Démystification de l'apprentissage automatique

Une formation ViaRezo

A propos de moi

Nicolas Fley

Etudiant en Césure à la Digital Tech Year™

Passionné par la data depuis quelque temps.

Ancien Responsable Communication de ViaRezo
Ancien Administrateur Système chez Hyris

nicolas.fley@student.ecp.fr

Pourquoi vouloir faire des Data Sciences ?

- **Un lien entre le conseil, les maths et l'informatique.**
- Des gains énormes dû aux process d'automatisation et de prise de décision (Dashboard/IA).
- **Le fantasme de l'intelligence artificielle.**
- C'est presque sexy de savoir différencier des chiens et des chats.
- **Ça rapporte vraiment beaucoup.**

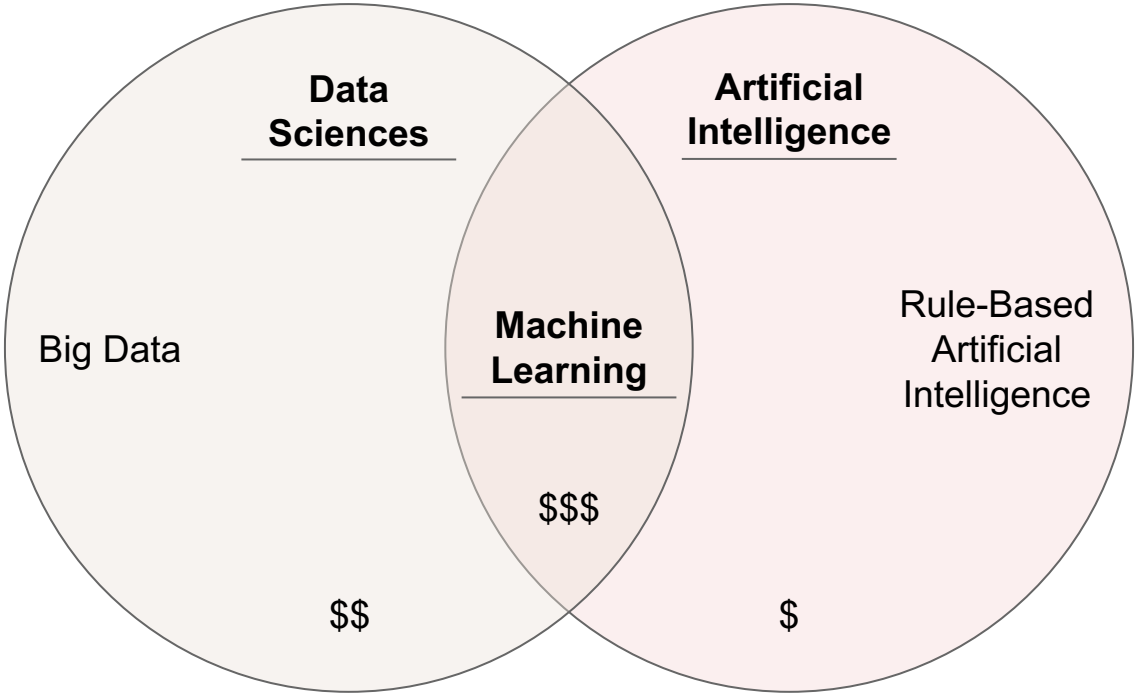
Plan de démystification

Plan de démystification 1h – 1h15

- C’est quoi le Machine Learning ?
- Rappel des principales notion associés.
- Comment construire mon modèle ?
- Les principaux algorithmes d’apprentissage.
- Explication des réseaux de neurones
- Petite démo

- **Pas de maths dans ces slides**

Artificial Intelligence vs Data Science vs Machine Learning



C’est quoi le machine learning ?

L'apprentissage automatique peut-être vu comme l’ensemble des techniques permettant à une machine d’apprendre à réaliser une tâche sans avoir à la programmer explicitement pour cela.

Arthur Samuel

Arthur Lee Samuel était un pionnier américain du jeu sur ordinateur, de l'intelligence artificielle et de l'apprentissage automatique.

C’est quoi le machine learning ?

- Différents types d’algorithmes

Supervisé

Croissant Chien

LR, KNN, SVM, RFT, NN...

Non Supervisé

Clustering, Anomaly detection

C’est quoi le machine learning ?

- Différents types d’algorithmes

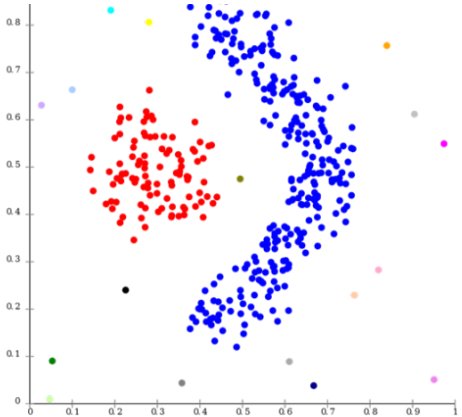
Supervisé

Croissant

Chien

LR, KNN, SVM, RFT, NN...

Non Supervisé

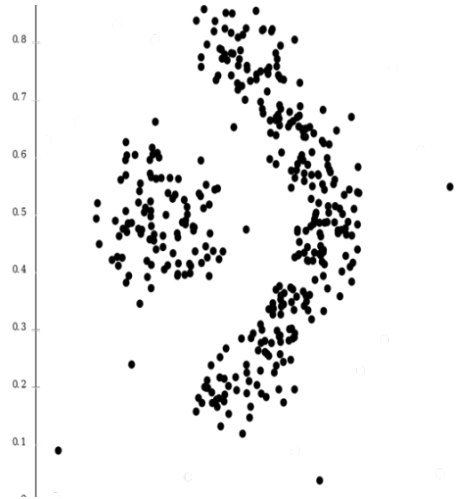


Clustering, Anomaly detection

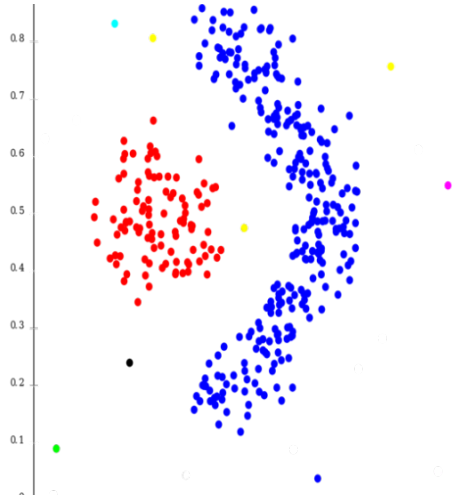
C’est quoi le machine learning ?

- Apprentissage non supervisé - entraînement

Clustering, Anomaly detection



Algorithmes
K-mean, density, ...



Les + : Pas besoin de Tag, visualisation souvent aisée, souvent très simple à comprendre et mettre en place
Les - : Très peu performant sur les problèmes complexes,

C’est quoi le machine learning ?

- Apprentissage supervisé - entraînement

Croissant

Chien

[1.409, 1.543, -1.348, 0.453, ...] Croissant
[1.109, 0.243, 1.138, -1.442, ...] Chien

[0,1,0,0,1,0,1,0,1,0,1,0,1,0,0,…] Croissant
[1,0,1,1,1,0,1,0,1,1,1,0,0,1,0,…] Chien

Modèle d’apprentissage

C’est quoi le machine learning ?

- Apprentissage supervisé - prédiction

[1.109, 0.243, 1.138, -1.442, ...]

[0,1,0,0,1,0,1,0,1,0,1,0,1,0,0,…]

Modèle d’apprentissage

Catégorie - score

Chien - 0.8 – (True)

Croissant - 0.2 – (False)

$$\begin{bmatrix} 0.8 \\ 0.2 \end{bmatrix}$$

Les + : Une plus grande liberté algorithmique, un vrai catalogue, suffit de faire son choix
Les - : Faire son choix, parfois comprendre le fonctionnement de l’algo

Principales notions/outils associés

Evaluation

Principales notions/outils associés

- Accuracy, Recall, F1-score

Accuracy

$$\frac{Nb\ Correct}{Nb\ Total} = \frac{TP + TN}{All}$$

Very Bad One

Recall (TPR)

$$\frac{Nb\ Normal\ Correct}{Nb\ Normal\ OK} = \frac{TP}{TP + FN}$$

Mouai

F1

$$\frac{2TP}{2TP + FP + FN}$$

Nice

200 images		Classe estimé	
		Chien	Croiss.
Classe réelle	Chien (150)	TP 140	FN 10
	Croiss. (50)	FP 20	TN 30

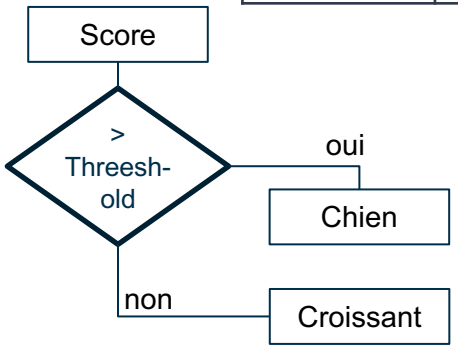
Principales notions/outils associés

- Threeshold

200 images		Classe estimé	
		Chien	Croiss.
Classe réelle	Chien (150)	TP 140	FN 10
	Croiss. (50)	FP 30	TN 20

