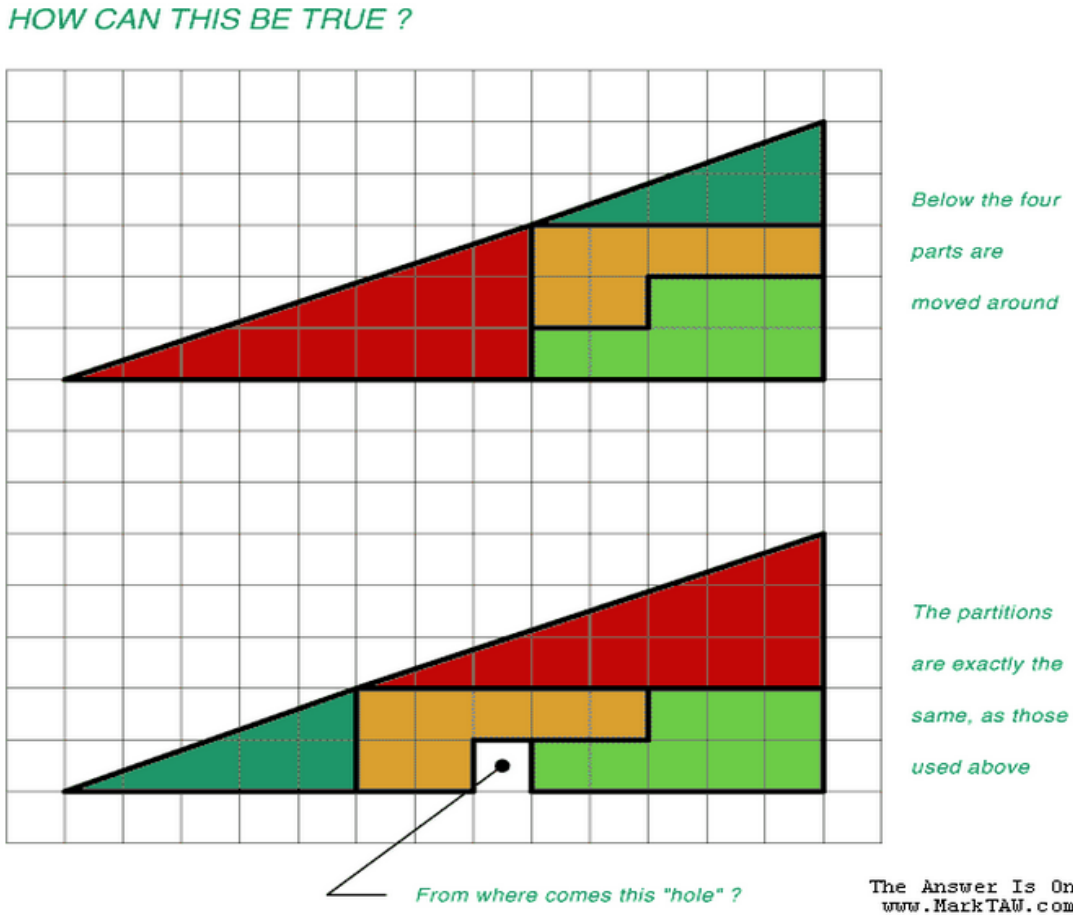
TIM KRASKA

LYING WITH STATISTICS AND VISUALIZATIONS



http://www.marktaw.com/blog/TheTriangleProblem.html

TEXT PROCESSING

6.S080 SOFTWARE SYSTEMS FOR DATA SCIENCE

TIM KRASKA

CASE STUDY FOR THIS CLASS

You work at Nickelback Inc.

Nickelback Inc recently downloaded every song text ever written (TB of data) to draw inspiration as they lately have trouble to produce a number 1 hit.

Now they want to create a system which enables them to search through this large collection of text and help them to write some songs.

Your task:

Task1: Design a system that efficiently finds all song texts contain certain keywords (e.g., “mountain” and “grass”)

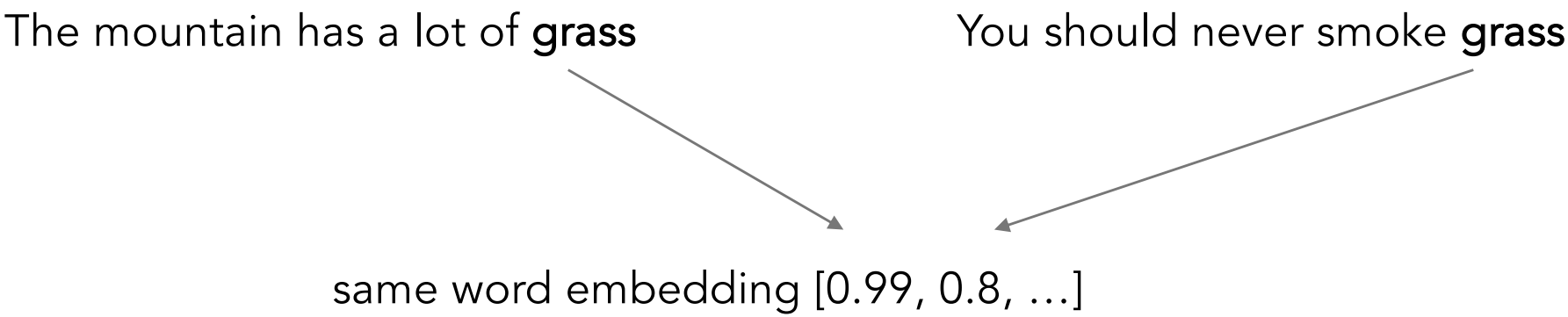
Task2: Create a simple ranking for the query results and enable that Nickelback can cluster the songs

Task3: Extend the system to allow search with sentiments (e.g., all happy songs, sad songs,…)

Task4: Extend the system further to find songs with the right meaning of “grass” (the green stuff in the football stadium)

Task5: Develop an assistant that helps Nickelback to write songs by predicting the next sentence

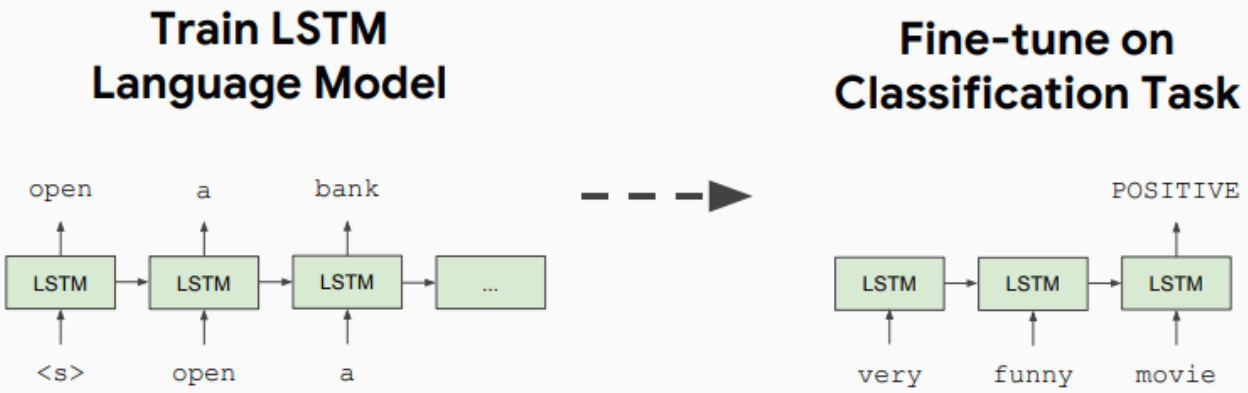
WHAT IS THE PROBLEM WITH WORD EMBEDDINGS?



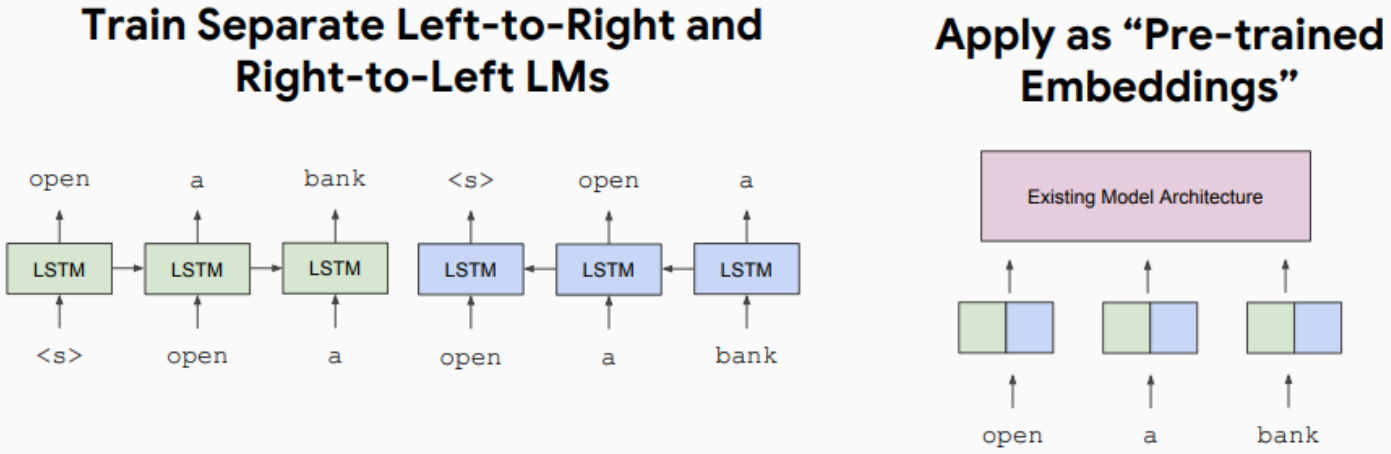
Solution: Train contextual representations on text corpus

LITTLE HISTORY

Semi-Supervised Sequence Learning, Google, 2015

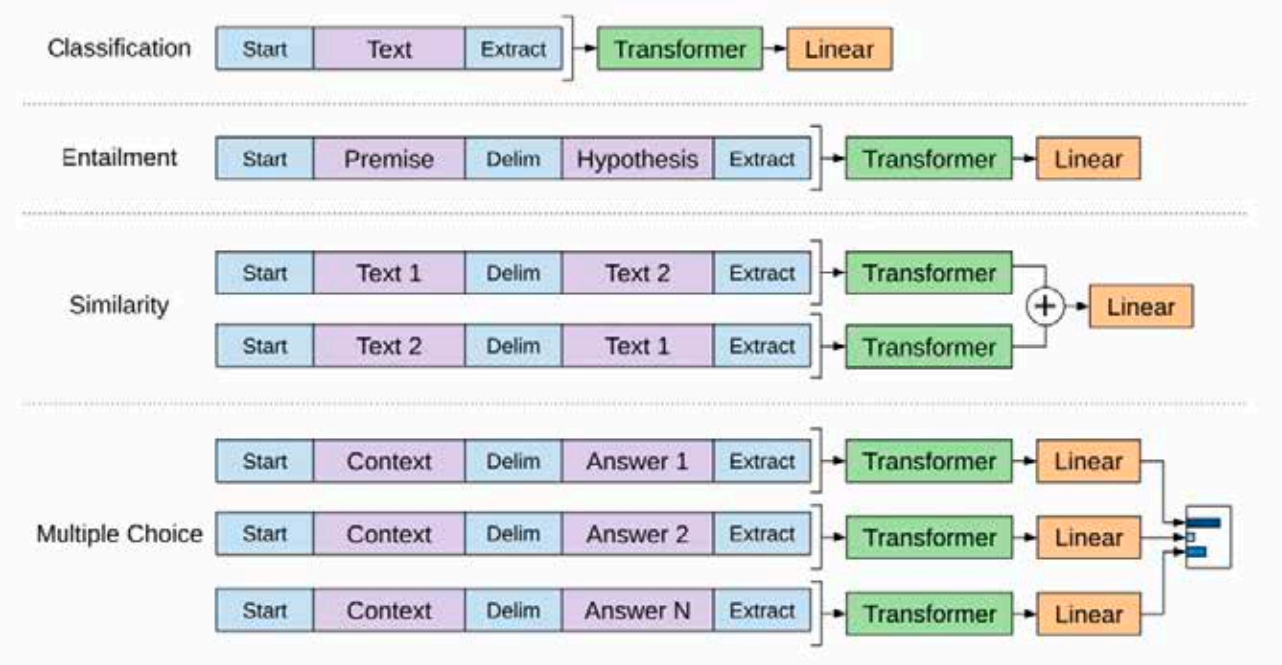


ELMo: Deep Contextual Word Embeddings, AI2 & University of Washington, 2017



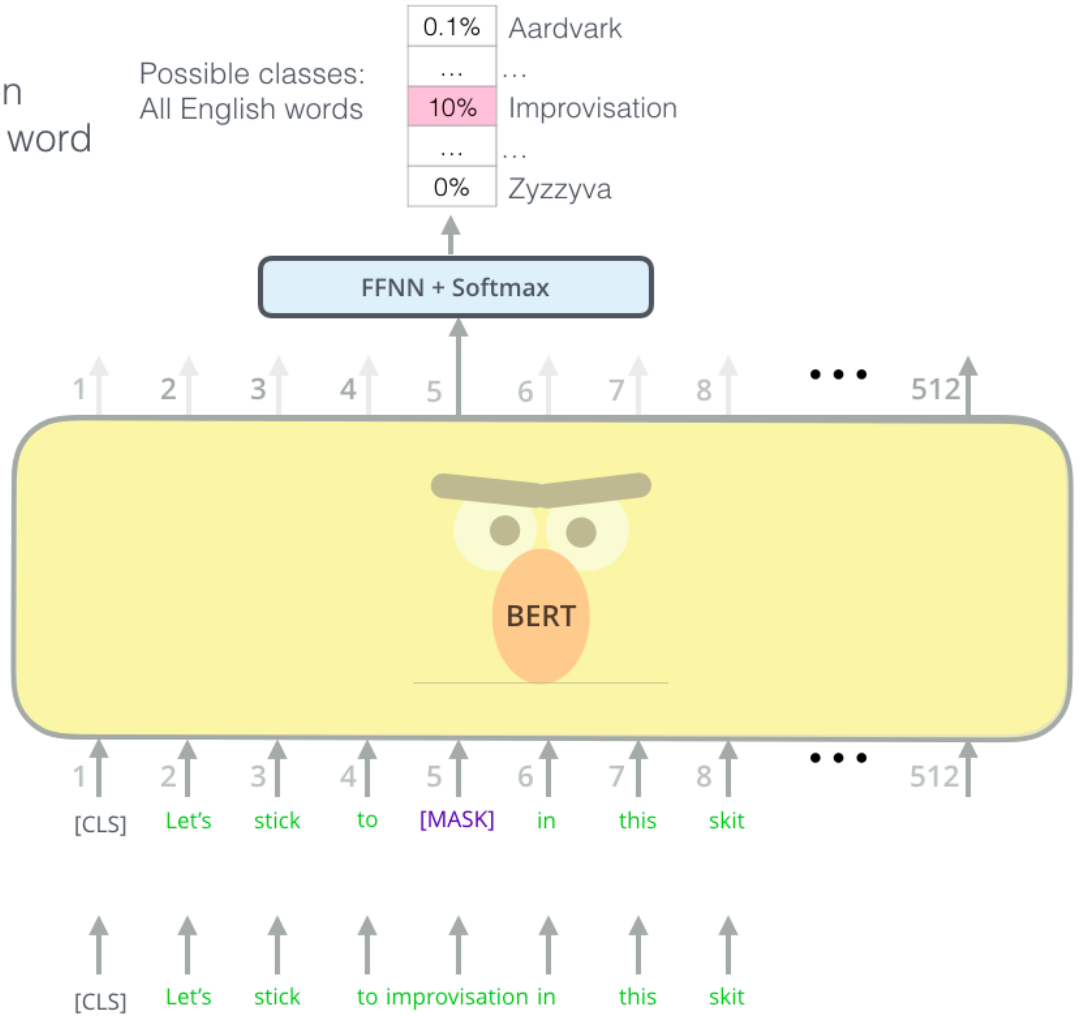
GPT

Improving Language Understanding by Generative Pre-Training, OpenAI, 2018 – Based on transformers/attention from "Attention is All You Need" Vaswani et al



BERT

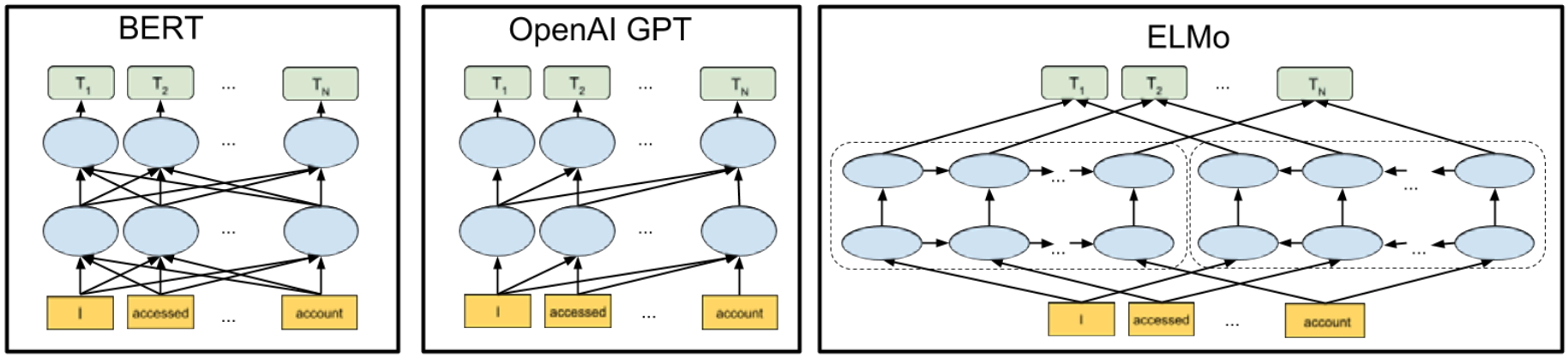
Use the output of the masked word's position to predict the masked word



Randomly mask 15% of tokens

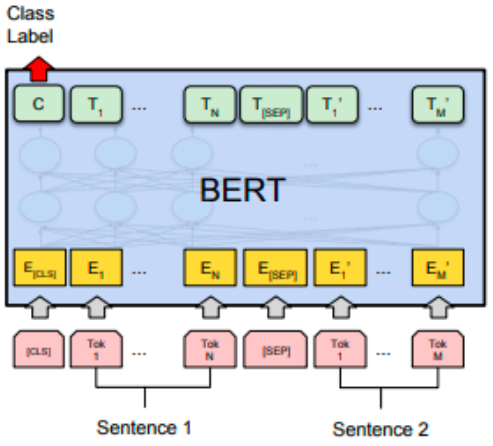
Input

BERT VS OPENAI GPT VS ELMO

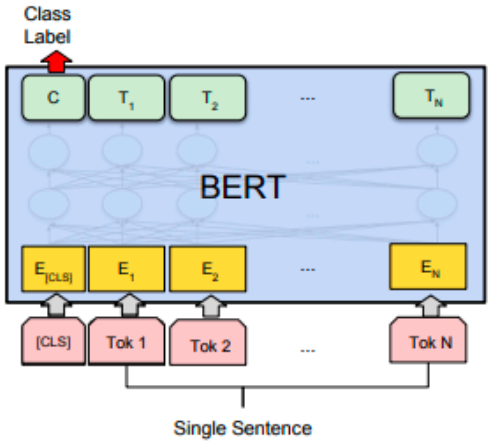


See also http://jalammar.github.io/illustrated-gpt2/

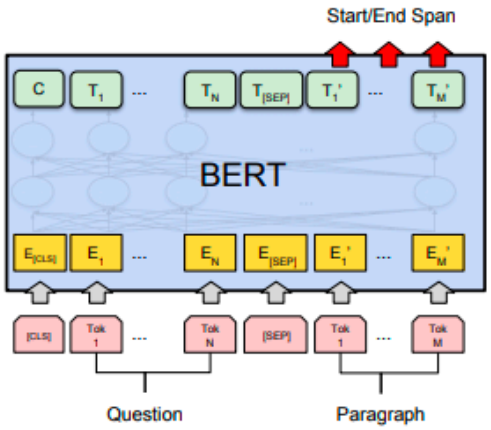
TASKS



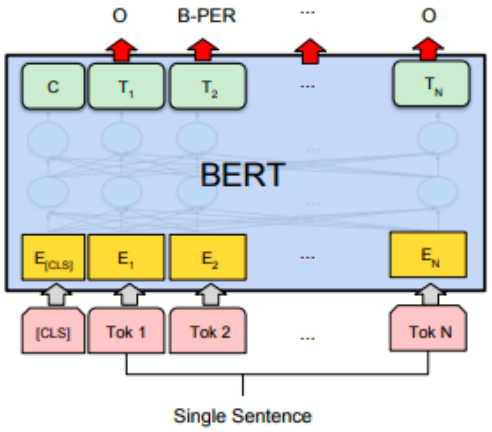
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA



