

# Classification et gestion d'incertitudes

Classification et gestion des incertitudes  
Violaine Antoine

GT Big Data, Clermont-Ferrand, France

14 Novembre 2024

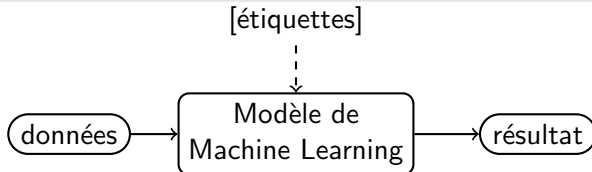
# Outline

- 1 Introduction
- 2 Clustering semi-supervisé et étiquettes imprécises
- 3 Classification avec étiquettes imprécises
- 4 Conclusion

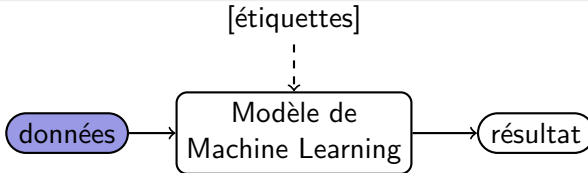
# Outline

- 1 Introduction
- 2 Clustering semi-supervisé et étiquettes imprécises
- 3 Classification avec étiquettes imprécises
- 4 Conclusion

# Les informations partiels sont partout en Machine learning !

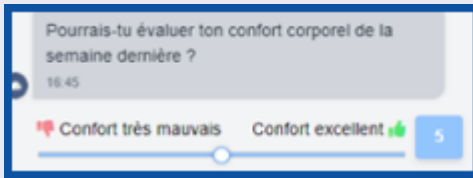


# Les informations partiels sont partout en Machine learning !

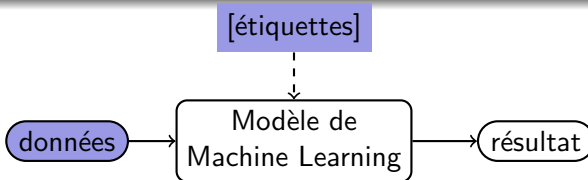


## Exemples

- Données
  - subjectivité de questionnaires, réponses aléatoires

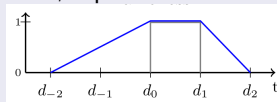


# Les informations partiels sont partout en Machine learning !



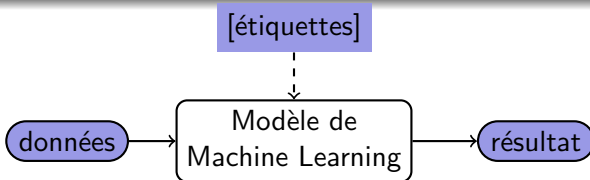
## Exemples

- Données
  - subjectivité de questionnaires, réponses aléatoires
- Etiquettes
  - états graduels
  - informations imprécises



7  
classe 1 ou 7

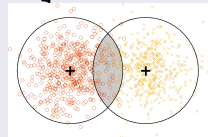
# Les informations partiels sont partout en Machine learning !



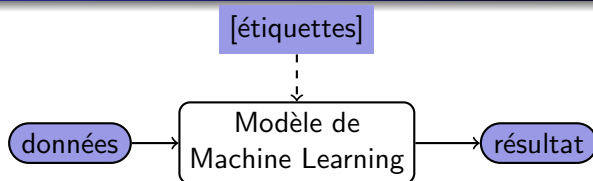
## Exemples

- Données
  - subjectivité de questionnaires, réponses aléatoires
- Etiquettes
  - états graduels
  - informations imprécises
- Résultats
  - incertitude sur la décision

**7**  
classe 1 ou 7



# Intérêt de conserver des informations partielles



## Entrées

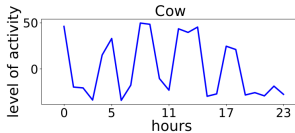
- permet de conserver tous les individus
- se rapproche plus de la véracité des données

## Sortie

- permet un postprocessing humain ou automatique
  - étude manuelle des outliers, des objets imprécis,...
  - apprentissage actif
  - ...

