

Computational text analysis: Survival guide

ESU 24 @ Cluj-Napoca
Jeremi Ochab, Artjoms Šeļa

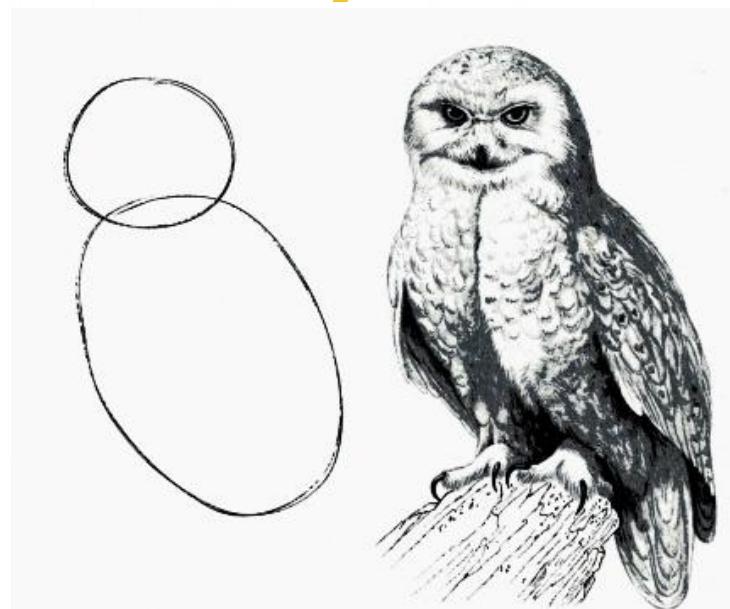


Fig 1. Draw two circles

Fig 2. Draw the rest of the damn Owl



Is there a path between C. Bronte and H. Melville in Manhattan?

- How to define your own space, populate it with texts and calculate routes between them?



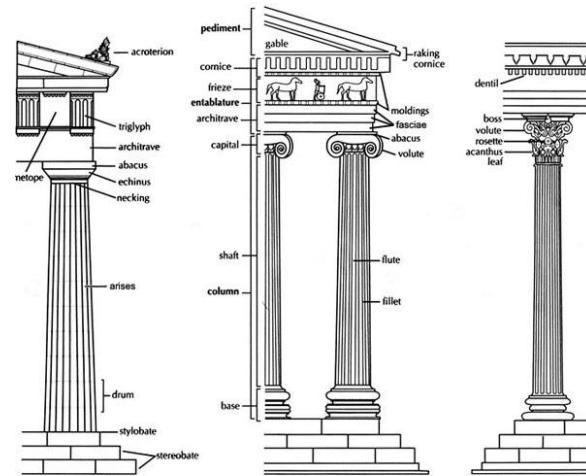
Is there a path between C.
Bronte and H. Melville in
Manhattan?

What is stylometry?

- A sub-field of computational text analysis that studies **differences** between texts
- Lutosławski 1897: method of “measuring stylistic affinities”

*Don't mix up with another “stylometry” –
“the art of measuring columns”!*

Stylometrie, f., **stylometry**, the art of measuring
columns (Säulenmeßkunst).



Stylometry and the 19th c. positivism

- New Shakespeare Society in 1850s
- Dating Dialogues of Plato (**Scottish** and **continental** schools, W. Lutoslawski)
- T.C. Mendenhall (style and spectral analysis)
- Math branch in 20th c.: G. Y. Yule
- Major shift: Mosteller & Wallace 1963



T.C. Mendenhall (1841-1924)
The characteristic curves of composition (1887)

T.C. Mendenhall: word lengths

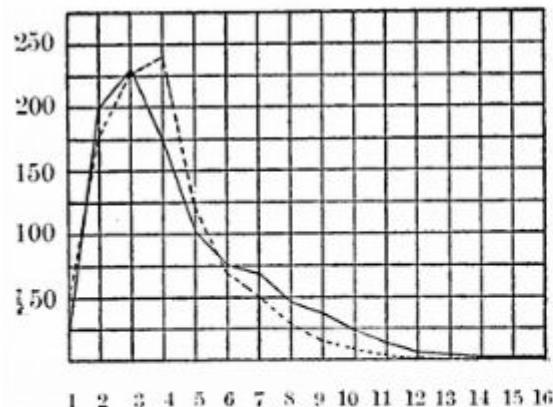


FIG. 1: Relative frequencies (per mille) of word-lengths measured by number of characters in works of W. Shakespeare (dashed) and F. Bacon (full line). Source: Mendenhall 1901: 104 (facsimile).

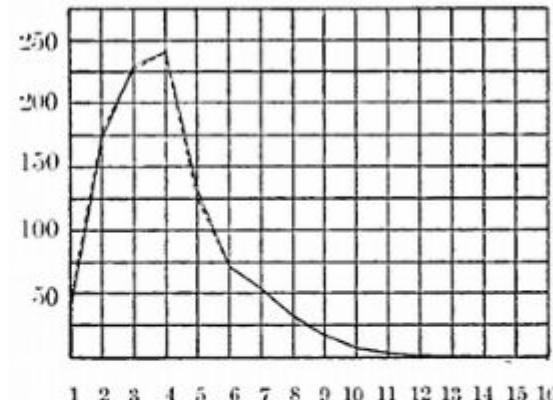
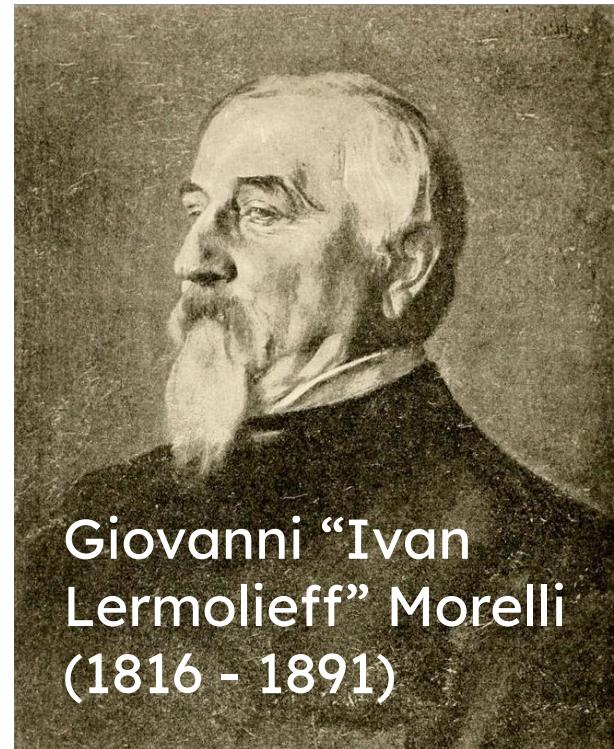
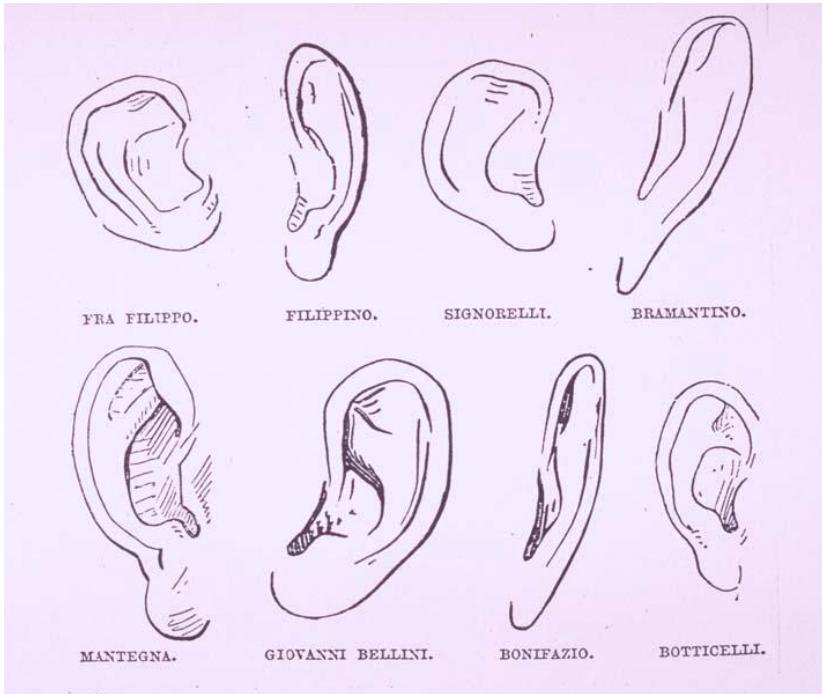


FIG. 2: Relative frequencies (per mille) of word-lengths measured by number of characters in works of W. Shakespeare (dashed) and C. Marlowe (full line). Source: Mendenhall 1901: 105 (facsimile).

Study of authorship in paintings

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



Giovanni "Ivan
Lermolieff" Morelli
(1816 - 1891)

“Anthropometrics” and fingerprints

- “Anthropometrics” of Alphonse Bertillon
- One dimension of measurements is not enough
- But combine, 2, 3...N, and individual *profiles* emerge



Alphonse Bertillon (1853-1914)

A model of text

- Differences between texts can be expressed in a multitude of ways
- Central question is **how to represent** a text so it could be placed on a quantitative scale? i.e. how to **model** it?
- Short answer: all representations are ‘wrong’, but some are useful (or more useful than others)

A model of text

- Differences between texts can be expressed in a multitude of ways
- Central question is **how to represent** a text so it could be placed on a quantitative scale? I.e. how to **model** it?
- Short answer: all representations are ‘wrong’, but some are useful (or more useful than others)
 - Word frequencies?
 - Algorithmically inferred topics?
 - Part of Speech tags?
 - Networks of character connections?
 - Embeddings?
 - Sentiment scores?

Silly things: bags of words

Mr. Sherlock Holmes, who was usually very late in the mornings, save upon those not infrequent occasions when he was up all night, was seated at the breakfast table. I stood upon the hearth-rug and picked up the stick which our visitor had left behind him the night before. It was a fine, thick piece of wood, bulbous-headed, of the sort which is known as a "Penang lawyer." Just under the head was a broad silver band nearly an inch across. "To James Mortimer, M.R.C.S., from his friends of the C.C.H.," was engraved upon it, with the date "1884." It was just such a stick as the old-fashioned family practitioner used to carry – dignified, solid, and reassuring.

Silly things: bags of words

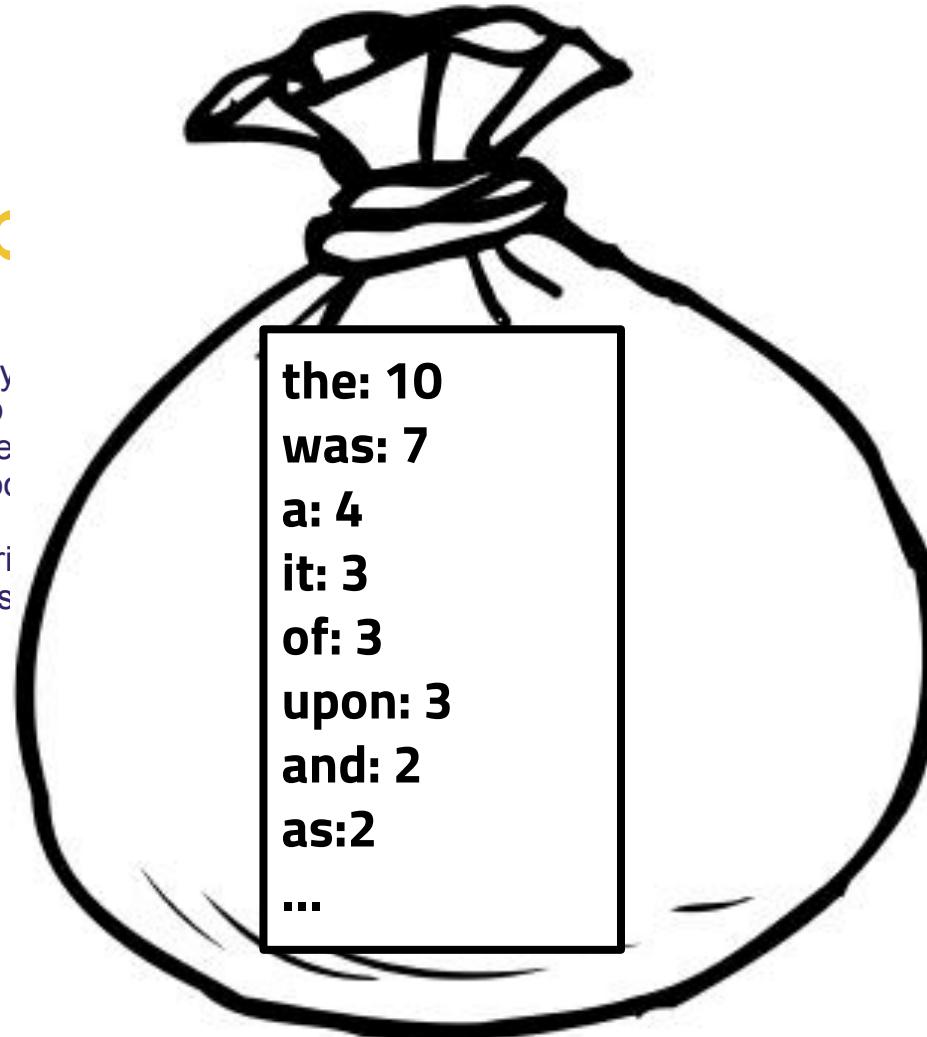
Mr. Sherlock Holmes, who w infrequent occasions when h upon the hearth-rug and pick before. It was a fine, thick pie "Penang lawyer." Just under James Mortimer, M.R.C.S., f date "1884." It was just such dignified, solid, and reassurir



he mornings, save upon those not seated at the breakfast table. I stood ur visitor had left behind him the night eaded, of the sort which is known as a silver band nearly an inch across. "To .C.H.," was engraved upon it, with the oned family practitioner used to carry –

Silly things: t

Mr. Sherlock Holmes, who was usually infrequent occasions when he was up upon the hearth-rug and picked up the before. It was a fine, thick piece of wood "Penang lawyer." Just under the head James Mortimer, M.R.C.S., from his friend date "1884." It was just such a stick as dignified, solid, and reassuring.



Silly things: bags of words

Mr. Sherlock Holmes, who w infrequent occasions when h upon the hearth-rug and pick before. It was a fine, thick pie "Penang lawyer." Just under James Mortimer, M.R.C.S., f date "1884." It was just such dignified, solid, and reassurir



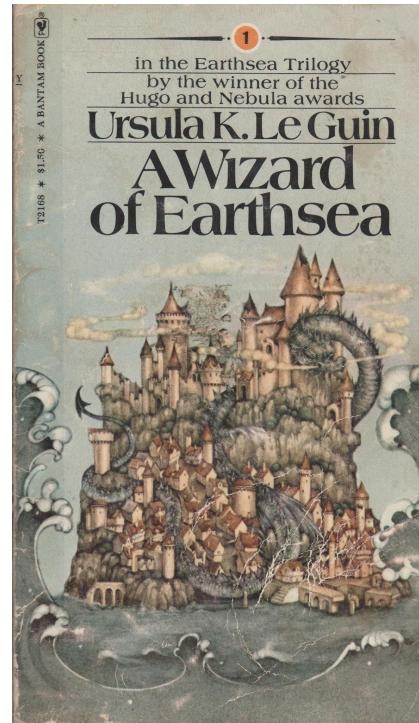
he mornings, save upon those not seated at the breakfast table. I stood ur visitor had left behind him the night eaded, of the sort which is known as a silver band nearly an inch across. "To .C.H.," was engraved upon it, with the oned family practitioner used to carry –

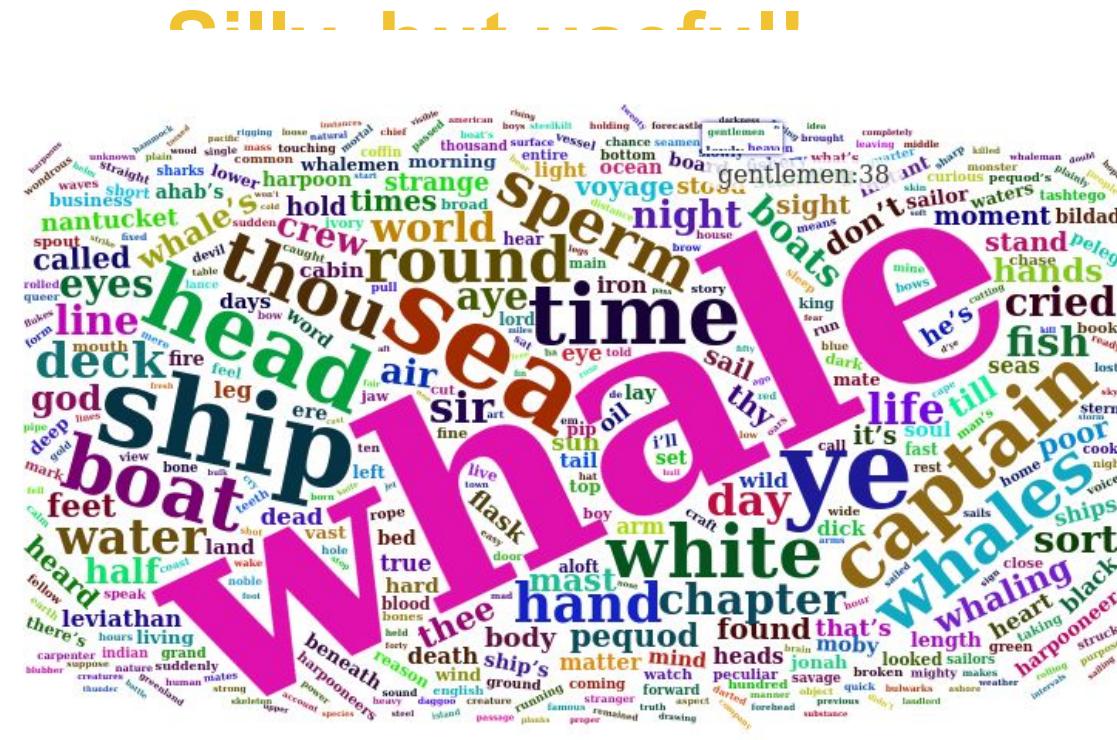
Doyle_Baskerville_p1: (10, 7, 4, 3, 3, 3, 2, 2,...)

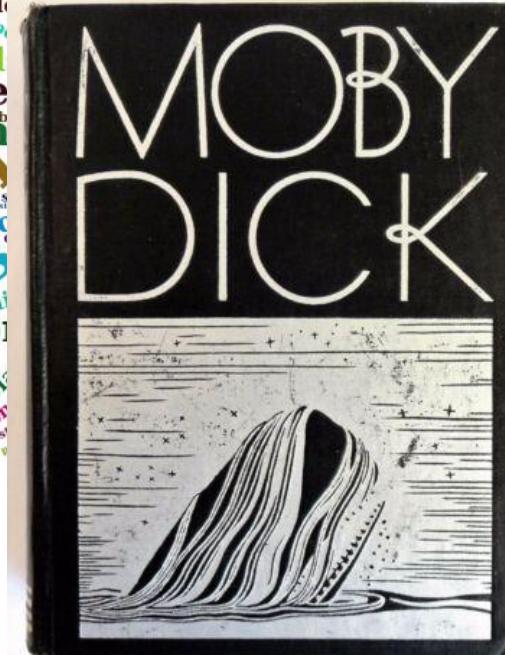
A colorful word cloud centered on the word "Sea". The words are arranged in a roughly circular pattern around the central word. The words and their associated meanings are:

- grey waves mage child learned
- wood house cold gont port
- true east jasper land north
- rain found sleep hills
- water spoke night
- sail lad run fell till boy red fire south
- dark power left set
- shadow lay lost
- spell sun rose evil call
- town dry ran sky town ran
- ogion boat day low hold
- days wind heart master
- time stood dead past staff fear
- names eyes serret voice looked door
- lord white sound friend mountain tower heard

grey waves mage child learned
wood house cold gont port
true east jasper land north
rain found sleep hills
water spoke night
sail lad run fell till boy red fire south
dark sun wise shadow lay lost
evil town ran dry sky spell sea light
ogion ran air court sat low hold
days wind boat day isle
heart stood master dead fear
time eyes past staff serret door
names stood voice tower heard
lord white sound friend mountain





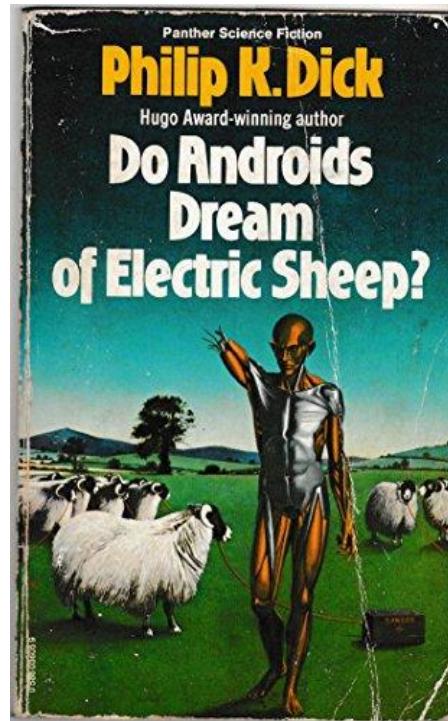


Silly, but useful!

A word cloud composed of numerous words in various colors (yellow, orange, red, green, blue, purple, pink, brown, black) and sizes. The words are arranged in a non-linear, overlapping pattern. Some words have small, faint text next to them, likely indicating a definition or part of speech. The words include: bottle, voice, miss, earth, kill, what's, late, deal, guess, mind, goat, garland, day, sat, set, ago, let's, legs, god, hear, hand, list, animal, found, roof, unit, held, san, feet, real, gray, cat, lot, test, car, die, sky, life, time, sir, tube, wait, tv, door, bounty, cage, job, scale, war, bed, talk, live, police, luft, feel, wife, table, owl, call, androids, left, false, animals, office, dead.