Chien min score : 0.5

(sinon croissant)



200 images		Classe estimé	
		Chien	Croiss.
Classe réelle	Chien (150)	TP 130 ↓	FN 20 ↑
	Croiss. (50)	FP 40 ↑	TN 10 ↓

Chien min score : 0.6

(sinon croissant)

Principales notions/outils associés

- AUCROC (un truc compliqué mais utile)

200 images		Classe estimé	
		Chien	Croiss.
Classe réelle	Chien (150)	144	6
	Croiss. (50)	48	2

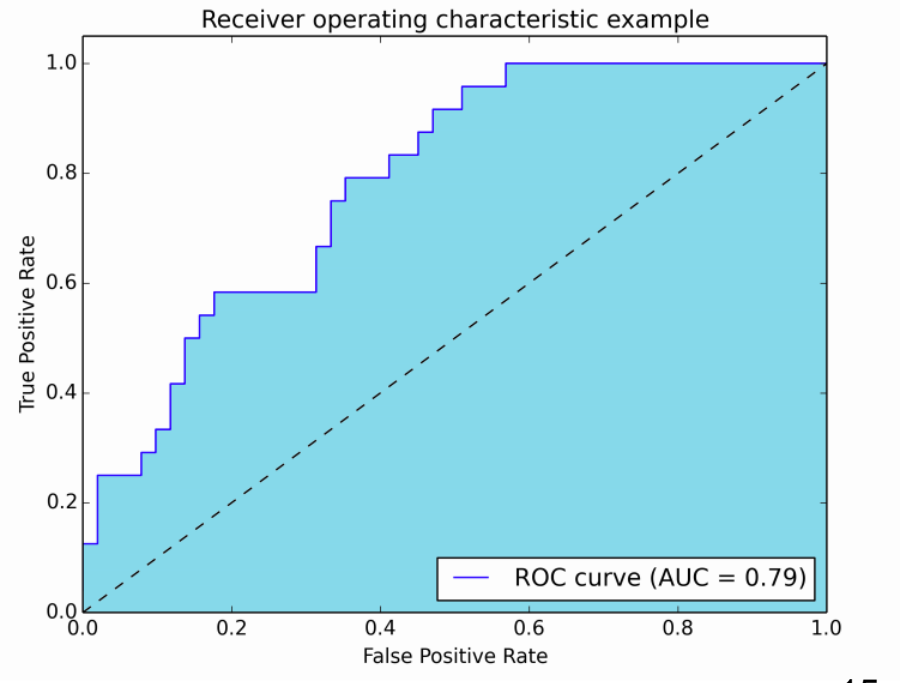
Threeshold : 0.1

200 images		Classe estimé	
		Chien	Croiss.
Classe réelle	Chien (150)	130	20
	Croiss. (50)	40	10

Threeshold : 0.6

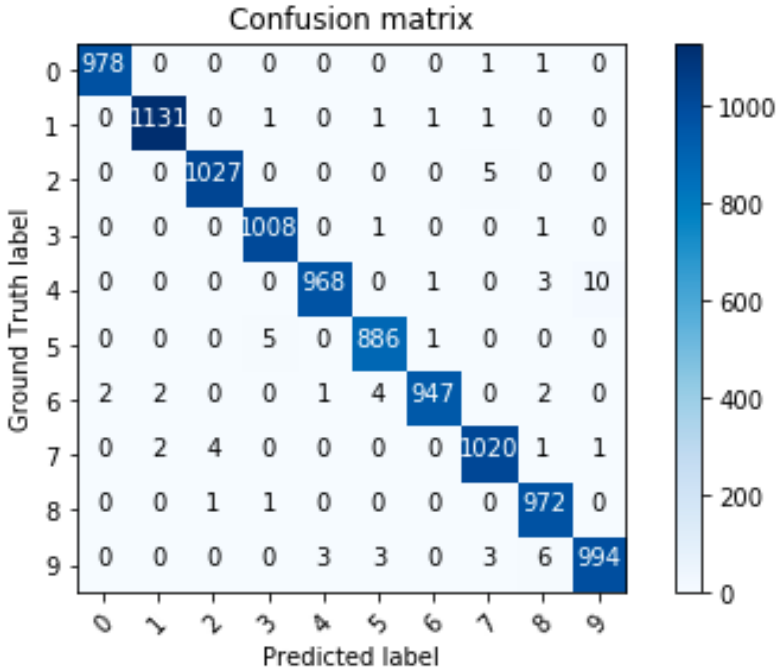
200 images		Classe estimé	
		Chien	Croiss.
Classe réelle	Chien (150)	10	140
	Croiss. (50)	2	48

Threeshold : 0.9



Principales notions/outils associés

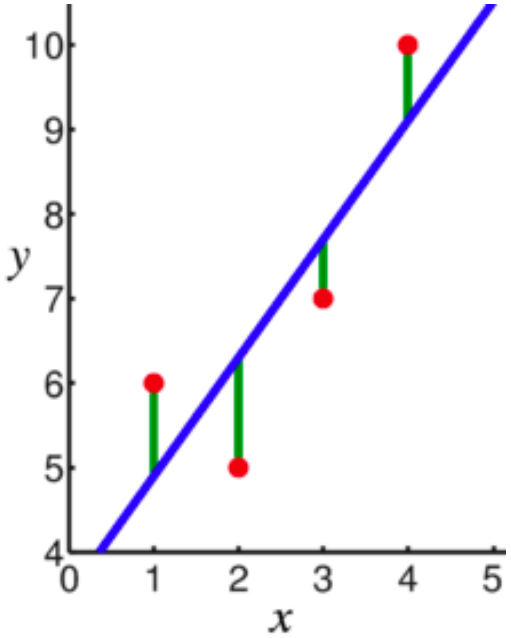
- Confusion Matrix



- Détection des catégories en difficulté par rapport aux autres
- Autres metrics (1 AUCROC/classe)

Principales notions/outils associés

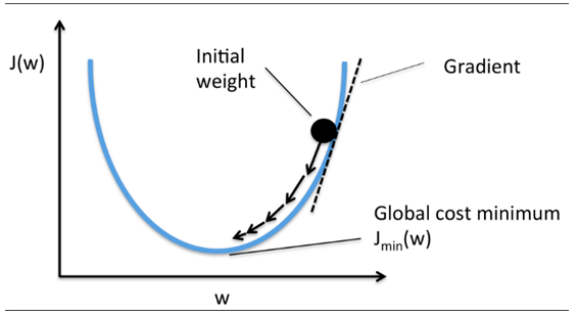
- Loss Function



$$J = Const \times \sum (vrai\ valeur\ - prediction)^2$$

Quadratic Loss Function

- L'objectif d'un algo utilisant une Loss Function est de minimiser cette dernière.

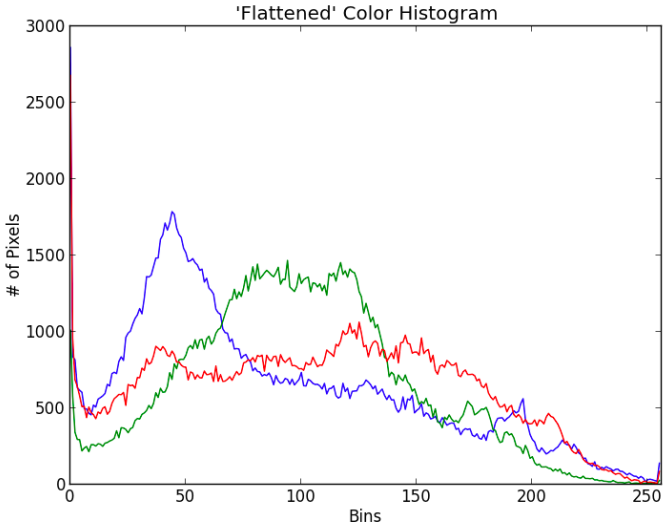


Entrainement

Principales notions/outils associés

- Différenciation des paramètres

Histogramme couleur/pixel



Mouai :/

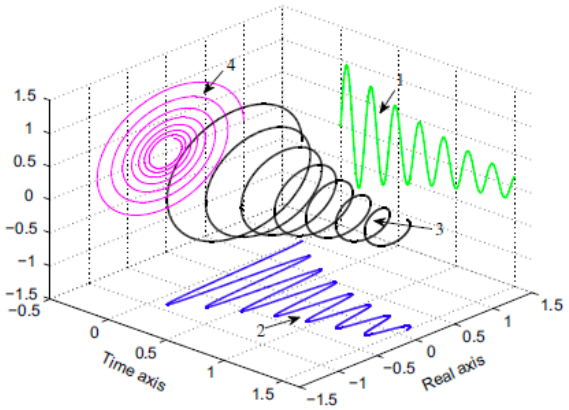
Extraction de points d'interêt

Cool :D

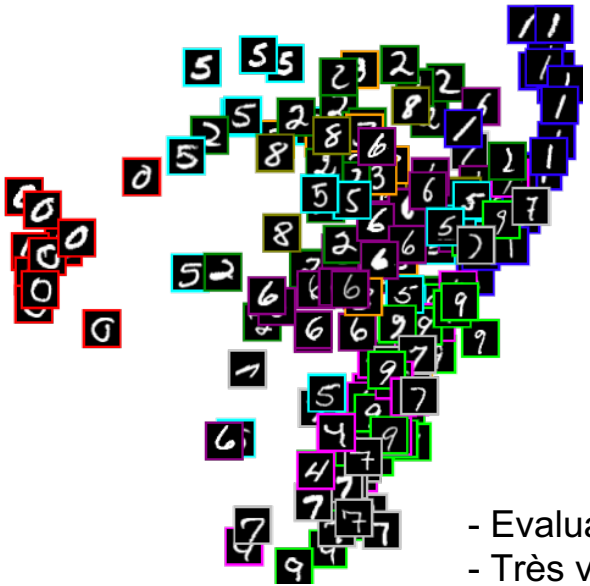
Visualisation

Principales notions/outils associés

- PCA (Principal Components Analysis)



Projection n dimensions
vers graph 2D/3D



- Evaluation du model,
- Très visuel => cool pour des non initiés (ton boss)

i dimension -> j dimensions => Tri des paramètres/features les plus « important(e)s »

Comment construire mon modèle ?

