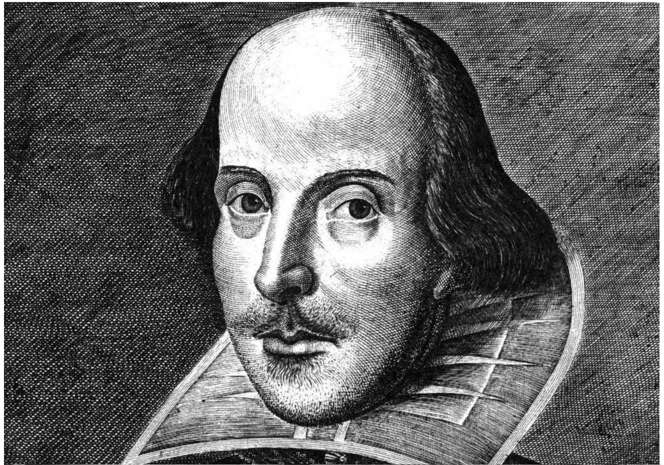
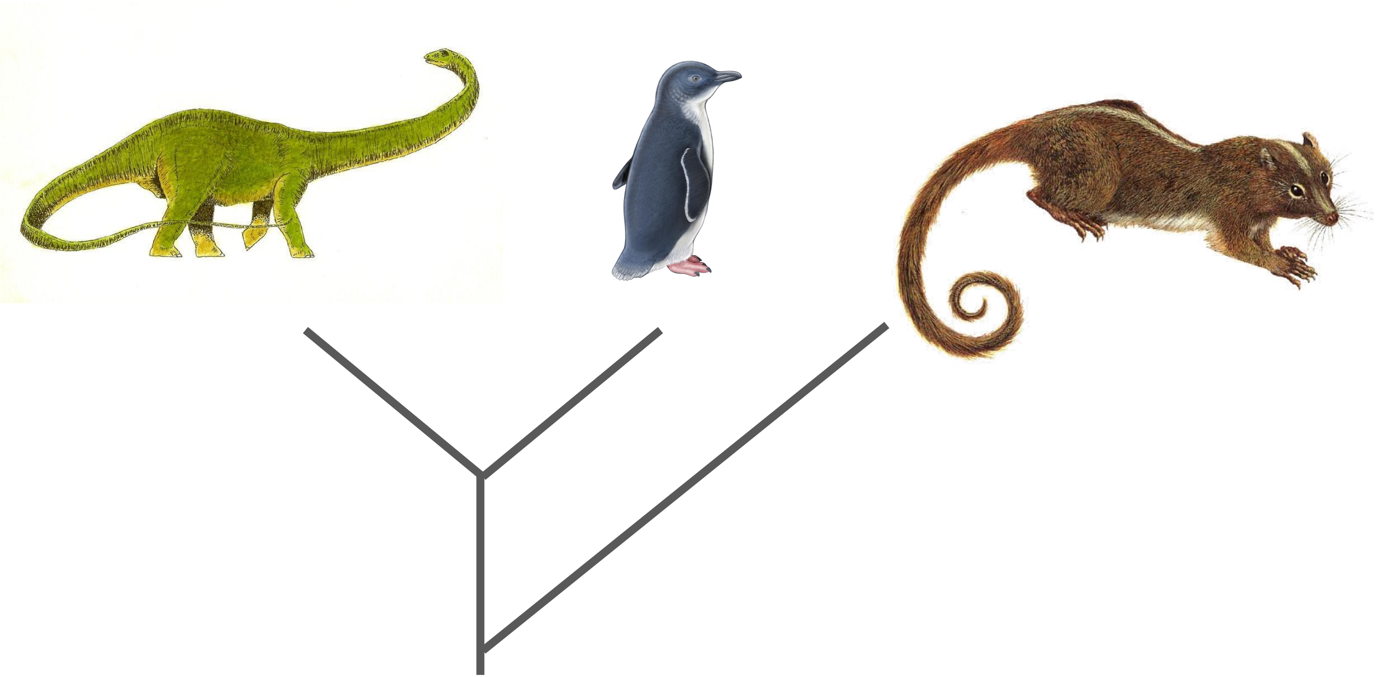
Stylometry and authorship attribution



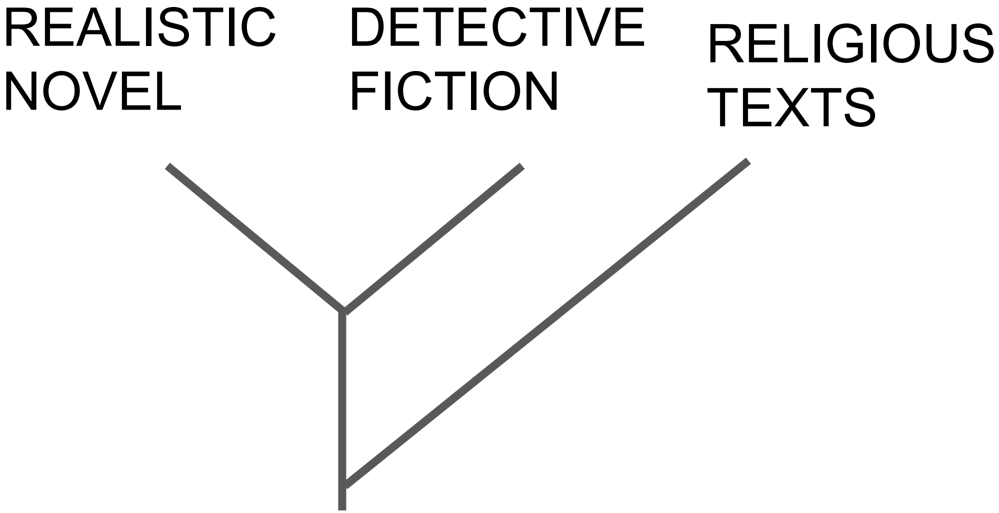
How similar? How different?

