The Role of Data Science in Statistics Education

Deborah Nolan
University of California, Berkeley

Main Message I:
Infusing Statistics Curricula with Data Science Makes Our Students Better Statisticians
AND
Infusing Data Science with Statistics Benefits Data Science.

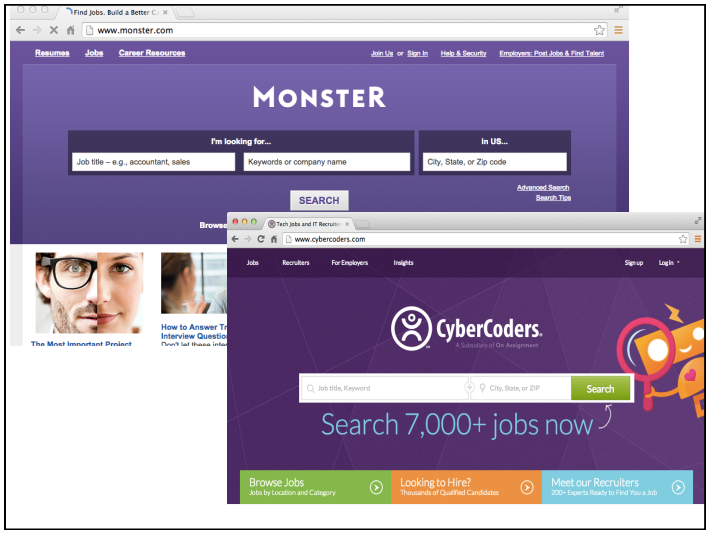
Main Message II:
In Statistics Context Matters. "Context" must include computational context.

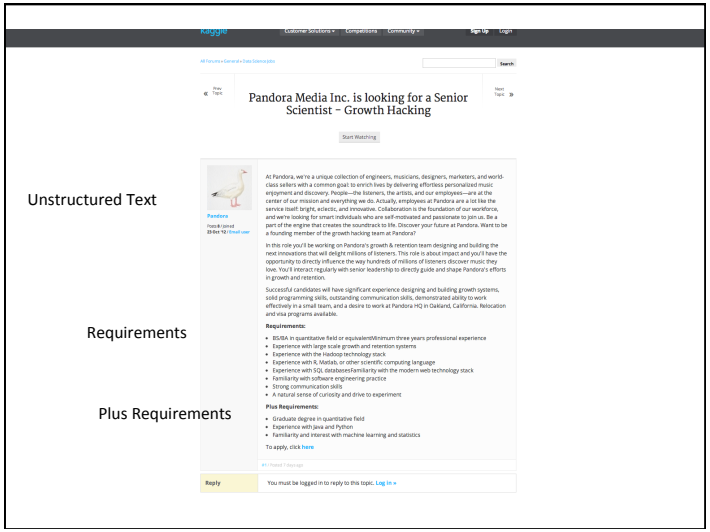
Outline

- Data Science & Statistics
- Tenets of statistical practice
- Course specifics
- Opportunities & Challenges
- Recommendations & Discussion

What is Data Science?

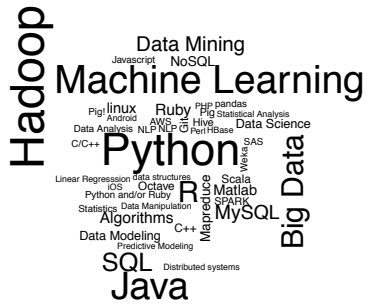
Let's Find Out!





What to do?

- Scrape thousands of listings for Data Science job postings
  - Site specific formats
  - Multiple pages of listings
- Extract skills, education, years experience from unstructured text
- Organize results into analyzable data



BUT –
These are technologies that employers want.

What We Did

- Had a problem of interest
- Proactively found data to help us answer a question
- Data were freely available on the Internet
- Collected data for purpose other than originally intended
- Visualization findings

Data Scientists say

- Three sexy aspects: (Driscoll via Howe)
  - Statistics
  - Visualization
  - Data Munging
- Separate Statistics From:
  - Visualization
  - Prediction
  - Machine Learning
  - EDA

Statistics and Computing

Statistician's Tool Box

Friedman 1997, Role of Statistics in the Data Revolution

- Statistics is being defined in terms of a set of tools… probability, real analysis, asymptotics
- The field of Statistics seems to be defined as the set of problems that can be successfully addressed with these and related tools…
- **Computing** has been one of the most glaring omissions in the set of tools.