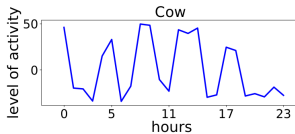


# Types de connaissances partielles



état 1 oestrus	état 2 boîterie	état 3 velage
état 4 blessure	état 5 mammite	state 6 apathie

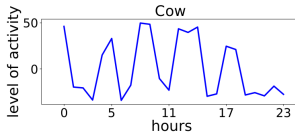
# Types de connaissances partielles



Décision ou connaissance stricte

état 1 oestrus	état 2 boîterie	état 3 velage
état 4 blessure	état 5 mammite	state 6 apathie

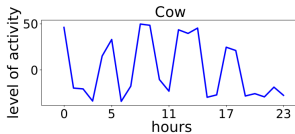
# Types de connaissances partielles



Imprécision

état 1 oestrus	état 2 boîterie	état 3 velage
état 4 blessure	état 5 mammite	state 6 apathie

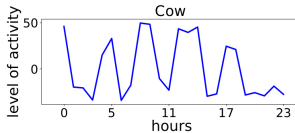
# Types de connaissances partielles



Incertitude

état 1 oestrus	état 2 boîterie	état 3 velage
état 4 blessure	état 5 mammite	state 6 apathie

# Types de connaissances partielles



Ignorance

état 1

oestrus

état 4

blessure

état 2

boîterie

état 5

mammite

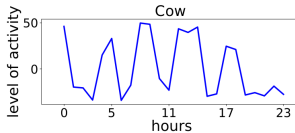
état 3

velage

state 6

apathie

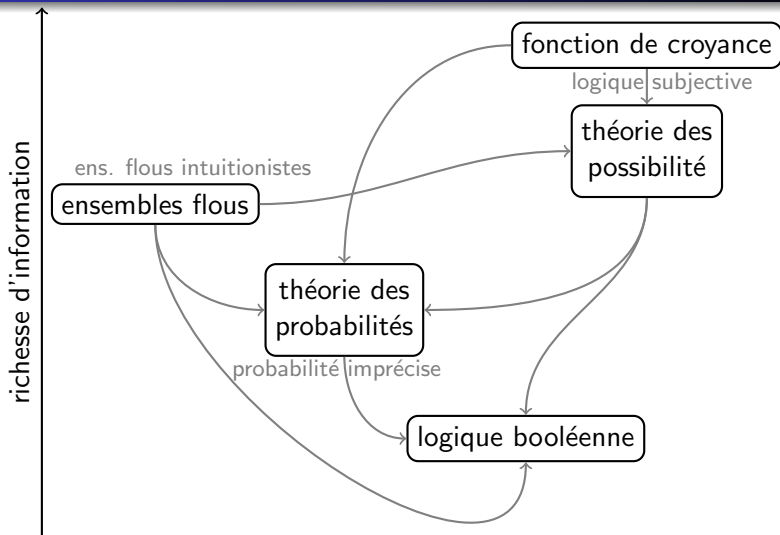
# Types de connaissances partielles



Conflit

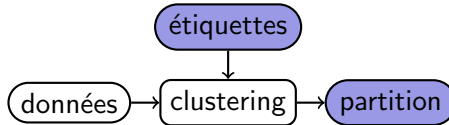
état 1 oestrus	état 2 boîterie	état 3 velage
état 4 blessure	état 5 mammite	state 6 apathie

# Modèles mathématiques pour les connaissances partielles



# Vers une définition plus flexible des étiquettes

## ① Travail sur le clustering semi-supervisé



[V. Antoine, J. Guerrero, G. Romero: *Possibilistic fuzzy c-means with partial supervision*. Fuzzy Sets Syst. 2022.]

