Computational text analysis: Survival guide

ESU 24 @ Cluj-Napoca

Jeremi Ochab, Artjoms Šeļa

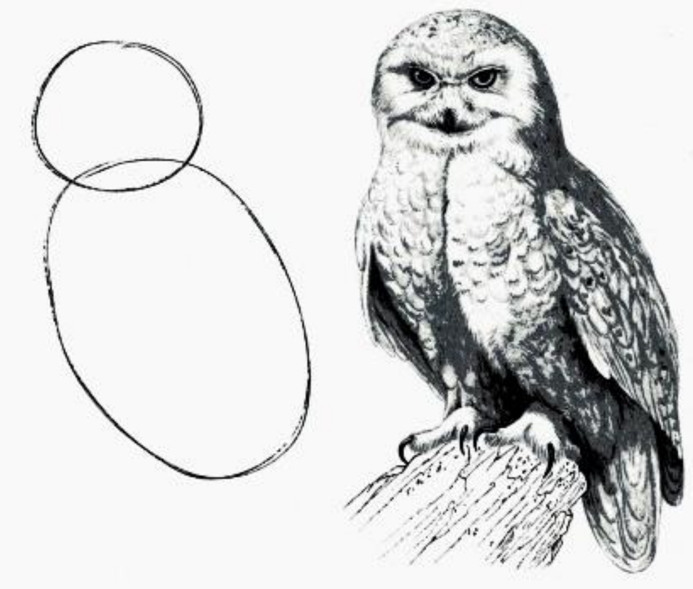


Fig 1. Draw two circles

Fig 2. Draw the rest of the damn Owl

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

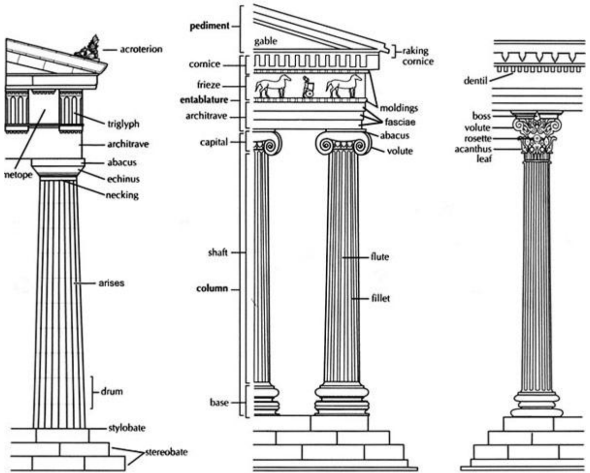
- How to define your own space, populate it with texts and calculate routes between them?

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

- <u>New Shakespeare Society</u> 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 1964

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

T.C. Mendenhall: word lengths

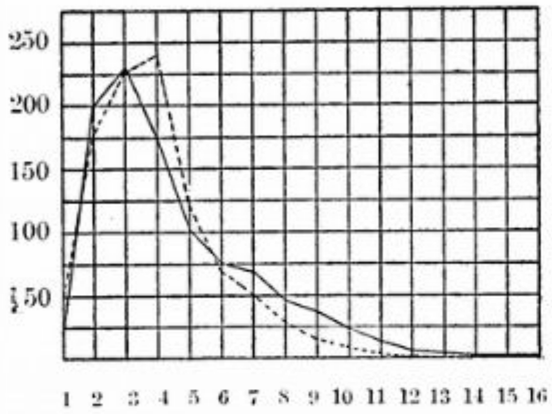


FIG. 1: Relative frequencies (per mille) of word-lengths measured by number of characters in works of W. Shakespeare (dashed) and F. Bacon (full line). Source: Mendenhall 1901: 104 (facsimile).

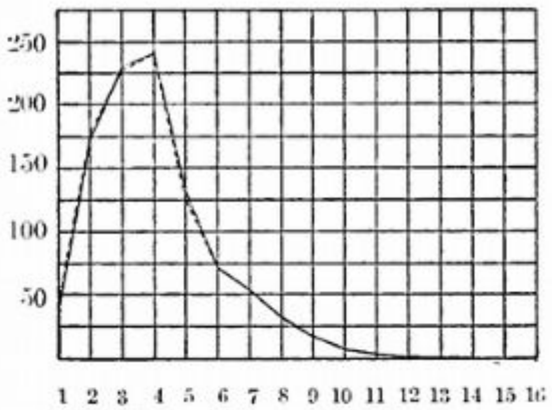
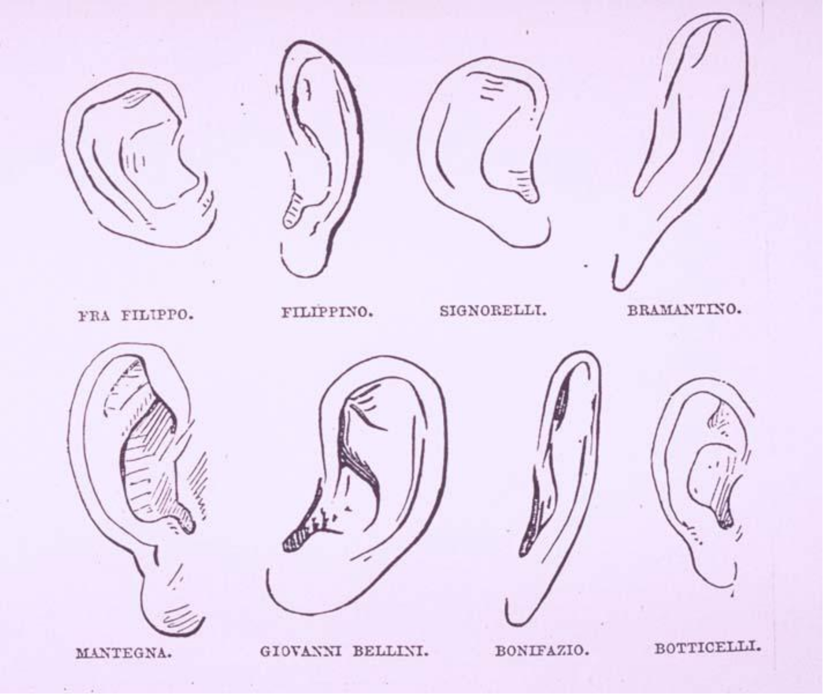
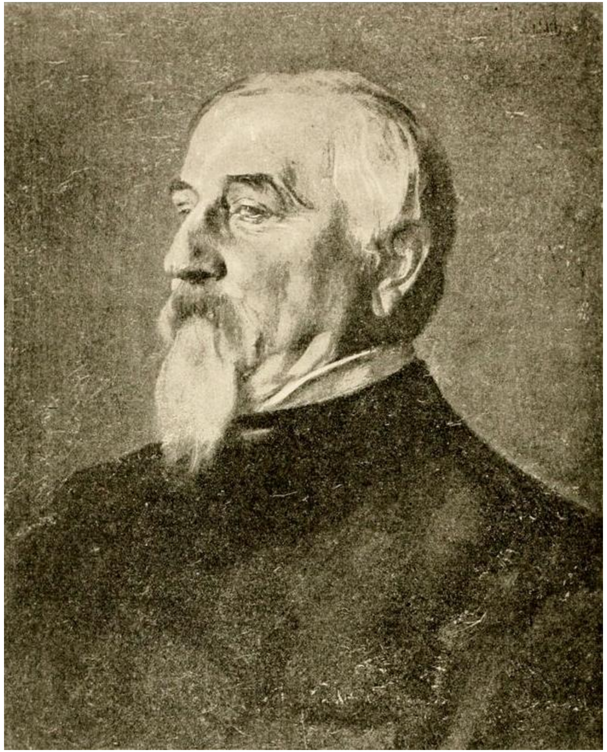


FIG. 2: Relative frequencies (per mille) of word-lengths measured by number of characters in works of W. Shakespeare (dashed) and C. Marlowe (full line). Source: Mendenhall 1901: 105 (facsimile).

Study of authorship in paintings

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





“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

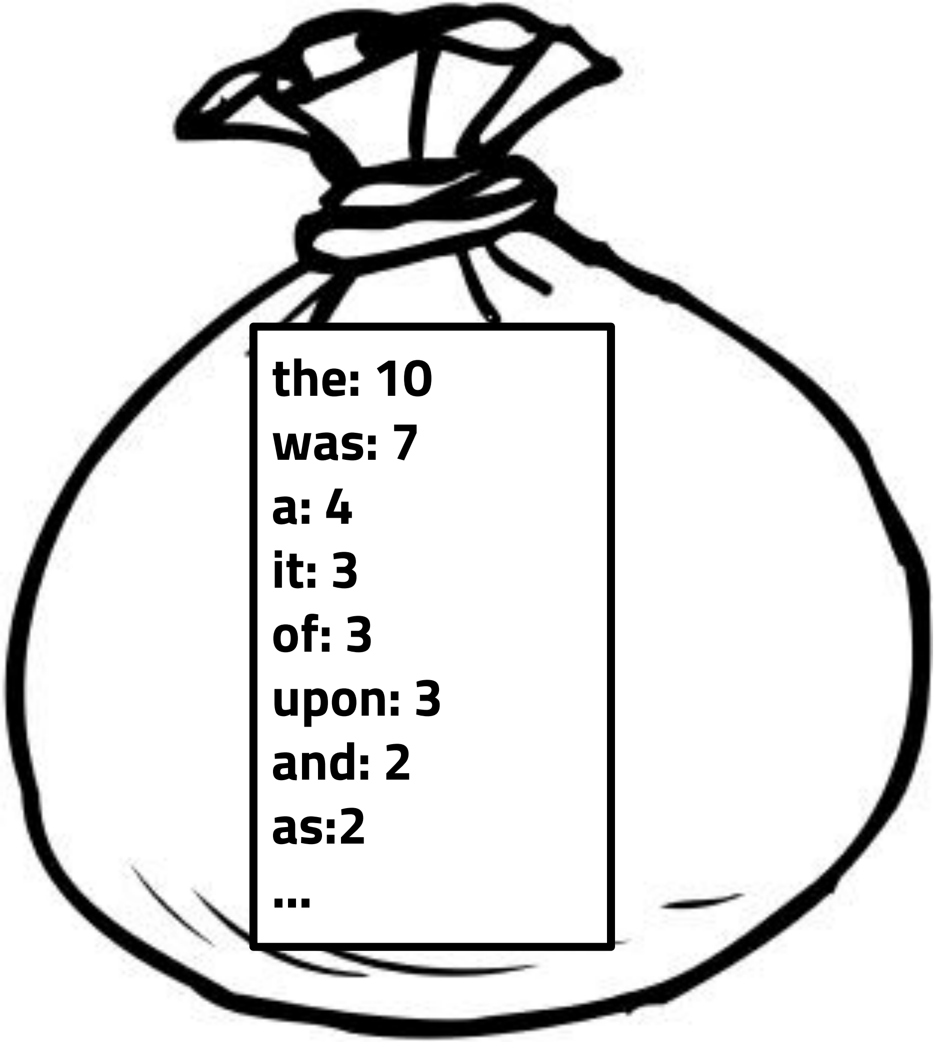
Mr. Sherlock Holmes, who was usually very late in the mornings, save upon those not infrequent occasions when he was up all night, was seated at the breakfast table. I stood upon the hearth-rug and picked up the stick which our visitor had left behind him the night before. It was a fine, thick piece 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 … ne mornings, save upon those not
infrequent occasions when h … seated at the breakfast table. I stood
upon the hearth-rug and pick … ur visitor had left behind him the night
before. It was a fine, thick pie … eaded, of the sort which is known as a
“Penang lawyer.” Just under … silver band nearly an inch across. “To
James Mortimer, M.R.C.S., f … .C.H.,” was engraved upon it, with the
date "1884." It was just such … oned family practitioner used to carry –
dignified, solid, and reassurir

Silly things: b

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 woo
“Penang lawyer.” Just under the head
James Mortimer, M.R.C.S., from his fri
date "1884." It was just such a stick as
dignified, solid, and reassuring.



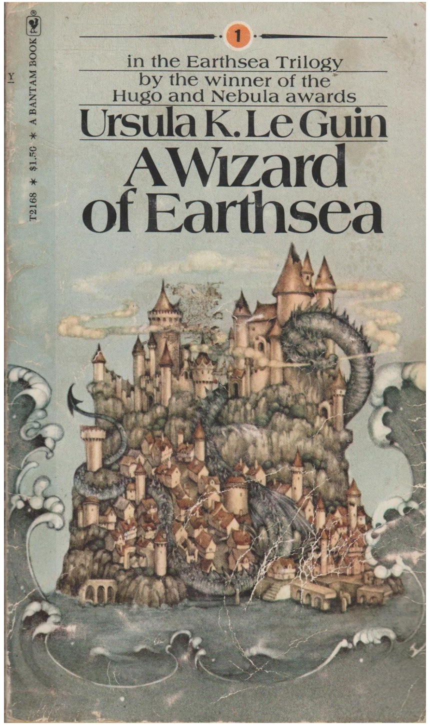
Silly things: bags of words

Mr. Sherlock Holmes, who w ne mornings, save upon those not
infrequent occasions when h seated at the breakfast table. I stood
upon the hearth-rug and pick ur visitor had left behind him the night
before. It was a fine, thick pie eaded, of the sort which is known as a
“Penang lawyer.” Just under silver band nearly an inch across. “To
James Mortimer, M.R.C.S., f .C.H.,” was engraved upon it, with the
date "1884." It was just such oned family practitioner used to carry –
dignified, solid, and reassurir

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

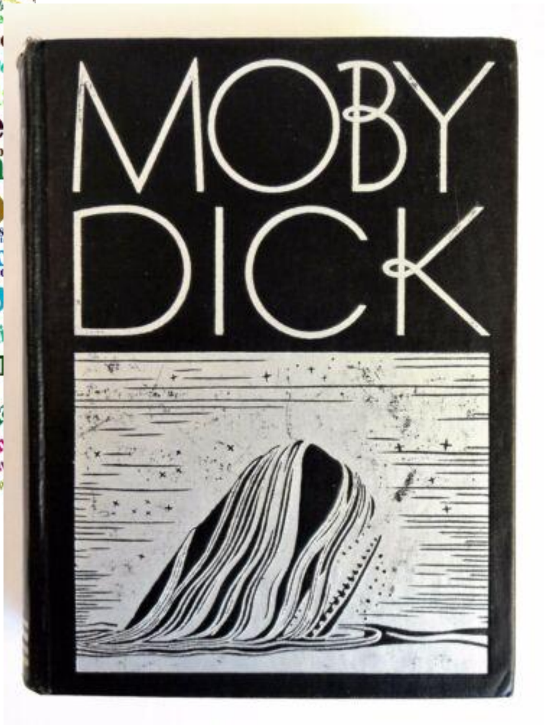
Silly, but useful!

Silly, but useful!



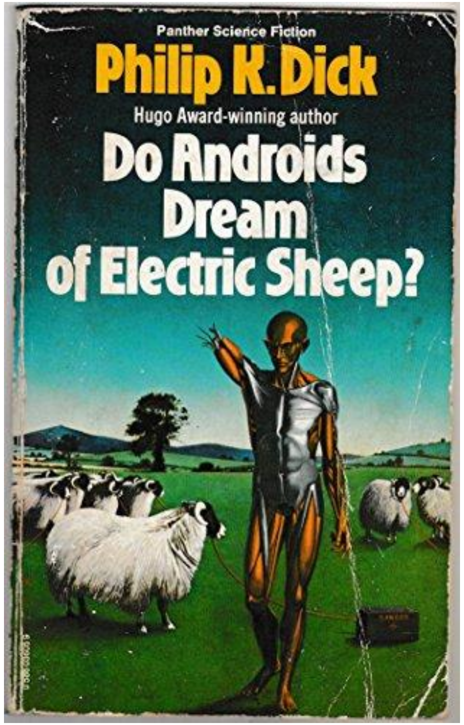
Silly, but useful!

Silly, but useful!



Silly, but useful!

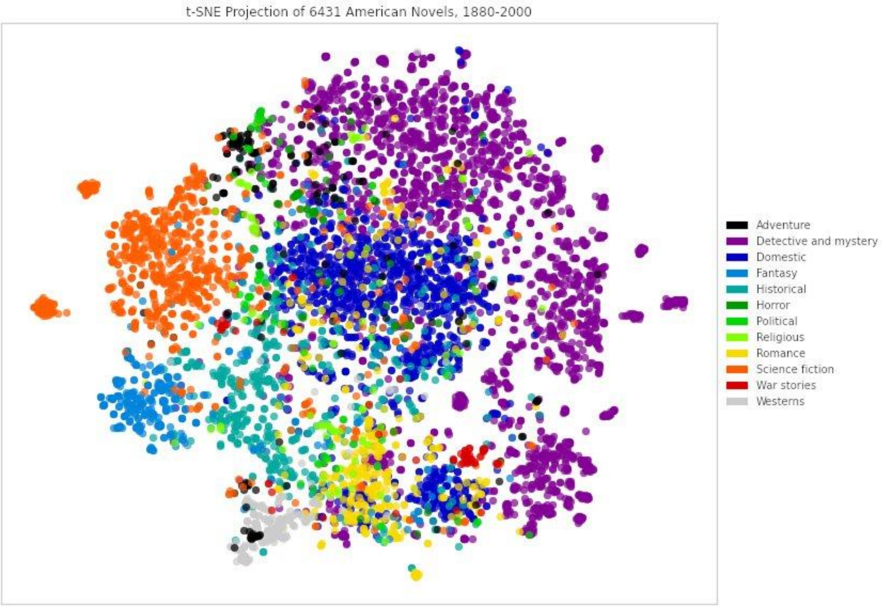
Silly, but useful!



