UN POINT SUR L'EXPLICABILITÉ ET L'INTERPRÉTABILITÉ EN (DEEP ...) MACHINE LEARNING

M. Serrurier
IRIT, Toulouse, France

12 novembre 2018

INTRODUCTION

LES DÉBUTS DE L'IA
// APPROCHE LOGIQUE VS APPROCHE CONNEXIONISTE

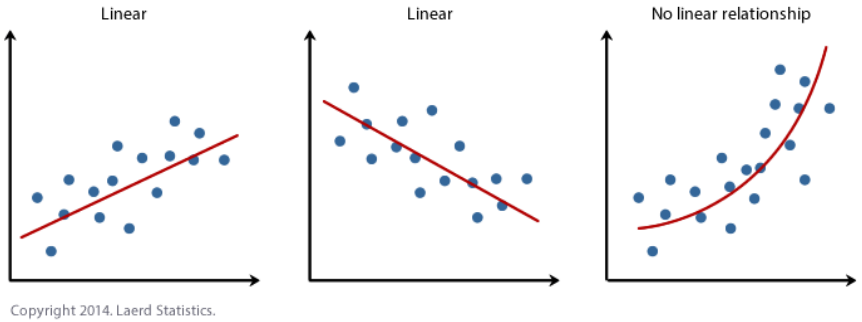
Acte de naissance : 1956, Darmouth College

Chaque aspect de l'apprentissage, ou toute autre caractéristique de l'intelligence peut en principe être décrit si précisément qu'il est possible de construire une machine pour le simuler

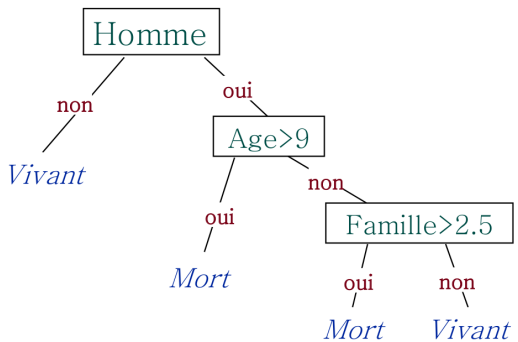
- Approche logique (John McCarthy)
  - Inspirée des travaux de Turing
  - Modéliser le raisonnement par la logique
- Applications : systèmes experts, recherche opérationnelle

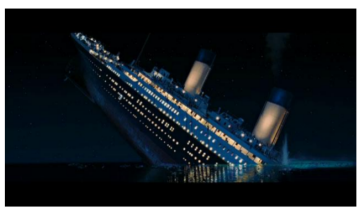
- Approche par schéma
  - Mc Culloch
  - Pitts
- Applications : Neurone artificiel
  - Perceptron
  - Perceptron multicouche
  - Deep learning

MACHINE LEARNING
// DIFFÉRENTS TYPES D'EXPLICABILITÉ



- Mathématique : Modèle linéaire
  - Pouvoir de représentation
  - Bornes de risque





- Informatique : Arbres de décision
  - Interprétable sous une forme logique
  - Lisible par un expert

Problème :

Sur des données complexes, les modèles d'apprentissage les plus efficaces sont aussi les moins explicables

- Cas pathologique : Deep learning
  - dépasse toutes les autres méthodes pour certains problèmes
  - Écosystème très développé et orienté vers les applications

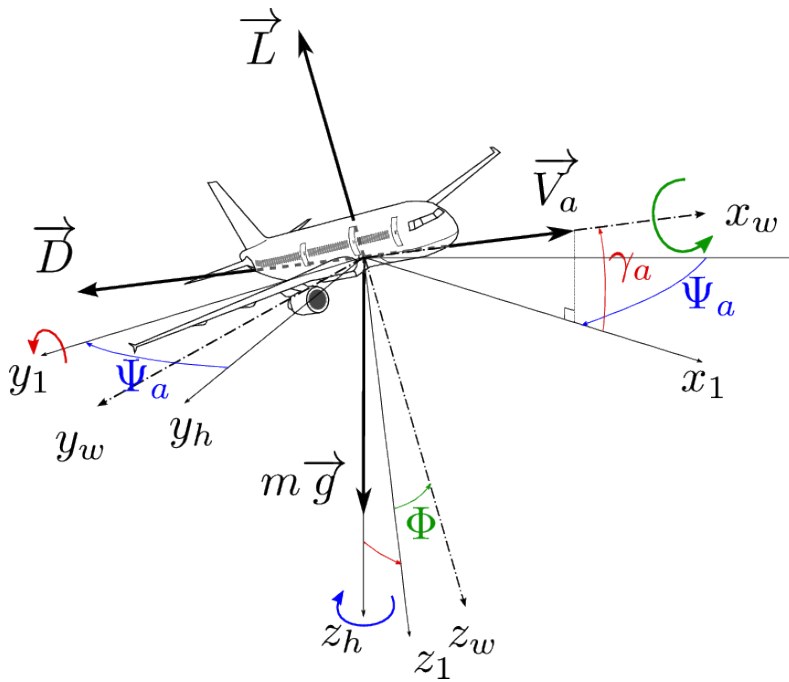
- Très peu de résultats sur l'explicabilité des réseaux profonds

