Statistical Abuses and Course Wrap Up

There Are Three Kinds of Lies

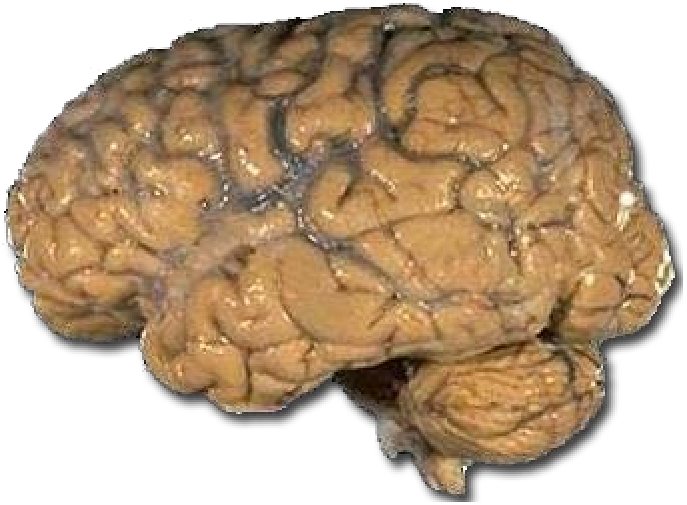
LIES

DAMNED LIES

and

STATISTICS

Humans and Statistics

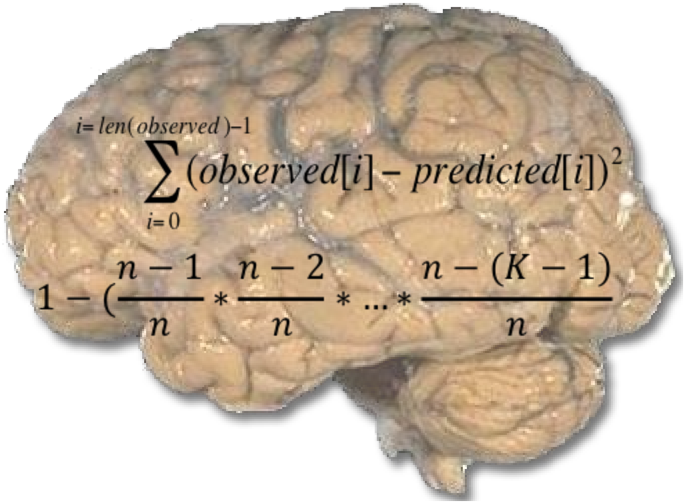


$$\sum_{i=0}^{i=len(observed)-1}(observed[i]-predicted[i])^2$$

$$1-(\frac{n-1}{n}*\frac{n-2}{n}*...*\frac{n-(K-1)}{n}$$

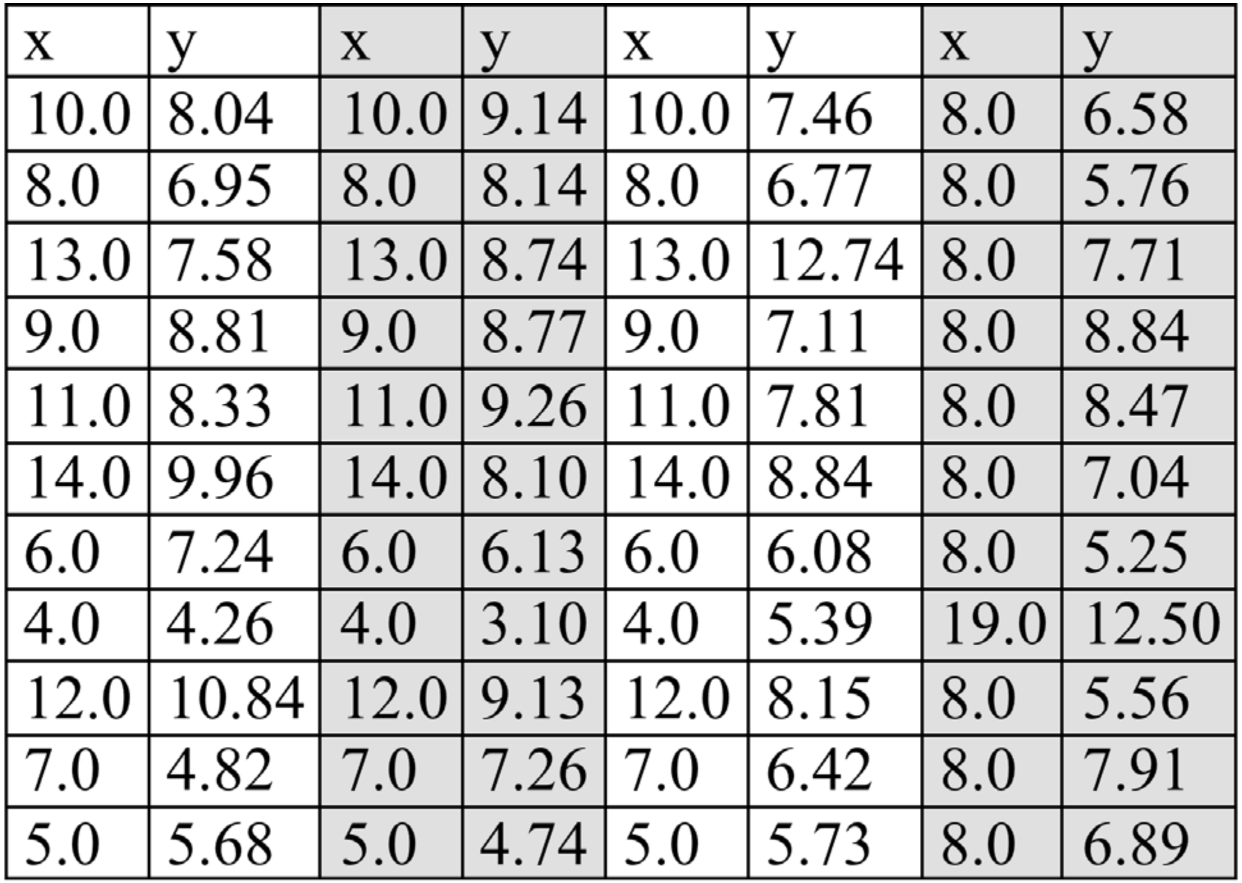
Humans and Statistics

"If you can't prove what you want to prove, demonstrate something else and pretend they are the same thing. In the daze that follows the collision of statistics with the human mind, hardly anyone will notice the difference." – *Darrell Huff*


$\sum_{i=0}^{i=len(observed)-1}(observed[i]-predicted[i])^2$
$1-(\frac{n-1}{n}*\frac{n-2}{n}*...*\frac{n-(K-1)}{n}$

