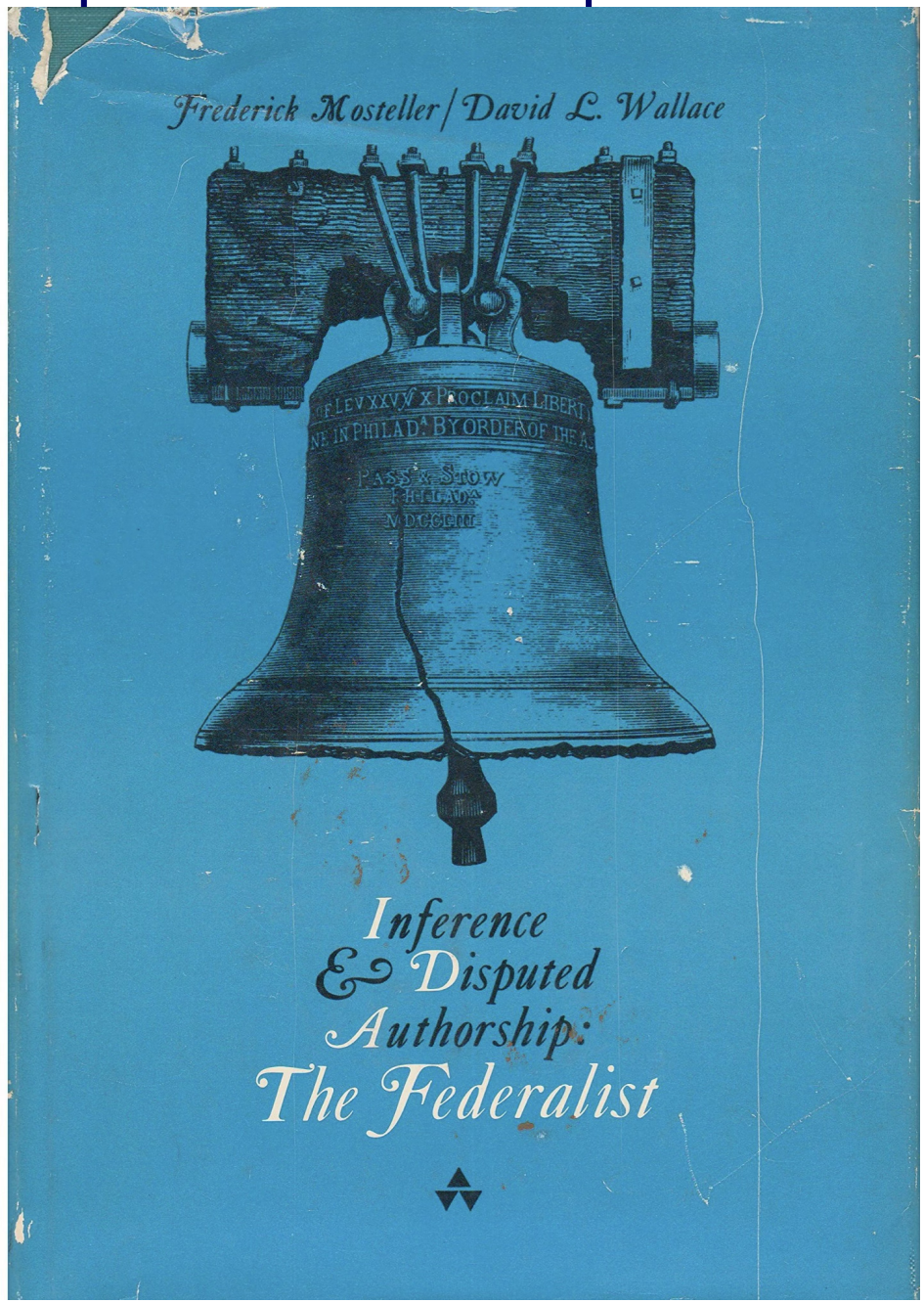
Text as Data

Meeting 3: Authorship Attribution

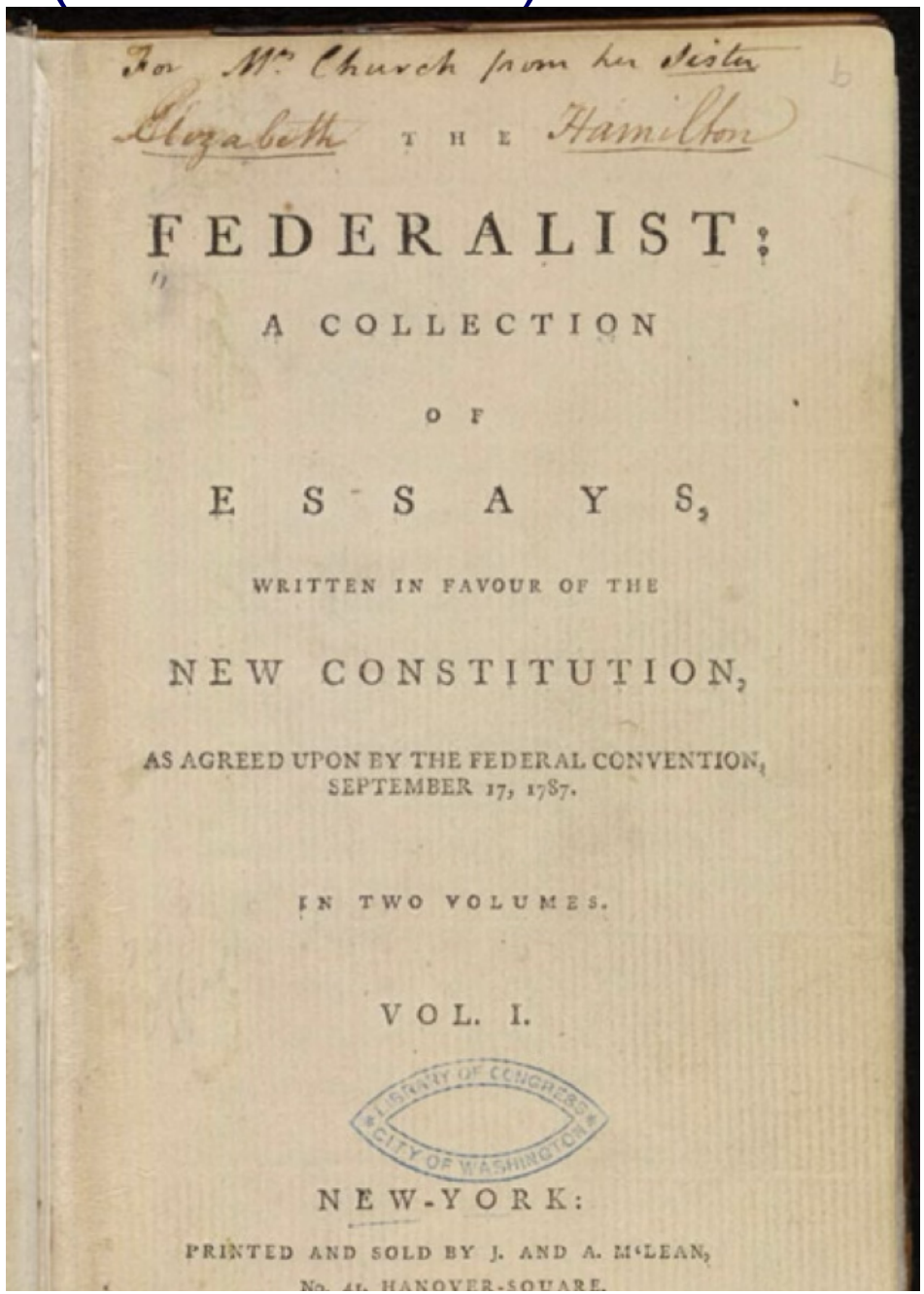
Petro Tolochko

Inference and Disputed Authorship: The Federalist



One of the first (if not the first) text-as-data study

One of the first (if not the first) text-as-data study



For Mrs. Church from her sister
Elizabeth THE Hamilton

FEDERALIST:

A COLLECTION

OF

ESSAYS,

WRITTEN IN FAVOUR OF THE

NEW CONSTITUTION,

AS AGREED UPON BY THE FEDERAL CONVENTION,
SEPTEMBER 17, 1787.

