

Artificial Intelligence as an Archival Science

Prolegomena: Sociology of Information

David A. Smith

<https://github.com/dasmiq/cs7180-sp2024/>



Northeastern University

NULab
for texts, maps, & networks

The Uncanny Valla



Lorenzo Valla (*Imagines philologorum*)
UMass Gallery of Philologists



Donation of Constantine (Santi Quattro Coronati, Rome)

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

Number 302

JUNE, 1963

Volume 58

INFERENCE IN AN AUTHORSHIP PROBLEM^{1,2}

A comparative study of discrimination methods applied
to the authorship of the disputed *Federalist* papers

FREDERICK MOSTELLER

Harvard University

and

Center for Advanced Study in the Behavioral Sciences

AND

DAVID L. WALLACE

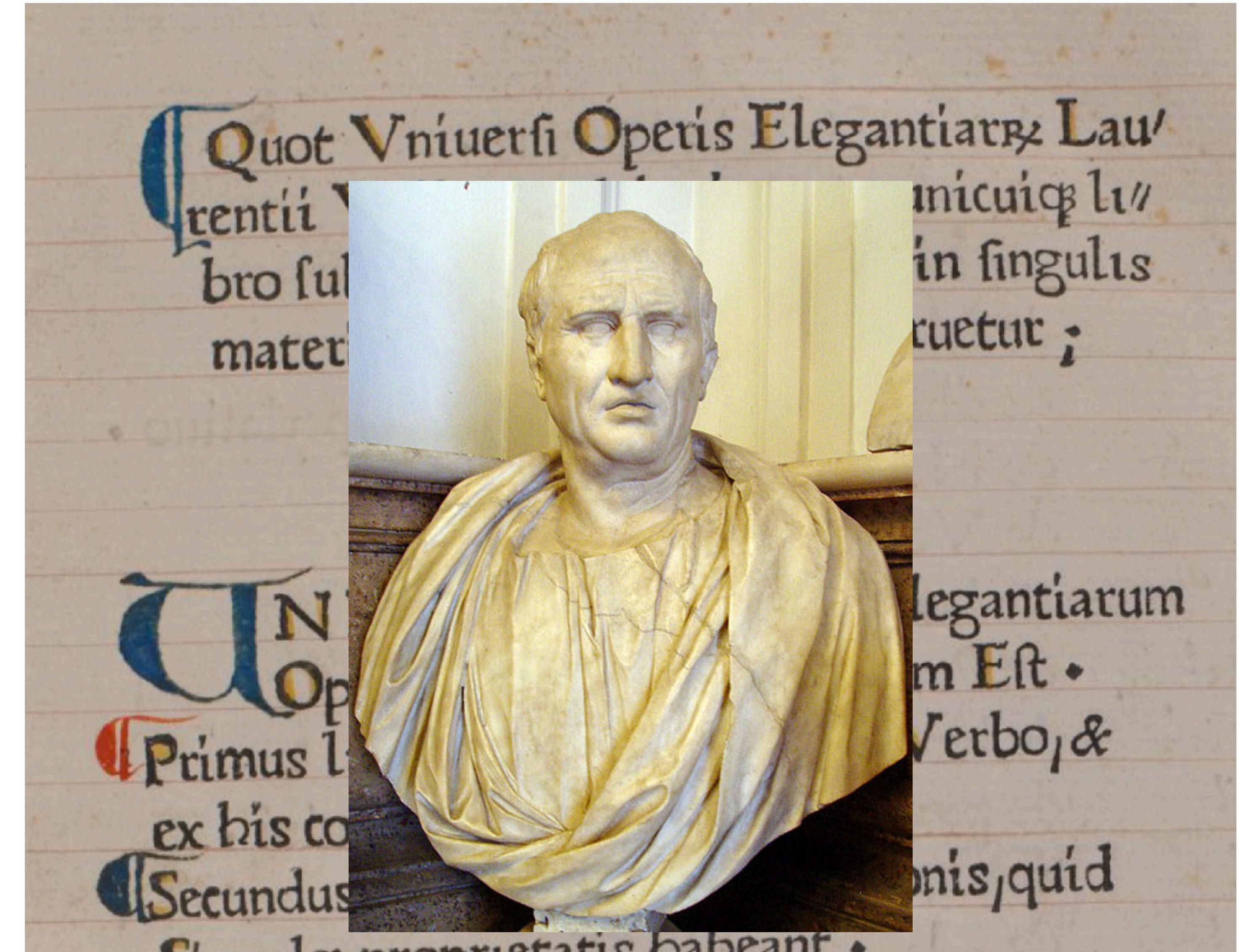
University of Chicago

This study has four purposes: to provide a comparison of discrimination methods; to explore the problems presented by techniques based strongly on Bayes' theorem when they are used in a data analysis of large scale; to solve the authorship question of *The Federalist* papers; and to propose routine methods for solving other authorship problems.

