From ‘time’ elastic distances to ‘time’ elastic inner vector spaces

-Prospective applications in ‘time’ series, sequence or text classification-

P.-F. Marteau, SEASIDE

Challenging issue

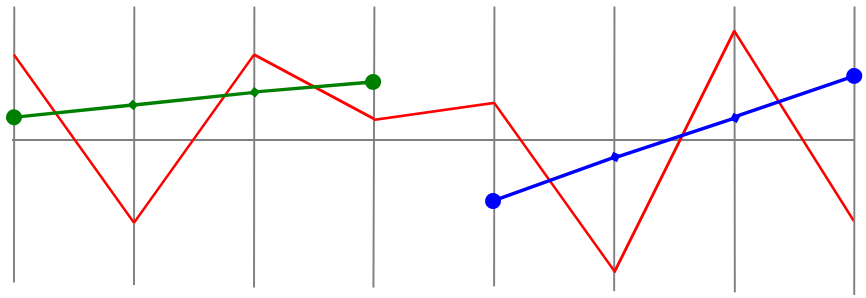
- Embedding complex data in an inner vector space (or normed space) gives access to a wide range of applications relying on main linear algebra results
  - PCA, SVD (LSI, LSA), 'Optimal' Dimension Reduction, Matching pursuit, data compression, etc.
- For sequential or 'time-stamped' data, elastic distances (such as Levenshtein Dist., DTW, LCSS, ERP, TWED) have proved to be superior in general to Euclidean or Minkowski distances in particular for discrimination or regression tasks
  - Unfortunately, elastic distances do not derive from a norm (an inner product)
- **Existence of 'time-elastic' inner products?**
  - Preserving the access to main linear algebra results
  - Keeping beneficing from 'time-elasticity' capability

Content

- Challenging issue
- ‘Time’-stamped data
- ‘Time elastic’ distance or pseudo distance
- ‘Time elastic inner products’ (TEIP)
- Prospective applications
- Conclusion, perspectives

‘times-tamped’ data

- U : sequence set, $A \in$ U is a finite sequence
- $A_1^n = A(1)A(2)...A(n)$ with $A(i) \in SxT$
- S : “spatial” set (numeric or symbolic) : spatial dimensions
- T : “time” set : timestamp dimension
- $A(i)=(a(i);t_{a(i)})$
- Examples

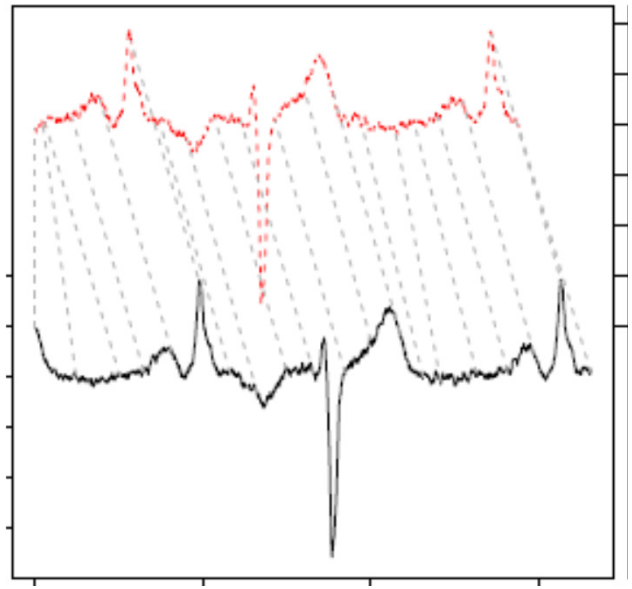


AGTCCGCGAATACAGGCTCGGT

<TEXT> *By itself, not a whole lot. Genome sequencing is often compared to "decoding," but a sequence is still very much in code. In a sense, a genome sequence is simply a very long string of letters in a mysterious language.*

‘Time’-elastic distances or pseudo distances

$$DTW(A_1^p, B_1^q) = \delta(a(p), b(q)) + Min \begin{cases} DTW(A_1^p, B_1^{q-1}) \\ DTW(A_1^{p-1}, B_1^{q-1}) \\ DTW(A_1^{p-1}, B_1^q) \end{cases}$$



- Recursive definition
- O(N^2)
- Nonlinearity (*Min, Max, δ(.;.)*),
  - ➔ Impossibility to derive from elastic distances inner (dot) products
  - ➔ Impossibility to derive a norm
  - ➔ Existence of ‘*elastic’ inner* vector space ?

Dot/Inner Products

- Euclidean dot product

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$$

- Inner product

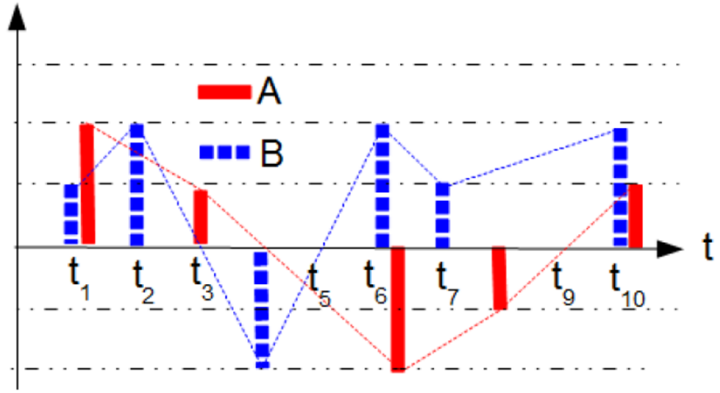
•Conjugate symmetry: $\langle x, y \rangle = \overline{\langle y, x \rangle}$.
Note that in $\mathbb{R}$, it is symmetric.

•Linearity in the first argument: $\langle ax, y \rangle = a\langle x, y \rangle$.
$\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$.

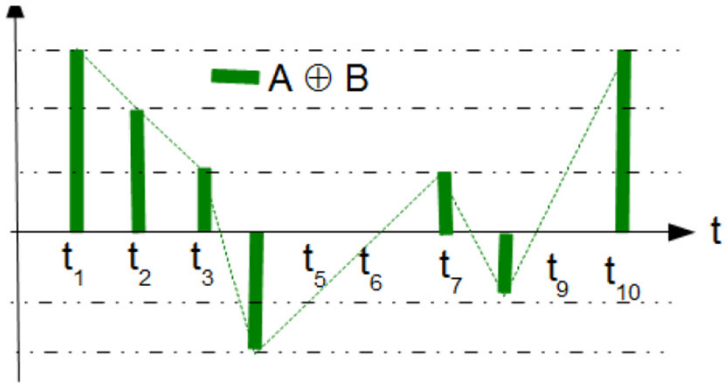
•Positive-definiteness: $\langle x, x \rangle \geq 0$
with equality only for $x = 0$.

