Philosophical Underpinnings of the Replication Crisis

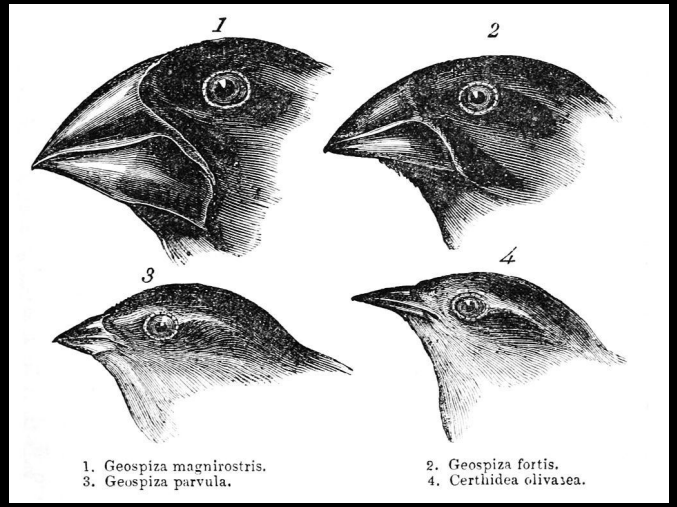
The Problem of Induction and Causal Inference

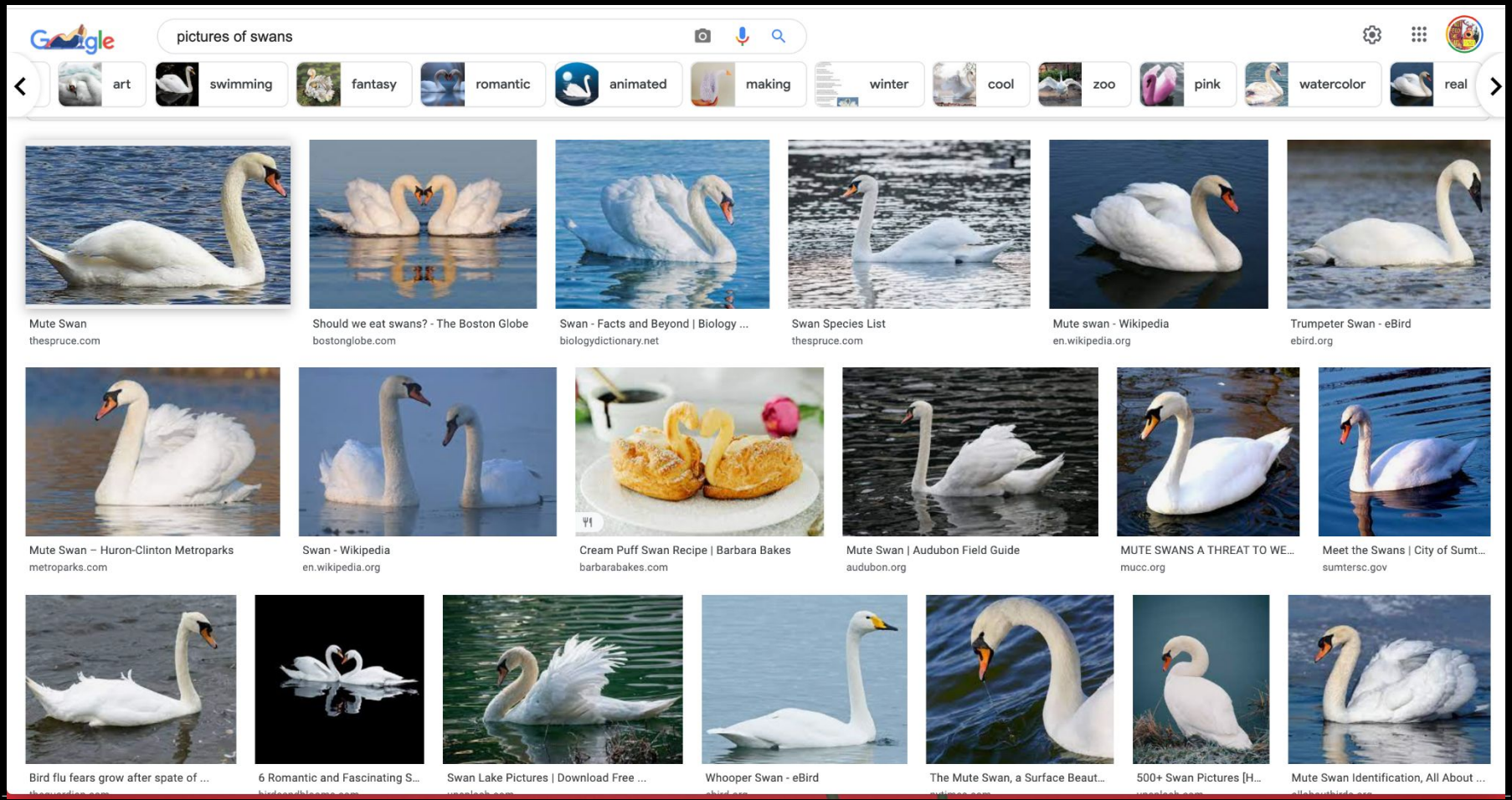
1/20/2022

Brief Review

- Canonical Papers Documenting the Replication Crisis
  - Why most published research findings are false (2005) *PLOS Medicine*
  - Estimating the reproducibility of psychological science (2015) *Science*
  - Evaluating the replicability of social science experiments in Nature and Science from 2010 to 2015 (2018) *NHB*
  - The Generalizability Crisis (2019) *Behavioral and Brain Sciences*
- Problems
  - Sample sizes chosen without regard for statistical power
  - Publication bias, file drawer problem
  - Type I and Type II errors, false positives, false negatives
  - P-hacking, HARKing (hypothesizing after results are known), misinterpreting p-values
- Tools to Partially Address and Evaluate Problems
  - Power calculations, pre-registration, open science
  - P-curves, z-curves, funnel plots, expected replication rate, metascience
  - Critical thinking

Drawing inspiration from birds




pictures of swans
art
swimming
fantasy
romantic
animated
making
winter
cool
zoo
pink
watercolor
real
Mute Swan
thespruce.com
Should we eat swans? - The Boston Globe
bostonglobe.com
Swan - Facts and Beyond | Biology ...
biologydictionary.net
Swan Species List
thespruce.com
Mute swan - Wikipedia
en.wikipedia.org
Trumpeter Swan - eBird
ebird.org
Mute Swan – Huron-Clinton Metroparks
metroparks.com
Swan - Wikipedia
en.wikipedia.org
Cream Puff Swan Recipe | Barbara Bakes
barbarabakes.com
Mute Swan | Audubon Field Guide
audubon.org
MUTE SWANS A THREAT TO WE...
mucc.org
Meet the Swans | City of Sumt...
sumtersc.gov
Bird flu fears grow after spate of ...
6 Romantic and Fascinating S...
Swan Lake Pictures | Download Free ...
Whooper Swan - eBird
The Mute Swan, a Surface Beaut...
500+ Swan Pictures [H...
Mute Swan Identification, All About ...

“Now, it is far from obvious, from a logical point of view, that we are justified in inferring universal statements from singular ones, no matter how numerous; for any conclusion drawn this way may always turn out to be false: no matter how many instances of white swans we may have observed, this does not justify the conclusion that all swans are white.”

