# ISTA 116: Statistical Foundations for the Information Age

Dr. Colin Dawson

The University of Arizona

22 August 2011

# Outline

- *Statistical Foundations for the Information Age*
- The "Information" in "Information Age" is all about **data**.
- What is (/are) data?

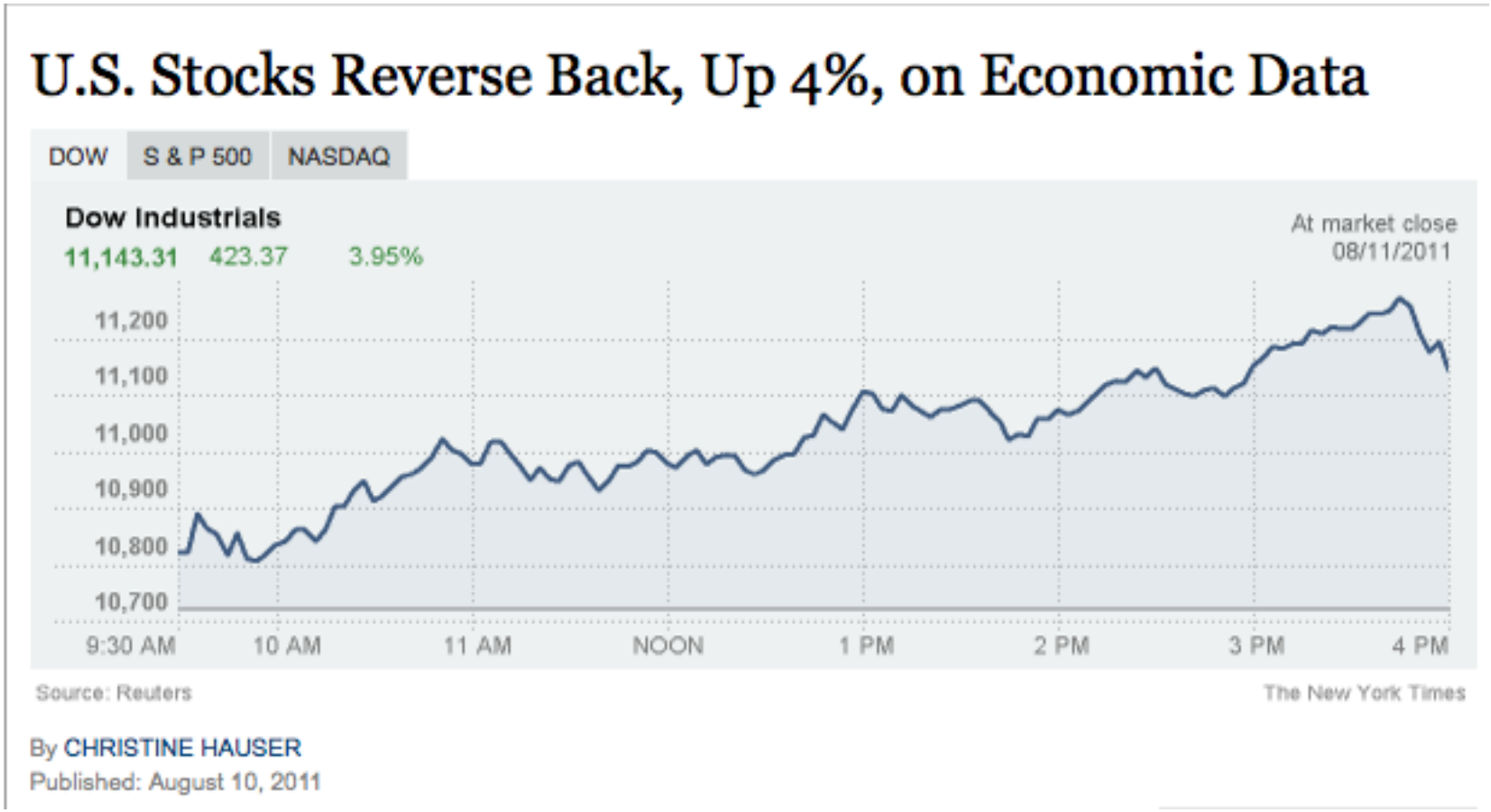



Figure: A recent front page headline at `nytimes.com`

Thursday's close was the first time that the S. & P. 500 had a change of at least 4 percent for four straight trading sessions since 2008. It closed up 51.88 points, or 4.63 percent, at 1,172.64.