$(U^*, \oplus, \otimes)$ provided $(S, \oplus_S, \otimes_S)$

$\otimes$ **Definition 2.3.** For all $A \in \mathbb{U}^*$ and all $\lambda \in \mathbb{R}$, $C = \lambda \otimes A \in \mathbb{U}^*$ is such that for all $i \in \mathbb{N}$ such that $0 \leq i \leq |A|$, $C(i) = (\lambda.a(i), t_{a(i)})$ and thus $|C| = |A|$.



$\oplus$ **Definition 2.4.** For all $(A, B) \in (\mathbb{U}^*)^2$, the addition of A and B



U^*: set of sequences defined on $(S-\{0_s\})xT$

From elastic distance to 'time' elastic products

$$DTW(A_1^p, B_1^q) = \delta(a(p), b(q)) + Min \begin{cases} DTW(A_1^p, B_1^{q-1}) \\ DTW(A_1^{p-1}, B_1^{q-1}) \\ DTW(A_1^{p-1}, B_1^q) \end{cases}$$

$$< A_1^p, B_1^q >_{tep} = \sum \begin{cases} \alpha \cdot < A_1^{p-1}, B_1^q >_{tep} \\ \beta \cdot < A_1^{p-1}, B_1^{q-1} >_{tep} + f(a(p), b(q)) \cdot g(t_{a(p)}, t_{b(q)}) \\ \alpha \cdot < A_1^p, B_1^{q-1} >_{tep} \end{cases}$$

$$< A, \Omega >_{tep} = < \Omega, A >_{tep} = < \Omega, \Omega >_{tep} = \xi$$

where Ω is the null sequence, ξ, α, β are real
and g is strictly positive

Provided the existence of the addition $\oplus$ and scalar multiplication $\otimes$ on the sequence set U, what are the conditions on *f, g,* α, β *and* ξ to make it an inner product?

‘Time’ elastic inner products

(U, ⊕, ⊗) provided (S, $\oplus_S$, $\otimes_S$)

*U** is the set of sequences whose elements are in $S\backslash\{0_S\}$
$(0_S,t)$ identifies the null sequence element for all *t*)

Theorem 2.1. $< .,. >_{tep}$ *is an inner product on* $(\mathbb{U}^*, \oplus, \otimes)$ *iff:*

i) $\xi = 0$.
ii) $g : (T \times T) \rightarrow \mathbb{R}$ *is symmetric and strictly positive,*
iii) f *is an inner product on* $(S, \oplus_S, \otimes_S)$*, if we extend the domain of* f *on* S *while setting* $f(0_S, 0_S) = 0$.
iv) $\alpha = 1$ *and* $\beta = -1$,

Note that according to this result, it is possible to embed sequences of various lengths or times series not-uniformly sampled and/or of various lengths in a unique ‘elastic’ vector space structure.

‘Time’ elastic inner products

Example of a ‘time elastic’ inner product

$$< A_1^p, B_1^q >_{teip} = \sum \begin{cases} < A_1^p, B_1^{q-1} >_{teip} \\ - < A_1^{p-1}, B_1^{q-1} >_{teip} + a(p)b(q) \cdot e^{-\nu.|t_{a_p} - t_{b_q}|} \\ < A_1^{p-1}, B_1^q >_{teip} \end{cases}$$

ν: stiffness parameter
ν=0, infinite elasticity
ν =∞, null elasticity

Recursively embedded elastic dimensions

$$< A_1^p, B_1^q >_{teip} = \sum \begin{cases} < A_1^{p-1}, B_1^q >_{tep} \\ - < A_1^{p-1}, B_1^{q-1} >_{tep} + g(t_{a(p)}, t_{b(q)}) \cdot < a(p), b(q) >_{teip(S)} \\ < A_1^p, B_1^{q-1} >_{tep} \end{cases}$$

