Le potentiel et les enjeux de l'intelligence artificielle

Présenté par
Christian Gagné
François Laviolette

Conférence annuelle de TELUS Santé

mars 2018

crdm_ul
CENTRE DE RECHERCHE
EN DONNÉES MASSIVES
DE L'UNIVERSITÉ LAVAL

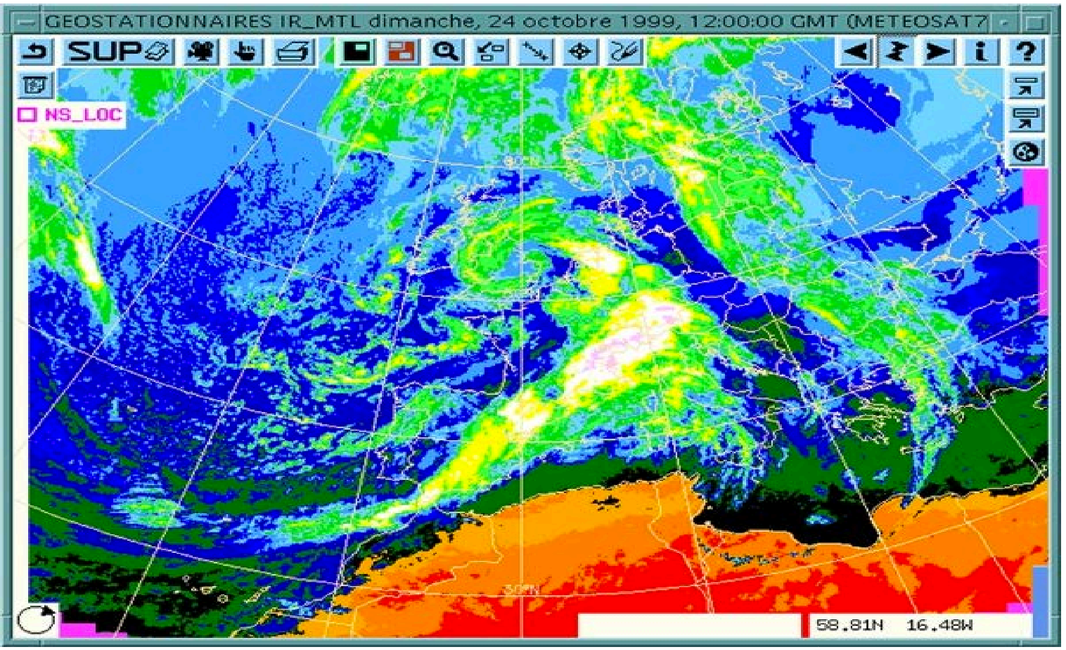
UNIVERSITÉ
LAVAL

Regulating the internet giants

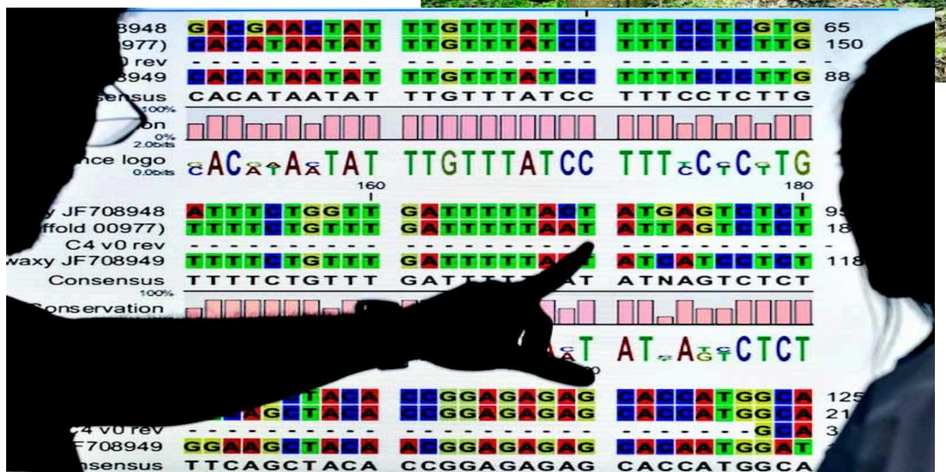
The world's most valuable resource is no longer oil, but data

The data economy demands a new approach to antitrust rules

Nous avons *de plus en plus* de façons nouvelles d'aller chercher *de plus en plus* de données!!!



Ma Conduite INTACT Ass.



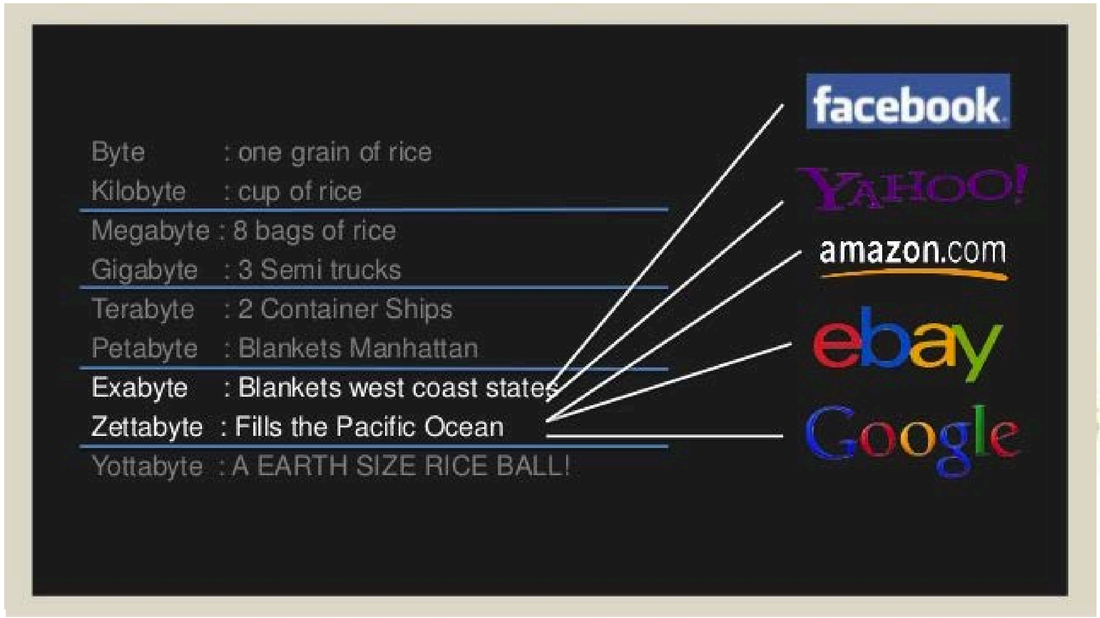
Qu’est-ce que le « Big Data » ?

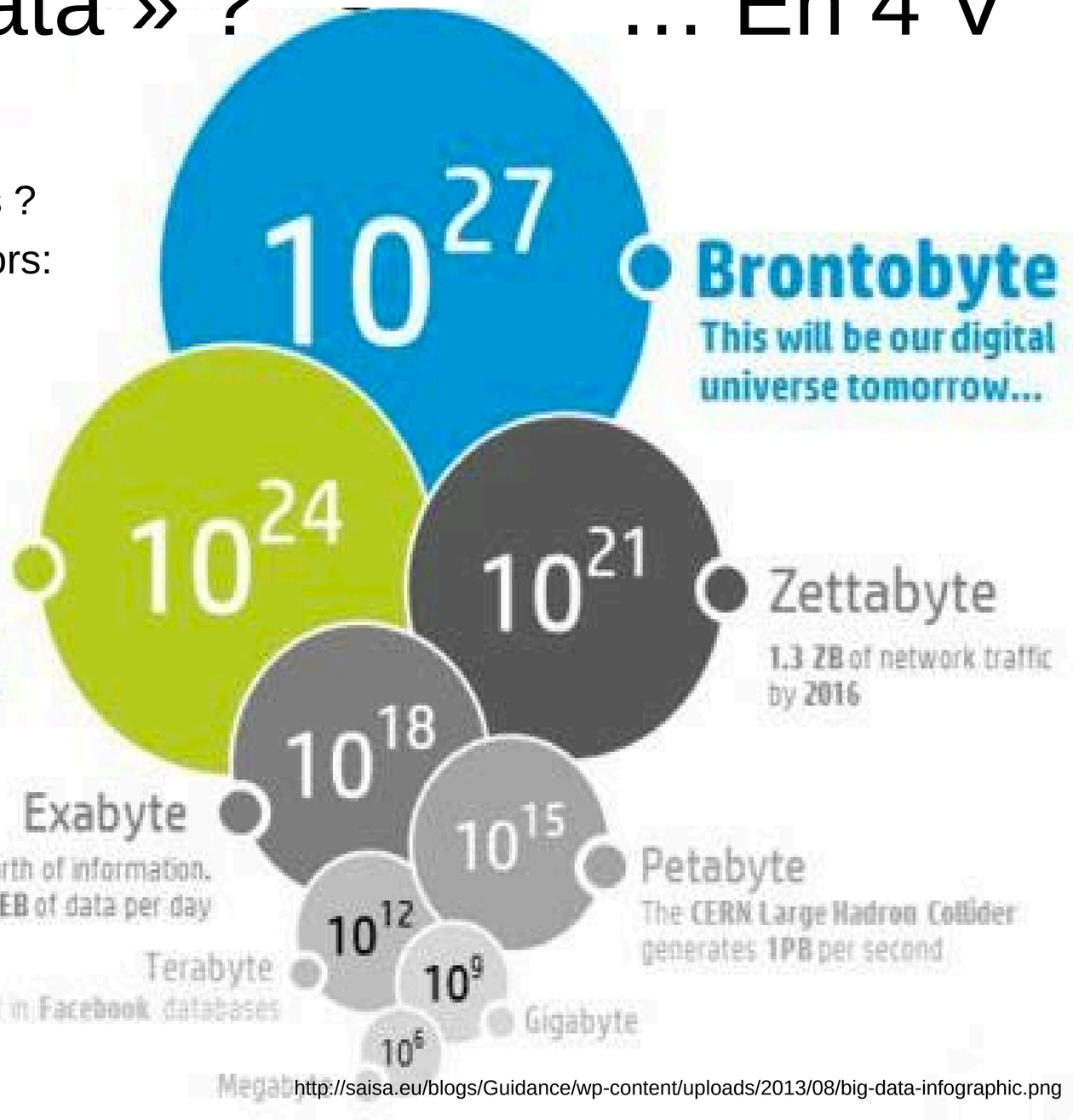
- D’abord, quel est le bon terme français?
  - Les Français parlent de mégadonnées
  - Nous avons choisi l’expression données massives,

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:





Comparison given by David Wellman
http://fr.slideshare.net/dwellman/what-is-big-data-24401517/

http://saisa.eu/blogs/Guidance/wp-content/uploads/2013/08/big-data-infographic.png

