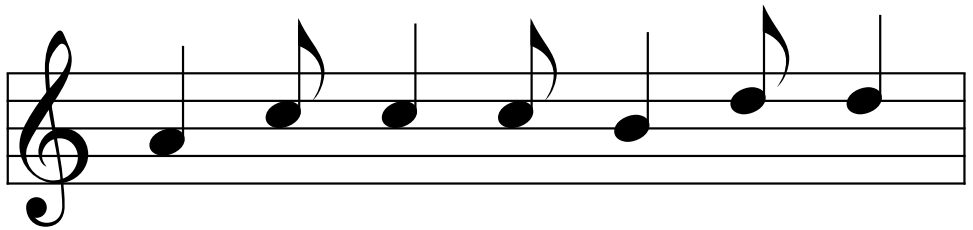


Trois ma-tous gris à mi-nuit...

Une introduction à la notion de *dissemblance analogique* et à son application à l'apprentissage.

Laurent Miclet, Arnaud Delhay, Sabri Bayoudh

IRISA Lannion-Projet Cordial.

Comptines (1).

❶ J'ai vu trois gros matous gris à minuit.

❷ J'ai vu trois matous gris à minuit pile.

❸ J'ai vu trois gros matous à minuit.

❹ J'ai vu ....

La question est de trouver ❹ telle que :

❶ est à ❷ comme ❸ est à ❹

Comptines (1).

J'ai	vu	trois	gros	matous	gris	à	minuit	
J'ai	vu	trois		matous	gris	à	minuit	pile
J'ai	vu	trois	gros	matous		à	minuit	
J'ai	vu	trois		matous		à	minuit	pile

Le secret : *aligner* les quatre séquences et appliquer les trois règles :

1	1	0
0	1	1
1	0	0
0	0	1

Comptines (2).

1. J'ai vu trois matous à minuit.
2. J'ai vu deux chatons à seize heures.
3. J'ai vu cinq taureaux à midi.
4. J'ai vu ....

Comptines (2).

Trois	matous	à	minuit
Deux	chatons	à	seize heures
Cinq	taureaux	à	neuf heures
Quatre	veaux	à	une heure

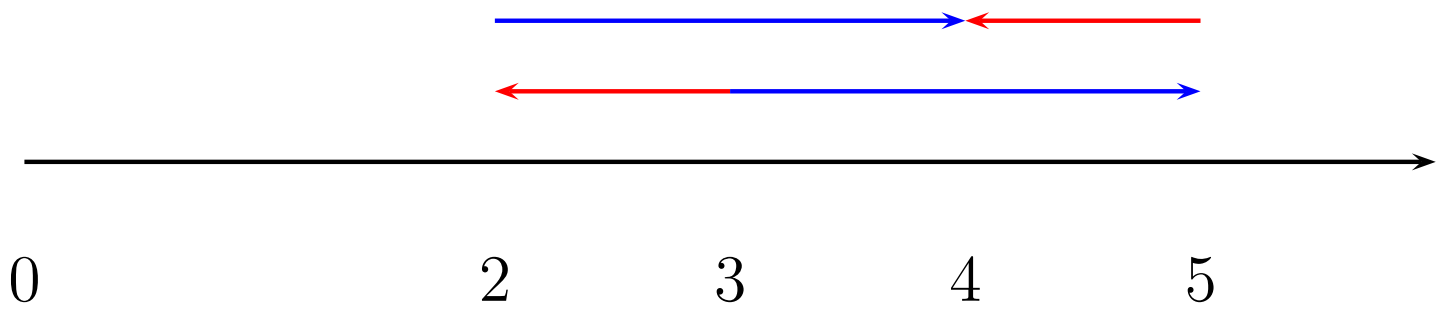
Le secret : l'alignement est déjà fait, il suffit de résoudre...

