



# Automatiser et démocratiser le ML : retour d'expérience

18/12/2019

clement.romac@weenove.fr



weenove

[www.weenove.fr](http://www.weenove.fr)

# Brève présentation

- Ma mission lors de mes **3 années d'alternance chez Weenove** : donner **accès au Machine Learning à nos utilisateurs**
- Accompagnement de sociétés dans des **projets de Data Science** durant mes expériences



# Fonctionnement de Biwee



Connexion à toutes vos sources de données, internes et externes



Plateforme intuitive et rapidité d'élaboration de vos tableaux



Partage des tableaux, gestion des accès et sécurité de vos données



Actualisation de vos informations en temps réel



Représentations graphiques variées et personnalisables



Interface utilisable sur PC, tablette et smartphone



# Clients, partenaires & éco-système

## Quelque références



## Quelque partenaires



## Ils nous soutiennent

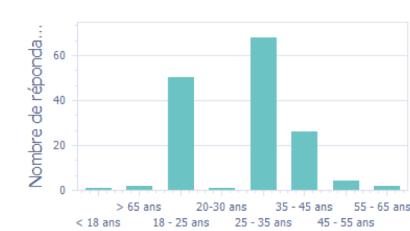


# Fonctionnement de BiWEE

**Tableau de bord**

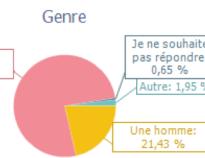
Source de données : Donnée enquête marketing ...

Filtre 2 : âge



Age Group	Nombre de répondants
< 18 ans	~5
18 - 25 ans	~50
25 - 35 ans	~65
35 - 45 ans	~25
45 - 55 ans	~5
> 65 ans	~5

Filtre 2 : genre



Genre	Pourcentage
Une femme	75,97 %
Une homme	21,43 %
Autre	1,95 %
Je ne souhaite pas répondre	0,65 %

Carte d'identité

Habitude de consommation

Habitude de consommation (2)

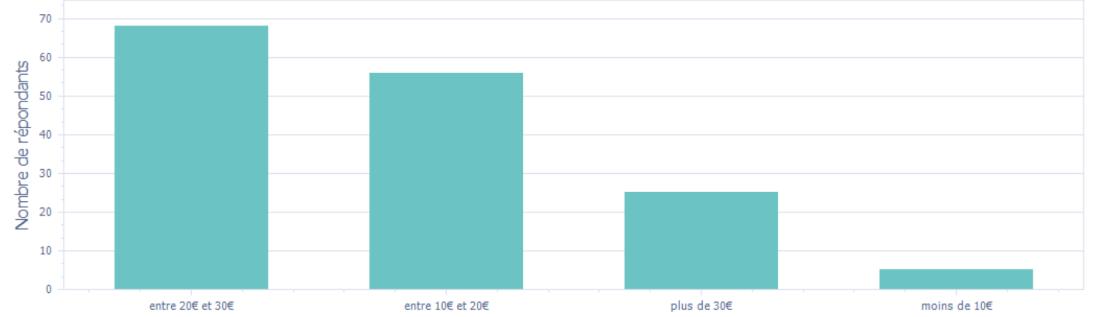
A propos de la librairie

Achetez-vous sur Internet ?



Statut	Pourcentage
Oui	47,40 %
Non	52,60 %

Combien dépensez vous lorsque vous vous déplacez dans une librairie ?



Spending Level	Nombre de répondants
entre 20€ et 30€	~68
entre 10€ et 20€	~55
plus de 30€	~25
moins de 10€	~5

Éléments graphiques



# Plan

- Le contexte du sujet
- Le Machine Learning automatisé, un aperçu
- Notre démarche chez Weenove
- Notre retour d'expérience



