L'utilisation de l'apprentissage profond dans l'analyse des séquences génomiques

Etienne Lord Ph.D
Chercheur en agronomie numérique
Agriculture et Agroalimentaire Canada -St-Jean-sur-Richelieu
Février 2021

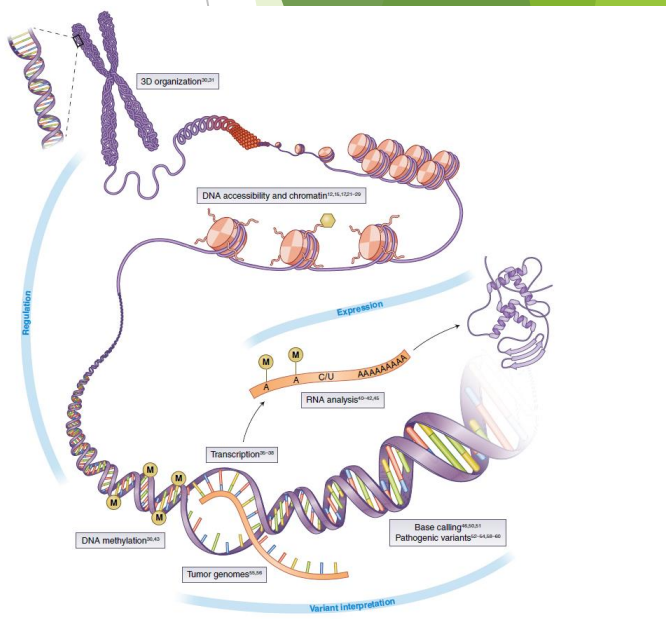
Agriculture and Agri-Food Canada | Agriculture et Agroalimentaire Canada

Buts de la présentation

- Démystifier l'apprentissage profond (*Deep Learning*) pour l'analyse de séquences génomiques ou protéiques.
- La création de modèles d'apprentissage profond avec Keras et R.

- Le code source requis pour suivre la présentation se retrouve sur le site:

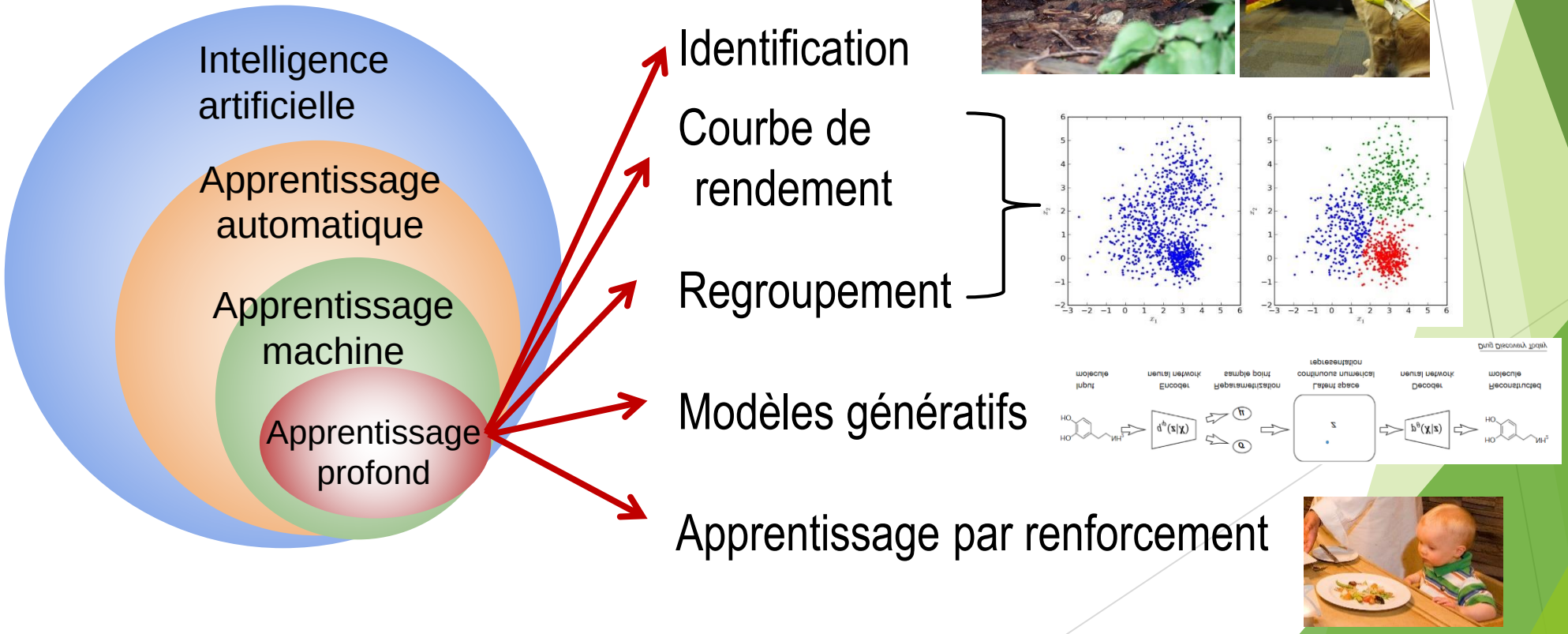
https://github.com/armadilloUQAM/BIF7002



Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., & Telenti, A. (2018). A primer on deep learning in genomics. *Nature genetics*, 1.

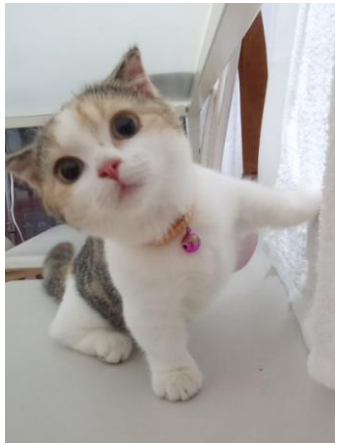
Apprentissage profond en bref

- Une science des **données** (*i.e. requière tous les cas possibles*)
- Une science de la **classification et des modèles**



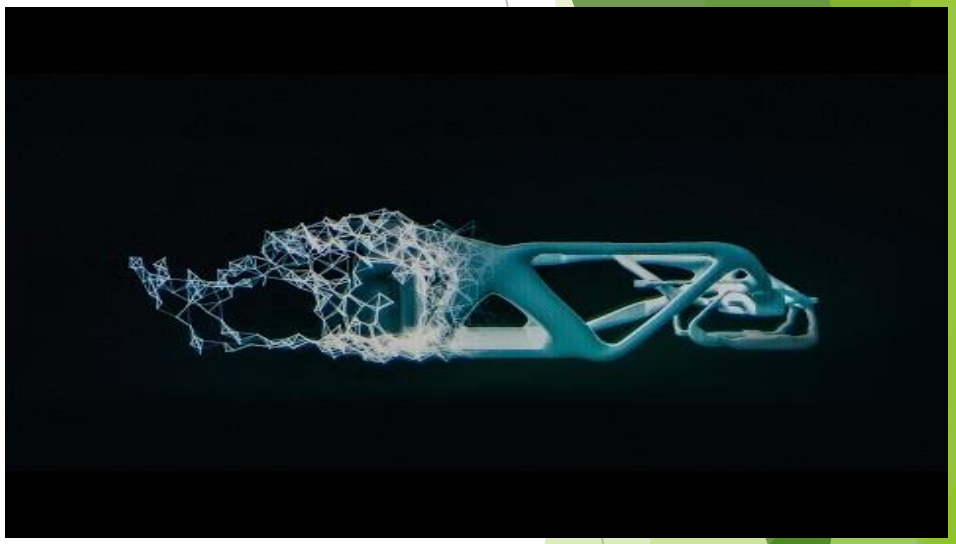
Apprentissage naturel

- Capacité à **stocker et à réagir à partir des expériences passées et à s'adapter aux nouvelles conditions.**
- C'est **une caractéristique des formes de vie naturelles.**
- Il est essentiel, au cours de la première étape de la vie, d'apprendre des faits fondamentaux : reconnaître sa voix, son visage, comprendre ce qui est dit, marcher, parler, ...



Apprentissage automatique

- Un programme informatique possède des capacités d'apprentissage si, tout en traitant des données représentatives, **il est capable de construire une représentation utile des données** afin de les utiliser ultérieurement.
- Il est ainsi capable de concevoir **un modèle*** pour la prédiction et la découverte de nouveaux faits



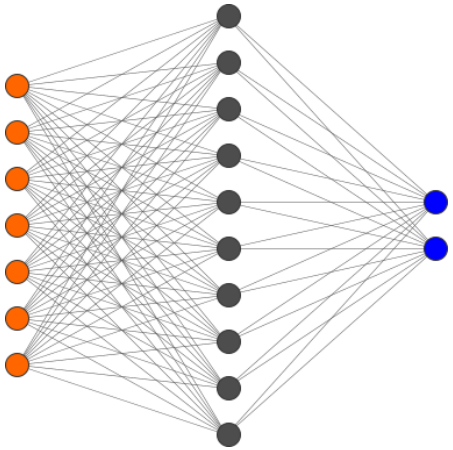
TED Talks: The incredible inventions of intuitive AI | Maurice Conti
https://youtu.be/aR5N2Jl8k14

*Un modèle est une description formelle des relations liant les données à tous les attributs qui les décrivent.

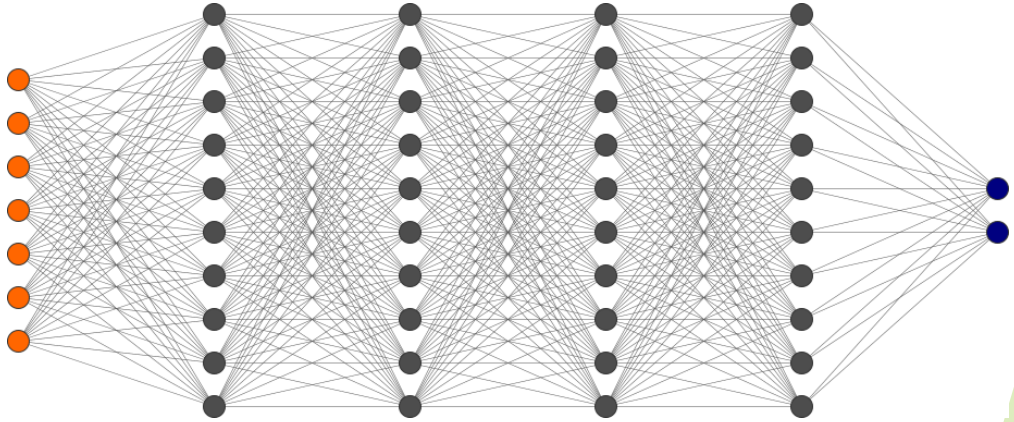
Apprentissage profond

- C'est **un apprentissage automatique** dans lequel **plusieurs couches** apprennent différentes caractéristiques de plus en plus **représentatives des données. – *François Chollet***

Réseaux de neurones simples



Apprentissage profond



Entrée des données

Traitement

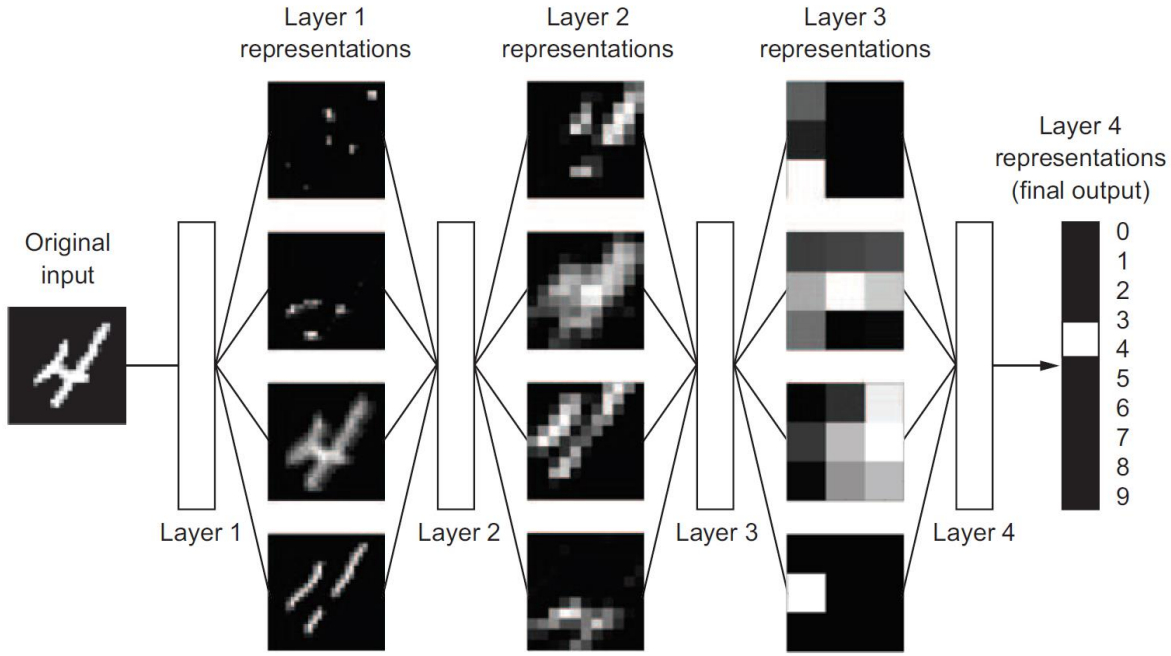
Sortie

https://www.deeplearningbook.org/

MNIST (Modified National Institute of Standards and Technology)

