Problématiques liées au machine learning

Dans la vraie vie...

Les Problémes théoriques

- Les données (trouées, incomplètes, non-représentatives, biais statistiques,...)
- Une fonction de perte (objectif) pas adaptée au besoin
- La fuite d'information dans le modèle
- La quasi non-interprétation du modèle
- Des modèles qui ne s'adaptent pas au changements
- Le surapprentissage

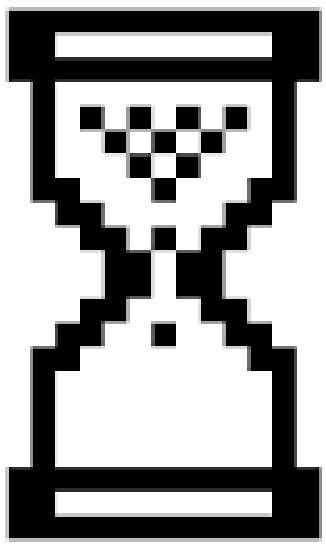
Les Problémes logistiques

Les **données** coûtent cher (récolte, nettoyage, #Data>1M).

Les Problémes logistiques

L'apprentissage d'un modèle prend **beaucoup** de **temps**

- Reconnaissance parole : 4 GPU $=>$ plusieurs jours, semaines
- Alphago : 3 semaines sur 5000 TPU ($\approx$ 30M$)
- On n'obtient pas de résultats concluant avec le premier run.
- Peu de visibilité sur le temps d'obtention d'une plus value



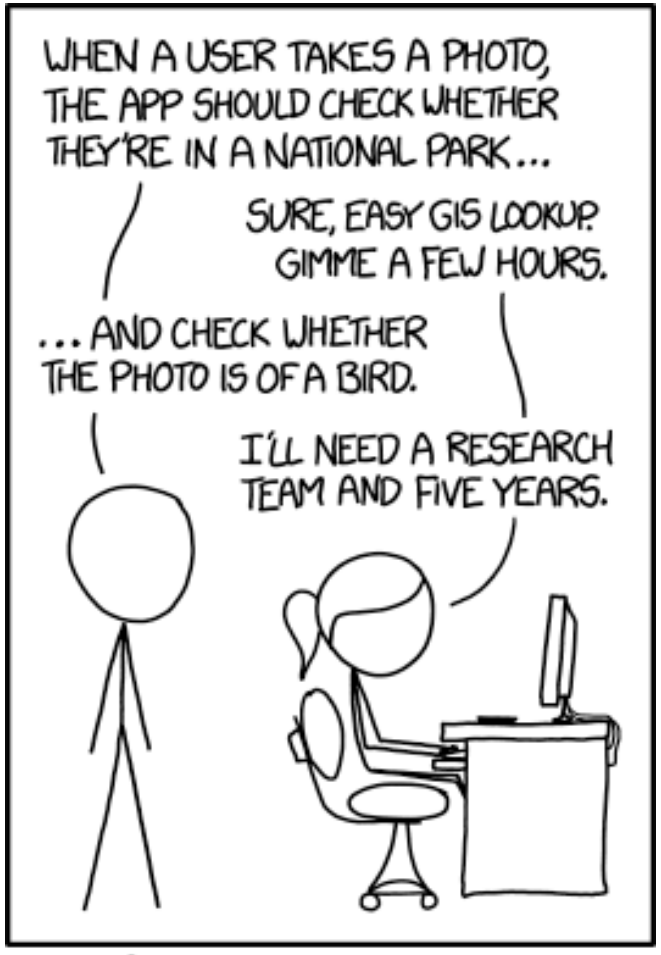
Les Problémes logistiques

L'utilisation de **GPU/TPU** coûte cher

...et ça ne s'arrange pas avec l'explosion du besoin par le minage de cryptomonnaies

Cours de l'action NVIDIA

Les Problémes de communication



Les Problémes logistiques

Les **ingénieurs en machine learning** (compétents) coûtent cher :

- git clone d'un papier + jouer avec data $\neq$ data-scientist
- au croisement de l'ingénérie (traitement de gros volumes de données, standard de développement) et des mathématiques (statistiques, algèbre, optimisation)
- Le nombre de nouveaux diplomés ne suit pas la demande
- Les meilleurs sont/vont dans une poignée d'entreprises ! (ou presque...)

