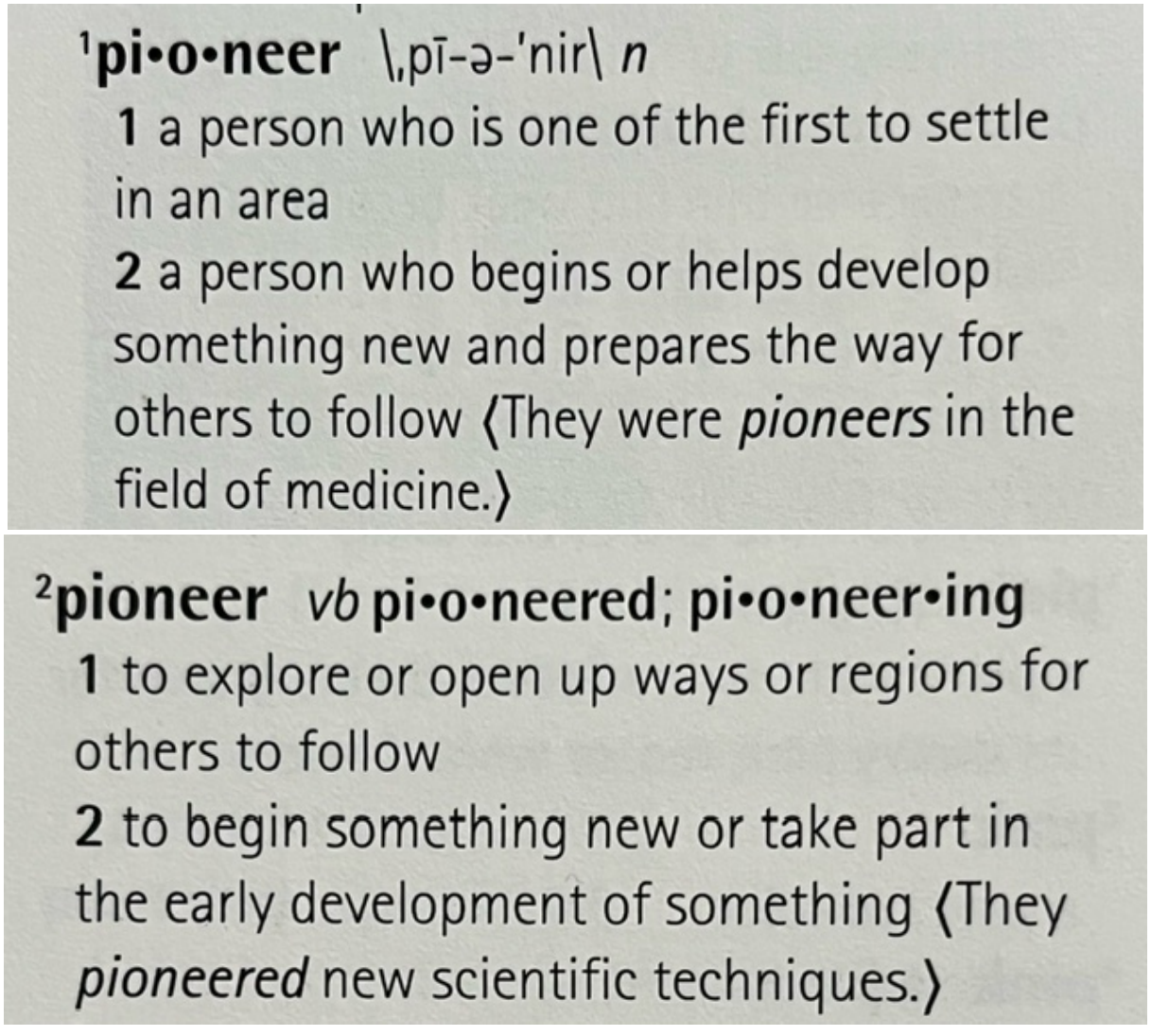
# Francisco Cantú as a Pioneer in the Application of Machine Learning (ML) to Political Science