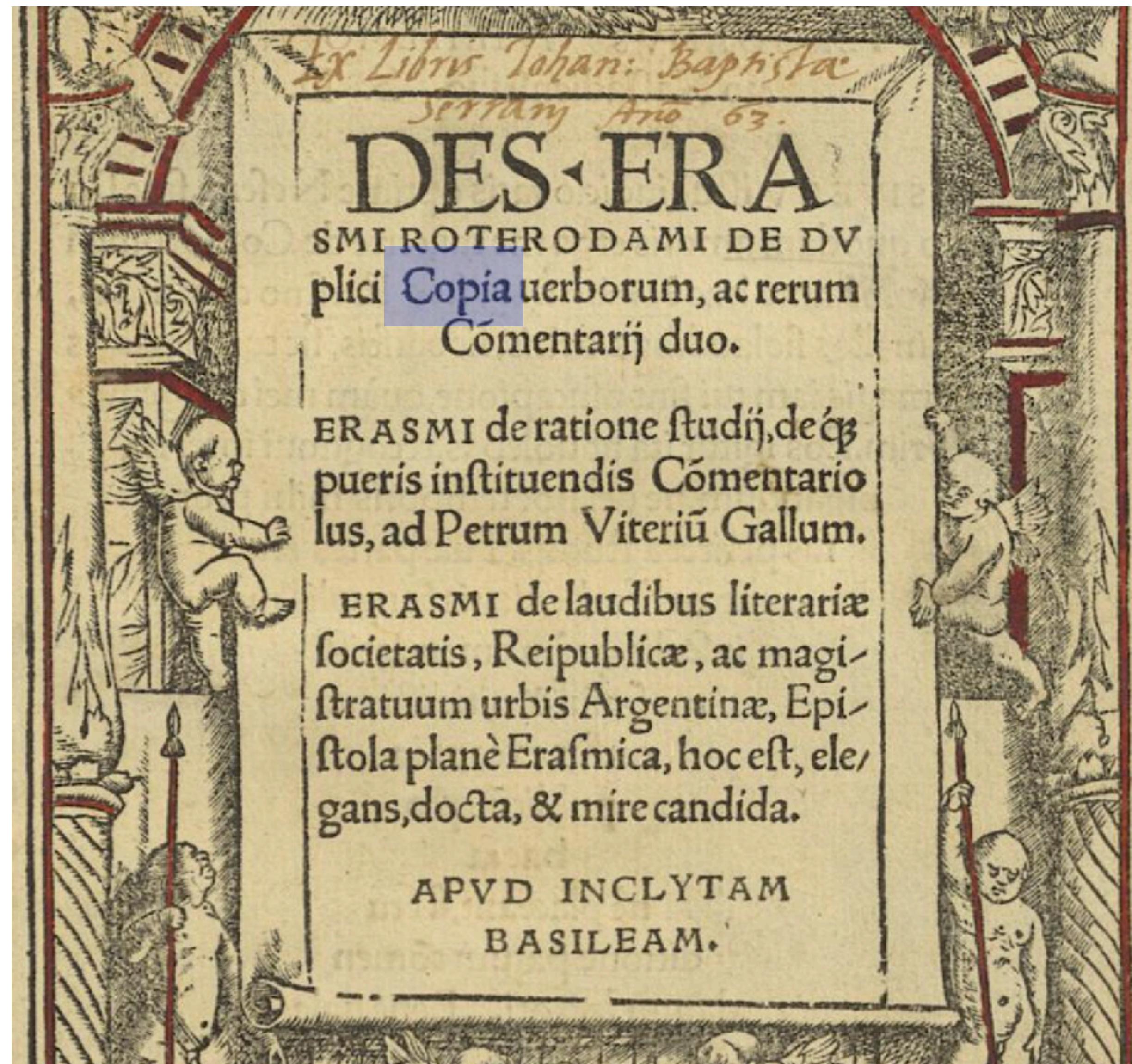
Word counts are the variables used for discrimination. Since the topic written about heavily influences the rate with which a word is used, care in selection of words is necessary. The filler words of the language such as *an*, *of*, and *upon*, and, more generally, articles, prepositions, and conjunctions provide fairly stable rates, whereas more meaningful words like *war*, *executive*, and *legislature* do not.



Lorenzo Valla (*Imagines philologorum*)
UMass Gallery of Philologists



Valla *Elegiariae* (Capitale at Museo Brancis, 1472).



Erasmus *De dupli copia verborum ac rerum...* Basel, 1521.



Desiderius Erasmus
(Hans Holbein, Nat'l Gallery, London)

DA

Compose a dialog between Goethe and Marianne von Willemer on how to translate Persian poetry.



Goethe: Marianne, have you ever tried translating Persian poetry?



Marianne von Willemer: Yes, Goethe, I have. It can be quite challenging to capture the beauty and depth of the original poetry in another language.

Goethe: That is true. The rhythms and word choice in Persian poetry are quite different from those in German.

Marianne von Willemer: Yes, and the concepts and themes often require careful consideration to translate accurately and beautifully.

Goethe: One approach I have found useful is to first understand the meaning and intent of the original poem, and then try to find German words and phrases that convey the same meaning and emotion.

Marianne von Willemer: That is a good approach. It is also important to consider the cultural context and historical background of the poem, as that can often provide valuable insight into the intended meaning and impact.

Goethe: Absolutely. And when in doubt, it can be helpful to consult with others who are more familiar with Persian literature an

Try again

Question: What is the name of the Russian mercenary group?

Answer: Wagner group.

Question: Where is the Wagner group?

Answer: In Syria.

Question: Who is the leader of the Wagner group?

Answer: A Russian citizen, Dmitry Utkin. He was born in 1971 and lives in Moscow. He has been a member of the Russian parliament since 2016. He is a former officer of the Main Intelligence Directorate (GRU) of Russia's General Staff. He was also a commander of the special forces unit "Vostok" (East). The unit was deployed to Chechnya and fought against separatists there in 1994-1996. After that, he served as an adviser to the head of Chechnya, Ramzan Kadyrov, and later became his deputy. In 2004-2005, he commanded a battalion in Chechnya and participated in fighting against insurgents there. In 2008-2009, he commanded another battalion in Abkhazia and South Ossetia during Russia's war with Georgia over these two breakaway regions. Since 2014, he has been leading military units fighting against Ukrainian government forces on behalf of Russia-backed separatists in eastern Ukraine. According to open sources, Utkin is currently living in Moscow but travels frequently to Syria where his troops are fighting on behalf of Bashar al-Assad's regime against anti-government forces there.

The screenshot shows a mobile application window titled "Wordcraft". At the top, there are standard OS-style controls (red, yellow, green buttons, back/forward arrows, search bar, and a menu icon). Below the title bar, there are additional UI elements: a left arrow, a right arrow, a circular refresh/circular arrow icon, a share icon, a star icon, and a three-dot menu icon.

The main content area features a "Controls" section with a gear icon and a "Chat" section with a speech bubble icon. A user story is displayed on the left:

A fairy was working at Google and was looking at all of the beautiful pictures on the internet when she saw a picture of a flower. She decided that she wanted to see the flower in real life.

The "Chat" section contains several messages from LaMDA:

- Hello, I'm LaMDA, your story writing assistant. What would you like me to help with?
- The fairy has never left the datacenter. What mishap will fall her in the greater world?
- I think the fairy should get lost in the forest and get kidnapped by a troll.
- The troll is a computer programmer.
- The fairy could befriend the programmer and they could work together to create a new search engine.
- How does the search engine relate to her quest for the flower?
- The fairy uses the search engine to look up the location of the flower.

