



Introduction à la Data Science

Alessandro Cerioni

Data Scientist

`alessandro.cerioni@etat.ge.ch`

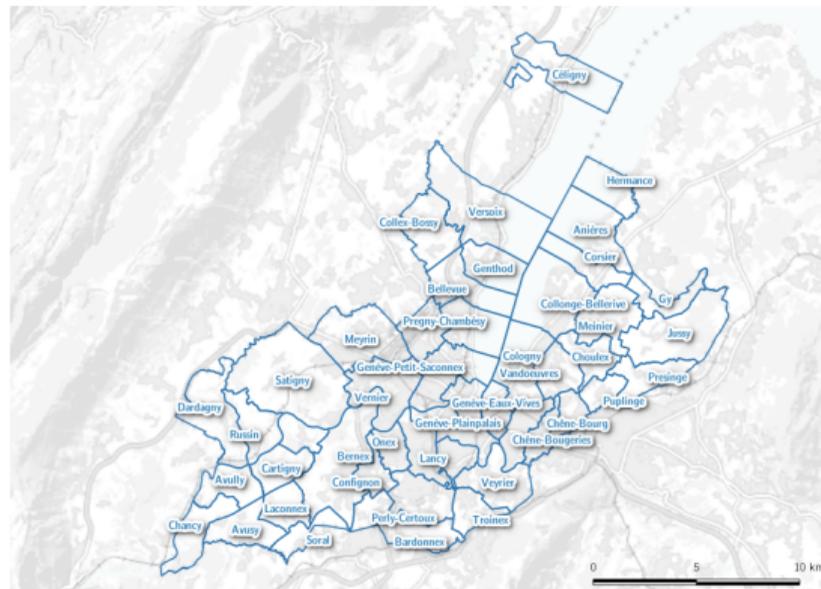
Qui suis-je ?

- 04/2011 : PhD en Physique Théorique, **Université de Bologne**
- 11/2011 → 03/2015 : post-doc à l'**ESRF** et au **CEA - Grenoble**
- 09/2015 → 10/2016 : formation continue à l'**INSA - Lyon, Mastère Spécialisé en Informatique**
- 11/2016 → 03/2020 : Chef de produit de la plate-forme data.grandlyon.com / Data Scientist à la **Métropole de Lyon**
- depuis le 1^{er} avril 2020 : Data Scientist à l'**Etat de Genève**



...et l'Etat de Genève ?

- 45 communes
- 282,49 km²
- 514 314 habitants en octobre 2022



Organisation

<https://www.ge.ch/document/organisation-etat>, <https://www.ge.ch/organisation>

<https://www.ge.ch/document/conseil-etat-legislature-2018-2023>



REPUBLIQUE
ET CANTON
DE GENEVE

POSE TENERIAS LUX

Département du territoire
Direction de l'information du territoire

06/11/2023 - Page 3

...et la Direction de l'Information du Territoire ?

- < 30 personnes

Missions historiques

- Cadastre, mensuration officielle
- Données du territoire
- Cartographie
- Expertise en systèmes d'information géographiques

Nouvelles missions

- Gouvernance de la donnée et de l'information
- Référentiel de données multi-morphes
- Packaging de données
- Qualité de la donnée (conseil, normalisation, ...)
- Promotion et innovation en matière de données



...et le Système d'Information du Territoire à Genève ?

*“Le Système d'Information du Territoire à Genève (SITG) est un organisme fondé sur un réseau de partenaires publics ayant pour but de **coordonner, centraliser et diffuser largement les données relatives au territoire genevois**. Les partenaires partageant entre eux et avec un large public les données géographiques produites dans le cadre de leurs missions.”*

- fondé en 1991
- 16 partenaires publics (CH, FR) : voir <https://ge.ch/sitg/le-sitg/organisation/partenaires>

<https://ge.ch/sitg>



Sommaire

1 Qu'est-ce que la Data Science ?

- Applications à l'Etat de Genève
- Intelligence Artificielle

2 Machine Learning

- Les principales méthodes
- Variance vs biais

3 Réseaux de neurones, Deep Learning

- Introduction
- Application à l'Etat de Genève : détection de piscines à partir d'images aériennes

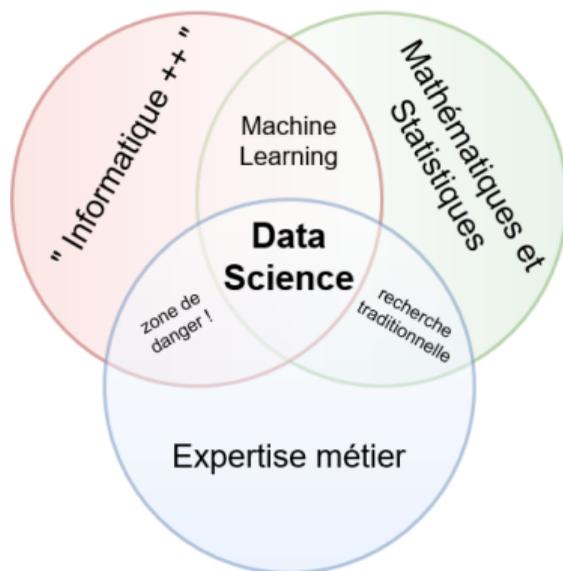


1. Qu'est-ce que la Data Science ?

...et l'Intelligence Artificielle ? ...et le Big Data ?

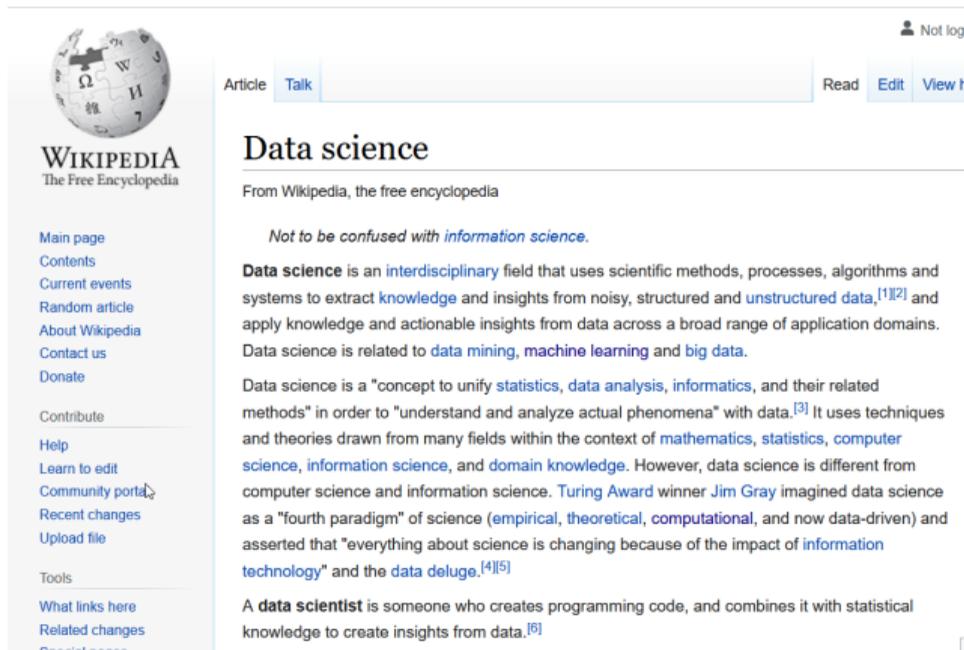


Qu'est-ce que la Data Science ?



Adaptation du diagramme publié à la page suivante : <https://tinyurl.com/m2qfavm>.

Qu'est-ce que la Data Science ?



The screenshot shows the Wikipedia article for "Data science". At the top left is the Wikipedia logo and navigation menu. The article title "Data science" is prominently displayed. Below the title, there is a sub-header "From Wikipedia, the free encyclopedia" and a note: "Not to be confused with information science." The main text defines data science as an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains. It is related to data mining, machine learning and big data. The text further explains that data science is a "concept to unify statistics, data analysis, informatics, and their related methods" in order to "understand and analyze actual phenomena" with data. It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, information science, and domain knowledge. However, data science is different from computer science and information science. Turing Award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational, and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge. A data scientist is someone who creates programming code, and combines it with statistical knowledge to create insights from data.

Source : https://en.wikipedia.org/wiki/Data_science



REPUBLIQUE
ET CANTON
DE GENEVE

POST TELEGRAMS LIJ

Département du territoire
Direction de l'information du territoire

06/11/2023 - Page 9

Data Science, pour faire quoi ?!