- Explicabilité et interprétabilité des modèles
  - Garanties mathématiques
  - Modèles éthiques
  - Interprétation des CNN
  - Apprentissage de représentation
- Explicabilité et interprétabilité des prédictions
  - Cartes d'activation
  - Approximation locale de modèle
  - Deep learning et logique
- Aspects Neurosciences, philosophiques et juridiques

EXPLICABILITÉ ET INTERPRÉTABILITÉ DES MODÈLES

EXPLICABILITÉ DES MODÈLES
// BESOIN DE PREUVES THÉORIQUES

- Problème pour les applications critiques ex :
  - Prédiction de trajectoire d'avion
  - Détection d'intrusions dans un système
  - Véhicules autonomes

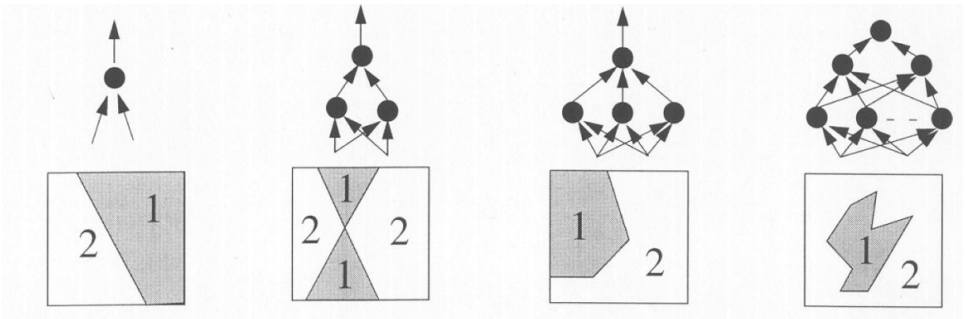


Problème :

Besoin de garantie sur les marges de sécurité des modèles

EXPLICABILITÉ DES MODÈLES
// QUELS TYPES DE GARANTIES ?

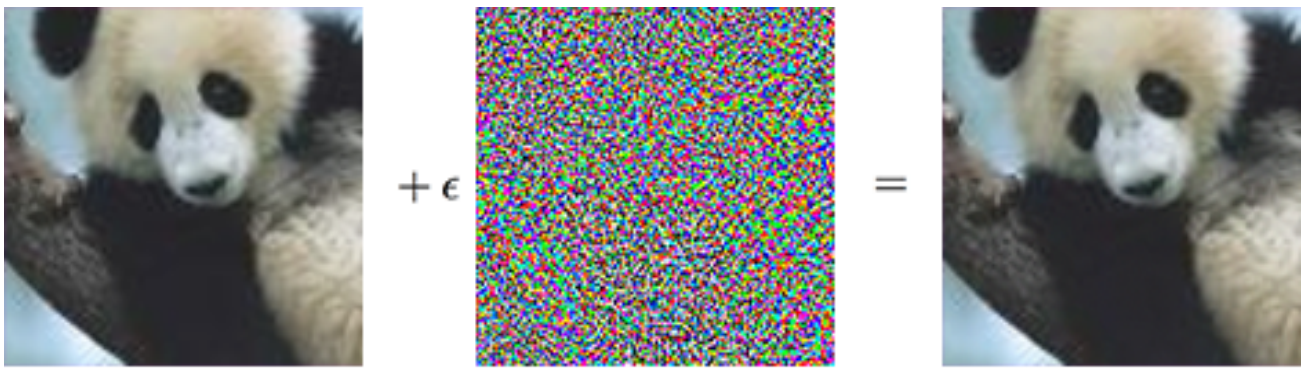
- Pouvoir de représentation



- optimisation
  - convergence/optimalité
  - vitesse de convergence
- Statistiques bornes sur les erreurs