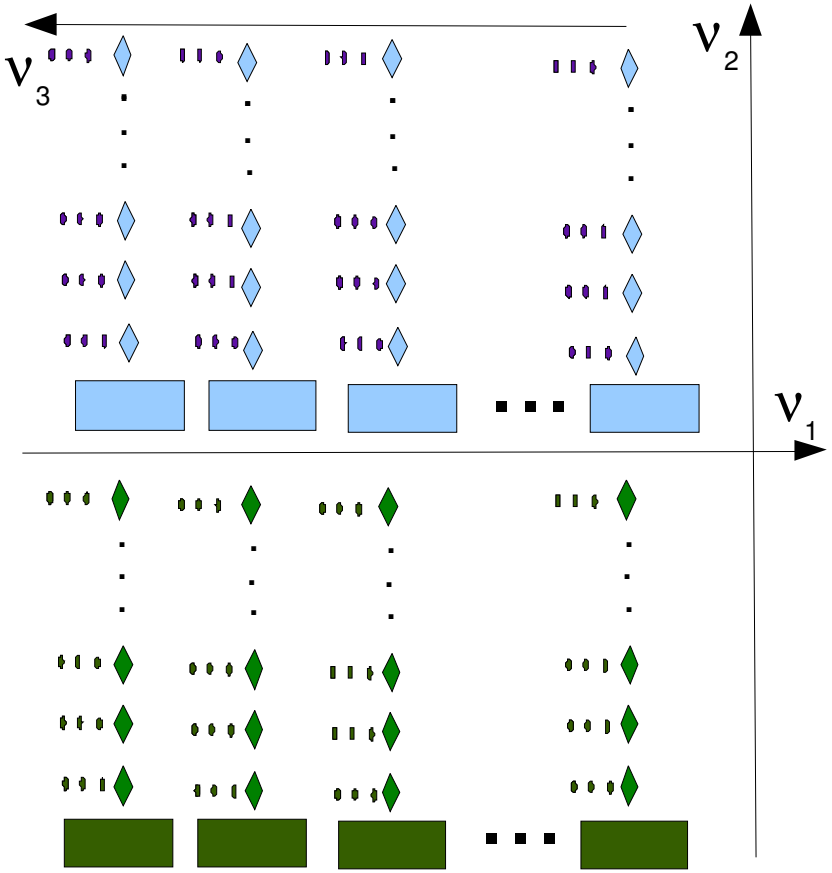
Paragraph level

Sentence Level

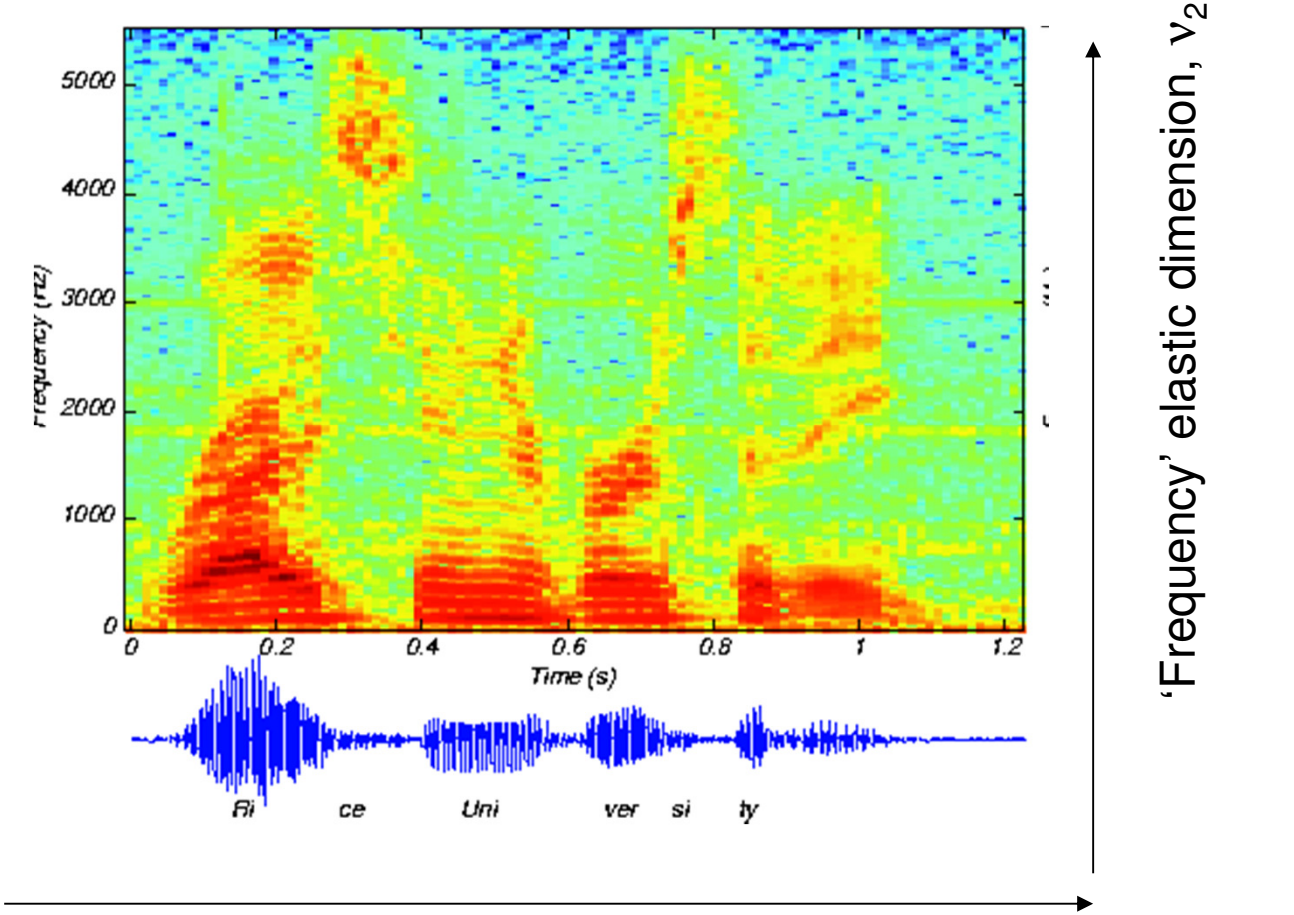
Word Level

Section Level

Document Level



Recursively embedded elastic dimensions



'Frequency' elastic dimension, ν_2

'Time' elastic dimension, ν_1

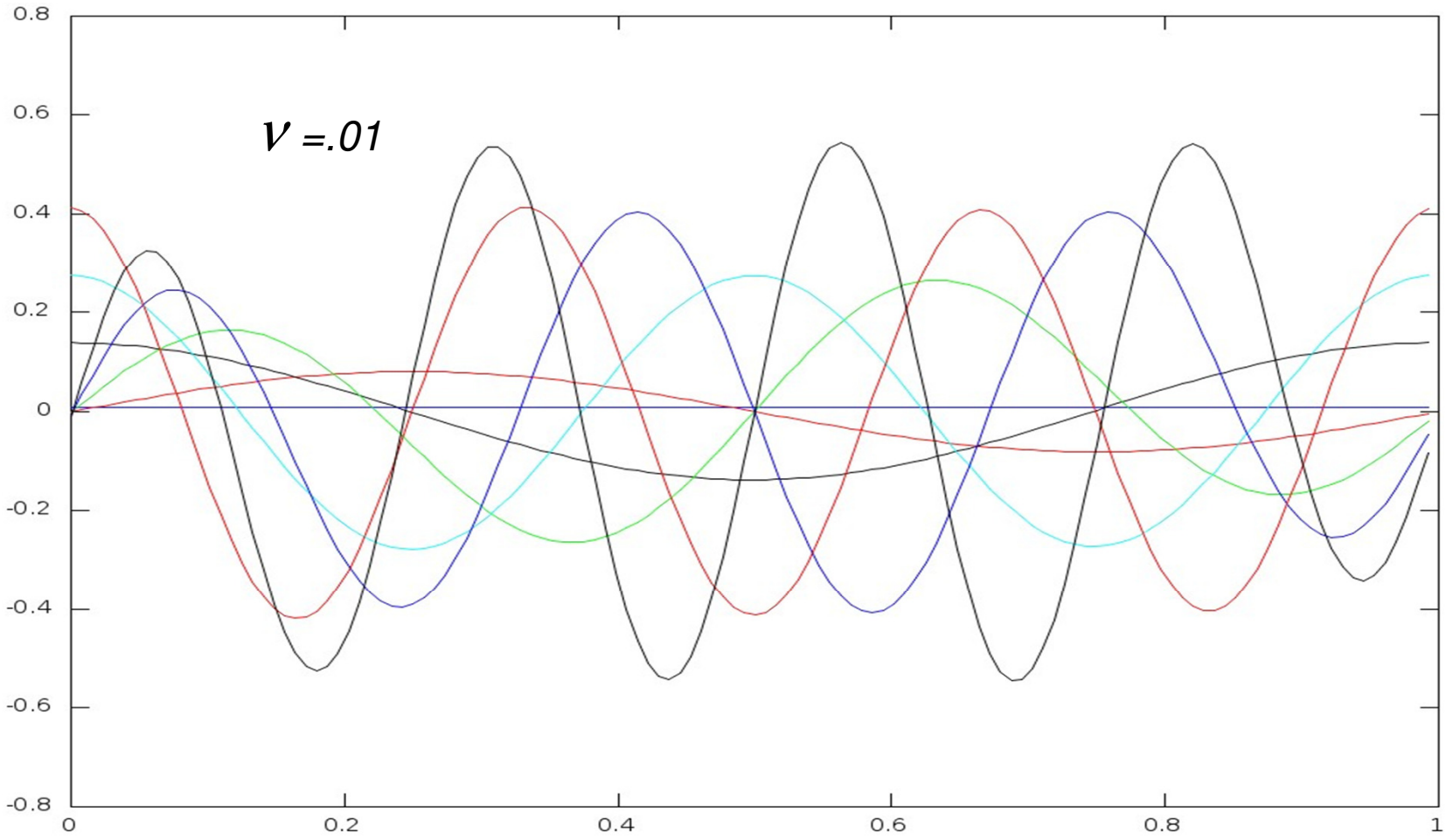
Elastic measures

Elastic norm: $$\left\| A_1^p \right\|_e = \sqrt{< A_1^p , A_1^p >_e}$$

Elastic distance $$eDist(A_1^p , B_1^q) = \left\| A_1^p \oplus (-B_1^q) \right\|_e$$

