CATIE

Solutions pour la société numérique

Machine Learning et ses applications

Zhe LI – Ingénieur Informatique@CATIE

Plan

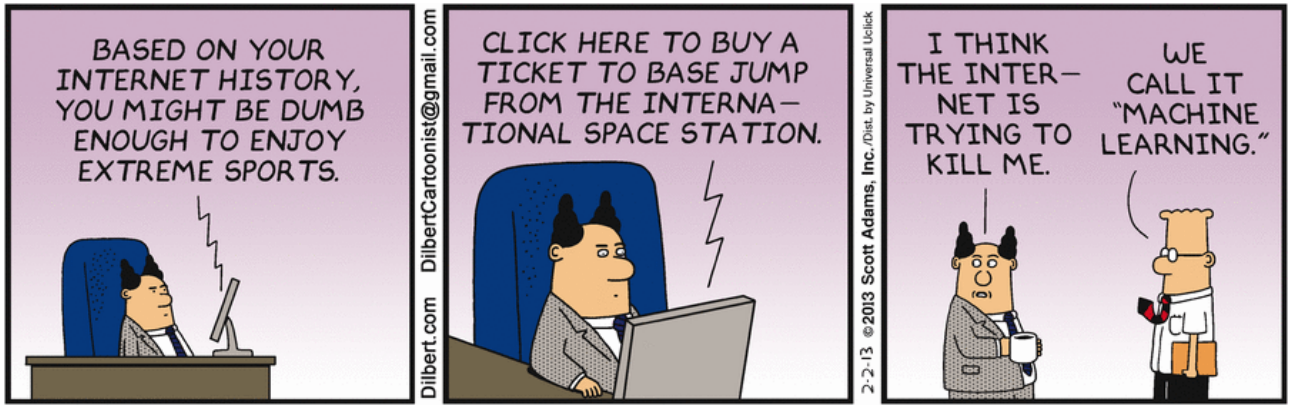
- Machine Learning
- Deep Learning
- Ethique et Machine Learning

Machine Learning : c'est quoi ?

Étudier et construire les programmes qui sont capables de :

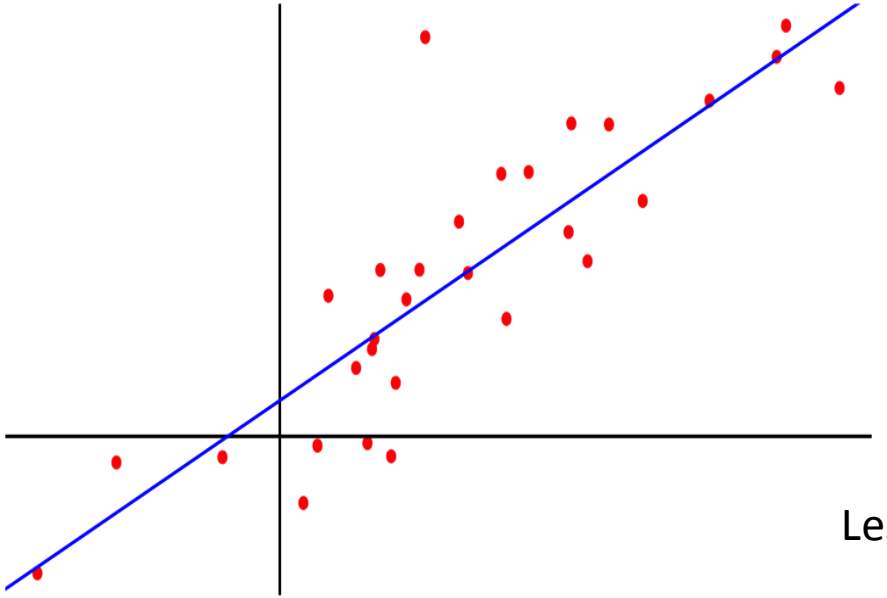
- Comprendre des échantillons de données en s'appuyant sur les techniques statistiques
- Déterminer la meilleure manière de modéliser les données
- Donner certaines prédictions pour les nouvelles données

