

# **STYLOMETRY**

Graz, 11 July 2024

Summer School “Computational Language  
Technologies for Medievalists”

# INTRODUCTION

---

# Stylometry and the Latin Middle Ages

- Latin literature, computer science, and medieval cultural history
  - Predoc
    - Collaborative Authorship in Twelfth-Century Latin Literature. A Stylometric Approach to Gender, Synergy and Authority (2015 - 2020).
  - Postdoc
    - Cross-Channel Stylistic Exchanges. A Stylometric Approach to the Impact of Mobility and Multilingualism on Medieval Latin Literature, 1000–1150 (2020 - 2023).
    - Stylometry for Aurality in Medieval Anglo-Latin Hagiography (900–1150): Performance, Community and Musicality (2023 - 2026).

# Stylometry and the Latin Middle Ages

- Method and challenges:
  - J. De Gussem, « Computational Stylistics and Medieval Texts », Routledge Resources Online - Medieval Studies (2023). DOI: 10.4324/9780415791182-RMEO391-1.

# HISTORY

---

# History

IND & BRAIN

## How a Computer Program Helped Show J.K. Rowling write A Cuckoo's Calling

Author of the *Harry Potter* books has a distinct linguistic signature

By Patrick Juola on August 20, 2013

JUOLA & ASSOCIATES, YOUR TEXT ANALYSIS EXPERTS.

### Specializing in disputed authorship of documents for court cases.

The age of seven. I'm becoming hateful towards all humans in general. Only because it seems so easy for people to get along and have empathy. Empathy? Only because it looks and feel sorry for people to witness yours. ~~Thank you all from the pit of my burning noseholes~~ Sincere for your letters and concern during the past years. For too much of an erratic, moody, baby. I don't have the ~~the~~ passion anymore andse remember: it's better to burn out than to fade away. peace love, Emptiness. Carl robin

Frances and Courtney, I'll beat your sister.

Please keep going Courtney!

for Frances  
for her life which will be so much happier  
without me... I LOVE YOU I LOVE YOU!!

Who is the real Italian novelist writing as Elena Ferrante?



Auteur Domenico Starnone. Beeld BELGAIMAGE

### Who Is Behind QAnon? Linguistic Detectives Find Fingerprints

Using machine learning, separate teams of computer scientists identified the same two men as likely authors of messages that fueled the viral movement.



Florian Cafiero and Jean-Baptiste Camps. *Affaires de style. Du cas Molière à l'affaire Grégory: la stylométrie mène l'enquête* (Paris: Le Robert 2022).

Jeroen De Gussem

# History

- So far no clear, all-encompassing historiography of the method
- National(istic) traditions
- The cradle is in the 19th century, but the earliest ‘stylometrists’ (avant la lettre) seem to have worked independently of each other
- ‘Protostyliometry’: quantitative descriptions of texts
  - Grammarians of Sanskrit in ancient India, such as Pāṇini (difficult to date, 6th/4th century BCE)
  - Alexandrians: 3rd century BCE–3rd century CE
  - Masoretes and the Old Testament: 6th-10th century CE
  - Cathedral schools and early universities: 11th-12th century CE
  - ...

# History

- **Lorenzo Valla (1407–1457)**
  - *De falso credita et ementita Constantini donatione*, §43, §48.
  - Exposed the donation charter of Constantine as a forgery
- “[...] Shut up the gross and monstrous braying of this ass. He is so enchanted by the sound of turgid vocabulary that he repeats the same things and regurgitates what he has already said. [...]”
- “But why do I look for any discretion or any learning in you, who are endowed with no talent and no literary taste? You say lights when you mean lamps, and transferred for the eastern territories when you mean transferred to the eastern territories.”



# History

**Leon Battista Alberti (1404–1472)**

*De componendis cyfris*

## Cryptography

“**From my calculations**, it turns out that in the case of poetry, the number of consonants exceeds the number of vowels by no more than an octave, while in the case of prose the consonants do not usually exceed the vowels by a ratio greater than a sesquialtera. If in fact we add up all the vowels on a page, let’s say there are 300, the overall sum of the consonants will be 400.”



LEON BATTISTA ALBERTI ARCHIT.

*J. Dijonii T.I.*

FIORENTINO

*J. Battista Sc. 48.*

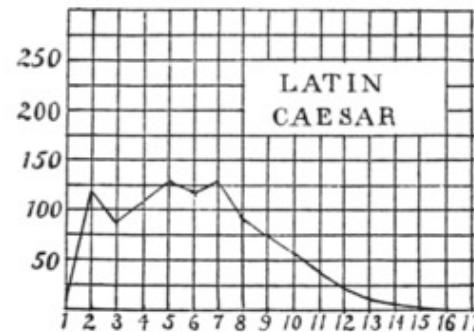
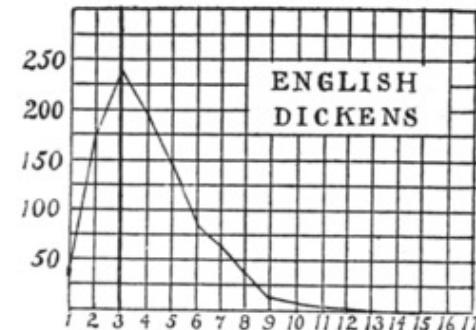
# History

- Wincenty Lutosławski (1863–1954)
  - the **stylème** to chronologically order Plato's texts.
- Thomas Mendenhall (1841–1924)
  - first to publish a stylometry paper applied to Charles Dickens in *Science* (1887), later (in 1901), he used “a simple counting machine.”

## SCIENCE.—SUPPLEMENT.

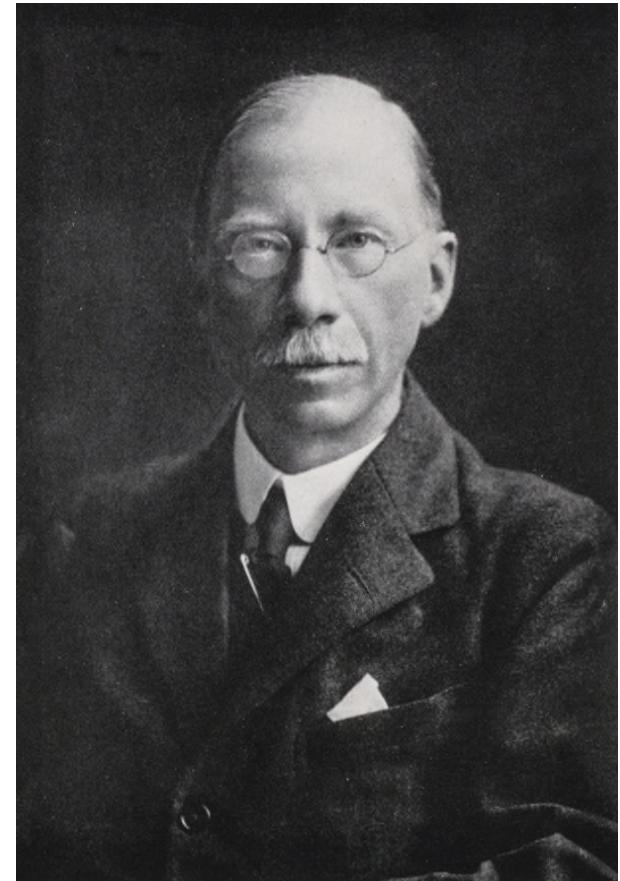
FRIDAY, MARCH 11, 1887.

