Probability Review

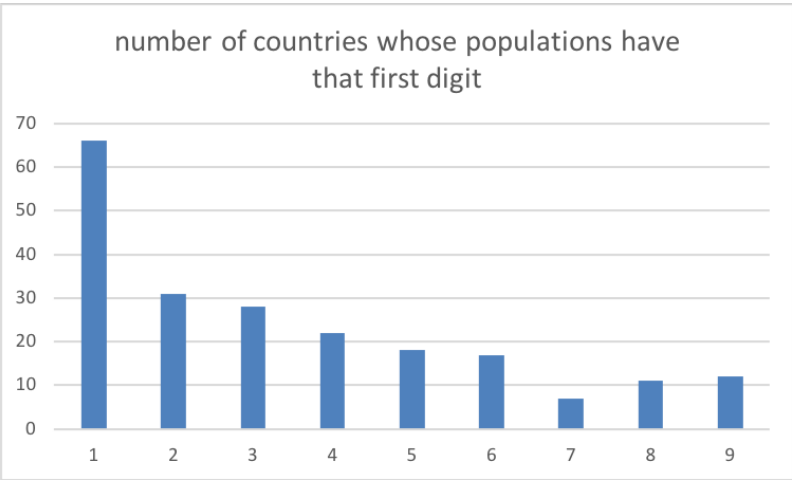
- probability: likeliness/chance: 0 to 1 (or 0 to 100%)
- algebra of probability
  - event will occur = 1 − probability it won't
  - if independent then probabilities multiply

  example: 3 people—NO shared birthday (ignoring leap years and assuming independence)—have probability: $\frac{364}{365}\frac{363}{365}$
  So probability of at least 2 of them sharing a birthday: $1 - \frac{364}{365}\frac{363}{365}$
- expected value can give us an idea of likely outcomes, e.g.
  $\frac{1}{16}$ # coin tossers for HHHH
  $.5 \cdot 85 + .05 \cdot 100 + .3 \cdot 75 + .15 \cdot 95$ for grade
  data might be in a decision matrix like in *Friend or Foe*
- law of large numbers
- Benford's law: first digit in many real-life data sets > 500 approximates a logarithmic trend where 1 occurs about 30% of the time while 9 less than 5%

Benford's Law?

Below is country population data from 2018. Does it satisfy Benford's Law?

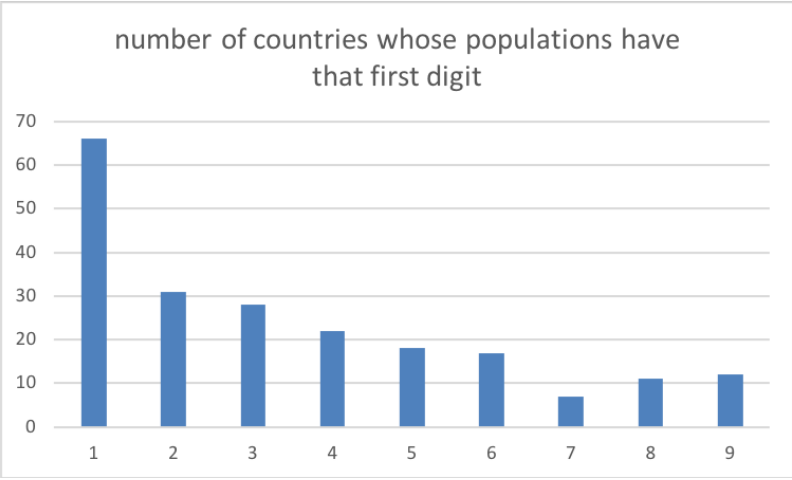
a) it fits perfectly
b) this proves it is fraudulent data and should lead to arrests
c) we should use a larger data set to have a better fit
d) what's a Benford's Law?
e) other

