Explicabilité de l’intelligence artificielle

Romain Giot

université de BORDEAUX

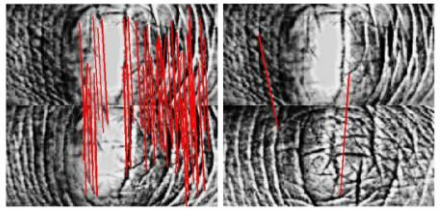
Préambule

Description de la présentation

Romain Giot – MCF HdR en informatique
IUT de Bordeaux / LaBRI

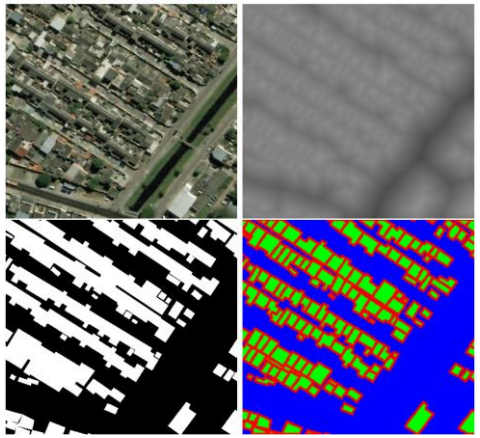
Authentification biométrique

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

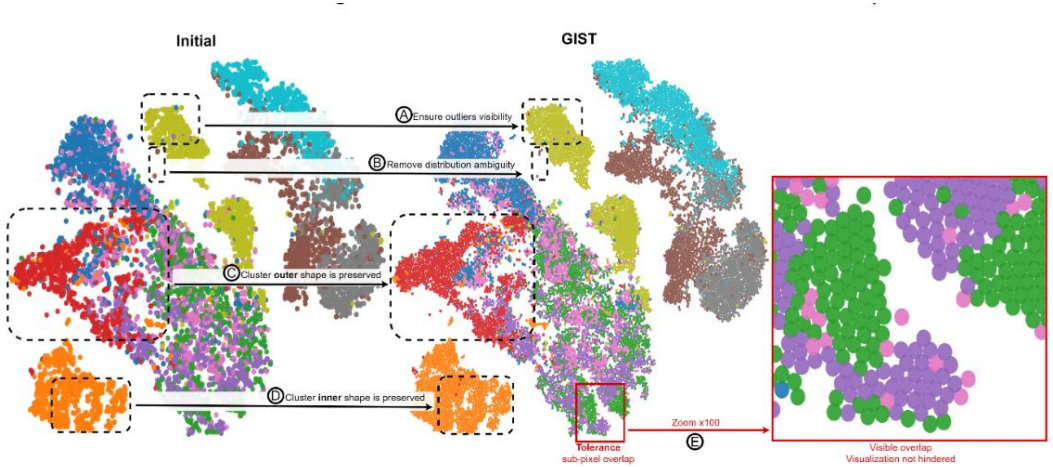
Classification hiérarchique, fusion d'informations

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

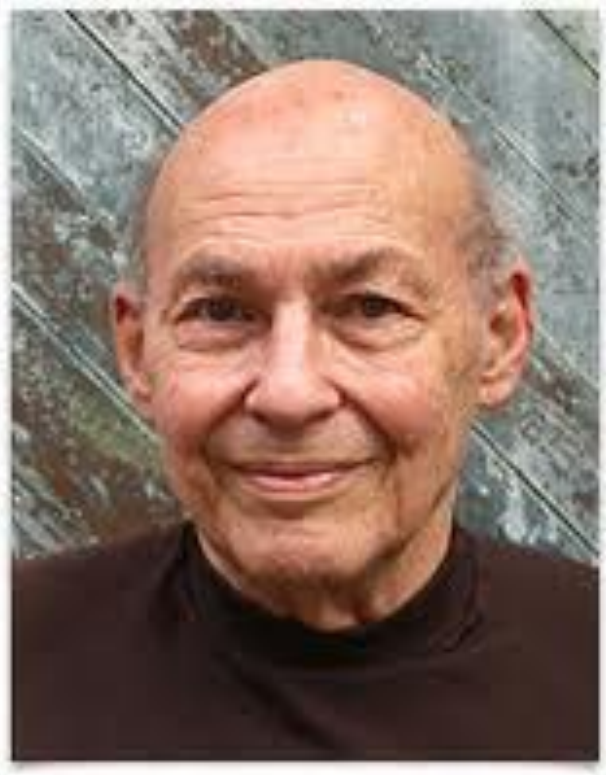
Introduction à l’intelligence artificielle

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 (1956)



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

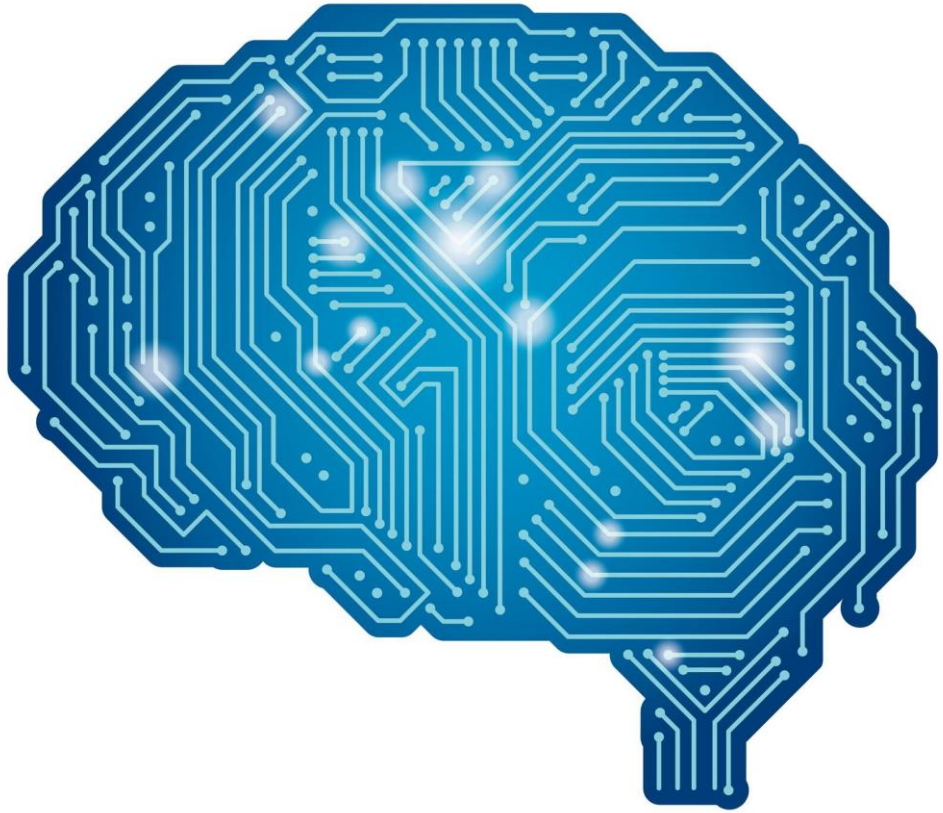
Dans les faits

Avant 2013

- Systèmes symboliques
- Optimisation basée sur des heuristiques
- Algorithmes codés en dur

Depuis 2013

- Apprentissage profond



https://fr.freepik.com/vecteurs-libre/illustration-concept-vecteur-intelligence-artificielle-isolee-fond-blanc_40268935.htm

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**

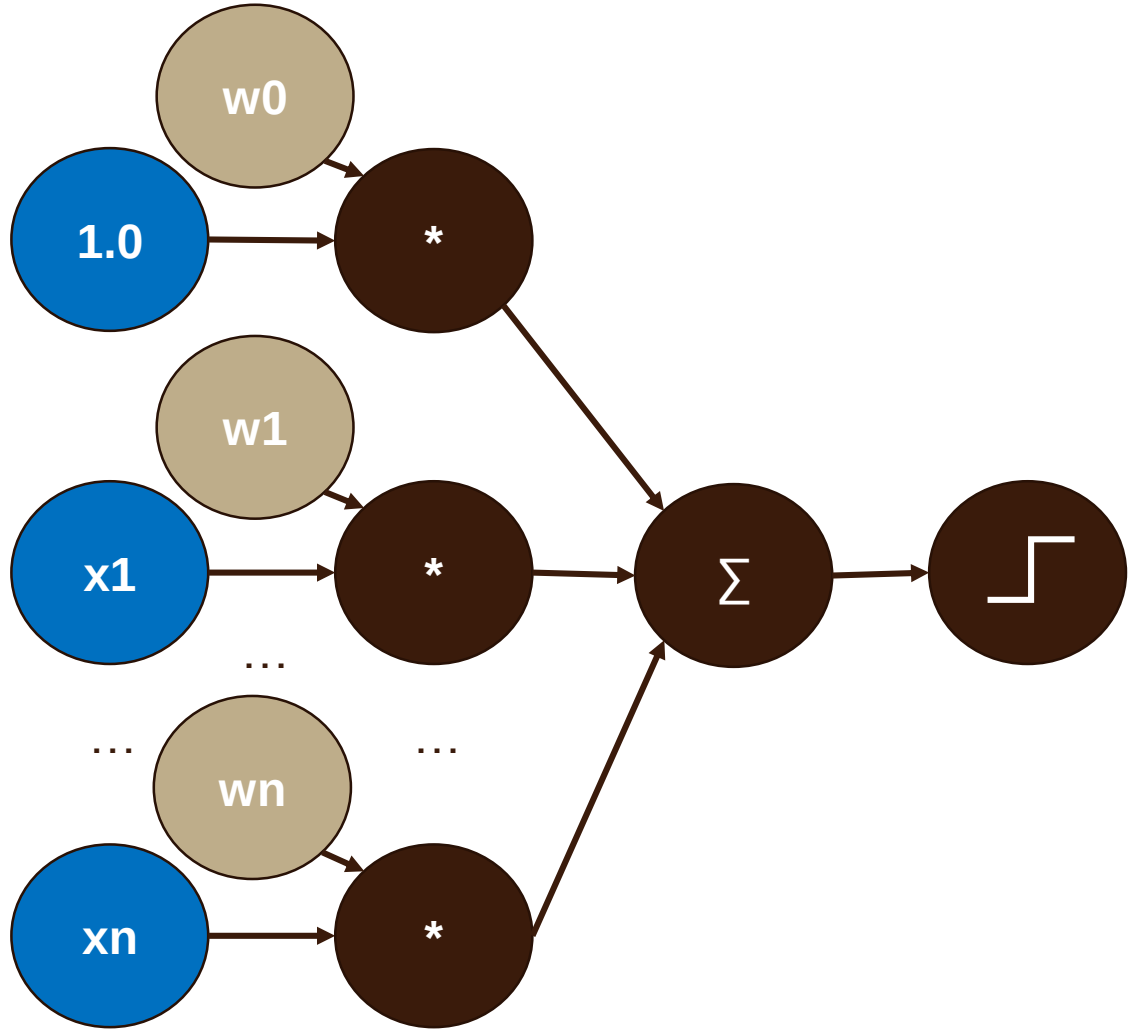
Interprétation

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

G.W. Leibniz, *Dissertatio de arte combinatoria*, 1666, *Sämtliche Schriften und Briefe* (Berlin: *Akademie Verlag*, 1923), A VI 1, p. 163; *Philosophische Schriften* (edited by Carl Immanuel Gerhardt), Bd. IV, p. 30;

1943 – Neurone formel de Warren McCulloch et Walter Pitts

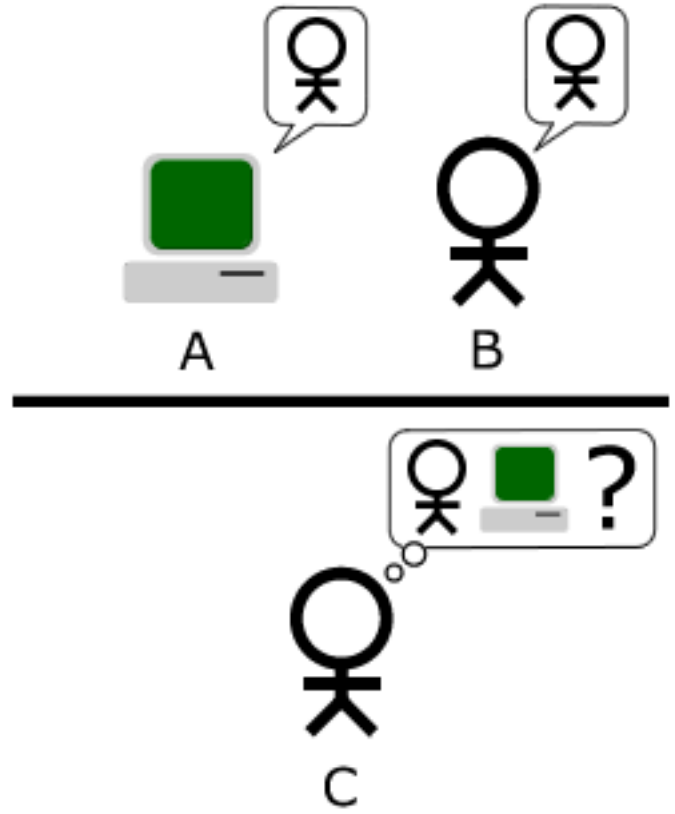
- Modélisation d'après des observations neurophysiologiques et anatomiques
- Somme pondérée binaire avec fonction d'activation à seuil



McCulloch, W. S., Pitts, W., *A Logical Calculus of the Ideas Immanent in Nervous Activity*, Bulletin of Mathematical Biophysics, vol. 5, pp. 115-133, 1943

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.
- Aujourd'hui : aucune machine n'a encore passé ce test



https://fr.wikipedia.org/wiki/Test_de_Turing

Alan Turing, « Computing machinery and intelligence », Mind, Oxford University Press, vol. 59, no 236, octobre 1950, p. 433-460

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

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

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

Dem.

$\vdash . *54{\cdot}26 . \supset \vdash :. \alpha = \iota`x . \beta = \iota`y . \supset : \alpha \cup \beta \in 2 . \equiv . x \neq y .$

$[*51{\cdot}231] \qquad \equiv . \iota`x \cap \iota`y = \Lambda .$

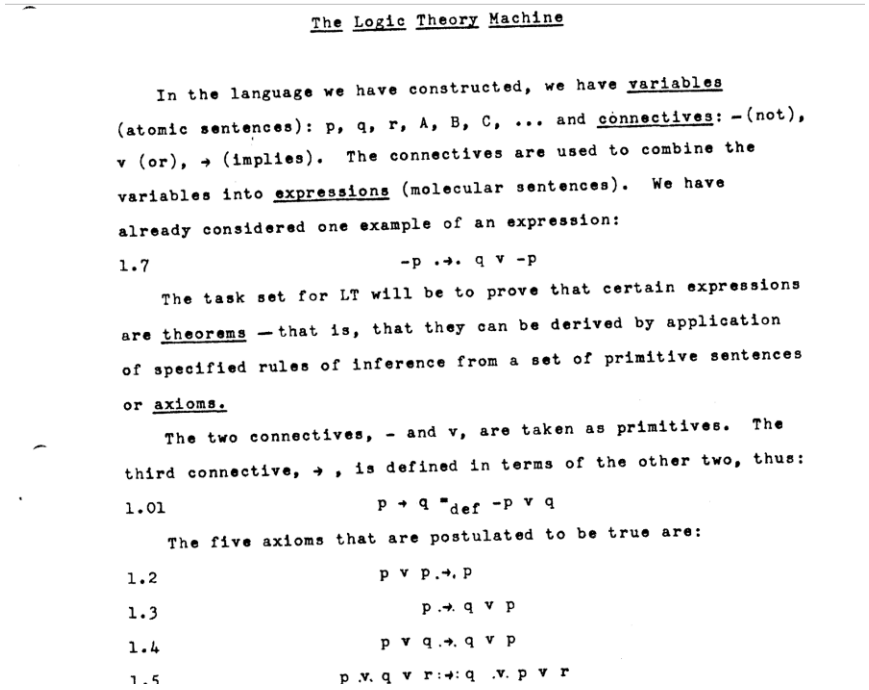
$[*13{\cdot}12] \qquad \equiv . \alpha \cap \beta = \Lambda \qquad (1)$

$\vdash . (1) . *11{\cdot}11{\cdot}35 . \supset$

$\vdash :. (\exists x, y) . \alpha = \iota`x . \beta = \iota`y . \supset : \alpha \cup \beta \in 2 . \equiv . \alpha \cap \beta = \Lambda \qquad (2)$

$\vdash . (2) . *11{\cdot}54 . *52{\cdot}1 . \supset \vdash .$ 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), → (implies). The connectives are used to combine the variables into expressions (molecular sentences). We have already considered one example of an expression:

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

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, → , is defined in terms of the other two, thus:

1.01 $\quad p \rightarrow q \ \overset{\text{def}}{=} \ -p \vee q$

The five axioms that are postulated to be true are:

1.2 $\quad p \vee p .\rightarrow. p$

1.3 $\quad p .\rightarrow. q \vee p$

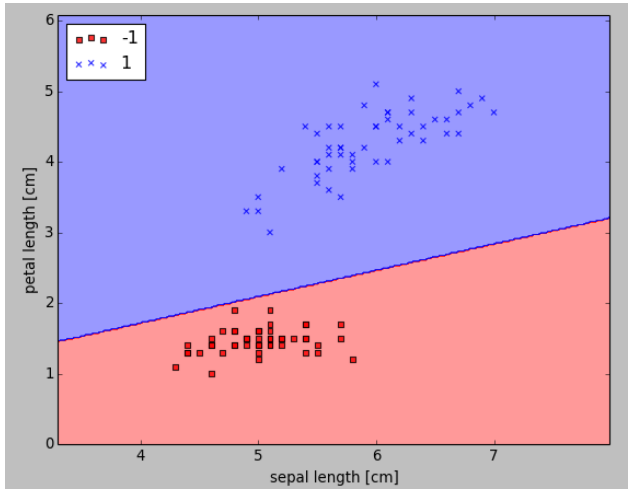
1.4 $\quad p \vee q .\rightarrow. q \vee p$

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

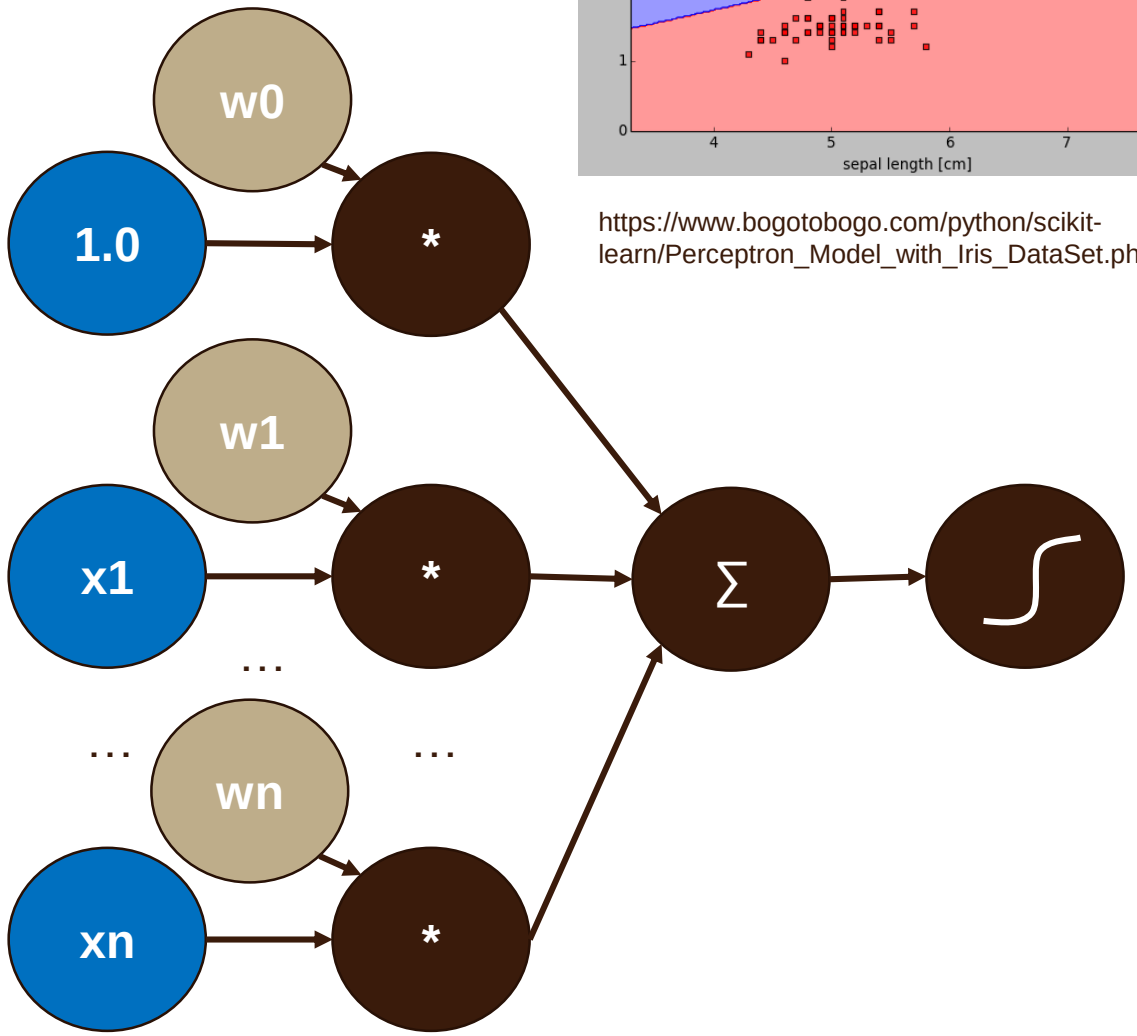
Newell, A., & Simon, H. (1956). The logic theory machine--A complex information processing system. *IRE Transactions on information theory*, *2*(3), 61-79.

1957 – Perceptron de Rosenblatt

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



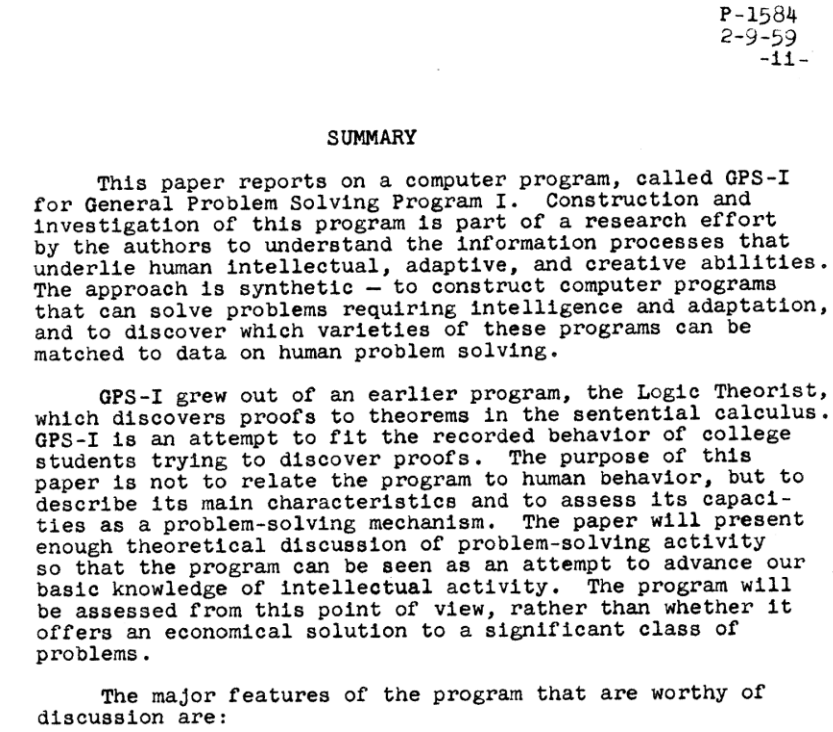
https://www.bogotobogo.com/python/scikit-learn/Perceptron_Model_with_Iris_DataSet.php



Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review, 65(6), 386.

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
-ii-

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:

Newell, A.; Shaw, J.C.; Simon, H.A., 1959. Report on a general problem-solving program. Proceedings of the International Conference on Information Processing. p. 256–264.