Think in terms of distances: what evolutionary path each species has undergone

Think in terms of distances: what evolutionary path each species has undergone



Continue thinking in terms of distances

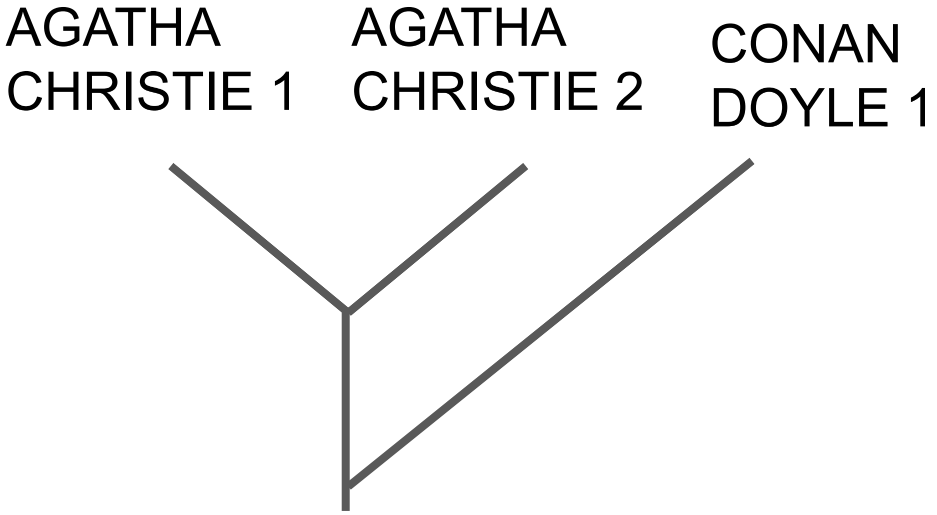


REALISTIC NOVEL

DETECTIVE FICTION

RELIGIOUS TEXTS

Continue thinking in terms of distances

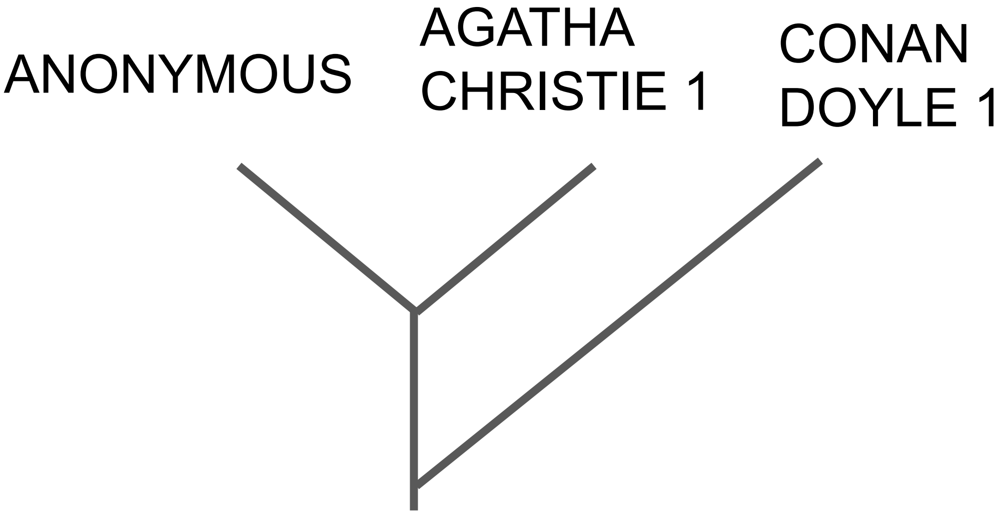


AGATHA CHRISTIE 1

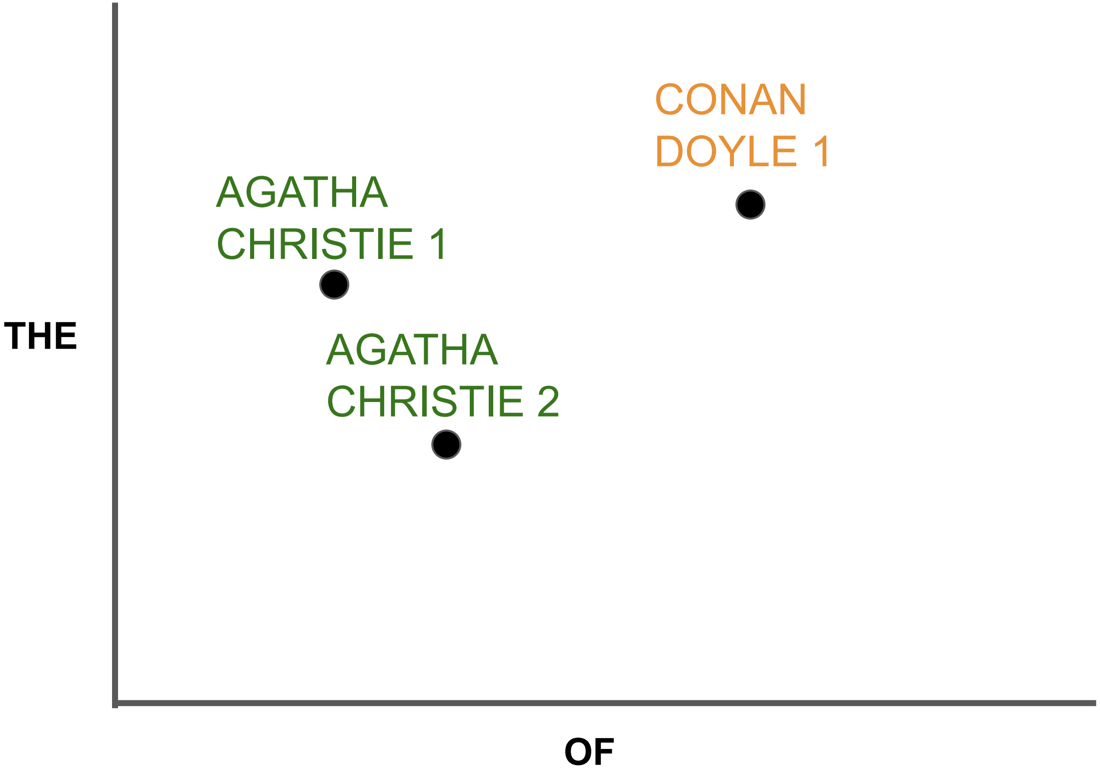
AGATHA CHRISTIE 2

CONAN DOYLE 1

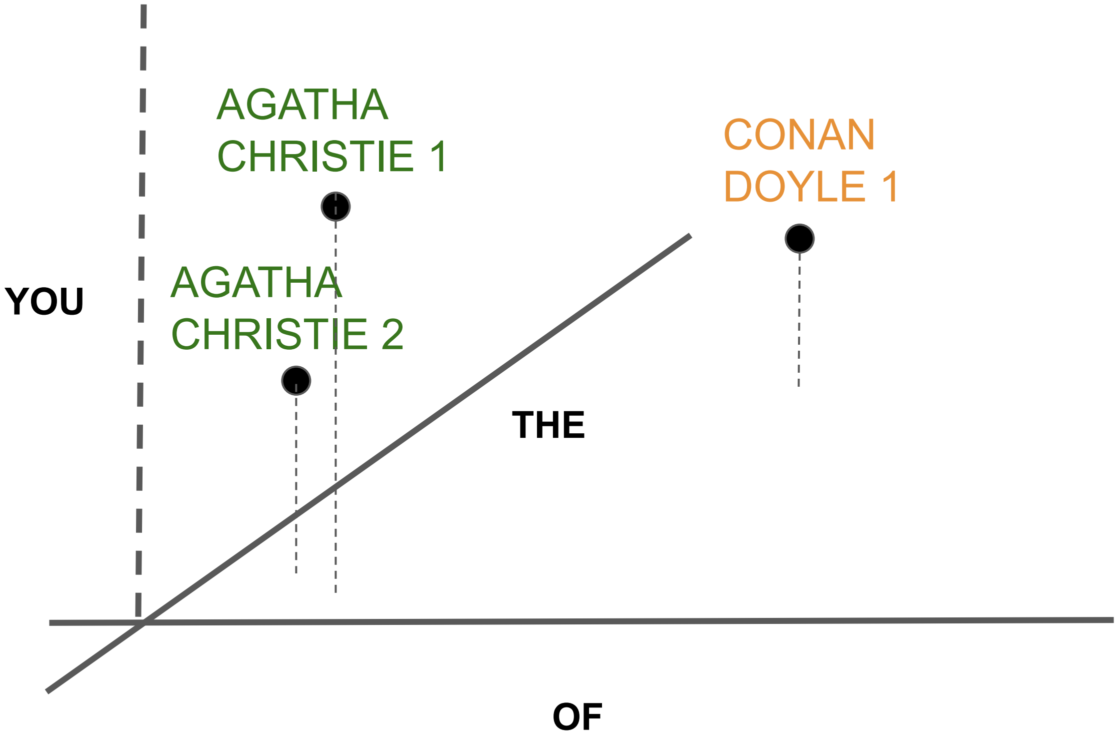
Continue thinking in terms of distances



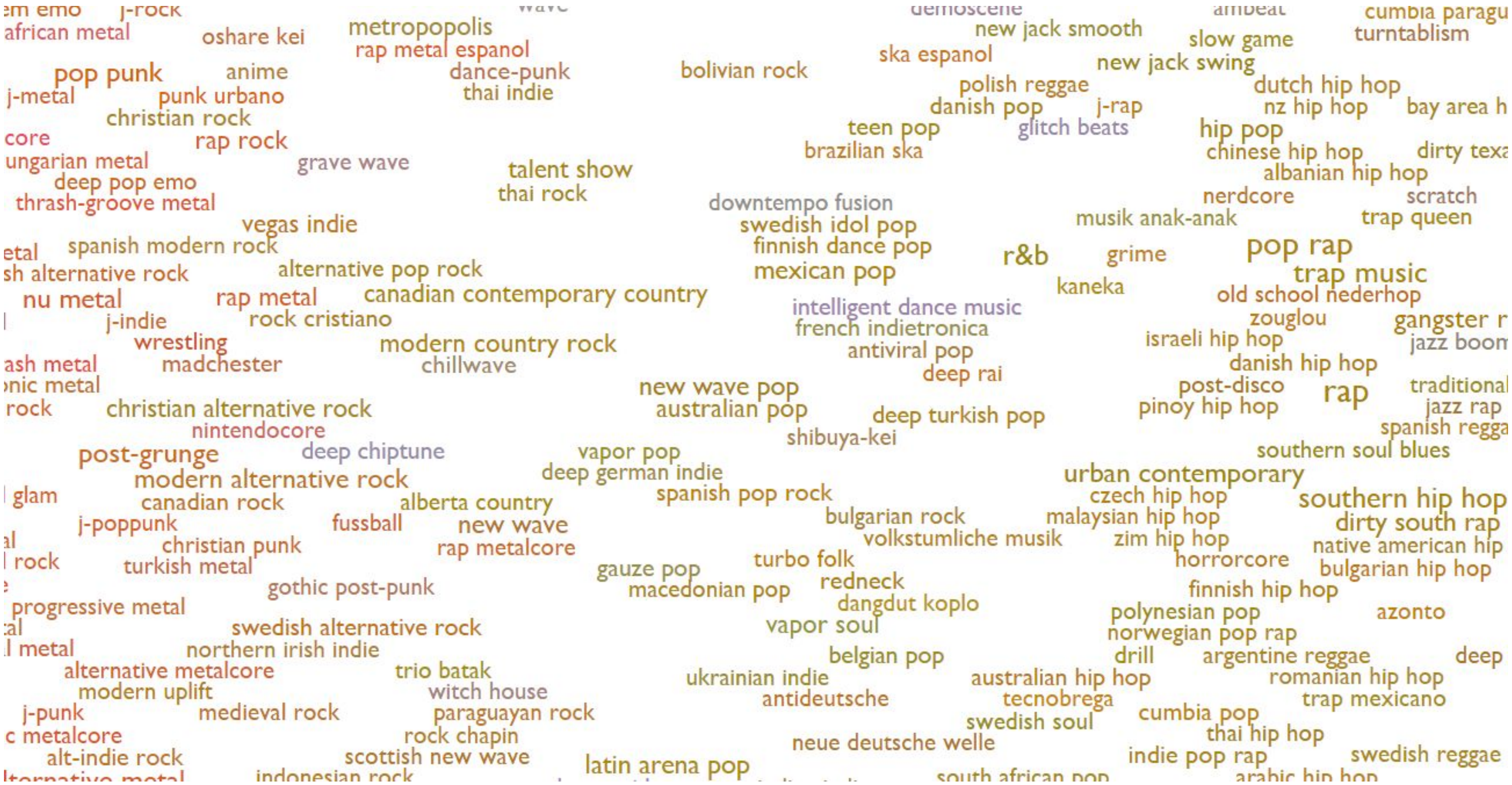
Continue thinking in terms of distances - 2D



Continue thinking in terms of distances - 3D



http://everynoise.com/



Stylometry tries to address the questions of style differences in a quantitative way

Stylometry tries to address the questions of style differences in a quantitative way

A “style” in this sense is some **set of quantifiable textual features that show unique behavior in different texts**

Stylometry tries to address the questions of style differences in a quantitative way

A “style” in this sense is some **set of quantifiable textual features that show unique behavior in different texts**

Many features were proposed: word lengths, sentence lengths, word frequencies, character n-grams, part-of-speech tags, versification features like rhythm and rhyme patterns…

Stylometry tries to address the questions of style differences in a quantitative way