Ce jeu de données est composé de 10 chiffres écrits à la main

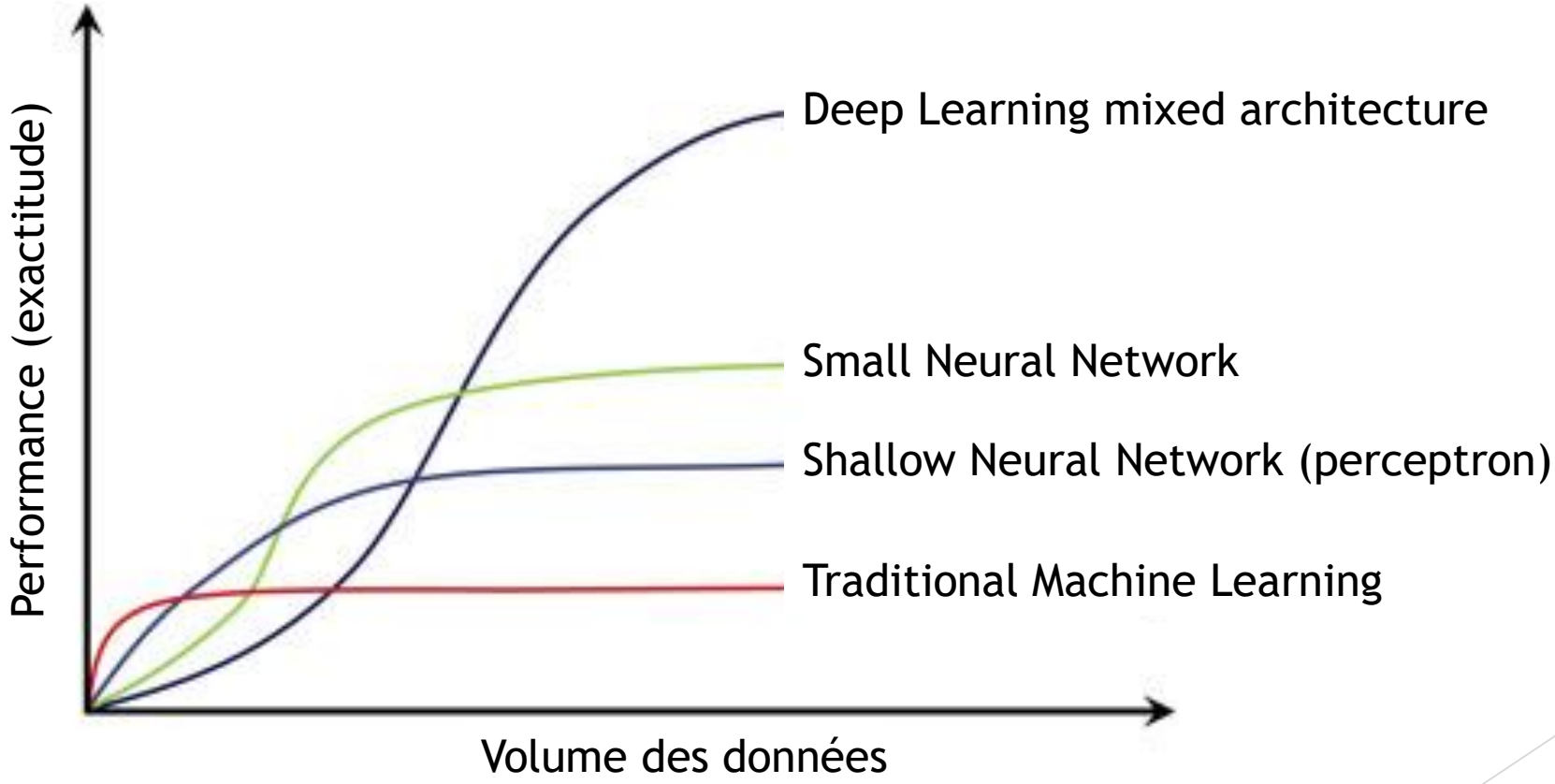
yann.lecun.com/exdb/mnist



Source: Deep Learning with Python, Francois Chollet

Apprentissage profond et méga-données

Est-ce que l'apprentissage profond est le bon outil ?

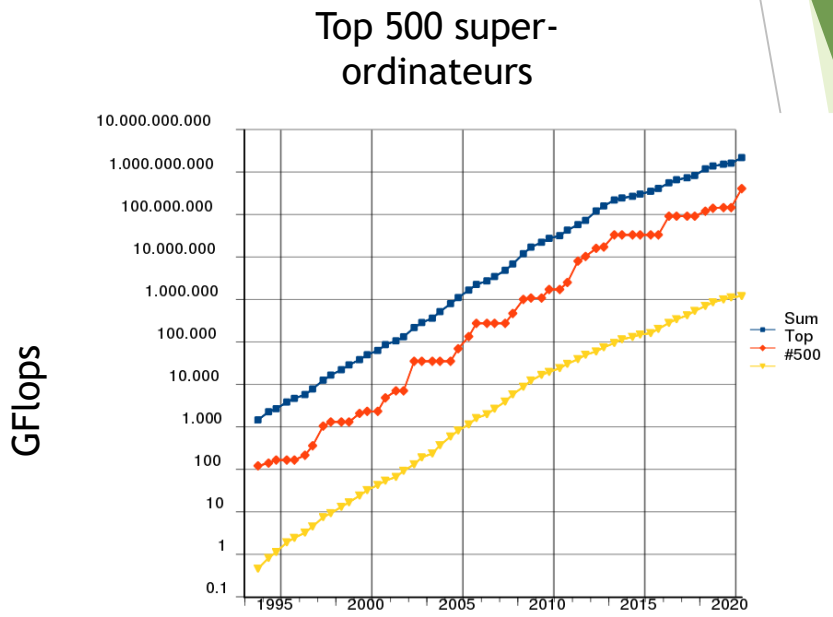


Tang, A., Tam, R., Cadrin-Chênevert, A., Guest, W., Chong, J., Barfett, J., ... & Poudrette, M. G. (2018). Canadian Association of Radiologists white paper on artificial intelligence in radiology. Canadian Association of Radiologists Journal, 69(2), 120-135.

Apprentissage profond

Pourquoi maintenant ?

1943 – **Mathématique des réseaux de neurones** - Walter Pitts & Warren McCulloch
1950 – Prédiction de l'apprentissage machine - Alan Turing
1959 – Découverte des cellules simples et complexes - David H. Hubel & Torsten Wiesel
...
1979-80 – Un réseau de neurones identifie des patrons - Kunihiko Fukushima
1986 – Amélioration de la rétro-propagation par David Rumelhart, Geoffrey Hinton, & Ronald J. Williams
1989 – **Identification de chiffres écrits à la main - Yann LeCun**
1993 – Une tâche 'very deep learning' est résolue - Jürgen Schmidhuber
1995 – Support vector machines - Corinna Cortes & Vladimir Vapnik
1**997 – Long short-term memory** - Schmidhuber & Sepp Hochreiter
1998 – Gradient-based learning - Yann LeCun
2009 – ImageNet - Fei Li
2014 – Generative Adversarial Networks (GAN) - Ian Goodfellow et co-auteurs



1995 → 2018
4 mois → 1 seconde
10 millénaires → 1 nuit

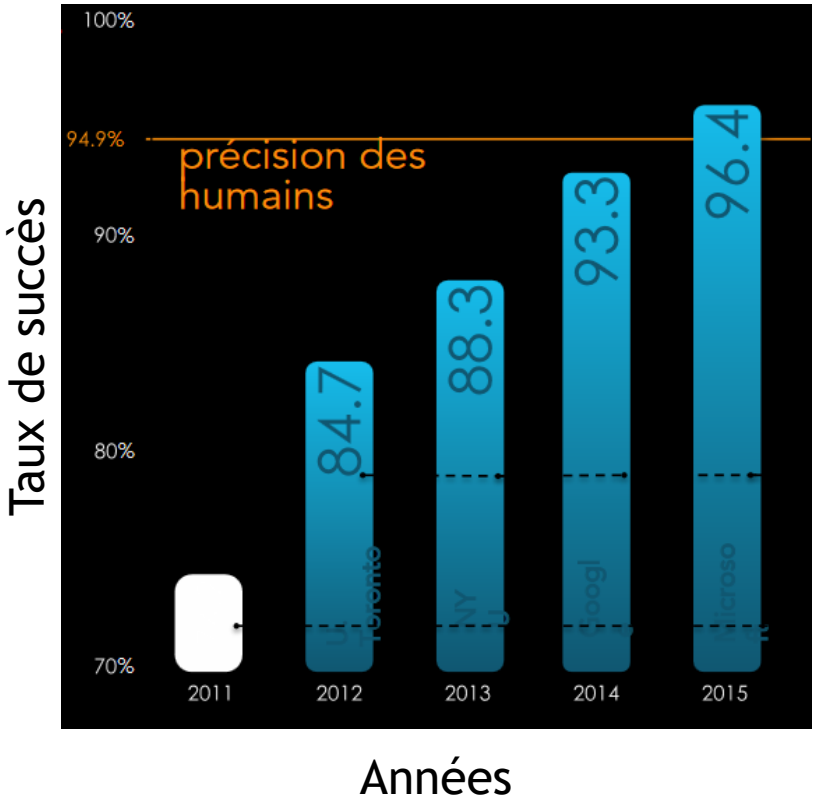
Nvidia RTX 3080
29.7 *tera*-FLOPS
900 $

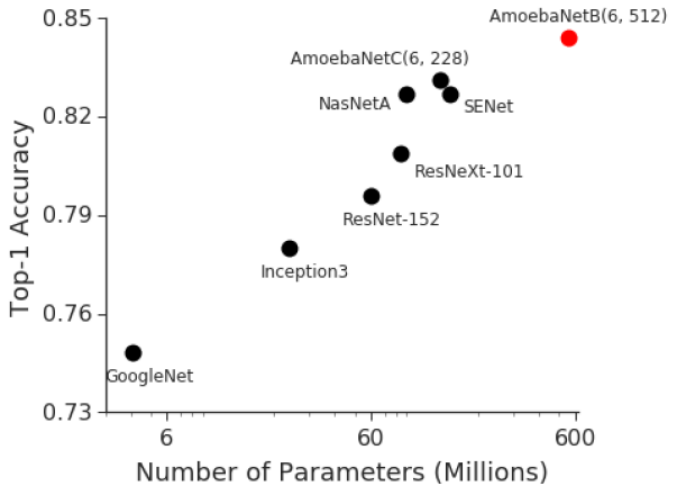
Source: Alain Tapp / IVADO - MILA 2018

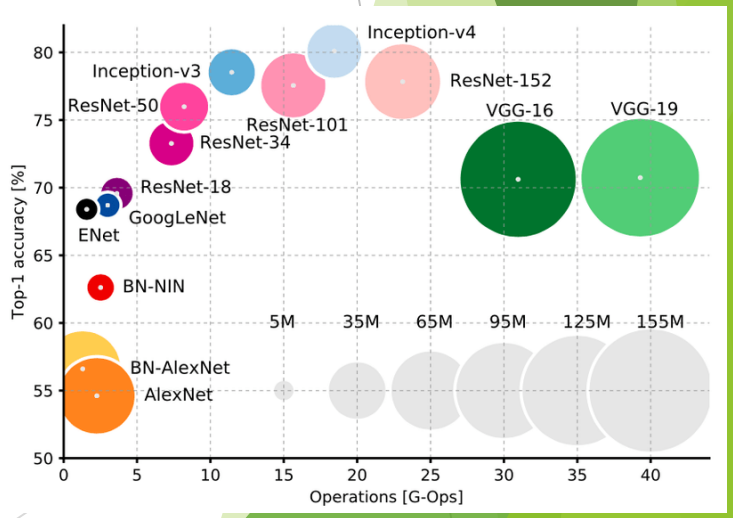
Date importante : 2015

Quand la machine dépasse l'homme dans la reconnaissance d'images

ImageNet: Catégorisation en 1000 classes, Top 5





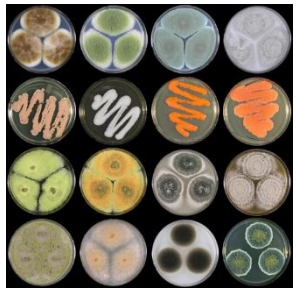


Canziani A, Paszke A, Culurciello E. An analysis of deep neural network models for practical applications. arXiv preprint arXiv:1605.07678. 2016 May 24.

Apprentissage profond en agriculture

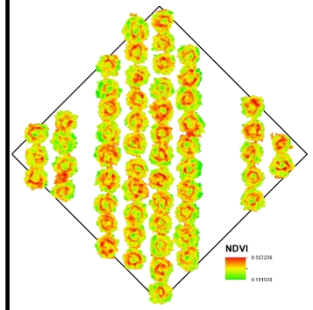
Différentes application de la reconnaissance d'images

Identification



e.g insectes, bactéries

Classification



e.g distribution de la taille des laitues

Quantification

e.g sévérité d'une maladie

Prédictions

Données météo →

Images → **Rendement**
Maladies

Données connexes →

Robots

- Détection automatique des rangs
- Reconstruction de la carte du terrain
- Désherbage automatique
- Application automatique de pesticides / herbicides

Singh, A., Ganapathysubramanian, B., Singh, A. K., & Sarkar, S. (2016). Machine learning for high-throughput stress phenotyping in plants. Trends in plant science, 21(2), 110-124.