Deux grandes familles d'applications :

- 1 automatiser, faciliter de certaines tâches autrement “fastidieuses”, e.g. lecture automatique de texte, détection d'objets, ...
- 2 analyses descriptives, diagnostiques, prédictives, prescriptives



La Data Science à l'Etat de Genève

Exemples d'applications - en interne

- 2020 Détection de piscines (DIT¹) - voir <https://tech.stdl.ch/PROJ-GEPOOL/>
- 2021 COVID-19, confinement et qualité de l'air : analyse d'impact (OCEV²)
- 2021- Mesure de la qualité de l'air par des capteurs *low cost* (OCEV)
- 2022- Mesure du bruit par des capteurs *low cost* (OCEV)
- 2023- Valorisation des données de trafic diffusées par Waze (OCT³)
- 2023- Valorisation des données de pluviométrie diffusées par Netatmo (OCEau⁴)

1 Direction de l'information du territoire

2 Office cantonal de l'environnement

3 Office cantonal des transports

4 Office cantonal de l'eau

Le “Swiss Territorial Data Lab” (STDL)



Swiss Territorial
Data Lab

Protection des données | Mentions légales  

[Home](#)

[Présentation](#)

[Nos projets](#)

[Notre équipe](#)

[Nos partenaires](#)

[Évènements](#)

[Soumettre un projet](#)

[Contact](#)

Le Swiss Territorial Data Lab est ...



voir <https://www.stdl.ch/>



REPUBLIQUE
ET CANTON
DE GENEVE

POEST TENORIAS LUX

Département du territoire
Direction de l'information du territoire

06/11/2023 - Page 12

La Data Science à l'Etat de Genève

Exemples d'applications - en collaboration avec le STDL

2020 Détection de piscines (DIT⁵) - voir <https://tech.stdl.ch/PROJ-GEPOOL/>

2021 Détection de panneaux thermiques (OCEN⁶) - voir <https://tech.stdl.ch/PROJ-TPNL/>

2021 Détection d'arbres isolés (OCAN⁷) - voir <https://tech.stdl.ch/PROJ-TREEDET/>

2023 Vectorisation des plans historiques du cadastre (DIT)

2023- Détection d'objets occupant les toits (DIT, OCAN, OCEN)

5 Direction de l'information du territoire

6 Office cantonal de l'énergie

7 Office cantonal de l'agriculture et de la nature

La Data Science à l'Etat de Genève

COVID-19, confinement et qualité de l'air

Questions métiers

- 1 Les mesures de confinement liées à la COVID-19, ont-elles eu un impact sur la qualité de l'air ?
- 2 Si oui, dans quelle mesure ? [analyse descriptive]
- 3 Par le biais de quels facteurs (ex. : trafic, chauffage, ...) ? [analyse diagnostique]

Approche adoptée

- "normalisation météorologique", voir <https://tinyurl.com/y2uz388d>

Données utilisées

- mesures de qualité de l'air : NO₂, O₃, PM_x (SABRA)
- météo : température, pression, humidité, précipitation, direction et vitesse du vent, ... (MétéoSuisse)
- trafic routier (OCT), trafic aérien (Aéroport de Genève)



Qualité de l'air : les stations de mesure du ROPAG⁸ (SABRA)



Station fixe



Capteur passif



Stations mobiles

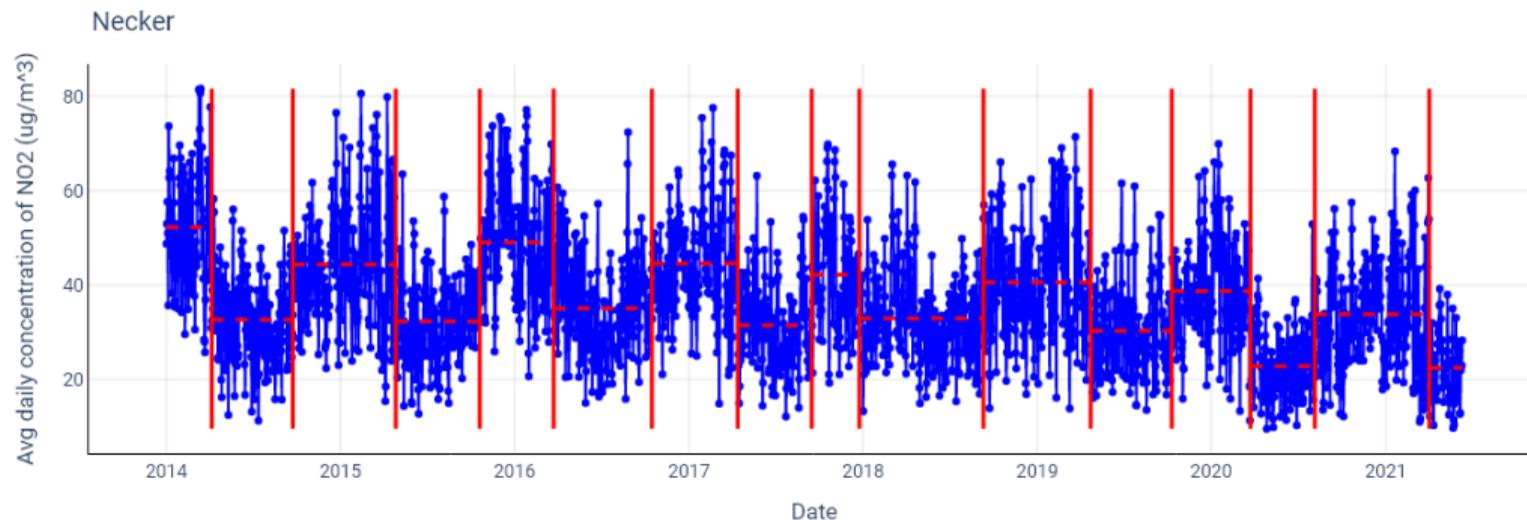


Pour plus d'infos, voir <https://tinyurl.com/j3tx6y6t>

8 Réseau d'observation de la pollution atmosphérique à Genève

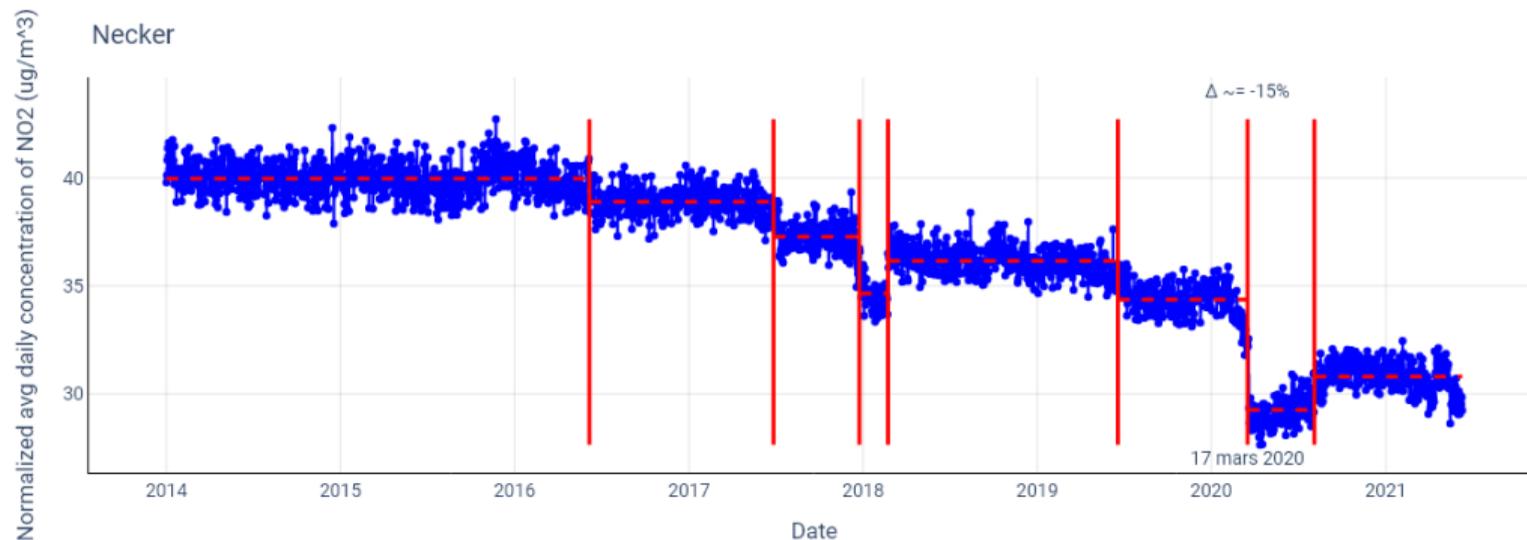
Normalisation météorologique

NO₂ @ Necker : concentration moyenne journalière, **AVANT** la normalisation météorologique



Normalisation météorologique

NO₂ @ Necker : concentration moyenne journalière, **APRÈS** la normalisation météorologique



