Stylometry with R

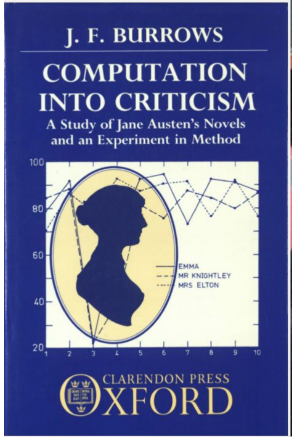
Part 1. Distances and uncertainty

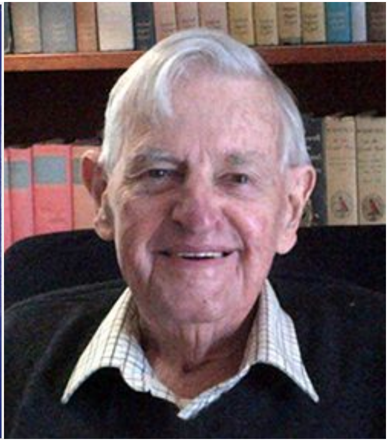
Joanna Byszuk and
Artjoms Šeļa

1. Quick recap of 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.**"

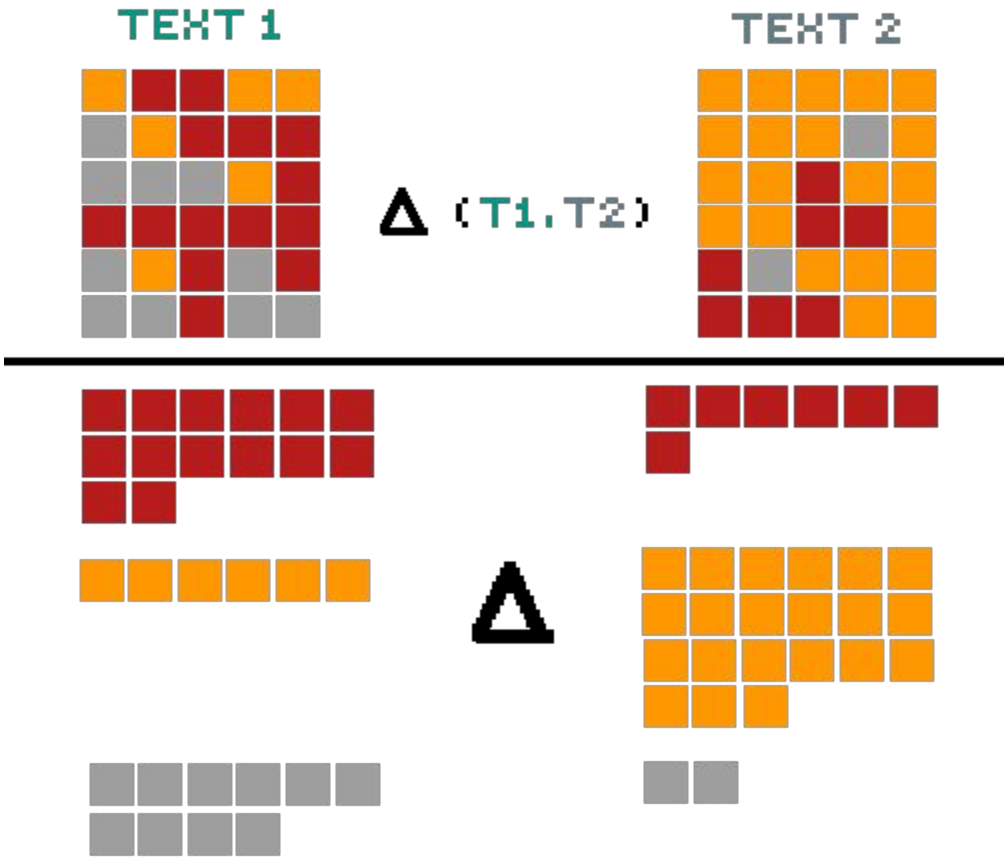
$$\Delta = \sum_{i=1}^{n} \frac{|z(x_i) - z(y_i)|}{n}$$

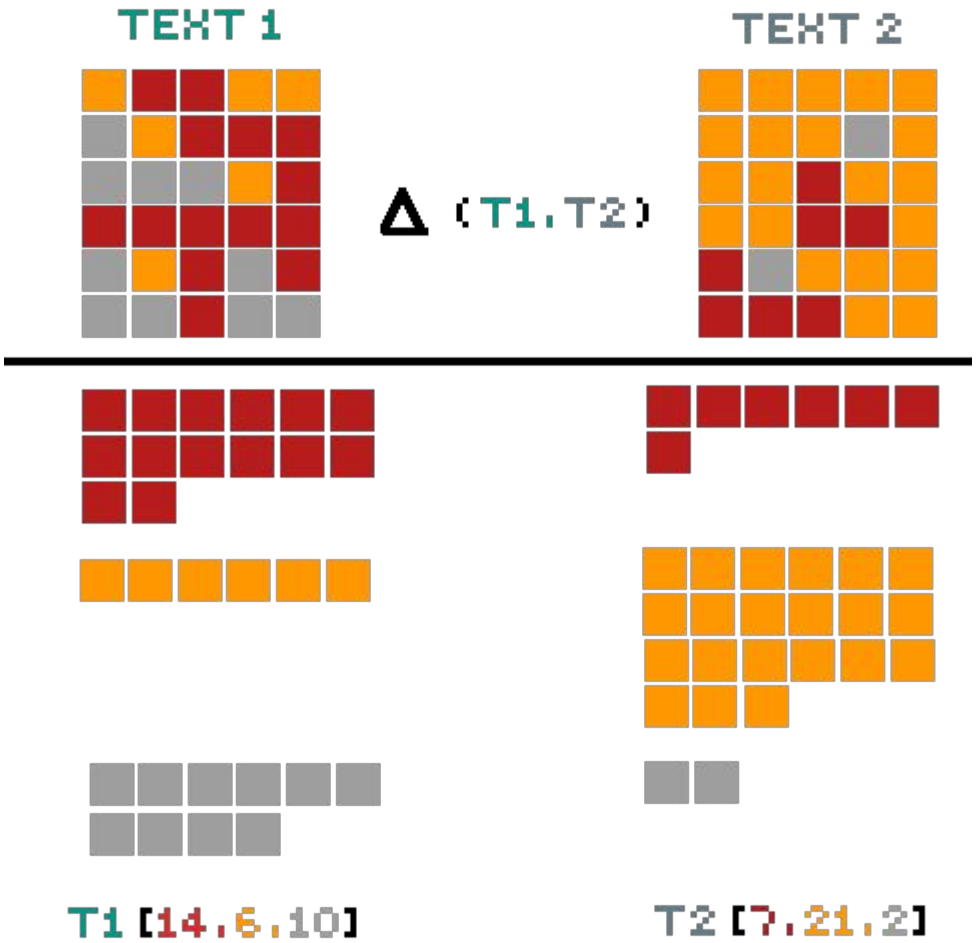




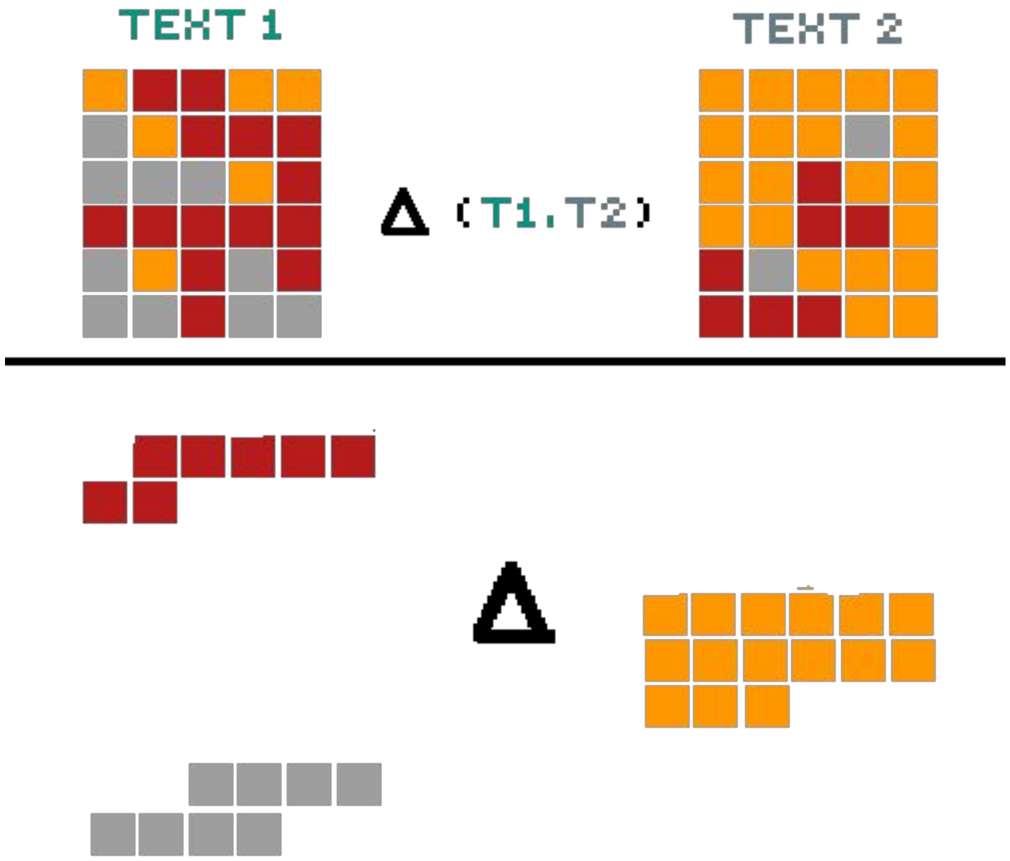
John Burrows (1928-2019)