(c) Question Answering Tasks: SQuAD v1.1



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

http://www.msmarco.org/leaders.aspx

MICROSOFT MACHINE READING COMPREHENSION DATASET (MS MARCO)

KeyPhrase Extraction(10/18/2019) ranked by F1 @3 on Eval

Rank	Model	Submission Date	Precision @1,@3,@5	Recall @1,@3,@5	F1 @1,@3,@5
1	**BERT (Base) Sequence Tagging** Si Sun (Tsinghua University), Chenyan Xiong (MSR AI), Zhiyuan Liu (Tsinghua University) [Code]	November 5th, 2019	0.484, 0.312, 0.227	0.255, 0.469, 0.563	0.321, **0.361**, 0.314
2	**Baseline finetuned on Bing Queries** MSMARCO Team	October 19th, 2019	0.397, 0.249, 0.149	0.215, 0.391, 0.391	0.267, **0.292**, 0.209
3	**Baseline** MSMARCO Team	October 19th, 2019	0.365, 0.237, 0.142	0.196, 0.367, 0.367	0.244, **0.277**, 0.198

Passage Retrieval(10/26/2018-Present) ranked by MRR on Eval

Rank	Model	Ranking Style	Submission Date	MRR@10 On Eval	MRR@10 On Dev
1	**Enriched BERT base + AOA index + CAS** Ming Yan of Alibaba Damo NLP	Full Ranking	August 20th, 2019	0.393	0.408
2	**W-Index retrieval + BERT-F re-rank** Zhuyun Dai of Carnegie Mellon University	Full Ranking	September 12th,2019	0.388	0.394
3	**Enriched BERT base + AOA index V1** Ming Yan of Alibaba Damo NLP	Full Ranking	May 13th, 2019	0.383	0.397

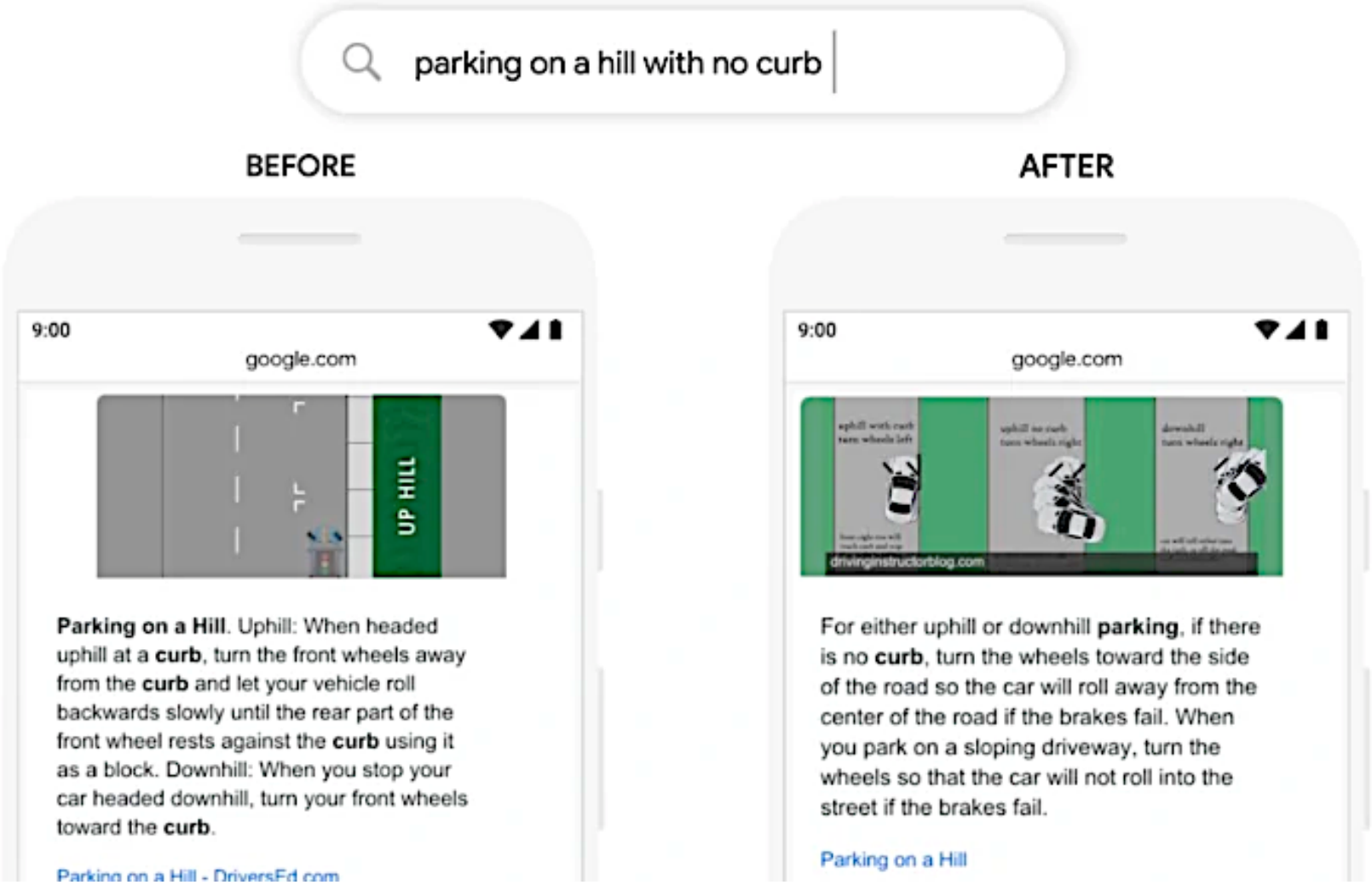
Q&A Task(03/01/2018-Present)

Rank	Model	Submission Date	Rouge-L	Bleu-1
1	**Multi-doc Enriched BERT** Ming Yan of Alibaba Damo NLP	June 20th, 2019	0.540	0.565
2	**Human Performance**	April 23th, 2018	0.539	0.485
3	**BERT Encoded T-Net** Y. Zhang, C. Wang, X.L. Chen	August 5th, 2019	0.526	0.539

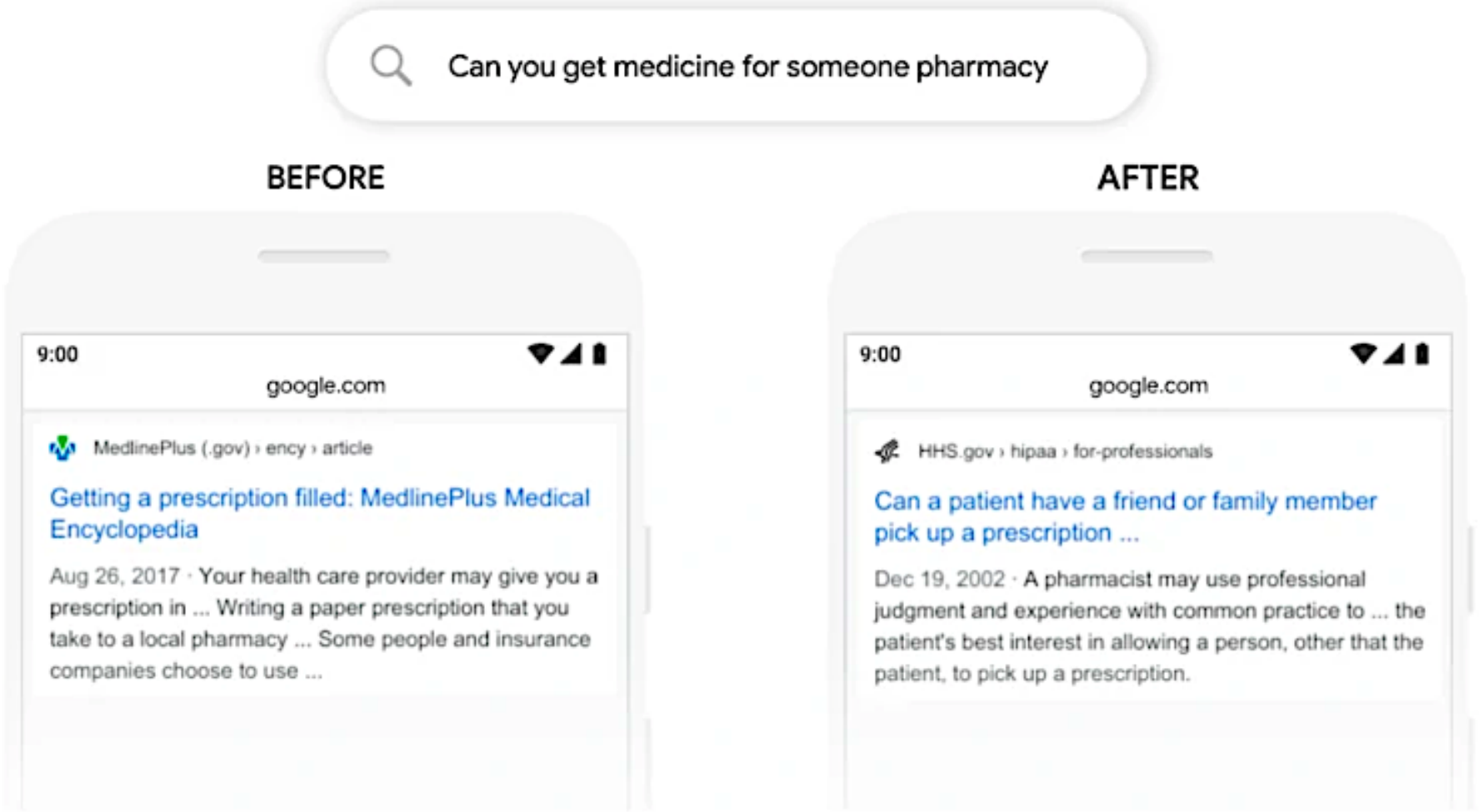
Q&A + Natural Language Generation Task(03/01/2018-Present)

Rank	Model	Submission Date	Rouge-L	Bleu-1
1	**Human Performance**	April 23th, 2018	0.632	0.530
2	**Masque NLGEN Style** NTT Media Intelligence Laboratories [Nishida et al. '19]	January 3rd, 2019	0.496	0.501
3	**BERT+ Multi-Pointer-Generator** Tongjun Li of the ColorfulClouds Tech and BUPT	June 11th,2019	0.495	0.476

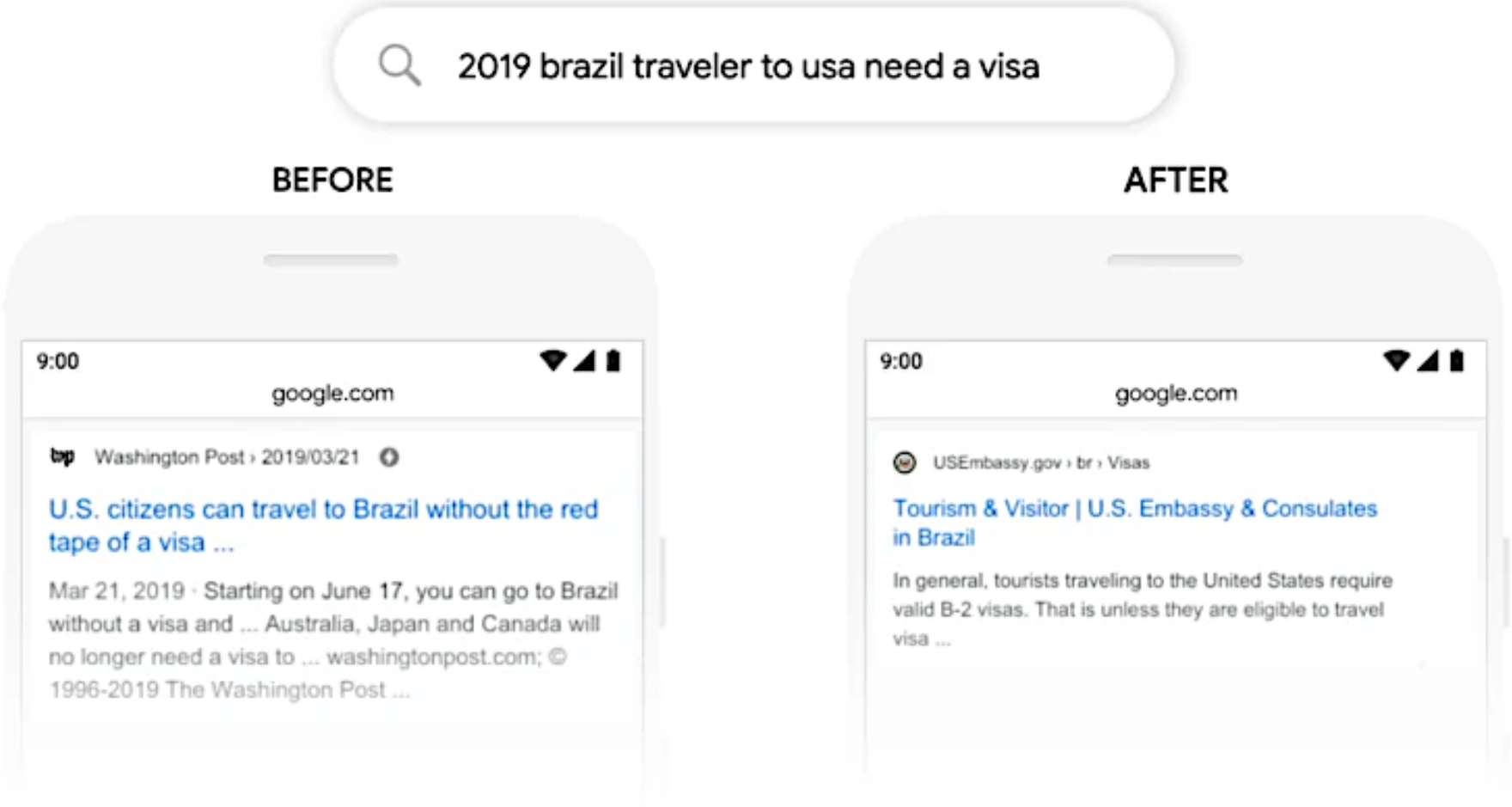
GOOGLE IS NOW USING BERT



GOOGLE IS NOW USING BERT

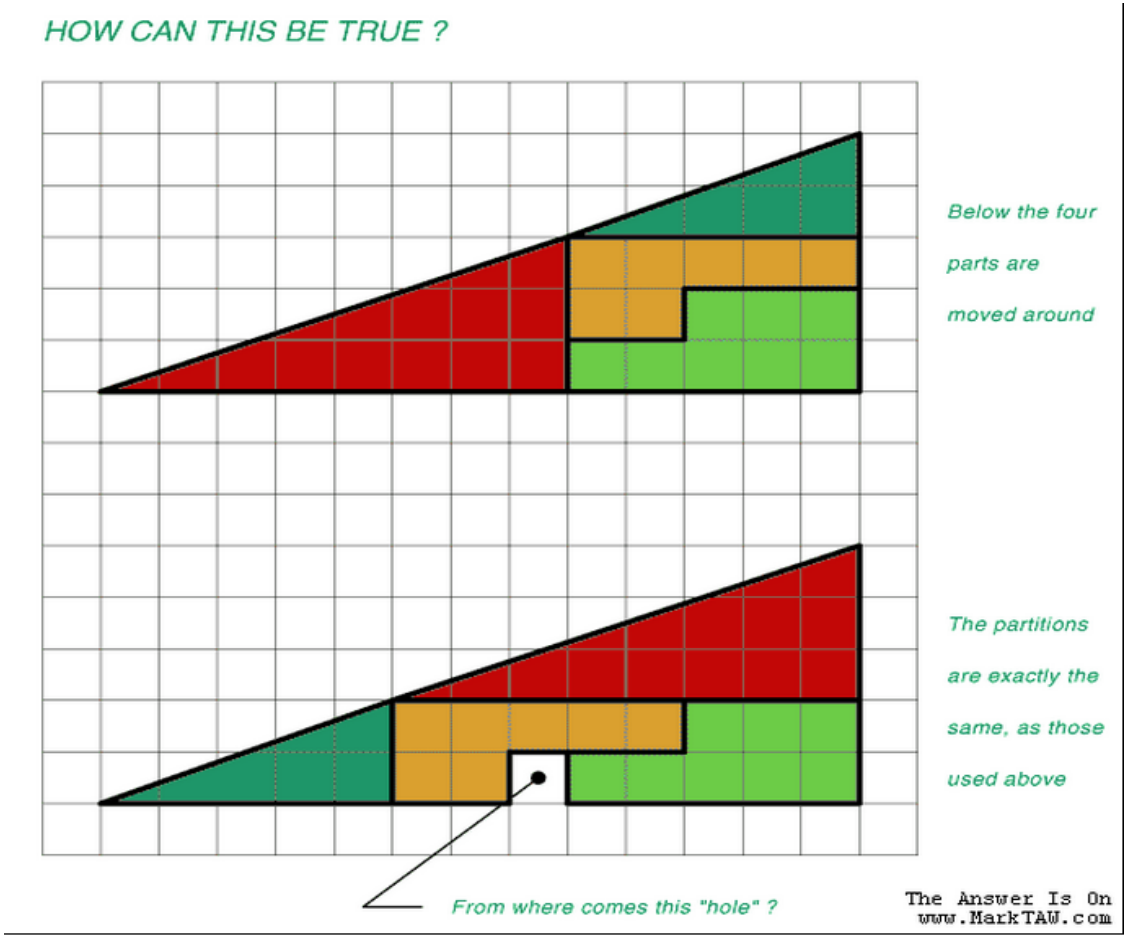


GOOGLE IS NOW USING BERT



TIM KRASKA

LYING WITH STATISTICS AND VISUALIZATIONS



http://www.marktaw.com/blog/TheTriangleProblem.html

SURVEYS

LYING WITH NUMBERS

"The average market salary for MIT graduates with 0-2 years of experience is $151,000 per year"

In how many ways can this be misleading?

https://www.paysa.com/blog/harvard-vs-mit-which-tech-grads-make-the-most/

SAMPLING BIAS

- More successful graduates are more likely to respond to surveys
  - They feel good about their earnings
  - Surveys are only sent to big companies
- How big is their sample size?
  - Not disclosed

- Tendency to exaggerate
  - Brag about your salary
  - School spirit, want your alma mater to rank highly
- Tendency to minimize
  - No one likes tax
- Do they cancel out each other?
  - No one knows!

LYING WITH NUMBERS

"The average market salary for MIT graduates with 0-2 years of experience is $151,000 per year"

In how many ways can this be misleading?

https://www.paysa.com/blog/harvard-vs-mit-which-tech-grads-make-the-most/

THE TERM "AVERAGE"

- Imagine a school with 5 alumni
  - Bill Doors: $1 million/year
  - Mark Bergkerzuck: $120k/year
  - Larry Sheet: $100k/year
  - Sergey Bin: $80k/year
  - Steve Baller: $80k/year
- Average can be mean, median, or mode; They can be totally different
- Mean: Evenly distributes the total among individuals
  - Can be unrepresentative when measurements are highly skewed
  - In our example: $276k/year
- Median: Value dividing distribution into two equal parts (50th percentile)
  - In our example: $100k/year
- Mode: Most frequently observed outcome (rarely reported with numeric data)
  - In our example: $80k/year

Slide by Larry Winner, "Overview of How to Lie with Statistics"

FINAL SENTENCE

"The average market salary for MIT graduates with 0-2 years of experience is $151,000 per year"

PayScale's methodology did not include alums with advanced degrees and only used data from graduates with bachelor's degrees. It also excluded self-employed and contract employees.

Because the salaries of graduates from elite schools vary extensively, the study has a relatively wide margin of error, the report stated.

