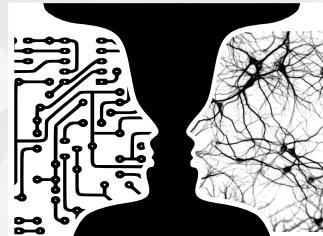


IFT-2004 -- MODÈLES ET LANGAGES DES BASES DE DONNÉES

**Le “Big Data” et l'intelligence artificielle...
qu'est-ce que c'est au juste ?**

*Présenté par
François Laviolette*



Groupe de
Recherche en
Apprentissage
Automatique de
Laval



Qu'est-ce que le « Big Data » ?

- D'abord, quel est le bon terme français?
 - En France on parle de *mégadonnées*
 - Je préfère l'expression *données massives*,

entre autre parce que nous pensons que le « Big Data »
n'est pas qu'un problème de quantité.

Qu'est-ce que le “Big Data”? ... En 4 V

- Volume

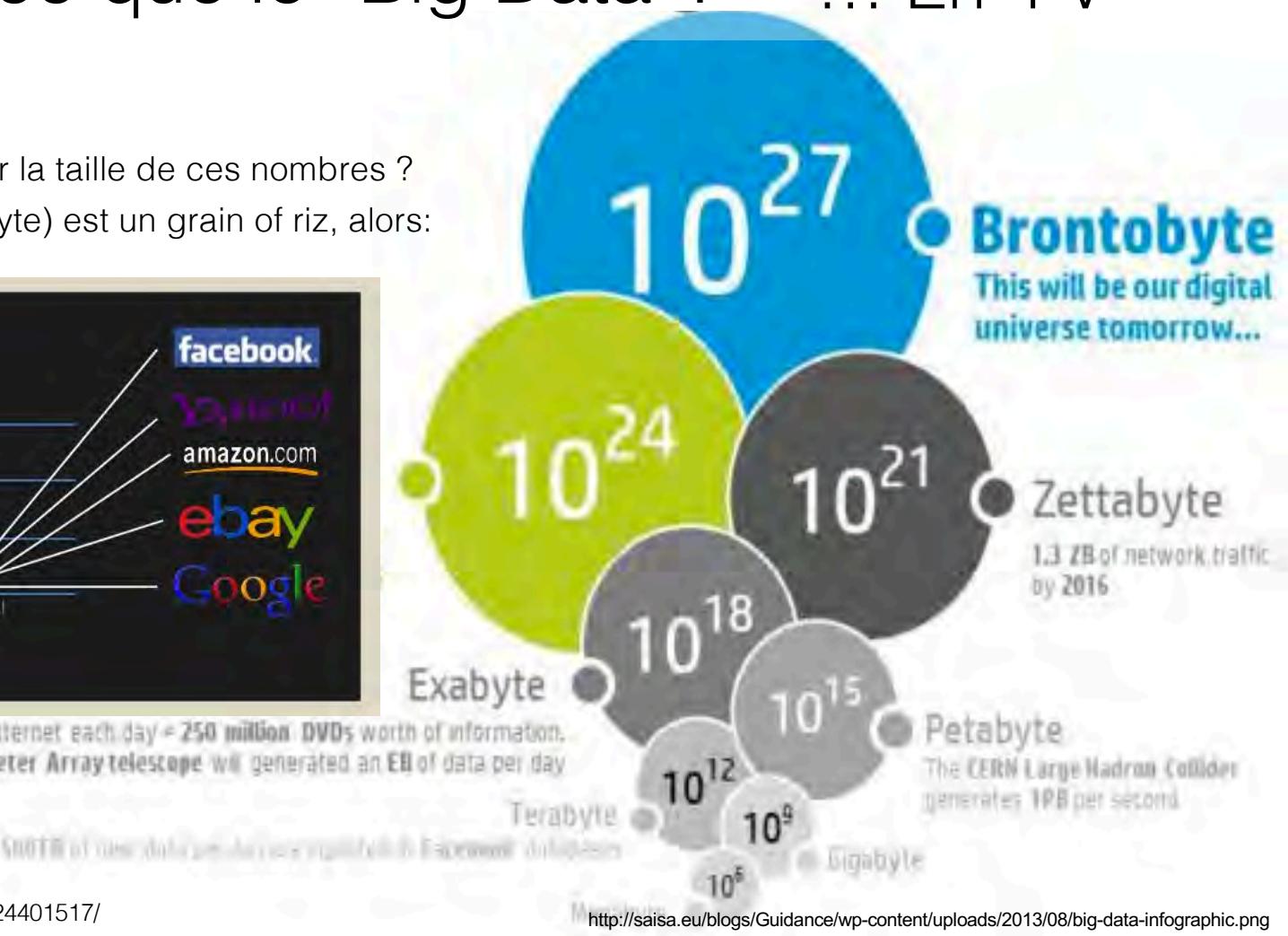
- Pouvons-nous nous figurer la taille de ces nombres ?
- Supposons qu'un octet (byte) est un grain of riz, alors:



1 EB of data is created on the internet each day = 250 million DVDs worth of information.
The proposed Square Kilometer Array telescope will generate an EB of data per day

Comparison given by David Wellman

<http://fr.slideshare.net/dwellman/what-is-big-data-24401517/>

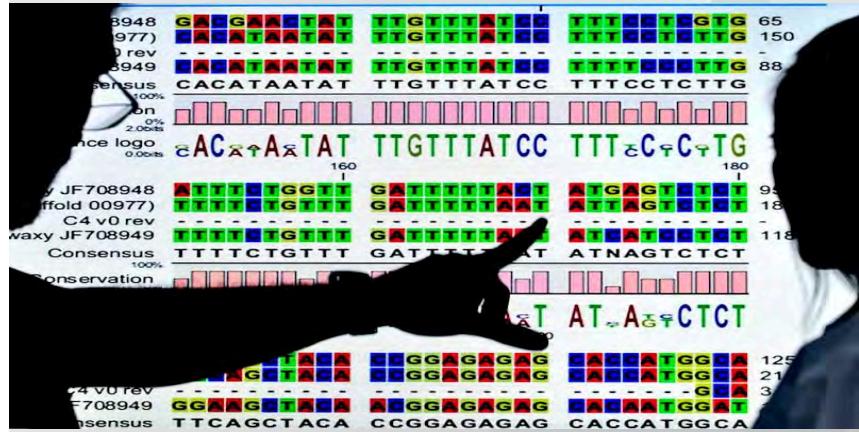


Qu'est-ce que le “Big Data”? ... En 4 V

- Volume
- Vélocité
- Variété
- Véracité

*Données provenant de diverses sources
non nécessairement structurées
Image, texte, données de senseurs, ...*

*Données provenant de projets differents,
avec des méthodologies non nécessairement compatibles*



Interprétation de données « omics »



des données d'imageries

et même des données environnementales



combinées avec des données cliniques

Exploiter la donnée environnementale

Les statistiques environnementales du gouvernement sont gérées à Orléans. Ce service vient d'inaugurer un espace ouvert aux start-up.

Carole Tribout
caroletribout@orange.fr

Après la French Tech, le Lab'D et l'AgroTech Val de Loire, voici le centre Green Tech Verte d'Orléans. Un espace, comme son nom anglophone ne l'indique pas, géré par le ministère des Transports et de l'Aménagement.

Il se situe au sein d'un site national, le Commissariat général au développement durable, qui accueille le pôle aménagement du service statistique du ministère, au 5, route d'Oliver, à Orléans. Y travaillent 70 salariés, sous la responsabilité d'Eric Bommati.

De 15 à 20 personnes

Le pôle vient de dénicher son espace documentation pour accueillir ce « datacenter ». C'est un espace de travail de 100 mètres carrés, raccordé à Internet (via une connexion à 10 Gbit/s) et à une centrale de données. Il sera dédié aux start-up innovantes qui ont envie de développer des données environnementales pour les domaines environnementaux, énergétiques, climatiques, chêne, et les start-up innovantes du concours que le centre vient de lancer (dir en encadré).

Il propose du haut débit, le Wi-Fi de bureau, de la téléconférence, une salle de réunion avec une 10-mètre-tablette et

une imprimante 3D.

Il a été mis en place

Un concours sur les pesticides

L'Institut national de l'environnement et du développement durable a lancé, hier, un concours national. Il s'agit d'imaginer de nouvelles solutions pour mieux visualiser les données concernant les pesticides dans les eaux souterraines. Les concurrents jusqu'à 16 juillet pour s'inscrire. Ils pourront apprendre les mesures de 2008 et 2014, faire un état des lieux de 2014, laisser un dossier de dossier le 16 février; les porteurs des projets retenus pourront gagner gratuitement de l'espace orléanais: 6.000 € pour le premier prix, 3.000 ou deuxième, 1.200 ou troisième. Même si le vainqueur n'a pas de solution à mettre en œuvre immédiatement, il peut la présenter au public, et il arrive, selon le référent Green Tech Verte Orléans, qu'un seul point de mesure recèle jusqu'à 40 molécules différentes, dont certaines

qui sont toxiques.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

Il existe aussi

des partenariats

avec des universités

et des laboratoires

de recherche.