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

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

$(p \wedge q) \vee \neg p$ => p = faux

$(p \wedge \neg p)$ => pas de solution

1962 Algorithme pour résoudre le problème SAT

La base des solveurs modernes pour :

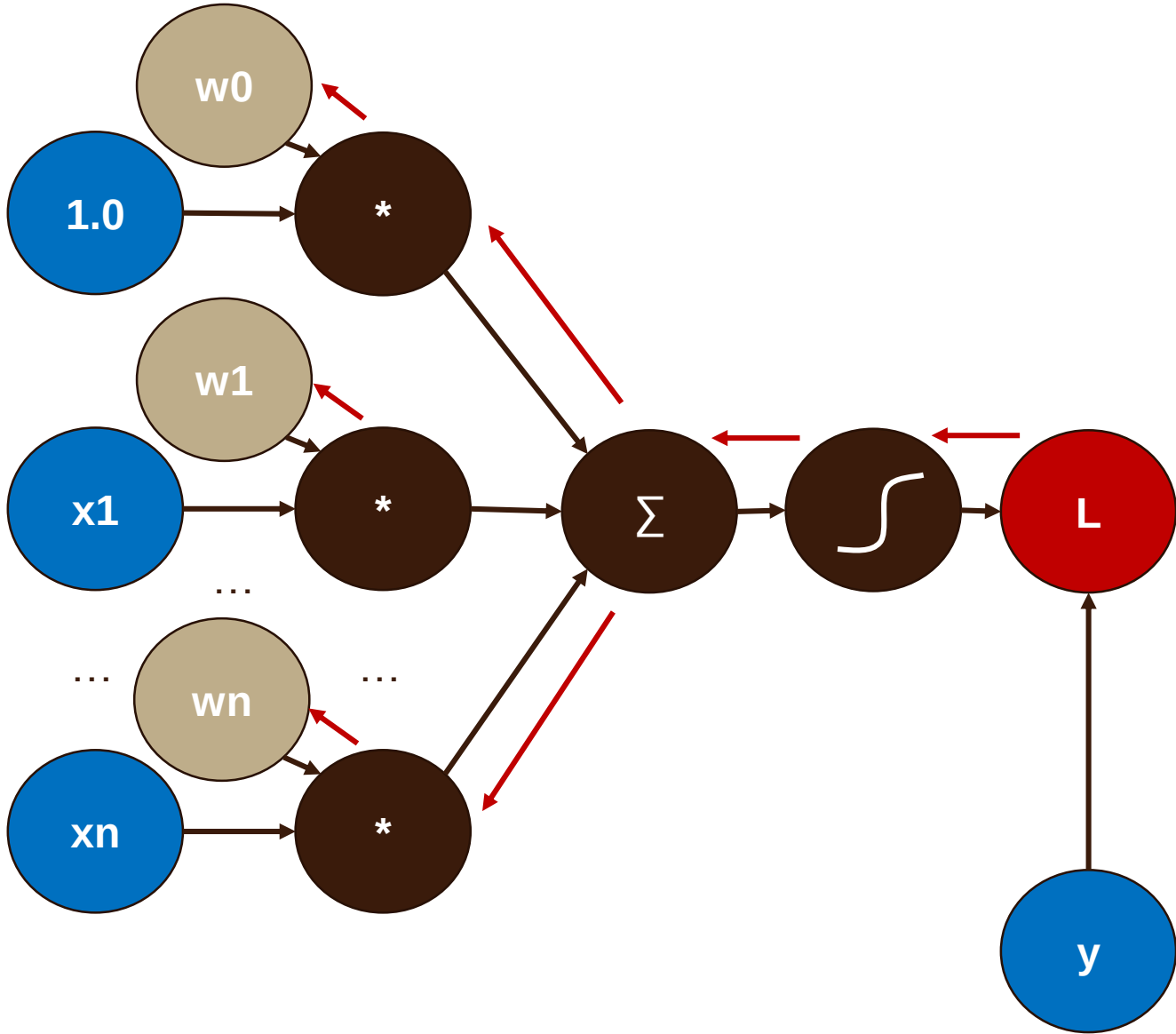
- Le diagnostique
- La planification
- La vérification de propriété de modèles

Martin Davis et Hillary Putnam, « A Computing Procedure for Quantification Theory », Communications of the ACM, 7, p. 201-215 (1960)

Martin Davis, George Logemann, and Donald Loveland. 1962. A machine program for theorem-proving. Commun. ACM 5, 7 (July 1962), 394–397.

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

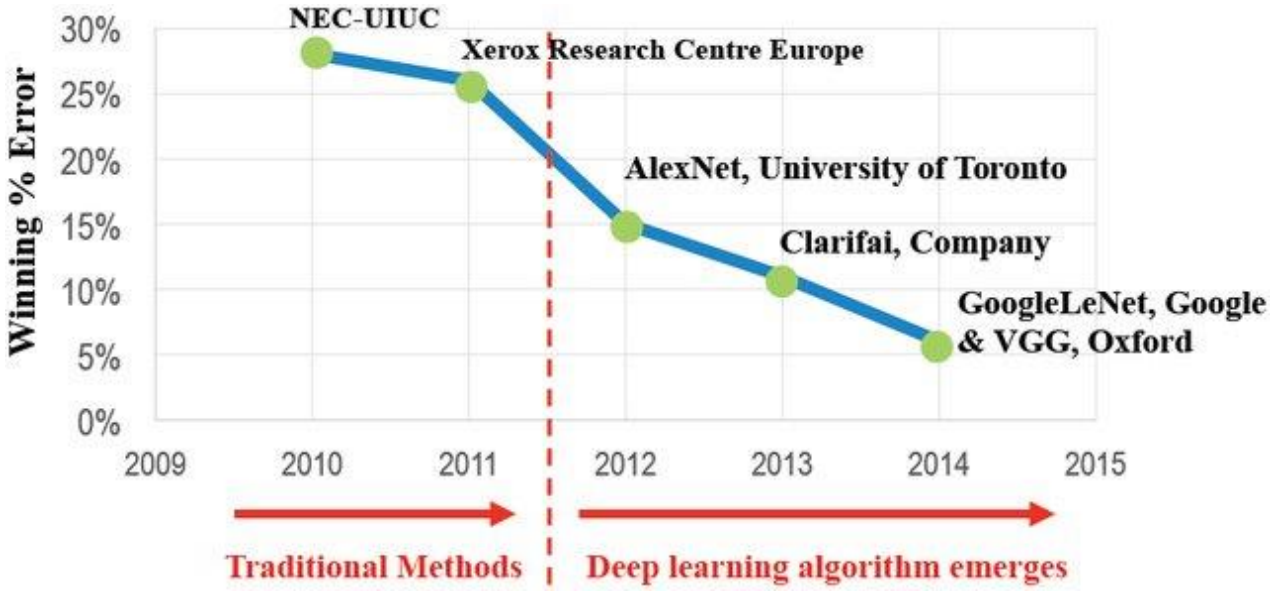
2005 – Première conduite autonome sur route inconnue

- **Règles**
  - **Conduite autonome < 10h**
  - **Technologie civile**
  - **Pas de contact avec les autres véhicules**
- **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



https://www.researchgate.net/publication/329975404_FPGA-Based_Accelerators_of_Deep_Learning_Networks_for_Learning_and_Classification_A_Review

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

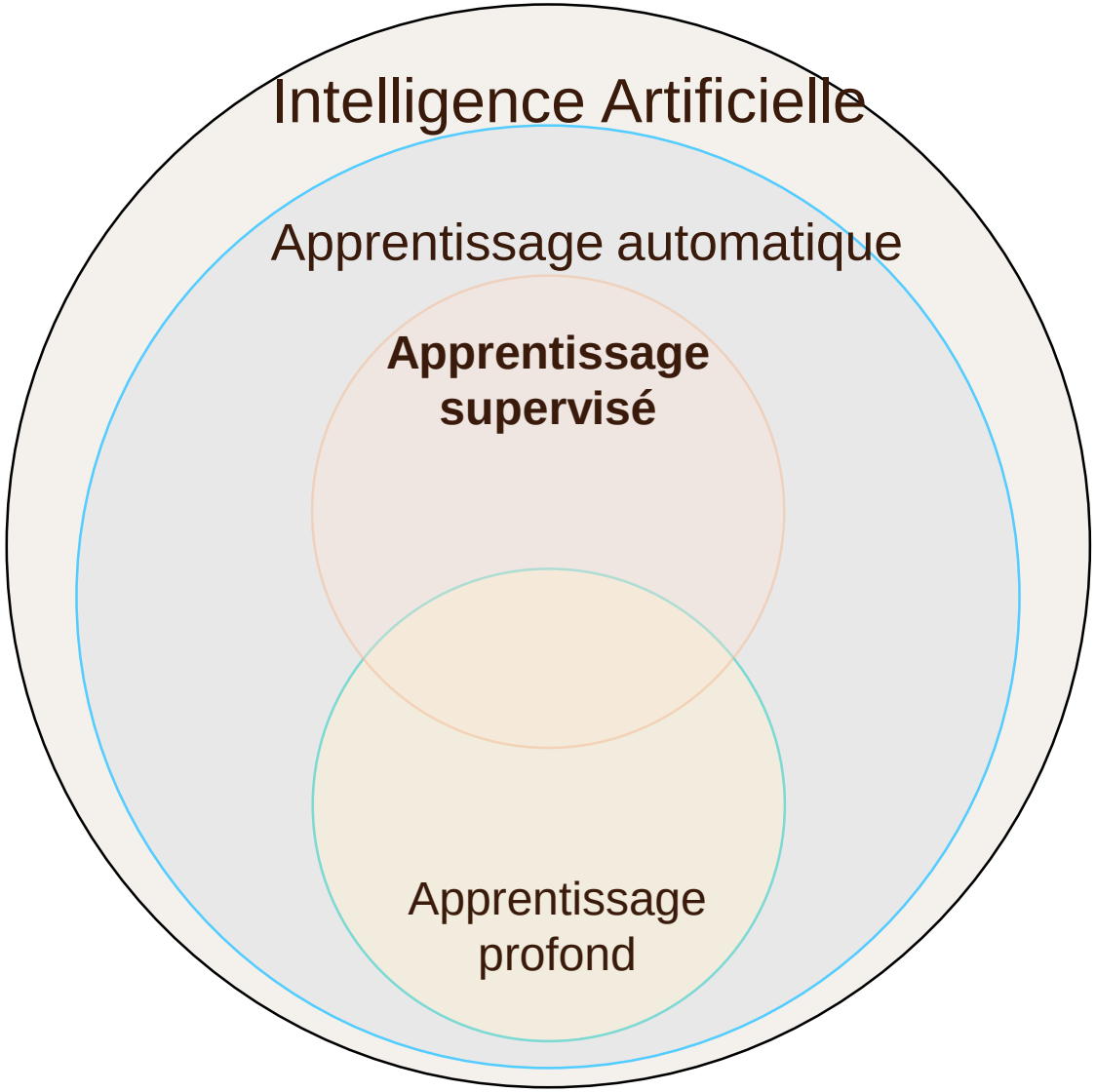
- Voiture intelligente
- Jeux Vidéos
- Œuvres artistiques
- Enseignement

https://www.nvidia.com/en-us/shield/sliders/ai-upscaling/

Généralités sur l’apprentissage supervisé

On se focalise sur l’aspect apprentissage supervisé et on ignore les autre aspects de l’intelligence artificielle

Positionnement de l’apprentissage supervisé



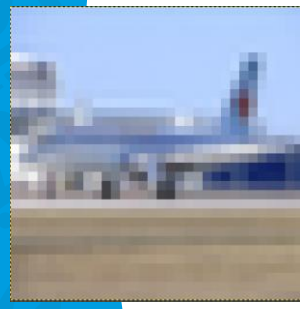
Un modèle de classification prédit une classe

Code écrit par un humain

Paramètres appris grâce aux données

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

X f_θ $\hat{y}$ $\hat{y}$



Donnée d'entrée

Classification

Modèle

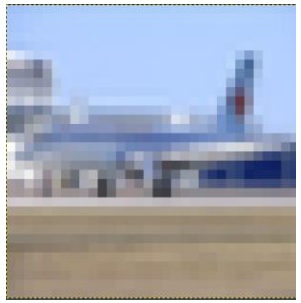
0

Résultat
numéro de la classe compréhensible par la machine

avion

Résultat
symbole compréhensible par un humain

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



Donnée d’entrée

Extraction des caractéristiques

Classification

Modèle

0

Résultat
(numéro de la classe)

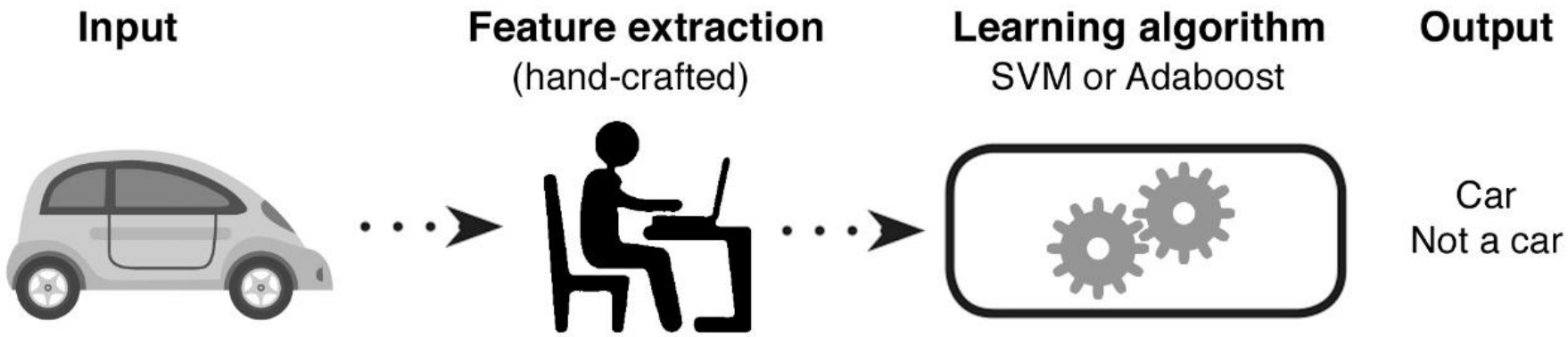
On maîtrise cette partie

On NE maîtrise PAS TOUJOURS cette partie

Généralités sur l’extraction de caractéristiques

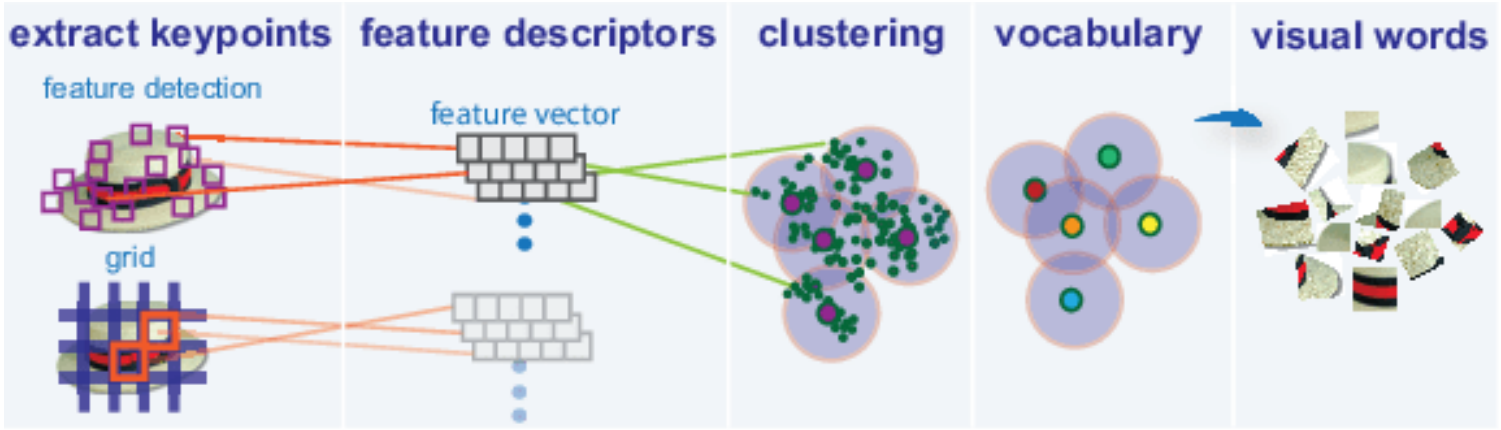
- Les données brutes ne peuvent pas être utilisées directement
- Il faut extraire des informations
- De nombreuses méthodes générales existent

Il est necessaire d’être expert des données pour extraire des caractéristiques pertinentes

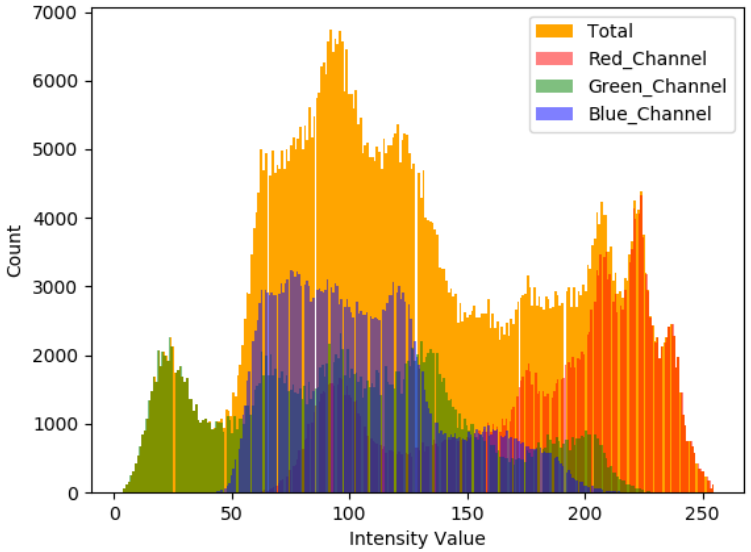


https://freecontent.manning.com/the-computer-vision-pipeline-part-4-feature-extraction/

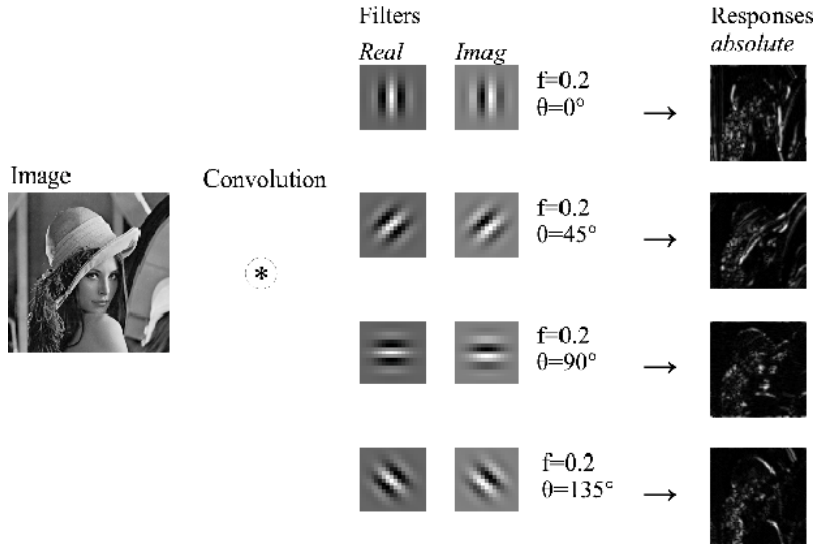
Extraction de caractéristique - Image



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



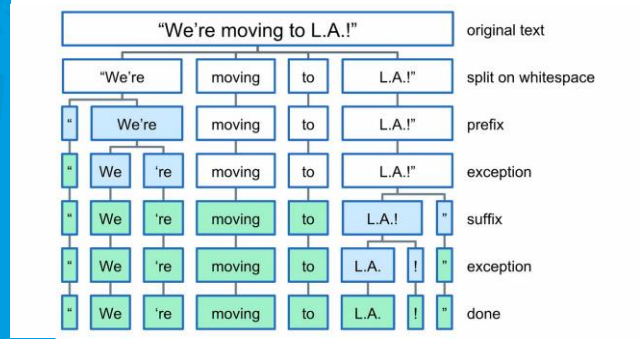
https://miro.medium.com/v2/resize:fit:640/1*Vd5OY8LRaybkFj2NjBbpbA.png

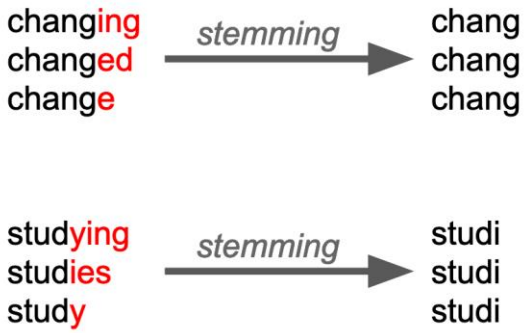


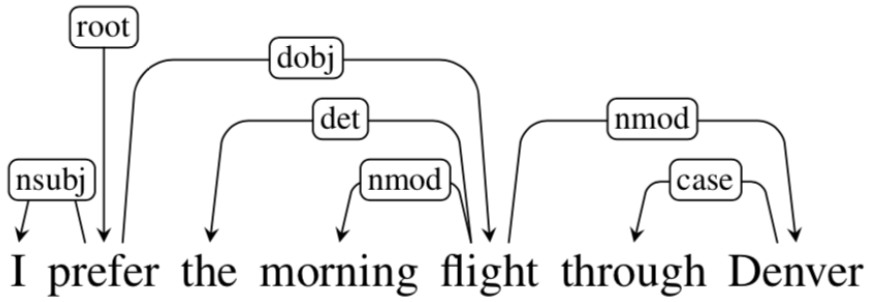
https://www.semanticscholar.org/paper/Gabor-features-in-image-analysis-K%C3%A4m%C3%A4r%C3%A4inen/b0bc670f8ffb469eec5253cf9d634e90fd6eda88

Extraction de caractéristique - Texte

Découpage → Recherche de radical → Dépendances





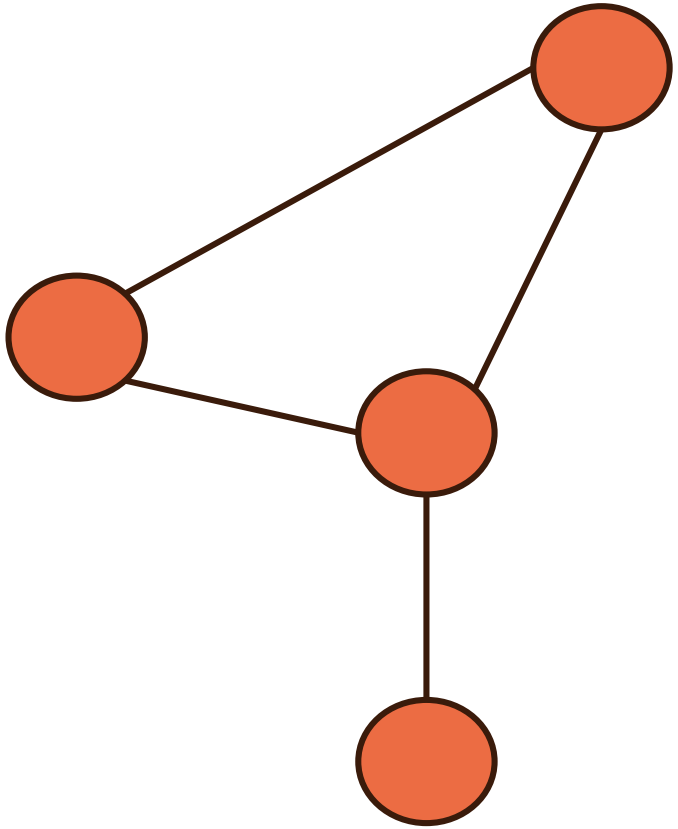


Extraction de caractéristique – Données tabulaires

- 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 necessaire d’être expert des données pour extraire des caractéristiques pertinentes

Donnée d’entrée

Extraction des caractéristiques

Classification

Modèle profond

0

Résultat (numéro de la classe)

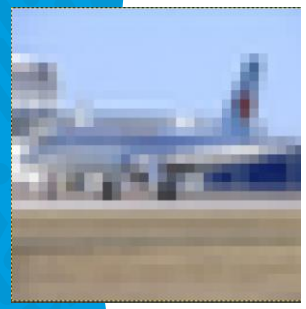
On NE maîtrise PLUS cette partie, elle est APPRISE