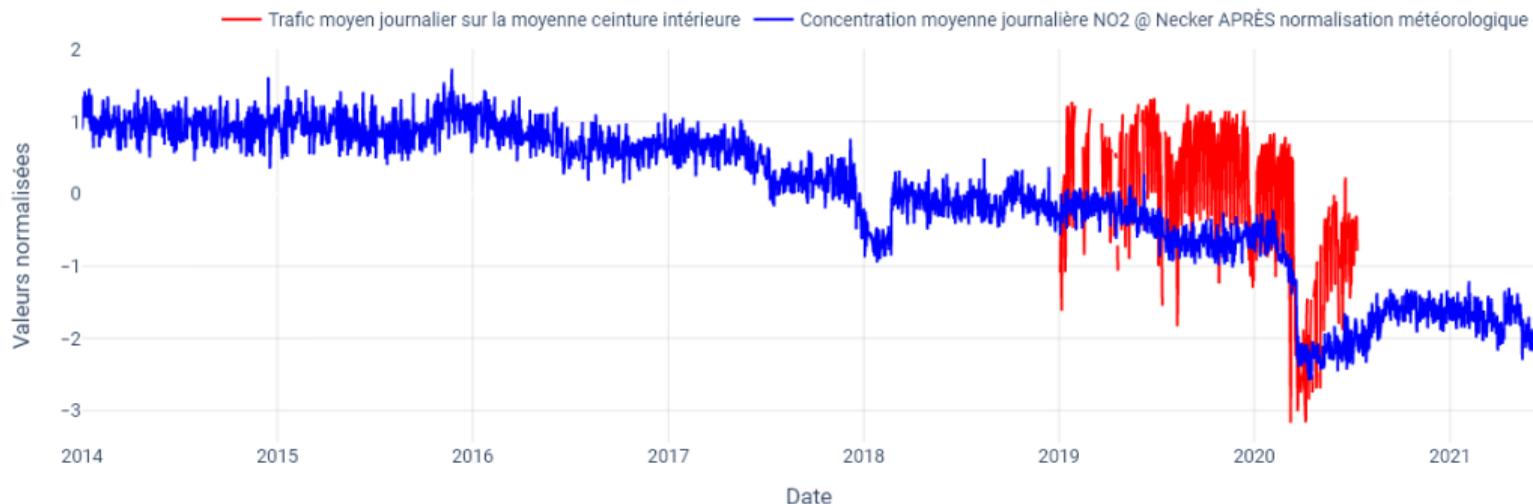
Oui, les mesures de confinement ont eu un impact sur la qualité de l'air !

voir aussi <https://tinyurl.com/5b29xy6b>, <https://tinyurl.com/newe65w3>



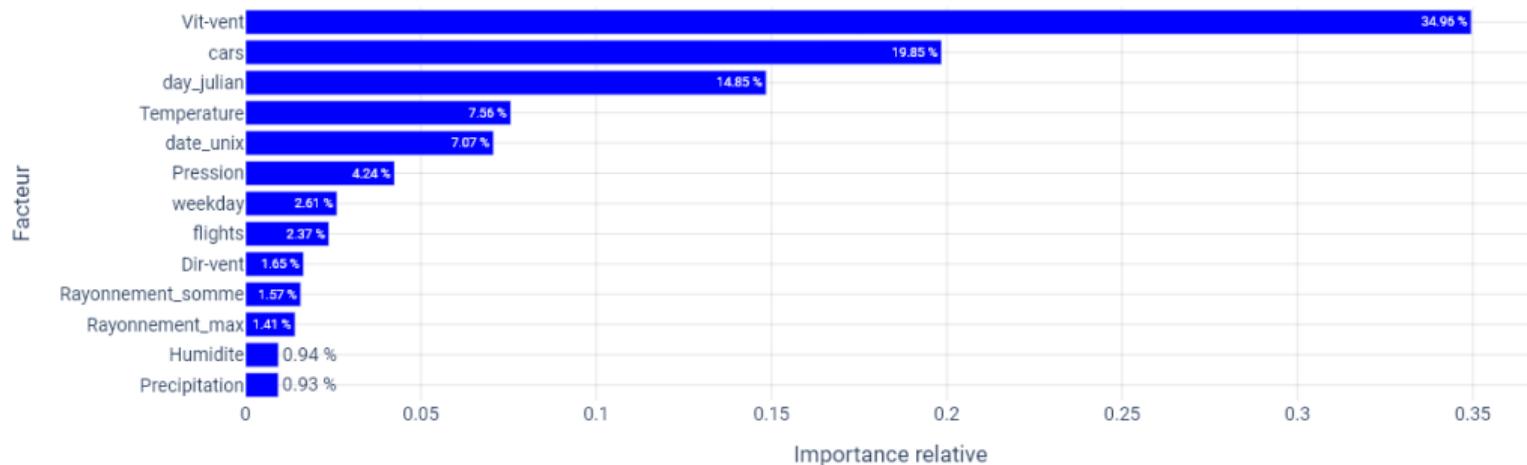
...par le biais de quels facteurs ?

NO₂ @ Necker : concentration moyenne journalière, **APRÈS** la normalisation météorologique



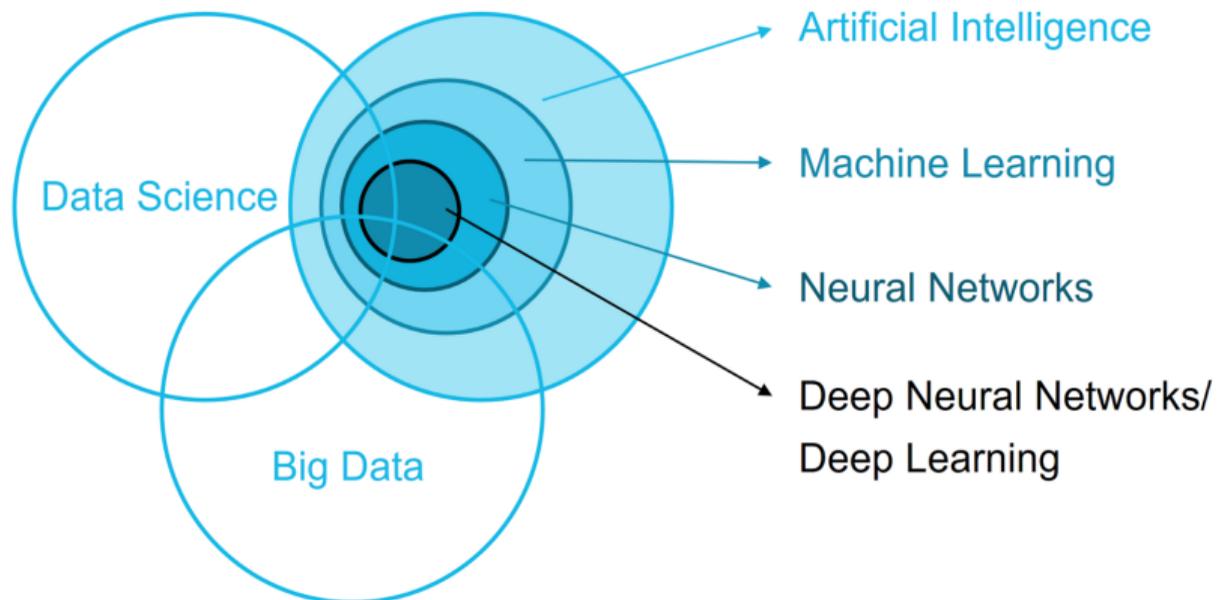
...par le biais de quels facteurs ?

NO₂ @ Necker : concentration moyenne journalière, **APRÈS** la normalisation météorologique



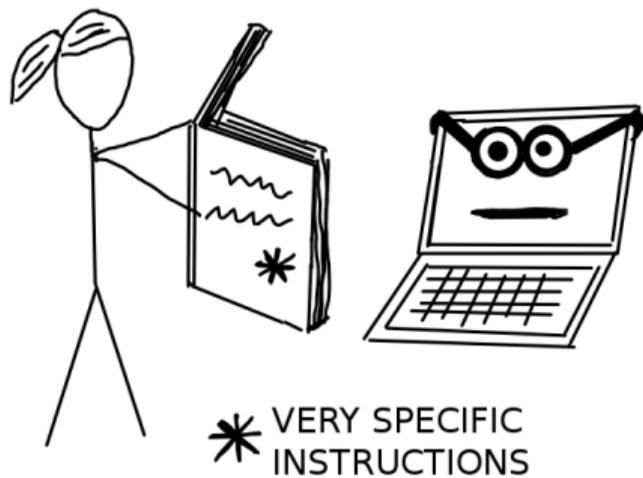
Voir aussi <https://www.grand-geneve.org/sites/default/files/fichiers/ggame/brochure-g2ame-hd.pdf>

Data Science, IA, Big Data

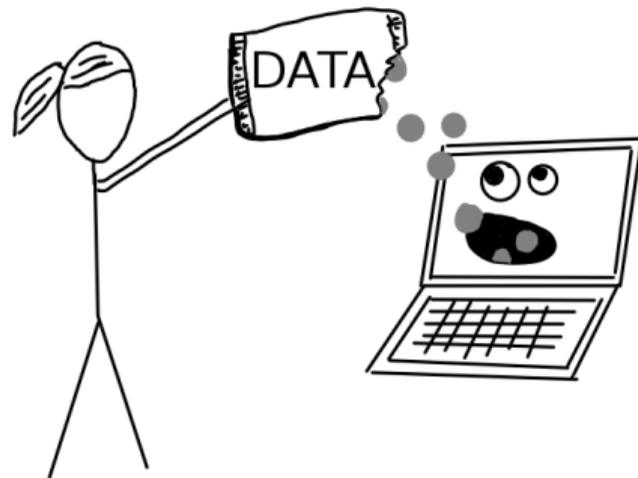


Source : <https://towardsdatascience.com/role-of-data-science-in-artificial-intelligence-950efedd2579>

Without Machine Learning



With Machine Learning

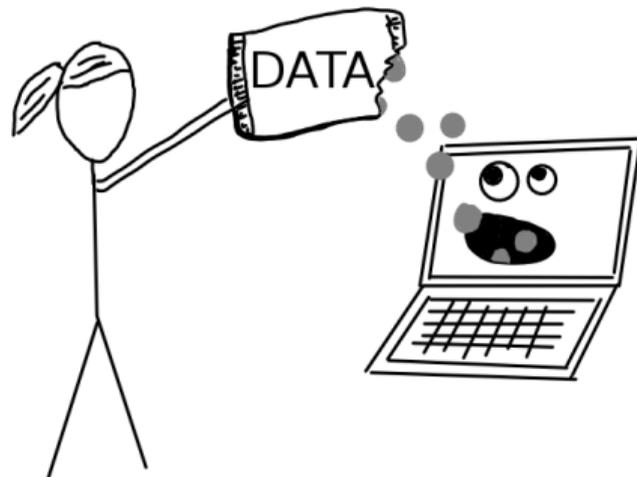


Source : <https://christophm.github.io/interpretable-ml-book/terminology.html>

Without Machine Learning



With Machine Learning



En réalité, des instructions (algorithmes) restent toujours nécessaires : celles permettant l'apprentissage automatique !