Anscombe’s Quartet

- Four groups each containing 11 x, y pairs

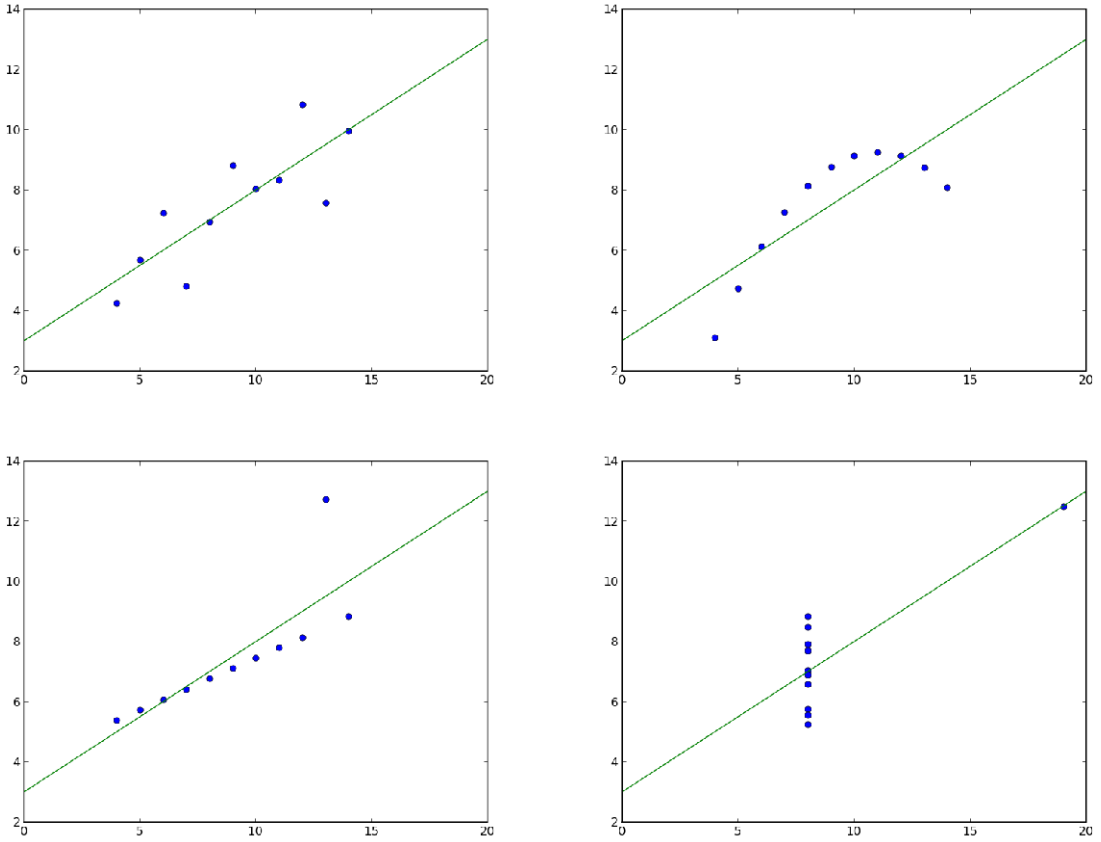


x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Summary Statistics

- Summary statistics for groups identical
  - Mean x = 9.0
  - Mean y = 7.5
  - Variance of x = 10.0
  - Variance of y = 3.75
  - Linear regression model: y = 0.5x + 3
- Are four data sets really similar?

