Lecture 13: Statistical Sins

(download slides and .py files from Stellar to follow along)

John Guttag

MIT Department of Electrical Engineering and Computer Science

Announcements

- Course evaluations still open
  - http://web.mit.edu/subjectevaluation/

There are Three Kinds of Lies

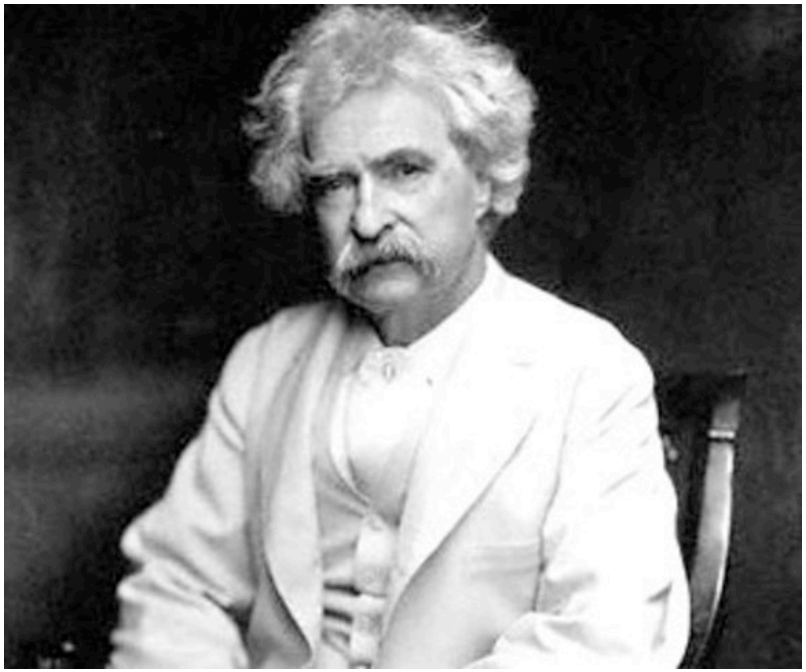
LIES

DAMNED LIES

and

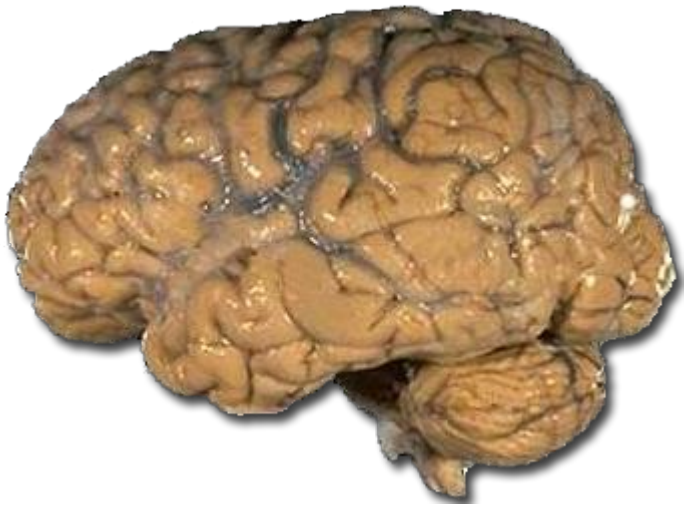
STATISTICS

Attributed To



Humans and Statistics

Human Mind



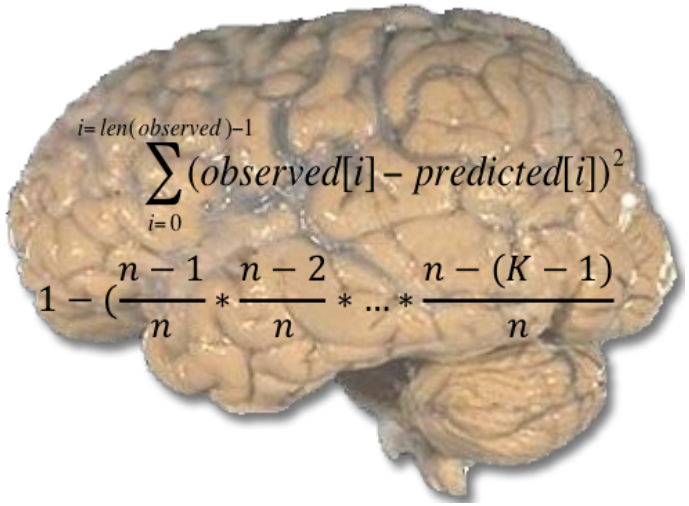
Statistics

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