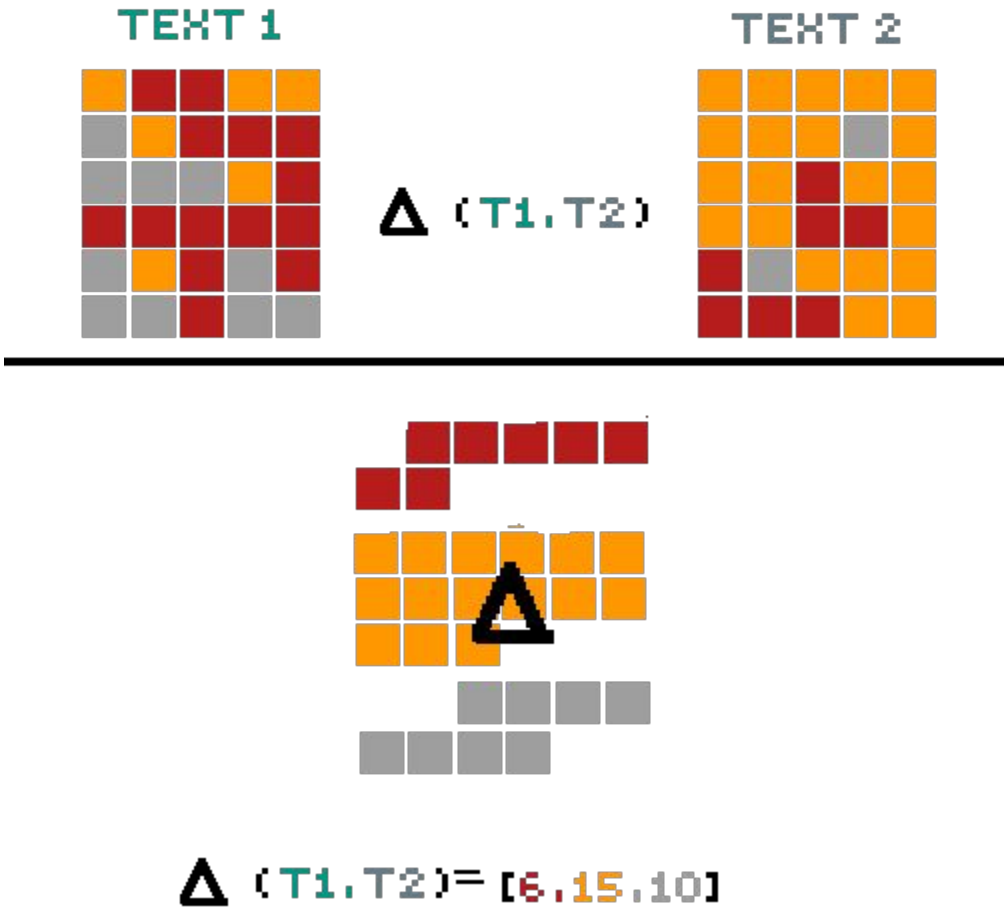

TEXT 1
Δ (T1,T2)
TEXT 2

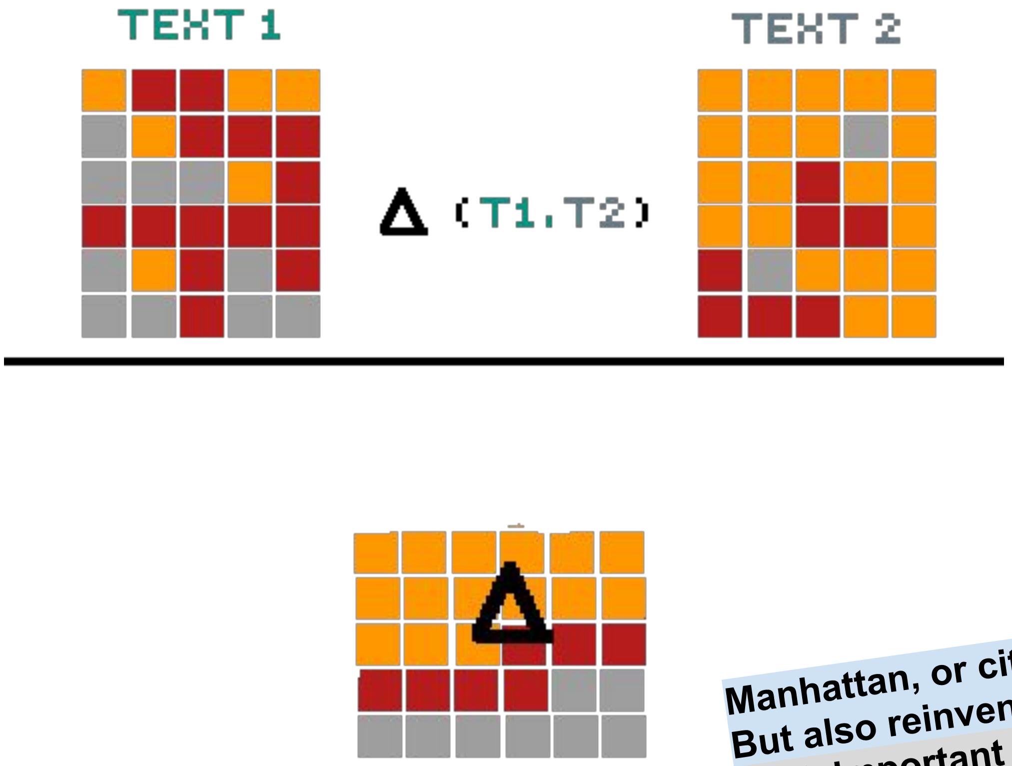

TEXT 1
TEXT 2
Δ (T1,T2)
Δ



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


TEXT 1
TEXT 2
Δ (T1,T2)
Δ

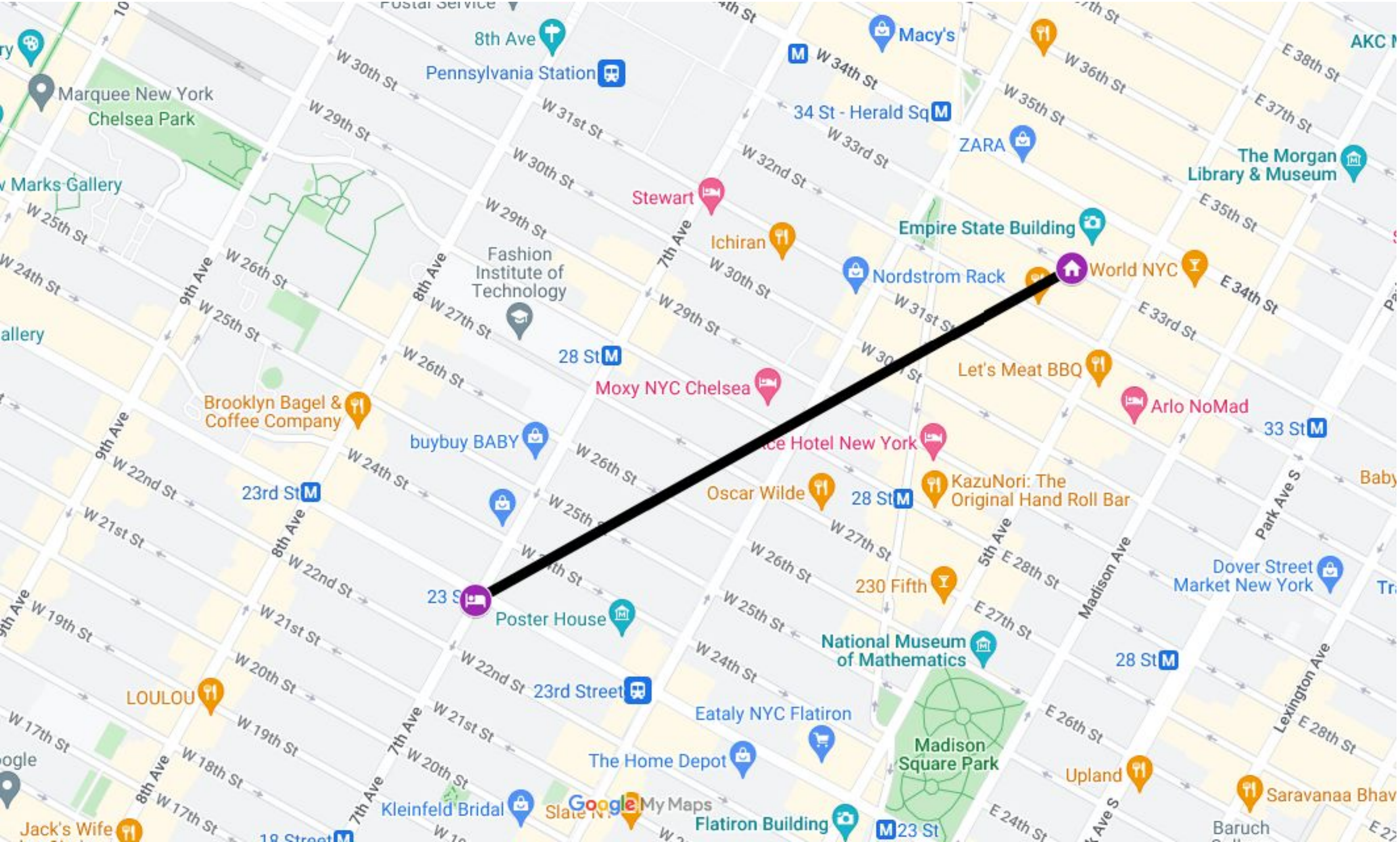

TEXT 1
TEXT 2
Δ (T1,T2)
Δ (T1,T2)= [6,15,10]



