

Explicabilité de l'intelligence artificielle

Romain Giot

université
de **BORDEAUX**

Préambule

**Description de la
présentation**

Romain Giot – MCF en informatique

IUT de Bordeaux / LaBRI

Authentification biométrique

rides du doigt, dynamique de frappe, modèles adaptatifs, multi-biométrie

Villes intelligentes

prédiction d'usage des systèmes de vélos libre service, segmentation d'images satellites

Analyse de documents et d'images

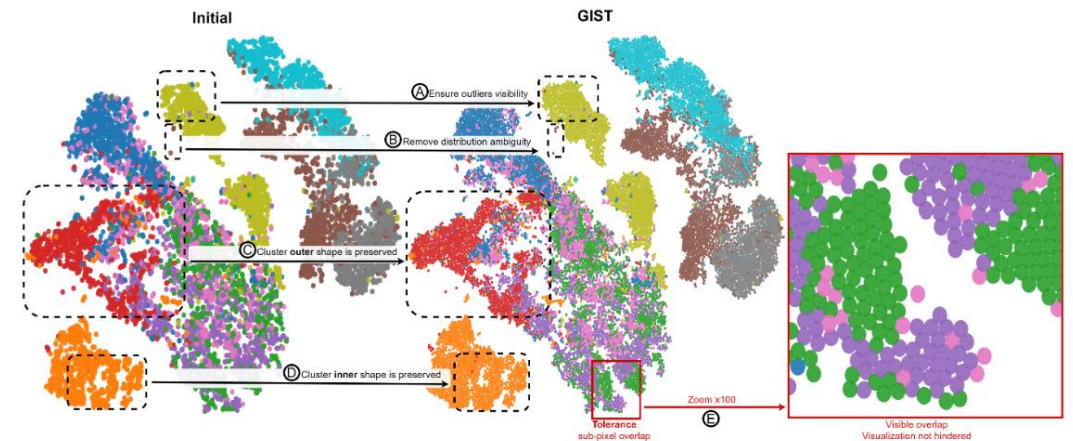
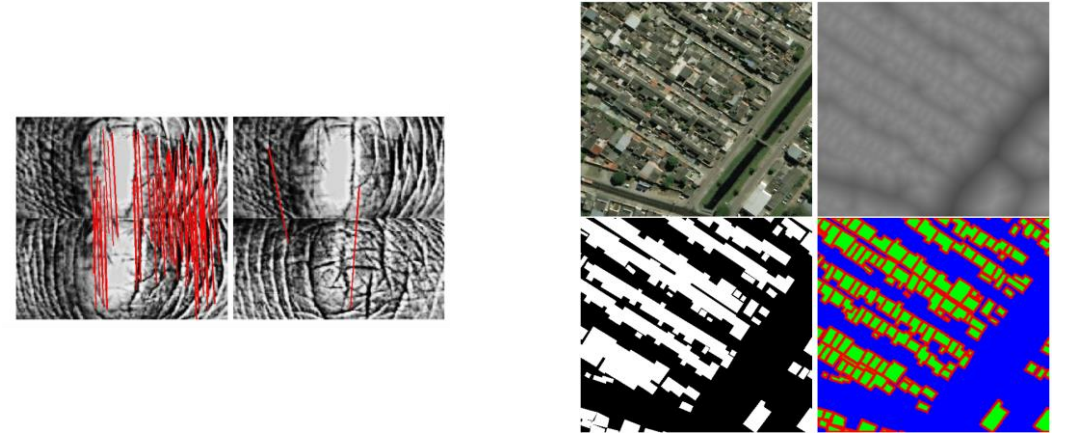
classification hiérarchique, fusion d'informations

Intelligence artificielle pour la visualisation d'informations

dessin de graphes, suppression de chevauchements, analyse de représentations visuelles

Visualisation d'informations pour l'intelligence artificielle

Explicabilité, détection d'erreurs dans les données



Organisation de la présentation

1. Introduction à l'intelligence artificielle
2. Généralités sur l'apprentissage supervisé
3. Evaluation de l'apprentissage supervisé
4. Exemples de modèles de classification
5. Explicabilité intrinsèque
6. Explicabilité après-coup
7. Sélection de travaux liés à l'explicabilité

Interrompez-moi quand vous voulez

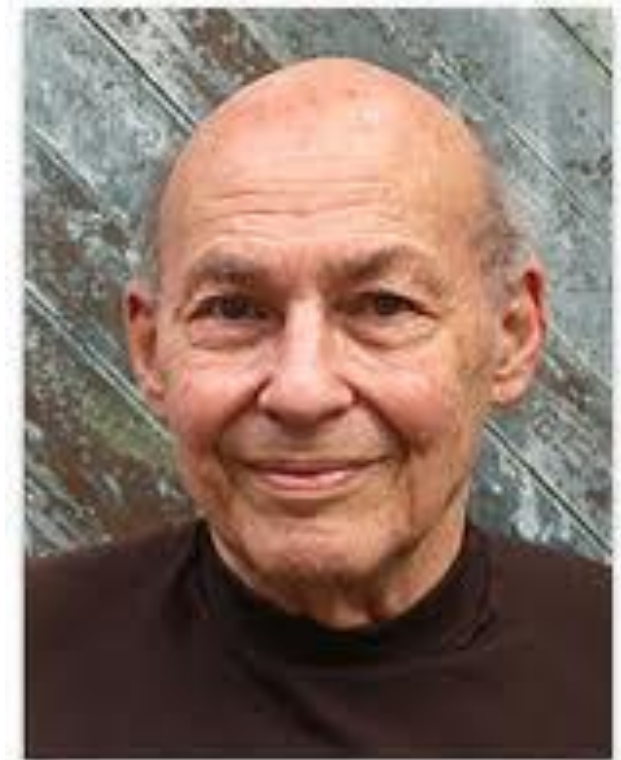
Introduction à l'intelligence artificielle

**Survol de
l'historique de
l'intelligence
artificielle**

Définition d'Intelligence Artificielle

La construction de programmes informatiques qui s'adonnent à des tâches qui sont, pour l'instant, accomplies de façon plus satisfaisante par des êtres humains car elles demandent des processus mentaux de haut niveau tels que : l'apprentissage perceptuel, l'organisation de la mémoire et le raisonnement critique.

Marvin Lee Misky



<https://history.computer.org/pioneers/minsky.html>

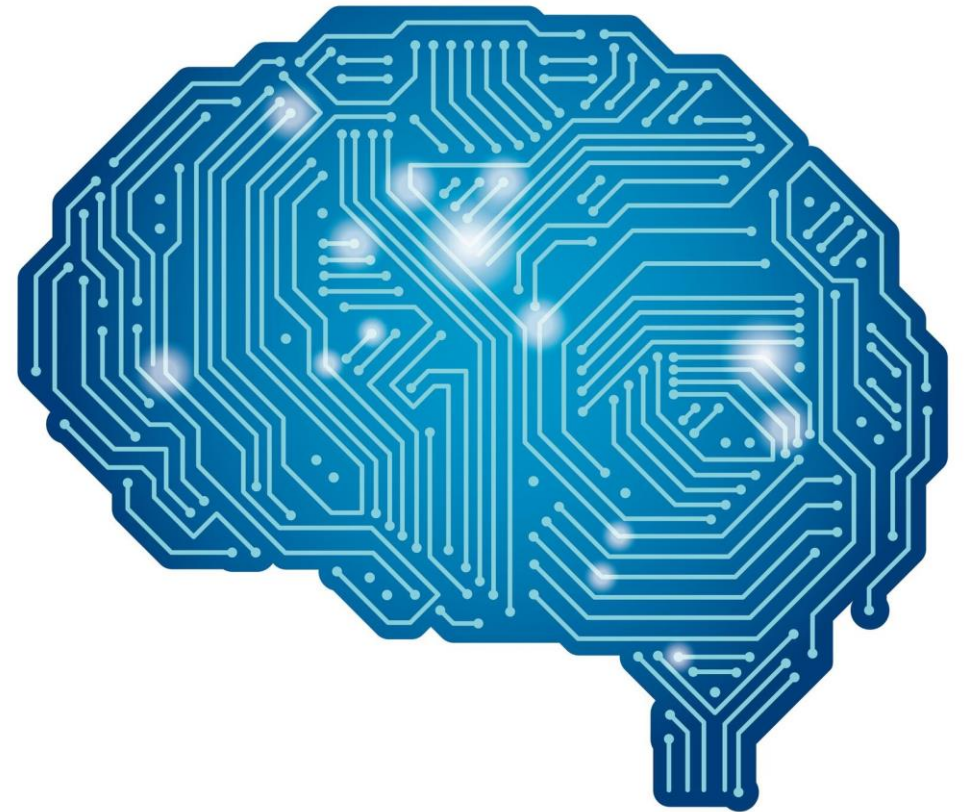
Dans les faits

Avant 2013

Systèmes symboliques, optimisation basée sur des heuristiques, algorithmes codés en dur

Depuis 2013

Apprentissage profond



1666 - *Calculus ratiocinator* de Gottfried Wilhelm Leibniz

Idée

une **méthode** qui permettrait de **démêler le vrai du faux** dans toute discussion dont les termes seraient exprimés dans une **langue philosophique universelle**

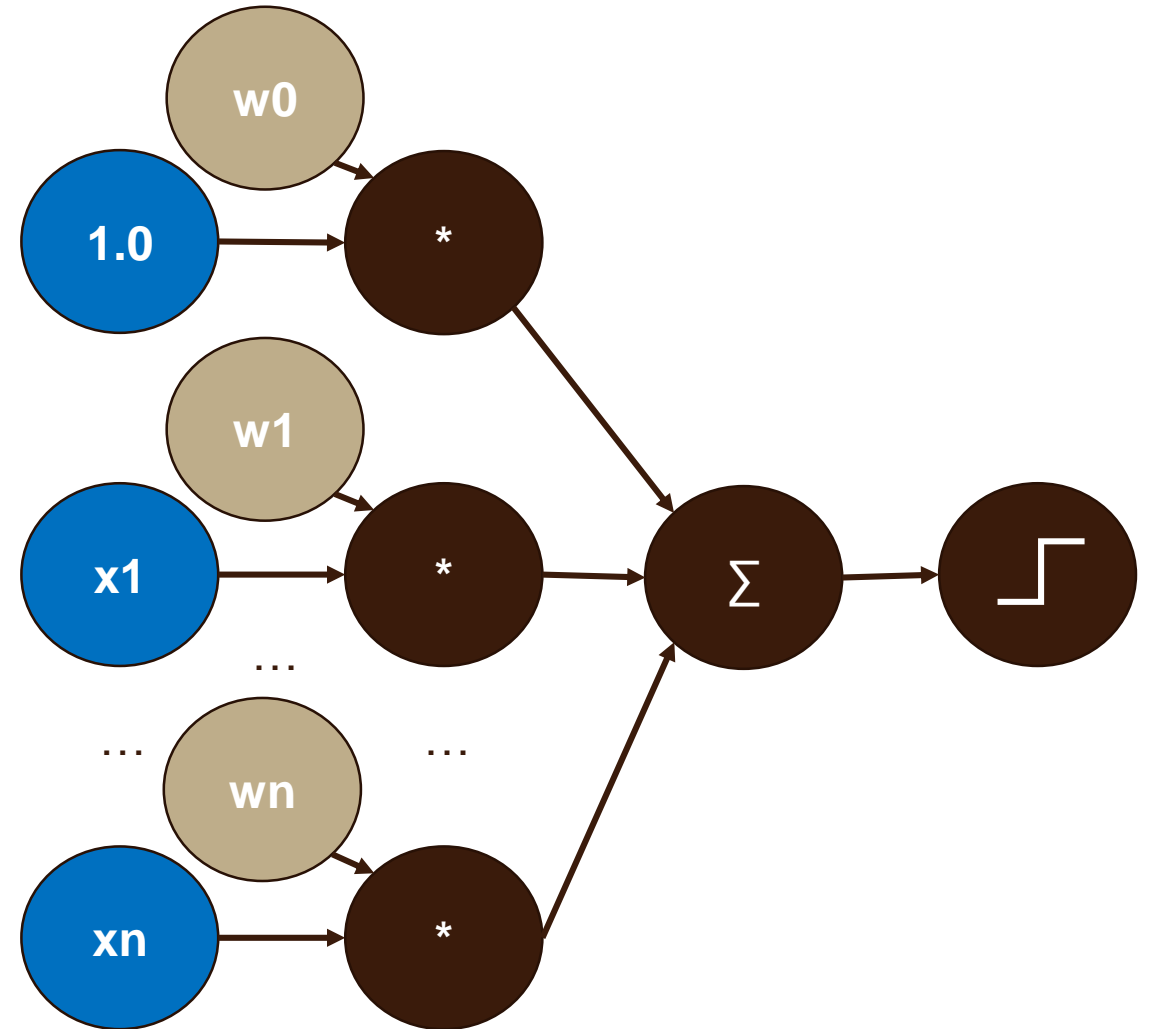
Interprétation

- méthode : algorithme implémenté sur une machine
- démêler le vrai du faux : intelligence artificielle
- langue philosophique universelle : langage de programmation



1943 – Neurone formel de Warren McCulloch et Walter Pitts

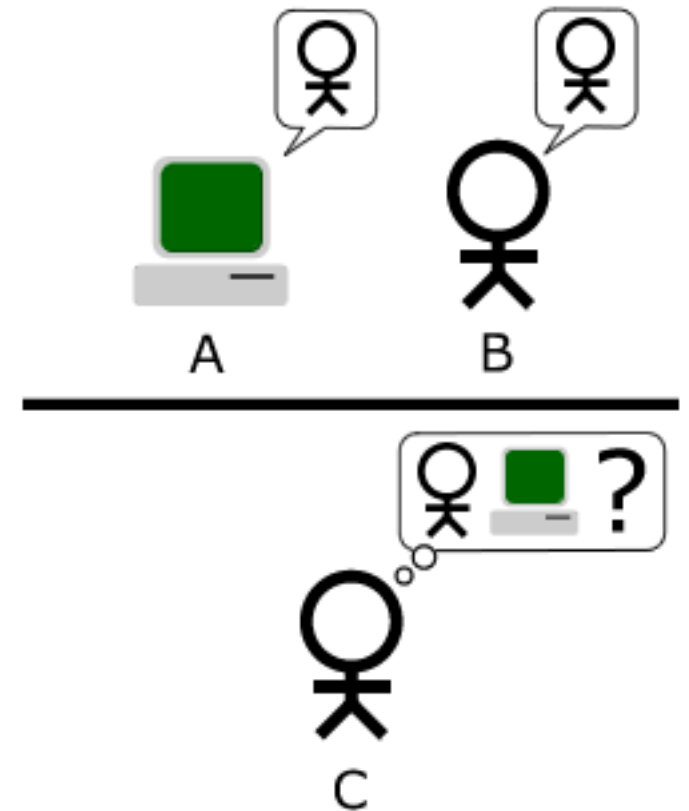
- Modélisation d'après des observations neurophysiologiques et anatomiques
- Somme pondérée binaire avec fonction d'activation à seuil



Domaine $\{0,1\}$ Domaine \mathbb{R} Domaine \mathbb{R} Domaine $\{0,1\}$

1950 – Test de Turing

- Proposition de test d'intelligence artificielle fondé sur la faculté d'une machine à imiter la conversation humaine.
- Prédiction : à l'an 2000, les machines avec 128 Mo de mémoire seront capables de tromper environ 30 % des juges humains durant un test de 5 minutes.
- Aucune machine n'a encore passé ce test



1955 – Logic theorist par Allen Newell, Herbert Simon et Cliff Shaw

- 1^{er} raisonnement symbolique
- Capable de prouver 38 des 52 théorèmes des *Principia Mathematica* de Whitehead et Russell

*54·43. $\vdash :: \alpha, \beta \in 1 . \supset : \alpha \cap \beta = \Lambda . \equiv . \alpha \cup \beta \in 2$

Dem.

$\vdash . *54·26 . \supset \vdash :: \alpha = t'x . \beta = t'y . \supset : \alpha \cup \beta \in 2 . \equiv . x \neq y .$
 [*51·231] $\equiv . t'x \cap t'y = \Lambda .$
 [*13·12] $\equiv . \alpha \cap \beta = \Lambda \quad (1)$

$\vdash . (1) . *11·11·35 . \supset$
 $\vdash :: (\forall x, y) . \alpha = t'x . \beta = t'y . \supset : \alpha \cup \beta \in 2 . \equiv . \alpha \cap \beta = \Lambda \quad (2)$

$\vdash . (2) . *11·54 . *52·1 . \supset \vdash . \text{Prop}$

From this proposition it will follow, when arithmetical addition has been defined, that $1 + 1 = 2$.

The Logic Theory Machine

In the language we have constructed, we have variables (atomic sentences): p, q, r, A, B, C, ... and connectives: - (not), v (or), \rightarrow (implies). The connectives are used to combine the variables into expressions (molecular sentences). We have already considered one example of an expression:

1.7 $-p \rightarrow q \vee -p$

The task set for LT will be to prove that certain expressions are theorems — that is, that they can be derived by application of specified rules of inference from a set of primitive sentences or axioms.

The two connectives, - and v, are taken as primitives. The third connective, \rightarrow , is defined in terms of the other two, thus:

1.01 $p \rightarrow q \text{ "def" } -p \vee q$

The five axioms that are postulated to be true are:

1.2 $p \vee p \rightarrow p$

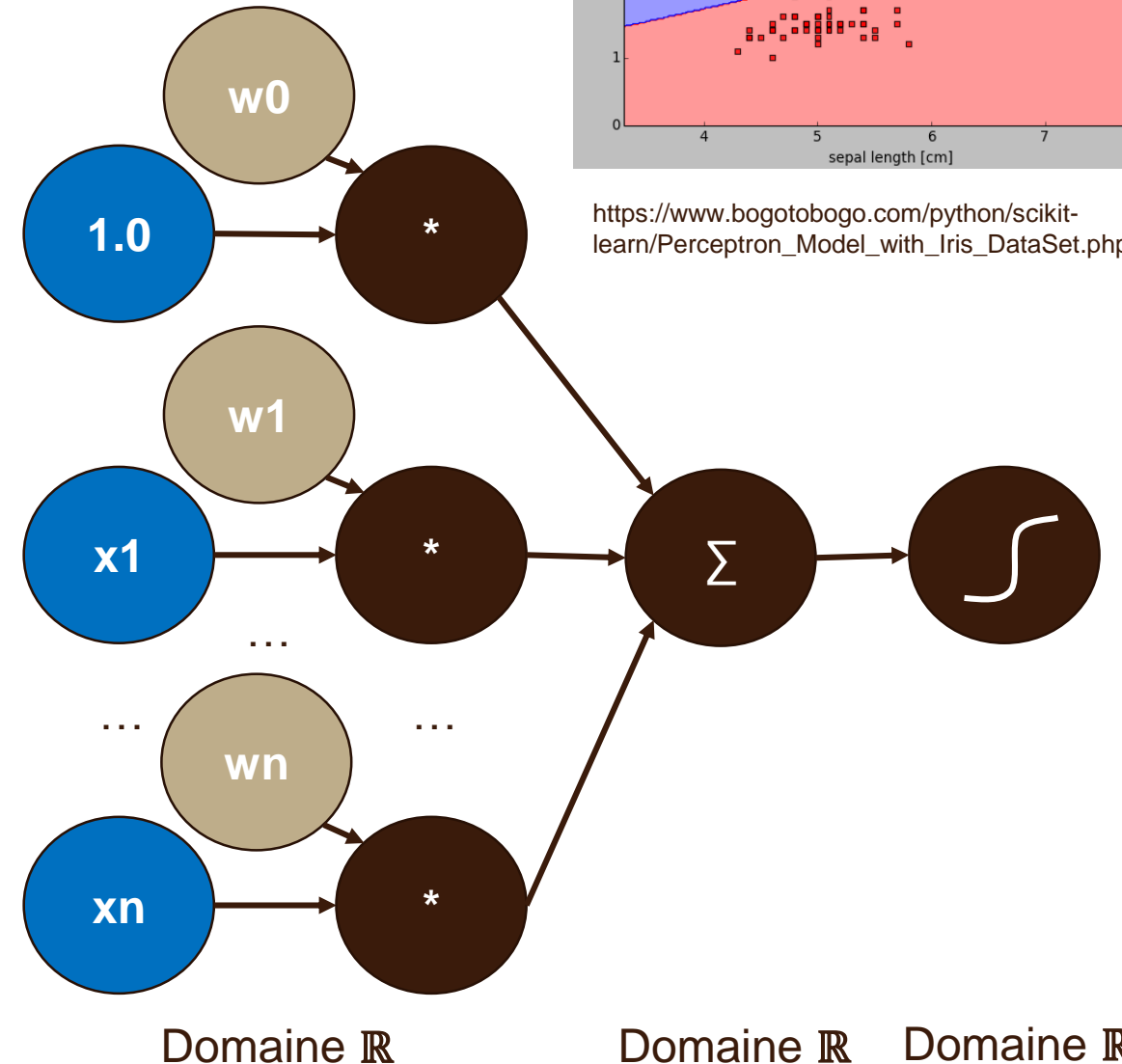
1.3 $p \rightarrow q \vee p$

1.4 $p \vee q \rightarrow q \vee p$

1.5 $p \vee q \vee r \rightarrow q \vee p \vee r$

1957 – Perceptron de Rosenblatt

- **Classifieur linéaire à deux classes**
- **Élément fondamental du réseau de neurones moderne**
- **Différence au neurone de McCulloch et Pitts**
 - Valeurs réelles plutôt que binaires
 - Accompagné d'une règle d'apprentissage



1959 – General Problem Solver de Herbert Simon, Cliff Shaw et Allen Newell

- solveur de problèmes universel : résout n'importe quel problème (simple) formalisé
- 1er programme séparant données et code

P-1584
2-9-59
-11-

SUMMARY

This paper reports on a computer program, called GPS-I for General Problem Solving Program I. Construction and investigation of this program is part of a research effort by the authors to understand the information processes that underlie human intellectual, adaptive, and creative abilities. The approach is synthetic – to construct computer programs that can solve problems requiring intelligence and adaptation, and to discover which varieties of these programs can be matched to data on human problem solving.

GPS-I grew out of an earlier program, the Logic Theorist, which discovers proofs to theorems in the sentential calculus. GPS-I is an attempt to fit the recorded behavior of college students trying to discover proofs. The purpose of this paper is not to relate the program to human behavior, but to describe its main characteristics and to assess its capacities as a problem-solving mechanism. The paper will present enough theoretical discussion of problem-solving activity so that the program can be seen as an attempt to advance our basic knowledge of intellectual activity. The program will be assessed from this point of view, rather than whether it offers an economical solution to a significant class of problems.