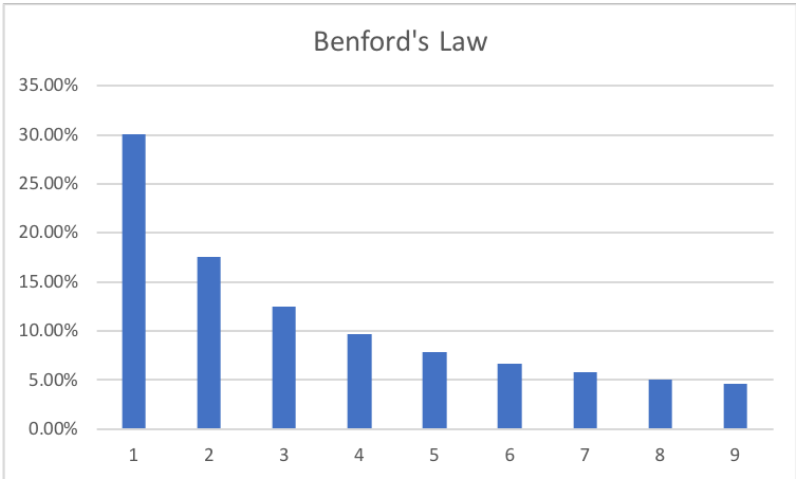


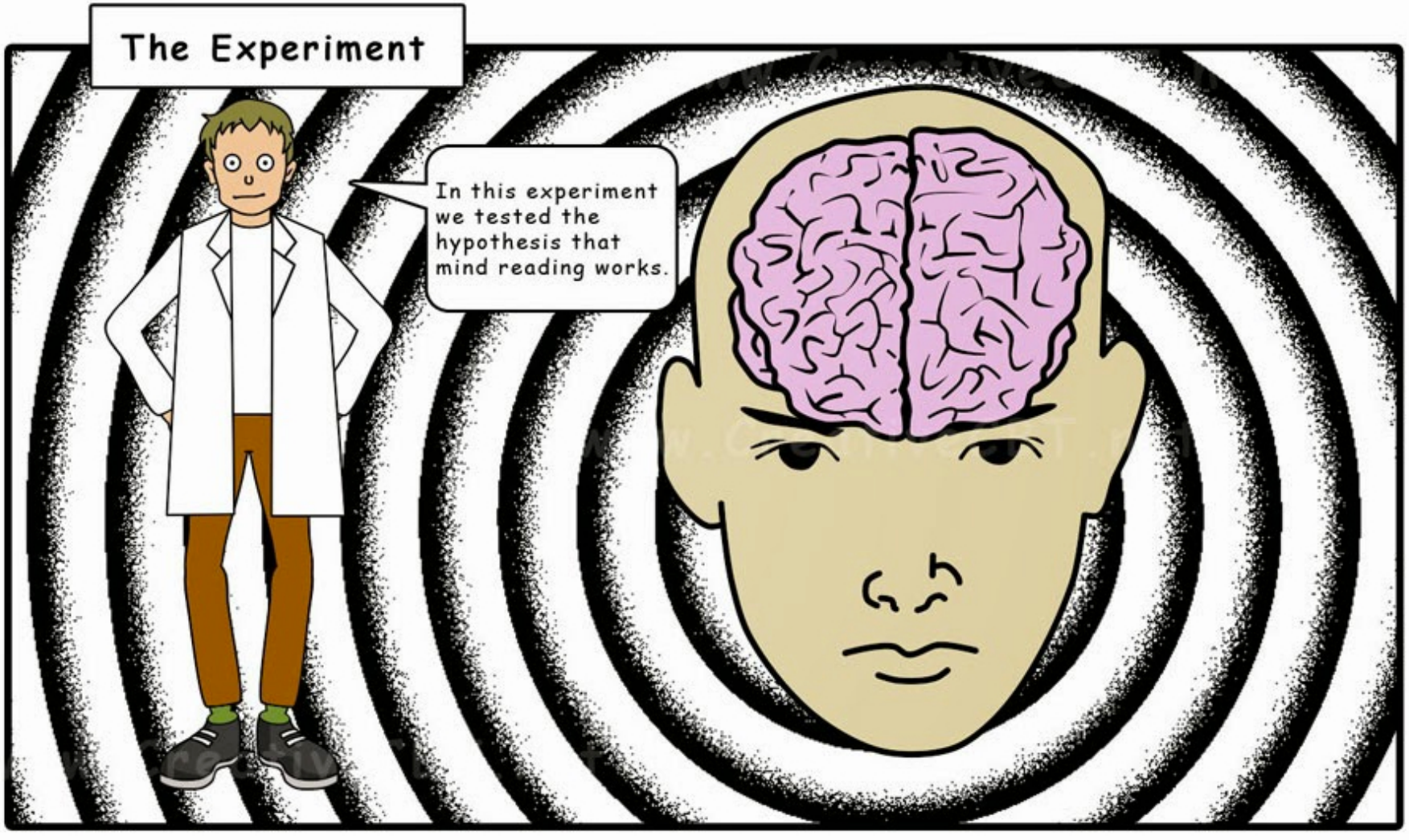
Benford's Law?

Below is country population data from 2018. Does it satisfy Benford's Law?

a) it fits perfectly
b) this proves it is fraudulent data and should lead to arrests
c) we should use a larger data set to have a better fit
d) what's a Benford's Law?
e) other