Proxies

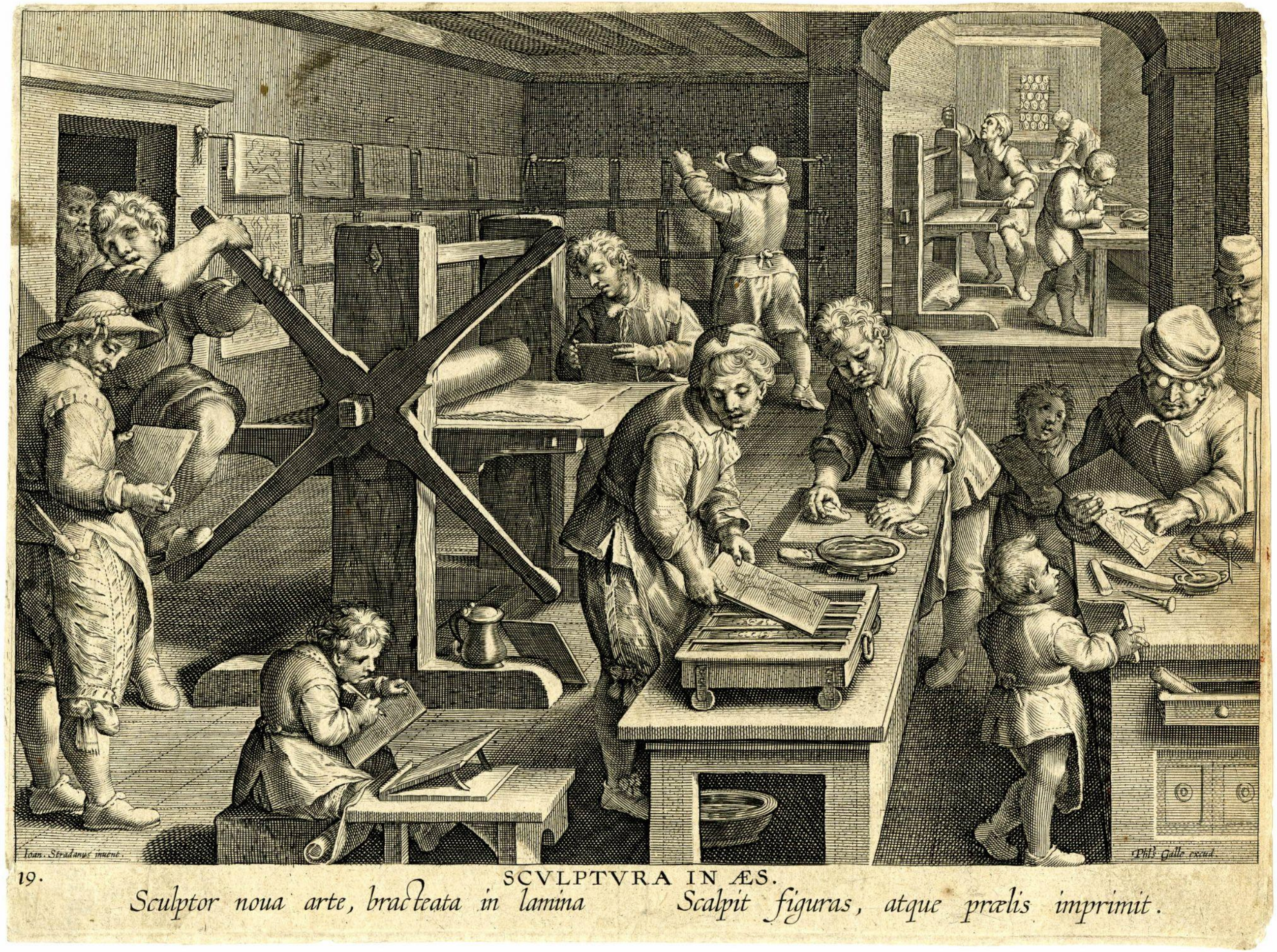
- Word frequencies may serve as a proxy to **things we care about** in texts
- Word frequencies are the result of word choice -> word choice is a result of **forces that organize texts** (cultural and social conditions)

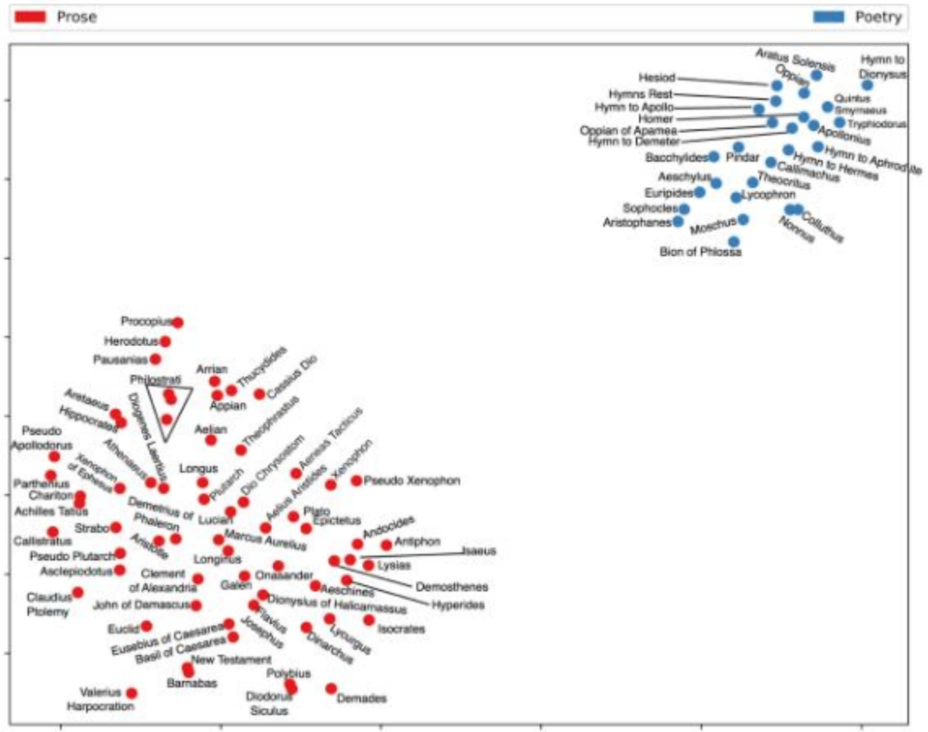
Text-dice rolling will be organized by multitude of forces



Can we tell apart…

prose from poetry?

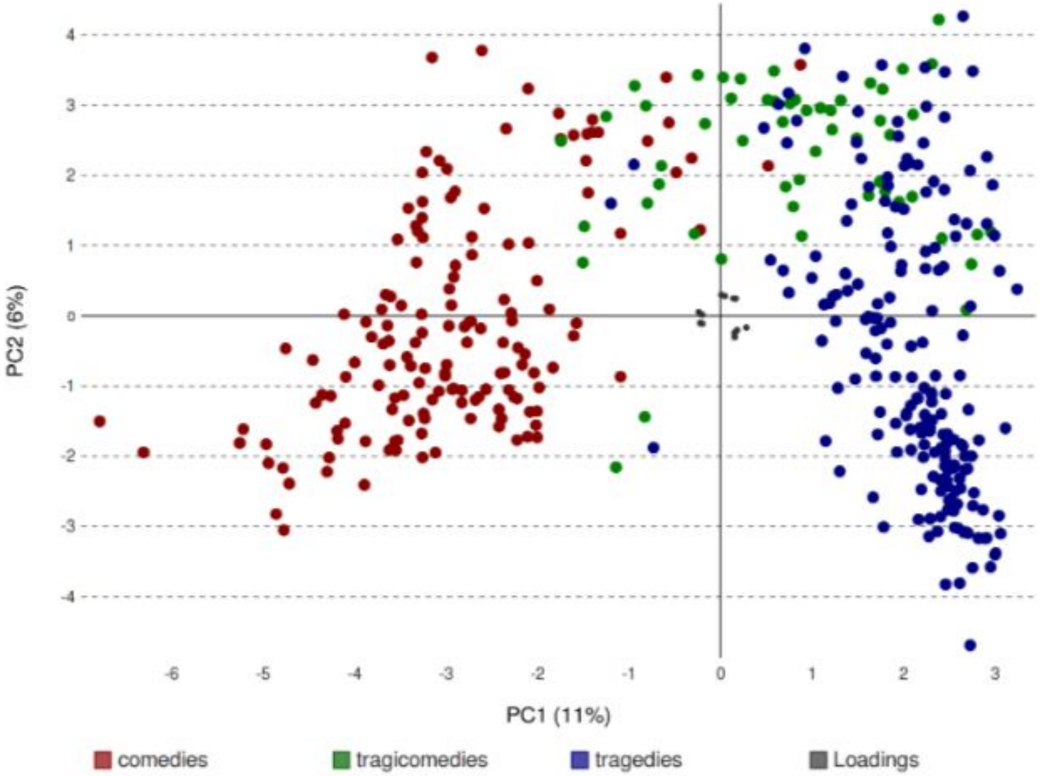




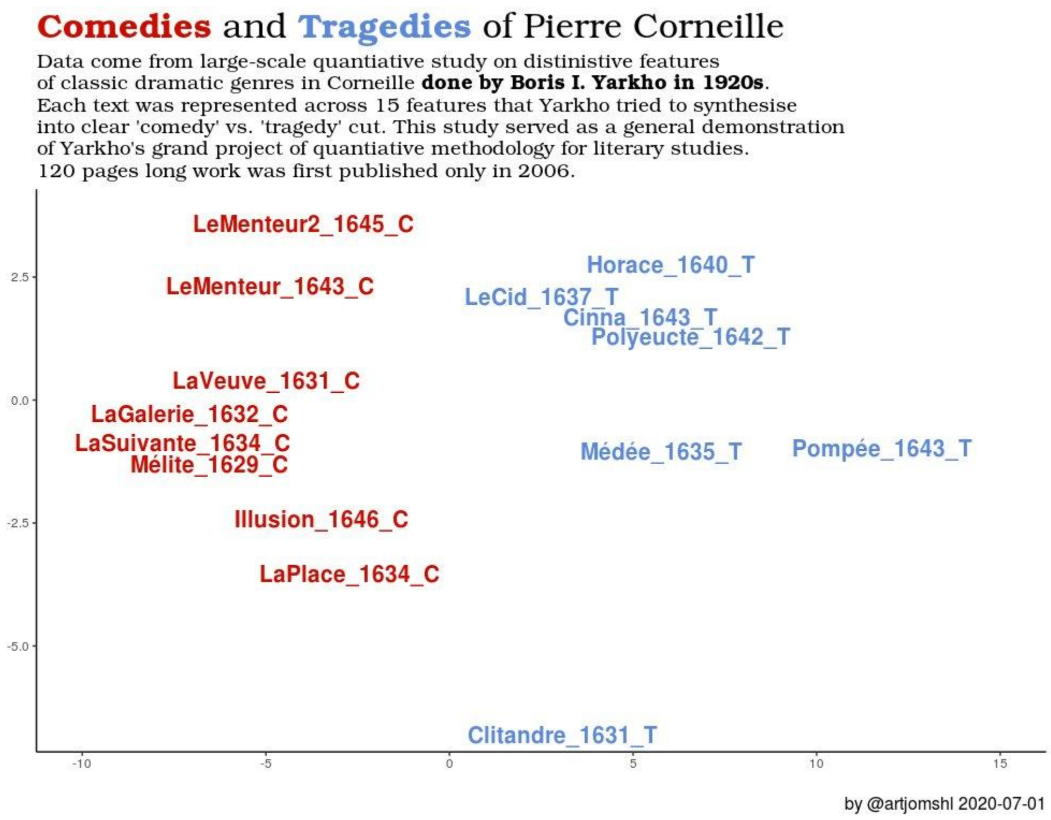
Storey & Mimno 2020: Like Two Pis in a Pod: Author Similarity Across Time in the Ancient Greek Corpus

Can we tell apart…

comedy from tragedy?



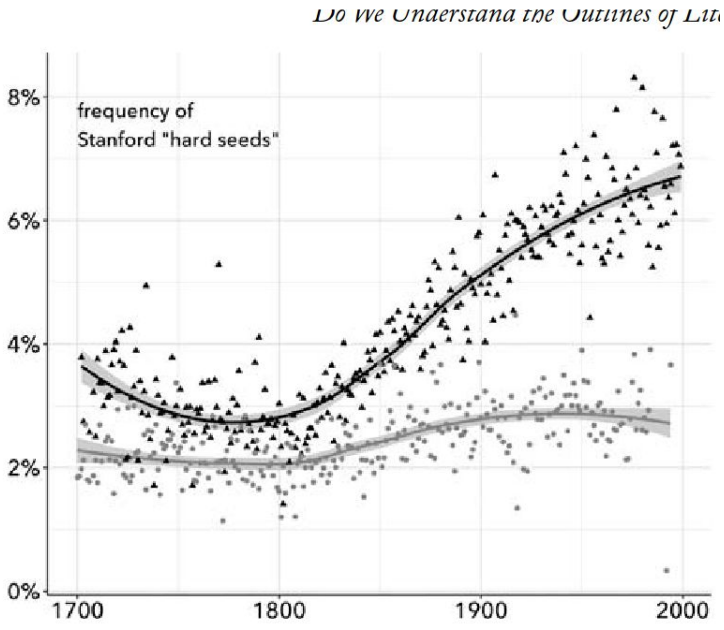
Schöch, C. 2017. Topic Modeling Genre…

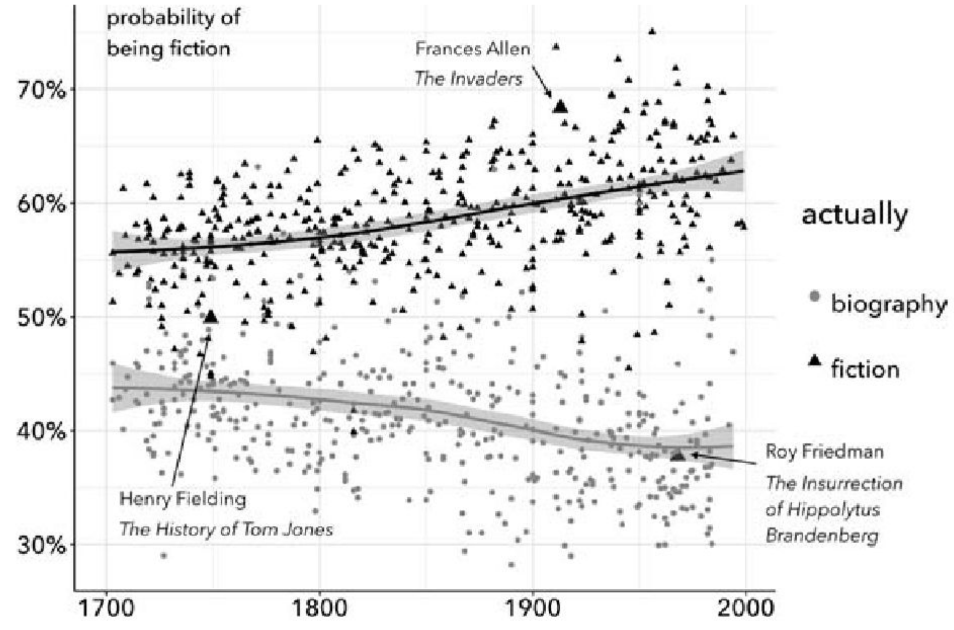


Boris Yarkho (1889-1942)

Can we tell apart…

Fiction from non-fiction?

