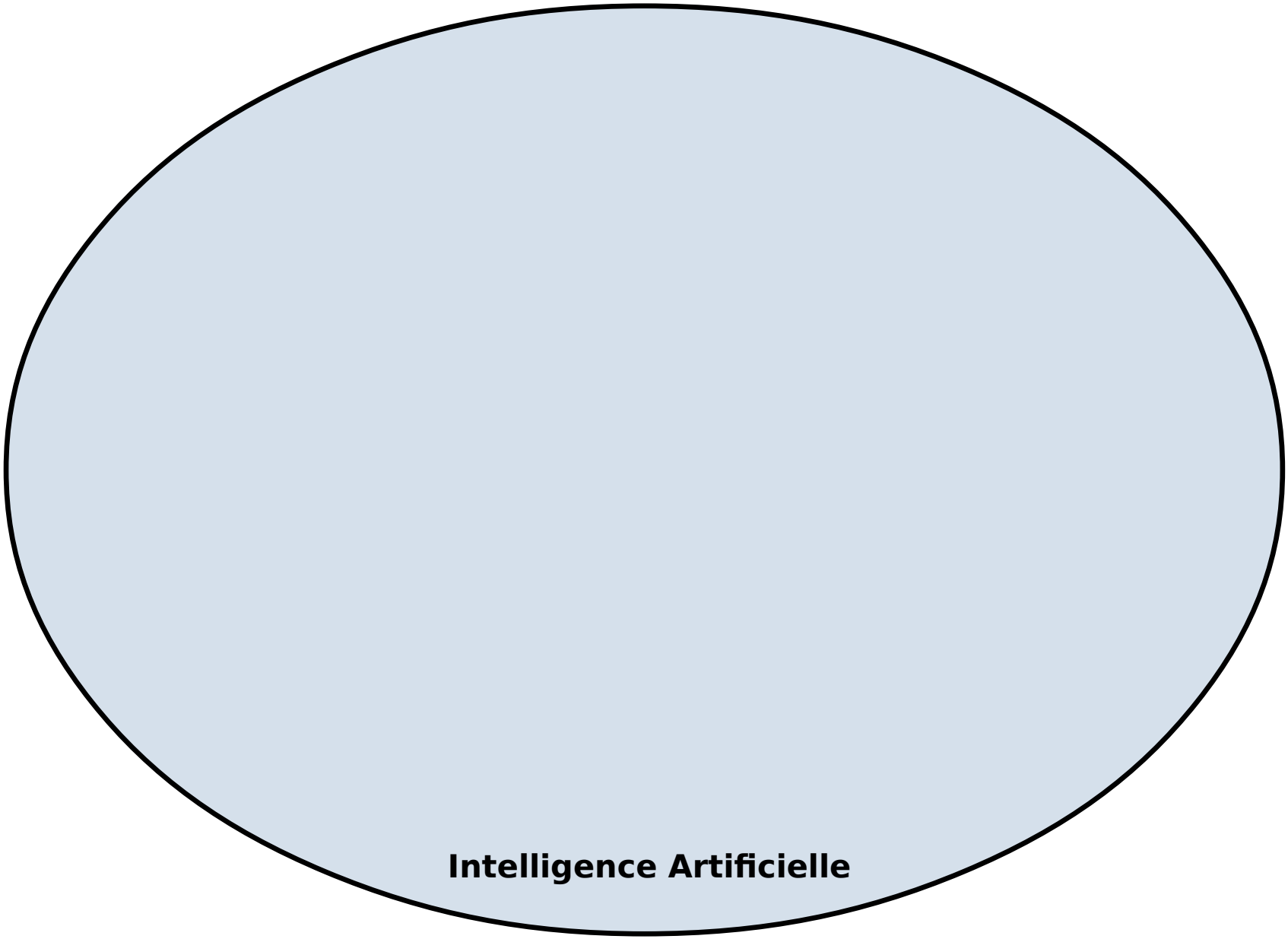
Apprentissage de réseaux de neurones à activations binaires avec garanties statistiques