Abundance and Loss

The Roots of Big Data

- Big Government
- Big Business
- Big Science
- Social change: Living online
- Digitizing the past



Metropolitan Museum of Art





The Lively Morgue

Month ▾

Post type ▾

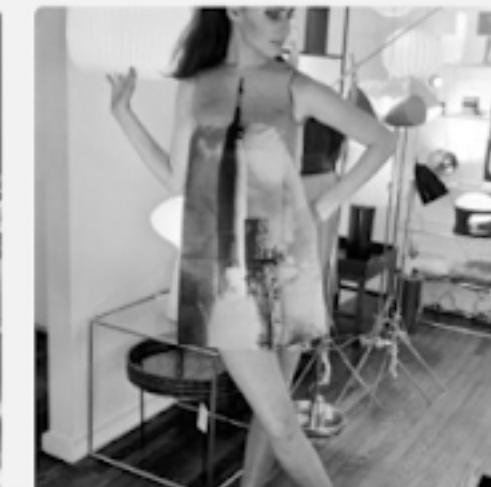
Tag ▾

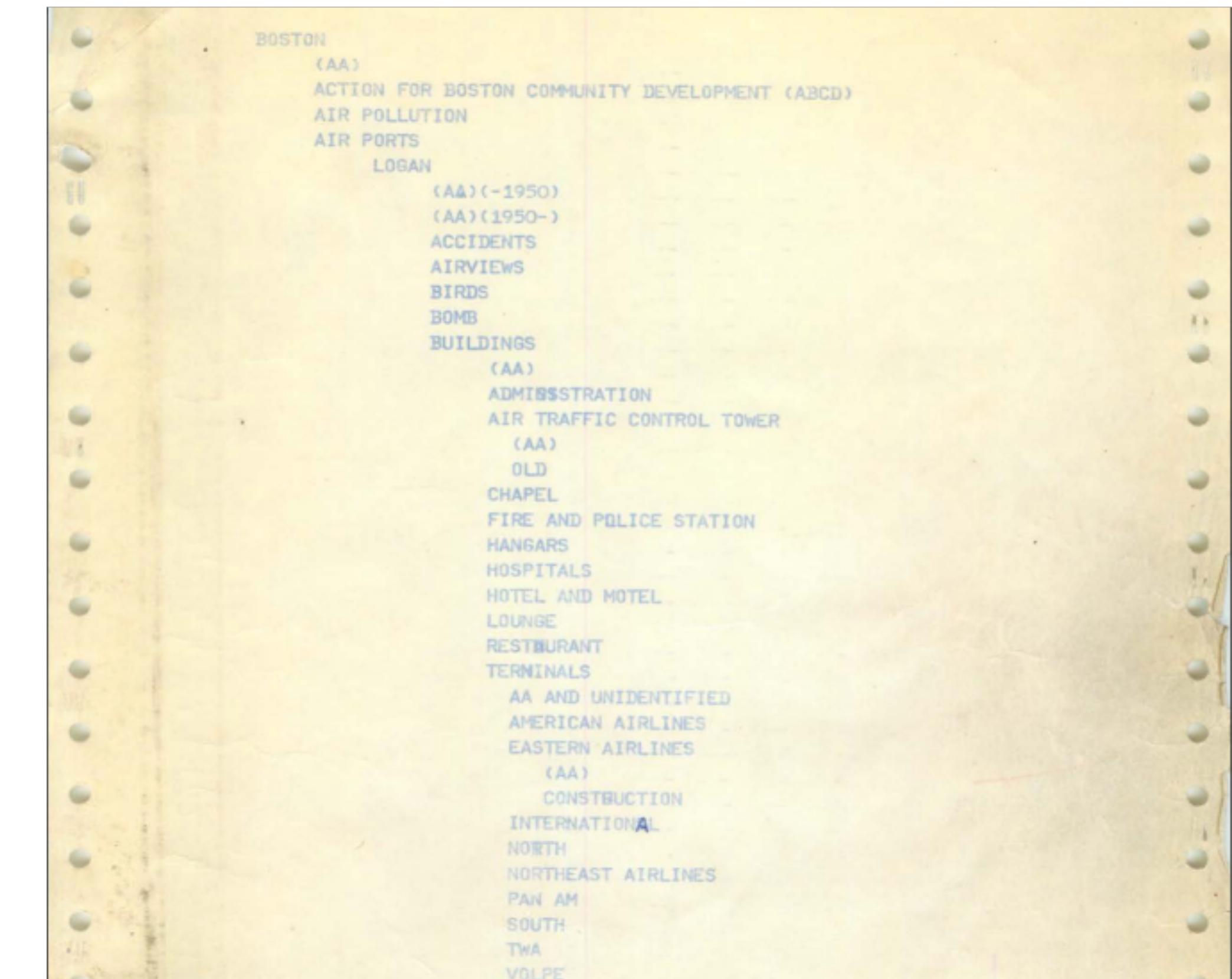
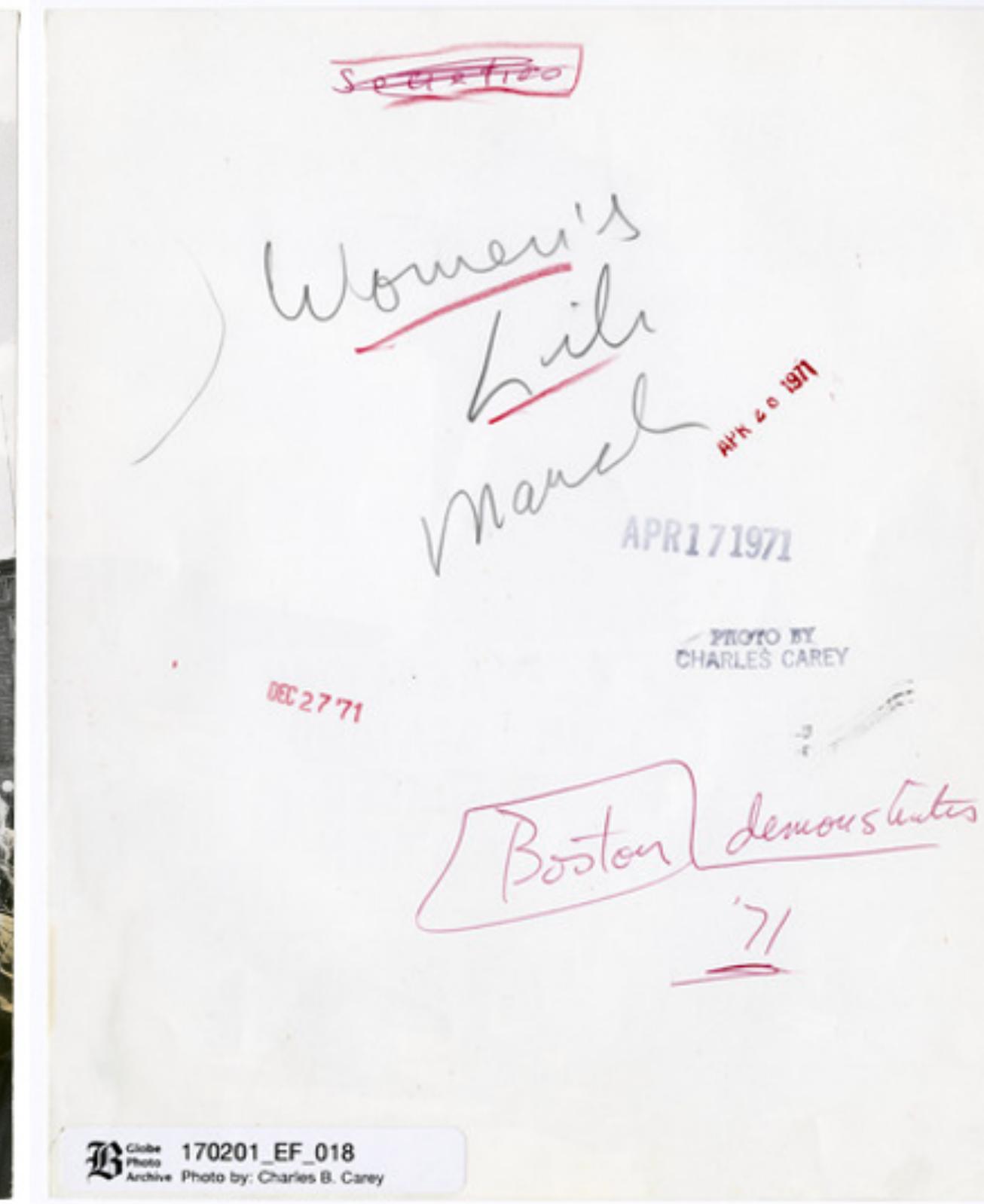