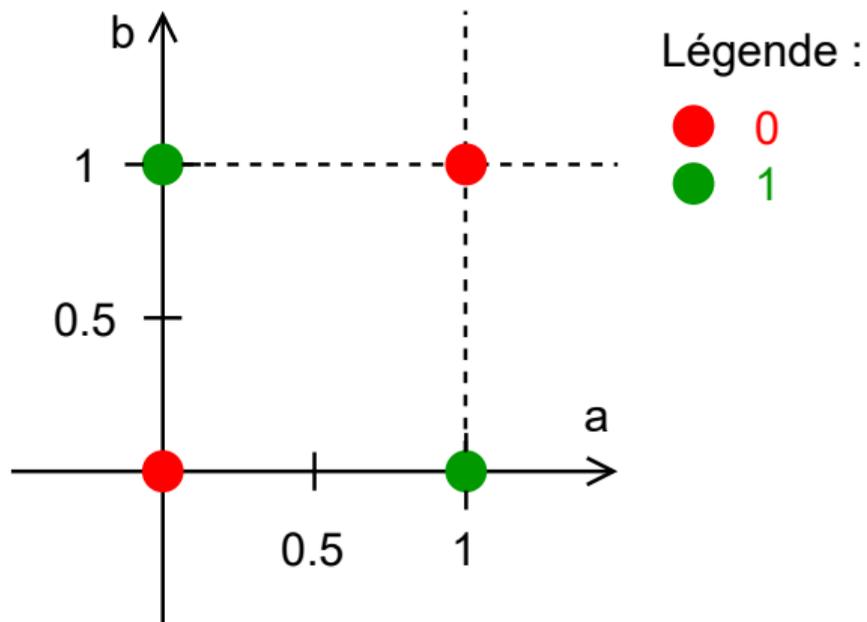
Source : <https://christophm.github.io/interpretable-ml-book/terminology.html>

IA sans/avec le Machine Learning : l'exemple de la fonction XOR

a	b	a XOR b
0	0	0
0	1	1
1	0	1
1	1	0

IA sans/avec le Machine Learning : l'exemple de la fonction XOR

a	b	a XOR b
0	0	0
0	1	1
1	0	1
1	1	0



IA **sans/avec** le Machine Learning : l'exemple de la fonction XOR

a	b	a XOR b
0	0	0
0	1	1
1	0	1
1	1	0

Implémentation **sans** le Machine Learning (en Python)

```
1 def xor(a, b):
2
3     if (a == 0 and b == 0) or (a == 1 and b == 1):
4         return 0
5     elif (a == 0 and b == 1) or (a == 1 and b == 0):
6         return 1
7     else:
8         return None
```

IA **sans/avec** le Machine Learning : l'exemple de la fonction XOR

a	b	a XOR b
0	0	0
0	1	1
1	0	1
1	1	0

Implémentation **avec** le Machine Learning (en Python)

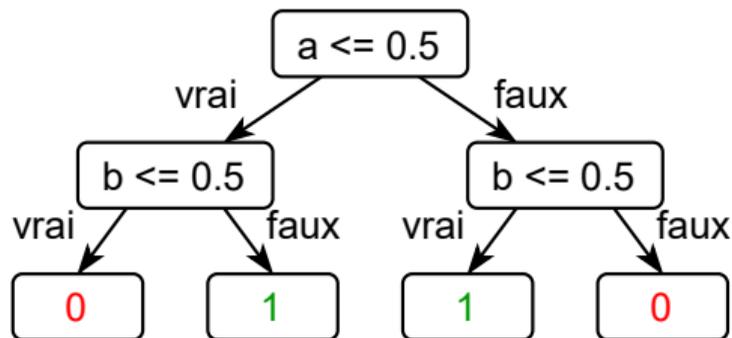
```
1 from sklearn import tree
2
3 def xor(a, b):
4
5     X = [[0, 0], [0, 1], [1, 0], [1, 1]]
6     Y = [0, 1, 1, 0]
7
8     classifieur = tree.DecisionTreeClassifier()
9     classifieur = classifieur.fit(X, Y)
10
11     return classifieur.predict([[a, b]])[0]
```

IA sans/avec le Machine Learning : l'exemple de la fonction XOR

a	b	a XOR b
0	0	0
0	1	1
1	0	1
1	1	0

Implémentation avec le Machine Learning

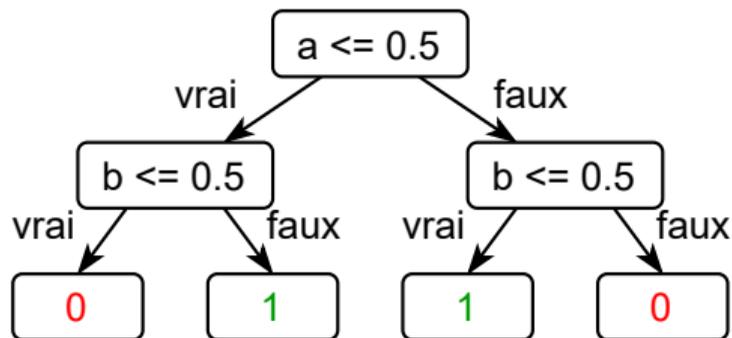
Arbre de décision obtenu par apprentissage automatique :



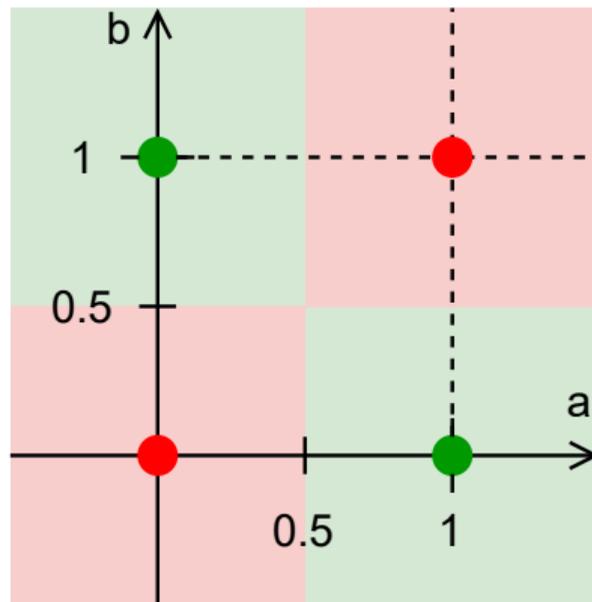
IA sans/avec le Machine Learning : l'exemple de la fonction XOR

Implémentation avec le Machine Learning

Arbre de décision obtenu par apprentissage automatique :



Surface de décision :



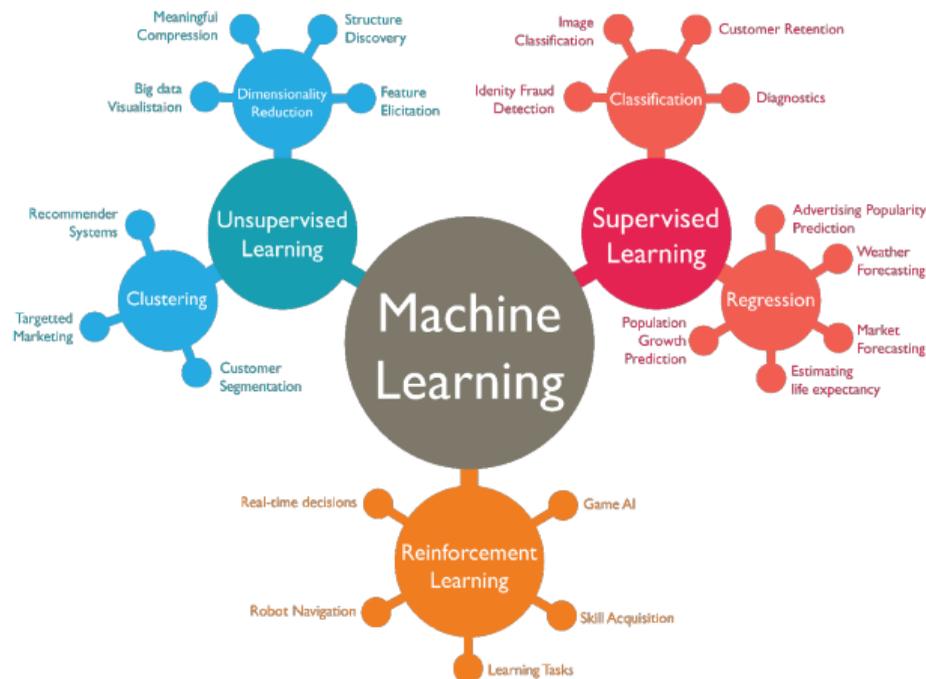
2. Machine Learning



«*Domaine d'étude qui donne aux ordinateurs la capacité
d'apprendre sans être explicitement programmés.*»