Qu’est-ce que le « Big Data » ? … En 4 V

- Volume
- Vélocité
- Variété
- Véracité

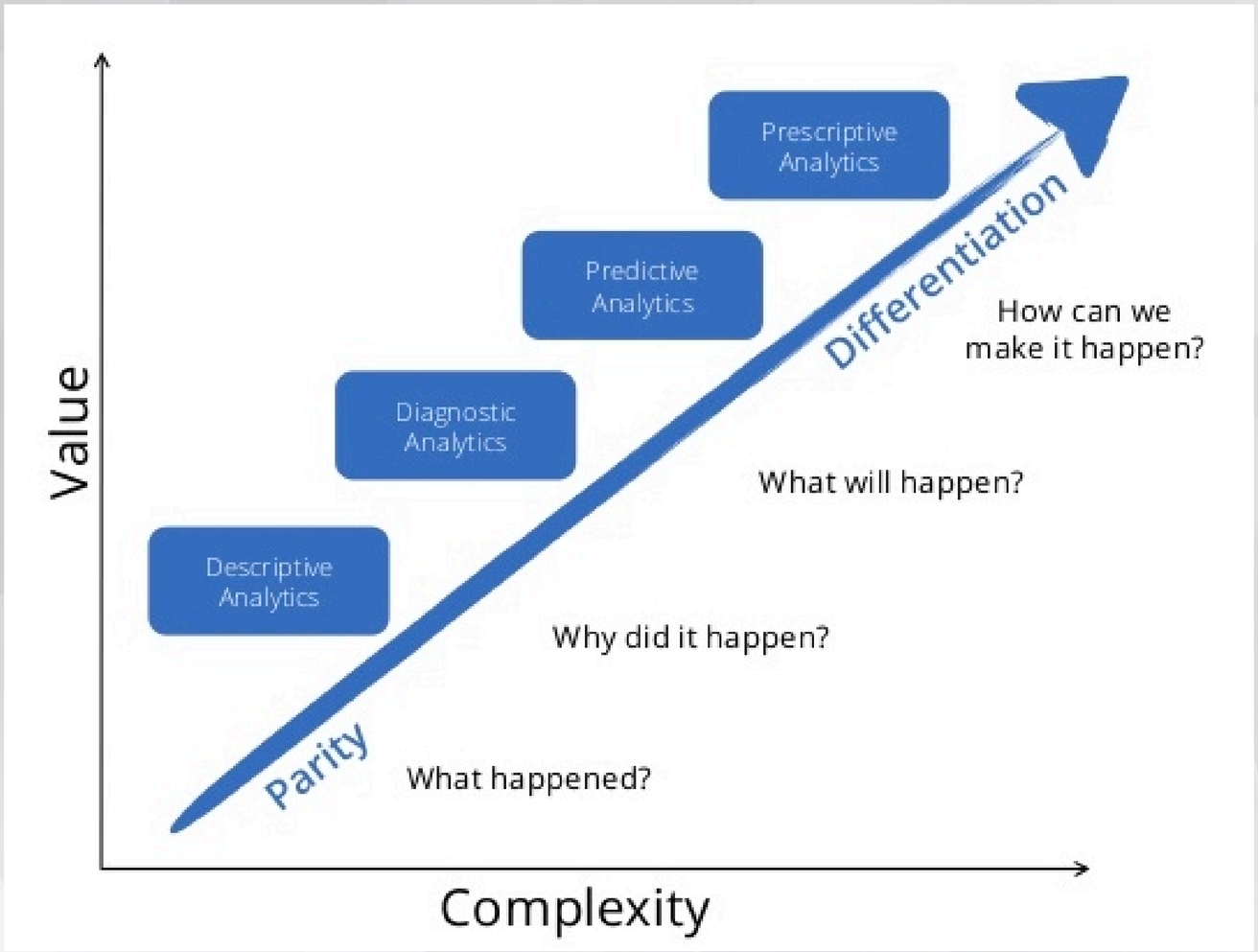
Données provenant de diverses sources non nécessairement structurées

Image, texte, données de senseurs, …

Données provenant de projets différents,

avec des méthodologies non nécessairement compatibles

Quelle est la Valeur qu’on peut tirer du « Big Data » ?

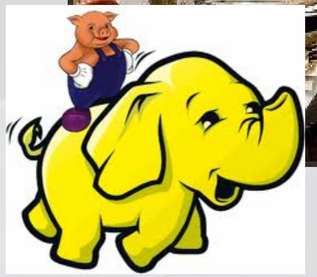


Big data is not about the size of the data, it’s about the value within data

(https://fr.slideshare.net/dwellman/what-is-big-data-24401517)

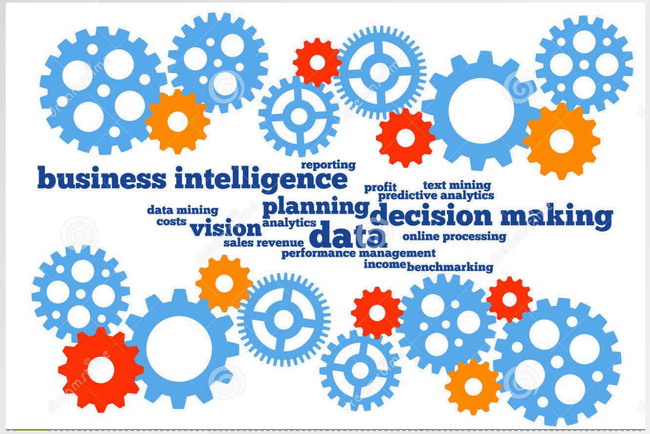
Les défis du « Big Data »

- Les données massives forcent le développement de nouvelles méthodes pour:
  - entreposer et retrouver la donnée
  - effectuer les analyses et autres calculs
  - visualiser l’information
  - réaliser les prises de décisions associées



Spark

Et pour toutes ces tâches, l’apprentissage machine et la recherche opérationnelle sont des outils de prédilection.

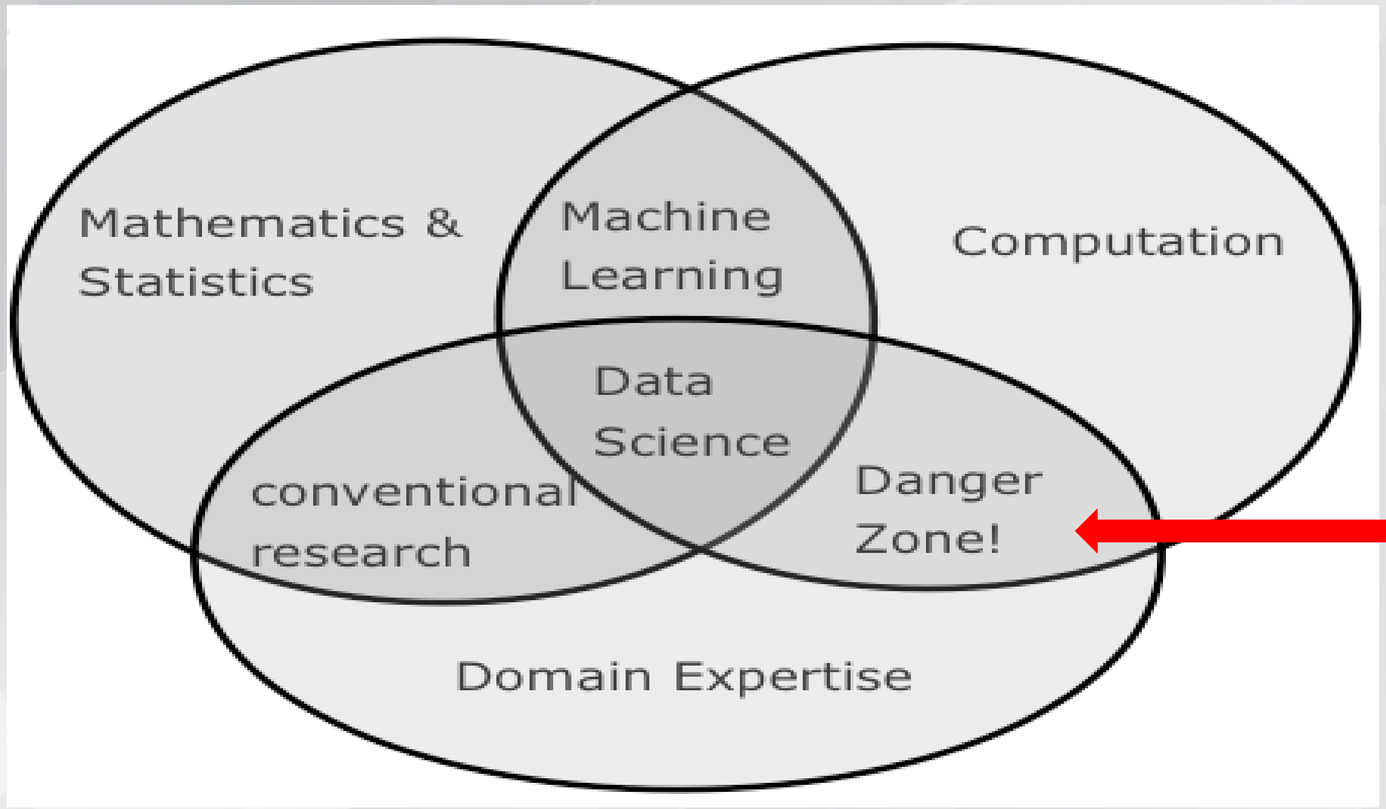


L’ apprentissage machine
vs
la recherche opérationnelle

Apprentissage machine:
Comprendre les informations

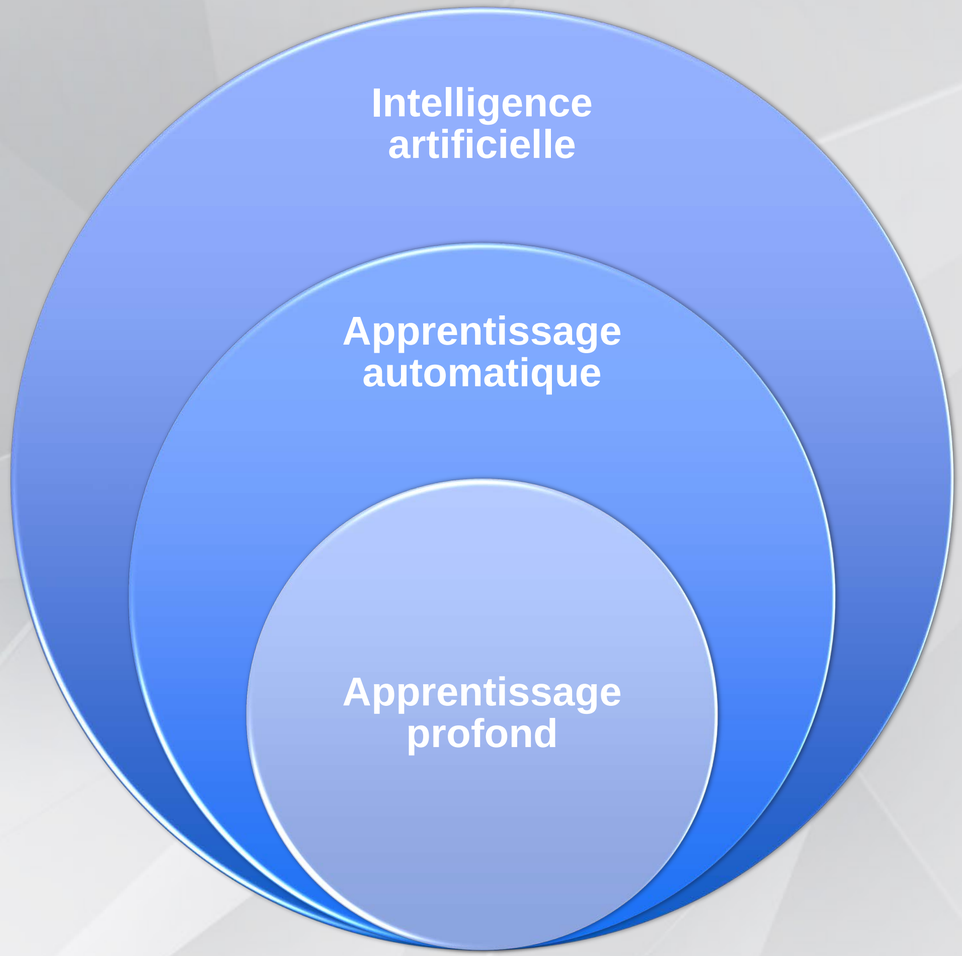
Recherche opérationnelle:
Optimise la prise de décision

L’ apprentissage machine
et
les données massives



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

L’intelligence artificielle et ses apprentissages



Source: « Why Deep Learning Matters and what’s next for Artificial Intelligence », Algorithmia, Novembre 2016