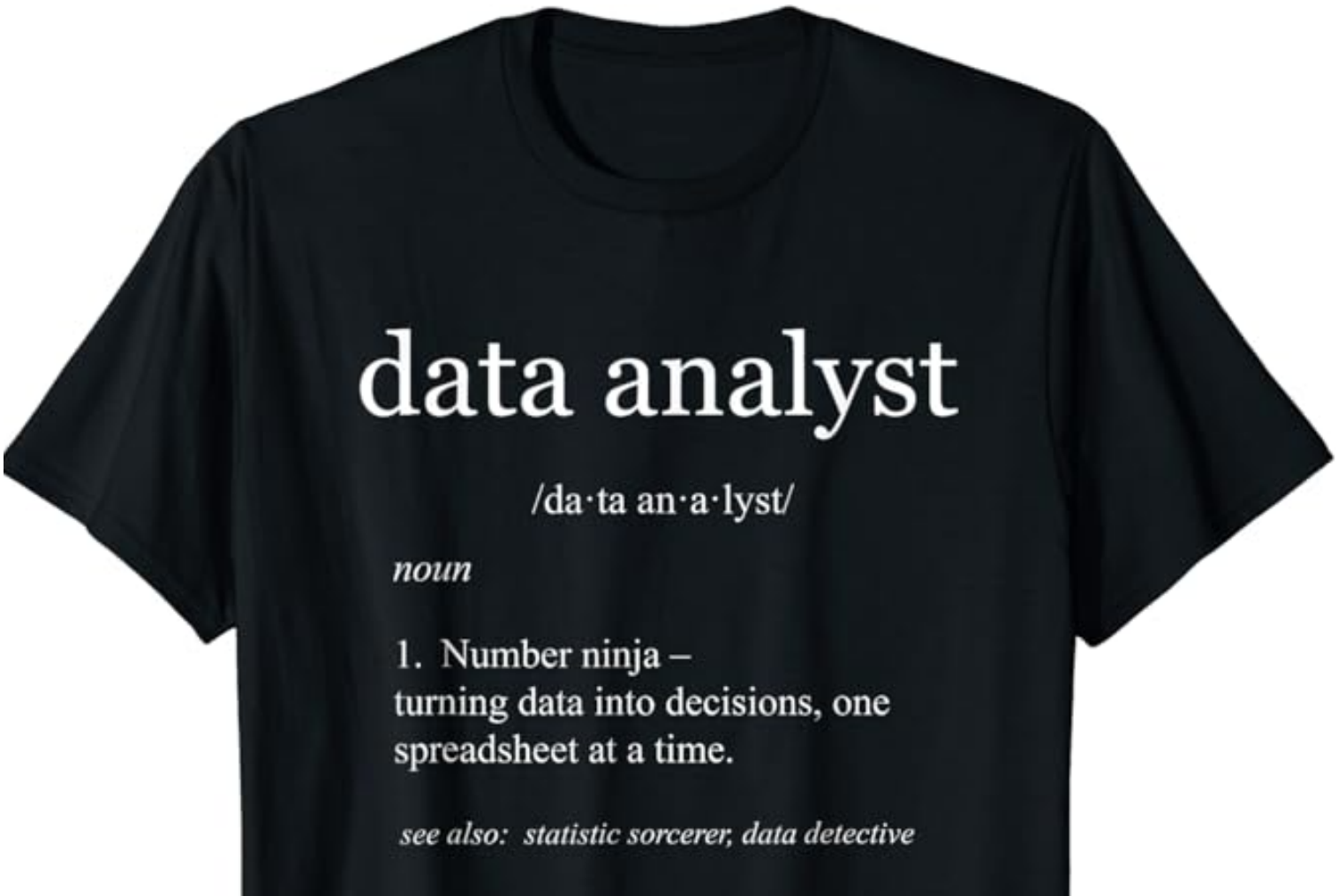




Benjamin Mathews www.creativecbt.net/2013/10/creative-cbt-comic-8-psychic.html

Data Analysis

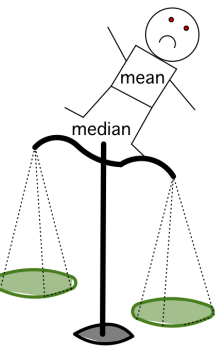
NSF: importance of “harnessing the data... at the local, state, national, and international levels to help unleash the power of data in the service of ... society.”

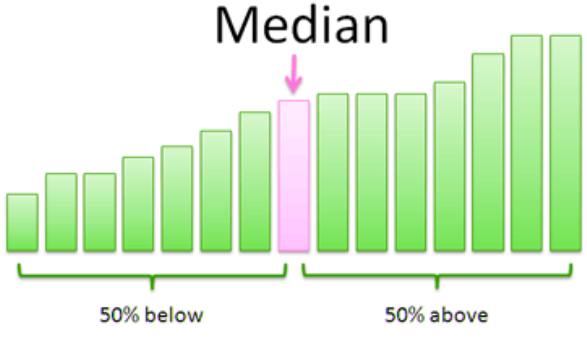


https://www.amazon.com/Analyst-Definition-Funny-Dictionary-Analytics/dp/B0CK57W4H3

Analyzing Central Tendency Measures

- **average**: mean is the middle of the data via “weight” or distance. Sum all the numbers and divide by how many
- **median**: middle of the ordered data, i.e. the midpoint of the distribution. A number and place, if there is no middle number, it is the average of the two nearest