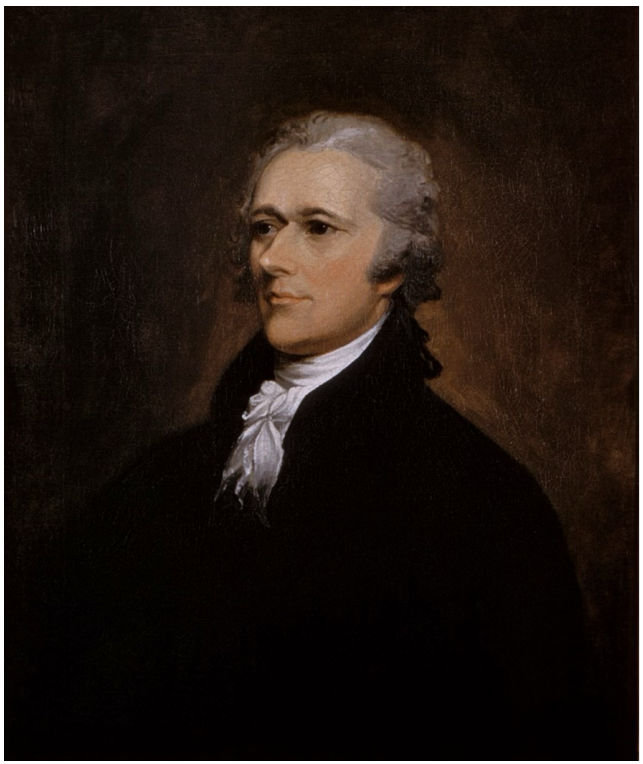
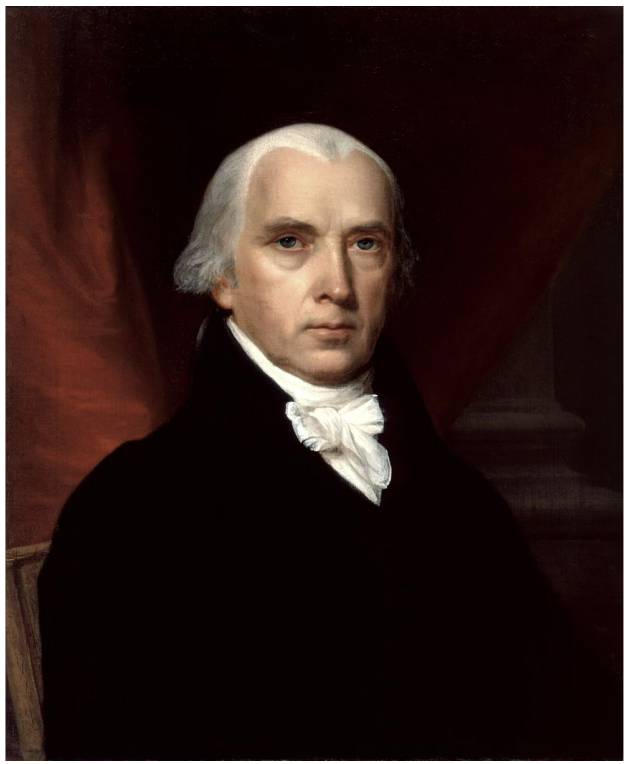
IN TWO VOLUMES.

VOL. I.

NEW-YORK:

PRINTED AND SOLD BY J. AND A. M'LEAN,
No. 41, HANOVER-SQUARE.