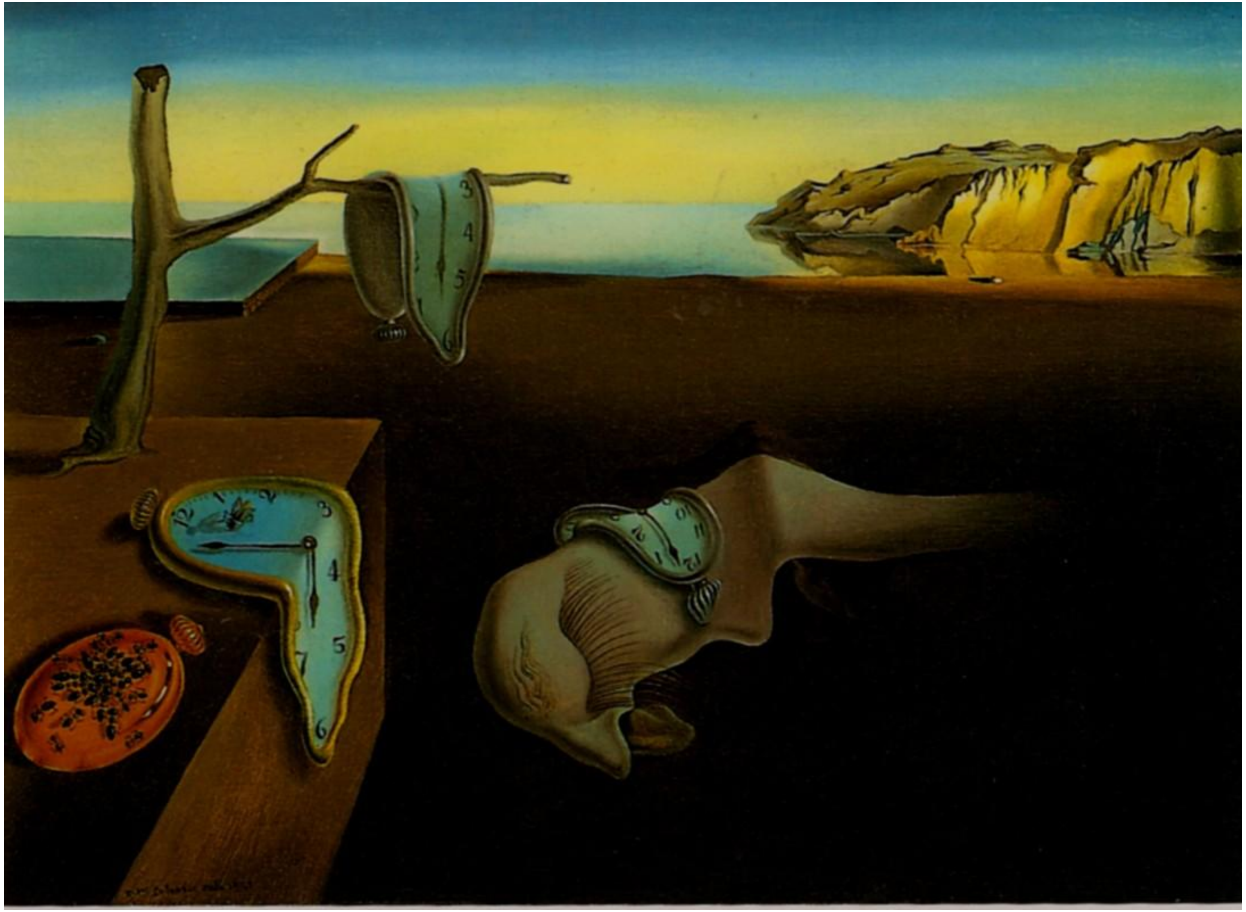


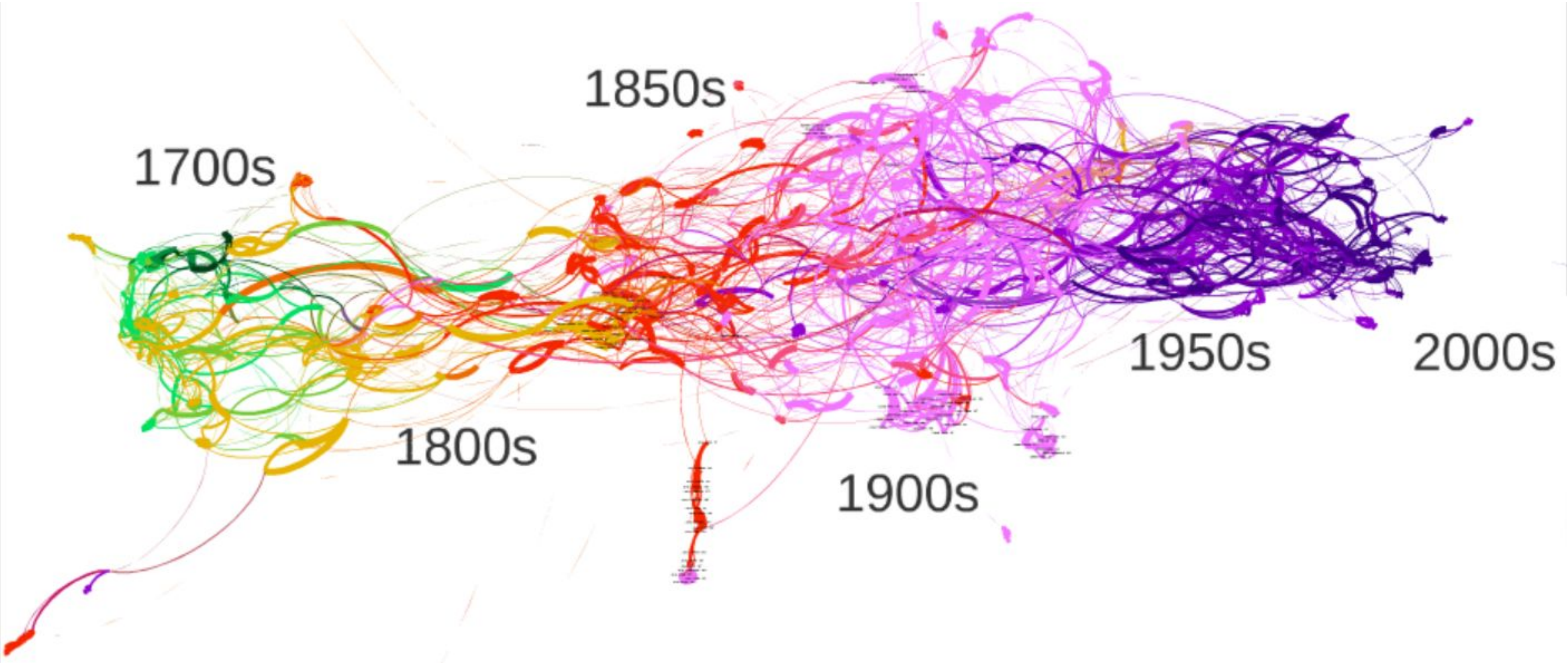


Underwood T. 2019 *Distant Horizons*

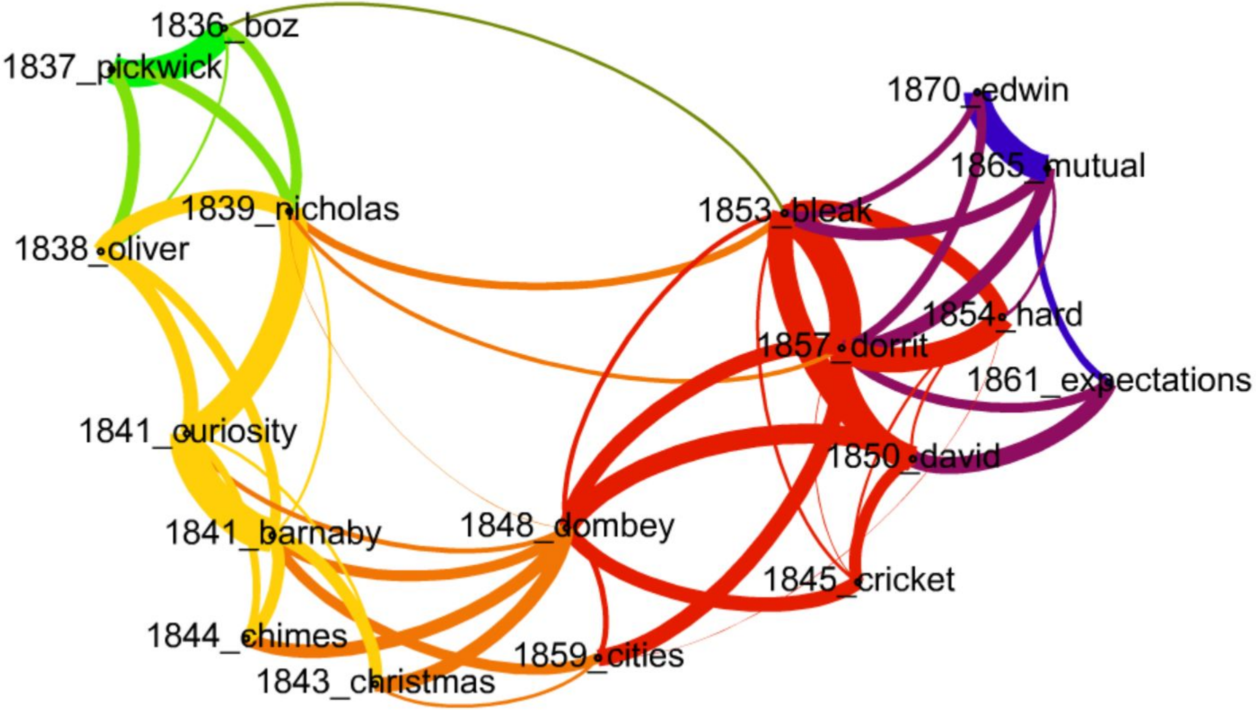
Can we tell apart…

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





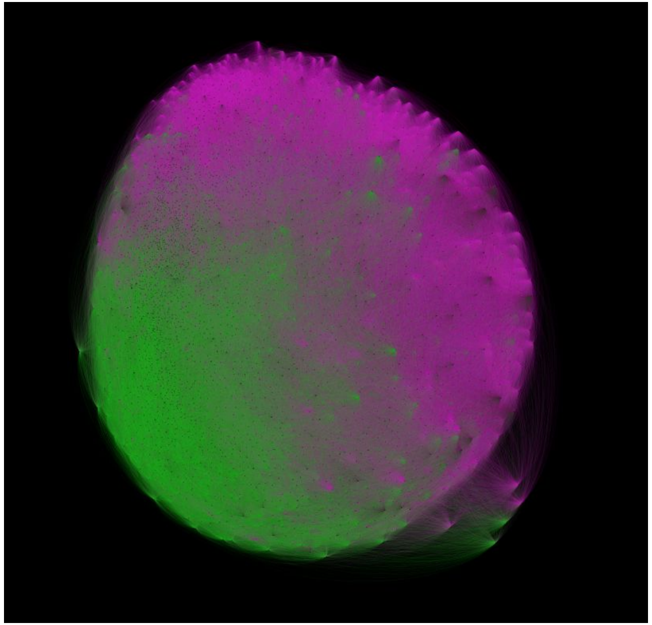
Rybicky 2016



Rybicky 2016

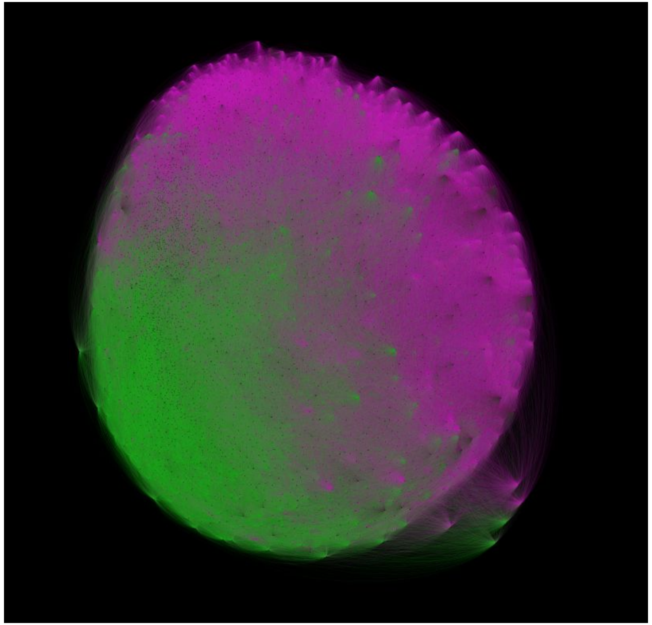
Can we tell apart…

a text written by <u>woman</u> from text written by <u>man</u>?

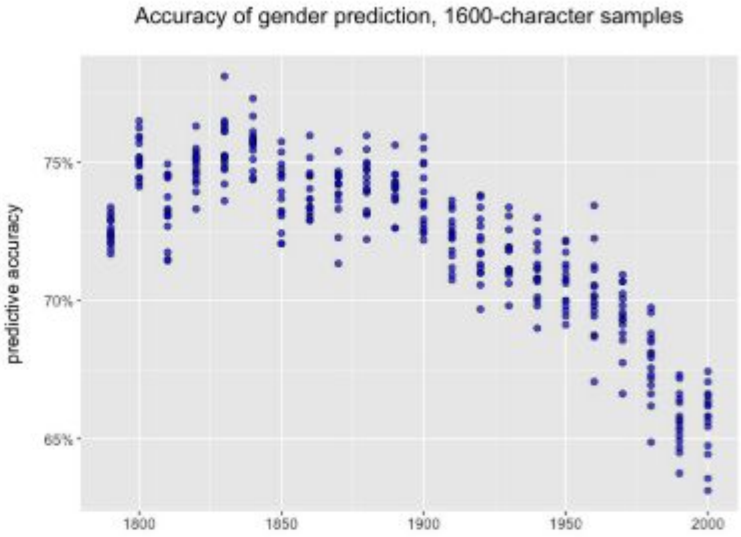


Gender is color-coded

Jockers 2013 *Macroanalysis*

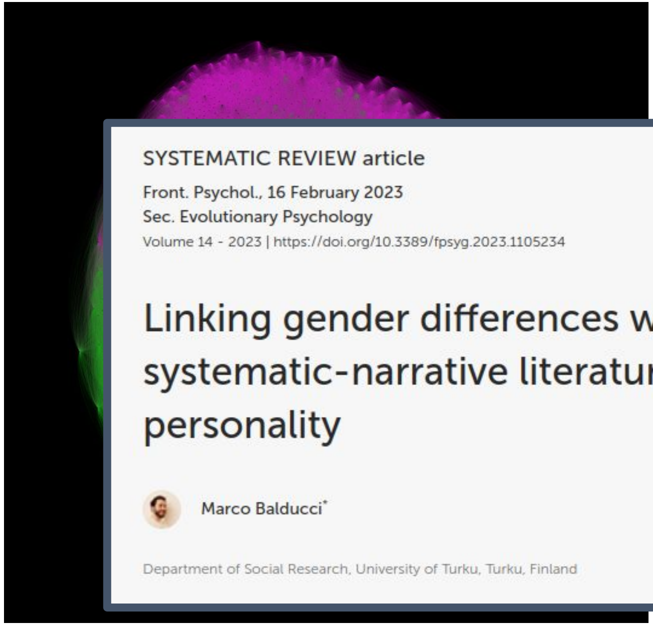


Jockers 2013 *Macroanalysis*



F/M characters recognizability decreases over time

Underwood et al. 2018



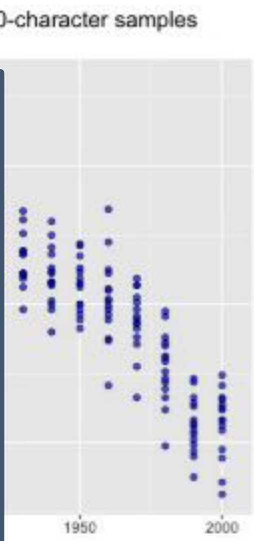
SYSTEMATIC REVIEW article

Front. Psychol., 16 February 2023
Sec. Evolutionary Psychology
Volume 14 - 2023 | https://doi.org/10.3389/fpsyg.2023.1105234

Linking gender differences with gender equality: A systematic-narrative literature review of basic skills and personality

Marco Balducci*

Department of Social Research, University of Turku, Turku, Finland



Gender is color-coded

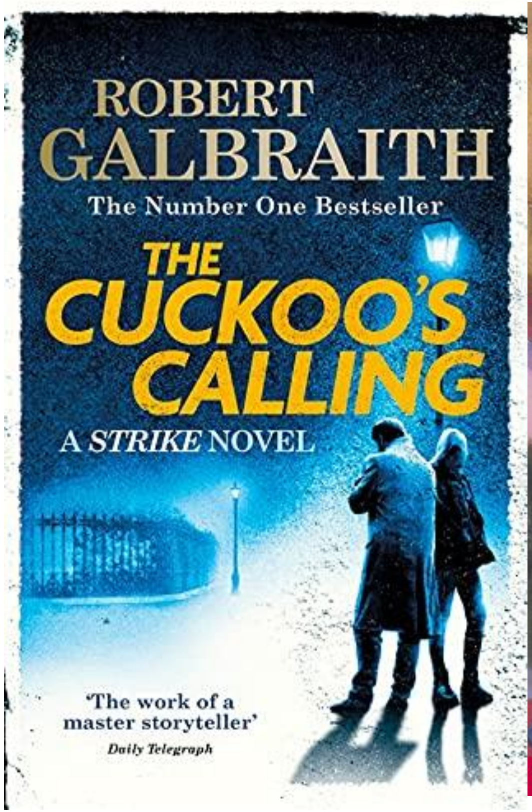
Jockers 2013 *Macroanalysis*

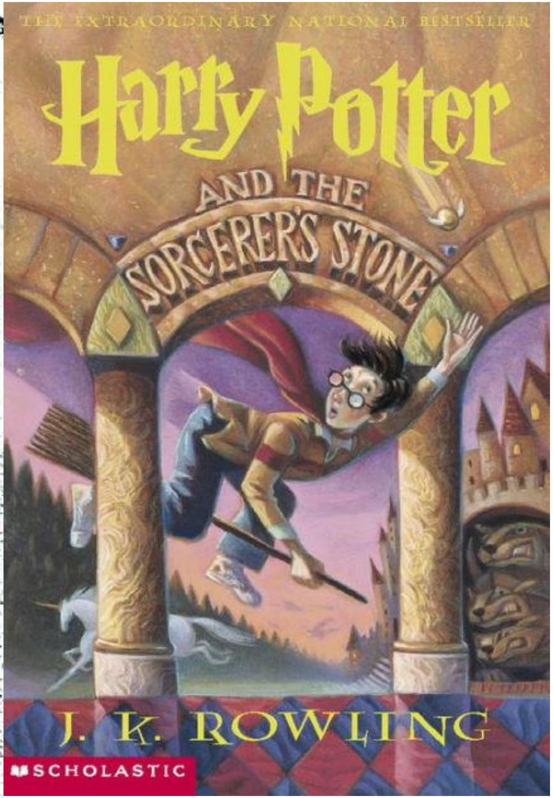
F/M characters decreases over time

Underwood et al. 2018

Can we tell apart…

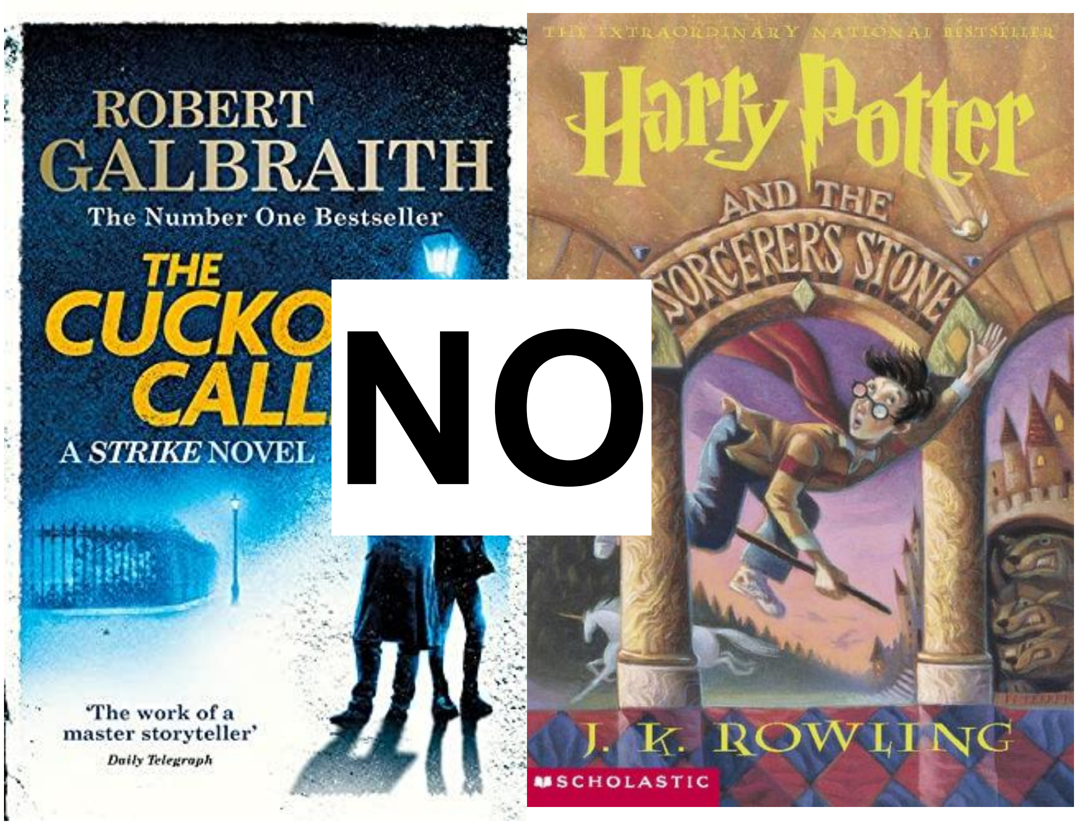
Robert Galbraith from J.K. Rowling?





Can we tell apart…

Robert Galbraith from J.K. Rowling?



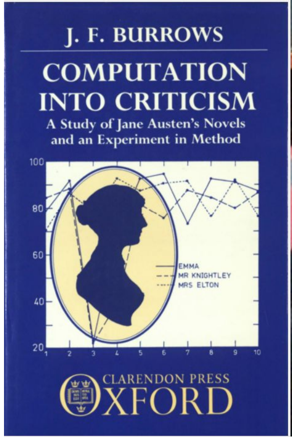
Curse & blessing of word frequencies

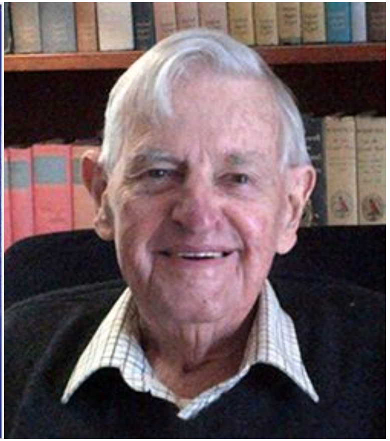
- **1. Authorship.** Two texts of the same author usually appear the closest to each other than to any outsider text
- **2. Modes of writing.** Fiction and nonfiction grew apart stylistically; Poetry books (esp. regularized verse) will always form a VERY exclusive party with themselves.
- **3. Genre.** 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)

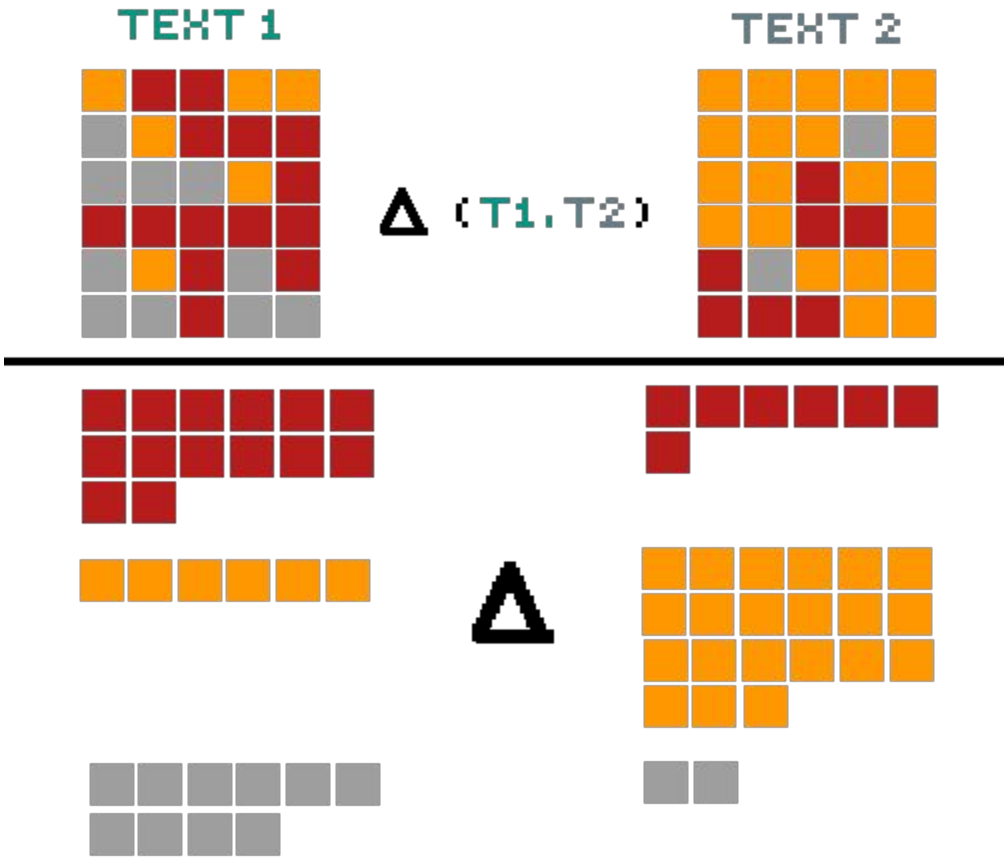
$$\Delta = \sum_{i=1}^{n} \frac{|z(x_i) - z(y_i)|}{n}$$

