When Data Lies: Scientific Fraud and the Power of Measurement

Gov 51: Data Analysis and Politics

Scott Cunningham

Harvard University

Week 4, Thursday
February 19, 2026

A Study That Changed the Conversation

Can a single conversation change someone's mind about gay marriage?

LaCour and Green Claimed a 20-Minute Conversation Could Shift Attitudes

LaCour and Green (2014), published in *Science*:

- Door-to-door canvassing experiment in Los Angeles
- **Treatment**: 20-minute conversations about same-sex marriage
- **Control**: conversations about recycling
- Canvassers were either gay or straight (randomized)
- Outcome: feeling thermometer (0–100) toward gay people
- Measured at baseline and multiple follow-up waves

Michael LaCour was a graduate student at UCLA.
Donald Green was a senior professor at Columbia.

The Results Were Striking — and Persistent

What LaCour reported:

- Large shifts in attitudes after a single conversation
- Effects *persisted* across follow-up waves
- Gay canvassers produced larger effects than straight canvassers
- Effects even spread to other household members

This was unusual: most persuasion effects fade within days.
LaCour's effects seemed to last for months.

The Study Received Enormous Attention

- Published in *Science* — one of the most prestigious journals in the world
- Covered by the *New York Times*, *Washington Post*, NPR, and more
- Cited by political campaigns and advocacy organizations
- LaCour became a rising star — received a job offer from Princeton

But some researchers noticed something odd...

Something Doesn’t Add Up

Broockman and Kalla Tried to Build on LaCour's Work

David Broockman (Stanford) and **Joshua Kalla** (Berkeley) wanted to run a similar experiment.

- They contacted LaCour for methodological details
- As they dug into the data, they found irregularities
- They recruited **Peter Aronow** (Yale, statistician) to help investigate

What followed was one of the most important examples of *scientific self-correction* in recent history.

Your Measurement Tools Can Detect Fabricated Data

The tools you learned this week and last week are enough:

- **Histograms** reveal the shape of a distribution
- **Summary statistics** (mean, SD) describe its center and spread
- **Correlation** measures how two variables move together
- **Comparing distributions** across datasets flags anomalies

If you know what real data looks like, fake data stands out.

Fabricated data tends to look "too clean" or to match the wrong source.

Regression to the Mean

Recall: Galton Noticed Something Strange About Heredity

Remember when we discussed **Sir Francis Galton** (1886) and his study of parents' and children's heights?

- Tall parents tend to have tall children — but **not as tall**
- Short parents tend to have short children — but **not as short**
- Each generation's extremes *regress* back toward the population average

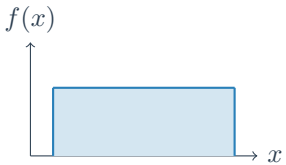
> Galton called this "regression toward mediocrity."
> We call it **regression to the mean**.

This is where the word "regression" in statistics comes from.

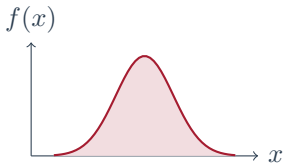
Think of Probability Like Spreading Butter on Bread

A **probability density function** (PDF) tells you how likely different values are. Imagine you have a fixed amount of butter and a slice of bread (the range of possible values).

Uniform distribution:
Spread the butter evenly. Every value is equally likely.

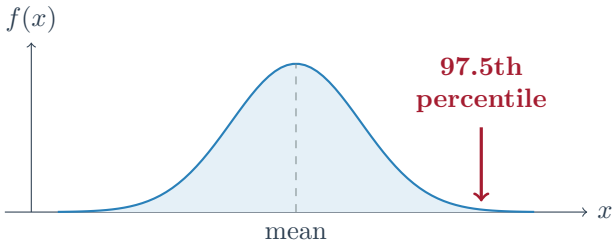


Normal distribution:
Pile the butter in the middle. Values near the center are much more likely.



Same amount of butter (total probability = 1). Different shapes.

Extreme Events Are Rare — and Unlikely to Repeat



LeBron James is at the 97.5th percentile of basketball ability. What are the odds his son is *even better*? Very small. The likely outcome: excellent — but closer to average.

Extreme values in one draw predict less extreme values in the next.

Surveys Are Noisy — Even for the Same Person

Now apply this to **feeling thermometers** (0–100).

Suppose you ask someone: "How warmly do you feel toward Harvard?"