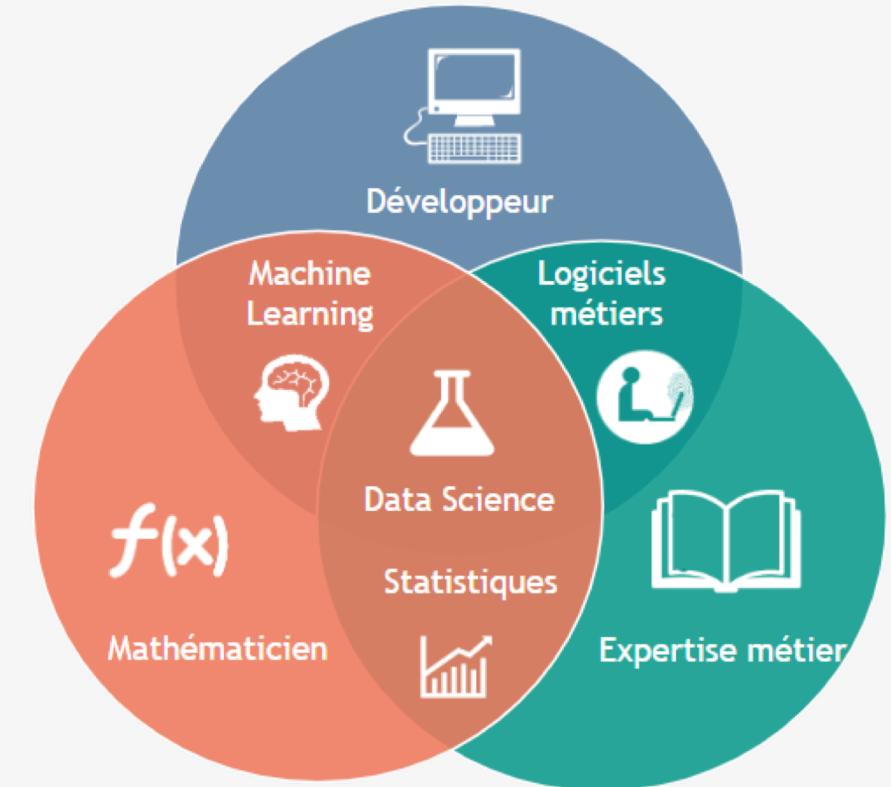
## Le contexte du sujet



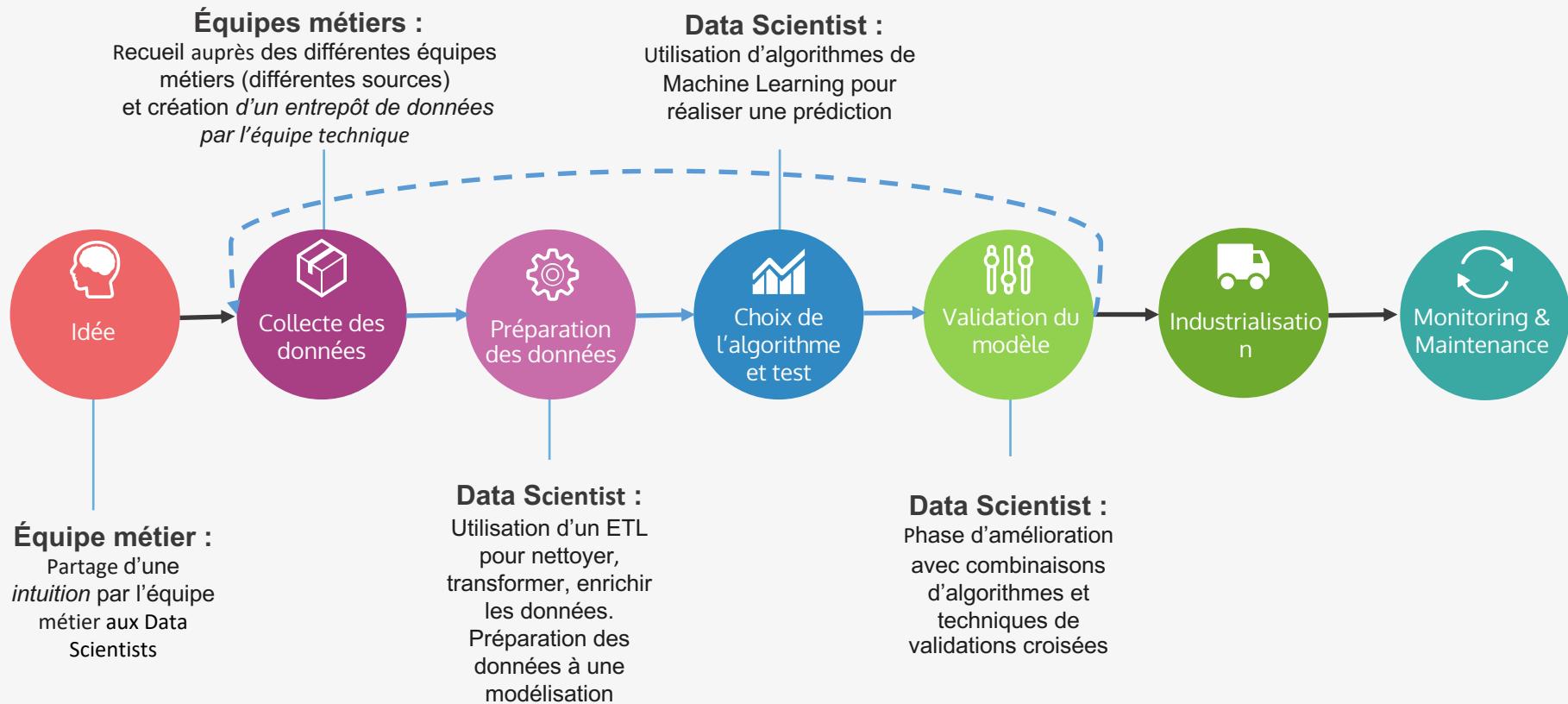
# Contexte - La Data Science

## Définitions

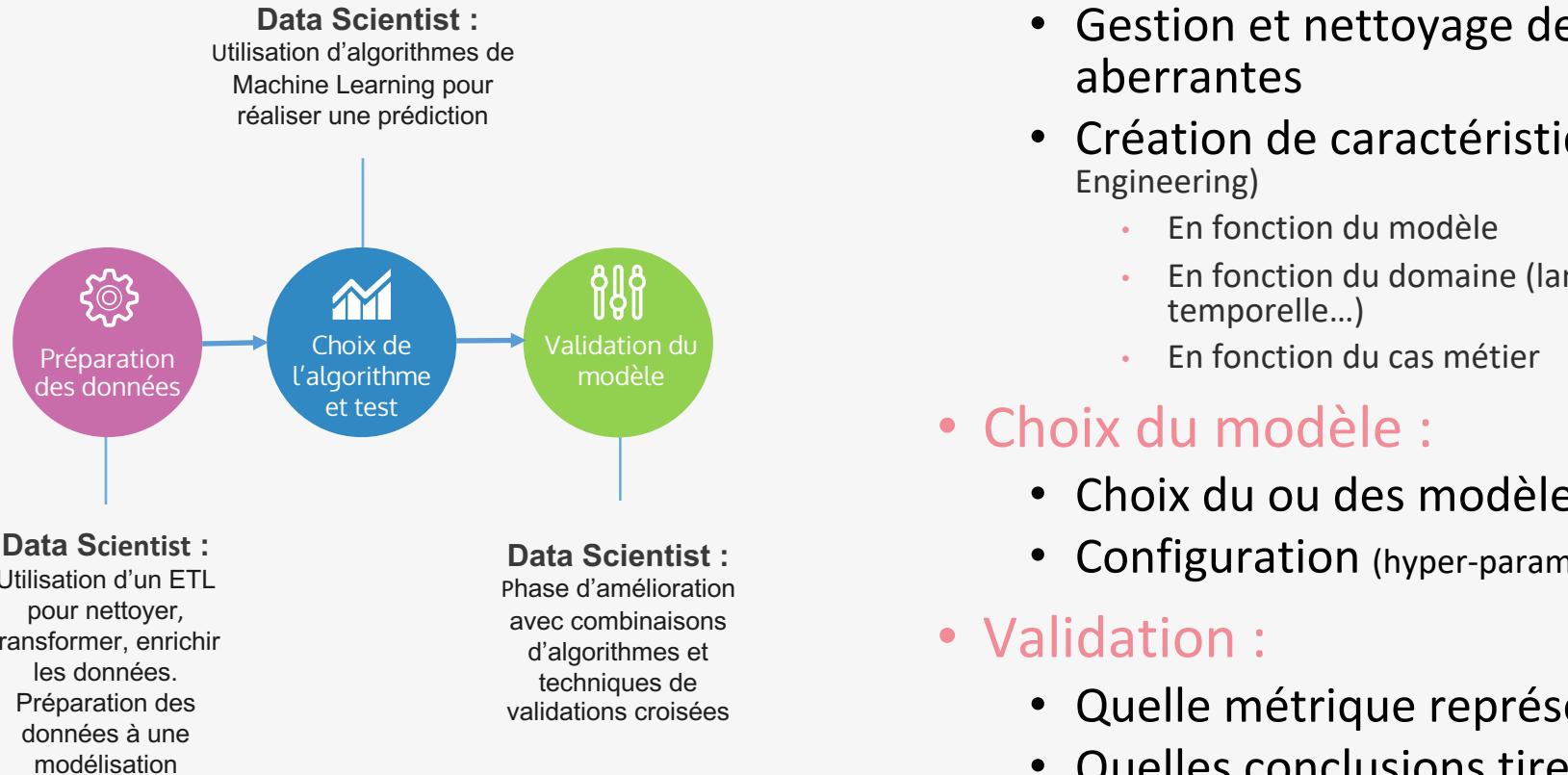
- **Données** : informations stockées par un ordinateur
- **Data Science** : Analyser / traiter les données pour en extraire de l'information (*Academic Press, 1995*)



# Contexte – Le pipeline d'un projet



# Contexte – Le pipeline d'un projet



## • Préparation des données (75% du temps selon Caruana, 2015) :

- Gestion et nettoyage des valeurs manquantes et aberrantes
