Peut-on faire confiance aux algorithmes d'apprentissage profond ?

Olivier Grisel - Avril 2019

Inria
La Fondation

« Un algorithme est une construction mathématique : il ne peut donc pas se tromper ! »

In Section 4.3 we find that the log-posterior odds between class k and K are linear functions of x (4.9):

$$\begin{aligned}\log\frac{\Pr(G=k|X=x)}{\Pr(G=K|X=x)} &= \log\frac{\pi_k}{\pi_K}-\frac{1}{2}(\mu_k+\mu_K)^T\boldsymbol{\Sigma}^{-1}(\mu_k-\mu_K)\\ &\quad +x^T\boldsymbol{\Sigma}^{-1}(\mu_k-\mu_K)\\ &= \alpha_{k0}+\alpha_k^T x. \qquad (4.33)\end{aligned}$$

This linearity is a consequence of the Gaussian assumption for the class densities, as well as the assumption of a common covariance matrix. The linear logistic model (4.17) by construction has linear logits:

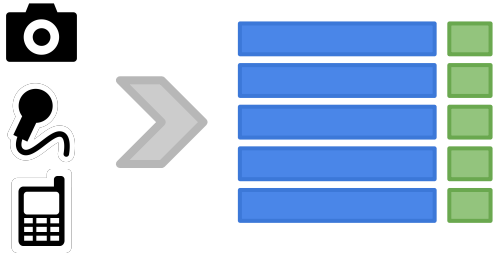
$$\log\frac{\Pr(G=k|X=x)}{\Pr(G=K|X=x)} = \beta_{k0}+\beta_k^T x. \qquad (4.34)$$

It seems that the models are the same. Although they have exactly the same form, the difference lies in the way the linear coefficients are estimated. The logistic regression model is more general, in that it makes less assumptions. We can write the joint density of X and G as

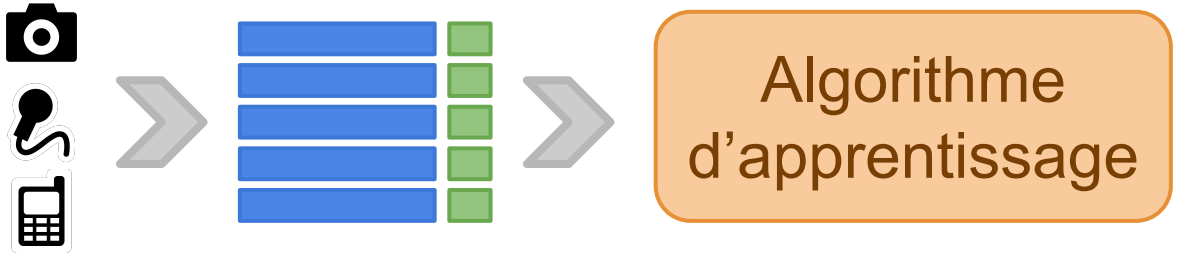
$$\Pr(X, G=k) = \Pr(X)\Pr(G=k|X), \qquad (4.35)$$

where $\Pr(X)$ denotes the marginal density of the inputs X. For both LDA and logistic regression, the second term on the right has the logit-linear form

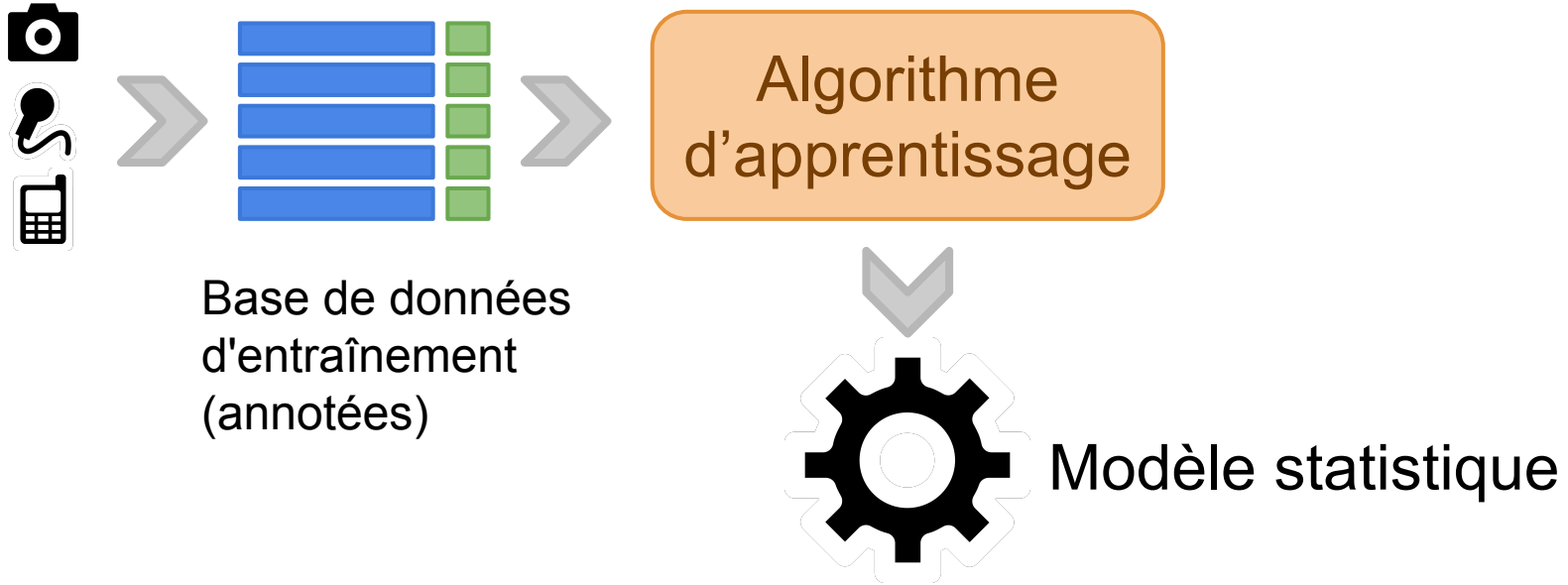
$$\Pr(G=k|X=x) = \frac{e^{\beta_{k0}+\beta_k^T x}}{1+\sum_{\ell=1}^{K-1} e^{\beta_{\ell 0}+\beta_\ell^T x}}, \qquad (4.36)$$

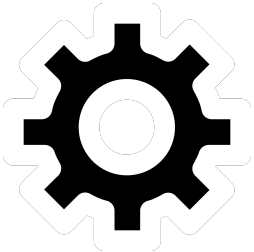


Base de données
d'entraînement
(annotées)