On NE maîtrise toujours PAS cette partie

Evaluation des modèles d’apprentissage supervisé

Il est nécessaire d’évaluer les modèles pour connaitre leurs performances

Pour rappel, voici comment fonctionne un classifieur



Donnée d'entrée

Classification

Modèle

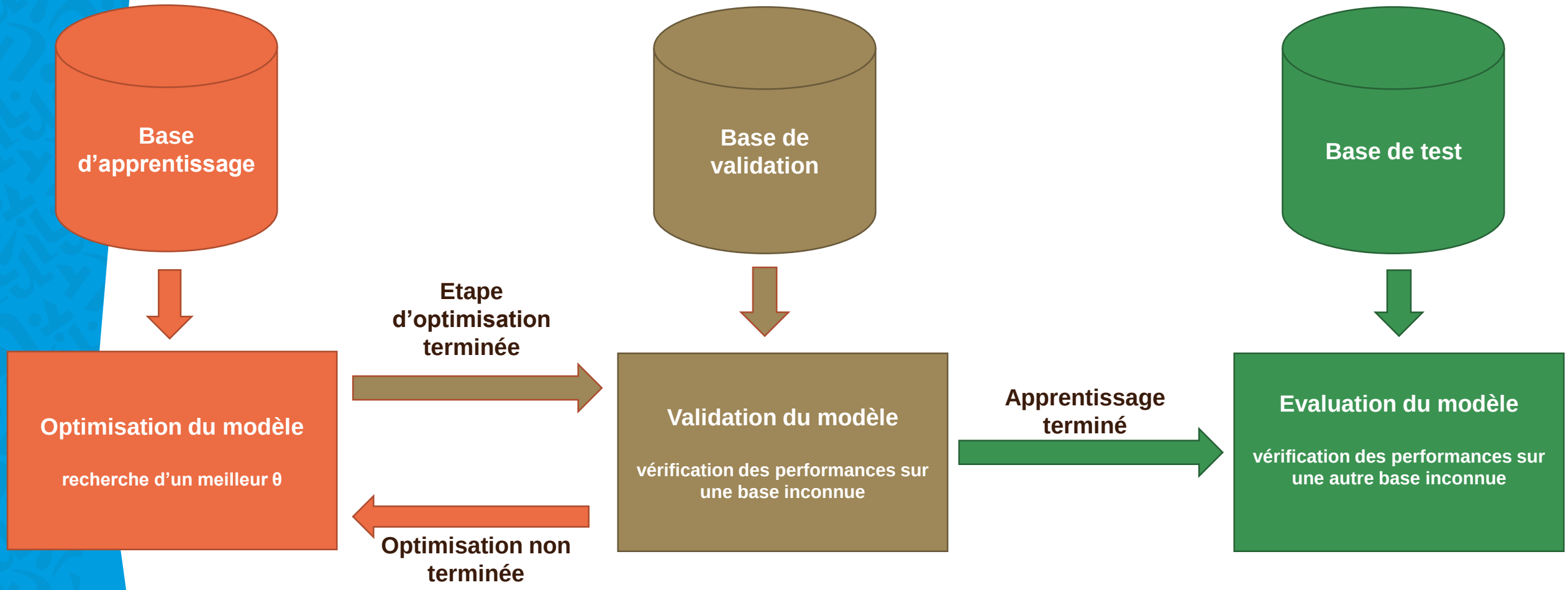
0

Résultat
(numéro de la classe)

avion

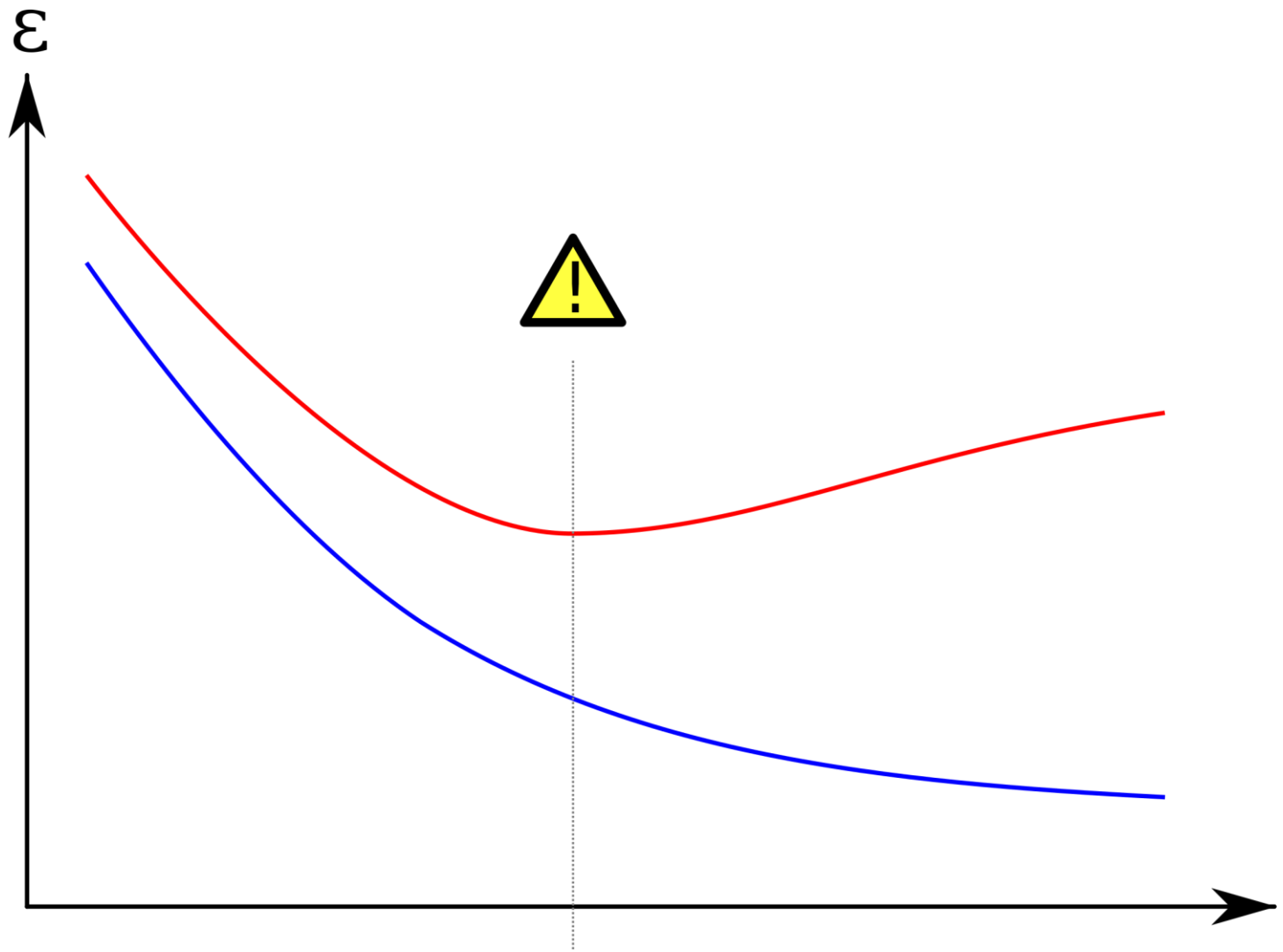
Résultat
(nom de la classe)

Aparté : apprentissage d'un modèle



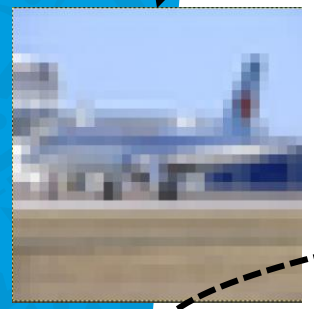
Les techniques dépendent bien entendu du type de modèle

Pourquoi autant de bases ?

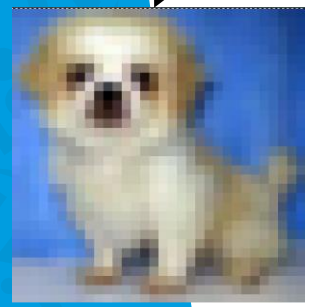


https://en.wikipedia.org/wiki/Overfitting

Evaluation d'un modèle de classification



...



Base d'évaluation étiquetée

Classification

Modèle

0

1

...

5

Résultats

0

0

...

5

Vérité terrain

Evaluation

Rapport d'évaluation

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

Vérité terrain \ 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

Vérité terrain \ 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

Chaque colonne représente les exemples classé dans une classe précise

Prédiction

Vérité terrain

	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

Vérité terrain	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

Vérité terrain

	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

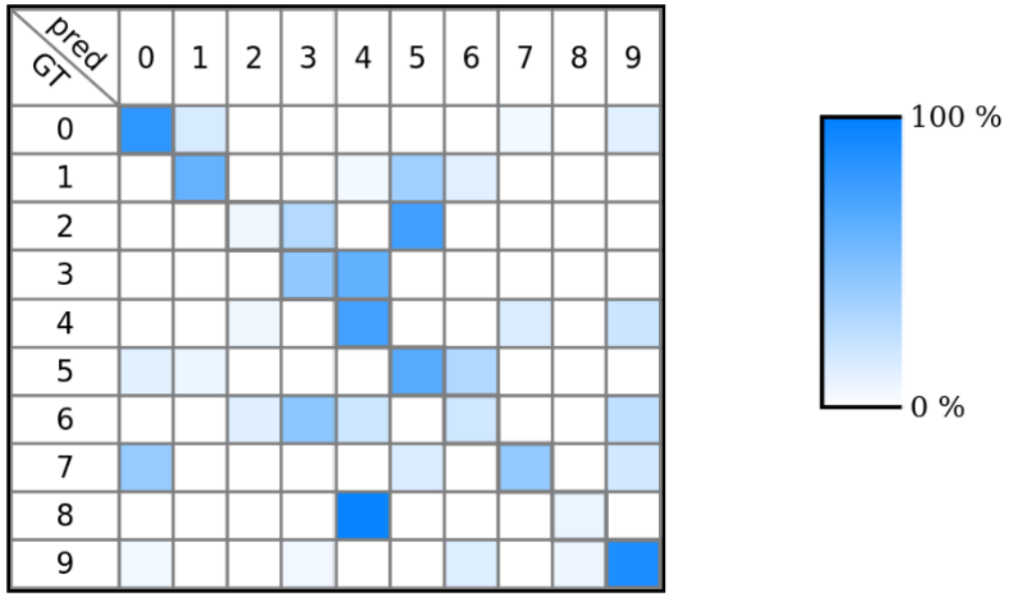
La matrice de confusion du système parfait est donc diagonale

Prédiction

Vérité terrain

	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



L. Pommé, R. Bourqui, R. Giot, and D. Auber, "Relative Confusion Matrix: Efficient Comparison of Decision Models," Proceedings of the 26th International Conference Information Visualisation (IV2022), 2022

Taux de reconnaissance

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

Prédiction

Vérité terrain	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

Précision(classe)

$$\frac{Vrais\ positifs\ de\ i}{Vrais\ positifs\ de\ i + Faux\ positifs\ de\ i}$$

Prédiction

Vérité terrain	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

Rappel(classe)

$$\frac{Vrais\ positifs\ de\ i}{Vrais\ positifs\ de\ i + Faux\ négatifs\ de\ i}$$

Prédiction

Vérité terrain	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(classe)

$$\frac{Vrais\ négatifs\ de\ i}{Vrais\ négatifs\ de\ i + Faux\ positifs\ de\ i}$$

Prédiction

Vérité terrain	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

Généralités sur l’évaluation

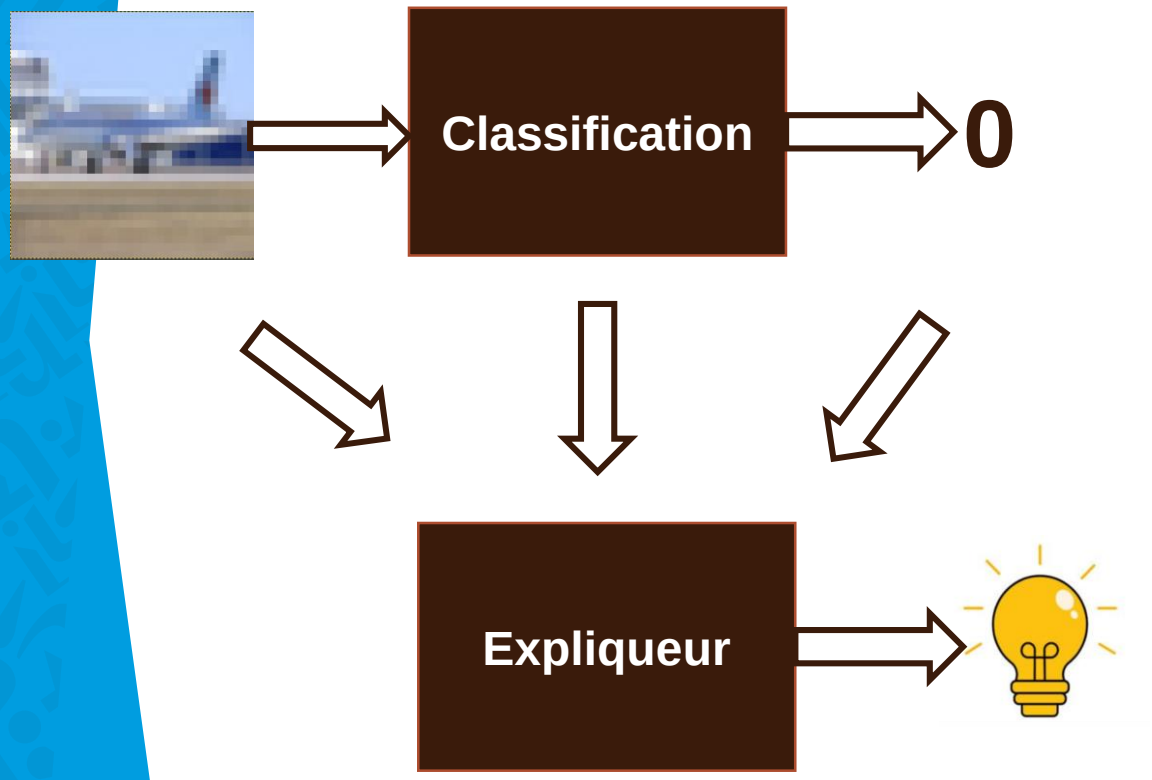
- **Prérequis**
  - **Modèle**
  - **Base de test de qualité**
- **Obtient**
  - **Métriques à interpréter et comparer**
  - **Une idée des performances**
  - **Possibilité de comparer des modèles relativement les uns aux autres**
- **Il manque**
  - **Raisons qui expliquent les réussites et les erreurs**

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 ?



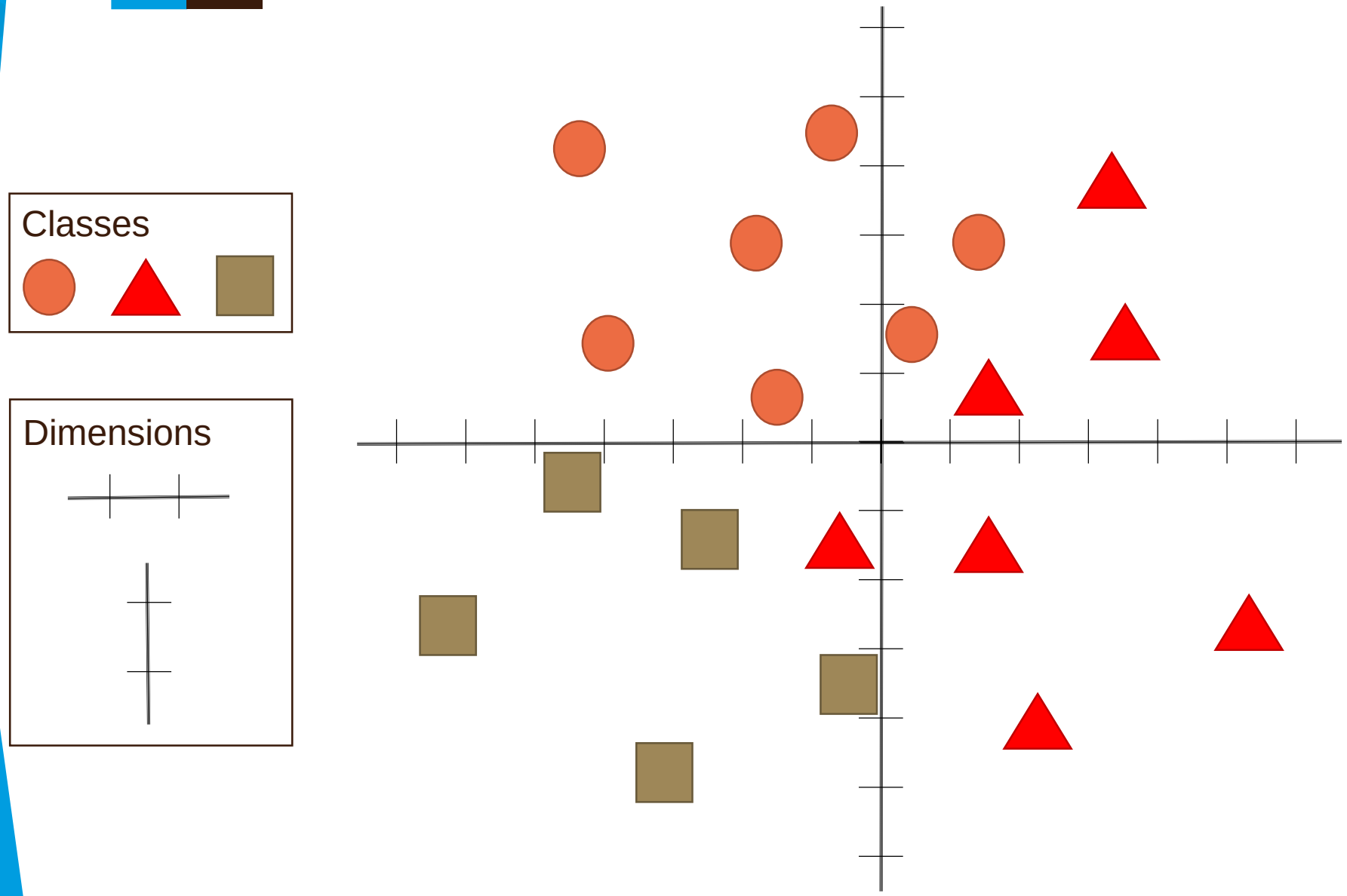
Méthodes après-coup

- 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

Quelques exemples de modèles simples de classification supervisée

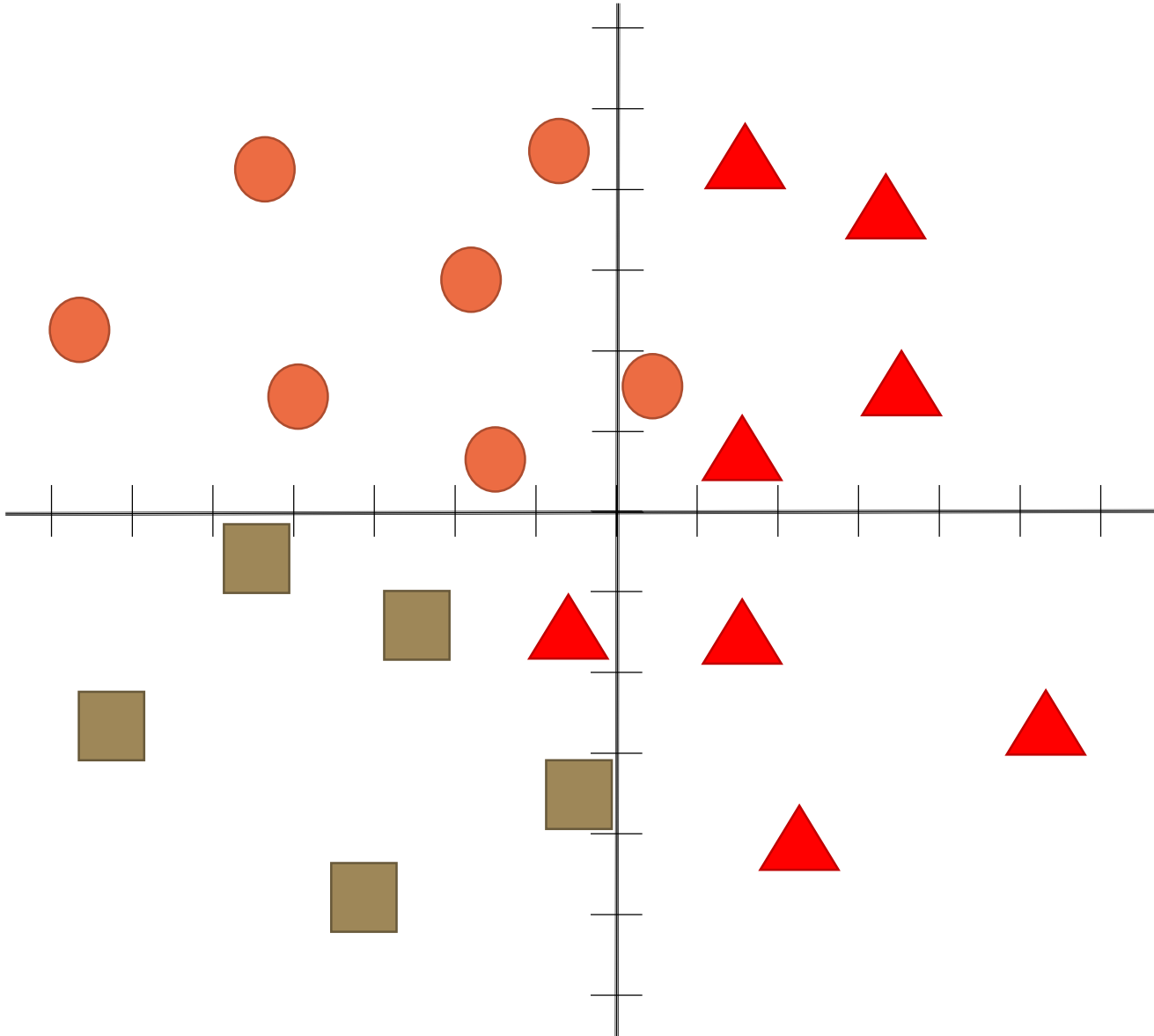
Description de quelques modèles avant de présenter les explications intrinsèques et après-coup

Classification de données en 2 dimensions

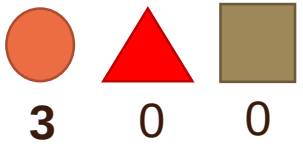
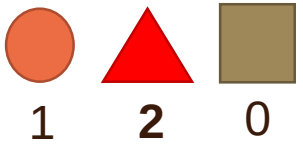
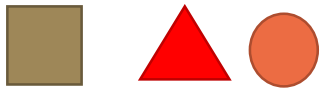
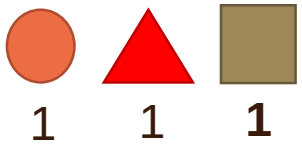


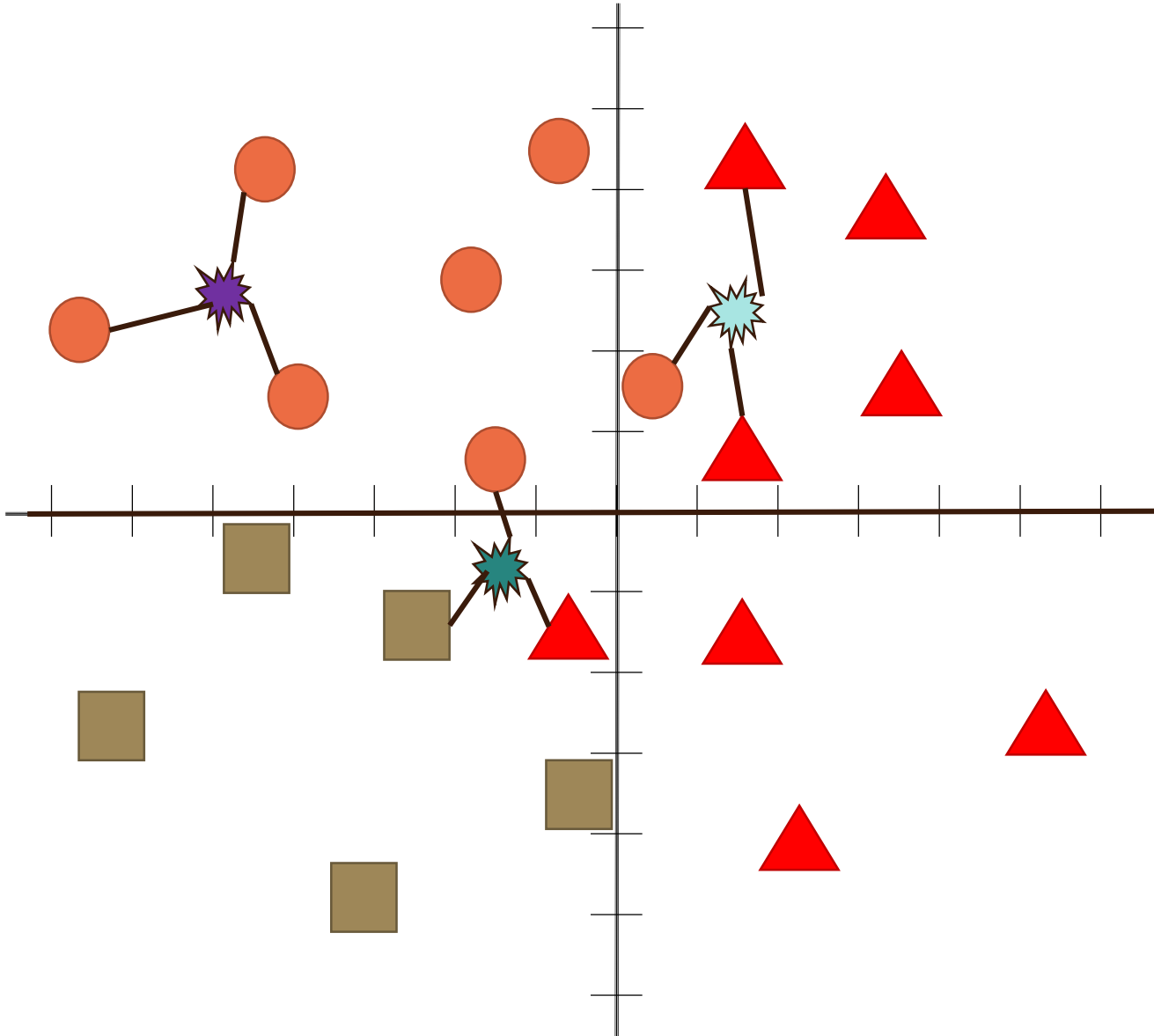
K plus proches voisins : apprentissage

Paramètres : ensemble des données étiquetées
Hyperparamètres : K
F : code qui recherche les exemples les plus proches
Optimiseur : aucun



