L'intelligence artificielle : potentiel et défis

Présenté par
François Laviolette

Journée de la recherche du FRQNT

Québec, 16 mai 2018

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

Groupe de Recherche en Apprentissage Automatique de 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

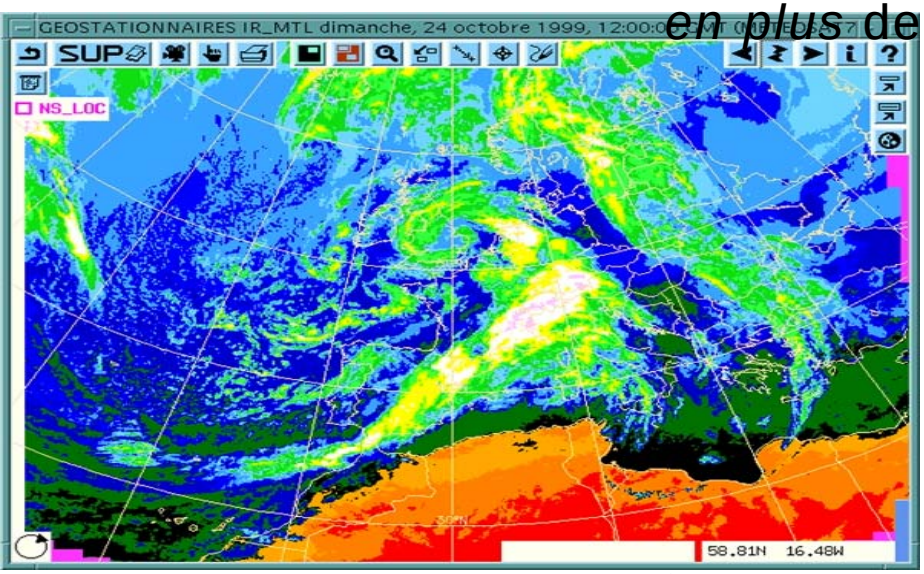
Advertisement

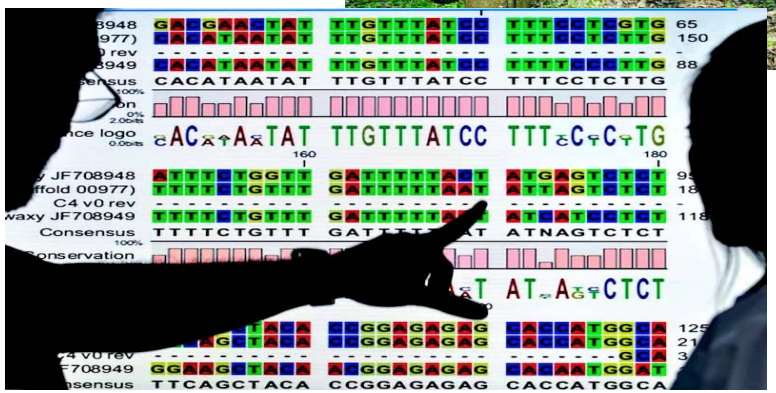
Not all intelligence is artificial.

Enjoy 12 weeks' access for just $12.

SUBSCRIBE

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





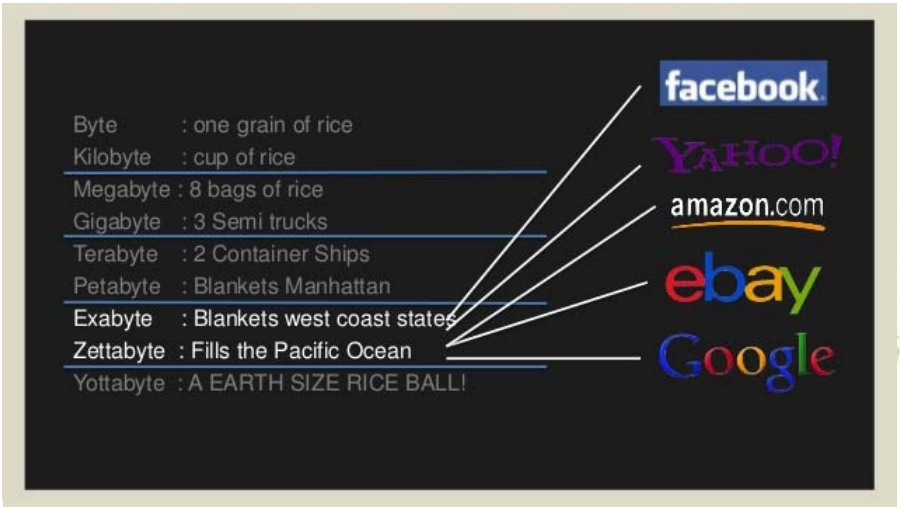
Qu’est-ce que le « Big Data » ?

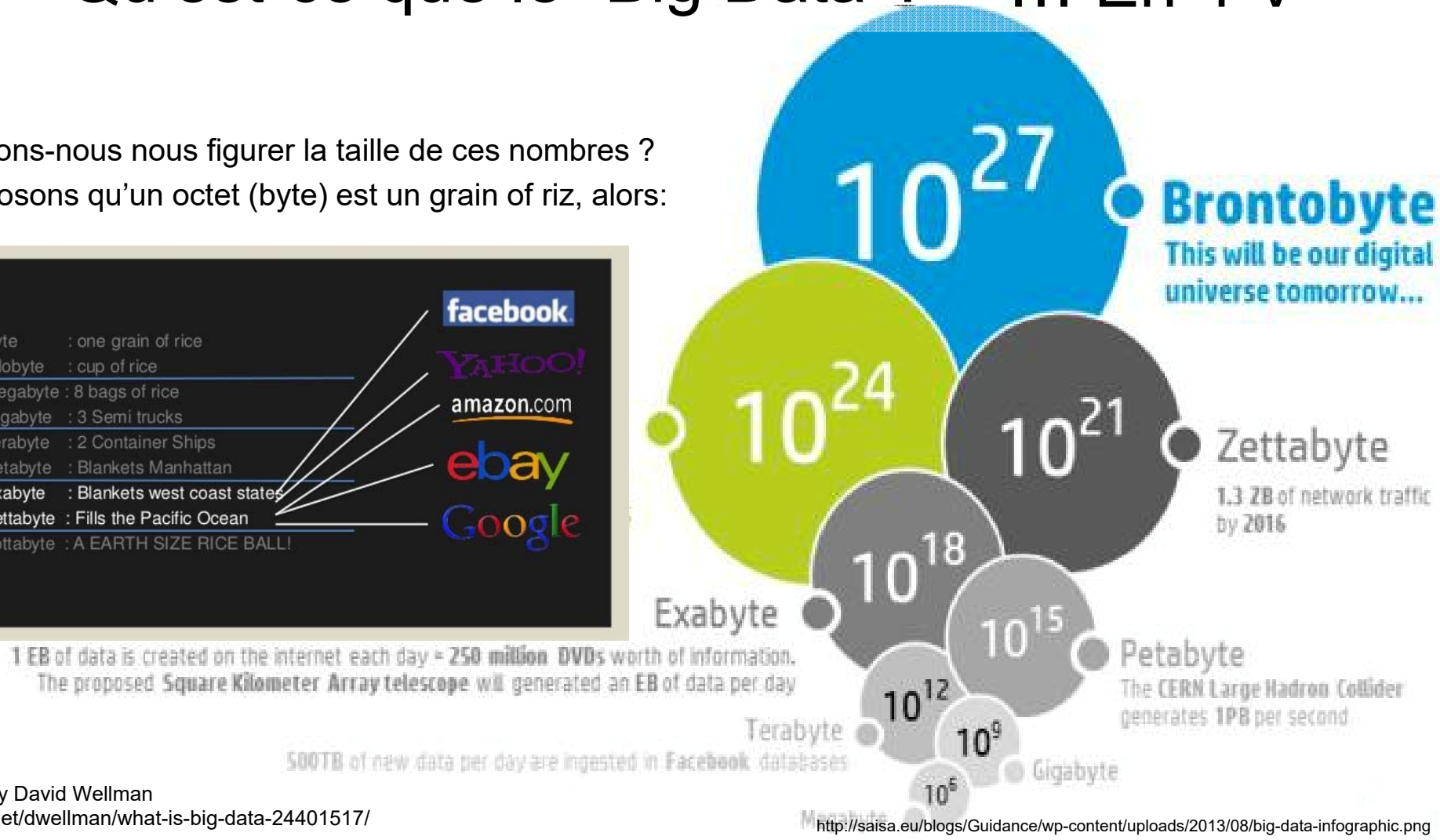
- D’abord, quel est le bon terme français?
  - Les Français parlent de *mégadonnées*
  - Nous avons choisi l’expression *données massives*,

    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

What is “Big Data”? … In 4 V

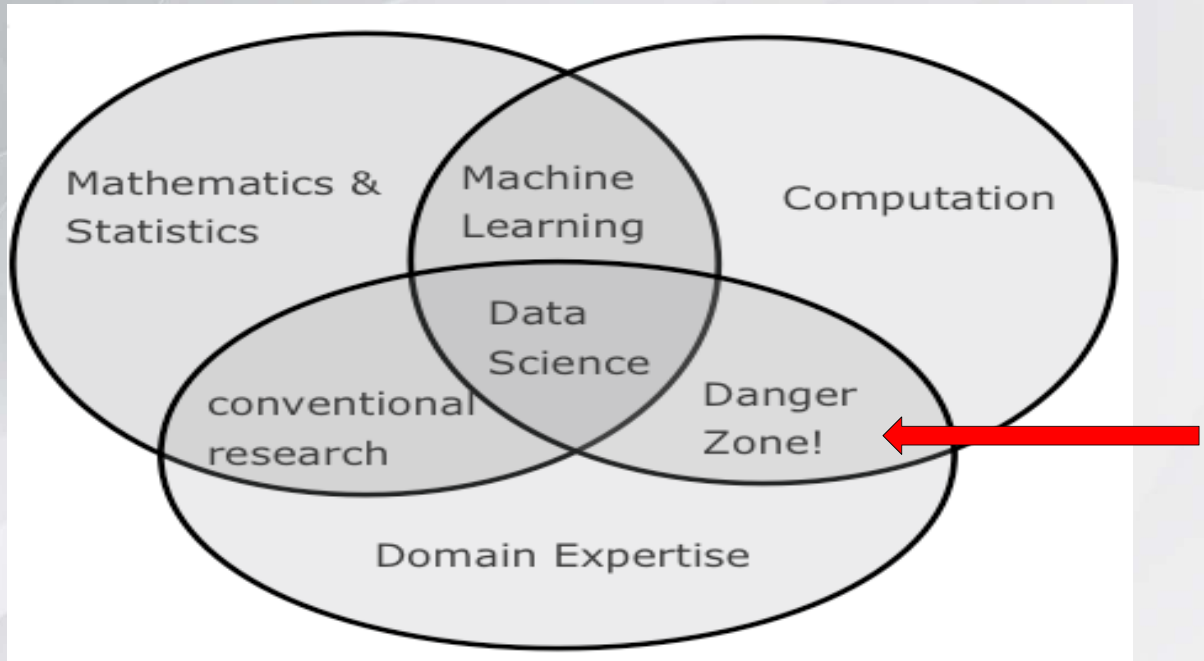
- Volume
- Vélocity
- Variety
- Véracity

Data gathered from different sources
Images, texts, sensor data, mouse clics …

Data that comes from different projects,
Based on different methodologies non necessarily compatible

Machine Learning
and
Big Data

Machine learning:
understand the information



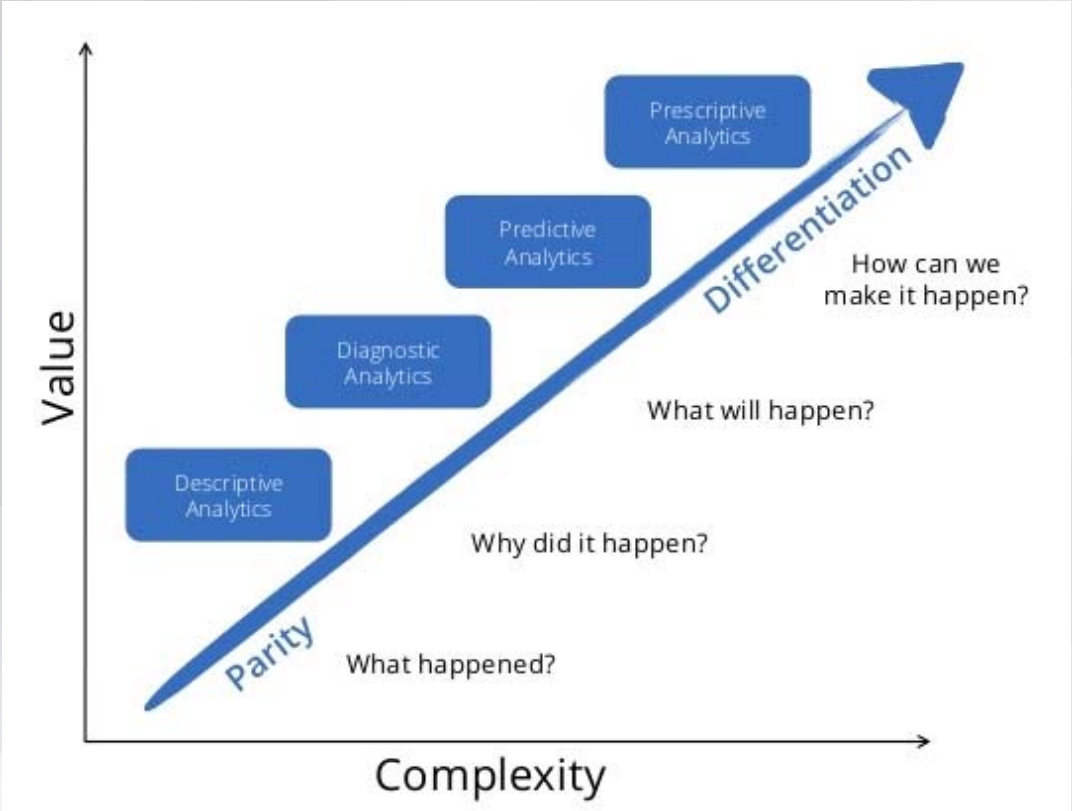
Le diagramme de Venn de Drew Conway sur le « Big Data »

L’ apprentissage automatique
vs
la recherche opérationnelle

Apprentissage automatique:
Comprendre les informations

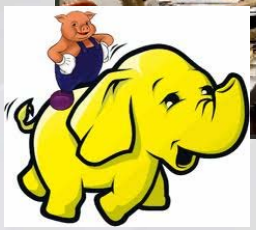
Recherche opérationnelle:
Optimise la prise de décision

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)

