Sciences des données à l'UBS

Etat des lieux et projets potentiels

Nicolas Courty
Professeur / Université de Bretagne Sud

Data Science et IA

Science des données (SD)

- Il semble utile de commencer par quelques définitions

- En termes très généraux, la science des données vise à l'extraction de connaissance d'ensembles de données

  - A l’interface de plusieurs domaines :

    - Mathématiques (statistique principalement),

    - la théorie de l'information et la technologie de l'information, notamment le traitement de signal, des modèles probabilistes, l'apprentissage automatique, l'apprentissage statistique,

    - la programmation informatique, l'ingénierie de données, la modélisation d'incertitude, le stockage de données, la compression de données et le calcul à haute performance

Principaux secteurs d’activités

- Déluge de données disponibles et à traiter/exploiter

Automobile

Transport, tourisme

Energie

Industrie

Médias

Finance

Secteur public

Santé

Pharmacie

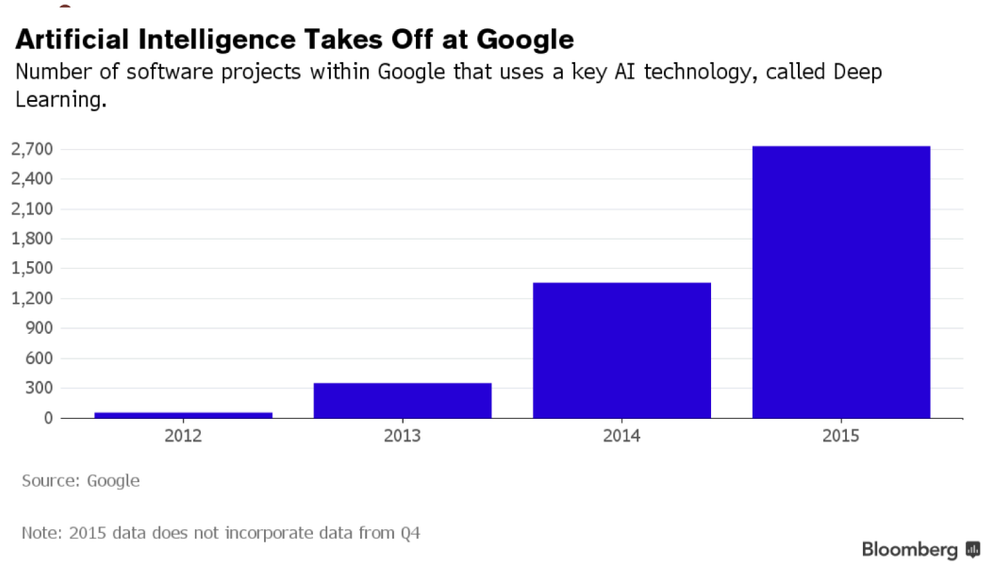
Agriculture

Biens de consommation

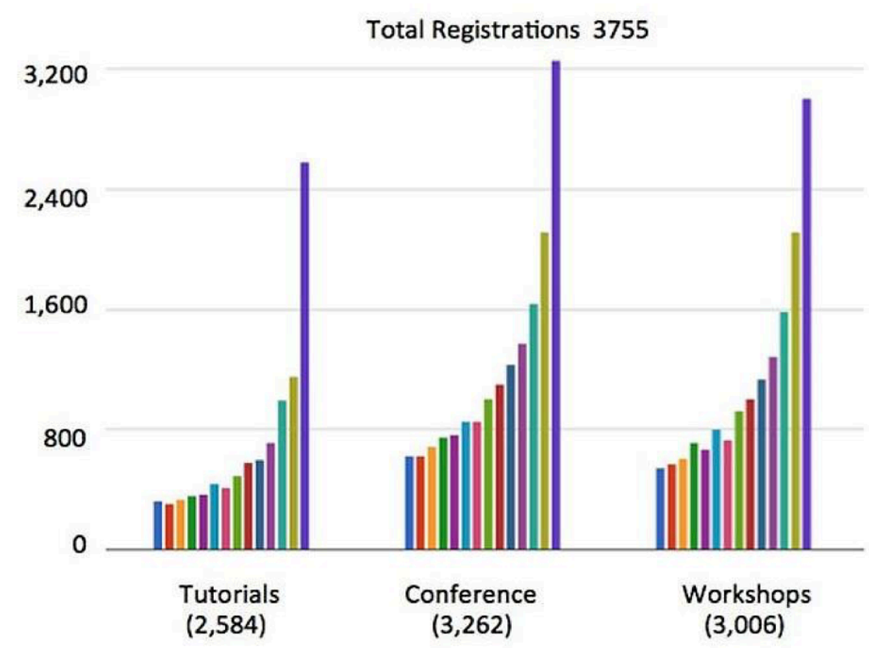
Télécommunications

Discipline en plein essor

- Explosion du volume de données disponibles (ère du numérique)
- Intérêt industriel fort porté par les GAFA (géants du web)
- 2016 : France AI, 2018 : mission Villani, 2019 : plan européen pour l'IA, plusieurs milliards d'euros pour rattraper le 'retard' sur les US et la Chine



Place de l'IA dans les produits Google (->2015)



Inscription à la conférence
majeure du domaine (jusqu'en 2015)
NIPS (Neural Information Processing system)

Qu’est ce que l’Intelligence Artificielle (IA) ?

- **Pas une définition unique !**

- **L’approche cognitive** : l’IA est la réalisation des programmes imitant dans leur fonctionnement l’esprit humain. Exemple : la compréhension du langage naturel

- **L’approche pragmatiste** : une IA est une boite noire intelligente qui réussit un certain nombre de tests. Exemple : Test de Turing

- **L’approche connexioniste** : une IA doit être capable d’apprendre à partir d’exemples et de généraliser (empiriquement) cet apprentissage

En général : l’IA vise à faire exécuter par l’ordinateur des tâches pour lesquelles l’humain est actuellement meilleur que la machine.

IA : quelques définitions utiles, liens avec le TLN

Historique

- **Formellement, la discipline démarre dans les années 1950 avec les travaux d'Alan Turing**

- **1950-1970 : balbutiements et premiers échecs.**

  - Des enseignements cependant: importance des connaissances non exprimées, étude de la représentation des connaissances, impossibilité de représenter 'toutes' les connaissances, nécessité de les rédiger sous forme générique

- **1970 - 1980 : Spécialisation**

  - compréhension du langage naturel, démonstration automatique de théorèmes, représentation des connaissances, programmation des jeux, résolution de problèmes

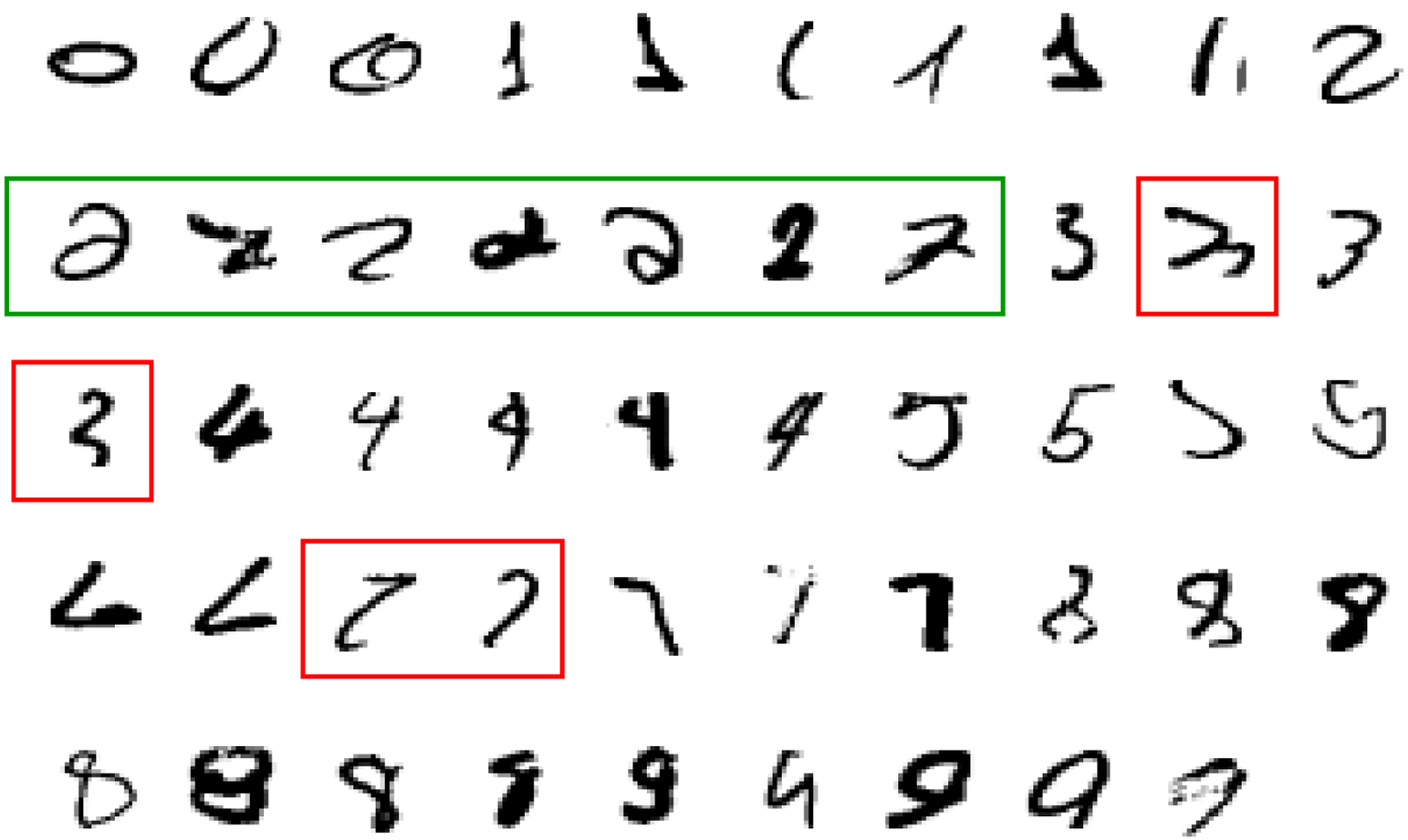
Historique

- **1980 - 2005 : hiver de l'IA, peu de succès sur des tâches en apparence simple**

Ce que vous voyez	Ce que la machine voit

Historique

- **1980 - 2005 : hiver de l'IA, peu de succès sur des tâches en apparence simple**
- Qu'est ce qu'un 2 ?



Historique

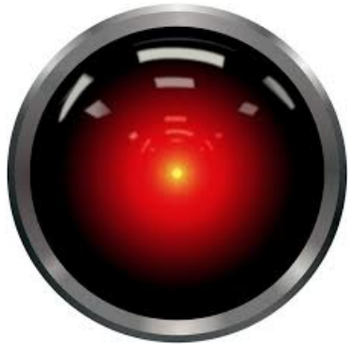
- **2005 - maintenant : l'essor**
  - **2 facteurs sont la source de ces progrès :**

    Données + capacité de calcul

  - **On dénote par machine learning ou apprentissage automatique toutes les méthodes qui permettent d'exploiter des données au travers d'algorithmes**
- **A partir de 2010 : avénement du deep learning**
  - **2016 : France AI, 2018 : mission Villani, 2019 : plan européen pour l'IA, plusieurs milliards d'euros pour rattraper le 'retard' sur les US et la Chine**
- **Maintenant - ? : la bulle ?**

Intelligence Artificielle Généralisée

- **Intelligence artificielle généralisée (IAG) désigne une machine capable d'appréhender n'importe quel type de problème avec un niveau d'intelligence équivalent ou supérieur à un humain**
  - On n'y est clairement pas encore !