Karl Popper (1935) *The Logic of Scientific Discovery,* p. 4


Absence of
Evidence is not
Evidence of
Absence

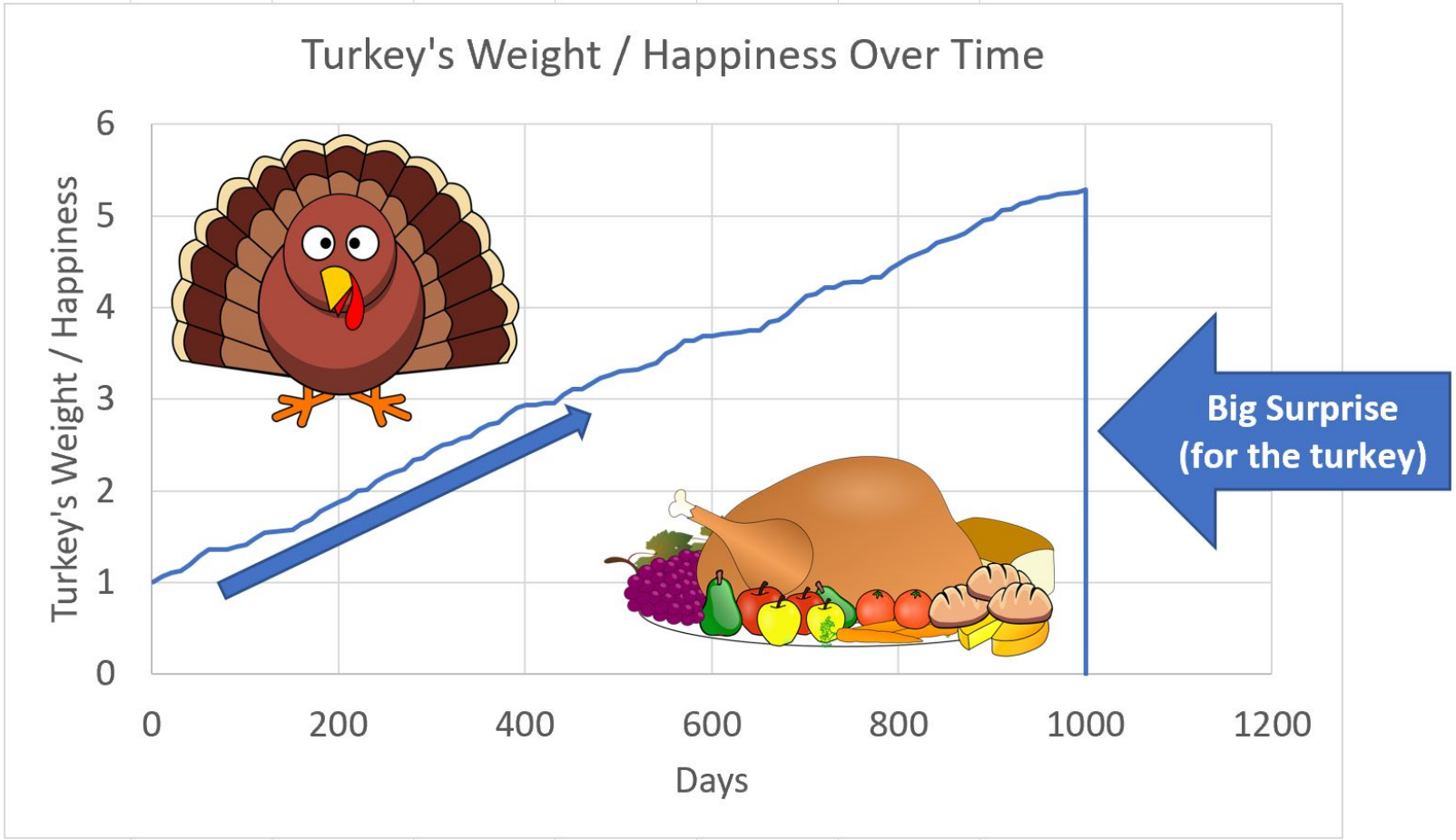
The Problem of Induction

(Sometimes called Hume's Problem)

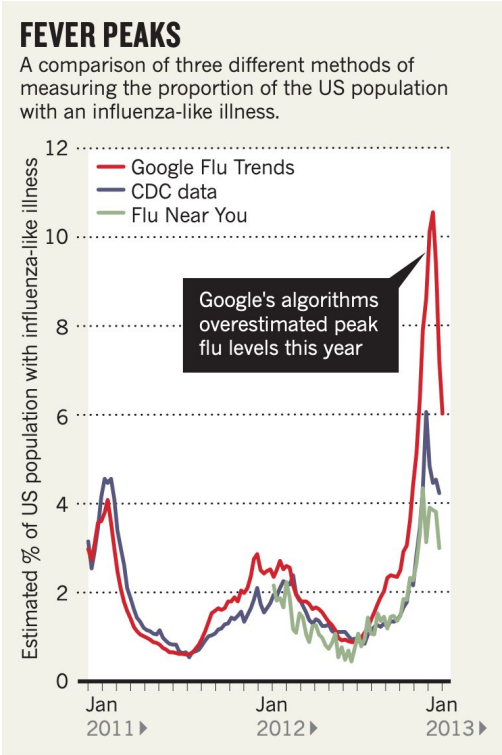

THE
BLACK SWAN MAN
10¢
THE SILENT PROTECTOR, THE MYSTERIOUS GUARDIAN ANGEL, THE FEARLESS FEATHERED HERO DEFENDING AGAINST THE EVIL AND DESTRUCTIVE FORCES OF RISK..!

Classic Mistake of Induction

The Turkey Problem


Turkey's Weight / Happiness Over Time
Turkey's Weight / Happiness
0
1
2
3
4
5
6
0
200
400
600
800
1000
1200
Days
Big Surprise
(for the turkey)

Another Classic Mistake: **Google Flu Trends**

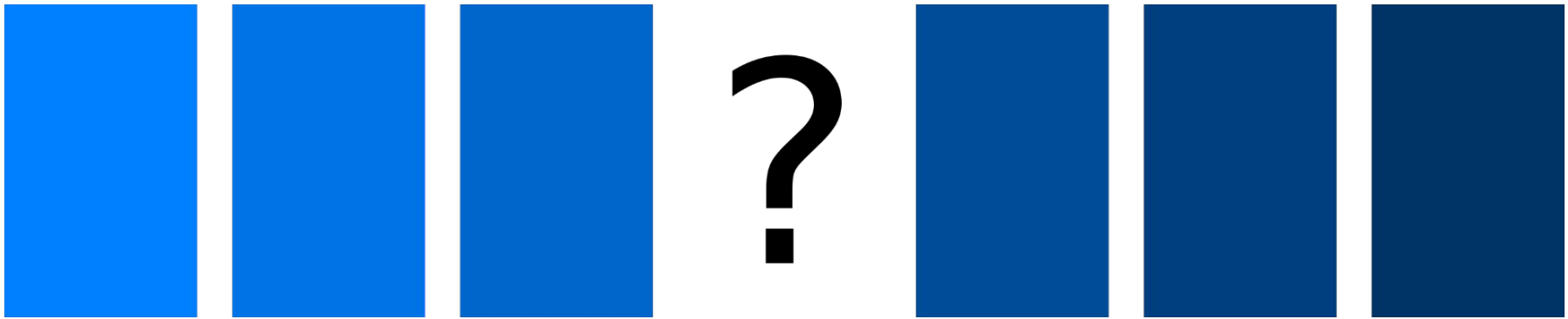


Source: Butler 2013, *Nature*

So……. how do we turn observations into rigorous science?

(Demarcation Problem)

Hume's Exception: The Missing Shade of Blue



Popper's Solution

Induction is a myth

Popper's Philosophy of Empirical Falsification

- Scientific theories should be falsifiable
- If all actual attempts to falsify the theory have failed so far, then we can consider the scientific theory highly corroborated
- Knowledge is created by conjecture and criticism. Scientific experiments are attempts at criticizing existing theories

Empirical Falsification

- **Claim**: All swans are white.
  - **Falsification**: One black swan

Empirical Falsification

- **Conventional wisdom**: facial expressions reveal emotions
- **Falsification**: Body cues, not facial expressions, discriminate between images of intense negative and positive emotions (Aviezer et al 2012)

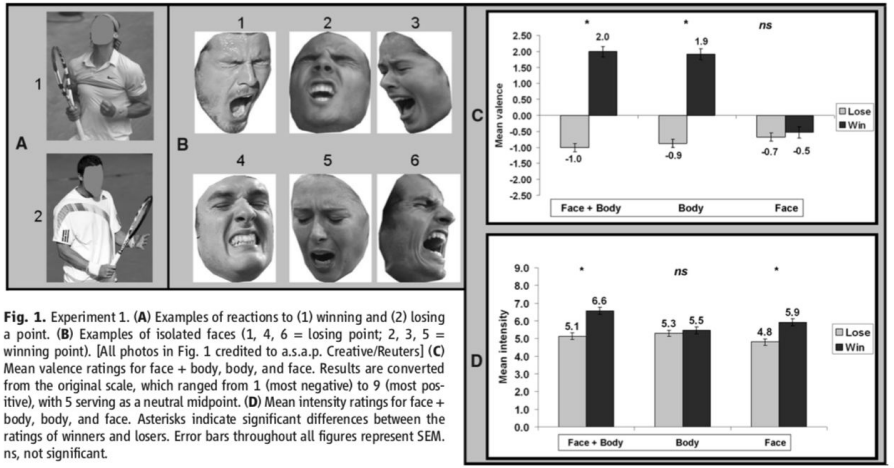


Fig. 1. Experiment 1. (**A**) Examples of reactions to (1) winning and (2) losing a point. (**B**) Examples of isolated faces (1, 4, 6 = losing point; 2, 3, 5 = winning point). [All photos in Fig. 1 credited to a.s.a.p. Creative/Reuters] (**C**) Mean valence ratings for face + body, body, and face. Results are converted from the original scale, which ranged from 1 (most negative) to 9 (most positive), with 5 serving as a neutral midpoint. (**D**) Mean intensity ratings for face + body, body, and face. Asterisks indicate significant differences between the ratings of winners and losers. Error bars throughout all figures represent SEM. ns, not significant.

“affective valence was higher for winners than losers when ratings were based on the face + body ($P < 0.0001$), or the body alone ($P < 0.0001$), but not when ratings were based on the face alone ($P > 0.3$)”

Aviezer et al 2012 *Science*

Empirical Falsification

- Scientifically-corroborated statement: Body cues, not facial expressions, discriminate between images of intense negative and positive emotions (Aviezer et al 2012)

- **How would you design an experiment to falsify this science-backed statement?**
  - Direct replication
  - Adapted replication