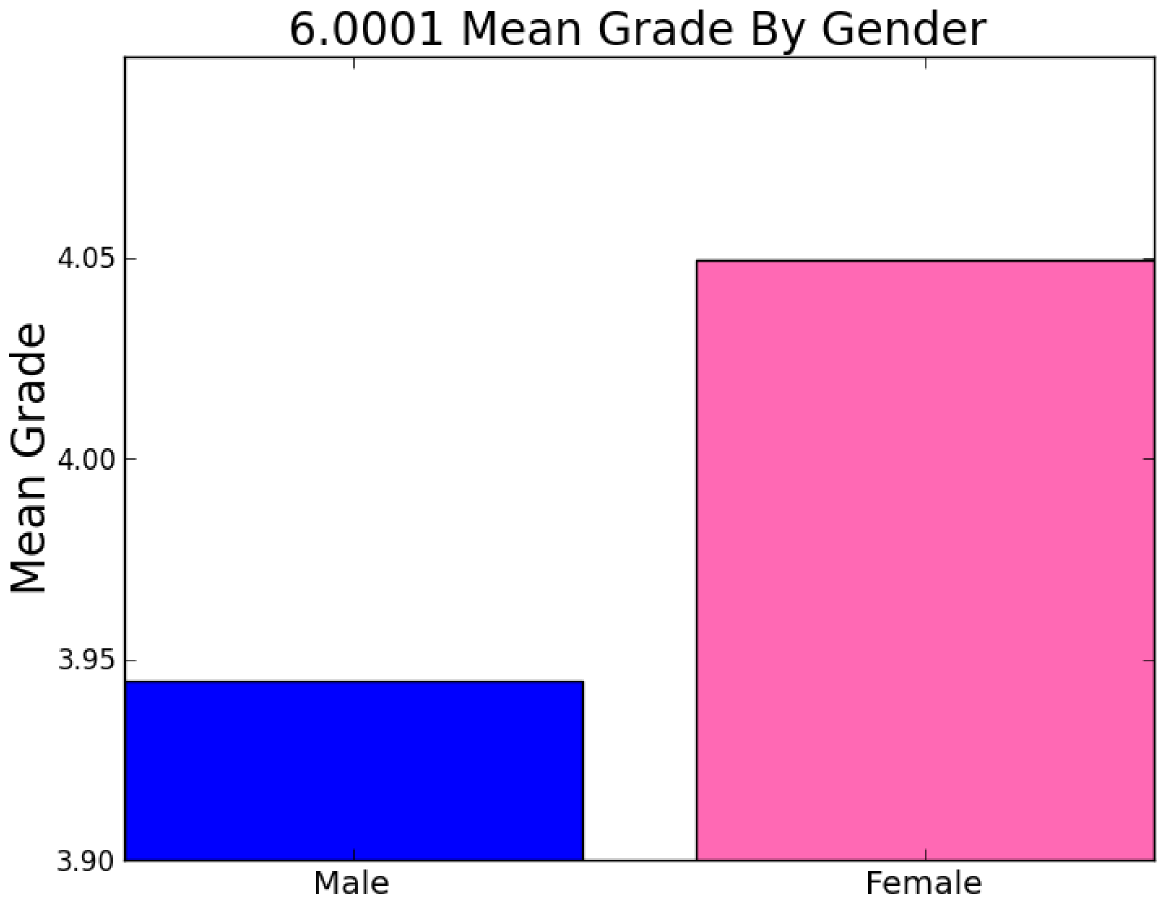
Let’s Plot the Data



