



Classical and Digital Approaches to Greek Palaeography
PATMOS · GREECE 14–18 SEPTEMBER 2026

Machine Learning for Manuscript Classification

Elpida Perdiki, PhD – Democritus University of Thrace

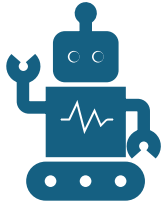




The scriptoria net

A word or two on terminology

Terms, terms, terms



Artificial intelligence

tasks of pattern recognition

associated with human intelligence



Machine learning (AI subset)

classification, prediction, or
recognition on larger datasets

Supervised: classified labelled data

Unsupervised: patterns for classification
on unlabelled data

Reinforcement: trial and error, learning
progress with rewards and penalties



Data mining

information extraction from big data

Data classification: supervised data
categorisation in classes

Data clustering: unsupervised data
organisation into clusters



What is Orange?

- _ University of Ljubljana
- _ Data mining
- _ Machine learning
- _ Python library
- _ GUI application

Quick disclaimer:

is it classification or clustering?

The John Chrysostom case

The **pop** culture of **John Chrysostom**

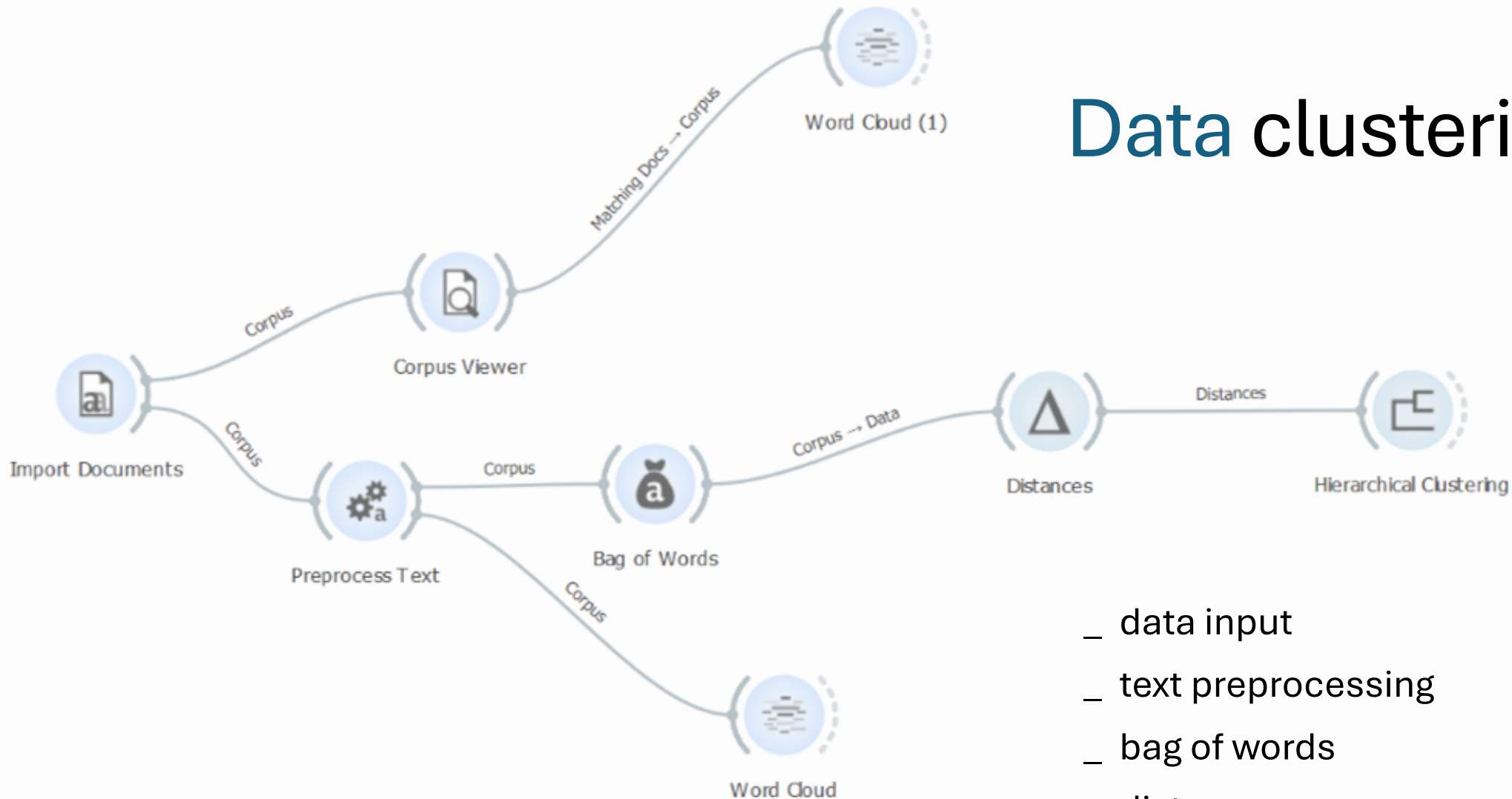
- Church Father, Archbishop of Constantinople, preacher
- **20,000** mss
- *double tradition*
 - short recension (older text, raw)
 - long recension (newer text, revised)