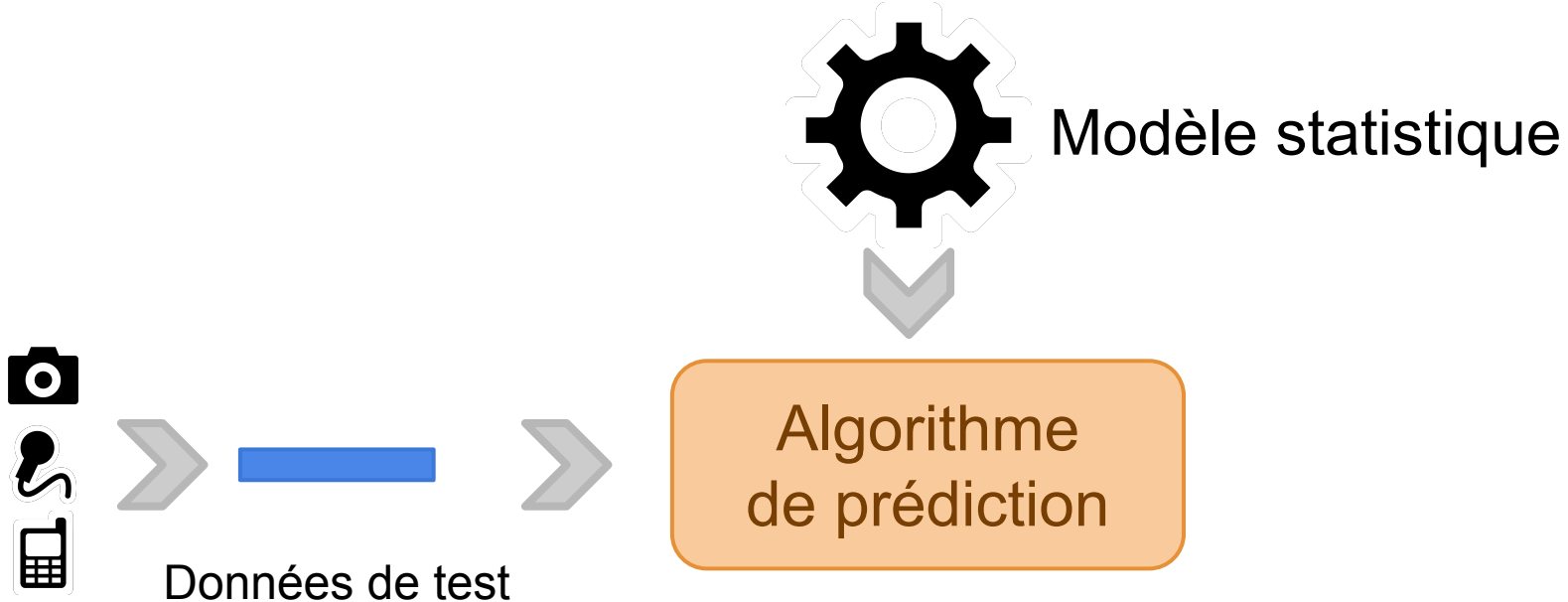


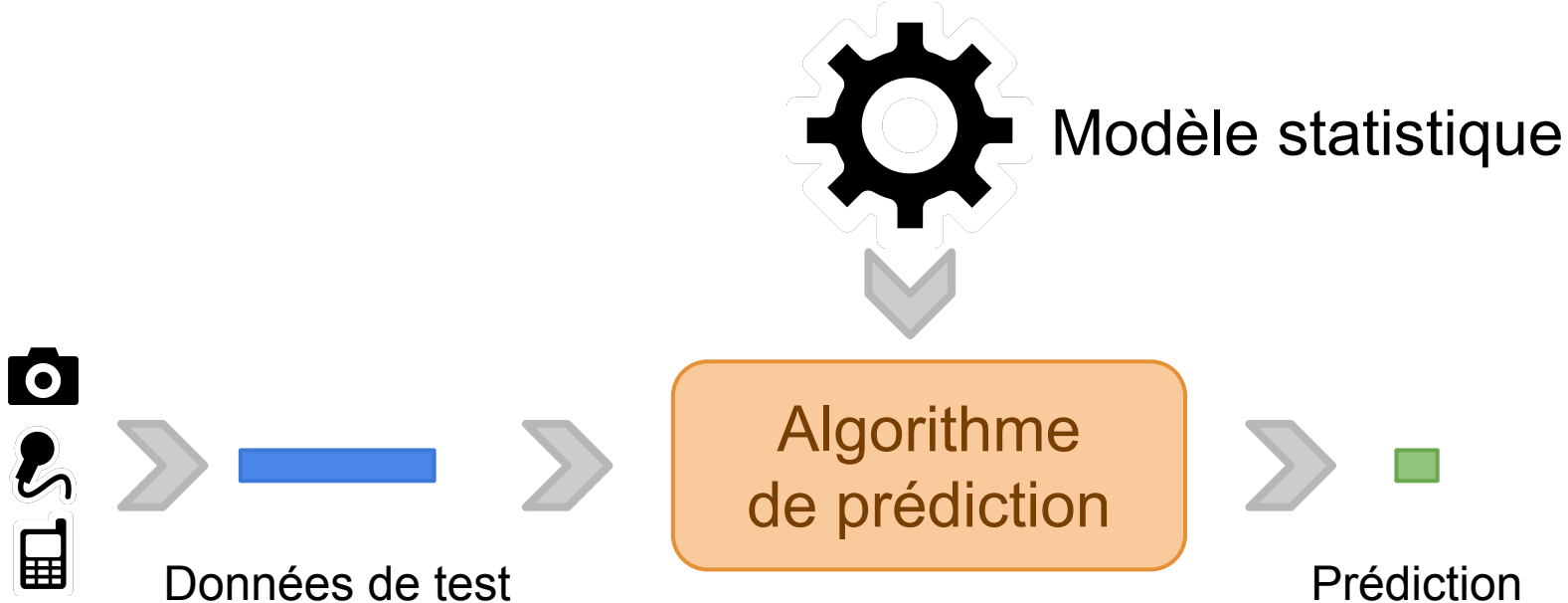
Base de données
d'entraînement
(annotées)

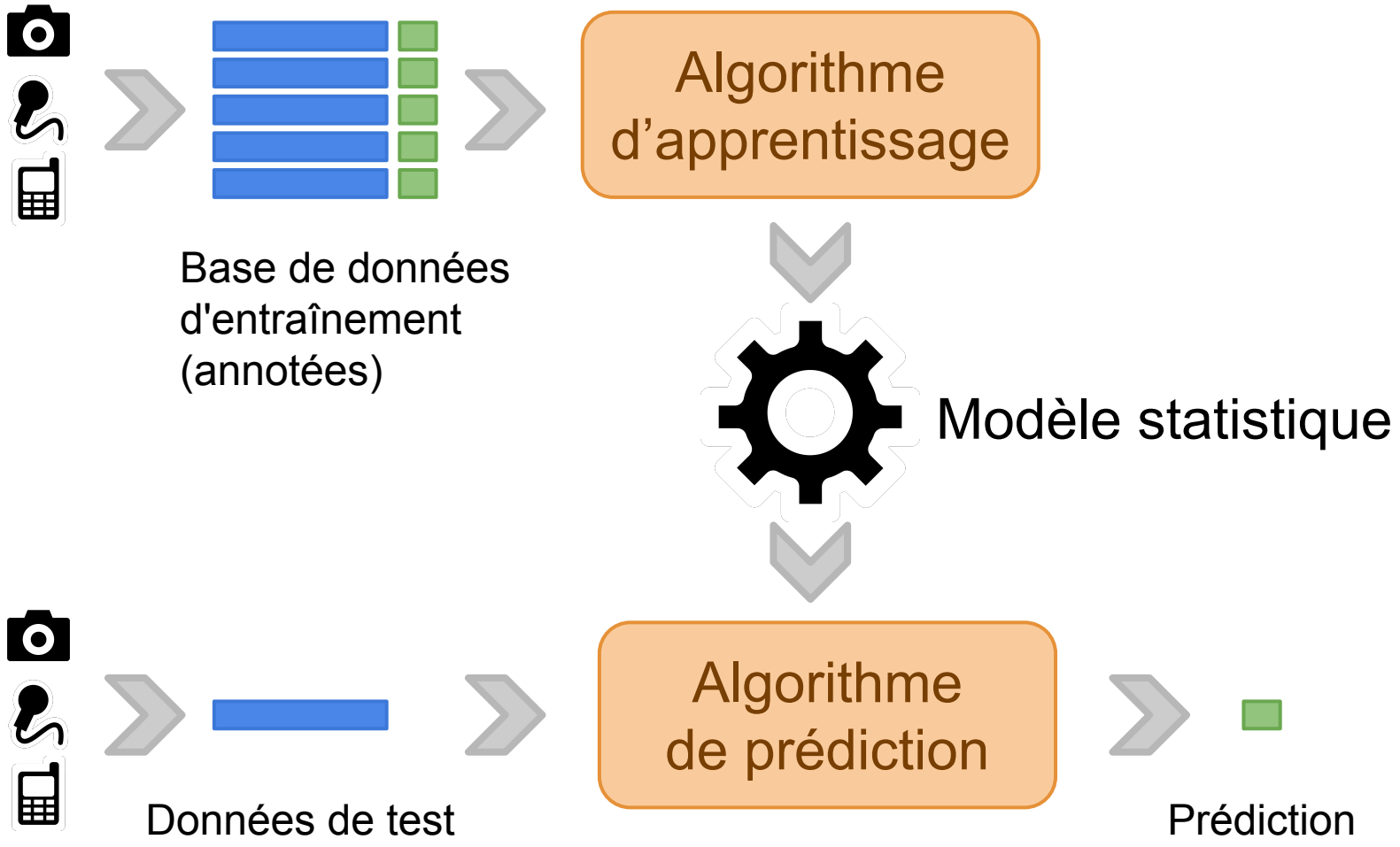

Algorithme
d’apprentissage
Base de données
d'entraînement
(annotées)
Modèle statistique



Modèle statistique


Modèle statistique
Algorithme
de prédiction
Données de test


Modèle statistique
Algorithme
de prédiction
Données de test
Prédiction


Algorithme
d'apprentissage
Base de données
d'entraînement
(annotées)
Modèle statistique
Algorithme
de prédiction
Données de test
Prédiction

Un «algorithme» fait des erreurs lorsque le modèle est mauvais.

Pourquoi obtient-on de mauvais modèles ?

Erreurs de mesures dans les entrées

Qualité des annotations de la base d'entraînement

Taille de la base d'entraînement

Mauvaise représentativité des données d’apprentissage

Mauvais a priori sur le contenu des données d'entraînement, présence de biais inconnus

Choix de modélisation inadaptés : choix de la famille de modèles, choix de objectif optimisé lors de l’apprentissage...

