

# Text as Data

From Craigslist to Congress

Scott Cunningham

Harvard University

Week 3, Tuesday

February 11, 2026

Last Thursday we learned to summarize data. Today we learn to create it.

**Thursday:** Descriptive statistics—means, medians, standard deviations

**Today:**

- ▷ Continuing our story with *real research*
- ▷ Line graphs: showing means over time or age
- ▷ Text as data: converting words into numbers
- ▷ A \$11 replication of a \$10,000 study

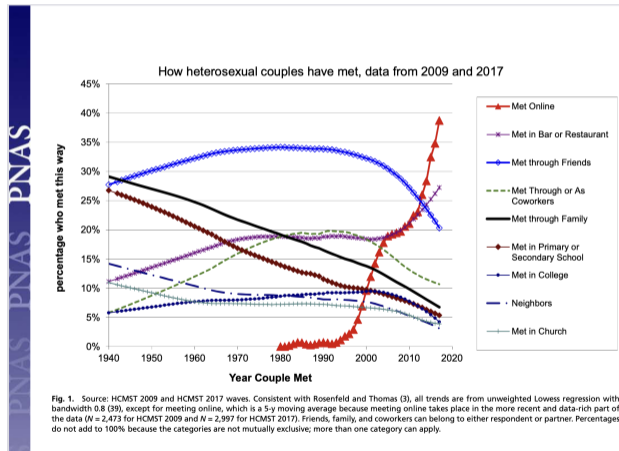
Same statistical tools, applied to messy real-world questions

The image features a minimalist design with four large, semi-transparent circles in soft pastel colors: pink, orange, teal, and light blue. These circles are arranged in a loose, asymmetrical pattern around the central text. The text 'Pictures Tell Stories' is written in a dark blue, serif font, centered horizontally and partially overlaid by the pink circle on the left.

**Pictures Tell Stories**

# How Heterosexual Couples Meet

Rosenfeld, Thomas & Hausen (2019, PNAS)



# The Felt Experience

## What People Actually Say

**Success:** *“Thanks to Tinder, I met the love of my life. We’ve been married seven years.”*

**Exhaustion:** *“Dating has become exhaustive in a way I never thought possible; the modern world has made people disposable.”*  
—Lauren

**Structural complaint:** *“A small percentage of male users receive a lot of attention. Often, they do not want to commit, and a hook-up culture prevails.”*

**The asymmetry:** *“One person holds stronger romantic feelings, hoping for commitment, while the other enjoys the perks without intending to deepen it.”*  
—Psychology Today

27% of 2024 marriages started on apps. But 78% of users report burnout.

# Recall: We can tell powerful stories with just means

## Last Thursday:

- ▷ Mean = “typical” value
- ▷ Compare means across groups
- ▷ The mean is a summary—a compression of data

## Today:

- ▷ Apply these tools to real research
- ▷ New visualization: **line graphs** (vs. histograms)
- ▷ Lines show how means change over time, age, or other continuous variables

# Case Study: Online Dating and the American Family

## My own research project

- ▷ 166,000 Craigslist Personals posts from Internet Archive
- ▷ Classified using GPT-4o-mini (\$10, a few hours)
- ▷ Research question: What were people looking for?

## Categories:

- ▷ **Romantic (R)**: Seeking long-term relationship
- ▷ **Casual (C)**: Seeking hookup or casual encounter

Craigslist Personals shut down in 2018—but the Wayback Machine preserved them

## Classifying Intent: Romantic vs. Casual

	R/C Ratio	Interpretation
Men seeking women	2.1	2× more romantic than casual
Women seeking men	5.7	6× more romantic than casual
<b>Gender gap</b>	<b>3.6</b>	Women much more romantic

**Key finding:** Romantic > Casual for both genders

If everyone wants romance, why does dating feel so hard?