phrase attribuée à Arthur Lee Samuel (MIT, Bell Labs, IBM, . . .), 1959

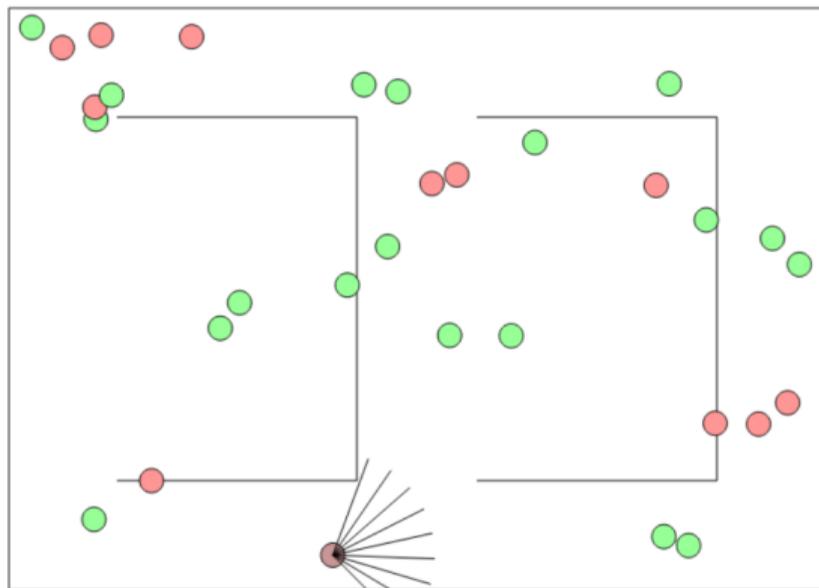
Machine Learning



Source : <https://www.wordstream.com/blog/ws/2017/07/28/machine-learning-applications>

Les principales méthodes de Machine Learning

Apprentissage par renforcement : démo



voir <https://cs.stanford.edu/people/karpathy/convnetjs/demo/rldemo.html>

Les principales méthodes de Machine Learning

Classification : exemple d'application



SPAM

vs.

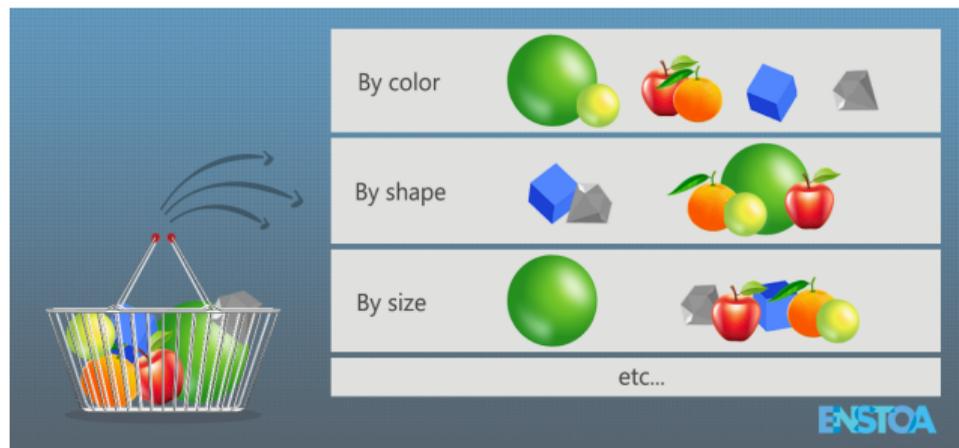


HAM

Source : <https://blog.codecentric.de/en/2016/06/spam-classification-using-sparks-dataframes-ml-zeppelin-part-1/>

Les principales méthodes de Machine Learning

Clustering : exemple d'application

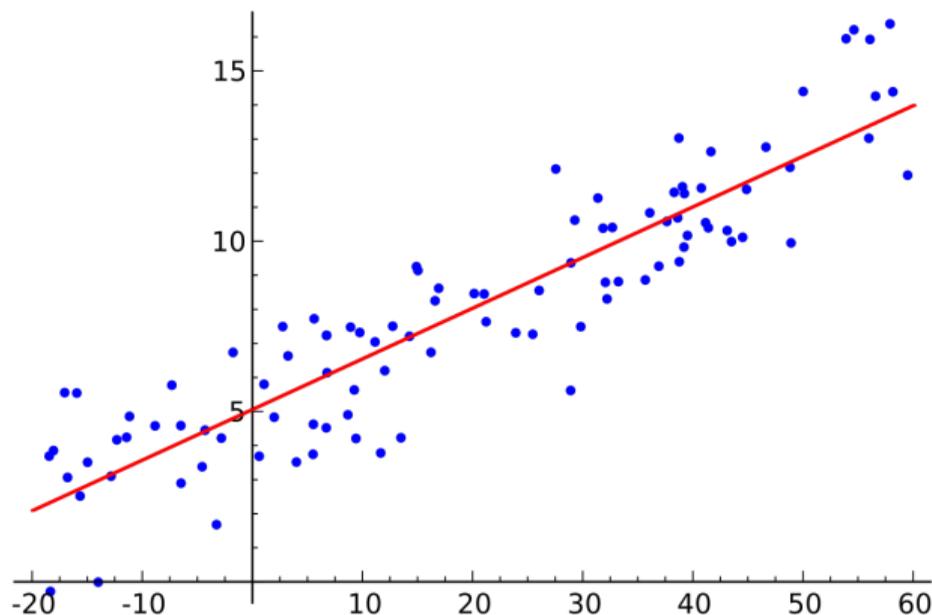


- Classification → étiquettes connues *a priori* (méthode supervisée)
- Clustering → étiquettes choisies *a posteriori* (méthode non-supervisée)

Source : <https://enstoa.com/blog/machine-learning-construction-how-clustering-data-can-improve-processes-part-2-of-2>

Les principales méthodes de Machine Learning

Régression



Source : https://en.wikipedia.org/wiki/Linear_regression



REPUBLIQUE
ET CANTON
DE GENEVE

POST TENEBRAS LUX

Département du territoire
Direction de l'information du territoire

06/11/2023 - Page 35

Régression : estimation vs prediction

Modèle de régression linéaire

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \varepsilon$$

X_1, X_2, \dots → variables indépendantes/explicatives

Y → variable dépendante/explicée

■ estimation (analyse descriptive)

données d'entraînement $\rightsquigarrow \beta_0, \beta_1, \beta_2, \dots$

■ prédiction (analyse prédictive)

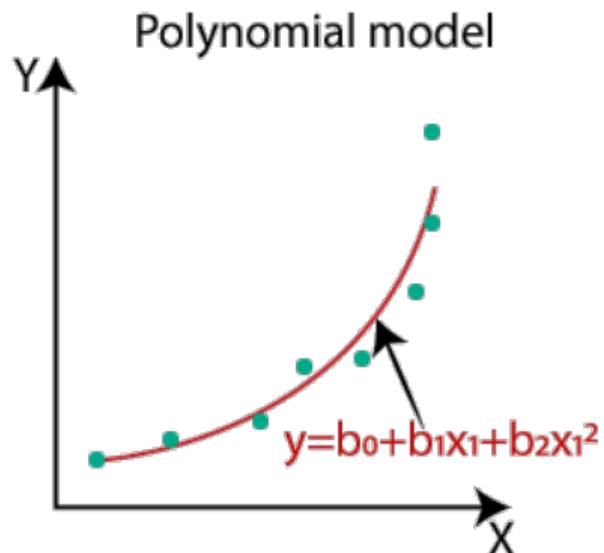
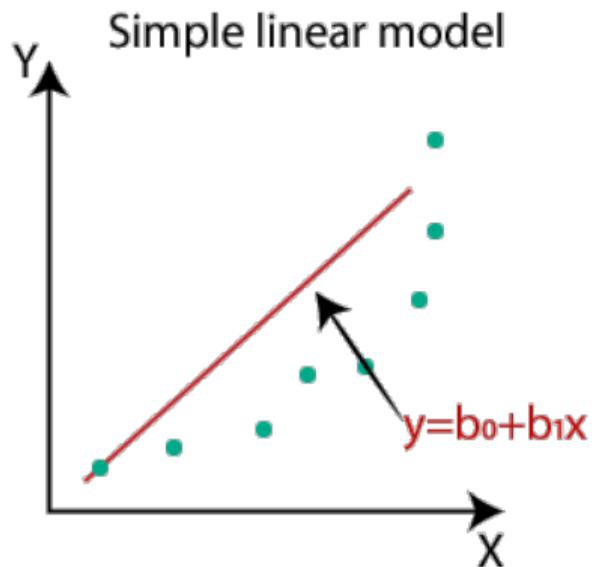
$\beta_0, \beta_1, \beta_2, \dots; X_1, X_2, \dots \notin$ données d'entraînement $\rightsquigarrow Y$

Source : <https://stats.stackexchange.com/questions/17773/what-is-the-difference-between-estimation-and-prediction#17789>