EXPLICABILITÉ DES MODÈLES
// ROBUSTESSE

- Analyse de sensibilité
- Robustesse aux attaques



- Image inverse

EXPLICABILITÉ DES MODÈLES
// MODÈLES ÉTHIQUE

- Biais dans les bases
  - Prédiction criminalité
  - Selection de CV

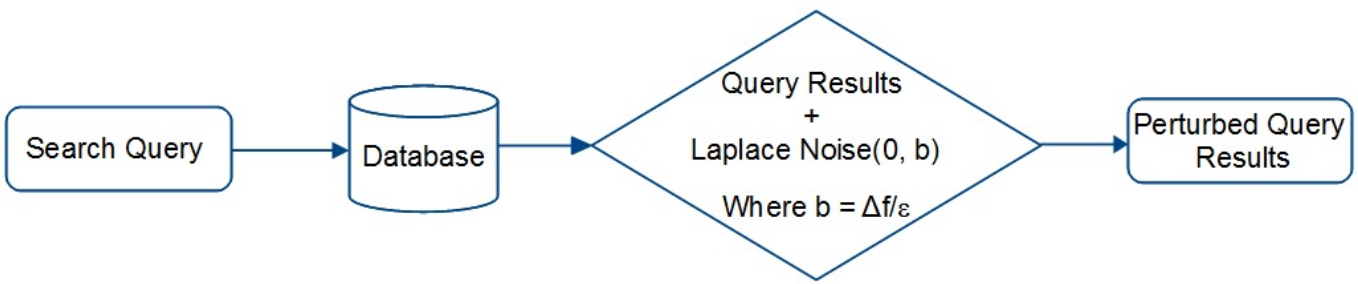
Intimité différentielle

Problème lié aux informations que l'on veut cacher, mais qui peuvent être trouvé çà l'aide d'algorithme d'apprentissage

- Machin learning et vie privée
  - Vie privée des utilisateurs
  - Secret industriel

EXPLICABILITÉ DES MODÈLES
// MODÈLES ÉTHIQUE

- Fair learning : Prédire Y grâce à X mais sans être biaisé par une variable A. Possibilités :
  - Modifier X
  - Contraindre l'apprentissage
- Intimité différentielle
  - anonymisation
  - Partage de données et d'algorithme sans divulguer des données sensibles

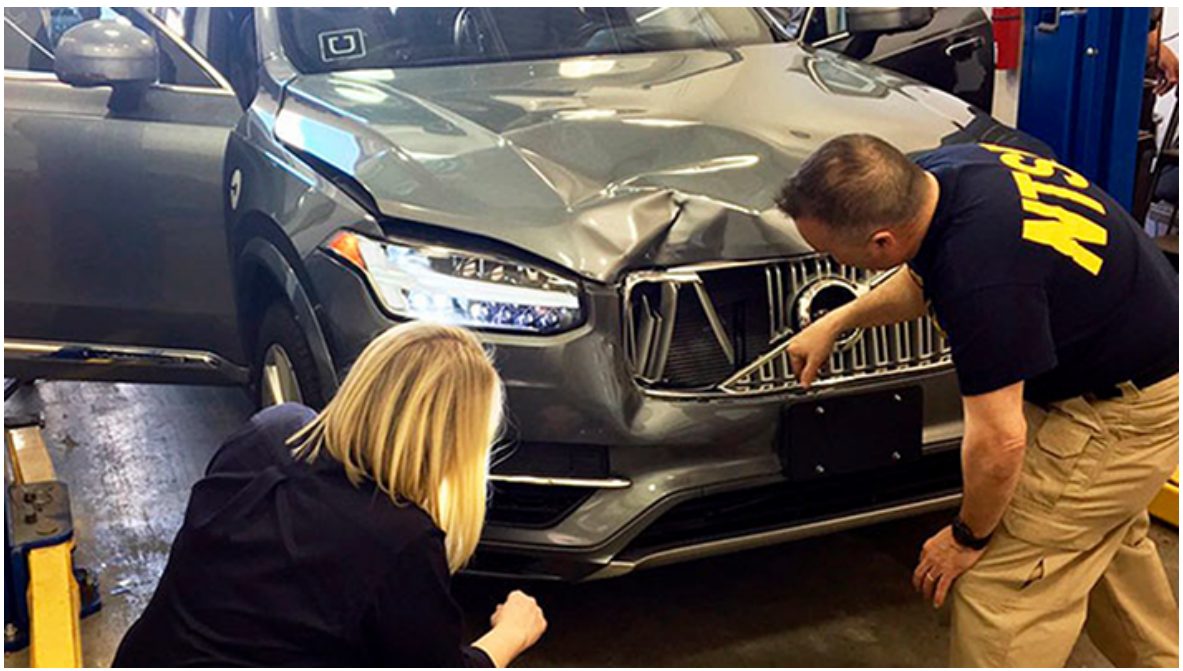


EXPLICABILITÉ ET INTERPRÉTABILITÉ DES PRÉDICTIONS

EXPLIQUER UNE PRÉDICTION
// ACCIDENT VOITURE AUTONOME

Question :