One of the first (if not the first) text-as-data study


The Federalist Papers

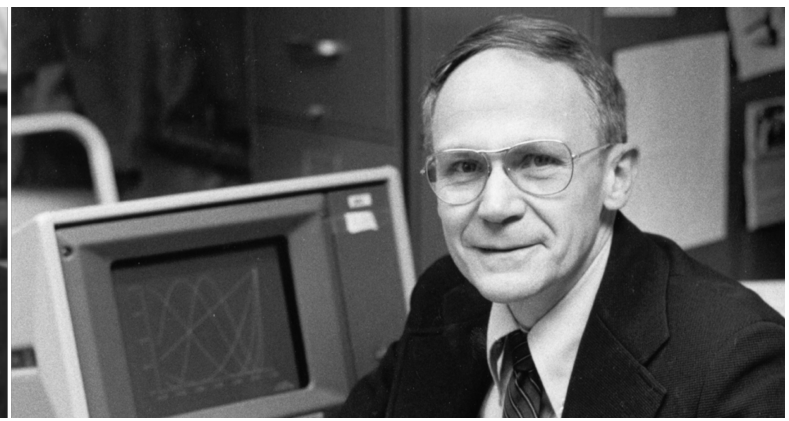
- 85 essays published in 1787–88 advocating ratification of the U.S. Constitution
- Written under the pseudonym **Publius**
- Main authors: **Alexander Hamilton**, **James Madison**, **John Jay**
- Historical mystery: who wrote which of the disputed essays?

The Authorship Puzzle

- 73 essays: authorship fairly certain
- 12 essays: disputed between Hamilton and Madison
- Historians debated authorship for 150+ years
- Provided an ideal test case for applying quantitative methods

Mosteller & Wallace (1964)

- First large-scale statistical study of text
- Used Bayesian methods and early computing power
- Demonstrated that statistical evidence can resolve historical debates

Text Analysis

Text Analysis

Dimension Reduction

- Remove the stopwords

Dimension Reduction

- Remove the stopwords
- Still too many words!

Dimension Reduction

- Remove the stopwords
- Still too many words!
- Remove all the words **but** the stopwords

Dimension Reduction

- Remove the stopwords
- Still too many words!
- Remove all the words **but** the stopwords
- Maybe there is information in them!

Simplified example from Grimmer et al., 2022

- Focus on:
  - “Man”
  - “By”
  - “Upon”

Simplified example from Grimmer et al., 2022

- Focus on:
  - "Man"
  - "By"
  - "Upon"
- The rates with which the authors use these words may indicate authorship

Word Rates

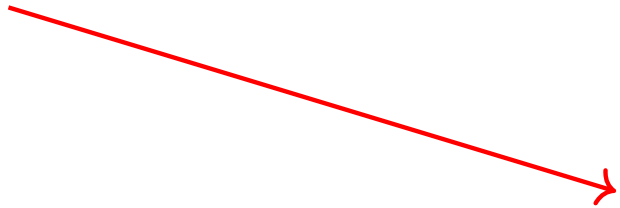
	man	by	upon
Hamilton	102	859	374
Madison	17	474	7
Jay	0	82	1