A “style” in this sense is some **set of quantifiable textual features that show unique behavior in different texts**

Many features were proposed: word lengths, sentence lengths, word frequencies, character n-grams, part-of-speech tags, versification features like rhythm and rhyme patterns…

Authorship attribution is a sub-problem of stylometry: **how unique patterns of textual features help to catch author identity**

Stylometry and the 19th century positivism

- New Shakespeare Society (feminine endings in Shakespeare’s verse)
- T.C. Mendenhall (style and spectral analysis)
- Dating *Dialogues* of Plato (German school and V. Lutoslawski)

Stylometry as a quest for the “fingerprint”, “individual handwriting”.

Sounds not too much like literary analysis, right?

Because it is not: more **criminalistics** and **forensics**

.

T.C. Mendenhall (1841-1924)

The characteristic curves of composition (1887)

T.C. Mendenhall: word lengths

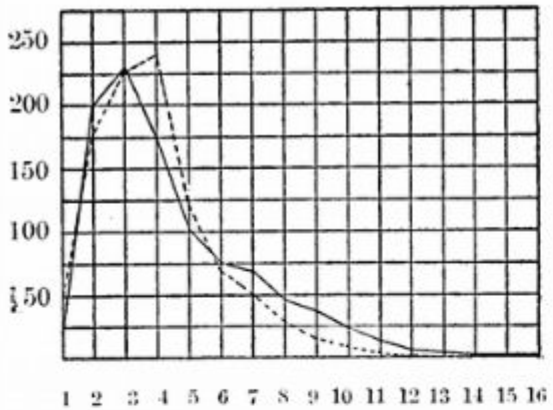


FIG. 1: Relative frequencies (per mille) of word-lengths measured by number of characters in works of W. Shakespeare (dashed) and F. Bacon (full line). Source: Mendenhall 1901: 104 (facsimile).

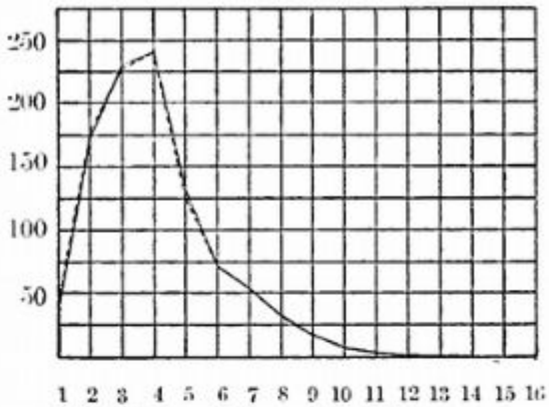
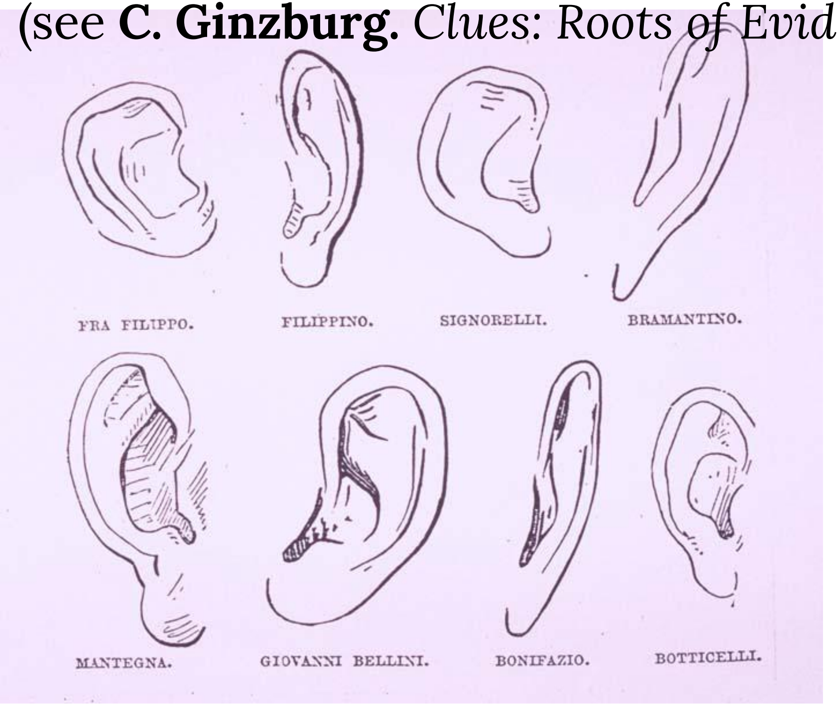
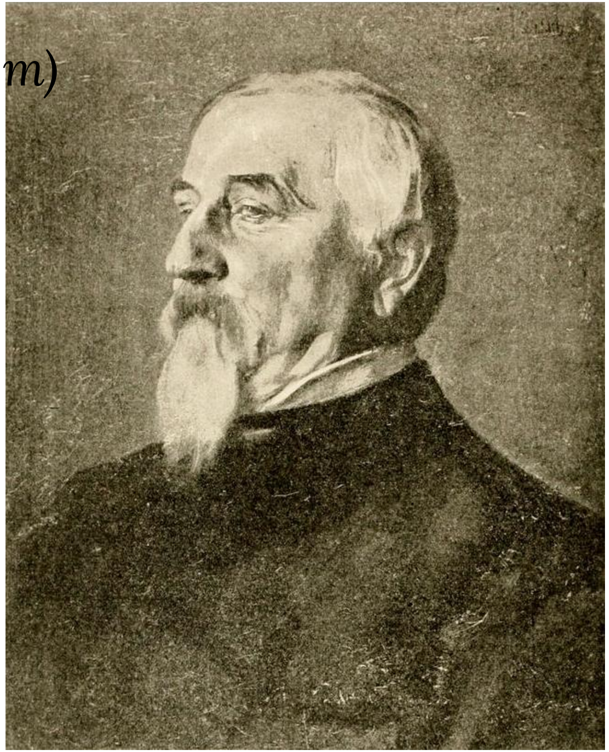


FIG. 2: Relative frequencies (per mille) of word-lengths measured by number of characters in works of W. Shakespeare (dashed) and C. Marlowe (full line). Source: Mendenhall 1901: 105 (facsimile).