Sebastian M. Saiegh

Empirical Implications of Theoretical Models (EITM)

2025 Francisco Cantú Memorial

# Pioneer



**[1]pi•o•neer** \ˌpī-ə-ˈnir\ *n*

**1** a person who is one of the first to settle in an area

**2** a person who begins or helps develop something new and prepares the way for others to follow ⟨They were *pioneers* in the field of medicine.⟩

**[2]pioneer** *vb* **pi•o•neered; pi•o•neer•ing**

**1** to explore or open up ways or regions for others to follow

**2** to begin something new or take part in the early development of something ⟨They *pioneered* new scientific techniques.⟩

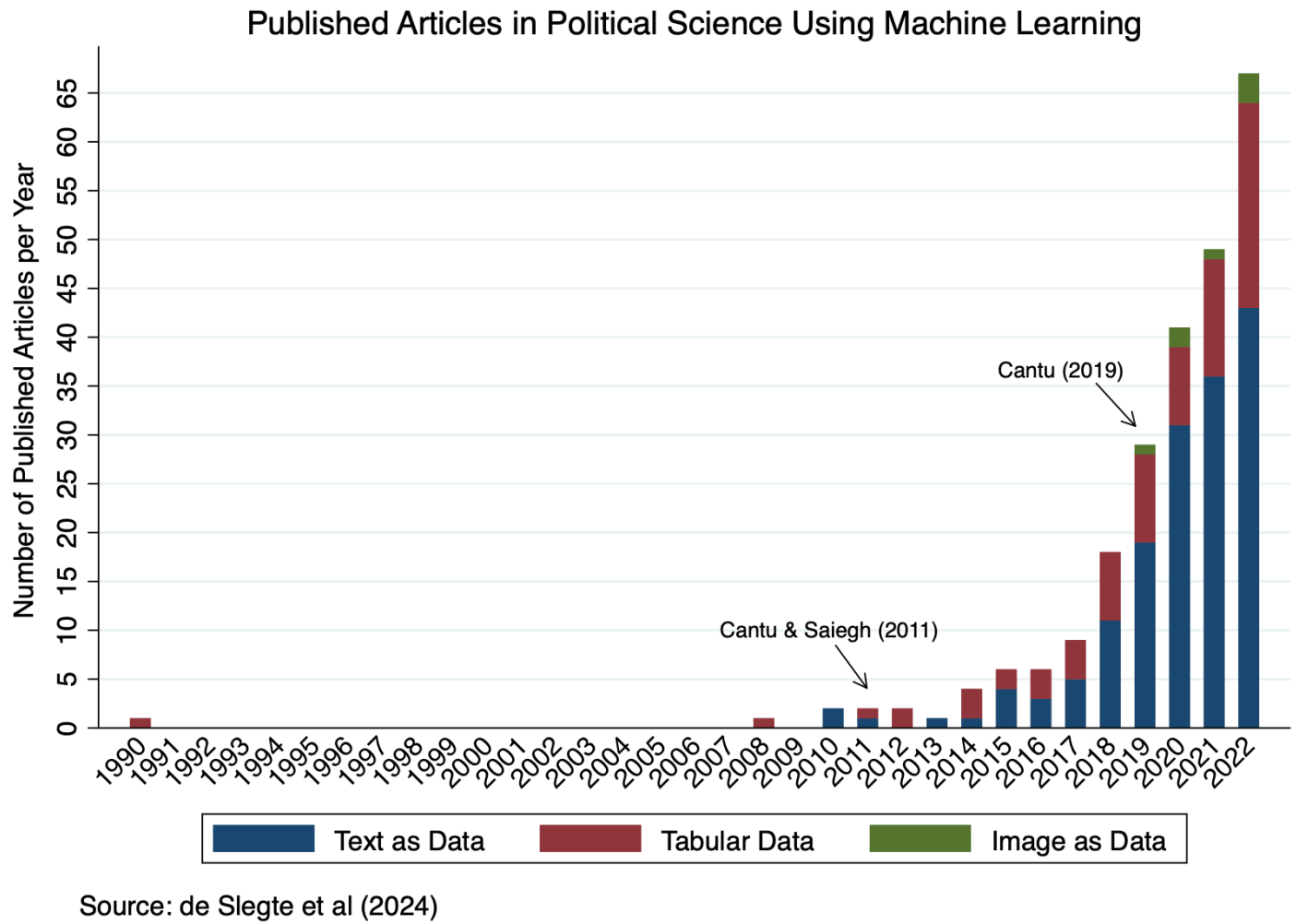
# Pioneer (cont.)

