

INTRODUCTION À LA CLASSIFICATION AUTOMATIQUE

Yannick Estève



ÉCOLE THÉMATIQUE
BIG DATA SPEECH
ROSCOFF - JUILLET 2018

CLASSIFICATION AUTOMATIQUE

- ▶ Procédé qui consiste à catégoriser des objets en fonction de leur similarité ou leur dissimilarité
- ▶ Les catégories peuvent être définies *a priori* ou bien construites durant le processus de classification
 - ▶ dans ce dernier cas, le nombre de catégories peut être déterminé *a priori*, ou pas
- ▶ La notion de similarité dépend du type d'objets à traiter

CLASSIFICATION AUTOMATIQUE ET PAROLE

- ▶ Un grand nombre de tâches en traitement automatique de la parole et du langage sont abordées comme des problèmes de classification automatique
 - ▶ Exemples :
 - ▶ reconnaissance de phonèmes
 - ▶ déterminer à quel phonème prédéfini une portion du signal audio correspond
 - ▶ segmentation et regroupement en locuteurs
 - ▶ déterminer les portions de signal de parole qui se ressemblent et celles qui diffèrent (aucune information a priori sur les locuteurs ni leur nombre)
 - ▶ classification d'appels téléphoniques
 - ▶ reconnaissance de la langue

CLASSIFICATION AUTOMATIQUE ET APPRENTISSAGE AUTOMATIQUE

► Deux principaux scénarios :

1. Nous disposons d'un ensemble d'exemples comportant pour chaque individu observé une association avec un groupe d'individus. Les groupes d'individus possibles sont déjà connus.

- ➡ But : associer chaque nouvel individu à un de ces groupes
- ➡ Moyens : algorithmes d'apprentissage automatique supervisé

CLASSIFICATION AUTOMATIQUE ET APPRENTISSAGE AUTOMATIQUE

► Deux principaux scénarios :

1. Nous disposons d'un ensemble d'exemples comportant pour chaque individu observé une association avec un groupe d'individus. Les groupes d'individus possibles sont déjà connus.

- ➡ But : associer chaque nouvel individu à un de ces groupes
- ➡ Moyens : algorithmes d'apprentissage automatique supervisé

2. Nous disposons d'un ensemble d'individus ayant différentes caractéristiques

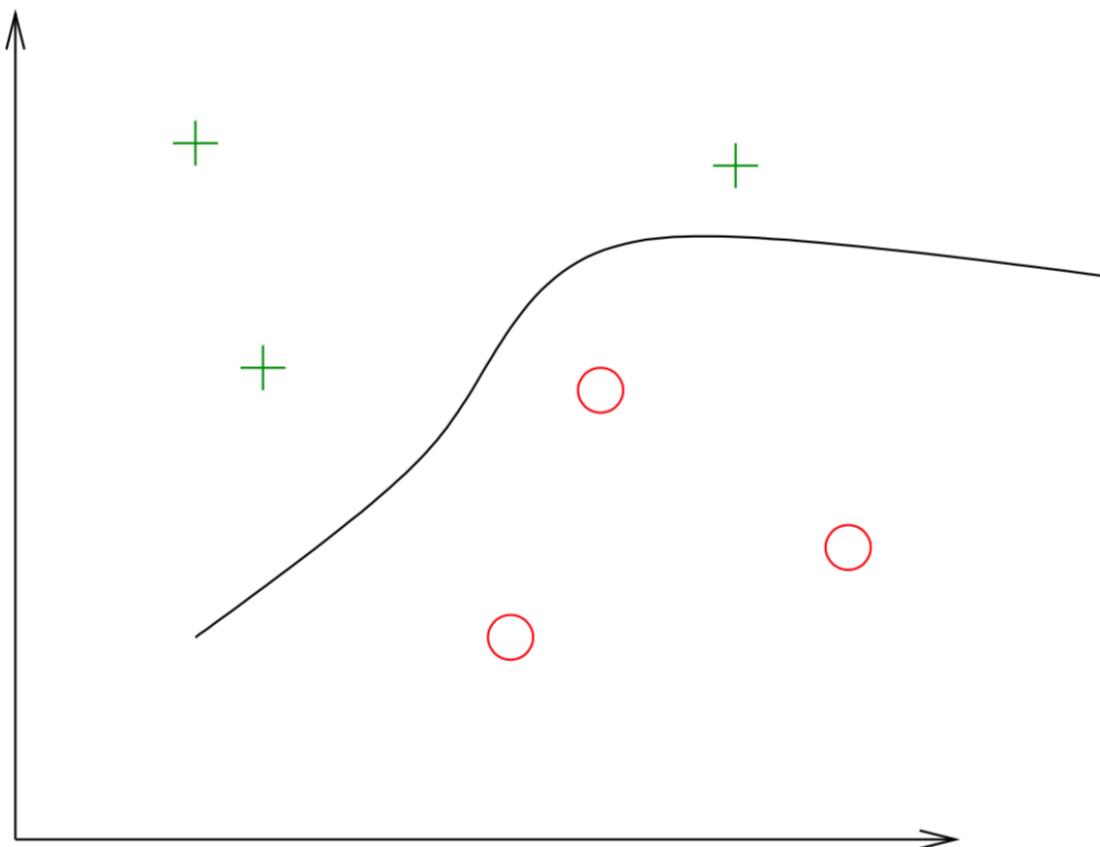
- ➡ But : regrouper ces individus au sein de groupe d'individus aux caractéristiques homogènes, et être capable d'appliquer ces regroupement à de nouveau individus
- ➡ Moyens : algorithmes de regroupement (*clustering*), apprentissage non supervisé

SCÉNARIO 1 : APPRENTISSAGE SUPERVISÉ ET CLASSIFICATION AUTOMATIQUE

- ▶ Soit une population divisée en q groupes homogènes d'individus différents
 - ▶ Pour chaque individu nous disposons des valeurs de p caractéristiques X_1, \dots, X_p
 - ▶ Nous ne savons pas quelles sont les valeurs de ces caractéristiques qui permettent d'affecter un individu dans un groupe
- ▶ Soit un nouvel individu w^* avec les valeurs X_{1*}, \dots, X_{p*} : à quel groupe appartient-il ?