Construire la confiance

Regarder ses données : inspection qualitative

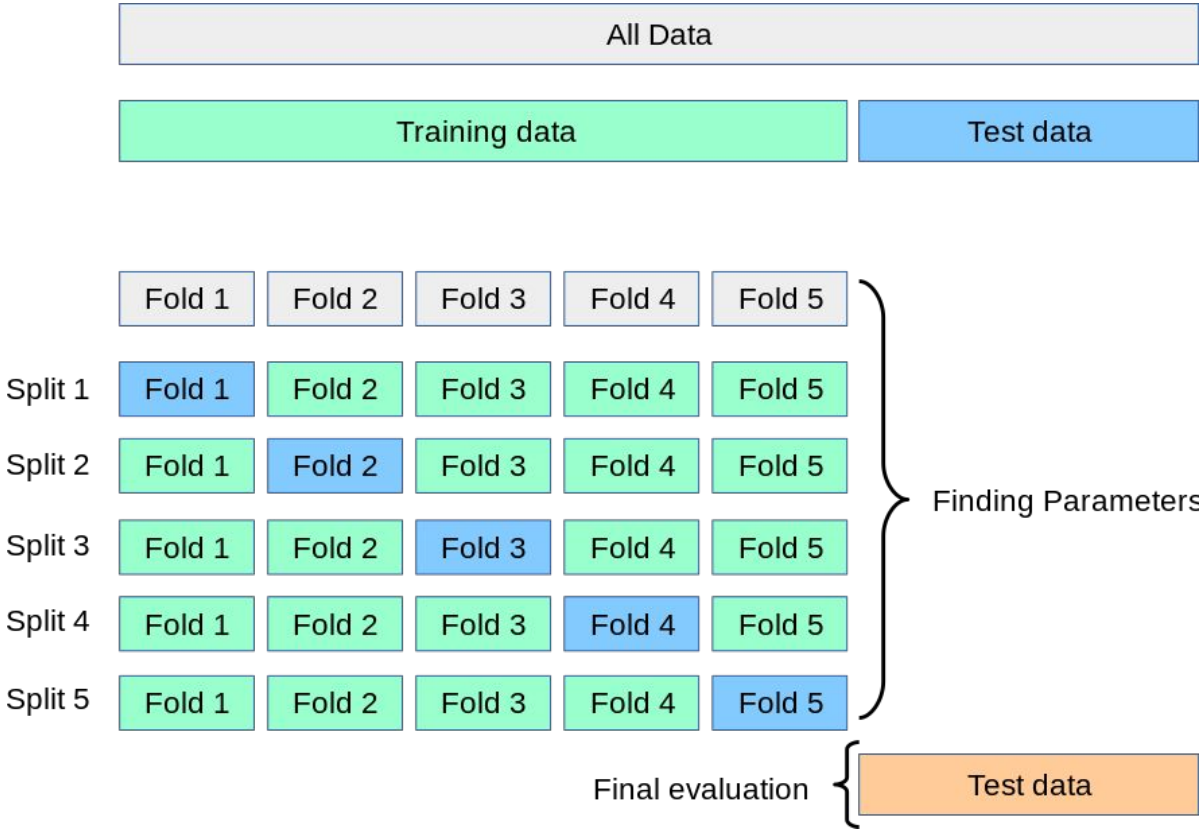
Mesurer la qualité des annotations (accord entre annotateurs)

Mesurer la performance en généralisation : validation croisée

Validation granulaire sur des sous-groupes de la population cible par âge, genre, origine ethnique ou sociale

Expliquer les décisions au cas par cas : ouvrir la boite noire

Estimer le niveau de confiance de chaque prédiction


All Data
Training data
Test data
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
Split 1
Split 2
Split 3
Split 4
Split 5
Finding Parameters
Final evaluation
Test data

Validation croisée

Une ceinture de sécurité pour le data-scientist

Attention à stratifier si la base d'entraînement n’est pas elle même i.i.d. : par sujet, par scanner ...

Ne permet pas de s’assurer que la base d'entraînement soit représentative des cas qui seront observés quand le modèle sera déployé en production.

Le cas particulier de l’apprentissage profond

Modèles en boîte “blanche”

TABLE 4.3. *Results from stepwise logistic regression fit to South African heart disease data.*

	Coefficient	Std. Error	Z score
`(Intercept)`	−4.204	0.498	−8.45
`tobacco`	0.081	0.026	3.16
`ldl`	0.168	0.054	3.09
`famhist`	0.924	0.223	4.14
`age`	0.044	0.010	4.52

source : http://web.stanford.edu/~hastie/ElemStatLearn/

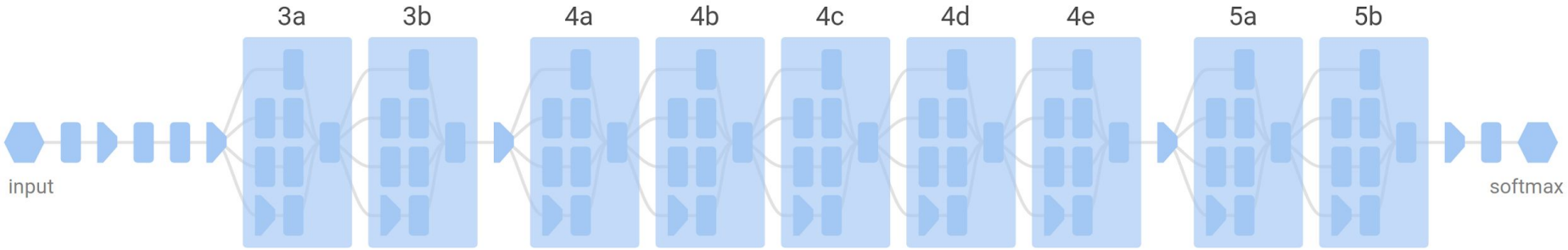
Limitation du modèle en boîte blanche

Une explication traditionnelle est compréhensible uniquement avec un faible nombre de variables décorrélées.

La détection d'objet dans les images implique un grand nombre de variables (des dizaines de milliers de pixels par images) très corrélées.

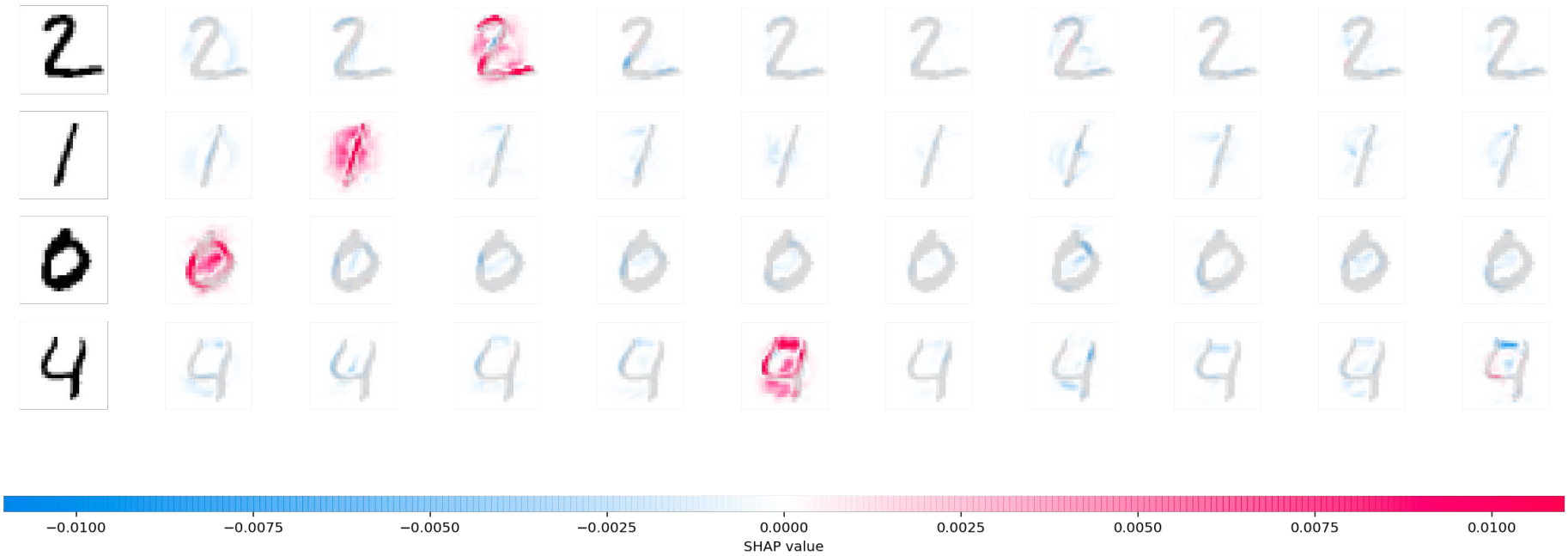
Les algorithmes d'apprentissage profond sont typiquement imbattables sur des données avec un grand nombre de variables avec des dépendances structurées (pixels d'une images, séquences de mesures temporelles…)

Qui ferait confiance à l'explication fournie par un modèle linéaire qui prédit sur 5 pixels bien choisis et qui se trompe la plupart du temps ?