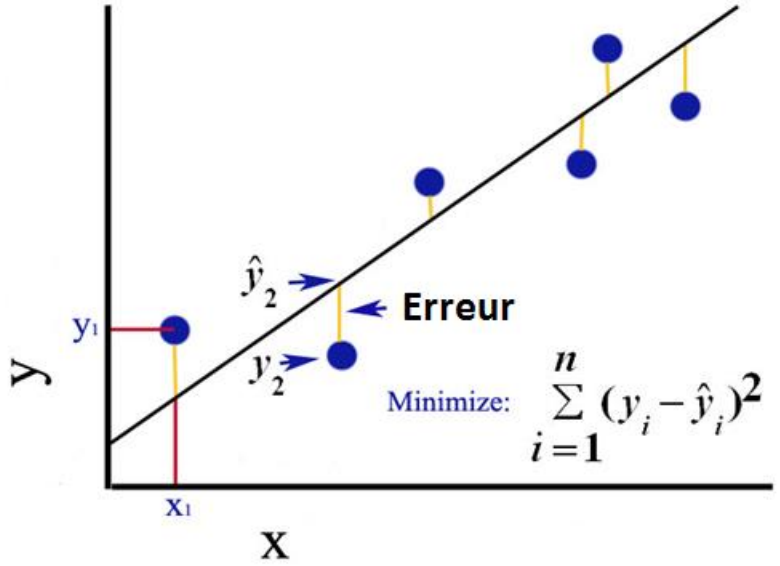
} Expérience



Apprentissage

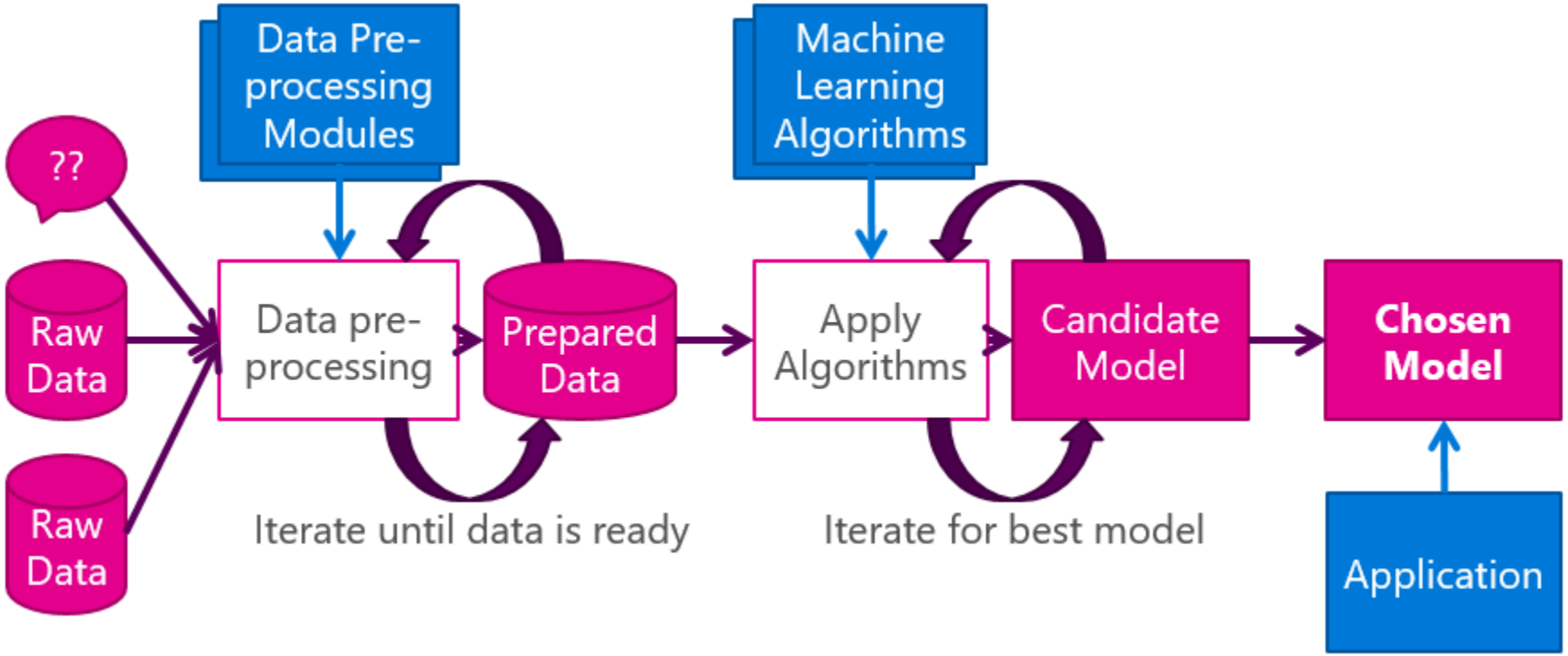
Pour un modèle donné, choisir les meilleurs paramètres possibles pour décrire les données





Les données sont modélisées : y = ax + b

Machine Learning Pipeline



https://blogs.msdn.microsoft.com/martinkearn/2016/03/01/machine-learning-is-for-muggles-too/

Préparer les données

C'est l'étape la plus importante, qui consomme beaucoup de temps

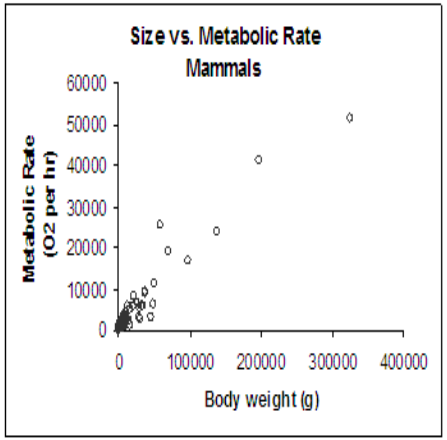
- Nettoyage des données
- Transformation des données
  - Données numériques
  - Données textuelles

https://www.linkedin.com/pulse/big-data-101-cleaning-script-fennel-aurora

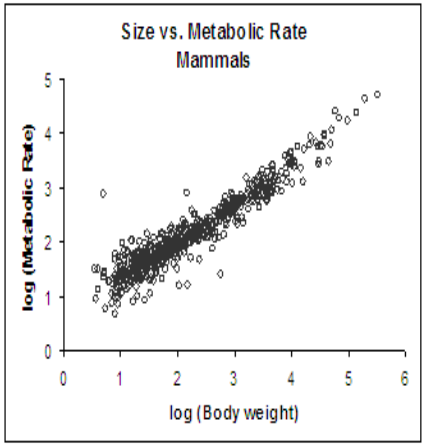
Transformation des données - Exemple

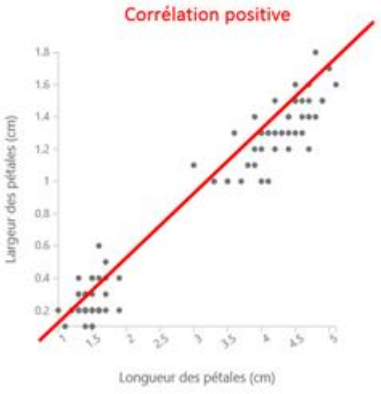
CATIE
Solutions pour la société numérique

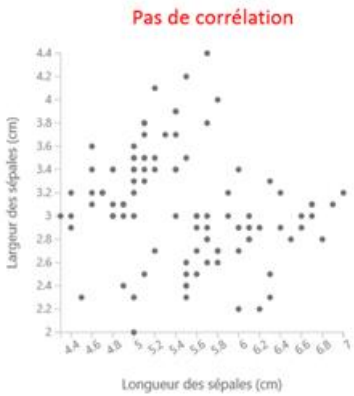
Longueur des sépales (cm)	Largeur des sépales (cm)	Longueur des pétales (cm)	Largeur des pétales (cm)	Prédiction
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5	3.6	1.4	0.2	Iris-setosa



Log
Trans-
form







Type d'apprentissage

CATIE
Solutions pour la société numérique

- L'apprentissage supervisé
- L'apprentissage non-supervisé
- L'apprentissage par renforcement