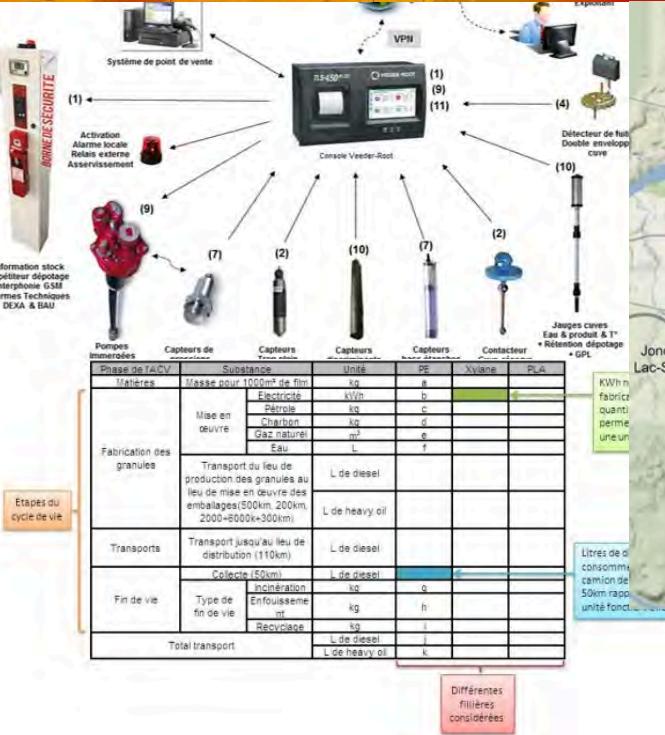
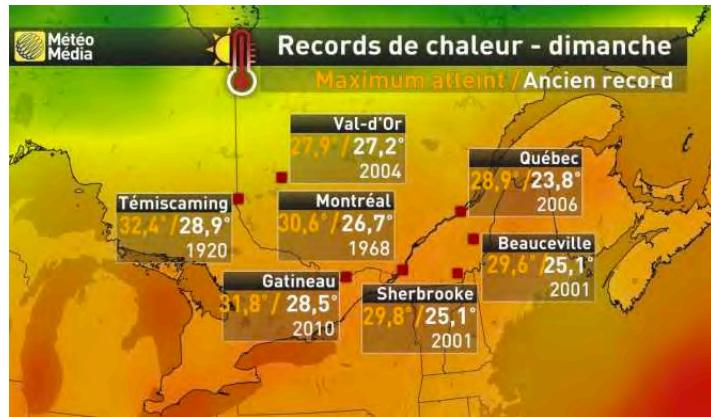
Il existe aussi

des partenariats

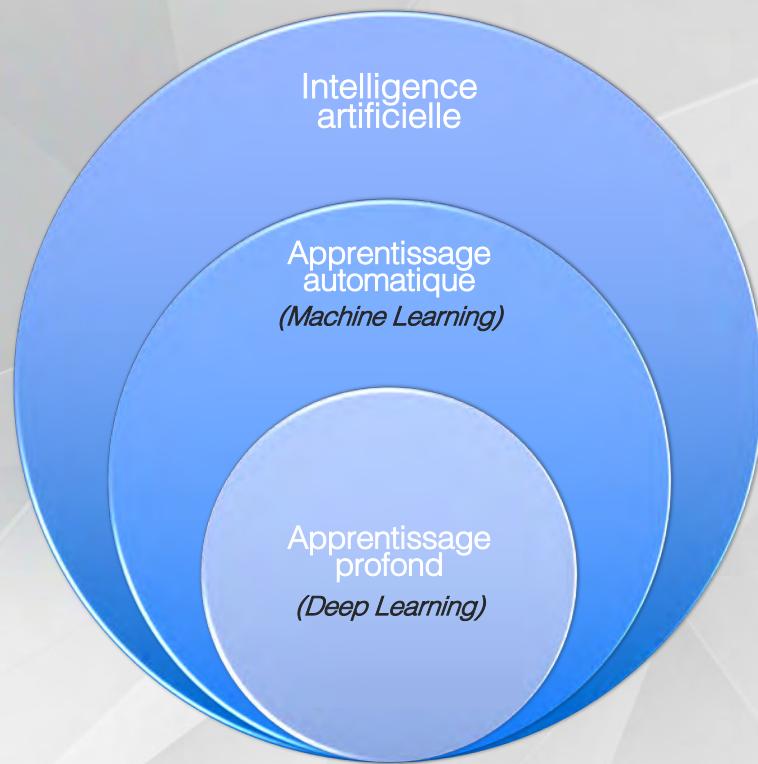
avec des universités

et des laboratoires

es hétérogènes et non structurées (suite)



L'intelligence artificielle et ses apprentissages

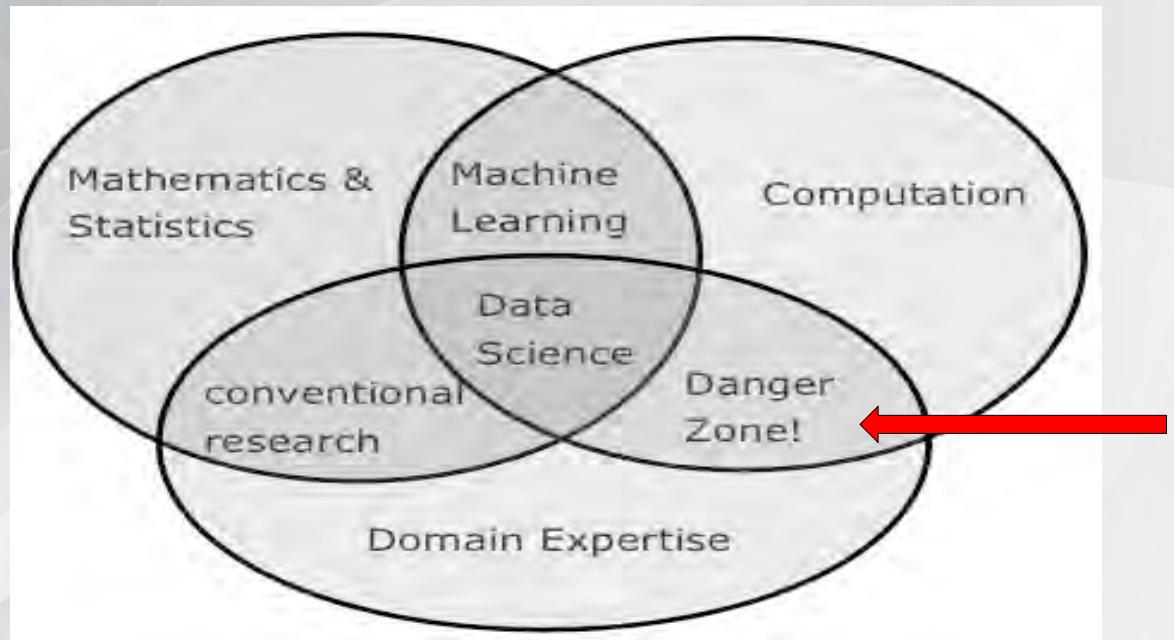


Source: [« Why Deep Learning Matters and what's next for Artificial Intelligence »,](#)
[Algorithmia, Novembre 2016](#)

L'apprentissage automatique (machine learning) et les données massives



Apprentissage automatique:
Comprendre les informations



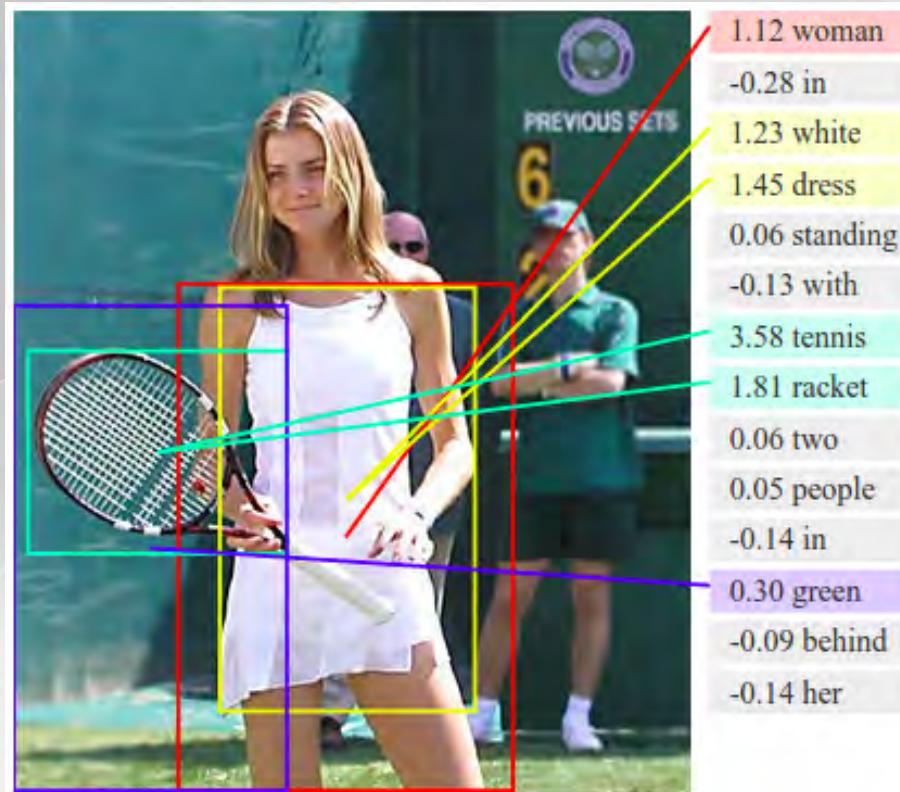
Le diagramme de Venn de Drew Conway sur le « Big Data »