While the adoption of machine learning in politcal science has surged in recent years, Francisco Cantú's *Fraudulent Democracy?* stands as a foundational milestone.

- Published in 2011, the groundbreaking article introduced several innovations:
  - Use synthetic data to train a fraud detection prototype
  - Application of supervised learning framework
  - Detect fraudulent practices using distribution of the digits in reported vote counts

Approach inspired later advancements, including his 2019 APSR article using convolutional neural networks to analyze Mexico's 1988 elections.

# Pioneer (cont.)



# Pioneer (cont.)

1. Schrodt (1990). Predicting interstate conflict outcomes using a bootstrapped ID3 algorithm, *Political Analysis*
2. Goldsmith, Chalup & Quinlan (2008). Regime Type and International Conflict, *Journal of Peace Research*
3. Quinn et al (2010). How to Analyze Political Attention with Minimal Assumptions and Costs, *American Journal of Political Science*
4. Hancock et al (2010). Social language processing: A framework for analyzing the communication of terrorists and authoritarian regimes, *Behavioral Sciences of Terrorism and Political Aggression*
5. Cantú & Saiegh (2011). Fraudulent Democracy?, *Political Analysis*

## Election Studies

1. Cantú & Saiegh (2011). Fraudulent Democracy?, *Political Analysis*
2. Clark, Morris & Lomax (2018). Estimating the outcome of UKs referendum on EU membership using e-petition data and machine learning algorithms, *Journal of Information Technology & Politics*
3. Hemsley et al (2018). Tweeting to the Target: Candidates' Use of Strategic Messages and Mentions on Twitter, *Journal of Information Technology & Politics*
4. Cantú (2019). The fingerprints of fraud: Evidence from Mexico's 1988 presidential election, *American Political Science Review*
5. Muchlinski (2020). We need to go deeper: measuring electoral violence using convolutional neural networks and social media, *Political Science Research and Methods*

