Données massives et intelligence artificielle,

leurs potentiels et enjeux en sciences médicales

Présenté par
François Laviolette

À la 8e journées d'étude des comités d'éthique de la recherche et de leurs partenaires

Organisée par le ministère de la Santé et des Services sociaux (MSSS)

20 octobre 2017

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

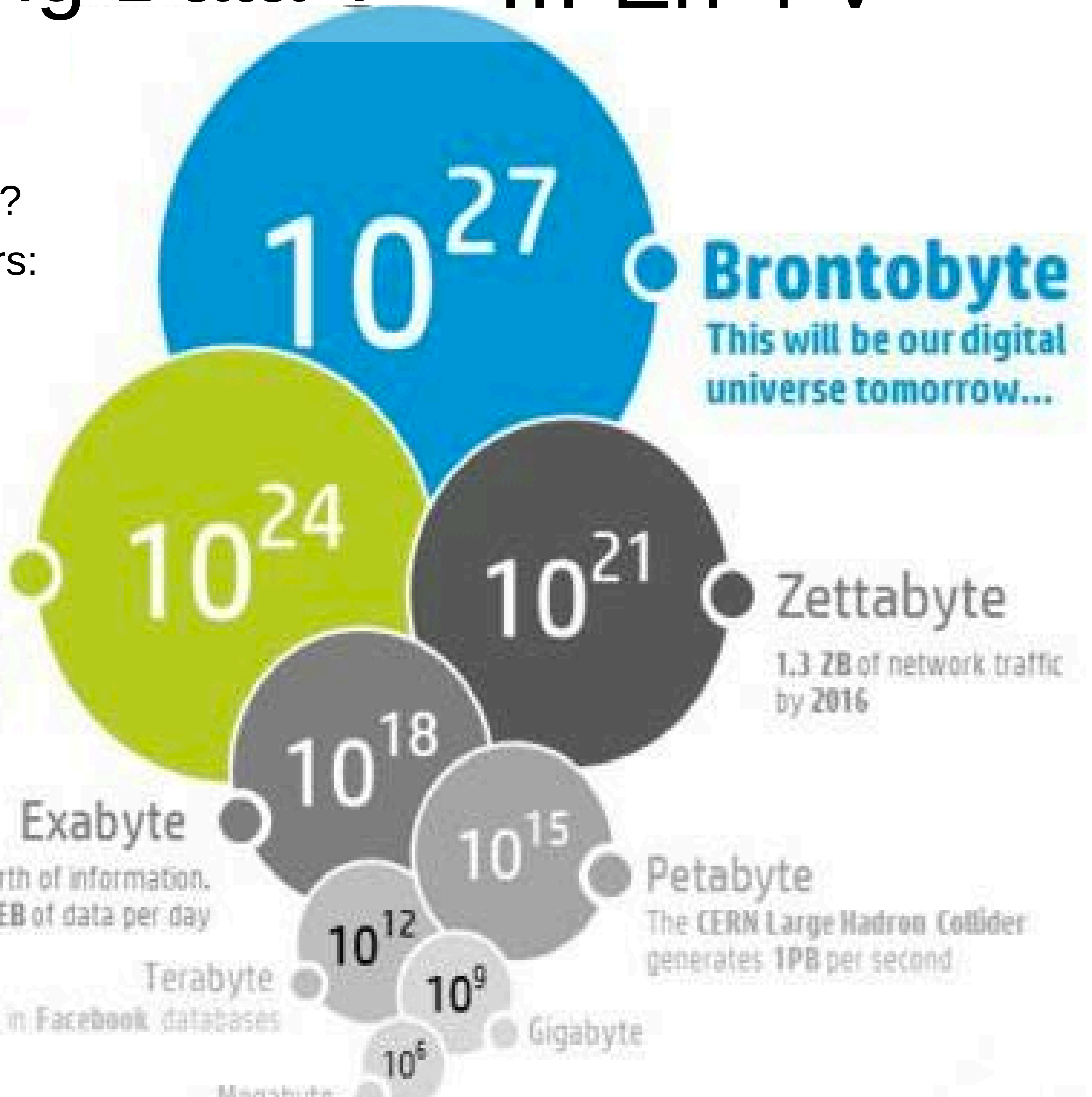
Qu’est-ce que le « Big Data » ?