Word Proportions

	man	by	upon
Hamilton	.076	.643	.28
Madison	.034	.952	.014
Jay	0	.988	.012

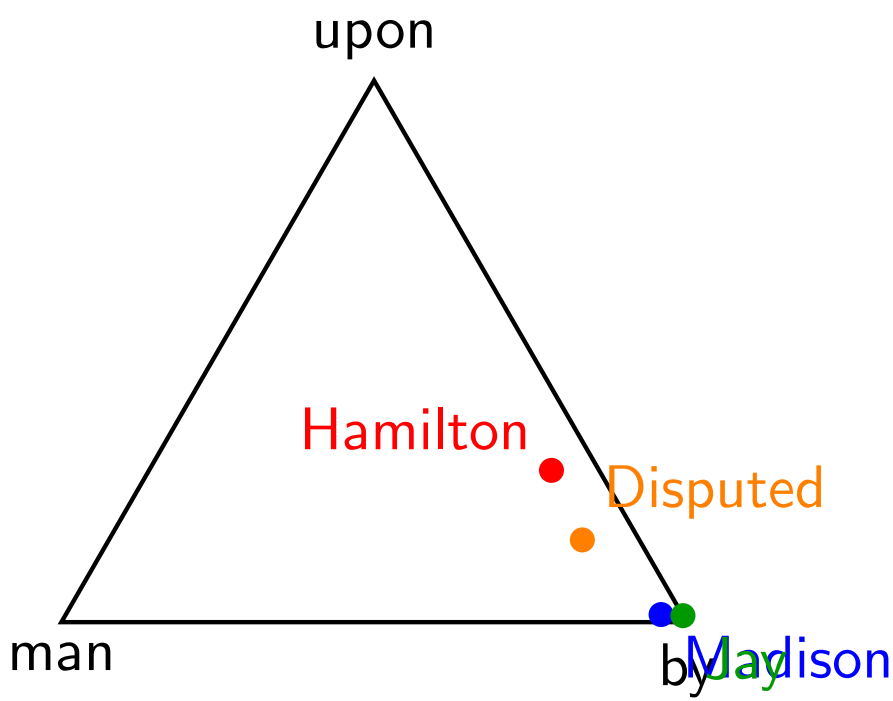
Word Proportions

Multinomial Model of Language

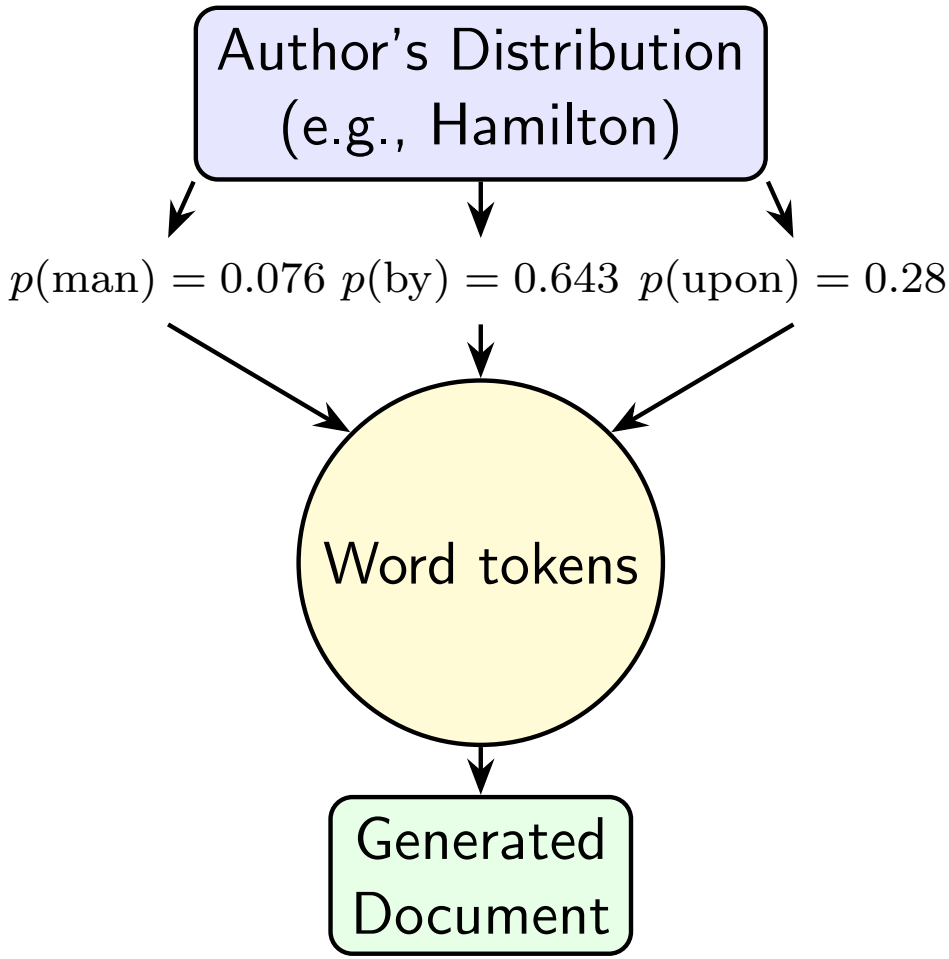


	man	by	upon
Hamilton	.076	.643	.28
Madison	.034	.952	.014
Jay	0	.988	.012

Word Proportions in a Probability Simplex



Multinomial Model of Language



Each document is generated by sampling words from the author's multinomial distribution.

From Counts to Probabilities

- Start with word counts for each author:

$$\text{Hamilton: } \{\text{man} = 102, \text{by} = 859, \text{upon} = 374\}$$

- Normalize counts to get probabilities:

$$p(\text{man}) = 0.076, \quad p(\text{by}) = 0.643, \quad p(\text{upon}) = 0.280$$

- These probabilities define the author's **multinomial distribution over words**.

Multinomial Likelihood

- For a new document, we observe counts:

$$x = (x_{\text{man}}, x_{\text{by}}, x_{\text{upon}})$$

- The probability of seeing x given author A is:

$$P(x \mid A) = \frac{N!}{\prod_i x_i!} \prod_i p_i^{x_i}$$

- $N = \sum_i x_i =$ total words in the document
- $p_i =$ probability of word i for author A

From Likelihood to Classification

- Compute $P(x \mid \text{Hamilton})$ and $P(x \mid \text{Madison})$
- Compare: which author makes the observed document more likely?
- Classification rule:

$$\hat{A} = \arg \max_{A \in \{\text{Hamilton, Madison, Jay}\}} P(x \mid A)$$

- This is essentially the **Naive Bayes classifier** for text.