Mais c'est un peu plus compliqué :

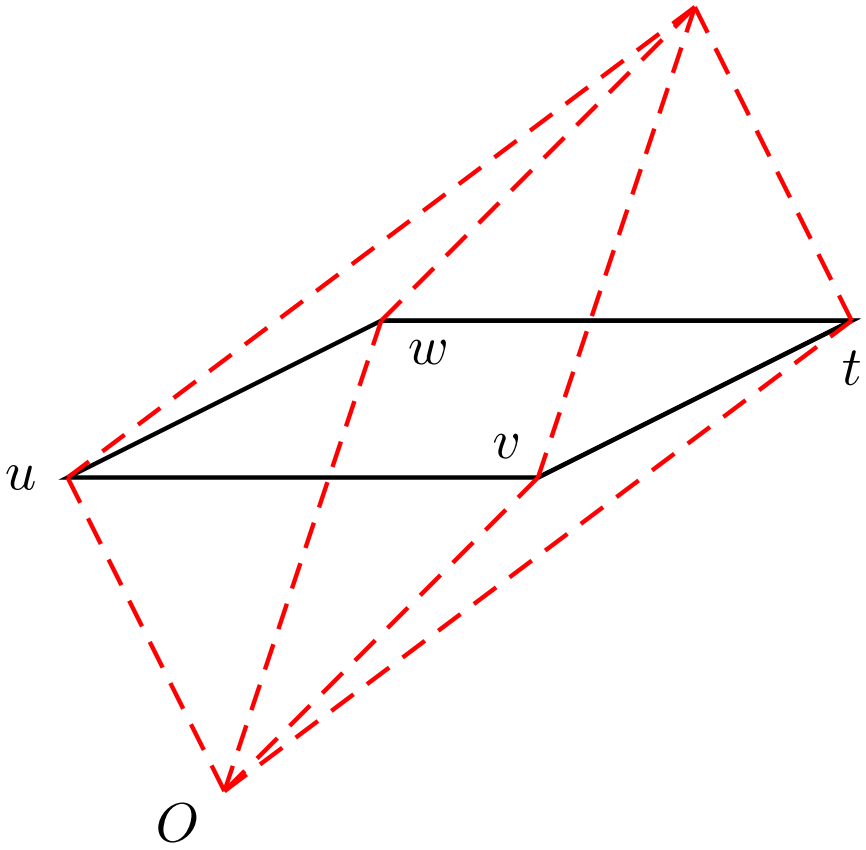
- chiffres
- animaux
- heures
- ...

Une résolution possible pour les chiffres.

Trois	:	Deux	::	Cinq	:	Quatre
3	:	2	::	5	:	4
3	:	5	::	2	:	4



Résolution dans un espace vectoriel.



t est défini par :

- $\overrightarrow{uv} = \overrightarrow{wt}$
- $\overrightarrow{uw} = \overrightarrow{vt}$
- $\overrightarrow{Ou} + \overrightarrow{Ot} = \overrightarrow{Ov} + \overrightarrow{Ow}$

Conséquence :

- $\Rightarrow \delta(u,v) = \delta(w,t)$
  $\delta(u,w) = \delta(v,t)$

Résolution dans un espace de traits binaires.

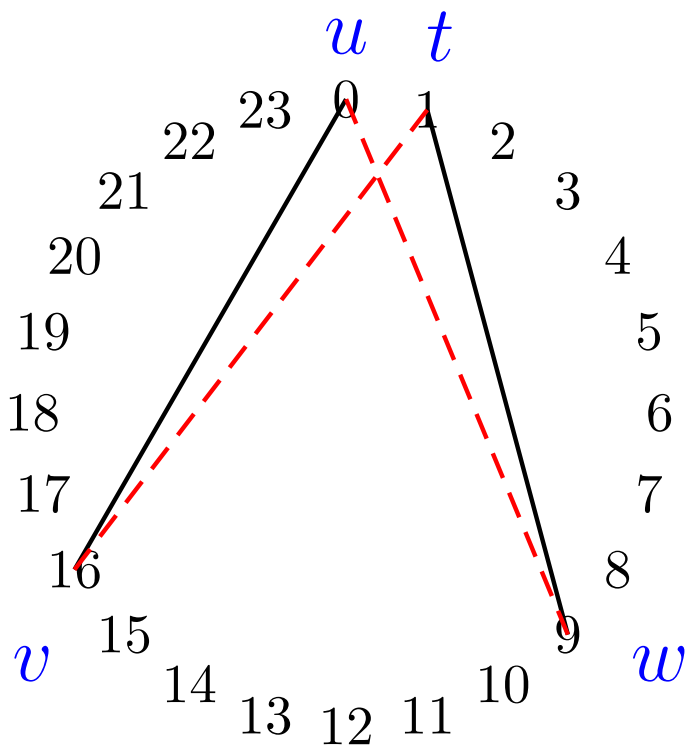
	oiseau	bovin	jeune	mâle	félin	mamm.	adulte
matous	0	0	0	1	1	1	1
chatons	0	0	1	0	1	1	0
taureaux	0	1	0	1	0	1	1
veaux	0	1	1	0	0	1	0

$$\delta_H(matous,chatons) = \delta_H(taureaux,veaux) = 3$$

$$\delta_H(matous,taureaux) = \delta_H(chatons,veaux) = 2$$

Résolution dans un espace cyclique.