Méthode

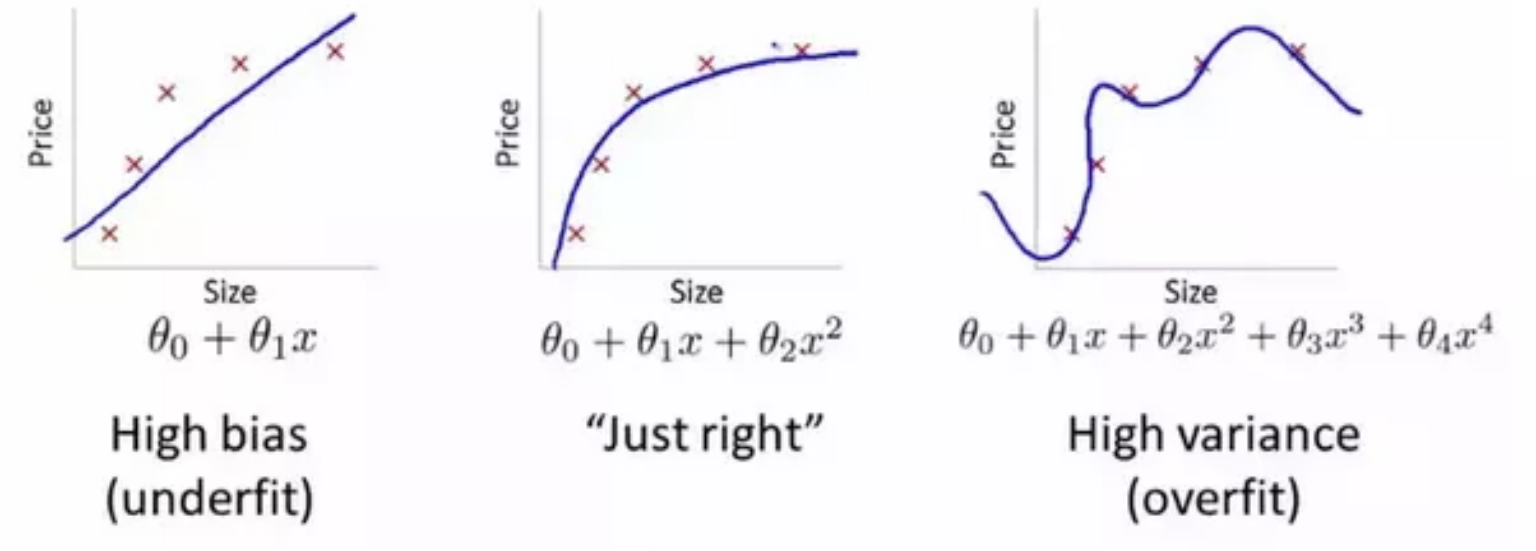
Les prérequis d'un bon modèle

- Visualiser la donnée

	2	3	4	5	6	7	8	9	10	11	...	257
Data Source	Country Name	Aruba	Andorra	Afghanistan	Angola	Albania	Arab World	United Arab Emirates	Argentina	Armenia	...	Vietnam
Unnamed: 4	1960	NaN	NaN	0.0460679	0.104357	1.25819	0.643654	0.118786	2.36747	NaN	...	0.215631
Unnamed: 5	1961	NaN	NaN	0.0536149	0.0847184	1.37419	0.68524	0.108937	2.44262	NaN	...	0.225435
Unnamed: 6	1962	NaN	NaN	0.0737813	0.216025	1.43996	0.760996	0.163355	2.52239	NaN	...	0.25876
Unnamed: 7	1963	NaN	NaN	0.0742514	0.206877	1.18168	0.875116	0.175712	2.31636	NaN	...	0.247579
Unnamed: 8	1964	NaN	NaN	0.0863165	0.216174	1.11174	0.999349	0.132651	2.53838	NaN	...	0.314058
Unnamed: 9	1965	NaN	NaN	0.101499	0.206089	1.1661	1.16617	0.14637	2.64171	NaN	...	0.343354
Unnamed: 10	1966	NaN	NaN	0.107674	0.265164	1.33306	1.2735	0.159359	2.79265	NaN	...	0.490532

Les prérequis d'un bon modèle - préparation

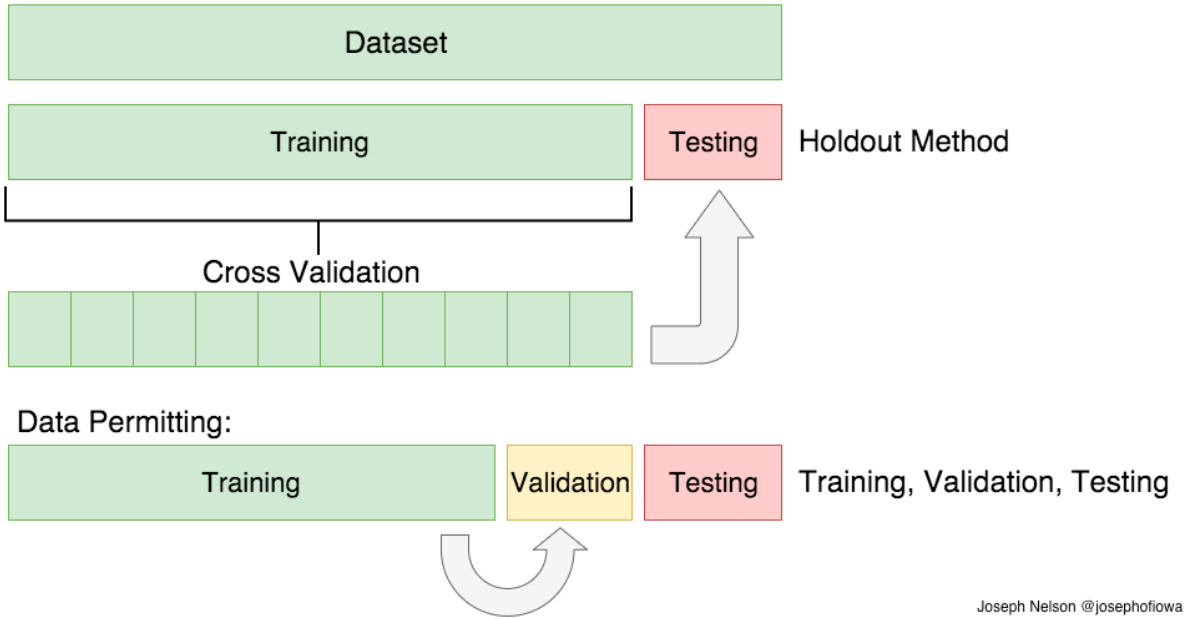
- Underfitting & Overfitting, choix des paramètres