The **proposed** workflow

Homilies on the Letter to Titus > Orange data mining > two recensions grouping
(Homilies 1 and 5)

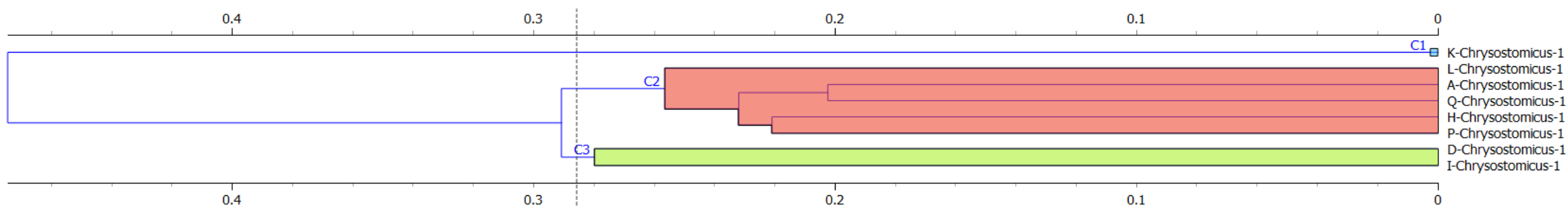
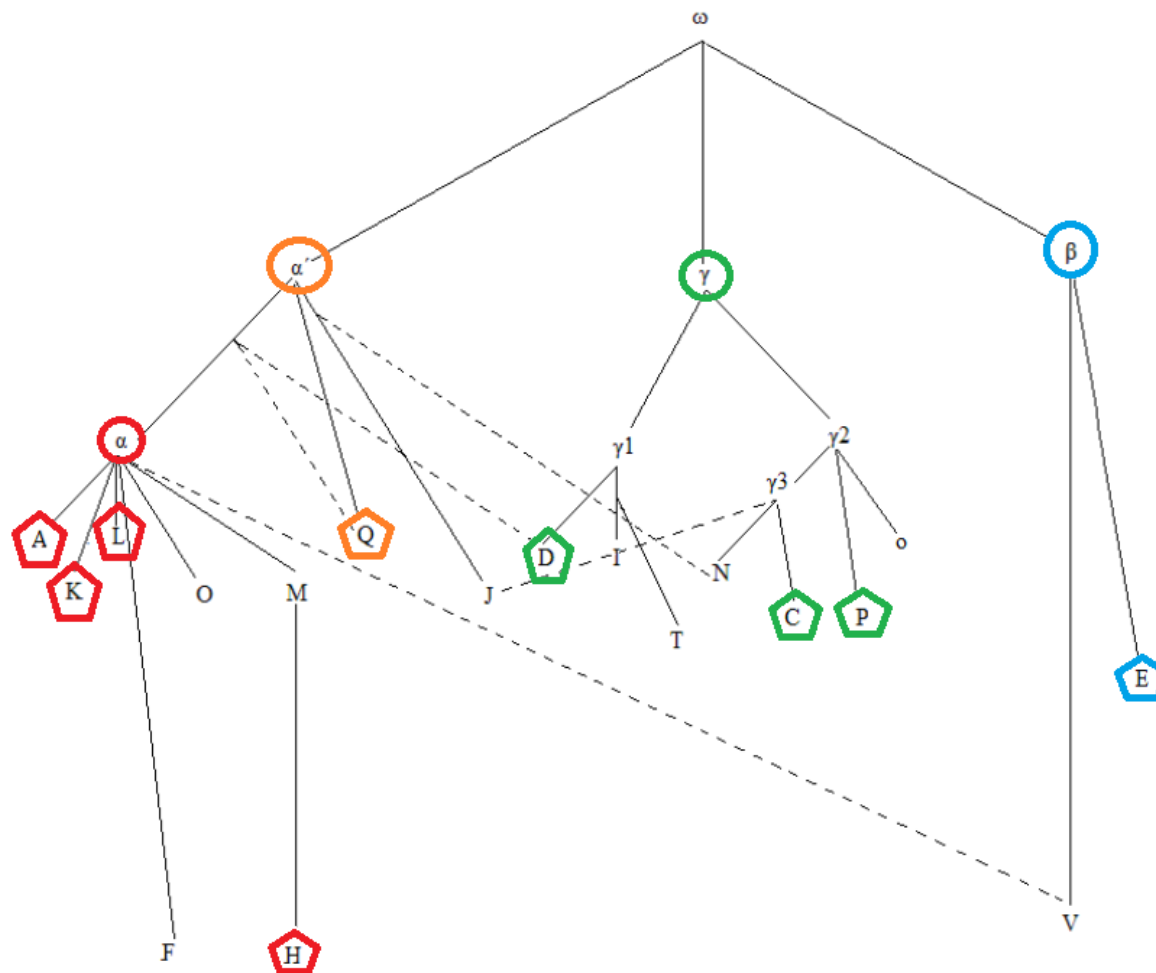
Data clustering