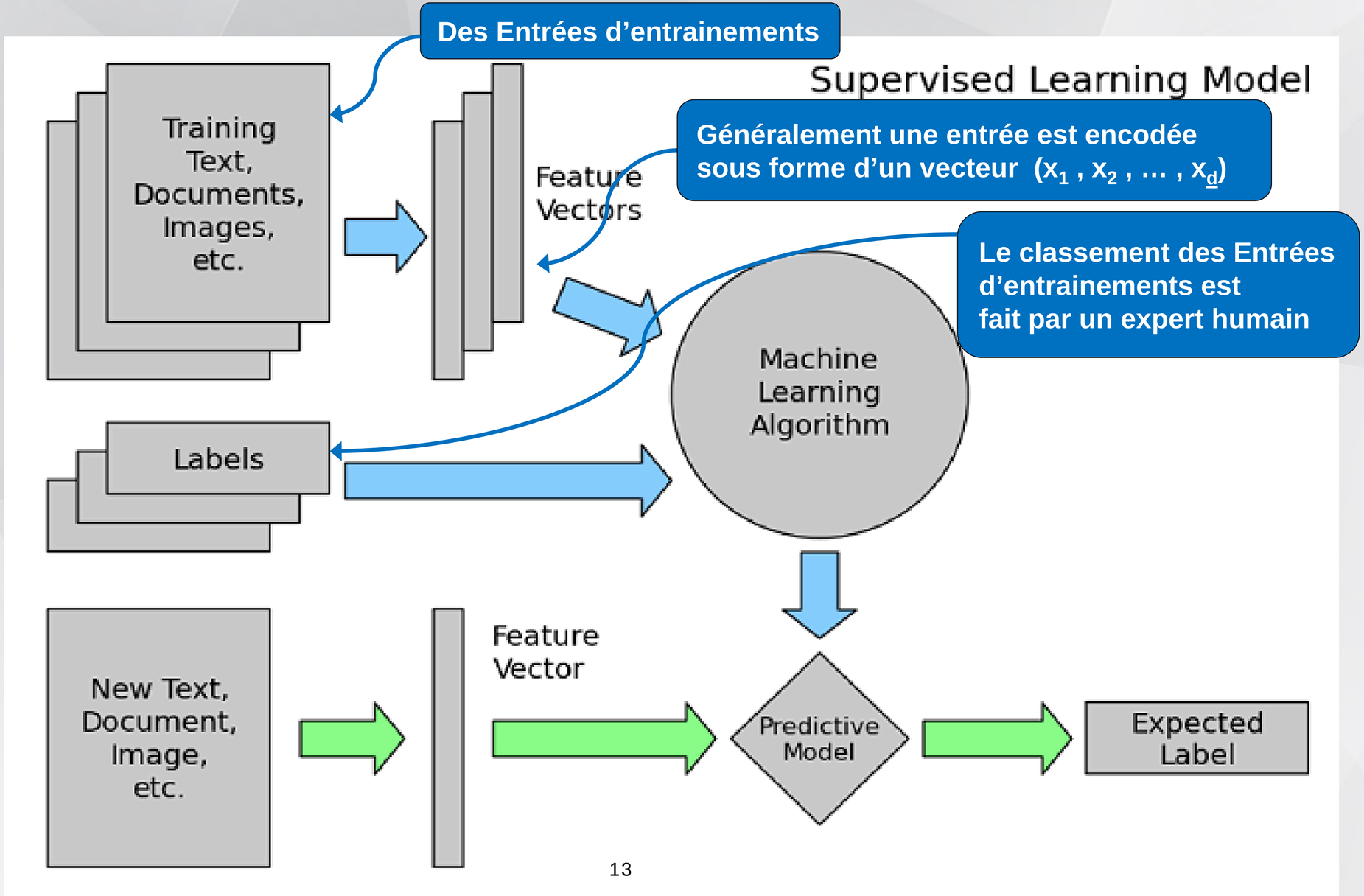
Apprentissage machine101

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

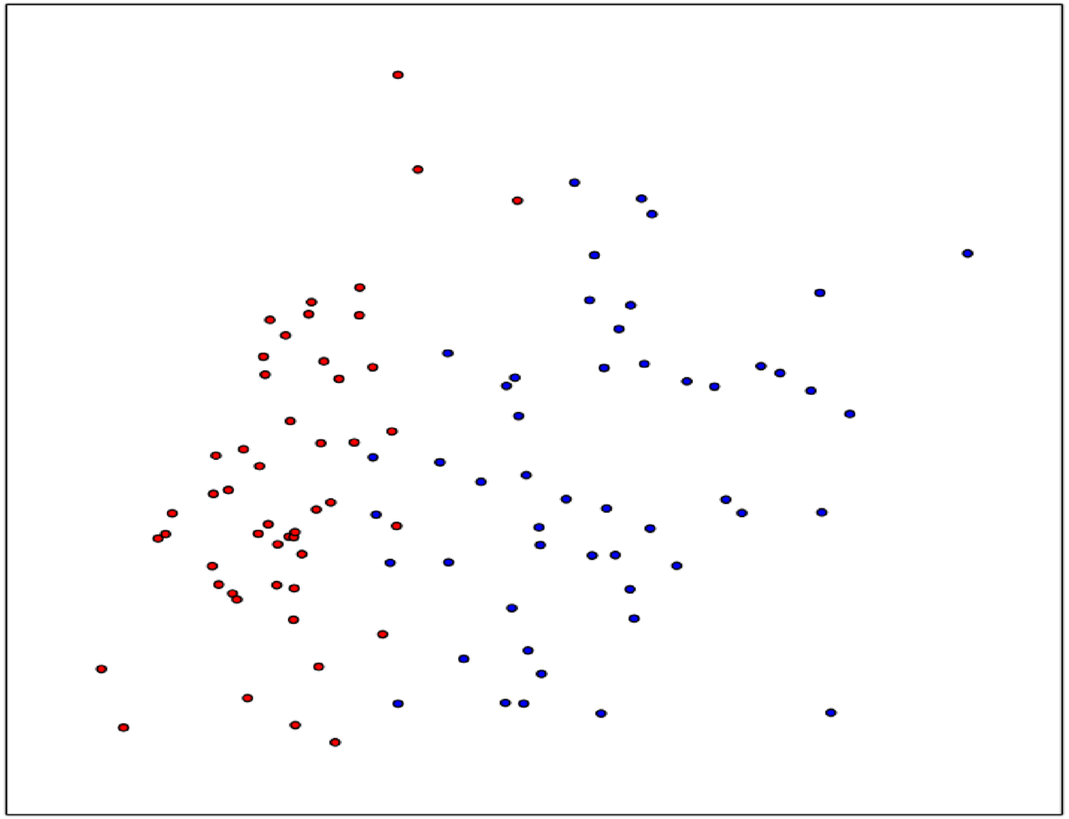
L’apprentissage supervisé

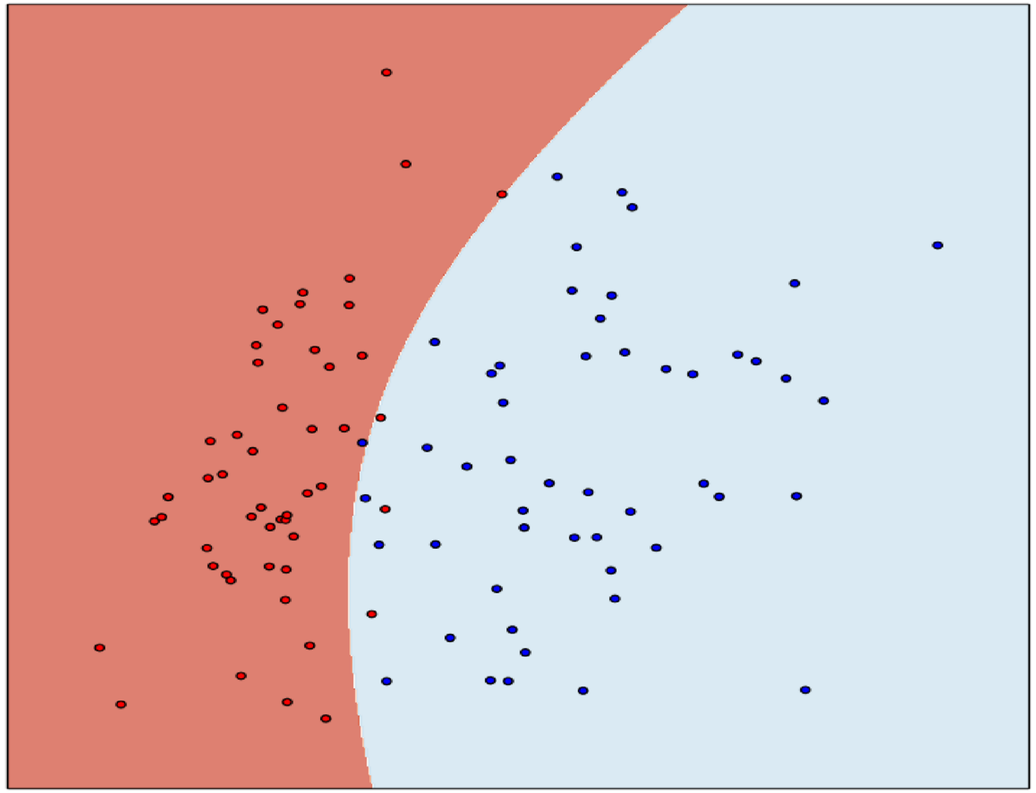


La tâche d’apprentissage en pratique

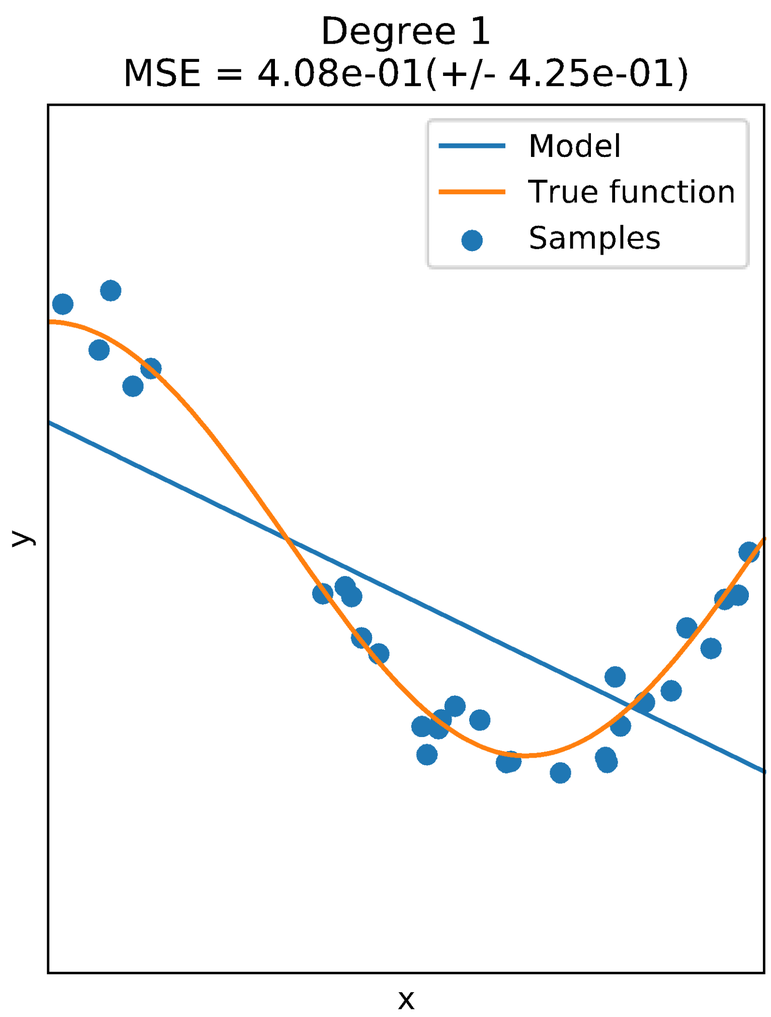
Chercher un classificateur *h* qui fait peu d’erreurs sur l’ensemble d’entraînement tout en évitant le surapprentissage (overfitting) qui résulterait d’une correspondance trop parfaite des données d’apprentissage

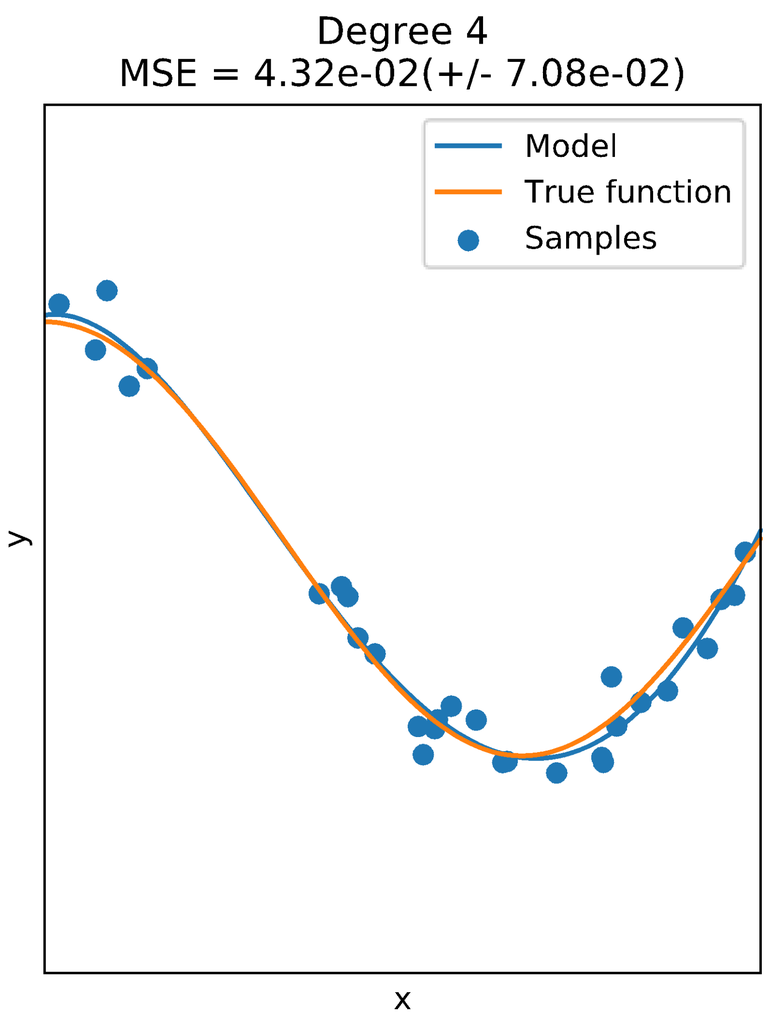
Le tout doit se calculer efficacement !!