mean word-length suggested itself. The new



# History

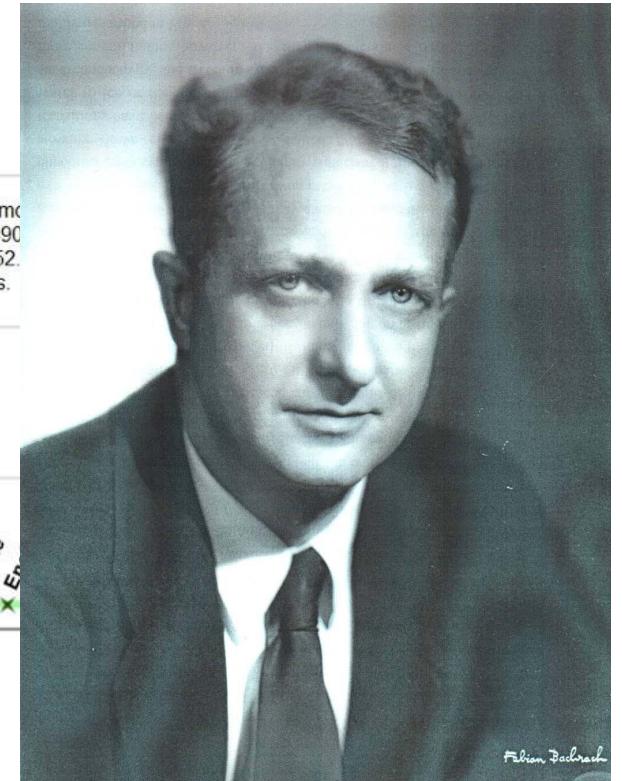
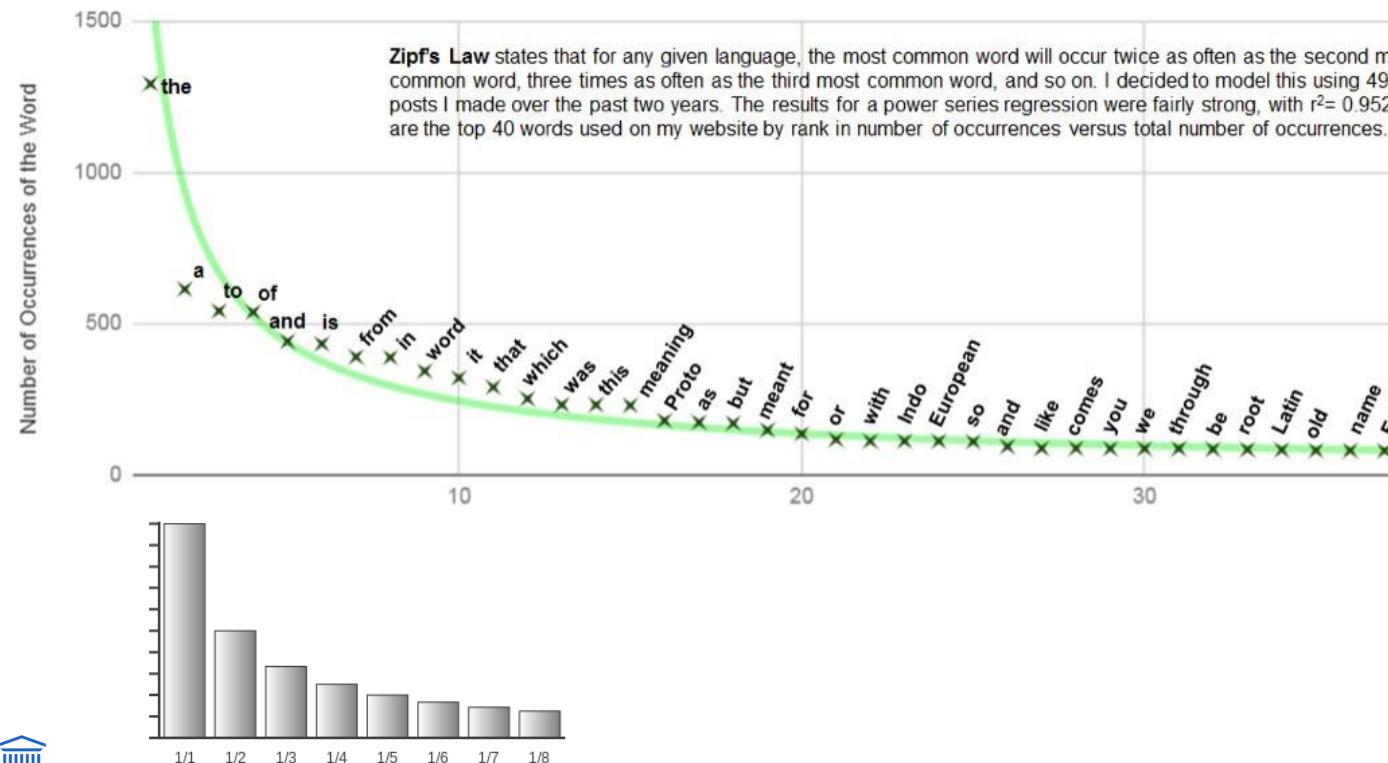
- Sentence length (1939)
  - Thomas à Kempis, *De imitatione Christi*
- Yule's 'Characteristic  $k$ ' (1944)
  - Lexical richness
  - How dense or extensive an author's vocabulary is



G. Udny **Yule**  
1871–1951  
Yule's  $K$ .

Jeroen De Gussem

# History



George Kingsley Zipf  
(1902–1950)  
Zipf's Law.

# History

- “One class of words that we use has been called function words—the filler words of the language, such as a, an, by, to, and that. Generally they include prepositions, conjunctions, pronouns, and certain adverbs, adjectives, and auxiliary verbs. [...]
- We like the function words rather well because many of them are not much influenced by the context of the writing [...].”



The Wallaces [left] with the Mostellers [right] at the University of Chicago in 1973.

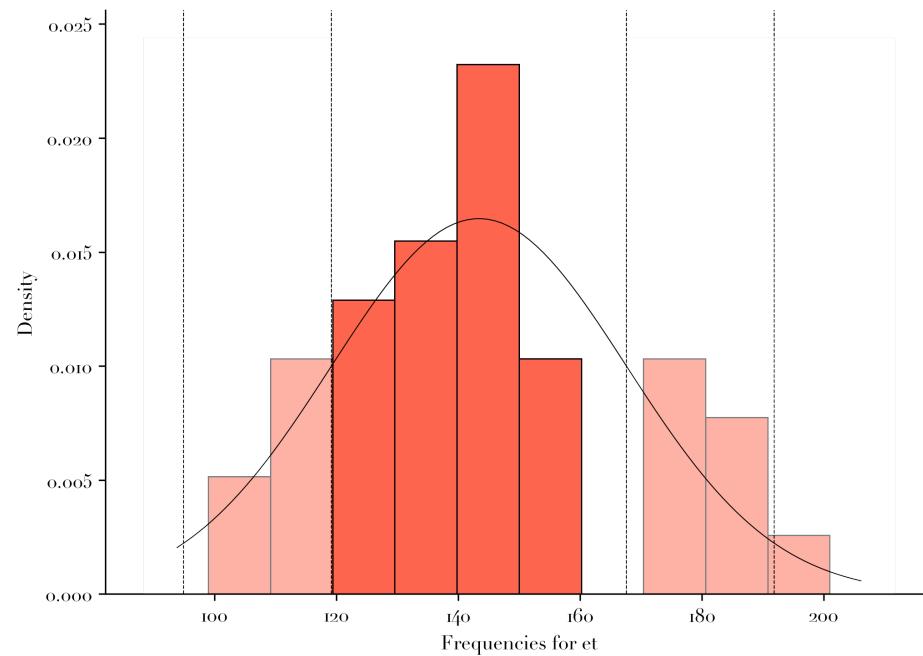


Example of IBM 7090.