- Monday: 95 (just got into a class they wanted)
- Wednesday: 90 (normal day)
- Friday: 99 (beautiful weather on campus)

Their "true" feeling might be around 90–95. But each measurement has **noise** — random day-to-day fluctuation.

If someone scores 99 today, they'll probably score lower next time — not because they changed, but because 99 was partly luck.

Real Repeated Measures Regress to the Mean

This has a testable implication for survey data:

- If you measure the *same people* twice on the same thermometer...
- The **correlation** between Wave 1 and Wave 2 should be high — but **not perfect**
- Typical test-retest correlations for feeling thermometers: $r \approx 0.5$ to 0.8
- A correlation near $r = 1.0$ would be suspicious

If repeated measures correlate **too perfectly**,
the data may not come from real human beings.

This is exactly what Broockman, Kalla, and Aronow checked.

What LaCour and Green Actually Claimed to Have Done

The *Science* paper reported **two separate experiments**:

- **Study 1**: Recruited a sample of LA County voters through snowball sampling. Randomly assigned them to a conversation about gay marriage (treatment) or recycling (control). Measured attitudes with a feeling thermometer at baseline and several follow-up waves.
- **Study 2**: Recruited a *different* sample of LA County voters the same way. Same experimental design, same feeling thermometer.

Two experiments, two independently recruited samples, same feeling thermometer question at each wave.

Why the Baseline Data from These Samples Should Differ

Study 1 and Study 2 surveyed **different people**. And neither group is a national sample.

Given what we know about regression to the mean:

- Feeling thermometers are **noisy** — ask the same person twice, you get different numbers
- Different people recruited through different networks will give different answers
- The distributions should look *similar* (same city, same question) but **not identical**

And they *definitely* should not match a national survey of tens of thousands of Americans.

Unless the data wasn't collected from real people at all.

LaCour's Thermometer Data Lined Up Perfectly — Three Times

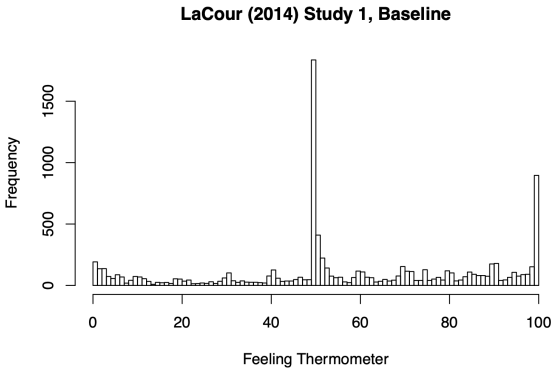
Irregularity #1: The finding that broke the case.

Broockman, Kalla, and Aronow compared the baseline feeling thermometer distributions across three datasets:

1. LaCour's **Study 1** baseline
2. LaCour's **Study 2** baseline
3. The **CCAP** (a national survey from 2012 that used the same thermometer question)

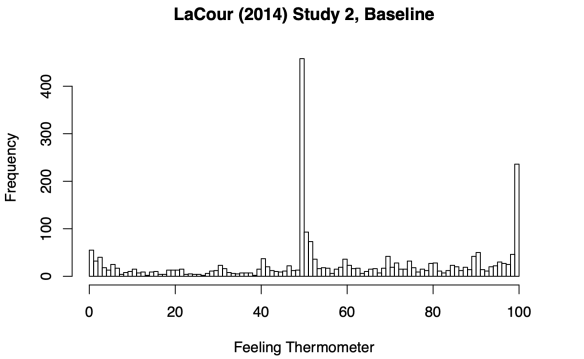
All three were statistically indistinguishable.
Three datasets that should differ — and they were identical.

LaCour's Study 1 Baseline Thermometer



Baseline feeling thermometer from LaCour's Study 1. Note the spike at 50 and at 100.

LaCour's Study 2 Baseline Thermometer — a "Different Sample"

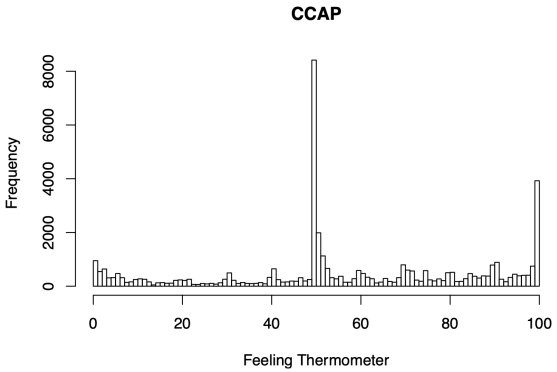



```
hist(ccap.therm, breaks=101, main="CCAP", xlab="Feeling Thermometer")
```

CCAP

Supposedly a different group of LA voters. Same spikes at 50 and 100. Same shape.