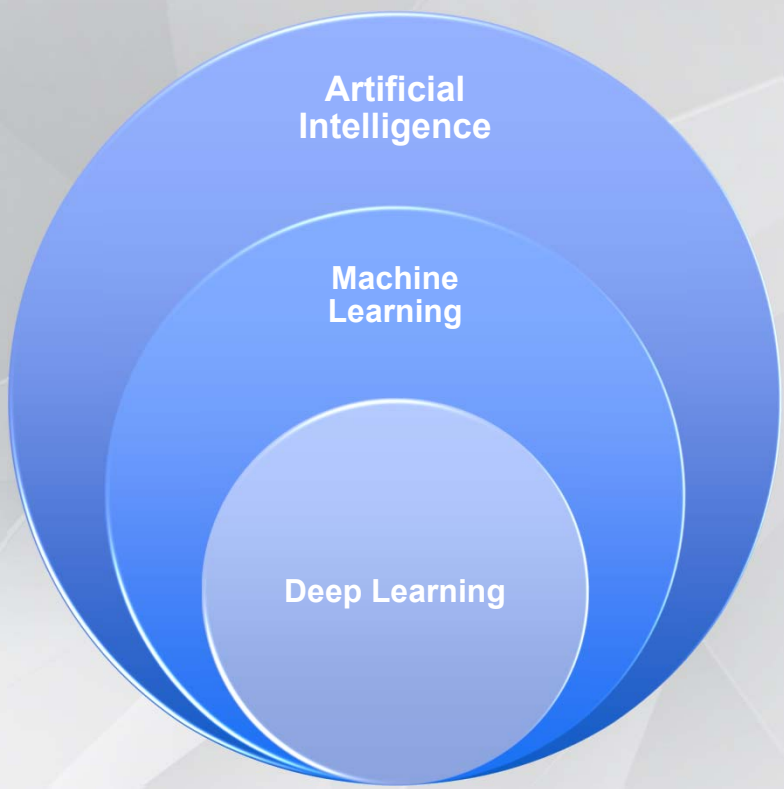
Les défis du “Big Data”



- 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

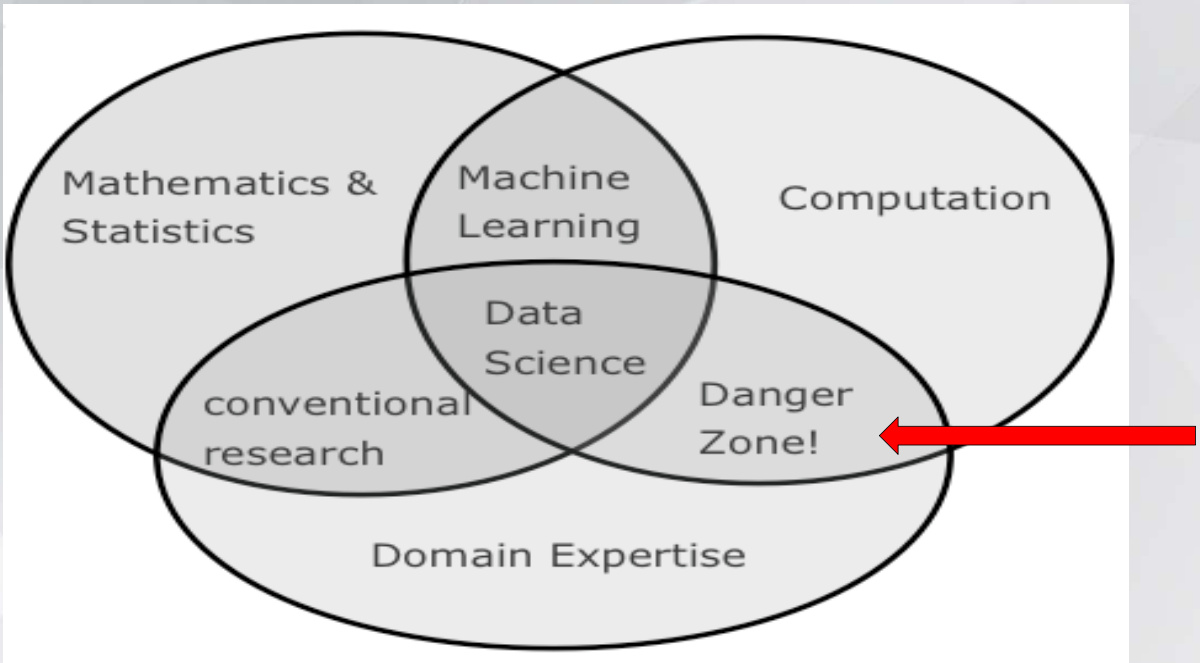
Et pour toutes ces tâches, <u>l’apprentissage machine</u> et la <u>recherche opérationnelle</u> sont des outils de prédilection.

Artificial intelligence and machine learning



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

L’ apprentissage automatique
et
les données massives



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

Machine learning 101

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

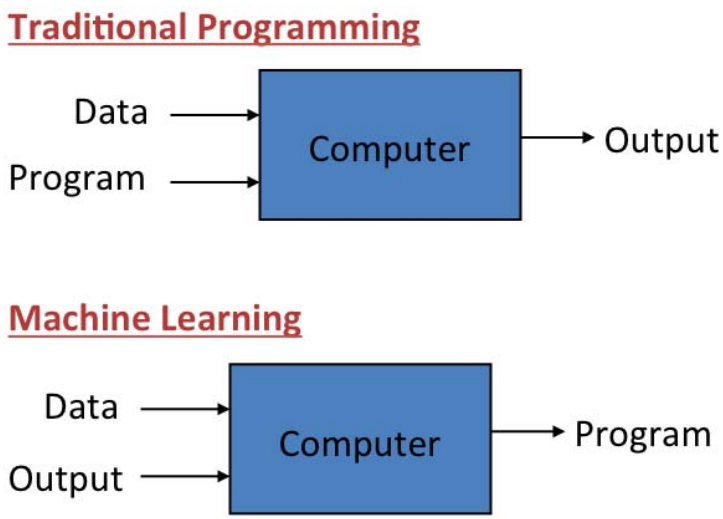
-Arthur Samuel (1959)

The machine learns

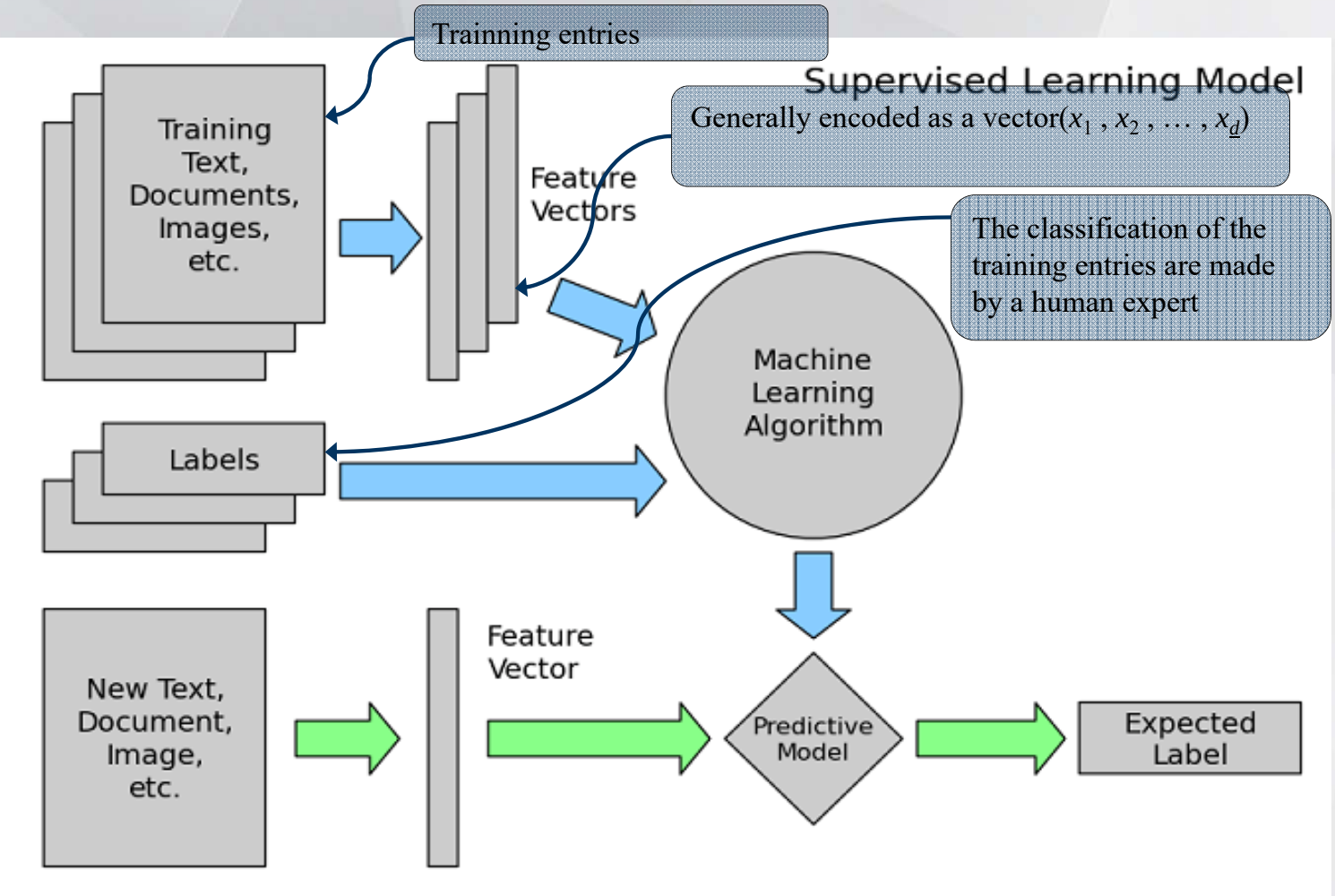
from examples or

from some interactions

with the environment

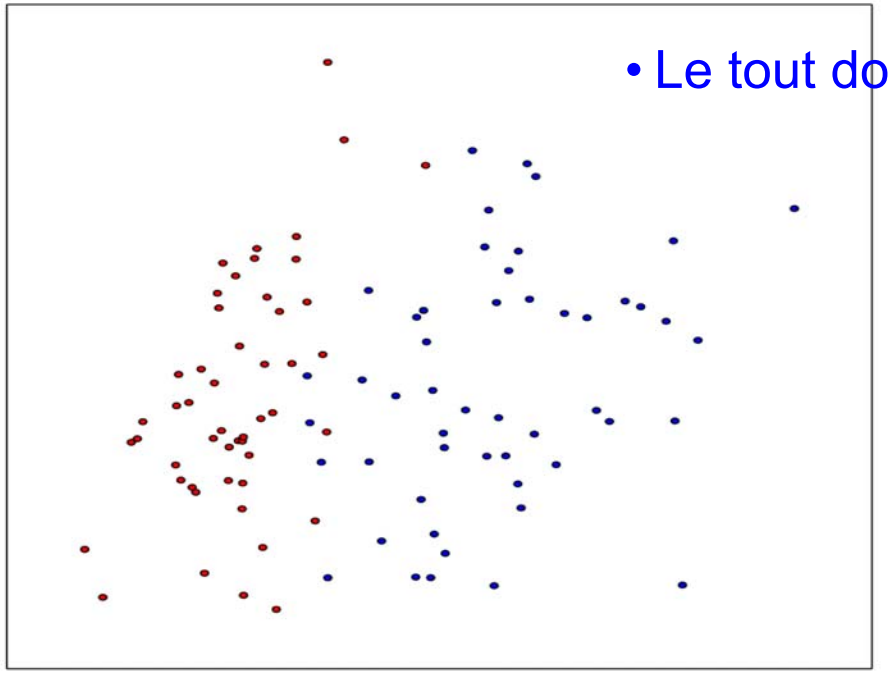


L'apprentissage supervisé

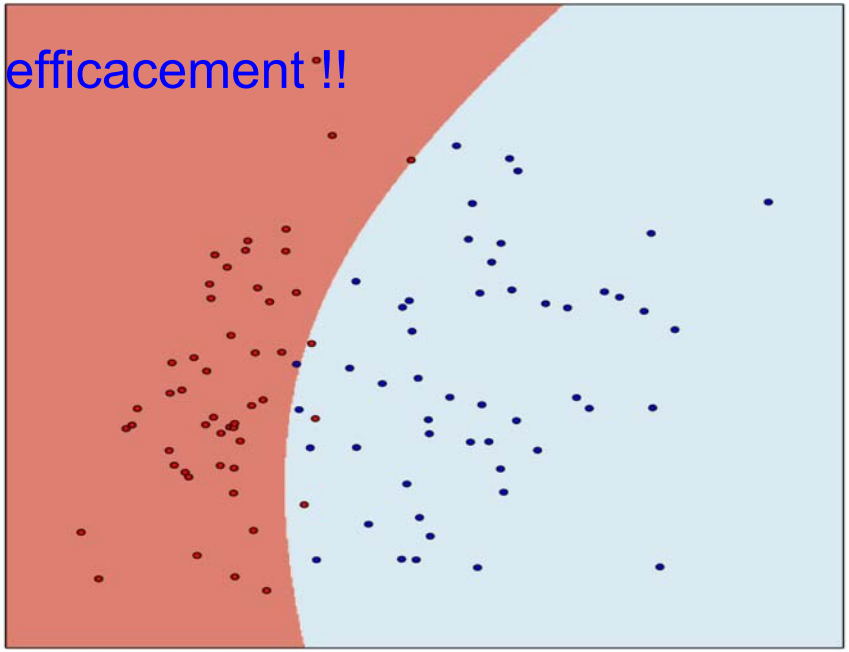


La tâche d'apprentissage en pratique

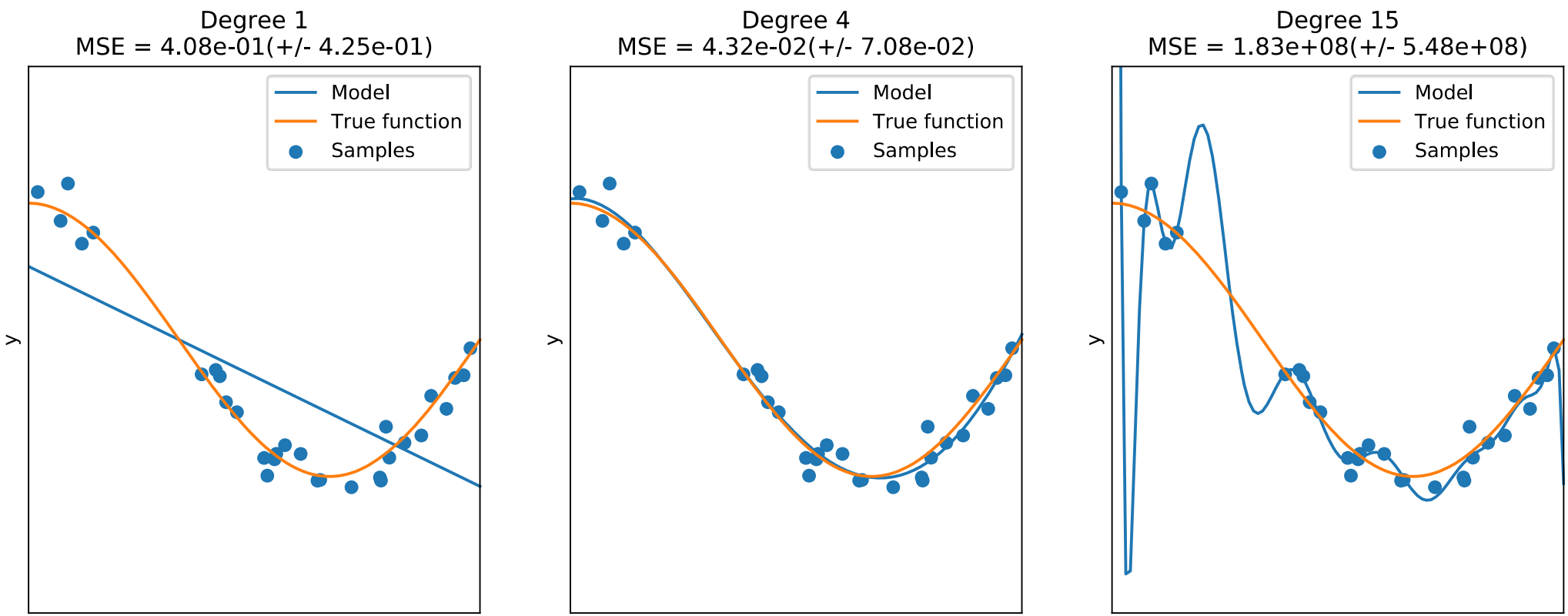
- Chercher un classificateur *h* qui fait peu d'erreurs sur l'ensemble d'entraînement tout en évitant le surapprentissage (overfitting) qui résulterait d'une correspondance trop parfaite des données d'apprentissage