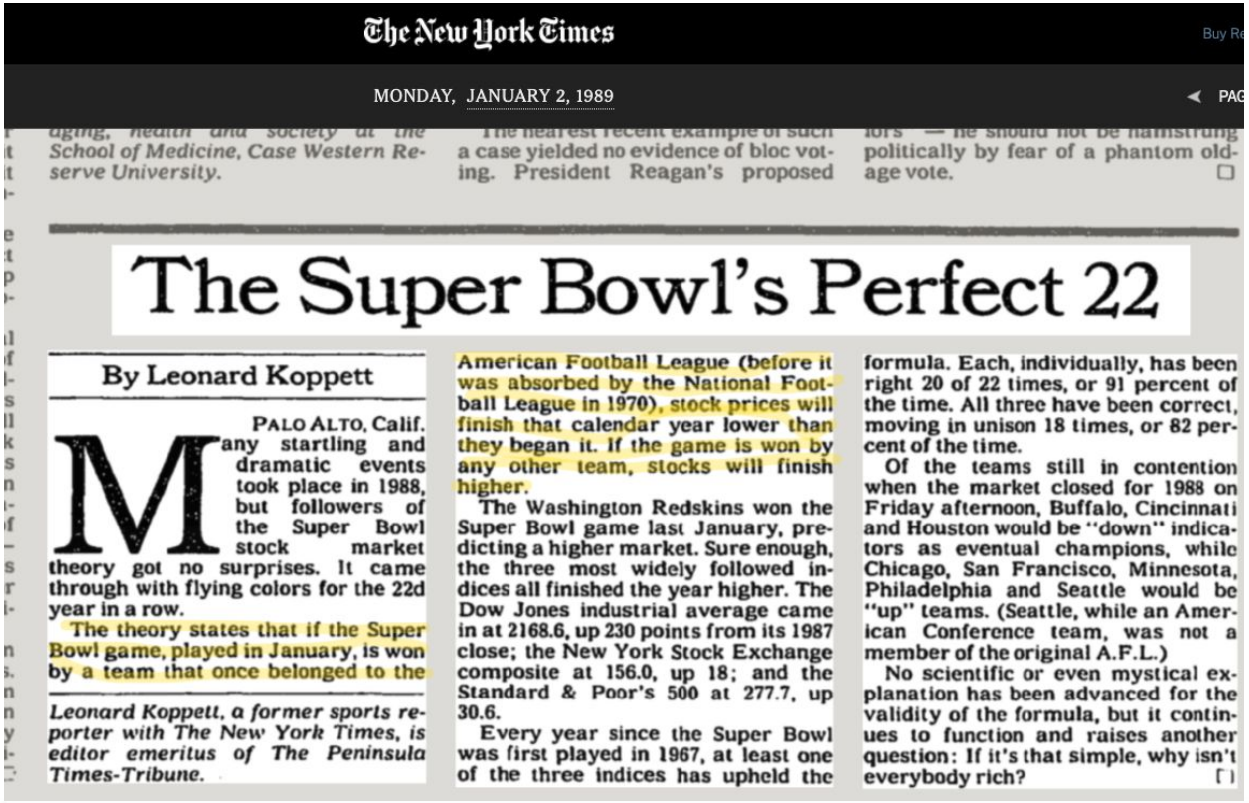
Counterfactual Thinking

Control: Full image showing the face and body

Treatment 1: Cropped image showing only the face

Treatment 2: Cropped image showing only the body

Lack of Counterfactual Thinking → Fooled by Randomness



The New York Times

MONDAY, JANUARY 2, 1989

aging, health and society at the School of Medicine, Case Western Reserve University.

The nearest recent example of such a case yielded no evidence of bloc voting. President Reagan's proposed

...lors" — he should not be hamstrung politically by fear of a phantom old-age vote.

The Super Bowl's Perfect 22

By Leonard Koppett

PALO ALTO, Calif. Many startling and dramatic events took place in 1988, but followers of the Super Bowl stock market theory got no surprises. It came through with flying colors for the 22d year in a row.

The theory states that if the Super Bowl game, played in January, is won by a team that once belonged to the American Football League (before it was absorbed by the National Football League in 1970), stock prices will finish that calendar year lower than they began it. If the game is won by any other team, stocks will finish higher.

The Washington Redskins won the Super Bowl game last January, predicting a higher market. Sure enough, the three most widely followed indices all finished the year higher. The Dow Jones industrial average came in at 2168.6, up 230 points from its 1987 close; the New York Stock Exchange composite at 156.0, up 18; and the Standard & Poor's 500 at 277.7, up 30.6.

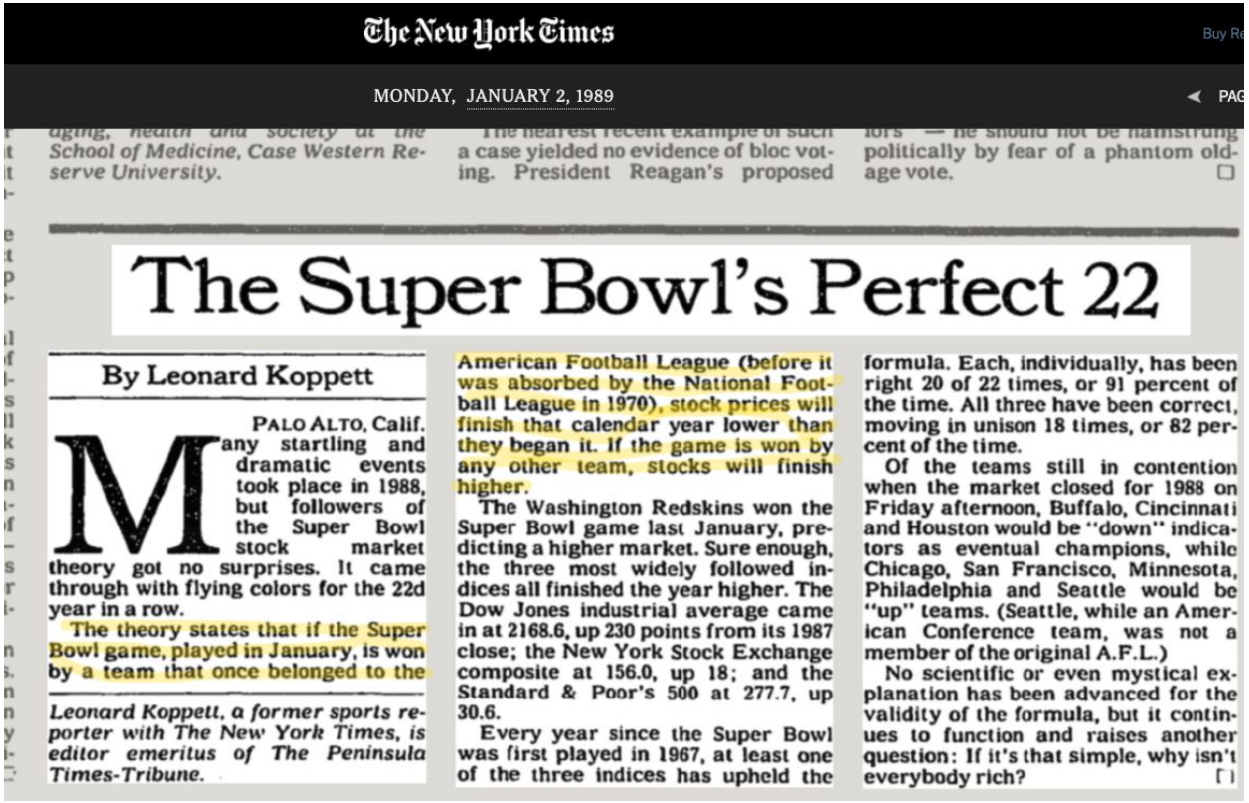
Every year since the Super Bowl was first played in 1967, at least one of the three indices has upheld the formula. Each, individually, has been right 20 of 22 times, or 91 percent of the time. All three have been correct, moving in unison 18 times, or 82 percent of the time.

Of the teams still in contention when the market closed for 1988 on Friday afternoon, Buffalo, Cincinnati and Houston would be "down" indicators as eventual champions, while Chicago, San Francisco, Minnesota, Philadelphia and Seattle would be "up" teams. (Seattle, while an American Conference team, was not a member of the original A.F.L.)

No scientific or even mystical explanation has been advanced for the validity of the formula, but it continues to function and raises another question: If it's that simple, why isn't everybody rich?

Leonard Koppett, a former sports reporter with The New York Times, is editor emeritus of The Peninsula Times-Tribune.

Lack of Counterfactual Thinking → Fooled by Randomness



The New York Times

MONDAY, JANUARY 2, 1989

aging, health and society at the School of Medicine, Case Western Reserve University.

The nearest recent example of such a case yielded no evidence of bloc voting. President Reagan's proposed

politically by fear of a phantom old-age vote.

The Super Bowl's Perfect 22

By Leonard Koppett

PALO ALTO, Calif. Many startling and dramatic events took place in 1988, but followers of the Super Bowl stock market theory got no surprises. It came through with flying colors for the 22d year in a row.

The theory states that if the Super Bowl game, played in January, is won by a team that once belonged to the American Football League (before it was absorbed by the National Football League in 1970), stock prices will finish that calendar year lower than they began it. If the game is won by any other team, stocks will finish higher.

The Washington Redskins won the Super Bowl game last January, predicting a higher market. Sure enough, the three most widely followed indices all finished the year higher. The Dow Jones industrial average came in at 2168.6, up 230 points from its 1987 close; the New York Stock Exchange composite at 156.0, up 18; and the Standard & Poor's 500 at 277.7, up 30.6.

Every year since the Super Bowl was first played in 1967, at least one of the three indices has upheld the formula. Each, individually, has been right 20 of 22 times, or 91 percent of the time. All three have been correct, moving in unison 18 times, or 82 percent of the time.