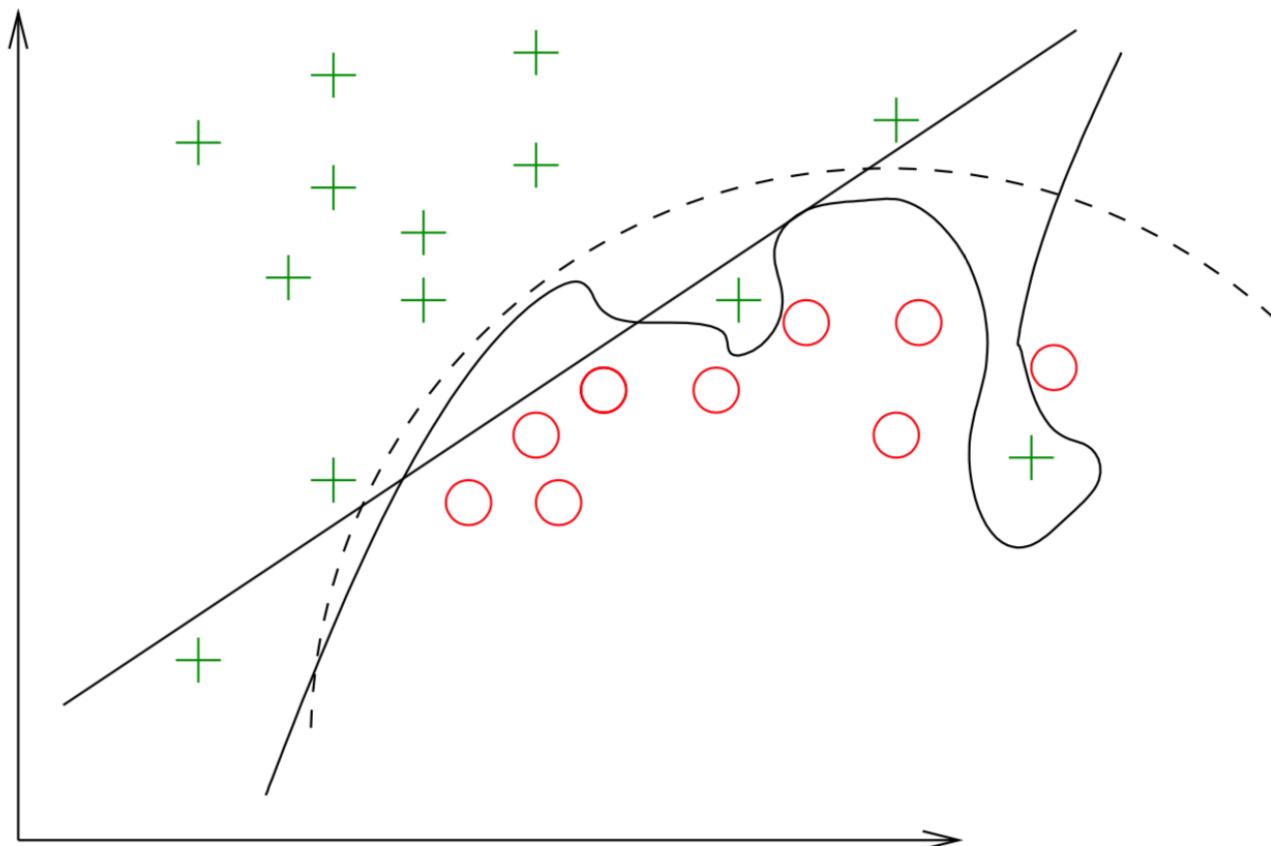
SCÉNARIO 1 : APPRENTISSAGE SUPERVISÉ ET CLASSIFICATION AUTOMATIQUE

- ▶ Soit une population divisée en $q=2$ groupes homogènes d'individus différents



SCÉNARIO 1 : APPRENTISSAGE SUPERVISÉ ET CLASSIFICATION AUTOMATIQUE

- ▶ Précision vs. surapprentissage



SCÉNARIO 1 : APPRENTISSAGE SUPERVISÉ ET CLASSIFICATION AUTOMATIQUE

Christophe Chesneau. Éléments de classification. Master. France. 2016. <cel-01252973v3>

Un professeur a décomposé sa classe en deux groupes de niveau.

Nous disposons de quelques informations concernant certains individus

	Maths	Physique	Ed Mus	Art Plas	Groupe
Boris	20	20	0	0	G1
Mohammad	8	8	12	12	G2
Stéphanie	20	20	0	0	G1
Jean	0	0	20	20	G2
Lilly	10	10	10	10	G1
Annabelle	2	2	18	18	G2

À quel groupe appartient Bob ?

	Maths	Physique	Ed Mus	Art Plas	Groupe
Bob	9	15	13	11	inconnu

→ Quelle est la probabilité que Bob appartienne à G1 (ou G2) ?

APPRENTISSAGE SUPERVISÉ ET CLASSIFICATION AUTOMATIQUE

► Exemple avec R

Contenu du fichier groupesclasse.txt :

i	x1	x2	x3	x4	y
1	20	20	0	0	G1
2	8	8	12	12	G2
3	20	20	0	0	G1
4	0	0	20	20	G2
5	10	10	10	10	G1
6	2	2	18	18	G2

	Maths	Physique	Ed Mus	Art Plas	Groupe
Boris	20	20	0	0	G1
Mohammad	8	8	12	12	G2
Stéphanie	20	20	0	0	G1
Jean	0	0	20	20	G2
Lilly	10	10	10	10	G1
Annabelle	2	2	18	18	G2

APPRENTISSAGE SUPERVISÉ ET CLASSIFICATION AUTOMATIQUE

- ▶ Exemple avec R
- ▶ Utilisation de SVM (*Support Vector Machine*)

```
#chargement des données
df <- read.table("groupesclasse.txt",header=T,sep=" ",row.names=1)

#calcul du modèle (apprentissage)
model <- svm(y ~ ., data = df)

#initialisation de la variable bob
bob <- data.frame(x1 = c(7), x2 = c(9), x3 = c(15), x4 = c(13), x5 = c(11))

#prédiction du groupe dans lequel Bob devrait être affecté
pred_bob <- predict(model,bob)

#affichage du groupe (la valeur de la variable pred_bob)
print(pred_bob)
```

