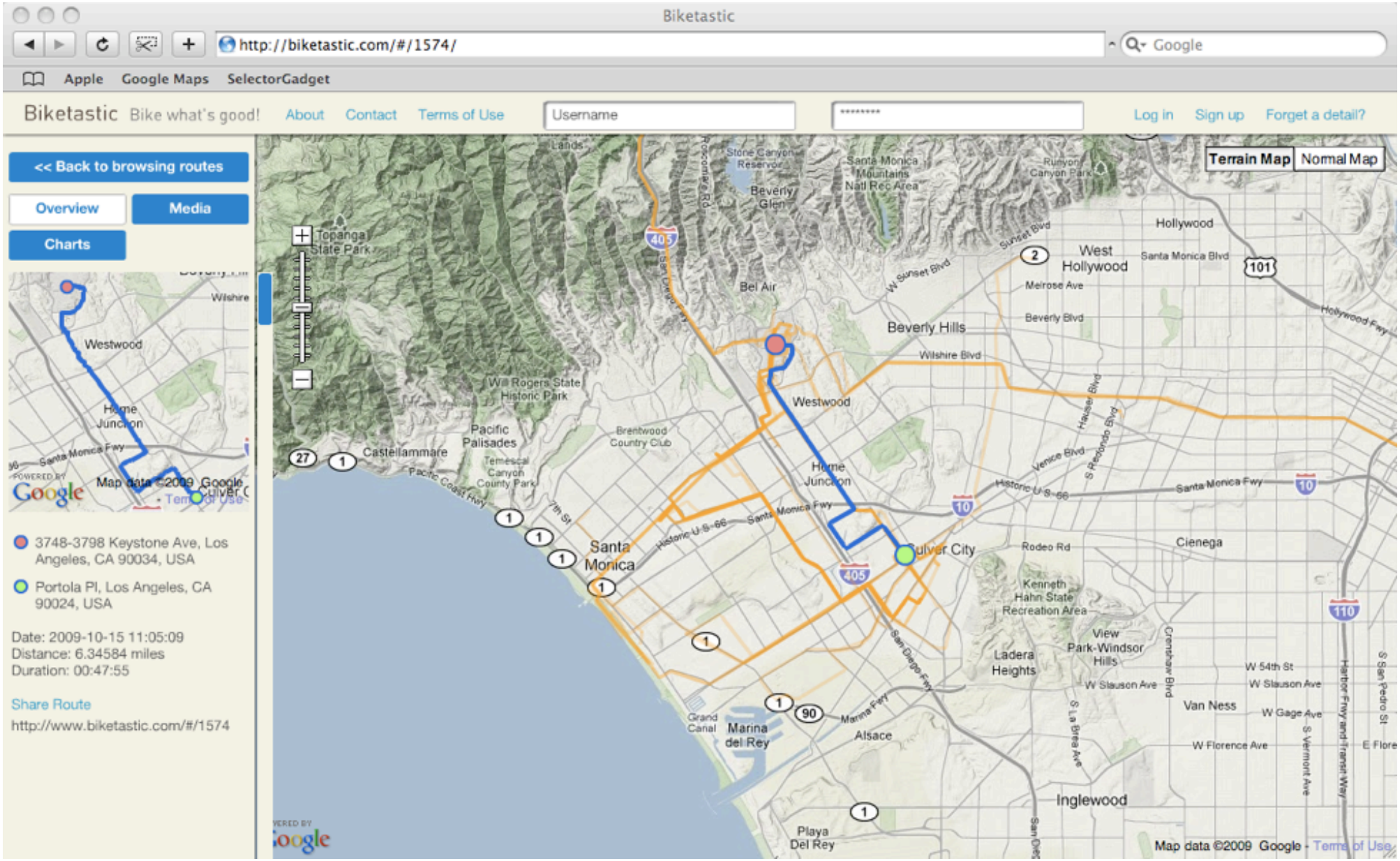
36 People Contributing in Santa Monica!
1194 Invasive Weeds Found!

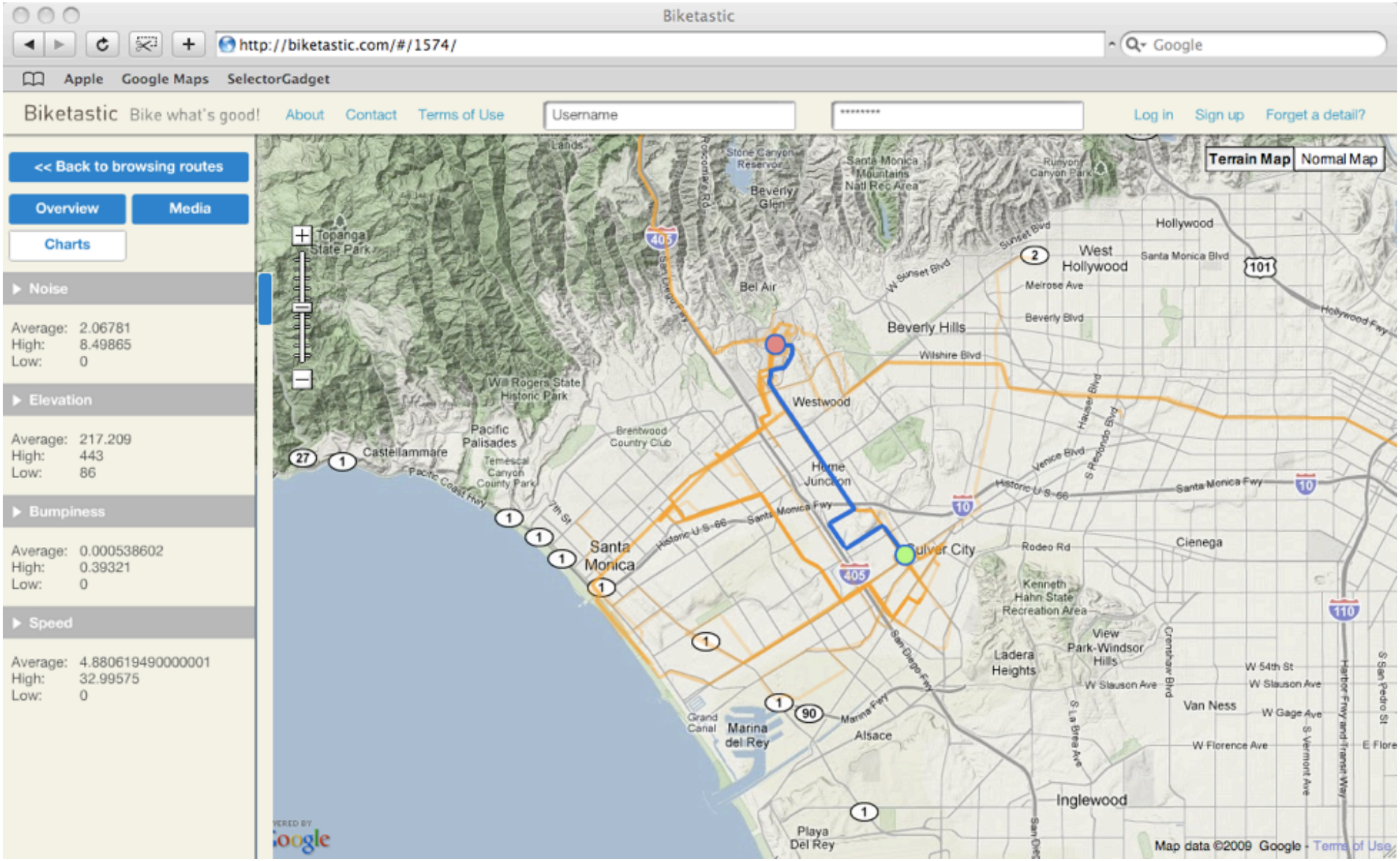
Weed of the week: *Spanish Broom*

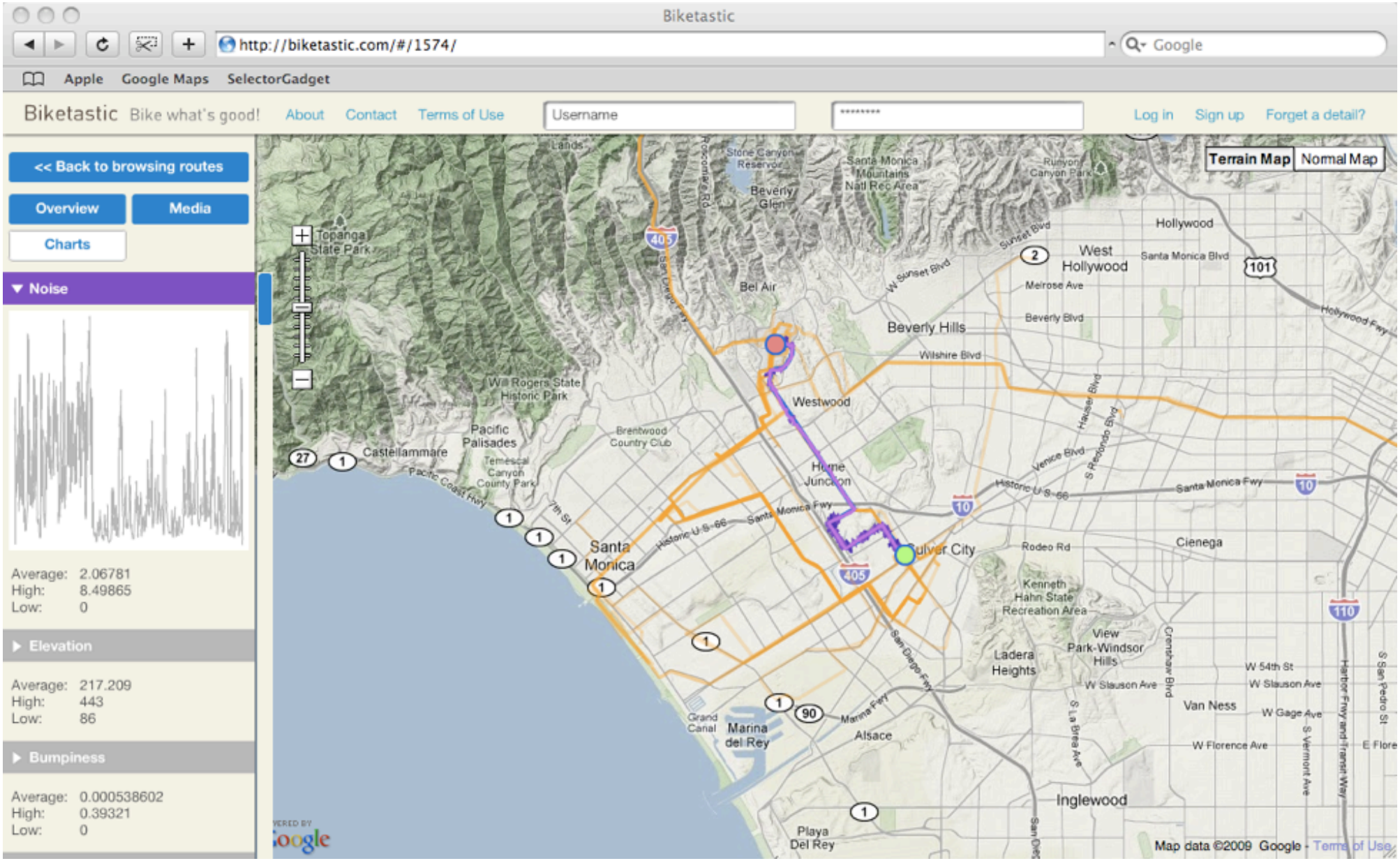
Spanish broom is a perennial shrub, 6 to 15 ft tall. Stems are green, cylindrical, no angled (rush-like). Leaves are small, less than half an inch, shed during summer drought - giving plant a stick-like appearance. Flowers are yello, pea-like and fragrant, about an inch in size. Flowers from early spring to the beginning of summer. Fruit is pod shaped, 2 to 4 inches long, generally appear starting in June or July. More

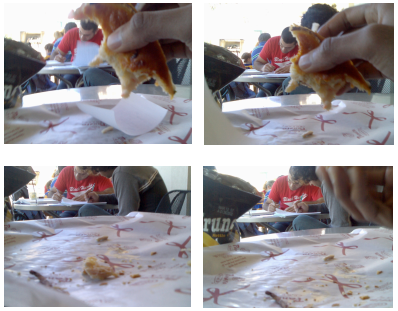
Master Gardener: Invasive plants harming area's native biodiversity
Marin Independent-Journal
Of the 29 highest-priority invasive **species** at PRNS (included in a draft **invasive plant** plan), 14 are escaped ornamental garden plants. ...
See all stories on this topic

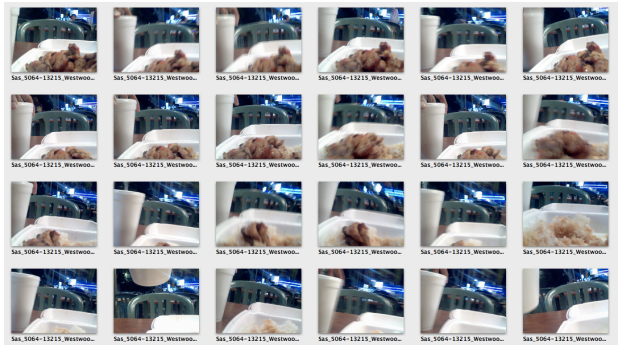
UCLA CENS


Biketastic
http://biketastic.com/#/1574/
Google
Apple
Google Maps
SelectorGadget
Biketastic
Bike what's good!
About
Contact
Terms of Use
Username
Log in
Sign up
Forget a detail?
<< Back to browsing routes
Overview
Media
Charts
Terrain Map
Normal Map
3748-3798 Keystone Ave, Los Angeles, CA 90034, USA
Portola Pl, Los Angeles, CA 90024, USA
Date: 2009-10-15 11:05:09
Distance: 6.34584 miles
Duration: 00:47:55
Share Route
http://www.biketastic.com/#/1574
Topanga State Park
Will Rogers State Historic Park
Pacific Palisades
Castellammare
Temescal Canyon County Park
Brentwood Country Club
Bel Air
Beverly Glen
Stone Canyon Reservoir
Santa Monica Mountains Natl Rec Area
Beverly Hills
West Hollywood
Hollywood
Westwood
Home Junction
Santa Monica
Culver City
Cienega
Rodeo Rd
Kenneth Hahn State Recreation Area
Ladera Heights
View Park-Windsor Hills
Van Ness
Inglewood
Marina del Rey
Grand Canal
Alsace
Playa Del Rey
Map data ©2009 Google - Terms of Use

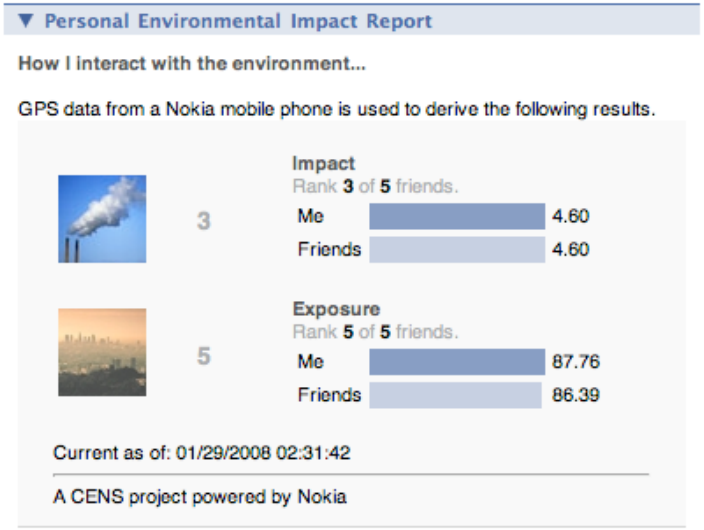

Biketastic
http://biketastic.com/#/1574/
Google
Apple
Google Maps
SelectorGadget
Biketastic
Bike what's good!
About
Contact
Terms of Use
Username
Log in
Sign up
Forget a detail?
<< Back to browsing routes
Overview
Media
Charts
Noise
Average: 2.06781
High: 8.49865
Low: 0
Elevation
Average: 217.209
High: 443
Low: 86
Bumpiness
Average: 0.000538602
High: 0.39321
Low: 0
Speed
Average: 4.880619490000001
High: 32.99575
Low: 0
Terrain Map
Normal Map
Topanga State Park
Will Rogers State Historic Park
Pacific Palisades
Castellammare
Temescal Canyon County Park
Brentwood Country Club
Bel Air
Beverly Glen
Stone Canyon Reservoir
Santa Monica Mountains Natl Rec Area
Beverly Hills
West Hollywood
Hollywood
Westwood
Home Junction
Culver City
Santa Monica
Marina del Rey
Grand Canal
Playa Del Rey
Alsace
Ladera Heights
View Park-Windsor Hills
Van Ness
Inglewood
Cienega
Kenneth Hahn State Recreation Area
Rodeo Rd
Melrose Ave
Beverly Blvd
Wilshire Blvd
Sunset Blvd
Santa Monica Blvd
Venice Blvd
Historic U.S. 66
Santa Monica Fwy
Marina Fwy
San Diego Fwy
Hollywood Fwy
Harbor Fwy and Transit Way
Crenshaw Blvd
W 54th St
W Slauson Ave
W Gage Ave
W Florence Ave
S La Brea Ave
S Vermont Ave
S San Pedro St
Map data ©2009 Google - Terms of Use


Biketastic
http://biketastic.com/#/1574/
Google
Apple
Google Maps
SelectorGadget
Biketastic
Bike what's good!
About
Contact
Terms of Use
Username
Log in
Sign up
Forget a detail?
<< Back to browsing routes
Overview
Media
Charts
Noise
Average: 2.06781
High: 8.49865
Low: 0
Elevation
Average: 217.209
High: 443
Low: 86
Bumpiness
Average: 0.000538602
High: 0.39321
Low: 0
Terrain Map
Normal Map
Topanga State Park
Will Rogers State Historic Park
Pacific Palisades
Castellammare
Temescal Canyon County Park
Brentwood Country Club
Bel Air
Beverly Glen
Stone Canyon Reservoir
Santa Monica Mountains Natl Rec Area
Runyon Canyon Park
Hollywood
West Hollywood
Beverly Hills
Westwood
Culver City
Santa Monica
Marina del Rey
Playa Del Rey
Inglewood
Cienega
Ladera Heights
View Park-Windsor Hills
Van Ness
Alsace
Grand Canal
Kenneth Hahn State Recreation Area
Sunset Blvd
Wilshire Blvd
Beverly Blvd
Melrose Ave
Santa Monica Blvd
Rodeo Rd
Venice Blvd
Historic U.S. 66
Santa Monica Fwy
Pacific Coast Hwy
7th St
San Diego Fwy
Marina Fwy
Hollywood Fwy
Harbor Fwy and Transit Way
Crenshaw Blvd
W 54th St
W Slauson Ave
W Gage Ave
W Florence Ave
S La Brea Ave
S Vermont Ave
S San Pedro St
Map data ©2009 Google - Terms of Use



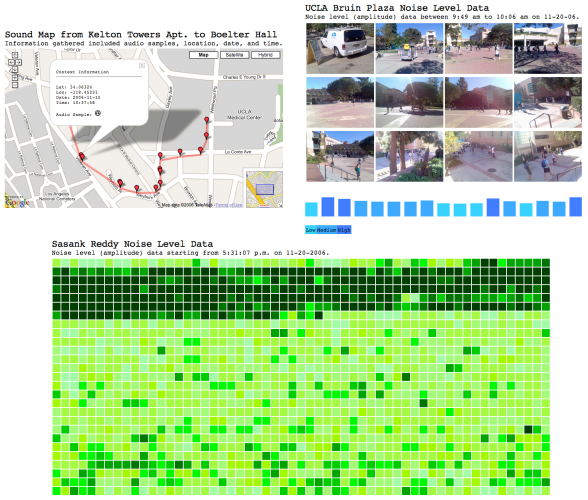


Diet Monitoring



Personal Environment
Impact Report

Walkability Study



Noise Mapping

Neighborhood
Asset Mapping

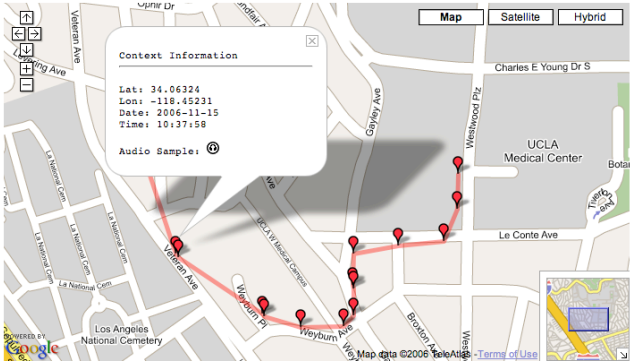
Walkability Studies

Sasank Reddy

Noise Mapping

Sound Map from Kelton Towers Apt. to Boelter Hall

Information gathered included audio samples, location, date, and time.



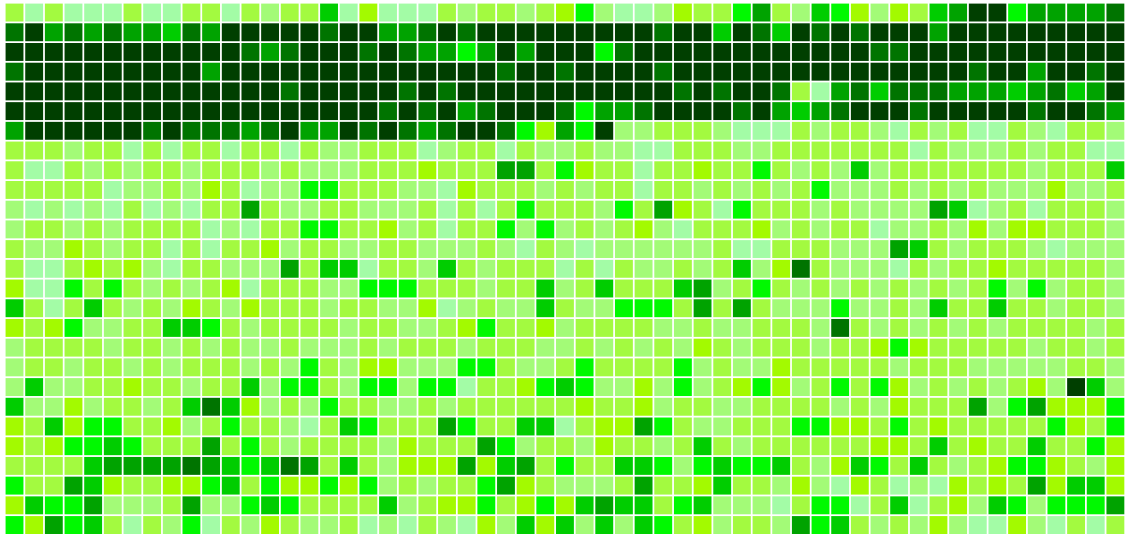
UCLA Bruin Plaza Noise Level Data

Noise level (amplitude) data between 9:49 am to 10:06 am on 11-20-06.

Low Medium High

Sasank Reddy Noise Level Data

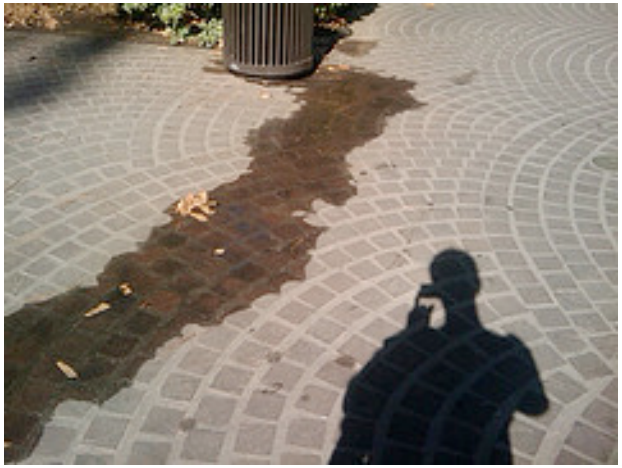
Noise level (amplitude) data starting from 5:31:07 p.m. on 11-20-2006.

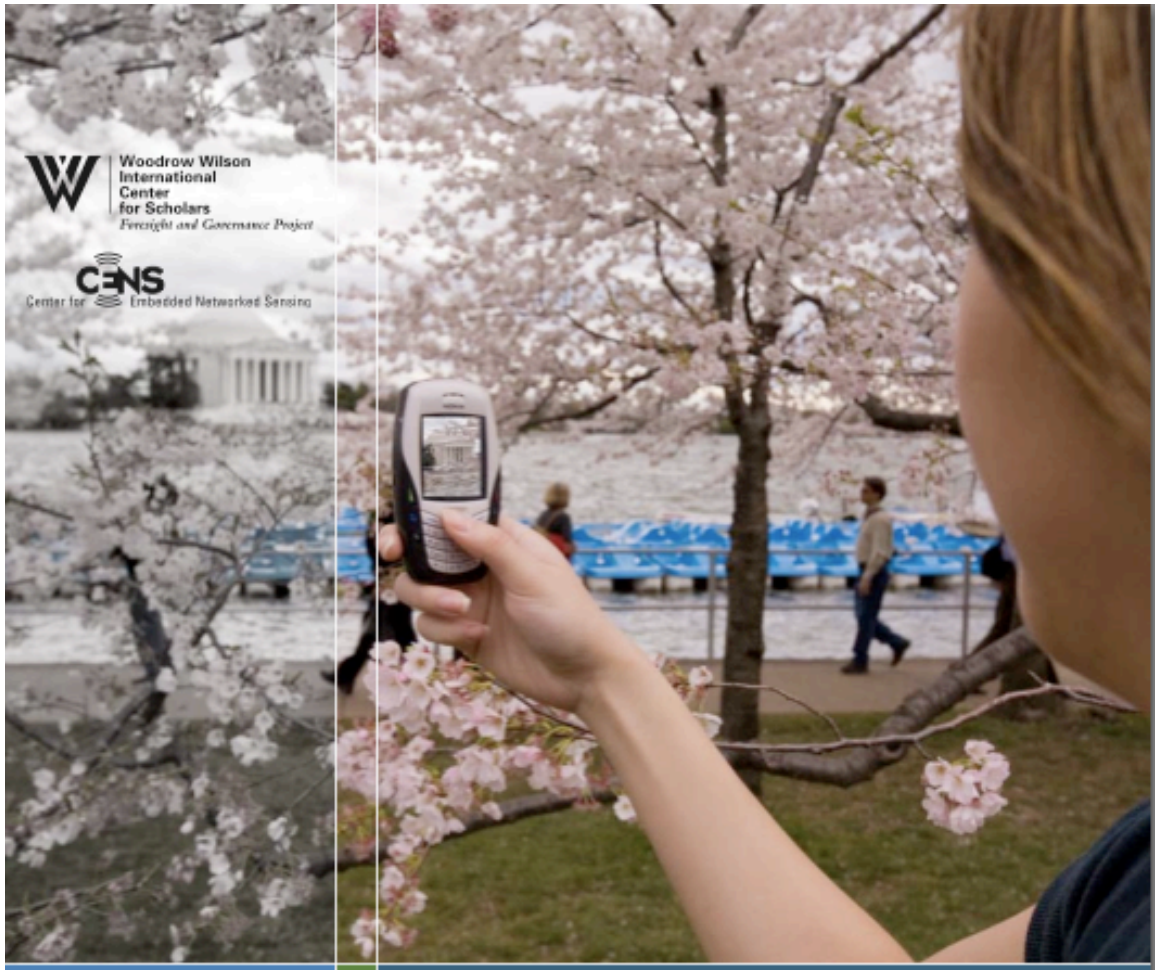


Sasank Reddy

WATERBUSTERS

Its your water





Woodrow Wilson International Center for Scholars
Foresight and Governance Project

CENS
Center for Embedded Networked Sensing

MAY 2009

WHITE PAPER

Participatory Sensing

A citizen-powered approach to illuminating the patterns that shape our world

Authored by: Jeffrey Goldman, Katie Shilton, Jeff Burke, Deborah Estrin, Mark Hansen, Nithya Ramanathan, Sasank Reddy, Vids Samanta, Mani Srivastava, and Ruth West

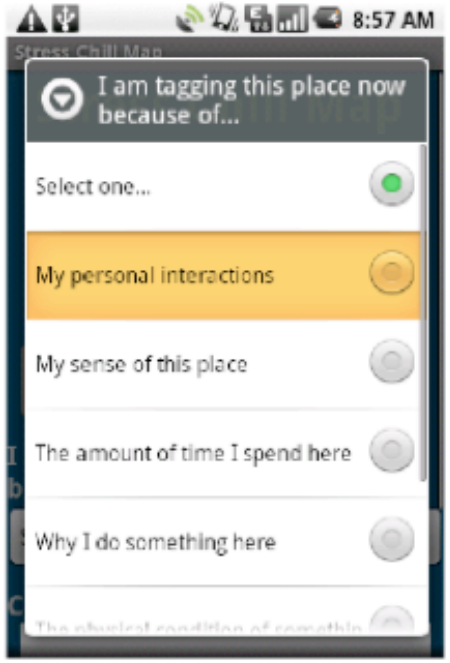
Background on Participatory Sensing (end)

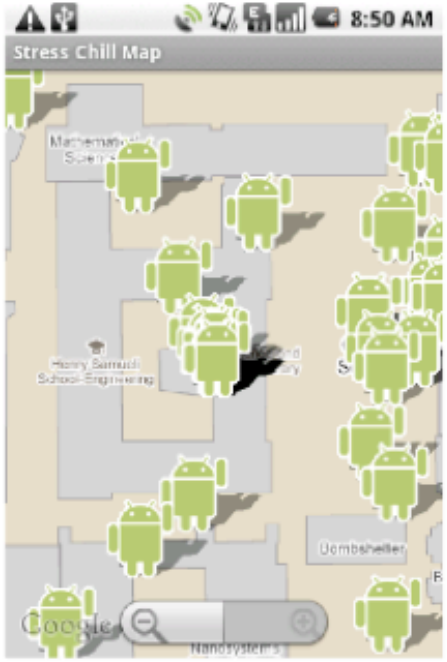
Unit 6:

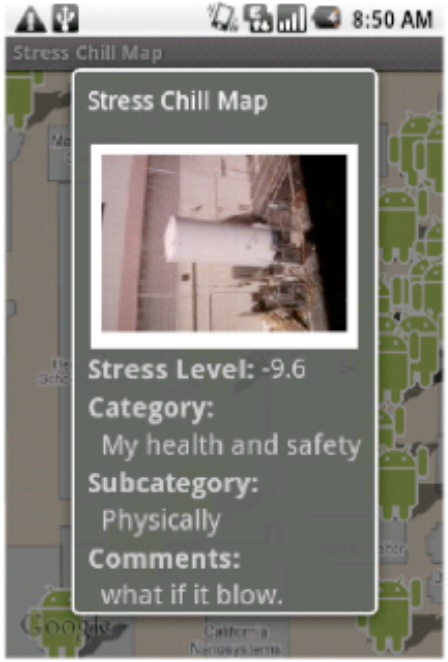
Participatory Urban Sensing

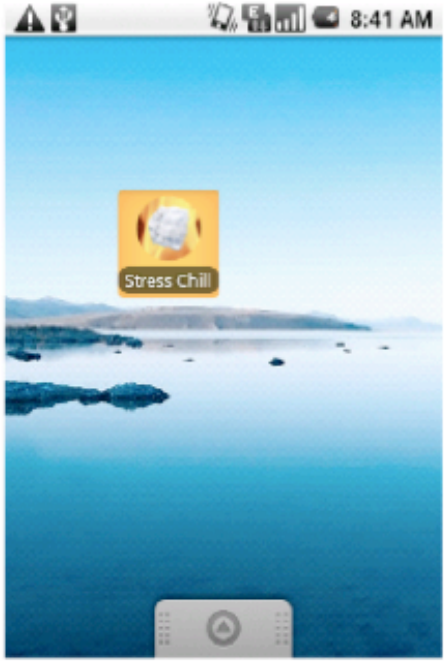

8:56 AM
Stress Chill Map
Stress Chill Map
Take a Picture
Stress 5.8 Chill
I am tagging this place now because of...
Select one...
Comments:


8:47 AM
Stress Chill Map
I am tagging this place now because of...
My personal interactions
My personal interactions here and now
Family
Comments:
Upload


8:57 AM
I am tagging this place now because of...
Select one...
My personal interactions
My sense of this place
The amount of time I spend here
Why I do something here


8:50 AM
Stress Chill Map


8:50 AM
Stress Chill Map
Stress Chill Map
Stress Level: -9.6
Category:
My health and safety
Subcategory:
Physically
Comments:
what if it blow.


8:41 AM
Stress Chill

Overview

With the Stress/Chill application, the students are able to make timestamped and geo-tagged observations consisting of one (essentially) continuous variable, one categorical variable, an image and text

At the end of the unit, the students will be asked to tell some kind of story using these data, suggesting the need to treat **time**, **space**, **images** and **text** as objects that can be manipulated and analyzed by a computer

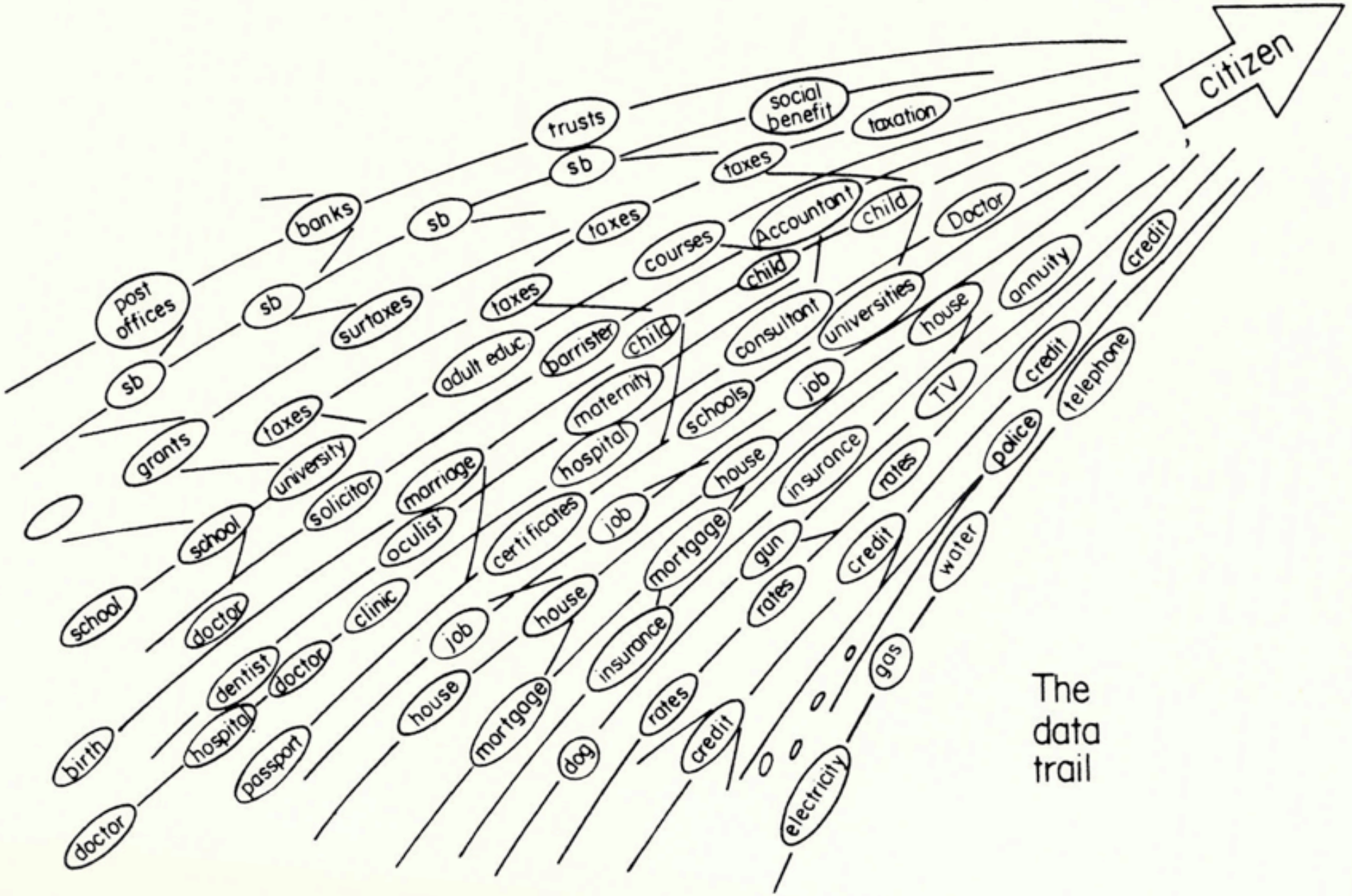
To achieve some depth, this story telling will inevitably lead the students to **outside data sources to help explain patterns** they see or to simply provide context for their discussions

Overview

In a traditional statistics class, we might present continuous and categorical data types and build up **a framework based largely in probability** theory for making statements about data

We avoided this waltz with probability, building instead on **the computational tools the students have been acquiring** all term -- Leaving mathematics behind turned out to be incredibly freeing and let us examine topics never discussed in an introductory class

In a sense, the pre-Mobilize curriculum is a first draft of something akin to a unit in **“data science”** -- We envision **a hybrid of statistical and computational thinking** that develops both the functional aspects of the tools students are learning as well as a rich critical component



From *Platform for Change* by Stafford Beer

Overview

Much of the critical piece we cover has emerged directly from **our CENS experience** in traditional embedded sensing as well as the more recent projects in participatory sensing

For example, early on, students learn about the **social implications of data** collection (some of which we have seen this morning already); they see that there is not “natural” view of a data set, but that **each representation both hides as well as reveals** information; and finally, they come to see that **data are, necessarily, imperfect records of reality**, depending heavily on the “measuring” device

All along, students are invited to question **why things look as they do**, from data formats to their computing languages

Overview

From a practical perspective, we felt that we had to prepare students for the **world of data that has emerged in the last five years** -- Their experience of computer mediated technologies is no longer the “search and find” model of the web from 1995

Instead, **data find the students** -- Realtime processes are exposed as services, with self-describing, humanly readable data formats fueling a new age of interoperability

In addition, more and more organizations are taking up **the banner of transparency and making their data available** to the public -- Learning to access, combine and manipulate varied data sources is practically **a demand for effective citizenship**

Overview: Software choice

While many of you are familiar with Python, we felt that its **overhead for effective story telling with data** was simply to great -- The data structures and subsequent programmatic techniques for accessing, combining and manipulating data were more about Python than the underlying information

Know that **we tried to make use of Python**, but in the end, we opted for a language that is vastly more expressive in terms of operations on and graphics from data -- **R is an open source tool** that has its beginnings at Bell Labs around the time Unix was being developed

It has a huge following among the “data science” crowd and is an incredibly active project -- The programming style (from **command line to small scripts to programs**) is very similar to Python, but the objects and methods for working with data are much more advanced

Overview: The lessons

With all that as background, let's consider the lessons -- We start with **simple data sets** that let us focus on **the syntax of the R language**, but over time we move to more and more complex data types

In the context of two surveys, we start with **"traditional" statistical data** like categorical and continuous variables -- We motivate graphical representations (bar plots, mosaic plots, histograms, box plots and scatter plots) as well as numerical summaries (mean, median, and quantiles)

Students learn to **ask questions of a data set** (How many cases? What are the variables? What kind of measure is each variable? Which cases are extreme? What do the distributions of the variables look like?), an approach that they will repeat even when the data become more complex

Overview: The lessons

1. Surveys (CDC, Public Agenda)

2. How a computer represents **time**, understanding temporal processes
   (Web access logs from each school)

3. How a computer represents **location**, understanding spatial processes
   (LA Bike Count, LA Metro, US Census, USGS Did You Feel it?)

4. (Optional) **Gridded** spatial data
   (CIESIN Gridded World Population Map)

5. **Images** as computational objects
   (CENS database of stills from web cams in Glacier National Park, Big Sur and Filmore, CA)

Overview: The lessons (cont.)