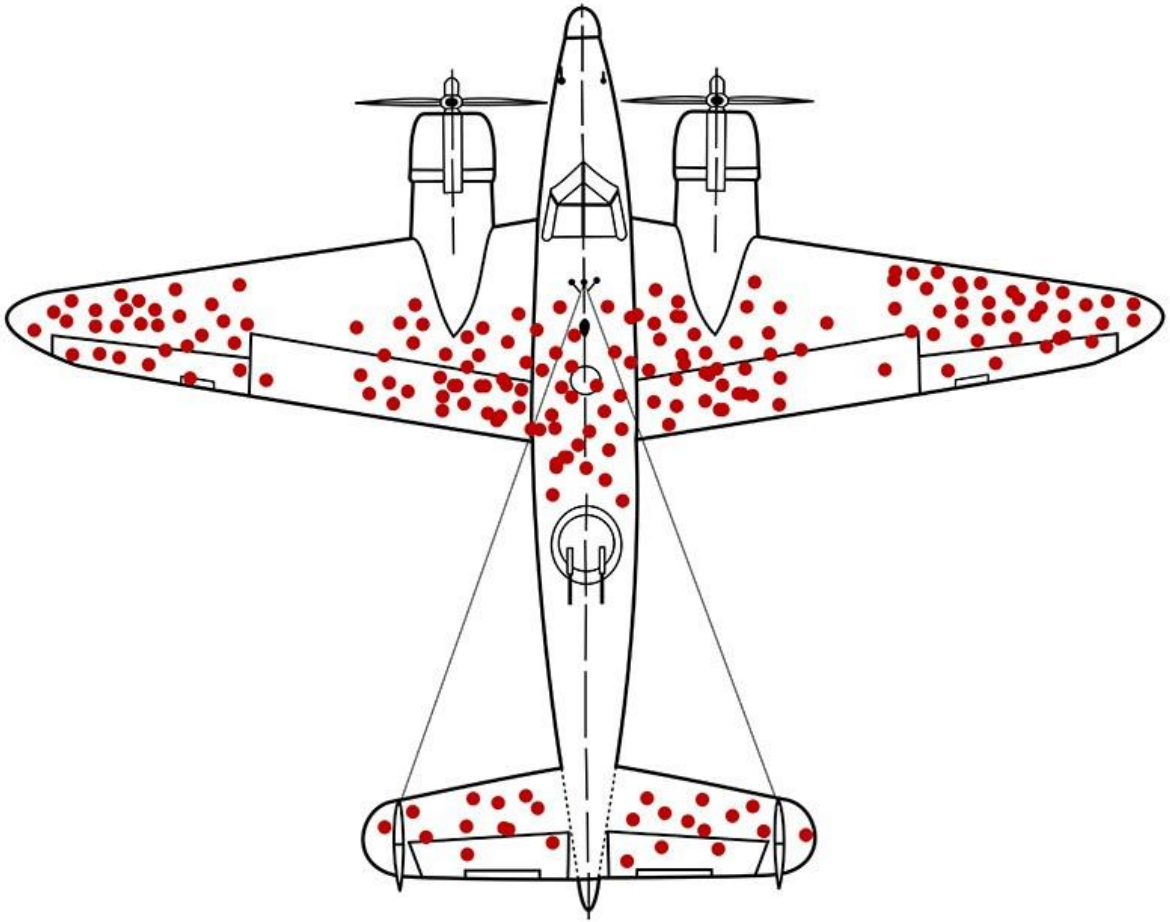
Of the teams still in contention when the market closed for 1988 on Friday afternoon, Buffalo, Cincinnati and Houston would be "down" indicators as eventual champions, while Chicago, San Francisco, Minnesota, Philadelphia and Seattle would be "up" teams. (Seattle, while an American Conference team, was not a member of the original A.F.L.)

No scientific or even mystical explanation has been advanced for the validity of the formula, but it continues to function and raises another question: If it's that simple, why isn't everybody rich?

Leonard Koppett, a former sports reporter with The New York Times, is editor emeritus of The Peninsula Times-Tribune.

Probability = $(22*21)/2^{22} = 0.01\%$

That's 1/10,000



If we had all 15 million high school students in the United States flip a fair coin 20 times, we would expect around ?? of them to get the exact same side all 20 times in a row.

If we had all 15 million high school students in the United States flip a fair coin 20 times, we would expect around 30 of them to get the exact same side all 20 times in a row.

If we asked these 30 students to flip a 21st time, there's a one in billion chance that they will all get the same side again. And, there's a 50% chance that at least 15 will get the same side again.

Imagine you are in a research experiment and you flip a coin and get heads 20 times in a row.

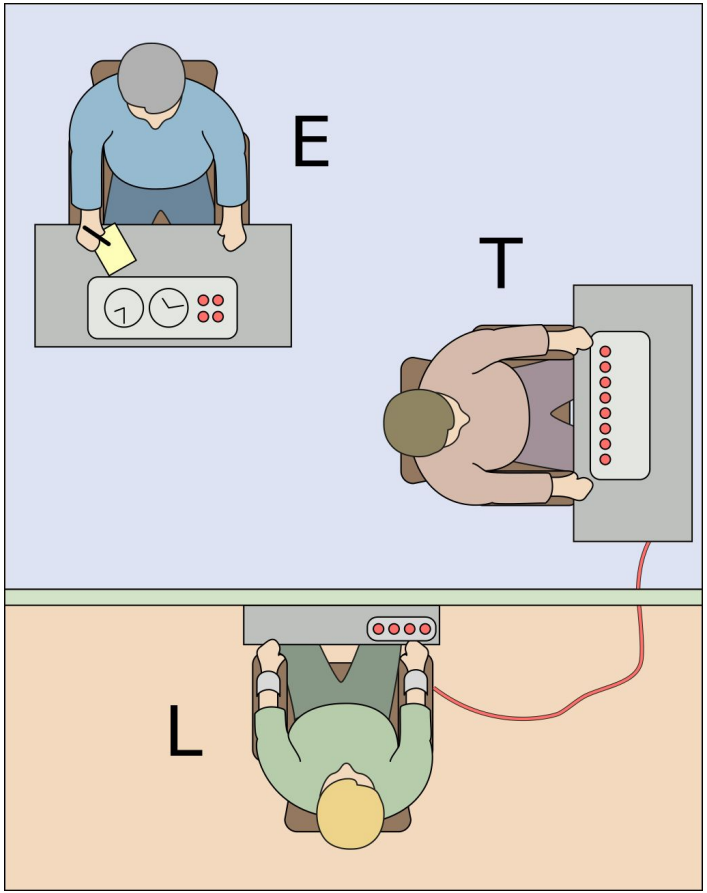
What is the likelihood you'll get heads on the 21st time?

The Ludic Fallacy

"the misuse of games to model real-life situations"

- Nassim Taleb (2007), *The Black Swan*

The Milgram Experiment

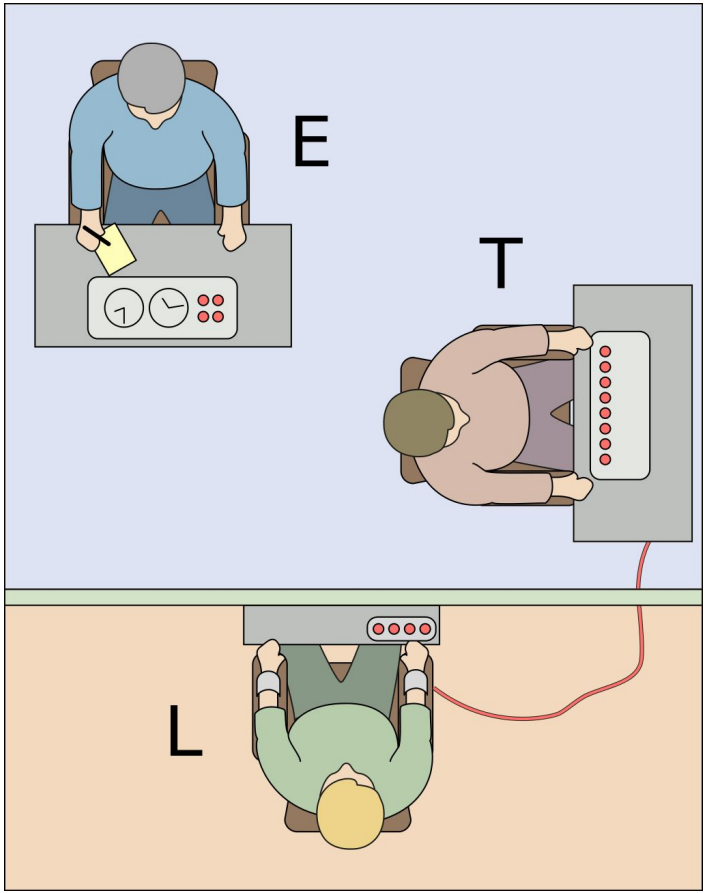


Milgram (1963)

The Milgram Experiment

This archival material indicates that in 18 of 23 variations of the experiment, the mean levels of shock for those who fully believed that they were inflicting pain were lower than for subjects who did not fully believe they were inflicting pain. These data suggest that the perception of pain inflated subject defiance and that subject skepticism inflated their obedience.

Perry et al 2019, *Social Psychology Quarterly*



Is this research ecologically valid?

The Linda Problem Kahneman & Tversky (1981)

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Which is more probable?

(a) Linda is a bank teller.
(b) Linda is a bank teller and is active in the feminist movement.

The Conjunction Fallacy

A fallacy when specific conditions are assumed to be more probable than a more general condition

The Linda Problem Revisited Hertwig & Gigerenzer (1999)

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

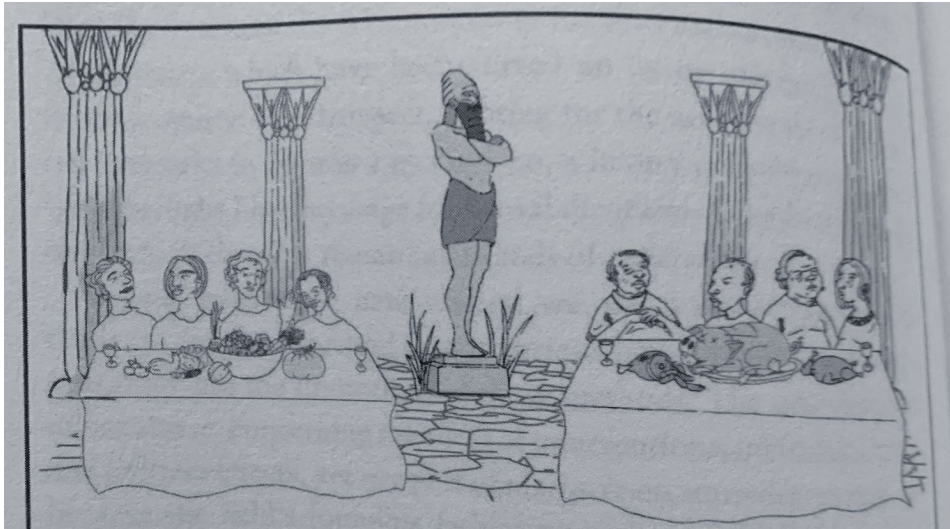
Which is more **probable**?