Date importante : 2016-2017

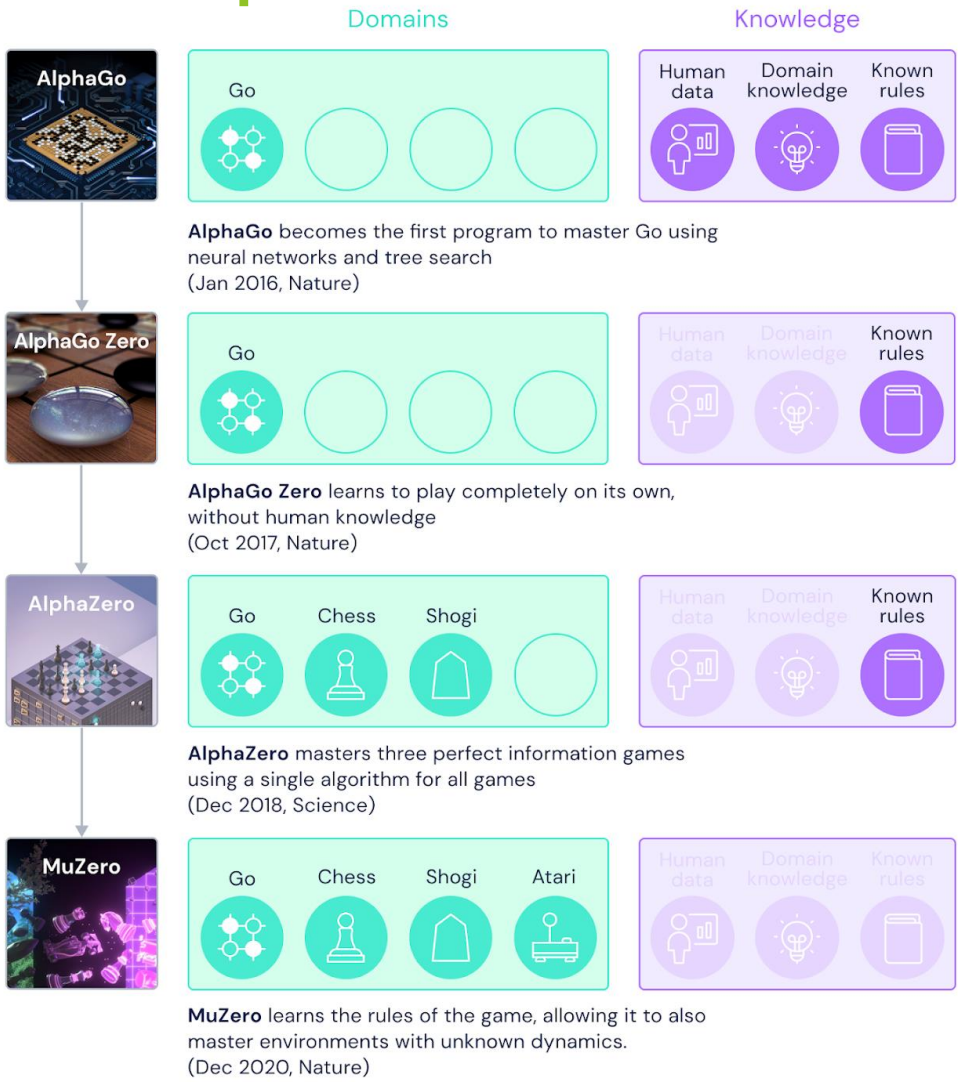
Quand la machine dépasse l'homme dans le jeu de Go

Ke Jie a affronté le programme de Google AlphaGo - sans succès.

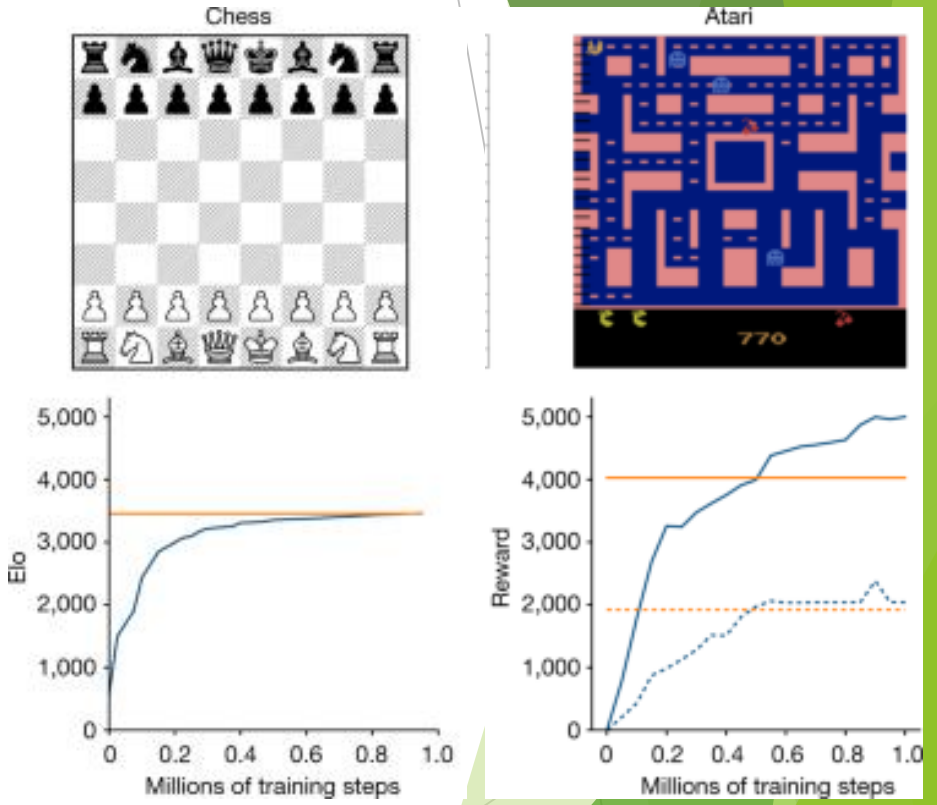
https://www.lapresse.ca/techno/actualites/201705/25/01-5101273-un-ordinateur-bat-le-meilleur-joueur-chinois-de-go.php

DeepMind (2016 à aujourd'hui)

Apprendre à partir de zéro



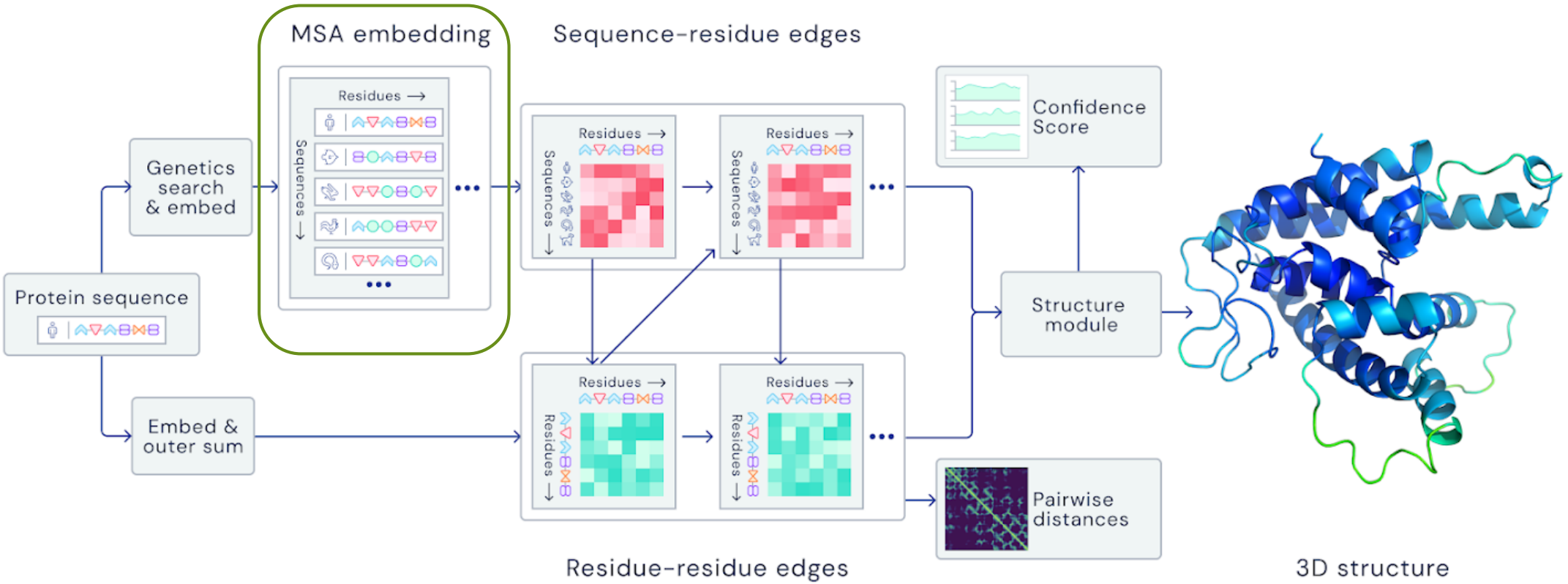
deepmind.com/blog/article/muzero-mastering-go-chess-shogi-and-atari-without-rules



Schrittwieser, Julian, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez et al. "Mastering atari, go, chess and shogi by planning with a learned model." *Nature* 588, no. 7839 (2020): 604-609.

AlphaFold (2018-2020)

Une révolution en bioinformatique

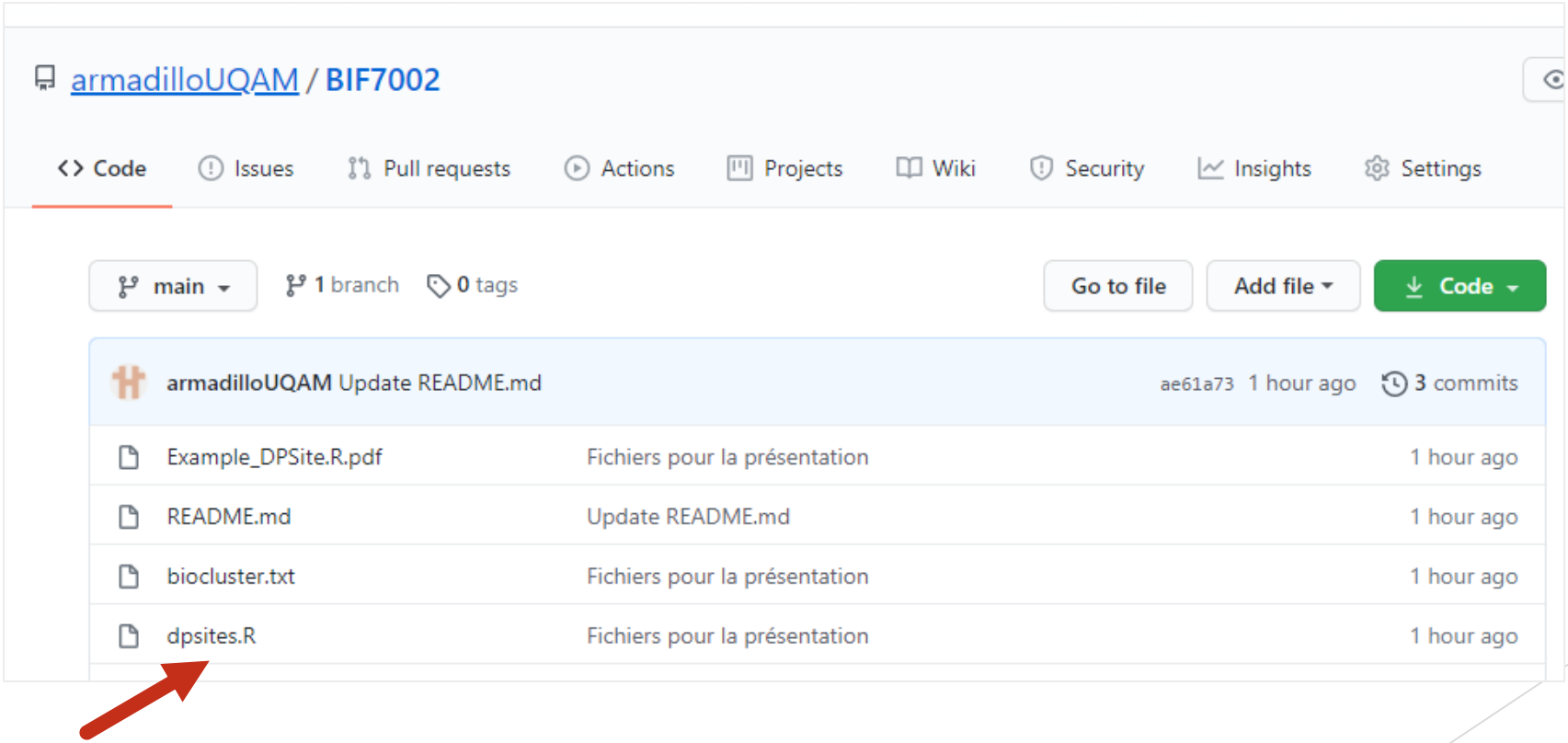


https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

Apprendre à utiliser l'apprentissage profond

(Télécharger maintenant le fichier dpsites.R et l'ouvrir)

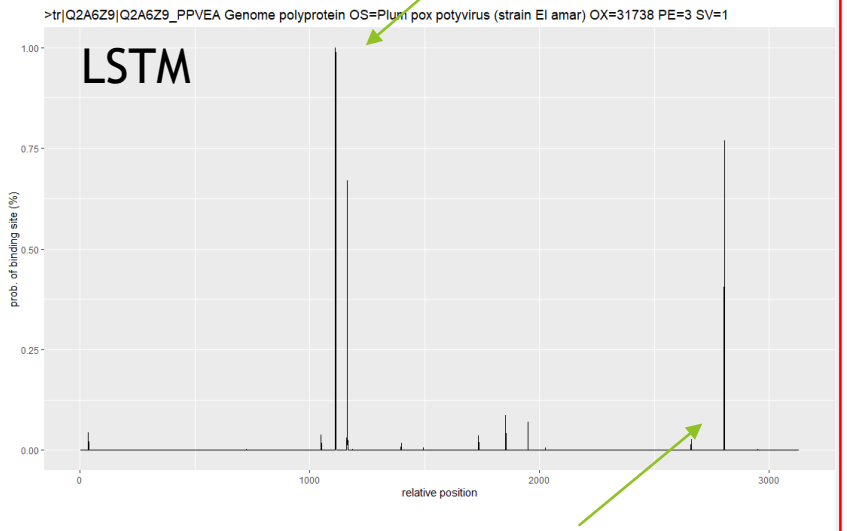
https://github.com/armadilloUQAM/BIF7002