CORRELATION VS CAUSATION

What conclusions can you make from this data?

Does going to MIT make you rich?

LYING WITH SURVEYS

Three questions you should ask after you read any paper:

1. Is there any bias in the sample set?
    a. Look for unconscious bias
    b. Look for conscious bias
2. What statistics are they actually talking about?
3. What conclusions can we make from their findings?

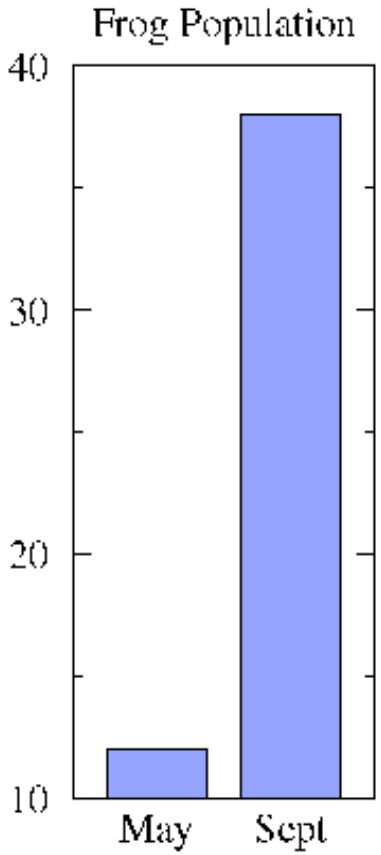
LYING WITH VISUALIZATIONS

LYING WITH VISUALIZATION

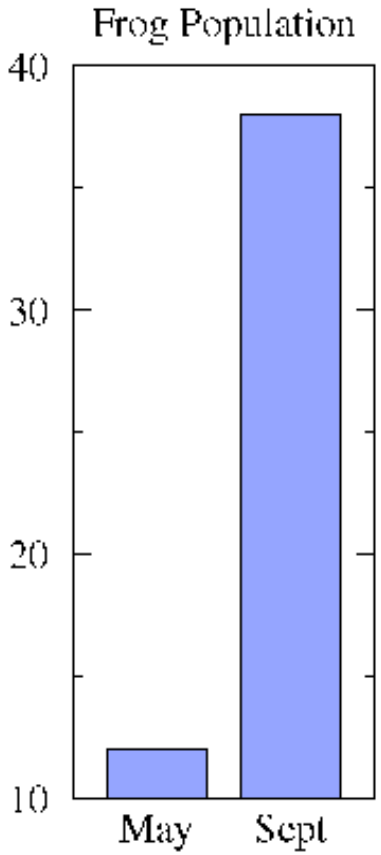
~~*seeing is believing*~~

"don't believe everything you see."

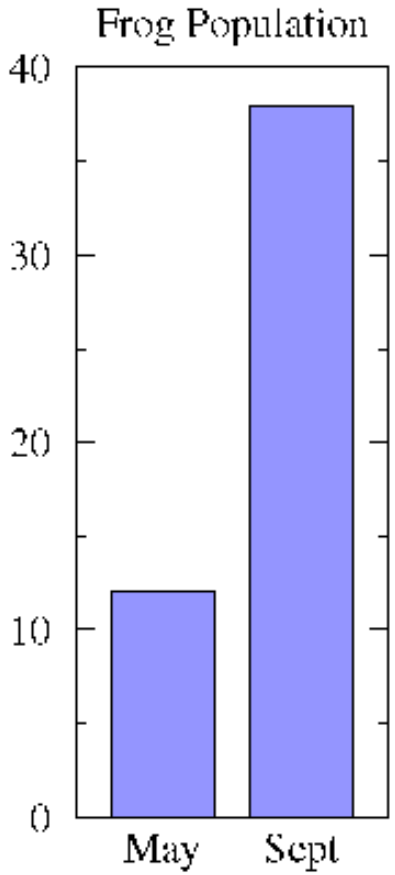
LYING WITH BAR CHARTS



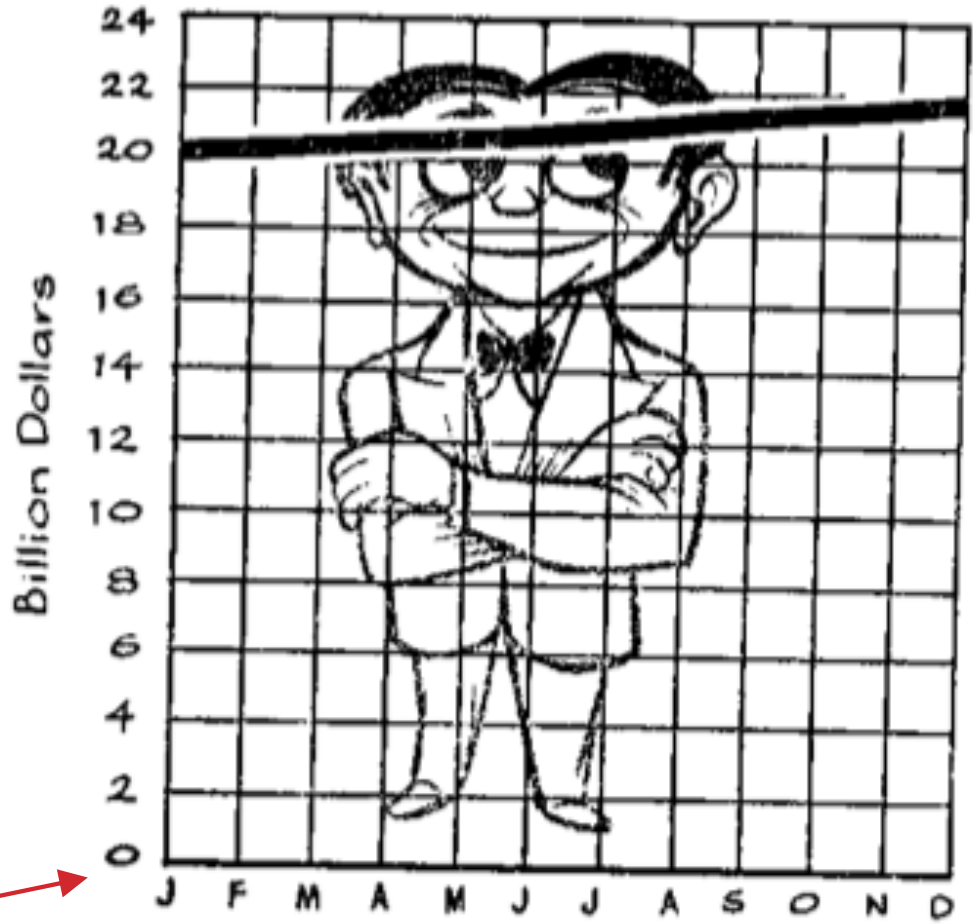
LYING WITH BAR CHARTS



VS

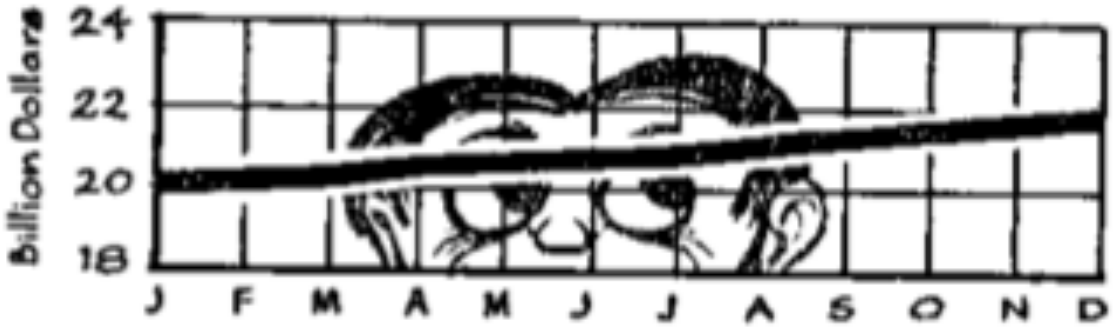


LYING WITH LINE CHART

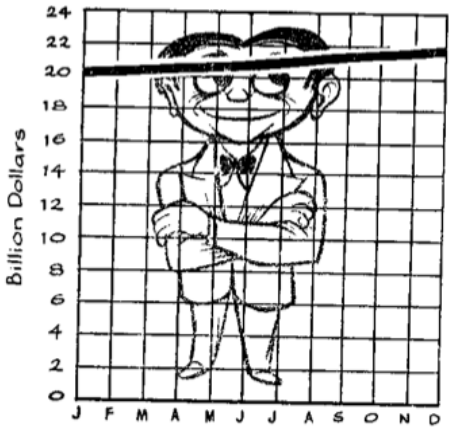


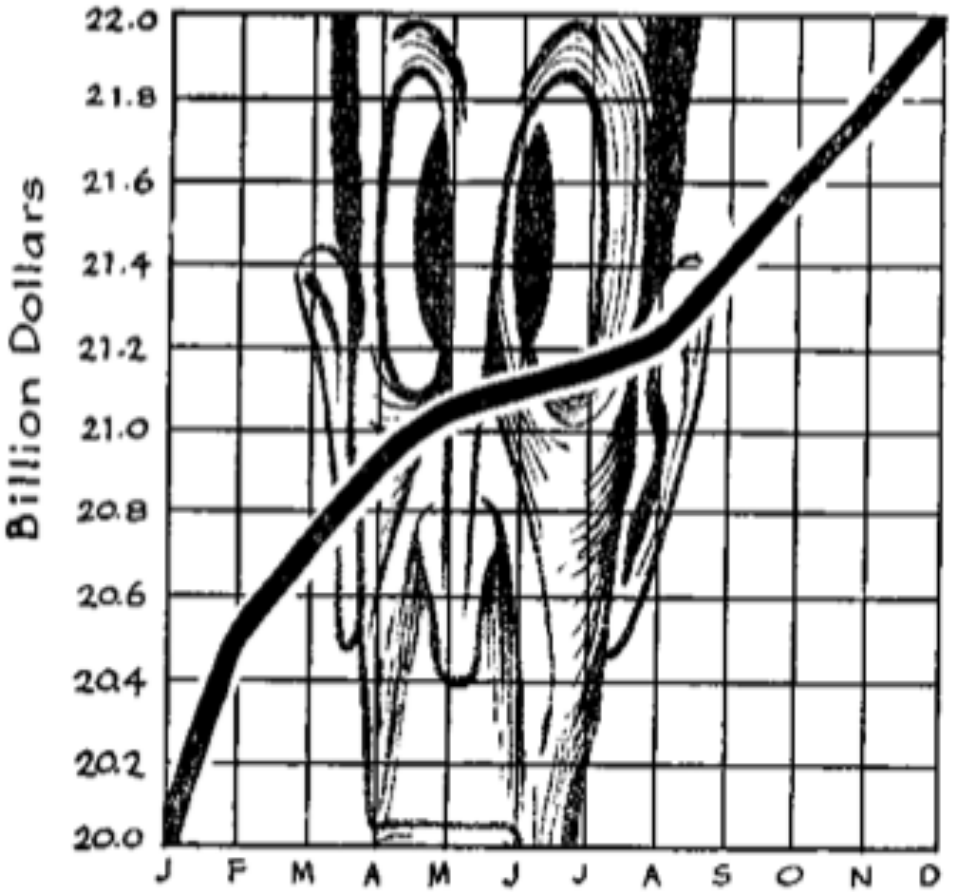
Zero line at the bottom

CHOP OFF THE BOTTOM



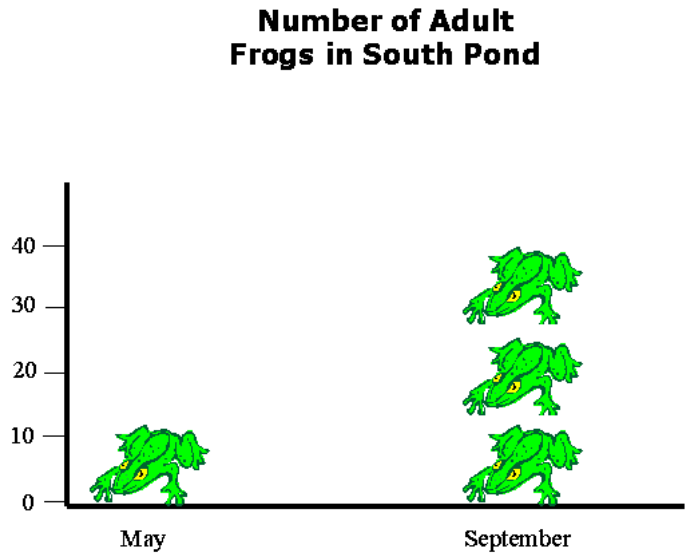
CHANGE THE PORTION OF Y-AXIS





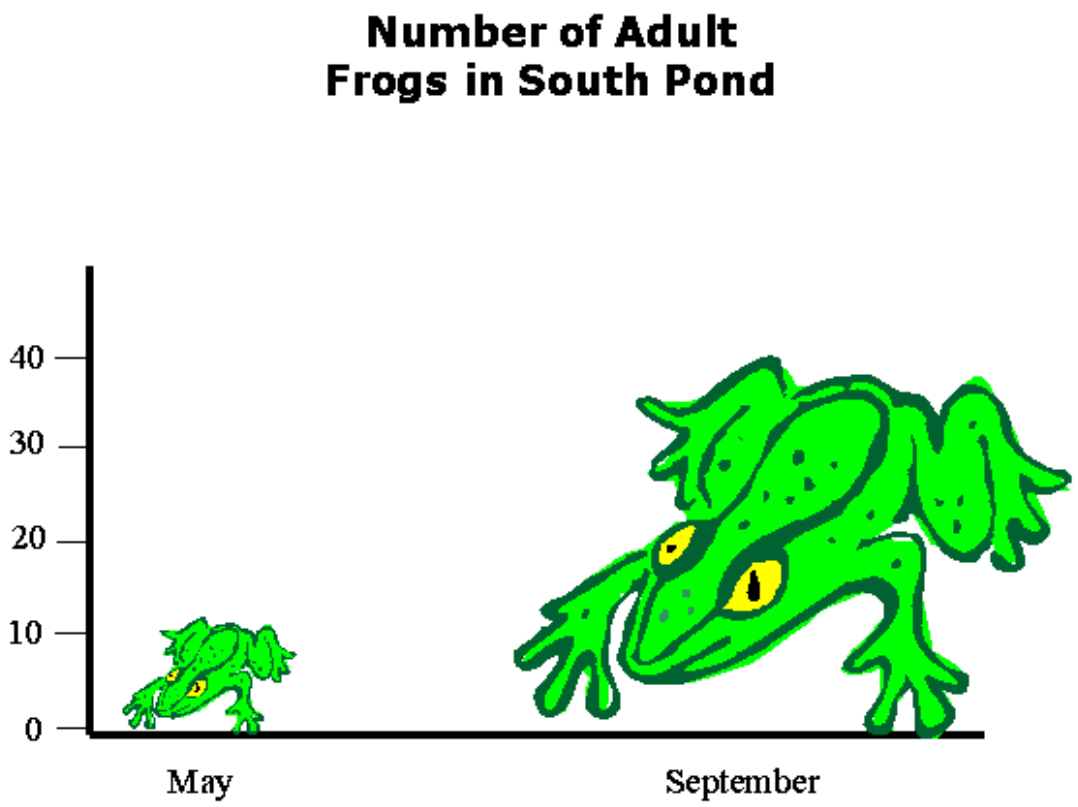
LYING WITH DIAGRAM

- Say that in a pond, there were
  - 13 Adult frogs in May
  - 39 Adult frogs in September
- Represented in a "stacked-frog" plot



LYING WITH DIAGRAM

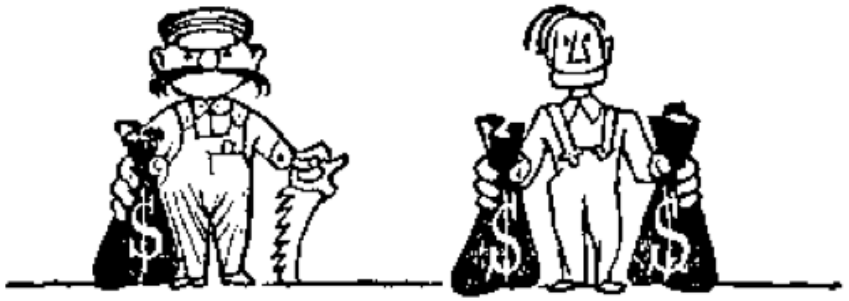
or we can represent in this way...



http://www.physics.csbsju.edu/stats/display.html

SOME MORE EXAMPLE

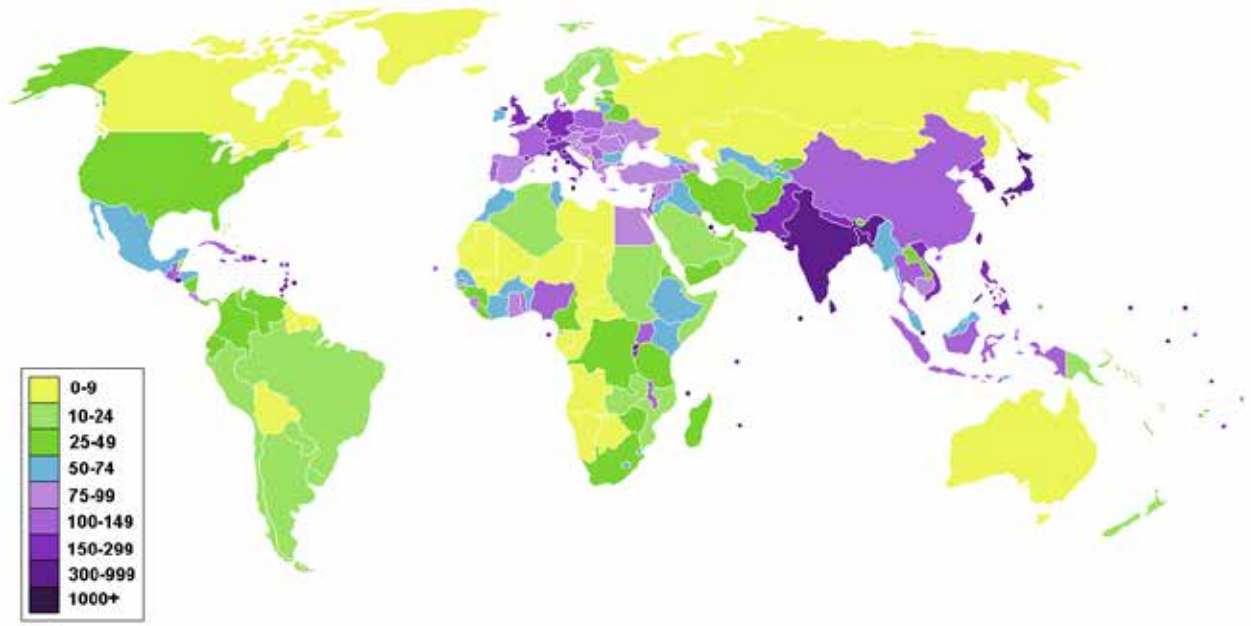
People at “A” get twice pay than people at “B”



