STYLOMETRY

…in detail

Theoretical implications

“It has been noted that the switch from content words to function words in authorship attribution studies has an interesting historic parallel in art-historic research. […] Giovanni Morelli (1816-1891) was among the first to suggest that the attribution of, for instance, a *Quattrocento* painting to some Italian master, could not happen based on ‘content’ […] Morelli thought it better to restrict an authorship analysis to discrete details such as ears, hands and feet: such fairly functional elements are naturally very frequent in nearly all paintings, because they are to some extent content-independent. […] the argument is often raised that the use of these [function] words would not be under an author’s conscious control during the writing process.”
(Kestemont, 2014)

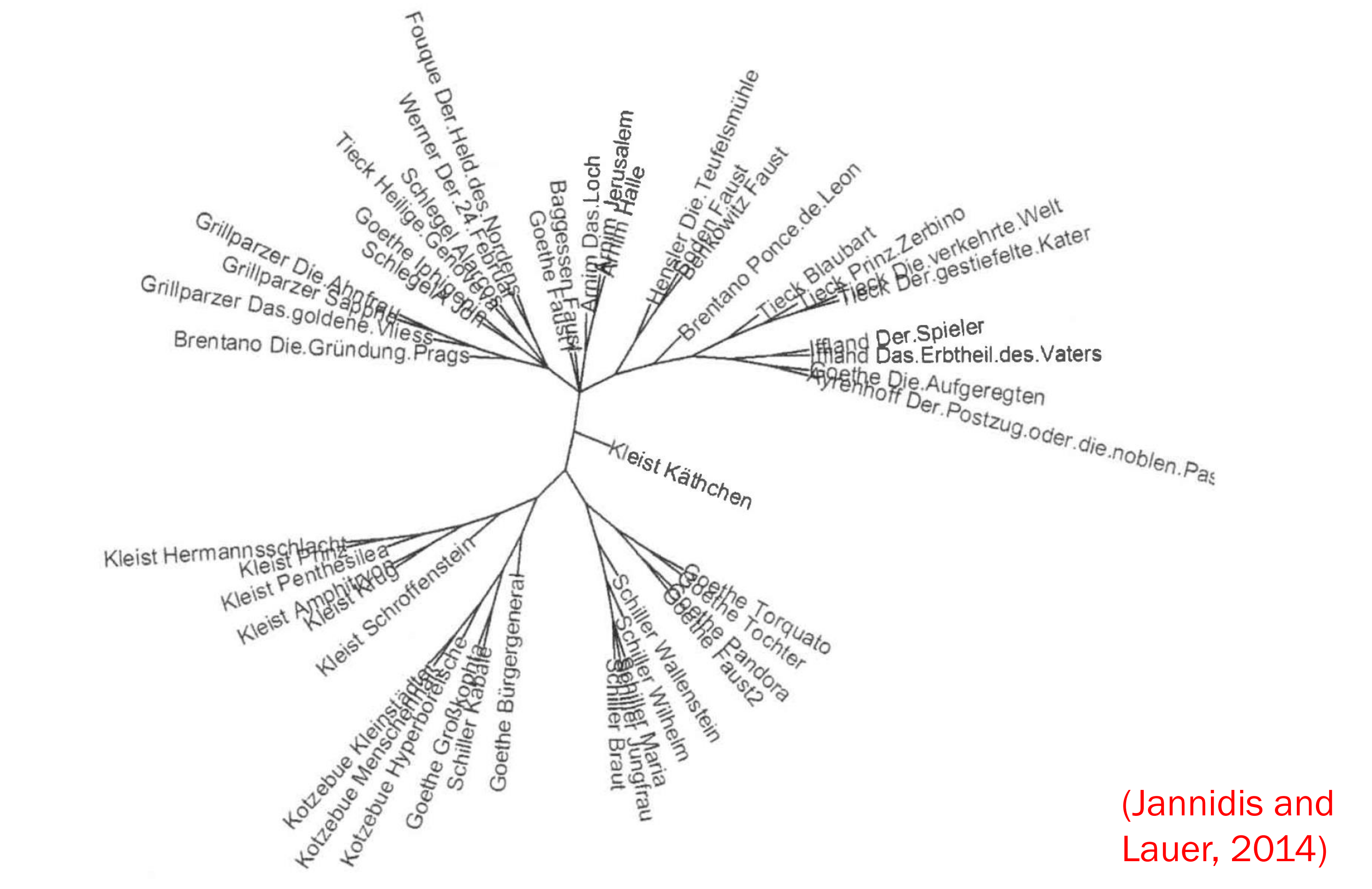
Theoretical Implications

"Style is a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively."

- style […] should be seen as a complex system, with features situated at different linguistic levels
- we conceive of stylistic features as explicitly defined and clearly identifiable.
- a certain style can be described using methods based on computing frequencies, relations, and distributions of features and relevant statistics [quantitative], as well as methods based on precise observation and description of individual occurrences [qualitative]

(Herrmann et al. 2015)

Applications…



(Jannidis and Lauer, 2014)

Applications…

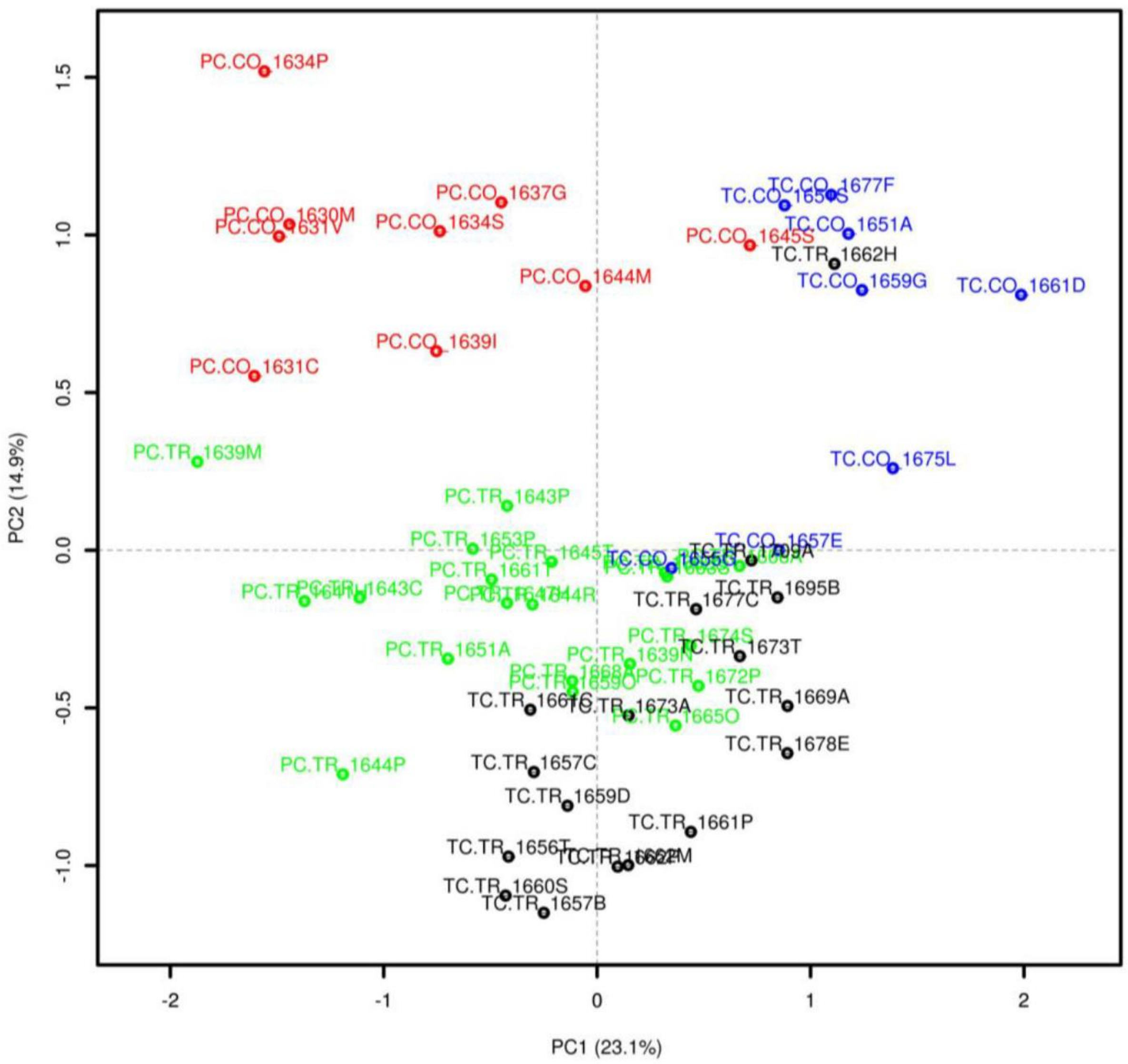
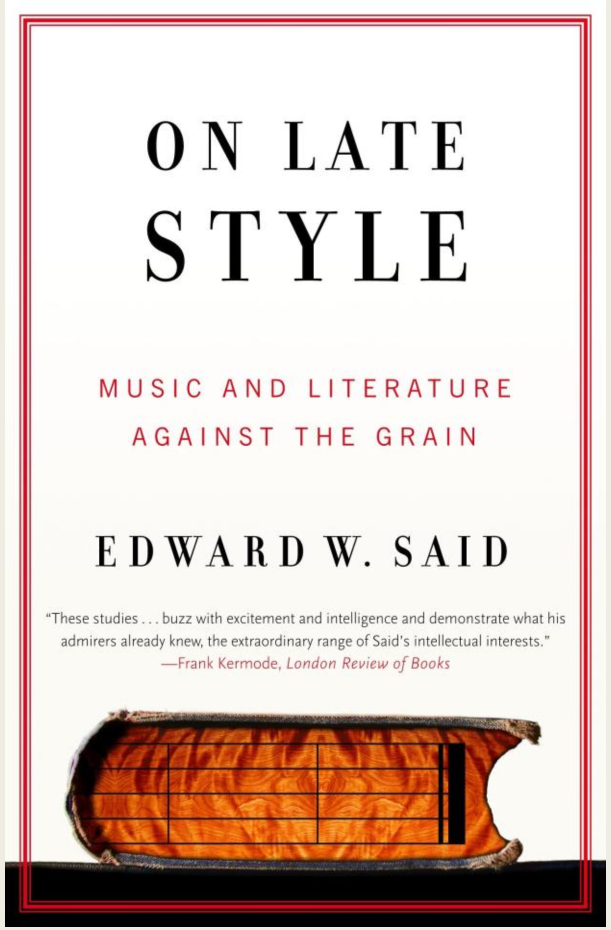


Abb. 8: Principal Component Analysis
(Kürzel: PC = Pierre Corneille, TC = Thomas Corneille, CO = comédies, TR = tragédies)

(Schöch, 2014)

Applications…



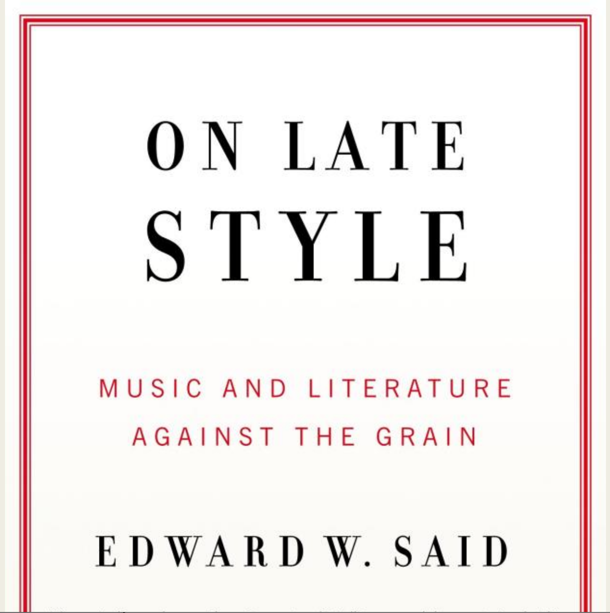
ON LATE STYLE

MUSIC AND LITERATURE AGAINST THE GRAIN

EDWARD W. SAID

"These studies . . . buzz with excitement and intelligence and demonstrate what his admirers already knew, the extraordinary range of Said's intellectual interests."
—Frank Kermode, *London Review of Books*

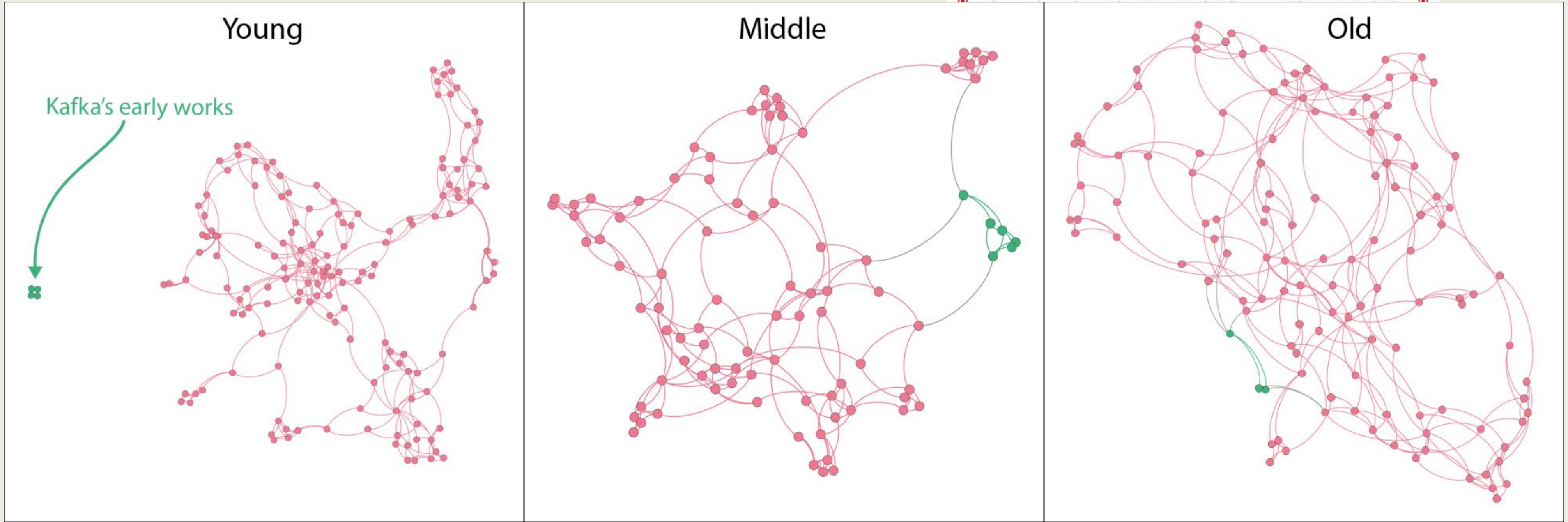
Applications…



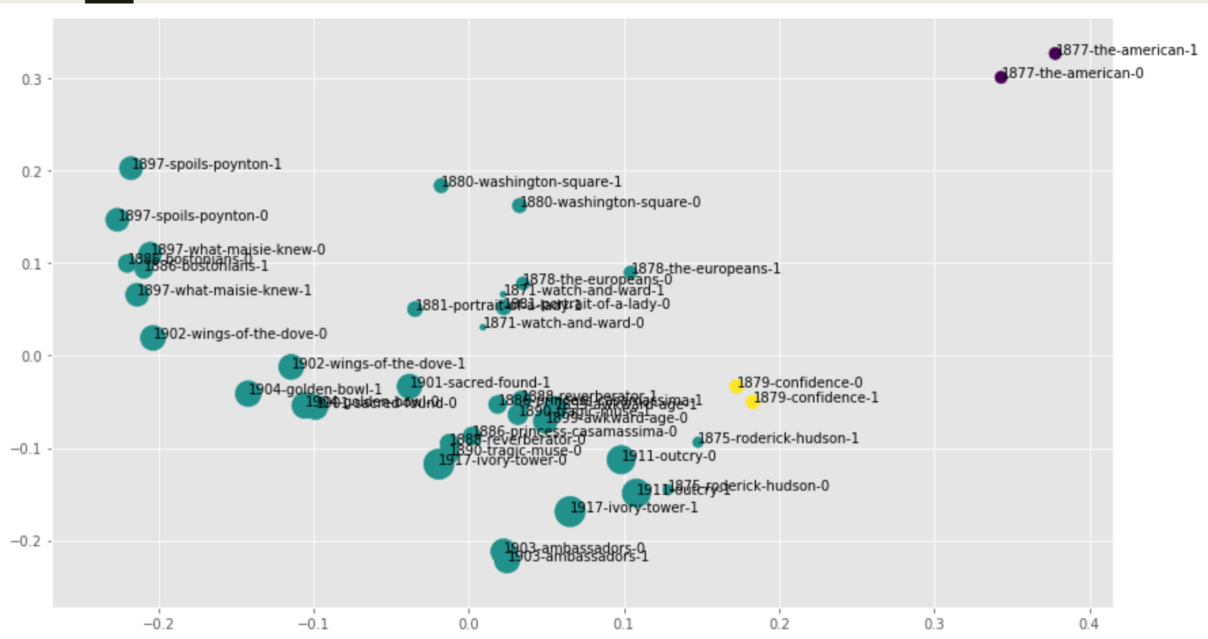
ON LATE STYLE

MUSIC AND LITERATURE AGAINST THE GRAIN

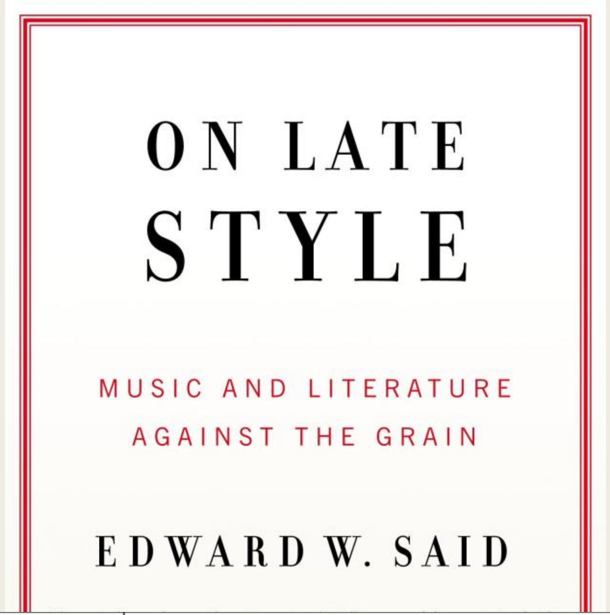
EDWARD W. SAID



(Rebora and Salgaro, 2018)



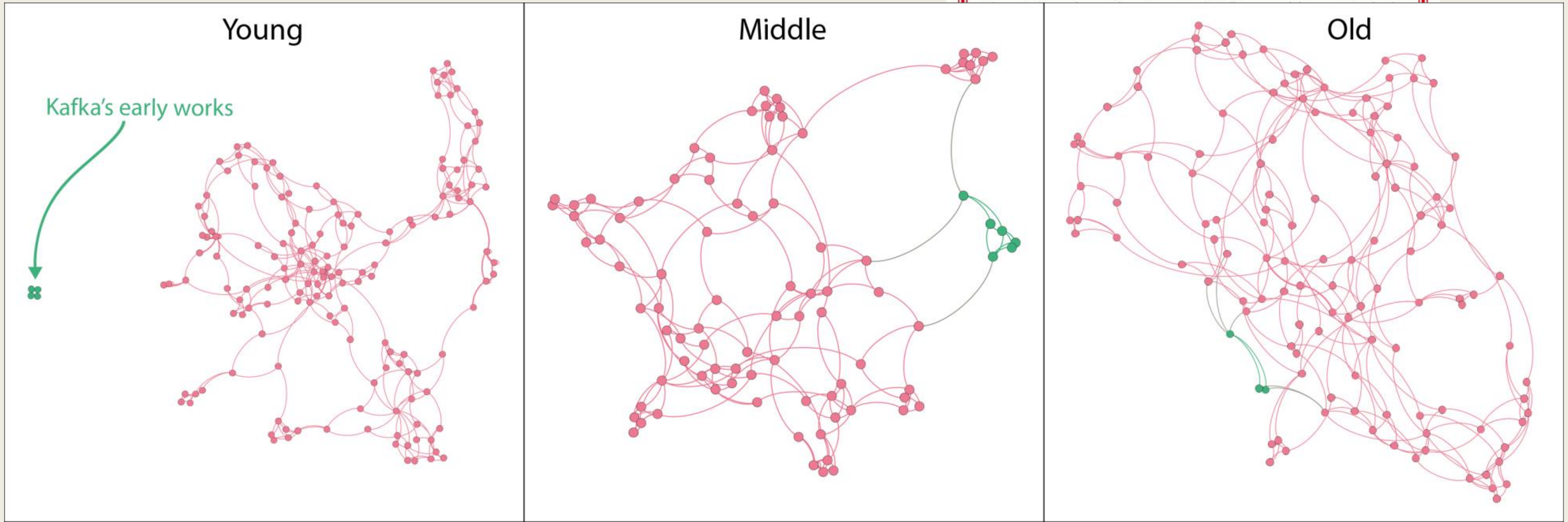
(Reeve, 2018)



ON LATE STYLE

MUSIC AND LITERATURE AGAINST THE GRAIN

EDWARD W. SAID

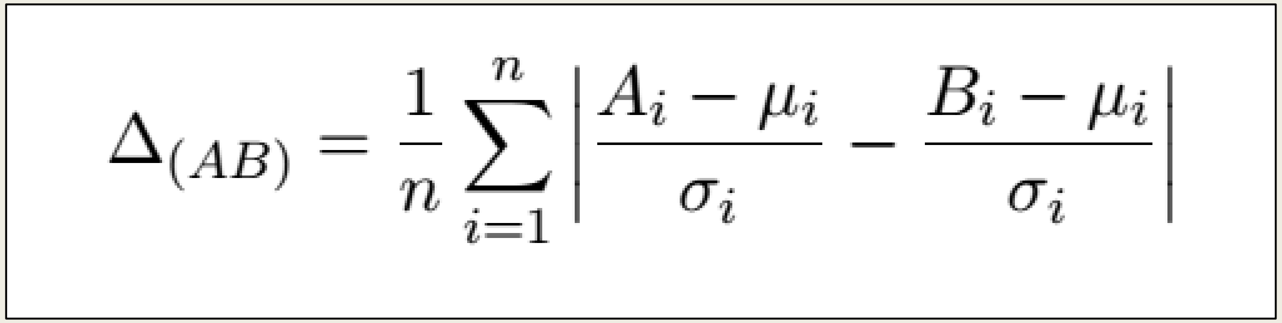


(Rebora and Salgaro, 2018)

A bit of mathematics…

The (many) distance measures

The (many) distance measures



$$\Delta_{(AB)} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - \mu_i}{\sigma_i} - \frac{B_i - \mu_i}{\sigma_i}\right|$$