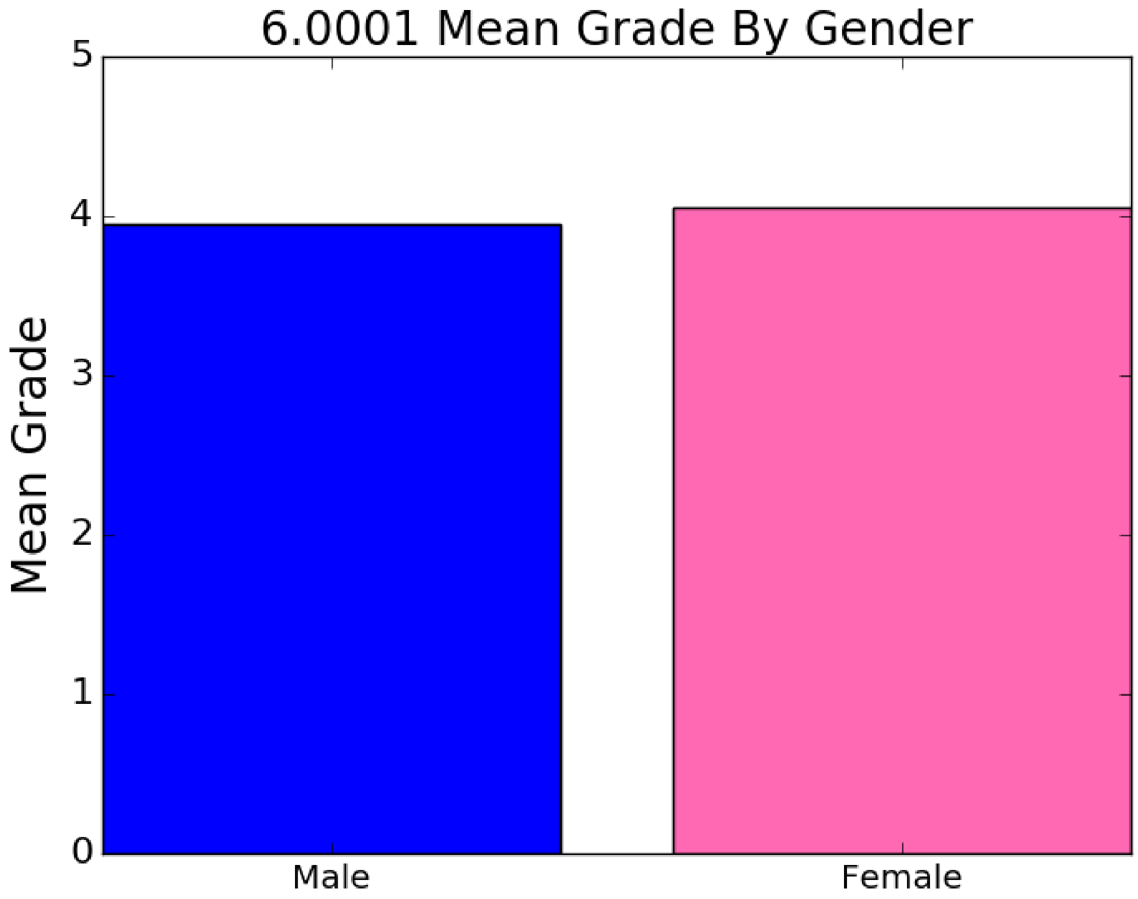
Moral: Statistics about the data is not the same as the data