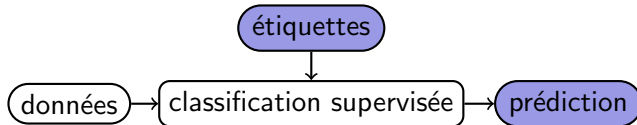
[V. Antoine, J. Guerrero, J. Xie: *Fast semi-supervised evidential clustering*. Int. J. Approx. Reason. 2021.]

[J. Xie, V. Antoine: *On a New Evidential C-Means Algorithm with Instance-Level Constraints*. SUM 2019.]

[V. Antoine, J. Guerrero, T. Boone, G. Romero: *Possibilistic clustering with seeds*. FUZZ-IEEE 2018.]

[V. Antoine, N. Labroche: *Semi-supervised Fuzzy c-Means Variants: A Study on Noisy Label Supervision*. IPMU 2018.]

## ② Travail sur la classification supervisée



[N. Wagner, V. Antoine, J. Koko, R. Lardy: *Fuzzy k-NN Based Classifiers for Time Series with Soft Labels*. IPMU 2020.]

présence d'incertitude



# Outline

- 1 Introduction
- 2 Clustering semi-supervisé et étiquettes imprécises
- 3 Classification avec étiquettes imprécises
- 4 Conclusion

# Clustering semi-supervisé

## Problématique du clustering

Aucune connaissance a priori

- comment définir la notion de similarité ?
- comment choisir une solution parmi plusieurs partition possible ?



# Clustering semi-supervisé

## Problématique du clustering

Aucune connaissance a priori

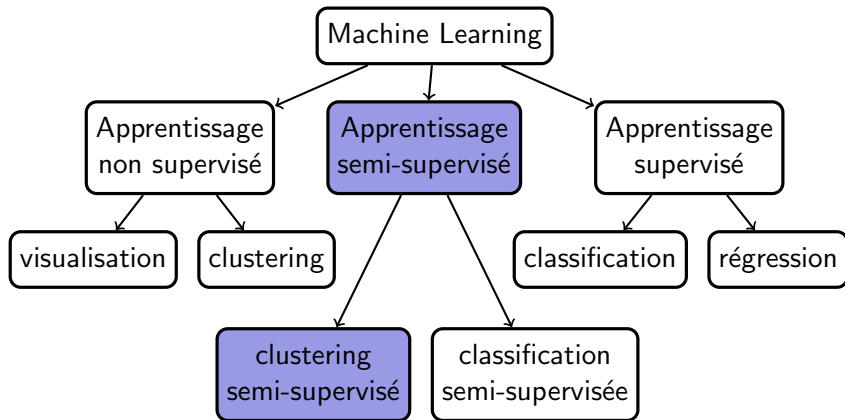
- comment définir la notion de similarité ?
- comment choisir une solution parmi plusieurs partition possible ?



## Information provenant de l'expert

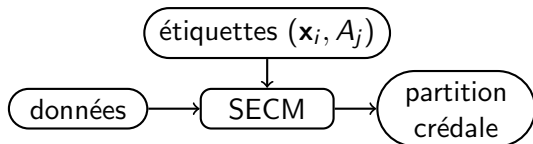
- étiquettes,
- contraintes par pair,
- classes équilibrées,...