The major features of the program that are worthy of discussion are:

1960 – Début du SAT (satisfaisabilité booléenne) par Davis et Putnam

1960 Algorithme pour prouver qu'un ensemble de clauses est satisfiable

$$(p \wedge q) \vee \neg p \Rightarrow p = \text{faux}$$
$$(p \wedge \neg p) \Rightarrow \text{pas de solution}$$

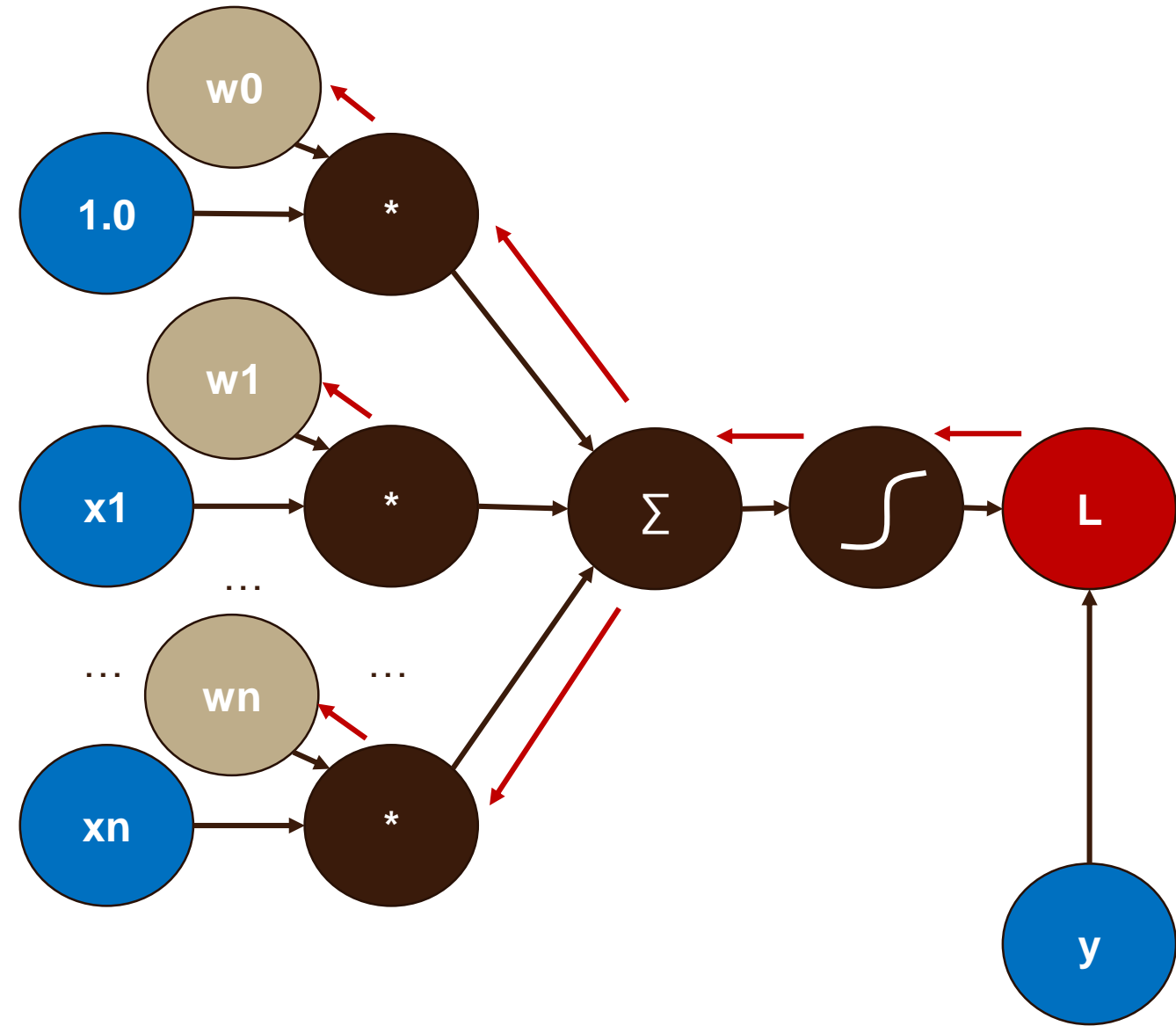
1962 Algorithme pour résoudre le problème SAT

La base des solveurs modernes pour :

- Le diagnostic
- La planification
- La vérification de propriété de modèles

1980 – Rétropropagation du gradient

- Inventé plusieurs fois par différentes personnes sur la même période
- Méthode actuelle d'apprentissage pour réseau de neurones
- Principe
 - Modification des poids de la sortie du graphe vers l'entrée en fonction d'une fonction de coût



1997 – Deep blue bat Kasparov

- Superordinateur de IBM (1.8m/1400kg)
- Bât le champion du monde Garry Kasparov
- Un bug a permis un coup très impressionnant qui a déstabilisé le champion



<http://www.intelligenceartificielle.org/details-deep+blue+super-ordinateur+d+ibm+-13.html>

2005 – Première conduite autonome sur route inconnue

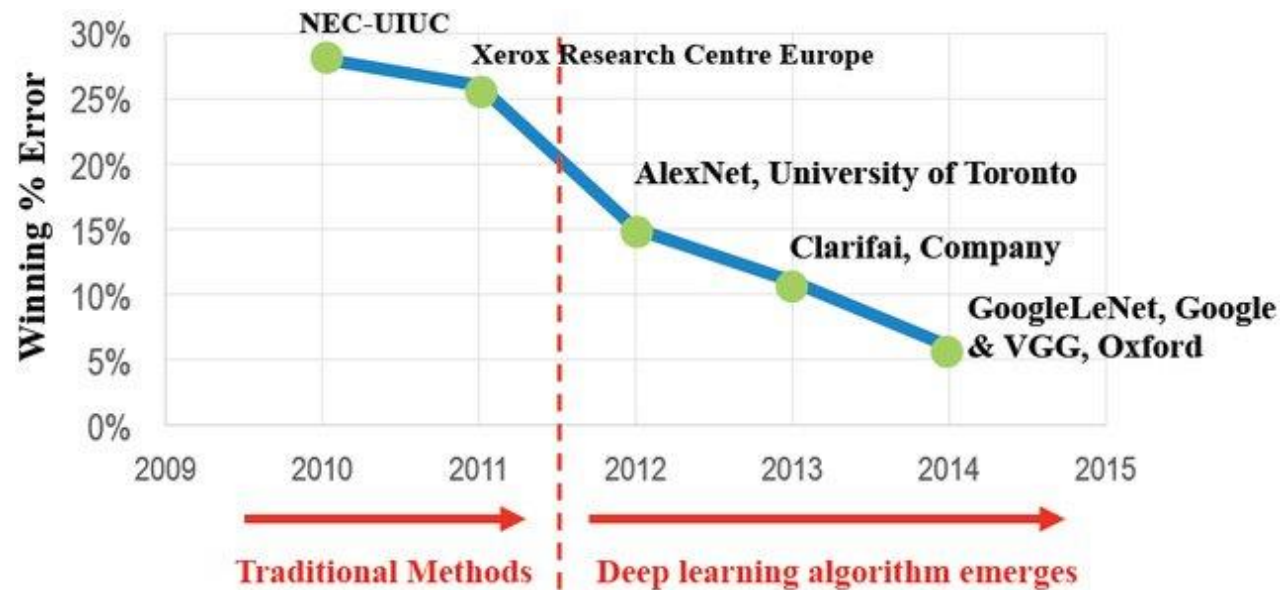
- 2004 - Darpa Grand Challenge (230km désert)
 - Aucune voiture ne termine
- 2005
 - 5 véhicules sur 23 terminent



<https://www.herox.com/blog/159-the-drive-for-autonomous-vehicles-the-darpa-grand>

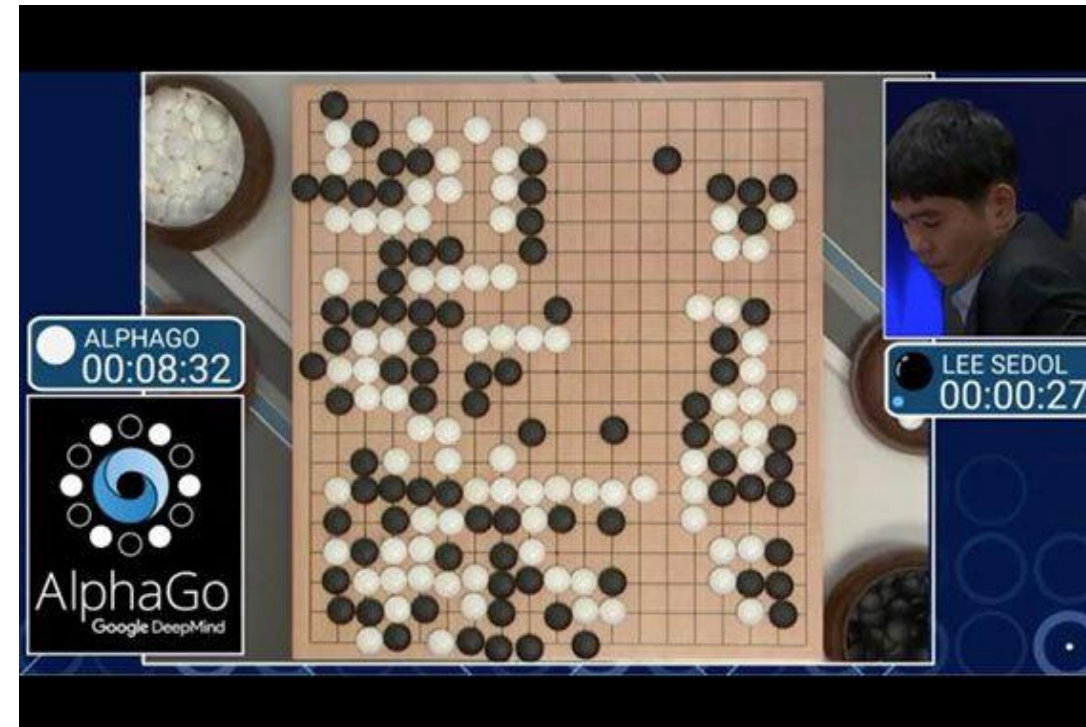
2012 – Apprentissage profond

- Réveil de techniques abandonnées 10 ans plus tôt
- AlexNet bat haut la main les concurrents utilisant des méthodes standards dans un concours de classification
- Depuis les performances ne font que s'améliorer



2015 – Alpha GO

- Programme de DeepMind
- 1er programme à battre différents joueurs professionnels (2015/2016/2017)
- Surpassé par Alpha Zero en 2017
- <https://github.com/leela-zero/leela-zero>



<https://www.lesaffaires.com/techno/technologie-de-l-information/la-victoire-d-alphago-contre-le-champion-du-jeu-de-go-restera-dans-l-histoire/585999>

L'intelligence artificielle dans la vie quotidienne

Aujourd'hui

- Traduction automatique
- Planification de trajet
- Assistant personnel
- Upscaling jeux ou vidéos

Demain

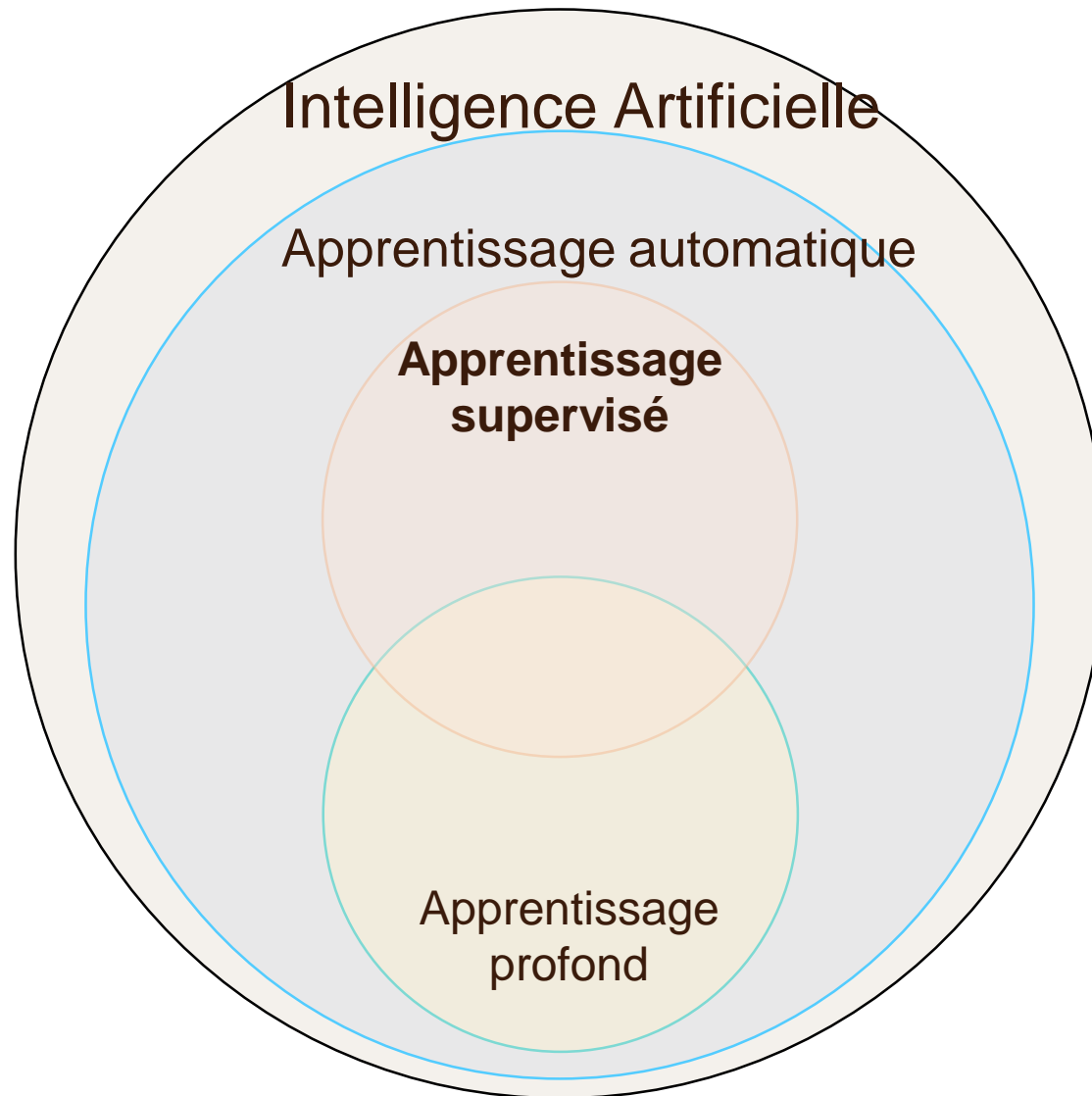
- Voiture intelligente
- Jeux Vidéos
- Œuvres artistiques
- Enseignement



Généralités sur l'apprentissage supervisé

**On se focalise sur
l'aspect
apprentissage
supervisé et on
ignore les autres
aspects de
l'intelligence
artificielle**

Positionnement de l'apprentissage supervisé

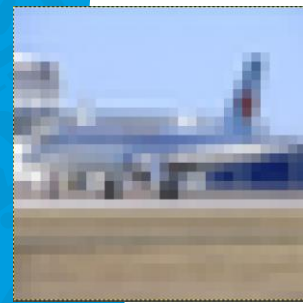
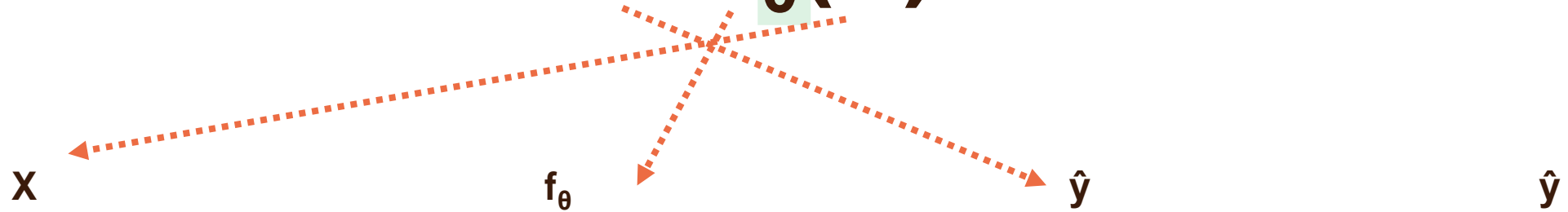


Un modèle de classification prédit une classe

Code écrit par un humain

Paramètres appris grâce aux données

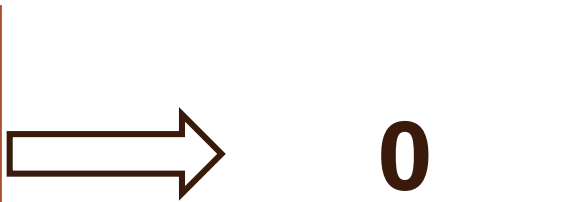
$$\hat{Y} = f_{\theta}(X)$$



Donnée d'entrée



Modèle

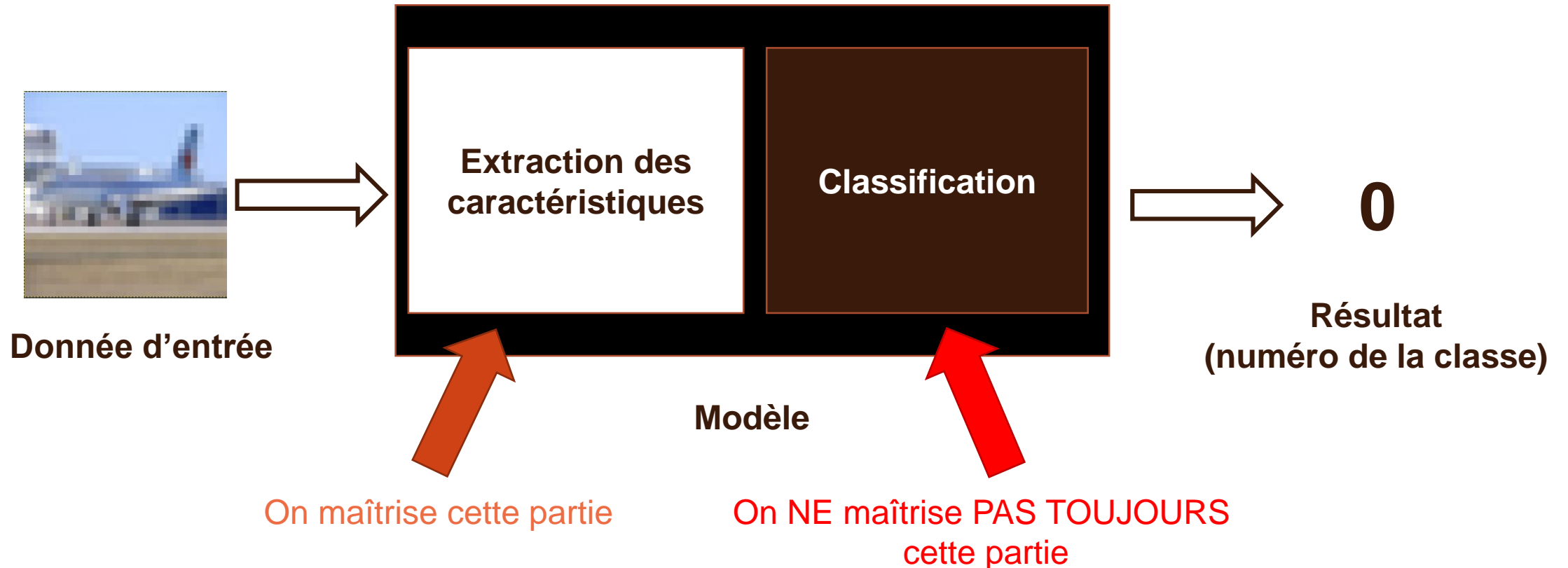


Résultat
numéro de la classe
compréhensible par la
machine



Résultat
symbole compréhensible
par un humain

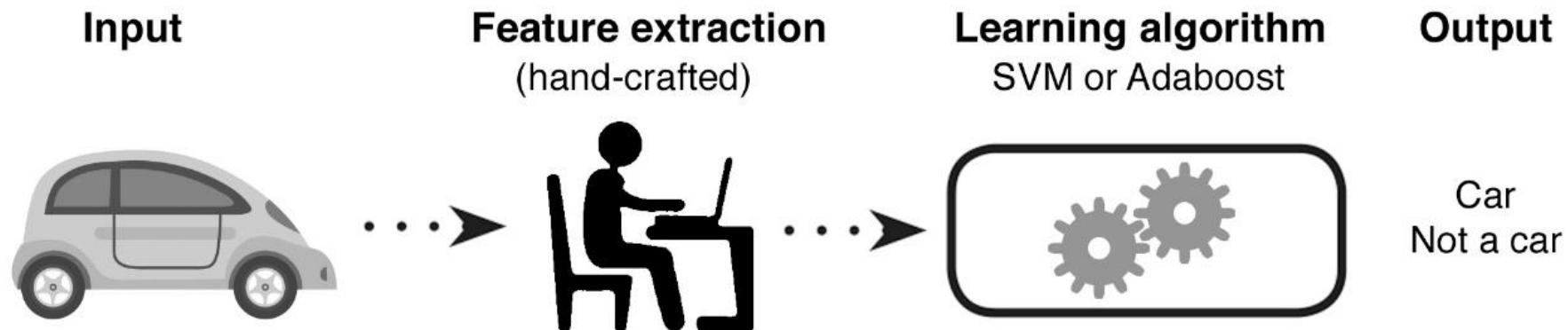
Les données d'entrée doivent être transformées pour être exploitables par la machine



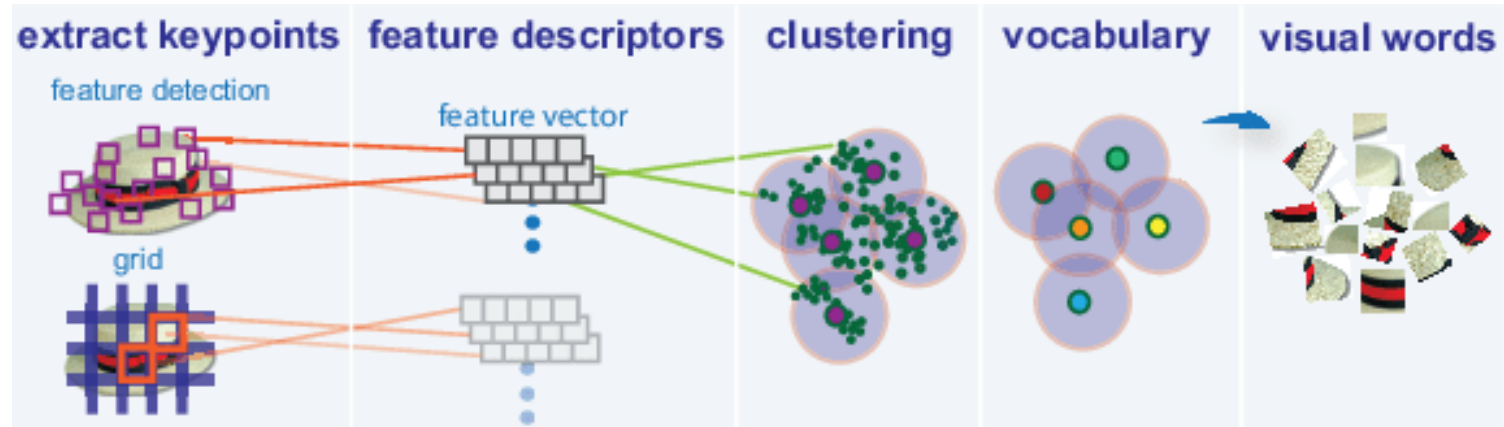
Généralités sur l'extraction de caractéristiques

- Les données brutes ne peuvent pas être utilisées directement
- Il faut extraire des informations
- De nombreuses méthodes générales existent

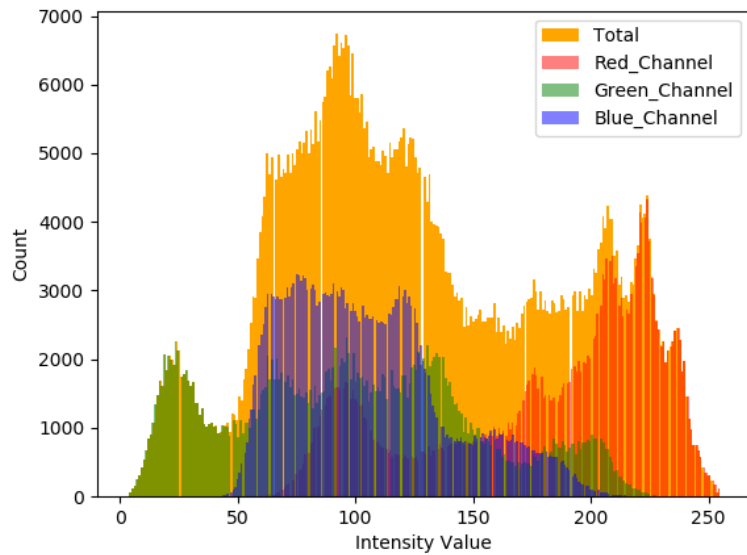
Il est nécessaire d'être expert des données pour extraire des caractéristiques pertinentes



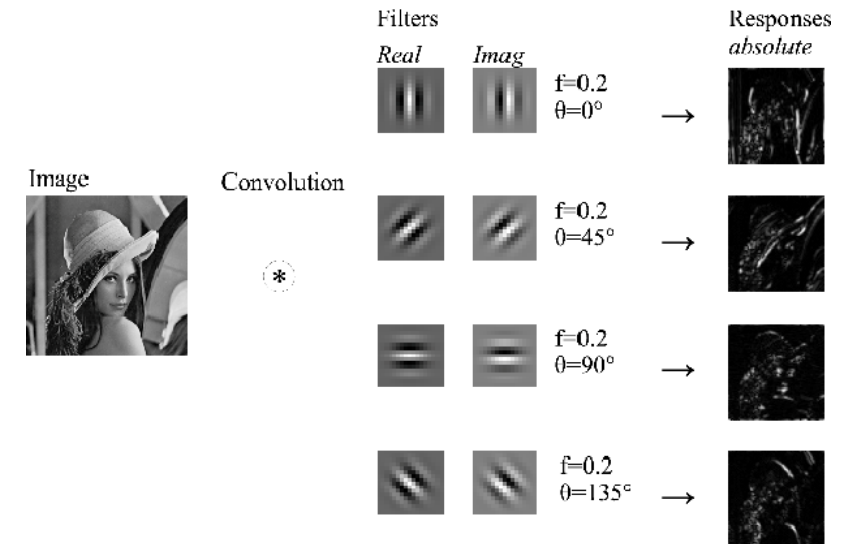
Extraction de caractéristique - Image



<https://www.mathworks.com/help/vision/ug/image-classification-with-bag-of-visual-words.html>



https://miro.medium.com/v2/resize:fit:640/1*Vd5OY8LRaybkFj2NjBbpbA.png



<https://www.semanticscholar.org/paper/Gabor-features-in-image-analysis-K%C3%A4m%C3%A4r%C3%A4inen/b0bc670f8ffb469eec5253cf9d634e90fd6eda88>

Extraction de caractéristique - Texte

Découpage



Recherche de radical



Dépendance
s



changing
changed
change

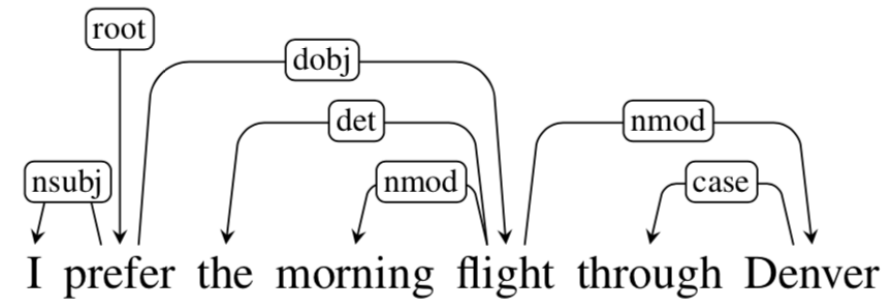
stemming

chang
chang
chang

studying
studies
study

stemming

studi
studi
studi



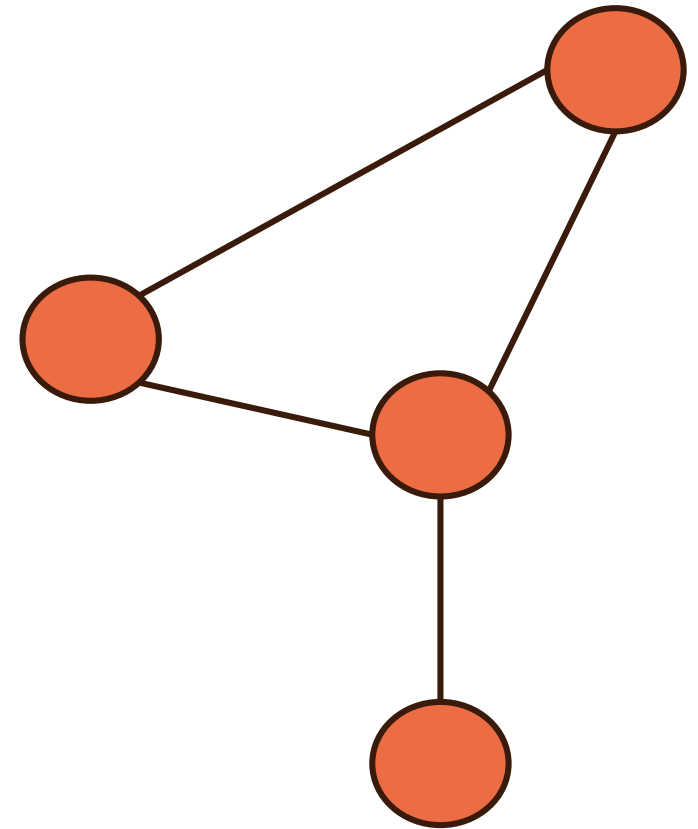
Extraction de caractéristique – Données tabulaires



- Normalisation de dates
- Gestion de données manquantes
- Agrégations diverses
- ...