Giovanni Morelli and study of authorship in paintings

(see **C. Ginzburg.** *Clues: Roots of Evidential Paradigm*)



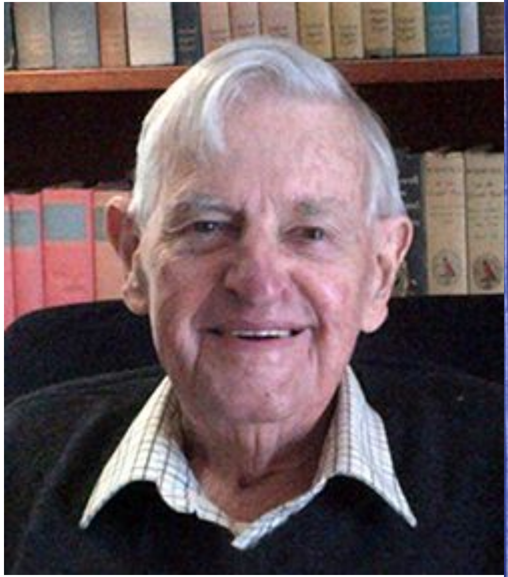


“Anthropometrics” and fingerprints

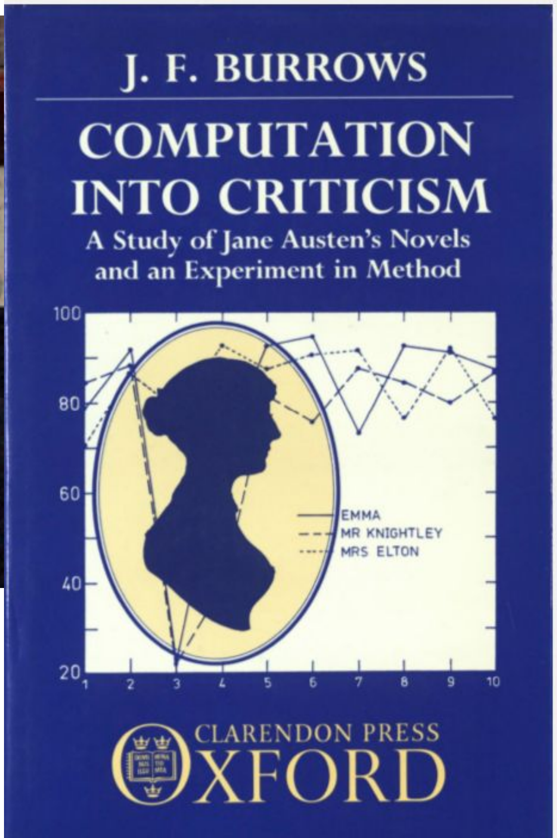
- “Anthropometrics” of Alphonse Bertillon: very prec
  on several sets of measurements.
- One dimension of measurements (e.g. height) coulc
  we combine 3 of them? 6? 12? Hundreds or even th
  modern stylometry actually do

Alphonse Bertillon (1853-1914)

Distances between texts: Burrow’s delta (2002)



John Burrows



J. F. BURROWS

COMPUTATION INTO CRITICISM

A Study of Jane Austen’s Novels and an Experiment in Method

CLARENDON PRESS
OXFORD

Delta-distance

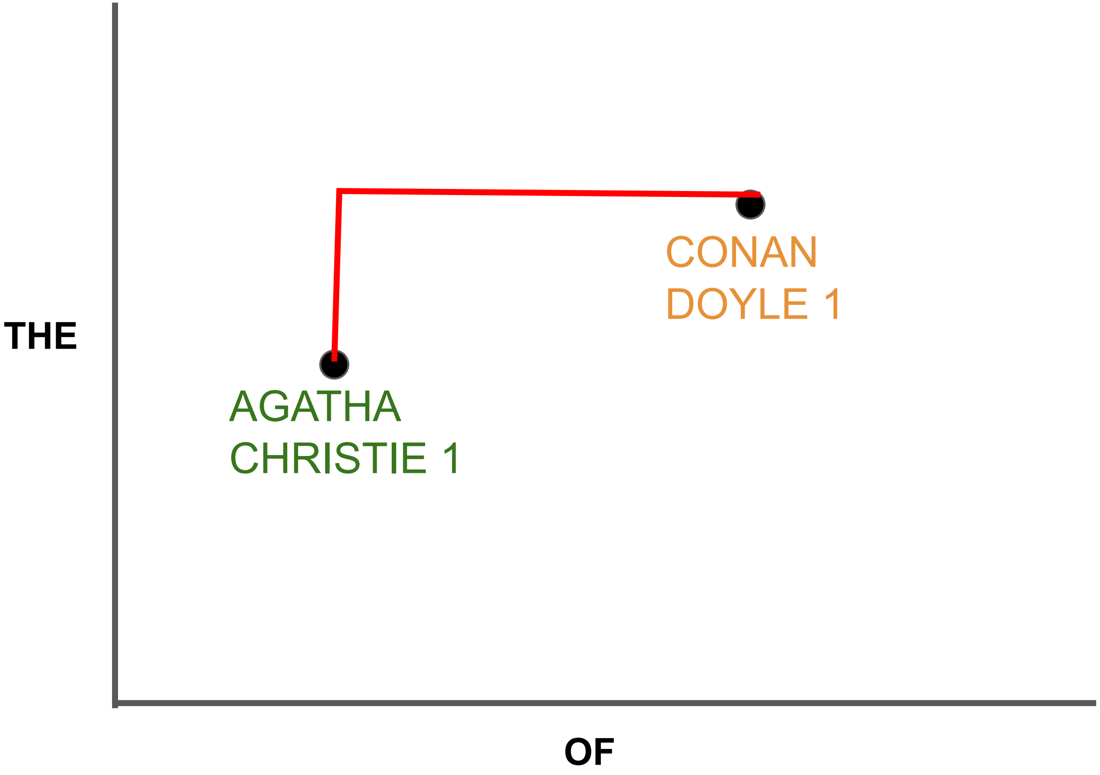
$$\Delta = \sum_{i=1}^{n} \frac{|z(x_i) - z(y_i)|}{n}$$

