

# DIY Computational Text Analysis with R

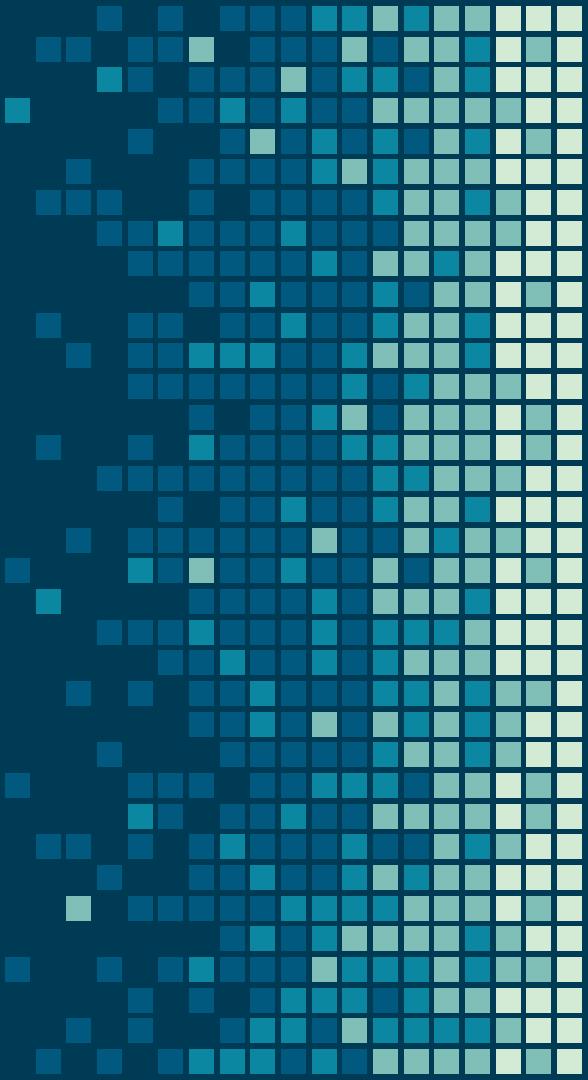
## – introductions

Joanna Byszuk & Jeremi Ochab

DHSI 2024

“DIY Computational Text Analysis  
with R”

Day One



# About us

10 Computational 01  
01 Stylistics 0101000  
11 Group 011010110



Computational Stylistics Group in Kraków & Digital Humanities Lab @Jagiellonian

Joanna Byszuk (researcher at IJP PAN),  
Jeremi Ochab (researcher at Jagiellonian University)

# About us

10 Computational  
01 Stylistics  
11 Group 0101000  
011010110



Research on:

- various aspects of stylometry (methods, languages, problems),
- natural language processing (e.g. HTR, OCR, Direct Speech),
- deep learning in humanities (distributional semantics, multimodal data analysis)

# Useful links

<https://computationalstylistics.github.io/>

<https://github.com/computationalstylistics/stylo>

<https://journal.r-project.org/archive/2016/RJ-2016-007/RJ-2016-007.pdf>

# Materials

[https://github.com/JoannaBy/DHSI2024-DIY/tree/  
main](https://github.com/JoannaBy/DHSI2024-DIY/tree/main) – notes, corpora, tasks

# Course plan

Monday, 10th June 2024

- 10:15-12:00 Introduction to stylometry and textual analysis
- 1:00-2.15 Introduction to supervised and unsupervised machine learning, designing a study and first experiments

# Course plan

Tuesday, 11th June 2024

- 9:00-12:00 Exploratory analysis – methods of dimension reduction and feature explainability
- 1:00-3.45 Cluster analysis, consensus trees and feature engineering

# Course plan

Wednesday, 12th June 2024

- 9:00-12:00 Classification methods and validation of the results
- 1:00-3.45 Sequential analysis of singular works

# Course plan

Thursday, 13th June 2024

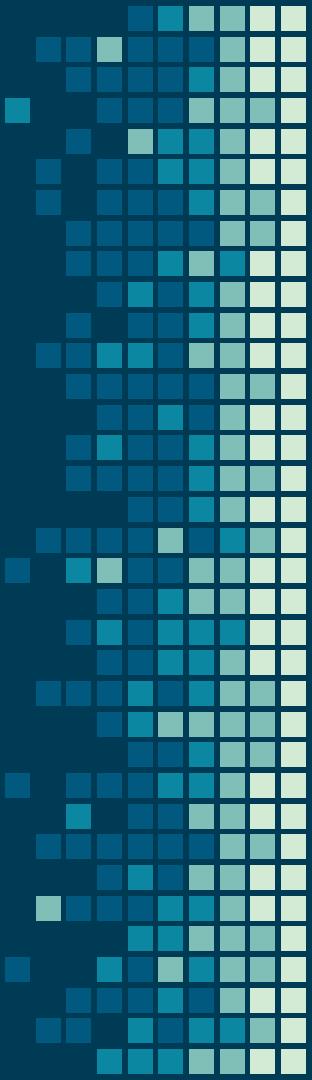
- 9:00-12:00 Network analysis
- 1:00-3.45 Designing serial experiments and adding more R

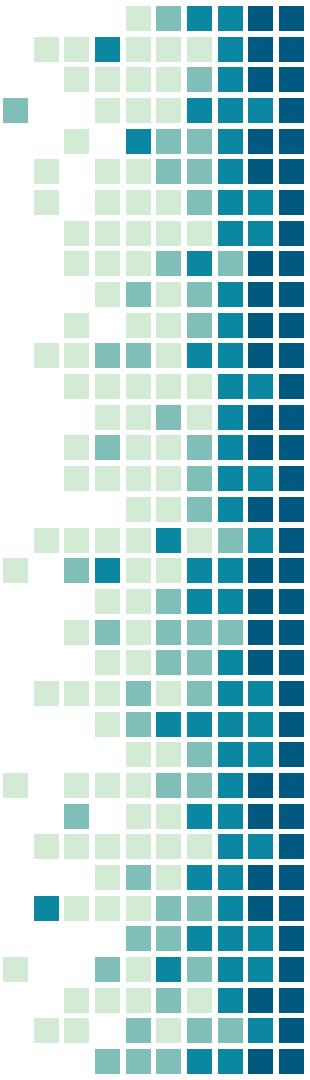
# Course plan

Friday, 14th June 2024

- 9:00-10:15 What Have We Learned?

And now to you





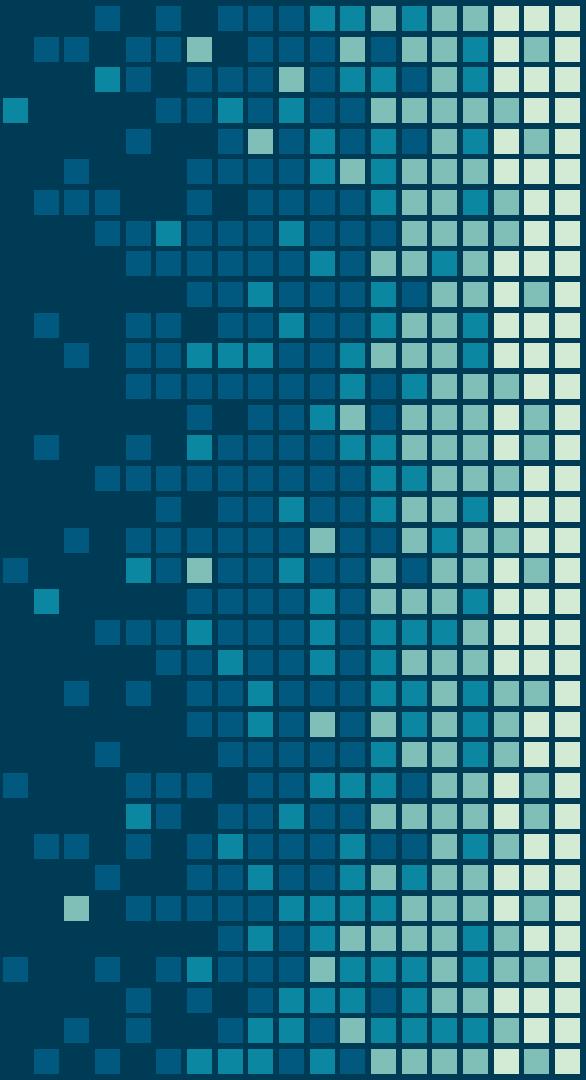
What's your name (and pronouns)?

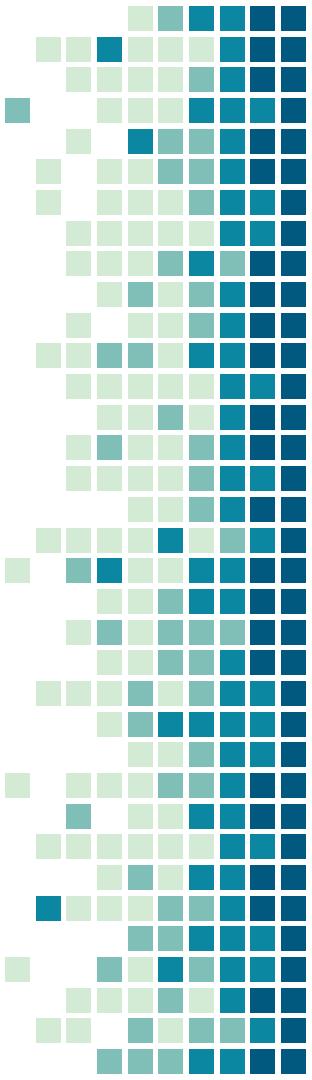
Where are you from?

Why are you interested in stylometry?

What's the project you'd like to do here?

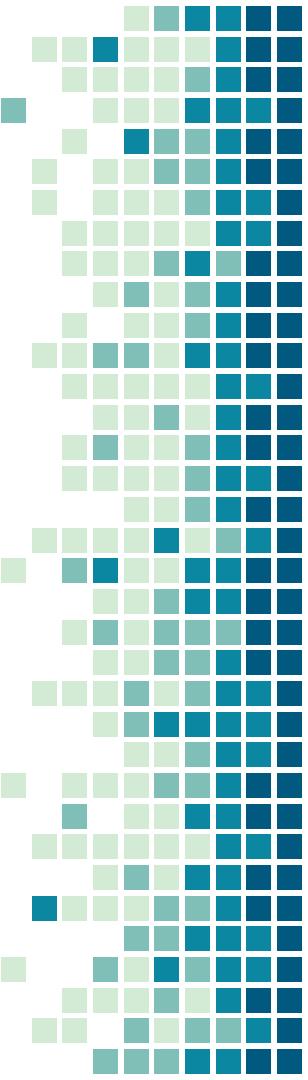
# Introduction to stylometry





# Distant reading

- **Analysing big literary collections** based on not text but metadata / research literature etc.  
(as conceptualized by Franco Moretti 2000)
- **Analysing language / literature “objectively”**
  - **at a distance**, looking at particular features within the texts  
(Mendenhall 1887, Lutosławski 1890)  
(but also Lorenzo Valla 1440, Augustus de Morgan 1851)



# What is stylometry?

Stylometry =  
use of quantitative methods  
to examine similarities and differences  
within a group of [texts]

# Stylometry is related to

- Computational & Forensic Linguistics
- Network Analysis
- Natural Language Processing

How does it work?