- **Nous sommes plutôt sur des schémas d'IA sélective, c'est à dire capable de bonnes performances sur des tâches précises**

Quelles tâches ?

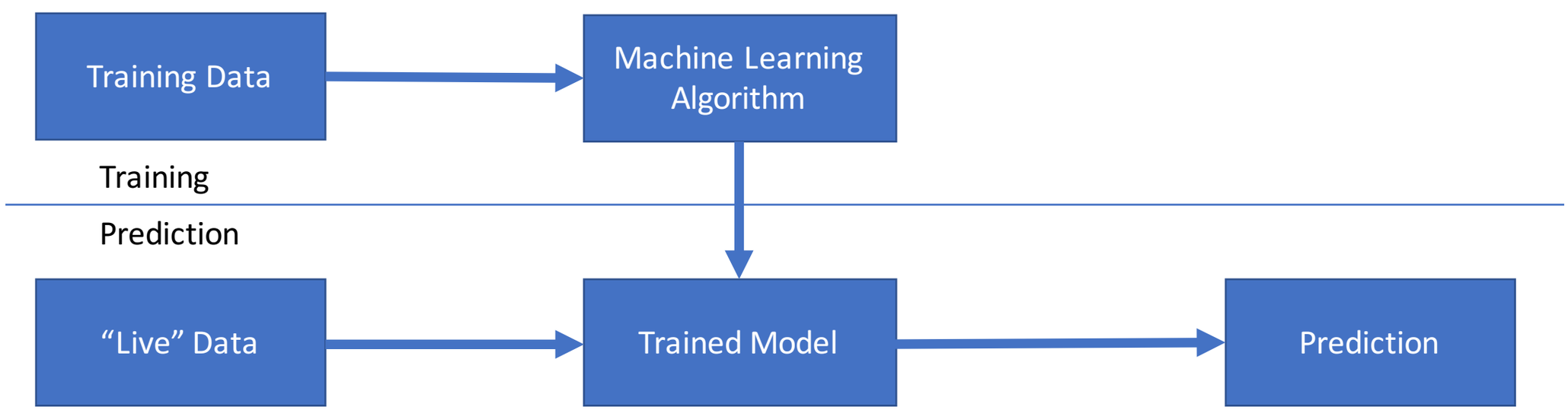
- **Vision par ordinateur (ex : reconnaissance/détection d'objets)**
- **Traitement du langage naturel (ex : traduction)**
- **Traitement des données audio (analyse/synthèse)**
- **Modèles prédictifs (météo, modèle financier, bourse)**
- **Analyse de données clients**
- **Systèmes de recommandation**
- **Robotique (contrôle)**
- **Etc.**

Machine et deep learning

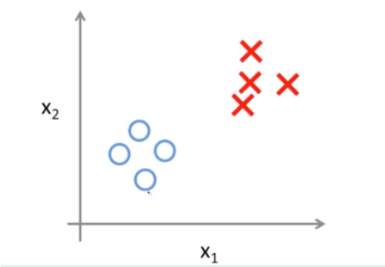
Machine learning

- **Apprendre sans avoir ‘expliciter’ les mécanismes que l’on cherche à reproduire**

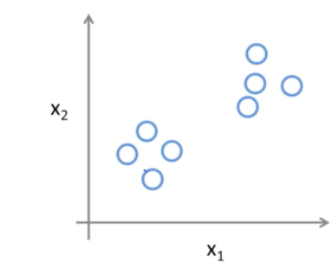
- **Plus concrètement, plusieurs catégories de technique sont utilisés pour prédire de l’information**



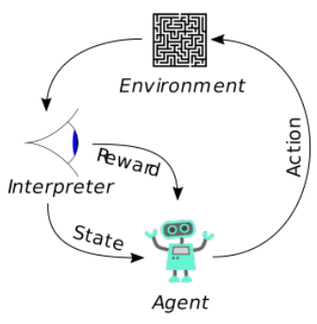
Les grands problèmes du machine learning