Elastic cosine $$eCos(A_1^p , B_1^q) = \frac{< A_1^p , B_1^q >_e}{\left\| A_1^p \right\|_e . \left\| B_1^q \right\|_e}$$

Sanity check : Gram-Schmidt orthogonalization of a sine basis



Prospective applications in sequence or text matching

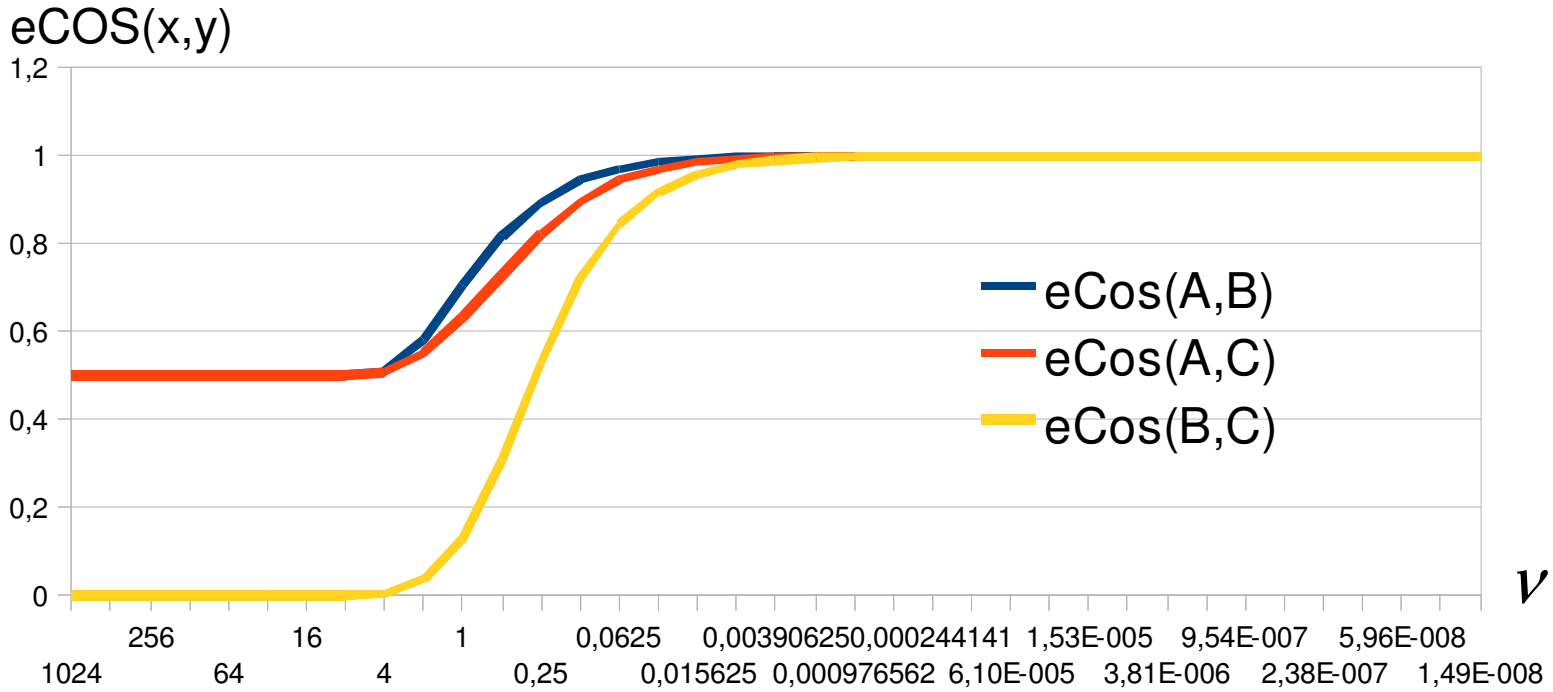
$$< A_1^p, B_1^q >_{teip_{tm}} = \sum \begin{cases} < A_1^{p-1}, B_1^q >_{teip_{tm}} \\ - < A_1^{p-1}, B_1^{q-1} >_{teip_{tm}} + e^{-\nu.|t_{a(p)} - t_{b(q)}|} \delta(a(p), b(q)) \\ < A_1^p, B_1^{q-1} >_{teip_{tm}} \end{cases} \quad (7)$$

where $a(p)$ and $b(q)$ are vectors whose coordinates identify words with weightings, $\delta(a, b) =< a, b >$ is the Euclidan inner product, and ν a *time stiffness* parameter.

Proposition 3.3. *For $\nu = 0$ and δ redefined as $\delta(a, b) = 1$ if $a = b$, 0 otherwise, the elastic inner product defined in Eq.7 coincides with the euclidean inner product between two vectors whose coordinates correspond to term frequencies observed into the A_1^p and B_1^q text sequences.*

Proposition 3.4: *if ν -> ∞, and if the two sequences in argument are uniformly sampled and have the same length, then the TEIP tends toward the Euclidean inner product*