Silly, but useful!

bottle
voice miss late deal guess mind
what's sat goat garland day
set ago let's legs god hear list
animal found
feet real gray cat roof unit held san
life test car it'll
tv time sir tube
door war wait die sky
bed talk job cage
police bount job scale
left owl call androids wife live
false animals office luft feel table
dead



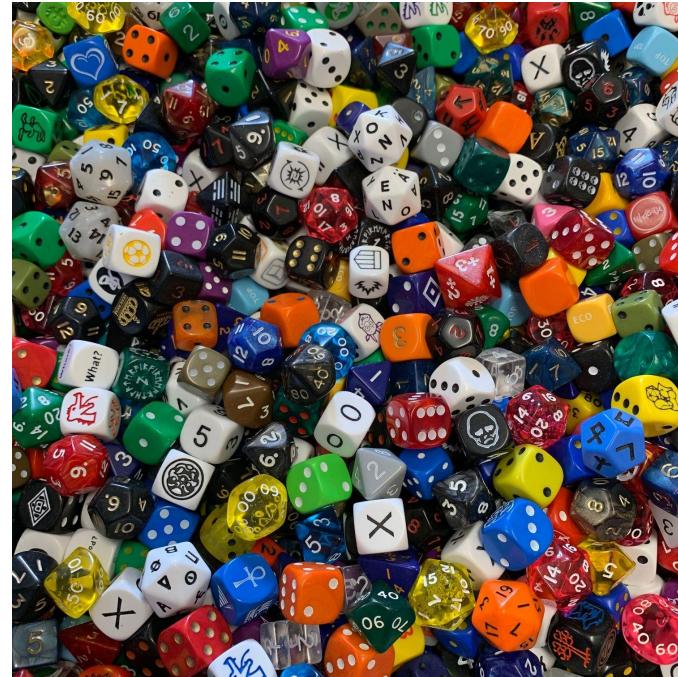
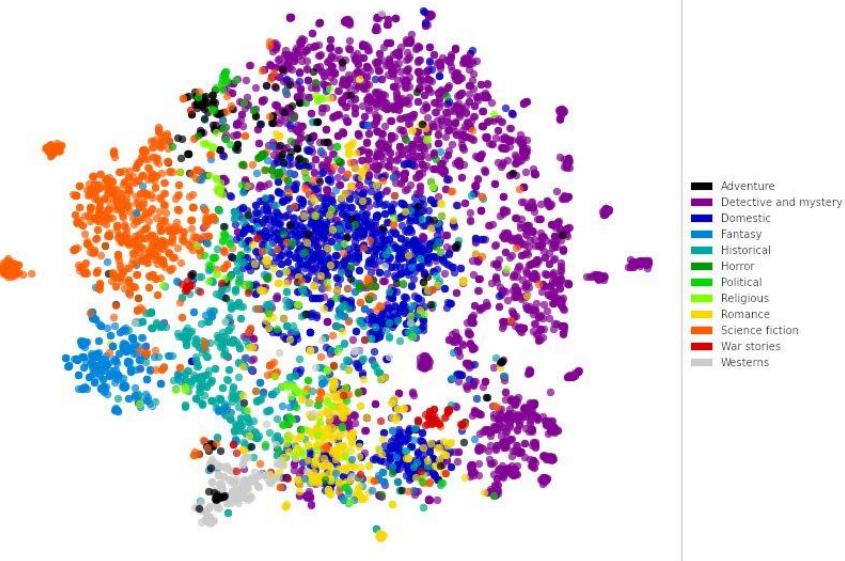
Proxies

- Word frequencies may serve as a proxy to **things we care about** in texts
- Word frequencies are the result of word choice -> word choice is a result of **forces that organize texts** (intention, cultural and social conditions)

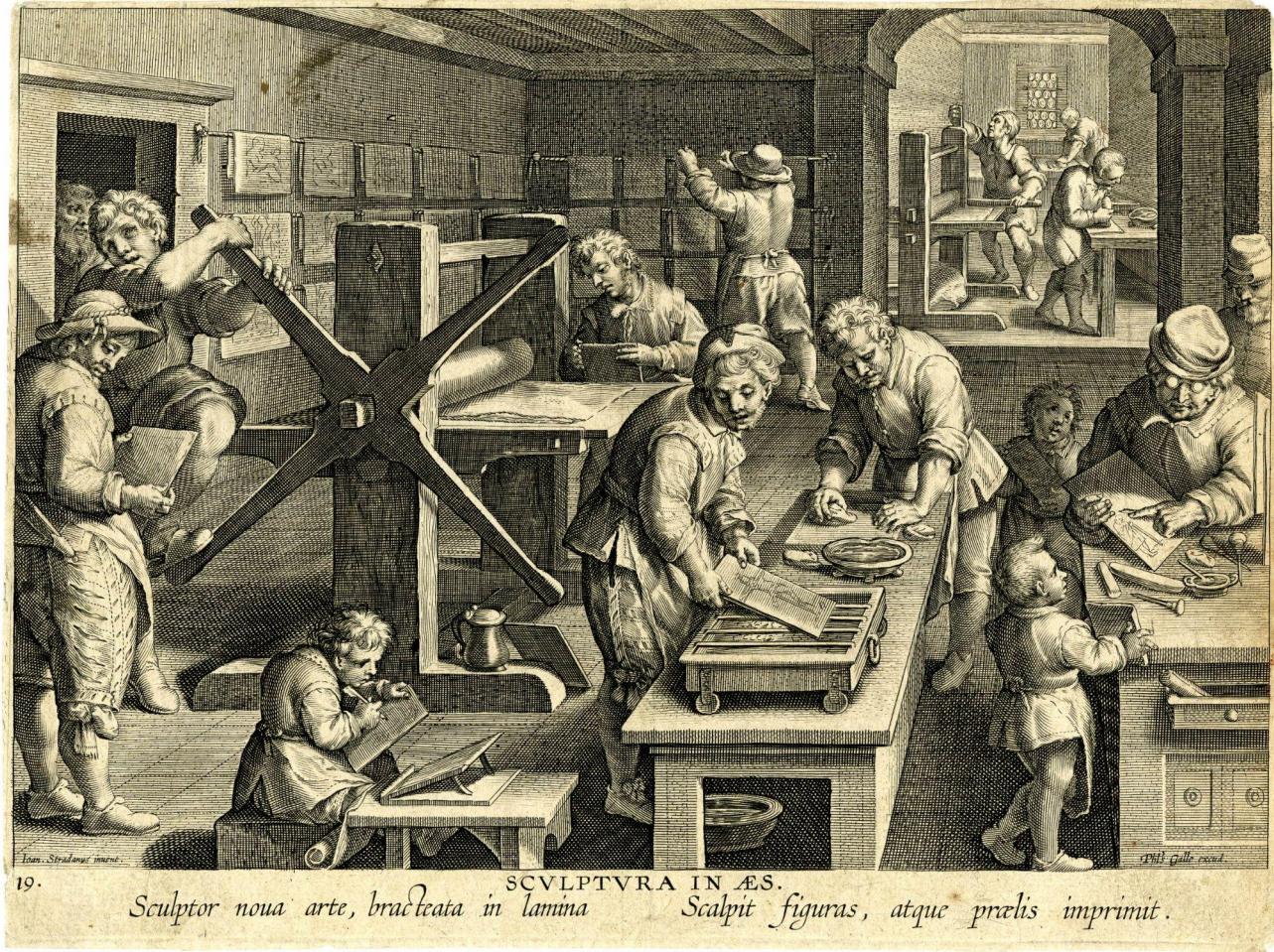


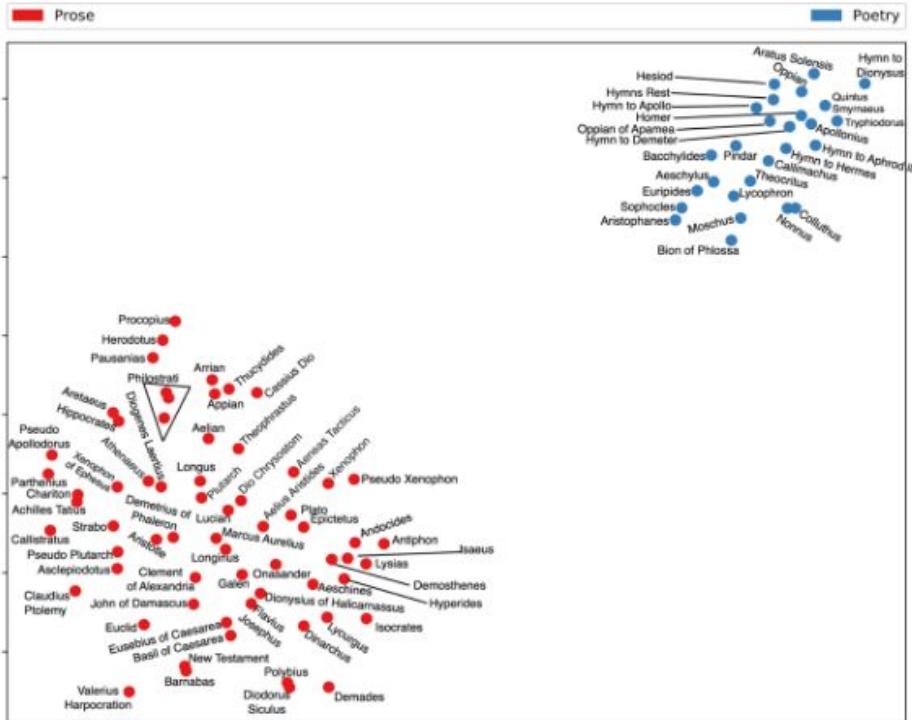
Text-dice rolling will be organized by multitude of forces

t-SNE Projection of 6431 American Novels, 1880-2000



Can we tell apart... **prose from poetry?**

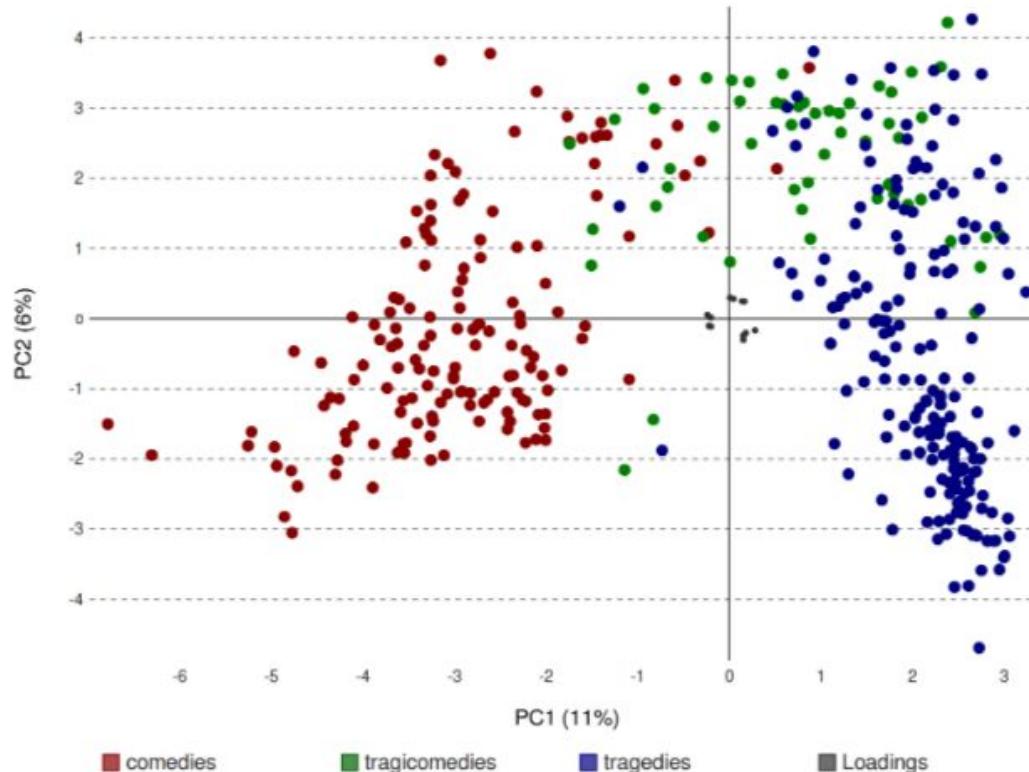




Storey & Mimno 2020: Like Two Pis in a Pod: Author Similarity Across Time in the Ancient Greek Corpus

**Can we tell
apart...
comedy
from
tragedy?**



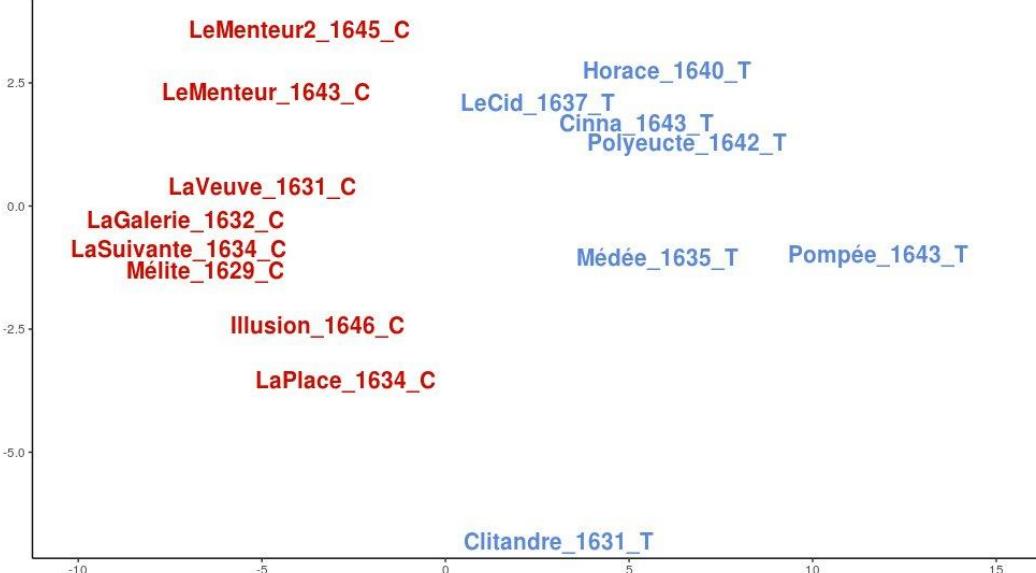


Schöch, C. 2017. Topic Modeling Genre...

Comedies and Tragedies of Pierre Corneille

Data come from large-scale quantitative study on distinctive features of classic dramatic genres in Corneille **done by Boris I. Yarkho in 1920s**.

Each text was represented across 15 features that Yarkho tried to synthesise into clear 'comedy' vs. 'tragedy' cut. This study served as a general demonstration of Yarkho's grand project of quantitative methodology for literary studies.
120 pages long work was first published only in 2006.

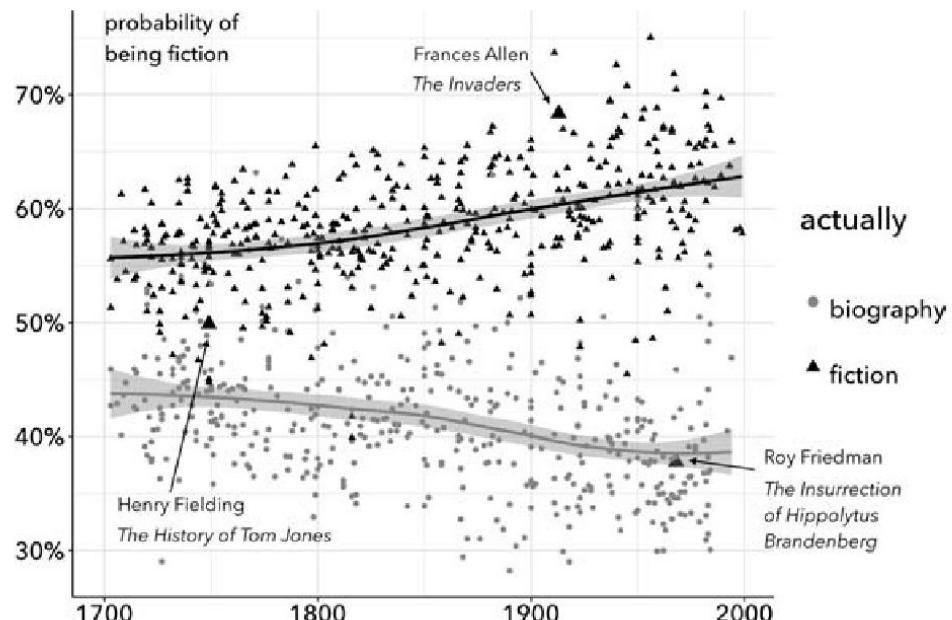
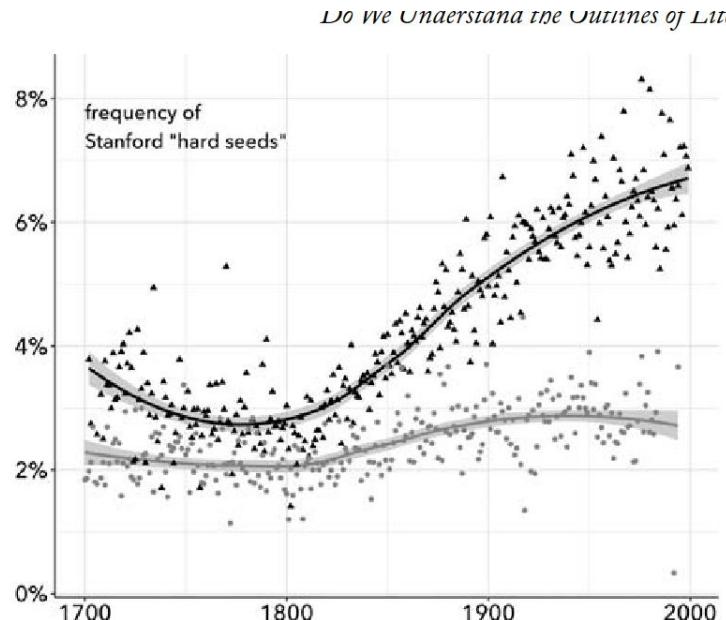


Boris Yarkho (1889-1942)

by @artjomshl 2020-07-01

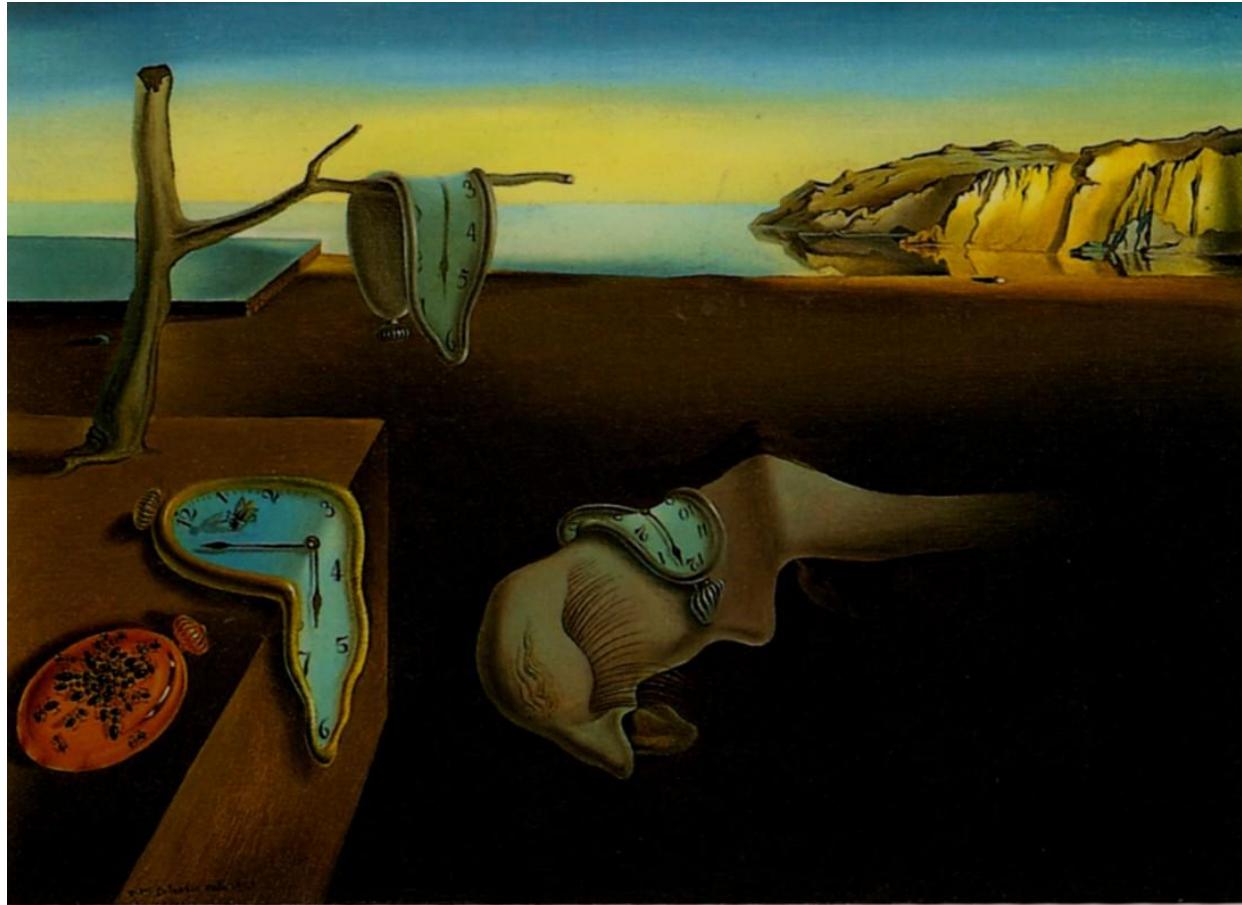
**Can we tell
apart...
Fiction from
non-fiction?**