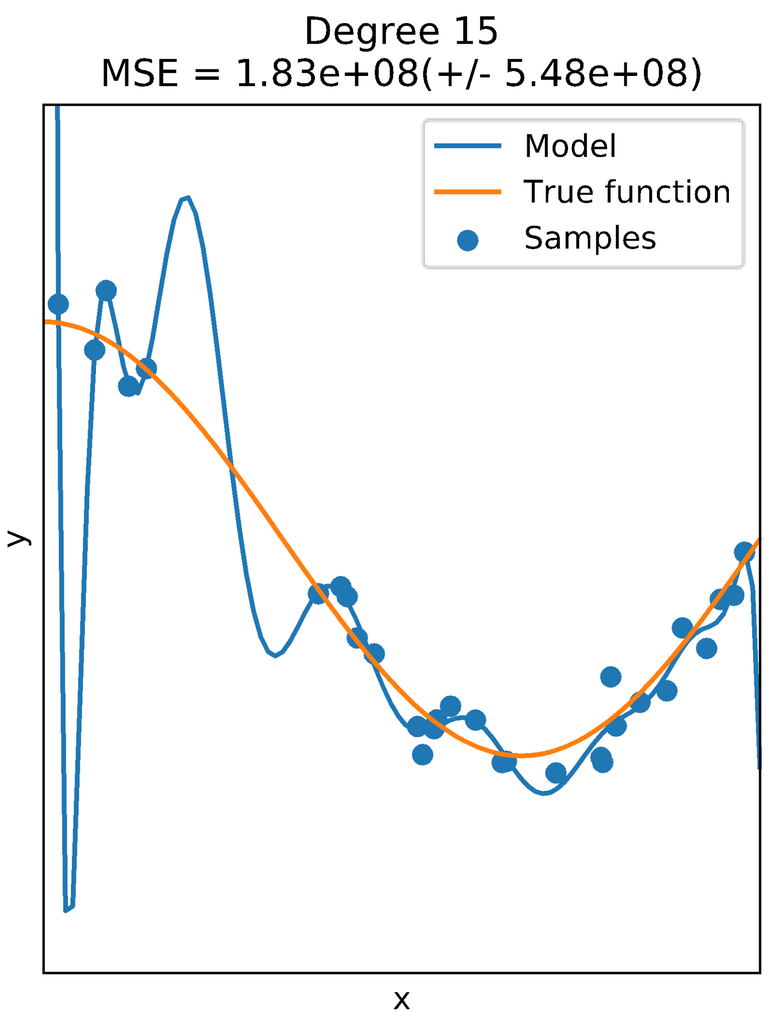




Un exemple de la problématique du sur/sous apprentissage







Un exemple d'algorithme d'apprentissage

Le réseau de neurones

Let consider a neural network architecture with one hidden layer

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

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

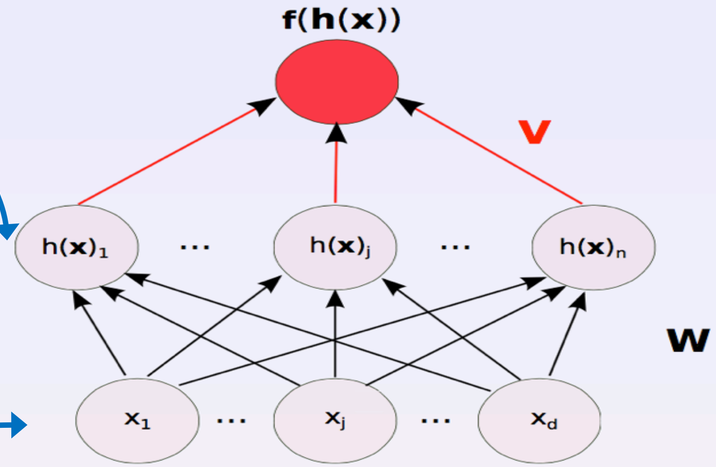
where $f_y(\mathbf{x})$ denotes the conditional probability that the neural network assigns $\mathbf{x}$ to class y.

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 $\mathbf{f}(\mathbf{h}(\mathbf{x}^s)) = y^s$
3. Update $\mathbf{W}$ towards $\mathbf{f}(\mathbf{h}(\mathbf{x}^s)) = y^s$

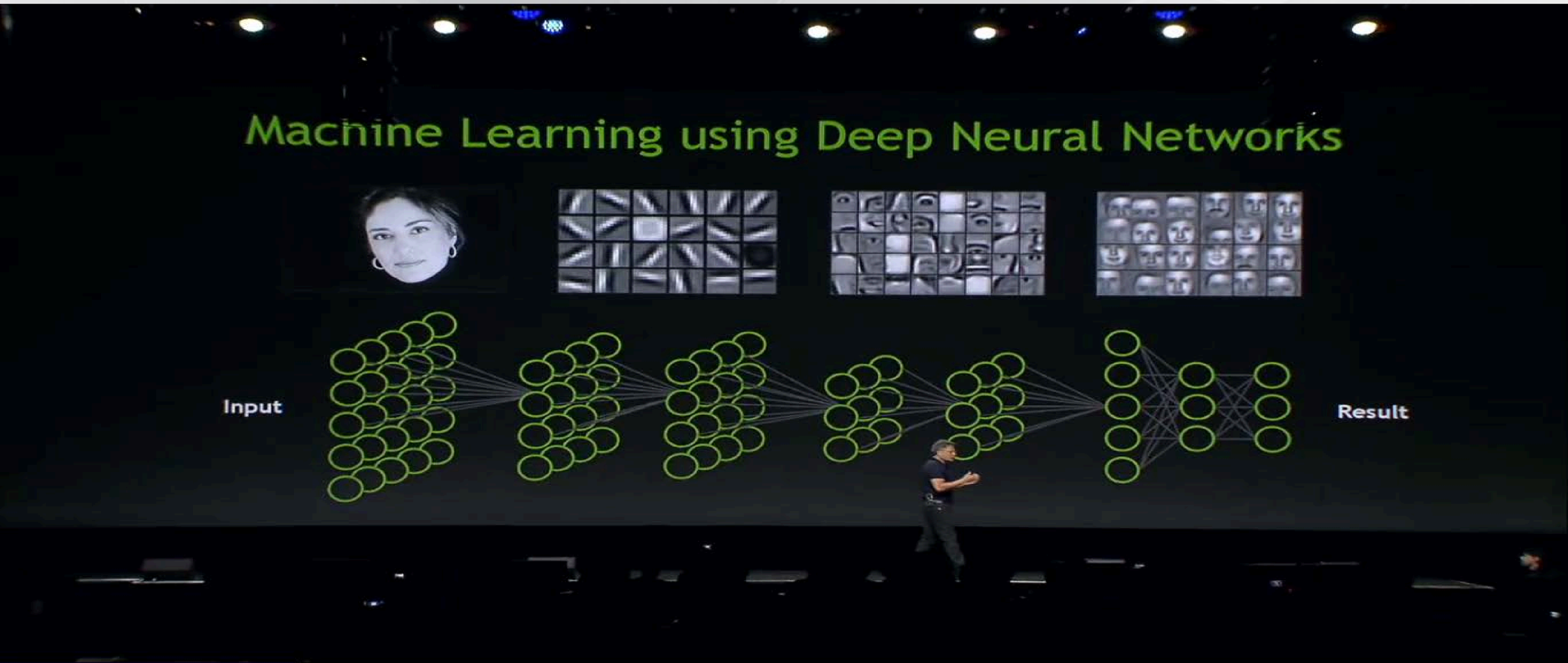
The hidden layer learns a **representation** $\mathbf{h}(\cdot)$ from which linear hypothesis $\mathbf{f}(\cdot)$ can **classify source examples**.

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



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

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

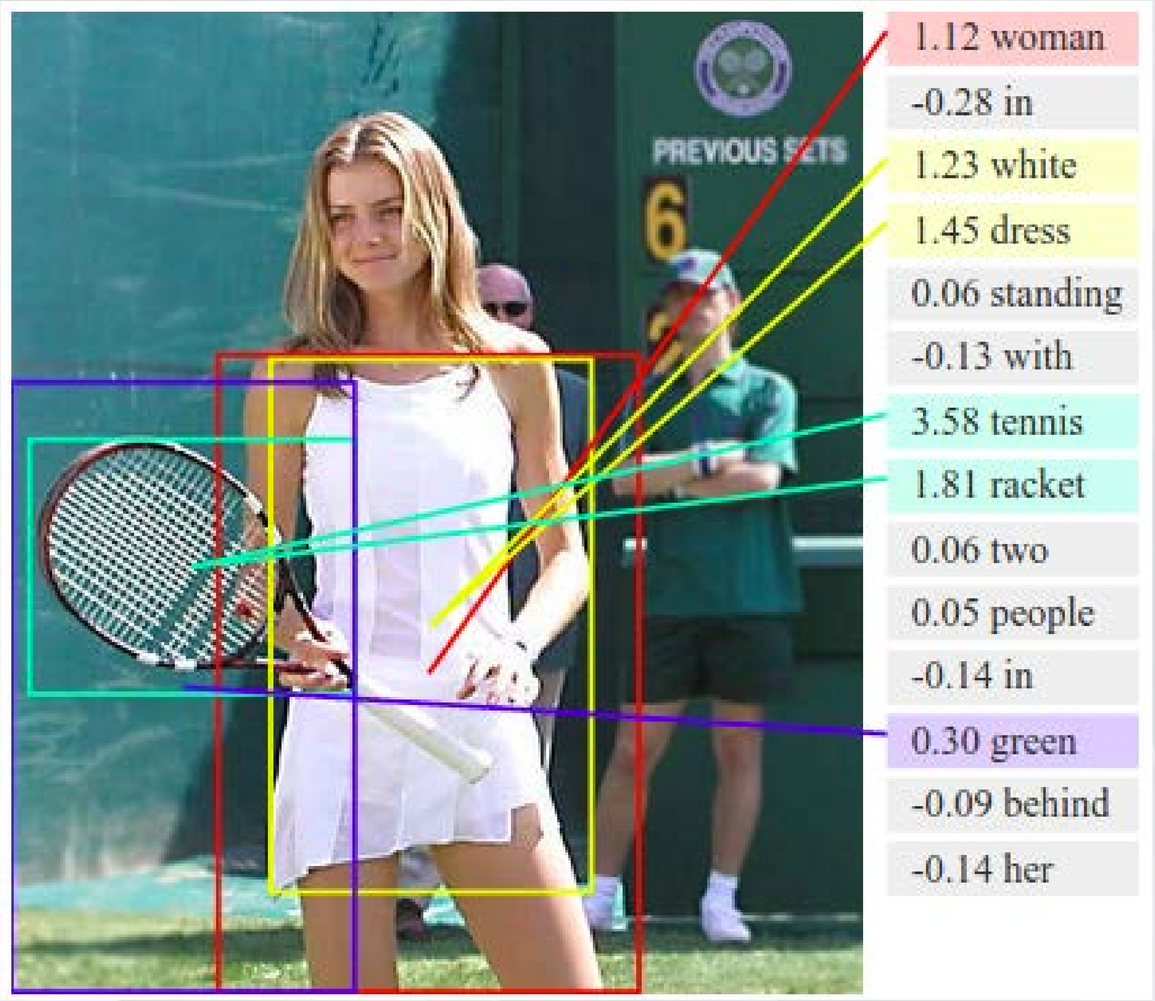


Étiquetage de scènes par des réseaux profonds

Une image vaut mille mots

[Farabet et al. ICML 2012, PAMI 2013]