# How does stylometry work?

corpus of texts

+

distance measure

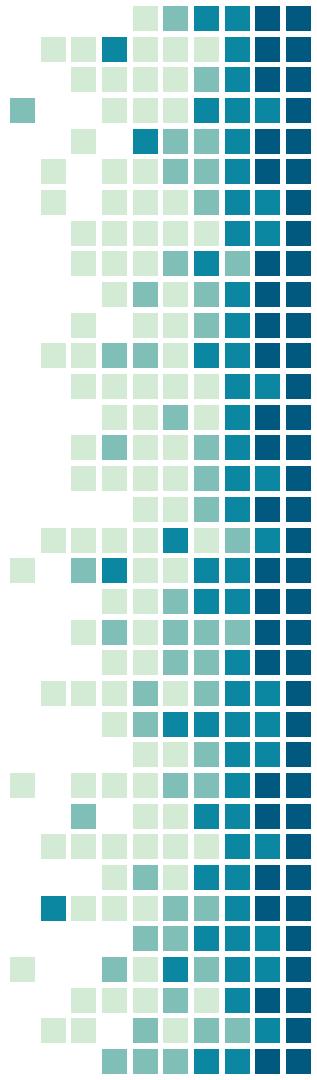
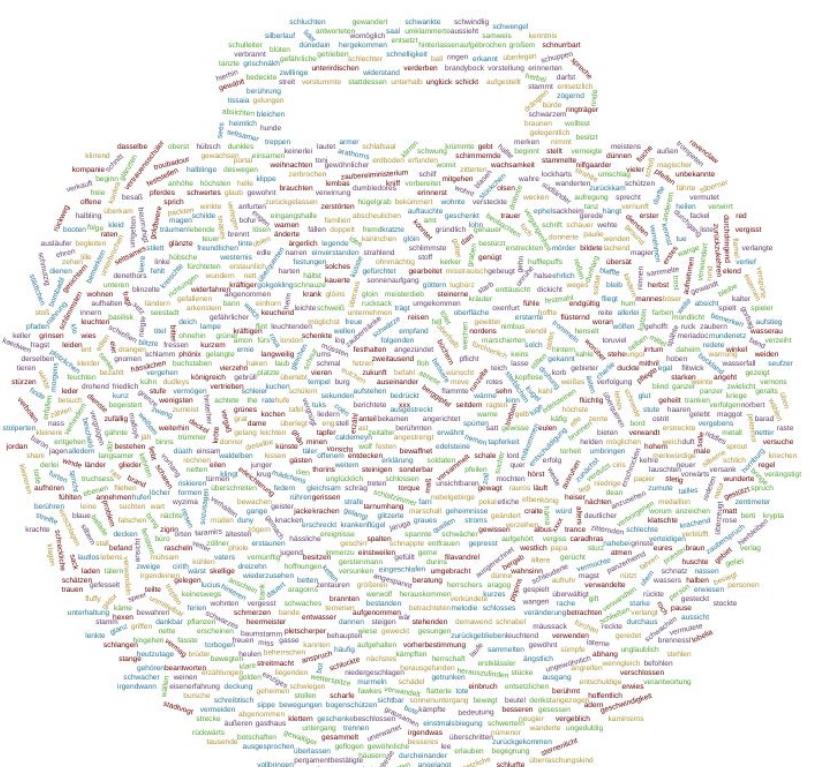
+

classification algorithm

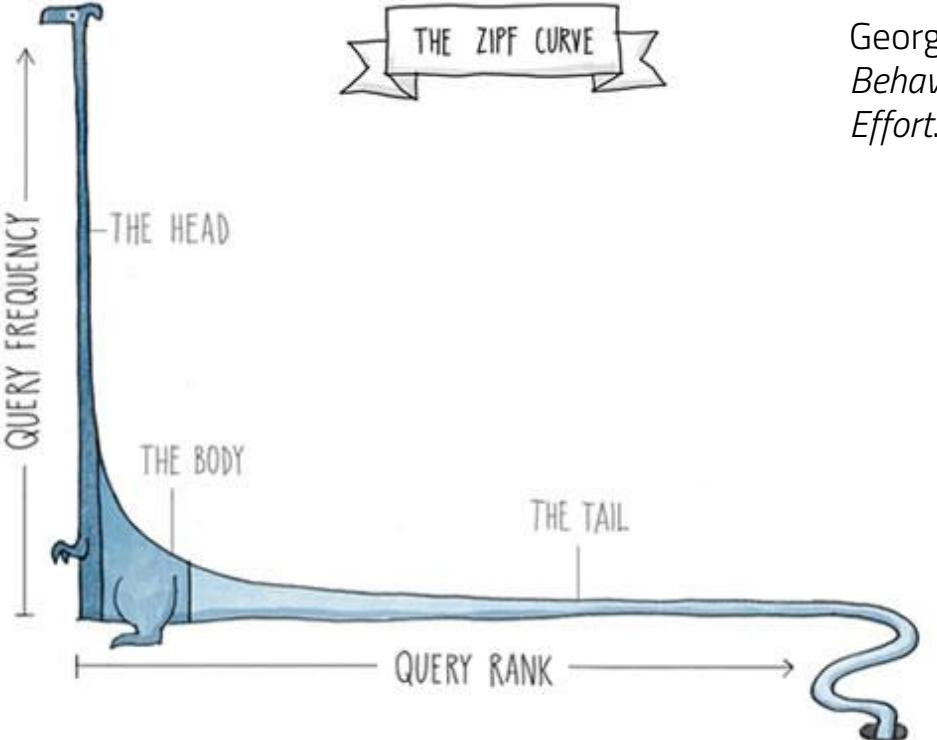
+

(visualisation)

# Text as bag-of-words



# Zipf's law

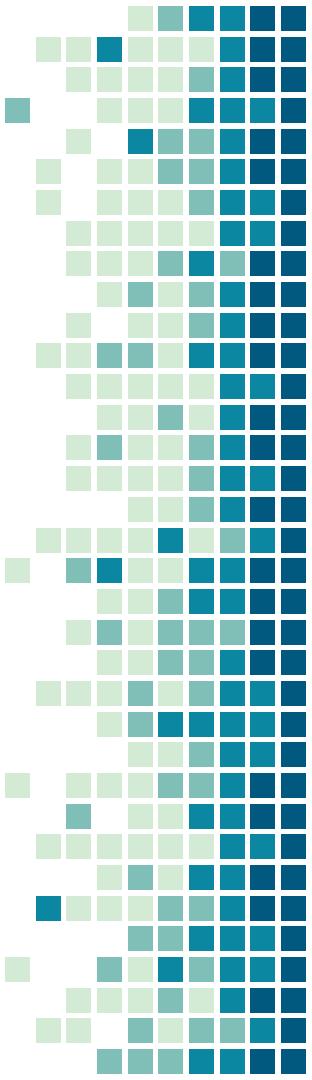


George K. Zipf (1949) *Human Behavior and the Principle of Least Effort*. Addison-Wesley.

# Zipf's law

Am I the only one around here  
that tries to do things with the  
least effort possible and  
expects a good result?!





# What words are at the top?

	Agnes	Tenant	Emma	Pride	Sense	Jane
the	2511	5929	5204	4330	4105	7835
and	2733	6705	4878	3577	3489	6618
to	2366	5594	5186	4136	4103	5152
of	1602	3734	4292	3609	3571	4359
i	2204	6075	3191	2064	1998	7165
a	1296	2792	3126	1948	2067	4467
in	911	2021	2174	1866	1948	2762
that	776	1909	1800	1577	1383	1655
he	659	2259	1811	1338	1112	1902
was	1000	1835	2400	1847	1861	2525
it	795	2280	2529	1532	1755	2403
you	760	2844	1999	1356	1191	2971
her	750	1760	2483	2224	2543	1714

	Agnes	Tenant	Emma	Pride	Sense	Jane
the	3.67471	3.54285	3.24344	3.55705	3.43227	4.18704
and	3.99959	4.00655	3.04026	2.93847	2.91722	3.53667
to	3.46251	3.34267	3.23222	3.39768	3.43060	2.75324
of	2.34444	2.23124	2.67503	2.96476	2.98579	2.32946
i	3.22543	3.63009	1.98882	1.69556	1.67057	3.82899
a	1.89662	1.66835	1.94831	1.60026	1.72826	2.38717
in	1.33320	1.20764	1.35496	1.53290	1.62876	1.47602
that	1.13563	1.14072	1.12187	1.29549	1.15635	0.88444
he	0.96441	1.34986	1.12872	1.09915	0.92977	1.01643
was	1.46344	1.09650	1.49582	1.51729	1.55602	1.34937
it	1.16344	1.36241	1.57622	1.25852	1.46739	1.28417
you	1.11222	1.69942	1.24589	1.11394	0.99582	1.58771
her	1.09758	1.05168	1.54755	1.82699	2.12625	0.91597

For two texts  $T$  and  $T_1$ , and for a set of  $n$  words,

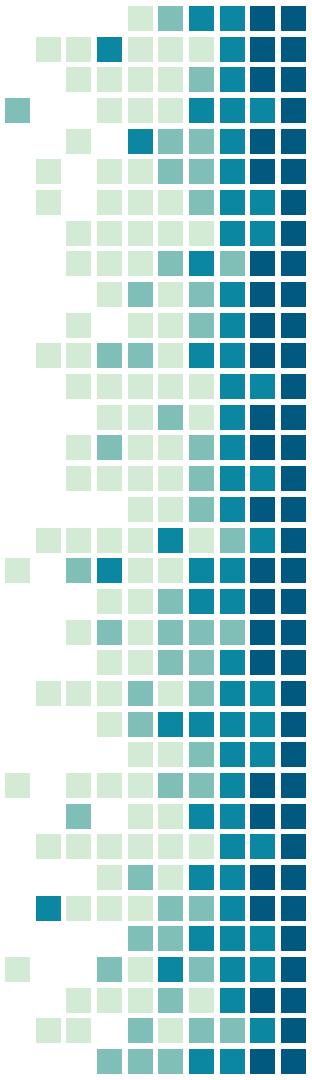
$$\Delta(T, T_1) = \frac{1}{n} \sum_{i=1}^n |z(f_i(T)) - z(f_i(T_1))|$$

Where  $z(f_x(T)) = \frac{f_x(T) - \mu_x}{\sigma_x}$ ;

$f_x(T)$  = raw frequency of word  $x$  in text  $T$ ;

$\mu_x$  = mean frequency of word  $x$  in a collection of texts;

$\sigma_x$  = standard deviation of frequency of word  $x$ .