- 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

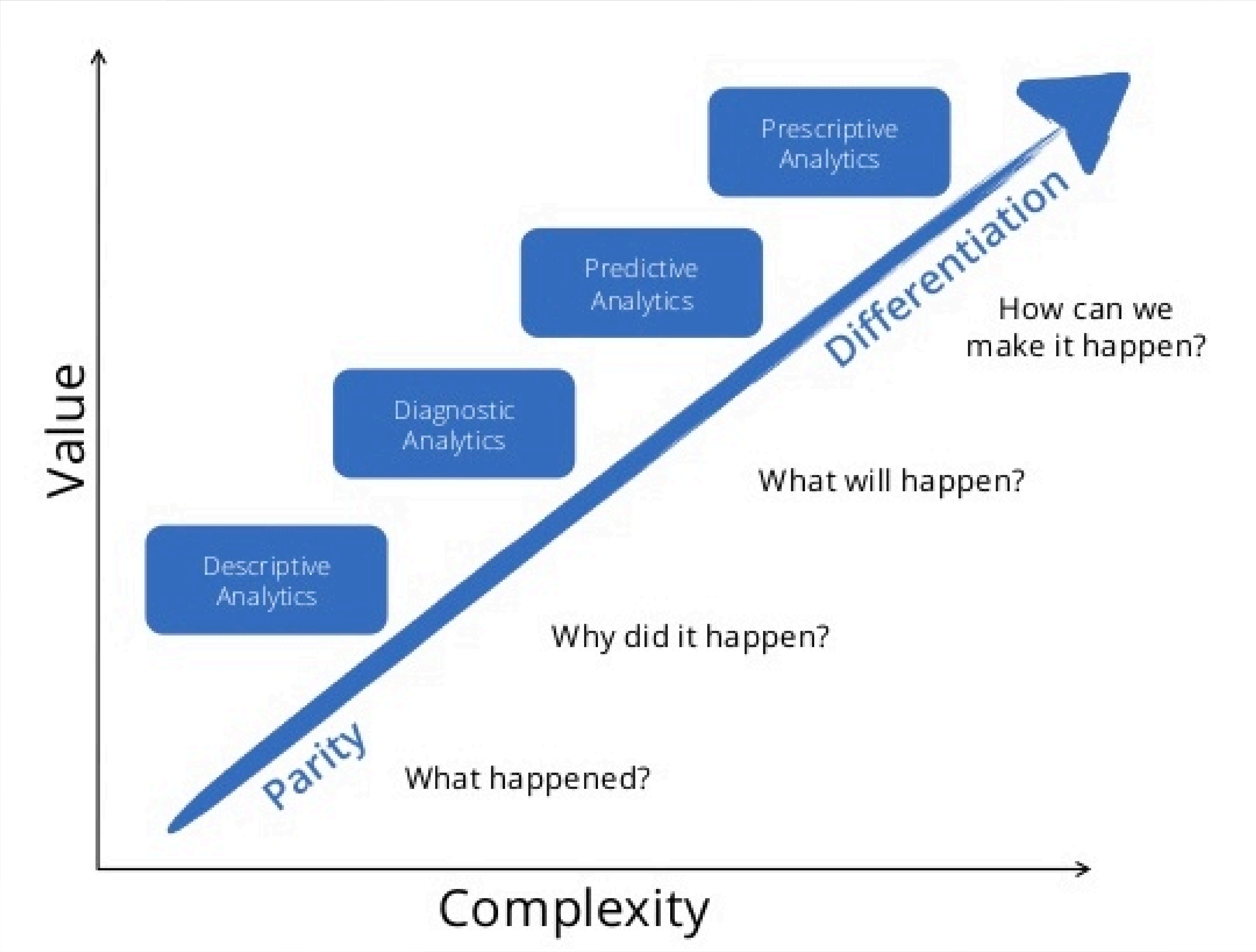
Qu’est-ce que le “Big Data”? … En 4 V

- Volume
- Vélocité
- Variété
- Véracité

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

Données provenant de projets de recherche différents,
avec des méthodologies non nécessairement compatibles

Quelle est la Valeur qu'on peut tirer du "Big Data"?



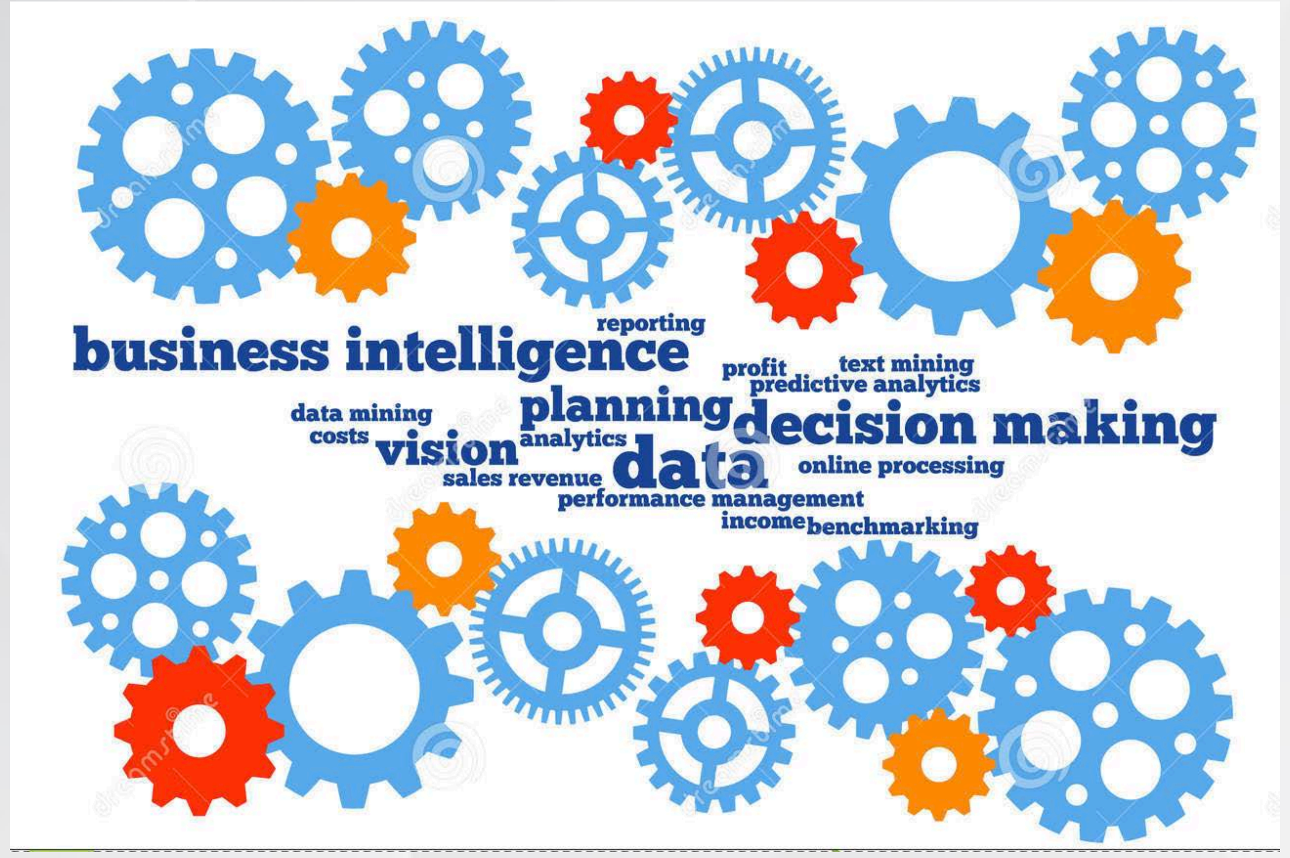
Big data is not about the size of the data, it's about the value within data

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

Les défis du “Big Data”