Moral: Use visualization tools to look at the data itself

Lying with Pictures

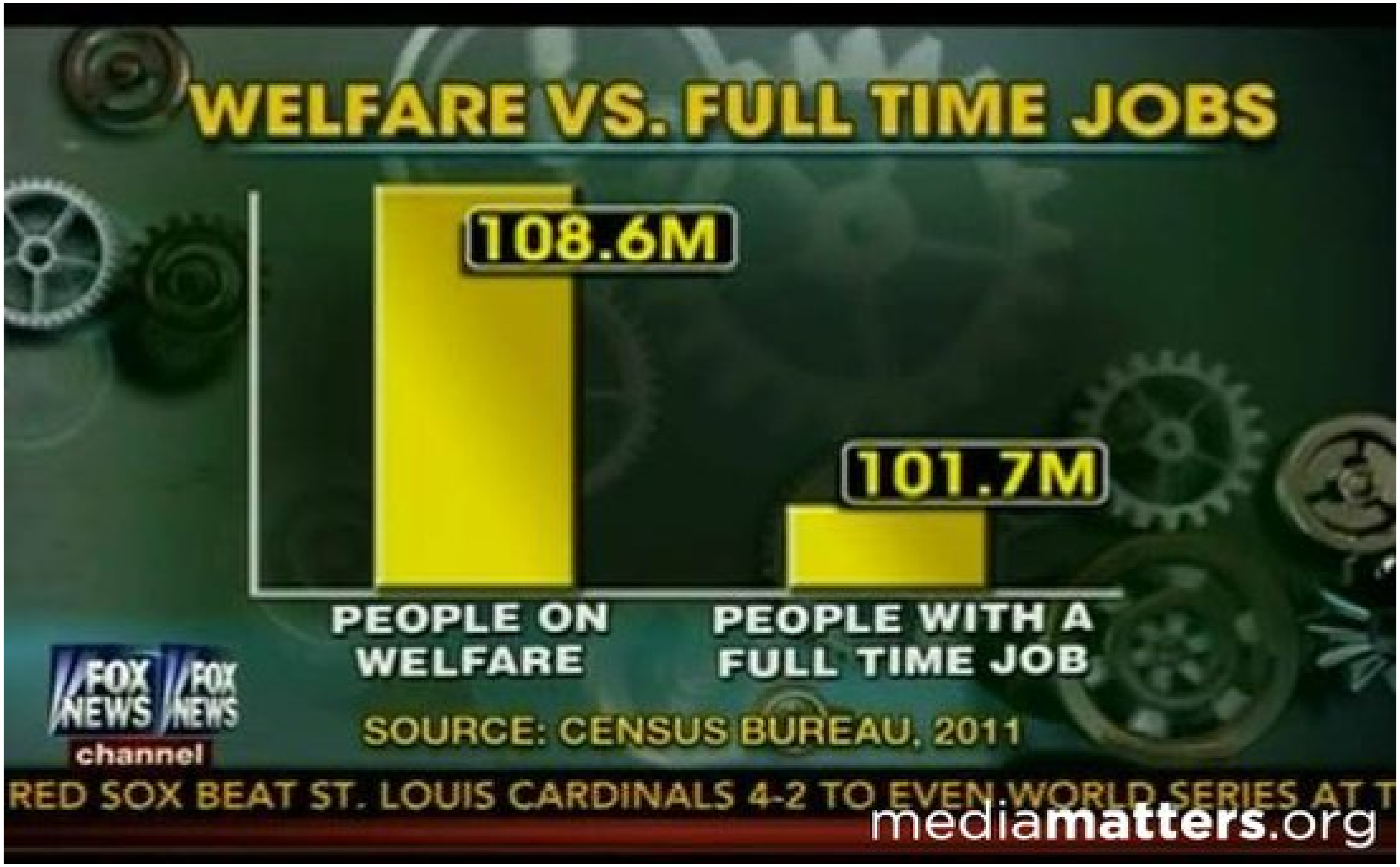


Telling the Truth with Pictures



Moral: Look carefully at the axes labels and scales

Lying with Pictures



Moral: Ask whether the things being compared are actually comparable

GIGO

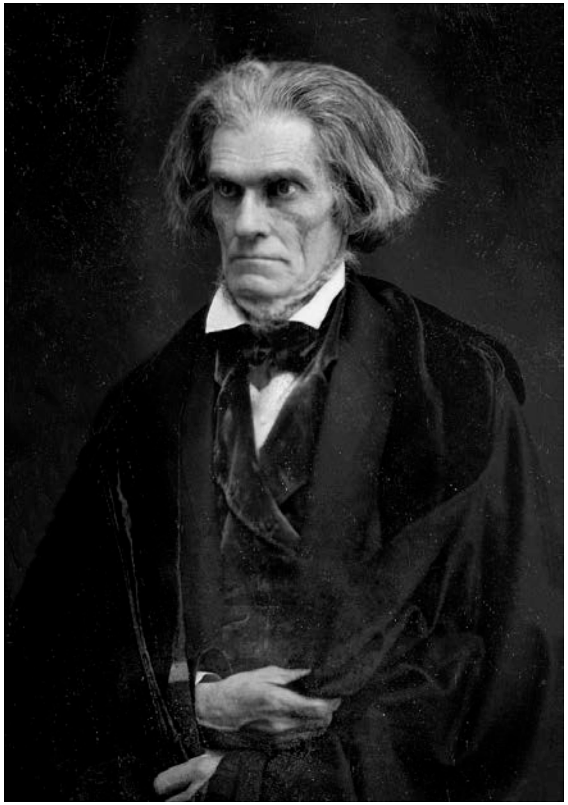
Garbage In, Garbage Out