= what's the difference in how two texts use a given feature, compared to its average use

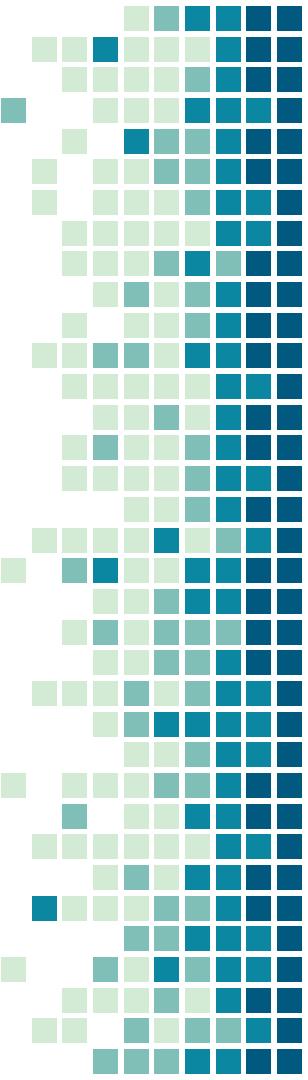
E.g. ....

	Agnes	Tenant	Emma	Pride	Sense	Jane
the	3.67471	3.54285	3.24344	3.55705	3.43227	4.18704
and	3.99959	4.00655	3.04026	2.93847	2.91722	3.53667
to	3.46251	3.34267	3.23222	3.39768	3.43060	2.75324
of	2.34444	2.23124	2.67503	2.96476	2.98579	2.32946
i	3.22543	3.63009	1.98882	1.69556	1.67057	3.82899
a	1.89662	1.66835	1.94831	1.60026	1.72826	2.38717
in	1.33320	1.20764	1.35496	1.53290	1.62876	1.47602
that	1.13563	1.14072	1.12187	1.29549	1.15635	0.88444
he	0.96441	1.34986	1.12872	1.09915	0.92977	1.01643
was	1.46344	1.09650	1.49582	1.51729	1.55602	1.34937
it	1.16344	1.36241	1.57622	1.25852	1.46739	1.28417
you	1.11222	1.69942	1.24589	1.11394	0.99582	1.58771
her	1.09758	1.05168	1.54755	1.82699	2.12625	0.91597

	Agnes	Pride	Jane	David	Mill	Tom	Clarissa
Tenant	0.81	1.07	0.88	0.92	0.98	1.16	1.1
Emma	1.12	0.78	1.28	1.15	1.2	1.25	1.24
Sense	1.14	0.69	1.24	1.16	1.25	1.13	1.21
Professor	1.06	1.21	0.69	0.94	1	1.27	1.3
Villette	1.07	1.26	0.65	0.91	0.96	1.28	1.3
Bleak	1.09	1.18	0.92	0.55	0.87	1.21	1.17
Hard	1.16	1.25	0.96	0.65	0.91	1.26	1.25
Wuthering	1.06	1.31	0.81	0.94	1.01	1.32	1.27
Adam	1.13	1.37	0.95	0.9	0.66	1.42	1.32
Middlemarch	1.01	1.1	0.99	0.87	0.65	1.17	1.12
Joseph	1.2	1.19	1.24	1.18	1.29	0.64	1.11
Pamela	1.15	1.24	1.27	1.19	1.26	1.11	0.67
Sentimental	1.38	1.53	1.23	1.22	1.29	1.42	1.38

# What words at the top? What relevance?

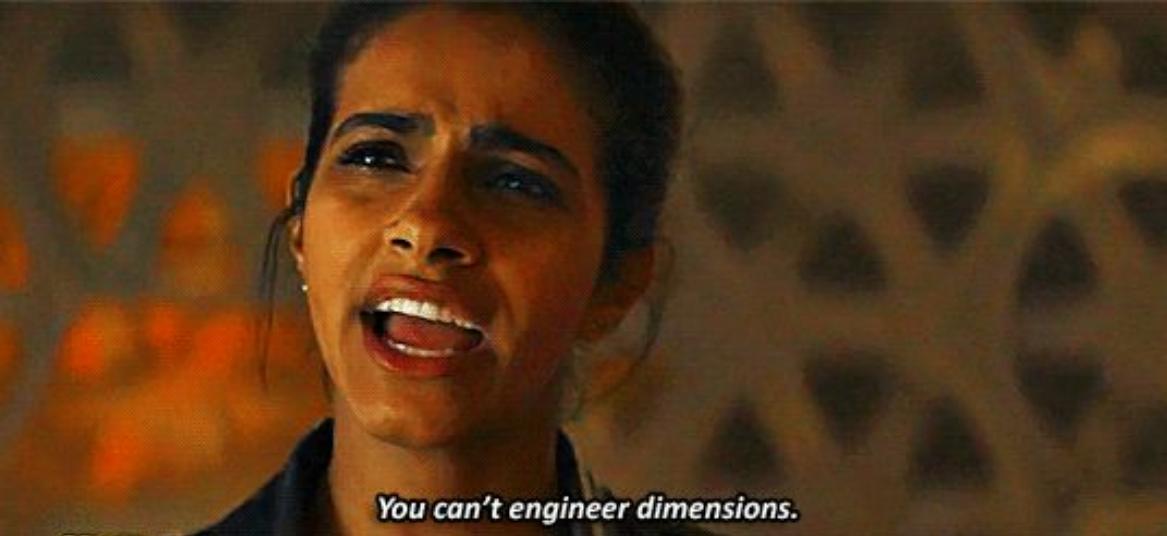
- Grammatical words occupy the top of the frequency list  
  
(Zipf, 1948)
- Grammatical words are strong predictors  
  
(Mosteller & Wallace, 1964)
- Therefore: top N words are strong predictors  
  
(numerous stylometrists around the world)



Cool, but how to  
actually analyse this?



Answer: dimension reduction



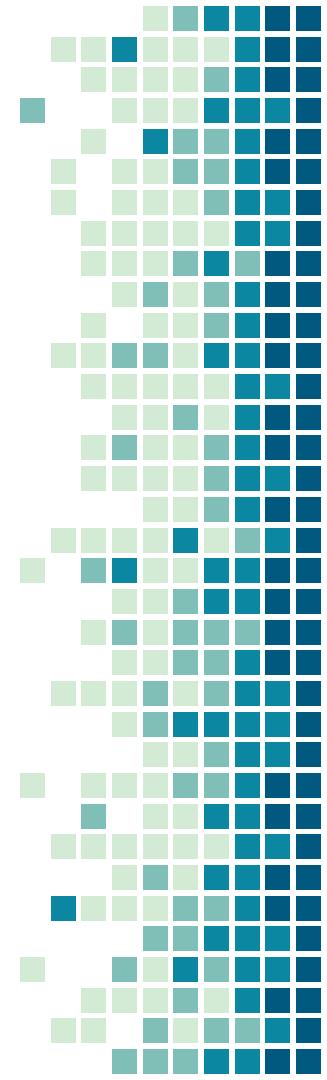
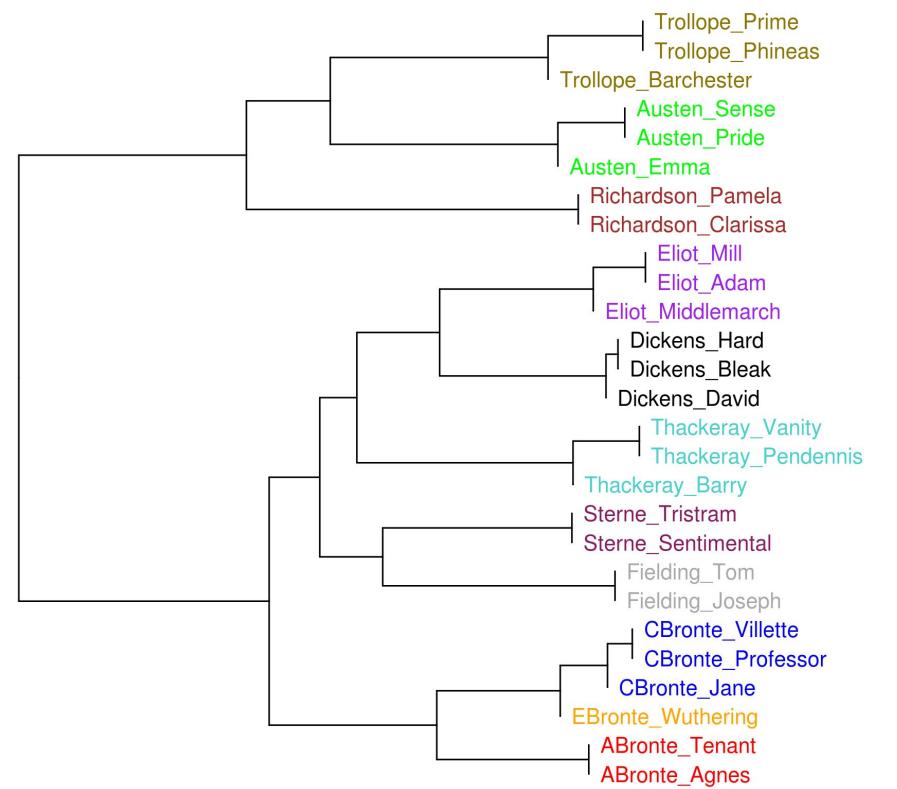
*You can't engineer dimensions.*