# The Gender Gap: A Theoretical Insight

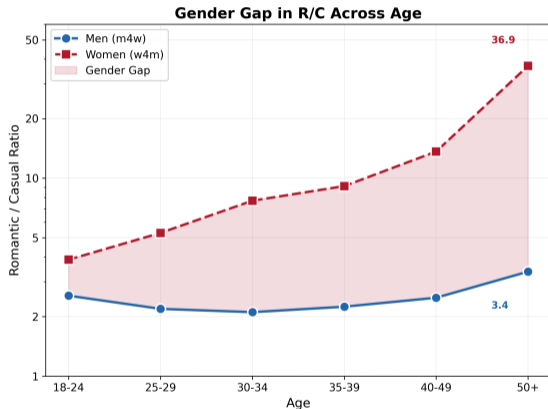
## The 3.6-point gap creates market imbalance

- ▷ Women are *much* more likely to seek romantic partners
- ▷ Men are *relatively* more likely to seek casual encounters
- ▷ This creates a “shortage” of romantic men

**Implication:** Women seeking romance face harder search

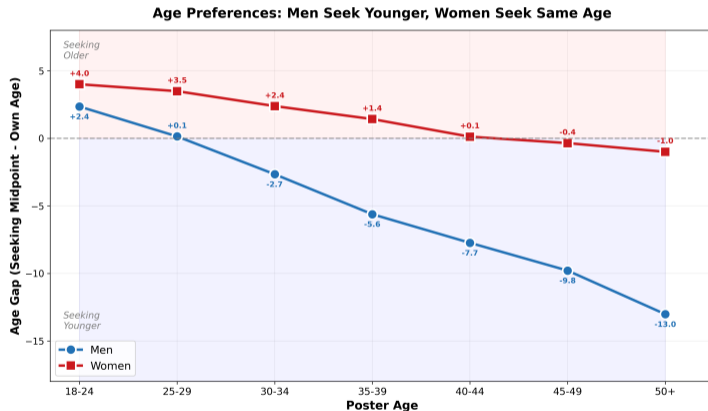
But this is just averages. How does it vary by age?

# The Same Story as a Picture



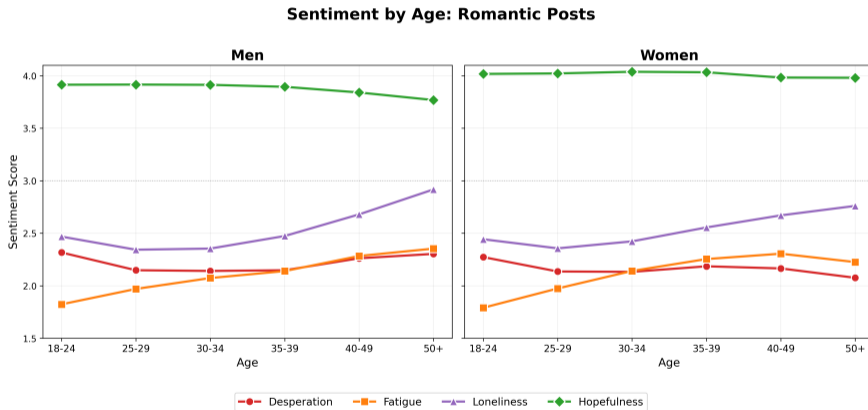
This is a **line graph**—great for showing means over age. The shaded gap makes the divergence visceral.

# Age Preferences: A Diverging Pattern



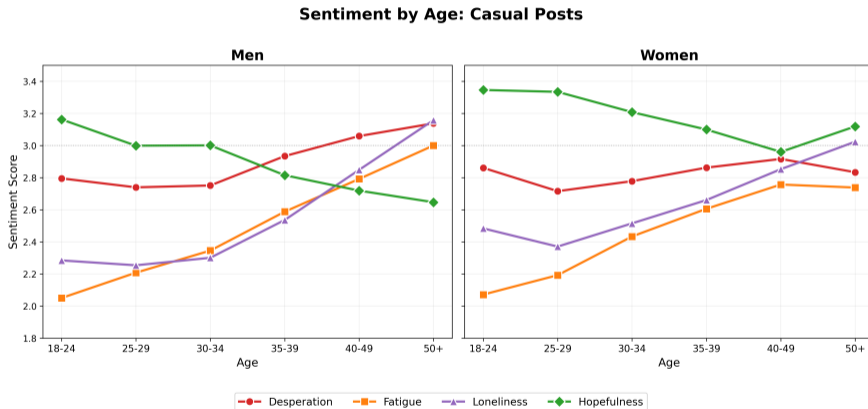
Men seek younger; women stay near their own age. A table couldn't show this pattern.