minuit **est à** seize heures **comme** neuf heures **est à** une heure



$$\delta(u,v) = \delta(w,t) \quad \text{et} \quad \delta(u,w) = \delta(v,t)$$

Comptines (3).

1. Trois gros matous gris à minuit.
2. Deux chatons noirs à seize heures pile.
3. Cinq gros taureaux gris à neuf heures.
4. ....

Le secret : aligner et résoudre *en même temps.*

Trois	gros	matous	gris	à	minuit	
Deux		chatons	noirs	à	seize heures	pile
Cinq	gros	taureaux	gris	à	neuf heures	
Quatre		veaux	noirs	à	une heure	pile

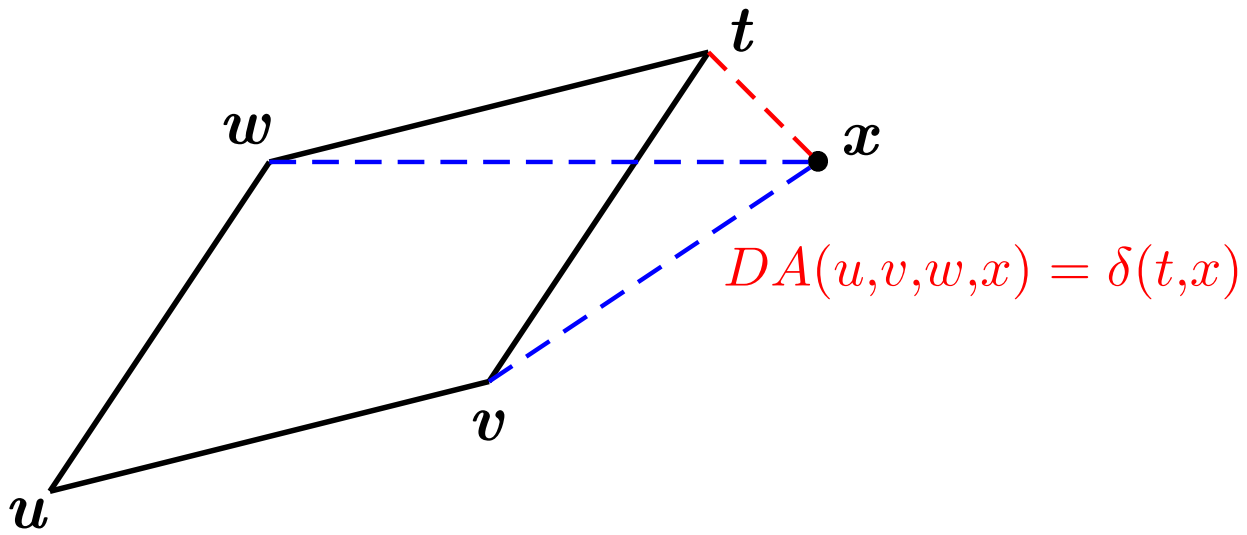
Résolution approximative.

Trois	gros	matous	gris	à	minuit	
Deux		chatons	noirs	à	seize heures	pile
Cinq	gros	taureaux	gris	à	neuf heures	
Quatre		veaux	noirs	à	une heure	pile
Trois		vaches		à	deux heures	pile
↑		↑			↑	

Dissemblance analogique :

A est à B à peu près comme C est à D.

Dissemblance analogique dans $\mathbb{R}^n$.



$$DA(u,v,w,x) = \delta(x,t) \quad \text{avec} \quad u : v :: w : t$$

Exemple.

Trois	gros	matous	gris	à	minuit	
Deux		chatons	noirs	à	seize heures	pile
Cinq	gros	taureaux	gris	à	neuf heures	
Trois		vaches		à	deux heures	pile

↑

$$DA(trois,deux,cinq,trois) = DA(3,2,5,3) = ?$$

On résoud $DA(3,2,5,t)$, ce qui donne : $t + 3 = 2 + 5$, soit $t = 4$.

