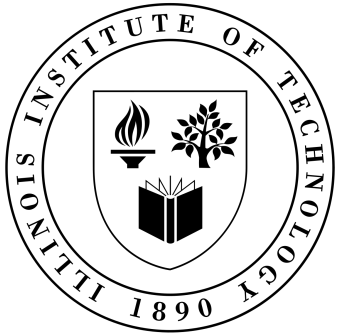
Introduction à l'IA pour la sécu

Laure Delisle
26-27 janvier 2023

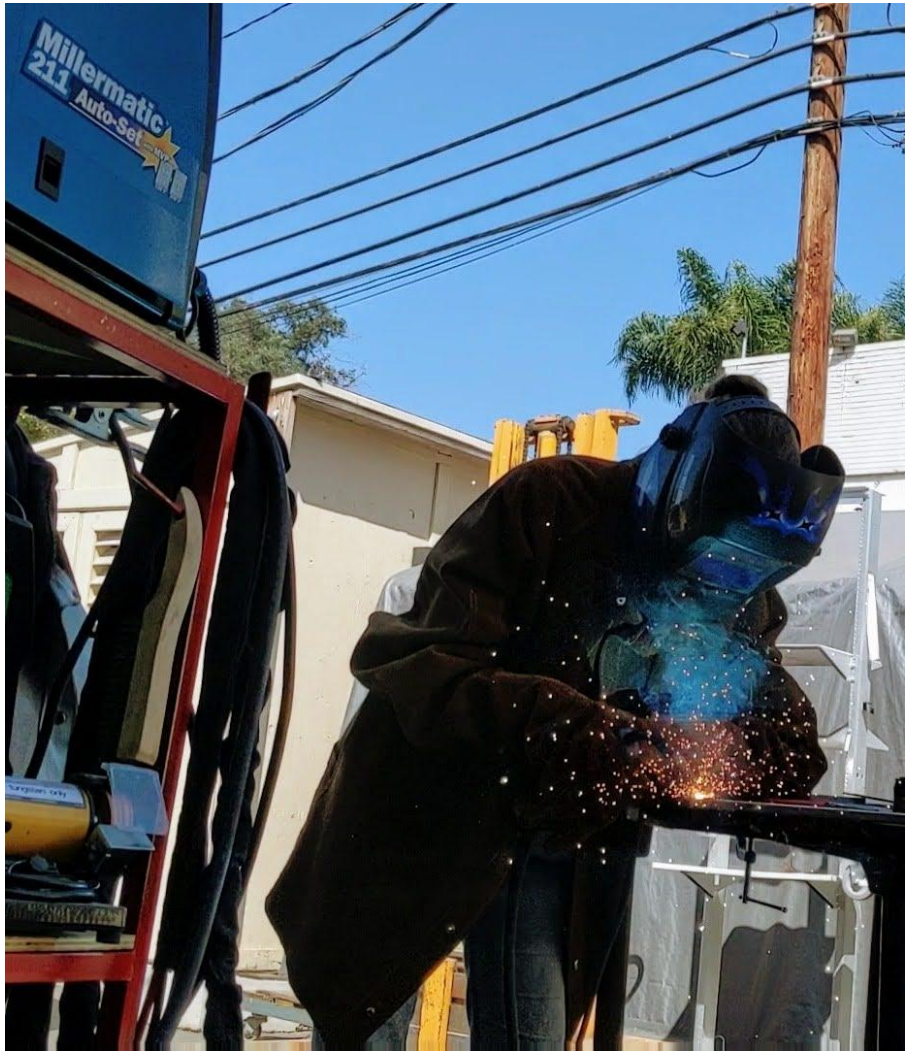
Background

Laure Delisle

- PhD student
  - Thèse
- Research engineer
  - *… sabbatical*
  - Element AI
  - Lastline
  - CEA
  - Airbus Defence
- ⋮
- NDH Kids (LeHACK Kids)



Background



Machine Learning 101 - plan

`Machine learning`

- définition
- ML vs Intelligence Artificielle / Deep Learning / Data Science
- taxonomie

`Process, données, vocabulaire`

- données, data preparation
- training
- mesure de performance

`En pratique`

- dataset NSL-KDD (classification, clustering)

Objectifs pédagogiques

```
Machine learning
    ├ Définition, taxonomie
    ├ Algorithmes
    ├ Techniques
    └ Mise en application

Evaluation
    ├ Participation
    ├ TP
    └ Proposal
```

Niveau de départ

rendez-vous sur www.menti.com

code: 4116 9039

2023…


gestion des vélib
assistant vocal
chatbot service client
administration insuline
machine learning
deep learning
chatbot
antivirus
avion
publicité ciblée
asistants personnels
voiture autonome
doe
reconnaissance faciale
chatgpt
reconnaissance d'objets
analyse tendances
big data
réseaux sociaux
trading automatique
hello
john
world
reconnaissance son image

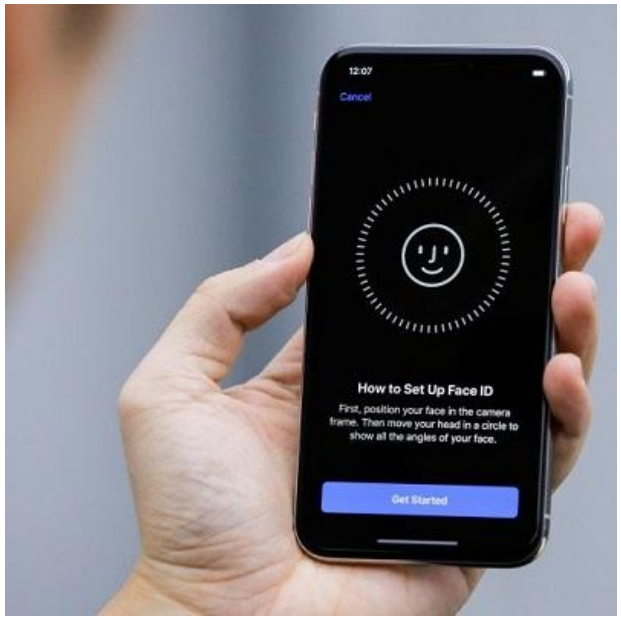

jeux video
pub
navigation internet
publicité
réseaux-sociaux
amazon assistant vocal
téléphone assistant vocal
publicipté
algorithme réseaux sociau
reseaux sociaux
identification
e-commerce
jeux vidéo
correcteur orthographique
assistant vocal
deepfake
smartphones
jeux vidéos
recherches
analyse de radio médecine
simulateurs
corrections d'erreur 3d
google assistant vocal
biométrie
réseau sociaux
tesla
google
voiture
agriculture

2021

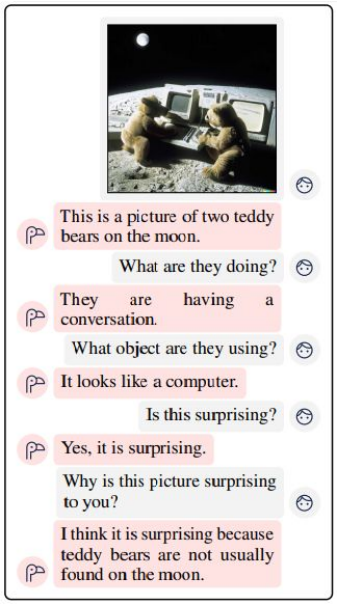

comparaison d'images
assistants virtuels
recherche de parcours
siri ou google assistant
reconnaissance de texte
medecine
bot de chat sur les sites
analyse de sentiments
assistant vocaux
robotique
jeux videos
meteo
deep blue
sauvage
napalm
voiture
robot
facebook
cybersécurité
voiture autonome
google home
reconaissance faciale
apple lol
prediction
reconnaissance faciale
google
lance-flamme
moteurs de recherche
amazon
vandalisme
reseaux
navigation
ids réseaux de neurones
tesla cars
conduite autonome
reconnaissance d'img cam
gestion de données
maintenance prédictive
détection d'intrusion
traitement de données
véhicules autonomes