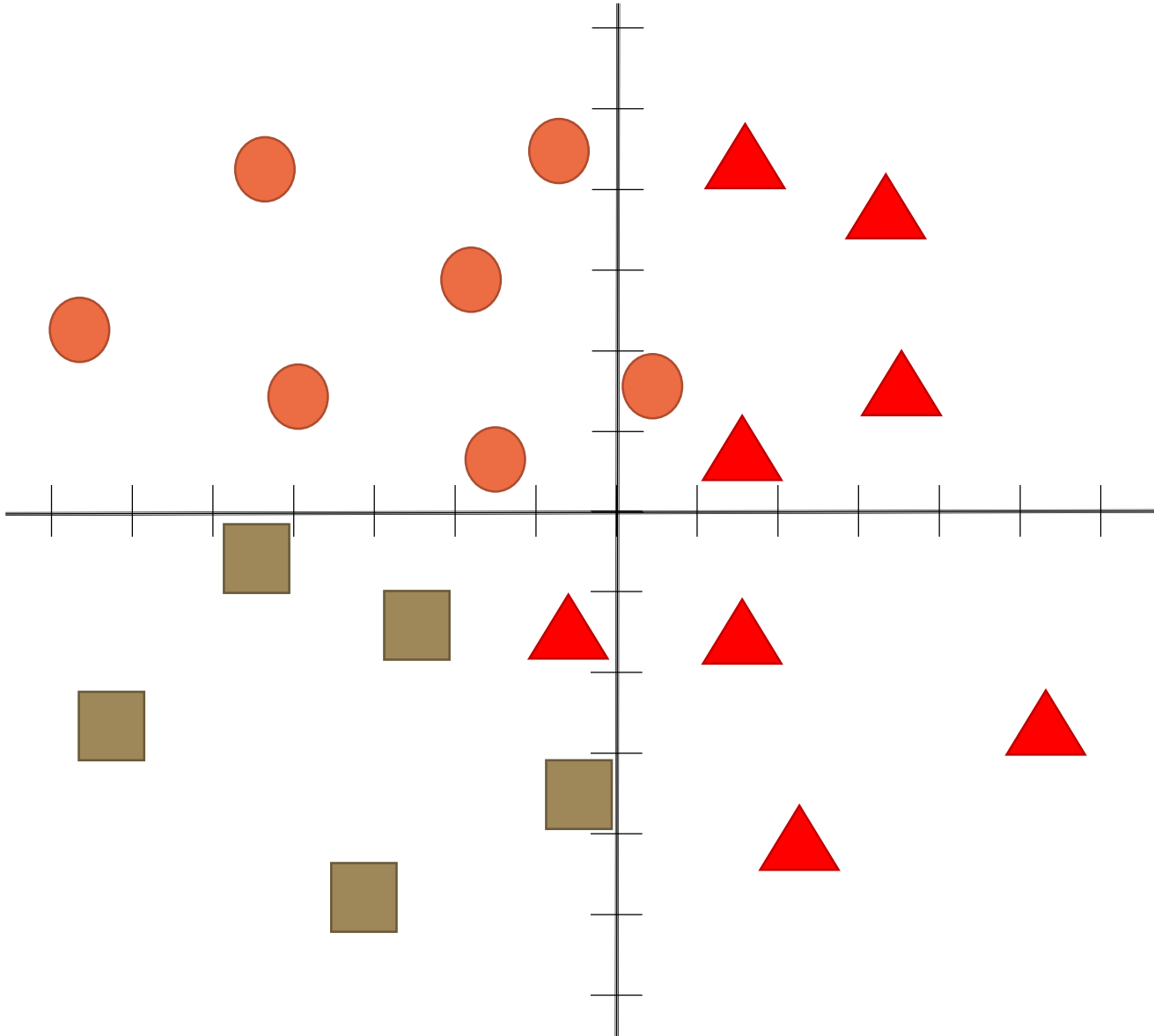
K plus proches voisins : inférence

Requête	K voisins	Comptage	Prédiction
		 **3** 0 0	
		 1 **2** 0	
		 1 1 **1**	

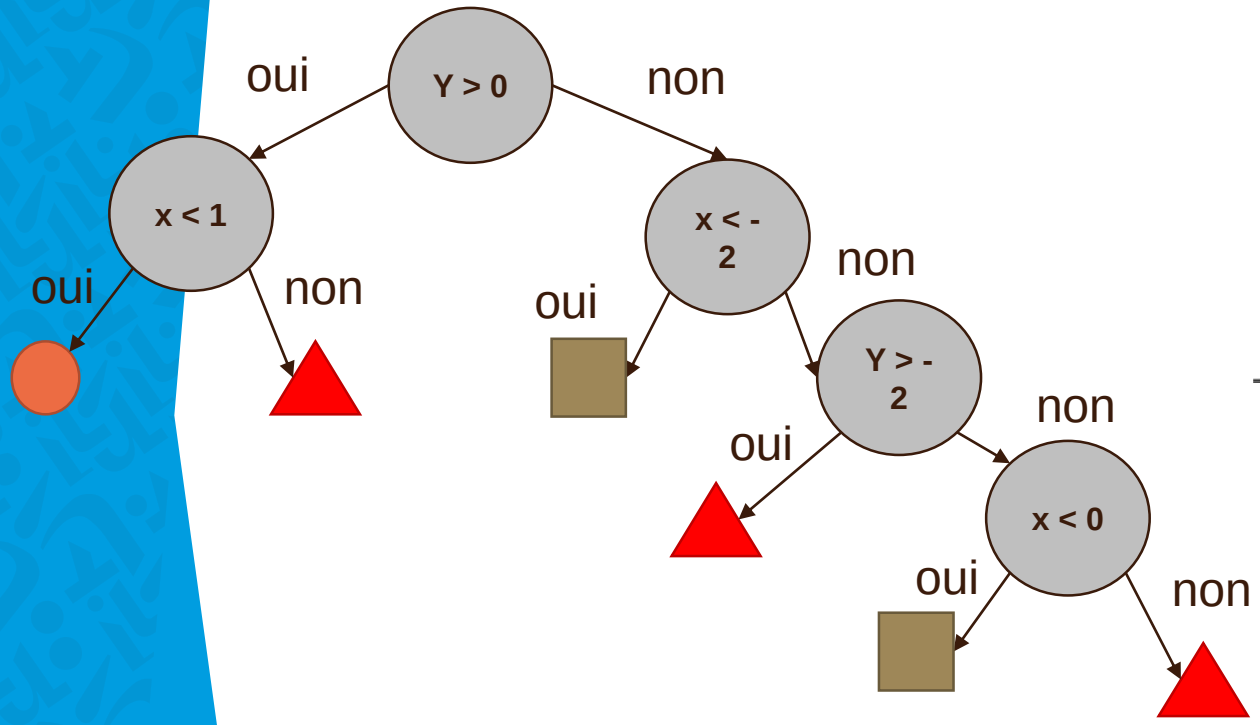


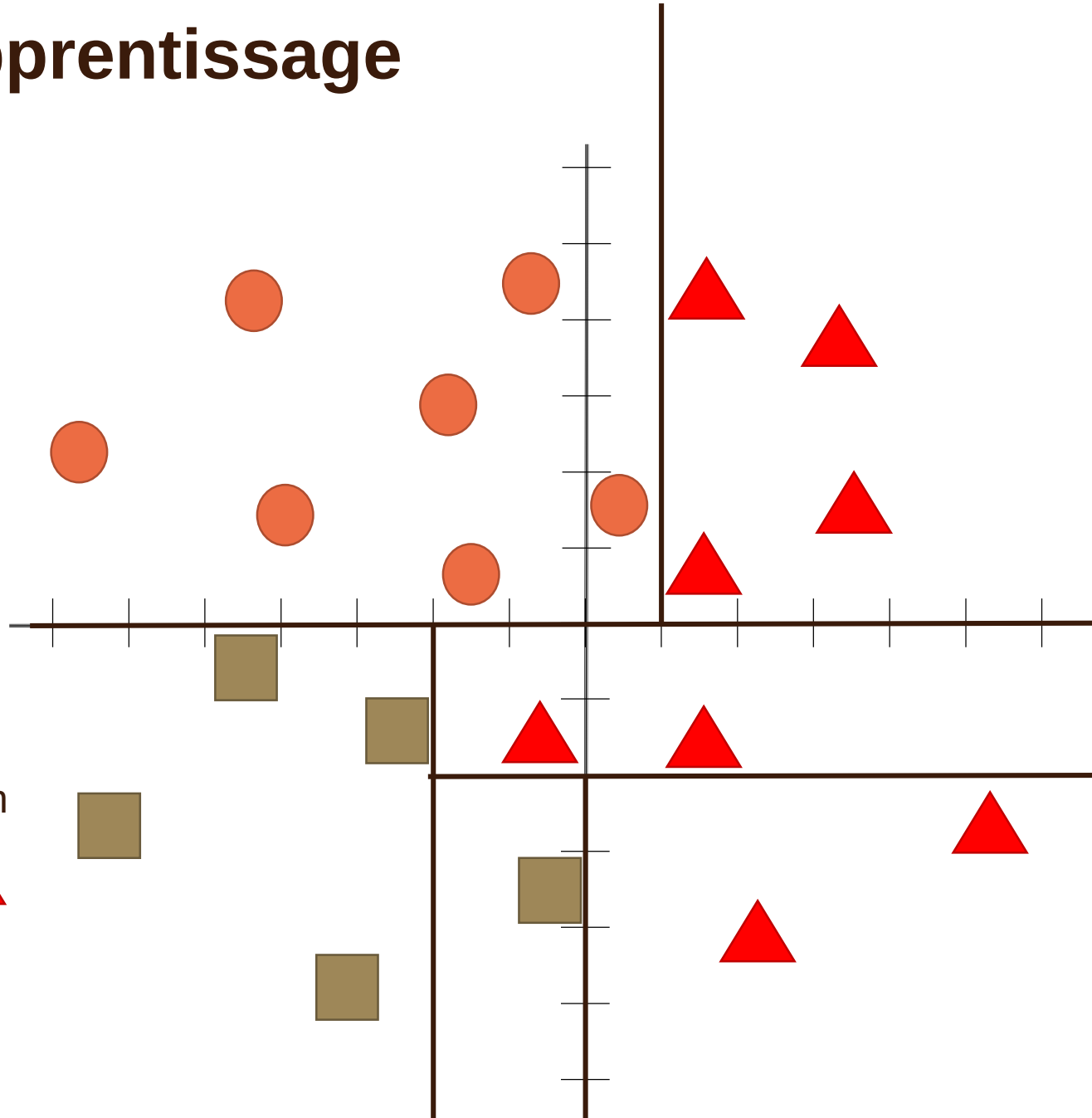
Arbre de décision : apprentissage

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

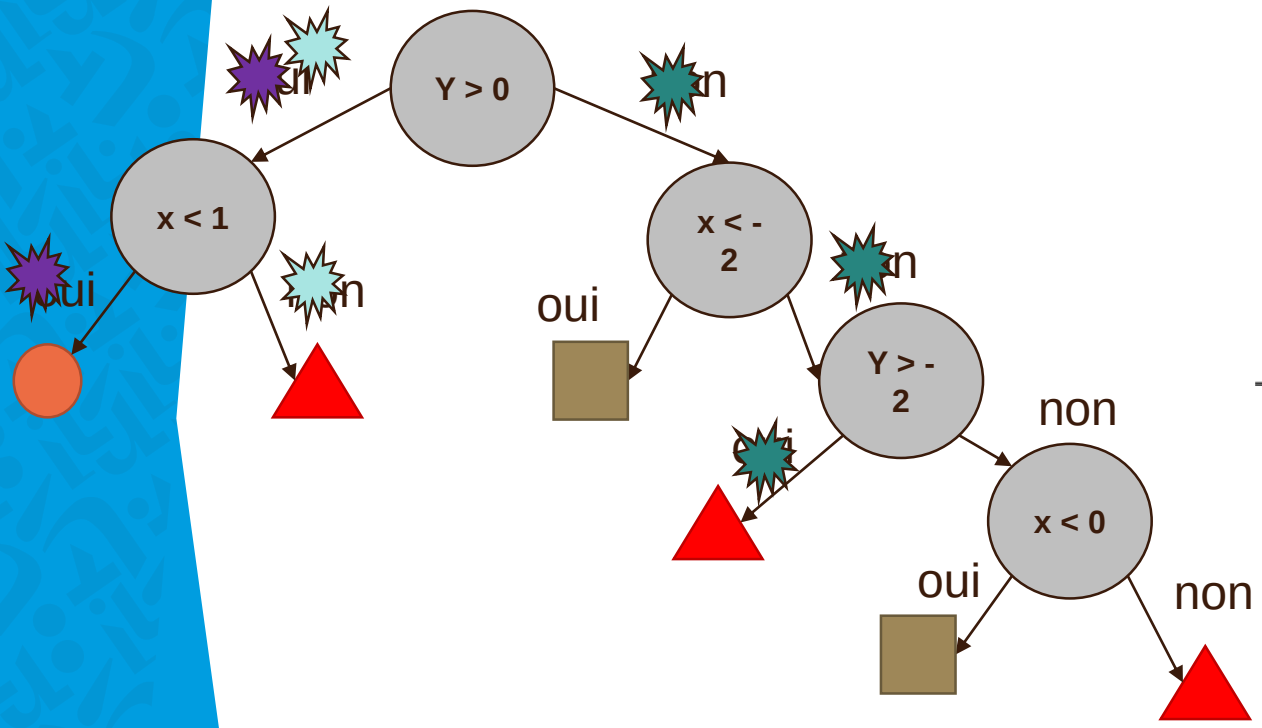


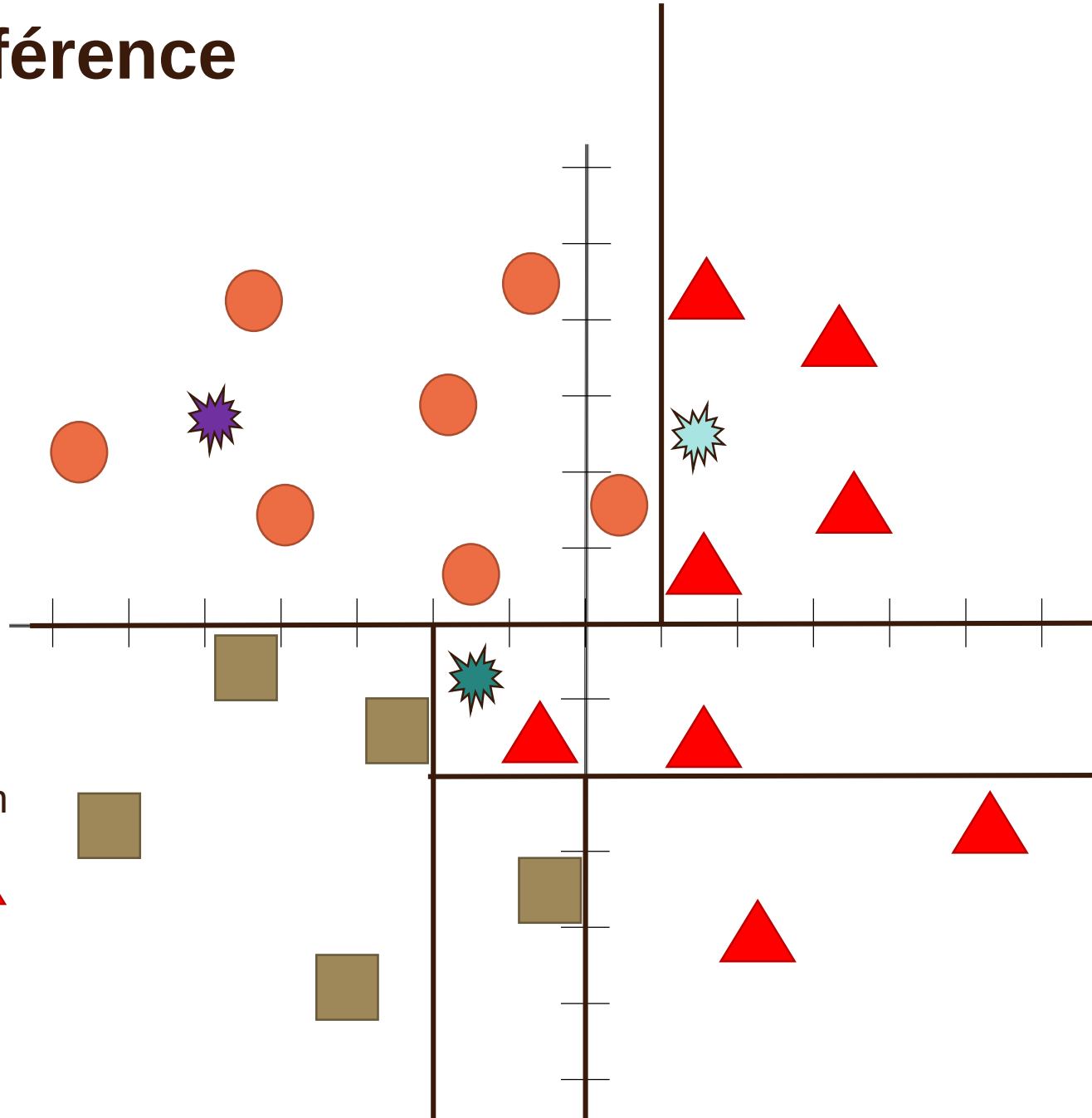
Arbre de décision : apprentissage





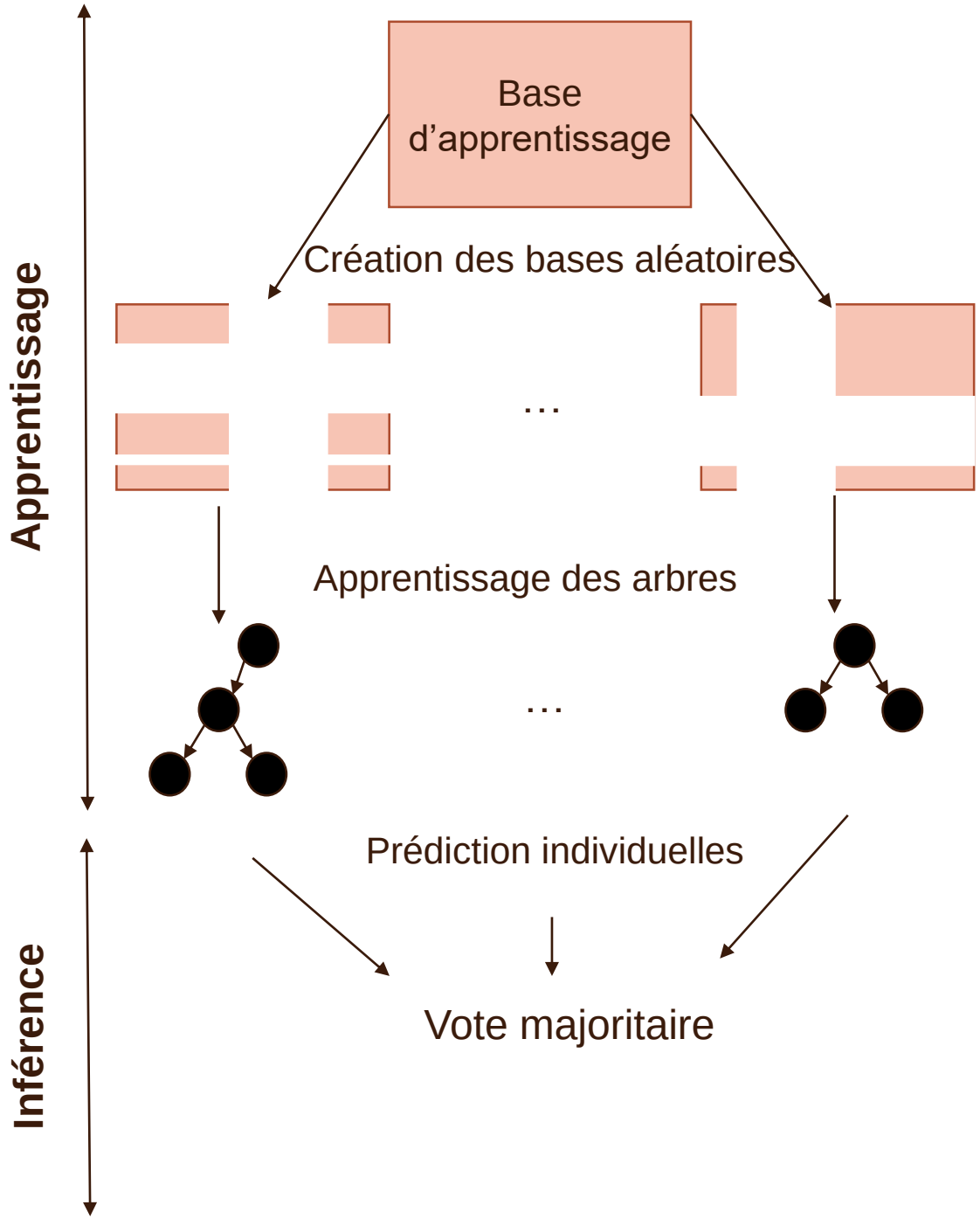
Arbre de décision : inférence





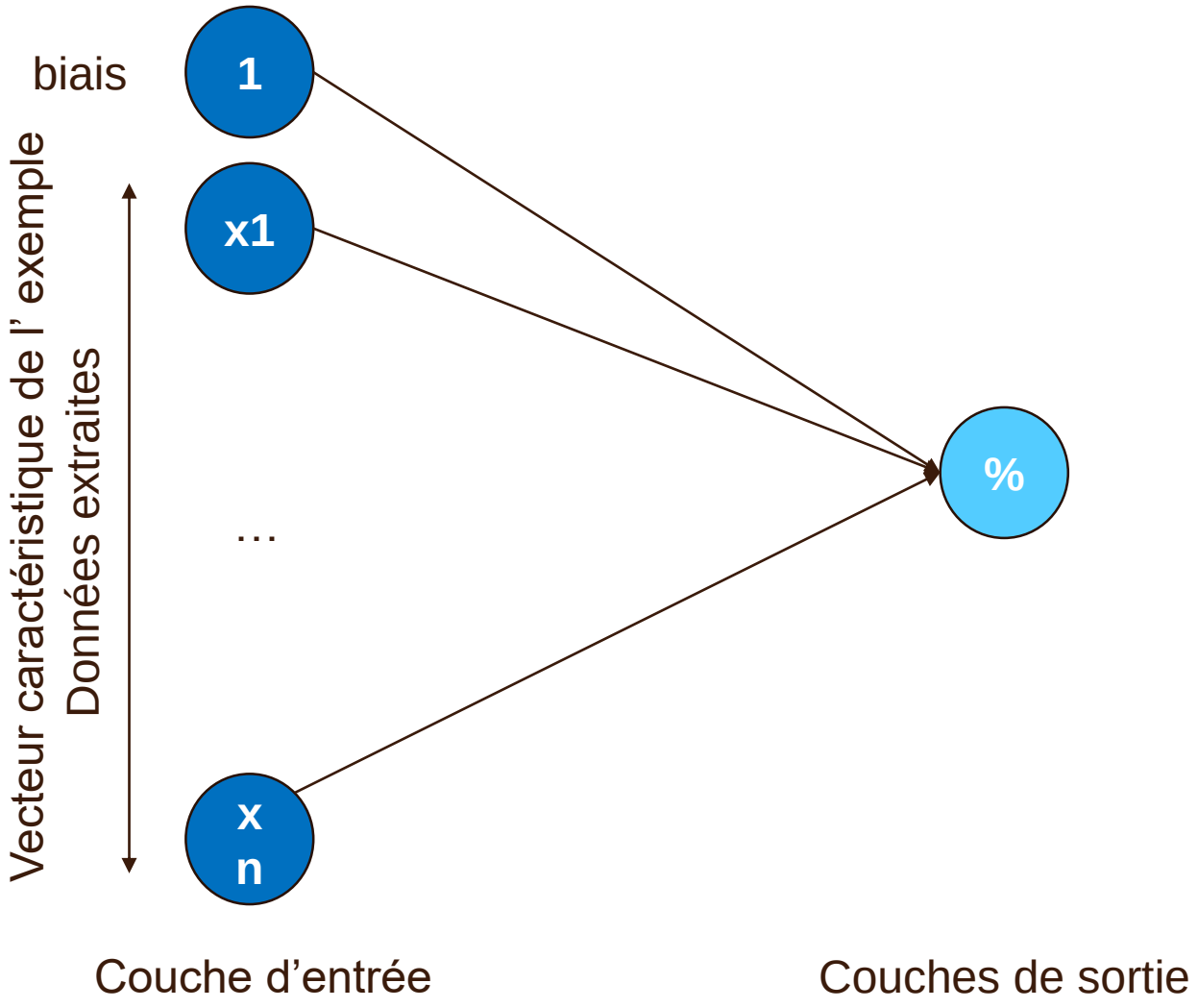
Forêt aléatoire

- **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**



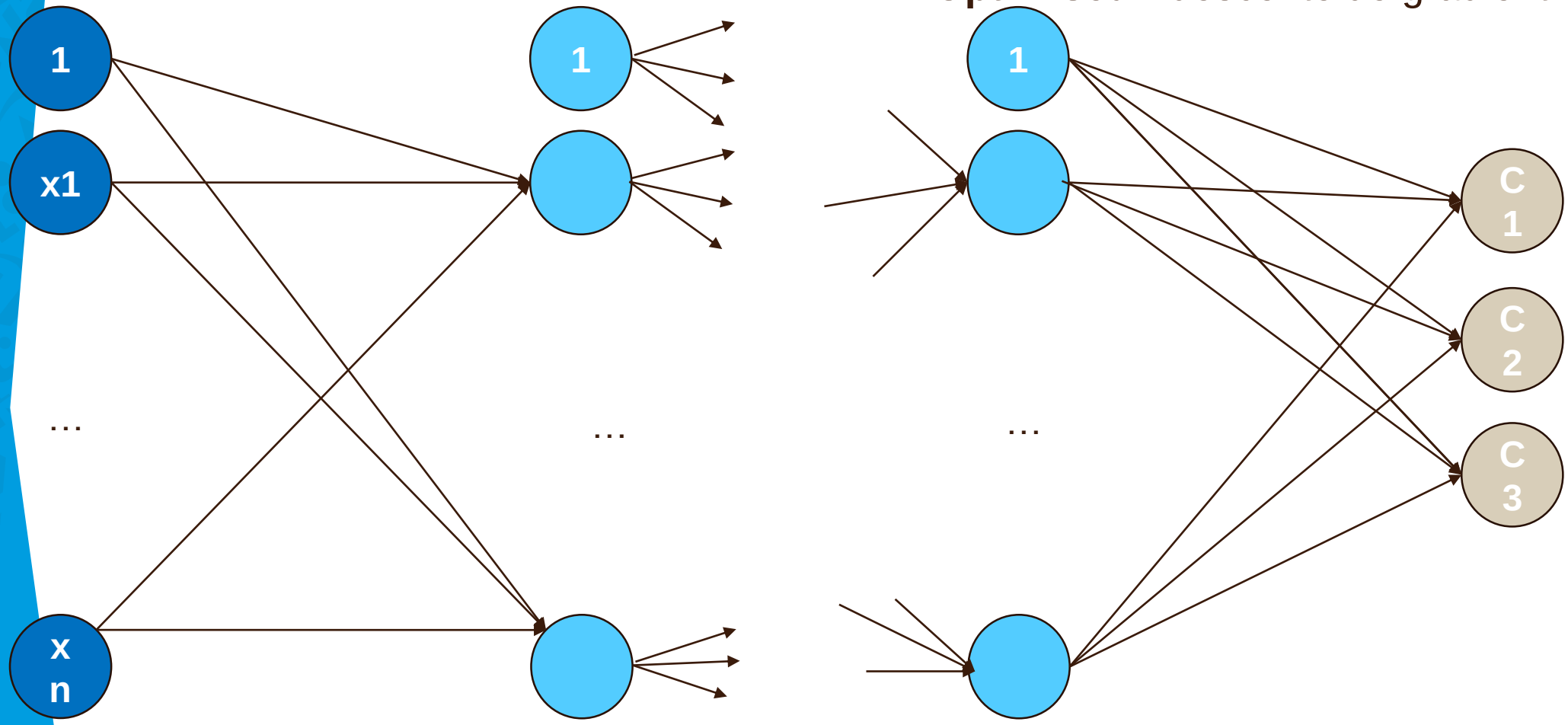
Régression logistique

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

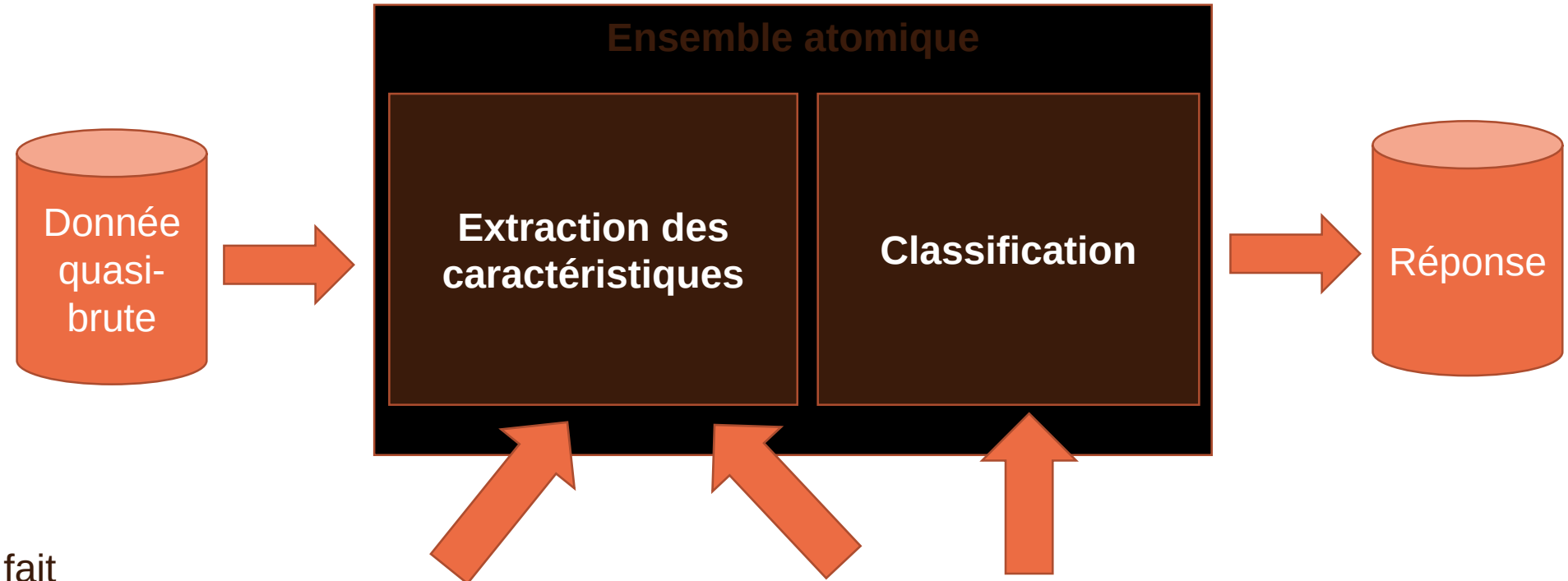


Réseau de neurones multi couches

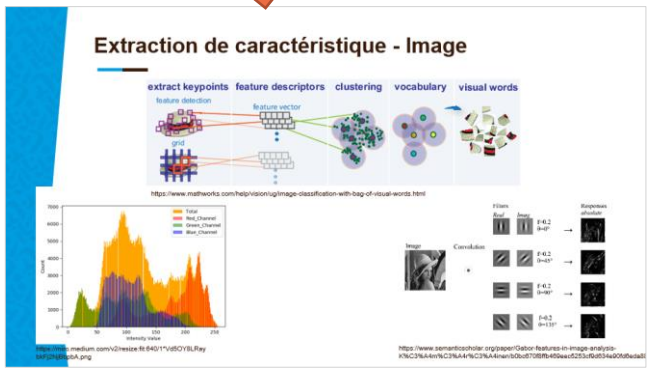
Paramètres : ensemble des poids (=arêtes)
Hyperparamètres : profondeur et largeur du rése
F: code qui exécute les opérations mathématique
du réseau
Optimiseur: descente de gradient

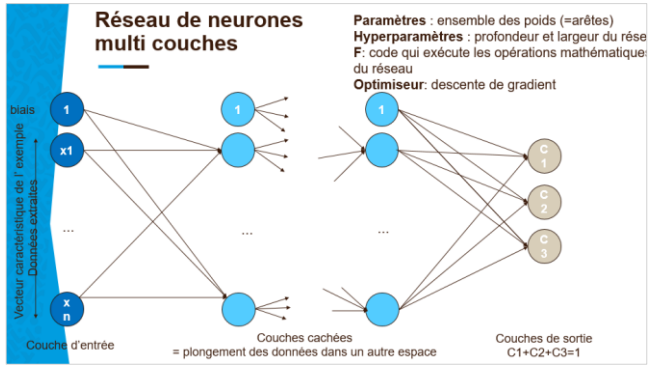


Réseau de neurones profond



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



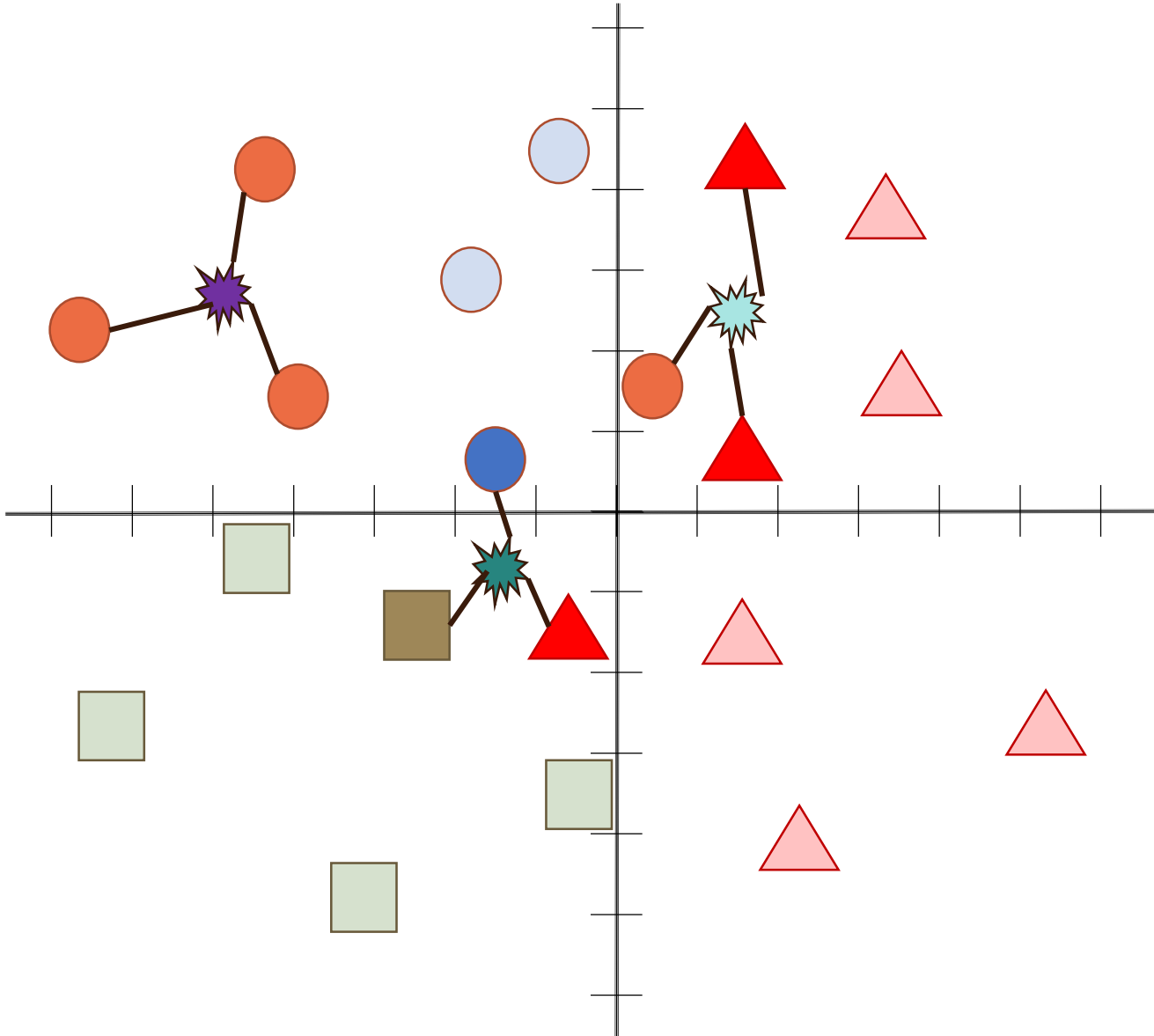


La classification est faite grâce à un MLP.
La plupart des poids sont ici