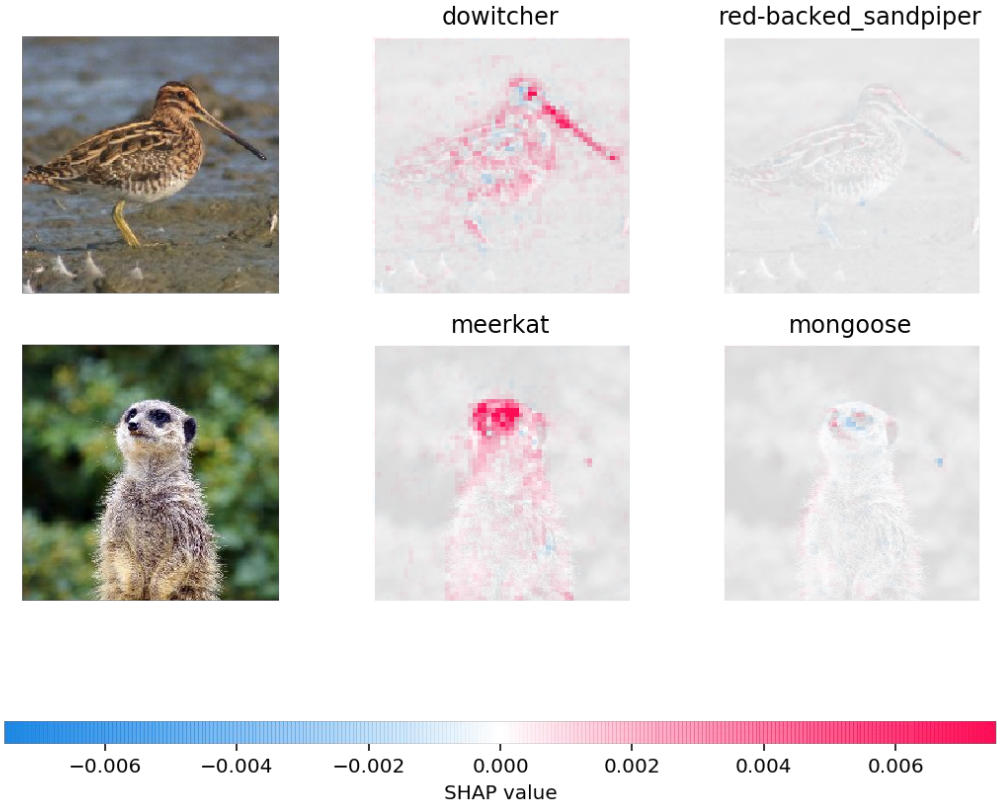


source : https://distill.pub/2019/activation-atlas/

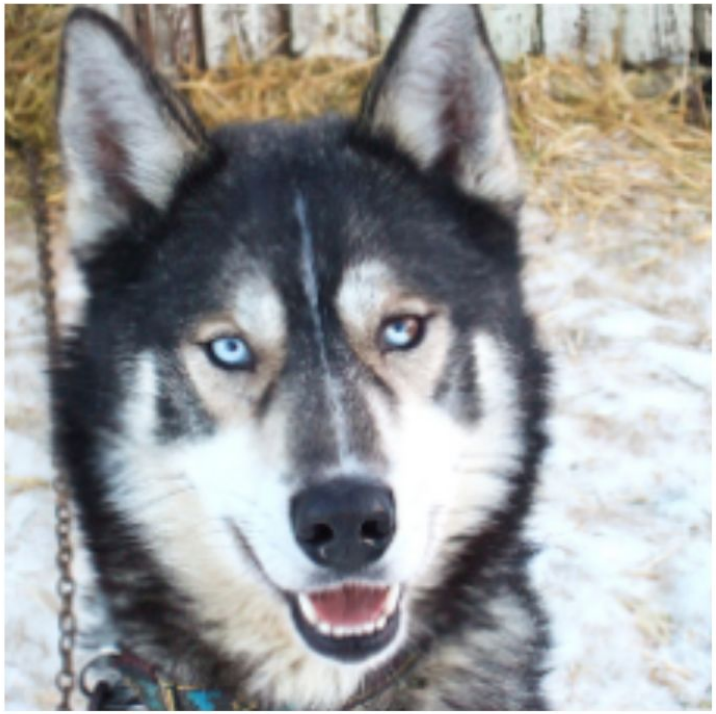
Expliquer une décision d’un modèle fortement non linéaire