- Crédit de caractéristiques pertinentes (ou Feature Engineering)
  - En fonction du modèle
  - En fonction du domaine (langage naturel, images, prévision temporelle...)
  - En fonction du cas métier

## • Choix du modèle :

- Choix du ou des modèles (capacité vs coût)
- Configuration (hyper-paramètres)

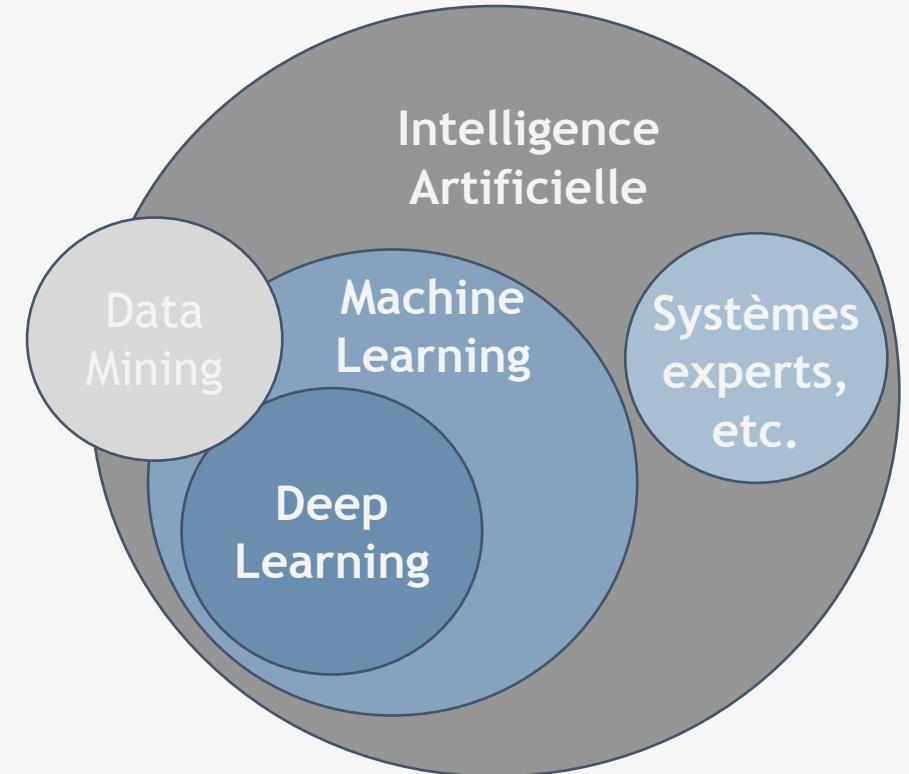
## • Validation :

- Quelle métrique représentative ?
- Quelles conclusions tirer d'un résultat ?