6. **Web services**, reinforcing lessons on time space and introducing text
   (Twitter scrapes with mentions of expressions about seasons)

7. (Optional) The hue circle and special spaces for plotting data, **clustering**
   (Google Image API)

8. (Optional) The **geometry** of high-dimensional data, multi-dimensional scaling
   (Online distance calculator, survey data revisited)

9. **Text** as a computational object
   (Presidential biographies from whitehouse.gov and the TREC Spam data set)

Section to be skipped: Tech motivation (start)

Approaching the introductory statistics course

In terms of background, it had been three years since I last taught this class, and in the meantime I had been rethinking the role that **computing and data technologies should play at the undergraduate level**

1. In 2005 and 2006 we hosted a **summer program for undergraduates** that emphasized computing and visualization

2. CENS introduced a core research focus in **participatory urban sensing**; this emphasized data collection and analysis by the public

3. Together with Co-PIs from Geography, Computer Science and Information Studies, we mapped out a field **"Data Science"** for the NSF IGERT call

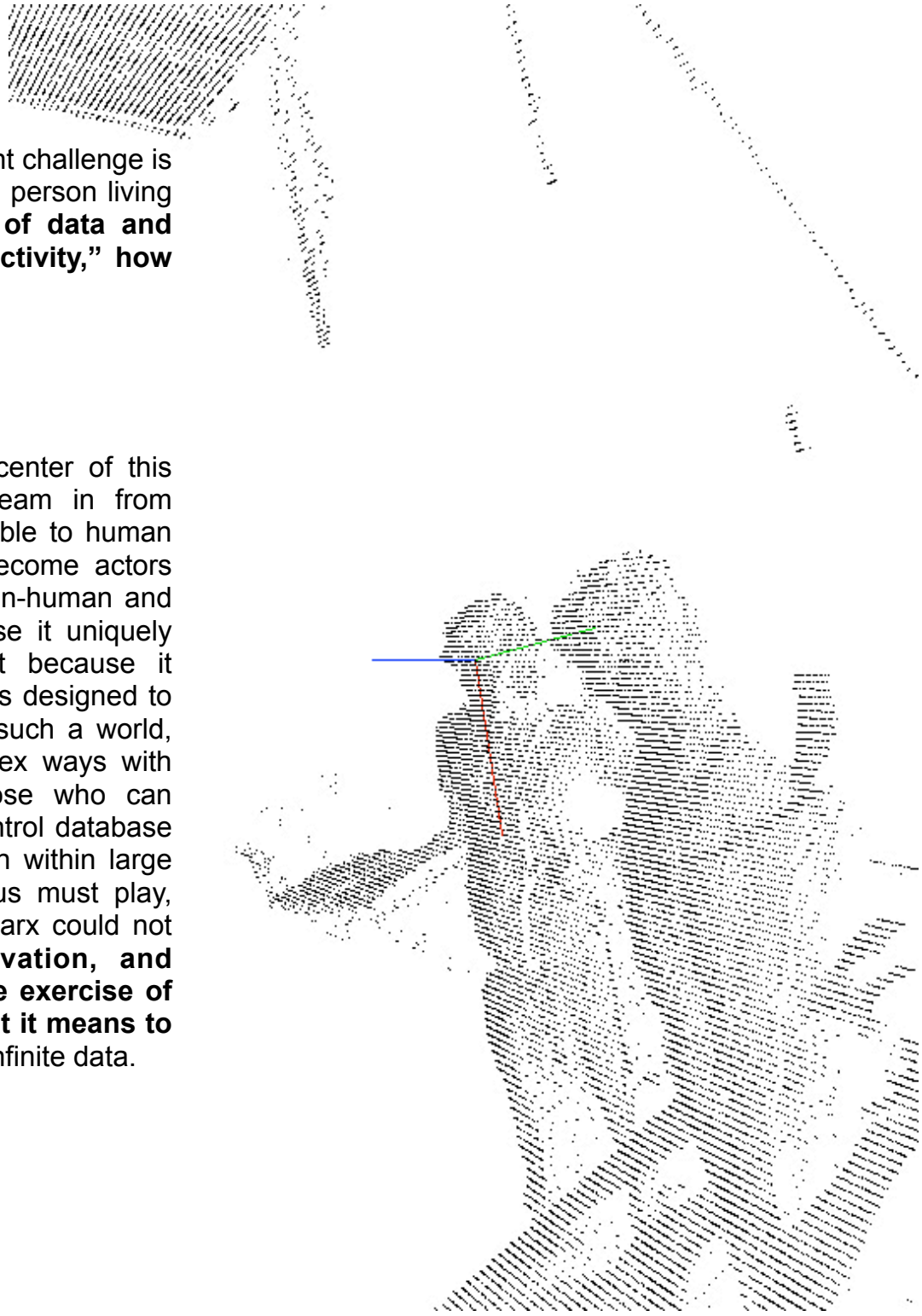
4. In my graduate courses (seminars, Statistics 202a) I refined a story in which the fate of statistics **was tied to that of information technologies**

The more interesting and at the end maybe more important challenge is how to represent the personal subjective experience of a person living in a data society. **If daily interaction with volumes of data and numerous messages is part of our new "data-subjectivity," how can we represent this experience in new ways?**

Lev Manovich , The Anti-Sublime Ideal in Data Art

...The human cannot reasonably be imagined as the center of this distributed networked system. Rather, data flows stream in from everywhere, and the vast majority of them remain invisible to human perception and direct control. Humans in this world become actors among many different kinds of agents, most of them non-human and non-biological. Human intelligence is prized not because it uniquely distinguishes us from other (biological) species, but because it occupies a niche within a complex ecology of interactions designed to optimize its capabilities and minimize its limitations. In such a world, technological expertise (which correlates only in complex ways with economic class) assumes increased importance. Those who can interrogate databases, and even more so those who control database design and the standards determining how they function within large network systems, set the rules by which the rest of us must play, whether we are conscious of it or not. In a way that Marx could not have imagined, data, capital, **technological innovation, and information flows are setting new parameters for the exercise of power and control, and new landscapes defining what it means to be human.** This is the real fear, and the real promise, of infinite data.

N. Katherine Hayles
Reality Mining, RFIDs and Real Fears about Infinite Data



Approaching introductory statistics

In each case, I rediscovered my love for statistics and the important **social-political-scientific-technical** roles that our discipline plays as the acknowledged science of data

It's almost a cliche at this point to say that **the success of statistics as a discipline depends on our ability to compute**; the corollary, however, being that our practice should make more explicit our dependence on information technologies and emphasize the need to continually track new programming languages and paradigms, new database technologies, new data formats

In terms of teaching, perhaps this sums things up...

Today, software and hardware together provide far more powerful factories than most statisticians realize, factories that many of today's most able young people find exciting and worth learning about on their own. Their interest can help us greatly, if statistics starts to make much more nearly adequate use of the computer. However, **if we fail to expand our uses, their interest in computers can cost us many of our best recruits, and set us back many years.**

Today, software and hardware together provide far more powerful factories than most statisticians realize, factories that many of today's most able young people find exciting and worth learning about on their own. Their interest can help us greatly, if statistics starts to make much more nearly adequate use of the computer. However, **if we fail to expand our uses, their interest in computers can cost us many of our best recruits, and set us back many years**.

John W. Tukey, The Technical Tools of Statistics, **1964**

Approaching introductory statistics

And with all that as background, I cracked open my copy of Samuels and Witmer, the book that I had helped to select a few years back

Let’s have a look...

Contents

CONTENTS

I INTRODUCTORY

II DIAGRAMS

III DISTRIBUTIONS

IV TESTS OF GOODNESS OF FIT, INDEPENDENCE AND HOMOGENEITY; WITH TABLE OF χ^2

V TESTS OF SIGNIFICANCE OF *MEANS*, DIFFERENCES OF MEANS, AND REGRESSION COEFFICIENTS

VI THE CORRELATION COEFFICIENT

VII INTRACLASS CORRELATIONS AND THE ANALYSIS OF VARIANCE

A little retro

Fisher's **Statistical Methods for Research Workers** was first published in **1925** and the table of contents is remarkably similar to the text I had been using for the previous two years of Statistics 13

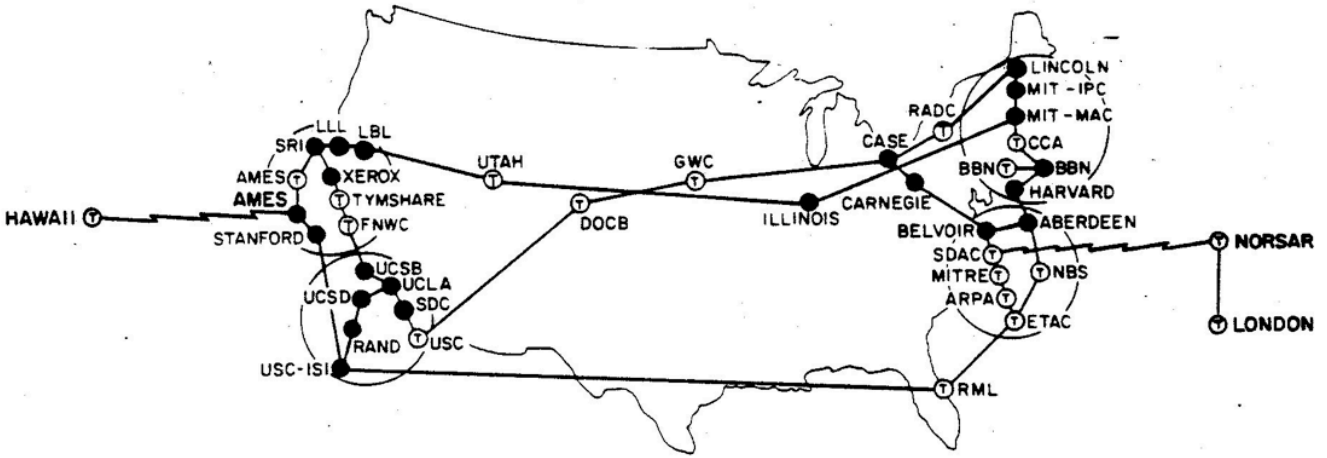
To be fair, my scan of SMRW was from the most recent edition which appeared in 1973; although that's still over 30 years old, perhaps there's nothing very different about statistics since then

Let's look back - first at the state of computing in 1973...

1972-1973

The first pocket calculator hits the market;
Texas instruments and HP will follow suit

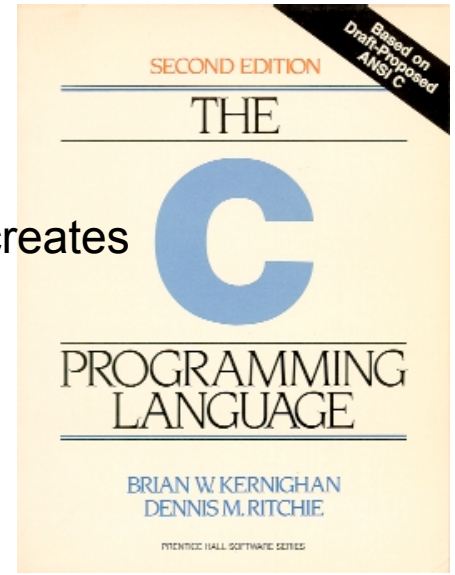
Xerox builds a personal computer (mouse, ethernet, GUI) but it's too expensive for the general public

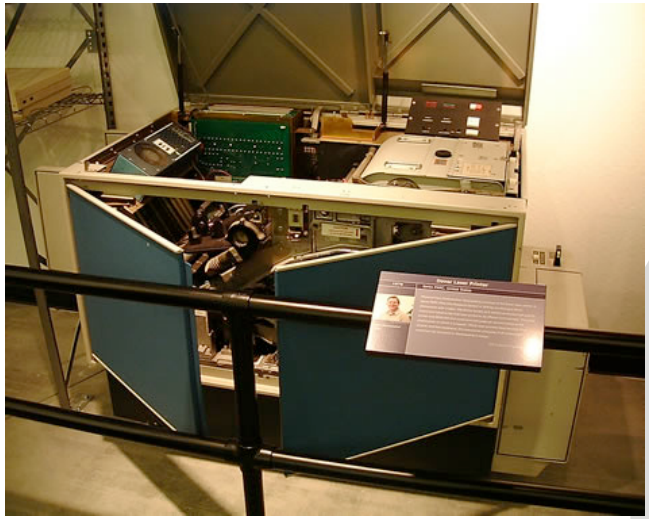


The ARPAnet connects 40 sites, the Internet consists of 25 computers

1972-1973

Dennis Richie at Bell Labs creates

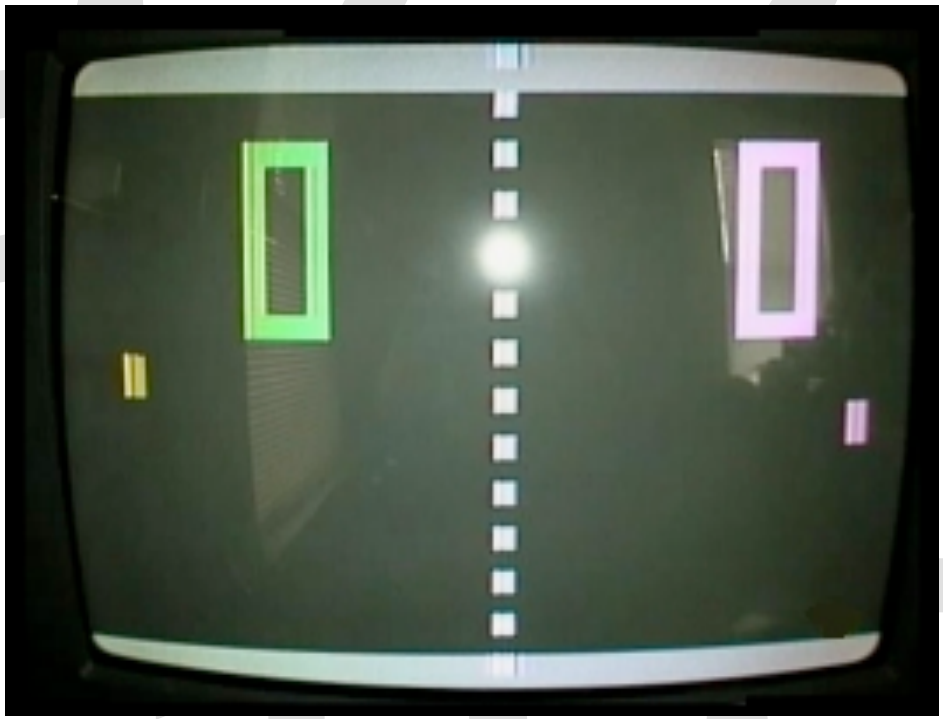




The first working
laser printer is produced

The Universal Product Code
(UPC) is developed

1973



PONG is born!

A little retro

Ah, PONG

Anyway, in the early 1970s we find the beginnings of many of the computing ideas we now take for granted; what about statistics and statistical computing in particular?

Let’s have a look...

THE ANNALS
of
STATISTICS

An Official Journal of
The Institute of Mathematical Statistics

Articles

Short Communications

Continued on back cover

Vol. 1, No. 1—January 1973

"Up to that time, the Annals had published not only papers in mathematical statistics,but also had been one of the main outlets for papers in probability theory. Now the editor, Ingram Olkin, felt that the theory of probability had developed into a subject that deserved its own journal. He persuaded the IMS to create a new journal, the Annals of Probability, and at the same time to broaden the scope of the old Annals by dropping the limiting adjective "mathematical,"so that it would become more welcoming to applied work. The first of these two endeavors was wholly successful, the second less so."

AMERICAN STATISTICAL ASSOCIATION

FOUNDED 1839

806 - 15th Street, N.W. • Washington, D. C. 20005 • (code 202) 393-3253

November 17, 1971

RECEIVED
NOV 22 1971
BIOMATHEMATICS

To: Paul Meier — Arthur P. Dempster
F. J. Anscombe — Wilfrid J. Dixon
Joseph M.Cameron — Michael D. Godfrey
John M. Chambers — Mervin E. Muller
Joseph F. Daly — Martin Schatzoff

From: Churchill Eisenhart, President

By vote of the membership, a new ASA Section on Statistical Computing has been established effective January 1, 1972. More than 1400 members of ASA voted for the establishment of the Section, which thus becomes the successor to the Committee on Computers in Statistics.

Will Dixon has agreed to be 1972 Chairman of the Section and Alan Forsythe will be Secretary. Art Dempster is the 1972 Program Chairman for the Section for the organization of sessions at the Annual Meetings, August 1972 in Montreal. If you have any suggestions for topics for these sessions, please write to Art Dempster at the Statistics Dept., Harvard Univ., 2 Divinity Ave., Cambridge, Mass. 02138.

On behalf of the Association, thank you for your service on the Committee and your leadership in providing the establishment of the new Section.

#3

Proposed Statistical Computing Section, ASA

Statistical Computing is growing rapidly. There is no basis in the Association for encouraging, organizing and integrating these activities into the Association. The section would also serve as a resource group and as liaison to computing societies.

If you are in favor of a section being organized for this purpose please sign below.

Edwin H. [illegible]
[illegible], UCLA
Emil H. Jebe
[illegible]
Douglas A. Zahn
Jeanne Copp
[illegible]
Eugene B. Cohen
Martin C. Weinrich
Etan Markowitz
Joseph J. Chmiel
George Dimberger
Max A. Woodbury
Norman R. [illegible]
Frederick Mosteller
Donald [illegible]
[illegible] McDonnell Douglas [illegible]

William J. Kennedy ISU
William Forsythe UCLA
Martin A. James
Betty Ullman
M G Ott
Nancy A. Symons
K. L. Lee
R. M. Elashoff
J A Hartigan
John M. Chambers
Andrés Petrasovits
Harold Grossman
[illegible]
Robert E. Pearson
Harry O. Levine
Jack D. Maye

[illegible] Haberman
Jack Bishop Jr. Dow Corning
Nader Fergany, UNC
Terry Pope
John Meyer
Martin Schatzoff
Robert F. Ling
Donald Guthrie
Jolayne Service
A K Chattopadhyay, Purdue University
J. D. Daly (Census)
M. Tarter, UC Berkeley.
[illegible] Chen
[illegible] Byers, Com-Share
Anne VanWagoner - com-share
Sister Ignatia Frye - Marygrove College
James W. Longley BLS
[illegible]
P. A. Lachenbruch
C A Blackwood
Charles W. Dunnett
Leo Katz
Albert E. Beaton
Roy H. Wampler, NBS
George Sadowsky - Urban Institute
James R. Thompson - Rice U.

Elizabeth Shuhany
Richard R. Pabst
S S Swaminathan
Ram Gnanadesikan
E. N. West U of Alberta
Richard Madsen
B. Nicholas Lauer (Iowa State)
R. L. Chamberlain (Iowa State)
Daniel Driscoll (Cortland State N.Y.)
Duane Meeter F.S.U.
S. Searle (Cornell)
Elliot Tanis (Hope College)
Gary Anderson (McMaster U.)
Krishna Saw [illegible] (G.D. Searle)
E. L. Battiste, IMSL
Suite 510, 6200 Hillcroft
Houston, Tex 77036
Peter Nemenyi, Virginia State College, Petersburg Va. 23803
D. M. DUNN

Bruce J. McDonald - ONR
[illegible] - CIMS: NYU.
W. Morven Gentleman - Univ. of Waterloo.
Charles E. Gates - Texas A&M
Nancy R. Mann Rocketdyne

Statistical Computing is growing rapidly. There is no basis in the Association for encouraging, organizing and integrating these activities into the Association. The section would also serve as a resource group and as liaison to computing societies.

Statistical computing circa 1972-1973

The ASA Section on Statistical Computing was officially founded on January 1, 1972 and sponsored an invited session at that year's JSM

In 1973, Francis and Heiberger asked the Section to form a committee to evaluate statistical packages, their report ultimately appearing in the American Statistician in 1975

Statistical Computing

This Department will carry articles of high quality on all aspects of computation in statistics.
Papers describing new algorithms, programs, or statistical packages will not contain listings of the program, although the completely documented program must be available from the author. Review of the paper will always include a running test of the program by the referee.
The description of a program or package in this Department should not be construed as an endorsement of it by the American Statistical Association or its Committees, nor is any warranty implied about the validity of the program.
The Editorial Committee will be pleased to confer with authors about the appropriateness of topics or drafts of possible articles.

Criteria and Considerations in the Evaluation of Statistical Program Packages

IVOR FRANCIS* AND RICHARD M. HEIBERGER,** AND PAUL F. VELLEMAN***

1. *Introduction*

Packages of computer programs for statistical analyses have proliferated in recent years and are now widely used, but not always widely understood. Packages can greatly assist statisticians by relieving them of tedious and error-prone computations, by making possible analyses of large data sets, and by providing a flexibility and versatility which can lead to more complete and searching analyses. In addition they are frequently used by non-statisticians to perform statistical analyses hitherto possible only with the collaboration of a statistician.

Few of these packages have ever been formally reviewed by statisticians. Consequently most users choose a package because of its availability or because of word-of-mouth recommendation. As packages have become more available, statisticians have become increasingly concerned about their impact because some contain errors, or are used for purposes other than those intended by their authors.

The first task the committee set itself was to compile a set of criteria and considerations that would be useful in discussing packages. In this task the three authors of the report were greatly aided by nearly one hundred people, including both package designers and users, who had responded to an early draft sent to some five hundred people.

At the Annual meetings of the ASA at St. Louis in August 1974, the Committee presented a report on its work to the Section on Statistical Computing, and to a session with J. M. Chambers, W. J. Dixon, and H. O. Hartley as invited discussants. Copies of the full report [1], containing a list of contributors, are available from the authors.

In this article, we summarize, with some minor modifications, the section of the report on criteria and considerations. The considerations discussed in the report include those suggested by these contributors. Some appear as firm statements about which we expect little controversy. Others are opinions

Statistical computing circa 1972-1973

Francis, Heiberger and Velleman divided their criteria into **user interface** (documentation, “control language”, data structures, printed output, cost, audience and pedagogy), **statistical effectiveness** (versatility and accuracy) and **implementation** (programmer’s documentation, extensibility, portability, source language)

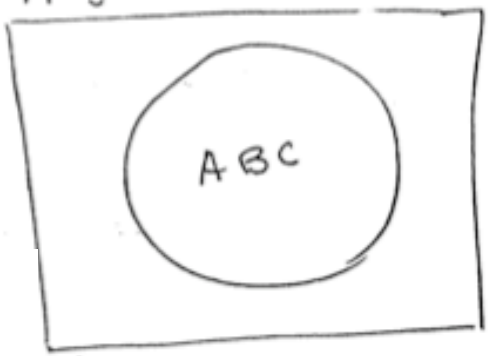
Keep in mind, however, that the early 1970s was the era of the mainframe computer, and the software market was dominated by a handful of “products” (BMDP, SPSS and SAS, for example)

And in 1976.... S

JHC

Algorithm Interface

0/0/76

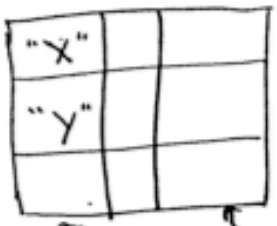


ABC: general (FORTRAN) algorithm

XABC: FORTRAN subroutine to provide interface between ABC & Language and/or utility programs

XABC (INSTR, OUTSTR)

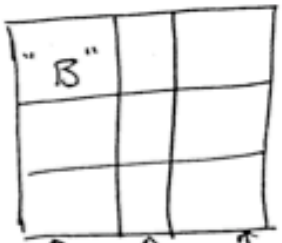
Input INSTR →



Pointers/Values

Argument Names or Blank

OUTSTR →



Pointers / Values

Types (Modes)

Result Names

Note: Names are meaningful to Algorithm, not necessarily to Language

The development of S

The image on the previous page is a scan of the first graphic produced at the May 5, 1976 meeting, and according to John Chambers:

> "The upper portion depicts the concept of an interface between a proposed user-level language and an 'algorithm,' which meant a Fortran-callable subroutine. The lower portion of the figure shows diagrammatically a hierarchical, list-like structure for data, the direct forerunner of lists with named components, structures with attributes and eventually classes with slots... ***The two portions of the sketch in fact lead to the themes of function calls and objects, in retrospect.***

By the end of 1976, Chambers and Rick Becker had produced a working version of their sketch

The program was initially referred to as "the system" locally, and attempts to come up with new names yielded lots of "unacceptable" candidates, but all of which had the letter "S" in common; and so, given the precedent of the recently-developed C language...

Approaching introductory statistics

All of this is a slight distraction, of course; but my point is that my beginning text was **completely silent** when it came to the uses, characteristics and demands of **modern data sources**

It painted statistics in an ahistorical, a-critical fashion; tables and formula were the end results of each chapter -- **statistics is not a living field**, but a set of procedures that can be applied to **a very narrow range of problems,** problems that dodged some of our really "big ideas"

Any nuance, any sense that statistics was **the product of human ingenuity,** was washed over with a kind of mathematical inevitability; there was no hint of **the debates in the field**, no sense of our **history**, and, ultimately, the students were **not prepared to question** the tools they were being offered

Approaching introductory statistics

And my students came to me with significant experience in certain kinds of data analysis, whether SW acknowledged it or not

We are all aware of the proliferation of visualization tools (popularized GIS platforms like Google Earth, for example) and analysis platforms (ManyEyes, Swivel), and even simple data collection frameworks via Twitter

Importantly, students entered the class with a profound understanding that data and data processing are important forces in their lives --
How do help them recognize that all of this is statistics?

Data and computing

The first part of the course emphasized **data and its essential character**; that a single data set can be **reformatted, reshaped, re-aggregated** to focus on different questions

We addressed the **“promiscuity” of data** and the ways in which data sets can be combined to address new questions; in future classes we will speak more directly of how data are formatted and the ease with which information can be extracted

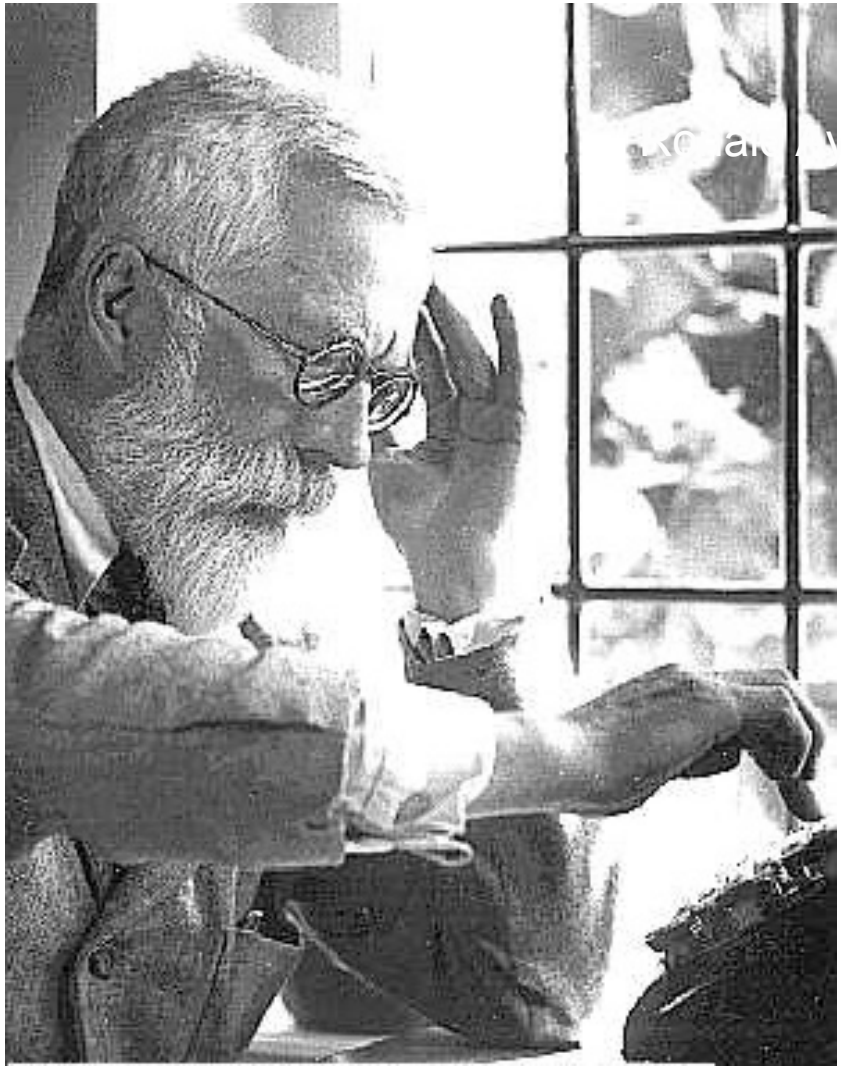
Data and computing

Throughout, we emphasized the idea that data collection (in the health sciences, in particular) is often **a kind of social exchange** and that we need to think carefully about what happens when we turn the world into bits

In other words, we went way beyond just the taxonomy of data types presented by S&W (which has its own interesting history...)

Statisticians and their computing tools...

Ronald Aylmer Fisher (1890-1962)



Fisher once commented that **“he had learned all he knew” over his hand calculator**; as we will see, these computations can reveal structures in data

John Wilder Tukey (1915 - 2000)

For Tukey, “[w]hat the statistician often needs is enough different ‘looks’ to have a good chance to learn about this real world.”

Rick Becker and John Chambers
standing over a terminal running S

BELL LABS NEWS

FOR THE PEOPLE OF BELL LABORATORIES

Vol. 20 No. 38 September 29, 1980

In this issue:

Statistical Software Sleuth Tackles Variety of Cases

Factors contributing to the characteristics of computer programs that affect the amount of necessary debugging and the feasibility of testing by telephone for hearing defects are among the various studies conducted at Bell Labs with the aid of a new software system—simply called *S*.

S was developed at Bell Labs specifically for interactive statistical data analysis, graphics and related scientific computing. Because *S* is interactive, users can see results immediately in order to determine what to do next. They can steer the analysis themselves, without extensive programming experience.

S is being used in a number of Bell System applications. Fran O'Neil, of the Digital Systems Applications Department at Merrimack Valley, used it to develop a model for evaluating the economics of alternative transmission facility and network plans. Tom Hamilton, of AT&T Long Lines' Accounts and Finance Organization at Bedminster, linked *S* with a four-color plotter to enhance the graphical representation of budgetary planning analyses. Last year Blan God-

Lee Wilkinson (Grammar of Graphics)

Well, here are two pix of the machine I built in 1978 and used to write SYSTAT. It was a Cromemco Z2 with 64K of RAM, 2MB of 7" disks, and a Cherry keyboard I mounted in a wood case. I used the Cromemco FORTRAN compiler and WordStar to write code. I used a Diablo printer and a Strobe drum plotter for graphics.

One of my treasured possessions is a plaque (not sure where it is) I won for best talk at National Computer Graphics Association (now defunct). Among all the Tectronix and HP scopes with wire-frame jet planes swooping around, I gave a talk on statistical graphics using my plotter. I noted the few spots on the slide due to raindrops while I photographed the resulting plot in my back yard from a stepladder. The audience loved it.

When Bill Eddy used to regale me with all the VAXes in his possession, I liked to say I had exclusive use of a machine as well, except I had to pay for it out of my own pocketbook. But I could take it to bed at night.

Ross Ihaka and Paul Murrell (co-creator of R and author of “R Graphics” and “Data Technologies”)


R Programming for Bioinformatics
Introduction to Data Technologies

Trevor Hastie (“The Elements of Statistical Learning” and “Generalized Additive Models”)

Dot and I contemplating a thorny problem (but I thinks its wunderground.com we are looking at)

Section to be skipped: Tech motivation (end)

Section to be skipped: History (start)

Statistical Computing in the 1960's

A number of statistical systems and programs already existed; BMD and P-Stat were in current use and GenStat was being developed

These systems grew out of **specific application areas** and tended to offer **pre-packaged analyses**

At the time, most statistics researchers would not be directly involved in analyzing data; programmers (read graduate students) would do the grubby work when necessary

The systems like BMD and SAS, for example, and PStat, to some extent, and GenStat's another good example, **all grew up in environments where statisticians were required to do some fairly committed routine analysis**. So BMD, of course comes from a biomedical field; SAS from several areas but medical, again; and GenStat comes from an agricultural background.

Now in all those situations, the statistics groups, amongst other duties, were expected to be doing kind of analysis to order. You know, the data would come along from a experiment, or a clinical trial, or other sources, and as part of the job of the statisticians to produce analysis. **Now often the analysis that they produced were relatively predetermined, or at least that's how it worked out.**

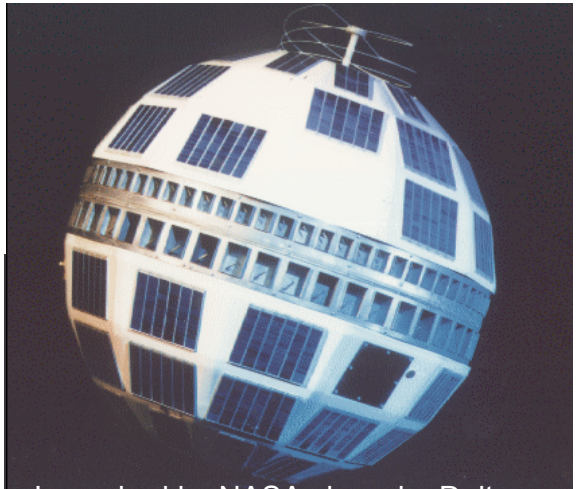
Interview with John Chambers, 2002

The mid 1960's at Bell Labs

Statistics Research at Bell Labs tackled large-scale data analysis projects with teams of researchers and "programmers"

Unlike much of the statistical computing of the day, this kind of work was not well suited to pre-packaged programs

Even then, AT&T was creating large-scale applications; data from TelStar, an early (1960s) communications satellite, involved tens of thousands of observations



Launched by NASA aboard a Delta rocket from Cape Canaveral on July 10, 1962, Telstar was the first privately sponsored space launch. A medium-altitude satellite, Telstar was placed in an elliptical orbit (completed once every 2 hours and 37 minutes), revolving at a 45 degree angle above the equator. Because of this, its transmission availability for transatlantic signals was only 20 minutes in each orbit.

Telstar relayed its first television pictures (of a flag outside its ground station in Andover, Maine) on the date of its launch. Almost two weeks later, on July 23, it relayed the first live transatlantic television signal. During that evening it also dealt with the first telephone call transmitted through space and successfully transmitted faxes, data, and both live and taped television, including the first live transmission of television across an ocean (to Pleumeur-Bodou, in France). John F. Kennedy, then President of the United States, gave a live transatlantic press conference via Telstar.

The mid 1960's at Bell Labs

During this period, John Tukey was also starting to formulate the beginnings of Exploratory Data Analysis

"Exploratory analysis is detective work -- numerical detective work -- or counting detective work -- or graphical detective work"

EDA is an approach to data analysis that postpones the usual assumptions about what kind of model the data follow with the more direct approach of allowing the data itself to reveal its underlying structure and model. EDA is not a mere collection of techniques; **EDA is a philosophy as to how we dissect a data set**; what we look for; how we look; and how we interpret.

Most EDA techniques are **graphical** in nature with a few quantitative techniques. The reason for the heavy reliance on graphics is that by its very nature **the main role of EDA is to open-mindedly explore**, and graphics gives the analysts unparalleled power to do so, enticing the data to reveal its structural secrets, and being always ready to gain some new, often unsuspected, insight into the data. In combination with the natural pattern-recognition capabilities that we all possess, graphics provides, of course, unparalleled power to carry this out.

Taken from `www.itl.nist.gov`

The mid 1960's at Bell Labs

> "Today, software and hardware together provide far more powerful factories than most statisticians realize, factories that many of today's most able young people find exciting and worth learning about on their own. Their interest can help us greatly, if statistics starts to make much more nearly adequate use of the computer. **However, if we fail to expand our uses, their interest in computers can cost us many of our best recruits, and set us back many years**."

The technical tools of Statistics, Nov 1964

The mid 1960's at Bell Labs

In previous decades, computers had matured significantly, but the access a user might have to these systems was limited; recall that in 1964, Bell Labs partnered with MIT and GE to create Multics (for Multiplexed Information and Computing Service)

> "Such systems must run continuously and reliably 7 days a week, 24 hours a day in a way similar to telephone or power systems, and must be capable of meeting wide service demands: from multiple man-machine interaction to the sequential processing of absentee-user jobs; from the use of the system with dedicated languages and subsystems to the programming of the system itself"

The mid 1960's at Bell Labs

While Bell Labs dropped out of the project in 1969, it did spark a lot of interest among researchers throughout the lab; John Chambers had just joined the lab and was starting to think about larger computing issues

He and a small group of statisticians began to consider how this kind of computing platform might benefit the practice of statistics

What should the computer do for us?

What should the computer do for us?

Let's take a moment and answer that question for ourselves; you now have a fair bit of experience collecting, organizing and analyzing data with simple shell tools and Python scripts

What operations would a computing environment designed for statisticians make easy? What kinds of data types should be "built-in"? How should one interact with it?

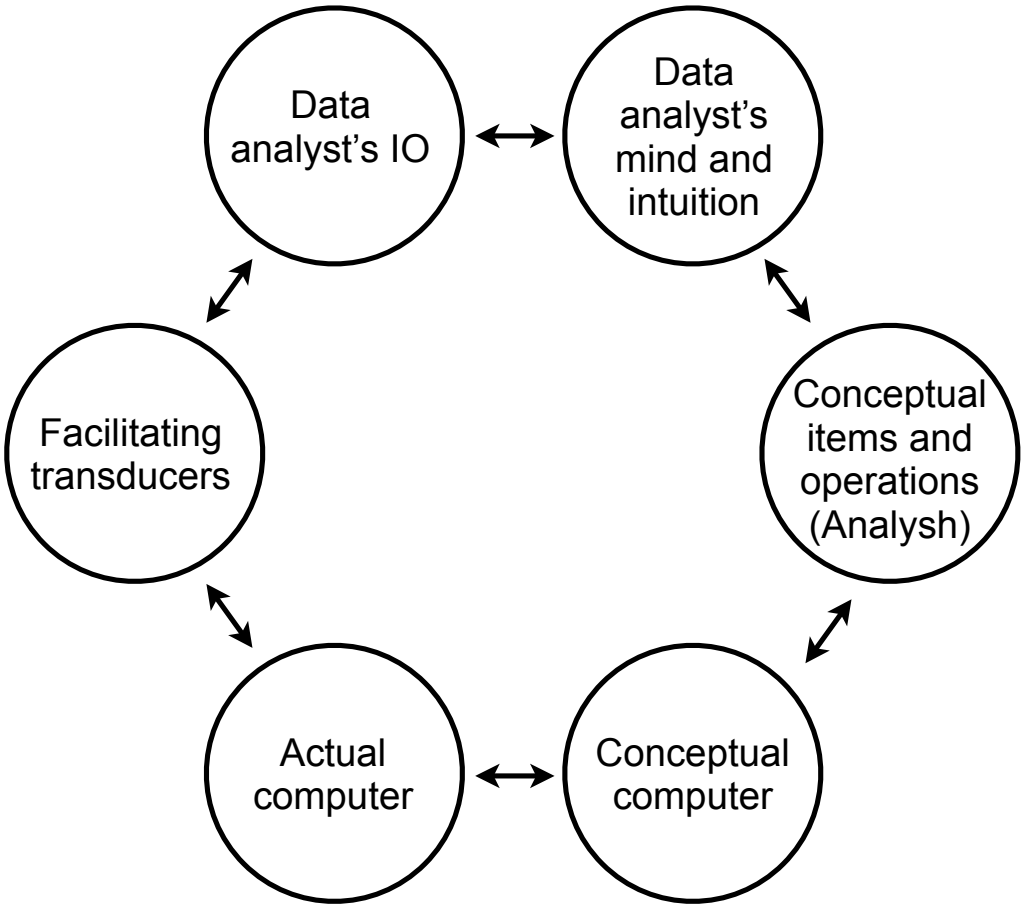
How do your answers change if you think instead about a someone looking at data, someone who might not be a statistician?

The mid 1960's at Bell Labs

What should the computer do for us?

In answering this question, the Bell Labs group considered the necessary components of a system that would "more naturally express what we do and allow us to interactively *program* analyses"

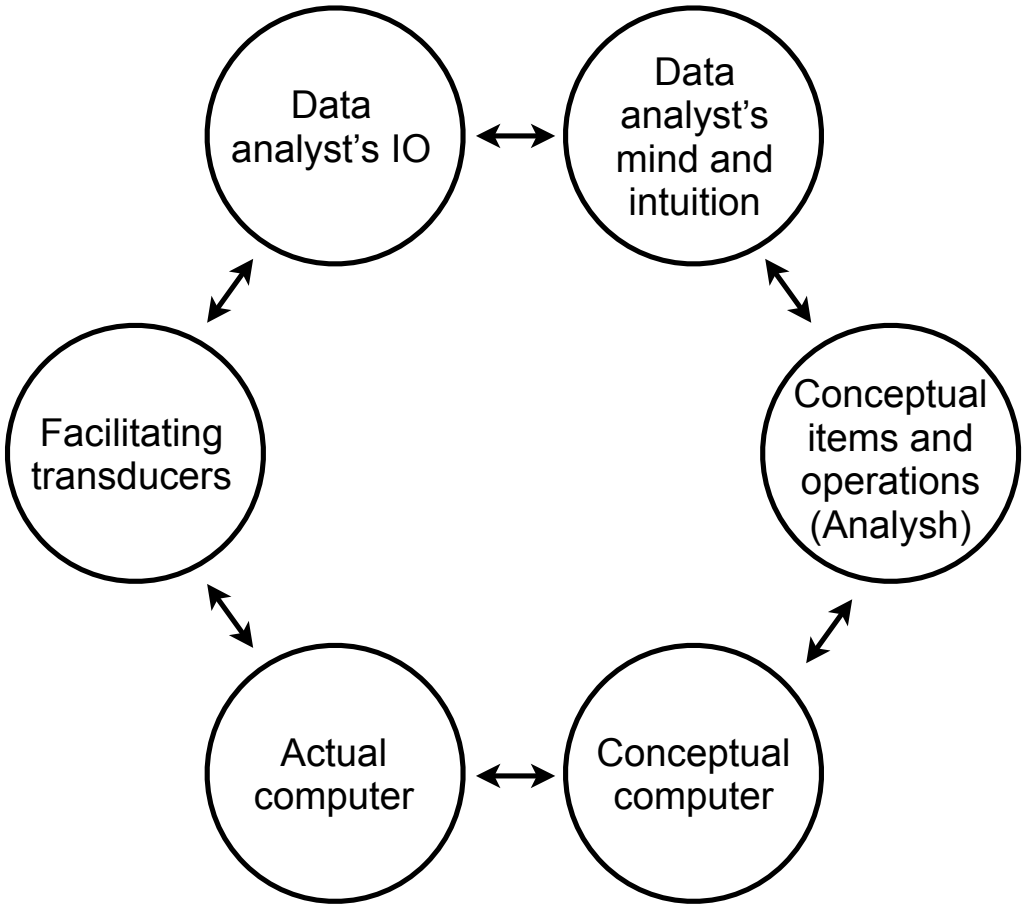
Tukey produced a memo that outlined the basic concepts; memos were the emails of the mid 1960s



Adapted from Chambers (2000)

Follow the arrows clockwise from the Mind and Intuition block. Tukey's notion is that data analysts have an arsenal of operations applicable to data, which they describe to themselves and to each other in a combination of mathematics and (English) words, for which he coins the term Analysh. These descriptions can be made into algorithms (my term, not his) -- specific computational methods, but not yet realized for an actual computer (hence the conceptual computer). Then a further mapping implements the algorithm, and running it produces output for the data analyst. The output, of course, stimulates further ideas and the cycle continues. (The facilitating transducers I interpret to mean software that allows information to be translated back and forth between internal machine form and forms that humans can write or look at -- a transducer, in general, converts energy from one form to another. So parsers and formatting software would be examples.)