1

G2



Bob devrait être affecté dans le groupe 2

Levels: G1 G2

APPRENTISSAGE SUPERVISÉ ET CLASSIFICATION AUTOMATIQUE

- ▶ Exemple avec R
- ▶ Utilisation de SVM (*Support Vector Machine*)

```
#chargement des données
df <- read.table("groupesclasse.txt",header=T,sep=" ",row.names=1)

#calcul du modèle (apprentissage)
model <- svm(y ~ ., data = df)

#initialisation de la variable bob
bob <- data.frame(x1 = c(7), x2 = c(9), x3 = c(15), x4 = c(13), x5 = c(11))

#prédiction du groupe dans lequel Bob devrait être affecté
pred_bob <- predict(model,bob)

#affichage des probabilités pour chaque groupe
head(attr(pred_bob, "probabilities"))
```

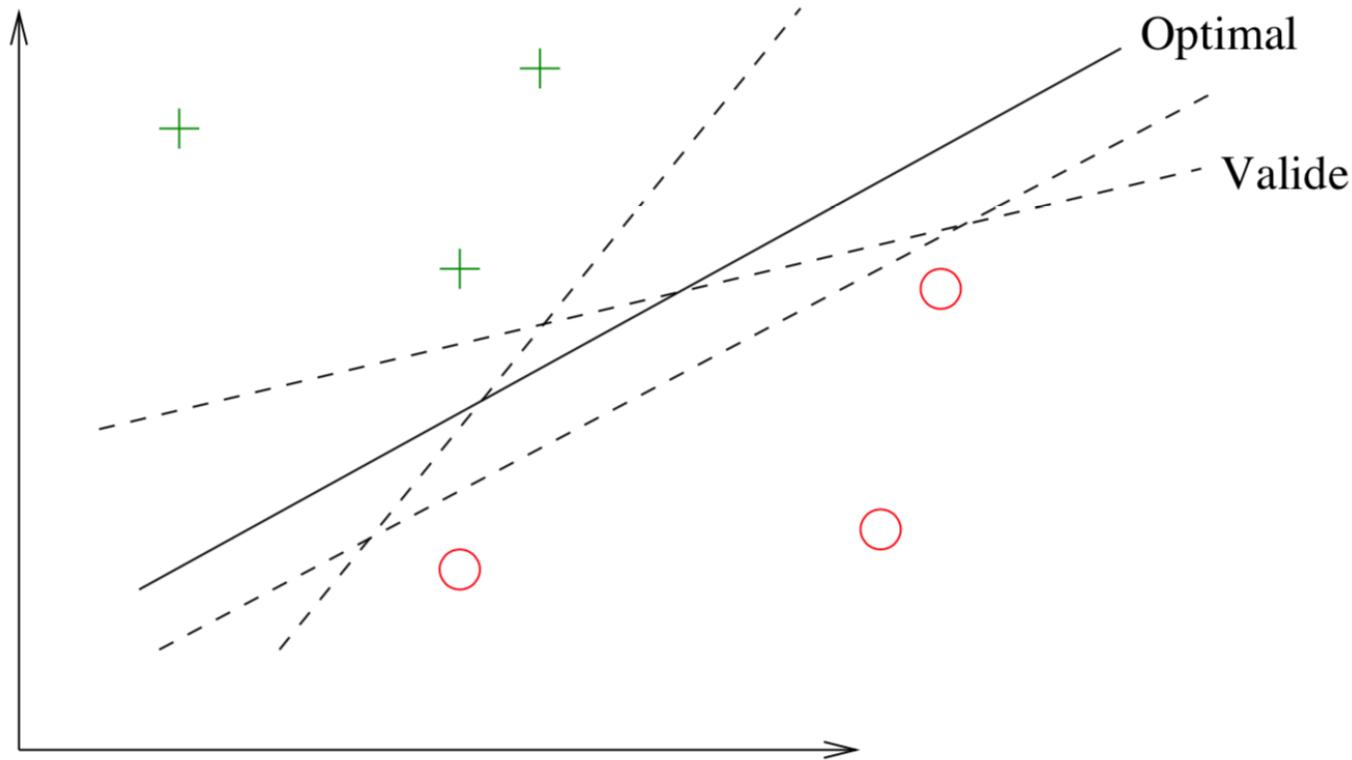
	G1	G2
1	0.4207464	0.5792536



Les probabilités d'affectation de Bob dans chacun des groupes

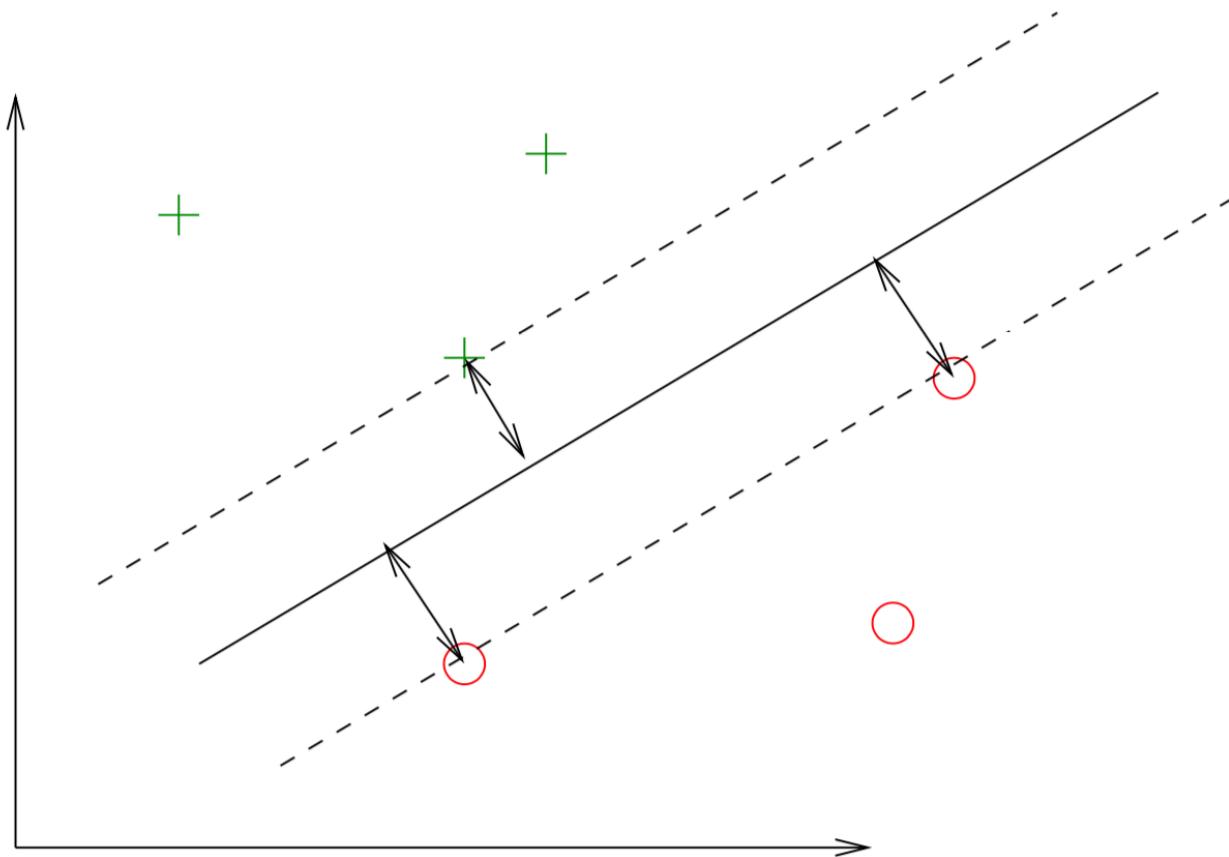
APPRENTISSAGE SUPERVISÉ ET CLASSIFICATION AUTOMATIQUE

► SVM : introduction intuitive



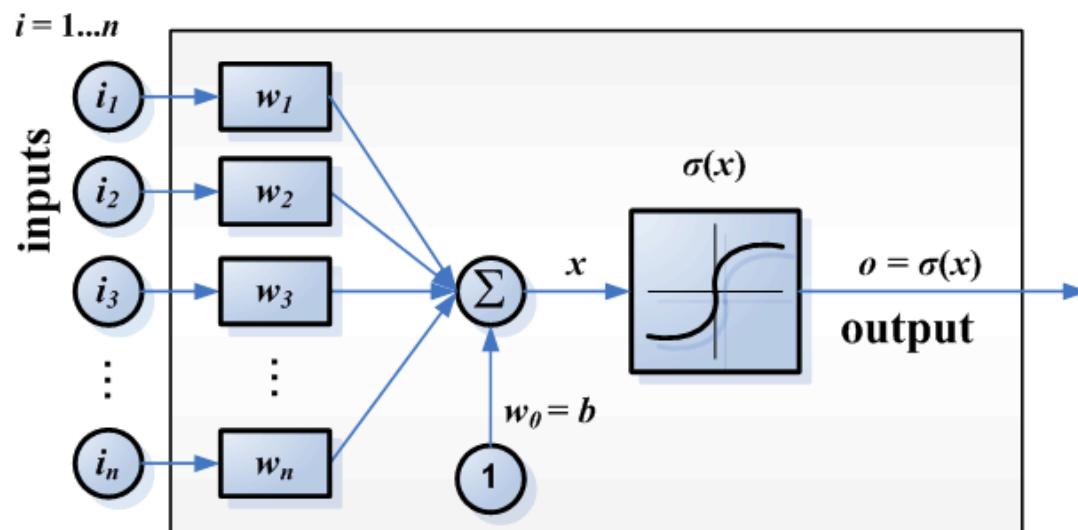
APPRENTISSAGE SUPERVISÉ ET CLASSIFICATION AUTOMATIQUE

► SVM : introduction intuitive



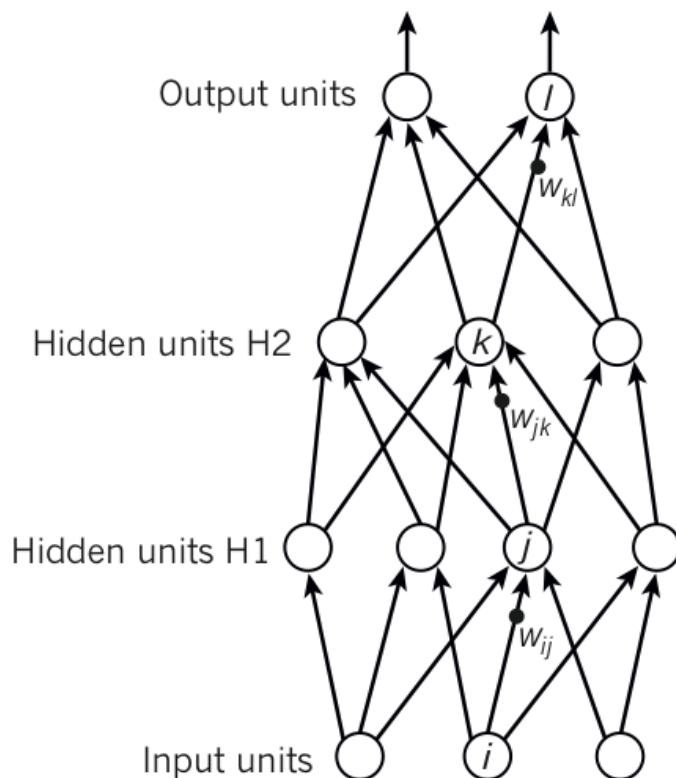
APPRENTISSAGE SUPERVISÉ ET CLASSIFICATION AUTOMATIQUE

- ▶ Réseau de neurones : introduction
 - ▶ un neurone artificiel : une simple fonction mathématique



