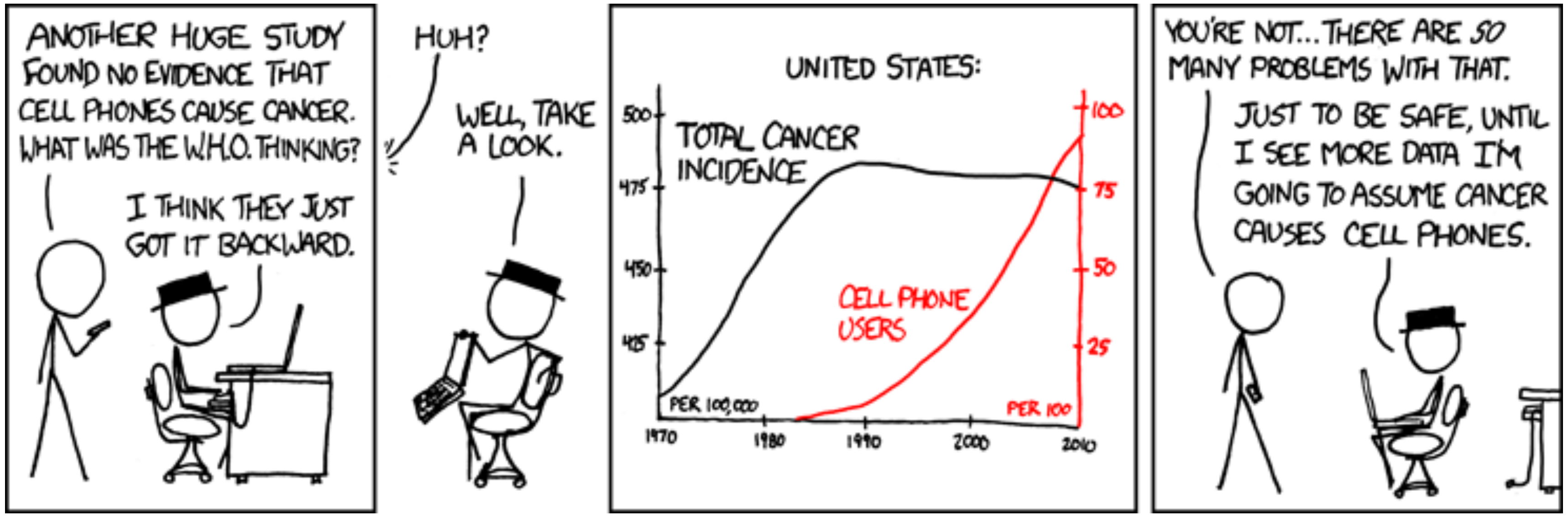
Lecture 8: Statistics

COSC 480 Data Science, Spring 2017
Michael Hay

Review course schedule

Disclaimer

- Three lectures on concepts related to statistics
  - Today: (descriptive) statistics
  - Next week: probability, hypothesis testing
- Not a substitute for an entire course (or two)!
- Goal: Illustrate some basic concepts, potential power and avoid common pitfalls


ANOTHER HUGE STUDY FOUND NO EVIDENCE THAT CELL PHONES CAUSE CANCER. WHAT WAS THE W.H.O. THINKING?
I THINK THEY JUST GOT IT BACKWARD.
HUH?
WELL, TAKE A LOOK.
UNITED STATES:
TOTAL CANCER INCIDENCE
CELL PHONE USERS
500
475
450
425
PER 100,000
100
75
50
25
PER 100
1970
1980
1990
2000
2010
YOU'RE NOT... THERE ARE SO MANY PROBLEMS WITH THAT.
JUST TO BE SAFE, UNTIL I SEE MORE DATA I'M GOING TO ASSUME CANCER CAUSES CELL PHONES.

https://xkcd.com/925/

Single variable

Say you are buying a house…

- Your agent could tell you, with perfect "honesty," that the "average" annual income in the neighborhood is
  - $150,000
  - $35,000
  - $10,000

http://yourfinancialblog.com/wp-content/uploads/2013/01/buy-house.jpg

Example from Rick Ord of UCSD: http://cseweb.ucsd.edu/~ricko/CSE3/Lie_with_Statistics.pdf
Credit: slide adapted with permission from Ashwin Machanavajjhalla and Jun Yang, Duke University