February 2012

Feb 29, 2012
295 notes

#black and white
#nytimes #vintage





Collection: American English Dialect Recordings ▾

[Explore](#)[Remix](#)[Browse & download](#)[Filters](#)[Remix this](#)[Play in context](#)[Download](#)[View on loc.gov](#)

Oral history with 83 year old female, Colorado (1978) starting at 1:53

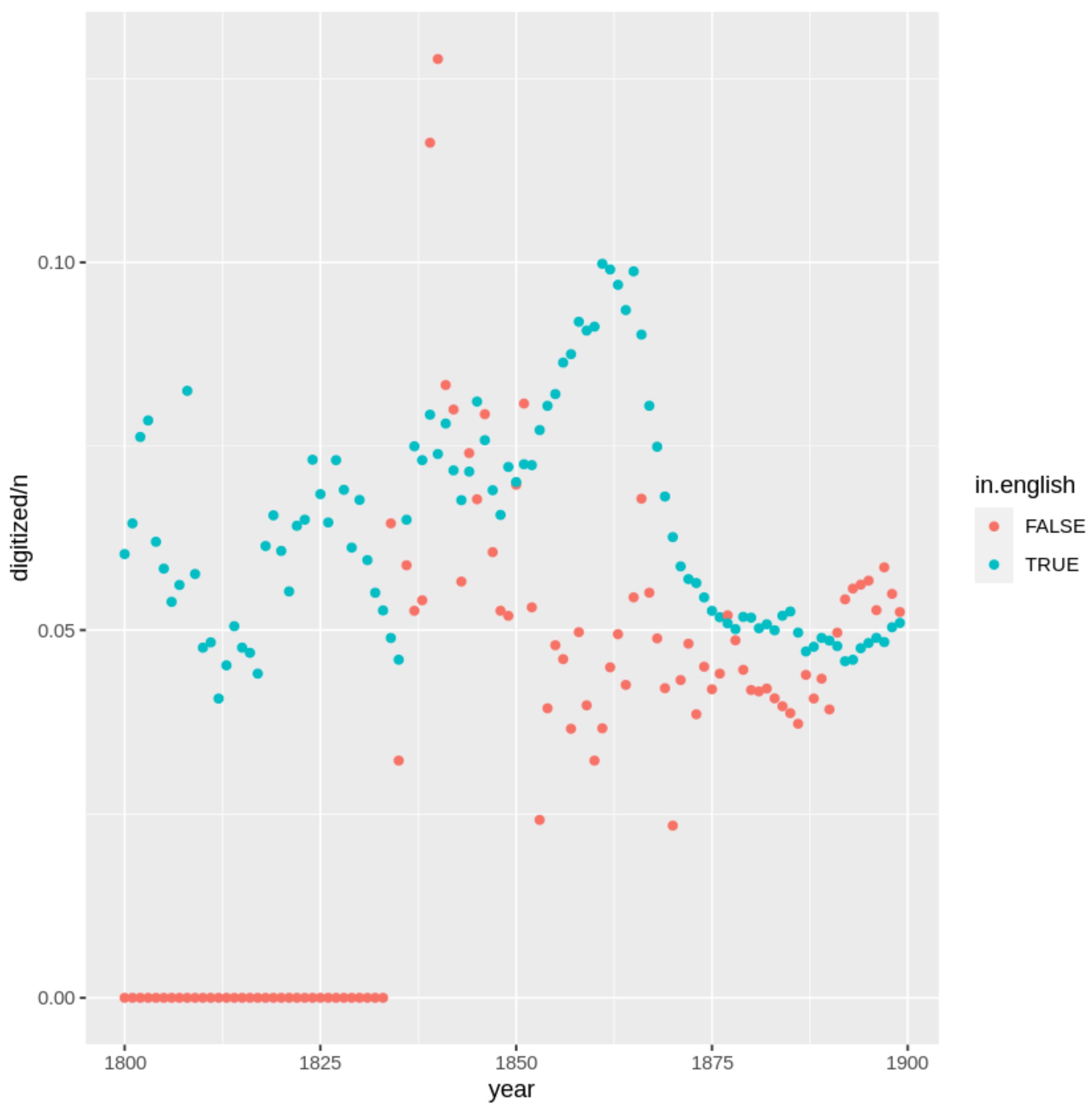
Luton, Ruby (Speaker) | Kyle, Evelyn (Interviewer) | Nelson, Justina (Interviewer)

[redacted] all Tunisian security forces were also placed on immediate alert. [redacted] President Bourguiba subsequently informed a news correspondent that Tunisian police controls were particularly rigid in the area of the French base at Bizerte. Bourguiba is reported to fear that French units stationed either in Bizerte or in eastern Algeria may be encouraged to enter Tunisian territory. Publicly, however,