- **Apprentissage supervisé :** à partir d'un jeu de données labellisées

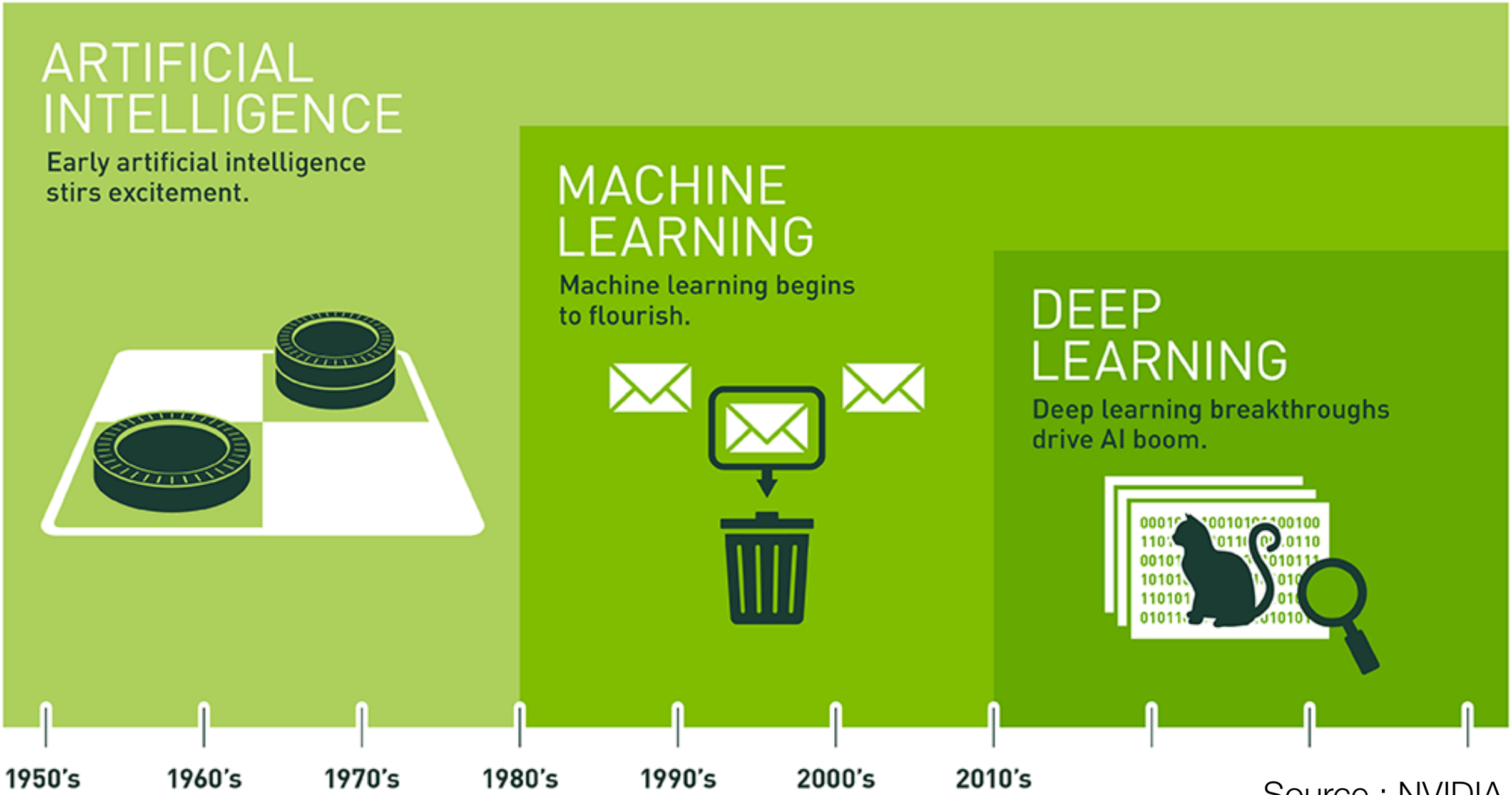


- **Apprentissage non-supervisé :** uniquement des données brutes



- **Apprentissage par renforcement :** mécanisme de feedback

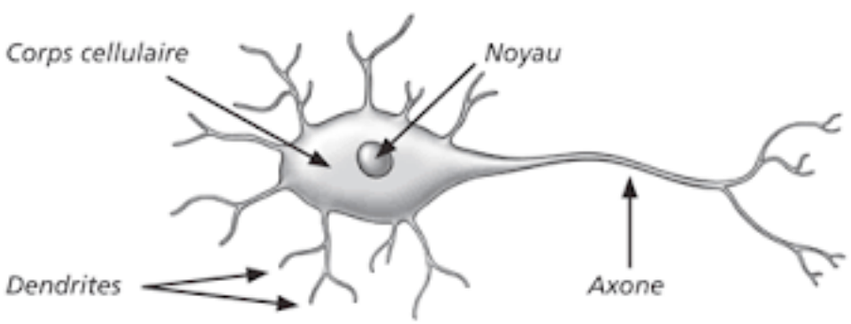
IA / Machine learning / Deep learning

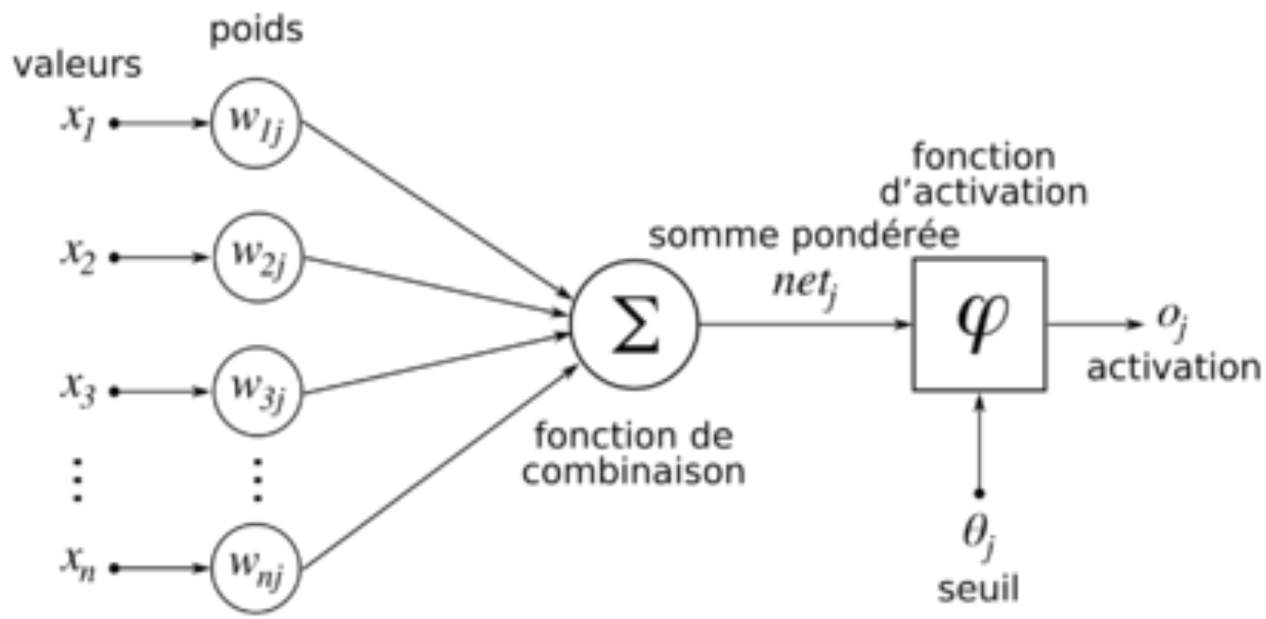


Source : NVIDIA

Deep learning

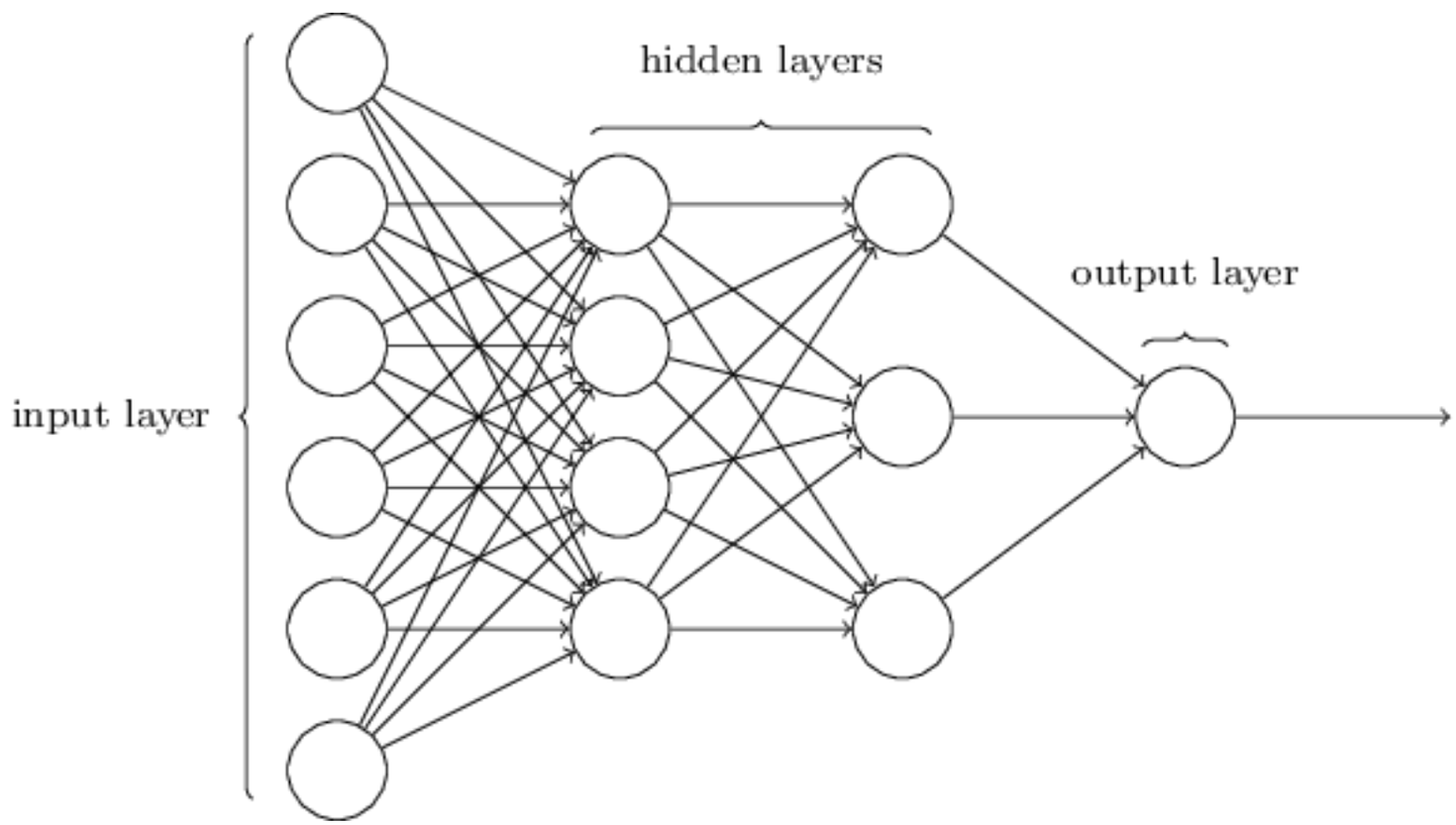
- **L'apprentissage profond ou 'deep learning' regroupe une catégorie de méthodes d'apprentissage basées sur des réseaux de neurones**
  - le neurone étant un modèle mathématique d'un neurone biologique





Architecture complexe

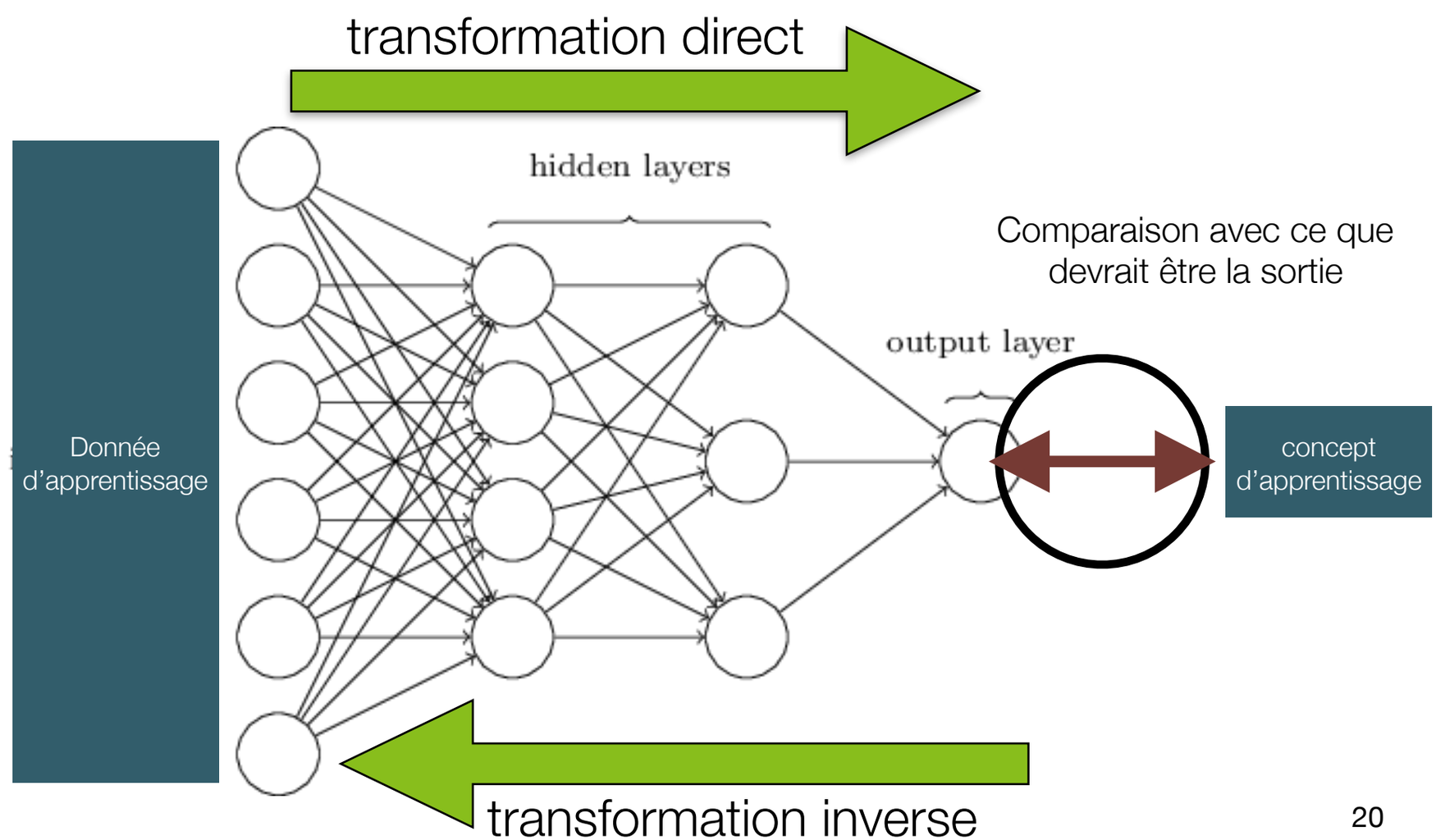
- **Combinaison de ces unités simples**



Donnée brute — — — — — — — — —> Concept

Apprentissage !

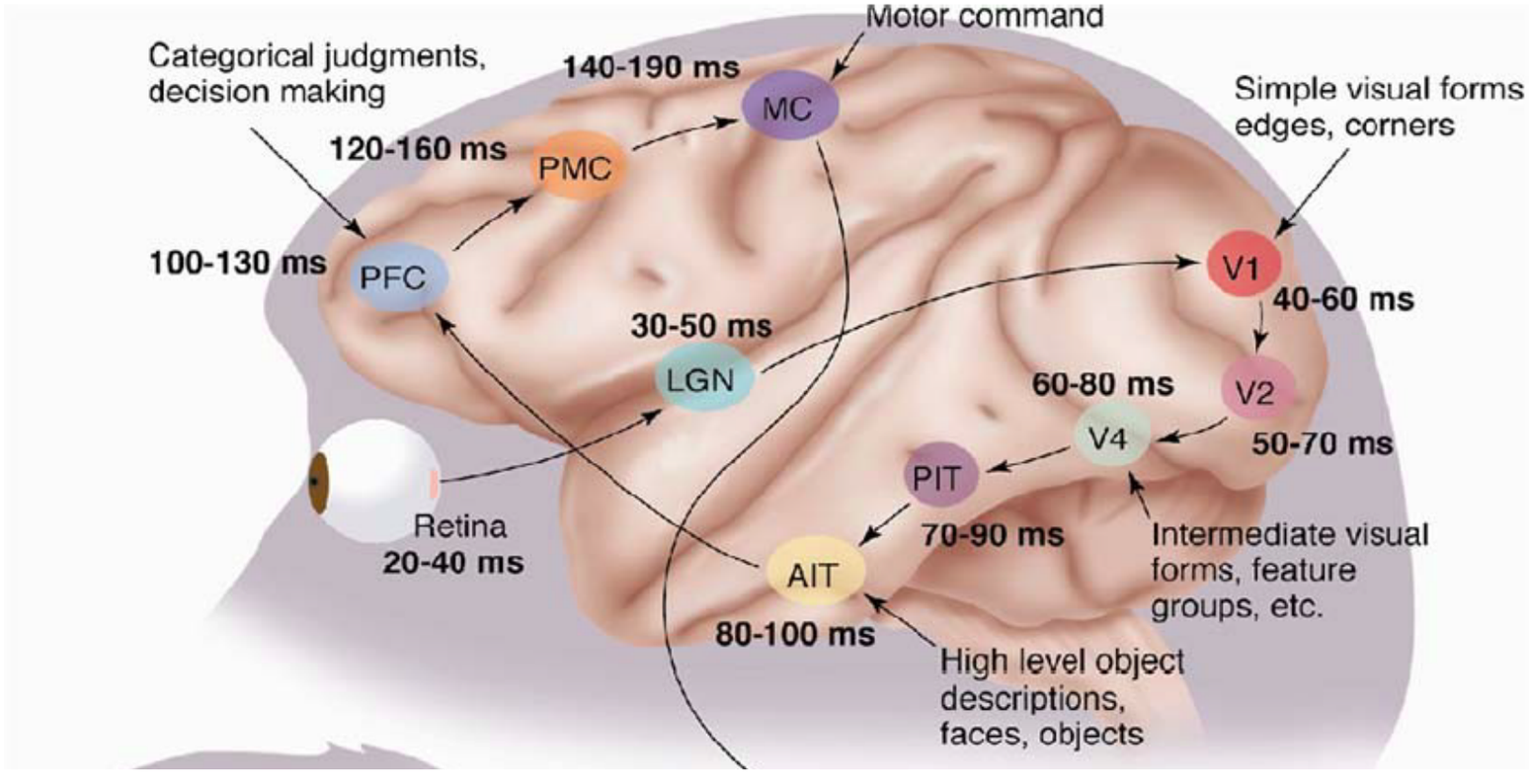
- **Les poids associées aux combinaisons sont calculés automatiquement à partir d'une base d'apprentissage (données/ concepts)**
    - *apprentissage dit supervisé*
- Rétropropagation !