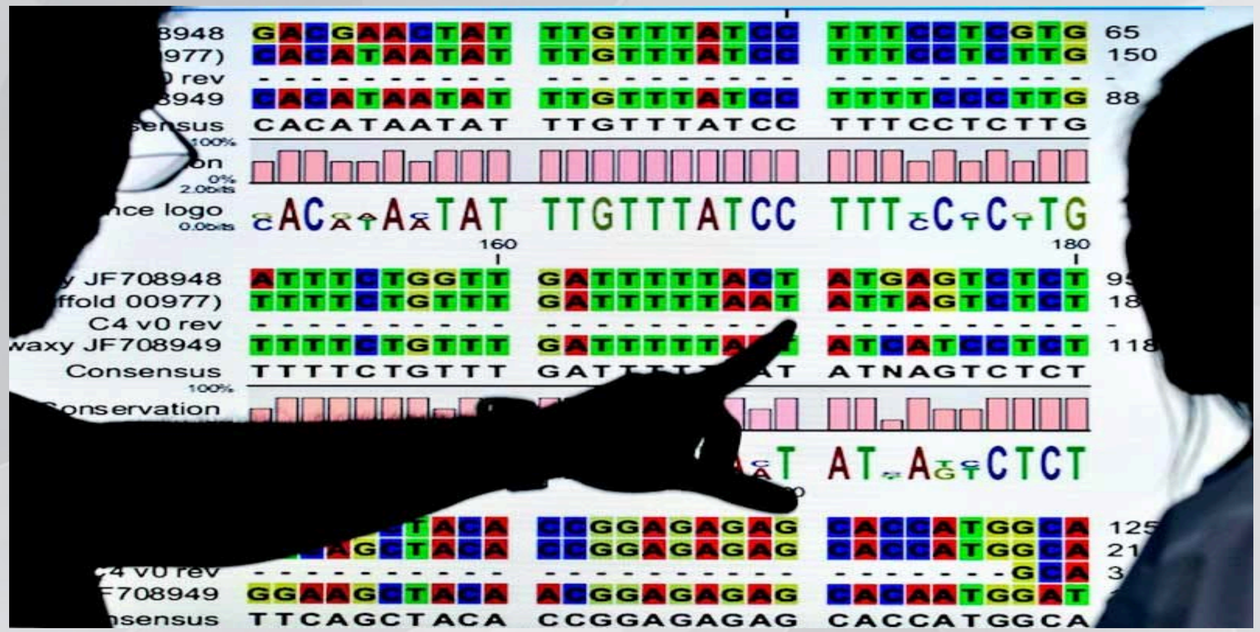
L’apprentissage machine, un outil pour « percevoir » les informations

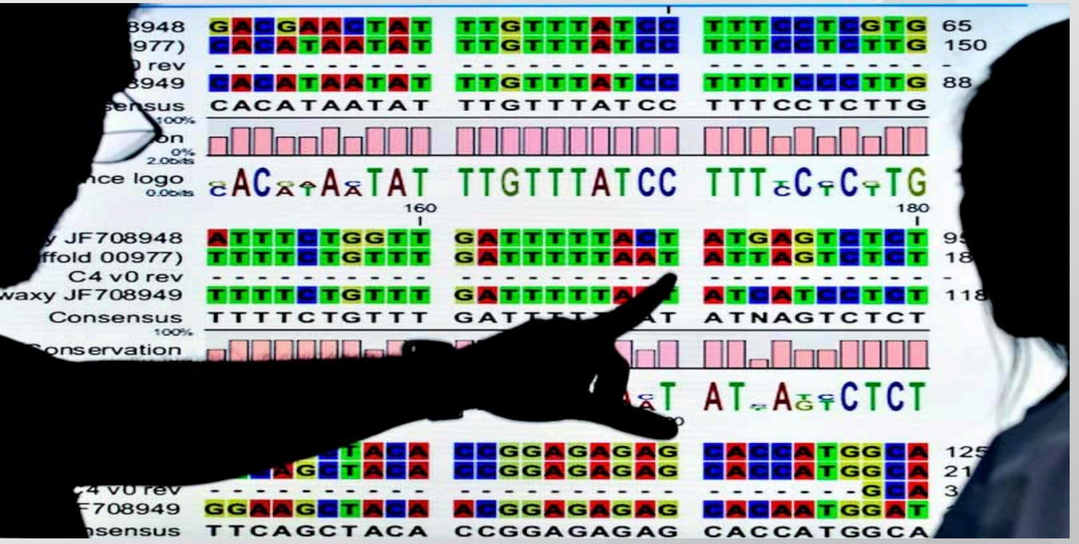


http://cs.stanford.edu/people/karpathy/deepimagesent/

En 2016, il y a eu un événement majeur en l’intelligence artificielle

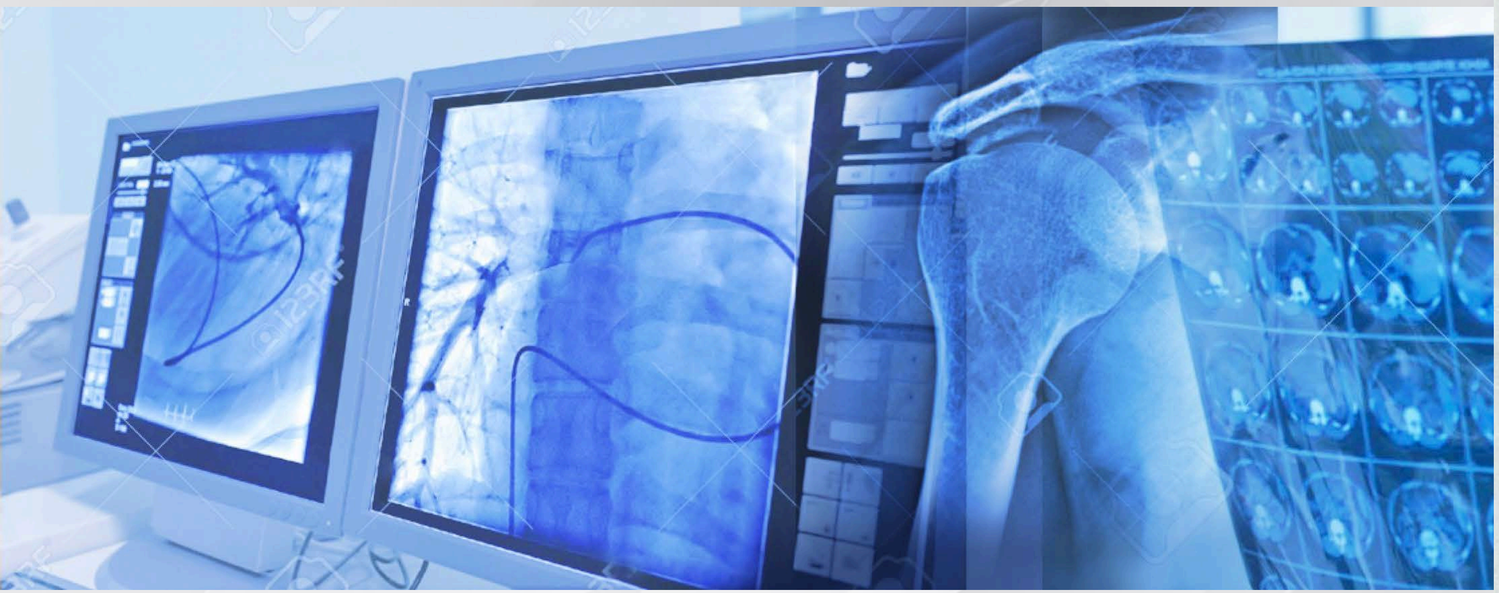
Interprétation de données « omics »





Interprétation de données « omics »

combinées avec des données cliniques



des données d’imageries

SCIENCES ■ Un nouvel espace « Green Tech Verte », dédié aux jeunes et aux start-up, vient d'ouvrir à Orléans

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
carole.tribout@centrefrance.com

Après la French Tech, le Lab'O et l'AgreenTech Valley, voici le « datacenter Green Tech Verte » d'Orléans. Un espace, comme son nom anglophone ne l'indique pas, géré par le ministère de l'Environnement.

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

De 15 à 20 personnes

Le pôle vient de déménager son espace documentation pour accueillir ce « datacenter ». C'est un espace de travail de 100 mètres carrés, ouvert gratuitement (sous convention) à une quinzaine de personnes intéressées par les données environnementales : des stagiaires, des chercheurs, et les start-up lauréates du concours que le centre vient de lancer (lire en encadré).

Il propose du haut débit, le Wi-Fi, de l'audio ou de la visioconférence, une salle de détente et une tisanerie. L'idée est de faire profiter les porteurs de projet des compétences présentes dans la maison. Et de faire avancer la science de la donnée (datascience).

INAUGURATION. Éric Bonmati, le responsable du pôle, a présenté l'espace de coworking aux invités, hier matin.

Un concours sur les pesticides

L'incubateur orléanais 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 ont jusqu'au 16 janvier pour s'inscrire. Ils pourront s'appuyer sur les mesures de 2.000 points en France, de 2007 à 2014. Le jury choisira une dizaine de dossiers le 16 février. Les porteurs des projets retenus pourront profiter gratuitement de l'espace orléanais. 6.000 € iront au premier prix, 3.000 au deuxième, 1.000 au troisième. Même si le volume de pesticides diminue, leur qualité augmente, et, finalement, la pollution ne décroît pas. Il arrive, selon le référent Green Tech Verte Orléans, qu'un seul point de mesure recèle jusqu'à 40 molécules différentes, dont certaines interdites depuis de nombreuses années !

Imaginer de nouveaux services citoyens

Car des millions d'informations sont enregistrées, notamment par les nouveaux compteurs intelligents, tels Linky. Que ce soit sur l'énergie, la faune, la flore, les risques naturels, les eaux… Et, le plus souvent, elles sont accessibles à tous. Encore faut-il le savoir et savoir comment les trouver.

Rendre ces données faciles d'utilisation : ce sera le rôle des experts du ministère. S'en servir pour imaginer de nouveaux services utiles aux citoyens : c'est ce qui est attendu des start-up.

Afin, par exemple, de mieux prévoir les inondations, de mieux répertorier les espèces, de mieux penser l'aménagement urbain. Ou encore, « on peut imaginer une application qui donne la qualité de l'air d'Orléans, ou de ses transports en commun… », cite, au hasard, Éric Bonmati. Tout est imaginable.

Cette nouvelle démarche entre dans la politique « Green Tech Verte », mise en œuvre, depuis février, par Ségolène Royal, ministre de l'Environnement (représentée, hier, par Serge Bossini, son directeur de la recherche). Il s'agit « d'accompagner la transition écologique et énergétique et de stimuler l'innovation ».