“On two occasions I have been asked [by members of Parliament], ‘Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?’ I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question.” – Charles Babbage (1791-1871)

GIGO in the 1840's

"The data on insanity revealed in this census is unimpeachable. From it our nation must conclude that the abolition of slavery would be to the African a curse."

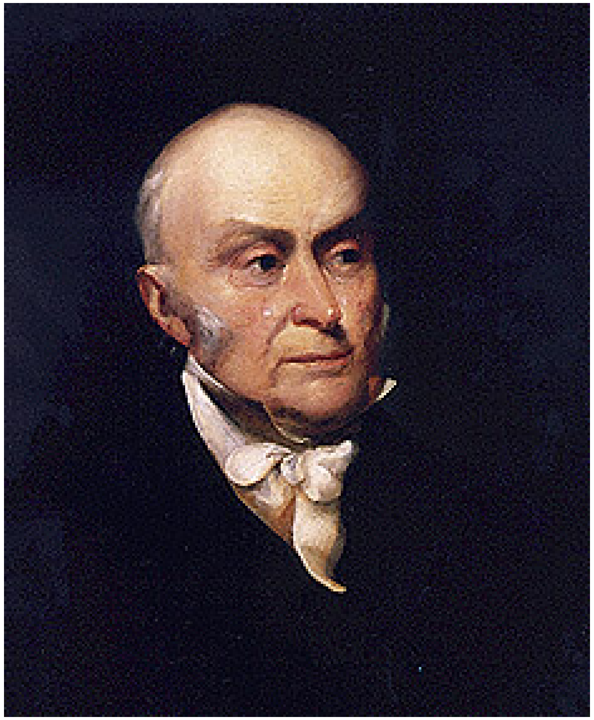
– John C. Calhoun
U.S. Secretary of State