# Clustering semi-supervisé



# Motivation

L'expert fournit des étiquettes imprécises  $A_j$



## Exemple d'annotation d'expert

$\omega_1$  pour les carrés,  $\omega_2$  pour les cercles,  $\omega_3$  pour les pentagones

	$\omega_1$	$\omega_2$	$\omega_3$	$A_j$
	✓	✗	✗	$\omega_1$
	✗	✓	✗	$\omega_2$
	?	?	✗	$\omega_{12} = \{\omega_1, \omega_2\}$

## Partition crédale $\mathbf{M} = (m_{ij})$

Chaque objet  $i$  a un degré de croyance  $m_i$  pour chaque sous- ensemble  $A_j \subseteq \Omega$

$$m_{ij} \in [0, 1], \sum_{A_j \subseteq \Omega} m_{ij} = 1$$

# Partition crédale $\mathbf{M} = (m_{ij})$

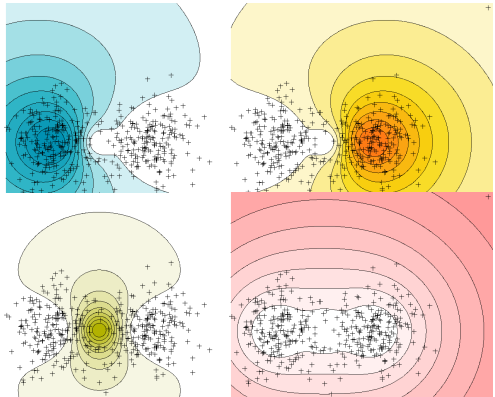
Chaque objet  $i$  a un degré de croyance  $m_i$  pour chaque sous- ensemble  $A_j \subseteq \Omega$

$$m_{ij} \in [0, 1], \sum_{A_j \subseteq \Omega} m_{ij} = 1$$

## Exemple

$\omega_1$  classe carré,  $\omega_2$  classe cercle

	$m_{i\emptyset}$	$m_{i\omega_1}$	$m_{i\omega_2}$	$m_{i\Omega}$
○	0	0	1	0
□	0	1	0	0
◻	0	0.9	0.1	0
◻	0	0	0	1
☆	1	0	0	0



# Cohérence entre étiquettes et partition crédale dure

	partition crédale							étiquette			
	$m_{i\omega_1}$	$m_{i\omega_2}$	$m_{i\omega_{12}}$	$m_{i\omega_3}$	$m_{i\omega_{13}}$	$m_{i\omega_{23}}$	$\Omega$	$A_j$			
○	1	0	0	0	0	0	0	$\omega_1$	++		
○	0	0	1	0	0	0	0	$\omega_1$	+		
○	0	0	0	0	0	0	1	$\omega_1$	=		
○	0	1	0	0	0	0	0	$\omega_1$	-		



# Cohérence entre étiquettes et partition crédale dure

	partition crédale							étiquette			
	$m_{i\omega_1}$	$m_{i\omega_2}$	$m_{i\omega_{12}}$	$m_{i\omega_3}$	$m_{i\omega_{13}}$	$m_{i\omega_{23}}$	$\Omega$	$A_j$			
OOOO	1	0	0	0	0	0	0	$\omega_1$	++		
	0	0	1	0	0	0	0	$\omega_1$	+		
	0	0	0	0	0	0	1	$\omega_1$	=		
	0	1	0	0	0	0	0	$\omega_1$	-		
DDDD	0	1	0	0	0	0	0	$\omega_{12}$	++		
	0	0	1	0	0	0	0	$\omega_{12}$	+		
	0	0	0	0	1	0	0	$\omega_{12}$	=		
	0	0	0	0	0	0	1	$\omega_{12}$	=		
	0	0	0	1	0	0	0	$\omega_{12}$	-		

## Cohérence entre étiquettes et partition crédale dure

	partition crédale							étiquette		r=1	
	$m_{i\omega_1}$	$m_{i\omega_2}$	$m_{i\omega_{12}}$	$m_{i\omega_3}$	$m_{i\omega_{13}}$	$m_{i\omega_{23}}$	$\Omega$	$A_j$		$T_{ij}$	
OOOO	1	0	0	0	0	0	0	$\omega_1$	++	1	
	0	0	1	0	0	0	0	$\omega_1$	+	1/2	
	0	0	0	0	0	0	1	$\omega_1$	=	1/3	
	0	1	0	0	0	0	0	$\omega_1$	-	0	
DDDD	0	1	0	0	0	0	0	$\omega_{12}$	++	1	
	0	0	1	0	0	0	0	$\omega_{12}$	+	$\sqrt{2}/2$	
	0	0	0	0	1	0	0	$\omega_{12}$	=	1/2	
	0	0	0	0	0	0	1	$\omega_{12}$	=	$\sqrt{2}/3$	
	0	0	0	1	0	0	0	$\omega_{12}$	-	0	