*Maybe you can't.*

# #1 Cluster analysis

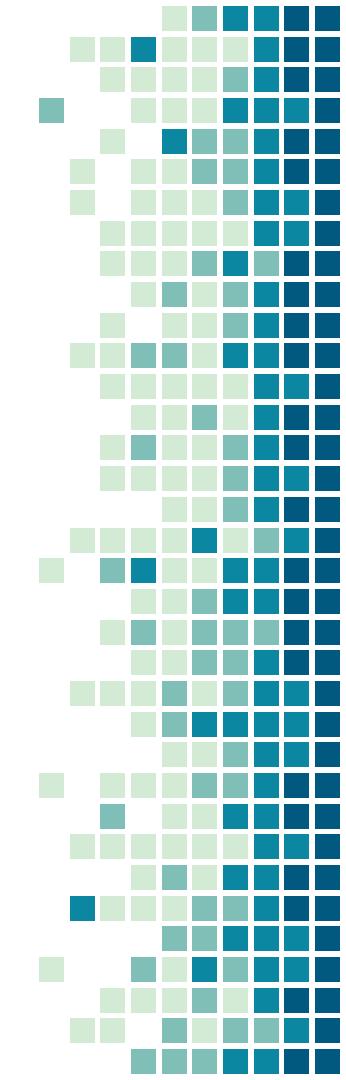
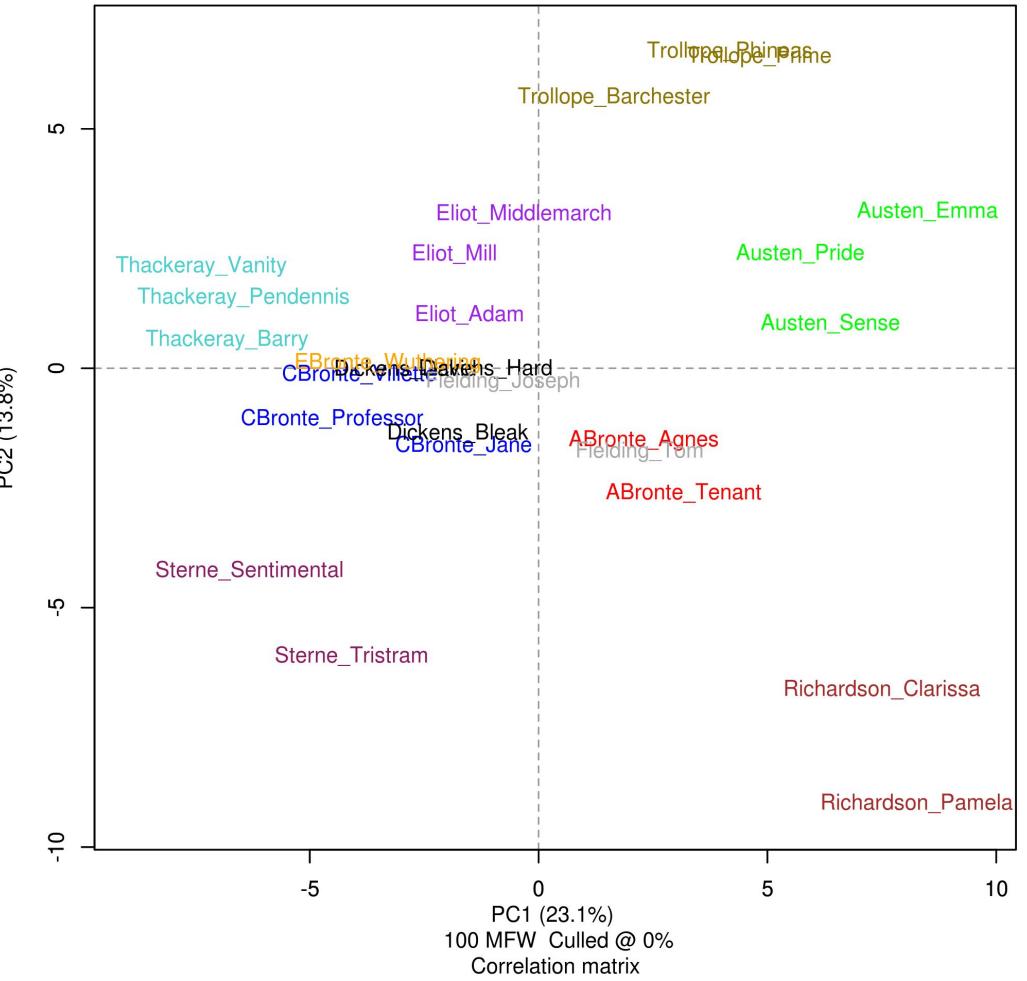
## ASmallCollection Cluster Analysis



#2 PCA

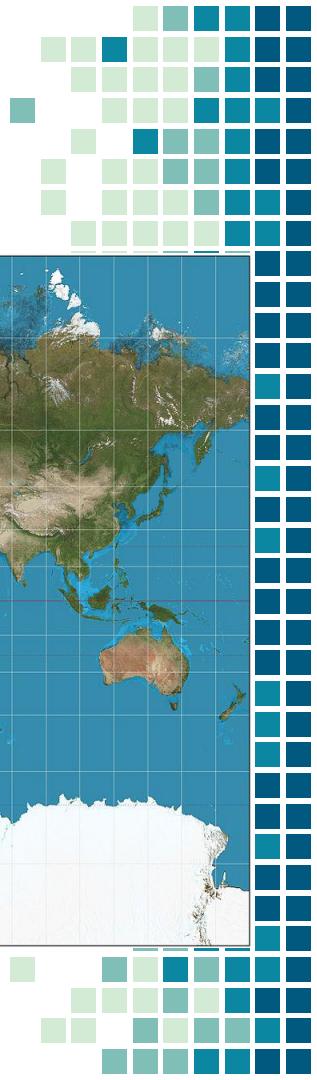
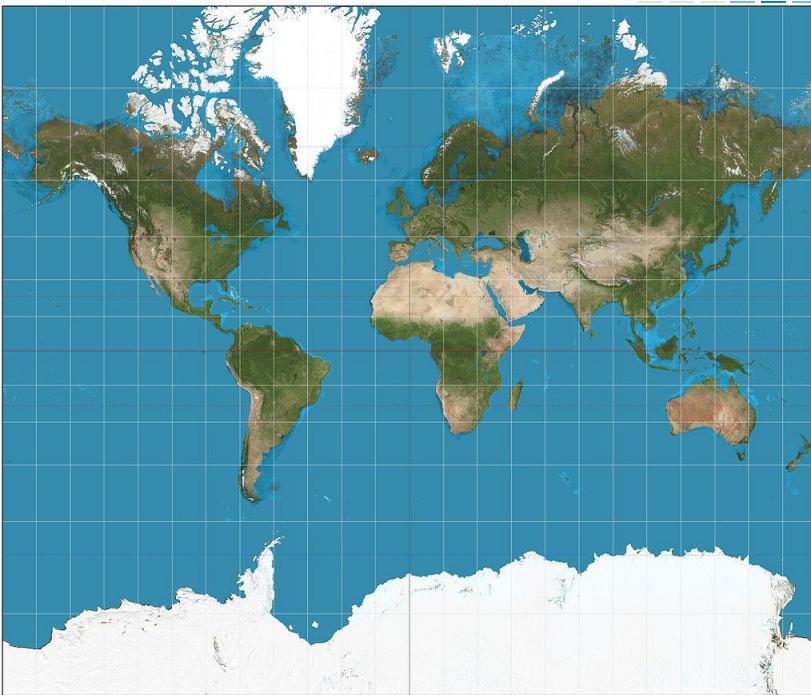
## ASmallCollection Principal Components Analysis

# PCA – Principal Component Analysis



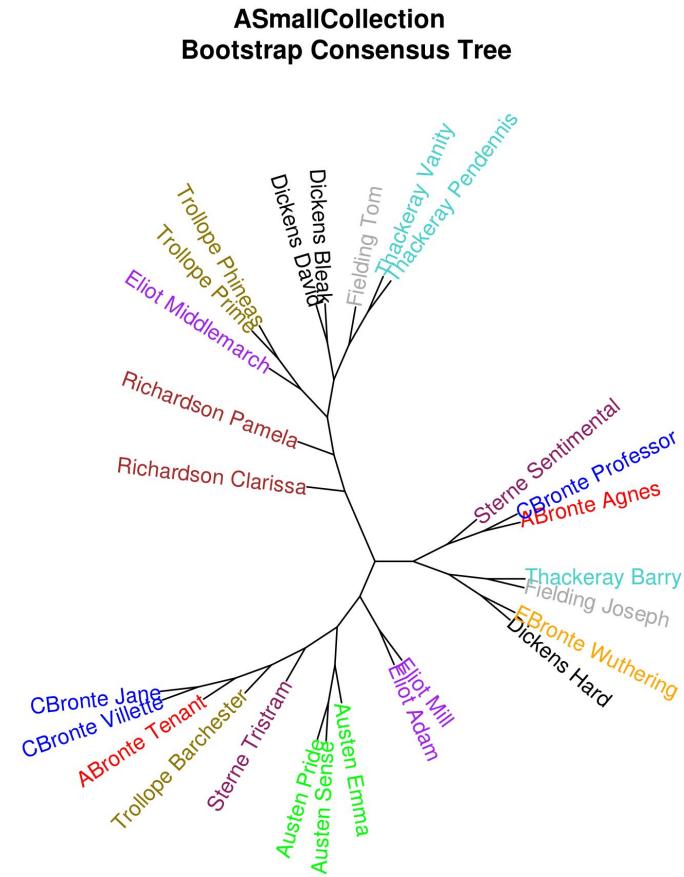
# Problems with dimension reduction

# Information loss

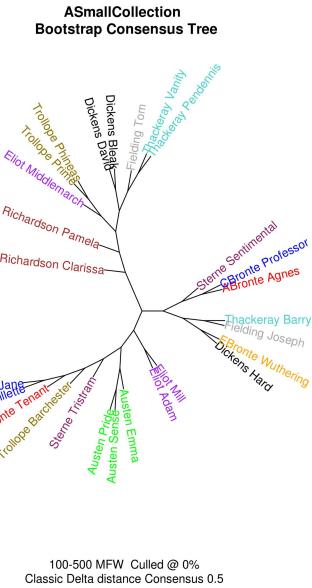
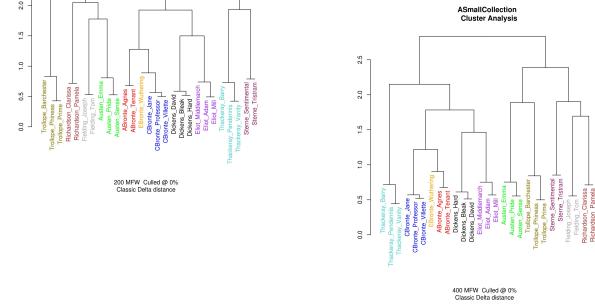
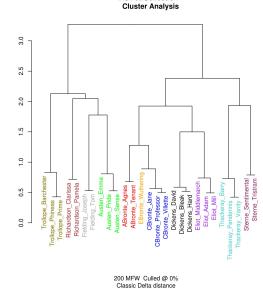
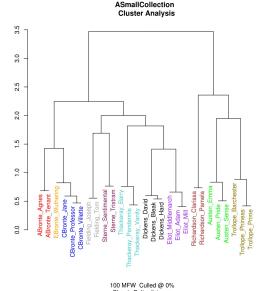