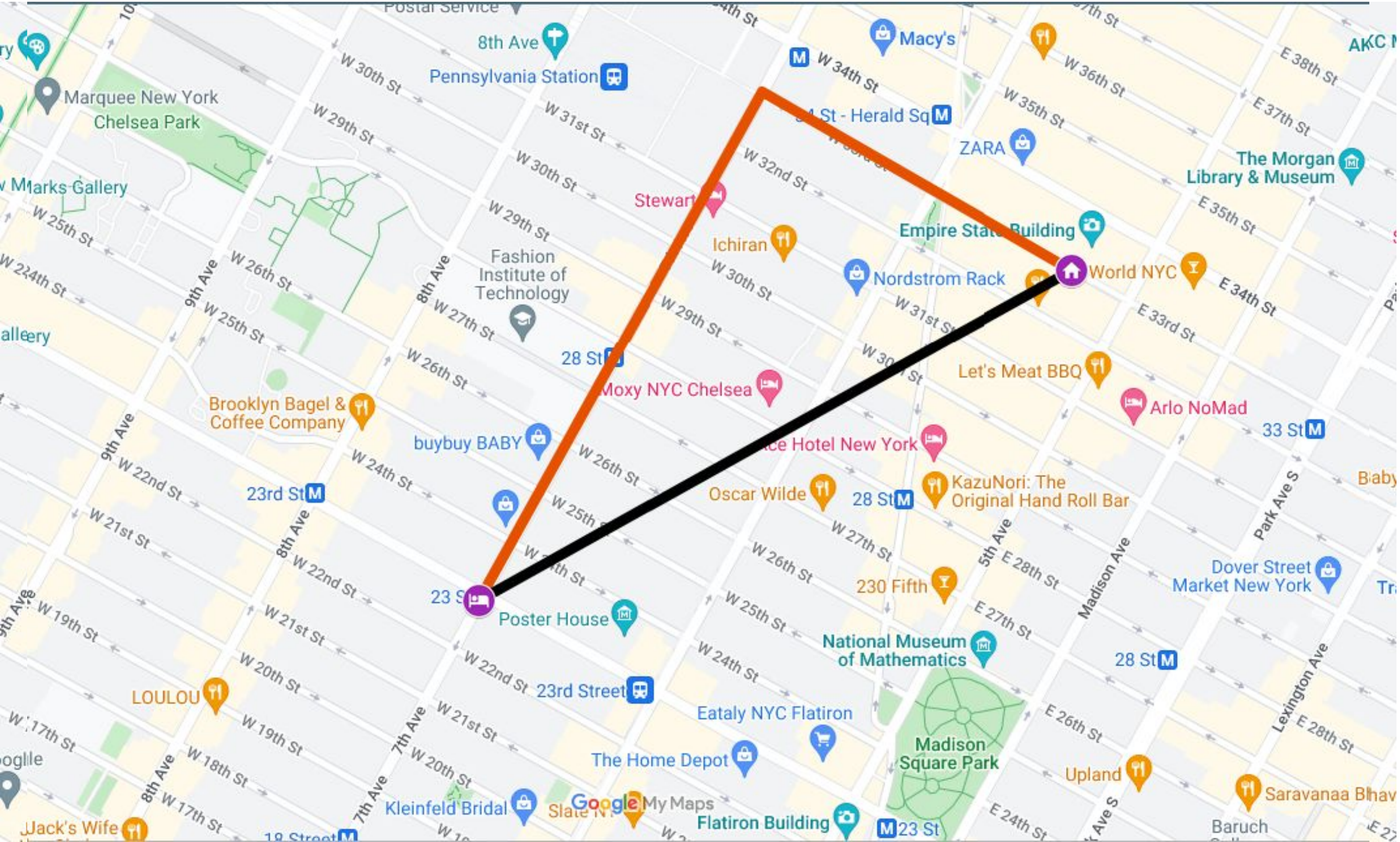
Manhattan, or city-block distance!
But also reinvented by Burrows
(with important adjustment)

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

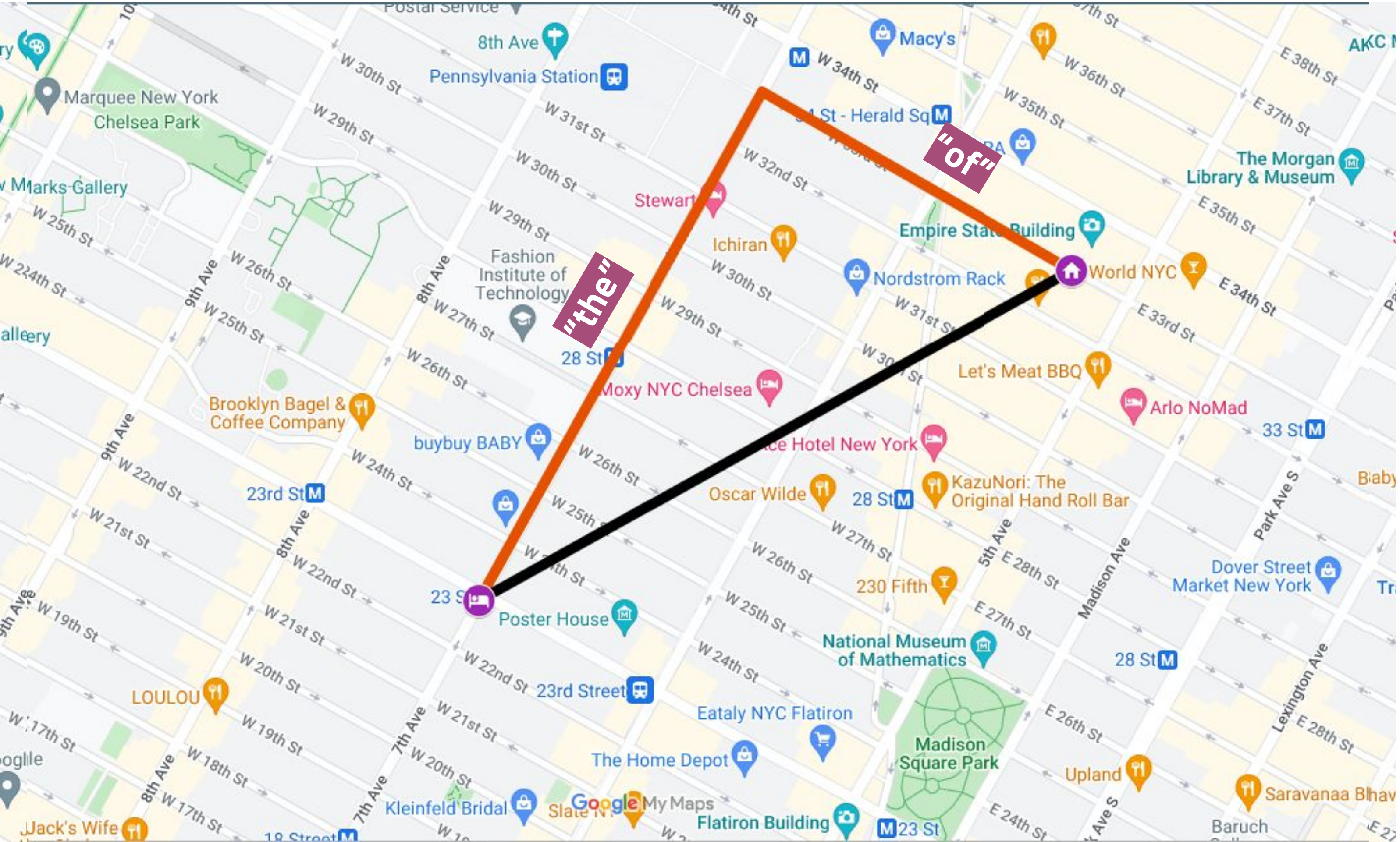
Petr Plecháč: https://versologie.cz/talks/2017chicago/



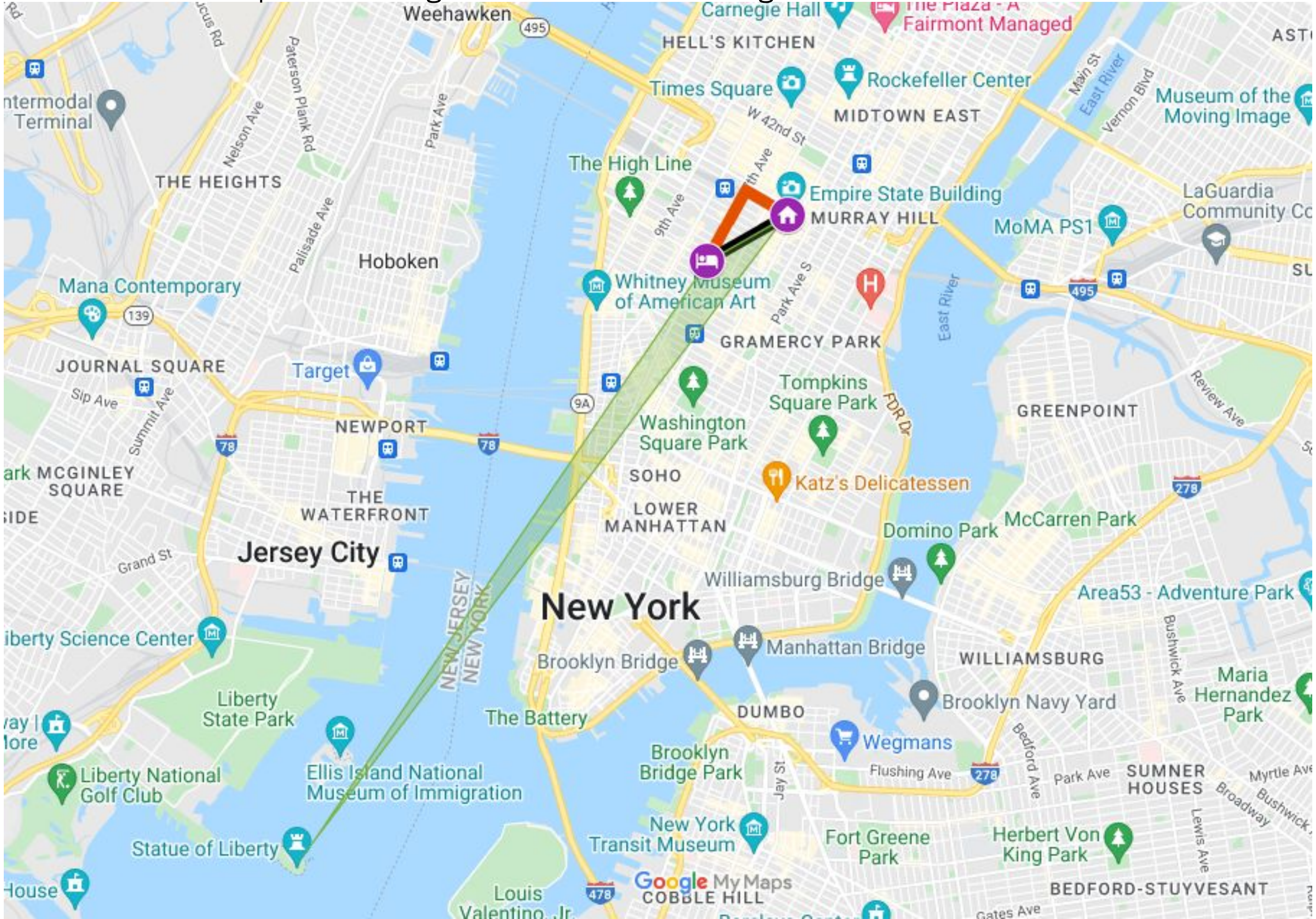
Petr Plecháč: https://versologie.cz/talks/2017chicago/