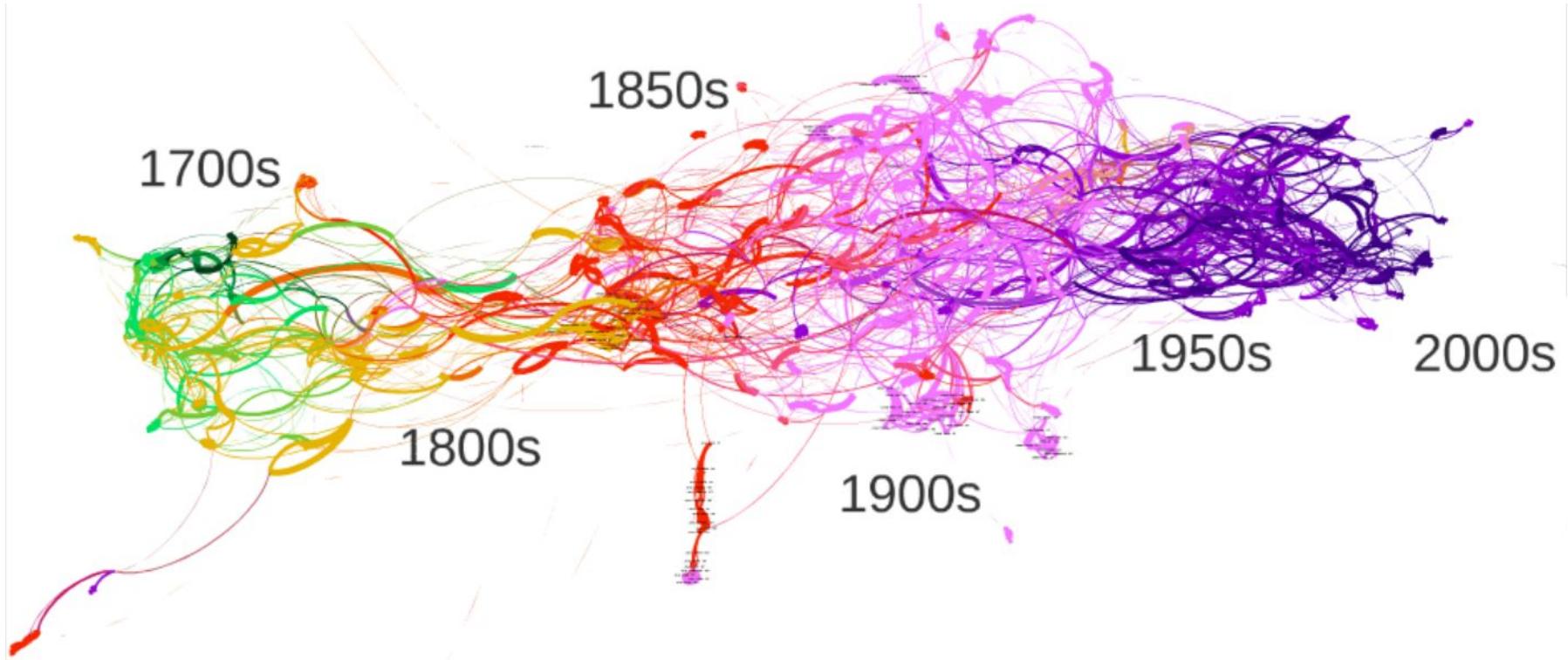




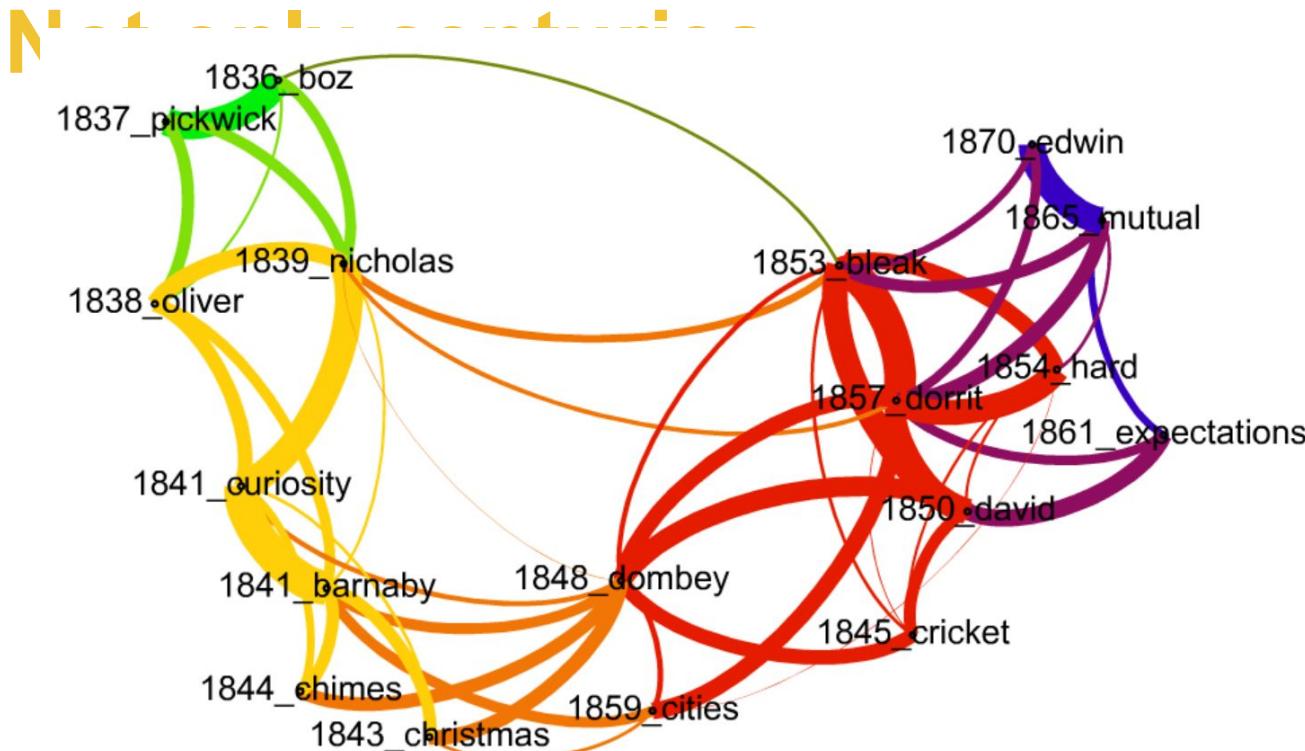
Can we tell
apart...

a 19th c. text
from 18th c.
text?





Rybicky 2016

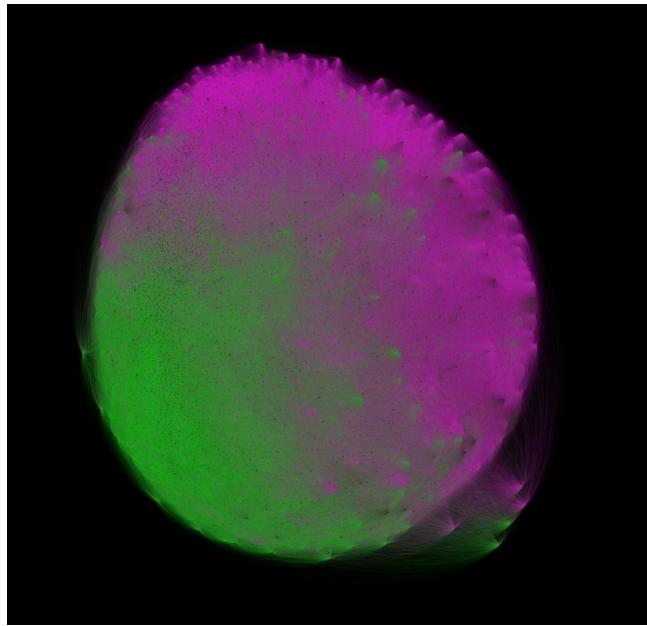


Rybicky 2016

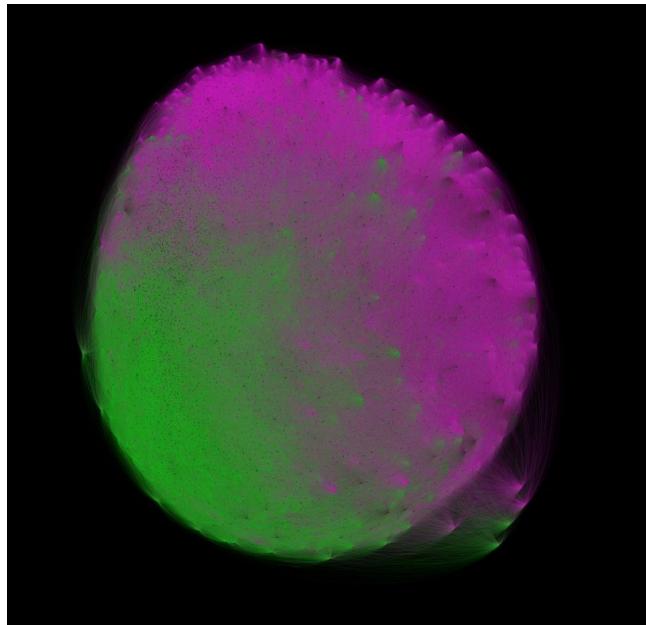
Can we tell
apart...

a text
written by a
woman from
text written
by a man?



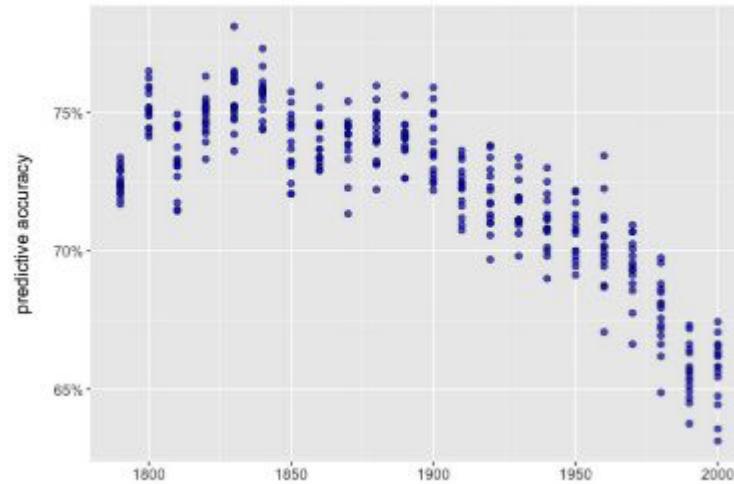


Jockers 2013 *Macroanalysis*



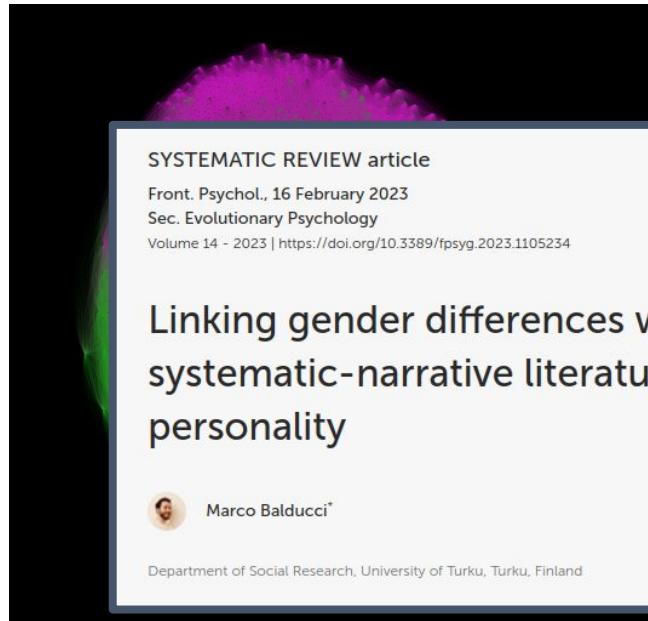
Jockers 2013 *Macroanalysis*

Accuracy of gender prediction, 1600-character samples

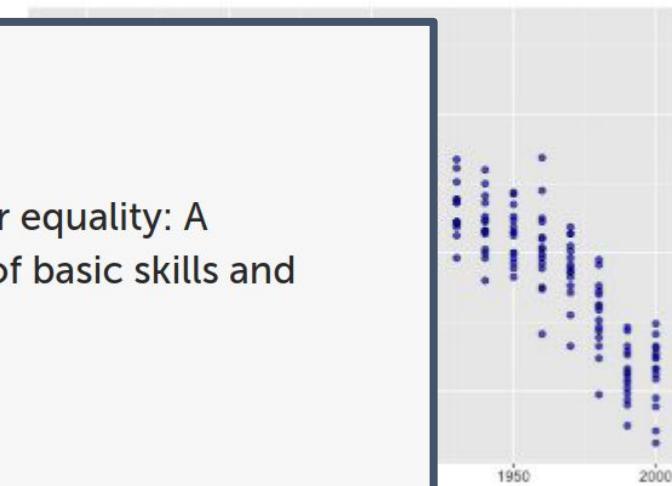


F/M characters recognizability decreases over time

Underwood et al. 2018



Accuracy of gender prediction, 1600-character samples



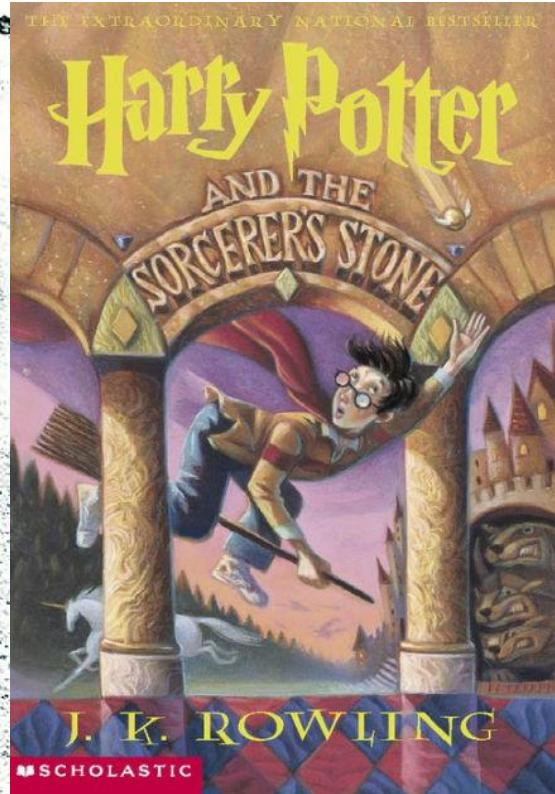
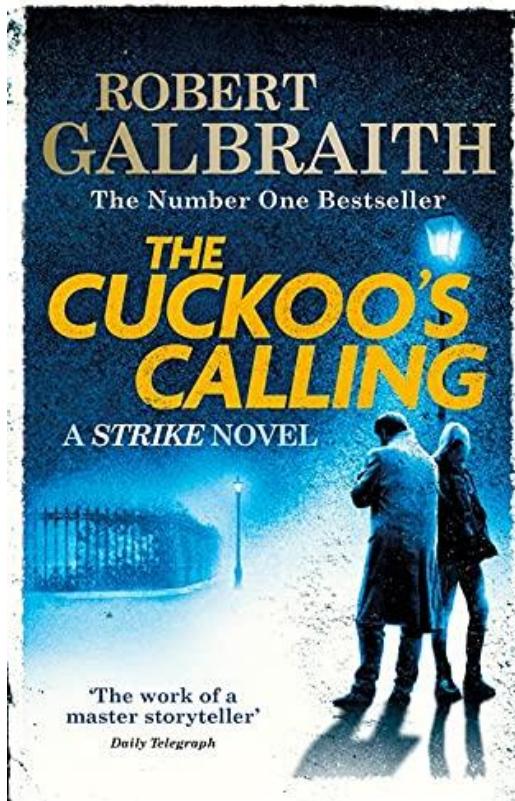
Jockers 2013 *Macroanalysis*

F/M characters recognizability decreases over time

Underwood et al. 2018

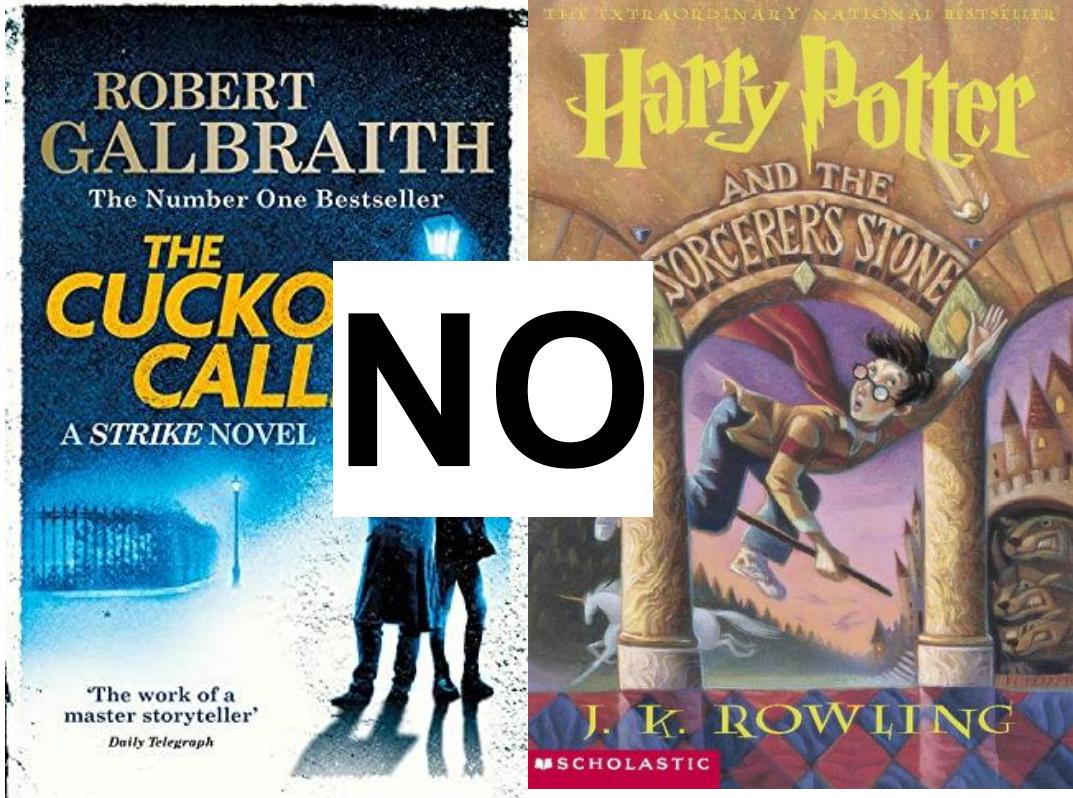
Can we tell
apart...

Robert
Galbraith
from J.K.
Rowling?



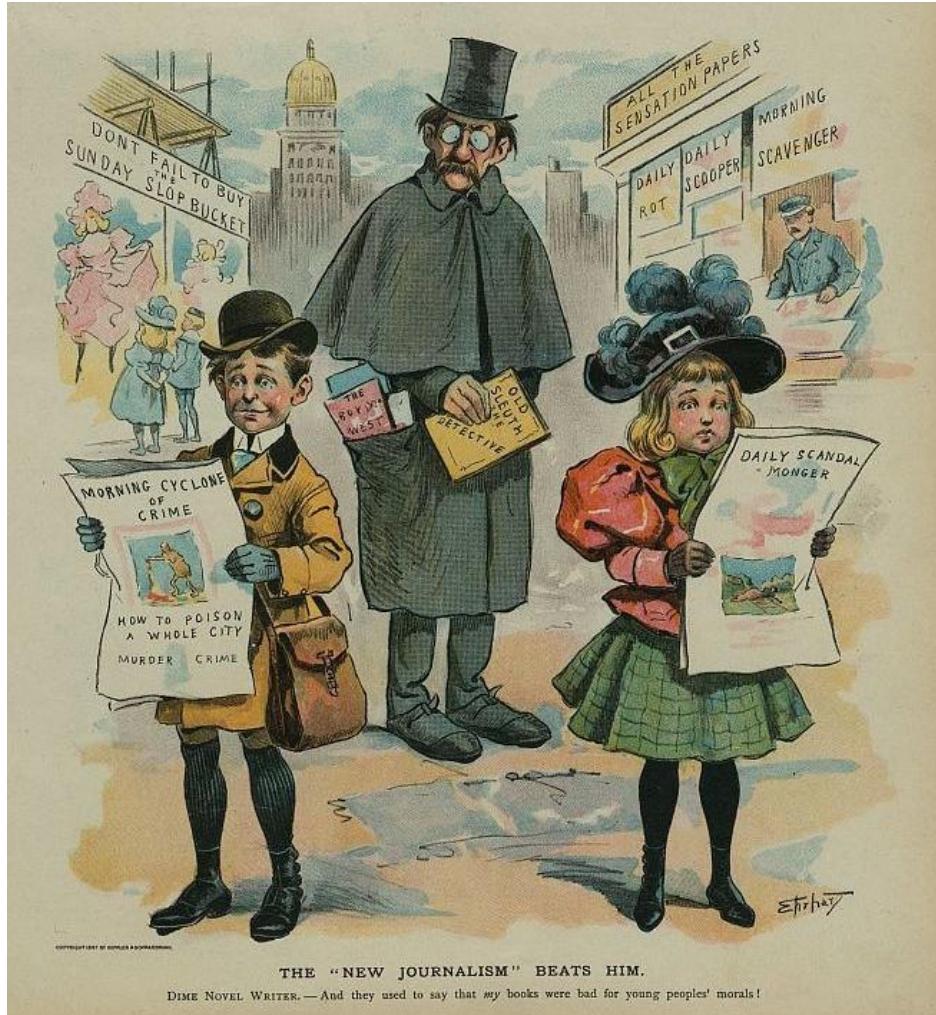
Can we tell
apart...

Robert
Galbraith
from J.K.
Rowling?



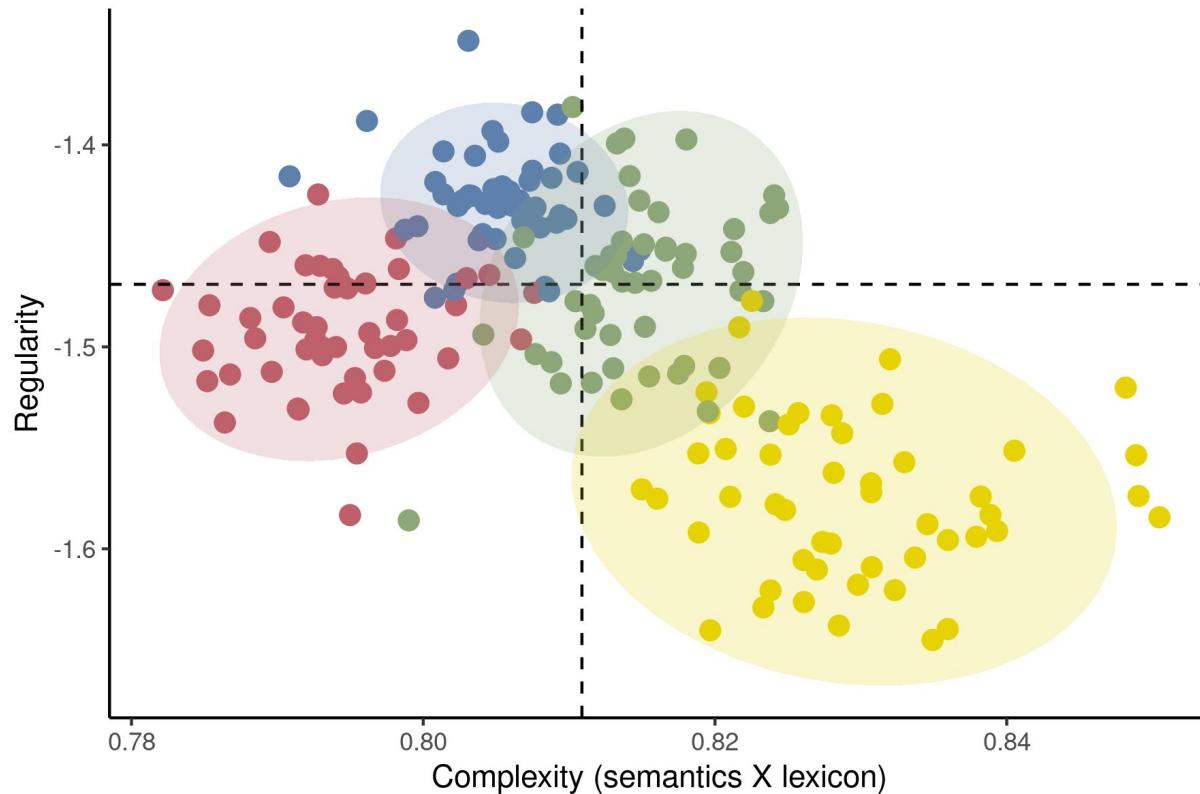
Can we tell apart...