# Gold standard

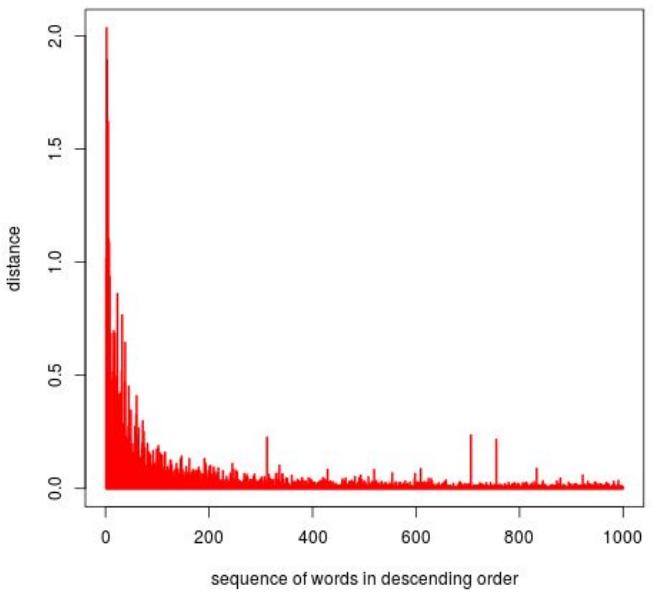
## Bootstrap consensus tree



# Bootstrap consensus tree - how?



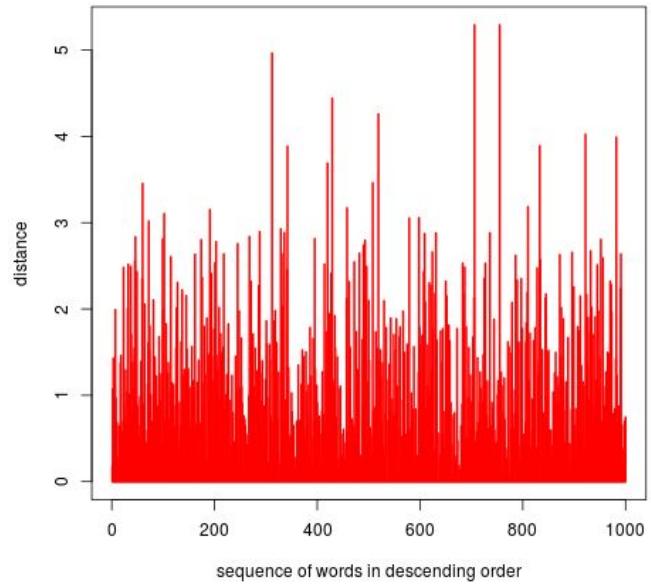
# Distance measure



Classic mathematical methods, e.g. Euclidean distance?

# Distance measure

## Classic Burrows distance!



For two texts  $T$  and  $T_1$ , and for a set of  $n$  words,

$$\Delta(T, T_1) = \frac{1}{n} \sum_{i=1}^n |z(f_i(T)) - z(f_i(T_1))|$$

$$\text{Where } z(f_x(T)) = \frac{f_x(T) - \mu_x}{\sigma_x};$$

$f_x(T)$  = raw frequency of word  $x$  in text  $T$ ;

$\mu_x$  = mean frequency of word  $x$  in a collection of texts;

$\sigma_x$  = standard deviation of frequency of word  $x$ .

# Classification algorithm

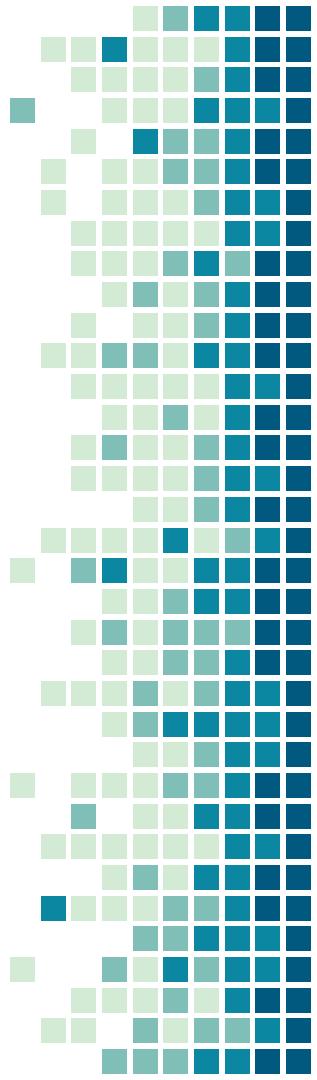
- k-nearest neighbors (v. good)
- Support Vector Machine (best)
- Nearest Shrunken Centroids
- Naive Bayes

## mini-deu Cluster Analysis



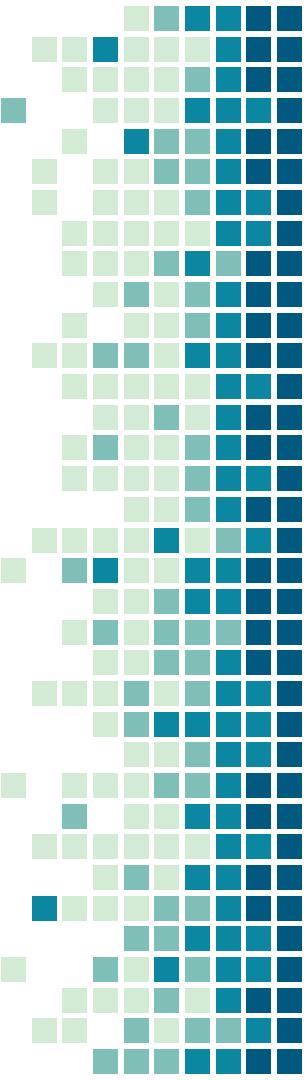
# Visualisation

Hierarchical  
clustering  
analysis  
(and many  
others)



2.0 1.5 1.0 0.5 0.0

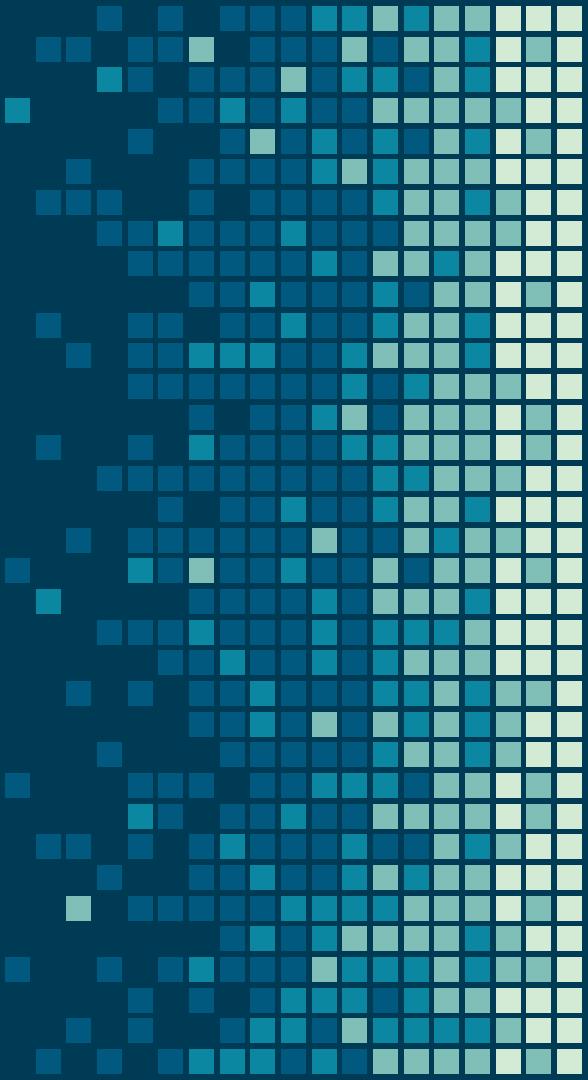
100 MFW Culled @ 0%  
Distance: wurzburg



# Applications of stylometry

- authorship attribution,
- tracing chronology,
- analysis of cross and inter genre relationships,
- big data analysis,
- style transfer and anonymization,
- ... and many others.

# Stylometry in exploration



# Making sense of a big American drama corpus

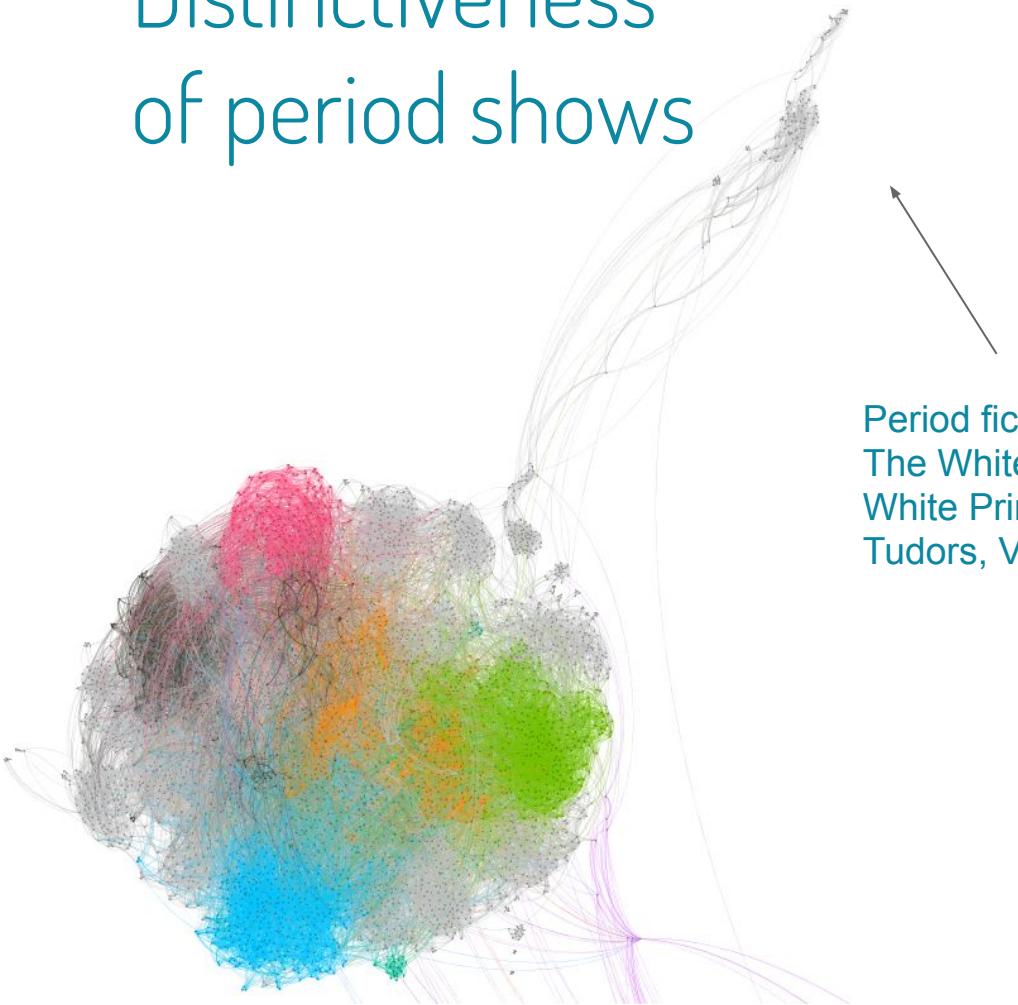
With Dennis Mischke  
(Potsdam U), Michał  
Choiński (Jagiellonian U)