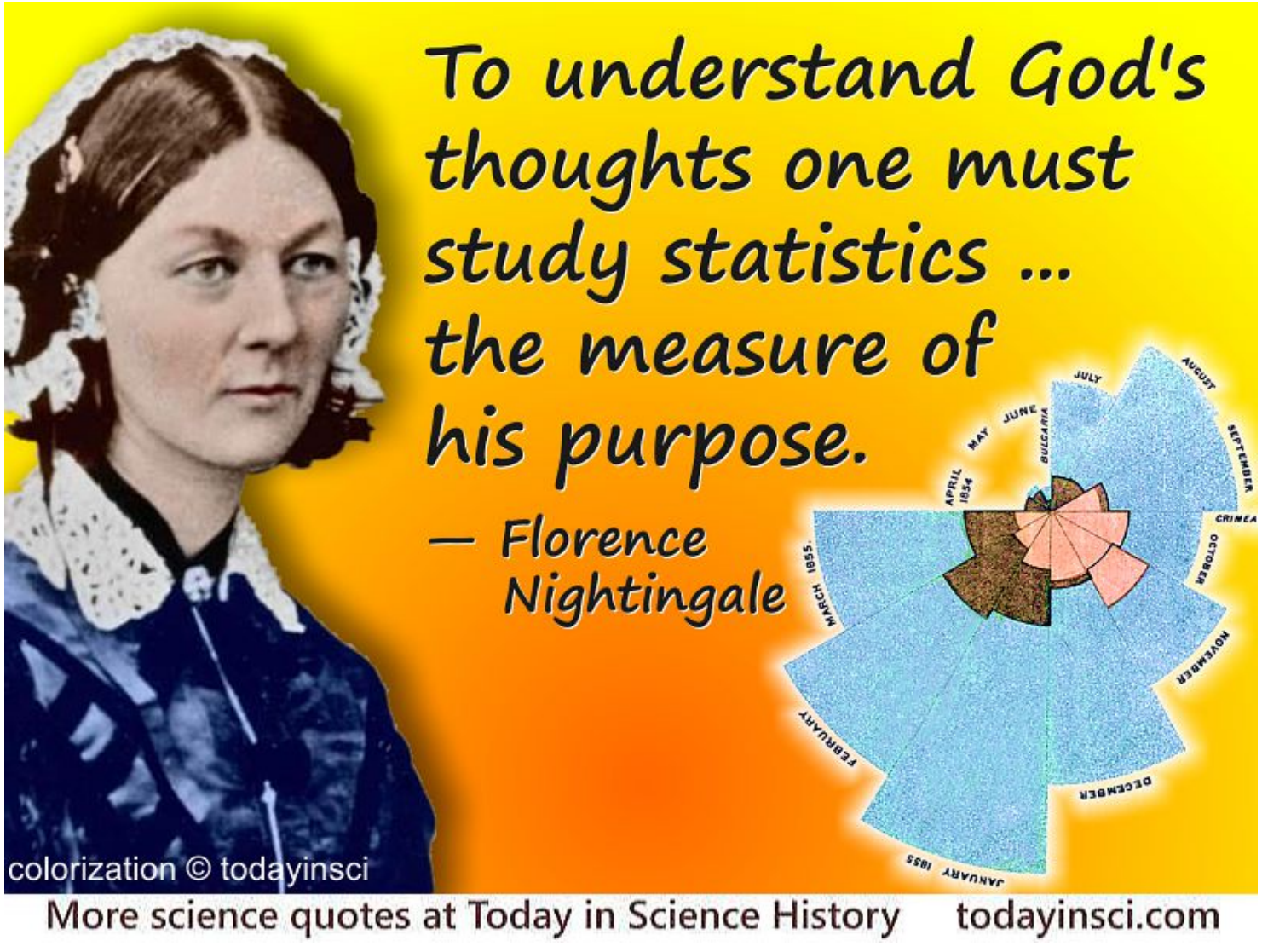




1. data set: 1 2 3 4 5
2. data set: 1 2 4 9
3. A realtor wants to advertise how inexpensive it is to live in an area. Should they use the mean or median. Why?

Picture credit: 1. self-created
2. http://www.testingtreatments.org/2015/02/04/more-than-average-confusion-about-averages/mean-mediocre-median-small/
3. https://byjus.com/maths/median/

Florence Nightingale–Visualization of Stats to Save Lives

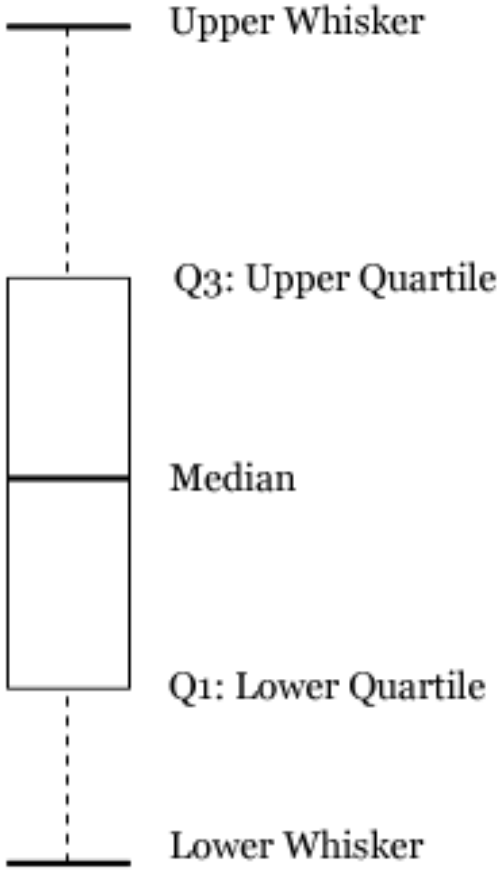


Picture credit:

todayinsci.com/N/Nightingale_Florence/NightingaleFlorence-StatisticsQuote800px.htm

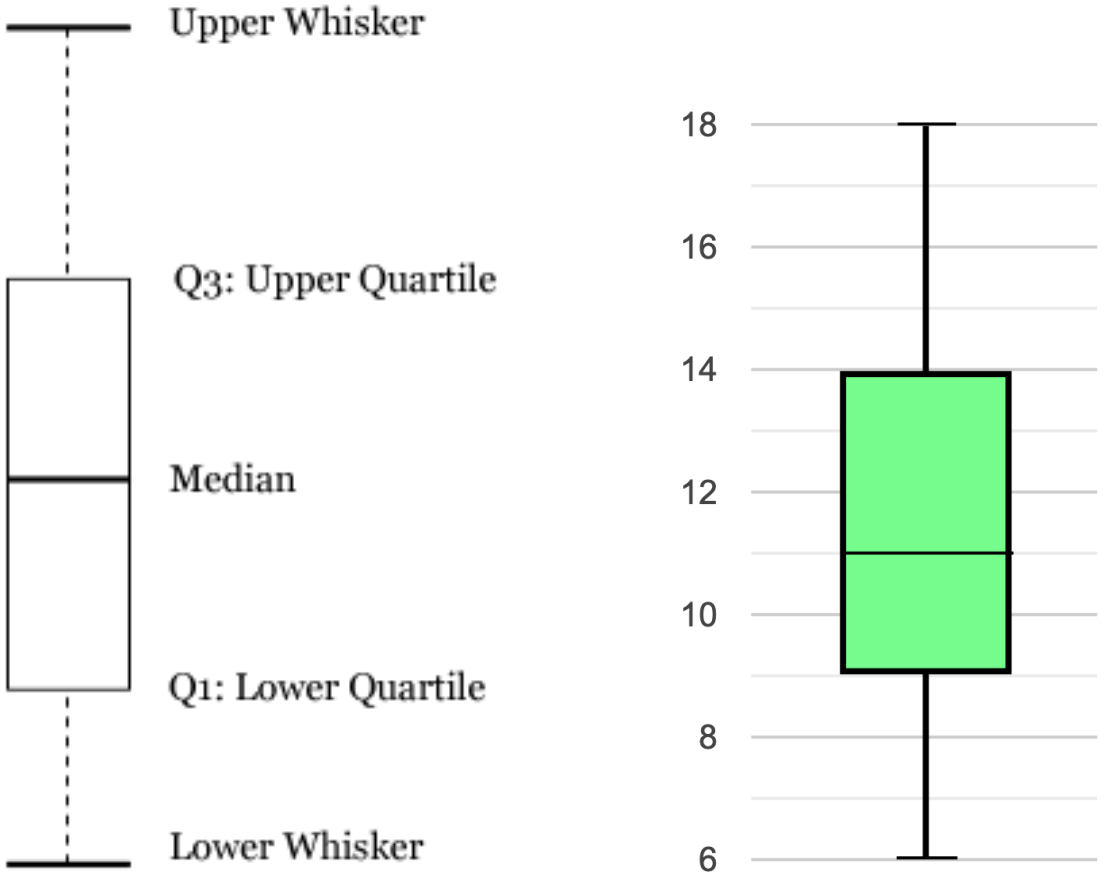
Visualizing the 5-Number Summary: Box plots