Explications intrinsèques

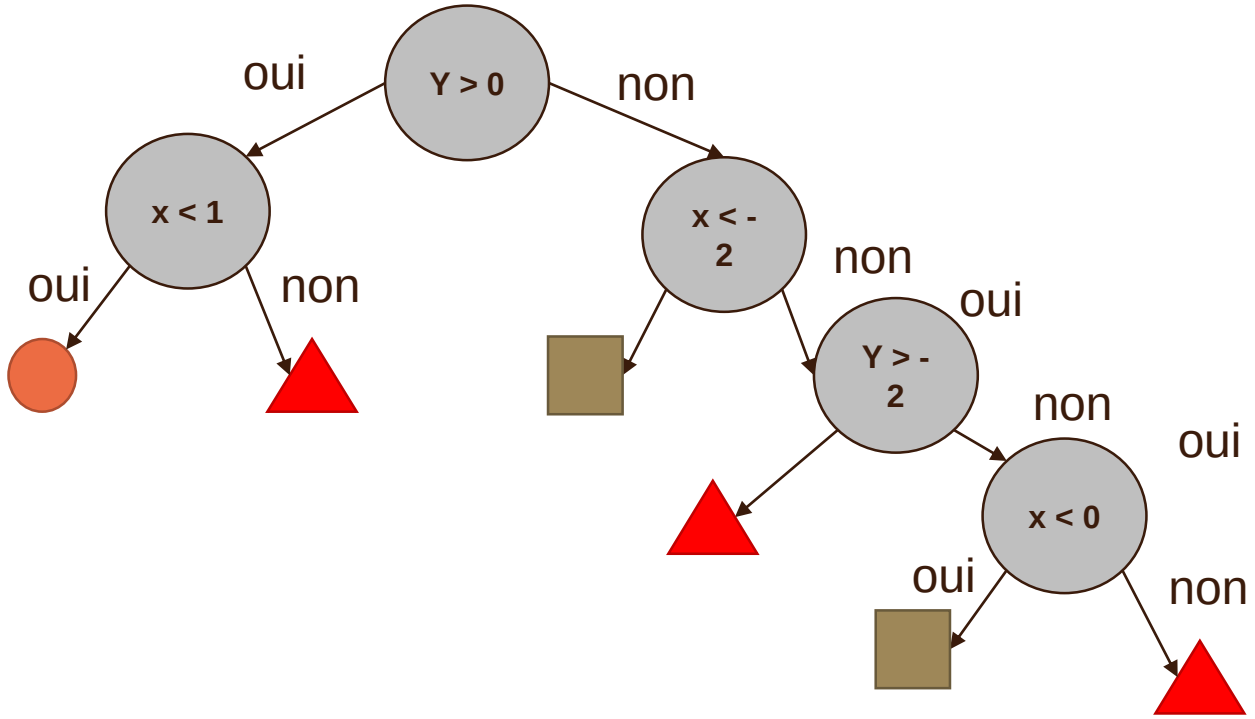
Les modèles proposent directement des explications

K plus proches voisins : représentation des plus proches voisins

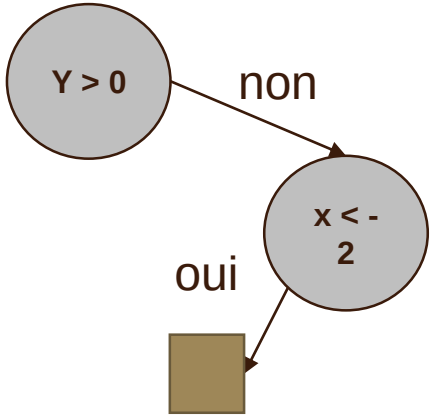


Arbre de décision : représentation de l'arbre ou du chemin emprunté par une donnée

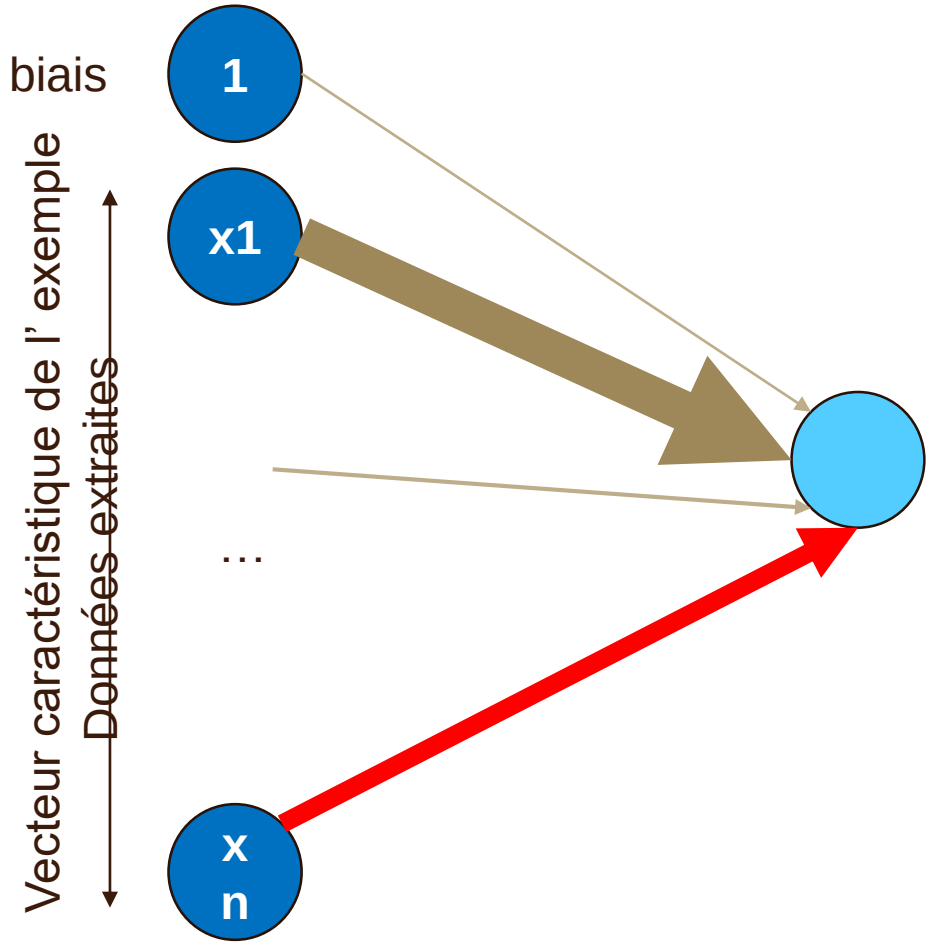
Explication globale



Explication locale



Régression logistique : importance des (valeur absolue des) poids



Limites des explications intrinsèques

- **Limitées à des modèles simples**
  - **Donc pas forcément les plus performants**
- **Ne passent pas l'échelle**
  - **Comment visualiser les plus proches voisins lorsqu'on ne connait 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 intéractive est une solution : illustration avec la régression logistique

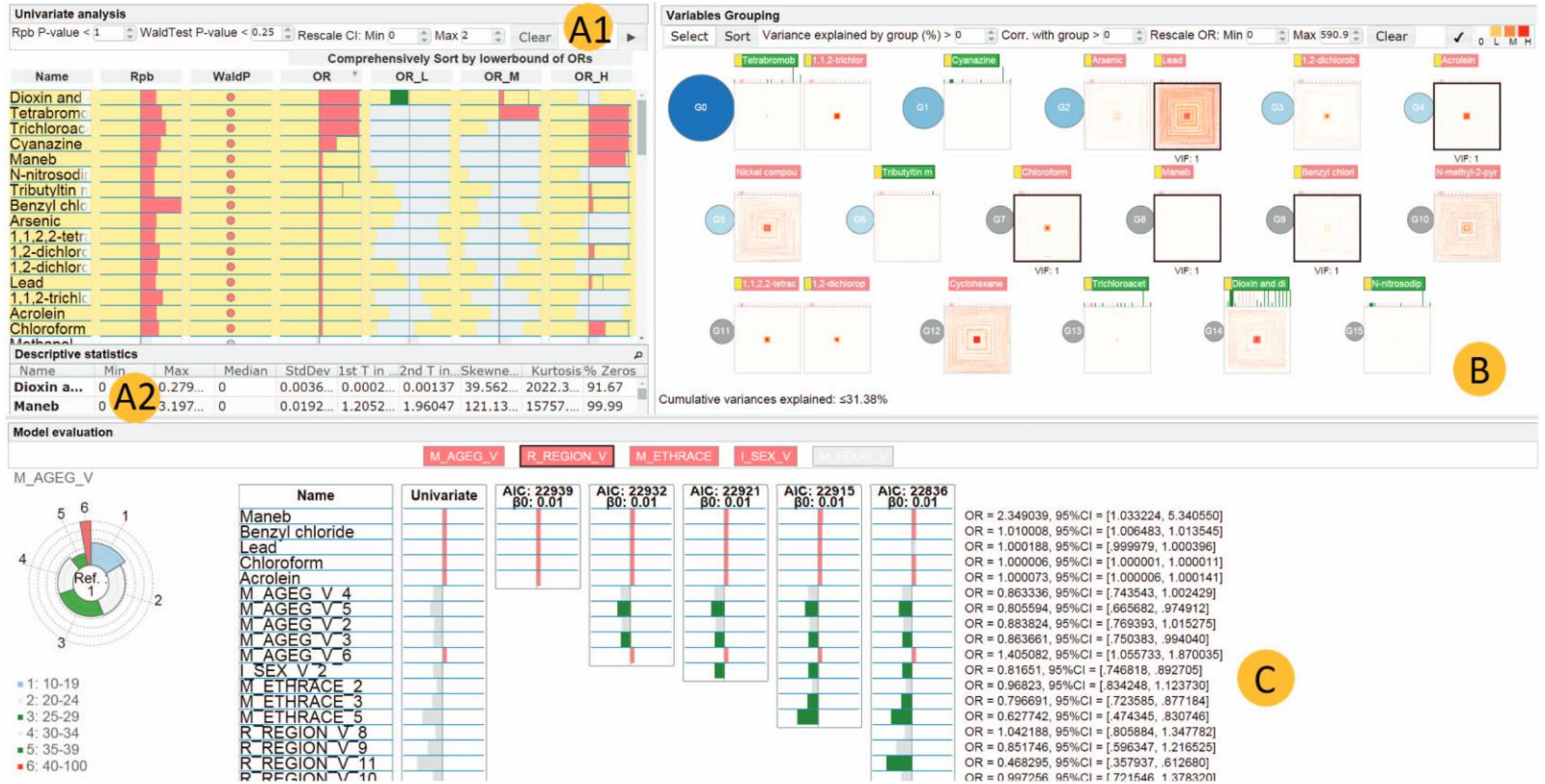


Figure 1: The interface of the prototype where the limb reduction defect is analyzed. (A1) The univariate analysis view. (A2) The descriptive statistics information panel. (B) The variable grouping view. (C) The model evaluation and comparison view.

Zhang, Chong, et al. "A visual analytics approach to high-dimensional logistic regression modeling and its application to an environmental health study." 2016 IEEE Pacific Visualization Symposium (PacificVis). IEEE, 2016.

