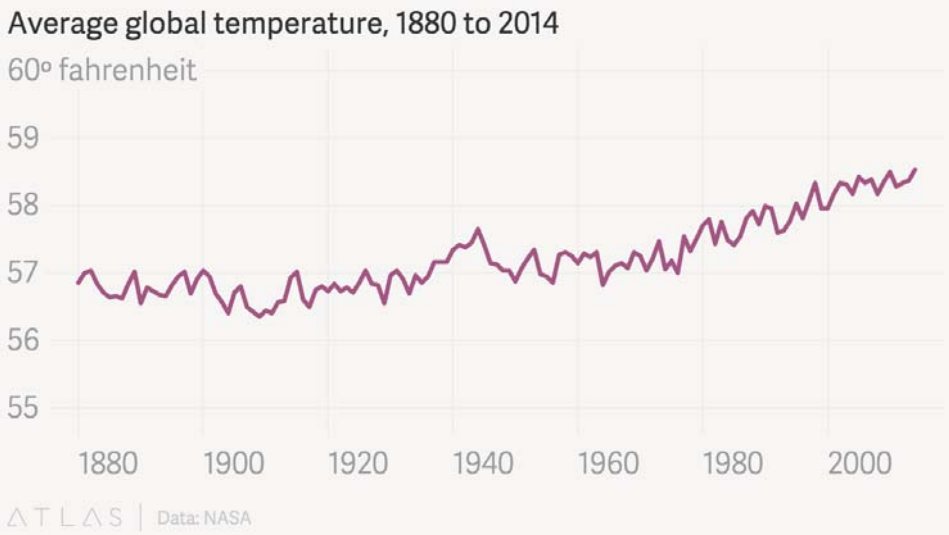
Lecture 15: Statistical Sins and Wrapup

Announcements

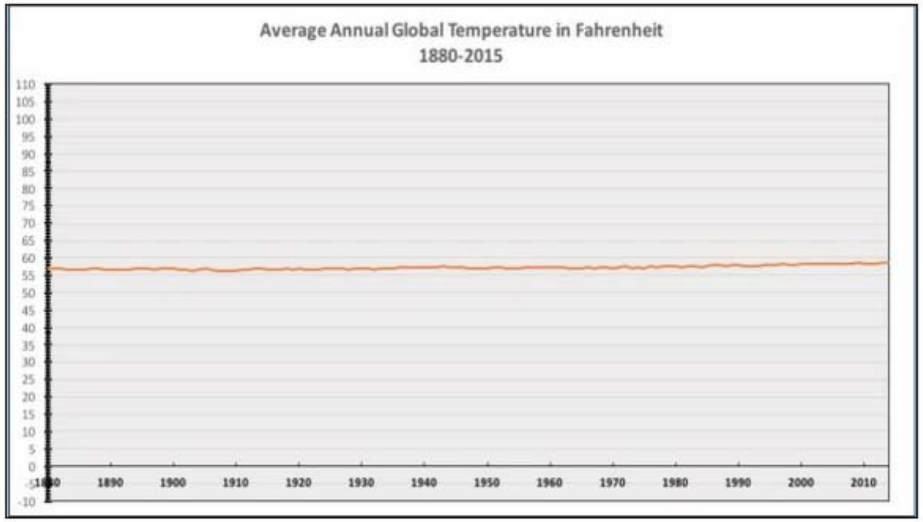
- Course evaluations
  - Online evaluation now through noon on Friday, December 16
- Final exam on Monday!

Global Warming, Fact or Fiction



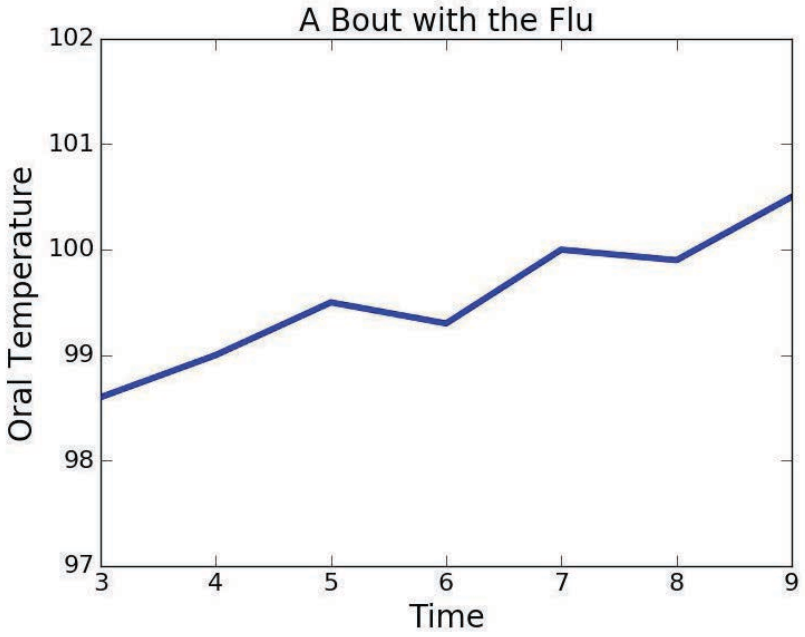
On Monday I said, beware of charts where the y-axis doesn't start at zero