Bio-mimétisme

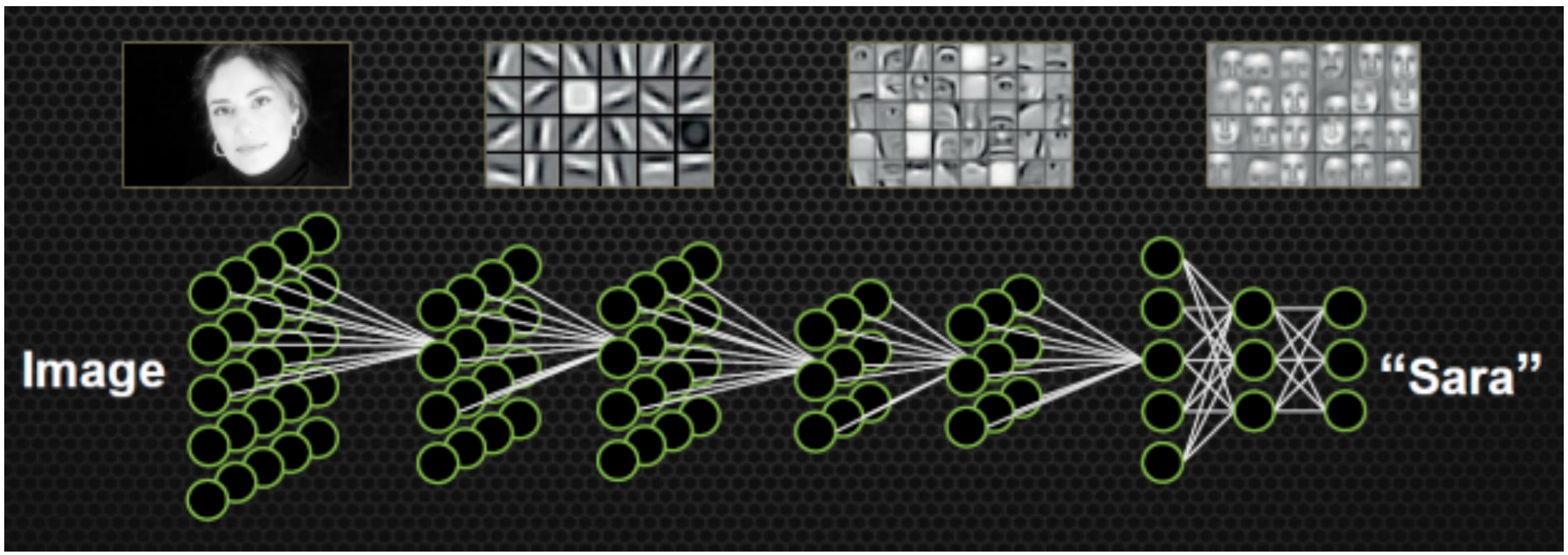
- **Outre le lien entre le neurone mathématique et le neurone biologique, un réseau de neurones imite (dans une certaine mesure) le fonctionnement du cerveau**

    - *exemple de la hiérarchie dans la vision humaine*



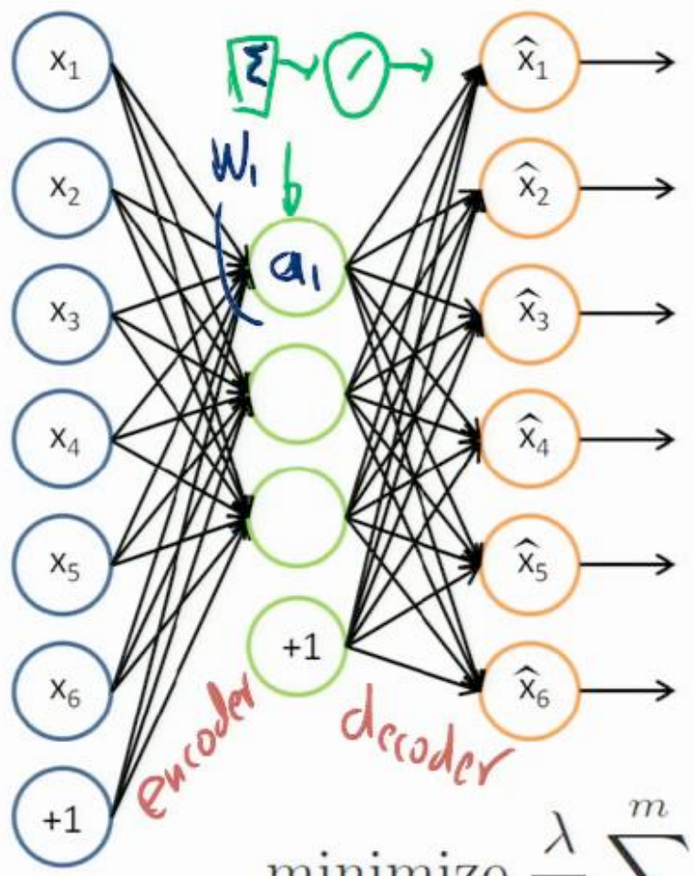
Réseau ‘convolutionnel’

- **Passages des primitives aux informations conceptuelles de haut niveau**



Modélisation du problème mathématique

On ne rentrera pas dans les détails aujourd'hui…



$$\overset{1\times 1}{a_1} = \underline{W}_1 \underline{x}^{(i)} \qquad \underline{a} W_1^T = \hat{x}$$

$$\underline{W}_1 \in \mathbb{R}^{1\times 7} \qquad \underline{x}^{(i)} \in \mathbb{R}^{7\times 1}$$

$$\underset{W}{\text{minimize}} \; \frac{\lambda}{m} \sum_{i=1}^{m} \underbrace{\|W^T \underbrace{[W x^{(i)}]}_{a^{(i)}} - x^{(i)}\|_2^2}_{\hat{x}^{(i)}} + \sum_{i=1}^{m} \sum_{j=1}^{k} g(W_j x^{(i)})$$

$$\underset{W,b,c}{\text{minimize}} \; \frac{\lambda}{m} \sum_{i=1}^{m} \|\underbrace{\sigma(W^T \sigma(W x^{(i)} + \underline{b}) + \underline{c})}_{\hat{x}^{(i)}} - \underline{x}^{(i)}\|_2^2 + \underbrace{\sum_{i=1}^{m} \sum_{j=1}^{k} g(W_j x^{(i)})}_{\text{regularis}}$$

e.g., smooth L_1 penalty: $g(.) := \log(\cosh(.))$

Ce qui a changé lors des dix dernières années

- **profondeur des architectures**
  - jusqu'à plusieurs centaines pour les derniers modèles
- **Algorithmes / théorie de la rétro-propagation**
- **distribution des calculs : utilisation de CPUs + GPUs (parallélisme)**

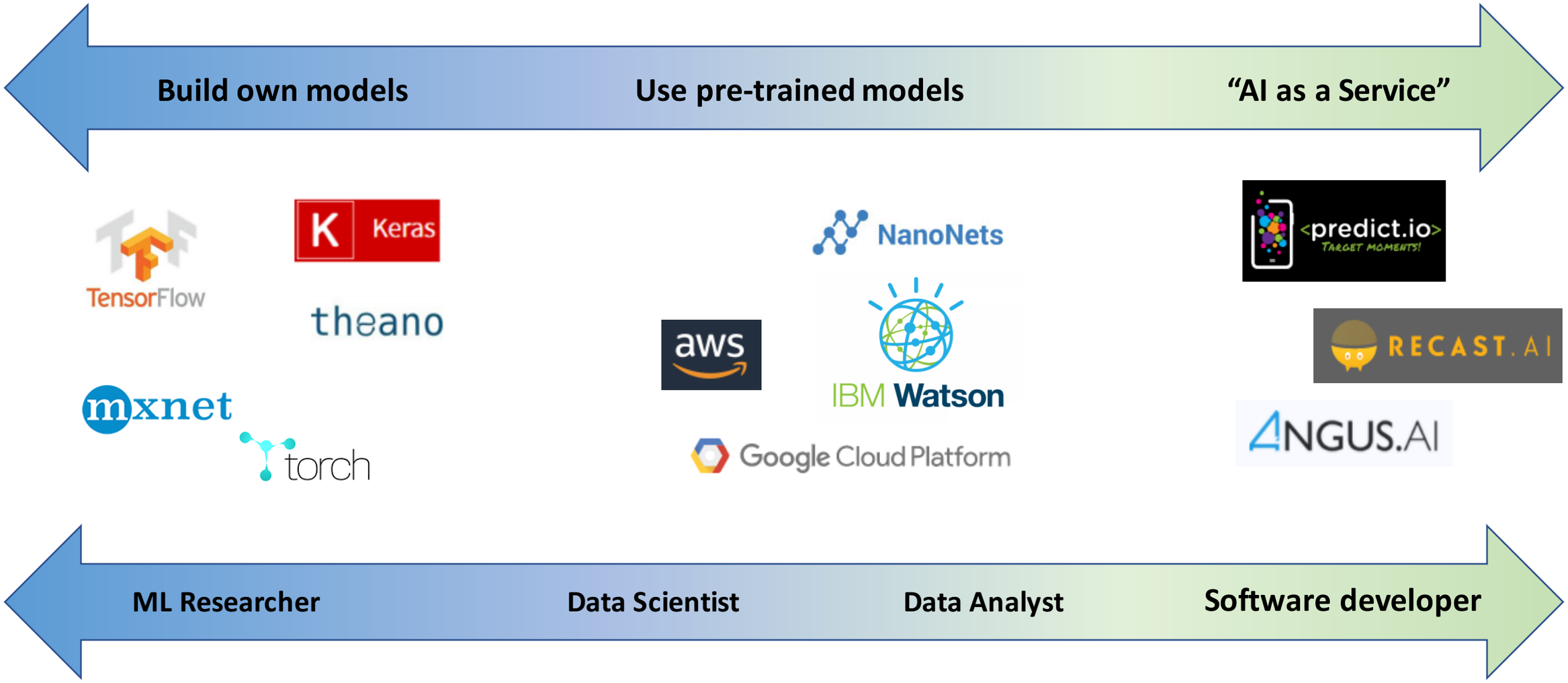
Google brain

Nvidia Tesla

IA opérationnelle en entreprise

Utilisation de l'IA dans un contexte opérationnel

- **Plusieurs approches / vues**



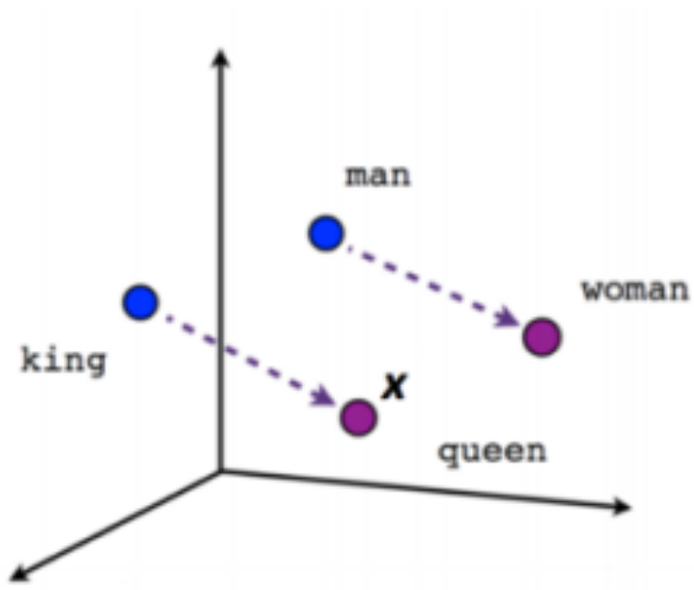
Clés de la réussite pour un projet basé IA