On the distances

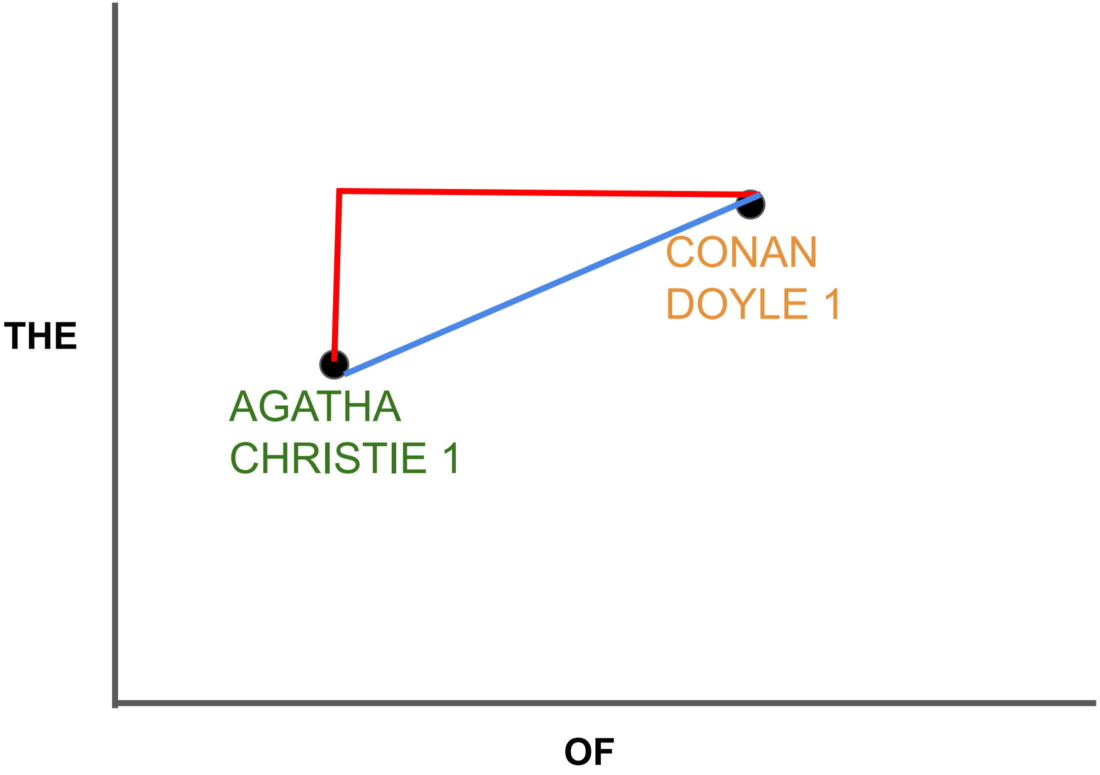
http://versologie.cz/talks/2017chicago/

Petr Plecháč (Czech Academy of Sciences)

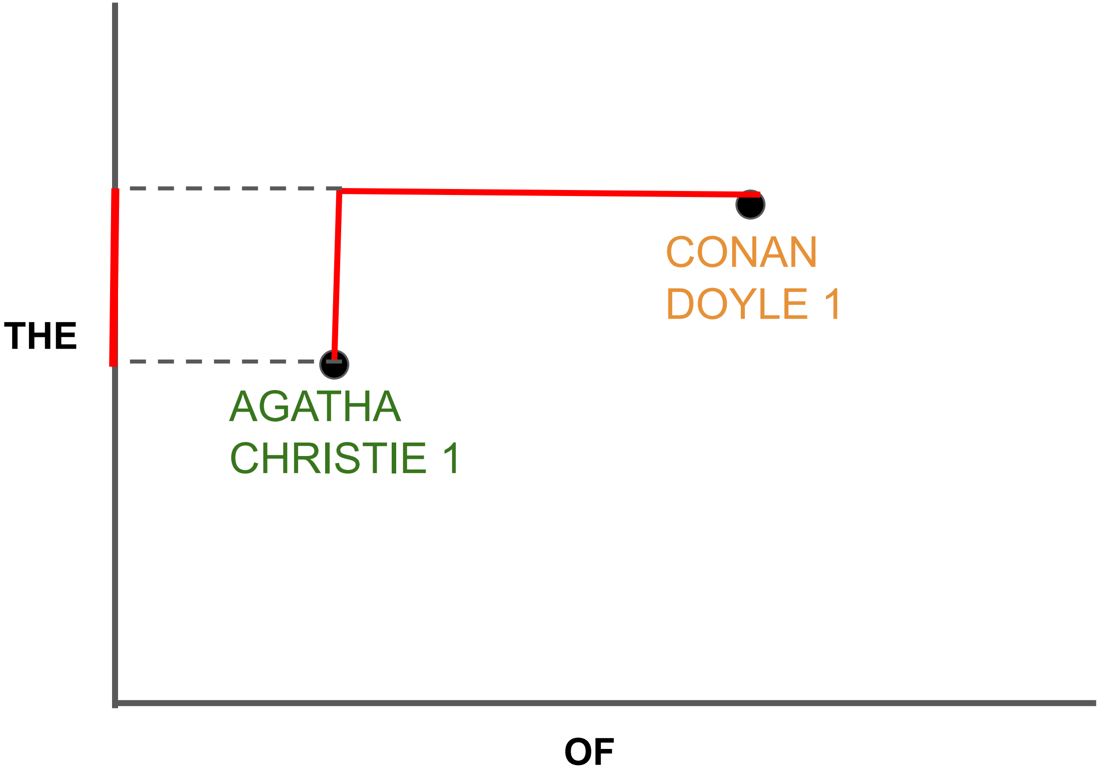
Manhattan (also Delta)