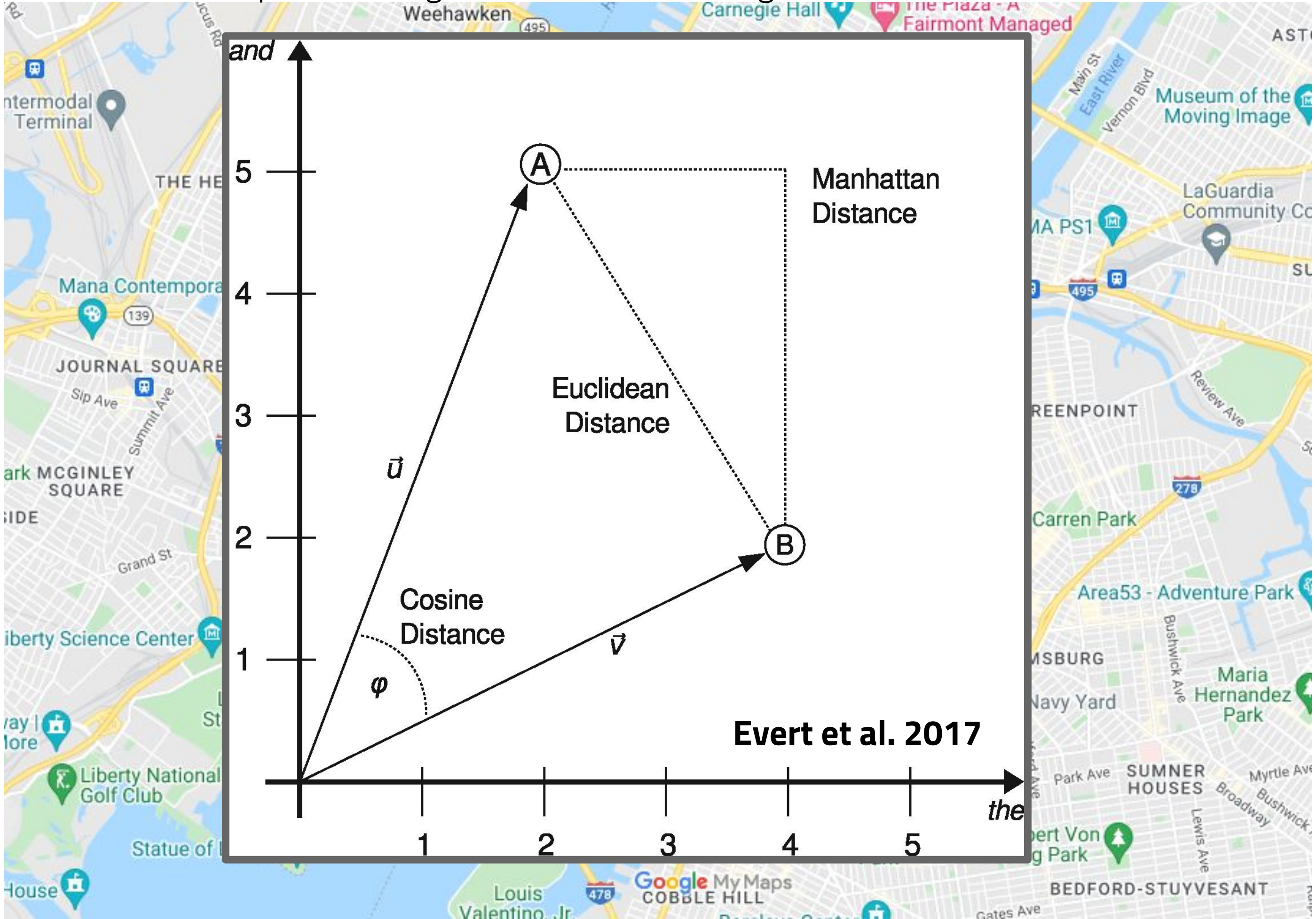
Petr Plecháč: https://versologie.cz/talks/2017chicago/



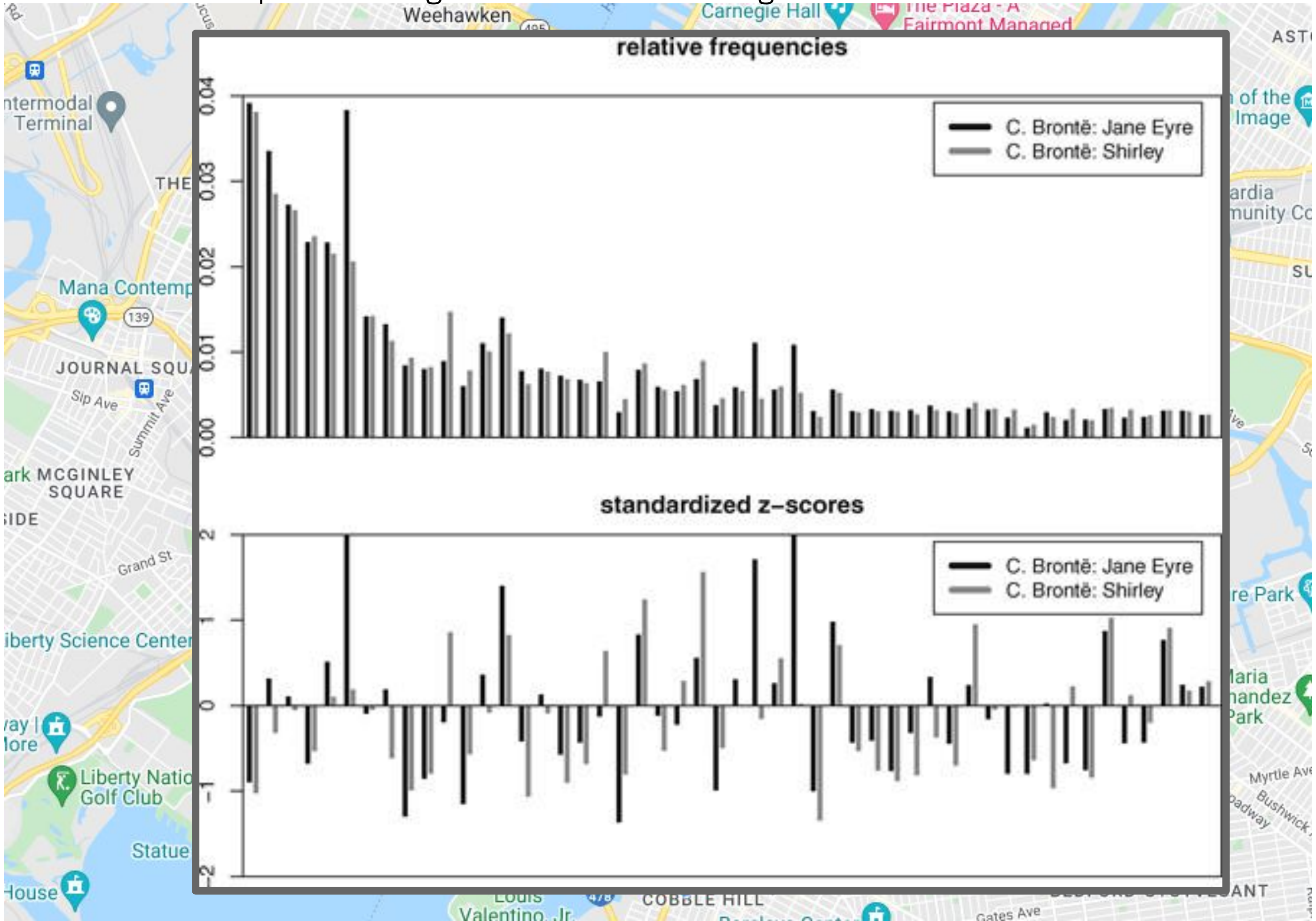
Petr Plecháč: https://versologie.cz/talks/2017chicago/



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



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



Note on distances for French

Burrow's Delta
(with Euclidean normalization)

Why Molière most likely did write his plays

FLORIAN CAFIERO AND JEAN-BAPTISTE CAMPS Authors Info & Affiliations

SCIENCE ADVANCES • 27 Nov 2019 • Vol 5, Issue 11 • DOI: 10.1126/sciadv.aax5489

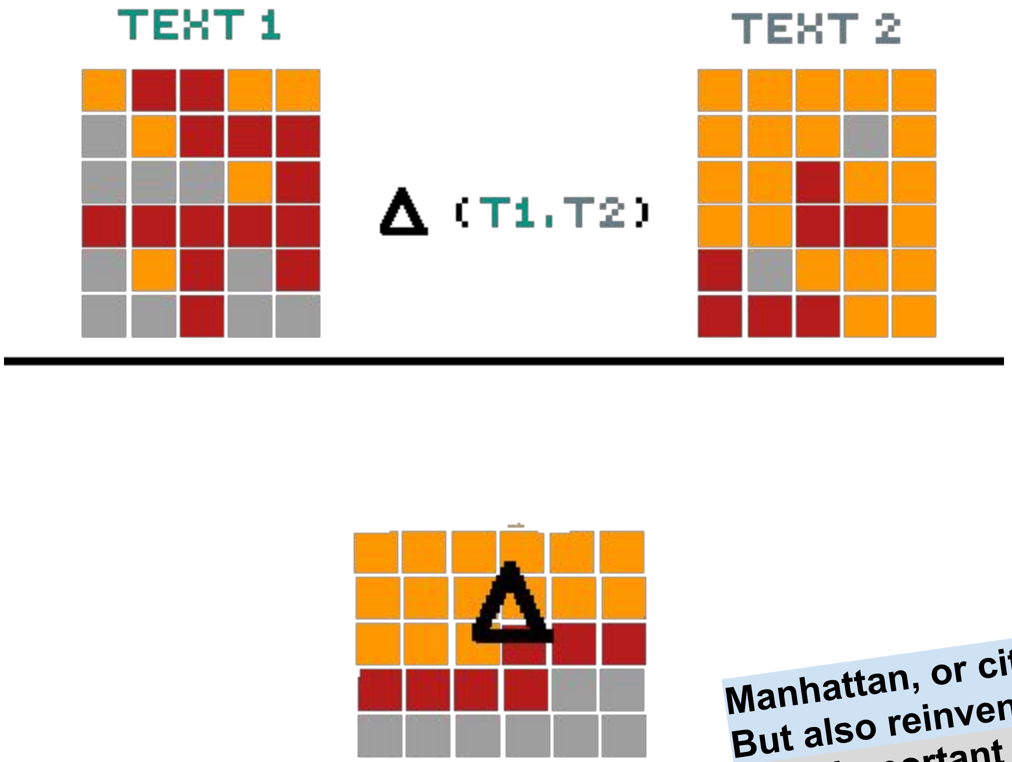
Cosine Delta
(Wurzburg)