The CCAP National Survey — a Completely Different Population



Tens of thousands surveyed nationally in 2012. Different population, same question. Same shape.

A Formal Test Confirms What the Pictures Show

The **Kolmogorov-Smirnov (KS) test** asks: could these two samples have come from the same distribution? It produces a test statistic D and a **p-value** (more on p-values next week).

```
ks.test(lacour.therm.study1, lacour.therm.study2)

## Two-sample Kolmogorov-Smirnov test
## D = 0.0139, p-value = 0.8458
```

Two things to notice:

- $D \approx 0.014$ — the biggest gap between the two distributions is only 1.4%. They are nearly identical.
- The p-value of 0.85 means we **cannot say they are different**.

But what happens when they compare LaCour's data to *other* surveys?

LaCour's Data Matches Only the CCAP — Nothing Else

The KS test found **no detectable difference** between LaCour's data and the CCAP ($D = 0.009$, $D = 0.014$).

But when they compared LaCour's data to *other* feeling thermometer surveys:

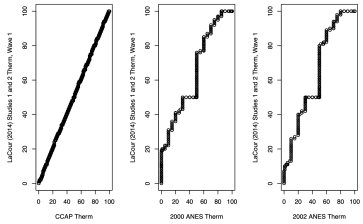
- vs. 2000 ANES: $D = 0.17$ (19× larger)
- vs. 2002 ANES: $D = 0.28$ (31× larger)
- vs. Philadelphia and Miami surveys: easily distinguishable

LaCour's data is indistinguishable from the CCAP — the one dataset he had access to — and nothing else.

Q-Q Plots: LaCour vs. CCAP and vs. ANES

Each dot compares a quantile of LaCour's data to the same quantile of another survey. If the distributions match, the dots fall on a **perfect 45-degree line**.

```
par(mfrow=c(1,3))
qqplot(ccap.therm, lacour.therm, xlab = "CCAP Therm",
       ylab = "LaCour (2014) Studies 1 and 2 Therm, Wave 1")
qqplot(anes2000, lacour.therm, xlab = "2000 ANES Therm",
       ylab="LaCour (2014) Studies 1 and 2 Therm, Wave 1")
qqplot(anes2002, lacour.therm, xlab = "2002 ANES Therm",
       ylab="LaCour (2014) Studies 1 and 2 Therm, Wave 1")
```



Left: LaCour vs. CCAP — perfect line. Middle & right: LaCour vs. ANES — clear deviations.

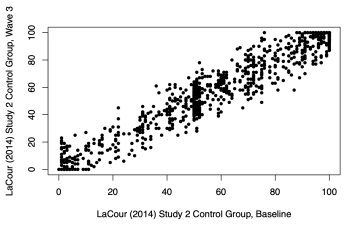
Three datasets that should differ all look identical. That doesn't happen with real human measurements.

How Did He Fake the Follow-Up Waves?

The Control Group's Responses Were Too Stable Over Time

Irregularity #2: The **control group** received no treatment. Their scores across waves should reflect only measurement noise.

```
plot(lacour.study2.controlgroup$Therm_Level.1,
     lacour.study2.controlgroup$Therm_Level.3, pch=20,
     xlab = "LaCour (2014) Study 2 Control Group, Baseline",
     ylab = "LaCour (2014) Study 2 Control Group, Wave 3")
```



Each dot is one person: baseline (x) vs. Wave 3 (y). The points hug the diagonal *too tightly* for a noisy instrument. Correlations across waves: $r = 0.93$ to 0.97.

If you feel "so-so" about something and have to pick a number from 0 to 100, what do you say?

Real Humans Heap on Round Numbers — Especially 50

Feeling thermometer responses show **heaping**: people cluster on round numbers. In LaCour's baseline data (which matches the CCAP), 231 out of 1,203 people answered exactly 50.

That's expected. If you're ambivalent, you say 50. Not 48, not 53 — fifty.