The (many) distance measures

$$\Delta_{(AB)} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - \mu_i}{\sigma_i} - \frac{B_i - \mu_i}{\sigma_i}\right|$$

$$\delta_{(AB)} = \sum_{i=1}^{n}|A_i - B_i|$$

The (many) distance measures

$$\Delta_{(AB)} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - \mu_i}{\sigma_i} - \frac{B_i - \mu_i}{\sigma_i}\right|$$

$$\delta_{(AB)} = \sum_{i=1}^{n}|A_i - B_i|$$

$$\delta_{(AB)} = \sqrt{\sum_{i=1}^{n}|(A_i)^2 - (B_i)^2|}$$

The (many) distance measures

$$\Delta_{(AB)} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - \mu_i}{\sigma_i} - \frac{B_i - \mu_i}{\sigma_i}\right|$$

$$\delta_{(AB)} = \sum_{i=1}^{n}|A_i - B_i|$$

$$\Delta_{(AB)} = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\left|\frac{(A_i)^2 - (B_i)^2}{\sigma_i}\right|}$$

$$\delta_{(AB)} = \sqrt{\sum_{i=1}^{n}|(A_i)^2 - (B_i)^2|}$$

The (many) distance measures

$$\Delta_{(AB)} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - \mu_i}{\sigma_i} - \frac{B_i - \mu_i}{\sigma_i}\right|$$

$$\delta_{(AB)} = \sum_{i=1}^{n}|A_i - B_i|$$

$$\Delta_{(AB)} = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\left|\frac{(A_i)^2 - (B_i)^2}{\sigma_i}\right|}$$

$$\delta_{(AB)} = \sqrt{\sum_{i=1}^{n}|(A_i)^2 - (B_i)^2|}$$

$$\cos\Delta = \frac{Z_A \cdot Z_B}{||Z_A|| \cdot ||Z_B||}$$

The (many) distance measures

$$\Delta_{(AB)} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i-\mu_i}{\sigma_i}-\frac{B_i-\mu_i}{\sigma_i}\right|$$

$$\delta_{(AB)} = \sum_{i=1}^{n}|A_i - B_i|$$

$$\Delta_{(AB)} = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\left|\frac{(A_i)^2-(B_i)^2}{\sigma_i}\right|}$$

$$\delta_{(AB)} = \sqrt{\sum_{i=1}^{n}|(A_i)^2-(B_i)^2|}$$

$$\cos\Delta = \frac{Z_A \cdot Z_B}{||Z_A||\cdot||Z_B||}$$

Delta (and Cosine) Distance

Delta (and Cosine) Distance

- Based on z-scores:

$$z = \frac{x - \mu}{\sigma}$$

Delta (and Cosine) Distance

- Based on z-scores:

$$z = \frac{x - \mu}{\sigma}$$

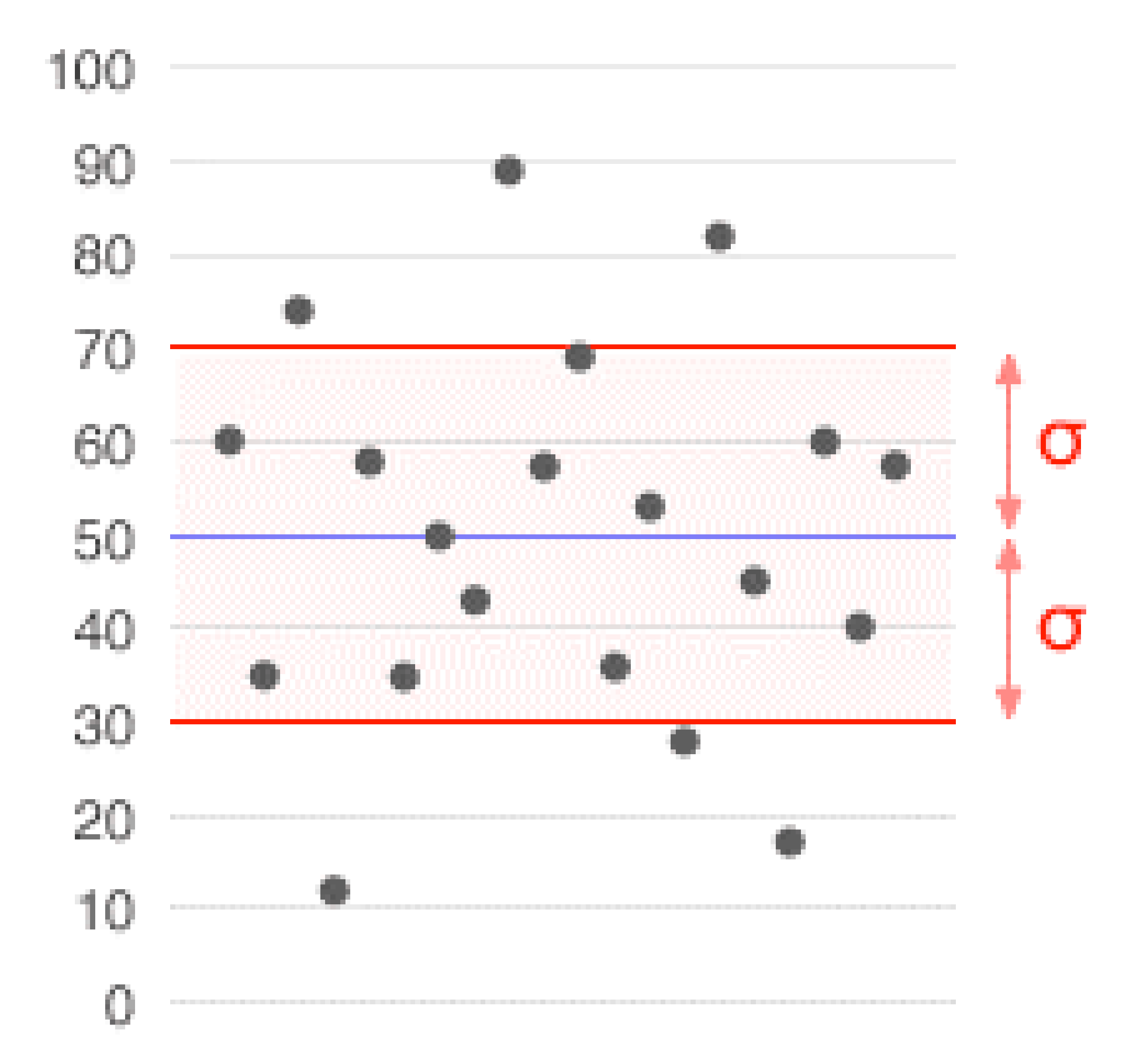
x is a number in a series
μ is the mean of the series
σ is the «standard deviation»

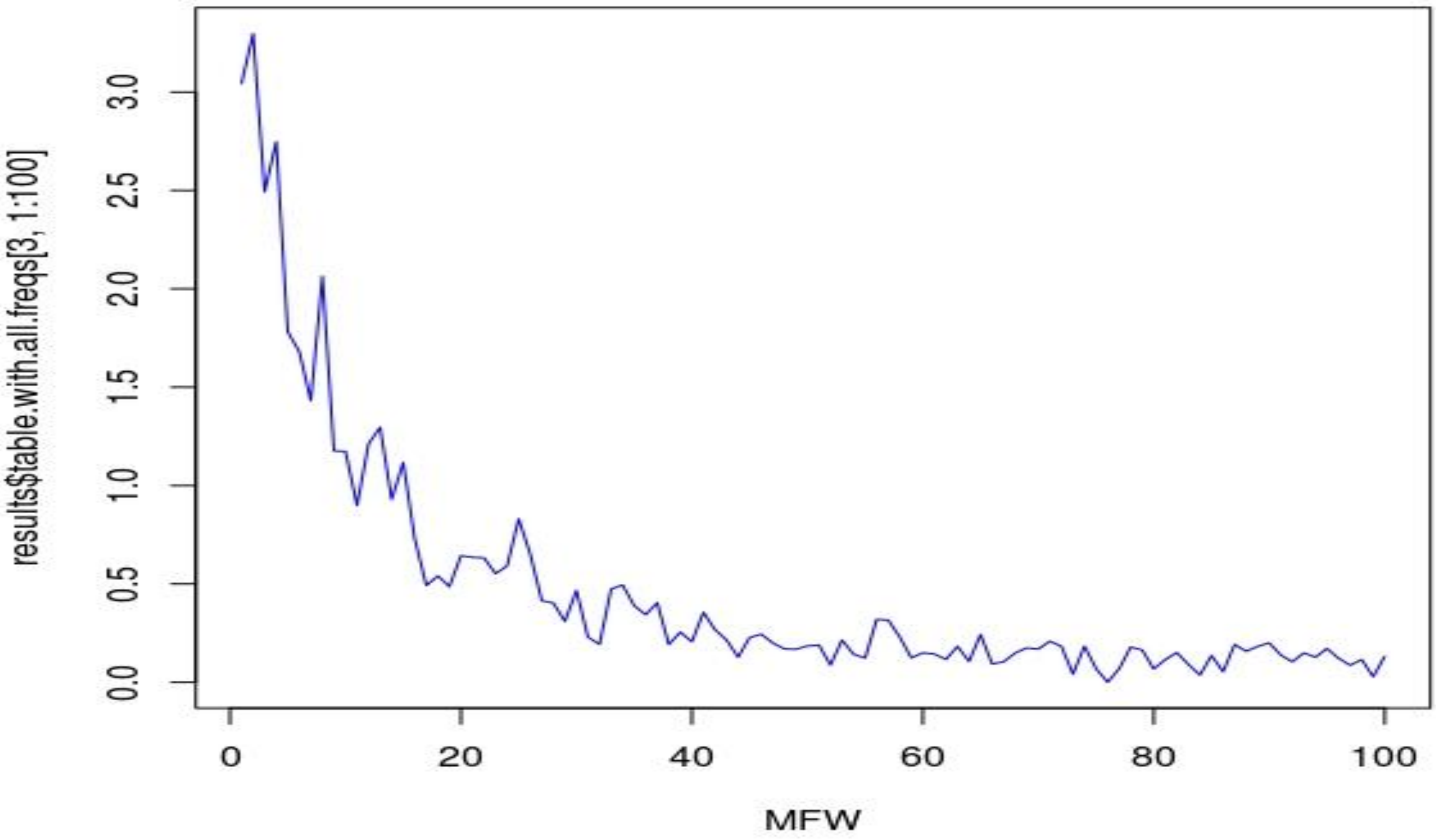
Delta (and Cosine) Distance

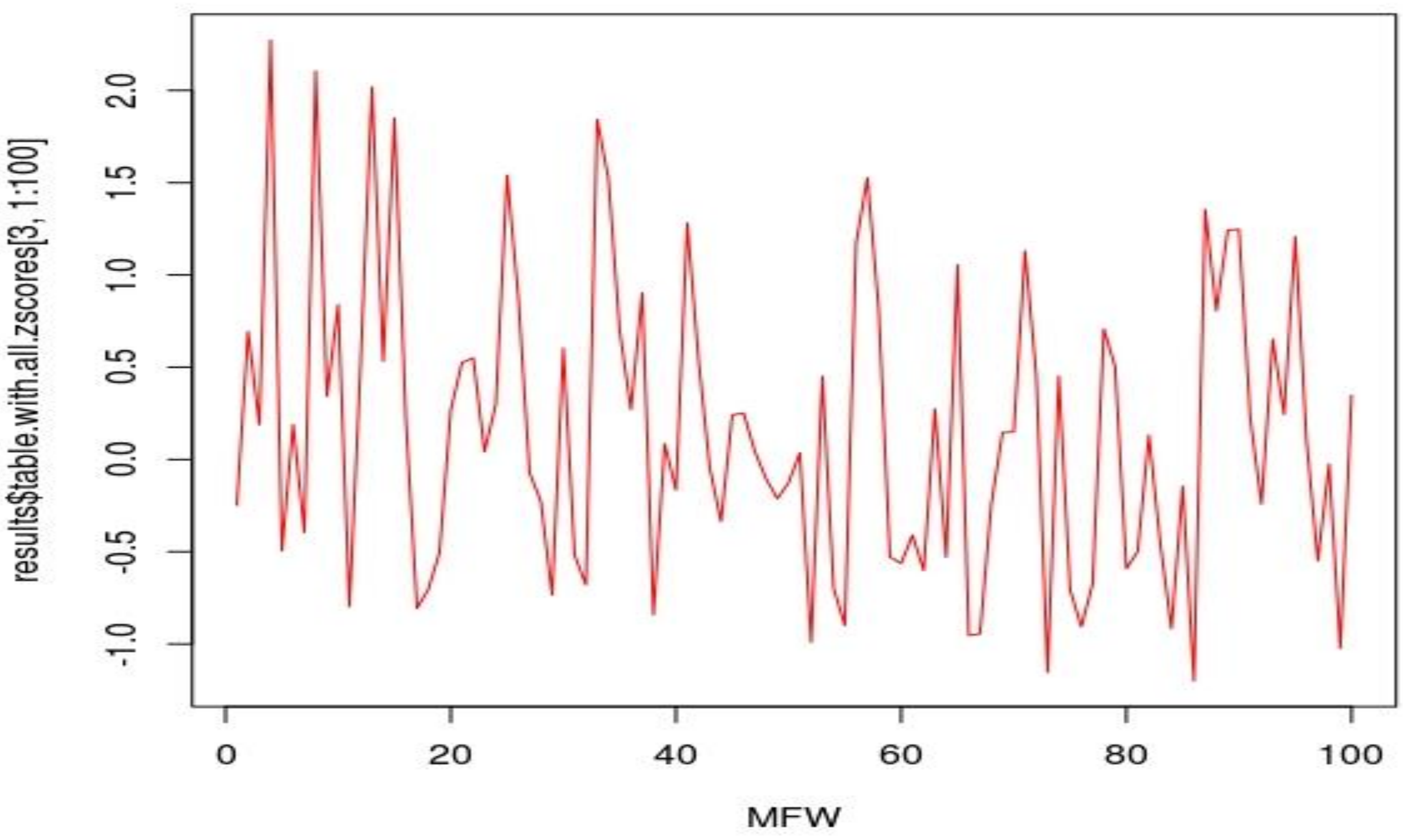
- Based on z-scores:

$$z = \frac{x - \mu}{\sigma}$$

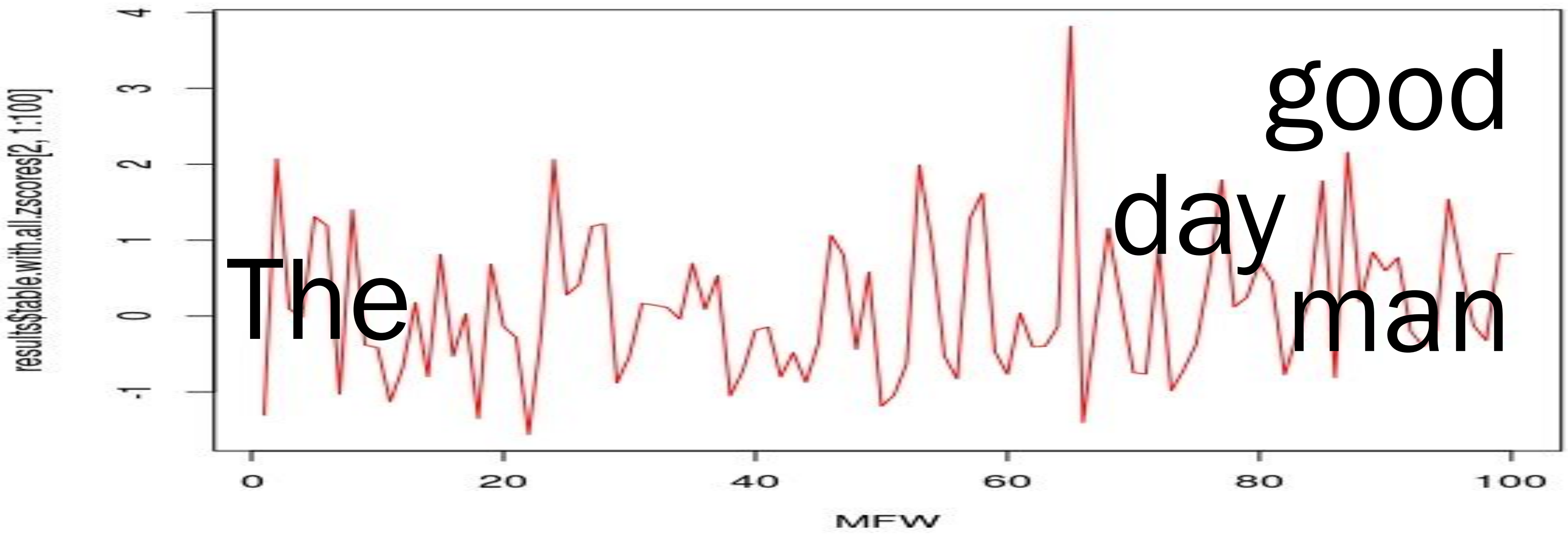
x is a number in a series
μ is the mean of the series
σ is the «standard deviation»




results$table.with.all.freqs[3, 1:100]
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
20
40
60
80
100
MFW


results$table.with.all.zscores[3, 1:100]
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
0
20
40
60
80
100
MFW

Zeta



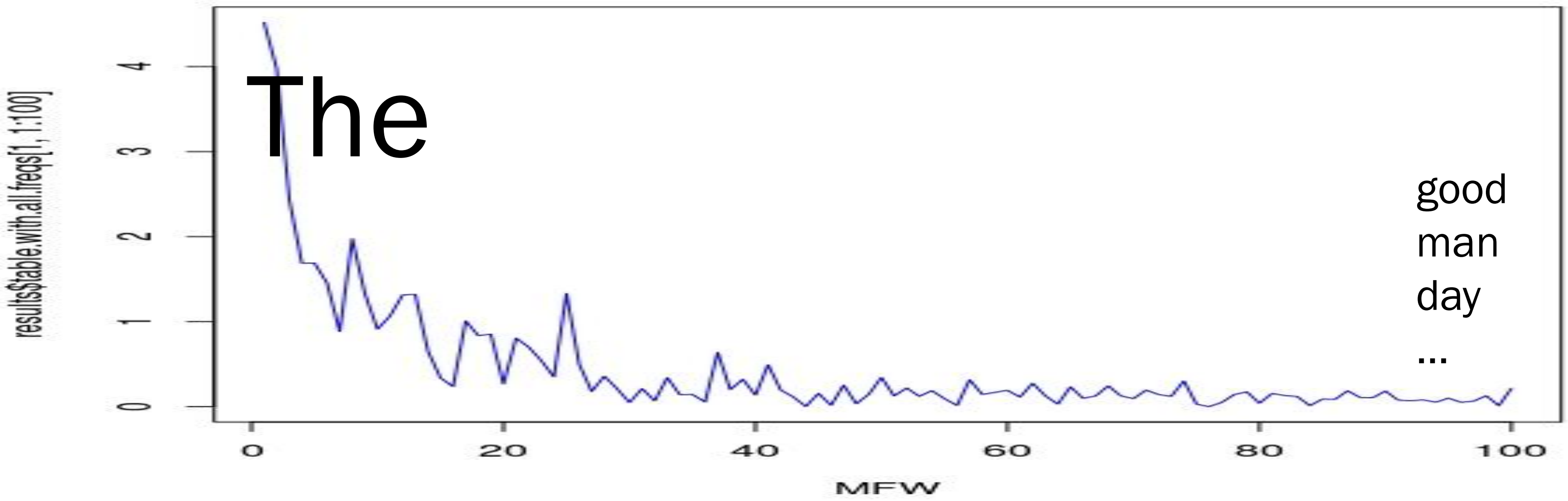
The

good

day

man

Not-Zeta



The

good
man
day
…

How the distance is calculated

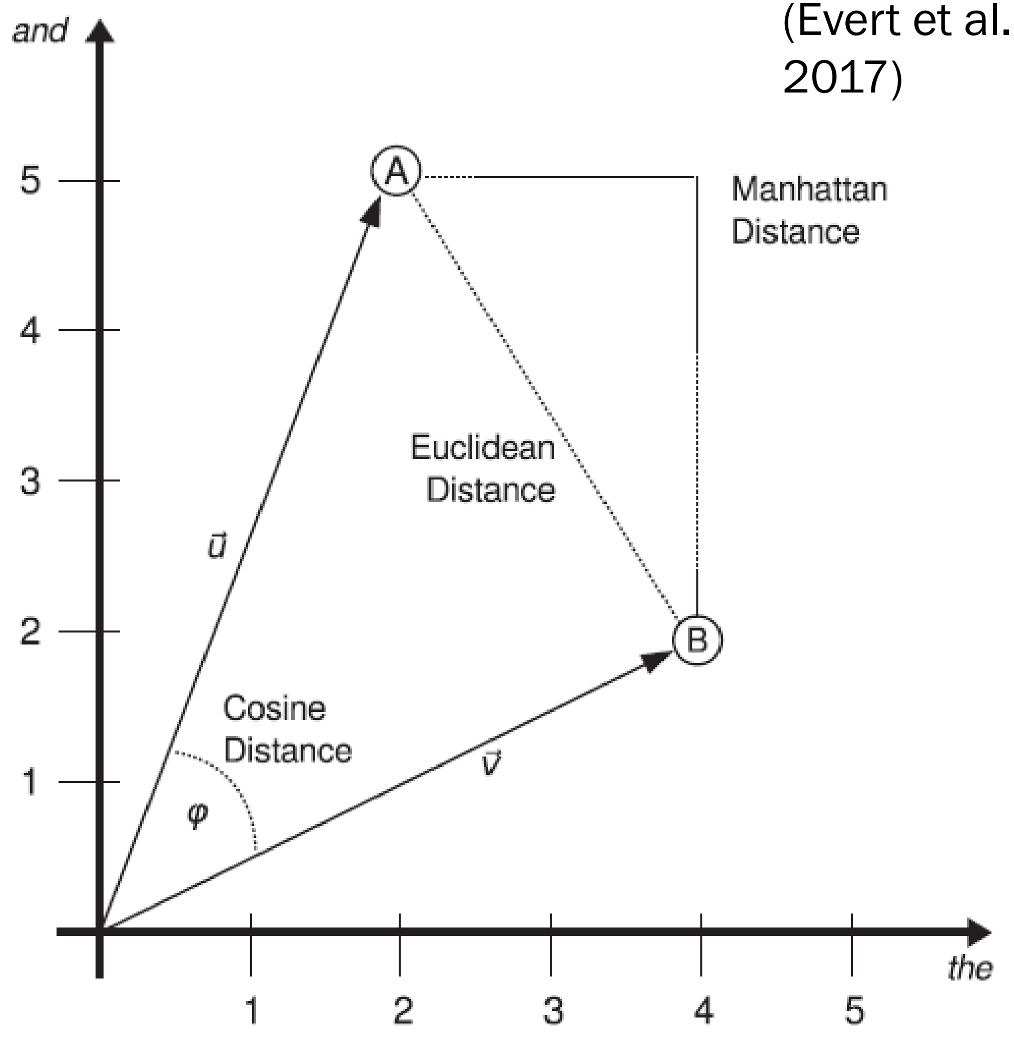
	text A	text B
and	5	2
the	2	4
of	3	5
in	0	1
for	1	0
…	…	…