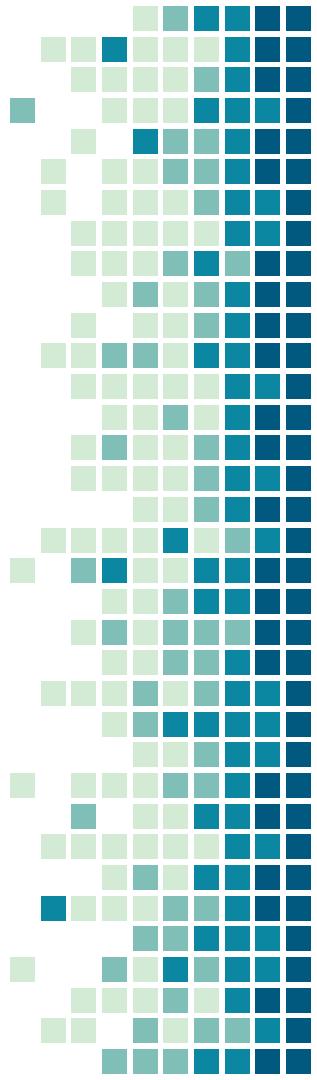
# Examining influences in TV series

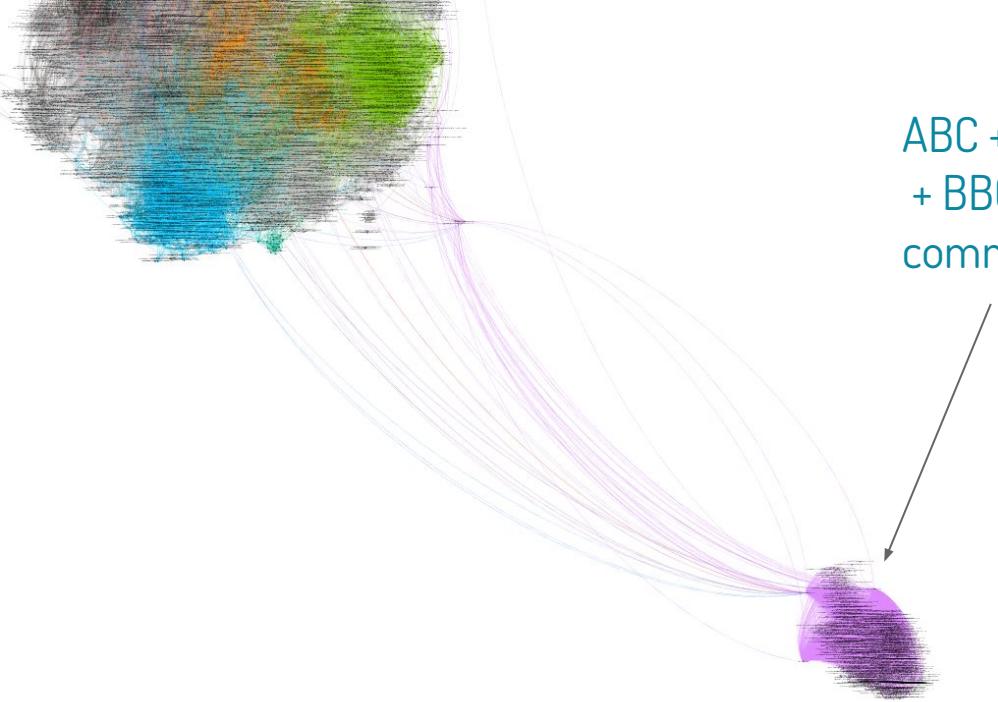
- 90 British and American shows
- Mix of genres
- Genre? Writer? Producer?

# Distinctiveness of period shows



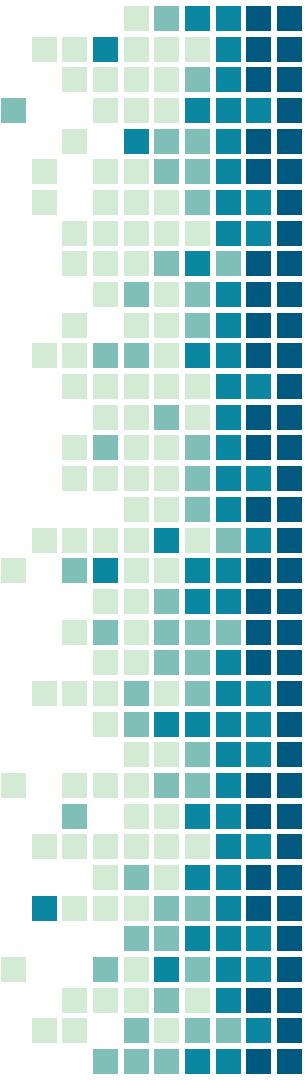
Period fiction:  
The White Queen, The  
White Princess, The  
Tudors, Victoria, etc.



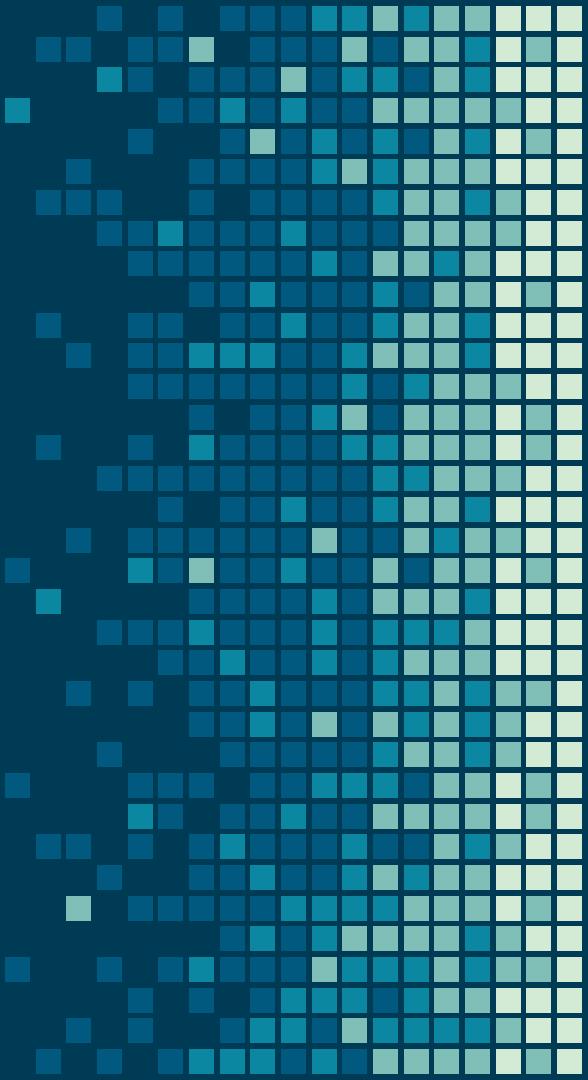


Production company?