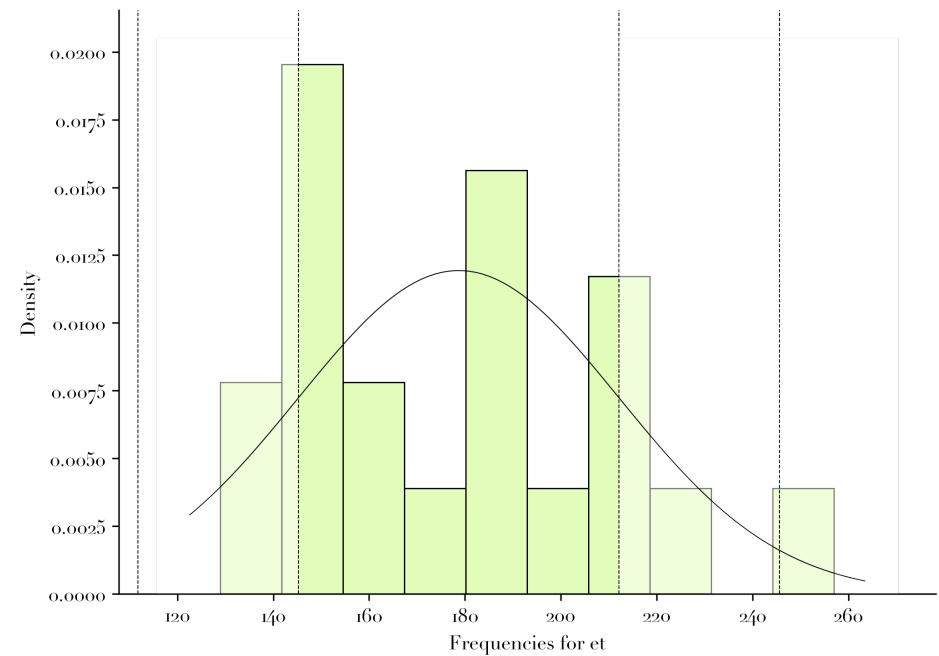
Jeroen De Gussem

# History

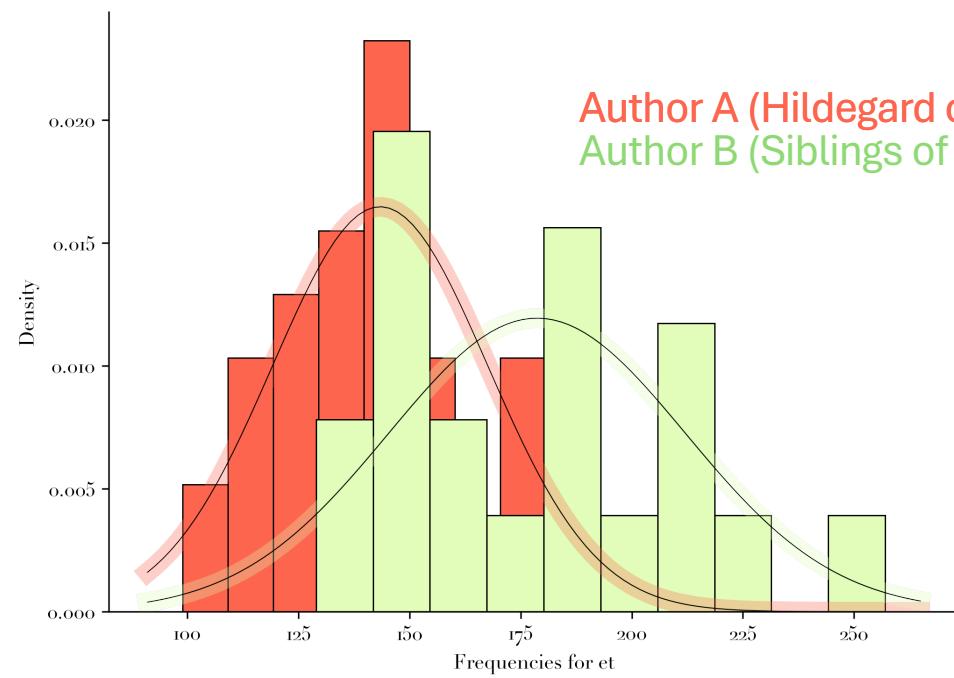
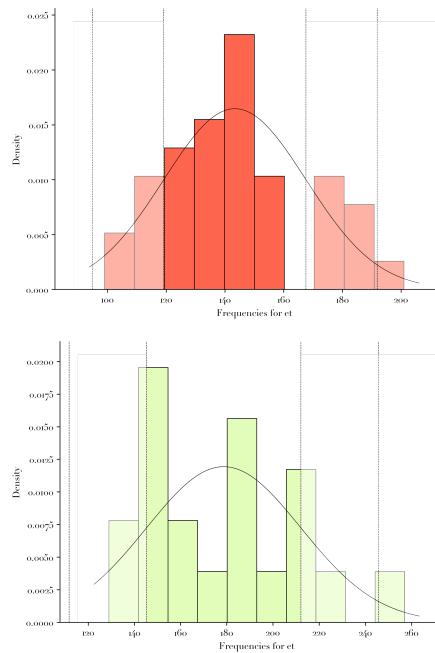
Author A (Hildegard of Bingen)



Author B (Siblings of Schönau)



# History



# **APPLICATIONS TO MEDIEVAL TEXTS?**

---

# Attribution and verification

- Starting as early as Yule and attribution of *De imitatione Christi* to Thomas of Kempen (Yule 1944).
- Single-unitary authorship of *Beowulf* (Neidorf, Krieger, et al. 2019).
- Attribution of *Guillaume d'Angleterre* to Chrétien de Troyes (Slautina 2009; Id. 2012).
- Dante Alighieri's letter to his patron Cangrande which is probably not entirely his (Downey et al. 2014).
- Disputed authorship of twelfth-century Latin *Epistolae duorum amantium* sometimes attributed to Abelard and Heloise (Cupiccia and Stella 2006).
- Authorship of thirteenth-century *Cronica Polonorum* by Venetian author of the *Translatio Sancti Nicolai* (Eder 2015; Kabala 2020).

# Scribal profiles and MS transmission

- Textual similarity between two redactions of the 13th-century Old Norse-Icelandic *Ljósvetninga* saga supports primacy of one redaction over the other (Macpherson and Tirosh 2020).
- Scribes abandon their individual spelling habits in favour of the ones from their exemplar, especially for words with initial line position in Middle English verse (Thaisen 2014).
- Gauge the stylistic impact of scribes on varying versions of Jacob of Maerlant's 13th-century *Scolastica* or Rhyming Bible (Van Dalen-Oskam 2012).

# Collaboration, apprenticeship, networks

- Investigation of impact of collaborator Guibert of Gembloux on Hildegard of Bingen's letter corpus and two suspect visions (Kestemont, Moens, and Deploige 2015).
- Stylistic impact of secretary Nicholas of Montiéramey on the sermons and letters of Bernard of Clairvaux (De Gussem 2017).
- Collaborative authorship of the *Vita Ædwardi regis* (Dockray-Miller, Drout, Kinkade and Valerio 2021).
- N-gram network analysis of recurrent formulas in Old English verse illustrate a single, densely connected social network (Battles 2019).

## Diatopic variation (~ dialectometry)

- Isolate the impact of local speech and writing habits in written Old French *chansons de geste* (Camps 2018).

## Stylochronometry

- Lutosławski's work on the chronological order of Plato's oeuvre (1898).
- 'Stylochronometry' to study the evolution of the copying practice of the individual scribes of the Carthusian Herne monastery in the Low Countries (Haverals and Kestemont 2020)

## Translatorship attribution

- Identifying the translator(s) of the West Saxon Gospels (Miranda-García, Calle-Martín and Marqués-Aguado 2008)
- Medieval Greek-Latin translations of Aristotle (Beullens, Haverals and Nagy, 2024)

# METHODOLOGY

## Using Jupyter notebooks



## NOTEBOOK 1

# 1.1-1.2 Preprocessing

- Fundamental for reliable stylometric analyses, especially crucial for medieval texts:
  - Medieval texts vary greatly, reflecting different editing principles, recensions, writing conventions, ... sometimes redundant for authorship and literary style analysis.
- Preprocessing Steps:
  - Removal and editing of irrelevant characters: punctuation, numerals, OCR errors, case-folding, titles, annotations, etc.
  - Tokenizing text into meaningful units, often word tokens.
  - Normalization/Standardization: Align orthographical and editorial conventions.
  - Disambiguation: Semantically distinguish homographs.
  - Stemming: Recover base stem or morphological root of word tokens.
  - Lemmatization: Transform word tokens to a standard dictionary form.
  - PoS Tagging and Parsing: Identify a token's part of speech and syntactic function.
  - Automated Scansion: Analyze prosodic units.

## NOTEBOOK 1

# 1.3 Sampling / Segmentation

- **Discrete:** Slices text into fixed segments, each following the previous one.
- **Rolling:** Uses a sliding window to create overlapping segments, providing a more sensitive scan of the text's stylistic profile.
- **Random:** Randomly selects sentences until a predefined sample length is reached, creating new combinations of the author's lexical distribution.
- **Generative:** Expands on random sampling by generating new text to imitate and extend the stylistic profile.