Future of Data Analysis

Tukey 1962 (Wilkinson, 2012)

- *Algorithmic* models as important as *algebraic* models
- Model building - a recursive endeavor
- Explore data for surprising insights,
- Data analysis would push the limits of existing computer systems.

Computational Science

SIAM Working Group Computational Science & Engineering Education, 2001

- "Computation is now regarded as an equal and indispensable partner, along with theory and experiment, in the advance of scientific knowledge"
- Computing is an essential, foundational skill for modern data analysis and statistics research

Mathematical Sciences 2025

US National Academy of Science, 2013

- Boundaries within math-sciences and between math-sciences and other subjects are eroding
- Growth areas in statistics are fostered by the explosion in capabilities for simulation, computation, and data analysis
- Computation is central to future research and training in our discipline
- Math-sciences should support scientific computing research

Frontiers in Massive Data Analysis

US National Academy of Science, 2013

"Statistical rigor is necessary to justify the inferential leap from data to knowledge, and many difficulties arise in attempting to bring statistical principles to bear on massive data."

Tenets of Statistical Practice

Practice of Statistics

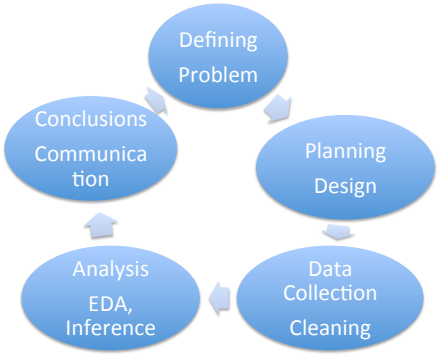
- Context matters
- Data Analysis Cycle
- Visualization
- Communication

How Do Computing and Data Science Enter the Picture?

Context Matters:

- The important difficulties and the whole point of statistics lies in the interplay between the context (the original questions) and the statistics (Speed '86)
- Every time the amount of data increases by a factor of ten, we should totally rethink how we analyze it (Friedman '97)

Data Analysis Cycle

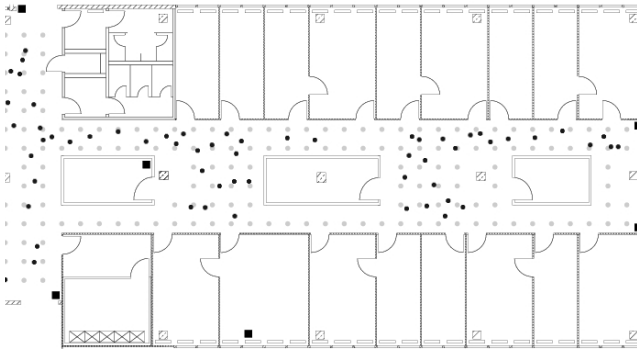


Wild & Pfannkuch ISR 1999

Indoor Positioning Systems – Predicting Location

Use Wi-Fi signals and networks to locate devices (and people) within buildings

Indoor Positioning Systems – Design



Indoor Positioning Systems – Data

```
# timestamp=2006-02-11 08:31:58
# usec=250
# minReadings=110
t=1139643118358;id=00:02:2D:21:0F:33;pos=0.0,0.0,0.0;degree=0.0;\
00:14:bf:b1:97:8a=-38,2437000000,3;\
00:14:bf:b1:97:90=-56,2427000000,3;\
00:0f:a3:39:e1:c0=-53,2462000000,3;\
00:14:bf:b1:97:8d=-65,2442000000,3;\
00:14:bf:b1:97:81=-65,2422000000,3;\
00:14:bf:3b:c7:c6=-66,2432000000,3;\
00:0f:a3:39:dd:cd=-75,2412000000,3;\
00:0f:a3:39:e0:4b=-78,2462000000,3;\
00:0f:a3:39:e2:10=-87,2437000000,3;\
02:64:fb:68:52:e6=-88,2447000000,1;\
02:00:42:55:31:00=-84,2457000000,1
```

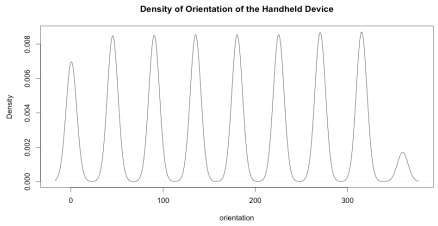
What software to look at the data?