# Contexte – Le Machine Learning

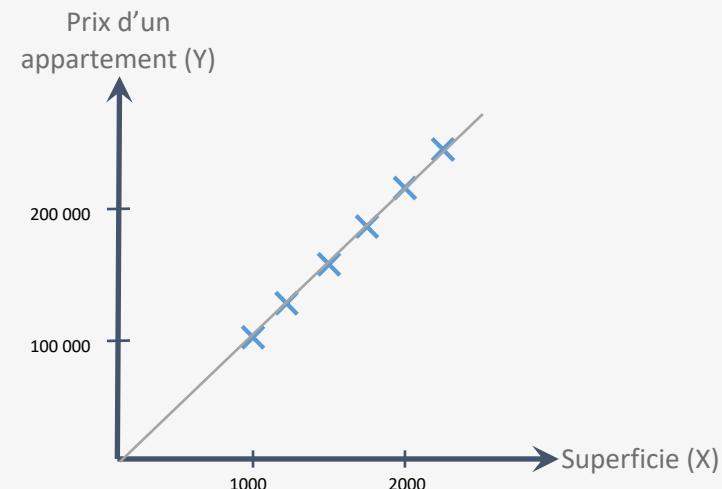
## Définitions

- **Machine Learning** : "Un programme informatique est dit apprenant d'une expérience E pour une tâche T avec sa performance mesurée par P, si sa performance à réaliser la tâche T mesurée par P s'améliore avec E". (Mitchell, 1997)



# Contexte – Le Machine Learning

X : Superficie (m <sup>2</sup> )	Y : Prix (en €)
2000	200 000
2500	250 000
1600	160 000
1200	120 000
2300	230 000
1800	180 000



$$\hat{Y}(X) = aX + b$$

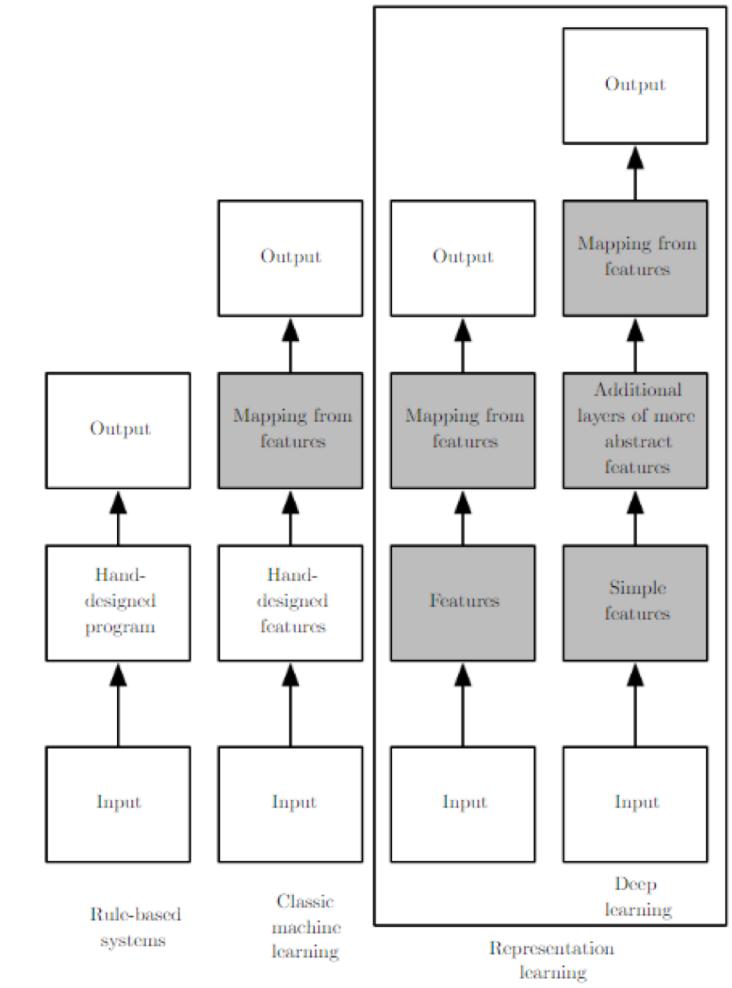
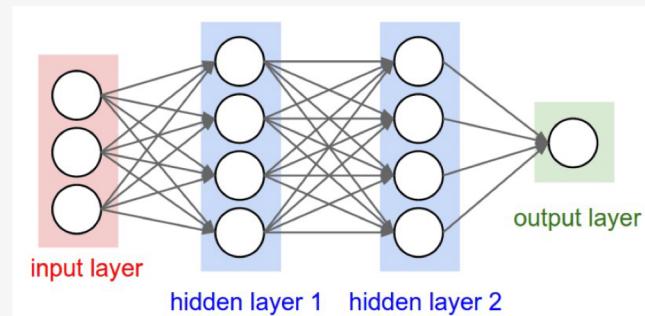
# Contexte – Le Machine Learning