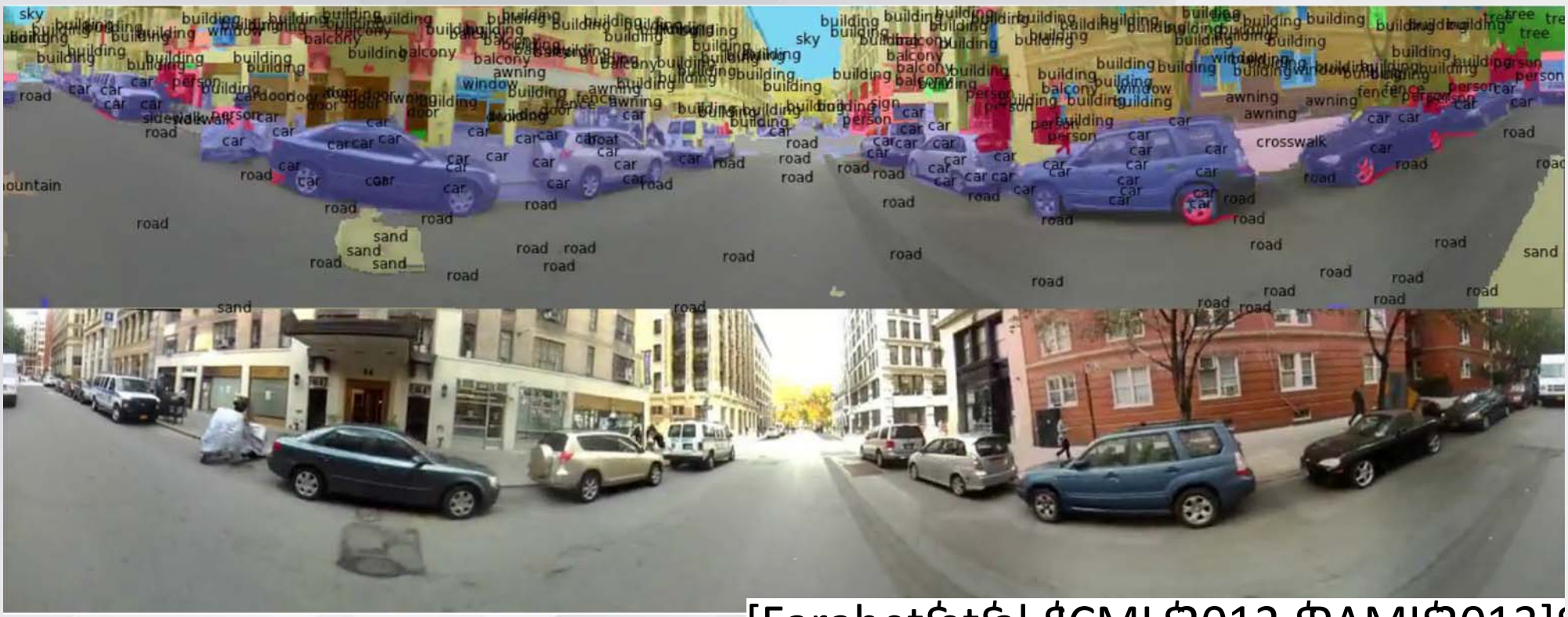
- Le tout doit se calculer efficacement !!



Un exemple de la problématique du sur/sous apprentissage

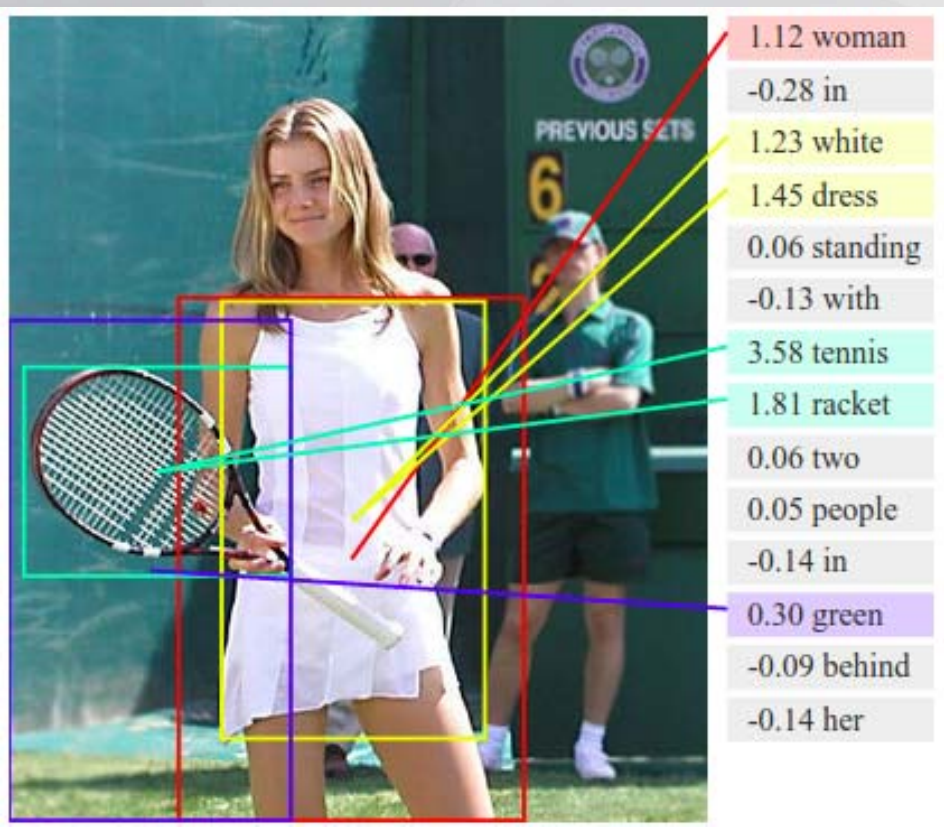


Scene labelling using deep learning

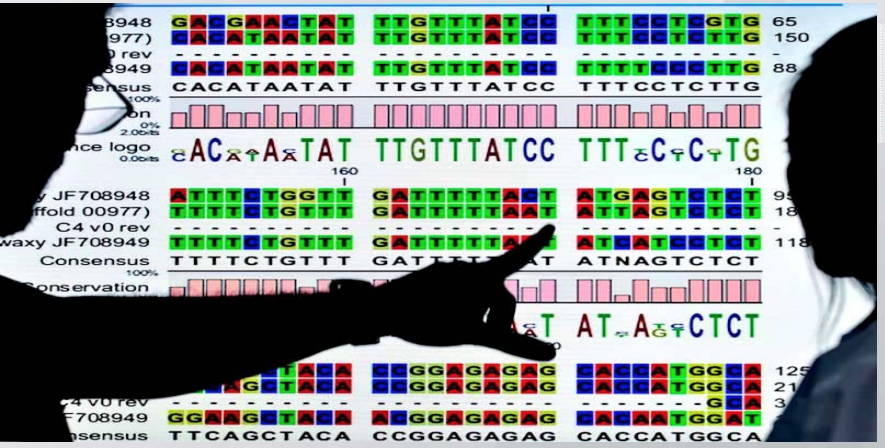


[Farabet et al. ICML 2012, PAMI 2013]

Machine learning, a tool to « understand » les information



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



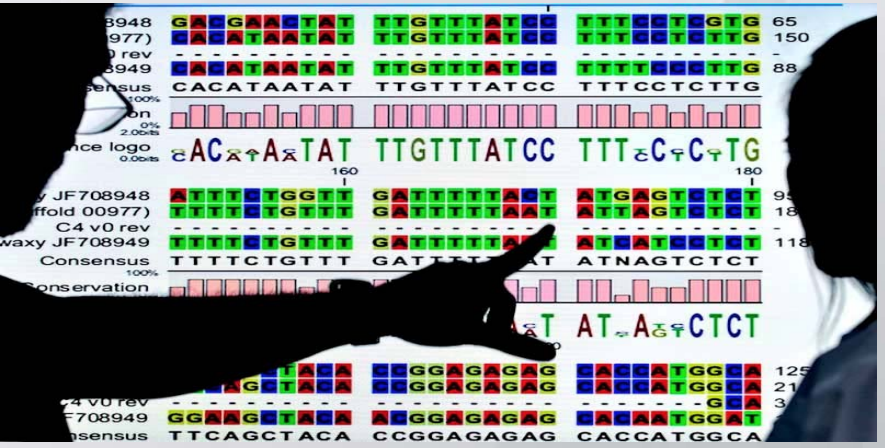
Big Data ? … In 4 V

Interpretation of « omics » data

- Volume
- Velocity
- Variety
- Veracity

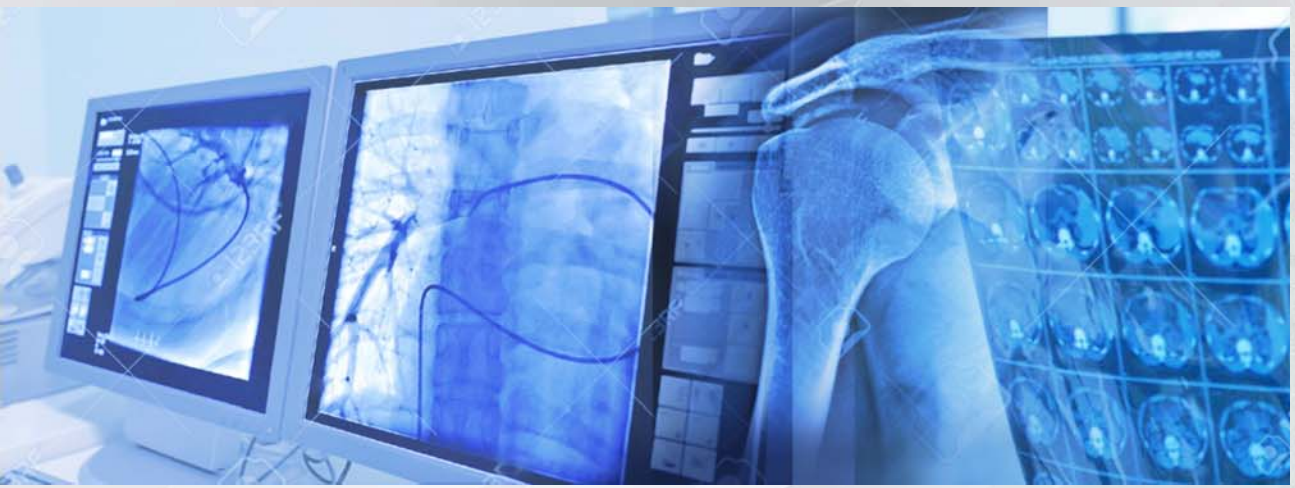
Data gathered from different sources
Images, texts, sensor data, mouse clics …

Data that comes from different projects,
Based on different methodologies non necessarily compatible

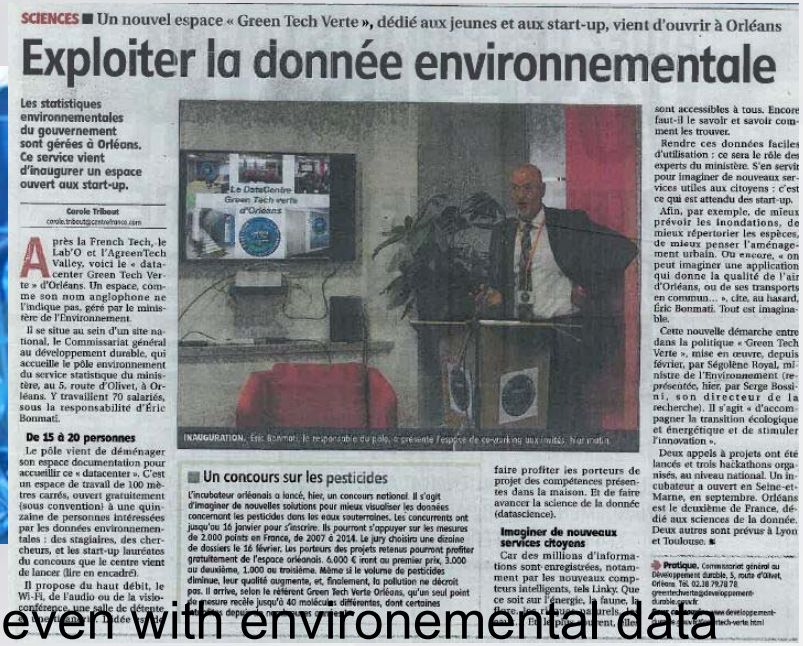


Interpretation « omic » data

combined with clinical data



medical images



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 [illegible]

INAUGURATION. Éric Bonmati, le responsable du pôle, présente l'espace de co-working aux invités. [illegible]

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 [illegible]

faire profiter les porteurs de projet des compétences présentes dans la maison. Et de faire avancer la science de la donnée (datascience).