### Mesure de cohérence

$$T_{ij} = T_i(A_j) = \sum_{A_\ell \cap A_j \neq \emptyset} \frac{|A_j \cap A_\ell|^{r/2}}{|A_\ell|^r} m_{i\ell}, \quad r \geq 0 \text{ un hyperparamètre}$$

## Cohérence entre étiquettes et partition crédale dure

	partition crédale							étiquette		r=1	r=0
	$m_{i\omega_1}$	$m_{i\omega_2}$	$m_{i\omega_{12}}$	$m_{i\omega_3}$	$m_{i\omega_{13}}$	$m_{i\omega_{23}}$	$\Omega$	$A_j$		$T_{ij}$	$T_{ij}$
OOOO	1	0	0	0	0	0	0	$\omega_1$	++	1	1
	0	0	1	0	0	0	0	$\omega_1$	+	1/2	1
	0	0	0	0	0	0	1	$\omega_1$	=	1/3	1
	0	1	0	0	0	0	0	$\omega_1$	-	0	0
DDDD	0	1	0	0	0	0	0	$\omega_{12}$	++	1	1
	0	0	1	0	0	0	0	$\omega_{12}$	+	$\sqrt{2}/2$	1
	0	0	0	0	1	0	0	$\omega_{12}$	=	1/2	1
	0	0	0	0	0	0	1	$\omega_{12}$	=	$\sqrt{2}/3$	1
	0	0	0	1	0	0	0	$\omega_{12}$	-	0	0

### Mesure de cohérence

$$T_{ij} = T_i(A_j) = \sum_{A_\ell \cap A_j \neq \emptyset} \frac{|A_j \cap A_\ell|^{r/2}}{|A_\ell|^r} m_{i\ell}, \quad r \geq 0 \text{ un hyperparamètre}$$

# Étude de l'hyperparamètre $r$

	$m_{iw_1}$	$m_{iw_2}$	$m_{iw_{12}}$	$m_{iw_3}$	$m_{iw_{13}}$	$m_{iw_{23}}$	$\Omega$	$A_j$	$r=1, T_{ij}$		$r=0, T_{ij}$	
OO	1	0	0	0	0	0	0	$\omega_1$	++	1	+	1
OO	0	0	1	0	0	0	0	$\omega_1$	+	1/2	+	1
OO	0	0	0	0	0	0	1	$\omega_1$	=	1/3	+	1
OO	0	1	0	0	0	0	0	$\omega_1$	-	0	-	0
DD	0	1	0	0	0	0	0	$\omega_{12}$	++	1	+	1
DD	0	0	1	0	0	0	0	$\omega_{12}$	+	$\sqrt{2}/2$	+	1
DD	0	0	0	0	1	0	0	$\omega_{12}$	=	1/2	+	1
DD	0	0	0	0	0	0	1	$\omega_{12}$	=	$\sqrt{2}/3$	+	1
DD	0	0	0	1	0	0	0	$\omega_{12}$	-	0	-	0

## Mesure de cohérence

- $r = 0 \Rightarrow$  ne pénalise pas les sous-ensembles de grandes cardinalités. Utile en cas de bruit dans les étiquettes.
- $r > 0 \Rightarrow$  pénalise les sous-ensembles de grandes cardinalités. Étiquettes certaines.