<u>Algorithme</u>	<u>Hyper-paramètres</u>	<u>Avantages</u>	<u>Inconvénients</u>
Régression (linéaire, polynomiale, logistique)	<ul style="list-style-type: none"><li>Degré si polynomiale</li><li>Type de régularisation</li><li>Coefficient de régularisation</li></ul>	<ul style="list-style-type: none"><li>Rapide</li><li>Explicable</li></ul>	<ul style="list-style-type: none"><li>Peu adaptée aux problèmes complexes et non-linéaires</li></ul>
Arbre de décision	<ul style="list-style-type: none"><li>Nombre minimum d'éléments de feuille</li><li>Profondeur maximum</li><li>Critère de séparation</li></ul>	<ul style="list-style-type: none"><li>Rapide</li><li>Explicable</li><li>Non-linéaire</li></ul>	<ul style="list-style-type: none"><li>Enclin au sur-apprentissage</li><li>Performances moyennes</li></ul>
Naive Bayes	<ul style="list-style-type: none"><li>Loi à priori si une variante est utilisée</li></ul>	<ul style="list-style-type: none"><li>Rapide</li><li>Nécessite peu d'exemples</li><li>Efficace</li></ul>	<ul style="list-style-type: none"><li>Hypothèse d'indépendance forte</li><li>Performances moyennes</li></ul>
SVM	<ul style="list-style-type: none"><li>Fonction de noyau</li><li>Coefficient de régularisation</li></ul>	<ul style="list-style-type: none"><li>Très performant</li><li>Applicable à tous les problèmes</li></ul>	<ul style="list-style-type: none"><li>Long et coûteux</li></ul>
K plus proches voisins	<ul style="list-style-type: none"><li>K</li></ul>	<ul style="list-style-type: none"><li>Rapide</li></ul>	<ul style="list-style-type: none"><li>Performances moyennes</li></ul>
Random Forest	<ul style="list-style-type: none"><li>Nombre d'arbres</li><li>Hyper-paramètres des arbres</li></ul>	<ul style="list-style-type: none"><li>Très performant</li><li>Applicable à tous les problèmes</li></ul>	<ul style="list-style-type: none"><li>Long et coûteux</li><li>Un peu enclin au sur-apprentissage</li></ul>
Gradient Boosting	<ul style="list-style-type: none"><li>Hyper-paramètres des arbres</li></ul>	<ul style="list-style-type: none"><li>Très performant</li><li>Applicable à tous les problèmes</li></ul>	<ul style="list-style-type: none"><li>Très long et coûteux</li></ul>

# Contexte – Le Deep Learning

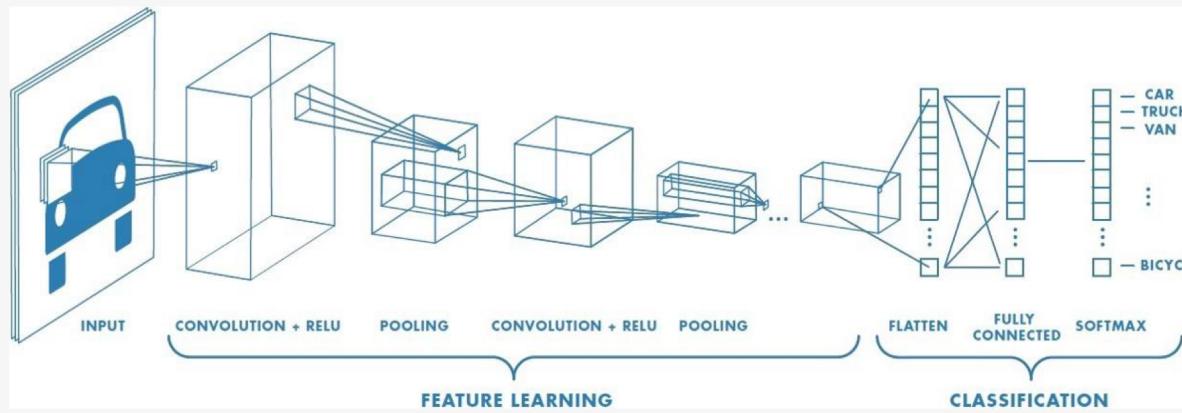
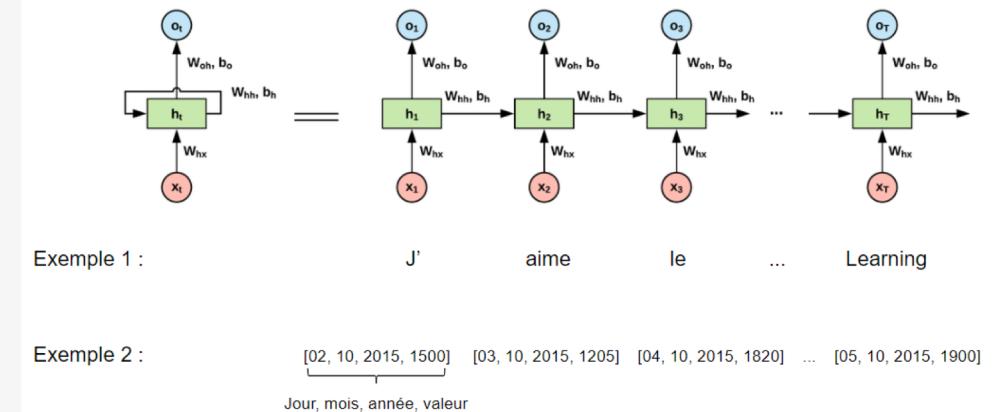
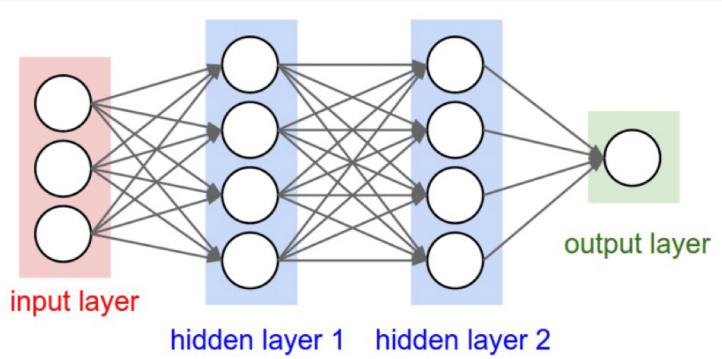
## Définitions