Are #s only at top?

What data structure to use?

Indoor Positioning Systems – Clean Data

EDA to validate orientation

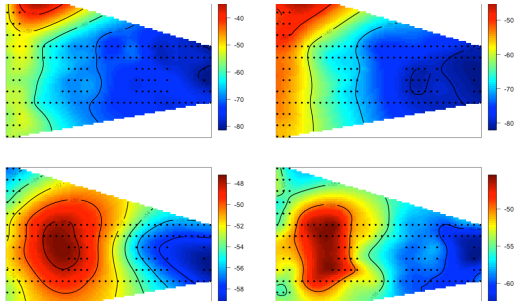


00:04:0e:5c:23:fc	00:0f:a3:39:dd:cd	00:0f:a3:39:e0:4b
418	145619	43508
00:0f:a3:39:e1:c0	00:0f:a3:39:e2:10	00:14:bf:3b:c7:c6
145862	19162	126529
00:14:bf:b1:97:81	00:14:bf:b1:97:8a	00:14:bf:b1:97:8d
120339	132962	121325
00:14:bf:b1:97:90	00:30:bd:f8:7f:c5	00:e0:63:82:8b:a9
122315	301	103

Too many MAC addresses

Indoor Positioning Systems – Deriving Variables

Connect MAC address to router location to compute distance



Indoor Positioning Systems – Modeling and Prediction

- Nearest Neighbor method
  - Distance is 5-dimesional signal strength space
  - Model selection to choose #neighbors
- Bayesian method
- Assessment
  - Training and test data
  - Model comparison

Data Analysis Cycle:

- Closer to data source
- Understand and derive data
- Techniques use to analyze data

Visualization

- Cleveland, Tufte, Wainer, Wilkinson, Brewer, Cook
- Visualization –
  - Make the data stand out
  - Facilitate comparisons
  - Information rich displays

CIA Factbook –

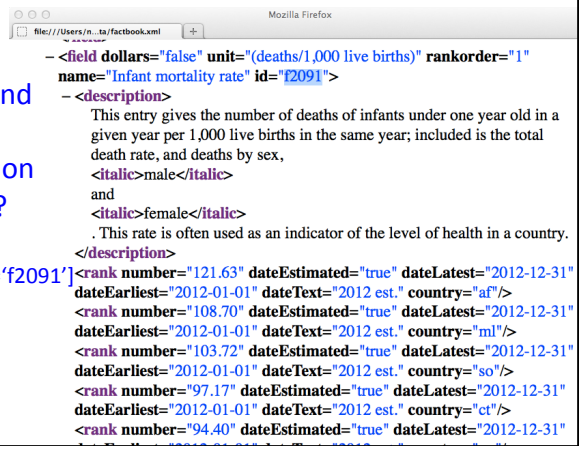
Interactive Visualizations

Mashups

CIA Factbook – Data

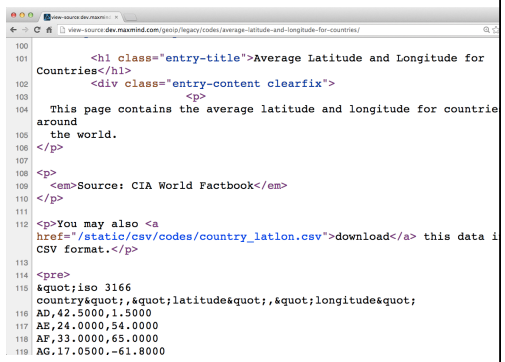
How to find the information we want?