JOURNAL ARTICLE

Understanding and explaining Delta measures for authorship attribution FREE

Stefan Evert, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, Thorsten Vitt

Digital Scholarship in the Humanities, Volume 32, Issue suppl_2, December 2017, Pages ii4–ii16, https://doi.org/10.1093/llc/fqx023



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

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

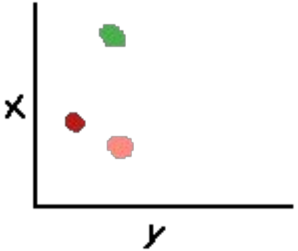
DISTANCE MATRIX

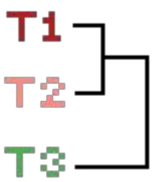
	T1	T2	T3
T1	0		
T2	0.3	0	
T3	0.7	0.9	0

DISTANCE MATRIX

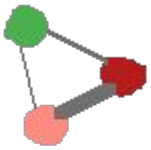
	T1	T2	T3
T1	0		
T2	0.3	0	
T3	0.7	0.9	0

MULTIDIMENSIONAL SCALING



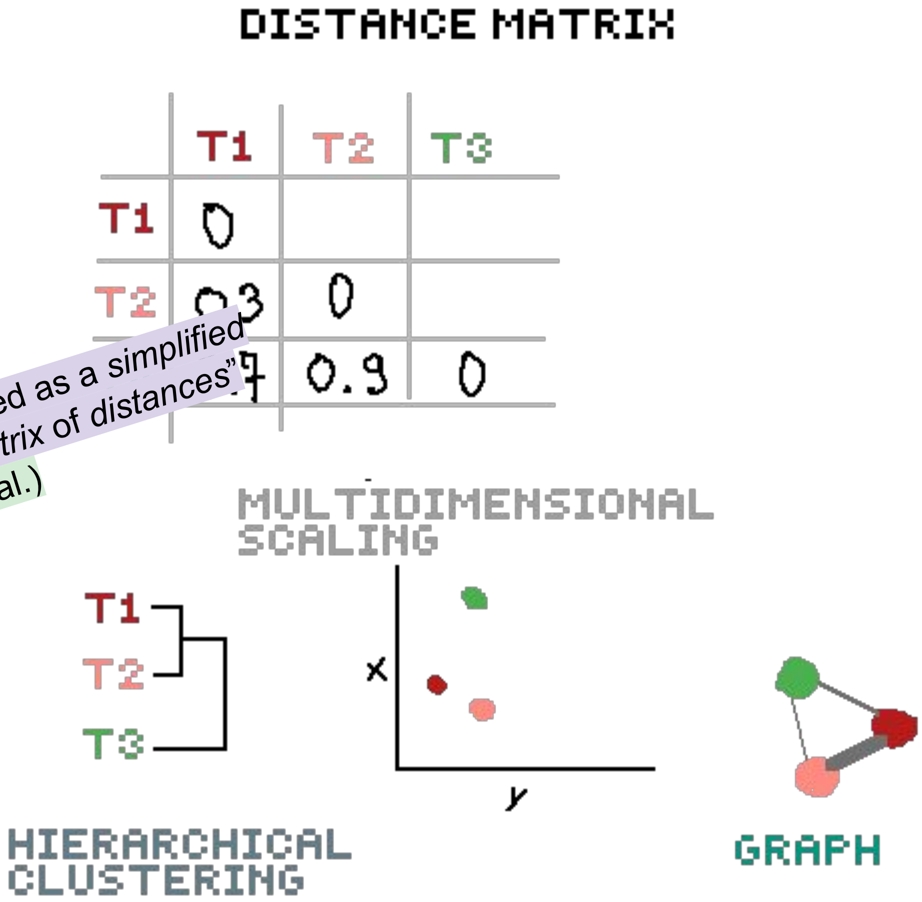


HIERARCHICAL CLUSTERING

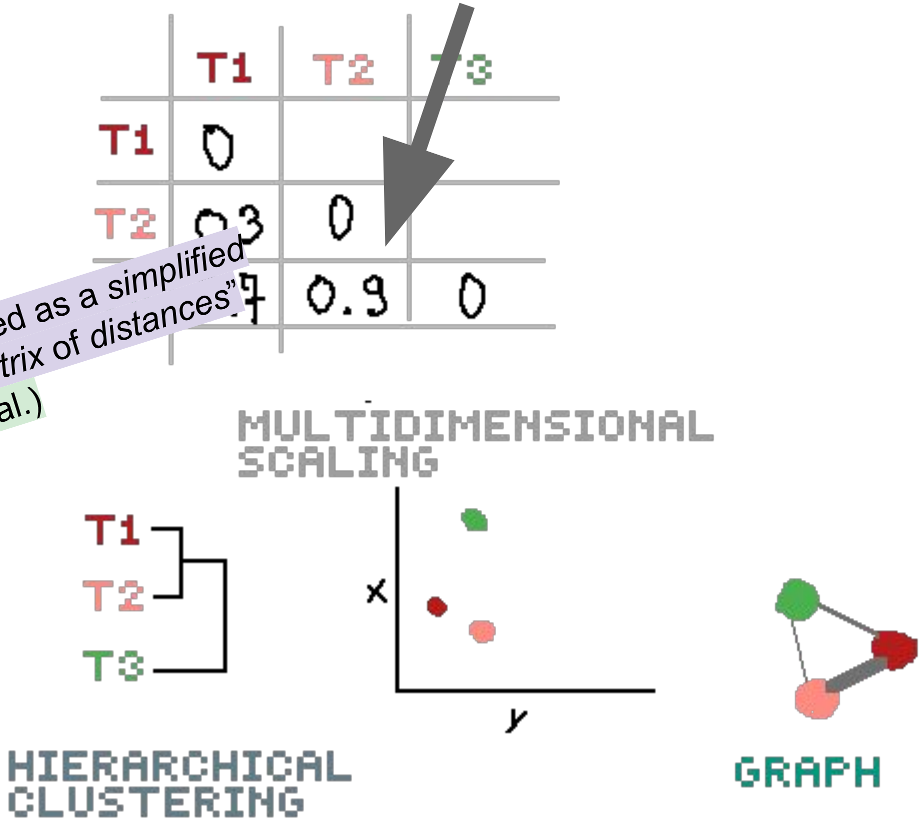


GRAPH

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



OK, but how much can I trust this distance measure?



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

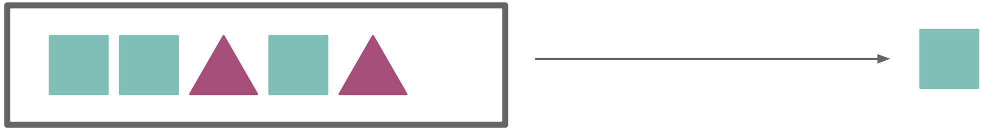
Sampling, bootstrapping, iterations!

Sidenote

Sampling without replacement:

Sidenote

Sampling without replacement:



Sidenote

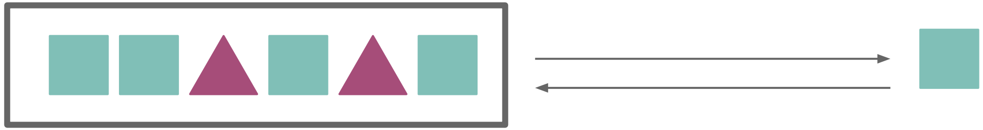
Sampling without replacement:

Sidenote

Sampling without replacement:

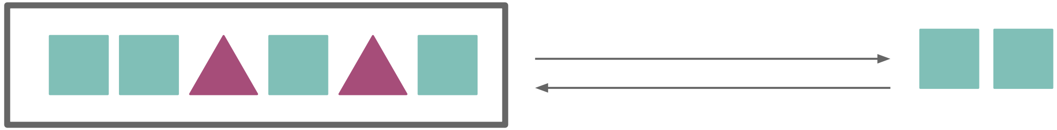
Sidenote

Sampling *with* replacement:



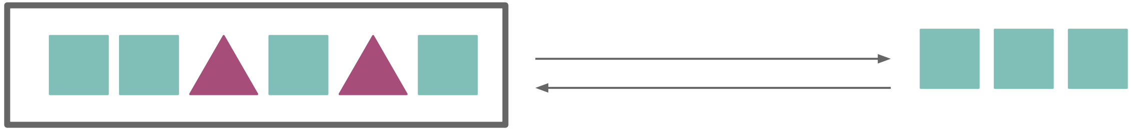
Sidenote

Sampling *with* replacement:



Sidenote

Sampling *with* replacement:



2. Sampling & bootstrapping

Sample: 

p(square) = 0.66

2. Sampling & bootstrapping

Sample: ■ ■ ▲ ■ ▲ ■

p(square) = 0.66

Resample 1: ■ ▲ ▲ ■ ▲ ■ 0.5

2. Sampling & bootstrapping

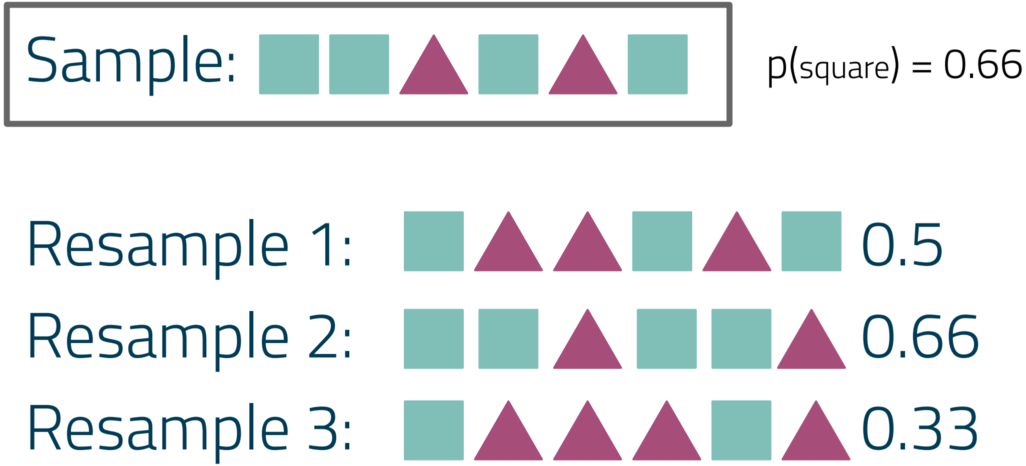
Sample: ■ ■ ▲ ■ ▲ ■

p(square) = 0.66

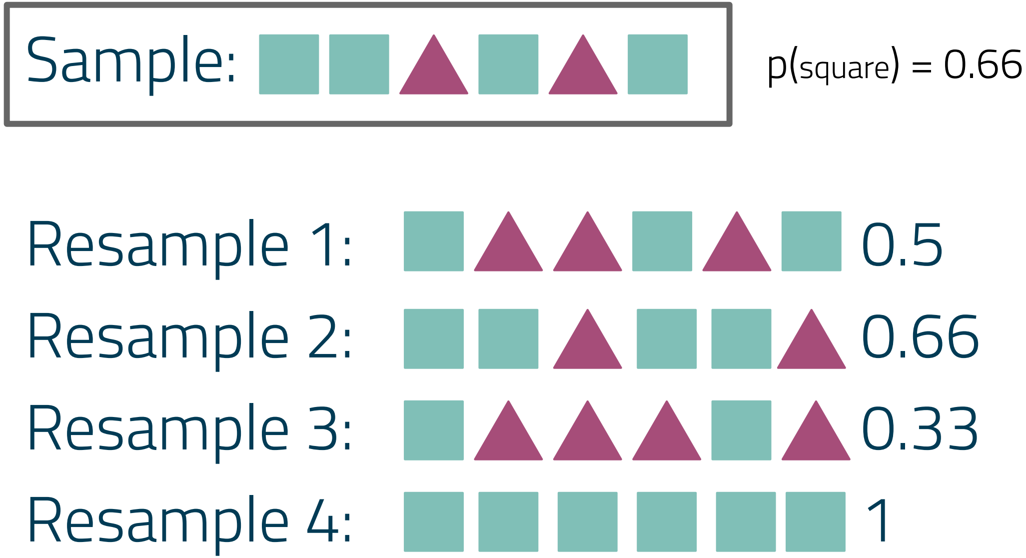
Resample 1: ■ ▲ ▲ ■ ▲ ■ 0.5

Resample 2: ■ ■ ▲ ■ ■ ▲ 0.66

2. Sampling & bootstrapping



2. Sampling & bootstrapping



2. Sampling & bootstrapping

Sample:

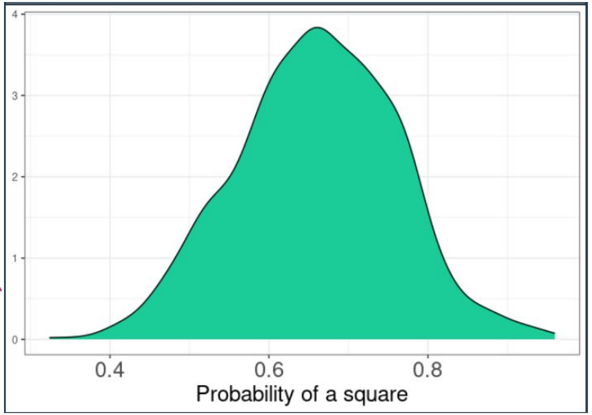
p(square) = 0.66

Resample 1:

Resample 2:

Resample 3:

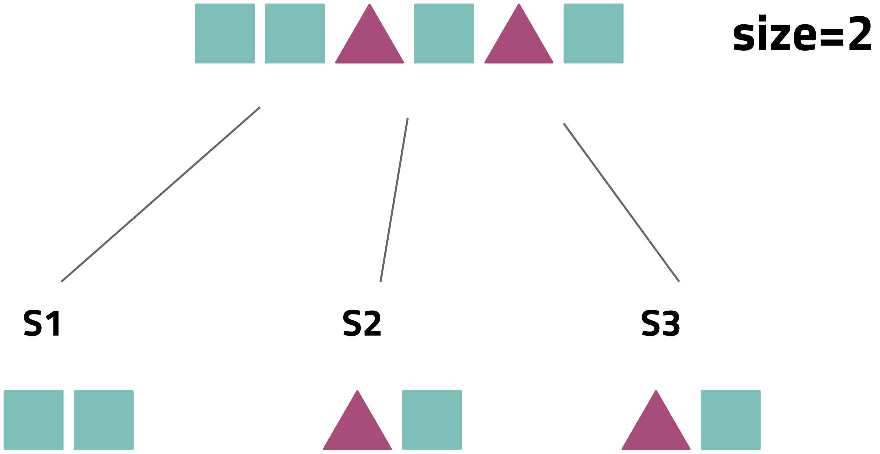
Resample 4:



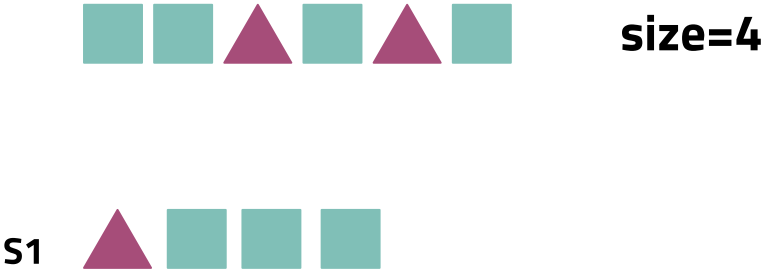
3. Estimating uncertainty in text similarity **(within *stylo*)**

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

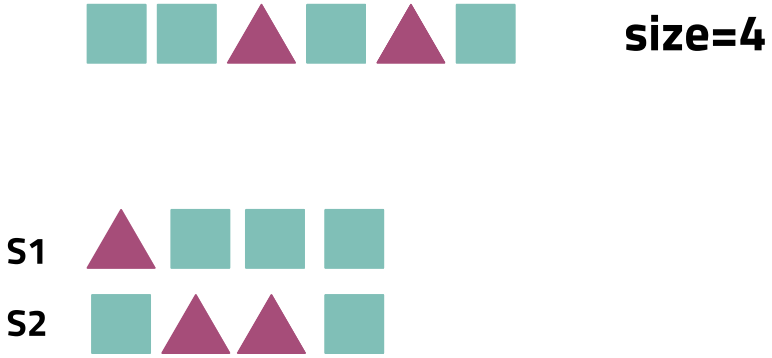
Normal vs. random sampling **(in stylo)**



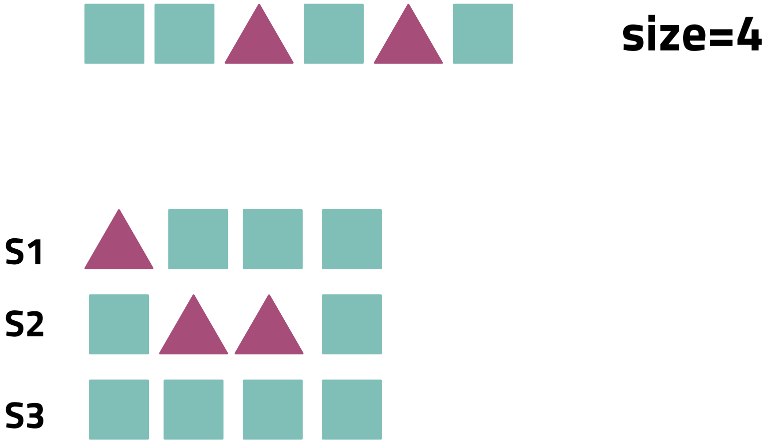
Normal vs. random sampling **(in stylo)**



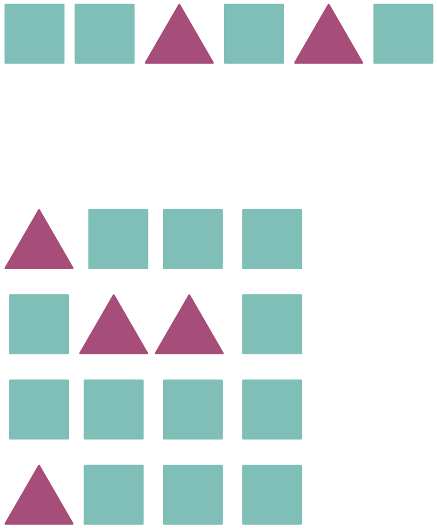
Normal vs. random sampling (in stylo)



Normal vs. random sampling (in stylo)