How the distance is calculated

	text A	text B
and	5	2
the	2	4
of	3	5
in	0	1
for	1	0
…	…	…

How the distance is calculated

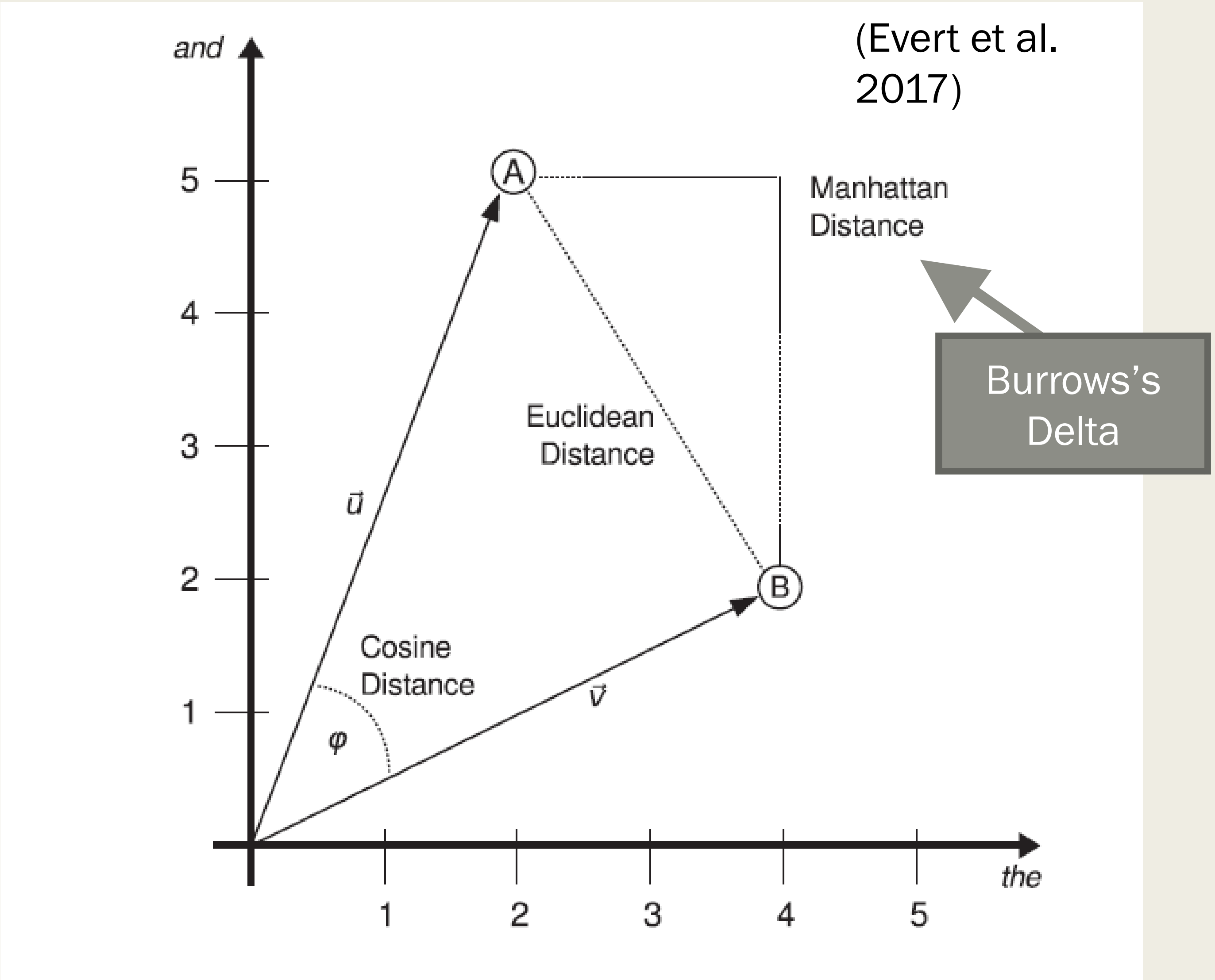
	text A	text B
and	5	2
the	2	4
of	3	5
in	0	1
for	1	0
…	…	…



(Evert et al. 2017)

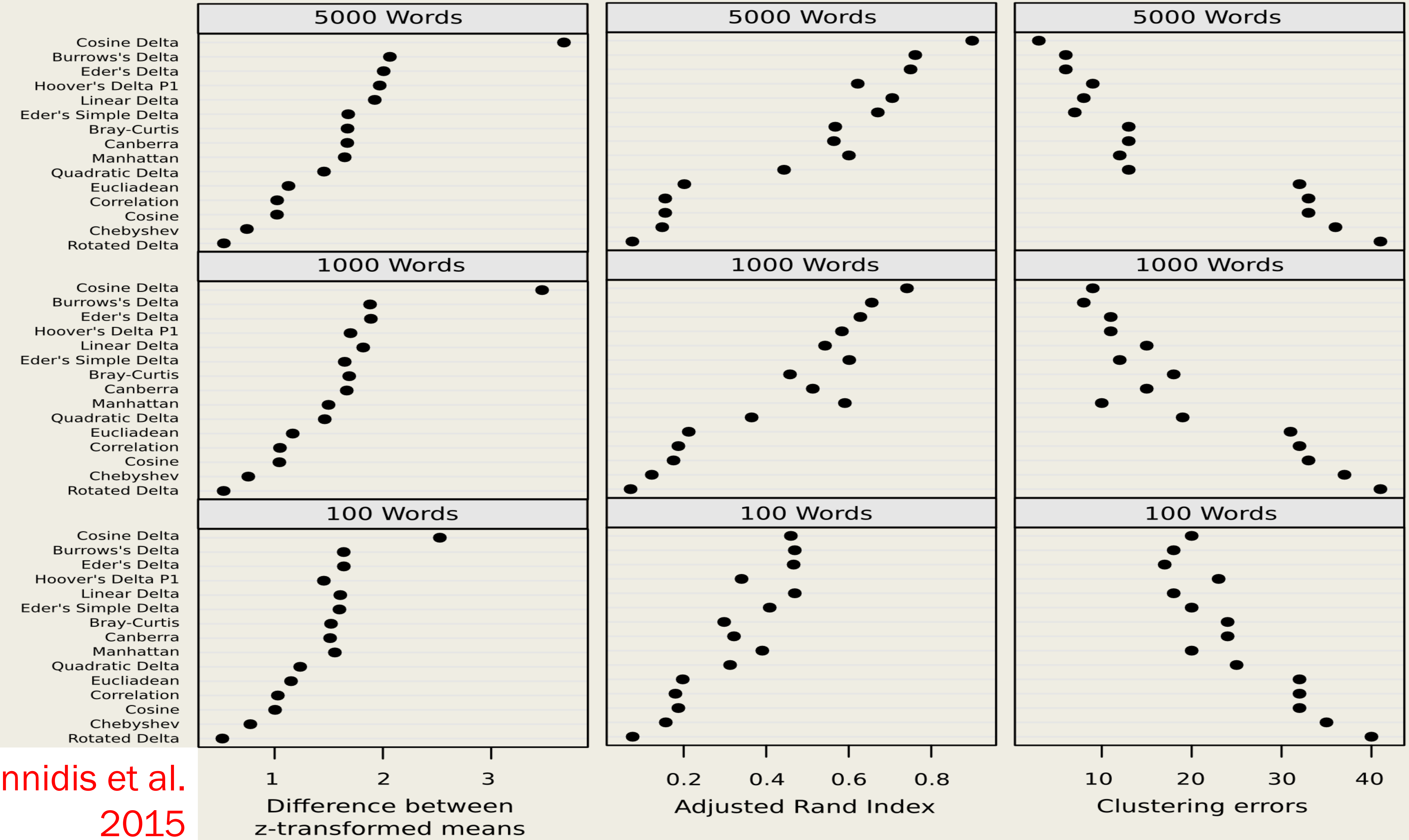
How the distance is calculated

	text A	text B
and	5	2
the	2	4
of	3	5
in	0	1
for	1	0
…	…	…

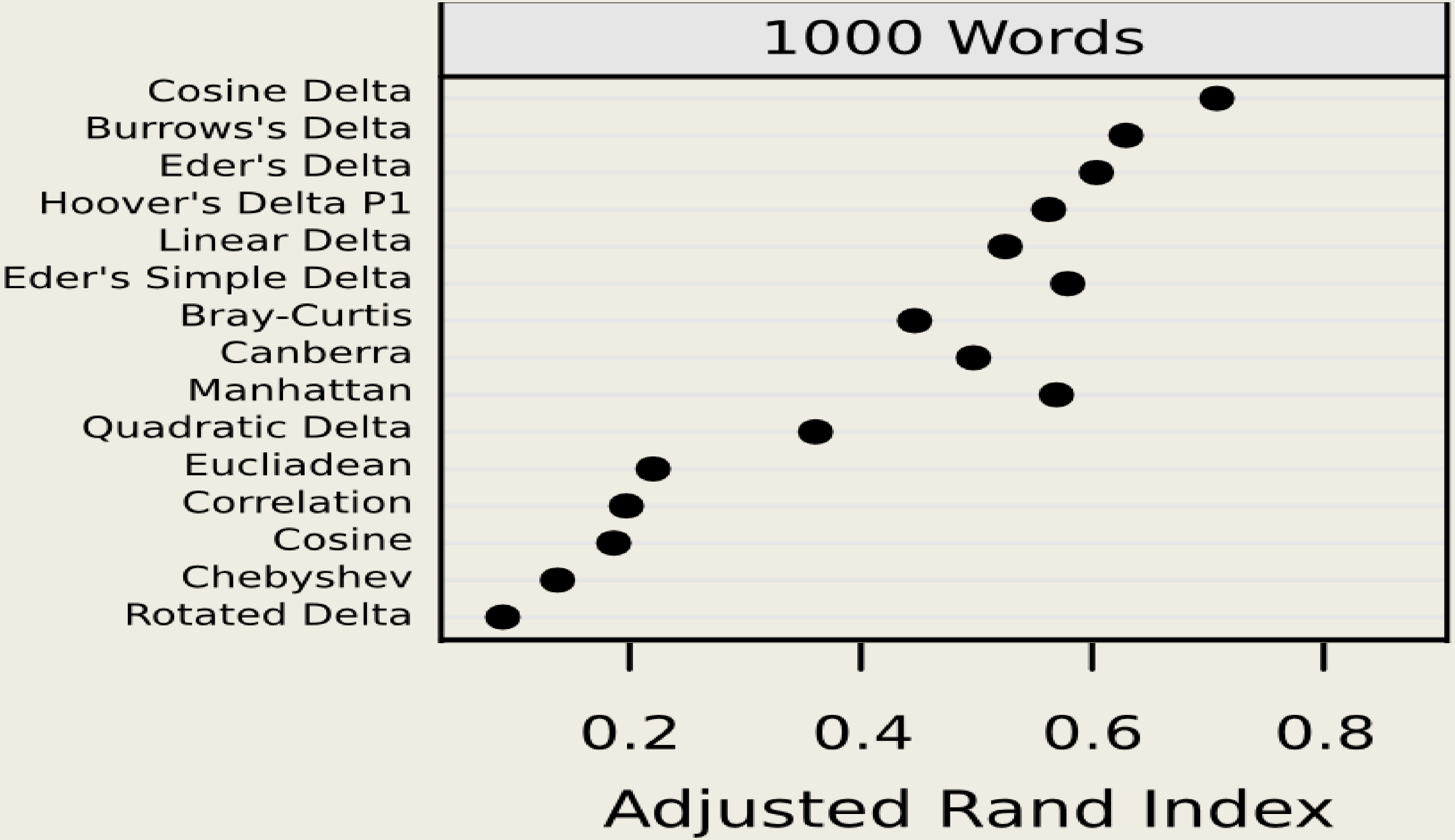


(Evert et al. 2017)

Burrows's Delta



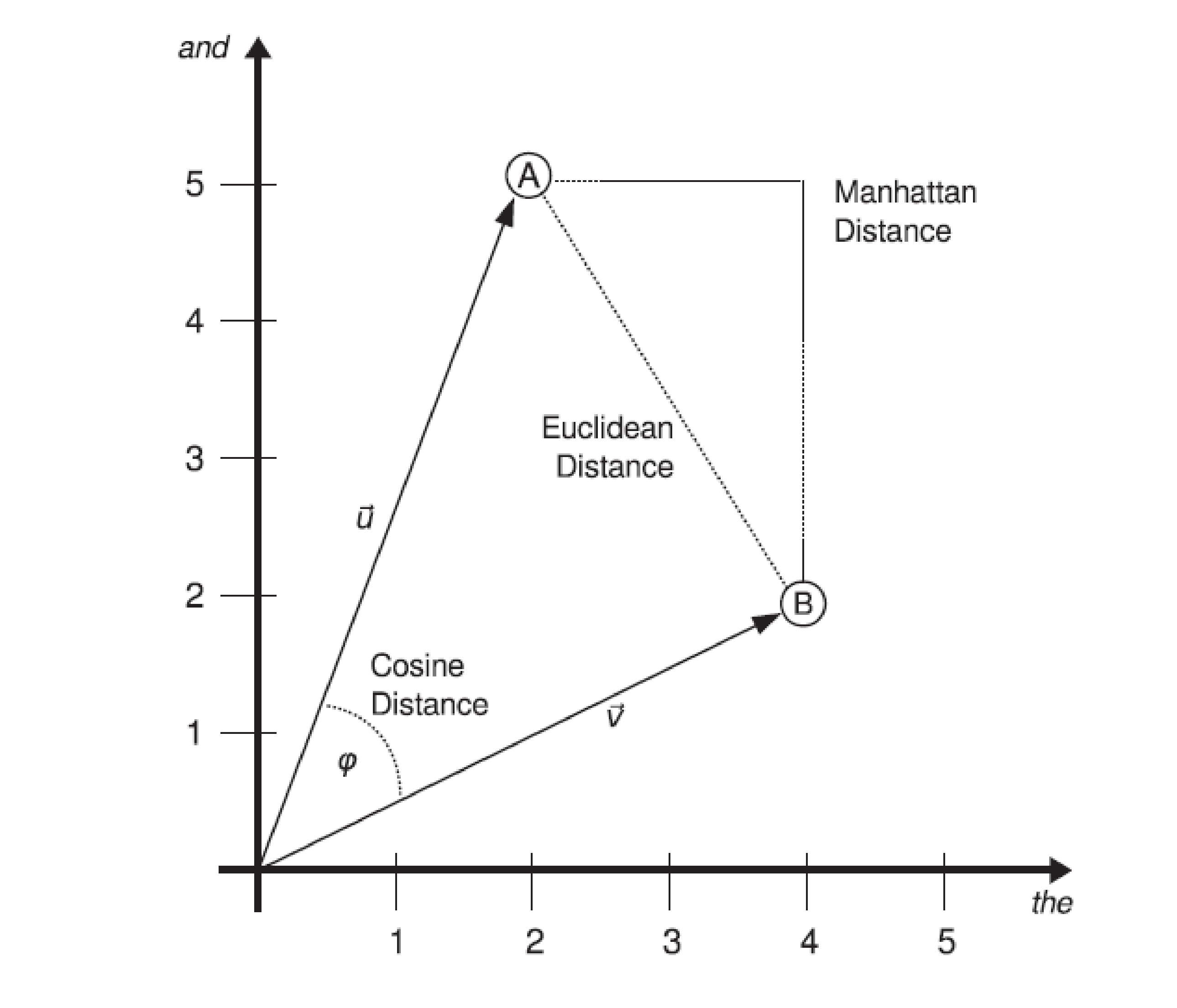
Jannidis et al. 2015


1000 Words
Cosine Delta
Burrows's Delta
Eder's Delta
Hoover's Delta P1
Linear Delta
Eder's Simple Delta
Bray-Curtis
Canberra
Manhattan
Quadratic Delta
Eucliadean
Correlation
Cosine
Chebyshev
Rotated Delta
0.2
0.4
0.6
0.8
Adjusted Rand Index

Delta and Cosine Delta

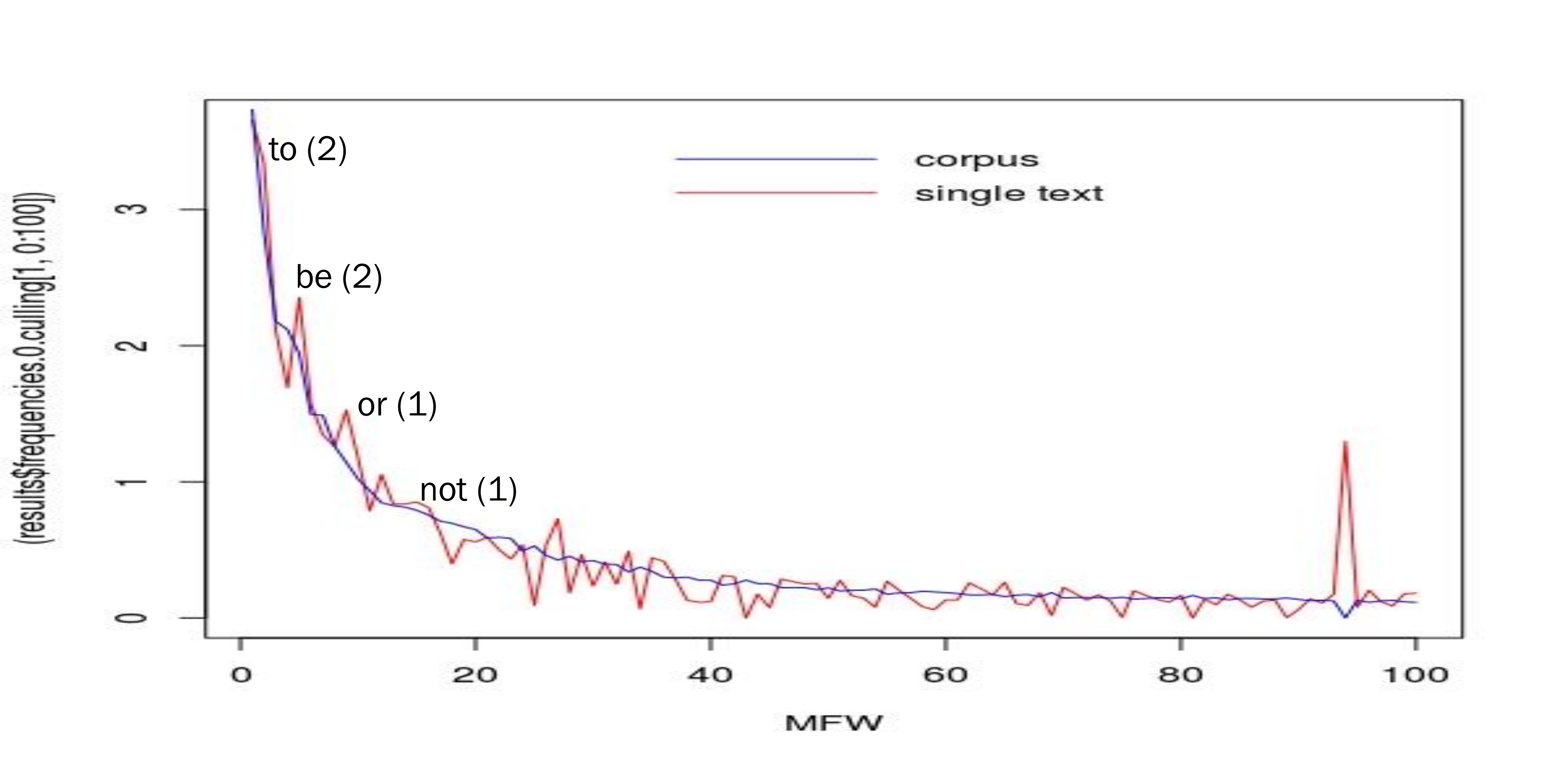
- They measure an angle (or a taxi drive) between two vectors (representing two texts)
- In a n-dimensional space (representing the most frequent words)
- The values are not frequencies, but z-scores

Delta and Cosine Delta

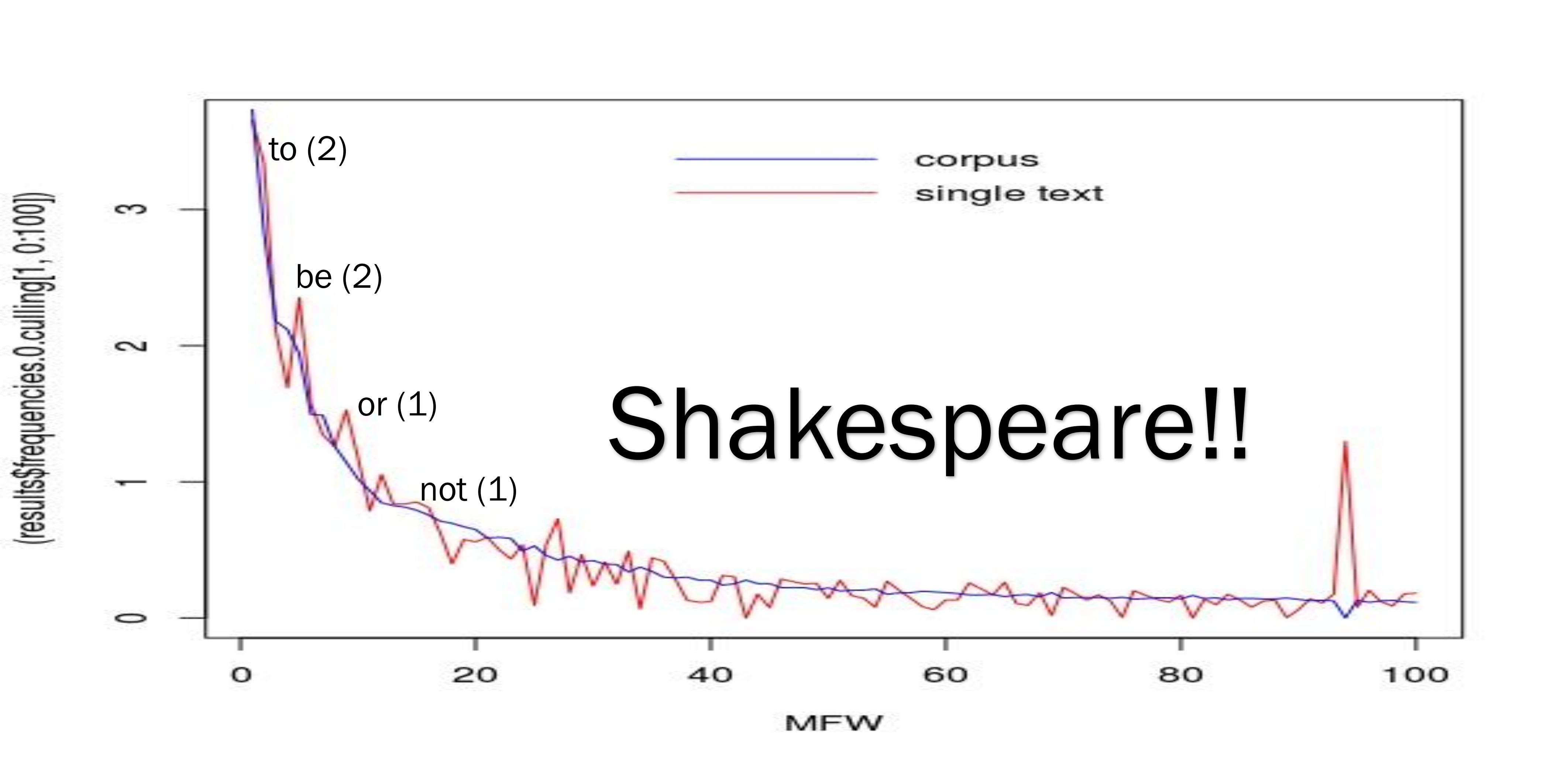


- They measure an angle (or a taxi drive) between two vectors (representing two texts)
- In a n-dimensional space (representing the most frequent words)
- The values are not frequencies, but z-scores