1. Excel can give us lo, q1, median, q3, hi. choose an axis and scale for your data: [6, 9, 11, 14, 18]
2. create a visual of these 5 numbers—plot with lines across
3. draw the box and whiskers



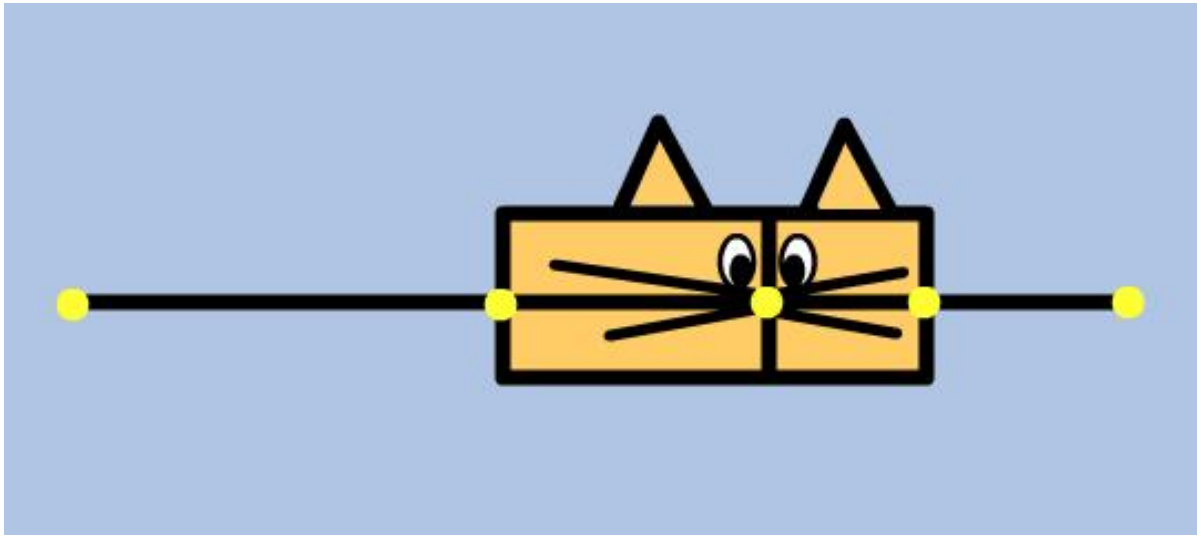
Visualizing the 5-Number Summary: Box plots

1. Excel can give us lo, q1, median, q3, hi. choose an axis and scale for your data: [6, 9, 11, 14, 18]
2. create a visual of these 5 numbers—plot with lines across
3. draw the box and whiskers



Picture credit: `http://informationandvisualization.de/blog/box-plot`
`https://www.statskingdom.com/boxplot-maker.html`