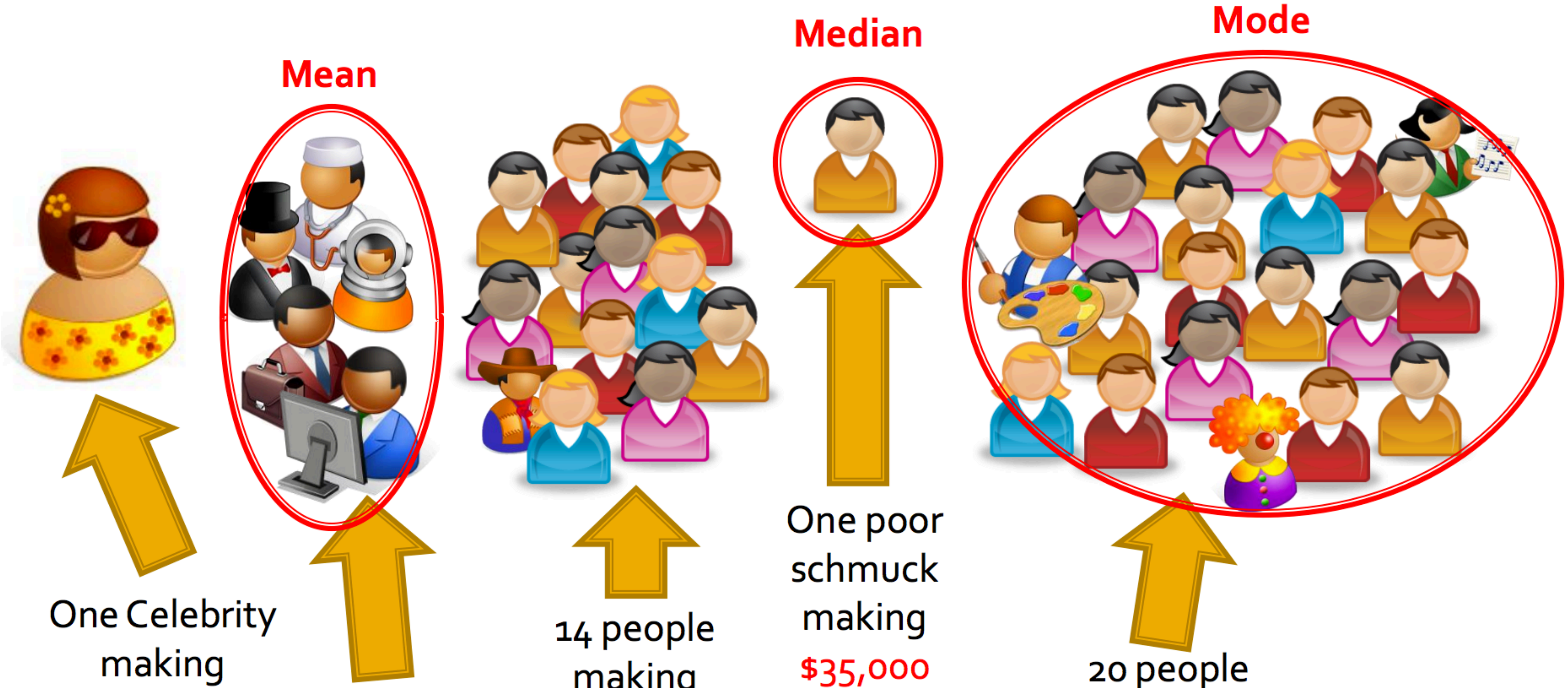
Measures of central tendency

- Collection a collection of data values x1, …, xN
- **Mean**: arithmetic average (sum / count)
- **Median**: the "middle" number when all values are sorted
- **Mode**: the most common value

So, *which is which*?

- $10,000
- $35,000
- $150,000

The reality is…



Mean

Median

Mode

One Celebrity
making
$4,465,000

Five
Professionals
making
$150,000

14 people
making
$50,000

One poor
schmuck
making
$35,000

20 people
making
$10,000

http://cseweb.ucsd.edu/~ricko/CSE3/Lie_with_Statistics.pdf

Exercise

Instructions: ~1 minute to think/answer on your own; then discuss with neighbors; then I will call on one of you

- Suppose you have some dataset of ages. All ages are constrained to be between [0, 100]. There are N ages in the dataset.
- If I were to change a single age, how much could it change each statistic (mean, median, mode)?
- Consider the *worst case* (largest change). So, you can choose the dataset and the modification to maximize change.

Exercise

Instructions: ~1 minute to think/answer on your own; then discuss with neighbors; then I will call on one of you

- Suppose you have some dataset of ages. All ages are constrained to be between [0, 100]. There are N ages in the dataset.

- If I were to change a single age, how much could it change each statistic (mean, median, mode)?

- Consider the *worst case* (largest change). So, you can choose the dataset and the modification to maximize change.