Taken from Chambers (2000)



Adapted from Chambers (2000)

The mid 1960's at Bell Labs

This group met again in the summer of 1965; Martin Wilk opened the meeting with the following

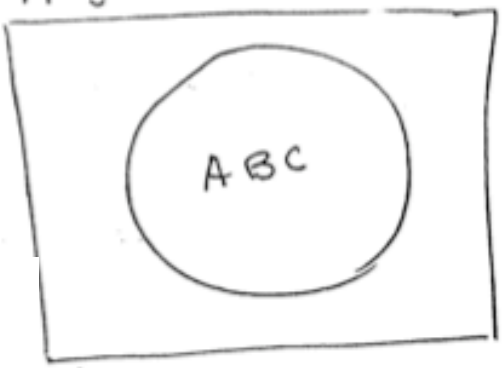
> "What do we want? We want to have easy, flexible, availability of basic or higher level operations, with convenient data manipulation, bookkeeping and IO capacity. We want to be able easily to modify data, output formats, small and large programs and to do this and more with a standard language adapted to statistical usage."

Unfortunately, the plans developed by this group remained just that, plans; it wasn't until a decade later that a second group met, and on **May 5, 1976, work started on what would become "S"**

JTHC

Algorithm Interface

5/5/76



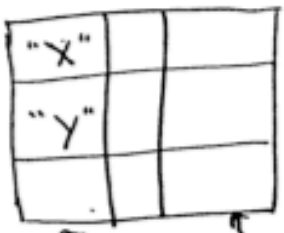
XABC

ABC: general (FORTRAN) algorithm

XABC: FORTRAN subroutine to provide interface between ABC & Language and/or utility programs

XABC (INSTR, OUTSTR)

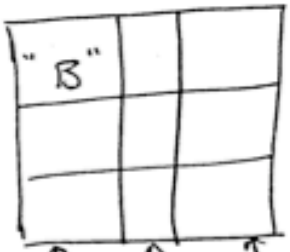
Input INSTR →



Pointers/Values

Argument Names or Blank

OUTSTR →



Pointers / Values

Types (Modes)

Result Names

Note: Names are meaningful to Algorithm, not necessarily to Language

The development of S

The image on the previous page is a scan of the first graphic produced at the May 5 meeting, and according to Chambers:

> "The upper portion depicts the concept of an interface between a proposed user-level language and an 'algorithm,' which meant a Fortran-callable subroutine. The lower portion of the figure shows diagrammatically a hierarchical, list-like structure for data, the direct forerunner of lists with named components, structures with attributes and eventually classes with slots... ***The two portions of the sketch in fact lead to the themes of function calls and objects, in retrospect.***

The development of S

By the end of 1976, John Chambers and Rick Becker had produced a working version of their sketch

The program was initially referred to as “the system” locally, and attempts to come up with new names yielded lots of “unacceptable” candidates, but all of which had the letter “S” in common; and so, given the precedent of the recently-developed C language...

Unix and S

Becker and Chambers made the most of the Unix project happening in parallel at the labs at the same time; by creating **a Unix version of S** (S Version 2), **their system became portable** (or at least it could go anywhere Unix could)

At that point, **AT&T started licensing both Unix and S**, with both university groups and “third-party” resellers in mind; this meant others could contribute to the development of the software

The development of S

About a decade after the first version of S, a complete revision of the language took place (S3, described in the “blue book”) which gave rise to an extensive modeling effort (described in the “white book”); for most of you, this version of S will most closely resemble R (to come)

One more decade and another version (S4, described in Chambers’ book “Computing with Data”) that brought with it a new class and method system

In 1998 Chambers won the ACM (Association for Computing Machinery) Software System Award; **S has “forever altered the way people analyze, visualize and manipulate data”**

And finally...

Ross Ihaka and Robert Gentleman (both at the University of Auckland at the time) wrote a reduced version of S for “teaching purposes”

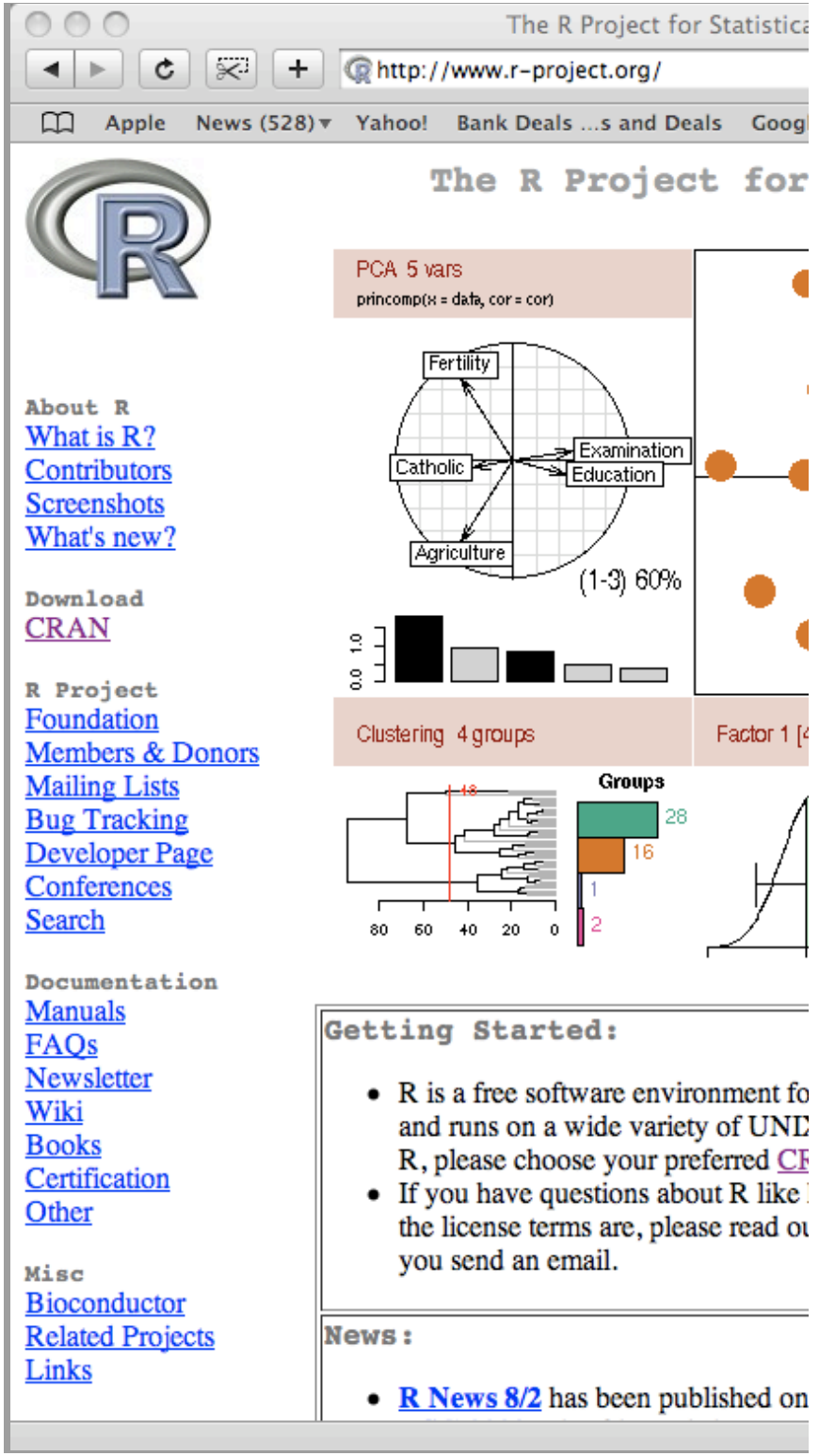
In 1995 Ross and Robert were persuaded to release the code for R under the GPL (the General Public License; it allows you to use it freely, distributed it, and even sell it, as long as the receiver has the same rights and the source code is freely available)

R is now administered by a core group of about a dozen people; but many more contribute code, and many many more program in R

The R environment

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes

- *an effective data handling and storage facility,*
- *a suite of operators for calculations on arrays, in particular matrices,*
- *a large, coherent, integrated collection of intermediate tools for data analysis,*
- *graphical facilities for data analysis and display either on-screen or on hardcopy, and*
- *a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.*

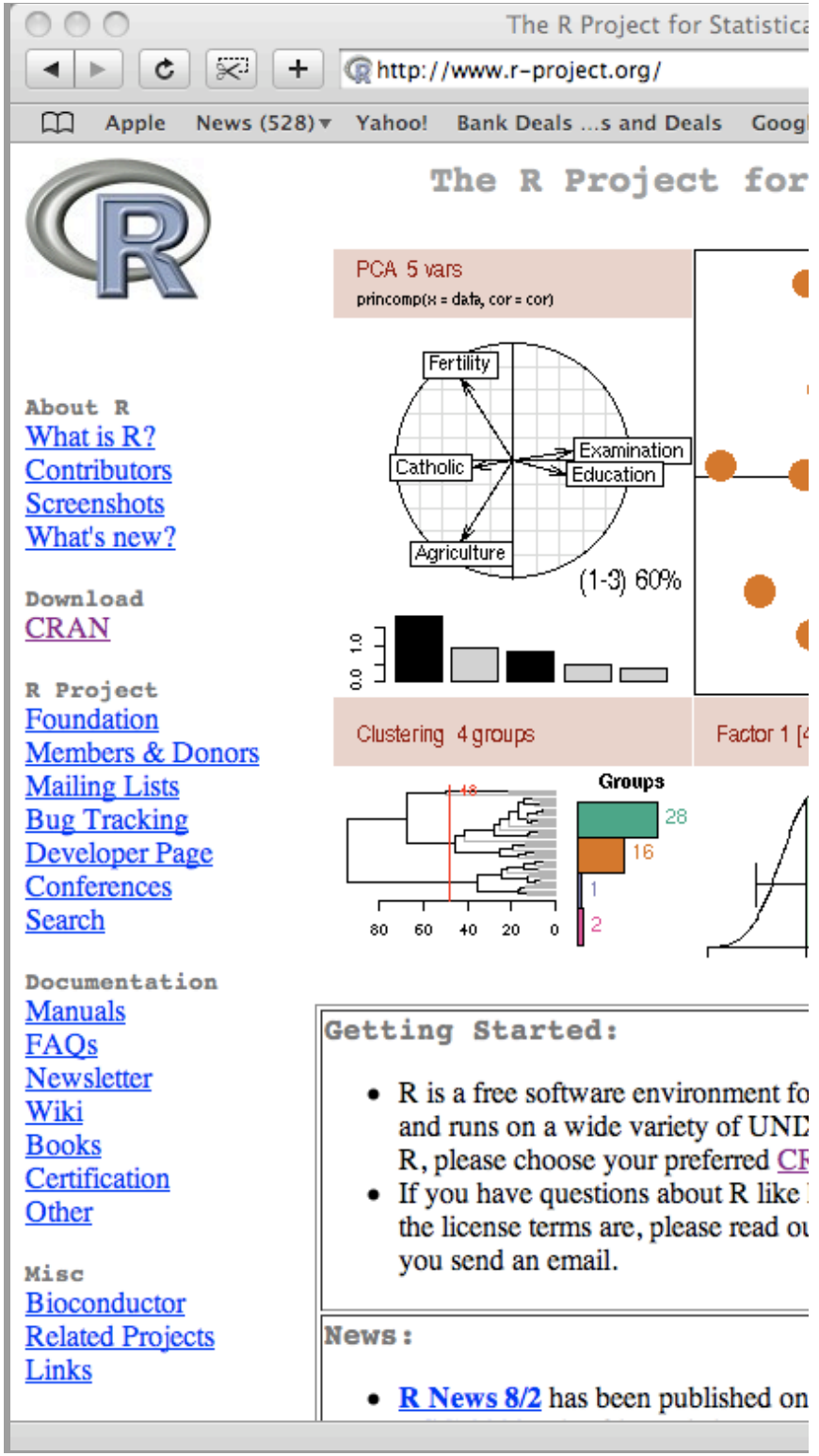


The R environment

The term "environment" is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software.

R, like S, is designed around a true computer language, and it allows users to add additional functionality by defining new functions. Much of the system is itself written in the R dialect of S, which makes it easy for users to follow the algorithmic choices made.

Many users think of R as a statistics system. We prefer to think of it of an environment within which statistical techniques are implemented. R can be extended (easily) via packages.



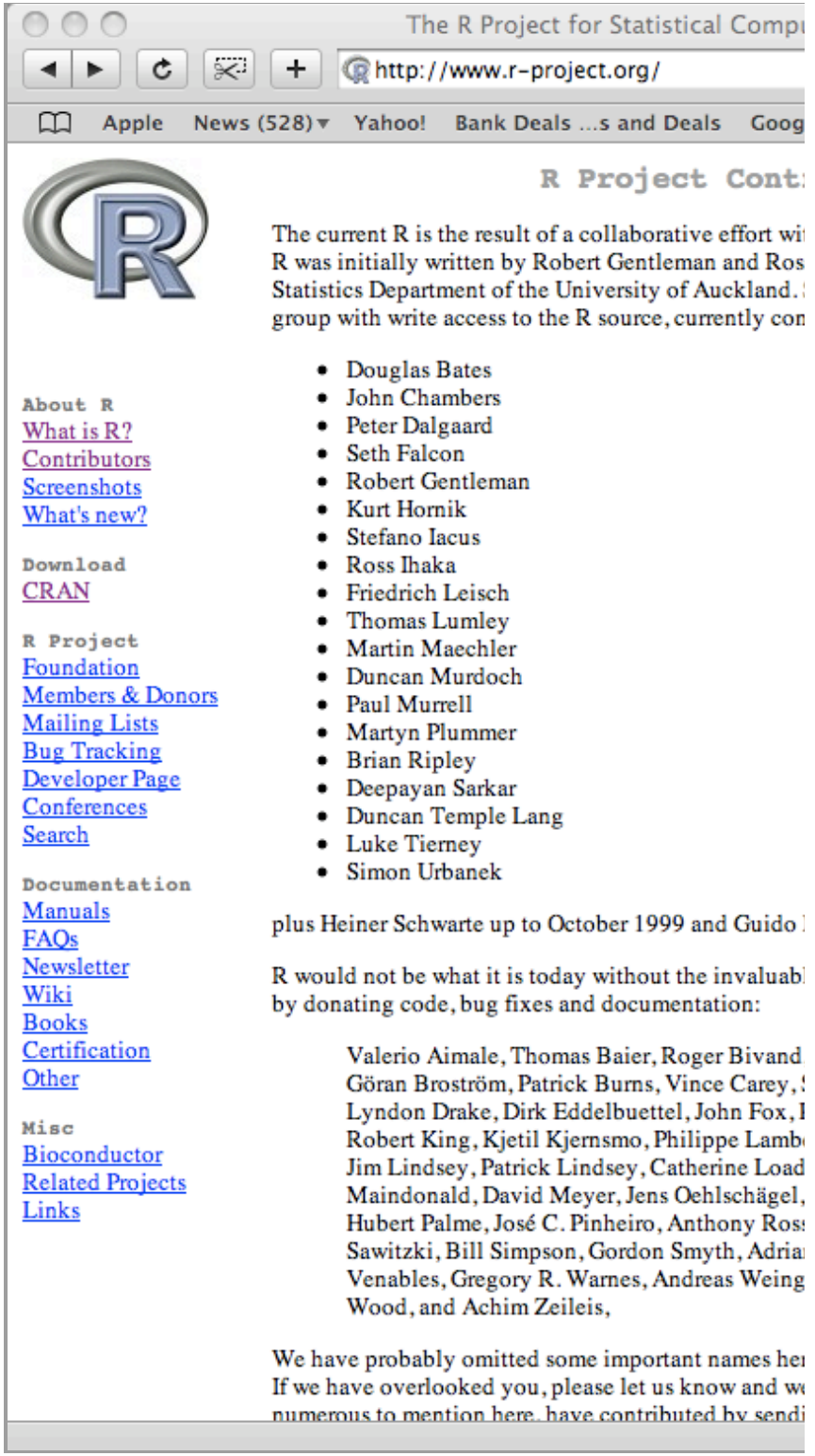
The R environment

While there is a fair bit of difference between how R and S are implemented, one of the most visible for users involves how code is shared

> *The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.*

In my mind, the ease with which packages can be published, accessed (installed) and auditioned is one of the real innovations of R

As an aside, what do (nearly) all of the names on the right have in common?



Section to be skipped: History (end)

Surveys

The simple tabular structure of a survey means that we can introduce R's basic **data frame along with the syntax for subsetting**

Simple **graphical displays and numerical summaries** are introduced as students try to make sense of the survey respondents -- We hope that the subject matter has enough intrinsic interest that this process can be conducted without much guidance

We close the lesson with some comments on **how data are published** -- That the choice of formats made by an organization can say a lot about intended use cases

ISSUE NO. 1:
ARE PARENTS AND STUDENTS READY FOR MORE MATH AND SCIENCE?

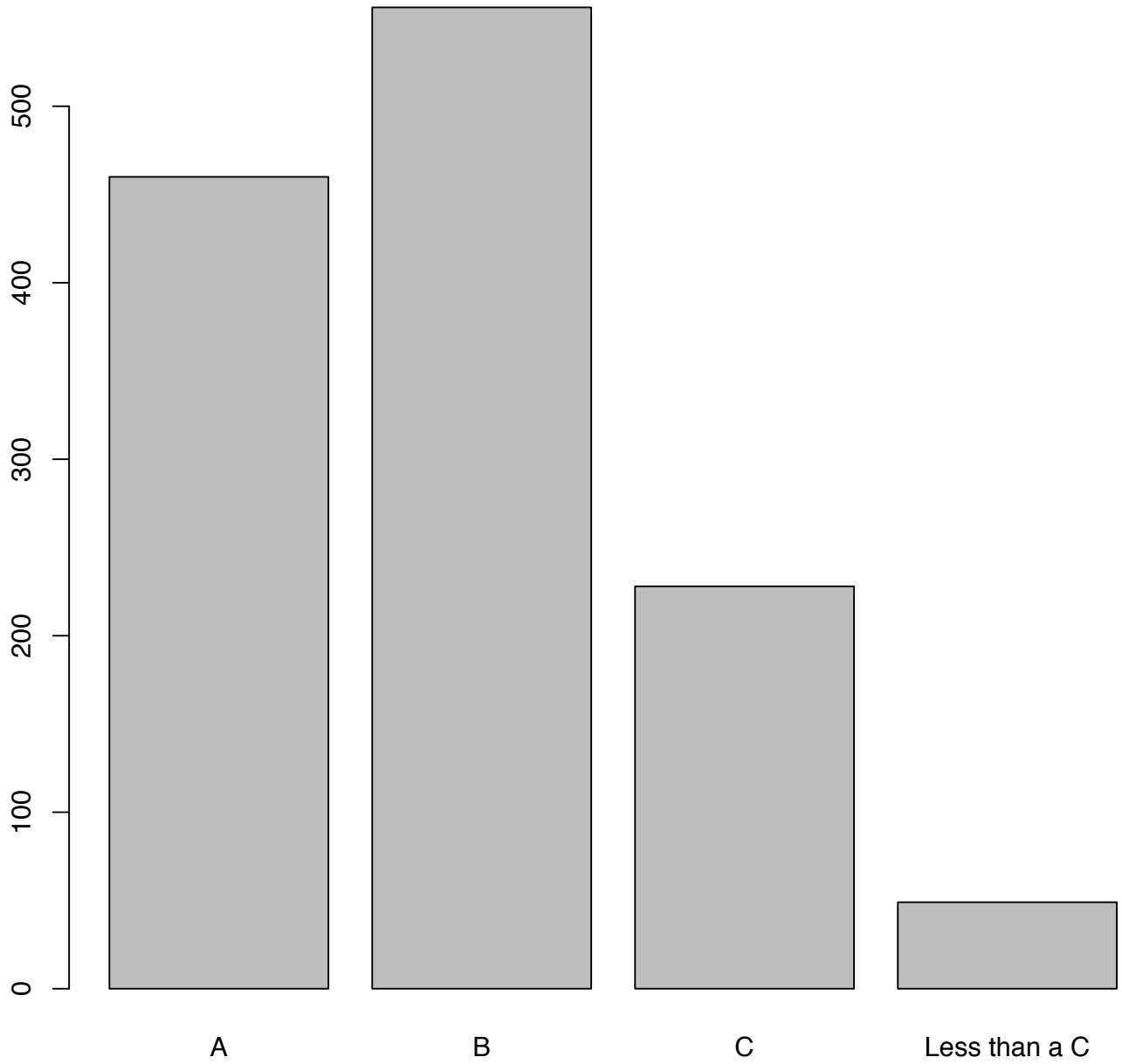
A Report from **Education Insights** at Public Agenda.

Funding for this report was provided by:
GE Foundation
Nellie Mae Education Foundation
The Wallace Foundation

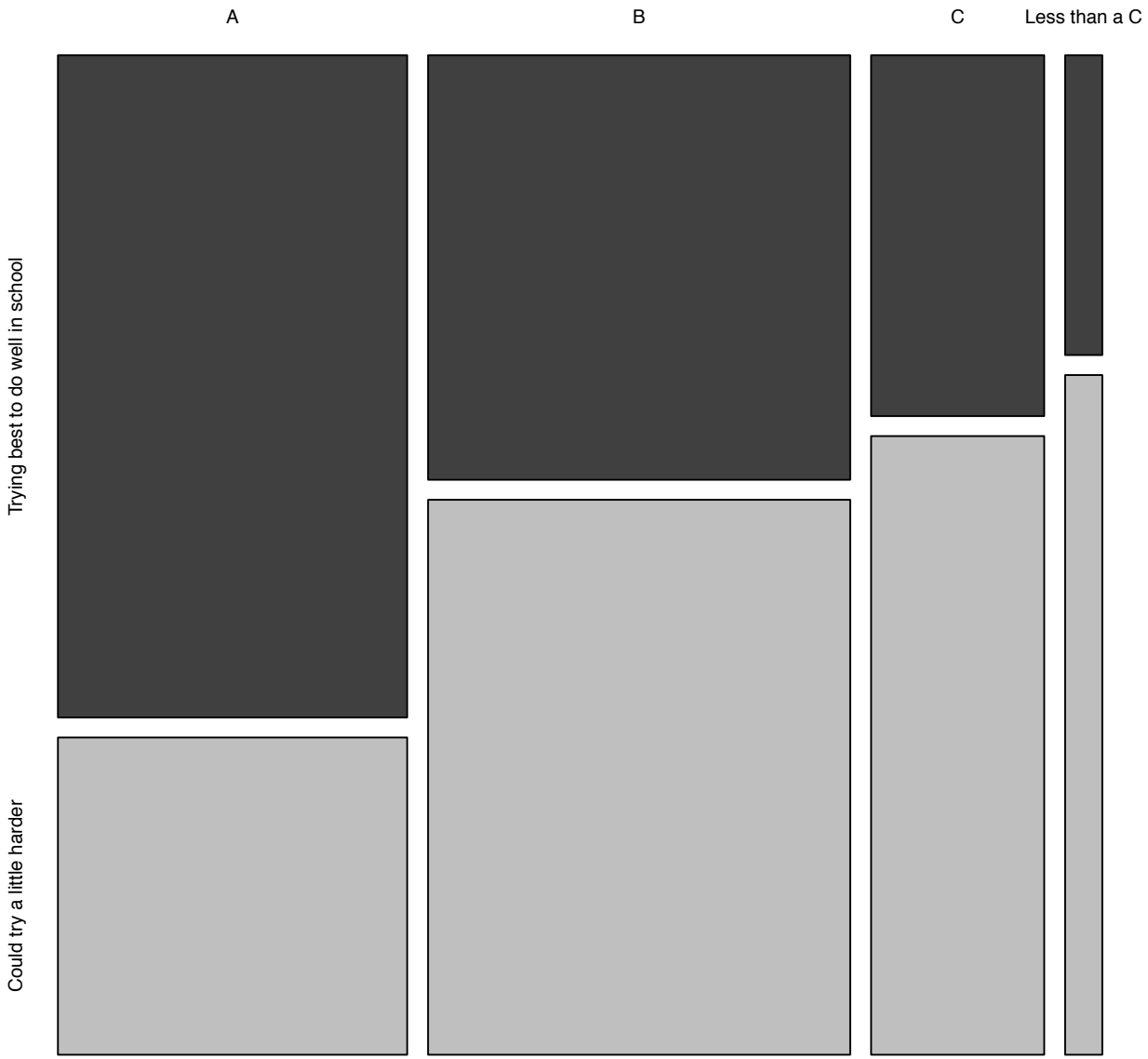
EDUCATION INSIGHTS

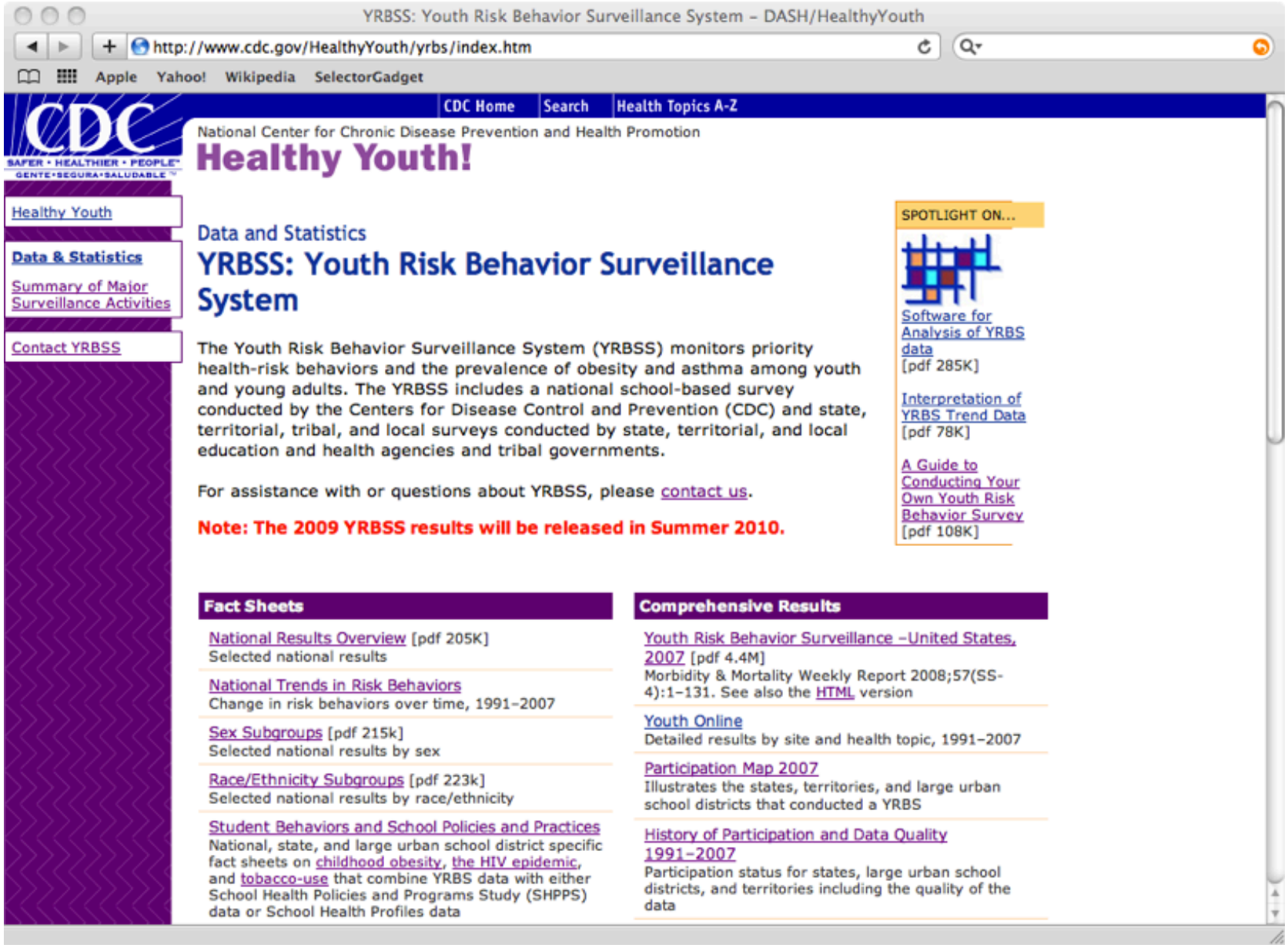
A Public Agenda Initiative to Build Momentum for Improving American Schools

```
# loading data and asking some simple questions

> load(url("http://mobilize.stat.ucla.edu/day1/data/reality_check.Rda"))

> class(survey)
[1] "data.frame"

> dim(survey)
[1] 1293    4

> head(survey)
            year                          effort                homework        grades
1    Ninth grade        Could try a little harder About the right amount             B
2 Eleventh grade Trying best to do well in school      Too much homework             B
3    Sixth grade Trying best to do well in school About the right amount             B
4    Ninth grade        Could try a little harder About the right amount             B
5    Ninth grade        Could try a little harder    Too little homework Less than a C
6   Eighth grade Trying best to do well in school      Too much homework             A

# having a look...

> table(survey$grades)

           A            B            C Less than a C
         460          556          228           49

> barplot(table(survey$grades))
> mosaicplot(table(survey$grade,survey$effort))

# what does “no relationship” look like?
> mosaicplot(table(survey$effort,sample(survey$grade)))
```


0
100
200
300
400
500
A
B
C
Less than a C

Grades v. Effort




YRBSS: Youth Risk Behavior Surveillance System – DASH/HealthyYouth
http://www.cdc.gov/HealthyYouth/yrbs/index.htm
Apple
Yahoo!
Wikipedia
SelectorGadget
CDC Home
Search
Health Topics A-Z
CDC
SAFER • HEALTHIER • PEOPLE
GENTE • SEGURA • SALUDABLE
National Center for Chronic Disease Prevention and Health Promotion
Healthy Youth!
Healthy Youth
Data & Statistics
Summary of Major Surveillance Activities
Contact YRBSS
Data and Statistics
YRBSS: Youth Risk Behavior Surveillance System
The Youth Risk Behavior Surveillance System (YRBSS) monitors priority health-risk behaviors and the prevalence of obesity and asthma among youth and young adults. The YRBSS includes a national school-based survey conducted by the Centers for Disease Control and Prevention (CDC) and state, territorial, tribal, and local surveys conducted by state, territorial, and local education and health agencies and tribal governments.
For assistance with or questions about YRBSS, please contact us.
Note: The 2009 YRBSS results will be released in Summer 2010.
SPOTLIGHT ON...
Software for Analysis of YRBS data [pdf 285K]
Interpretation of YRBS Trend Data [pdf 78K]
A Guide to Conducting Your Own Youth Risk Behavior Survey [pdf 108K]
Fact Sheets
National Results Overview [pdf 205K]
Selected national results
National Trends in Risk Behaviors
Change in risk behaviors over time, 1991–2007
Sex Subgroups [pdf 215k]
Selected national results by sex
Race/Ethnicity Subgroups [pdf 223k]
Selected national results by race/ethnicity
Student Behaviors and School Policies and Practices
National, state, and large urban school district specific fact sheets on childhood obesity, the HIV epidemic, and tobacco-use that combine YRBS data with either School Health Policies and Programs Study (SHPPS) data or School Health Profiles data
Comprehensive Results
Youth Risk Behavior Surveillance –United States, 2007 [pdf 4.4M]
Morbidity & Mortality Weekly Report 2008;57(SS-4):1–131. See also the HTML version
Youth Online
Detailed results by site and health topic, 1991–2007
Participation Map 2007
Illustrates the states, territories, and large urban school districts that conducted a YRBS
History of Participation and Data Quality 1991–2007
Participation status for states, large urban school districts, and territories including the quality of the data

4. Variable Documentation

This section describes how the race/ethnicity, overweight, obese, and dichotomous variables were generated from the original survey questions.

4.1 Race/Ethnicity

The 2007 YRBS uses the following two questions to determine race/ethnicity:

Q4 Are you Hispanic or Latino?
- A. Yes
- B. No

Q5 What is your race? (Select one or more responses.)
- A. American Indian or Alaska Native
- B. Asian
- C. Black or African American
- D. Native Hawaiian or Other Pacific Islander
- E. White

Ethnicity (Q4) is scanned as a single column variable with either A or B as valid responses. Race (Q5) is the only YRBS question that permits the selection of more than one response. It is a "check all that apply" type question and is scanned as an eight-column character variable. If the student selected "A", then the first column contains an "A". If they selected "B", then the second column contains a "B", and so on.

To maintain comparability with results prior to 2005, which used a single question to ascertain race/ethnicity, *Q4* and *Q5* are combined to create the two-column *raceeth* variable. If the student selected "B" for *Q4* and only one response option for Q5 then *raceeth* is set to a number between "1" and "5" depending on the Q5 option selected. If they selected "A" for Q4 and no response for Q5, *raceeth* is set to "6" indicating "Hispanic/Latino". If they selected "A" for Q4 and one or more responses for Q5, then *raceeth* is set to "7" indicating "Multiple – Hispanic/Latino". If they selected "B" for Q4 and more than one response for Q5, then *raceeth* is set to "8" indicating "Multiple – Non-Hispanic/Latino". *Raceeth* is set to "missing" if they answered "B" to Q4 and left Q5 blank, or they left Q4 blank regardless of the response for Q5, or if Q4 or Q5 is out of range.

2007 YOUTH RISK BEHAVIOR SURVEY

United States High School Survey
Data Users Manual

Q4: Ethnicity	*Q5: Race*	*Raceeth* (Values and Labels)
B	A	1 (American Indian/Alaskan Native)
B	B	2 (Asian)
B	C	3 (Black or African American)
B	D	4 (Native Hawaiian or Other Pacific Islander)
B	E	5 (White)
A	Missing	6 (Hispanic/Latino)
A	1 or more responses	7 (Multiple– Hispanic/Latino)
B	2 or more responses	8 (Multiple– Non-Hispanic/Latino)
B	Missing	Missing
Missing	Missing or any response	Missing
Out of range	Out of range	Missing

4.2 BMI Percentile, Overweight, and Obese

Overweight (QNROVWGT) and *Obese* (QNOVWGT) status for the YRBS is determined using a SAS program provided by CDC's Division of Nutrition and Physical Activity (DNPA). The SAS program generates BMI and BMI percentile for age and sex based on the 2000 CDC Growth Charts. The student's BMI percentile determines *Overweight* and *Obese* status. More information about the SAS program can be found on DNPA's website - www.cdc.gov/nccdphp/dnpa/growthcharts/resources/sas.htm.

Age (Q1), Sex (Q2), Height (Q6), and Weight (Q7) are input into the SAS program. The units on the YRBS for these variables differ from the units required by the SAS program. Prior to using the SAS program, the YRBS units are converted as follows:

Variable	YRBS	SAS
Age	A. 12 years old or younger	A → 150 months*
	B. 13 years old	B → 162 months
	C. 14 years old	C → 174 months
	D. 15 years old	D → 186 months
	E. 16 years old	E → 198 months
	F. 17 years old	F → 210 months
	G. 18 years old or older	G → 222 months
Sex	1 = Female	1 = Male
	2 = Male	2 = Female
Height	Meters	Centimeters
Weight	Kilograms	Kilograms

*Use the mid-point of the year to determine age in months. For example, if the student answered 12 years old or younger, use (12.5 years *12 = 150 months.)

BMI and BMI percentile are output by the SAS program and available on the public use YRBS data file. When BMI percentile is at or above the 85th percentile and below the 95th percentile for BMI by age and sex, the student is considered overweight, and QNROVWGT is set to "1". The student is considered obese, and QNOVWGT is set to "1", when BMI percentile is at or above the 95th percentile for BMI by age and sex. QNROVWGT and QNOVWGT are mutually exclusive.

4.3 Dichotomous Variables

There are two types of dichotomous variables - **QN#** and **QNword**. The dichotomous variables present the percentage of students answering the predetermined response(s) of interest (ROI). Students who answered the ROI(s) are in the numerator. The denominator is either all students or a subset of students who have indicated in the current survey they participate in a selected activity or behavior. Students must have provided valid data to be included in any dichotomous variable calculations. Therefore students with missing responses or who had their answers subverted are not included. The variables are created and added to the master datasets during editing.

4.3.1 QN# Variables: Each question has a corresponding dichotomous variable. The name of the dichotomous variable corresponds to the standard question number. For example, the dichotomous variable for Q10 is named QN10. The table below provides the question and response options used for each standard Q# variable and related QN# variable. The bolded responses indicate the ROIs for that question. The ROIs are set to "1" for the QN# variables; the remaining responses are set to "2" or to "missing" for the QN# variable. The numerator and denominator are further defined below the responses. The summary text for each QN# variable is also listed.

Q8:	When you rode a bicycle during the past 12 months, how often did you wear a helmet?	
	A.	I did not ride a bicycle during the past 12 months
	B.	**Never wore a helmet**
	C.	**Rarely wore a helmet**
	D.	Sometimes wore a helmet
	E.	Most of the time wore a helmet
	F.	Always wore a helmet
QN8:	Numerator:	Students who answered B or C for Q8
	Denominator:	Students who answered B, C, D, E, or F for Q8
	Summary text:	Among students who rode a bicycle during the past 12 months, the percentage who never or rarely wore a bicycle helmet

Q9:	How often do you wear a seat belt when riding in a car driven by someone else?	
	A.	**Never**
	B.	**Rarely**
	C.	Sometimes
	D.	Most of the time
	E.	Always
QN9:	Numerator:	Students who answered A or B for Q9
	Denominator:	Students who answered A, B, C, D, or E for Q9
	Summary text:	Percentage of students who never or rarely wore a seat belt when riding in a car driven by someone else
Q10:	During the past 30 days, how many times did you ride in a car or other vehicle driven by someone who had been drinking alcohol?	
	A.	0 times
	B.	**1 time**
	C.	**2 or 3 times**
	D.	**4 or 5 times**
	E.	**6 or more times**
QN10:	Numerator:	Students who answered B, C, D, or E for Q10
	Denominator:	Students who answered A, B, C, D, or E for Q10
	Summary text:	Percentage of students who rode one or more times during the past 30 days in a car or other vehicle driven by someone who had been drinking alcohol
Q11:	During the past 30 days, how many times did you drive a car or other vehicle when you had been drinking alcohol?	
	A.	0 times
	B.	**1 time**
	C.	**2 or 3 times**
	D.	**4 or 5 times**
	E.	**6 or more times**
QN11:	Numerator:	Students who answered B, C, D, or E for Q11
	Denominator:	Students who answered A, B, C, D, or E for Q11
	Summary text:	Percentage of students who drove a car or other vehicle one or more times during the past 30 days when they had been drinking alcohol
Q12:	During the past 30 days, on how many days did you carry a weapon such as a gun, knife, or club?	
	A.	0 days
	B.	**1 day**
	C.	**2 or 3 days**
	D.	**4 or 5 days**
	E.	**6 or more days**
QN12:	Numerator:	Students who answered B, C, D, or E for Q12
	Denominator:	Students who answered A, B, C, D, or E for Q12
	Summary text:	Percentage of students who carried a weapon such as a gun, knife, or club on one or more of the past 30 days

Time

This lessons starts with **a small review about data formats** (binary v. ASCII, fixed width v. delimited, humanly readable, self describing and so on) -- We also have a small discussion on purposeful data collection

We then transition to **the web access logs from their school's web site** -- Here they see a different kind of data collection as well as a data file that needs to be **parsed to extract information (regular expressions)**

The bulk of this lesson, however, is about **data structures to represent the time of an event** (in this case, the hit to a web server)

We include a discussion of logarithms (sorry) to help with viewing certain kinds of access data

We close on a comment about how R can be used for **automated browsing** and the idea of **realtime data analysis**

```
> library(hexView)
> viewRaw("yrbs07.sav",nbytes=100)

 0  :  1f 8b 08 00 00 00 00 00 00 03 ec bd 3d b0 2c c9 95  |  ............=.,..
17  :  df 57 f7 dd be f7 be c1 9b 21 a0 37 d8 19 72 29 82  |  .W.......!.7..r).
34  :  4f 9e 4c 61 29 41 92 b1 81 19 90 da 5d 47 8a 58 46  |  O.La)A......]G.XF
51  :  28 82 8a a0 83 c0 02 a4 01 71 36 66 21 45 50 32 d4  |  (........q6f!EP2.
68  :  0e 6d d9 92 29 d2 a6 bd b6 6c c9 66 c0 5c 5b 32 64  |  .m..)....l.f.\[2d
85  :  e0 6b 40 ec 40 af fb 55 d6 3b f5 3b e7 7f 32        |  .k@.@..U.;.;.2
```

Deeplinks Archives

April, 2010
March, 2010
February, 2010
January, 2010
December, 2009
November, 2009
October, 2009
September, 2009
More Archives

Blog Categories

Accessibility for the Reading Disabled
Analog Hole
Announcement
Anonymity
Anti-Counterfeiting Trade Agreement
Bloggers' Rights
Broadcast Flag
Broadcasting Treaty
CALEA
Call To Action
Cell Tracking
Coders' Rights Project

JANUARY 27TH, 2010

Help EFF Research Web Browser Tracking

Announcement by Peter Eckersley

What fingerprints does your browser leave behind as you surf the web?