Extraction de caractéristiques - Graphes

- **Détection de communautés**
- **Calcul de chemins**
- **Calcul de plongement**
- **Métriques diverses**
- **Recherche de motifs**
- ...



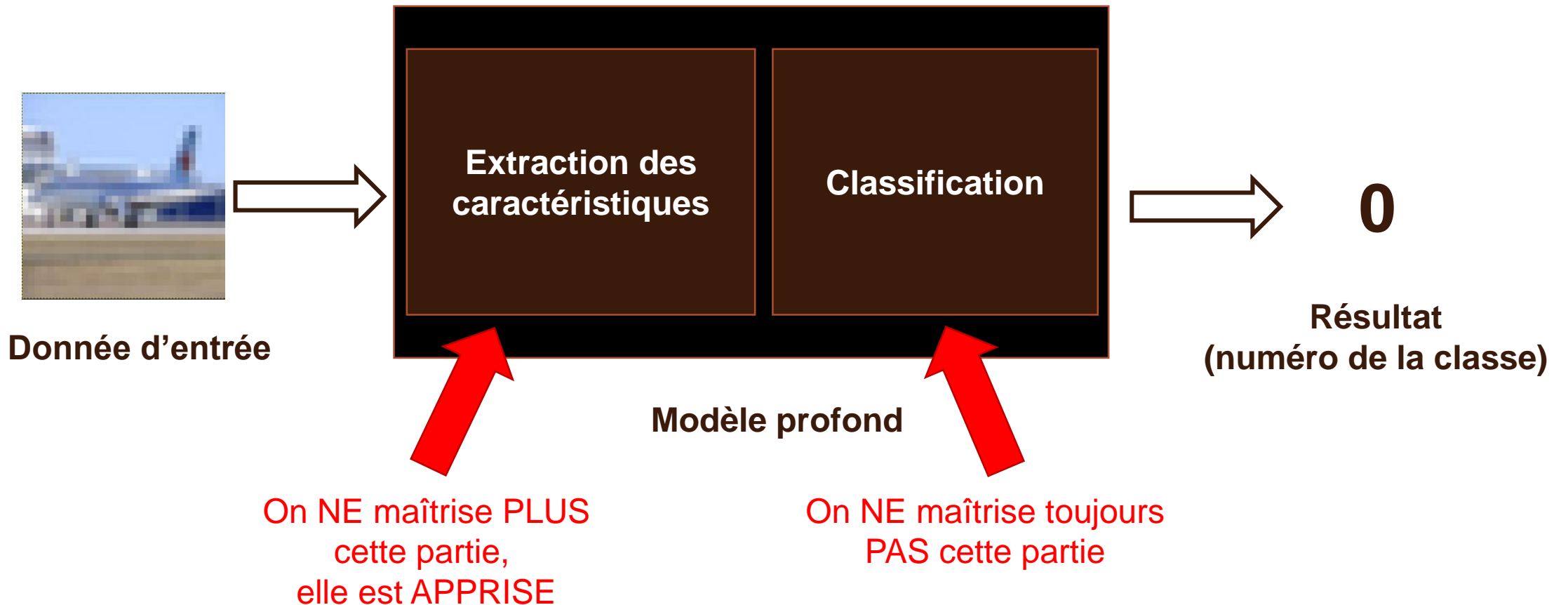
Extraction de caractéristique – Données structurées



- **Opérations spécifiques aux données**

Avec l'apprentissage profond, l'extraction devient automatique et non maîtrisée

Il n'est plus nécessaire d'être expert des données pour extraire des caractéristiques pertinentes



Evaluation des modèles d'apprentissage supervisé

**Il est nécessaire
d'évaluer les
modèles pour
connaitre leurs
performances**

Pour rappel, voici comment fonctionne un classifieur



0



avion

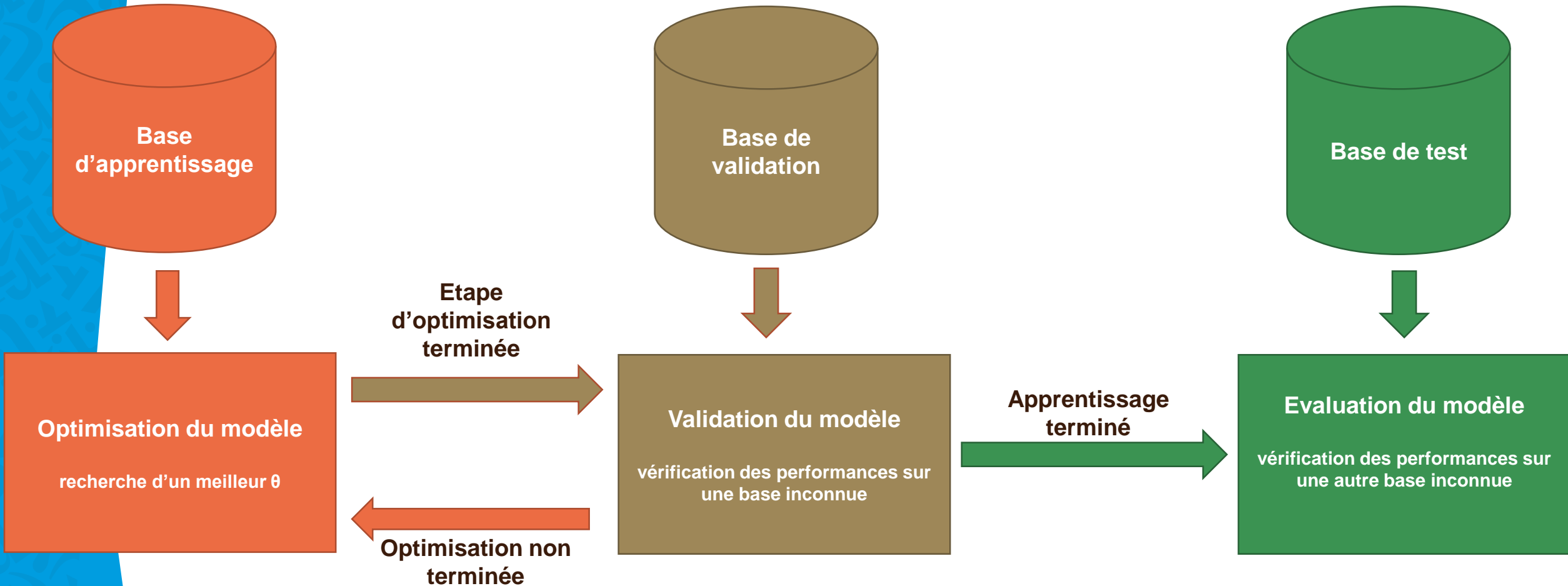
Donnée d'entrée

Modèle

Résultat
(numéro de la classe)

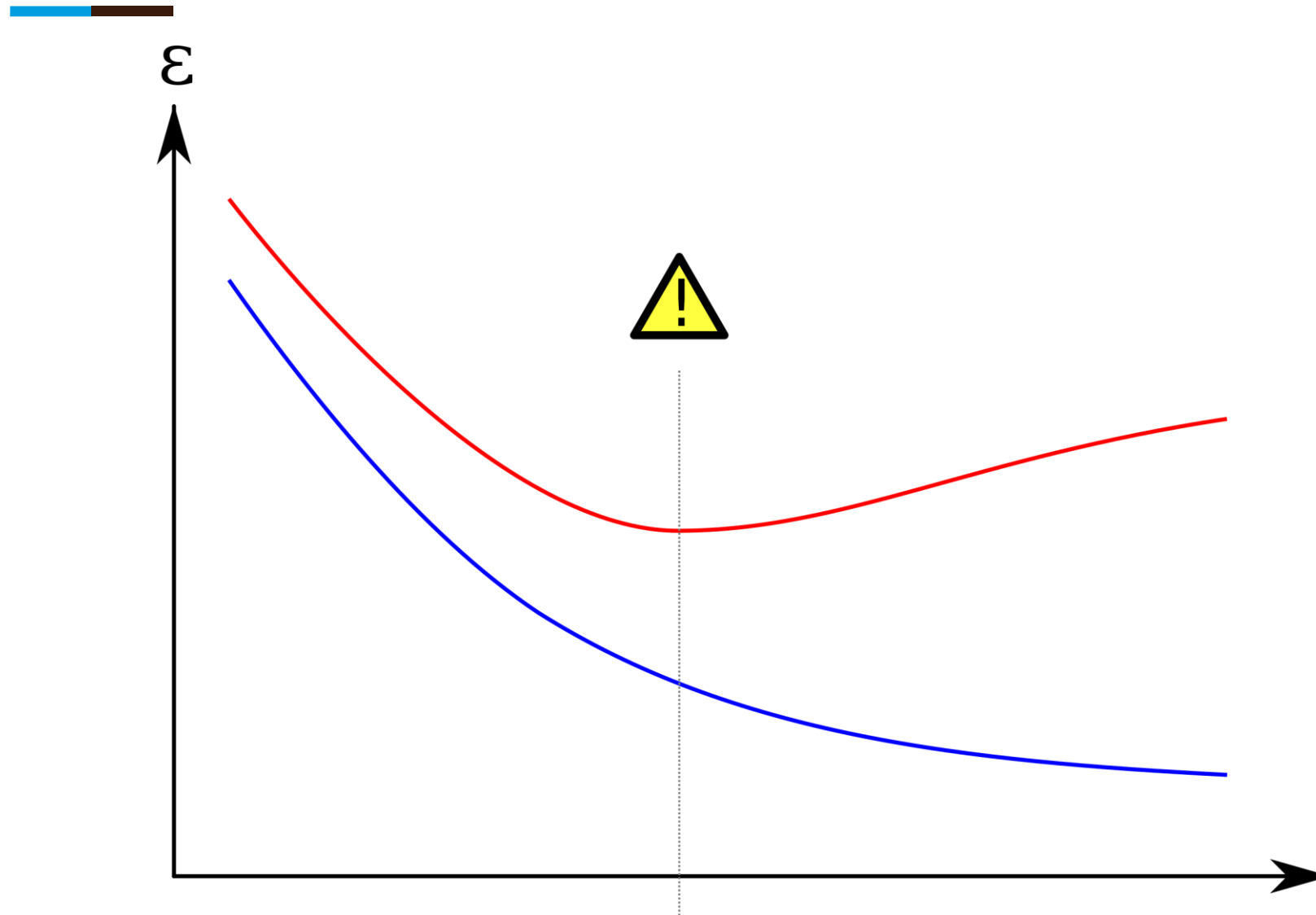
Résultat
(nom de la classe)

Aparté : apprentissage d'un modèle

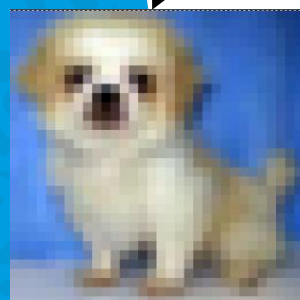
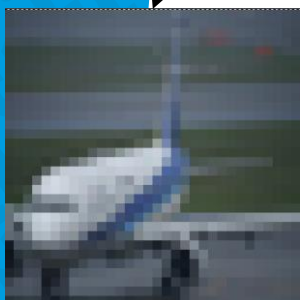
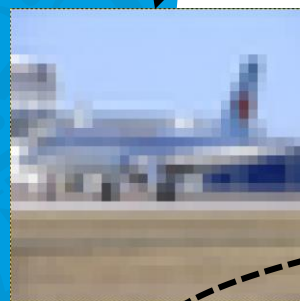


Les techniques dépendent bien entendu du type de modèle

Pourquoi autant de bases ?



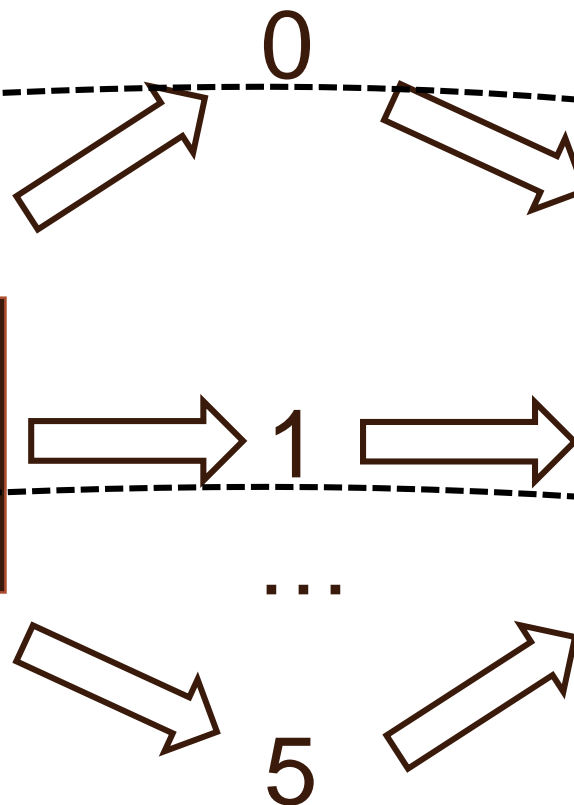
Evaluation d'un modèle de classification



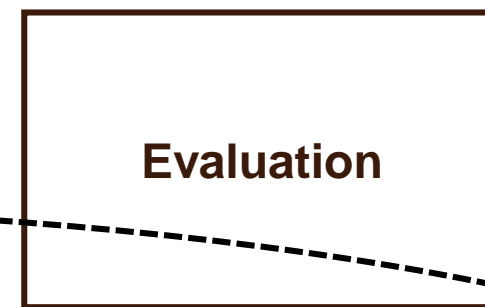
Base d'évaluation
étiquetée



Modèle



Résultats



Vérité
terrain

La matrice de confusion : un ensemble de compteurs pour calculer différentes métriques

```
[[ 828, 13, 12, 11, 18, 0, 2, 4, 85, 27],  
 [ 10, 910, 0, 5, 1, 1, 0, 1, 11, 61],  
 [ 47, 1, 708, 64, 88, 14, 63, 4, 8, 3],  
 [ 3, 4, 16, 768, 33, 93, 50, 19, 4, 10],  
 [ 10, 0, 39, 43, 788, 12, 57, 43, 6, 2],  
 [ 2, 0, 10, 137, 29, 777, 8, 33, 0, 4],  
 [ 7, 2, 10, 54, 29, 7, 888, 1, 1, 1],  
 [ 24, 2, 14, 39, 76, 17, 4, 818, 2, 4],  
 [ 27, 14, 0, 7, 3, 0, 3, 0, 9333, 14],  
 [ 19, 64, 1, 7, 2, 1, 1, 0, 18, 887]])
```

La matrice de confusion : un ensemble de compteurs pour calculer différentes métriques



Prédiction

	avion	voiture	oiseau	chat	cerf	chien	grenouille	cheval	bateau	camion
avion	828	13	12	11	18	0	2	4	85	27
voiture	10	910	0	5	1	1	0	1	11	61
oiseau	47	1	708	64	88	14	63	4	8	3
chat	3	4	16	768	33	93	50	19	4	10
cerf	10	0	39	43	788	12	57	43	6	2
chien	2	0	10	137	29	777	8	33	0	4
grenouille	7	2	10	54	29	7	888	1	1	1
cheval	24	2	14	39	76	17	4	818	2	4
bateau	27	14	0	7	3	0	3	0	9333	14
camion	19	64	1	7	2	1	1	0	18	887

Chaque ligne représente les exemples appartenant à une classe précise



Prédiction

	avion	voiture	oiseau	chat	cerf	chien	grenouille	cheval	bateau	camion	total
avion	828	13	12	11	18	0	2	4	85	27	1000
voiture	10	910	0	5	1	1	0	1	11	61	1000
oiseau	47	1	708	64	88	14	63	4	8	3	1000
chat	3	4	16	768	33	93	50	19	4	10	1000
cerf	10	0	39	43	788	12	57	43	6	2	1000
chien	2	0	10	137	29	777	8	33	0	4	1000
grenouille	7	2	10	54	29	7	888	1	1	1	1000
cheval	24	2	14	39	76	17	4	818	2	4	1000
bateau	27	13	0	7	3	0	3	0	933	14	1000
camion	19	64	1	7	2	1	1	0	18	887	1000

Vérité terrain

Chaque colonne représente les exemples classés dans une classe précise



Prédiction

	avion	voiture	oiseau	chat	cerf	chien	grenouille	cheval	bateau	camion	total
avion	828	13	12	11	18	0	2	4	85	27	1000
voiture	10	910	0	5	1	1	0	1	11	61	1000
oiseau	47	1	708	64	88	14	63	4	8	3	1000
chat	3	4	16	768	33	93	50	19	4	10	1000
cerf	10	0	39	43	788	12	57	43	6	2	1000
chien	2	0	10	137	29	777	8	33	0	4	1000
grenouille	7	2	10	54	29	7	888	1	1	1	1000
cheval	24	2	14	39	76	17	4	818	2	4	1000
bateau	27	13	0	7	3	0	3	0	933	14	1000
camion	19	64	1	7	2	1	1	0	18	887	1000
total	977	1009	810	1135	1067	922	1076	923	1068	1013	10000

Les cellules de la diagonale représentent le nombre d'exemples bien classés



Prédiction

	avion	voiture	oiseau	chat	cerf	chien	grenouille	cheval	bateau	camion	total
avion	828	13	12	11	18	0	2	4	85	27	1000
voiture	10	910	0	5	1	1	0	1	11	61	1000
oiseau	47	1	708	64	88	14	63	4	8	3	1000
chat	3	4	16	768	33	93	50	19	4	10	1000
cerf	10	0	39	43	788	12	57	43	6	2	1000
chien	2	0	10	137	29	777	8	33	0	4	1000
grenouille	7	2	10	54	29	7	888	1	1	1	1000
cheval	24	2	14	39	76	17	4	818	2	4	1000
bateau	27	13	0	7	3	0	3	0	933	14	1000
camion	19	64	1	7	2	1	1	0	18	887	1000
total	977	1009	810	1135	1067	922	1076	923	1068	1013	10000

Les autres représentent les erreurs de classification



Prédiction

	avion	voiture	oiseau	chat	cerf	chien	grenouille	cheval	bateau	camion	total
avion	828	13	12	11	18	0	2	4	85	27	1000
voiture	10	910	0	5	1	1	0	1	11	61	1000
oiseau	47	1	708	64	88	14	63	4	8	3	1000
chat	3	4	16	768	33	93	50	19	4	10	1000
cerf	10	0	39	43	788	12	57	43	6	2	1000
chien	2	0	10	137	29	777	8	33	0	4	1000
grenouille	7	2	10	54	29	7	888	1	1	1	1000
cheval	24	2	14	39	76	17	4	818	2	4	1000
bateau	27	13	0	7	3	0	3	0	933	14	1000
camion	19	64	1	7	2	1	1	0	18	887	1000
total	977	1009	810	1135	1067	922	1076	923	1068	1013	10000

Vérité terrain

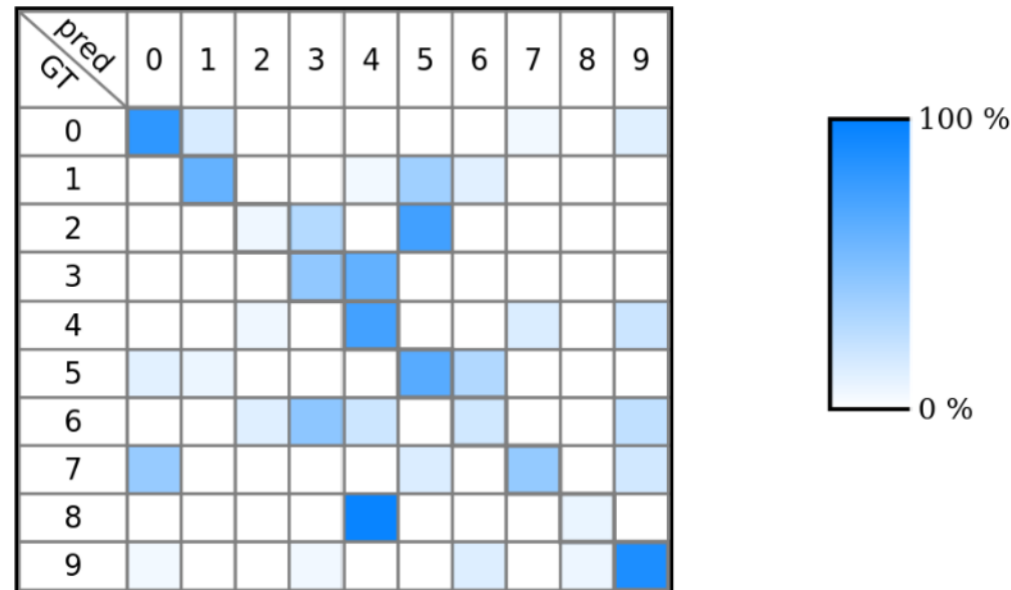
La matrice de confusion du système parfait est donc diagonale



Prédiction

	avion	voiture	oiseau	chat	cerf	chien	grenouille	cheval	bateau	camion	total
avion	1000	0	0	0	0	0	0	0	0	0	1000
voiture	0	1000	0	0	0	0	0	0	0	0	1000
oiseau	0	0	1000	0	0	0	0	0	0	0	1000
chat	0	0	0	1000	0	0	0	0	0	0	1000
cerf	0	0	0	0	1000	0	0	0	0	0	1000
chien	0	0	0	0	0	1000	0	0	0	0	1000
grenouille	0	0	0	0	0	0	1000	0	0	0	1000
cheval	0	0	0	0	0	0	0	1000	0	0	1000
bateau	0	0	0	0	0	0	0	0	1000	0	1000
camion	0	0	0	0	0	0	0	0	0	1000	1000
total	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	10000

Mais nous regardons rarement la matrice de confusion sous forme textuelle



Taux de reconnaissance

$$\frac{\sum_i \text{Vrais positifs de } i}{\text{Nb d'exemples}}$$

Prédiction

	avion	voiture	oiseau	chat	cerf	chien	grenouille	cheval	bateau	camion	total
avion	828	13	12	11	18	0	2	4	85	27	1000
voiture	10	910	0	5	1	1	0	1	11	61	1000
oiseau	47	1	708	64	88	14	63	4	8	3	1000
chat	3	4	16	768	33	93	50	19	4	10	1000
cerf	10	0	39	43	788	12	57	43	6	2	1000
chien	2	0	10	137	29	777	8	33	0	4	1000
grenouille	7	2	10	54	29	7	888	1	1	1	1000
cheval	24	2	14	39	76	17	4	818	2	4	1000
bateau	27	13	0	7	3	0	3	0	933	14	1000
camion	19	64	1	7	2	1	1	0	18	887	1000
total	977	1009	810	1135	1067	922	1076	923	1068	1013	10000

Vérité terrain

Précision $Précision(i) = \frac{\text{Vrais positifs de } i}{\text{Vrais positifs de } i + \text{Faux positifs de } i}$

Prédiction

	avion	voiture	oiseau	chat	cerf	chien	grenouille	cheval	bateau	camion	total
avion	828	13	12	11	18	0	2	4	85	27	1000
voiture	10	910	0	5	1	1	0	1	11	61	1000
oiseau	47	1	708	64	88	14	63	4	8	3	1000
chat	3	4	16	768	33	93	50	19	4	10	1000
cerf	10	0	39	43	788	12	57	43	6	2	1000
chien	2	0	10	137	29	777	8	33	0	4	1000
grenouille	7	2	10	54	29	7	888	1	1	1	1000
cheval	24	2	14	39	76	17	4	818	2	4	1000
bateau	27	13	0	7	3	0	3	0	933	14	1000
camion	19	64	1	7	2	1	1	0	18	887	1000
total	977	1009	810	1135	1067	922	1076	923	1068	1013	10000

Vérité terrain

Rappel

$$\text{Rappel}(i) = \frac{\text{Vrais positifs de } i}{\text{Vrais positifs de } i + \text{Faux négatifs de } i}$$

Prédiction

	avion	voiture	oiseau	chat	cerf	chien	grenouille	cheval	bateau	camion	total
avion	828	13	12	11	18	0	2	4	85	27	1000
voiture	10	910	0	5	1	1	0	1	11	61	1000
oiseau	47	1	708	64	88	14	63	4	8	3	1000
chat	3	4	16	768	33	93	50	19	4	10	1000
cerf	10	0	39	43	788	12	57	43	6	2	1000
chien	2	0	10	137	29	777	8	33	0	4	1000
grenouille	7	2	10	54	29	7	888	1	1	1	1000
cheval	24	2	14	39	76	17	4	818	2	4	1000
bateau	27	13	0	7	3	0	3	0	933	14	1000
camion	19	64	1	7	2	1	1	0	18	887	1000
total	977	1009	810	1135	1067	922	1076	923	1068	1013	10000

Spécificité

$$\text{Spécificité}(i) = \frac{\text{Vrais négatifs de } i}{\text{Vrais négatifs de } i + \text{Faux positifs de } i}$$

Prédiction

	avion	voiture	oiseau	chat	cerf	chien	grenouille	cheval	bateau	camion	total
avion	828	13	12	11	18	0	2	4	85	27	1000
voiture	10	910	0	5	1	1	0	1	11	61	1000
oiseau	47	1	708	64	88	14	63	4	8	3	1000
chat	3	4	16	768	33	93	50	19	4	10	1000
cerf	10	0	39	43	788	12	57	43	6	2	1000
chien	2	0	10	137	29	777	8	33	0	4	1000
grenouille	7	2	10	54	29	7	888	1	1	1	1000
cheval	24	2	14	39	76	17	4	818	2	4	1000
bateau	27	13	0	7	3	0	3	0	933	14	1000
camion	19	64	1	7	2	1	1	0	18	887	1000
total	977	1009	810	1135	1067	922	1076	923	1068	1013	10000

Vérité terrain

Généralités sur l'évaluation

- **Prérequis**
 - **Modèle**
 - **Base de test de qualité**
- **Obtient**
 - **Métriques à interpréter et comparer**
 - **Une idée des performances**
 - **Possibilité de comparer des modèles relativement les uns aux autres**
- **Il manque**
 - **Raisons qui expliquent les réussites et les erreurs**

Comment expliquer les décisions du classifieur ?