APPRENTISSAGE SUPERVISÉ ET CLASSIFICATION AUTOMATIQUE

► Réseau de neurones : introduction



$$y_l = f(z_l)$$

$$z_l = \sum_{k \in H2} w_{kl} y_k$$

$$y_k = f(z_k)$$

$$z_k = \sum_{j \in H1} w_{jk} y_j$$

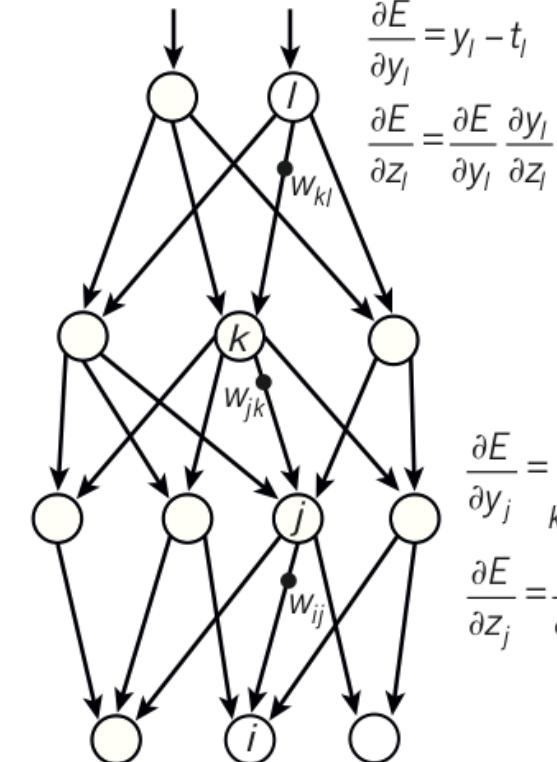
$$y_j = f(z_j)$$

$$z_j = \sum_{i \in \text{Input}} w_{ij} x_i$$

Compare outputs with correct answer to get error derivatives

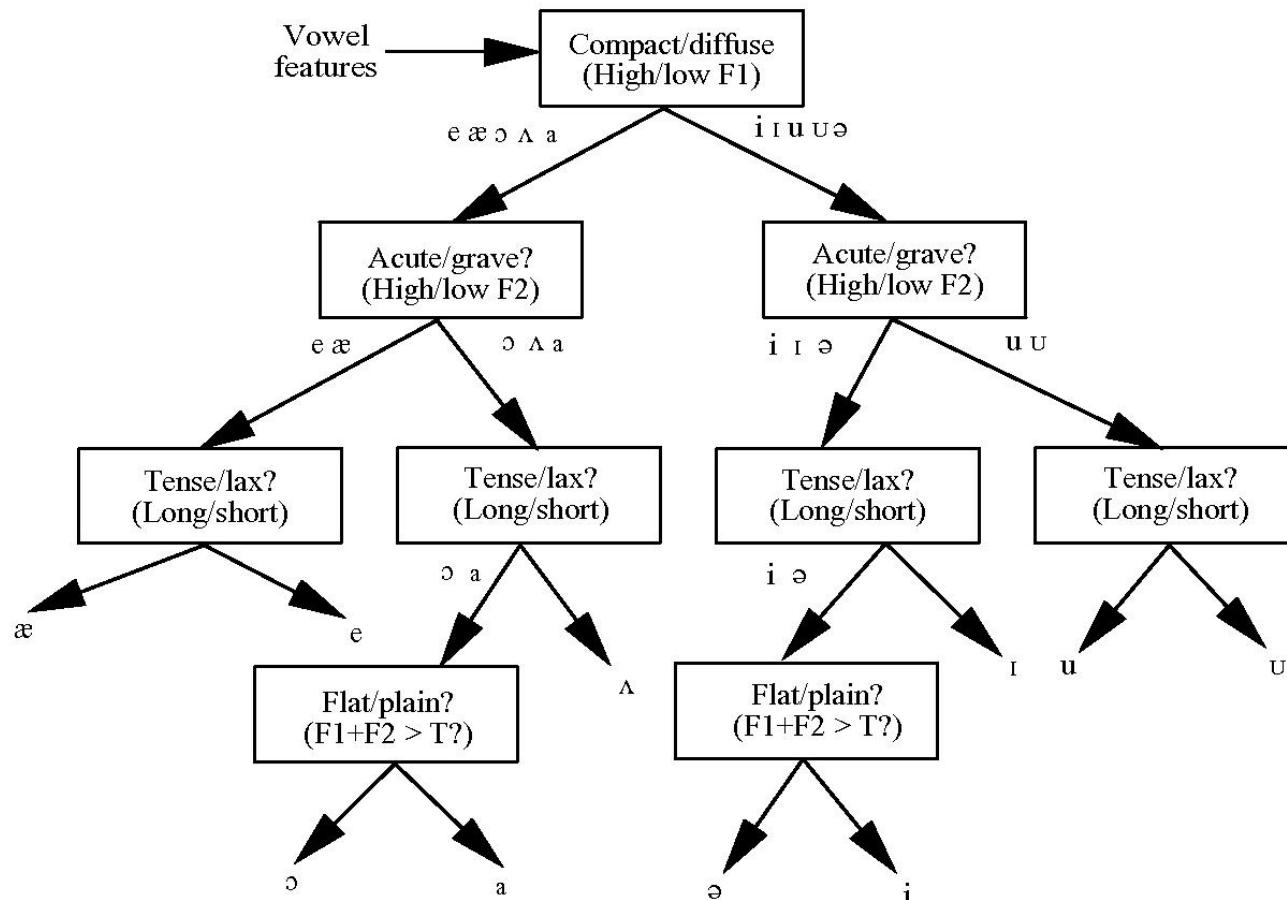
$$\frac{\partial E}{\partial y_k} = \sum_{l \in \text{out}} w_{kl} \frac{\partial E}{\partial z_l}$$

$$\frac{\partial E}{\partial z_k} = \frac{\partial E}{\partial y_k} \frac{\partial y_k}{\partial z_k}$$