Traditionally, people assume they can prevent a website from identifying them by disabling cookies on their web browser. Unfortunately, this is not the whole story.

When you visit a website, you are allowing that site to access a lot of information about your computer's configuration. Combined, this information can create a kind of fingerprint — a signature that could be used to identify you and your computer. But how effective would this kind of online tracking be?

EFF is running an experiment to find out. Our new website Panopticlick will anonymously log the configuration and version information from your operating system, your browser, and your plug-ins, and compare it to our database of five million other configurations. Then, it will give you a uniqueness score — letting you see how easily identifiable you might be as you surf the web.

Adding your information to our database will help EFF evaluate the capabilities of Internet tracking and advertising companies, who are already using techniques of this sort to record people's online activities. They develop these methods in secret, and don't always tell the world what they've found. But this experiment will give us more insight into the privacy risk posed by browser fingerprinting, and help web users to protect themselves.

To join the experiment:
http://panopticlick.eff.org/

Mechanics of web browsing

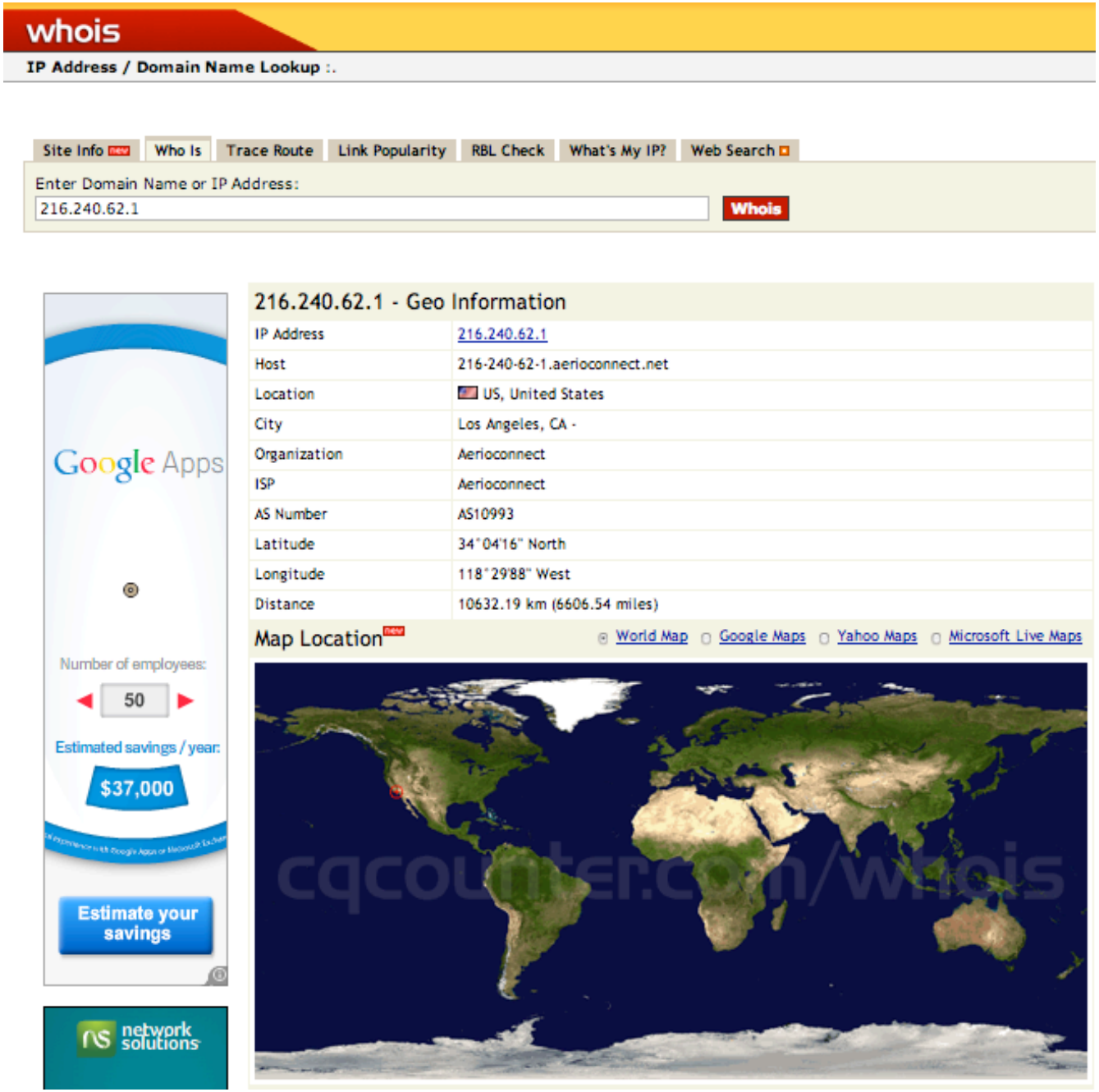
Each school was provided with a month of their web access logs (yes, a little old school, but it dovetails with the web page design unit that comes before this one) -- Recall the beauty that is automated logging (which we contrast with “purposeful” data collection)

```
216.240.62.* - [01/Feb/2010:12:28:15 -0800] "GET /South_East_HS/index.htm HTTP/
1.1" 200 14635 "http://notebook.lausd.net/portal/page?
_pageid=33,54194&_dad=ptl&_schema=PTL_EP&school_code=8881" "Mozilla/4.0
(compatible; MSIE 7.0; Windows NT 5.1;
[xSP_2:e4603403d6e74a45a470377e59d3859a_213]; 797805673603)"
```

```
216.240.62.* - [01/Feb/2010:12:28:15 -0800] "GET /South_East_HS/Images/Banner/
silver_Home.png HTTP/1.1" 200 3880 "http://www.lausd.k12.ca.us/South_East_HS/
index.htm" "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1;
[xSP_2:e4603403d6e74a45a470377e59d3859a_213]; 797805673603)"
```

```
216.240.62.* - [01/Feb/2010:12:28:15 -0800] "GET /South_East_HS/IEmystyle.css
HTTP/1.1" 200 6919 "http://www.lausd.k12.ca.us/South_East_HS/index.htm" "Mozilla/
4.0 (compatible; MSIE 7.0; Windows NT 5.1;
[xSP_2:e4603403d6e74a45a470377e59d3859a_213]; 797805673603)"
```

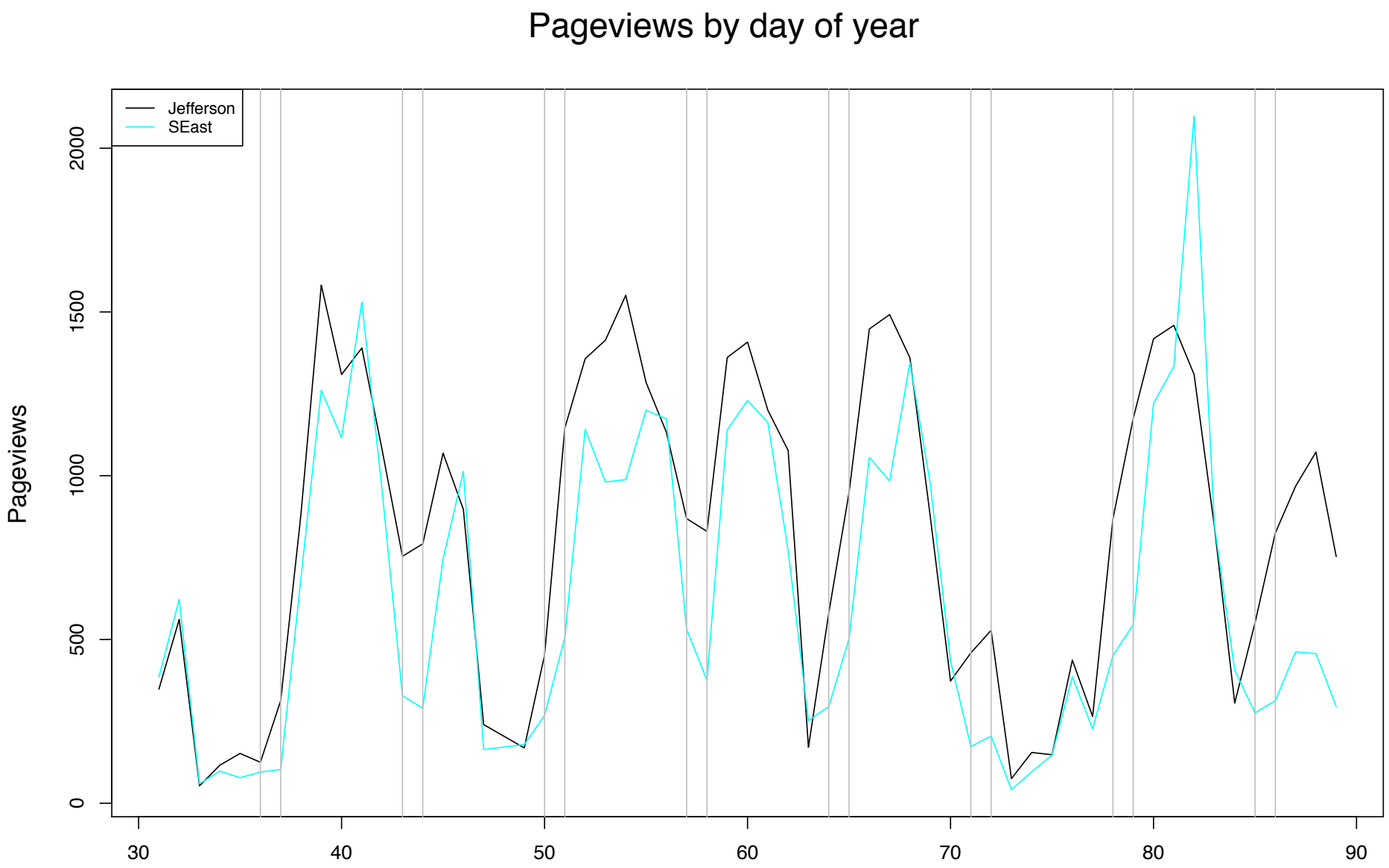
We walk the students through the meaning of many of these fields and then boil things down into an R data structure (telling them a bit about regular expressions along the way)

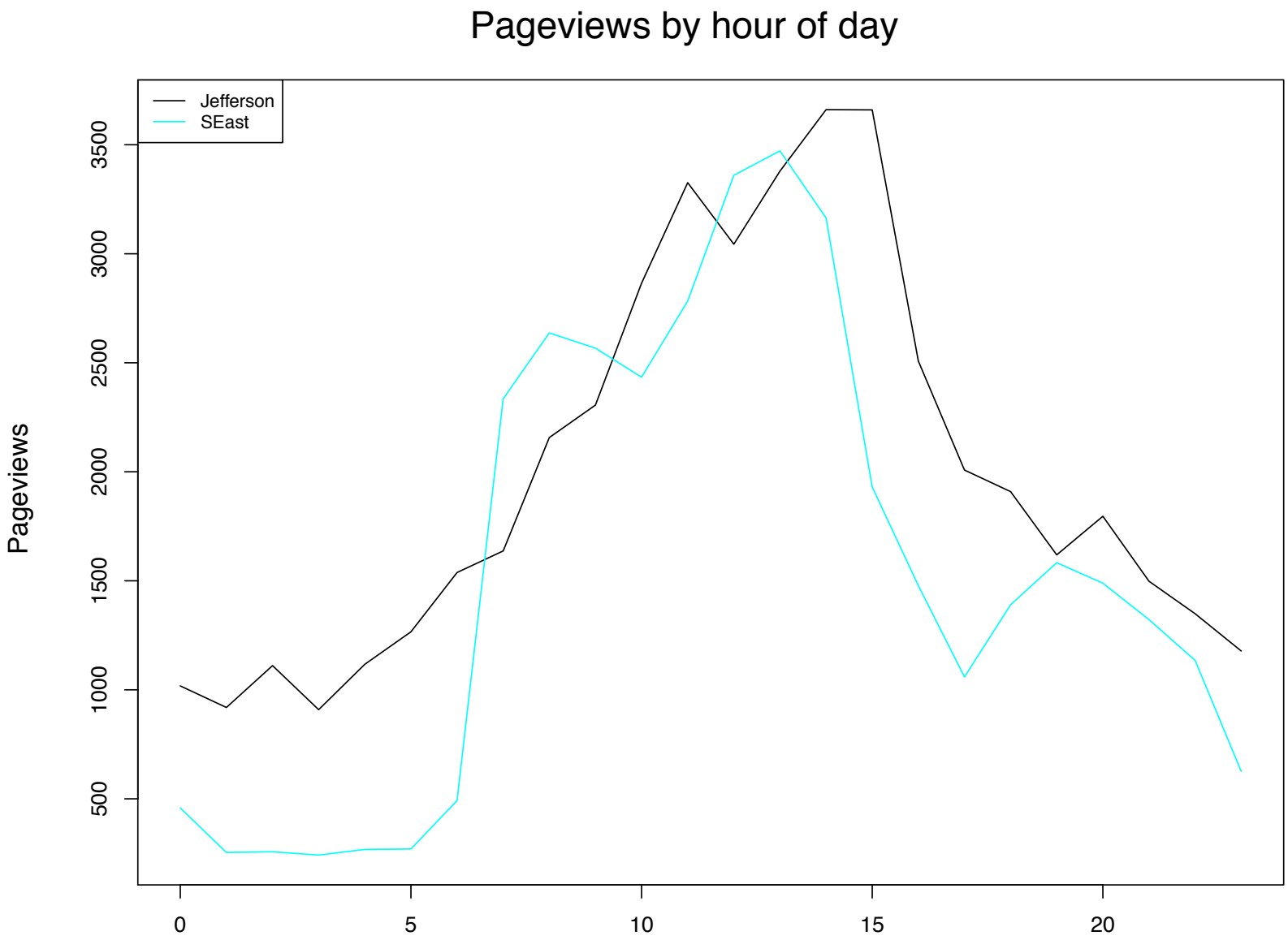


http://cqcounter.com/whois/

```
> load(url("http://mobilize.stat.ucla.edu/day3/data/sehs.Rda"))

> head(sehs)
             ip                  date                                                request code bytes
1  10.180.200.* 01/Feb/2010:11:56:08      /South_East_HS/Images/roundedcornr_1_tr.png  404   241
2  10.180.200.* 01/Feb/2010:11:56:08              /South_East_HS/roundedcornr_2_r.png  404   241
5  10.180.200.* 01/Feb/2010:11:56:23 /South_East_HS/Images/Banner/white_ContactUs.png  200  4723
6  10.180.200.* 01/Feb/2010:11:56:25              /South_East_HS/roundedcornr_2_r.png  404   253
7  10.180.200.* 01/Feb/2010:11:56:25      /South_East_HS/Images/roundedcornr_1_tr.png  404   253
11 10.180.200.* 01/Feb/2010:11:56:27                          /South_East_HS/favicon.ico  404   253
                                        referrer
1  http://www.lausd.k12.ca.us/South_East_HS/index.htm
2  http://www.lausd.k12.ca.us/South_East_HS/index.htm
5  http://www.lausd.k12.ca.us/South_East_HS/index.htm
6  http://www.lausd.k12.ca.us/South_East_HS/index.htm
7  http://www.lausd.k12.ca.us/South_East_HS/index.htm
11                                                 -

agent
1
Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)
2
Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)
5          Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727;
InfoPath.2; OfficeLiveConnector.1.3; OfficeLivePatch.0.0; .NET CLR 3.0.4506.2152; .NET CLR 3.5.30729)
6  Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; GTB6.4; .NET CLR 1.1.4322; .NET CLR 2.0.50727;
InfoPath.2; OfficeLiveConnector.1.3; OfficeLivePatch.0.0; .NET CLR 3.0.4506.2152; .NET CLR 3.5.30729)
7  Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; GTB6.4; .NET CLR 1.1.4322; .NET CLR 2.0.50727;
InfoPath.2; OfficeLiveConnector.1.3; OfficeLivePatch.0.0; .NET CLR 3.0.4506.2152; .NET CLR 3.5.30729)
11 Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; GTB6.4; .NET CLR 1.1.4322; .NET CLR 2.0.50727;
InfoPath.2; OfficeLiveConnector.1.3; OfficeLivePatch.0.0; .NET CLR 3.0.4506.2152; .NET CLR 3.5.30729)
```

```
# create a date object
> just_dates <- strptime(sehs$date,format="%d/%b/%Y:%H:%M:%S")
> class(just_dates)
[1] "POSIXt"  "POSIXlt"

> head(just_dates)
[1] "2010-02-01 11:56:08" "2010-02-01 11:56:08" "2010-02-01 11:56:23" "2010-02-01 11:56:25"
[5] "2010-02-01 11:56:25" "2010-02-01 11:56:27"

> just_dates[1]
[1] "2010-02-01 11:56:08"
> just_dates[100]
[1] "2010-02-01 12:00:52"
> just_dates[1] < just_dates[100]
[1] TRUE

> just_dates[100] - just_dates[1]
Time difference of 4.733333 mins

> just_dates[100]
[1] "2010-02-01 12:00:52"

# add a second
> just_dates[100] + 1
[1] "2010-02-01 12:00:53 PST"

# ... or a minute...
> just_dates[100] + 60
[1] "2010-02-01 12:01:52 PST"

# ... or an hour...
> just_dates[100] + 60*60
[1] "2010-02-01 13:00:52 PST"

# ... or a day.
> just_dates[100] + 24*60*60
[1] "2010-02-02 12:00:52 PST"
```

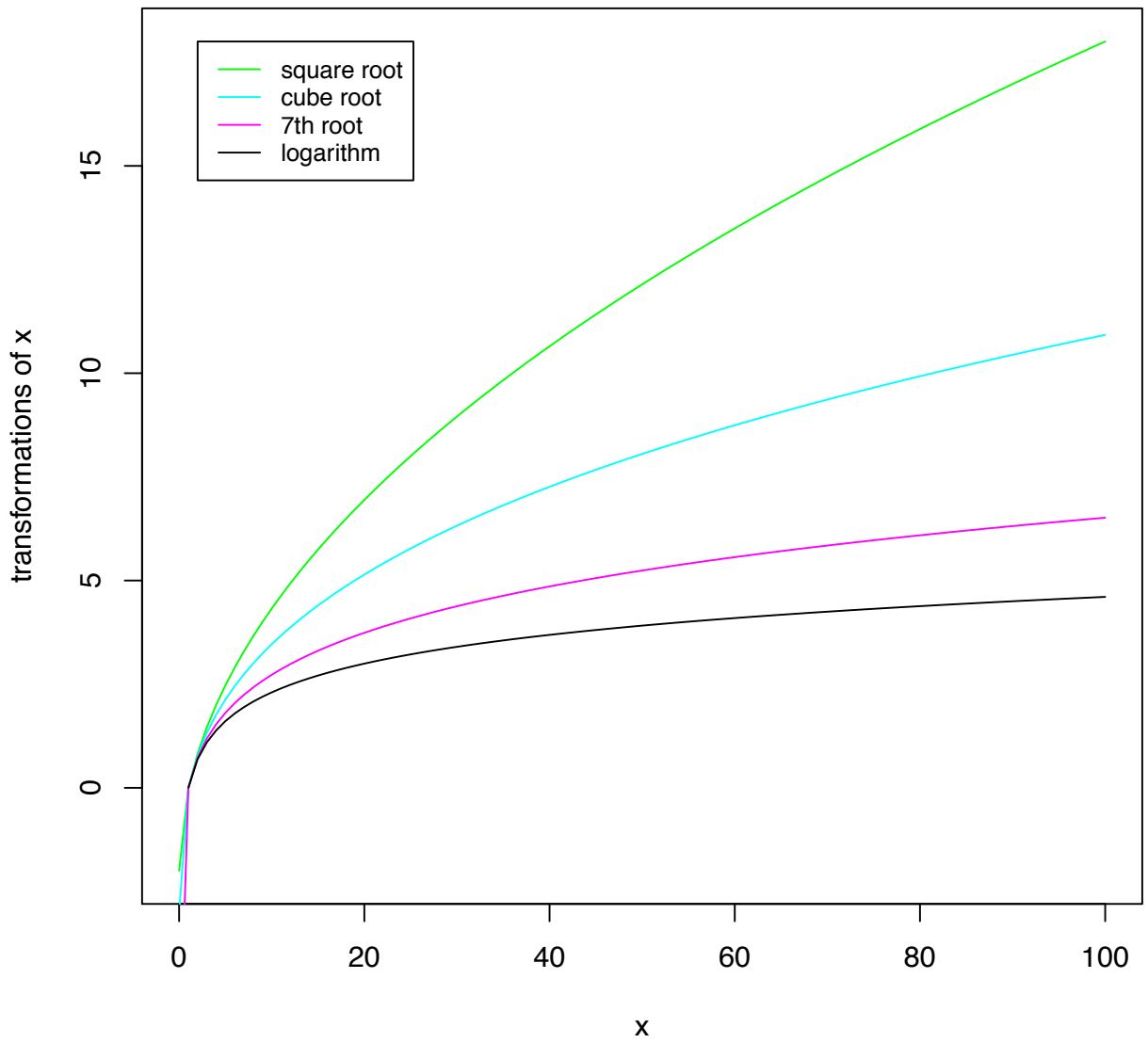

Pageviews by day of year
Jefferson
SEast
Pageviews
0
500
1000
1500
2000
30
40
50
60
70
80
90


Pageviews by hour of day
Jefferson
SEast
Pageviews
500
1000
1500
2000
2500
3000
3500
0
5
10
15
20

/error.html	8895
/Jefferson_HS/lscomp.htm	6233
/Jefferson_HS/index.htm	3738
/Jefferson_HS/tdmacwin.htm	1067
/Jefferson_HS/sped.htm	806
/Jefferson_HS/tdar.htm	797
/Jefferson_HS/alum.htm	632
/Jefferson_HS/lessons/shayes/index.html	591
/Jefferson_HS/sports.htm	546
/Jefferson_HS/lessons/shayes/ap.html	538
/Jefferson_HS/dept.htm	532
/Jefferson_HS/td_7zip.html	496
/Jefferson_HS/standards/sampl_syllabus.html	441
/Jefferson_HS/jtour.htm	433
/Jefferson_HS/handbk/thandbook.htm	426
/Jefferson_HS/tdexcel.htm	320
/Jefferson_HS/tchdocs.htm	314
/Jefferson_HS/scifair.htm	303

/error.html	12745
/South_East_HS/index.htm	5604
/South_East_HS/Faculty.htm	1488
/South_East_HS/Bell_Schedules.htm	1215
/South_East_HS/Calendar_Student.htm	1141
/South_East_HS/Current_Events.htm	987
/South_East_HS/Sports.htm	955
/South_East_HS/PhotoStory.htm	905
/South_East_HS/Academies.htm	698
/South_East_HS/College_Center.htm	661
/South_East_HS/Academy_VAPA.htm	656
/South_East_HS/Academy_California_.htm	645
/South_East_HS/Clubs.htm	608
/South_East_HS/Academy_ArchitectureEngineering.htm	602
/South_East_HS/Academy_BusinessFinance.htm	602
/South_East_HS/Academy_TechMedia.htm	601
/South_East_HS/Academy_JusticeLaw.htm	588
/South_East_HS/Departments.htm	543

```
> bytes_per_day <- aggregate(sehs$bytes,by=list(wday=sehs$wday,yday=sehs$yday),sum)
> plot(bytes_per_day$yday,bytes_per_day$x/1048576,type="l")
> abline(v=bytes_per_day$yday[bytes_per_day$wday==0])
> abline(v=bytes_per_day$yday[bytes_per_day$wday==6])

> from_google <- subset(sehs,grepl("google",referrer))
> dim(from_google)
[1] 2329    7

> head(from_google$referrer)

[1] "http://www.google.com/search?q=southeast+high+school&ie=UTF-8&oe=UTF-8&hl=en&client=safari"
[2] "http://www.google.com/search?source=ig&hl=en&rlz=&q=south+east+high+school+south+gate
+california&aq=9&aqi=g10&oq=south+east+high+"
[3] "http://www.google.com/search?hl=en&source=hp&q=SOUTH+EAST+HIGH+SCHOOL&aq=f&aqi=g10&oq="
[4] "http://www.google.com/search?hl=en&source=hp&q=south+east+high+school&aq=0&aqi=g10&oq=south+east+hi"
[5] "http://www.google.com/hws/dell/afe?hl=en&s=http://ww.southeasthighschool.com/"
[6] "http://www.google.com/search?q=south+east+high+school
+california&ie=utf-8&oe=utf-8&aq=t&rls=org.mozilla:en-US:official&client=firefox-a"

> not_google <- subset(from_google,
+                      !grepl("south",referrer,ignore.case=TRUE) &
+                          !grepl("east",referrer,ignore.case=TRUE))

> tail(not_google$referrer)

[1] "http://www.google.com/cse?cx=012518883040023400225%3A30biird4qzs&cof=FORID%3A11&q=high+school+
%236&sa=Search&ad=w9&num=10&rurl=http%3A%2F%2Fwww.lausd.net%2Fsearch%2Findex.html%3Fcx
%3D012518883040023400225%253A30biird4qzs%26cof%3DFORID%253A11%26q%3Dhigh%2Bschool%2B%25236%26sa%3DSearch"
[2] "http://www.google.com/cse?cx=012518883040023400225%3A30biird4qzs&cof=FORID%3A11&q=maria
+che&sa=Search&ad=w9&num=10&rurl=http%3A%2F%2Fwww.lausd.net%2Fsearch%2Findex.html%3Fcx
%3D012518883040023400225%253A30biird4qzs%26cof%3DFORID%253A11%26q%3Dmaria%2Bche%26sa%3DSearch"
[3] "http://www.google.com.au/search?q=atlas+protocol&sourceid=ie7&rls=com.microsoft:en-
US&ie=utf8&oe=utf8&redir_esc=&ei=nuyyS4OZNY6OkQXilZioBA"
[4] "http://www.google.com/search?hl=en&client=safari&rls=en-us&q=%22diana+rendon%22+san+diego&start=40&sa=N"
[5] "http://www.google.com/cse?cx=012518883040023400225%3A30biird4qzs&cof=FORID%3A11&q=mr+downey
+jr&sa=Search&ad=w9&num=10&rurl=http%3A%2F%2Fwww.lausd.net%2Fsearch%2F%3Fcx
%3D012518883040023400225%253A30biird4qzs%26cof%3DFORID%253A11%26q%3Dmr%2Bdowney%2Bjr%26sa%3DSearch"
[6] "http://www.google.com/search?hl=en&q=rhian+donadelle&aq=f&aqi=&aql=&oq=&gs_rfai="
```


square root
cube root
7th root
logarithm
transformations of x
0
5
10
15
0
20
40
60
80
100
x

```
# R and automated browsing...

# The discussion of web access logs inevitably leads to questions of
# human v. robot actions. We show how R can be used to follow links
# on a page.

> library(XML)
> page <- htmlTreeParse("http://www.lausd.net/Jefferson_HS/index.htm",useInternalNodes=TRUE)
> getNodeSet(page,"//a[@href]")

> old_ops <- options()
> old_ops$HTTPUserAgent
[1] "R (2.11.0 x86_64-apple-darwin9.8.0 x86_64 darwin9.8.0)"

# Changing the user agent...

> options(HTTPUserAgent = "A program I made up.")
> new_ops <- options()

> new_ops$HTTPUserAgent
[1] "A program I made up."
```

Space

This unit starts with a study of a data set published by **LA Bike Count** -- We begin with questions about how one **describes location**, moving from addresses to longitude and latitude

Students then make simple maps of the US, California and LA -- Then focus on the transportation system in LA

We discuss three basic spatial objects: **Points, lines and regions** and illustrate them with Metro and Census data

We close with a largely self-guided analysis of the USGS “Did you feel it?” data set for the recent Baja California earthquake -- Which gives us an opportunity to compare CSV to XML formats

Results from the 2009 City of Los Angeles Bicycle and Pedestrian Count

Table 14: PM Count Locations
Pedestrian & Bicycle Data

PM Count Intersections	PM Ped	PM Bike
1st & Alameda	241	62
4th & Wilton	87	48
7th & Figueroa	1,979	216
8th & La Brea	272	72
9th & Pacific	160	58
Alvarado & 7th	625	150
Ballona Creek	181	353
Broadway Bridge	26	63
Eagle Rock & Colorado	246	53
Echo Park & Sunset	1,369	121
Figueroa & Pasadena	253	93
Florence & Graham	1,526	119
Fountain & Vermont	518	110
Glendale & Park	189	65
Hollywood & Highland	1,377	93
Hoover & McClintock	711	977
Idaho & Bundy	263	105
Kittridge & De Soto	191	56
LA River @ Baum Bridge	46	164
Lankershim & Vineland	213	97
Lincoln & Bluff Creek	32	35
Long Beach & Los Flores	224	39
Los Feliz & Riverside	124	97
Manchester & Hoover	407	73
National & Overland	147	42
Santa Monica & Highland	411	103
Santa Monica & Westwood	370	153
Sepulveda & Ohio	167	138
Sunset & Hyperion	542	145
Topanga & Burbank	98	35
Venice & National	287	118
Washington & Mildred Ave	154	391
Washington & Compton	621	90
Westholme & Wilshire	162	52
Westwood & Le Conte	3,806	220
Wilshire & Western	2,303	155
Woodman & Orange Line Station	133	49
York & Ave 50	180	40
Totals	20,635	5,043

Get Lat Lon – find the latitude and longitude of a point on a map

http://www.getlatlon.com/

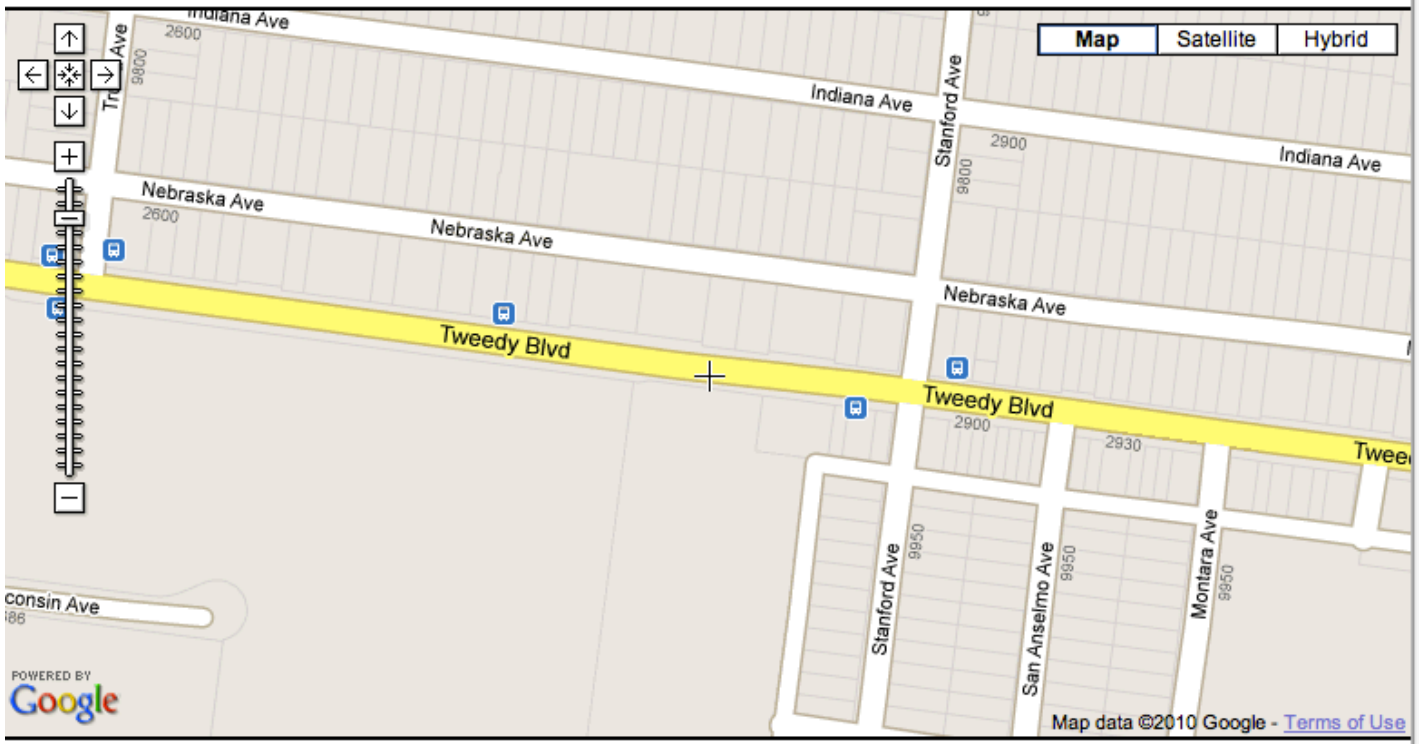
Google

Apple Yahoo! Wikipedia SelectorGadget

Get Lat Lon

Find the latitude and longitude of a point on a map.

Place name: 2720 Tweedy Boulevard Sc Zoom to place Zoom to my location (by IP)

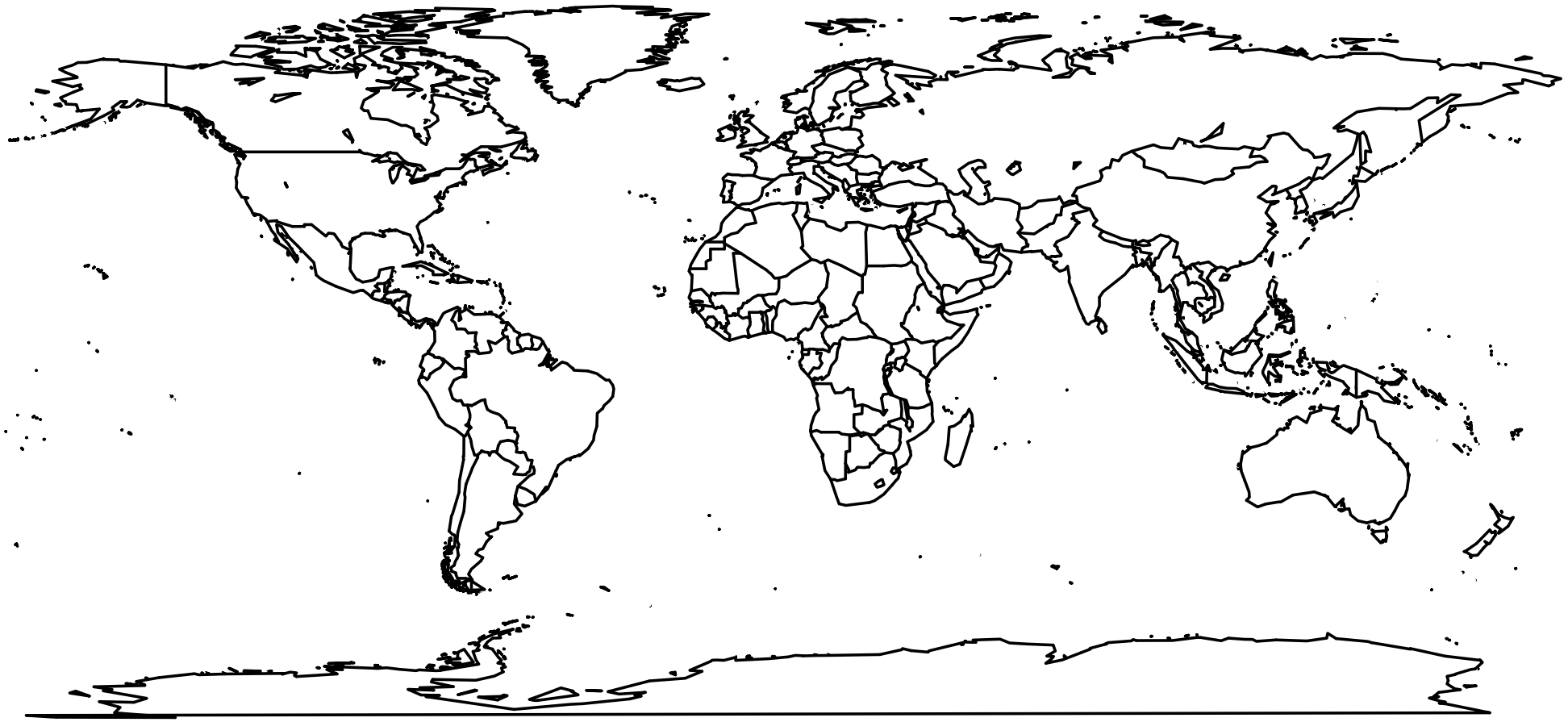


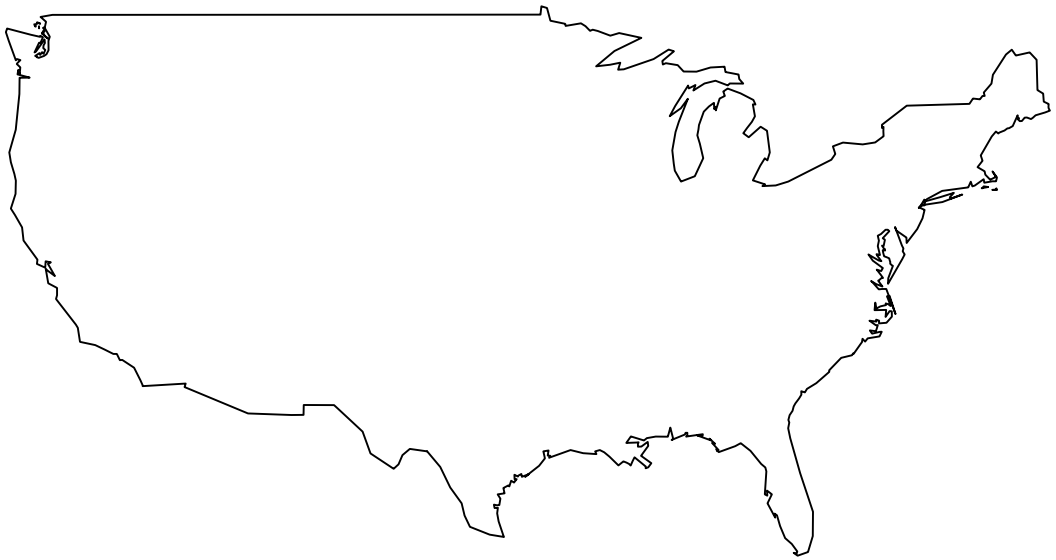
Latitude, Longitude: 33.9459274, -118.2216349

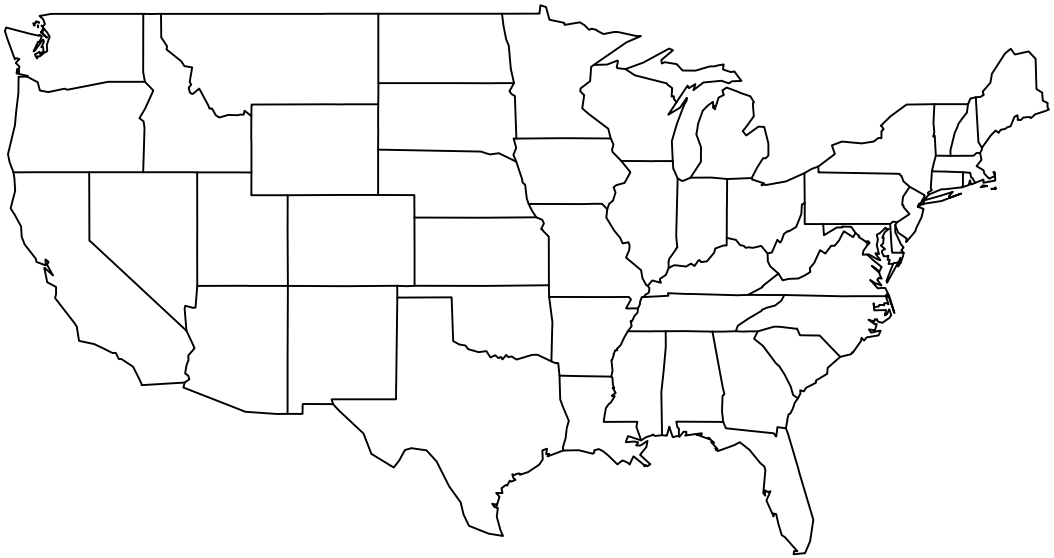
WKT: POINT(-118.2216349 33.9459274)

