How to Lie with Statistics

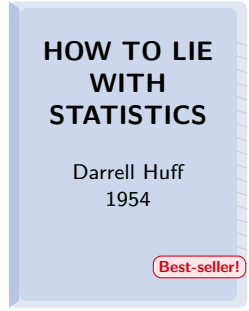
A Field Guide to Statistical Deception

(So You Can Spot It, Not Do It)

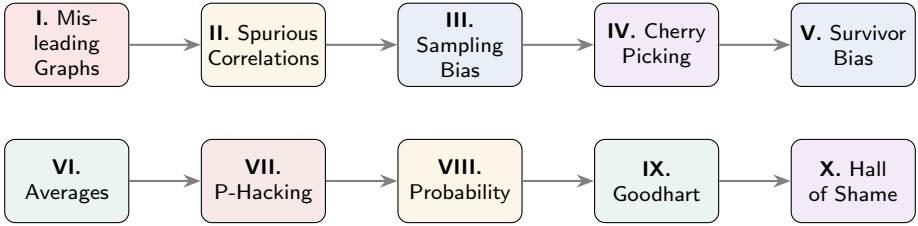
Why This Lecture?

> *"There are three kinds of lies: lies, damned lies, and statistics."*
>
> — Mark Twain (1907)

Goal: Build your statistical skepticism. After this lecture, misleading charts will **physically hurt** to look at.



Roadmap of Deception



✎ **Rule #1:** "If you torture the data long enough, it will confess to anything." — Ronald Coase

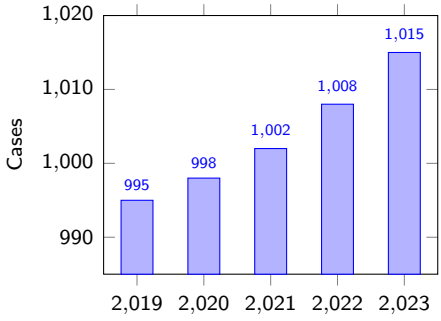
Part I

Misleading Graphs

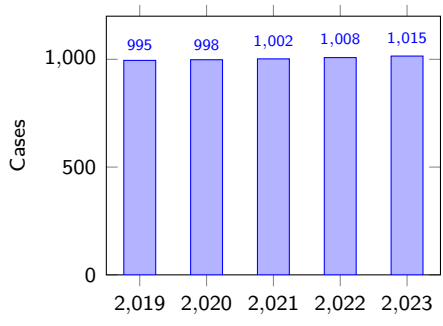
The most popular lie

Trick #1: The Truncated Y-Axis

"Crime is SKYROCKETING!"



Same data, honest axis



A 2% change looks like a 1000% change. Always check where the y-axis starts!

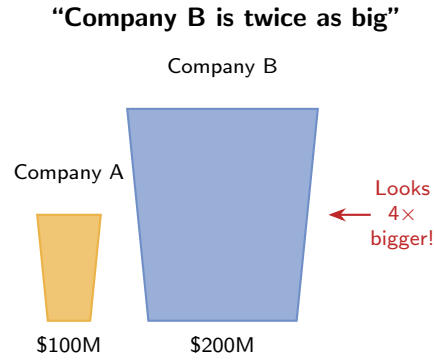
Trick #2: The 3D Pictograph

The problem with 3D:

- Perspective makes back slices look **smaller**
- Front slices appear **larger**
- Tilting distorts all proportions
- Looks fancy $\Rightarrow$ harder to read

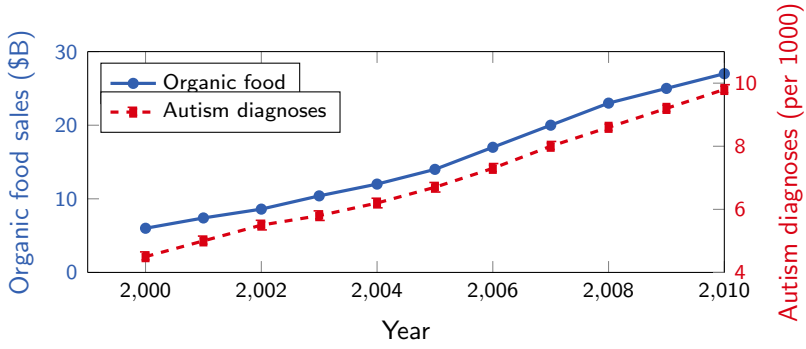
The pictograph trick:

- Double the *height* of an icon
- But it also doubles the *width*
- $2\times$ difference $\rightarrow$ looks $4\times$ bigger (area)
- In 3D: looks $8\times$ bigger (volume!)



Trick #3: The Dual-Axis Conspiracy

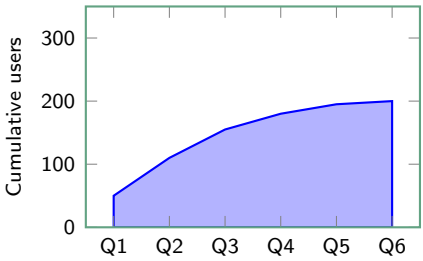
The trick: each line gets its *own* y-axis with its *own* scale. Crank the scales so the curves visually overlap — and any two trends look like they "move together," even if they're unrelated.



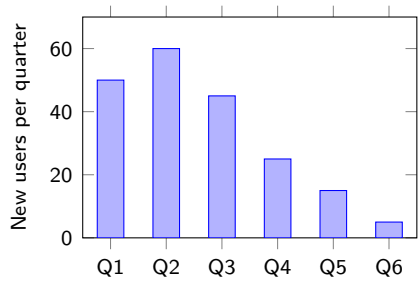
Two lines going up $\neq$ one causes the other. Almost everything trends upward over time.

Trick #4: The Cumulative Switcheroo

"Always growing!" (cumulative)



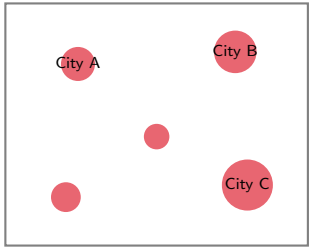
Reality (per quarter)



Cumulative charts can only go up. A startup dying at 5 users/quarter still has "200 total users and counting!"

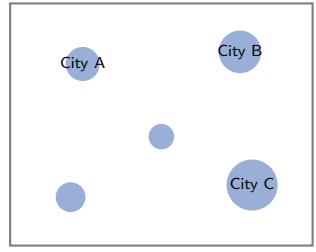
Trick #5: The Misleading Map

"Crime Hotspot Map"



"These cities are dangerous!"