- **Deep Learning** : Mouvement qui regroupe les techniques d'apprentissage par « couches de représentations ».
- **Réseau de neurones** : Algorithme de Machine Learning basé sur la connexion de neurones artificiels



Source : <http://www.deeplearningbook.org>

# Contexte – Le Deep Learning



Source : <https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53>

# Contexte – L'application de la Data Science

## Les champs d'application

- GAFA
  - Traduction, reconnaissance faciale, ciblage publicitaire, reconnaissance vocale...
- Partout où il y a des données
  - Affiner son recrutement (RH)
  - Prévision de chiffre d'affaires (Finances)
  - Anomalies dans des tests de résistance de pièces pour satellites (Industrie)

## Les problématiques

- Besoin de Data Scientists expérimentés
- Coûts d'un projet pour une petite structure
- Pas toujours facile de faire appel à un prestataire

# Le Machine Learning automatisé



# AutoML – Le domaine scientifique

## Objectifs

- Automatiser :
  - la **préparation des données**,
  - le **choix et la configuration d'un algorithme**
  - la **validation du modèle**
- Rendre le Machine Learning accessible à tous

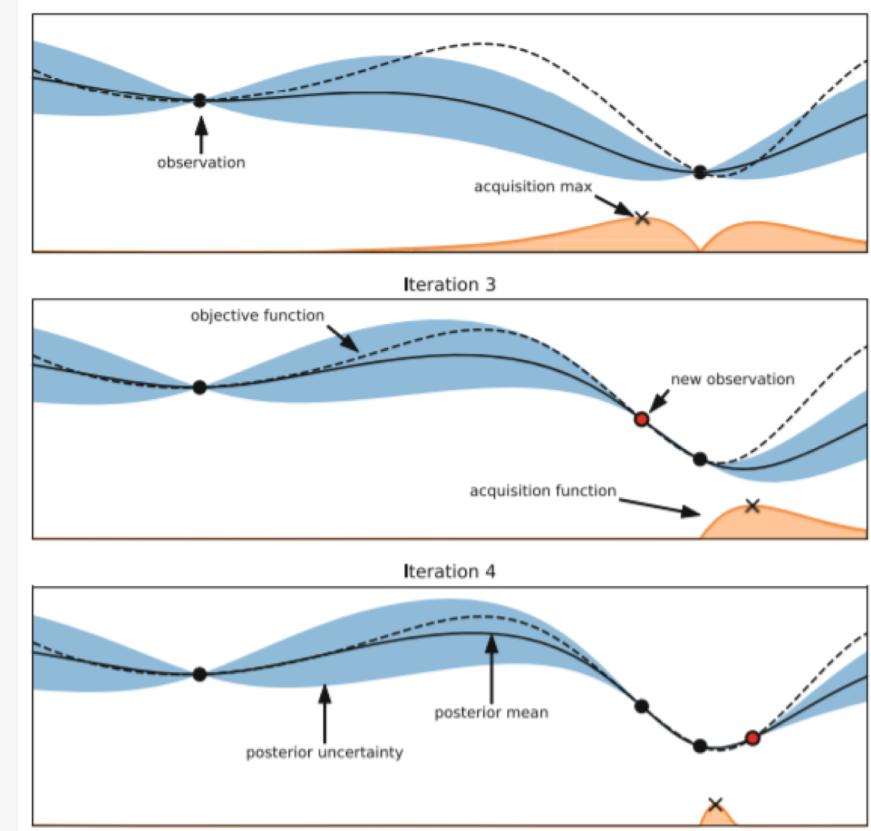
## Histoire

- Premières méthodes datent des années 90
- Engouement à partir de l'atelier AutoML de ICML 2014
- Poussé en grande partie par Google

# AutoML – Les méthodes

## Optimisation d'hyper-paramètres (HPO)

- Approche le choix des hyper-paramètres comme un **problème d'optimisation d'une fonction « boîte noire »**
- Certaines méthodes considèrent dans les hyper-paramètres les **méthodes de préparation et le choix de l'algorithme**
- Les méthodes à base d'**optimisation bayésienne** sont aujourd'hui les plus performantes
- Autres méthodes :
  - Algorithmes génétiques
  - Apprentissage par renforcement



Source : Hutter (2019)

# AutoML – Les méthodes

## Meta-Learning

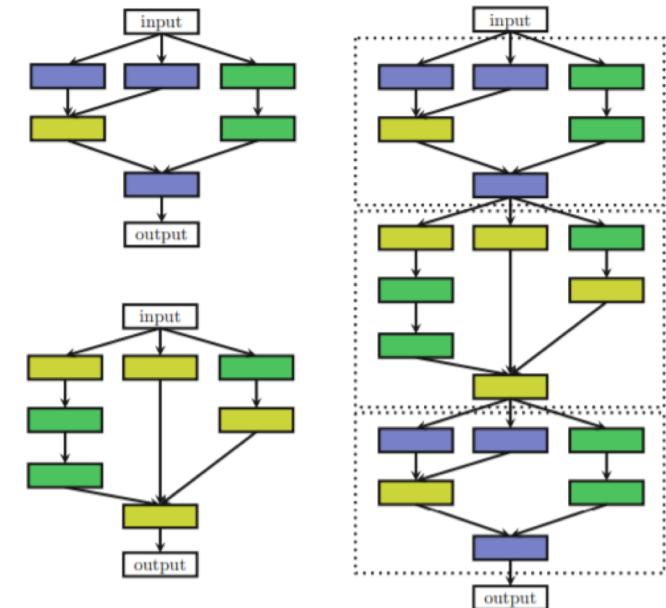
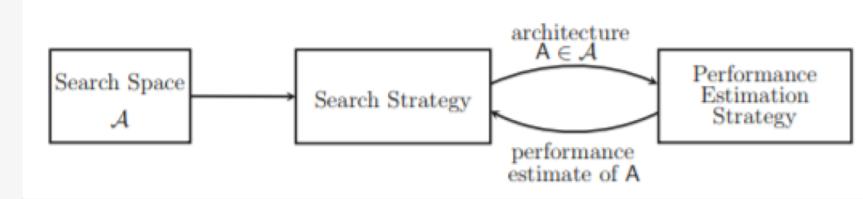
- Principe : capitaliser sur l'expérience des projets précédents
- Création de Meta-Data sur les projets
  - Sur le projet en lui-même (Meta-Features)
  - Sur les configurations testées
  - Sur les résultats obtenus
- Utiliser les Meta-Data pour un nouveau projet :
  - Trouver les projets qui ressemblaient à notre projet en cours et utiliser leur configuration
  - Apprendre un modèle de Machine Learning capable de prédire la bonne configuration pour un projet
  - Initialiser les hyper-paramètres avec ceux des projets les plus ressemblants à notre projet en cours