# Fraudulent Democracy? (2011)

## Fraudulent Democracy? An Analysis of Argentina's *Infamous Decade* Using Supervised Machine Learning

**Francisco Cantú and Sebastián M. Saiegh**

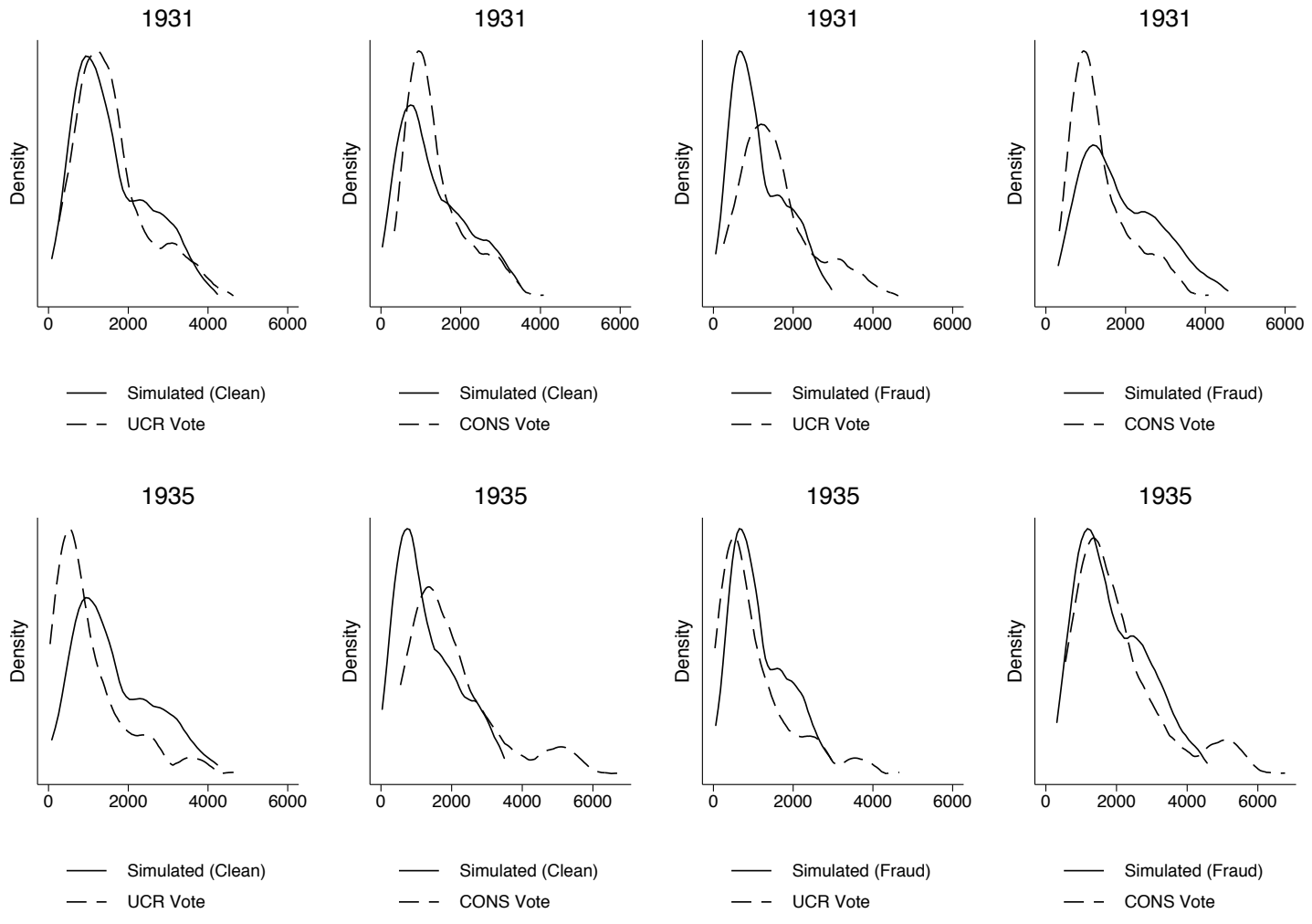
*Department of Political Science, University of California, San Diego, CA 92093*
*e-mail: ssaiegh@ucsd.edu (corresponding author)*

In this paper, we introduce an innovative method to diagnose electoral fraud using vote counts. Specifically, we use synthetic data to develop and train a fraud detection prototype. We employ a naive Bayes classifier as our learning algorithm and rely on digital analysis to identify the features that are most informative about class distinctions. To evaluate the detection capability of the classifier, we use authentic data drawn from a novel data set of district-level vote counts in the province of Buenos Aires (Argentina) between 1931 and 1941, a period with a checkered history of fraud. Our results corroborate the validity of our approach: The elections considered to be irregular (legitimate) by most historical accounts are unambiguously classified as fraudulent (clean) by the learner. More generally, our findings demonstrate the feasibility of generating and using synthetic data for training and testing an electoral fraud detection system.

