



Fondamentaux du Machine Learning



8-18 juin 2021
Formation CNRS – IFSeM



L'équipe enseignante



Alexandre Boucaud

ingénieur de recherche
au laboratoire APC



Sylvain Caillou

ingénieur de recherche
au L2IT

Agenda

- Horaires sur les 2 semaines
 - du mardi au vendredi
 - 9h – 12h30
 - Semaine 1 – bases du ML
 - Semaine 2 – deep learning
-

Modalités

1/2

- alternance de cours, cours TP et TP encadrés
- TP sur notebooks avec correction
- point en fin de journée pour passer en revue les questions / difficultés



Modalités

2/2

- webcam sur Zoom suggérée pendant les cours
- pauses régulières pour les questions
- utilisation du chat sur Citadel pour garder une trace des échanges / questions



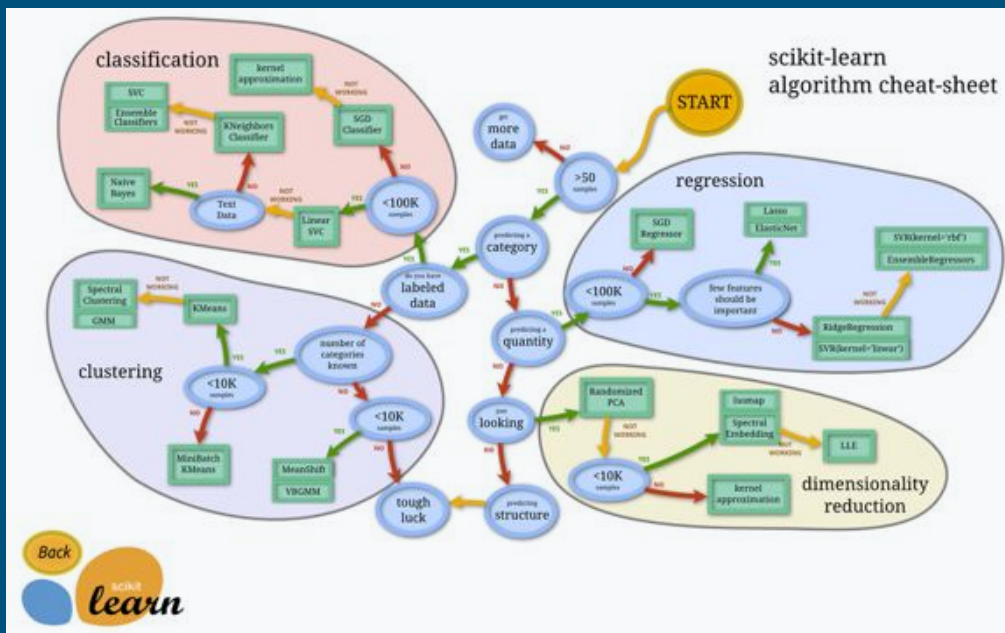
Sommaire

1. qu'est ce que le ML ?
2. d'où vient sa popularité ?
3. panorama des usages actuels
4. ce que vous allez apprendre dans ce cours



1. Qu'est ce que le ML ?

un ensemble d'algorithmes



Liste non exhaustive :

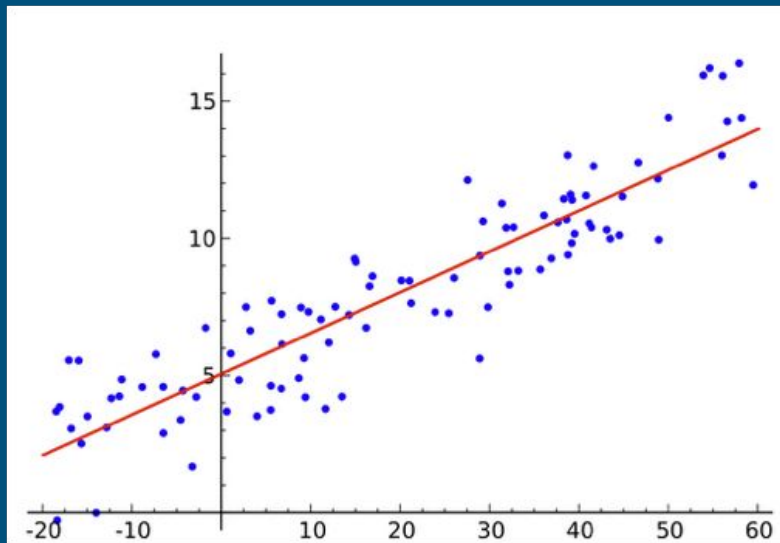
- Random Forests
- Nearest Neighbours
- Support Vector Machines
- (Deep) Neural Networks
- KMeans
- Gaussian Mixture Models
- Principal Component Analysis

Le machine learning c'est..

un ensemble d'**algorithmes**

qui produisent une **tâche précise**

régression

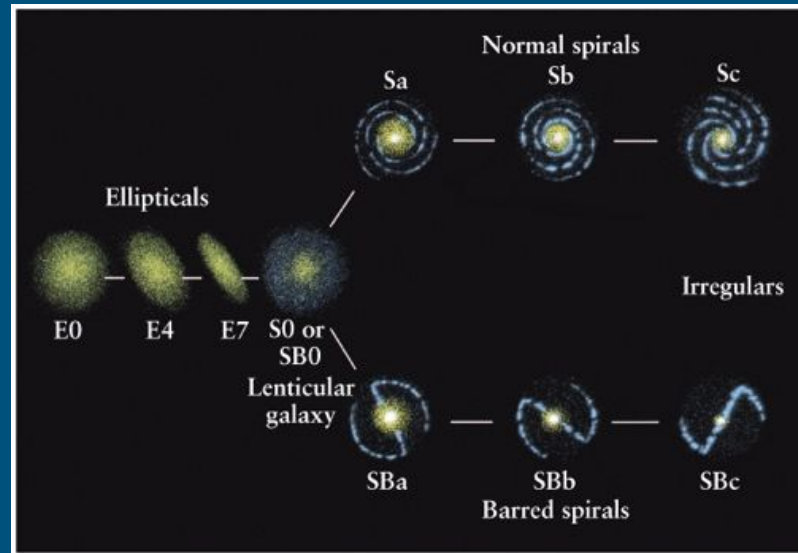


Le machine learning c'est..