# Sentiment by Age: Romantic Posts



Facets keep 8 lines readable. Each panel = one sentiment dimension.

# Sentiment by Age: Casual Posts



Completely different pattern. The *category* changes the story.

# The Market Facing Older R-Type Women

Consider a **40-year-old woman** seeking a romantic relationship:

- 1. The pool of R-type men her age is small**
  - ▷ Men's  $R/C \approx 2.6$  at age 40–49; Women's  $R/C \approx 15$
- 2. R-type men her age prefer younger women**
  - ▷ Men 40–49 prefer women  $\approx 9$  years younger on average
- 3. C-type men targeting her are increasingly desperate**
  - ▷ Casual men's desperation rises sharply with age

**Her options:** Keep searching (delays childbearing) · Lower standards (unstable) · Give up

This is data synthesis—turning statistics into human stories

# What Made These Pictures Work?

Six principles:

1. **One idea per panel:** Don't overload
2. **Facets for complexity:** 8 lines  $\rightarrow$  8 small panels
3. **Color with purpose:** Men vs. women, not decoration
4. **Shading for gaps:** Makes divergence *visceral*
5. **White space:** Let the data breathe
6. **Direct labels:** No legend decoding required

Your figures should be readable without the text around them



# Behind the Scenes: How We Built This Dataset

# We used “robots” to scrape the Internet Archive

## The problem:

- ▷ Craigslist Personals shut down in 2018
- ▷ But the Wayback Machine preserved them

## The solution: Web scraping

- ▷ Automated collection of web pages
- ▷ Write code that visits pages and extracts data
- ▷ Runs while you sleep

166,000 posts collected over a few days

# Claude Code wrote the scraping code for us

## Modern AI tools can write code:

- ▷ I described what I wanted in plain English
- ▷ Claude Code wrote the Python scraping code
- ▷ I reviewed it, tested it, ran it

## This is how research is done now:

- ▷ AI as research assistant
- ▷ You provide the *ideas*, AI provides the *implementation*
- ▷ But you must understand enough to verify

# But how did we classify 166,000 posts?

## The challenge:

- ▷ Reading them all would take years
- ▷ Hiring humans is expensive (\$0.10–\$1 per post = \$16,000–\$166,000)
- ▷ We needed automation

## The solution: Large Language Models (LLMs)

- ▷ GPT-4o-mini can classify text
- ▷ Cost: \$10 for all 166,000 posts
- ▷ Time: A few hours

# Text is data—if you can convert it to numbers

## The fundamental insight:

- ▷ You can't take an average of *words*
- ▷ But you CAN **count** words (word clouds = frequencies)
- ▷ And you CAN **classify** text into categories

## Categories become numbers:

- ▷ Romantic = 1, Casual = 0
- ▷ Now you can compute: mean, proportion, trend

Classification is the bridge between text and statistics



# A Brief History of Text Analysis

*Who wrote the disputed Federalist Papers?*

# The Federalist Papers: A 175-Year Mystery

## The setup:

- ▷ 85 essays published 1787–1788 under pseudonym “Publius”
- ▷ Goal: Persuade New York to ratify the Constitution
- ▷ Authors: Alexander Hamilton, James Madison, John Jay

## The problem:

- ▷ 51 known Hamilton, 14 known Madison, 5 known Jay
- ▷ **12 disputed:** Both Hamilton and Madison claimed them
- ▷ Hamilton died in 1804 duel; Madison lived until 1836

# Mosteller and Wallace (1963) solved it with statistics

## The researchers:

- ▷ Frederick Mosteller (Harvard Statistics)
- ▷ David Wallace (University of Chicago)