But in the follow-up waves, the heaping at 50 *disappears*:

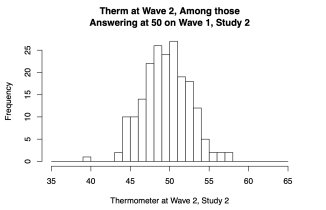
	Answered exactly 50	Total
Wave 1 (baseline)	231	1,203
Wave 2	34	1,039
Wave 3	23	1,055
Wave 4	20	1,066

If these were real people answering real surveys, round-number heaping should persist.

Clue #1: The 50-Heapers Got Scattered by a Bell Curve

Among the 231 people who answered exactly 50 at baseline, their Wave 2 answers:

```
hist(lacour.study2.controlgroup$Therm_Level.2[lacour.study2.controlgroup$Therm_Level.1==50],
     breaks=seq(from=35,to=65,by=1),
     main = "Therm at Wave 2, Among those\nAnswering at 50 on Wave 1, Study 2",
     xlab = "Thermometer at Wave 2, Study 2")
```



That's a **normal distribution** — the bell curve — centered on 50. Real humans don't scatter like this; they keep heaping on round numbers. But R's `rnorm()` generates exactly this shape. The authors started to suspect someone was adding computer-generated noise to the baseline.

Clue #2: Impossible Values Create a New Pile-Up at Zero

The normal distribution stretches from $-\infty$ to $+\infty$. So adding `rnorm()` noise creates impossible values: $5 + (-9) = -4$, or $98 + 7 = 105$. But thermometers only go 0–100. The fabricator has to truncate — replace anything < 0 with 0 and anything > 100 with 100. That creates a pile-up at the boundaries:

	Answered exactly 0	Total
Wave 1 (baseline)	1	1,203
Wave 2	38	1,039
Wave 3	38	1,055
Wave 4	45	1,066

Only 1 person said 0 at baseline. Then suddenly 38–45 people say 0 in every follow-up wave.

The Recipe: Baseline + Normal Noise + Truncation

The bell-curve scattering and the impossible zeros pointed to a hypothesis. LaCour generated his follow-up data mechanically:

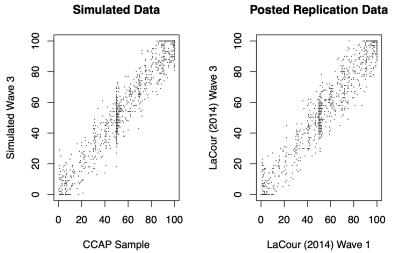
```
# Start with CCAP data as the baseline
wave3 <- round(ccap.therm + rnorm(n, mean = 0, sd = 8.4))

# Truncate: thermometers only go 0 to 100
wave3[wave3 < 0]   <- 0
wave3[wave3 > 100] <- 100
```

This three-line procedure explains *every* pattern they found:

- Baseline matches CCAP → because it *is* the CCAP data
- Heaping at 50 disappears → `rnorm()` smooths it out
- Heaping at 0 appears → truncation of negative values
- Correlations too high → small σ means little noise added

Simulated Data Looks Identical to LaCour's "Real" Data



Left: data generated by the three-line recipe. Right: LaCour's actual posted data. A KS test cannot distinguish them ($D = 0.03$, $p = 0.22$).

They reverse-engineered *exactly* how he faked it.

The Reckoning

Broockman, Kalla, and Aronow Published Their Findings in May 2015

Their working paper, "Irregularities in LaCour (2014)," documented **multiple** statistical anomalies — including everything we just walked through:

1. Baseline distributions matched the CCAP exactly
2. Control group wave-to-wave correlations were too high
3. Heaping patterns changed in ways consistent with `rnorm()` + truncation
4. The entire dataset could be reproduced with three lines of code
5. The survey firm LaCour named denied any involvement
6. LaCour could not produce the raw data

Each irregularity was damning alone.
Together, they were conclusive.

Green Requested Retraction; *Science* Retracted in June 2015

- Donald Green (the senior author) immediately requested retraction
- Green had not collected the data himself — he trusted LaCour
- *Science* retracted the paper without LaCour's agreement
- LaCour's Princeton job offer was rescinded
- The survey firm LaCour claimed to have used denied any involvement

LaCour never admitted to fabrication, but could not produce any evidence the study was real.