Which conveys a more accurate impression?

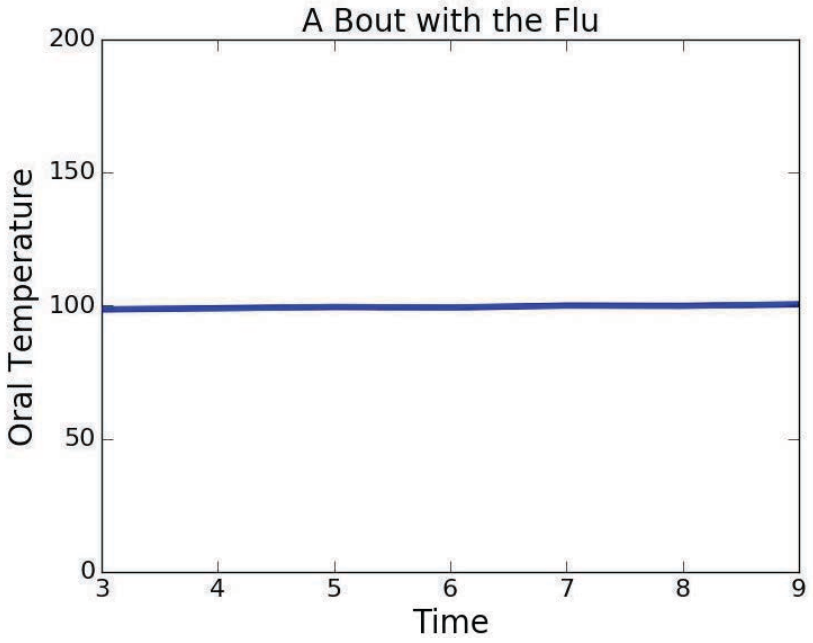


National Review, December 2015

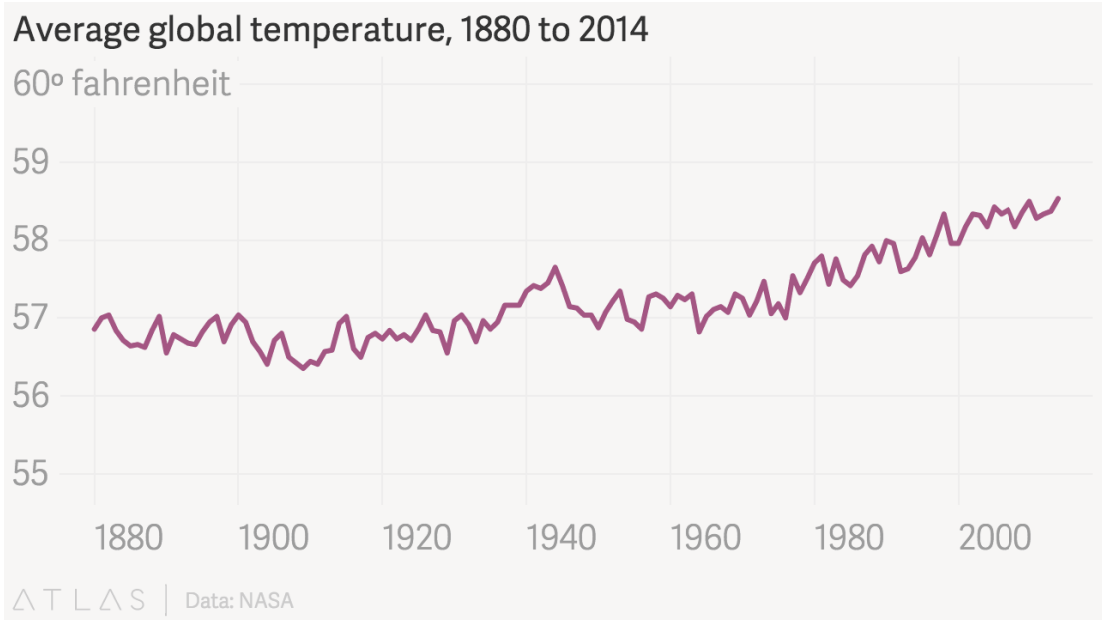
Fever and the Flu



Moral: Truncate the y-axis to eliminate preposterous values.