https://github.com/slundberg/shap



https://github.com/slundberg/shap



(a) Husky classified as wolf

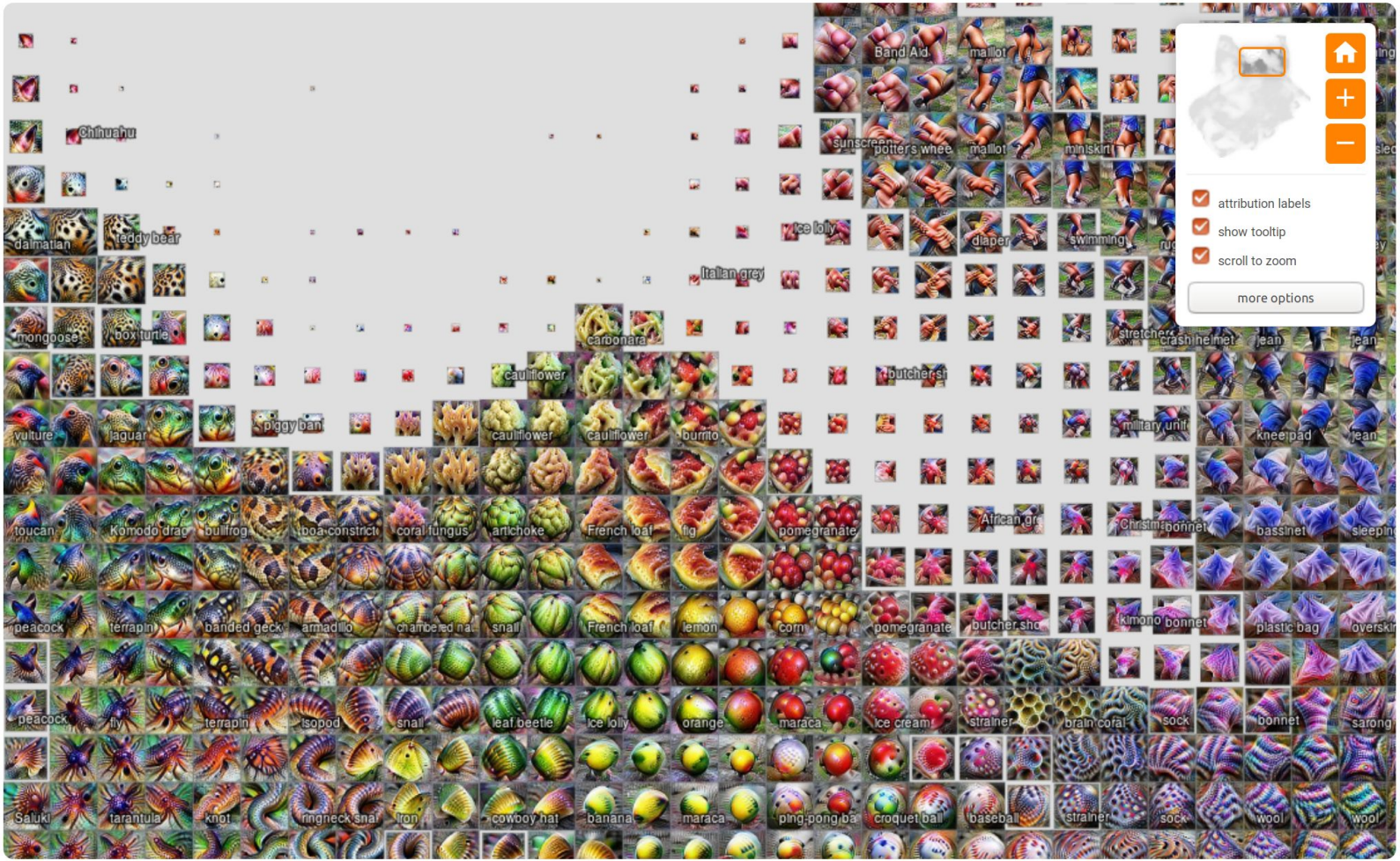
(b) Explanation

https://arxiv.org/abs/1602.04938

Ouvrir la boîte noire

Layer

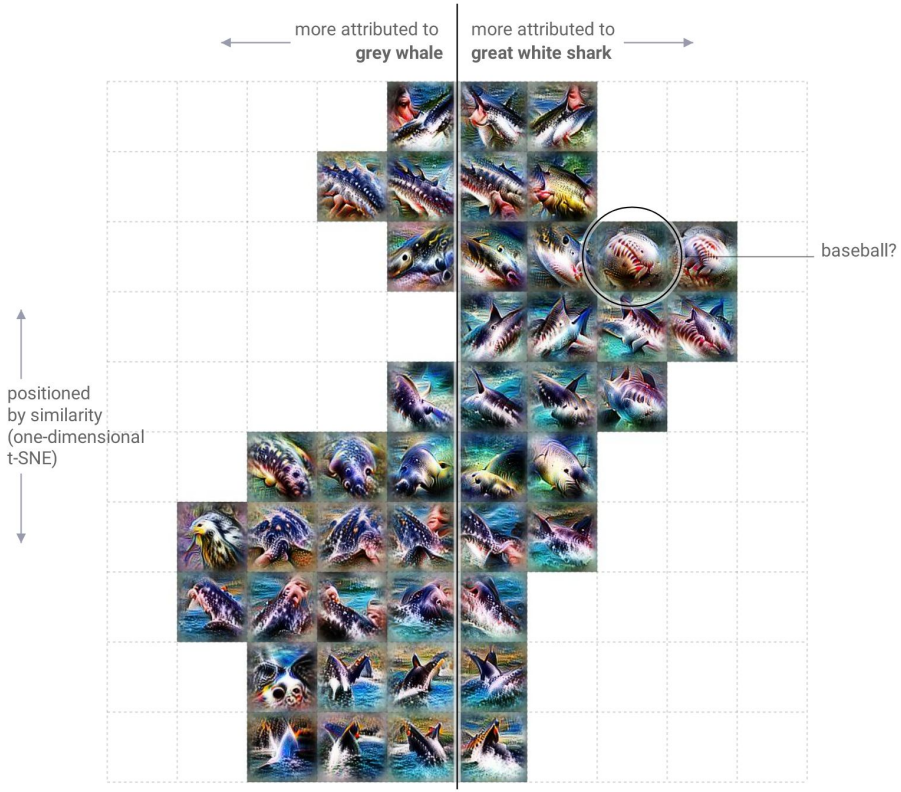
- MIXED3A
- MIXED3B
- MIXED4A
- MIXED4B
- MIXED4C
- **MIXED4D**
- MIXED4E
- MIXED5A
- MIXED5B



attribution labels
show tooltip
scroll to zoom
more options

Layer

- MIXED3A
- MIXED3B
- MIXED4A
- MIXED4B
- MIXED4C
- MIXED4D
- MIXED4E
- MIXED5A
- MIXED5B



source : https://distill.pub/2019/activation-atlas/

1.	**grey whale**	**91.0%**
2.	killer whale	7.5%
3.	great white shark	0.7%
4.	gar	0.4%
5.	sea lion	0.1%
6.	tiger shark	0.1%

1.	**great white shark**	**66.7%**
2.	baseball	7.4%
3.	grey whale	4.1%
4.	sombrero	3.2%
5.	sea lion	3.1%
6.	killer whale	2.7%

1.	**baseball**	**100.0%**
2.	rugby ball	0.0%
3.	golf ball	0.0%
4.	ballplayer	0.0%
5.	drum	0.0%
6.	sombrero	0.0%

source : https://distill.pub/2019/activation-atlas/

Pourquoi ouvrir la boite noire ?

Analyse des activations : mise en évidence de biais dans le fonctionnement du modèles (principalement utile pour les chercheurs)

Analyser la cause d'erreurs de prédiction en **phase de développement**

- Debugging des choix de modélisation
- Mise en évidence de biais dans la base d'entraînement

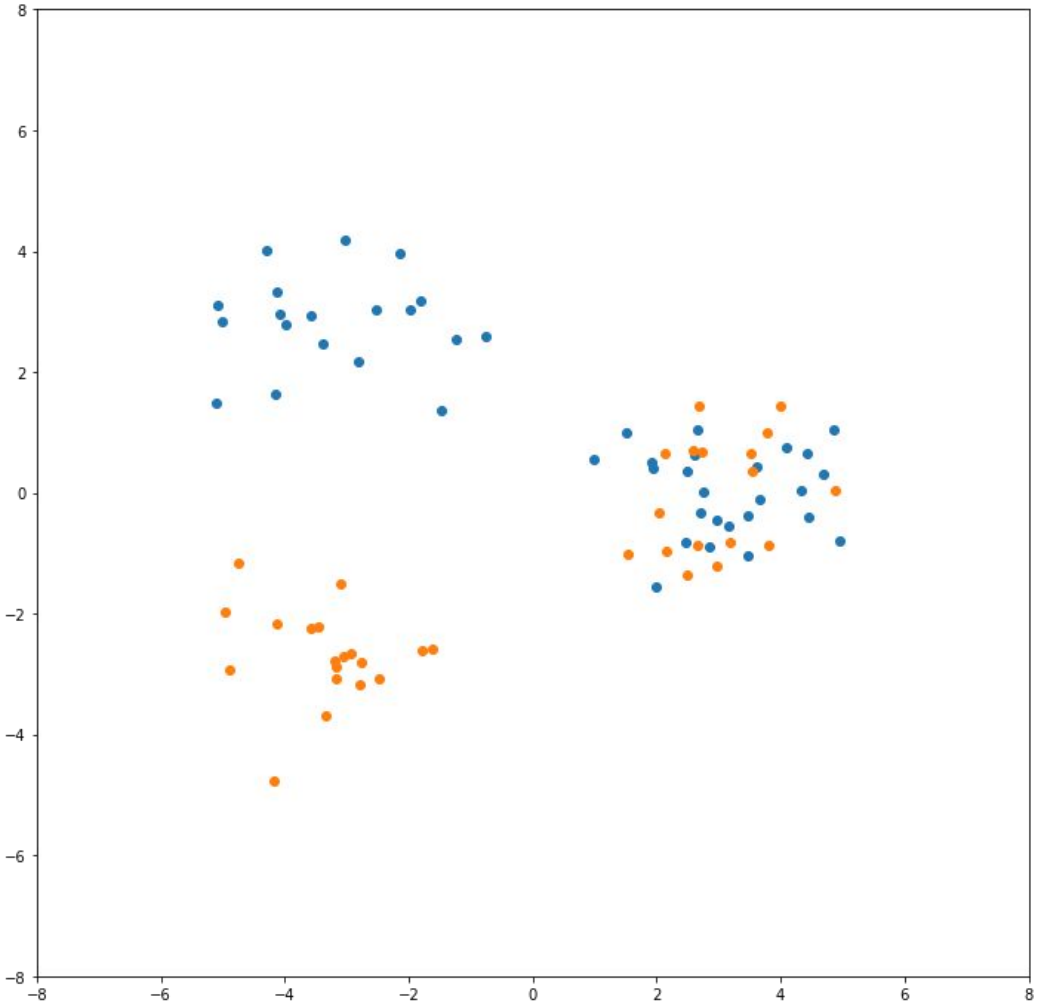
Analyser les décisions prises **en production**

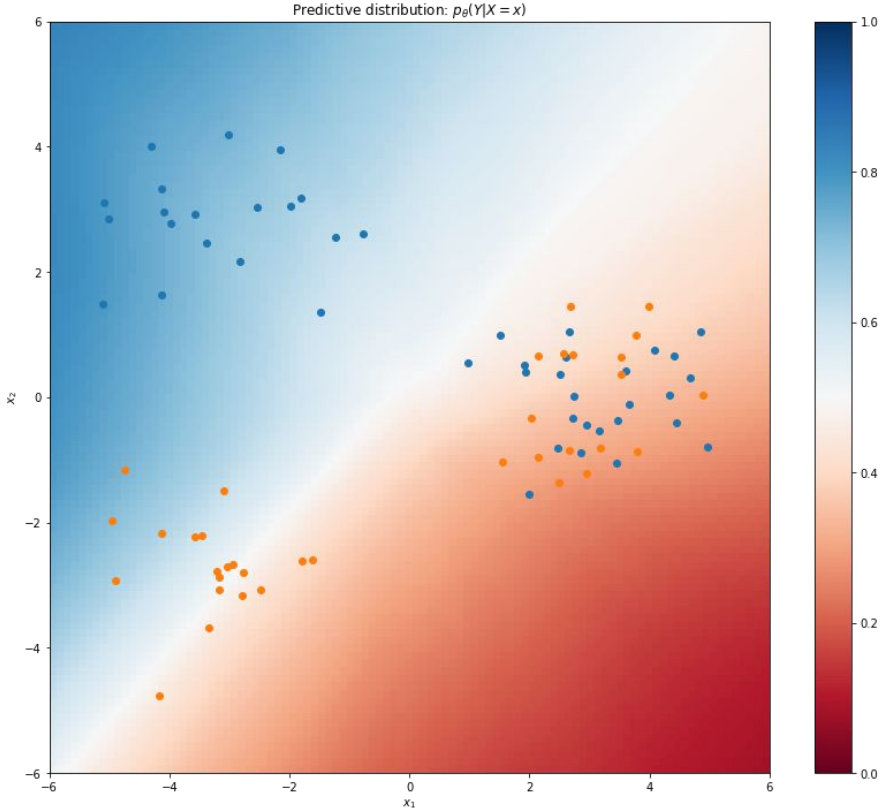
- Vérifier que le modèle prédit pour de "bonnes raisons"
- Analyser la cause de décisions incertaines (prédictions à faible confiance)

Peut-on faire confiance à la confiance ?