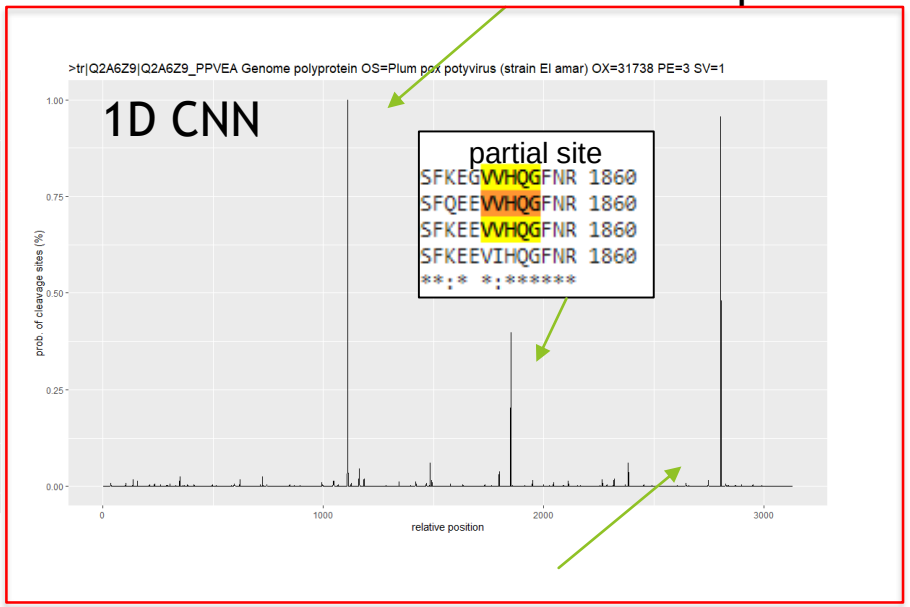


dpsites.R

```
tr|Q2A6Z9|Q2A6Z9_PPVEA          DTPTADEFLEYVEGENPDLSKYAADLIGEGQVVVHQSKRESQANLERVVAFVALVMMLFD 1140
sp|Q84934|POLG_PPVSK            CTPTADEFLEYVQGENPDLSKYAEDLIGDGQVVVHQSKRDSQANLERVAAFVALVMMLFD 1140
tr|A0A2Z5UG20|A0A2Z5UG20_PPVD   CTPTADEFLEYVGSENPDLLKHAEDLIGGWSVVVHQSKRDSQANLERVVAFVALVMMLFD 1140
sp|P17766|POLG_PPVNA            CTPTADEFLEYVGDENPDLLKHAEDLIGDGQVVVHQSKRDSQANLERVVAFVALVMMLFD 1140
                                 ********** .***** *:* **** .********:********.***********
```

DPSITES.R – Recherche de motifs dans les séquences





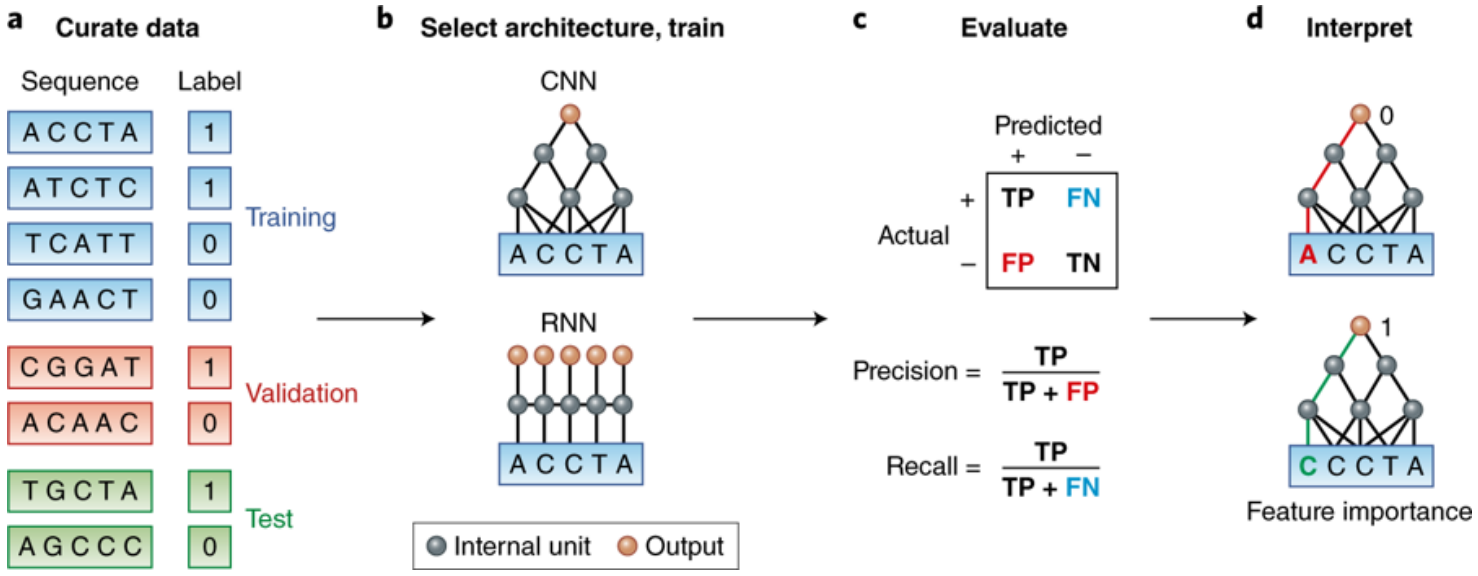

```
tr|Q2A6Z9|Q2A6Z9_PPVEA          KDGKAPYIAETALKKLYTDTEASETEIERYLEAFYSDLTDEDESNVVVHQADEKEDDEEE 2820
sp|Q84934|POLG_PPVSK            KDGKAPYIAETALRKLYTDSEASETEIERYLEAFYNDVDDSLDSNIVIHQADEEED-DEE 2819
tr|A0A2Z5UG20|A0A2Z5UG20_PPVD   KDGKAPYIAETALKKLYTDTEASETEIERYLEAFYNDINDDGESNVVVHQADERED-EEE 2819
sp|P17766|POLG_PPVNA            KDGKAPYIAETALKKLYTDTEASETEIERYLEAFYDNINDDGESNVVVHQADERED-EEE 2819
                                 ************:*****:**************.:: *. :**:*:*****.** :**
```

Largement inspiré de l'article de Zou (2018)

PERSPECTIVE
https://doi.org/10.1038/s41588-018-0295-5
nature genetics

A primer on deep learning in genomics

James Zou[1,2,3*], Mikael Huss[4,5], Abubakar Abid[3], Pejman Mohammadi[6,7], Ali Torkamani[6,7] and Amalio Telenti[6,7*]

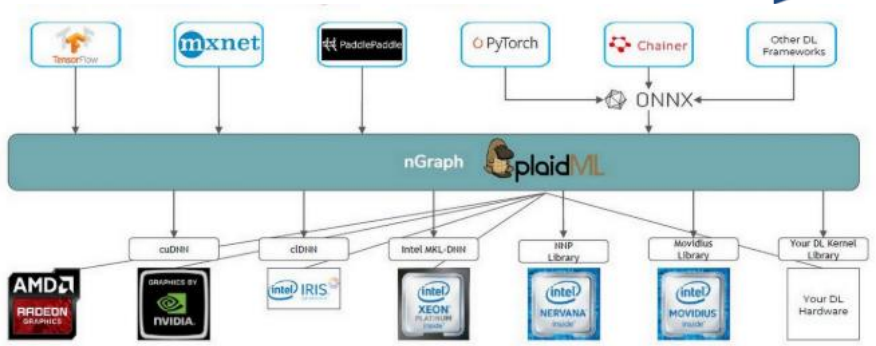


Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., & Telenti, A. (2018). A primer on deep learning in genomics. *Nature genetics*, 1.

Librairies pour l'apprentissage profond

Plusieurs compagnies ont leur propre plateformes

	Languages	Tutorials and training materials	CNN modeling capability	RNN modeling capability	Architecture: easy-to-use and modular front end	Speed	Multiple GPU support	Keras compatible
Theano	Python, C++	++	++	++	+	++	+	+
Tensor-Flow	Python	+++	+++	++	+++	++	++	+
Torch	Lua, Python (new)	+	+++	++	++	+++	++	
Caffe	C++	+	++		+	+	+	
MXNet	R, Python, Julia, Scala	++	++	+	++	++	+++	
Neon	Python	+	++	+	+	++	+	
CNTK	C++	+	+	+++	+	++	+	

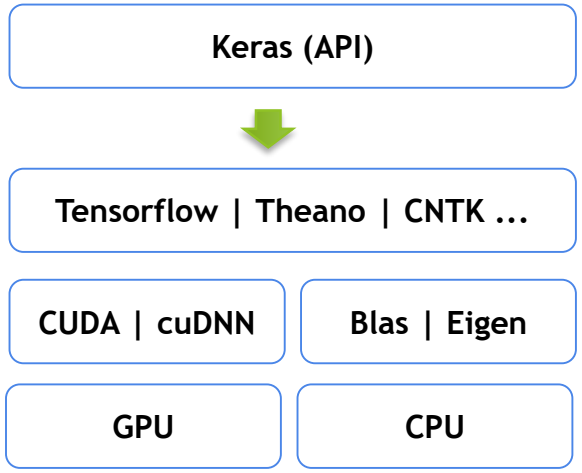


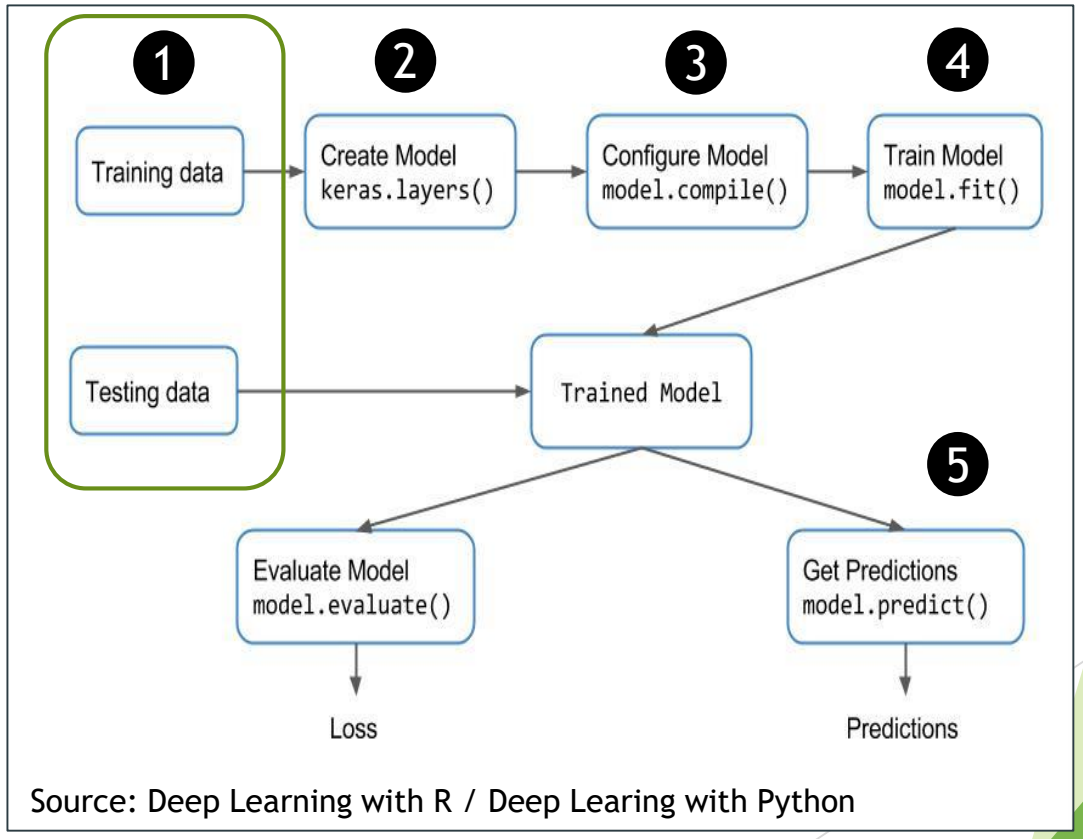
https://github.com/plaidml/plaidml

Gaétan Marceau Caron: École d'hiver MILA 2018
http://www.svds.com/getting-started-deep-learning/

Création de modèles avec Keras

5 étapes pour créer une classification de sites





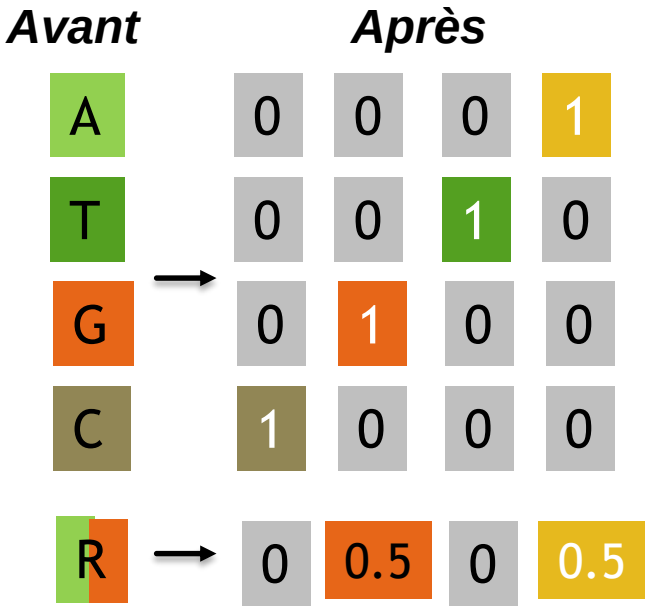
Source: Deep Learning with R / Deep Learing with Python

Encodage des données génomiques

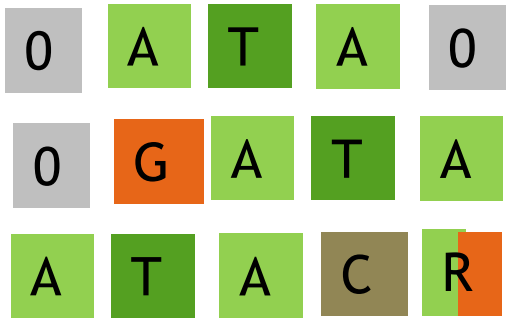
1

(ligne 47 / dpsites.R)