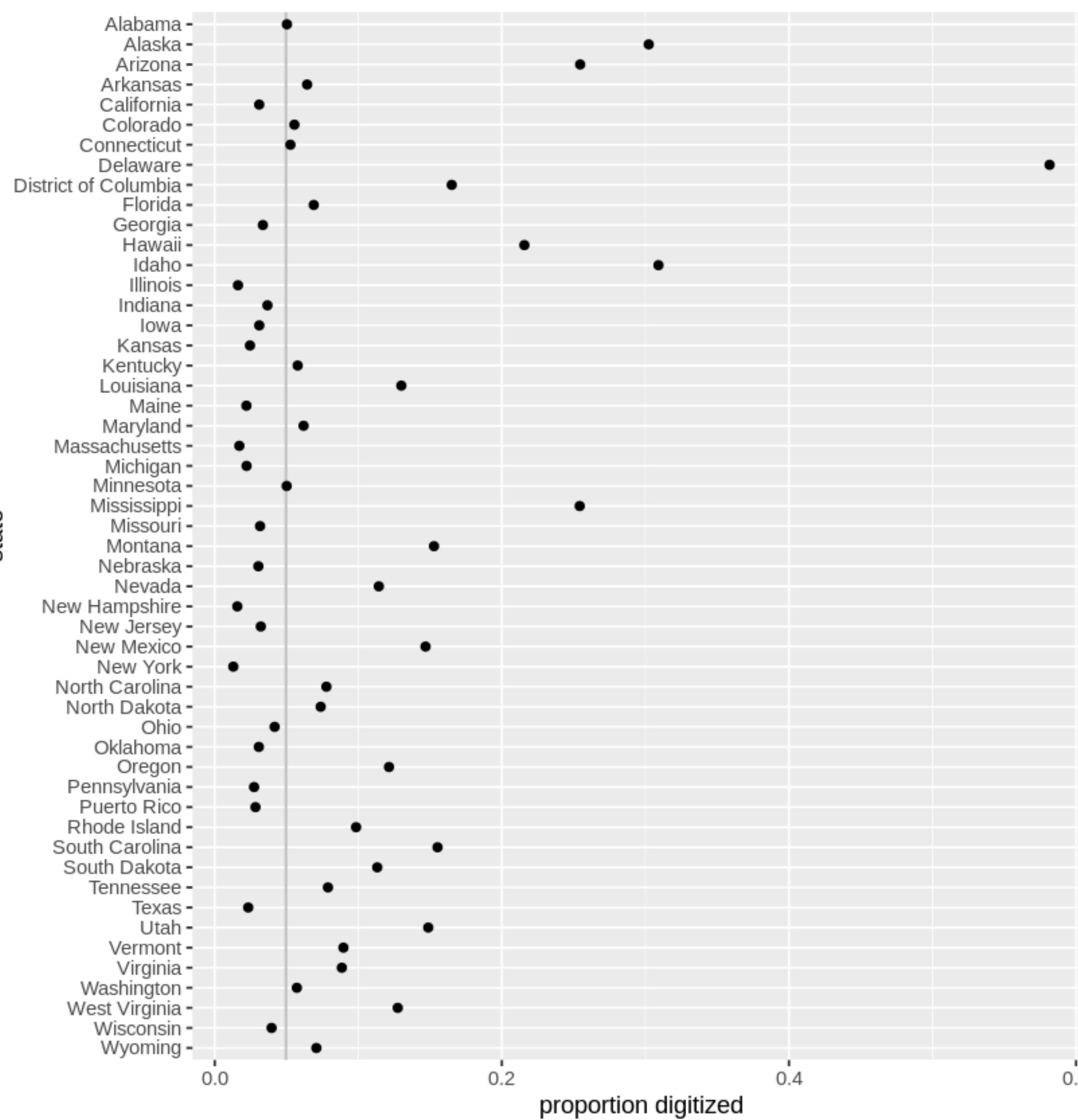
~~SECRET~~ [redacted]