Different writing groups (journals)?



Social production of form

Contemporary poetry samples from different social spaces / platforms



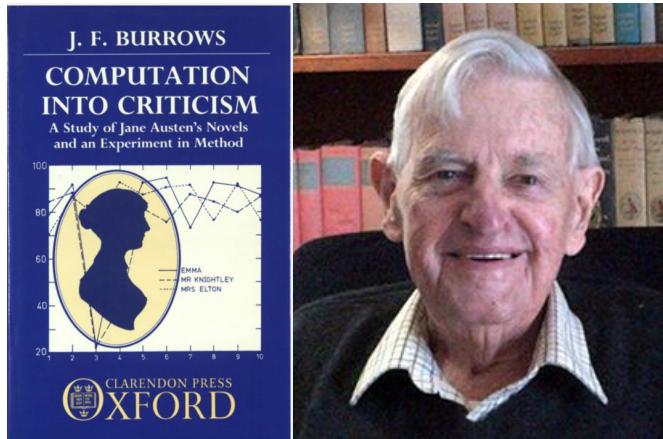
Curse & blessing of word frequencies

- **1. Authorship.** Two texts of the same author usually appear the closest to each other than to any outsider text
- **2. Modes of writing.** Fiction and nonfiction grew apart stylistically; Poetry books (esp. regularized verse) will always form a VERY exclusive party with themselves.
- **3. Genre.** Well-formed fiction genres (detective/mystery, sci-fi,
- **4. Chronology.** Global language change: Each generation of writers adopt slightly different version of language than the previous one (also: spelling conventions)
- **5. Social.** Be aware of a historical gap between women and men writing (socially constructed)
- ...

Burrows' Delta

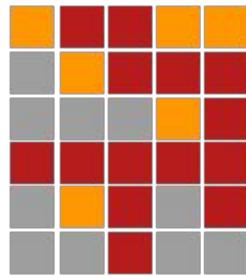
“Wealth of variables, many of which may be weak discriminators, almost always offer more tenable results than a smaller number of strong ones. [...] At all events, **a distinctive ‘stylistic signature’ is usually made up of many tiny strokes.**” (Burrows 2002)

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

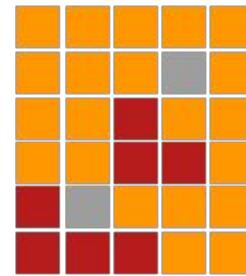


John Burrows (1928-2019)

TEXT 1

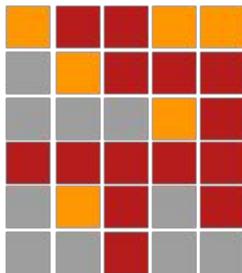


TEXT 2

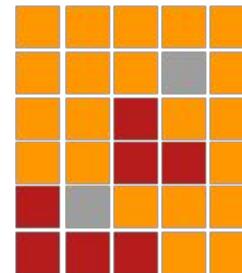


$$\Delta(T_1, T_2)$$

TEXT 1



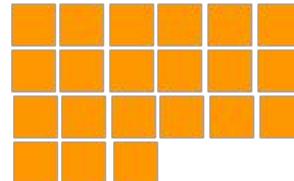
TEXT 2



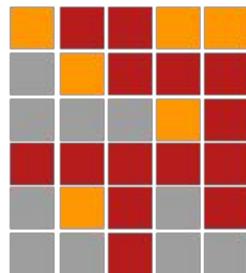
$$\Delta(T_1, T_2)$$



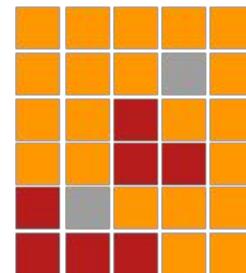
$$\Delta$$



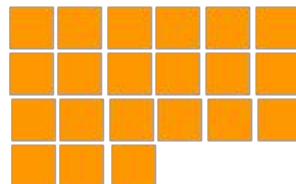
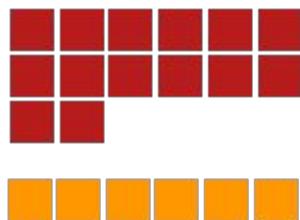
TEXT 1



TEXT 2



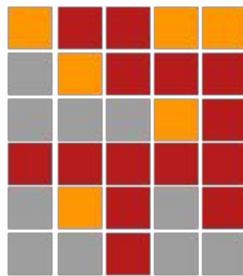
$$\Delta(T_1, T_2)$$



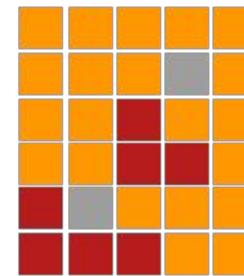
T1 [14, 6, 10]

T2 [7, 21, 2]

TEXT 1



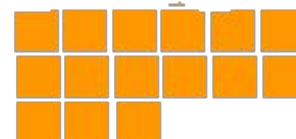
TEXT 2



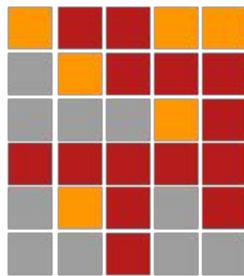
$$\Delta(T_1, T_2)$$



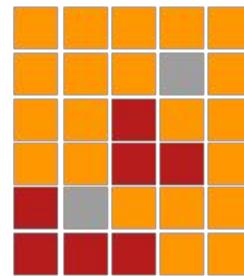
$$\Delta$$



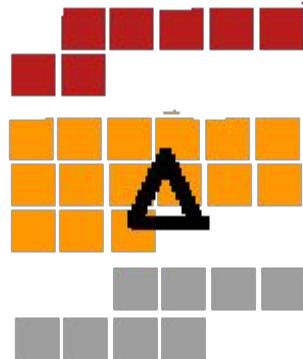
TEXT 1



TEXT 2

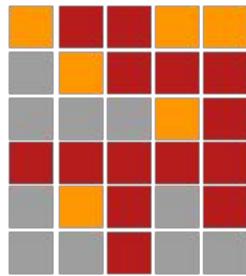


$$\Delta(T_1, T_2)$$

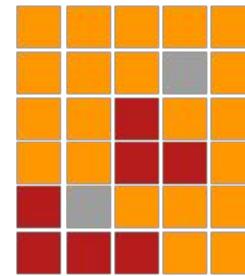


$$\Delta(T_1, T_2) = [6, 15, 10]$$

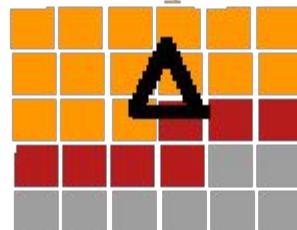
TEXT 1



TEXT 2



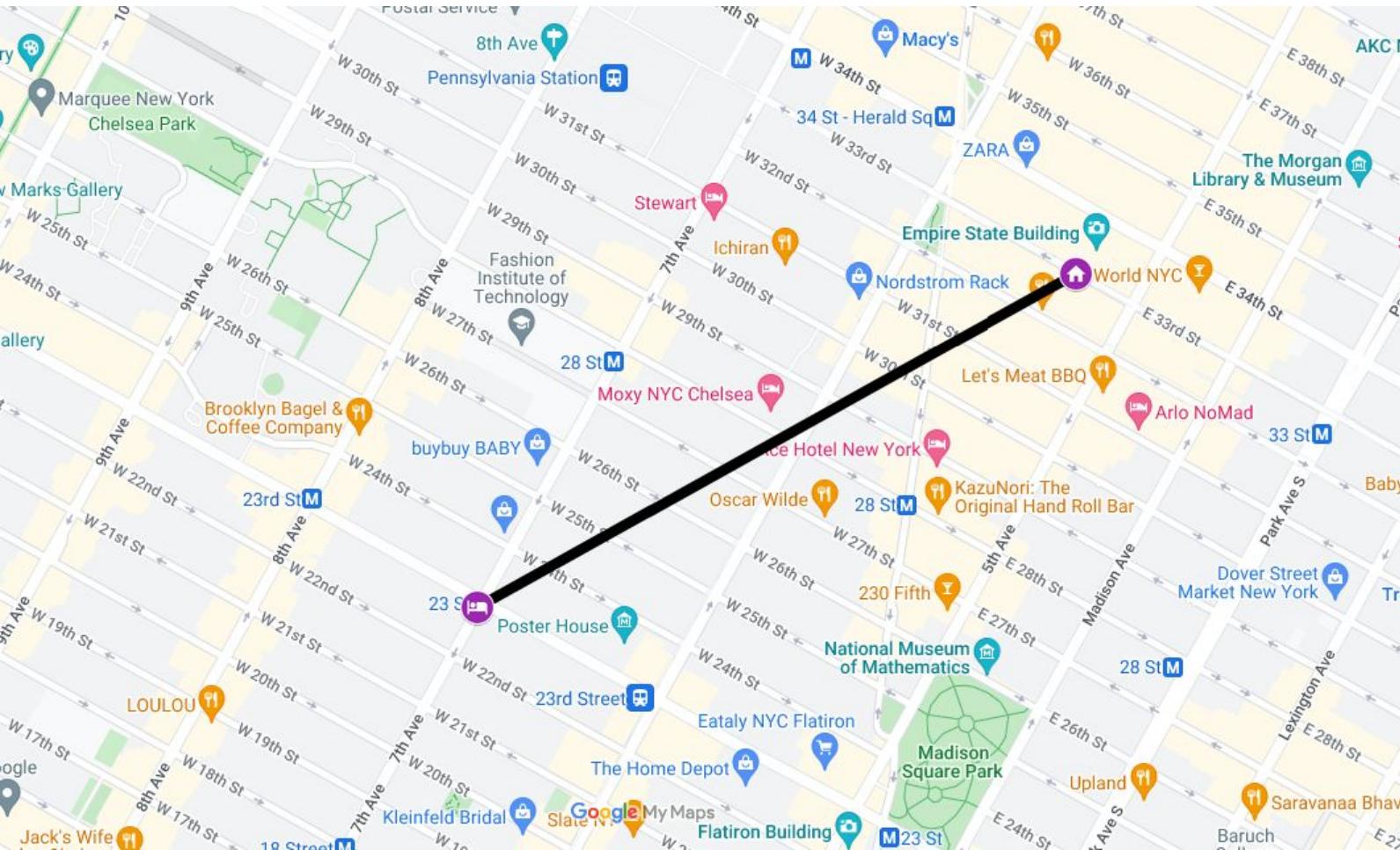
$$\Delta(T_1, T_2)$$



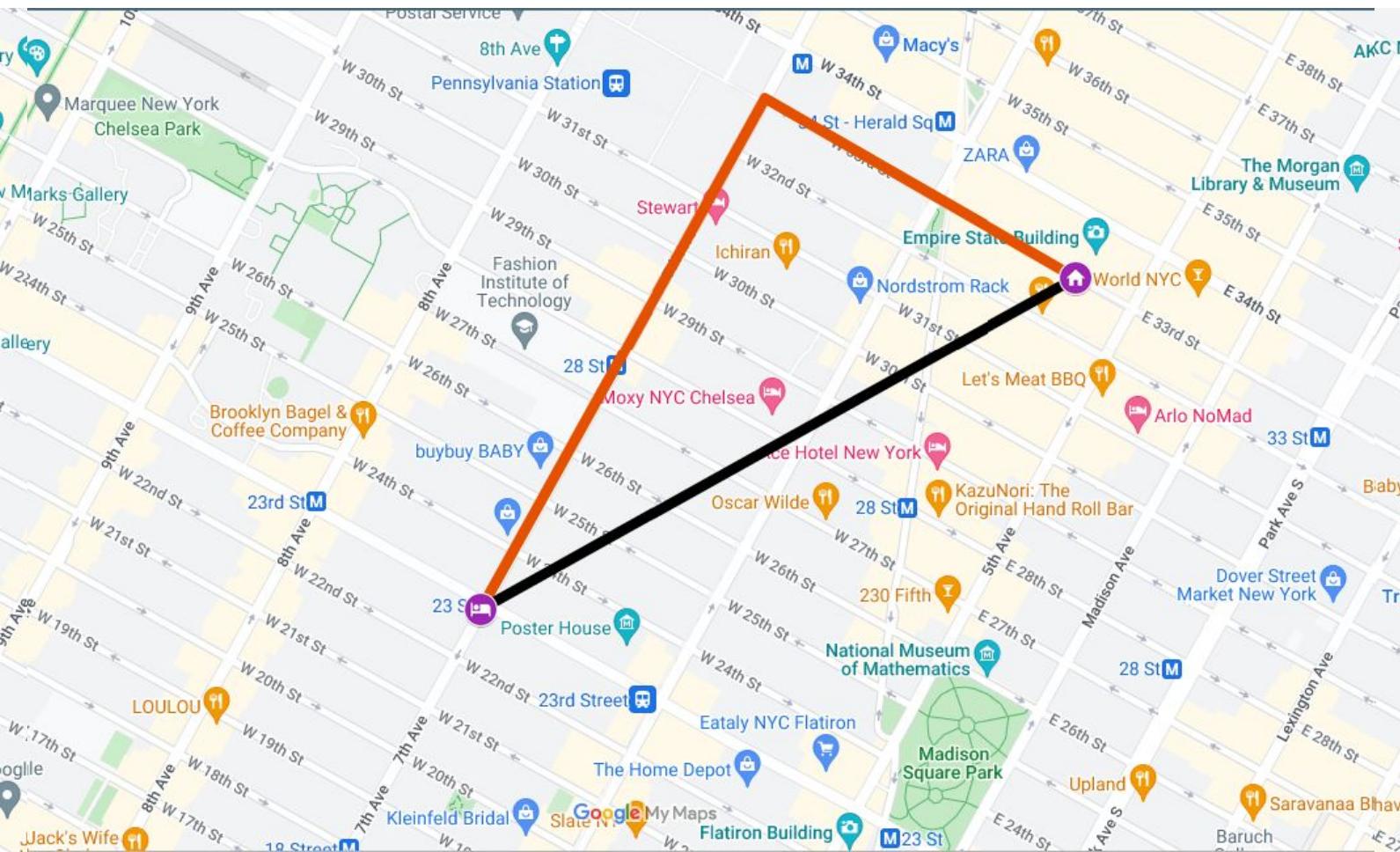
$$\Delta(T_1, T_2) = 7 + 15 + 8 = 30$$

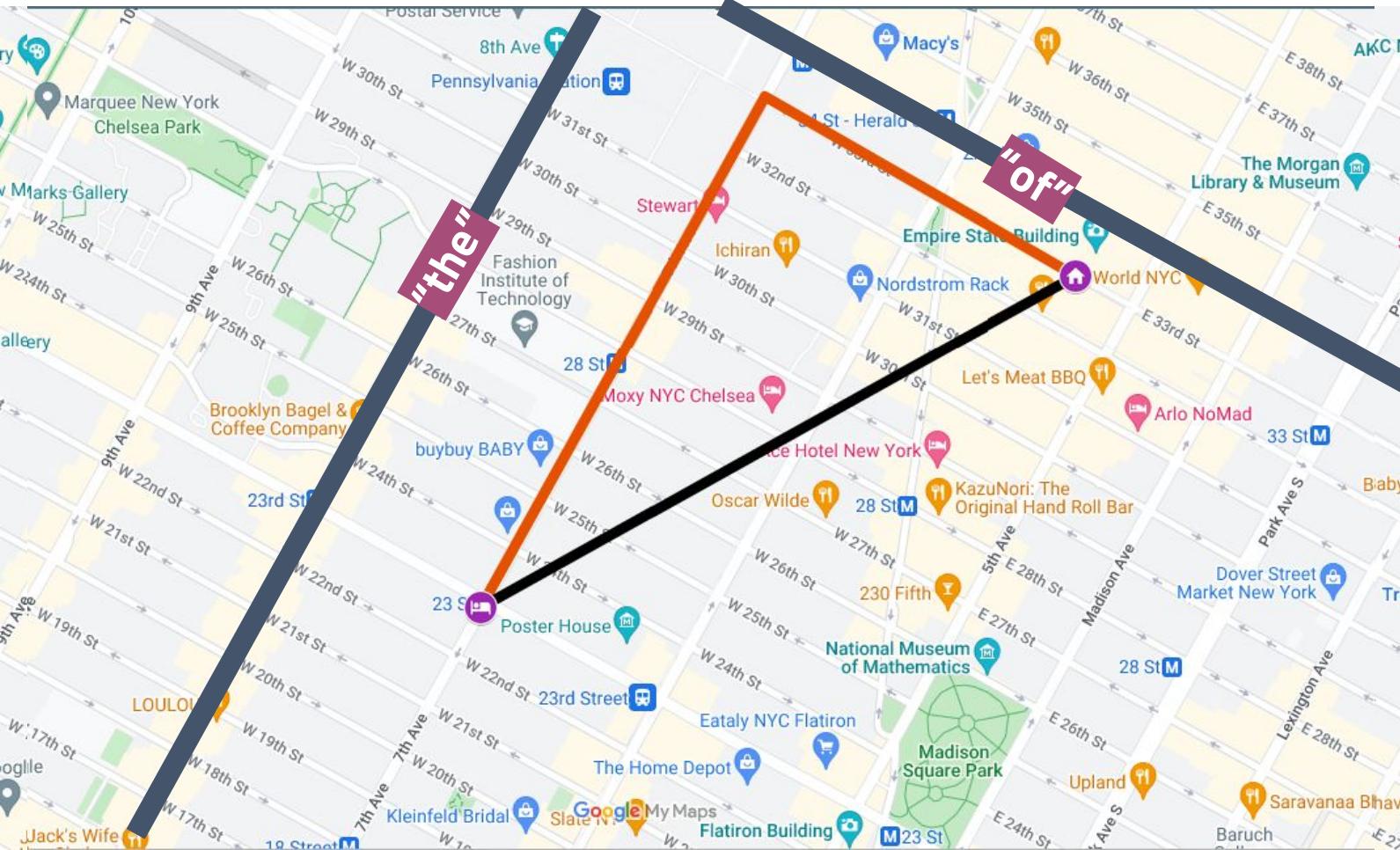
Manhattan, or city-block distance!
Reinvented by Burrows!
(with important adjustment)
Delta distance = Manhattan with scaled features

Petr Plecháč: <https://versologie.cz/talks/2017chicago/>



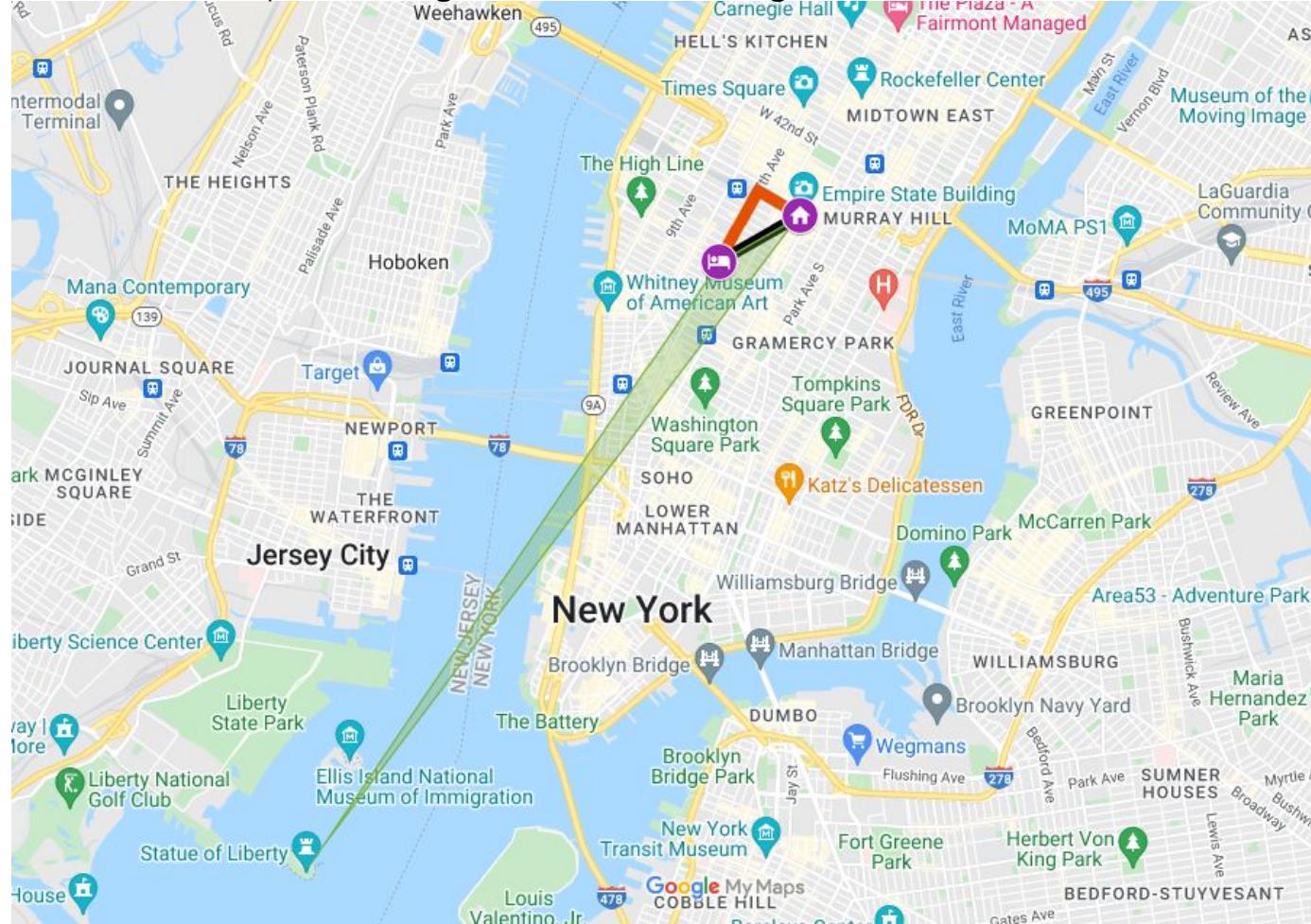
Petr Plecháč: <https://versologie.cz/talks/2017chicago/>