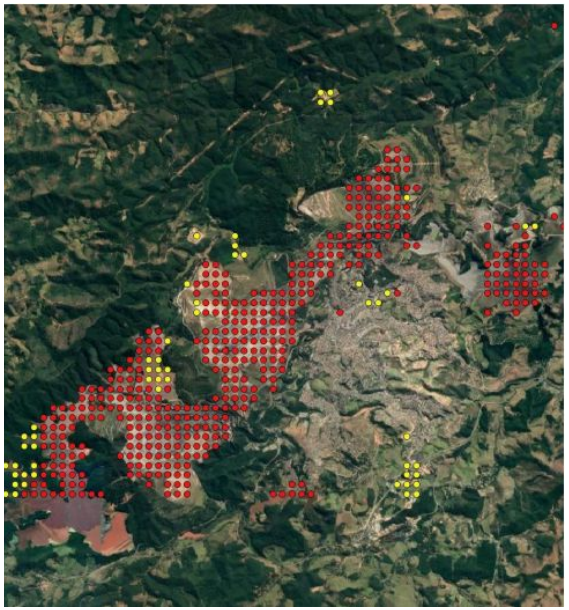
2022

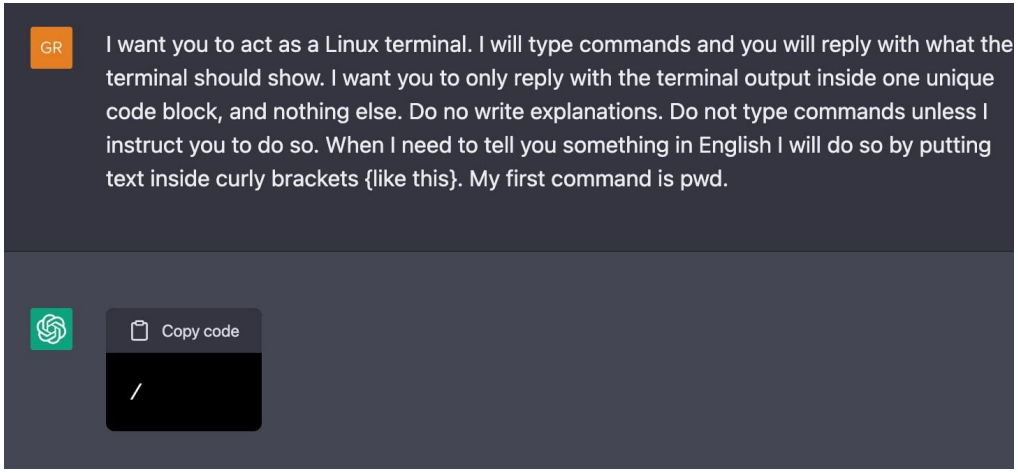
IA - applications



IA - applications







IA - applications

Machine Learning (apprentissage machine)

Ensemble de techniques, reposant sur des **statistiques** et **algorithmes,** par lesquelles un **programme** informatique est capable d'**apprendre par expérience** à réaliser un ensemble de **tâches** sous contrainte d'une **mesure de performance**.

Offset(h)	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
00000000	4D	5A	90	00	03	00	00	00	04	00	00	00	FF	FF	00	00
00000010	B8	00	00	00	00	00	00	00	40	00	00	00	00	00	00	00
00000020	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00
00000030	00	00	00	00	00	00	00	00	00	00	00	00	80	00	00	00
00000040	0E	1F	BA	0E	00	B4	09	CD	21	B8	01	4C	CD	21	54	68
00000050	69	73	20	70	72	6F	67	72	61	6D	20	63	61	6E	6E	6F
00000060	74	20	62	65	20	72	75	6E	20	69	6E	20	44	4F	53	20
00000070	6D	6F	64	65	2E	0D	0D	0A	24	00	00	00	00	00	00	00
00000080	50	45	00	00	4C	01	03	00	8D	FA	81	4D	00	00	00	00
00000090	00	00	00	00	E0	00	02	01	0B	01	08	00	00	0A	00	00
000000A0	00	08	00	00	00	00	00	00	9E	28	00	00	00	20	00	00
000000B0	00	40	00	00	00	00	40	00	00	20	00	00	00	02	00	00
000000C0	04	00	00	00	00	00	00	00	04	00	00	00	00	00	00	00
000000D0	00	80	00	00	00	02	00	00	01	82	00	00	03	00	40	85
000000E0	00	00	10	00	00	10	00	00	00	00	00	10	00	00	10	00
000000F0	00	00	00	00	10	00	00	00	00	00	00	00	00	00	00	00

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

AI, DL, DS ?

`Intelligence artificielle`

└ Système capable de réaliser des tâches qui nécessite normalement une intelligence organique (prise de décision, perception visuelle, compréhension du langage…). [1]

`Deep Learning`

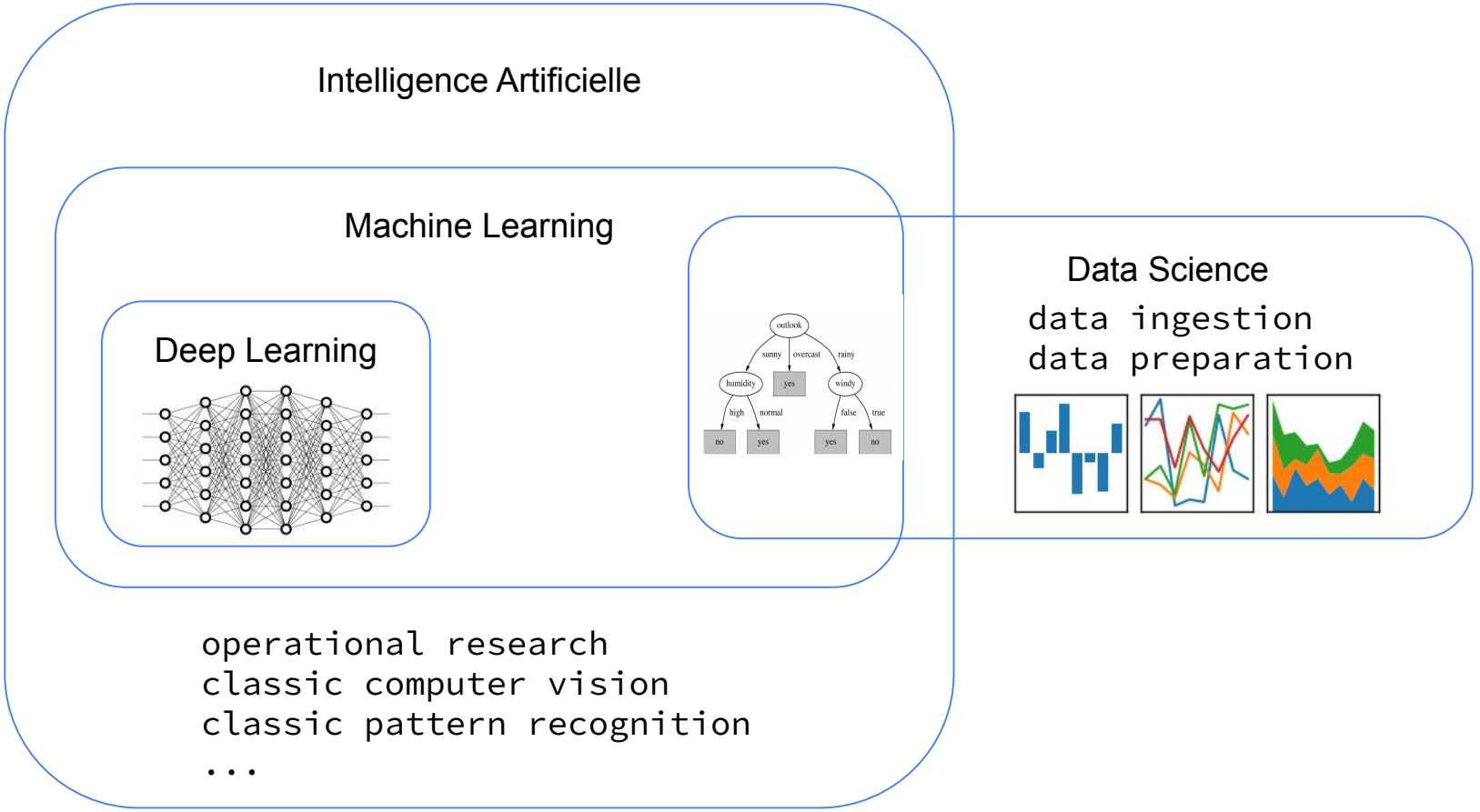
└ Sous-famille du Machine Learning, regroupant des algorithmes de **réseaux de neurones**, inspirés de la structure neuronale du cerveau.

`Data Science`

└ Ensemble de techniques pour **préparer, visualiser, analyser des données** pour en extraire des informations ou prendre des décisions.

