ISTA 116
Statistical Foundations for the Information Age

Mihai Surdeanu

SISTA: University of Arizona

January 9, 2013

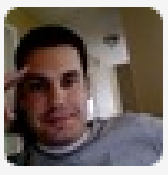
Why should I take this course?

Why?

How would you like to build algorithms that:

- Discover life-threating cases in Medicaid applications.
- "Debug humans" for healthcare.
- Discover flu trends using aggregated data from Google search queries.
- Recommend stocks to buy/sell using Twitter. (???)
- Use Twitter data to forecast political elections.
- Recommend similar items on eBay.
- Decide which ads to display for Google searches.
- Match employers and job applicants on Linkedin.
- Recommend music one will like on iTunes.

Data Science



Josh Wills @josh_wills

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

November 20, 2012

Data Scientist among best new jobs in USA

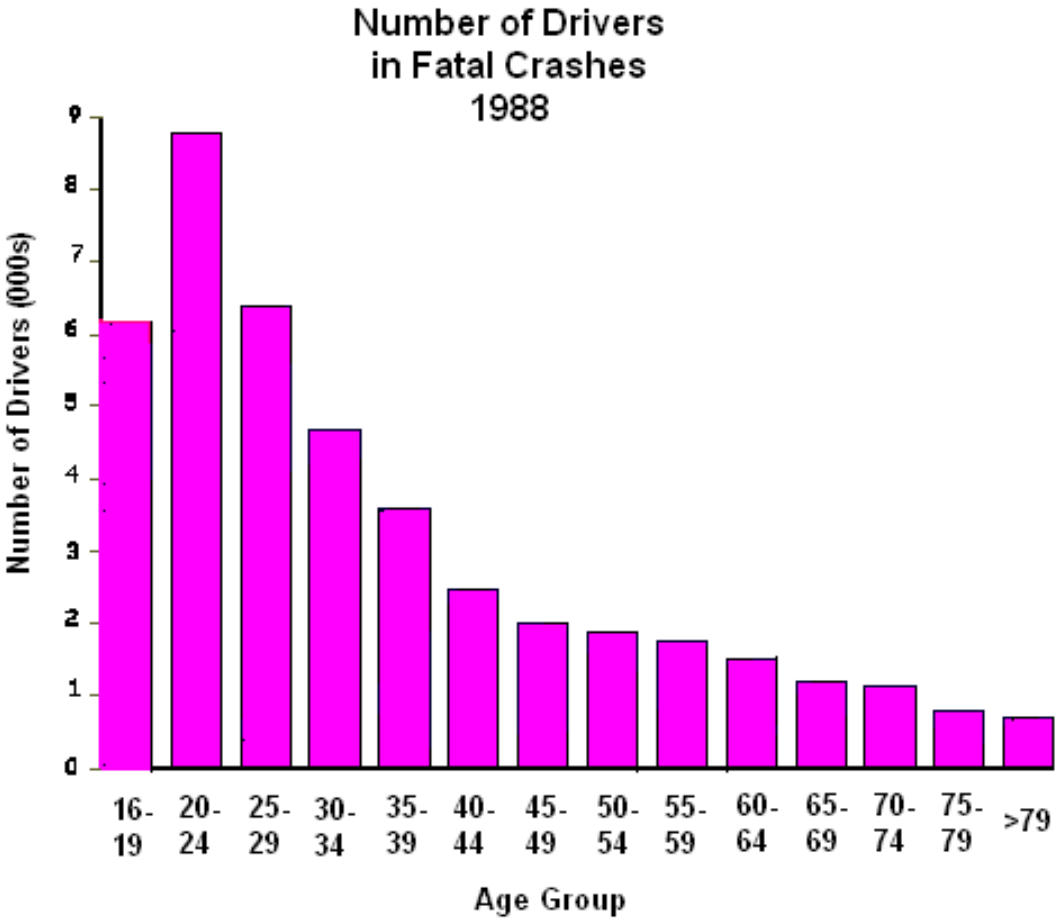
According to CNN Money. Video Game Designer, Social Media Manager and jobs in the green energy sector also made the list. The Harvard Business Review agrees: last month, they called 'Data Scientist' the "sexiest job of the 21st century".

Data Science

Be an intelligent consumer of data.