# Synthetic Data

- We generate vote counts for 100 simulated districts, each containing two competing parties, $i \in \{A, B\}$.
- Fraud is simulated in the following way:
  - In each district, we take away a fixed proportion $\gamma$ of party $A$'s votes and give $\delta(\gamma V_{Aj})$ votes to party $B$ (with $\gamma > 0$, $\delta > 0$).
- We calibrate our simulations using the 1931 and 1935 elections: $\alpha_A = 400$, $\alpha_B = 320$, $\gamma = 0.3$, and $\delta = 1.2$ provide the best fit between the simulated and real data.
- We generate 10,000 electoral contests. Each one is treated as a Bernoulli trial with probability of success/failure, $p = .5$.
- For each contest, we record:
  1. The mean of the first digit of party $i$'s votes in every district.
  2. The frequency of the number 1 as the first significant digit of party $i$'s votes in every district.

# Calibration



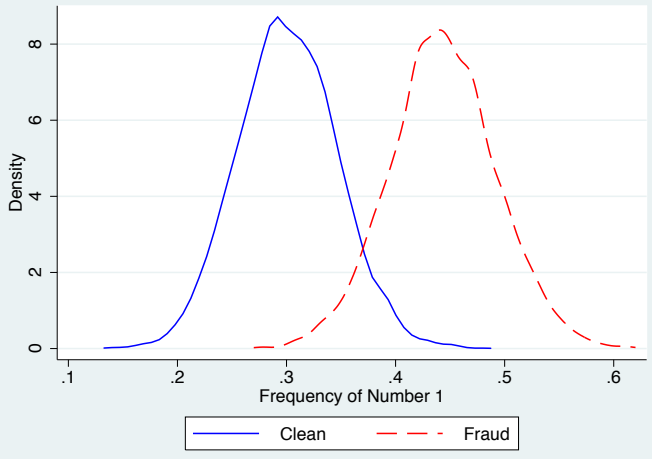
# Classification using Naive Bayes

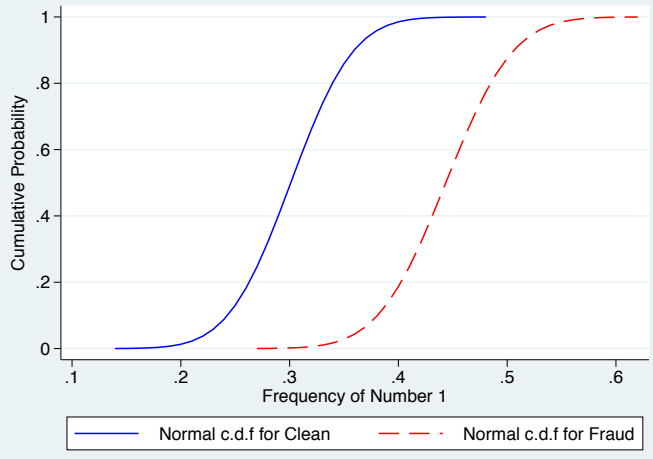
- The classification problem consists of finding the class with maximum probability given a set of observed attribute values.
- Following Bayes' theorem, the posterior probability of class $y$ can be written as:

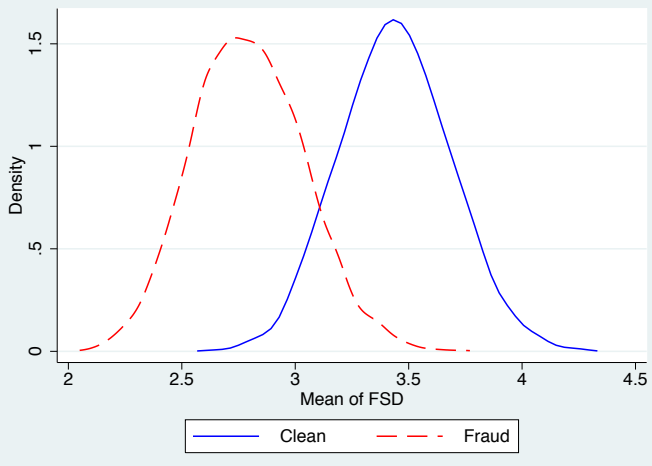
$$p(y|\mathbf{x}) = \frac{p(y)}{p(\mathbf{x})} \prod_{i=1}^{m} p(x_i|y),$$