complexity, number of params

Les prérequis d'un bon modèle - préparation

- Cross Validation



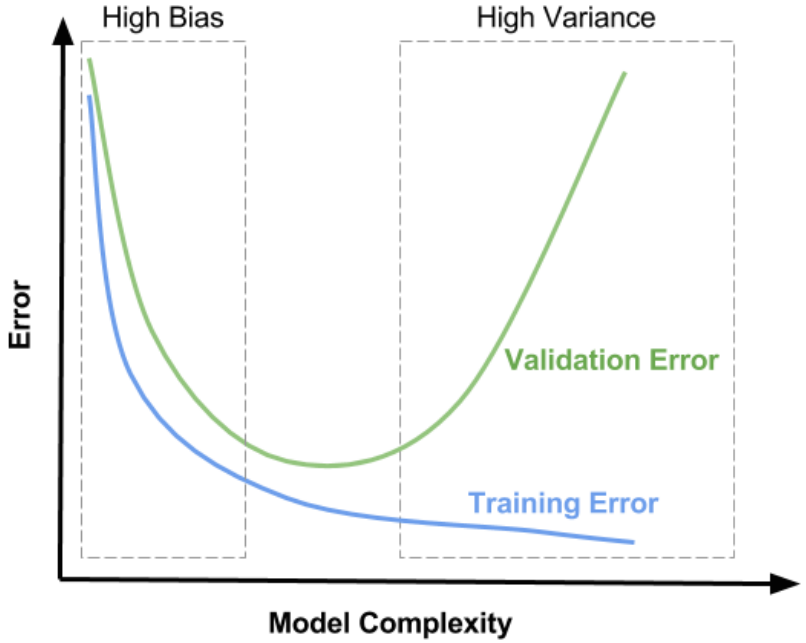
Les prérequis d’un bon modèle - modèle

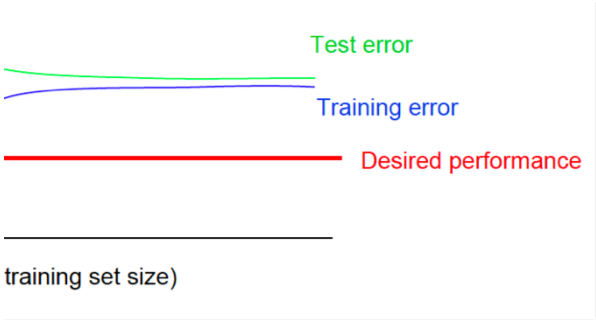
- Choix du modèle

Pas forcément de bon modèle au départ
Il faut en tester plusieurs

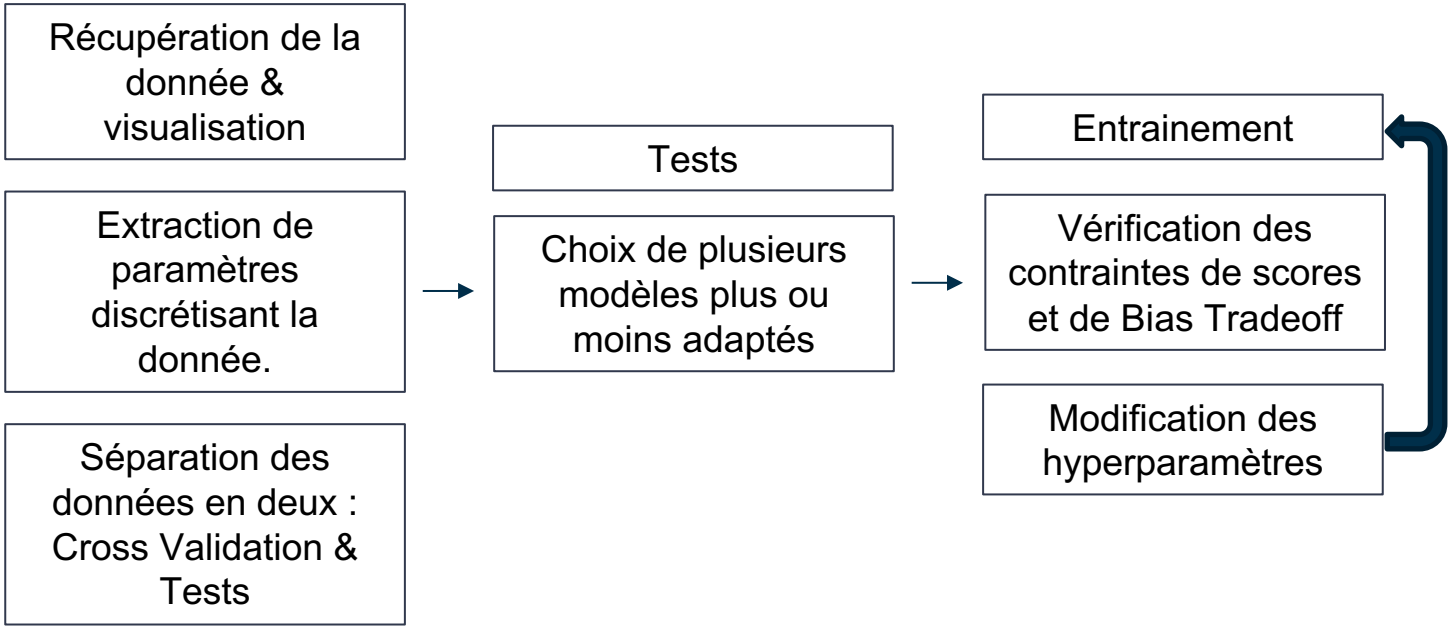
Les prérequis d'un bon modèle - amélioration

- Bias & Variance tradeoff (en fait c'est pareil que la slide d'avant)





Les prérequis d’un bon modèle

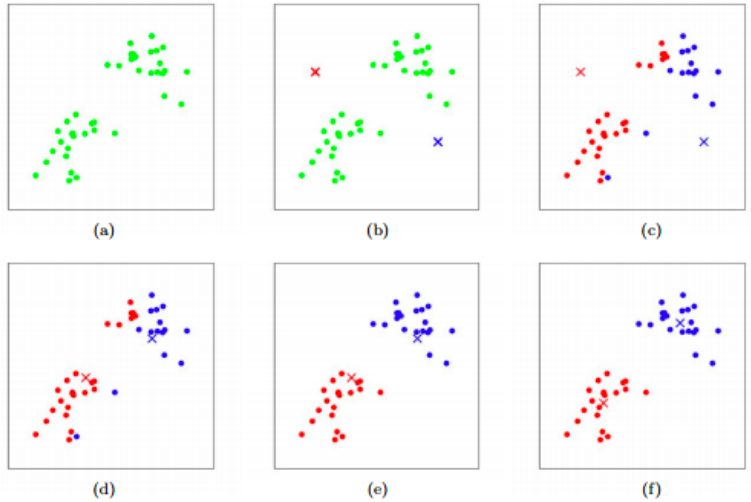


Les principaux modèles d’apprentissage.

Unsupervised

Les principaux modèles d'apprentissage.

- Clustering

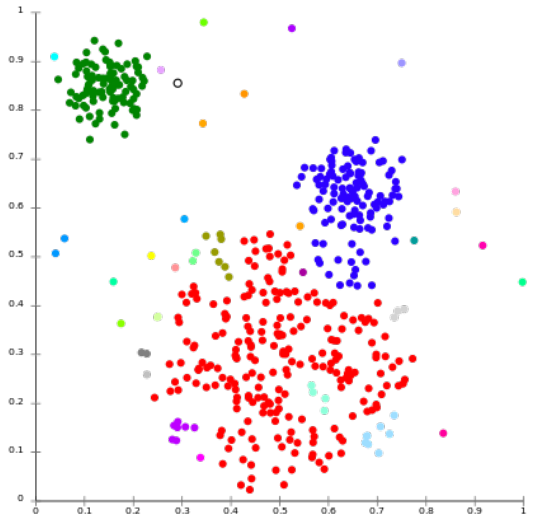


K moyens +
initialisation aléatoire

- Simple à mettre en place
- Nécessite que les points soient différenciable

Les principaux modèles d’apprentissage.

- Anomaly Detection



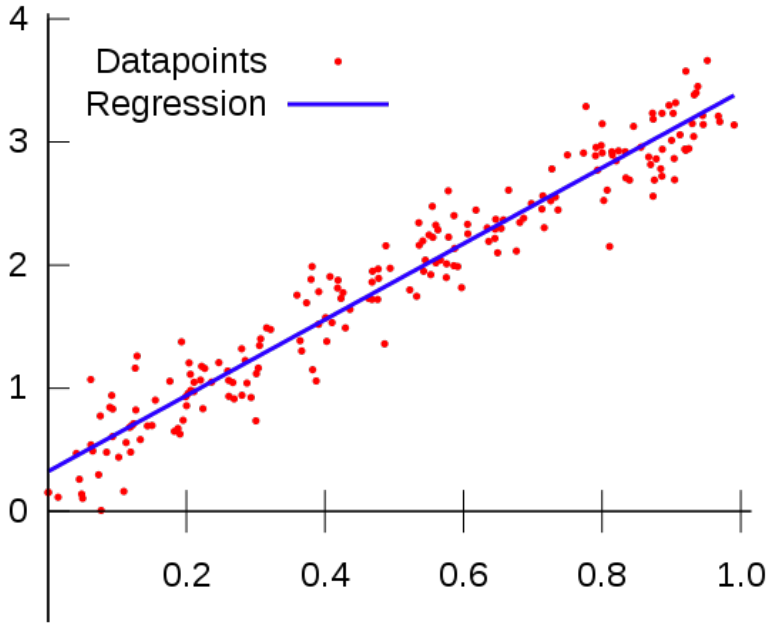
Density based

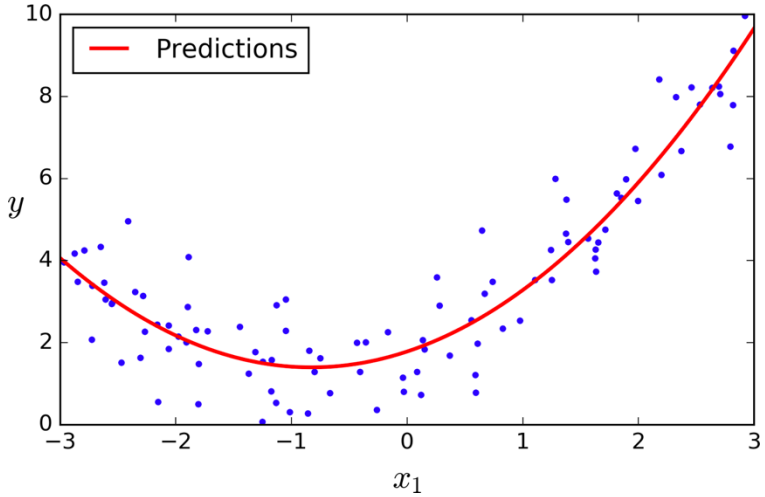
- Simple à mettre en place
- Gestion du risque complexe, faire attention aux contraintes mathématiques associé

Basic ones

Les principaux modèles d'apprentissage.