- **Des besoins clairement identifiés (quel est le problème que l'on cherche à résoudre ?)**
- **Une définition précise des produits/services associés avec des objectifs évaluables**
- **Avoir des données (beaucoup !) à disposition**
- **Etre capable de construire ou de trouver le bon modèle associé à ce problème**
- **Ajustements et tests dynamiques du produit - faire évoluer les performances tout au long de sa durée de vie**

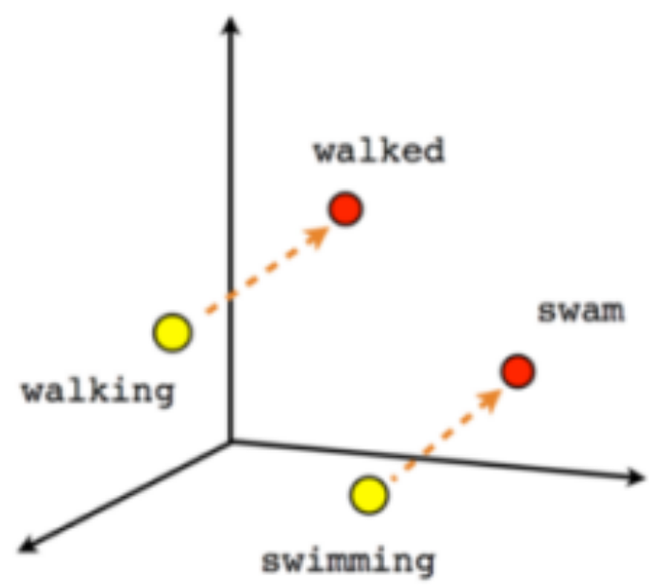
Traitement de la langue naturelle et deep learning
Une alliance naturelle ?

Plongements

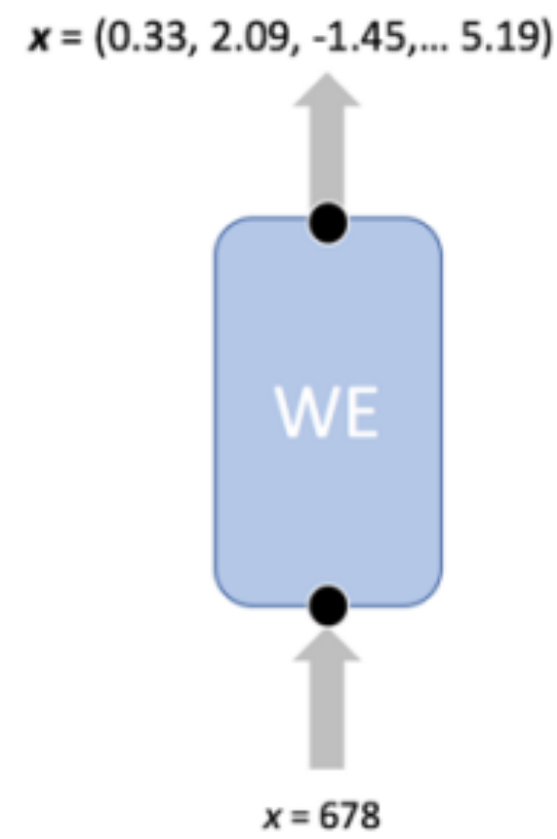
- Liens texte / numérique
- Vectoriser les mots