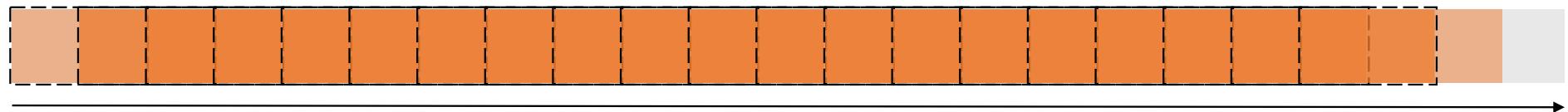
## NOTEBOOK 1

# 1.3 Discrete Sampling



## NOTEBOOK 1

# 1.3 Rolling Sampling



Progress of the text

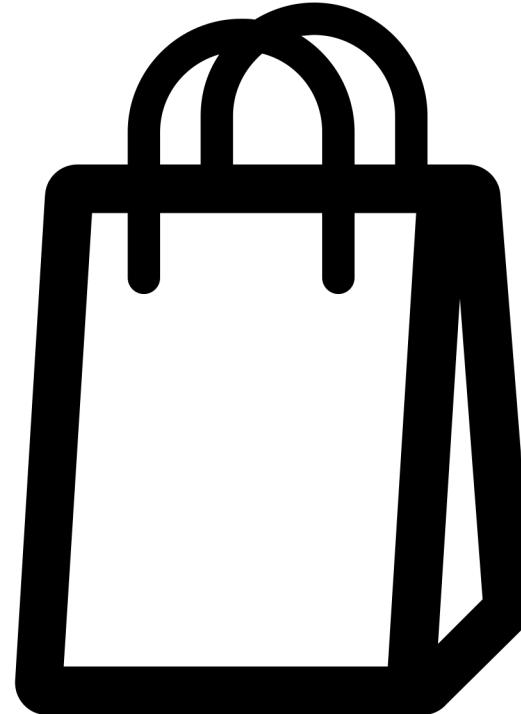
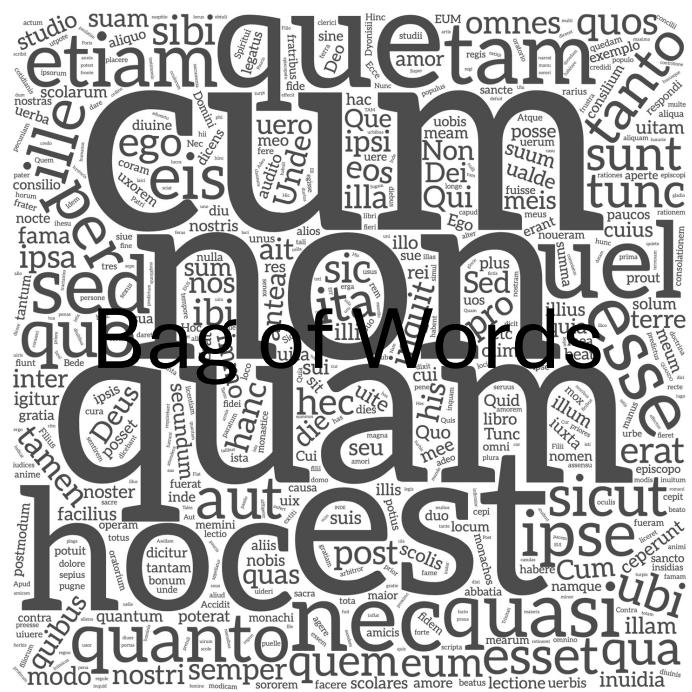
## NOTEBOOKS 1 & 2

# 1.4–1.5 / 2.1 Features

- Most frequent words (MFW)
- Function words, e.g. ['et,' 'in,' 'quoque,' 'ad,' ...]
- Character  $n$ -grams, e.g. ['-ur-', '-in-', '-et-', '-us-', ... ]
- Syntactic tags (PoS tags)
- Sentence length (?)
- Prosodic features, rhythm and metre (?)
- Semantic features (?): word embeddings (notebook 1)

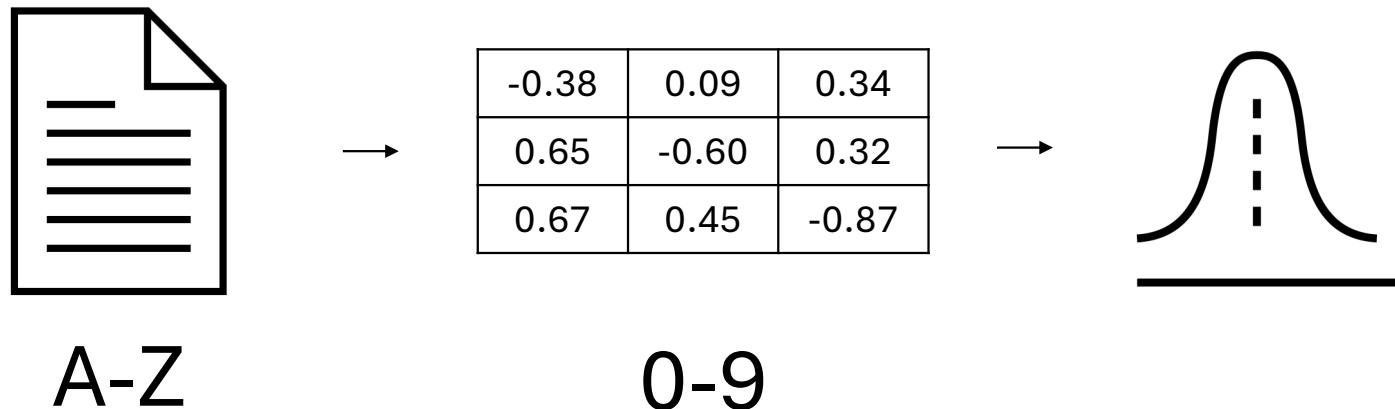
# NOTEBOOK 2

## 2.1 Feature extraction and vectorization



## NOTEBOOK 2

# 2.1 Feature extraction and vectorization



## NOTEBOOK 2

# 2.1 Feature extraction and vectorization

AUTHOR A			AUTHOR B		
	segment 1	segment 2	segment 3	segment 4	segment 5
et	116	114	131	130	123
in	156	119	134	130	88
est	52	37	45	48	31
ad	46	29	43	63	43
qui	19	11	12	22	14
cum	30	62	58	82	45
ut	31	13	27	36	30
me	33	22	30	33	34
non	20	35	30	29	32
e	41	45	23	32	22
					41

## NOTEBOOK 2

# 2.1 Feature extraction and vectorization

AUTHORA	
	segment 1
et	116
in	156
est	52
ad	46
qui	19
cum	30
ut	31
me	33
non	20
de	41



*difference  
distance*

AUTHORB	
	segment 4
	130
	130
	48
	63
	22
	82
	36
	33
	29
	32

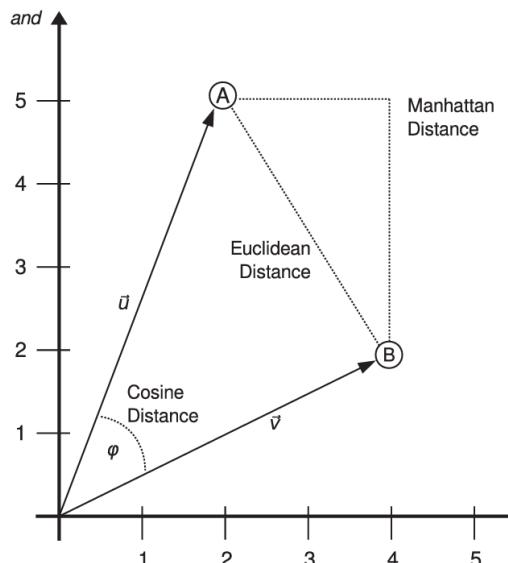
*n* ROWS  
*n* variables  
= features  
= dimensions

*n* COLUMNS  
*n* samples  
= observations  
= instances

Jeroen De Gussem

## NOTEBOOK 2

### 2.2.1 Distance Measures



- **Manhattan Distance:** Measures distance along axes at right angles, like navigating city blocks.
- **Euclidean Distance:** Straight-line distance, like "the crow flies."
- **Cosine Distance:** Measures the cosine of the angle between vectors.

**Fig. 2** Different vector distances between two example documents A and B illustrated in two-dimensional space

Evert, Stefan, et al. “Understanding and Explaining Delta Measures for Authorship Attribution.” *Digital Scholarship in the Humanities* 32.suppl\_2 (2017): ii4-ii16, at p. ii7.

# NOTEBOOK 2

## (2.2.2 Scaling)

In [13]:

```
from sklearn.preprocessing import StandardScaler
import numpy as np

np.set_printoptions(suppress=True, precision=2) # suppresses scientific notation

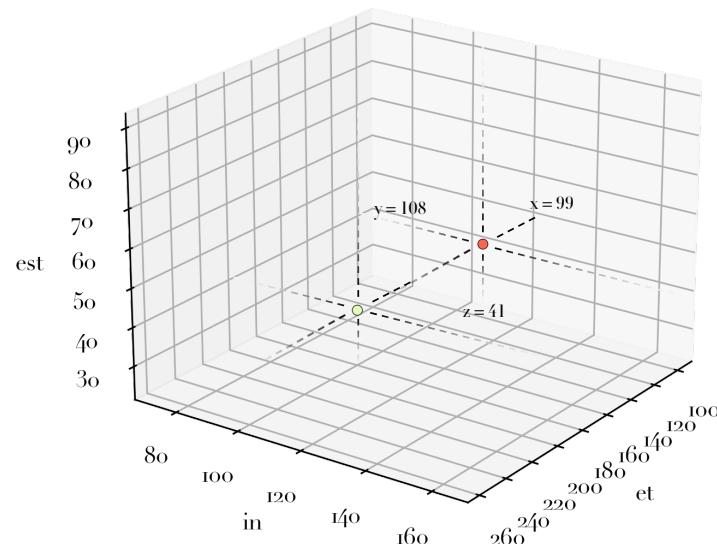
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

print(X_scaled)
```

```
[[ 1.73 -1. -0.71 ... -1.14  1.37 -0.36]
 [ 0.02 -0.44  1.28 ...  0.53 -0.6   0.43]
 [-0.93 -1.2   3.68 ... -0.62  0.48  0.35]
 ...
 [-0.65  1.14 -0.41 ... -0.05  0.31 -0.26]
 [-1.49  1.82 -0.85 ... -0.14  1.13 -0.36]
 [-0.61  1.14  0.27 ... -0.14  2.05  0.29]]
```