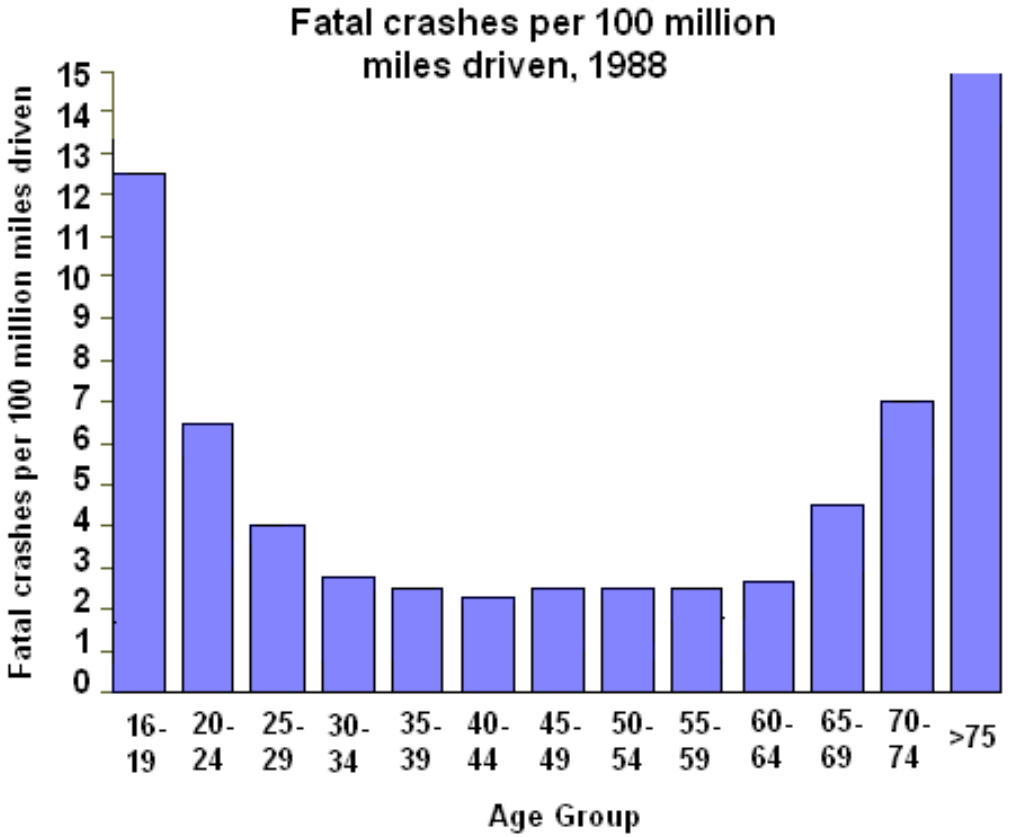
(More on this soon)

Bar Chart: Fatal Crashes



Graph is based on data from this study: Williams, Allan F., Ph.D., and Oliver Carston, Ph.D., "Driver Age and Crash Involvement," Am J Public Health 1989; 79: 326-327.

Bar Chart, same data



Graph is based on data from this study: Williams, Allan F., Ph.D., and Oliver Carston, Ph.D., "Driver Age and Crash Involvement," Am J Public Health 1989; 79: 326-327.