# Semi-supervised evidential clustering: SECM

## Idée globale

Si  $\mathbf{x}_i \in A_j \Rightarrow T_{ij}$  doit être élevé

## Fonction objectif à minimiser

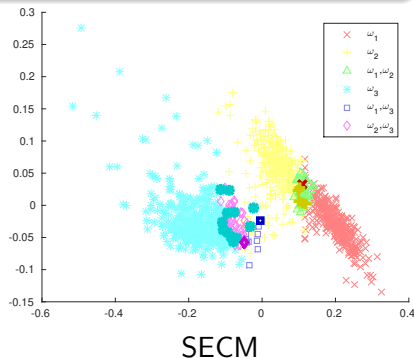
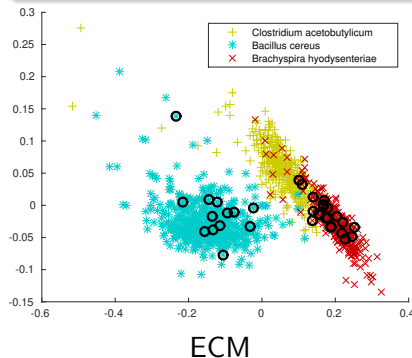
$$J_{SECM} = (1 - \gamma)J_{ECM} + \gamma \sum_{i=1}^n \sum_{A_j \subseteq \Omega, A_j \neq \emptyset} b_{ij}(1 - T_{ij})$$

tel que  $b_{ij} = \begin{cases} 1 & \text{si } \mathbf{x}_i \text{ est contraint avec } A_j, \\ 0 & \text{sinon.} \end{cases}$

# Application en génomique

## Jeu de données tetragen

Séquences d'ADN dont les plus grandes ont été divisées en plusieurs objets  $\Rightarrow$  génération d'étiquettes



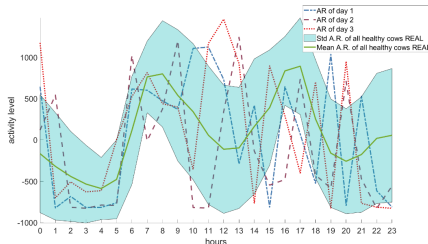
# Outline

- 1 Introduction
- 2 Clustering semi-supervisé et étiquettes imprécises
- 3 Classification avec étiquettes imprécises**
- 4 Conclusion

# Application agricole

## Objectif

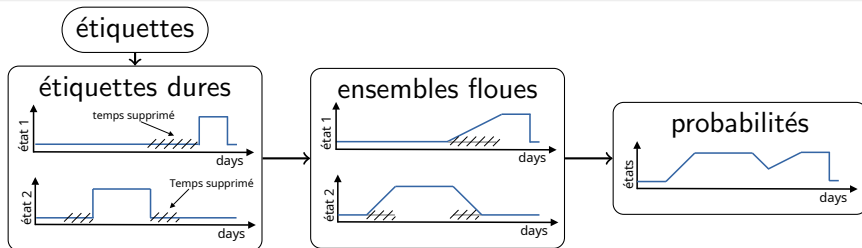
- Détection d'anomalies sur le niveau d'activité de vaches laitières
  - séries temporelles univariés par heure
  - grande variabilité intra-classes
  - étiquettes manuelles par jour