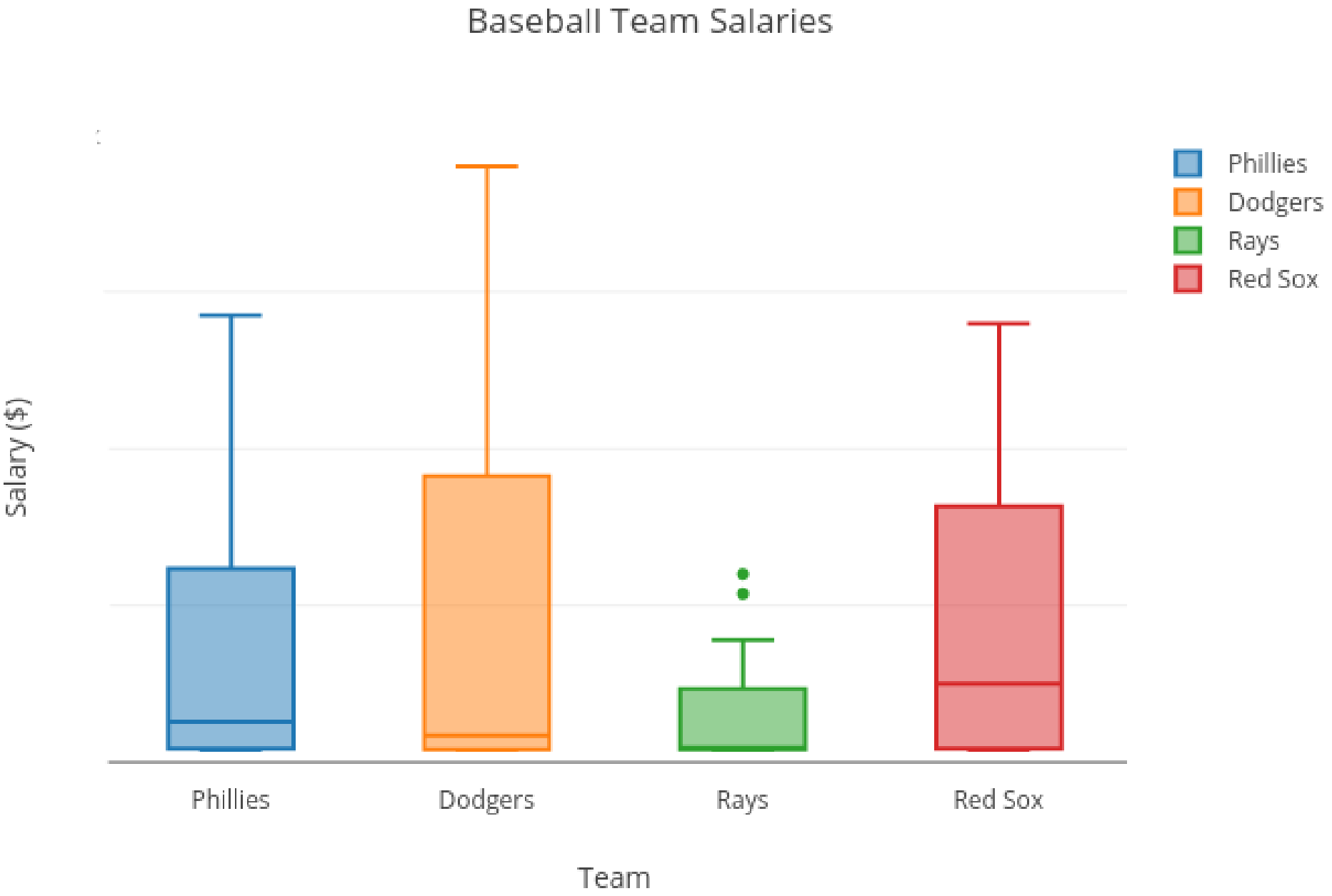
Analyzing Box and Whisker Plots



Picture credit:

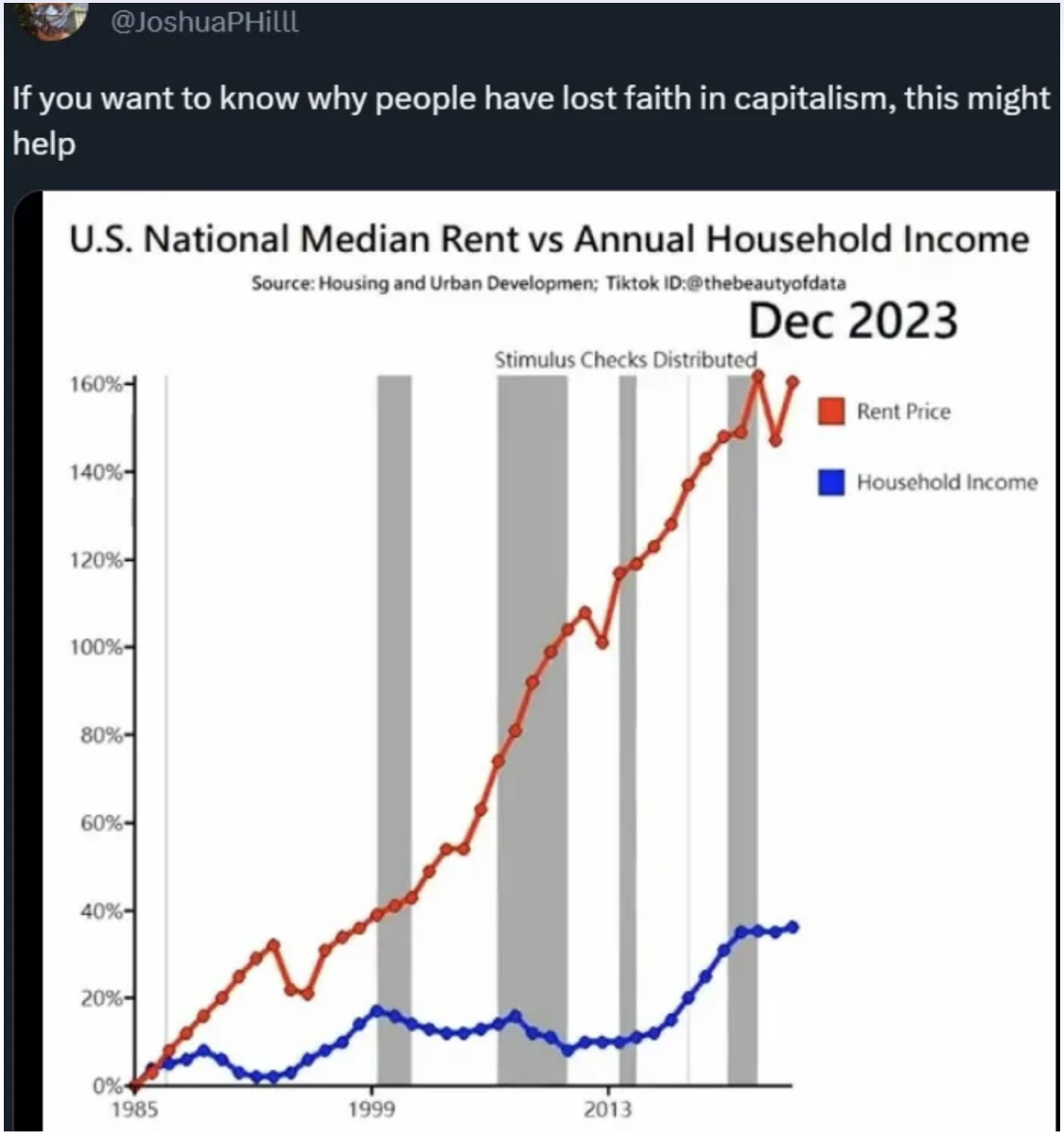
https://oursland.edublogs.org/2015/02/26/hss-id-a-12-and3-box-and-whiskers-plot/