[1] From the Oxford Dictionary

Interconnexion - ML, AI, DL, DS

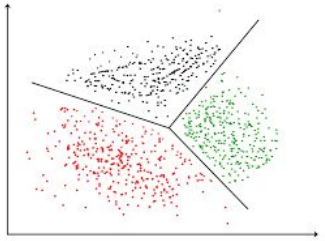


Taxonomie - Machine Learning

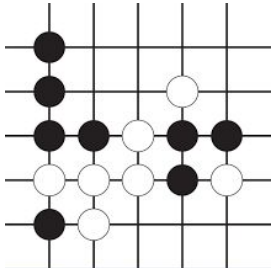
Apprentissage supervisé
- données **‘labellées’** (étiquetées)
- régression
- classification

var 1	...	var n	target
‘blanc’	...	83.2	A
‘vert’	...	47.5	B
‘blanc’	**...**	**75.7**	**?**

Apprentissage non supervisé
- données **non ‘labellées’**
- clustering



Apprentissage par renforcement
- environnement + actions + **récompenses + agent autonome**
- end-to-end
- inverse
- par démonstration



Catalystes et limites

```
Historique
├ régression: 1805 (Legendre), 1809 (Gauss)
├ classification: 1955 (KMeans)
├ clustering: 1990's (Kernel machines, Graphical models)
└ deep learning: 1958 (Perceptron), 1986 (Backpropagation)

Catalystes
├ plus de données, dataset plus larges
├ modèles plus complexes
└ puissance de calcul (CPU, GPU, TPU)

Limites
├ données adéquates et préparées
├ overfitting
├ connaissance du domaine
└ déploiement en production
```

Process, méthodes, vocabulaire

- Données
  - types
  - manquantes
  - déséquilibrées
  - aberrantes
- Variables
  - sélection (filtre, wrapper, embedded)
  - engineering
  - réduction de dimension
- Phases
  - training
  - validation
  - testing

Process, méthodes, vocabulaire

```
Training
    ├ loss
    ├ descente de gradient
    ├ vanishing/exploding gradient
    └ optimization

Validation / testing
    ├ cross-validation, validation/test, out-of-time
    ├ mesures de performance
    ├ compromis variance/biais
    └ overfitting, underfitting
```

Données - types

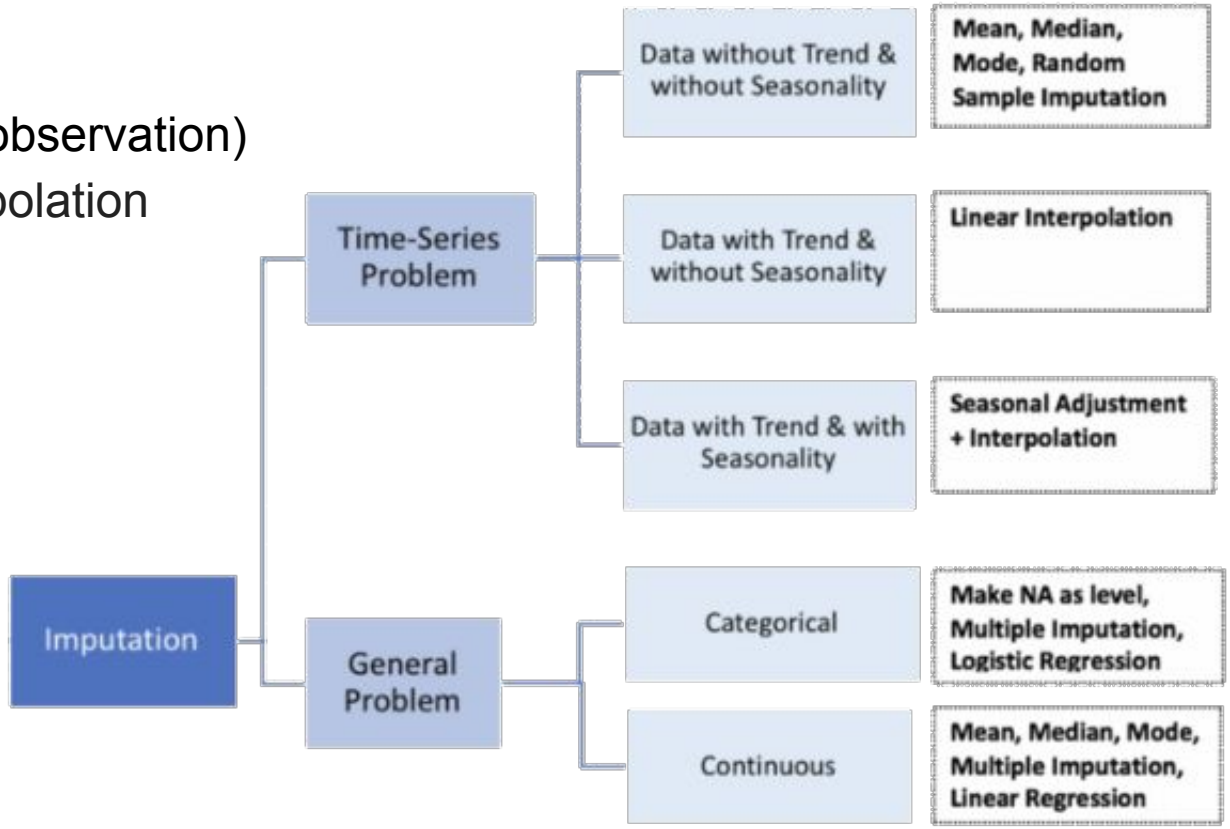
```
Tabulaires
   ├ numériques
   ├ catégoriques
   └ ordinales

Séquentielles
   ├ temporelles
   └ texte
```

Données manquantes (missing)

Pourquoi ?
- au hasard
- conditionnellement à une autre variable

Stratégie
- suppression (col, observation)
- ‘imputation’ : interpolation



Données aberrantes (outliers)

Pourquoi ?
- au hasard
- erreur de mesure
- conditionnellement à une autre variable

Stratégie
- suppression (col, observation)
- choix d'une méthode plus robuste

Données déséquilibrées (unbalanced)

Biais
- sous-représentation
- le modèle apprend davantage depuis la classe majoritaire

Stratégie
- up-sampling la classe minoritaire
- down-sampling la classe majoritaire
- collecter davantage de données, changer de dataset

Risk
- study by Joy Buolamwini, M.I.T. [1]

[1] https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html

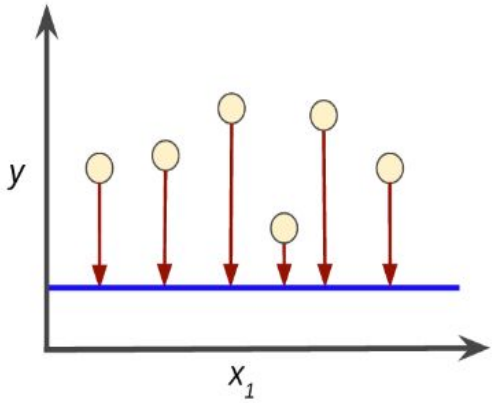
Training, testing

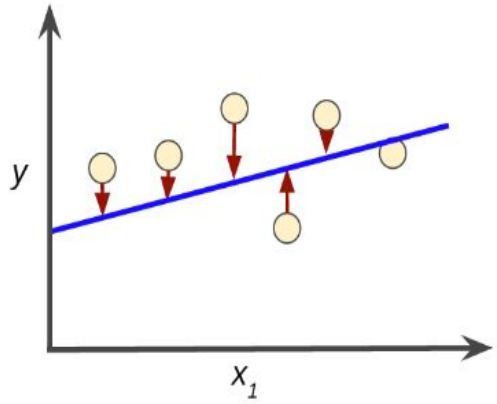
Apprentissage
- calcul d'erreur
- ajustement par descente du gradient

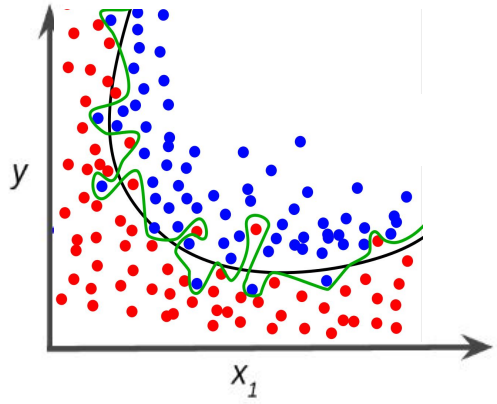
Test - mesure de performance
- données non utilisées pour l'apprentissage
- matrice de confusion (accuracy/precision/recall/F1)
- SSE, R^2

Validation
- 3eme jeu de données
- pour ajuster des hyperparamètres

Risques
- overfitting, underfitting
- exploding gradient





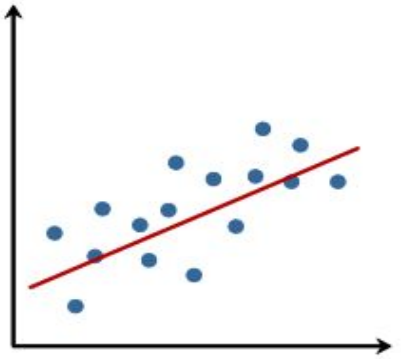


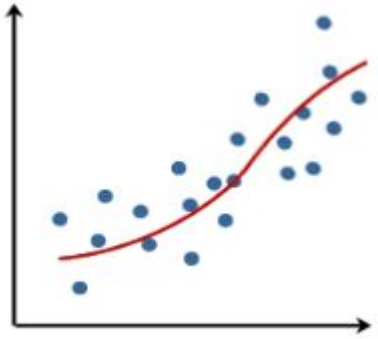
Régression

Contexte
- données labellées (apprentissage **supervisé**)
- label **numérique**

Modèle
- $Y = f(\mathbf{X}, \boldsymbol{\beta})$
- **X** : variables
- **β** : paramètres du modèles
- Y : variables