On calcule la distance entre $x = 3$ et $t = 4$, qui vaut 1. Donc,

$$DA(3,2,5,3) = 1$$

Dissemblance analogique dans $\mathbb{B}^n$.

u	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
v	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
w	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
t	0	0	1	1	1	1	?	?	?	?	0	0	0	0	1	1
x	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
DA	0	1	1	0	1	0	2	1	1	2	0	1	0	1	1	0
							1			1						

Exemple.

Trois	gros	matous	gris	à	minuit	
Deux		chatons	noirs	à	seize heures	pile
Cinq	gros	taureaux	gris	à	neuf heures	
Trois		vaches		à	deux heures	pile

↑

Exemple (suite).

	oiseau	bovin	jeune	mâle	félin	mamm.	adulte
matous	0	0	0	1	1	1	1
chatons	0	0	1	0	1	1	0
taureaux	0	1	0	1	0	1	1
veaux	0	1	1	0	0	1	0
vaches	0	1	0	0	0	1	1

$$DA(matous,chatons,taureaux,vaches) = \delta_H(veaux,vaches) = 2$$

Exemple complètement différent.

	oiseau	bovin	jeune	mâle	félin	mamm.	adulte
chatons	0	0	1	0	1	1	0
matous	0	0	0	1	1	1	1
taureaux	0	1	0	1	0	1	1
	0	1	?	?	0	1	?
vaches	0	1	0	0	0	1	1

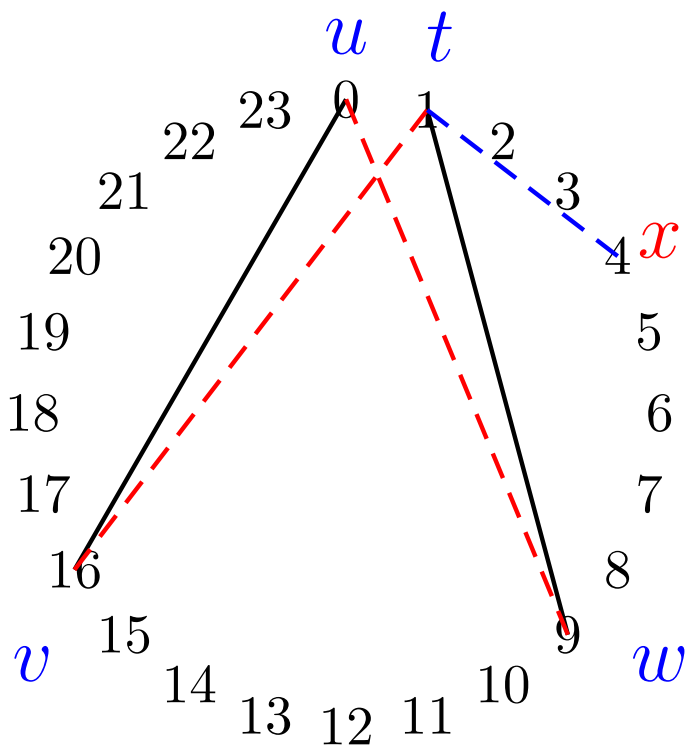
$DA(chatons,matous,taureaux,vaches) = \delta_H(???,vaches) = 1+2+1 = 4$

chaton : matous :: taureaux : ?

n'a pas de solution, mais on peut calculer

$DA(chatons,matous,taureaux,vaches)$

Dissemblance analogique dans un groupe cyclique.



Exemple.

$$DA(\text{minuit, seize heures, neuf heures, quatre heures}) = ?$$

On résoud

$$\text{minuit} : \text{seize heures} :: \text{neuf heures} : t$$

Ce qui donne (il y a toujours une solution unique)

$$t = \text{une heure}$$

La distance entre "quatre heures" et "une heure" est la longueur de la corde du cercle entre 4 et 1, soit (pour un rayon 1) environ 0.8

Pourquoi choisir cette mesure? Pour qu'elle soit une *distance*, en particulier pour qu'elle vérifie l'inégalité triangulaire.

Les axiomes de l'analogie.