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

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*



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

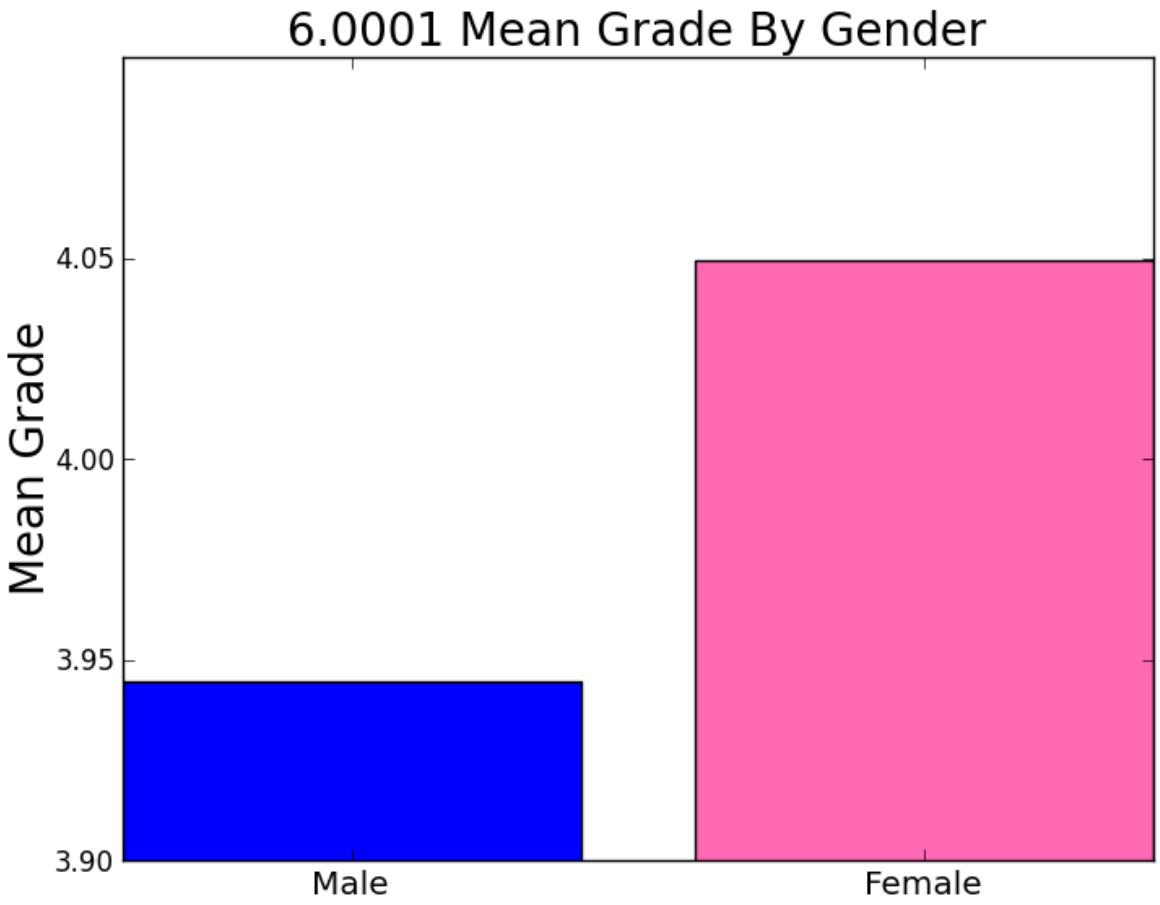
- Summary statistics for groups identical
  - Mean x = 9.0
  - Mean y = 7.5
  - Variance of x = 10.0
  - Variance of y = 3.75
  - Linear regression model: y = 0.5x + 3
- Are four data sets really similar?

Let's Plot the Data

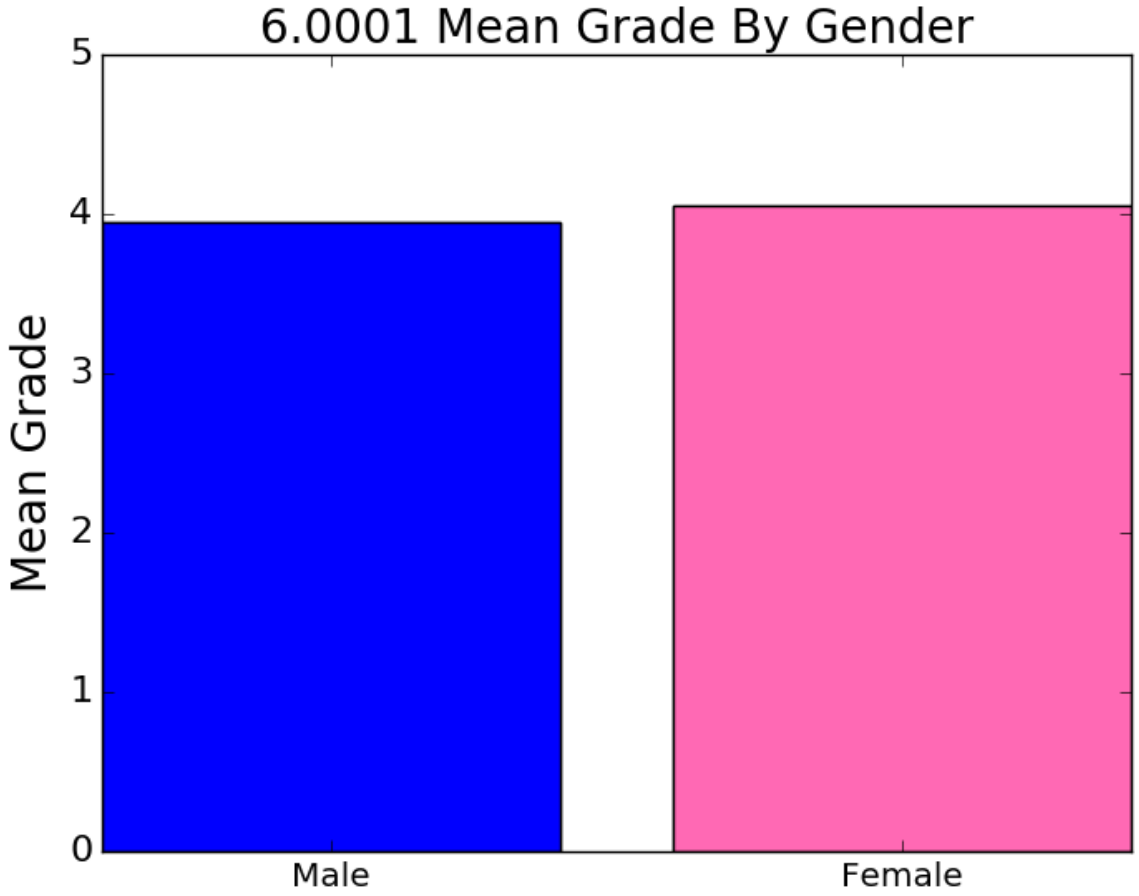
Moral: Statistics about the data is not the same as the data