(a) Linda is a bank teller.
(b) Linda is a bank teller **and** is active in the feminist movement.

What’s at stake?

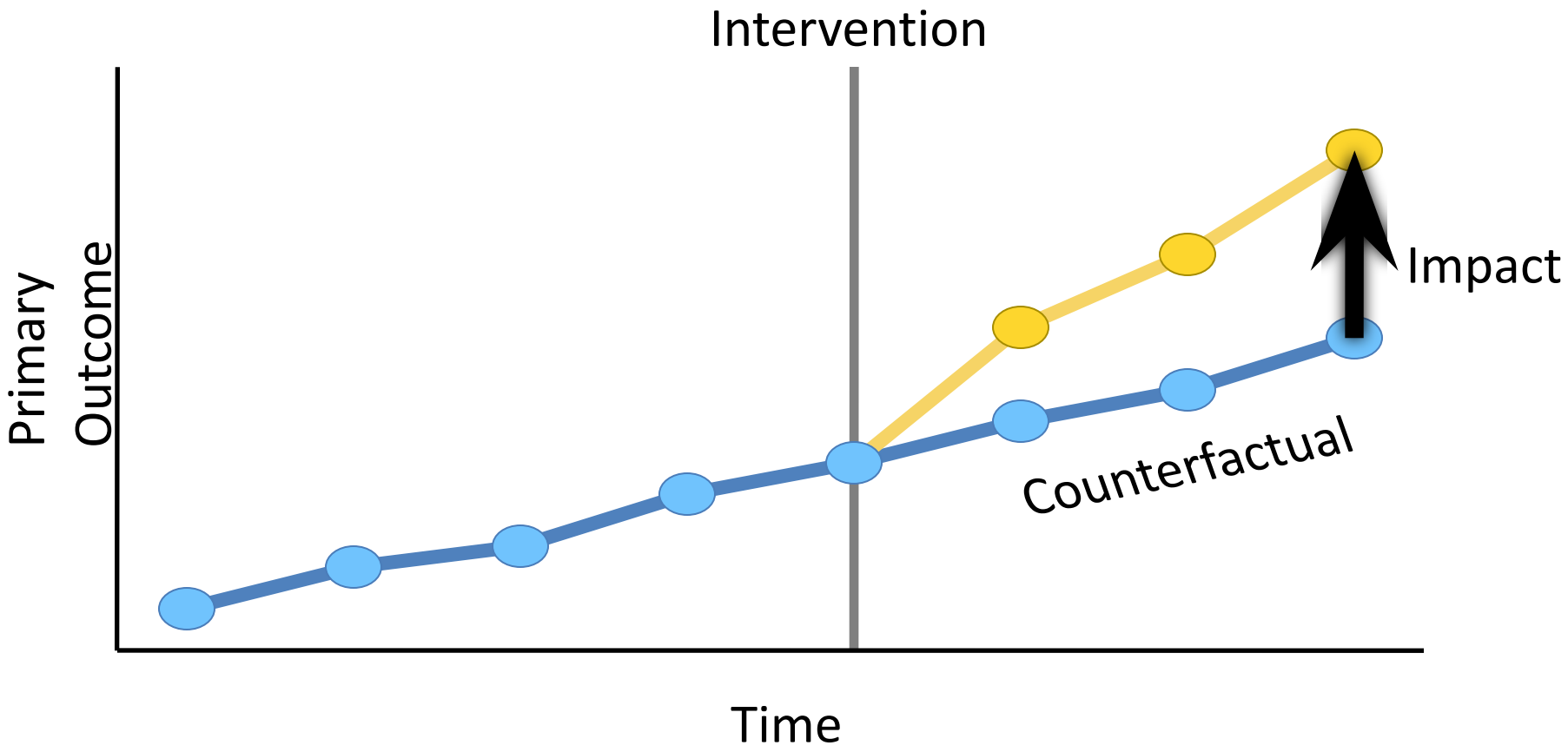
Counterfactual Thinking
in the Potential Outcomes Framework

(The Rubin-Neyman Causal Model)

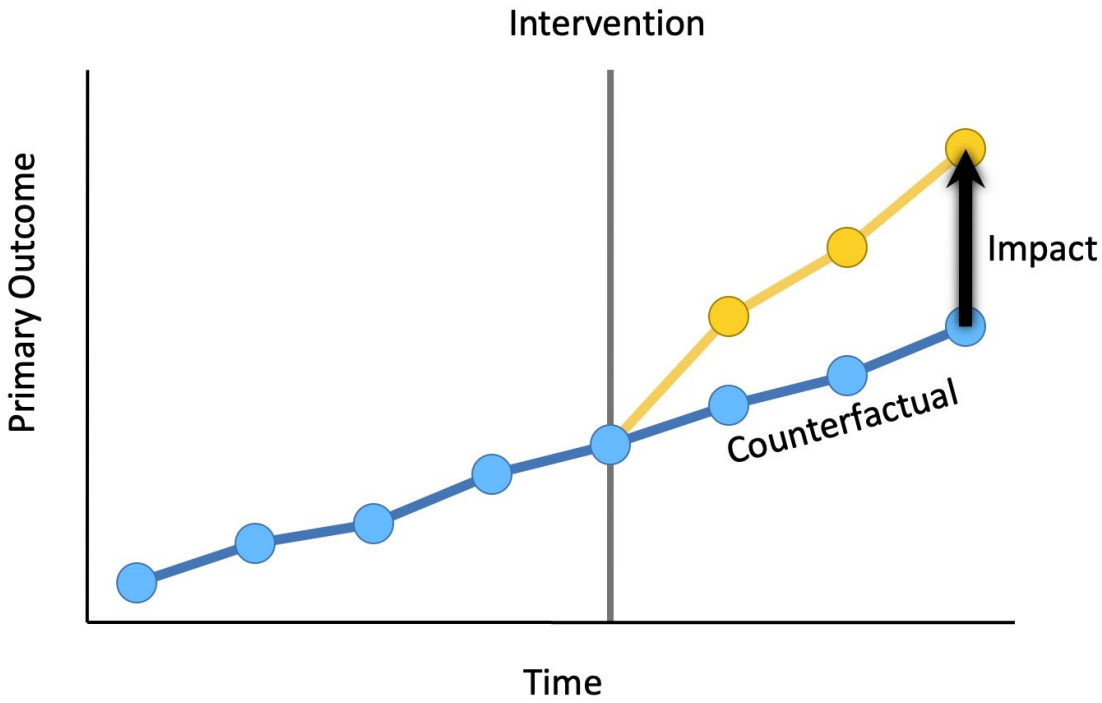


The biblical story of Daniel, often cited as the first controlled experiment. Daniel (third from left?) realized that a proper comparison of two diets could only be made when they were given to two groups of similar individuals, chosen in advance. King Nebuchadnezzar (rear) was impressed with the results. (*Source:* Drawing by Dakota Harr.)