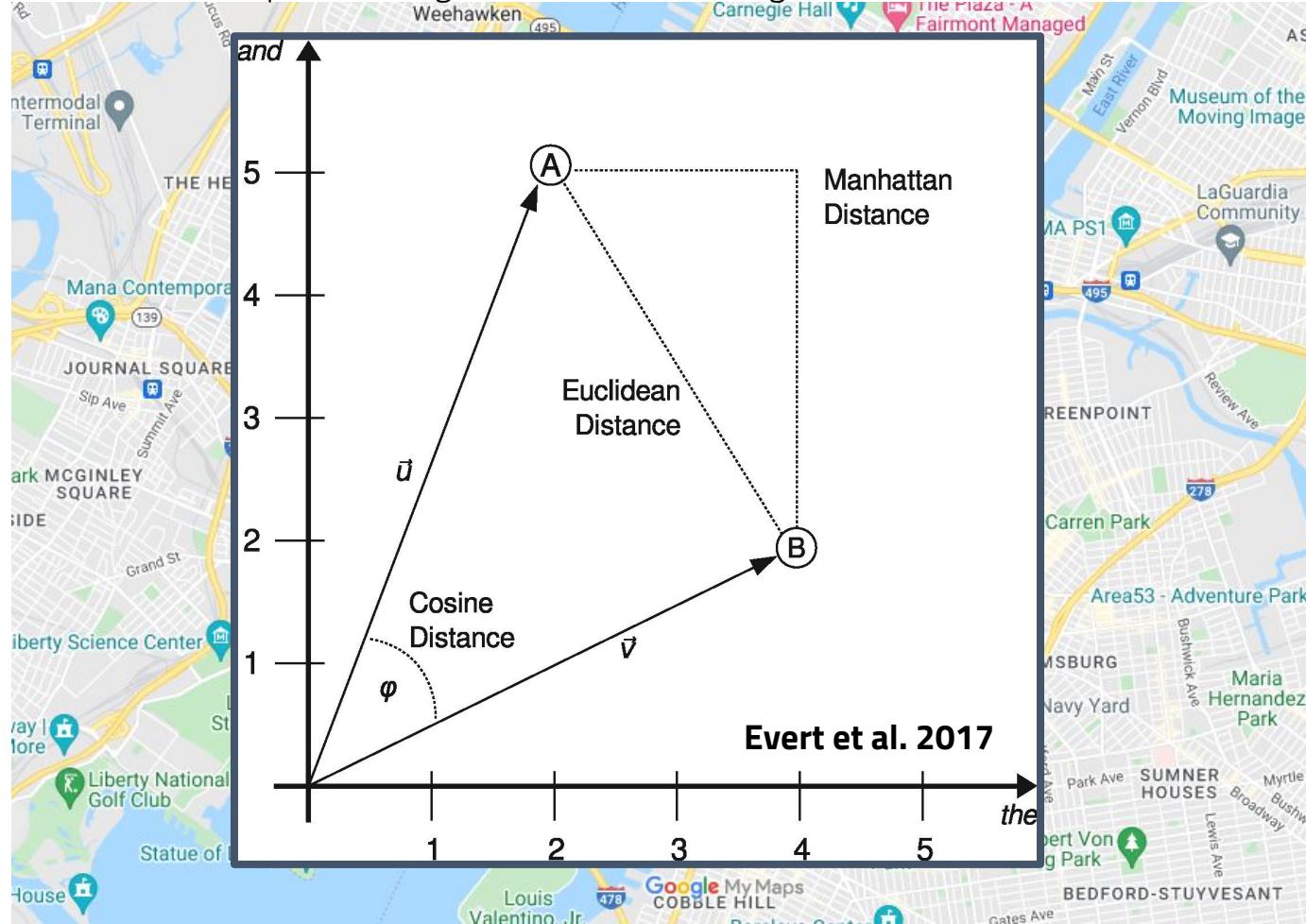


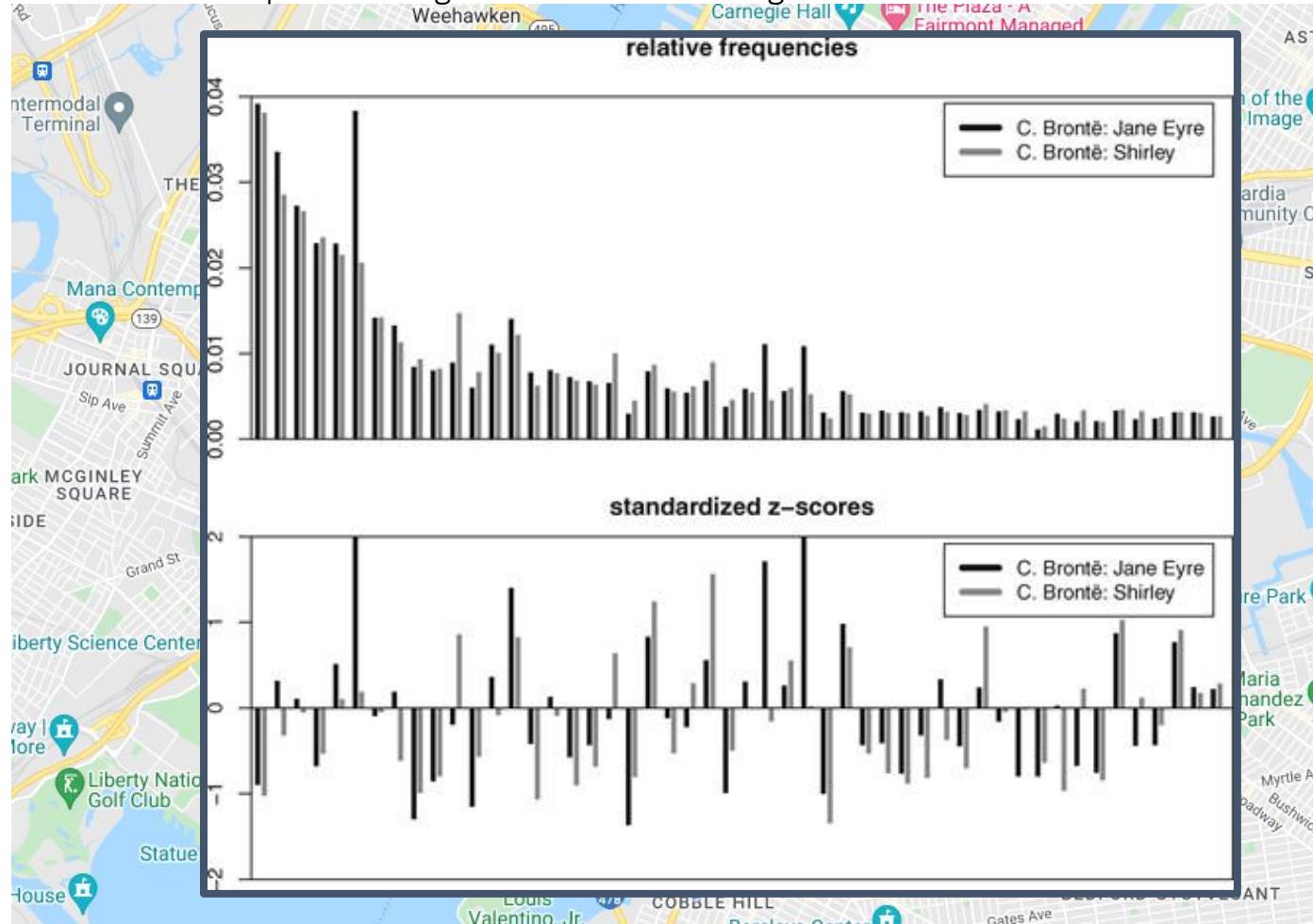




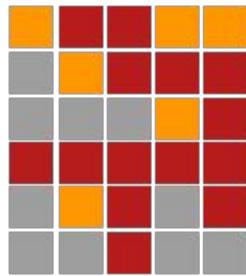
Petr Plecháč: <https://versologie.cz/talks/2017chicago/>



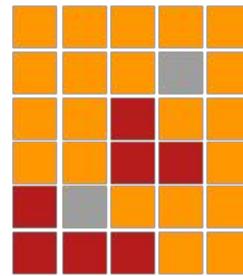




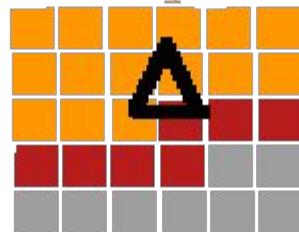
TEXT 1



TEXT 2



$$\Delta(T_1, T_2)$$



Manhattan, or city-block distance!
But also reinvented by Burrows
(with important adjustment)

$$\Delta(T_1, T_2) = 7 + 15 + 8 = 30$$

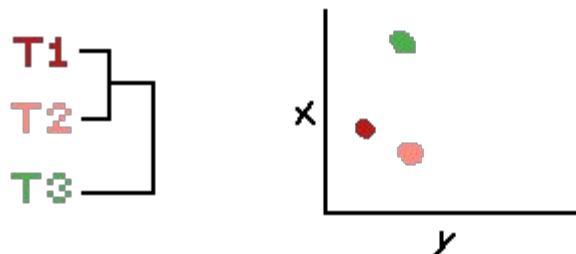
DISTANCE MATRIX

	T1	T2	T3
T1	0		
T2	0.3	0	
T3	0.7	0.9	0

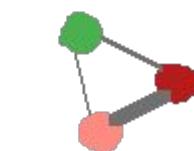
DISTANCE MATRIX

	T1	T2	T3
T1	0		
T2	0.3	0	
T3	0.7	0.9	0

MULTIDIMENSIONAL SCALING



HIERARCHICAL CLUSTERING



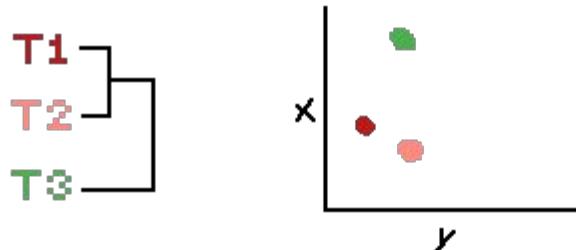
GRAPH

DISTANCE MATRIX

	T1	T2	T3
T1	0		
T2	0.3	0	
T3	0.7	0.9	0

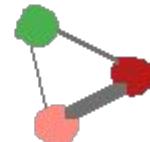
"A tree can be viewed as a simplified
description of a matrix of distances"
(Cavalli-Sforza et al.)

MULTIDIMENSIONAL SCALING



HIERARCHICAL CLUSTERING

GRAPH



Difference & similarity

Bonus: trees? DIY!

	T_1	T_2	T_3
Y_4^1	0,1	0,2	0,7
Y_4^2	0,15	0,2	0,65
X_4^1	0,3	0,5	0,2
X_4^2	0,4	0,4	0,2

Bonus: trees? DIY!

	A_1^1	A_1^2	X_1^1	X_1^2
A_1^1	0			
A_1^2	0.1	0		
X_1^1	0.8	0.9	0	
X_1^2	1	1	0.2	0

TL;DR Multivariate text analysis

1. **Feature space:** count things in multidimensional Manhattan of your design (MFWs, POS, etc...)

TL;DR Multivariate text analysis

1. **Feature space:** count things in multidimensional Manhattan of your design (MFWs, POS, etc...)
2. **Distance measure:** estimate differences between texts (each text == counted things)

TL;DR Multivariate text analysis

1. **Feature space:** count things in multidimensional Manhattan of your design (MFWs, POS, etc...)
2. **Distance measure:** estimate differences between texts (each text == counted things)
3. **Mapping relationships:** trees, projections, networks...

TL;DR Multivariate text analysis

1. **Feature space:** count things in multidimensional Manhattan of your design (MFWs, POS, etc...)
2. **Distance measure:** estimate differences between texts (each text == counted things)
3. **Mapping relationships:** trees, projections, networks...

`stylo` package does that in R!

Keys, features, and clusters

“Keyness” & “aboutness”

Keyword: ‘a word which occurs with unusual frequency in a given text [...] by comparison with a reference corpus of some kind’ (Scott 1996)

NB! The concept of “key” is essentially dependent on the context

“Keyness” & “aboutness”

Keyword: ‘a word which occurs with unusual frequency in a given text [...] by comparison with a reference corpus of some kind’ (Scott 1996)

NB! The concept of “key” is essentially dependent on the context

‘any **difference** in the linguistic character of two corpora **will leave its trace** in differences between their **word frequency lists**’ (Kilgarriff 1997)

“Keyness” & “aboutness”

Keyword: ‘a word which occurs with unusual frequency in a given text [...] by comparison with a reference corpus of some kind’ (Scott 1996)

NB! The concept of “key” is essentially dependent on the context

‘any **difference** in the linguistic character of two corpora **will leave its trace** in differences between their **word frequency lists**’ (Kilgarriff 1997)

Two main approaches: **EFFECT-SIZE DIFFERENCE** & **SIGNIFICANCE TESTING**

“Keyness” & “aboutness”

Keyword: ‘a word which occurs with unusual frequency in a given text [...] by comparison with a reference corpus of some kind’ (Scott 1996)

NB! The concept of “key” is essentially dependent on the context

‘any **difference** in the linguistic character of two corpora **will leave its trace** in differences between their **word frequency lists**’ (Kilgarriff 1997)

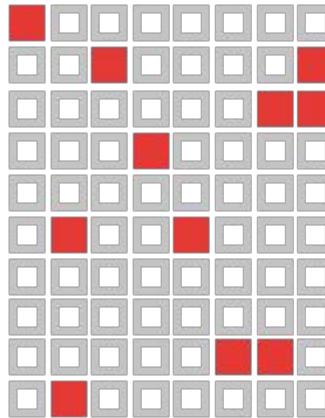
Two main approaches: **EFFECT-SIZE DIFFERENCE** & **SIGNIFICANCE TESTING**

NOT PAIRWISE: WORD DISPERSION MEASURES

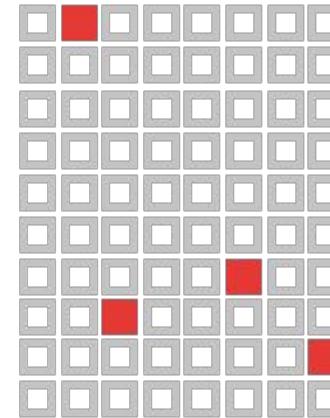
EFFECT-SIZE DIFFERENCE

HOW MUCH
DIFFERENCE

C1
("STUDY")



C2
("REFERENCE")

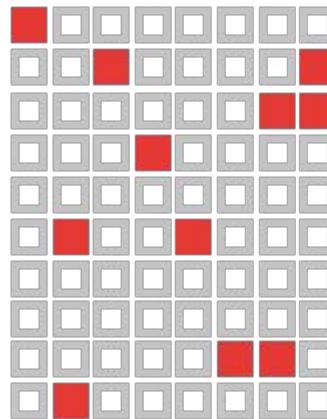


■ WIZARD

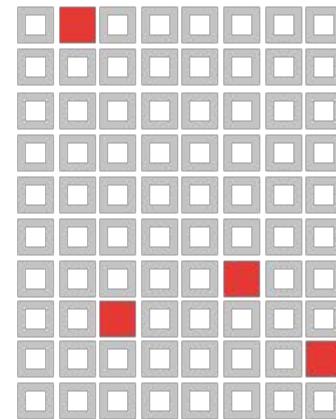
□ OTHER WORDS

HOW USE OF **WIZARD** DIFFERS
ACROSS C1 & C2?

C1
("STUDY")



C2
("REFERENCE")



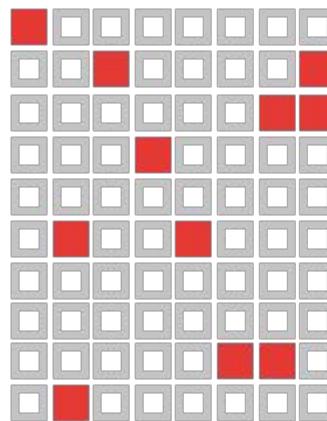
C1: ■■■■■■■■■■

$$P(\text{WIZARD}, C1) = 11/80 = 0.1375$$

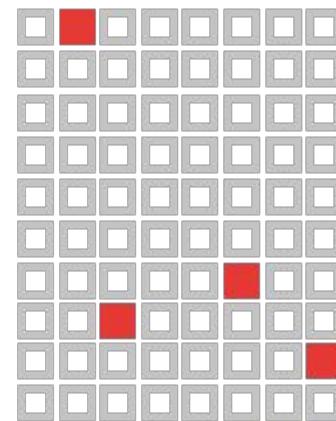
C2: ■■■■

$$P(\text{WIZARD}, C2) = 4/80 = 0.05$$

C1
("STUDY")



C2
("REFERENCE")



C1: ■■■■■■■■■■

RAW

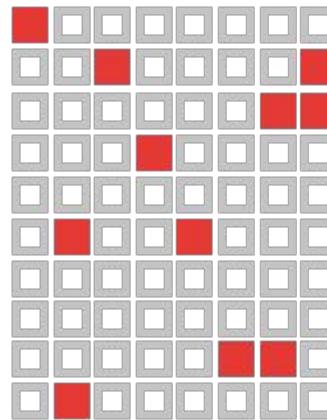
$$P(\text{WIZARD}, C1) = \underline{\underline{11/80 = 0.1375}}$$

C2: ■■■■

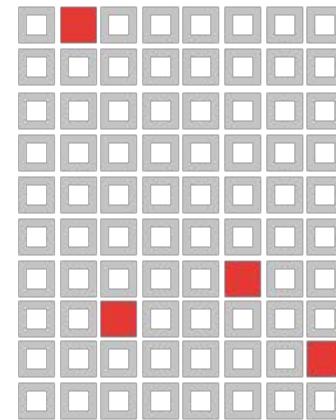
NORMALIZED

$$P(\text{WIZARD}, C2) = \underline{\underline{4/80 = 0.05}}$$

C1
("STUDY")



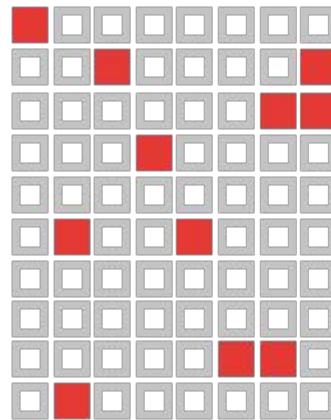
C2
("REFERENCE")



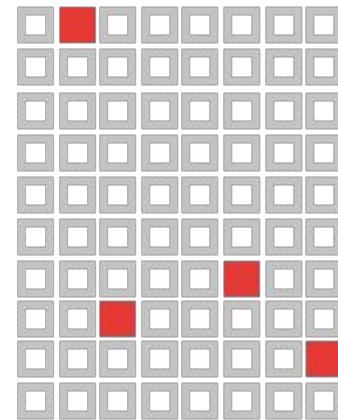
SIMPLE RATIO:

$$R = \frac{P(WIZARD, C1)}{P(WIZARD, C2)} = 2.75$$

C1
("STUDY")



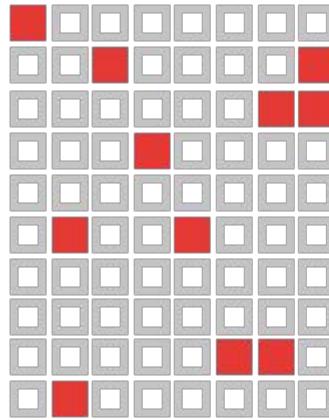
C2
("REFERENCE")



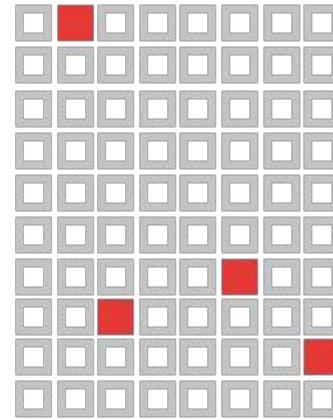
SIMPLE RATIO:

$$R = \frac{C1}{C2} = \frac{10}{4} = 2.5$$

C1
("STUDY")



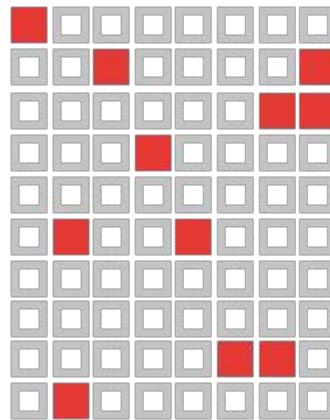
C2
("REFERENCE")



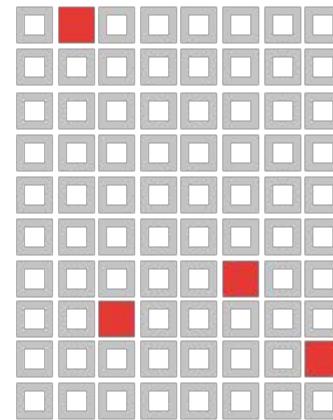
DIFFERENCE COEFFICIENT:

$$DC = \frac{P(WIZARD, C1) - P(WIZARD, C2)}{P(WIZARD, C1) + P(WIZARD, C2)} = 0.467$$

C1
("STUDY")



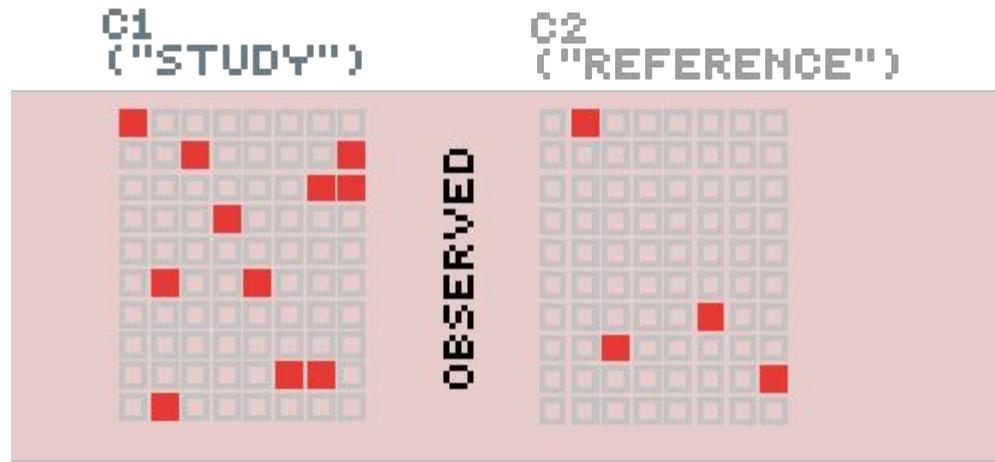
C2
("REFERENCE")



DIFFERENCE COEFFICIENT:

$$DC = \frac{\text{11-4}}{\text{11+4}} = 0.467$$

The equation shows the calculation of the Difference Coefficient (DC). The numerator is represented by a row of 4 red squares above a horizontal line. The denominator is represented by a row of 14 red squares below the same horizontal line. The result is given as 0.467.



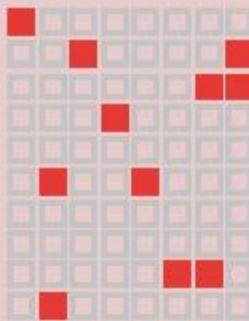
WHAT IF BOTH FREQUENCIES ARE DRAWN
FROM THE SAME DISTRIBUTION?
(= NO DIFFERENCE)

Significance testing: how (un)probable is the absence of any difference? (H_0)

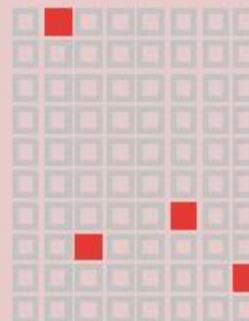
Methods: χ^2 (chi-squared test), log-likelihood...