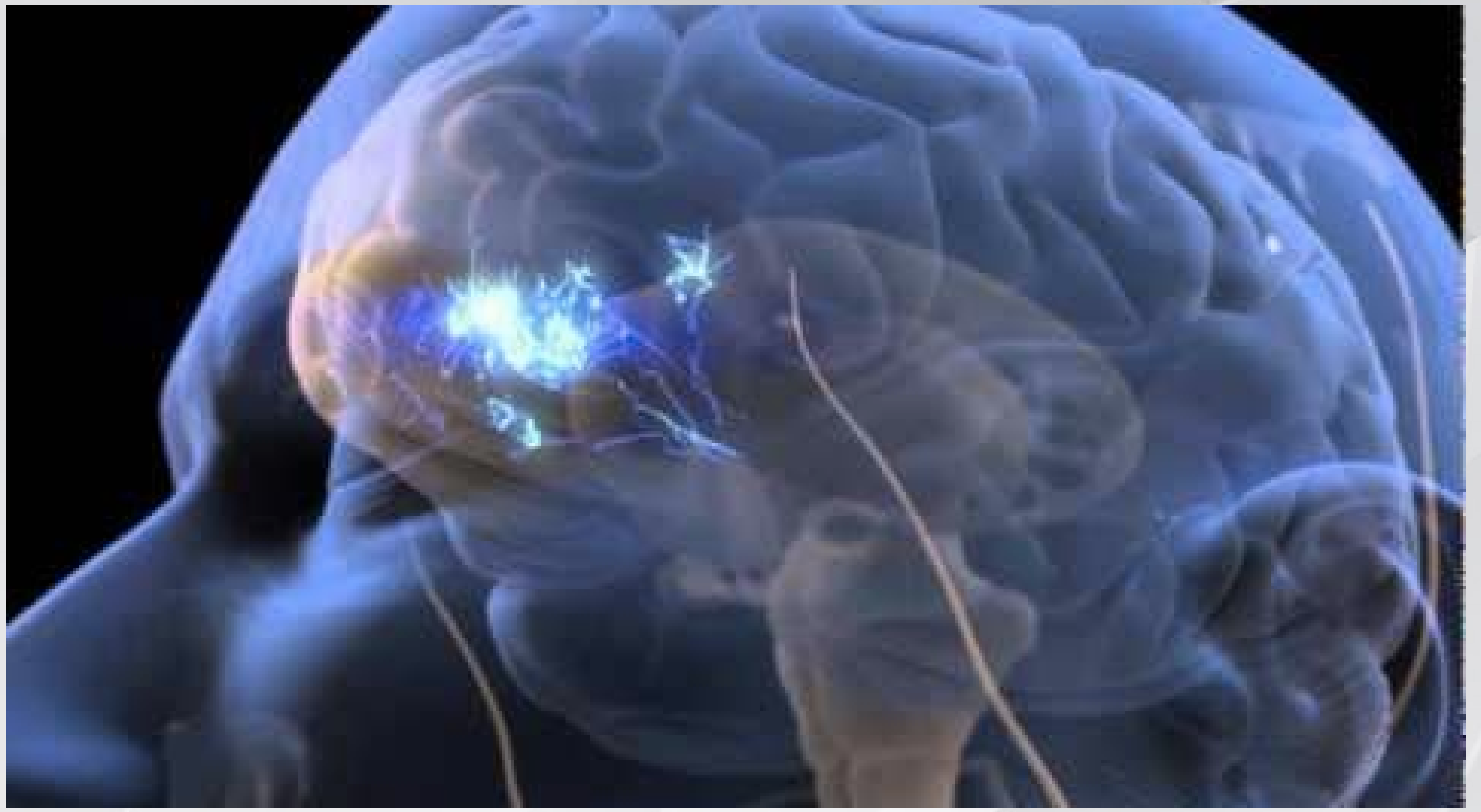
- Les données massives forcent le développement de nouvelles méthodes pour:
  - entreposer et retrouver la donnée
  - effectuer les analyses et autres calculs
  - visualiser l’information
  - réaliser les prises de décisions associées

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



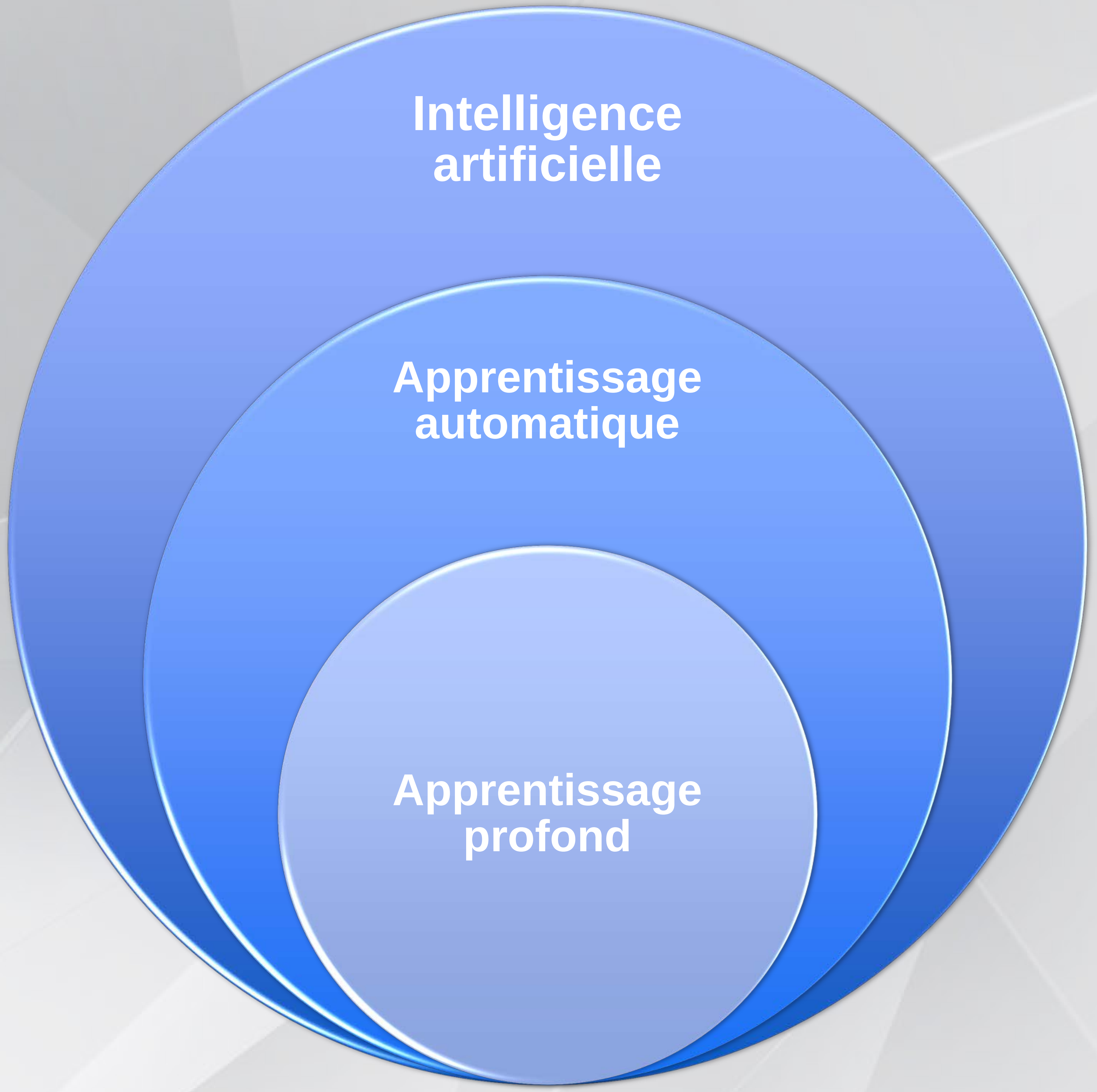
L' apprentissage automatique
vs
la recherche opérationnelle