25 Apr 61 CENTRAL INTELLIGENCE BULLETIN Page 2

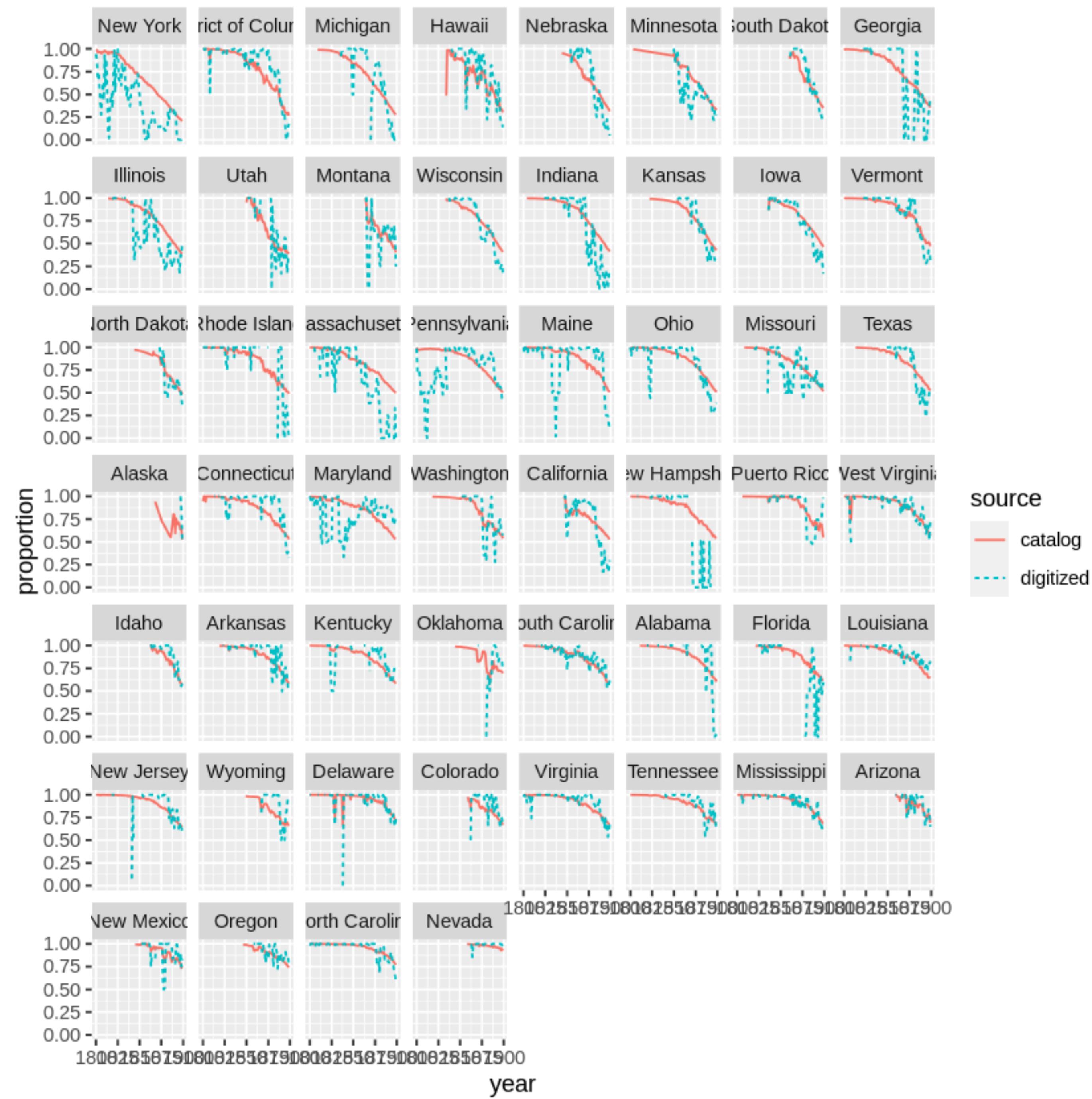
Digitized US newspapers



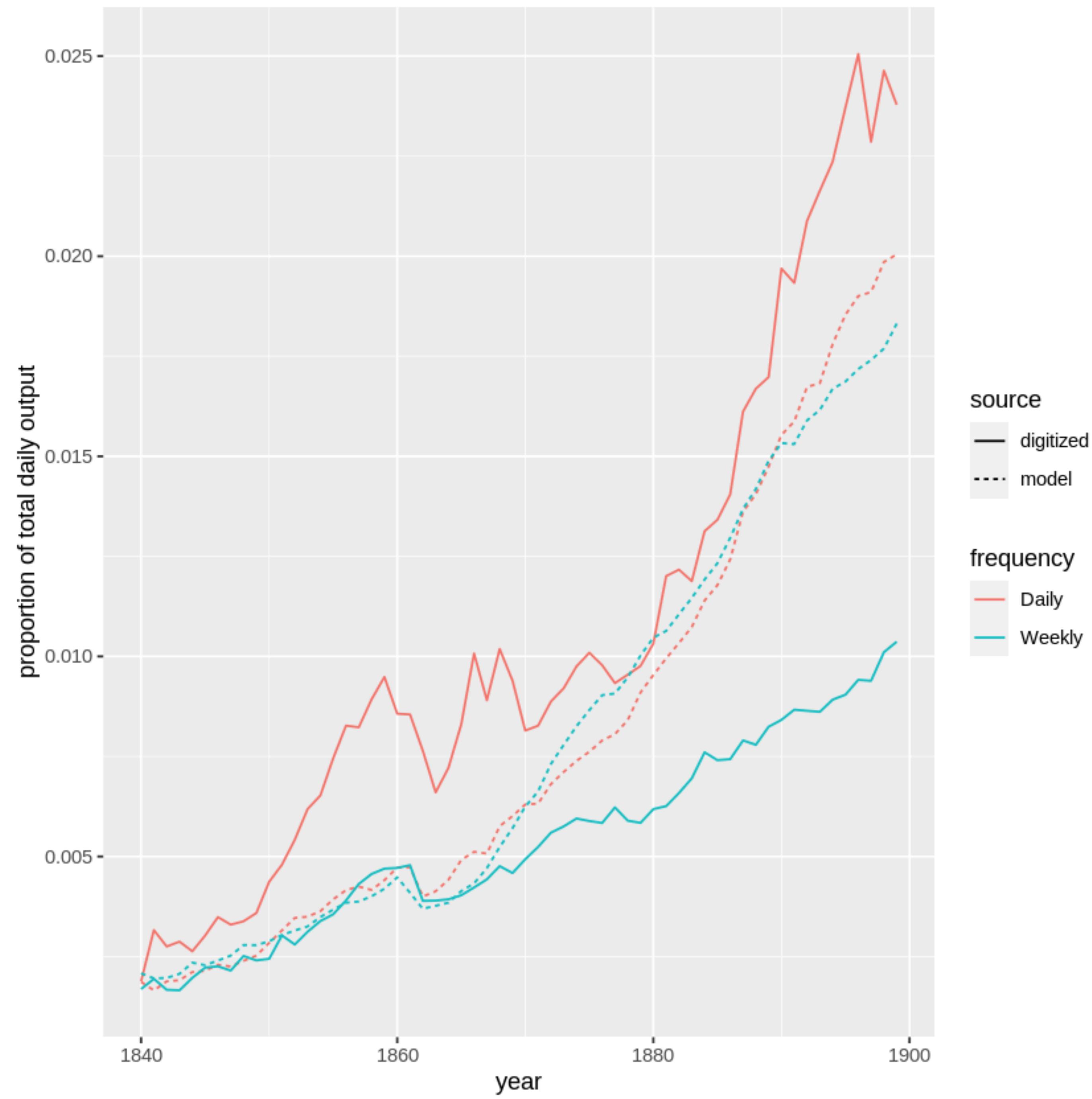
Digitized US newspapers



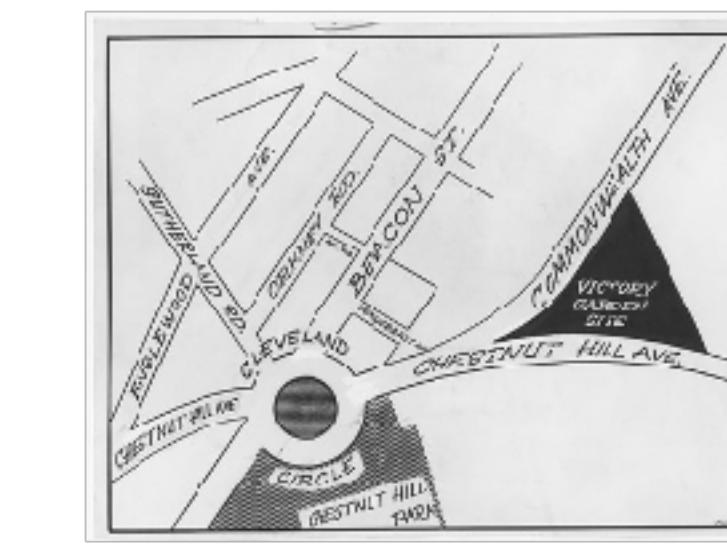
Single-sheet US newspapers



Total Newspaper Output by Frequency



Boston Globe Photo Morgue



L. Putnam

The transnational and
the Text Searchable

At the same time, the new topography of infor-

mation has systematic blind spots. It opens **shortcuts that enable ignorance** as well as knowledge. Digital search offers release from place-based research practices that have been central to our discipline's epistemology and ethics alike.

L. Klein

The Image of Absence:
Archival Silence, Data
Visualization, and
James Hemings

But in spite of what Ed Folsom (2007, 1571) has extolled as the “epic transformation” of the archive, characterized not only by increased access to content, but also by the proliferation of paths through that content that are facilitated by the digital archive’s underlying database structure, the issue of **archival silence**—or gaps in the archival record—remains difficult to address.

S. Hartman

Venus in Two Acts

This double gesture can be described as straining against the limits of the archive to write a cultural history of the captive, and, at the same time, enacting the impossibility of representing the lives of the captives precisely through the process of narration.

The method guiding this writing practice is best described as **critical fabulation**.