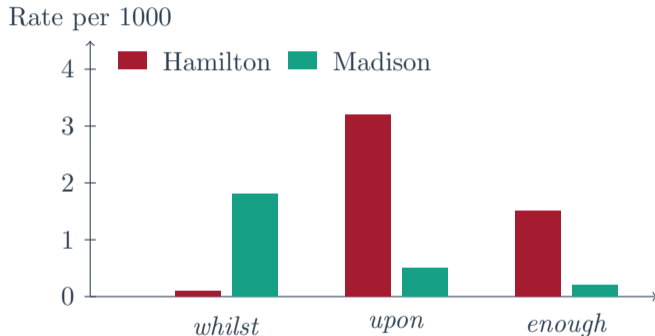
## Key insight: Function words reveal authorship

- ▷ Not *what* you write, but *how* you write
- ▷ Words like: *whilst, upon, enough, by, to*
- ▷ These are unconscious stylistic fingerprints

## Why function words?

- ▷ Content words vary by topic
- ▷ Function words are stable across topics

## Hamilton and Madison had different word patterns



**Hamilton never used “whilst”; Madison did.** Unconscious fingerprints.

# All 12 disputed papers were written by Madison

## Bayesian analysis:

- ▷ Odds ratios  $> 1000:1$  for Madison on disputed papers
- ▷ Federalist 51 (“Ambition must counteract ambition”): **Madison**
- ▷ This is now the historical consensus

Statistical analysis resolved a question historians couldn't

## Another Famous Case: Primary Colors (1996)

### The mystery:

- ▷ Anonymous novel about Clinton-like presidential candidate
- ▷ Bestseller, made into a movie
- ▷ Who wrote it?

### Stylometric analysis:

- ▷ Vassar professor analyzed function words, sentence patterns
- ▷ Identified Joe Klein (Newsweek columnist)
- ▷ Klein initially denied, then admitted

Same method as Federalist: signatures in style, not content

# The key insight: signatures in style, not content

## Why this works:

- ▷ People leave unconscious fingerprints in how they write
- ▷ Function words are stable across topics
- ▷ Content words change; style doesn't

## This is the foundation of:

- ▷ Authorship attribution
- ▷ Plagiarism detection
- ▷ Forensic linguistics

# From Authorship to Attitude

Can we measure what politicians *believe*  
from how they *speak*?



# Measuring 140 Years of Immigration Rhetoric

# Card et al. analyzed 305,000 political speeches (1880–2020)

## Data sources:

Source	Records	Time Period
Congressional speeches	290,800	1880–2020
Presidential communications	14,195	1880–2021
<b>Total</b>	<b>304,995</b>	<b>140 years</b>

**Span:** Chinese Exclusion Act (1882) to present

Card, Boustan, Abramitzky et al. (PNAS 2022)

# The Research Questions

1. Has immigration rhetoric changed over time?
2. Do Republicans and Democrats differ?
3. Has rhetoric about specific nationalities changed?
4. When did polarization emerge?

**Method:** Classify each speech as Pro, Anti, or Neutral

# Human annotators created the training data

## Annotation setup:

- ▷ 5 Princeton annotators (grad students, undergrads)
- ▷ 7,626 speech segments labeled
- ▷ Cost: ~\$10,000+ just for this labeling

## Labels:

- ▷ Pro-immigration
- ▷ Anti-immigration
- ▷ Neutral

This is expensive, slow, but creates the “ground truth”

# The Classification Scale

*Each speech gets placed on this line*



The aggregate measure:

$$\text{Average tone} = \% \text{ Pro} - \% \text{ Anti}$$

Range: -100 (all anti) to +100 (all pro)

# RoBERTa learned to classify from human examples

**RoBERTa:** A fine-tuned neural network

**Process:**

1. Trained on 7,626 labeled examples
2. Applied to remaining 305,000 speeches
3. Output: Pro/Anti/Neutral probabilities for each

**Performance:**

- ▷ ~65% accuracy on tone classification
- ▷ But humans only agreed at  $\alpha = 0.48$ !
- ▷ So 65% is actually quite good