Étiquetage de scènes par des réseaux profonds



[Farabet et al. ICML 2012, PAMI 2013]

L'apprentissage machine, un outil pour « percevoir » les informations



<http://cs.stanford.edu/people/karpathy/deepimagesent/>

L'apprentissage machine, un outil pour « percevoir » les informations



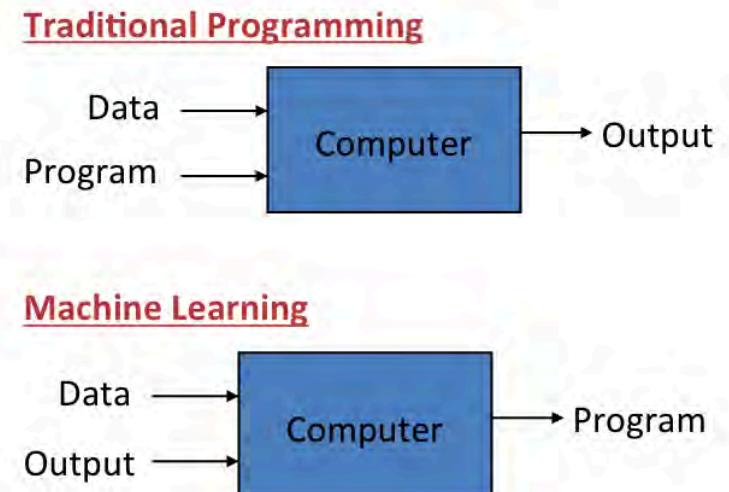
<https://www.archimag.com/reseaux-sociaux/2016/08/22/intelligence-artificielle-facebook-fasttext-textes-open-source>

Apprentissage automatique 101

“Field of study that gives computers the ability to learn without being explicitly programmed.”

-Arthur Samuel (1959)

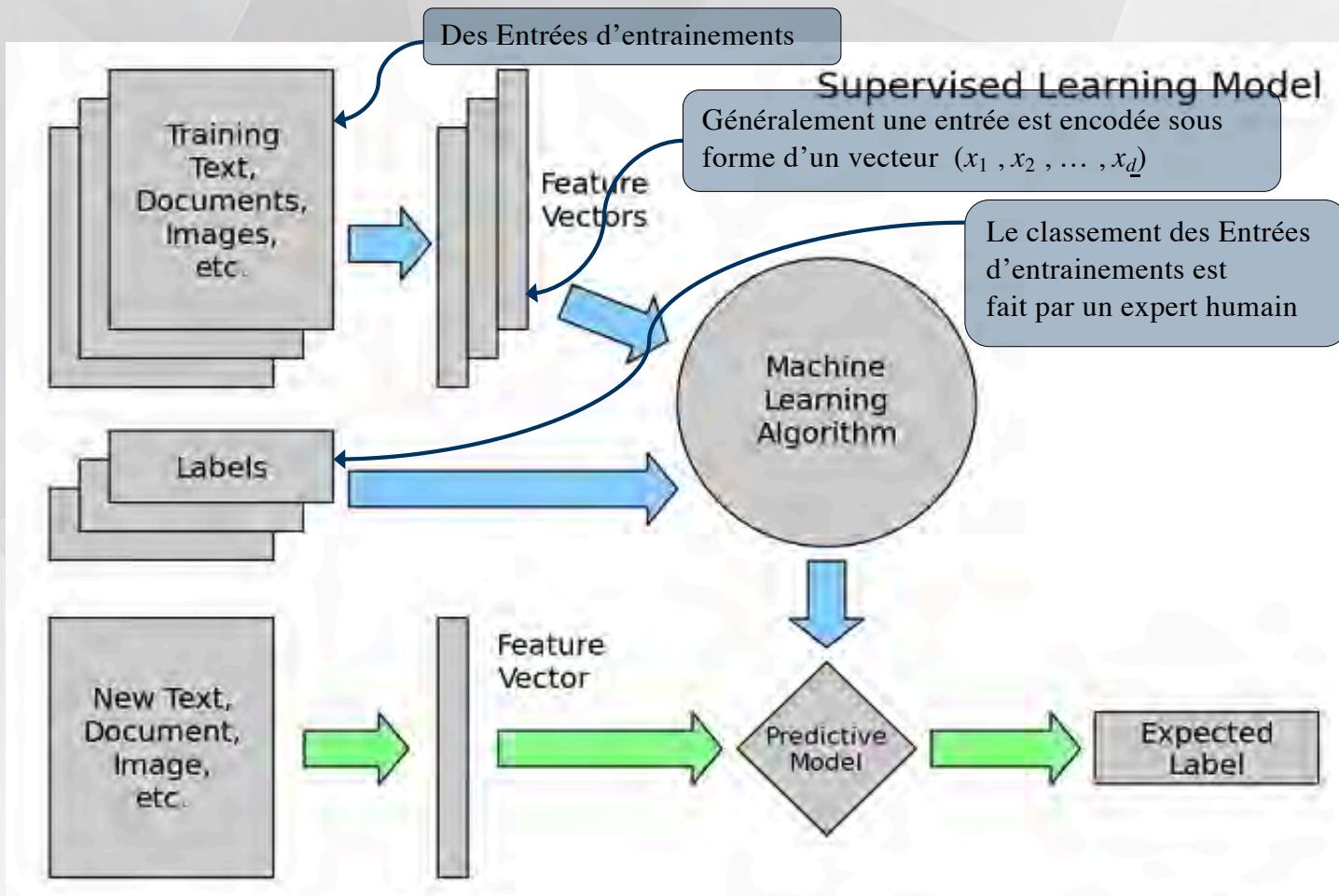
*L'apprentissage se fait
à partir d'exemples ou
d'interactions avec l'environnement*



Si on souhaite chercher à valoriser les données en situation « Big Data », en gros il y a deux cas

- Possible
 - Le problème est bien défini: on sait ce qu'on cherche
Exemples:
 - Netflix
 - Geovoxel
 - Watson
 - Très difficile
 - Il y a de l'information dans nos données,
 - ... On veut aller la chercher
 - Quelle information au juste ?
 - Toute l'information !!!!
 - Oui, mais là ...

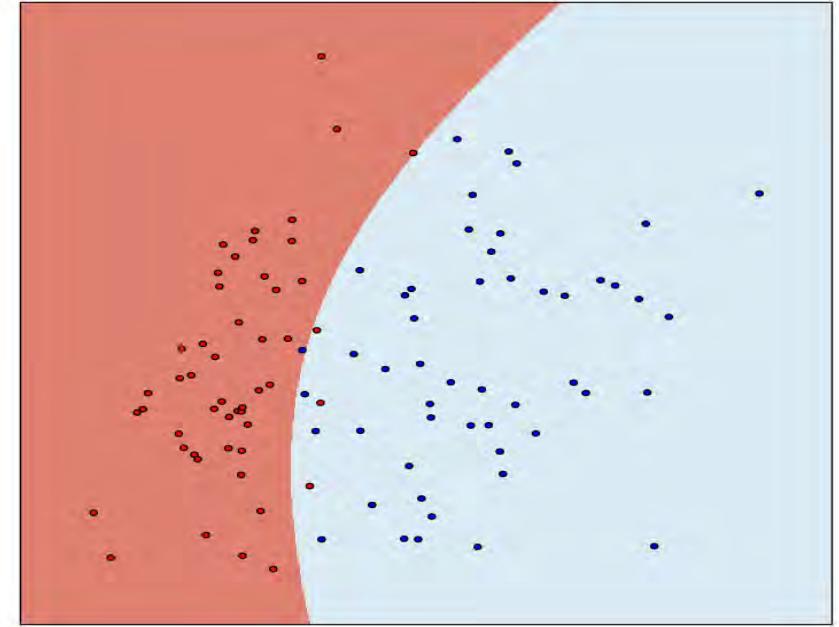
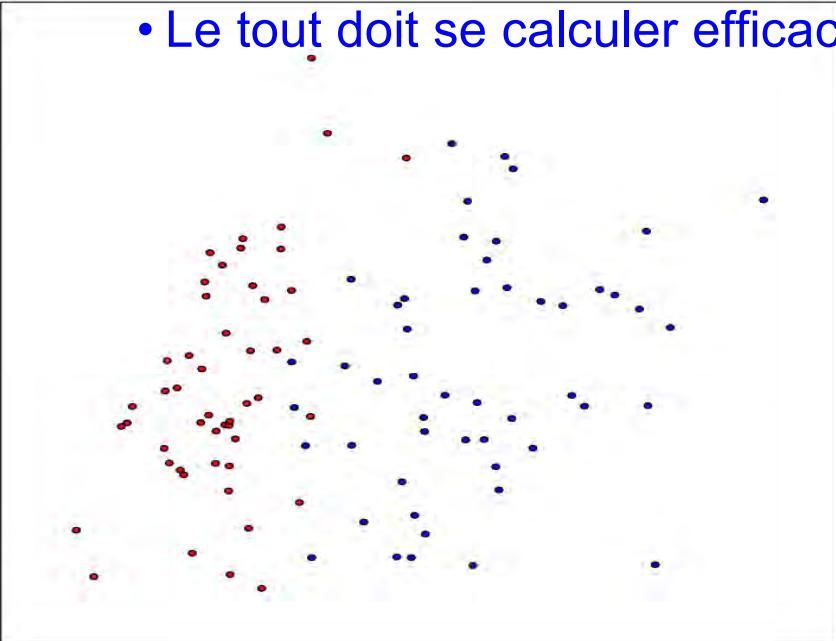
L'apprentissage supervisé