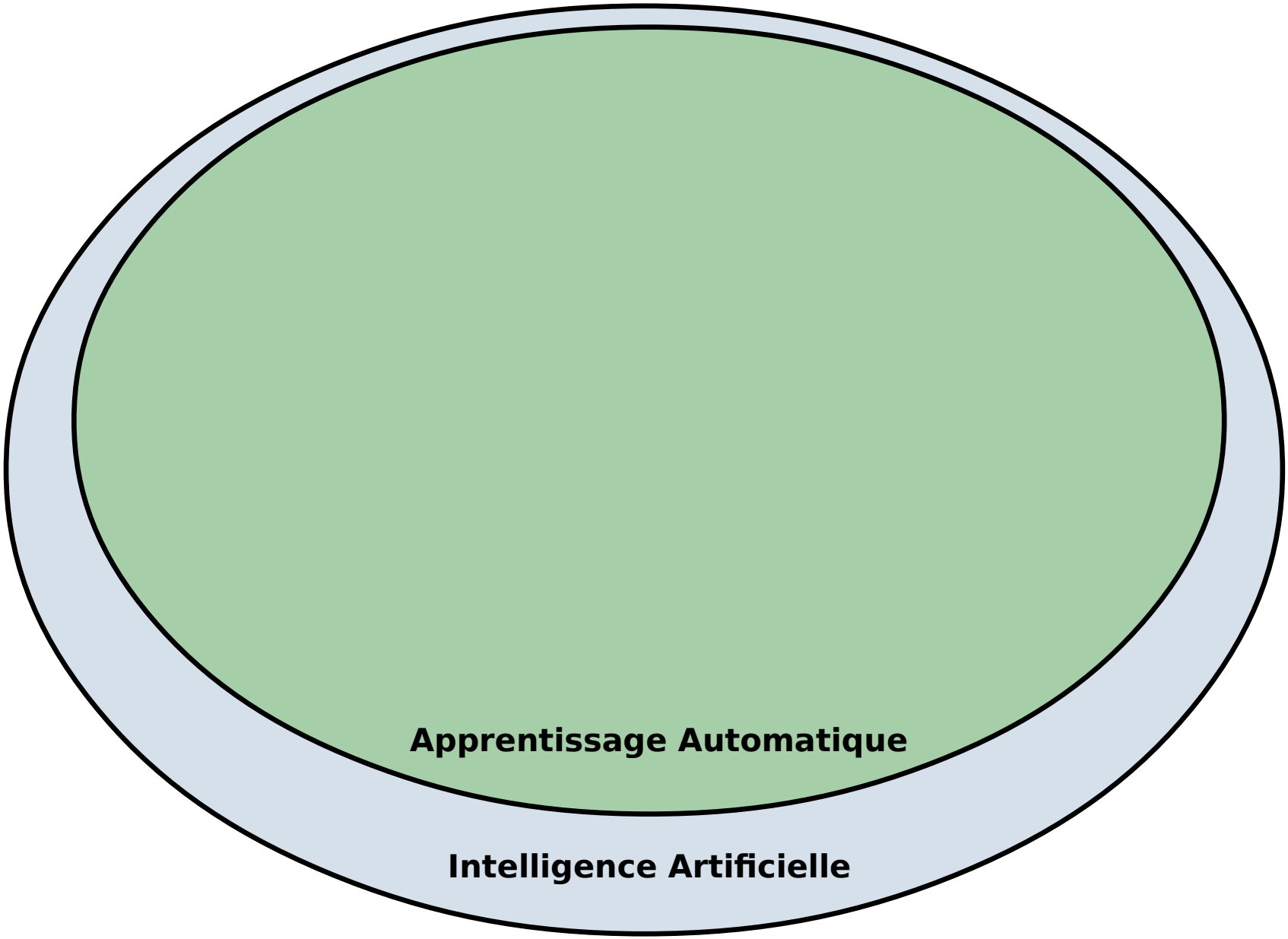
Gaël Letarte

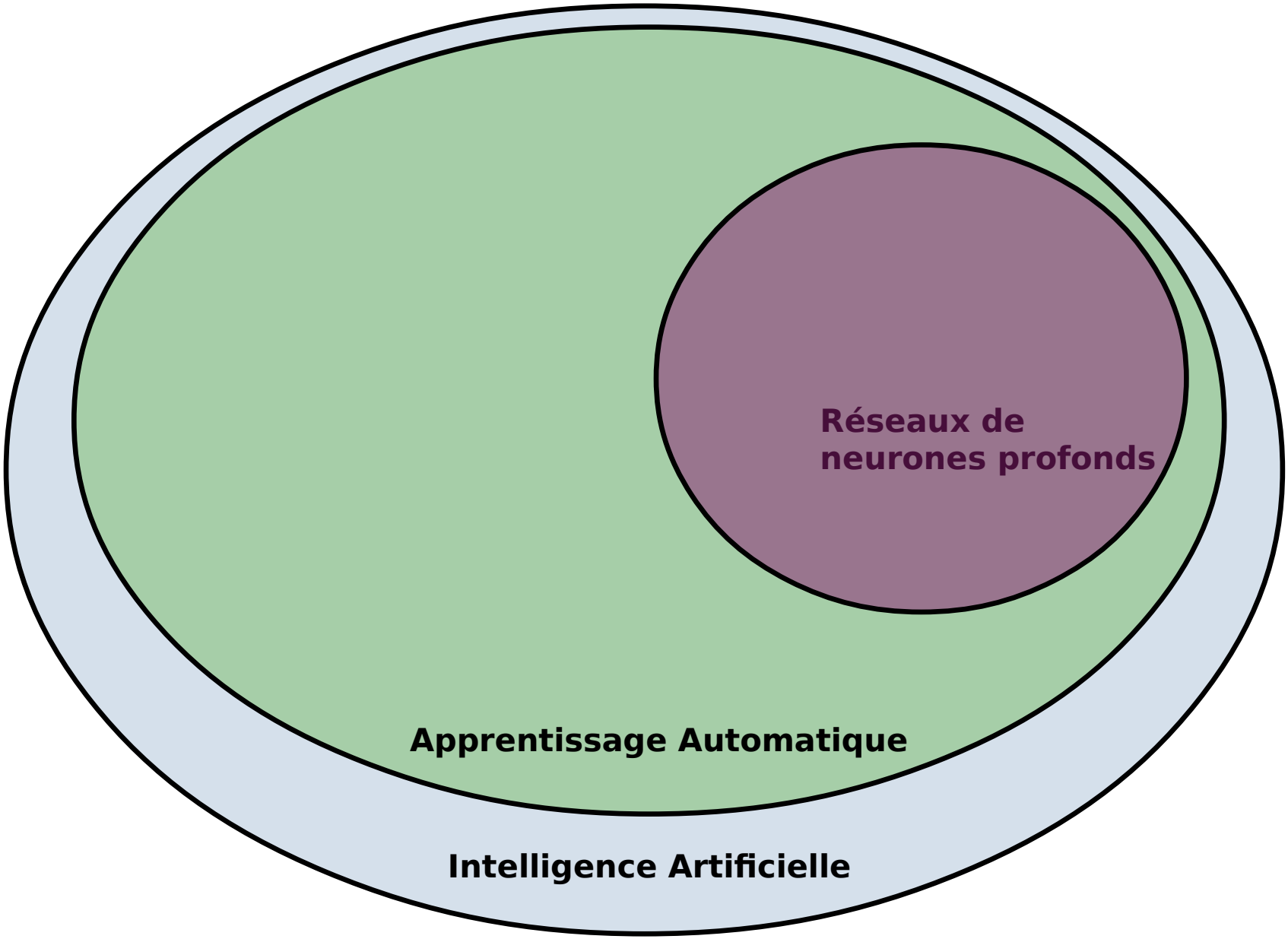
En collaboration avec Pascal Germain, Benjamin Guedj et François Laviolette

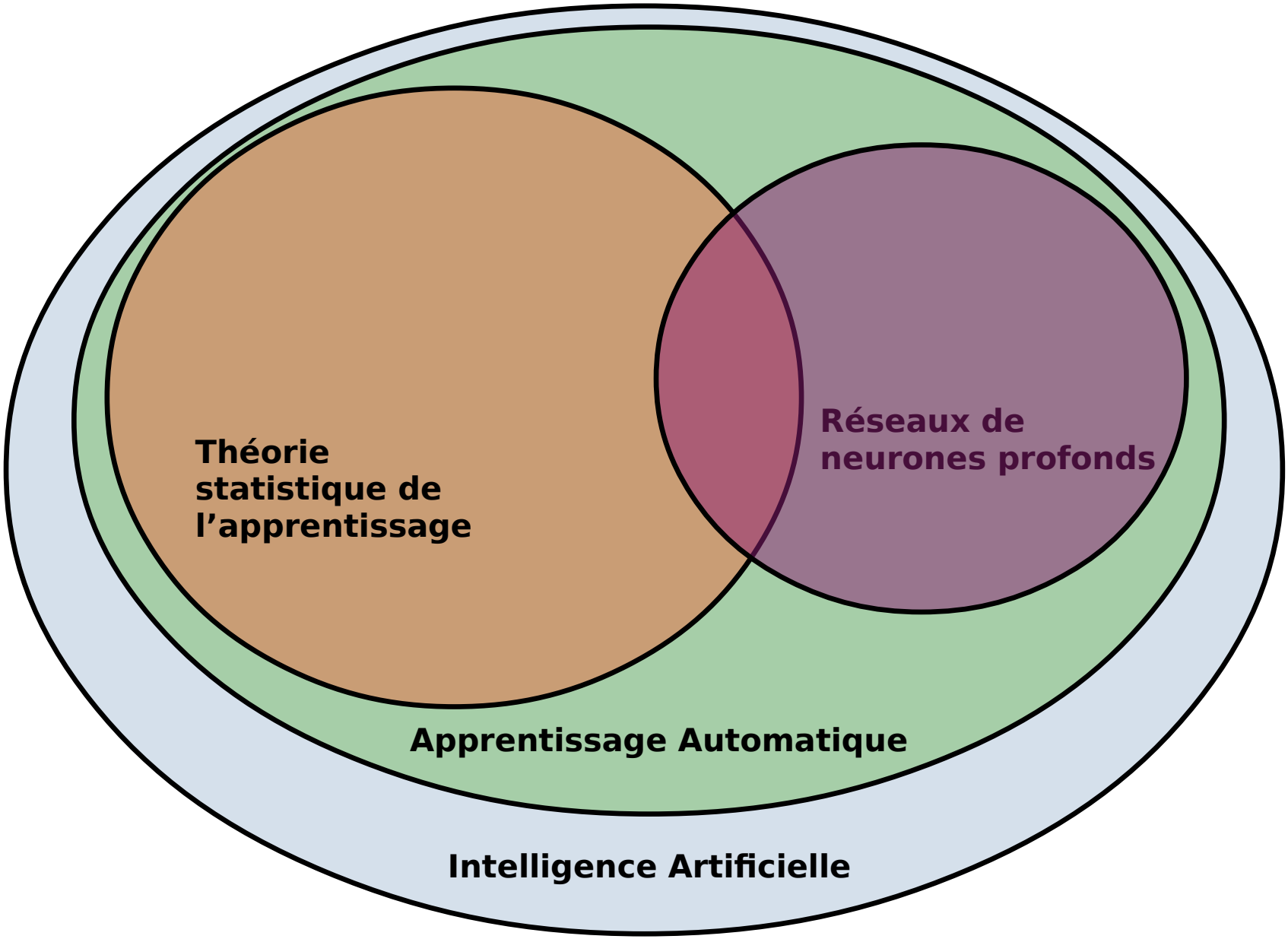
Département d'informatique et de génie logiciel, Université Laval