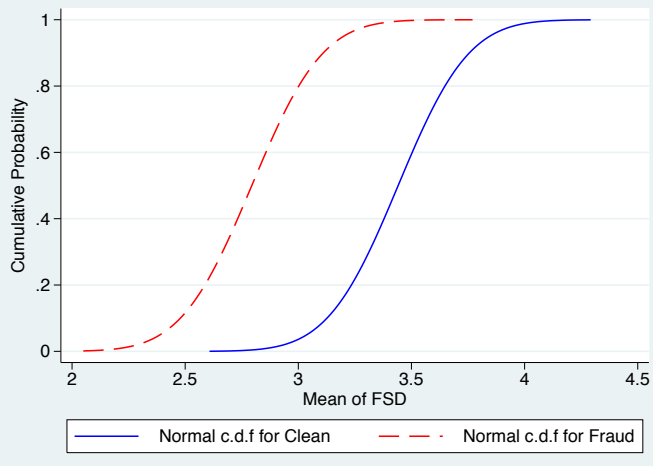
- Independence rarely holds in real-world applications; yet NB is accurate and efficient even if this assumption is violated.
- We use the class-conditional densities from our synthetic data to classify elections.

# Classification (cont.)



# Classification (cont.)

- In the 1931 elections, the frequency of the number 1 as the first significant digit (FSD) of the beneficiary of fraudulent practices is 0.4, and the mean of the FSD is 3.77
- The probability that the 1931 election was clean is thus:

$$\frac{p(c)p(x_1 \leq 0.4|c)p(x_2 \leq 3.77|c)}{p(c)p(x_1 \leq 0.4|c)p(x_2 \leq 3.77|c) + p(c')p(x_1 \leq 0.4|c')p(x_2 \leq 3.77|c')},$$

  where $x_1$ denotes the frequency of the number 1, and $x_2$ denotes the mean of the FSD.
- Given a prior assignment of probabilities $p(c) = p(c') = \frac{1}{2}$,

$$p(c|\mathbf{x}) = \frac{(.5)(1)(.9108)}{(.5)(1)(.9108) + (.5)(.1954)(1)} \approx 0.823$$

# Classification (cont.)

**Table 4**
**Classification of Buenos Aires' Elections (1931-1941)**

| Election | $p(clean) = p(fraud)$ | $p(clean\|\mathbf{x})$ | $p(fraud\|\mathbf{x})$ | $log\frac{p(y=1\|\mathbf{x})}{p(y=0\|\mathbf{x})}$ | Classification |
|---|---|---|---|---|---|
| Validation Set (Seed Data) | | | | | |
| **1931** | 0.5 | 0.823 | 0.176 | -1.539 | Clean |
| **1935** | 0.5 | 0.054 | 0.945 | 2.845 | Fraudulent |
| Test Set | | | | | |
| **1940** | 0.5 | 0.756 | 0.243 | -1.135 | Clean |
| **1941** | 0.5 | 0.080 | 0.919 | 2.441 | Fraudulent |

Notes: This table reports the classification of the elections in our validation set (top panel) and in our test set (bottom panel) obtained using the NB learning algorithm.

# Legacy of Francisco Cantú

Francisco's role as a pioneer in political science research should be acknowledged. His work:

- Shaped trajectory of field with new methodologies
- Inspired and motivated other researchers
- Influenced future studies and methodologies

He will be remembered and cited long after this initial work, leaving a lasting legacy in the scientific literature.