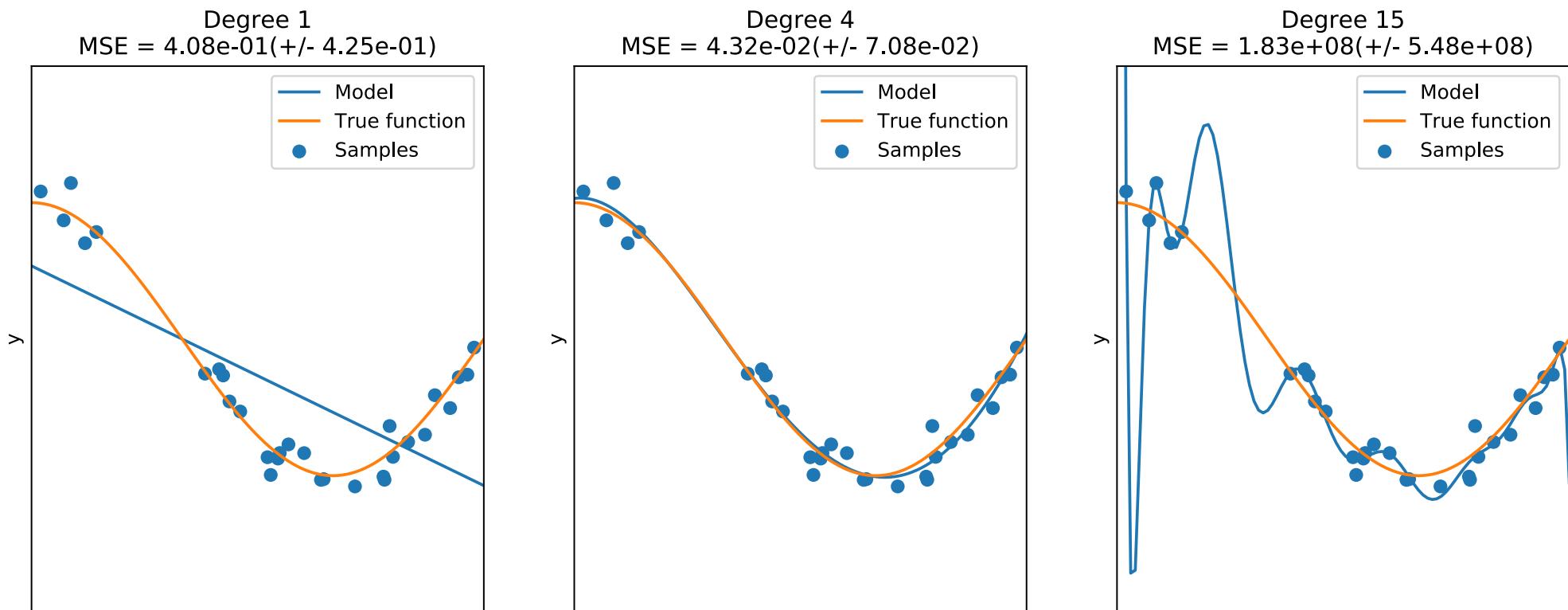
La tâche d'apprentissage en pratique

- Chercher un prédicteur h qui fait peu d'erreurs sur l'ensemble d'entraînement tout en évitant le surapprentissage (overfitting) qui résulterait d'une correspondance trop parfaite des données d'apprentissage

• Le tout doit se calculer efficacement !!



Un exemple de la problématique du sur/sous apprentissage



Un exemple d'algorithme d'apprentissage

Le réseau de neurones

Let consider a neural network architecture with one hidden layer

$$h(\mathbf{x}) = \text{sigm}(\mathbf{b} + \mathbf{W}\mathbf{x}), \quad \text{and} \quad f(h(\mathbf{x})) = \text{softmax}(\mathbf{c} + \mathbf{V}h(\mathbf{x})).$$

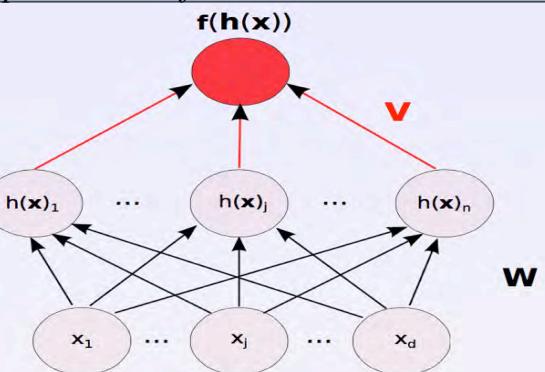
$$\min_{\mathbf{W}, \mathbf{V}, \mathbf{b}, \mathbf{c}} \left[\underbrace{\frac{1}{m} \sum_{i=1}^m -\log (f_{y_i^s}(\mathbf{x}_i^s))}_{\text{source loss}} \right].$$

NN find a new encoding of the data,
a more suitable representation for the task.

where $f_y(\mathbf{x})$ denotes the conditional probability that
the neural network assigns \mathbf{x} to class y .

Given a **source sample** $S = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^m \sim (\mathcal{D}_S)^m$,

1. Pick a $\mathbf{x}^s \in S$
2. Update \mathbf{V} towards $f(h(\mathbf{x}^s)) = y^s$
3. Update \mathbf{W} towards $f(h(\mathbf{x}^s)) = y^s$



Recall: entries are encoded as a vector
 (x_1, x_2, \dots, x_d)

The hidden layer learns a **representation** $h(\cdot)$ from which linear hypothesis $f(\cdot)$ can **classify source examples**.

Les réseaux de neurones (suite)

- Se réduit à un problème d'optimisation (non convexe)
- Peut être résolu par descente de gradient
- Danger de surapprentissage ... quoique
- Peut être difficile à bien entraîner pour les non initiés ... c
- Est présentement le meilleur choix d'algorithme d'apprentissage dans nombre de situations
 - Reconnaissance d'images
 - Traitement vidéo
 - Traitement de la langue naturelle
 - Reconnaissance vocale
 - Alpha go
 - Etc...

Let consider a neural network architecture with one hidden layer

$$h(x) = \text{sigm}(\mathbf{b} + \mathbf{W}x), \quad \text{and} \quad f(h(x)) = \text{softmax}(\mathbf{c} + \mathbf{V}h(x)).$$

$$\min_{\mathbf{W}, \mathbf{V}, \mathbf{b}, \mathbf{c}} \left[\frac{1}{m} \sum_{i=1}^m -\log(f_{\mathcal{D}}(x_i^*)) \right].$$

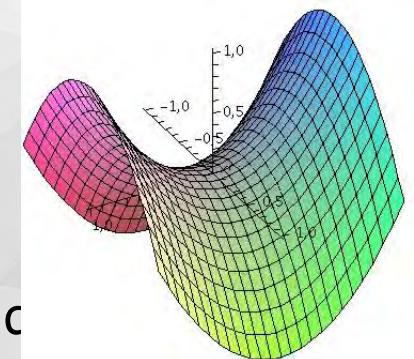
source loss

where $f_i(x)$ denotes the conditional probability that the neural network assigns x to class y .