ABC + a little Showtime  
+ BBC's Spooks, no  
common genre or topic

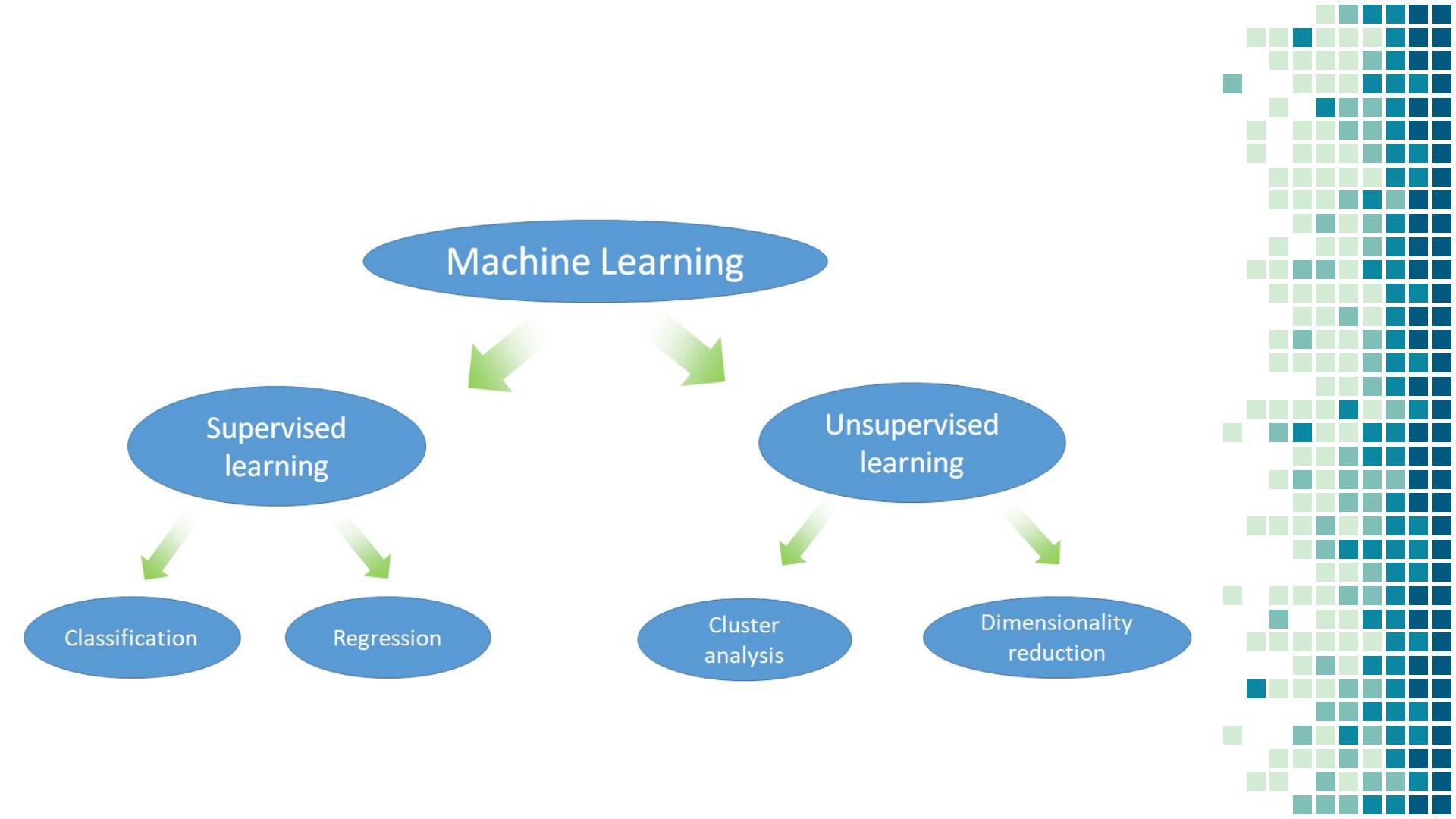


# Stylometry in authorship attribution



# Classification in Machine Learning





# Unsupervised classification



# Unsupervised learning

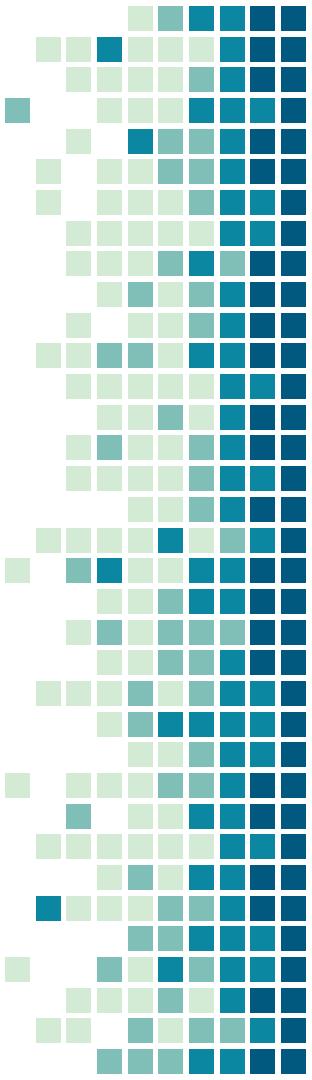
is a set of techniques that allow you to infer models to **extract knowledge of data sets where a priori is unknown.**

E.g.

- Cluster analysis
- Dimensionality reduction (PCA, MDS)

# Supervised classification





# Supervised learning

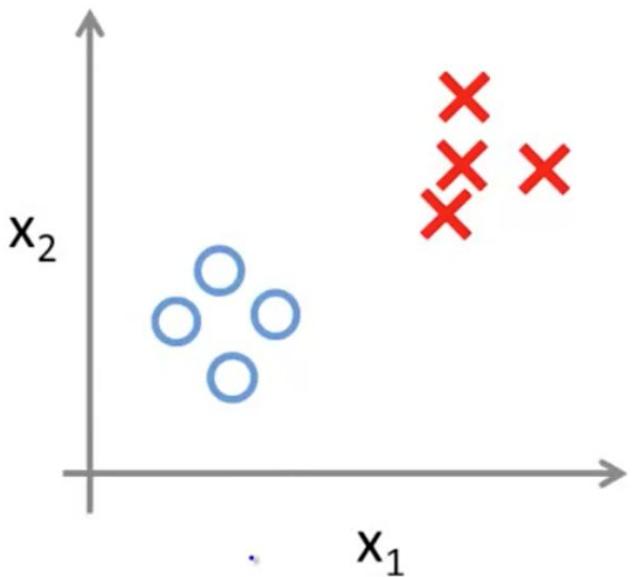
is a set of techniques that allows **future predictions** based on **behaviors or characteristics analyzed in historical data**.

E.g.

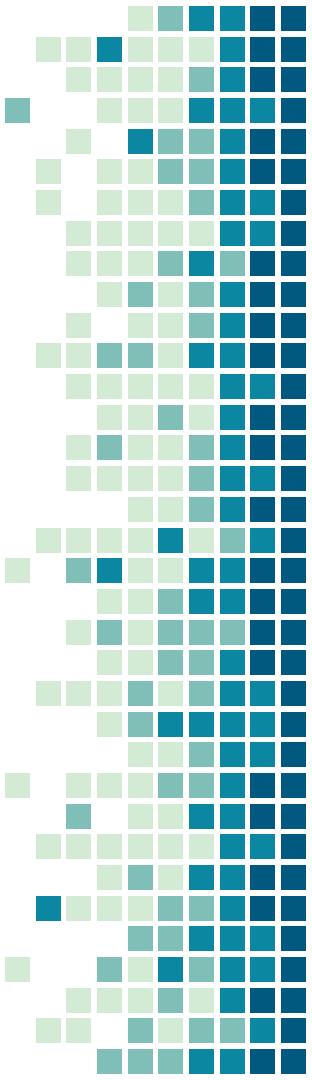
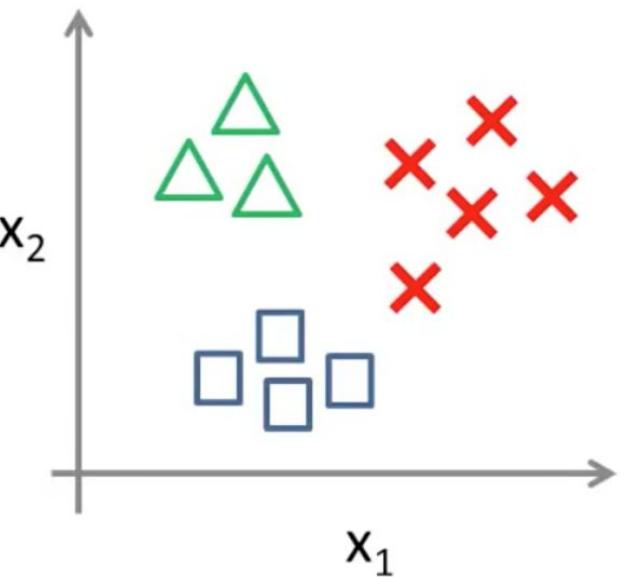
- Regression algorithms (linear regression, neural networks)
- Classification algorithms (logistic regression, Naive Bayes, Support Vector Machines, Random Forest)

# Types of classification

Binary classification:

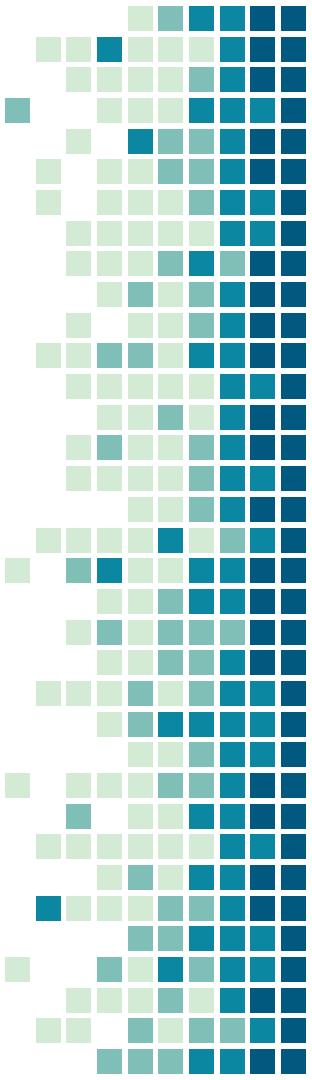


Multi-class classification:



# Cross-validation

- Verifying the quality by performing classification on a series of subsets of the main corpus
- E.g. "leave-one-out" = doing classification for all the cases of "corpus – one text", comparing results



# Attribution vs verification

## Attribution:

- Determining who of the known candidates from the closed set authored a given work

## Verification:

- Determining **IF** one of the known candidates from the open set authored a given work

# Classic authorship problems

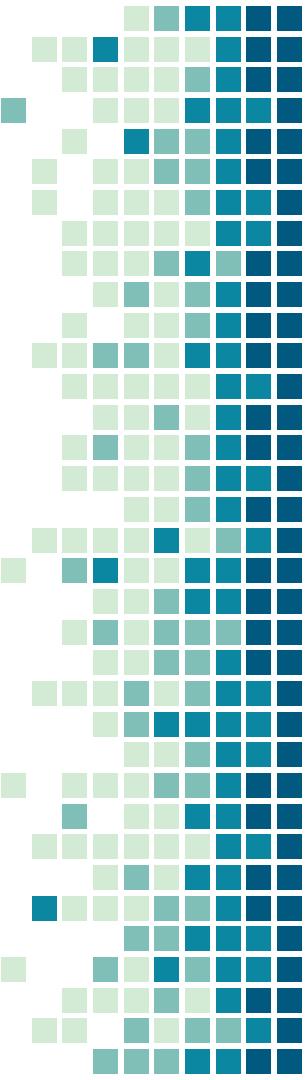
Federalist papers, JK Rowling



# Federalist Papers as an attribution case

- "A series of essays, anonymously published defending the document to the public"
- (Lin-Manuel Miranda 2015)
- 85 texts authored by: Alexander Hamilton (51?), James Madison (29?) and John Jay (5)
  - 12 letters of disputed authorship determined by stylometry

Mosteller, F., and D. L. Wallace (1964). Applied Bayesian and Classical Inference:  
The Case of The Federalist Papers. (and numerous other studies)