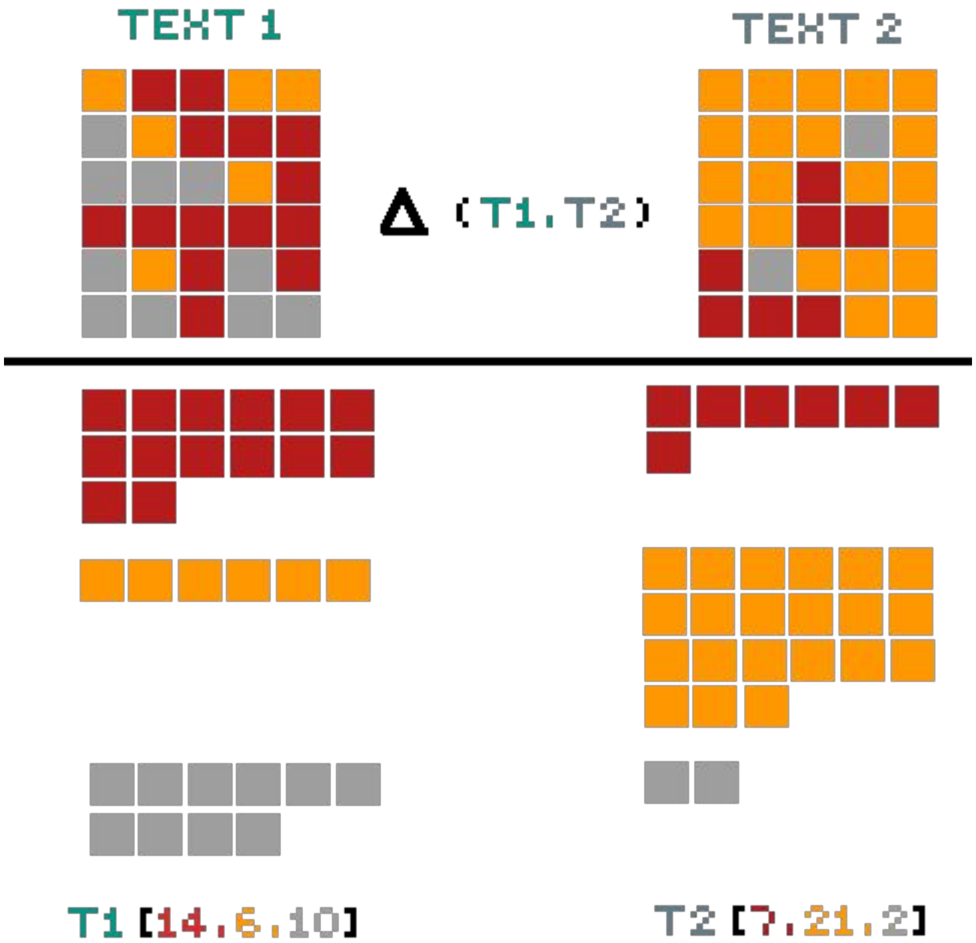


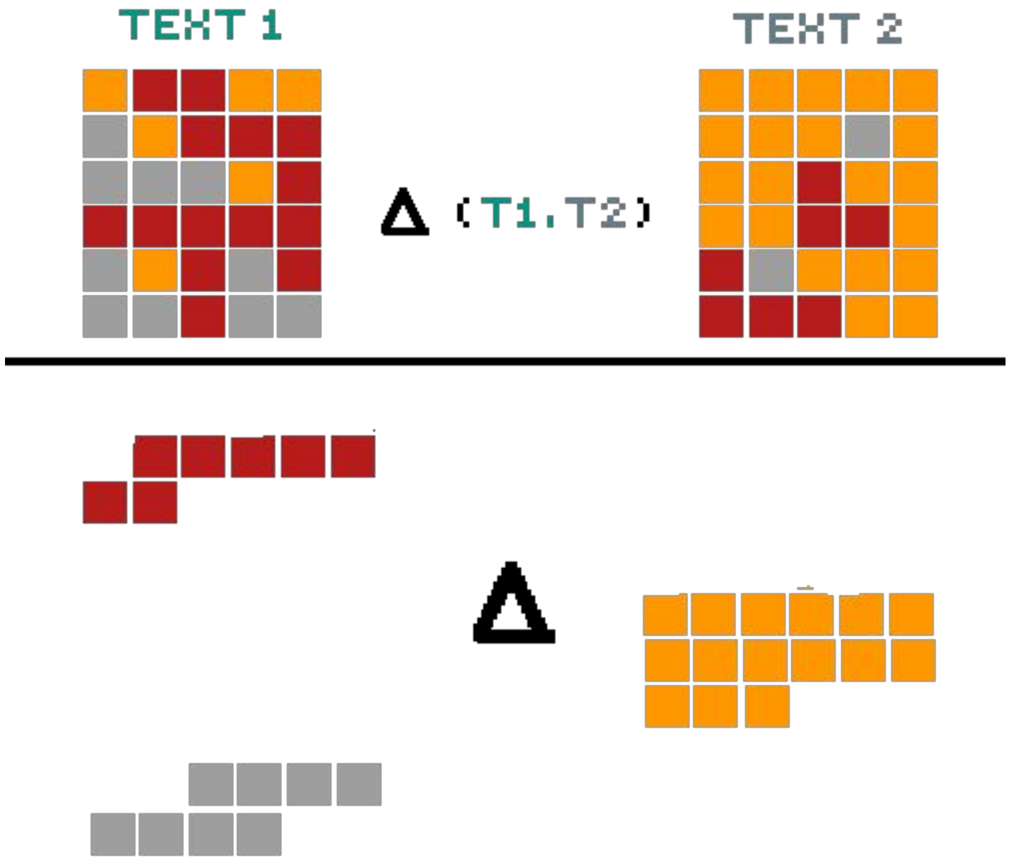


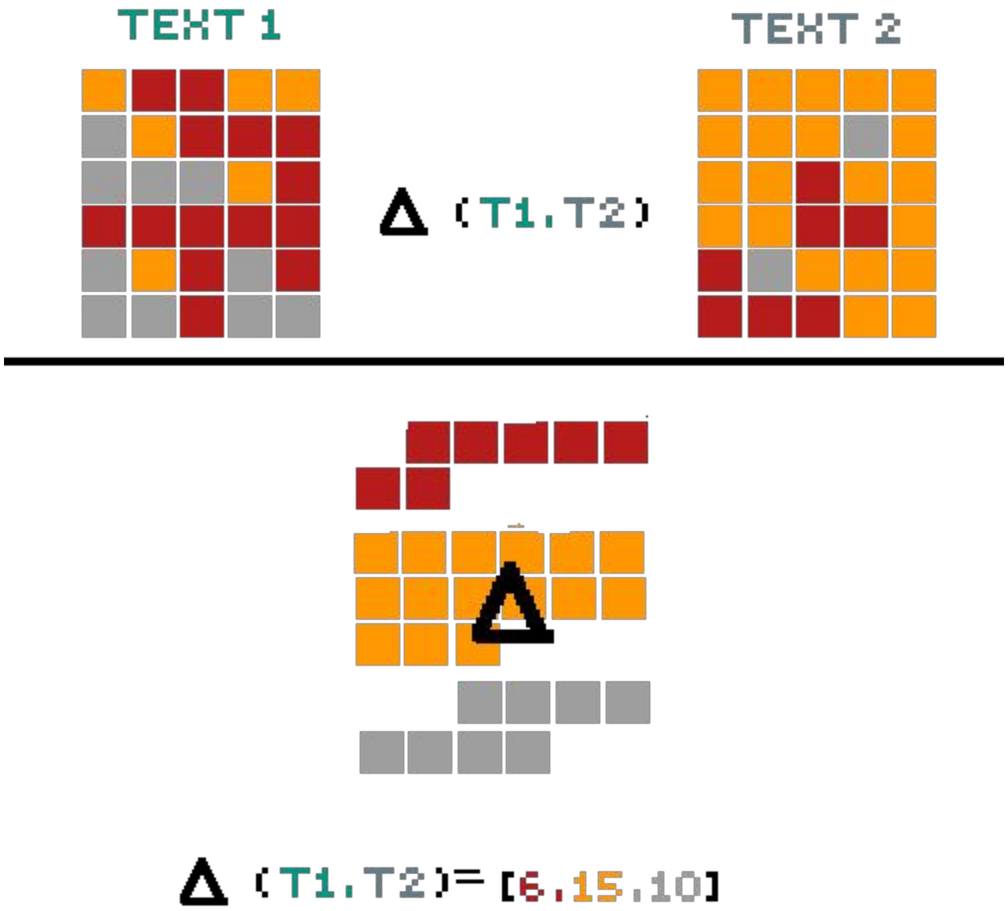
John Burrows (1928-2019)

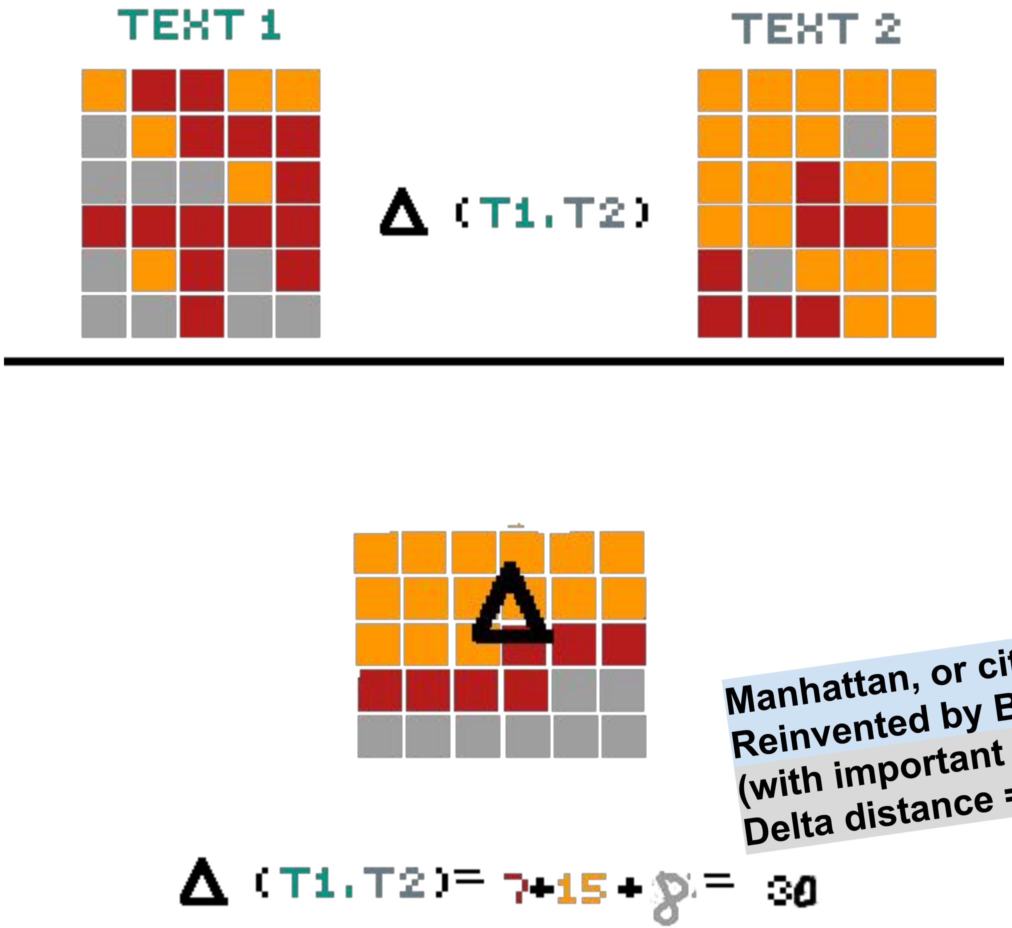

TEXT 1
Δ (T1,T2)
TEXT 2


TEXT 1
TEXT 2
Δ (T1,T2)
Δ


TEXT 1
TEXT 2
Δ (T1,T2)
T1 [14,6,10]
T2 [7,21,2]

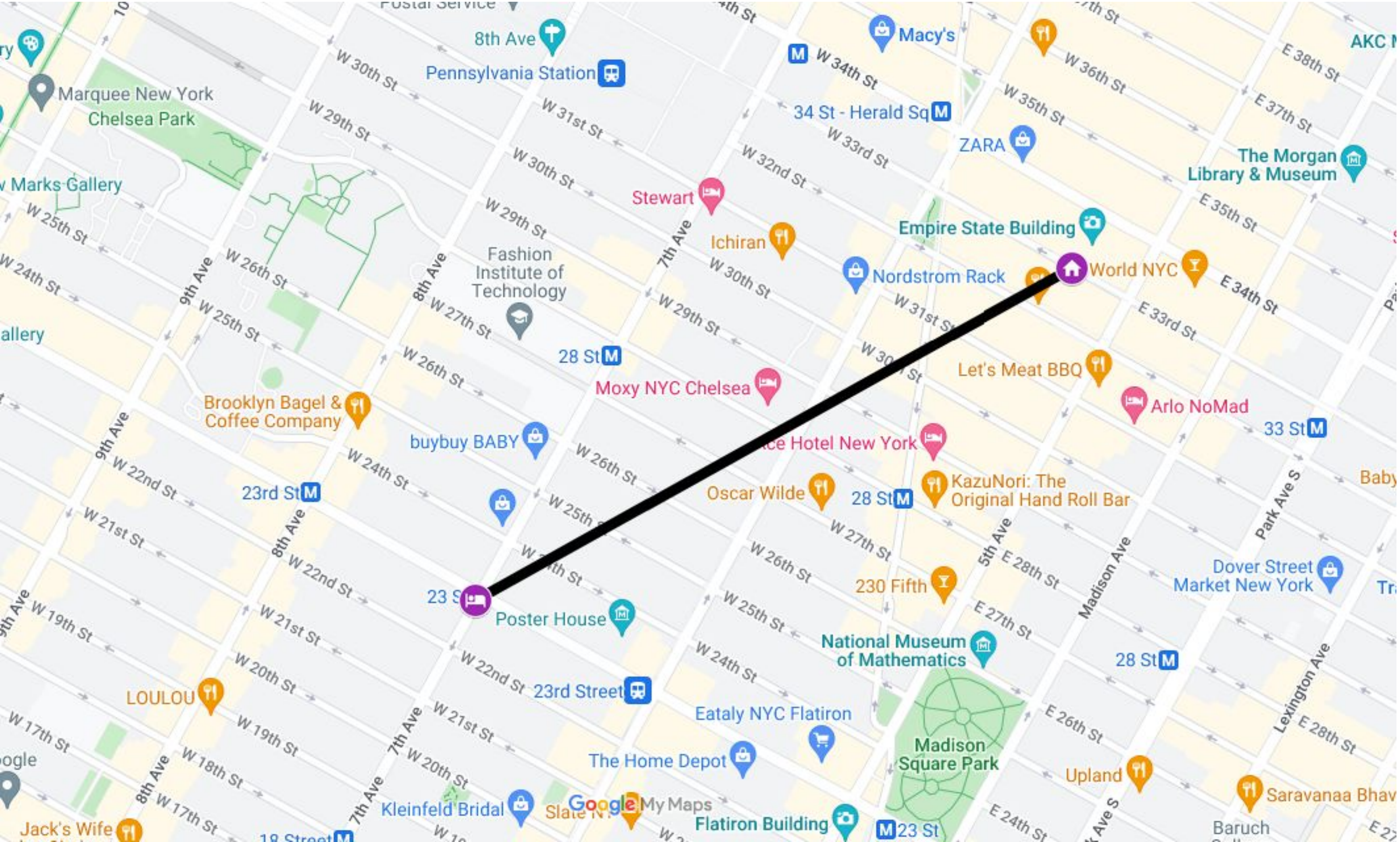

TEXT 1
TEXT 2
Δ (T1,T2)
Δ


TEXT 1
TEXT 2
Δ (T1,T2)
Δ (T1,T2)= [6,15,10]

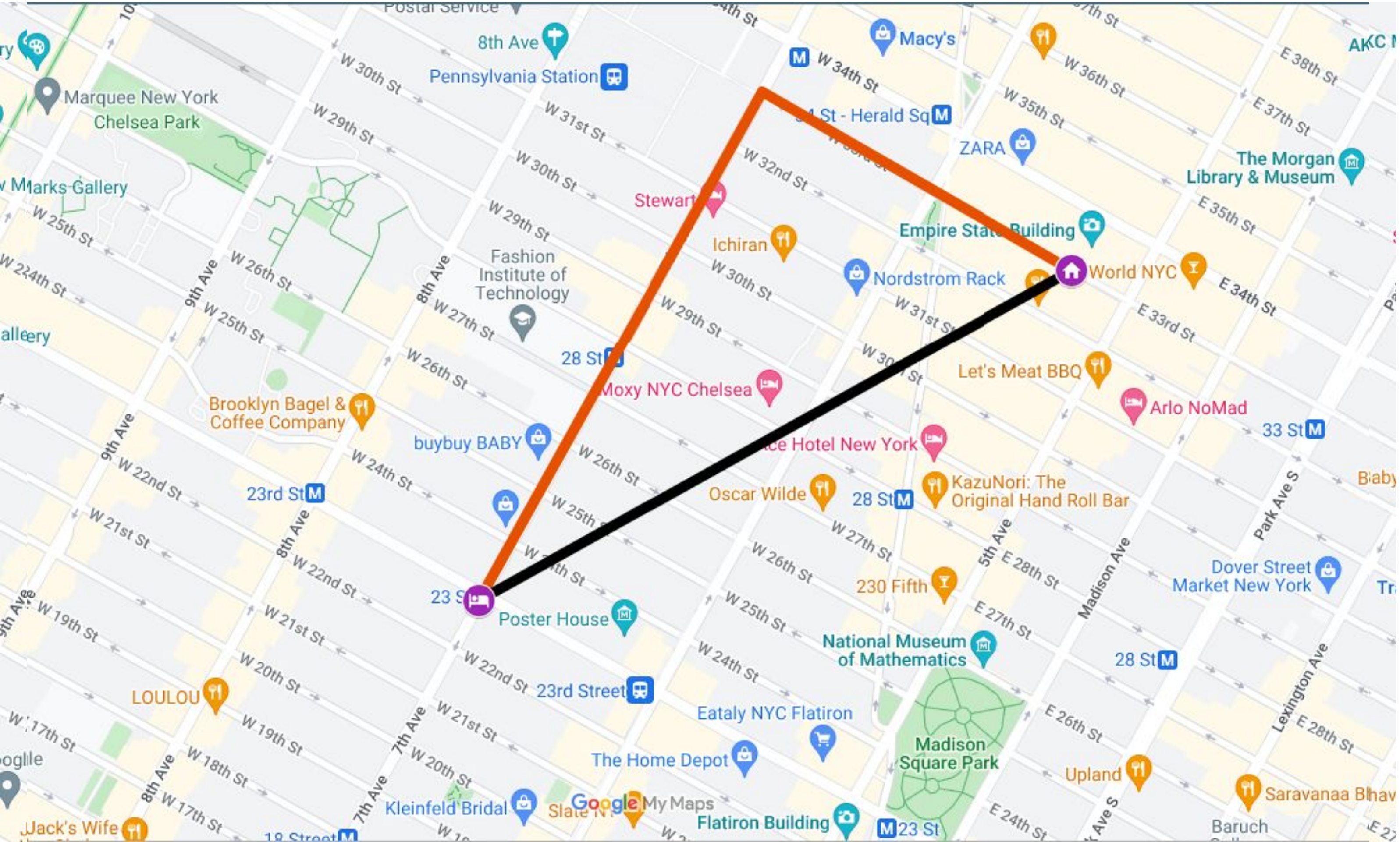


Manhattan, or city-block distance!
Reinvented by Burrows!
(with important adjustment)
Delta distance = Manhattan with scaled featur

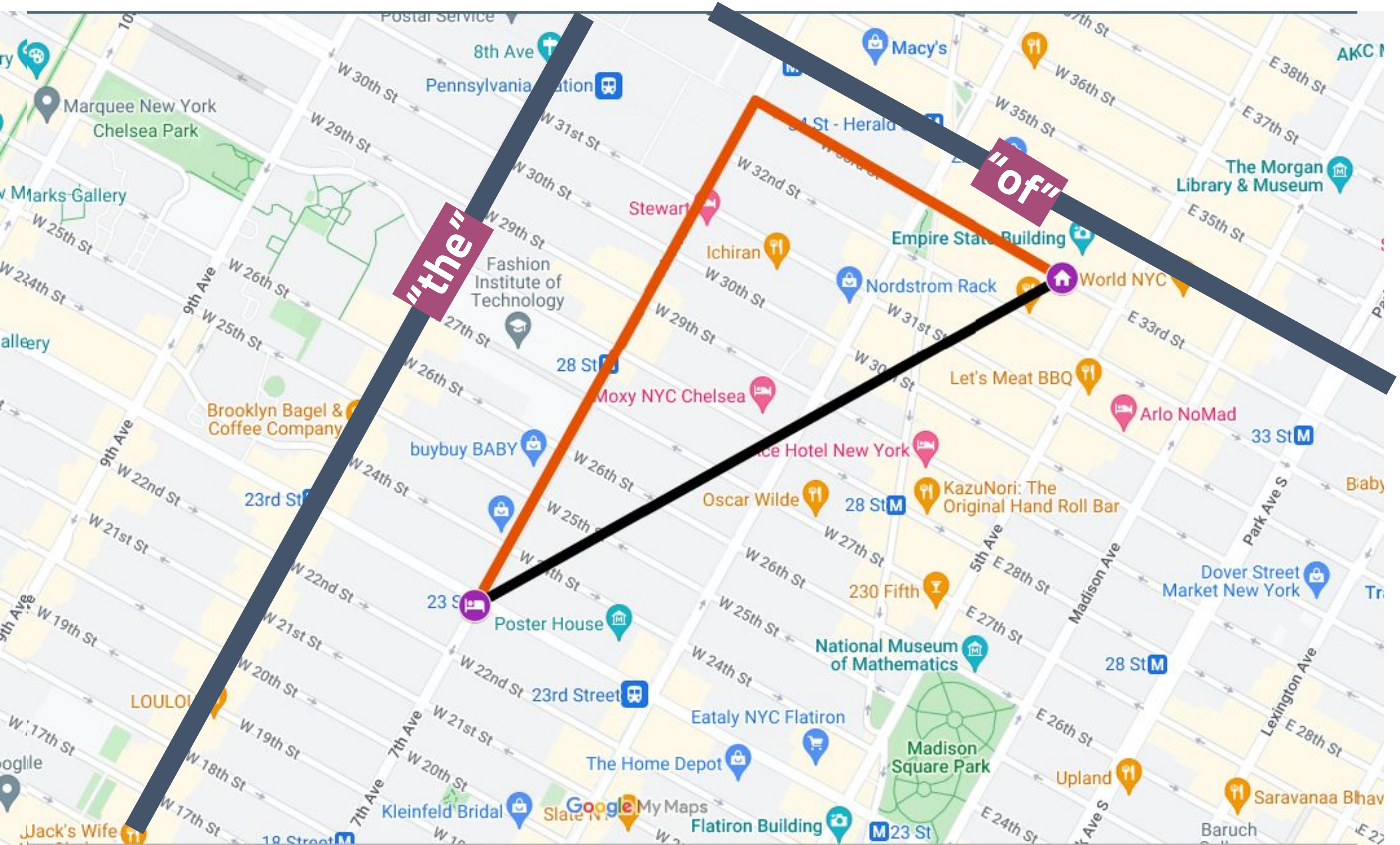
Petr Plecháč: https://versologie.cz/talks/2017chicago/