Social Security Spending

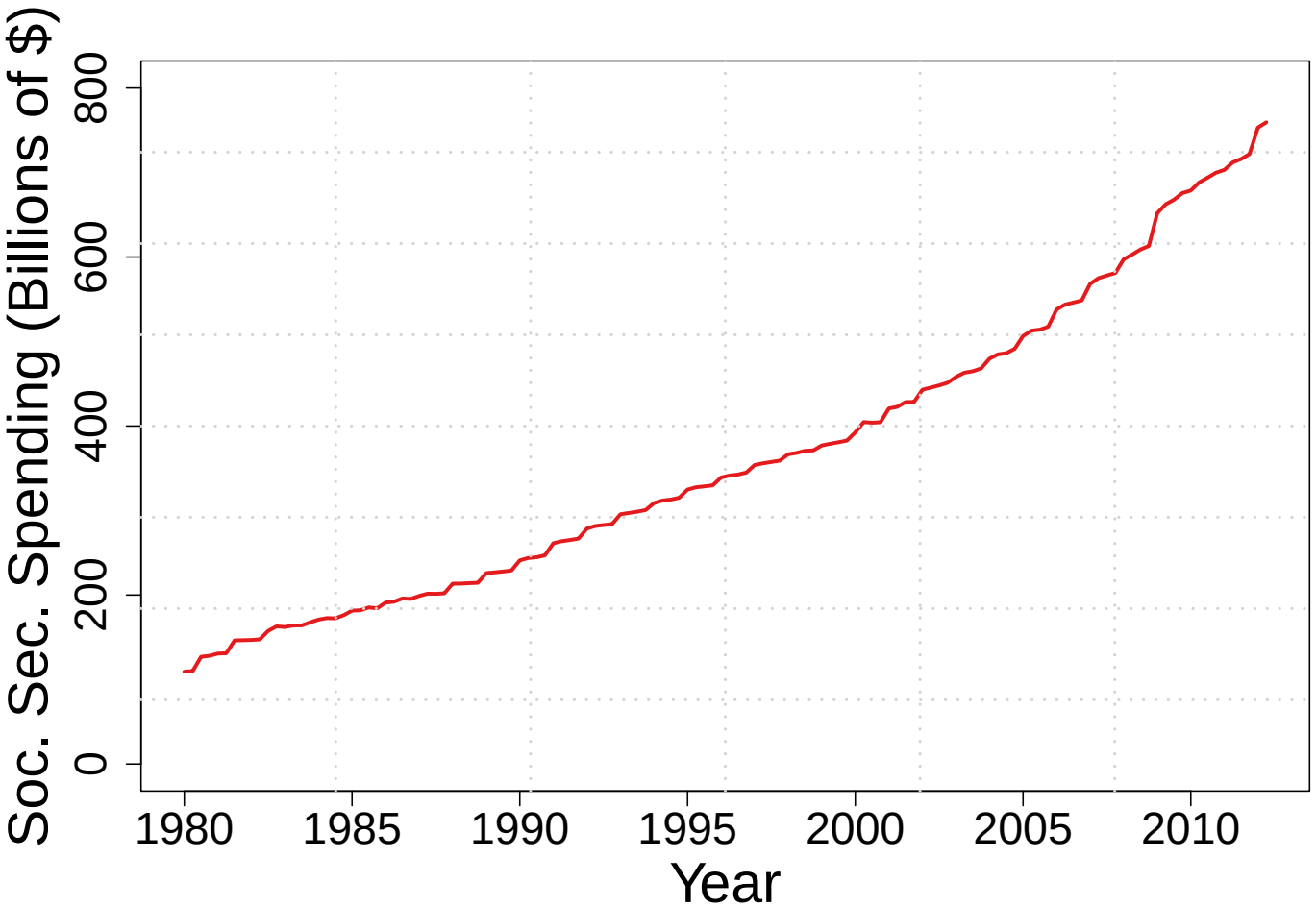


Figure : Total Social Security Spending, 1980-present

Social Security Spending

Three problems with this:

- Lots more people now than in 1960
- Meaning of $1 is not constant
- Economic output (GDP) per person increases over time, even after inflation-adjustment, so the value of what each person pays in to the system increases, too.