Toy experiment

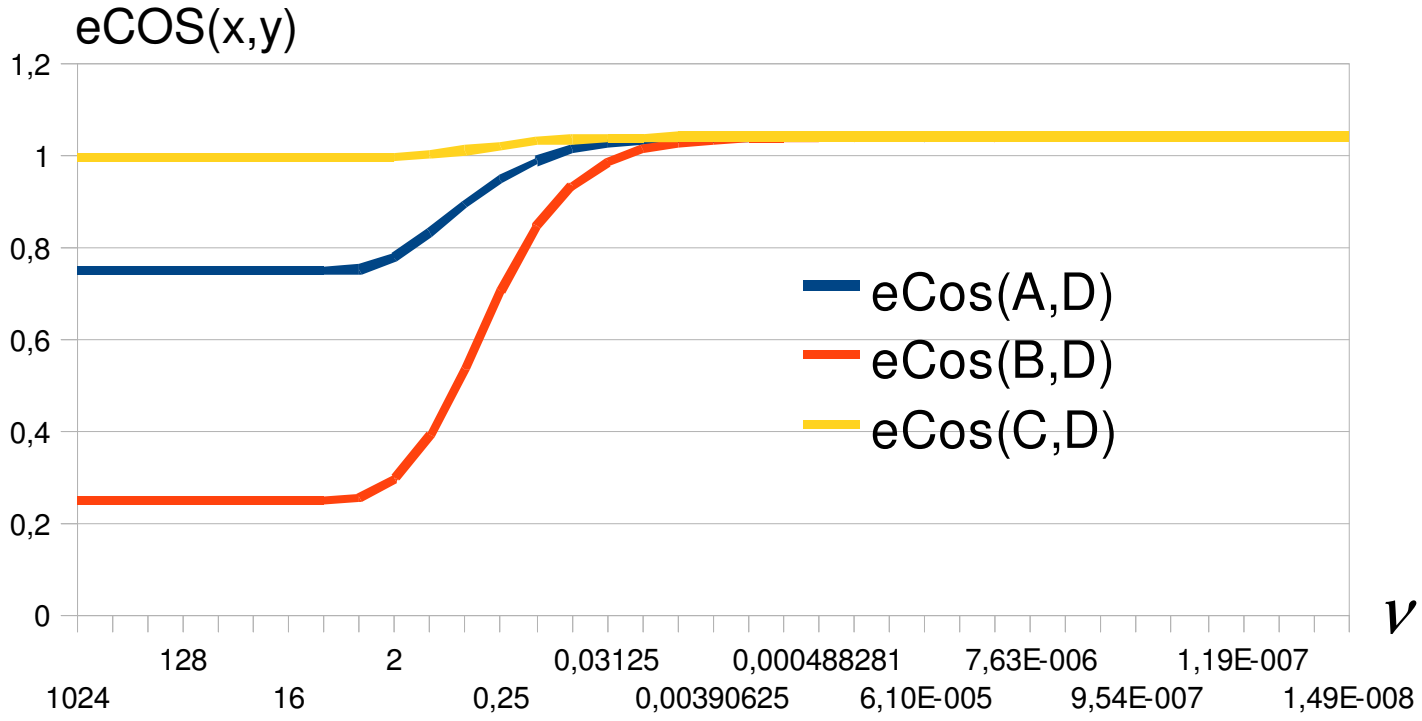


$$eCos(A_1^p, B_1^q) = \frac{< A_1^p, B_1^q >_e}{\|A_1^p\|_e . \|B_1^q\|_e}$$

$$\delta(x, y) = 1 \text{ if } x = y, \ 0 \text{ otherwise}$$

A	a	b	a	b	a	b	a	b
B	a	a	a	a	b	b	b	b
C	b	b	b	b	a	a	a	a

Toy experiment

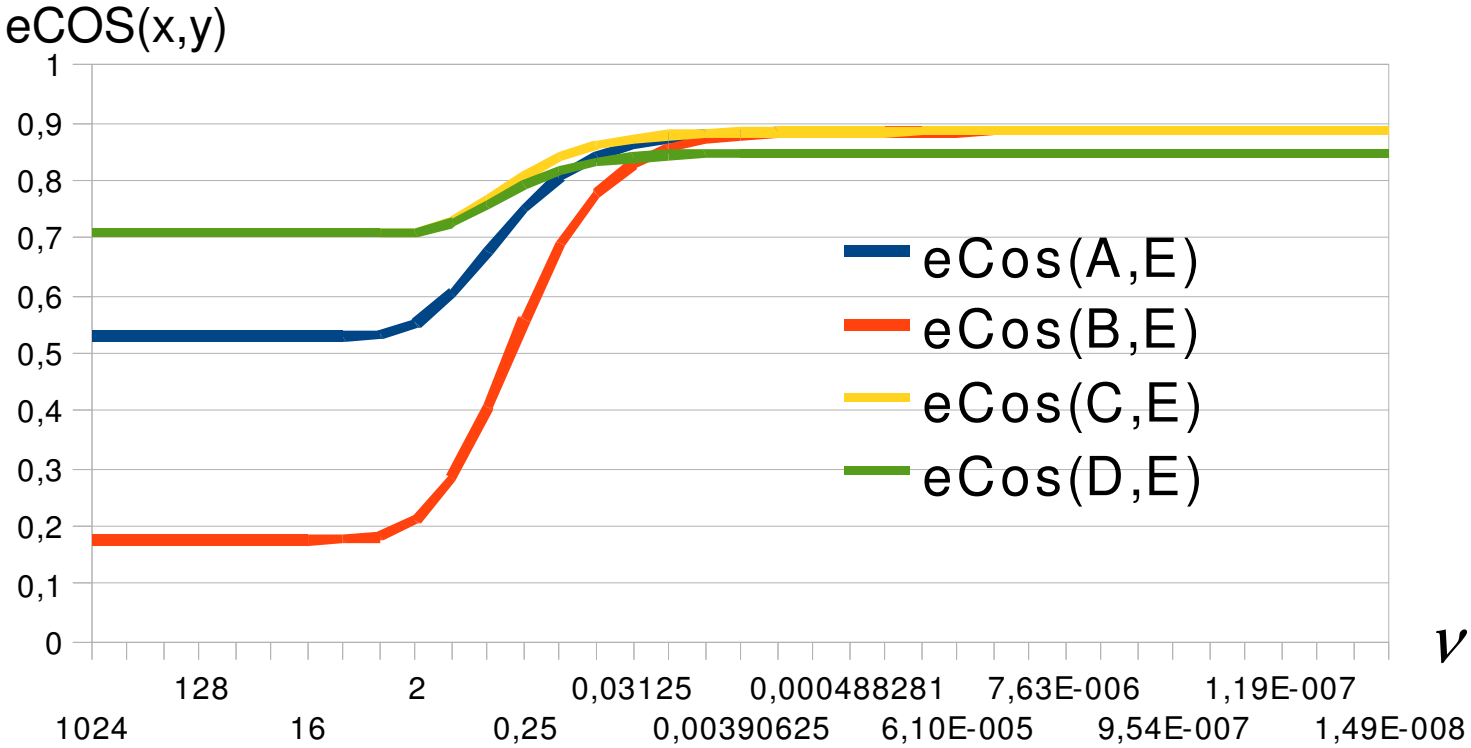


$$eCos(A_1^p, B_1^q) = \frac{< A_1^p, B_1^q >_e}{\|A_1^p\|_e . \|B_1^q\|_e}$$

$$\delta(?, x) = \delta(x, ?) = 1$$

A	a	b	a	b	a	b	a	b
B	a	a	a	a	b	b	b	b
C	b	b	b	b	a	a	a	a