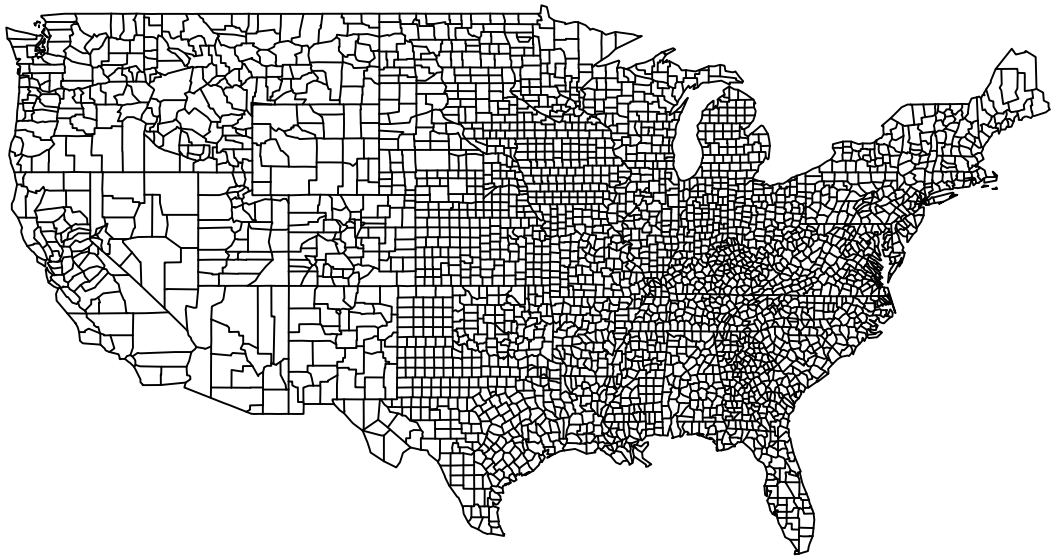
Google Maps zoom level: 17

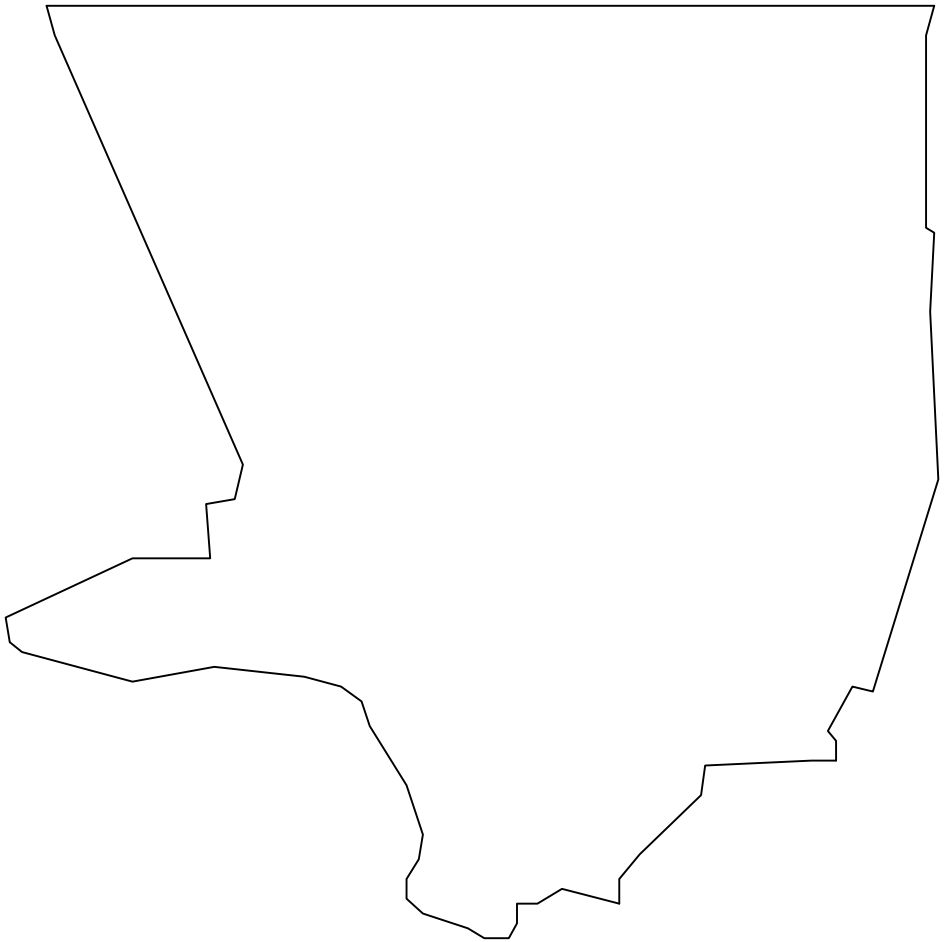
Timezone: America/Los_Angeles





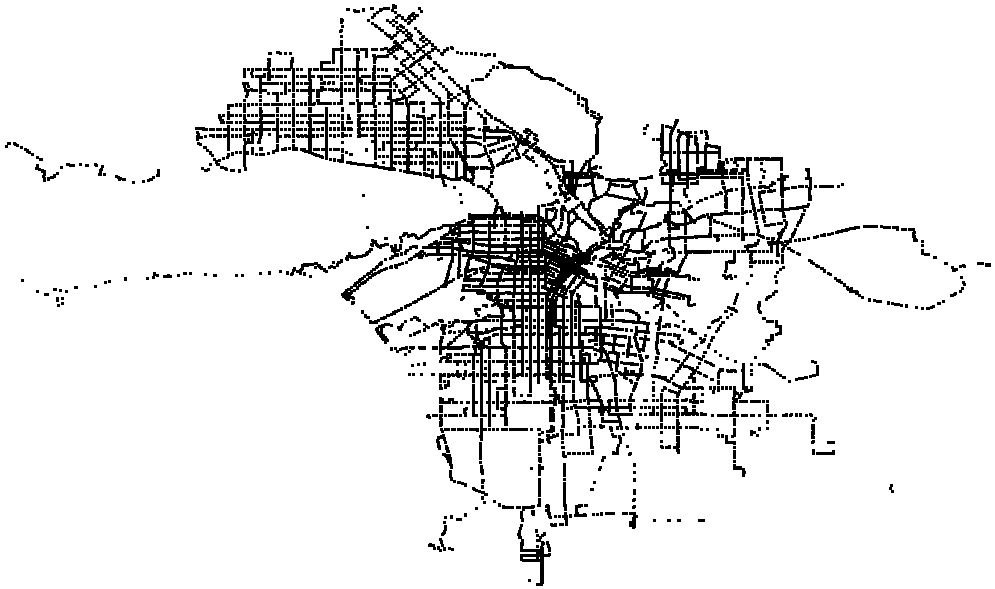






```
> library(maps)

> map('world')
> map('world','Mexico')
> map('world','Canada')

# The call to the function map() specifies first the database (in this
# case 'world') and then the optional names of regions to plot. The
# built-in databases are called 'world', 'usa', 'state' and 'county',
# with the last three referring to the United States. Here are some
# simple examples.

> map('state')
> map('state','california')

> map('county')
> map('county','california')
```

Spatial objects: Points –– Bus Stops in Los Angeles



```
# Spatial data: Points

> load(url("http://mobilize.stat.ucla.edu/day4/data/buses.Rda"))
> class(bus_stops)
[1] "SpatialPointsDataFrame"
attr(,"package")
[1] "sp"

> dim(bus_stops)
[1] 15536     3
> names(bus_stops)
[1] "stopnum" "Along"   "At"

> bbox(bus_stops)
            min        max
long -118.86083 -117.80931
lat    33.70685   34.32634

> summary(bus_stops)
Object of class SpatialPointsDataFrame
Coordinates:
            min        max
long -118.86083 -117.80931
lat    33.70685   34.32634
Number of points: 15536
Data attributes:
    stopnum           Along                 At
 1      :    1   BROADWAY:  256              :  312
 1,000  :    1   SUNSET  :  248    BROADWAY  :   98
 1,009  :    1   VERMONT :  230    WESTERN   :   82
 1,010  :    1   WILSHIRE:  209    FIGUEROA  :   78
 1,011  :    1   WESTERN :  196    WASHINGTON:   71
 1,012  :    1   OLYMPIC :  195    VERMONT   :   69
 (Other):15530   (Other) :14202   (Other)    :14826
```

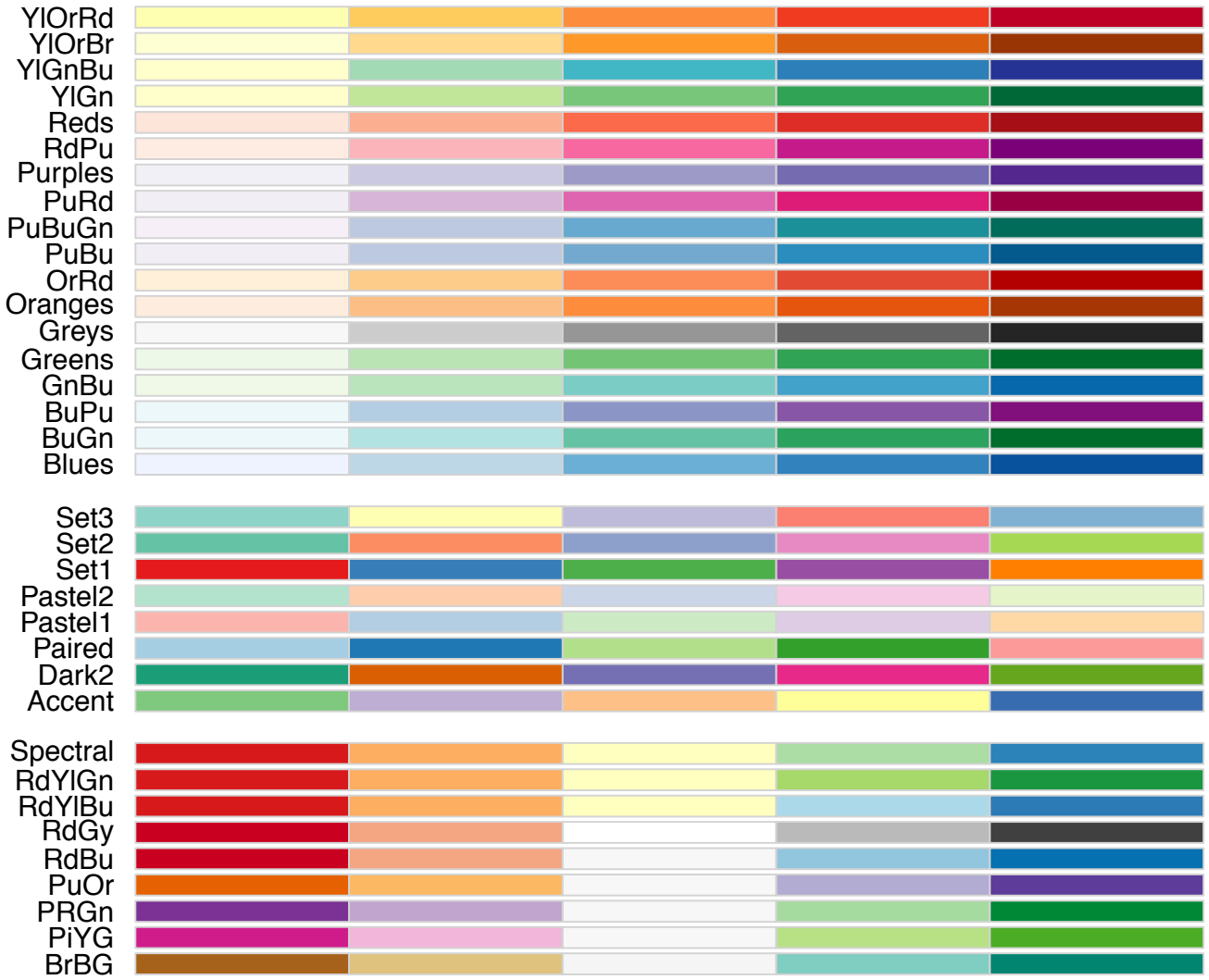
```
> class(bus_stops$Along)
[1] "factor"

> head(bus_stops$Along)
[1] PARAMOUNT JEFFERSON 117TH     43RD      120TH     120TH
865 Levels: 10231 SUNLAND 10238 SUNLAND 103RD 10TH 111TH 117TH 119TH ... ZOO

> tail(sort(table(bus_stops$Along)))

 OLYMPIC  WESTERN WILSHIRE  VERMONT   SUNSET BROADWAY
     195      196      209      230      248      256

> plot(bus_stops,pch=".")
> plot(bus_stops[bus_stops$Along=="SUNSET",],pch=".",col="blue",add=TRUE)
> plot(bus_stops[bus_stops$Along=="VERMONT",],pch=".",col="green",add=TRUE)

# Making a SpatialPointsDataFrame

> labike <- read.csv(url("http://mobilize.stat.ucla.edu/day4/data/labike.csv"),head=TRUE)

> new_labike <- SpatialPointsDataFrame(labike[,c(2,3)],labike[,-c(2,3)])
> class(new_labike)

> dim(new_labike)
> names(new_labike)

> new_labike[1:10,]

# Again, we have now separated the coordinate information from the data
# about each survey location. You can return it to a data frame with the
# command

> as.data.frame(new_labike)
```

Spatial objects: Lines –– Bus Routes crossing downtown

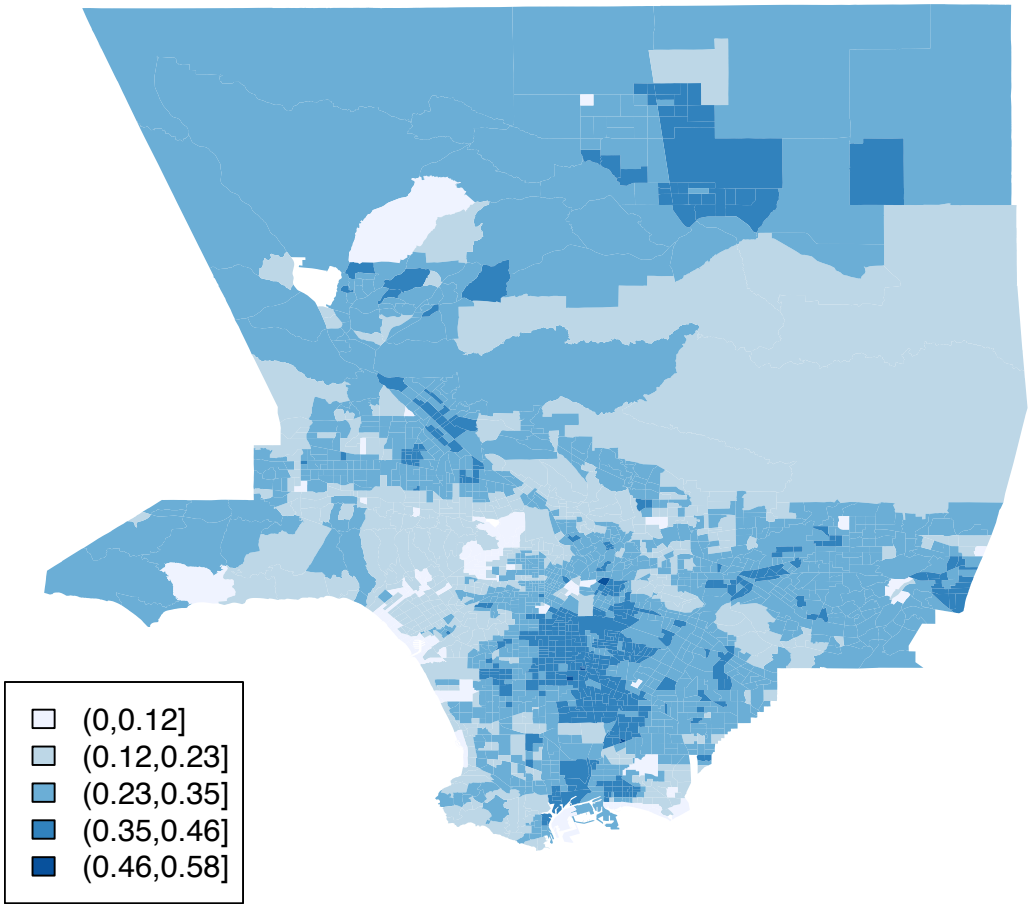
```
# Spatial data: Lines

# Download http://mobilize.stat.ucla.edu/day4/Metro.zip

> bus_routes <- readShapeSpatial("Metro/LocalCBD1209")

> class(bus_routes)
[1] "SpatialLinesDataFrame"
attr(,"package")
[1] "sp"

> dim(bus_routes)
[1] 291   4

> names(bus_routes)
[1] "VAR_ROUTE" "VAR_IDENT" "VAR_DIREC" "VAR_DESCR"

# Each row refers to a bus route and we can plot selected routes as follows...

> plot(bus_routes[1,])

> bus_routes$VAR_DESCR[1]
[1] End-End local June 2001
214 Levels: -378- to LASWEL -378- to LONMYR -70A- to TER028 ... WILVEO to 7THMAP

# ... or we can plot them all
> plot(bus_routes)
```

```
# Spatial data: Shapes

# Download the file http://mobilize.stat.ucla.edu/day4/data/Census.zip

> census <- readShapeSpatial("Census/Census2000")

> class(census)
[1] "SpatialPolygonsDataFrame"
attr(,"package")
[1] "sp"

> dim(census)
[1] 2052   35
> names(census)
 [1] "SP_ID"      "OBJECTID"   "STATE_FIPS" "CNTY_FIPS"  "STCOFIPS"
 [6] "TRACT"      "FIPS"       "POP2000"    "WHITE"      "BLACK"
[11] "AMERI_ES"   "ASIAN"      "HAWN_PI"    "OTHER"      "MULT_RACE"
[16] "HISPANIC"   "AGE_65_UP"  "MED_AGE"    "AVE_HH_SZ"  "OWNER_OCC"
[21] "RENTER_OCC" "MEDHHINC"   "PERCAPIN"   "AGE_UND18"  "RATE_BPOV"
[26] "NODIPLOMA"  "PROFESNLS"  "POPOVER25"  "WORK_OVR18" "WORK_CAR"
[31] "WORK_PUBLI" "WORK_MOTOR" "WORK_BIKE"  "WORK_PED"   "WORK_OTHER"
```

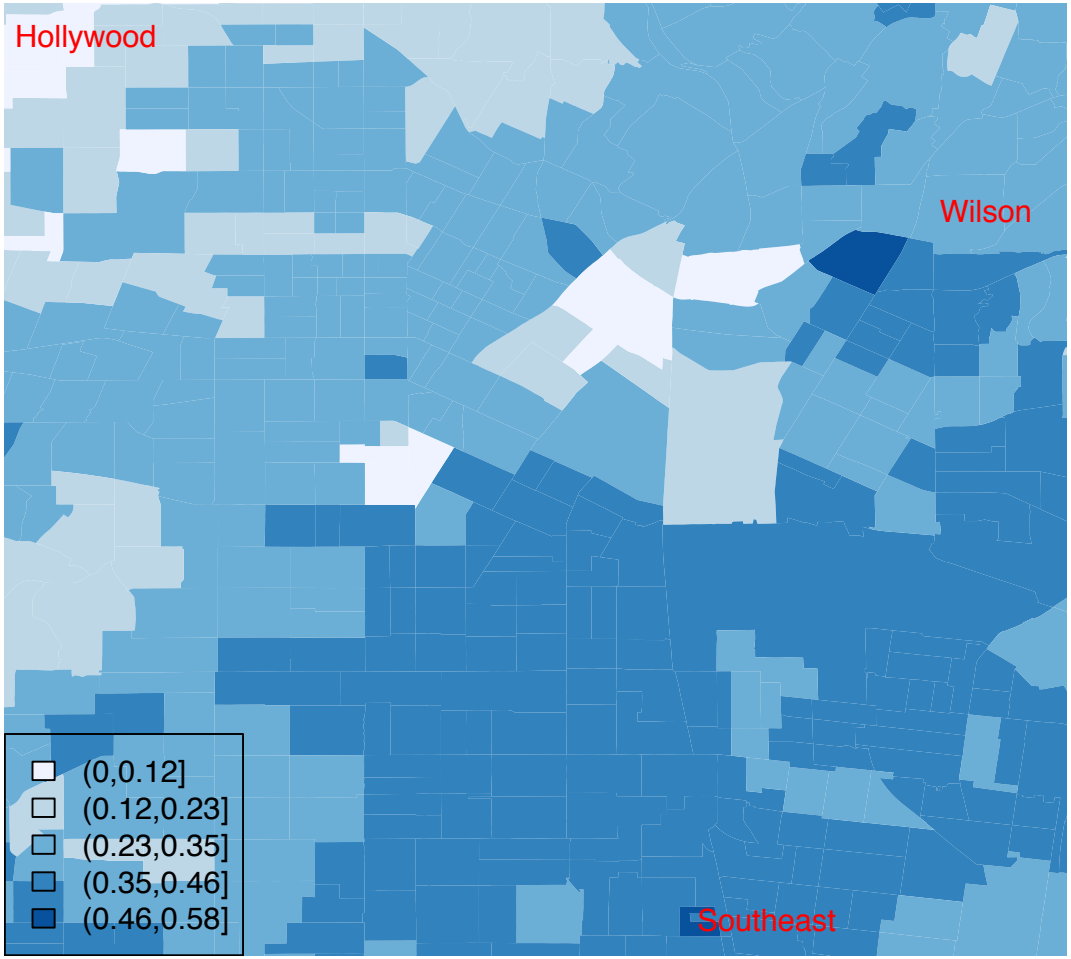
```
# ... grab locations of 3 LAUSD high schools (from getlatlon.com)

> tmp <- data.frame(
                longitude=c(-118.183523,-118.340061,-118.2216349),
                 latitude=c(34.06927,34.099182,33.9459274),
                    names=c("Wilson","Hollywood","Southeast"))

> hs <- SpatialPointsDataFrame(tmp[,c(1,2)],tmp[,3,drop=F])
> under18 <- census$AGE_UND18/census$POP2000

# slice the under18 variable...

> levs <- cut(under18,5,dig.lab=2)
> table(levs)

# ... pick a palette
> library(RColorBrewer)
> display.brewer.all(n=5)
> pal <- brewer.pal(5,"Blues")

> bb <- bbox(hs)
> plot(census,col=pal[levs],xlim=bb[1,],ylim=bb[2,],border=FALSE)

> legend("bottomleft",fill=pal,legend=levels(levs))
> text(hs,labels=hs$names,col="red")
> title("Proportion of People Under 18 by Census Tract")
```


YlOrRd
YlOrBr
YlGnBu
YlGn
Reds
RdPu
Purples
PuRd
PuBuGn
PuBu
OrRd
Oranges
Greys
Greens
GnBu
BuPu
BuGn
Blues
Set3
Set2
Set1
Pastel2
Pastel1
Paired
Dark2
Accent
Spectral
RdYlGn
RdYlBu
RdGy
RdBu
PuOr
PRGn
PiYG
BrBG

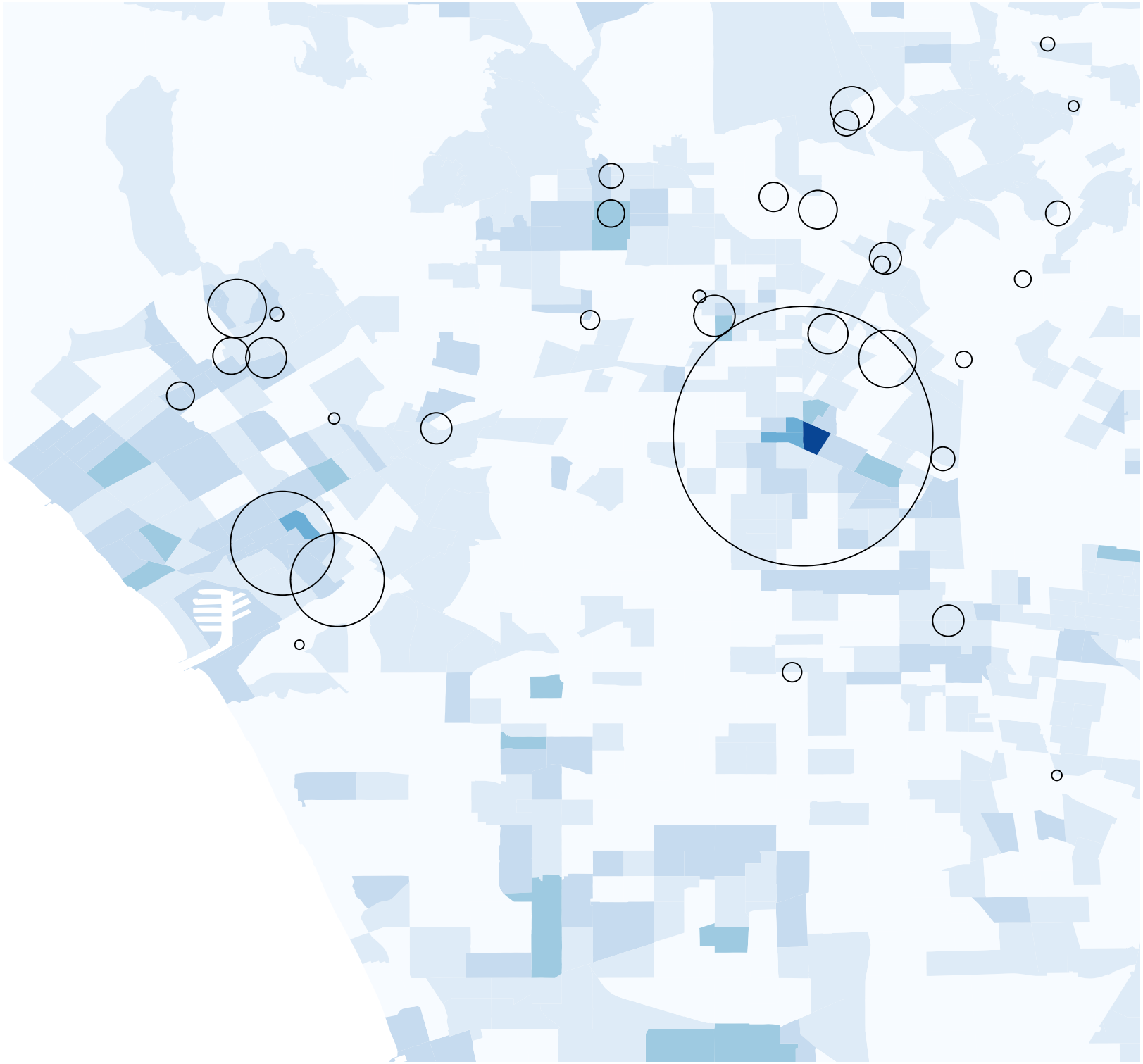
Proportion of People Under 18 by Census Tract

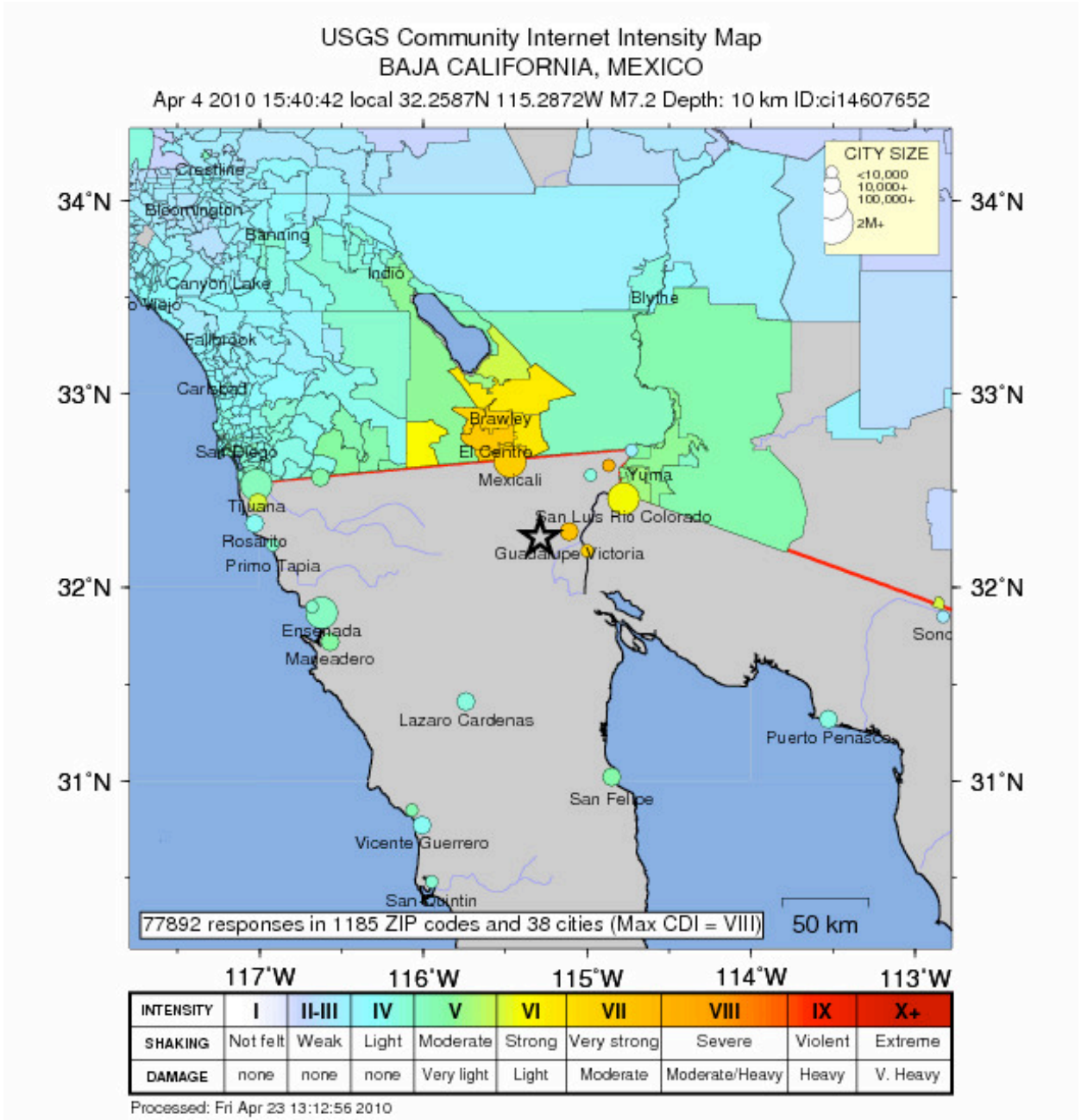


Proportion of People Under 18 by Census Tract



```
# And now the LA Bike Count data

> labike <- read.csv(url("http://mobilize.stat.ucla.edu/day4/data/labike.csv"),head=TRUE)

> levs <- cut(sqrt(census$WORK_BIKE),8)

> pal <- brewer.pal("Blues",n=8)

> plot(census,col=pal[levs],border=FALSE,xlim=c(-118.5,-118.2),ylim=c(33.9,34.1))

> symbols(labike$long,labike$lat,circles=labike$bike_count_pm,add=T)
```




USGS Community Internet Intensity Map
BAJA CALIFORNIA, MEXICO
Apr 4 2010 15:40:42 local 32.2587N 115.2872W M7.2 Depth: 10 km ID:ci14607652
CITY SIZE
<10,000
10,000+
100,000+
2M+
Crestline
Bloomington
Banning
Indio
Canyon Lake
Blythe
Fallbrook
Carlsbad
Brawley
San Diego
El Centro
Mexicali
Yuma
Tijuana
San Luis Rio Colorado
Rosarito
Primo Tapia
Guadalupe Victoria
Ensenada
Maneadero
Lazaro Cardenas
Puerto Penasco
San Felipe
Vicente Guerrero
San Quintin
77892 responses in 1185 ZIP codes and 38 cities (Max CDI = VIII)
50 km
34°N
33°N
32°N
31°N
117°W
116°W
115°W
114°W
113°W
INTENSITY I II-III IV V VI VII VIII IX X+
SHAKING Not felt Weak Light Moderate Strong Very strong Severe Violent Extreme
DAMAGE none none none Very light Light Moderate Moderate/Heavy Heavy V. Heavy
Processed: Fri Apr 23 13:12:56 2010

```
"30214",1.0,1,2890,33.4826,-84.4876,0,"Fayetteville","GA"
"37043",1.0,1,2629,36.5004,-87.2298,0,"Clarksville","TN"
"38119",1.0,1,2385,35.0788,-89.8450,0,"Memphis","TN"
"59102",1.0,1,1584,45.7744,-108.5816,0,"Billings","MT"
"59937",1.0,1,1762,48.5016,-114.5806,0,"Whitefish","MT"
"66762",1.0,1,1974,37.3931,-94.7060,0,"Pittsburg","KS"
"75249",1.0,1,1726,32.6443,-96.9625,0,"Dallas","TX"
"77030",2.0,1,1921,29.7083,-95.4009,0,"Houston","TX"
```

```
<cdidata>
  <cdi type="zip">
    <location name="30214">
      <cdi>1</cdi>
      <nresp>1</nresp>
      <dist>2890</dist>
      <lat>33.4826399272</lat>
      <lon>-84.4876145299</lon>
      <name>Fayetteville</name>
      <state>GA</state>
    </location>
    <location name="37043">
      <cdi>1</cdi>
      <nresp>1</nresp>
      <dist>2629</dist>
      <lat>36.5004094308</lat>
      <lon>-87.2297647075</lon>
      <name>Clarksville</name>
      <state>TN</state>
    </location>
    <location name="38119">
      <cdi>1</cdi>
      <nresp>1</nresp>
      <dist>2385</dist>
      <lat>35.0787618693</lat>
      <lon>-89.8449722897</lon>
      <name>Memphis</name>
      <state>TN</state>
    </location>
      ...
   </cdi>
</cdidata>
```

```
> baja <- read.csv(url(
  "http://earthquake.usgs.gov/earthquakes/dyfi/events/ci/14607652/uscdi_zip.txt"),head=FALSE)
> class(baja)
> dim(baja)
> names(baja)

> head(baja)
```

The data set "baja" has 1,219 rows, each referring to a different zip code. Notice that the variables are given names like "V1" becuase R did not receive any more descriptive information about them. We can repair this now.

```
> names(baja) <- c("zipcode","cdi","nresp","dist",
                   "latitude","longitude","hmm","city","state")
> head(baja)

> plot(baja$dist,baja$cdi)
```

(This plot works best if the graphics window is wider than it is tall.) What do you see here? Does it make sense?

We'll now have a simple look at the data using the "centers" of the ZIP codes, the longitude and latitude values...

```
> library(maps)
> map('usa')
> points(baja$long,baja$lat,pch=".")
```

... or focus just on California.

```
> map('states','california')
> points(baja$long,baja$lat,pch=".")
```

There are clearly more zip codes with participants reporting near the epicenter of the quake, but we see a cluster near big cities like Los Angeles -- This is because there are simply more ZIP codes in urban areas. Now, let's recreate the map of intensities by ZIP codes. To start, we'll need information about ZIP codes -- What we want is a shapefile describing their boundaries. The Census Bureau distributes these data at the web site below.

http://www.census.gov/geo/www/cob/z52000.html

Find California's shapefile (zt06_d00_shp.zip) and download. When you unpack it, you should have a folder called "zt06_d00_shp" and inside you will see the familiar ".shp", ".dbf" and ".shx" suffixes.

```
> library(maptools)
> zip <- readShapePoly("zt06_d00_shp/zt06_d00")

> class(zip)
> dim(zip)
> names(zip)
```

Again, we have a SpatialPolygonsDataFrame where each of the 2,490 rows refers to a different ZIP code in California and...

```
> plot(zip)
```

should produce a plot of them all. We can look at any individual ZIP code by subsetting.

```
> plot(zip[zip$NAME=="90069",])
> title("Home, W Hollywood, CA")
```

Now, pick your zip code and make a similar plot. Next, we reduce our "baja" ZIP code data set to include just those that are in California. Recall the logical operator "%in%".

```
> baja$zipcode %in% zip$ZCTA
```

The TRUE values correspond to "Did you feel it?" respondent zip codes that are contained in our data set of California ZIP codes. Now, we can use this to subset "baja".

```
> dim(baja)
> baja <- baja[baja$zipcode %in% zip$ZCTA,]

> dim(baja)
```

Now, let's form the colors that we'll assign to the zip codes in "zip". Start by creating a new variable of NA's.

```
> intensity <- rep(NA,nrow(zip))
> intensity[match(baja$zipcode,zip$ZCTA)] <- baja$cdi
```

The function match() will return values between 1 and the length of zip$ZCTA. For example, the first entry of

```
> match(baja$zipcode,zip$ZCTA)
```

is 1983. That means that baja$zipcode[1] is also zip$ZCTA[1983]. The code above, therefore, associates the "cdi" data in "baja" with the right row in zip. Now, let's create some colors for the "cdi" data.

```
> library(RColorBrewer)
> levs <- cut(intensity,9)
> table(levs)
```

We see that the levels are not very intuitive. Given that the "cdi" values are numbers like 1, 2, 2.1, 2.2... 4, 4.1, 4.2, ...

```
> table(intensity)
```

we might want to hand-craft our intervals a little. Instead of letting R decide how the intervals should be defined, we can provide the breaks. Here we set them to be 0, 1, 2, ... , 8. We then count the number of "intensity" values between 0 and 1, between 1 and 2 and so on.

```
> levs <- cut(intensity,breaks=c(0:8))
> table(levs)
```

We now have 8 different levels so we need 8 colors.

```
> pal <- brewer.pal(8,"Blues")
> cols <- pal[levs]
```

```
> cols
```

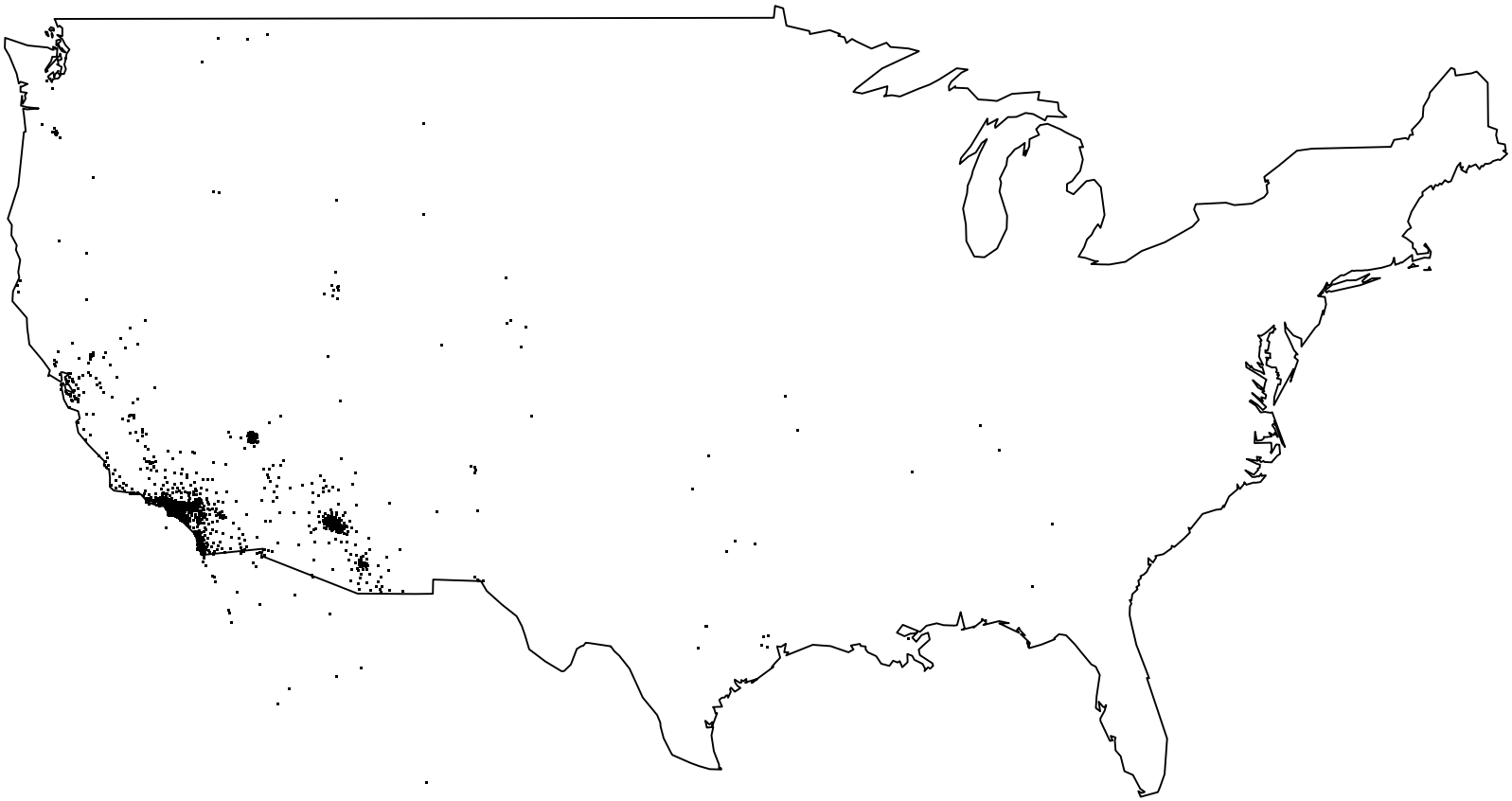
The "cols" variable has as many entries as there are rows in "zip" -- Each ZIP code has a color. Notice, however, that there are a number of NA's in the "cols" variable. This is because some ZIP codes had no respondents in the "Did you feel it?" data. For example, people living in extreme northern California probably didn't even notice the quake happened. If the colors are NA, then R will plot them as white. We will color these ZIP codes gray instead.