Apprentissage automatique:
Comprendre les informations



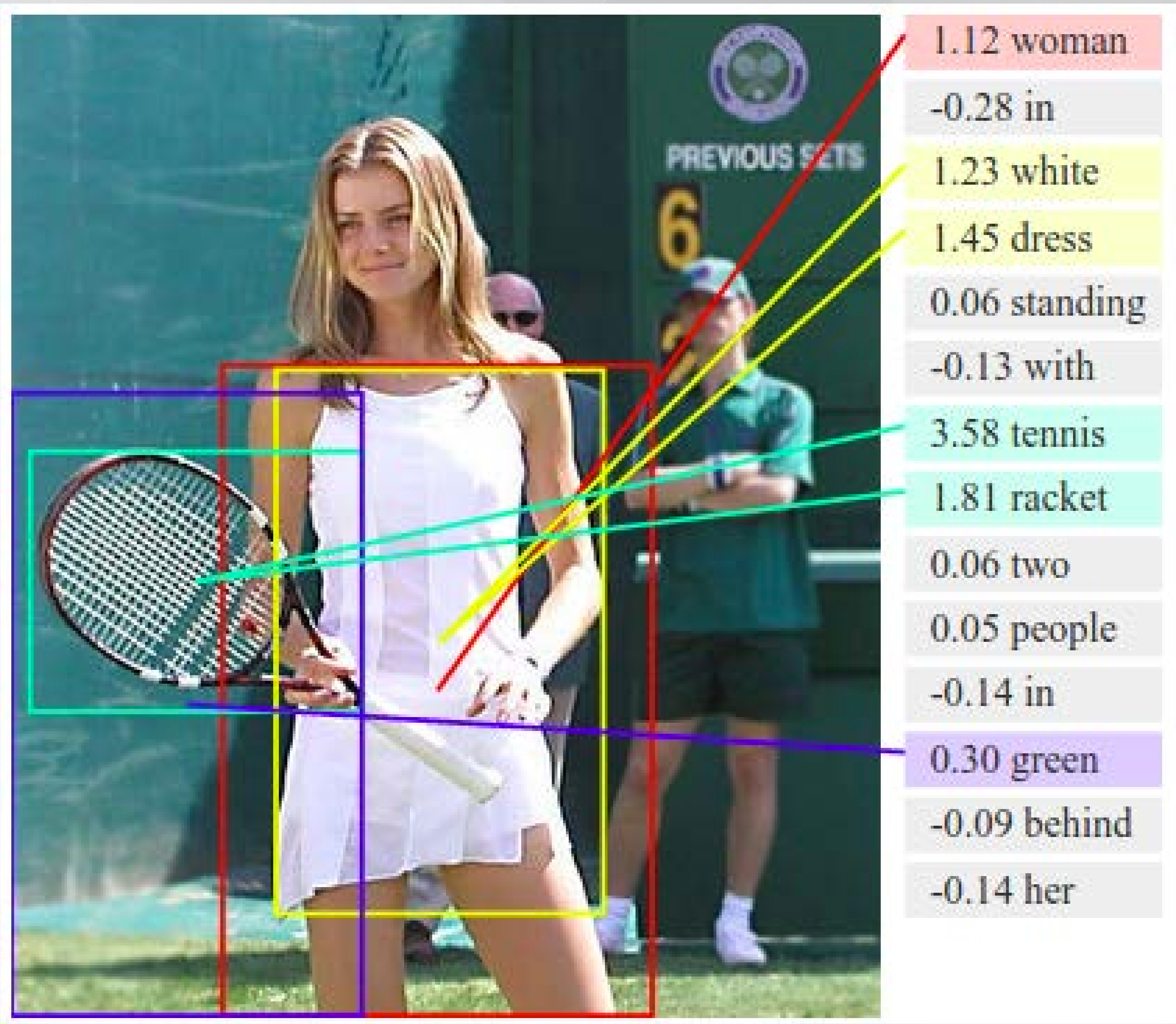
Recherche opérationnelle:
Optimise la prise de décision