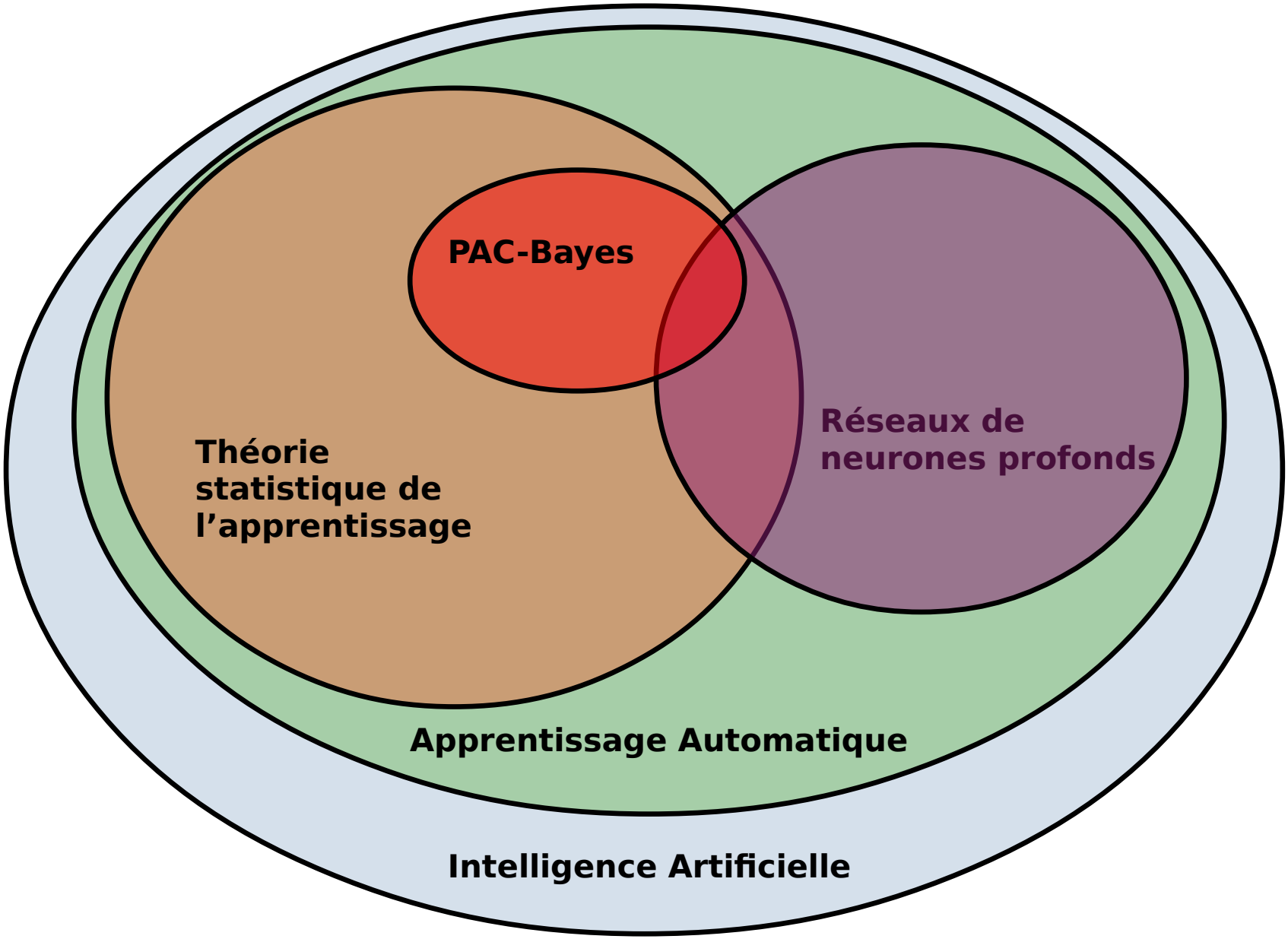
17 Janvier, 2020

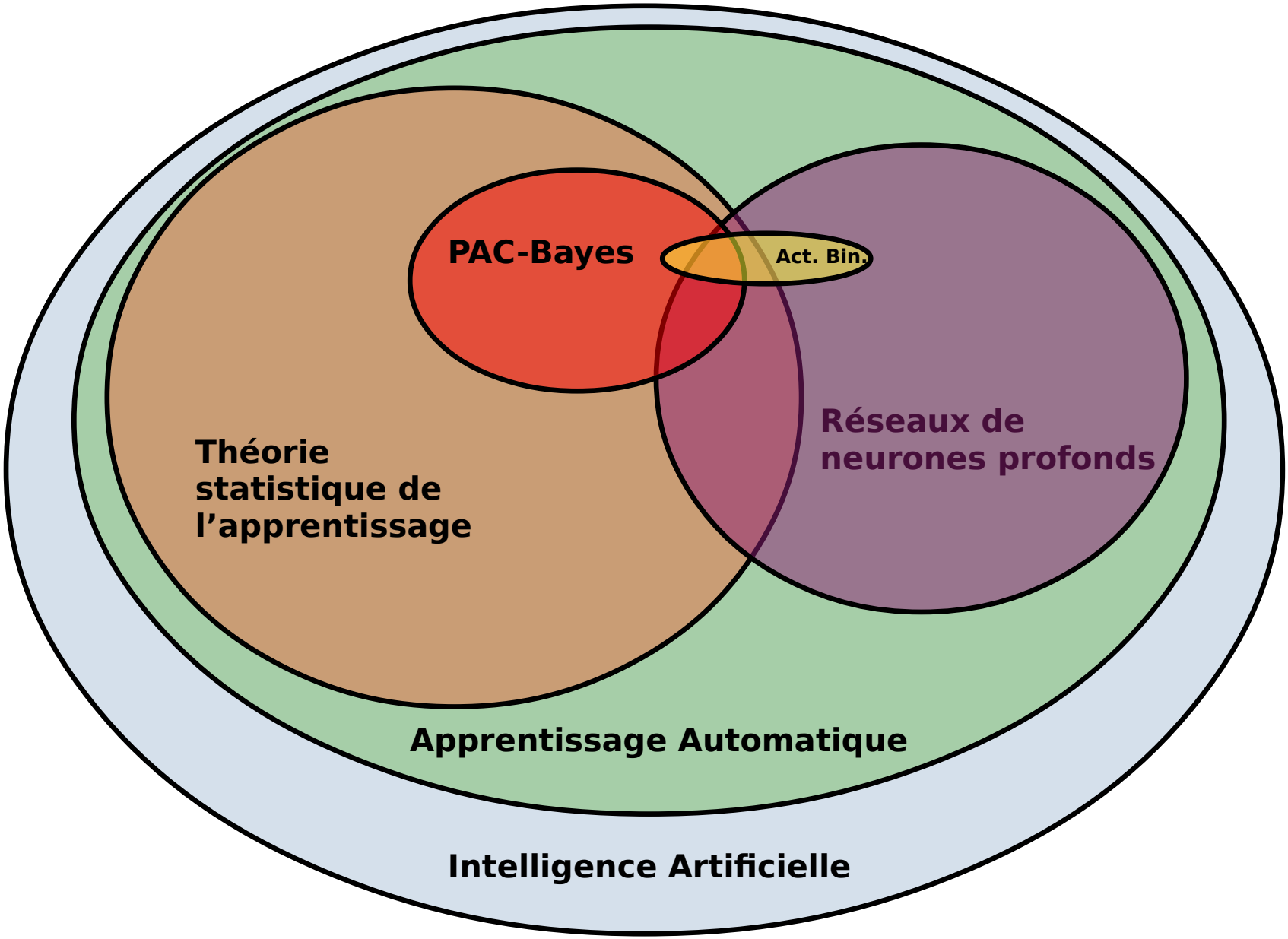

Intelligence Artificielle

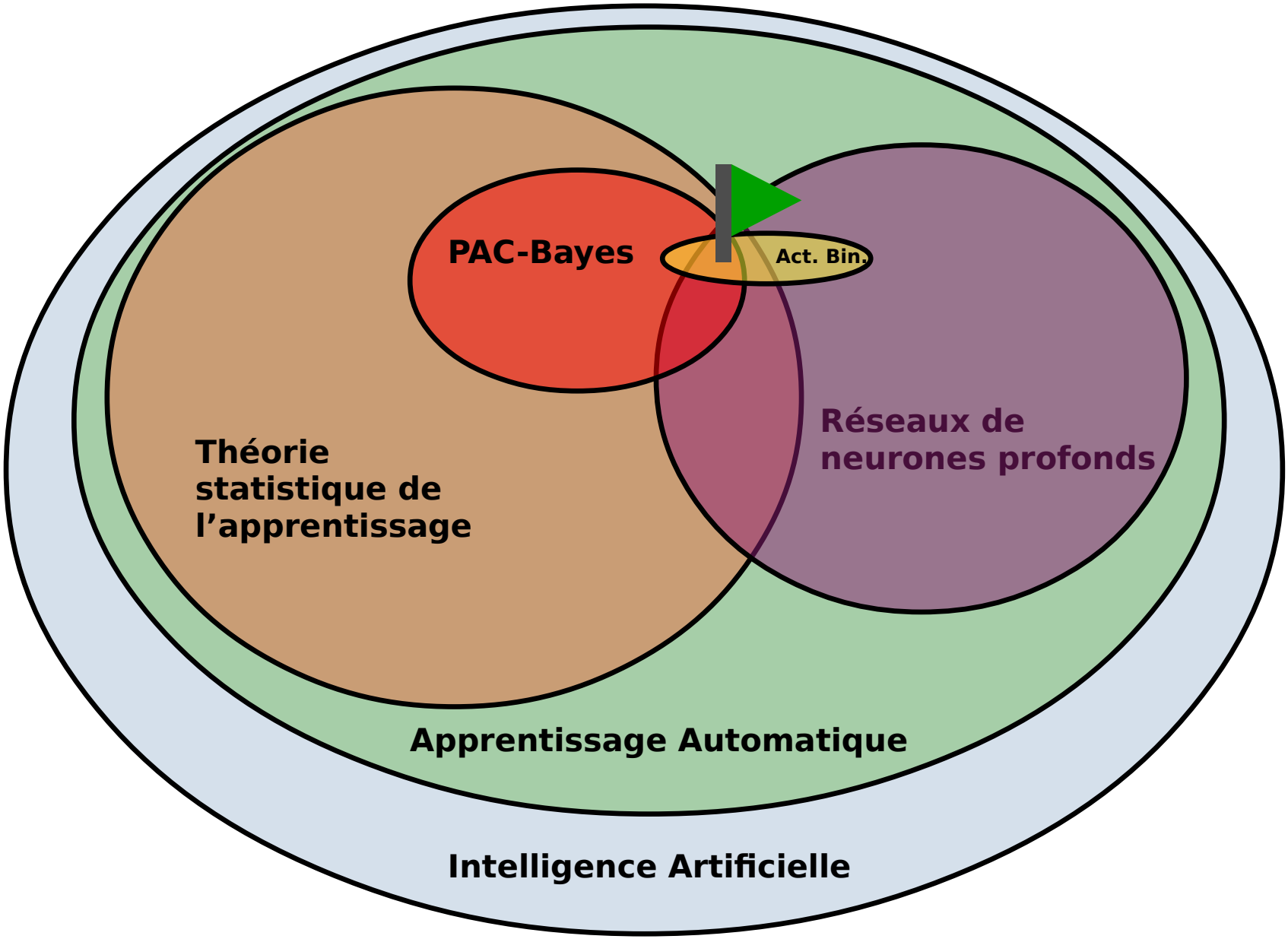

Apprentissage Automatique
Intelligence Artificielle


Réseaux de
neurones profonds
Apprentissage Automatique
Intelligence Artificielle