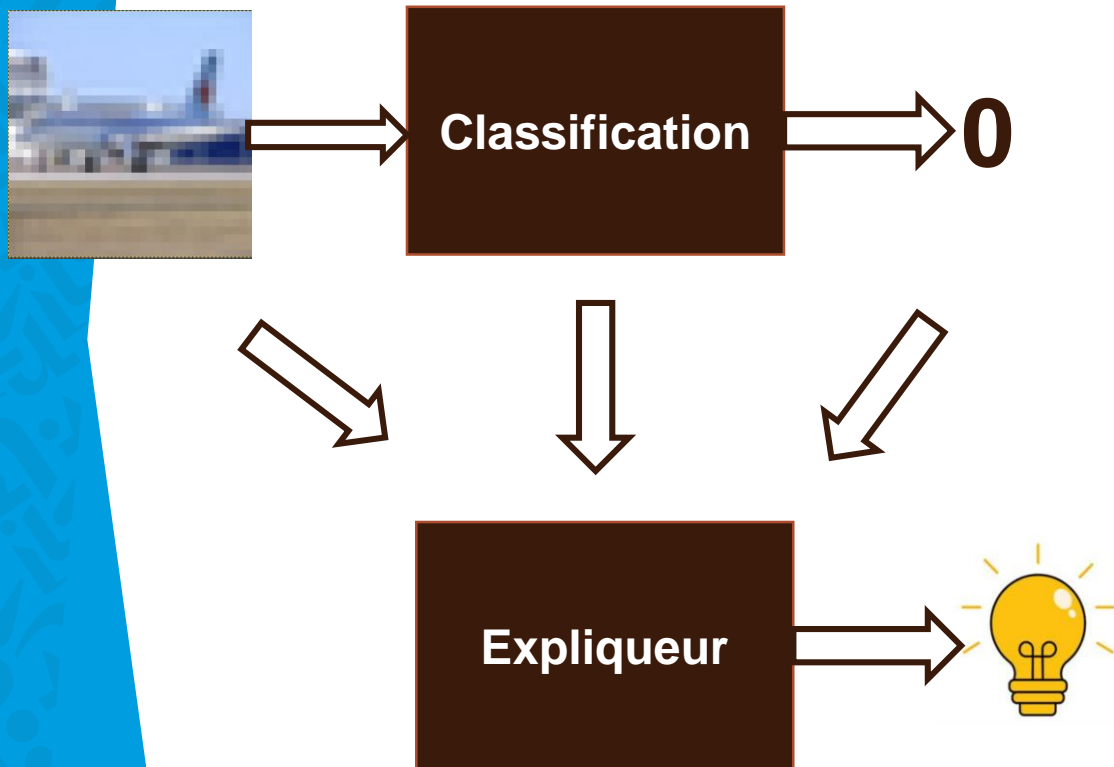
Méthodes intrinsèques

- Par construction, les modèles expliquent la décision
- En général il s'agit de modèles simples sur des données simples



Comment expliquer les décisions du classifieur ?

Méthodes après-coup



- On se sert de différentes informations pour supposer ce que le modèle a fait
- On ne sait pas toujours évaluer la qualité de l'explication


modèles simples de classification supervisée

**Description de
quelques modèles
avant de présenter
les explications
intrinsèques et
après-coup**

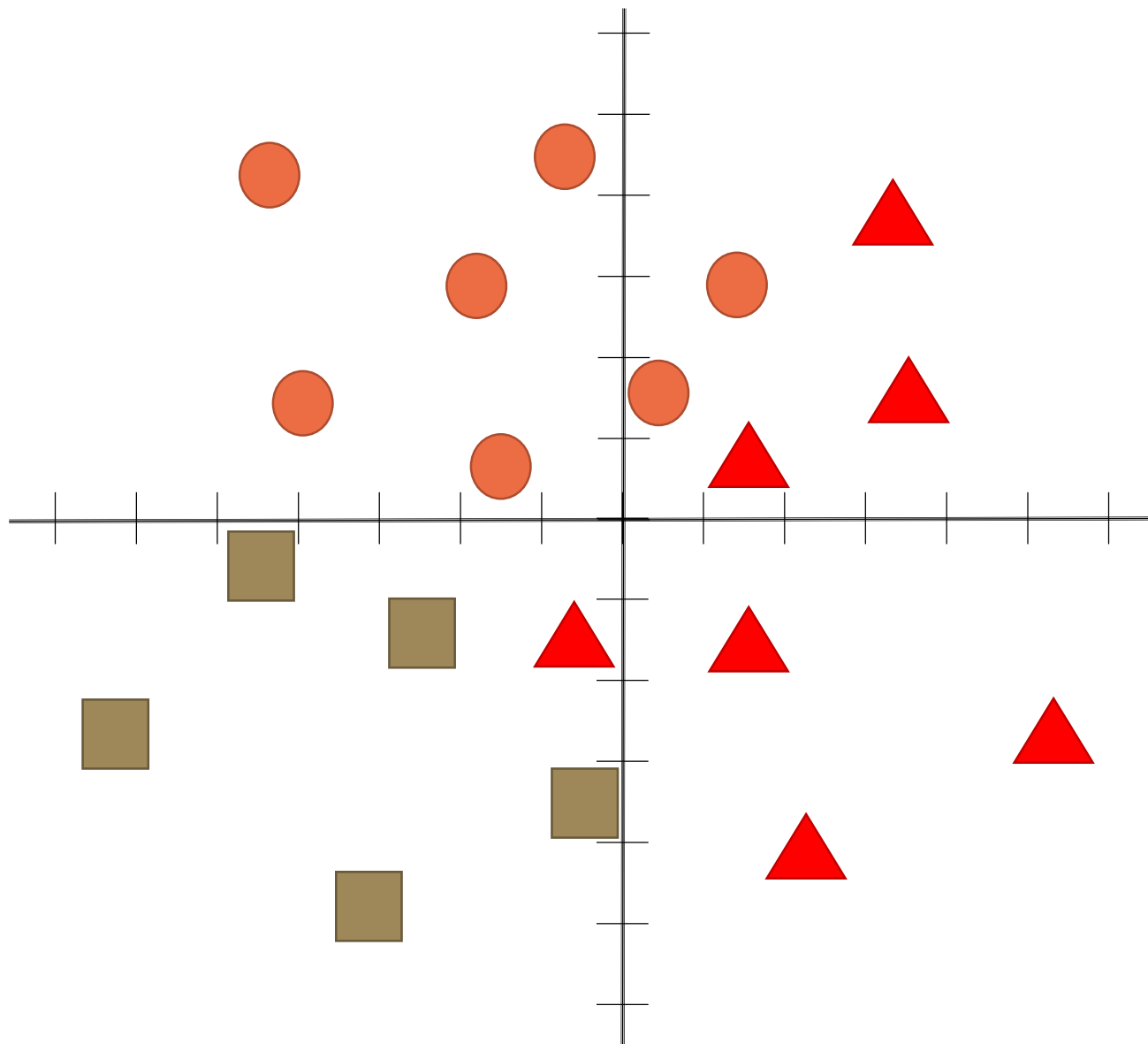
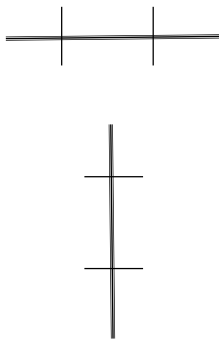
Classification de données en 2 dimensions



Classes



Dimensions



K plus proches voisins : apprentissage

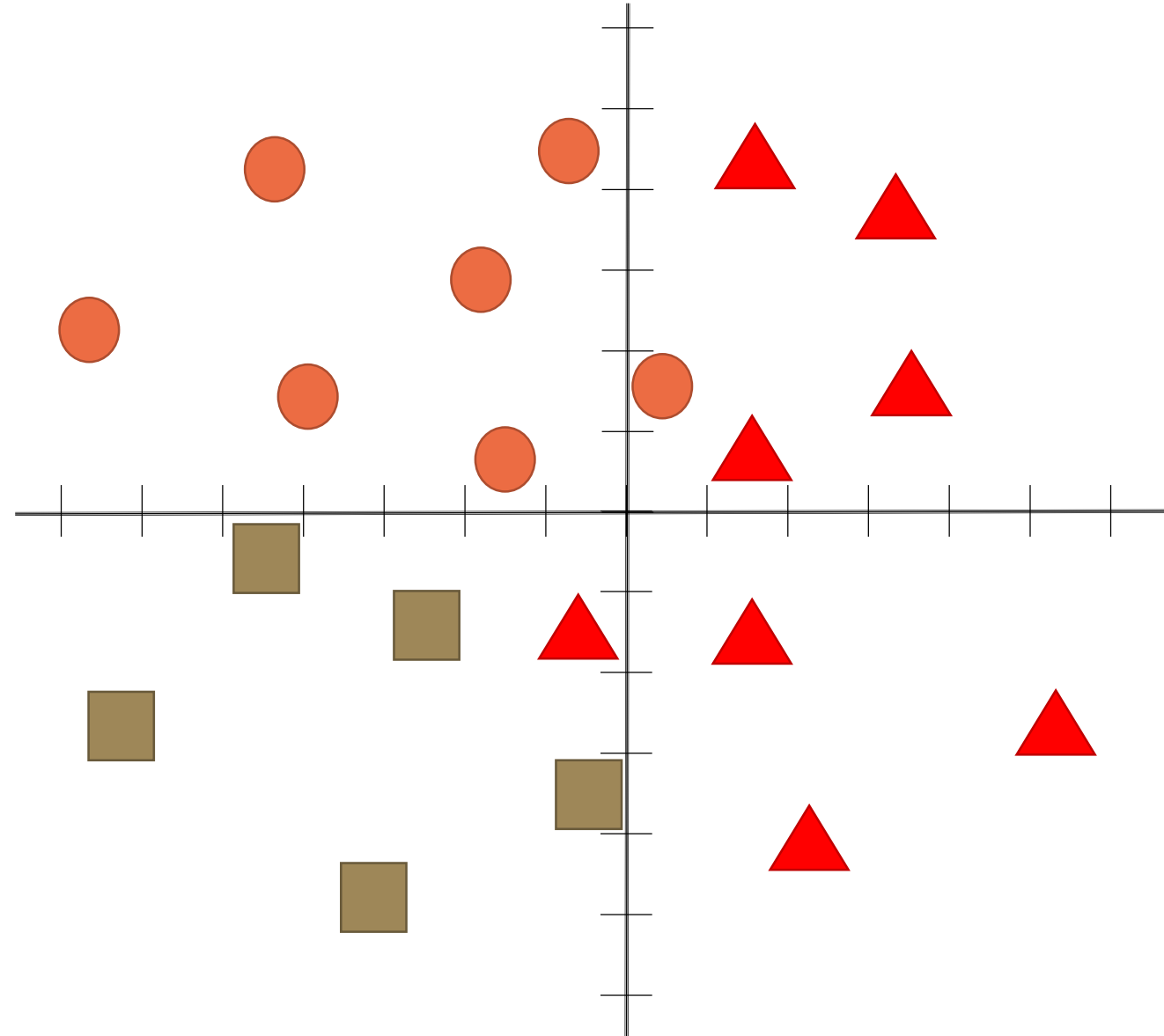


Paramètres : ensemble des données étiquetées

Hyperparamètres : K

F : code qui recherche les exemples les plus proches

Optimiseur : aucun



K plus proches voisins : inférence

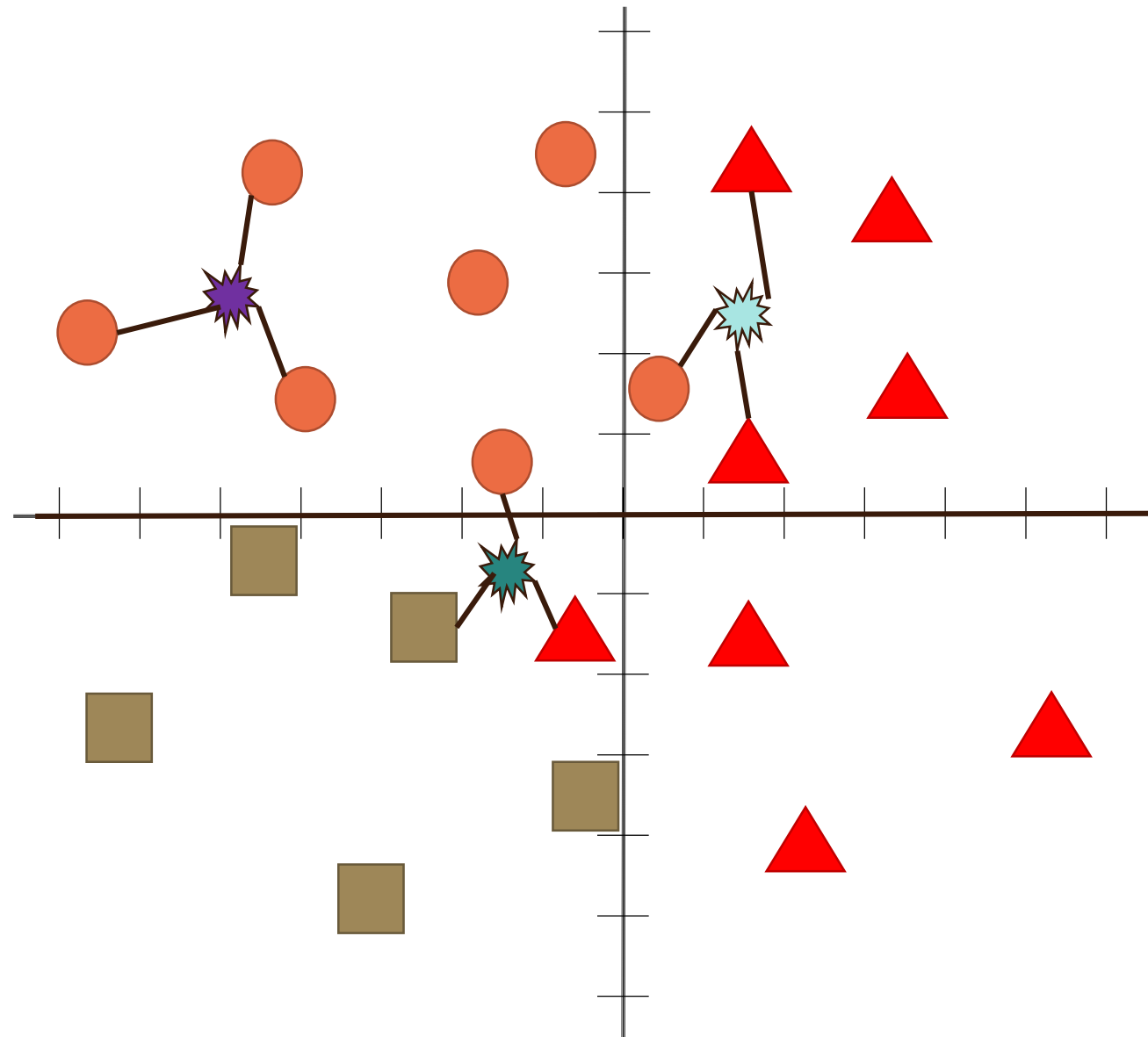
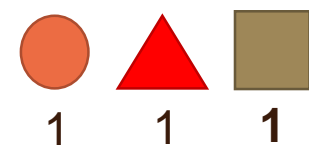
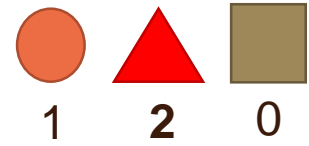
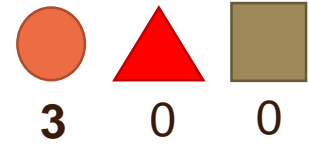


Requête

K voisins

Comptage

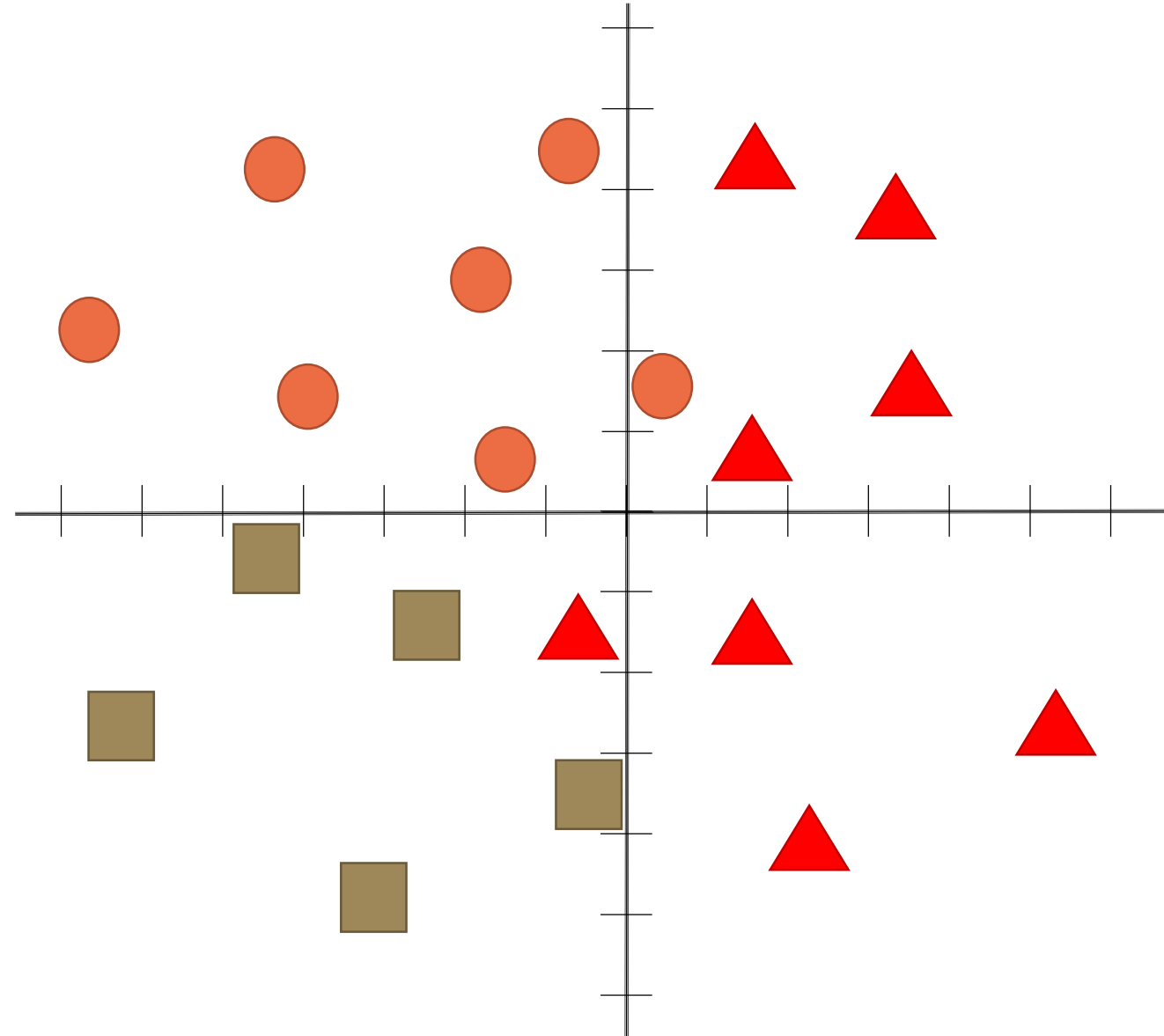
Prédiction



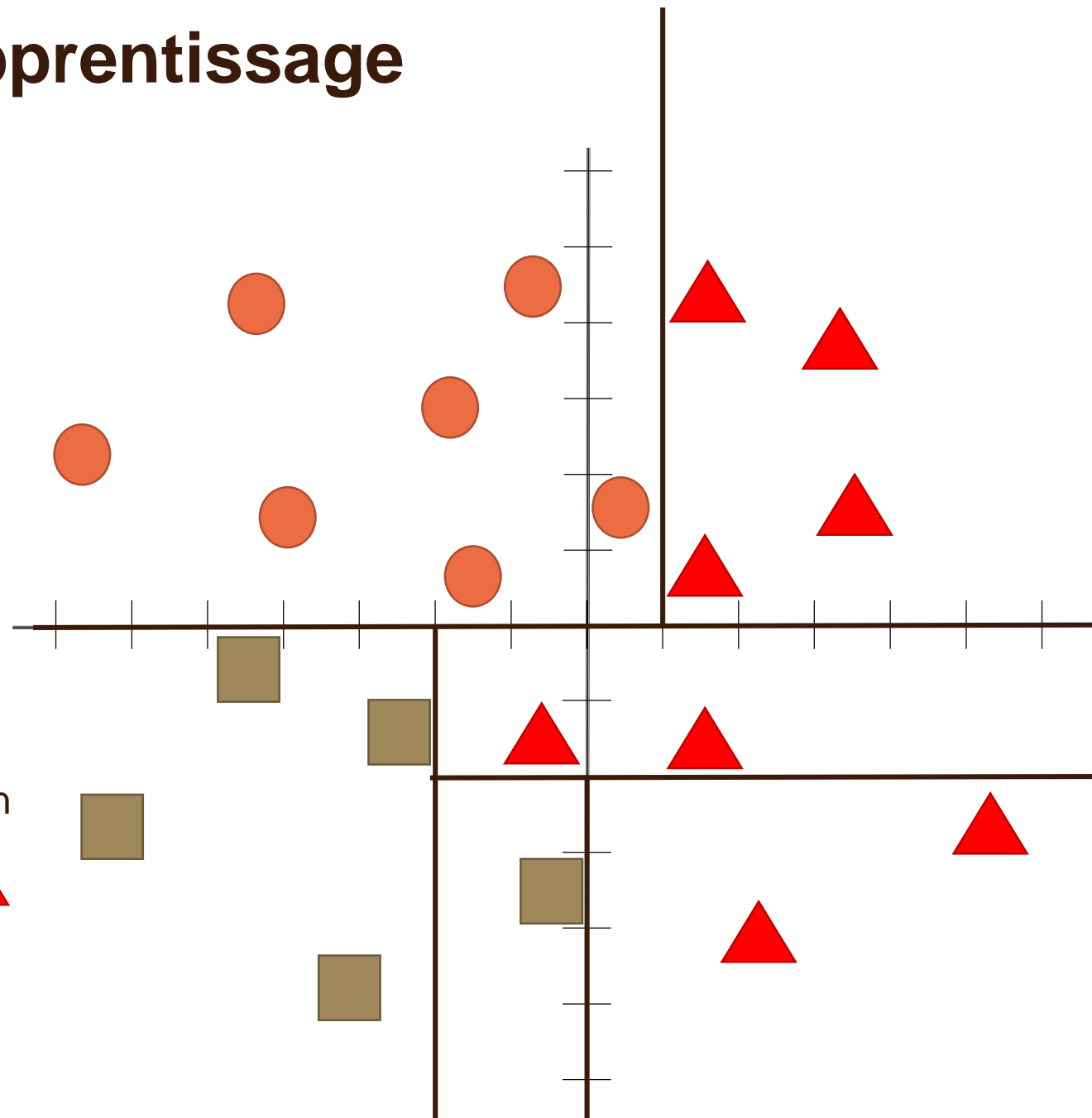
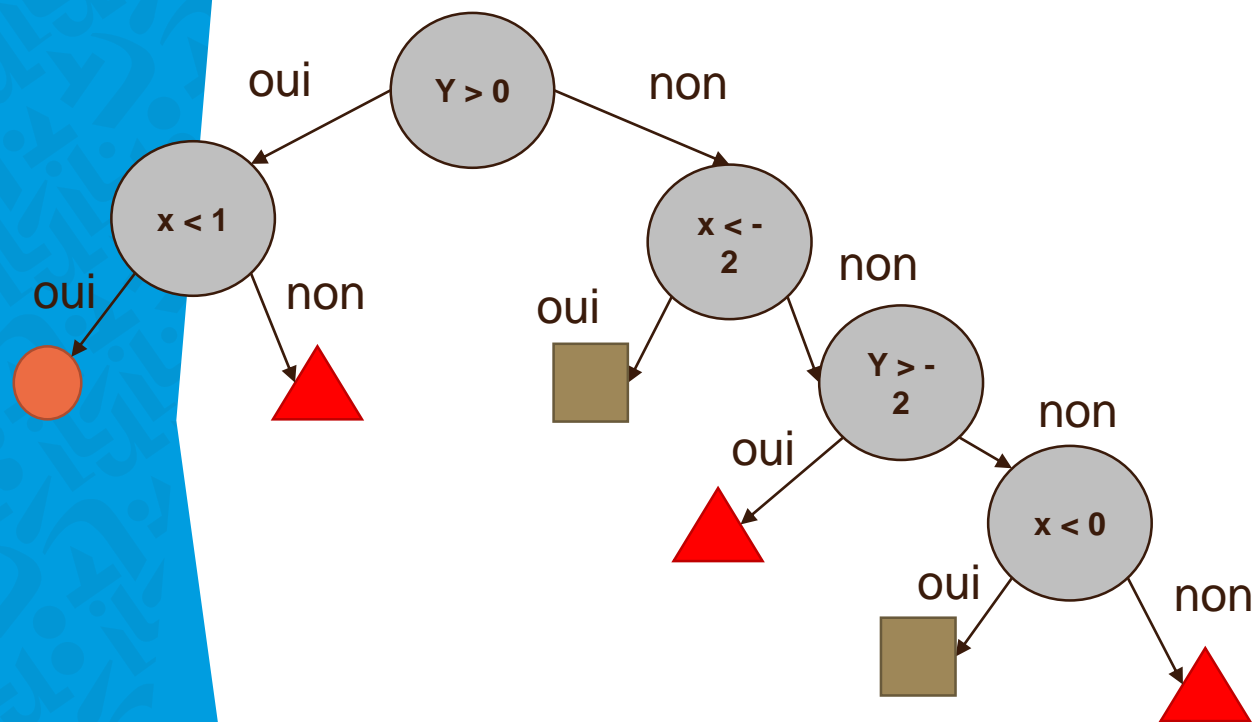
Arbre de décision : apprentissage