Social Security Spending

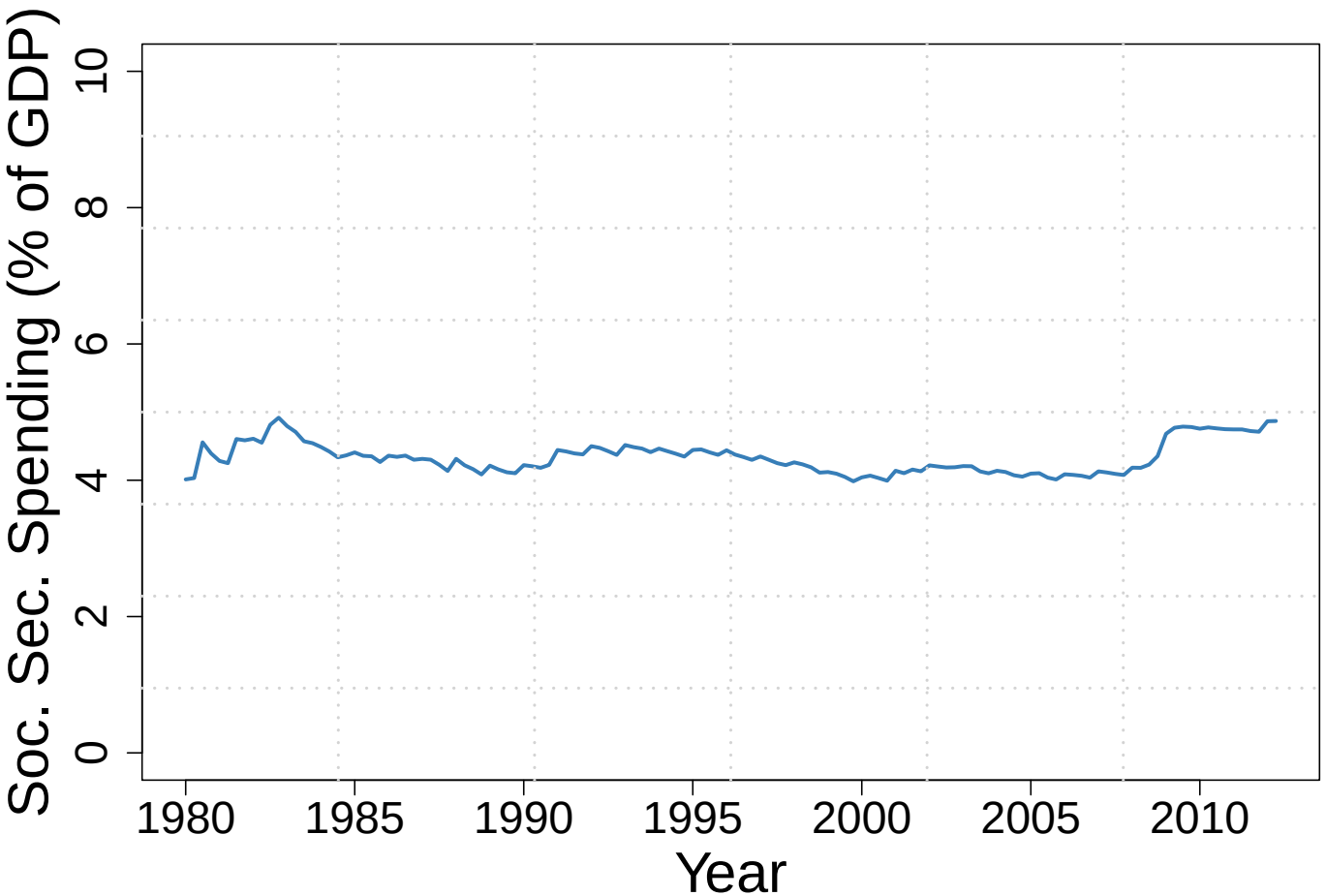


Figure : Social Security Spending as % of GDP, 1980-present

Map of Death

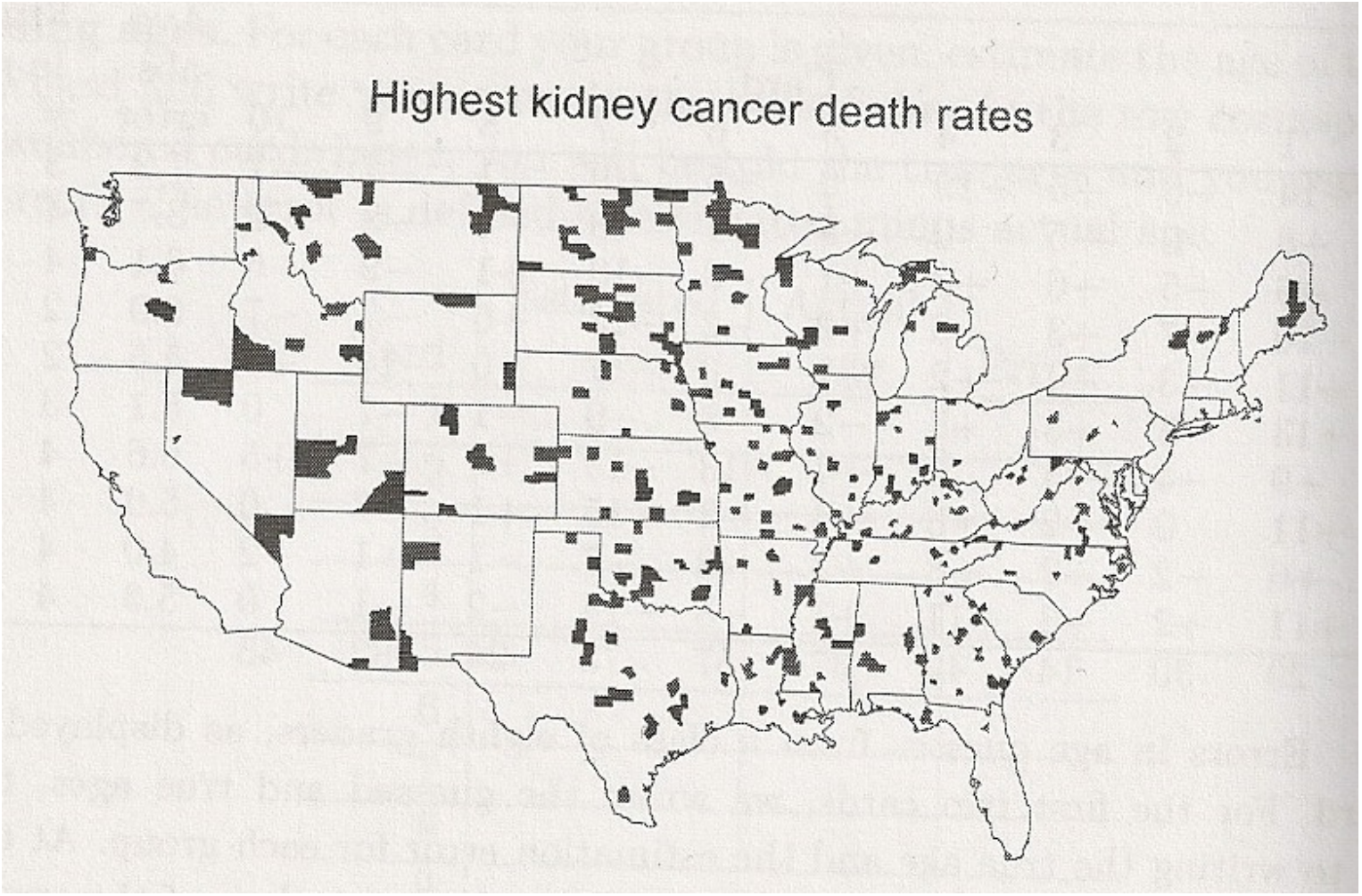


Figure : The counties with kidney cancer death rates in the top 10% nationally (from Gelman and Nolan, 2002)