D	b	b	?	b	a	?	a	a

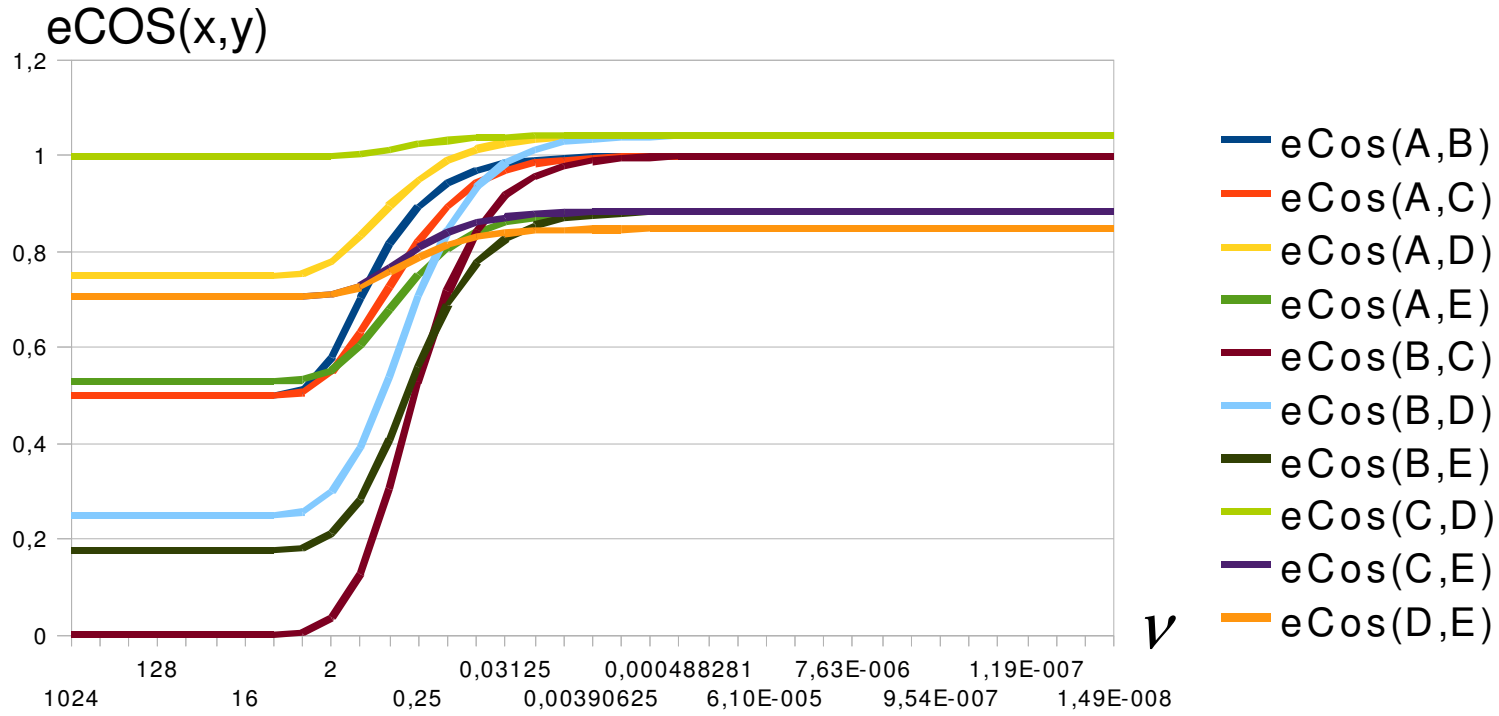
Toy experiment



$$eCos(A_1^p, B_1^q) = \frac{< A_1^p, B_1^q >_e}{\|A_1^p\|_e . \|B_1^q\|_e}$$

A	a	b	a	b	a	b	a	b
B	a	a	a	a	b	b	b	b
C	b	b	b	b	a	a	a	a
D	b	b	?	b	a	?	a	a
E	b	b	a	a				

Toy experiment



A	a	b	a	b	a	b	a	b
B	a	a	a	a	b	b	b	b
C	b	b	b	b	a	a	a	a
D	b	b	?	b	a	?	a	a
E	b	b	a	a				

CPER: support Invent'IST
IRISA, IETR, VALORIA, IRMAR, LMAM

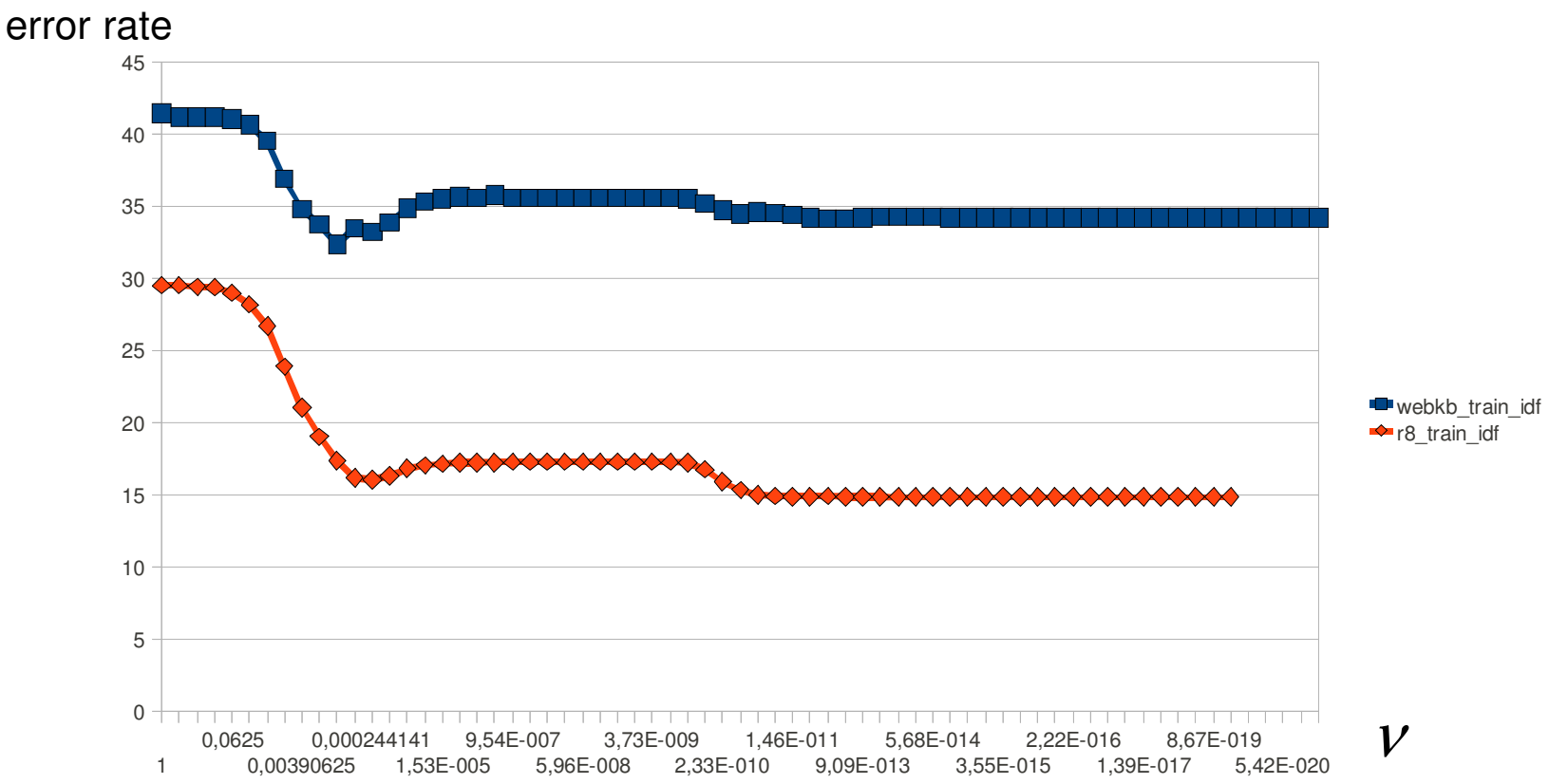
Platform for "Massive" Data Processing : 250 k€

