

Computing in the Undergraduate Statistics Curriculum

Deborah Nolan *UC Berkeley*

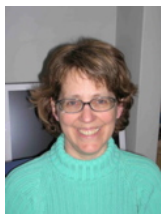
With

Duncan Temple Lang *UC Davis*

Mark Hansen *UCLA*

Partially funded by NSF 0618865

Who we are:



Experience in Industry, Computing,
Pedagogy
Undertaken significant curriculum review
Developed and taught new computing-
related courses

Our Perspective:

- Good computing skills are essential to good data analysis skills
- Modern statistical methods are computationally intensive, the mathematical understanding comes later
- Computing provides insight and understanding for statistical concepts in a constructive and tangible manner
- Students need to express ideas via computation with the same facility as with math

3

2000 Curriculum Review ASA Undergraduate Statistics Education Initiative

- More weight on data in data analysis
- Traditional math stat needs statistical thinking
- Traditional math stat needs computer intensive methods
- Field needs to be redefined to include data management and computer skills

4

Motivation: Computing Today

- Faster computers, parallel computing, data collecting devices, Web 2.0, PDAs (iphone, ipod, ...), multi-media - changing our world
- Science and statistics are changing a great deal, becoming multi- & inter-disciplinary
- Generation, collection, and dissemination of data very different, and very large
 - Web portals, databases, data technologies (XML, SOAP, HTTP requests, ...)

5

Motivation: Computing Today

- Computing is an essential tool with which we study, do data analysis, disseminate ideas
- Just as students learn how to use twitter, gmail, facebook, ... they should also learn the vocabulary & tools of scientific computing to be scientists
- Opportunity to teach modern statistical methods within a computing class to ensure students see new & exciting material

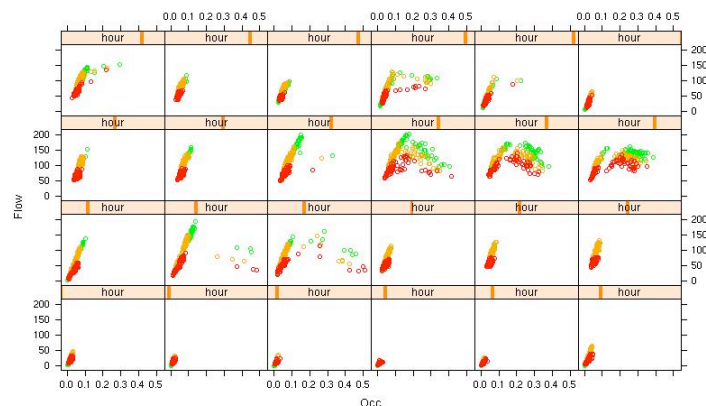
6

A Course: Concepts in Computing with Data

- Perform practical analyses with real, large, problem-driven data
- EDA in modern era with computing
- Participate in the entire data analysis cycle: data acquisition, cleaning, shaping, analysis, modeling, simulation/model checking, reporting
- Think statistically in approaching all of these aspects of data analysis, not just the modeling part
- Acquire computational tools, learn how to use them creatively, gain confidence to continue to learn about and embrace future technologies
- De-emphasize replication of existing well-tested algorithms; scaffold on existing software and functionality

7

Visualization: LA traffic at all hours



8

Text manipulation: Spam Assassin

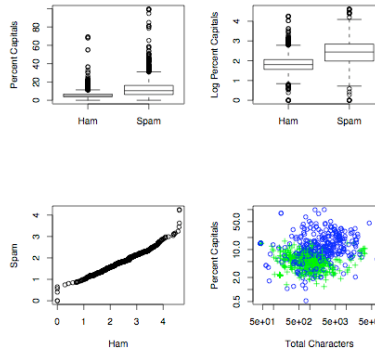
Return-Path: whisper@oz.net
 Delivery-Date: Fri Sep 6 20:53:36 2002
 From: whisper@oz.net (David LeBlanc)
 Date: Fri, 6 Sep 2002 12:53:36 -0700
 Subject: [Spambayes] Deployment
 In-Reply-To: <LNBLJKPBEHFEDALKOLCIEJABCA8.tim.one@comcast.net>
 Message-ID: <GCEDKONBLEFPPADDJCOCEHJENAA.whisper@oz.net>

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

Of course, with Outblat, 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.