From the Federalist Papers to Naive Bayes

- Mosteller & Wallace (1964) used the **multinomial model of language**:

$$P(x \mid A) = \frac{N!}{\prod_i x_i!} \prod_i p_i^{x_i}$$

- This is exactly the likelihood function used in a **multinomial Naive Bayes classifier**.
- Modern text classification (spam detection, sentiment, authorship) often relies on the same idea:

$$\hat{A} = \arg\max_A \; P(A) \prod_i p_i^{x_i}$$

- Key differences today:
  - More words (full vocabulary, not just "man/by/upon").
  - Additive smoothing to handle zeros.
  - Scaling to millions of documents with efficient implementations.

Bayesian Uncertainty: Dirichlet–Multinomial

- Prior: $\mathbf{p}_A \sim \text{Dirichlet}(\boldsymbol{\alpha})$
- Data (counts): $\mathbf{c}_A \quad \Rightarrow \quad$ Posterior:

$$\mathbf{p}_A \mid \mathbf{c}_A \sim \text{Dirichlet}(\boldsymbol{\alpha} + \mathbf{c}_A)$$

- Posterior predictive for a new document x:

$$P(x \mid \mathbf{c}_A, \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + N)} \prod_{i=1}^{K} \frac{\Gamma(\alpha_i + c_{A,i} + x_i)}{\Gamma(\alpha_i + c_{A,i})}$$

- Authors: $P(A \mid x) \propto P(x \mid \mathbf{c}_A, \boldsymbol{\alpha})\, P(A)$

Credible Intervals You Can Report

- Word-level: 95% CI for $p_{A,i}$ from Dirichlet draws.
- Word-level contrast: CI for log-odds ratio via Beta sampling.
- Document-level: CI for $P(A \mid x)$ via posterior predictive draws.
- Decision robustness: distribution of log-likelihood differences across draws.