- useful for making comparisons between subgroups using side-by-side box plots
- easy to see the middle 50% of the data in the box...
- central value (median), spread...



https://plot.ly/~nebeirne/13/baseball-team-salaries.png

Comparing and Ranking Countries—Your Data



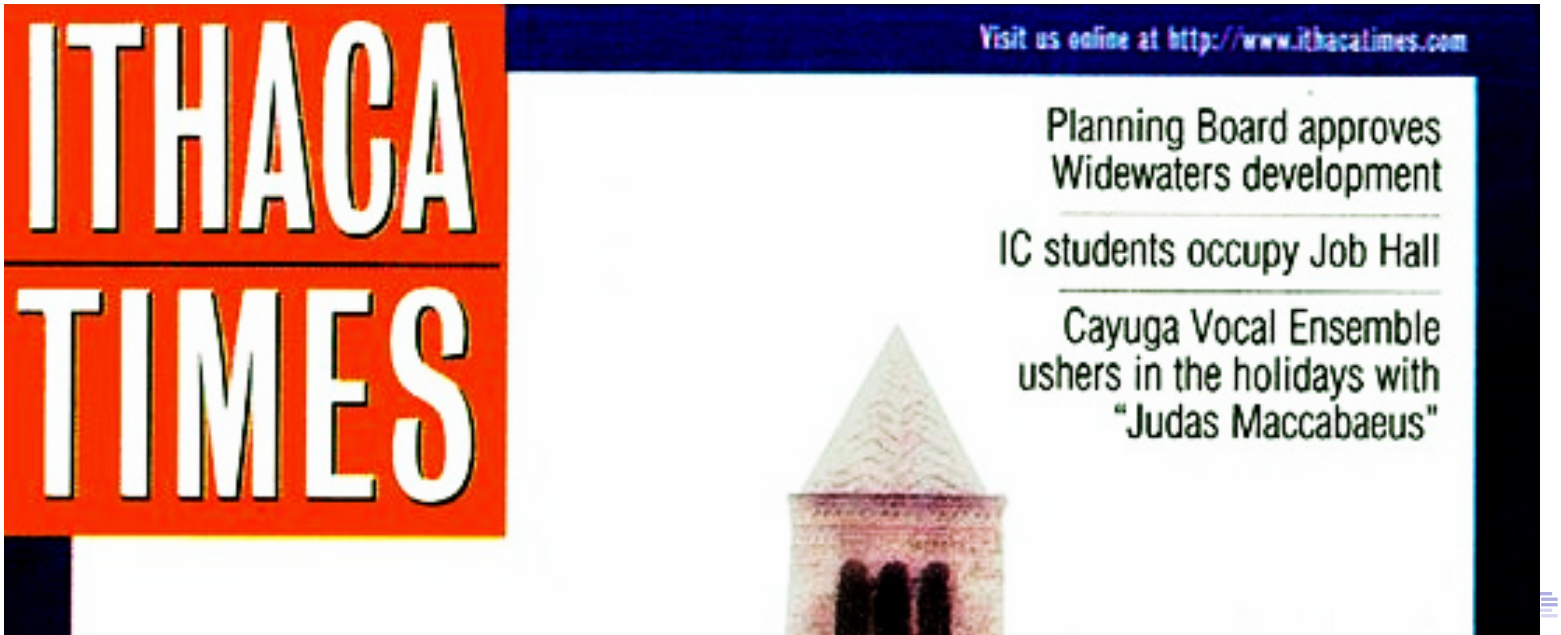
household income has been adjusted for inflation but rent price was not!