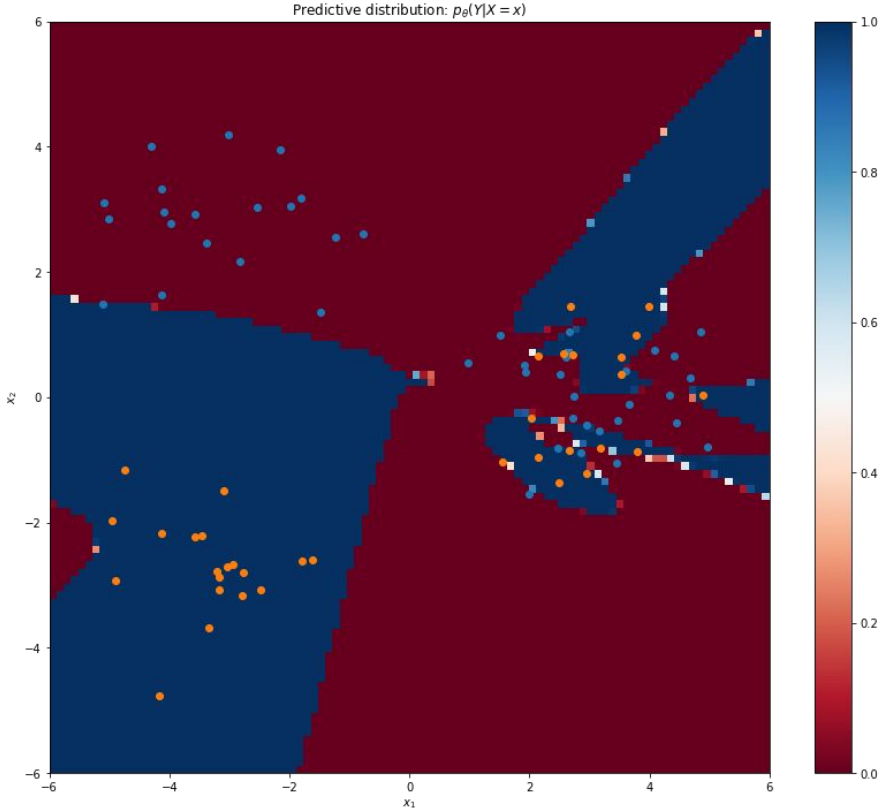
« Il y a des inconnus connus, c'est-à-dire, qu'il y a des choses que nous savons que nous ne savons pas. »

Donald Rumsfeld

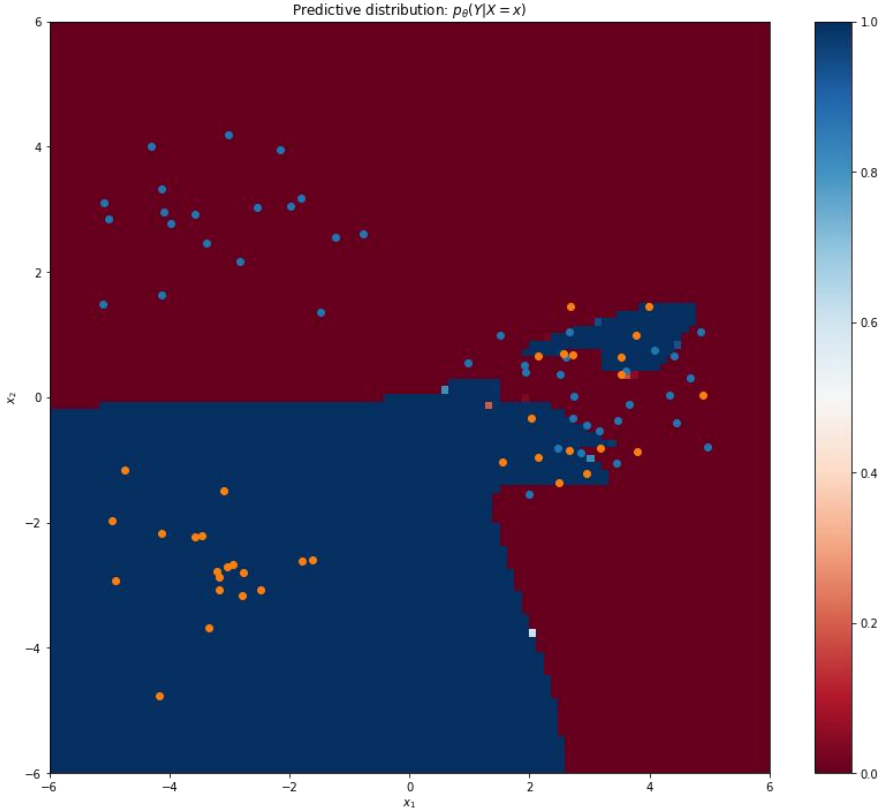

8
6
4
2
0
−2
−4
−6
−8
−8
−6
−4
−2
0
2
4
6
8



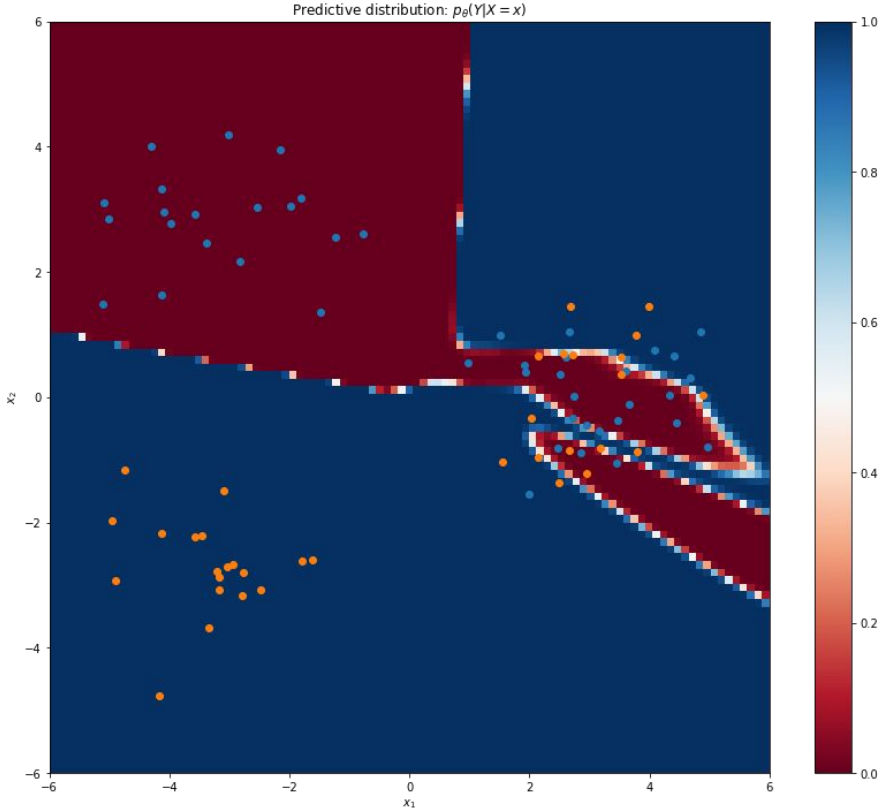
Réseau de neurones non-entraîné



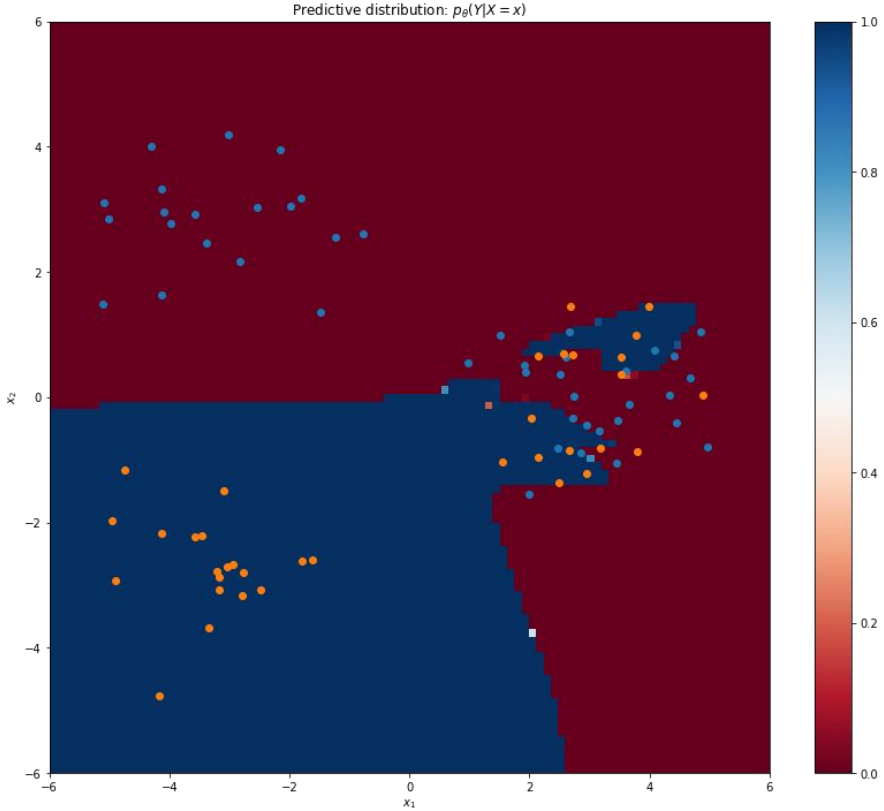
Modèle trop confiant: MLE avec l-BFGS sans régularisation