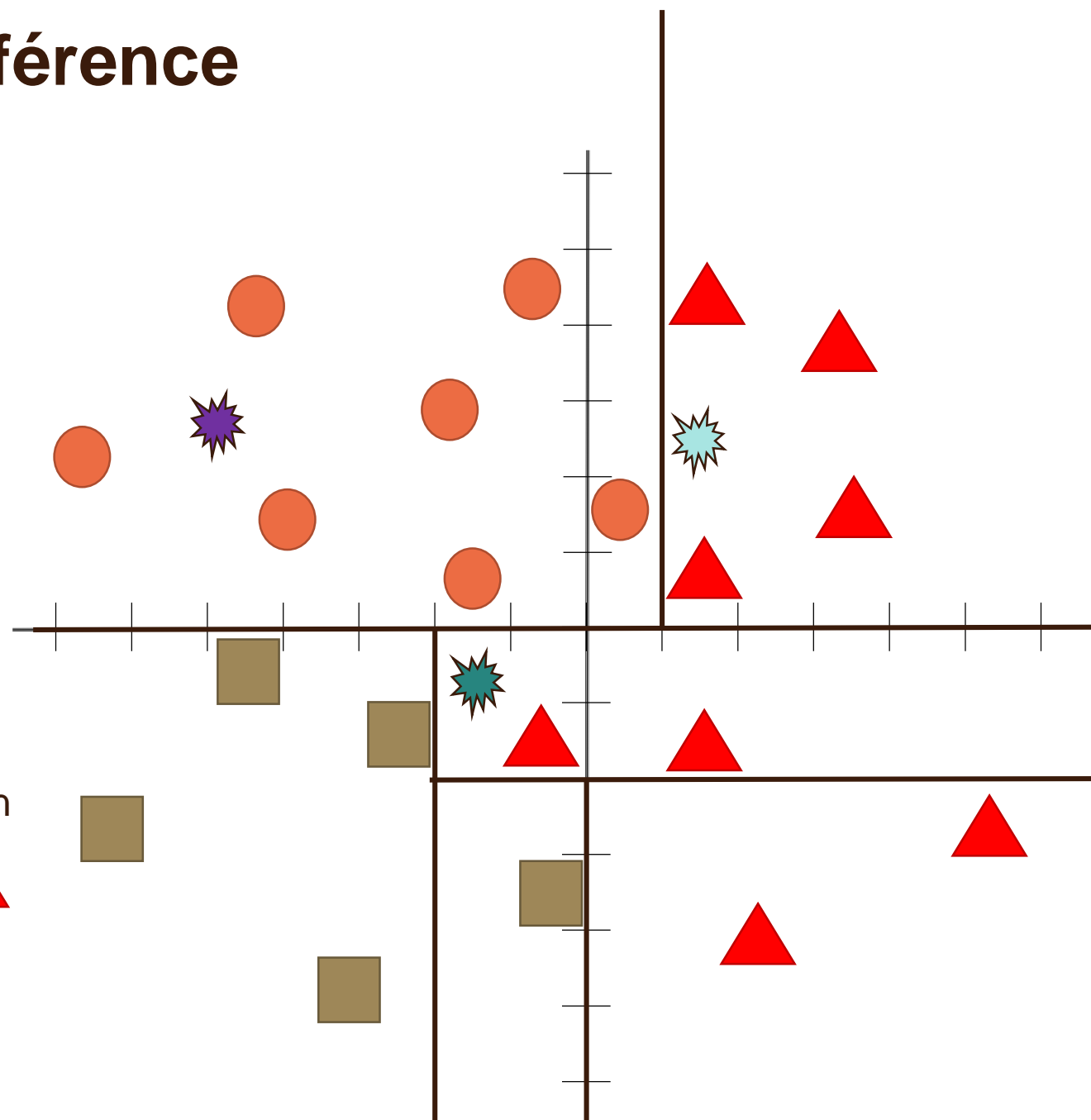
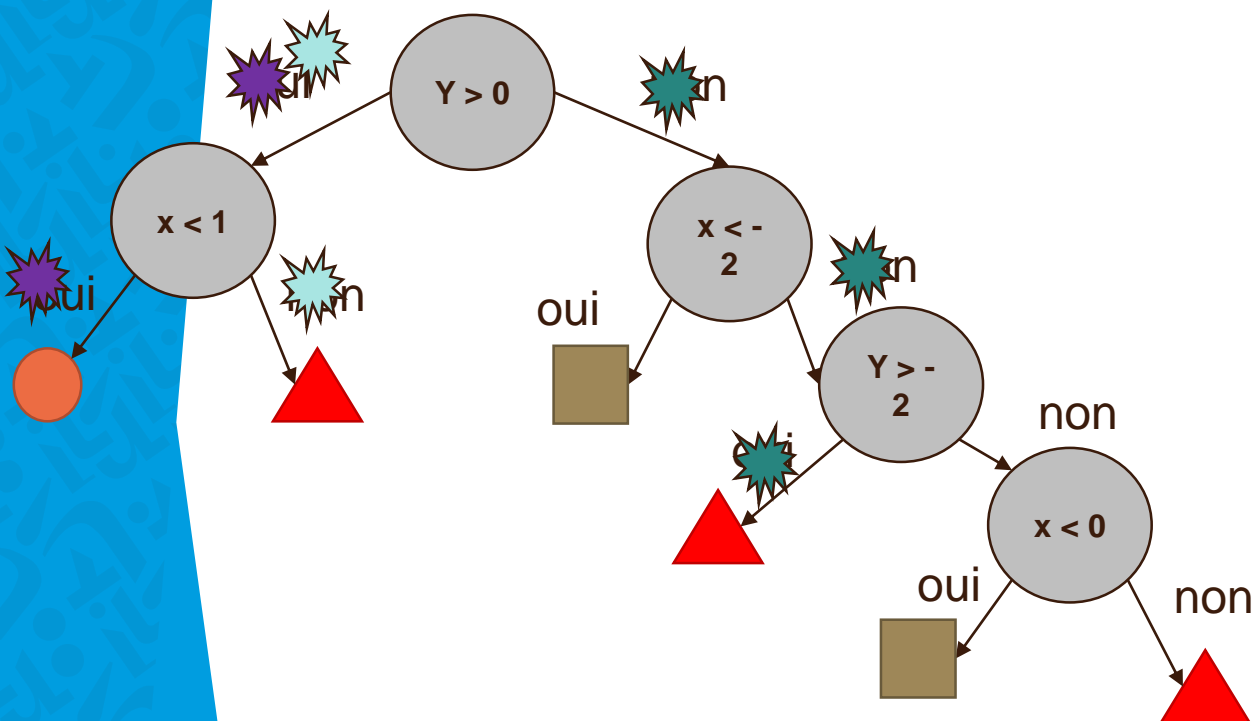
Paramètres : définition de l'arbre
F: code qui exécute un arbre
Hyperparamètres : configuration de l'optimiseur
Optimiseur: algorithme de construction de l'arbre



Arbre de décision : apprentissage

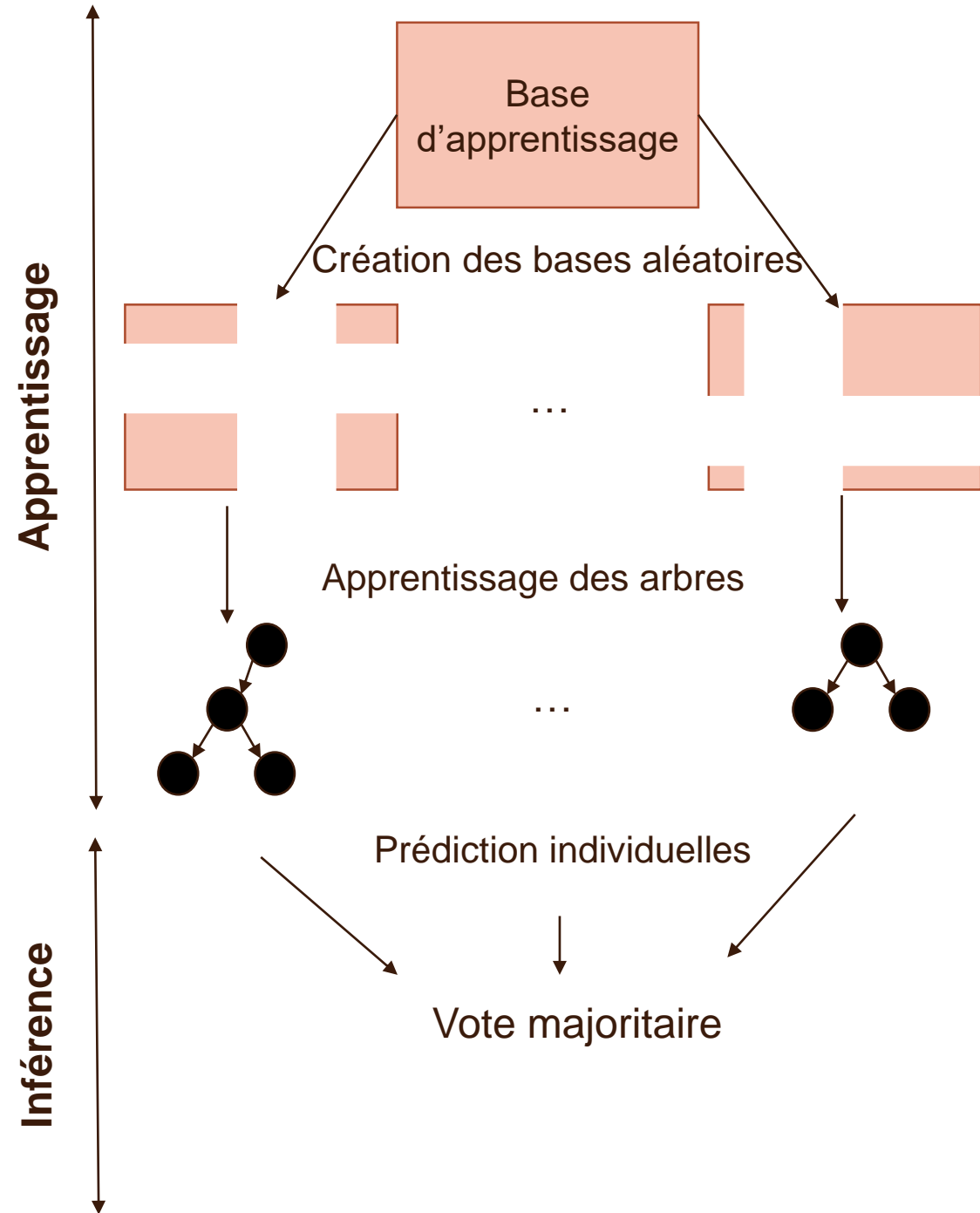


Arbre de décision : inférence



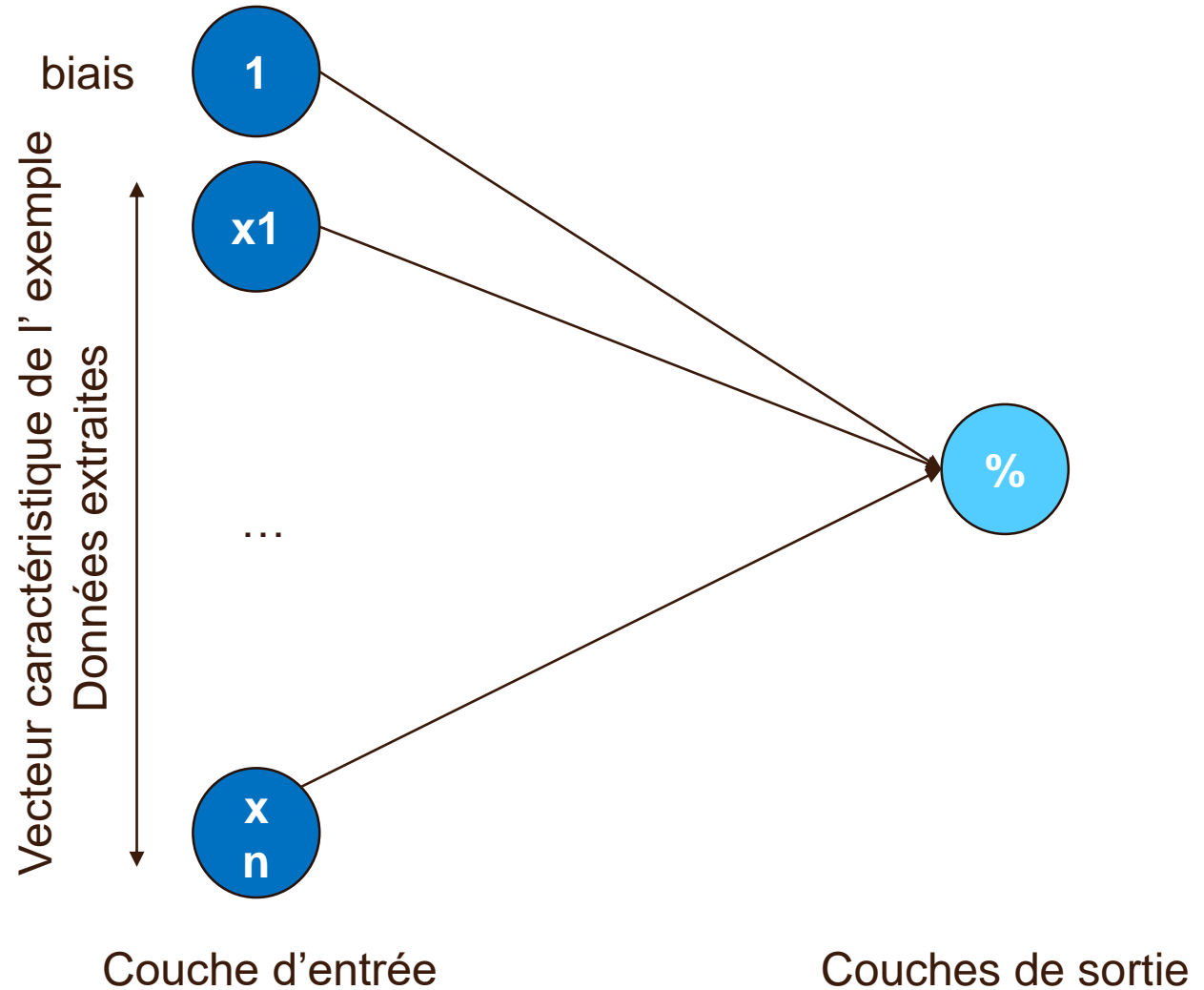
Forêt aléatoire

- Ensemble d'arbres
- Chaque arbre
 - Est appris avec un sous ensemble des exemples
 - Est appris avec un sous ensemble des caractéristiques
 - A des performances médiocres
- La forêt
 - Agrège les résultats de ses arbres
 - A de très bonnes performances

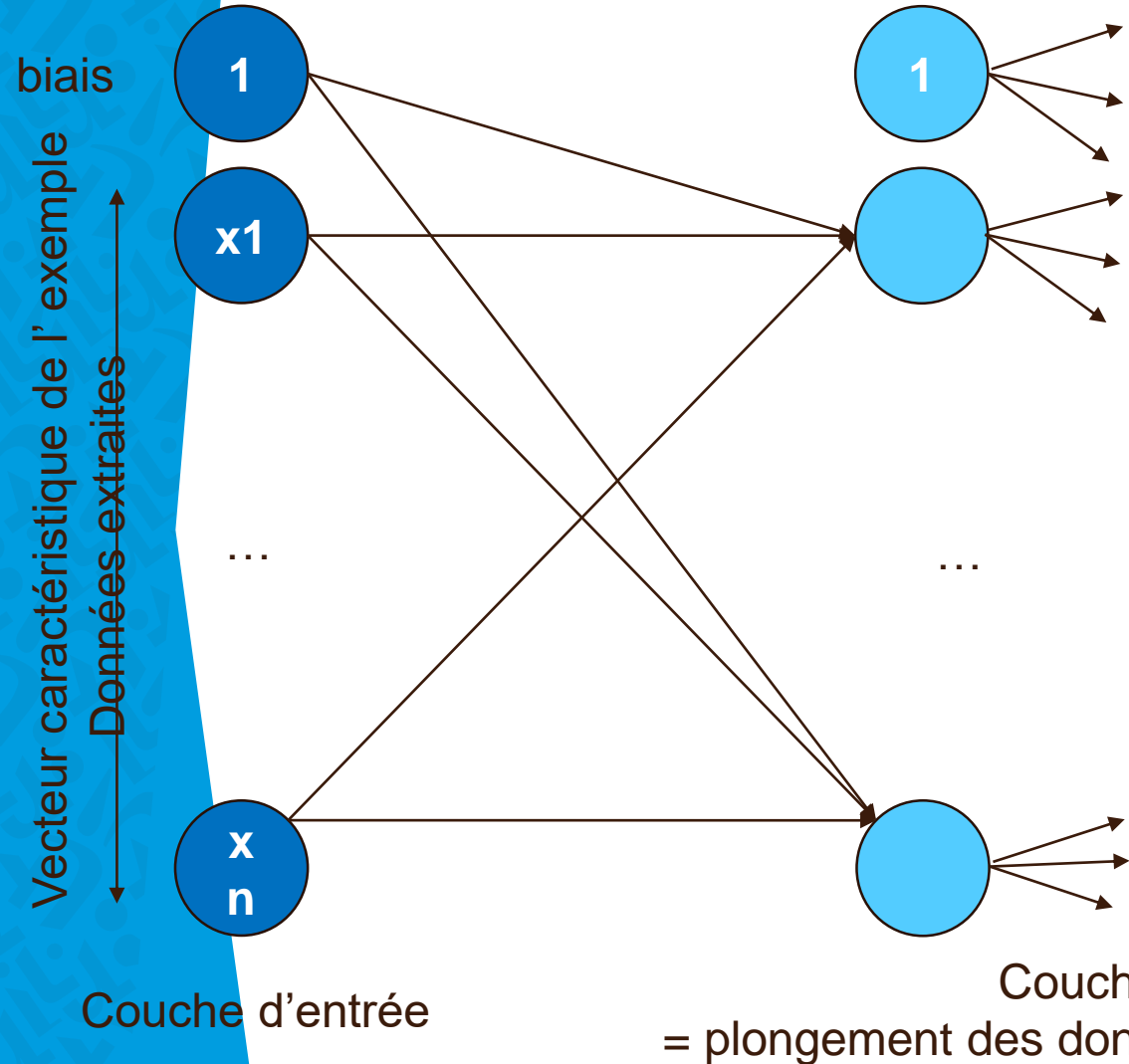


Régression logistique

- Paramètres** : ensemble des poids
- Hyperparamètres** : fonction de régularisation
- F**: code qui exécute la somme pondérée et l'activation
- Optimiseur**: descente de gradient, méthode de newton, optimiseurs hadock



Réseau de neurones multi couches

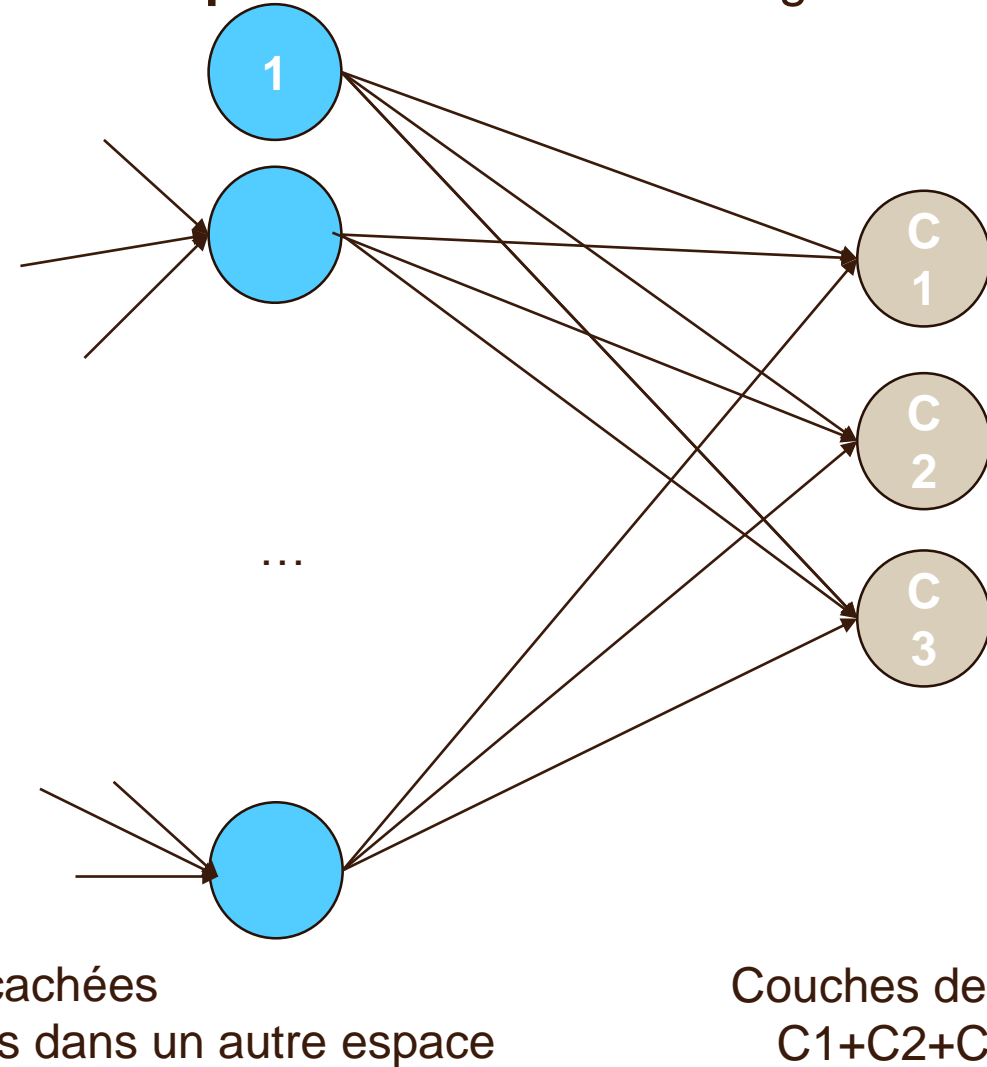


Paramètres : ensemble des poids (=arêtes)

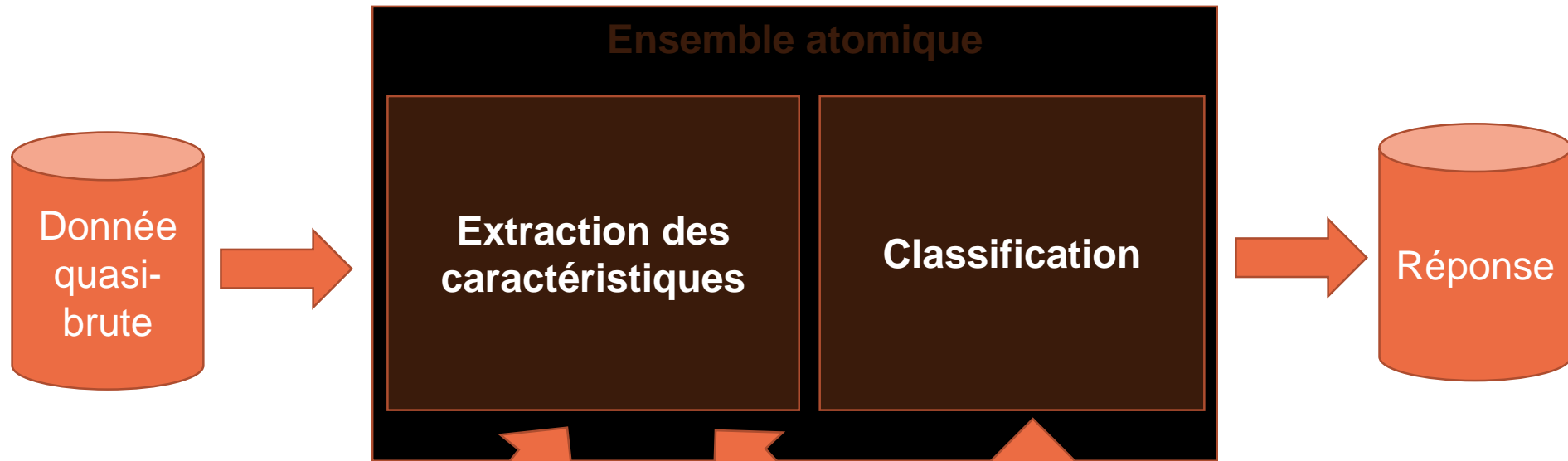
Hyperparamètres : profondeur et largeur du réseau

F: code qui exécute les opérations mathématiques du réseau

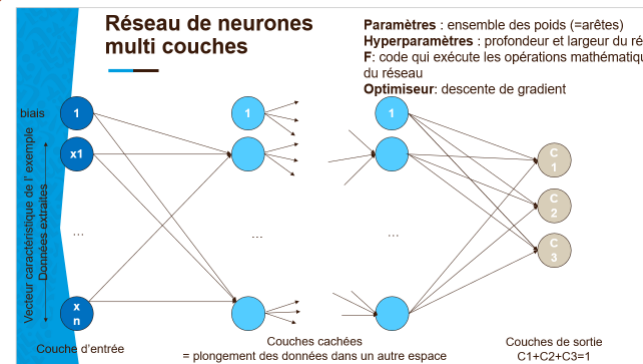
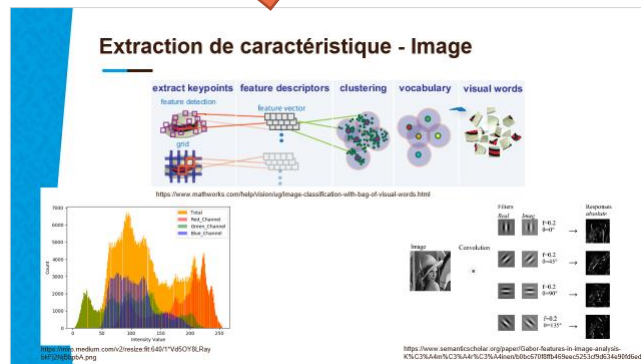
Optimiseur: descente de gradient