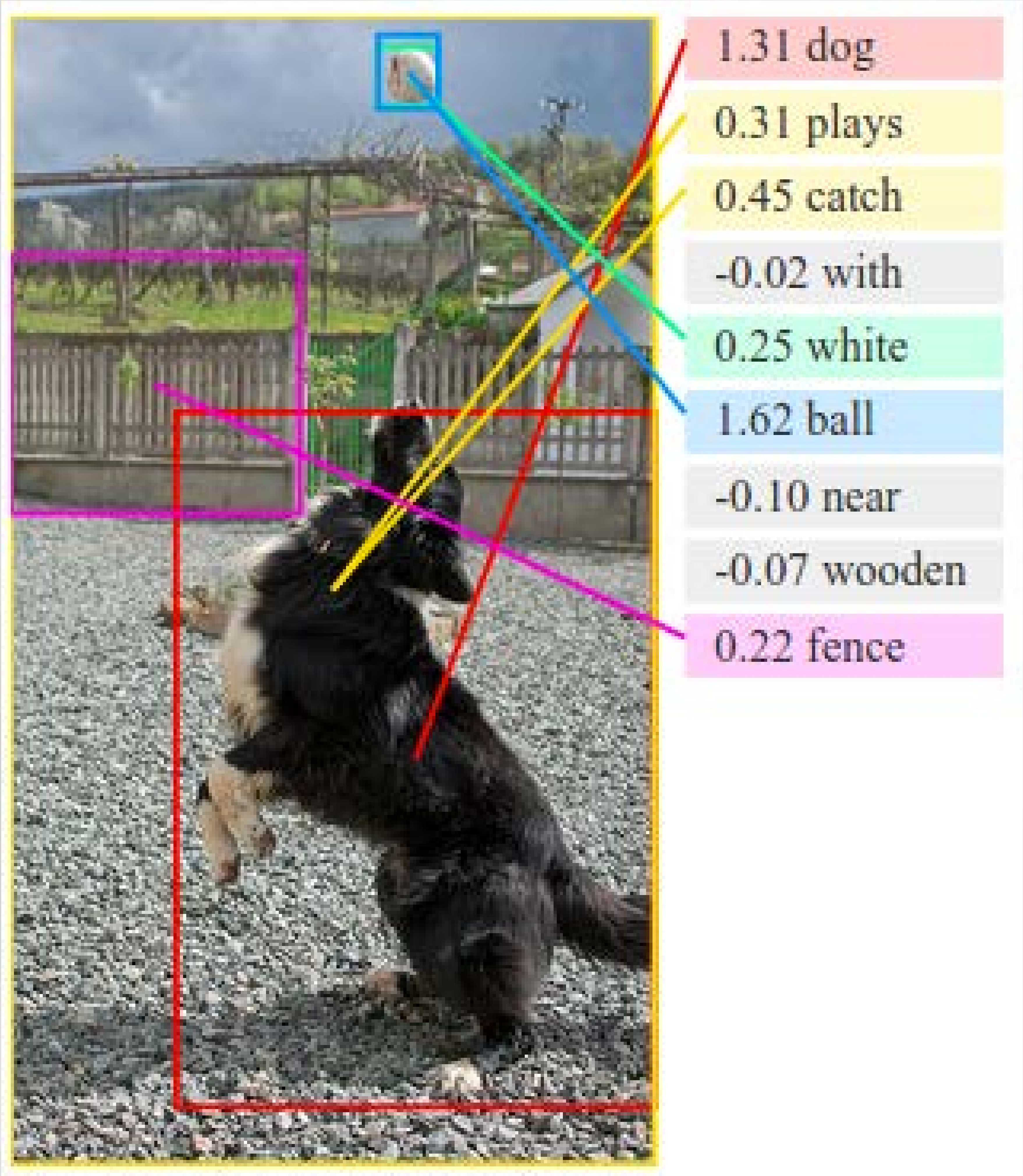
L’intelligence artificielle et ses apprentissages



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

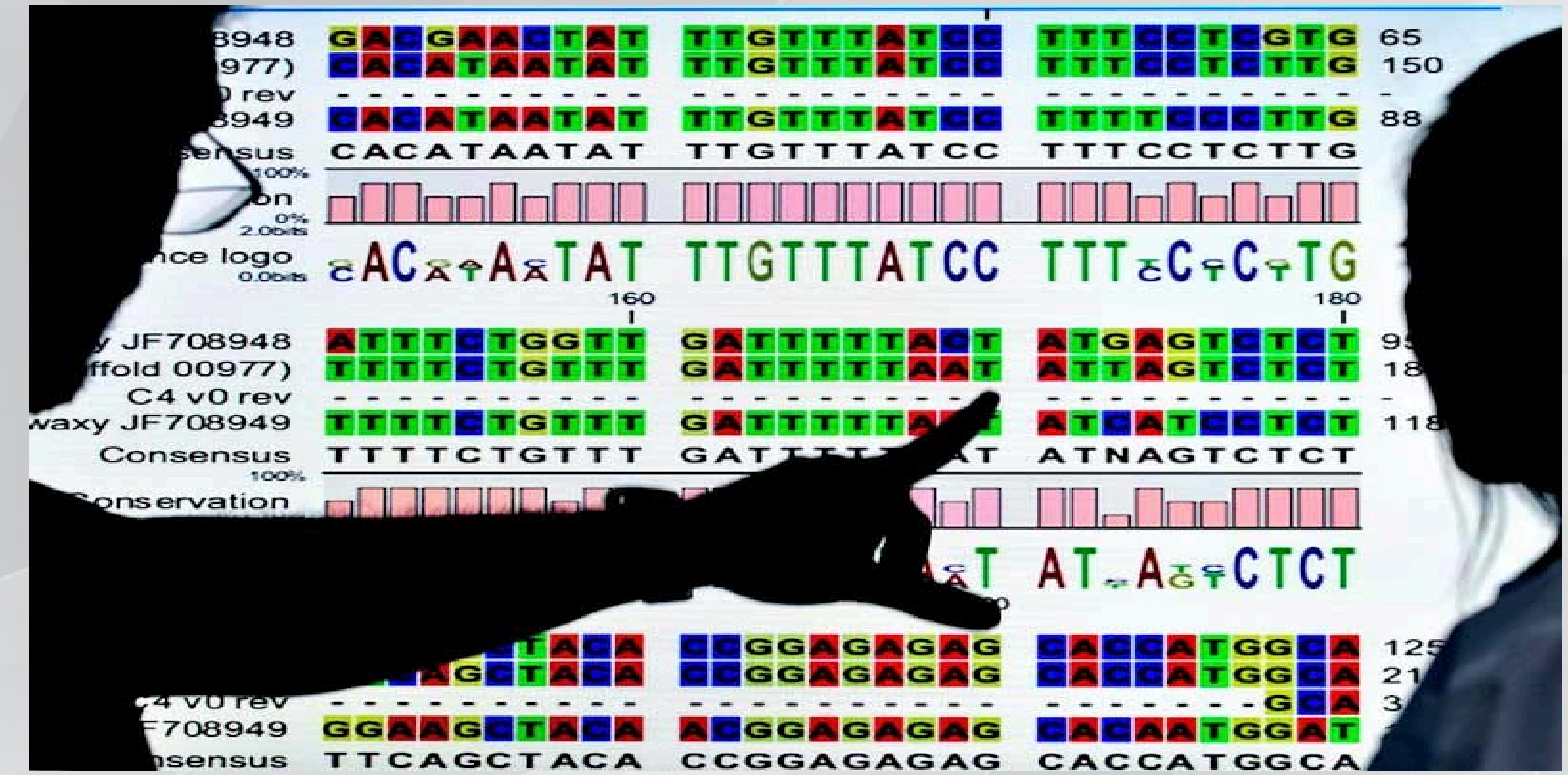
Interprétation automatique de contenu d’images