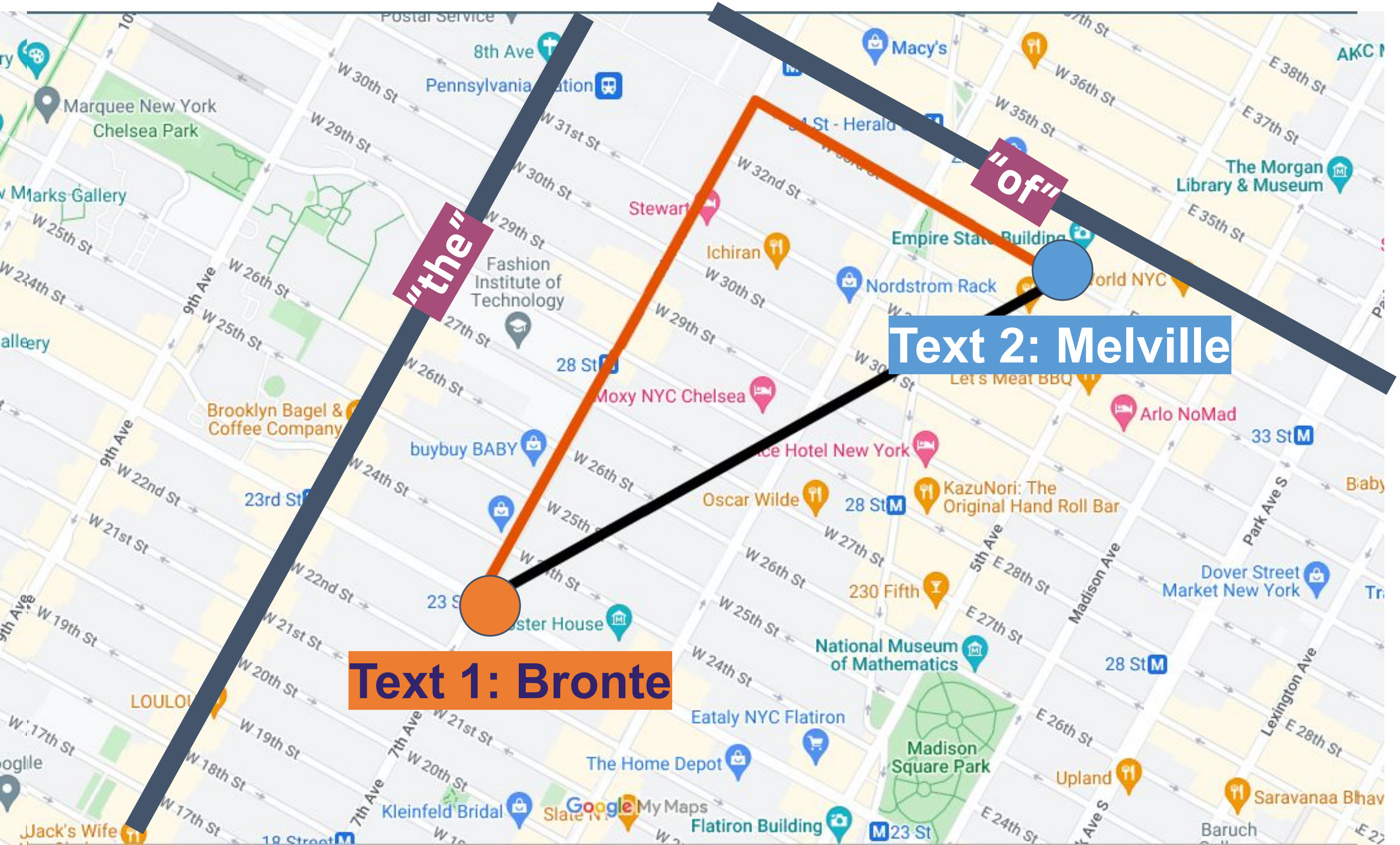
Petr Plecháč: https://versologie.cz/talks/2017chicago/



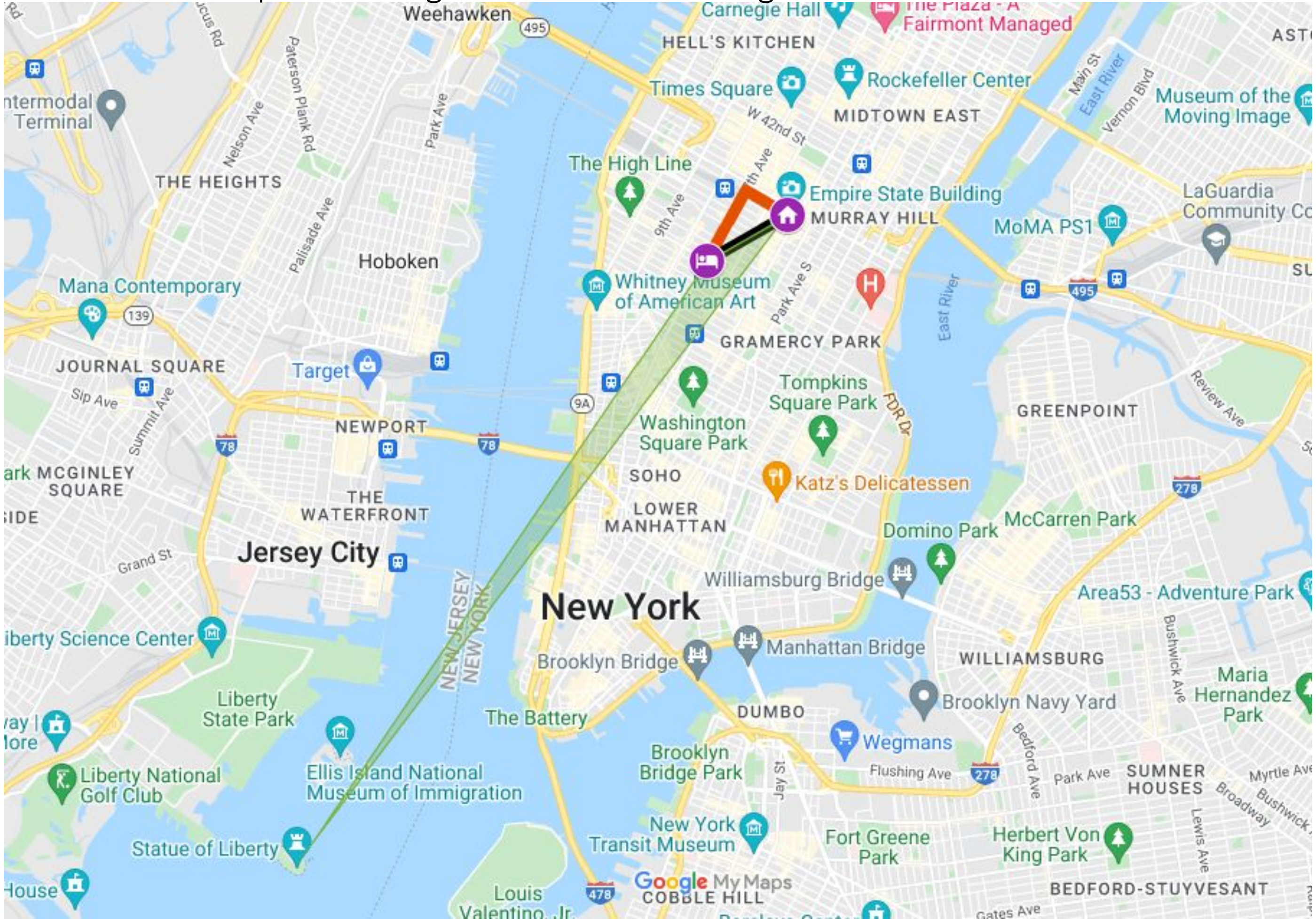
Petr Plecháč: https://versologie.cz/talks/2017chicago/



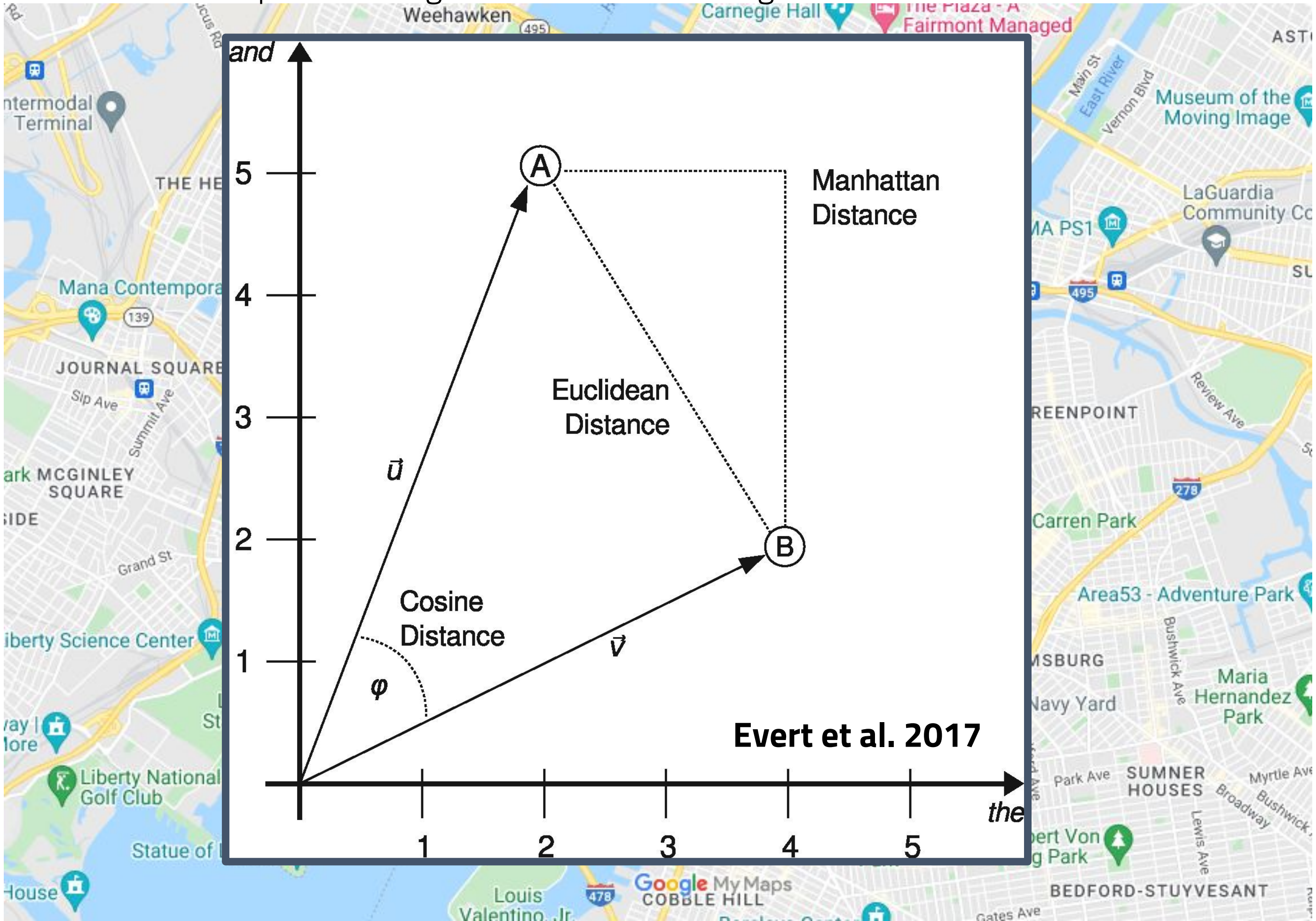
Petr Plecháč: https://versologie.cz/talks/2017chicago/



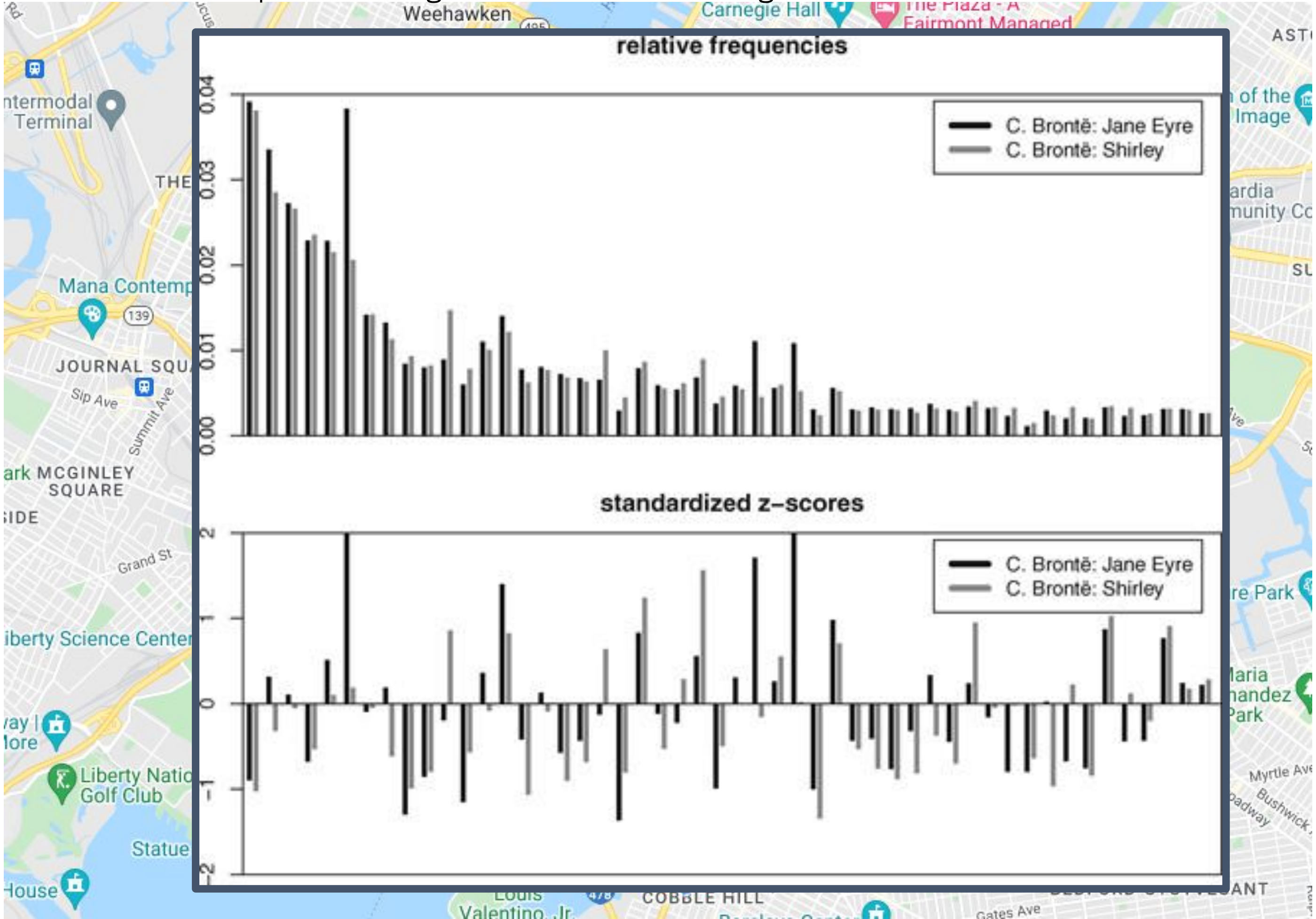
Petr Plecháč: https://versologie.cz/talks/2017chicago/

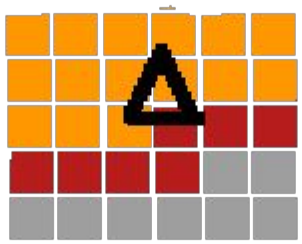


Petr Plecháč: https://versologie.cz/talks/2017chicago/



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





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

$$\Delta(T1,T2) = 7 + 15 + 8 = 30$$

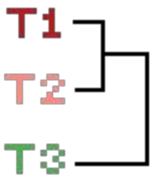
DISTANCE MATRIX

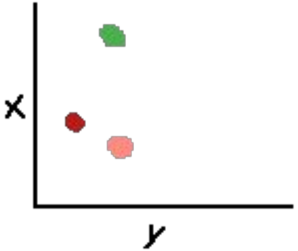
	T1	T2	T3
T1	0		
T2	0.3	0	
T3	0.7	0.9	0

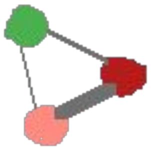
DISTANCE MATRIX

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

MULTIDIMENSIONAL SCALING







HIERARCHICAL CLUSTERING

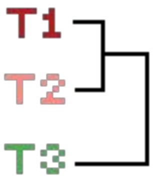
GRAPH

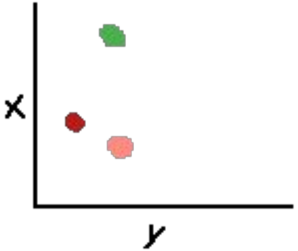
DISTANCE MATRIX

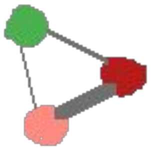
	T1	T2	T3
T1	0		
T2	0.3	0	
T3	[illegible]	0.9	0

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

MULTIDIMENSIONAL SCALING







HIERARCHICAL CLUSTERING

GRAPH

Bonus: trees? DIY!

	T_1	T_2	T_3
$ЯЧ^1$	0,1	0,2	0,7
$ЯЧ^2$	0,15	0,2	0,65
$XЧ^1$	0,3	0,5	0,2
$XЧ^2$	0,4	0,4	0,2

Bonus: trees? DIY!