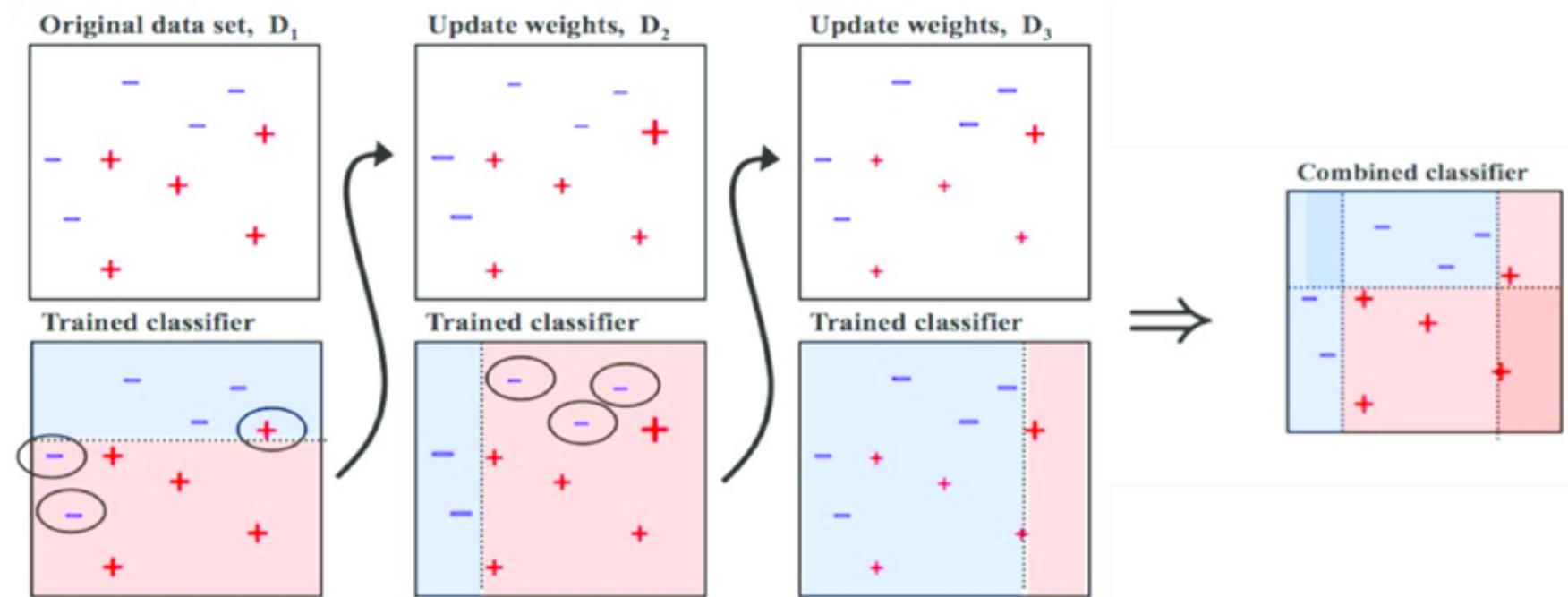
APPRENTISSAGE SUPERVISÉ ET CLASSIFICATION AUTOMATIQUE

- ▶ Autres approches de classification utilisables (non exhaustif) :
- ▶ Arbres de classification automatique



APPRENTISSAGE SUPERVISÉ ET CLASSIFICATION AUTOMATIQUE

- ▶ Autres approches de classification utilisables (non exhaustif) :
- ▶ AdaBoost : utilisation de classifieurs simples (e.g arbres binaires)

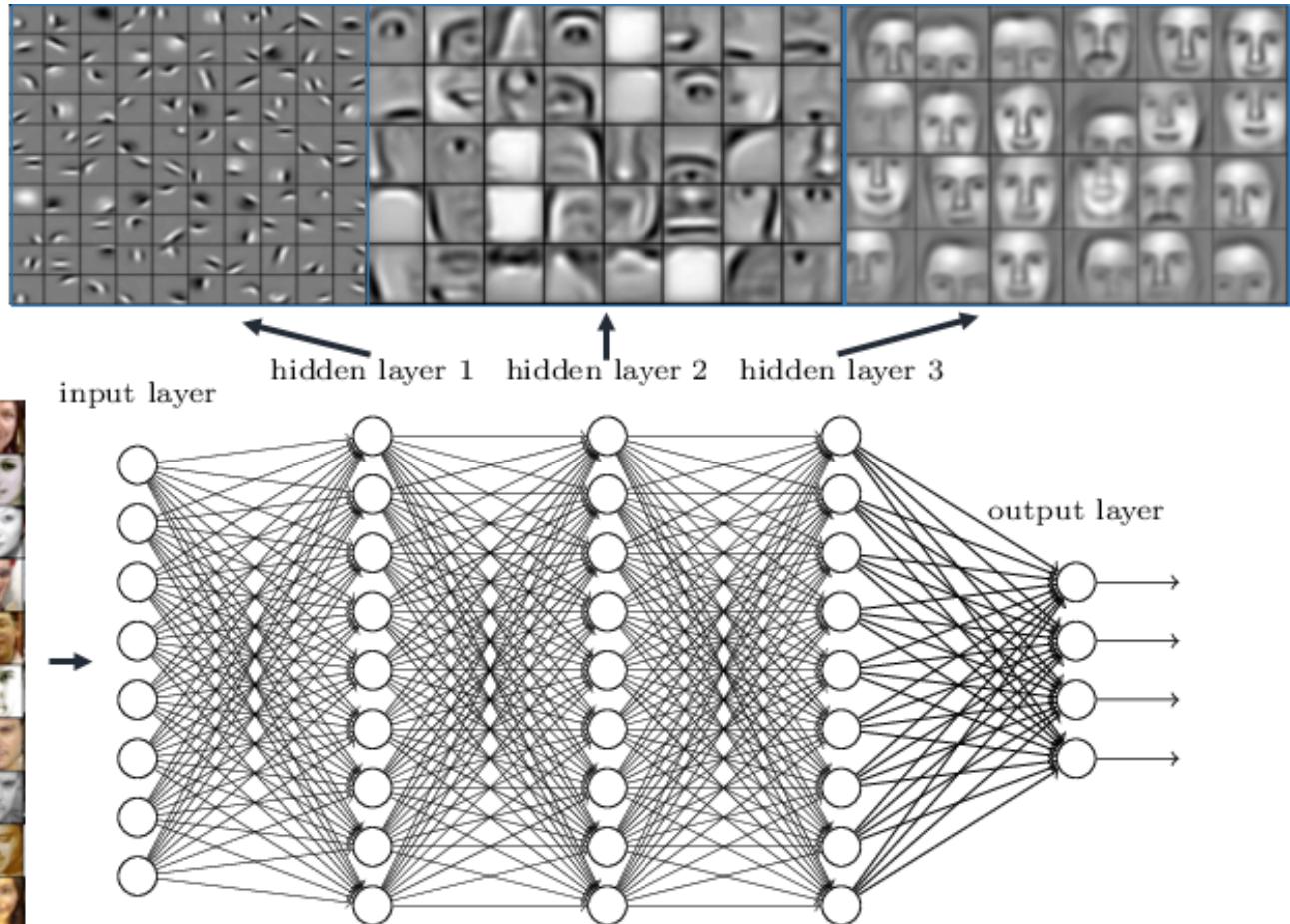


CLASSIFICATION AUTOMATIQUE NON SUPERVISÉE (CLUSTERING)

- ▶ Performance vs Interprétabilité
 - ▶ Arbre de classification
 - ▶ Adaboost
 - ▶ SVM
 - ▶ Réseau de neurones ?

REPRÉSENTATION INTERMÉDIAIRE D'IMAGES

Deep neural networks learn hierarchical feature representations



APPRENTISSAGE SUPERVISÉ ET CLASSIFICATION AUTOMATIQUE

- ▶ Métriques d'évaluation
- ▶ **Précision**, Rappel, f-mesure

La **précision** permet de répondre à la question suivante :

Quelle proportion d'identifications positives était effectivement correcte ?