Green's Retraction Letter to *Science* — May 19, 2015

Memo

May 19, 2015

To: Gilbert Chin
From: Donald Green
Re: Retraction of "LaCour, Michael J., and Donald P. Green. 2014. When Contact Changes Minds: An Experiment on Transmission of Support for Gay Equality. *Science*. 346(6215): 1366-1369."

I write to request a retraction of the above Science report. Last weekend, two UC Berkeley graduate students (David Broockman, and Josh Kalla) who had been working on a research project patterned after the studies reported in our article brought to my attention a series of irregularities that called into question the integrity of the data we present. They crafted a technical report with the assistance of Yale professor, Peter Aronow, and presented it to me last weekend. The report is attached. I brought their report to the attention of Lynn Vavreck, Professor of Political Science at UCLA and Michael LaCour's graduate advisor, who confronted him with these allegations on Monday morning, whereupon it was discovered that the on-line survey data that Michael LaCour purported to collect could not be traced to any originating Qualtrics source files. He claimed that he deleted the source file accidentally, but a Qualtrics service representative who examined the account and spoke with UCLA Political Science Department Chair Jeffrey Lewis reported to him that she found no evidence of such a deletion. On Tuesday, Professor Vavreck asked Michael LaCour for the contact information of survey respondents so that their participation in the survey could be verified, but he declined to furnish this information. With respect to the implementation of the surveys, Professor Vavreck was informed that, contrary to the description in the Supplemental Information, no cash incentives were offered or paid to respondents, and that, notwithstanding Michael LaCour's funding acknowledgement in the published report, he told Professor Vavreck that he did not in fact accept or use grant money to conduct surveys for either study, which she independently confirmed with the UCLA Law School and the UCLA Grants Office. Michael LaCour's failure to produce the raw data coupled with the other concerns noted above undermines the credibility of the findings.

I am deeply embarrassed by this turn of events and apologize to the editors, reviewers, and readers of Science.

Should Green have caught the fraud earlier?
What does this tell us about trust and verification in science?

The Real Experiment

There Were Always Red Flags Beyond the Statistics

Even before the forensic analysis, some things didn't add up:

- LaCour claimed to have raised **millions of dollars** from donors to fund the survey — as a PhD student
- He could not document where the money came from
- The survey firm he named (**uSamp**) denied any involvement
- He could not produce the raw data files when asked

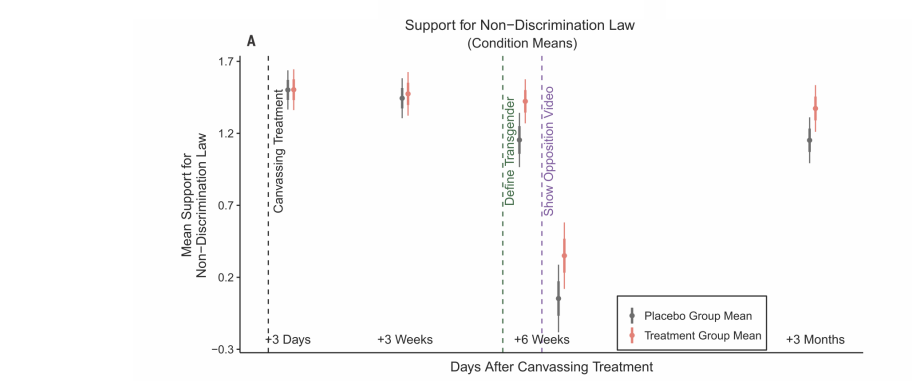
None of these alone proves fraud. But combined with the statistical evidence, the picture was clear.

Broockman and Kalla Then Did the Study for Real

After exposing the fraud, Broockman and Kalla actually ran the experiment LaCour claimed to have done. **Broockman and Kalla (2016)**, published in *Science*:

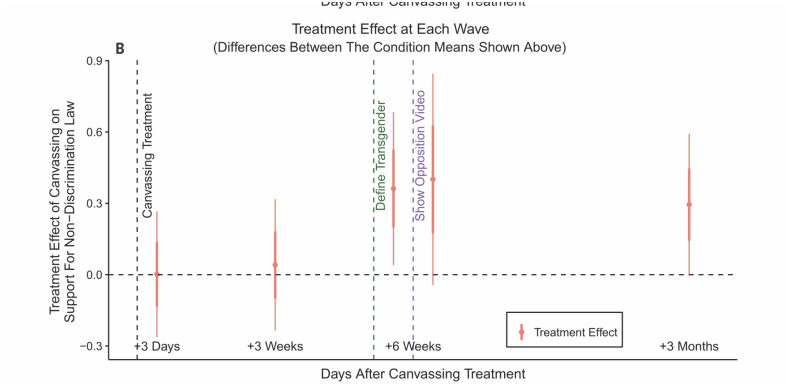
- Door-to-door canvassing in South Florida about **transgender prejudice**
- Recruited 68,378 voters; 1,825 completed the baseline survey
- Treatment ($n = 913$) vs. placebo ($n = 912$)
- 56 canvassers (15 transgender, 41 non-transgender)
- **Pre-registered**, verified survey firm, open data — all hands on the table