Measuring Impact



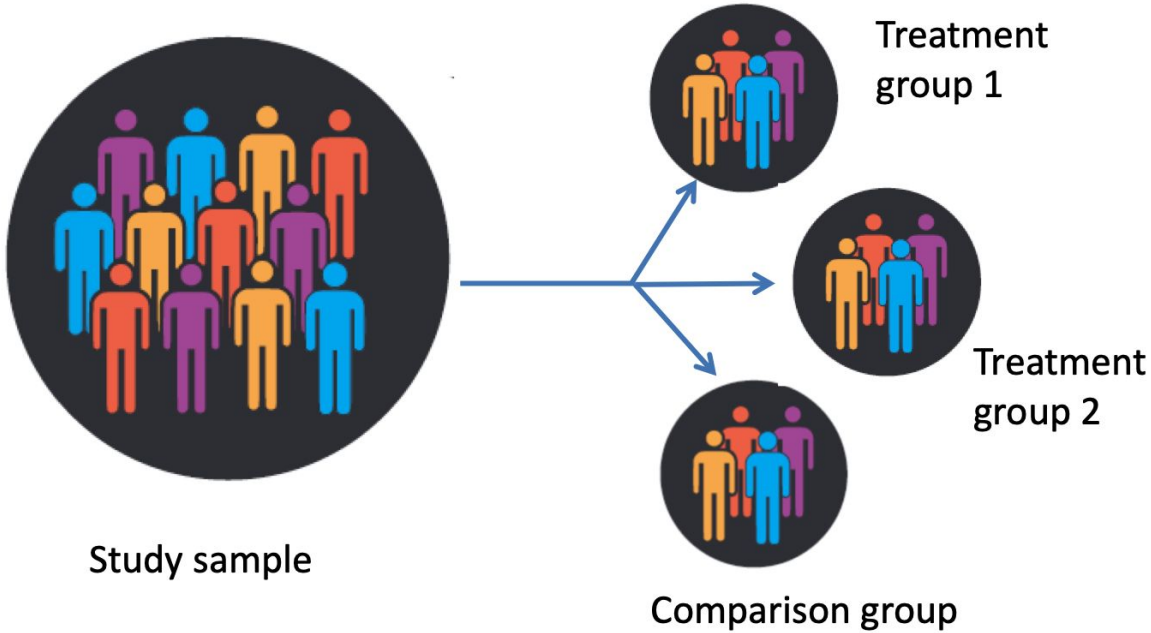
Measuring Impact



Source: http://runningres.com/lecture-notes

Creating good counterfactuals and avoiding selection bias

Randomization creates groups with similar characteristics

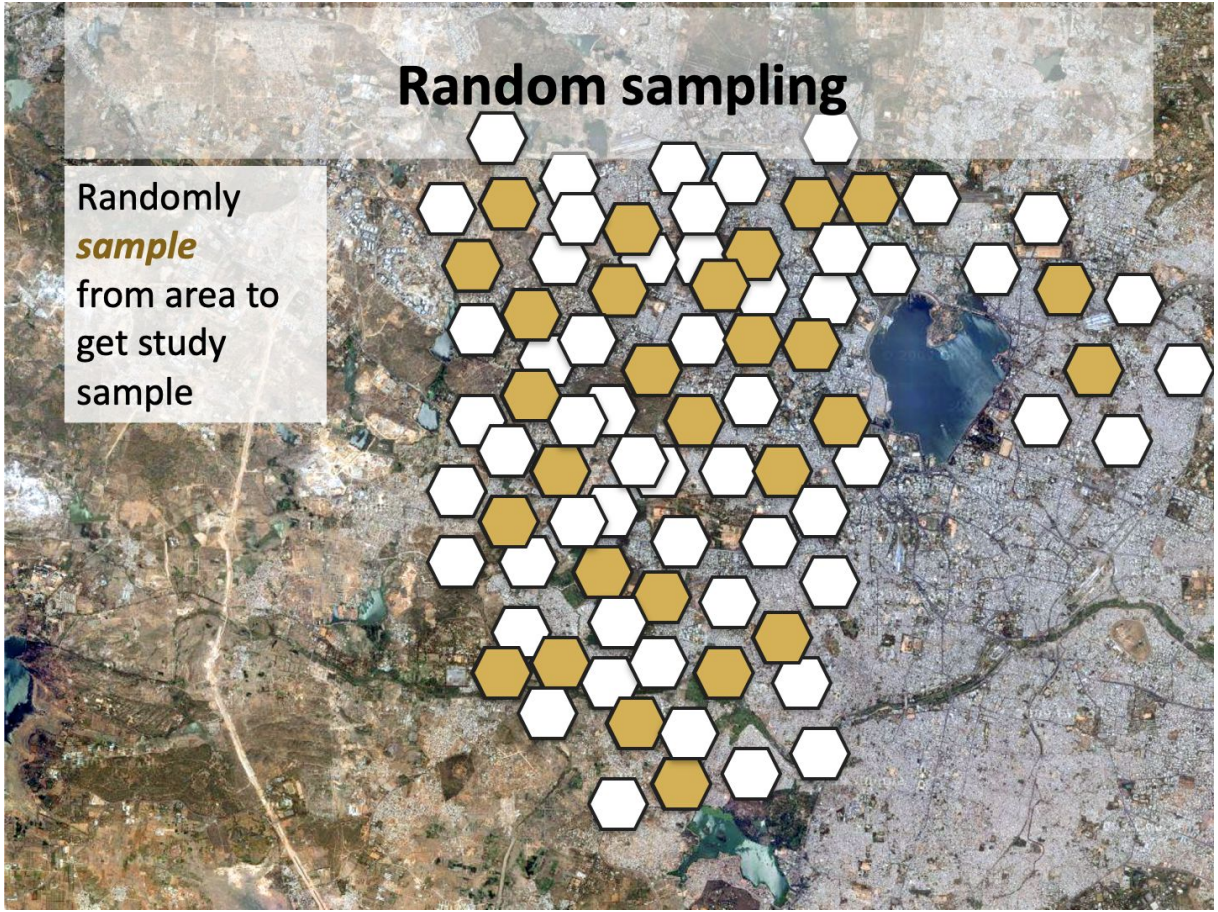


Source: http://runningres.com/lecture-notes

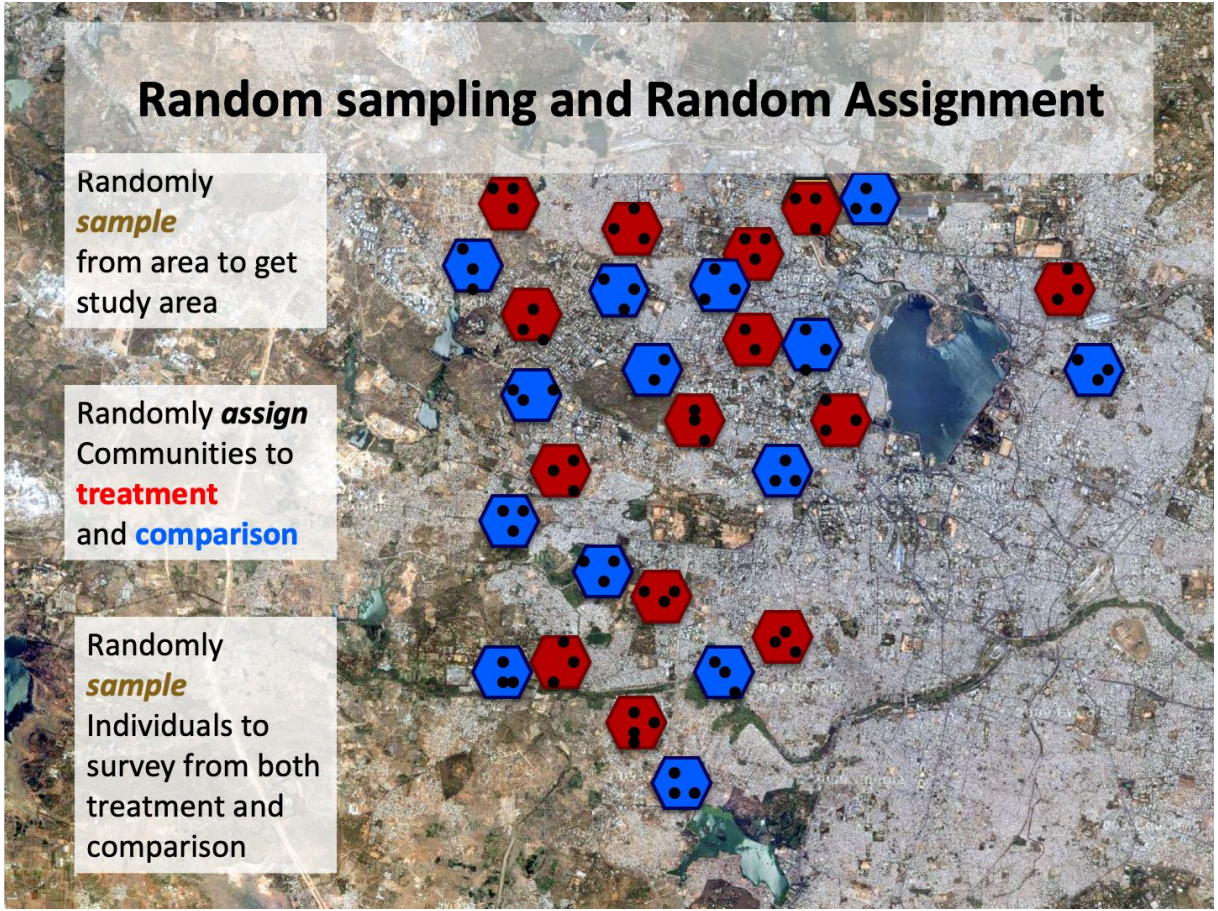
Random assignment vs. random sampling

- Random assignment
  - Units (people, schools etc.) are randomly assigned to different groups (e.g. treatment and comparison)
  - Creates two or more comparable groups
  - Basis of randomized evaluation
- Random sampling
  - Want to measure the characteristics of a group (e.g. average height)
  - Measure a random sample of the group
  - Often used during randomized evaluations, especially group level randomization

Source: http://runningres.com/lecture-notes

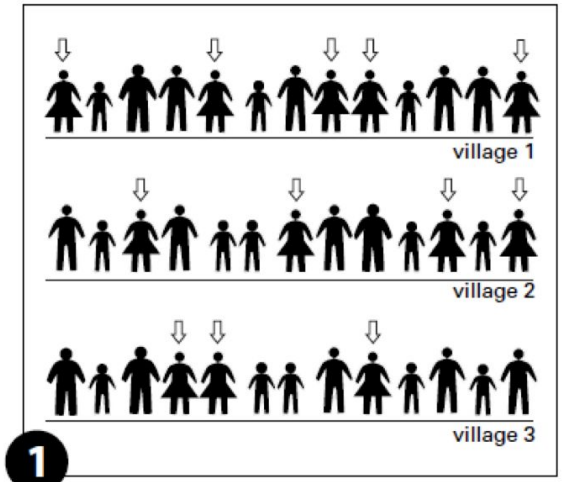


Source:
http://runningres.com/lecture-notes



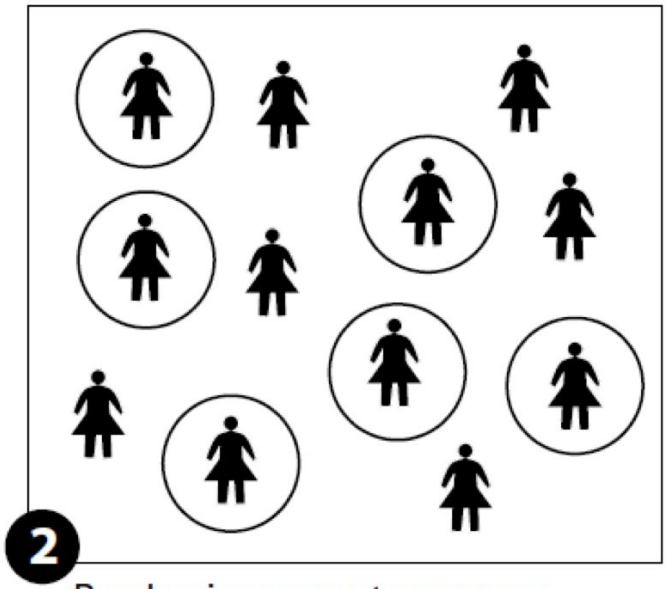
Source: http://runningres.com/lecture-notes