Male-Female



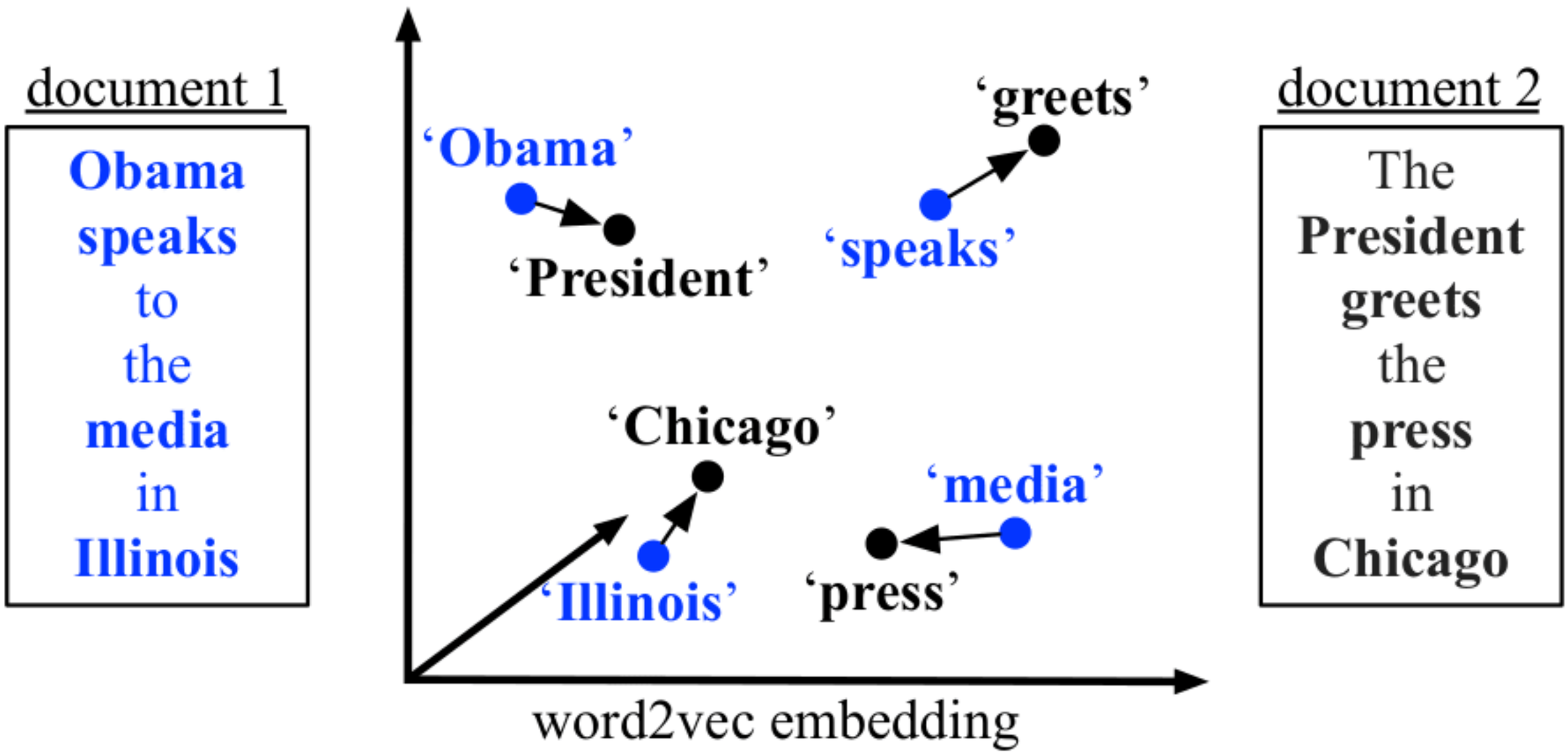
Verb tense



queen

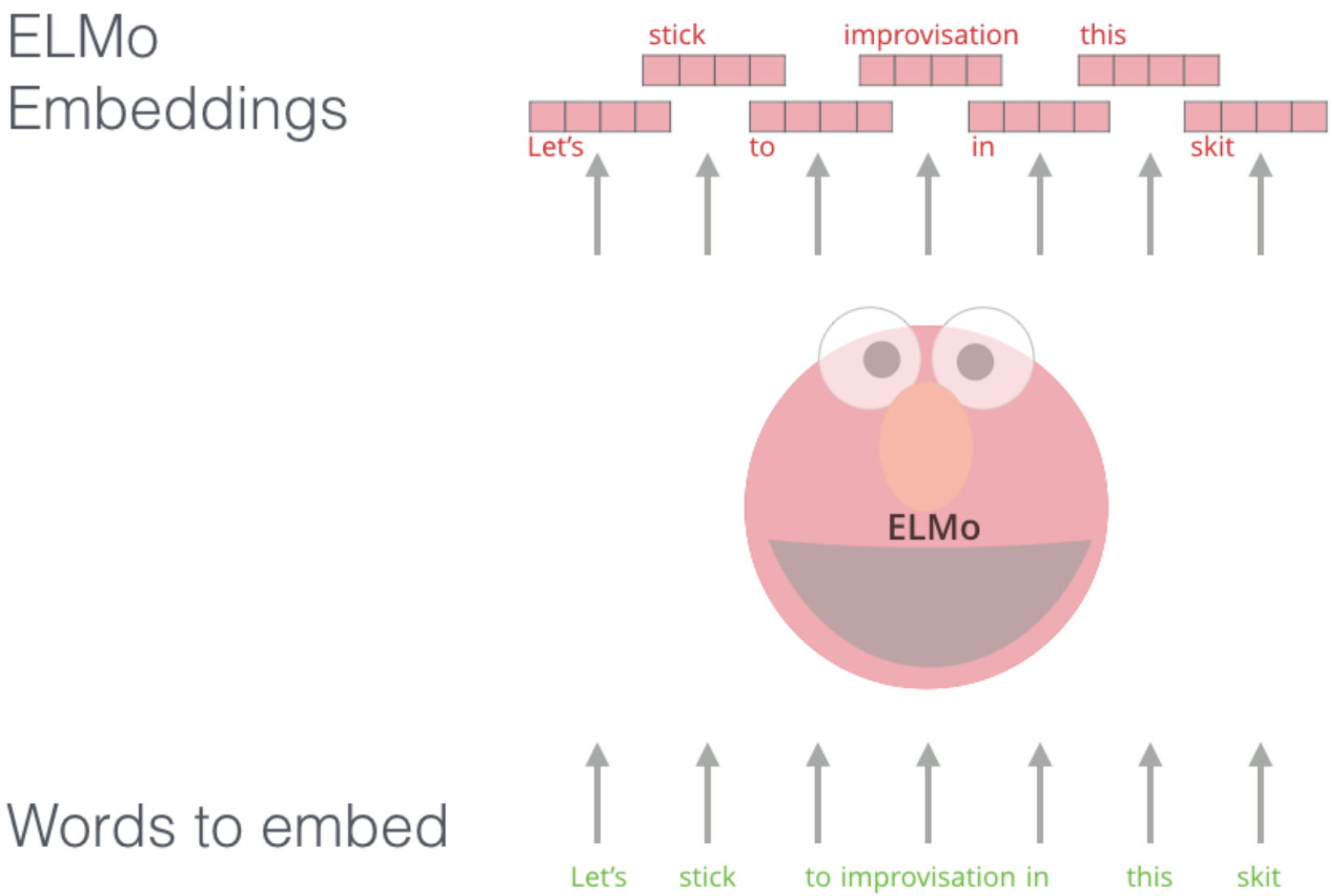
Plongements

- Similarité entre groupes de mots



Une révolution : la contextualisation

Une révolution : la contextualisation

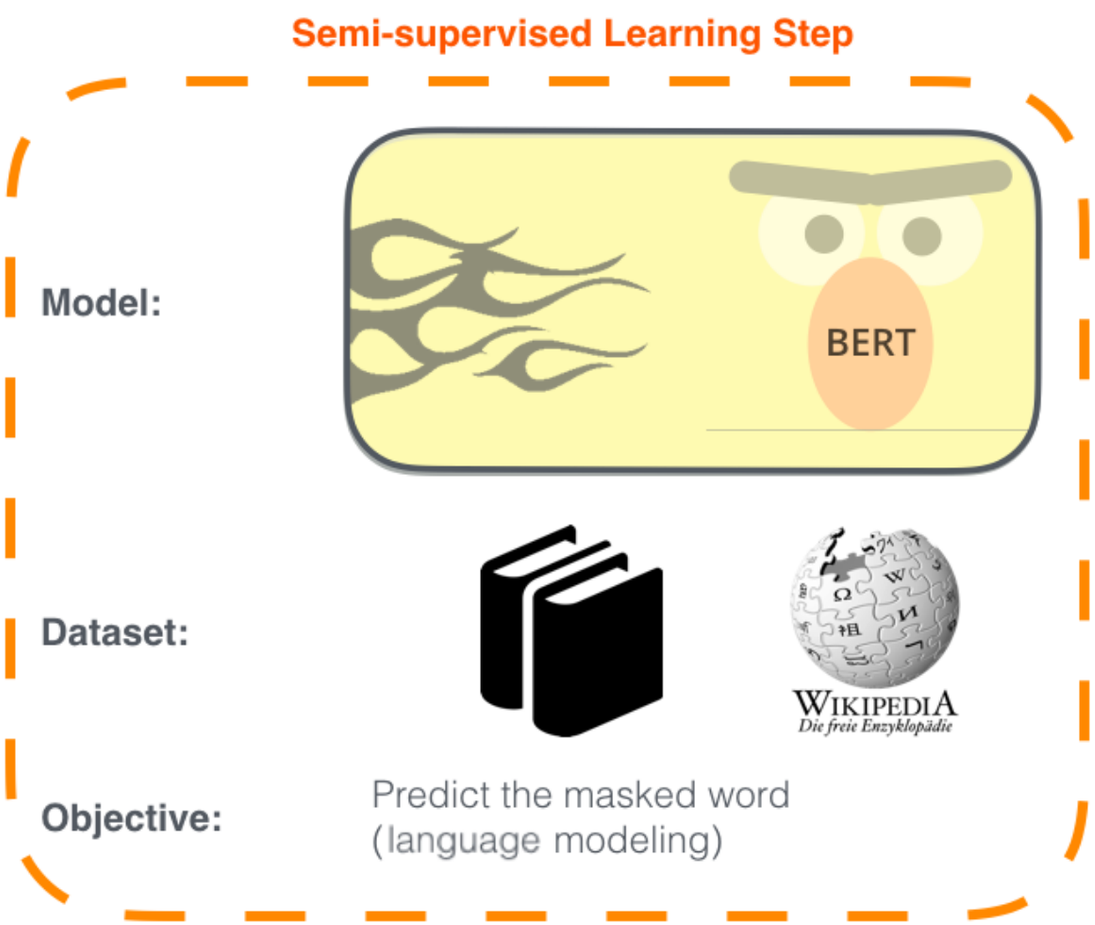


Avec des réseaux de neurones récurrents

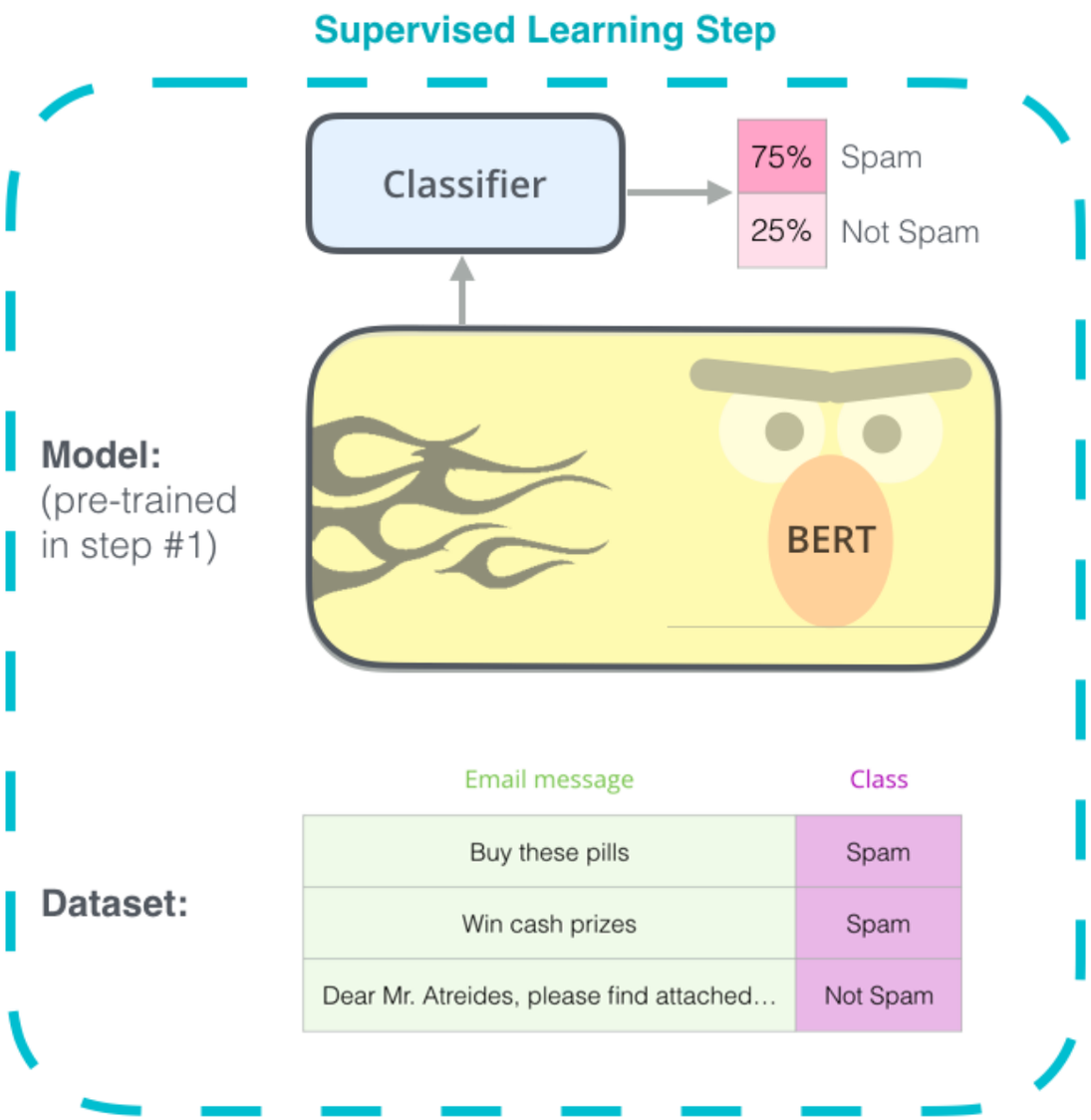
ELMo, BERT et consort : apprentissage par transfert

1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.

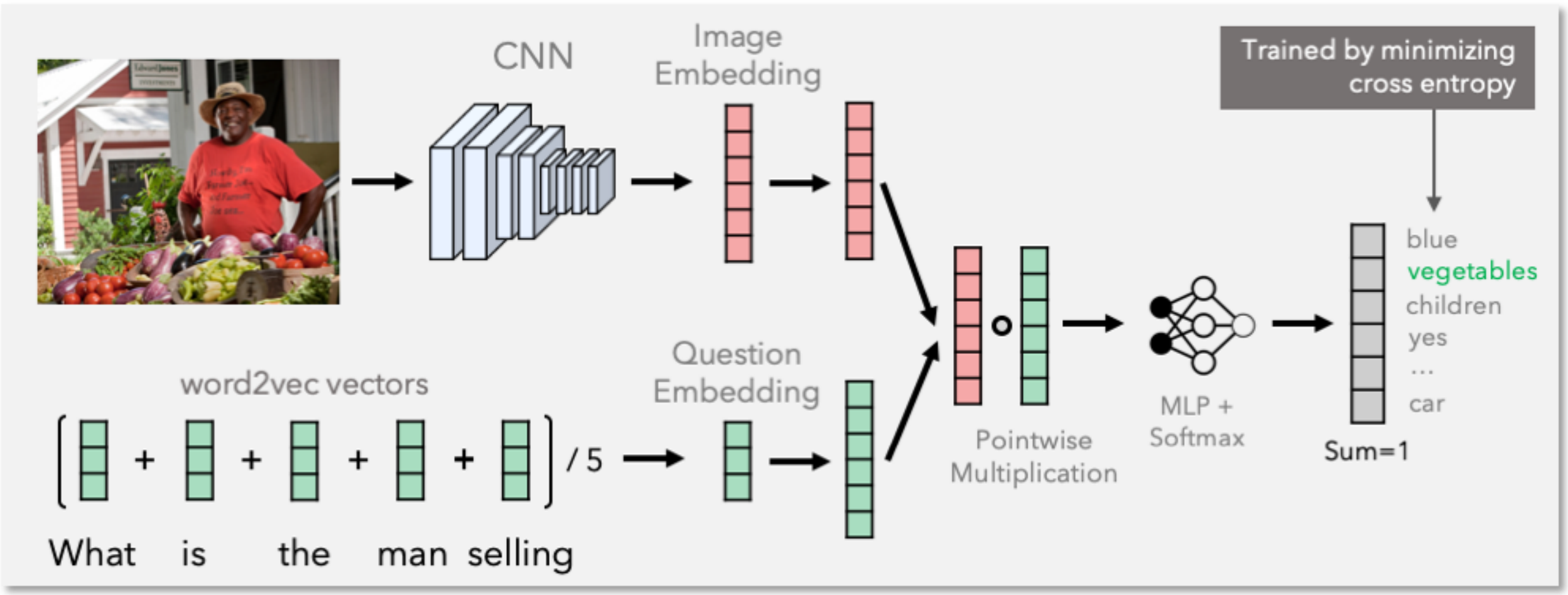


2 - Supervised training on a specific task with a labeled dataset.