un ensemble d'**algorithmes**

qui produisent une **tâche précise**

classification

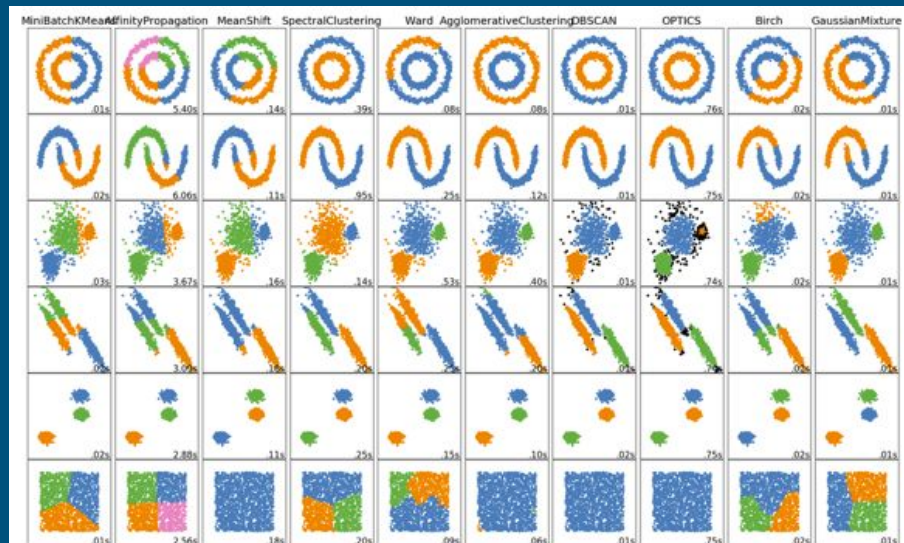


Le machine learning c'est..

un ensemble d'**algorithmes**

qui produisent une **tâche précise**

clustering



Le machine learning c'est..

un ensemble d'**algorithmes**

qui produisent une **tâche précise**

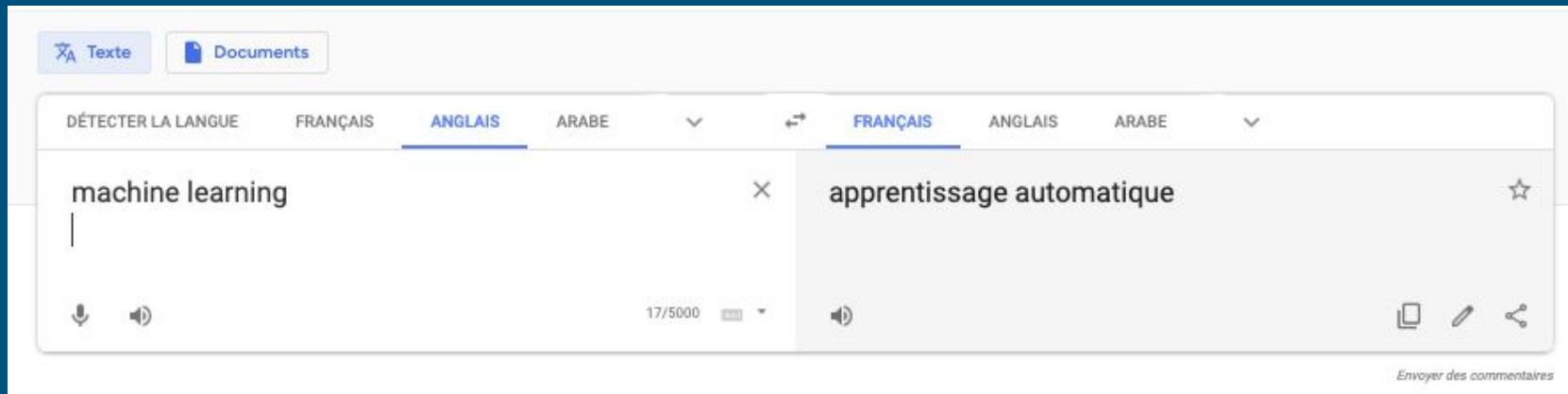
détection d'objets



Le machine learning c'est..

un ensemble d'**algorithmes**

qui produisent une **tâche précise**



Le machine learning c'est..

un ensemble d'**algorithmes**

qui produisent une **tâche précise**

avec des données

$$y = f(x)$$

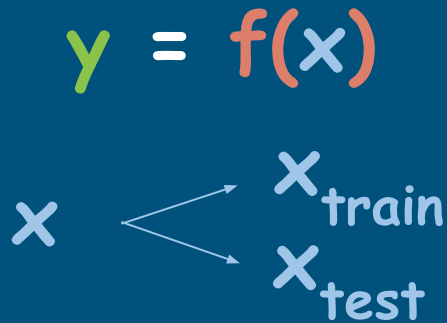
Autrement dit..

le machine learning fait
de l'**inférence** à partir de données

$$y = f(x)$$

Dans la pratique

1. séparation des données



2. entraînement

$$f(x_{\text{train}}) \rightleftharpoons Y_{\text{train}}$$

3. évaluation

$$f(x_{\text{test}}) = ? Y_{\text{test}}$$

Deux familles principales

Apprentissage supervisé

> action prédictive

permet de **répondre à une question** spécifique (régression, classification, etc)

nécessite des exemples où la solution est **connue** pour entraîner l'algorithme

se comporte comme de l'interpolation

Apprentissage non-supervisé

> action descriptive

permet de **découvrir des structures** dans les données (clustering, réduction de dimensions)

ne nécessite pas de données labellisées pour entraîner l'algorithme

peut s'utiliser **en amont de la classif/reg.**

Apprentissage supervisé (*supervised*)



problème de reconnaissance des
chiffres à partir d'images
(base de données MNIST)

<http://yann.lecun.com/exdb/mnist/>

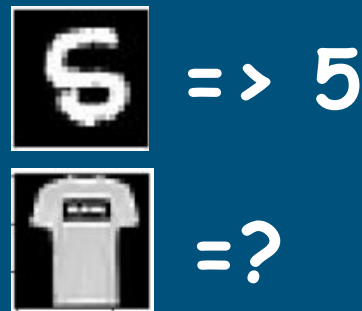
Apprentissage supervisé (*supervised*)



Entrainement :



Test :



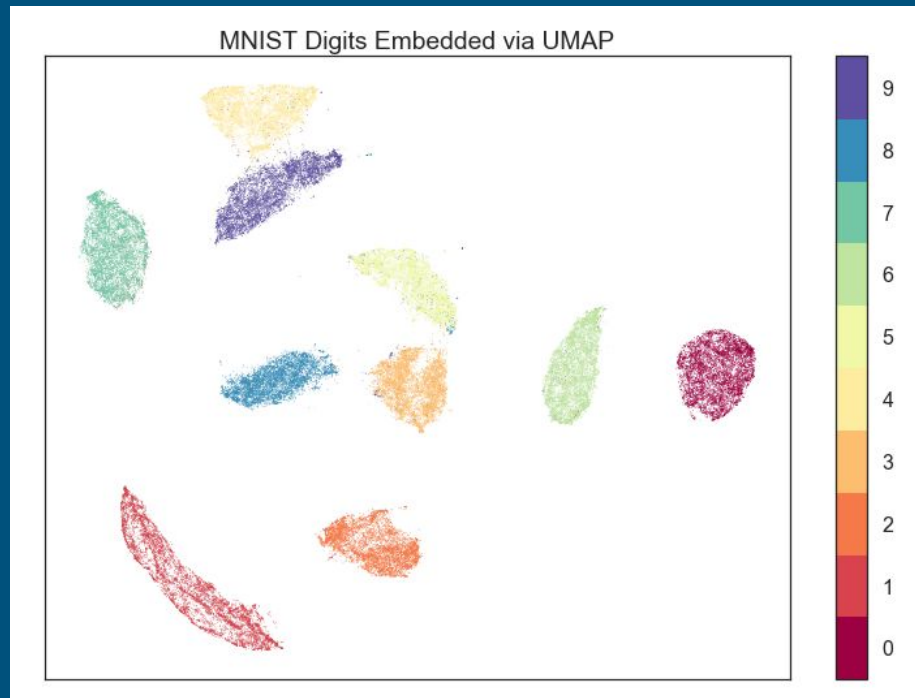


que dira un algorithme entraîné sur des images de chihuahua ?