## NOTEBOOK 2

# 2.1 Feature extraction and vectorization (3.1 PCA)



difference distance

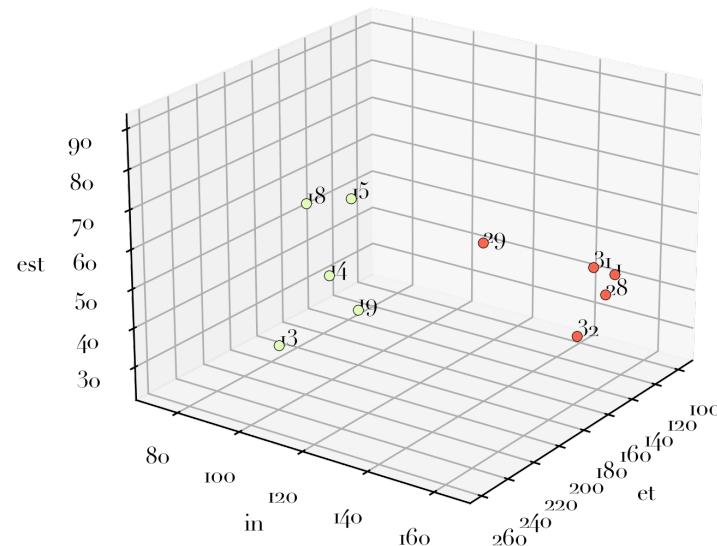
	AUTHORA segment 1	AUTHORB segment 4
et	116	130
in	156	130
est	52	48
ad	46	63
qui	19	22
cum	30	82
ut	31	36
me	33	33
non	20	29
de	41	32

n ROWS  
n variables  
= features  
= dimensions

n COLUMNS  
n samples  
= observations  
= instances

## NOTEBOOK 2

# 2.1 Feature extraction and vectorization (3.1 PCA)



AUTHORA	
	segment 1
et	116
in	156
est	52
ad	46
qui	19
cum	30
ut	31
me	33
non	20
de	41

*difference  
distance*

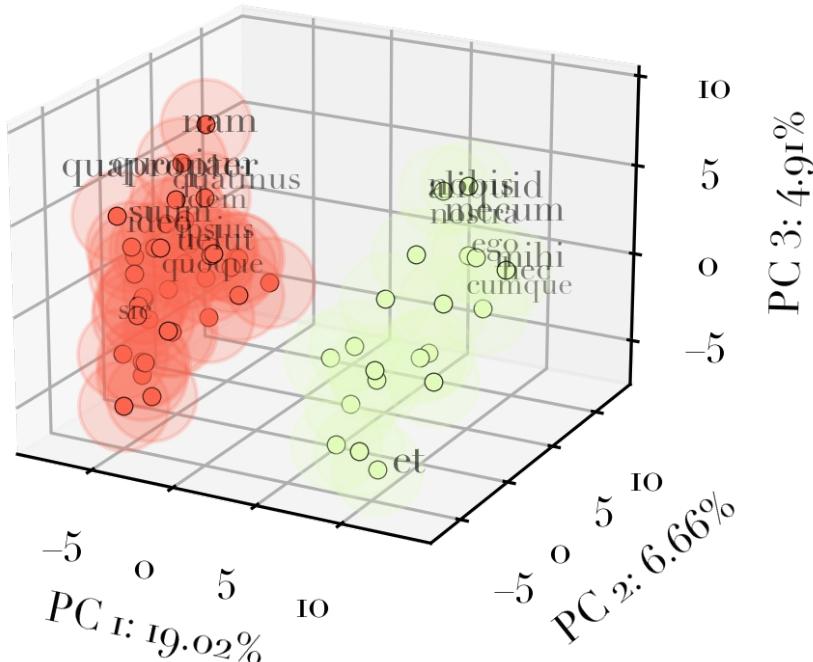
AUTHORB	
	segment 4
130	
130	
48	
63	
22	
82	
36	
33	
29	
32	

*n ROWS  
n variables  
= features  
= dimensions*

*n COLUMNS  
n samples  
= observations  
= instances*

## NOTEBOOK 3

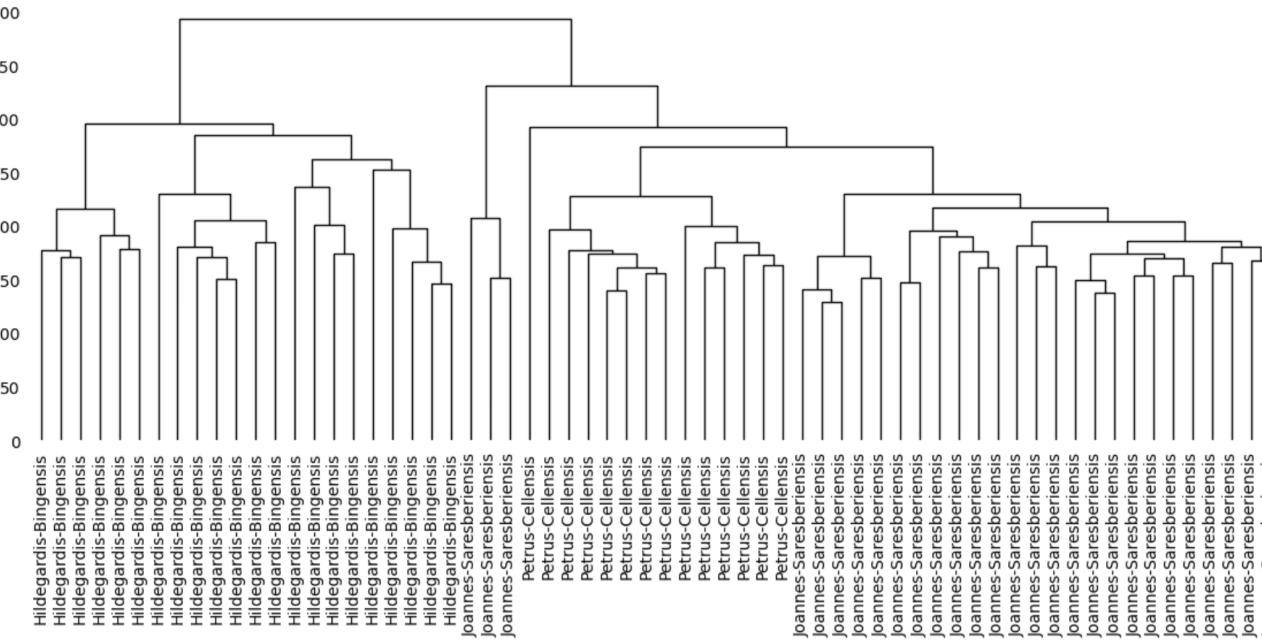
# 3.1. Principal Components Analysis (PCA)



	AUTHOR A			AUTHOR B		
	segment 1	segment 2	segment 3	segment 4	segment 5	segment 6
et	116	114	131	130	123	104
in	156	119	134	130	88	80
est	52	37	45	48	31	55
ad	46	29	43	63	43	28
qui	19	11	12	22	14	18
cum	30	62	58	82	45	66
ut	31	13	27	36	30	19
me	33	22	30	33	34	43
non	20	35	30	29	32	21
de	41	45	23	32	22	41