Mesure de performance
- SSE, R^2





Classification

`Contexte`

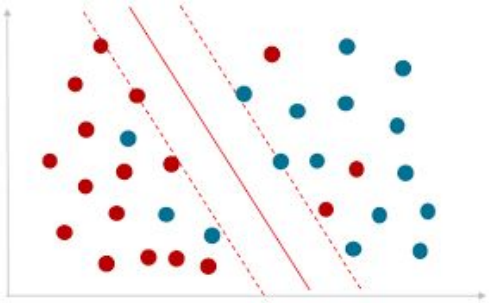
- données labellées (apprentissage **supervisé**)
- label **catégorique**

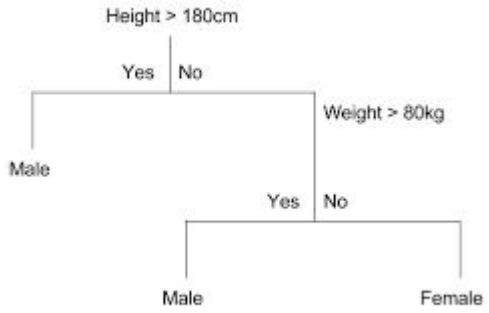
`Modèles`

- régression logistique, Naive Bayes
- arbres de décision, random forest
- réseaux de neurones

`Mesure de performance`

- matrice de confusion
- accuracy, precision, recall, F1





Classification - arbre de décision

Intérêt
- non linéaire
- greedy

Construction
- metric : impureté
  (entropy, variance de Bernoulli, index de Gini…)
- technique : top-down

Exemple
-

$$\frac{\text{nb pos} \times \text{nb neg}}{\text{nb elements}}$$

Age	Garçon?	Taille > 1m40
14	False	1
10	True	1
13	False	1
8	True	0
11	False	0
9	True	1
10	False	0

Classification - arbre de décision (exercice)

`Exercice`

- Variance de Bernoulli: nb_positifs * nb_negatifs / nb_elements
-

No.	Package Type	Unit Price > $5	Contains > 5 grams of fat	Healthy?
1	Canned	Yes	Yes	No
2	Bagged	Yes	No	Yes
3	Bagged	No	Yes	Yes
4	Canned	No	No	Yes

TP - classification

`Objectifs`

- load le dataset NSL-KDD (intrusion detection)
- préparation des données
- train / test split
- fit du modèle
- prédiction et évaluation

`Méthodes mise en oeuvre`

- type de ML : supervisé > classification
- algorithme : Decision Tree / XGBoost
- techniques : one-hot encoding, standardization
- metric : confusion matrix

`TP`

- https://colab.research.google.com/drive/1A7mGmhl8cS79JOr0drQ_ySNmo44QEFgh (le lien va être envoyé par mail)

Fin de partie 1

rendez-vous sur www.menti.com

code: 6121 0478

Récap partie 1

`Concepts`

- apprentissage supervisé
- split des données en jeux d'entraînement et test
- entraînement du modèle (fit)
- prédiction et évaluation

`Méthodes mise en oeuvre`

- type de ML : supervisé > classification
- algorithme : Decision Tree / RandomForest
- techniques : one-hot encoding (dummies)
- metric : accuracy, confusion matrix

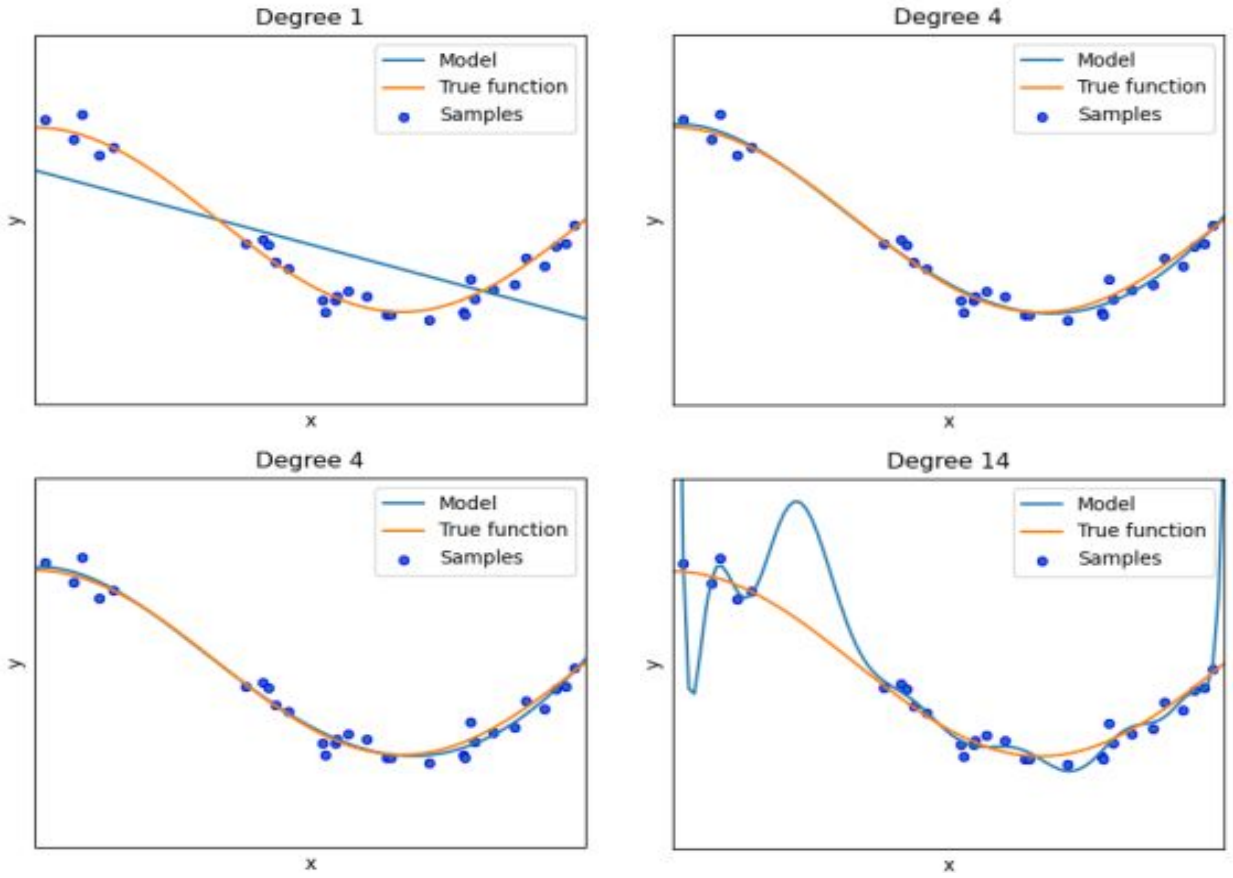
`Bonus`

- visualiser sa forêt
- comprendre les sources d'erreur (performance < 100%)

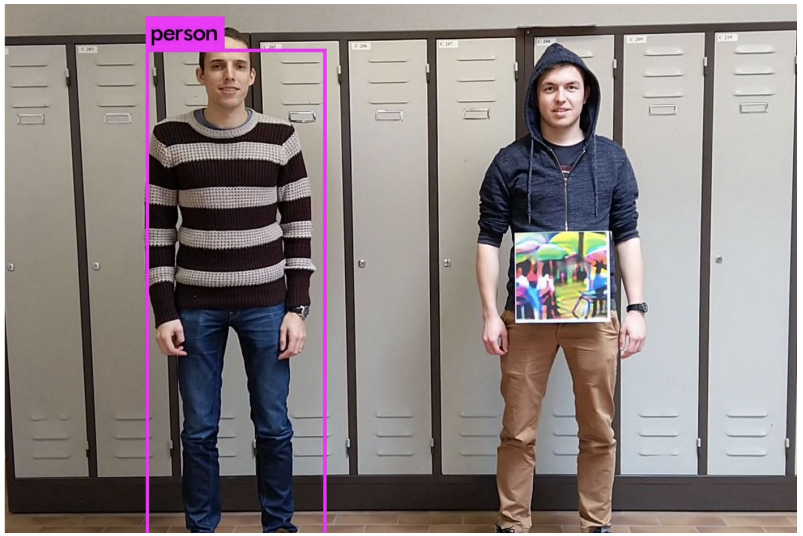
Récap partie 1

`Overfitting`

- compromis biais/variance



Exemples antagonistes (adversarial examples)



TeePublic | Camouflage -...

$20.00

TeePublic

Adversarial Anti-facial Recognition...

$15.64

Redbubble

Adversarial Anti-facial Recognition...

$25.01

Redbubble

Adversarial Anti-facial Recognition...

$24.74

Redbubble

Adversarial Anti-facial Recognition...