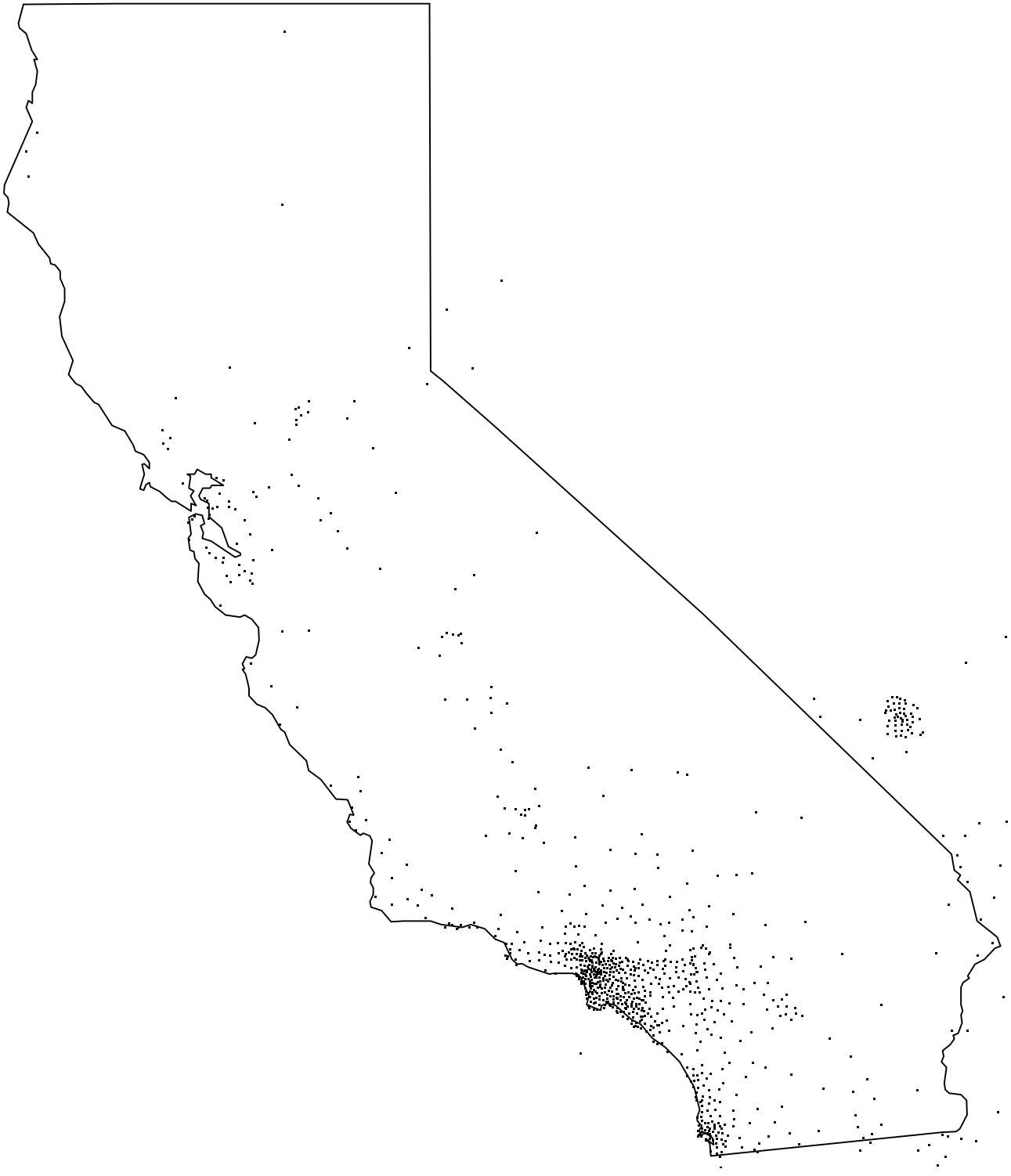
```
> cols[is.na(cols)] <- gray(0.9)
```

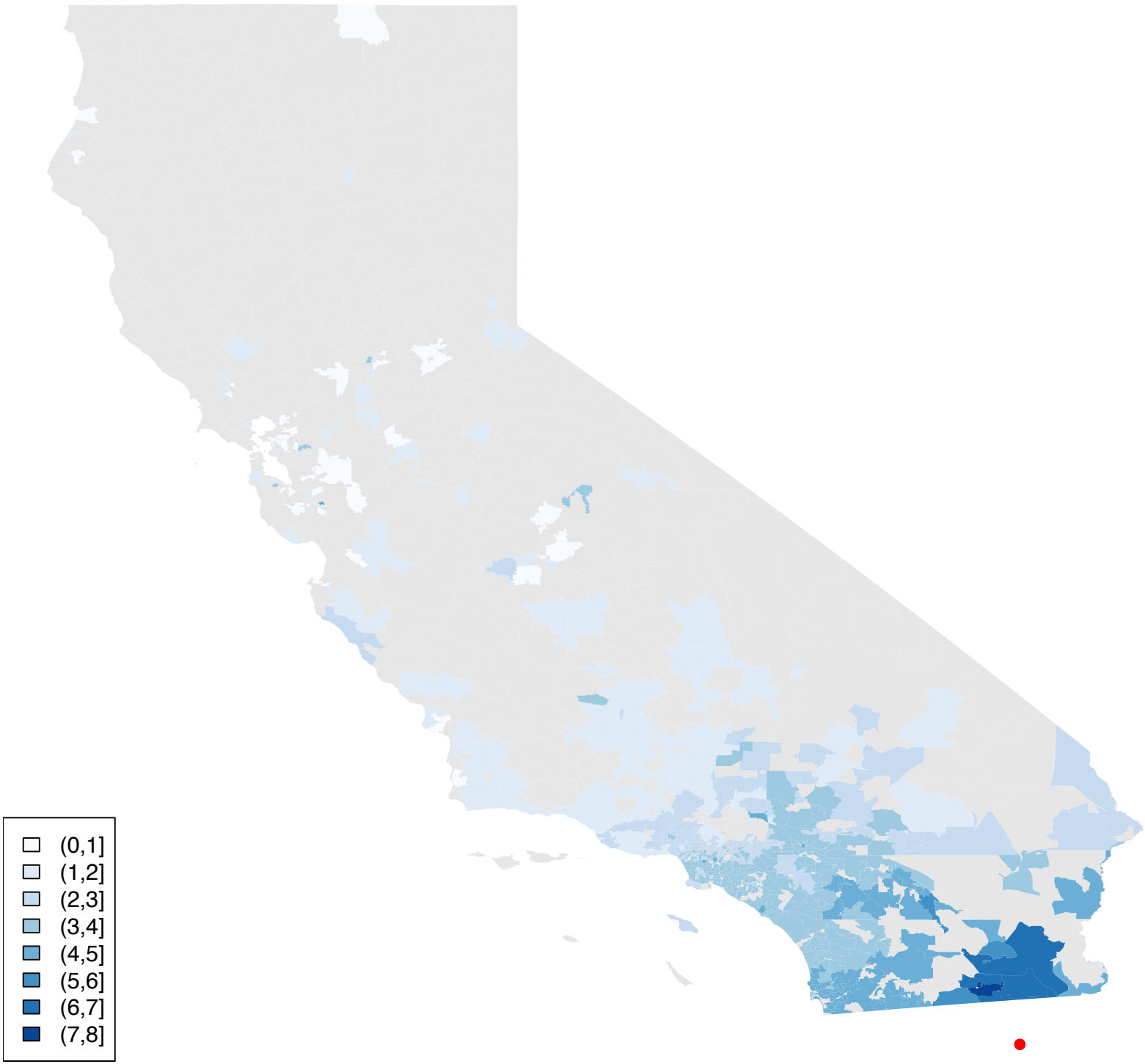
And now the plot!

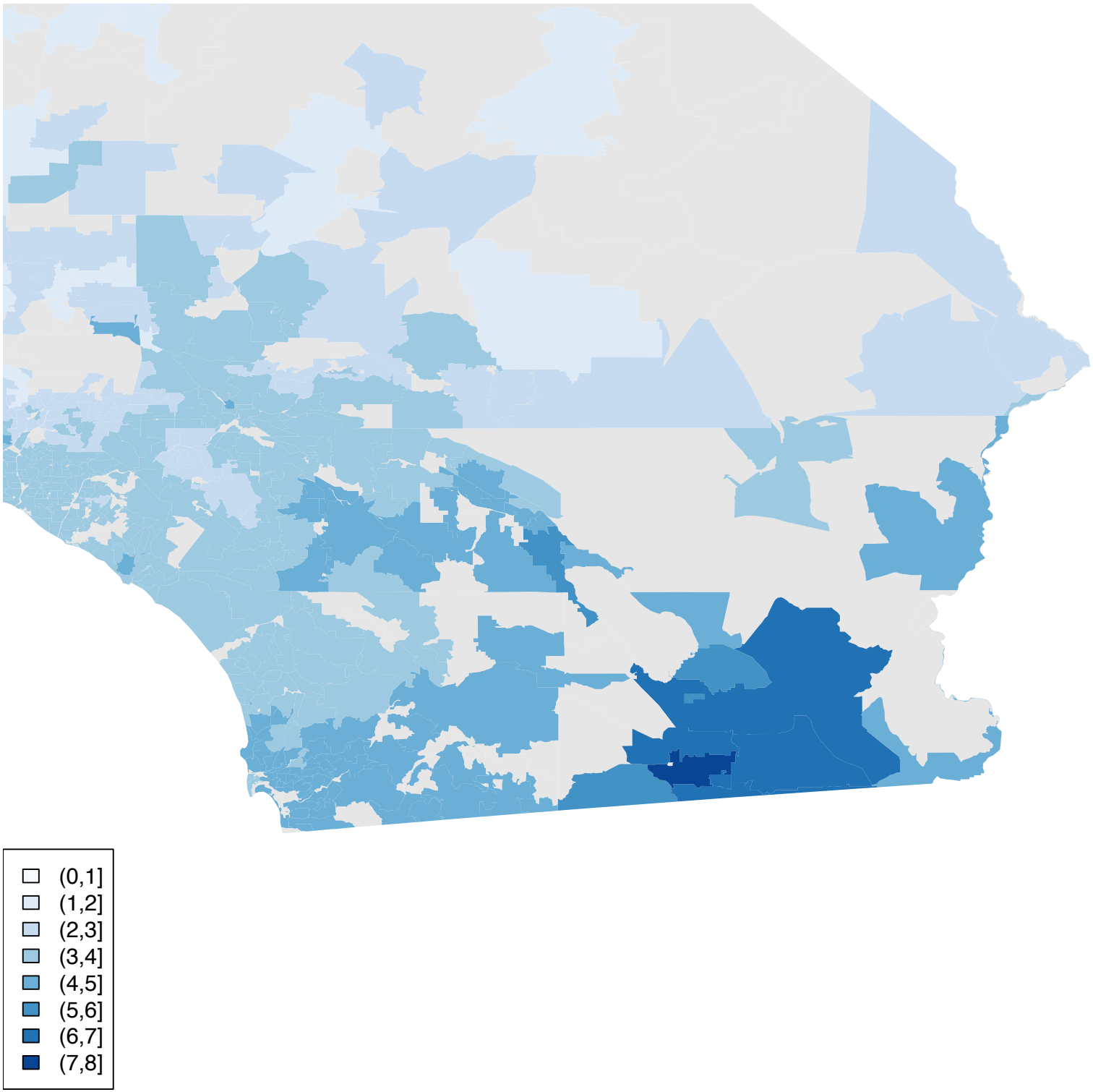
```
> plot(zip,col=cols,border=FALSE)
> points(-115.29,32.26,pch=19,col="red")
```

```
> legend("bottomleft",fill=pal,legend=levels(levs))
> title("Statewide CDI responses")
```






(0,1]
(1,2]
(2,3]
(3,4]
(4,5]
(5,6]
(6,7]
(7,8]

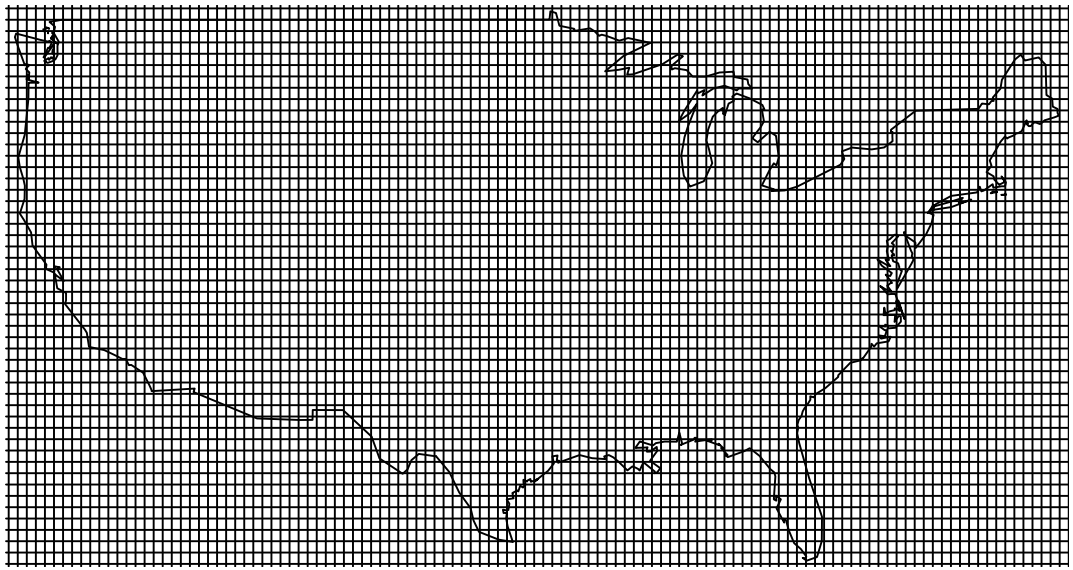

(0,1]
(1,2]
(2,3]
(3,4]
(4,5]
(5,6]
(6,7]
(7,8]

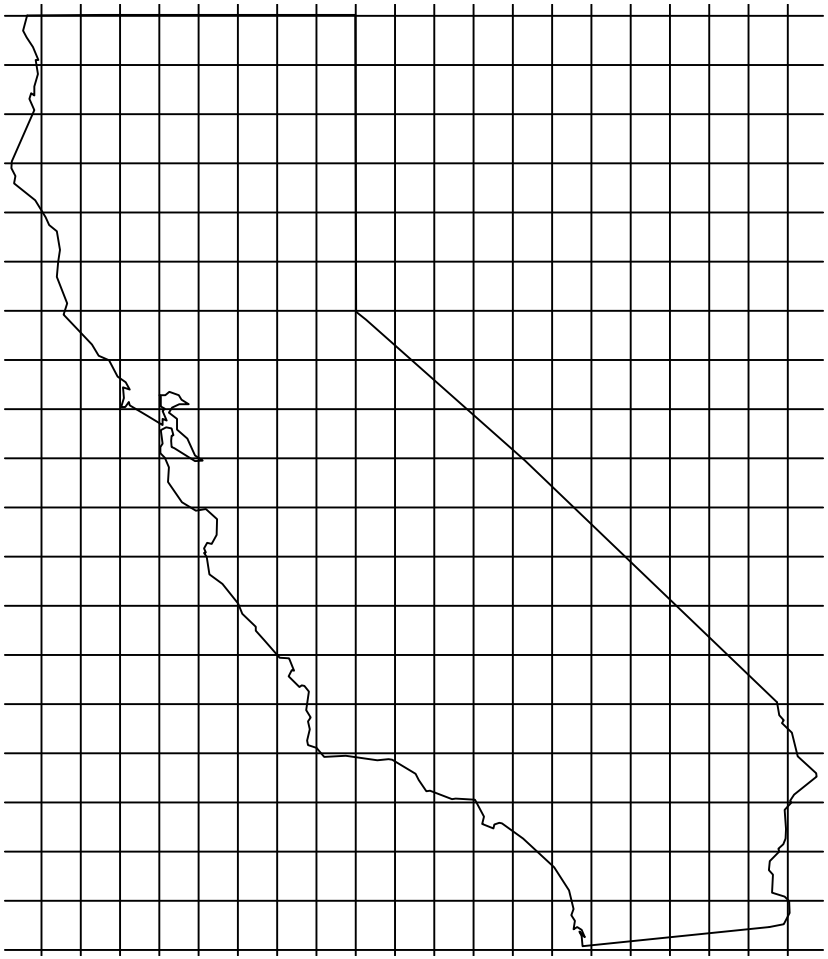
```
# Reading in the XML version of the data...

> library(XML)
> library(RCurl)

> u <- url("http://earthquake.usgs.gov/eqcenter/dyfi/events/nc/71299286/us/cdi_zip.xml")

> x <- getURL(u)
> class(x)
> x
> bajax <- xmlParse(x)
> bajax <- xmlToDataFrame(sfx,nodes=sfx['//location'])

# Putting points on a GoogleMap

> library(R2GoogleMaps)
> center = c(mean(locations$lat),mean(locations$lon))
> code = gmarker(locations$lat,locations$lon, addOverlay = TRUE)
> googleMapsDoc(code, center, zoom = 11, dim = c(750, 700),
                file = "first_try.html")
```

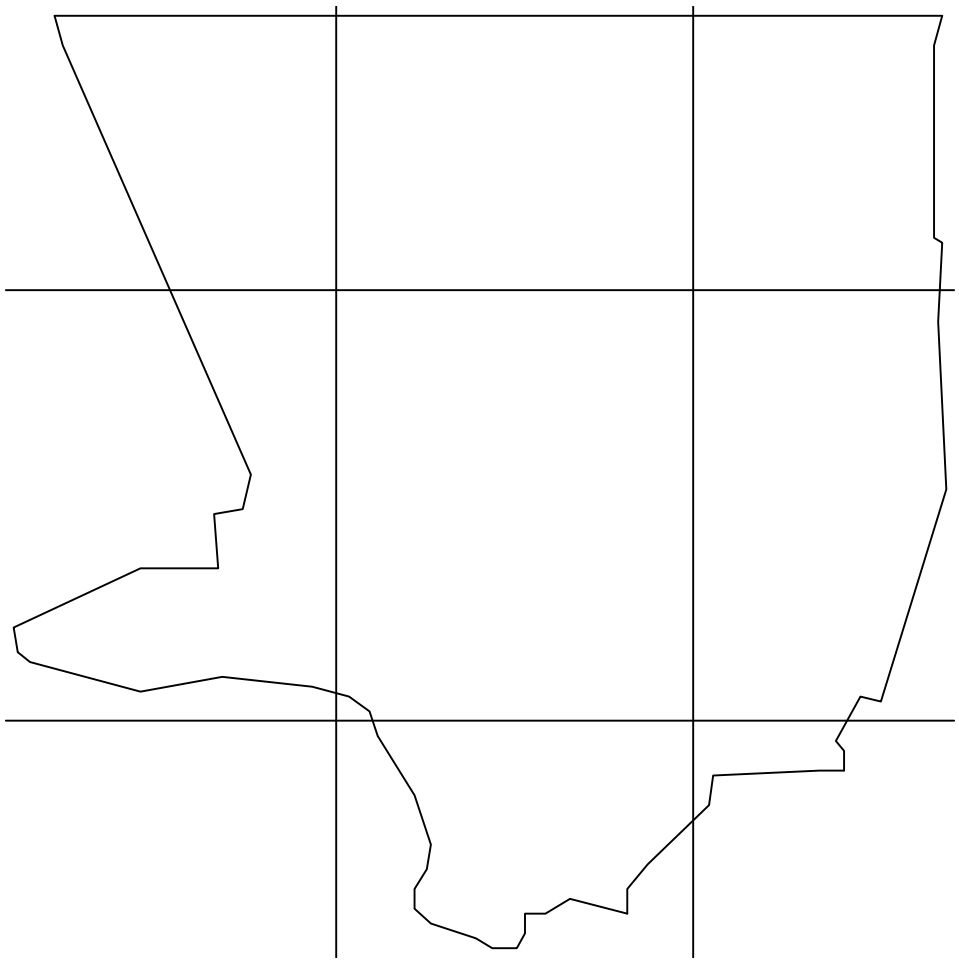
Gridded spatial data

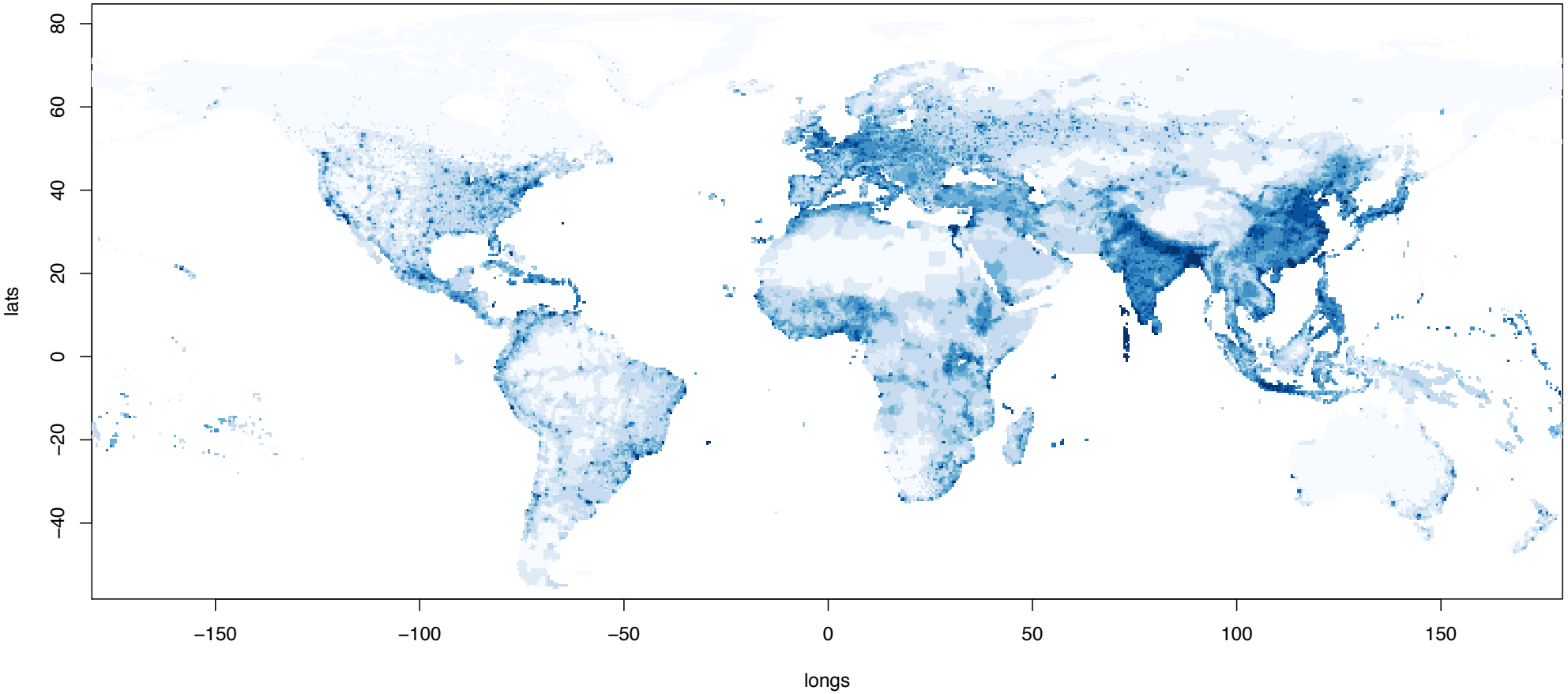
We talk about the **US Census and about censuses worldwide** -- These data have been aggregated by CIESIN at Columbia University (uh, here!) to provided a gridded (in longitude and latitude) map of the world's population

The grids will let us transition to images -- They also let us look at yet another web service publishing (this time) spatial data








lats
80
60
40
20
0
−20
−40
−150
−100
−50
0
50
100
150
longs

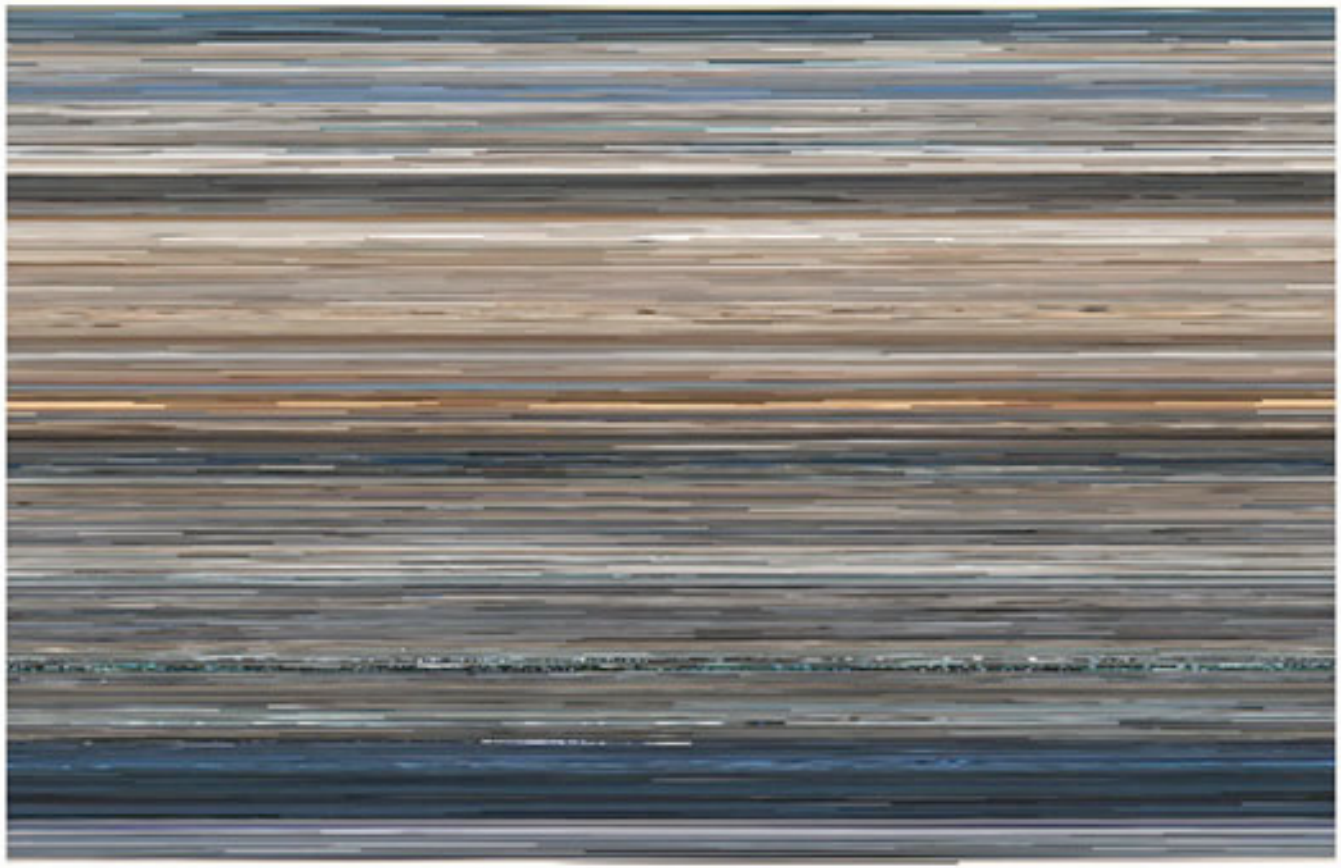
Images

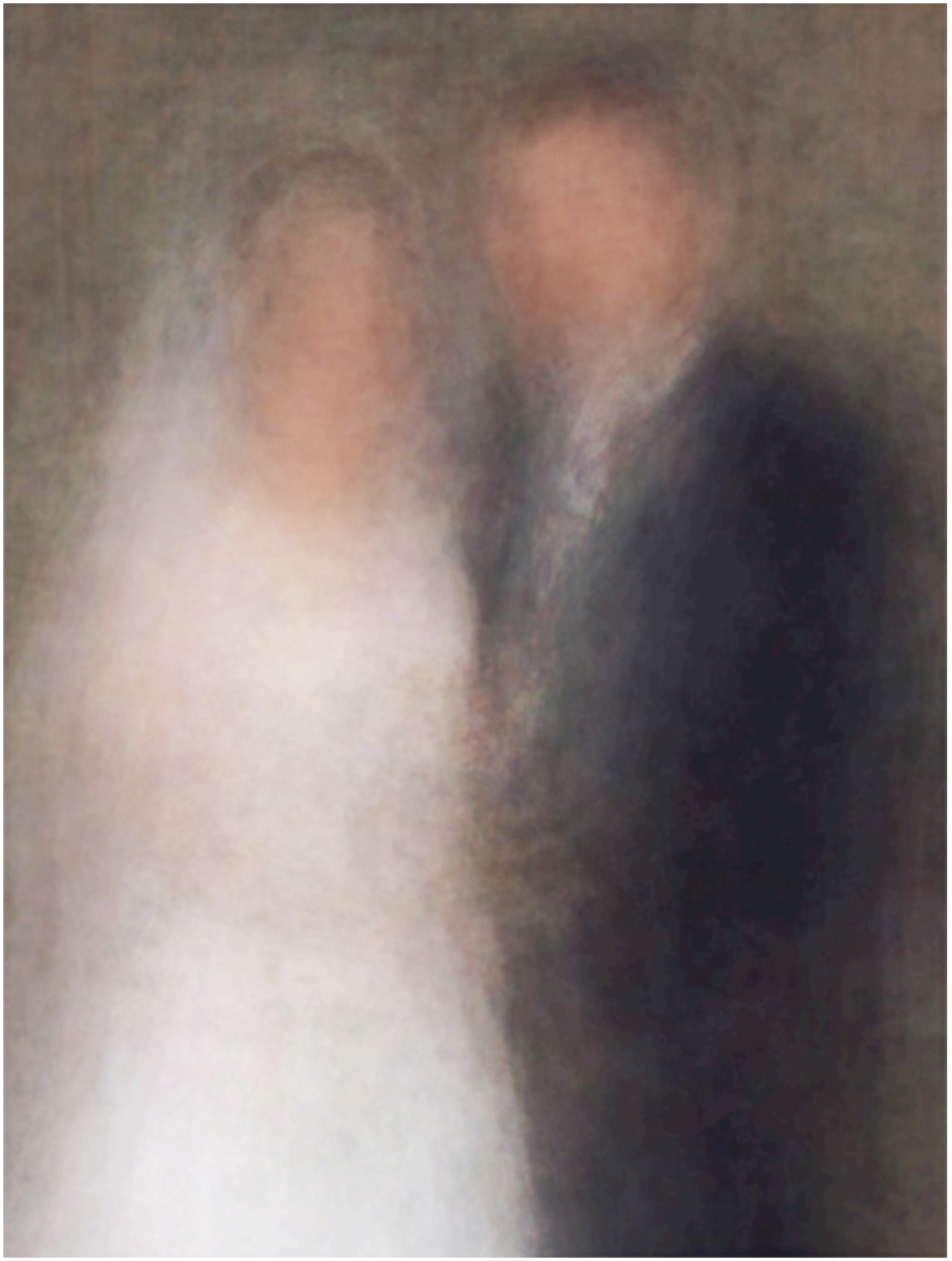
We begin with the structure of an image using RGB (essentially a grid x 3, or in R an array)

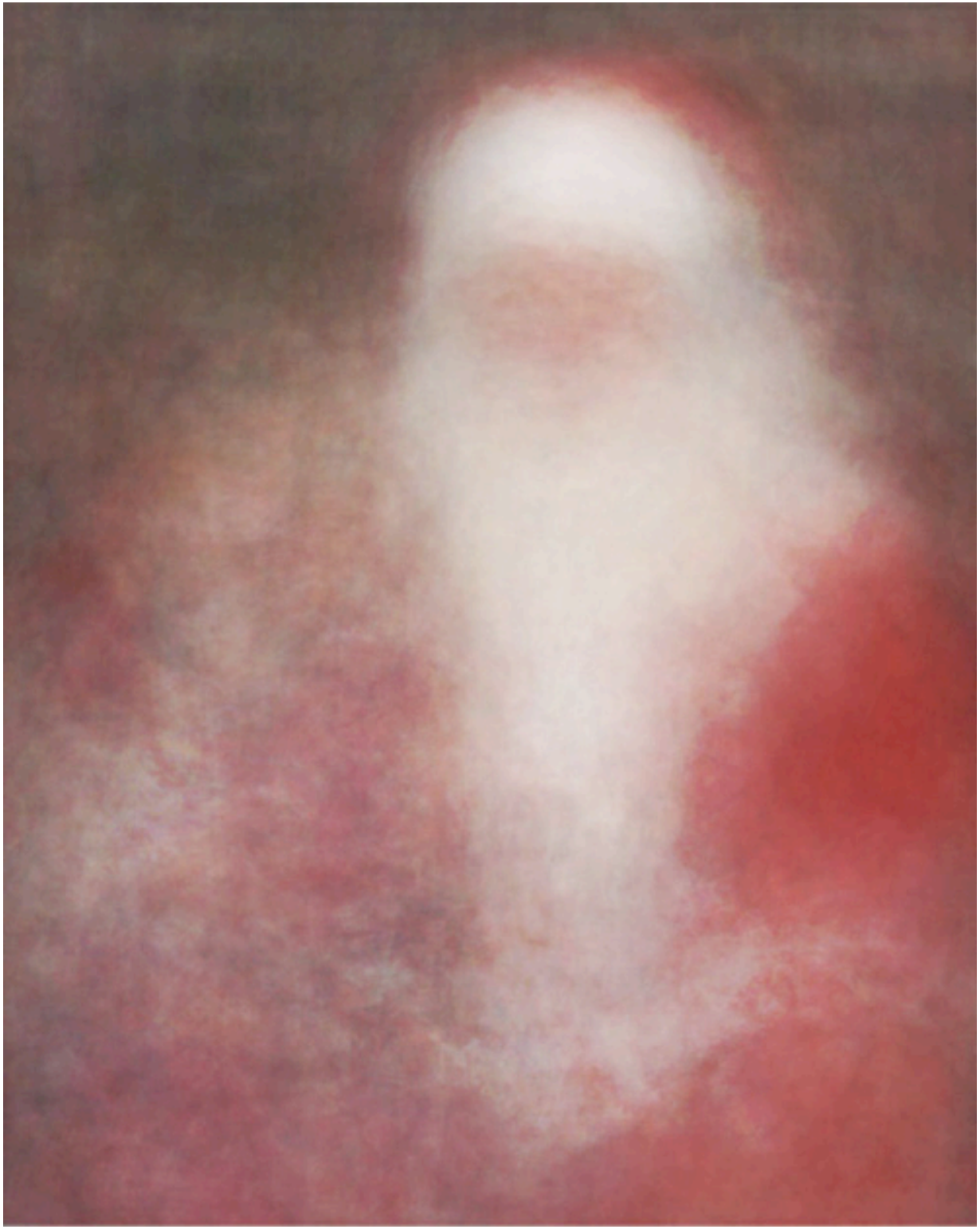
Students learn to extracting color bands from the images (the R, G and B), crop an image and then compute average colors

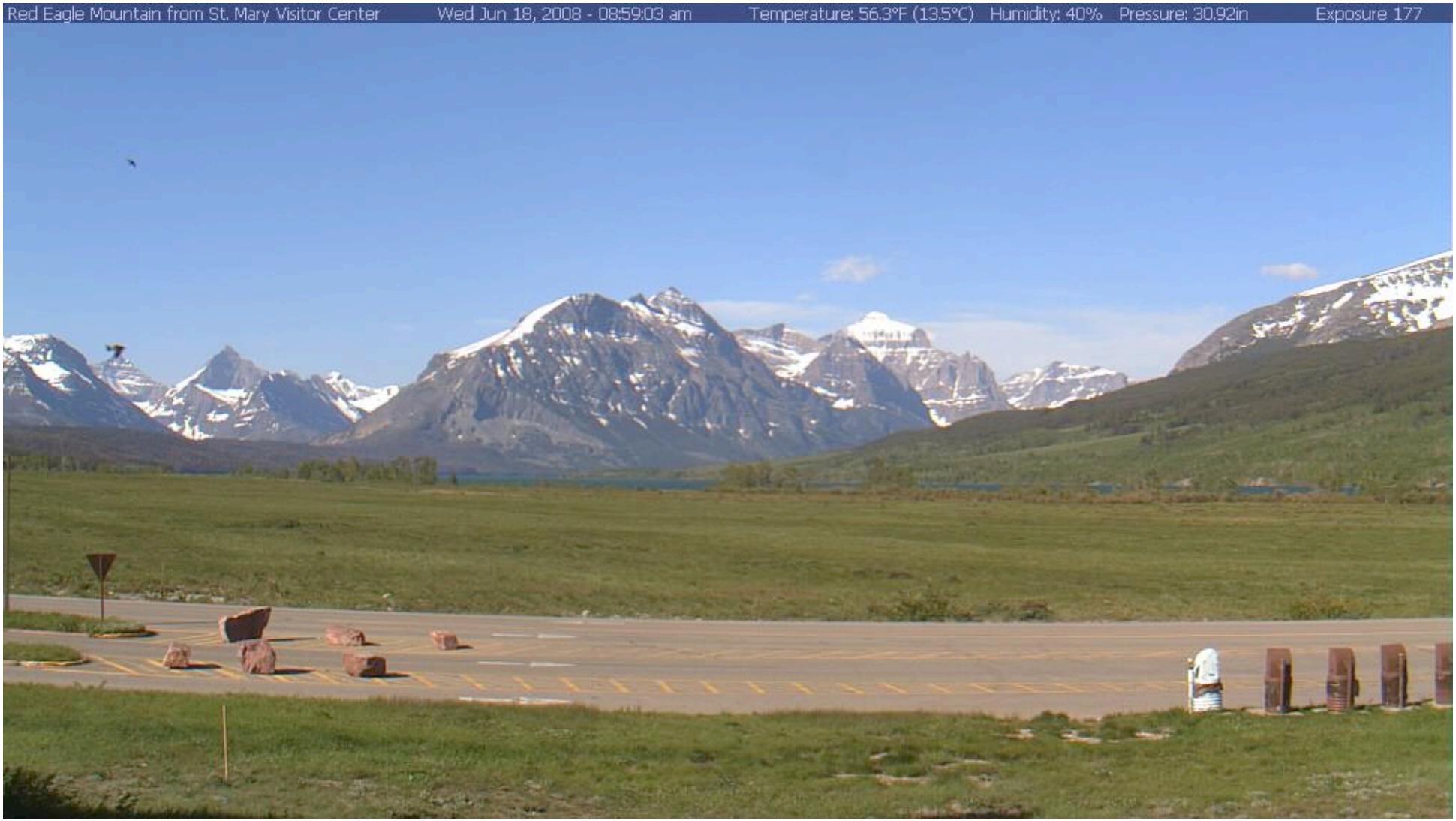
Ultimately, they will consider a large number of images from one of three web cams and plot some function of the amount of green in each image to track the onset of spring at the location