Apprentissage non-supervisé (*unsupervised*)

visualisation des images MNIST
dans un espace 2d
avec l'algorithme UMAP

<https://github.com/lmcinnes/umap>



Ce qu'il faut retenir..

le machine learning c'est une **boîte à outils** pour les chercheurs

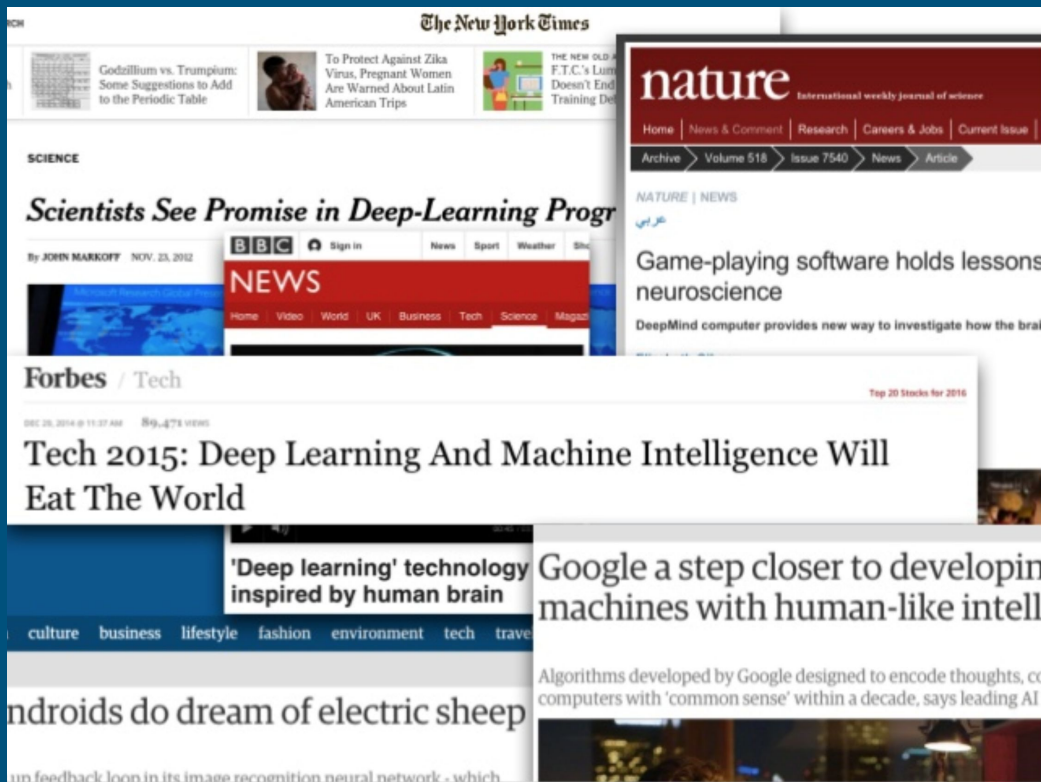
il existe une grande variété d'algorithmes et de modèles mais chacun d'entre eux est **plus adapté** pour une **tâche précise**

résoudre un problème avec du ML est surtout une question organisationnelle

- définition du problème
- caractérisation et pré-traitement des données

2. D'où vient sa popularité ?

En 2021, le machine learning est partout



mais il est surtout plébiscité par le terme IA

Quel est le lien entre intelligence artificielle et machine learning ?

Tout dépend de qui en parle..

les chercheurs détestent l'expression
"intelligence artificielle"

les industriels adorent



IA, la révolution n'a pas encore eu lieu - Michael Jordan (2018)

De quand datent ses premières utilisations ?

1957-69
dawn

techniques / tricks

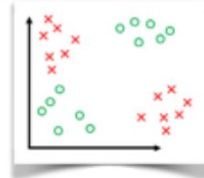
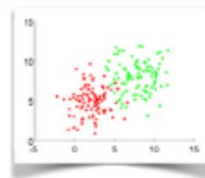
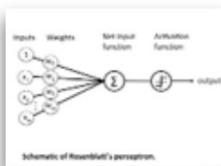
hardware