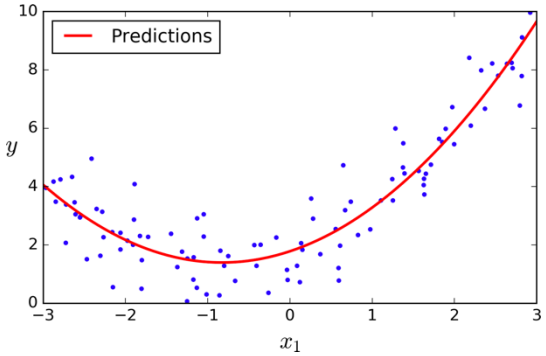
- Régression linéaire & polynomial



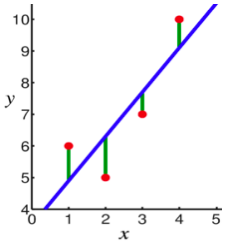


Les principaux modèles d'apprentissage.

- Régression linéaire & polynomial



$$y = ax^2 + bx + c$$



$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} \longrightarrow J = f(a,b,c)$$

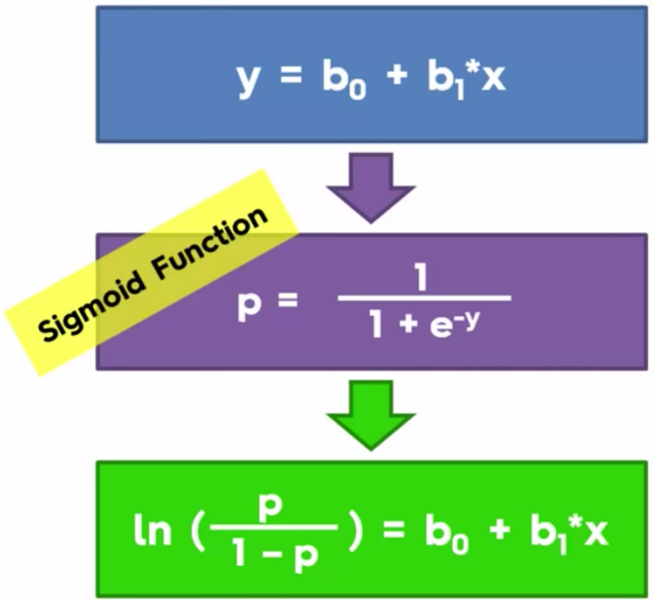
On cherche à minimiser J

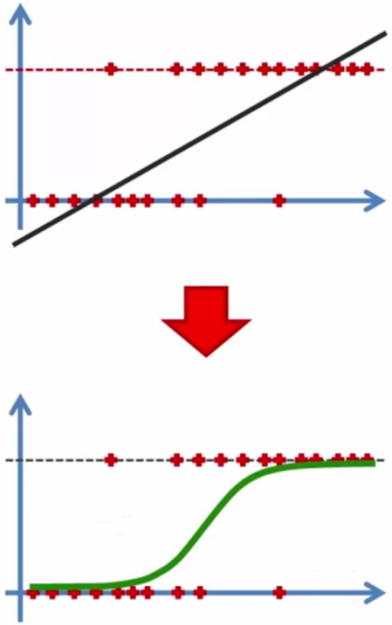
Trouver a,b,c tel que J=0

Utilisation de la méthode des gradient pour trouver ce minimum local

Les principaux modèles d'apprentissage.

- Régression logistique





- Convergence plus stable
- Convergence plus rapide
- Résous quelques problèmes de matheux

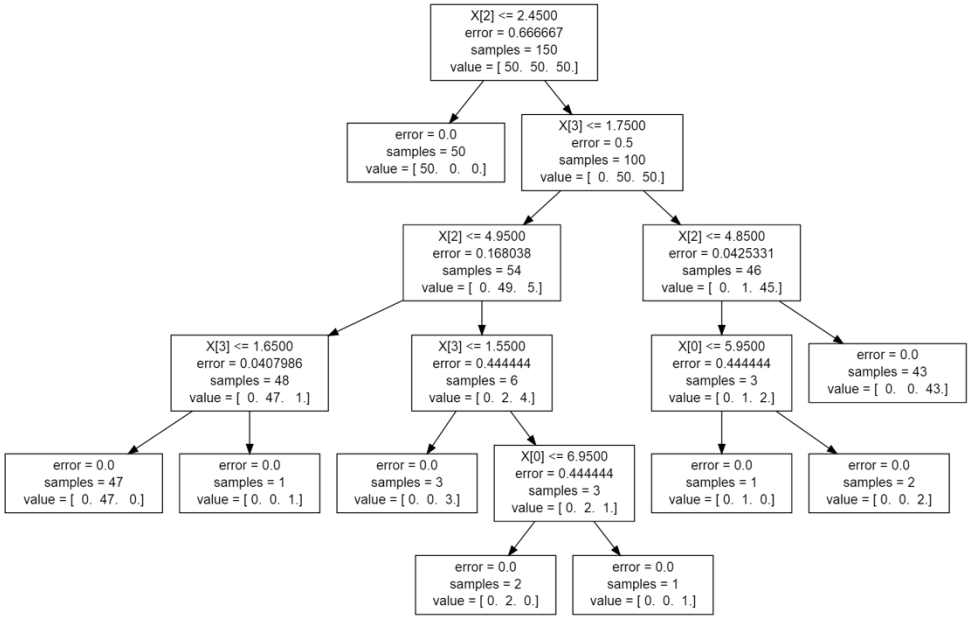
- Base du machine learning d'aujourd'hui (Réseaux de neurones notamment)

Les principaux modèles d’apprentissage.

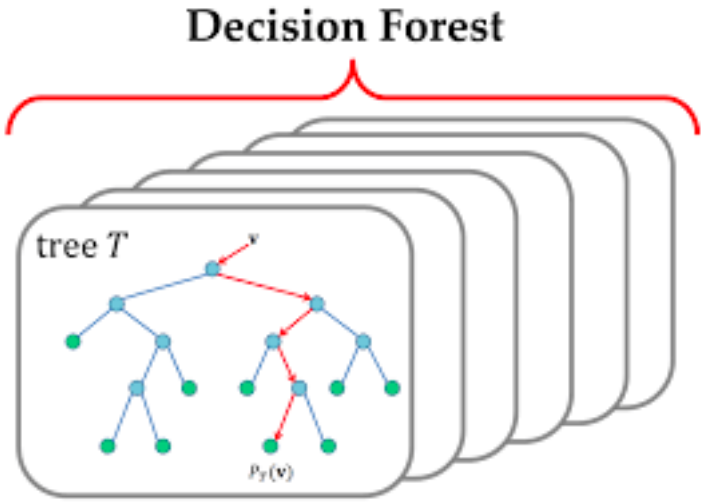
Most used ones

Les principaux modèles d'apprentissage.

- Random Tree Forest (Forêt d'arbres décisionnels)



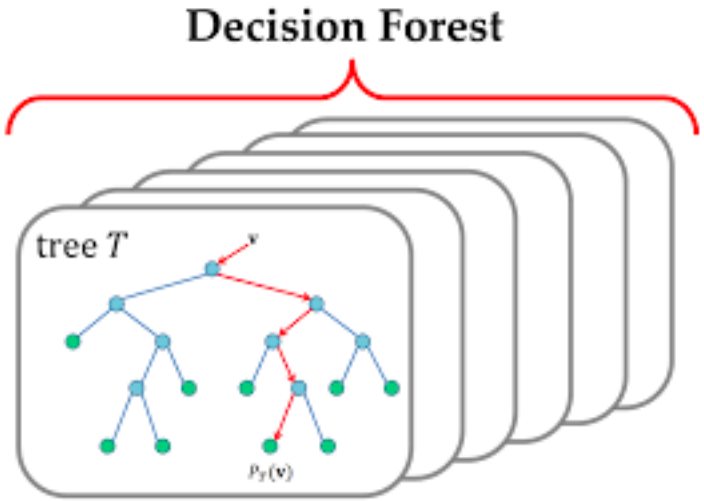
Conditions True / False sur les paramètres



Régressions sur l'ensemble des arbres

Les principaux modèles d’apprentissage.