Source : http://jalammar.github.io/illustrated-bert/

Couplage textes / images



Interprétabilité et IA

- **Beaucoup de méthodes (surtout le deep learning) fonctionnent comme des boites noires à l'heure actuelle**
  - Il est difficile d'interpréter le lien fait entre l'entrée et la sortie du modèle
- **Cela peut poser un certain nombre de problèmes (surtout lorsque l'humain est dans la boucle)**
- **Le traité de régulation des données de l'UE est assez ambigu sur ce point (GDPR : EU General Data Protection Regulation)**
  - Tout un chacun doit pouvoir demander des explications sur des décisions qui lui sont associées

Quelques (exemples de) limitations

Equité / Biais en IA

- **Certaines des tâches automatisées par l'IA ont un impact sociétal faible (systèmes de recommendation, réseaux sociaux, etc.) mais des algorithmes spécialisés peuvent au contraire avoir un impact fort (médecine, justice, finance)**

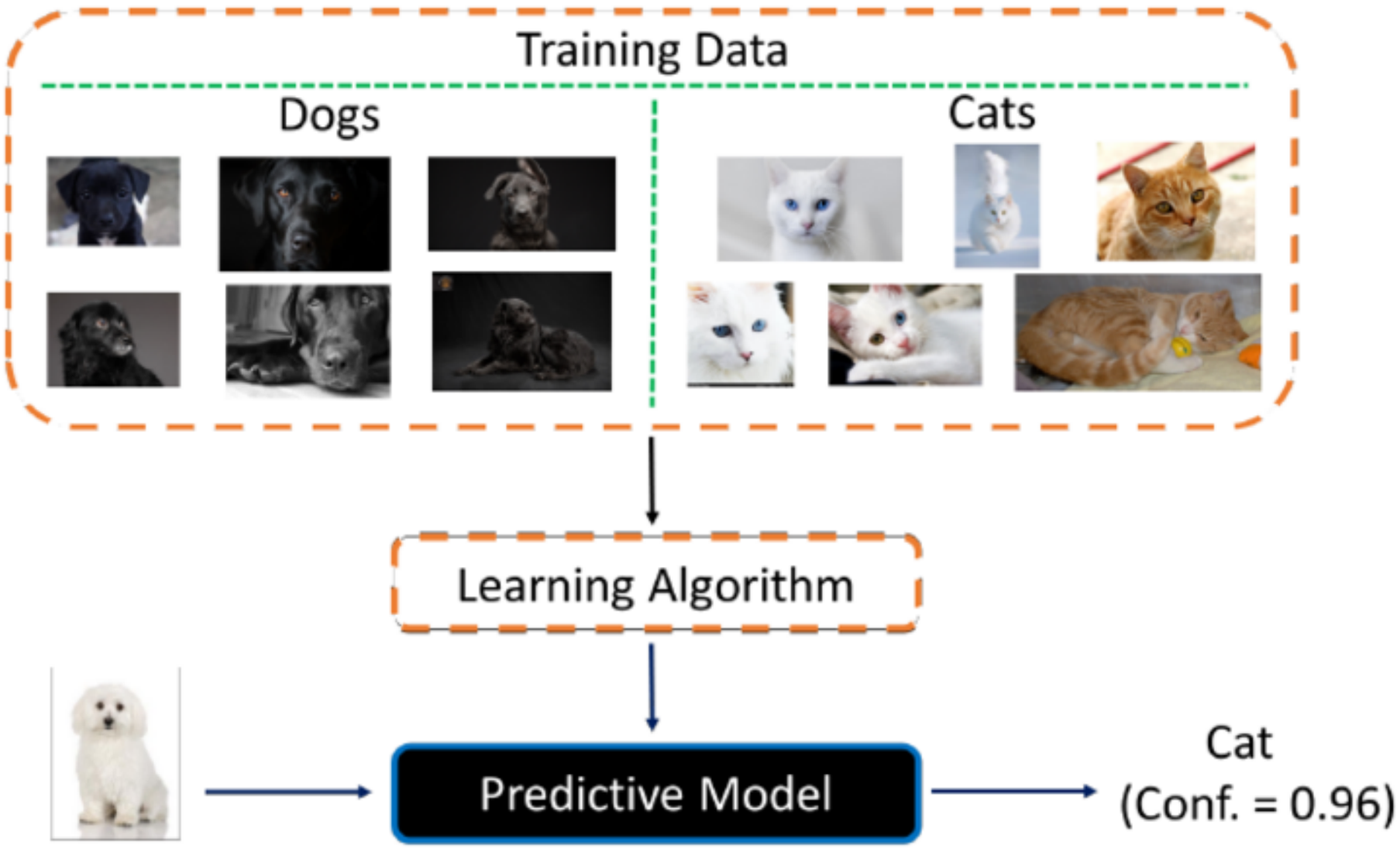
- **Les conséquences de l'utilisation d'IA dans ces domaines peuvent être très importantes**

- **Cependant il est observé que certaines méthodes renforce des biais (genre, raciaux, etc.)**

  - Pourquoi ? Les bases d'apprentissage ne sont pas équilibrés

  - Exemple en 2015 de Google image qui avait taggé 2 photos de personnes noires avec le label 'gorille'

Exemple de biais



RH et IA

- https://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html

The Upshot

FOLLOW US:
GET THE UPSHOT IN YOUR INBOX

ROBO RECRUITING

Can an Algorithm Hire Better Than a Human?

Claire Cain Miller @clairecm JUNE 25, 2015

117

Hiring and recruiting might seem like some of the least likely jobs to be automated. The whole process seems to need human skills that computers lack, like making conversation and reading social cues.

But people have biases and predilections. They make hiring decisions, often unconsciously, based on similarities that have nothing to do with the job requirements — like whether an applicant has a friend in common, went to the same school or likes the same sports.

That is one reason researchers say traditional job searches are broken. The question is how to make them better.

A new wave of start-ups — including Gild, Entelo, Textio, Doxa and GapJumpers — is trying various ways to automate hiring. They say that software can do the job more effectively and efficiently than people can. Many people are beginning to buy into the idea. Established headhunting firms like Korn Ferry are incorporating algorithms into their work, too.

If they succeed, they say, hiring could become faster and less expensive, and

RECENT COMMENTS

Mayurakshi Ghosh January 7, 2016
Hi Claire, excellent article and really insightful facts on algorithm recruitment. I completely agree how you mentioned the role played by...

Deborah Bishop July 2, 2015
This is a very interesting article. Perhaps distinguishing between different aspects in the process of bringing talent into your...

Yeti July 2, 2015
Including talents that would be rejected by the subjective biased boss or colleagues does not fuarantee their integration. They will be...

SEE ALL COMMENTS WRITE A COMMENT

Data Science à l’UBS

3 laboratoires identifiés sur cette thématique

IRISA
Vannes

Institut de Recherche en Informatique et Systèmes aléatoires
UMR 6074
Basée sur les campus de Rennes (35), Vannes (56), Lannion (22) et Brest (29).
Tutelles: CNRS, ENS Rennes, Inria, INSA Rennes, UBS, Université de Rennes 1, IMT Atlantique et CentraleSupélec.

LMBA
Vannes

Laboratoire de Mathématiques de Bretagne Atlantique
UMR 6205
Basée sur les campus de Vannes (56), Lorient (56) et Brest (29).
Tutelles: CNRS, Université de Bretagne Sud, UBO

LabSticc

Laboratoire des sciences et techniques de l'information, de la communication et de la connaissance
UMR 6285
Basée sur les campus de Vannes (56), Lorient (56), Quimper (29) et Brest (29).
Tutelles: CNRS, IMT Atlantique, UBO, UBS, ENSTA Bretagne et ENIB

Contributions de l'IRISA Vannes à la SD

IRISA
Vannes

- 2 équipes majoritairement impliquées : EXPRESSION et Obélix
- Equipe **Expression** (basée sur Vannes et Lannion) https://www.irisa.fr/fr/equipes/expression :
  - Responsable : Pierre-François Marteau
  - Effectifs hors membre associés :
    - Vannes : 3 PRU, 4 MCF, Lannion : 5 MCF
    - Non permanents : 1 Post-doc, 1 IR, 10 Doctorants
  - Thème de recherche : exploitation de données liées au langage (discours, gestes, expressions faciales, textes, log, etc.)
    - Données complexes, potentiellement symboliques, numériques, structurées

Contributions de l'IRISA Vannes à la SD

IRISA
Vannes

- 2 équipes majoritairement impliquées : EXPRESSION et Obélix

- Equipe **Expression** (basée sur Vannes et Lannion) :

  - Projet de recherche : méthodologies, formalisation et implémentation d'algorithmes de machine learning liés à ce type de média

  - Applications à la cybersécurité : analyse de comportements, détection d'anomalies dans les systèmes cyber physiques

  - Participation à des projets ANR, actions de transfert industriel, collaborations internationales

  - Communautés de publication : IEEE, ACM, appliquées aux média expressifs, ou fondements du machine learning

Contributions de l'IRISA Vannes à la SD

IRISA
Vannes

- 2 équipes majoritairement impliquées : EXPRESSION et Obélix
- Equipe **Obélix** (basée sur Vannes et Rennes) https://www-obelix.irisa.fr/ :
  - Responsable : Sébastien Lefèvre
  - Effectifs hors membre associés :
    - Vannes : 2 PRU, 4 MCF, Rennes : 1 DR CNRS, 1 MCF
    - Non permanents : 4 Post-doc, 10 Doctorants
  - Thème de recherche : Intelligence Artificielle pour l'Observation de la Terre (IA4EO)

Contributions de l'IRISA Vannes à la SD

IRISA
Vannes

- 2 équipes majoritairement impliquées : EXPRESSION et Obélix

- Equipe **Obélix** (basée sur Vannes et Rennes)
  https://www-obelix.irisa.fr/ :

  - Projet de Recherche : contributions méthodologiques au machine learning (principalement deep learning) et appliquées aux problèmes de télédétections

  - Participation à des projets ANR PRC, Astrid, … (ex: ANR OATMIL portée par UBS, sur l'application de la théorie du transport optimal au machine learning)

    - Liens avec la cybersécurité (ANR DGA, MOOC UBS sur les liens IA/cyber)

  - actions de transfert industriel (ex : CNES, Thalès), nombreuses collaborations internationales (ex : ETH Zurich, U. Wageningen)

  - Communautés de publication : IEEE, appliquées à la télédétection, ou fondements du machine learning (NIPS, ICML, ICLR : principales conférences en machine et deep learning)

Contributions du LMBA Vannes à la SD

LMBA
Vannes

- Les 3 équipes du LMBA sont impliquées :
  - Géométrie et topologie,
  - Systèmes dynamiques, probabilités et statistique,
  - Analyse, phénomènes stochastiques et applications
- Et contribuent au projet en SD du LMBA

Contributions du LMBA Vannes à la SD

LMBA
Vannes

- 3 axes transverses thématiques sont identifiés :
  - Analyse et fusion de différentes sources de données massives
    - Effectifs : 1 MCF UBS, 2 MCF et 2 PR hors UBS
  - Apprentissage statistique
    - Effectifs : 3 PR UBS, 3 MCF UBS, 1 MCF hors UBS
  - Traitement des images mathématique et deep learning
    - Effectifs : 3 PR UBS, 1 MCF UBS
- Participation à des projets ANR, nombreuses actions de transfert industriel