	ЯЧ^{1}	ЯЧ^{2}	XЧ^{1}	XЧ^{2}
ЯЧ^{1}	0			
ЯЧ^{2}	0.1	0		
XЧ^{1}	0.8	0.9	0	
XЧ^{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…

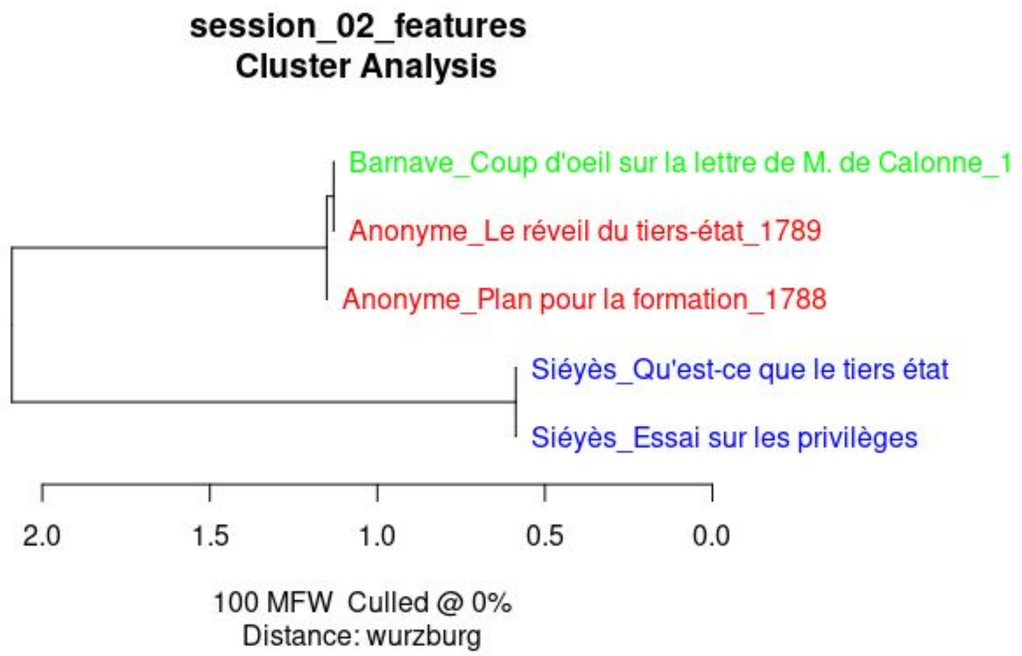
TL;DR Multivariate text analysis

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…

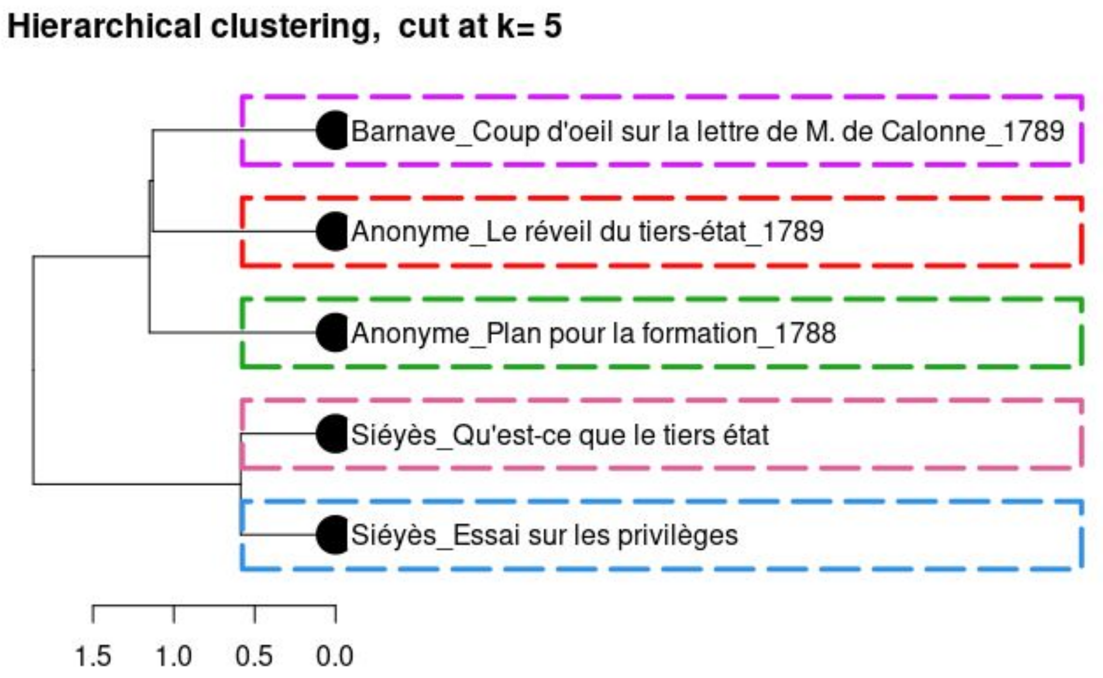
`stylo` package does that in R!

Clustering and features

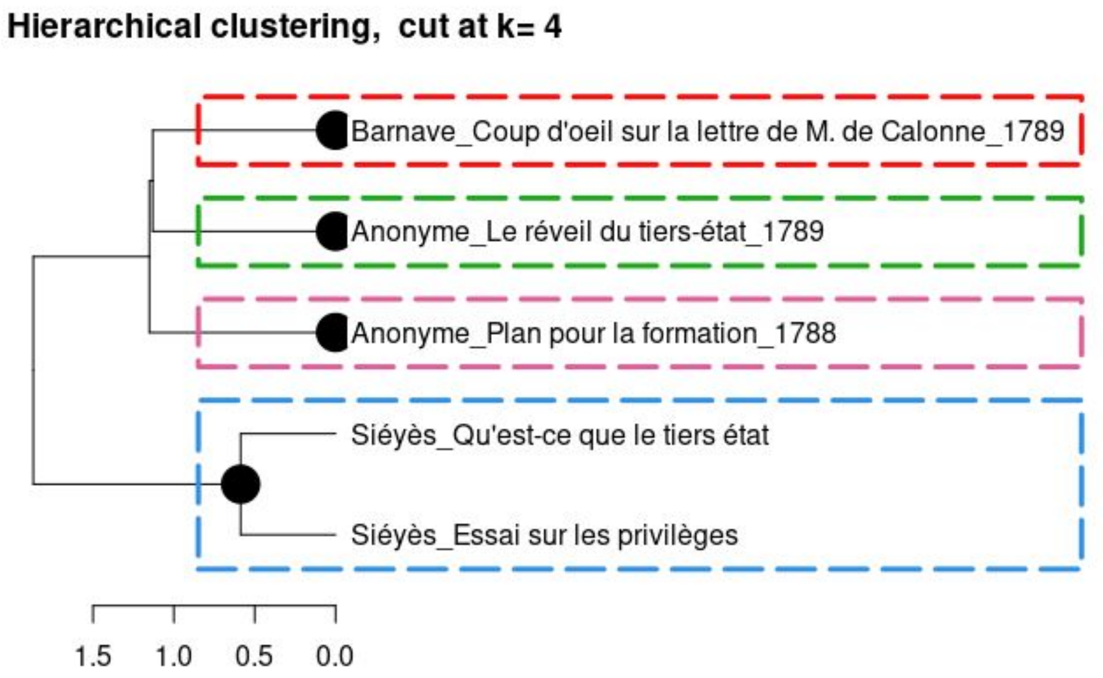
Cluster analysis: grouping suggestion



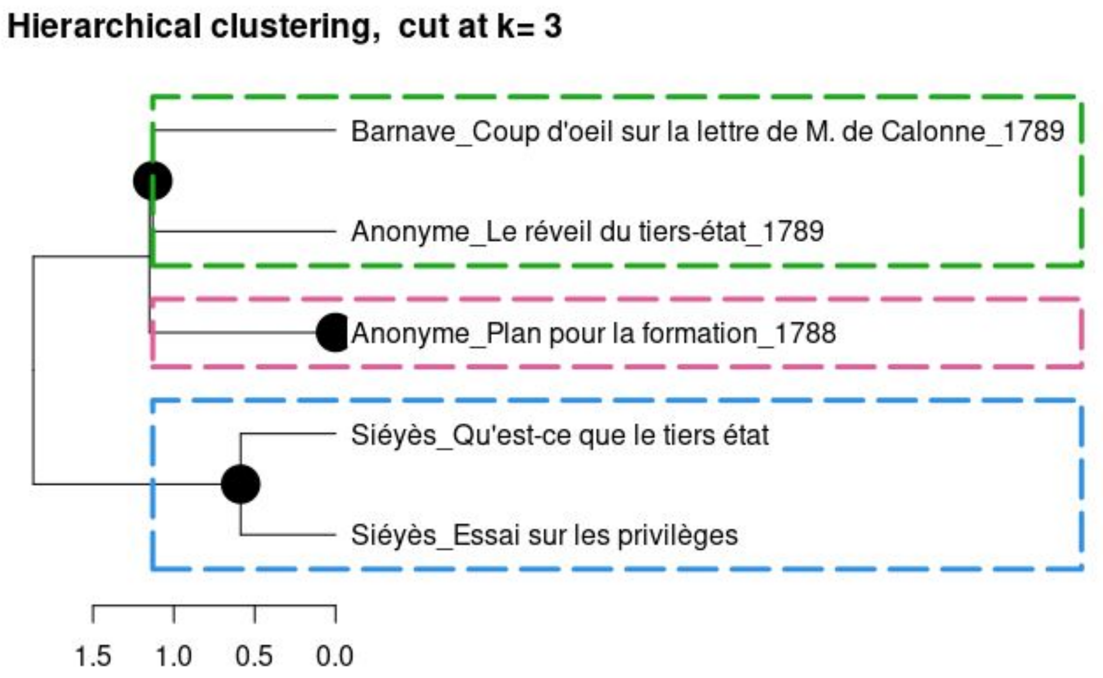
Cutting trees by groups!



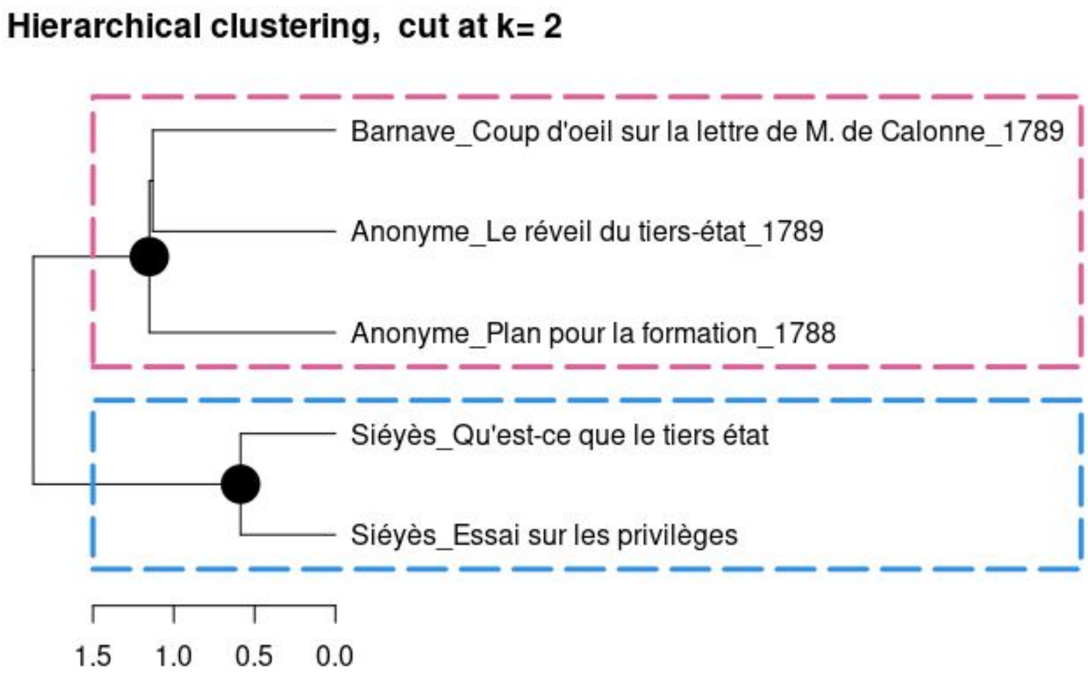
Cutting trees by groups!



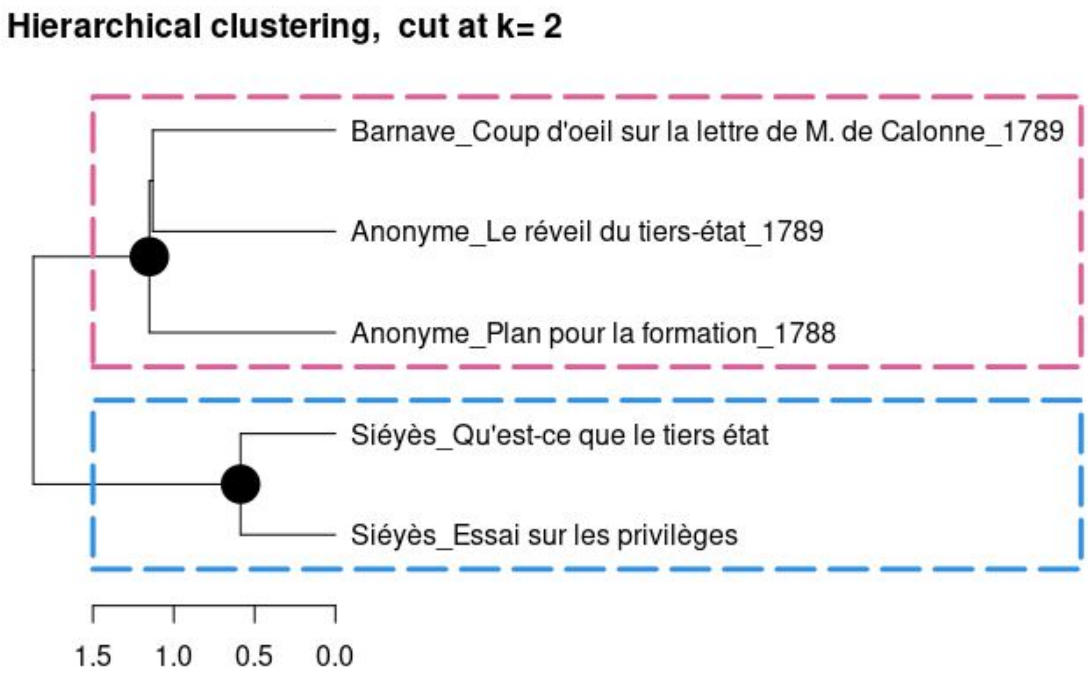
Cutting trees by groups!



Cutting trees by groups!

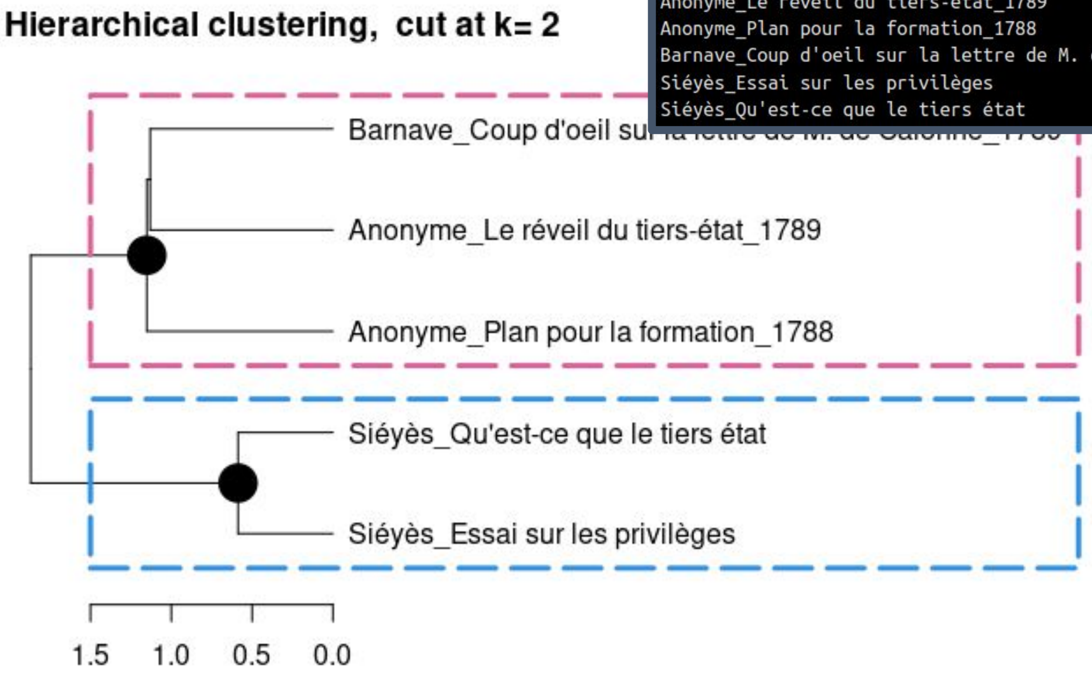


Think of new ‘pink’ and ‘blue’ classes



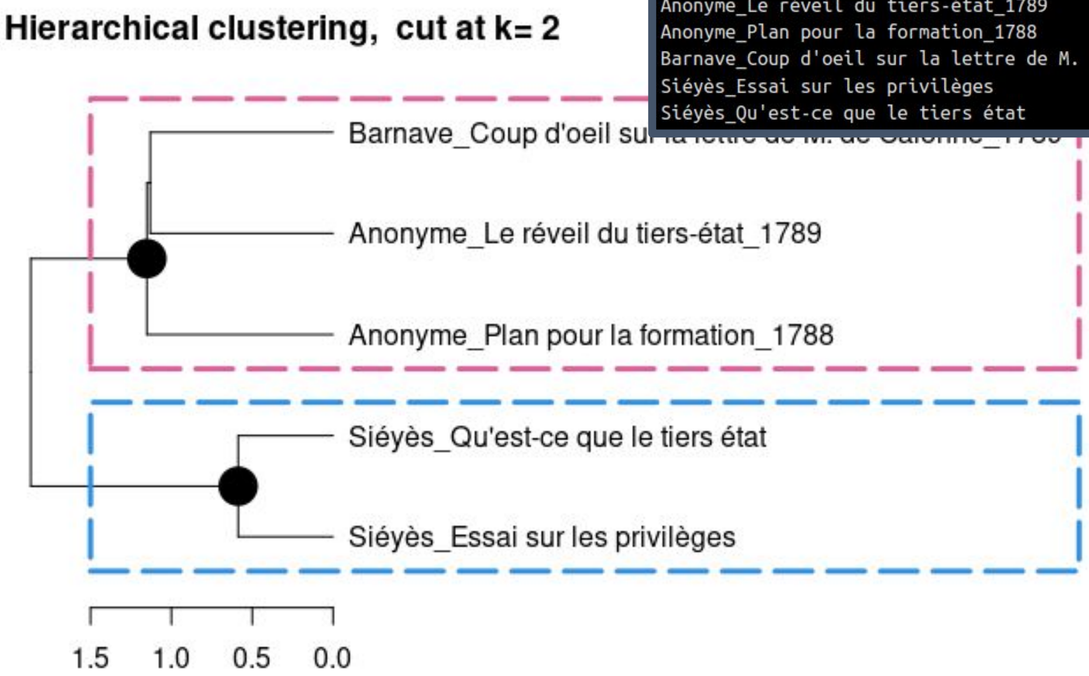
Classes are driven by word frequencies!

	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 privilèges	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

	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 privilèges	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

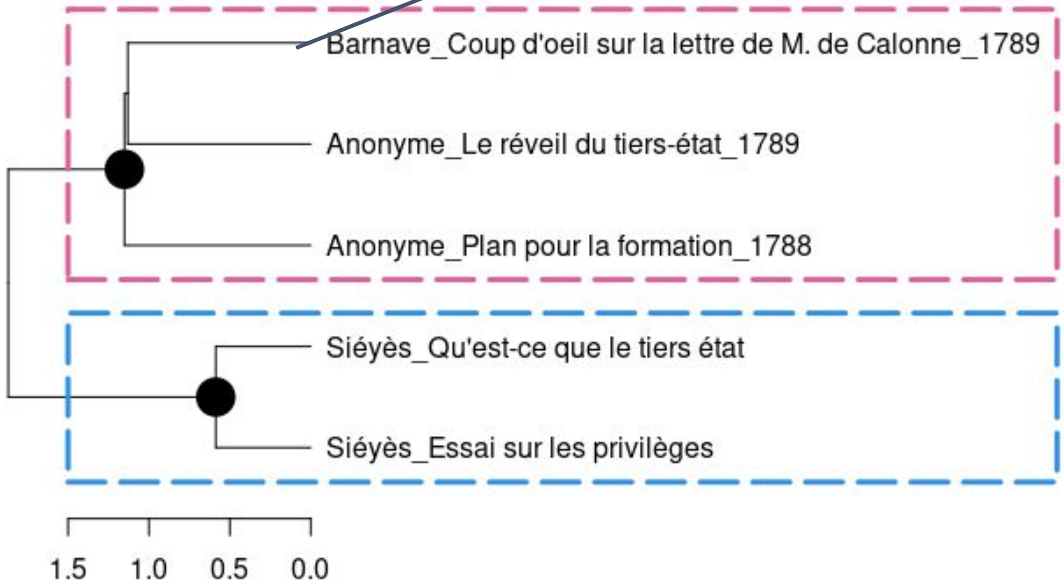


You an use 'emergent' class information from the tree to define corpora and check **which features differ** across clusters.