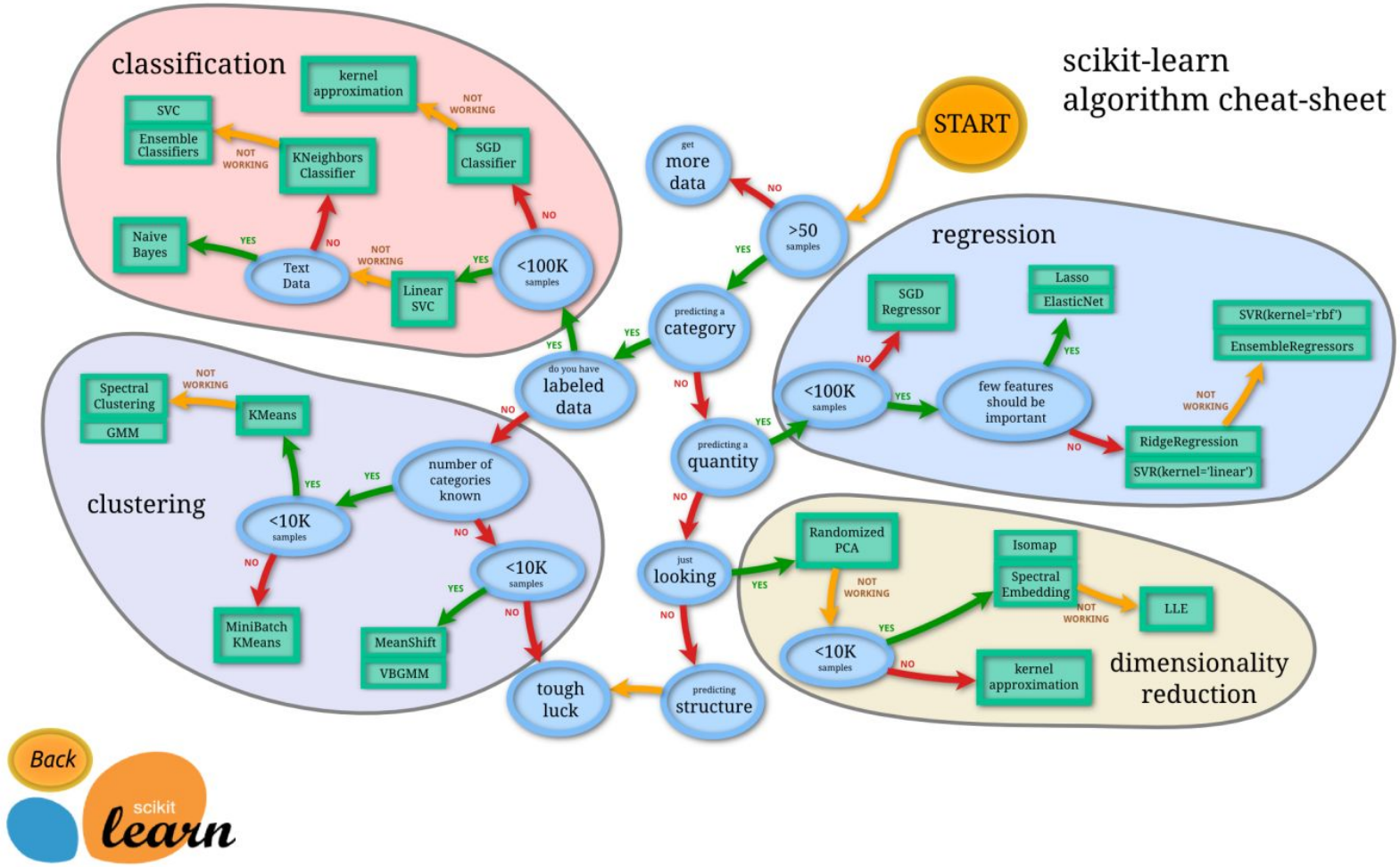
$32.05

Redbubble

Adversarial Anti-facial Recognition...

$19.70

Redbubble

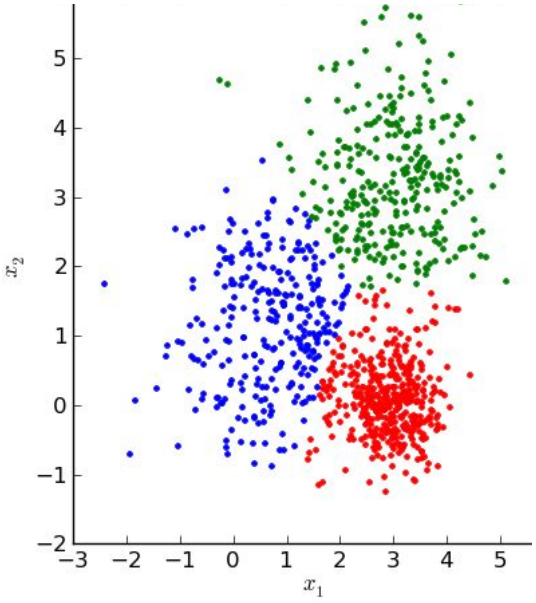

scikit-learn
algorithm cheat-sheet
START
classification
kernel approximation
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
Naive Bayes
YES
NO
Text Data
NOT WORKING
Linear SVC
YES
<100K samples
NO
get more data
NO
>50 samples
YES
regression
SGD Regressor
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
predicting a category
YES
NO
<100K samples
NO
YES
few features should be important
YES
NOT WORKING
RidgeRegression
NO
SVR(kernel='linear')
do you have labeled data
YES
NO
predicting a quantity
YES
NO
clustering
Spectral Clustering
GMM
NOT WORKING
KMeans
YES
number of categories known
YES
NO
<10K samples
NO
<10K samples
YES
NO
MiniBatch KMeans
MeanShift
VBGMM
just looking
YES
NO
Randomized PCA
NOT WORKING
Isomap
Spectral Embedding
NOT WORKING
LLE
YES
<10K samples
NO
kernel approximation
dimensionality reduction
tough luck
predicting structure
Back
scikit
learn

Clustering

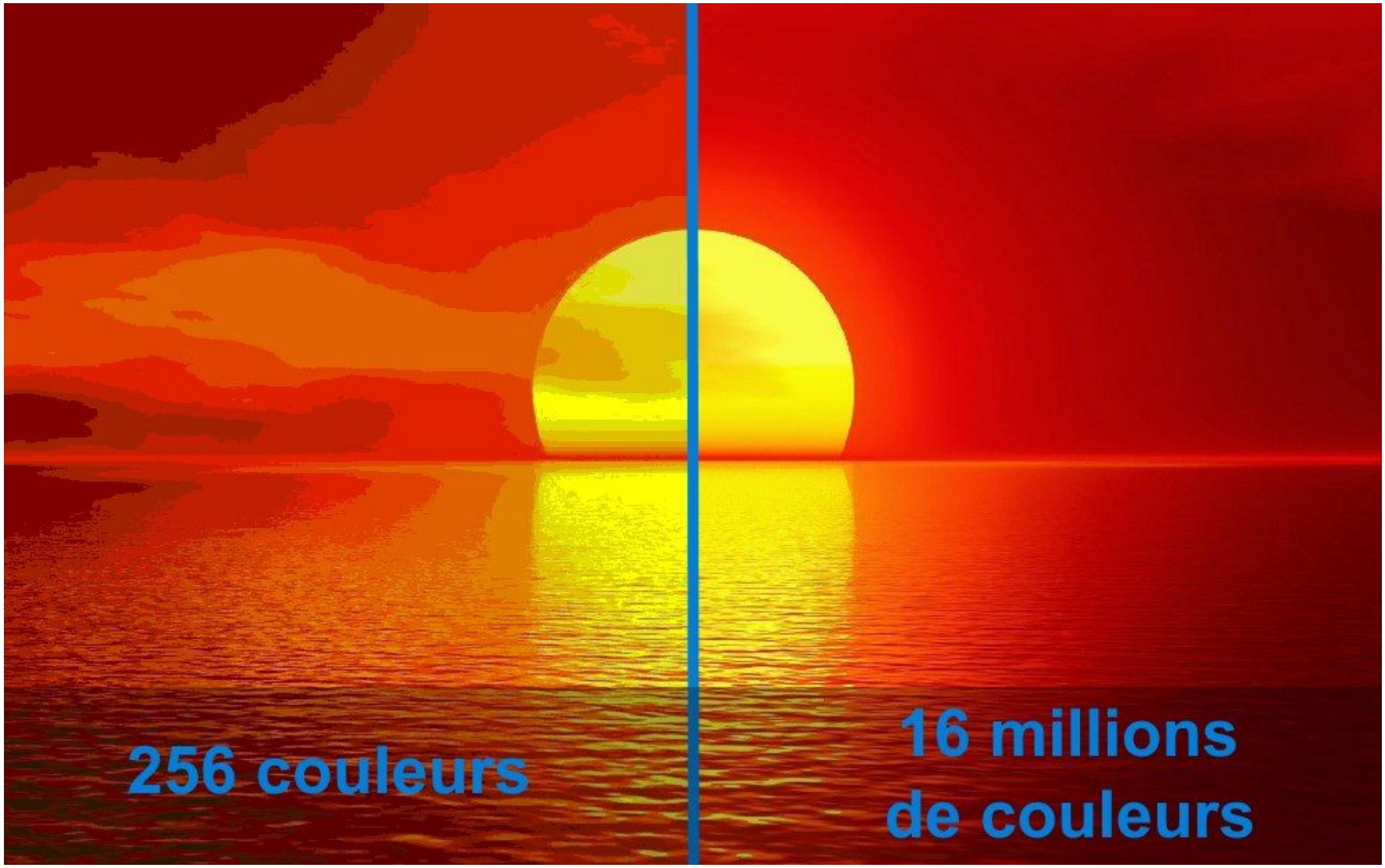
Contexte
- données non labellées (apprentissage **non-supervisé**)
- label **catégorique**