It was also the first time that the Dow Jones industrial average closed with a net change of 400 points or more for four straight sessions. It closed 423.37 points higher, or 3.9 percent, at 11,143.31.

# Disapproval Rate for Congress at Record 82% After Debt Talks


Andrew Burton/Getty Images

Watching the House vote. A record 82 percent of Americans disapprove of the way Congress is handling its job, a poll found.

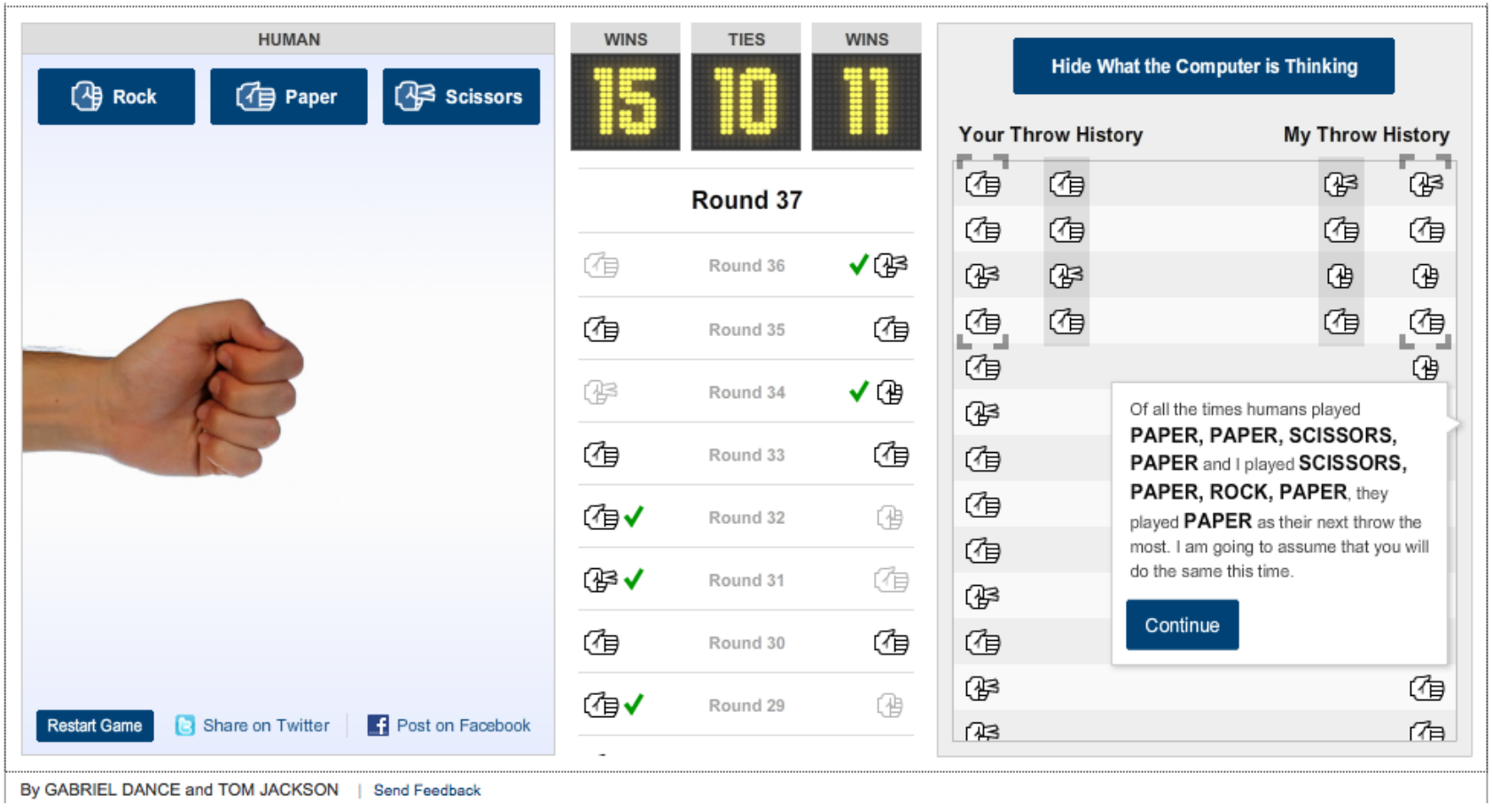

ON PASSAGE
CAPITOL HILL
2,350.50

RunningWithJoy.com

**More about...**

Creepy Pictures »
Wow »
Fish Pictures »
Animals and Facts »
5K Training Schedule »
Marathon Training »
Running Training »
Half Marathon »

About these links


HUMAN
Rock
Paper
Scissors
Restart Game
Share on Twitter
Post on Facebook
WINS
15
TIES
10
WINS
11
Round 37
Round 36
Round 35
Round 34
Round 33
Round 32
Round 31
Round 30
Round 29
Hide What the Computer is Thinking
Your Throw History
My Throw History
Of all the times humans played PAPER, PAPER, SCISSORS, PAPER and I played SCISSORS, PAPER, ROCK, PAPER, they played PAPER as their next throw the most. I am going to assume that you will do the same this time.
Continue
By GABRIEL DANCE and TOM JACKSON | Send Feedback

## Definition

**Data** consists of observations made in a systematic way that can be quantified and used to make decisions.

**Statistics** is the discipline of summarizing and interpreting data in mathematically justified ways.

**Goal:** Become an intelligent consumer of data.

- Understand the **variability** inherent in data.
- **Evaluate** and **interpret** data presented numerically and graphically.
- 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.