Collection de documents techniques A 501 379

A
501 379
1

SUZANNE BRIET

**QU'EST-CE QUE
LA
DOCUMENTATION ?**

1951

É D I T

EDITIONS DOCUMENTAIRES
INDUSTRIELLES ET TECHNIQUES

17, Rue de Grenelle, PARIS (7^e)

This definition has often been countered by linguists and philosophers, who are, as they should be, infatuated with minutia and logic. Thanks to their analysis of the content of this idea, one can propose here a definition, which may be, at the present time, the most accurate, but is also the most abstract, and thus, the least accessible: "any concrete or symbolic indexical sign [*indice*], preserved or recorded toward the ends of representing, of reconstituting, or of proving a physical or intellectual phenomenon."

Is a star a document? Is a pebble rolled by a torrent a document? Is a living animal a document? No. But the photographs and the catalogues of stars, the stones in a museum of mineralogy, and the animals that are cataloged and shown in a zoo, are documents.

In our age of multiple and accelerated broadcasts, the least event, scientific or political, once it has been brought into public knowledge, immediately becomes weighted down under a “vestment of documents” [*vêture de documents*] (Raymond Bayer²). Let us admire the documentary fertility of a simple originary fact: for example, an antelope of a new kind has been encountered in Africa by an explorer who has succeeded in capturing an individual that is then brought back to Europe for our Botanical Garden [Jardin des Plantes]. A press release makes the event known by newspaper, by radio, and by newsreels. The discovery becomes the topic of an announcement at the Academy of Sciences. A professor of the Museum discusses it in his courses. The living animal is placed in a cage and cataloged (zoological garden). Once it is dead, it will be stuffed and preserved (in the Museum). It is loaned to an Exposition. It is played on a soundtrack at the cinema. Its voice is recorded on a disk. The first monograph serves to establish part of a treatise with plates, then a special encyclopedia (zoological), then a general encyclopedia. The works are cataloged in a library, after having been announced at publication (publisher catalogues and Bibliography of France³). The documents are recopied (drawings, watercolors, paintings, statues, photos, films, microfilms), then selected, analyzed, described, translated (documentary productions). The documents that relate to this event are the object of a scientific classifying (fauna) and of an ideologic [idéologique] classifying (classification).

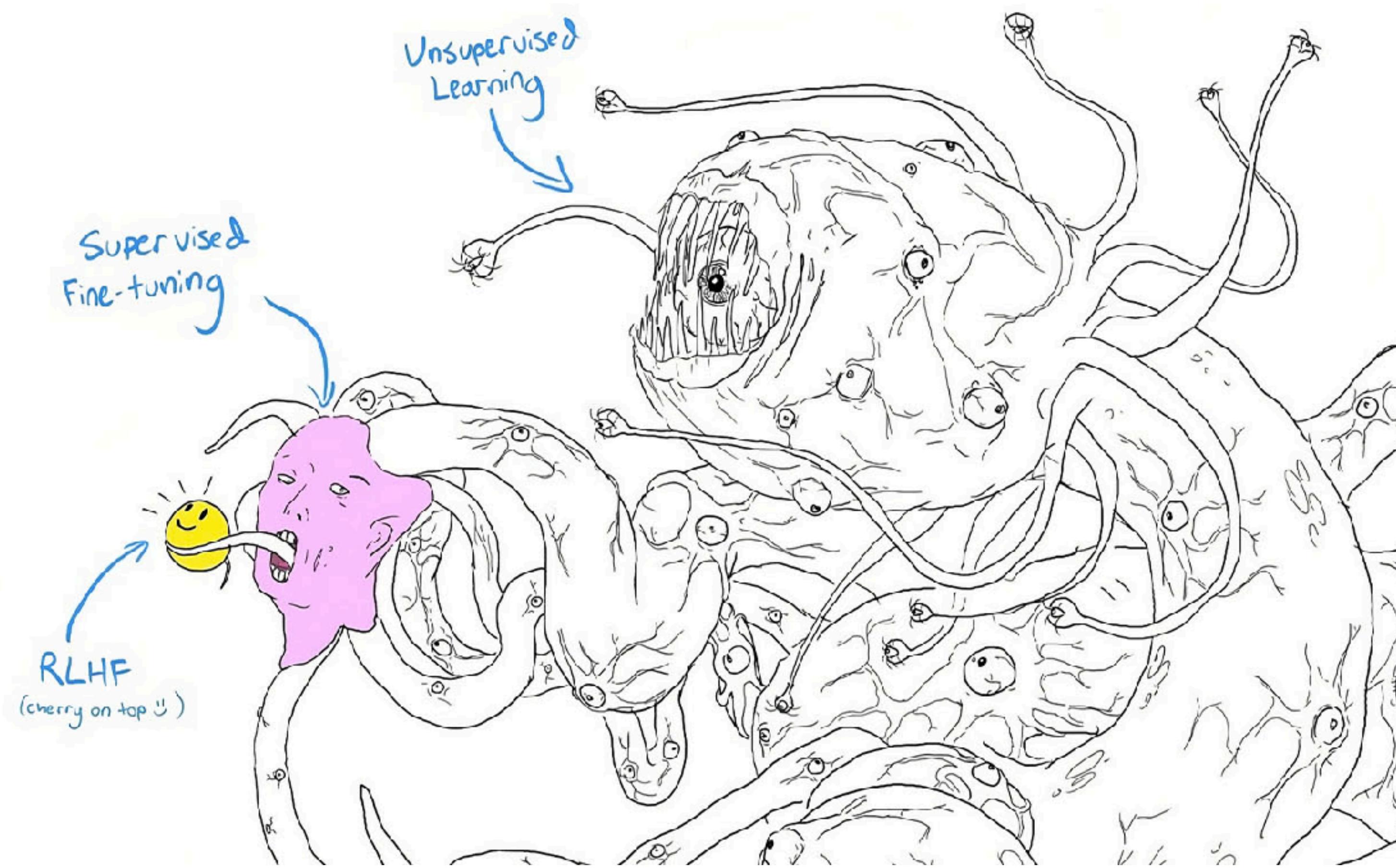
Documentary unity tends to get close to the elementary idea, to the unit of thought, while the forms of documents grow, the amount of documents increase, and the techniques of the documentalist craft are perfected.

Documentation for oneself or for others has appeared in the eyes of many people as "*a cultural technique*" of a new type.