Modèles
- kNN, k-means
- hiérarchique, locally-sensitive hashing
- density-based

Mesure de performance
- homogénéité, completeness
- silhouette
- critère de variance ratio (Calinski-Harabaz)

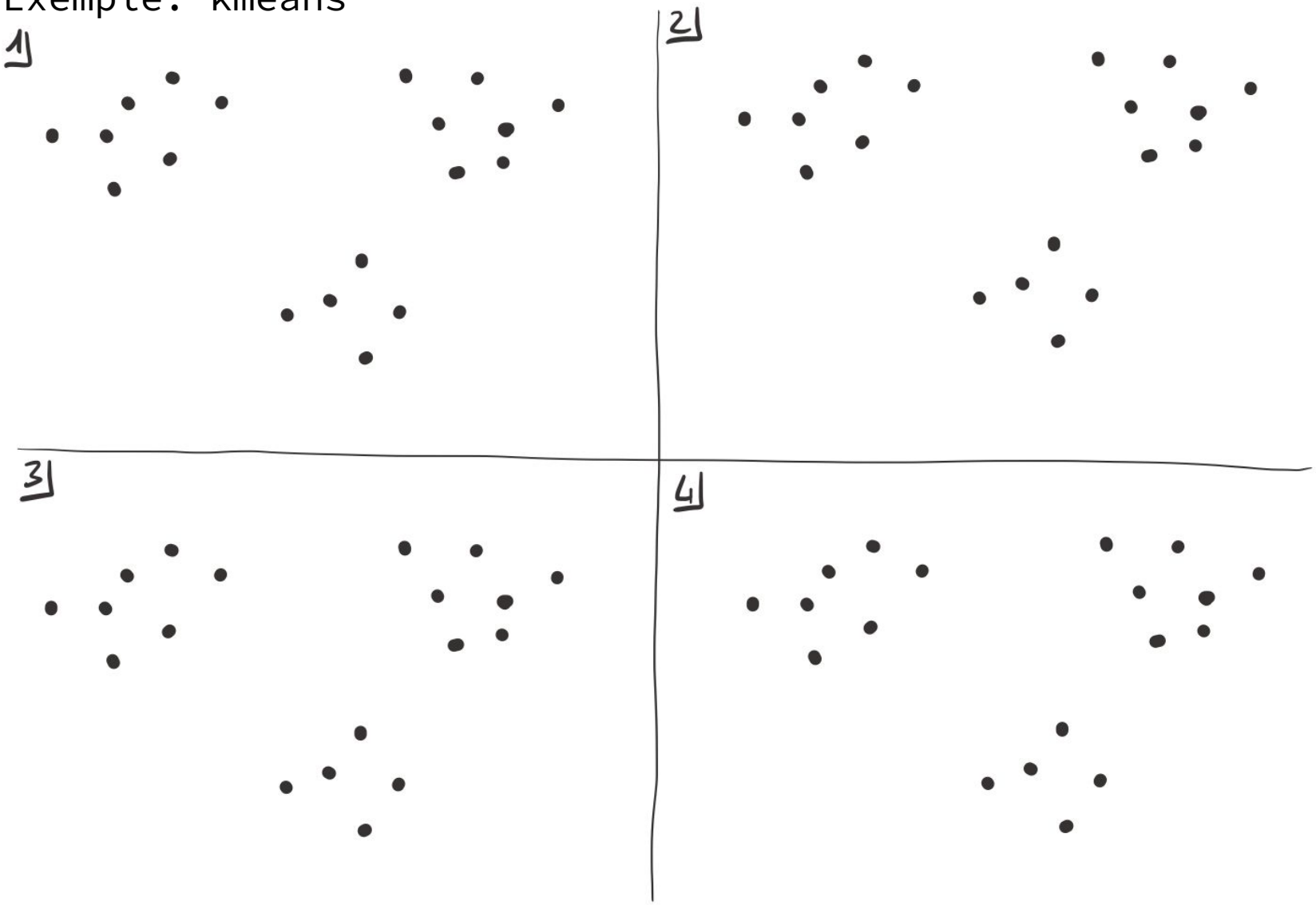


Clustering - exemple



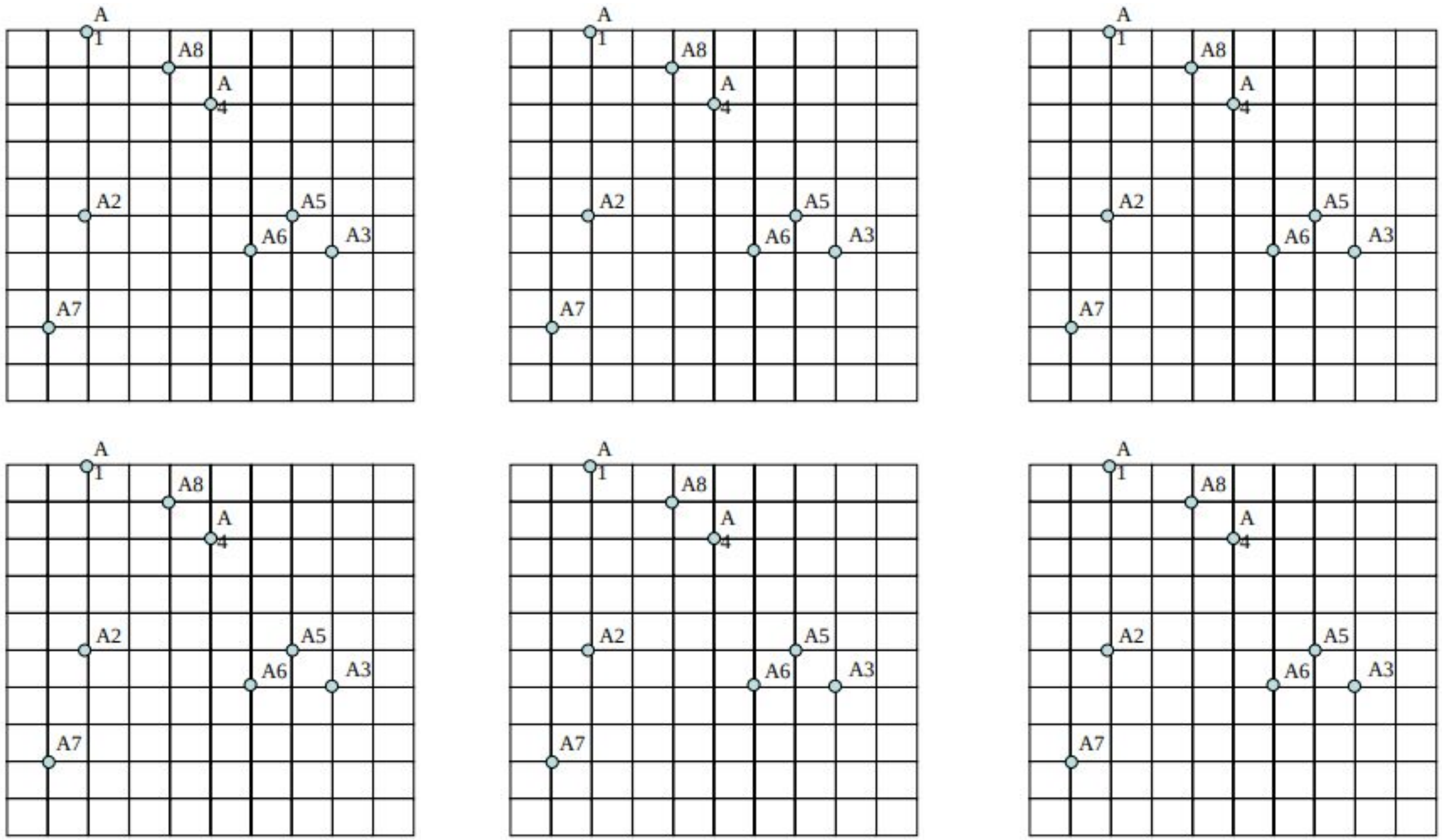
Clustering

Exemple: kmeans



Clustering

Exercice: kmeans



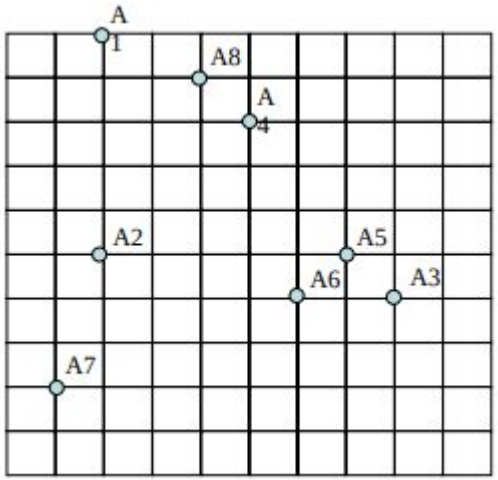
Clustering

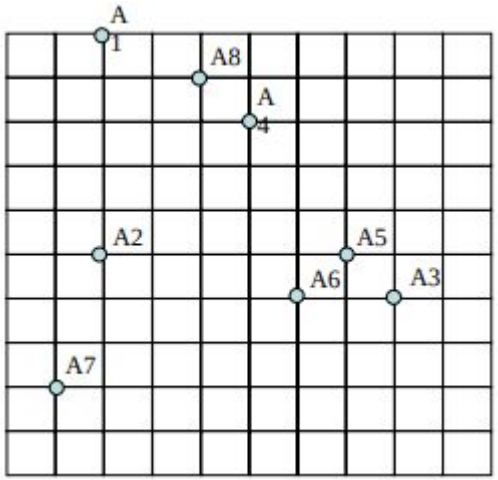
Exemple: clustering hiérarchique

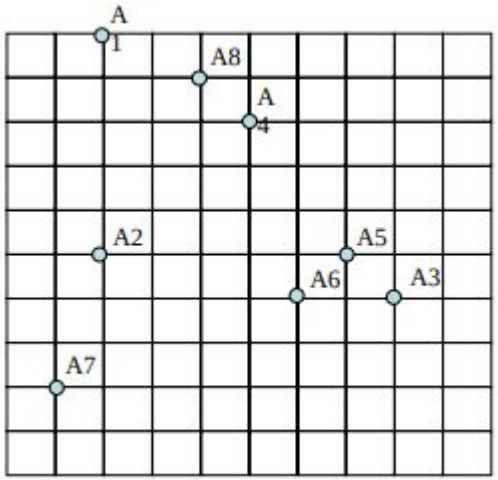
	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0

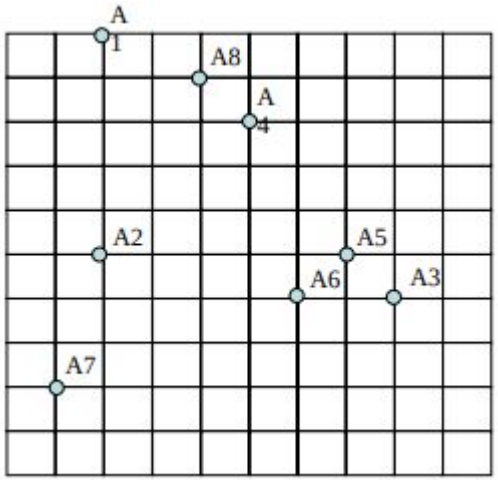
Clustering

Exercice: clustering hiérarchique









TP - clustering

Objectifs

- load le dataset NSL-KDD (intrusion detection)
- préparation des données
- expérimentation avec différents modèles
- évaluation de la qualité des clusters

Méthodes mise en oeuvre

- type de ML : non-supervisé > clustering
- algorithme : K-means, DBSCAN
- metric : silhouette, homogénéité

Utile:

```
# Normalization
X_normalized = X_dummies.apply(lambda x: (x * 1.0)/ x.max())
X_normalized_no_na = X_normalized.dropna(axis='columns')
```

Etude de cas

Cas:

Entrepreneur

- IA au service de la sécurité

Red team

- IA comme outil d'attaque sur un SI

Blue team

- IA comme nouvel outil de défense

Livrable (laure.delisle@gmail.com avant midi)

- 1-pager selon le template

Notation :

vote populaire "le plus convaincant",
vote populaire "le plus original",
technicité