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 [illegible] 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 ».

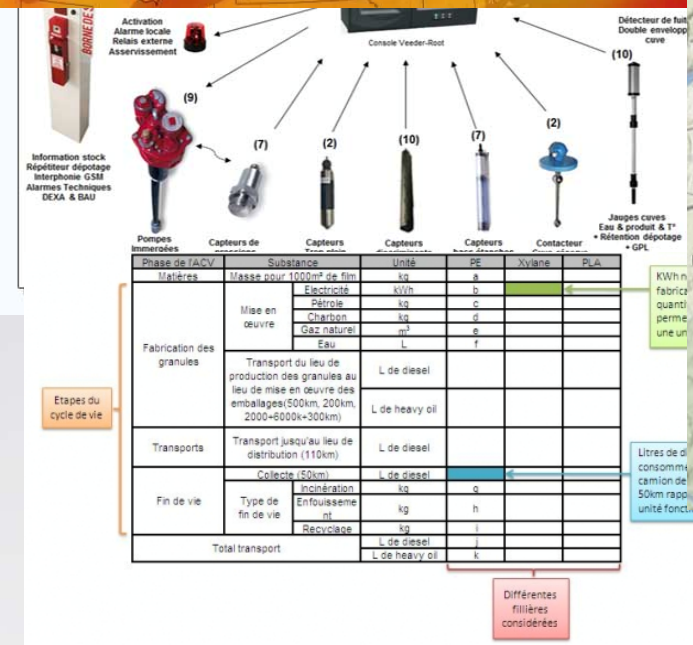
Deux appels à projets ont été lancés et trois hackathons organisés, au niveau national. Un incubateur a ouvert en Seine-et-Marne, en septembre. Orléans est le deuxième de France, dédié aux sciences de la donnée. Deux autres sont prévus à Lyon et Toulouse. ■

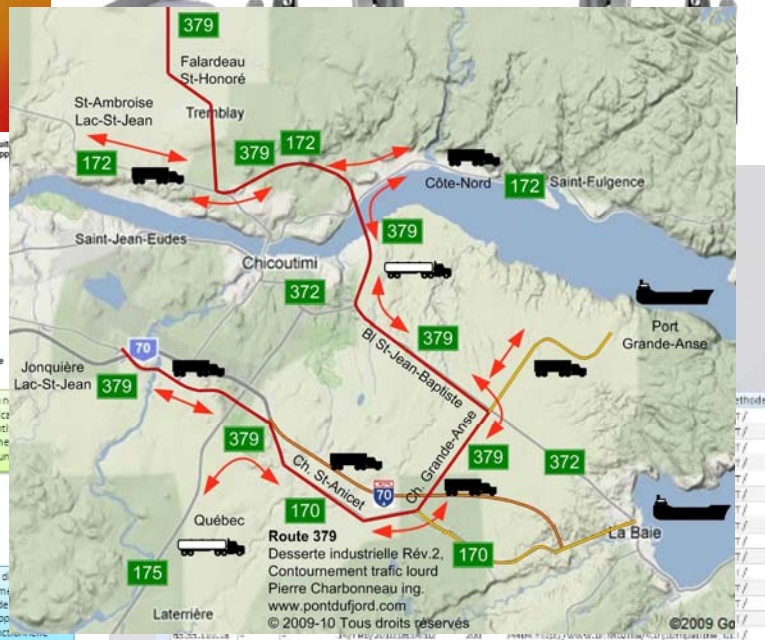
Pratique. Commissariat général au développement durable, 5, route d'Olivet, Orléans. Tél. 02.38.79.78.78. greentechvertedeveloppement-durable.gouv.fr

Pour [illegible] développement-durable.gouv.fr/green-tech-verte.html

and even with environemental data

2e example of a situation with heterogenous and unstrctured data





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

An example of a learning algorithm

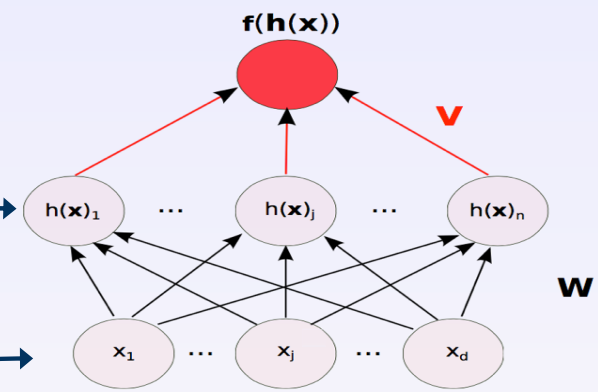
The neural network

Let consider a neural network architecture with one hidden layer

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

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

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



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

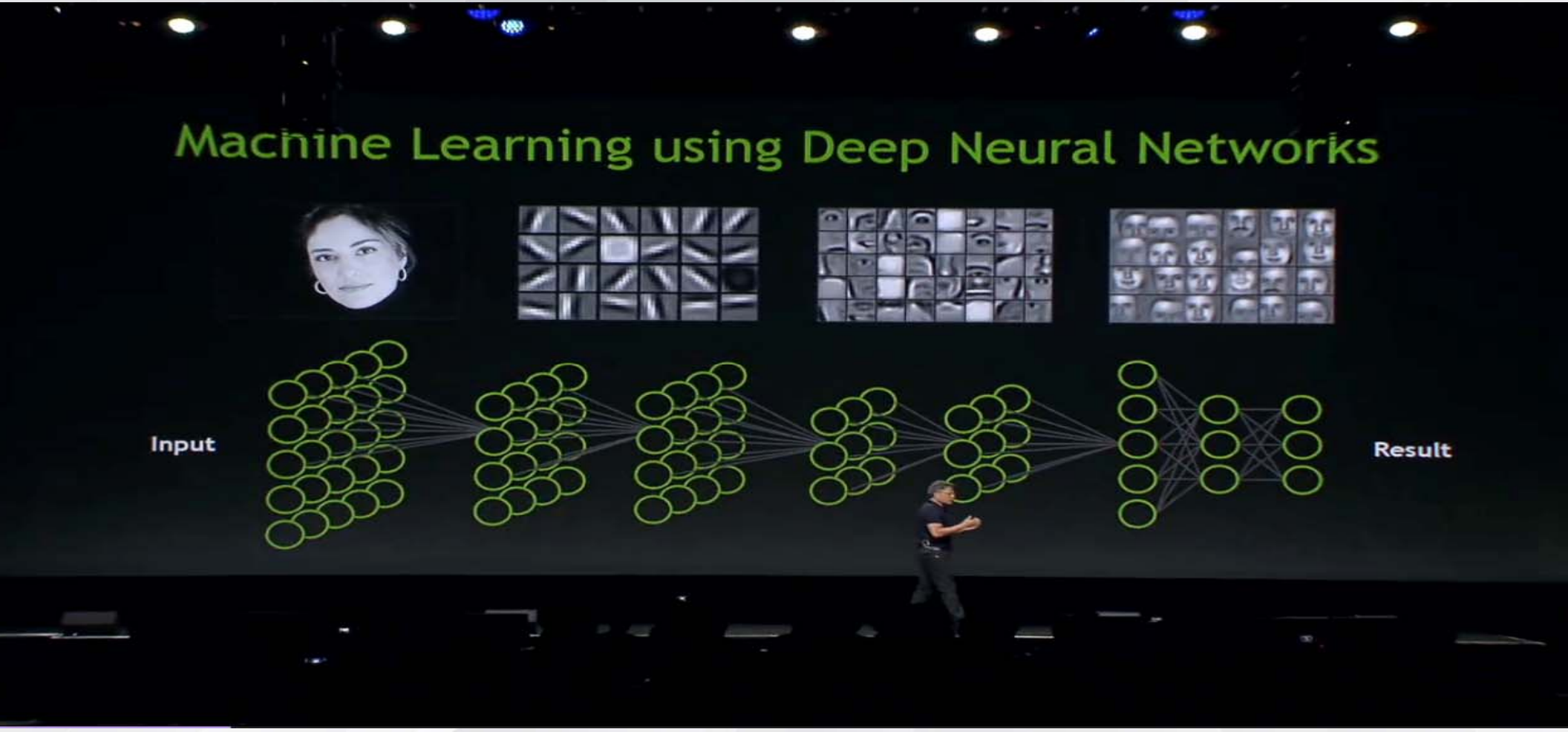
Given a source sample $S = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^m \sim (\mathcal{D}_S)^m$,

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 $\mathbf{f}(\mathbf{h}(\mathbf{x}^s)) = y^s$

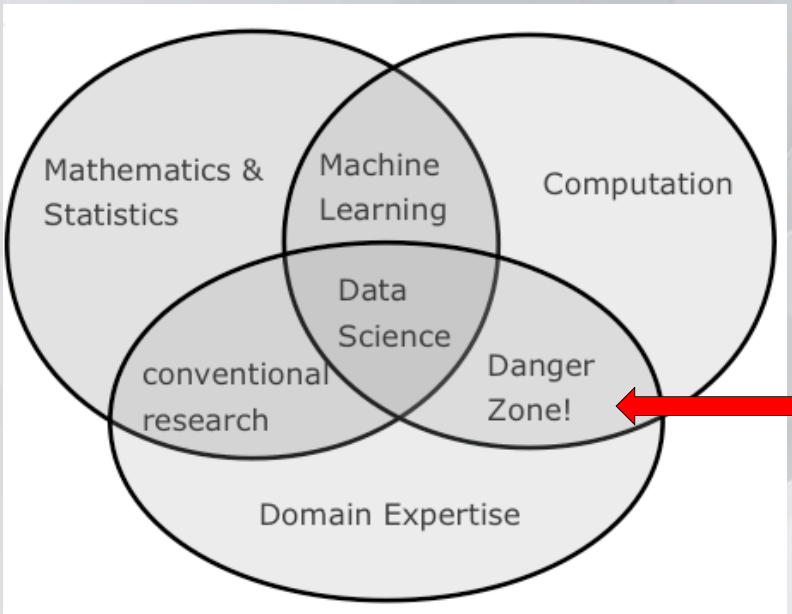
Recall: entries are encoded as a vector $(x_1, x_2, \ldots, x_d)$

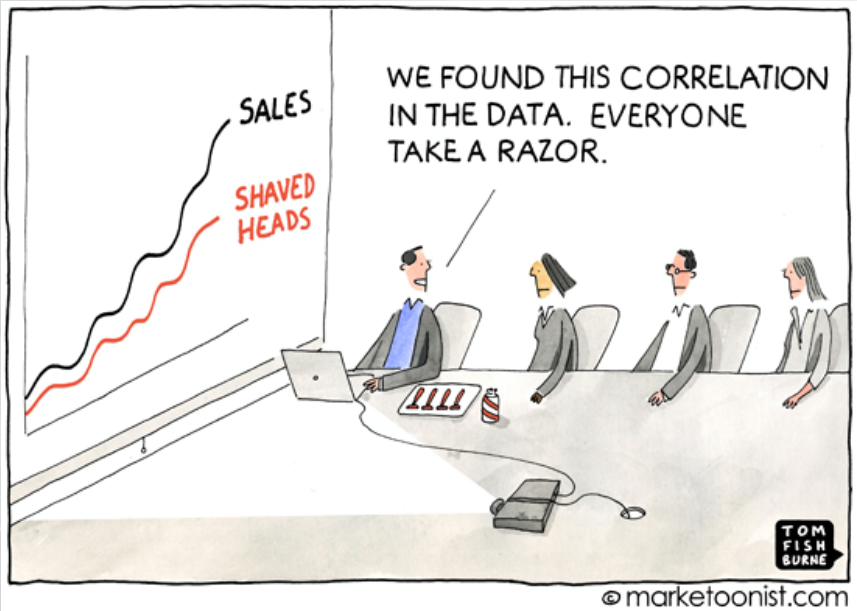
The hidden layer learns a **representation** $\mathbf{h}(\cdot)$ from which linear hypothesis $\mathbf{f}(\cdot)$ can **classify source examples**.

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



The danger zone of AI





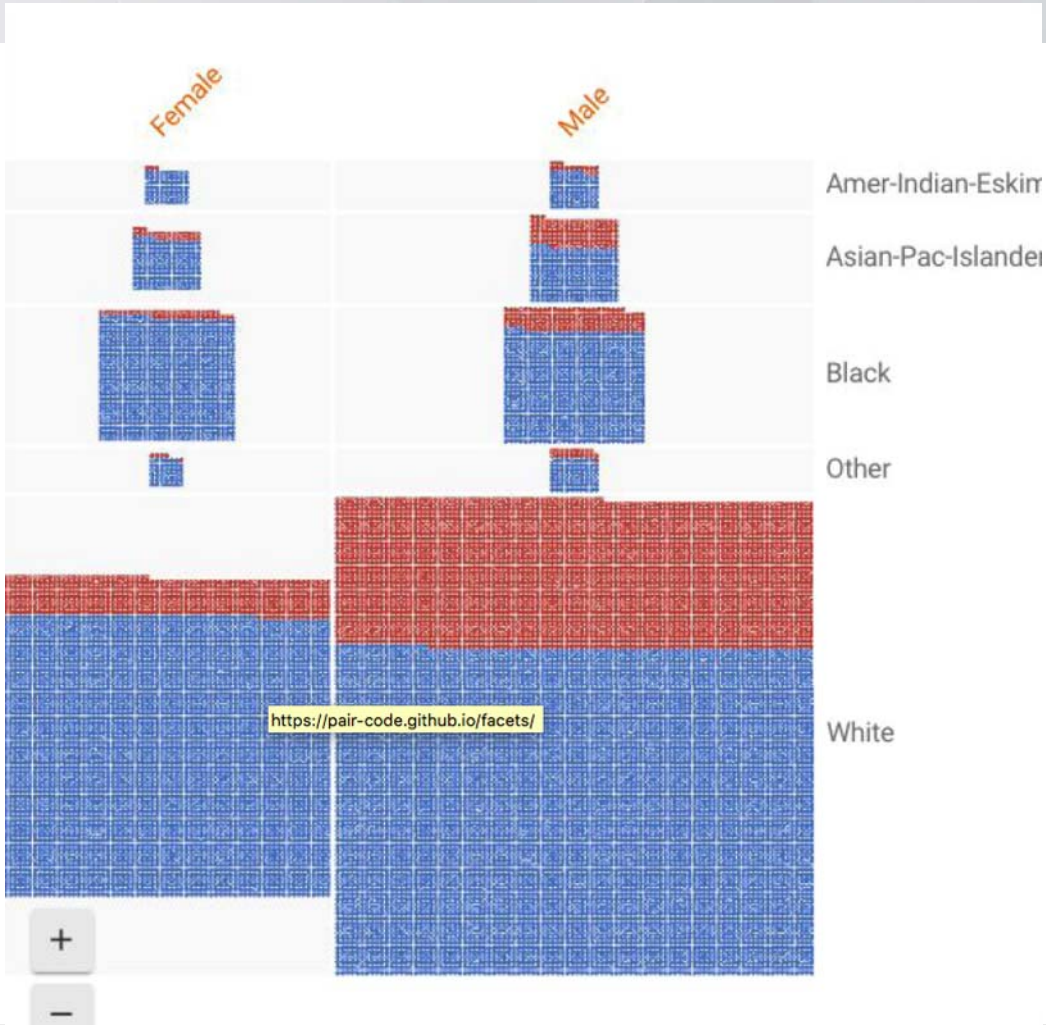
En apprentissage machine, les choses peuvent mal aller