Population Density Map



Same pattern = just where people live

Most "heatmaps" are just population maps in disguise.
Total counts $\neq$ rates. Always normalize by population (use *per capita*).

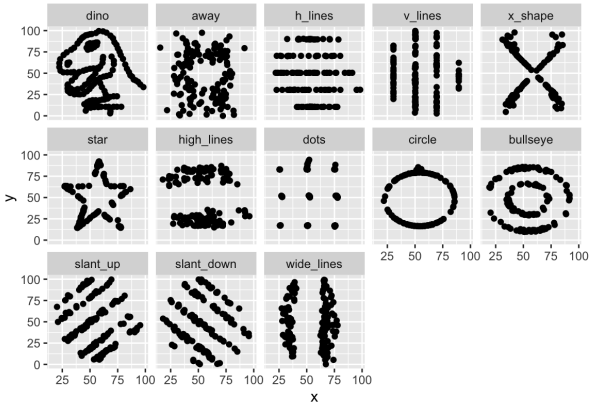
The Datasaurus Dozen: Always Plot Your Data!

All 13 datasets have:

- Same mean of X: 54.26
- Same mean of Y: 47.83
- Same std dev of X: 16.76
- Same std dev of Y: 26.93
- Same correlation: -0.06

Summary statistics can hide anything.

Always visualize your data before drawing conclusions!



(Based on Matejka & Fitzmaurice, 2017; extends Anscombe's Quartet from L2)

Part II

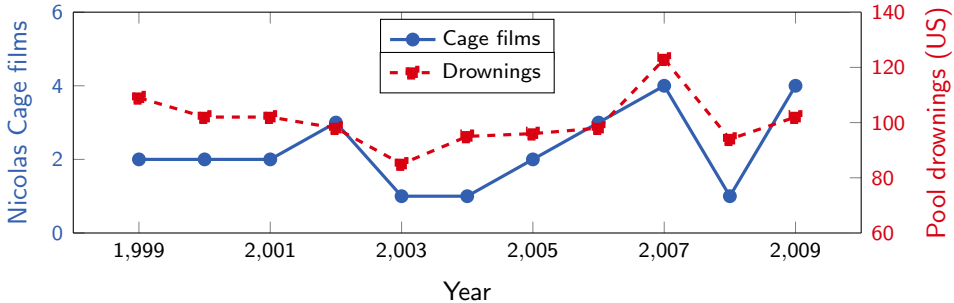
Spurious Correlations

Correlation $\neq$ Causation (the fun version)

Nicolas Cage Causes Drowning

The setup: test enough time series against each other and *some* will correlate by pure chance.

The proof: Nicolas Cage's films vs. US pool drownings, $r = 0.67$.



"Statistically significant!" Source: Tyler Vigen, *Spurious Correlations*

More Spurious Correlations (Tyler Vigen)

Cheese Consumption

vs

Deaths from getting tangled in bedsheets while sleeping

$r = 0.95$

(yes, that's a real CDC accident category)

US Science Spending

vs

Suicides by Hanging

$r = 0.99$

(your tax dollars at work?)

Margarine Consumption

vs

Divorce Rate in Maine

$r = 0.99$

(I Can't Believe It's Not Divorce!)

Internet Explorer Share

vs

Murder Rate in US

$r = 0.998$

(okay, this one might be causal)

Local version — "Tashir Pizza causes chess grandmasters": Tashir Pizza outlets in Yerevan rose 2010–2024 alongside Armenia's chess grandmaster count. Both go up because the country grew. *(That's why everything correlates with everything over time.)*

Why Does This Happen?

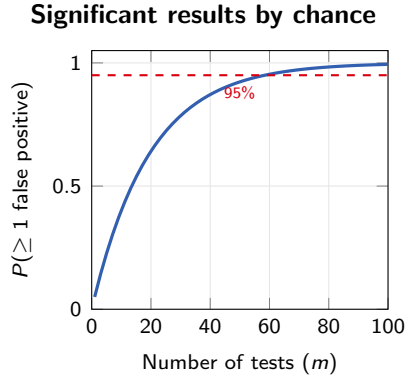
The multiple comparisons trap:

- Tyler Vigen tested ∼25,000 variable pairs
- At $\alpha = 0.05$: expect ∼1,250 "significant" results **by chance alone**
- Only show the best ones ⇒ instant "discovery"

The formula for nonsense:

$$P(\text{at least 1 false positive}) = 1 - (1 - \alpha)^m$$

With $m = 100$ tests at $\alpha = 0.05$:
$P = 1 - 0.95^{100} = 0.994$ (99.4%!)



The Ecological Fallacy: Chocolate ⇒ Nobel Prizes?

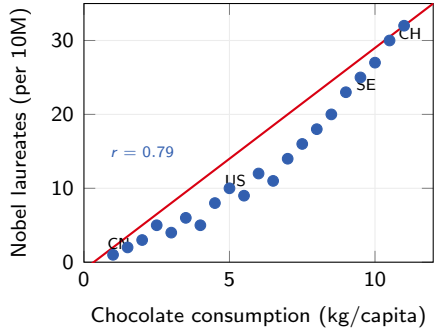
Messerli (2012), published in NEJM:

Countries that consume more chocolate per capita produce more Nobel laureates.

$r = 0.79$ (**"highly significant"**)

The ecological fallacy: Switzerland eats lots of chocolate *and* wins lots of Nobels — true of the country. But the actual laureates may not eat any! **Group-level pattern ≠ individual-level pattern.** You can't transfer a fact from one to the other.

(Lurking variable: rich countries fund both chocolate *and* research labs.)



Part III

Sampling Bias & Loaded Questions

Garbage in, garbage out

The Literary Digest Disaster (1936)

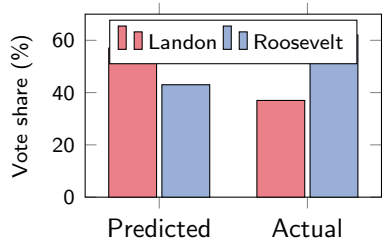
The biggest polling failure in history:

- *Literary Digest* polled **2.4 million** Americans
- Predicted: **Landon** wins (57%)
- Actual: **Roosevelt** wins (62%)

What went wrong?