//field[@id='f2091']



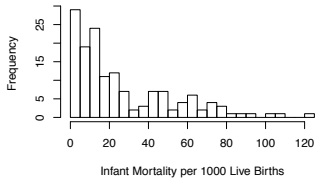
CIA Factbook – Web Scraping

Search the Web to find geographic info to scrape and mash

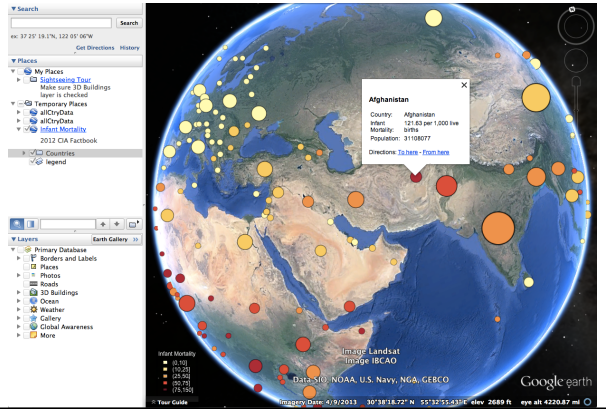


CIA Factbook – Data

How to convert numeric values with skewed distribution into colors?



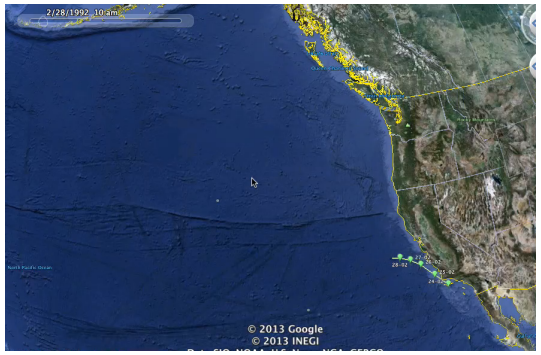
CIA Factbook – Google Earth as a Canvas for Plotting



Virtual Earth Browser: Google Earth

- Scale appropriately as zoom in and out
- Augment your data with others' – contributions from public & govt organizations
- Filter data with folders
- Additional information in pop-up windows
- Animation when associate time with location

Elephant Seal Movements



Model for spatial-temporal relationships

- Combine animation in time and virtual earth browser.
- Google Earth is a 3D plotting canvas
- Extend Formula Language in R to express this relationship

  ~ longitude + latitude @ time | condition
- KML – structured representation of data that GE uses for rendering (*RKML package* (Nolan & Temple Lang) creates KML from data frames)

Communication

- Statisticians communicate through:
  - Writing
  - Visualization
  - Mathematics
- Plus Code –
  - well documented code & pseudo-code are forms of communication

Reproducible Computation

- Well documented code
- Runnable code
- Code connected to the plots, tables, results in publication
- Data provenance
- Sweave, knitr, Rmd

Reproducible Research

- Document database
- Capture ideas – tangents, dead ends, what ifs
- Project document into different views
  - Abstract
  - Research paper
  - Tech Report
- Teach data analysis – uncover the thought process of an expert to use in teaching

Core++:

- Context matters – the scientific context *and* the computational context
- Data Analysis Cycle – Understand data and the techniques we use to analyze data. Good computing skills are essential to good data analysis
- Visualization- Complex & Interactive
- Communication – written language, math, visualization, *and* computation

How should we prepare students for the expanding role of technology and its uses across STEM fields?

The Statistics Course: A Ptolemaic Curriculum? (Cobb 2007)

- What we teach is largely the technical machinery of numerical approximations based on the normal distribution and its many subsidiary cogs.
- This machinery was once necessary. These days we have no excuse.
- We need to recognize that the computer revolution in statistics education is far from over

- In the beginning we taught mathematics and called it statistics; much of this was probability. Then with the help of computers, we started to teach data analysis and statistical modelling; this was fine apart from one feature: it was largely context free. (Speed, 1986)
- Traditional statistics courses do "not attempt to teach what we do, and certainly not why we do it… these courses are caught in a time warp that bores teachers and subsequently bores students." (Efron, 2003)