Les principales méthodes de Machine Learning

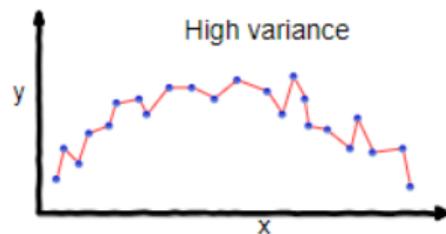
Régression



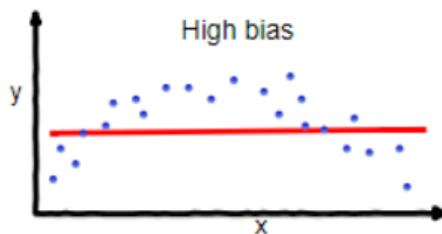
Source : <https://www.javatpoint.com/machine-learning-polynomial-regression>

Les principales méthodes de Machine Learning

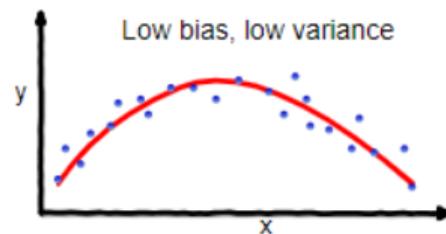
Régression : variance vs biais



overfitting



underfitting



Good balance

Source : <https://tinyurl.com/y5ulgmsm>, <https://tinyurl.com/y6poz6wm>

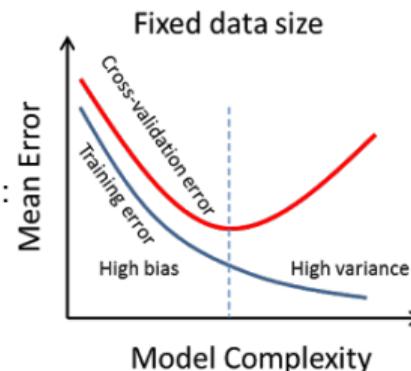
Variance vs biais

Comment éviter le piège ?

1 En découpant les données d'entraînement ; typiquement

- 70% → entraînement
- 15% → validation croisée
- 15% → test

2 En traçant les courbes d'erreur en fonction de différents paramètres, par ex. :

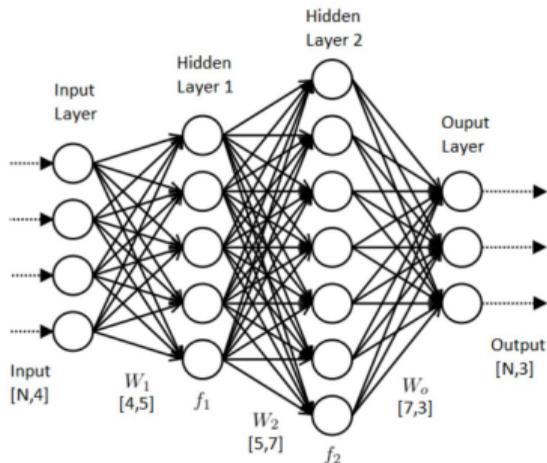


3 ...

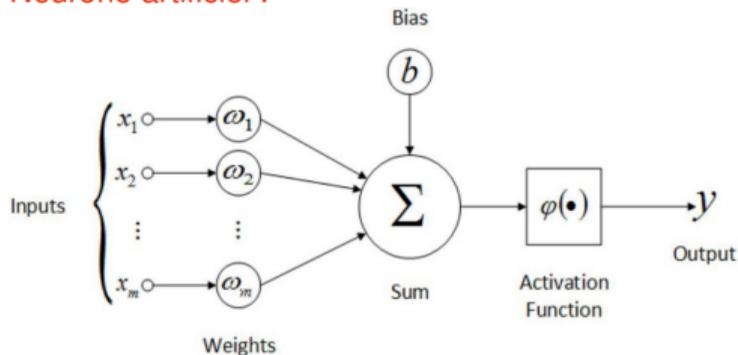
3. Réseaux de neurones, Deep Learning



Réseaux de neurones



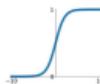
Neurone artificiel :



Fonctions d'activation :

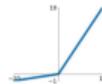
Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



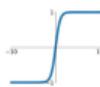
Leaky ReLU

$$\max(0.1x, x)$$



tanh

$$\tanh(x)$$



Maxout

$$\max(w_1^T x + b_1, w_2^T x + b_2)$$

ReLU

$$\max(0, x)$$



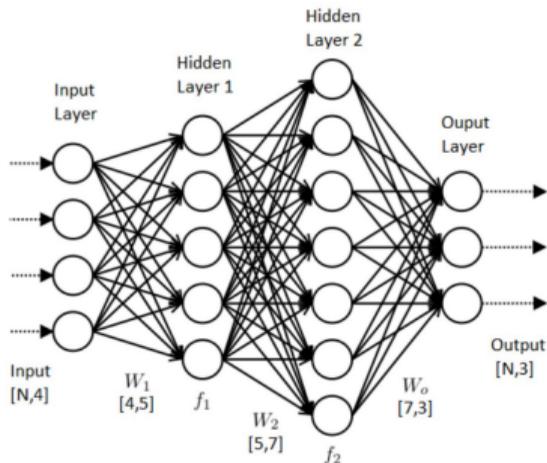
ELU

$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



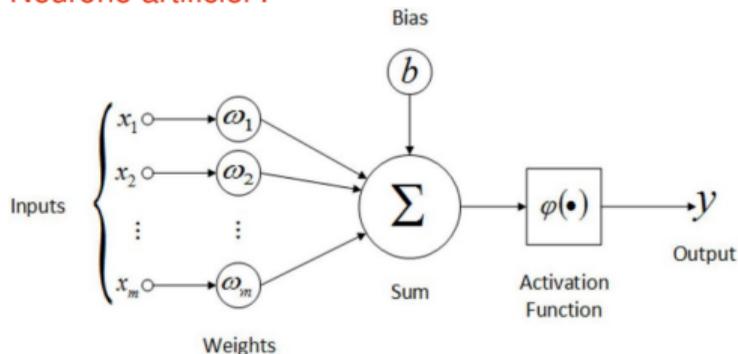
Sources : <https://tinyurl.com/447srfmc>, <https://tinyurl.com/hxmd6h8c>, <https://tinyurl.com/umah5s4b>

Réseaux de neurones



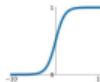
On parle de "Deep Learning" quand le réseau est profond, c'est-à-dire quand il est composé par un nb. élevé de couches.

Neurone artificiel :

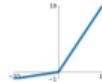


Fonctions d'activation :

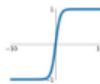
Sigmoid
 $\sigma(x) = \frac{1}{1+e^{-x}}$



Leaky ReLU
 $\max(0.1x, x)$



tanh
 $\tanh(x)$



Maxout
 $\max(w_1^T x + b_1, w_2^T x + b_2)$

ReLU
 $\max(0, x)$



ELU
 $\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$

Sources : <https://tinyurl.com/447srfmc>, <https://tinyurl.com/hxmd6h8c>, <https://tinyurl.com/umah5s4b>

IA sans/avec le Machine Learning : l'exemple de la fonction XOR

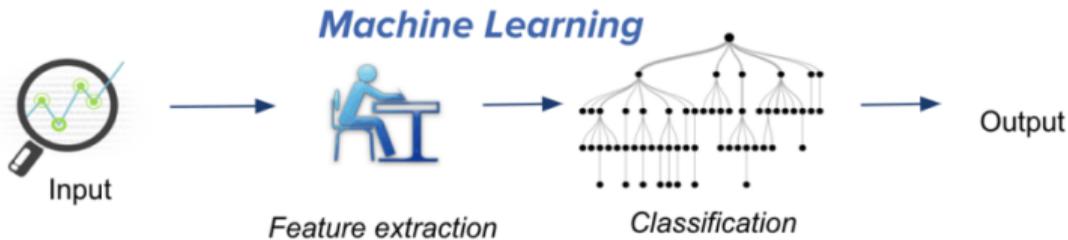
a	b	a XOR b
0	0	0
0	1	1
1	0	1
1	1	0

Implémentation avec un réseau de neurones (ML)

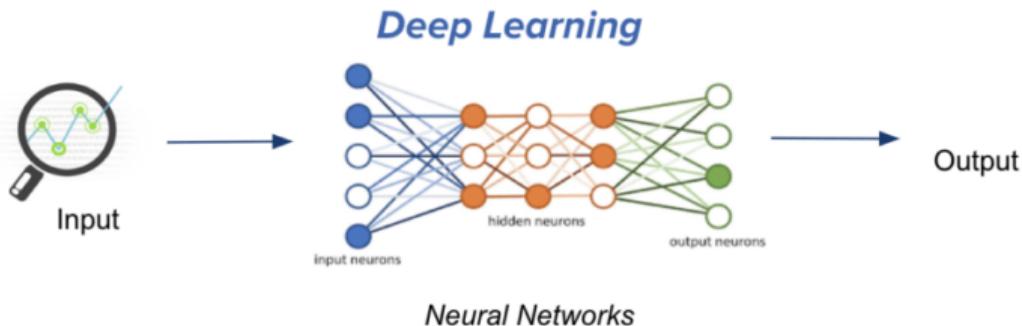
Démo interactive : <https://tinyurl.com/576svffb>

Voir <https://tinyurl.com/bvtd4s49>

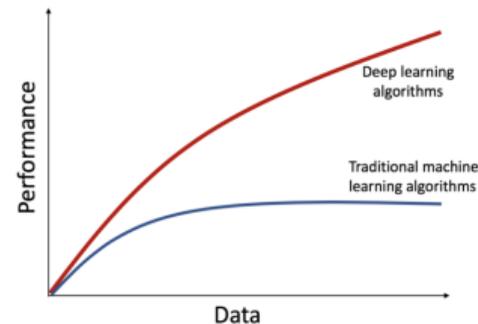




Traditional machine learning uses hand-crafted features, which is tedious and costly to develop.