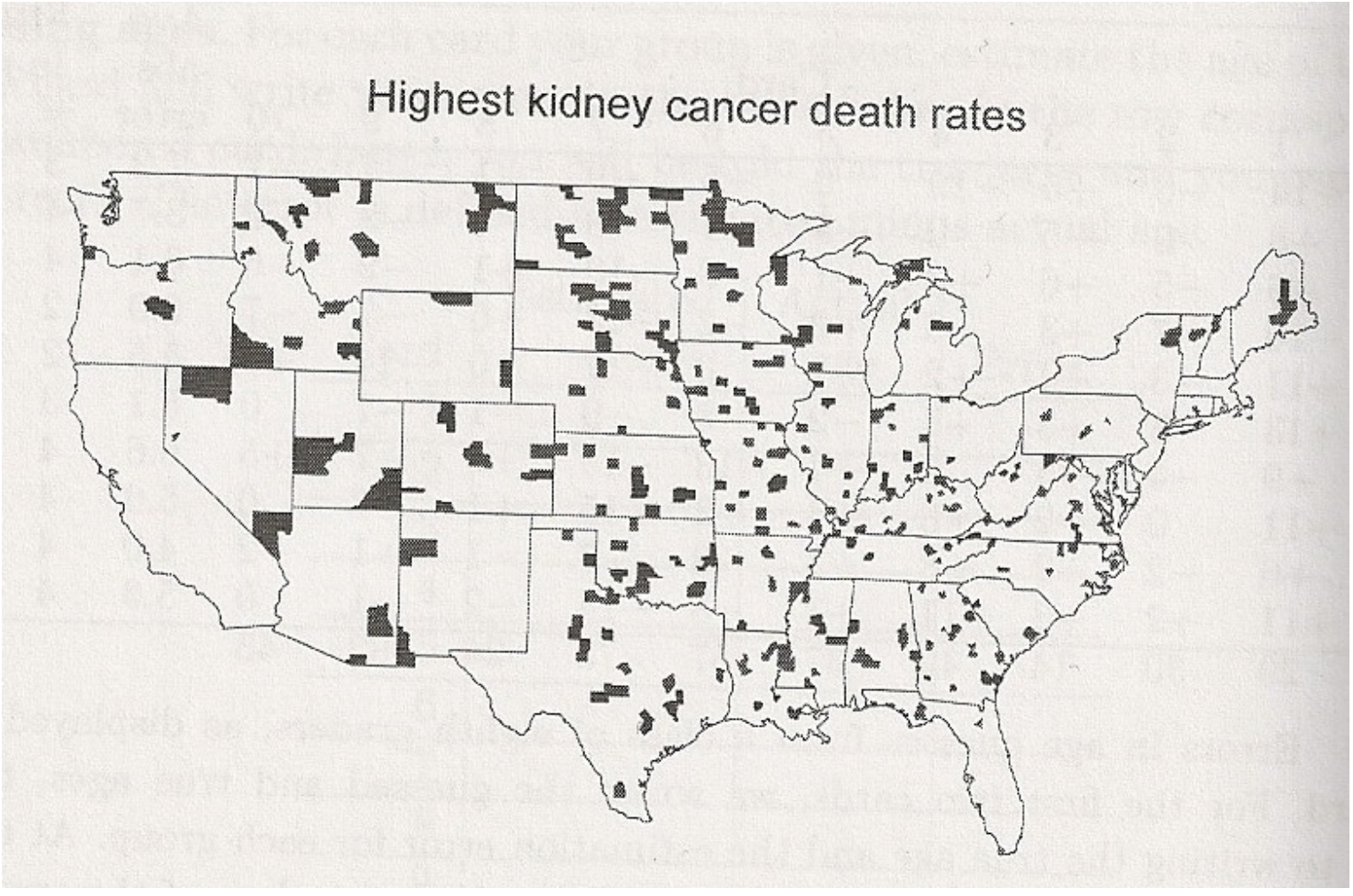



Figure: The counties with kidney cancer death rates in the top 10% nationally (from Gelman and Nolan, 2002)