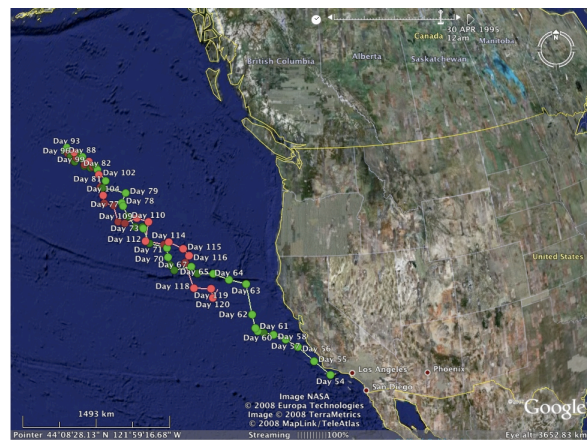
David LeBlanc
 Seattle, WA USA

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



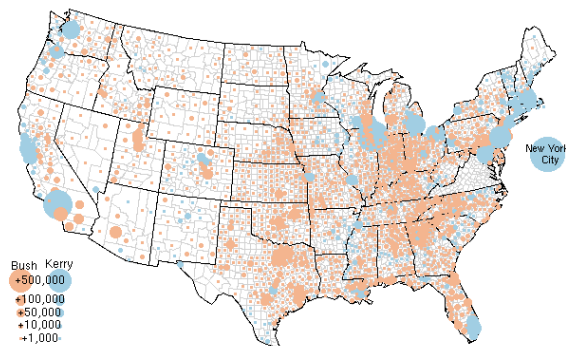
9

XML: Elephant seal migration



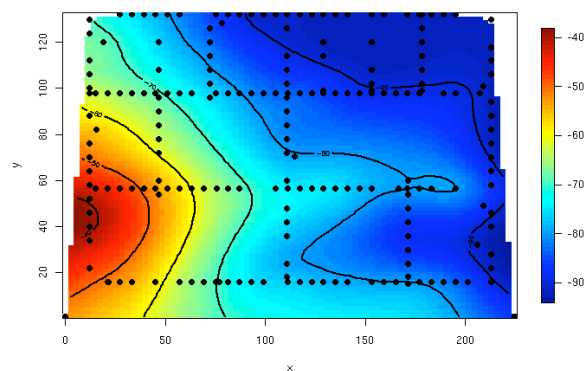
10

Web Data: County Map 2004 US Presidential Election



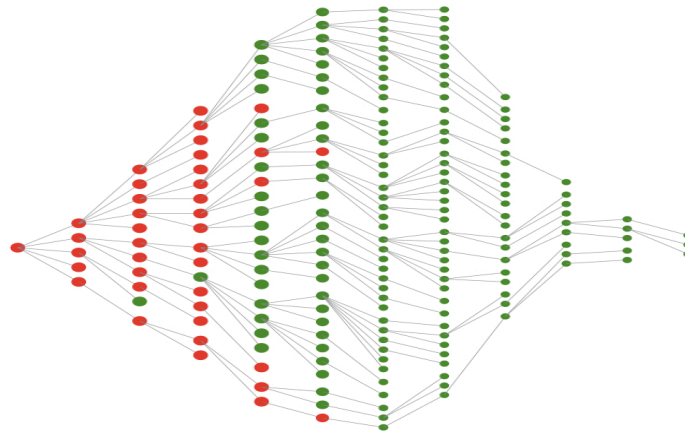
11

Algorithms - Nearest neighbor: Wireless geolocation



12

Simulation - Birth and Assassination Process



13

Why Data Technologies?

- Students get to do things and be creative
- Sense of achievement in analyzing data that are topical in an informal manner.
- Exposure to research and a sense of statistics in action
- Learn practical tools that can be used in real settings
- Connect statistics to other interests and work

14

Computing Concepts

- Computational thinking - Programming concepts
- Data visualization
- Text manipulation - Regular expressions
- Data technologies - databases
- Web 2.0 - XML
- Computationally intensive methods
- Simulation
- Event handling and GUI development

15

Software

- R- statistical software
- Shell commands
- Regular expressions
- SQL - Structured Query Language
- HTML - Hypertext Markup Language
- XML - Extensible Markup Language
- wxWidgets - Toolkit for graphical user interfaces

16

Course Work

- Participation in-class and on-line discussion
- Six short computing assignments
- Two projects completed in groups of 3
- Written final exam

17

One student's feedback:

I have to admit I was pretty naïve as to what I would get out of STAT 133 ... I thought the class ... should only be for CS majors.

STAT 133 was by far the most challenging, rewarding, and influential class I took at Berkeley.

I am currently working at a consulting firm that specializes in statistical and economic research and data analysis for large corporations. ... Every day I work with data, and STAT 133 gave me the tools and foundation to succeed in my current position and gave me the confidence to land the job in the first place.

I believe that this course contributes significantly to a student's understanding of the role computers will have in their future lives in the workforce or in graduate school.

18

NSF Funded Project:

- Sample course syllabi and curricula:
<http://www.stat.berkeley.edu/~statcur>
- Textbook under development
- Faculty development workshops
- Help others introduce changes at their institutions
- Faculty interested in piloting new courses -
statcur@stat.berkeley.edu

19

Obstacles:

- Institutional: Curriculum development slow and narrow in focus
- Views
 - Computing can be self taught and picked up as you go
 - Computing is just a skill and should not be part of the curriculum
- **Faculty training** : We were not taught this; it's not natural for us like math

20

- **Tukey** ('62) Statistics need to be regarded as a set of problems - problems that pertain to data
- **Breiman** ('77) ASA/IMS Conference on the Analysis of Large Complex Data Sets
- **Friedman** ('97) Statistics is being defined by a set of tools
 - Probability, real analysis, asymptotics,..
 - Computing has been the most glaring omission from the set of tools

21

Identified Computing Areas

- Concepts, languages and environments for programming with data
- Data Technologies
- Data and scientific visualization
- Computational statistics and numerical algorithms
- Simulation

22

Examples of Data

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

23