INRAE



# Gestion des étiquettes floues



## Connaissances a priori

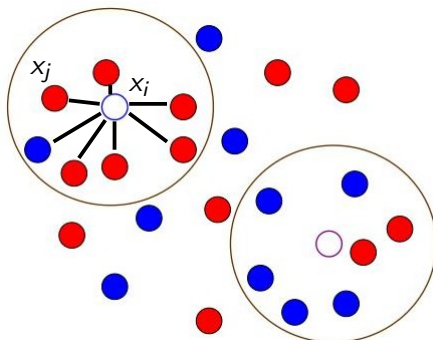
Type d'événement	zone floue avant	P = 1	zone floue après
rétentions placentaire	de -48 à 0	de 0 à 24	de 24 à 48
Velages	de -48 à 0	de 0 à 24	de 24 à 48
Oestrus	de -12 à 0	de 0 à 24	de 24 à 36
Boiteries	de -48 à -12	de -12 à 24	de 24 à 48
Mammites	de -48 à 0	de 0 à 24	de 24 à 48
Autres maladies	de -48 à 0	de 0 à 24	de 24 à 48
Injectons LPS	p = 0	de 12 à 24	de 24 à 48
Acidose	de -12 à 12	de 12 à 24	de 24 à 60
Changement parc	p = 0	de 12 à 24	de 24 à 48
Autre perturbations	p = 0	de 12 à 24	de 24 à 48

# Fuzzy kNN

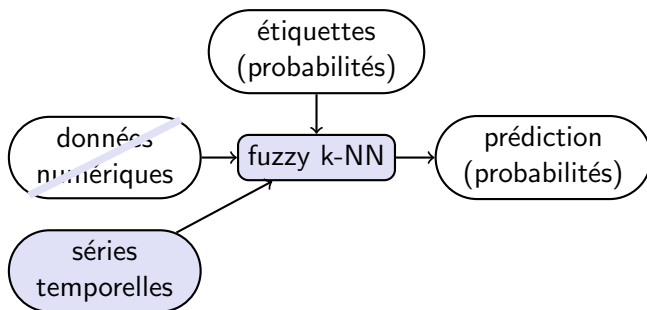
Soient

- $p_{jk}$  la probabilité pour  $x_j$  d'appartenir à la classe  $k$
- $d_{ij}$  la distance entre les points  $x_i$  et  $x_j$
- $\mathcal{V} = \{x_j, ..\}$  les  $k$  voisins de  $x_i$

$$p_{ik} = f_{x_j \in \mathcal{V}}(dist(x_i, x_j), p_{jk})$$



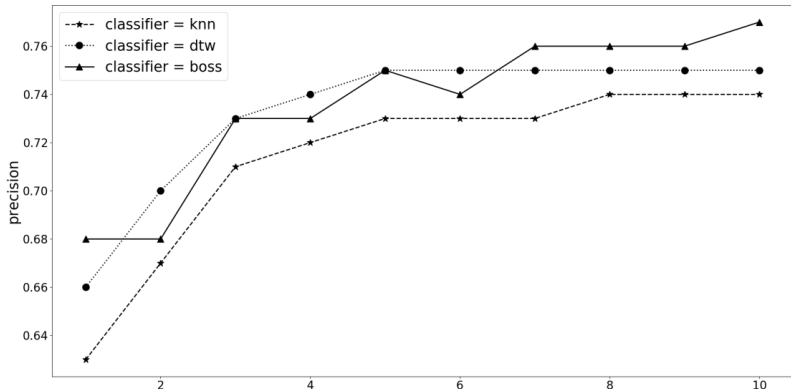
# Gestion d'une série temporelle avec Fuzzy kNN



## Modifications

- mesure DTW
- technique Bag of SFA Symbols (BOSS)

# Résultats



Précision obtenue par les classifieurs fuzzy k-NN, F-DTW et F-BOSS en fonction du nombre de voisins  $k$  avec la stratégie 3 et  $\mu = 0,3$  sur le jeu de données FreezerSmallTrain.

# Outline

- 1 Introduction
- 2 Clustering semi-supervisé et étiquettes imprécises
- 3 Classification avec étiquettes imprécises
- 4 Conclusion**

# Conclusion

## La modélisation des incertitudes en science des données

### Avantage

- permet une représentation plus proche de la réalité des données
- apporte une information riche en sortie
  - évite les erreurs
  - facilite l'interaction avec un expert
  - permet un post-traitement efficace

### Inconvénient

- peu adapté au temps réel
  - cas de prise de décision rapide et automatique
- la diffusion des incertitudes peut amener à trop d'imprécision

Merci