Moral: Use visualization tools to look at the data itself

Lying with Pictures



Telling the Truth with Pictures

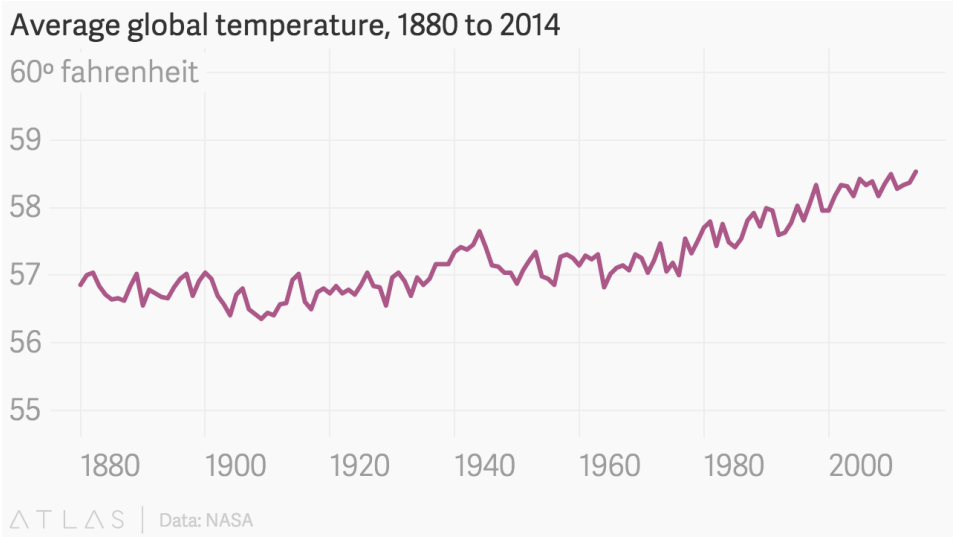


Moral: Look carefully at the axes labels and scales

Lying with Pictures

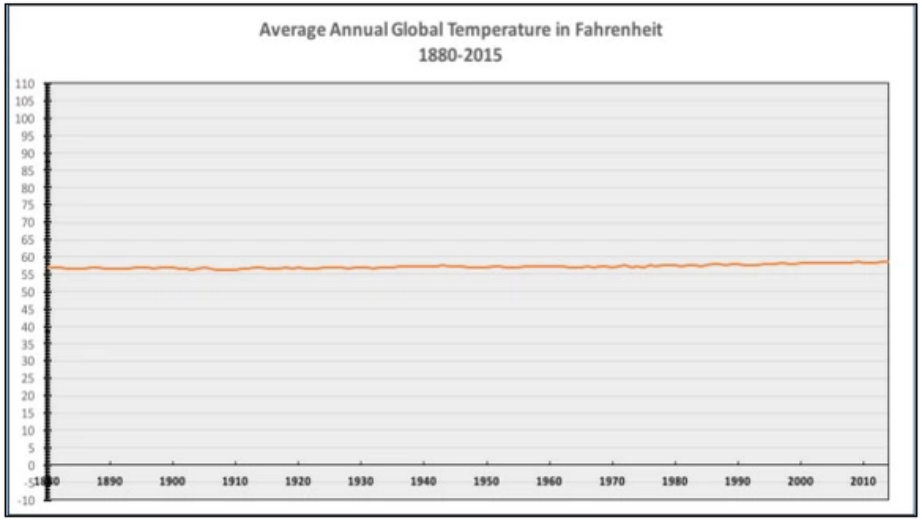
Moral: Ask whether the things being compared are actually comparable

Should the y-Axis Always Start at 0?



Which conveys a more accurate impression?

Moral: truncate the y-axis to remove preposterous values



National Review, December 2015

GIGO

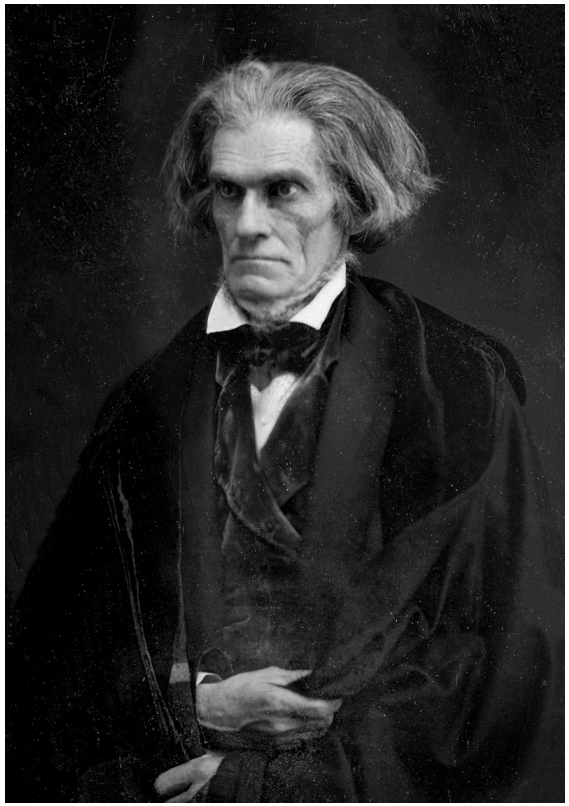
Garbage In, Garbage Out

"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."

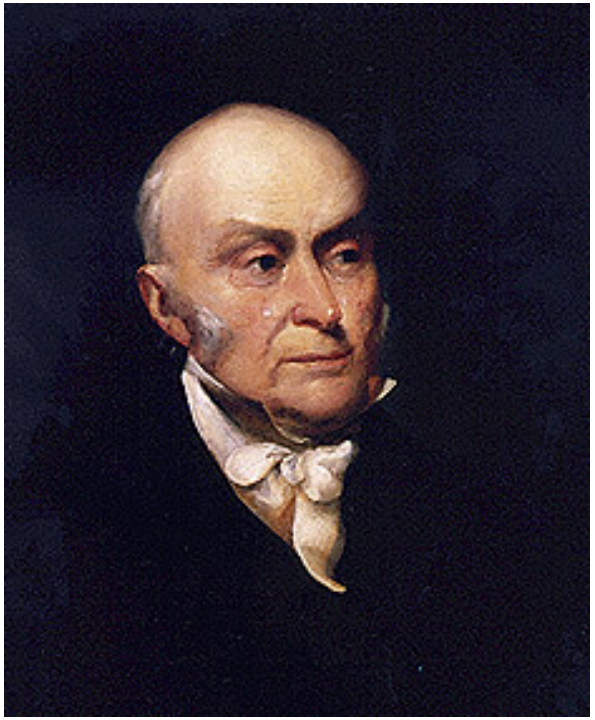
– John C. Calhoun
U.S. Secretary of State



