Developing Assignments

Challenges

- Challenging and time consuming to create rich, realistic assignments with suitable pedagogical activities.
- Multiple iterations to find a data set, carve out activities at the right level.
- Ideally, we want to borrow from actual research projects as case studies.
- Modern statistical methods are computationally intensive, the mathematical understanding comes later

Goals

- Ideally, we want to borrow from actual research projects as case studies.
- Include modern statistical methods that are computationally intensive yet intuitive
- Have students think statistically in approaching all aspects of data analysis, not just the modeling
- Build on multiple topics to model the complexity of real applications and to promote a sense of accomplishment

Goals, ctd.

- Expect the students to learn some things on their own, i.e. that they will have to find resources outside of those provided in the classroom.
- Participate in the entire data analysis cycle

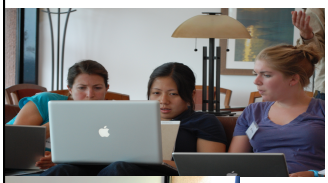
Data Analysis Process

- decompose the problem
- identify key components
- abstract and formulate different strategies
- connect the original problem to the statistical framework
- choose and apply methods
- reconcile the limitations of the solution
- communicate findings

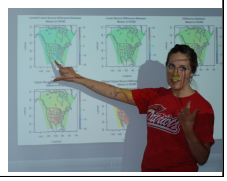
Data Sources for Projects

- Supplementary materials to TAS article lists a dozen datasets and simulation studies
- One excellent source – Explorations in Statistics Research Workshop
- Last year's workshop – contribution from participants - create case studies with a focus on computing
- Data Expo

Explorations in Statistics Research



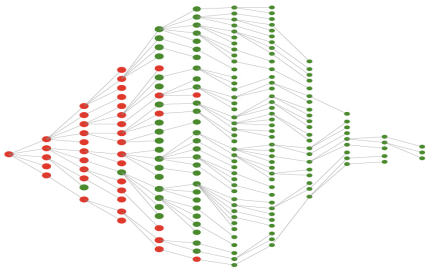
The seven day workshop is designed so that students get a sense of how statisticians approach large, complex problems.



Examples of Data

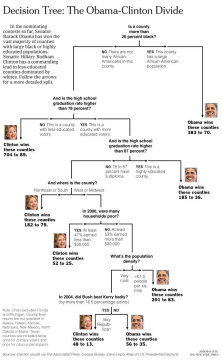
- Baseball database
- Wireless geo-location
- Elephant seal foraging
- CA traffic
- Census/Geographic/Election results
- State of the union addresses
- Spam/Ham Spam Assassin

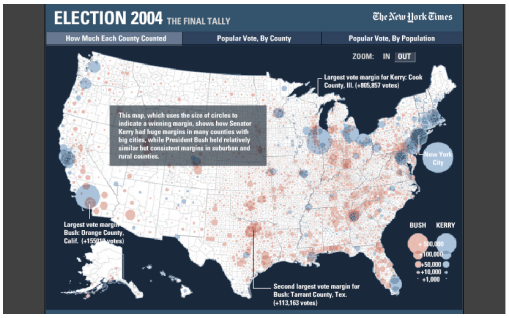
Birth and Assassination Process



Example Project

- Questions
  - Are there differences between the counties that go for Clinton and those that go for Obama?
  - Where are the Clinton-counties located?
  - How well can you predict the primary results for the upcoming primaries?





Data

- 2000 Census data at the county level: Excel files
  - cc07_tabB1.xls: population, land area, and density
  - cc07_tabB3.xls: race, age, gender
  - cc07_tabB4.xls: HS degree, Bachelor's degree, foreign born population, persons in poverty, HH income >$75,000
- 2004 Presidential election: countyVotes2004.txt scraped from CNN website
- 2008 Democratic Primary: DemPrimary2008.txt scraped from CNN Website
- Geographic data: counties.gml latitude and longitude
- Auxiliary data:
  - top 30 cities in country
  - mapping of states to regions
  - mapping of townships to counties for New England States

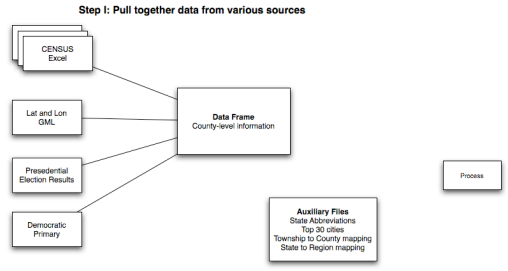
Background

The CNN site presentation of the data changed from 2004 to 2008.