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 privilèges	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



Cluster 2	Cluster 1
est	fe
sont	
sans	
sa	
sur	
et	
ses	

pute, lequel fe 1 Anemblée de TéMion ou ârrondiiïement. 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 ï'Éféétion 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 ârrondiiïement. 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 ï'Éféétion 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 … élion de,
Monte- e … Lé
IMrocès-

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

`seetrees` package (very unfinished)

Installation

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

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

Example

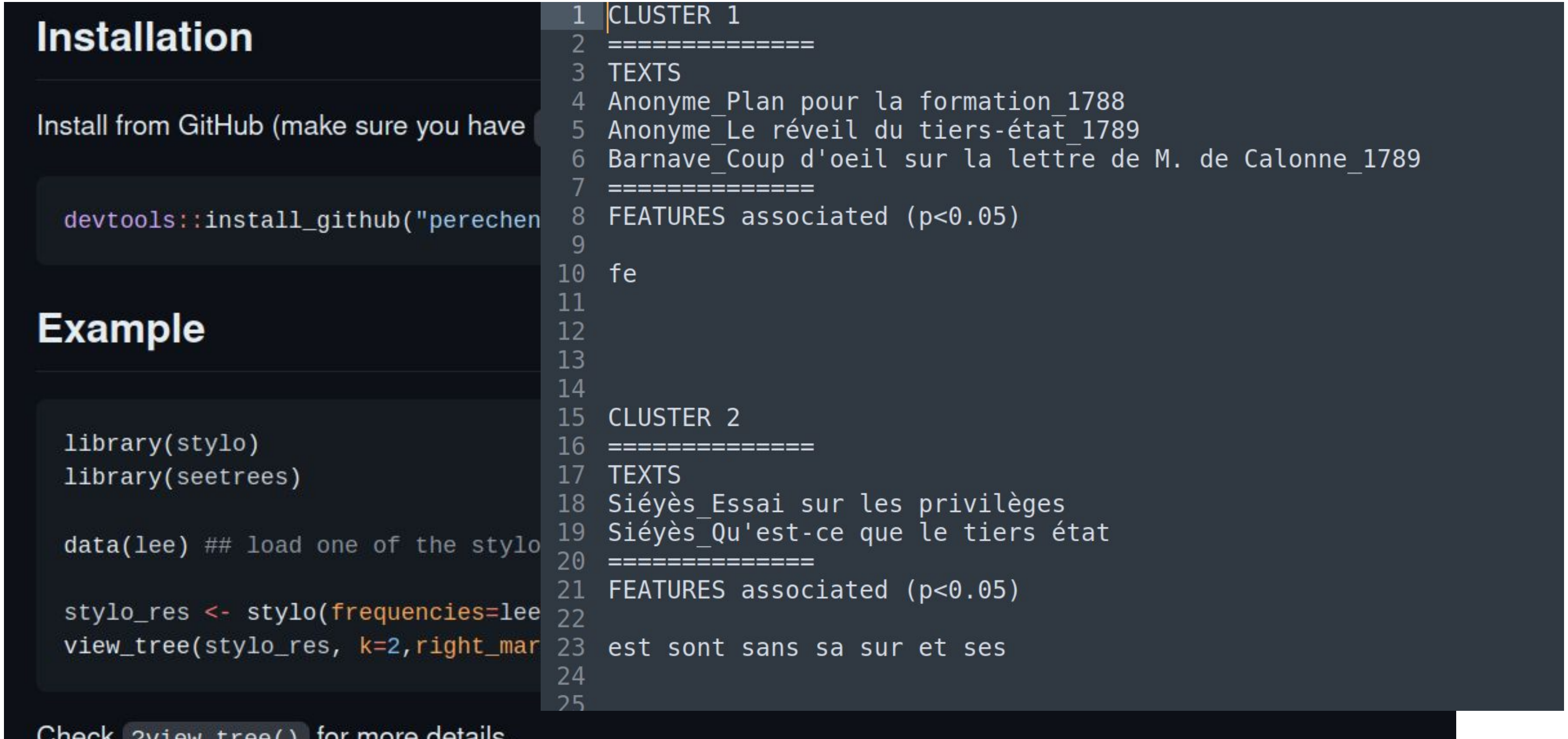
```
library(stylo)
library(seetrees)

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

stylo_res <- stylo(frequencies=lee
view_tree(stylo_res, k=2,right_mar
```

Check `?view_tree()` for more details

```
CLUSTER 1
===============
TEXTS
Anonyme_Plan pour la formation_1788
Anonyme_Le réveil du tiers-état_1789
Barnave_Coup d'oeil sur la lettre de M. de Calonne_1789
===============
FEATURES associated (p<0.05)

fe




CLUSTER 2
===============
TEXTS
Siéyès_Essai sur les privilèges
Siéyès_Qu'est-ce que le tiers état
===============
FEATURES associated (p<0.05)

est sont sans sa sur et ses
```

Sampling and uncertainty

Sidenote

Sampling without replacement:

Sidenote

Sampling without replacement:

Sidenote

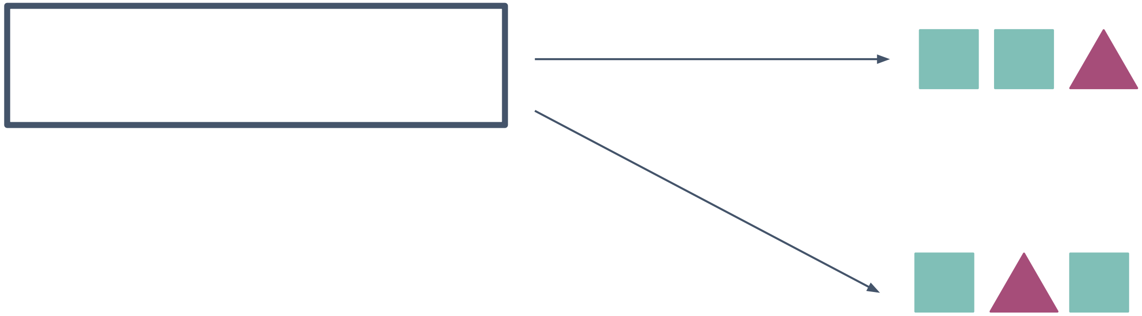
Sampling without replacement:

Sidenote

Sampling without replacement:

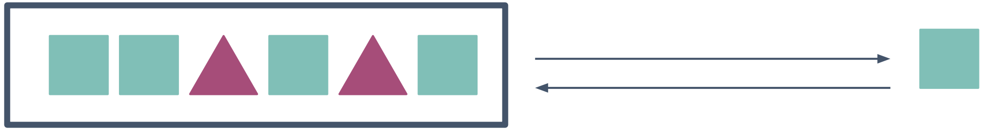
Sidenote

Sampling without replacement:



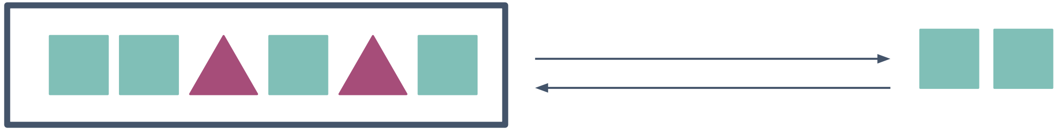
Sidenote

Sampling *with* replacement:



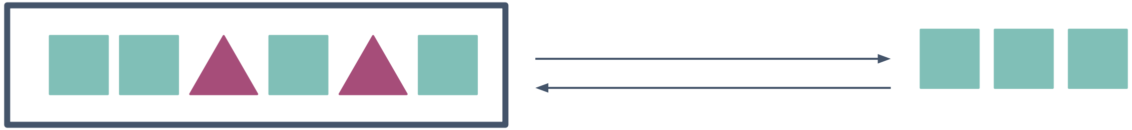
Sidenote

Sampling *with* replacement:



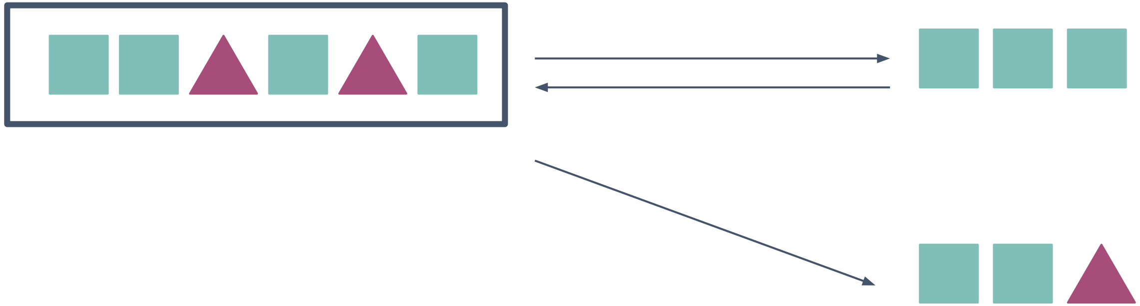
Sidenote

Sampling *with* replacement:



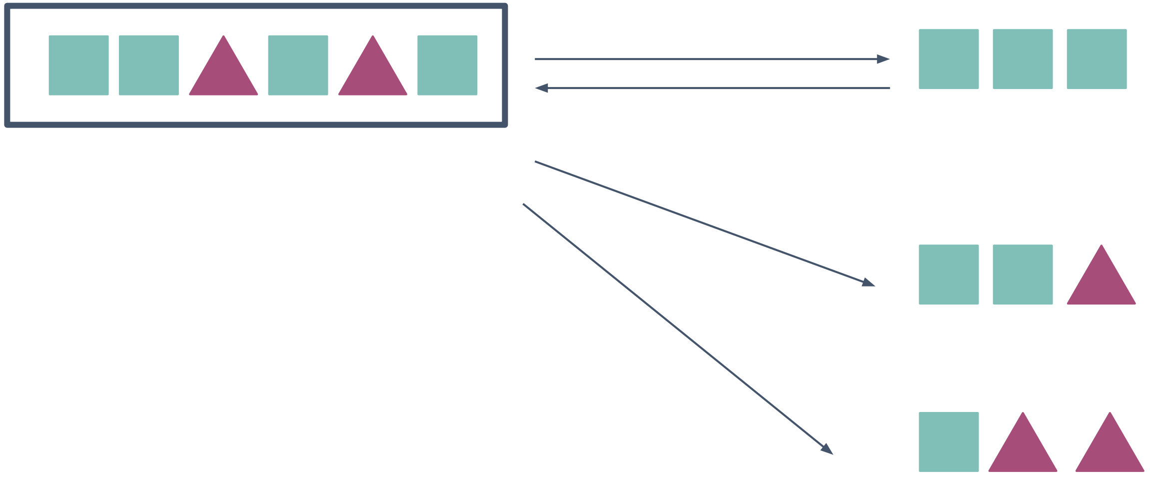
Sidenote

Sampling *with* replacement:



Sidenote

Sampling *with* replacement:



3. Estimating 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

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

2. Sampling & bootstrapping

Sample:  p(square) = 0.66

Resample 1:  0.5

2. Sampling & bootstrapping

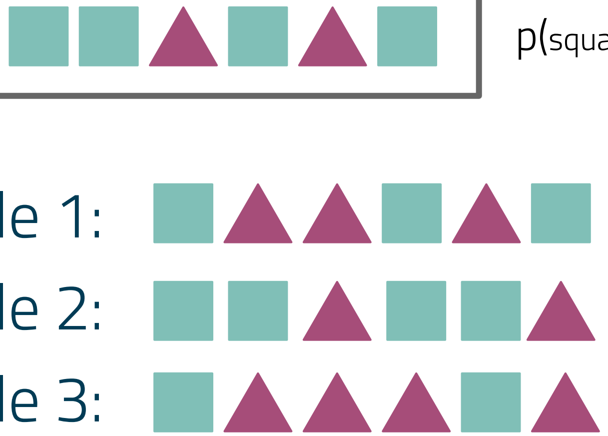
Sample: 

p(square) = 0.66

Resample 1:  0.5

Resample 2:  0.66

2. Sampling & bootstrapping



Sample:

p(square) = 0.66

Resample 1: 0.5

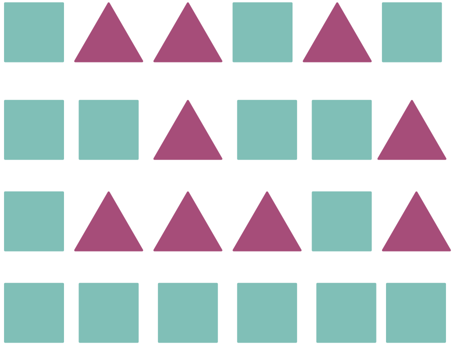
Resample 2: 0.66

Resample 3: 0.33

2. Sampling & bootstrapping

Sample: 

p(square) = 0.66



Resample 1: 0.5

Resample 2: 0.66

Resample 3: 0.33

Resample 4: 1

2. Sampling & bootstrapping

Sample:

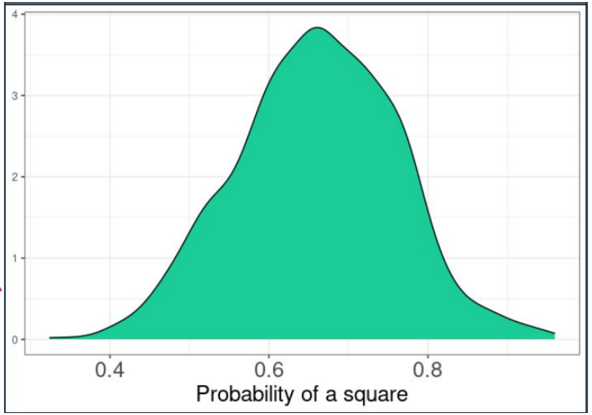
p(square) = 0.66

Resample 1:

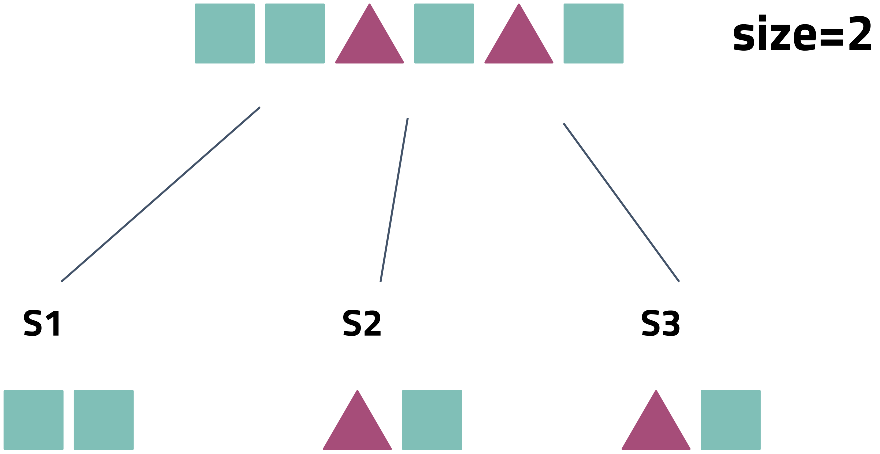
Resample 2:

Resample 3:

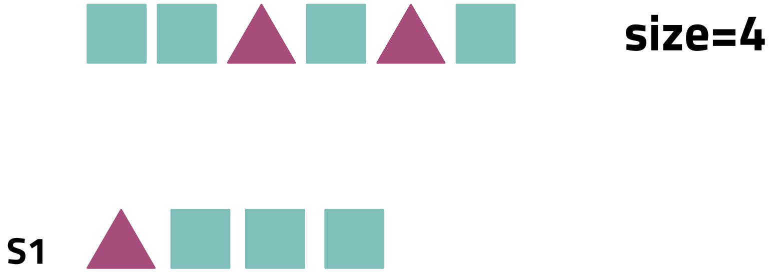
Resample 4:



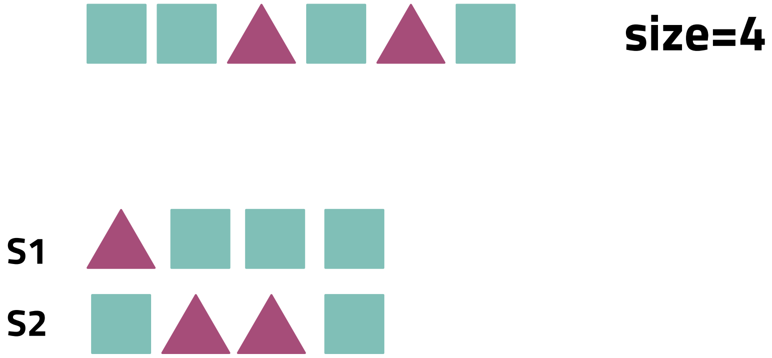
Normal vs. random sampling (in stylo)



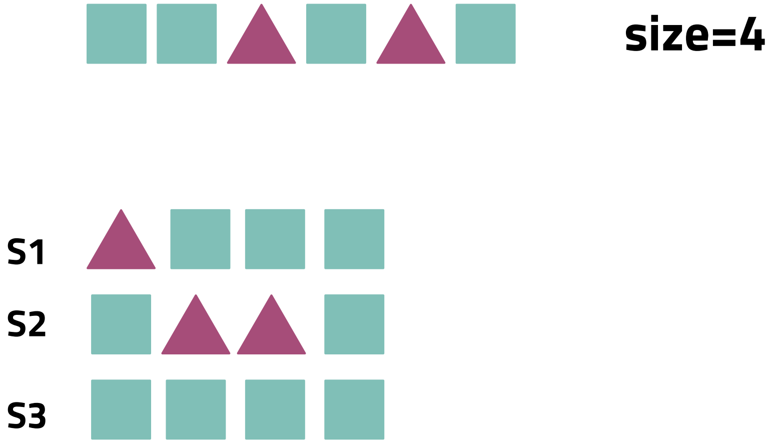
Normal vs. random sampling (in stylo)



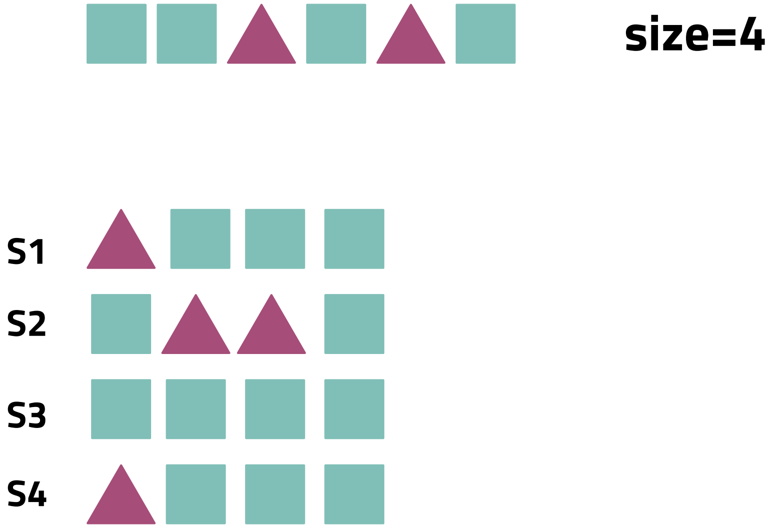
Normal vs. random sampling (in stylo)



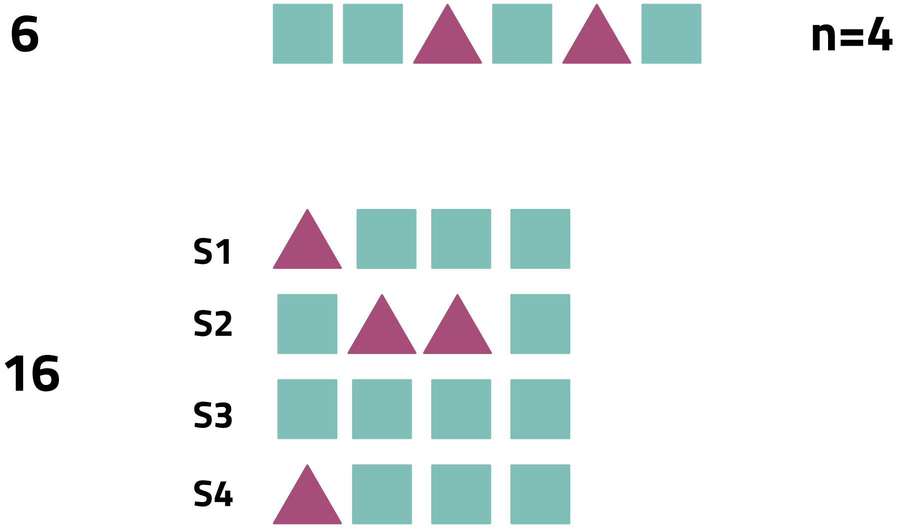
Normal vs. <u>random</u> sampling (in stylo)



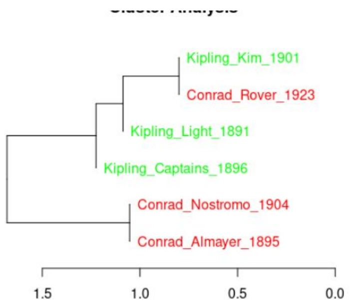
Normal vs. random sampling (in stylo)



Normal vs. random sampling (in stylo)