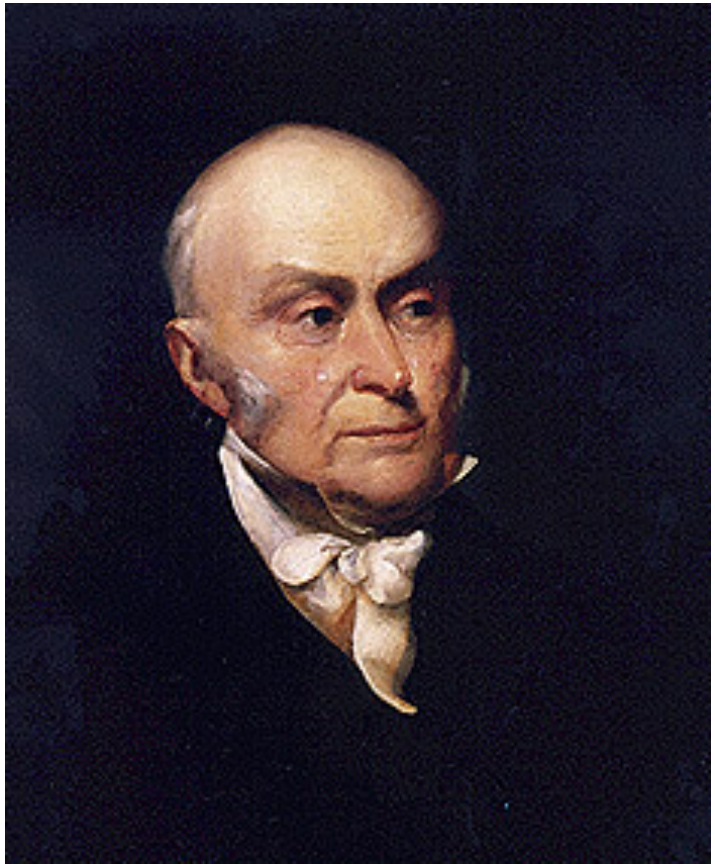
GIGO in the 1840's

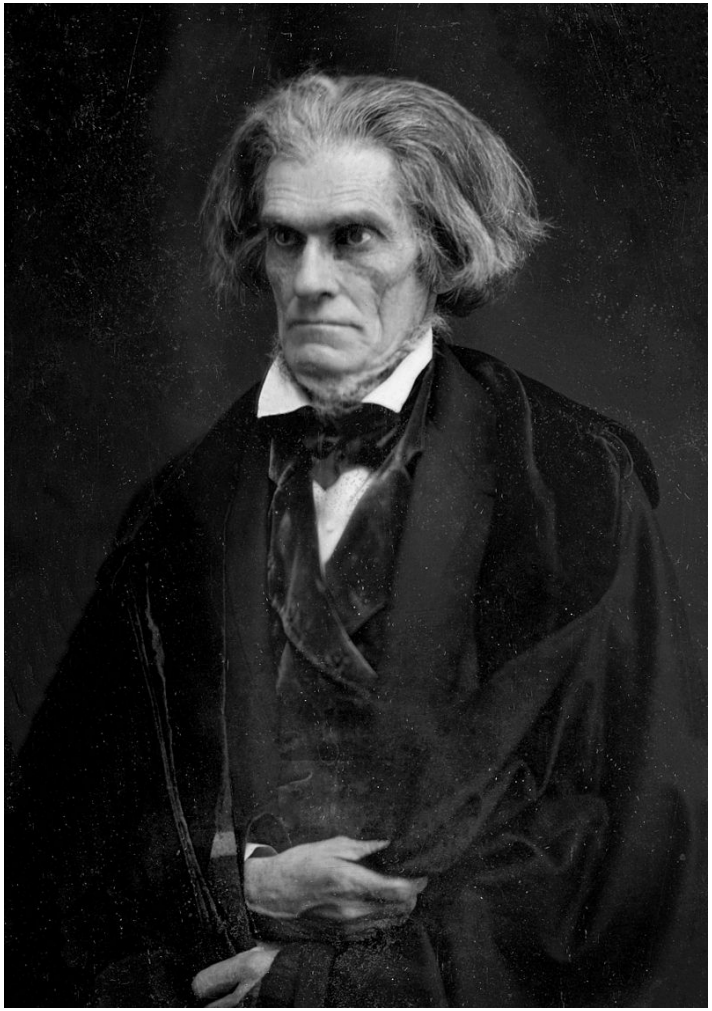
"Atrocious misrepresentations have been made on a subject of deep importance."

– 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."