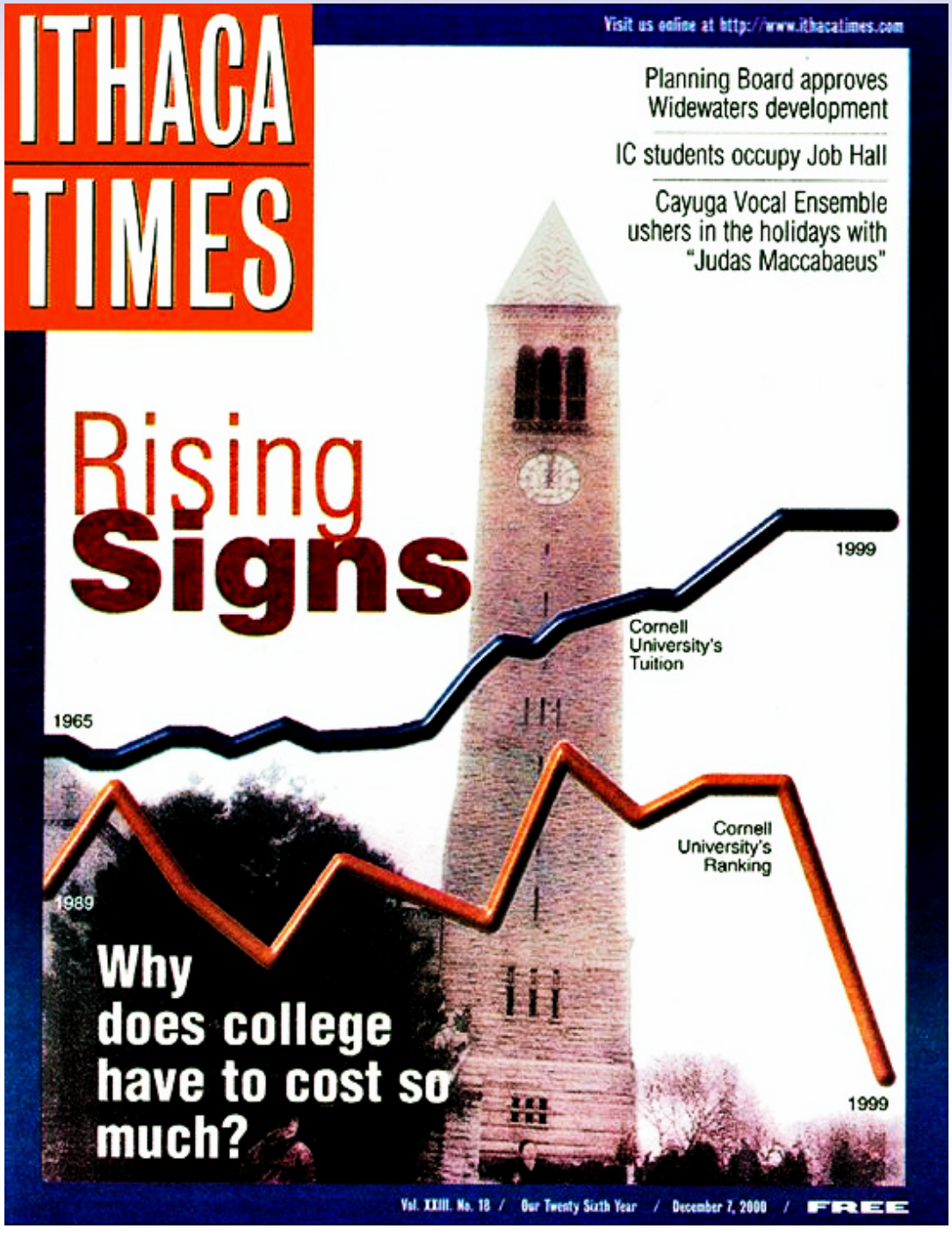
Worst Graph I've Ever Seen

Dave Bock submitted the cover of a magazine:

- tuition at Cornell University over 35 years
- ranking of Cornell University over 11 years

unbalanced data
axes manipulation
improper scaling
untrue correlation




ITHACA TIMES
Visit us online at http://www.ithacatimes.com
Planning Board approves Widewaters development
IC students occupy Job Hall
Cayuga Vocal Ensemble ushers in the holidays with "Judas Maccabaeus"
Rising Signs
1999
Cornell University's Tuition
1965
Cornell University's Ranking
1989
1999
Why does college have to cost so much?
Vol. XXIII. No. 18 / Our Twenty Sixth Year / December 7, 2000 / FREE

Is it appropriate for scientists and mathematicians today to use data obtained under what we agree were morally reprehensible conditions, like the syphilis experiment in Tuskegee, Alabama or many experiments in Nazi Germany?

a) yes
b) only in certain circumstances
c) no

Write down what you think are the strongest argument(s) from the “yes” and “no” sides. Pollev and share.

Is it appropriate for scientists and mathematicians today to use data obtained under what we agree were morally reprehensible conditions, like the syphilis experiment in Tuskegee, Alabama or many experiments in Nazi Germany?

a) yes
b) only in certain circumstances
c) no

Write down what you think are the strongest argument(s) from the "yes" and "no" sides. Pollev and share. Baruch C. Cohen:

- physiological responses of tortured victims
- biases of Nazi doctors' political aspirations and their enthusiasm for eugenics—not peer reviewed nor replicated
- moral hypocrisy: denounce and condemn the doctors and their experiments vs. use data to possibly benefit others—legitimacy is indirectly conferred
- EPA barred the use of Nazi data

Transplanting a murdered person's heart without consent? Forensic evidence obtained during illegal search and seizure?

Institutional Review Board and Ideal Collection

Belmont Report (1978):

- Respect for Persons (consent, privacy, additional safeguards for those vulnerable to coercion or undue influence)
- Beneficence (maximize benefits as you minimize risks)
- Justice (benefits and burdens of research are equitably distributed)

Reproducibility, consensus, and random sampling if possible

Picture credit: http://med.stanford.edu/content/dam/sm-news/images/2016/06/Replication.jpg

A research group wants to study the effectiveness of a quercetin supplement, and has contacts at the Watauga County Detention Center, Rikers Island in New York, and Silivri Penitentiaries Campus in Turkey. Quercetin is a flavonoid found in fruits and vegetables that is a strong antioxidant. Do you think Appalachian State University's Institutional Research Board (IRB) will approve such research?

a) yes and I have a good reason why
b) yes but I'm unsure of why
c) no but I'm unsure of why not
d) no and I have a good reason why not
e) other

Kat is making measurements in lab and is confident that they have set it up properly. When Kat tries to do the required calculations to verify the formulas in the book, the data seems wrong. So Kat does the mathematical calculations to determine what a correct set of data would be and simply changes the data to match the calculations.

a) I feel strongly that Kat's actions are ok
b) I somewhat feel that Kat's actions are ok
c) I somewhat feel that Kat's actions are problematic
d) I feel strongly that Kat's actions are problematic
e) other

Kat is making measurements in lab and is confident that they have set it up properly. When Kat tries to do the required calculations to verify the formulas in the book, the data seems wrong. So Kat does the mathematical calculations to determine what a correct set of data would be and simply changes the data to match the calculations.

a) I feel strongly that Kat's actions are ok
b) I somewhat feel that Kat's actions are ok
c) I somewhat feel that Kat's actions are problematic
d) I feel strongly that Kat's actions are problematic
e) other

falsification?
reproducibility, consensus and random sampling if possible

Comparing and Ranking Countries—Your Data