Deux appels à projets ont été lancés et trois hackathons organisés, au niveau national. Un incubateur a ouvert en Seine-et-Marne, en septembre. Orléans est le deuxième de France, dédié aux sciences de la donnée. Deux autres sont prévus à Lyon et Toulouse. ■

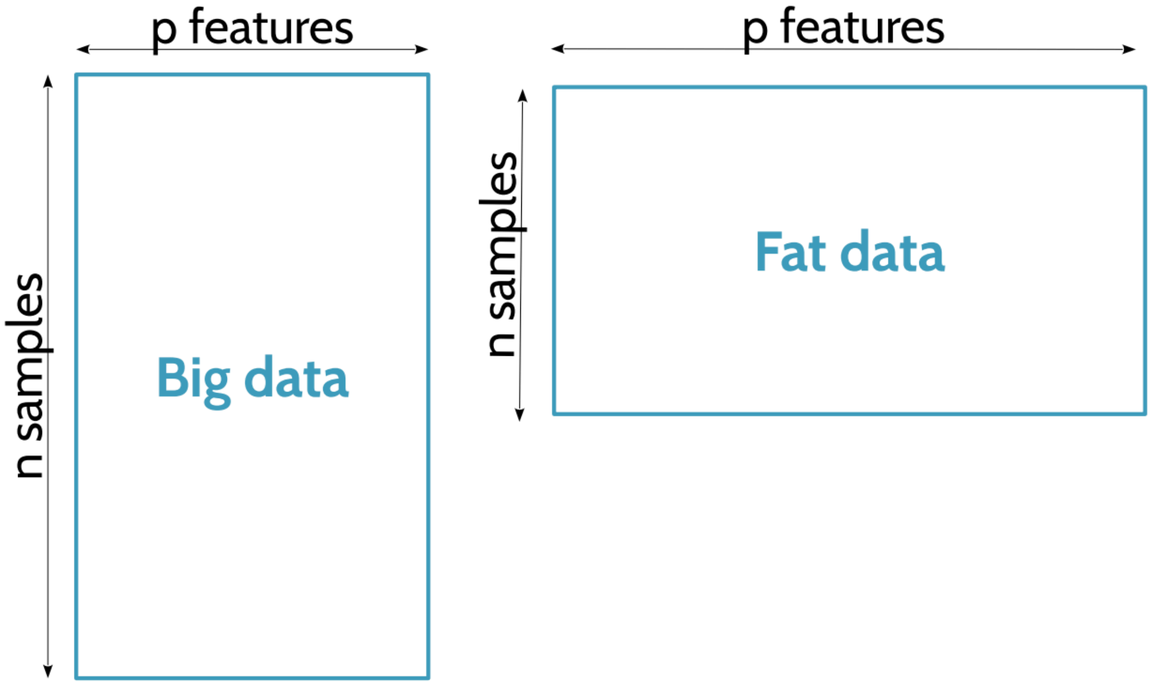
Pratique. Commissariat général au Développement durable, 5, route d'Olivet, Orléans. Tél. 02.38.79.78.78. greentechverte@developpement-durable.gouv.fr

Pour consulter : www.developpement-durable.gouv.fr/Greentech-verte.html

et même des données environnementales

Qu’est-ce que le « Fat Data » ?

Un cas particulier du paradigm du « Big Data »

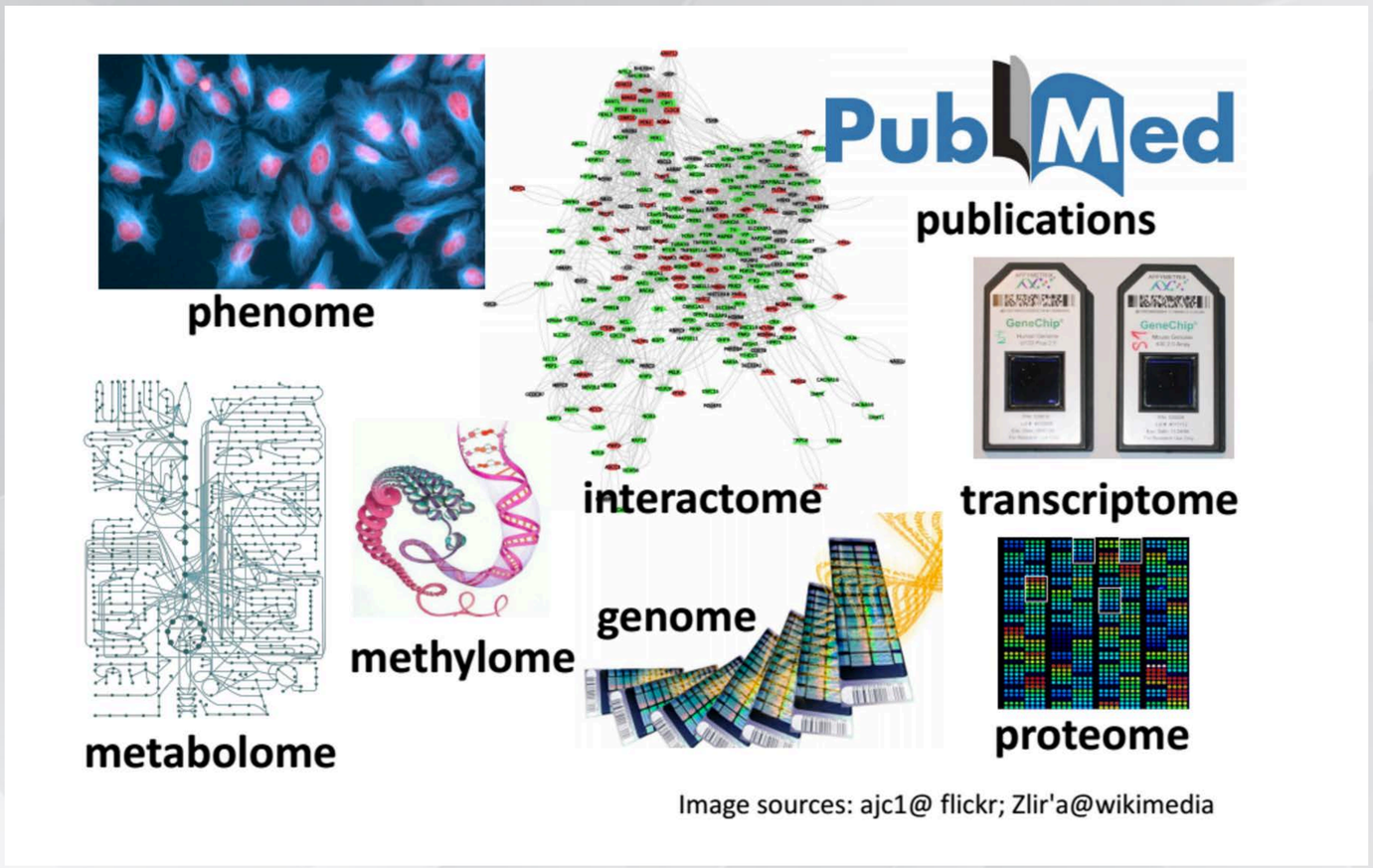


E.g. **G**enome-**W**ide **A**ssociation **S**tudies (GWAS):

- $p = 10^5 - 10^7$ **S**ingle **N**ucleotide **P**olymorphisms (SNPs)
- $n = 10^2 - 10^4$ samples.

Slide taken from: http://cazencott.info/dotclear/public/publications/2016-01-13-azencott.pdf

En science de la vie, le « Fat Data » est un véritable enjeu



Et le manque d’accès aux données ne fait qu’empirer les choses

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." »

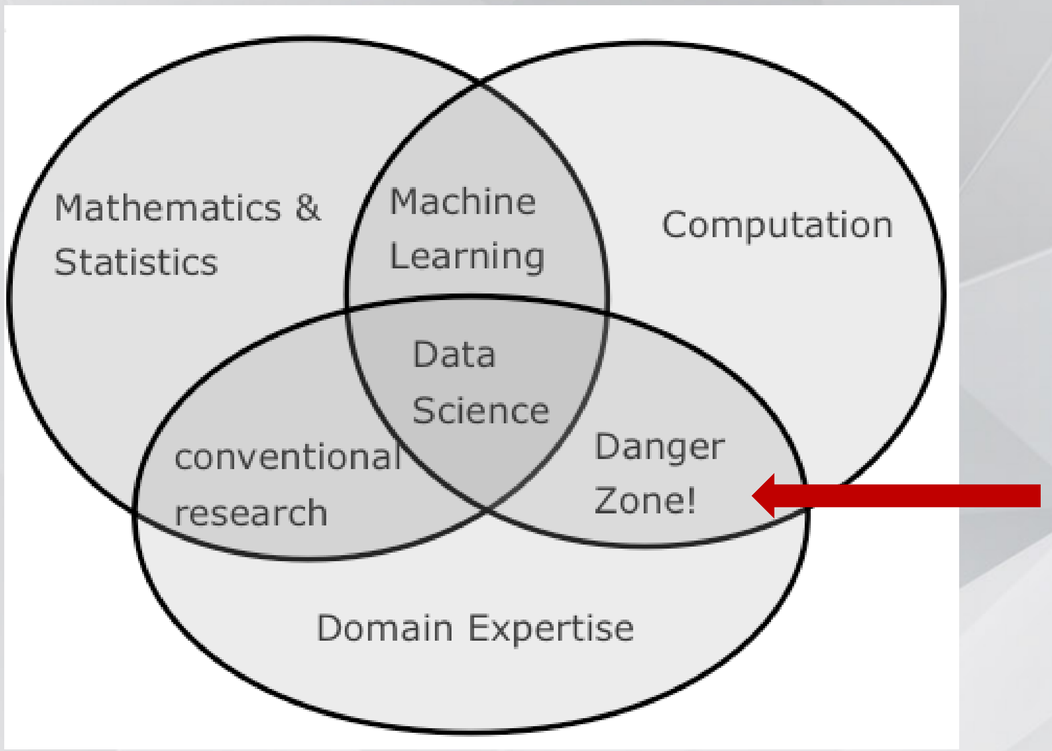
Droit individuel à la confidentialité et intérêt collectif

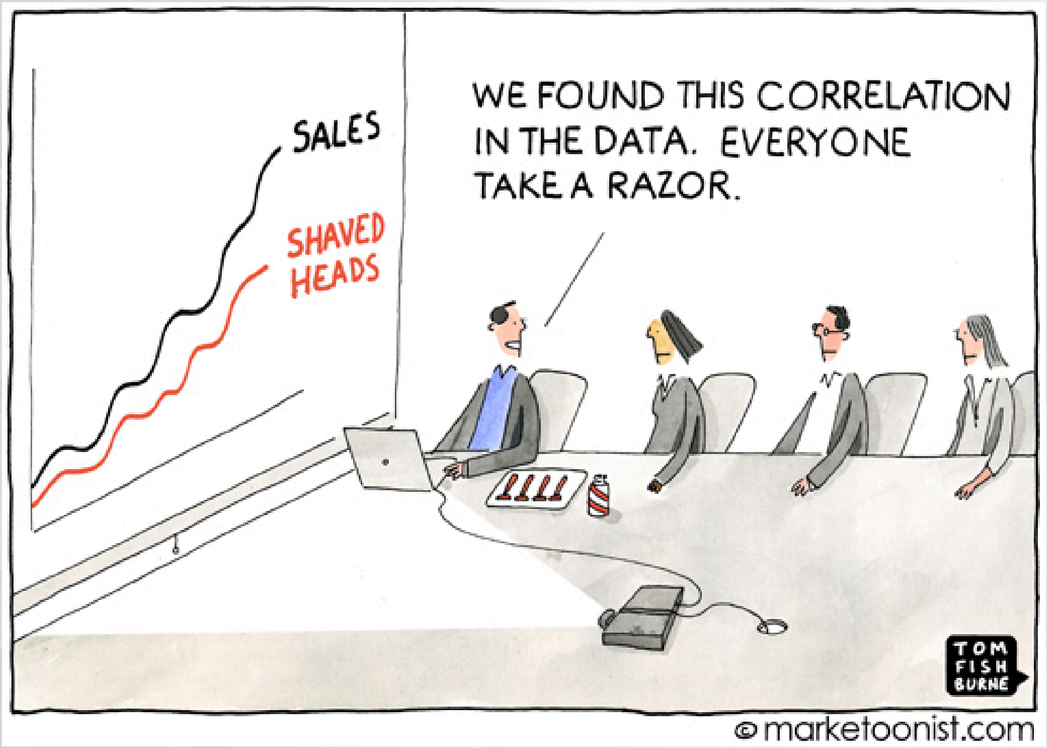
- Renoncer aux données, c’est se couper de grandes possibilités !
- On doit chercher un compromis entre la protection du citoyen et l’intérêt collectif.
- On doit aussi avoir le réflexe de conserver nos données et les voir comme « bien public »
  - La SAAQ a renoncé à son projet « Ajusto »
  - Les données 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 24h

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 accès sous forme de données ouvertes
  - Prendre toutes les précautions pour que que les données ne puisent « 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 coups »
  - 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