La précision peut être définie comme suit :

$$\text{Précision} = \frac{VP}{VP + FP}$$

APPRENTISSAGE SUPERVISÉ ET CLASSIFICATION AUTOMATIQUE

- ▶ Métriques d'évaluation
- ▶ Précision, **Rappel**, f-mesure

Le **rappel** permet de répondre à la question suivante :

Quelle proportion de résultats positifs réels a été identifiée correctement ?

Mathématiquement, le rappel est défini comme suit :

$$\text{Rappel} = \frac{VP}{VP + FN}$$

APPRENTISSAGE SUPERVISÉ ET CLASSIFICATION AUTOMATIQUE

- ▶ Métriques d'évaluation
 - ▶ Précision, Rappel, **f-mesure**

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

APPRENTISSAGE SUPERVISÉ ET CLASSIFICATION AUTOMATIQUE

- ▶ Protocole expérimental : train, dev, test
 - ▶ limiter les biais
 - ▶ apprentissage sur un corpus (train)
 - ▶ optimisation sur un autre corpus (dev)
 - ▶ test final sur test

SCÉNARIO 2 : CLASSIFICATION AUTOMATIQUE NON SUPERVISÉE (CLUSTERING)

- ▶ On considère n individus $\Gamma = \{\omega_1, \dots, \omega_n\}$ extraits au hasard d'une population. Pour chacun d'entre eux, on dispose de p valeurs de p caractères X_1, \dots, X_p .

- ▶ Les données sont de la forme :

	X_1	\dots	X_p
ω_1	$x_{1,1}$	\dots	$x_{p,1}$
\vdots	\vdots	\dots	\vdots
ω_n	$x_{1,n}$	\dots	$x_{p,n}$

- ▶ où, pour tout $(i,j) \in \{1, \dots, n\} \times \{1, \dots, p\}$, $x_{j,i} = X_j(\omega_i)$ est l'observation du caractère X_j sur l'individu ω_i .

SCÉNARIO 2 : CLASSIFICATION AUTOMATIQUE NON SUPERVISÉE (CLUSTERING)

- ▶ Objectif : Partant des données, l'objectif est de regrouper/ classer les individus qui se ressemblent le plus/qui ont des caractéristiques semblables.

	X_1	\dots	X_p
ω_1	$x_{1,1}$	\dots	$x_{p,1}$
\vdots	\vdots	\dots	\vdots
ω_n	$x_{1,n}$	\dots	$x_{p,n}$

SCÉNARIO 2 : CLASSIFICATION AUTOMATIQUE NON SUPERVISÉE (CLUSTERING)

- ▶ Méthodes : Pour atteindre l'objectif, plusieurs méthodes sont possibles. Parmi elles, il y a
 - ▶ l'algorithme de Classification Ascendante Hiérarchique (CAH),
 - ▶ l'algorithme des centres mobiles,
 - ▶ ...

	X_1	\dots	X_p
ω_1	$x_{1,1}$	\dots	$x_{p,1}$
\vdots	\vdots	\dots	\vdots
ω_n	$x_{1,n}$	\dots	$x_{p,n}$

CLASSIFICATION AUTOMATIQUE NON SUPERVISÉE (CLUSTERING)

- ▶ Généralement, approches qui s'appuie sur la notion de distance
 - ▶ approches aussi dites 'par modèles de distance'
- ▶ On appelle distance sur un ensemble M toute application $d : M^2 \rightarrow [0, \infty[$ telle que
 - pour tout $(x, y) \in M^2$, on a **$d(x, y) = 0$ si et seulement si $x = y$**
 - pour tout $(x, y) \in M^2$, on a **$d(x, y) = d(y, x)$**
 - pour tout $(x, y, z) \in M^3$, on a **$d(x, y) \leq d(x, z) + d(z, y)$** .

CLASSIFICATION AUTOMATIQUE NON SUPERVISÉE (CLUSTERING)

► Exemples de distances

Exemple 1 : distance euclidienne : Soient $m \in \mathbb{N}^*$, $x = (x_1, \dots, x_m) \in \mathbb{R}^m$ et $y = (y_1, \dots, y_m) \in \mathbb{R}^m$. On appelle distance euclidienne entre x et y la distance :

$$d(x, y) = \sqrt{\sum_{i=1}^m (x_i - y_i)^2}.$$

Exemple 2 : distance de Manhattan : Soient $m \in \mathbb{N}^*$, $x = (x_1, \dots, x_m) \in \mathbb{R}^m$ et $y = (y_1, \dots, y_m) \in \mathbb{R}^m$. On appelle distance de Manhattan entre x et y la distance :

$$d(x, y) = \sum_{i=1}^m |x_i - y_i|.$$

CLASSIFICATION AUTOMATIQUE NON SUPERVISÉE (CLUSTERING)

- ▶ Algorithme des centres mobiles (k means) :
 - ▶ L'algorithme des centres mobiles vise à classer une population Γ en q classes. Cela se fait de manière automatique ; il n'y a pas de lien hiérarchique dans les regroupements contrairement à l'algorithme CAH. Il est le mieux adapté aux très grands tableaux de données.

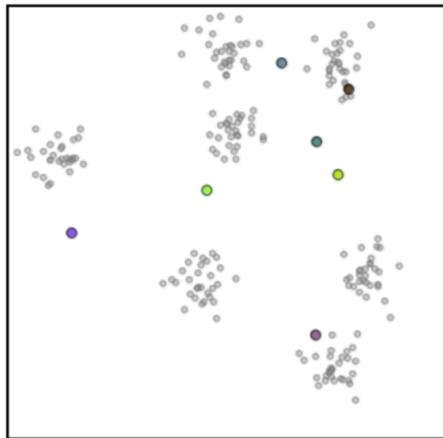
CLASSIFICATION AUTOMATIQUE NON SUPERVISÉE (CLUSTERING)

L'algorithme des centres mobiles avec la méthode de Lloyd (la plus standard) est décrit ci-dessous :

- On choisit q points au hasard dans \mathbb{R}^p . Ces points sont appelés centres.
- On calcule le tableau de distances entre tous les individus et les q centres.
- On forme alors q groupes de la manière suivante : chaque groupe est constitué d'un centre et des individus les plus proches de ce centre que d'un autre. On obtient une partition \mathcal{P}_1 de Γ .
- On calcule le centre de gravité de chacun des q sous-nuages de points formés par les q groupes. Ces q centres de gravité sont nos nouveaux q centres.
- On calcule le tableau de distances entre tous les individus et les nouveaux q centres.
- On forme alors q groupes, chaque groupe étant constitué d'un centre et des individus les plus proches de ce centre que d'un autre. On a une nouvelle partition \mathcal{P}_2 de Γ .
- On itère la procédure précédente jusqu'à ce que deux itérations conduisent à la même partition.