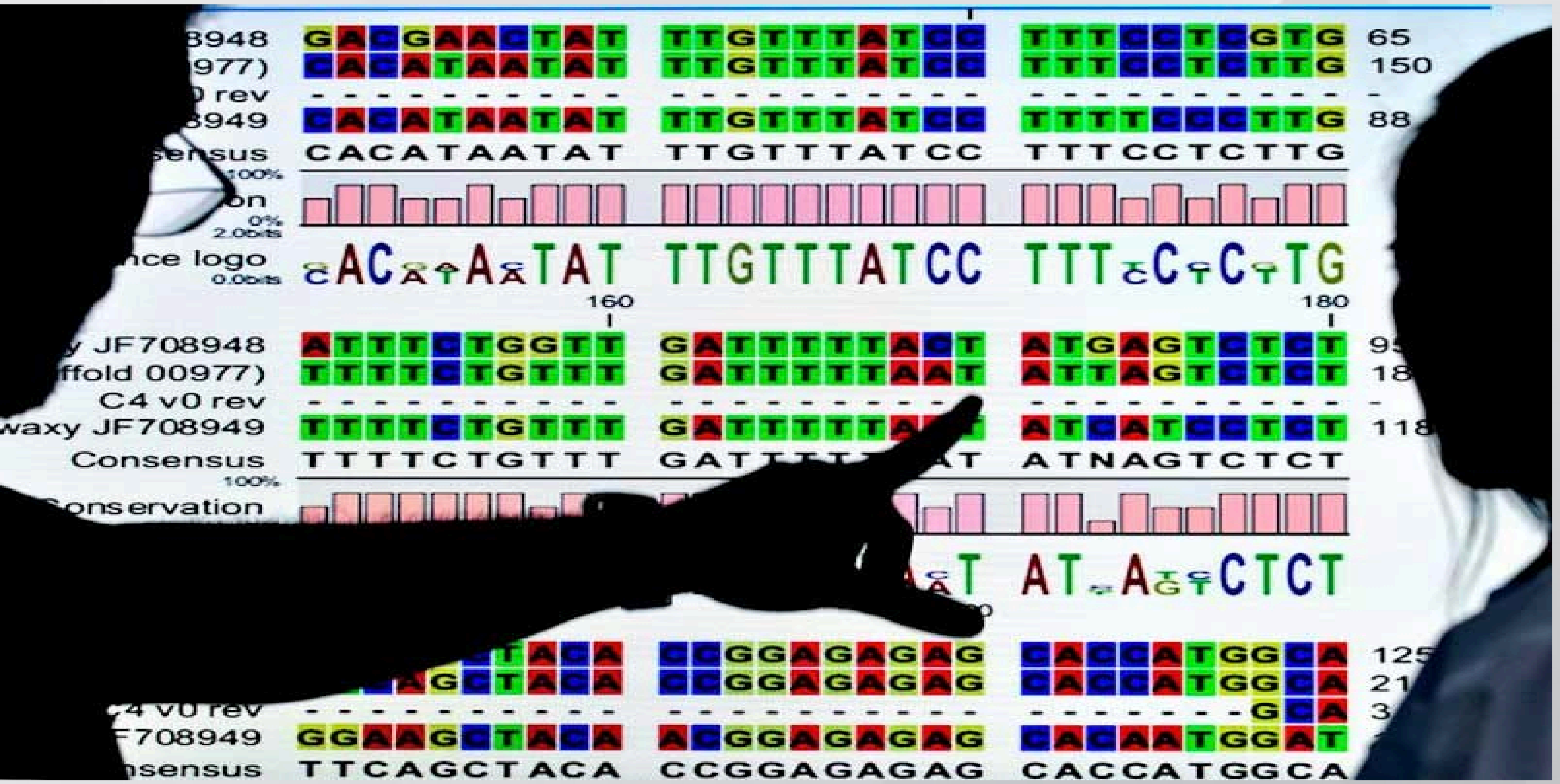




http://cs.stanford.edu/people/karpathy/deepimagesent/

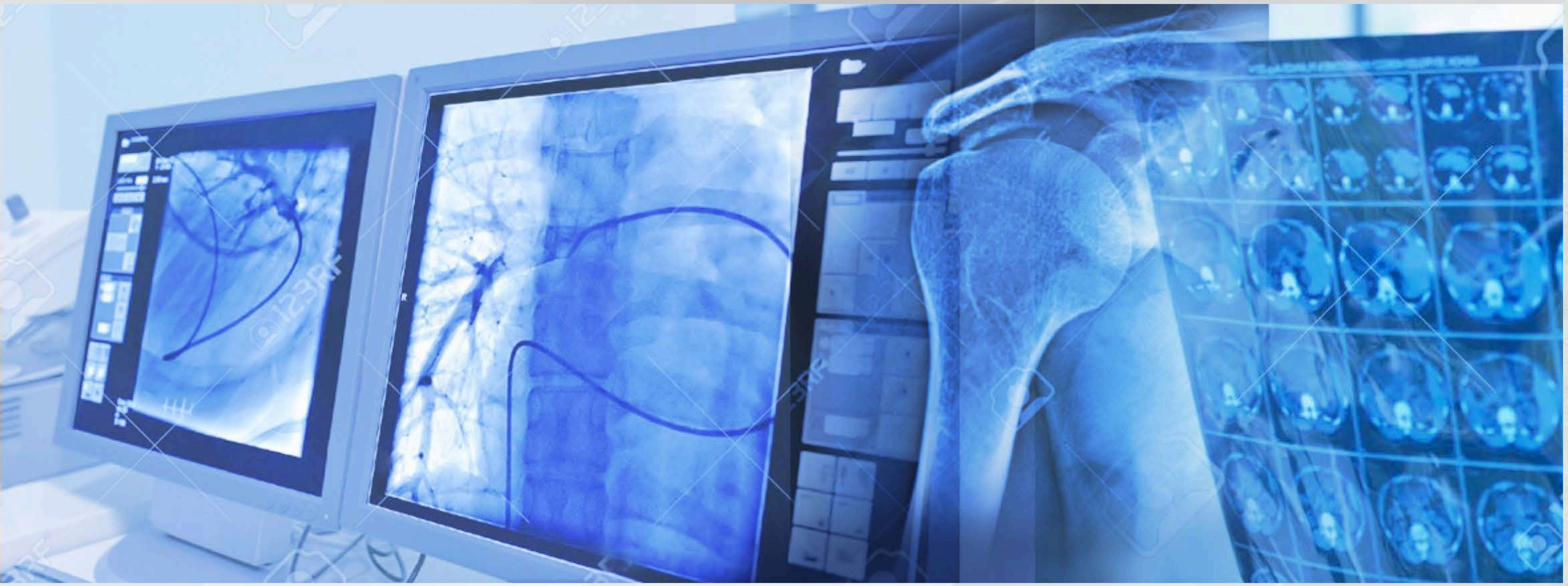
Interprétation de données « omics »





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).