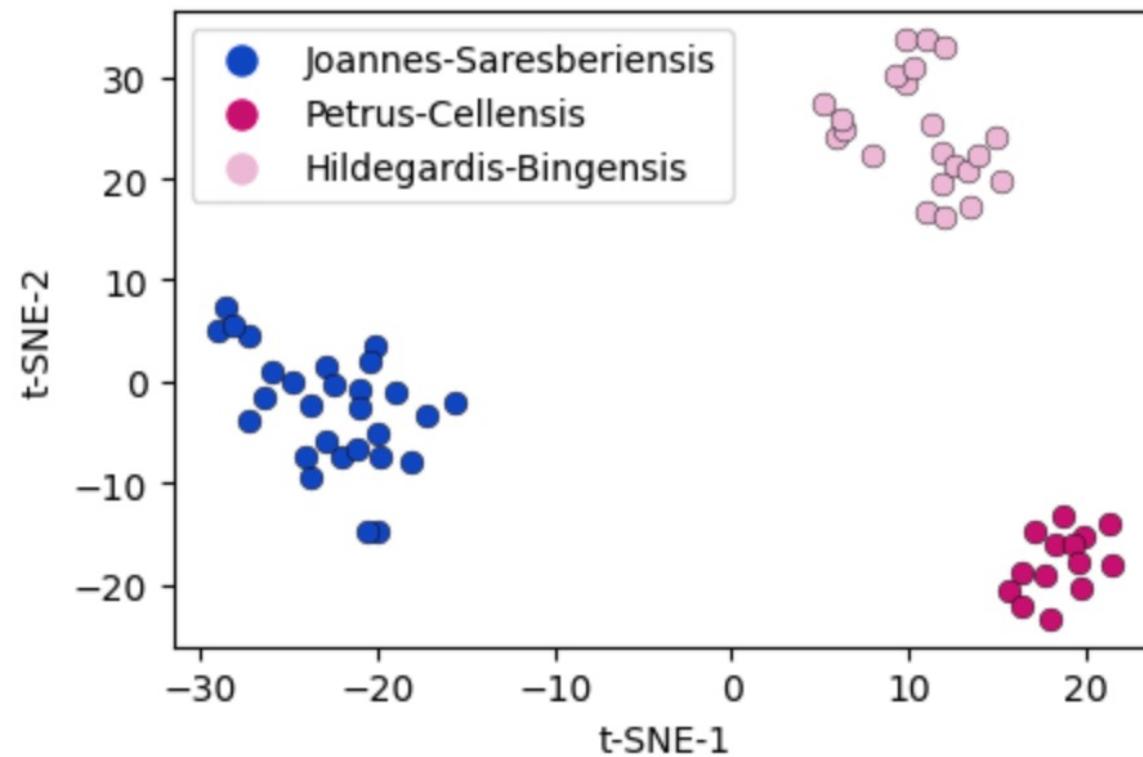
## NOTEBOOK 3

# 3.2 Dendrogram Analysis



# NOTEBOOK 3

## 3.3 T-SNE



# **SUGER OF SAINT-DENIS & THE DONATION CHARTER OF CHARLEMAGNE**

---

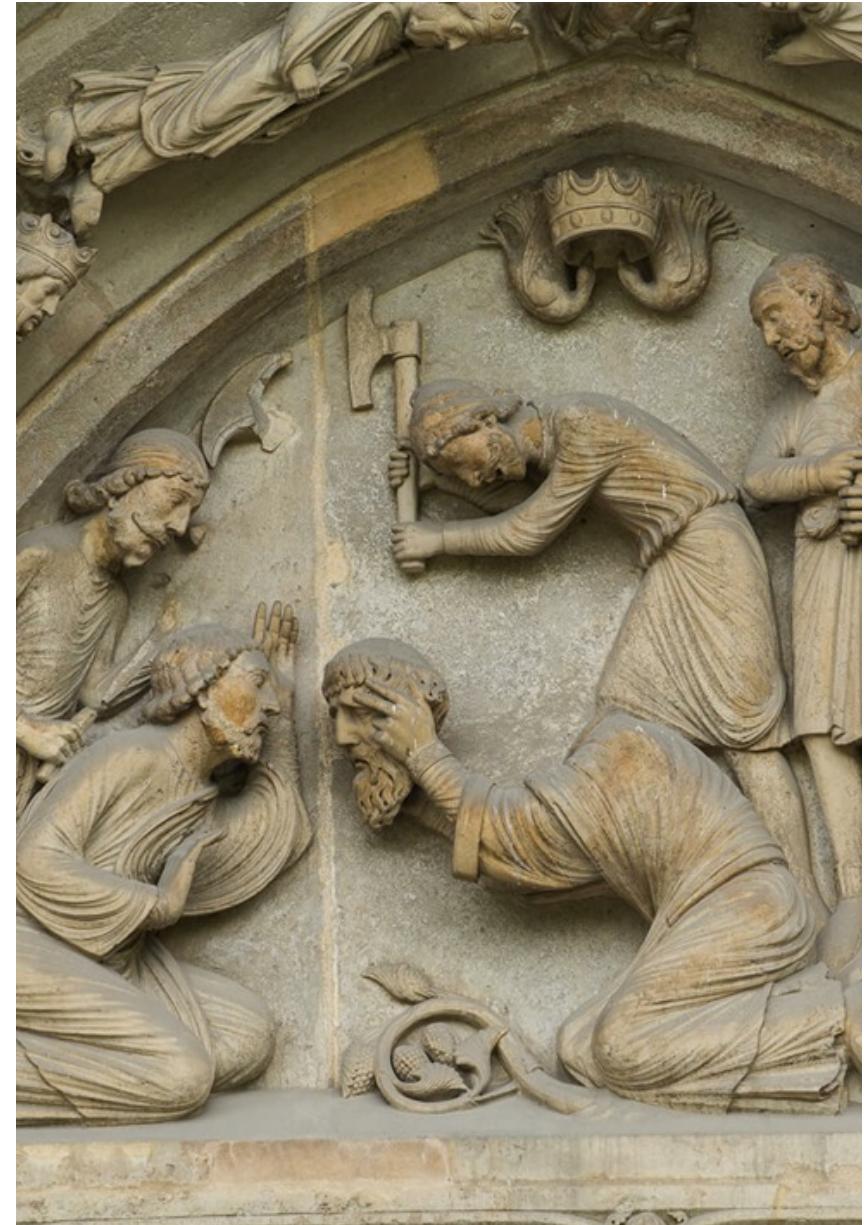
# Suger of Saint-Denis (1080/1–1151)

- Abbot of St-Denis from 1122 onward
- Counsellor and advisor to French Capetian kings Louis VI (1081–1137) and Louis VII (1120–1180).
  - Regency during the Second Crusade
- Royal diplomat — the Saint-Denis scriptorium stood in service / favour of the French monarchy



# Donation of Charlemagne

- Supposed 9th-century donation charter of Saint-Denis, allegedly written under the name of emperor Charlemagne.
  - Claims that Charlemagne granted Saint-Denis significant privileges, including the right to crown French kings and ordain bishops.
  - “head of all the kingdom’s churches” (*caput omnium ecclesiarum regni nostri*)
- Published by a monk of Saint-Denis, Jacques Doublet in 1625.
- Exact origins, date, and authorship of the document remain unclear, ranging from the 11th to the 17th century.



# Forgery in Saint-Denis

- Saint-Denis was notorious for producing forgeries and blending legend with reality.
  - The abbey's history of falsification and forgery attracted premodern scholars like Lorenzo Valla (1405/7–1457) and Jean Mabillon (1632–1707), who worked to distinguish authentic documents from fakes.
  - The St-Denis diplomas play a role in the famous discussions concerning document authentication between the Maurists and the Bollandists/Jesuits at the end of the 17th century / beginning of the 18th century.



DOM JEAN MABILLON,

R. Benedictin de la Congregaōn de S<sup>t</sup>. Maur

Né à S<sup>t</sup>. Pierre-mont, Diocèse de Reims, le 23. Novembre 1632.

Mort à Paris à l'Abb. S. Germain, le 27 Décemb. 1707. Agé de 76 ans.

# Charlemagne: Legends and Cult

- *Chansons de geste* around Charlemagne in the 12th century, and Charlemagne's canonization (1165)
- Saint-Denis claimed relics of the Passion (nails and crown of thorns) which Charlemagne supposedly brought from the East, enhancing its prestige and political influence. This connection with Charlemagne was used to assert spiritual sovereignty over other French churches, notably rivaling Reims.
  - *Descriptio qualiter* (late 11<sup>th</sup> century)
  - *Vita Karoli Magni* (1166)
  - *Pseudo-Turpin* (late 11<sup>th</sup>/early 12<sup>th</sup> c.) ~ direct verbal parallel; yet precedence to D Kar 286 unclear
  - *Pèlerinage de Charlemagne* (c. 1109–50)

# Three candidates?

- **Suger of Saint-Denis**
  - Has other forgeries on his name (expulsion of Argenteuil nuns)
  - Intimate knowledge of the chartrier
  - Wrote *vitae* for both French kings Louis VI and son Louis VII
- **Odo of Deuil**
  - Suger's disciple and successor
  - Abbot during C's canonization (1165)
  - Ties to French monarchy and wrote Crusade history in entourage of Louis VII
  - Morally inferior?
- **William of Saint-Denis**
  - Secretary and biographer of Suger
  - Excellent Latin writer (*Apologetic Dialogue*)

# Three candidates?

Author	Title (or <i>incipit</i> )	Ed.	Length
Suger of Saint-Denis	<i>De consecratione</i>	F. Gasparri	4,873 w
	<i>De rege Ludovico</i>	Ibid.	1,929 w
	<i>Epistolae</i>	Ibid.	6,615 w
	<i>Fragmentum vitae Ludovici</i>	Ibid.	1,405 w
	<i>Iunioris</i>		
	<i>Gesta Sugerii (De admin.)</i>	Ibid.	9,998 w
	<i>Vita Ludovici grossi</i>	H. Waquet	23,539 w
Odo of Deuil	<i>De profectione Ludovici VII in Orientem</i>	V. Berry	15,167 w
William of Saint-Denis	<i>Ad quosdam ex suis comonachis</i>	E.-R. Labande	884 w
	<i>De morte Sugerii</i>	A. Lecoy de La Marche	1,429 w
	<i>Dyalogum apologeticum</i>	A. Wilmart	12,435 w
	<i>Vita Sugerii</i>	F. Gasparri	5,399 w

Table 7.1: Corpus for candidate authors of D Kar 286: Suger of Saint-Denis, Odo of Deuil and William of Saint-Denis. Full bibliographical references are to be found in the primary sources given on pp. 377ff.