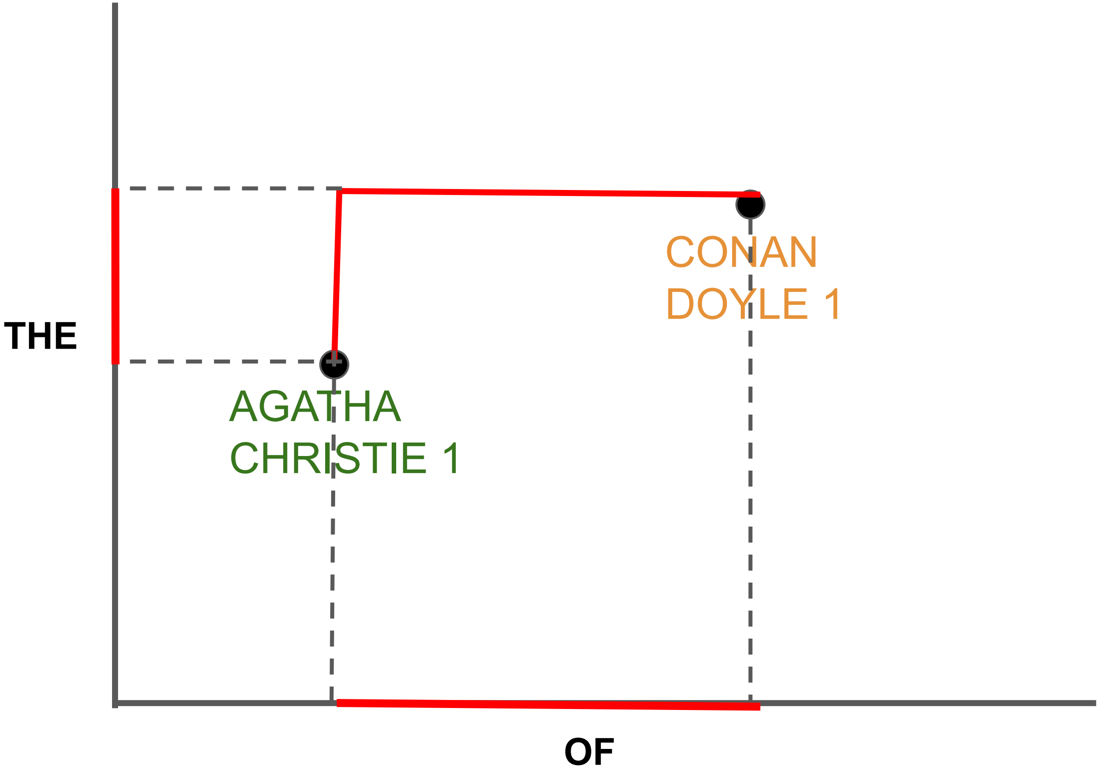
Manhattan + Euclidean

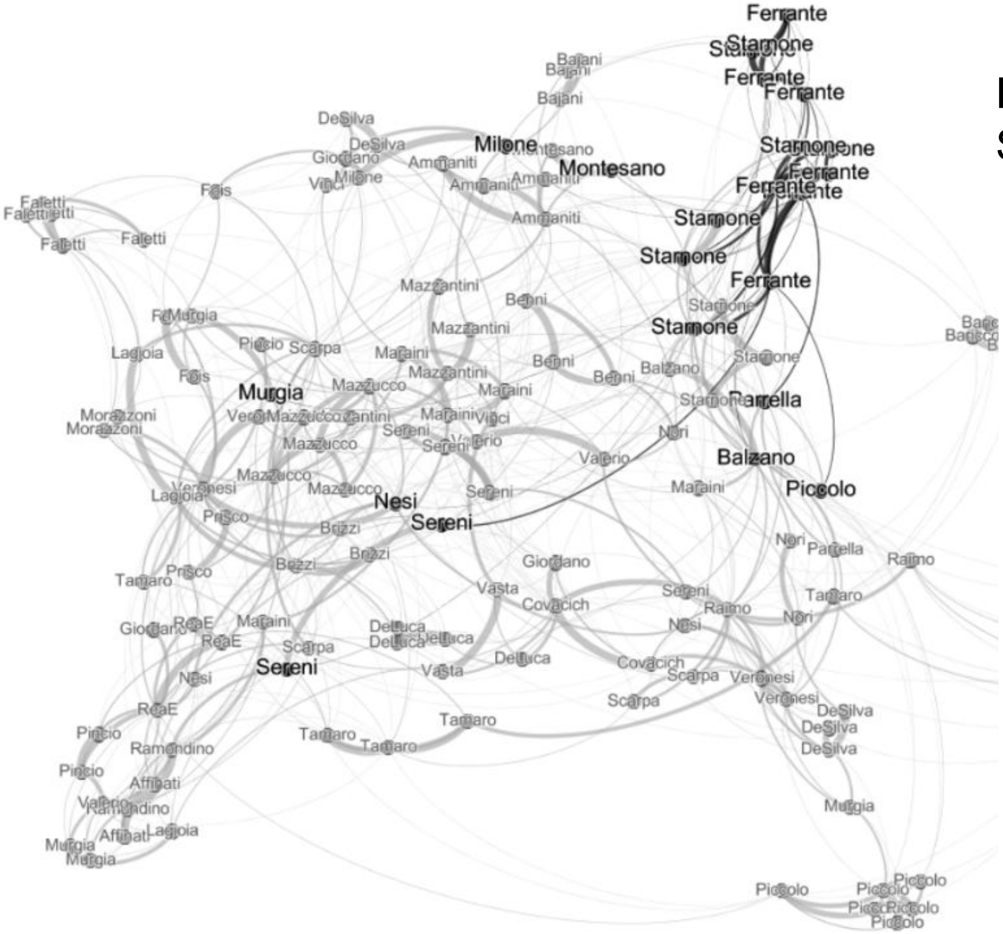


Manhattan (also Delta)



Manhattan (also Delta)





Elena Ferrante / Anita Raja / Domenico Starnone?

- Bootstrap Consensus Network based on Delta measure. 3 nearest neighbours of each novel, connected
- 150 Italian novels late 20th - early 21st century
- From 100 to 1000 MFWs (bootstrap, increment 100)
- List of closest candidates (Starnone the closest)

Eder 2018

“stylo” package for R

Eder, M., Rybicki, J. and Kestemont, M. (2016).
Stylometry with R: a package for computational text analysis. R Journal 8(1): 107-121.

Delta-distance

$$\Delta = \sum_{i=1}^{n} \frac{|z(x_i) - z(y_i)|}{n}$$

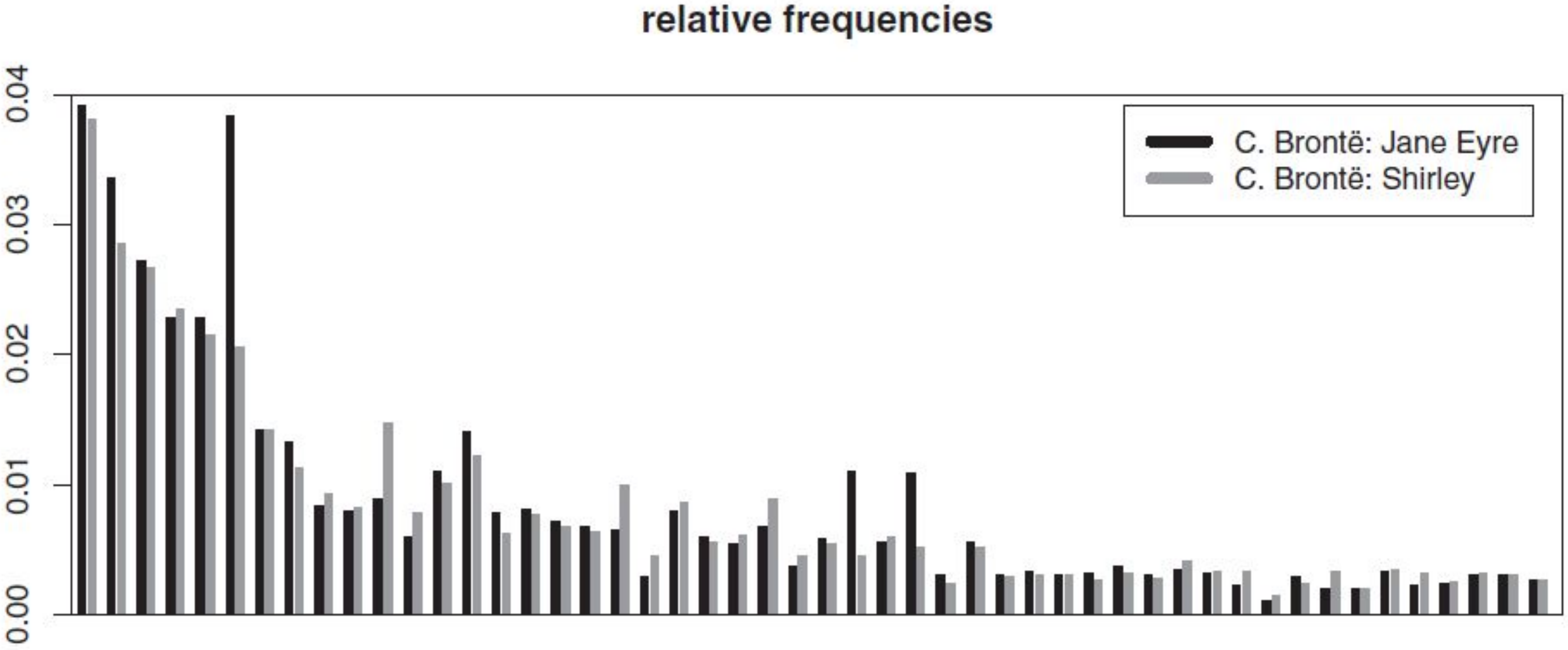
Normalization

- To overcome this effect a scaling of values is performed. We compute standardised **z-scores** for relative frequencies

$$z(x_i) = \frac{x_i - \mu(x)}{\sigma(x)}$$

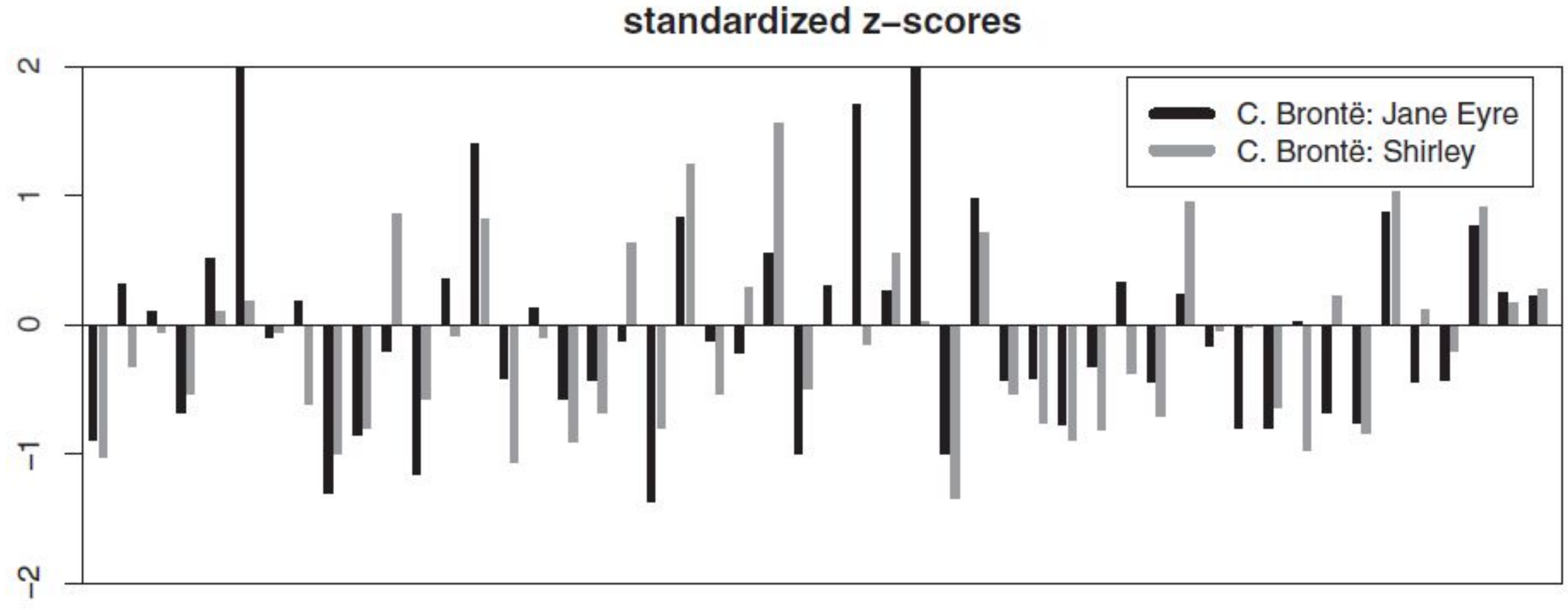
- ...which means z-score of some frequency = this frequency **minus** mean frequency of all observations (μ) **and divide** by standard deviation