- Random Tree Forest (Forêt d'arbres décisionnels)

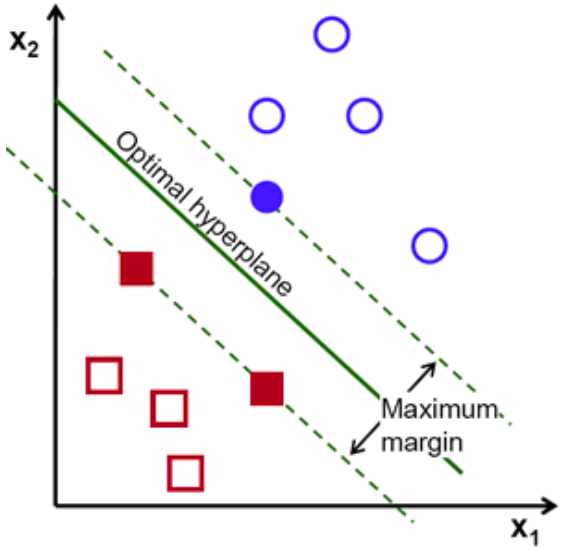


- Simple à mettre en œuvre
- Simple à comprendre
- Très visuel

- Ne permet pas de résoudre efficacement tout les problèmes (notamment les plus complexes)
- Abandonné pour la plupart des problèmes actuels

Les principaux modèles d'apprentissage.

- Suport Vector Machine



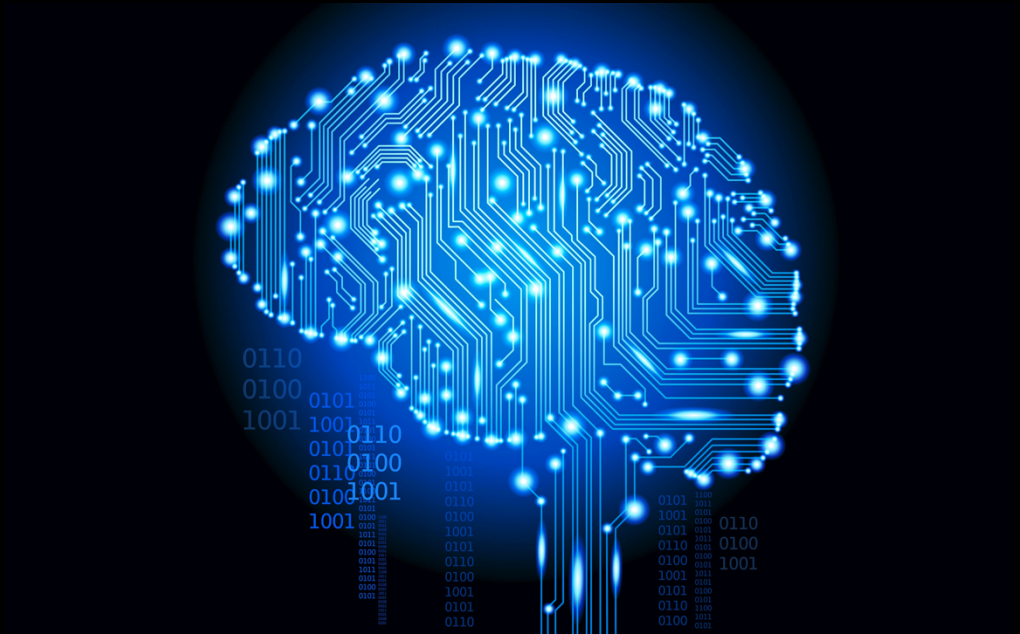
- Très efficace encore aujourd'hui
- Une des meilleur façon de séparer des vecteurs

- Compliqué à optimiser (faut faire des maths)
- Pas forcément hyper visuel

Les principaux modèles d’apprentissage.

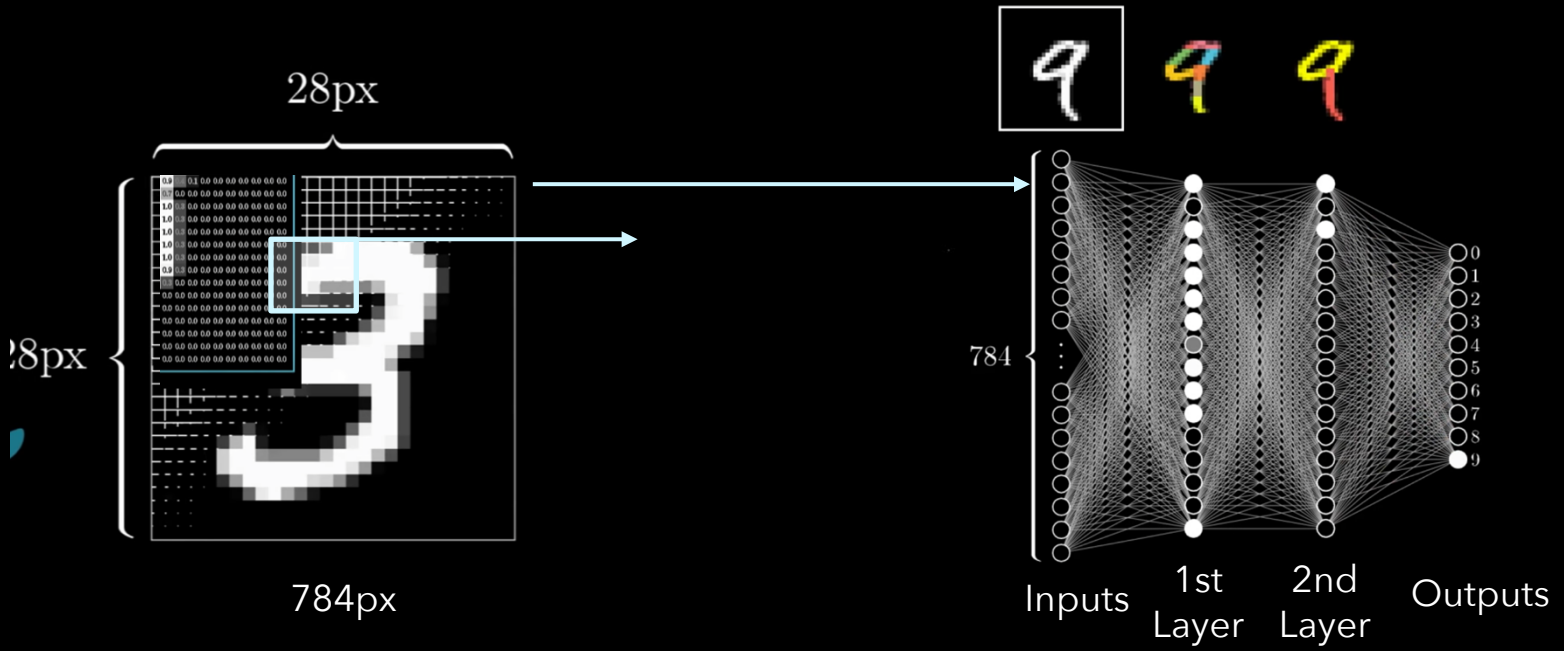
- Réseaux de neurones

But, what is a Neural network ?
https://www.youtube.com/watch?v=aircAruvnKk



Les principaux modèles d'apprentissage.

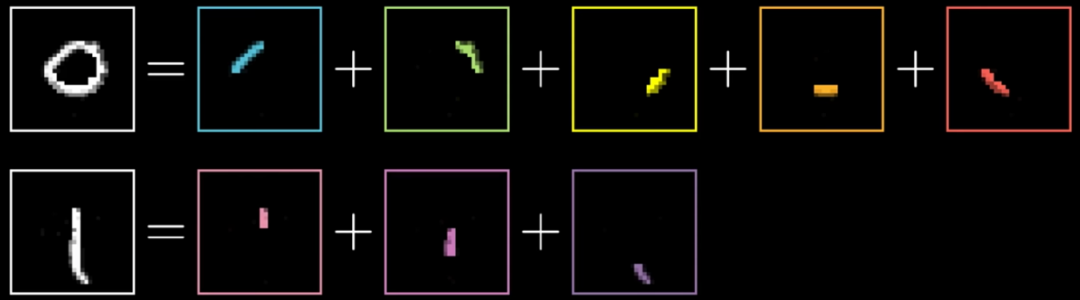
- Réseaux de neurones



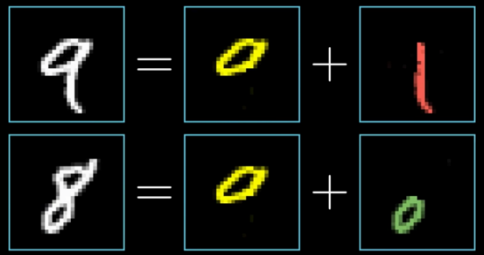
Les principaux modèles d’apprentissage.