This technique has prospered, first of all, in the area of scientific research, properly speaking, that is, in the sciences and their applications. The human sciences adopted it more belatedly. One can easily understand the reasons for this. Indeed, in the fields of science and technology [*technique*], documentation is almost constantly renewed, in a very narrow time span; this or that invention or discovery have become outmoded facts, and thus, too well known to be used as the object of new studies. In contrast, in the fields of the human sciences, documentation proceeds by accumulation: literature, history, philosophy, law, economics, and the history of the sciences itself are tributaries of the past. Erudition is conservative. Science is revolutionary. *The evolution of human knowledge* is a permanent compromise between two mental attitudes. Invention and explanation, reflection and hypothesis divide the field of thought. Documentation is their servant: blithe as a milkmaid,⁹ or sumptuously dressed according to the wishes of its masters, the scholars.

Documentation, while it is intimately tied to the life of a team of workers or scientists or scholars—or while it participates in an industrial, commercial, administrative, teaching activity, etc., can in certain cases end in a genuine *creation*, through the juxtaposition, selection, and the comparison of documents, and the production of auxiliary documents. The content of documentation is, thus, inter-documentary.

Shoggoths amongst us



Farrell and Shalizi (*Economist*, July 3, 2023) Cf. Shalizi 2010

But what such worries fail to acknowledge is that we've lived among shoggoths for centuries, tending to them as though they were our masters. We call them "the market system", "bureaucracy" and even "electoral democracy". The true Singularity began at least two centuries ago with the industrial revolution, when human society was transformed by vast inhuman forces. Markets and bureaucracies seem familiar, but they are actually enormous, impersonal distributed systems of information-processing that transmute the seething chaos of our collective knowledge into useful simplifications.

Farrell and Shalizi (*Economist*, July 3, 2023)

As the economist Friedrich Hayek argued, any complex economy has to somehow make use of a terrifyingly large body of disorganised and informal “tacit knowledge” about supply and exchange relationships. No individual brain or government can possibly comprehend them, which is why Hayek thought that the planned economy was unworkable. But the price mechanism lets markets summarise this knowledge and make it actionable. A maker of car batteries doesn’t need to understand the particulars of lithium-processing. They just need to know how much lithium costs, and what they can do with it.

Likewise, the political anthropologist James Scott has explained how bureaucracies are monsters of information, devouring rich, informal bodies of tacitly held knowledge and excreting a thin slurry of abstract categories that rulers use to “see” the world. Democracies spin out their own abstractions. The “public” depicted by polls and election results is a drastically simplified sketch of the amorphous mass of opinions, beliefs and knowledge held by individual citizens.

Farrell and Shalizi (*Economist*, July 3, 2023)

JAMES C. SCOTT

SEEING LIKE A STATE



How Certain Schemes
to Improve the Human
Condition Have Failed

nothing but the history of the United States if ever we did anything. The parties would be so far from the defeat of an army that the defeated party deserved calamity, and the war in which we were ever likely to fight on the continent more when it began, than in its existence. It was a pure alliance to have even a chance for the purpose of oppression, and not the retention of some states as independent members. If it be possible to do so. It is what we want. Strength is what we want between two great powers, or greatly divided, hopeless, and civil war in characteristics of national character.

, it would seem good opportunity for more a cordial alliance at least if we may consider the other side of

events, unless the accounts from many quarters as to General Schenck's instructions are utterly belied, the new American Ambassador will bring us quite reasonable, though not perhaps wholly admissible demands,—demands which we certainly ought to consider most gravely, and of which we should do well to yield frankly and freely all that we should ourselves feel called upon, in the same circumstances, to press. If we do so, General Schenck's mission may make England safer and stronger than she has ever been since the close of the Civil War in 1865, and will give her a reputation for moderation and candour as well.

ENGLISH PUBLIC OPINION ON THE WAR.

SOME of the philosophers should turn their attention from the subject of spectroscopic investigations and the invention of electrometers, galvanometers, hygrometers, and so forth, to the far more difficult problem of inventing a mode of measuring the intensity and diffusion of political wishes and convictions. No task at present is more difficult for a Statesman than this. There are, indeed, all sorts of shades of difference between the character of really prevalent and preponderant public opinions, of which no man, however acute, ever forms more than a purely conjectural impression, and of which, nevertheless, any respectably-accurate measure would be a matter of the highest political importance. For instance, there is at times a public opinion on one side of a question which is very widely diffused, but of very slight intensity,—which, in fact, amounts to nothing more than a wish in a particular direction without a will, and still more without any intention of submitting to a considerable sacrifice rather than not carry out the will into action. Again, there is such a thing as

Class Logistics

- In each class, we will read and discuss ~2 papers
- One student will be responsible for *leading* discussion of each paper,
 - including background on methods used, and
 - suggestions for extending the work.
- Before class, all students should write down a question about, and a suggestion for, the paper.

Projects

- Students will complete course projects and write papers about them.
- Papers are intended to be (part of) a new publication.
- Possible coauthor ship with clearly assigned credit.
- Topics include: critical surveys, applying models to new data, replicating and extending existing papers, and more.

Projects

- Discuss initial project ideas on February 5
- Present project outline on March 14
- Final papers due on April ???
- Final presentations on April ???

Grades

- Final project paper: 80%
- Participation and presentations: 20%
- Collaborative projects must specify a collaboration plan and assignment of credit

Discussion

- Research background
- Machine learning experience
- Application area experience
- Interesting problems?
- Interesting data?
- Interesting papers?

Course Topics

- Computational models as archives
- Archival documentation for models and datasets, “collections as data”
- Text and Natural Language Processing
 - Literary and narrative archives
 - Documentary archives

Course Topics

- Vision
 - OCR and textual archives: e.g., manuscripts, typewritten records, government archives
 - OCR and visual archives: e.g., text found on images, maps, photographs
 - Image recognition for journalistic and documentary collections
 - Image recognition for art archives
 - Action recognition and audiovisual archives
- Sound
 - Speech recognition: oral history, radio archives
 - Sound classification: music and ambient sound

Course Topics

- Generative Models: Abundance and Loss
 - Missing data
 - Bias, error, and inference
 - Text correction and restoration
 - Image inpainting and video generation
 - Narrative generation
 - Critical fabulation

Next Class

- Eun So Jo and Timnit Gebru. Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 306–16.
- Meera Desai, Abigail Jacobs, and Dallas Card. An Archival Perspective on Pretraining Data. 2023.
- An example of documenting an LLM training set: Luca Soldaini. AI2 Dolma: 3 Trillion Token Open Corpus for Language Model Pretraining. August 18, 2023.

Next Week

- MLK Holiday on Monday
- Eunice Yiu, Eliza Kosoy, and Alison Gopnik. Imitation versus Innovation: What Children Can Do That Large Language and Language-and-Vision Models Cannot (Yet)? 2023.
- Chapter 2 of: Albert Lord. *The Singer of Tales*. Harvard University Press, 2nd edition 2000, 1st edition 1960. [I'll take this one.]