Answer: mean could change by 100/N. Mode and median *could* change by 100 *but only for a pathological dataset where the mode/median is probably not a meaningful statistic.* In general, median is more robust to outliers than mean.

http://www.travelvivi.com/wp-content/uploads/2010/02/surcharges.jpg

Sample of airline passengers, mean(weight) = 155.3

http://www.dailydosemd.com/wp-content/uploads/2012/11/marathon-646.jpg

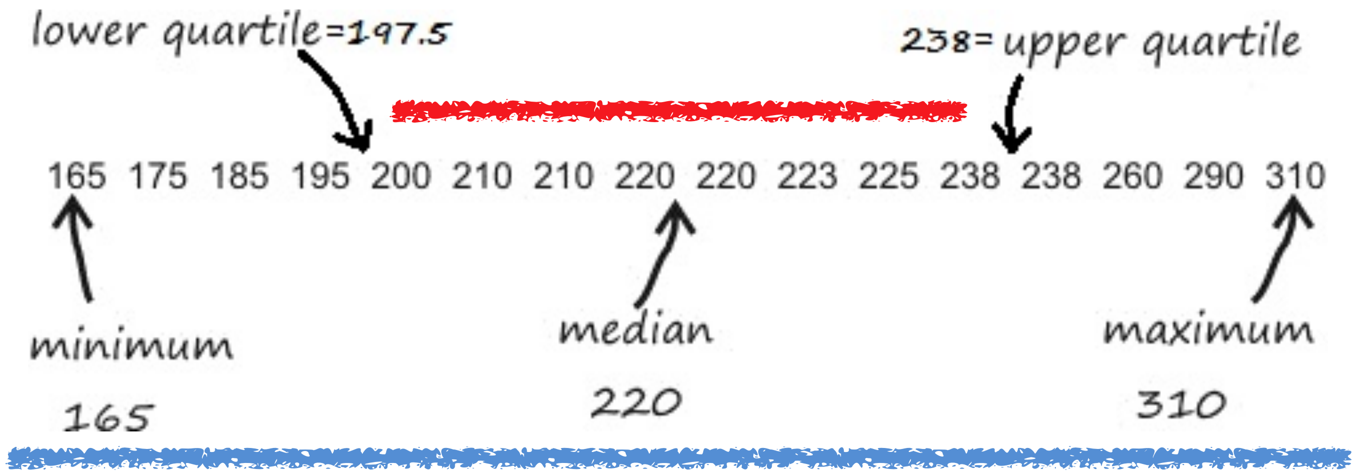
Sample of marathon runners, mean(weight) = 155.5

According to mean, weights are nearly *identical*. **What's missing?**

Stdev in python notebook

Measures of dispersion

- Variance: captures average squared difference from *mean*.
- Standard deviation: square root of variance (why useful?)
- **Interquartile range**: captures spread around *median*.
- **Range**: captures full spread (including outliers)



⬅ 5 number summary

Student A scored 50 points below mean on exam.

Student B scored 50 points above mean on exam.