Caveat! Text Length

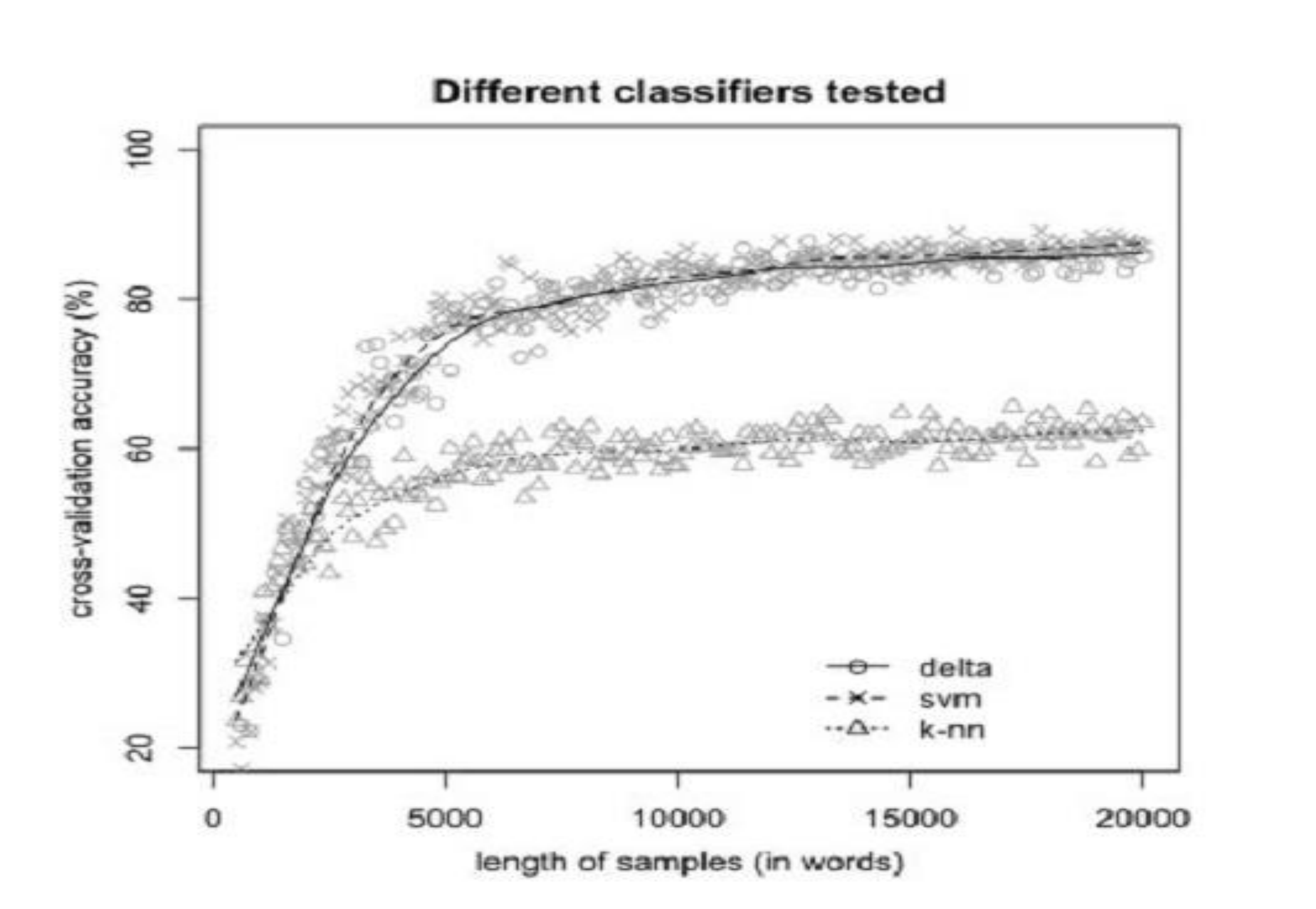


Caveat! Text Length



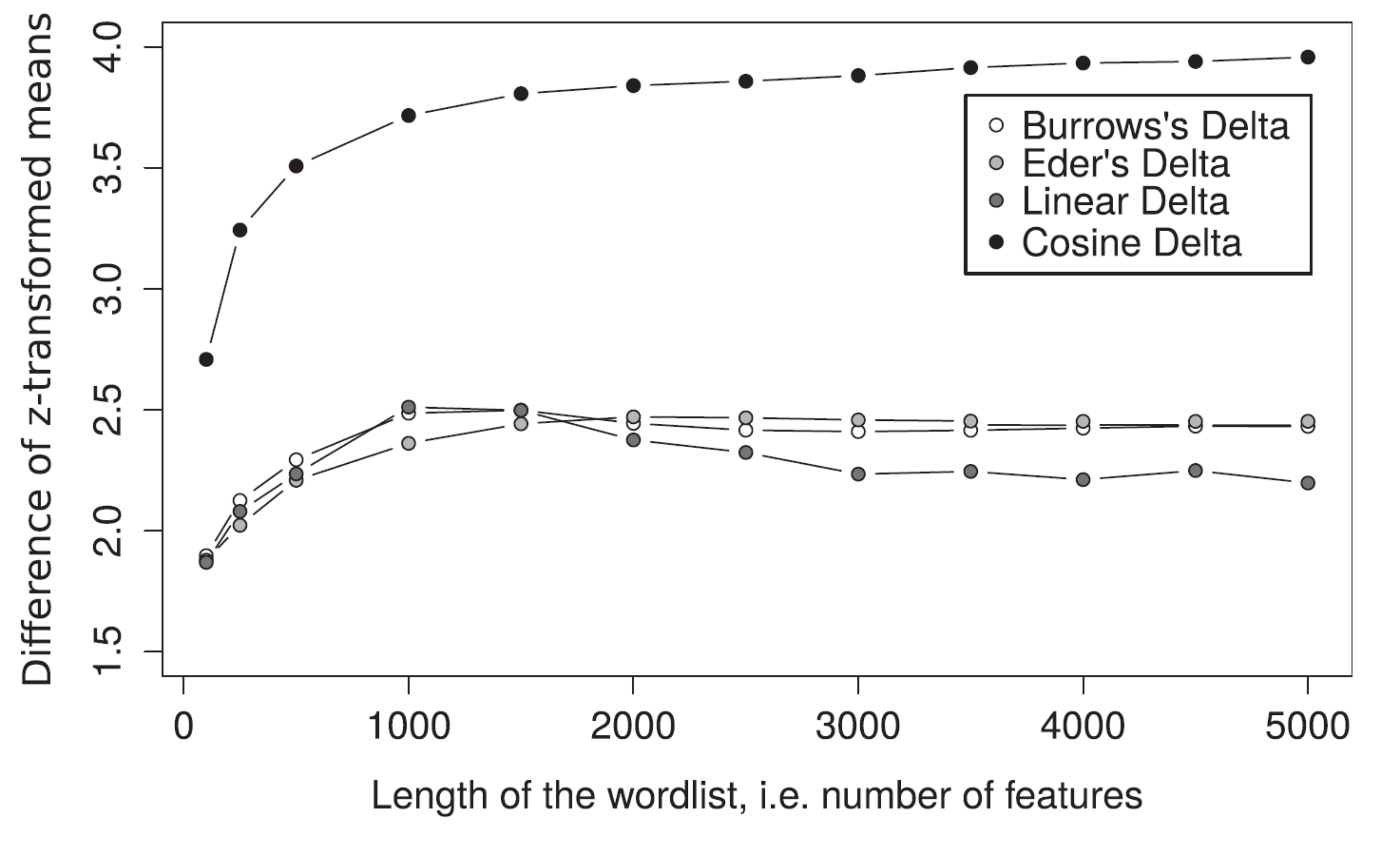
Shakespeare!!

Caveat! Text Length



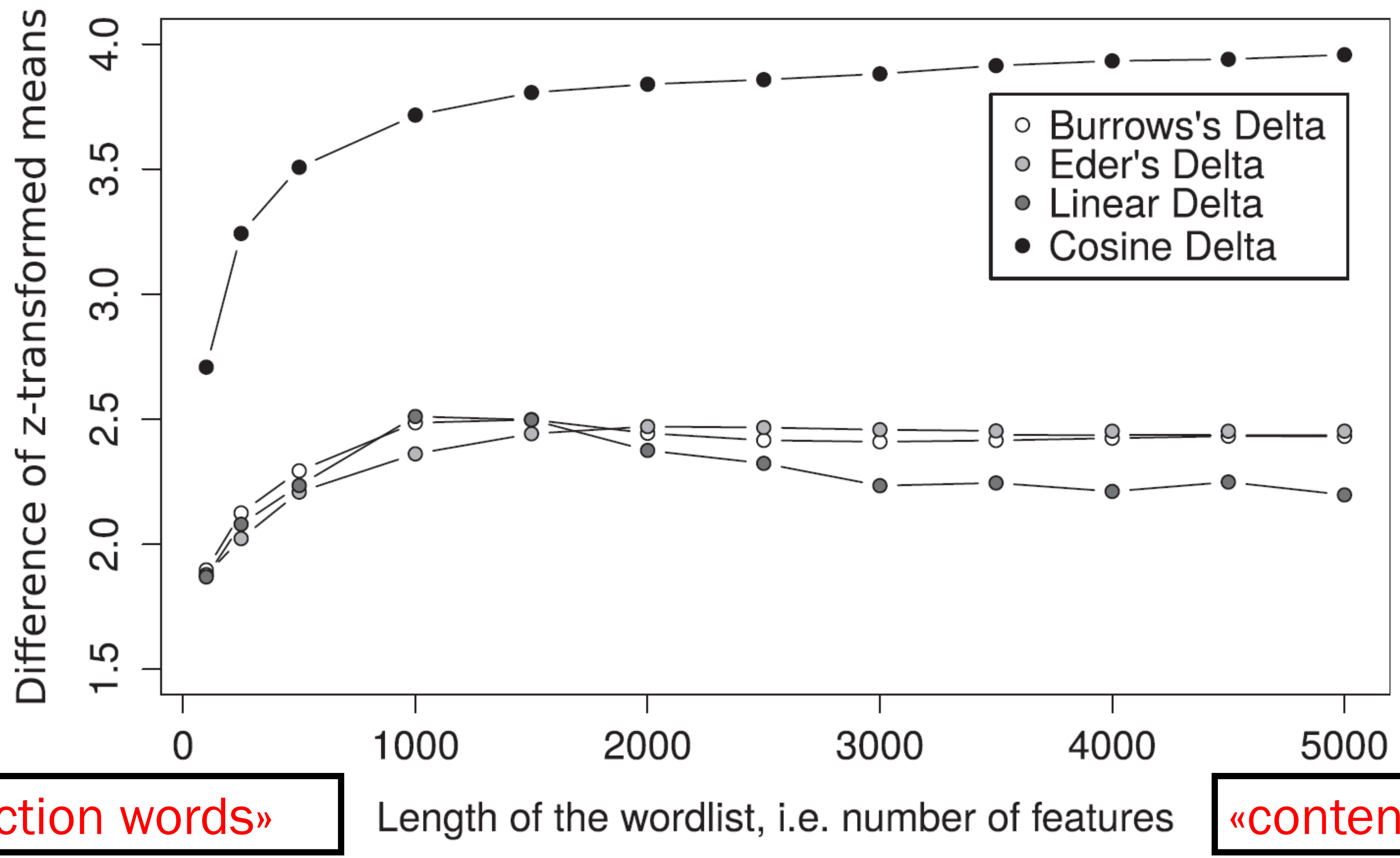
Minumun text length for a reliable stylometric analysis is about 5,000 words (Eder 2015)

Caveat(2)! How many MFW?



About 2,000 MFW produce the best results (Evert et al. 2017)

Caveat(2)! How many MFW?

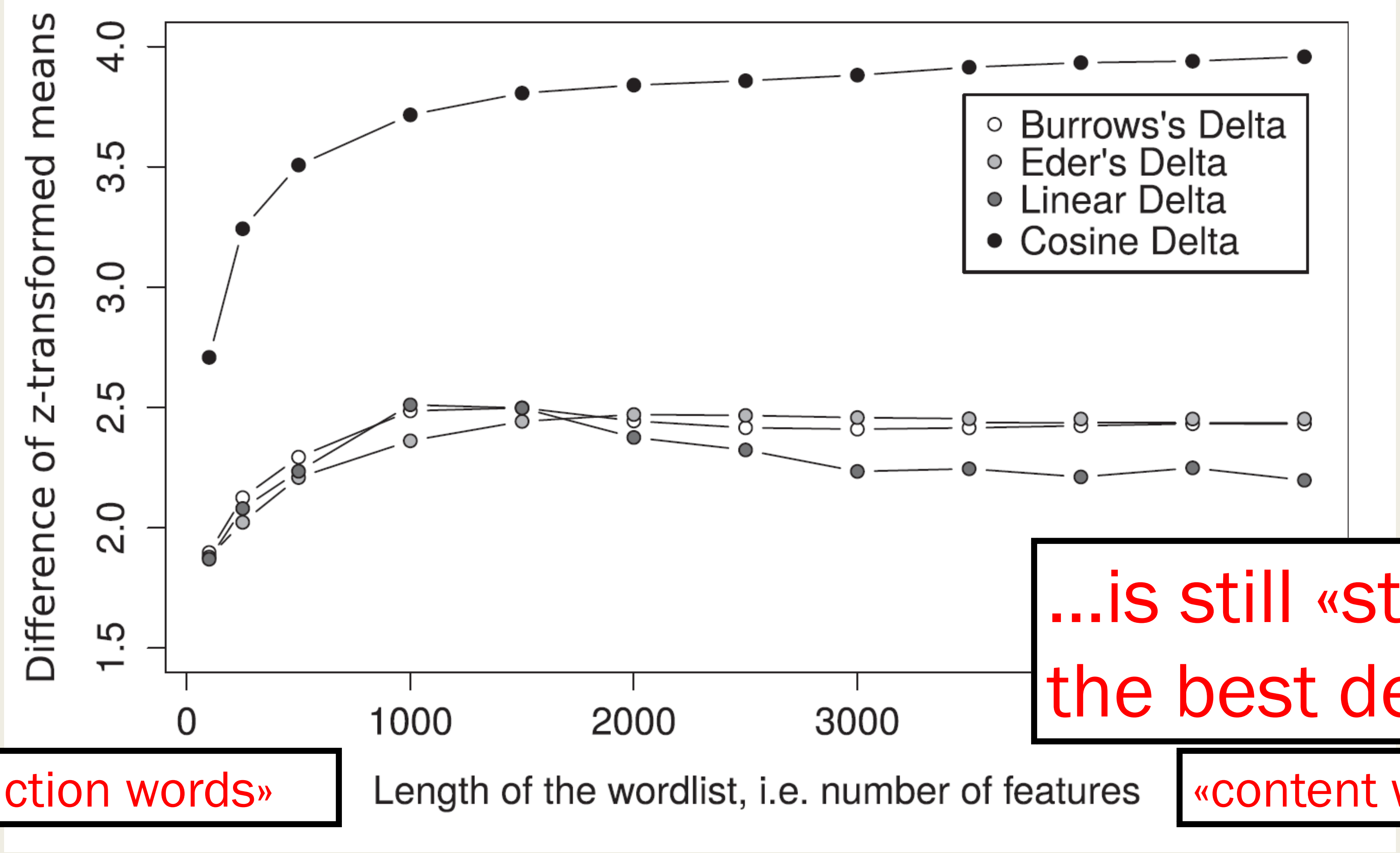


«function words»

«content words»

About 2,000 MFW produce the best results (Evert et al. 2017)

Caveat(2)! How many MFW?



«function words»

«content words»

...is still «stylometry» the best definition?

About 2,000 MFW produce the best results (Evert et al. 2017)

Not only most frequent words...

Not only most frequent words…

Not only most frequent words…

- Character n-grams frequency

Not only most frequent words…

- Character n-grams frequency
- Word n-grams frequency

Not only most frequent words…

- Character n-grams frequency
- Word n-grams frequency
- Word skip-grams frequency

Not only most frequent words…

- Character n-grams frequency
- Word n-grams frequency
- Word skip-grams frequency
- POS-tags n-grams frequency

Not only most frequent words…

- Character n-grams frequency
- Word n-grams frequency
- Word skip-grams frequency
- POS-tags n-grams frequency
- Syntactic labels n-grams frequency

Not only most frequent words…

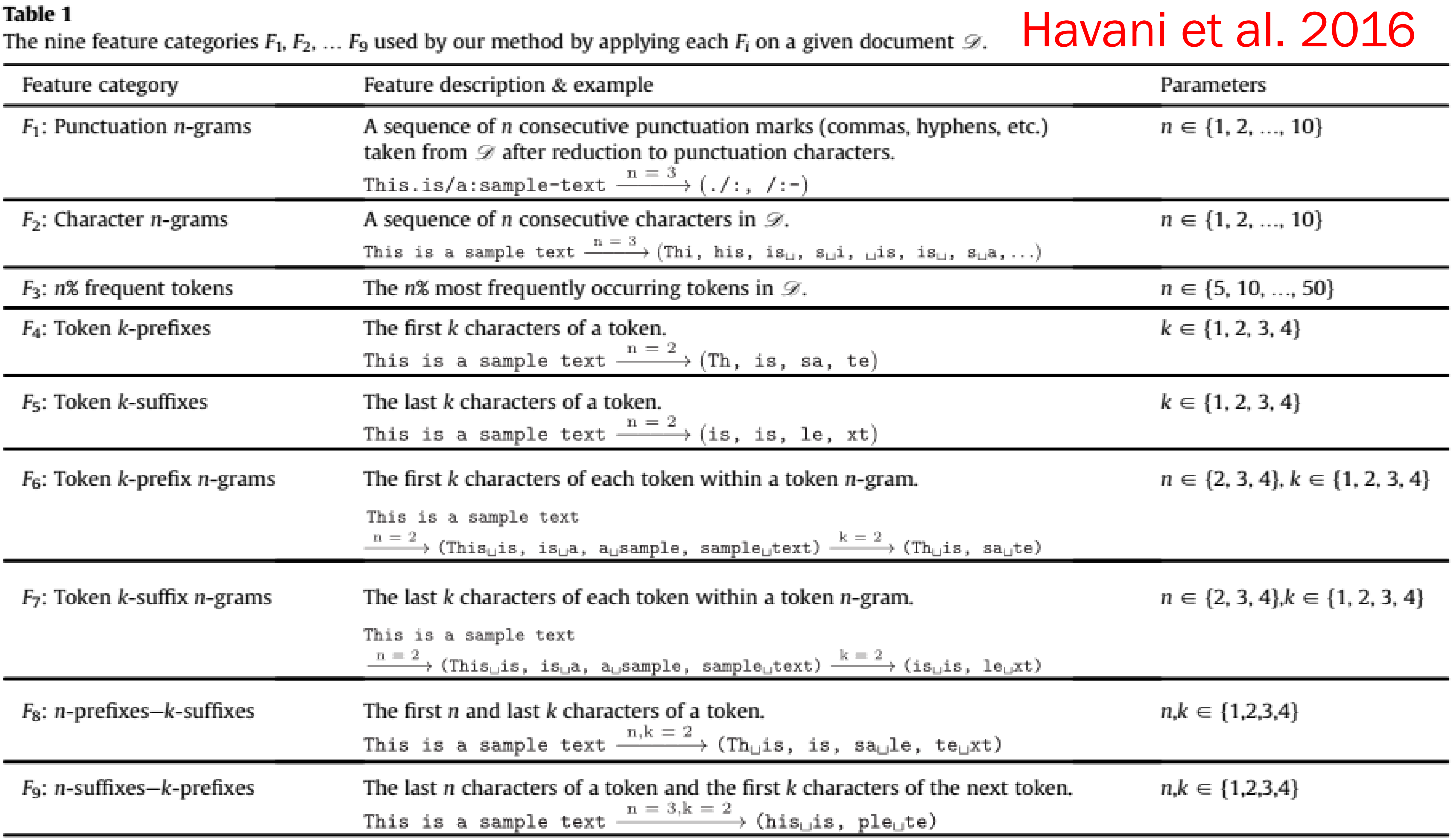
- Character n-grams frequency
- Word n-grams frequency
- Word skip-grams frequency
- POS-tags n-grams frequency
- Syntactic labels n-grams frequency
- Word length / sentence length

Not only most frequent words…

- Character n-grams frequency
- Word n-grams frequency
- Word skip-grams frequency
- POS-tags n-grams frequency
- Syntactic labels n-grams frequency
- Word length / sentence length
- Punctuation (n-grams) frequency

Not only most frequent words…

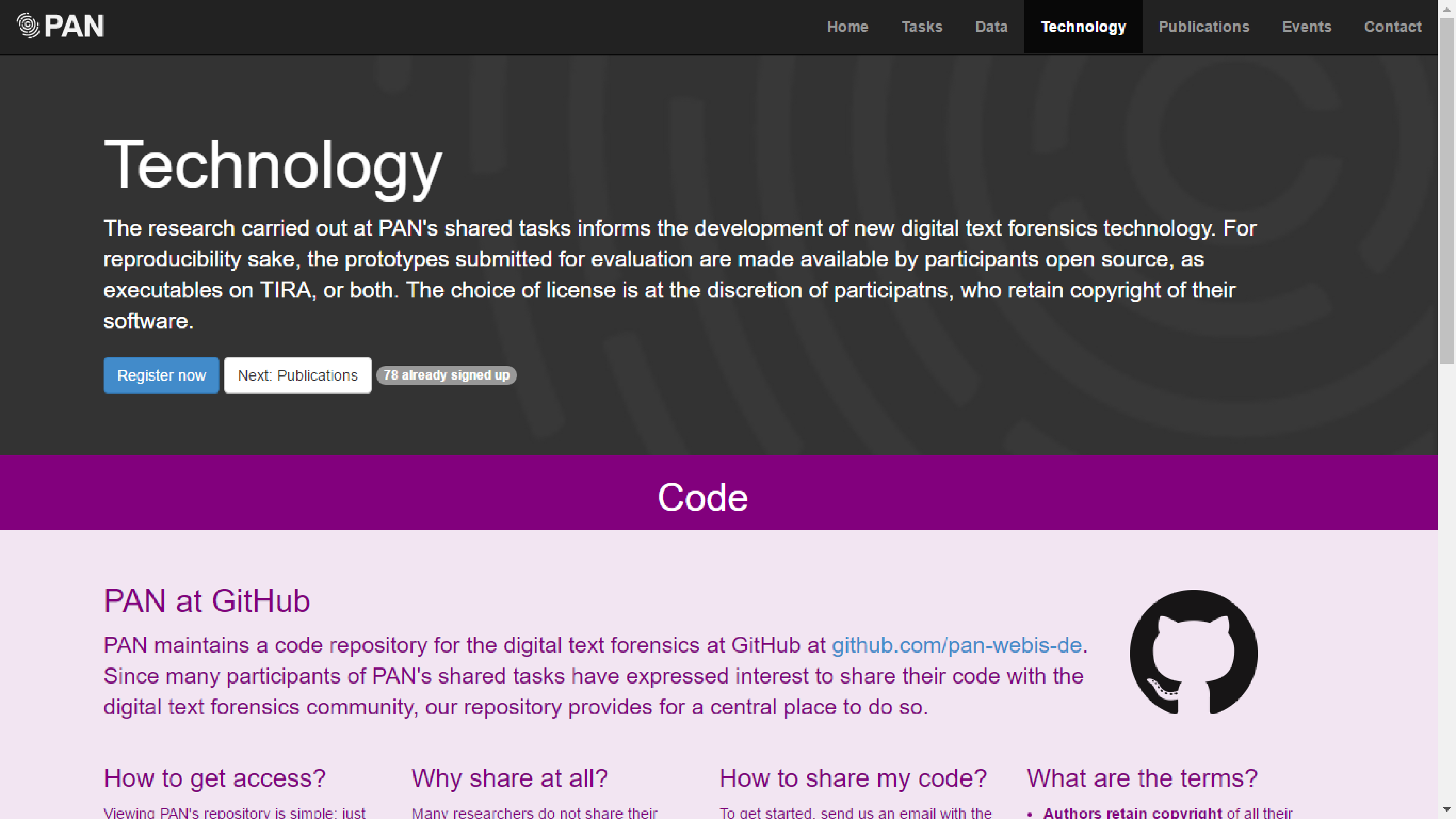
- Character n-grams frequency
- Word n-grams frequency
- Word skip-grams frequency
- POS-tags n-grams frequency
- Syntactic labels n-grams frequency
- Word length / sentence length
- Punctuation (n-grams) frequency
- …



Havani et al. 2016

Table 1
The nine feature categories $F_1, F_2, \ldots F_9$ used by our method by applying each F_i on a given document $\mathscr{D}$.