Encodage de l'ADN



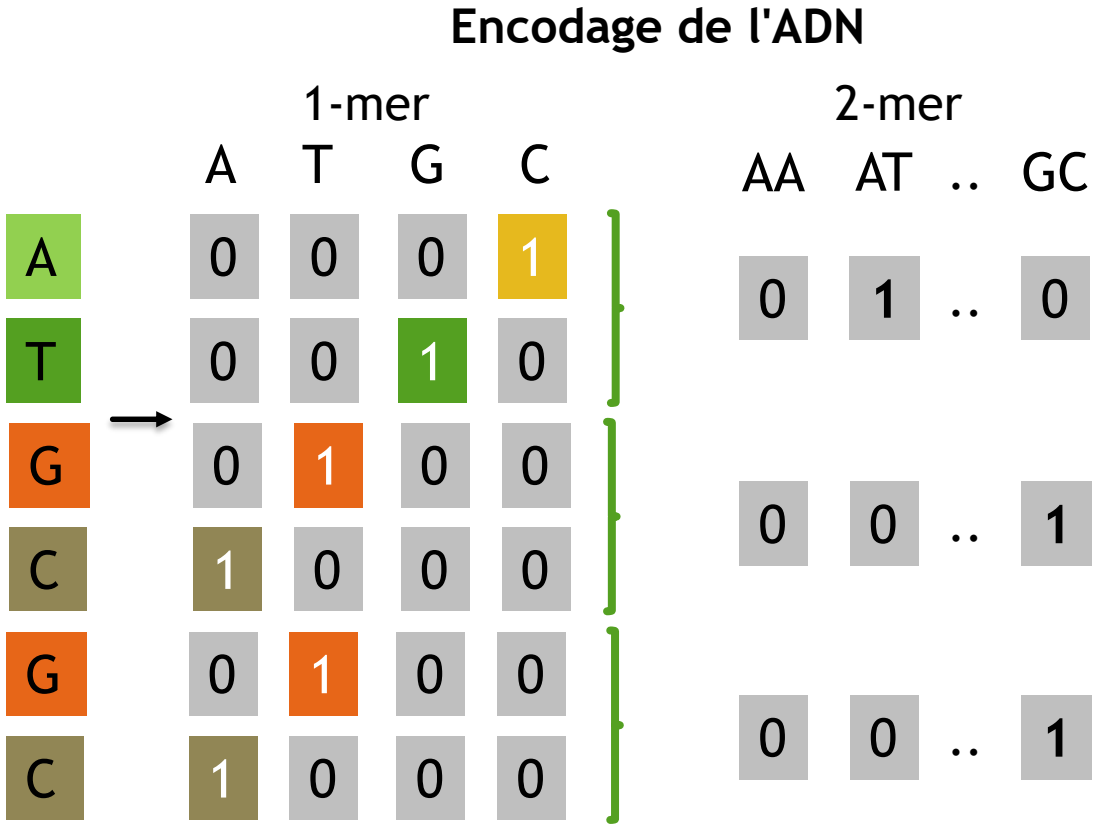
Ajouts de zéros si les séquences sont de différentes tailles

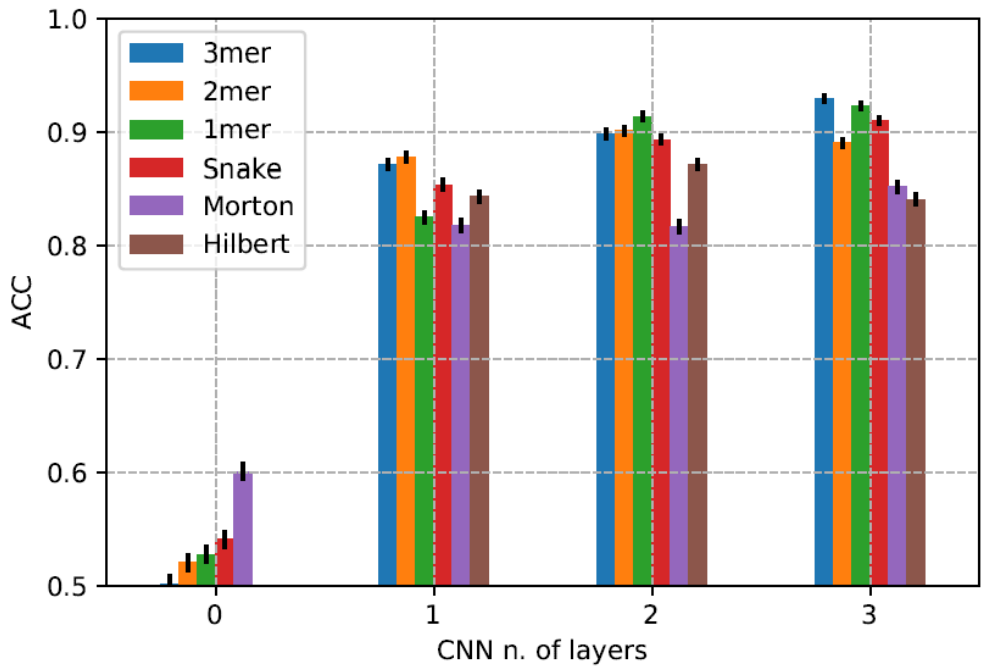


Lopez-del Rio, A., et al. (2020). Effect of sequence padding on the performance of deep learning models in archaeal protein functional prediction. *Scientific reports*, *10*(1), 1-14

Encodage des données génomiques

Aussi avec des *k*-mer





Noviello, T. M. R., Ceccarelli, F., Ceccarelli, M., & Cerulo, L. (2020). Deep learning predicts short non-coding RNA functions from only raw sequence data. *PLoS computational biology*, *16*(11), e1008415.

Modèles en apprentissage profond

Il y a un vocabulaire pour chaque types de couches

2

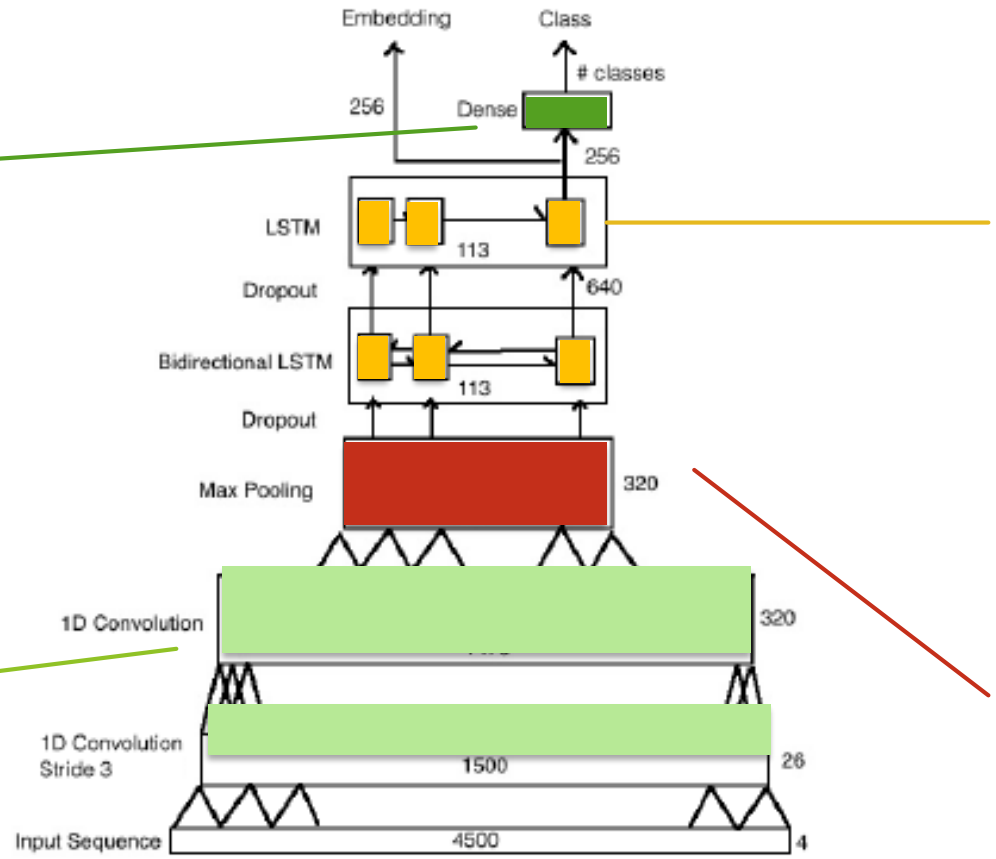
(ligne 66/ dpsites.R)

Dense ou Fully connected

Couche de base connectée où chaque neurone est connecté avec le neurone de la couche précédente.

Convolutional

Fait partie du **réseau neuronal convolutif** (CNN). Couches spéciales où sont formés des **filtres de taille déterminée**. Bon pour reconnaître différents **patrons**.



Senter, J. K., Royalty, T. M., Steen, A. D., & Sadovnik, A. (2019, November). Unaligned Sequence Similarity Search Using Deep Learning. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 1892-1899). IEEE.

LSTM/GRU

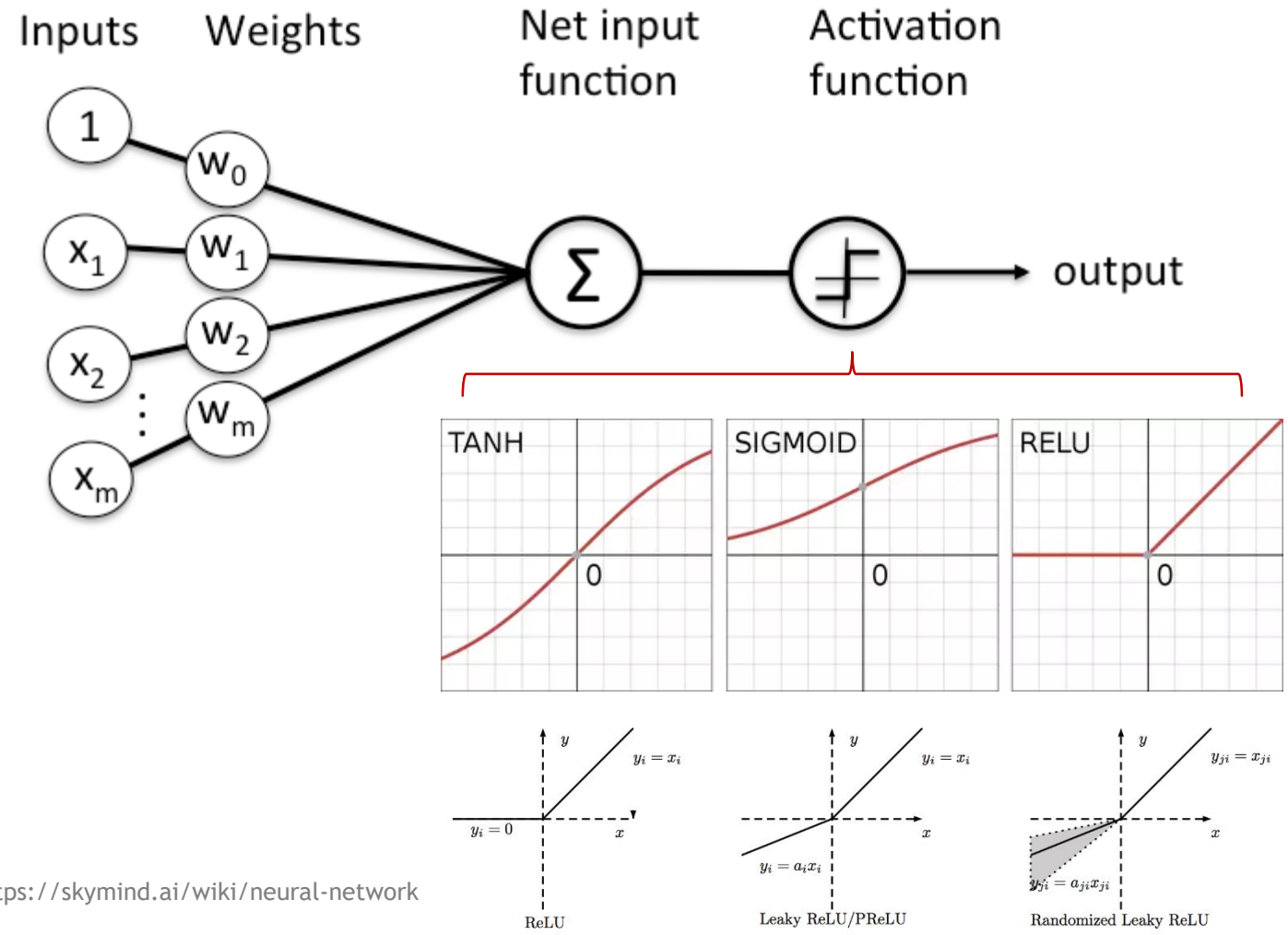
Les **cellules mémoires** font partie du réseau neuronal récurrent (RNN). Ces cellules sont normalement regroupées en série afin de se ***souvenir*** d'un schéma, par exemple, dans une série chronologique.

Pool/Unpool

Couche qui regroupe/dégroupe certaines données, ce qui simplifie les résultats.

Dense or Fully connected layer

Les principaux nœuds du réseau



https://skymind.ai/wiki/neural-network

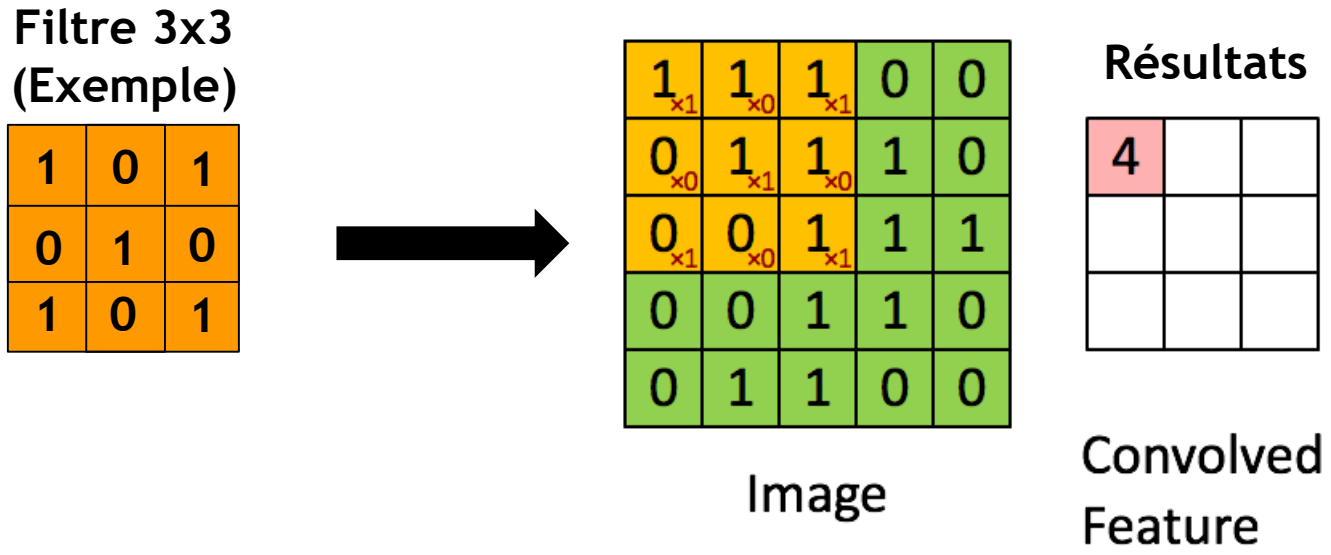
Couche dense

Une couche dans laquelle **chaque entrée est connecté à une sortie** avec un **poid appris**. Dans celle-ci, une opération linéaire est effectuée.