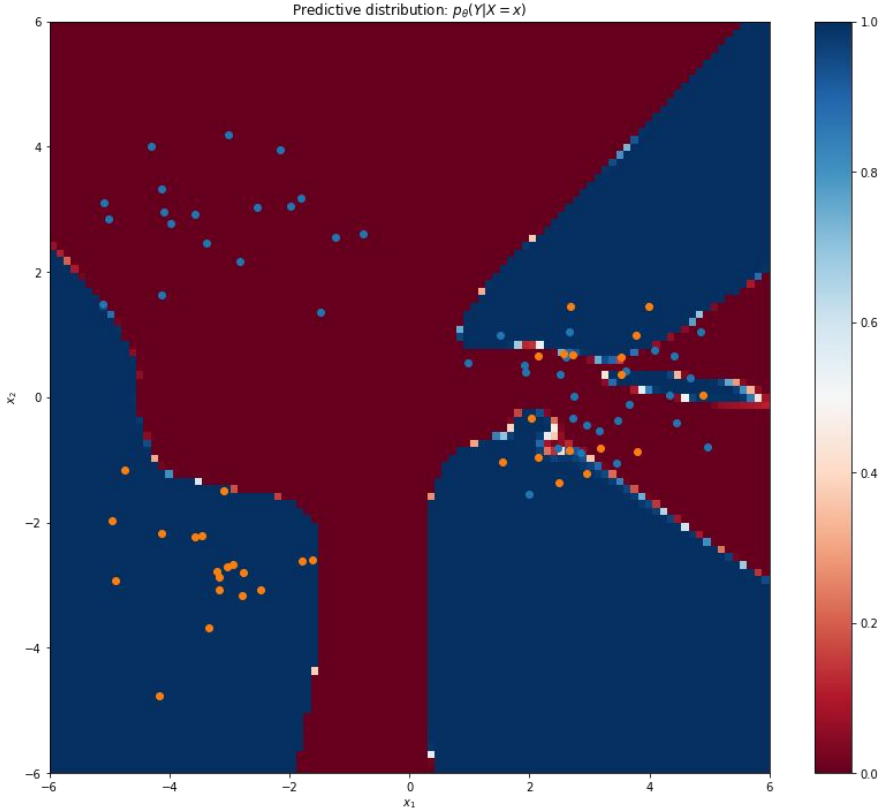
Modèle trop confiant (entraîné sur un échantillon aléatoire)



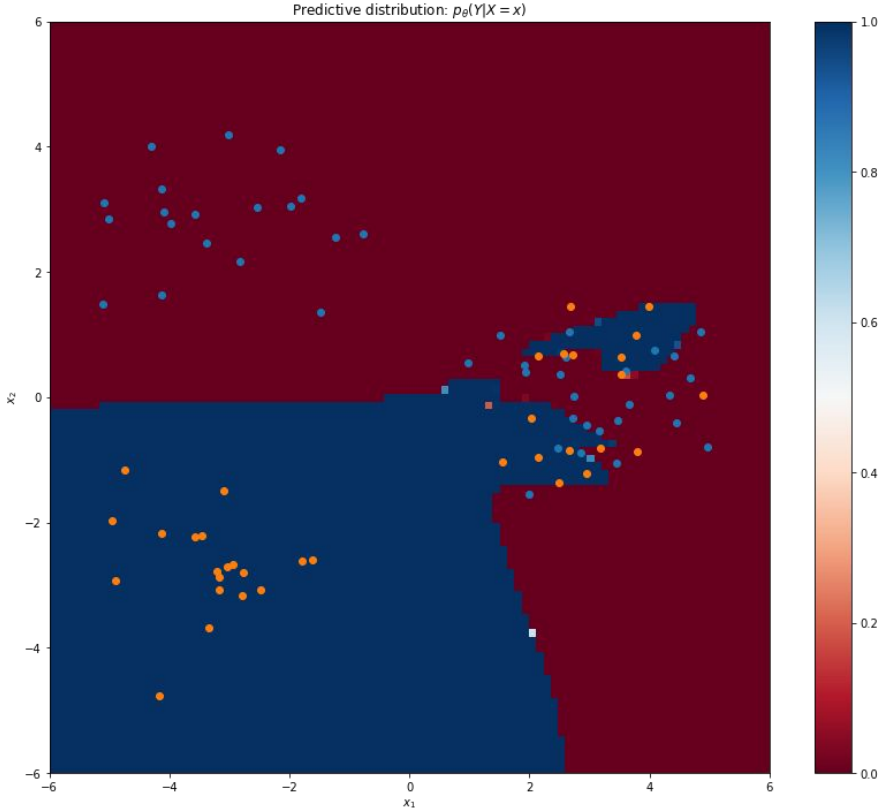
Modèle trop confiant (entraîné sur un échantillon aléatoire)



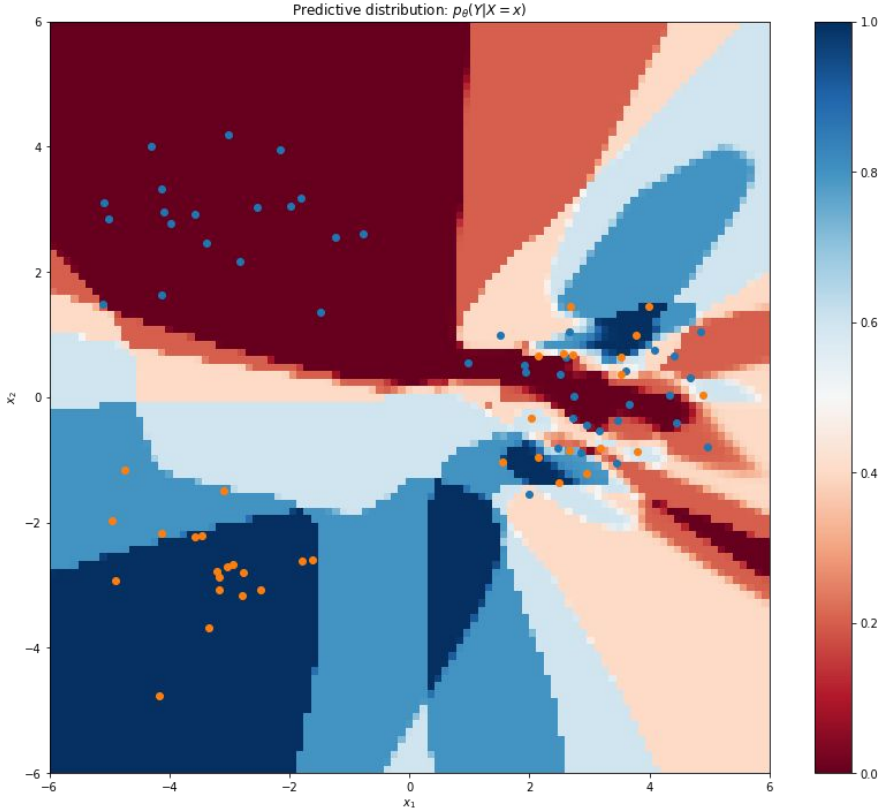
Modèle trop confiant (entraîné sur un échantillon aléatoire)