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 »

Un défis de l'intelligence artificielle

Équité

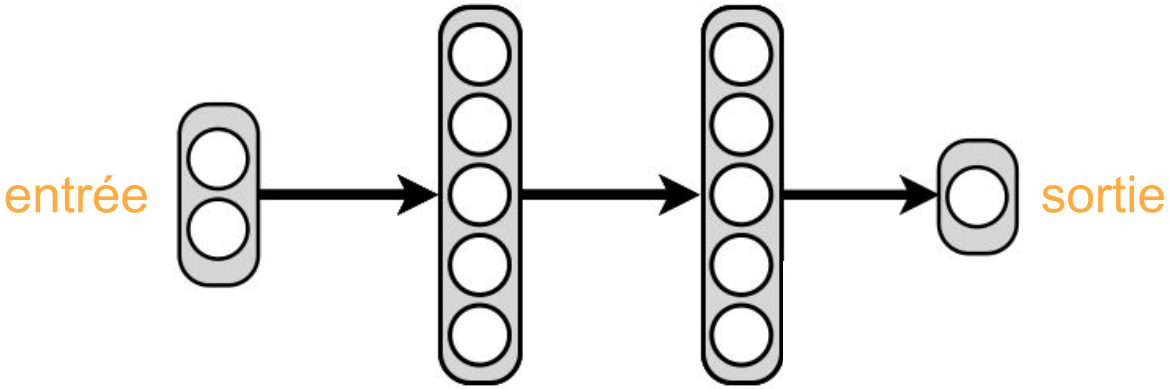


L'IA est aussi « bonne » que les données sur lesquelles elle a été entrainée

Autres défis de l'intelligence artificielle

Équité

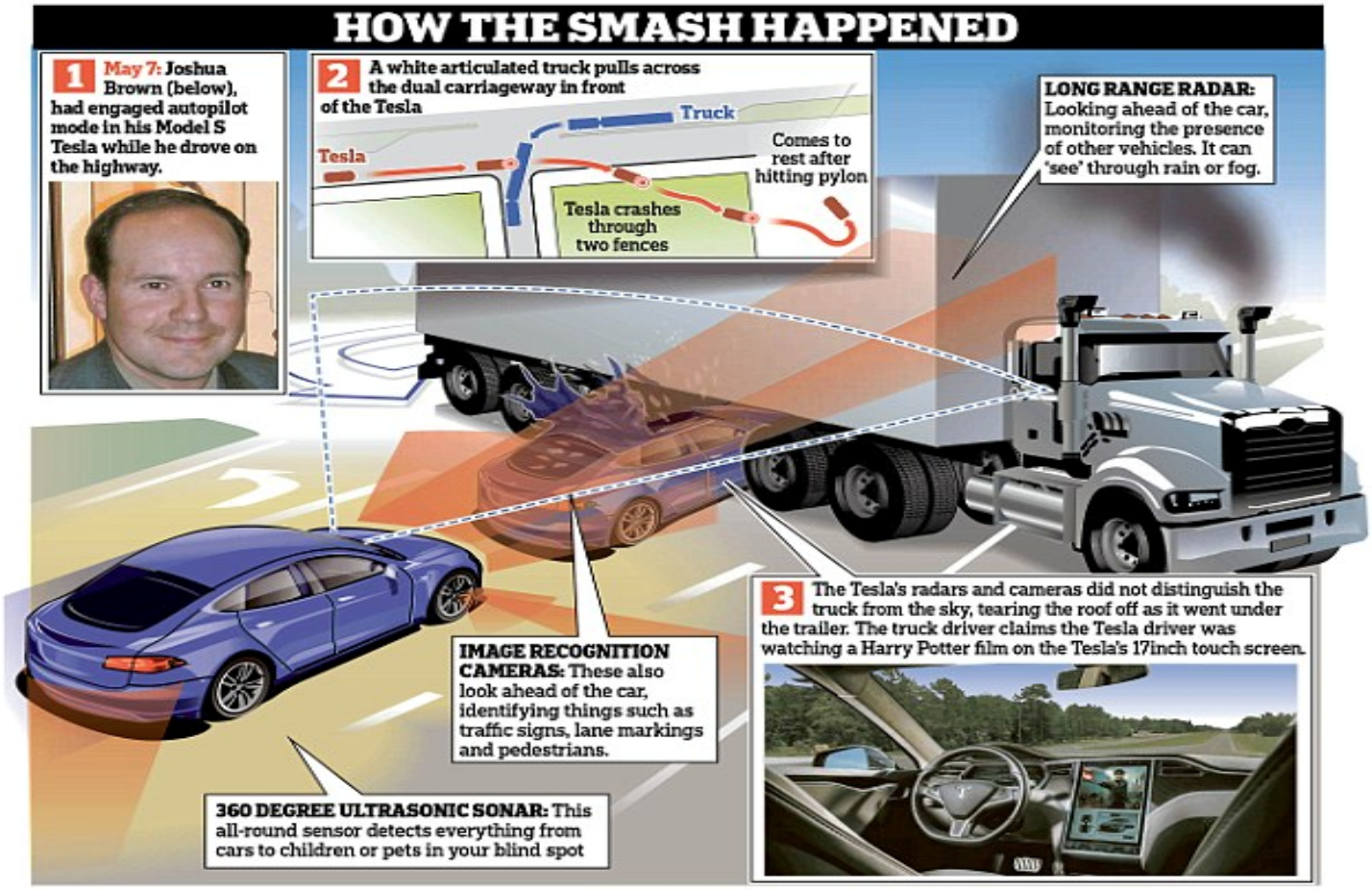
Une solution possible au manque d'équité:



Domain-Adversarial Training of Neural Networks
Ganin, Ustinova, Ajakan, Germain, Larochelle, Laviolette, Marchand, Lempitsky, 2017

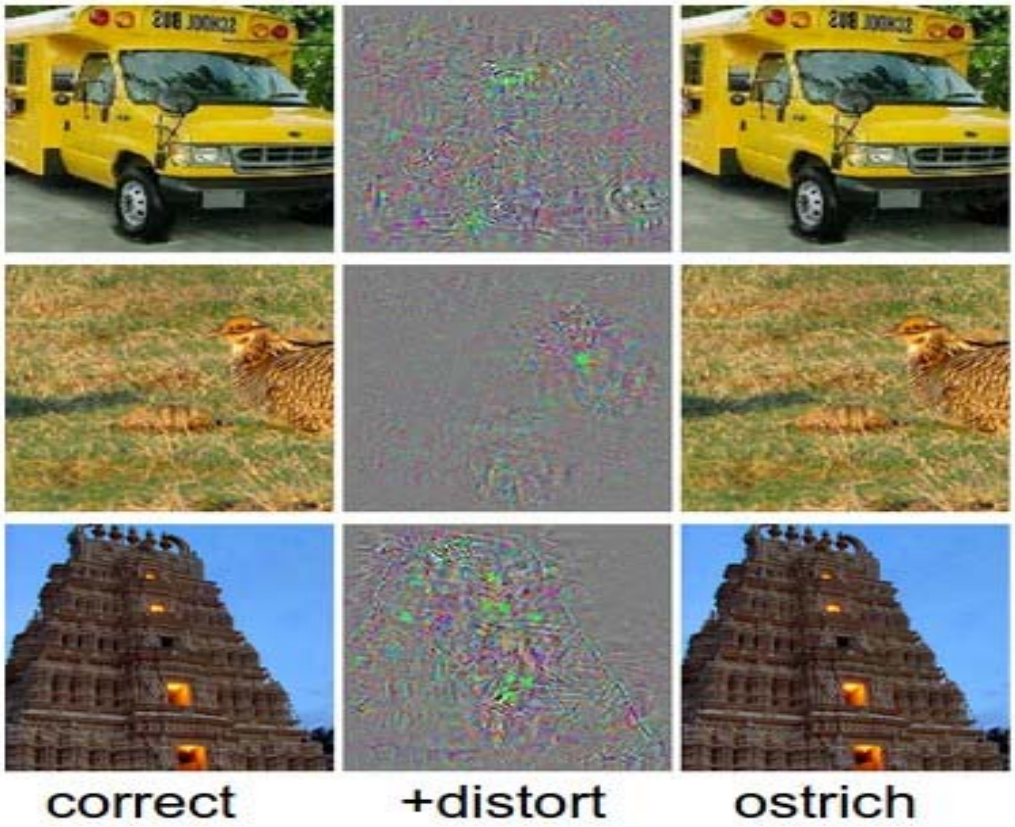
En apprentissage automatique, les choses peuvent mal aller