**What Did They Find?**

# Overall sentiment is more positive today

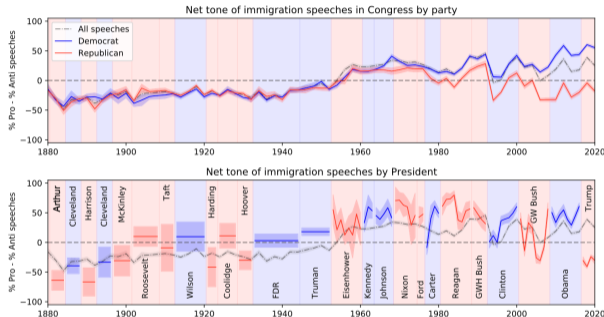
**Surprising!** Given current discourse, you might expect the opposite.

**Three eras:**

- 1. 1880–1940:** Consistently negative (quota era)
- 2. 1940–1965:** Shift toward positive (WWII to Immigration Act)
- 3. 1965–present:** Net positive on average

But this masks important variation by party...

# Immigration rhetoric: Overall trend (1880–2020)



**Fig. 1.** Evolution of attitudes toward immigration expressed in congressional speeches and presidential communications. Average tone is computed as the percentage of proimmigration speeches minus the percentage of antiimmigration speeches, where proimmigration means valuing immigrants and favoring less restricted immigration and vice versa. *Top* and *bottom* show the overall tone using all congressional speeches about immigration (black dashed line, with bands showing plus or minus two SDs based on the estimated proportions and number of speeches). *Top* also shows separate plots for speeches by Democrats and Republicans in Congress. (Due to limitations of the data, about 15% of speeches do not have a named speaker or party affiliation.) *Bottom* shows the corresponding estimates for each president, showing the overall average for a president's tenure when there are insufficient data to show annual variation. Note that most modern presidents have been more favorable toward immigration than the average member of Congress. By contrast, Donald Trump appears to be the most antiimmigration president in nearly a century. Similarly, congressional Republicans over the past decade have framed immigration approximately as negatively as the average member of Congress did a century earlier.

Average tone = % Pro – % Anti. Source: Card et al. (PNAS 2022)

# But the parties have polarized dramatically

**Through the 1970s:** Both parties roughly similar

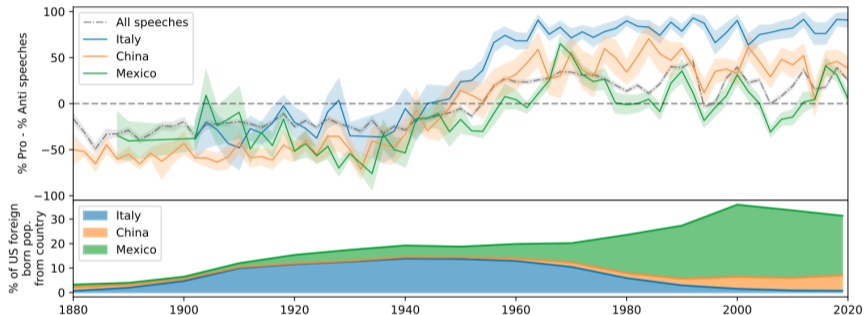
**Then divergence:**

- ▷ Democrats: increasingly positive
- ▷ Republicans: increasingly negative

**Today:** Historic extremes

- ▷ Democrats: most pro-immigration ever
- ▷ Republicans: as negative as the 1920s quota era

# Partisan polarization is at historic levels



**Fig. 2.** Average tone of immigration speeches when considering only those speeches that mention the country or nationality for each of the three most frequently mentioned nationalities (*Top*) and the percent of the US foreign-born population from each of these countries over time (*Bottom*). Despite the midcentury increase in proimmigration attitudes applying to all groups, a gap in tone by group persists to the present day, with Mexican immigrants being consistently framed more negatively than others and Italian immigrants being framed especially positively. These trends are mirrored in broader regional patterns for Europe, Asia, and Latin American and the Caribbean (*SI Appendix*).