LYING WITH MAPS

Choropleth map

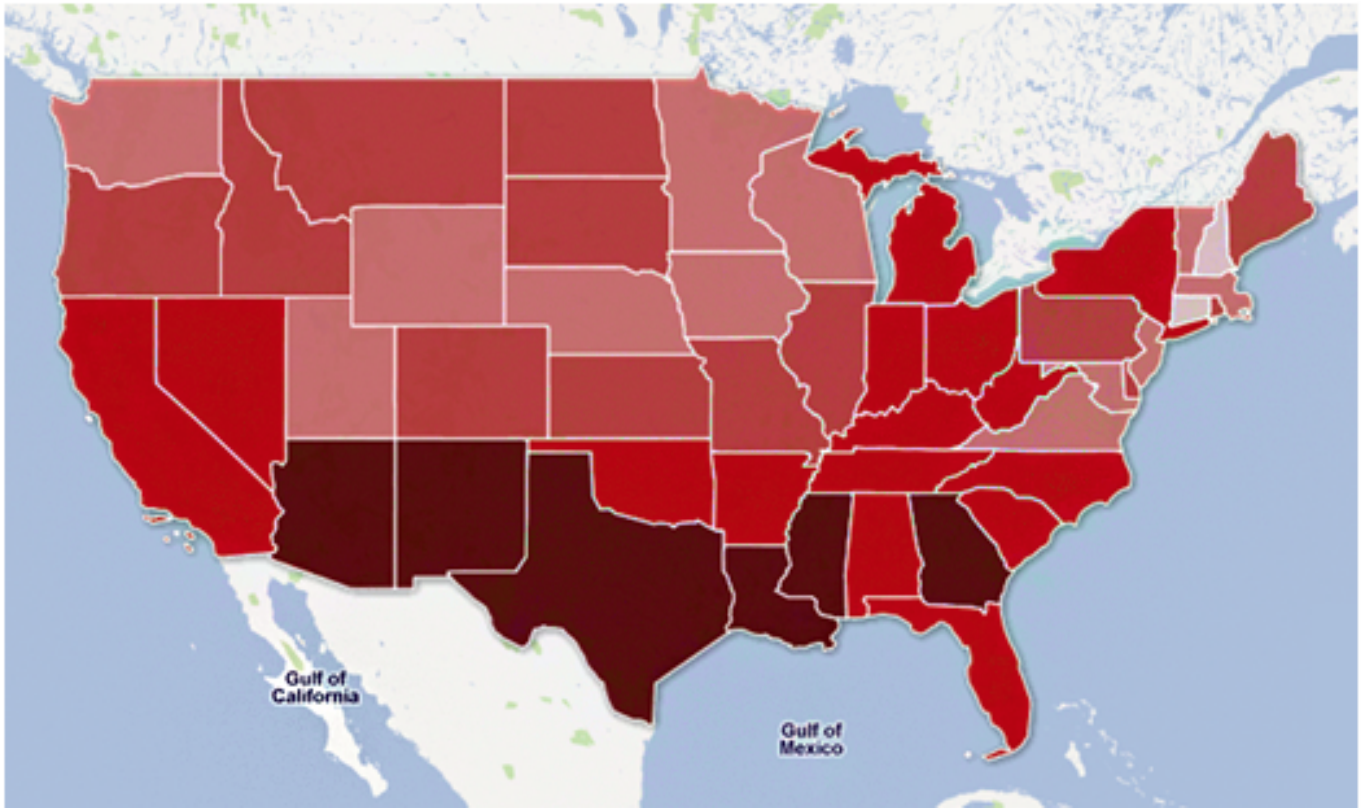
- provides an easy way to visualize measurement varies across geographic area



http://upload.wikimedia.org/wikipedia/commons/1/17/World_population_density_map.png

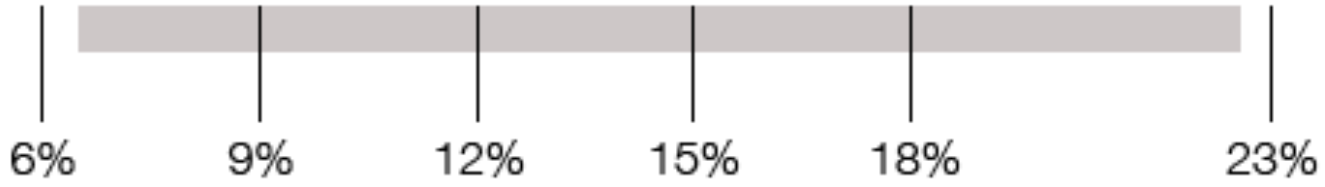
LYING WITH CHOROPLETH MAP

US poverty map from Guardian data blog

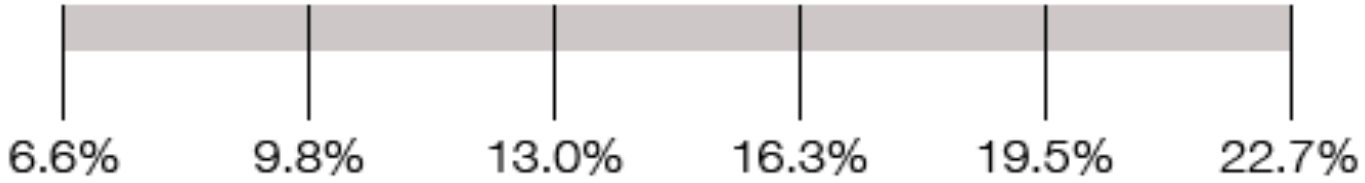


LYING WITH CHOROPLETH MAP

- Poverty data range from 6.6% to 22.7%
  - Unequally distributed

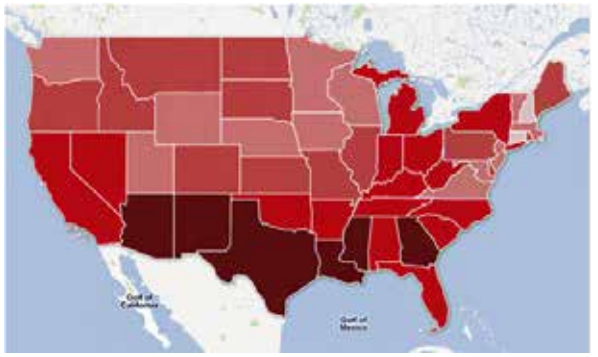


If we are measuring inequality, perhaps we should at least use equally distributed classes



http://vis4.net/blog/posts/choropleth-maps/

CHOICE OF COLOR

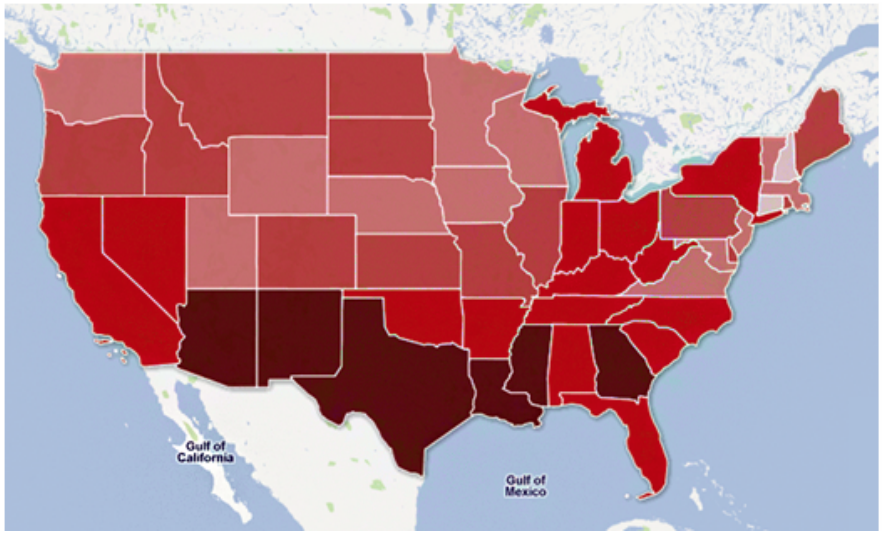


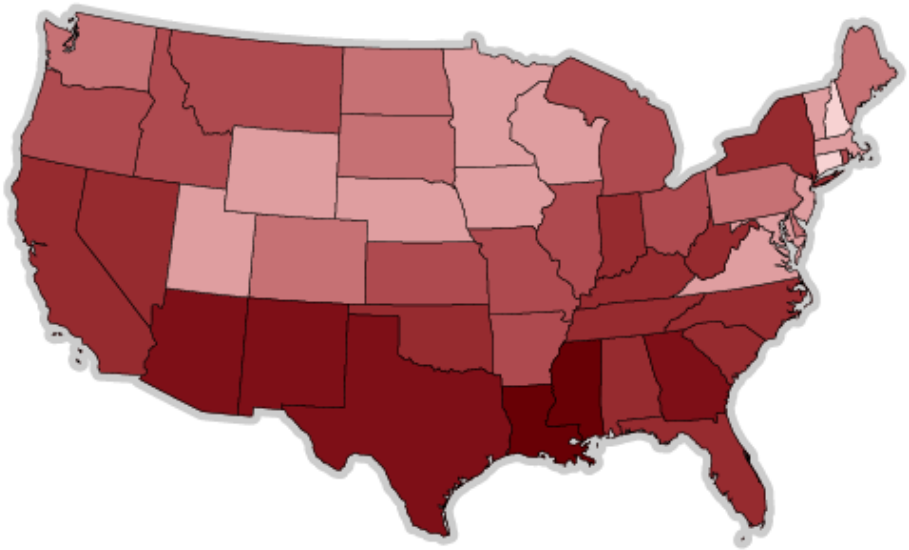
vs

http://vis4.net/blog/posts/choropleth-maps/

LYING WITH CHOROPLETH MAP

With equally distributed classes and equidistant colors from a HSV gradient



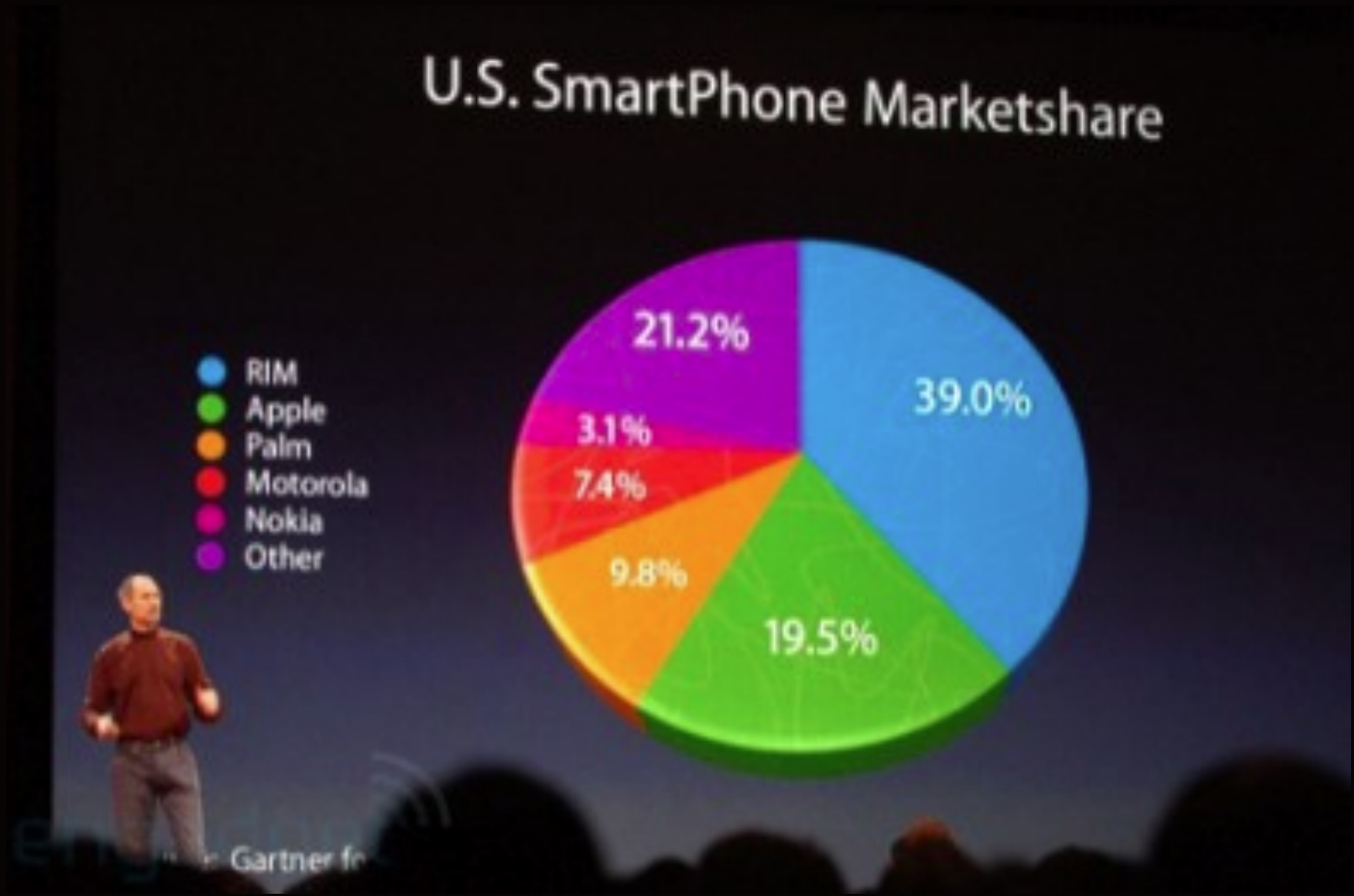


http://vis4.net/blog/posts/choropleth-maps/

LYING WITH CHOROPLETH MAP

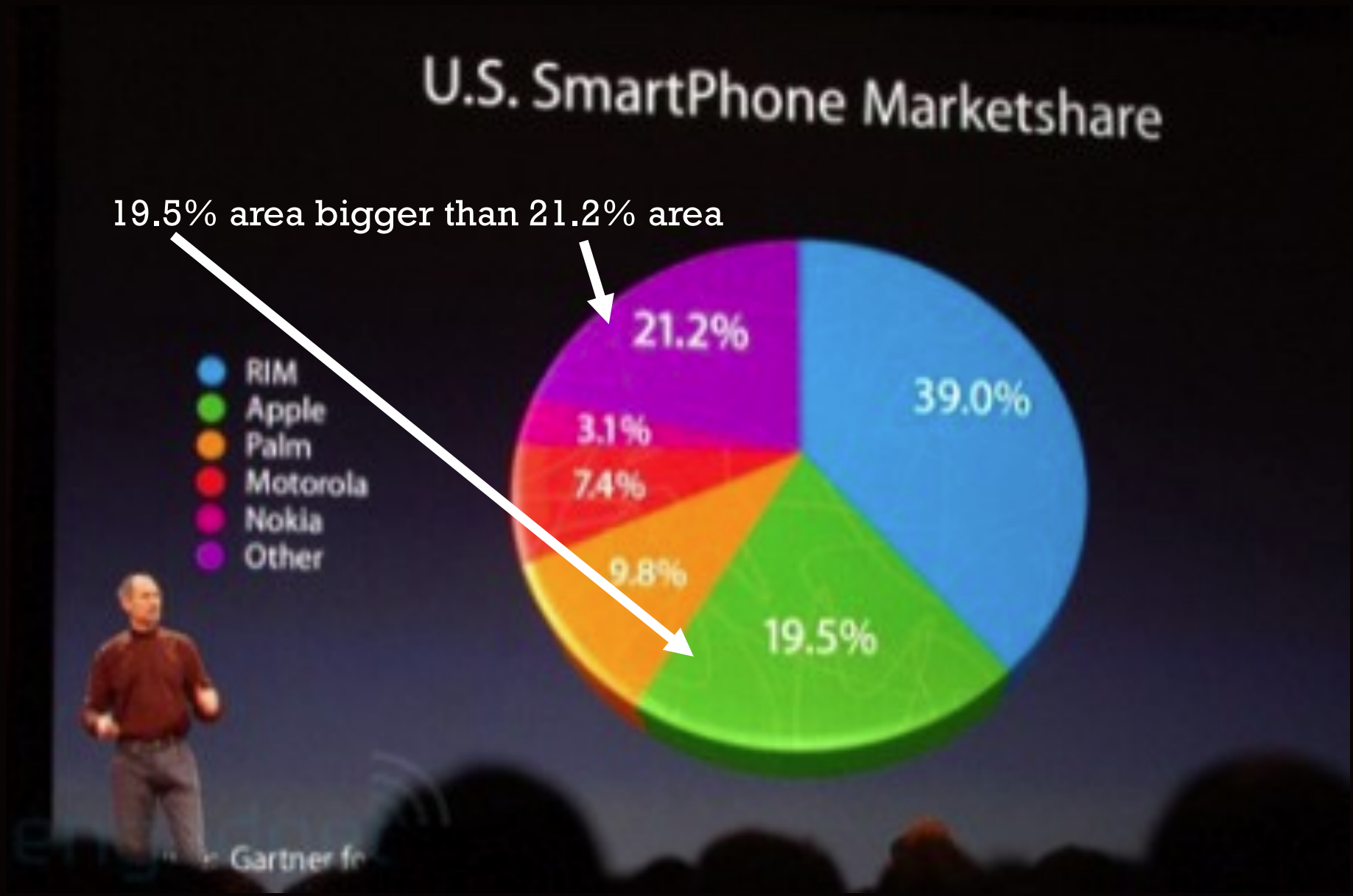
- When look at any choropleth map, be aware of
  - How they categorize the classes
  - How they choose the colors
- Choropleth map classification based on
  - Equal-intervals
  - quantile classing; each class has equal number quantity
  - Iterative algorithm to find “natural breaks”

TODAY'S EXAMPLE (APPLE WWDC 2008)



http://www.theguardian.com/technology/blog/2008/jan/21/iesdamnliesandstevejobs

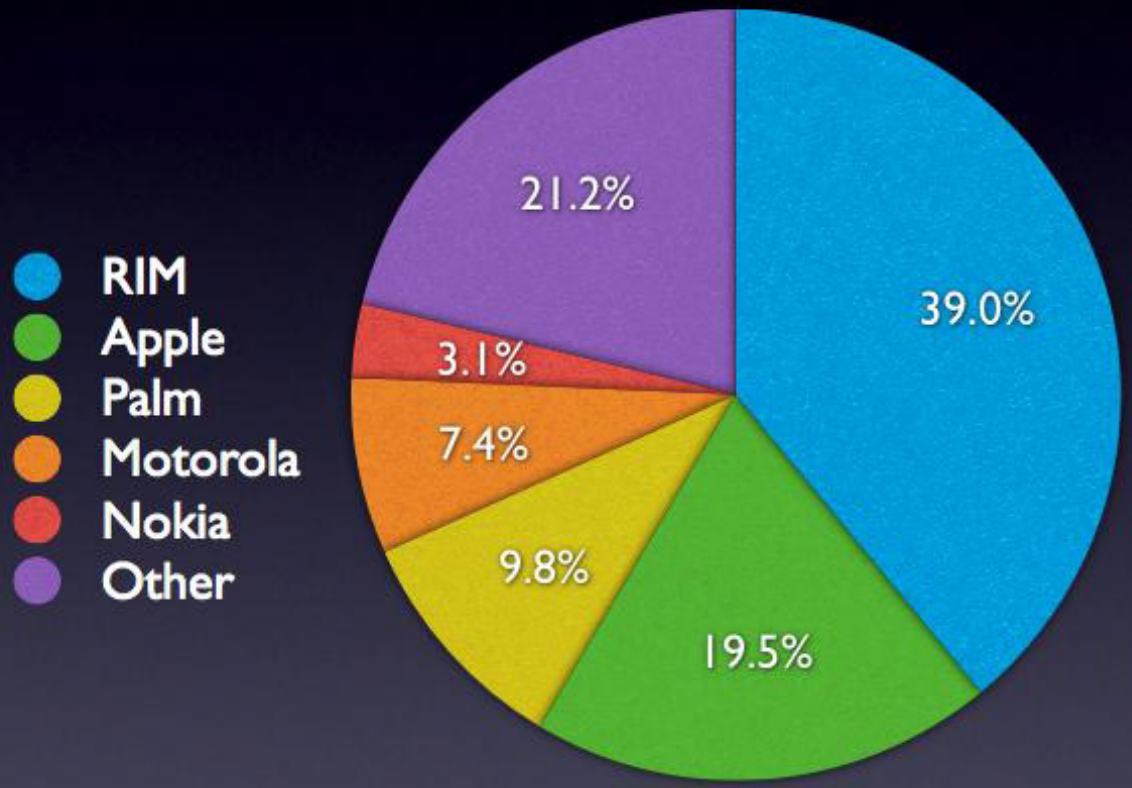
TODAY'S EXAMPLE (APPLE WWDC 2008)



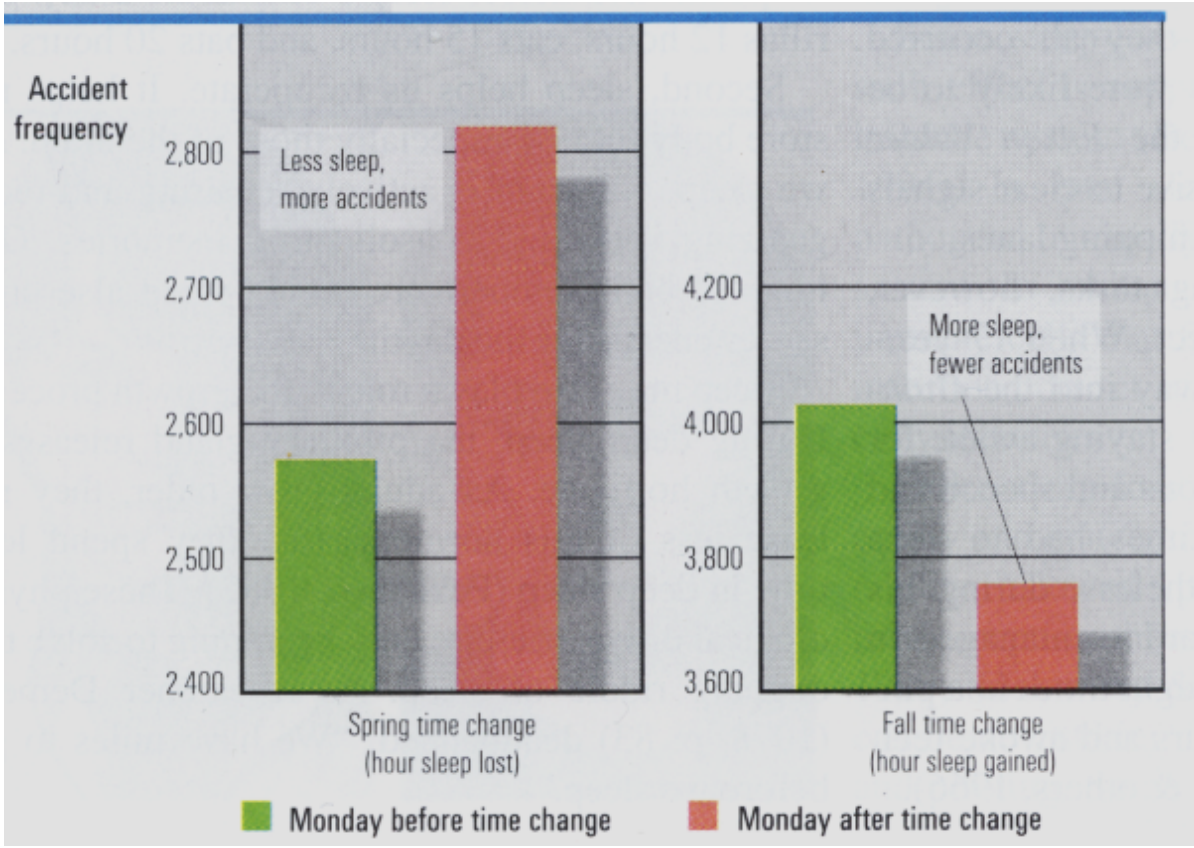
http://www.theguardian.com/technology/blog/2008/jan/21/liesdamnliesandstevejobs

TODAY'S EXAMPLE (APPLE WWDC 2008)

U.S. SmartPhone Marketshare



CLICKER



The above graphic was copied from a book about sleep research. The bars try to summarize the number of traffic accidents in Canada before and after daylight-savings time adjustments for the years 1991 and 1992 (combined). The goal of the graph is to suggest a correlation between lost sleep and traffic accidents.