C1
("STUDY")

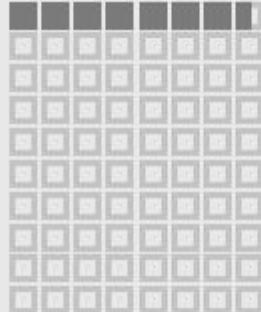
C2
("REFERENCE")



OBSERVED



????

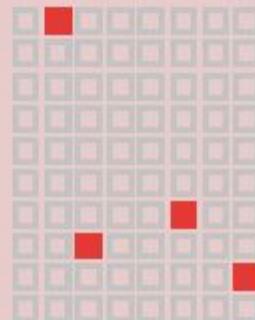
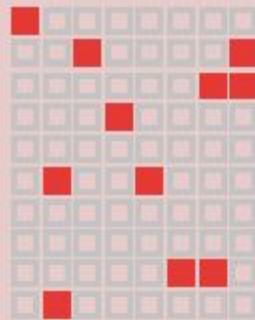


EXPECTED

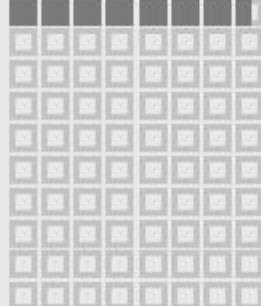
C1
("STUDY")

C2
("REFERENCE")

OBSERVED



EXPECTED

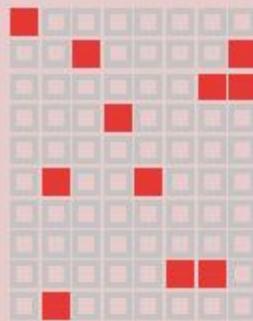


CHI-SQUARED
WC1:

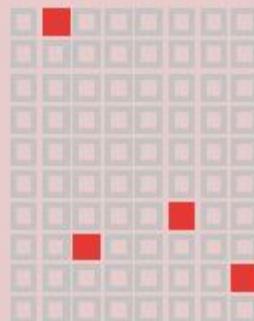
$$\frac{\text{OBSERVED} - \text{EXPECTED}}{\text{EXPECTED}}^2$$

C1
("STUDY")

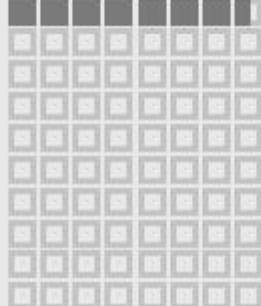
C2
("REFERENCE")



OBSERVED



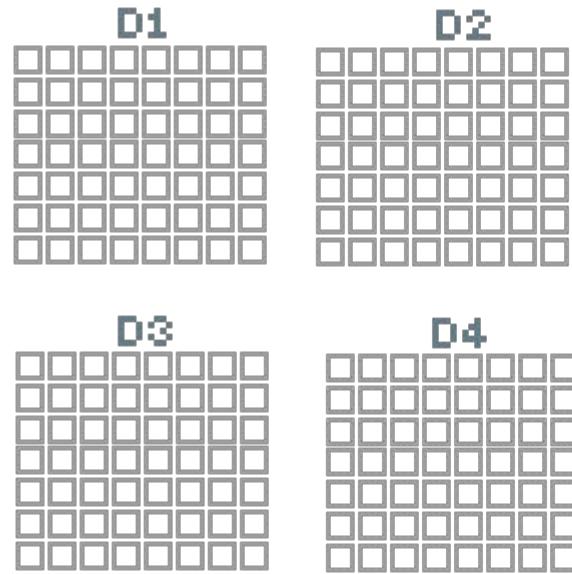
EXPECTED



CHI-SQUARED
WC1:

$$\frac{(\text{OBSERVED} - \text{EXPECTED})^2}{\text{EXPECTED}}$$

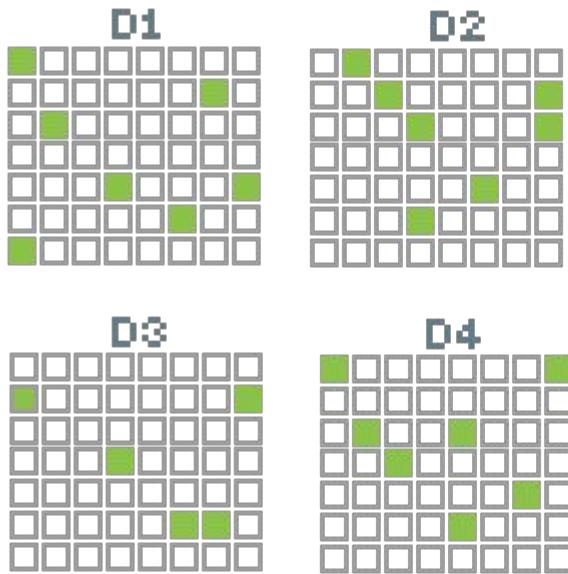
FINAL CHI-SQUARED:
 $\text{SUM}(\text{CHI-SQUARED FOR EACH WORD})$



TERM FREQ
X
INVERSE
DOCUMENT
FREQ

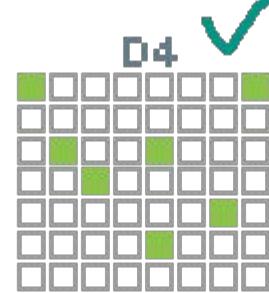
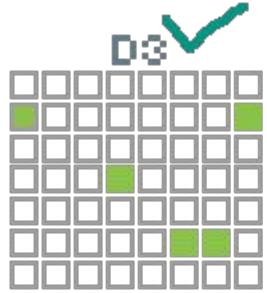
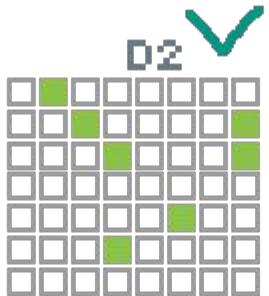
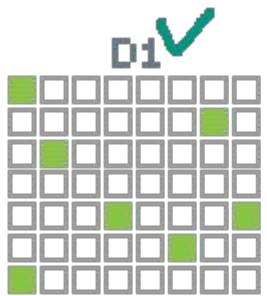
TF X IDF

Dispersion: words across documents



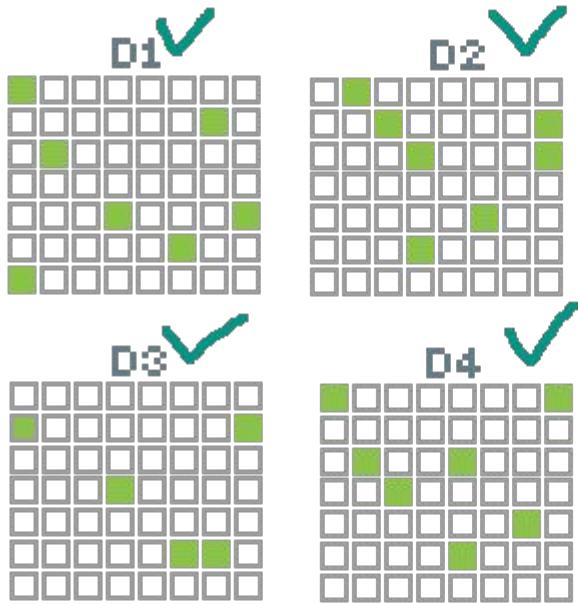
TERM FREQ
X
INVERSE
DOCUMENT
FREQ

TF X IDF



TERM FREQ
✗
INVERSE
DOCUMENT
FREQ

TF ✗ IDF

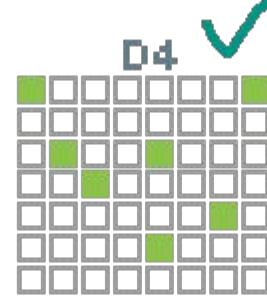
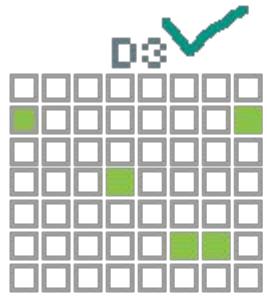
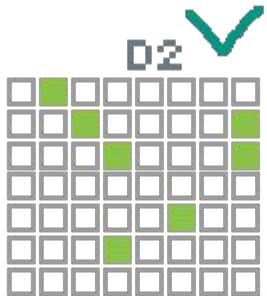
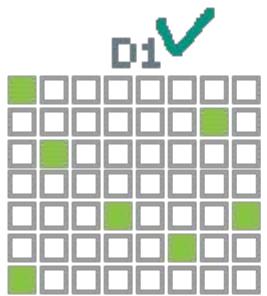


TERM FREQ
X

INVERSE
DOCUMENT
FREQ

TF X IDF

$$TFIDF(WD1) = ? \times \log \frac{\text{NUM OF DOCUMENTS}}{\text{NUM OF 'WORD' IN DOCUMENTS}}$$

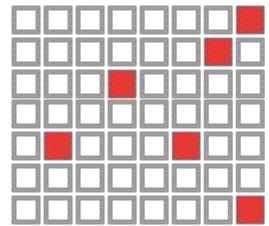


TERM FREQ
✗
INVERSE
DOCUMENT
FREQ

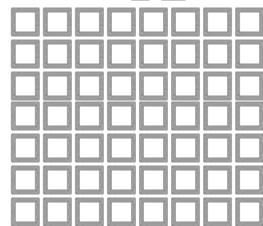
TF ✗ IDF

$$TFIDF(WD1) = 7 \times \log(\frac{4}{4}) = 0$$

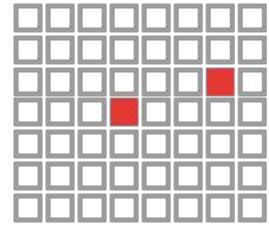
D1



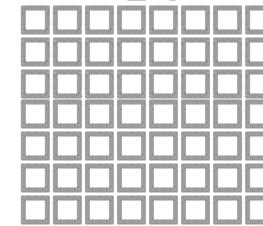
D2



D3

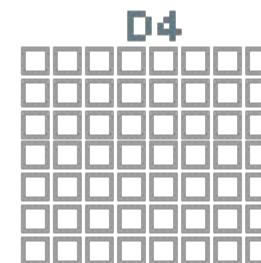
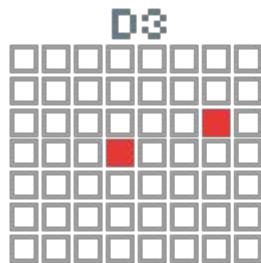
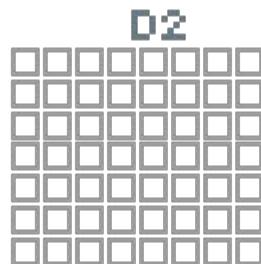
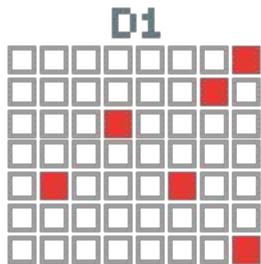


D4



TERM FREQ
X
INVERSE
DOCUMENT
FREQ

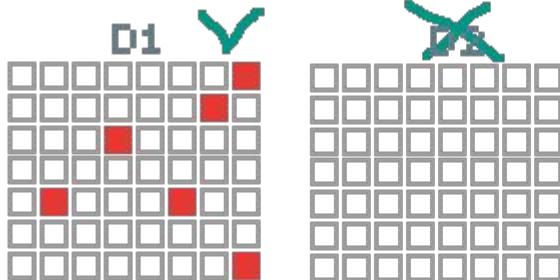
TF X IDF



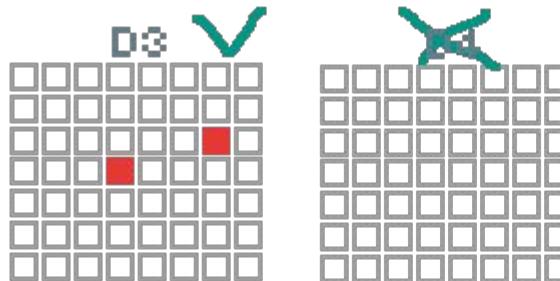
TERM FREQ
X
INVERSE
DOCUMENT
FREQ

TF X IDF

$$TFIDF(WD1) = 6 \times \log \frac{4}{2} = 4.16$$



TERM FREQ
X
INVERSE
DOCUMENT
FREQ



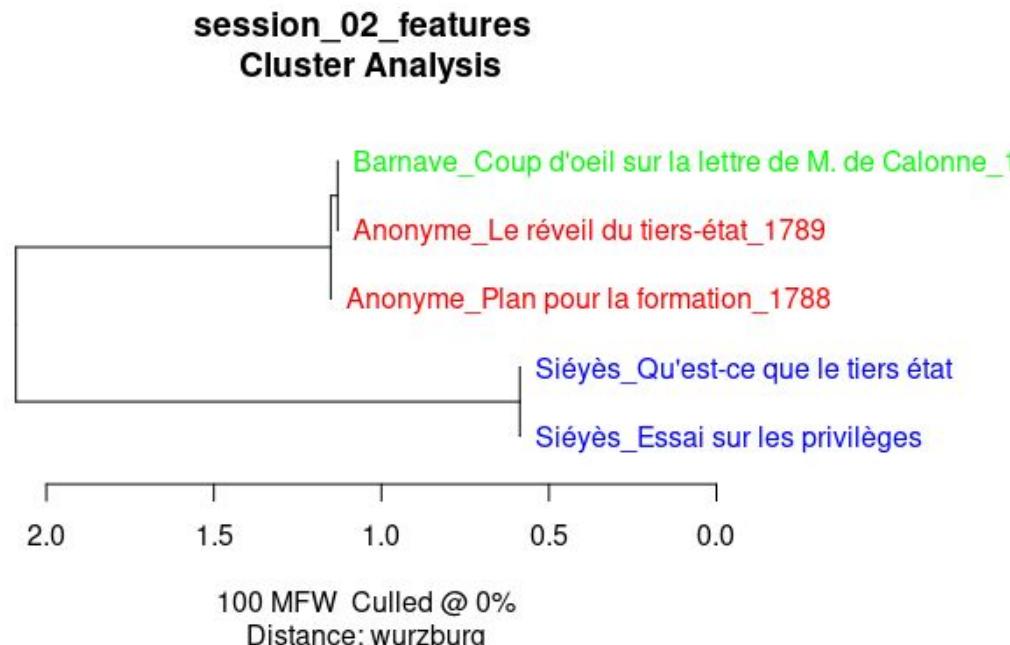
TF X IDF

$$TFIDF(WD1) = 6 \times \log \frac{4}{2} = 4.16$$

$$TFIDF(WD2) = 2 \times \log \frac{4}{2} = 1.38$$

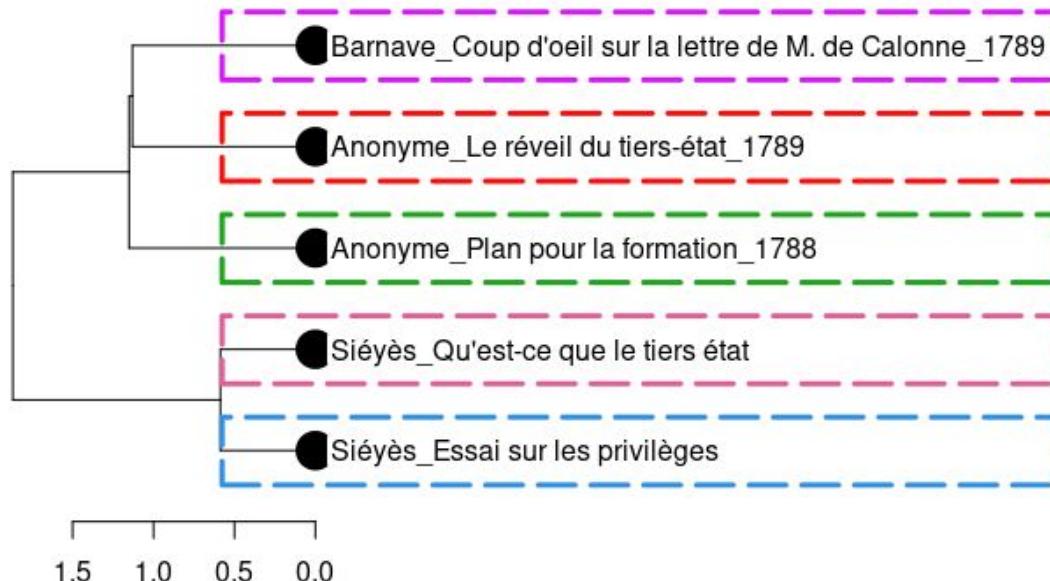
Words behind trees

Cluster analysis: grouping suggestion



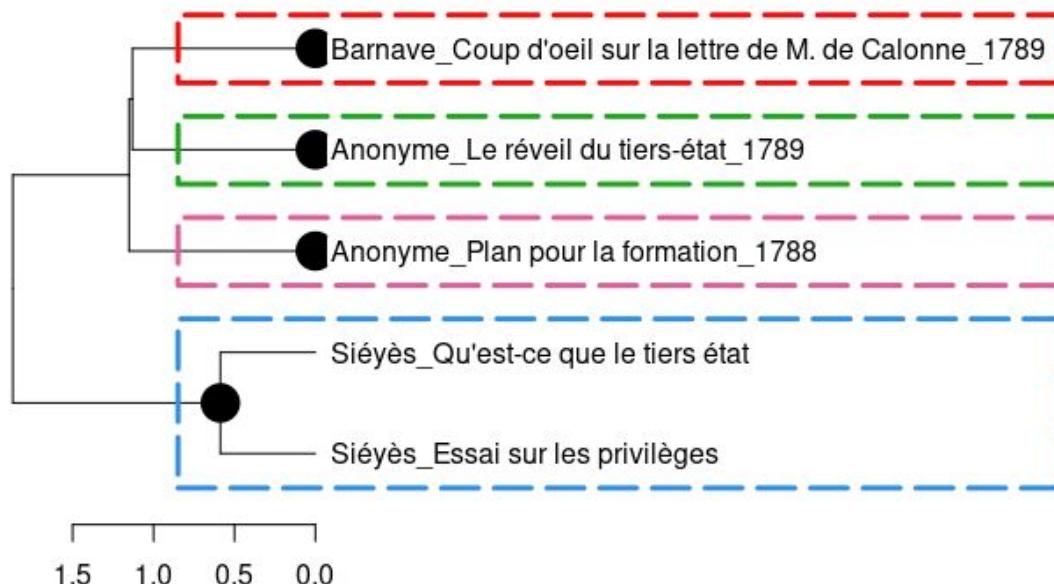
Cutting trees by groups!

Hierarchical clustering, cut at k= 5



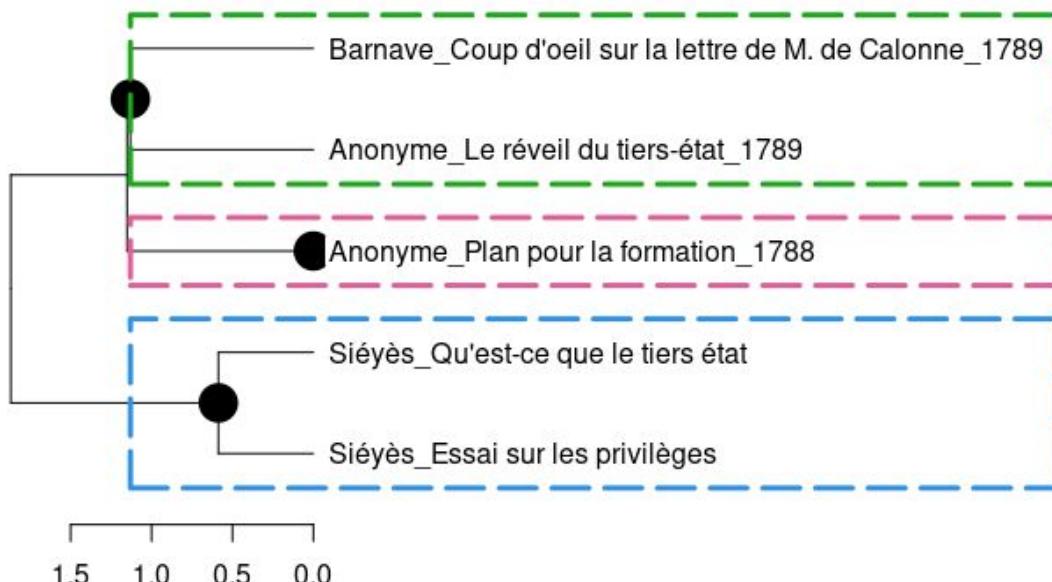
Cutting trees by groups!

Hierarchical clustering, cut at k= 4



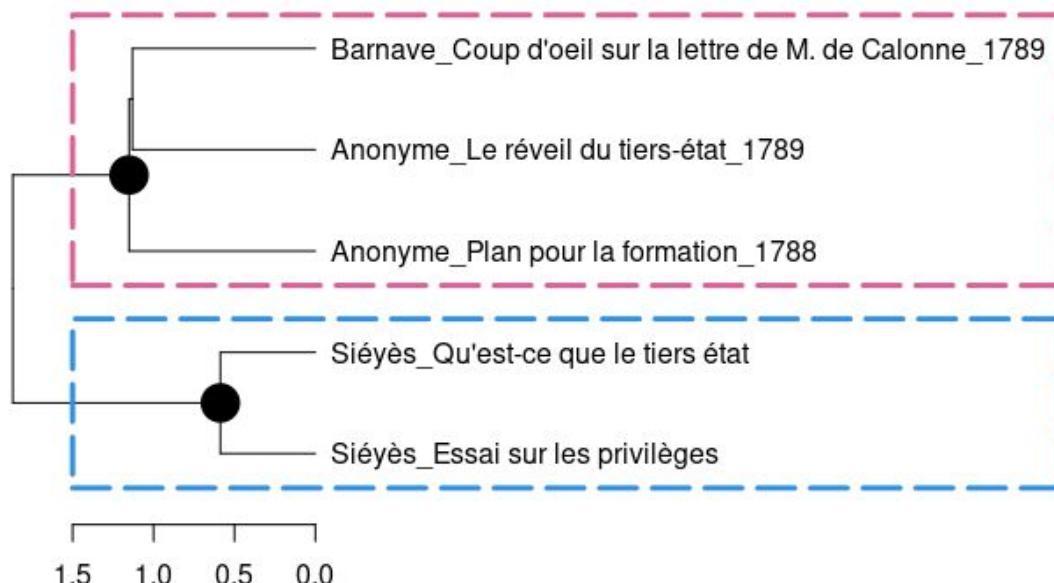
Cutting trees by groups!