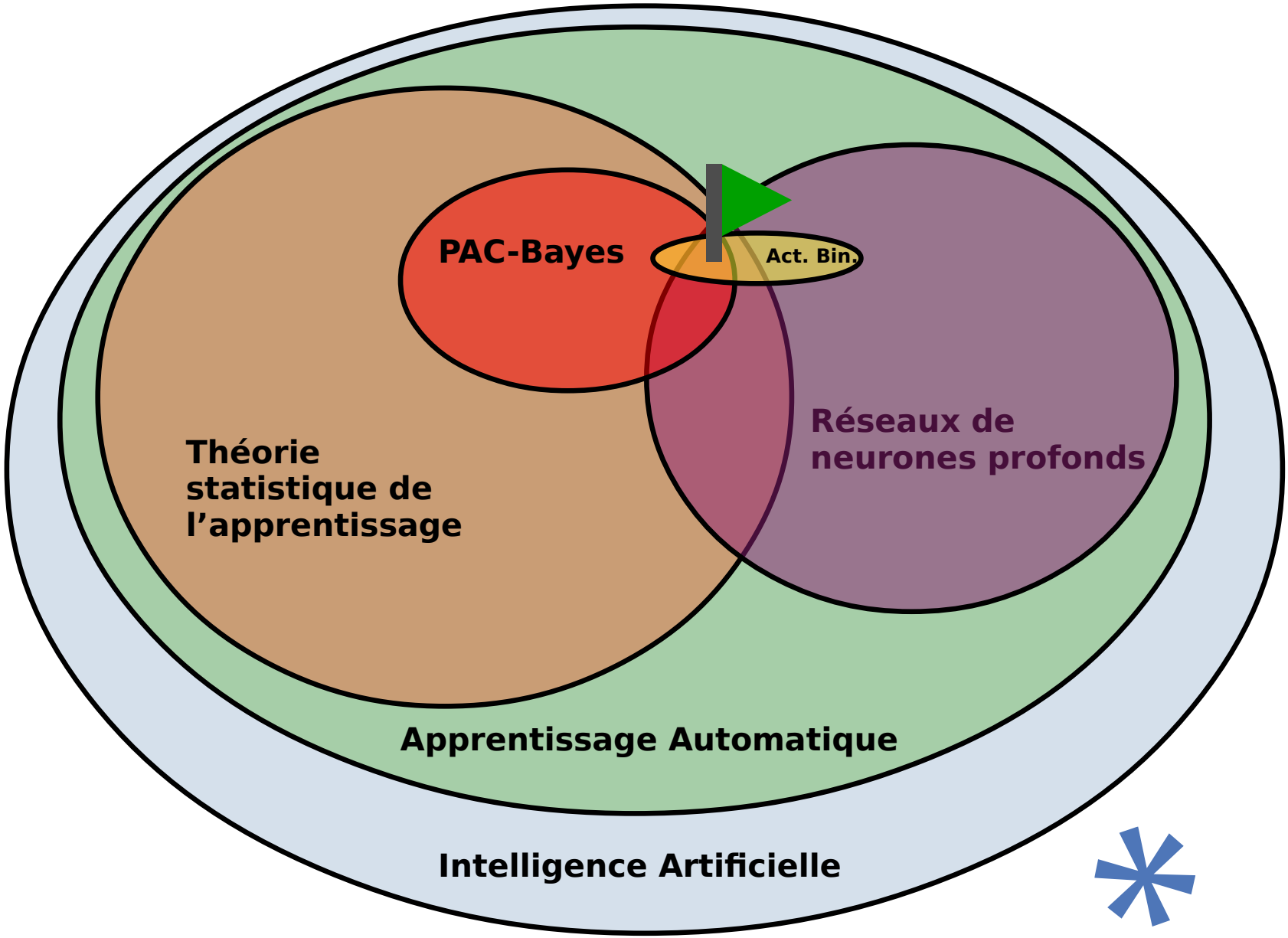

Théorie
statistique de
l'apprentissage
Réseaux de
neurones profonds
Apprentissage Automatique
Intelligence Artificielle


PAC-Bayes
Théorie
statistique de
l'apprentissage
Réseaux de
neurones profonds
Apprentissage Automatique
Intelligence Artificielle