Map of Somewhat Less Death

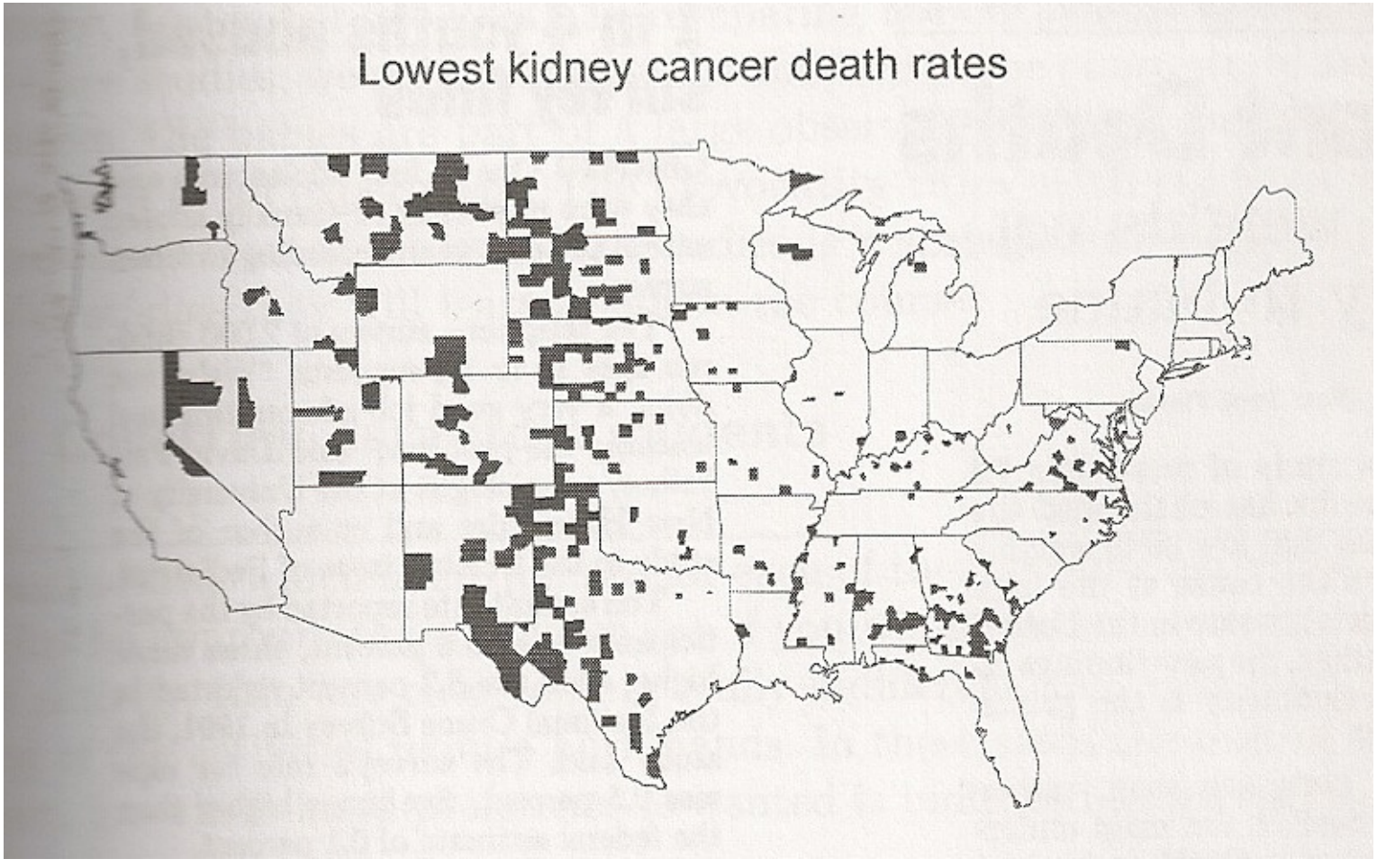


Figure : The counties with kidney cancer death rates in the **bottom** 10% nationally (from Gelman and Nolan, 2002)

Confusing Maps of Death

- Both the highest and lowest rates (as a percentage of cases) occur in low population counties: fewer cases, easier to get values near 0% and 100%, by *random chance*.

Undead Detection

- Suppose a test for Latent Zombieism (affecting 1 in 10,000 people) is 99% accurate in both directions: 99% of zombies-to-be test positive, and 99% of the unafflicted test negative.
- If you test positive, what is the probability you will become a zombie?

Undead Detection

Suppose 1 million people are tested.

	True Positive	**True Negative**	**Total**
Test Positive			
Test Negative			
Total			1,000,000

Undead Detection

About 1 in 10,000 will have LZ.

	True Positive	**True Negative**	**Total**
Test Positive			
Test Negative			
Total	100	999,900	1,000,000

Undead Detection

Of those, about 99% will test positive.

	True Positive	**True Negative**	**Total**
Test Positive	99		
Test Negative	1		
Total	100	999,900	1,000,000

Undead Detection

Of the rest, about 99% will test negative.

	True Positive	**True Negative**	**Total**
Test Positive	99	9,999	10,998
Test Negative	1	989,901	989,902
Total	100	999,900	1,000,000

Of those who test positive, less than 1% will actually become zombies! That is, until the rest start getting bitten...