Clicker: Find 3 problems with this visualization

LYING THROUGH AGGREGATIONS

	Smoker?	
	Yes	No
Dead	107	132
Alive	174	175
Total	281	307
% Dying	38.1%	43.0%

(data adapted from Appleton et al. 1996, Am. Stat.)

MOTIVATING EXAMPLE: SMOKING & SURVIVAL

20-year follow-up study, Wickham in UK (Tunbridge et al. 1977)

1972-1974, one-in-six survey of the electoral roll, largely concerned with thyroid disease and heart disease

For simplicity, consider women aged 45 to 75 at the start of the study

- Smoking status: current smoker (Y/N)
- 20-year survival info: determined for all women in the study

	Smoker?	
	Yes	No
Dead	107	132
Alive	174	175
Total	281	307
% Dying	38.1%	43.0%

Protective effect of smoking?

(data adapted from Appleton et al. 1996, Am. Stat.)

SMOKING & SURVIVAL (CON'T)

	Age Group					
	45-54		55-64		65-74	
	Smoker?		Smoker?		Smoker?	
	Yes	No	Yes	No	Yes	No
Dead	27	12	51	40	29	101
Alive	103	66	64	81	7	28
Total	130	78	105	121	36	129
% Dying	20.8%	15.4%	48.6%	33.1%	80.6%	78.3%

Consider 10-year ranges: 45-54,55-64,65-75

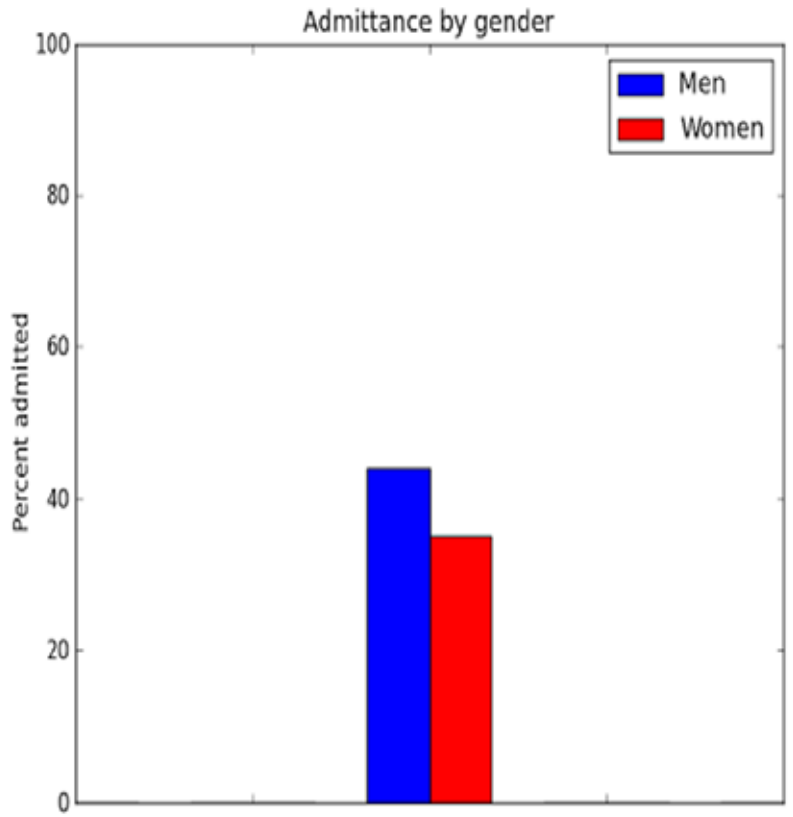
Non-smoking group does better in each case!

GENDER BIAS, OR NOT?

1973, UC Berkeley was afraid to be sued for discrimination against women in graduate school admissions

Percent acceptance: Male vs Female,

44% vs. 35%



GENDER BIAS, OR NOT? (CONT'D)

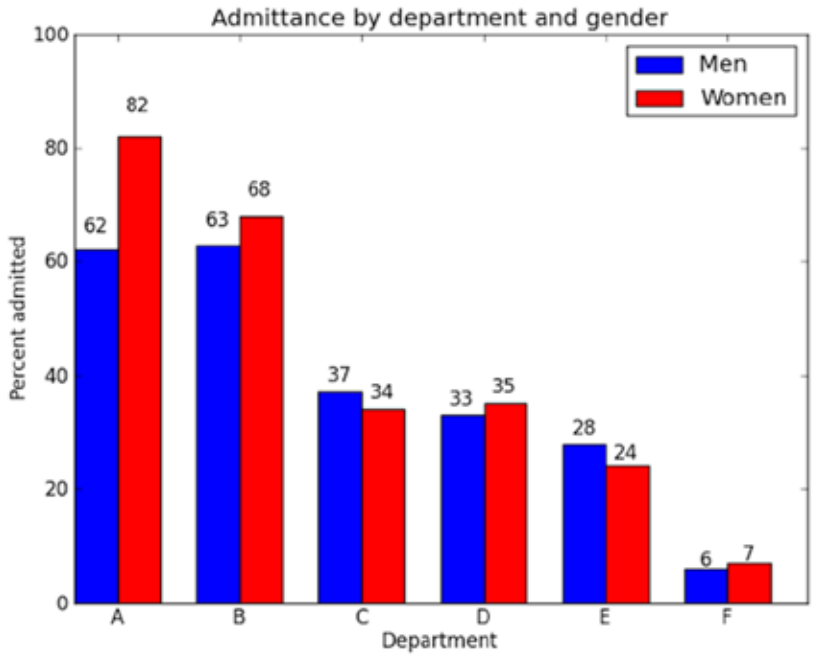
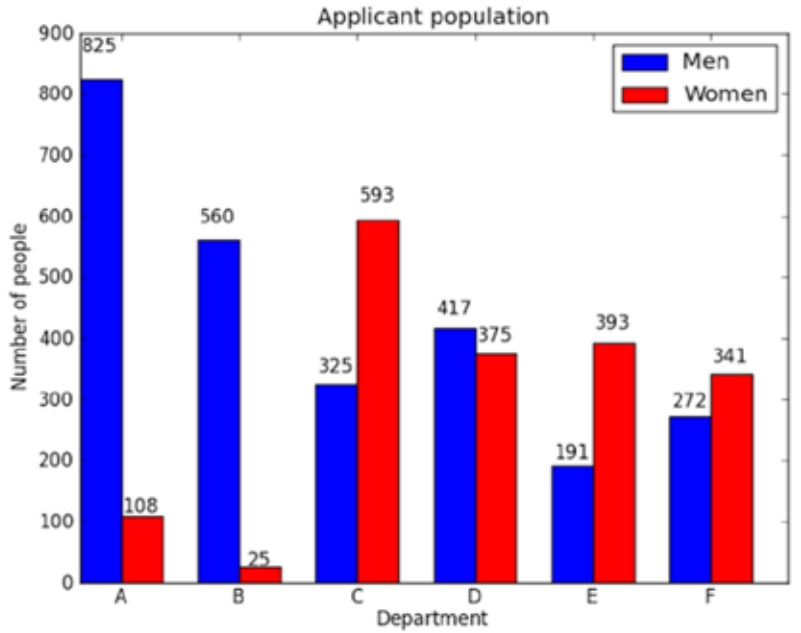


Table 2. Admissions data by sex of applicant for two hypothetical departments. For total, $\chi^2 = 5.71$, d.f. $= 1$, $P = 0.19$ (one-tailed).

Applicants	Outcome				Difference	
	Observed		Expected			
	Admit	Deny	Admit	Deny	Admit	Deny
	Department of machismatics					
Men	200	200	200	200	0	0
Women	100	100	100	100	0	0
	Department of social warfare					
Men	50	100	50	100	0	0
Women	150	300	150	300	0	0
	Totals					
Men	250	300	229.2	320.8	20.8	− 20.8
Women	250	400	270.8	379.2	− 20.8	20.8



P. J. Bickel, E. A. Hammel, J. W. O'Connell. (1975). Sex Bias in Graduate Admissions: Data from Berkeley. Science 187, (4175). pp. 398-404

SIMPSON PARADOX

If we have

$$\frac{a}{b} < \frac{A}{B} \text{ and } \frac{c}{d} < \frac{C}{D},$$

is it also true that

$$\frac{a+c}{b+d} < \frac{A+C}{B+D}?$$

Not necessarily! Note that

$$\frac{1}{3} < \frac{3}{8} \text{ and } \frac{5}{8} < \frac{2}{3}$$

but

$$\frac{1+5}{3+8} > \frac{3+2}{8+3}$$

Be aware of the dangers of ignoring a covariate that is correlated to an outcome variable and an explanatory one.

Simpson, E.H. (1951). "The interpretation of Interaction in Contingency Tables", *Journal of the Royal Statistical Society*, B, 13, 238-241.

MORE RECENT EXAMPLE

TechCrunch

Google found it paid men less than women for the same job

by Megan Rose Dickey https://techcrunch.com/2019/03/04/google-found-it-paid-men-less-than-women-for-the-same-job/

Wired

Are men at Google Paid less than women? Not Really

by Natasha Tiku https://www.wired.com/story/men-google-paid-less-than-women-not-really/

IMPRESSIVE FIGURES

UP-TOS

"runs up to 10x faster"

(https://www.digitalengineering247.com/article/altair-optistructruns-up-to-10x-faster-on-nvidia-gpus)

"lasts up to 5x longer"

(https://ca.crest.com/en-ca/products/crest-complete-whitening-plus-scope-outlast-toothpaste)

"cleans up to 10x better"

(https://www.youtube.com/watch?v=Yx9iCKKzYR4)

"Schism consistently outperforms simple partitioning schemes, …, reducing the cost of distributed transactions **up to 30%**"

Carlo Curino, Yang Zhang, Evan P. C. Jones, Samuel Madden: *Schism: a Workload-Driven Approach to Database Replication and Partitioning*. PVLDB 3(1): 48-57 (2010)

What is the problem with up-tos?

Example (from Colton)
Sex and race distribution of 158 cases of *abdominal aortic aneurysms (AAA)* at metropolitan hospitals in a Southern city

Sex & Race	#AAA
White Males	93
AA Males	30
White Females	22
AA Females	13

Author's conclusion: Incidence of AAA is almost 3 times more frequent in Whites than African-Americans.

Clicker: Do you see a potential problem?

Example (from Colton)
Sex and race distribution of 158 cases of *abdominal aortic aneurysms (AAA)* at metropolitan hospitals in a Southern city

Sex & Race	#AAA
White Males	93
AA Males	30
White Females	22
AA Females	13

Author's conclusion: Incidence of AAA is almost 3 times more frequent in Whites than African-Americans.

Do you see a potential problem?

This fallacy is known as a lack of denominators

EXCEPTION FALLACY

4 out of 6 members of the math team representing Canada at the 2018 International Math Olympiad were from Ontario.

Clicker: Does Ontario have the best K-12 math curriculum in Canada?

a) Yes

b) No

c) Impossible to say

d) Scooby-doo

EXCEPTION FALLACY

4 out of 6 members of the math team representing Canada at the 2018 International Math Olympiad were from Ontario.

Does Ontario have the best K-12 math curriculum in Canada?

Note that Ontario has 40% of the nation's population.

DO YOU SEE A PROBLEM WITH THIS LIST?

Example: The Top 10 All Time Grossing Films (in Millions – US)

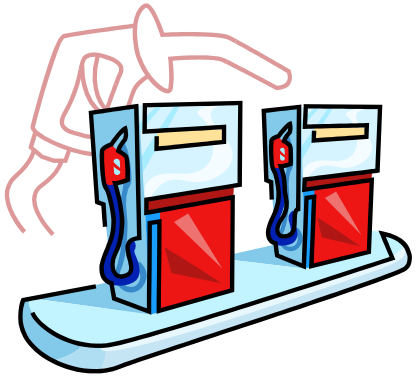
1) **Avatar (2009): $760**
2) **Titanic(1997): $658**
3) **Marvel's the Avengers (2012): $588**
4) **The Dark Knight (2008): $533**
5) **Star Wars I: The Phantom Menace (1999) $474**
6) **Star Wars IV: A New Hope (1977): $460**
7) **The Dark Knight Rises (2012) $449**
8) **Shrek 2 (2011): $441**
9) **E.T. The Extra-Terrestrial (1982): $435**
10) **The Hunger Games: Catching Fire (2013): $424**

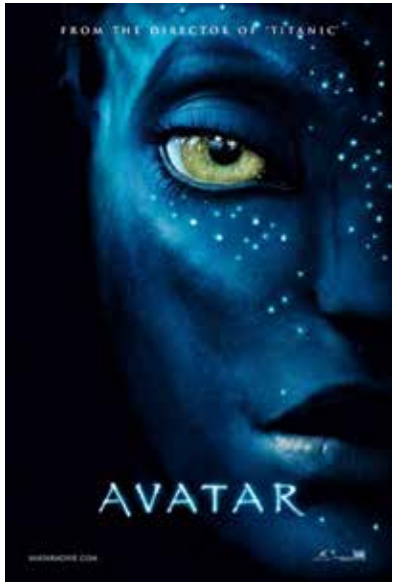
REAL VS. NOMINAL VARIABLES

Nominal Variables are in terms of a current dollars. For example, you're starting salary after college might be $50,000 per year.

Real variables are in terms of some fixed commodity. Real variables measure purchasing power. If a gallon of gas costs $2.00, then we can calculate your "real" income.

$$\textbf{Real Income} = \frac{\textbf{Nominal Income}}{\textbf{Price}} = \frac{\textbf{\$50,000}}{\textbf{\$2.00}} = \textbf{25,000}$$





In 2009, a gallon of gas cost $3.50

$$\text{Real Gross} = \frac{\text{Nominal Gross}}{\text{Price}} = \frac{\$749\text{M}}{\$3.50} = 214\text{M}$$

(Gallons of Gas)

In 1977, a gallon of gas cost $.62

$$\text{Real Income} = \frac{\text{Nominal Gross}}{\text{Price}} = \frac{\$460\text{M}}{\$.62} = 742\text{M}$$

(Gallons of Gas)

The Top 10 All Time Grossing Films– **Inflation Adjusted** (Millions of 2000 Dollars)

1) Gone With the Wind (1939): $1,689
2) Star Wars Episode IV(1977): $960
3) The Sound of Music(1965): $768
4) ET: The Extraterrestrial(1982): $764
5) The Ten Commandments (1956): $706
6) Titanic (1997): $691
7) Jaws (1975): $690
8) Dr. Zhivago (1965): $669
9) The Exorcist (1973): $596
10) Snow White (1937): $587

Notes: Avatar falls to #14 ($516), a movie ticket in 1939 was $0.23

SELF-DRIVING CARS

[Tesla] said Autopilot-enabled cars had covered 130 million miles without a fatality, compared to a national average of one fatality every 94 million miles. **Musk says it would be "morally reprehensible" to delay its rollout.**

Tesla's Cars Have Driven 140M Miles on Autopilot. Here's How --- Wired, 17 Aug 2016

Clicker: what is the problem with this statement?

SELF-DRIVING CARS

[Tesla] said Autopilot-enabled cars had covered 130 million miles without a fatality, compared to a national average of one fatality every 94 million miles. **Musk says it would be "morally reprehensible" to delay its rollout.**

Tesla's Cars Have Driven 140M Miles on Autopilot. Here's How --- Wired, 17 Aug 2016

A RAND Corporation report concluded that fatalities and injuries are so rare that it would require an automated car to drive as many as hundreds of billions of miles before its performance could be fairly compared with statistics from the much larger population of human drivers.

https://www.technologyreview.com/s/601849/teslas-dubious-claims-about-autopilots-safety-record/

SERVICE UP-TIME

A fictitious school bus status update website claims 99.9% uptime.

Is this good?

SERVICE UP-TIME

A fictitious school bus status update website claims 99.9% uptime.

Is this good?

What if in the morning hours (5am - 8am) of a big snowstorm day, the website is down due to too much traffic.

With three such snowstorms a year, the website is down 9 hours out of a total of 365 x 24 hours per year.

9/8760 is roughly 0.1%. But the website is down when you most need it to be up!