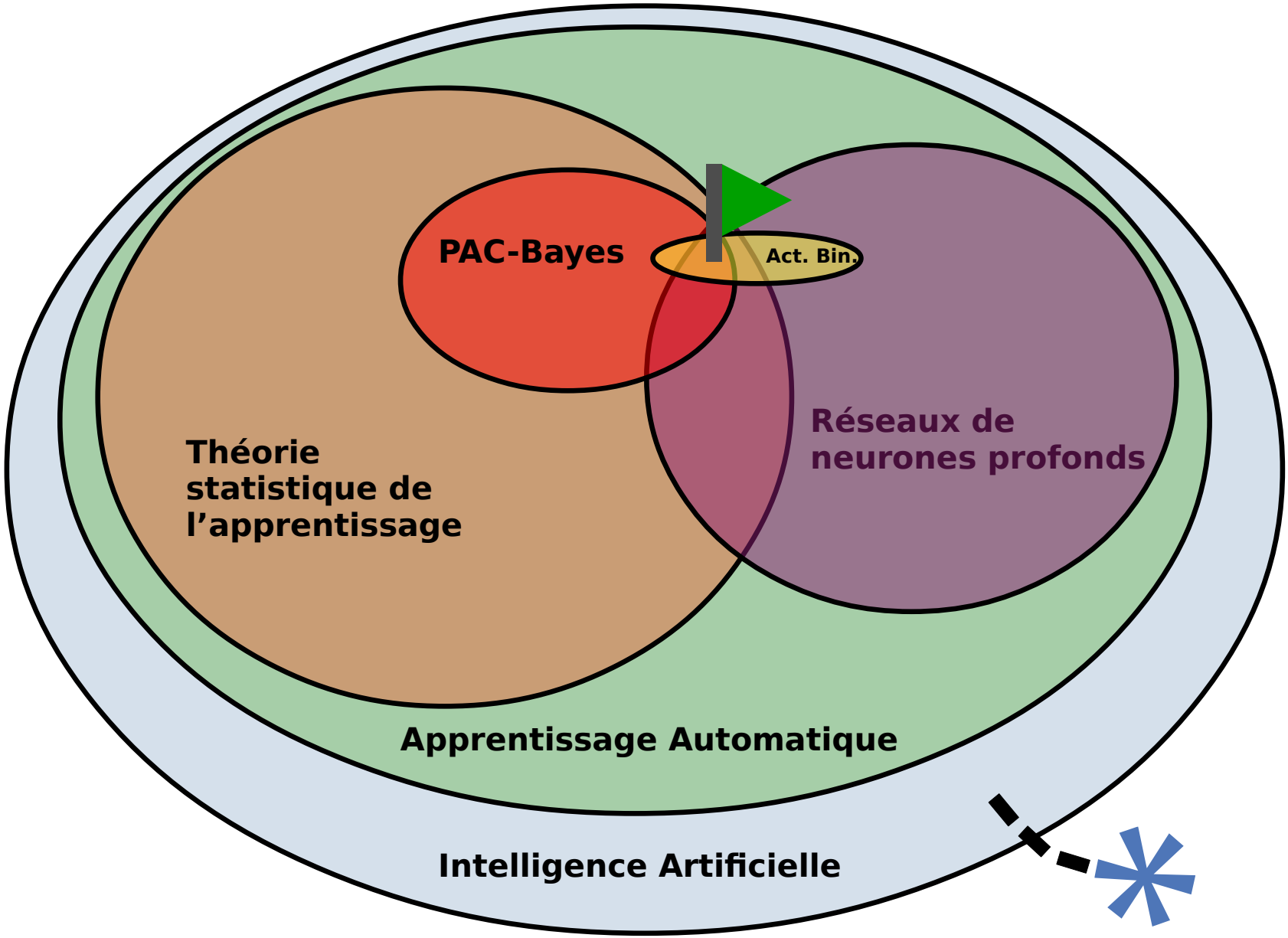

PAC-Bayes
Act. Bin.
Théorie
statistique de
l'apprentissage
Réseaux de
neurones profonds
Apprentissage Automatique
Intelligence Artificielle


PAC-Bayes
Act. Bin.
Théorie
statistique de
l'apprentissage
Réseaux de
neurones profonds
Apprentissage Automatique
Intelligence Artificielle