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 de plus en plus, 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. ■

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

■ Un concours sur les pesticides

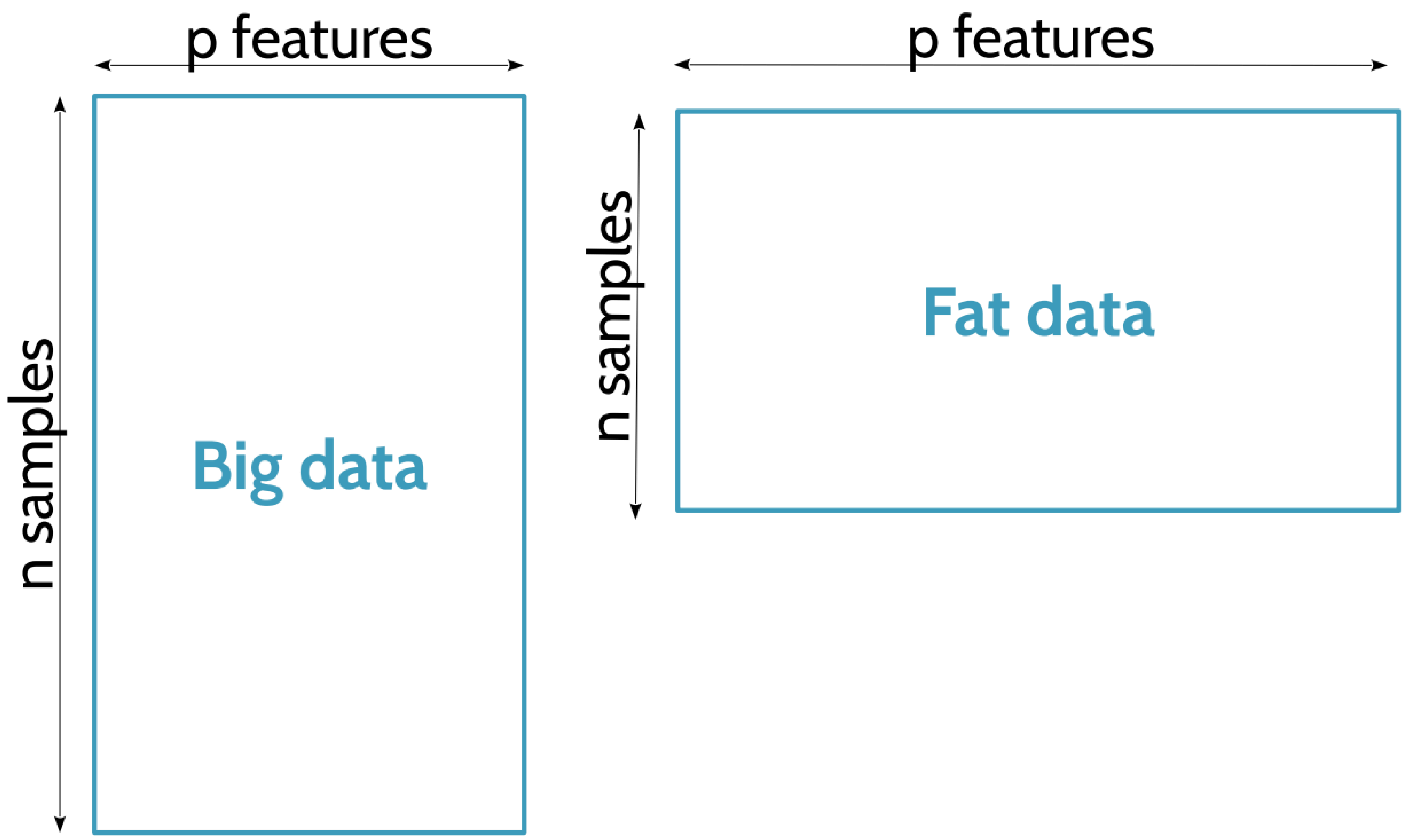
L'incubateur orléanais a lancé, hier, un concours national. Il s'agit d'imaginer de nouvelles solutions pour mieux visualiser les données concernant les pesticides dans les eaux souterraines. Les concurrents ont jusqu'au 16 janvier pour s'inscrire. Ils pourront s'appuyer sur les mesures de 2.000 points 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 !