data

perceptron

early mainframes

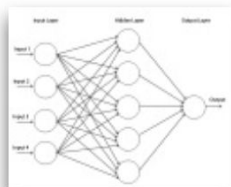
toy linear, small images, XOR



Golden age dans les années 90

1986-95
golden age

early NNs



workstations



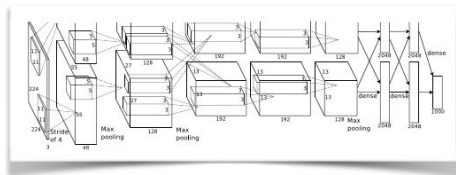
MNIST



Le renouveau après une période calme

2006-
deep learning

deep NNs



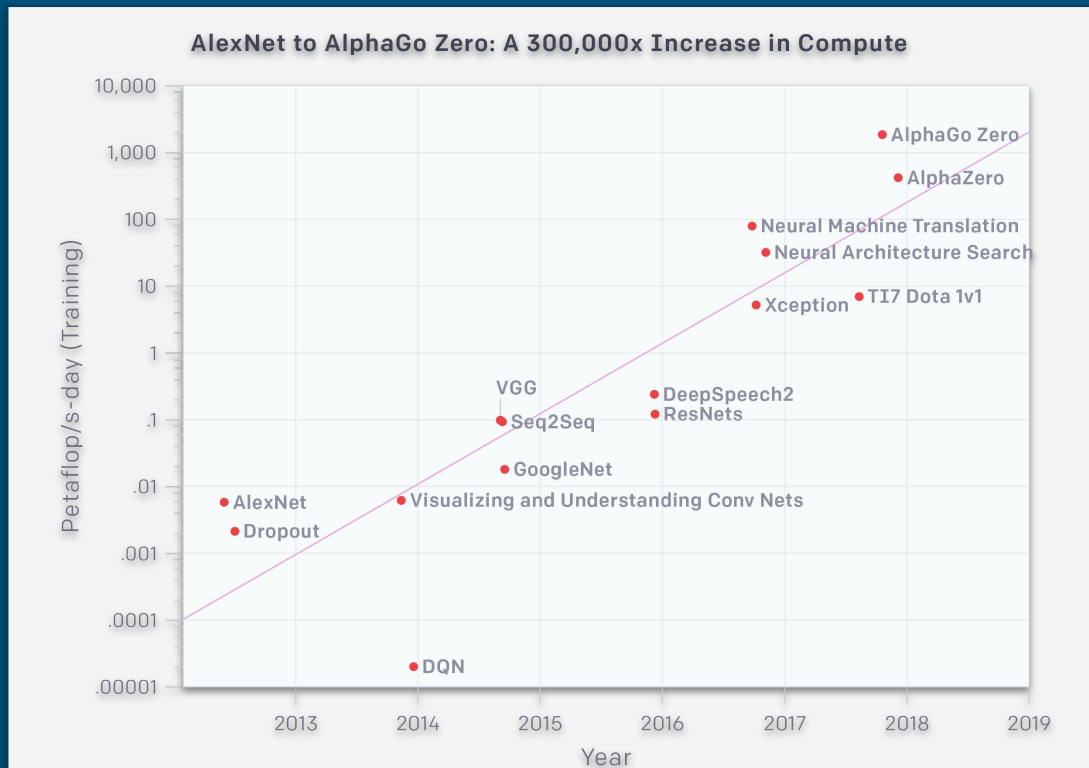
GPU, TPU, Intel Xeon Phi



Imagenet



Une utilisation sans cesse décuplée



les raisons du succès

- du **matériel** spécialisé de plus en plus performant
 - une grande disponibilité des **données**
 - une recherche stimulée sur les nouveaux **algorithmes**
 - un écosystème **open-source**
-

le matériel

hardware

- CPU dédiés
- GPU - graphics processing units
- TPU - tensor processing units



les données

data



les algorithmes

algorithms / models

Apprentissage..

- supervisé
- non-supervisé
- par récurrence
- par renforcement
- actif
- semi-supervisé
- ...



les logiciels libres

open-source software



theano



Microsoft
CNTK

PYTORCH



Caffe2

dmlc
mxnet

gensim

spaCy

3. Panorama des usages actuels

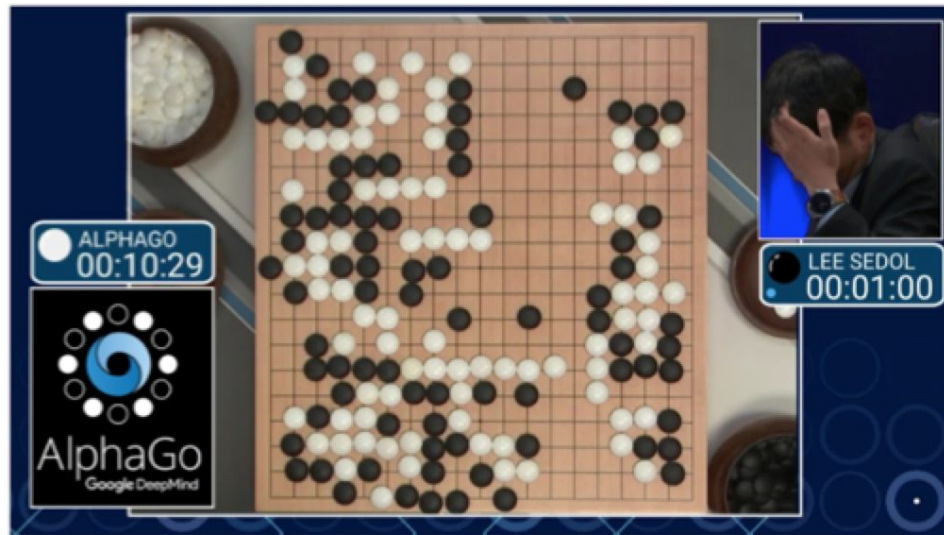
Parmi les applications courantes..

- traduction
 - du langage
 - image > texte
 - texte > image (génération)
- synthèse
 - de la parole
 - des images
 - des sons
- assistants personnels
- détections
 - reconnaissance des formes
 - reconnaissance faciale
 - aide à l'imagerie médicale
 - fraude (anomalies)
- voitures autonomes
- jeux
- minage de crypto-monnaies

Examples



Style transfer - Gatys (2015)




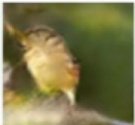








AlphaGo - DeepMind (2017)



Alexa - Amazon (2017)



Skin cancer diagnostic – Stanford (2017)

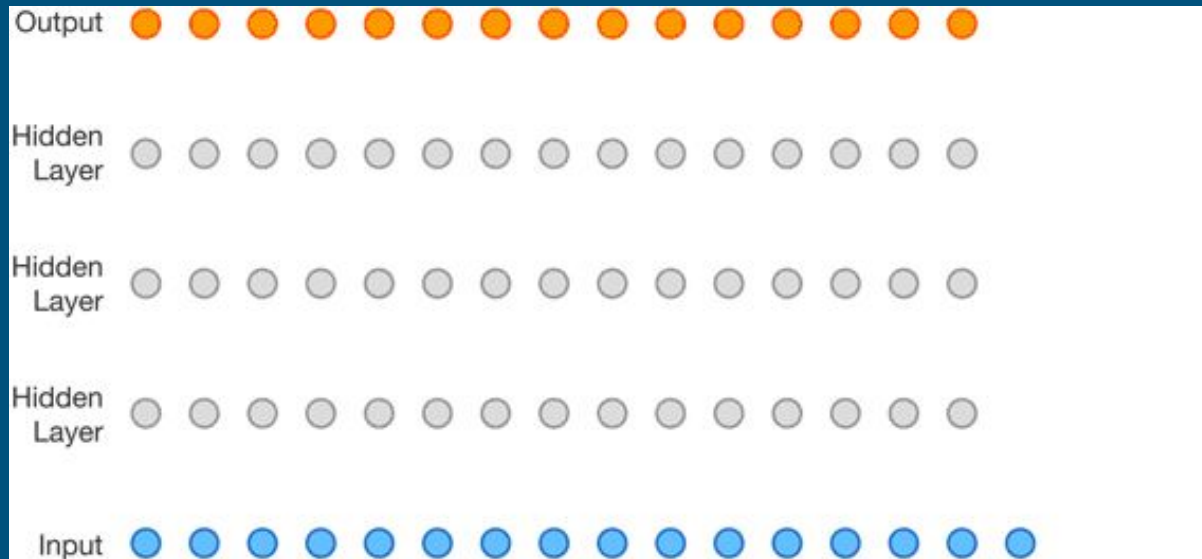
Text description	This bird is blue with white and has a very short beak	This bird has wings that are brown and has a yellow belly	A white bird with a black crown and yellow beak	This bird is white, black, and brown in color, with a brown beak	The bird has small beak, with reddish brown crown and gray belly	This is a small, black bird with a white breast and white on the wingbars.	This bird is white black and yellow in color, with a short black beak
Stage-I images							
Stage-II images							

StackGAN v2 – Zhang (2017)

Colorisation d'image (2017)