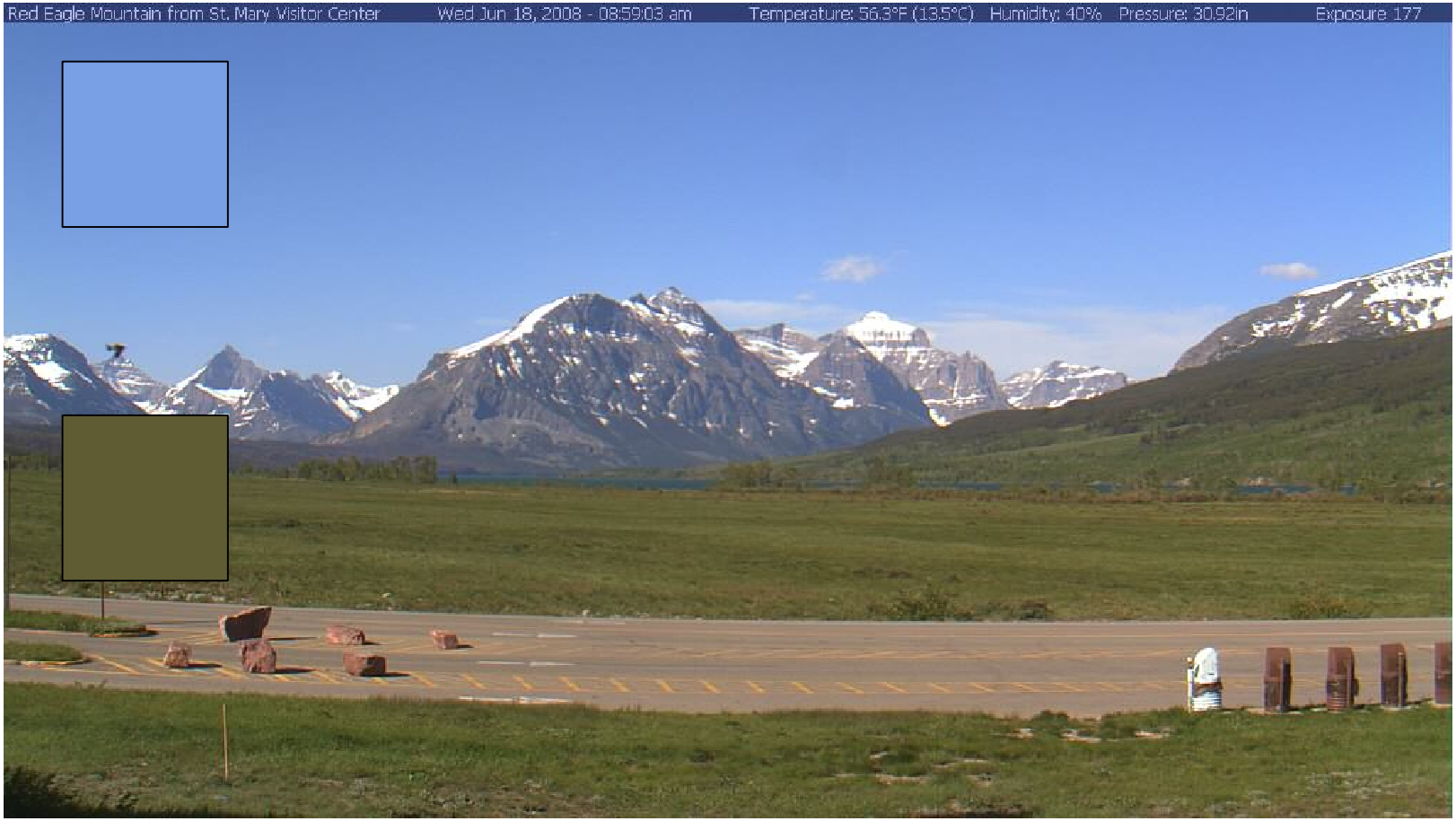
In addition to talking about basic subsetting again, students learn about objects and methods in R and apply their skills with time series data



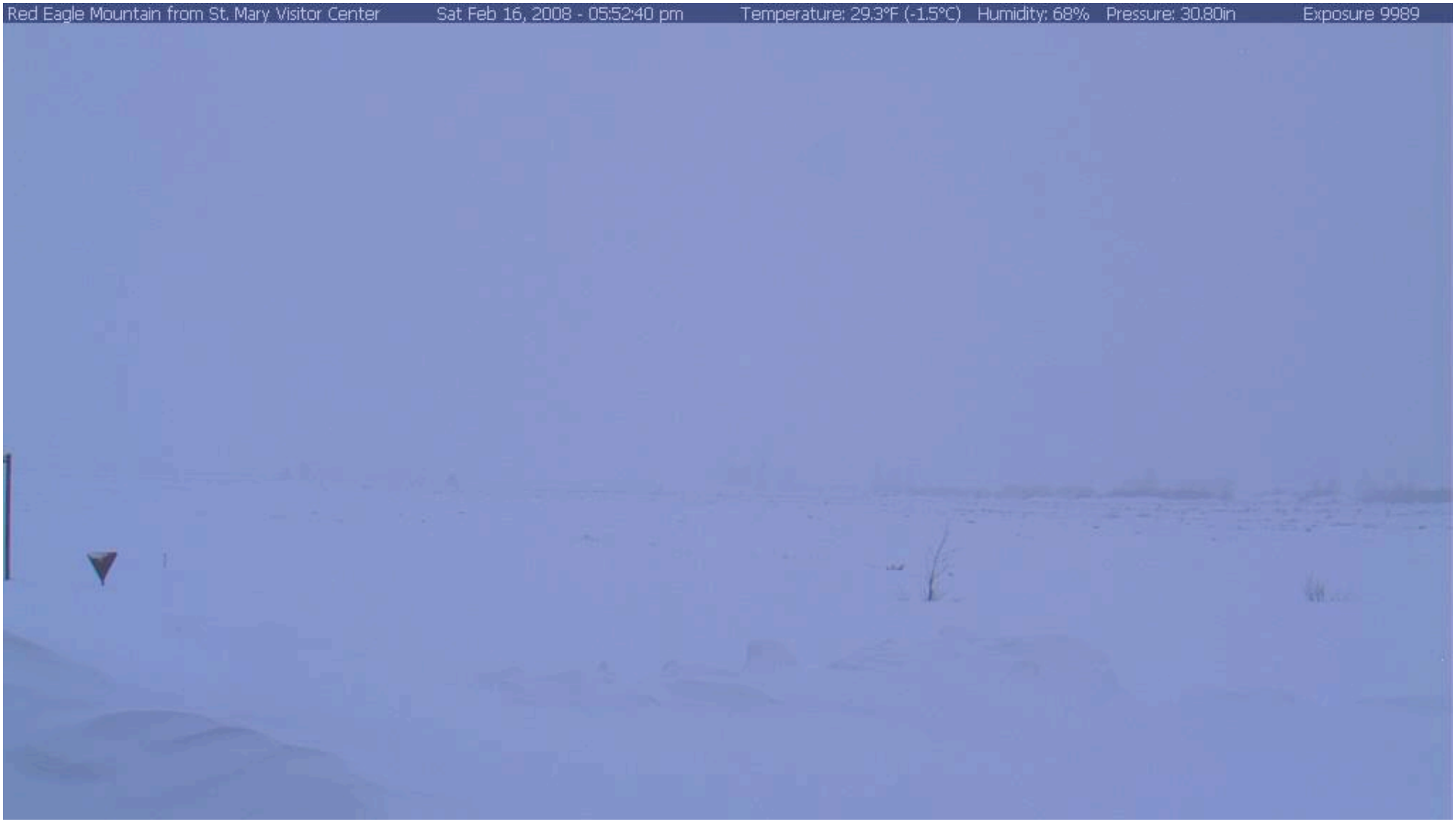




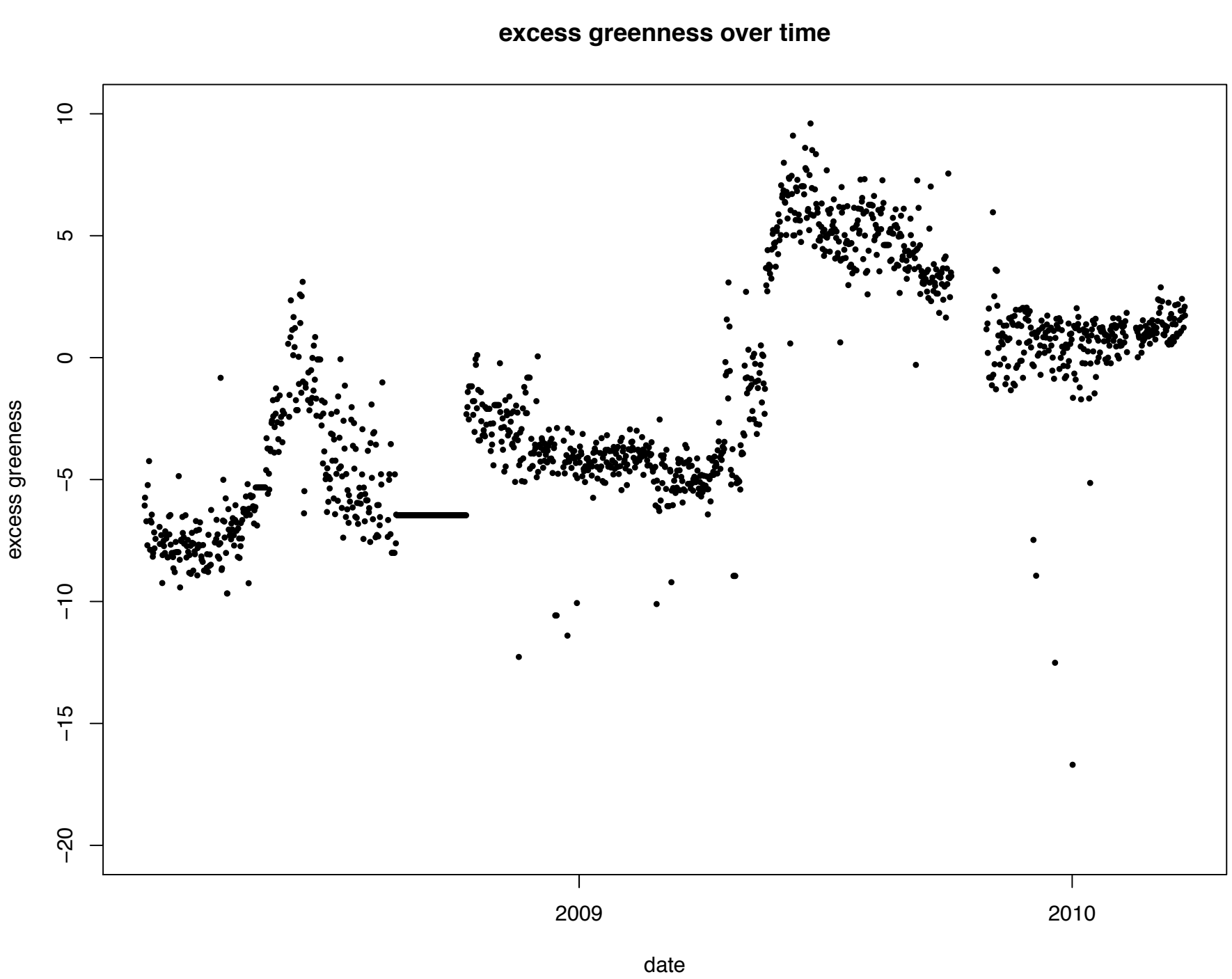

Red Eagle Mountain from St. Mary Visitor Center
Wed Jun 18, 2008 - 08:59:03 am
Temperature: 56.3°F (13.5°C)
Humidity: 40%
Pressure: 30.92in
Exposure 177


Red Eagle Mountain from St. Mary Visitor Center
Wed Jun 18, 2008 - 08:59:03 am
Temperature: 56.3°F (13.5°C)
Humidity: 40%
Pressure: 30.92in
Exposure 177

```
> library(biOps)
> files <- list.files("webcam_824",full.names=TRUE)

> class(files)
[1] "character"

> length(files)
[1] 1436

> head(files)
[1] "webcam_824/824_2008-02-14_08_00_04.jpg"
[2] "webcam_824/824_2008-02-14_14_00_04.jpg"
[3] "webcam_824/824_2008-02-15_14_00_04.jpg"
[4] "webcam_824/824_2008-02-16_08_00_04.jpg"
[5] "webcam_824/824_2008-02-16_14_00_04.jpg"
[6] "webcam_824/824_2008-02-17_08_00_04.jpg"

> cam <- readJpeg(files[6])

> plot(cam)

> mean(2*imgGreenBand(cam)-imgRedBand(cam)-imgBlueBand(cam))
[1] -47.69011
```


Red Eagle Mountain from St. Mary Visitor Center
Sat Feb 16, 2008 - 05:52:40 pm
Temperature: 29.3°F (-1.5°C)
Humidity: 68%
Pressure: 30.80in
Exposure 9989

```
> shortfiles <- list.files("webcam_824")
> format <- "824_%Y-%m-%d_%H_%M_%S.jpg"

> times <- strptime(shortfiles,format)
> range(times)
[1] "2008-02-14 08:00:04 PST" "2010-03-25 14:00:36 PDT"

> excess_green <- rep(0,length(files))

> for(i in 1:length(files)){
+      cam <- readJpeg(files[i])
+      excess_green[i] <- mean(2*imgGreenBand(cam)-imgRedBand(cam)-imgBlueBand(cam))
+      print(i)
+ }
[1] 1
[1] 2
[1] 3
[1] 4
[1] 5
[1] 6
[1] 7
[1] 8
[1] 9

...
[1] 1432
[1] 1433
[1] 1434
[1] 1435
[1] 1436

> plot(times, excess_green)
```


excess greenness over time
excess greeness
10
5
0
−5
−10
−15
−20
2009
2010
date


Nepenthe-Big Sur 11/18/2008 14:59:52

Text (first pass)

This lesson starts with a discussion about how, from a data collection perspective, there can be many ways to solve a problem -- Here we look at the advance of spring again, but through the lens of comments posted to Twitter

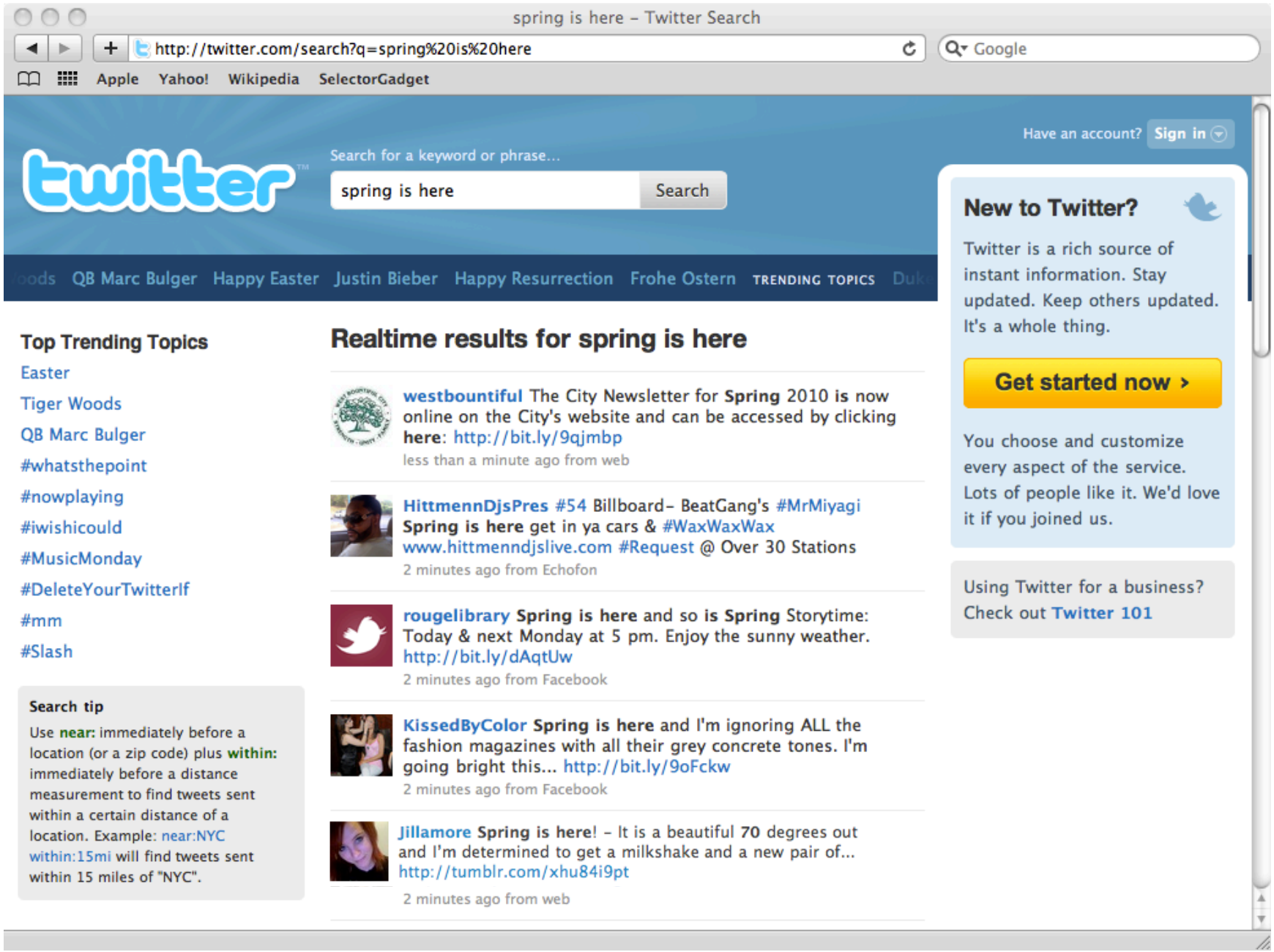
This lets us comment on the secondary uses of data and reconnect to the discussion of “purposeful data collection”

Students will pull data from the Weather Underground and add it to their plots of comment counts from Twitter

Time comes up (again), introducing seconds since an epoch rather than a calendar-based approach

Students make maps of “local” twitter comments and compare them to the regions’s weather


Twitter
http://twitter.com/
Google
Apple
Yahoo!
Wikipedia
SelectorGadget
Have an account? Sign in
twitter
Search for a keyword or phrase...
Search
Discover what's happening right now, anywhere in the world
TRENDING TOPICS
Easter
Justin Bieber
Who's Your Daddy
Tiger Woods
Happy Easter
Happy Resurrecti
New to Twitter?
Twitter is a rich source of instant information. Stay updated. Keep others updated. It's a whole thing.
Get started now ›
You choose and customize every aspect of the service. Lots of people like it. We'd love it if you joined us.
Using Twitter for a business? Check out Twitter 101
See who's here
Friends and industry peers you know. Celebrities you watch. Businesses you frequent. Find them all on Twitter.
Top tweets
View all ›
stl_cardinals Happy Opening Day! Go Cardinals! #stlcards
3 hours ago
Steph83Ga THREE children lost a father. TWO parents lost a son. EIGHT siblings lost a brother. MILLIONS of fans lost a hero. Justice 4 Michael Jackson
3 hours ago
KennyHamilton Don't worry guys, Justin is about to hit the stage. The show is running a bit late. Gotta let our band set up.
3 hours ago
© 2010 Twitter
About Us
Contact
Blog
Status
Goodies
API
Business
Help
Jobs
Terms
Privacy
Language: English


spring is here – Twitter Search
http://twitter.com/search?q=spring%20is%20here
Google
Apple
Yahoo!
Wikipedia
SelectorGadget
Have an account? Sign in
Search for a keyword or phrase...
spring is here
Search
QB Marc Bulger
Happy Easter
Justin Bieber
Happy Resurrection
Frohe Ostern
TRENDING TOPICS
Top Trending Topics
Easter
Tiger Woods
QB Marc Bulger
#whatsthepoint
#nowplaying
#iwishicould
#MusicMonday
#DeleteYourTwitterIf
#mm
#Slash
Search tip
Use near: immediately before a location (or a zip code) plus within: immediately before a distance measurement to find tweets sent within a certain distance of a location. Example: near:NYC within:15mi will find tweets sent within 15 miles of "NYC".
Realtime results for spring is here
westbountiful The City Newsletter for Spring 2010 is now online on the City's website and can be accessed by clicking here: http://bit.ly/9qjmbp
less than a minute ago from web
HittmennDjsPres #54 Billboard- BeatGang's #MrMiyagi Spring is here get in ya cars & #WaxWaxWax www.hittmenndjslive.com #Request @ Over 30 Stations
2 minutes ago from Echofon
rougelibrary Spring is here and so is Spring Storytime: Today & next Monday at 5 pm. Enjoy the sunny weather. http://bit.ly/dAqtUw
2 minutes ago from Facebook
KissedByColor Spring is here and I'm ignoring ALL the fashion magazines with all their grey concrete tones. I'm going bright this... http://bit.ly/9oFckw
2 minutes ago from Facebook
Jillamore Spring is here! - It is a beautiful 70 degrees out and I'm determined to get a milkshake and a new pair of... http://tumblr.com/xhu84i9pt
2 minutes ago from web
New to Twitter?
Twitter is a rich source of instant information. Stay updated. Keep others updated. It's a whole thing.
Get started now ›
You choose and customize every aspect of the service. Lots of people like it. We'd love it if you joined us.
Using Twitter for a business? Check out Twitter 101

Jill Bull (Jillamore) on Twitter

http://twitter.com/Jillamore

Google

Apple Yahoo! Wikipedia SelectorGadget

twitter

Have an account? Sign in

Hey there! Jill Bull is using Twitter.

Twitter is a rich source of instantly updated information. Join today and **follow @Jillamore** to start receiving these tweets.

Get started now ›

Already using Twitter from a phone? Complete signup here.

Jillamore

Name Jill Bull
Location ÜT: 40.360171,-74.079609
Bio Love Friends, Love Music, Love Family, Love Life.

13 following | 18 followers | 0 listed

Tweets 172

Favorites

Following

RSS feed of Jillamore's tweets

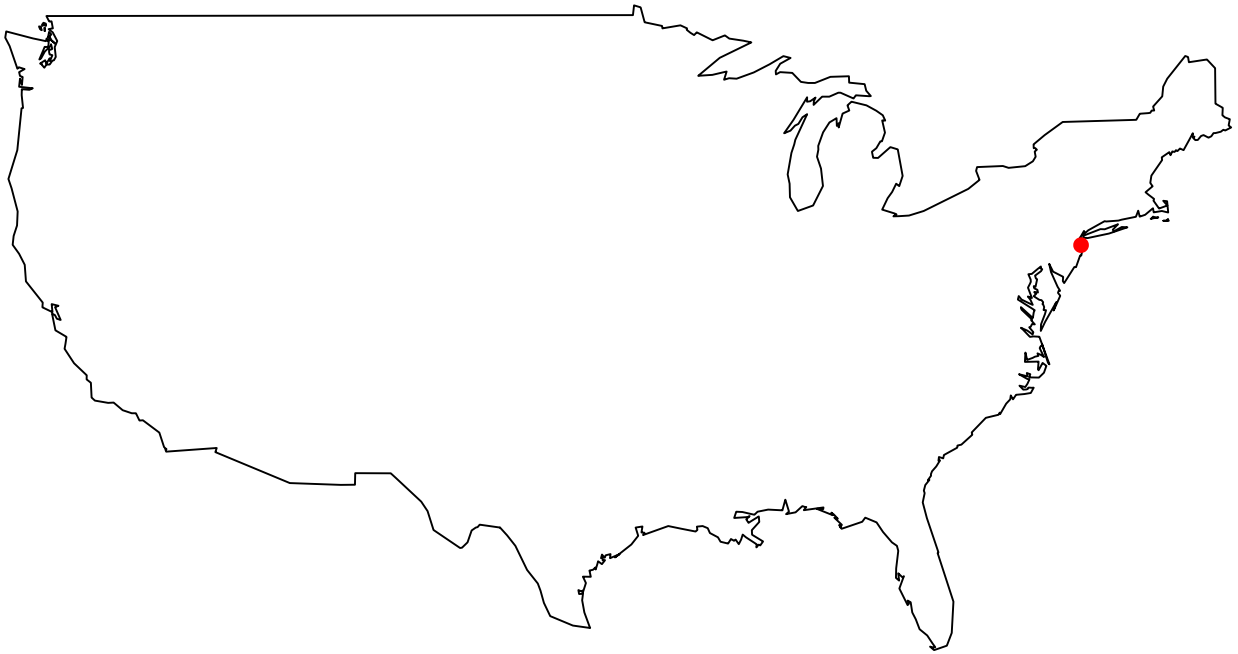
Spring is here! – It is a beautiful 70 degrees out and I'm determined to get a milkshake and a new pair of... http://tumblr.com/xhu84i9pt
7:09 AM Apr 2nd via Tumblr

I love it when its 70 degrees out! Especially since nj has had like 8 blizzards and like 20 days of rain in a row :p
6:57 AM Apr 2nd via UberTwitter

Sleeeping :)
9:23 PM Apr 1st via UberTwitter

So just chillin at my sisters tattoo shop wanting more tattoos
5:00 PM Apr 1st via UberTwitter

@antoboros howve you been? And when are you coming back to nj or nyc?
3:43 PM Apr 1st via UberTwitter



Welcome to Weather Underground : Weather Underground

http://www.wunderground.com/

Google

Apple Yahoo! Wikipedia SelectorGadget

Welcome to Weather Underground! Sign In or Create an Account. Edit my Page Preferences.

Full Screen NEW! - Mobile - iPhone - Lite - Download

eather Underground
underground.com

Search: Zip or City, State, Airport Code, Country | Weather Conditions | Go

Local Weather	Maps & Radar	Severe Weather	Photos & Video	Blogs	Travel & Activities	Resources
My Locations	Radar	Tropical & Hurricane	Photo Galleries	Dr. Jeff Masters	Ski & Snow	Severe Weather
History Data	Satellite	Storm Reports	World View	Meteorology Blogs	Sports	Climate Change
Weather Stations	WunderMap™	U.S. Severe Alerts	Webcams	Member Blogs	Road Trip Planner	About Maps & Radar

ocent Cities
ed Bank, NJ
ttsburgh, PA
ew York, NY
acier National
rk, MT
g Bear Lake, CA
Edit My Favorites

underPhotos
Congratulations to rigirl for winning the nithsonian Magazine photo contest!

Browse All Photos

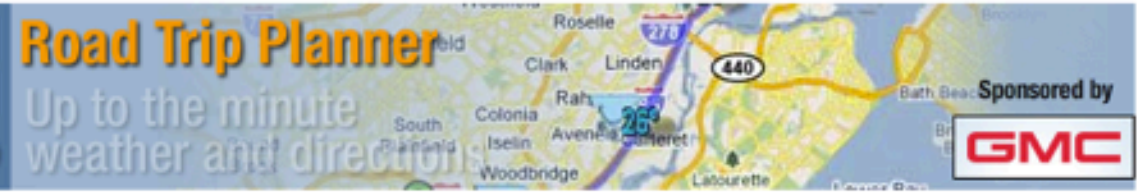
underMap

View WunderMap

ebsite Spotlight
eather Maps

start enter city or zip
destination enter city or zip
get directions and weather

Road Trip Planner
Up to the minute weather and directions



Sponsored by GMC

US

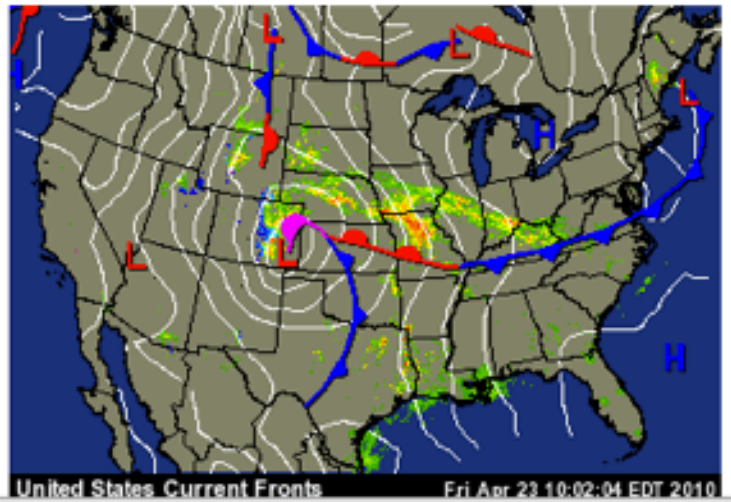
Select a map below or scroll down for more options.

Maps

Current | Forecast | Aviation | Models | WunderMap

US Fronts

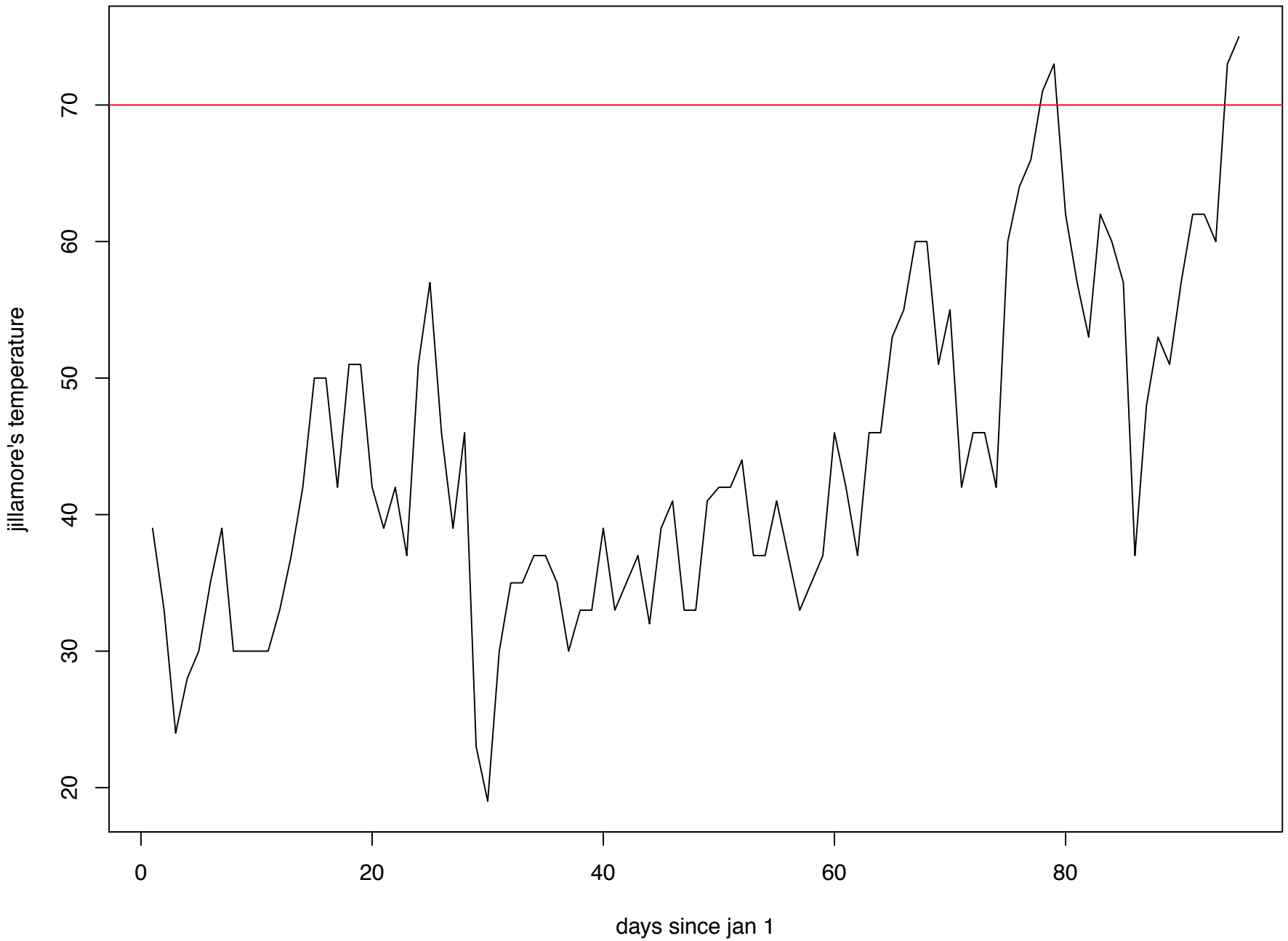
- Temperature
- Heat Index
- Windchill
- Humidity
- Radar
- Dew Point
- Wind
- Visibility
- Visible Satellite
- Satellite
- **Fronts**
- Snow Depth
- Precipitation
- Jet Stream



Road Trip Planner:

Going on a trip or to an outdoor event? Our new and improved Trip Planner helps you plan ahead

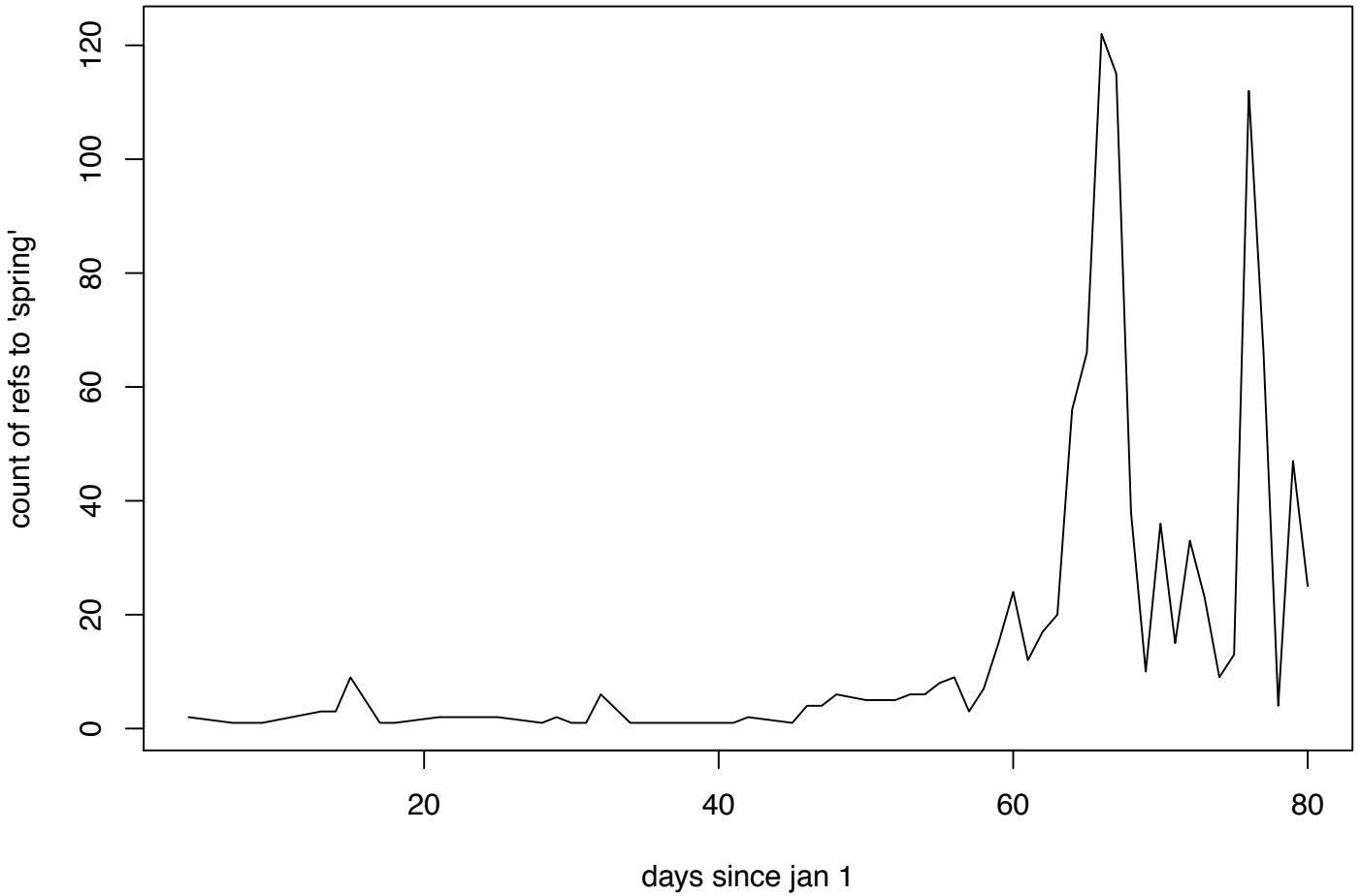
```
jillamore_weather <- read.csv(url(
   paste("http://www.wunderground.com/history/airport/KBLM/2010/1/1/CustomHistory.html?”,
   dayend=5&monthend=4&yearend=2010&req_city=NA&req_state=NA&req_statename=NA&format=1"),
   sep=””),head=T,comment.char="<")

>  plot(jillamore_weather$Max.Temperature,type="l")

> abline(h=70,col="red")
> sum(jillamore_weather$Max.Temperature>71)
[1] 3
```


jillamore's temperature
20
30
40
50
60
70
0
20
40
60
80
days since jan 1

```
<entry>
    <id>tag:search.twitter.com,2005:11484800853</id>
    <published>2010-04-02T15:09:26Z</published>
    <link href="http://twitter.com/Jillamore/statuses/11484800853" ... />
    <title>Spring is here! - It is a beautiful 70 degrees out and ... </title>
    <content type="html">Spring &lt;b&gt;is&lt;/b&gt; &lt;b&gt;here&lt;/b&gt;! ...
    <updated>2010-04-02T15:09:26Z</updated>
    <link href="http://a1.twimg.com/profile_images/366032002/Photo_29_normal.jpg" .
    <twitter:geo>
    </twitter:geo>
    <twitter:metadata>
      <twitter:result_type>recent</twitter:result_type>
    </twitter:metadata>
    <twitter:source>&lt;a href=&quot;http://www.tumblr.com/&quot; ... </twitter:sou
    <twitter:lang>en</twitter:lang>
    <author>
      <name>Jillamore (Jill xxxx)</name>
      <uri>http://twitter.com/Jillamore</uri>
    </author>
</entry>
```

```
> load(url("http://mobilize.stat.ucla.edu/day6/twitter.Rda"))
>
> class(seasons)
[1] "data.frame"

> dim(seasons)
[1] 429819      6

> head(seasons[,1:5])
     created       username  longitude latitude    search_term
1 1238837766    missouriBNN  -91.83183 37.96425 Spring is here
2 1238837550    Paul Joseph  -84.97169 34.76862 Spring is here
3 1238836314 Gonzalo Garcia  -73.98695 40.75605 Spring is here
4 1238836062          Aixa          NA       NA Spring is here
5 1238834899         Terry  -122.44256 47.24454 Spring is here
6 1238833105  Blair Johnson  -79.39098 43.64058 Spring is here

> head(seasons[,6])
[1] "Fat Jack's Erratic Rants: WARM WEATHER AND GAS SAVINGS: BUY A MOTORCYCLE: Spring is here. or at least on i
[2] "&quot;Spring is here. And I am a flower. with nothing interesting to say.&quot; Who said it? :-)"
[3] "Ants are starting to invade my bathroom. Guess that means spring is here. Grrrrr"
[4] "Im so bored spring is here.and so far it's going bad! :/"
[5] "Guess Spring is here and must figure out what type of veg. to plant and more flowers."
[6] "@Royboi pretty good. finally feeling like spring is here. feeling all randy and ready for patio drinkin!"
```

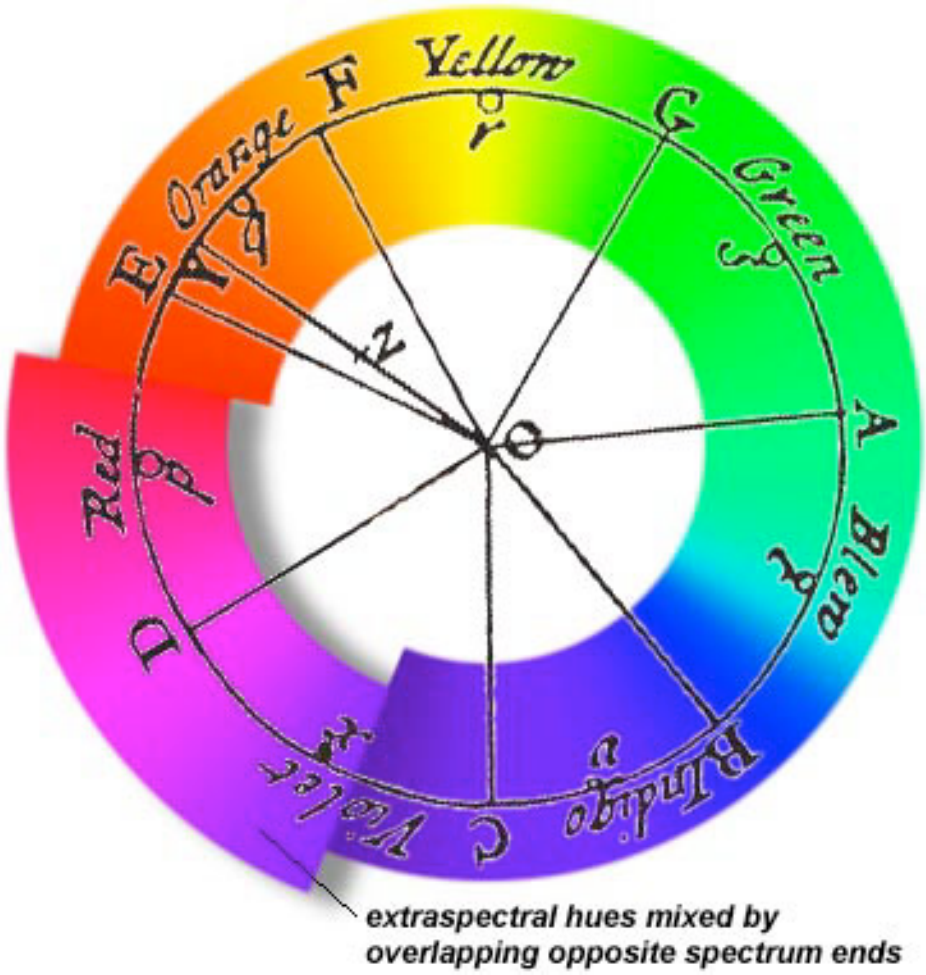
```
> table(seasons$search_term)

                  Blooming                  Bud burst
                     73472                        135
                  budbreak                  Budburst
                       200                        219
             buds breaking              Buds bursting
                        37                         84
          Cadono le foglie               Fall is here
                        33                      14590
                  Fiorendo      Fiori della primavera
                        10                          4
               Fiori nuovi                   Fiorire
                         6                        154
                 floreando                  florecer
                       237                       2076
                 Flowering               Foglie nuove
                     21826                         12
              germogliando               Germogliare
                         1                         32
                 Germoglio       Gli alberi diventano
                      1233                          3
                 it's fall                 It's spring
                     12947                      26594
               it's summer                 it's winter
                     50295                      26881
                  its fall                  its spring
                     10320                      18905
                its summer                  its winter
                     41868                      22263
           La primavera ha          Le foglie cambiano
                       221                          1
       Le foglie diventano               Le foglie sono
                         2                          9
          le foglie stanno  Le foglie stanno cambiando
                         3                          5
Le foglie stanno diventando                   Leaf out
                        51                       6175
               Leafing out                 Leaves are
                       211                      27741
         leaves are turning             Leaves changing
                       262                       2664
             Leaves falling                 New flower
                      2966                       4230
                New flowers                 New leaves
                      3153                       2162
            Newly flowering                     retoar
                         4                          2
                 Sbocciando                  Sbocciare
                         9                         69
                 Spring has           Spring has sprung
                     20778                      11450
              Spring is here               Spring is now
                     22873                        341
```