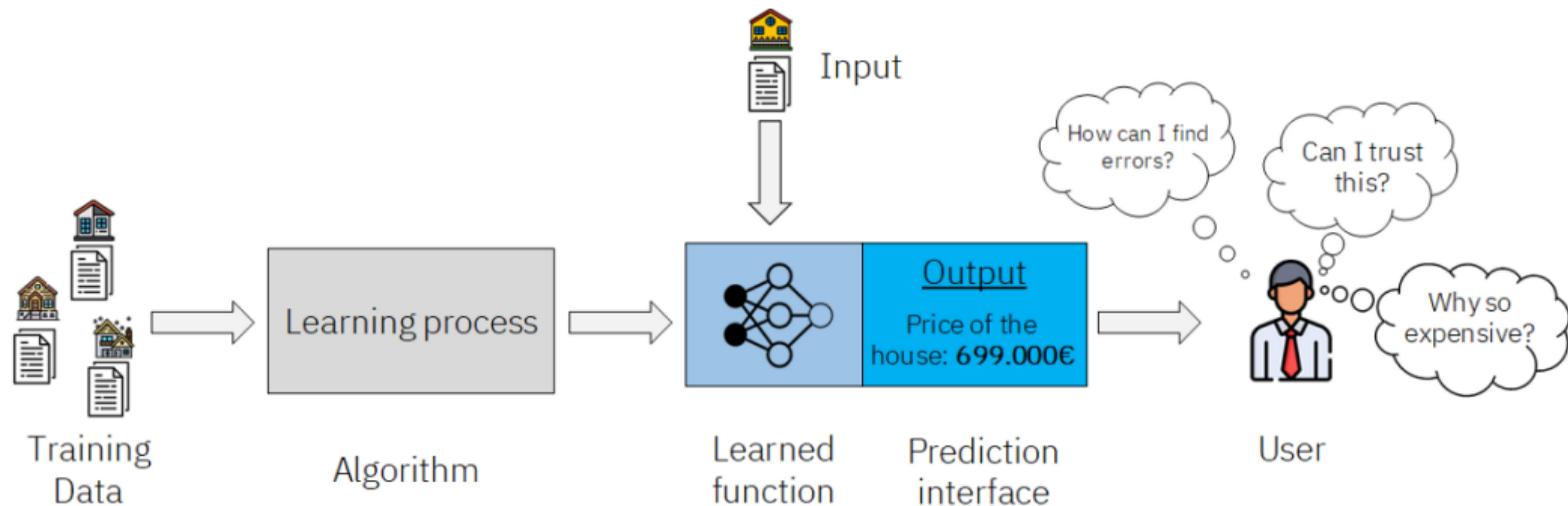


Deep learning learns hierarchical representation from the data itself, and scales with more data.



Source : <https://bluehexagon.ai/blog/what-is-deep-learning-and-how-is-it-different-from-machine-learning/>

XAI = eXplainable Artificial Intelligence



Source : <https://towardsdatascience.com/explainable-artificial-intelligence-14944563cc79>

Deep Learning

Quelques applications

- reconnaissance automatique du son et de la parole
- traitement automatique du langage naturel (NLP, ...) - par exemple ChatGPT
- robotique au sens large, ex. : jeux (échecs, Go - <https://tinyurl.com/qvkwao>, ...)
- “art” (?!) par ordinateur <https://github.com/jcjohnson/neural-style>, <https://tinyurl.com/y85u36ue>
- “image super-resolution”, voir <https://github.com/idealo/image-super-resolution>
- **vision par ordinateur (images, vidéos, nuage de points)**
 - détection d'objets
 - conduite autonome
 - ...
- ...

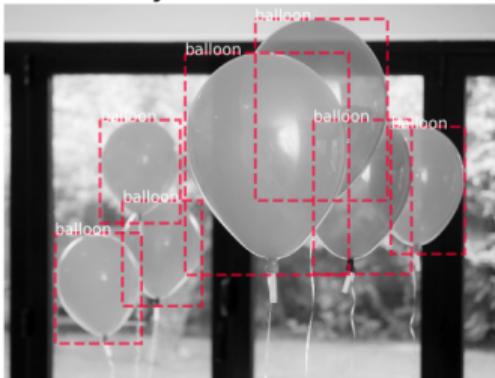
Classification



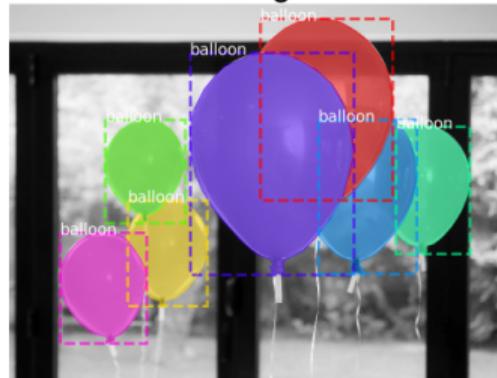
Semantic Segmentation



Object Detection

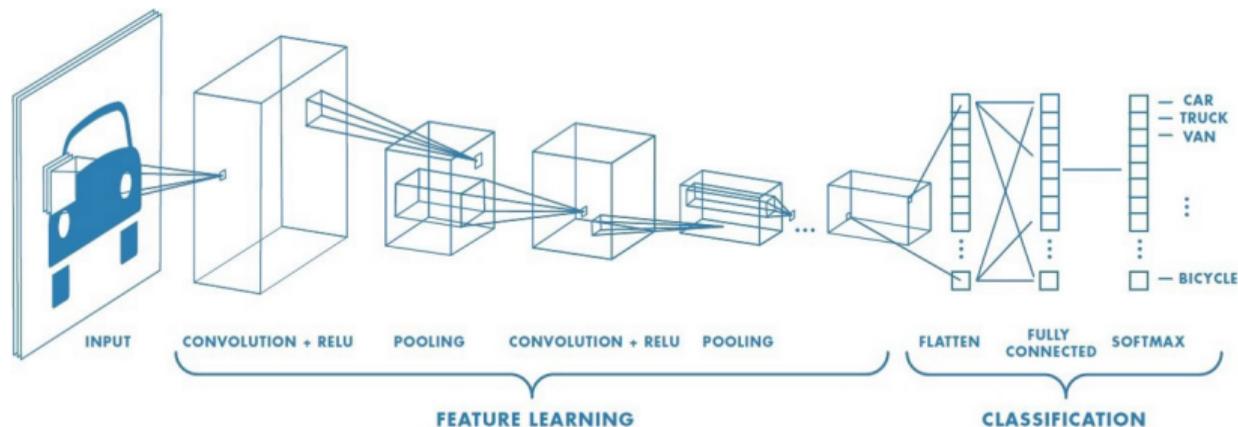


Instance Segmentation



Deep Learning

Réseaux de neurones convolutifs



Pour plus d'infos : <https://tinyurl.com/knt8p8m2>, <https://tinyurl.com/3t6xx29j>

Démo : https://adamharley.com/nn_vis/, <https://poloclub.github.io/cnn-explainer/>

Source : <https://tinyurl.com/y9mmosuq>

Deep Learning

Vision Transformers (ViT)

- 2017 “Attention is all you need”,
A. Vaswani *et al.* (Google)
- 2020 “An image is worth 16x16 words : Transformers for image recognition at scale”,
A. Dosovitskiy *et al.* (Google)

Transformers have become the model of choice in NLP due to their computational efficiency and scalability. In computer vision, convolutional neural network (CNN) architectures remain dominant, but some researchers have tried combining CNNs with self-attention. The authors experimented with applying a standard Transformer directly to images and found that when trained on mid-sized datasets, the models had modest accuracy compared to ResNet-like architectures. However, when trained on **larger datasets**, the Vision Transformer (ViT) achieved excellent results and approached or surpassed the state of the art on multiple image recognition benchmarks.

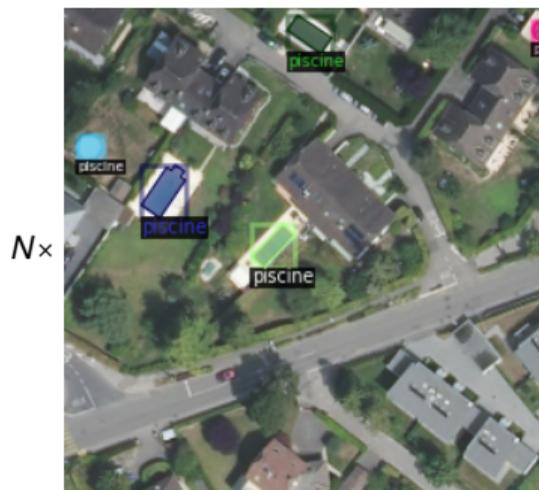
Source : <https://tinyurl.com/yrs9dbrf>



Deep Learning

Application à l'Etat de Genève : **détection de piscines à partir de photos aériennes**

apprentissage



prédiction



Pour plus d'informations, voir <https://tech.stdl.ch/PROJ-GEPOOL/>

La Data Science à l'Etat de Genève

Détection de piscines à partir de photos aériennes

Préambule

- Les piscines fixées durablement au sol doivent être cadastrées.