Convolutional layer

Ils forment la base des convolutional neural networks (CNN)

2



Convolution

Une couche dans laquelle des **filtres appris** sont appliqués sur chaque entrées. L'opération inverse est la **déconvolution**.

http://deeplearning.stanford.edu/wiki/index.php/Feature_extraction_using_convolution

Convolutional Neural Networks

Différentes couches = différentes caractéristiques

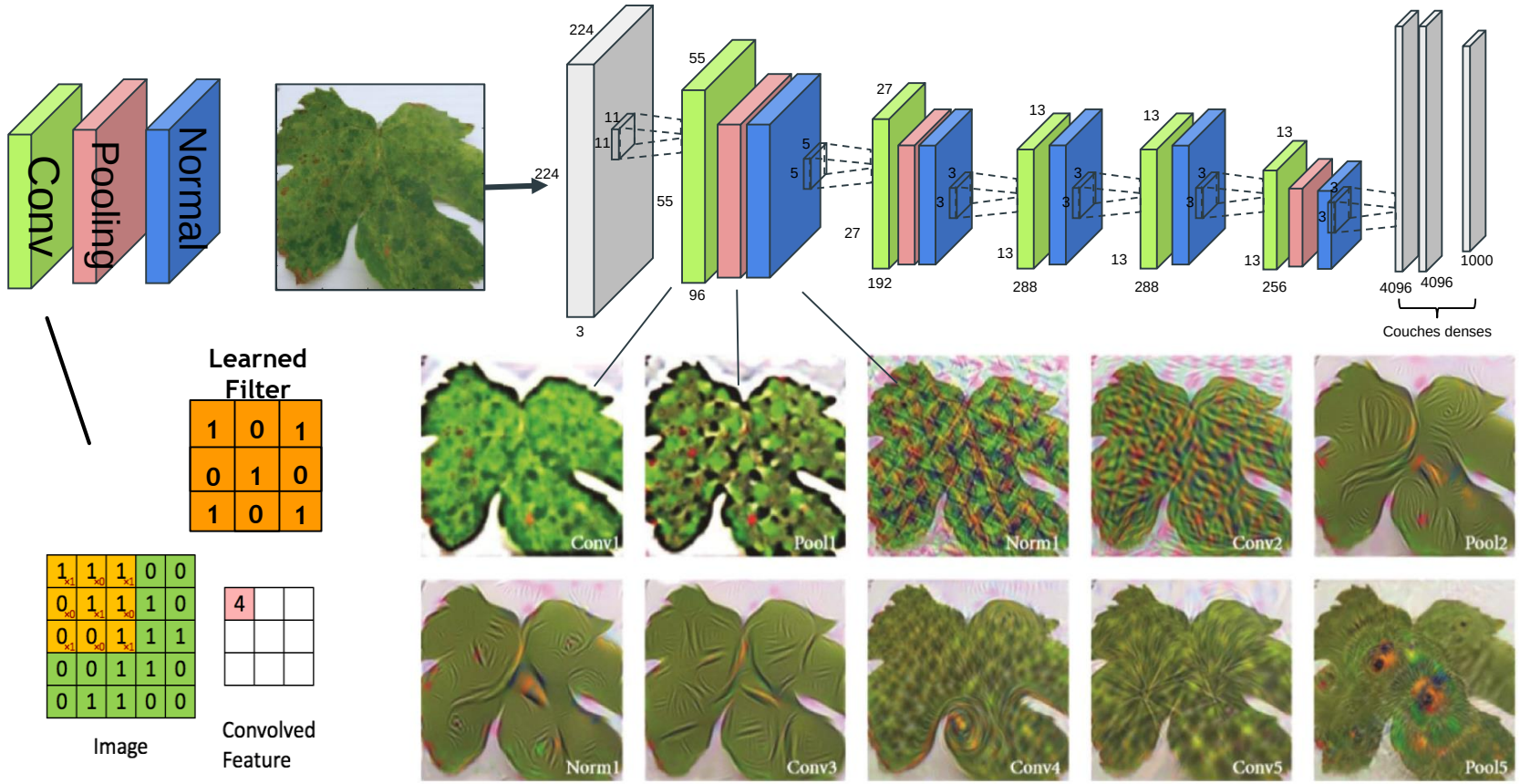


Fig. 2. Visualization of the output layers images after each processing step of the CaffeNet CNN (i.e. convolution, pooling, normalization) at a plant disease identification problem based on leaf images.
Source: Sladojevic et al. (2016).

Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, *147*, 70-90

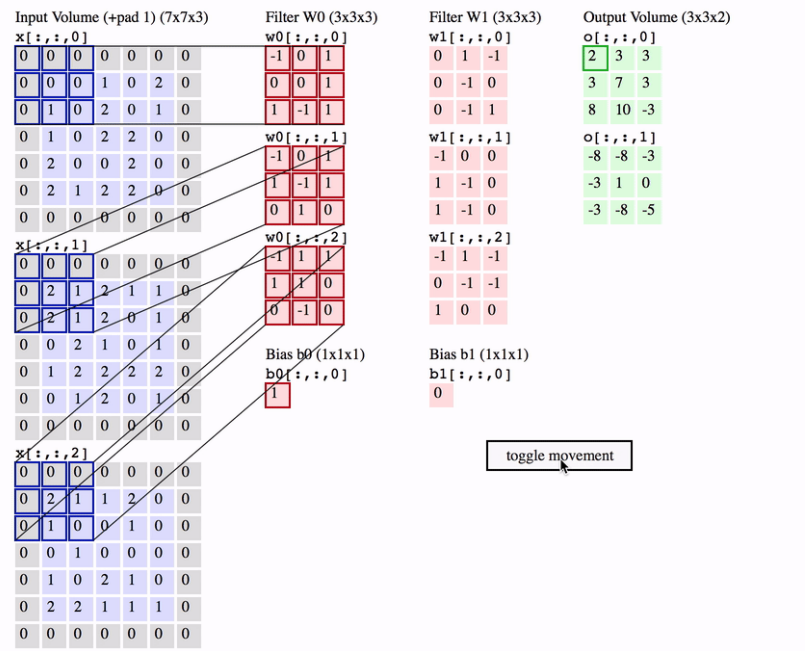
Convolution sur une image en couleurs

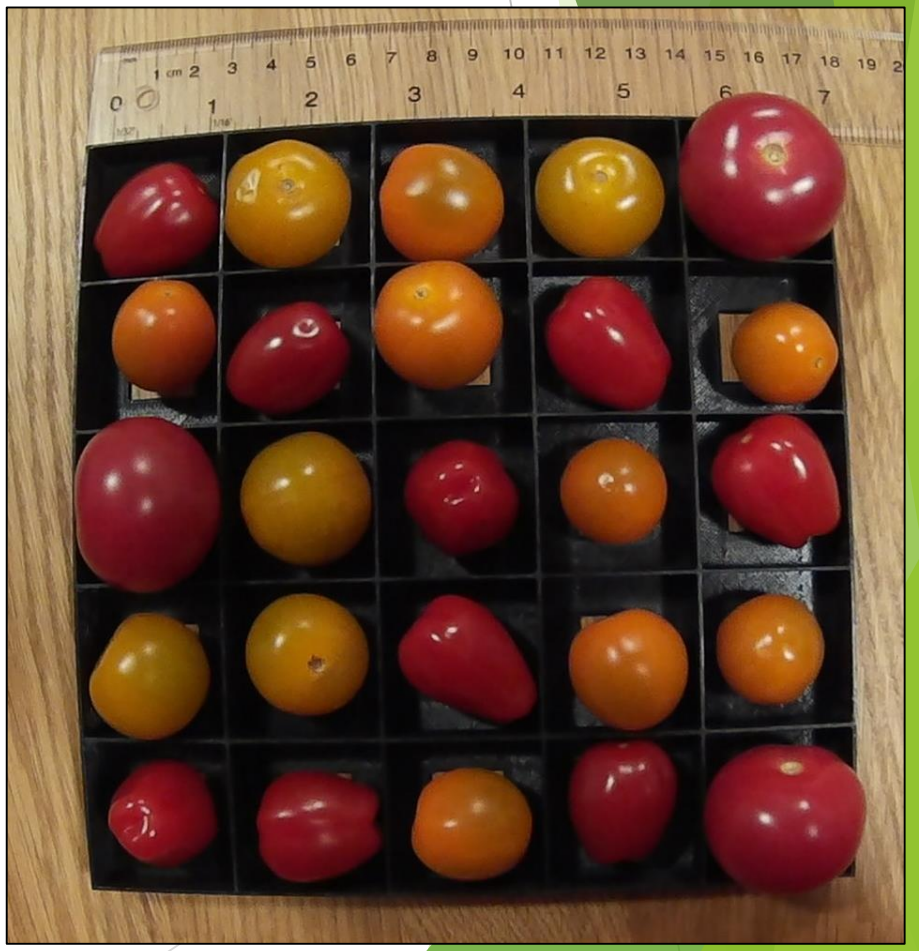
On appelle ce type de couche 3D-CNN

Rouge

Vert

Bleu



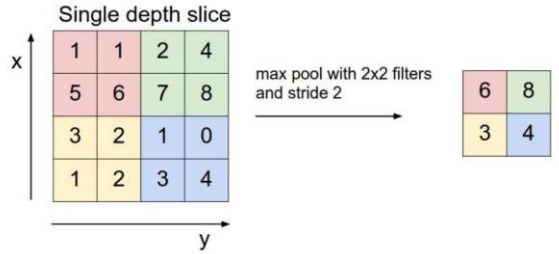


Pooling / Unpooling layer

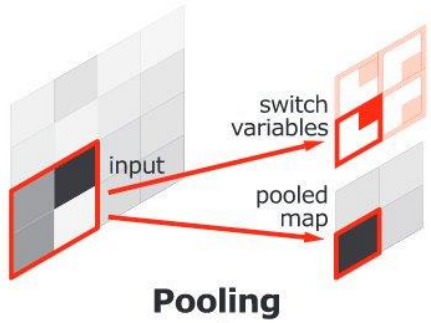
Opération qui réduit le nombre de paramètres du réseau

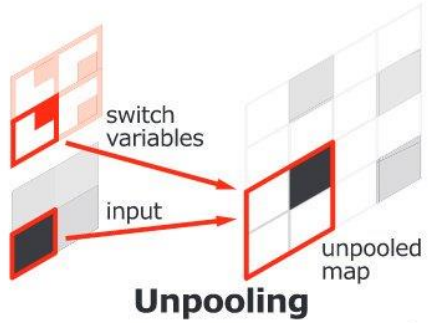
2

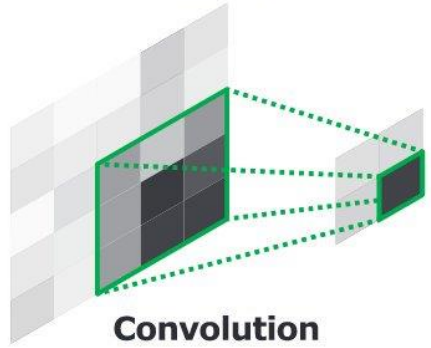
“Max Pooling”

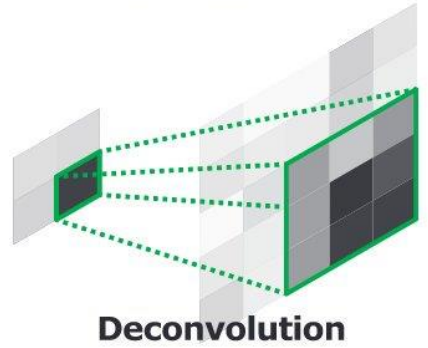


- Réduit le nombre de paramètres
- Ajoute de la résistance au bruit







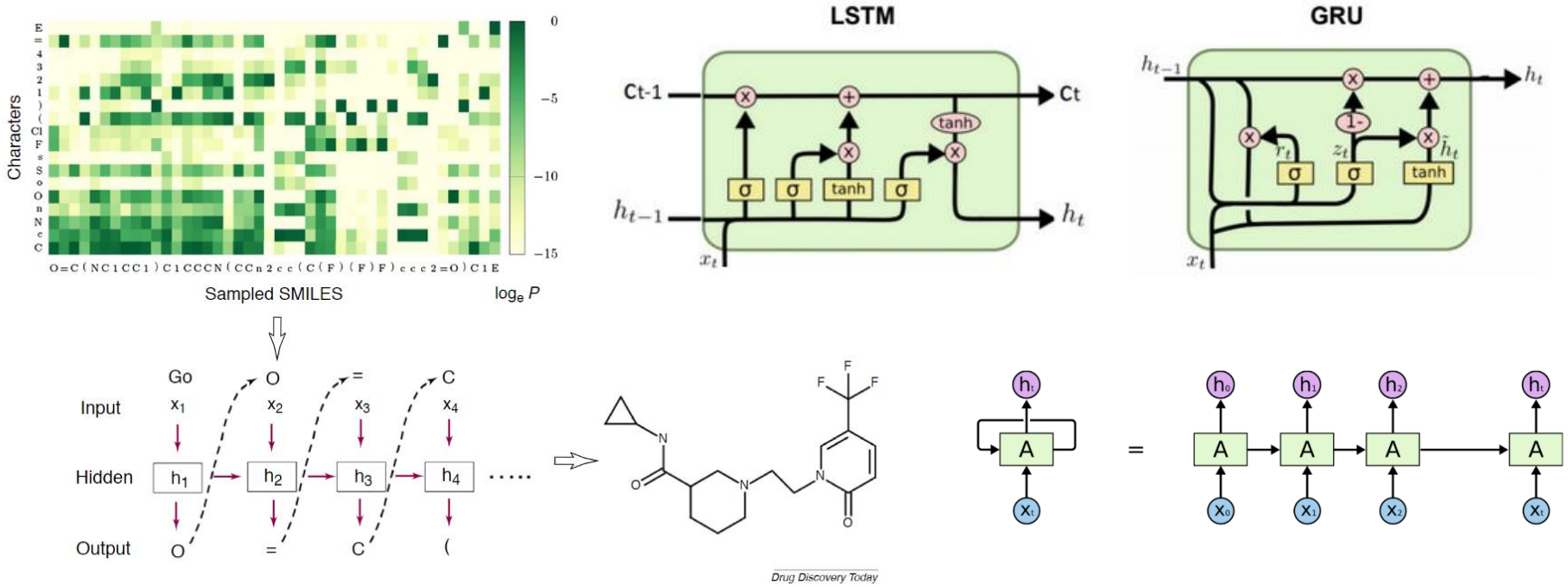


Pooling

Une couche de noeuds permettant de réduire les dimension appliquant une **somme** ou **moyenne** sur les valeurs en entrée

LSTM / GRU layer

Capables de conserver des informations séquentielles. Elle est utilisée dans les réseaux neuronaux récurrents (RNN).