Steps to randomization: 1



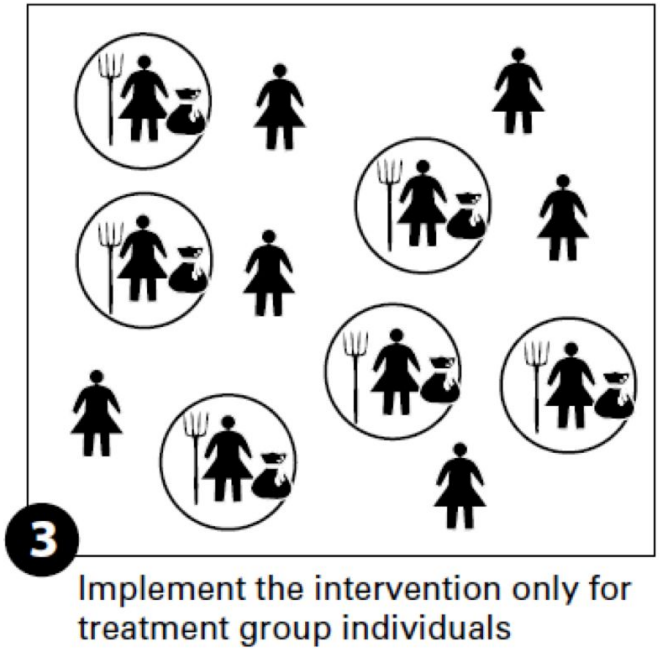
Source: http://runningres.com/lecture-notes

Steps to randomization: 2



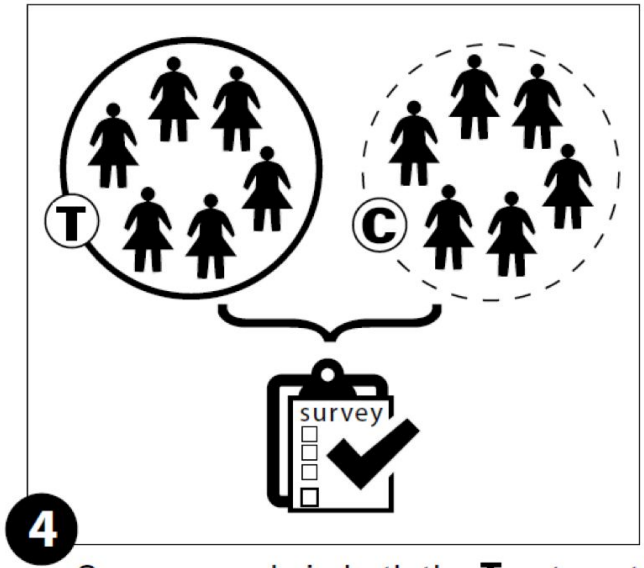
Source: http://runningres.com/lecture-notes

Steps to randomization: 3



Source: http://runningres.com/lecture-notes

Steps to randomization: 4



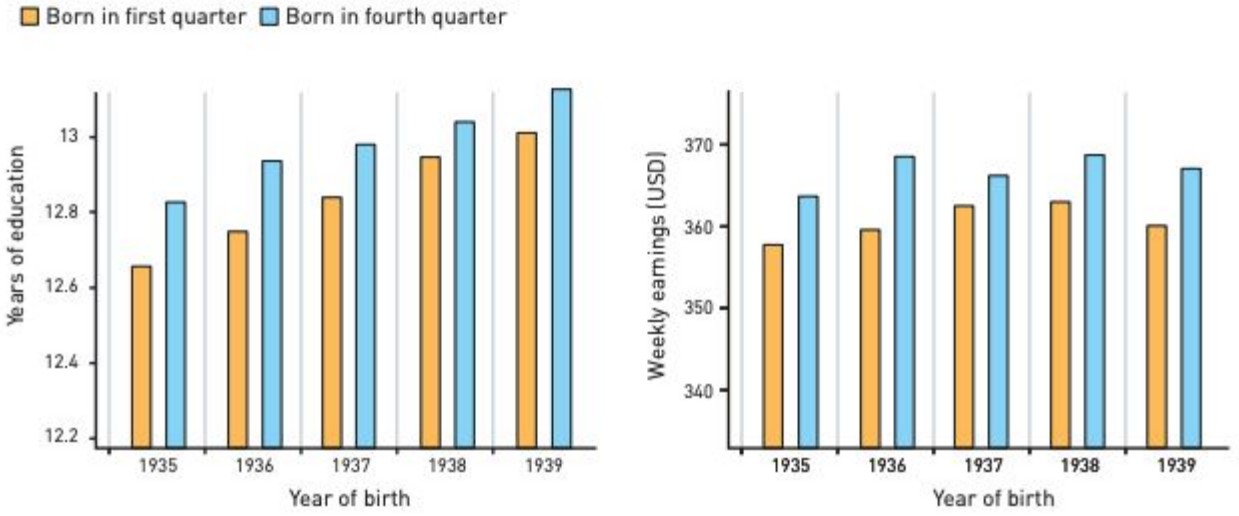
Survey people in both the **T**reatment Group (get program) and the **C**omparison Group (no program)

Source: http://runningres.com/lecture-notes

Natural Experiments

People born late in the year have more years of education and higher incomes

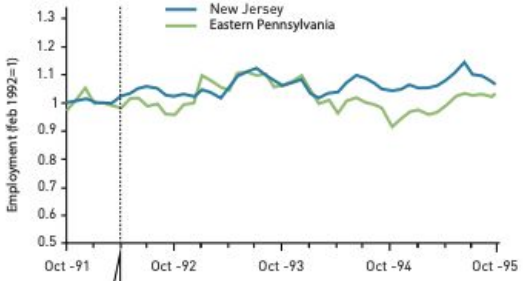
Additional years of education have a positive effect on income. The figure uses data from Angrist and Krueger (1991).



The effect of increasing the minimum wage

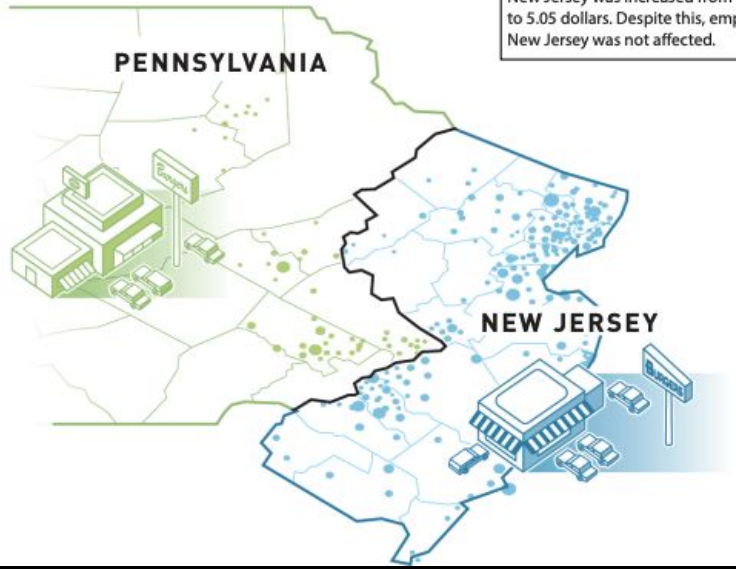
Card and Krueger used a natural experiment to study how increasing the minimum wage affects employment.

The researchers identified a treatment group (restaurants in New Jersey) and a control group (restaurants in eastern Pennsylvania) to measure the effect of increasing the minimum wage.



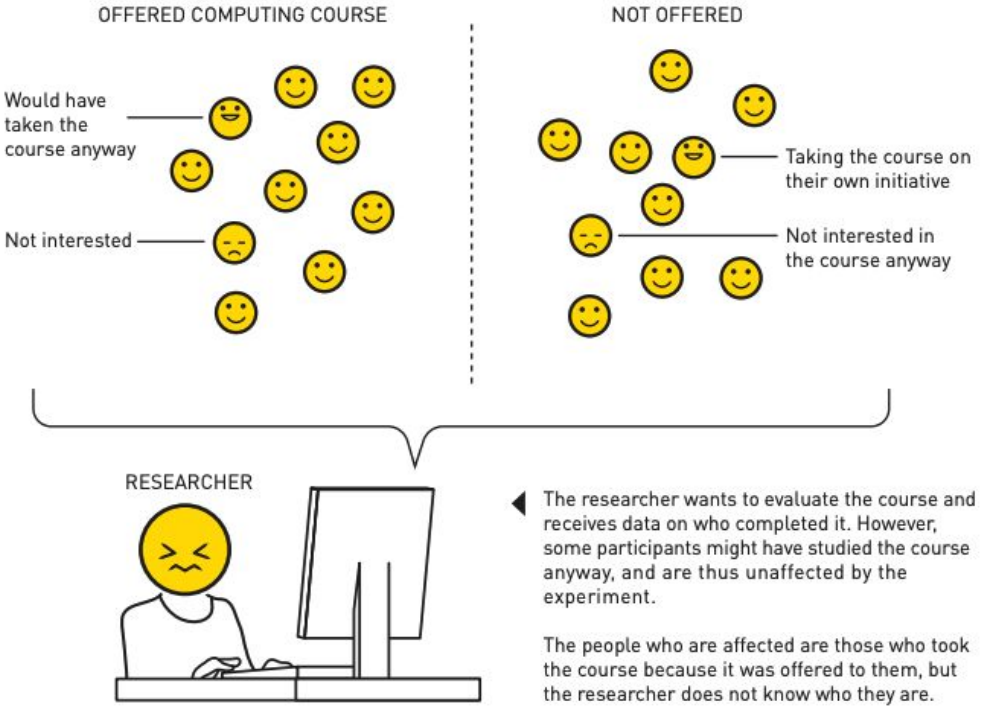
1 April 1992: The hourly minimum wage in New Jersey was increased from 4.25 dollars to 5.05 dollars. Despite this, employment in New Jersey was not affected.