Apprentissage supervisé

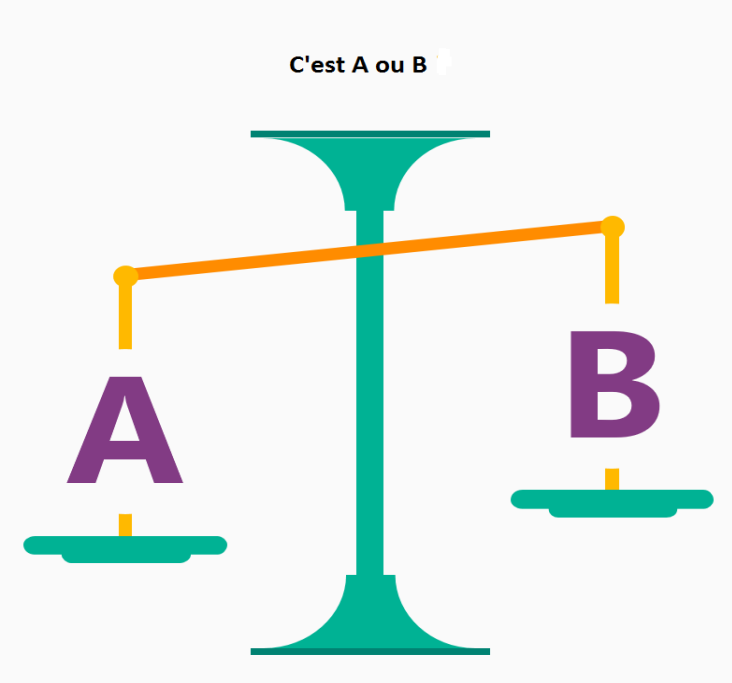
CATIE
Solutions pour la société numérique

- Les données sont déjà « labélisées »
- L'algorithme permet de trouver la loi déterminant l'output en fonction des inputs
- Random Forest, KNN, SVM, Régression linéaire, Régression Logistique…

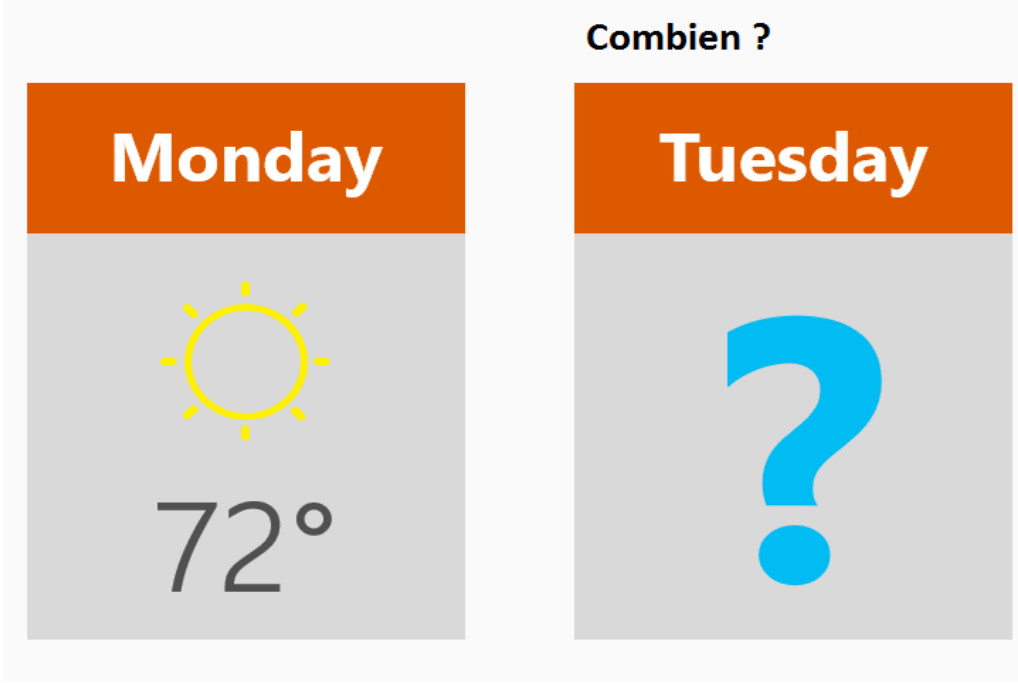
Apprentissage supervisé

CATIE
Solutions pour la société numérique

Classification



Régression



https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-for-beginners-the-5-questions-data-science-answers

Apprentissage non-supervisé

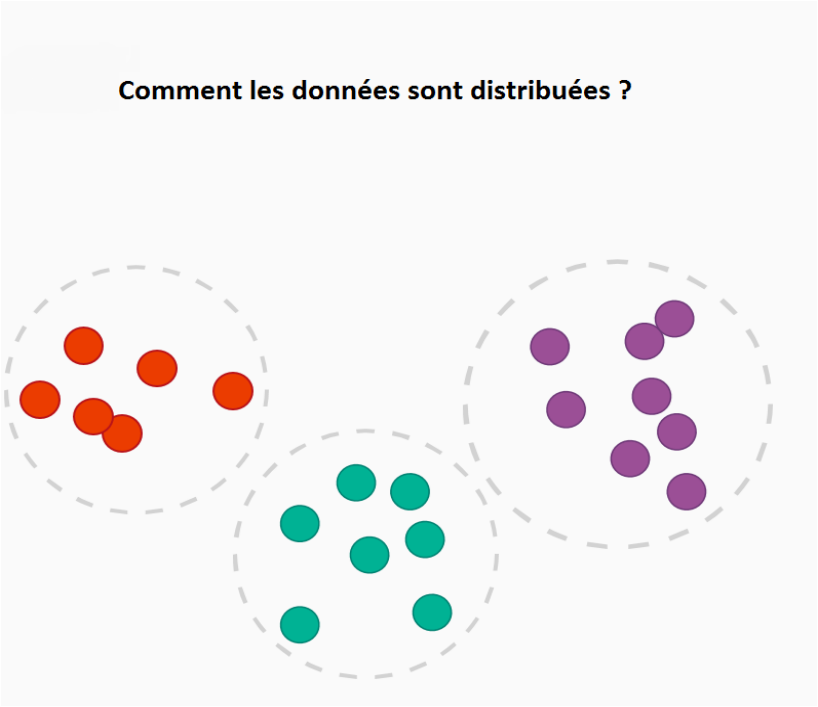
CATIE
Solutions pour la société numérique