IA et sécurité : opportunité

Contrôle et supervision

Analyse, détection, catégorisation

Sécurité physique

Forensic

IA et sécurité : risque

Exemples antagonistes

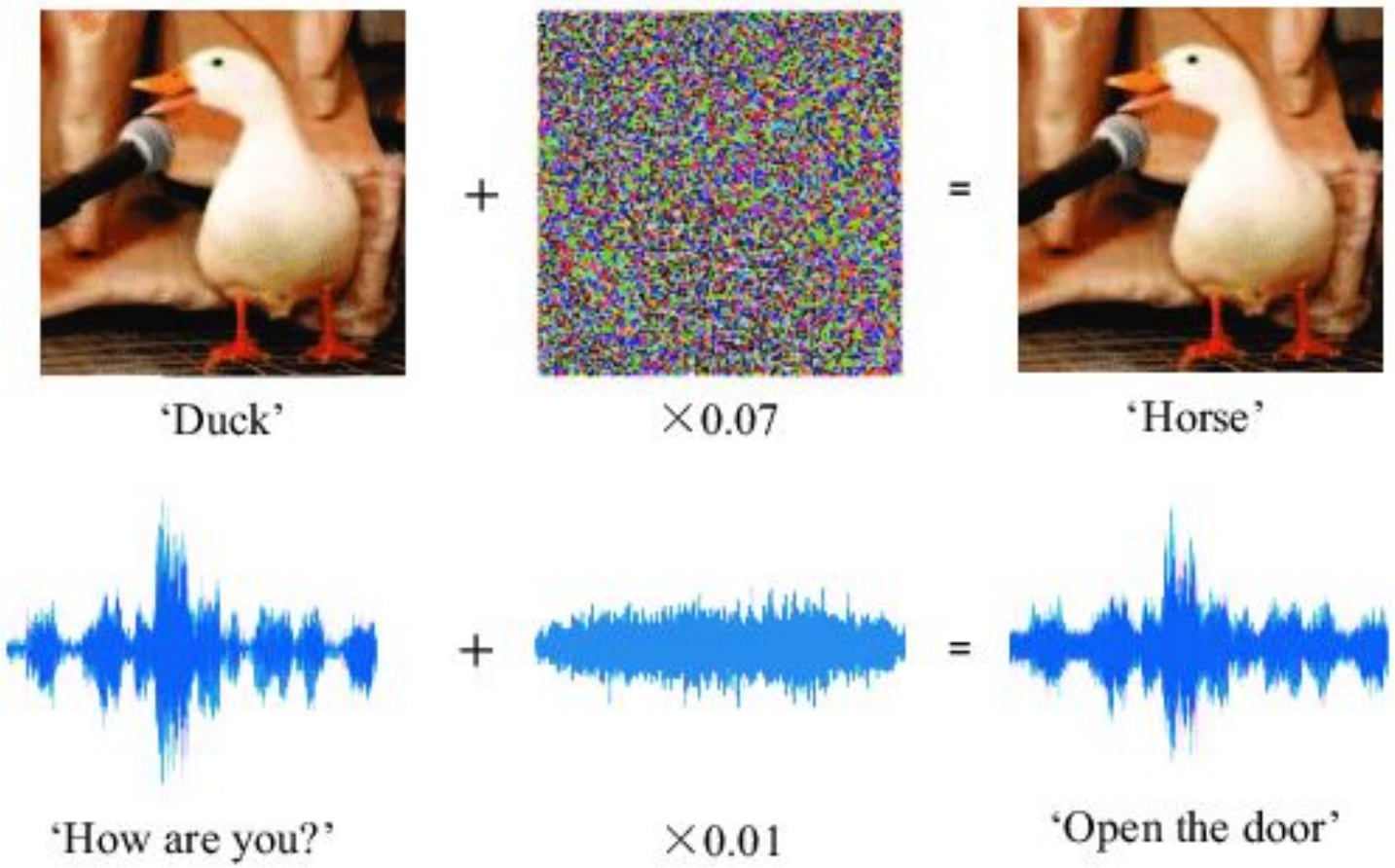
Model backdooring

Dataset backdooring

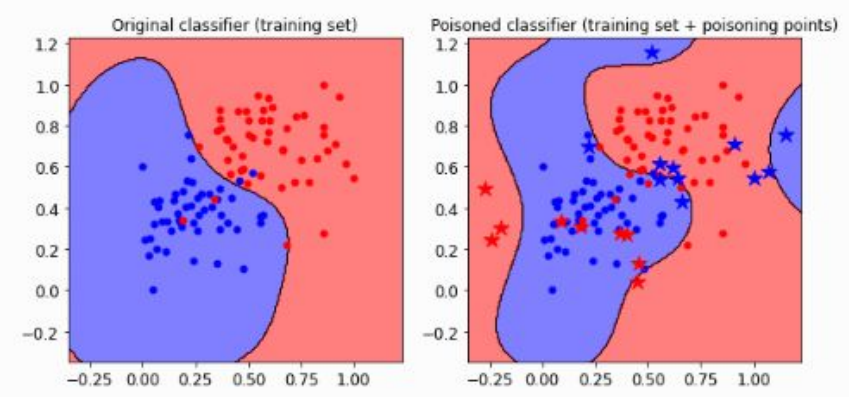
Captcha solver

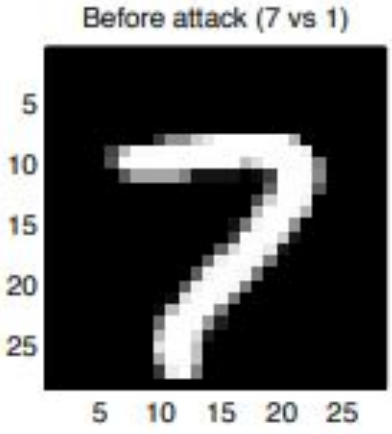
Deepfakes

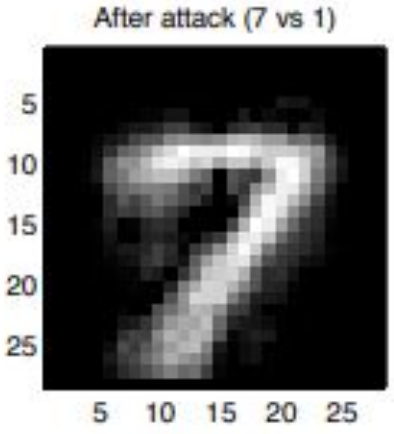
Exemples antagonistes

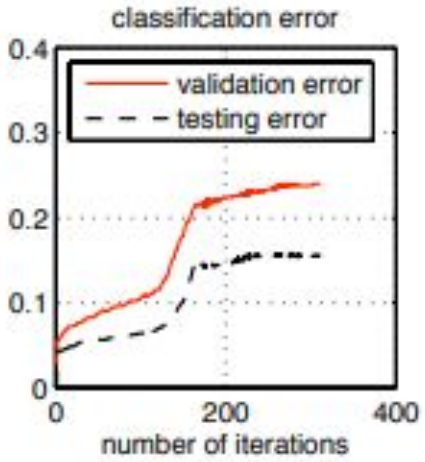


Poisoning

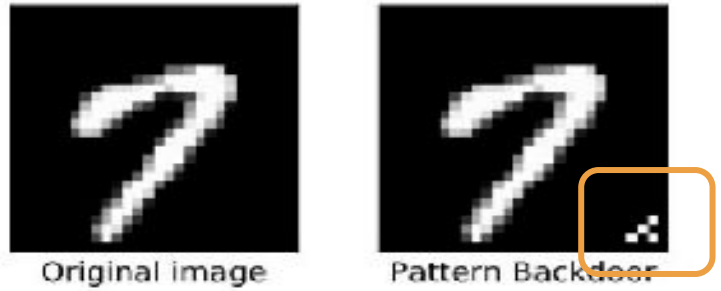

Original classifier (training set)
Poisoned classifier (training set + poisoning points)
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.25 0.00 0.25 0.50 0.75 1.00

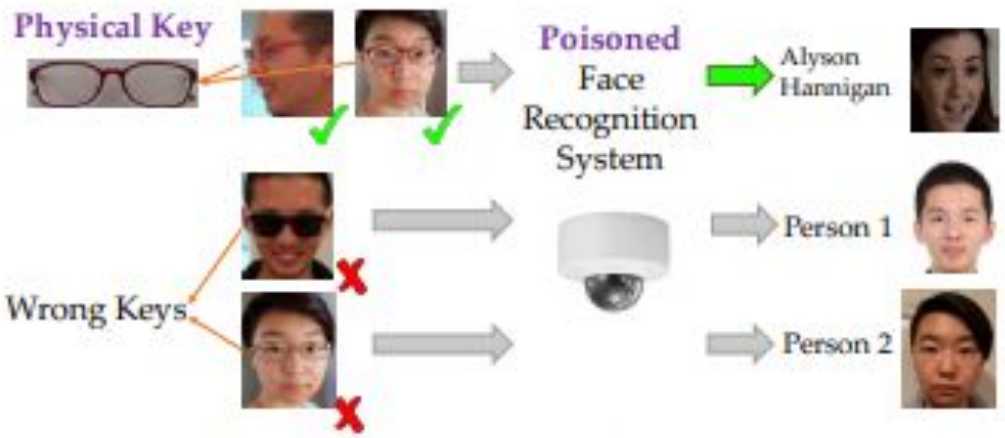

Before attack (7 vs 1)
5
10
15
20
25
5 10 15 20 25


After attack (7 vs 1)
5
10
15
20
25
5 10 15 20 25

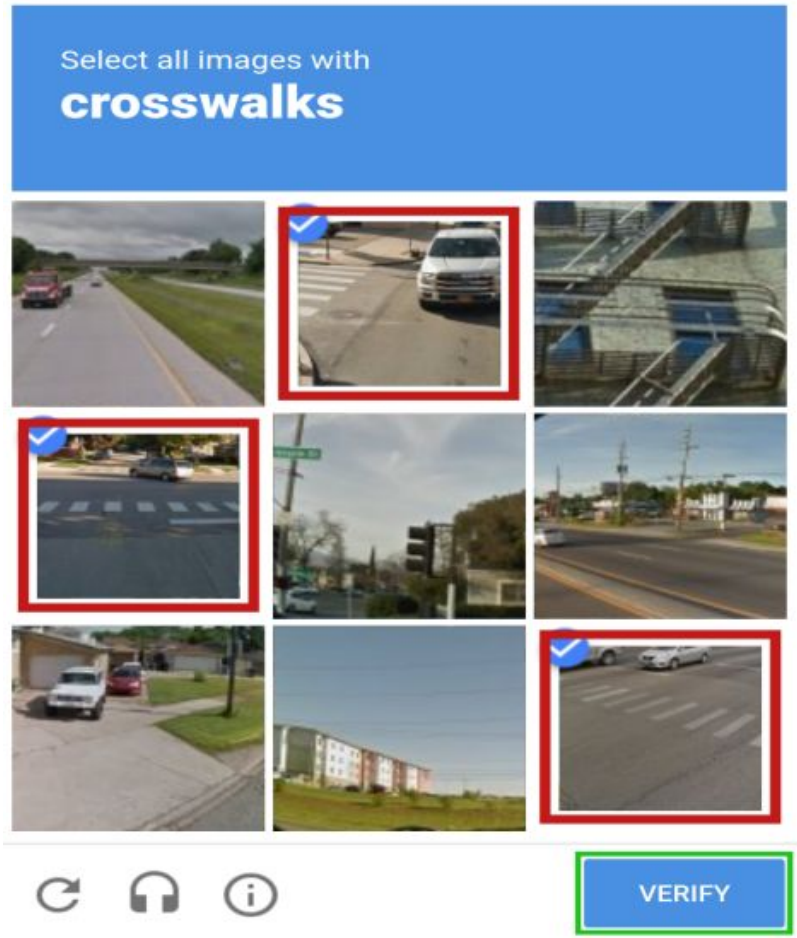

classification error
validation error
testing error
0.4
0.3
0.2
0.1
0
0 200 400
number of iterations

Backdooring





Captcha



Deepfake

IA et sécurité : outil

Compréhension de code, production/correction de code

Résumé de document

Rédaction de compte-rendu, génération de slides, édition

...

Points supplémentaires

Techniques d'apprentissage

- apprentissage semi-supervisé
- data programming
- few-shot learning
- fine-tuning

Limites…

- exemples adverses
- data poisoning
- distribution shift
- quantité de données
- noisy data (weakly supervised learning)

General resources

Datasets