- Réseaux de neurones

1st Layer
(precise features)

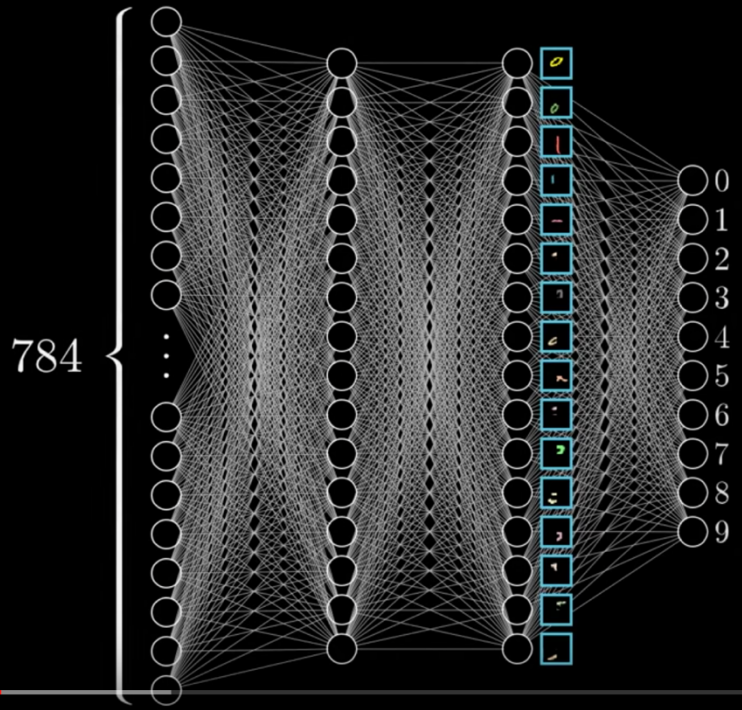


2nd Layer
(more general features)



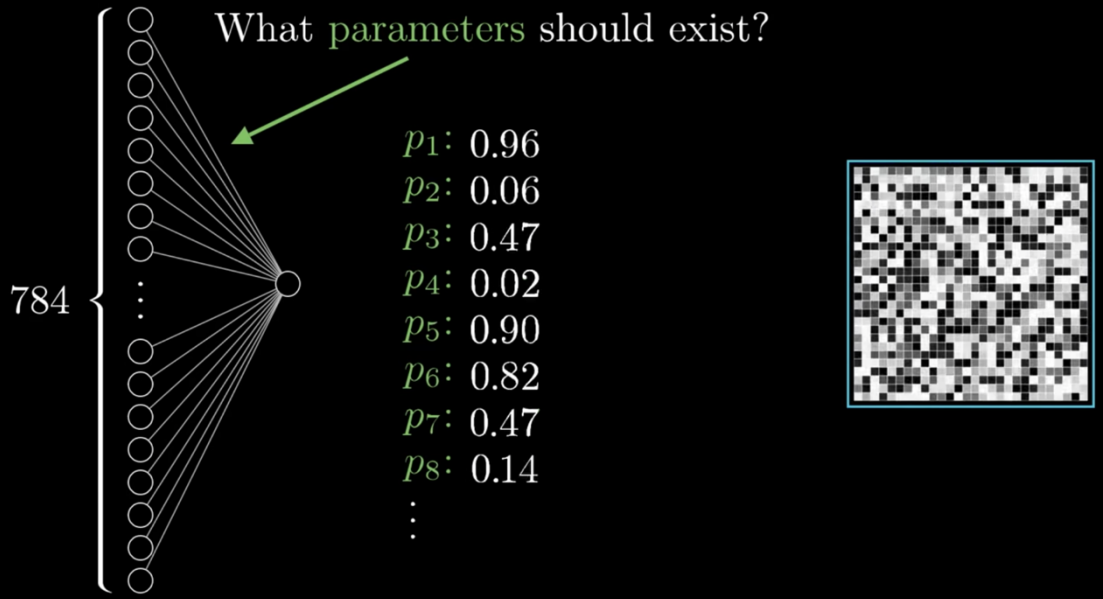
Les principaux modèles d’apprentissage.

- Réseaux de neurones



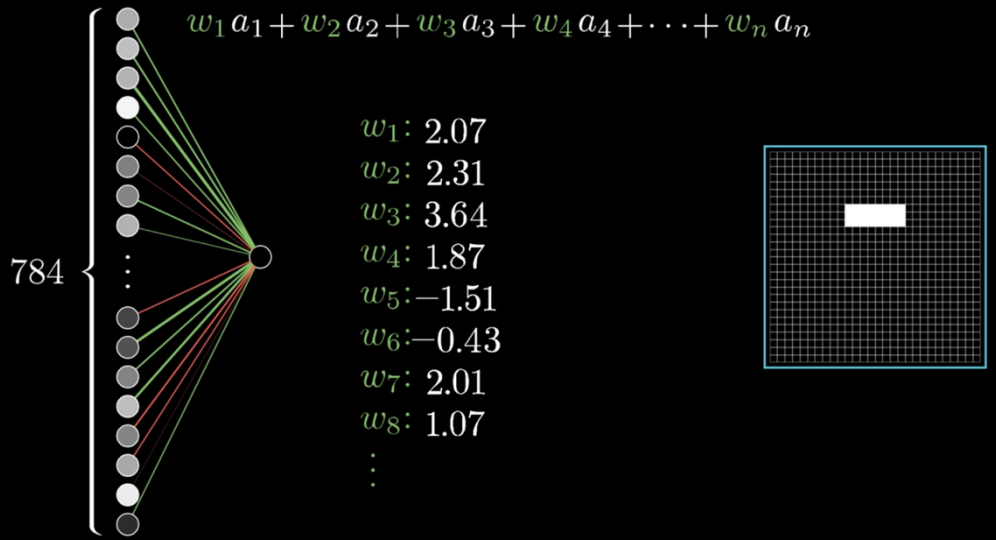
Les principaux modèles d’apprentissage.

- Réseaux de neurones



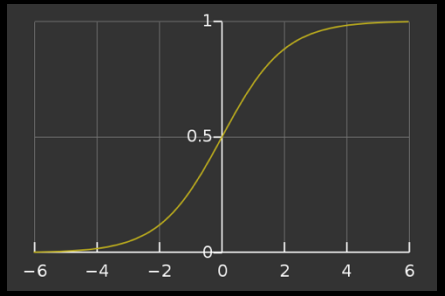
Les principaux modèles d’apprentissage.

- Réseaux de neurones

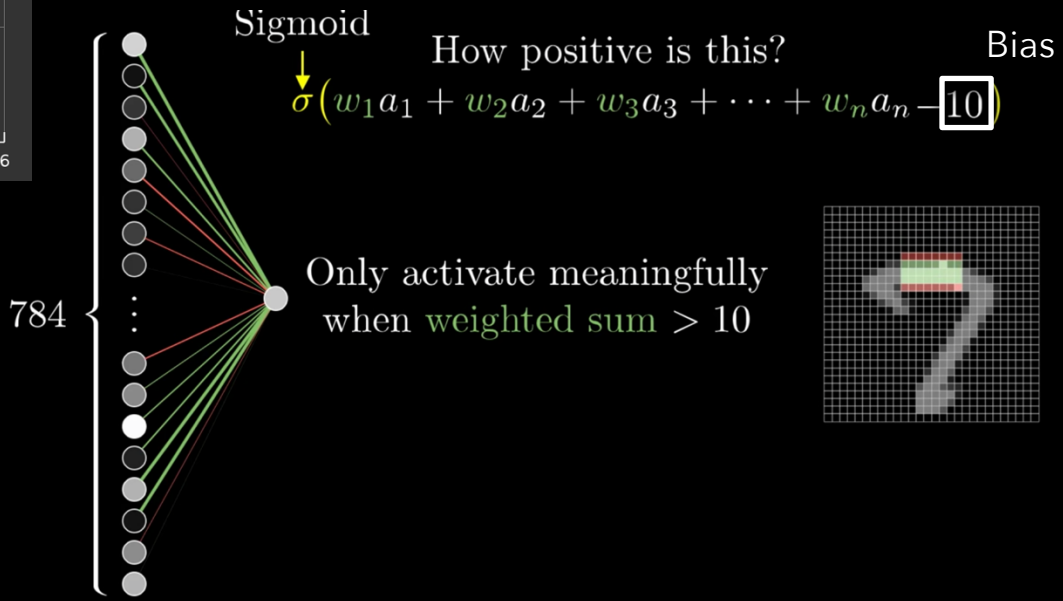


Les principaux modèles d'apprentissage.

- Réseaux de neurones



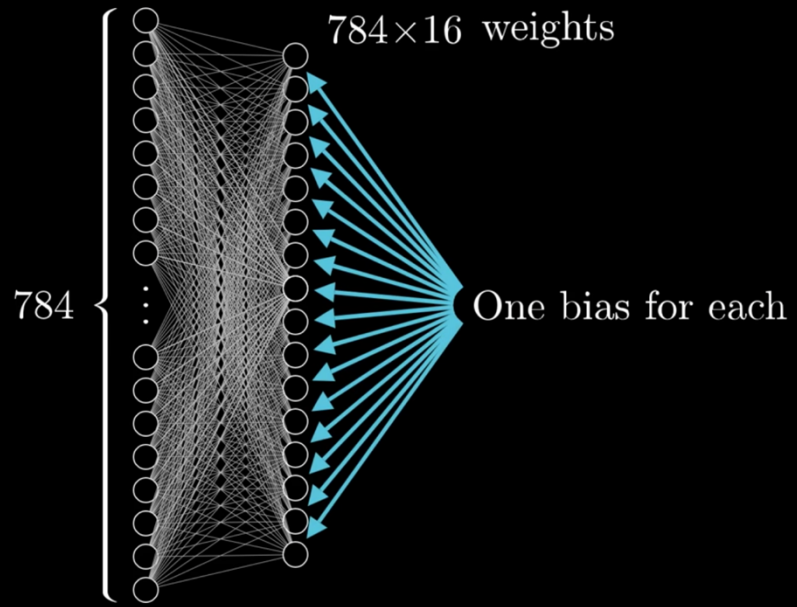
« Activation Function »



Le neurone prend une valeur proche de 1 si et seulement si « assez » de produits w.a sont positif (et proche de 1).