Feature category	Feature description & example	Parameters
F_1: Punctuation n-grams	A sequence of n consecutive punctuation marks (commas, hyphens, etc.) taken from $\mathscr{D}$ after reduction to punctuation characters. `This.is/a:sample-text` $\xrightarrow{n = 3}$ `(./:, /:-)`	$n \in \{1, 2, \ldots, 10\}$
F_2: Character n-grams	A sequence of n consecutive characters in $\mathscr{D}$. `This is a sample text` $\xrightarrow{n = 3}$ `(Thi, his, is␣, s␣i, ␣is, is␣, s␣a, ...)`	$n \in \{1, 2, \ldots, 10\}$
F_3: n% frequent tokens	The n% most frequently occurring tokens in $\mathscr{D}$.	$n \in \{5, 10, \ldots, 50\}$
F_4: Token k-prefixes	The first k characters of a token. `This is a sample text` $\xrightarrow{n = 2}$ `(Th, is, sa, te)`	$k \in \{1, 2, 3, 4\}$
F_5: Token k-suffixes	The last k characters of a token. `This is a sample text` $\xrightarrow{n = 2}$ `(is, is, le, xt)`	$k \in \{1, 2, 3, 4\}$
F_6: Token k-prefix n-grams	The first k characters of each token within a token n-gram. `This is a sample text` $\xrightarrow{n = 2}$ `(This␣is, is␣a, a␣sample, sample␣text)` $\xrightarrow{k = 2}$ `(Th␣is, sa␣te)`	$n \in \{2, 3, 4\}$, $k \in \{1, 2, 3, 4\}$
F_7: Token k-suffix n-grams	The last k characters of each token within a token n-gram. `This is a sample text` $\xrightarrow{n = 2}$ `(This␣is, is␣a, a␣sample, sample␣text)` $\xrightarrow{k = 2}$ `(is␣is, le␣xt)`	$n \in \{2, 3, 4\}$,$k \in \{1, 2, 3, 4\}$
F_8: n-prefixes–k-suffixes	The first n and last k characters of a token. `This is a sample text` $\xrightarrow{n,k = 2}$ `(Th␣is, is, sa␣le, te␣xt)`	n,$k \in \{1,2,3,4\}$
F_9: n-suffixes–k-prefixes	The last n characters of a token and the first k characters of the next token. `This is a sample text` $\xrightarrow{n = 3, k = 2}$ `(his␣is, ple␣te)`	n,$k \in \{1,2,3,4\}$



Technology

The research carried out at PAN's shared tasks informs the development of new digital text forensics technology. For reproducibility sake, the prototypes submitted for evaluation are made available by participants open source, as executables on TIRA, or both. The choice of license is at the discretion of participatns, who retain copyright of their software.

Code

PAN at GitHub

PAN maintains a code repository for the digital text forensics at GitHub at github.com/pan-webis-de. Since many participants of PAN's shared tasks have expressed interest to share their code with the digital text forensics community, our repository provides for a central place to do so.

How to get access?

Viewing PAN's repository is simple: just

Why share at all?

Many researchers do not share their

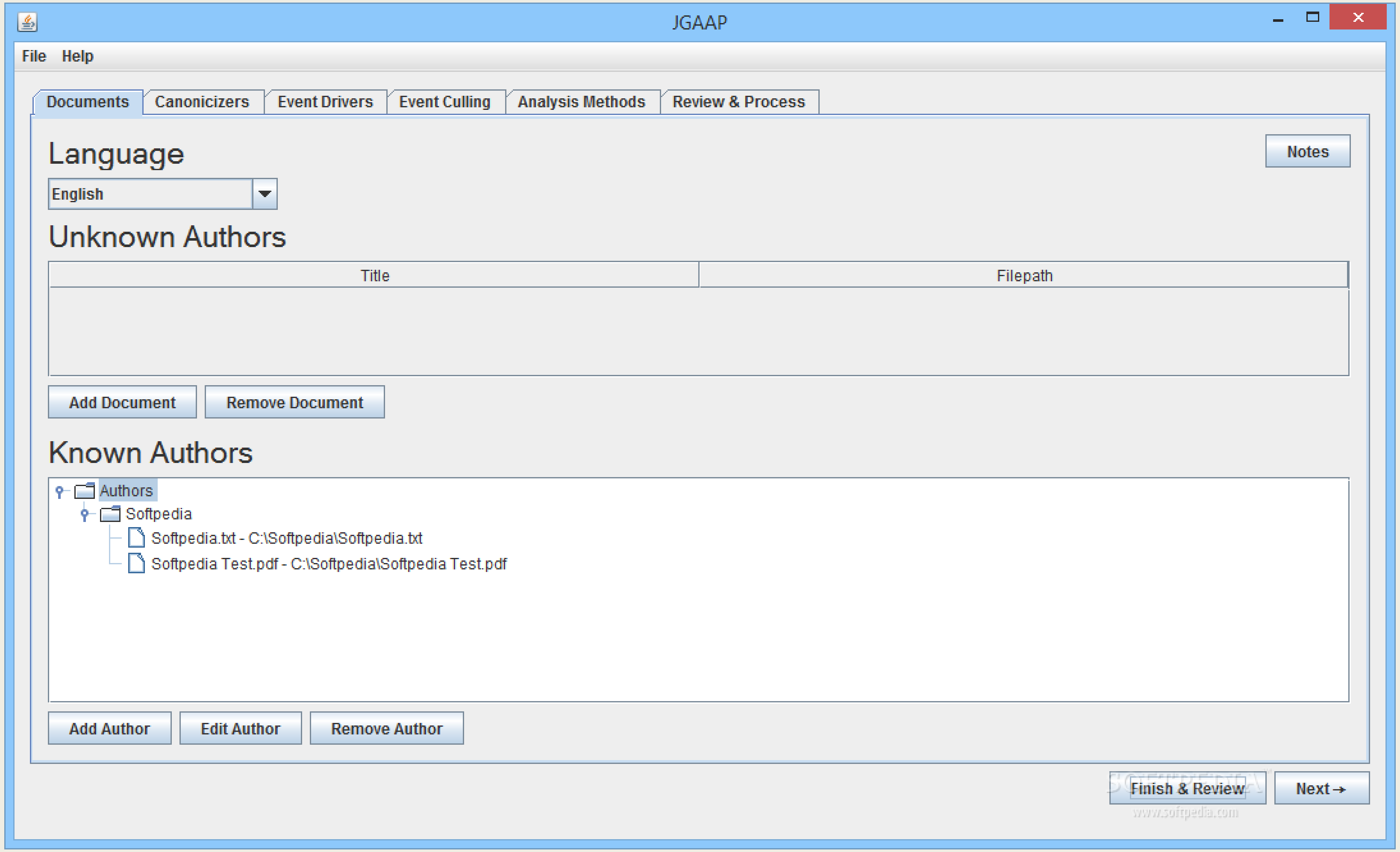
How to share my code?

To get started, send us an email with the

What are the terms?

- **Authors retain copyright** of all their

JGAAP -> authorship attribution with thousends of features!



Language

English

Unknown Authors

Title	Filepath

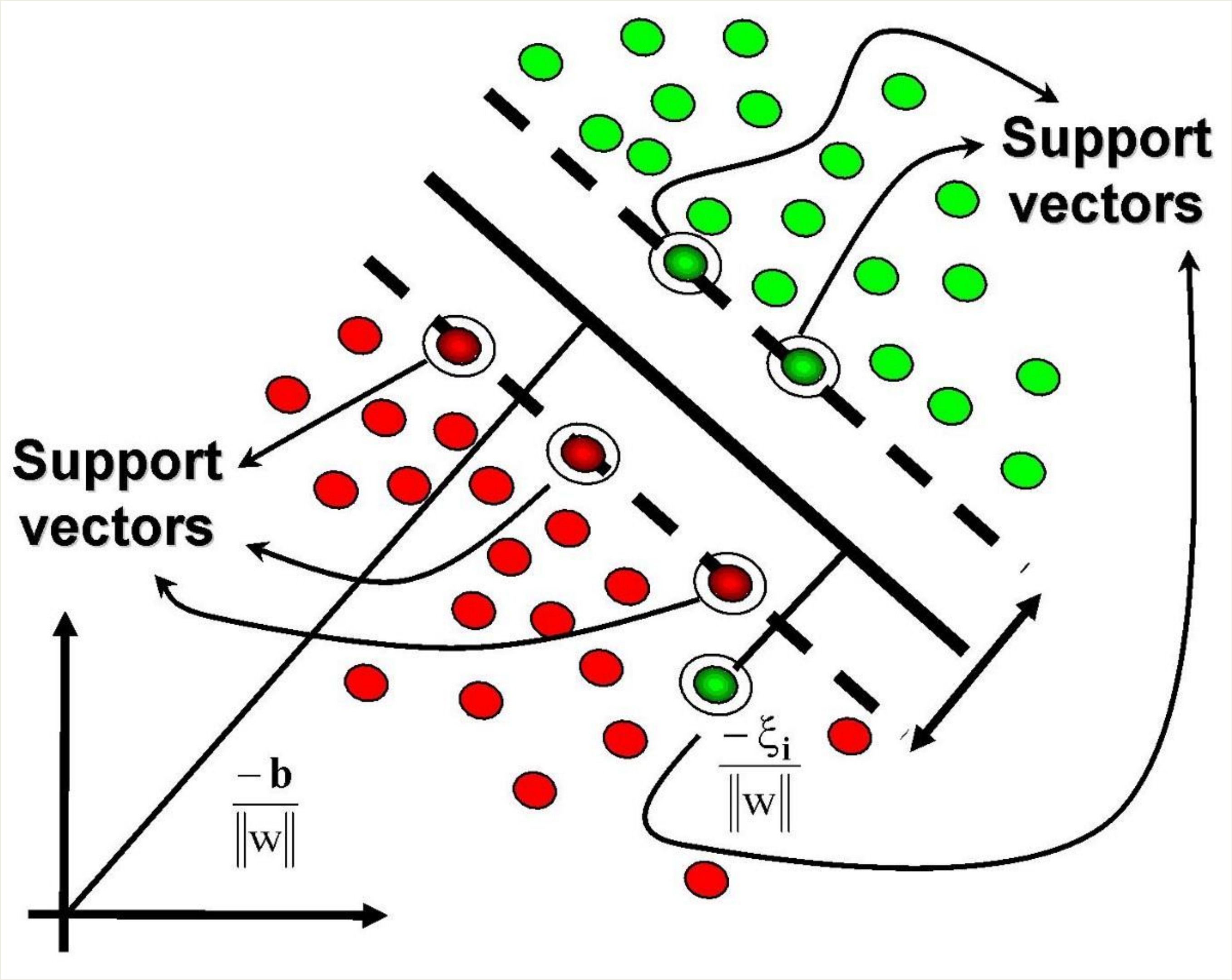
Known Authors

- Authors
  - Softpedia
    - Softpedia.txt - C:\Softpedia\Softpedia.txt
    - Softpedia Test.pdf - C:\Softpedia\Softpedia Test.pdf

Not only distance measures...

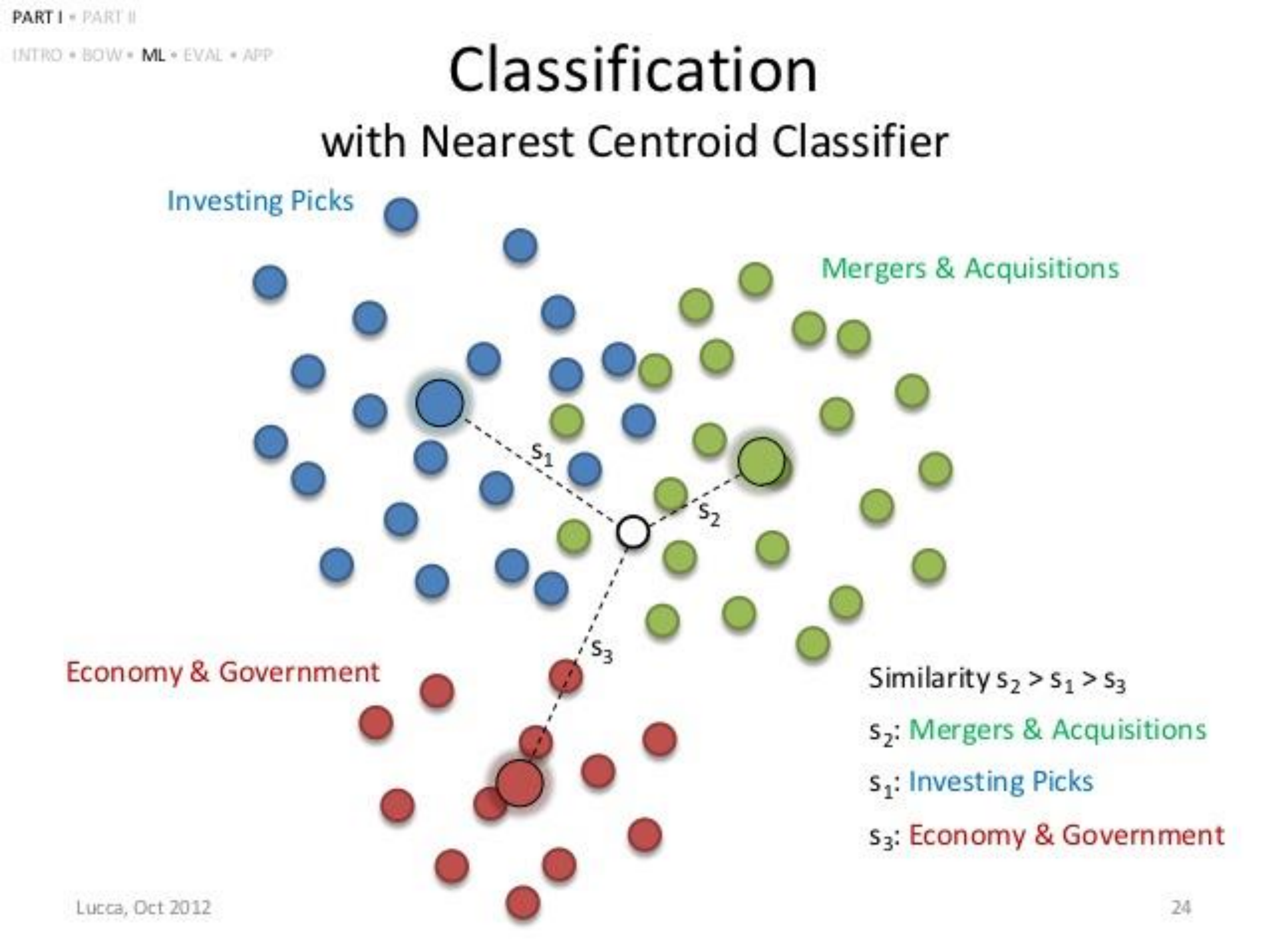
Machine Learning

- Support Vector Machines


Support vectors
Support vectors
$\frac{-\mathbf{b}}{\|\mathrm{w}\|}$
$\frac{-\xi_{\mathbf{i}}}{\|\mathrm{w}\|}$

Machine Learning

- Nearest Shrunken Centroids



PART I • PART II

INTRO • BOW • ML • EVAL • APP

Classification

with Nearest Centroid Classifier

Similarity $s_2 > s_1 > s_3$

s_2: Mergers & Acquisitions

s_1: Investing Picks

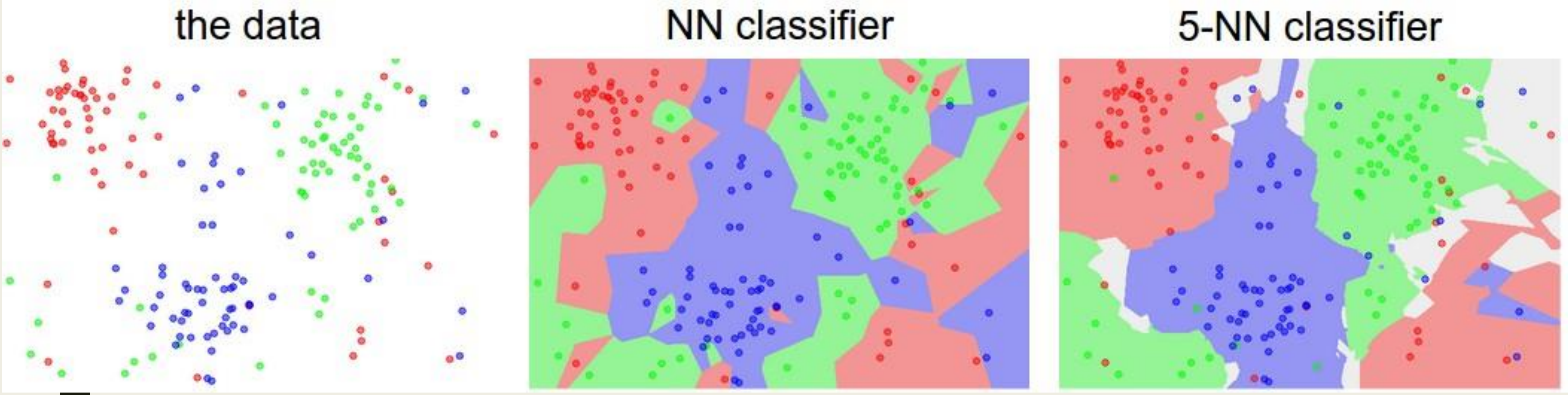
s_3: Economy & Government

Lucca, Oct 2012

24

Machine Learning

- k-NN



Machine Learning <-> Distance Measures

Machine Learning <-> Distance Measures

- Instead of calculating the distances between all texts in the corpus…

Machine Learning <-> Distance Measures

- Instead of calculating the distances between all texts in the corpus…
- The corpus is divided in two parts:
  training set
  and test set

Machine Learning <-> Distance Measures

- Instead of calculating the distances between all texts in the corpus…
- The corpus is divided in two parts:
  training set
  and test set
- The algorithms «learn» to distinguish the authors by working on the training set

Machine Learning <-> Distance Measures

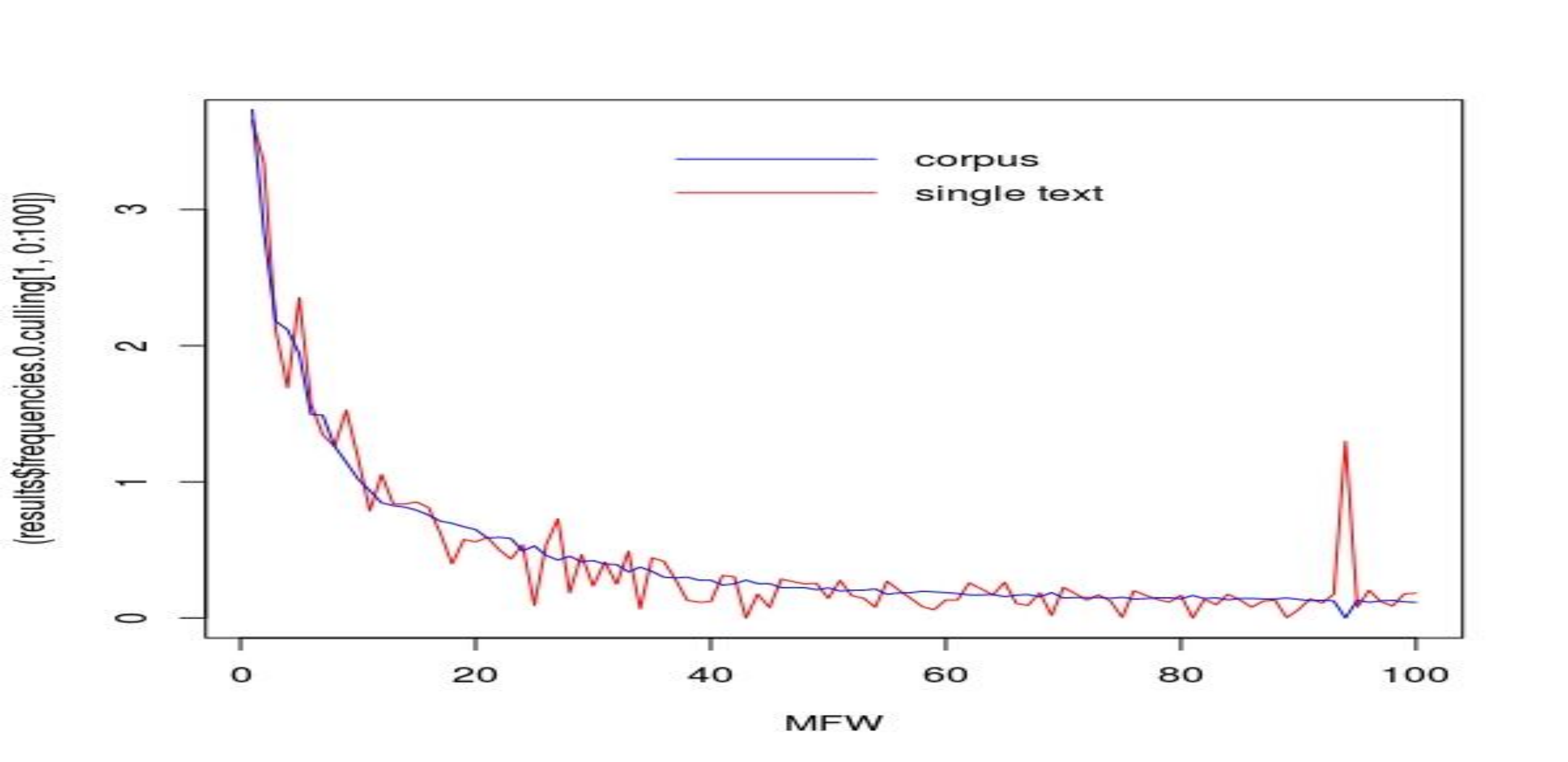
- Instead of calculating the distances between all texts in the corpus…
- The corpus is divided in two parts:
  training set
  and test set
- The algorithms «learn» to distinguish the authors by working on the training set
- …and they are «tested» on the test set