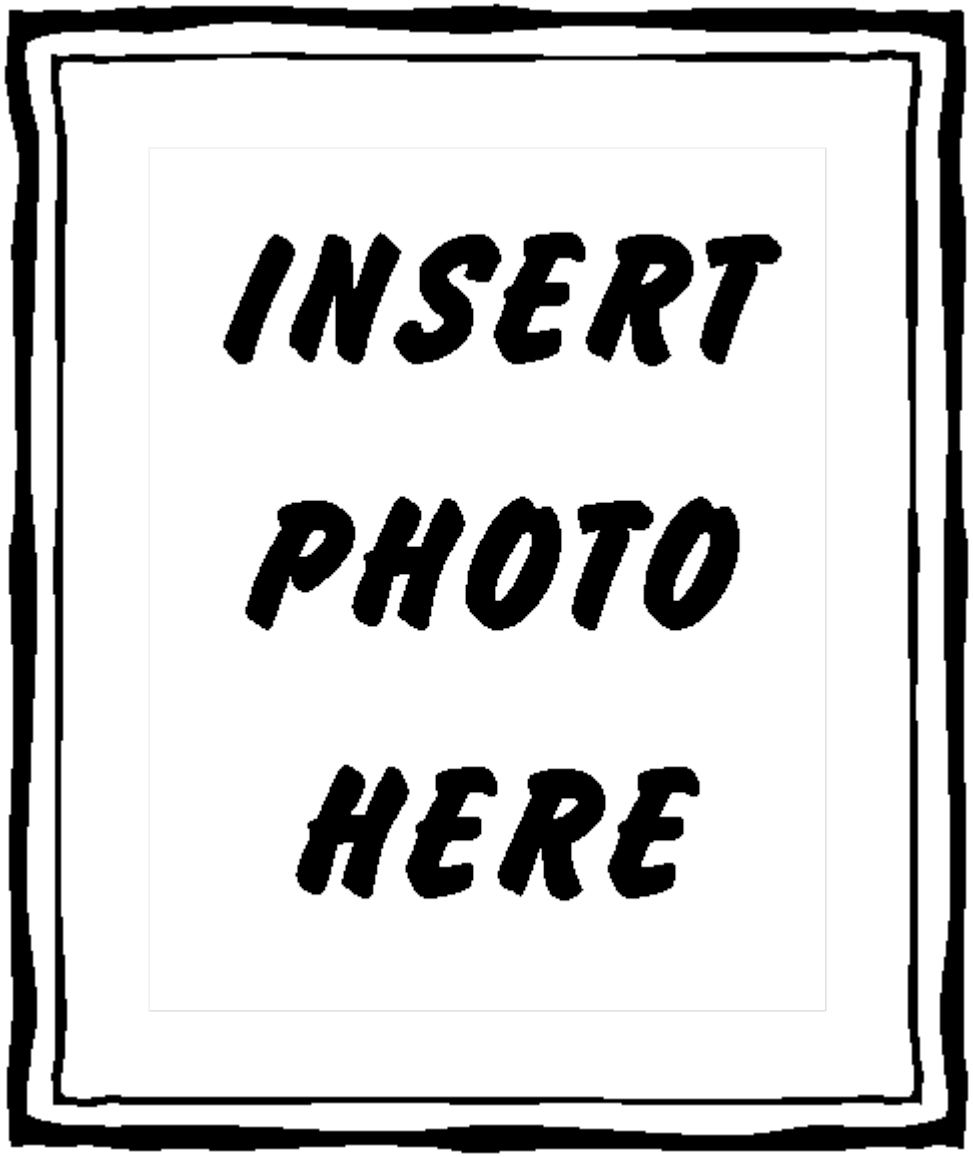
Was it the case that the measurement errors are unbiased and independent of each of other, and therefore almost identically distributed on either side of the mean?

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.

A Thing of the Past?



A WW II Fighter Plane

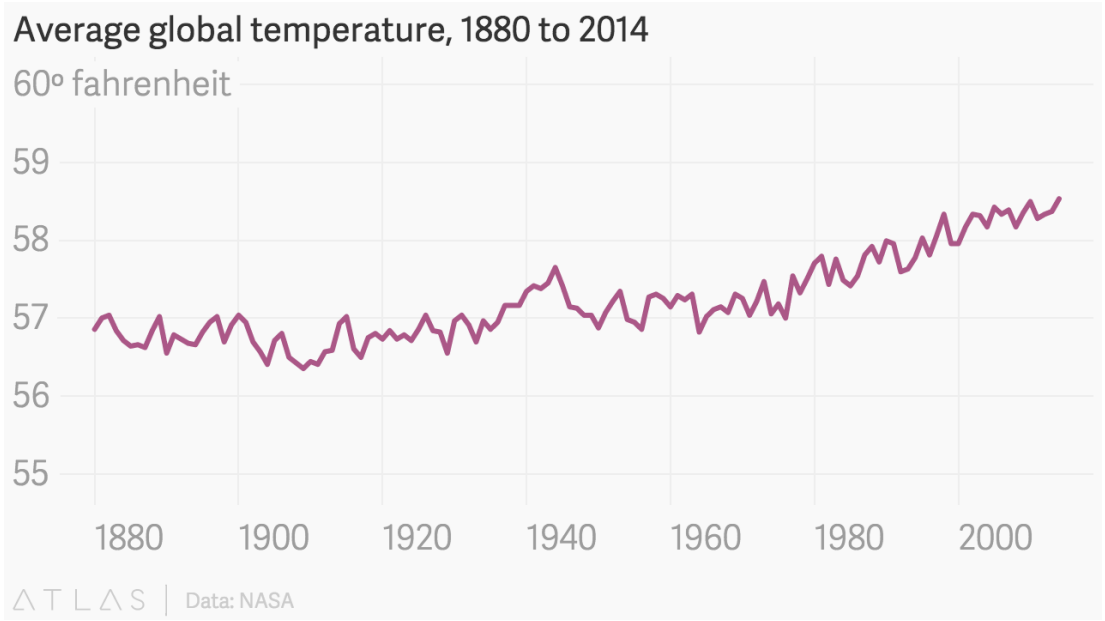
Sampling

- All statistical techniques are based upon the assumption that by sampling a subset of a population we can infer things about the population as a whole
- As we have seen, *if random sampling is used*, one can make meaningful mathematical statements about the expected relation of the sample to the entire population
- Easy to get random samples in simulations
- Not so easy in the field, where some examples are more convenient to acquire than others