Dr Nicolas Bonnel

Implementation on GPU andcluster (today ~20 nodes, perspective 32 nodes)

Early results in Text Mining

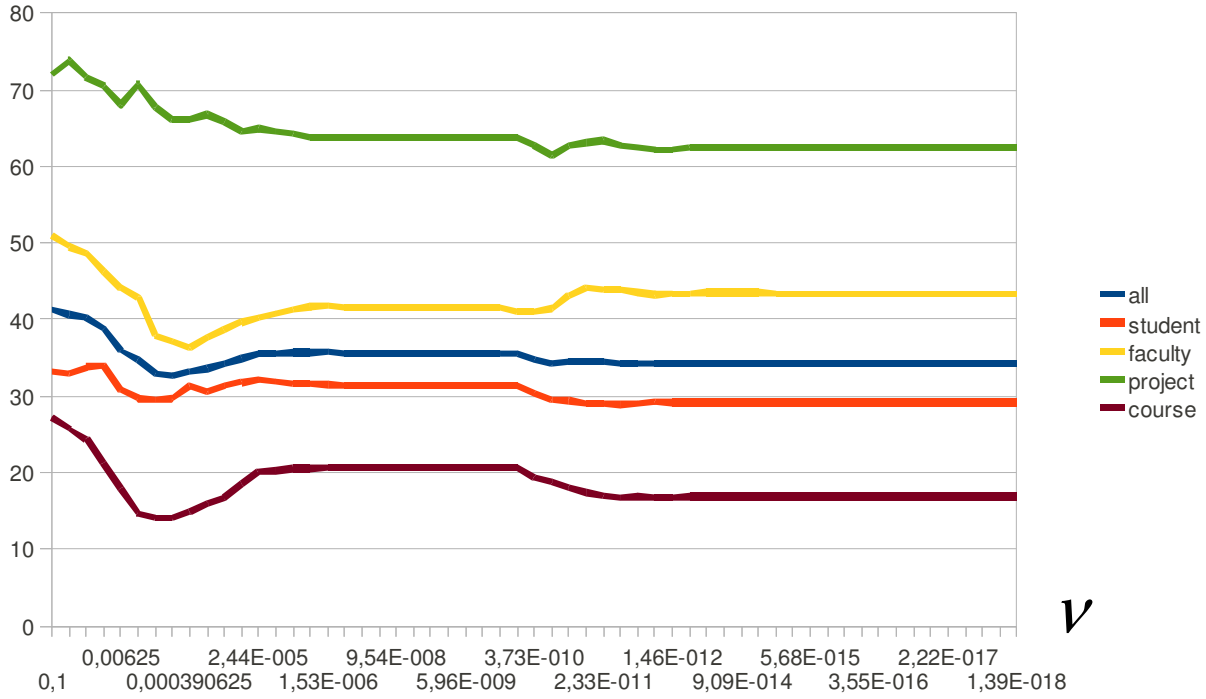
Text classification experiment: Reuters 21578, WebKB datasets



Early results in Text Mining

Text classification experiment: WebKB classes

Classif. error rate



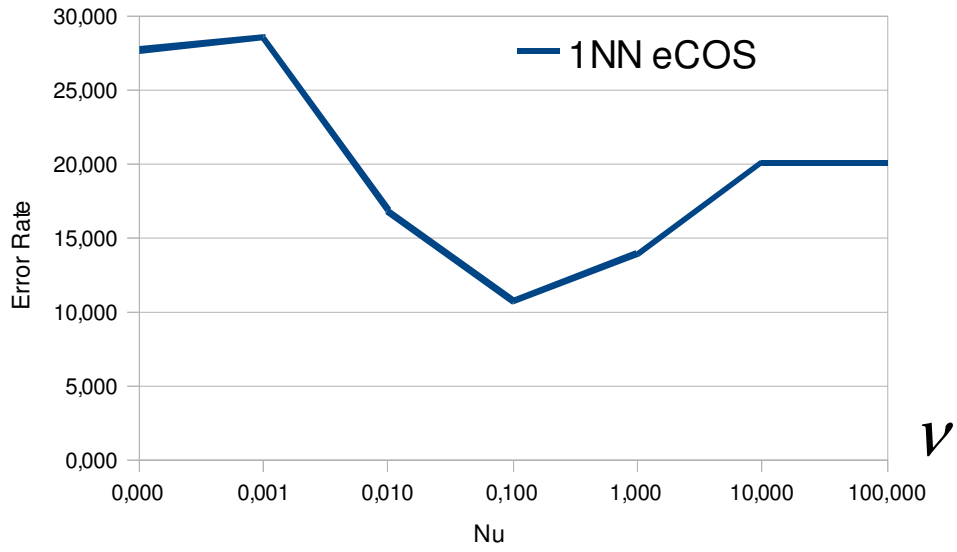
Early results in sequence classification

- **Protein Classification Benchmark Collection (ICGEB/EMBNet)**
  - 3-phosphoglycerate kinase (3PGK) protein sequences

ID	POSITIVE TEST	POSITIVE TRAIN	NEGATIVE TEST	NEGATIVE TRAIN
Archaea_Crenarchaeota	4	11	53	63
Archaea_Euryarchaeota	11	4	53	63
Bacteria_Actinobacteridae	5	68	31	27
Bacteria_Firmicutes	35	38	31	27
Bacteria_Proteobacteria	30	43	31	27
Eukaryota_Alveolata	4	39	44	44
Eukaryota_Euglenozoa	5	38	44	44
Eukaryota_Fungi	10	33	44	44
Eukaryota_Metazoa	12	31	44	44
Eukaryota_Viridaeplanta	8	35	44	44