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

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 problème

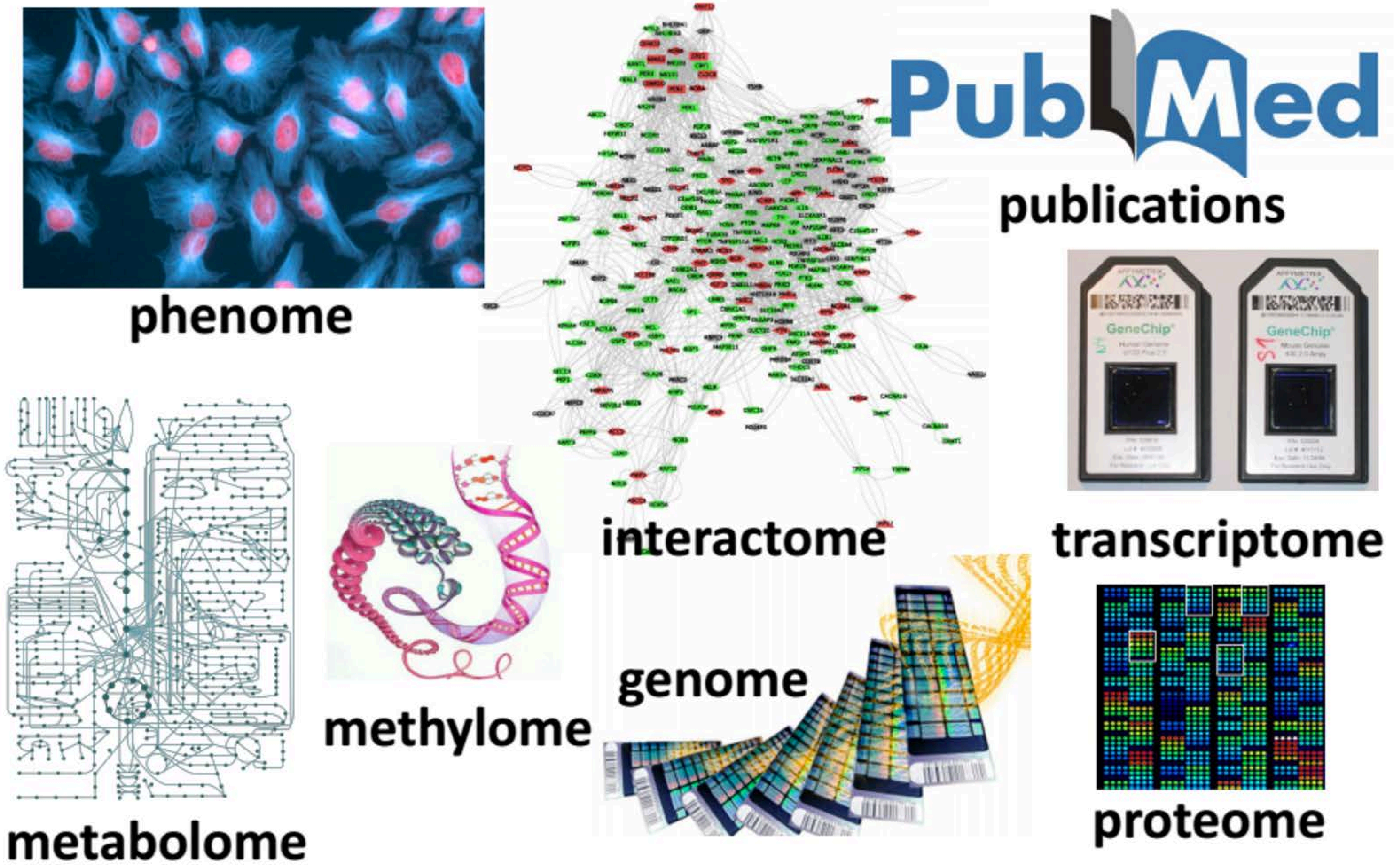


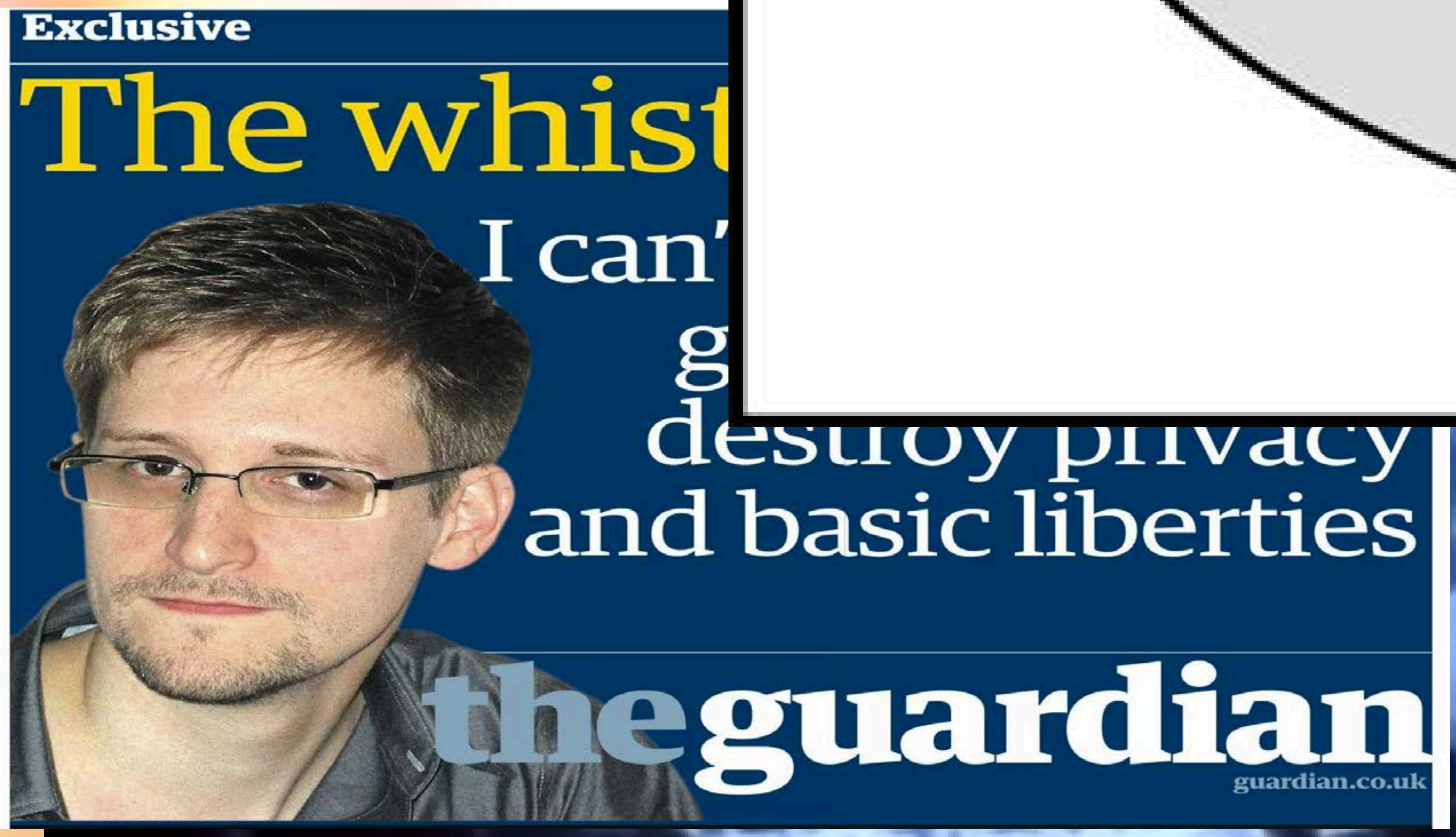
Image sources: ajc1@ flickr; Zlir'a@wikimedia

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

Doit-on a … a » ?

Non si:

« BIG B…



En fait, 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." »

En fait, on ne peut espérer une confidentialité parfaite

- Mais 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 renonce 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!!!

Alors on fait quoi avec nos données « sensibles » du monde de la santé ?

- 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
  - 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


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