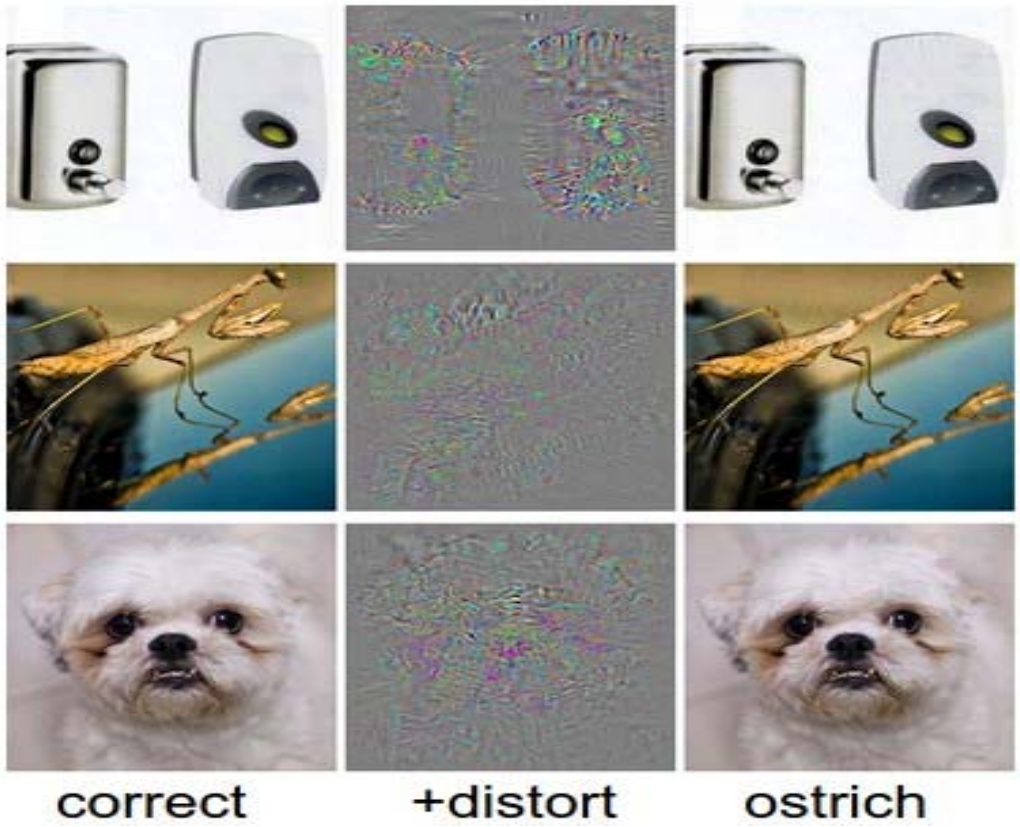
Evènements rares



Autres défis de l'intelligence artificielle

La robustesse





Autres défis de l'intelligence artificielle

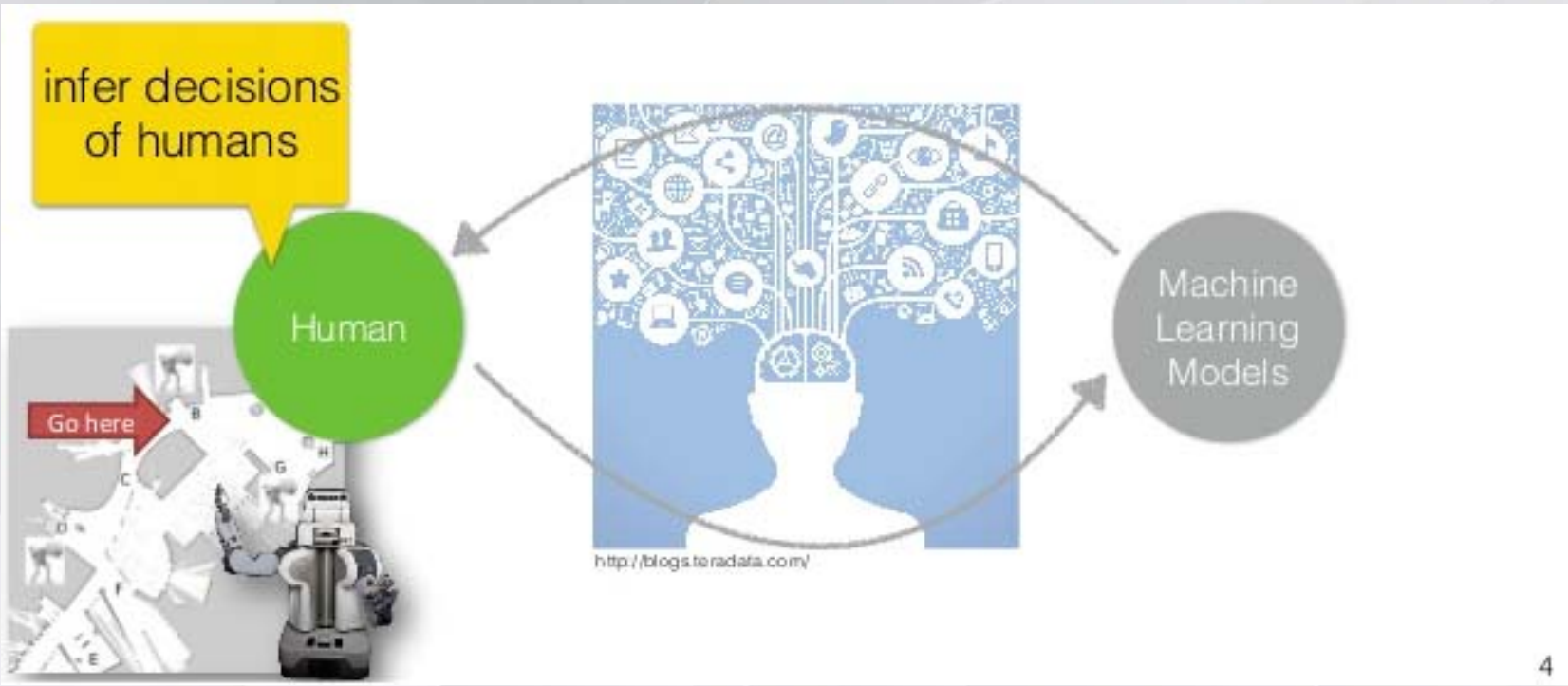
La robustesse

Le cas de TAY, l'intelligence artificielle "innocente" de Microsoft

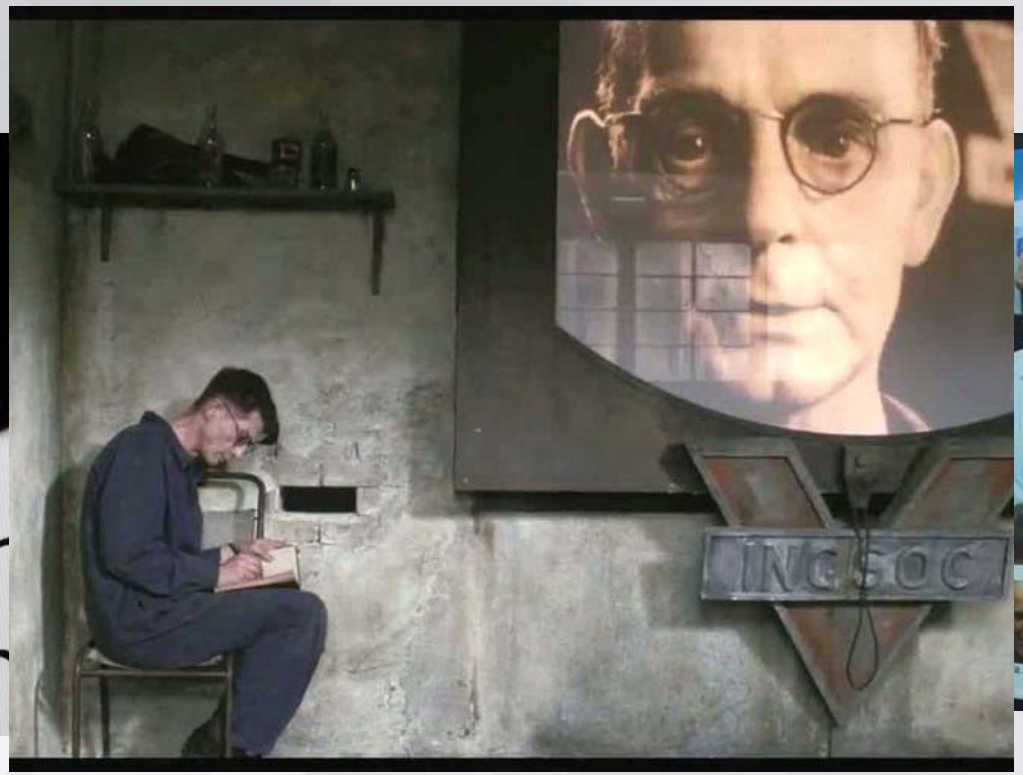
Autre défi de l'intelligence artificielle

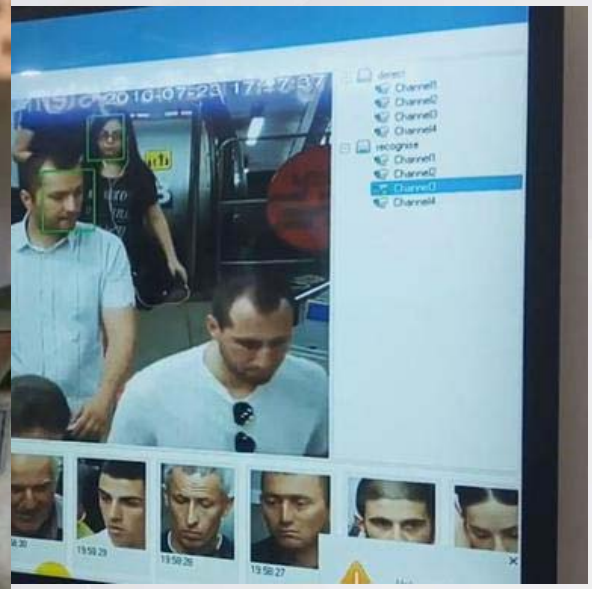
L'interprétabilité

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



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." »

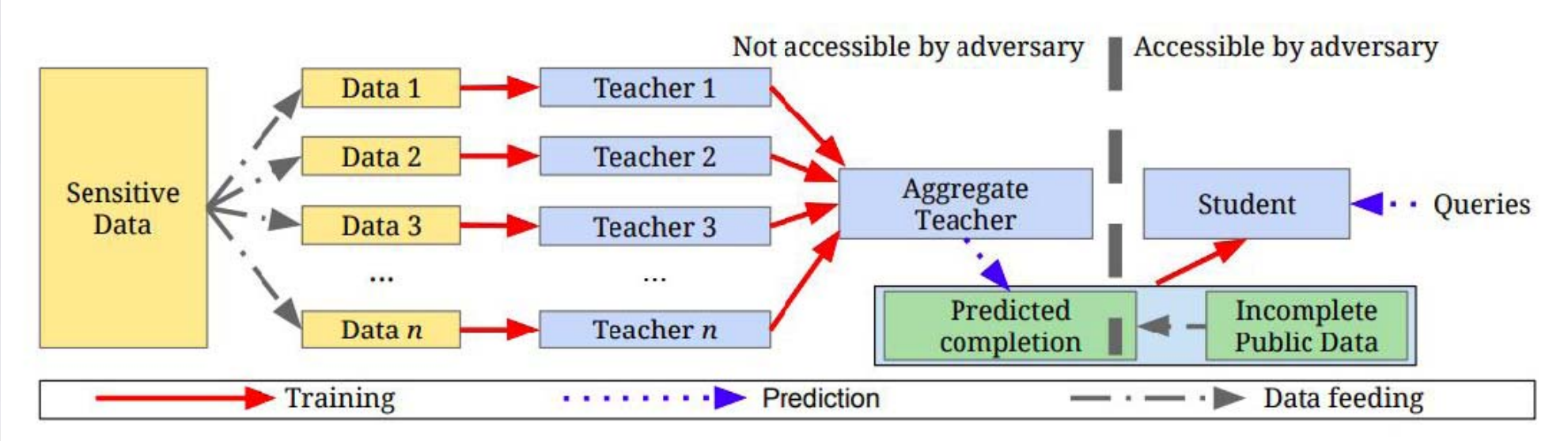
Droit individuel à la confidentialité et intérêt collectif

- Renoncer aux données, c'est se couper de grandes possibilités !
- On doit chercher un compromis entre la protection du citoyen et l'intérêt collectif.
- a
- 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

Autres défis de l'intelligence artificielle

La confidentialité des données

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

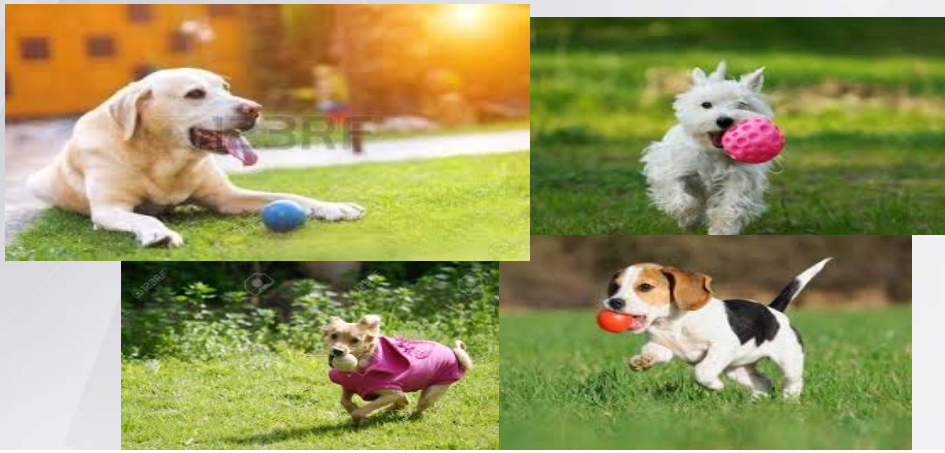


Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data
Papernot, Abadi, Erlingsson, Goodfellow, Talwar, 2017

Machine Learning and Big Data danger zone

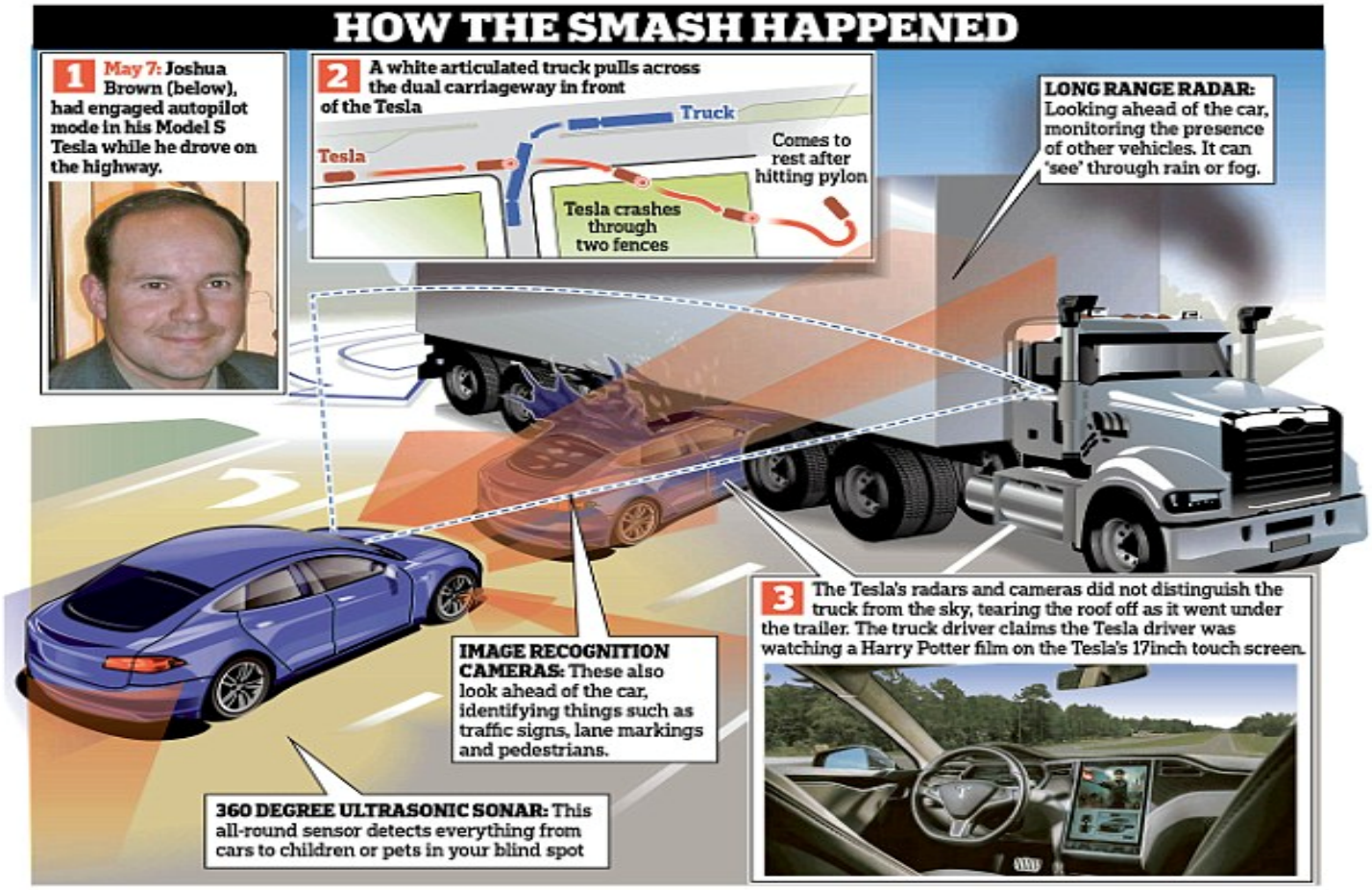
Correlation is not causality

How the data has been collected?
Data is supposed to be obtained iid.
i.e., each example of the training data has been obtained out of a unique (unknown) probability distribution and independently of all the other obtained example
Idem for the examples “to come”.



Machine Learning and Big Data danger zone

Rare events



AI challenges

- **Confidentiality:**

Tank's to Hugo Larochelle for some of the future figure

Even if we anonymise the data, we can not inssure confidentiality

- The Netflix case

- The Sweeney-2000 case
  - Medical information on 135 000 employees of Massachusetts.
  - An anonymised version shared for research purposes.
  - No personal information, but some individual characteristics are fulfill.
  - By crossing with a voting list, Dr. Latyana Sweeney identified William Weld, the state governor and therefore has access to its medical record.