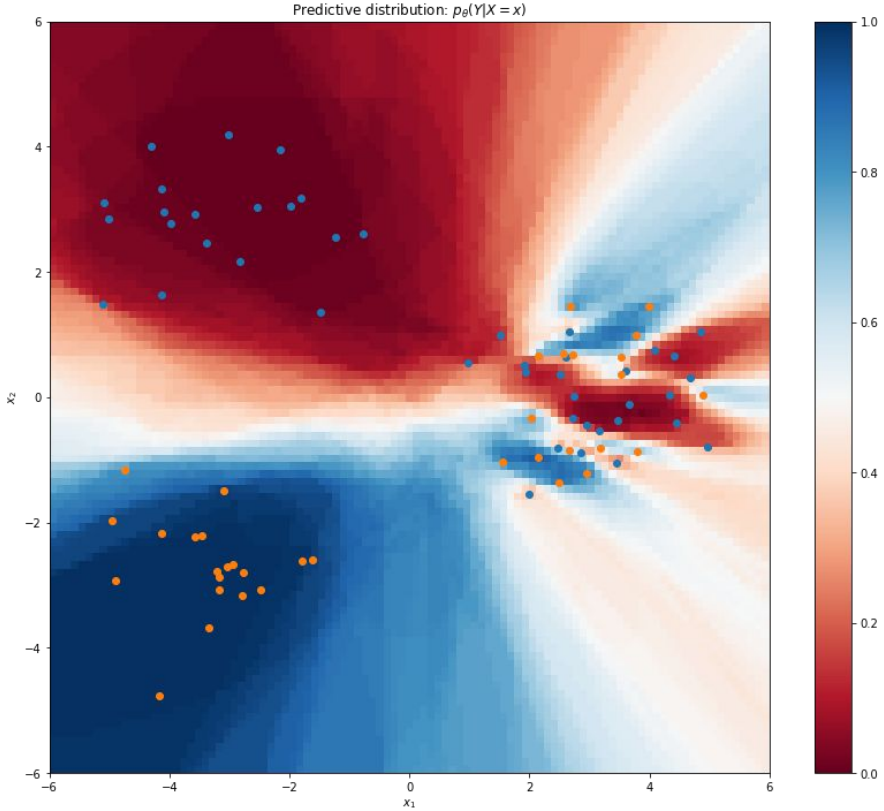
Modèle trop confiant (entraîné sur un échantillon aléatoire)



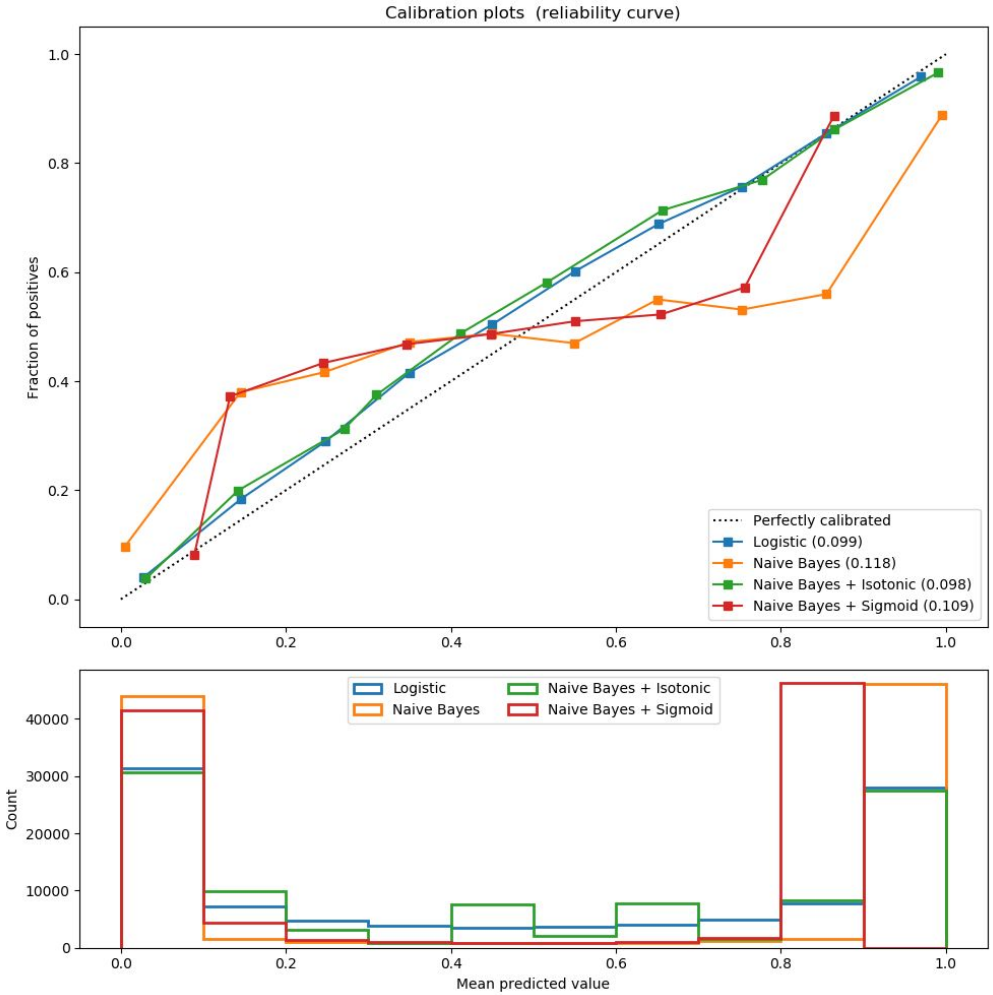
Modèle trop confiant (entraîné sur un échantillon aléatoire)



Ensemble de 5 modèles



Ensemble de 100 modèles


Calibration plots (reliability curve)
Fraction of positives
Perfectly calibrated
Logistic (0.099)
Naive Bayes (0.118)
Naive Bayes + Isotonic (0.098)
Naive Bayes + Sigmoid (0.109)
Count
Logistic
Naive Bayes
Naive Bayes + Isotonic
Naive Bayes + Sigmoid
Mean predicted value

Avoir confiance dans la confiance des prédictions

Évaluer la calibration avec les courbes de calibration

Corriger les problèmes de calibration

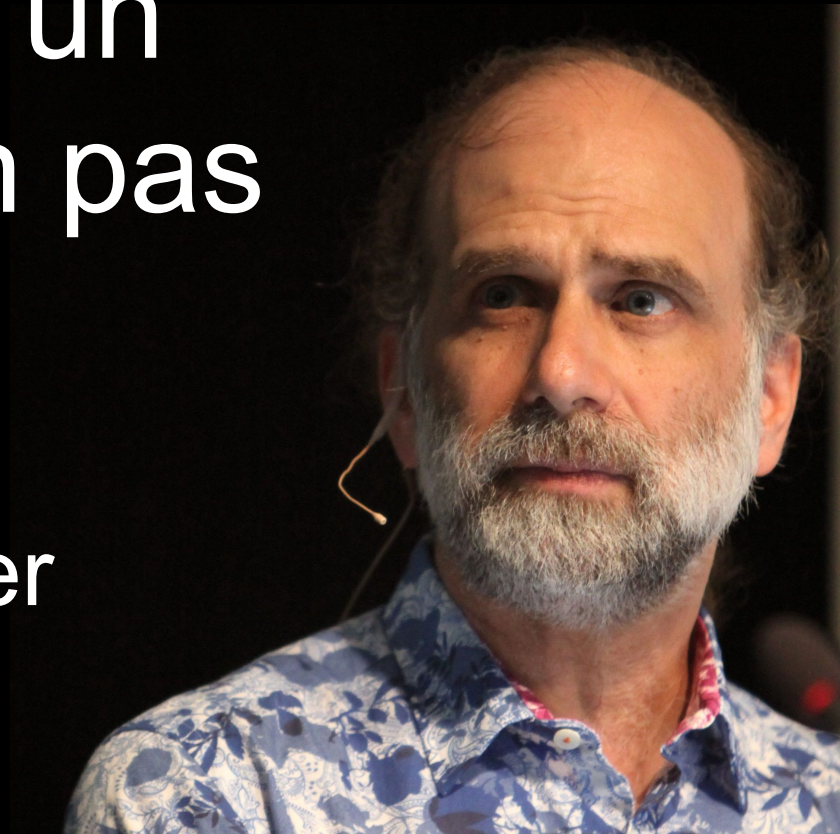
- Calibration a posteriori (régression isotonique)
- Régularisation (learning rate + early stopping, dropout, weight decay…)
- Ensemble (Bagging)
- Bayesian Deep Learning (Bayes by Backprop, Noisy Adam, Cyclic SG-HMC)

Détection de la nouveauté et des anomalies : p(y|x) vs p(x)

Conclusion

« La sécurité est un processus et non pas un produit. »

Bruce Schneier



Conclusions

Éducation : des concepteurs, régulateurs et des opérateurs sur les pièges courants et bonnes pratiques dans la conception et l'utilisation de systèmes à base de ML/DL.

Mise en place de **procédures de validation** systématiques : qualité des annotations, couverture de la base d'entrainement, mesure de la performance en cross-validation avec une stratification de la population, mesure de calibration de l'indice de confiance.

Améliorer la **transparence** : outils d'explication des décisions pour le data scientist, publication des courbes de performances, prédictions granulaires avec niveau de confiance, open-source, partage des données...

Merci pour votre attention !

@ogrisel