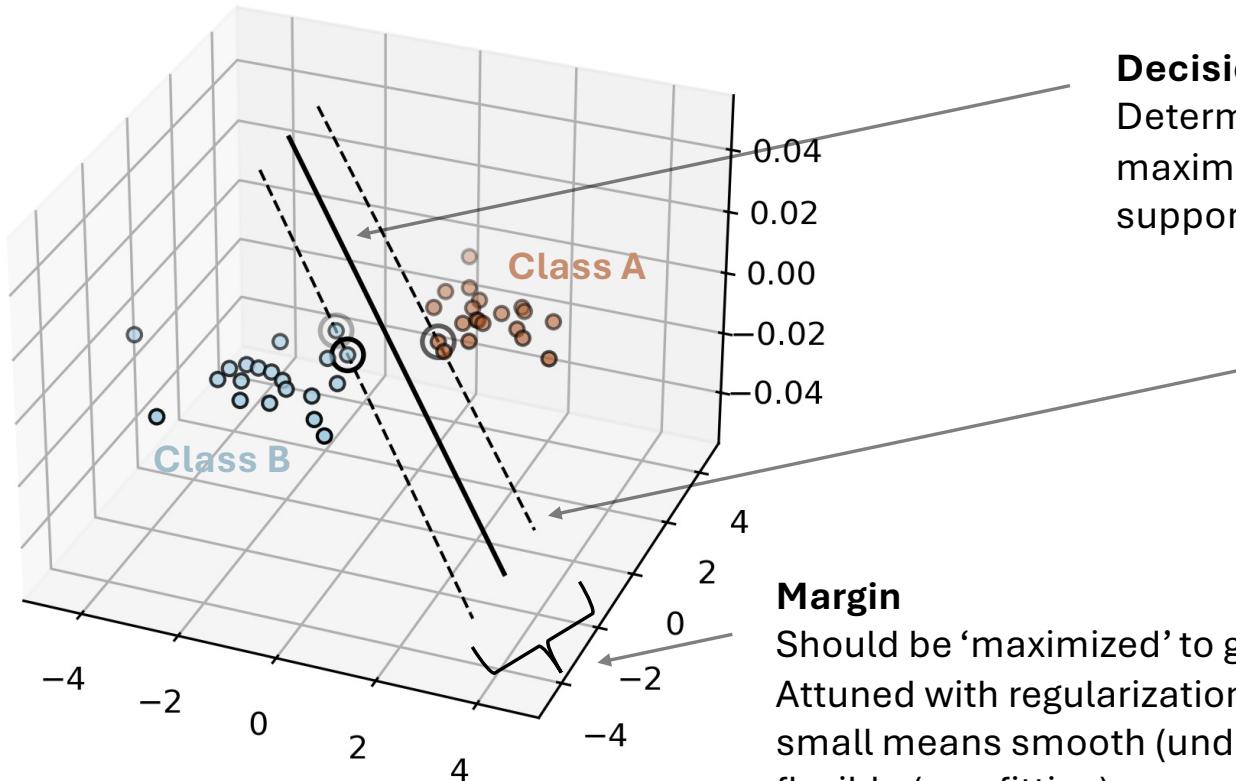
## NOTEBOOK 4

# 4. Supervised Classification

- Training data:
  - X: feature inputs, from text segments
  - y: class outputs, from 3 candidate authors
- During training process, model 'observes' correctly labelled X-y pairs and registers correlations between feature patterns (X) and class labels (y).
- After training, model can predict labels for new, unclassified texts.
- Can assign classes via hard decisions or probability scores.
- Parameter combinations (feature set, vector length, etc.) can be evaluated.
  - Enables fine-tuning and optimization of parameters.

## NOTEBOOK 4

# 4.1 SVM – Support Vector Machine



### Decision boundary

Determined by the hyperplane that maximizes the margin between the support vectors

### Support vectors

Constituted by the nearest data points from each class.

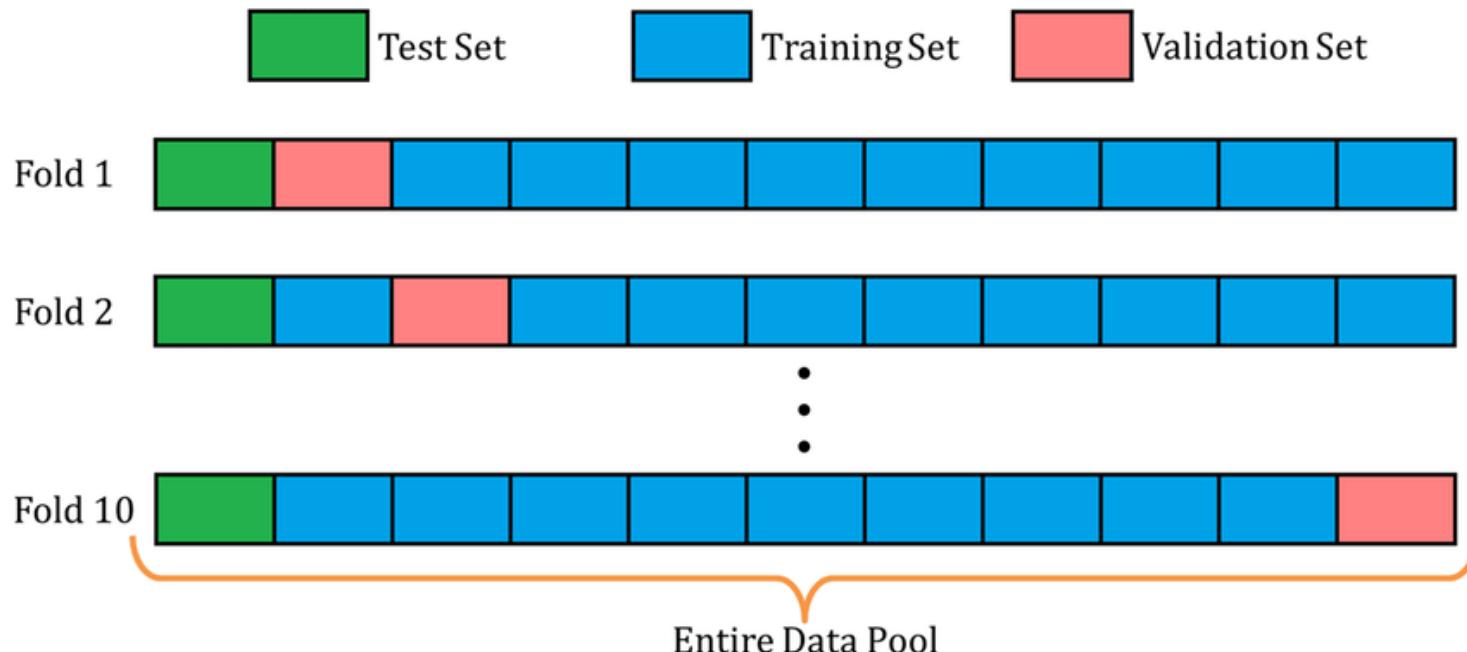
### Margin

Should be ‘maximized’ to guarantee separability. Attuned with regularization parameter **C**, where small means smooth (underfitting) and large means flexible (overfitting).