Keyness analysis

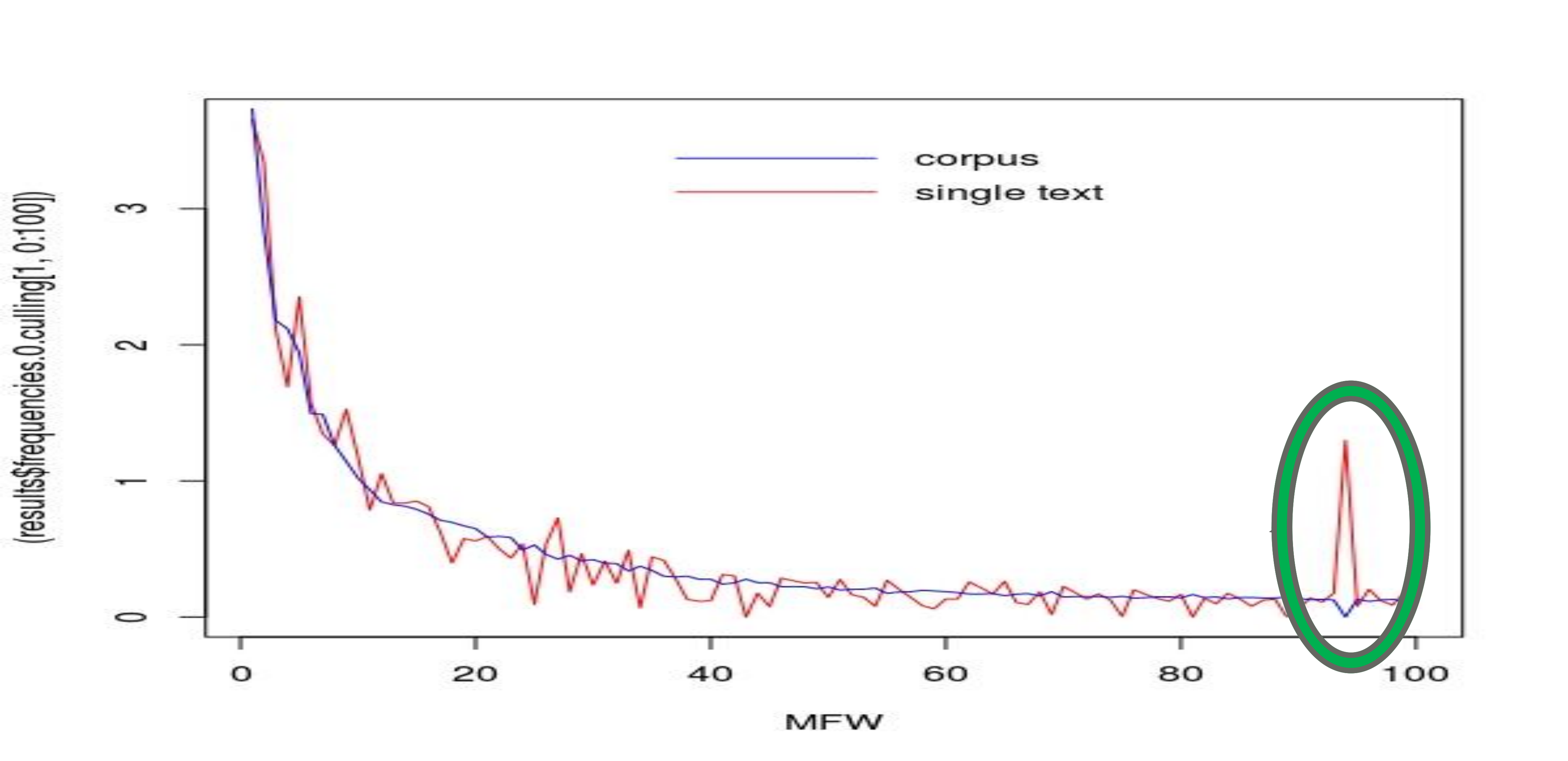
Keyness Analysis

“This established measure of corpus stylistics (cf. Hoover et al., 2015) compares the frequencies of single words included in some text (collection) with those obtained in a (normally larger) reference corpus. It outputs a long list of words that deviate statistically from that reference corpus (cf. Rayson, 2012; Scott & Tribble, 2006). Here, the reference corpus acts as a statistical ‘norm’ against which the word use in the text(s) under scrutiny may be compared. The examined words, depending on whether they deviate positively or negatively, are thus “over-” or “under-represented” with regard to that norm.” (Herrmann 2017)

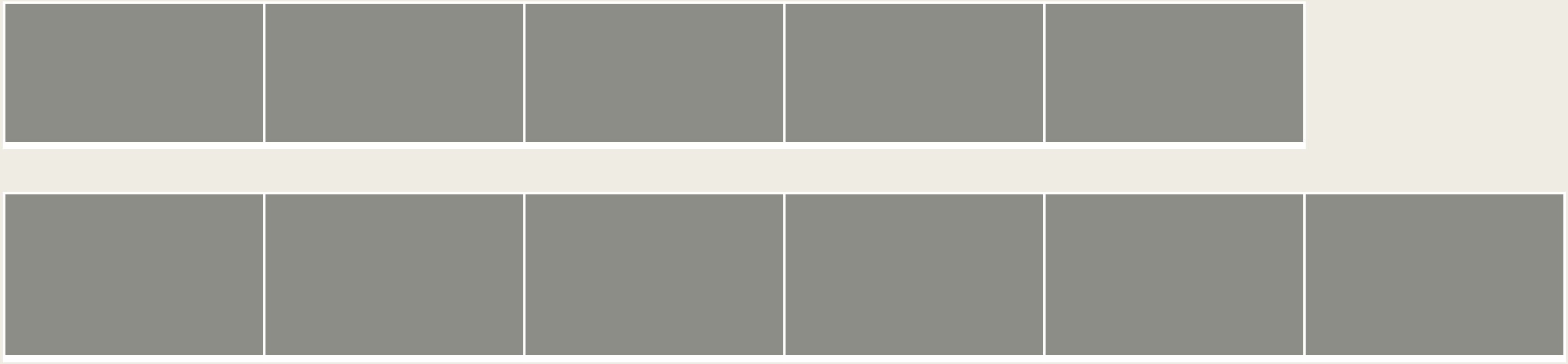
Keyness Analysis



Keyness Analysis



Zeta Analysis



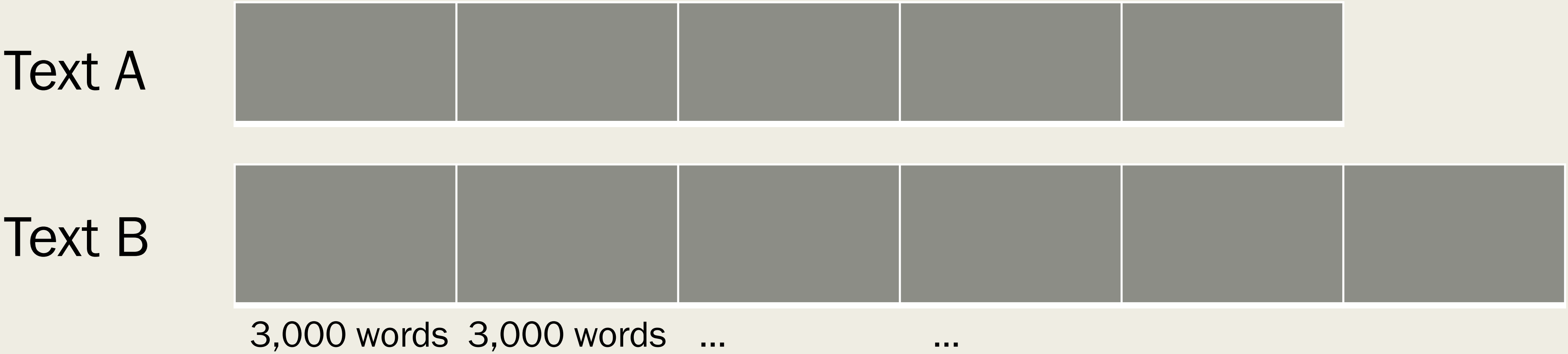
Text A

Text B

3,000 words 3,000 words … …

Zeta Analysis

Pick up a word:
«fou» (for example)



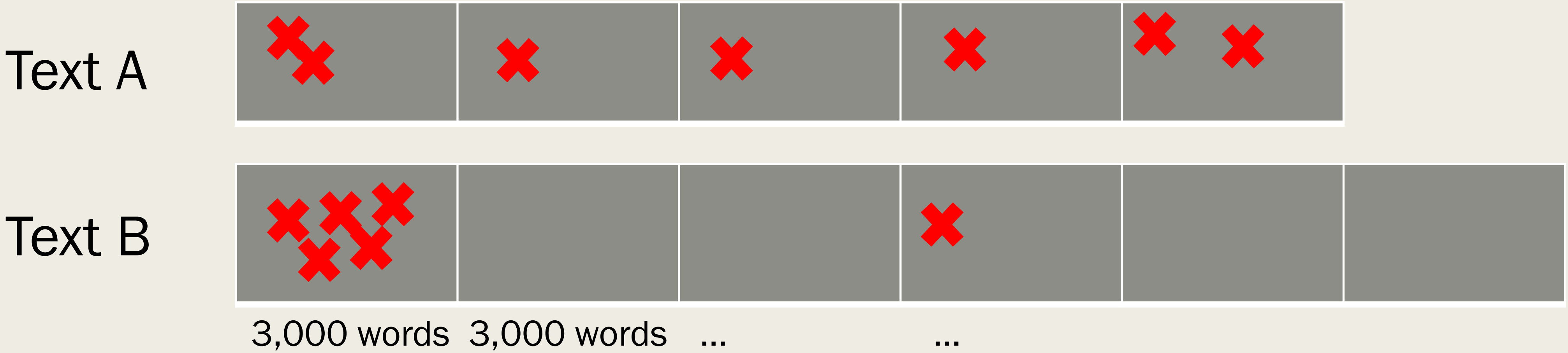
Text A

Text B

3,000 words 3,000 words … …

Zeta Analysis

Pick up a word:
«fou» (for example)



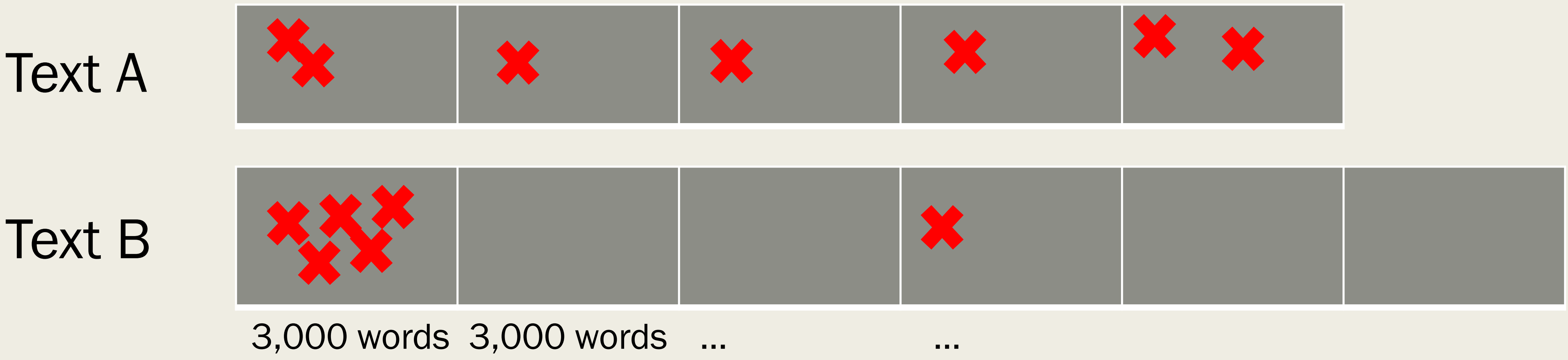
Text A

Text B

3,000 words 3,000 words … …

Zeta Analysis

Pick up a word:
«fou» (for example)

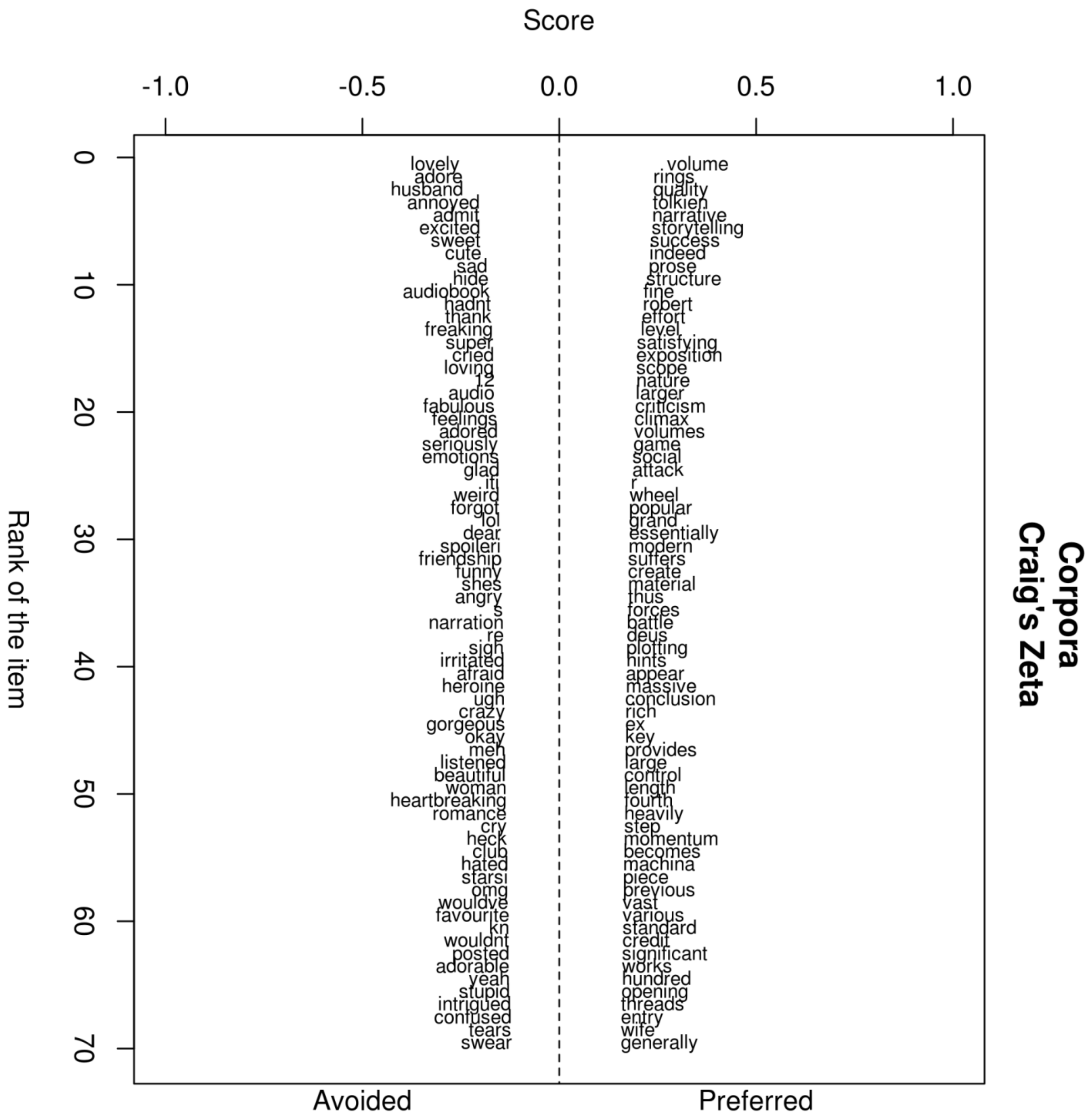


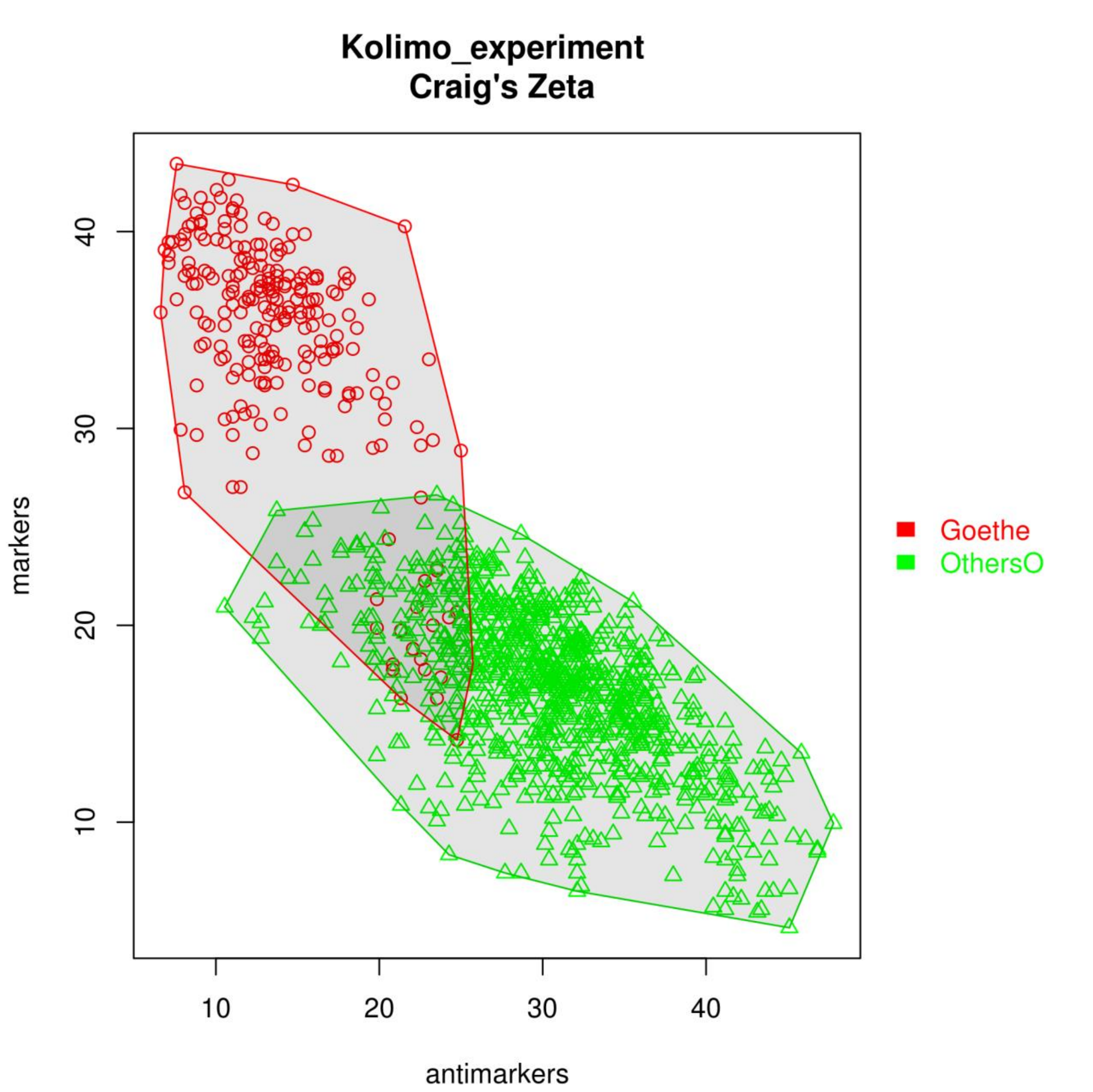
Text A

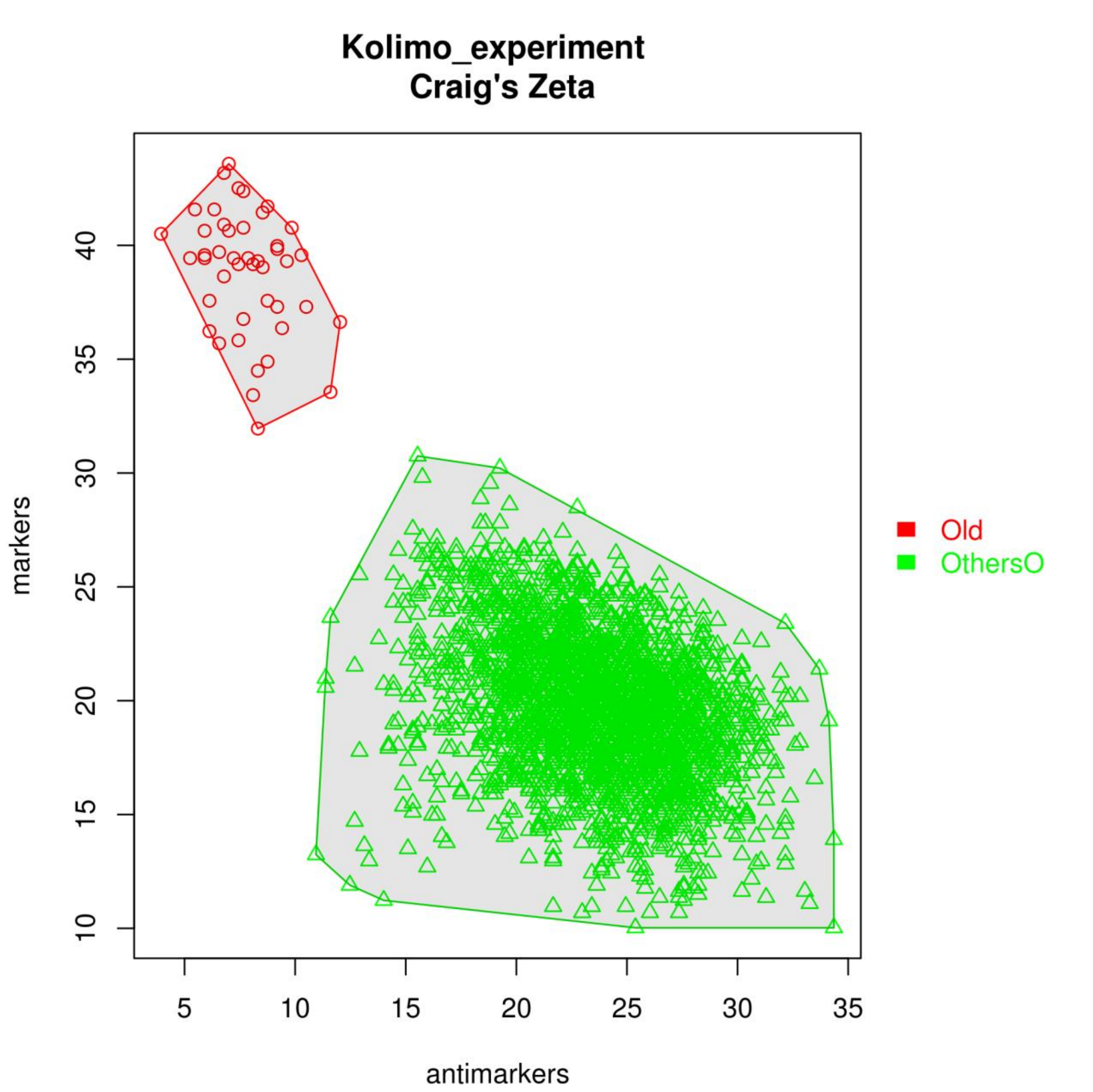
Text B

3,000 words 3,000 words … …

- Count in how many slices of the text appears the word «fou»
- Calculate the proportion
  Text A: 1 (100%); text B: 0.33 (33%)
- Subtract the two values
  (so the word «fou» has Zeta = 0.66 for Text A)
- Repeat the operation for all the words in the two texts