CLASSIFICATION AUTOMATIQUE NON SUPERVISÉE (CLUSTERING)

(données issues de 7 lois normales bidimensionnelles, classification avec 7 centres)

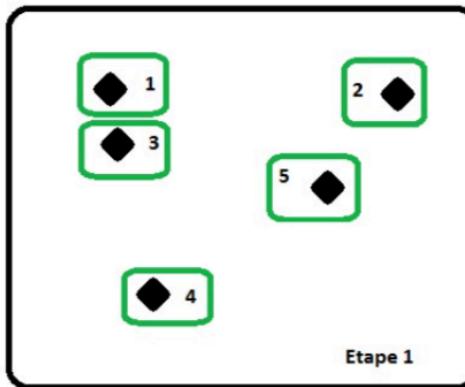


CLASSIFICATION AUTOMATIQUE NON SUPERVISÉE (CLUSTERING)

- ▶ Algorithme des centres mobiles (k means)
- ▶ Remarque importante : La classification des individus dépend du choix des centres initiaux. Plusieurs méthodes existent pour choisir judicieusement ces centres.

CLASSIFICATION AUTOMATIQUE NON SUPERVISÉE (CLUSTERING)

- ▶ Principes de la Classification Ascendante Hiérarchiques :
 - ▶ processus itératif
 - ▶ regroupe les 2 éléments les plus proches à chaque itération



DÉTECTION ET CARACTÉRISATION DE LA PAROLE SPONTANÉE

- ▶ Projet EPAC
- ▶ Segments de parole + transcription automatique
- ▶ Traits caractéristiques :
 - ▶ 'prosodiques' : durée de voyelle, débit phonémique, ...
 - ▶ 'linguistiques' : répétition de mots, sacs de n-grams, ...
 - ▶ 'ASR' : mesure de confiance

DÉTECTION ET CARACTÉRISATION DE LA PAROLE SPONTANÉE

TABLE II
 PRECISION AND RECALL IN THE CLASSIFICATION OF THE SPEECH
 SEGMENTS ACCORDING TO 3 CATEGORIES: *prepared speech, low
 spontaneity AND high spontaneity*

prepared speech					
Feat.	ling(ref)	ling(asr)	acou(asr)	all(asr)	all+global(asr)
Prec.	56	53.0	56.3	57.8	62.1
Recall	64.1	61.8	58.3	61.7	64.2
low spontaneous					
Feat.	ling(ref)	ling(asr)	acou(asr)	all(asr)	all+global(asr)
Prec.	43.8	40.7	44.0	45.5	49.2
Recall	37.7	31.7	41.3	40.5	44.2
high spontaneous					
Feat.	ling(ref)	ling(asr)	acou(asr)	all(asr)	all+global(asr)
Prec.	65.2	58.0	59.7	65.5	69.3
Recall	65.9	62.4	61.6	68.8	74.6

TRAITEMENT DE SÉQUENCES

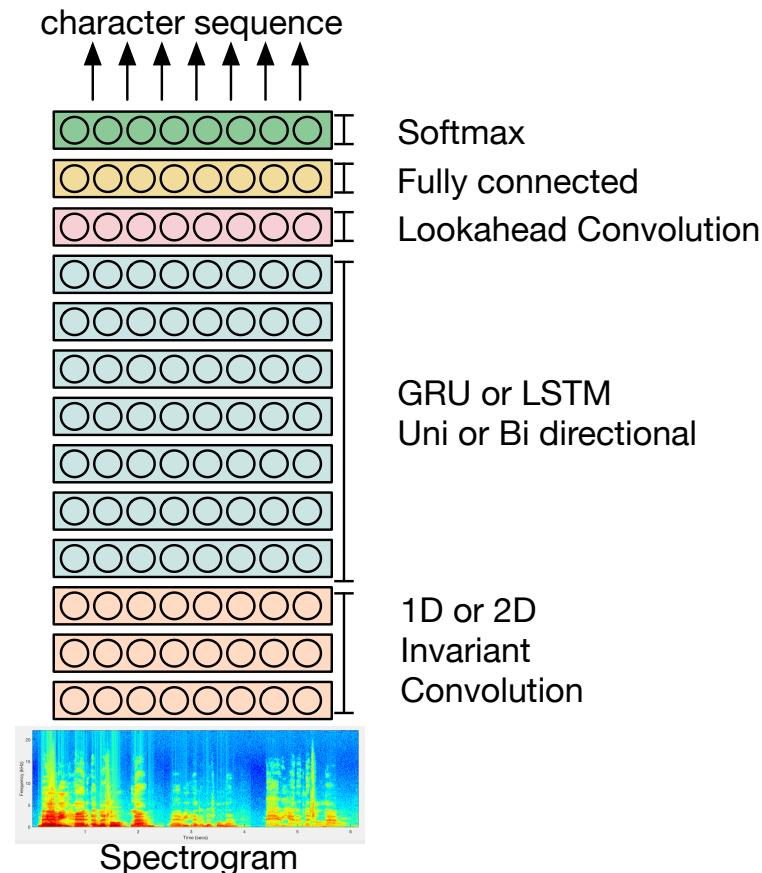
- ▶ Reconnaissance de la parole : HMM, mais aussi pur neuronal
- ▶ Reconnaissance d'entités nommées, par exemple utilisation de CRF

TRAITEMENT DE SÉQUENCES

word sequence le sculpteur césar est mort hier à paris à l' âge de soixante dix sept ans

word sequence
with EN tags le sculpteur <**pers** césar > est mort <**time** hier > à <**loc** paris > à l' âge de
<**amount** soixante dix sept ans >

starred mode:
only EN tags, EN
values, and stars * <**pers** césar > * <**time** hier > * <**loc** paris > * <**amount** soixante dix sept ans >



MERCI !

- ▶ Christophe Chesneau. Éléments de classification. Master. France. 2016. <cel-01252973v3>