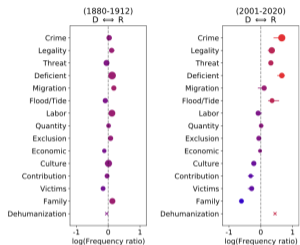
Note the crossing point and widening gap. Source: Card et al. (PNAS 2022)

# Trump broke from historical presidential patterns

## A historical first:

- ▷ First modern president more anti-immigration than own party
- ▷ Most anti-immigration president in 140 years
- ▷ Broke pattern where presidents were moderating voices

# Presidential rhetoric: Trump vs. history



**Fig. 3.** Relative usage frequency for each of 14 frames by Republicans compared to Democrats, both for the late 19th/early 20th century (*Left*) and the past two decades (*Right*). Farther to the left on each plot represents more frequent usage by Democrats and vice versa (plotted as log frequency ratio). Circle size represents the overall prominence of the frame in speeches about immigration, relative to all speeches. To ensure the robustness of these findings, we leave out each word in turn from each frame and show the full range of possible values obtained using horizontal lines (not visible when the full range is contained within the circle). “Dehumanization” is an aggregation of metaphorical categories (see *Measuring Dehumanization*). Compared to the absence of polarization a century ago, certain frames today are disproportionately used by Republicans (“crime,” “legality,” “threats,” “deficiency,” and “flood/tide”) and Democrats (“family,” “victims,” “contributions,” and “culture”). Republicans also show significantly higher use of implicit dehumanizing metaphors like “animals” and “cargo.”

Presidential tone relative to Congress. Source: Card et al. (PNAS 2022)

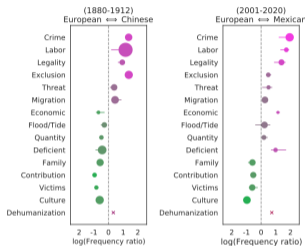
# Mexican immigrants framed most negatively

## Rhetoric varies by nationality:

- ▷ **Chinese:** Negative during exclusion era (1882–1943)
- ▷ **Italian:** Now positive (was negative in early 1900s)
- ▷ **Mexican:** Persistently most negative

**Key insight:** Italians “became white”—Mexicans haven’t (yet?)

# Rhetoric by immigrant nationality



**Fig. 4.** Relative usage frequency for each of 14 frames in speeches mentioning Chinese vs. European immigrants in the late 19th/early 20th century (Left) and those mentioning Mexican vs. European immigrants in the 21st century (Right). Farther to the left on each plot represents greater usage in speeches mentioning European groups. Circle size represents the overall frequency of the frame in the relevant speeches relative to all speeches. Horizontal lines show the minimum and maximum values of the log ratio obtained when leaving out each term in the corresponding lexicon. In turn, “Dehumanization” is an aggregation of the six metaphorical categories. There is a strong correlation between how Mexican immigrants are framed today and how Chinese immigrants were framed a century earlier, relative to European immigrants of the corresponding time period, in terms of both the explicit frames emphasized and a significantly higher usage of dehumanizing metaphors for mentions of the non-European groups.

Tone by immigrant nationality mentioned. Source: Card et al. (PNAS 2022)

## Is There a Cheaper Way?

\$10,000 and weeks of work...

or \$11 and an afternoon?



# Replicating with Modern LLMs

# I replicated this study for \$11 in 4.5 hours

## The approach:

- ▷ Used GPT-4o-mini via OpenAI Batch API
- ▷ Zero-shot classification (no training!)
- ▷ Same task: classify as Pro, Anti, or Neutral
- ▷ Claude Code wrote the pipeline

**The question:** Do we get the same results?

# The Process



**Batch API:** Submit all at once, get results hours later (50% off)

## The Cost Comparison

Approach	Cost	Time
Human annotation + RoBERTa	\$10,000+	Weeks
GPT-4o-mini Batch API	\$11	4.5 hours

**Cost reduction:** 99.9%

**Time reduction:** 99%+

But do we get the same answers?