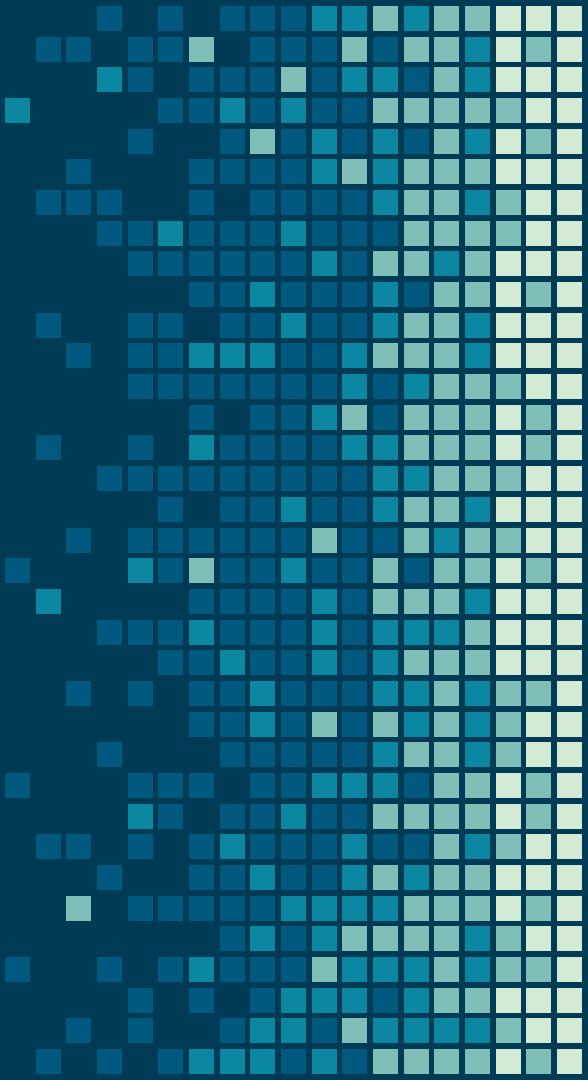
# JK Rowling or Robert Galbraith?

- Who wrote “The Cuckoo's Calling”?
- Study by Patrick Juola (2013)

“comparing against Rowling's own The Casual Vacancy, Ruth Rendell's The St. Zita Society, P.D. James' The Private Patient and Val McDermid's The Wire in the Blood.... Of the 11 sections of Cuckoo, six were closest (in distribution of word lengths) to Rowling, five to James.”

- Confirmed by the author

# Stylometry beyond text



# Measuring style in dance

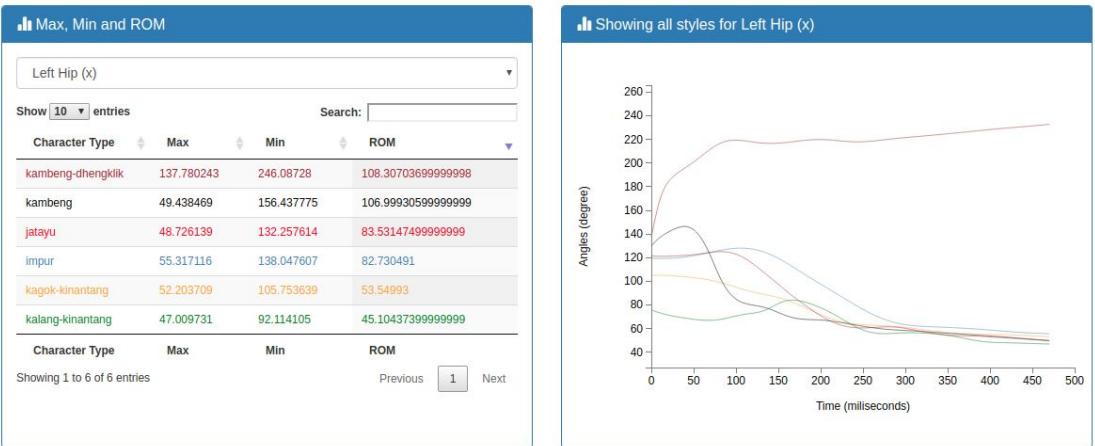
M. Escobar Varela and L. Hernández-Barraza. '[Digital Dance Scholarship: Biomechanics and culturally-situated dance analysis](#)' in Digital Scholarship in the Humanities (2019).

## Questions:

- use the biomechanical toolkit to address questions relevant to dance scholars
- identify the biomechanical markers of different character types for male dancers in the dramatic Sendratari form of Yogyakarta

# Measuring style in dance

Comparison Matrix (all styles for one joint)



The best discriminator of humble versus proud qualities is the left hip (on the x plane), where higher ROM correlates with a humble quality and a lower ROM correlates with a proud quality.

Escobar Varela, M., Hernández-Barraza, L. „Digital Dance Scholarship: Biomechanics and Culturally Situated Dance Analysis”. Digital Scholarship in the Humanities. <https://doi.org/10.1093/lhc/fqy083>.

Screenshot my own from: <https://villaorlado.github.io/dance/html/compareall.html>

# Stylometry of literary papyri

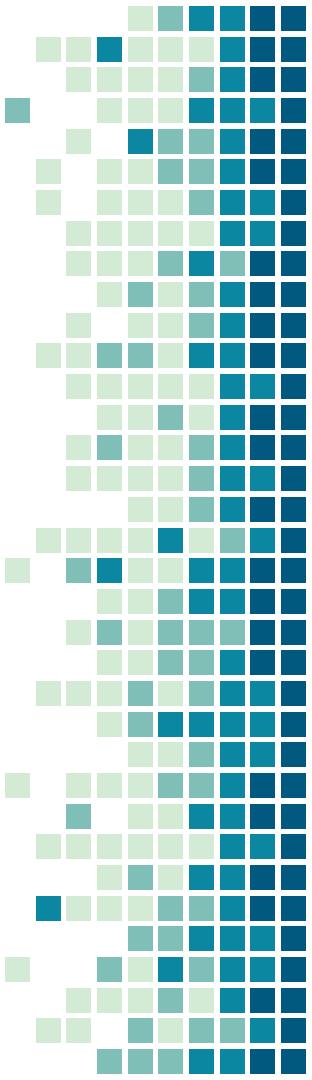
Ochab J.K., Essler H. **Stylometry of literary papyri**. In: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage - DATeCH2019. ACM Press; 2019. p. 139–42. Available from: <http://dl.acm.org/citation.cfm?doid=3322905.3322930>

## Goals:

- authorship attribution
- automatic genre classification

And for documentary papyri:

- automatic extraction of formulaic expressions
- automatic genre classification
- supplementation of missing metadata
- enhancement of metadata and annotation



# Stylometry of papyri

Ochab J.K., Essler H. **Stylometry of literary papyri**. In: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage - DATeCH2019. ACM Press; 2019. p. 139–42. Available from: <http://dl.acm.org/citation.cfm?doid=3322905.3322930>

## Study:

- 298 texts from Digital Corpus of Literary Papyrology (DCLP).
- The metadata from the Leuven Database of Ancient Books (LDAB)
- 66 authors

# Stylometry of papyri

Ochab J.K., Essler H. **Stylometry of literary papyri**. In: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage - DATeCH2019. ACM Press; 2019. p. 139–42. Available from: <http://dl.acm.org/citation.cfm?doid=3322905.3322930>

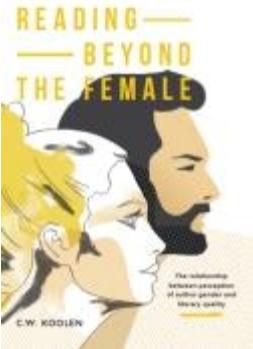
## Findings:

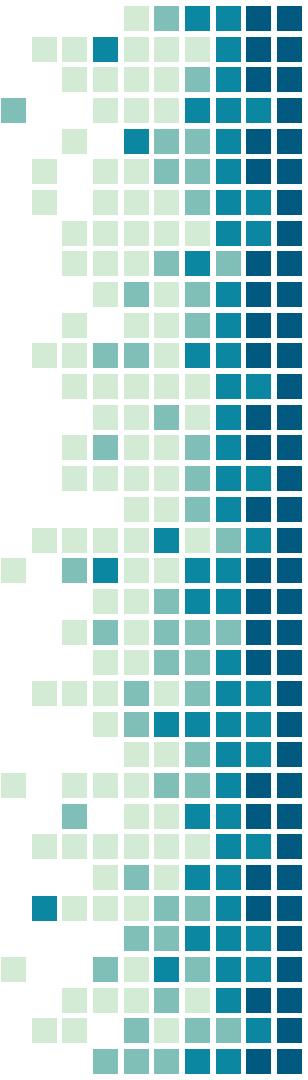
- successful classification correlating text regularization (scribes' impact?)
- good chances of automatic genre distinction

# Stylometry of literary quality

C.W. Koolen – Reading Beyond The Female

What is the relation between gender and perceived literary quality?





# And many more...

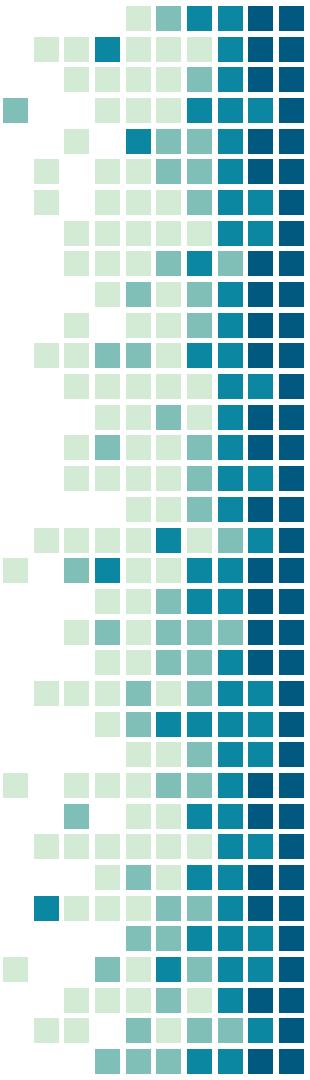
- Comic books stylometry (Alexander Dunst)
- Music stylometry (e.g. Andrew Brinkman)
- Cinemetrics
- Multimodal stylometry?

Let's try to do that!

[https://computationalstylistics.github.io/stylo\\_nutshell/#main-functions-stylo](https://computationalstylistics.github.io/stylo_nutshell/#main-functions-stylo)

# Getting started





Set working directory:

Command line:

```
setwd("the/path/to/my/favourite/folder")
```

RStudio users: find your directory in the Files panel,  
then use *Menu > More > Set as Working Directory*

Windows users: use *Menu > File > Change directory*

NEXT:

Type: `library(stylo)`

And then:

`stylo()`