Comment expliquer l'erreur d'une voiture autonome



EXPLICABILITÉ DES PRÉDICTIONS

// COMMENT VOIT UN RÉSEAU DE NEURONES

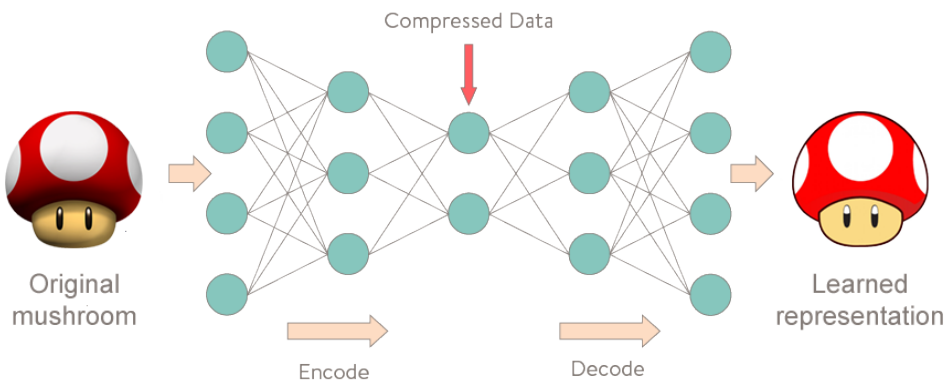
EXPLICABILITÉ DES PRÉDICTIONS
// RÉSEAU CONVOLUTIONNEL

- déterminer quels sont les pixels d'une image qui impactent le plus la classification :
  - analyse locale de sensibilité
  - carte d'activation basée sur le gradient

Principe :

Apprendre une représentation de faible dimension des données

- Auto-encoder
- Variational Auto-encoder
- Aprentissage adverse
  - Generative Adverserial Nerworks (GAN)
  - Bi-GAN
  - Info-GAN

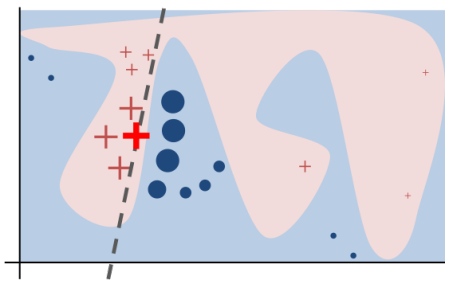


EXPLICABILITÉ DES PRÉDICTIONS
// APPROXIMATION

Principe :

Approximer localement un réseau avec un modèle linéaire

- LIME - Local Interpretable Model-Agnostic Explanations
  - Perturbation autour de l'exemple à prédire
  - Calcul d'un séparateur linéaire pour ces exemples perturbés
  - Interprétation à partir du modèle linaire
- Aproche agnostique : ne dépends pas du modèle que l'on veut approximer



EXPLICABILITÉ DES PRÉDICTIONS
// RÉSEAUX LOGIQUES

Principe :