Relative frequencies curve (Zipf's curse)



Evert et al. 2017

Normalized frequencies

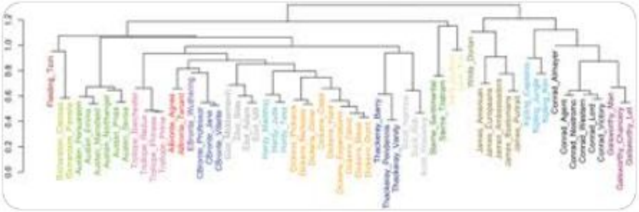


Evert et al. 2017

some deep stats!!

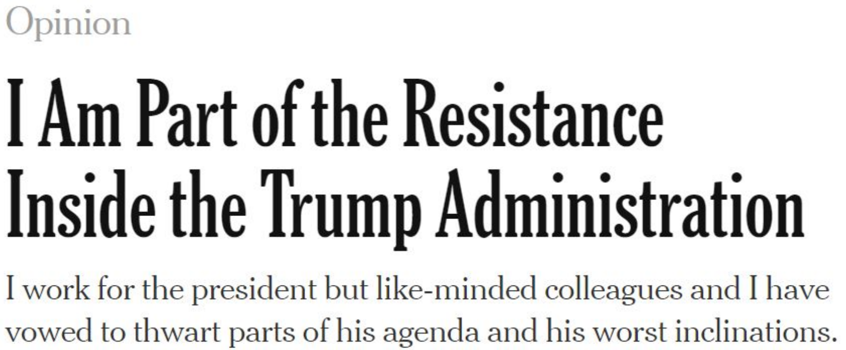
WikiLeaks ✓
@wikileaks

Based upon our statistical analysis of t
language used in the New York Times
anonymous Op Ed, the author is likely to be
an older (58%), conservative (92%) male (66-
87%). Sources should protect themselves by
consulting "adverserial stylometry" and
"forensic author profiling".



6:49 PM - 6 Sep 2018

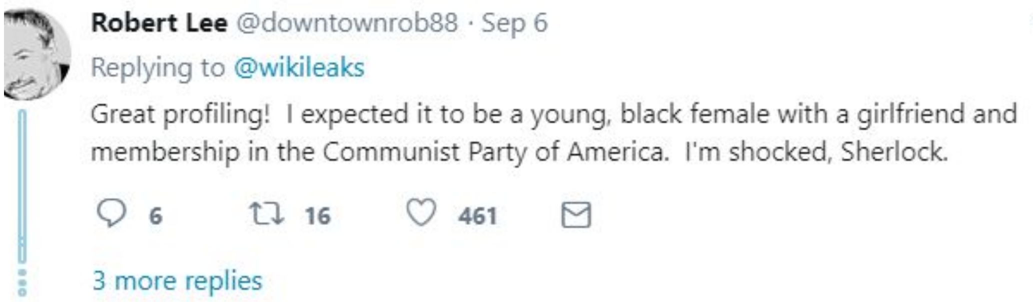
2,095 Retweets **3,889** Likes



Opinion

I Am Part of the Resistance Inside the Trump Administration

I work for the president but like-minded colleagues and I have vowed to thwart parts of his agenda and his worst inclinations.



Robert Lee @downtownrob88 · Sep 6

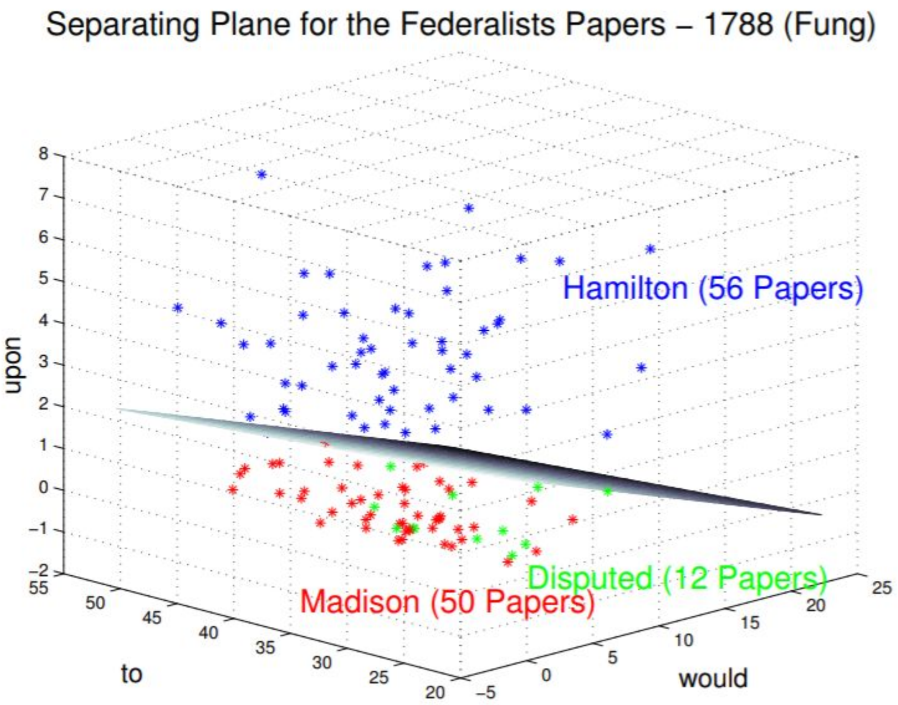
Replying to @wikileaks

Great profiling! I expected it to be a young, black female with a girlfriend and membership in the Communist Party of America. I'm shocked, Sherlock.