Race and the Death Penalty

Data from 1981 Florida Homicide Cases

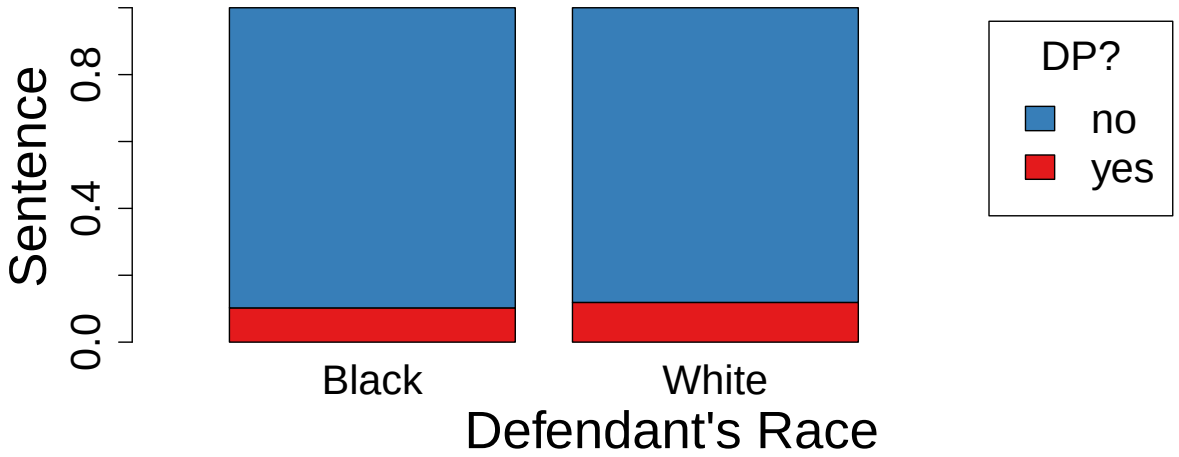


Figure : Proportions of Death Sentences Conditioned on Defendant's Race

- The rate of death sentences is similar for black and white defendants (slightly higher for white defendants):

Race and the Death Penalty

- However, notice what happens when we separate the data by both victim and defendant race.

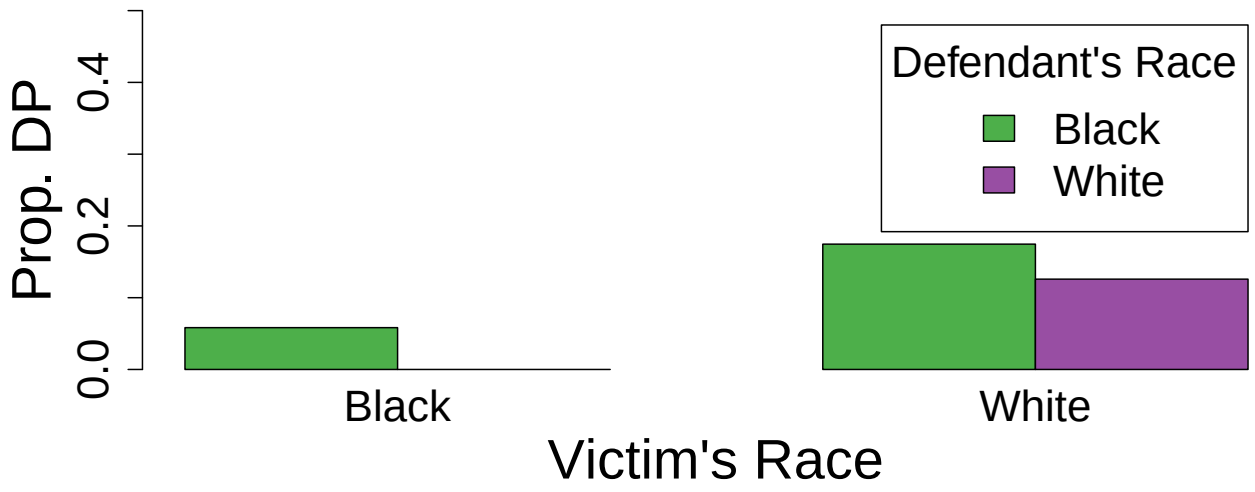
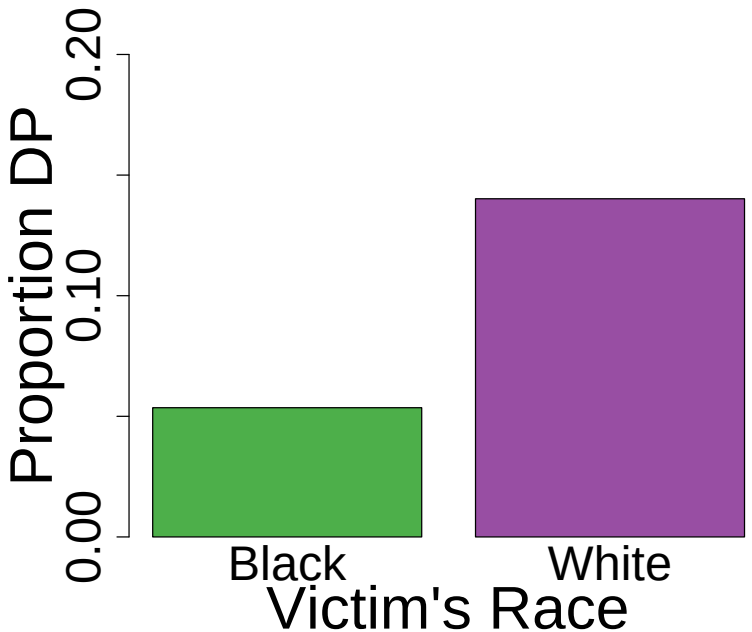
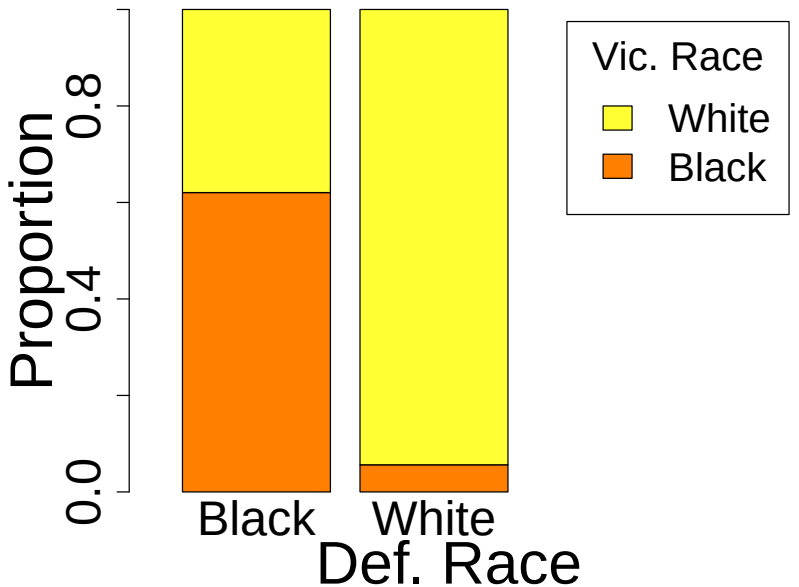