Une analogie sur un ensemble X est une relation sur X^4, *i. e.* un sous-ensemble $\mathcal{A} \subset X^4$. Quand $(A,B,C,D) \in \mathcal{A}$, les quatre éléments A, B, C et D sont *en analogie*, ce qui s'écrit "la relation analogique $A : B :: C : D$ est vraie", ou simplement $A : B :: C : D$, ce qui se lit "A est à B comme C est à D". Pour tout quadruplet en analogie, il est nécessaire que les axiomes suivants soient vérifiés :

Symétrie de la relation "comme" : $C : D :: A : B$

Echange des termes moyens :: $A : C :: B : D$

Un troisième axiome (*le déterminisme*) définit les solutions des équations triviales :

$$A : A :: B : X \quad \Rightarrow \quad X = B$$
$$A : B :: A : X \quad \Rightarrow \quad X = B$$

Conséquences.

- Avec ces axiomes, on démontre que cinq autres formulations sont équivalentes à $A : B :: C : D$:

$$B : A :: D : C \qquad D : B :: C : A \qquad C : A :: D : B$$

$$D : C :: B : A \qquad \text{et} \qquad B : D :: A : C$$

- Une analogie s'exprime donc de huit façons équivalentes.
- Il y a seulement trois analogies possibles entre quatre termes, dont les formes canoniques sont :

$$A : B :: C : D \qquad A : C :: D : B \qquad A : D :: B : C$$

Propriétés de la dissemblance analogique.

Dans les trois types d'ensembles où nous l'avons définie, la dissemblance analogique possède les propriétés suivantes :

1. $\forall(u,v,w,x) \in X^4,\ AD(u,v,w,x) = 0 \Leftrightarrow u : v :: w : x$
2. $\forall(u,v,w,x) \in X^4,\ AD(u,v,w,x) = AD(w,x,u,v) = AD(u,w,v,x)$
3. $\forall(u,v,w,x,z,t) \in X^6,\ AD(u,v,w,x) \leq AD(u,v,z,t) + AD(z,t,w,x)$
4. En général, $\forall(u,v,w,x) \in X^4,\ \ AD(u,v,w,x) \neq AD(v,u,w,x)$

Des alphabets aux séquences.

- On sait résoudre des équations analogiques dans trois types d'ensembles.
- On a défini une dissemblance analogique avec de bonnes propriétés dans ces trois cas.
- On va s'intéresser maintenant à l'analogie et à la dissemblance analogique entre séquences composées d'éléments de l'un quelconque de ces ensembles.
- La technique consiste à aligner optimalement les séquences pour résoudre "colonne par colonne".
- L'alphabet doit contenir un symbole particulier "$\backsim$" qui signifie "absence de lettre" : $a \backsim B \equiv\backsim aB \backsim\equiv aB$.

Un exemple.

Soit $\Sigma' = \{a,b,\alpha,\beta,A,B,\backsim\}$ avec les analogies

$a : b :: A : B$, $a : \alpha :: b : \beta$ et $A : \alpha :: B : \beta$.

Les quatre séquences $aBA\beta$, $\alpha bBAB$, $ba\alpha$ et βbaA forment une analogie par l'alignement :

$$\begin{array}{ccccc} a & \backsim & B & A & \beta \\ | & | & | & | & | \\ \alpha & b & B & A & B \\ | & | & | & | & | \\ b & \backsim & a & \backsim & \alpha \\ | & | & | & | & | \\ \beta & b & a & \backsim & A \end{array}$$

Définition de cet alphabet par des traits binaires.

	a	$\overline{a}$	b	$\overline{b}$	m	$\overline{m}$	M	$\overline{M}$	g	$\overline{g}$
a	1	0	0	1	1	0	0	1	0	1
b	0	1	1	0	1	0	0	1	0	1
A	1	0	0	1	0	1	1	0	0	1
B	0	1	1	0	0	1	1	0	0	1
α	1	0	0	1	0	1	0	1	1	0
β	0	1	1	0	0	1	0	1	1	0
$\backsim$	0	0	0	0	0	0	0	0	0	0

Distance de Hamming associée.

	a	b	A	B	α	β	$\backsim$
a	0	4	4	8	4	8	5
b	4	0	8	4	8	4	5
A	4	8	0	4	4	8	5
B	8	4	4	0	8	4	5
α	4	8	4	8	0	4	5
β	8	4	8	4	4	0	5
$\backsim$	5	5	5	5	5	5	