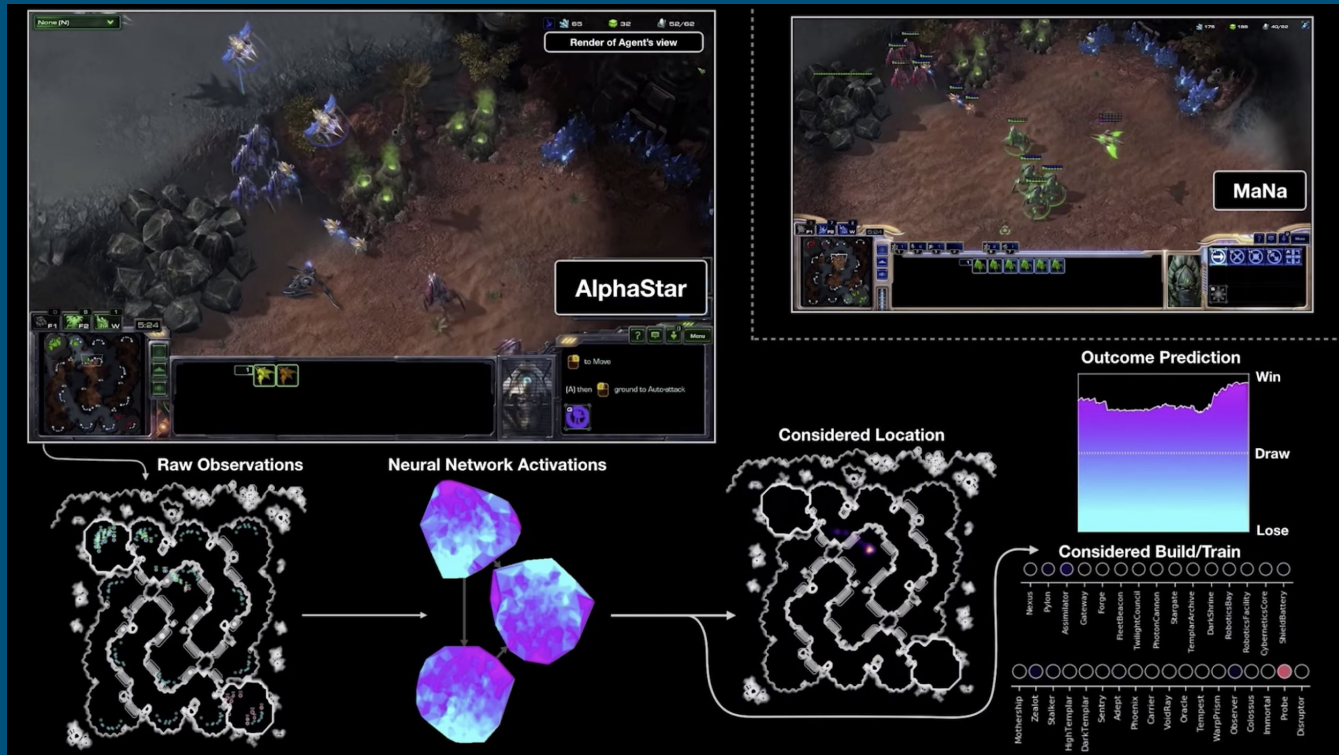
Synthèse de la parole (2017)



Synthèse des visages (2018)



Stratégie en temps réel (2019)



Efficient text parser/generator (2020)

GPT-3, l'intelligence artificielle qui a appris presque toute seule à presque tout faire

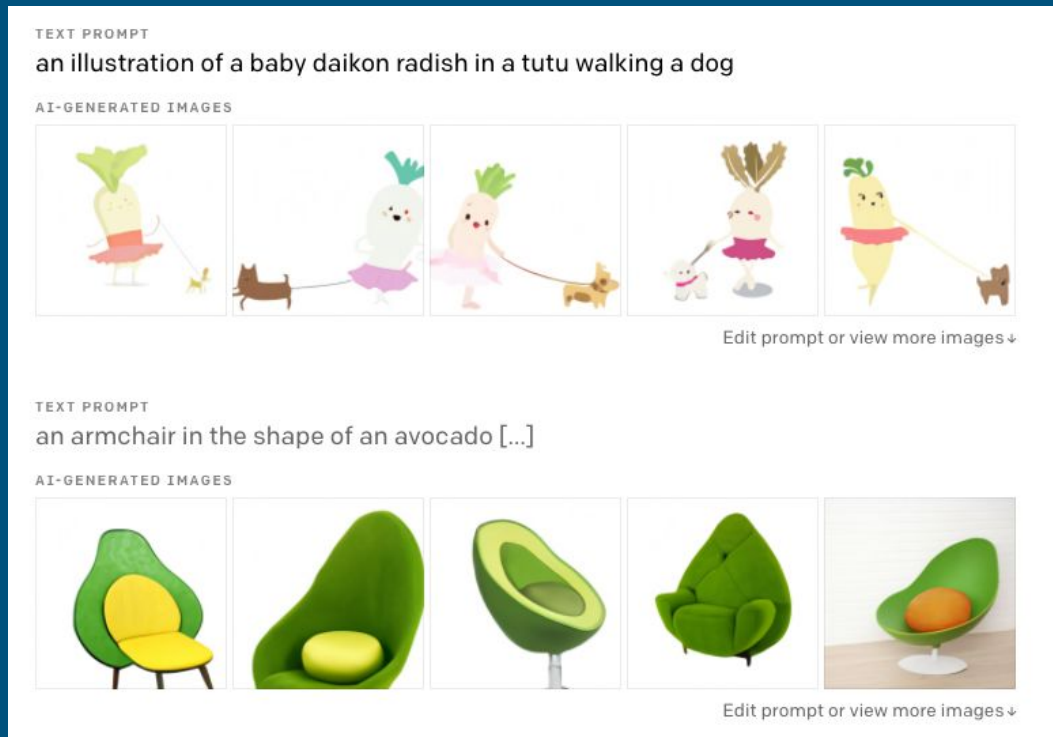
L'entreprise américaine OpenAI exploite le plus gros réseau de neurones artificiels au monde, effectuant une grande variété de tâches avec des résultats souvent bluffants, mais à la qualité imprévisible.