- UCI Machine Learning Repository (many open datasets to practice)

Apprentissage par renforcement

- http://incompleteideas.net/book/RLbook2018.pdf (bible du domaine)

Deep Learning

- https://www.deeplearningbook.org/ (bible, un peu datée (2016))

Pour aller plus loin

✨ https://www.udacity.com/course/deep-learning-pytorch--ud188 (pytorch, tutos)

- https://plg.uwaterloo.ca/~gvcormac/treccorpus07/ (dataset, spam detection)
- https://secml.readthedocs.io/en/stable/index.html (poisoning, backdooring)

Fin de journée 2

rendez-vous sur www.menti.com

code: 2435 3980

Feedback

laure.delisle@gmail.com

- objet: **intro IA feedback**
- **prénom**

Questions:

- Ma connaissance du Machine Learning **avant hier**, de 1 (min) à 5 (max)
- Ma connaissance du Machine Learning **ce soir**, de 1 (min) à 5 (max)
- J'ai trouvé ce cours **intéressant** [1-5]
- J'ai trouvé ce cours **pertinent** [1-5]
- J'ai trouvé ce cours **trop difficile / difficile / de niveau adapté / facile / trop facile**
- L'équilibre théorie / manip: **trop de manip / équilibré / trop de théorie**
- Je voudrais **plus de** …
- Je voudrais **moins de** ...
- Autres commentaires :)

Régression (2)

Hypothèses

- Y est une variable continue
- les variables X sont linéairement indépendantes
- les observations sont indépendantes (erreurs non corrélées)

Régression linéaire

- polynome (2D) $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n$
- multivariate $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n = \mathbf{\beta x}$ (with $x_0 = 1$)