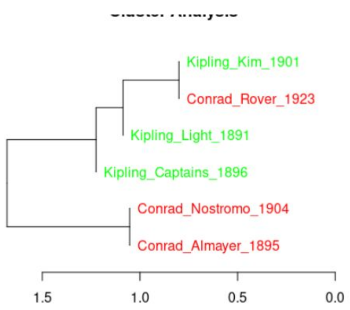


4. Consensus trees

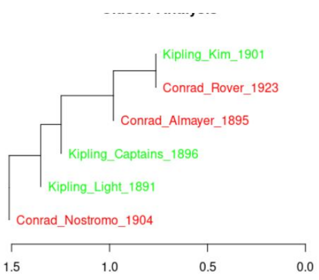


Feature set 1

4. Consensus trees

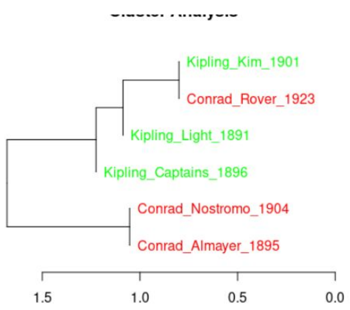


Feature set 1

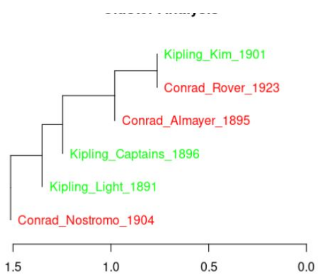


Feature set 2

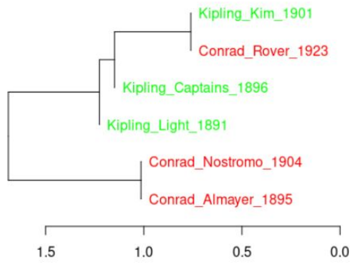
4. Consensus trees



Feature set 1

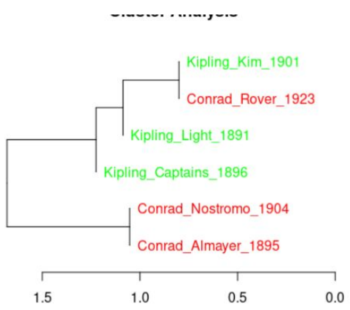


Feature set 2

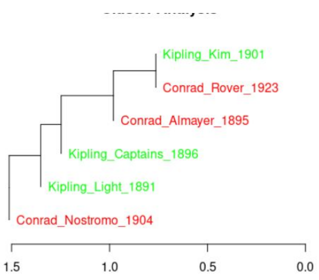


Feature set 3

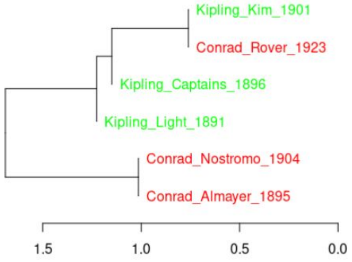
4. Consensus trees



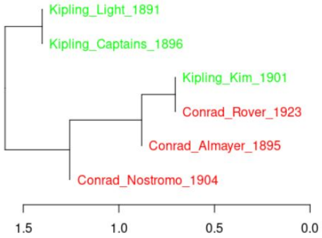
Feature set 1



Feature set 2

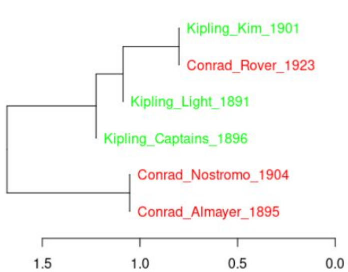


Feature set 3

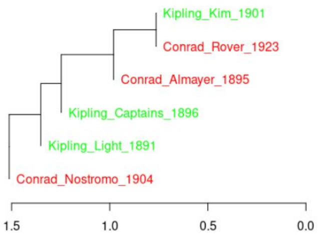


Feature set 4

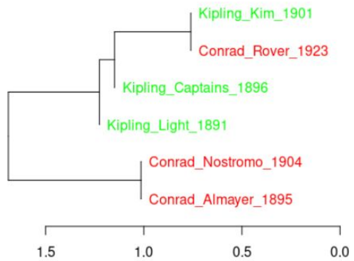
4. Majority rule (>50% of branches)



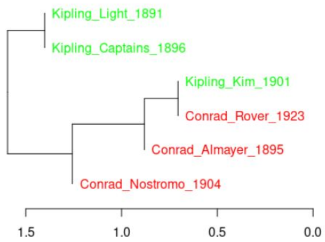
Feature set 1



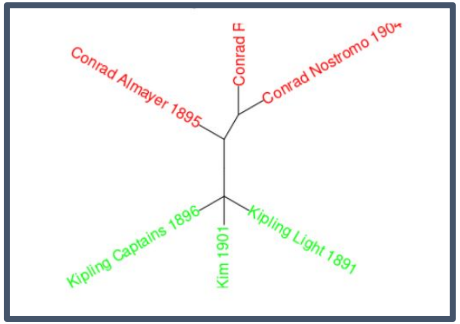
Feature set 2



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”:

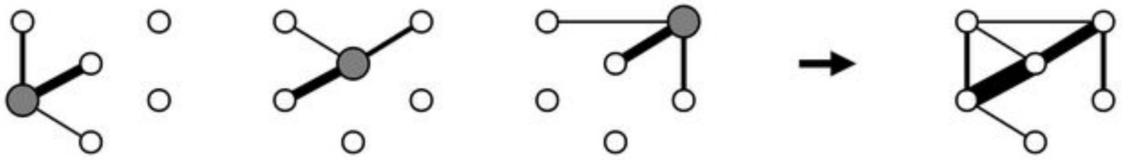
- MFW length
- Culling strength
- Text themselves (take samples from texts)

….

But the possibilities are limitless

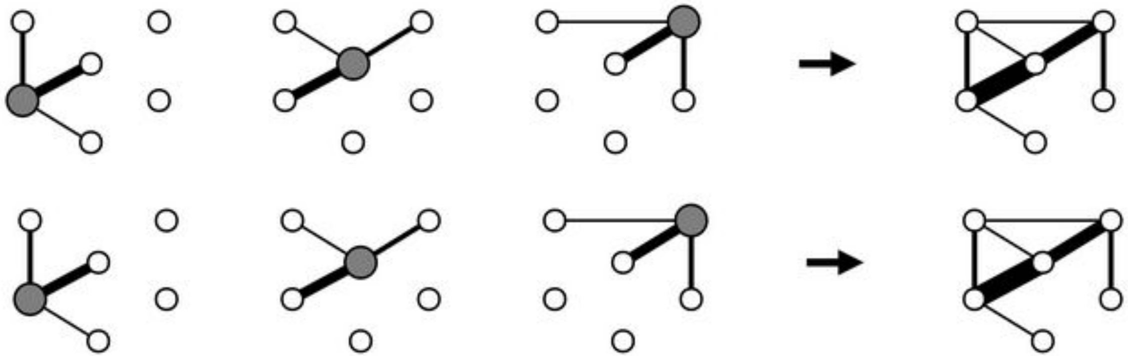
6. Consensus networks

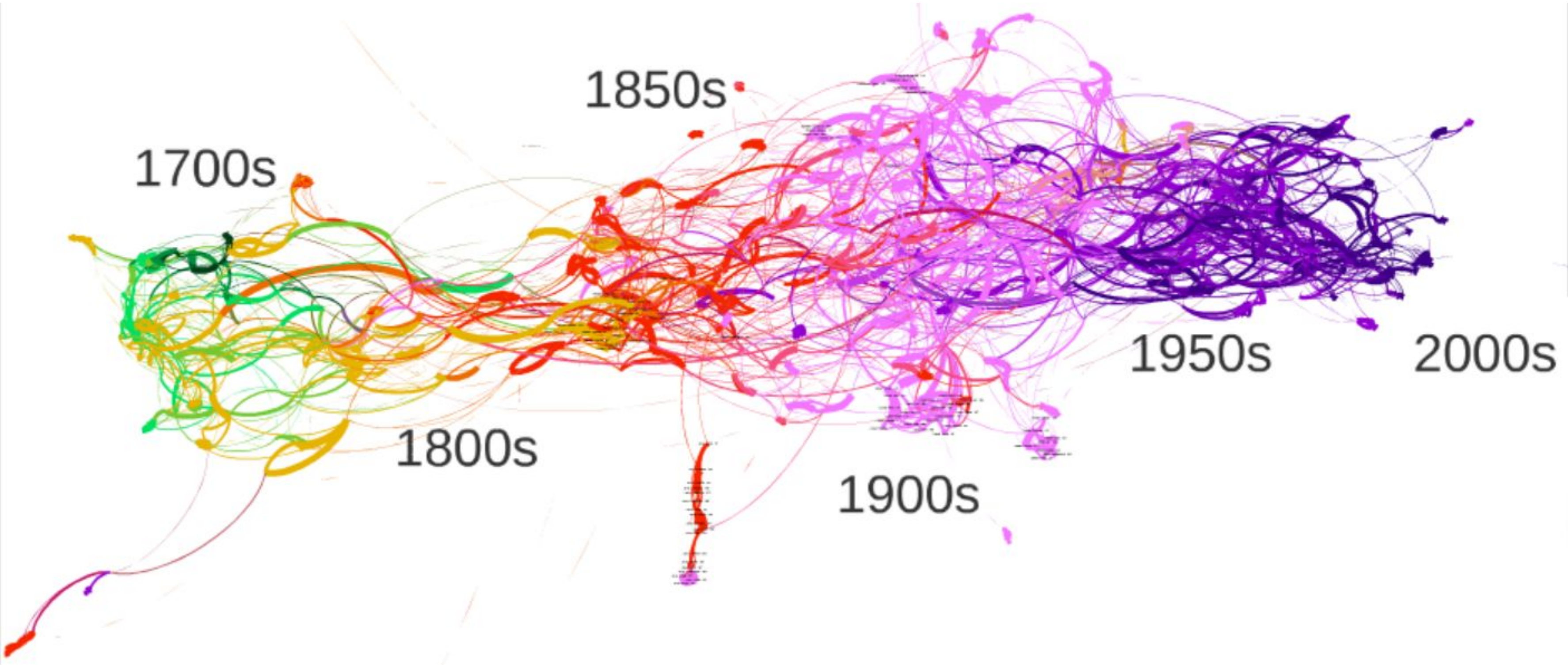
1. Look at the neighbours!



6. Consensus networks

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



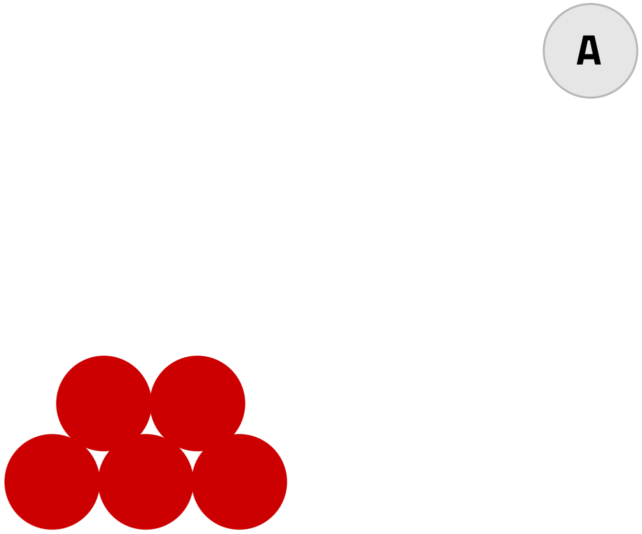

1700s
1850s
1800s
1900s
1950s
2000s

- 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**

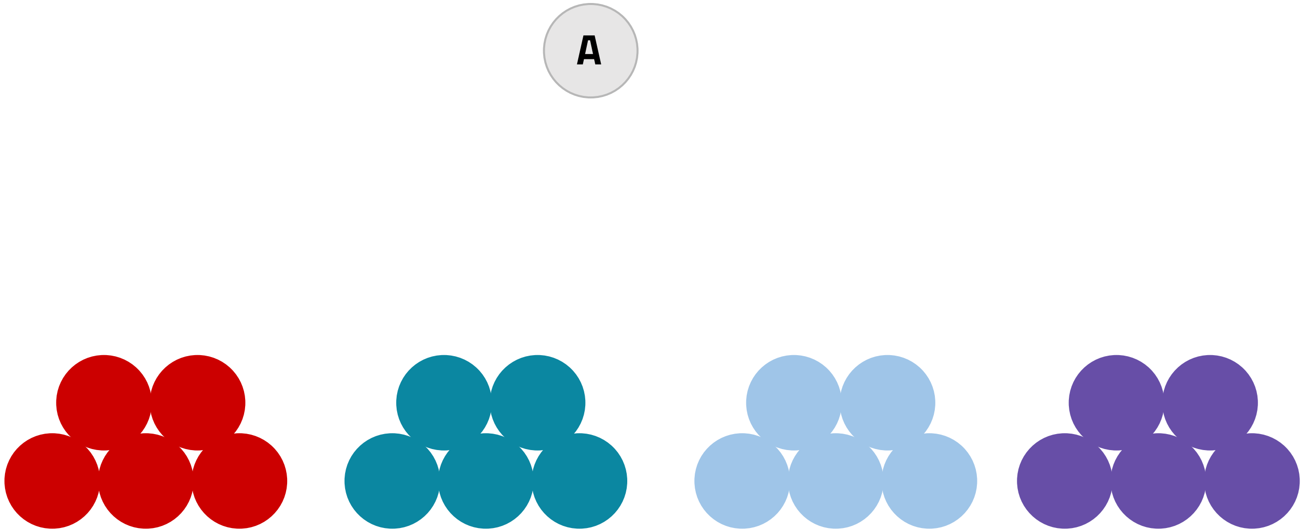
6. General imposters

A

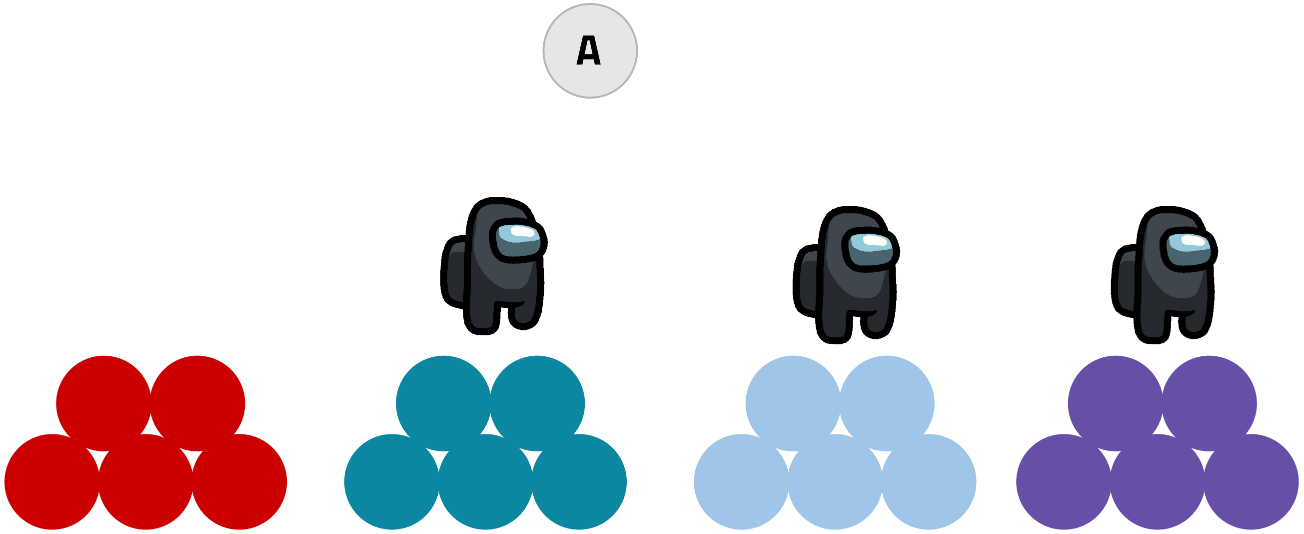
6. General imposters



6. General imposters



6. General imposters

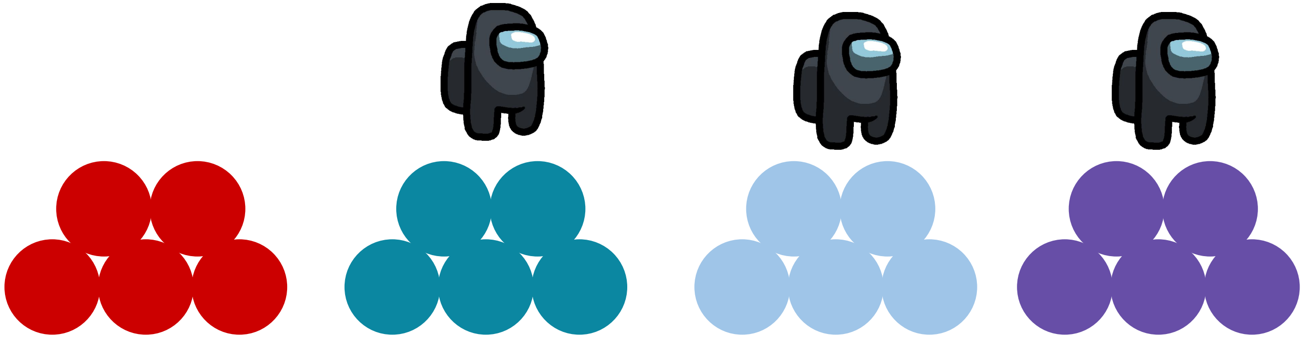


6. General imposters

Random samples

Random features

Random imposters

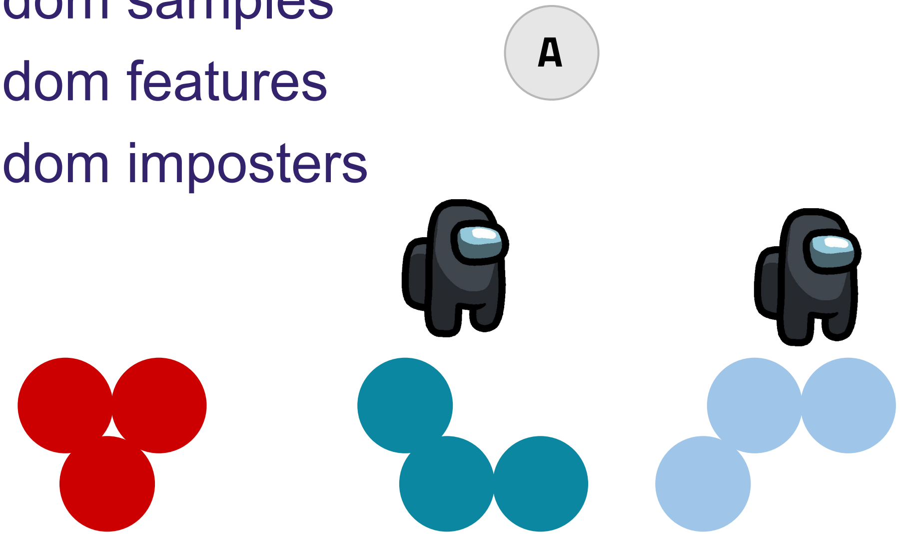


6. General imposters

Random samples

Random features

Random imposters

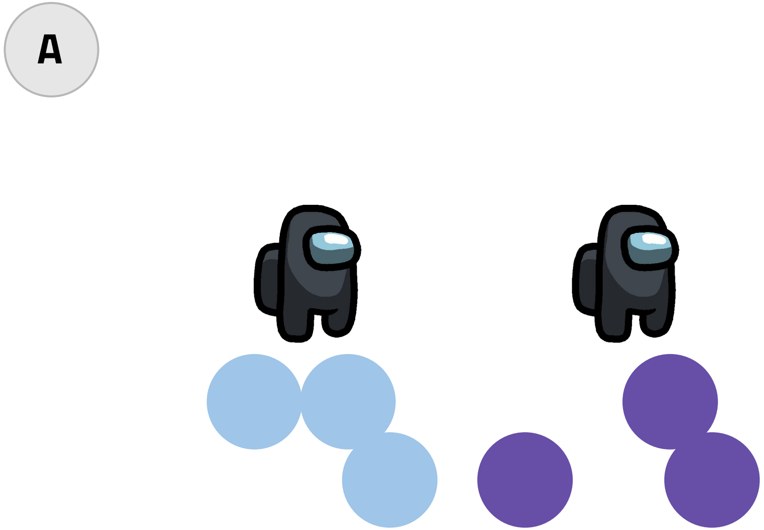


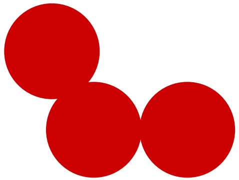
6. General imposters

Random samples

Random features

Random imposters





6. General imposters