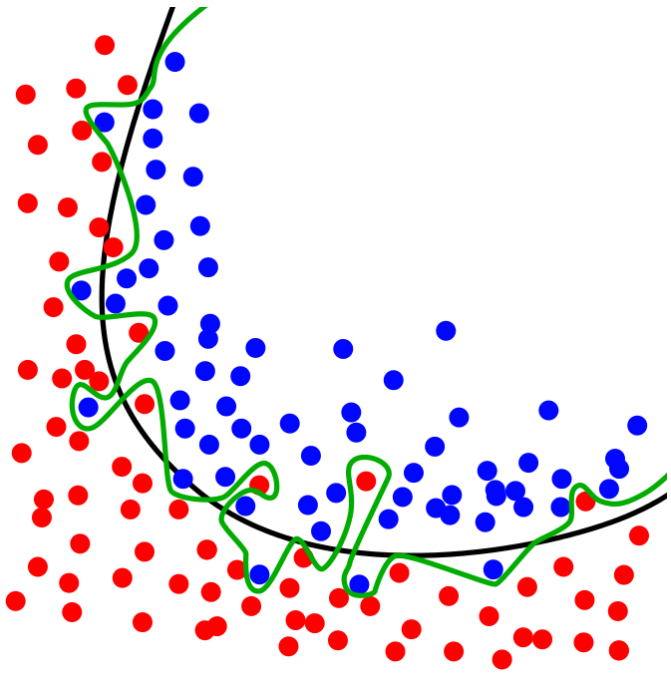
Expertise en Machine Learning

- Savoir transcrire le besoin en modèle d'apprentissage.
- Quels modèle sur quelles données ?
  - Forme : Nombre de couches ? de quelle taille ? quelle astuce ? ...
  - L'algorithme d'optimisation (SGD, adaboost, adam,...)
  - Méthodes de régularisation (norme des paramètres dans la loss, bruitage, dropout, ...)

Régularisation

$\approx$

empêcher le surapprentissage



Les Problémes de communication

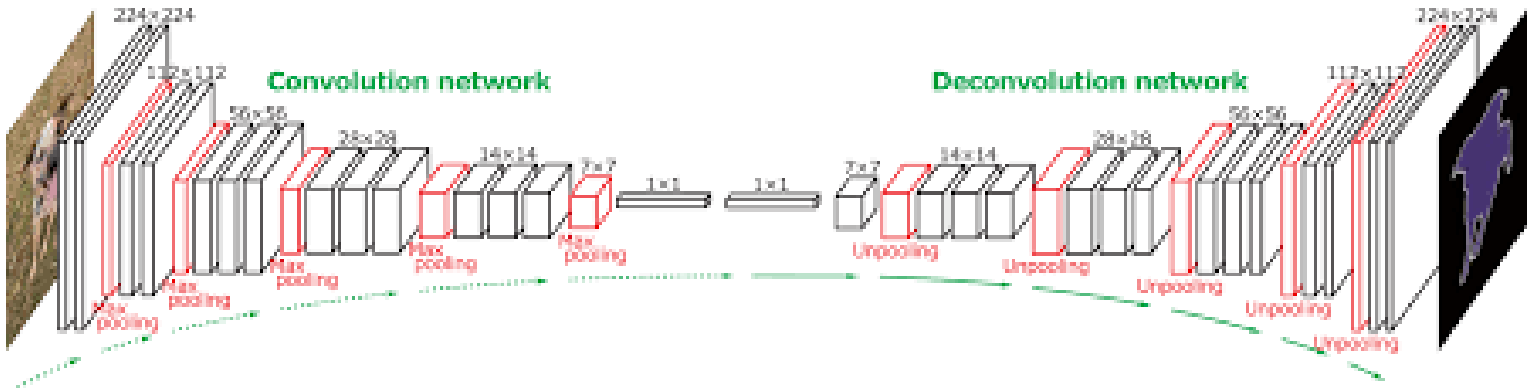
Utilisateurs prêts à accepter un algorithme qui fait des erreurs ?

Les Problémes de communication

Utilisateurs prêts à accepter un algorithme qui fait des erreurs ?

Les Problémes de communication

Utilisateurs prêts à accepter un algorithme qui n'est pas interprétable ?
Réseaux de neurones <=> enchainement de multiplications matricielles avec des millions de paramètres.