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



size=4

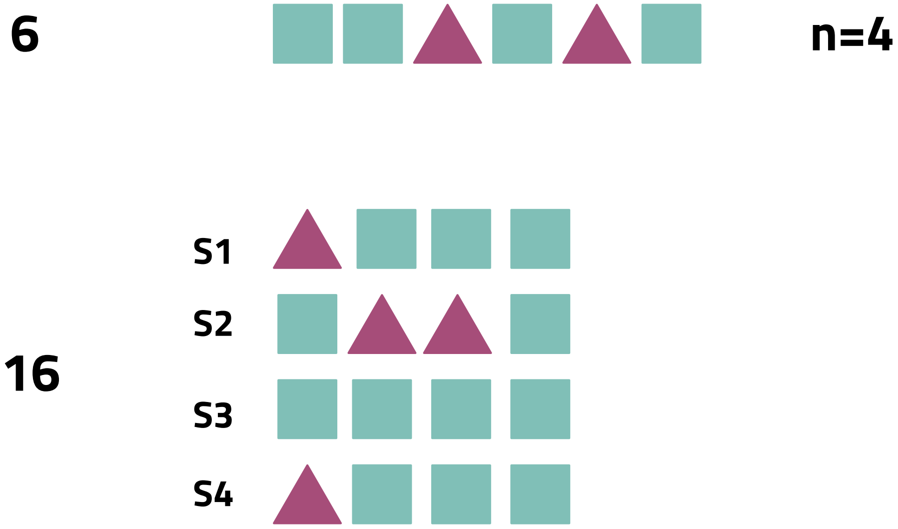
S1

S2

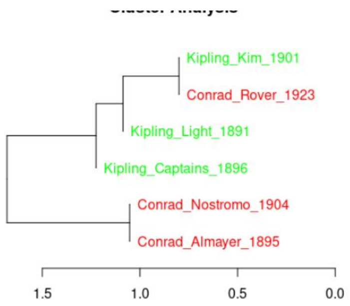
S3

S4

Normal vs. random sampling **(in stylo)**

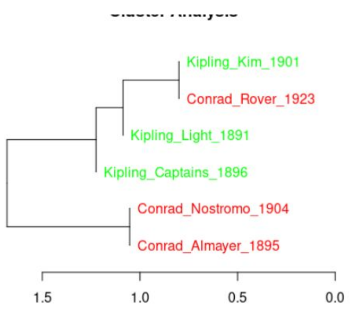


4. Consensus trees

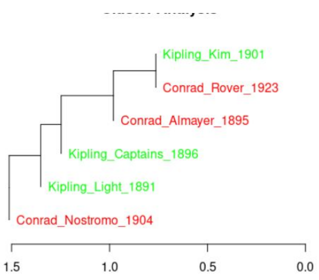


Feature set 1

4. Consensus trees

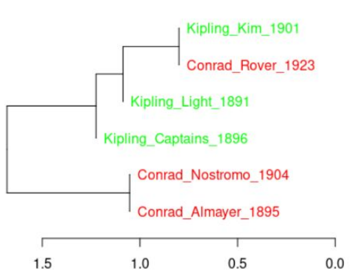


Feature set 1

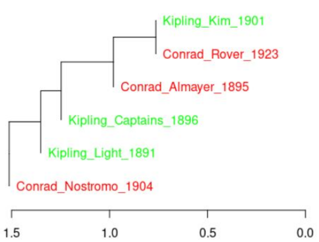


Feature set 2

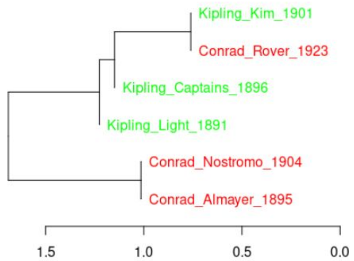
4. Consensus trees



Feature set 1

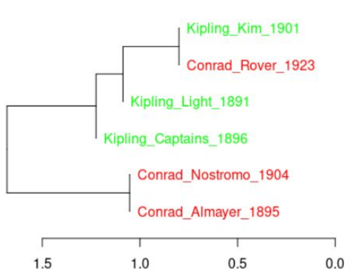


Feature set 2

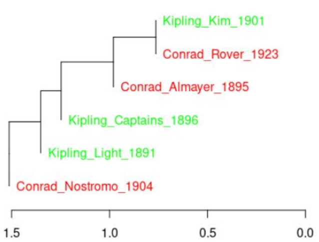


Feature set 3

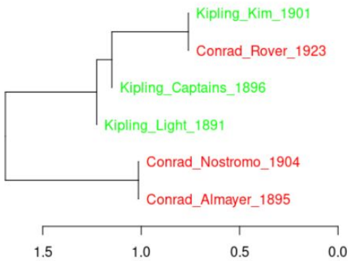
4. Consensus trees



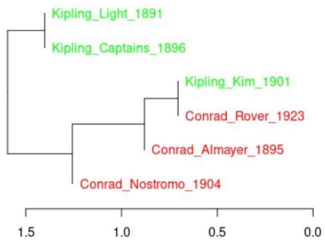
Feature set 1



Feature set 2

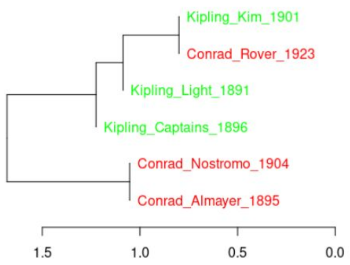


Feature set 3

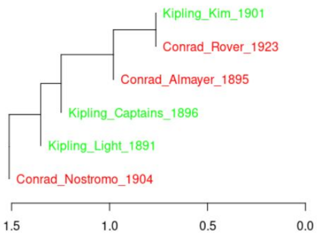


Feature set 4

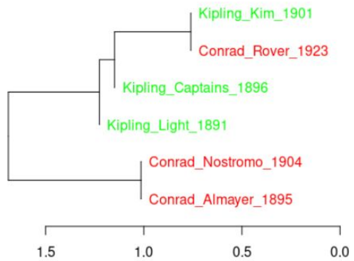
4. Majority rule (>50% of branches)



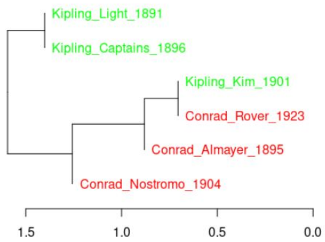
Feature set 1



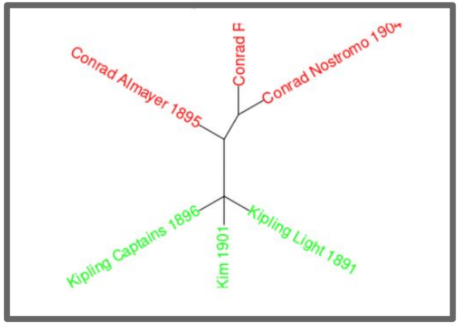
Feature set 2



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”:

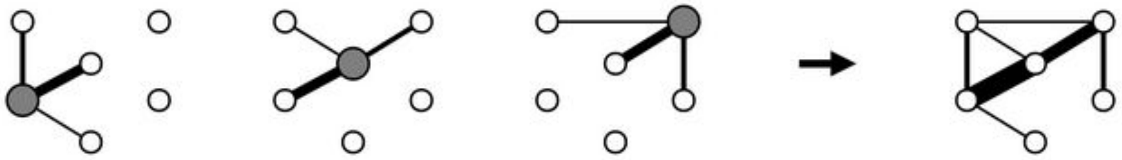
- MFW length
- Culling strength
- Text themselves (take samples from texts)

….

But the possibilities are limitless

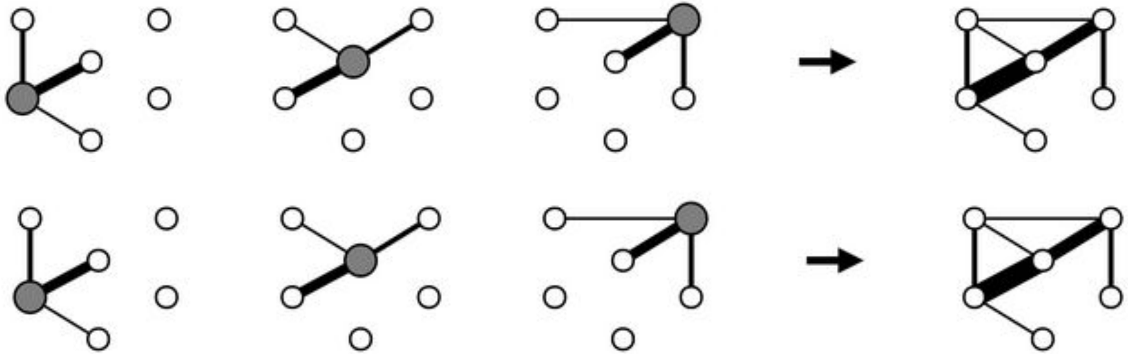
6. Consensus networks

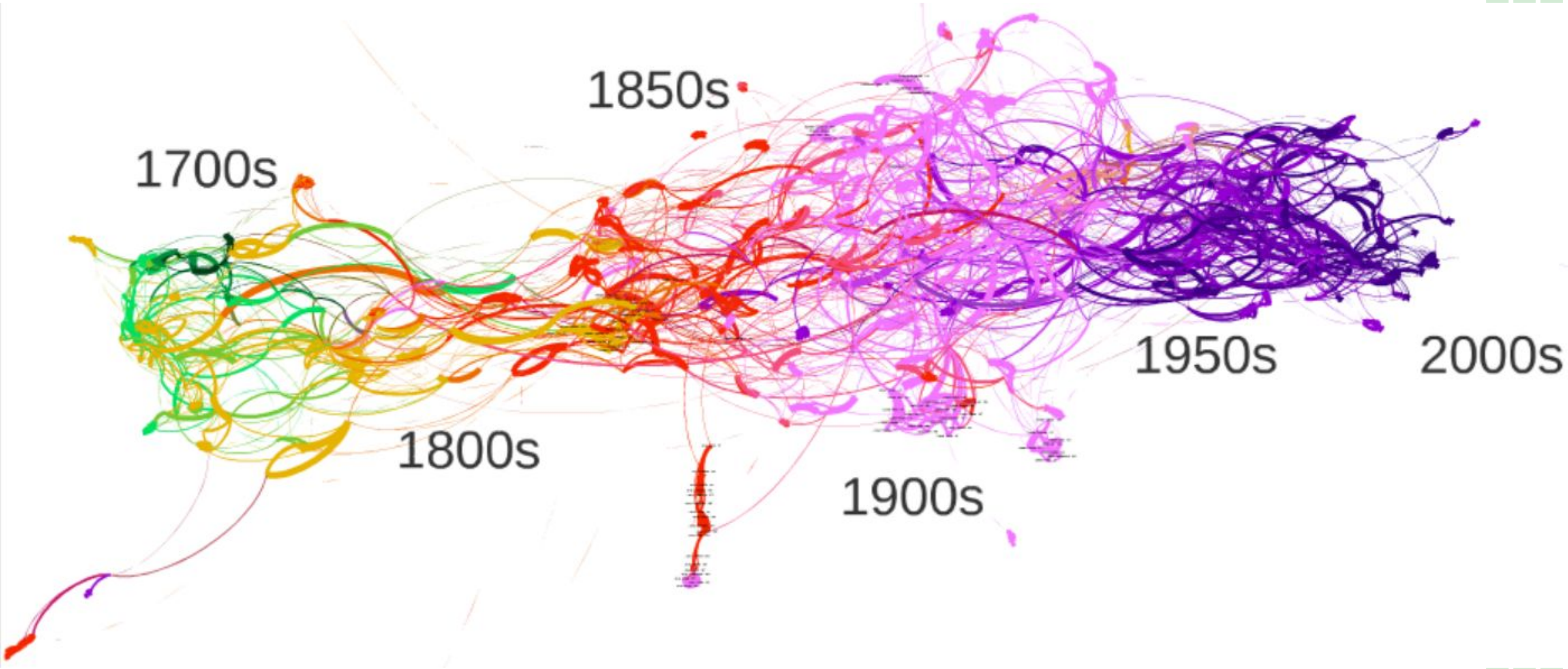
1. Look at the neighbours!