LSTM/GRU

Couches permettant de **retenir de l'information séquentielle** (temporelle). Le **GRU** est la version la plus récente.

http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T. (2018). The rise of deep learning in drug discovery. *Drug discovery today*, *23*(6), 1241-1250.

Le model dans notre code

2

dpsites.R

```
model <- keras_model_sequential() %>%
  layer_conv_1d(filters = 32, kernel_size = 5,
  input_shape = c(nchar(sequence$Sites[1]), 22)) %>%
  layer_max_pooling_1d(pool_size =4) %>%
  layer_flatten() %>%
  layer_dense(units = 16, activation = "relu") %>%
  layer_dense(units = 2, activation = "softmax")


(lignes 66 à 72…)
```

Compilation du modèle

3

(ligne 74 / dpsites.R)

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d (Conv1D)              (None, 5, 32)             3552
_________________________________________________________________
max_pooling1d (MaxPooling1D) (None, 1, 32)             0
_________________________________________________________________
flatten (Flatten)            (None, 32)                0
_________________________________________________________________
dense (Dense)                (None, 16)                528
_________________________________________________________________
dense_1 (Dense)              (None, 2)                 34
=================================================================
Total params: 4,114
Trainable params: 4,114
Non-trainable params: 0
```

summary(model)

dpsites.R
1D CNN
32 filters of size 5

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm (LSTM)                  (None, 9, 32)             7040
_________________________________________________________________
flatten (Flatten)            (None, 288)               0
_________________________________________________________________
dense (Dense)                (None, 16)                4624
_________________________________________________________________
dense_1 (Dense)              (None, 2)                 34
=================================================================
Total params: 11,698
Trainable params: 11,698
Non-trainable params: 0
```

summary(model)

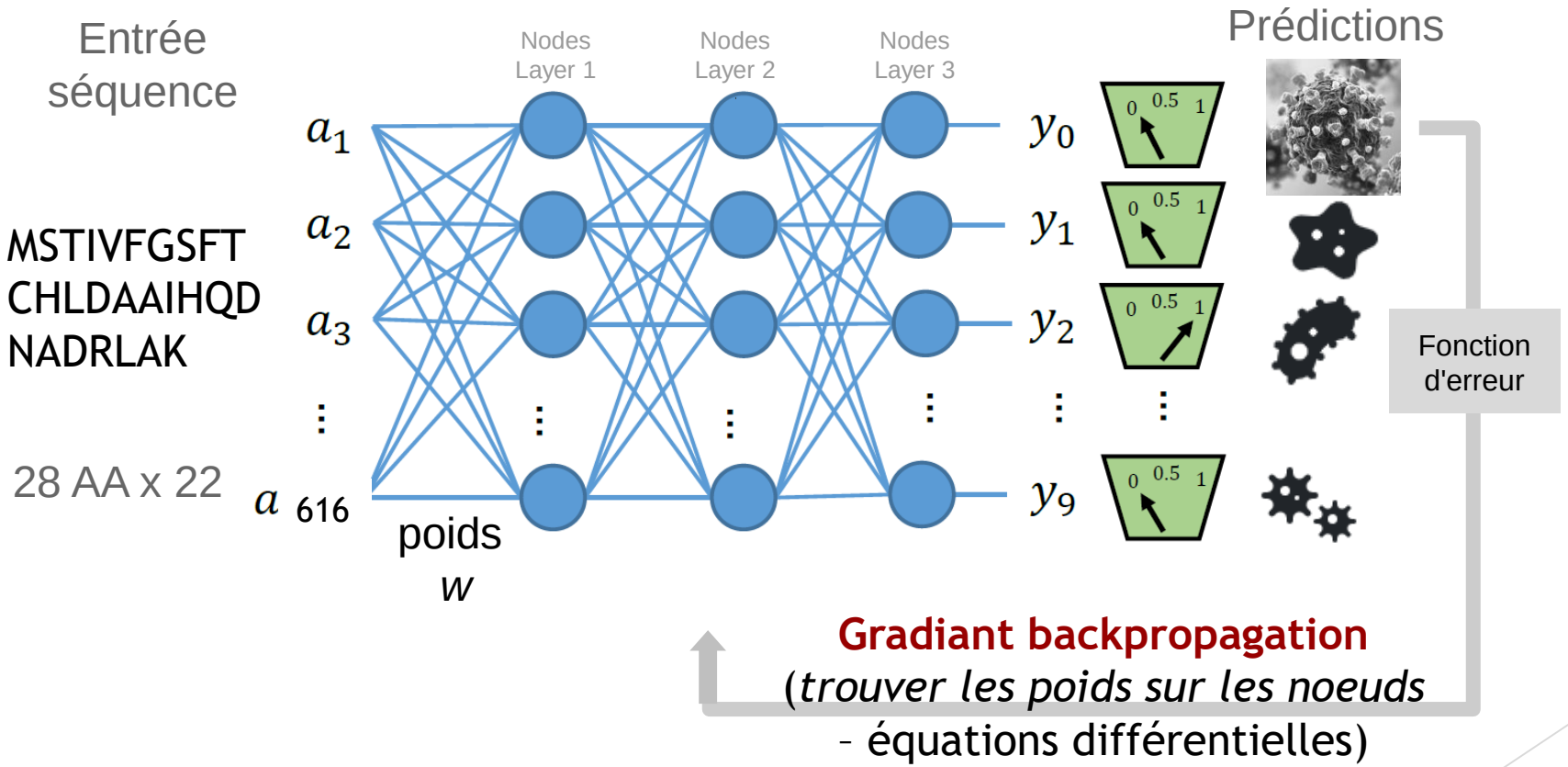
dpsites_LSTM.R
LSTM
32 filters of size 9

Entraînement

Dans l'entraînement, on minimise l'erreur du modèle

4

(ligne 85 / dpsites.R)



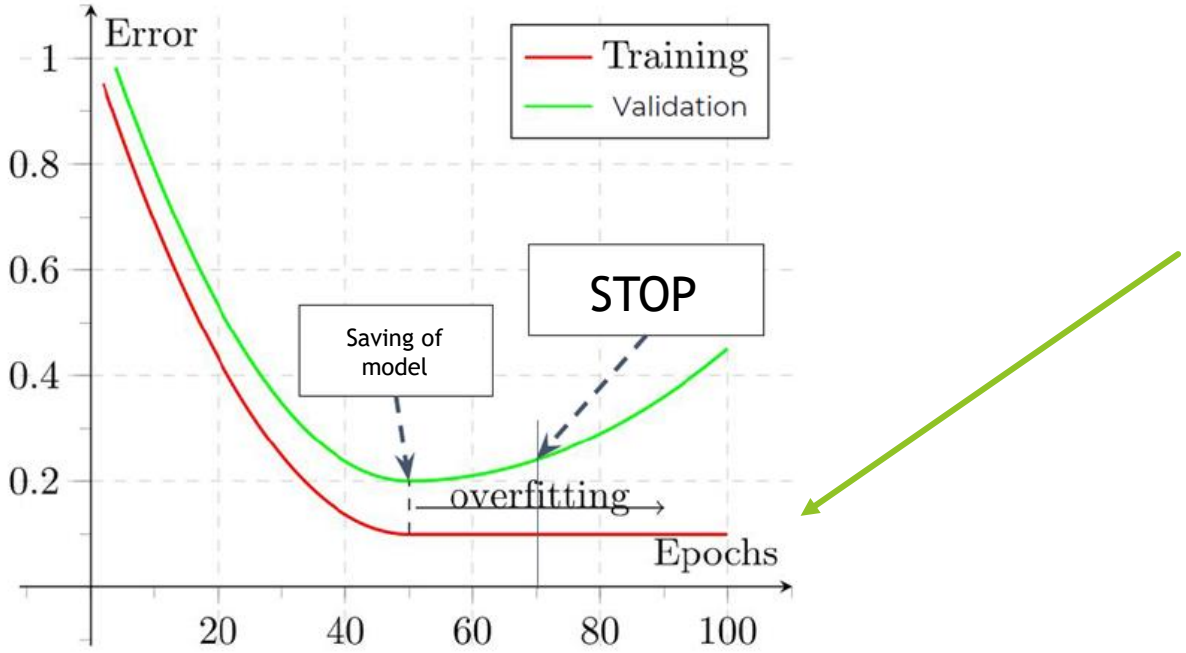
Fonction d'erreur

Fonction de différence entre les données **d'entraînement** et de **validation** que l'on cherche à minimiser dans l'apprentissage machine *e.g.* MSE, Binary cross-entropy

Rumelhart, David E.; Hinton, Geoffrey E.; Williams, Ronald J. (8 October 1986). "Learning representations by back-propagating errors". Nature. 323 (6088): 533–536.

Division des données

Avant de commencer, on va préserver certaines données



Epoch

Une **itération de l'algorithme** de minimisation de l'erreur sur l'ensemble des données.

4

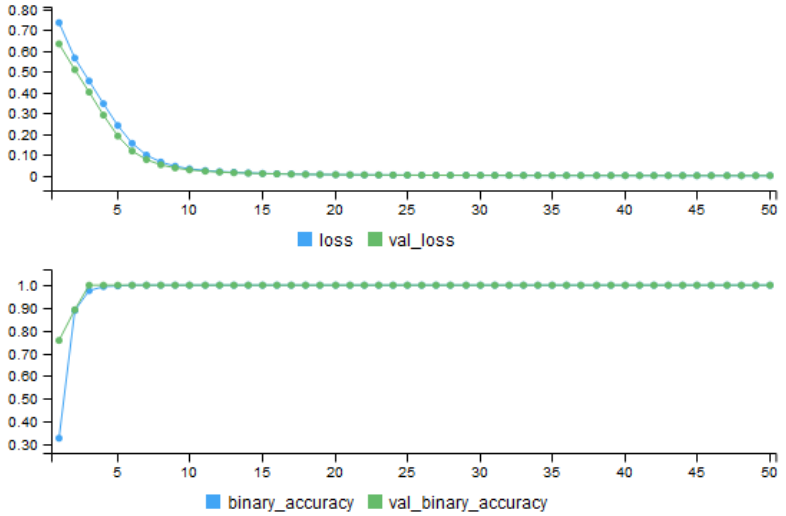
(ligne 85 / dpsites.R)

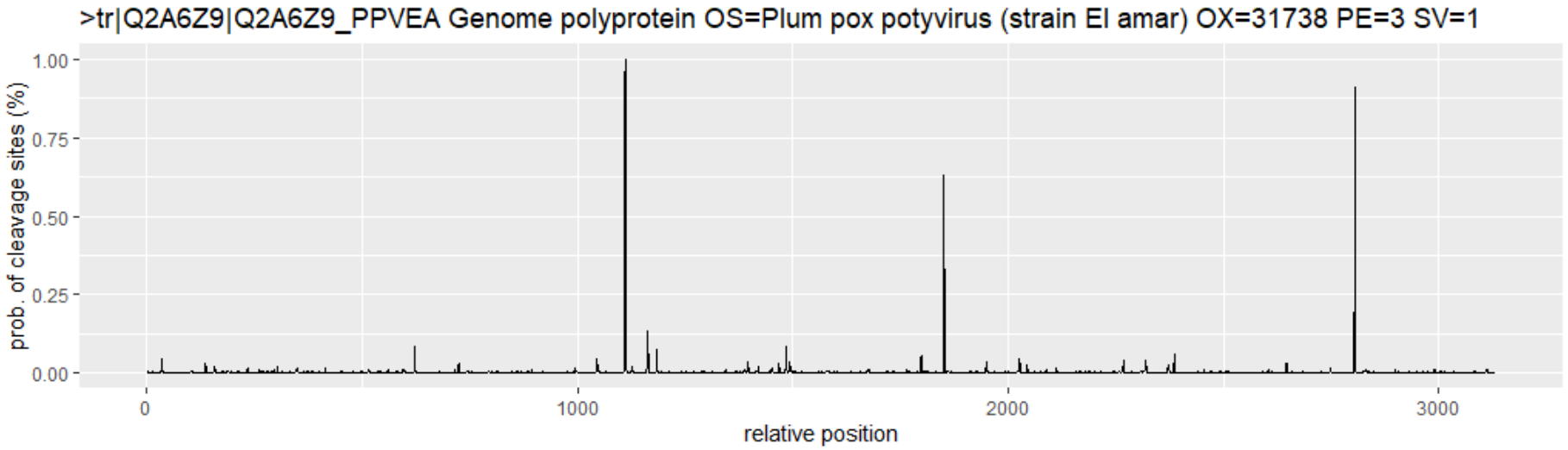
Prédictions

5

(ligne 95 / dpsites.R)

```
binary_accuracy: 1.0000 - val_loss: 7.3469e-04 - val_binary_accura
Epoch 48/50
331/331 [==============================] - 0s 105us/sample - loss:
binary_accuracy: 1.0000 - val_loss: 7.0492e-04 - val_binary_accura
Epoch 49/50
331/331 [==============================] - 0s 106us/sample - loss:
binary_accuracy: 1.0000 - val_loss: 6.7839e-04 - val_binary_accura
Epoch 50/50
331/331 [==============================] - 0s 105us/sample - loss:
binary_accuracy: 1.0000 - val_loss: 6.5290e-04 - val_binary_accura
```