Autres sphère où il faut être prudent avec l’IA





Le diagramme de Venn de Drew Conway sur le Big Data

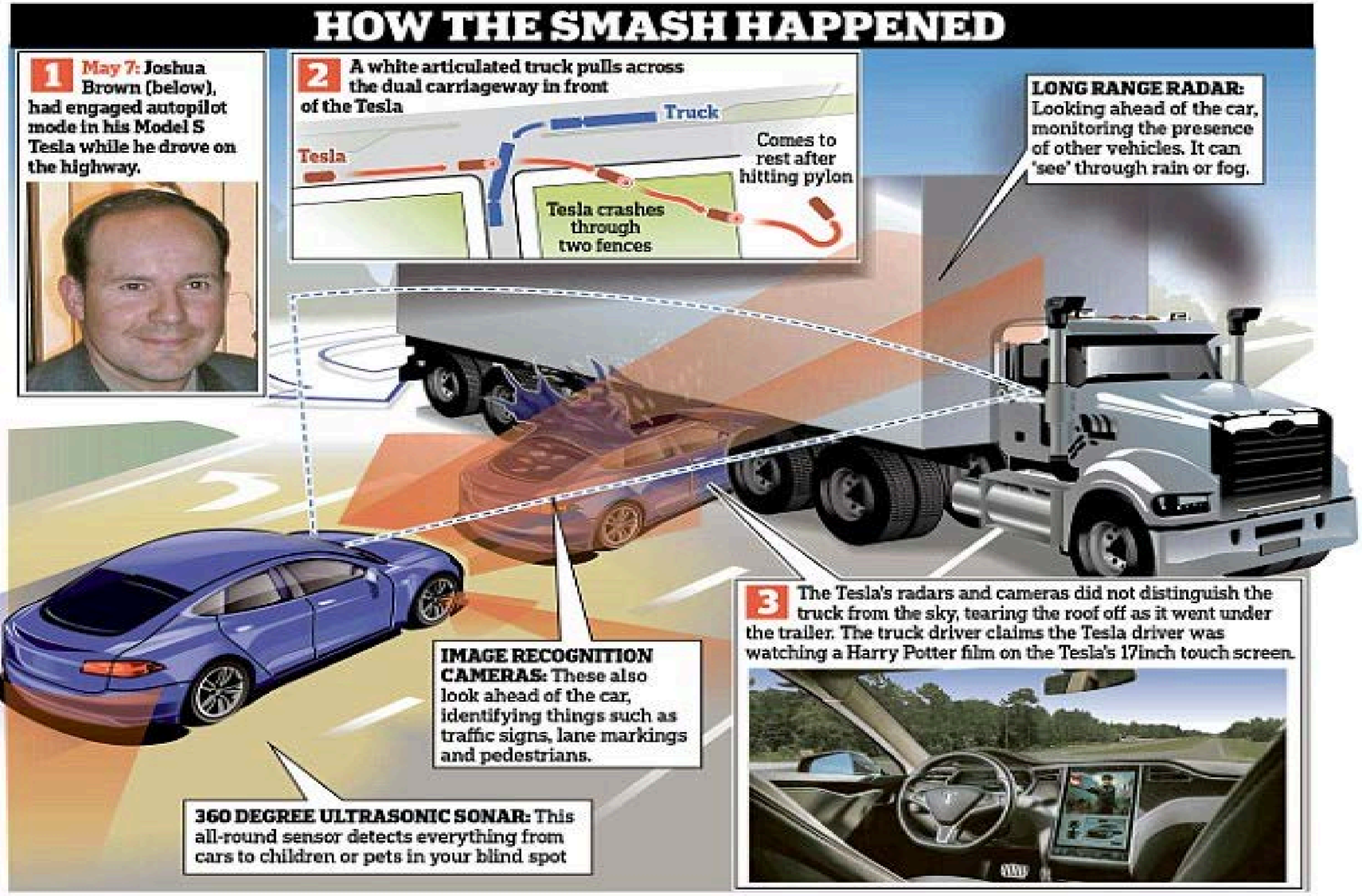
En apprentissage machine, les choses peuvent mal aller

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’entrainement 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 »

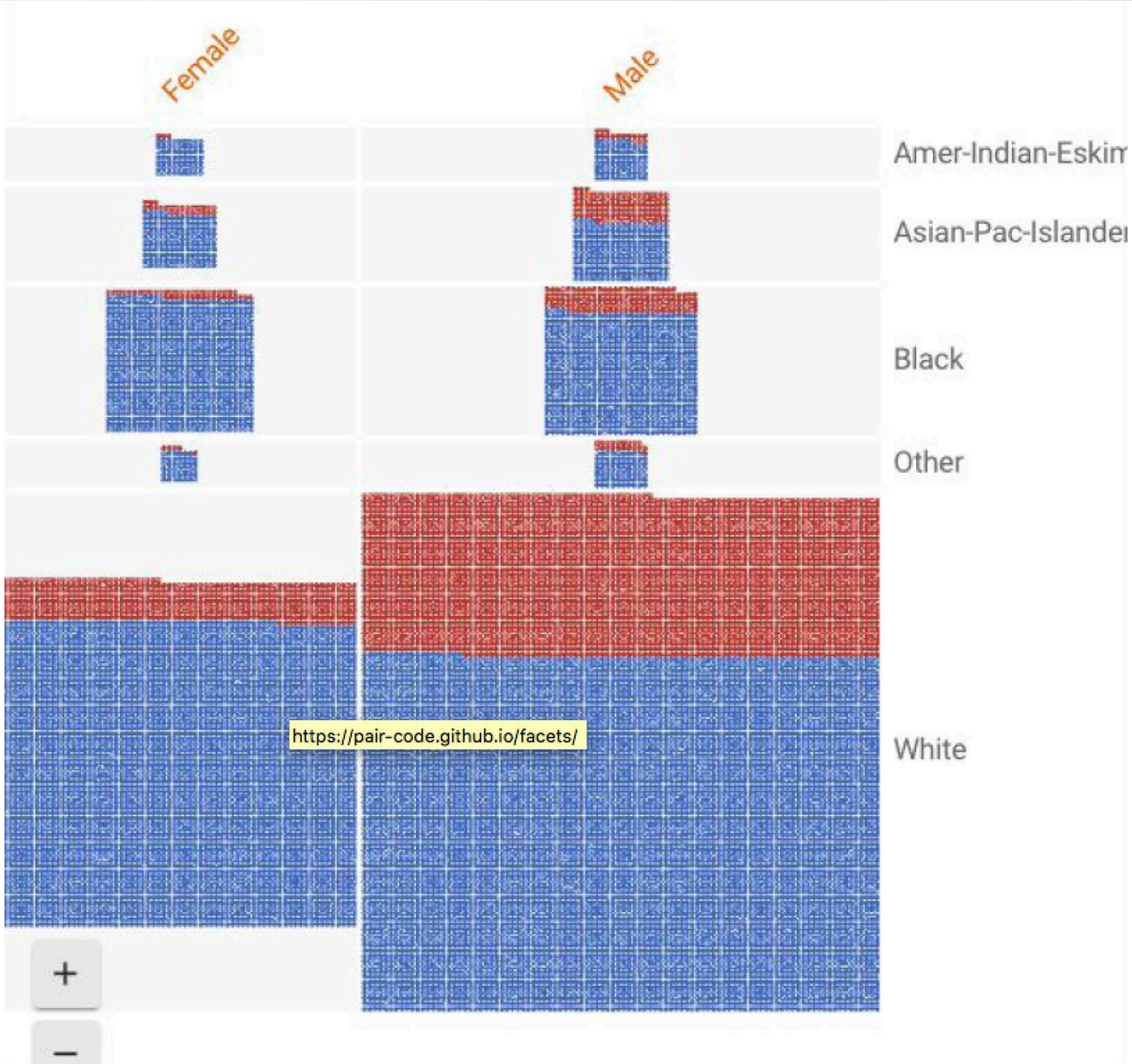
En apprentissage machine, les choses peuvent mal aller

Les évènements rares



Autres défis de l’intelligence artificielle

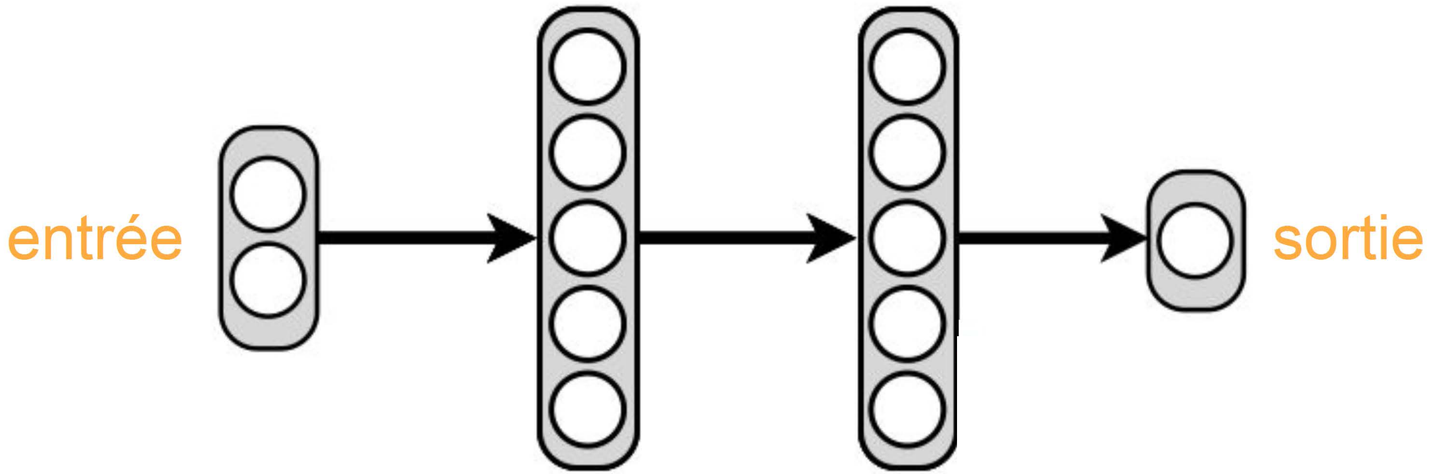
L’équité



L’IA est aussi « bonne » que les données sur lesquelles elle a été entrainée

Autres défis de l’intelligence artificielle
L’équité

Une solution possible au manque d’équité:

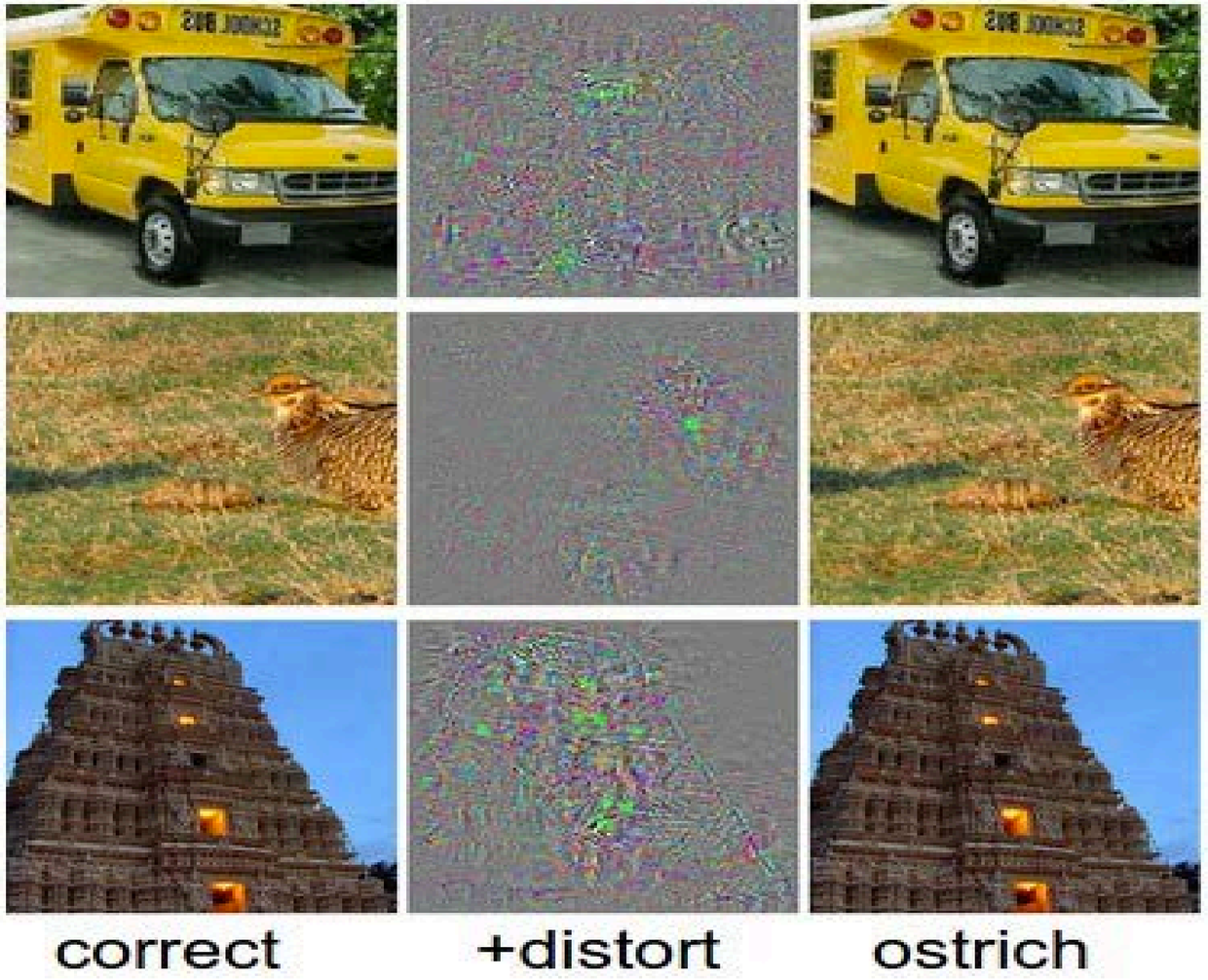


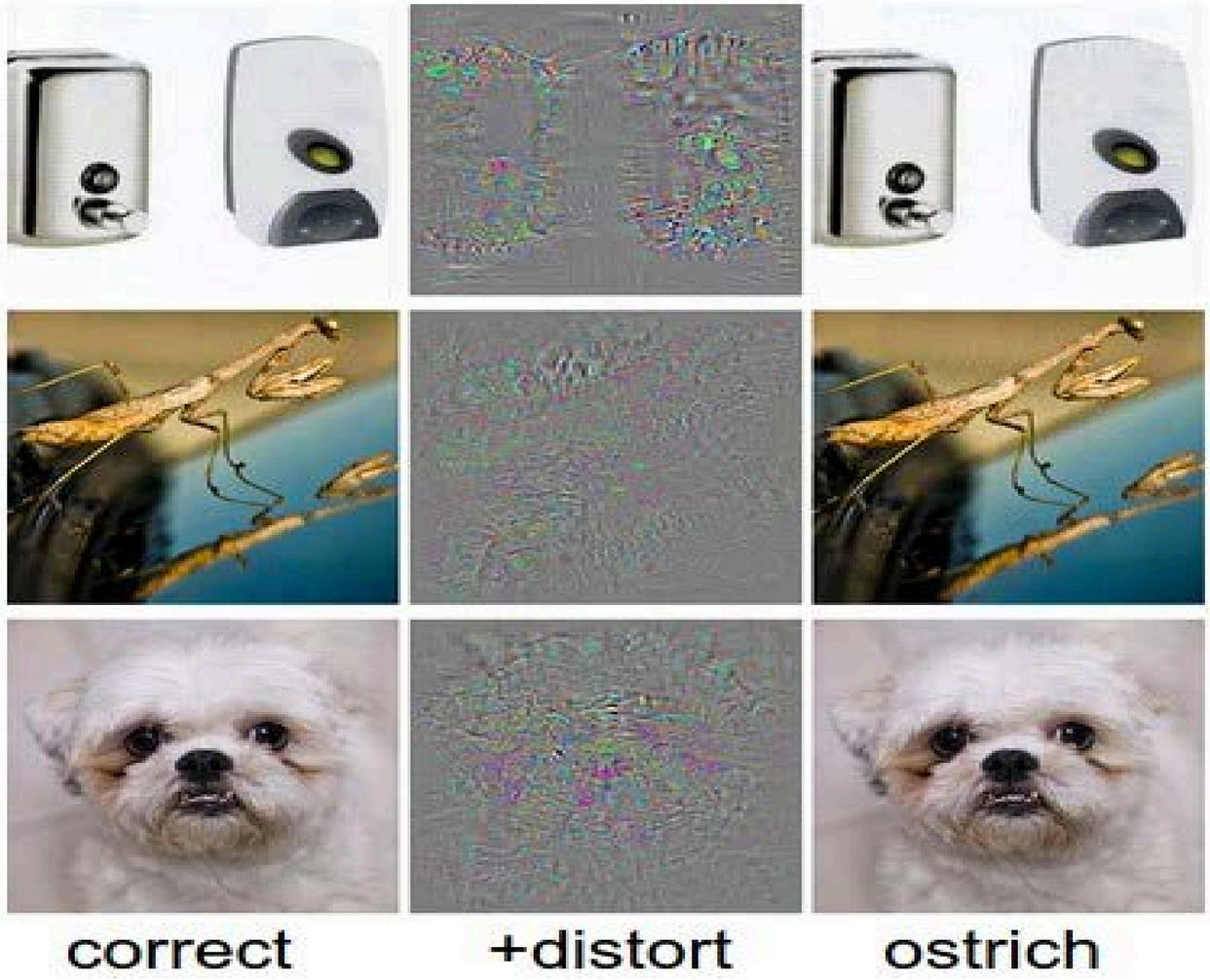
Domain-Adversarial Training of Neural Networks

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

Autres défis de l’intelligence artificielle

La robustesse





Autres défis de l’intelligence artificielle

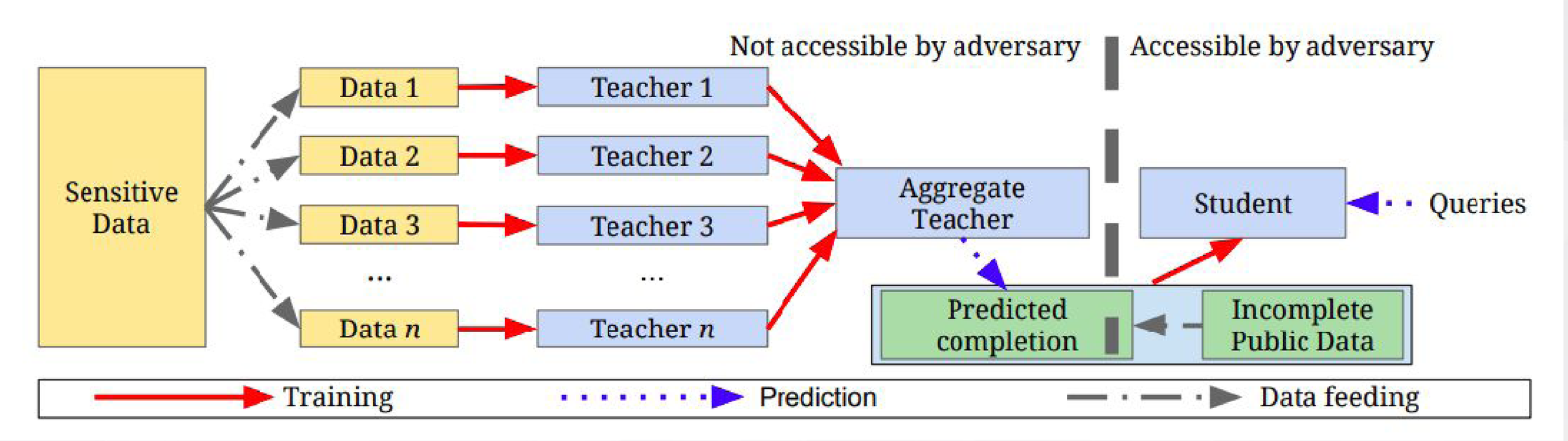
La robustesse

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

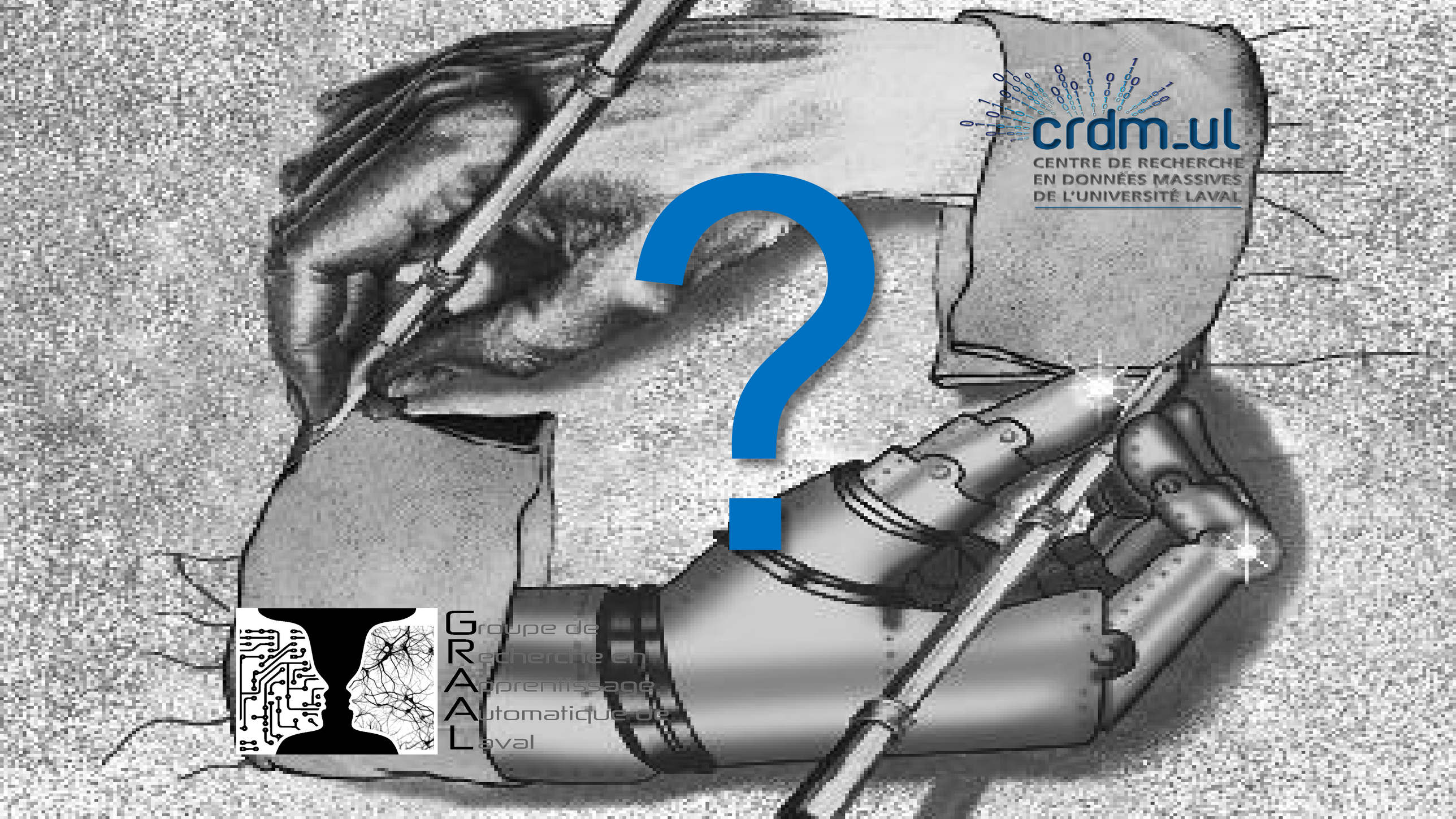
Autres défis de l'intelligence artificielle

La confidentialité des données

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



Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data
Papernot, Abadi, Erlingsson, Goodfellow, Talwar, 2017


crdm_ul
CENTRE DE RECHERCHE
EN DONNÉES MASSIVES
DE L'UNIVERSITÉ LAVAL
GRAAL
Groupe de
recherche en
apprentissage
automatique de
Laval