# How well did they agree?

**Overall agreement:** 69%

**Context:**

- ▷ Human annotators only agreed at  $\alpha = 0.48$
- ▷ So 69% is actually quite good
- ▷ The LLM performs as well as a typical human annotator

But *where* do they disagree? This tells us something important.

# The Transition Matrix: Just the Diagonal

Diagonal = agreement rates

RoBERTa	LLM Classification		
	PRO	NEUTRAL	ANTI
PRO	<b>63%</b>	—	—
NEUTRAL	—	<b>85%</b>	—
ANTI	—	—	<b>51%</b>

**Reading:** When RoBERTa said PRO, LLM agreed 63% of the time

**Notice:** NEUTRAL has highest agreement (85%)

## The Transition Matrix: PRO Row

What happened when RoBERTa said PRO?

RoBERTa	LLM Classification		
	PRO	NEUTRAL	ANTI
PRO	<b>63%</b>	<b>33%</b>	<b>4%</b>
NEUTRAL	—	85%	—
ANTI	—	—	51%

**Key insight:** When LLM disagreed with PRO, it usually said NEUTRAL

Polarity flips (PRO  $\rightarrow$  ANTI) are rare: only 4%

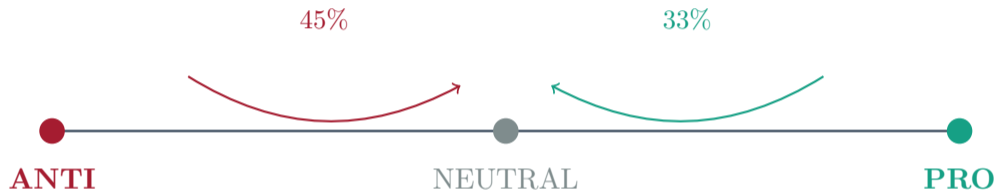
# The Transition Matrix: Full Picture

RoBERTa	LLM Classification		
	PRO	NEUTRAL	ANTI
PRO	63%	33%	4%
NEUTRAL	7%	85%	8%
ANTI	4%	45%	51%

**Key insight:** NEUTRAL absorbs from both tails

- ▷ PRO → NEUTRAL: 33%
- ▷ ANTI → NEUTRAL: 45%
- ▷ Polarity flips (PRO ↔ ANTI): only ~4%

## Return to the Classification Line



The LLM pushes uncertain cases toward the middle

It's more conservative—when in doubt, say NEUTRAL

# The Key Question

How is it possible to change so many classifications...  
...and yet the average trends stay the same?

Think about this. It's counterintuitive.

# The Comparison: Original vs. LLM

## Original (RoBERTa)

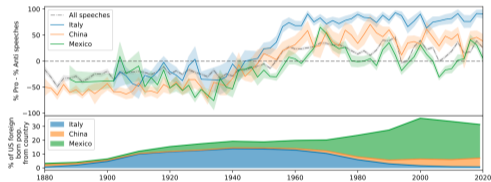
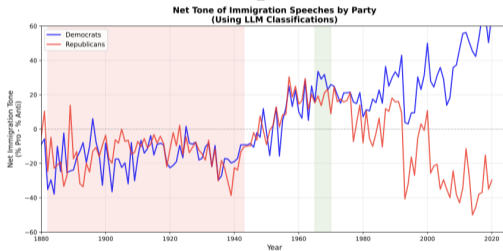


Fig. 2. Average tone of immigration speeches when considering only those speeches that mention the country or nationality for each of the three most frequently mentioned nationalities (Top) and the percent of the US foreign-born population from each of these countries over time (Bottom). Despite the midcentury increase in proimmigration attitudes applying to all groups, a gap in tone by group persists to the present day, with Mexican immigrants being consistently framed more negatively than others and Italian immigrants being framed especially positively. These trends are mirrored in broader regional patterns for Europe, Asia, and Latin American and the Caribbean (SI Appendix).

## LLM Replication



Same story! Partisan polarization, same timing, same direction

# Why? Symmetric Noise Removal

**The answer:**

When you remove equal amounts from both tails...  
...the mean doesn't change!