- Sampled from **phone books** and **car registrations**
- 1936: only the **wealthy** had phones/cars
- Wealthy $\Rightarrow$ Republican $\Rightarrow$ Landon

2.4M biased $<$ 50K representative



Lesson: A huge biased sample is worse than a small representative one.

Local — "the Yerevan-poll trap": A TV show asks 2,000 Yerevan pedestrians, gets 72% for X, headlines it "Armenia chooses X." But Yerevan is only ∼37% of the country and skews younger/more urban. Same mistake as 1936, different city.

Loaded Survey Questions

Loaded version:

"Do you agree the government should **stop wasting** taxpayer money on failed programs?"

Result: **89% agree**

Same topic!

Neutral version:

"Should the government change its current spending on social programs?"

Result: **51% agree**

Leading:

"Do you support **helping children** get a better education?"

Result: **98% yes** (who'd say no?)

Same policy!

Specific:

"Do you support the $50B Education Reform Act funded by a 2% income tax increase?"

Result: **34% yes**

Anchoring: Random Numbers Change Your Answers

Kahneman & Tversky — the wheel of fortune:

Subject spins a wheel *(secretly rigged to land on 10 or 65)*, then is asked: "Is the % of African countries in the UN higher or lower than this number? Now give your actual guess."

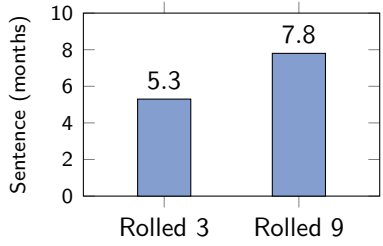
Results:

- Wheel = **10** ⇒ avg guess **25%**
- Wheel = **65** ⇒ avg guess **45%**

Subjects knew the wheel was random. It still shifted their answer 20 points.

Real-world abuse: "Was this worth \$500? \$300? \$199?" The \$500 anchor makes \$199 feel like a deal.

German judges experiment:



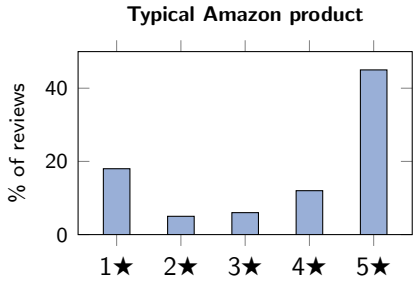
German judges read a shoplifting case, then rolled rigged dice showing **3** or **9**. They knew the dice were arbitrary — those who rolled 9 sentenced avg **7.8 months**, those who rolled 3 gave **5.3 months**.

Even professionals anchor on numbers they know are random.

Local — Vernissage: "100k dram, for you 50k." You feel like 50% off. The 100k was an anchor. The carpet was always ∼30k.

The J-Curve of Reviews

What you'd expect vs. what you get



The **J-shape**: lots of 5s, some 1s, almost nobody in the middle.

Why does this happen?

Extreme emotions motivate action. "Meh" doesn't.

Purchase bias: people who buy already liked the category.

Post-purchase rationalization: "I spent $X, so it must be good."

Fake reviews: sellers buy 5-star, competitors buy 1-star.

A 4.2-star product might be mediocre. The silent middle is invisible.

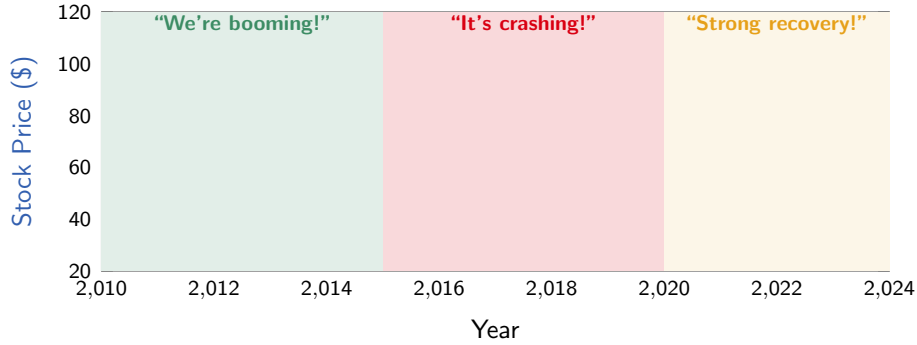
Part IV

Cherry-Picking

Choosing your data like choosing your friends

The Art of Selective Presentation

Below: **one** stock's price from 2010 to 2024. Three different commentators look at the *same* chart, each highlighting a different window. Watch how the story flips depending on where the window starts and ends.



Same dataset, three contradictory stories. **Eurovision example:** "Armenia is rising!" (2008–2010 top-10 streak) vs. "Armenia is collapsing!" (2018–2024). Same country, same data, opposite stories.

Cherry-Picking in the Wild

Climate change denial:

"Global warming stopped in 1998!"

1998 was an extreme El Niño year. Starting from the hottest year makes any trend look flat.

Vaccine "research":

"Look at these 3 studies that show harm!"

Ignoring 3,000 studies that show safety. This is also called the **file-drawer problem**.

How to spot it:

✓ Ask: *"Why this time period?"*

✓ Ask: *"What about the full dataset?"*

✓ Check if the baseline is an outlier

✓ Look for **pre-registration** of analysis

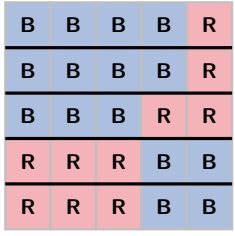
✓ Be suspicious of round-number start dates ("since 2000")

✓ Request the **raw data**

Gerrymandering: Cherry-Picking Boundaries

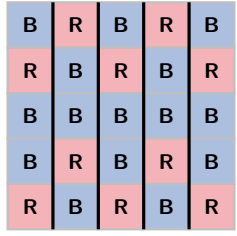
Same 25 voters, 5 districts, 3 different maps: ■ = 15 Blue ■ = 10 Red

Proportional



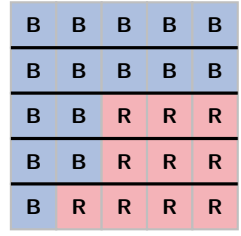
3B–2R (fair)

Blue gerrymander



5B–0R (Blue sweep!)

Red gerrymander



2B–3R (Red wins!)

Same voters, same preferences — completely different outcomes.

The person who draws the boundaries controls the result.