CONTROL GROUP TREATMENT GROUP



Local average treatment effect

Joshua Angrist and Guido Imbens showed how natural experiments can be used to arrive at precise conclusions about cause and effect. Natural experiments differ from clinical trials as the researcher does not have complete control over who receives the treatment.



The researcher wants to evaluate the course and receives data on who completed it. However, some participants might have studied the course anyway, and are thus unaffected by the experiment.

The people who are affected are those who took the course because it was offered to them, but the researcher does not know who they are.

Natural experiment shows hungry judges give harsher sentences

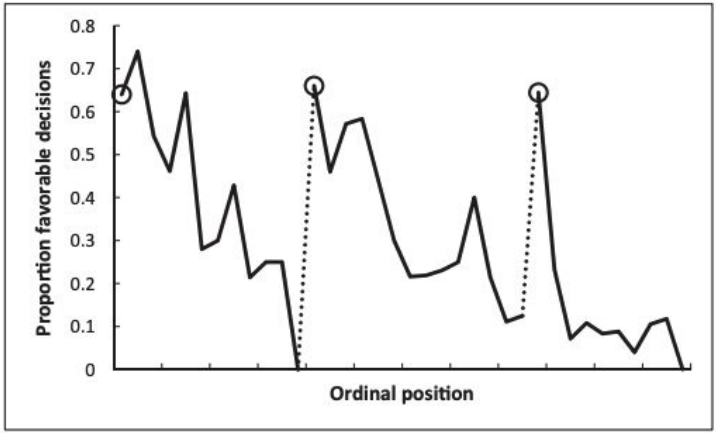


Fig. 1. Proportion of rulings in favor of the prisoners by ordinal position. Circled points indicate the first decision in each of the three decision sessions; tick marks on *x* axis denote every third case; dotted line denotes food break. Because unequal session lengths resulted in a low number of cases for some of the later ordinal positions, the graph is based on the first 95% of the data from each session.

Source: Danziger et al 2011

Fooled by non-randomness

LETTER

Overlooked factors in the analysis of parole decisions

Keren Weinshall-Margel and John Shapard

[a]*Israeli Courts Research Division, The Supreme Court, Jerusalem 91950, Israel; and*

[b]*Federal Judicial Center, Washington, DC 20002*

Source: Danziger et al 2011

Non-Random Variation in Shift Preferences

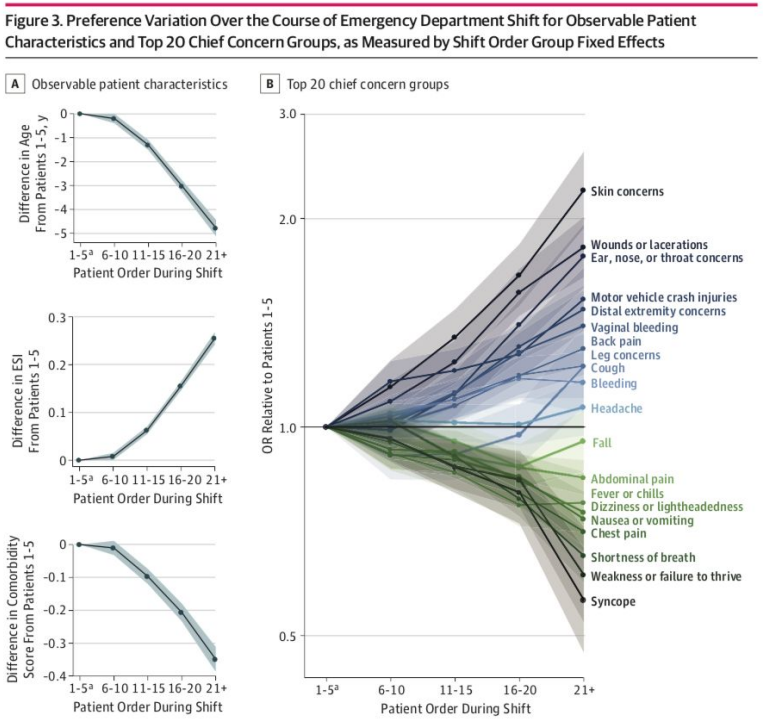


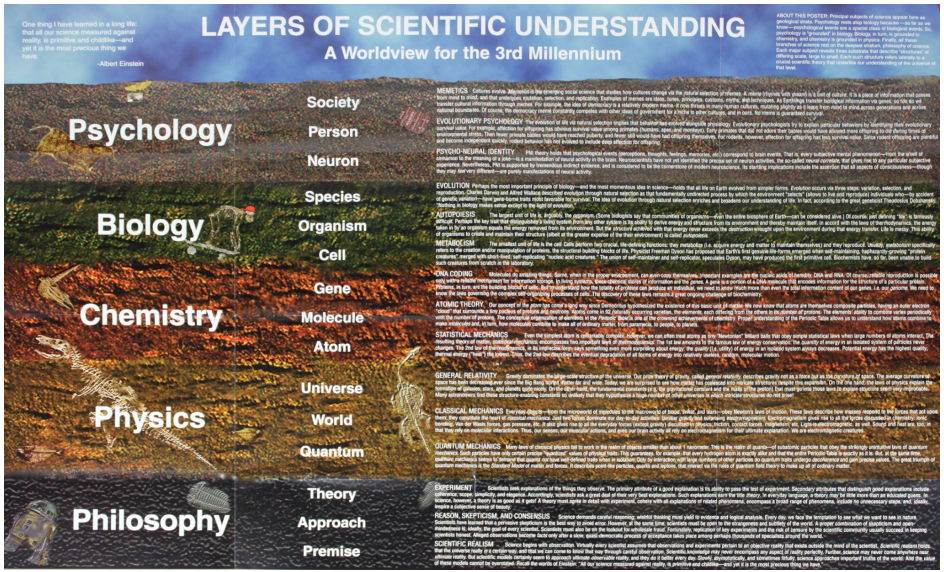
Figure 3. Preference Variation Over the Course of Emergency Department Shift for Observable Patient Characteristics and Top 20 Chief Concern Groups, as Measured by Shift Order Group Fixed Effects

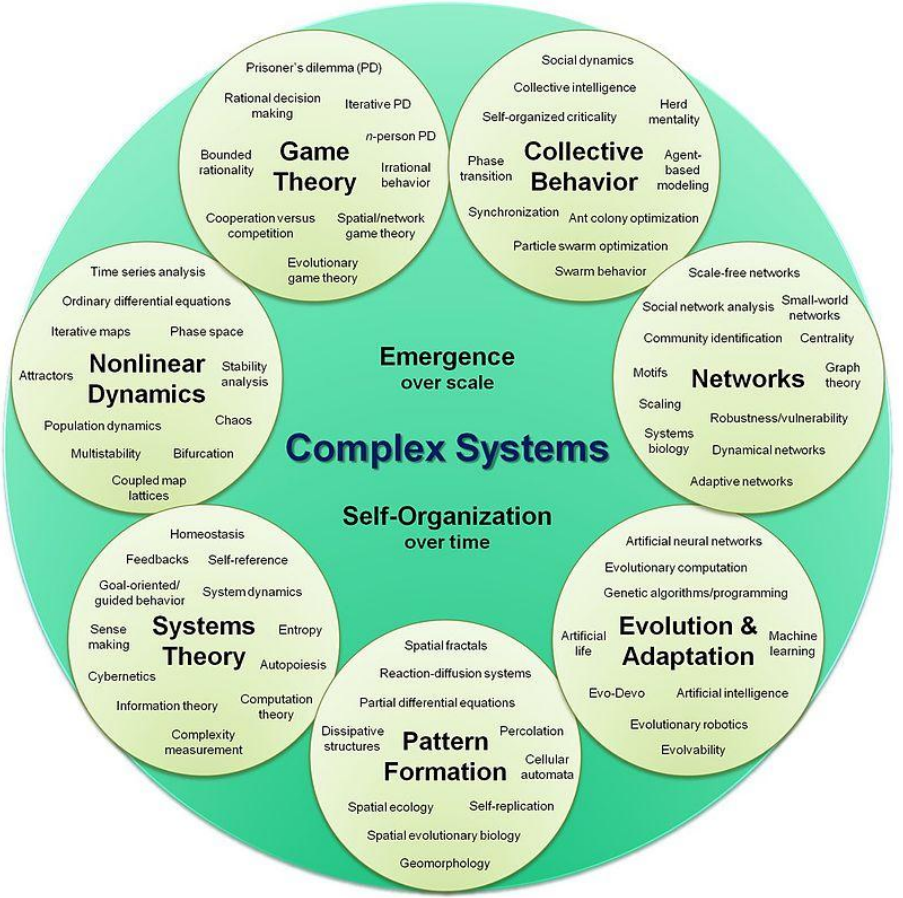
Patients in each shift were binned according to the order in which they were seen during a shift, up to 5 shift order groups: patients 1 to 5, 6 to 10, 11 to 15, 16 to 20, and 21 and more. The *F* score was $F_4 = 292.3$ ($P < .001$) for the differences in age, $F_4 = 563.4$ ($P < .001$) for Emergency Severity Index score (ESI), and $F_4 = 125.1$ ($P < .001$) for comorbidity score. Shaded areas indicate 95% CIs.

[a] Patients 1-5 was the reference group.

Source: Chang and Obermeyer 2020

Consider Complexity
Resist Reduction





Source: https://wonderfest.org/layers-of-scientific-understanding-poster/

Recap

- Black swans and the problem of induction
- Karl Popper's solution: empirical falsification
- Empirical falsification via counterfactual thinking
- Survivorship bias and probabilistic thinking
- Keeping ecological validity in mind
- Formalizing counterfactual thinking with the Potential Outcomes Framework
- Randomized experiments for causal inference
- Approximating randomized assignment with clever Nobel Prize-winning insights
- Considering complexity and resisting reduction