Given a source sample $\mathcal{S} = \{(x_i^*, y_i^*)\}_{i=1}^m \sim (\mathcal{D}_S)^m$,

1. Pick a $x^* \in \mathcal{S}$
2. Update \mathbf{V} towards $f(h(x^*)) = y^*$
3. Update \mathbf{W} towards $f(h(x^*)) = y^*$

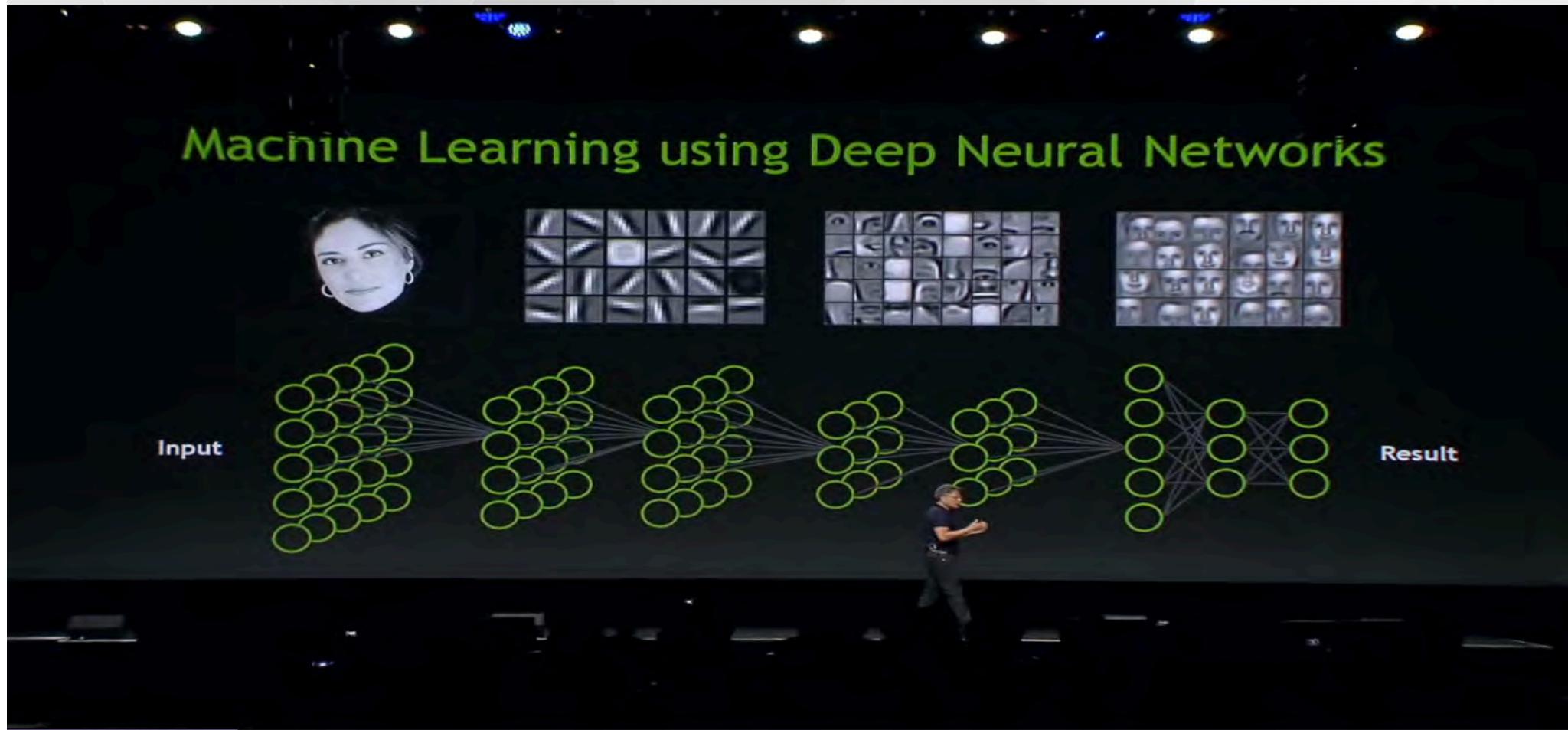
The hidden layer learns a representation $h(\cdot)$ from which linear hypothesis $f(\cdot)$ can classify source examples.



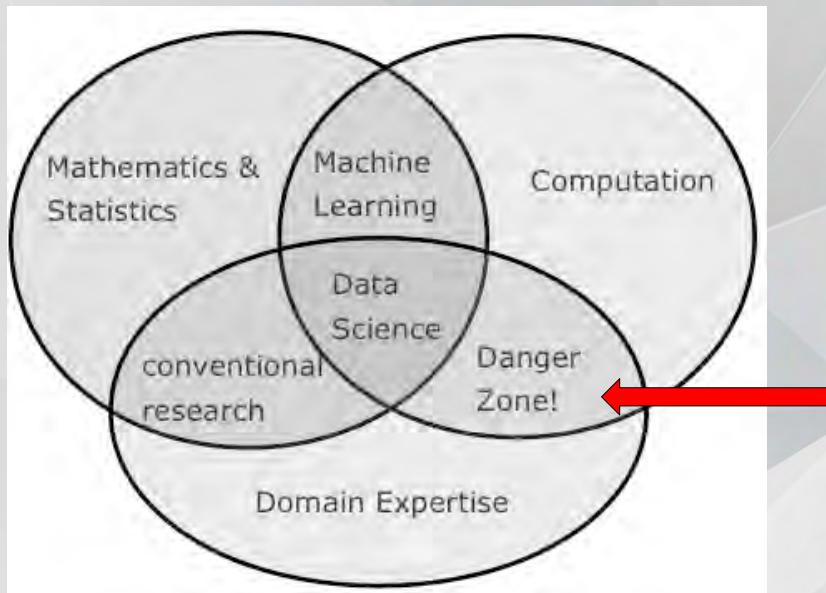
- Est présentement le meilleur choix d'algorithme d'apprentissage dans nombre de situations

- Reconnaissance d'images
 - Traitement vidéo
 - Traitement de la langue naturelle
 - Reconnaissance vocale
 - Alpha go
 - Etc...

Un réseau de neurones apprend une représentation des données qui « rend » la tâche à accomplir plus facile



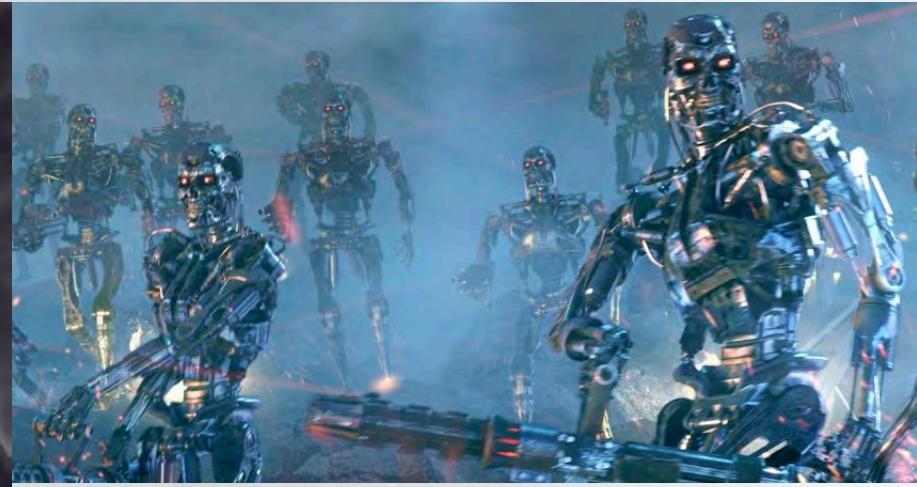
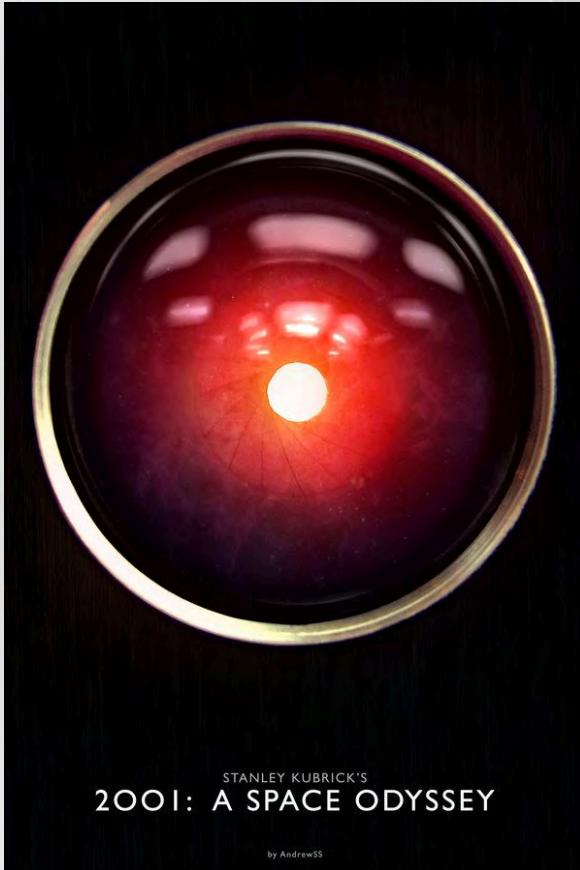
L'apprentissage automatique & les données massives