Le Monde (nov 2020)

<https://beta.openai.com>

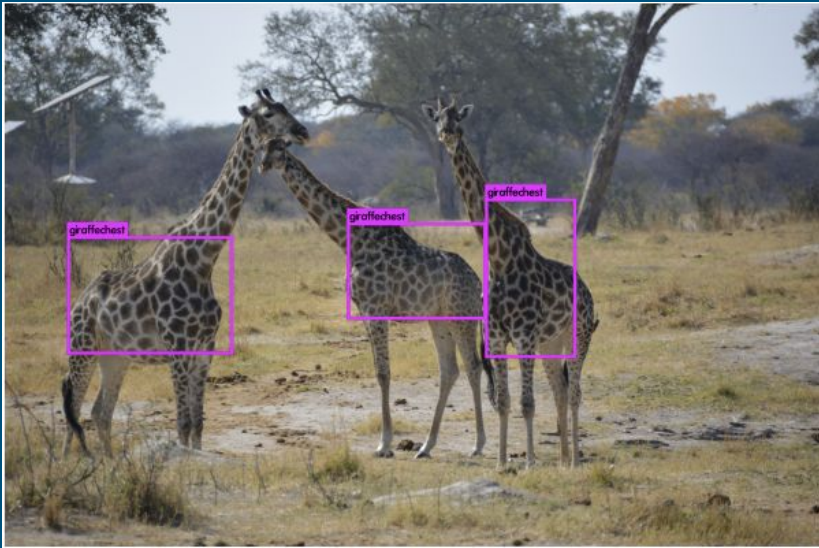
<https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html>

Image generation from text (2021)



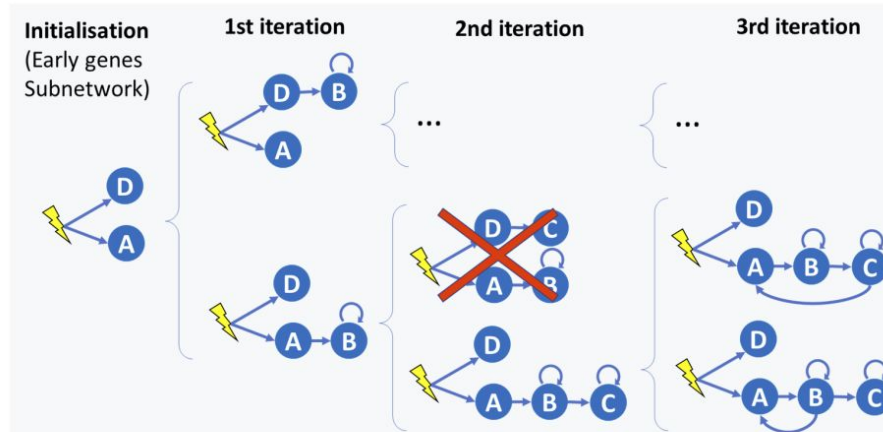
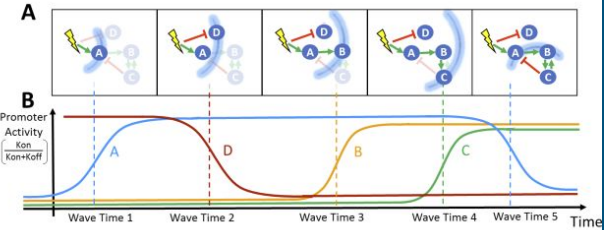
et en sciences..

Détection et suivi de girafes

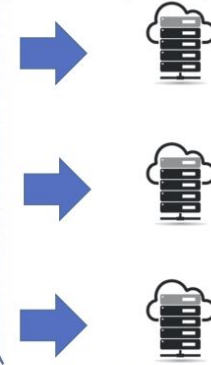


Prédiction de mutations génétiques

WASABI SPLITS & PARALLELIZES GRN INFERENCE PROBLEM

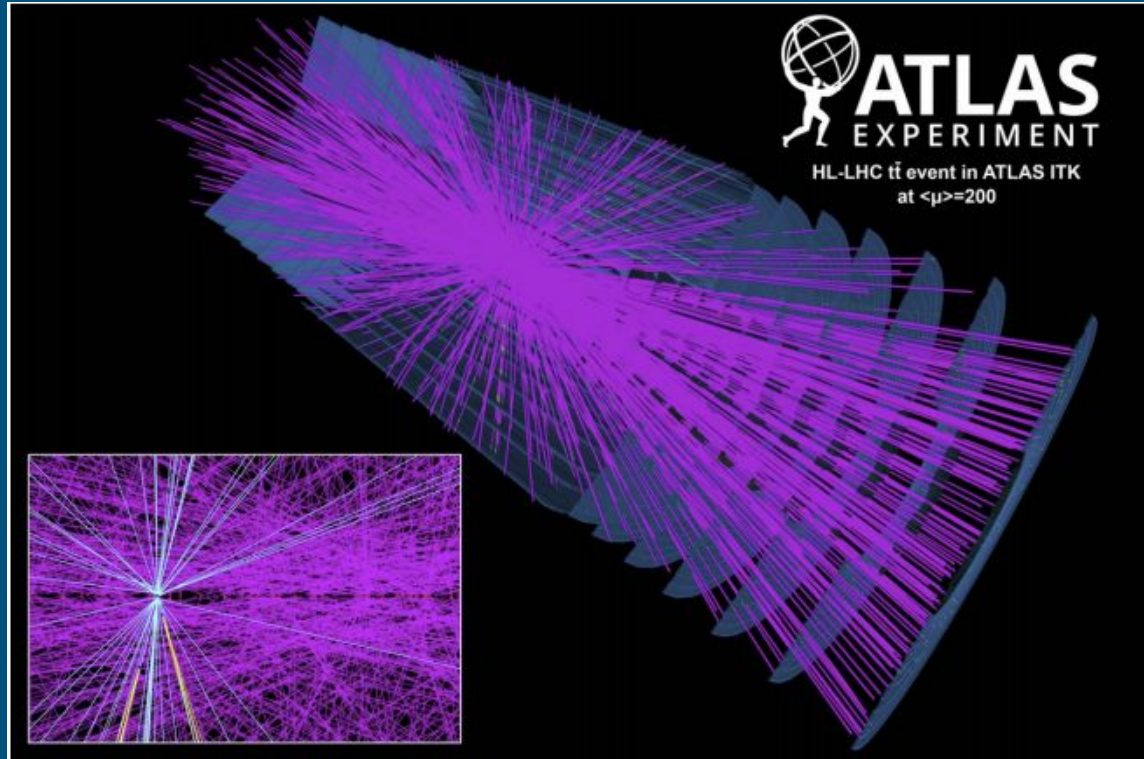


Cloud computing
(1000 cores / 15 days)