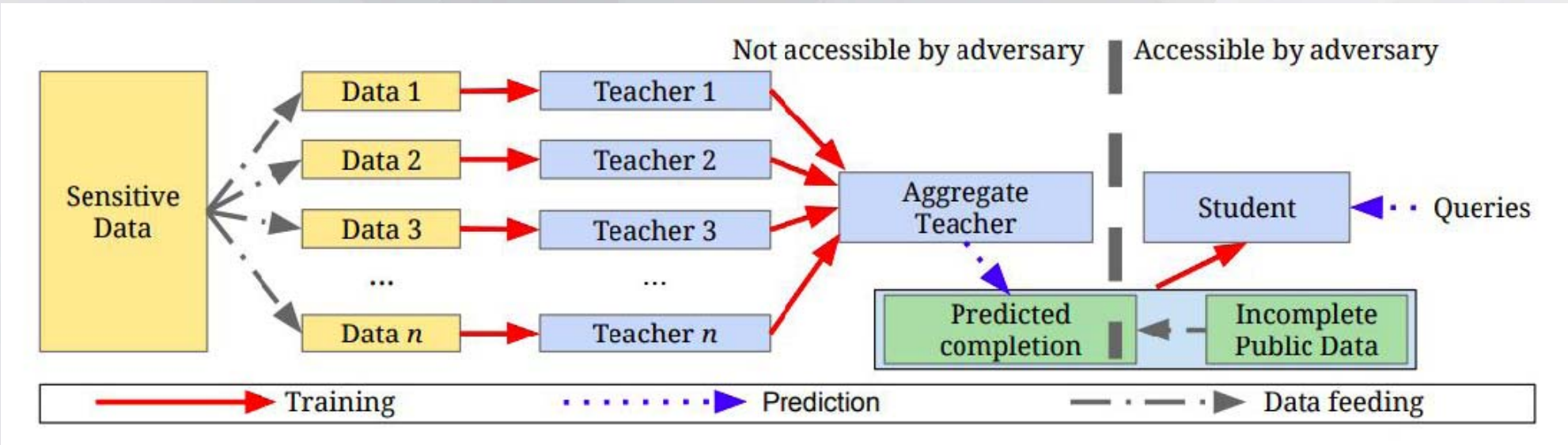
« According to the Cambridge Voter list, six people had his particular birth date; only three of them were men; and, he was the only one in his 5-digit ZIP code." »

AI challenges

- **Confidentiality (cont):** the data is « compressed » in a neural network

Tank's to Hugo Larochelle for some of the future figure

Confidentiality

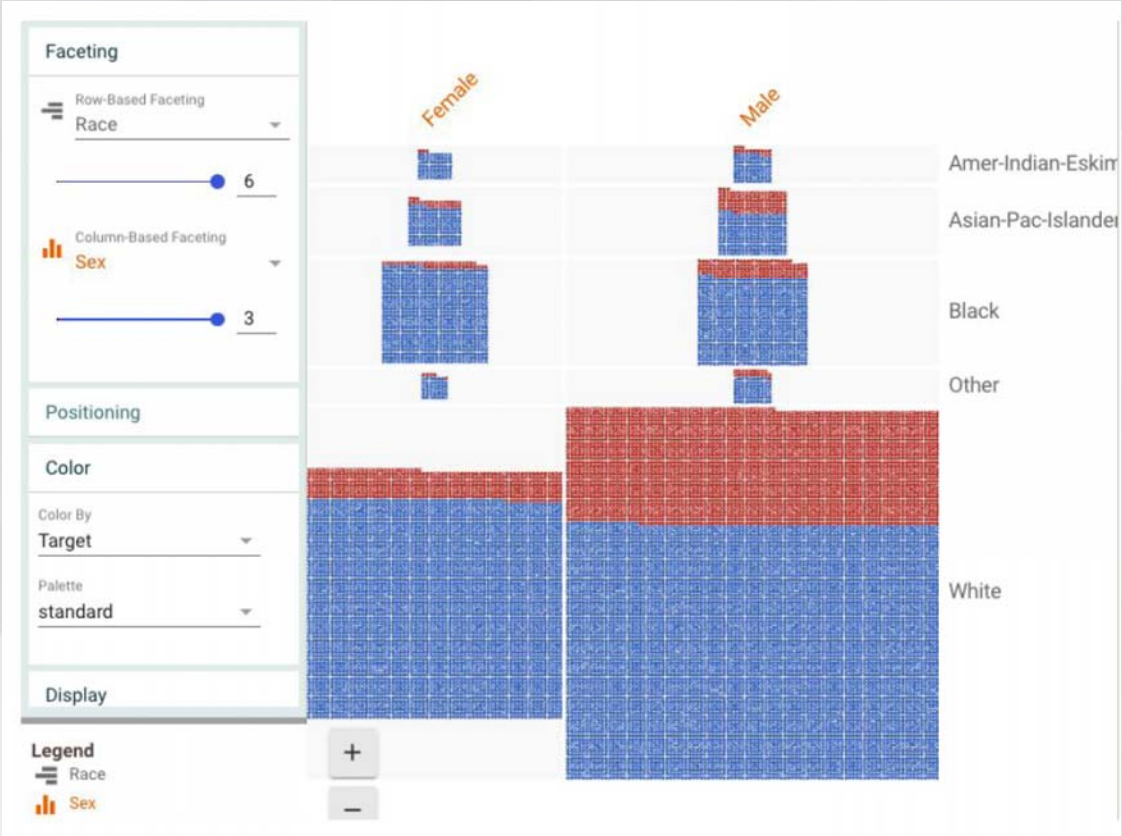


Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data
Papernot, Abadi, Erlingsson, Goodfellow, Talwar, 2017

AI challenges

- Confidentiality: the data is « compressed » in a neural network
- **Equity:** if the data do not well represent the population, so will the associated AI

Equity

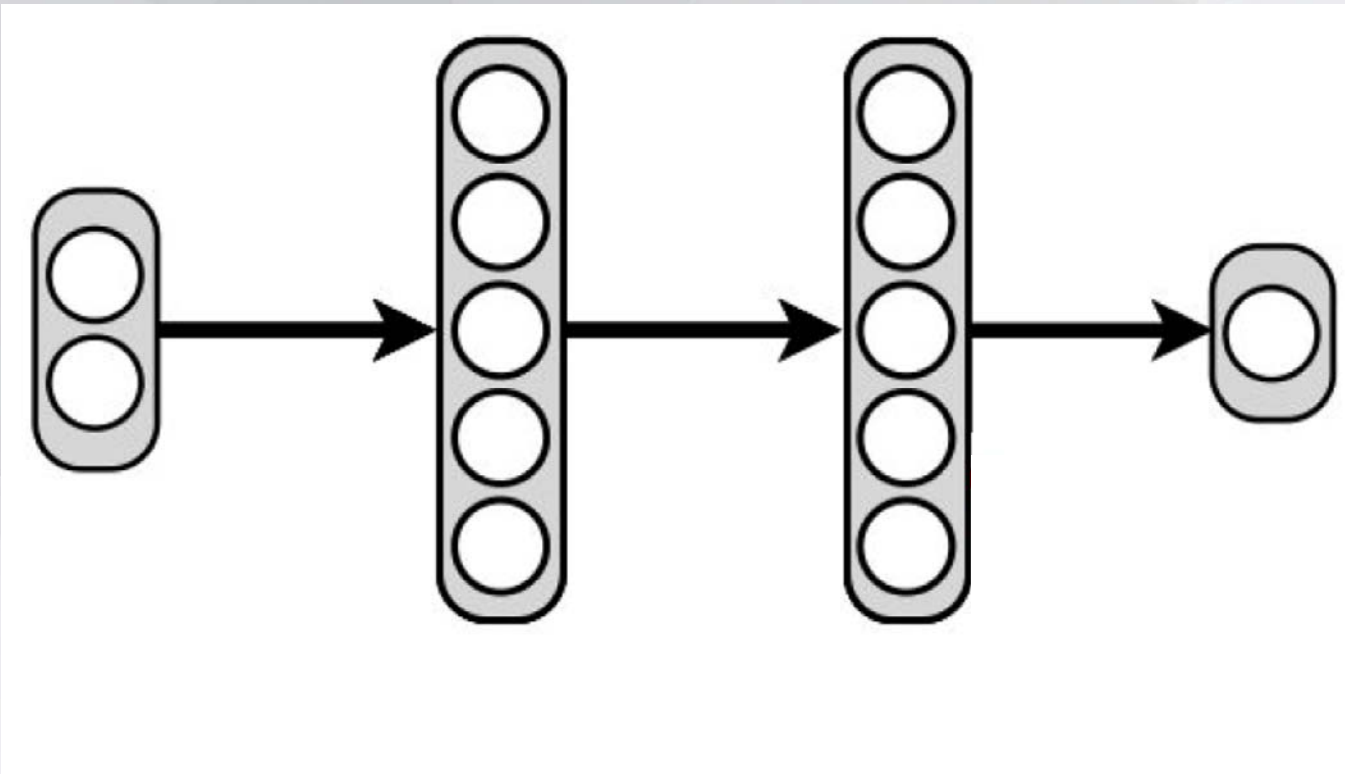


The AI can be better
Than the data on which it has been trained

Facets Dive, People+AI Research Initiative, 2017

Equity

Input



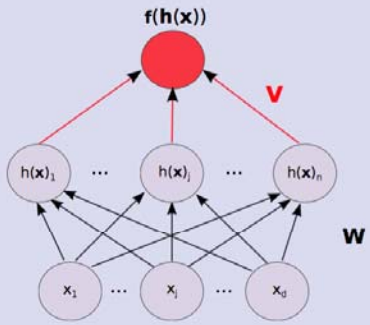
Output

Information not to be taken into account

Domain-Adversarial Training of Neural Networks
Ganin, Ustinova, Ajakan, Germain, Larochelle, Laviolette, Marchand, Lempitsky, 2017

An adversarial setting

Standard Neural Network (NN)



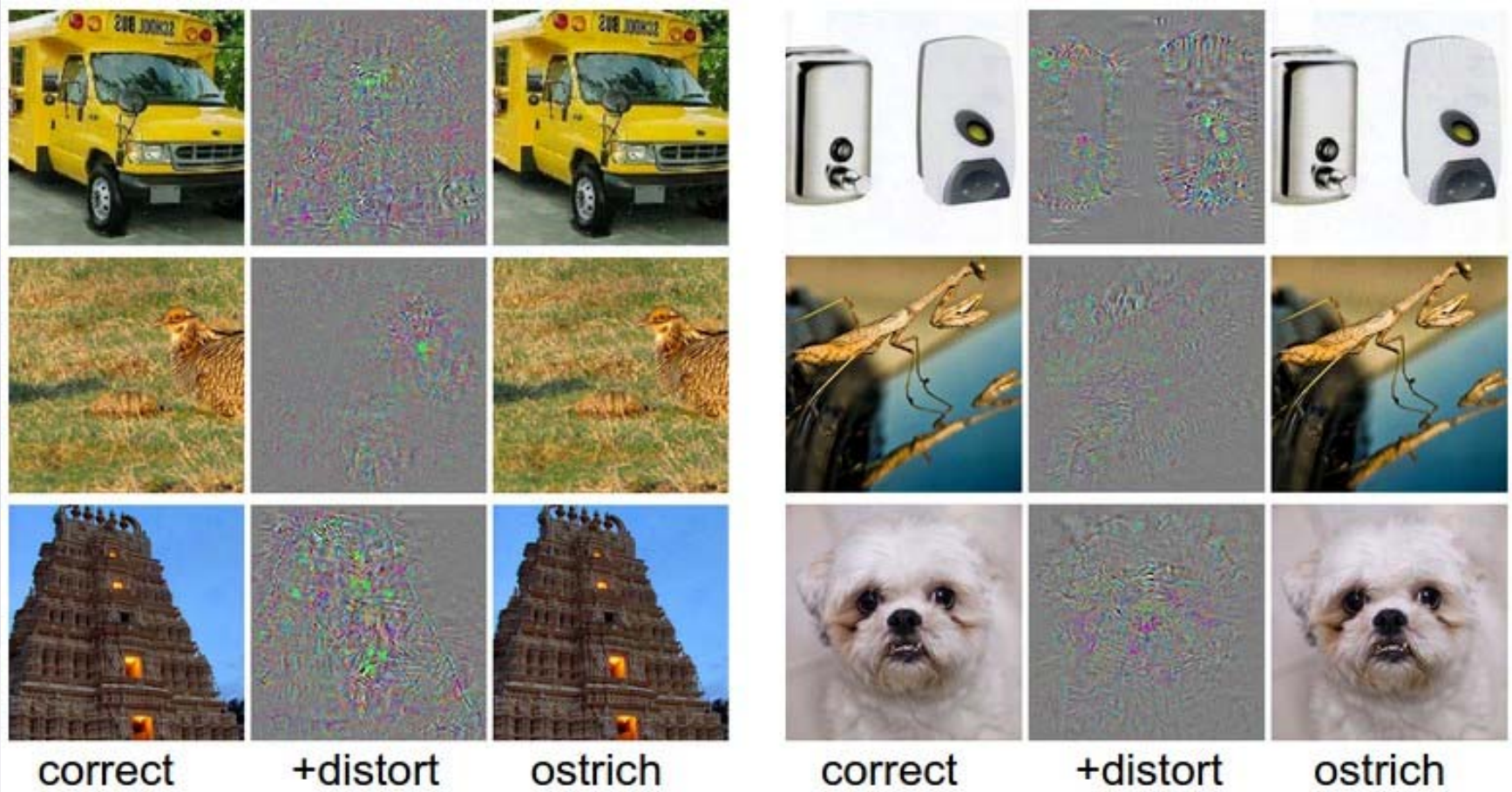
Domain-Adversarial Neural Networks (DANN)

FL et al.

AI challenges

- Confidentiality: the data is « compressed » in the neural network
- Equity: if the data do not well represent the population, so will the associated IA
- **Robustness:** The AI can be fooled in presence of “altered data”