Réseau de neurones profond



Un réseau fait l'extraction
De nombreuses astuces sont utilisées pour diminuer le nombre de poids

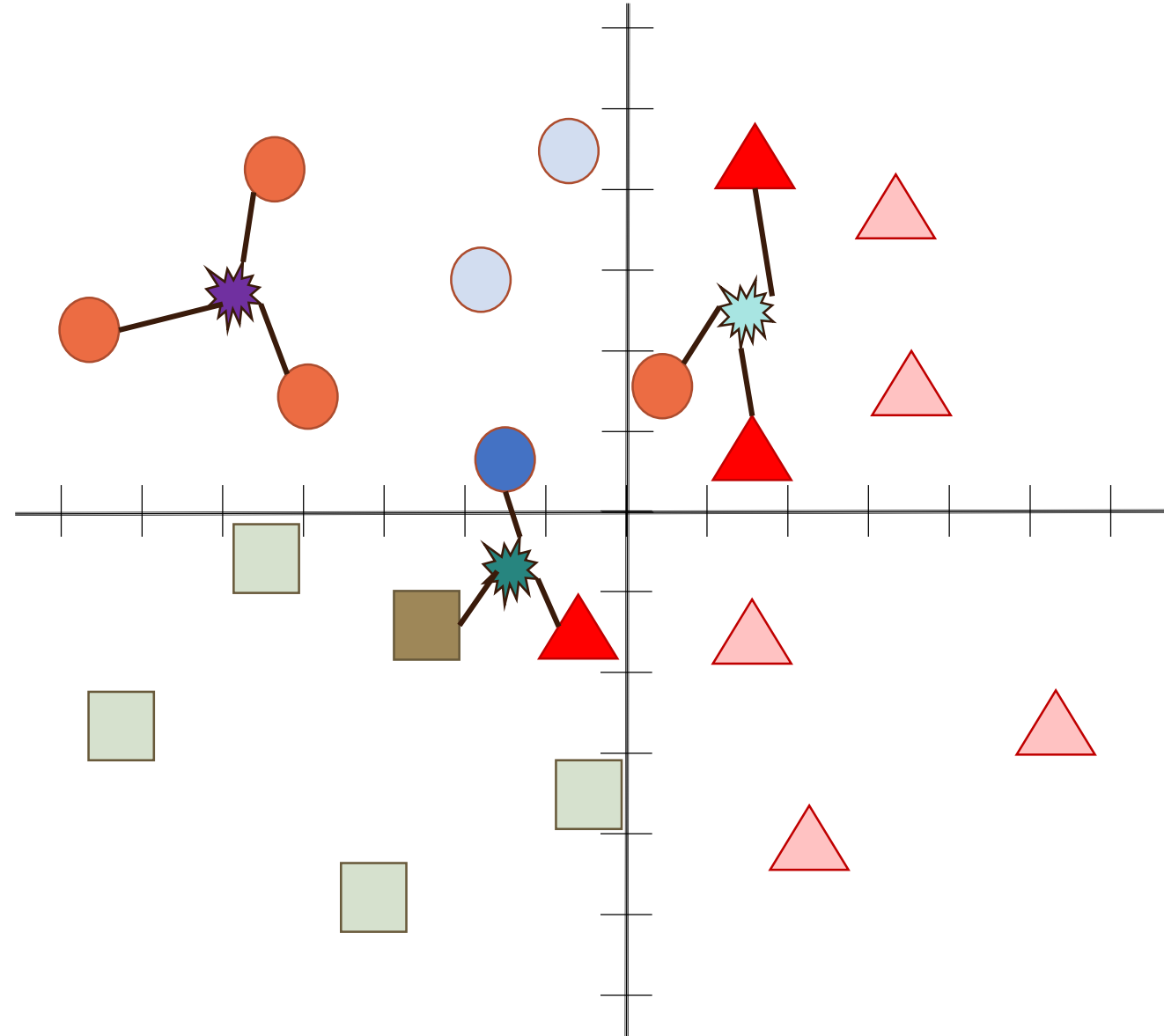
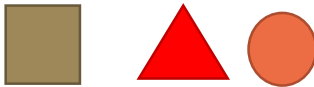
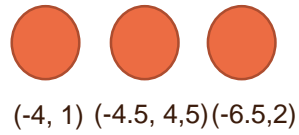


La classification est faite grâce à un MLP.
La plupart des poids sont ici

Explications intrinsèques

**Les modèles
proposent
directement des
explications**

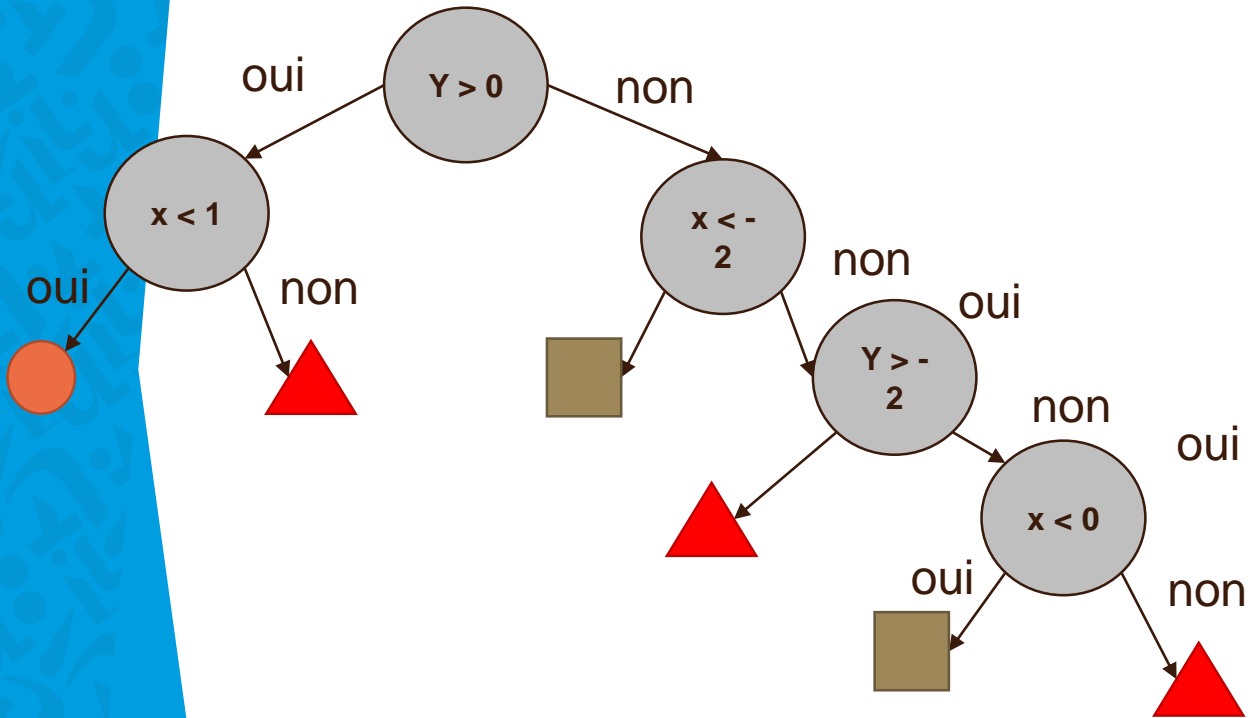
K plus proches voisins : représentation des plus proches voisins



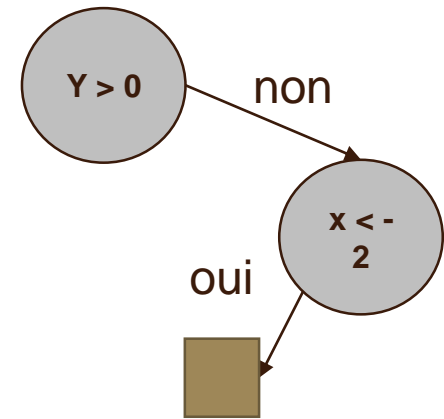
Arbre de décision : représentation de l'arbre ou du chemin emprunté par une donnée



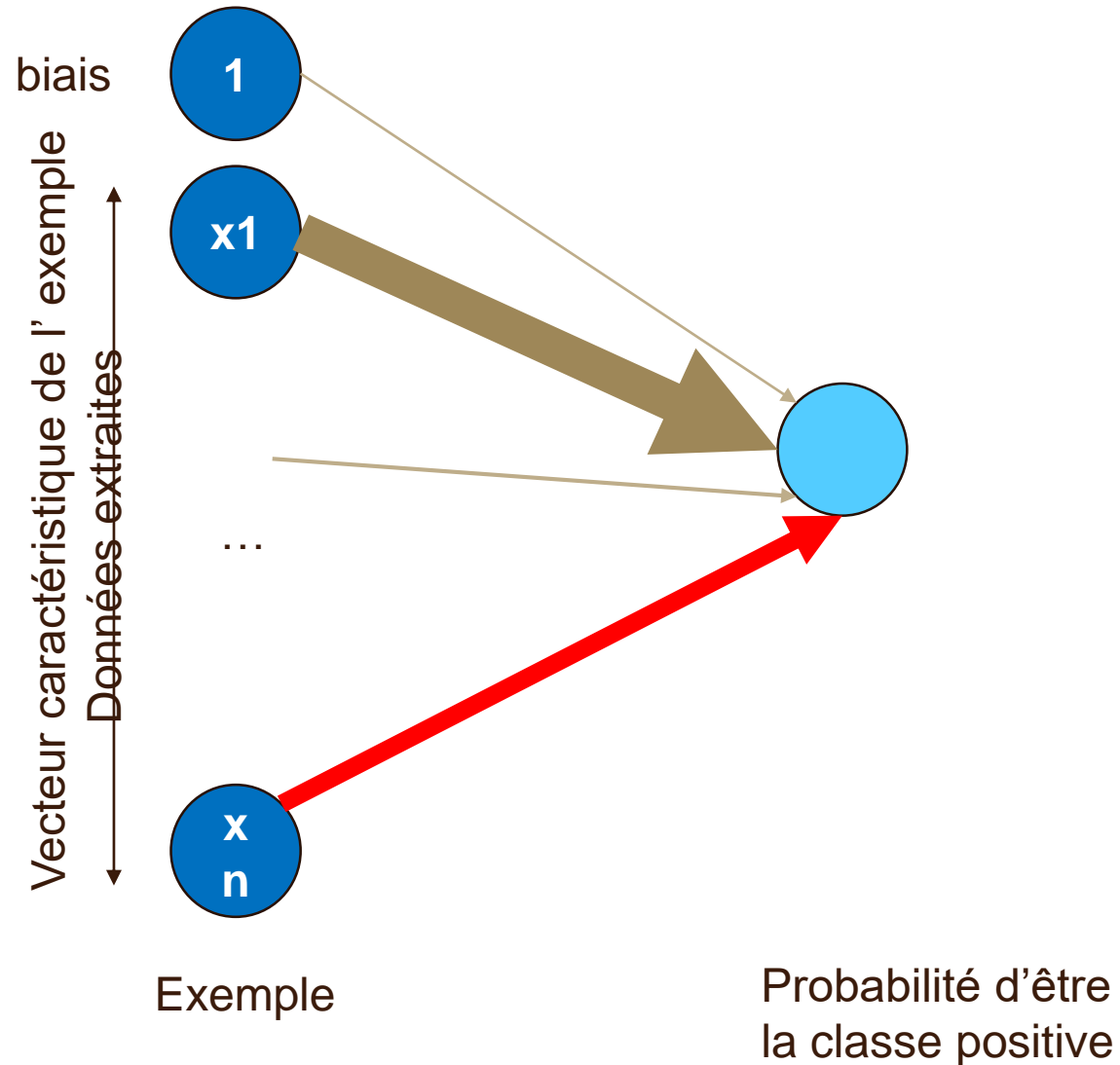
Explication globale



Explication locale



Régression logistique : importance des (valeur absolue des) poids



Limites des explications intrinsèques

- **Limité à des modèles simples**
 - Donc pas forcément les plus performants
- **Ne passe pas l'échelle**
 - Comment visualiser les plus proches voisins lorsqu'on ne connaît pas la sémantique des caractéristiques et que la donnée n'est pas une image ?
 - Comment comprendre un arbre de décision trop grand ?
 - Comment lister les poids lorsqu'il y en a trop ?
 - ...

L'analyse visuelle et interactive est une solution : illustration avec la régression logistique



Figure 1: The interface of the prototype where the limb reduction defect is analyzed. (A1) The univariate analysis view. (A2) The descriptive statistics information panel. (B) The variable grouping view. (C) The model evaluation and comparison view.

L'analyse visuelle et interactive est une solution : illustration avec les arbres de décision

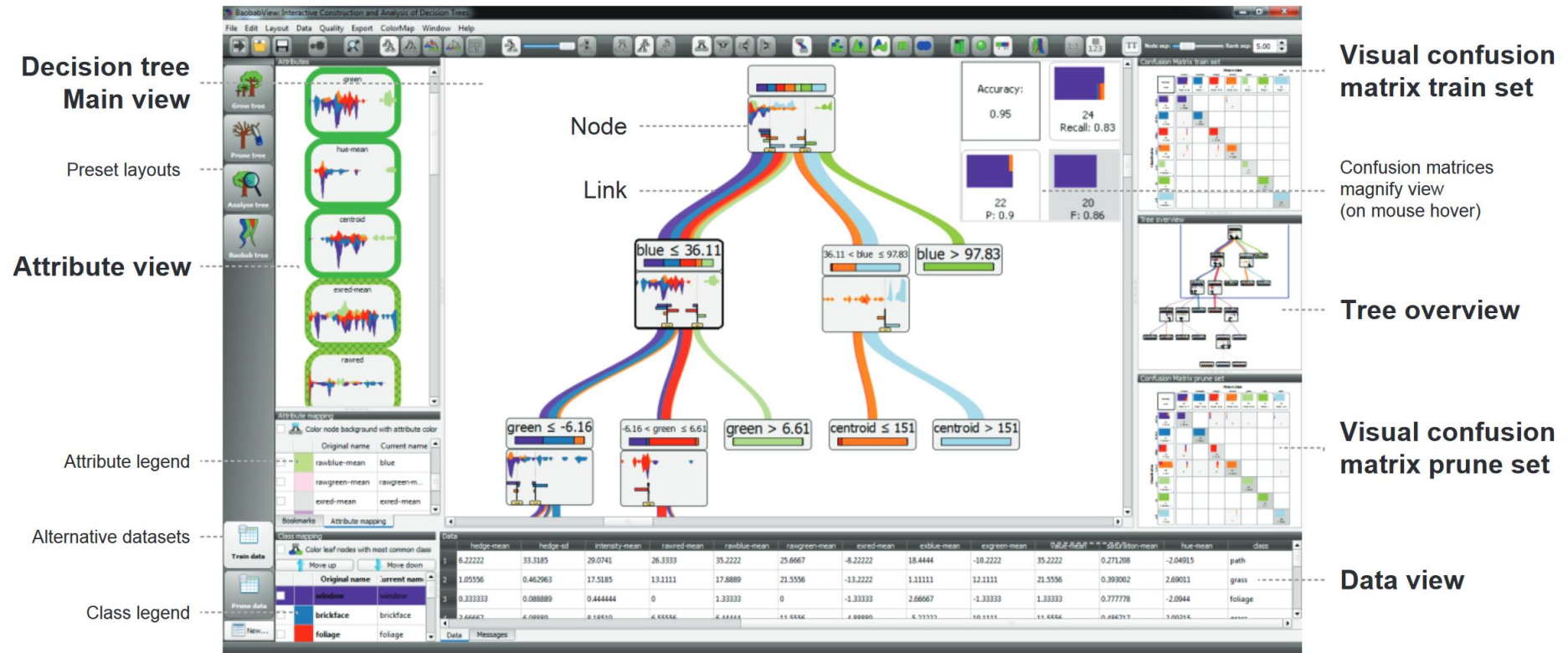


Figure 3: Interface of the interactive decision tree construction software with according proposed decision tree visualization; Based on adapted node-link diagram. Nodes contain important decision tree components. Links are visualized as a stream of items.

L'analyse visuelle et interactive est une solution : illustration avec les forêts aléatoires

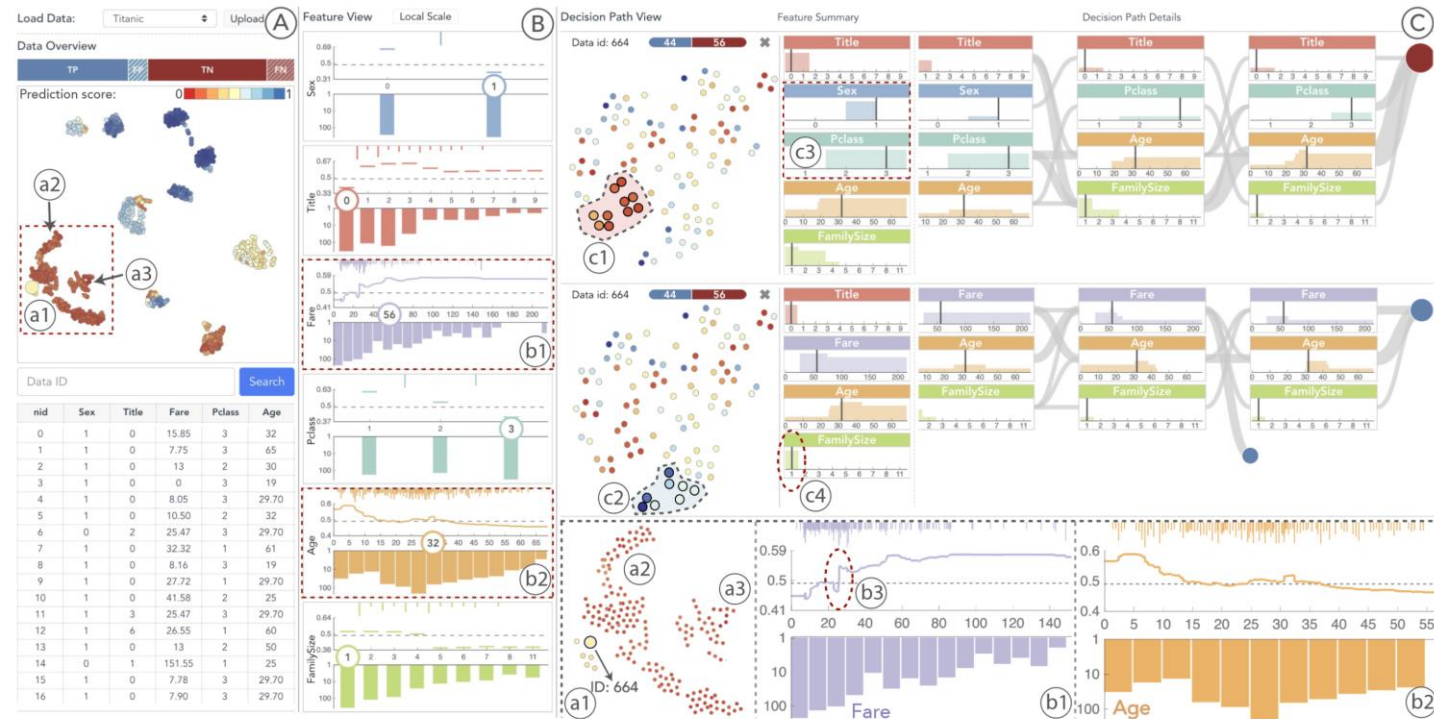


Figure 1. Using iForest to interpret random forests with Titanic dataset: (A) a Data Overview displaying an overview of how random forests classify data; (B) a Feature View depicting the relationships between features and predictions from various perspectives; (C) a Decision Path View revealing the underlying working mechanisms by enabling users to audit and compare different decision paths. iForest allows users to interpret random forests from various perspectives. For example, users can compare the negative decision paths (c1) against the positive ones (c2) to examine the most significant reasons for generating different results.

Discussion sur les explications intrinsèques

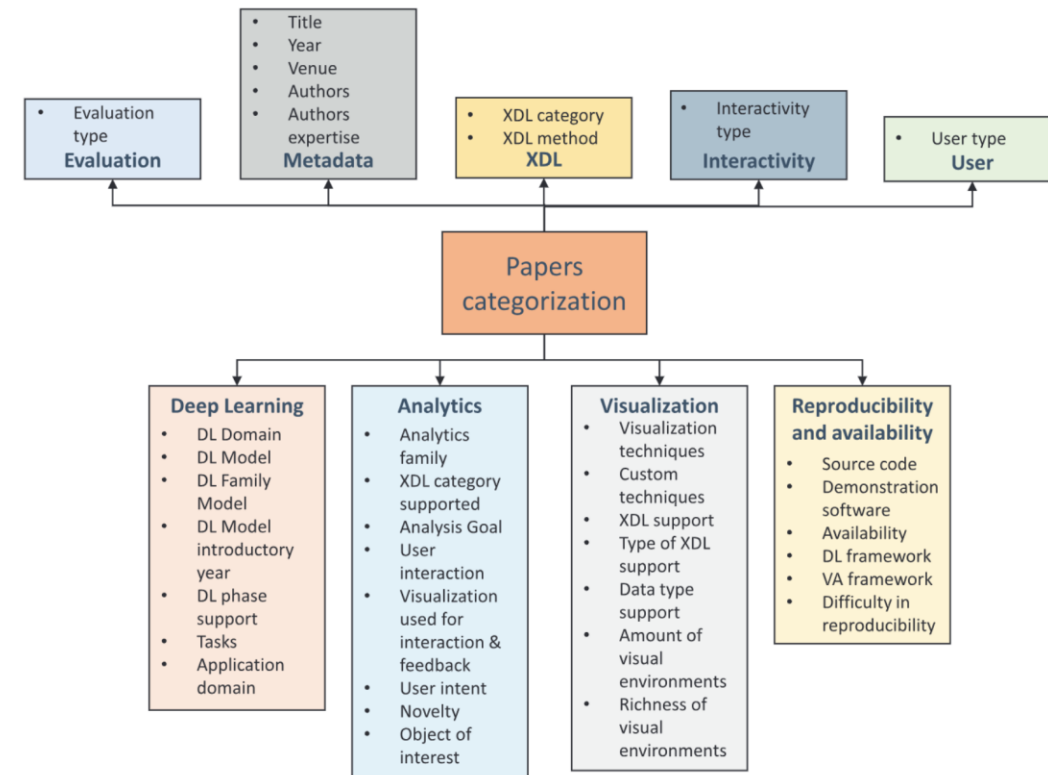
- **Certaines familles de modèles sont auto-explicables**
 - Globalement sur l'ensemble des données
 - Localement pour un exemple particulier
- **Cependant, elles passent rarement l'échelle**
- **L'analyse visuelle et interactive peut aider à mieux appréhender le modèle**
 - Il est nécessaire de définir les tâches exactes à résoudre
 - La courbe d'apprentissage de tels outils est souvent longue

Explications après-coup

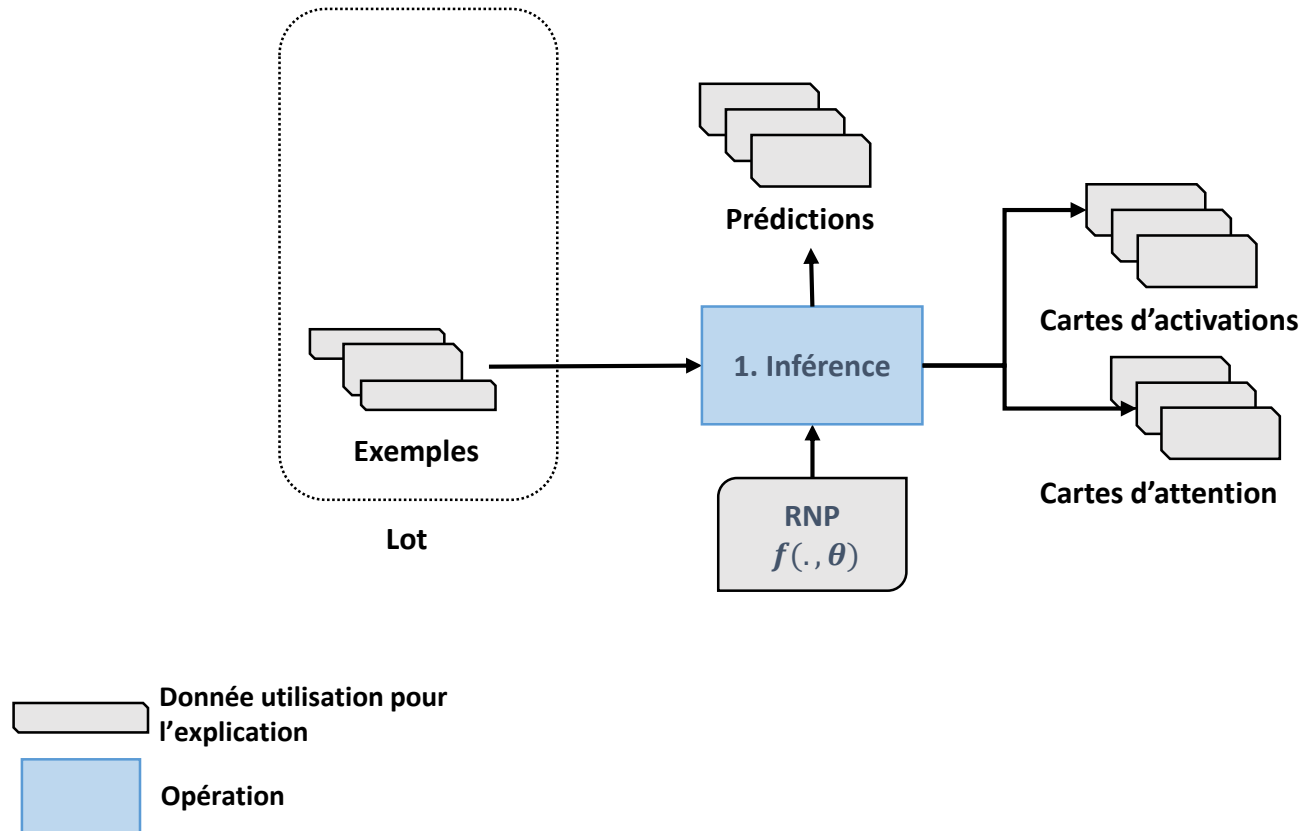
Des post-traitements cherchent à comprendre le modèle. On se focalise sur l'apprentissage profond dans cette partie