- In 2004, it was available in HTML tables, and could be acquired through programmatic calls to the Web and post-processing the HTML.
- In 2008, the tables were generated on the fly in javascript and json. We used a packet sniffer to figure out the correct URL to call. Then we programmatically called the URL, received json in return, and processed it with RJson package to get plain text.

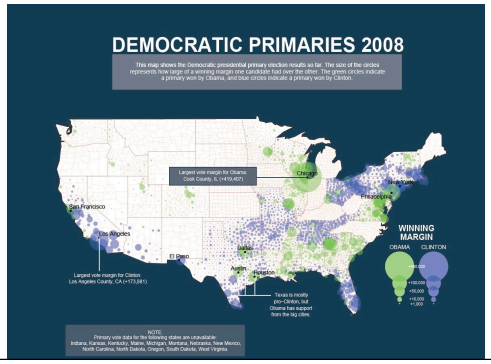
Logistics

- Final Project
- Intermediate deadlines over 3 week period
- Students worked in groups of 3-5
- Turned in: 8-10 page write up; 6 plots; all code (documented).

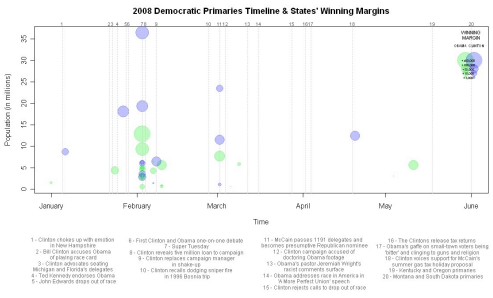


- **Step II: Model Fitting and Map making**
  - Provide plots for comment.
  - This group of students followed the color schemes of the NYT, and they added markers that make it clear that Obama tends to get support in the big cities.
- **Step III: Final Write-up**
  - Students turn in, report with graphics embedded, appendix of how the data frame was created, and code. Samples will be posted on the web.

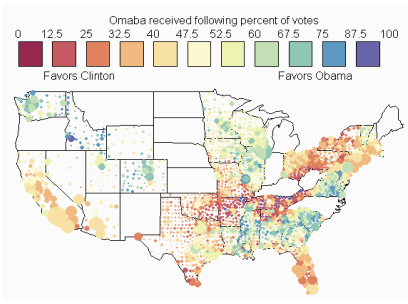
This group of students followed the color schemes of the NYT, and they added markers that make it clear that Obama tends to get support in the big cities.



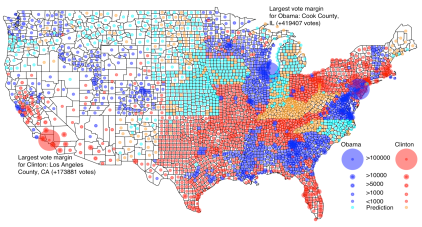
This group of students examined how the state primaries unfolded in time. The size of the circle denotes the vote margin, the color denotes the winner, and there are many markers to indicate political events.



These students display the percentage victory through color, in addition to showing the size of the victory through area of the circle.



These students used their recursive partitioning results to predict the Obama/Clinton wins and place the predictions on the map. It shows the Obama advantage in the deep south, the Clinton advantage in the Appalachians, and the close race in the Midwest.



Goal of HWs

- Practice with writing code, designing functions, and improving code
- Reinforce use of statistical thinking in all aspects of statistical computing
- Vary format of work to be handed in
- Drills vs. Practice
- Opportunity to bring in data analysis
- Variants of an assignment
- Part of a project

Brief Descriptions of HW

- What are skills needed?
- What are the expectations?