Score
-1.0
-0.5
0.0
0.5
1.0
Rank of the item
0
10
20
30
40
50
60
70
lovely
adore
husband
annoyed
admit
excited
sweet
cute
sad
hide
audiobook
hadnt
thank
freaking
super
cried
loving
12
audio
fabulous
feelings
adored
seriously
emotions
glad
iti
weird
forgot
lol
dear
spoileri
friendship
funny
shes
angry
s
narration
re
sigh
irritated
afraid
heroine
ugh
crazy
gorgeous
okay
men
listened
beautiful
woman
heartbreaking
romance
cry
heck
club
hated
starsi
omg
wouldve
favourite
kn
wouldnt
posted
adorable
yeah
stupid
intrigued
confused
tears
swear
volume
rings
quality
tolkien
narrative
storytelling
success
indeed
prose
structure
fine
robert
effort
level
satisfying
exposition
scope
nature
larger
criticism
climax
volumes
game
social
attack
r
wheel
popular
grand
essentially
modern
suffers
create
material
thus
forces
battle
deus
plotting
hints
appear
massive
conclusion
rich
ex
key
provides
large
control
length
fourth
heavily
step
momentum
becomes
machina
piece
previous
vast
various
standard
credit
significant
works
hundred
opening
threads
entry
wife
generally
Avoided
Preferred
Craig's Zeta
Corpora


Kolimo_experiment
Craig's Zeta
markers
40
30
20
10
antimarkers
10
20
30
40
🟥 Goethe
🟩 OthersO


Kolimo_experiment
Craig's Zeta
markers
40
35
30
25
20
15
10
antimarkers
5
10
15
20
25
30
35
🟥 Old
🟩 OthersO

Log-likelihood

Log-likelihood

- …is an hypotesis-based test

Log-likelihood

- …is an hypotesis-based test
- “[…] rather than two groups of texts characterized by different word rates, this hypothesis claims that there is, in fact, a single group. Words are examined one at a time; those words for which this hypothesis seems most wrong will be counted as distinctive” (Riddell 2015)

Log-likelihood formula

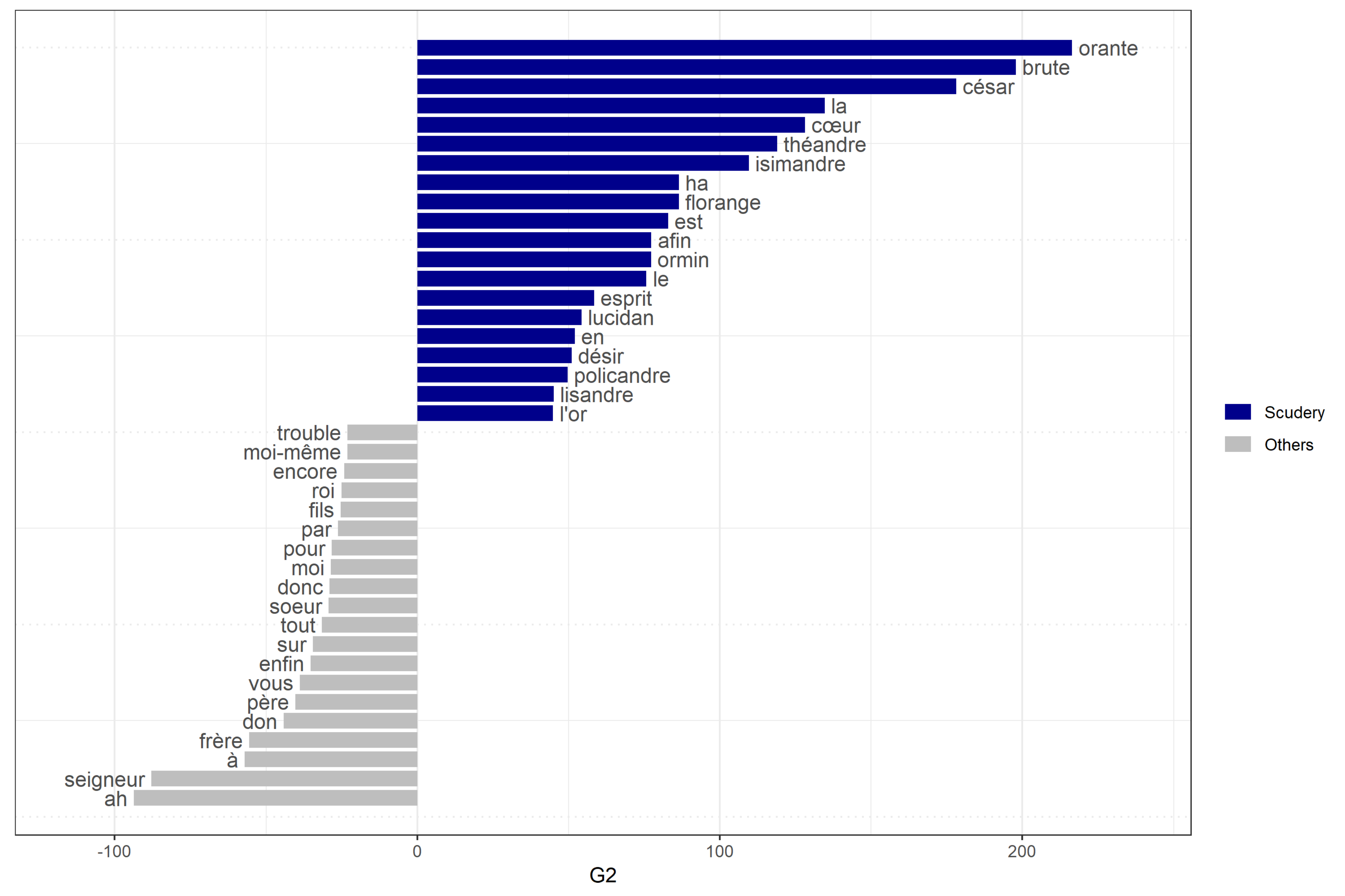
$$\sum_i O_i \times \ln \frac{O_i}{E_i}$$

O

	«fou»	Not «fou»
Text a	11	388592
Text b	96	445265

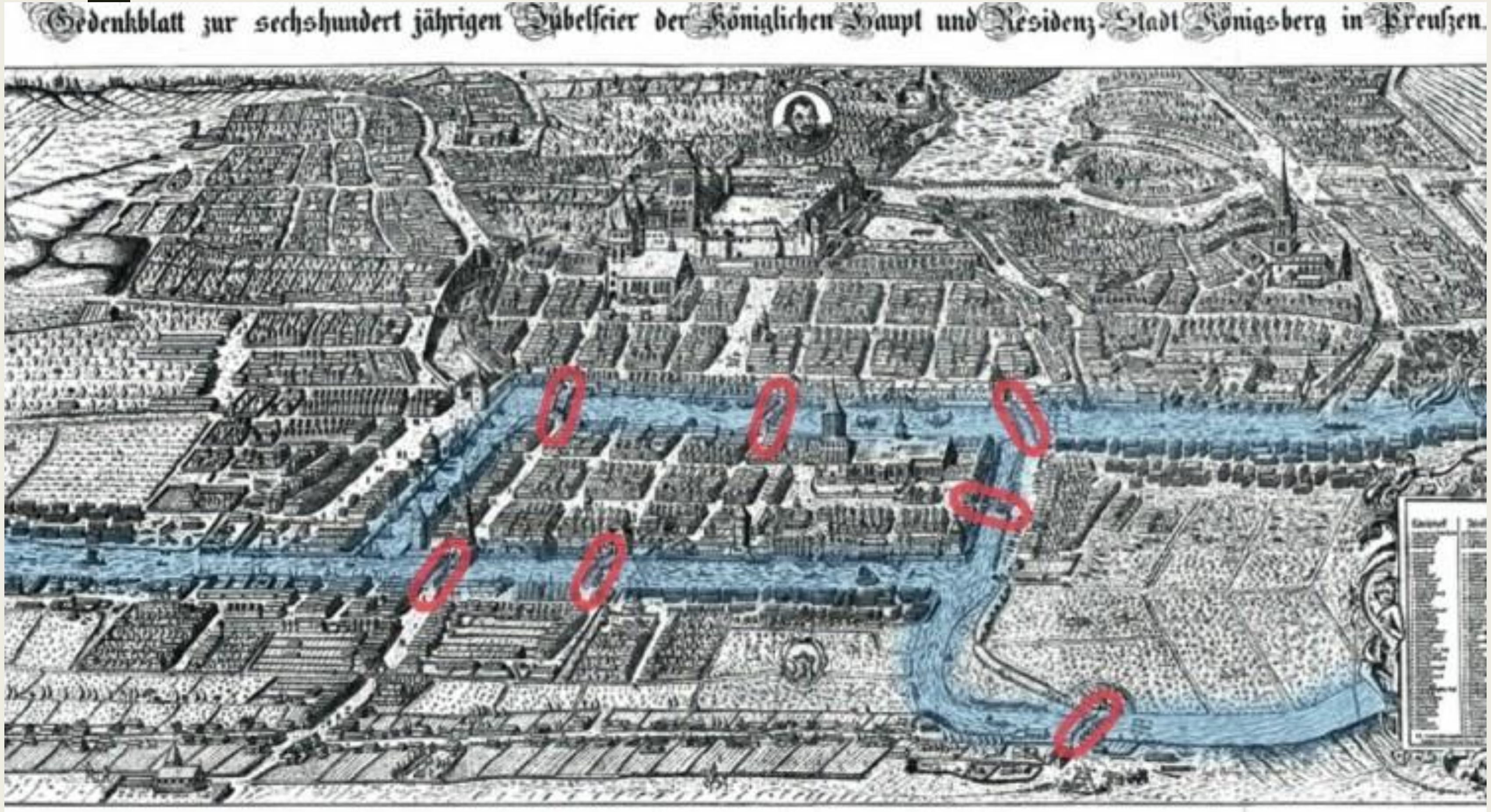
E

	«fou»	Not «fou»
Text a	48.06	388553
Text b	51.94	445303


orante
brute
césar
la
cœur
théandre
isimandre
ha
florange
est
afin
ormin
le
esprit
lucidan
en
désir
policandre
lisandre
l'or
trouble
moi-même
encore
roi
fils
par
pour
moi
donc
soeur
tout
sur
enfin
vous
père
don
frère
à
seigneur
ah
-100
0
100
200
G2
Scudery
Others

Network analysis

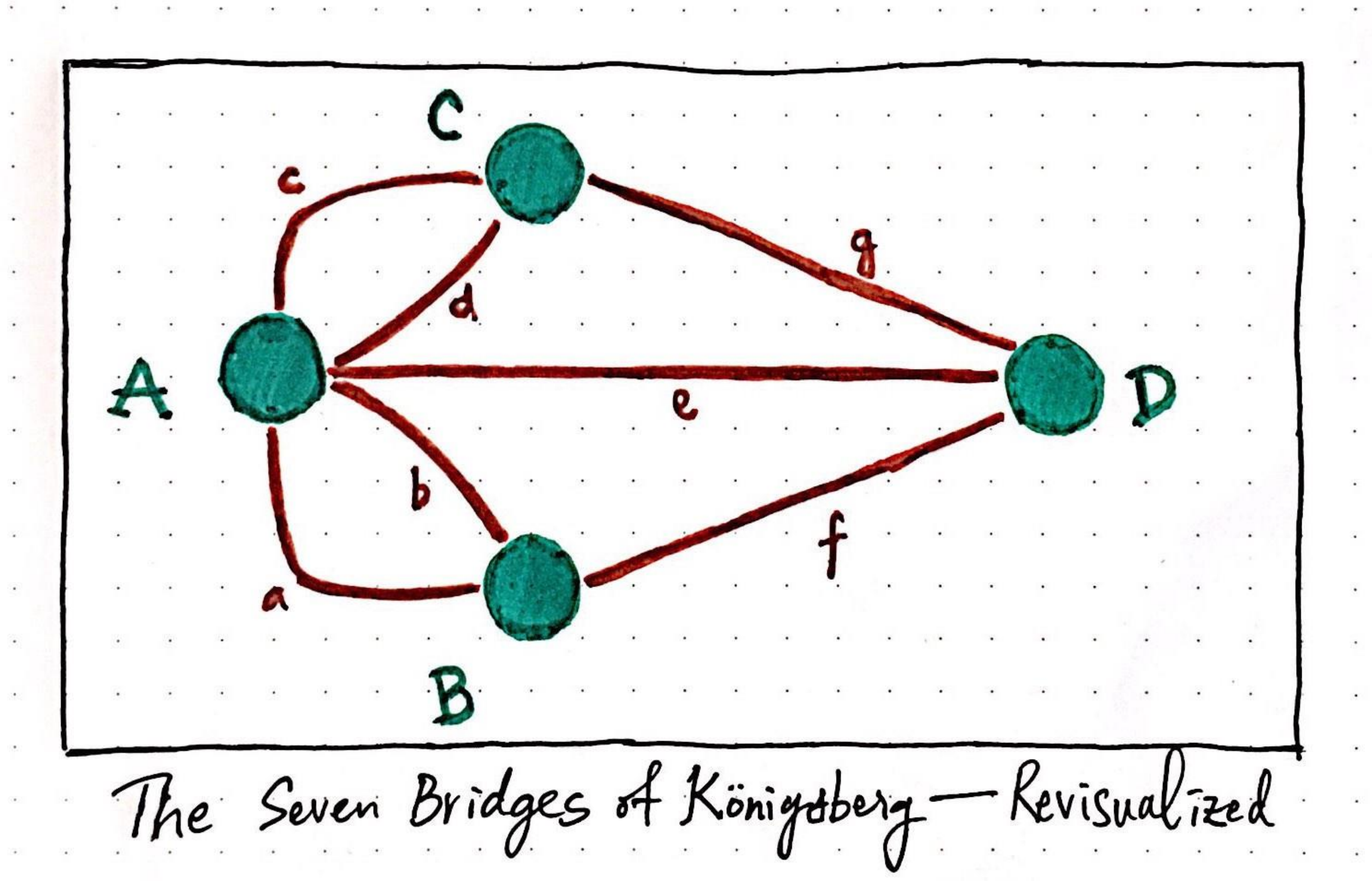
Seven Bridges of Königsberg


Gedenkblatt zur sechshundert jährigen Jubelfeier der Königlichen Haupt und Residenz-Stadt Königsberg in Preußen.

“The problem was to devise a walk through the city that would cross each of those bridges once and only once.
[...]
Euler proved that the problem has no solution.”

(Wikipedia)

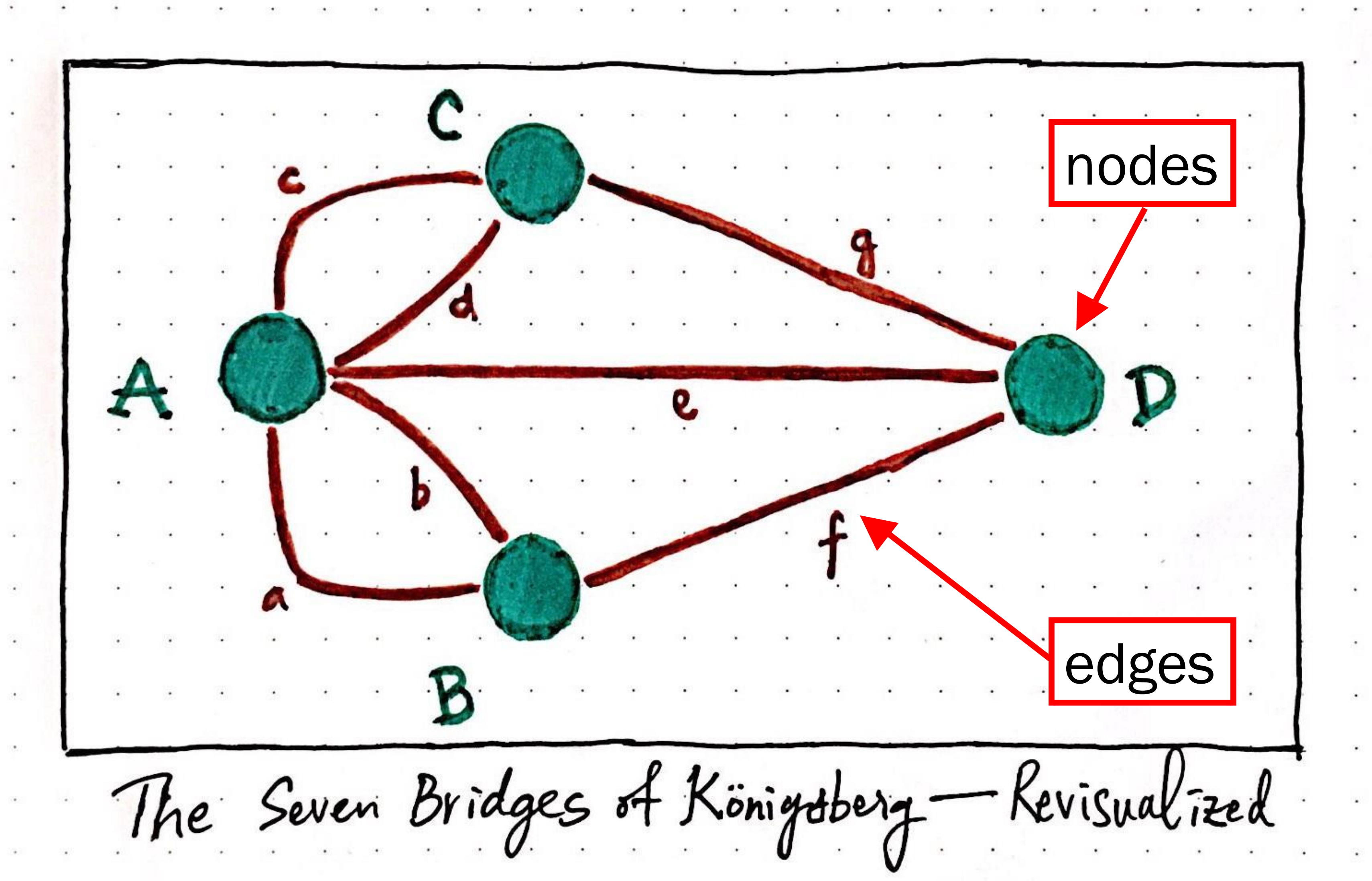
Seven Bridges of Königsberg



"The problem was to devise a walk through the city that would cross each of those bridges once and only once.
[...]
Euler proved that the problem has no solution."

(Wikipedia)

Seven Bridges of Königsberg

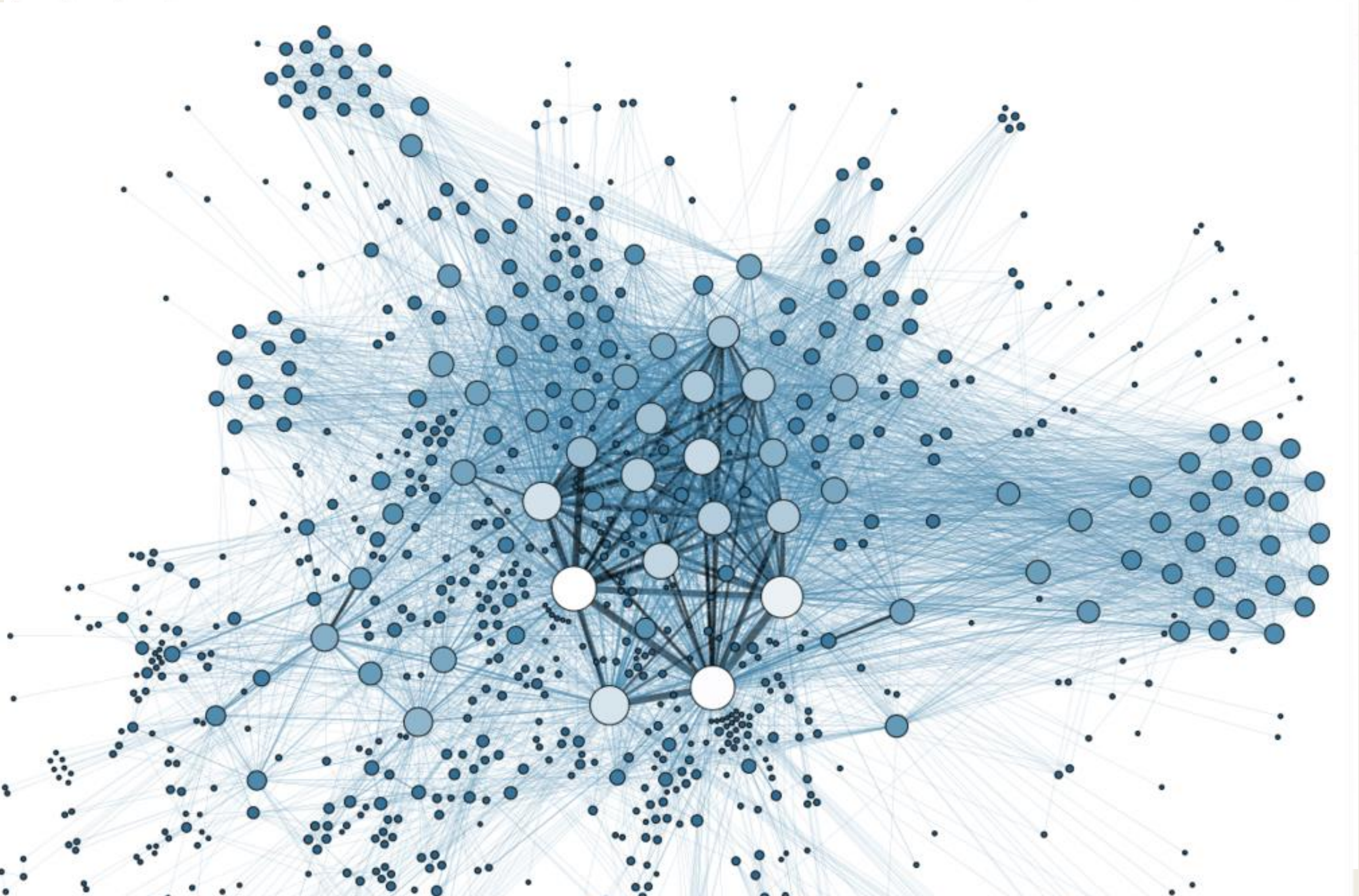


The Seven Bridges of Königsberg — Revisualized

“The problem was to devise a walk through the city that would cross each of those bridges once and only once.
[...]
Euler proved that the problem has no solution.”