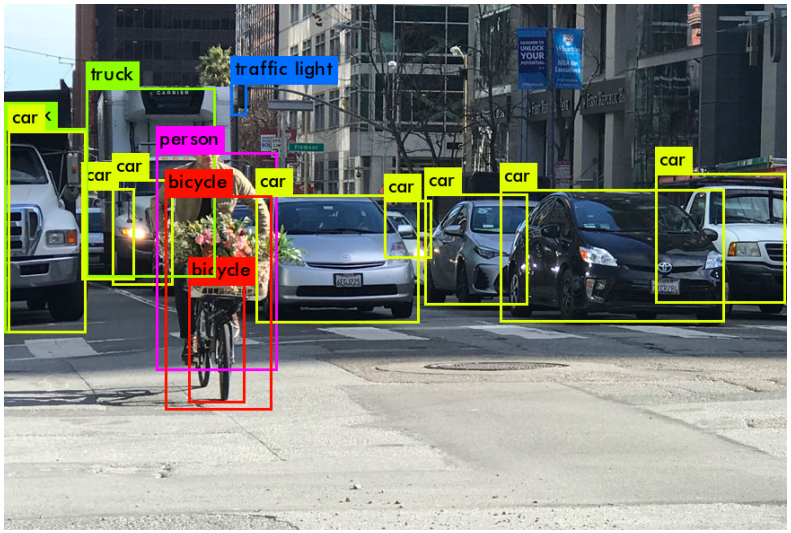
Carte


PAC-Bayes
Act. Bin.
Réseaux de
neurones profonds
Théorie
statistique de
l'apprentissage
Apprentissage Automatique
Intelligence Artificielle

Carte

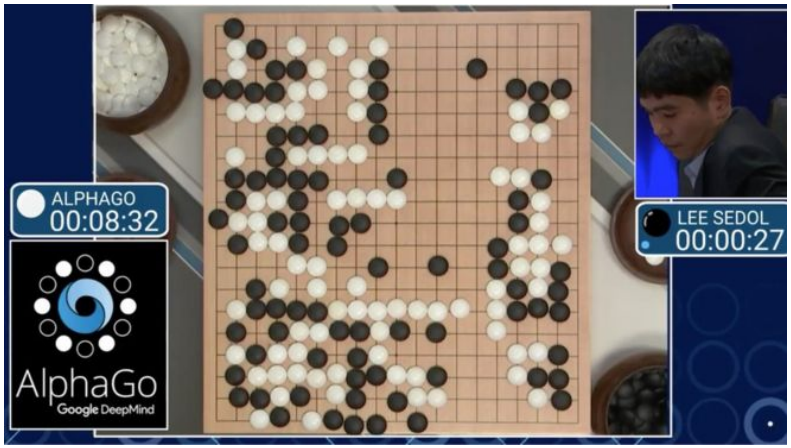

PAC-Bayes
Act. Bin.
Théorie
statistique de
l'apprentissage
Réseaux de
neurones profonds
Apprentissage Automatique
Intelligence Artificielle

Des succès!



Crédit: *https://medium.com/@jonathan_hui/*

Crédit: *www.google.com*



Crédit: *https://hicomm.bg/*

Crédit: *https://entertainment.ie/*

Des échecs...

The New York Times

Apple Card Investigated After Gender Discrimination Complaints

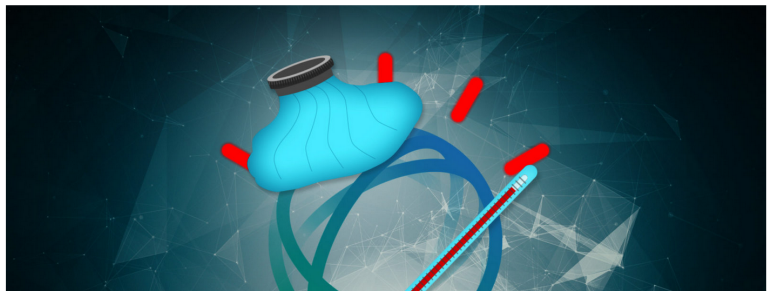
A prominent software developer said on Twitter that the credit card was "sexist" against women applying for credit.

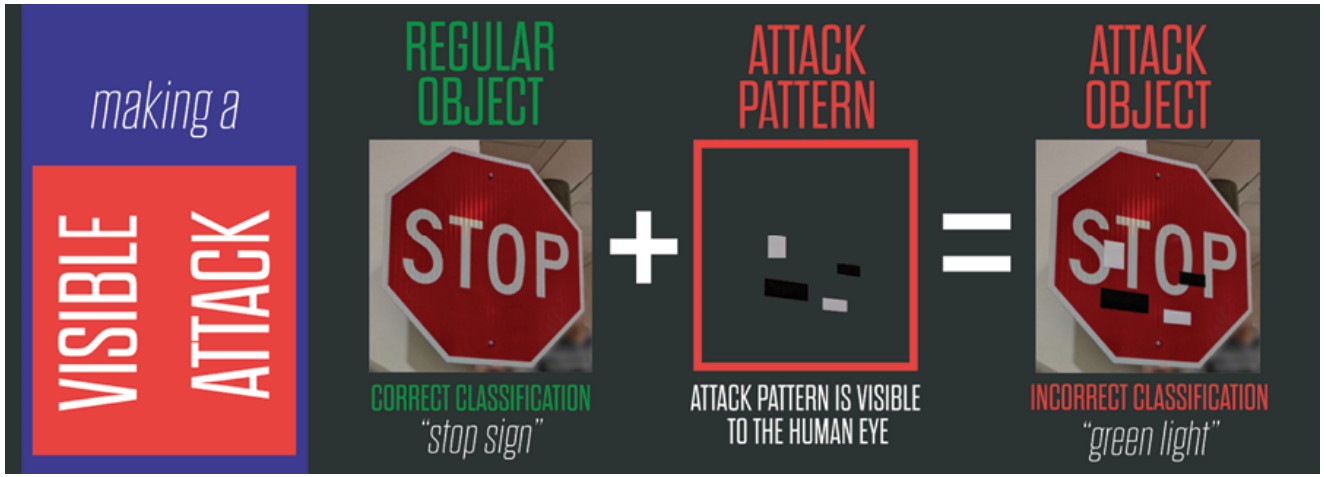
EXCLUSIVE

STAT+

IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show

By CASEY ROSS @caseymross *and* IKE SWETLITZ / JULY 25, 2018





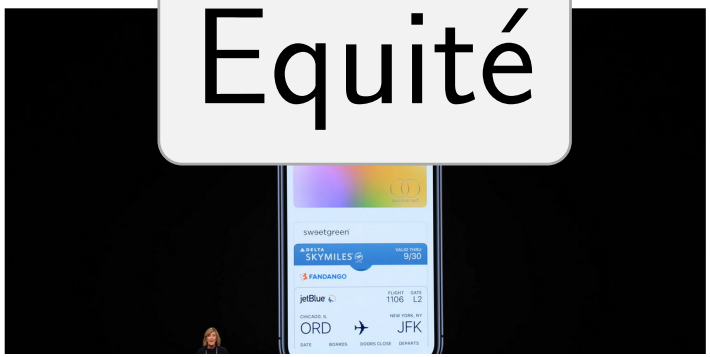
Crédit: *Comiter, Marcus. "Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It."*

Des garanties?

The New York Times

Apple Card Investigated After Gender Discrimination Complaints

A prominent software developer said on Twitter that the credit card was "sex … it.