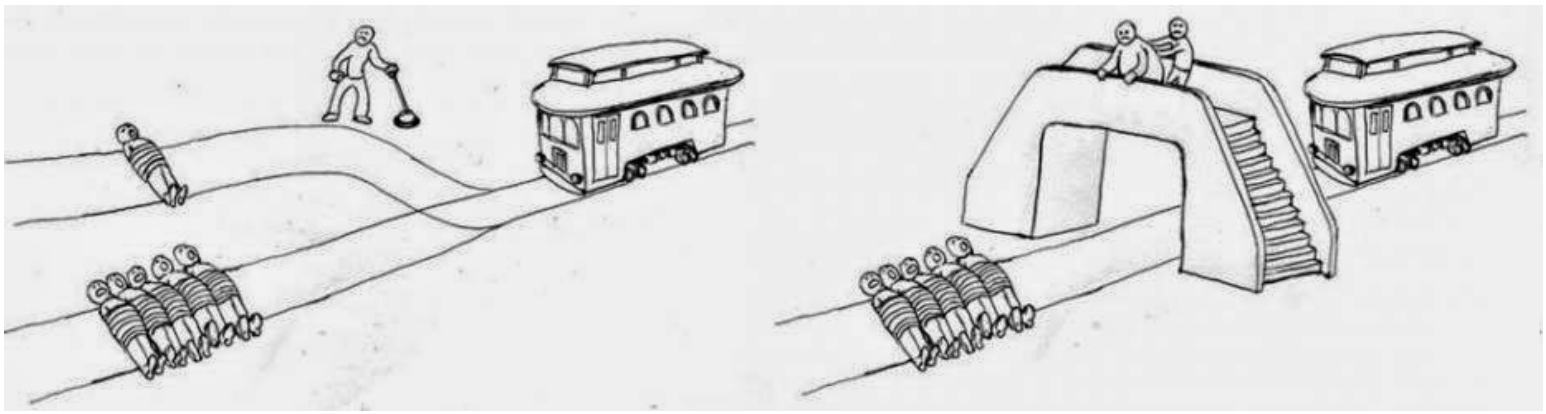
Utiliser un réseau de neurones pour apprendre des règles en logique du premier ordre

- TensorLog
  - Codage des données logique sous forme matricielle
  - Inférence par calcul de gradient
- Permet de considérer de l'information apriori
- Ne marche que sur des problèmes simples

NEUROSCIENCE, PHILOSOPHIE ET JURIDIQUE

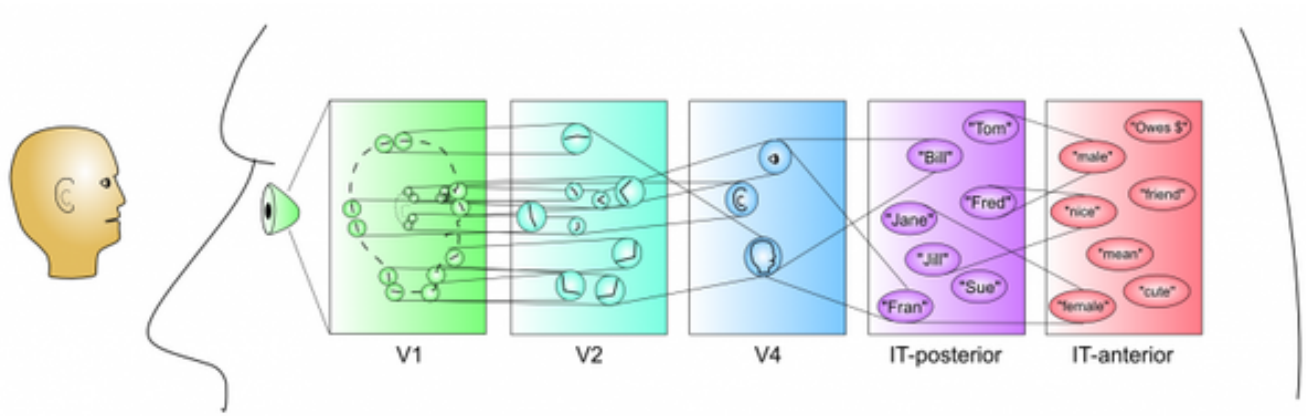
AU-DELÀ DU MACHINE LEARNING
// ASPECTS PHILOSOPHIQUES

- Comment définir l'interprétabilité
- Qu'est qu'un algorithme éthique
- Acceptabilité des décisions algorithmiques



AU-DELÀ DU MACHINE LEARNING
// NEUROSCIENCE

- Les réseaux de neurones ont-ils une interprétation biologique
- Les réseaux de neurones peuvent-ils aider à la compréhension du cerveau



AU-DELÀ DU MACHINE LEARNING
// ASPECTS JURIDIQUE

- Droits des algorithmes
  - Libéralisation des données publiques
  - Droit à l'explicabilité
- problèmes :
  - Comment définir l'explicabilité en droit
  - comment définir les responsabilités pour des décisions prises par des algorithmes

Exemple :

Facebook et le divorce

CONCLUSIONS

- Apprentissage artificiel capable de répondre à des problèmes difficiles
- L'aspect boîtes noires des algorithmes d'apprentissage limite leur applicabilité à des applications critiques
- Besoin important
  - De garantie formelle
  - De faire le lien avec les outils formels d'intelligence artificielle
- Déborde sur d'autres domaines
  - philosophie
  - droit
  - neurosciences