Contributions du LMBA Vannes à la SD

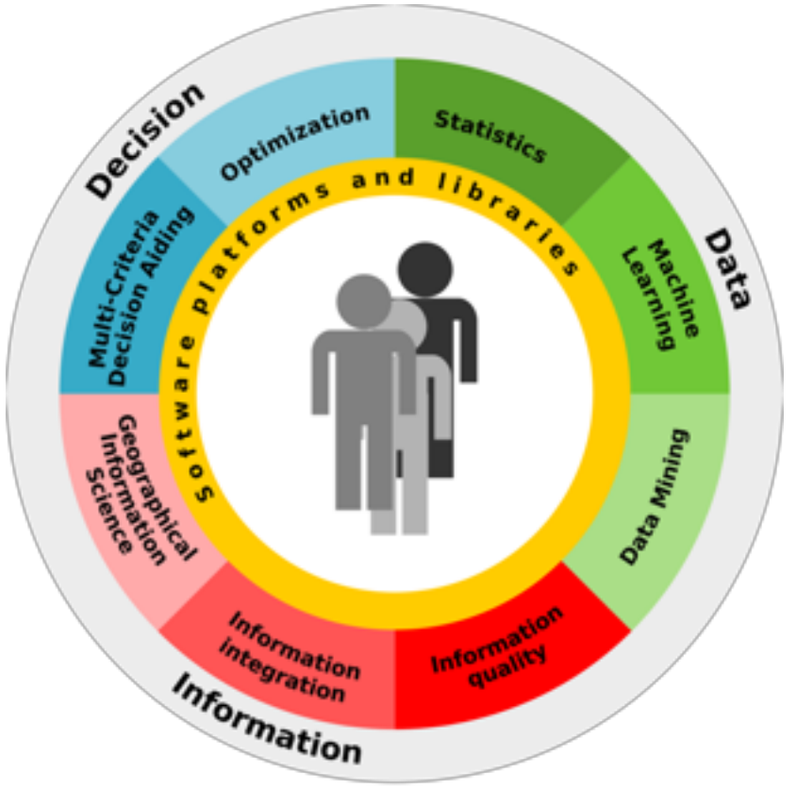
LMBA
Vannes

- Projet de recherche sur le couplage données-modèles
- 3 axes dans ce projet
  - Jumeaux numériques : modèles de prédiction basés sur les données
  - Ajustement : identification des modèles statistiques par les données
  - Traitement mathématique des images

Contributions du Lab-Sticc à la SD

LabSticc

- Au travers de l'équipe DECIDE, au travers des aspects suivants :
  - Recherche opérationnelle et optimisation de tournées multi-véhicules et de capteurs mobiles
  - Modélisation et optimisation de données spatio-temporelles
  - Aide multi-critère à la décision
  - Sciences de l'Information Géographique



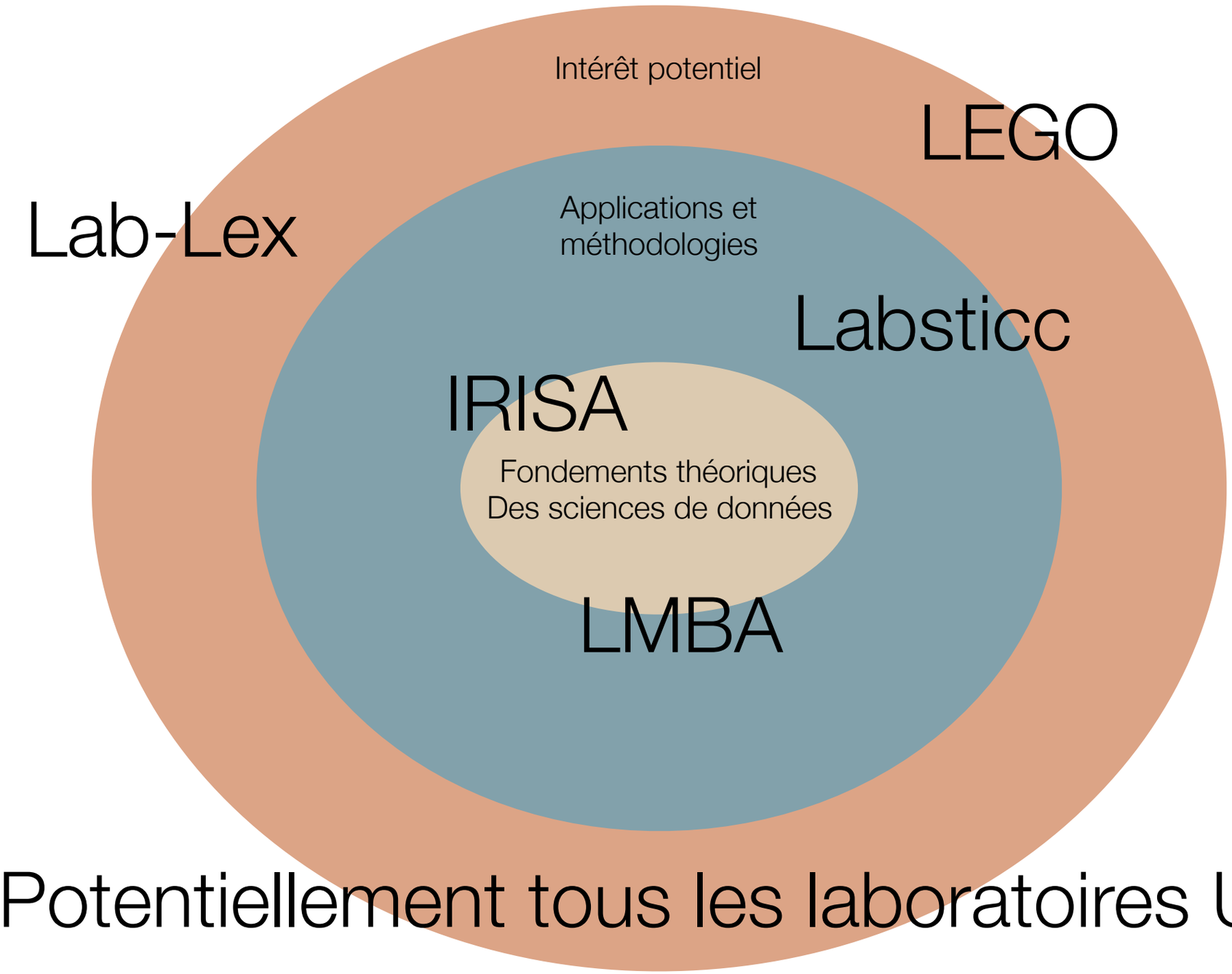
DECIDE

Contributions du LabSticc à la SD

LabSticc

- Domaines applicatifs :
  - Aide à la décision dans les systèmes d'informations géographiques au service de l'environnement (par exemple à partir de capteurs fixes ou mobiles)
  - Fouille de données appliquée à la classification et modélisation d'états physiologiques et cognitifs
  - Aide à la décision, fouille de données et optimisation dans l'industrie du futur
- Effectif total : 5 PR, 10 MC, 1 IR CNRS, 3 IE, 2 post-doc, 11 doctorants
- Effectifs UBS (depuis le site web) :
  - 3 PRU, 2 MC, 2IE, 1IR CNRS, 2 doctorants

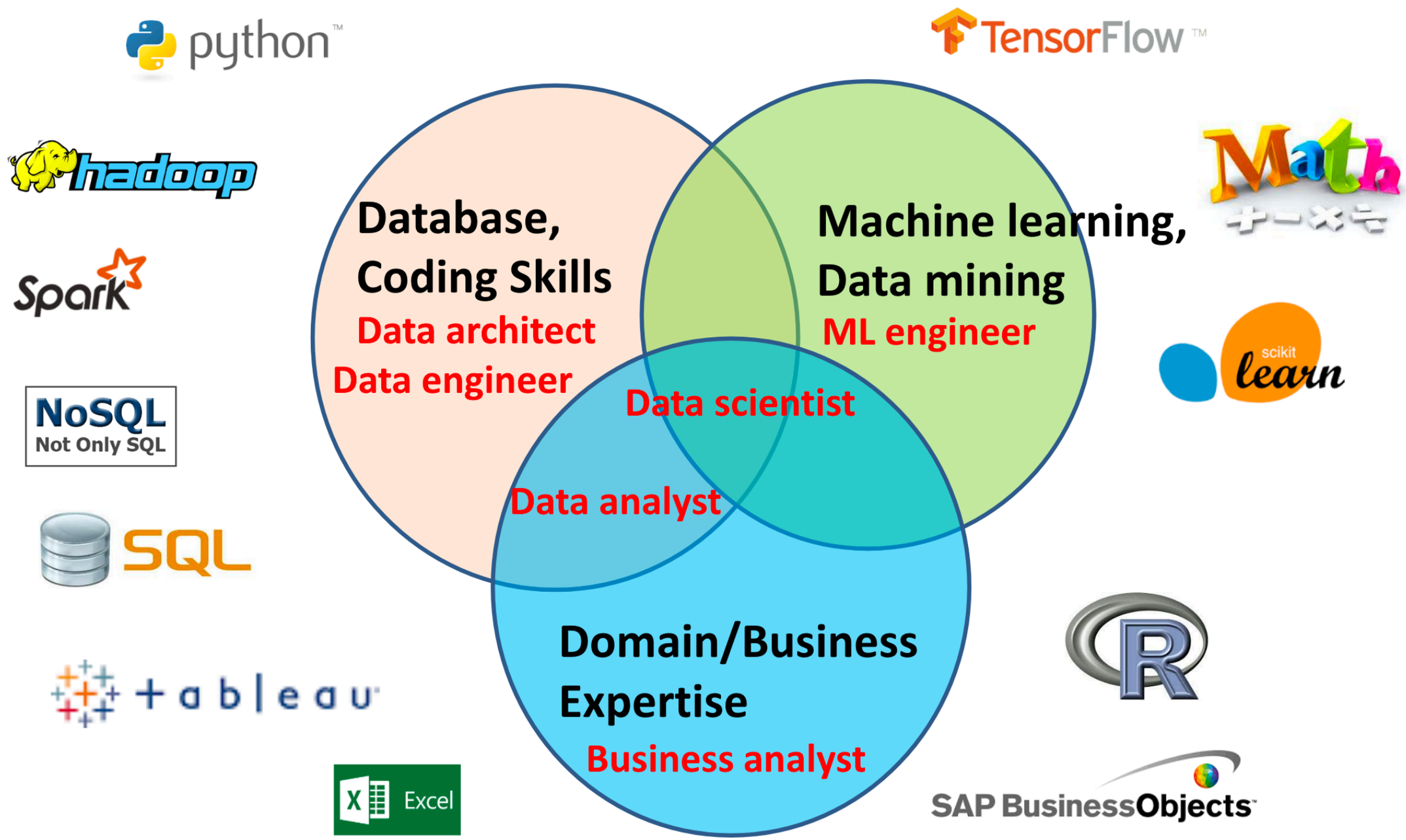
Différents niveaux de contribution



Potentiellement tous les laboratoires UBS

Data Science à l’UBS
Volet formations

Quelques orientations professionnelles



Formations de l'UBS dans le domaine SD

- IUT STID (coeur de formation)
- 3 Masters du DMIS (UFR SSI) sont concernés
  - Master DSMS (Data Science et Modélisation Statistique)
  - Master Ingénierie Mathématiques (à distance)
  - Master Informatique parcours AIDN (Applications Interactives et données numériques)
- Cursus Master Ingénierie (CMI) en Data science (formation en 5 ans)
- Master Erasmus Mundus Copernicus : 'Geodata Science'

COPERNICUS MASTER IN DIGITAL EARTH

Merci pour votre attention