Part V

Survivorship Bias

The dead don't talk (about their data)

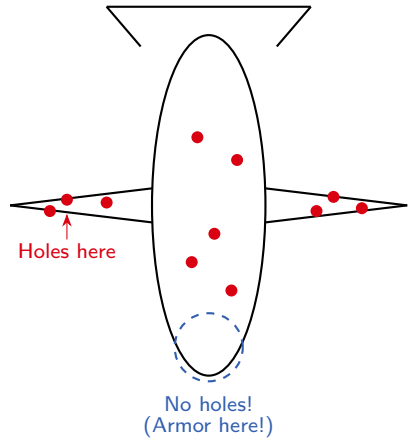
Abraham Wald and the Missing Bullet Holes

WWII, 1943:

- Bombers return with bullet holes
- Military wants to add armor where holes are
- Statistician Abraham Wald says: **No!**

The holes show where planes can be hit and survive.

Armor the places with *no holes* — those are where planes were hit and **never came back**.



You only see the survivors.

The dead planes are missing from your data.

Survivorship Bias Everywhere

Mutual Funds

"Our fund returned 15%!"

The 50 funds that **lost** money were quietly closed. You only see winners.

~60% of funds are dissolved within 15 years

Startups

"Drop out like Gates and Zuckerberg!"

You don't hear from the millions who dropped out and **failed**.

~90% of startups fail

Music

"Practice and you'll be a star!"

Millions practiced just as hard. You only interview the famous ones.

Success = skill + luck + timing

Buildings

"They built things to last back then!"

No, the shoddy old buildings already collapsed. Only the good ones survived.

Selection, not quality

Local version — the 1000-year-old churches: "Geghard, Tatev, Khor Virap — still standing after a millennium. They built things differently back then!" No — they built **thousands** of small chapels and most collapsed in the 1679 or 1988 earthquakes, or quietly turned to rubble over centuries. We see only the survivors. The dead chapels never made it onto a postcard.

The WWI Helmet Paradox

1916: British Army introduces the Brodie steel helmet.

Result: Head injury rates *went up*.

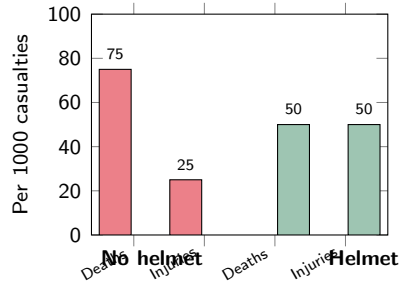
Military brass considered **removing** the helmets!

The explanation:

Before helmets: hit in head ⇒ **dead** (counted as "killed", not "head injury")

After helmets: hit in head ⇒ **survived** (now counted as "head injury")

The helmets worked! Deaths went down. But the *denominator changed*.



Head injuries "doubled" — but only because soldiers **stopped dying**.

Lesson: When an intervention changes who survives, rates can be misleading.

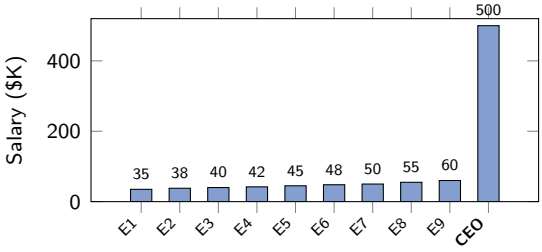
Part VI

The Average Lie

Mean, median, and the art of misdirection

Which "Average" Do You Mean?

10 employees at a company:



Mean: \$91,300

Median: \$46,500

Mode: ∼\$40K range

The company says:
"Average salary is \$91K!"

Reality:
9 out of 10 earn $<$ \$60K
The CEO drags the mean up

Always ask: Mean or median?

Yerevan IT salaries. One CTO at PicsArt on \$25k/month drags the "average" up. The *median* junior dev is ∼\$1.2k. 90% of the workforce lives below "average."

Simpson's Paradox: The Ultimate Plot Twist

UC Berkeley Admissions, 1973

Overall	Men: 44% admitted	Women: 35% admitted	**Bias?!**
Dept A	Men: 62%	Women: 82%	✓
Dept B	Men: 63%	Women: 68%	✓
Dept C	Men: 37%	Women: 34%	≈
Dept D	Men: 33%	Women: 35%	✓

Women applied to more competitive departments!
In most departments, women did as well or slightly better.

The Will Rogers Phenomenon

"When the Okies left Oklahoma for California, they raised the average IQ of *both states*." — Will Rogers

How it works:

- Move the *worst* of Group A to Group B
- A's average goes **up** (lost its weakest)
- If that member $>$ B's average, B goes **up** too!

Both groups "improved" but **no one got better!**

Better diagnostics reclassify patients, creating the illusion of progress.

Cancer stage migration:

Early stage: 90, 92, 88, 85
Mean: **88.8**

↓ Reclassify 85

Late stage: 30, 25, 20
Mean: **25.0**

Early: 90, 92, 88
Mean: **90.0** (↑)

Late: 30, 25, 20, 85
Mean: **40.0** (↑)

Percentage Tricks

Trick 1: The Round Trip

Stock goes **up 50%**:
\$100 $\rightarrow$ \$150

Then **down 50%**:
\$150 $\rightarrow$ \$75

You lost 25%!

Up 50% + down 50% $\neq$ break even

Trick 2: The Triple Cut

"We reduced errors by 50%, three times in a row!"

$100 \rightarrow 50 \rightarrow 25 \rightarrow 12.5$

That's **87.5%** reduction, not 150%.

Percentages don't add.

The "200% more" trap: Product A has 1g of protein. Product B has 3g. B has "200% more" (or "3$\times$ as much"). These sound different but mean the same thing. Advertisers pick whichever sounds more impressive.

Regression to the Mean

Daniel Kahneman's flight instructor story:

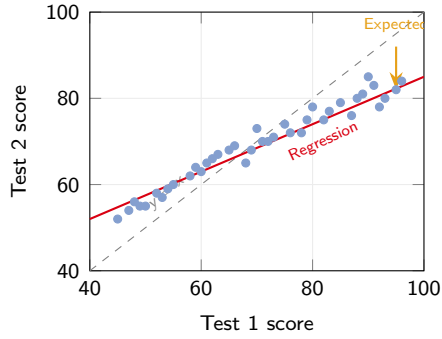
Instructors noticed: after **praise** for a good landing, the next one was usually **worse**. After **punishment** for a bad landing, the next was **better**.

"See? Punishment works!"

No. Extreme performances are followed by average ones — regardless of praise or punishment.