Robustness



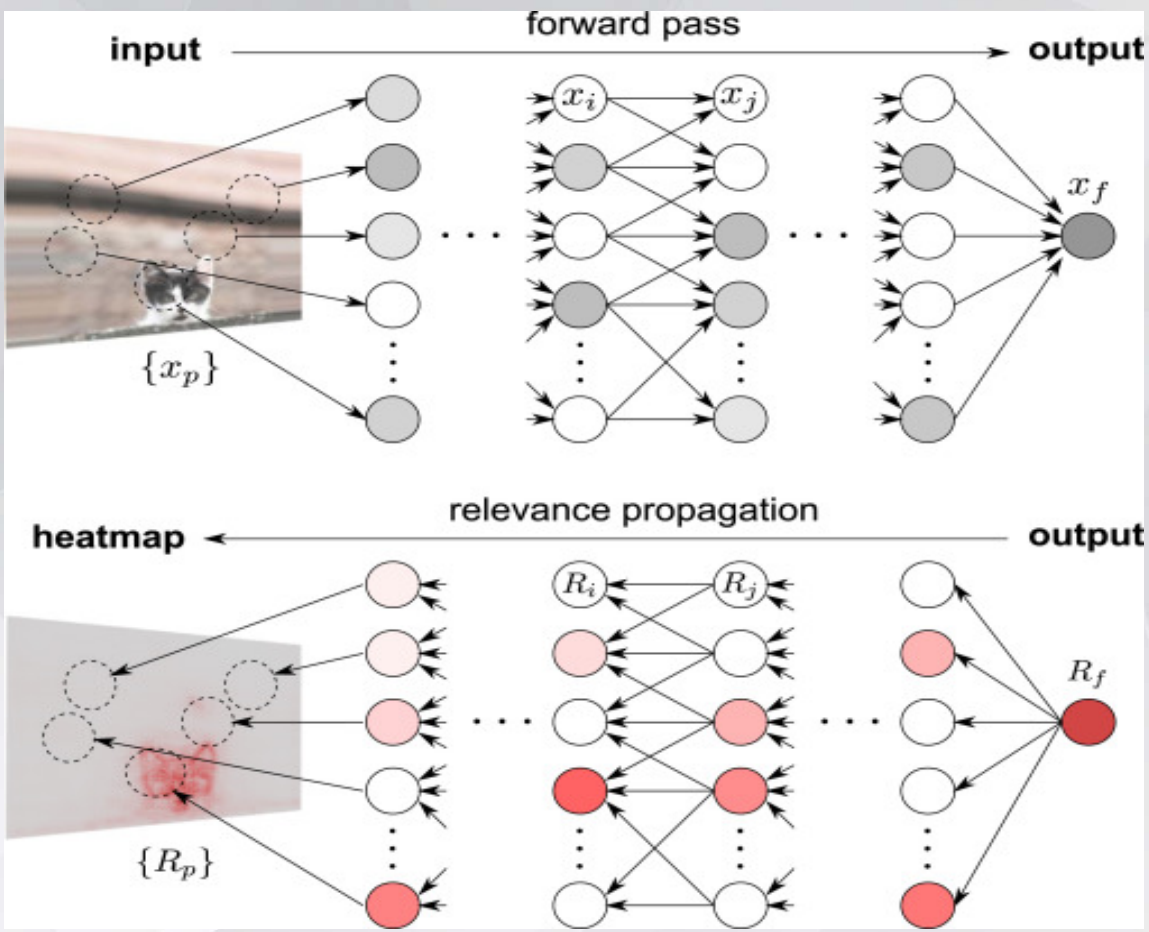
Robustness

The case of TAY, the “innocent” AI Chatbot of Microsoft

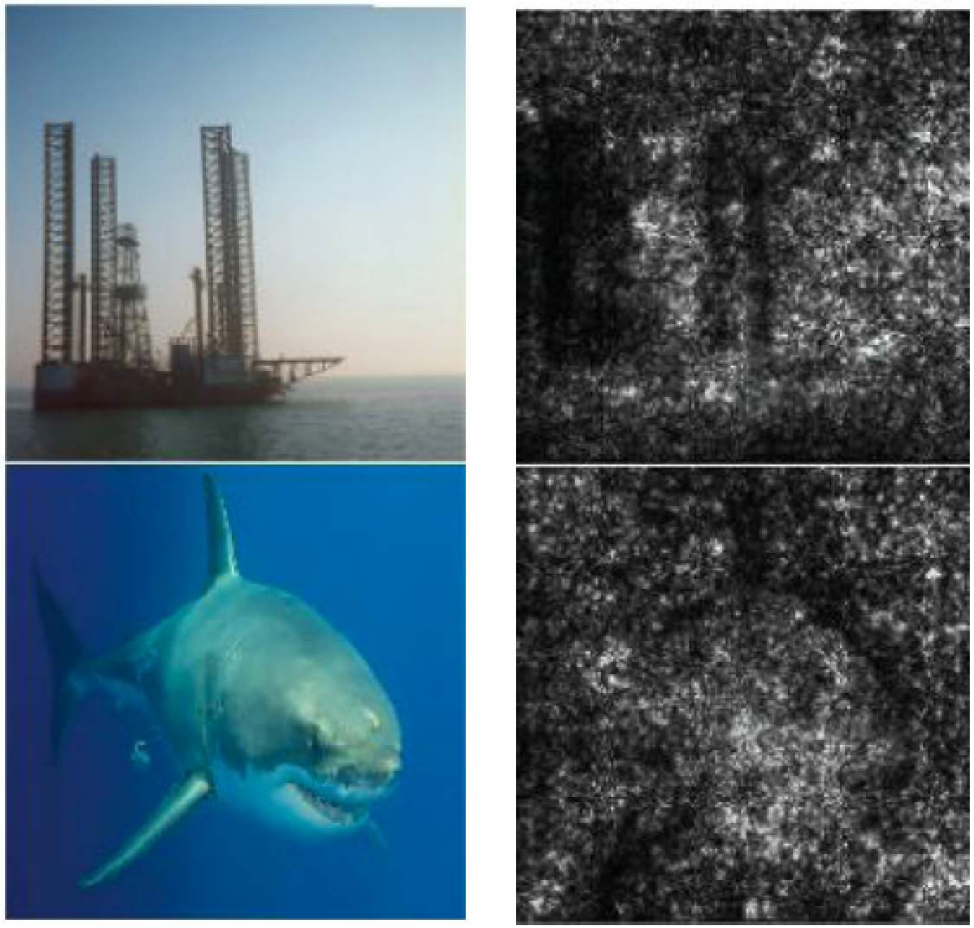
AI challenges

- Confidentiality: the data is « compressed » in the neural network
- Equity: if the data do not well represent the population, so will the associated IA
- Robustness: The AI can be fooled in presence of “altered data”
- **Interpretability:** Can we figure out the “reasonning” of an AI

Interpretability



Interpretability



A Learning algorithm for genomic data

FL et

Definition

- k-mer: a sequence of k nucleotides
- $\mathcal{K}$: the set of all k-mers that are in at least one genome of $\mathcal{S}$

Each genome is represented by its **k-mer profile**, which is a binary vector indicating the presence (1) or absence (0) of each k-mer

$$\mathcal{K} = \left\{ \begin{array}{cccc} \text{CAGATA} & \text{GATAGA} & \text{GAACAG} & \text{CGATGA} \\ \text{AGATAG} & \text{AGAACA} & \text{ATAGAA} & \text{CCGGCT} \\ \text{AACAGC} & \text{TAGAAC} & \text{TTTCGG} & \text{AAATAC} \end{array} \right\}$$

$$\mathbf{x} = \text{CAGATAGAACAGC}$$

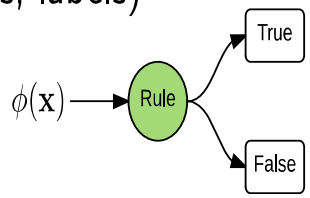
$$\phi(\mathbf{x}) = \begin{array}{|c|c|c|c|c|c|c|c|c|c|c|c|} \hline 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 \\ \hline \end{array}$$

CAGATA	TTTCGG	AGATAG	GATAGA	CGATGA	AACAGC	ATAGAA	CCGGCT	TAGAAC	GAACAG	AGAACA	AAATAC

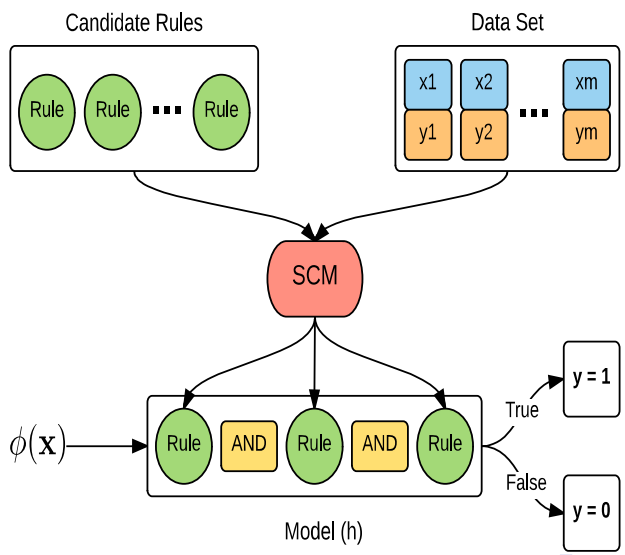
Thank's to Alexandre Drouin, for the slides

KOVER

- **Input:** A set of **boolean-valued rules** (presence/absence of k-mers) and a **data set** (examples, labels)



- **Objective:** Find the **shortest** conjunction (logical-AND) or disjunction (logical-OR) of rules that **most accurately predicts** the labels

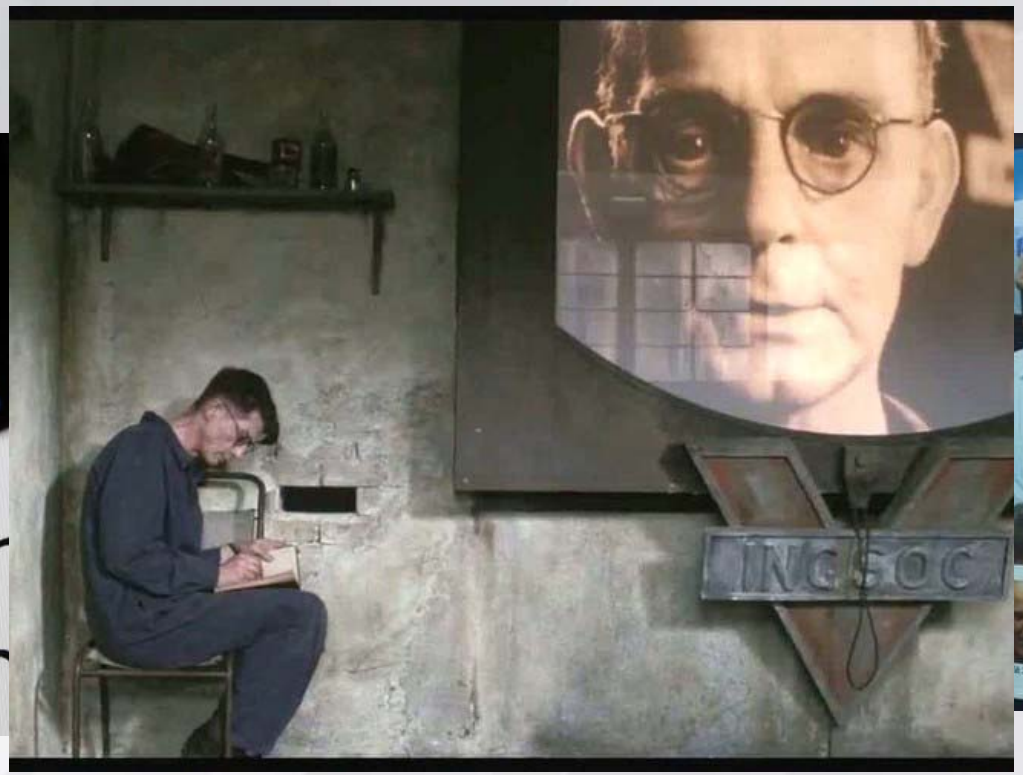


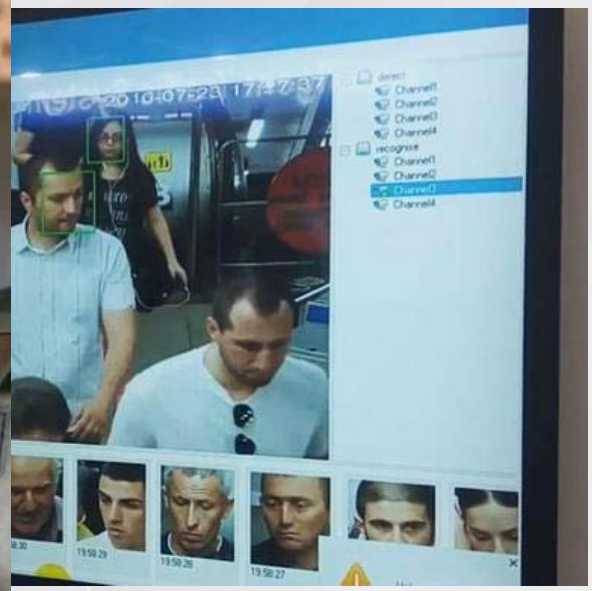
Another challenge of AI

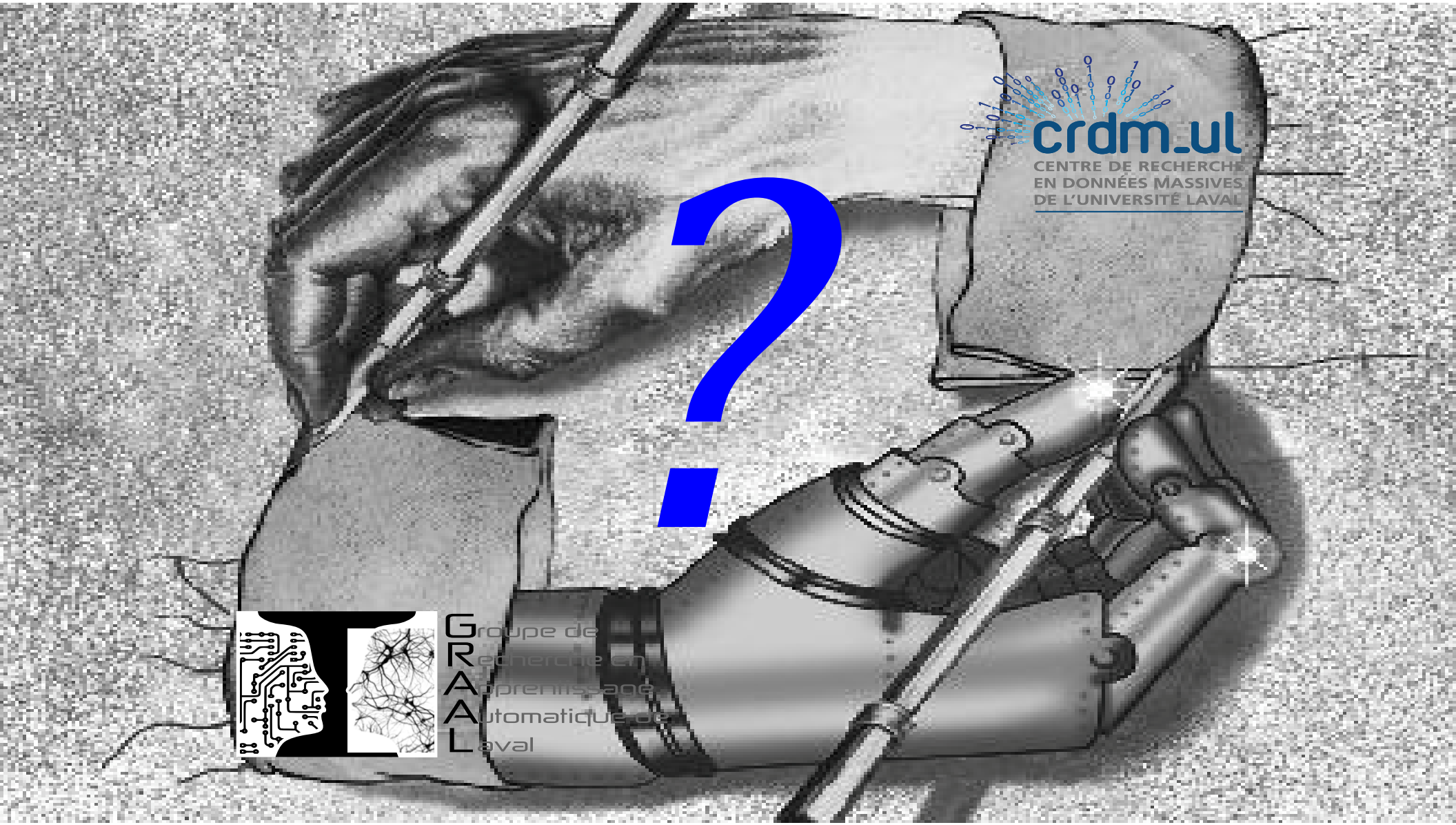
Prevent possible « bad impact » on our society

AI and sensible data

Ethic and social acceptatbility have to be in the loop






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