Hierarchical clustering, cut at k= 3



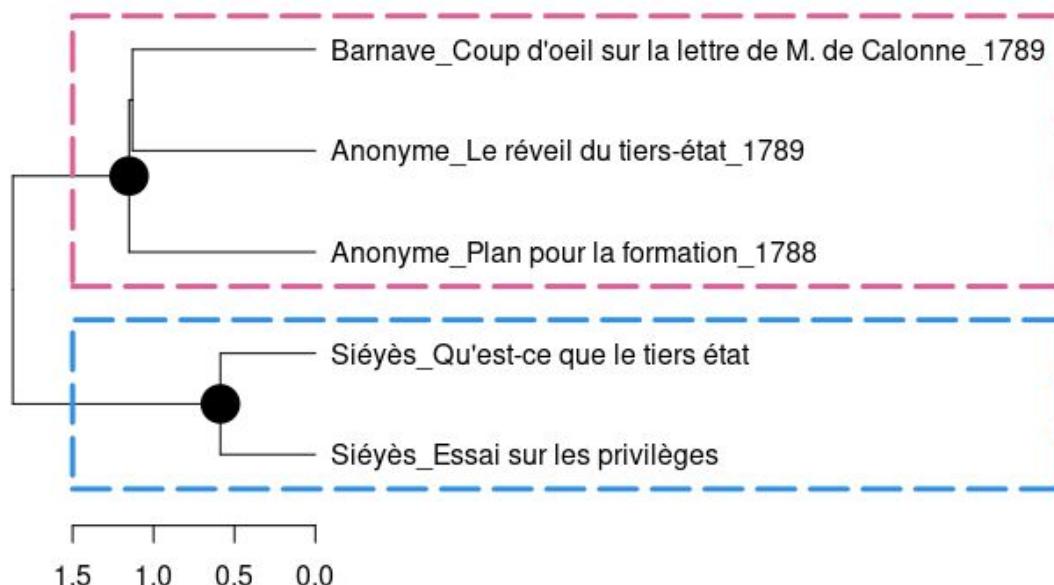
Cutting trees by groups!

Hierarchical clustering, cut at k= 2



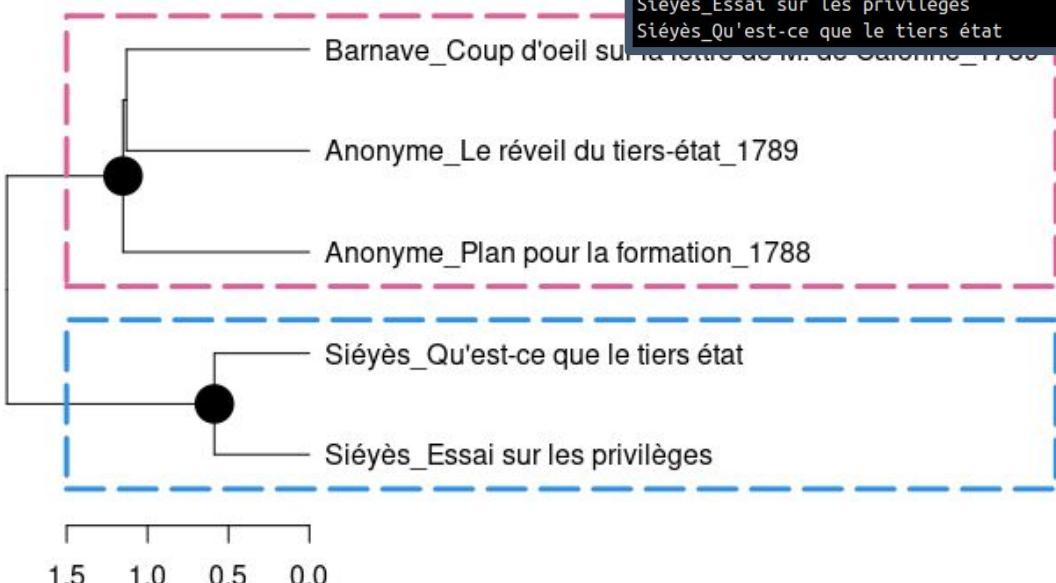
Think of new ‘pink’ and ‘blue’ as suggested classes

Hierarchical clustering, cut at k= 2



Classes are driven by word frequencies!

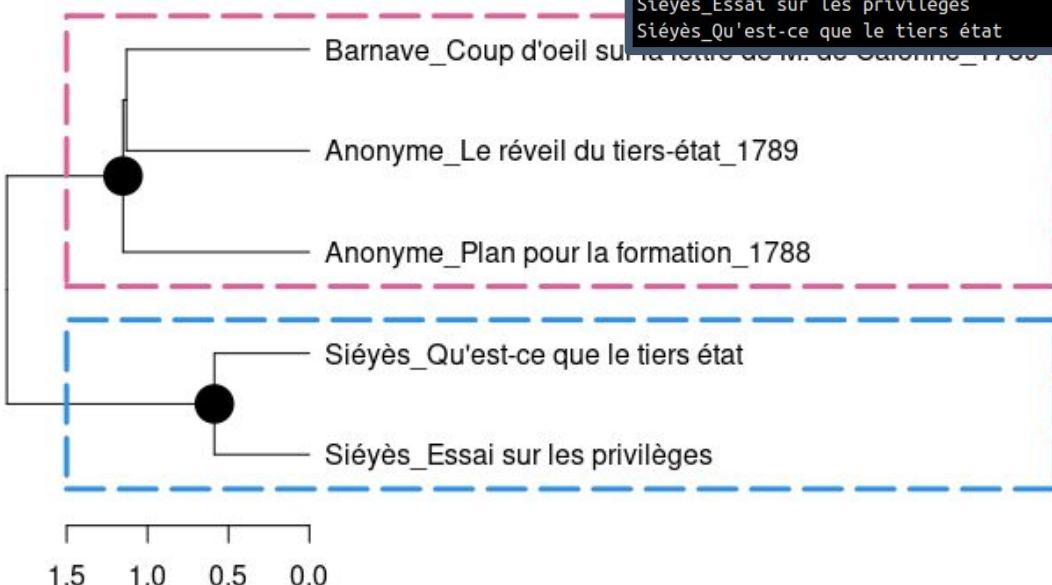
Hierarchical clustering, cut at k= 2



	de	la	les	l	à	le
Anonyme_Le réveil du tiers-état_1789	4.9919255	3.1490453	2.9068111	2.0993635	1.8333808	1.9663722
Anonyme_Plan pour la formation_1788	4.3847242	3.1117397	4.6676096	1.6030174	1.3672796	1.7444602
Barnave_Coup d'oeil sur la lettre de M. de Calonne_1789	4.2275472	3.0196766	2.8443405	2.2209234	1.1494253	2.1235145
Siéyès_Essai sur les priviléges	4.3275072	2.729227	2.4114403	2.0095336	2.3460136	1.8412936
Siéyès_Qu'est-ce que le tiers état	3.9608393	2.9340656	2.3968003	1.874459	2.0416082	1.8983375

Feature ~ cluster association

Hierarchical clustering, cut at k= 2



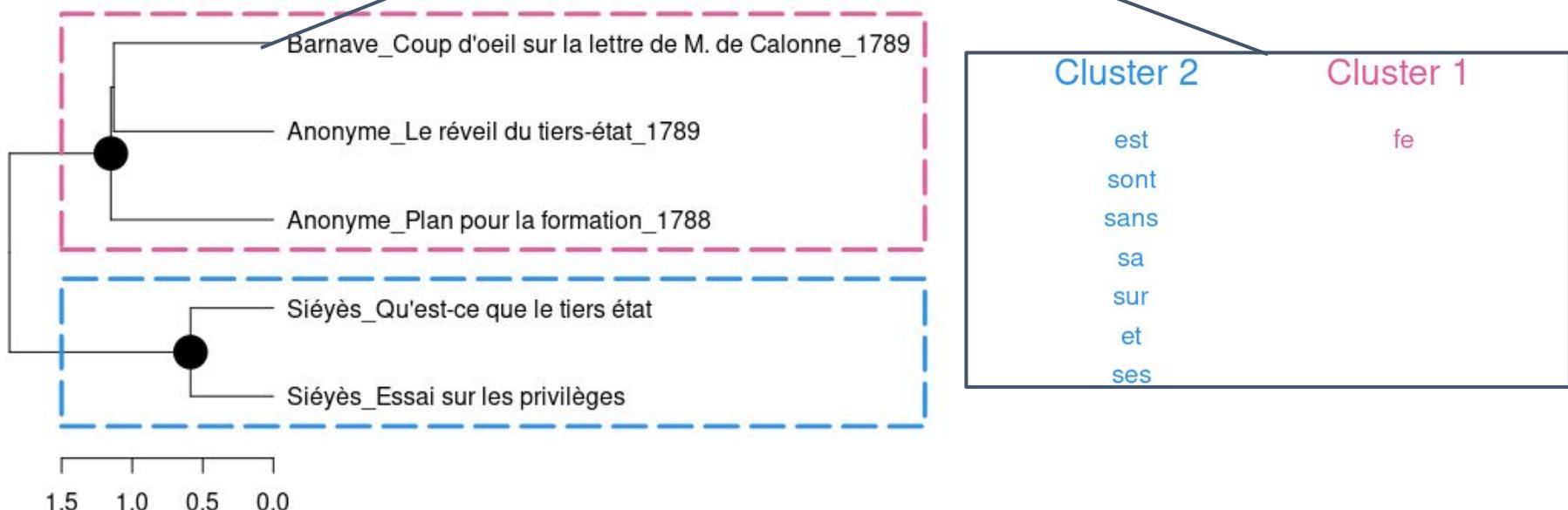
	de	la	les	l	à	le
Anonyme_Le réveil du tiers-état_1789	4.9919255	3.1490453	2.9068111	2.0993635	1.8333808	1.9663722
Anonyme_Plan pour la formation_1788	4.3847242	3.1117397	4.6676096	1.6030174	1.3672796	1.7444602
Barnave_Coup d'oeil sur la lettre de M. de Calonne_1789	4.2275472	3.0196766	2.8443405	2.2209234	1.1494253	2.1235145
Siéyès_Essai sur les priviléges	4.3275072	2.729227	2.4114403	2.0095336	2.3460136	1.8412936
Siéyès_Qu'est-ce que le tiers état	3.9608393	2.9340656	2.3968003	1.874459	2.0416082	1.8983375

You can use 'emergent' class information from the tree to define corpora and check **which features differ** across clusters.

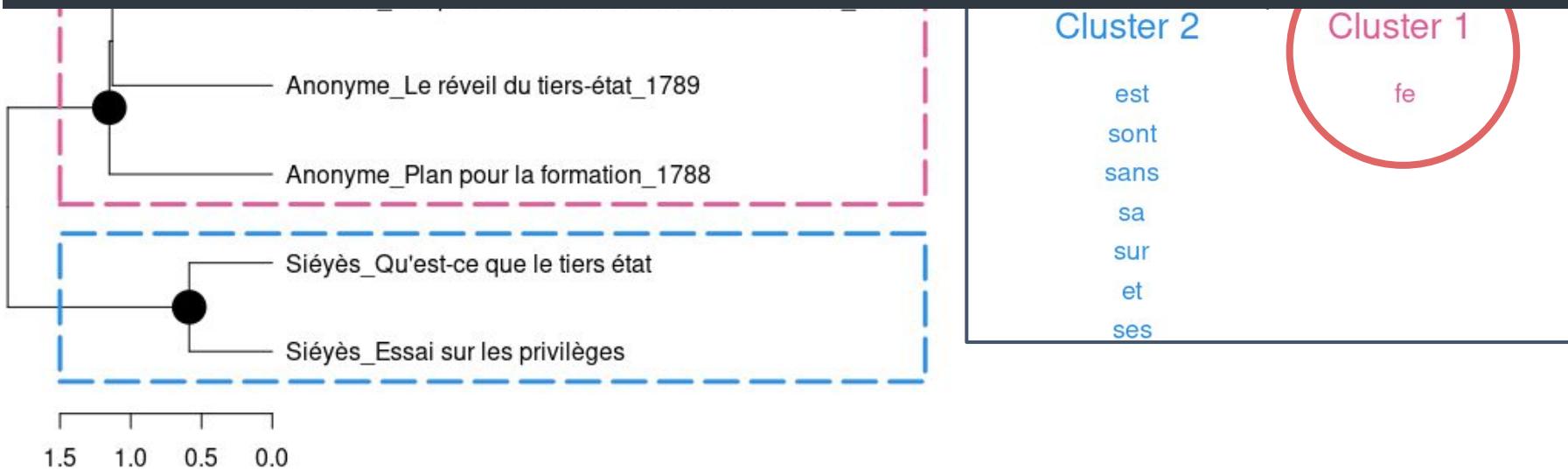
Think about it as **keyness** problem.

	de	la	les	l	à	le
Anonyme_Le réveil du tiers-état_1789	4.9919255	3.1490453	2.9068111	2.0993635	1.8333808	1.9663722
Anonyme_Plan pour la formation_1788	4.3847242	3.1117397	4.6676096	1.6030174	1.3672796	1.7444602
Barnave_Coup d'oeil sur la lettre de M. de Calonne_1789	4.2275472	3.0196766	2.8443405	2.2209234	1.1494253	2.1235145
Siéyès_Essai sur les priviléges	4.3275072	2.729227	2.4114403	2.0095336	2.3460136	1.8412936
Siéyès_Qu'est-ce que le tiers état	3.9608393	2.9340656	2.3968003	1.874459	2.0416082	1.8983375

Hierarchical clustering, cut at k= 2



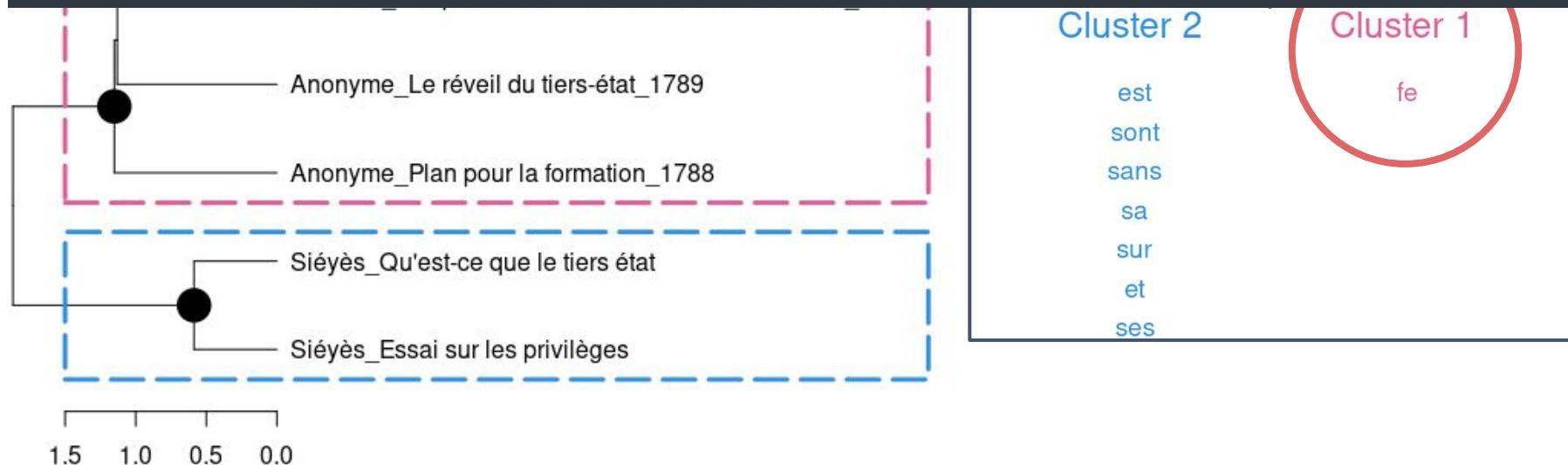
pute, lequel fe 1 Anemblée de Témion ou arrondiïïement. Ces Députés ne pourront être élus que parmi les propriétaires domiciliés ou parmi les forains qui auront des propriétés dans le lieu payant cinquante livres de charges réelles & pour être élu, il ne fera pas néceflaire d'être présent à l'Afl'emblée.. Les Etars indiquer 6ht les chefs-lieux des arrondiflemens ailleurs que dans les villes qui Ont des Députés particuliers \$ &. pouf la première convocation, les Députes de i'Efëetion de Grenoble fe réuniront a Vitille ceux de l'Election de Vienne, à Bouroin,; ceux de l'Elefliôn de Romans à Beaurepaire; ceux de l'Ele&ion de Valence, Chabeuil ceux de l'Eleftiqn de Gap, ^Charges ceux de l'Eleélion de, Monte- entr*eux les^ Députés qui devront repréfeii* -t|r du içMiûtib; aux EtatliLé IMrocès-verbai fera envoyé au Secrétaire* le nom des



pute, lequel fe 1 Anemblée de Témion ou arrondiïïement. Ces Députés ne pourront être élus que parmi les propriétaires domiciliés ou parmi les forains qui auront des propriétés dans le lieu payant cinquante livres de charges réelles & pour être élu, il ne fera pas néceflaire d'être présent à l'Afl'emblée.. Les Etars indiquer 6ht les chefs-lieux des arrondiflemens ailleurs que dans les villes qui Ont des Députés particuliers \$ &. pouf la première convocation, les Députes de i'Efëetion de Grenoble fe réuniront a Vitille ceux de l'Election de Vienne, à Bouroin,; ceux de l'Elefliôn de Romans à Beaurepaire; ceux de l'Ele&io

OCR: becomes **more** of a problem when text sources are different

H
Monte- e
IMrocès - Versai sera envoié au Secrétaire - le nom des



``seetrees` package (very unfinished)`

Installation

Install from GitHub (make sure you have `devtools` package):

```
devtools::install_github("perechen/seetrees")
```

Example

```
library(stylo)
library(seetrees)

data(lee) ## load one of the stylo datasets

stylo_res <- stylo(frequencies=lee,gui=F)
view_tree(stylo_res, k=2,right_margin=12) ## redraws a dendrogram based on distance matrix, cuts it
```

Check `?view_tree()` for more details.

‘seetrees` package (very unfinished)

Installation

Install from GitHub (make sure you have

```
devtools::install_github("perechen/
```

Example

```
library(stylo)
library(seetrees)

data(lee) ## load one of the stylo objects

stylo_res <- stylo(frequencies=lee)
view_tree(stylo_res, k=2,right_margin=10)
```

```
1 CLUSTER 1
2 =====
3 TEXTS
4 Anonyme_Plan pour la formation_1788
5 Anonyme_Le réveil du tiers-état_1789
6 Barnave_Coup d'oeil sur la lettre de M. de Calonne_1789
7 =====
8 FEATURES associated (p<0.05)
9
10 fe
11
12
13
14
15 CLUSTER 2
16 =====
17 TEXTS
18 Siéyès_Essai sur les priviléges
19 Siéyès_Qu'est-ce que le tiers état
20 =====
21 FEATURES associated (p<0.05)
22
23 est sont sans sa sur et ses
24
25
```

Check `?view_tree()` for more details.

Sampling and uncertainty

Sidenote

Sampling without replacement:



Sidenote

Sampling without replacement:



Sidenote

Sampling without replacement:



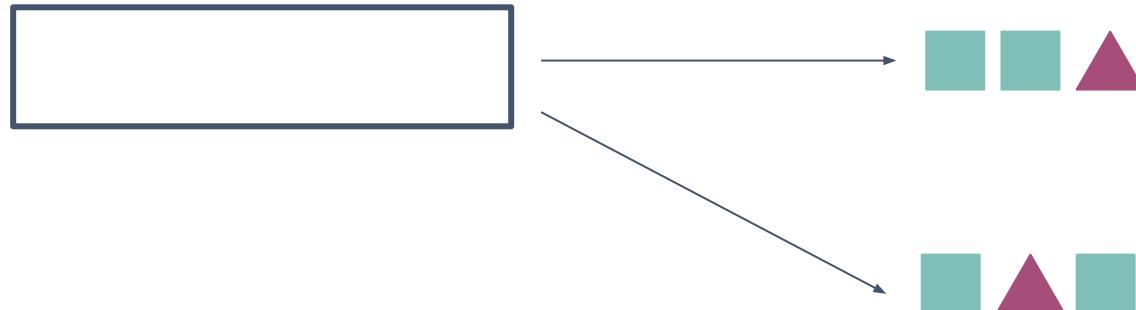
Sidenote

Sampling without replacement:



Sidenote

Sampling without replacement:



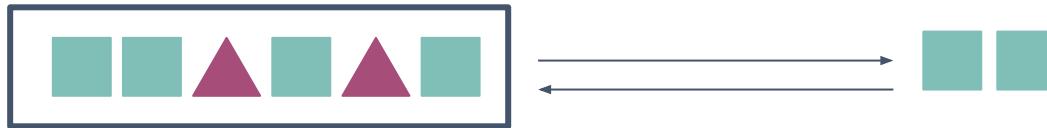
Sidenote

Sampling ***with*** replacement:



Sidenote

Sampling ***with*** replacement:



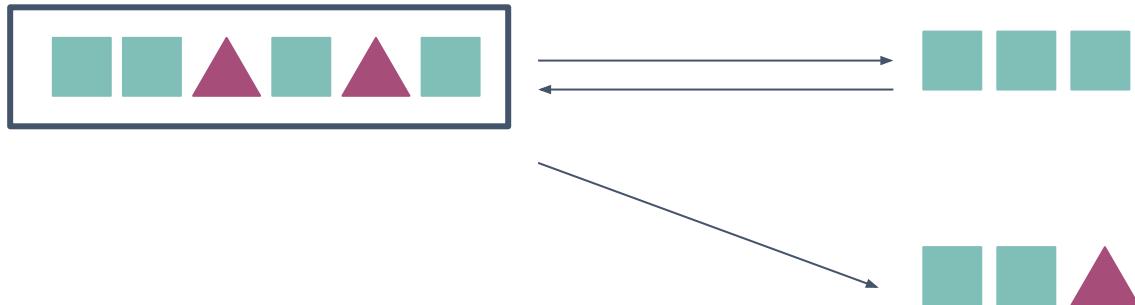
Sidenote

Sampling ***with*** replacement:



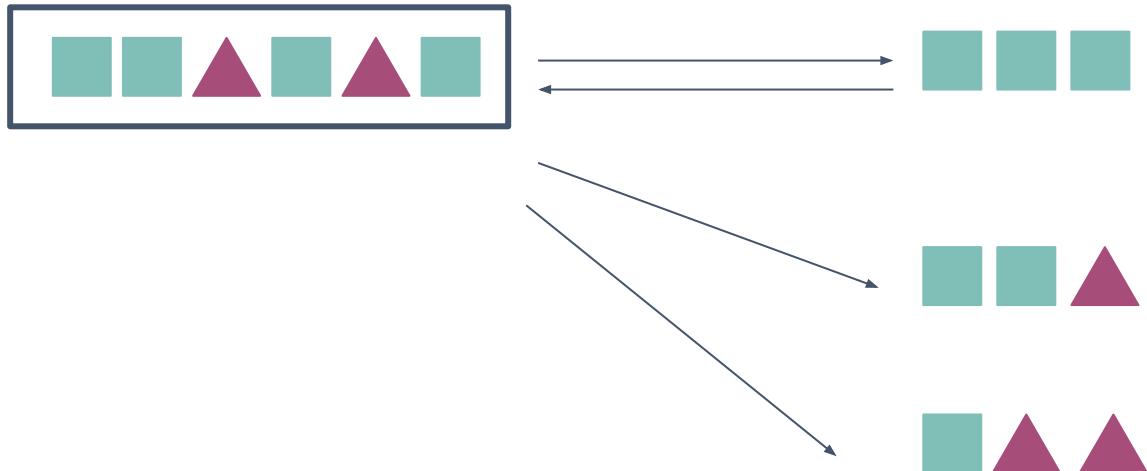
Sidenote

Sampling ***with*** replacement:



Sidenote

Sampling ***with*** replacement:



Uncertainty in text similarity (within *stylo*)

- Random sampling tricks
- (Bootstrap) consensus trees (Eder 2013)
- (Bootstrap) consensus networks (Eder 2017)
- General Imposters (Kestemont et al. 2016)

2. Sampling & bootstrapping

Sample:



$$p(\text{square}) = 0.66$$

2. Sampling & bootstrapping

Sample:



$$p(\text{square}) = 0.66$$

2. Sampling & bootstrapping

Sample:  $p(\text{square}) = 0.66$

Resample 1:  0.5

2. Sampling & bootstrapping

Sample:  $p(\text{square}) = 0.66$

Resample 1:  0.5

Resample 2:  0.66

2. Sampling & bootstrapping

Sample:  $p(\text{square}) = 0.66$

Resample 1:  0.5

Resample 2:  0.66

Resample 3:  0.33

2. Sampling & bootstrapping

Sample:  $p(\text{square}) = 0.66$

Resample 1:  0.5

Resample 2:  0.66

Resample 3:  0.33

Resample 4:  1

2. Sampling & bootstrapping

Sample:



$$p(\text{square}) = 0.66$$

Resample 1:



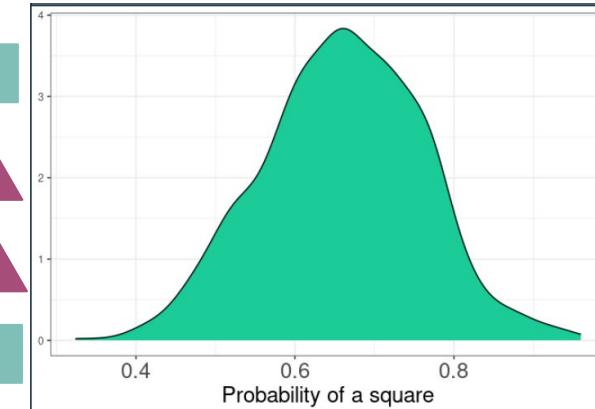
Resample 2:



Resample 3:



Resample 4:



2. Sampling & bootstrapping

Sample:  $p(\text{square}) = 0.66$

Resample 1:  0.5

2. Sampling & bootstrapping

Sample:  $p(\text{square}) = 0.66$

Resample 1:  0.5

Resample 2:  0.66

2. Sampling & bootstrapping

Sample:  $p(\text{square}) = 0.66$

Resample 1:  0.5

Resample 2:  0.66

Resample 3:  0.33

2. Sampling & bootstrapping

Sample:  $p(\text{square}) = 0.66$

Resample 1:  0.5

Resample 2:  0.66

Resample 3:  0.33

Resample 4:  1

2. Sampling & bootstrapping

Sample:



$$p(\text{square}) = 0.66$$

Resample 1:



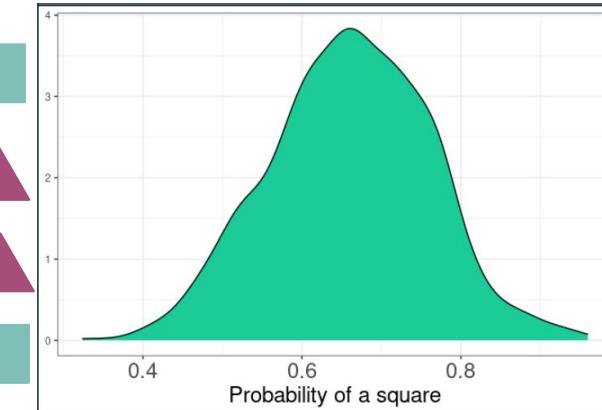
Resample 2:



Resample 3:



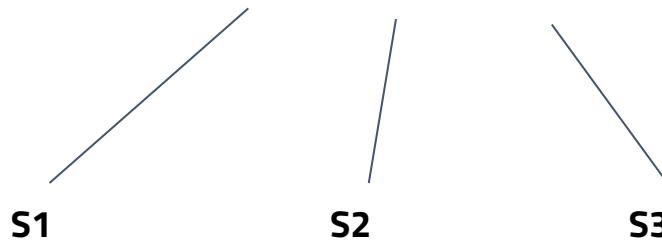
Resample 4:



Normal vs. random sampling (in stylo)



size=2



Normal vs. random sampling (in stylo)



size=4



s1

Normal vs. random sampling (in stylo)



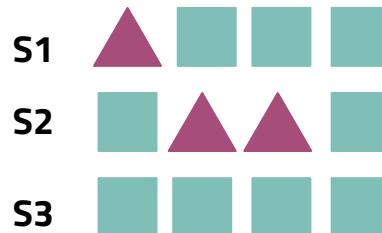
size=4



Normal vs. random sampling (in stylo)



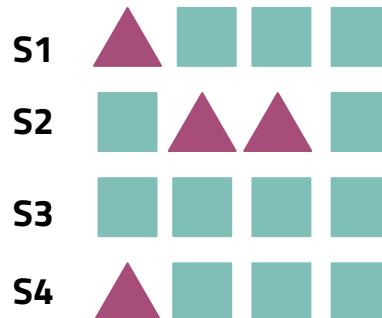
size=4



Normal vs. random sampling (in stylo)



size=4



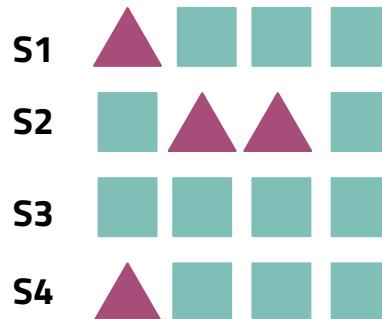
Normal vs. random sampling (in stylo)

6

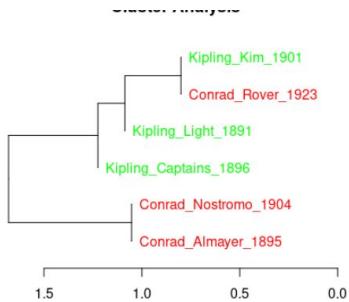


n=4

16

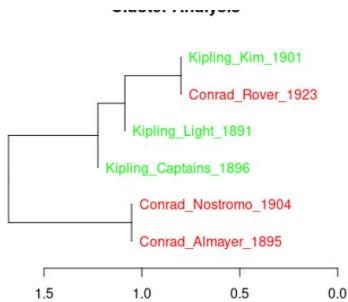


4. Consensus trees

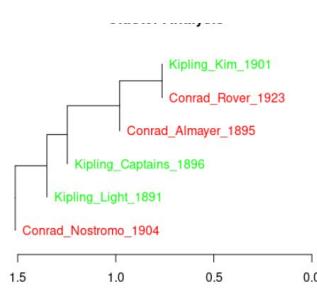


Feature set 1

4. Consensus trees

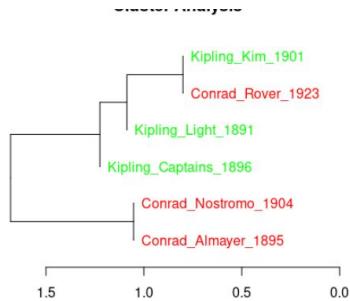


Feature set 1

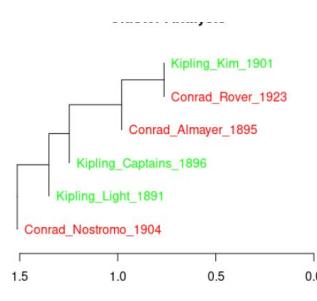


Feature set 2

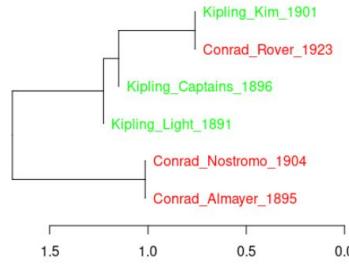
4. Consensus trees



Feature set 1

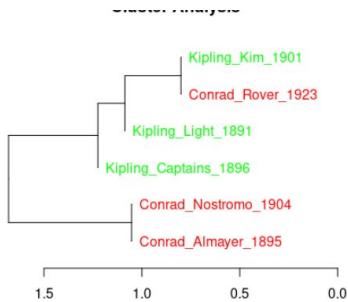


Feature set 2

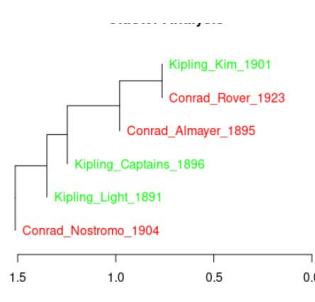


Feature set 3

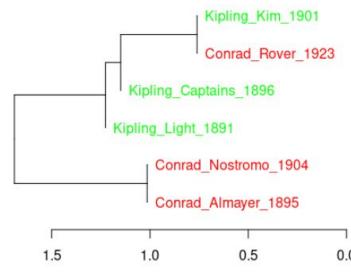
4. Consensus trees



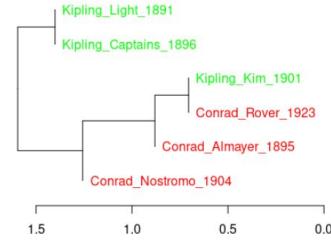
Feature set 1



Feature set 2

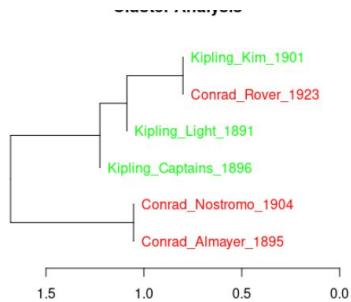


Feature set 3

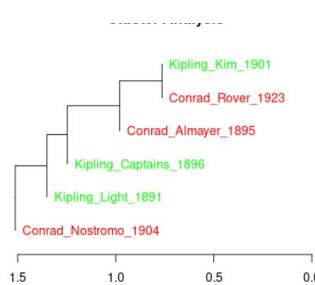


Feature set 4

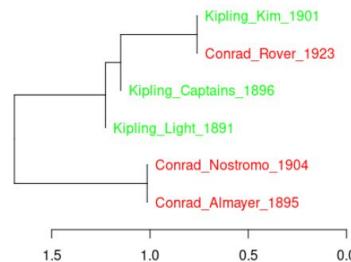
4. Majority rule (>50% of branches)



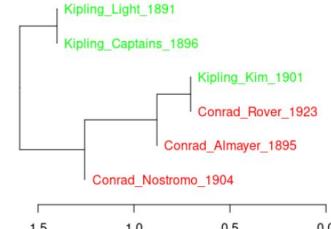
Feature set 1



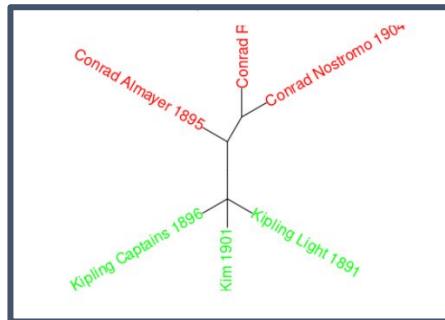
Feature set 2



Feature set 3



Feature set 4



5. Consensus trees

Using `stylo()` off the shelf you can “bootstrap”:

- MFW length
- Culling strength
- Text themselves (take samples from texts)

5. Consensus trees

Using `stylo()` off the shelf you can “bootstrap”:

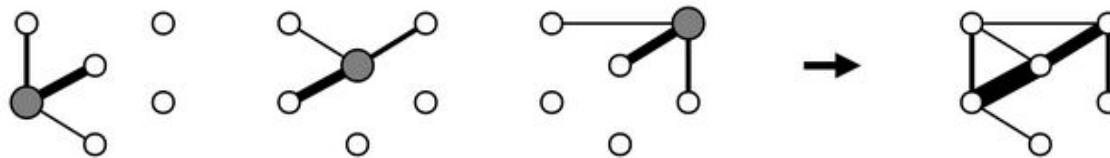
- MFW length
- Culling strength
- Text themselves (take samples from texts)

....

But the possibilities are limitless

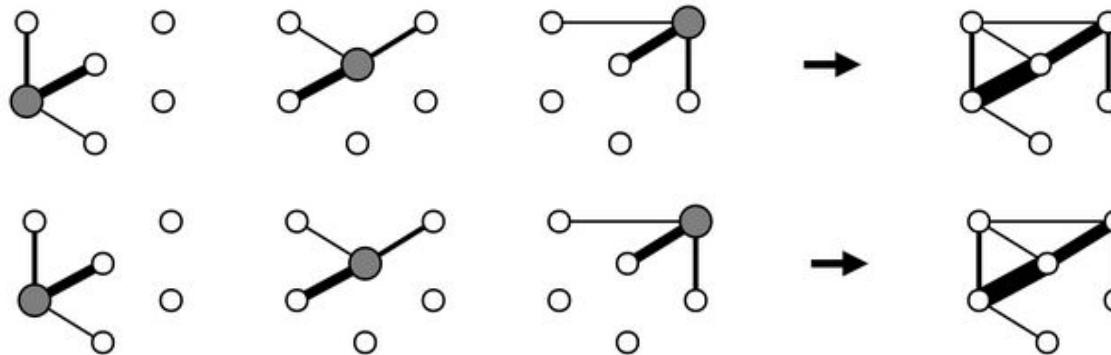
6. Consensus networks

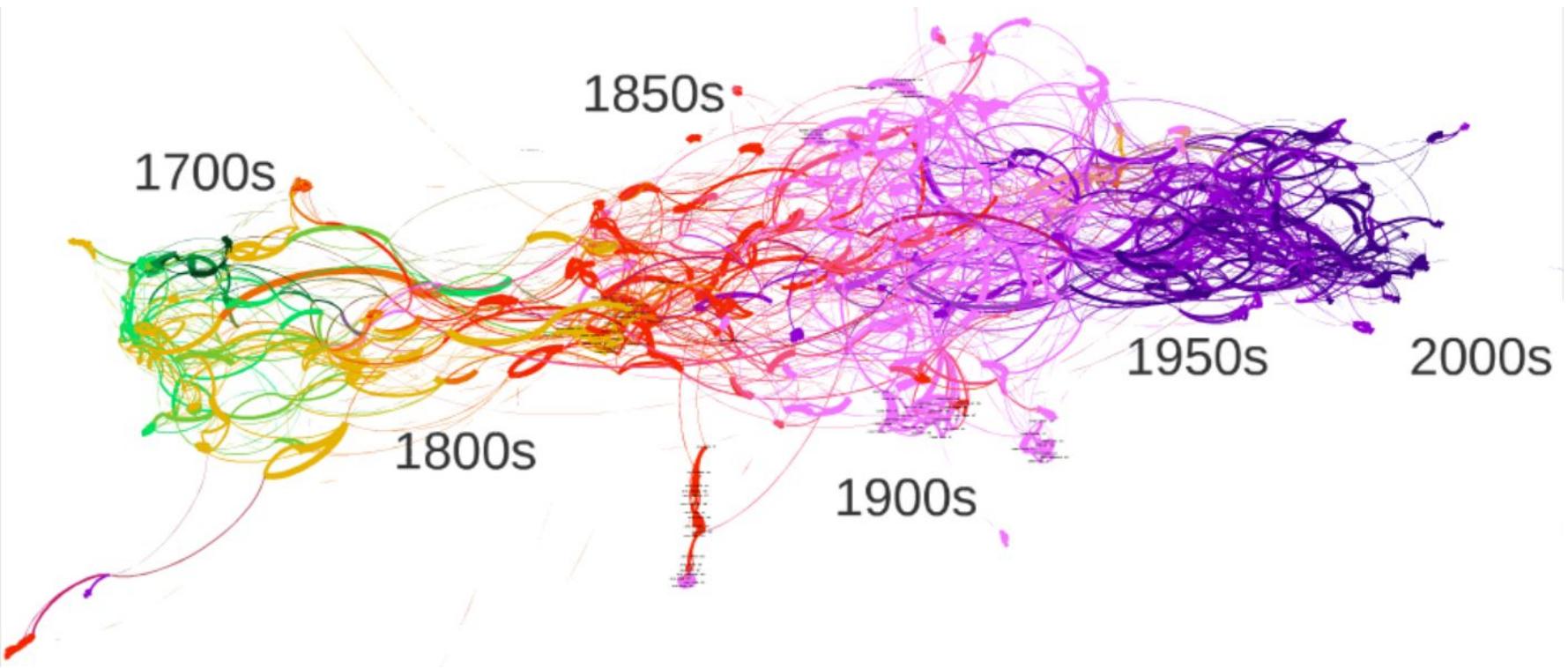
1. Look at the neighbours!



6. Consensus networks

1. Look at the neighbours!
2. Then look at the neighbours many times!



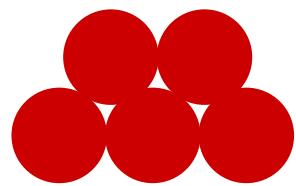


- Try using `stylo.network()` (alpha version!)
- Or brave the depths of Gephi
- Or work with networks from R!
 - Best tutorial I know:
 - **<https://kateto.net/network-visualization>**

6. General imposters

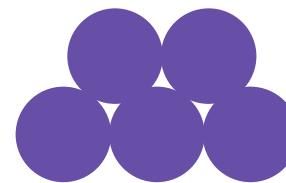
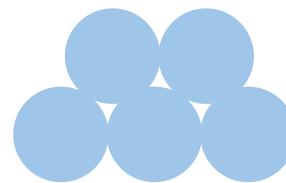
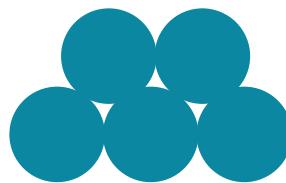
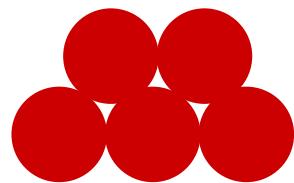


6. General imposters

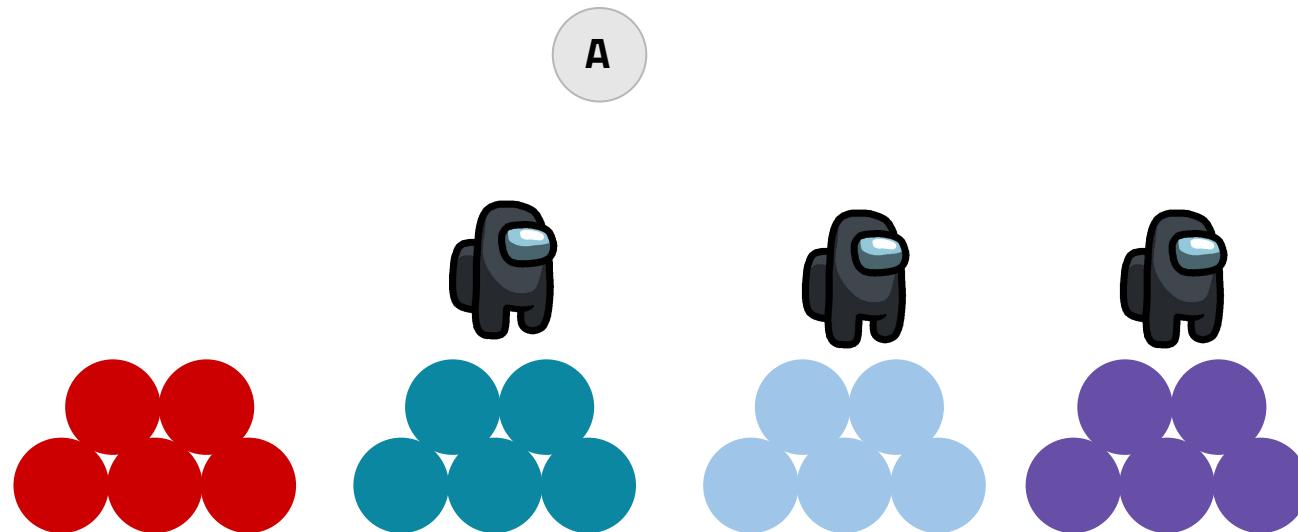


6. General imposters

A



6. General imposters



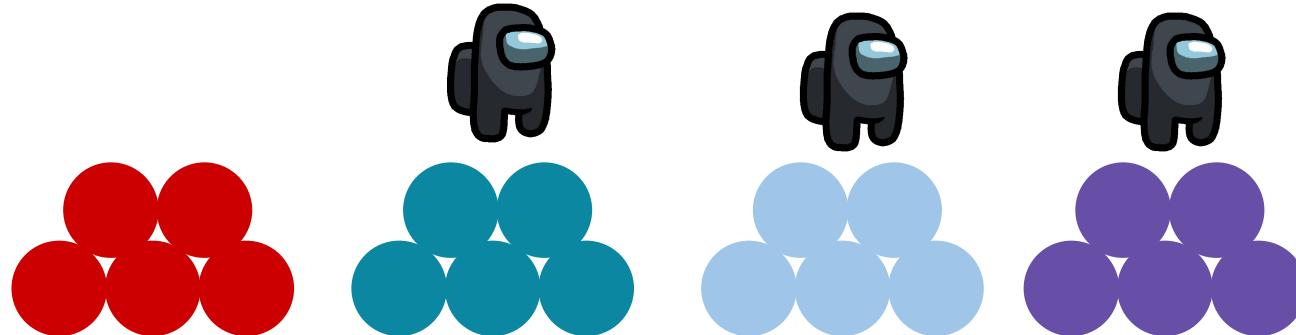
6. General imposters

Random samples

Random features

Random imposters

A



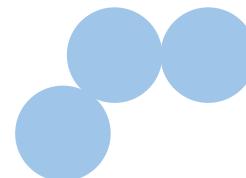
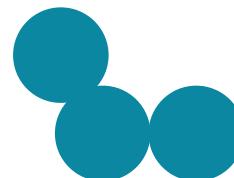
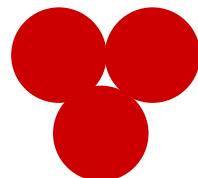
6. General imposters

Random samples

Random features

Random imposters

A



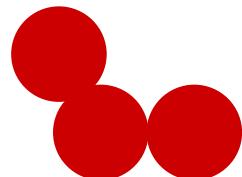
6. General imposters

Random samples

Random features

Random imposters

A



6. General imposters