Il peut exister plusieurs solutions exactes.

$$
\begin{array}{cccc}
a & \backsim & b & \backsim \\
| & | & | & | \\
\backsim & \backsim & \alpha & \backsim \\
| & | & | & | \\
a & a & b & b \\
| & | & | & | \\
\backsim & a & \alpha & b
\end{array}
\quad \text{et} \quad
\begin{array}{cccc}
a & \backsim & b & \backsim \\
| & | & | & | \\
\alpha & \backsim & \backsim & \backsim \\
| & | & | & | \\
a & a & b & b \\
| & | & | & | \\
\alpha & a & \backsim & b
\end{array}
$$

Comment calculer la dissemblance?

Il faut aligner les quatre séquences en cherchant le meilleur alignement, celui qui minimise la somme des dissemblances analogiques dans les colonnes.

a	$\backsim$	b	$\backsim$
|	|	|	|
$\backsim$	$\backsim$	α	$\backsim$
|	|	|	|
a	a	b	b
|	|	|	|
$\backsim$	a	A	b
0	0	4	0

a	$\backsim$	b	$\backsim$
|	|	|	|
α	$\backsim$	$\backsim$	$\backsim$
|	|	|	|
a	a	b	b
|	|	|	|
$\backsim$	a	A	b
5	0	5	0

Comment faire?

- L'algorithme que nous proposons est une généralisation du calcul de la distance d'édition.
- Il parcourt les quatre séquences en synchronie par programmation dynamique.
- Il trouve la (ou les) solution(s) optimale(s) en remplissant un tableau à quatre dimension. Chaque case est remplie en calculant le minimum sur quinze cases "précédentes".

Utilisation en apprentissage.

- Le cas simple : un ensemble d'exemples (disons, de séquences) supervisés par des classes.
- $\mathcal{S} = \{x_i, U(x_i)\}$ pour $i = 1, m$
- Connaissant y, comment trouver $U(y)$?
- Chercher dans $\mathcal{S}$ le triplet (u^*, v^*, w^*) tel que $DA(u^*, v^*, w^*, y)$ soit minimale.
- Résoudre l'équation analogique $U(u^*) : U(v^*) :: U(w^*) : T$
- Affecter T à $U(y)$.

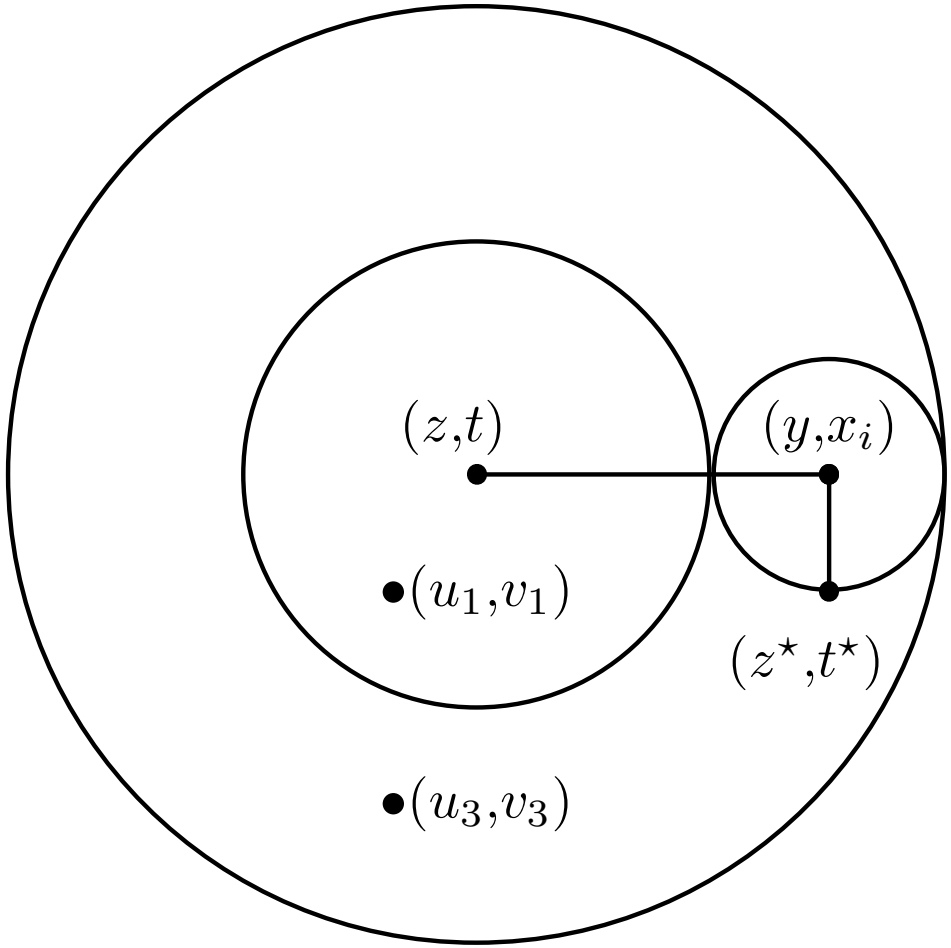
Quelques problèmes immédiats.