Non-representative Sampling

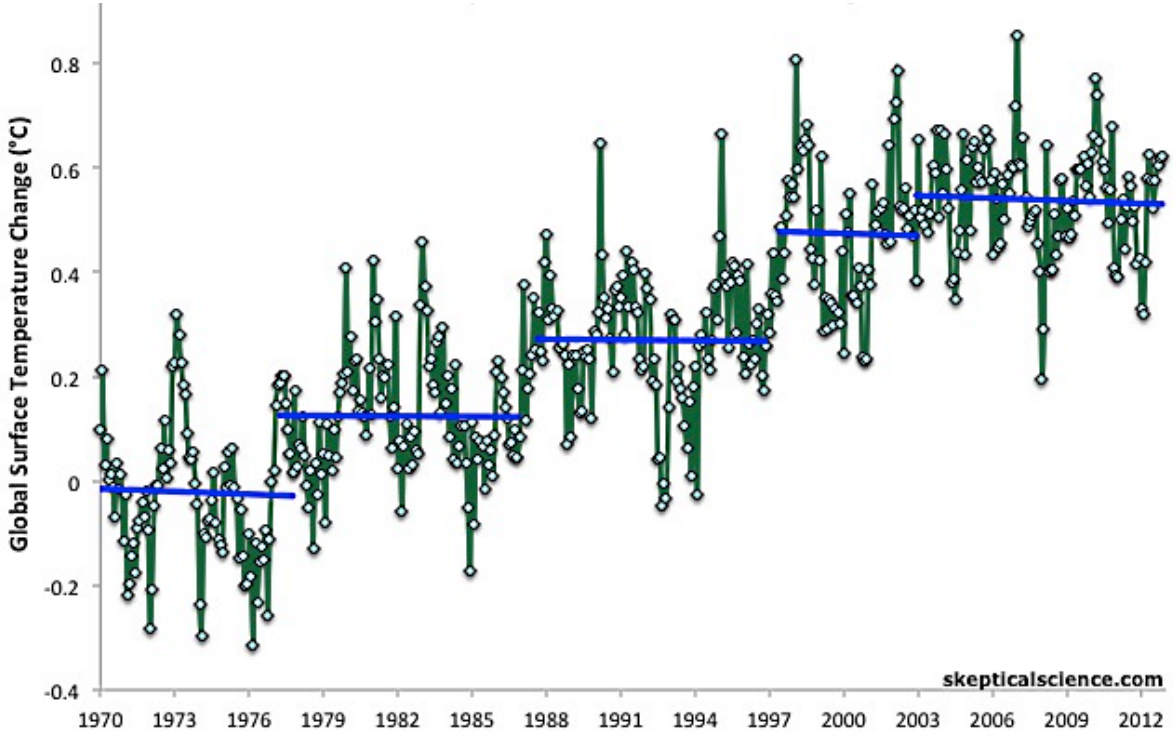
- "Convenience sampling" not usually random, e.g.,
  - Survivor bias, e.g., course evaluations at end of course or grading final exam in 6.0002 on a strict curve
  - Non-response bias, e.g., opinion polls conducted by mail or online
- When samples not random and independent, we can still do things like compute means and standard deviations, but **we should not draw conclusions from them** using things like the empirical rule and central limit theorem.
- Moral: Understand how data was collected, and whether assumptions used in the analysis are satisfied. If not, be wary.

The Myth of Global Warming



Chose an interval consistent with phenomenon being considered.

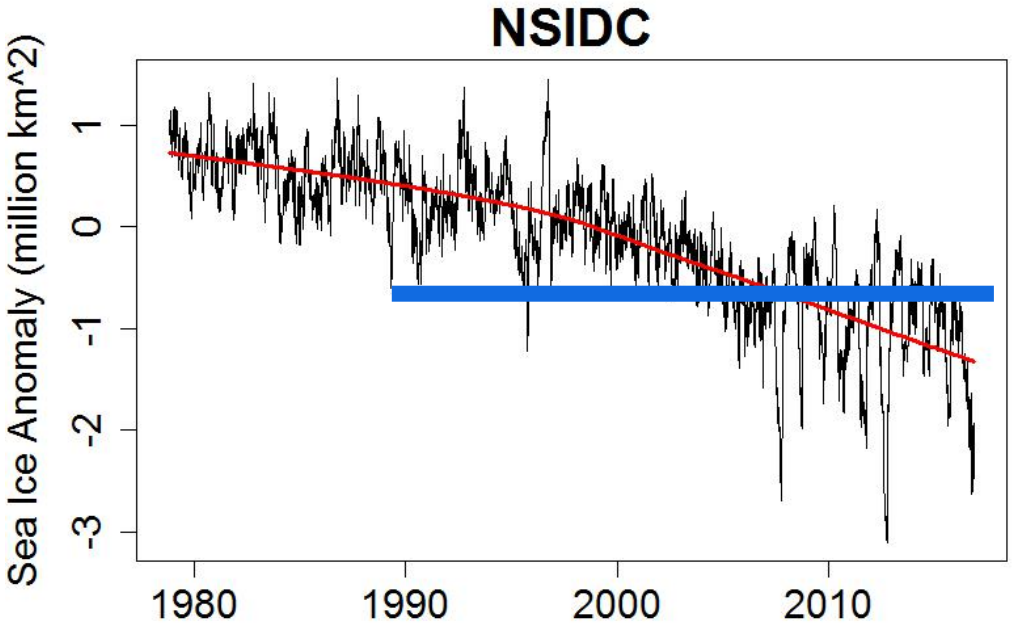
The Myth of Global Warming



Moral: Don't confuse fluctuations with trends.

But At Least the Arctic Ice Isn't Melting

"Yesterday, April 14th, the Arctic had more sea ice than it had on April 14, 1989 – 14.511 million square kilometres vs 14.510 million square kilometres, according to the National Snow and Ice Data Center of the United States, an official source." *Lawrence Solomon, Financial Post, April 15, 2013*



Moral

A Suitable Holiday Gift?

A Comforting Statistic

- 99.8% of the firearms in the U.S. will not be used to commit a violent crime in any given year
- How many privately owned firearms in U.S.?
- ~300,000,000
- 300,000,000*0.002 = 600,000

A Not So Comforting Statistic

"Mexican health officials suspect that the swine flu outbreak has caused more than 159 deaths and roughly 2,500 illnesses." CNN, April 29, 2009

How many deaths per year from seasonal flu in U.S.?