- Aucun label n'est fourni à l'algorithme
- Il doit découvrir la structure caractéristique de l'input sans assistance humaine

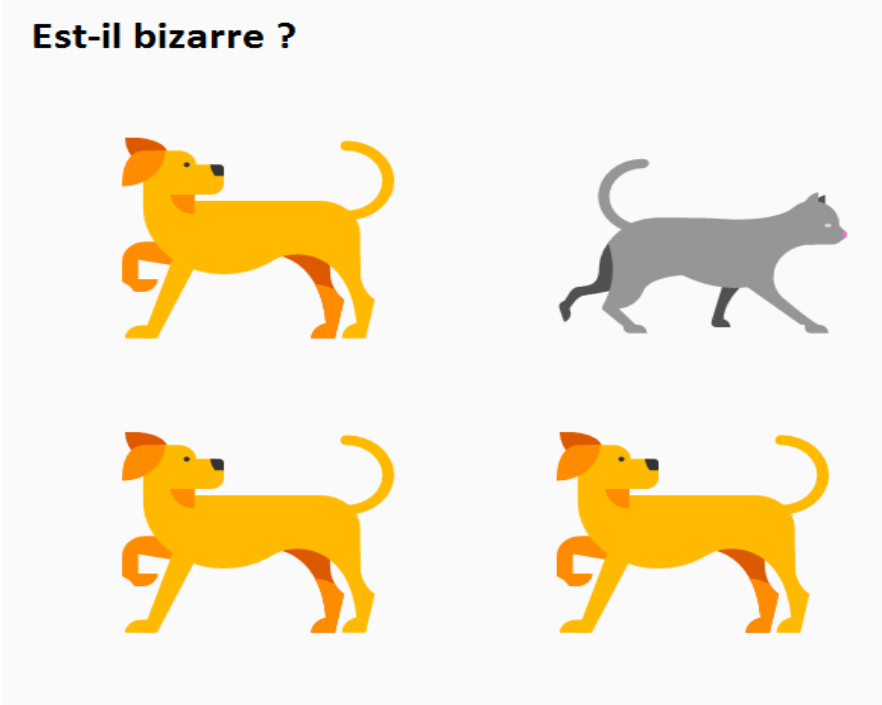
Apprentissage non-supervisé

CATIE
Solutions pour la société numérique

Clustering (K-Means, DBScan…)



Détection d'anomalie



https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-for-beginners-the-5-questions-data-science-answers

Apprentissage semi-supervisé

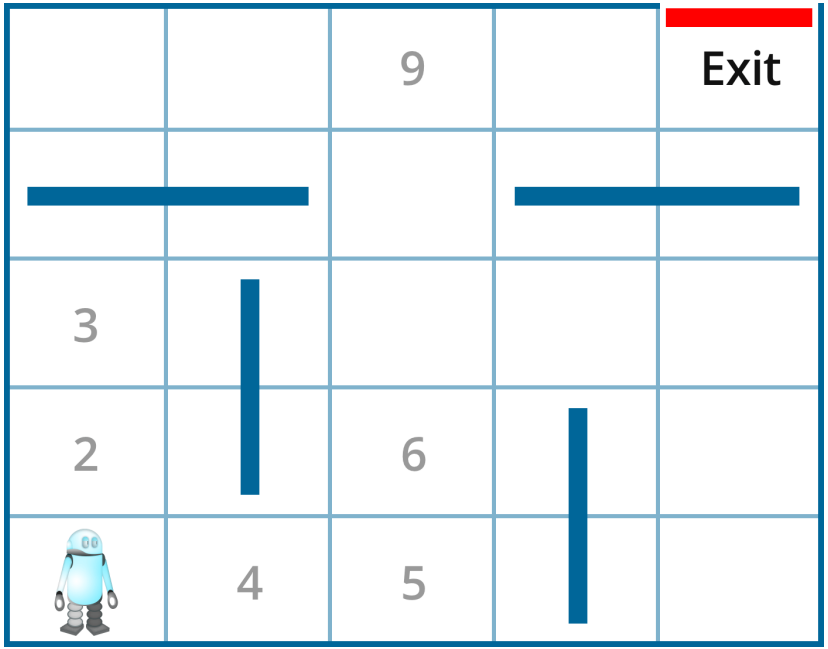
CATIE
Solutions pour la société numérique

- Utiliser un ensemble de données étiquetées et non-étiquetées

- Un intérêt provient du fait que l'étiquetage de données nécessite l'intervention d'un utilisateur humain. Lorsque les jeux de données deviennent très grands, cette opération peut s'avérer fastidieuse

Apprentissage par renforcement

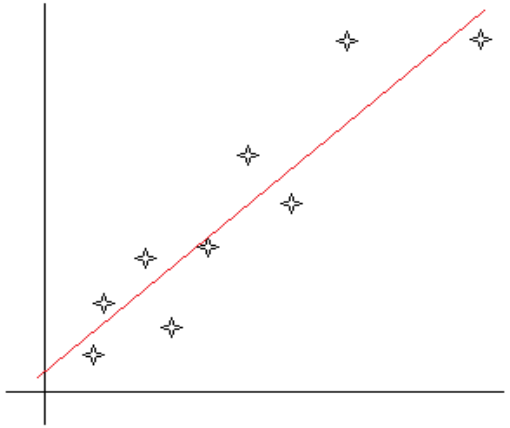
- Considérer un agent autonome, plongé au sein d'un environnement, et qui doit prendre des décisions en fonction de son état courant. En retour, l'environnement procure à l'agent une récompense, qui peut être positive ou négative



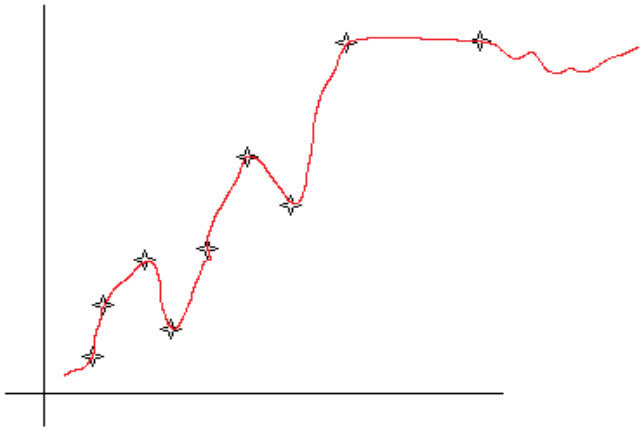
https://www.oreilly.com/ideas/reinforcement-learning-explained