Etat de l'art de l'analyse visuelle et interactive pour l'explicabilité de l'apprentissage profond

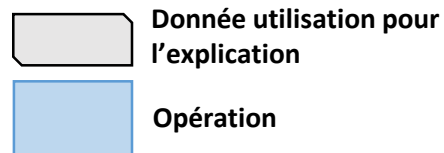
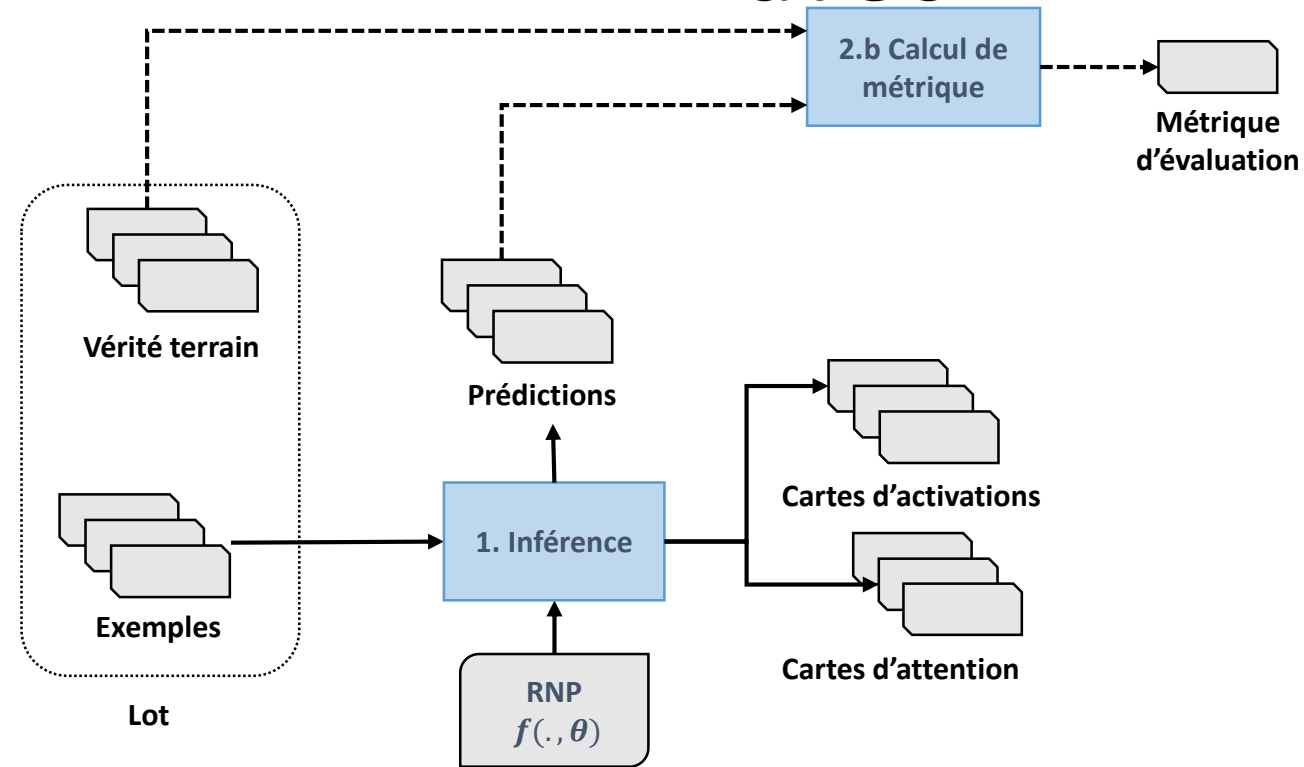
- Analyse de la littérature
- Collection de 196 articles
- Conservation de 67 d'entre eux
- Analyse suivant 38 caractéristiques



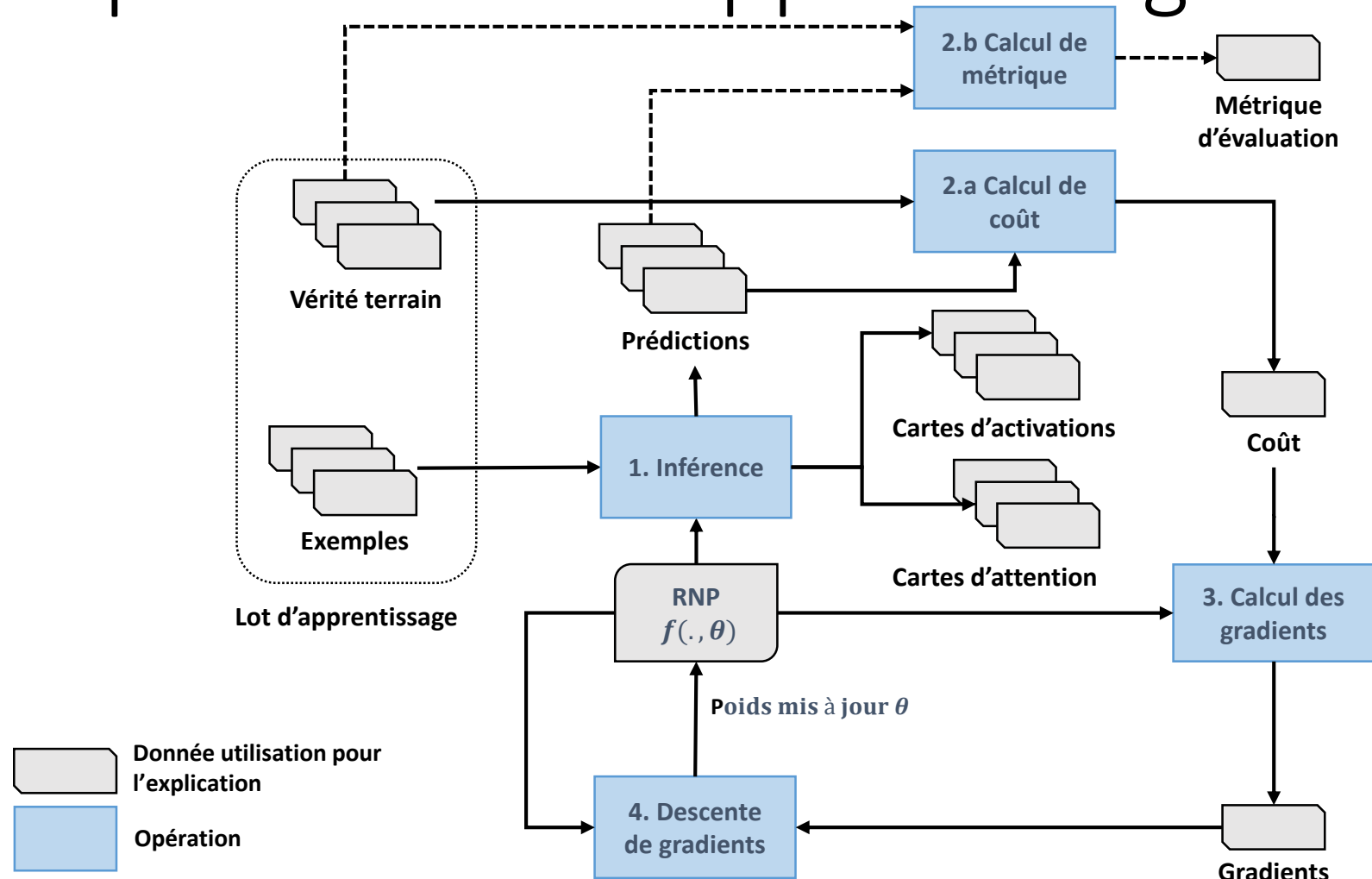
Ensemble des données accessibles pour expliquer : à l'inférence **sans** vérité terrain



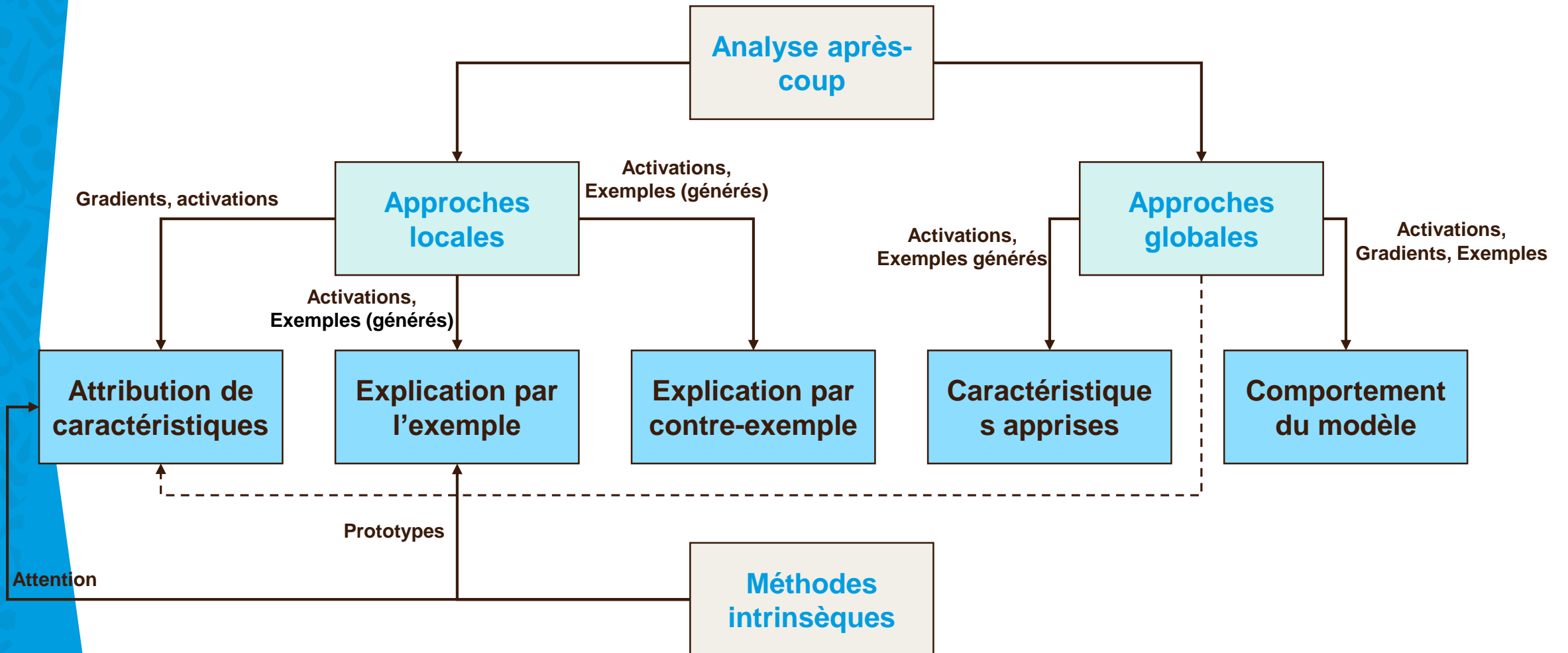
Ensemble des données accessibles pour expliquer : à l'inférence **avec** vérité terrain



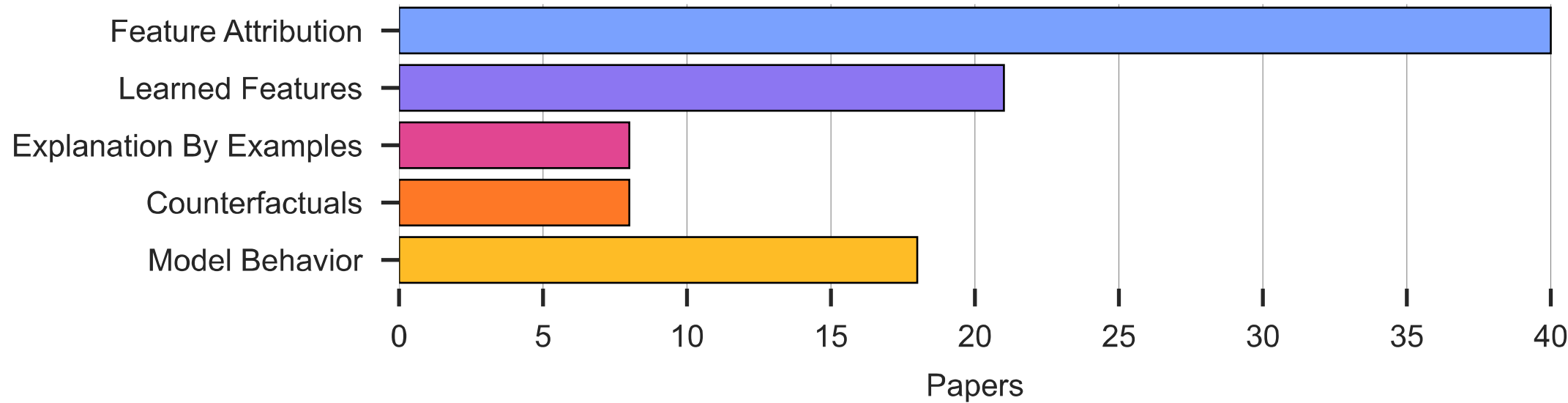
Ensemble des données accessibles pour expliquer : lors de l'apprentissage



Taxonomie des méthodes d'explication



Nette sur-représentation des outils utilisant l'attribution de caractéristiques

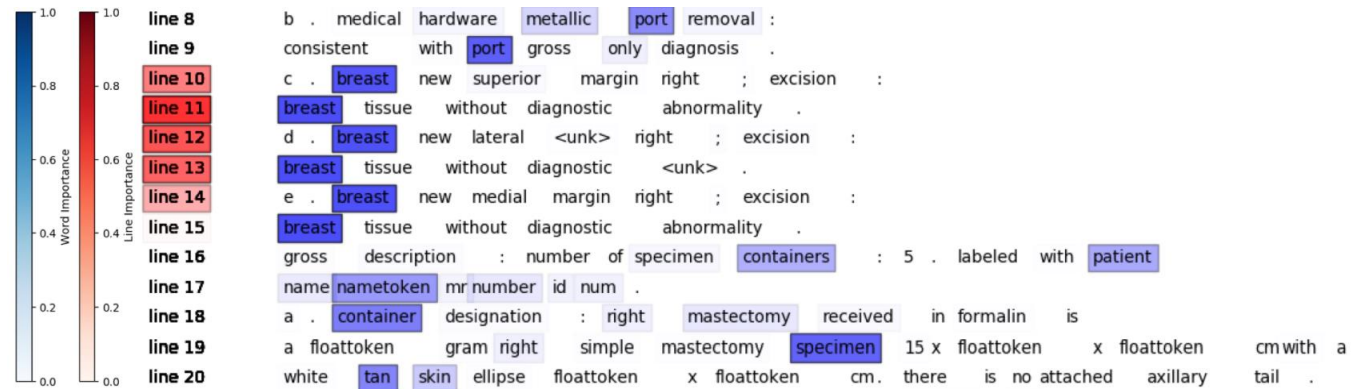


Attribution de caractéristiques

- Sur quelles données d'entrée le modèle se focalise ?
 - Attribution de score sur chaque donnée d'entrée
- Catégorie la plus étudiée dans la littérature
- Supporte différents types de données
- Utilise différents types de représentations



Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.



Chae J., Gao S., Ramanathan A., Steed C. A., Tourassi G.: Visualization for Classification in Deep Neural Networks. Tech. Rep., Oak Ridge National Laboratory (ORNL), Oak Ridge, TN, United States, Oct. 2017

Attribution de caractéristiques : plusieurs tâches sont réalisables

- Vérification de la robustesse du modèle
- Analyse des erreurs
- Exploration de données
- Contrôle de l'amélioration du modèle
- Analyse « quoi si ? »
- Détection de biais
- ...

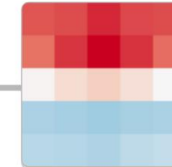


Park, Cheonbok, et al. "Vatun: Visual analytics for testing and understanding convolutional neural networks." *Eurographics Conference on Visualization (EuroVis)-Short Papers*. 2021.

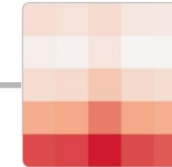
Caractéristiques apprises

- **Qu'est-ce que les données ont appris durant l'apprentissage ?**
 - Génère les caractéristiques qui excitent fortement un neurone
- **Principaux défis**
 - Gérer les milliers de neurones et couches
- **Solutions**
 - Abstractions et partitionnement
 - Encodage spécifiques aux données

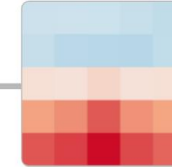
Windows (4b:237)
excite the car detector
at the top and inhibit
at the bottom.



Car Body (4b:491)
excites the car
detector, especially at
the bottom.



Wheels (4b:373) excite
the car detector at the
bottom and inhibit at
the top.



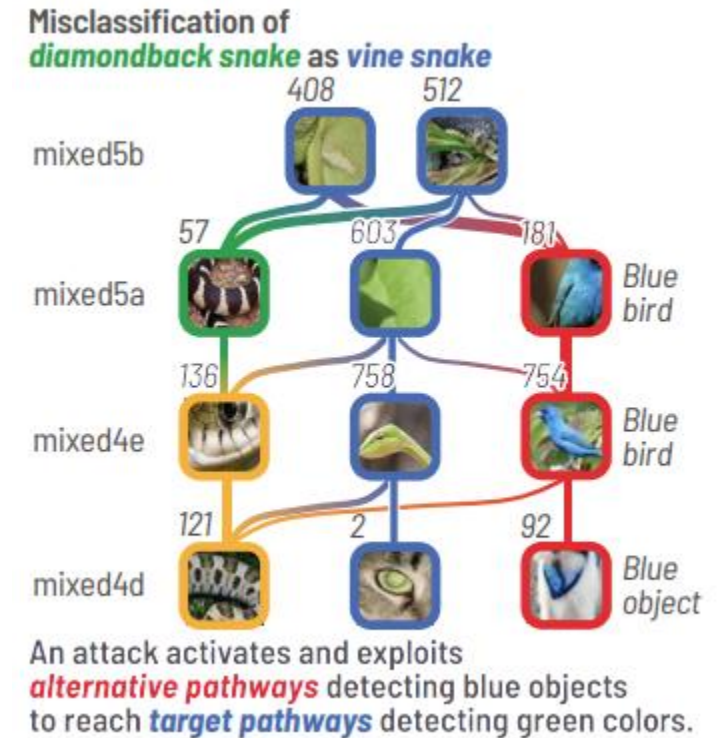
● positive (excitation)
● negative (inhibition)



A **car detector (4c:447)**
is assembled from
earlier units.

Caractéristiques apprises : plusieurs tâches sont possibles

- Diagnostiquer les erreurs lors de l'apprentissage
- Vérifier la connaissance apprise
- Comparer les modèles
- Explorer le rôles des neurones et couches
- Comprendre les attaques adversariales
- ...



Das N., Park H., Wang Z. J., Hohman F., Firstman R., Rogers E., Chau D. H. P.: Bluff: Interactively deciphering adversarial attacks on deep neural networks. In Proceedings of the 2020 IEEE Visualization Conference (VIS)

Explication par exemples

- **Quels exemples sont considérés similaires par le modèle ?**
 - **Caractéristiques, activations, représentation, sortie, ...**
- **Buts**
 - **Compréhension de la représentation de la donnée par le réseau**
 - **Prédictions**

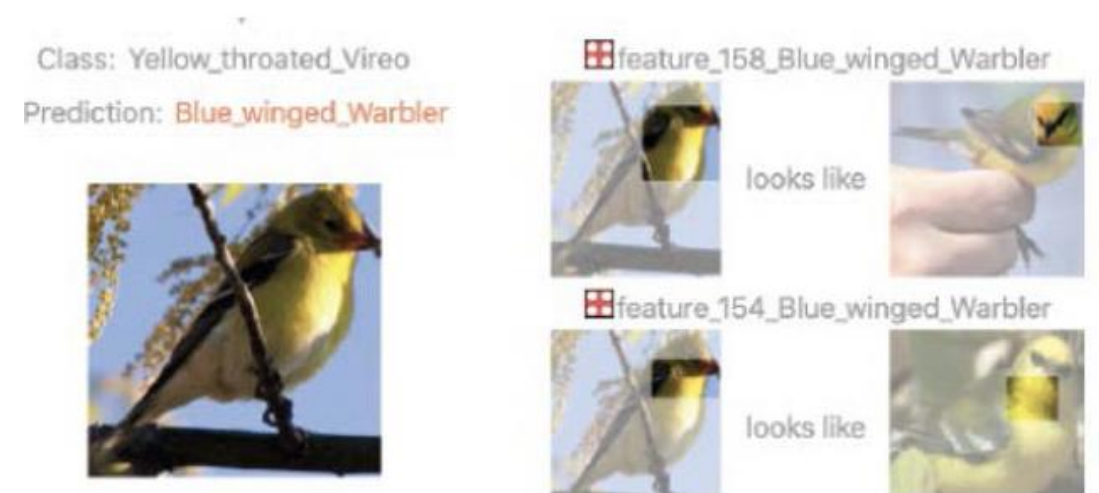


Similaire à



Explication par exemples : plusieurs tâches sont réalisables

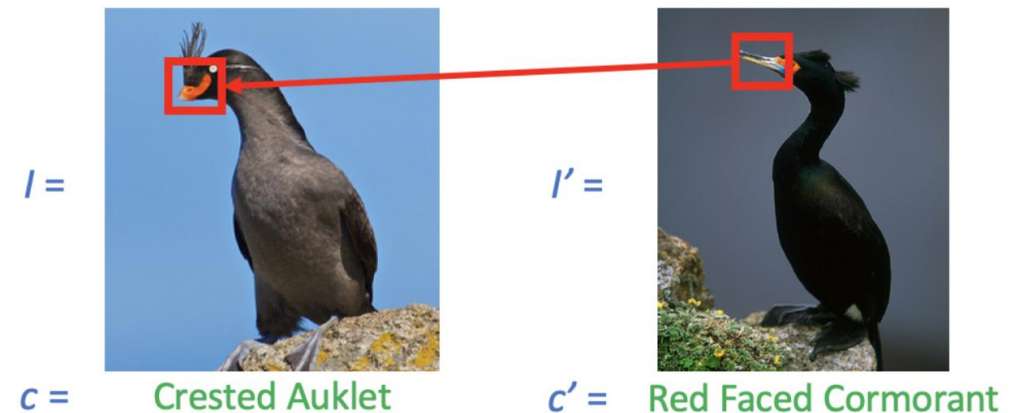
- Identification d'erreurs
- Correction des hyperparamètres de l'apprentissage
- Amélioration de l'architecture
- Comparaison de prédictions
- Analyse « quoi si ? »



Chan G. Y.-Y., Bertini E., Nonato L. G., Barr B., Silva C. T.: Melody: Generating and visualizing machine learning model summary to understand data and classifiers together. arXiv preprint arXiv:2007.10614 (July 2020).

Explication par contre-exemple

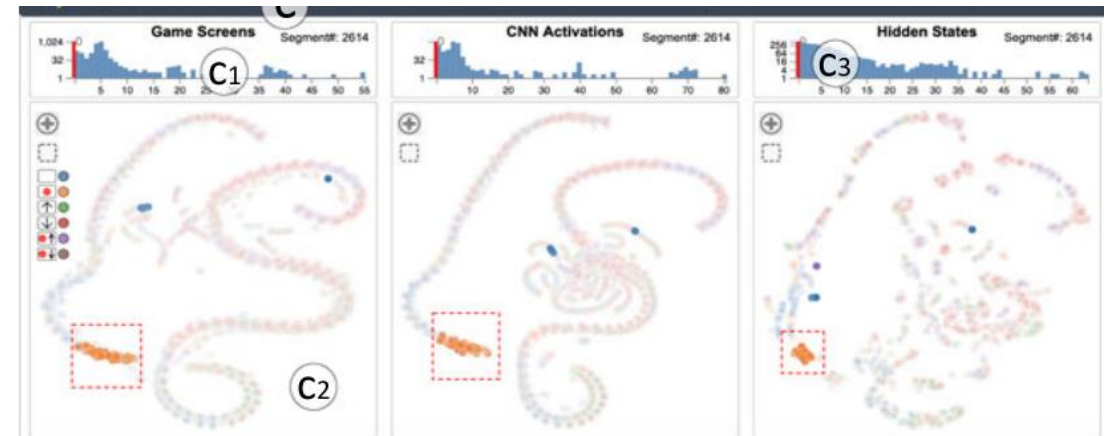
- Que faut-il changer pour obtenir une prédiction différente ?
 - Proposition d'un exemple légèrement différent générant une autre prédiction
 - Ou proposition d'un changement à effectuer



Goyal, Yash, et al. "Counterfactual visual explanations." *International Conference on Machine Learning*. PMLR, 2019.