What Real Data Looks Like: Group Means Over Time



The **dots** are group means (you know these). The **lines** are confidence intervals (next week). The groups start similar — but *not identical*. Then they diverge after treatment.

The Treatment Effect: Differences Between Groups



Each dot is the difference between treatment and placebo means. The dashed line at 0 means "no effect." The effect persists through 3 months — with *real* data this time.

The Real Results Were More Modest — and More Interesting

What Broockman and Kalla found with real data:

- A single 10-minute conversation *did* durably reduce transphobia
- Effects lasted at least 3 months
- Both transgender and non-transgender canvassers were effective
- The effects were real but **smaller** than what LaCour fabricated

Given the damage LaCour's fraud did to this research question, it was even more important that someone do it correctly.

> The real science is more nuanced —
> and more interesting — than the fraud.

This Keeps Happening

Francesca Gino: Data Fabrication at Harvard Business School

In 2023, the blog **Data Colada** — run by three behavioral scientists — found evidence of data manipulation in multiple studies by Francesca Gino, a tenured professor at Harvard Business School.

- Harvard conducted an internal investigation
- The investigation confirmed the findings
- Gino's tenure was revoked — one of the first times in Harvard's history
- She was removed from HBS

The tools Data Colada used? The same ones we've been discussing: looking at distributions, checking for patterns that shouldn't be there, comparing datasets.

MIT Economics: A PhD Student's Data Couldn't Be Verified

In 2025, MIT's Economics Department investigated a preprint about AI and scientific discovery by a second-year PhD student.

- Professors Daron Acemoglu and David Autor were acknowledged in the paper
- MIT's Committee on Discipline reviewed the work
- MIT stated it had "no confidence in the provenance, reliability or validity of the data"
- The student is no longer affiliated with MIT

The paper was never published, but it had already begun influencing public discussions about AI.

What This Means for You

Data Science Is Rhetoric — Telling the Truth Well

The work of data science is **rhetoric**: telling the truth in a persuasive way, so that the receiver updates their beliefs *appropriately*.

This happens through individual studies. But it also happens in the aggregate — through the accumulated weight of *all* the studies. When someone fabricates data, they corrupt that shared body of evidence.

And there's a subtle danger on the receiving end, too: we are most vulnerable to fraud when the results **confirm what we already believe**. LaCour's story was appealing — a single conversation changes minds about gay marriage? People *wanted* that to be true. Our radar drops when results feel good.

> The studies we most want to believe are the ones
> that deserve the most scrutiny.

Fraud Breaks Trust — and Trust Is Everything

Aristotle identified **ethos** — the credibility of the speaker — as the first principle of persuasion. Before anyone listens to your evidence, they have to trust you.

If data science is going to improve how societies make decisions, then the people doing this work have to be the ones the world looks to and trusts — *especially* when the findings are hard to hear.

That means being the person who would not lie, not because lying is illegal, but because the whole enterprise depends on it. Every fabricated dataset makes the next honest result harder to believe.

> The most important thing a data scientist protects
> is not the data. It's the trust.

But Notice What Caught It: The Simple Things

Means | **Variance** | **Correlation** | **Normal Dist.**

Histograms. Summary statistics. Comparing distributions. The normal distribution and what it implies about real human data.

These are the tools you already have. And they were enough to unravel one of the biggest frauds in social science.

Lies leave fingerprints.

Looking Ahead

Next week: Prediction and Linear Regression (QSS Chapter 4)

- How do we predict outcomes from data?
- The linear regression model
- Least squares estimation

Problem Set 2 is due Thursday, March 5.

PS2 uses the LaCour and CCAP datasets from today's lecture. You'll reproduce the forensic analysis yourself.

Good measurement doesn't just describe data. It protects the integrity of science itself.

Questions?