Les Problémes Ethiques

Grande-Bretagne : Predpol (abandonné), National Data Analytics Solution (NDAS) en mars 2019

Les Problémes Ethiques

Détection mensonge à la frontière (iBorderCtrl)

Projet pilote en Hongrie, Lettonie et Grèce

Les Problémes Ethiques

Chine 170 Millions de caméras (450 Millions en 2020)

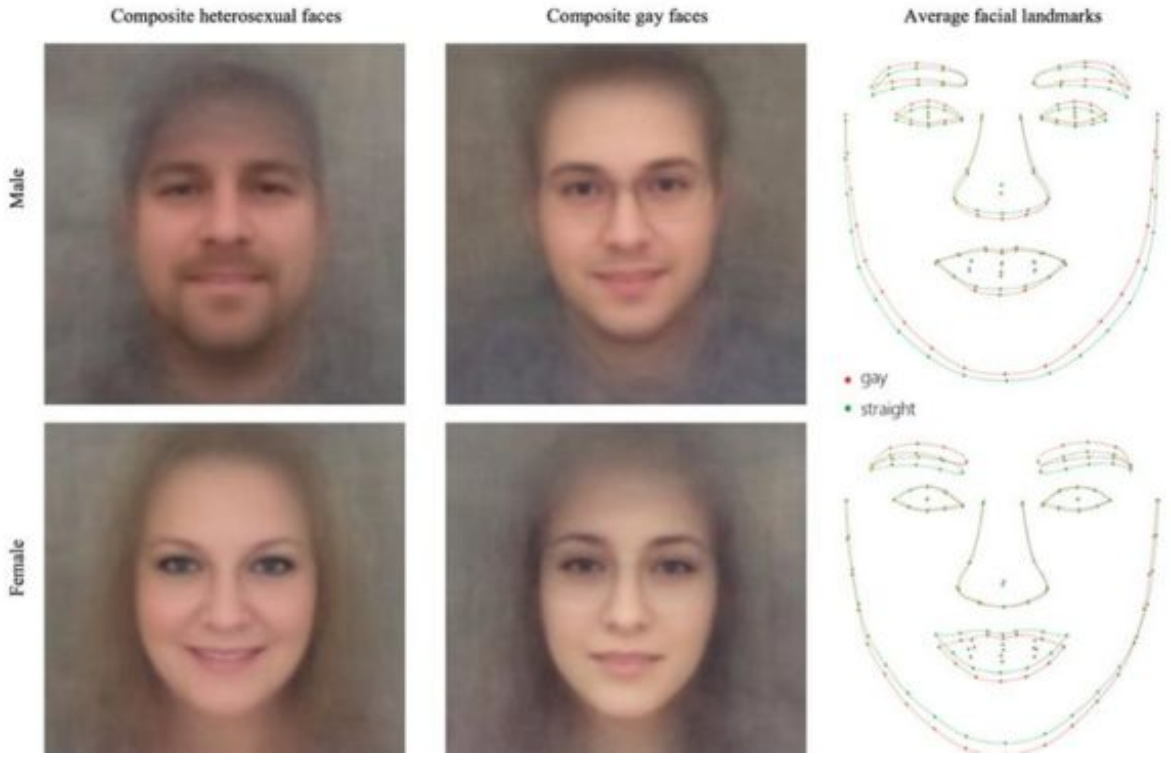
Les Problémes Ethiques

Les Problémes Ethiques

Automatisation du processus de recrutement

Les Problémes Ethiques

Préférence sexuelles à partir d'une photo ?



Les Problémes Ethiques

Évaluer une Baby-sitter en fonction de son activité en ligne ?

Les Problémes Ethiques

La chine traque ses musulmans du Xinjiang :
Classement automatique en 36 catégories à partir de toutes les données possibles.

Les Problémes Ethiques

OpenAI : GPT-2 est 'trop' performant pour la génération de texte

Example One

System Prompt:

Deutsche Bank lays off 4,000 people.

Model completion:

Out of the 40,000 employees at Deutsche Bank, only about 4,000 will be asked to leave, CEO John Cryan said.

The bank is slated to announce as soon as this week the formal creation of a new capital structure, one of the most important changes in its history.

According to the Financial Times, some of the key changes that the bank has to put in place include:

A new common equity tier one capital structure: Instead of a separate "core" and "non-core" capital, there will be a common tier one capital. The core and non-core capital will be separate.

Instead of a separate "core" and "non-core" capital, there will be a common tier one capital. The core and non-core capital will be separate. The bank's asset-based funding ratio: The ratio that compares a bank with its assets to its liabilities will be lowered from a minimum of 10 per cent to 7 per cent.

Les Problémes Ethiques

Deepfake : 'Zao'