- _ data input
- _ text preprocessing
- _ bag of words
- _ distances
- _ hierarchical clustering

The main takeaway

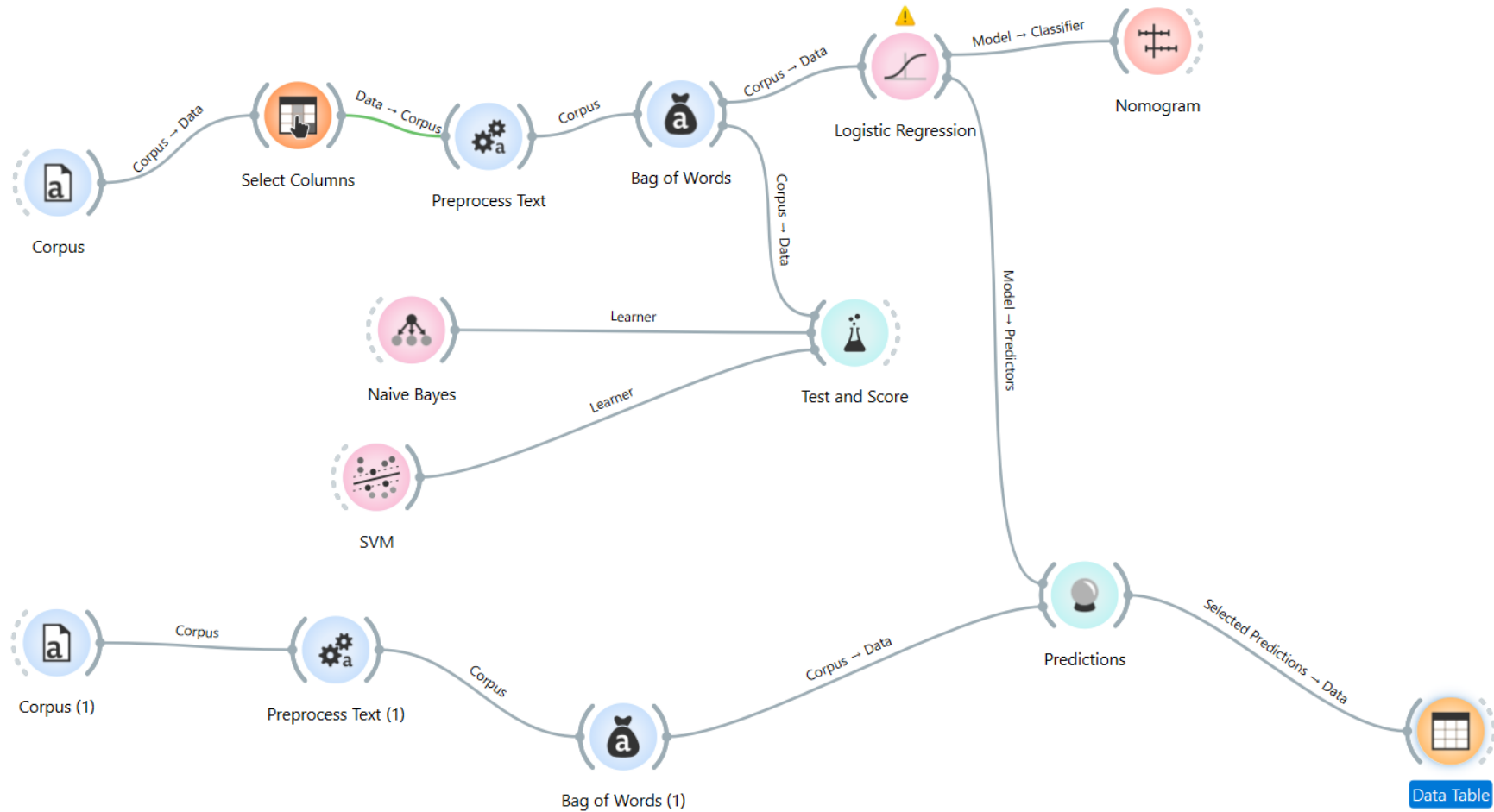
successful even with
HTR data ($\leq 20\%$ CER)



Do not ask *what* but *how*

not *what* is the *stemma codicum* but *how* do we connect the witnesses

Is classification now possible?