Sur-apprentissage

Le modèle se comporte alors comme une table contenant tous les échantillons utilisés lors de l'apprentissage et perd ses pouvoirs de prédiction sur de nouveaux échantillons

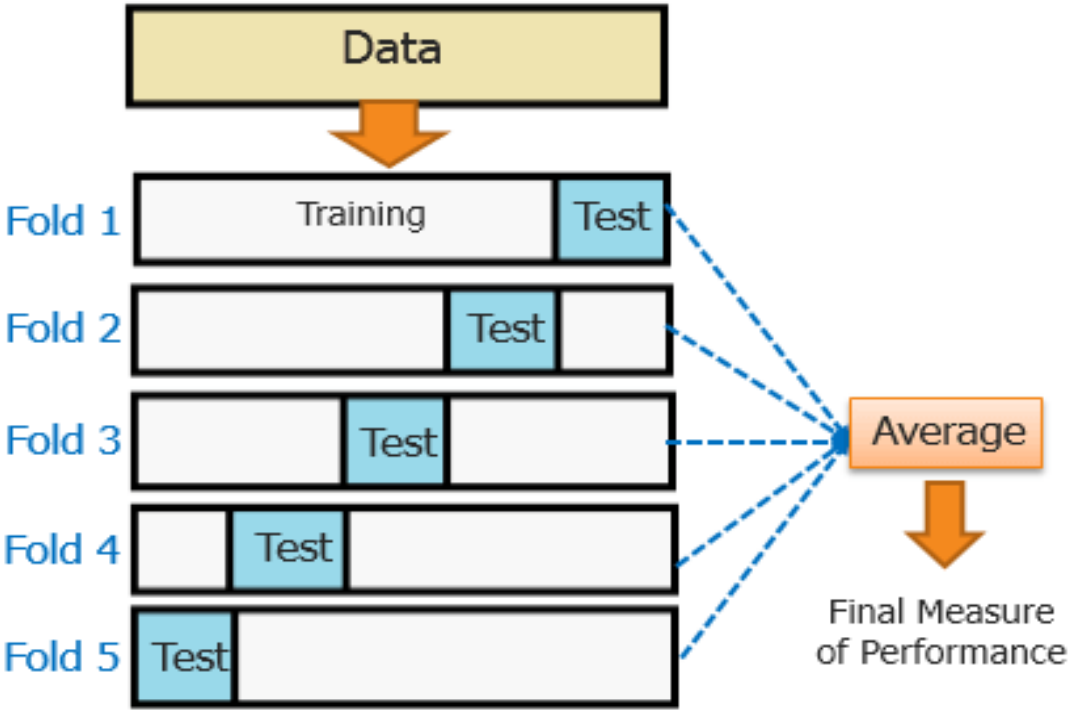


Apprentissage correct



Sur-apprentissage

Cross validation



Evaluation de la performance de l'algorithme

CATIE
Solutions pour la société numérique

Root Mean Squared Error (RMSE) :

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Confusion matrix:

n=165	**Predicted: NO**	**Predicted: YES**	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

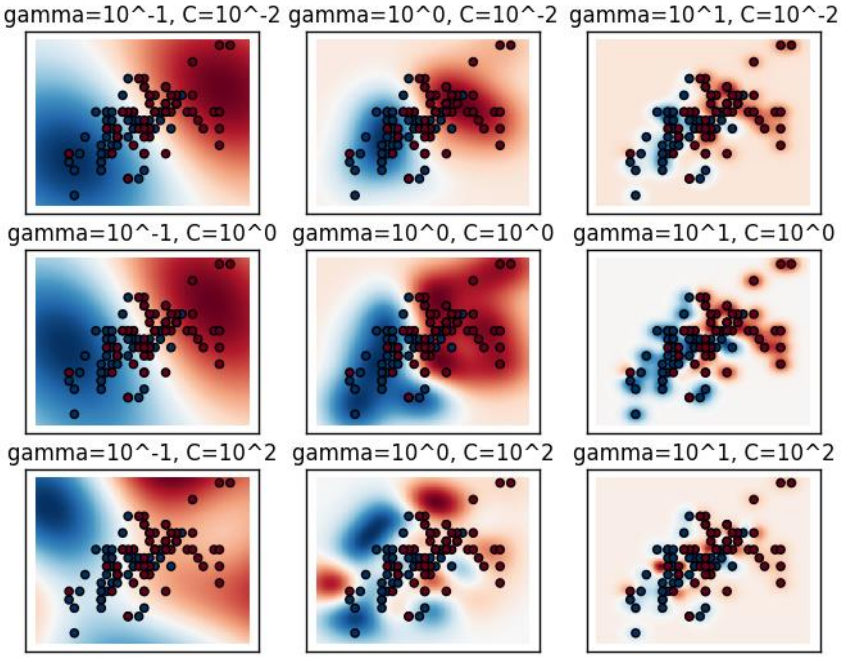
Silhouette, AIC, BIC ….

Trouver les meilleurs paramètres de l'algorithme

En général, un algorithme possède plusieurs paramètres (hyperparameters)