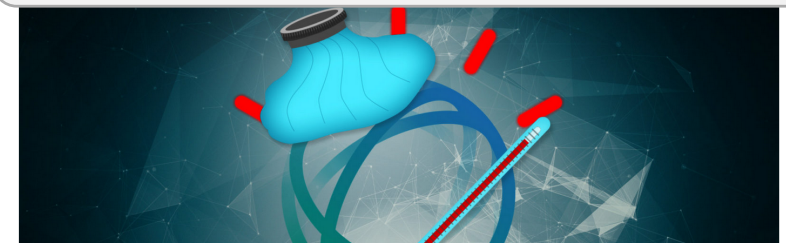
Équité

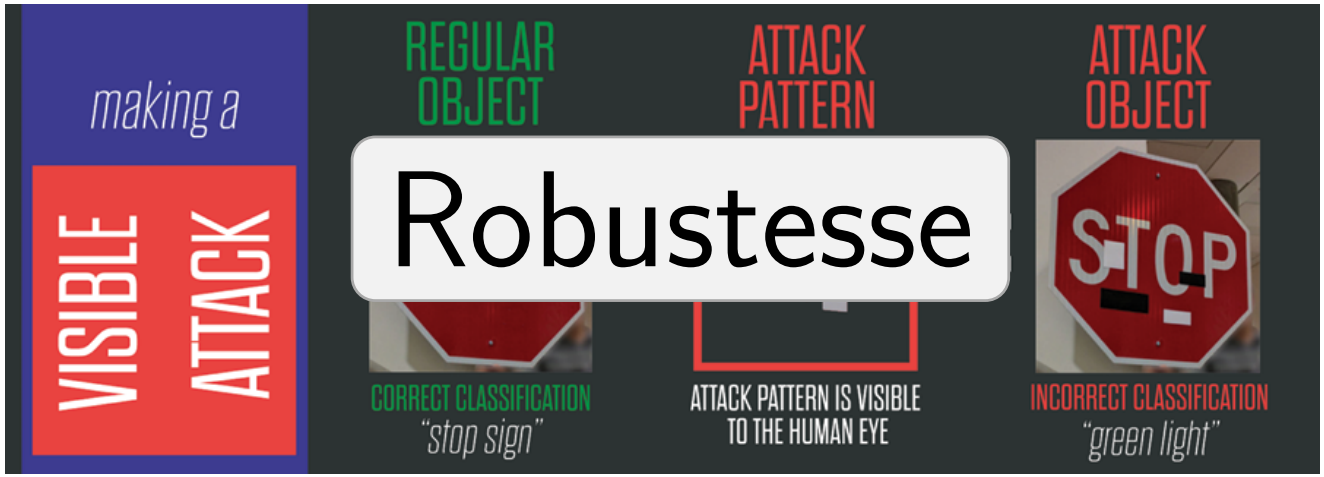
EXCLUSIVE

STAT+

IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal

Généralisation

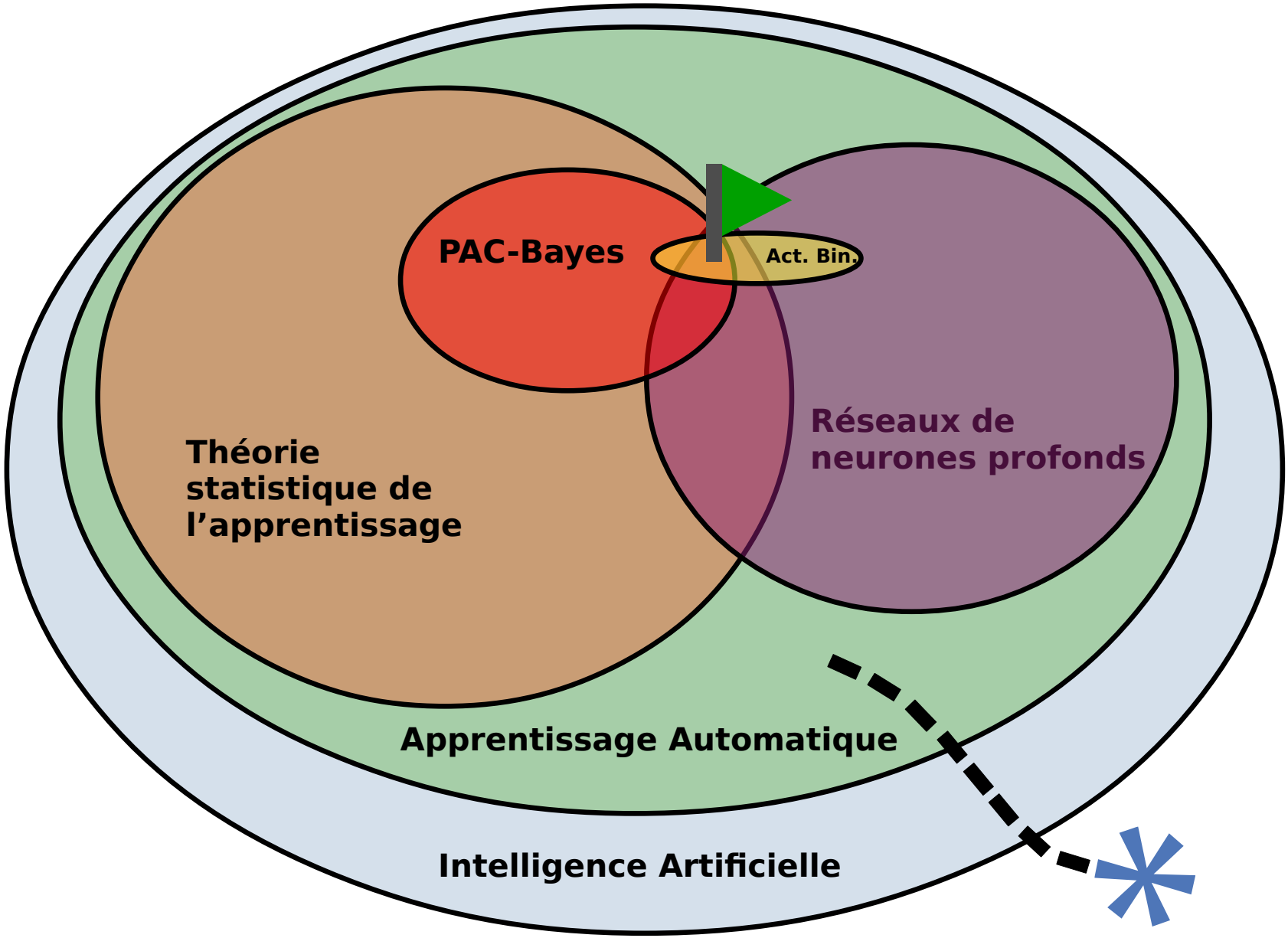




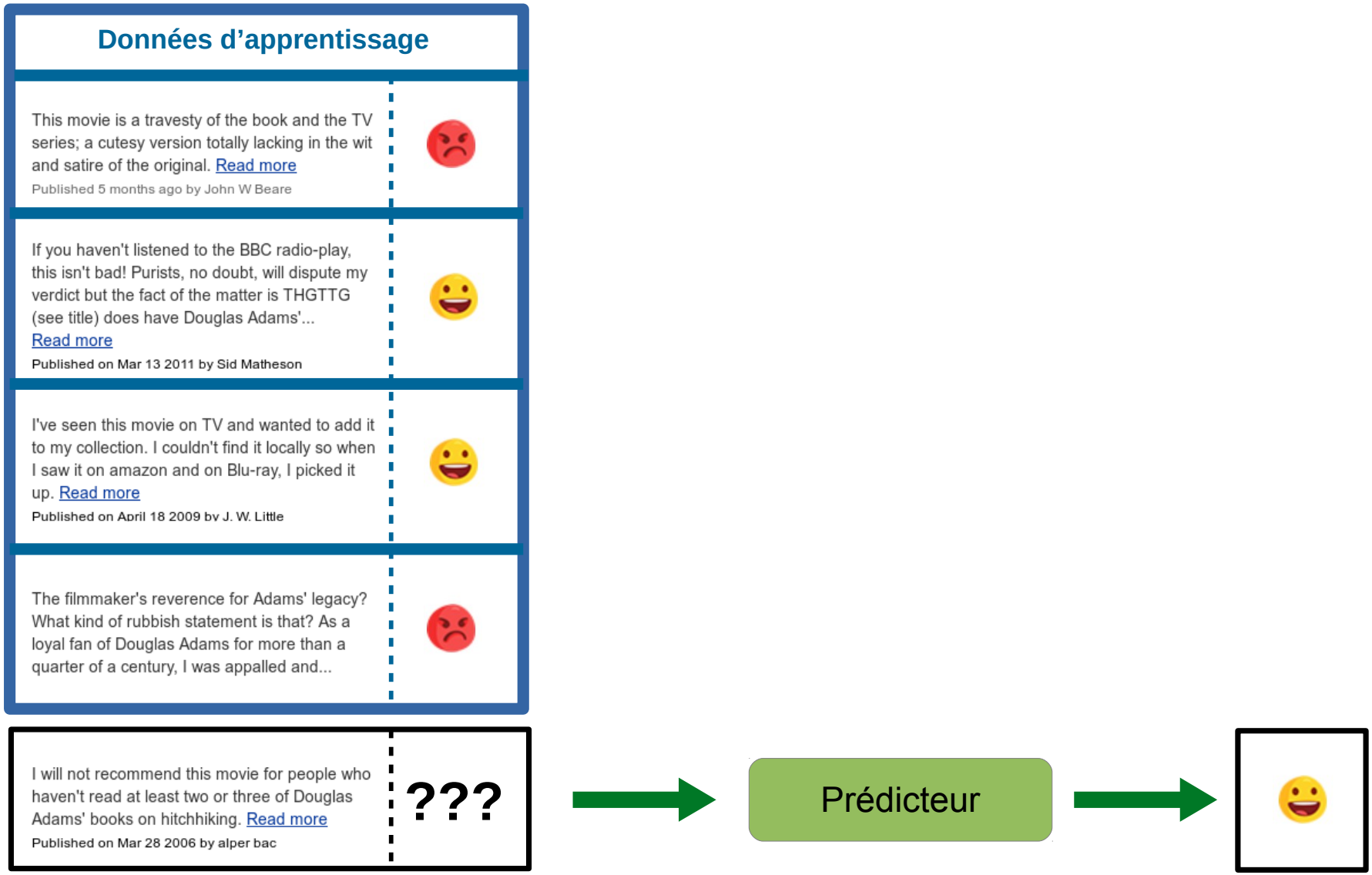
Robustesse