Student A is dismayed, student B overjoyed. **Are these reasonable reactions?**

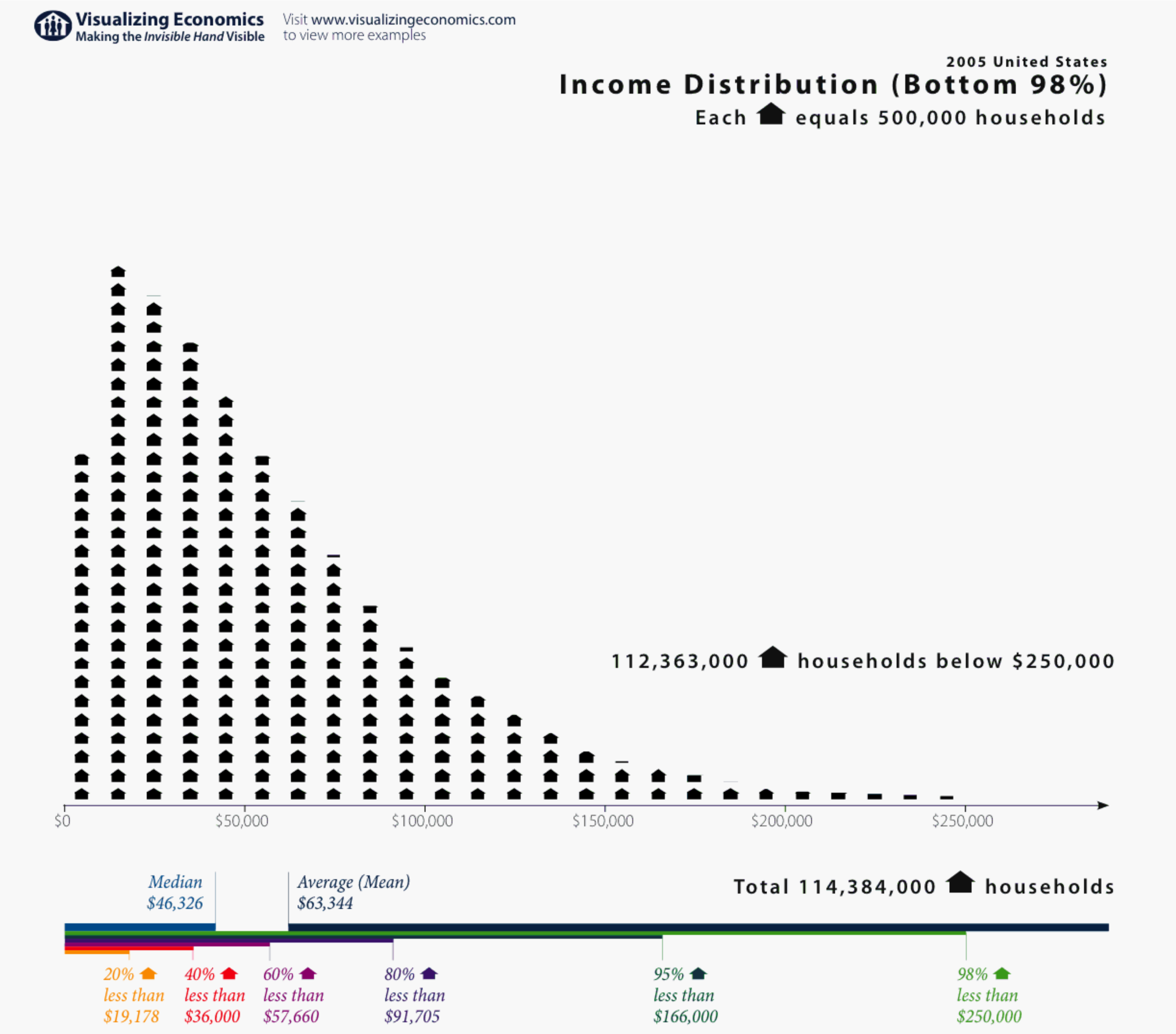
Dance icon by Andrew Doane

Visualizations

- Statistics *summarize* entire dataset into small set of numbers. Information is **lost**.

- Use *visualization to* see entire data distribution.

Histograms



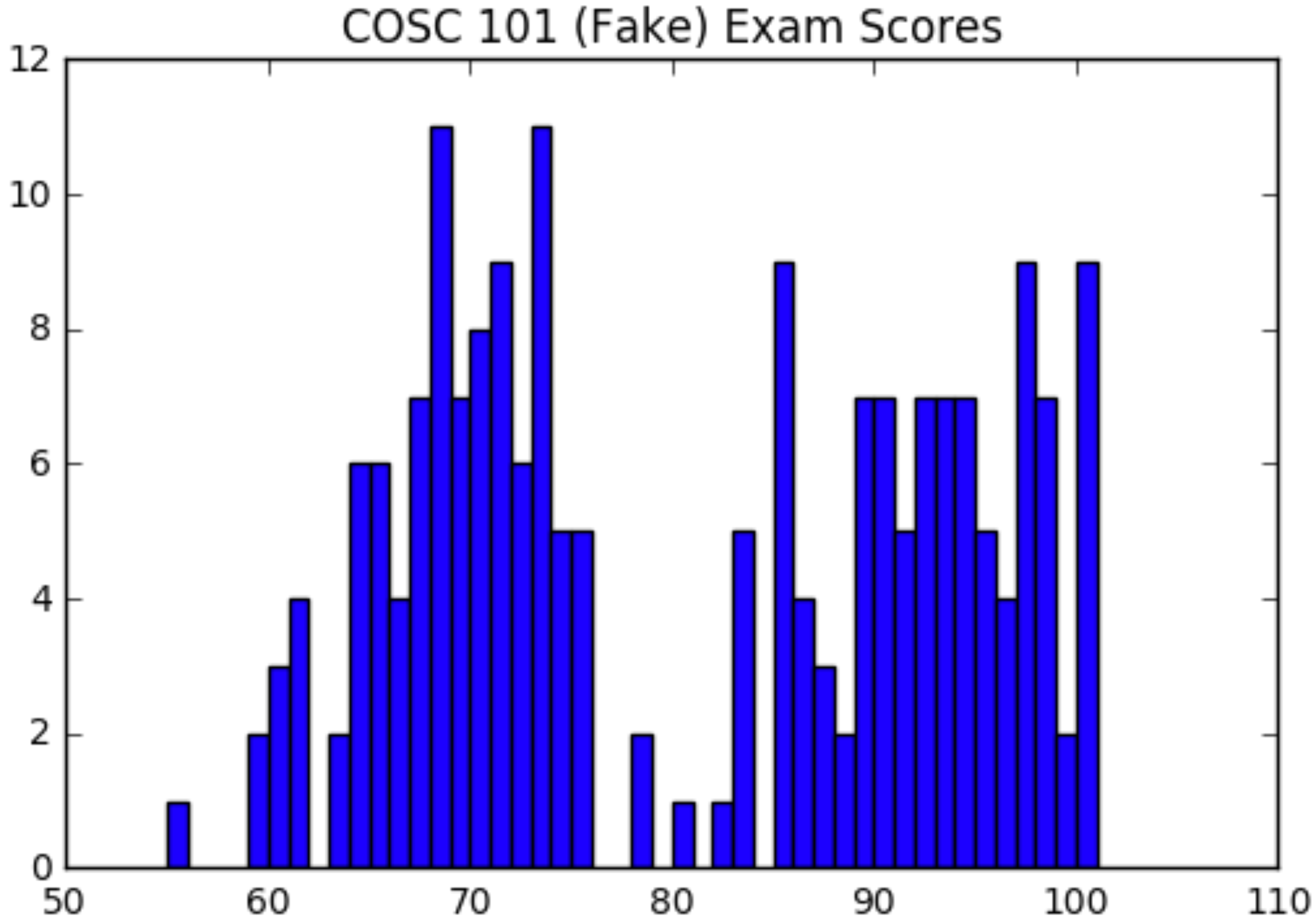
http://visualizingeconomics.com/blog/2006/11/05/2005-us-income-distribution