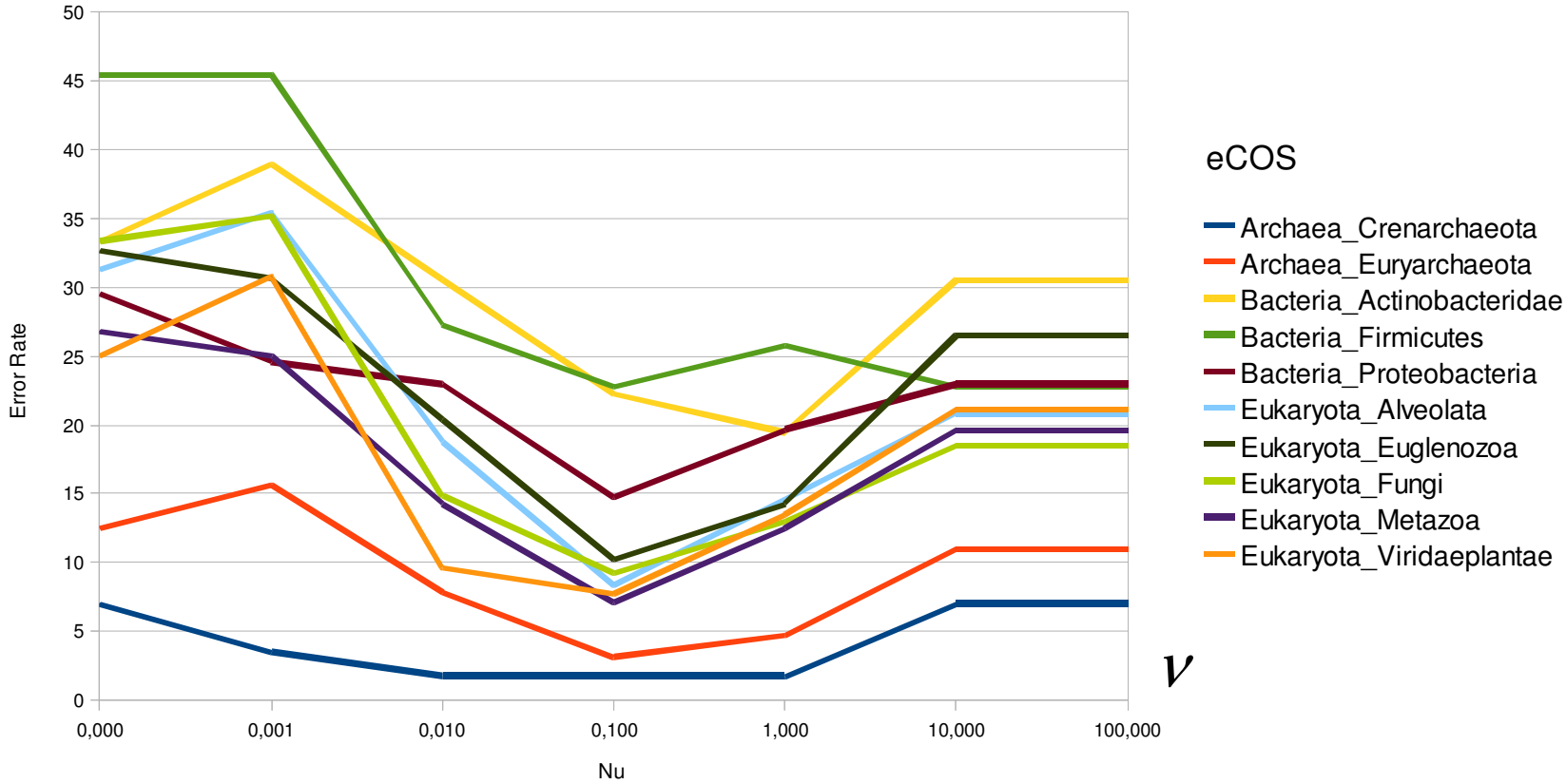
Early results in sequence classification

–3-phosphoglycerate kinase (3PGK) protein sequences



Early results in sequence classification

3-phosphoglycerate kinase (3PGK) protein sequences



Average AUC values for the 10 classification tasks

1NN	BLAST	SW	NW	LA	PRIDE
SCOP95_supfam_fam_1	0,70	0,84	0,84	0,79	0,89
SCOP95_supfam_fam_kfold_2	0,95	0,98	0,98	0,97	0,97
SCOP95_fold_supfam_3	0,57	0,65	0,67	0,58	0,85
SCOP95_fold_supfam_kfold_4	0,90	0,94	0,95	0,91	0,96
SCOP95_class_fold_5	0,60	0,61	0,63	0,57	0,78
SCOP95_class_fold_kfold_6	0,92	0,92	0,93	0,88	0,93
CATH95_H_S_7	0,98	0,99	0,99	0,99	0,95
CATH95_H_S_kfold_8	0,97	0,98	0,98	0,98	0,95
CATH95_T_H_9	0,57	0,65	0,67	0,61	0,77
CATH95_T_H_kfold_10	0,90	0,93	0,93	0,91	0,91
CATH95_A_T_11	0,55	0,57	0,58	0,54	0,67
CATH95_A_T_kfold_12	0,90	0,91	0,92	0,89	0,85
CATH95_C_A_13	0,65	0,71	0,70	0,69	0,56
CATH95_C_A_kfold_14	0,91	0,90	0,91	0,86	0,81
Average	**0,79**	**0,83**	**0,83**	**0,80**	**0,85**

SW: Smith-Waterman
NW: Needleman-Wunsch
LA: Local Alignment kernel
PRIDE : Probability of similarity (Histogram of distances between motifs)

AUC: Area under the Curve
(**ROC** curve)

1NN	ν=1	ν=0,1	ν=0,05	ν=0,01	ν=0,001	Max	Rank
SCOP95_supfam_fam_1	0,78	0,82	**0,83**	0,84	0,81	0,89	4
SCOP95_supfam_fam_kfold_2	0,91	0,95	**0,95**	0,94	0,91	0,97	5
SCOP95_fold_supfam_3	0,75	0,79	**0,79**	0,76	0,70	0,85	2
SCOP95_fold_supfam_kfold_4	0,89	0,93	**0,93**	0,91	0,86	0,95	4
SCOP95_class_fold_5	0,59	0,66	**0,67**	0,67	0,64	0,80	2
SCOP95_class_fold_kfold_6	0,80	0,88	**0,89**	0,85	0,80	0,90	5
CATH95_H_S_7	0,95	0,97	**0,97**	0,96	0,92	0,98	5
CATH95_H_S_kfold_8	0,95	0,97	**0,97**	0,96	0,94	0,98	4
CATH95_T_H_9	0,73	0,79	**0,79**	0,77	0,70	0,85	1
CATH95_T_H_kfold_10	0,89	0,93	**0,93**	0,90	0,84	0,94	1
CATH95_A_T_11	0,58	0,62	**0,62**	0,61	0,60	0,77	2
CATH95_A_T_kfold_12	0,82	0,88	**0,88**	0,85	0,80	0,90	5
CATH95_C_A_13	0,38	0,46	**0,50**	0,63	0,64	0,71	6
CATH95_C_A_kfold_14	0,77	0,83	**0,84**	0,81	0,76	0,85	5
Average	**0,77**	**0,82**	**0,83**	**0,82**	**0,78**	**0,88**	**3,64**

Conclusion, perspectives

- The existence of ‘time-elastic’ products seems to be proved given the existence of $\oplus$ and $\otimes$
  - Elastic inner products generalize somehow the Euclidean inner product and the classical vector model defined for textual information retrieval
  - Several ‘time-elastic’ dimensions can be managed recursively
  - Leads to positive definite kernels (not the case for DTW-like distances)
- Early results on time series or sequential data classification show some potentiality that need to be consolidated
- Generalization to simultaneously elastic dimension? (can **T** be multidimensional?)
- ‘time-elastic’ Linear Algebra has to be further developed