Traffic flow on highways in CA

- Read documentation to figure out which function to use to read data
- Explore relationship between 2 variables
- Hand in pdf file – paragraph, code, plots

Wireless Geo-location

- Practice investigating characteristics of data
- Convert from one data structure to another
- Use different structures for different types of analyses

Deconstruct - Reconstruct

- Use common terminology for describing a plot and its components
- Evaluate effectiveness of a plot
- Identify the message in the plot
- Find additional data/markers to improve plot
- Familiarity with basic plotting software
- Turn in HTML

Birth and Assassination Process

- Build simulation from standard RNG functions
- Experience with function writing
- Understand randomness
- Use EDA to test functions
- Use DOE to explore the characteristics of the process

State of Union Address

- Modern methods:
  - Text mining: K-L distance and S-J metric
  - Term frequency; document frequency
  - MDS, Hierarchical clustering
- Experience with function writing
- Understand randomness
- Use EDA to test functions
- Use DOE to explore the characteristics of the process

Cross-validation

- Modern methods:
  - Text mining: K-L distance and S-J metric
  - Term frequency; document frequency
  - MDS, Hierarchical clustering
- Experience with function writing
- Understand randomness
- Use EDA to test functions
- Use DOE to explore the characteristics of the process

Cross-validation

- Experience with function writing
- Profiling code to speed it up
- Well written code -

Student plots from HW & projects

Spam Assassin

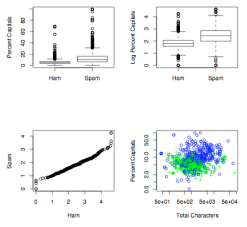
Return-Path: whisper@oz.net
Delivery-Date: Fri Sep 6 20:53:36 2002
From: whisper@oz.net (David LeBlanc)
Date: Fri, 6 Sep 2002 12:53:36 -0700
Subject: [Spambayes] Deployment
In-ReplyTo:<LNBBLJKPBEHFEDALKDLCIEJABCAB.tim.one@comcast.net>
Message-ID: <GCEDKDNBLEFPPADDJCDECEHJENAA.whisper@oz.net>

You missed the part that said that spam is kept in the "eThunk" and was viewable by a simple viewer for final disposition?

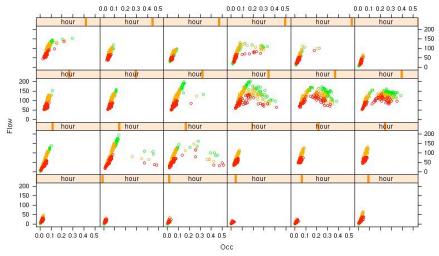
Of course, with Outbloat, you could fire up PythonWin and stuff the Spam into the Junk Email folder... but then you loose the ability to retrain on the user classified ham/spam.

David LeBlanc
Seattle, WA USA

> -----Original Message-----
> From: spambayes-bounces=whisper=oz.net@python.org
> [mailto:spambayes-bounces=whisper=oz.net@python.org]On Behalf

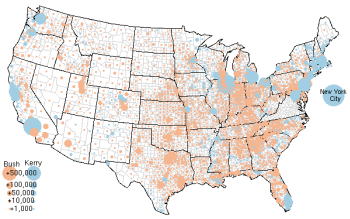


LA traffic at all hours



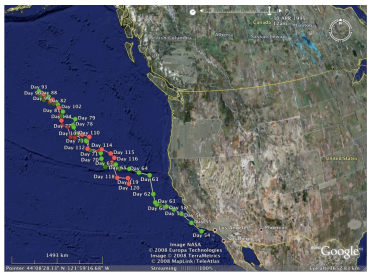
33

County Map
2004 US Presidential Election



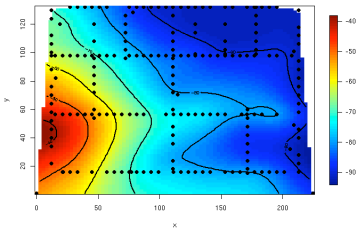
34

Elephant seal migration



35

Wireless geolocation



36