L’analyse visuelle et intéractive est une solution : illustration avec les arbres de décision

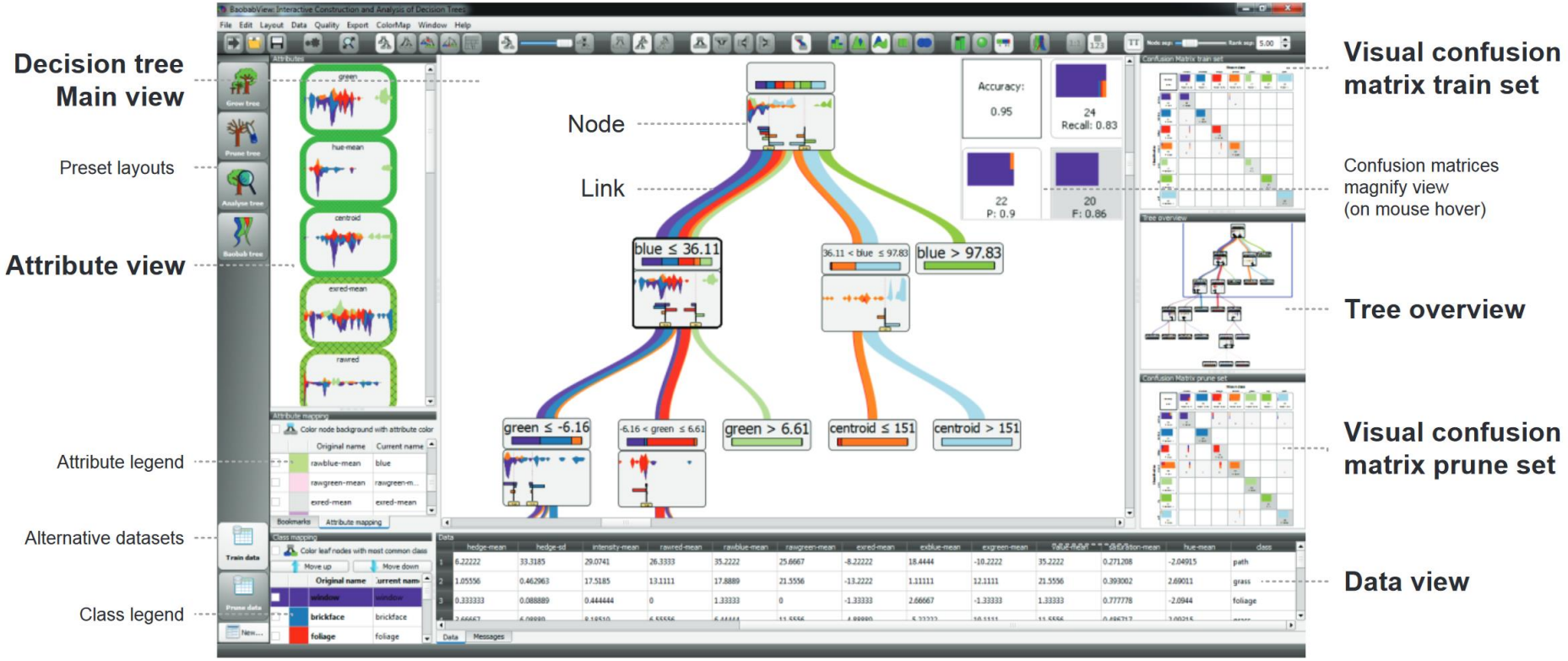


Figure 3: Interface of the interactive decision tree construction software with according proposed decision tree visualization; Based on adapted node-link diagram. Nodes contain important decision tree components. Links are visualized as a stream of items.

Van Den Elzen, Stef, and Jarke J. Van Wijk. "Baobabview: Interactive construction and analysis of decision trees." *2011 IEEE conference on visual analytics science and technology (VAST)*. IEEE, 2011.

L’analyse visuelle et intéractive 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.

Zhao, Xun, et al. "iforest: Interpreting random forests via visual analytics." *IEEE transactions on visualization and computer graphics* 25.1 (2018): 407-416.

Discussion sur les explications intrinsèques

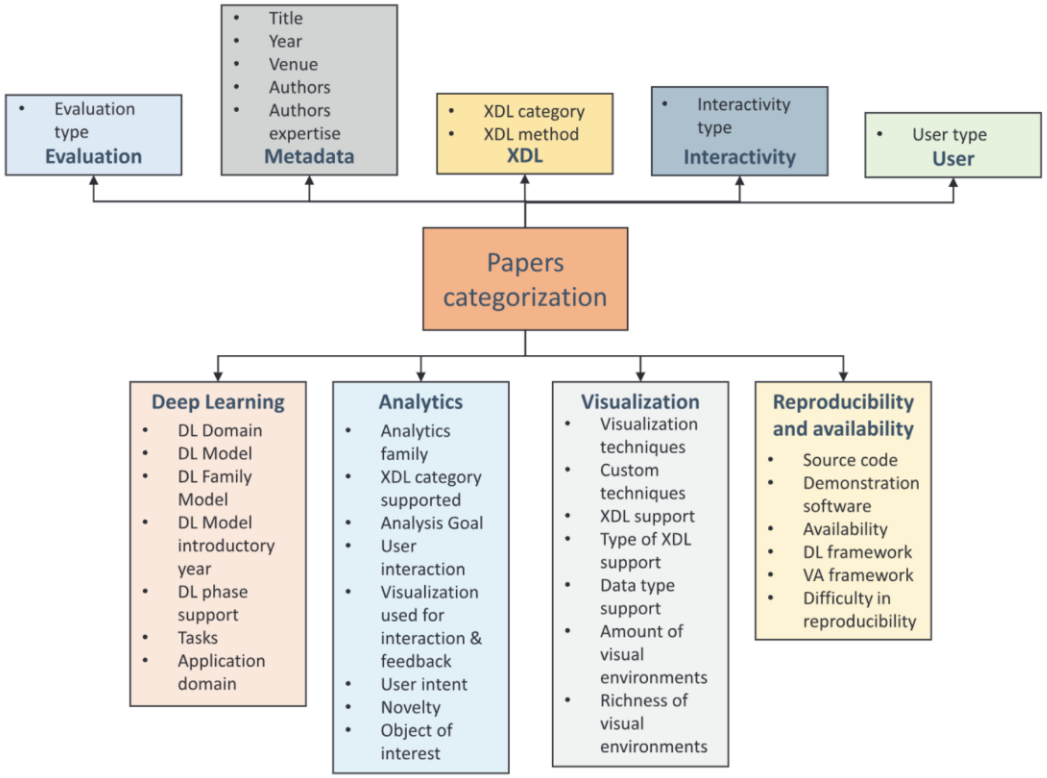
- Certaines familles de modèles sont auto-explicables
  - Globalement sur l'ensemble des données
  - Localement pour un exemple particulier
- Cependant, elles passent rarement l'échelle
- L'analyse visuelle et interactive peut aider à mieux appréhender le modèle
  - Il est nécessaire de définir les tâches 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 intéractive 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



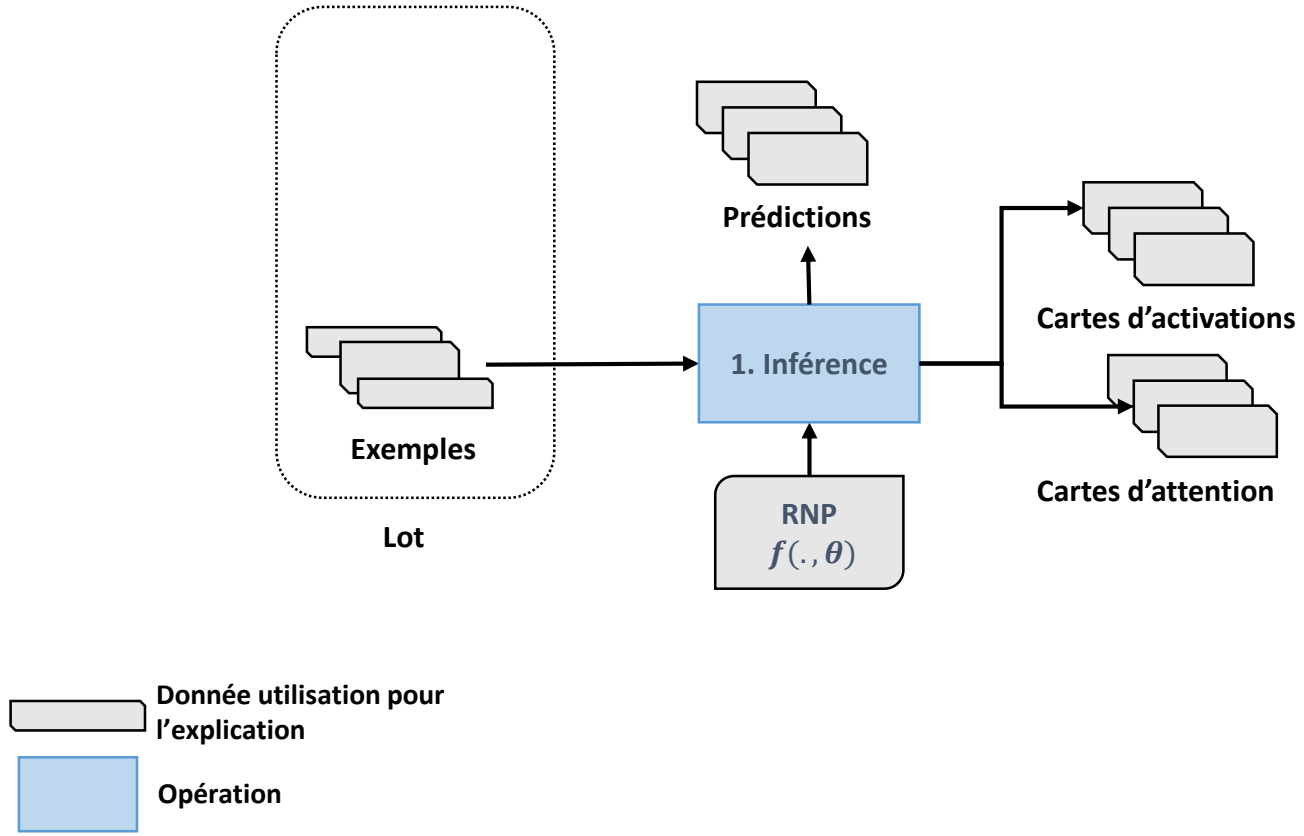
La Rosa, B., Blasilli, G., Bourqui, R., Auber, D., Santucci, G., Capobianco, R., Bertini, E., Giot, R. & Angelini, M. (2023, February). State of the art of visual analytics for explainable deep learning. In *Computer Graphics Forum* (Vol. 42, No. 1, pp. 319-355).

Critères de comparaison

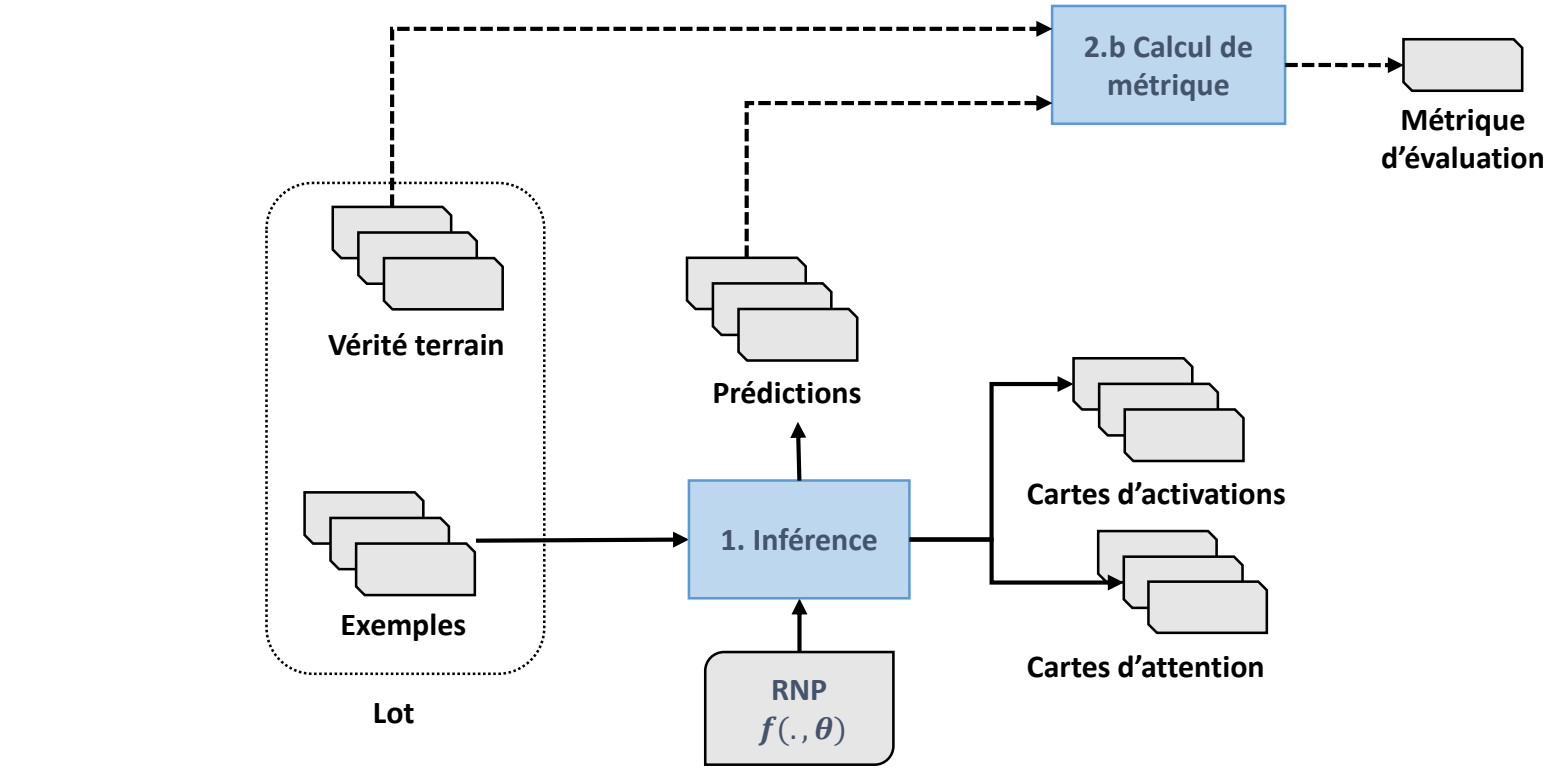
#	System Name	Reference	Venue	Year	Category					Application Domain	DL Model	Users			Interactivity			Phase		Evaluation					Code
					FA	LF	EE	CE	MB			A	T	E	P	I	M	TR	TE	Q-USt	F-USt	F-CSt	CSt	USc	
01	DG-Viz	[LYY*20]	JMIR	2020	■					Medical	RNN			👤		📊			■				✓		✓
02	RetainVis	[KCK*19]	TVCG	2018	■					Medical	RNN			👥		📊	⚙		■				✓		✓
03	V-Awake	[CWGvW19]	CGF	2019	■					Medical	Other			👥	👁				■			✓			✓
04	ViSFA	[WWM20]	ArXiv	2020	■					Domain agnostic	RNN			👥	👁			■						✓	✓
05	VisLRPDesigner	[HJZ*21]	CGF	2021	■					XDL	CNN		📈			📊	⚙		■	✓	✓			✓	✓
06	VisQA	[JKV*22]	TVCG	2021	■					Domain agnostic	Transformers	👥				📊	⚙		■	✓		✓			
07	DeepVID	[WGZ*19]	TVCG	2019	■					Domain agnostic	Other	👥					⚙	■				✓			
08	GLANCE	[vdBCR*20]	IOVS	2020	■					Medical	CNN			👤	👁				■	✓	✓	✓			
09	IVDAS	[HSL*21]	JVIS	2021	■					BioInformatics	CNN		📈			📊			■				✓		
10	GANViz	[WGYS18]	TVCG	2018	■					Domain agnostic	Generative	👥			👁			■			✓		✓		✓
11	explAIner	[SSSE19]	TVCG	2019	■					Domain agnostic	Variable		📈		👁			■	■		✓				✓
12	AttributionHeatmaps	[WONM18]	IEEE Big Data	2018	■					Domain agnostic	RNN		📈		👁			■					✓		✓
13	NNVA	[HLW*19]	TVCG	2020	■					BioInformatics	Variable			👤		📊	⚙		■			✓			
14	NJM-Vis	[JCM20]	IUI	2020	■					Domain agnostic	Variable		📈		👁				■		✓		✓		
15	BERTViz	[Vig19]	ACL-SD	2019	■					Domain agnostic	Transformers		📈			📊			■					✓	✓
16	–	[DWSZ20]	PacificVIS	2020	■					Domain agnostic	RNN			👤		📊			■	✓				✓	✓
17	Dodrio	[WTC21]	ACL-SD	2021	■					Domain agnostic	Transformers		📈		👁				■					✓	✓
18	Attention Flows	[DWB21]	TVCG	2020	■					Domain agnostic	Transformers	👥				📊			■		✓		✓		✓
19	–	[CHS20]	Vis.Inform.	2020	■					Linguistics	CNN		📈			📊			■				✓		
20	–	[HCC*20]	Distill	2020	■					Game agent	DRL	👥			👁				■					✓	✓
21	TSViz	[SMM*19]	IEEE Access	2019	■					Domain agnostic	CNN	👥			👁				■					✓	✓
22	–	[CGR*17]	VADL	2017	■					Domain agnostic	RNN	👥			👁			■						✓	
23	AttViz	[ŠSE*21]	EACL	2021	■					Domain agnostic	Transformers		📈		👁				■					✓	
24	–	[MFH*21]	TVCG	2020		■				Domain agnostic	CNN		📈		👁				■		✓		✓		✓
25	DeepEyes	[PHG*18]	TVCG	2017		■				Domain agnostic	CNN	👥			👁			■						✓	
26	ProtoViewer	[ZDXR20]	VIS	2020		■				Domain agnostic	CNN	👥			👁				■					✓	
27	Deep View	[ZXZ*17]	ICML-Viz	2017		■				Domain agnostic	CNN	👥			👁			■						✓	
28	CNNComparator	[ZHP*17]	VADL	2017		■				Domain agnostic	CNN		📈		👁				■					✓	
29	DeepVis	[YCN*15]	ICML	2015		■				Domain agnostic	CNN	👥			👁				■					✓	
30	ShapeShop	[HHC17]	CHI	2017		■				Domain agnostic	Variable		📈			📊			■					✓	
31	Bluff	[DPW*20]	VIS	2020		■				Domain agnostic	CNN	👥			👁				■					✓	✓
32	NeuroCartography	[PDD*22]	TVCG	2021		■				Domain agnostic	CNN		📈		👁				■	✓				✓	✓
33	Summit	[HPRC20]	TVCG	2019		■				Domain agnostic	CNN	👥			👁				■					✓	
34	TopoAct	[RCPW21]	CGF	2021		■				Domain agnostic	Variable		📈		👁				■				✓		
35	CNNVis	[LSL*17]	TVCG	2016		■				Domain agnostic	CNN	👥			👁				■				✓		
36	RNNVis	[MCZ*17]	VIS	2017		■				Domain agnostic	RNN		📈			📊			■		✓		✓		
37	DeepCompare	[MMD*19]	CG&A	2019		■				Domain agnostic	Variable		📈		👁				■			✓	✓		
38	What-If Tool	[WPB*19]	TVCG	2019				■		Domain agnostic	Variable		📈			📊			■				✓		✓
39	DECE	[CMQ21]	TVCG	2020				■		Domain agnostic	Variable		📈			📊		■	■		✓		✓		✓
40	–	[BDME20]	VizSec	2020					■	Cybersecurity	Variable	👥			👁				■				✓		
41	GNNLens	[JWW*22]	TVCG	2020					■	Domain agnostic	GNN	👥			👁				■		✓		✓		✓
42	ActiVis	[KAKC18]	TVCG	2017					■	Domain agnostic	Variable		📈			📊			■			✓	✓		
43	DRLIVE	[WZY*22]	TVCG	2021					■	Game agent	DRL		📈			📊			■			✓			✓
44	SCANViz	[WZY20]	PacificVIS	2020					■	Domain agnostic	Generative	👥				📊	⚙	■				✓			✓
45	–	[SW17]	VADL	2017	■	■				Domain agnostic	CNN		📈		👁				■				✓		
46	SANVis	[PCN*19]	VIS	2019	■	■				Domain agnostic	Transformers	👥				📊			■				✓		
47	Seq2Seq-Vis	[SGB*19]	TVCG	2018	■		■			Domain agnostic	Other		📈			📊	⚙		■				✓		
48	exBERT	[HSG20]	ACL-SD	2020	■		■			Domain agnostic	Transformers		📈			📊			■			✓			✓
49	ProtoSteer	[MXC*20]	TVCG	2019	■		■			Domain agnostic	Other			👥			⚙		■			✓			✓
50	PolicyExplainer	[MSHB22]	PacificVIS	2022	■			■		Domain agnostic	DRL			👥		📊			■		✓		✓		✓
51	Recast	[WSP*21]	ACM-HCI	2021	■			■		Linguistics	Transformers			👥		📊			■		✓				
52	MultiRNNExplorer	[SWJ*20]	PacificVIS	2020	■				■	Environment	RNN		📈		👁				■				✓		
53	DQNViz	[WGSY19]	TVCG	2018	■				■	Game agent	DRL	👥			👁			■				✓			
54	DRLViz	[JVW20]	CGF	2020	■				■	Game agent	DRL	👥			👁				■		✓				
55	–	[ZZM16]	ICML	2016	■				■	Game agent	DRL	👥			👁				■				✓		
56	VATUN	[PYN*21]	EuroVis	2021	■				■	Domain agnostic	CNN		📈			📊			■				✓		✓
57	M2Lens	[WHJ*22]	TVCG	2021	■				■	Linguistics	Variable			👥	👁				■			✓			✓
58	Blocks	[BJY*18]	TVCG	2017		■			■	Domain agnostic	CNN	👥			👁			■					✓		
59	AEVis	[LLS*18]	TVCG	2020		■			■	Domain agnostic	CNN	👥			👁				■				✓		
60	TNNVis	[NHP*18]	PacificVIS	2018		■			■	Domain agnostic	CNN	👥				📊		■	■				✓		
61	CNN2DT	[JLL*19]	JVIS	2020		■			■	Domain agnostic	CNN		📈		👁				■		✓		✓		✓
62	DynamicsExplorer	[HLvB*20]	PacificVIS	2020			■		■	Robotics	DRL	👥			👁				■					✓	
63	LSTMVis	[SGPR18]	TVCG	2017			■	■	■	Domain agnostic	RNN		📈			📊	⚙		■		✓			✓	✓
64	USEVis	[JTH*21]	Vis.Inform.	2021	■	■	■			Domain agnostic	Transformers		📈		👁				■			✓			
65	Melody	[CBN*20]	ArXiv	2020	■		■	■		Domain agnostic	Variable	👥				📊			■					✓	
66	HypperSteer	[WM20]	ArXiv	2020	■			■	■	Domain agnostic	RNN		📈			📊			■					✓	
67	NLIZE	[LLL*19]	TVCG	2018	■		■	■		Linguistics	Other	👥				📊	⚙		■		✓		✓		✓

Visualizations																	Analytics Family														Analytics Subject	Analytics Novelty
Heatmap	Node-Link	Line/Area	Bar	Image	Scatter	Words	Table	Matrix	Sankey	Pie	PCP	Radial	Treemap	Boxplot	Tree	Similarity Analysis	Feature Influence	Aggregation/Summarization	Clustering	What-if Analysis	Model/Framework	Search/Mining	Saliency Analysis	Activation Maximization	Statistical Analysis	Semantic Analysis	Pattern Analysis					
	●		●		■					■										■	■									Input	●
	■		■		■		■									■					■									Input/Output	●
■		■		■		■	■							■		■															Temporal stages	
		■	■	■		■		■									■														Features	●
■			■	■		■		■											■												Image	●
■	■			■				■											■												Attention Heads	○
■		■		■						■											■										Image	●
■		■		■																											Image	
■	■	■	■	■	■		■														■										Image	
■	■	■		■	■												■														Image	○
■	■	■	■	■	■															■											All	○
■	■	■					■	■													■										Input	
■	■	■	■				■				■						■														Input	○
■	■		■			■	■										■														Words	○
■	■					■																									Attention Heads	
■	■					■																									Sentences/Words	○
■						■	■										■														Words	●
	■		■	■		■				■	■						■	■													Words	●
			■	■		■											■														Words	●
■			■	■																											Image	
	■		■														■														Input/CNN Filters	●
	■		■								■							■													Sentences/Words	
	■				■	■													■												Words	○
	■		■				■			■				■						■											Network	●
	■		■		■		■	■					■																		Filters	●
			■					■										■													Prototypes	●
	■																■		■												Neurons	●
	■		■														■						■								Input/Model	
	■		■					■														■									Images	○
	■																					■									Activations	
	■		■						■													■									Input	○
	■		■						■								■														Neurons	●
	■		■						■								■														Neurons	●
	■																■														Activations	○
	■		■						■									■													Neurons	○
	■		■				■		■						■				■												Hidden states	○
	■						■		■				■										■								Activations	
	■						■																■								Input	○
	■						■										■					■									Input	○
	■						■															■									Input	●
	■																						■								Activations	○
	■																						■								Activations	○
	■																						■								Activations	○
	■																						■								Internals	○
	■																						■								Activations	
	■																						■								Internals/Input	○
	■																						■								Internals	
																							■								Input/Activations	○
	■		■																				■								Attention Heads	○
	■																						■								Input/Attention Heads	○
	■		■																				■								Words	○
	■		■																				■								Sentences	○
	■																						■								Internals	●
	■																						■								Internals/Input	○
	■		■																				■								Features/Hidden units	●
	■																						■								Internals/Input	
	■																						■								Internals	
	■																						■								Image	
	■																						■								Input	
	■																						■								Features	●
	■																						■								Neurons	●
	■																						■								Internals	●
	■																						■								nan	
	■																						■								Model	○
	■																						■								Hidden States	
	■																						■								Hidden States	
	■																						■								Attention Heads	○
	■																						■								Internals	●
	■																						■								Internals	
	■																						■								Words	○