(Wikipedia)

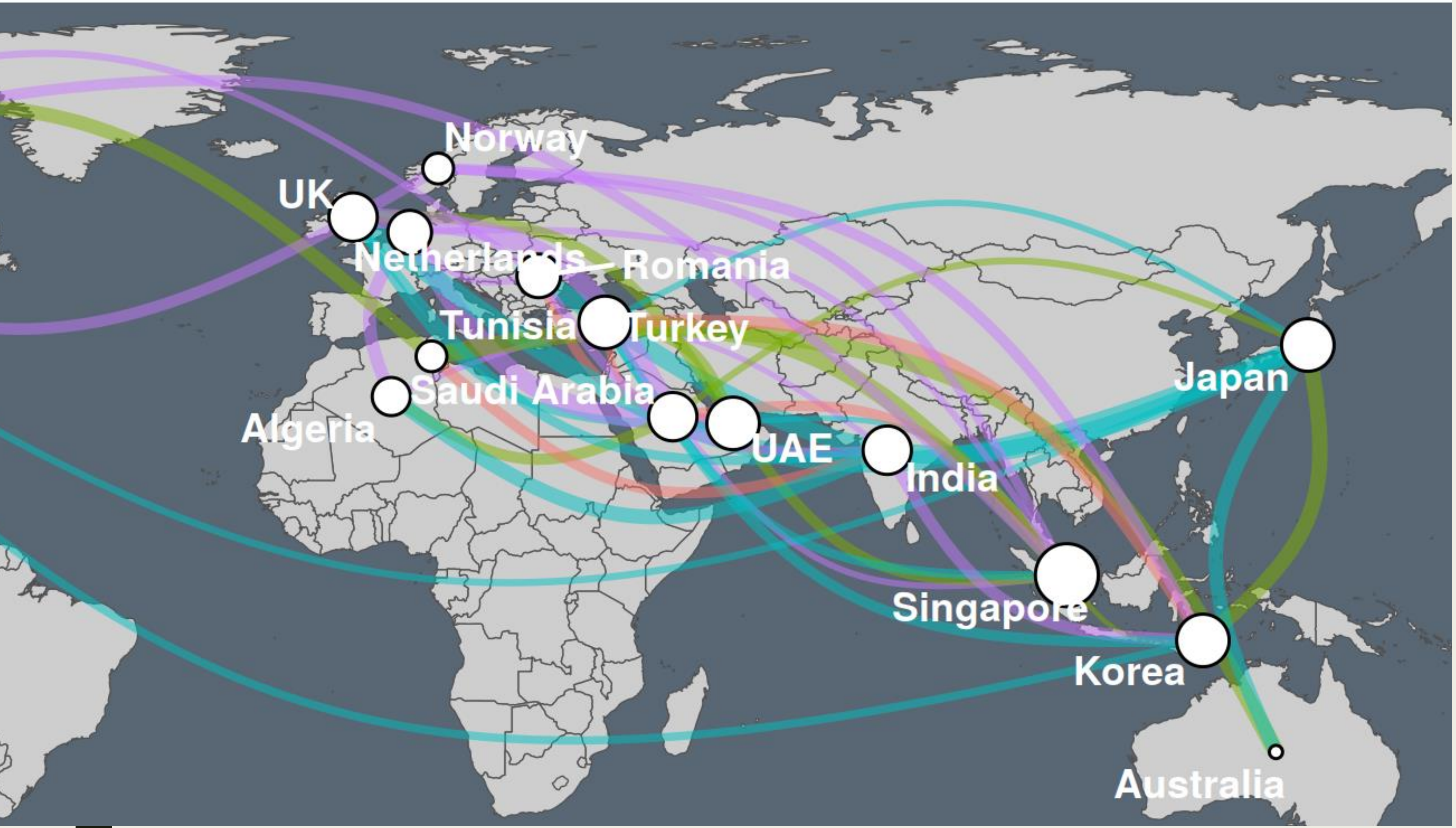
Seven Bridges of Königsberg



“The problem was to devise a walk through the city that would cross each of those bridges once and only once.
[...]
Euler proved that the problem has no solution.”

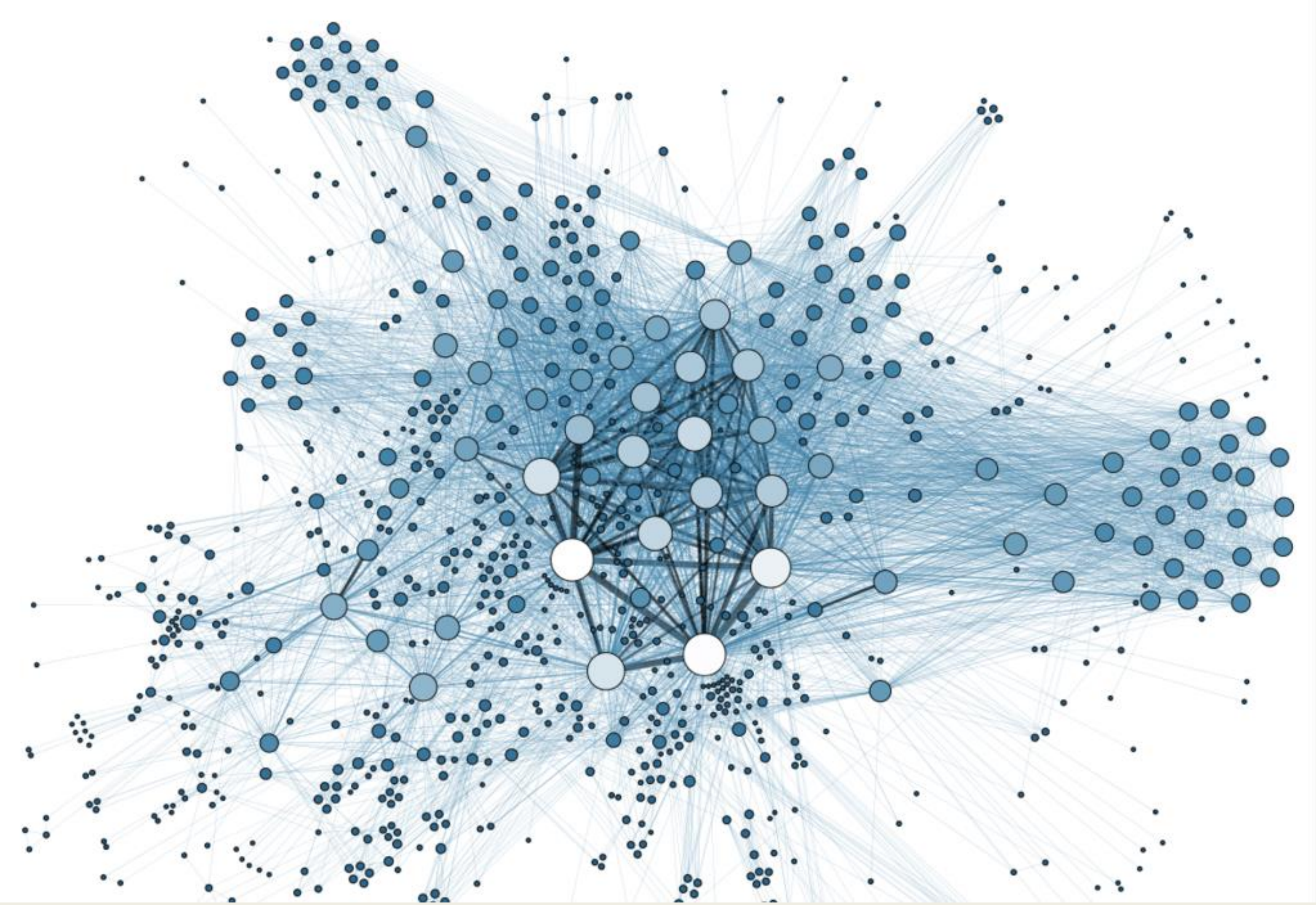
(Wikipedia)

Geo-coded networks



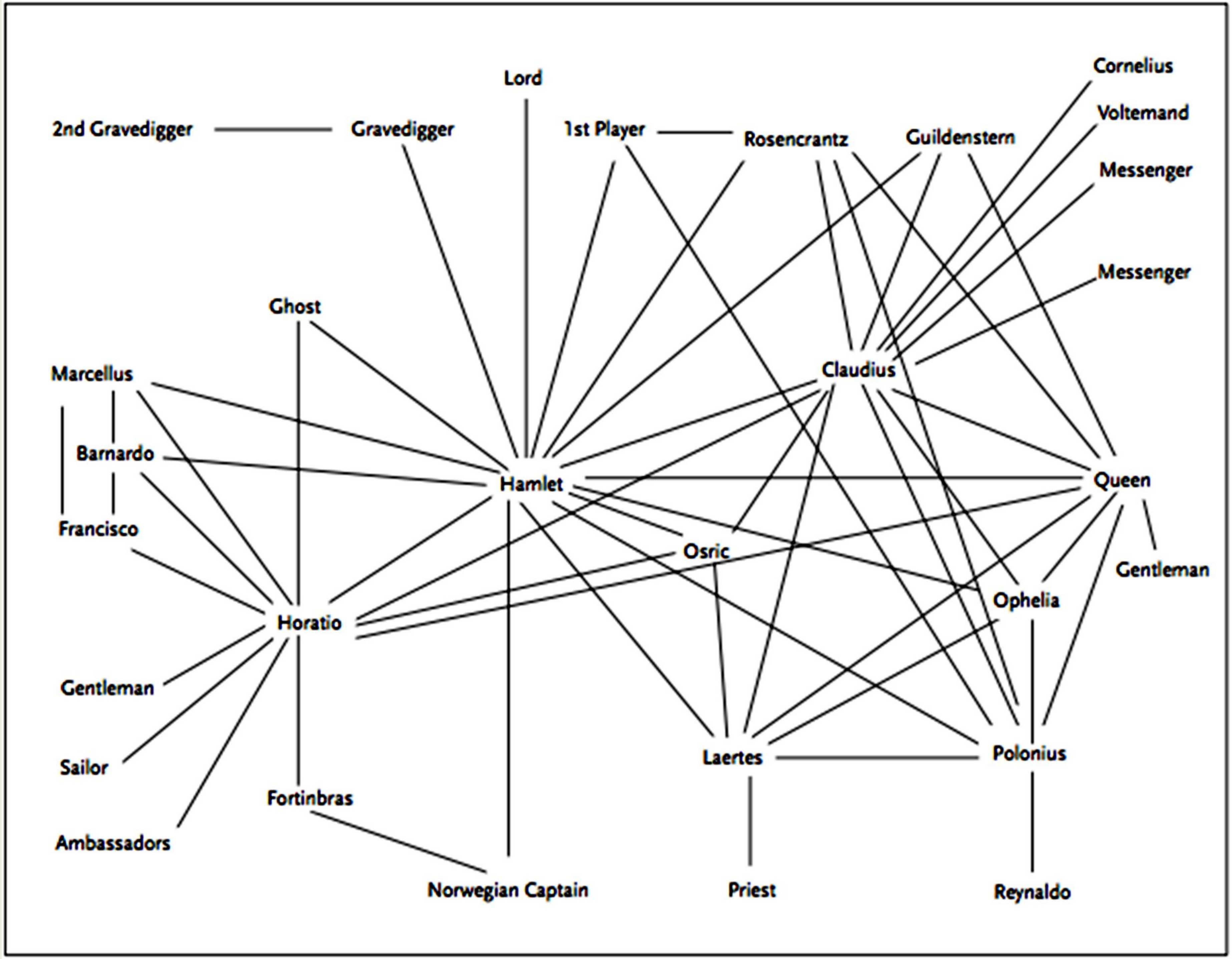
The position of the nodes is fixed on the map

«Bootstap» networks



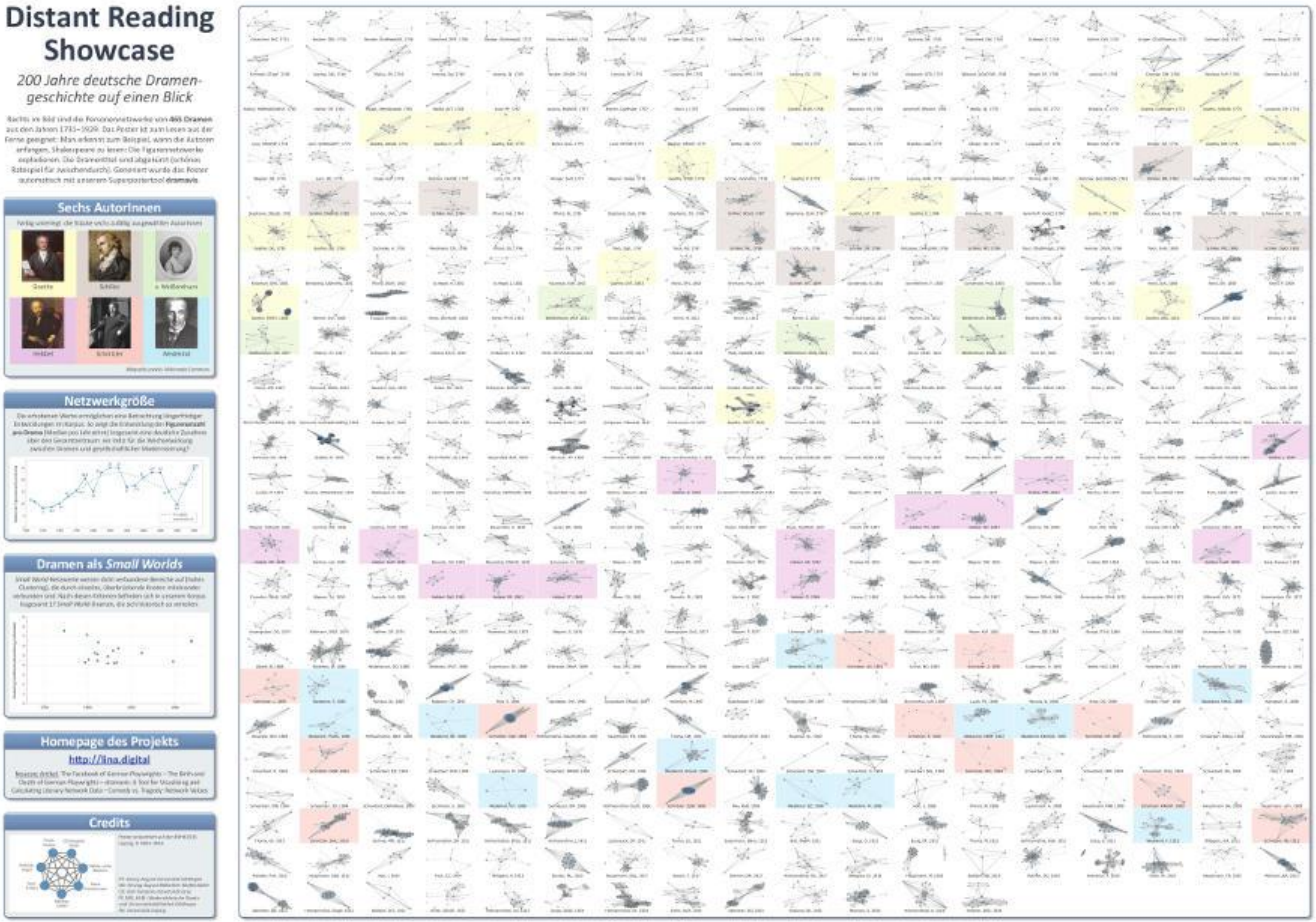
The position of the nodes is determined by the strength of their connections
(i.e. by the edge's «weight»)

Character Networks



(cf. Moretti 2011)

Network Analysis of 200 years of (German) theater



Distant Reading Showcase

200 Jahre deutsche Dramengeschichte auf einen Blick

Rechts im Bild sind die Personennetzwerke von 465 Dramen aus den Jahren 1731–1929. Das Poster ist zum Lesen aus der Ferne geeignet: Man erkennt zum Beispiel, wann die Autoren anfangen, Shakespeare zu lesen: Die Figurennetzwerke explodieren. Die Dramentitel sind abgekürzt (schönes Ratespiel für zwischendurch). Generiert wurde das Poster automatisch mit unserem Superpostertool **dramavis**.

Sechs AutorInnen

Netzwerkgröße

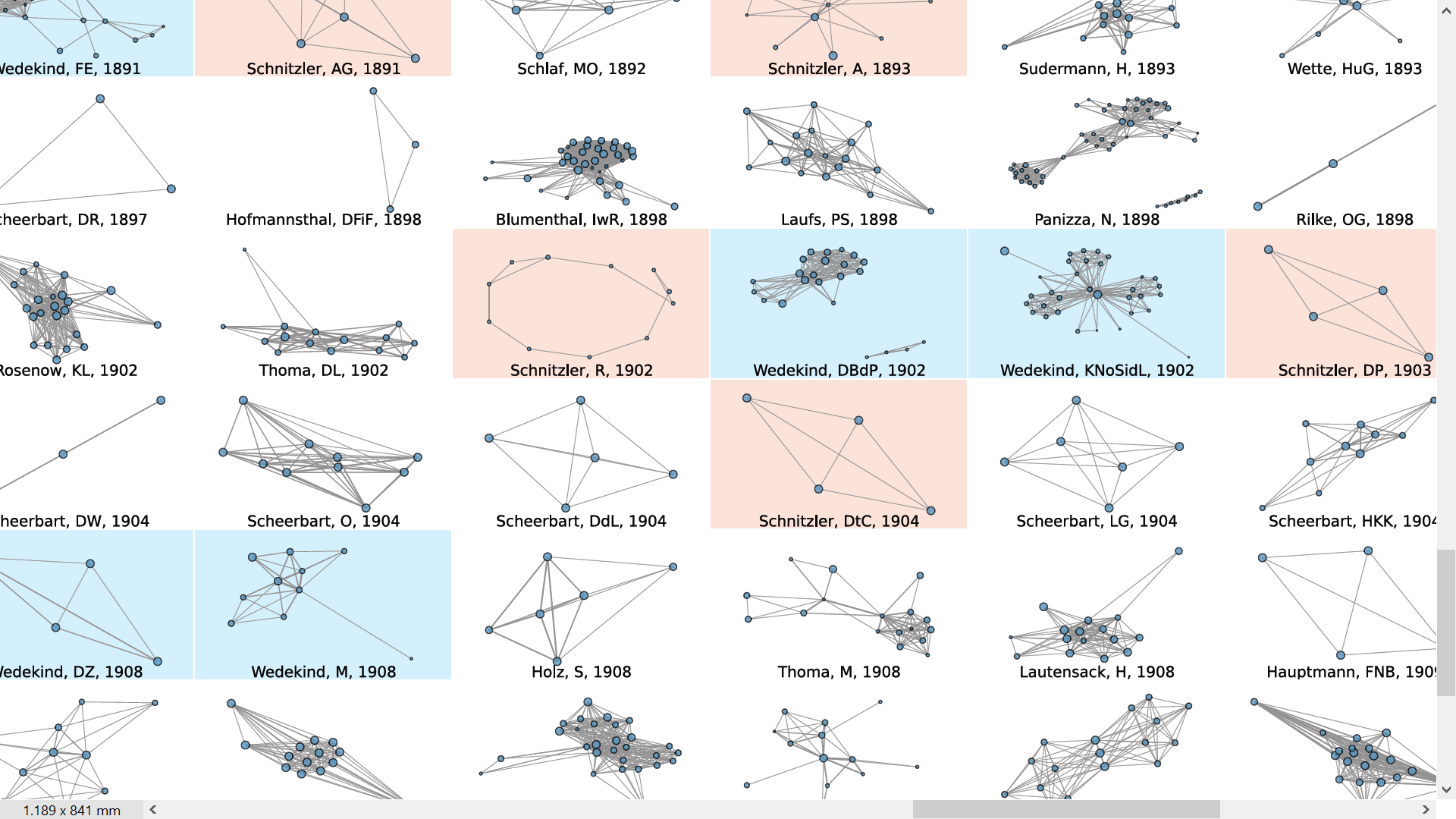
Dramen als *Small Worlds*

Homepage des Projekts

http://lina.digital

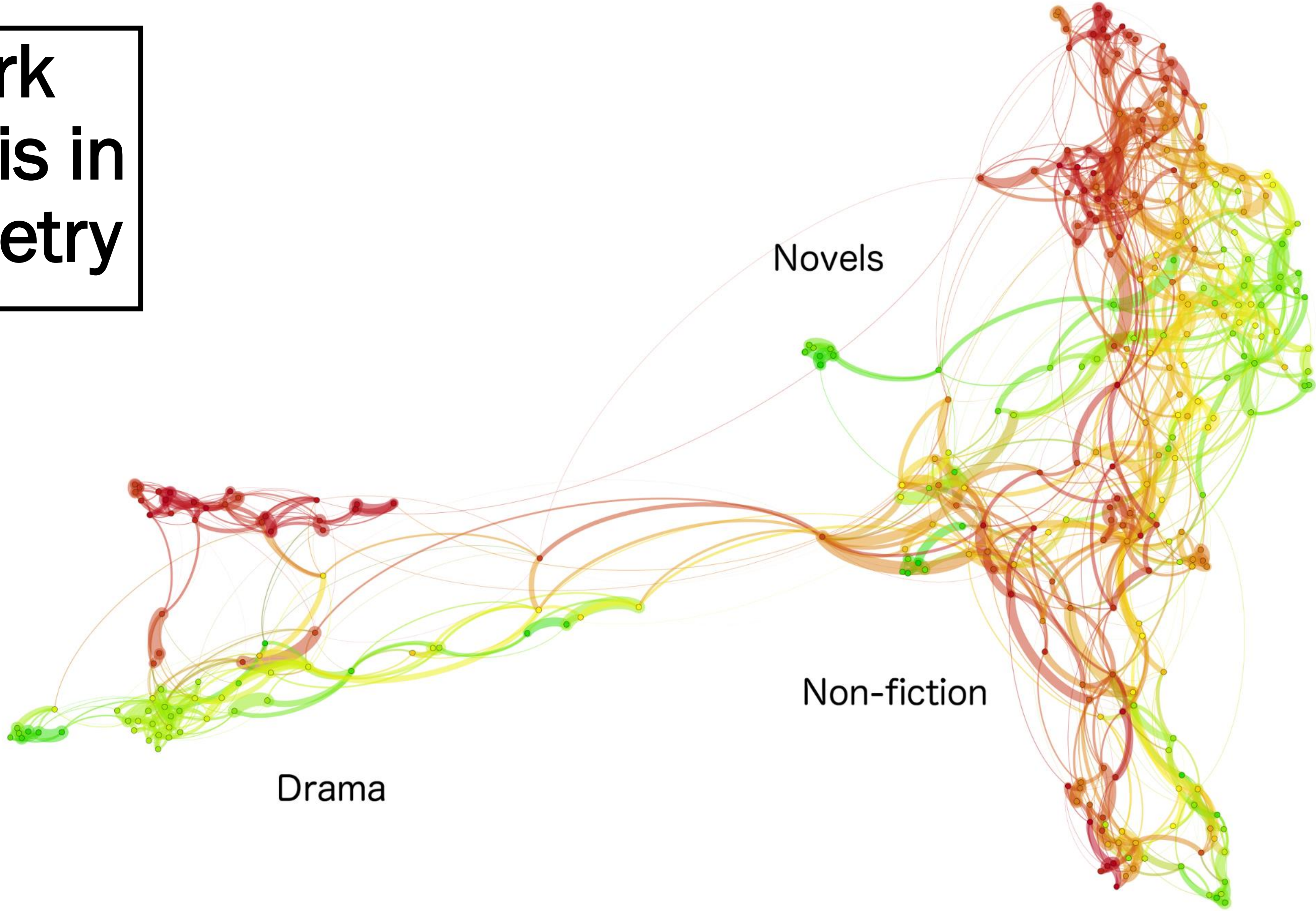
Credits

(Fischer et al. 2016)



Wedekind, FE, 1891
Schnitzler, AG, 1891
Schlaf, MO, 1892
Schnitzler, A, 1893
Sudermann, H, 1893
Wette, HuG, 1893
heerbart, DR, 1897
Hofmannsthal, DFiF, 1898
Blumenthal, IwR, 1898
Laufs, PS, 1898
Panizza, N, 1898
Rilke, OG, 1898
Rosenow, KL, 1902
Thoma, DL, 1902
Schnitzler, R, 1902
Wedekind, DBdP, 1902
Wedekind, KNoSidL, 1902
Schnitzler, DP, 1903
heerbart, DW, 1904
Scheerbart, O, 1904
Scheerbart, DdL, 1904
Schnitzler, DtC, 1904
Scheerbart, LG, 1904
Scheerbart, HKK, 1904
Wedekind, DZ, 1908
Wedekind, M, 1908
Holz, S, 1908
Thoma, M, 1908
Lautensack, H, 1908
Hauptmann, FNB, 190
1.189 x 841 mm

Network Analysis in stylometry


Novels
Non-fiction
Drama