P-VALUE

A New Study shows: A Glass Of Red Wine Is The Equivalent To An Hour At The Gym [Fox News 02/15 and others]

http://www.huffingtonpost.co.uk/2016/01/08/a-glass-of-red-wine-is-the-equivalent-to-an-hour-at-the-gym-says-new-study_n_7317240.html

A new study shows: Secret to winning a Nobel prize?
Eat More Chocolate [Time 10/12]

Scientists find the secret of longer life for men

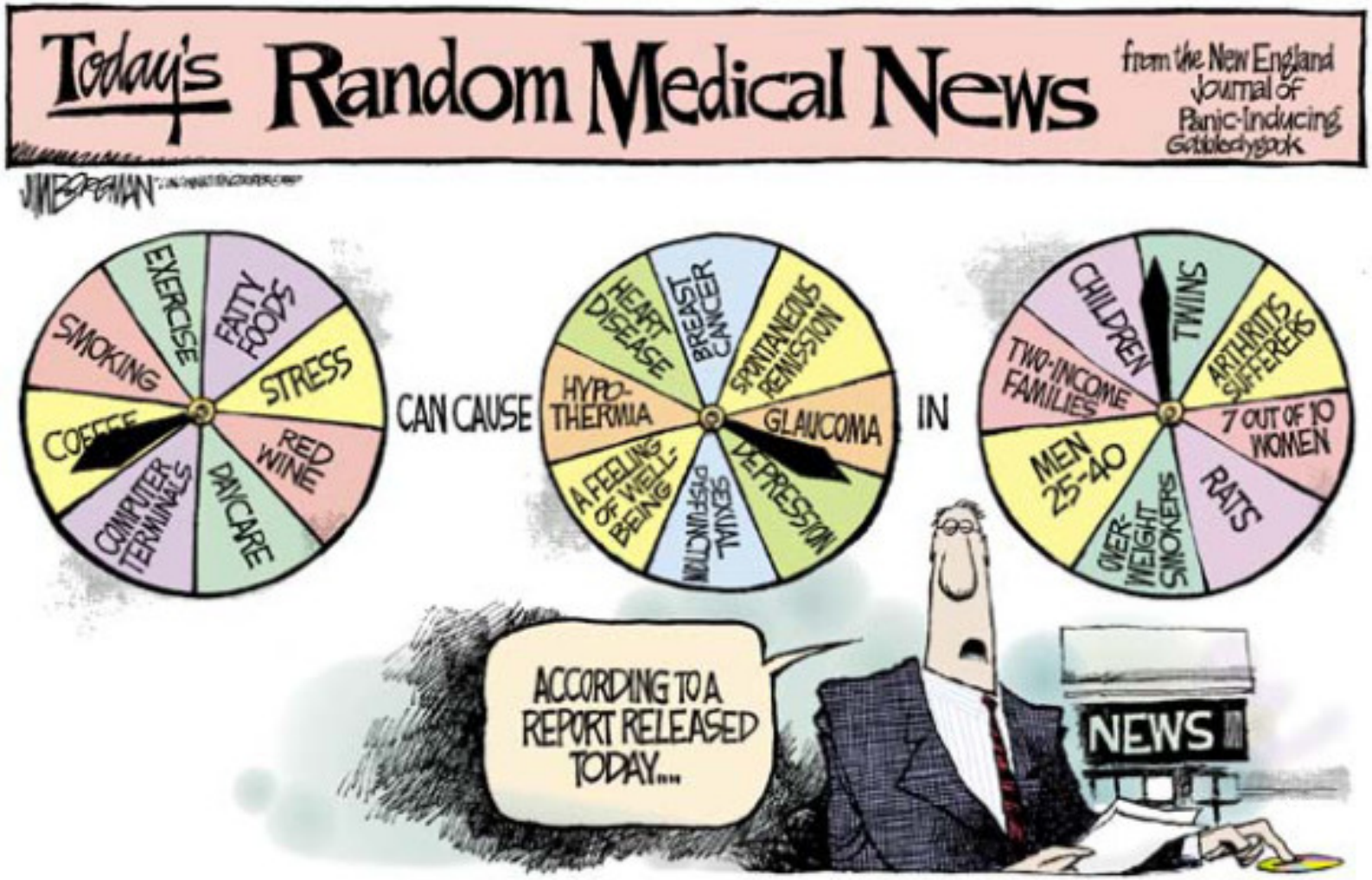
[Daily Mail UK, 09/12]

http://www.dailymail.co.uk/sciencetech/article-2207981/Scientists-secret-living-life-men-bad-news-Castration-key.html

Scientists find the secret of longer life for men (The bad news: castration is the key) [Daily Mail UK, 09/12]

http://www.dailymail.co.uk/sciencetech/article-2207981/Scientists-secret-living-life-men-bad-news-Castration-key.html

Data Dredging



STATISTICAL TEST

Beer Consumption Increases Human Attractiveness to Malaria Mosquitoes

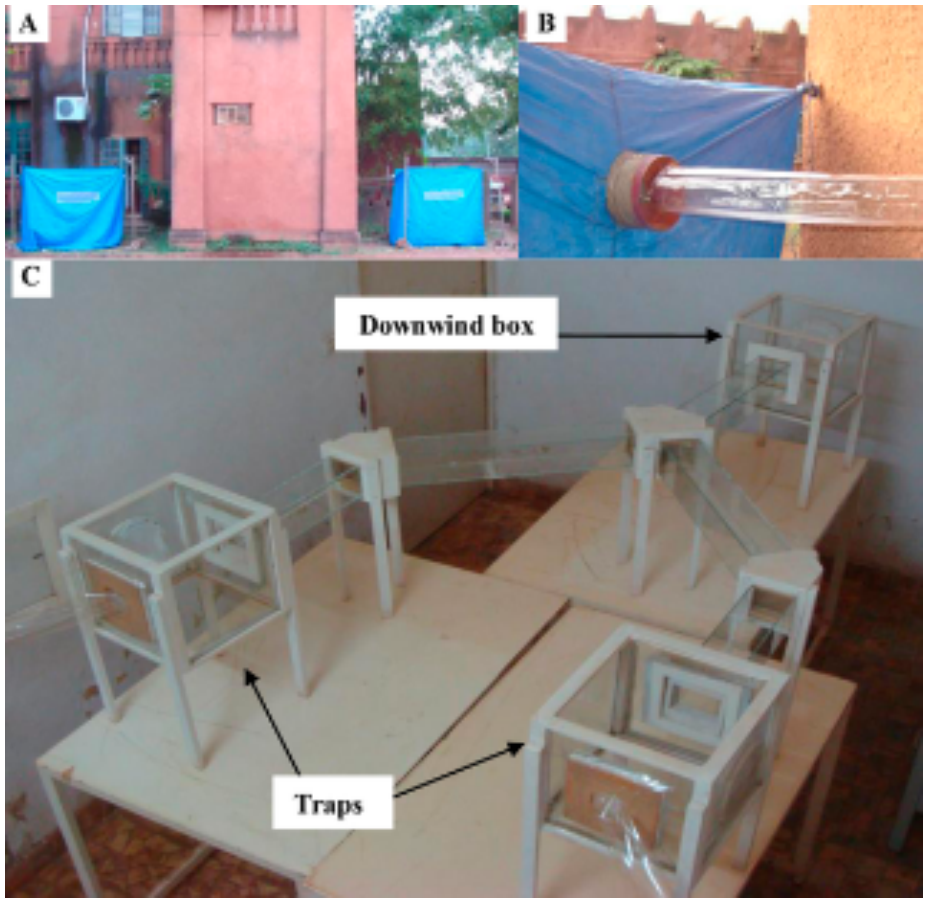
Beer (25):

27 20 21 26 27 31 24 21 20 19
23 24 28 19 24 29 18 20 17 31
20 25 28 21 27
Mean: 23.6

Water (18):

21 22 15 12 21 16 19 15 22 24
19 23 13 22 20 24 18 20
Mean: 19.2

Is a difference of 4.4 significant?



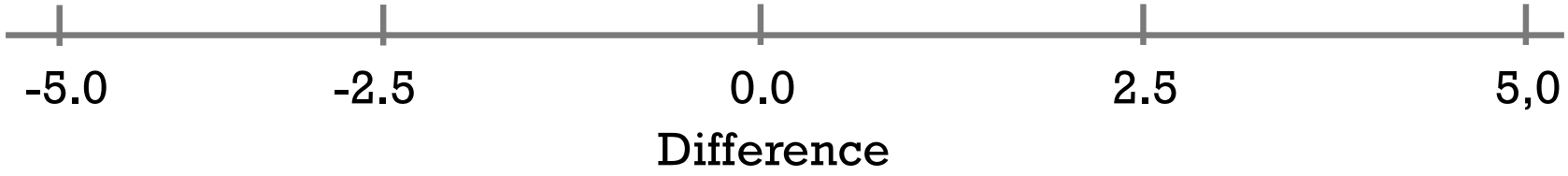
Beer Consumption #Human Attractiveness to Malaria Mosquitoes
Thierry Lefèvre , Louis-Clément Gouagna, Kounbobr Roch Dabiré, Eric Elguero, Didier Fontenille, François Renaud, Carlo Costantini, Frédéric Thomas

PERMUTATION TEST

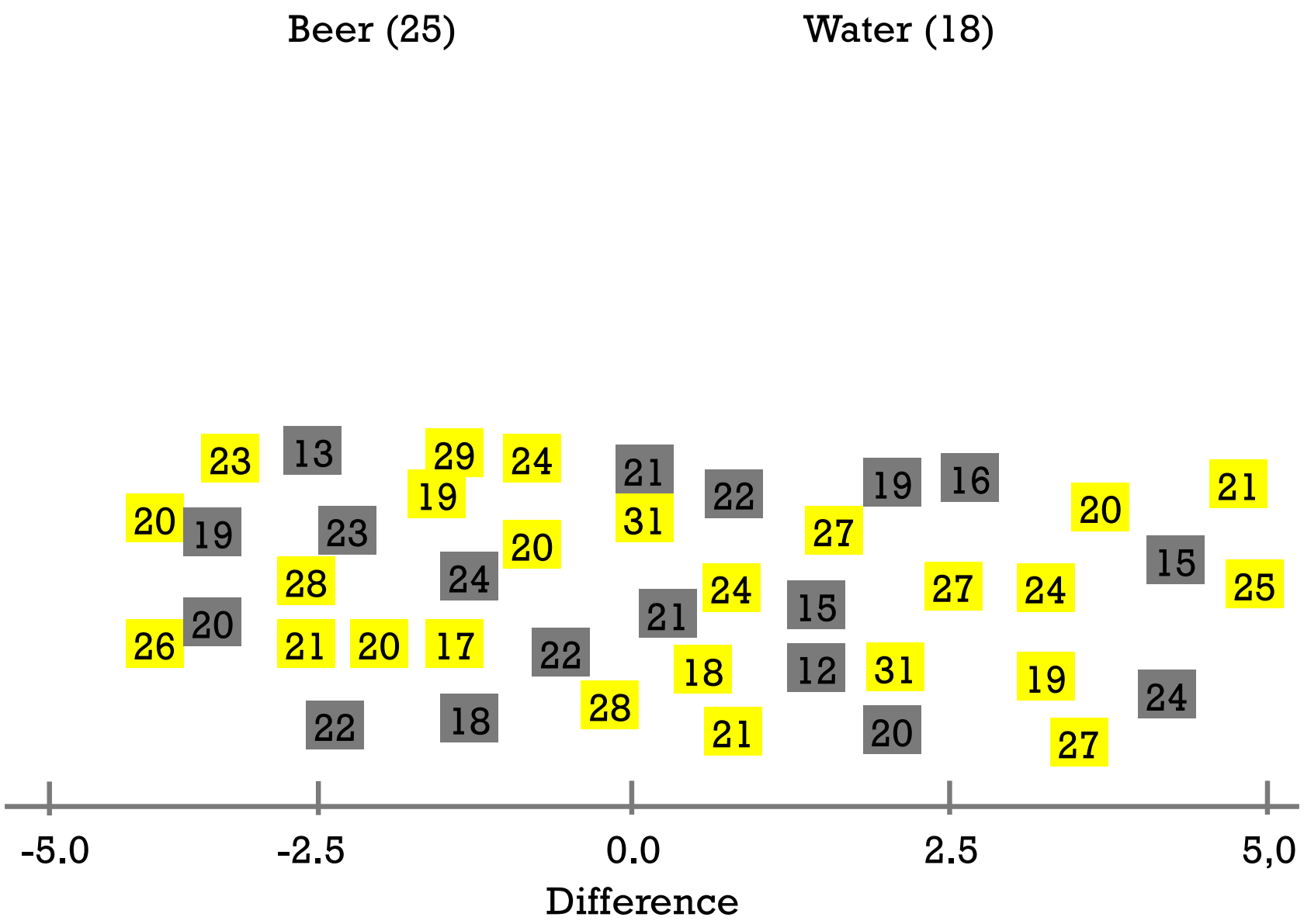
Beer (25)				
27	23	20	31	29
20	24	25	24	18
21	28	28	21	20
26	19	21	20	17
27	24	27	19	31