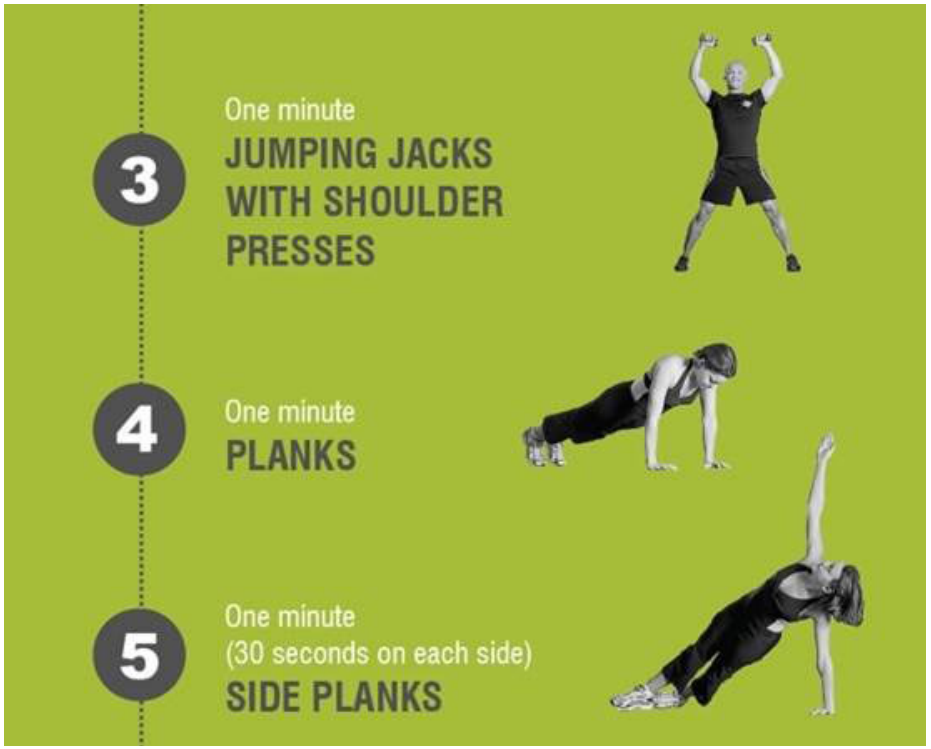
About 36,000

Moral: Context Matters!

Relative to What?

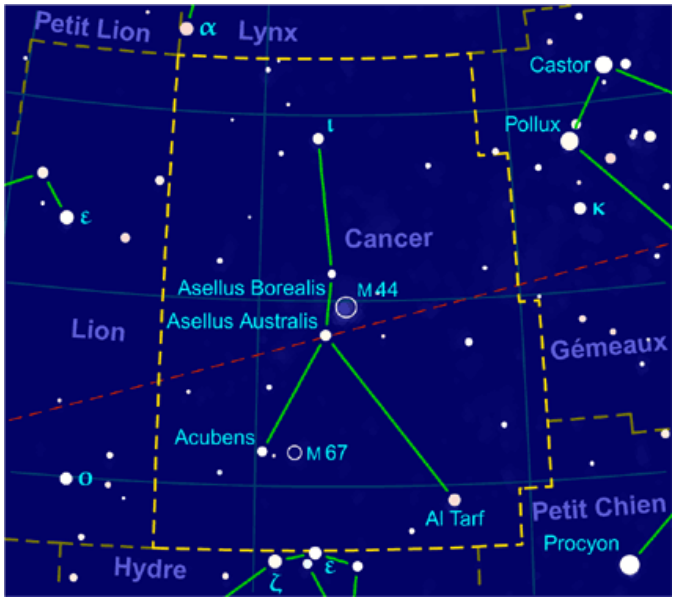
- Skipping lectures increases your probability of failing 6.0002 by 50%
- From 0.5 to 0.75?
- From 0.005 to 0.0075?
- Moral: Beware of percentage change when you don't know the denominator

Five Minute Break



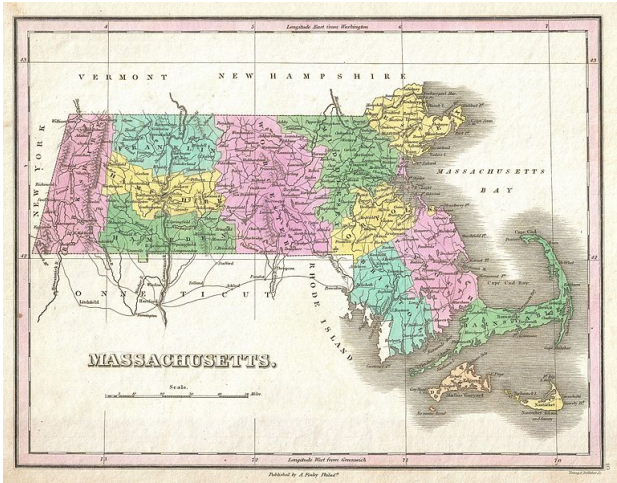
Cancer Clusters

- A **cancer cluster** is defined by the CDC as "a greater-than-expected number of cancer cases that occurs within a group of people in a geographic area over a period of time"
- About 1000 "cancer clusters" per year are reported to health authorities in the U.S.
- Vast majority, but not all, are deemed not significant



A Hypothetical Example

- Massachusetts is about 10,000 square miles
- About 36,000 new cancer cases per year
- An attorney partitioned state into 1000 regions of 10 squares miles each, and looked at distribution of cases
  - Expected number of cases per year per region: 36
- Discovered that region 111 had 30% more new cancer cases than expected over a 3 year period!
- How worried should residents be?



https://commons.wikimedia.org/wiki/File:1827_Finley_Map_of_Massachusetts_-_Geographicus_-_Massachusetts-finley-1827.jpg

How Likely Is it Just Bad Luck?

```
def findProb(numCommunities, community,
             casesPerTrial, multiple, numTrials):
    numTimesOver = 0
    threshhold = (casesPerTrial/numCommunities)*multiple
    for t in range(numTrials):
        cases = [0]*numCommunities
        for i in range(casesPerTrial):
            cases[random.choice(range(numCommunities))] += 1
        if cases[community] > threshhold:
            numTimesOver += 1
    prob = round(numTimesOver/numTrials, 4)
    print('Est. prob. of it being a random event =', prob)

#Initialize constants
numCasesPerYear = 36000
numYears = 3
stateSize = 10000
communitySize = 10
numCommunities = stateSize//communitySize
multiple = 1.3

casesPerTrial = numCasesPerYear*numYears
findProb(numCommunities, 111, casesPerTrial, multiple, 100)
```

Try 10,000 Trials

- Estimated probability: 0.0005 !
- Seems like time to check the water supply and air quality in area 111?

But, Are We Asking the Right Question?