Visualizations

- Statistics *summarize* entire dataset into small set of numbers. Information is **lost**.

- Use *visualization* as sanity check on statistics.

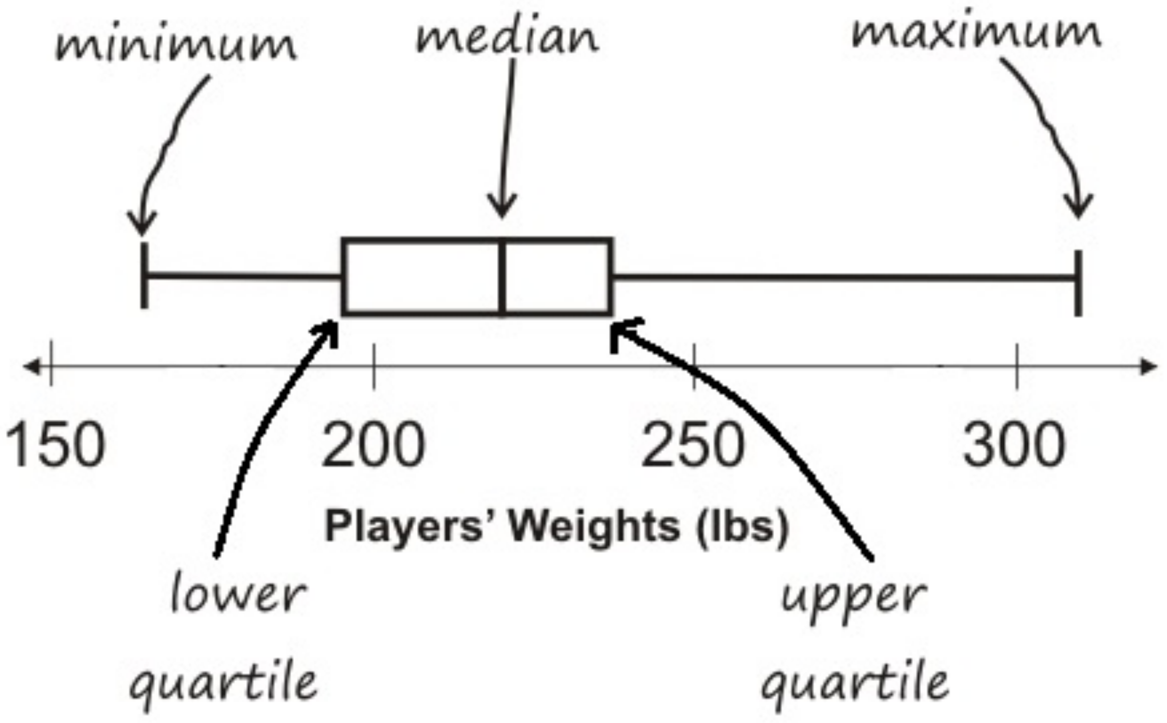
Median exam score = 81



Sanity check: is the median a useful statistic for this data?

Box (and whisker) plots

- A visual display of the 5-number summary.
- *Note*: sometimes whiskers extend only to 5th and 95th percentile.



Demo using notebook

http://illuminations.nctm.org/Lesson.aspx?id=2643

More than one variable

Demo using notebook

Correlation

- Covariance:
  - Whereas variance measures how single variable deviates from its mean
  - Covariance measures how two variables vary *in tandem* from their means.
- Correlation
  - Is covariance after data is *rescaled*.