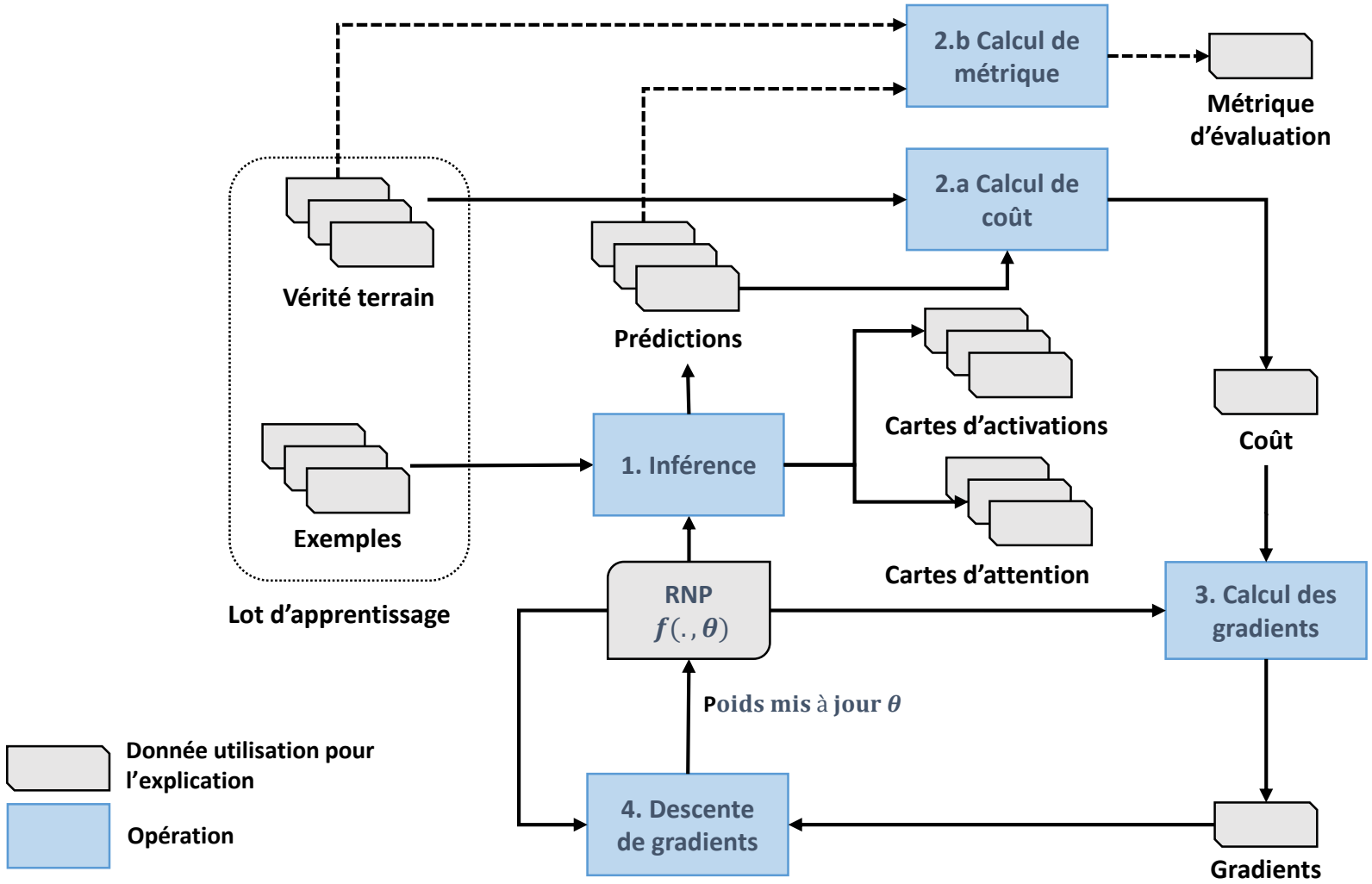
Ensemble des données accessibles pour expliquer : à l'inférence sans vérité terrain



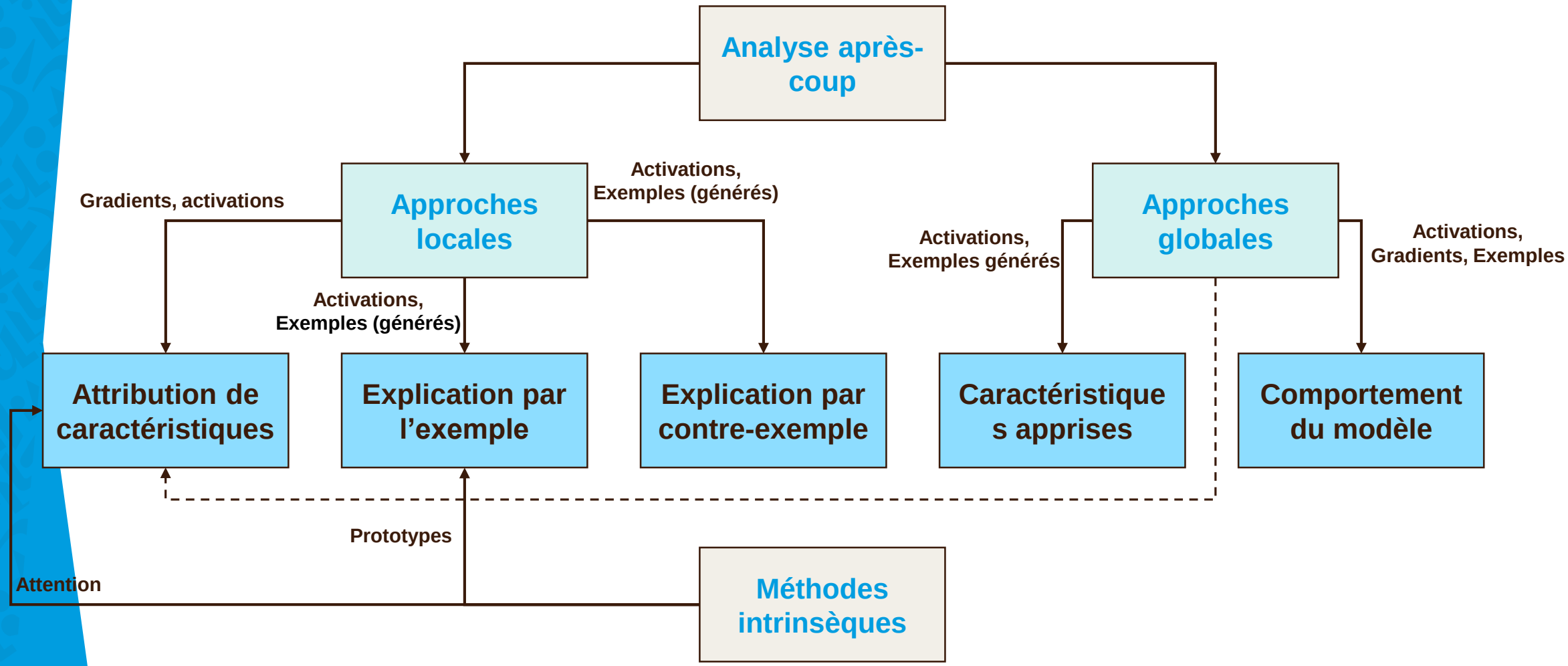
Ensemble des données accessibles pour expliquer : à l'inférence avec vérité terrain



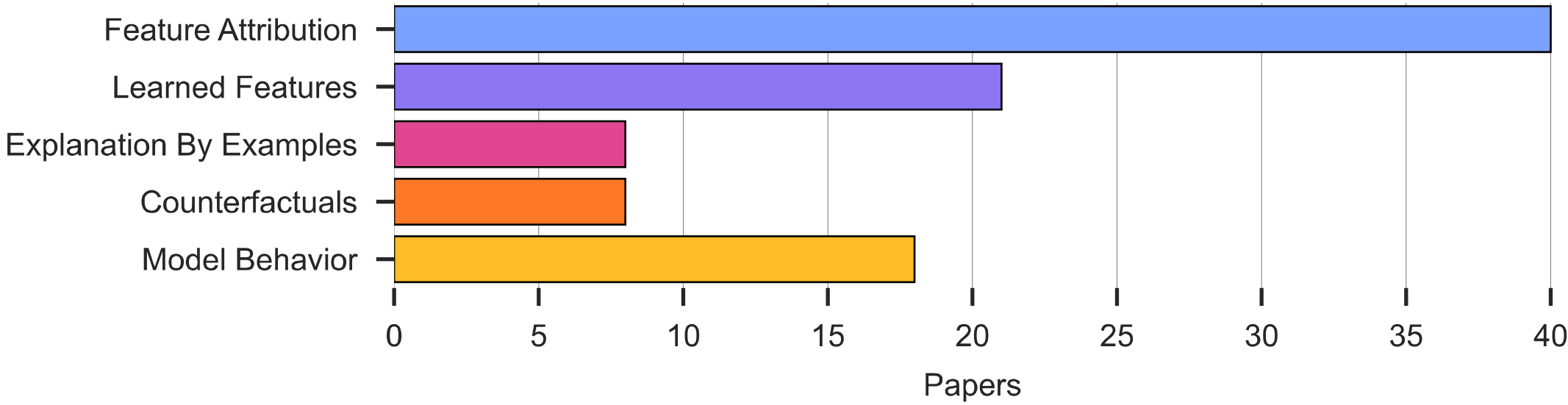
Ensemble des données accessibles pour expliquer : lors de l'apprentissage



Taxonomie des méthodes d'explication



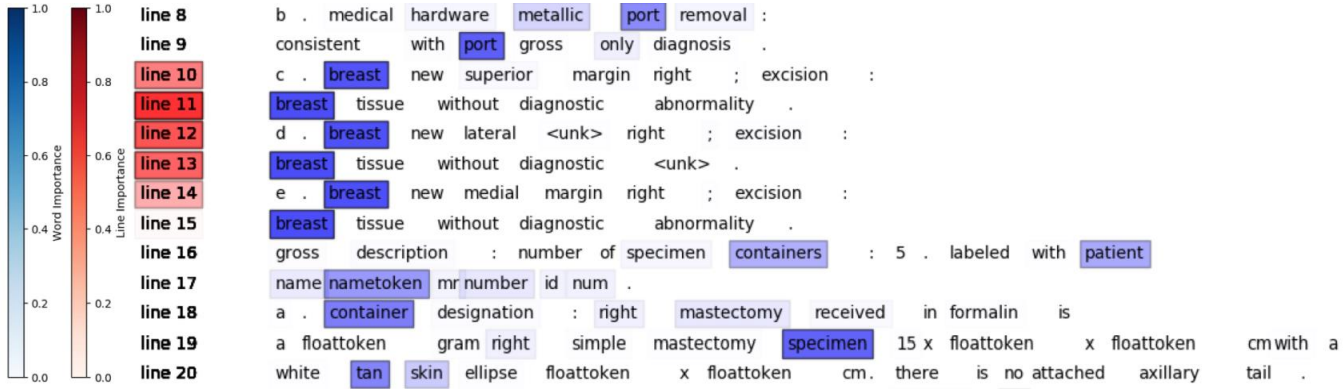
Nette sur-représentation des outils utilisant l’attribution de caractéristiques



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**

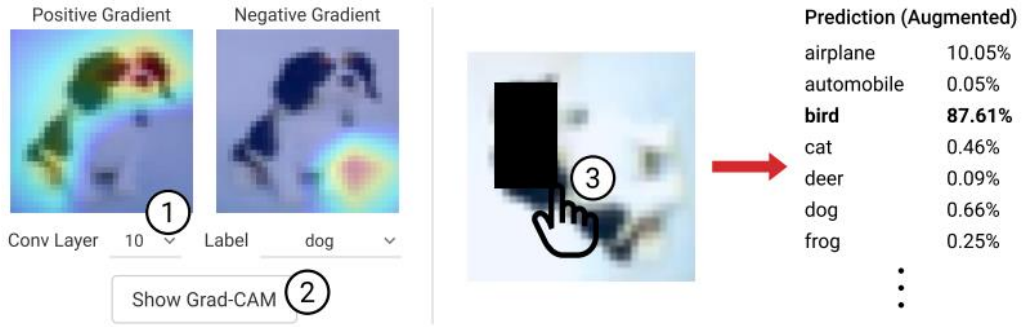
Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.



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

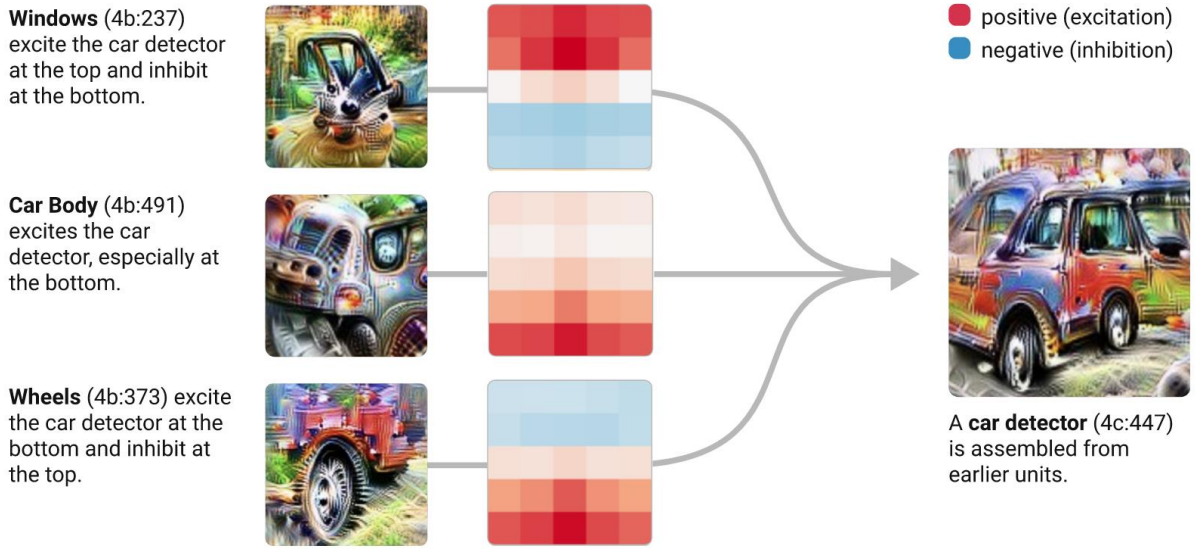
- Vérification de la robustesse du modèle
- Analyse des erreurs
- Exploration de données
- Contrôle de l’amélioration du modèle
- Analyse « quoi si ? »
- Détection de biais
- …



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

Caractéristiques apprises

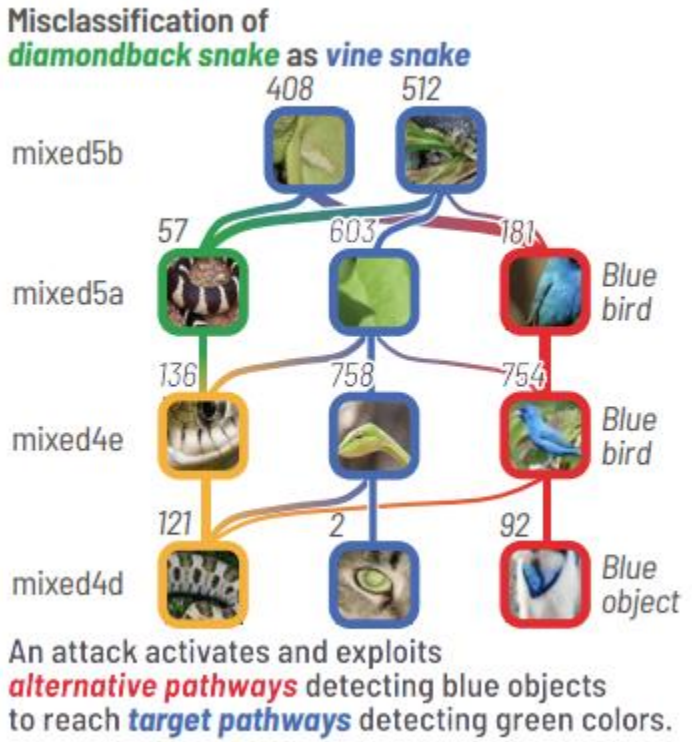
- **Qu’est-ce que les données ont appris durant l’apprentissage ?**
  - **Génère les caractéristiques qui excitent fortement un neurone**
- **Principaux défis**
  - **Gérer les milliers de neurones et couches**
- **Solutions**
  - **Abstractions et partitionnement**
  - **Encodage spécifiques aux données**



Olah, et al., "Zoom In: An Introduction to Circuits", Distill, 2020.

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)

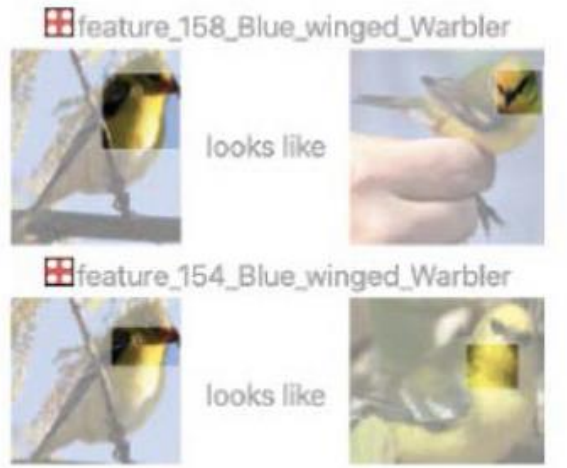
Explication par exemples

- **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

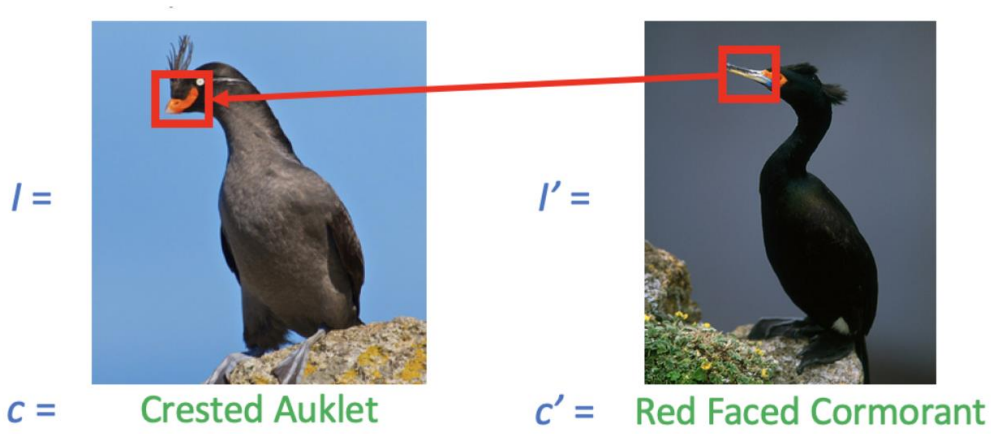
- Identification d’erreurs
- Correction des hyperparamètres de l’apprentissage
- Amélioration de l’architecture
- Comparaison de prédictions
- Analyse « quoi si ? »



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

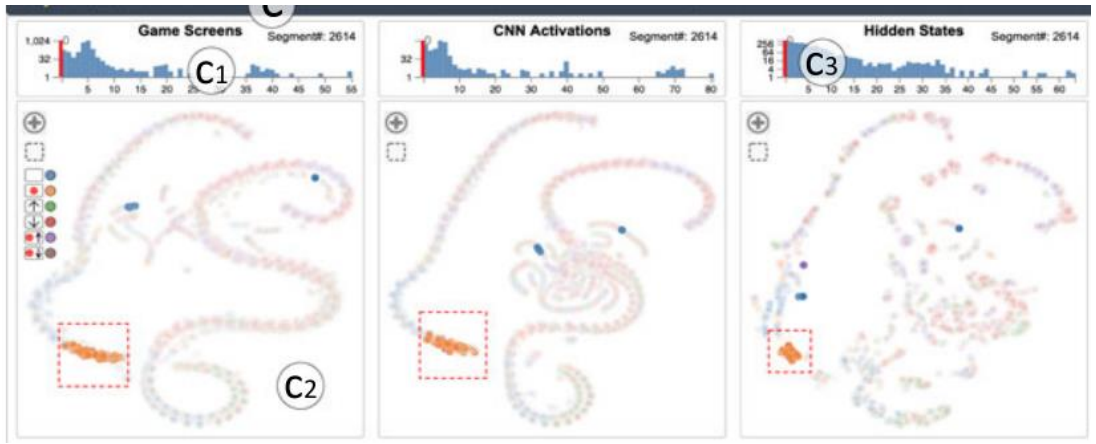
- **Que faut-il changer pour obtenir une prédiction différente ?**
  - **Proposition d'un exemple légèrement différent générant une autre prédiction**
  - **Ou proposition d'un changement à effectuer**



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

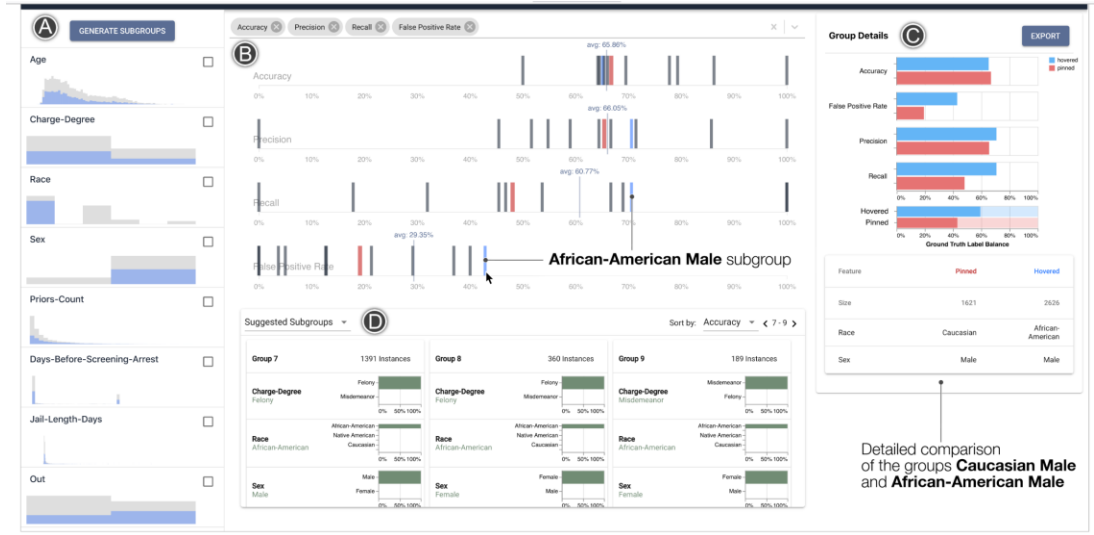


Autres aspects importants

Le modèle n’est pas la seule partie importante à analyser

Les données sont plus importantes que les modèles

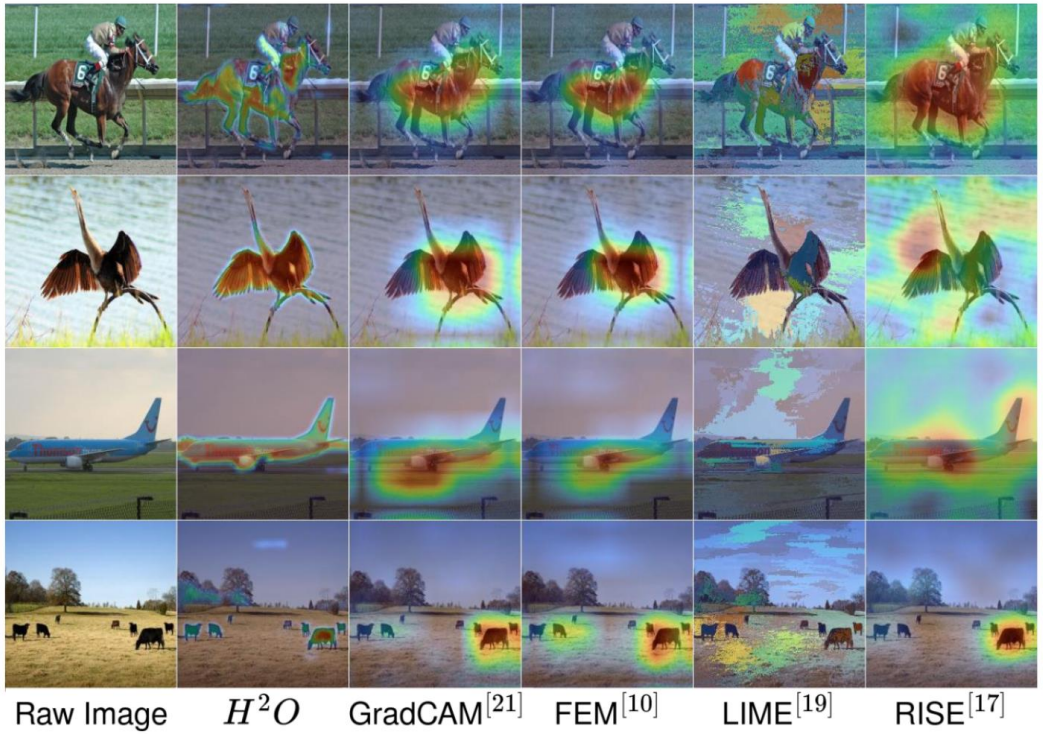
- Les performances des modèles dépendent principalement des données d'apprentissage
- Ces données peuvent souffrir de différents biais
  - Il est important de les connaître
  - Rarement pris en compte
- Ces données peuvent être sujet à oubli
  - Il est important de pouvoir les retirer de la connaissance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



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 ?