Only 4 errors out of 21 instances

bow-feature hidden include skip-normalization title	ms_id	text (GT)	logistic Regression	{...}
	True	True		
1	A-Chrysostomic...	ΠΤΟΥΕΝΑΤΙΟΙΣ...	long	recension (shor...
2	D-Chrysostomi...	Τ ΛΟΣΗΣ ΣΕΡΙΩ...	short	recension (shor...
3	H-Chrysostomi...	†ΘΠΠΘΤΑΙ ΟΙ...	long	recension (shor...
4	I-Chrysostomic...	οὕτως καὶ ἡμεῖ...	long	recension (shor...
5	K-Chrysostomic...	Α ΤΡΥ IN ΑΟΟΙ...	short	recension (shor...
6	L-Chrysostomic...	ΑΛΟΓΟΣ ΠΕΩΤ...	long	recension (shor...
7	Q-Chrysostomi...	+ ΛΟΓΟΣ ΠΡΩΣ...	short	recension (shor...
8	A-Chrysostomic...	τοὺς οἰκέτας λέ...	long	recension (shor...
9	D-Chrysostomi...	ΛΥΓΕ΄ΠΠΠΠΠΠ...	short	recension (shor...
10	H-Chrysostomi...	※ ΛΟΓΟΣ Ε ※...	long	recension (shor...
11	I-Chrysostomic...	ΟΜΙΛΙΑ †ΠΤΟΥ...	short	recension (shor...
12	K-Chrysostomic...	+ ΛΟΓΟΣ Ε :~ΠΠ...	long	recension (shor...
13	L-Chrysostomic...	+ ΛΟΓΟΣ Ε †ΠΠ...	long	recension (shor...
14	Q-Chrysostomi...	ΛΟΓΟΣ ΕΠΠΠΠ...	short	recension (shor...
15	A-Chrysostomic...	ΠΤΟΥΕΝΑΤΙΟΙΣ...	long	recension (shor...
16	D-Chrysostomi...	Τ ΛΟΣΗΣ ΣΕΡΙΩ...	long	recension (shor...
17	H-Chrysostomi...	†ΘΠΠΘΤΑΙ ΟΙ...	long	recension (shor...
18	I-Chrysostomic...	οὕτως καὶ ἡμεῖ...	long	recension (shor...
19	K-Chrysostomic...	Α ΤΡΥ IN ΑΟΟΙ...	long	recension (shor...
20	L-Chrysostomic...	ΑΛΟΓΟΣ ΠΕΩΤ...	long	recension (shor...
21	Q-Chrysostomi...	+ ΛΟΓΟΣ ΠΡΩΣ...	short	recension (shor...)

Spoiler alert

Why we cannot rely on it

Where's the catch?

Test and Score - Orange

☒ Cross validation

Number of folds: 5

☒ Stratified

☐ Cross validation by feature

☐ Random sampling

Repeat train/test: 10

Training set size: 66 %

☒ Stratified

☐ Leave one out

☐ Test on train data

☐ Test on test data

Evaluation results for target (None, show average over classes)

Model	Train	Test	AUC	CA	F1	Prec	Recall	MCC	Spec	LogLoss
Naive Bayes	0.372	0.117	0.583	0.571	0.571	0.595	0.571	0.167	0.595	2.717
SVM	0.505	0.512	0.354	0.571	0.416	0.327	0.571	0.000	0.429	0.764
Logistic Regression	0.201	0.118	0.604	0.571	0.571	0.571	0.571	0.125	0.554	0.692

overfitting

Compare models by: Area under ROC curve

☐ Negligible diff.: 0.1

	Naive Bayes	SVM	Logistic Regression
Naive Bayes		0.500	0.114
SVM	0.500		0.169
Logistic Regression	0.886	0.831	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

14 | 3x14



Part of the inhumanity of the computer is that, once it is competently programmed and working smoothly, it is completely honest.

Asimov Isaac. 1971. "Who Needs Money?" in *Change! Seventy-One Glimpses of the Future*.