Comportement du modèle

- Comment le modèle se comporte dans un scénario donné ?
- But
 - Explorer le rôle des différentes couches
 - Visualiser la logique du traitement du réseau
 - Extraire des règles de décision
- Repose majoritairement sur des outils d'analyse visuelle interactive



Autres aspects importants

**Le modèle n'est pas
la seule partie
importante à
analyser**

Les données sont plus importantes que les modèles

- Les performances des modèles dépendent principalement des données d'apprentissage
- Ces données peuvent souffrir de différents biais
 - Il est important de les connaître
 - Rarement pris en compte
- Ces données peuvent être sujet à oubli
 - Il est important de pouvoir les retirer de la connaissance du modèle
 - **Thème de recherche émergent**

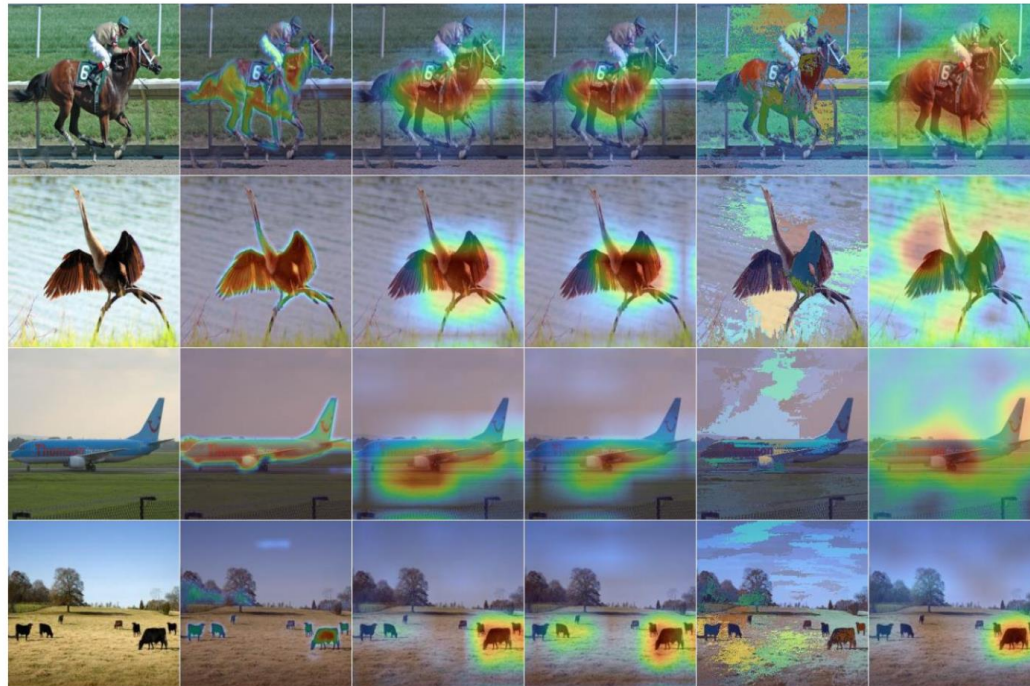


Cabrera, Ángel Alexander, et al. "FairVis: Visual analytics for discovering intersectional bias in machine learning." *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 2019.

Il est nécessaire d'augmenter la confiance aux modèles et leurs explications

Quelle est la vérité terrain d'une explication ?

Comment empêcher la contrefaçon d'explications ?



Raw Image H^2O GradCAM^[21] FEM^[10] LIME^[19] RISE^[17]



Il est nécessaire que le grand public comprenne les limites des systèmes

Les masses de données

- Collecte des données
- Propriété intellectuelle

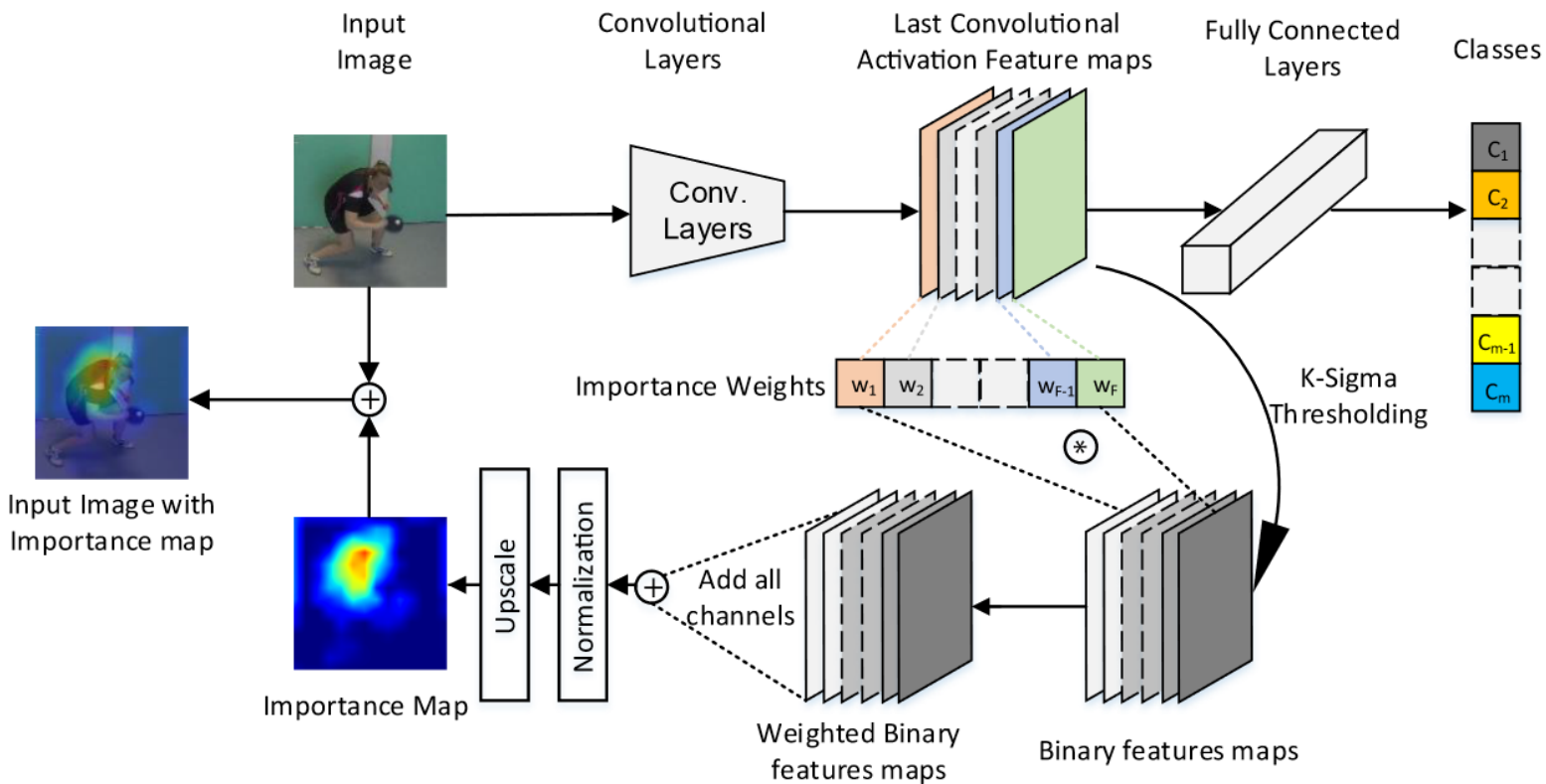
Fiabilité des systèmes

- Ils peuvent se tromper
- Ils peuvent mieux fonctionner dans certains contextes que d'autres
- On ne sait pas toujours les expliquer

Sélection de travaux sur l'explicabilité

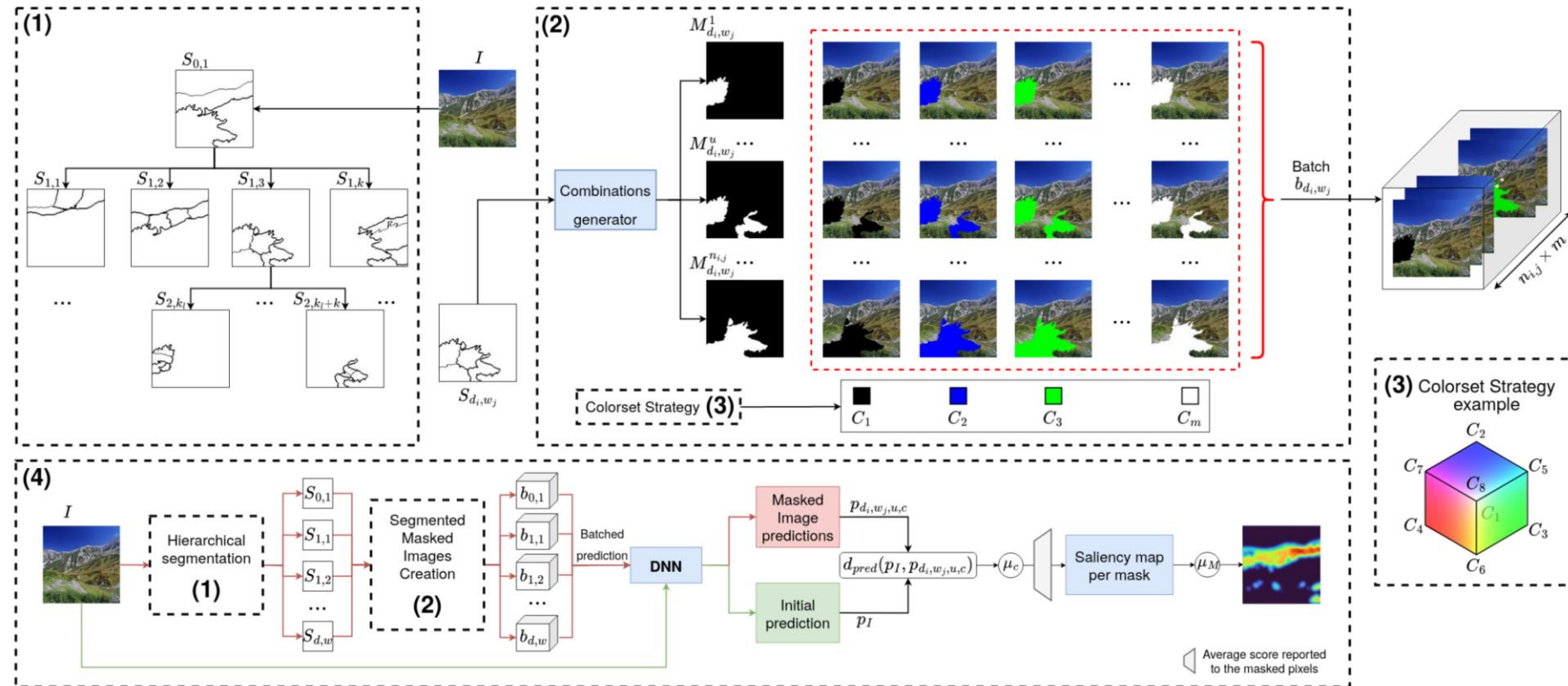


Features Explanation Method

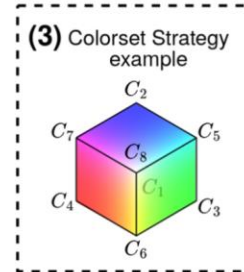


- Attribution de caractéristiques
- Explication locale
- Indépendant de la classe
- Boite blanche
- Activations dans les réseaux de neurones

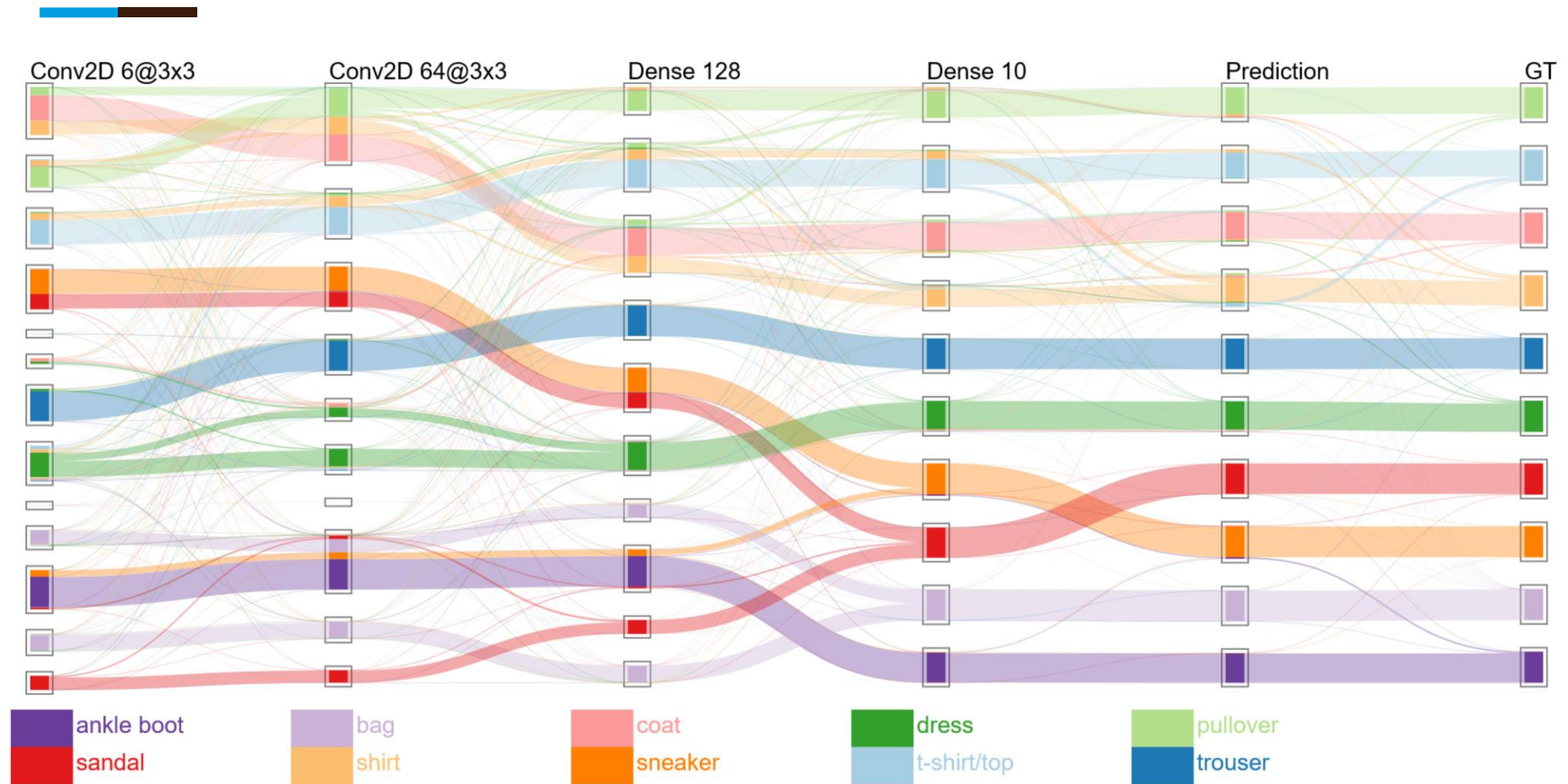
H2O: Heatmap by Hierarchical Occlusion



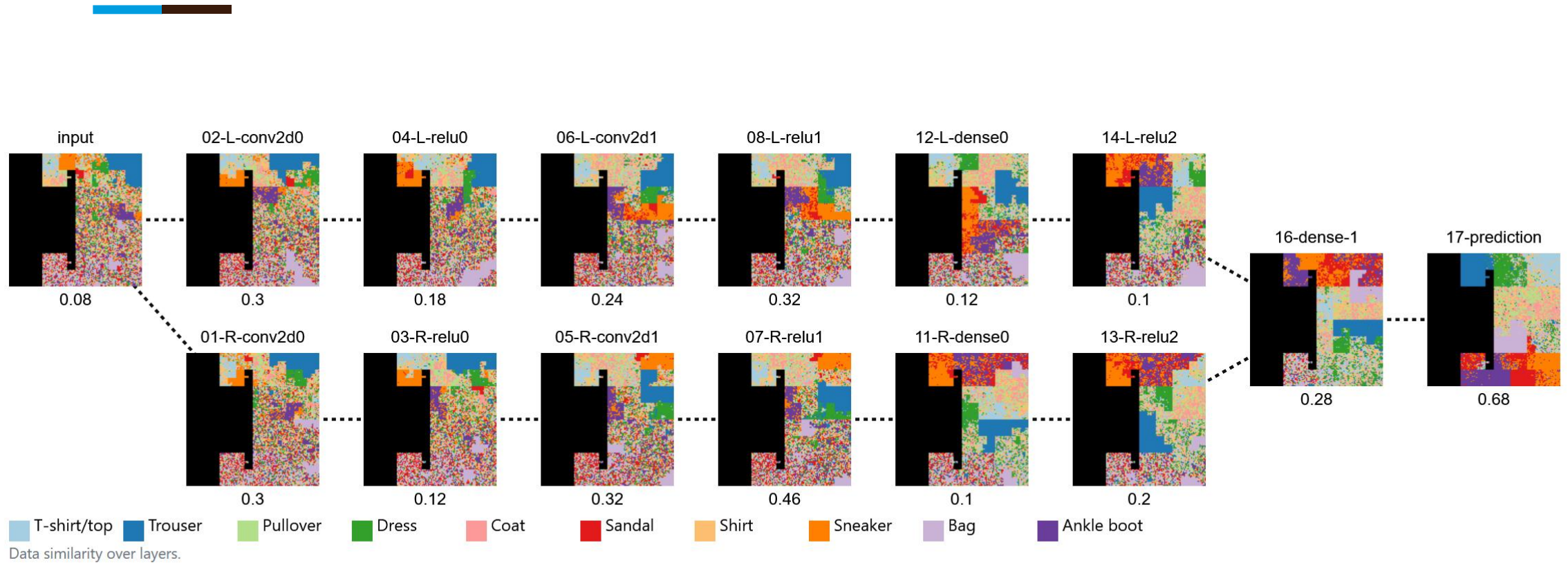
- Attribution de caractéristiques
- Explication locale
- Indépendant de la classe
- Boite noire



Deep Dive into Deep Neural Networks with Flows

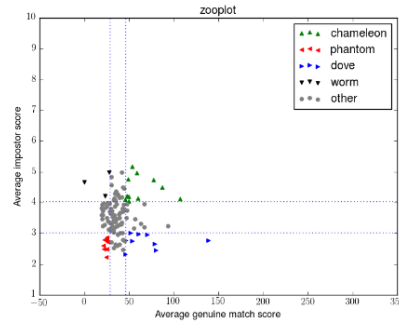


Samples Classification Analysis Across DNN Layers with Fractal Curves

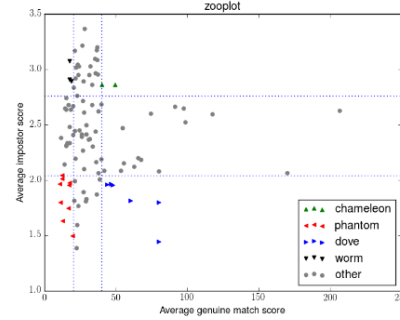


Analyse du comportement du modèle

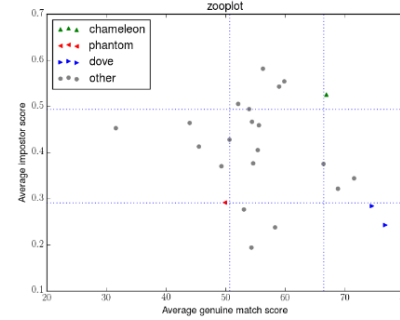
Zoo Graph: a New Visualisation for Biometric System Evaluation



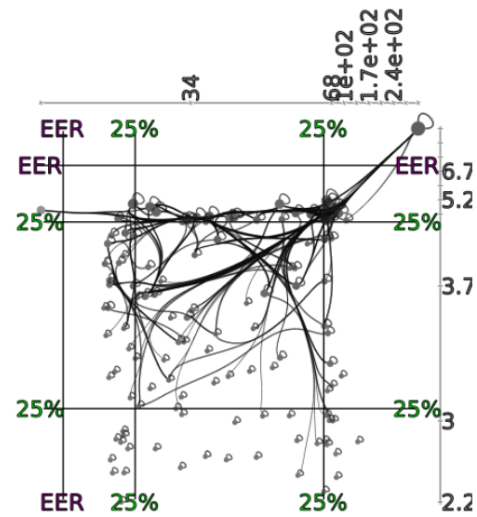
(a) Zoo Plot/ AR



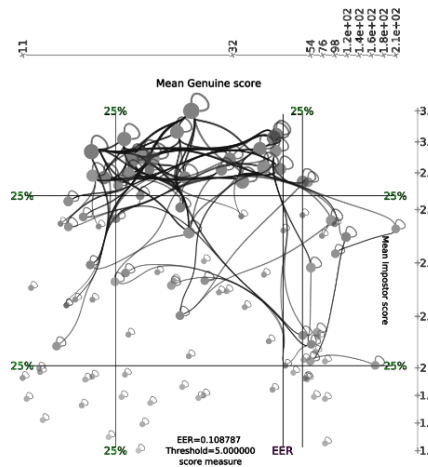
(b) Zoo Plot/ ENSIB



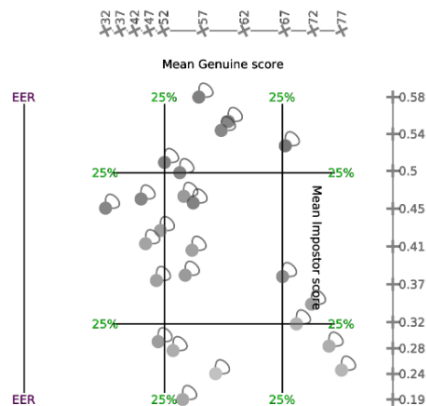
(c) Zoo Plot/ veins



(d) Zoo Graph/ AR



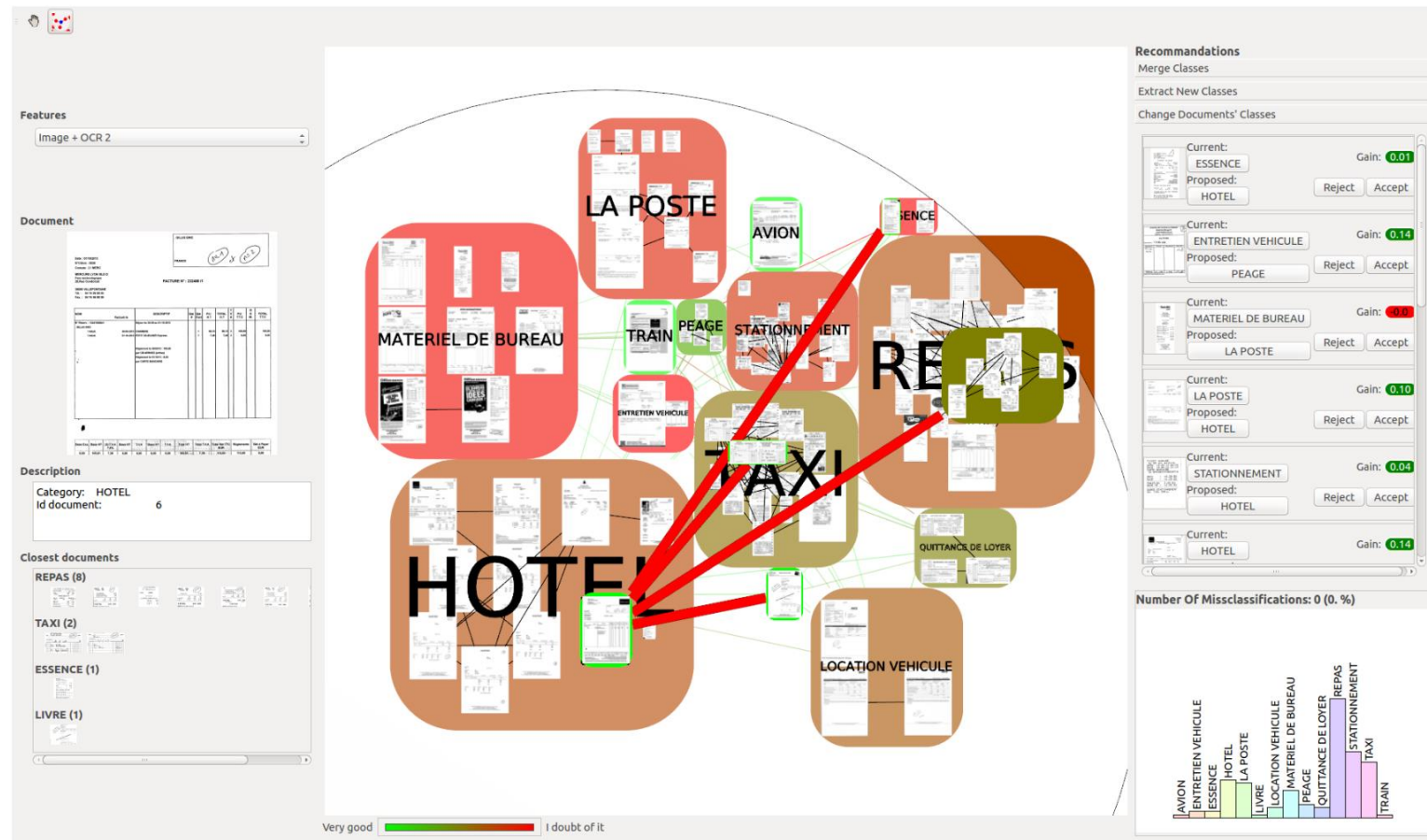
(e) Zoo Graph/ ENSIB



(f) Zoo Graph/ veins

Figure 2. Illustration of the difference between the baseline method (*Zoo Plot*) and the proposal (*Zoo Graph*) for 3 biometric databases.

Visual Graph Analysis for Quality Assessment of Manually Labelled Documents Image Database



Conclusion rapide



Conclusion

- **Les modèles d'apprentissage font des erreurs**
- **Nous sommes capables de les évaluer, mais pas forcément comprendre leurs erreurs**
- **Les méthodes actuelles d'explicabilité ont des limites**
 - **Elles imposent toujours à l'utilisateur d'inférer l'information pertinente**
 - **Il faut ajouter de la sémantique aux explications**
 - **Elles traitent majoritairement de méthodes à base d'images**

Perspectives de recherche

- **Inclusion de sémantique dans les explications**
- **Hybrider les approches intrinsèques et après-coup**
- **Simplifier les systèmes d'explication pour les rendre grand public**
- **Et les sciences sociales ?**