The Sports Illustrated "Cover Jinx":
Athletes featured after a great season tend to do worse next year. Not a curse — just regression to the mean.

Extreme scores regress to average



The Law of Small Numbers

Which US counties have the lowest kidney cancer rates?

Answer: Small, rural, sparsely populated counties.

Which counties have the highest rates?

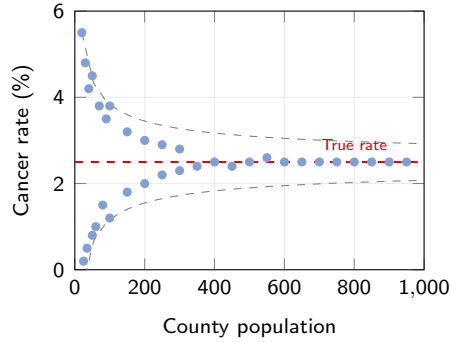
Also small, rural, sparsely populated counties!

Why? Small samples produce **extreme results in both directions**.

A county with 100 people: 0 or 3 cases = 0% or 3% rate.

A county with 1M people: rate is stable near the true mean.

Sample mean variance shrinks with n



Classic funnel plot: extremes are always small-n

The Dunning-Kruger Effect

Kruger & Dunning, 1999 (Cornell): Gave students a grammar test, then asked each: "what percentile do you think you scored in?"

Result, sorted by actual quartile:

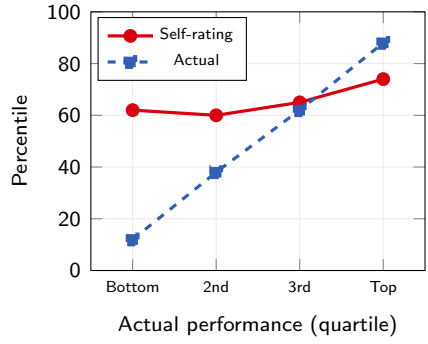
- Bottom 25% scored **12th** percentile, rated themselves **62nd**

  $\Rightarrow$ **overconfident by 50 points**
- Top 25% scored **88th**, rated themselves **74th**

  $\Rightarrow$ **underconfident by 14 points**

The story everyone tells: "Incompetent people don't know they're incompetent. Experts modestly underestimate themselves."

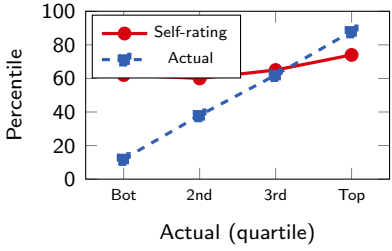
The classic D-K chart



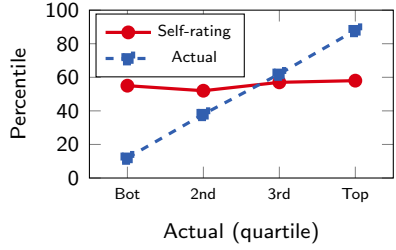
The gap between lines is "overconfidence" at the bottom, "underconfidence" at the top.

Plot Twist: The Same Pattern from Pure Noise

Real D-K data (1999)



Random numbers — no real effect



Same fan pattern. But the right chart came from 1000 *random numbers* for self-rating, with **zero relationship** to actual score.

Why does it look identical? When you sort people by actual score, the bottom quartile is by definition the low scorers. Their self-rating is independent, so its average is just the population mean ($\sim$55). Bottom-actual (12) vs. mean self-rating (55) = "overconfidence by 43 points" — from nothing.

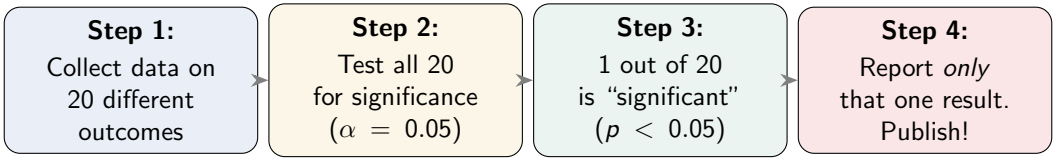
This is **regression to the mean** doing all the work. There may still be a real D-K effect, but *this chart cannot prove it* — it would look the same if there weren't one.

Part VII

P-Hacking & Bad Science

How to get published with zero real findings

The P-Hacking Playbook



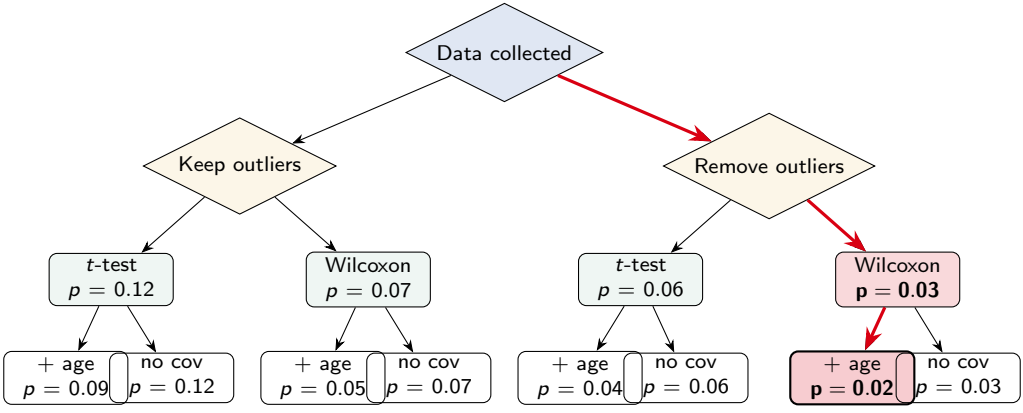
The XKCD Jelly Bean Experiment:

Scientists test if jelly beans cause acne. Test 20 colors. Green jelly beans show $p < 0.05$.

Headline: "Green Jelly Beans Linked to Acne!"

Expected false positives with 20 tests: $20 \times 0.05 = 1$. You just found the noise.

The Garden of Forking Paths



16 possible analyses $\Rightarrow$ high chance of finding $p < 0.05$ somewhere

The Texas Sharpshooter Fallacy

The metaphor:

1. Shoot randomly at a barn
2. Find a cluster (by chance)
3. Paint the bullseye around it
4. Claim you're a sharpshooter!