Explications

1. Chargement des séquences d'entraînement (224 pour chacune des classes).
2. Création d'un *convulational neural network* (CNN).
3. Entraînement du CNN à distinguer entre les bons et mauvais sites de clivage pour une protéase.
4. Application du modèle entraîné à un groupe de séquences protéiques inconnues pour détecter des sites de clivage potentiels.

```
P3-_CAJ43126.1  QVVVHQSKR      1
P3-_CAJ43126.1  STNFLPNTF      0
P3-_CDO19252.1  QVVVHQSKR      1
P3-_CDO19252.1  SEKFSAILE      0
P3-_CDO19251.1  QVVVHQSKR      1
P3-_CDO19251.1  QSMGKLFCK      0
```

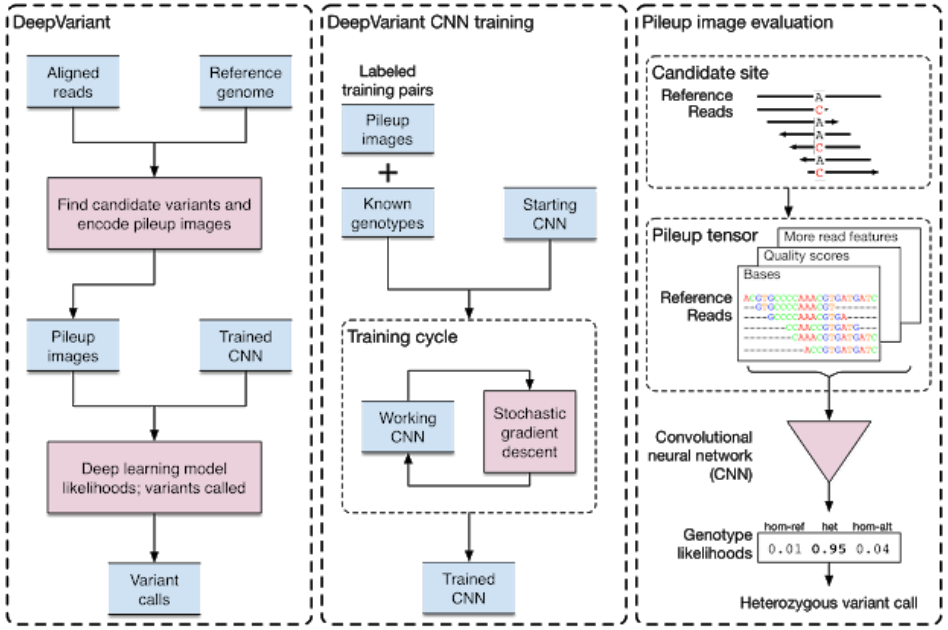
Autres concepts

Quels sont les autres applications de l'apprentissage profond dans l'analyse de séquences

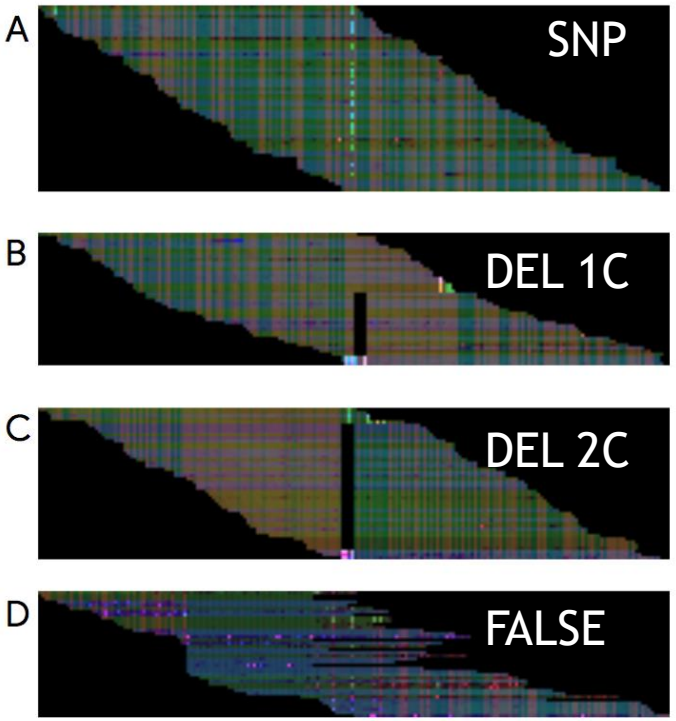
Google deep variant

Utilisation pour la détections d'INDELs et SNPs

Google deep variant est un CNN

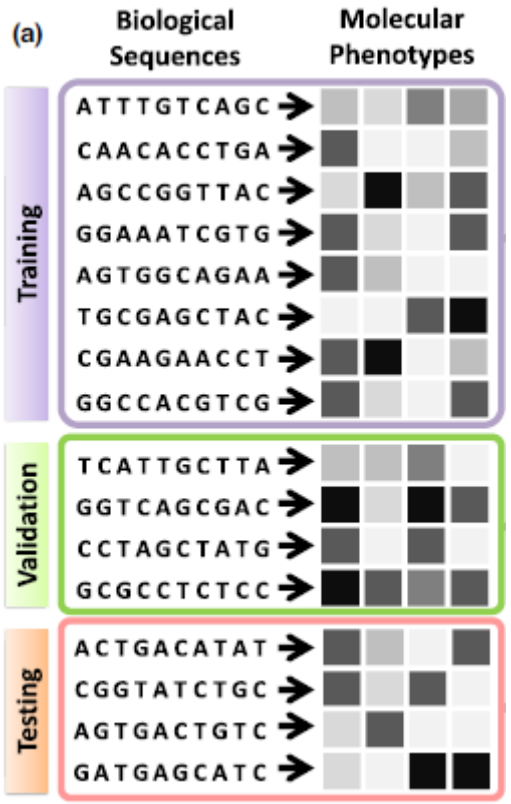


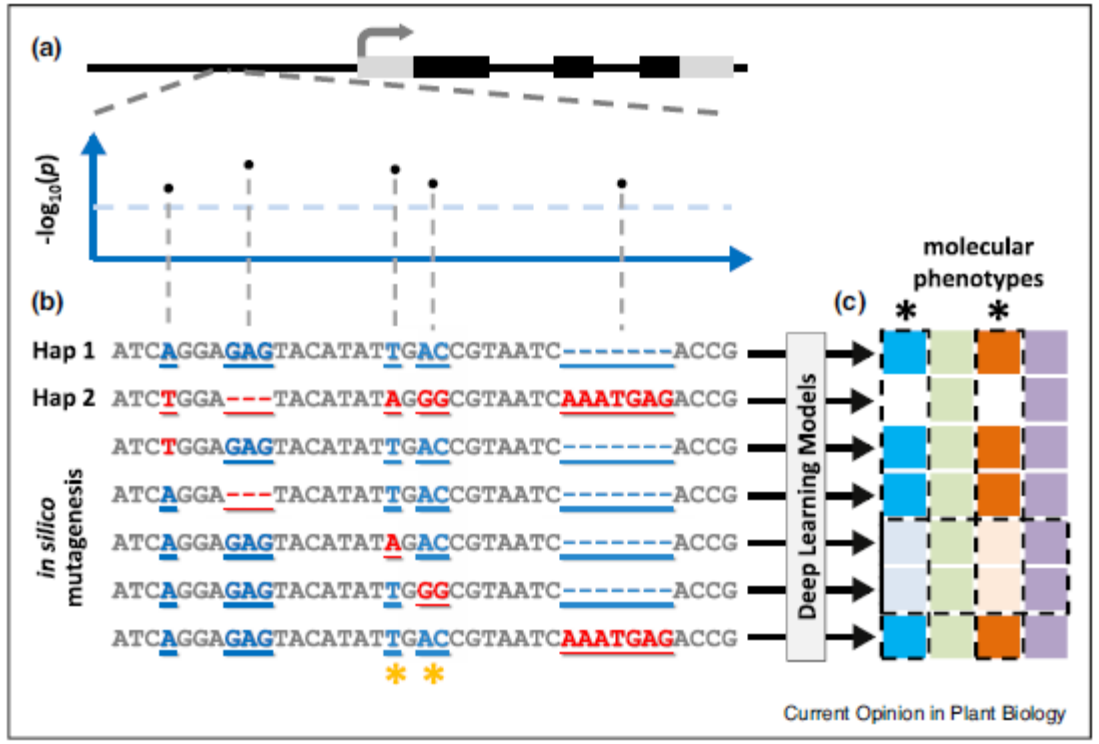
Pileup sequence images



https://ai.googleblog.com/2017/12/deepvariant-highly-accurate-genomes.html
https://github.com/google/deepvariant

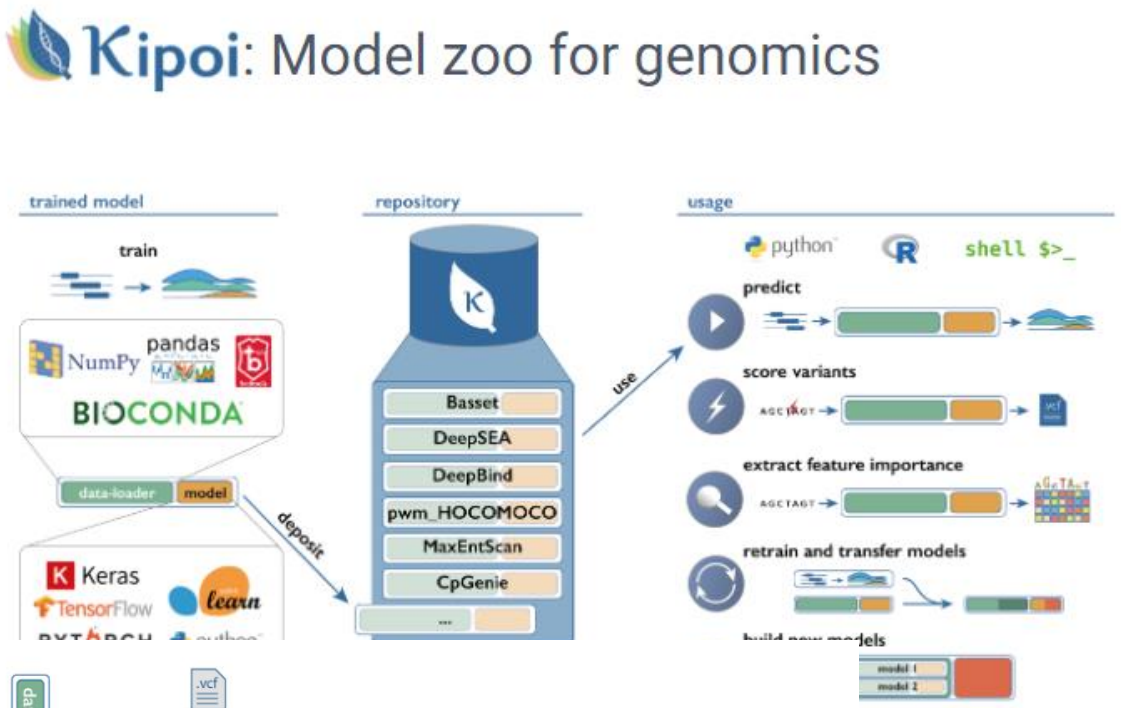
Prédictions du phénotypes à partir de données génomiques

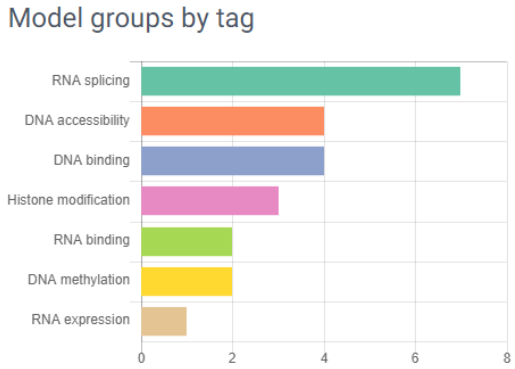


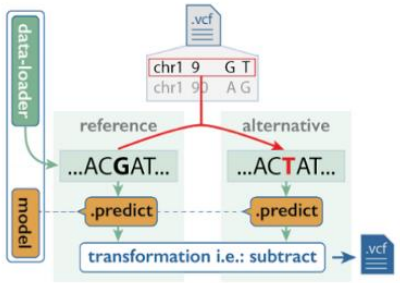


Wang, H., et al. (2020). Deep learning for plant genomics and crop improvement. *Current Opinion in Plant Biology*, *54*, 34-41.

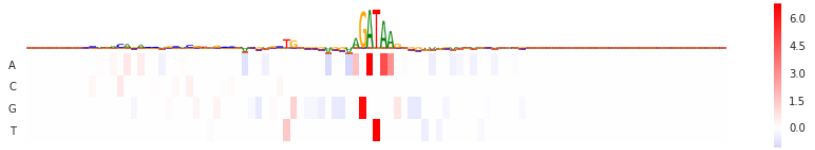
Modèles déjà entraînés








```
# Annotate VCF file with
# variant scores
kipoi veff score_variants \
  <Model> \
  --dataloader_args='{
    "fasta_file": "hg38.fa"}' \
  --vcf_path 'input.vcf' \
  -o 'annotated.vcf'
```



https://medium.com/@zigaavsec/kipoi-utilizing-machine-learning-models-for-genomics-e75b7e26a56c
https://github.com/kipoi/kipoi

Références

- https://ai.googleblog.com/
- https://openai.com/
- Wang, H., et al. (2020). Deep learning for plant genomics and crop improvement. *Current Opinion in Plant Biology*, *54*, 34-41. (Bon review des concepts clés)
- Zhang, Y., Zhou, X., & Cai, X. (2020). Predicting Gene Expression from DNA Sequence using Residual Neural Network. *bioRxiv*. https://doi.org/10.1101/2020.06.21.163956 (nouveau modèle de transfert pour la prédiction de l'expression des gènes)
- Zou, J., et al. (2018). A primer on deep learning in genomics. Nat Genet. 2019 Jan;51(1):12-18. doi: 10.1038/s41588-018-0295-5. (code original pour la détection de caractéristiques génomiques).
- Lopez-del Rio, A., et al. (2020). Effect of sequence padding on the performance of deep learning models in archaeal protein functional prediction. *Scientific reports*, *10*(1), 1-14. (encodage des sequences pour l'apprentissage profond)
- Vu, D., Groenewald, M., & Verkley, G. (2020). Convolutional neural networks improve fungal classification. Scientific reports, 10(1), 1-12. (utilisation de different modèles (CNN, BLAST, RDP et DBN pour la classification de séquences)