**The LLM is removing “noise”—uncertain classifications**

- ▷ PRO → NEUTRAL: removes from positive tail
- ▷ ANTI → NEUTRAL: removes from negative tail
- ▷ Roughly symmetric → mean preserved

The signal remains; the noise is removed

## A Simple Example

**Original classifications:**

$-2, -1, 0, +1, +2 \rightarrow \text{Mean} = 0$

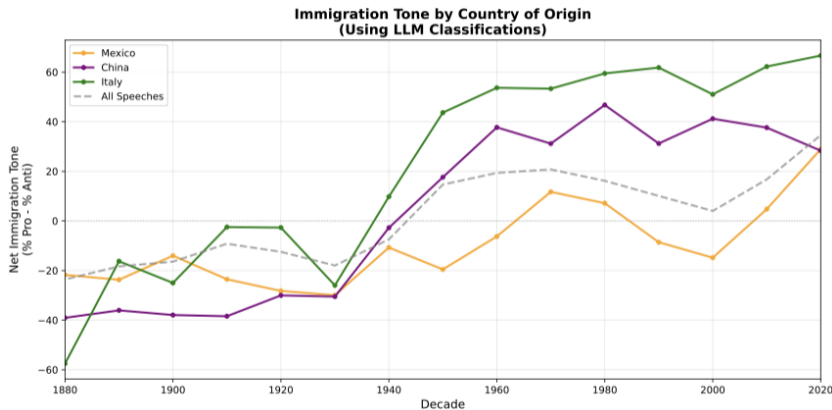
**After symmetric noise removal:**

$-1, 0, +1 \rightarrow \text{Mean} = 0$

**Same average, fewer observations**

The extremes were uncertain; removing them doesn't change the center

# The Ordering Also Holds



Italy > China > Mexico—same as original



# Making Line Graphs in R

# Lines are great for means over time

## When to use what:

- ▷ **Histograms:** Distribution of one variable
- ▷ **Bar charts:** Comparing categories
- ▷ **Line graphs:** Trends over time, age, or continuous variable

## Lines show:

- ▷ Change and direction
- ▷ Comparisons between groups
- ▷ Patterns over continuous scales

# R Code for Line Graphs

```
library(tidyverse)

# Compute means by year and party
trends <- speeches %>%
  group_by(year, party) %>%
  summarize(avg_tone = mean(tone))

# Plot line graph
ggplot(trends, aes(x = year, y = avg_tone, color = party)) +
  geom_line(size = 1.2) +
  labs(title = "Immigration Rhetoric Over Time",
       x = "Year", y = "Average Tone (% Pro - % Anti)") +
  theme_minimal()
```

# Key ggplot Elements for Lines

- ▷ `geom_line()` for the lines
- ▷ `color = variable` for grouping (party, gender, etc.)
- ▷ `facet_wrap()` for panels (like sentiment plots)
- ▷ `geom_ribbon()` for shaded areas (like the gender gap)

You'll practice this in section and on the problem set



# What We Learned

# Key Takeaways

## 1. Means and lines tell powerful stories

The Craigslist data reveals market imbalances

## 2. Text becomes data through classification

Categories  $\rightarrow$  numbers  $\rightarrow$  statistics

## 3. Authors leave stylistic fingerprints

Function words, not content words

## 4. Modern LLMs can replicate expensive methods cheaply

\$11 vs. \$10,000

## 5. Robust findings survive different methods

If the signal is real, symmetric noise removal preserves it

# For Your Final Project

**I'm leaving this for you to consider:**

- ▷ Maybe you want to analyze different speeches
- ▷ Maybe you want to classify something else entirely
- ▷ The tools are accessible now

**Come see me:** I can show you the code

Office hours: Wednesdays 2–4pm

# Resources

**Paper:** Card et al. (PNAS 2022)

Will be on the exam—read it carefully

**Book:** Leah Boustan, *Streets of Gold*

Available at Harvard Coop

**My replication:** Substack series (link on course website)

Shows the full pipeline



Text is data. LLMs are cheap.  
Go measure something that matters.