The Myth of Global Warming

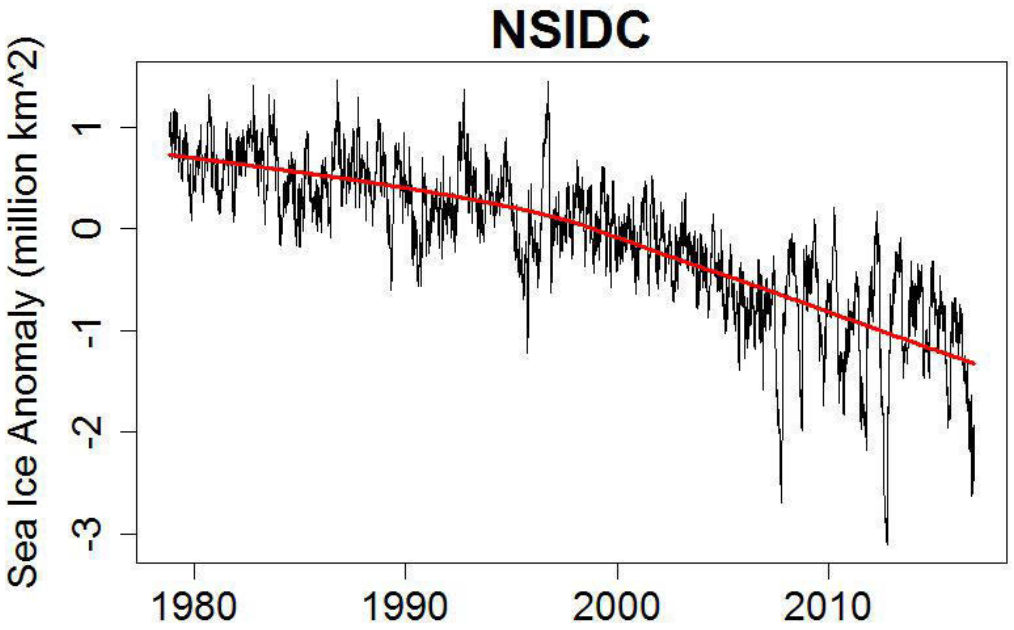


Moral: Don't confuse fluctuations with trends.

Chose an interval consistent with phenomenon being considered.

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 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

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

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 are deemed not significant

A Hypothetical Example

- Massachusetts is about 10,000 square miles
- About 36,000 new cancer cases per year
- Attorney partitioned state into 1000 regions of 10 squares miles each, and looked at distribution of cases
  - Expected number of cases per region: 36
- Discovered that region 111 had 143 new cancer cases over a 3 year period!
  - More than 32% greater than expected
- How worried should residents be?

How Likely Is it Just Bad Luck?

```
numCasesPerYear = 36000
numYears = 3
stateSize = 10000
communitySize = 10
numCommunities = stateSize//communitySize

numTrials = 100
numGreater = 0
for t in range(numTrials):
    locs = [0]*numCommunities
    for i in range(numYears*numCasesPerYear):
        locs[random.choice(range(numCommunities))] += 1
    if locs[111] >= 143:
        numGreater += 1
prob = round(numGreater/numTrials, 4)
print('Est. probability of region 111 having\
 at least 143 cases =', prob)
```

How Likely Is it Just Bad Luck?

```
anyRegion = 0
for trial in range(numTrials):
    locs = [0]*numCommunities
    for i in range(numYears*numCasesPerYear):
        locs[random.choice(range(numCommunities))] += 1
    if max(locs) >= 143:
        anyRegion += 1
print(anyRegion)
aProb = round(anyRegion/numTrials, 4)
print('Est. probability of some region  having\
 at least 143 cases =', aProb)
```

A variant of cherry picking called

multiple hypothesis testing

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

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
- Presentation 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
  - 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, 6.006, 6.034
- Find and interesting UROP
- Minor in CS
- Major in CS

6.0002 Introduction to Computational Thinking and Data Science
Fall 2016