6 16 461

3 more replies

12 disputed Federalists papers

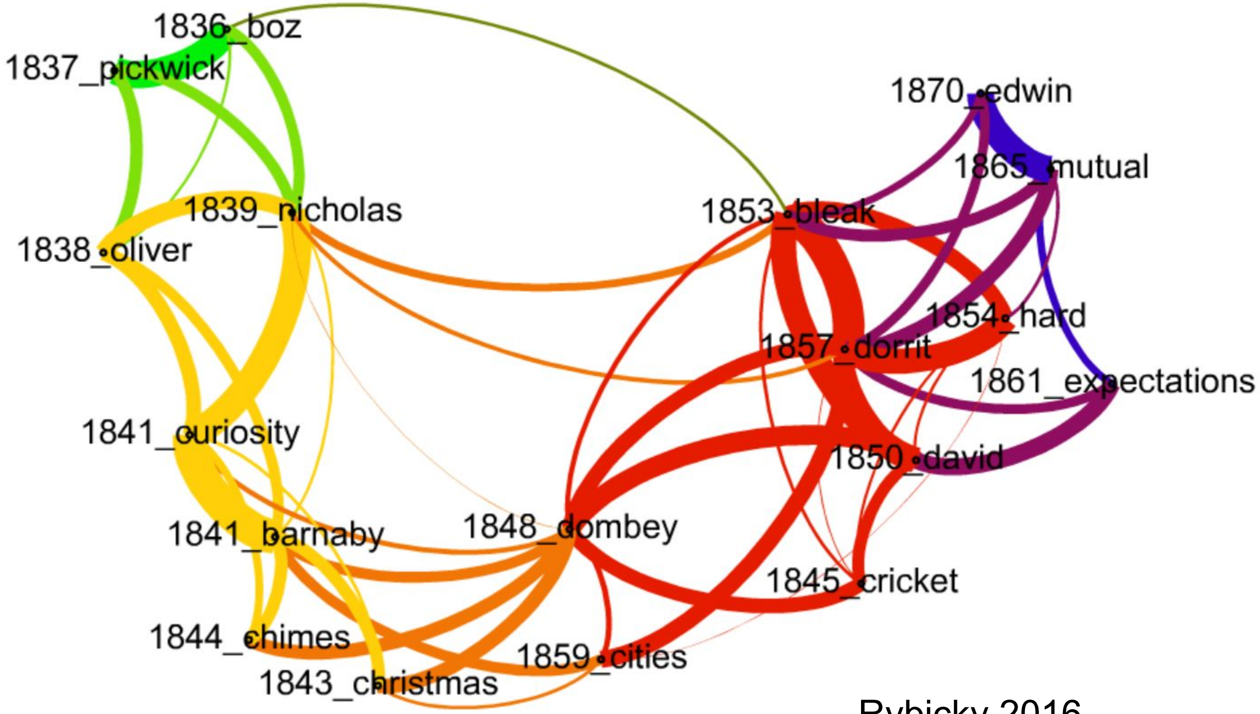


Glenn (2003). The disputed federalist papers: SVM feature selection via concave minimization

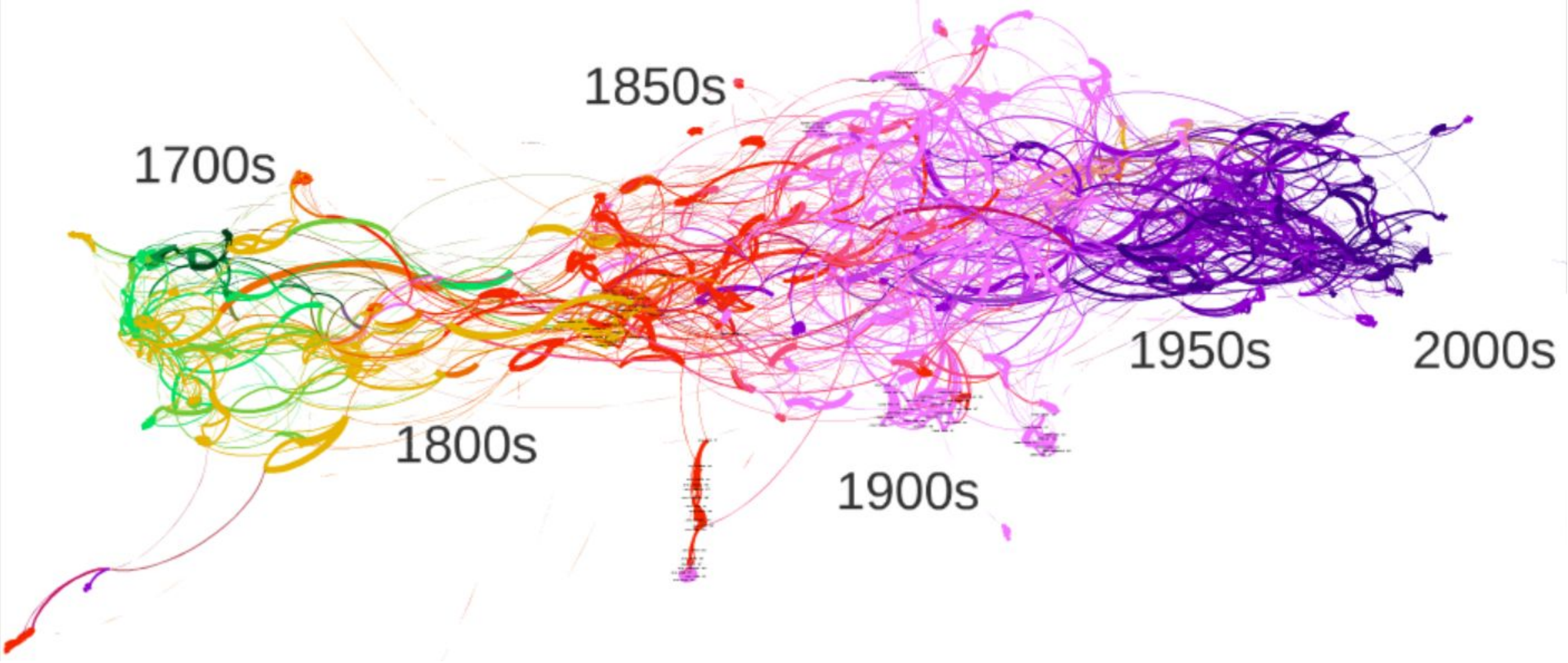
Fig. 3. Obtained Hyperplane in 3 dimensions

- Language provide enough variability for largely unconscious and unique patterns of writing
- We basically are constantly “emanating” our fingerprint and not only through natural languages
- Chat rooms (Inches et al. 2013), programming code, even texts written in artificial languages (elven, dothraki, klingon) exhibit strong-enough authorial signal to figure out probable authors
- Beware of “forensic authorial profiling”!

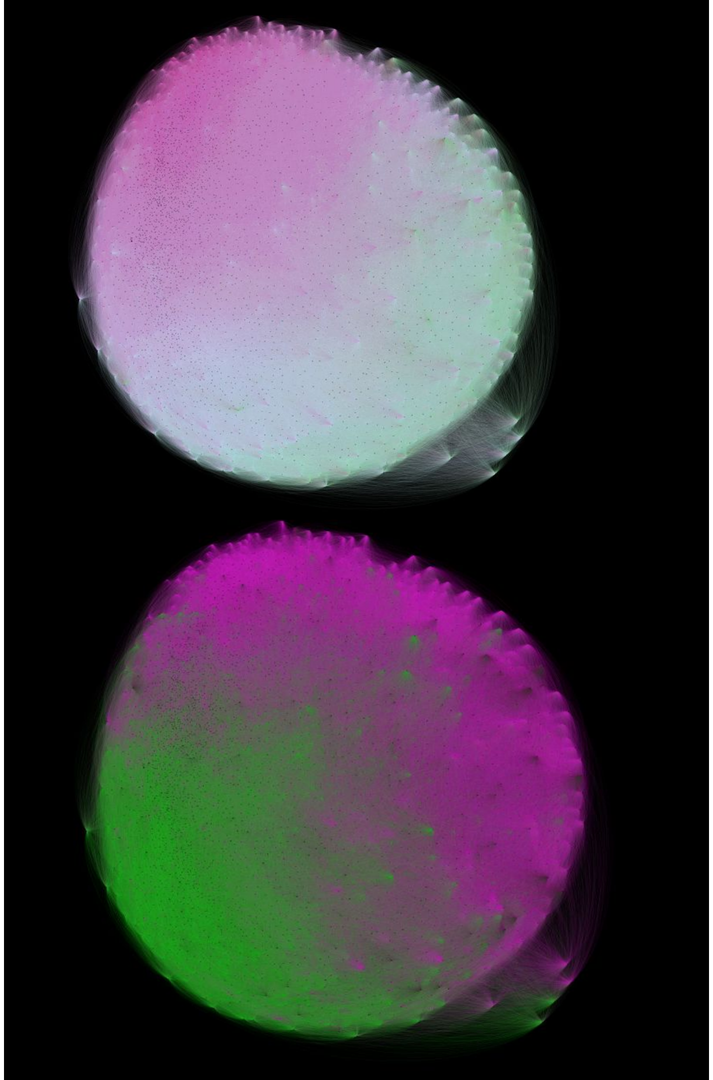
What lexical frequencies signal us?



Rybicky 2016



Rybicky 2016



Global slow changes?

- 1 m. connections of novels, based on distance measurements using most frequent words & topics (themes)
- Colors: 1) time, 2) gender
- Jockers *Macroanalysis* (2013)
- plots redone in 2019

Resources (corpora)

From stylo guys: https://computationalstylistics.github.io/resources/

English, Latin, Russian, Italian, Shakespeare etc.

.txtLab **Novel450** dataset: https://txtlab.org/data-sets/

150 ger, 150 fr, 150 eng