In statistics:

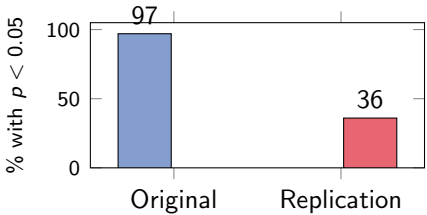
1. Collect lots of data
2. Find an interesting pattern
3. Construct a hypothesis that "predicts" it
4. Publish as if you predicted it all along

The fix: **Pre-registration.** State your hypothesis *before* seeing the data.

The Replication Crisis

Open Science Collaboration (2015):

- Replicated 100 psychology studies
- Originally: 97% had $p < 0.05$
- Replication: only **36%** replicated



Famous non-replications:

✗ **Power posing** makes you confident
(Carney et al., 2010 — failed to replicate)

✗ **Ego depletion** drains willpower
(Baumeister — multi-lab failures)

✗ **Priming** with elderly words
makes you walk slower
(Bargh, 1996 — failed to replicate)

Absence of Evidence $\neq$ Evidence of Absence

What the study says:

"We found **no evidence** that treatment X improves outcomes"

($p = 0.15$, $n = 30$, 4-week study)

$\neq$

What the headline says:

"Study **proves** treatment X **doesn't work**!"

(Completely different claim!)

"No evidence" could mean: the study was too **small**, too **short**, measured the **wrong outcome**, or the effect is **real but small**. Only a well-powered study with a tight confidence interval around zero gives genuine evidence of absence.

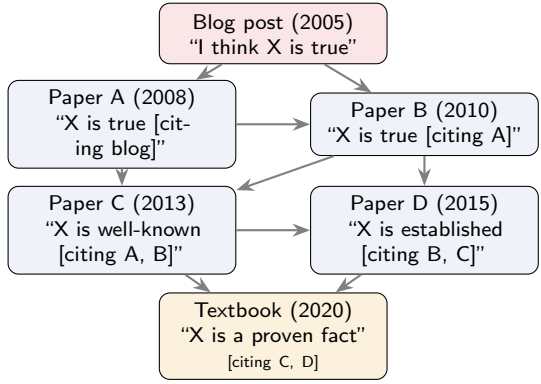
The Woozle Effect: Citation $\neq$ Proof

Named after Winnie-the-Pooh: Pooh and Piglet follow footprints in the snow, getting excited as tracks multiply. They're following *their own footprints*.

In science:

1. Someone makes an **unsupported claim**
2. A paper **cites** it as background
3. A third paper cites the second
4. Now it's "well-established" — but nobody ever **tested** it!

"63% of statistics are made up." Try to find the source. You can't.



Part VIII

Probability Illusions

Your intuition is lying to you

The Base Rate Trap

A disease test:

- Disease prevalence: 1 in 1,000
- Test sensitivity: 99%
- False positive rate: 5%

You test positive. What's the chance you're sick?

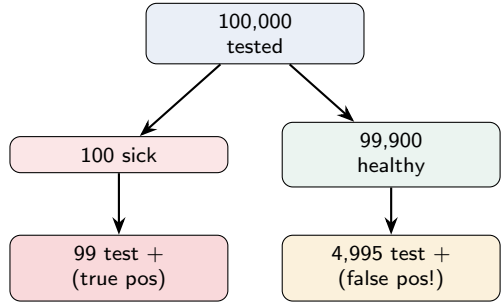
Most people guess: **95%**

Actual answer:

$$P(\text{sick}|+) = \frac{0.001 \times 0.99}{0.001 \times 0.99 + 0.999 \times 0.05}$$

$$= \frac{0.00099}{0.05094} \approx \mathbf{1.9\%}$$

Out of 100,000 people:



99 real vs 4,995 false!

Misleading Conditional Probabilities

"This drug doubles your risk of cancer!"

Baseline risk: 1 in 10,000

New risk: 2 in 10,000

Relative risk: $2\times$ (scary!)

Absolute risk: $+0.01\%$ (meh)

"9 out of 10 dentists recommend!"

How many were surveyed? 10?

What were they comparing to?

"Recommend" $\neq$ "this is the best"

Maybe they said "sure, it's fine"

Always ask: Relative or absolute risk? What's the baseline? What's the sample size?

Scary Percentages from Tiny Numbers

"SHARK ATTACKS UP 200%!"

Last year: **1** attack
This year: **3** attacks

Technically true. Completely meaningless.
Your odds: 1 in 3.7 million.

"CRIME SOARS 100% IN SMALLTOWN!"

Last year: **2** burglaries
This year: **4** burglaries

"Doubled!" Or: 2 extra incidents in a town of 5,000.

Rule of thumb: When you see a scary percentage, always ask for the **base rate**. A 200% increase from 1 is just 3. A 0.1% increase from 1,000,000 is 1,000.

The Birthday Problem

How many people do you need in a room for a 50% chance that two share a birthday?

Most people guess: **183** (half of 365)

Actual answer: **23**

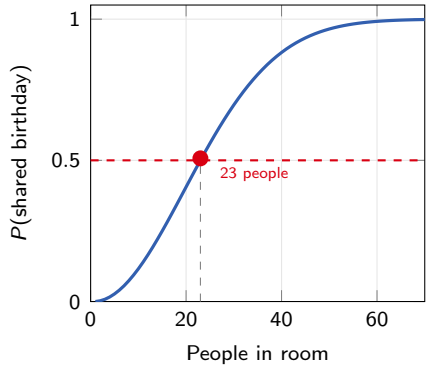
Why? You're not matching to *one* person — you're checking *all pairs*:

$\binom{23}{2} = 253$ pairs!

$P(\text{no match}) = \frac{365}{365} \cdot \frac{364}{365} \cdots \frac{343}{365}$

$= 0.493 \Rightarrow P(\text{match}) = 0.507$

The "Davit" version. Birthdays are uniform (1/365); Armenian names are not. Davit, Narek, Hayk, Tigran, Aram cover ∼15–20% of male births. In a class of 25 boys, a name collision is way more likely than the birthday math suggests.



The Gambler's Fallacy

Monte Carlo Casino, August 18, 1913. The roulette ball landed on **black 26 times in a row**. After the 15th black, gamblers piled money on red. "It's *due*!"

What's P(red on the next spin)?

The Gambler's Fallacy

Monte Carlo Casino, August 18, 1913. The roulette ball landed on **black 26 times in a row**. After the 15th black, gamblers piled money on red. "It's *due*!"

What's P(red on the next spin)?