Exercise

Instructions: ~1 minute to think/answer on your own; then discuss with neighbors; then I will call on one of you

- Complete the analogy:

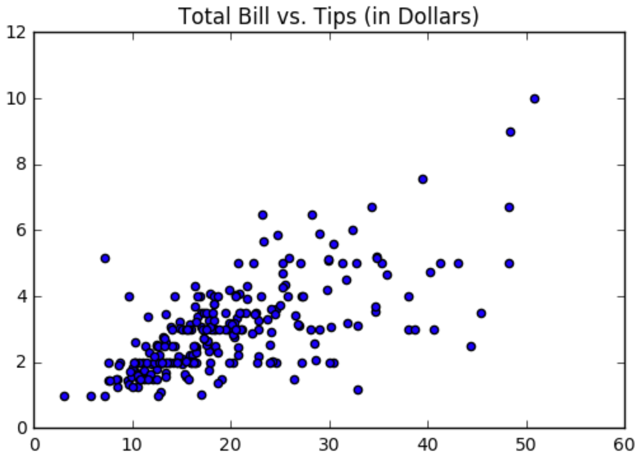
- Mean, standard deviation and correlation are all related. All based on means/averages. Sensitive to outliers.

- Median, percentiles are related. Both based on position in sorted order. Robust to outliers.

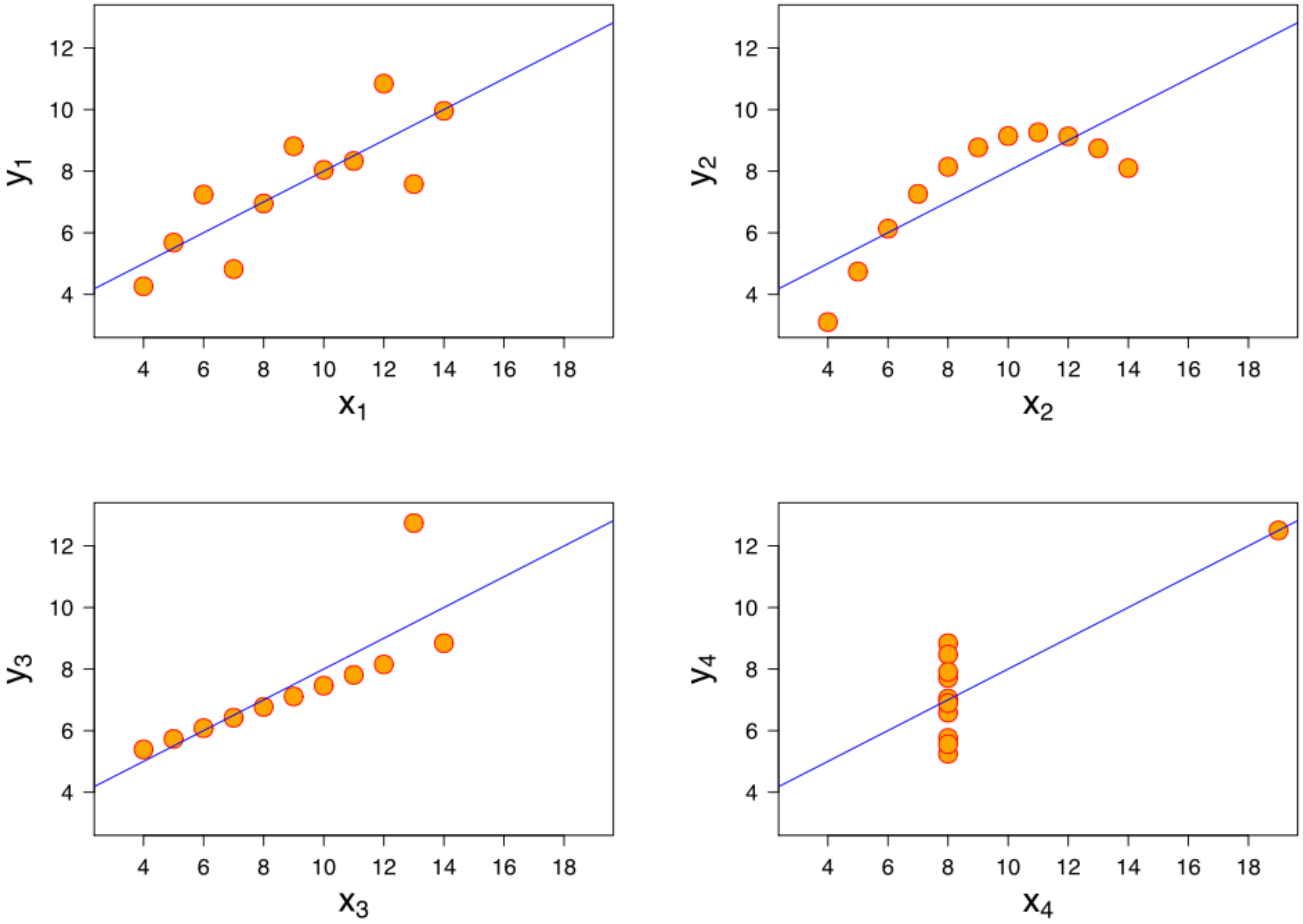
- Can you design a measure that is analogous to correlation but robust to outliers?