Changer ces paramètres pourrait influencer le résultat

- GridSearch
- RandomSearch

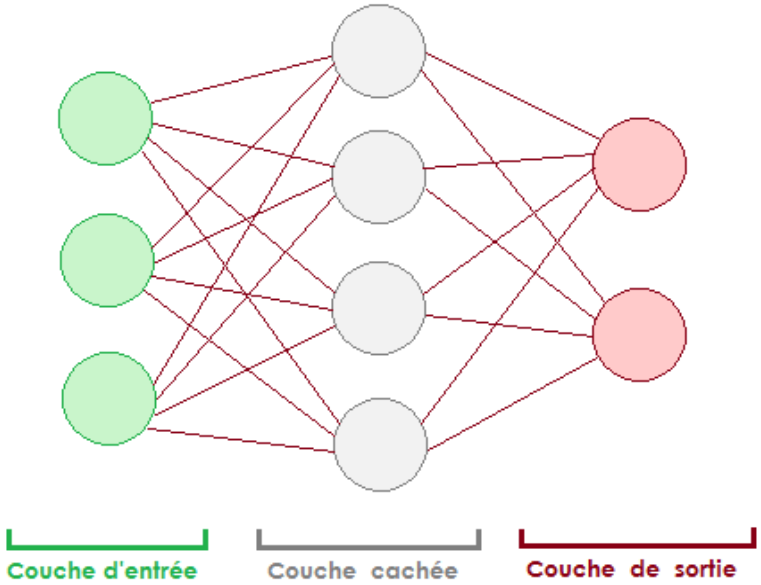


http://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html

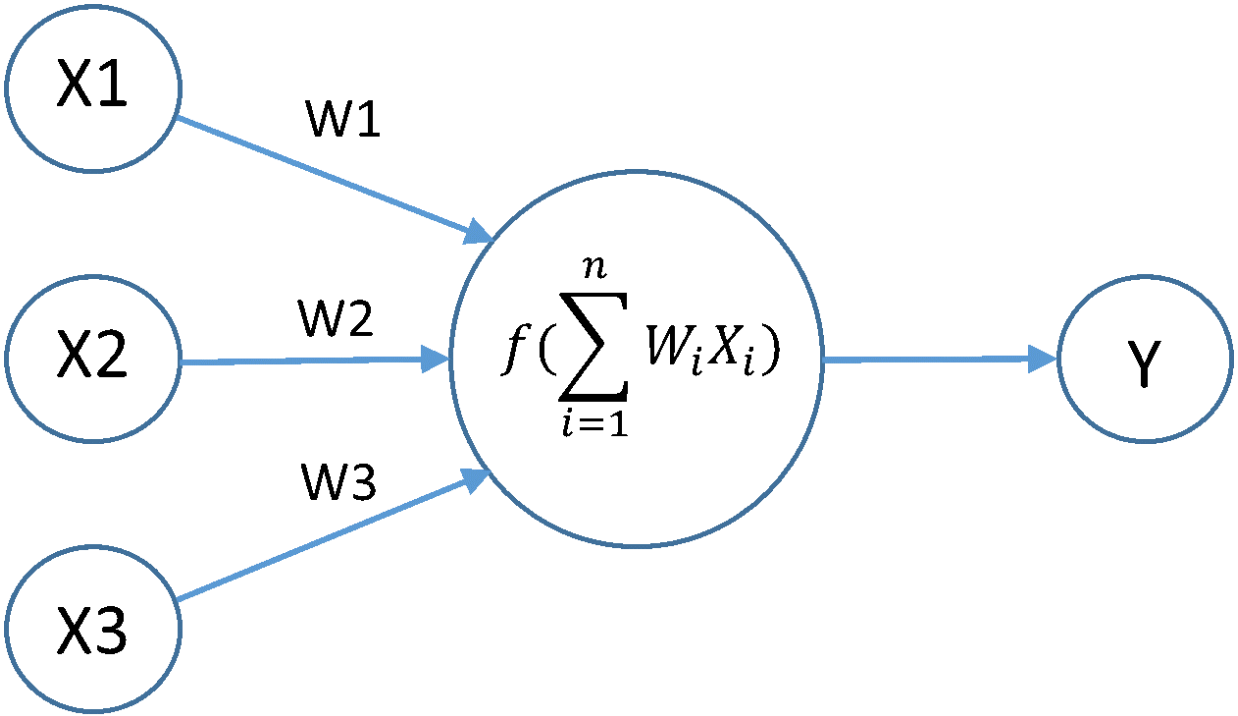
CATIE
Solutions pour la société numérique

Deep Learning

Réseaux de neurones



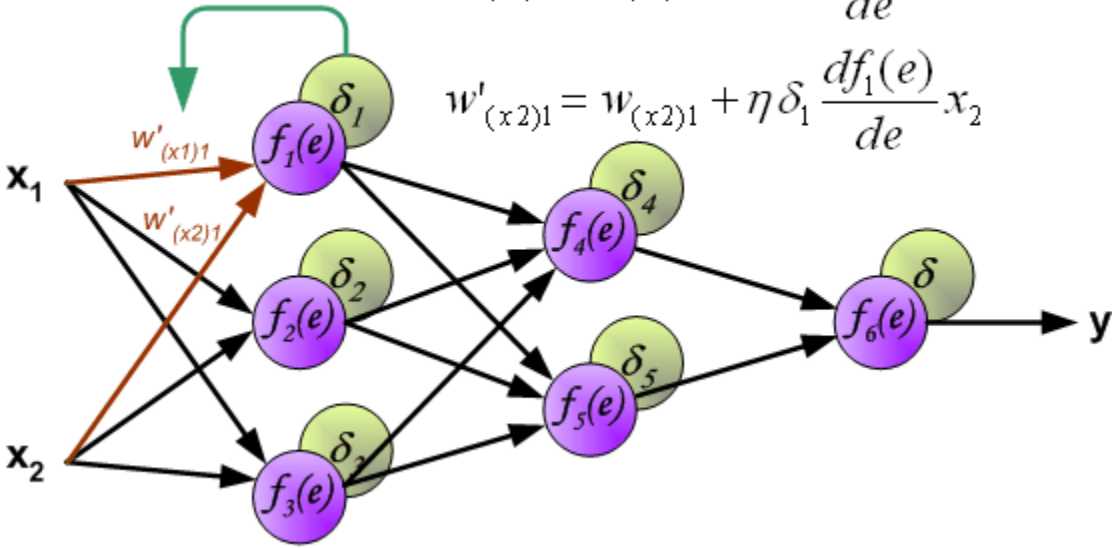
Réseaux de neurones



Réseaux de neurone – rétro-propagation

$$w'_{(x1)1} = w_{(x1)1} + \eta \delta_1 \frac{df_1(e)}{de} x_1$$

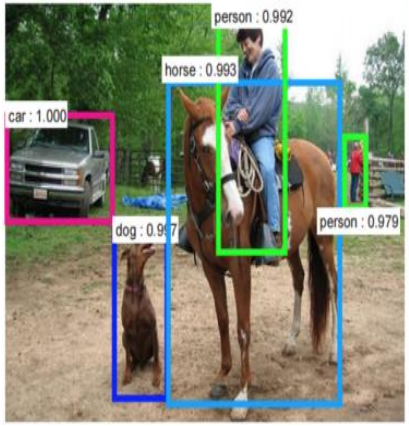
$$w'_{(x2)1} = w_{(x2)1} + \eta \delta_1 \frac{df_1(e)}{de} x_2$$