6. Consensus networks

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



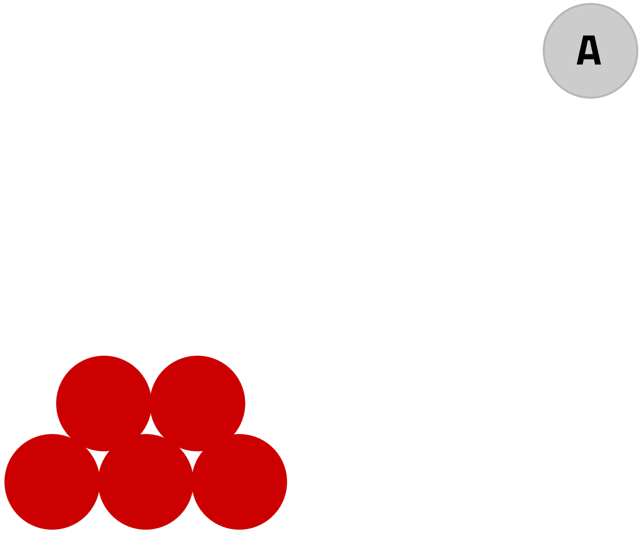

1700s
1800s
1850s
1900s
1950s
2000s

- 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**

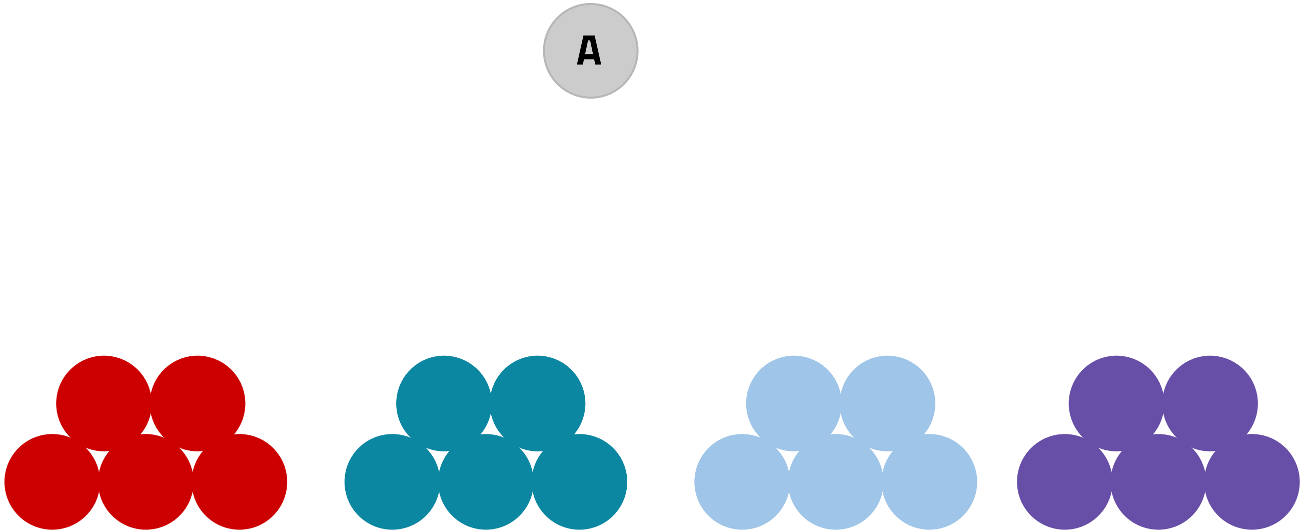
6. General imposters

A

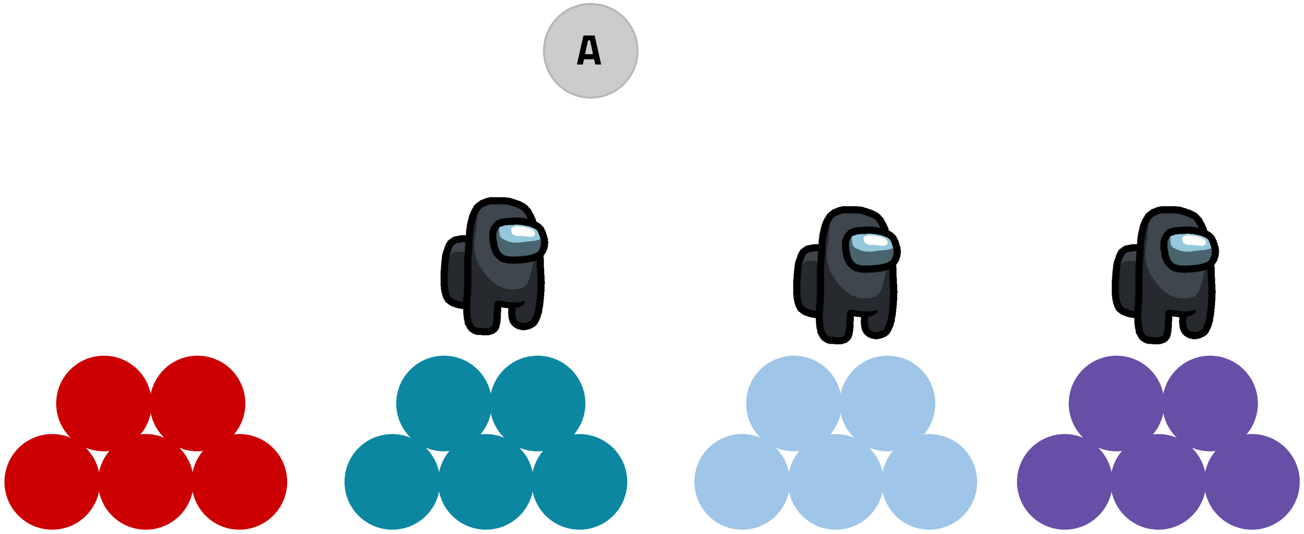
6. General imposters



6. General imposters



6. General imposters



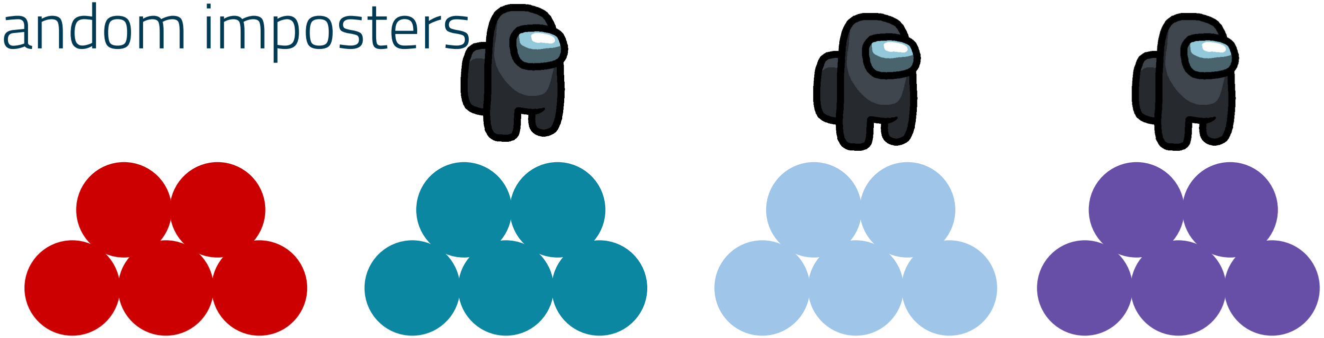
6. General imposters

Random samples

A

Random features

Random imposters

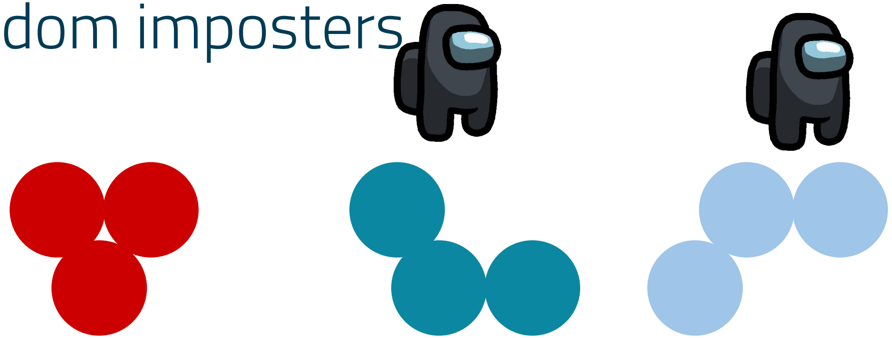


6. General imposters

Random samples

Random features

Random imposters

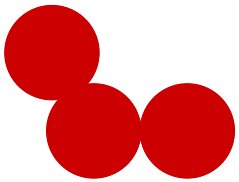


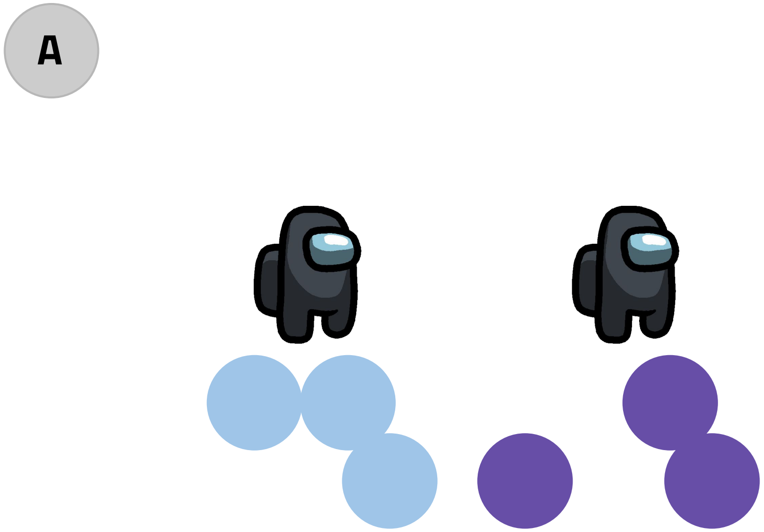
6. General imposters

Random samples

Random features

Random imposters





6. General imposters