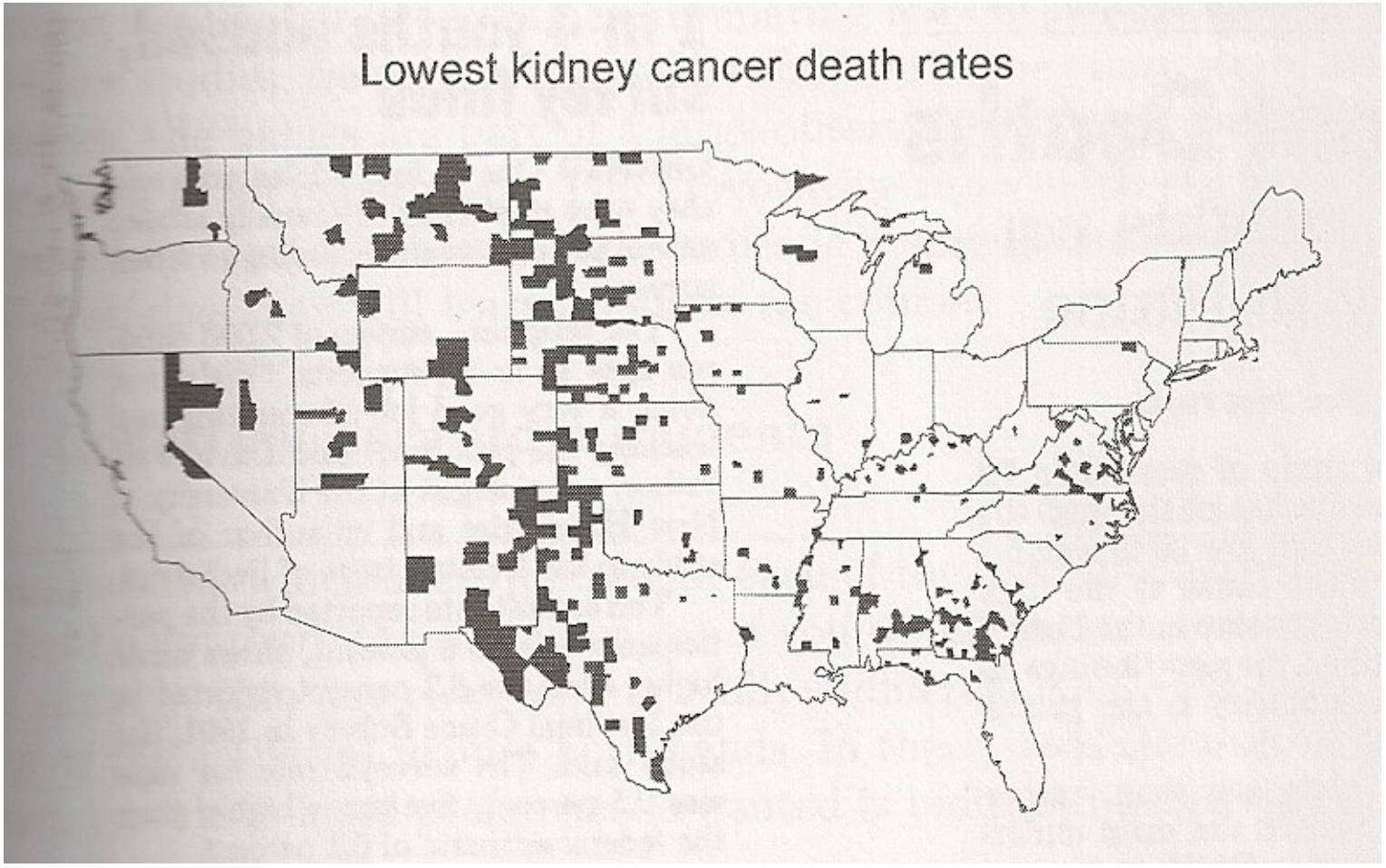



Figure: The counties with kidney cancer death rates in the **bottom** 10% nationally (from Gelman and Nolan, 2002)

- 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*.
- Highlights importance of understanding variability and probability.

- Suppose a test for a somewhat rare disease (affecting 1 in 10,000 people) is 99% accurate: 99% of sick people test positive, and 99% of healthy people test negative.
- If you test positive, what is the probability you have the disease?

Suppose 1 million people are tested.

| | **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 |

Two 50 minute lectures, and one 1 hr. 50 minute lab per week.

- Lecture: Monday and Wednesday, 1:00-1:50, this room.
- Lab: Various times on Tuesdays or Wednesdays.
- Me: Dr. Colin Dawson (`cdawson@email.arizona.edu`)
  - Office Hours: W 2-4 and by appointment (Gould-Simpson 850)
- Lab instructors:
  - Jeff Berry (`jjberry@email.arizona.edu`)
    - Office Hours: Th 2:30-3:30, F 11:00-1:00 (Douglass 304)
  - Derek Green (`dtgreen@cs.arizona.edu`)
    - Office Hours: M 10-12 and 2-3 (Gould-Simpson 721)
  - Philip Sparks (`pspark@email.arizona.edu`)
    - Office Hours: M 3-4 and Th 12-1 (Gould-Simpson 930a)
  - Anu Venkatesh (`anuvenk@email.arizona.edu`)
    - Office Hours: T 2-4 (Gould-Simpson 930a)

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.

## Note!

- Course info and syllabus at `d2l.arizona.edu`
- If you're using a laptop, follow the links on d2l to install `R` before you get to lab!
- Web quiz on the syllabus is due this Friday via d2l!