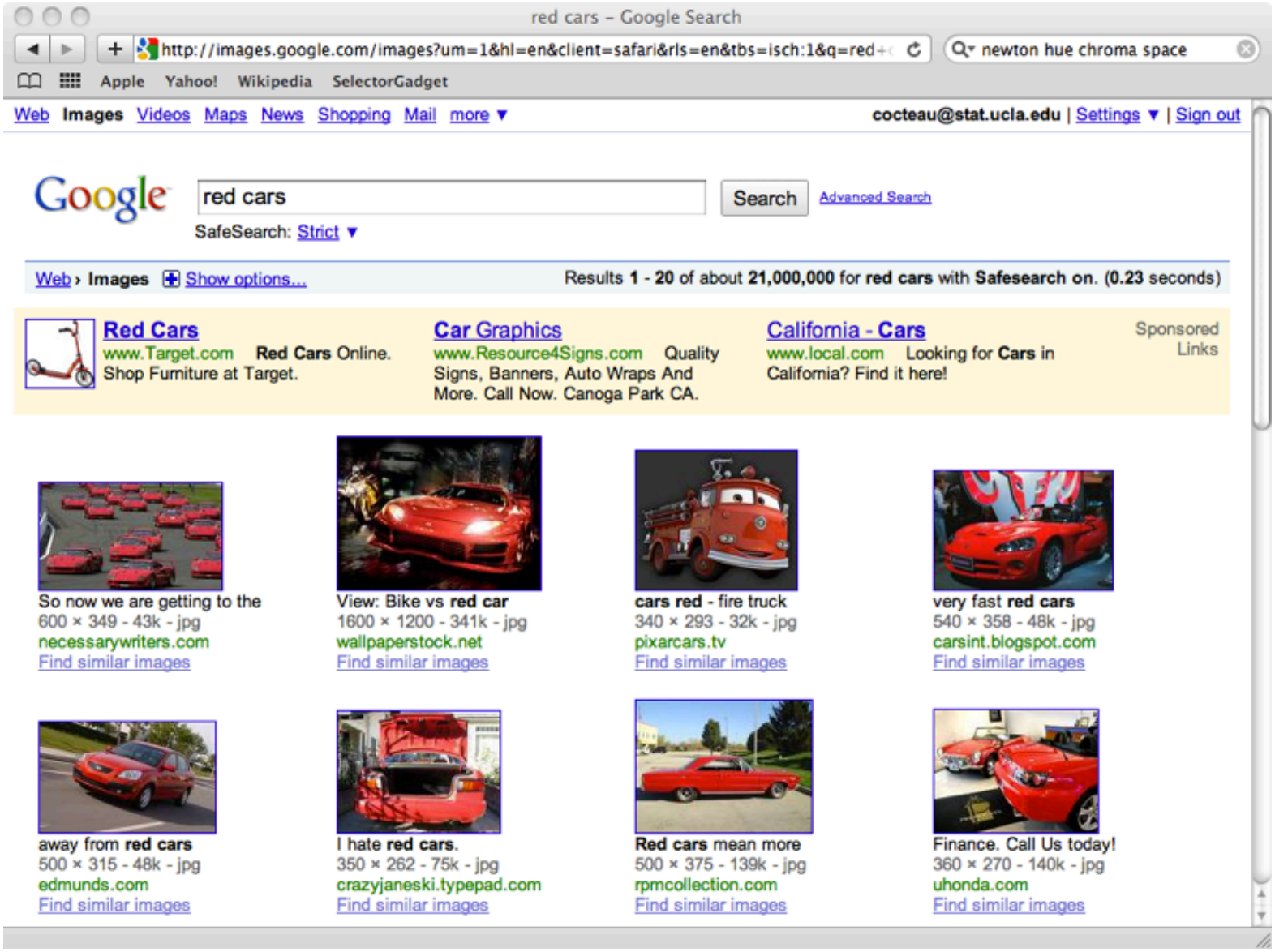

count of refs to 'spring'
0
20
40
60
80
100
120
20
40
60
80
days since jan 1

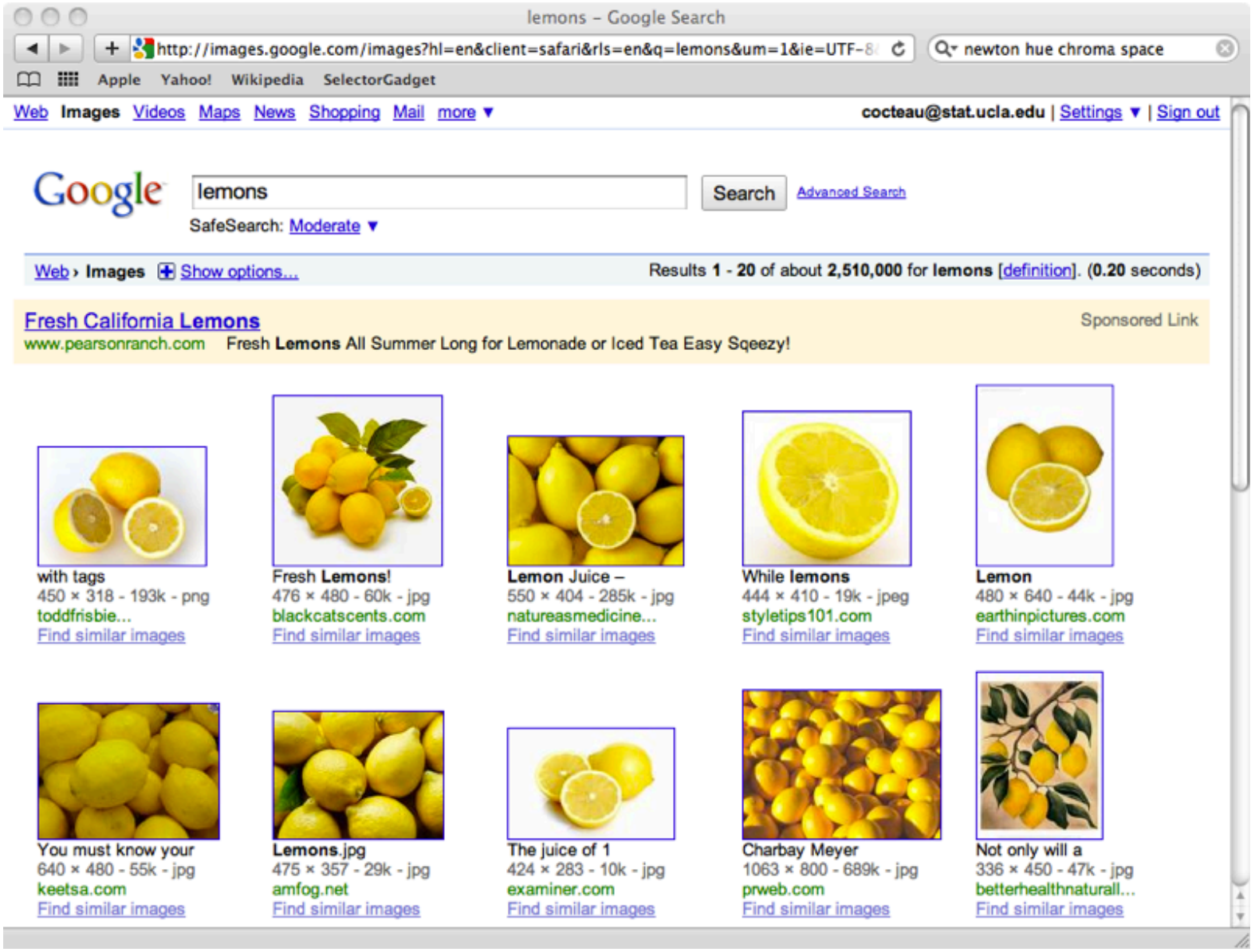
The hue circle

Here we try to motivate the use of a new “space” or plane on which to plot data -- here the hue-chroma space

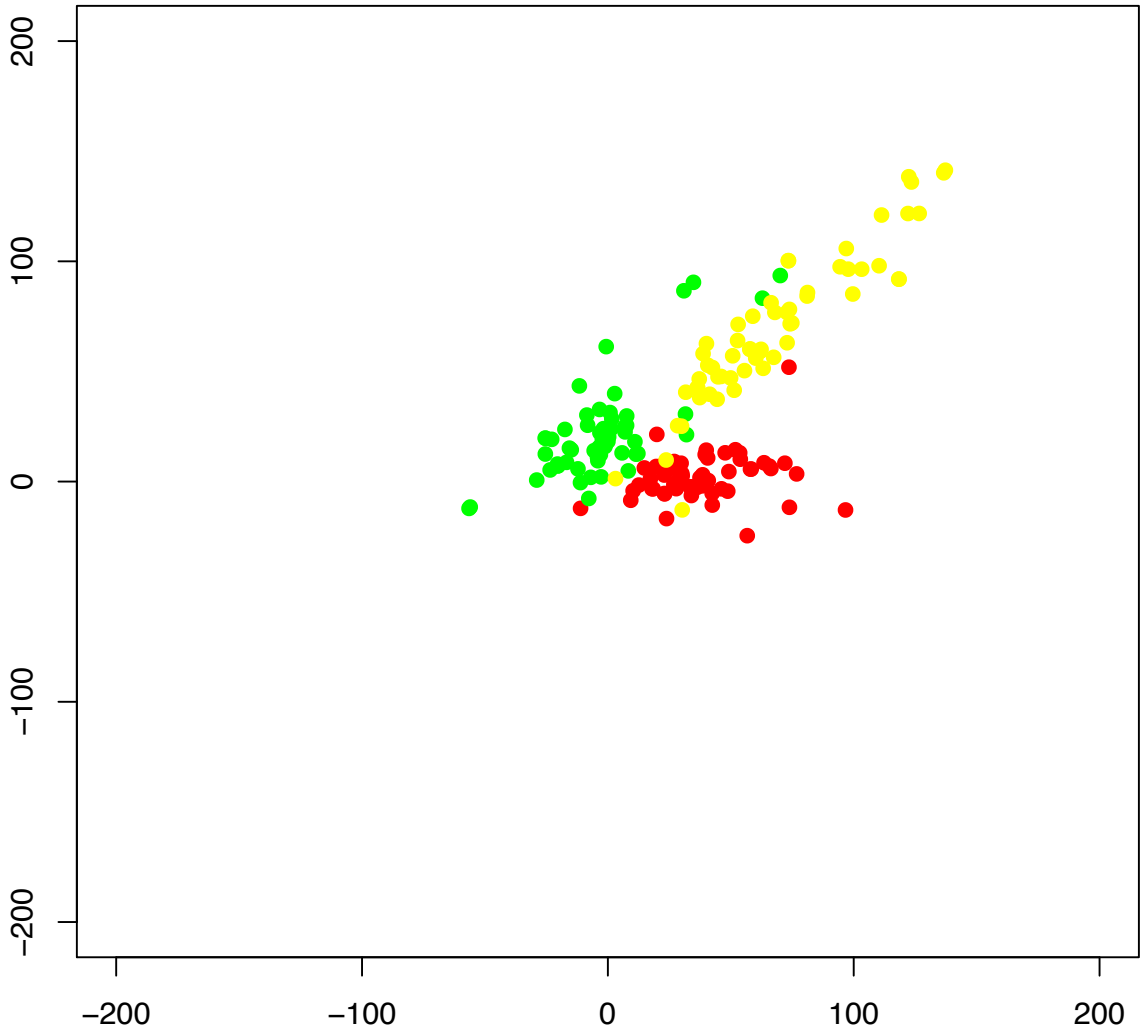
In addition, we kick the tires on Google Images API, pulling images of one color and seeing how they cluster in this hue chroma space




red cars - Google Search
http://images.google.com/images?um=1&hl=en&client=safari&rls=en&tbs=isch:1&q=red
newton hue chroma space
Apple Yahoo! Wikipedia SelectorGadget
Web Images Videos Maps News Shopping Mail more ▼
cocteau@stat.ucla.edu | Settings ▼ | Sign out
Google
red cars
Search
Advanced Search
SafeSearch: Strict ▼
Web › Images Show options...
Results 1 - 20 of about 21,000,000 for red cars with Safesearch on. (0.23 seconds)
Red Cars
www.Target.com Red Cars Online. Shop Furniture at Target.
Car Graphics
www.Resource4Signs.com Quality Signs, Banners, Auto Wraps And More. Call Now. Canoga Park CA.
California - Cars
www.local.com Looking for Cars in California? Find it here!
Sponsored Links
So now we are getting to the
600 × 349 - 43k - jpg
necessarywriters.com
Find similar images
View: Bike vs red car
1600 × 1200 - 341k - jpg
wallpaperstock.net
Find similar images
cars red - fire truck
340 × 293 - 32k - jpg
pixarcars.tv
Find similar images
very fast red cars
540 × 358 - 48k - jpg
carsint.blogspot.com
Find similar images
away from red cars
500 × 315 - 48k - jpg
edmunds.com
Find similar images
I hate red cars.
350 × 262 - 75k - jpg
crazyjaneski.typepad.com
Find similar images
Red cars mean more
500 × 375 - 139k - jpg
rpmcollection.com
Find similar images
Finance. Call Us today!
360 × 270 - 140k - jpg
uhonda.com
Find similar images


lemons – Google Search
http://images.google.com/images?hl=en&client=safari&rls=en&q=lemons&um=1&ie=UTF-8
newton hue chroma space
Apple Yahoo! Wikipedia SelectorGadget
Web Images Videos Maps News Shopping Mail more ▼
cocteau@stat.ucla.edu | Settings ▼ | Sign out
Google
lemons
Search
Advanced Search
SafeSearch: Moderate ▼
Web › Images Show options...
Results 1 - 20 of about 2,510,000 for lemons [definition]. (0.20 seconds)
Fresh California Lemons
Sponsored Link
www.pearsonranch.com Fresh Lemons All Summer Long for Lemonade or Iced Tea Easy Sqeezy!
with tags
450 × 318 - 193k - png
toddfrisbie...
Find similar images
Fresh Lemons!
476 × 480 - 60k - jpg
blackcatscents.com
Find similar images
Lemon Juice –
550 × 404 - 285k - jpg
natureasmedicine...
Find similar images
While lemons
444 × 410 - 19k - jpeg
styletips101.com
Find similar images
Lemon
480 × 640 - 44k - jpg
earthinpictures.com
Find similar images
You must know your
640 × 480 - 55k - jpg
keetsa.com
Find similar images
Lemons.jpg
475 × 357 - 29k - jpg
amfog.net
Find similar images
The juice of 1
424 × 283 - 10k - jpg
examiner.com
Find similar images
Charbay Meyer
1063 × 800 - 689k - jpg
prweb.com
Find similar images
Not only will a
336 × 450 - 47k - jpg
betterhealthnaturall...
Find similar images

Hue Circle: Red Camaros, Lemons and Central Park in Summertime



Geometry lesson

We are now quite late in the unit and we extend the idea from the hue circle lesson that a novel transformation could let us see clusters in a data set with more than 3 variables

We introduce multidimensional scaling by looking at distances between data points -- From a distance matrix we move to a plot that, again, let’s us see clustering

Free Map Tools

Maps you can make use of...

Navigate:

Popular Map Tools
- ZIP Codes Inside a Radius
- How Far Can I Travel
- How Far is it Between
- Radius From UK Postcode
- UK Postcode Map
- Measure Distance
- Area Calculator
- Radius Around Point
- Distance Between UK Postcodes
- UK Postcodes Inside Radius

Map Resources
- Download UK Postcodes

About
News
Contact
FAQ's

How Far is it Between

This tool can be used to find the distance between two named points on a map. You can decide which two points to measure and then find out the distance between them as the crow flies and distance when driving. Type in the names of the places below and click the Show button.

Options

City, Country City, Country

From boston, Massachussets to Seattle, Wa Usa Show

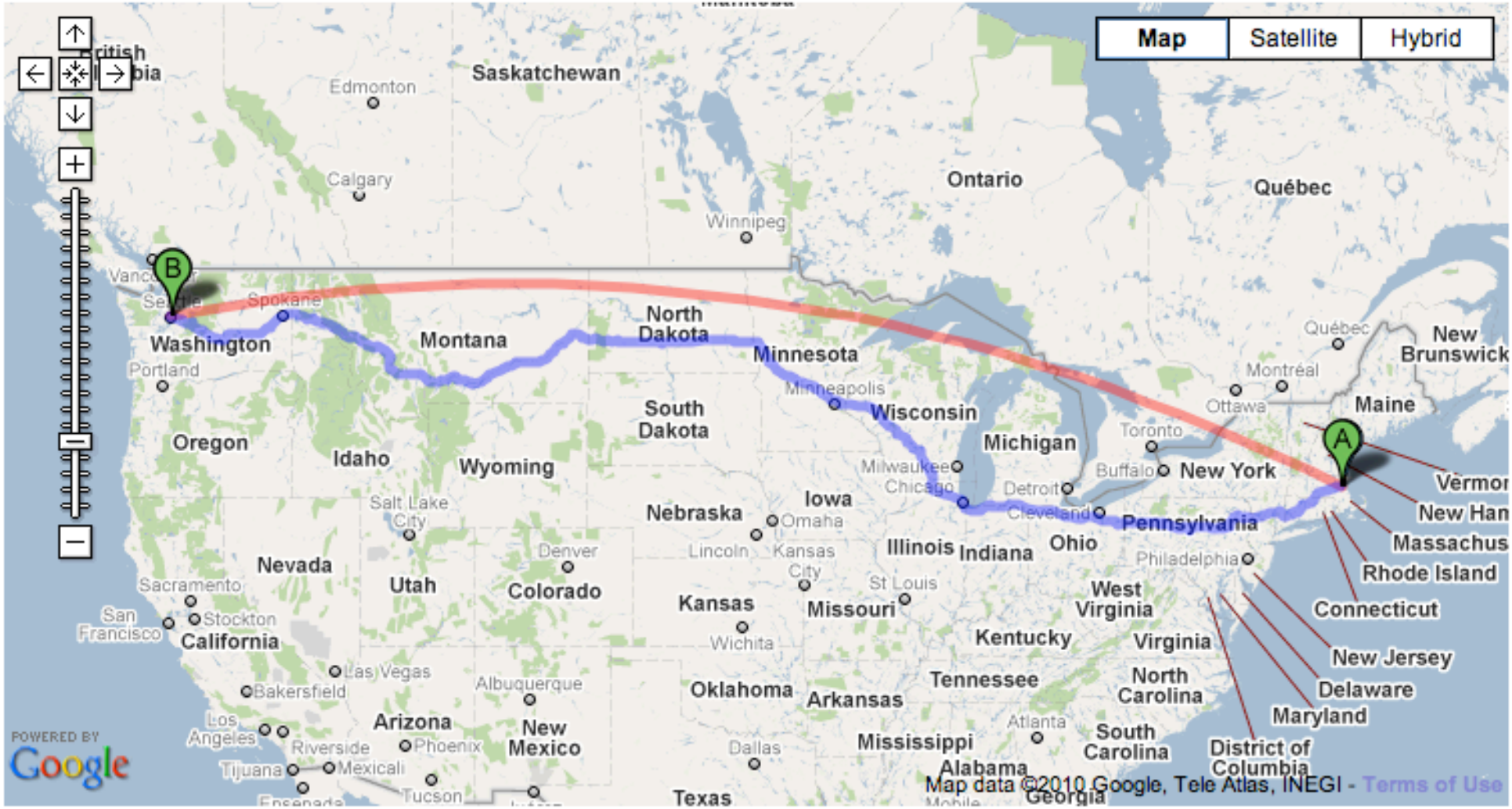
Measure in : (•) miles ( ) km

Distance as the Crow Flies : 2486.106

Distance by Land Transport : 3053.084

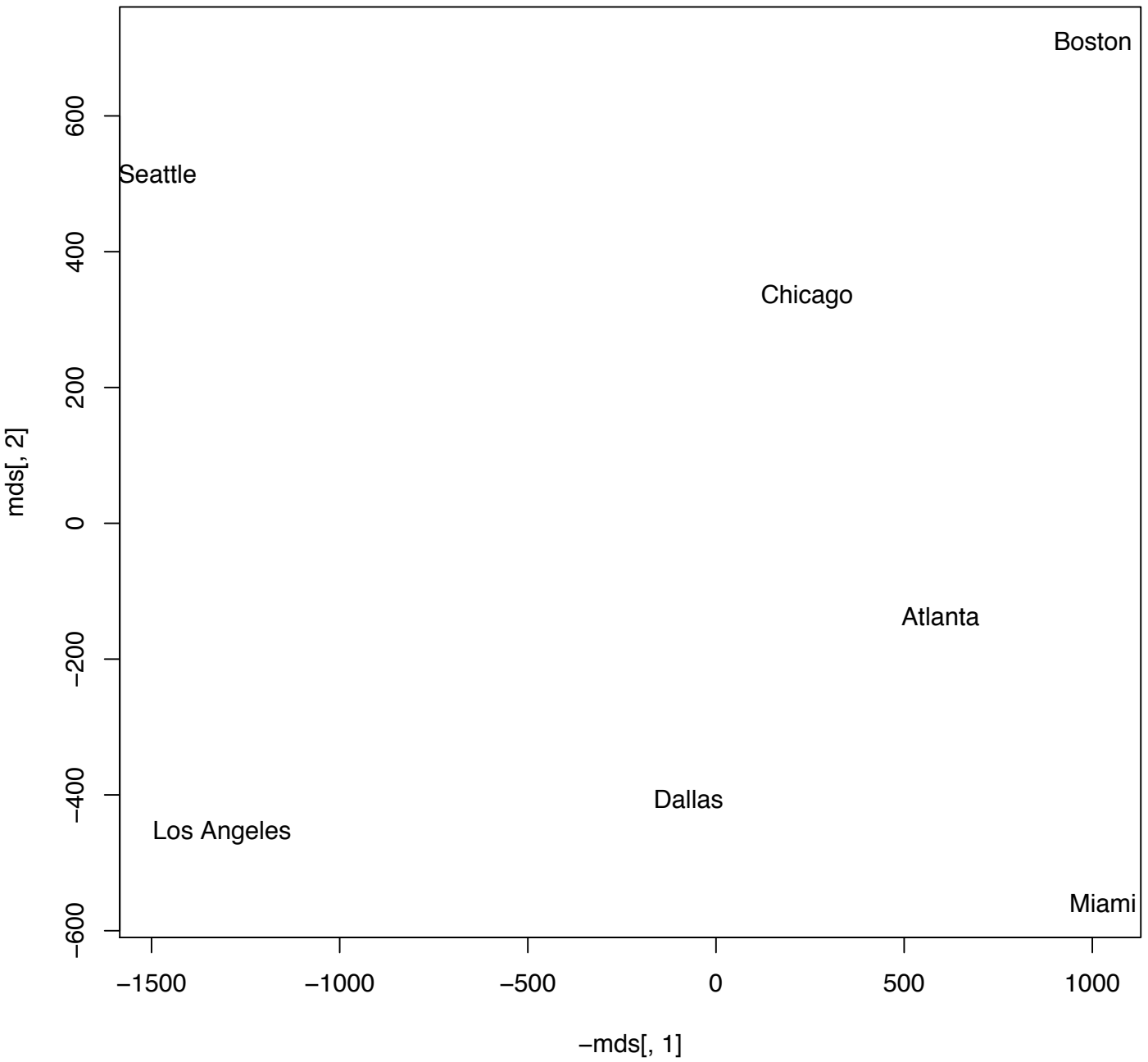
You can link to this result : **How Far is it Between Boston, Massachussets and Seattle, Wa Usa**
http://www.freemaptools.com/how-far-is-it-between-boston_-massachussets-and-seattle_-wa-usa.htm

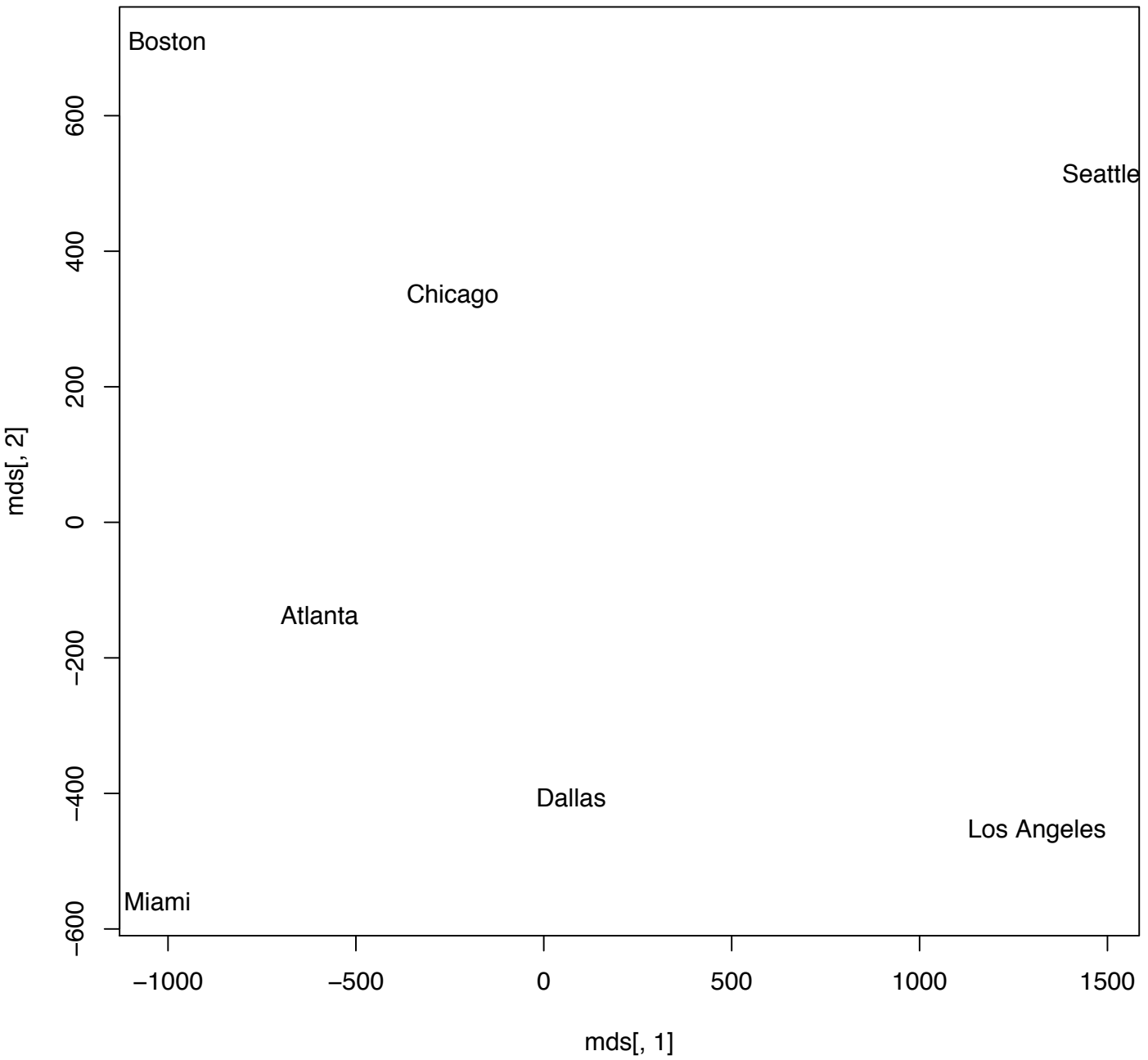
Map Showing the Distance Between boston, Massachussets and Seattle, Wa Usa


Map
Satellite
Hybrid
Saskatchewan
Edmonton
Calgary
Ontario
Québec
Winnipeg
B
A
Spokane
Washington
Montana
North Dakota
Minnesota
Portland
Minneapolis
Wisconsin
Michigan
Toronto
Ottawa
Montréal
Québec
New Brunswick
Maine
Oregon
Idaho
Wyoming
South Dakota
Milwaukee
Chicago
Detroit
Buffalo
New York
Cleveland
Pennsylvania
Salt Lake City
Nebraska
Omaha
Iowa
Lincoln
Kansas City
Illinois
Indiana
Ohio
Philadelphia
Nevada
Utah
Colorado
Denver
St Louis
Missouri
West Virginia
Sacramento
San Francisco
Stockton
California
Kansas
Wichita
Kentucky
Virginia
Connecticut
New Jersey
Delaware
Maryland
Las Vegas
Bakersfield
Albuquerque
Oklahoma
Arkansas
Tennessee
North Carolina
Los Angeles
Riverside
Phoenix
Arizona
New Mexico
Atlanta
South Carolina
Mississippi
Dallas
Alabama
District of Columbia
Tijuana
Mexicali
Tucson
Texas
Georgia
POWERED BY Google
Map data ©2010 Google, Tele Atlas, INEGI - Terms of Use

Seattle to Seattle	0.000
Seattle to Boston	2486.106
Seattle to Los Angeles	961.261
Seattle to Atlanta	2179.149
Seattle to Chicago	1733.626
Seattle to Miami	2732.169
Seattle to Dallas	1680.809
Boston to Seattle	2486.106
Boston to Boston	0.000
Boston to Los Angeles	2591.118
Boston to Atlanta	935.785
Boston to Chicago	848.793
Boston to Miami	1257.655
Boston to Dallas	1549.078
Los Angeles to Seattle	961.261
Los Angeles to Boston	2591.118
Los Angeles to Los Angeles	0.000
Los Angeles to Atlanta	1932.464
Los Angeles to Chicago	1742.325
Los Angeles to Miami	2335.113
Los Angeles to Dallas	1237.771
Atlanta to Seattle	2179.149
Atlanta to Boston	935.785
Atlanta to Los Angeles	1932.464
Atlanta to Atlanta	0.000
Atlanta to Chicago	588.811
Atlanta to Miami	606.023
Atlanta to Dallas	719.548

Chicago to Seattle	1733.626
Chicago to Boston	848.793
Chicago to Los Angeles	1742.325
Chicago to Atlanta	588.811
Chicago to Chicago	0.000
Chicago to Miami	1191.169
Chicago to Dallas	804.676
Miami to Seattle	2732.169
Miami to Boston	1257.655
Miami to Los Angeles	2335.113
Miami to Atlanta	606.023
Miami to Chicago	1191.169
Miami to Miami	0.000
Miami to Dallas	1109.769
Dallas to Seattle	1680.809
Dallas to Boston	1549.078
Dallas to Los Angeles	1237.771
Dallas to Atlanta	719.548
Dallas to Chicago	804.676
Dallas to Miami	1109.769
Dallas to Dallas	0.000


Boston
Seattle
Chicago
Atlanta
Dallas
Los Angeles
Miami
mds[, 2]
600
400
200
0
−200
−400
−600
−1500
−1000
−500
0
500
1000
−mds[, 1]


Boston
Seattle
Chicago
Atlanta
Dallas
Los Angeles
Miami
600
400
200
0
−200
−400
−600
mds[, 2]
−1000
−500
0
500
1000
1500
mds[, 1]

Text

Finally, we close with text -- We make liberal use of R's text mining capabilities to show the students how to compute with text

We transform a collection of president's biographies, identifying words, removing punctuation, "stemming" and deleting common words like "a" and "the"

Students see how word frequency can be used to differentiate classes of documents, and multidimensional scaling is used again to make a plot of the different biographies

We close with an example of text classification -- Spam v. Ham


WORDCOUNT
PREVIOUS WORD
NEXT WORD
r article achieve prevent waste inclu
1513 1514 1515 1516 1517 1518
CURRENT WORD
FIND WORD:
BY RANK:
REQUESTED WORD: R
RANK: 1513
86800 WORDS IN ARCHIVE
ABOUT WORDCOUNT
WordCount™ ©2003 Jonathan Harris | Number27 | Help

TextArc.org Home

http://textarc.org/

Apple Yahoo! Wikipedia SelectorGadget

TextArc

An alternate way to view a text

A TextArc is a visual represention of a text—the entire text (twice!) on a single page. A funny combination of an index, concordance, and summary; it uses the viewer's eye to help uncover meaning. Here are more detailed overviews of the interactive work and the prints.

Hamlet *(Try it!)*

Alice in Wonderland

Thousands of Texts

Still Screen Images

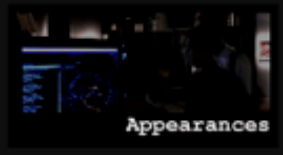
Reactions

Print Editions

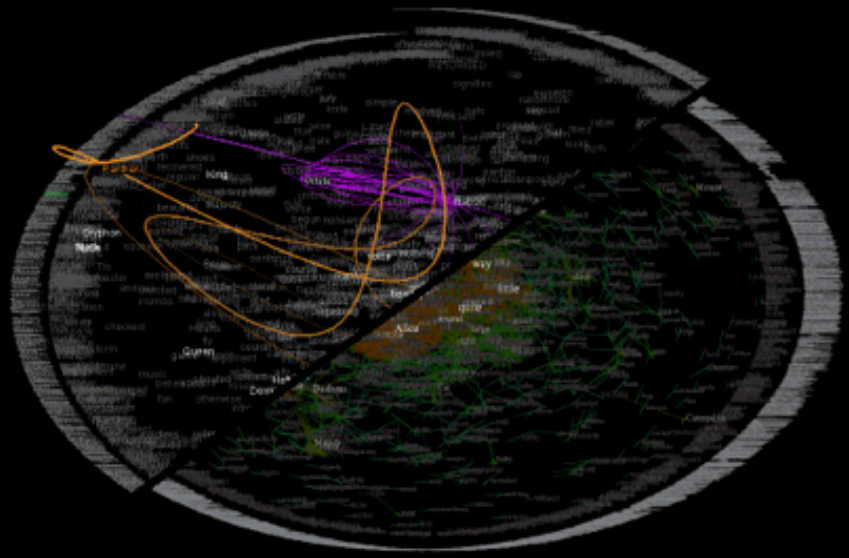
Applications

Other Work

Those Involved



Appearances



** Computing with text

In this short addendum, we will outline some of the ways in which words and books and poems can be made into computable objects -- Objects to which we can apply our new-found data analysis skills. When you think about a book, for example, it usually has a fairly predictable structure. There are chapters which are made up of paragraphs which are made up of sentences which are made up of words. Research areas with names like "stylometrics" attempt to say something quantitative about an author's work, for example, by computing the average number of words per sentence or the average number of letters per word written by an author. Some authors write in short, choppy sentences, while others craft sentences that are over a page long, adding phrase after phrase. Some authors choose simple vocabulary, while others prefer long, complex words. Statistics of this kind can not only point out interesting ways to think about the differences between authors, but they can even be used to help us figure out who wrote texts if their author is unknown or uncertain. One of the earliest analyses of this kind was of the famed Federalist Papers, a collection of documents describing the philosophy and motivation behind our system of government. The papers are thought to be written by Alexander Hamilton, James Madison and/or John Jay. In the mid 1960s, a group of statisticians considered a number of novel statitics to differentiate the writing styles of the three men.

Now, the counts of the different words in a document have also been used to characterize something about the document's subject. In Italo Calvino's novel "If on a winter's night a traveller" we learn how one of his characters "reads" a novel:

Now, the counts of the different words in a document have also been used to characterize something about the document's subject. In Italo Calvino's novel "If on a winter's night a traveller" we learn how one of his characters "reads" a novel:

> She explained to me that a suitably programmed computer can read a novel in a few minutes and record the list of all the words contained in the text, in order of frequency. "That way I can have an already completed reading at hand," Lotaria says, "with an incalculable saving of time. What is the reading of a text, in fact, except the recording of certain thematic recurrences, certain insistences of forms and meanings? (p. 186)
>
> "Words that appear eighteen times: boys, cap, come, dead, eat, enough, evening, French, go, handsome, new, passes, period, potatoes, those, until…
>
> "Don't you already have a clear idea what it's about?" Lotario says. "There's no question: it's a war novel, all action, brisk writing, with a certain underlying violence.

The idea that the frequency with which words appear in a document might reflect something of its content has real-world applications. For example, the spam filter busy intercepting junk e-mail for you is working on the frequency of words in each message. If a message makes too many references to "Viagra" or "won" or "Visa", there is a strong suspicion that the e-mail is spam.

```
> captions <- readLines(
       url("http://mobilize.stat.ucla.edu/day10/data/captions.txt"))

> head(captions)
[1] "unhealthy obsessions. - 22/04/2010" "power lunch"
[3] "Me and Brooke"                      "Frosty Morning"
[5] "Chill Day."                         "Lily"

> capcorp <- Corpus(VectorSource(captions))
> capcorp
A corpus with 2172 text documents

> inspect(capcorp[1:3])
A corpus with 10 text documents

The metadata consists of 2 tag-value pairs and a data frame
Available tags are:
  create_date creator
Available variables in the data frame are:
  MetaID

[[1]]
unhealthy obsessions. - 22/04/2010

[[2]]
power lunch

[[3]]
Me and Brooke

new_capcorp = tm_map(capcorp,tolower)
new_capcorp = tm_map(new_capcorp,removeWords,stopwords())
new_capcorp = tm_map(new_capcorp,removePunctuation)
```

The Presidents

18th Century

1. George Washington
2. John Adams

19th Century

3. Thomas Jefferson
4. James Madison
5. James Monroe
6. John Quincy Adams
7. Andrew Jackson
8. Martin Van Buren
9. William Henry Harrison
10. John Tyler
11. James K. Polk
12. Zachary Taylor
15. James Buchanan
16. Abraham Lincoln
17. Andrew Johnson
18. Ulysses S. Grant
19. Rutherford B. Hayes
20. James Garfield
21. Chester A. Arthur
22. Grover Cleveland
23. Benjamin Harrison
24. Grover Cleveland

Barack Obama | The White House

http://www.whitehouse.gov/about/presidents/barackobama

gary king harvard

Apple Yahoo! Wikipedia SelectorGadget

OUR PRESIDENTS

1. George Washington
2. John Adams
3. Thomas Jefferson
4. James Madison
5. James Monroe
6. John Quincy Adams
7. Andrew Jackson
8. Martin Van Buren
9. William Henry Harrison
10. John Tyler
11. James K. Polk
12. Zachary Taylor
13. Millard Fillmore
14. Franklin Pierce
15. James Buchanan
16. Abraham Lincoln
17. Andrew Johnson
18. Ulysses S. Grant
19. Rutherford B. Hayes
20. James Garfield
21. Chester A. Arthur
22. Grover Cleveland
23. Benjamin Harrison
24. Grover Cleveland
25. William McKinley
26. Theodore Roosevelt

Barack Obama

Barack H. Obama is the 44th President of the United States.

His story is the American story — values from the heartland, a middle-class upbringing in a strong family, hard work and education as the means of getting ahead, and the conviction that a life so blessed should be lived in service to others.

With a father from Kenya and a mother from Kansas, President Obama was born in Hawaii on August 4, 1961. He was raised with help from his grandfather, who served in Patton's army, and his grandmother, who worked her way up from the secretarial pool to middle management at a bank.

After working his way through college with the help of scholarships and student loans, President Obama moved to Chicago, where he worked with a group of churches to help rebuild communities devastated by the closure of local steel plants.

He went on to attend law school, where he became the first African—American

STAY CONNECTED

Facebook YouTube

Twitter Vimeo

Flickr iTunes

MySpace LinkedIn

LATEST NEWS & UPDATES

Read the Blog

WHITE HOUSE PHOTO GALLERY

View the Gallery

THE VICE PRESIDENT of the UNITED STATES

MIDDLE CLASS