Visualizations

- Statistics *summarize* entire dataset into small set of numbers. Information is **lost**.
- Use *visualization…*
  - *…* to see entire data distribution
  - … as *sanity check* on statistics



Anscombe's quartet



Four datasets: all have same means, variances, covariances.

https://en.wikipedia.org/wiki/Anscombe's_quartet

Visualizations

- Statistics *summarize* entire dataset into small set of numbers. Information is **lost**.
- Use *visualization* as sanity check on statistics.

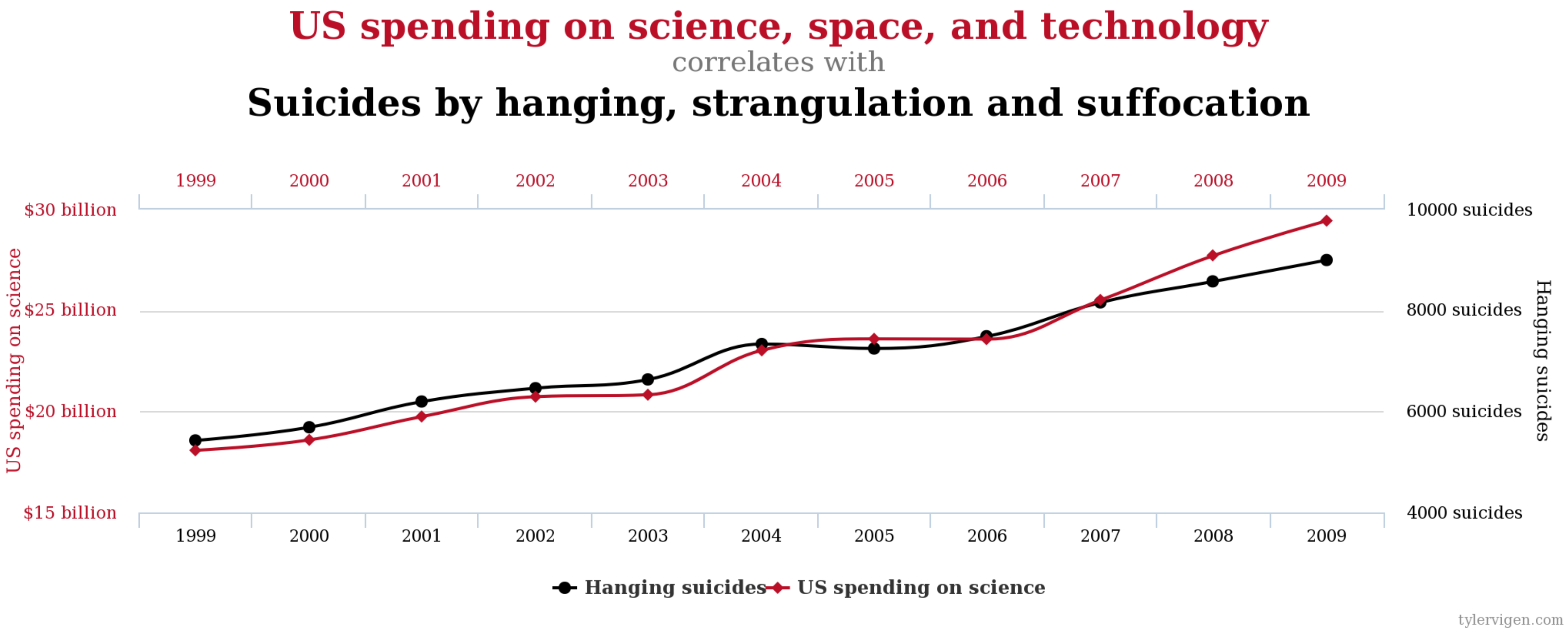
Pitfalls

Caveat: list is not exhaustive

What you *choose* to measure

https://www.youtube.com/watch?v=I1M_8ByLMJQ

Correlation ≠ Causation



http://www.tylervigen.com/spurious-correlations

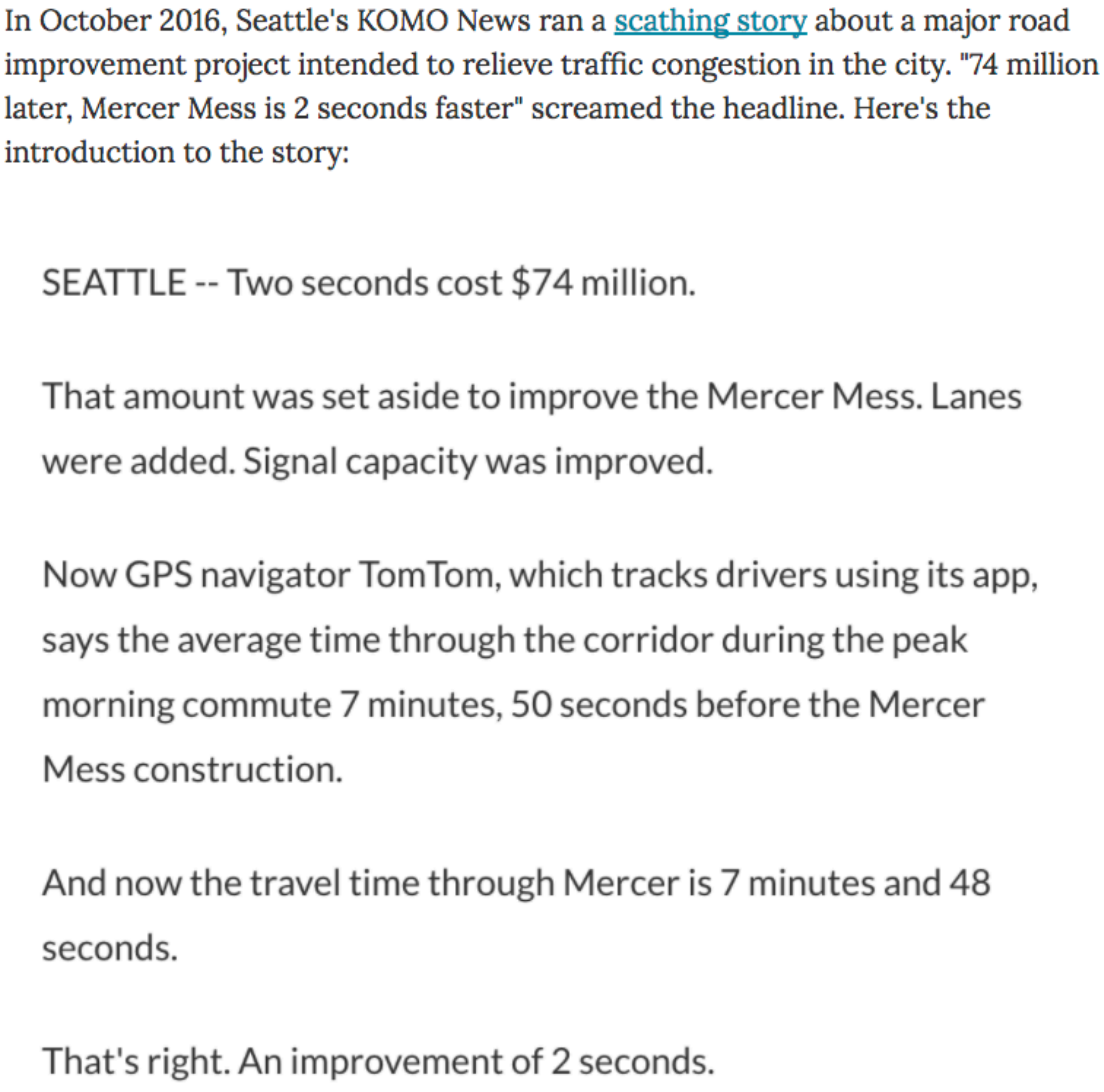
Simpson's paradox

http://vudlab.com/simpsons/

Exercise

Instructions: ~1 minute to think/ answer on your own; then discuss with neighbors; then I will call on one of you

- Critique this statistical analysis



In October 2016, Seattle's KOMO News ran a scathing story about a major road improvement project intended to relieve traffic congestion in the city. "74 million later, Mercer Mess is 2 seconds faster" screamed the headline. Here's the introduction to the story:

> SEATTLE -- Two seconds cost $74 million.
>
> That amount was set aside to improve the Mercer Mess. Lanes were added. Signal capacity was improved.
>
> Now GPS navigator TomTom, which tracks drivers using its app, says the average time through the corridor during the peak morning commute 7 minutes, 50 seconds before the Mercer Mess construction.
>
> And now the travel time through Mercer is 7 minutes and 48 seconds.
>
> That's right. An improvement of 2 seconds.

http://callingbullshit.org/case_studies/case_study_traffic_improvements.html

Summary

- Descriptive statistics useful tool for *summarizing* data.
    - Central tendency: mean, median, mode
    - Dispersion: variance, standard deviation, IQR, range
    - Correlation: covariance, correlation
- Summaries, by definition, have *less information*.
- Use visualization to "see" all of the data, sanity check statistics.
- Beware of pitfalls including measurement choice, correlation ≠ causation, Simpson's paradox.