A variant of cherry picking called
multiple hypothesis testing

The Texas Sharpshooter

(modified) CC-BY Image Courtesy of Putneypics

Machine Learning Can Make It All Worse

- Possible to commit the usual statistical sins
  - Plus some new ones
- People are trusting black-box models
  - Assume data is valid
  - Assume distribution of training data is representative of test data
  - Assume that labels capture what we actually care about
  - …
- Imagine Robo Judge
  - A model to predict whether somebody released from prison will commit a crime

Hypothetical Scenario

- Data
  - Features: Lots of descriptive information about thousands of people who have been released from jail
  - Label: Whether or not they were they convicted of a crime after they were released
- Logistic regression model: Estimates probability of label given feature vector
- Usage: Use probability to decide whether or not an individual should be paroled

What are the kinds of questions that need to be asked?

Some Questions

- Is the training population representative of all of the test populations to which it will be applied
  - E.g., if trained using data acquired during a recession, should it be used during an economic boom?
- Are we predicting the appropriate thing>
  - Labels capture who was convicted of a crime, not who committed one
    - Some sub-populations more likely to be arrested
    - Some sub-populations more likely to be convicted

Labels are often proxies for things we really want to predict

The Bottom Line

- When drawing inferences from data, skepticism is merited.
- But remember, skepticism and denial are different.
- "Doubt, indulged and cherished, is in danger of becoming denial; but if honest, and bent on thorough investigation, it may soon lead to full establishment of the truth." – Ambrose Bierce

Parting Words About 6.0002

Four Ways to Study the World

- Theory
  - Mathematical thinking and skills
- Physical experiments
  - Experimental design and skills
- Computational experiments
  - Computational thinking and skills
- Data-intensive
  - Computer science and statistics

6.0002 Themes

- Computational experiments
  - Modeling
  - Algorithmic thinking
  - Testing and debugging
  - Presenting and evaluating results
- Extracting information from data
  - Visualizing data
  - Manipulating data
  - Statistical analysis
  - Machine learning

6.0002 Major Topics

- Optimization problems
- Stochastic thinking
- Modeling aspects of the world
- Becoming a better programmer
  - Exposure to a few extra features of Python and some useful libraries
  - Practice, practice, practice

Optimization Problems

- Many problems can be formulated in terms of
  - Objective function
  - Set of constraints
- Greedy algorithms often useful
  - But may not find optimal solution
- Many optimization problems inherently exponential
  - But dynamic programming often works
  - And memoization a generally useful technique
- Examples: knapsack problems, graph problems, curve fitting, clustering

Stochastic Thinking

- The world is (predictably) non-deterministic
- Thinking in terms of probabilities is often useful
- Randomness is a powerful tool for building computations that model the world
- Random computations useful even when for problems that do not involve randomness
  - E.g., integration

Modeling the World

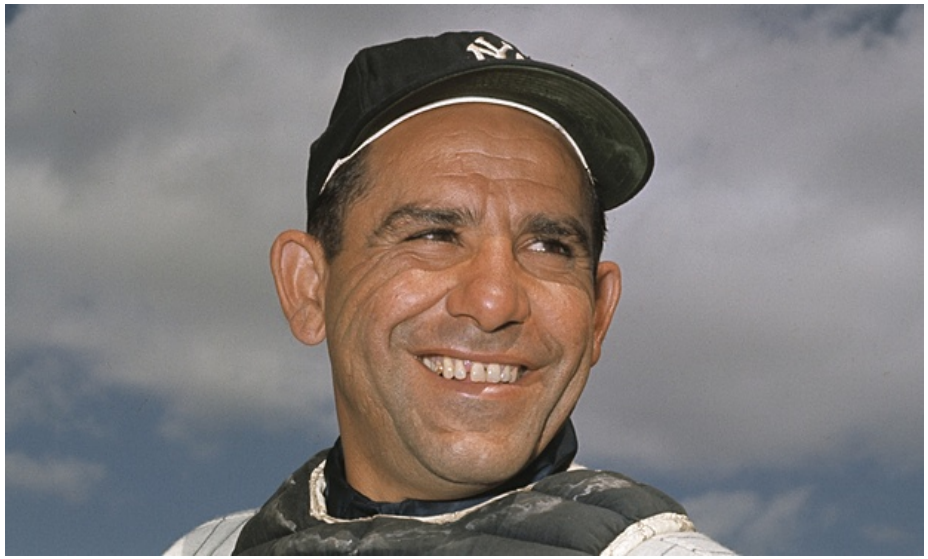
- Models always inaccurate
  - Provide abstractions of reality
- Deterministic models, e.g., graph theoretic
- Statistical models
  - Simulation models: Monte Carlo simulation
  - Models based on sampling
    - Characterizing accuracy is critical
      - Central limit theorem
      - Empirical rule
    - Machine learning
      - Unsupervised and supervised
- Making sense of data
  - Plotting
  - Good and bad practices

What's Next for You?

- Many of you have worked very hard
  - Rest of the staff and I appreciate it
- Only you know your return on investment
  - Take a look at early problem sets (esp. if you were in 6.0001)
  - Think about what you'd be willing tackle now
- Remember that you can write programs to get answers
- There are other CS courses you are prepared to take
  - 6.009, 6.005, 6.0006, 6.036
- Find an interesting UROP
- Minor in CS
- Major in 6.3, 6.2, 6-7, 6-14, , 11-6, 18C

Has CS Peaked?

- **"It's tough to make predictions, especially about the future." – Yogi Berra**



Predicting the Future of Computing

- "I think there is a world market for maybe five computers." - Thomas Watson, chairman of IBM, 1943

- "Computers in the future may weigh no more than 1.5 tons." - *Popular Mechanics*, forecasting the relentless march of science, 1949

Predicting the Future of Computing

- "I have traveled the length and breadth of this country and talked with the best people, and I can assure you that data processing is a fad that won't last out the year." - The editor in charge of business books for Prentice Hall, 1957

- "There is no reason anyone would want a computer in their home." - Ken Olson, president, chairman & founder of Digital Equipment Corp., 1977

MIT Thinks the Best Is Yet to Come