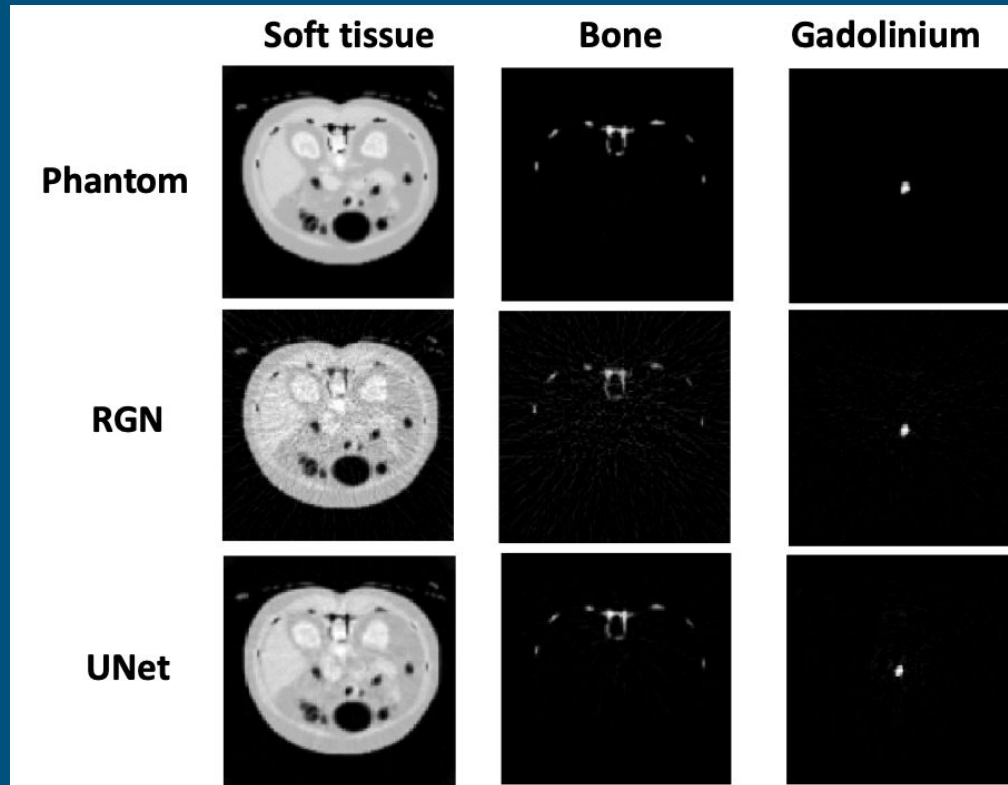


+ 
Machine Learning

Détection d'anomalies dans les calorimètres



Reconstruction et débruitage d'image méd.



Résoudre des équations et intégrales

DEEP LEARNING FOR SYMBOLIC MATHEMATICS

Guillaume Lample*
Facebook AI Research
glample@fb.com

François Charton*
Facebook AI Research
fcharton@fb.com

ABSTRACT

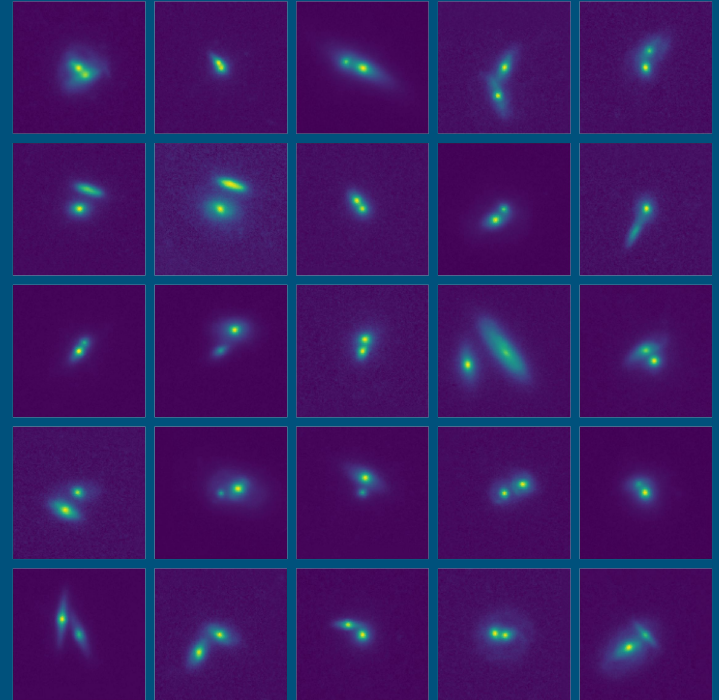
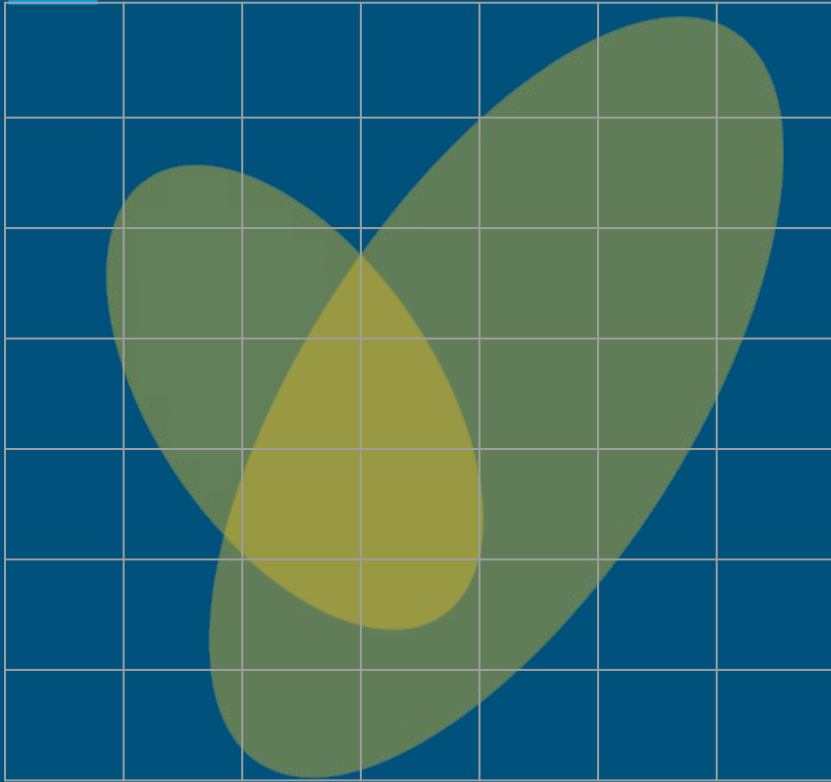
Neural networks have a reputation for being better at solving statistical or approximate problems than at performing calculations or working with symbolic data. In this paper, we show that they can be surprisingly good at more elaborated tasks in mathematics, such as symbolic integration and solving differential equations. We propose a syntax for representing mathematical problems, and methods for generating large datasets that can be used to train sequence-to-sequence models. We achieve results that outperform commercial Computer Algebra Systems such as Matlab or Mathematica.

Simulation de galaxies

A vous de trouver l'image simulée..