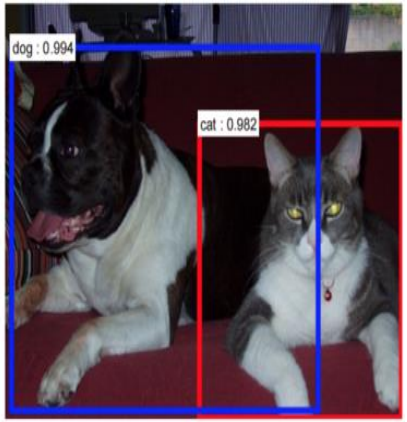


http://galaxy.agh.edu.pl/~vlsi/AI/backp_t_en/backprop.html

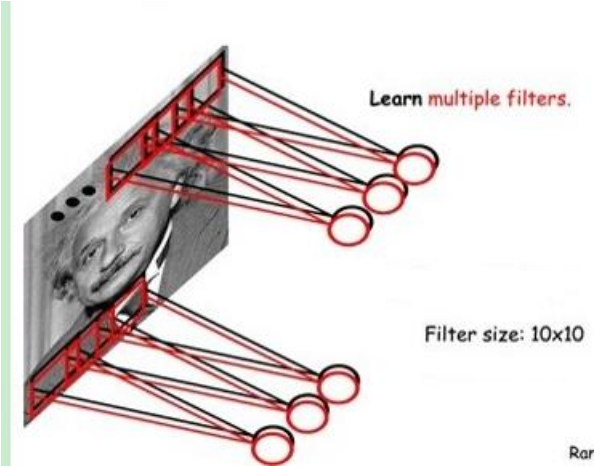
Deep learning – Réseaux de neurones convolutif