Le diagramme de Venn de Drew Conway sur le Big Data



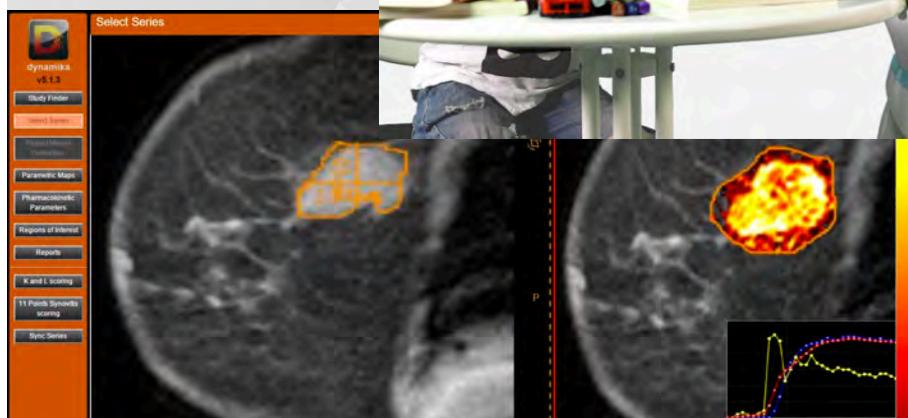
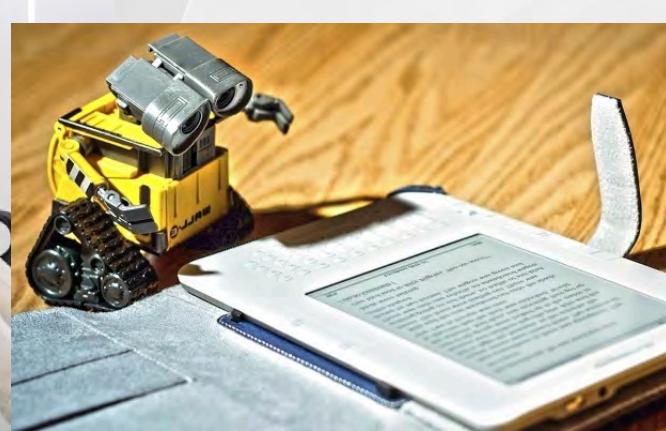
Les dérives potentielles de l'IA, ça pourrait être quoi ?



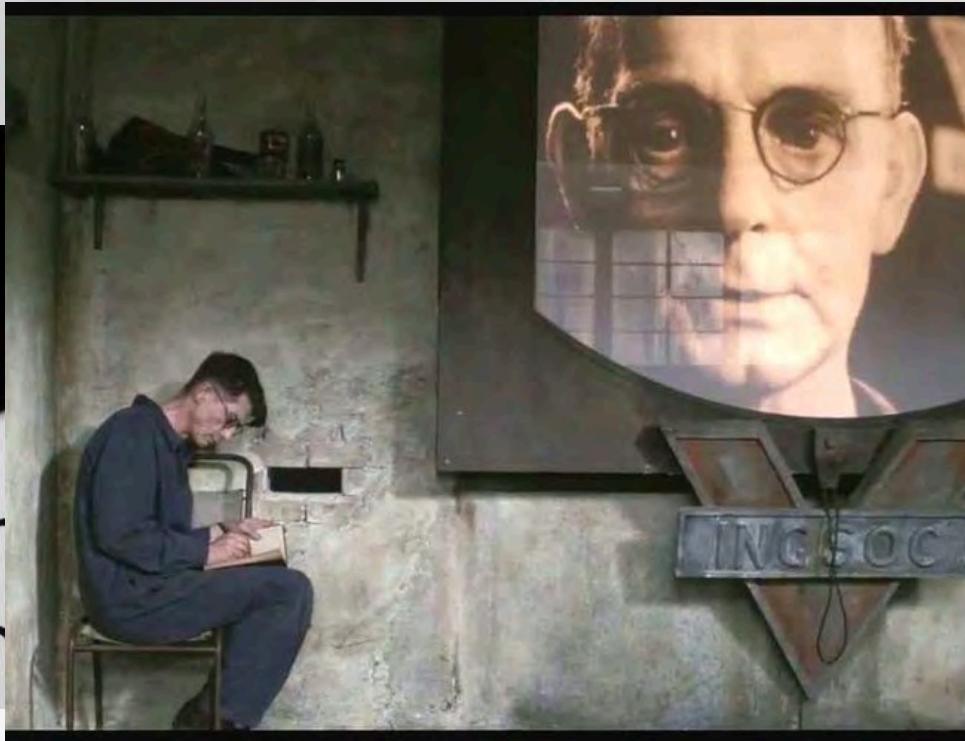
“ THE DEVELOPMENT OF ARTIFICIAL INTELLIGENCE CAN BECOME BOTH THE MOST POSITIVE AND THE MOST TERRIBLE FACTOR FOR MANKIND. WE MUST BE AWARE OF THE DANGER IT POSES ”

Il y a d'autres possibilités de mauvais usages de l'IA qui devraient nous préoccuper, et ce dès maintenant

I'intelligence artificielle dans le futur



L'intelligence artificielle et les données personnelles



En dénominalisant, on ne peut espérer une confidentialité parfaite

- Le cas de Netflix
- Le cas Sweeney-2000

- Informations médicales sur 135 000 employés de l'état du Massachusetts.
- Version anonyme partagée pour la recherche.
- Aucune information personnelle, mais certaines caractéristiques individuelles.
- À l'aide d'une liste des voteurs, Dr. Latyana Sweeney identifie William Weld, alors gouverneur de l'état, et obtient donc accès à son historique médical.

« According to the Cambridge Voter list, six people had his particular birth date; only three of them were men; and, he was the only one in his 5-digit ZIP code. »



Droit individuel à la confidentialité et intérêt collectif

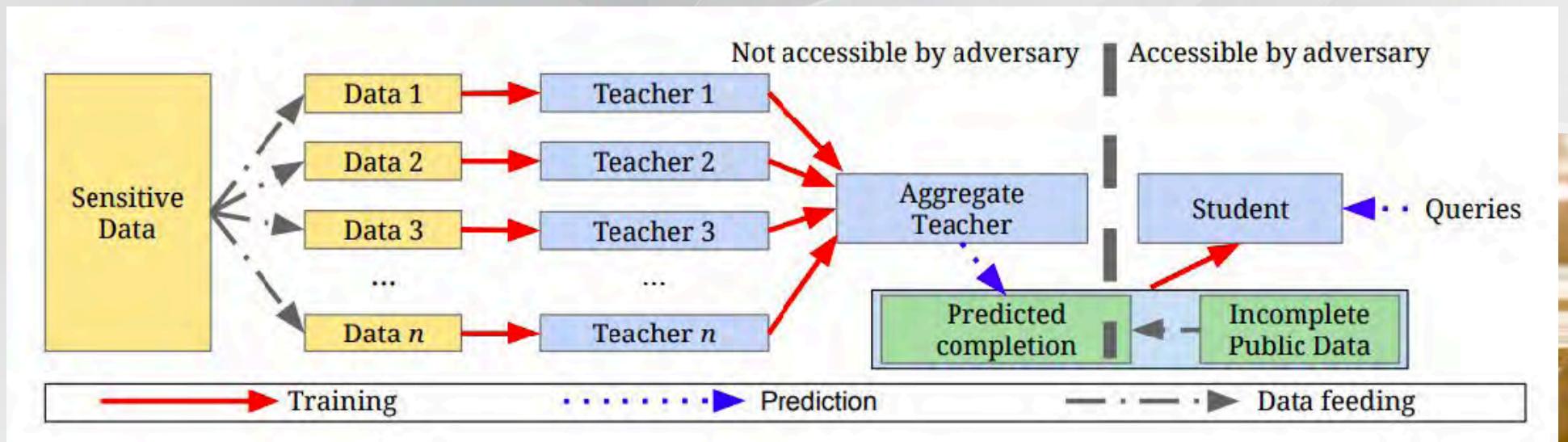
- Renoncer aux données, c'est se couper de grandes possibilités !
- On doit chercher un compromis entre la protection du citoyen et l'intérêt collectif.
- On doit aussi avoir le réflexe de conserver nos données et les voir comme « bien public »
 - La Société d'assurance automobile du Québec (SAAQ) a renoncé à son projet « Ajusto »
 - Les données des produits scannés en épicerie appartiennent à une compagnie privée
 - **Les sciences de la vie représentent un bien public encore plus précieux!!!**
 - Les données des appareils des soins intensifs sont effacées après 72h

Alors on fait quoi avec nos données « sensibles »?