Questions métiers

- 1 Les **informations cadastrales** sur les piscines, sont-elles à jour ?
- 2 Y a-t-il, sur le canton, des **piscines cadastrables non déclarées** ?

N.B. : Le cadastre des piscines est publié en Open Data, [🔗 https://ge.ch/sitg/fiche/1836](https://ge.ch/sitg/fiche/1836).

Réponse basée sur la Data Science

- 1 Détection à partir de photos aériennes, en utilisant le **Deep Learning**

N.B. : Des campagnes d'acquisition de photos aériennes sont menées avec une certaine régularité (2019, 2018, 2016, 2015, ...). Les données sont publiées en Open Data, ex. : [🔗 https://ge.ch/sitg/fiche/3137](https://ge.ch/sitg/fiche/3137).

La Data Science à l'Etat de Genève

Détection de piscines à partir de photos aériennes : **les principales étapes**

- 1 **Préparation** des données d'entraînement (~ 300 Mo)
- 2 **Entraînement** d'un réseau de neurones artificiel
(optimisation des paramètres du réseau)
- 3 **Evaluation qualitative et quantitative** des prédictions
- 4 Ajustement des “**hyper-paramètres**” du réseau
- 5 **Réitérations** des étapes 2, 3, 4
- 6 Enfin : **détection** sur tout le territoire, **livraison** des résultats, etc.

Pour plus d'informations, voir [🔗 https://tech.stdl.ch/PROJ-GEPOOL/](https://tech.stdl.ch/PROJ-GEPOOL/). Pour le code, voir [🔗 https://github.com/swiss-territorial-data-lab/object-detector](https://github.com/swiss-territorial-data-lab/object-detector).



La Data Science à l'Etat de Genève

Détection de piscines à partir de photos aériennes : **les principales étapes**

1. Préparation des données

- 1 Définition d'une **zone d'intérêt** :
surface couverte par les parcelles cadastrales, sans le Lac Léman
- 2 Définition d'une **stratégie de maillage** de cette zone : "Slippy Map Tiles"^a
- 3 Export des **images aériennes sous forme de tuiles** :
images de 256x256 pixels, géo-référencées (format GeoTIFF)
- 4 Division des images en **trois sous-groupes** :
 - 1 entraînement - 70%
 - 2 validation croisée - 15%
 - 3 test - 15%
- 5 Génération des "**masques de segmentation**"

a voir <https://developers.planet.com/tutorials/slippy-maps-101/>

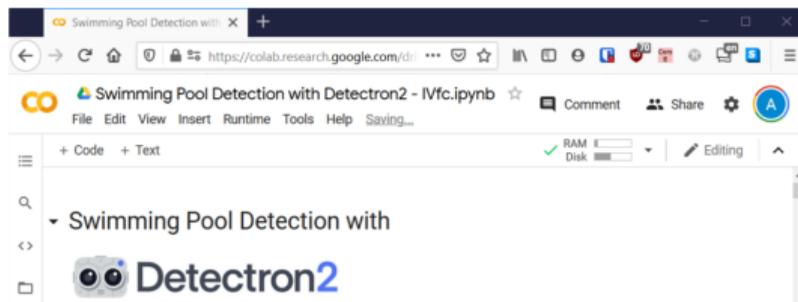


La Data Science à l'Etat de Genève

Détection de piscines à partir de photos aériennes : **les principales étapes**

2. Entraînement d'un réseau de neurones

- Utilisation de Detectron2^a en mode “boîte noire” (ou plutôt “grise”^b ?!)



^a Bibliothèque pour le langage Python développée par Facebook et distribuée en Open Source, implémentant plusieurs architectures de réseaux de neurones à l'état de l'art. Voir <https://github.com/facebookresearch/detectron2>.

^b voir <https://tinyurl.com/yxqutgd6>

La Data Science à l'Etat de Genève

Détection de piscines à partir de photos aériennes : **les principales étapes**

3. Evaluation **qualitative** et **quantitative** des prédictions

- 1 Géoréférencement des prédictions
- 2 Comparaison avec la “vérité terrain”
 - vrais positifs
 - **faux positifs : justes ? erronés ?**
 - (vrais négatifs)
 - **faux négatifs : justes ? erronés ?**
- 3 Calcul de **métriques** : précision, rappel, f1-score
voir https://en.wikipedia.org/wiki/Precision_and_recall

La Data Science à l'Etat de Genève

Détection de piscines à partir de photos aériennes : **les principales étapes**

3. Evaluation **qualitative** et **quantitative** des prédictions

faux positif, mais... probablement vrai ?



faux négatif, mais... vraisemblablement vrai !



La Data Science à l'Etat de Genève

Détection de piscines à partir de photos aériennes : **les principales étapes**

3. Evaluation **qualitative** et **quantitative** des prédictions

faux positif, vraiment faux



faux négatif, vraiment faux

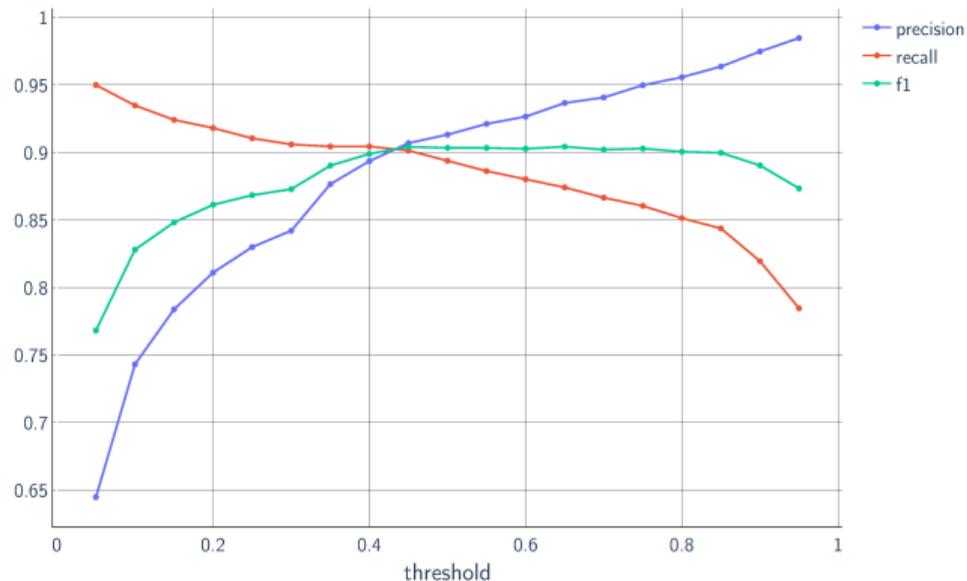


Détection par Deep Learning

Les principales étapes : 3. Evaluation qualitative et quantitative des prédictions

Dépendance du choix du **seuil d'acceptabilité** : $5\% \leq \text{threshold} \leq 95\%$.

Exemple sur le jeu de validation croisée :



Détection par Deep Learning

Les principales étapes : [6. Détection sur tout le Canton, livraison des résultats, retour utilisateur, ...](#)

threshold = 0.97 \Rightarrow 562 nouveaux objets sont détectés, dont :

- 128 "autres objets"
- 211 piscines pas cadastrables
- 223 piscines cadastrables



Résumé

1 Qu'est-ce que la Data Science ?

- relation avec d'autres disciplines
- relation avec l'IA et le Big Data

2 Machine Learning

- différence par rapport à des méthode SANS apprentissage
- ⚠ variance vs biais

3 Réseaux de neurones, Deep Learning

- différences par rapport au Machine Learning sans réseaux de neurones
- ⚠ "Explainable AI"; importance de l'expertise métier
- vaste champ d'application, y compris sur les données territoriales

Merci pour votre attention, et...



REPUBLIQUE
ET CANTON
DE GENEVE

POSI TENERIAS LUX

Département du territoire
Direction de l'information du territoire

06/11/2023 - Page 61

...attention aux pièges !



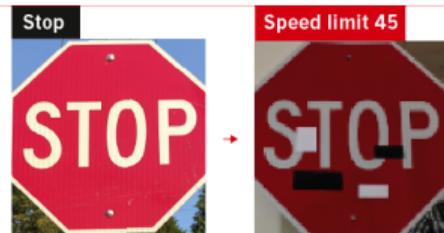
Mot-clé : réseaux antagonistes génératifs, voir <https://tinyurl.com/rauza45a>

Sources : <https://tinyurl.com/3de8kam3>, <https://tinyurl.com/2fuxvxy6>

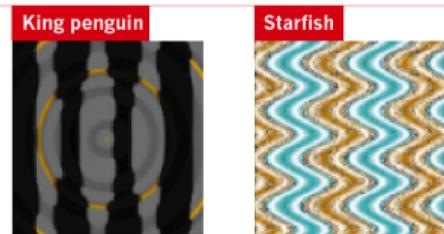
FOOLING THE AI

Deep neural networks (DNNs) are brilliant at image recognition — but they can be easily hacked.

These stickers made an artificial-intelligence system read this stop sign as 'speed limit 45'.



Scientists have evolved images that look like abstract patterns — but which DNNs see as familiar objects.



©nature

Mot-clé : "adversarial patch", voir <https://arxiv.org/abs/1712.09665>