## NOTEBOOK 4

### 4.1.1 train\_test\_split (cross-validation)

4.1.1 Preparing the Dataset for Training → `train_test_split`



## NOTEBOOK 4

### 4.1.2 Evaluation

- **Accuracy** — *"How often did we correctly attribute the text segment to a given author?"*
  - Correct attributions / total predictions
- **Precision** — *"When we positively identified a text segment as written by a given author, how often was that true?"*
  - True positive identifications / total positive identifications
- **Recall** — *"Were we able to attribute all text segments of a given author to that author?"*
  - Correct positive identifications / total actual positives
- **F1-score** — Harmonic mean of precision and recall

## NOTEBOOK 4

### 4.1.2 Evaluation

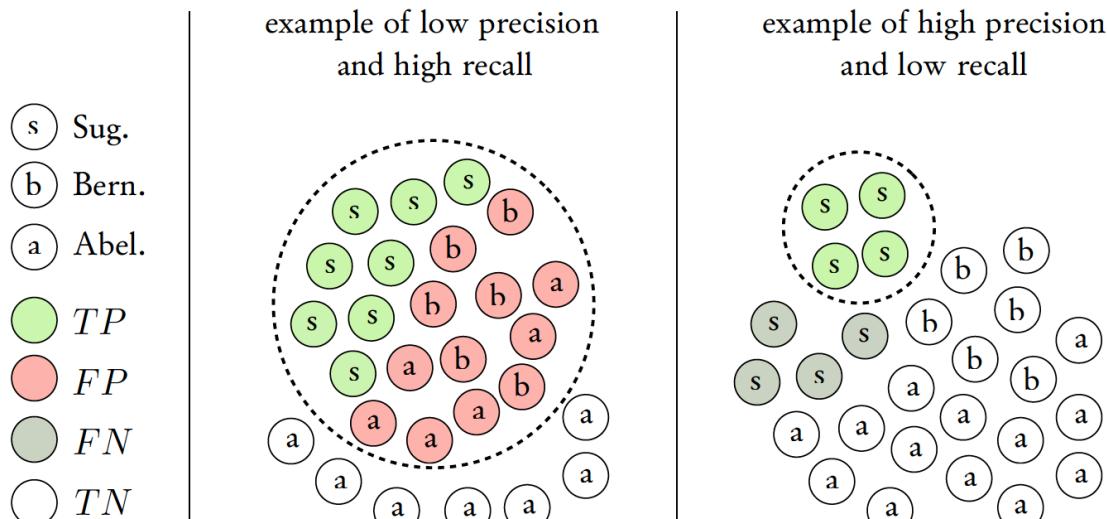


Figure A.13: Visual intuition of thresholding, i.e. what happens when the trade-off between precision and recall is imbalanced.

## NOTEBOOK 4

# SVM for Suger, Odo, and William

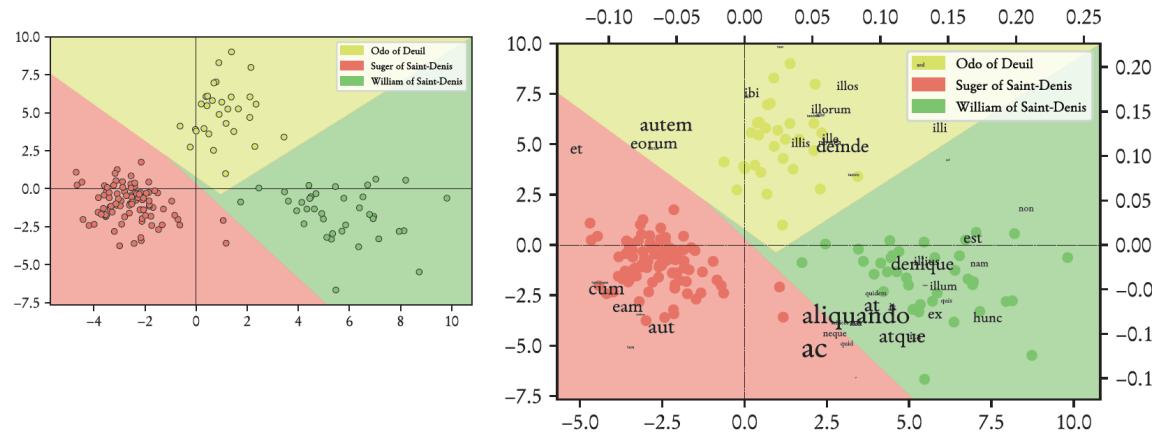


Figure 7.1: PCA plot giving a 2D intuition of the decision boundaries drawn by the best-scoring SVM classifier (whose settings can be consulted in table 7.2). The background colours indicate the respective ‘zones’ of each of the candidate authors Odo (yellow), Suger (red) and William (green). The right-hand figure shows the 50 (out of 250) loadings. Only the highest ranking loadings on both PC’s have been visualized (the font size expresses the weight). Settings:  $s-l = 500$  | *type* = most-frequent function words |  $n = 250$  (50 most important are visualized) | *vect.* = standard-scaled tfidf-weighted frequencies | *expl. var.* = 8.26%.

# Bibliography

- Battles, Paul. "Using N-gram Analysis to Map Intertextual Networks in Old English Verse." *Digital Philology: A Journal of Medieval Cultures* 8.2 (2019): 155-191.
- Cafiero, Florian, and Jean-Baptiste Camps. *Affaires de style. Du cas Molière à l'affaire Grégory: la stylométrie mène l'enquête*. Le Robert, 2022.
- Camps, Jean-Baptiste. "Manuscripts in Time and Space: Experiments in Scriptometrics on an Old French Corpus." *Corpus-Based Research in the Humanities CRH-2* (2018): 55-64.
- Cupiccia, Matilde. "Progressi nello studio del cursus: i metodi statistici e il caso di Eloisa e Abelardo." *Filologia Mediolatina* 5 (1998): 37-48.
- De Gussem, Jeroen. "Computational Stylistics and Medieval Texts." *Routledge Resources Online - Medieval Studies* (2023). DOI: 10.4324/9780415791182-RMEO391-1.
- De Gussem, Jeroen. "Bernard of Clairvaux and Nicholas of Montiéramey: Tracing the Secretarial Trail with Computational Stylistics." *Speculum* 92.S1 (2017): S190-S225.
- Dockray-Miller, Mary, et al. "The Author and the Authors of the 'Vita Ædwardi Regis': Women's Literary Culture and Digital Humanities." *Interfaces: A Journal of Medieval European Literatures* 8 (2021): 160-213.
- Downey, Sarah, et al. "Lexomic Analysis of Medieval Latin Texts." *The Journal of Medieval Latin* (2014): 225-275.
- Eder, Maciej. "In Search of the Author of *Chronica Polonorum* Ascribed to Gallus Anonymus: A Stylometric Reconnaissance." *Acta Poloniae Historica* 112 (2015).
- Kabala, Jakub. "Computational Authorship Attribution in Medieval Latin Corpora: The Case of the Monk of Lido (ca. 1101–08) and Gallus Anonymous (ca. 1113–17)." *Language Resources and Evaluation* 54.1 (2020): 25-56.
- Kestemont, Mike, Sara Moens, and Jeroen Deploige. "Collaborative Authorship in the Twelfth Century: A Stylometric Study of Hildegard of Bingen and Guibert of Gembloux." *Digital Scholarship in the Humanities* 30.2 (2015): 199-224.

- Lutoslawski, Wincenty. "Principes de stylométrie appliqués à la chronologie des œuvres de Platon." *Revue des études grecques* 11 (1898): 61-81.
- MacPherson, Michael, and Yoav Tirosh. "A Stylometric Analysis of Ljósvetninga saga." *Gripla* 31 (2020).
- Mosteller, Frederick, and David L. Wallace. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. 2nd ed., Springer Series in Statistics, Springer Verlag, 1984 (1964).
- Neidorf, Leonard, et al. "Large-Scale Quantitative Profiling of the Old English Verse Tradition." *Nature Human Behaviour* 3.6 (2019): 560-567.
- Slautina, Maria. "The Disputed Authorship of a Medieval Text: A New Solution to the Attribution Problem in the Case of Guillaume d'Angleterre." *Authorship* 2.1 (2012): 1–16.
- Slautina, Maria. *L'attribution du roman Guillaume d'Angleterre par la méthode de reconnaissance des formes*. Diss. Caen, 2009.
- Stella, Francesco. "Analisi informatiche del lessico e individuazione degli autori nelle "Epistolae duorum amantium" (XII secolo)." 2006.
- Thaisen, Jacob. "Initial Position in the Middle English Verse Line." *English Studies* 95.5 (2014): 500-513.
- Van Dalen-Oskam, Karina. "The Secret Life of Scribes. Exploring Fifteen Manuscripts of Jacob van Maerlant's *Scolastica* (1271)." *Literary and Linguistic Computing* 27.4 (2012): 355-372.
- Yule, G. Udny. *The Statistical Study of Literary Vocabulary*. Cambridge University Press, 1944.