Still $18/37 \approx 48.6\%$**.** The wheel has no memory — each spin is independent. "But it HAS to be red now!" is the **Gambler's Fallacy**. **They lost millions.**

Phantom Patterns: Your Brain Sees What Isn't There

Which sequence is "more random"?

A: H T H T H T H T T H

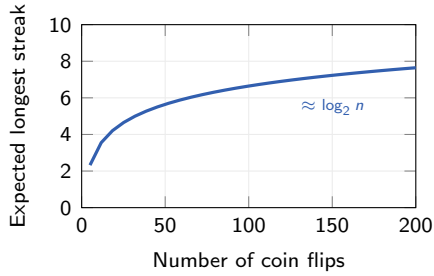
B: H H H T T H T T T H

Most people pick A. **But B is more realistic!**

Real randomness produces **streaks and clusters**. Our brains are wired to see patterns — even in pure noise.

Fraud detection: Fabricated data looks "too uniform." Real data has clumps and gaps.

Longest expected streak in n flips



In 100 flips, expect ∼7 in a row. In 1000: ∼10.

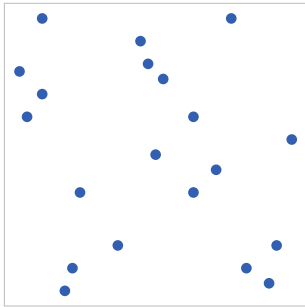
Normal, not suspicious!

Stock "technical analysis" finds patterns in random walks. Studies show it works no better than chance.

The Clustering Illusion

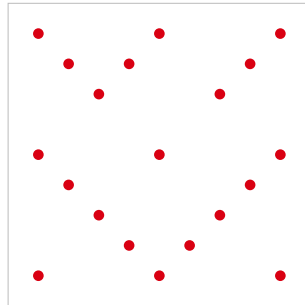
Which is random?

A



Which is random?

B



A is random. B is "too uniform" — humans placed those. True randomness has **clusters and voids**. "Cancer clusters" near power lines? Often just what randomness looks like.

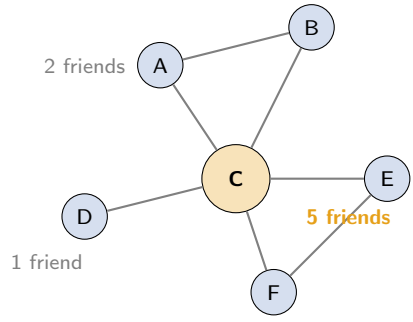
The Friendship Paradox

On average, your friends have more friends than you do.

Sounds depressing. But it's just math:

- Popular people appear in *more* friend lists
- So they're **oversampled** when you average "my friends' friend counts"
- This is **sampling bias**, not a personal failing!

Useful application: To detect an epidemic early, monitor random people's *friends* — they're more connected and get sick sooner.



C appears in everyone's friend list
$\Rightarrow$ oversampled

Benford's Law: The Fraud Detector

First digits of real-world data are NOT uniform.

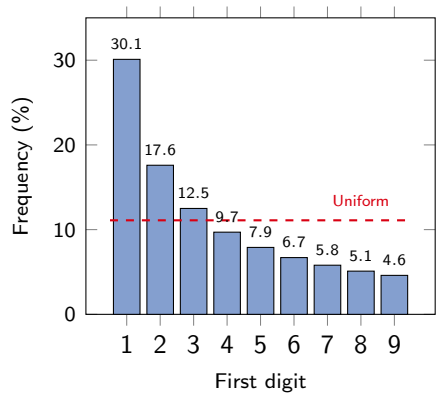
City populations, stock prices, electricity bills, river lengths, tax returns — they all follow:

$$P(d) = \log_{10}\left(1 + \frac{1}{d}\right)$$

Fraud detection: If someone invents numbers, they tend to use digits uniformly (∼11% each).

Real data: 30% start with 1!

Mismatch ⇒ likely fabricated.



Berkson's Paradox: Why Do Attractive People Seem Mean?

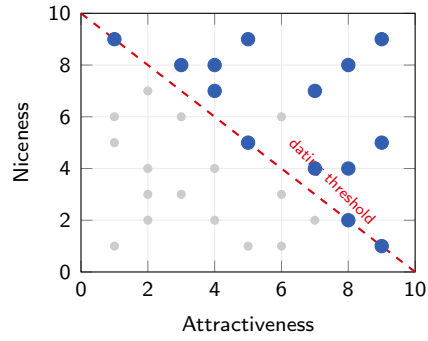
Observation: Among people you date, attractiveness and niceness seem **negatively** correlated.

Reality: In the full population, there's **no correlation** at all!

The trick: You only date people above some threshold on *attractiveness* + *niceness*.

This **conditioning on a collider** creates a spurious negative correlation:

- Very attractive + dating ⇒ don't *need* to be nice
- Very nice + dating ⇒ don't *need* to be attractive



Gray: everyone. **Blue**: your dating pool.

Negative slope appears *only* in the blue dots!

The Inspection Paradox: Why Everything Seems Worse

Buses come every 10 min on average.
Your average wait should be 5 min, right?

Nope. It's longer.

Why? You're more likely to arrive during a *long* gap than a short one, because long gaps take up more time.

The same trick: Marshrutka #253/46

Why is it *always* packed?

A packed marshrutka holds 4× more people than an empty one. So when you board, you're 4× more likely to be in the packed one.

Yours isn't always packed — you're just always in the packed ones.

Bus arrival gaps:

Mean gap = 10 min, but you landed in the 25 min gap (most probable!)

The Paradox of Unanimity

A police lineup: 6 witnesses all identify the same suspect.

Intuition: 100% agreement = strong evidence!

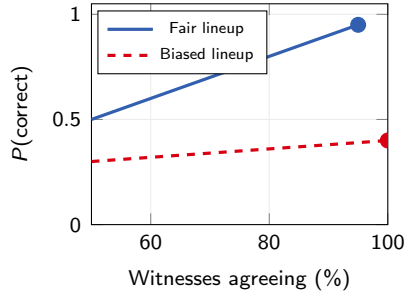
Reality: Perfect agreement is *suspicious*.

Why? Each witness has ∼10% error rate. Probability all 6 agree *correctly*: $0.9^6 = 0.53$.

But if the lineup is **biased**, all 6 agree trivially: $P \approx 1$.

Unanimity signals a systematic flaw.