-
- Pistes de réflexions
 - distinguer un accès aux données restreint à des cliniciens et à des chercheurs d'institutions reconnues et un accès sous forme de données ouvertes
 - Prendre toutes les précautions pour que les données ne puissent « sortir » et pour que seuls les algorithmes « voient » les données
 - S'il y a un ou des partenaires privés associés au projet, on partage nos découvertes, pas nos données
 - s'assurer que la population est bien au courant de comment les données sont colligées, sécurisées et de ce qui est fait avec. Mettre le citoyen « dans le coup »
 - prévoir à l'avance comment gérer une situation où il y aurait fuites de données afin de protéger au mieux les individus qui verraient leurs vies privées ainsi compromises

Malgré tout ça, des problèmes subsistent

Un réseau de neurones encode les données sur lesquelles il a été entraîné



En apprentissage automatique, les choses peuvent mal aller

Corrélation n'est pas causalité

Comment les données ont-elle été collectées ?

Les données doivent être obtenues de façon iid.
i.e., chaque exemple des données d'entraînement est supposé
avoir été obtenue par une pige d'une certaine distribution inconnue et
qui soit indépendante des autres données obtenues

Idem pour les exemples « à venir »



En apprentissage automatique, les choses peuvent mal aller

Corrélation n'est pas causalité

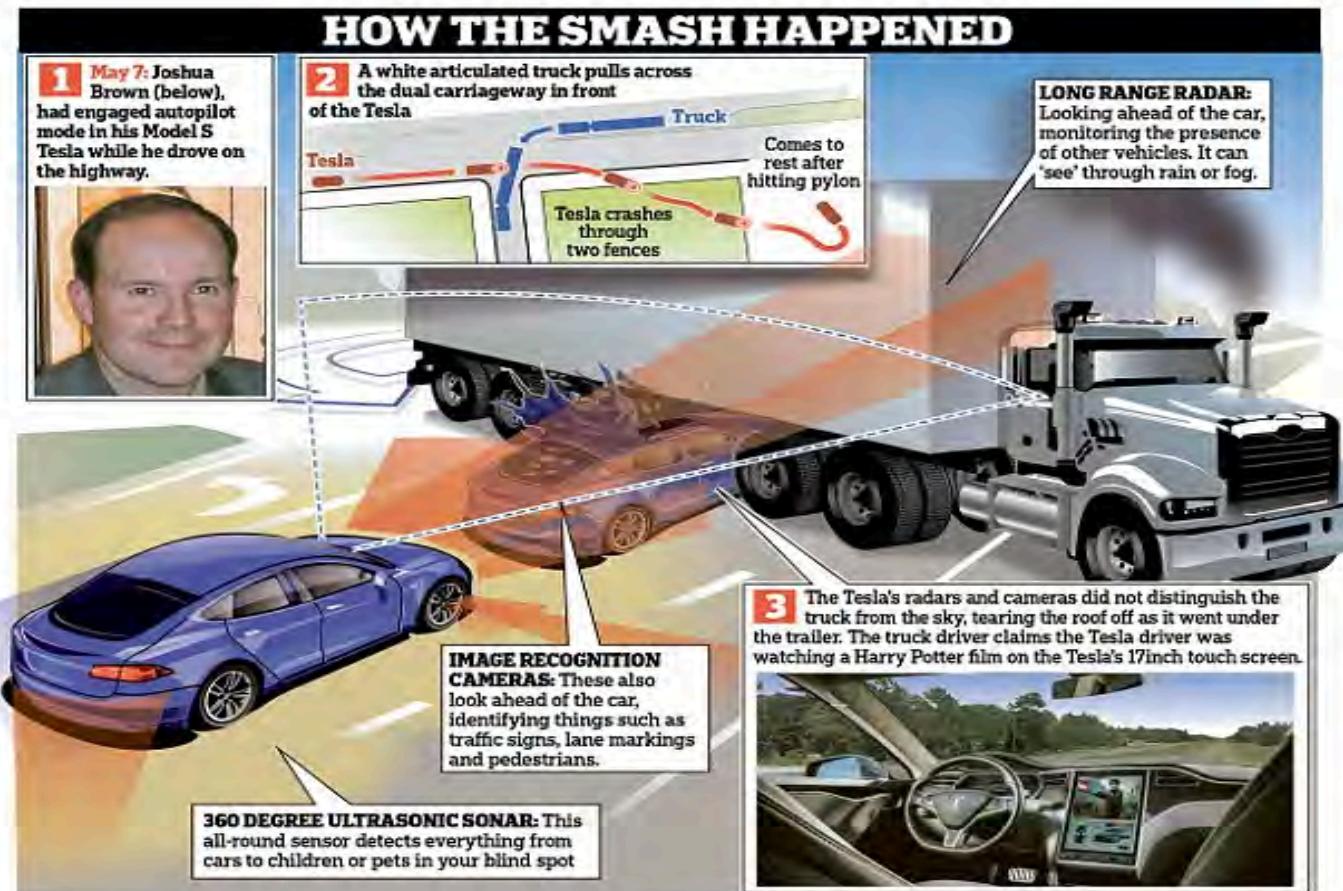
La recherche basée sur les données massives et l'IA est « data driven », ce qui est très différents de la recherche traditionnelle

On peut envisager qu'il y a ici un problème pour les comités d'éthiques car on ne peut dire dès le départ d'un projet tout ce qu'on y fera, le tout pouvant dépendre de ce qu'on trouvera dans les données



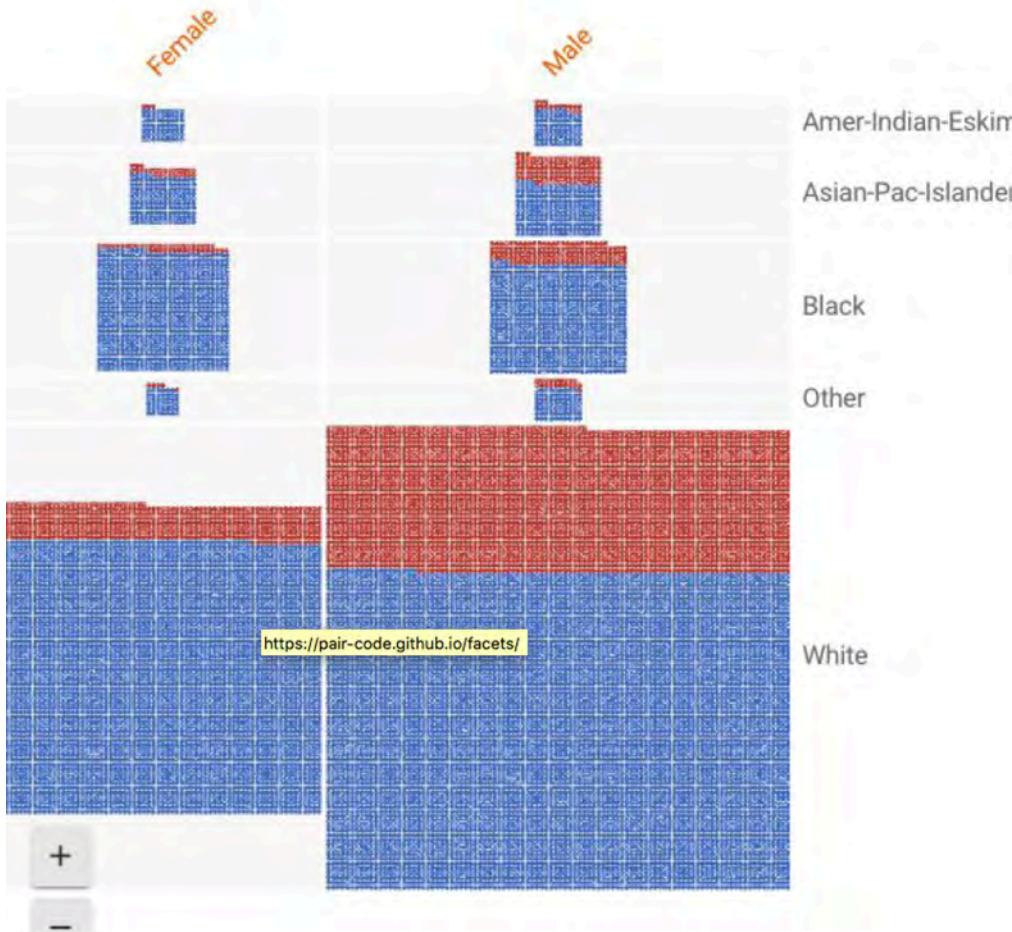
En apprentissage automatique, les choses peuvent mal aller