Water (18)			
21	19	16	24
22	23	19	18
15	13	15	20
12	22	22	
21	20	24	

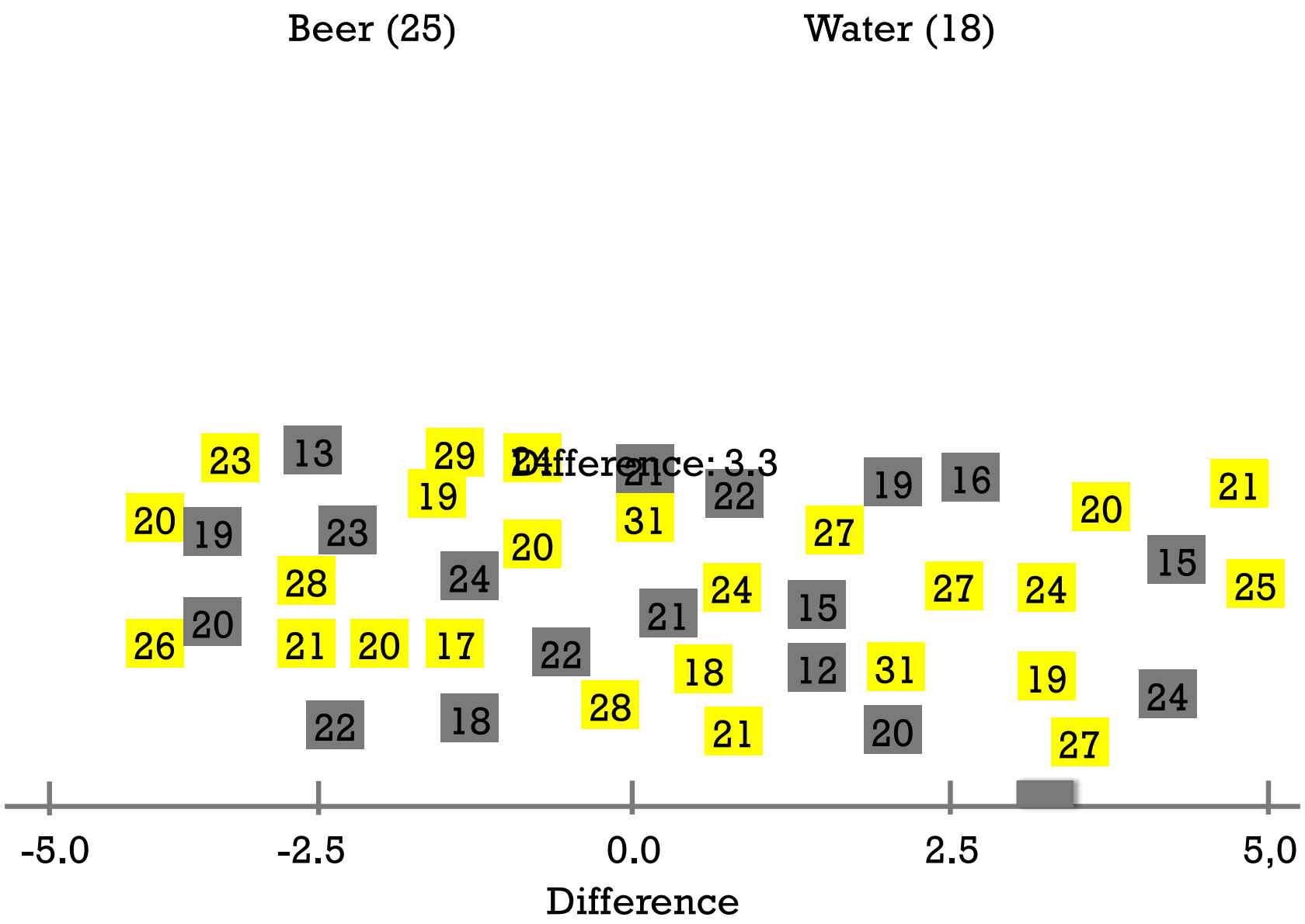
Difference: 4.4



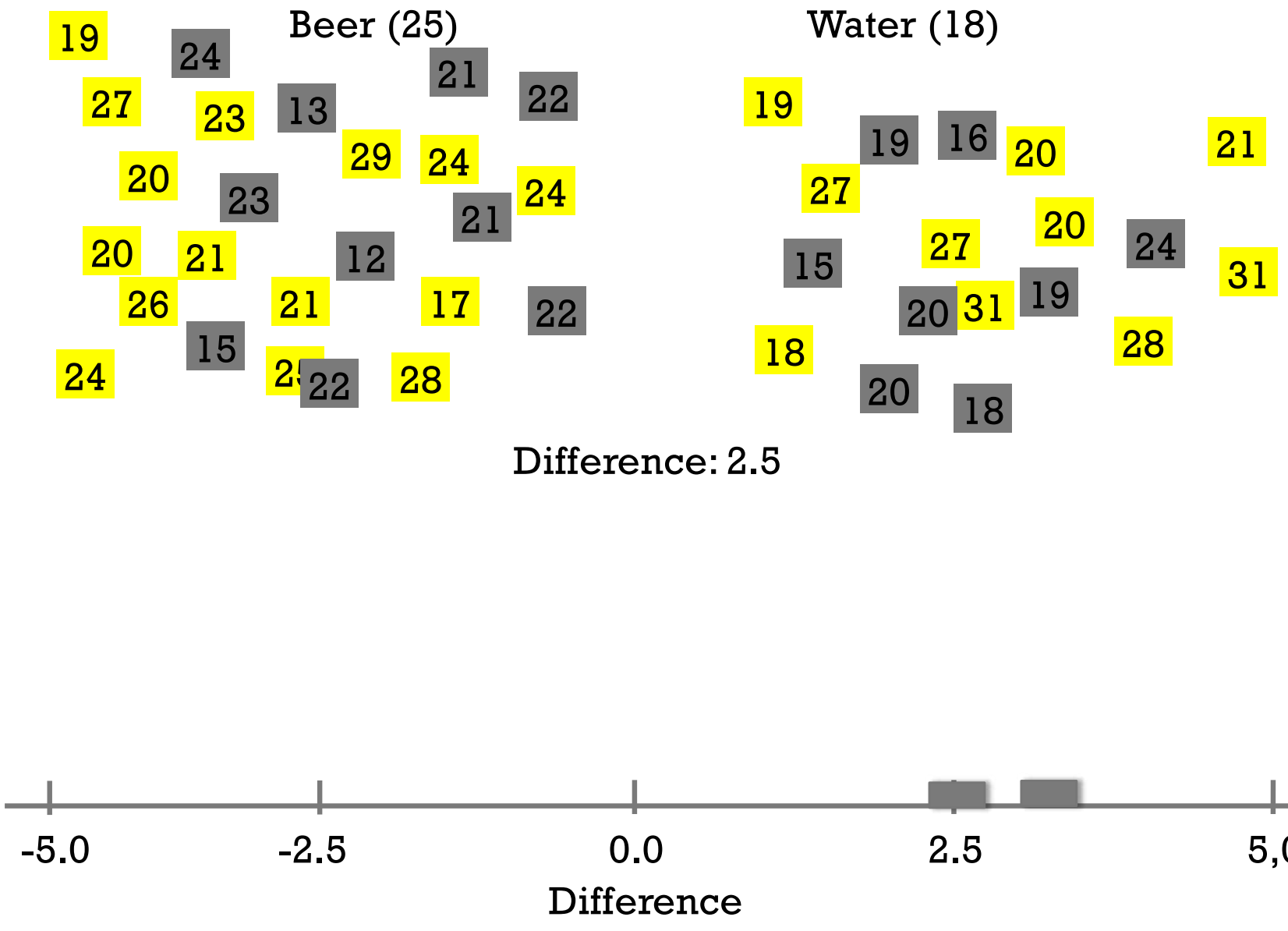
PERMUTATION TEST



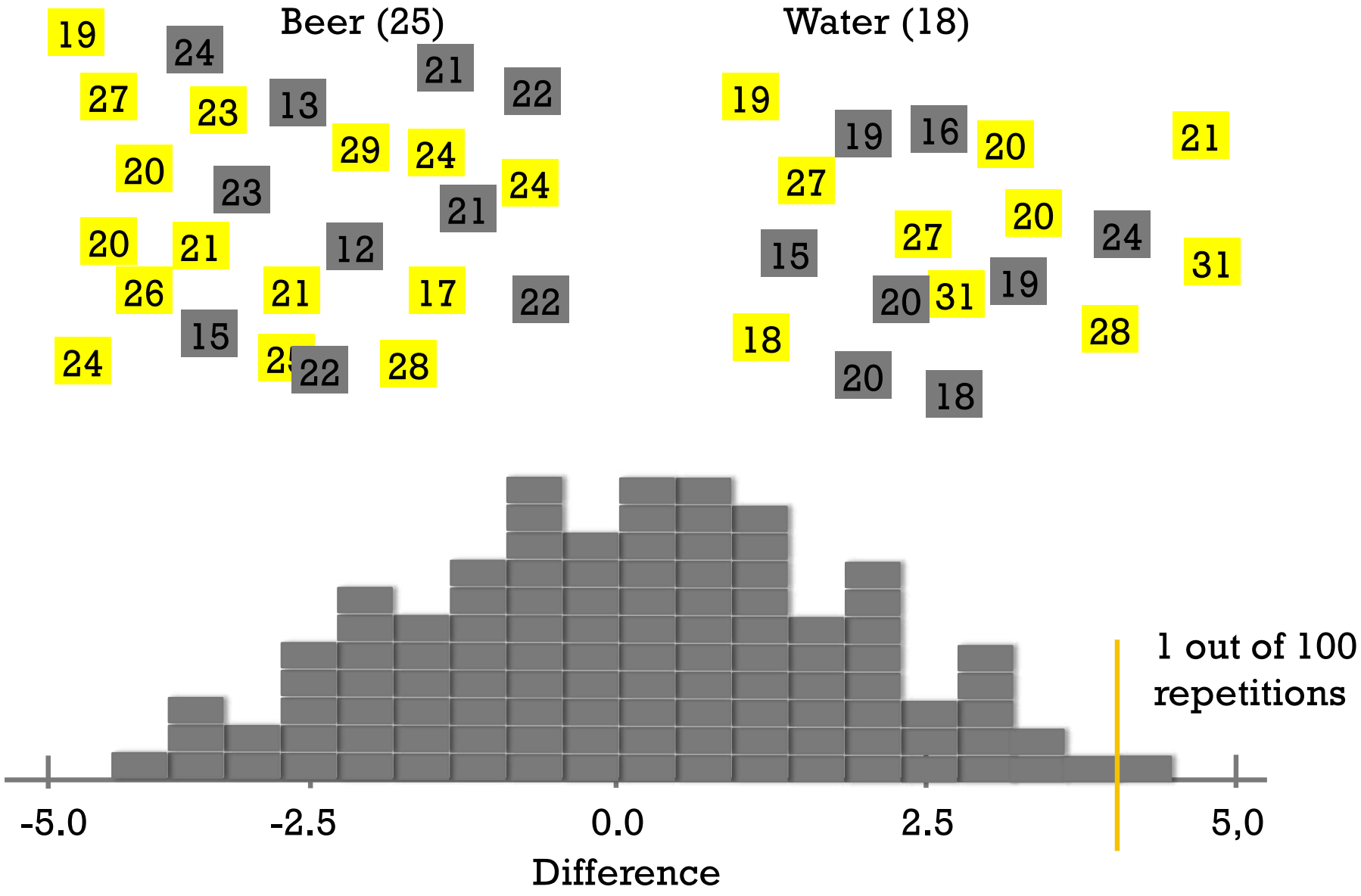
PERMUTATION TEST



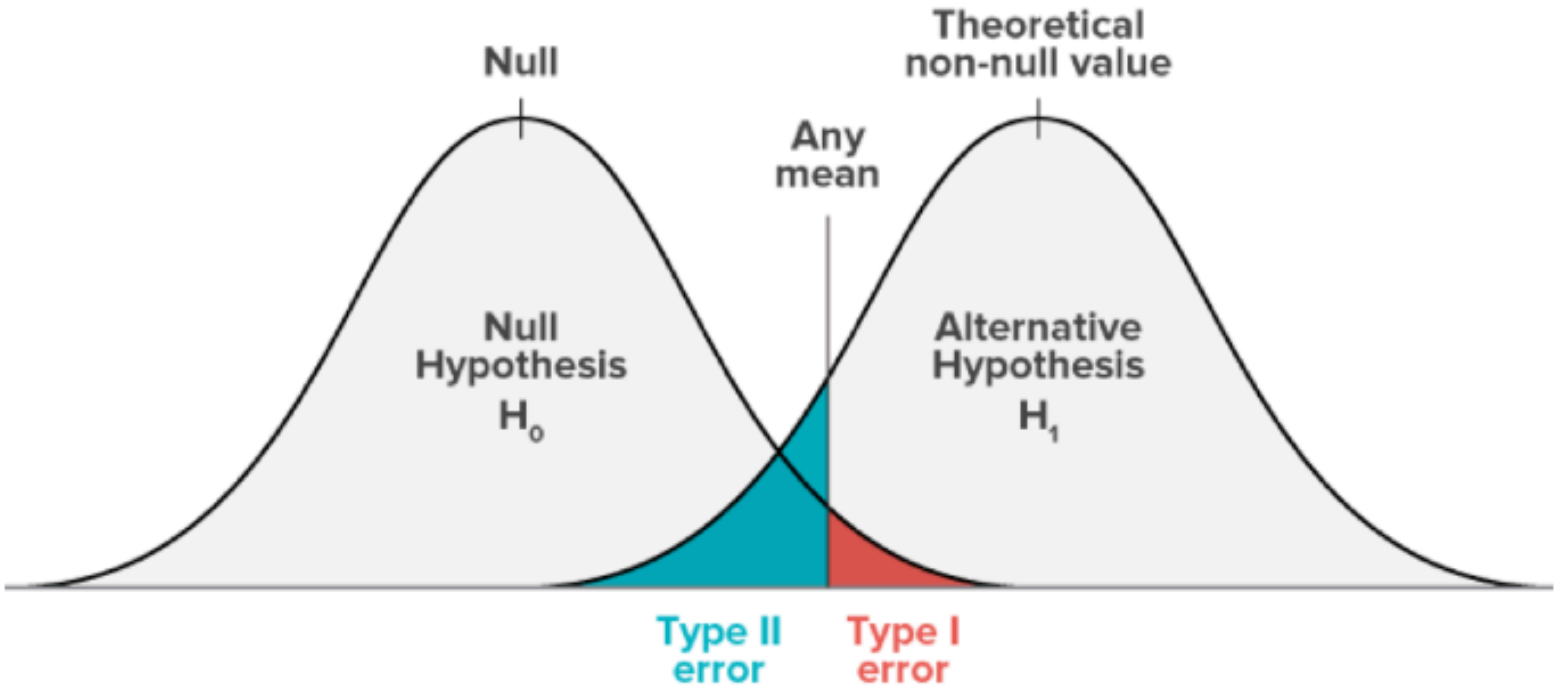
PERMUTATION TEST



PERMUTATION TEST



TYPE I VS TYPE II ERROR



The p value (Type I error) is the probability to obtain an effect equal to or more extreme than the one observed presuming the null hypothesis of no effect is true

P-VALUE HAS PROBLEMS!

BASIC AND APPLIED SOCIAL PSYCHOLOGY, 37:1–2, 2015

ISSN: 0197-3533 print/1532-4834 online
DOI: 10.1080/01973533.2015.1012991

Routledge
Taylor & Francis Group

Editorial

David Trafimow and Michael Marks
New Mexico State University

The *Basic and Applied Social Psychology* (BASP) 2014 Editorial emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would be not required to perform it (Trafimow, 2014). However, to allow authors a grace period, the Editorial stopped short of actually banning the NHSTP. The purpose of the present Editorial is to announce that the grace period is over. From now on, BASP is banning the NHSTP.

With the banning of the NHSTP from BASP, what are the implications for authors? The following are anticipated questions and their corresponding answers.

Question 1. *Will manuscripts with p-values be desk rejected automatically?*

Answer to Question 1. No. If manuscripts pass the a strong case for rejecting it, confidence intervals do not provide a strong case for concluding that the population parameter of interest is likely to be within the stated interval. Therefore, confidence intervals also are banned from BASP.

Bayesian procedures are more interesting. The usual problem with Bayesian procedures is that they depend on some sort of Laplacian assumption to generate numbers where none exist. The Laplacian assumption is that when in a state of ignorance, the researcher should assign an equal probability to each possibility. The problems are well documented (Chihara, 1994; Fisher, 1973; Glymour, 1980; Popper, 1983; Suppes, 1994; Trafimow, 2003, 2005, 2006). However, there have been Bayesian proposals that at least somewhat circumvent

FICTIOUS EXAMPLE: BRAIN CANCER

Hypothesis: Brain cancer causes a headache

Data shows $p<0.01$ (considered very significant)

FICTIOUS EXAMPLE: BRAIN CANCER

Hypothesis: Brain cancer causes a headache

Data shows $p<0.01$ (considered very significant)

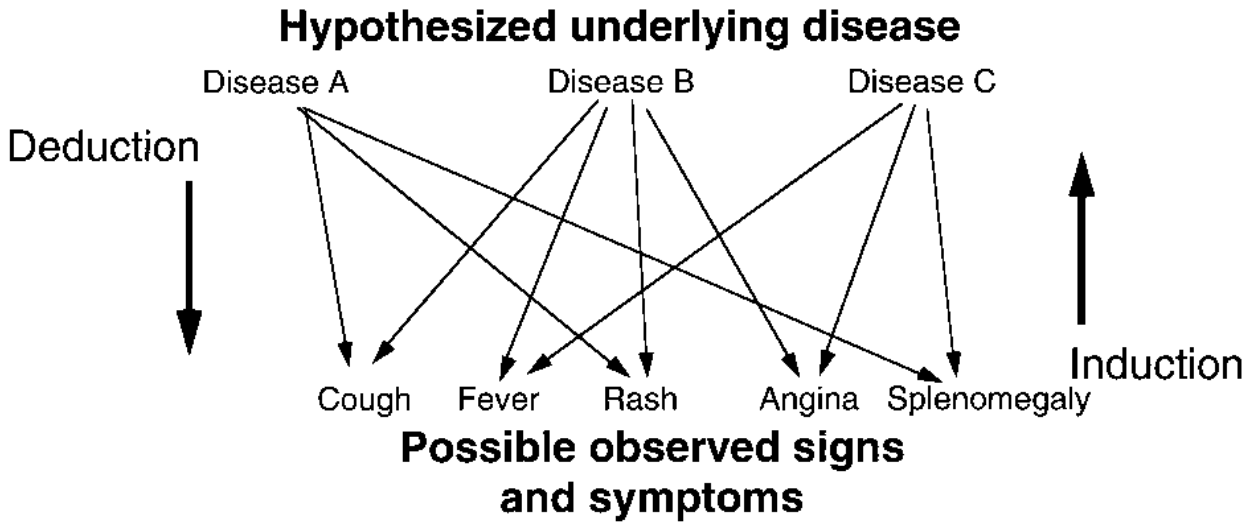
If you have a headache,
how likely is it that you have brain cancer?

FICTIOUS EXAMPLE: BRAIN CANCER

Hypothesis: Brain cancer causes a headache

Data shows $p<0.01$ (considered very significant)

If you have a headache,
how likely is it that you have brain cancer?



CLICKER:
WHAT IS THE INTERPRETATION OF P < 0.05

A) The chances are greater than 1 in 20 that a difference would be found if the study were repeated.

B) The probability is less than 1 in 20 that a difference this large could occur by chance alone.

C) The probability is greater than 1 in 20 that a difference this large could occur by chance alone.

D) The chance is 95% that the study is correct

E) None of the above

MISCONCEPTION 1

"In my experience teaching many academic physicians, when physicians are presented with a single-sentence summary of a study that produced a surprising result with $P = 0.05$, the overwhelming majority will confidently state that there is a 95% or greater chance that the null hypothesis is incorrect.

What is wrong with this?

Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med. 1999;130:995-1004.

MISCONCEPTION 1

"In my experience teaching many academic physicians, when physicians are presented with a single-sentence summary of a study that produced a surprising result with $P = 0.05$, the overwhelming majority will confidently state that there is a 95% or greater chance that the null hypothesis is incorrect.

This is an understand- able but categorically wrong interpretation because the *P* value is calculated on the assumption that the null hypothesis is true. It cannot, therefore, be a direct measure of the probability that the null hypothesis is false. This logical error reinforces the mistaken notion that the data alone can tell us the probability that a hypothesis is true. "

Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med. 1999;130:995-1004.

MISCONCEPTION #1

"If P=.05, the null hypothesis has only a 5% chance of being true"

Let us suppose we flip a penny four times and observe four heads, two-sided P =.125. This does not mean that the probability of the coin being fair is only 12.5%.

Steven Goodman: "A Dirty Dozen: Twelve P-Value Misconceptions"

MISCONCEPTION #2

A non significant difference (eg, P .05) means there is no difference between groups.

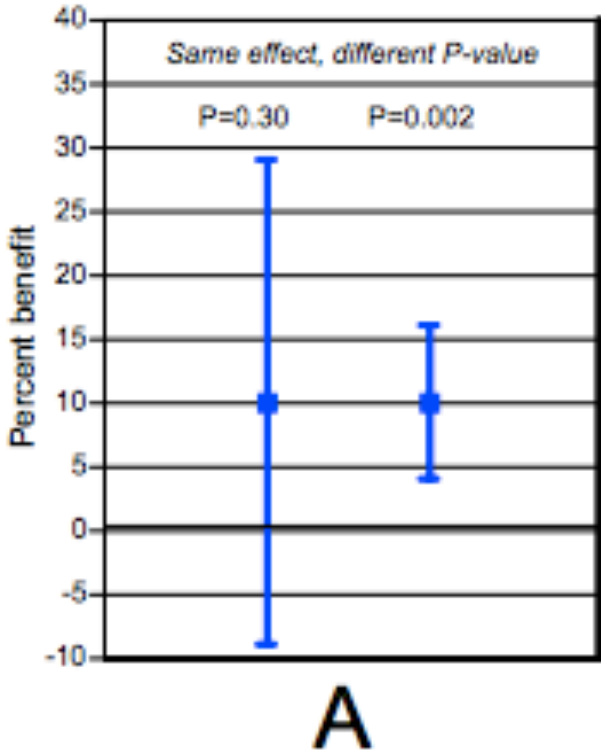
- A non significant difference only means the null effect is statistically consistent with the observation
- It does not make the null effect most likely
- In fact, the observed effect best explains the effect regardless the significance.

Steven Goodman: "A Dirty Dozen: Twelve P-Value Misconceptions"

MISCONCEPTION #3

A statistically significant finding is (clinical) important

The P value carries no information about the magnitude of an effect, which is captured by the effect estimate and confidence interval.



MISCONCEPTION #4

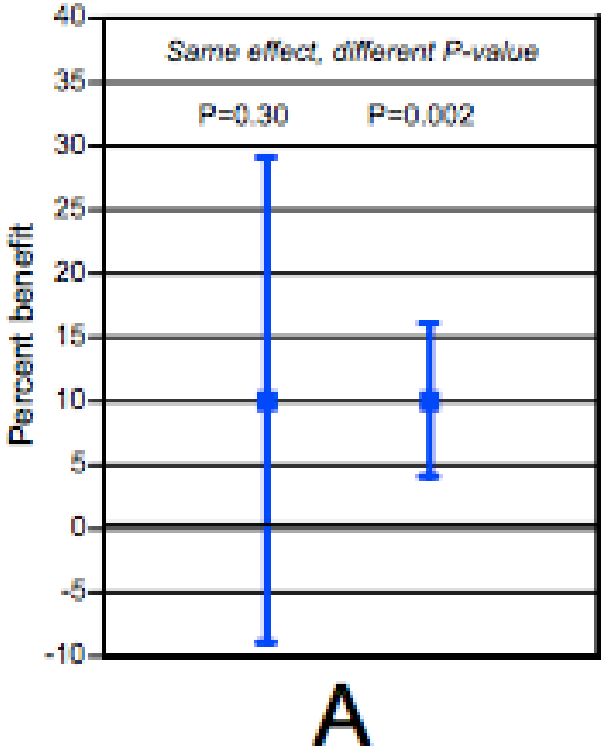
"Studies with P values on opposite sides of .05 are conflicting"

H_0:Drug T has no effect

H_1:Drug T has a positive effect

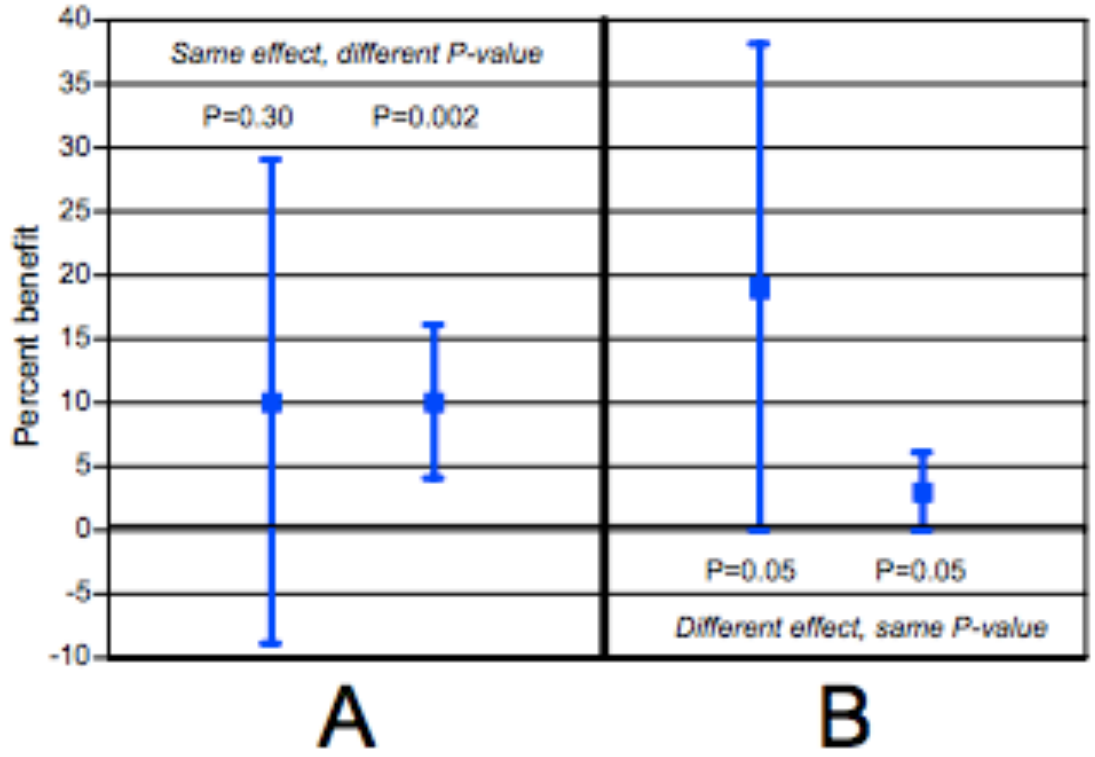
Study I: P=0.3

Study II: P=0.002



MISCONCEPTION #5

Studies with the same P value provide the same evidence against the null hypothesis



P-VALUE HAS PROBLEMS!

BASIC AND APPLIED SOCIAL PSYCHOLOGY, 37:1–2, 2015

ISSN: 0197-3533 print/1532-4834 online
DOI: 10.1080/01973533.2015.1012991

Routledge
Taylor & Francis Group

Editorial

David Trafimow and Michael Marks
New Mexico State University

The *Basic and Applied Social Psychology* (BASP) 2014 Editorial emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would be not required to perform it (Trafimow, 2014). However, to allow authors a grace period, the Editorial stopped short of actually banning the NHSTP. The purpose of the present Editorial is to announce that the grace period is over. From now on, BASP is banning the NHSTP.

With the banning of the NHSTP from BASP, what are the implications for authors? The following are anticipated questions and their corresponding answers.

Question 1. *Will manuscripts with p-values be desk rejected automatically?*

Answer to Question 1. No. If manuscripts pass the a strong case for rejecting it, confidence intervals do not provide a strong case for concluding that the population parameter of interest is likely to be within the stated interval. Therefore, confidence intervals also are banned from BASP.