Source : Hutter (2019)

# AutoML – Les méthodes

## Neural Architecture Search (NAS)

- Objectif : Trouver une architecture optimale pour un réseau de neurones
- Définition d'un espace de recherche
  - Aujourd'hui presque essentiellement centré sur les réseaux de neurones convolutifs
  - Peut être initialisé par du Meta-Learning
- Les mêmes méthodes que celles pour l'optimisation d'hyper-paramètres sont utilisées
- Grosse problématique de temps d'apprentissage



Source : Hutter (2019)

# AutoML – L'application des méthodes

## Automatisation de la préparation

- Certaines méthodes d'optimisation d'hyper-paramètres prennent en compte des méthodes de préparation (bien que très primaires)
- Le Meta-Learning peut permettre de réutiliser des méthodes ayant marché par le passé

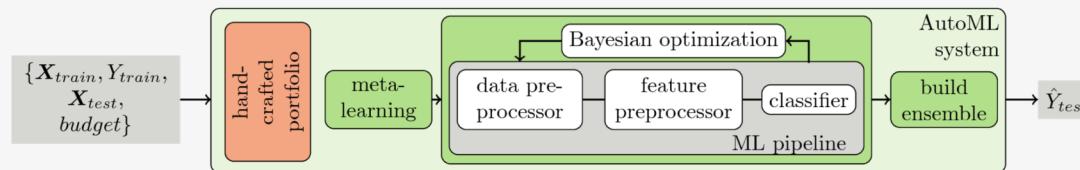
## Automatisation de l'entraînement

- Machine Learning :
  - Le choix du modèle peut être intégré à l'optimisation d'hyper-paramètres ou fait avec du Meta-Learning
  - Pour les hyper-paramètres de l'algorithme :
    - Initialisation avec du Meta-Learning
    - Optimisation bayésienne
- Deep Learning :
  - L'architecture :
    - Optimisation bayésienne (+ Meta-Learning)
  - Les autres hyper-paramètres :
    - Meta-Learning

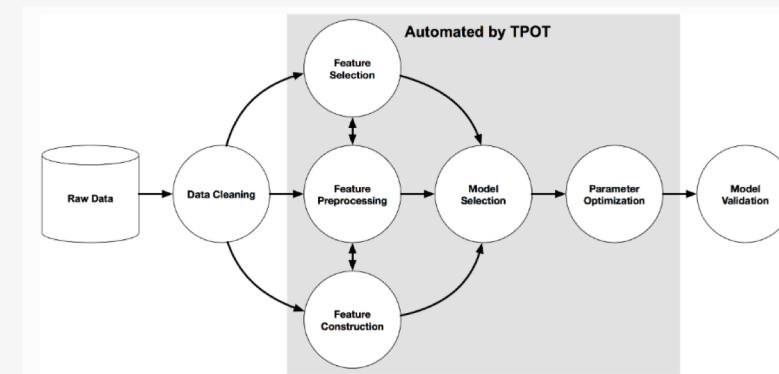
# AutoML – Les outils

## Libres

- Auto-sklearn



- TPOT



- Auto-Keras

## Propriétaires

- Data Robot

- Un des leaders
- Flou sur les méthodes utilisées

- Google Cloud AutoML

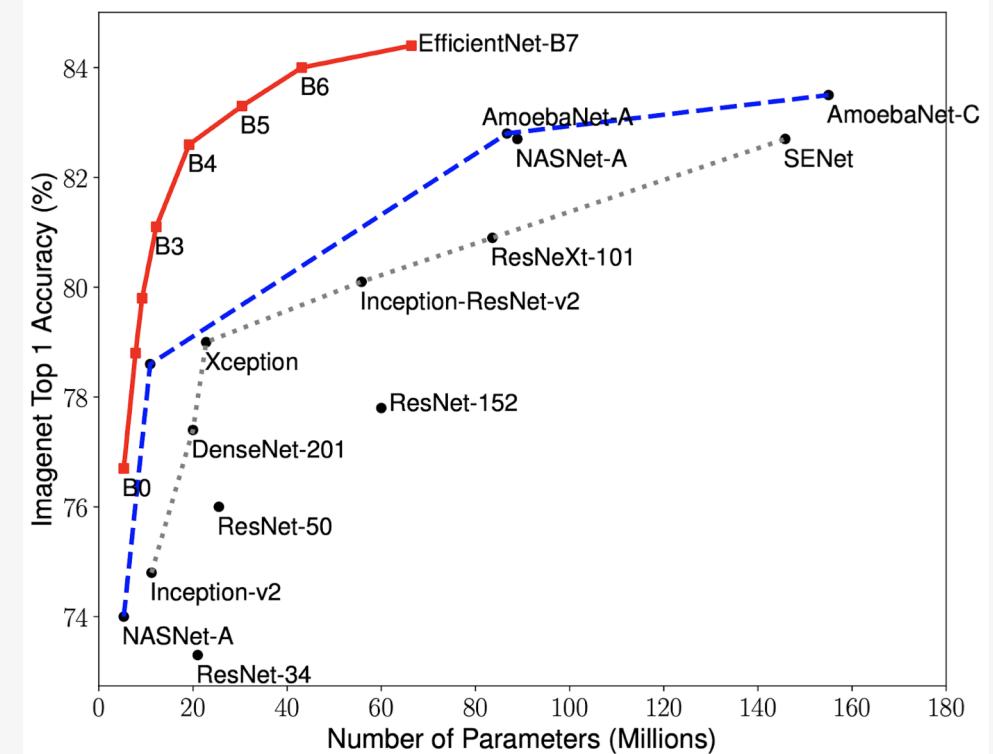
- Seulement en bêta
- Classification d'images, classification de texte, traduction