Figure : Proportions of Death Sentences Conditioned on Both Victim's and Defendant's Race

Race and the Death Penalty

- How can this be?





- The DP is applied more often for white victims; and most homicides involve same-race individuals.
- There were almost twice as many white victims.

Simpson's "Paradox"

- This reversal of a trend when a third variable is involved is called **Simpson's Paradox** (we'll see more examples like this in coming weeks).

Goal: Become an intelligent consumer of data.

- Understand the **variability** inherent in data.
- **Evaluate** and **interpret** data presented numerically and graphically.
- Learn to **present** data in meaningful and understandable ways.
- Become conversant in the basics of **probability**: the mathematical discipline that connects a concrete hypothesis about some phenomenon to what to *expect* the data to look like.
- Use tools of probability to evaluate the strength of the **evidence** for hypotheses (i.e., how surprising would this data be if the hypothesis were *false*?)

Essential Info

Me: **Mihai Surdeanu**

email: msurdeanu@email.arizona.edu

Office: Gould-Simpson 811

Office Hours: M/W 9:15 – 10:30

Website: `http://www.surdeanu.info/mihai/teaching/ista116`

Lab Instructors

Nathan Dykhuis
ndykhuis@cs.arizona.edu
Office: Gould-Simpson 918
Office Hours: M/W 3:30 – 4:30

Richard Reilly
richardr@email.arizona.edu
Office: Gould-Simpson 228
Office Hours: Tue 2:30 – 4 and F 12:30 – 2

Lab Time and Place

- Tues 12:00 – 1:50 – ECE 229 (with Richard)
- Tues 2:00 – 3:50 – McClelland Park 102 (with Nathan)
- Tues 4:00 – 5:50 – Shantz 338 (with Nathan)

Topics

Part I: Summarizing and Visualizing Data

- Descriptive statistics, creating and reading graphs
- In lab, learn the statistical package `R` for computation and plotting