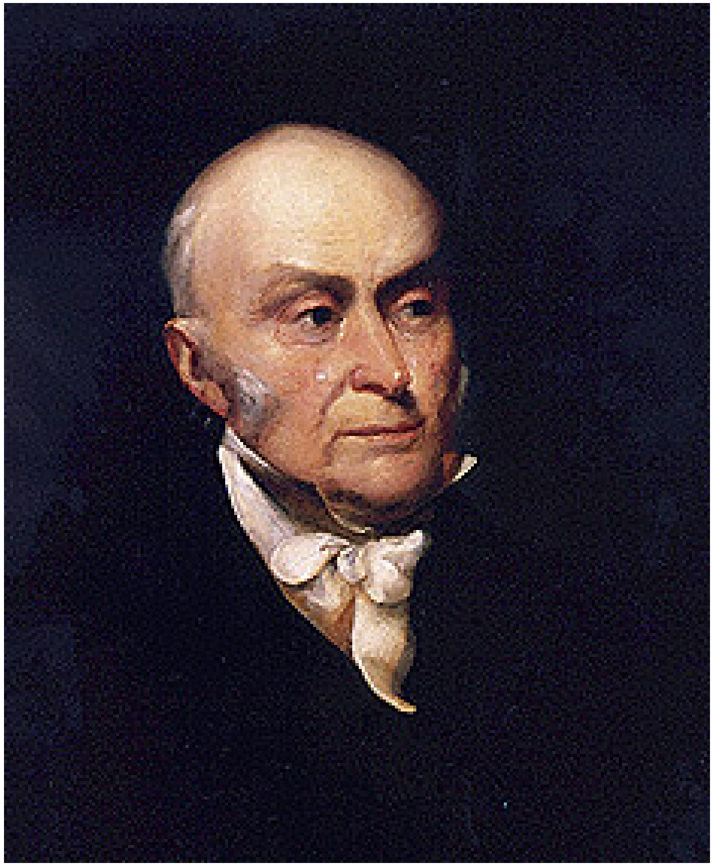
GIGO in the 1840's

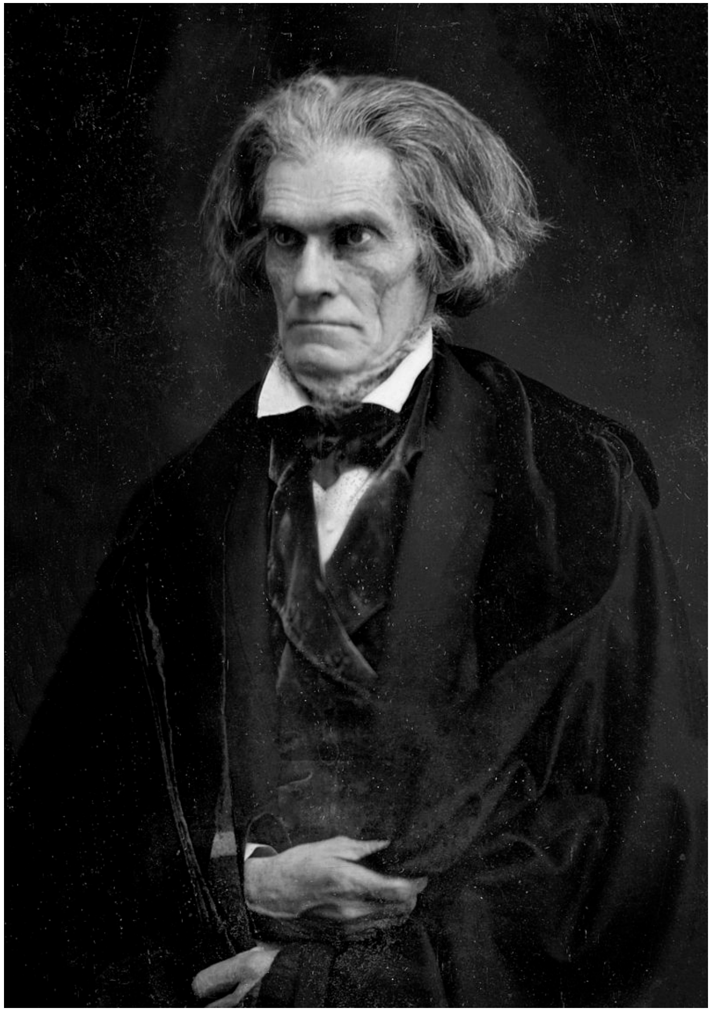
"Atrocious misrepresentations have been made on a subject of deep importance."

– John Quincy Adams
U.S. Representative from Massachusetts
(and former President)



Who Are Going to Believe?





Calhoun’s Response to Errors in Data

“there were so many errors they balanced one another, and led to the same conclusion as if they were all correct.”

Was it the case that the measurement errors are unbiased and independent of each of other, and therefore almost identically distributed on either side of the mean?

No, later analysis showed that the errors were not random but systematic.

“it was the census that was insane and not the colored people.”—James Freeman Clarke

Moral: Analysis of bad data can lead to dangerous conclusions.