Pommé, Luc-Etienne, Romain Bourqui, and Romain Giot. "H²O: Heatmap by Hierarchical Occlusion." *CBMI 2023*. 2023.

Dombrowski, Ann-Kathrin, et al. "Explanations can be manipulated and geometry is to blame." *Advances in neural information processing systems* 32 (2019).

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

Les masses de données

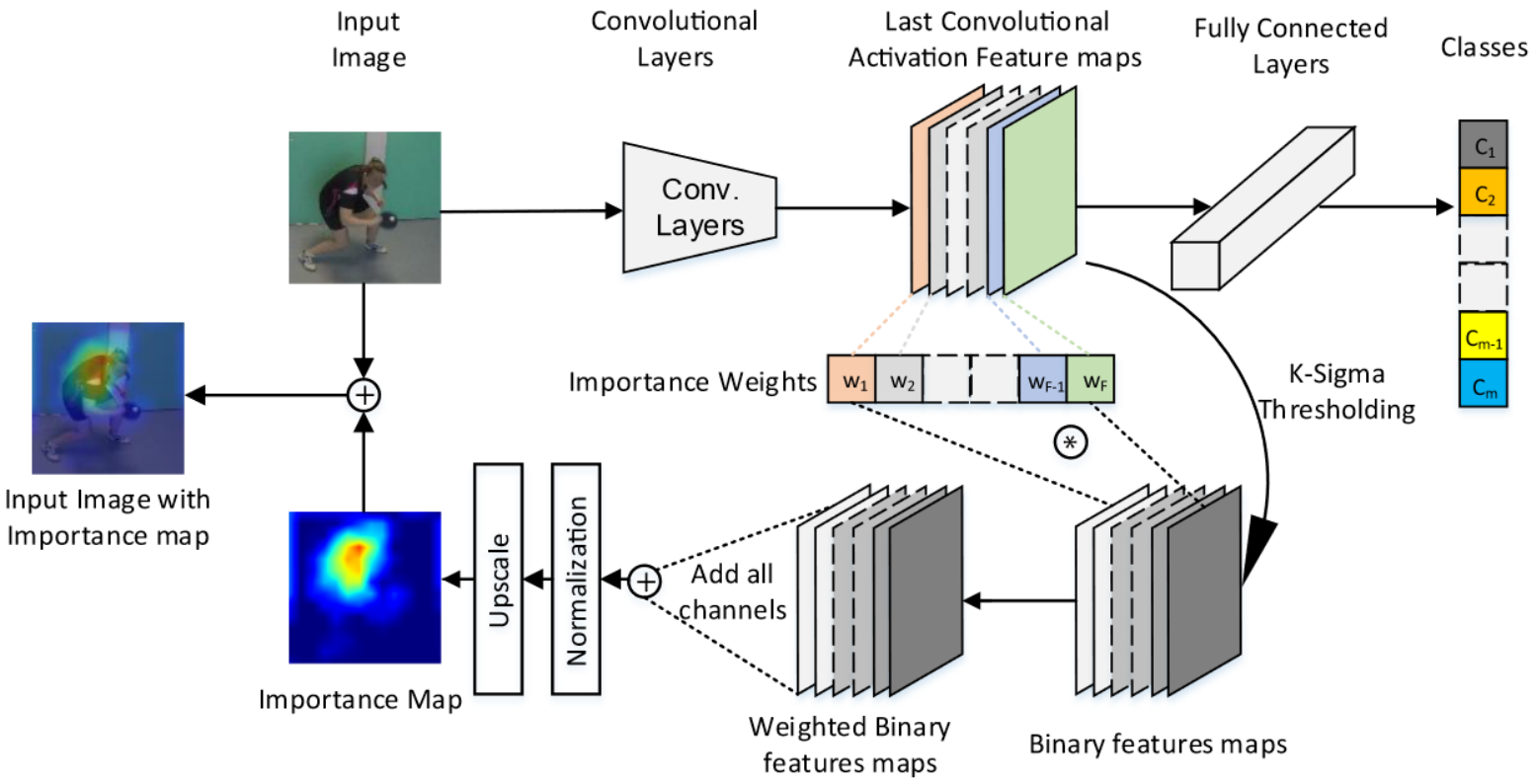
- Collecte des données
- Propriété intellectuelle

Fiabilité des systèmes

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

Sélection de travaux sur l’explicabilité

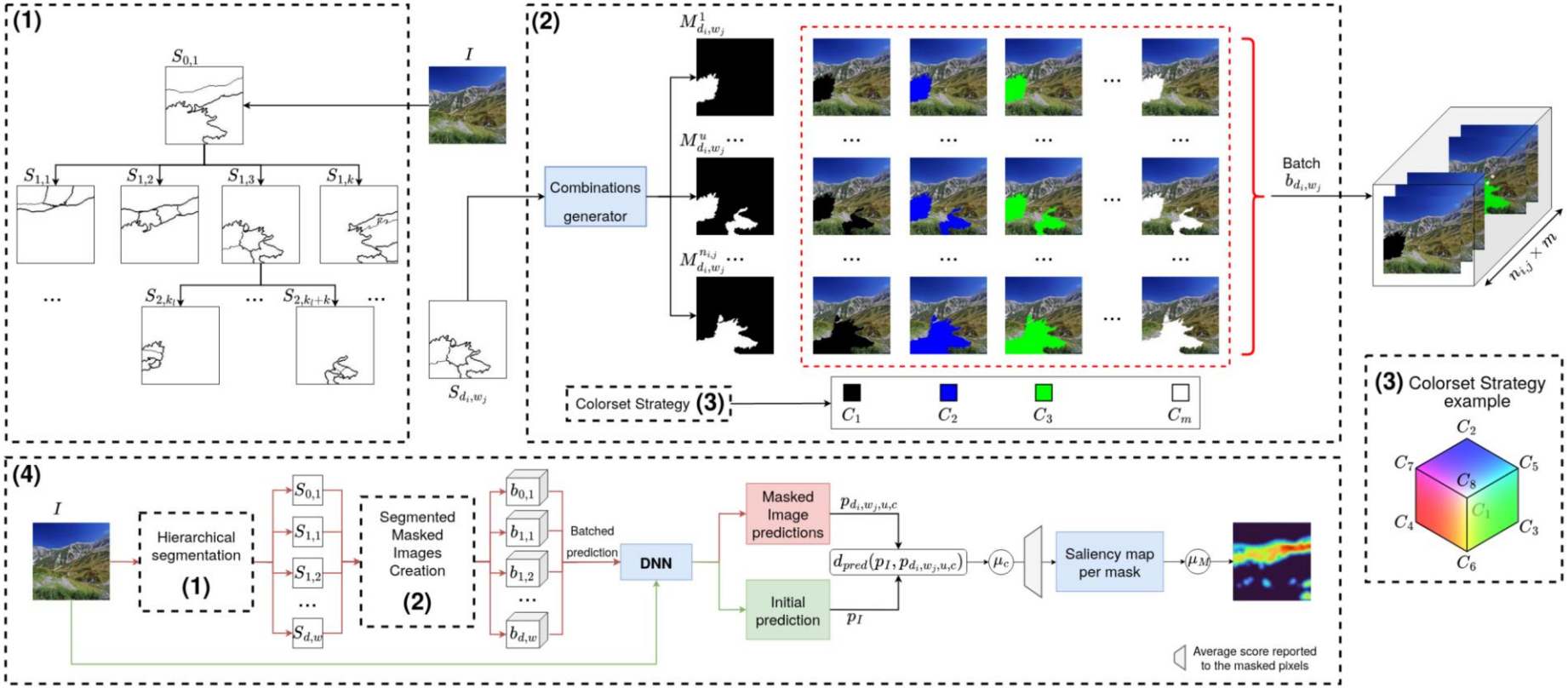
Features Explanation Method



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

Fuad, K. A. A., Martin, P. E., Giot, R., Bourqui, R., Benois-Pineau, J., & Zemmari, A. (2020, November). Features understanding in 3D CNNS for actions recognition in video. In 2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA) (pp. 1-6). IEEE.

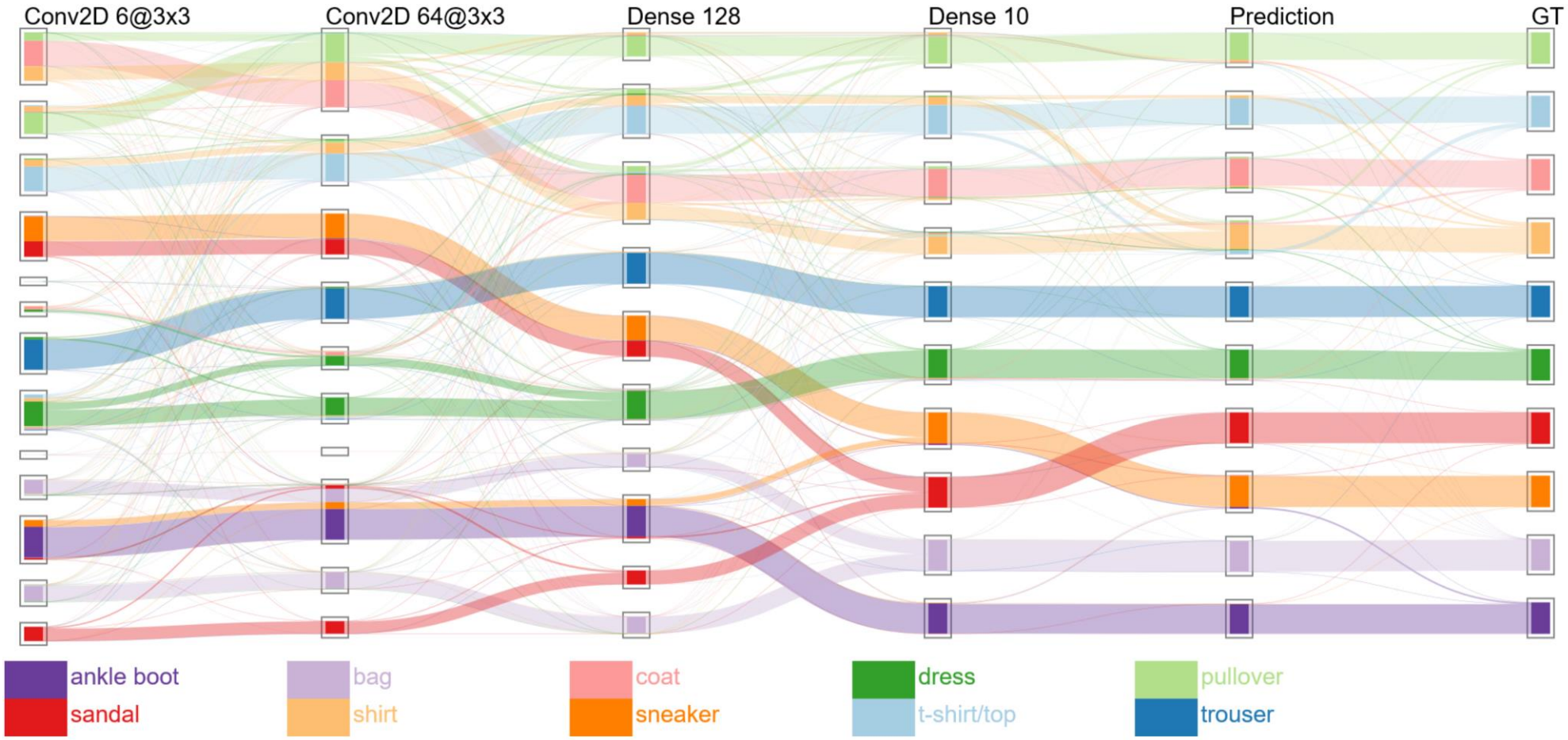
H2O: Heatmap by Hierarchical Occlusion



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

Pommé, Luc-Etienne, Romain Bourqui, and Romain Giot. "H²O: Heatmap by Hierarchical Occlusion." *CBMI 2023*. 2023.

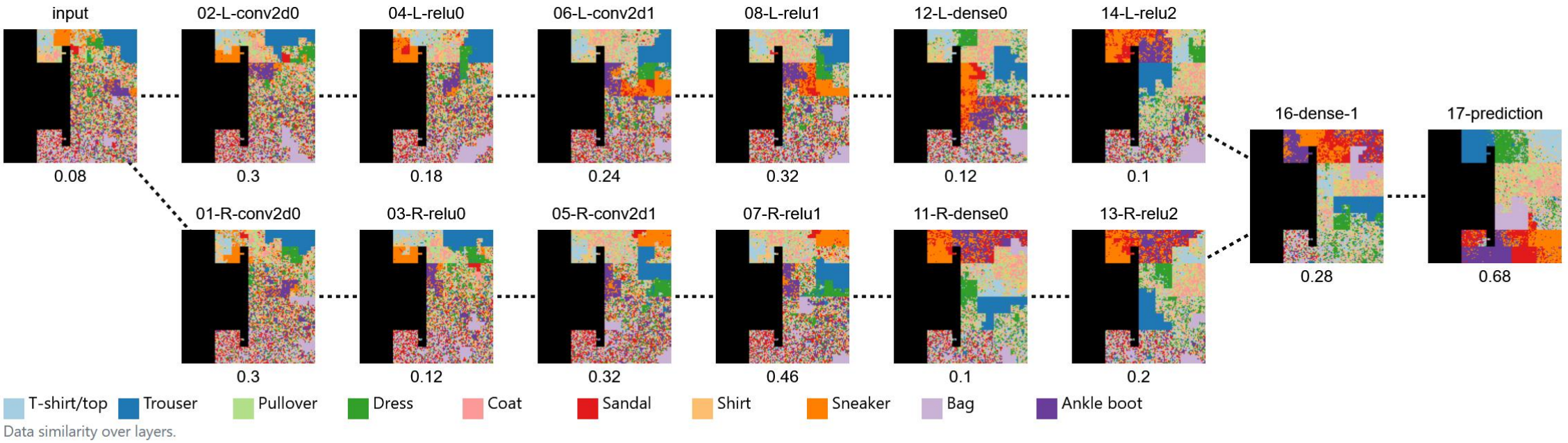
Deep Dive into Deep Neural Networks with Flows



A. Halnaut, R. Giot, R. Bourqui, D. Auber. Deep Dive into Deep Neural Networks with Flows. Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020): IVAPP, Feb 2020, Valletta, Malta. pp.231-239

Analyse du comportement du modèle

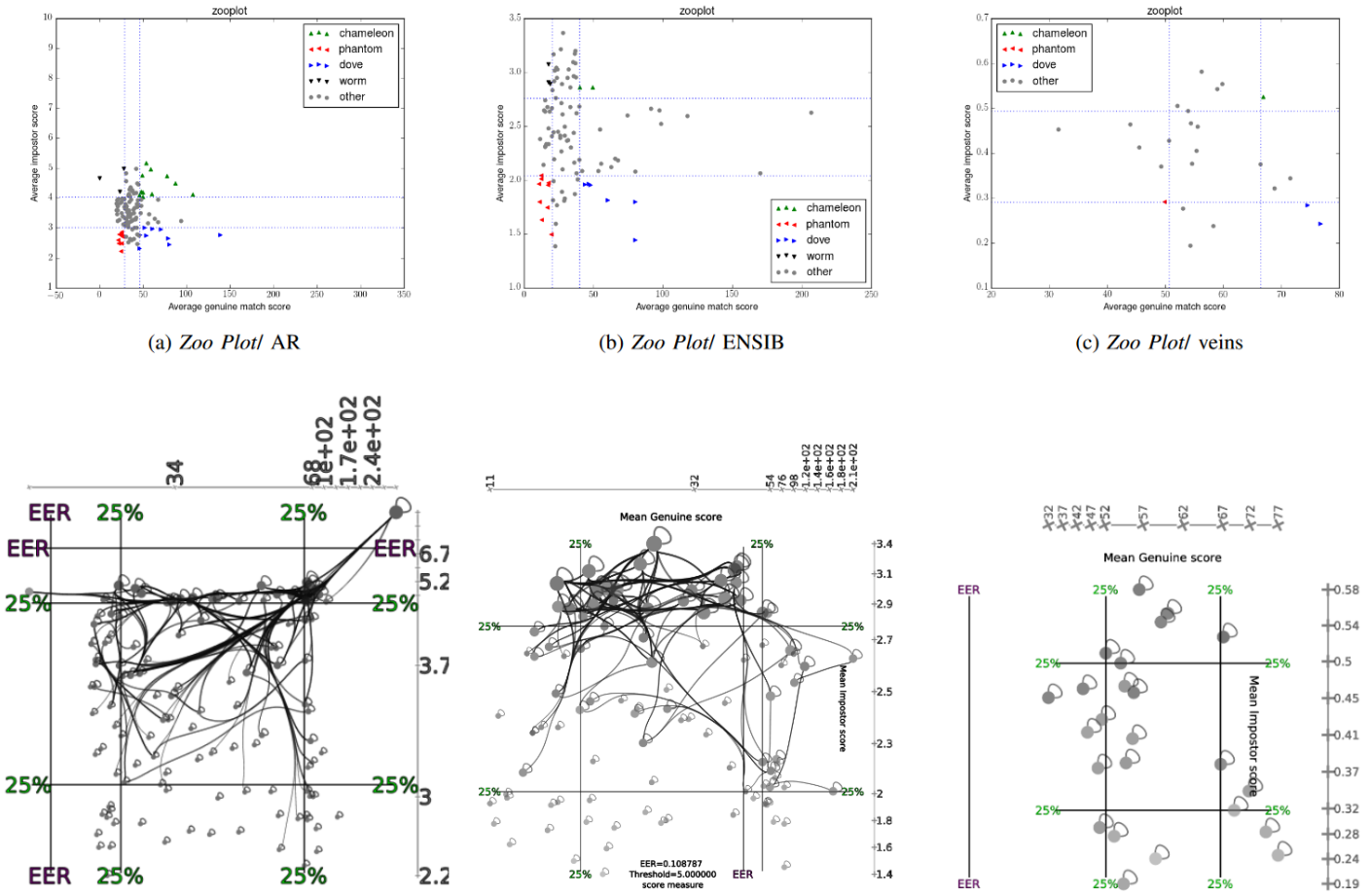
Samples Classification Analysis Across DNN Layers with Fractal Curves



Analyse du comportement du modèle

A. Halnaut, R. Giot, R. Bourqui, D. Auber. Samples Classification Analysis Across DNN Layers with Fractal Curves. ICPR 2020's Workshop Explainable Deep Learning for AI, Jan 2021, Milano (virtual), Italy.

Zoo Graph: a New Visualisation for Biometric System Evaluation



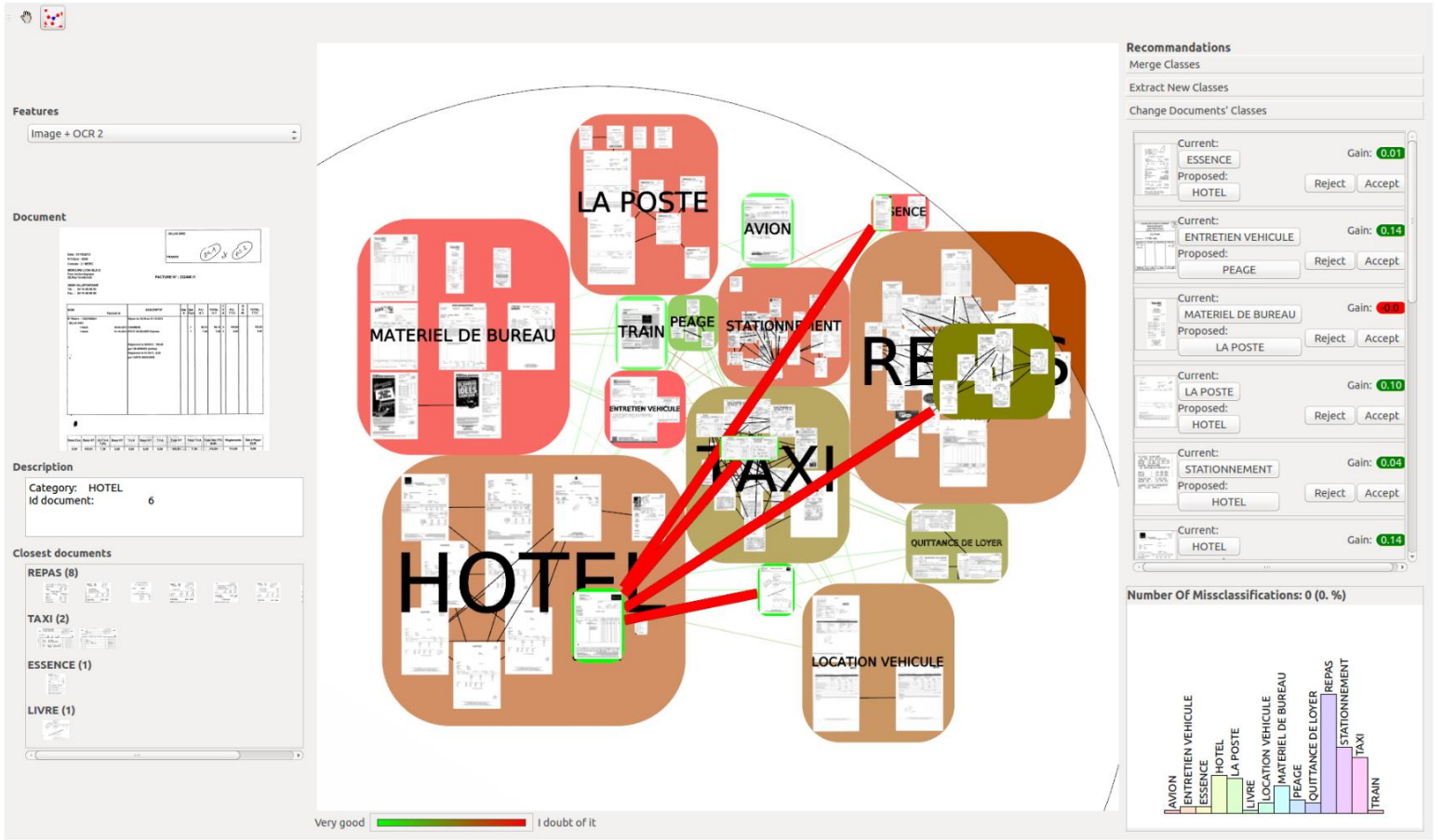
(a) *Zoo Plot*/ AR (b) *Zoo Plot*/ ENSIB (c) *Zoo Plot*/ veins

(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.

Giot, Romain, Romain Bourqui, and Mohamad El-Abed. "Zoo graph: a new visualisation for biometric system evaluation." 2016 20th International Conference Information Visualisation (IV). IEEE, 2016.

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



Giot, Romain, et al. "Visual graph analysis for quality assessment of manually labelled documents image database." 2015 13th International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2015.

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