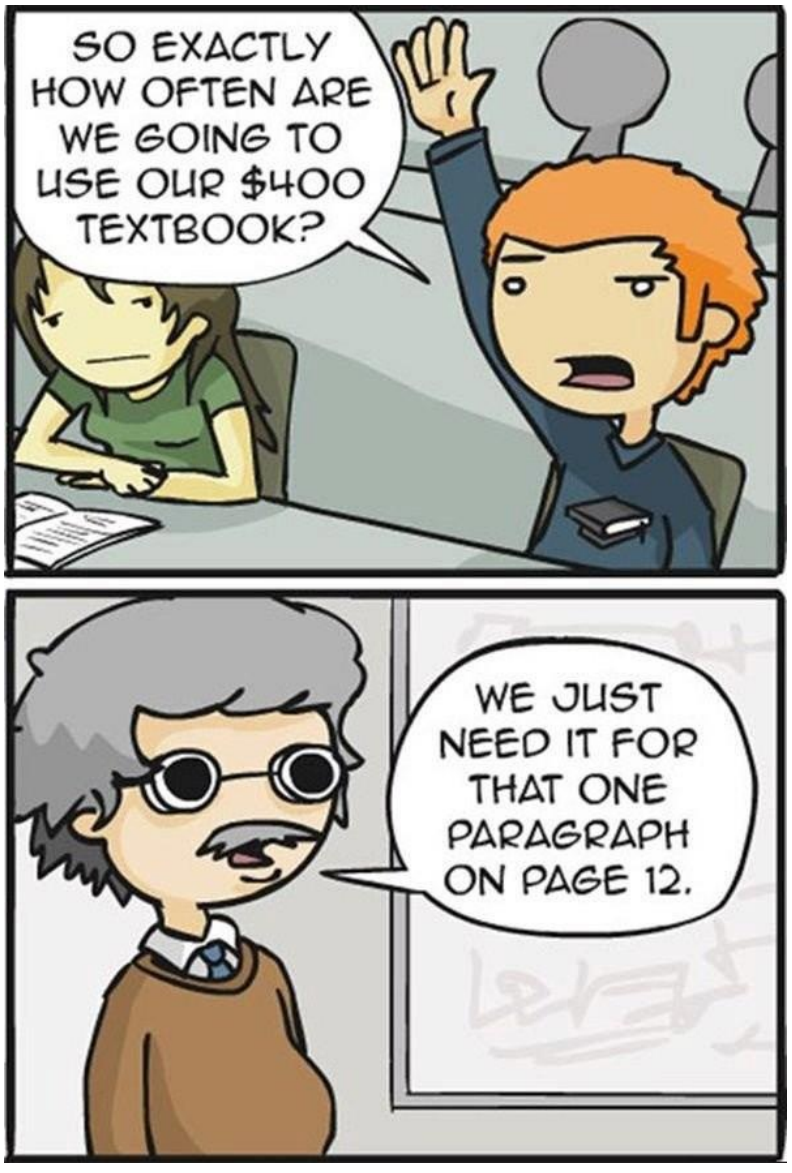
Part II: Basics of Probability

- Common probability distributions, random sampling, the behavior of sample means.
- Use `R` in lab to simulate random phenomena.

Part III: Foundations of Statistical Inference

- Defining a statistical hypothesis
- The logic of hypothesis testing
- Evaluating how "surprising" the data would be

Textbooks

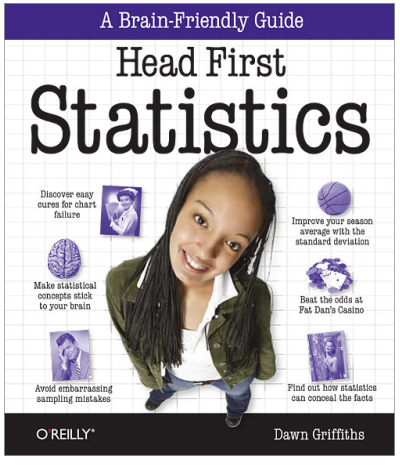

SO EXACTLY HOW OFTEN ARE WE GOING TO USE OUR $400 TEXTBOOK?
WE JUST NEED IT FOR THAT ONE PARAGRAPH ON PAGE 12.

Textbooks

NOTE:

All of the books for this course are available online for free through the links given below.

Main Text



Griffiths, D. (2009). *Head First Statistics*. Sebastopol, CA: O'Reilly Media, Inc.

Available online through the University of Arizona bookstore:

`http://proquest.safaribooksonline.com.ezproxy2.library.arizona.edu/book/statistics/9780596527587`

Using R

Owen, W. J. (2010). *The R Guide, ver. 2.5*. Department of Mathematics and Computer Science, University of Richmond, Richmond, VA.

Available for free online at:

`http://cran.r-project.org/doc/contrib/Owen-TheRGuide.pdf`

Additional References

See the website

Grades

Component	**Weight**
Homework	300 pts
In-Class Quizzes	150 pts
Web Assignments	50 pts
Term Paper	100 pts
Midterm Exam	200 pts
Final Exam	200 pts
Total	1000 pts

Homework (30%)

- Roughly every 2 weeks
- Use computational techniques learned in lab to explore concepts learned in lectures
- Turn in electronically via d2l (lab instructors will clarify formatting, etc.). Due dates all on Fridays.
- You may consult with other students, but **you must write up your own solutions** and **you may not copy answers** (see syllabus for details).

Quizzes (15%)

- Short (5 min.) quizzes on basic definitions, etc., given toward the beginning of every(ish) class. First one is next Wednesday.
- Miss no more than 2 quizzes in first 3 weeks, or 1/3 of total to avoid administrative drop!

Web Assignments (5%)

- Shorter HW assignments done via d2l that do not use R
- Due in weeks when no full HW is due (also Fridays)
- Similar to exam questions
- Multiple attempts allowed

Term Paper (10%)

- One 4-6 page paper due toward end of semester: April 26
- Find a dataset yourself and explore a question of interest

Exams (40%)

- Mixed format (MC, fill-in, written response)
- Midterm is March 6th
- Final is May 3rd
- Higher score counts for more (see syllabus)

To Dos

Before first lab:

In labs we will be using RStudio which can be downloaded here: http://rstudio.org/ if you would like to install it on your home machine or laptop.

As soon as possible:

Read info on website:
http://www.surdeanu.info/mihai/teaching/ista116
Take web assignment 1 in D2L (due next week)