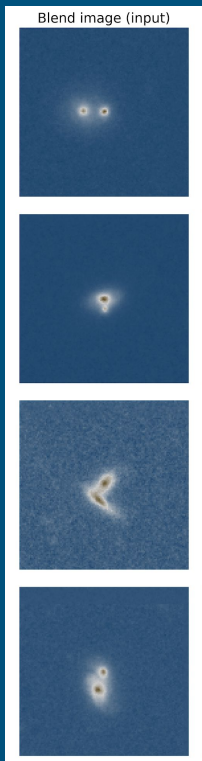


Séparation de galaxies

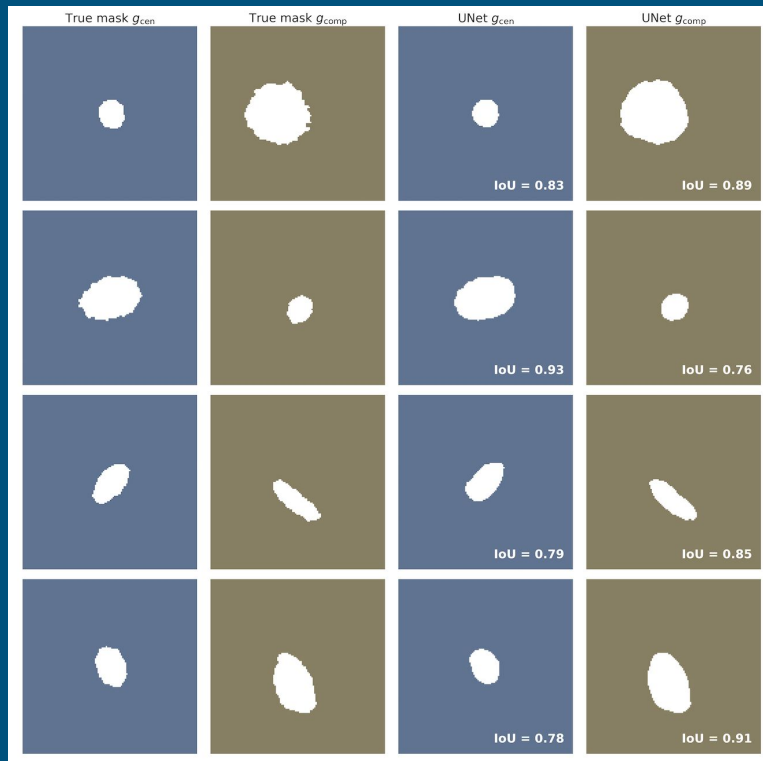


Séparation de galaxies

INPUT IMAGES
(TEST SET)



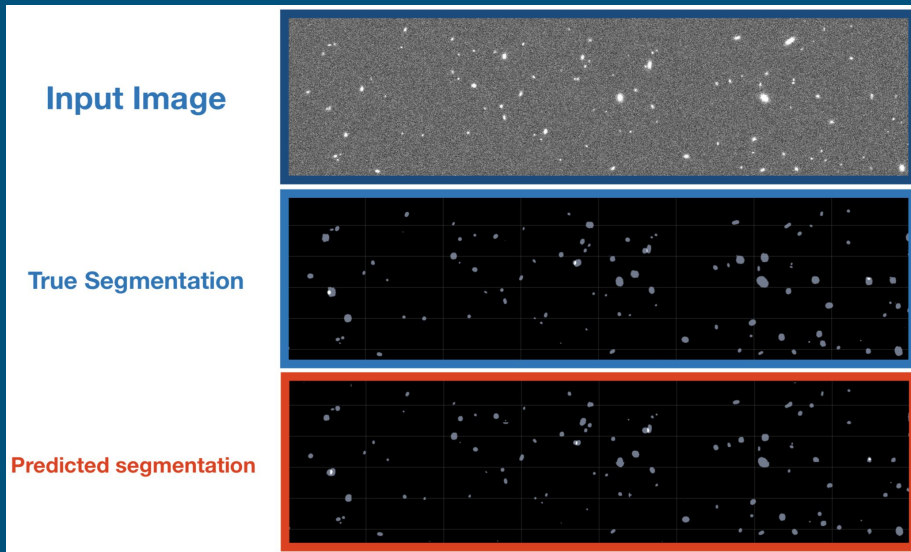
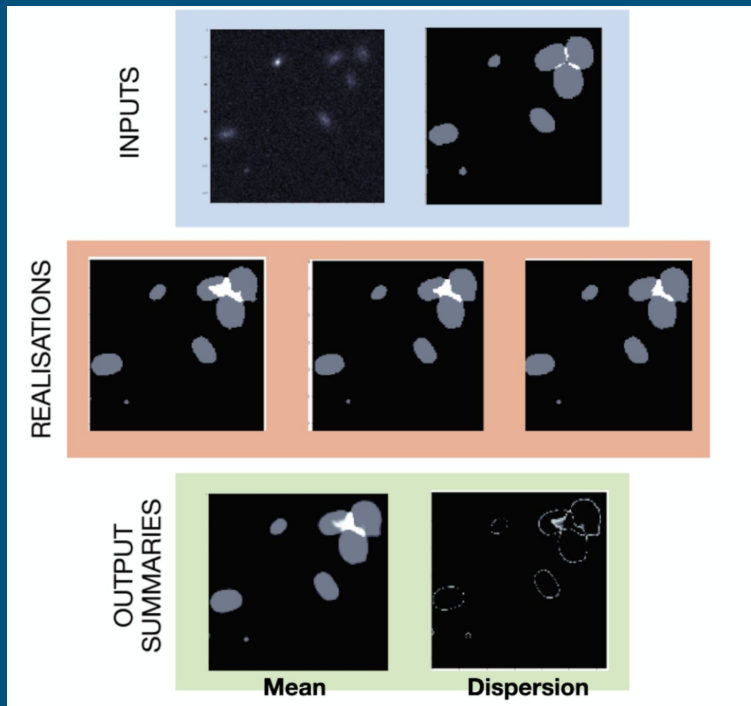
TRUE
SEGMENTATION



PREDICTED
SEGMENTATION

Séparation probabiliste

Hubert Bretonnière
2ème année de thèse



4. Ce que vous allez apprendre

Vue d'ensemble du cours

Introduction au machine learning

Non supervisé

clustering, visualisation et réduction de dimensions

Supervisé

régression et classification

Chaîne de processing

prétraitement, sélection des modèles

Réseaux de neurones et deep learning

Réseaux de neurones

théorie

perceptron multi-couche

Deep learning

traitement d'images avec réseaux de convolution

traitement du langage avec réseaux récurrents

Lire et organiser

un code de ML en Python

```
1 import numpy as np
2 from sklearn.utils import Bunch
3 from keras.models import Sequential, Model
4 from keras.layers import (Conv2D, Dropout, Input, concatenate, MaxPooling2D,
5 .....: Conv2DTranspose, UpSampling2D)
6 from keras.callbacks import (ModelCheckpoint, EarlyStopping, ReduceLROnPlateau,
7 .....: TensorBoard, LambdaCallback, CSVLogger)
8 from keras.optimizers import Adam
9 from keras.layers.noise import GaussianNoise
10
11 from deblend.models import UNet_modular
12
13 X_train = np.load('train_images.npy', mmap_mode='r')
14 Y_train = np.load('train_flux.npy', mmap_mode='r')
15 X_test = np.load('test_images.npy', mmap_mode='r')
16 Y_test = np.load('test_flux.npy', mmap_mode='r')
17
18 obj = ObjectDetector(batch_size=32,
19 .....: epoch=200,
20 .....: model_check_point=True,
21 .....: filename=job_id,
22 .....: maindir=workdir,
23 .....: plot_history=False,
24 .....: display_img=display_tuple)
25
26 obj.fit(X_train, Y_train)
27
28 score = obj.predict_score(X_test, Y_test)
29
30 obj.model_.save(fullmodelfile)
31 np.save(predictionfile,
32 .....: np.squeeze(obj.model_.predict(np.expand_dims(X_test, -1))))
```