- Que signifie l'analogie sur les étiquettes ?
- Comment traiter les solutions triviales multiples ?
- Comment traiter les solutions non triviales différentes ?
- Comment traiter des supervisions plus complexes ?
  - Par exemple : transcription orthographique-phonétique.
- Comment diminuer la complexité, qui est *a priori* en $\mathcal{O}(m^3)$?
- Quelles sont les données adaptées à un tel type d'apprentissage ?
- Comment affecter un coût à l'insertion et à la destruction d'une lettre ?

Quelques raisons d'être optimiste.

- On connaît des alphabets qui sont "naturellement" adaptés.
  - Alphabet phonétique (défini par des traits).
  - Code de Freeman (groupe cyclique).
  - Distances entre protéines.
- On connaît des applications effectives en linguistique (morphologie, traduction).
- On peut adapter des méthodes rapides de recherche du plus proche voisin.
- On peut (peut-être) utiliser les méthodes de plongement des espaces finis métriques dans les espaces normés de petite dimension, qui commencent à être utilisées pour la recherche approximative du plus proche voisin.

Une idée pour la recherche rapide.



(u_2,v_2)

$$DA(y,x_i,u_1,v_1) \geq DA(y,x_i,z^*,t^*)$$

$$DA(y,x_i,u_2,v_2) \geq DA(y,x_i,z^*,t^*)$$

Une idée pour l'apprentissage de l'analogie entre étiquettes de classes.

- Dans l'ensemble d'apprentissage, corréler les analogies entre objets et les analogies entre étiquettes.
- Appliquer cette corrélation à la décision.

Des idées pour améliorer l'apprentissage pour les objets "vecteurs binaires".

- Dans l'ensemble d'apprentissage, corréler les analogies entre traits et les analogies entre étiquettes.
- Appliquer cette corrélation à la décision.

ou

- Dans l'ensemble d'apprentissage (à deux classes), chercher quels sont les traits les plus "analogiquement" discriminants.
- Pondérer la dissemblance analogique par trait en fonction de cette qualité.

Expériences : données SPECT.

- Number of Instances: 267
- Number of Attributes: 23 (22 binary + 1 binary class)
- Attribute Information:
  - OVERALL DIAGNOSIS: 0,1 (class attribute, binary)
  - F1: 0,1 (the partial diagnosis 1, binary)
  - ...
  - F22: 0,1 (the partial diagnosis 22, binary)
- dataset is divided into:
  - training data (”SPECT.train” 80 instances)
  - testing data (”SPECT.test” 187 instances)

Expériences : données SPECT.

Class Distribution			
Class	Entire data	training dataset	testing dataset
0	55	40	15
1	212	40	172

Résultats									
PPV		ID3		Analogie		An. pondérée		CLIP4
11	4	12	3	11	4	10	5	
76	97	51	121	61	112	53	120	
57.8%		71.1%		65.7%		69.5%		86.1%

Bibliographie.

- Thèse de N. Stroppa, sous la direction de F. Yvon.
  Vendredi 4 novembre après-midi, à l'ENST Paris.
- Yves Lepage : `www.slc.atr.jp/~lepage/`
- Rapports internes INRIA de A. Delhay et L. Miclet (2003, 2004, 2005).
- A. Delhay, L. Miclet (2005). *Analogie entre séquences. Définition, calcul et utilisation en apprentissage supervisé.* RIA. Vol 19/4-5, pp.683-712.