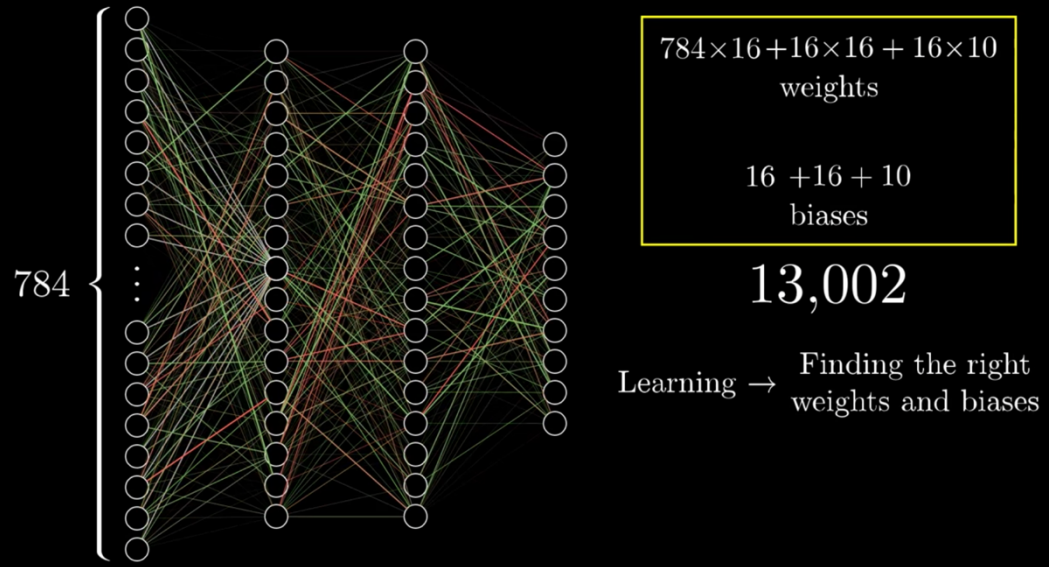
Les principaux modèles d’apprentissage.

- Réseaux de neurones



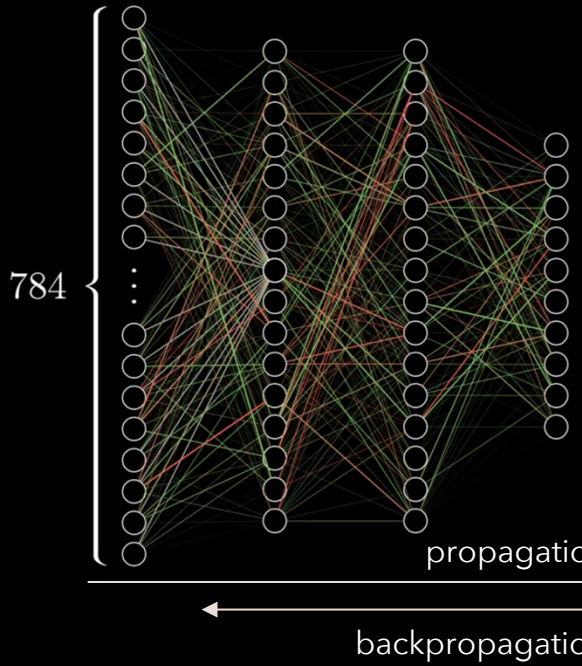
Les principaux modèles d’apprentissage.

- Réseaux de neurones



Les principaux modèles d'apprentissage.

- Réseaux de neurones – Backpropagation

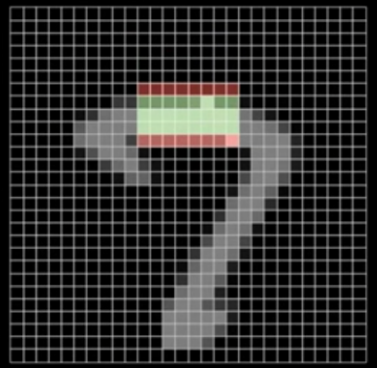


$784 \times 16 + 16 \times 16 + 16 \times 10$
weights

$16 + 16 + 10$
biases

13,002

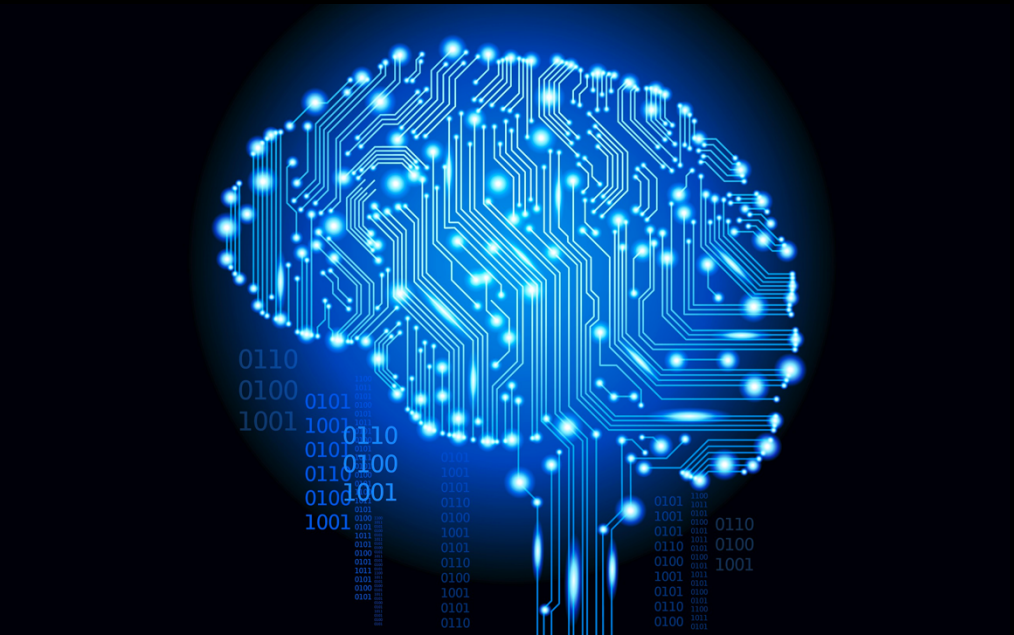
Learning → Finding the right weights and biases



C'est en fait la convergence
de cette cycle qui va créer ces motifs

Les principaux modèles d’apprentissage.

- Réseaux de neurones – Backpropagation

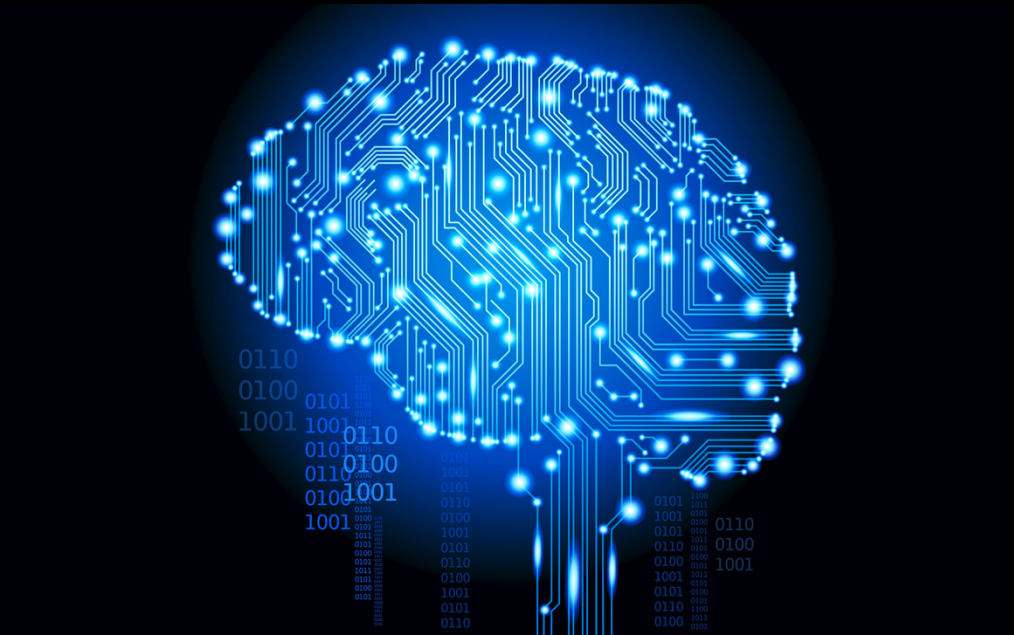


- Peut répondre très efficacement à des problématiques très complexes juste avec de la donnée brut taguée.
- Possibilité d’utiliser des modèles déjà entrainé pour les adapter à sa problématique

- Demande énormément de data (plus le problème, le réseau est complexe, plus il faut de données)
- Difficile de comprendre les choix effectué
- Difficile de visualiser les résultats
- Beaucoup de ressources nécessaires pour améliorer l’état de l’art

Les principaux modèles d’apprentissage.

- Réseaux de neurones – Demo



https://goo.gl/8qwvJf