Bayesian procedures are more interesting. The usual problem with Bayesian procedures is that they depend on some sort of Laplacian assumption to generate numbers where none exist. The Laplacian assumption is that when in a state of ignorance, the researcher should assign an equal probability to each possibility. The problems are well documented (Chihara, 1994; Fisher, 1973; Glymour, 1980; Popper, 1983; Suppes, 1994; Trafimow, 2003, 2005, 2006). However, there have been Bayesian proposals that at least somewhat circumvent

P-HACKING (ALSO DATA DREDGING, DATA FISHING, DATA SNOOPING, DATA BUTCHERY)

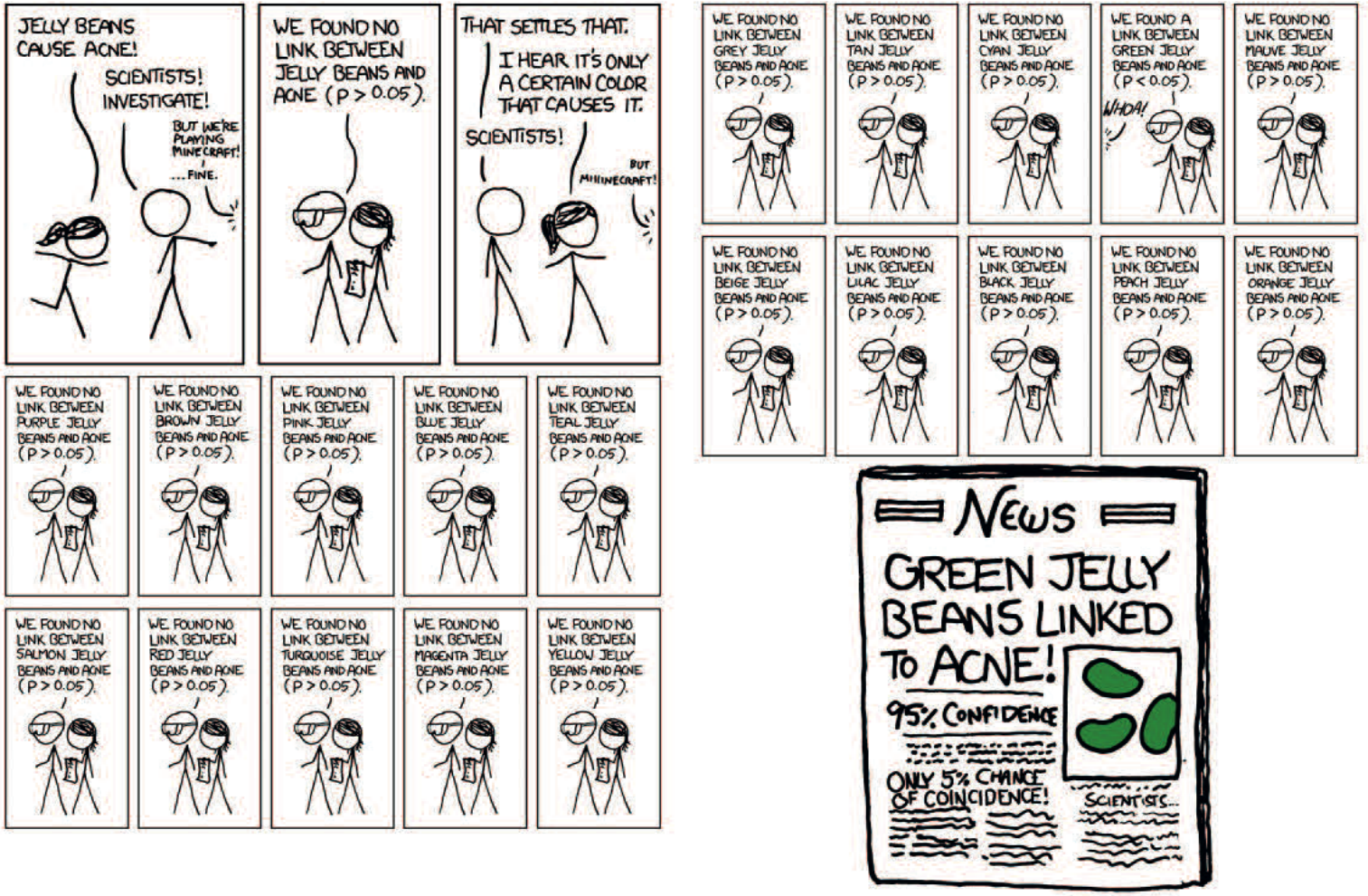


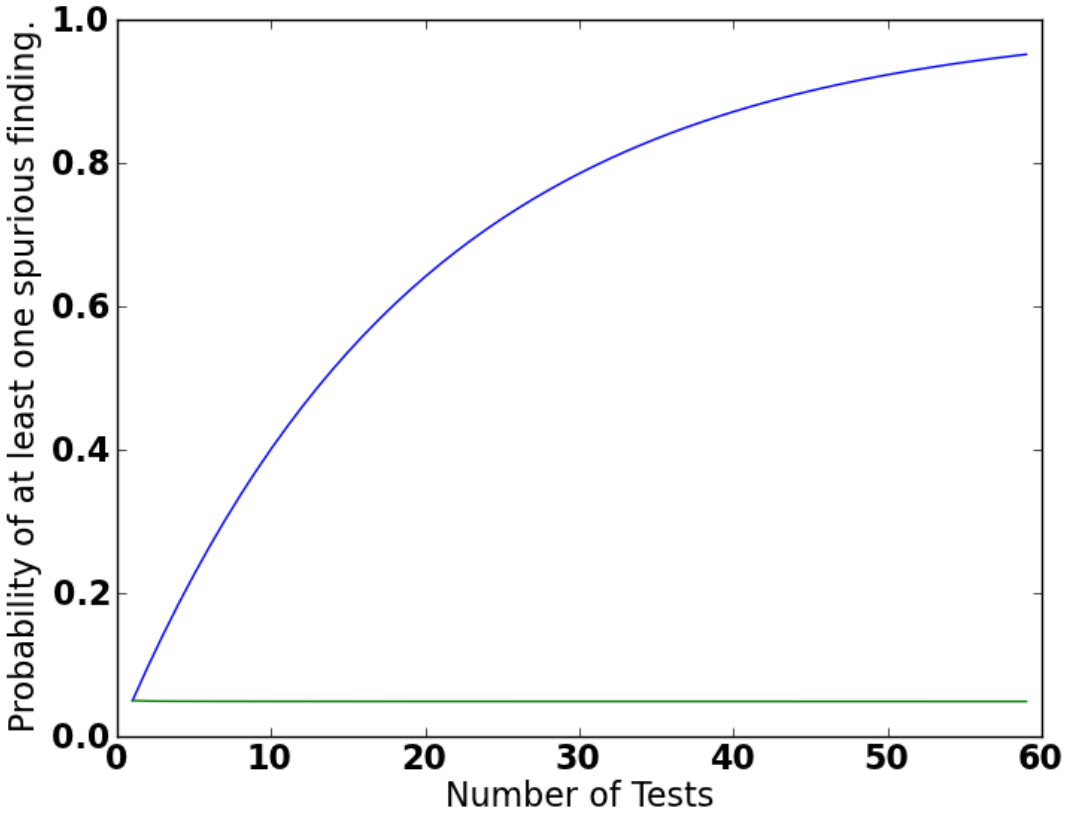
Figure 1. There is no overall effect of jelly beans on acne. Bummer. How about subgroups? Often subgroups are explored without alerting the reader to the number of questions at issue. Courtesy xkcd, http://xkcd.com/882/

P(detecting an effect when there is none) = $\alpha = 0.05$

P(detecting an effect when it exists) = $1 - \alpha$

P(detecting an effect when it exists on every experiment) = $(1 - \alpha)^k$

P(detecting an effect when there is none on <u>at least one</u> experiment) = $1 - (1 - \alpha)^k$



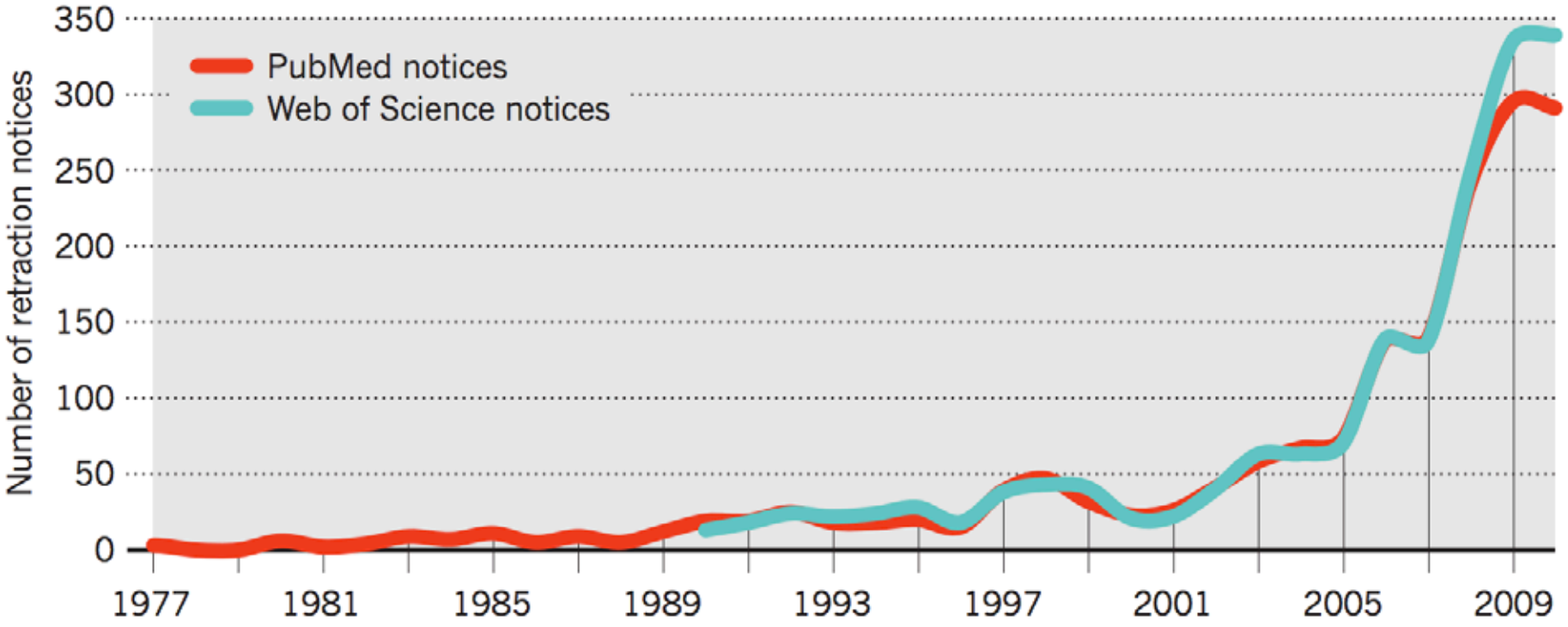
$\alpha = 0.05$

"Familywise Error Rate"

MISTAKES AND FRAUD

2001 – 2011:

- 10X increase in retractions
- only 1.44X increase in papers

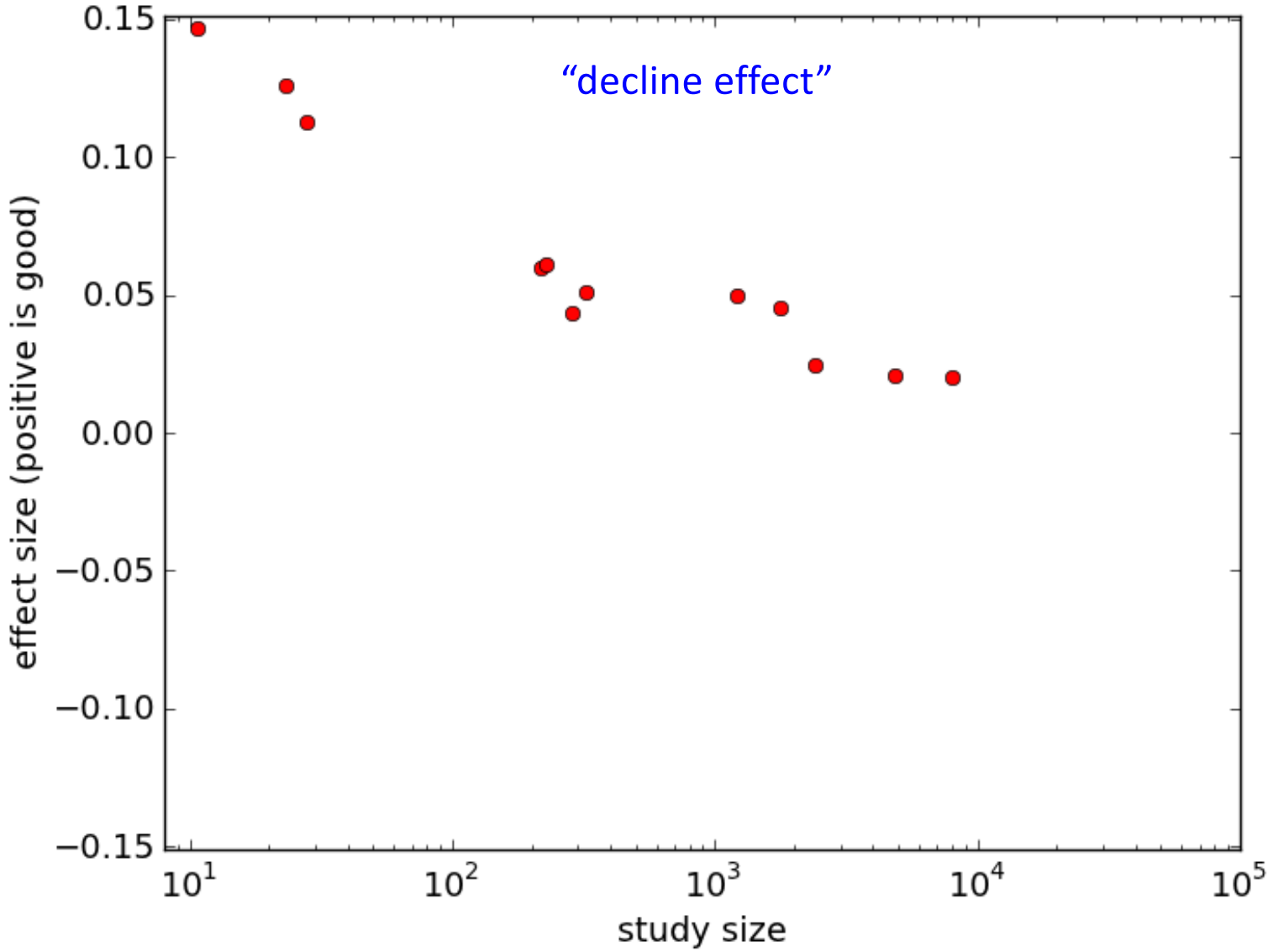


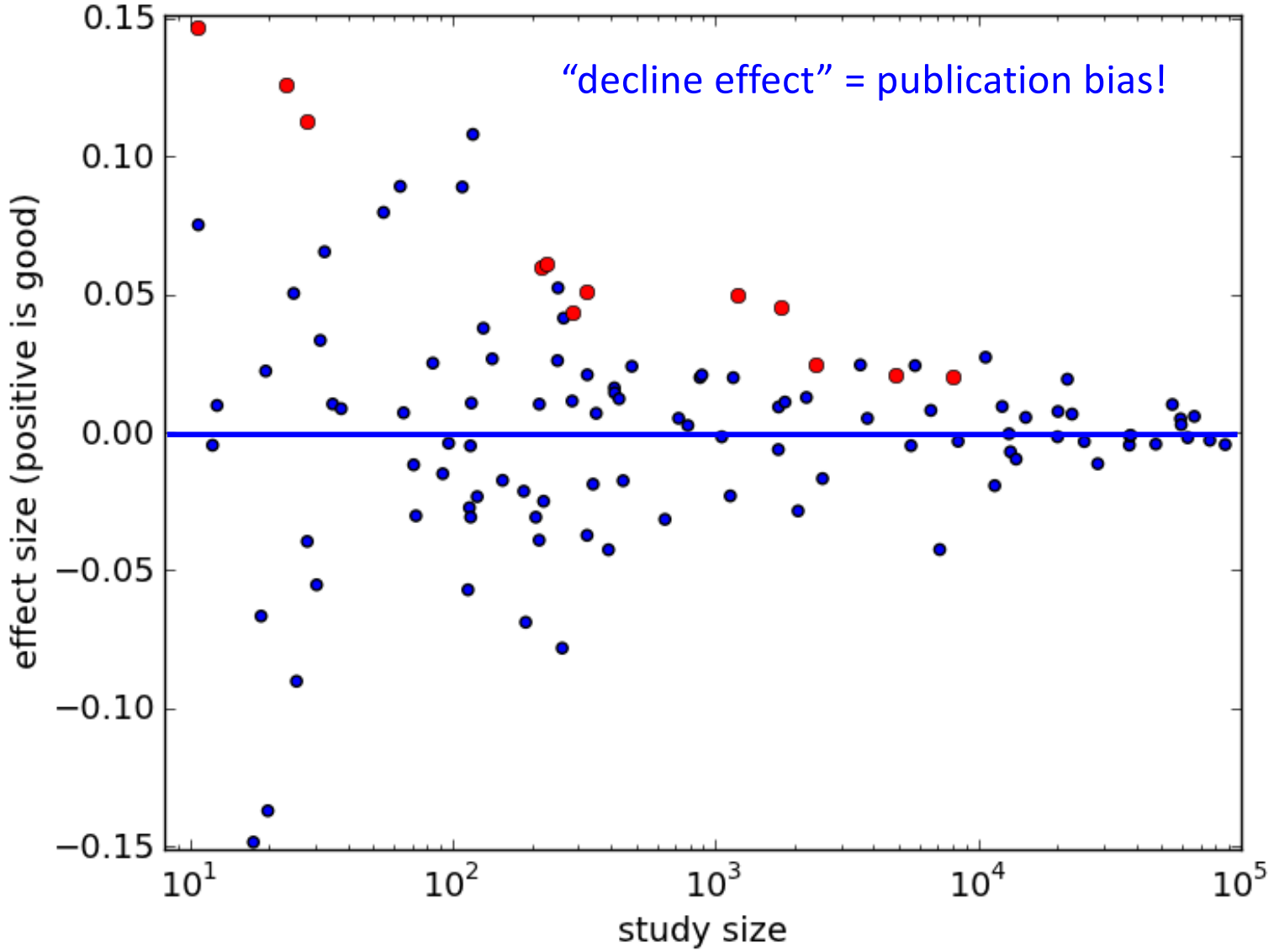
Richard Van Noorden, 2011, Nature 478

The Rise of the Retractions

http://www.nature.com/news/2011/111005/pdf/478026a.pdf

PUBLICATION BIAS




"decline effect" = publication bias!
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
effect size (positive is good)
10^1
10^2
10^3
10^4
10^5
study size

FAMILY-WISE ERROR RATE CORRECTIONS

Bonferroni Correction

- Just divide by the number of hypotheses

$$\alpha_c = \frac{\alpha}{k}$$

Šidák Correction

- Asserts independence

$$\alpha = 1 - (1 - \alpha_c)^k$$

$$\alpha_c = 1 - (1 - \alpha)^{\frac{1}{k}}$$

MANY ANALYSTS, ONE DATA SET

MANY ANALYSTS, ONE DATA SET

Variations in Analytic Choices Affect Results

Abstract:

"Twenty-nine teams involving 61 analysts used the same data set to address the same research question: whether soccer referees are more likely to give red cards to dark-skin-toned players than to light- skin-toned players. Analytic approaches varied widely across the teams, and the estimated effect sizes ranged from 0.89 to 2.93 (*Mdn* = 1.31) in odds-ratio units. Twenty teams (69%) found a statistically significant positive effect, and 9 teams (31%) did not observe a significant relationship. Overall, the 29 different analyses used 21 unique combinations of covariates. Neither analysts' prior beliefs about the effect of interest nor their level of expertise readily explained the variation in the outcomes of the analyses. …. Crowdsourcing data analysis, a strategy in which numerous research teams are recruited to simultaneously investigate the same research question, makes transparent how defensible, yet subjective, analytic choices influence research results."

CLOSING THOUGHTS

"It is easy to lie with statistics,
but it is easier to lie without them."

attributed to Frederick Mosteller (1916-2006)

REFERENCES

- How to lie with Statistics - Darrell Huff
- How to lie with Maps - Mark Monmonier
- http://www.sciencebasedmedicine.org/psychology-journal-bans-significance-testing/
- Nuzzo R: Scientific method: statistical errors. Nature. 2014 Feb 13;506(7487)
- Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med. 1999;130:995-1004.
- Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. Ann Intern Med. 1999;130:1005-13.