CATIE
Solutions pour la société numérique





AlphaGo

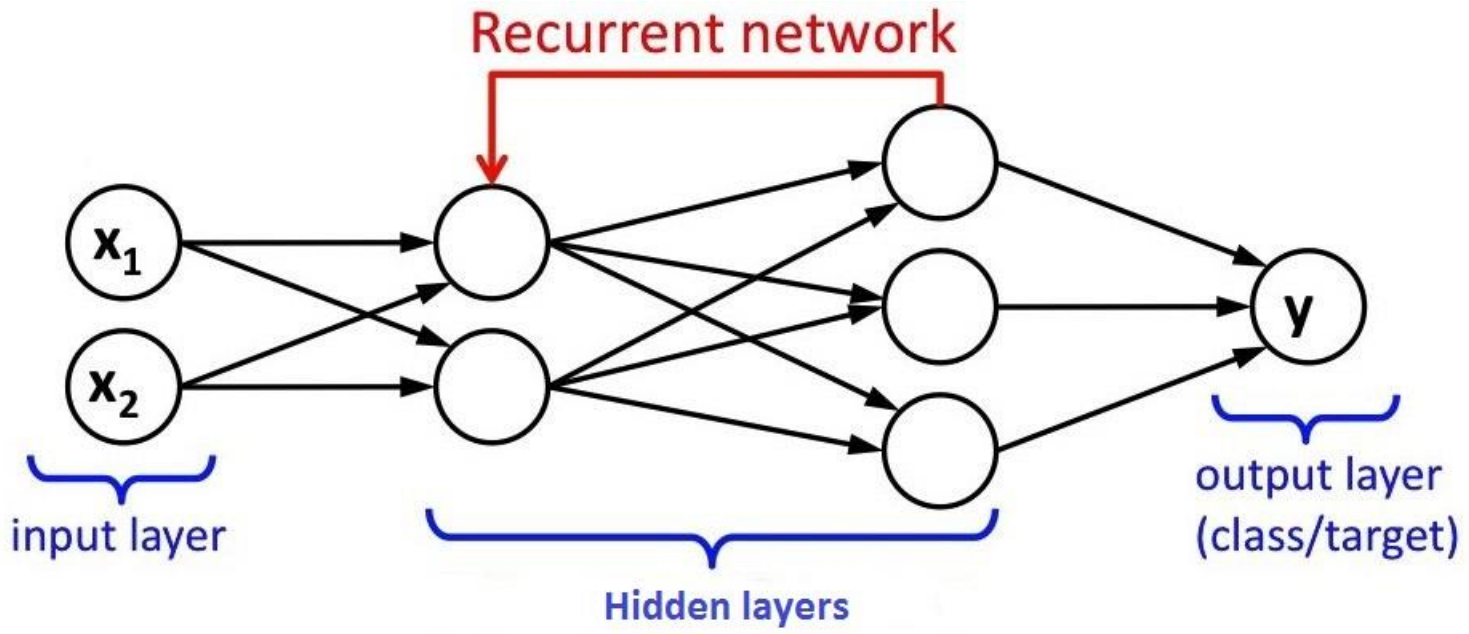


Deep learning – Réseaux de neurones récurrent

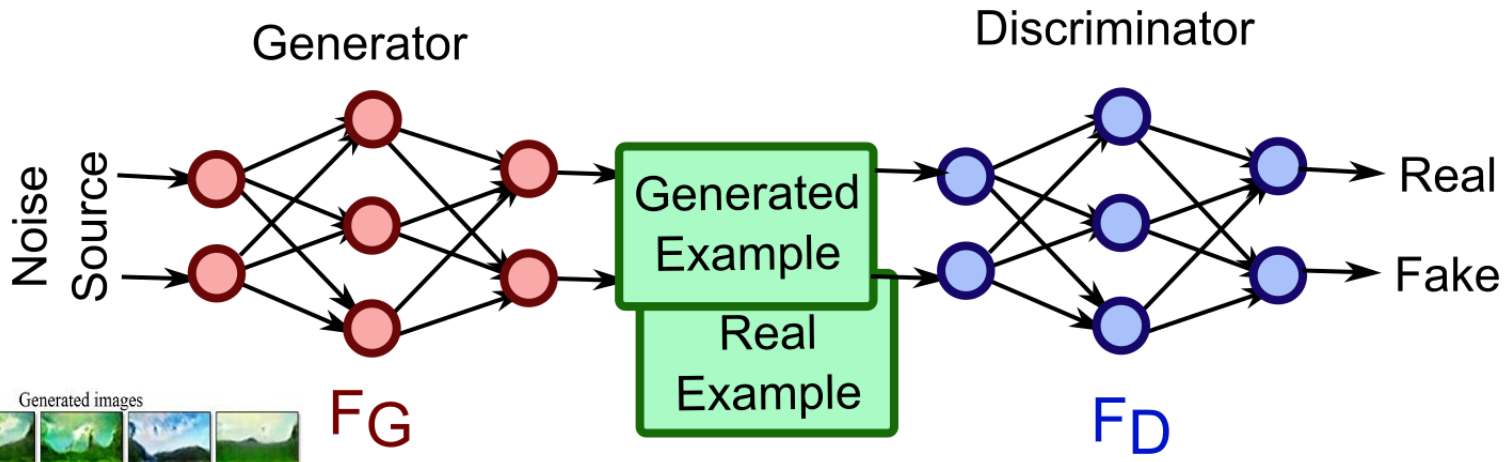
CATIE
Solutions pour la société numérique

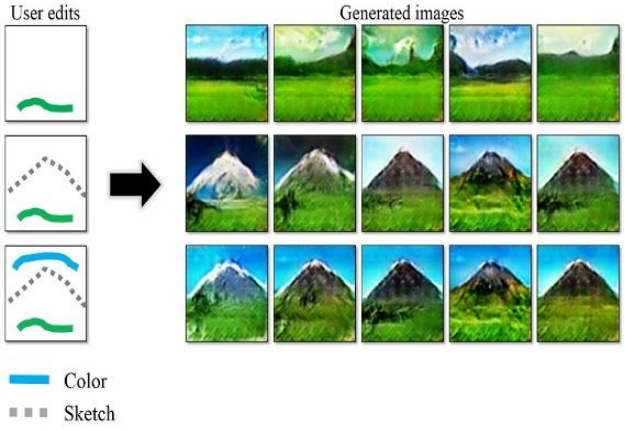
- Composition de musique
- Traduction des langues
- Reconnaissance automatique de la parole
- Synthèse vocale
- Chatbot
- …

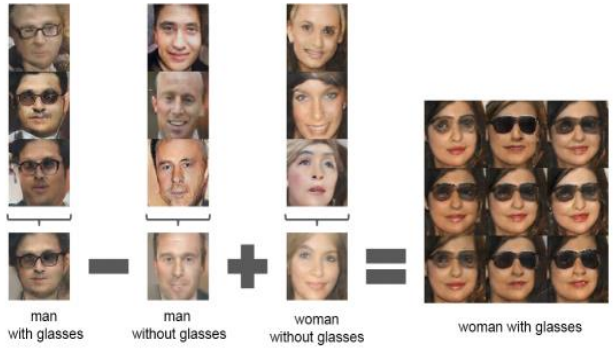
Deep learning – Réseaux de neurones récurrent



Deep learning – Generative Adversarial Networks







Les langages, les frameworks…

CATIE
Solutions pour la société numérique

Keras

TensorFlow

theano

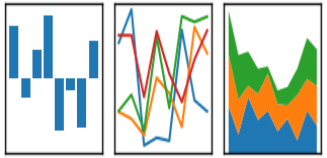
Ray

DASK

data
iku

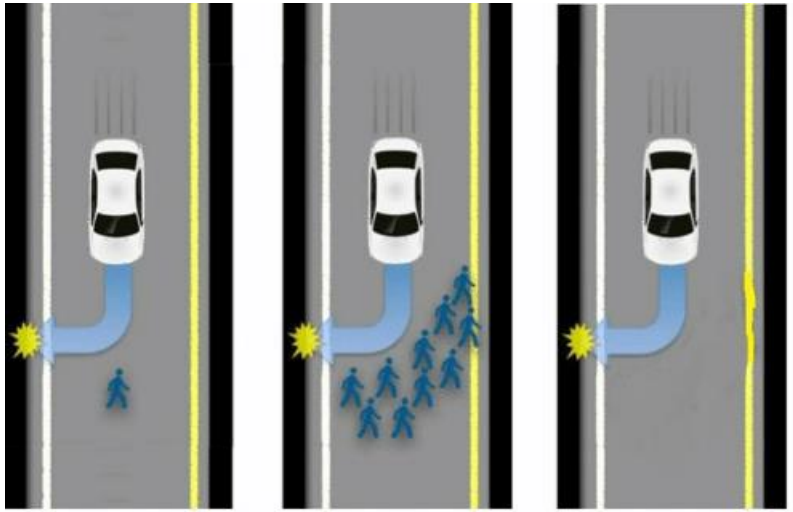
pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



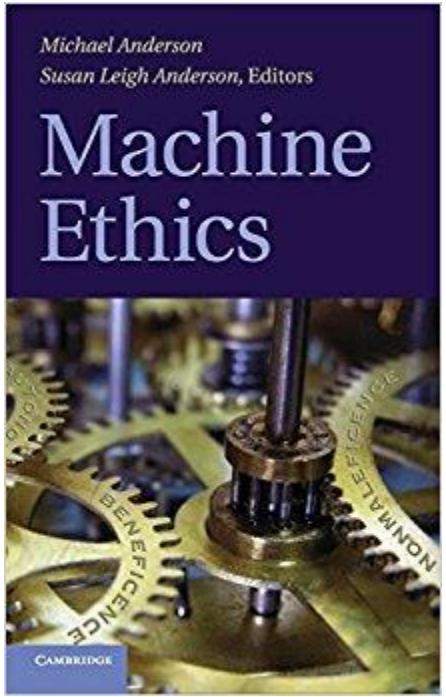
Ethique et Machine Learning

CATIE
Solutions pour la société numérique



Should A Self-Driving Car Kill Its Passengers In The Given Scenarios?

https://www.technologyreview.com/s/542626/why-self-driving-cars-must-be-programmed-to-kill/



CATIE

Solutions pour la société numérique

Merci pour votre attention

Questions ?