Concepts in Computing with Data

AKA Data Science for Statisticians

Developed with Duncan Temple Lang

Preparation for work/research

Our Students Need –

- Technical skills to engage in collaborative research and problem solving with data
- To be ready to engage in and succeed at statistical inquiry
- Confidence to overcome computational challenges to carry out a comprehensive data analysis
- Communicate through writing code, as well as writing reports

Concepts in Computing with Data

- Core requirement along with probability and theoretical statistics

Year	Enrollments
2004-2005	40
2013-2014	400
2014-2015	750 (expected)

- Prereqs: NONE (sophomore standing strongly recommended)
- Majors: Stat, Math, Engineering, plus others

Philosophy

- Use the computer expressively to conduct statistical analysis of data
- Use existing software rather than build routines from the ground up.
- Statistical Thinking in the context of computing with data
- Work closely with "original" data

Data Analysis Cycle

- Data ACQUISITION – I/O, string manipulation
- Data CLEANING – verification, manipulation
- Data ORGANIZATION – data frames, databases, XML
- Data ANALYSIS – fit and assess statistical models, conduct exploratory data analysis
- Data SIMULATED – simulation studies to understand behavior of data
- Data REPORTING – presentation graphics, reports

Statistical Concepts

- Basic statistical numeracy
  - Variability, data reduction, method of comparison
- Graphics
  - Elements and principles of graphing
- Computationally intensive methods, e.g.,
  - Classification trees, spline smoothing
- Simulation tools
  - Monte Carlo, bootstrap, cross-validation

Computational Concepts

- Structured Programming
  - Function writing
  - Control flow
  - Environments
- Debugging and efficiency
- Data technologies
  - SQL and relational algebra
  - XML, XPath, KML, SVG
  - regular expressions and text manipulation
- Event Driven Programming – HTML5, JavaScript

Student Work **with Data**

- Problems with data (authentic, problem driven)
- Raw Data -> Analyzable Data
- EDA in modern era ***with*** computing
- Computer intensive methods & simulation
- Model the art of learning new technologies
- Case-based

Opportunities & Challenges

New Modes for Learning:

- Students need to learn how to learn about new technologies
- Multiple venues for learning –
  - MOOCs,
  - Case studies and practical experience
  - Specialized short courses
  - Web resources, e.g., Stacked Overflow

Materials & Delivery

- Training has focused on code recipes and GUI applications, not on computational thinking
- Technology evolving quickly, we need a new paradigm for educating our students
  - Dearth of educational materials
  - Textbook teaching is slow to respond and sole viewpoint

Access to Technology

- Easy to get lost in the latest technology and miss the point of computational problem solving
- High-end technology is not readily available

US

- Many faculty do not have the skill set, time, confidence to keep current
- Culture shift in our community to recognize essential role of computing in our field

What Can You Do?

Keep Up To Date

- Enroll in an online course in statistical computing and modern data sciences, e.g.,

Roger Peng's "Data Science Specialization"

Bill Howe's "Introduction to Data Science"

- Attend a Summer school in computing/data science
- Partner with faculty who can help you bridge the gap

Advocate for Change

- Hire faculty who can bridge this gap
- Revise your undergraduate curriculum in a BIG way
- Send signal to others that computing and data science are important to the success of statistics

Contribute to Resources

*Introduction to Data Technologies (*Murrell, 2009, https://www.stat.auckland.ac.nz/~paul/ItDT/)

Data Science Specialization - Peng et al, 2014, http://jhudatascience.org/

Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving (2015, Nolan & Temple Lang)

Conclusions & Discussion

Critical point in statistics

Computing is an increasingly vital part of statistics in this era of

- Ubiquitous data availability & sources.
- Increased volume and complexity of data.
- New and ever-evolving Web technologies.
- Increased relevance of data analysis in all fields, done by non-statisticians

We must integrate computing into our statistics programs at a significant and serious level to enable our students to:

- Have the essential skills needed to engage in collaborative research
- Have the confidence needed to meet computational challenges in comprehensive data analyses
- Engage in and succeed at statistical inquiry

Discussion Points

- Why not just take traditional CS courses?
- What do we eliminate from statistics curricula to fit in this new material?
- Technologies are constantly changing so what do we teach?
- How is data science different from data analysis?
- Is data science simply vocational training?
- What are the core concepts of data science?