- Microsoft AzureML

- Bibliothèque Python

## Résultats – Les réponses apportées

- Les performances d'outils tels que Auto-sklearn, TPOT ou Data Robot donnent des résultats similaires à ceux de Data Scientists sur certains cas
- Les résultats de **NAS** sur de la classification d'images peuvent être très bon
- Les outils tels que Data Robot simplifient grandement l'accès au Machine Learning au travers d'une **interface intuitive**
- Peut **guider le Data Scientist** et lui faire gagner du temps



Source : Tan and Le (2019)

## Résultats – Les réponses en suspend

- La préparation des données automatisée n'est que très primaire (elle nécessite de l'expertise métier)
- La prévision temporelle est un peu délaissée
- Les méthodes de NAS sont très peu matures
- La plupart des outils sont sous forme de bibliothèque de code
- Les outils sous forme d'interface requièrent encore énormément d'expertise

## Résultats – Les nouvelles questions

- Comment réduire les **ressources de calcul** nécessaires ?
- Mettre à disposition le résultat ne suffit pas ! Comment aider l'utilisateur à comprendre son modèle et ses résultats ?
- Tirer les bonnes conclusions à partir de **métriques** n'est pas chose facile (même pour un Data Scientist). Comment guider l'utilisateur dans l'interprétation des performances de son modèle ?

## Notre démarche chez Weenove



# Weenove – Notre démarche

- Donner accès à du Machine Learning **de manière totalement automatisé** directement au sein des **tableaux de bord**
- Objectif :
  - S'adapter à un **maximum de cas**
  - Proposer une solution nécessitant le **minimum de connaissances**



# Weenove – Notre démarche

Pour cela :

- Interface intuitive et compréhensible
- Automatisation maximum :
  - Ne demander que la **colonne à prédire**
  - Exécution **déportée et scalable** (Docker)
  - **Déploiement automatique** (AzureML)
  - Automatisation du **traitement des données et de l'entraînement** (Scikit-Learn, Random Search)



<https://www.youtube.com/watch?v=bfNiQISjoUE&feature=youtu.be>

- **11 cas concrets** depuis 2 ans (dont 6 menés au bout)
  - 4 prévisions temporelles
    - 6 mois d'historique de consommation électrique toutes les heures (3600 lignes)
  - 1 régression
  - 1 classification (binaire)
    - 278 000 lignes, 30 colonnes

- Processus :

- Discussion sur les problématiques et les données stockées
- Premières explorations de données
- Lancement de Biwee AutoML
- Analyse des résultats, traitement des données à la main si nécessaire

# Weenove – Notre retour d'expérience



- Les résultats sont très encourageants
- Les problématiques de l'AutoML sont réelles
- La préparation des données est un des gros freins
- Les temps d'entraînement peuvent être problématiques (même avec du Random Search)
- Autres problématiques : données mal réparties

# Weenove – Notre retour d'expérience



- Les résultats doivent être **interprétables** :
  - Gros frein à l'adoption du ML en général
  - Des solutions arrivent : **SHAP**
  - Loin d'être encore totalement interprétable (même pour un Data Scientist !)

# Weenove – La suite

- Continuer à améliorer notre interface
- Déploiement à plus d'utilisateurs
- Amélioration de notre module prédictif
  - Meta-Learning (Intelligence Collective)
  - Données OpenData (exemple météo)
  - Améliorations du choix des hyper-paramètres

## Conclusion



- Le nombre de projets de Data Science explosent
- Les étapes d'un projet peuvent potentiellement être automatisées
- Le domaine du Machine Learning automatisé tente de résoudre cette problématique
- Depuis 2014, des progrès ont été faits mais il reste encore du chemin pour la mise à disposition à tous
- Les progrès en Machine Learning automatisé vont peut-être faire évoluer le métier de Data Scientist



# Bibliographie

- Academic Press, I.  
1995. Data science and its application. tokyo : Academic press.
- Mitchell, T. M.  
1997. Machine Learning, McGraw-Hill series in computer science. McGraw-Hill.
- Caruana  
2015. Research opportunities in automl.
- Bergstra, J. and Y. Bengio  
2012. Random search for hyper-parameter optimization. 13 :281– 305.
- Hutter, Kotthoff, Vanschoren  
2019. Automated machine learning : Methods, systems, challenges.
- Tan, M. and Q. V. Le  
2019. EfficientNet : Rethinking model scaling for convolutional neural networks.



# Questions ?

**MERCI !**