More agreement ≠ more reliability



When you see 100% agreement, ask:

is the process fair, or is it rigged?

Part IX

Goodhart's Law

When the measure becomes the target

Goodhart's Law

"When a measure becomes a target,
it ceases to be a good measure."

— Charles Goodhart, 1975

Soviet Nail Factory

Target: # of nails ⇒ millions of tiny, useless nails

Target: weight of nails ⇒ one giant nail

The Cobra Effect

British India: bounty on dead cobras ⇒ people *bred* cobras for the bounty

Bounty cancelled ⇒ cobras released ⇒ *more* cobras!

Teaching to the Test

KPI: test scores ⇒ schools teach only what's on the test

Students score higher but learn less

Hospital Waits

Target: reduce ER wait times ⇒ patients held in ambulances outside

Wait "starts" when you enter the ER

Local hypothetical — chess in schools: Armenia has mandatory chess in 2nd–4th grade. Imagine a target: "avg student rating must rise 5%/year." Schools drill tournament openings instead of teaching reasoning. Metric climbs; the real goal — better thinking — gets worse.

The McNamara Fallacy

> "The first step is to measure whatever can be easily measured. The second is to disregard what can't be easily measured. The third is to presume that what can't be measured isn't important. The fourth is to say that what can't be measured doesn't exist."
>
> — Daniel Yankelovich

Named after Robert McNamara, US Secretary of Defense during Vietnam.

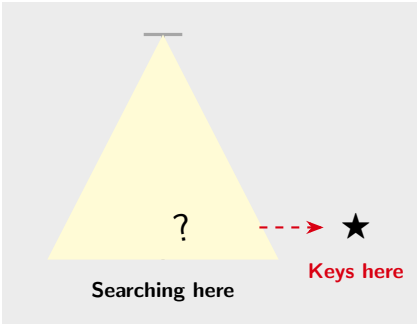
Vietnam War: Measured success by **body count**. High body count = "winning."

The US "won" every metric and lost the war.

Policing: **Arrest counts** as KPI ⇒ officers arrest easy targets, ignore hard cases.

Academia: **Paper count** as KPI ⇒ salami-slicing research into minimum publishable units.

The Streetlight Effect



"Why are you searching here?"
"Because the **light is better**."

In research and policy:

Economics: GDP is easy to measure, so we optimize it. Wellbeing, inequality, sustainability? Harder to measure ⇒ ignored.

Medicine: We study diseases with available data (hospital records), not the ones killing people in places without hospitals.

ML: We benchmark on ImageNet because it exists, not because it represents real-world vision tasks.

Policing: More police in an area ⇒ more arrests ⇒ "See, that area has more crime!" ⇒ send more police. Feedback loop.

Part X

Hall of Shame

Real-world statistical crimes

Real-World Statistical Crimes

Challenger Disaster (1986)

Engineers warned: O-rings fail in cold. Chart only showed flights *with* failures. Full data clearly showed the trend.

7 astronauts died from a chart flaw

The "Hot Hand" Debate

"Streak players score more." **1985:** No, it's an illusion! **2015:** Actually yes — original study had a statistical bias!

30 years of "debunking" was wrong

UK COVID Deaths Chart

"Deaths within 28 days of positive test." Never removed recovered patients from denominator.

Made fatality rate look much higher

Andrew Wakefield (1998)

"MMR vaccine causes autism." $n = 12$. Undisclosed conflicts. Data fabricated. Paper retracted. License revoked.

Anti-vax movement persists 25+ years

The Prosecutor's Fallacy

Sally Clark case (1999, UK):

- Two of her babies died (SIDS)
- Expert witness: "The probability of two SIDS deaths is 1 in 73 million"
- She was **convicted of murder**

The error:

- $P(\text{evidence}|\text{innocent})$ is small
- But that's NOT $P(\text{innocent}|\text{evidence})$!
- Double murder of own babies is *also* extremely rare

She was **exonerated** in 2003.

The expert was found guilty of misconduct.

The Fallacy:

$$P(\text{data}|\text{innocent}) \neq P(\text{innocent}|\text{data})$$

This is exactly **Bayes' theorem**:

$$P(H|\text{data}) = \frac{P(\text{data}|H)\,P(H)}{P(\text{data})}$$

You must consider the **prior** $P(H)$ — how common is double murder?

Your Statistical Lie Detector

10 Questions to Ask Every Chart, Claim, or Study (each links back to a Part)

1. **Where does the y-axis start? Is the scale honest?** → Parts I, IV
2. **Is it showing relative or absolute numbers?** → Parts VI, VIII
3. **What's the sample size? Who was sampled?** → Part III
4. **Mean or median? (And does it matter here?)** → Part VI
5. **Could there be a confounder or Simpson's paradox?** → Part VI
6. **Is this correlation or causation?** → Part II
7. **How many comparisons were made? (Multiple testing?)** → Part VII
8. **Who is missing from the data? (Survivorship bias?)** → Part V
9. **Who funded the study? What's their incentive?** → Parts IX, X
10. **Has it been replicated?** → Part VII

Homework: Spot the Lie

1. **Find a misleading graph** in the news or on social media. Explain what's wrong with it and redraw it honestly.

2. **Go to** `tylervigen.com/spurious-correlations` and pick your favorite. Explain *why* these variables are correlated (hint: what are the lurking variables or coincidences?).

3. **Mean vs. median salary:** A company reports "average employee compensation is \$120K." You discover that 2 executives earn \$500K and 48 other employees earn \$X each, and the median salary is \$65K. Find \$X. Which average should the company report and why?

4. **The base rate problem:** An AI system detects shoplifters with 99.5% accuracy and a 1% false positive rate. In a mall with 10,000 visitors/day, where 0.1% actually shoplift, how many innocent people get flagged? Should the mall use this system?

Resources

- **Darrell Huff** — *How to Lie with Statistics* (1954)
- **Tyler Vigen** — *Spurious Correlations* (tylervigen.com)
- **Carl Bergstrom & Jevin West** — *Calling Bullshit* (2020)
- **XKCD** — *Significant* (xkcd.com/882)
- **Edward Tufte** — *The Visual Display of Quantitative Information*
- **Open Science Collaboration** — "Estimating the Reproducibility of Psychological Science" (Science, 2015)
- **Alberto Cairo** — *How Charts Lie* (2019)

Questions?

"It is easy to lie with statistics.
It is hard to tell the truth without them."

— Andrejs Dunkels

Stay skeptical. Stay curious.