Événements rares



Défis en lien avec l'intelligence artificielle

Équité



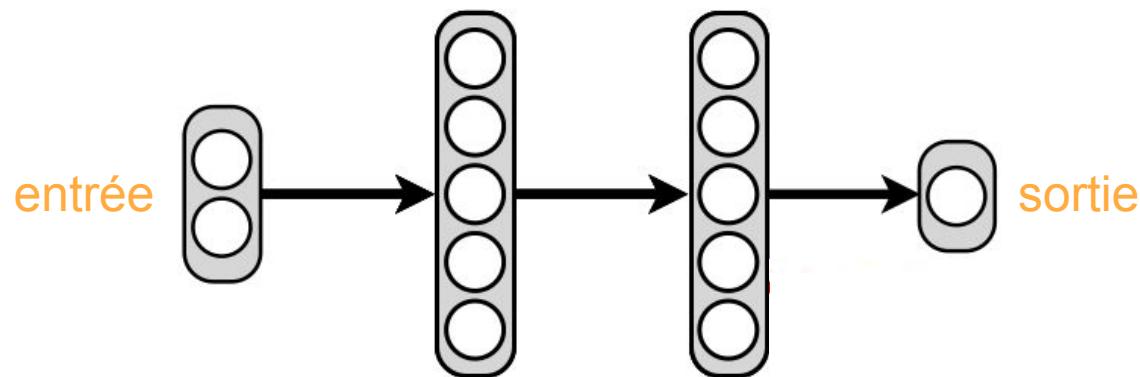
**L'IA est aussi « bonne »
que les données sur
lesquelles elle a été
entraînée**

**En situation de « gouvernance
algorithme » ou pour toute forme de
décisions critiques qu'on relègue à une
IA, c'est essentiel !!!**

Défis en lien avec l'intelligence artificielle

Équité

Une solution possible au manque d'équité:



Domain-Adversarial Training of Neural Networks

Ganin, Ustinova, Ajakan, Germain, Larochelle, Laviolette, Marchand, Lempitsky, 2017

Autres défis de l'intelligence artificielle

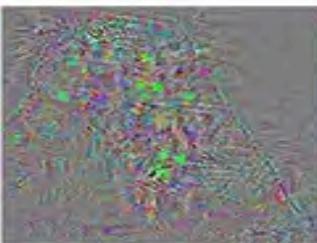
La robustesse aux attaques atagonistes

Le cas de TAY, l'intelligence artificielle “innocente” de Microsoft



Autres défis de l'intelligence artificielle

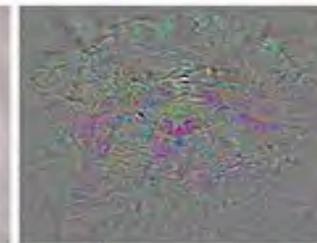
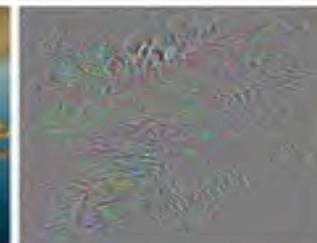
La robustesse aux attaques antagonistes



correct

+distort

ostrich



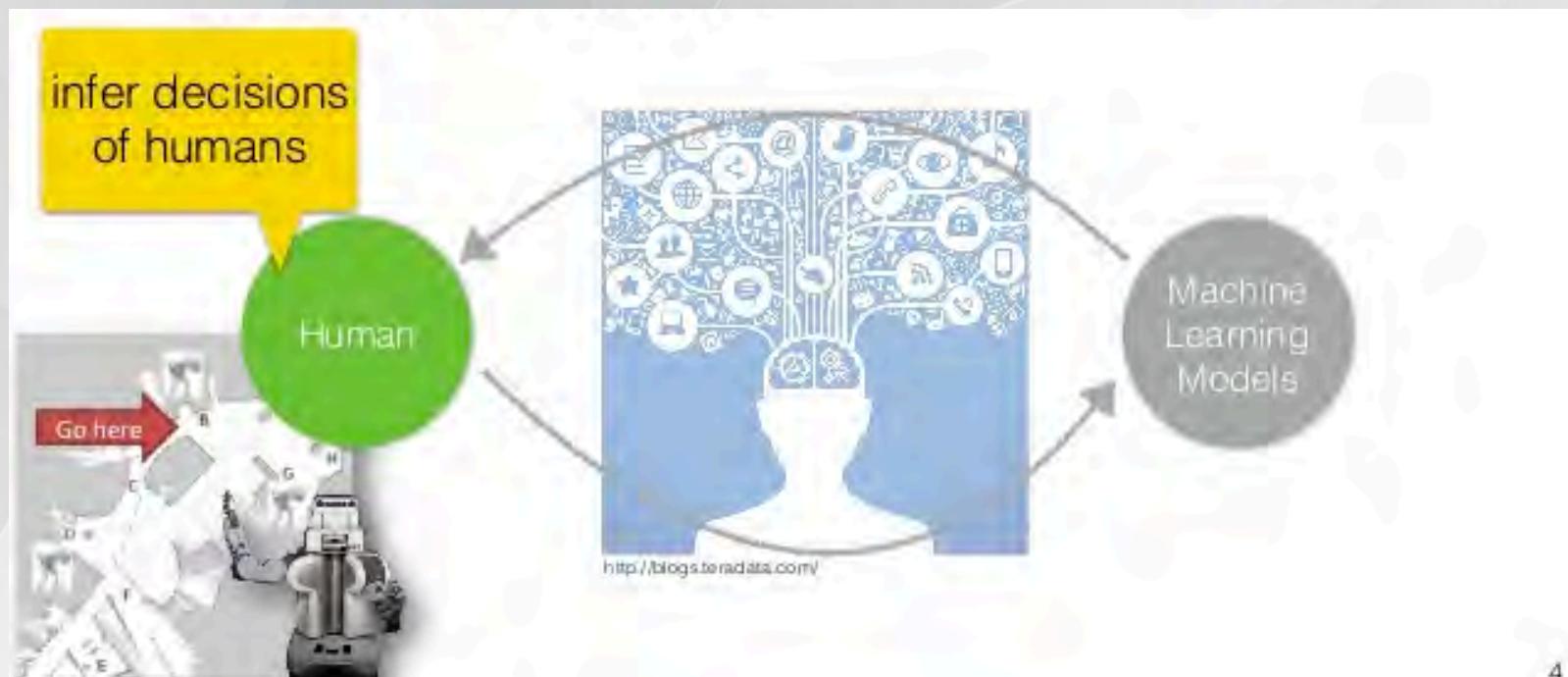
correct

+distort

ostrich

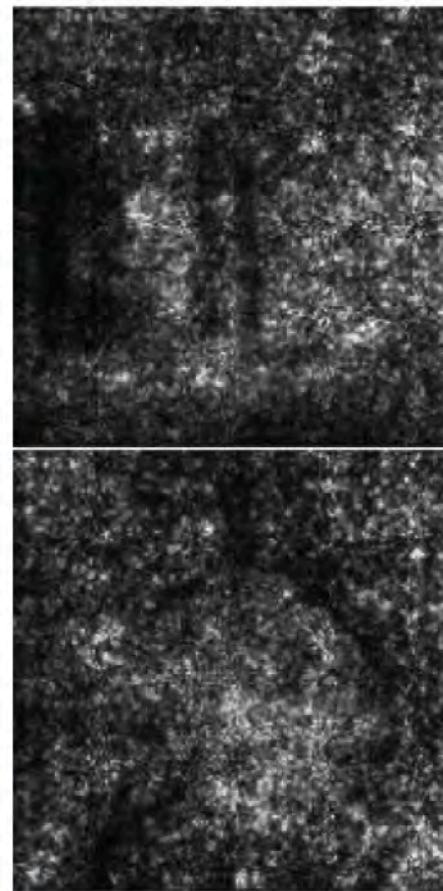
Autre défi de l'intelligence artificielle L'interprétabilité

Pour certaines tâches, l'humain a besoin de comprendre la décision de l'IA



Autres défis de l'intelligence artificielle

L'interprétabilité



Autres défis de l'intelligence artificielle

Prévoir ses impacts sur la société



Questions qu'on est en droit de se poser concernant l'IA

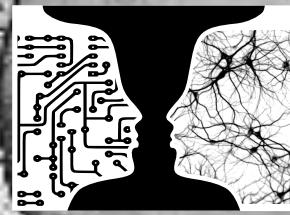
- Qu'est-ce qu'on se donne le droit de faire (et qu'est-ce qu'on s'interdit de faire) avec l'IA ?
 - Si on se limite, on s'enlève des possibilités
 - Si on ne se limite pas, on ouvre la voie à des dérives
- Quels vont être les impacts sociaux directs et indirects de l'IA?
- Jusqu'où sommes-nous prêts à laisser l'IA prendre des décisions pour nous ?
- ...



Pour un développement responsable de l'IA il serait souhaitable de

- 1. Élaborer un cadre éthique pour le développement et le déploiement de l'IA ;**
- 2. Orienter la transition numérique afin que tous puissent bénéficier de cette révolution technologique ;**
- 3. Ouvrir un espace de dialogue pour réussir collectivement un développement inclusif, équitable et écologiquement soutenable de l'IA.**





Groupe de
Recherche en
Apprentissage
Automatique de
Laval



crdm.ul
CENTRE DE RECHERCHE
EN DONNÉES MASSIVES
DE L'UNIVERSITÉ LAVAL