Crédit: *Comiter, Marcus. "Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It."*

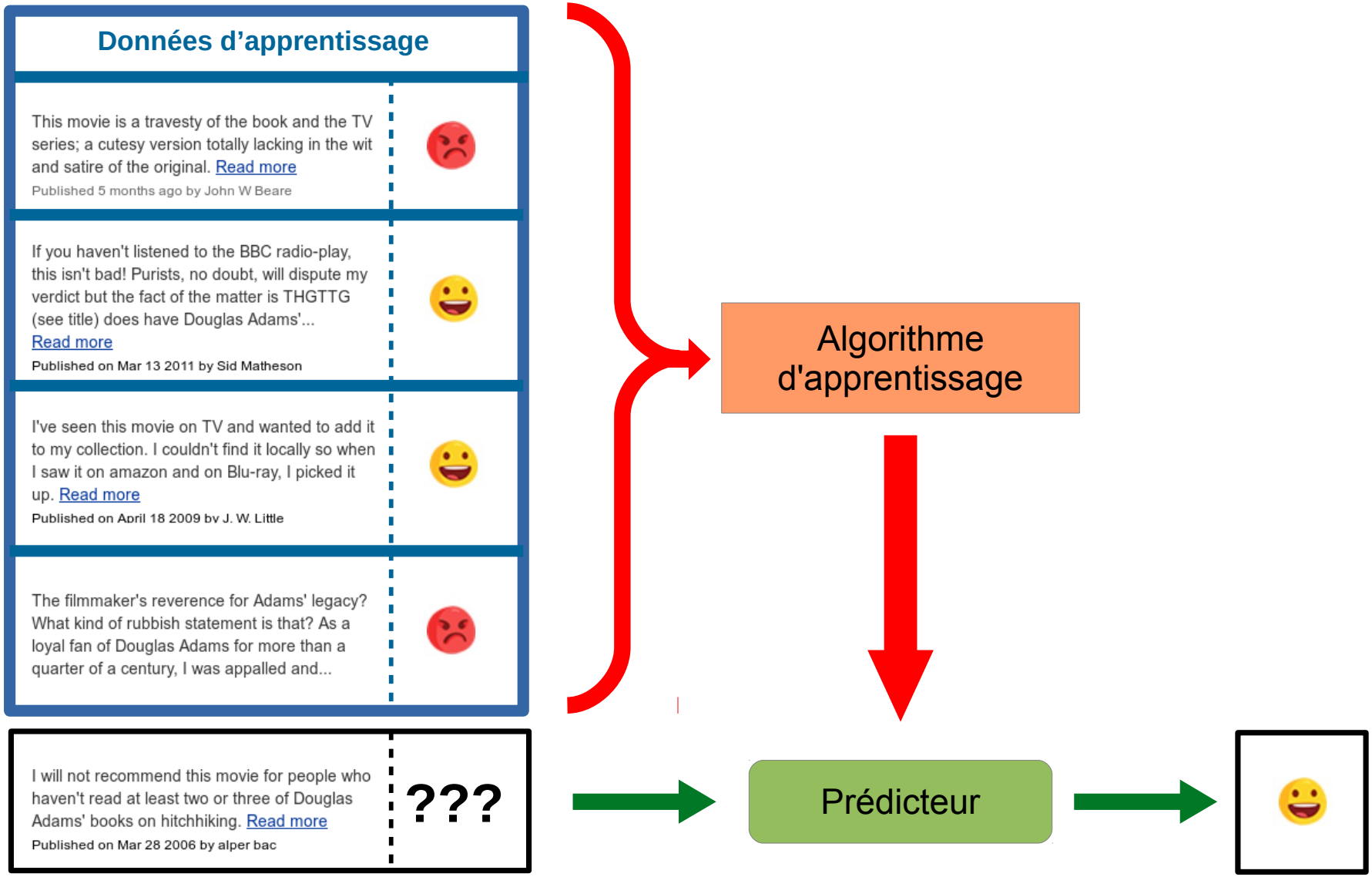
Carte


PAC-Bayes
Act. Bin.
Réseaux de
neurones profonds
Théorie
statistique de
l'apprentissage
Apprentissage Automatique
Intelligence Artificielle

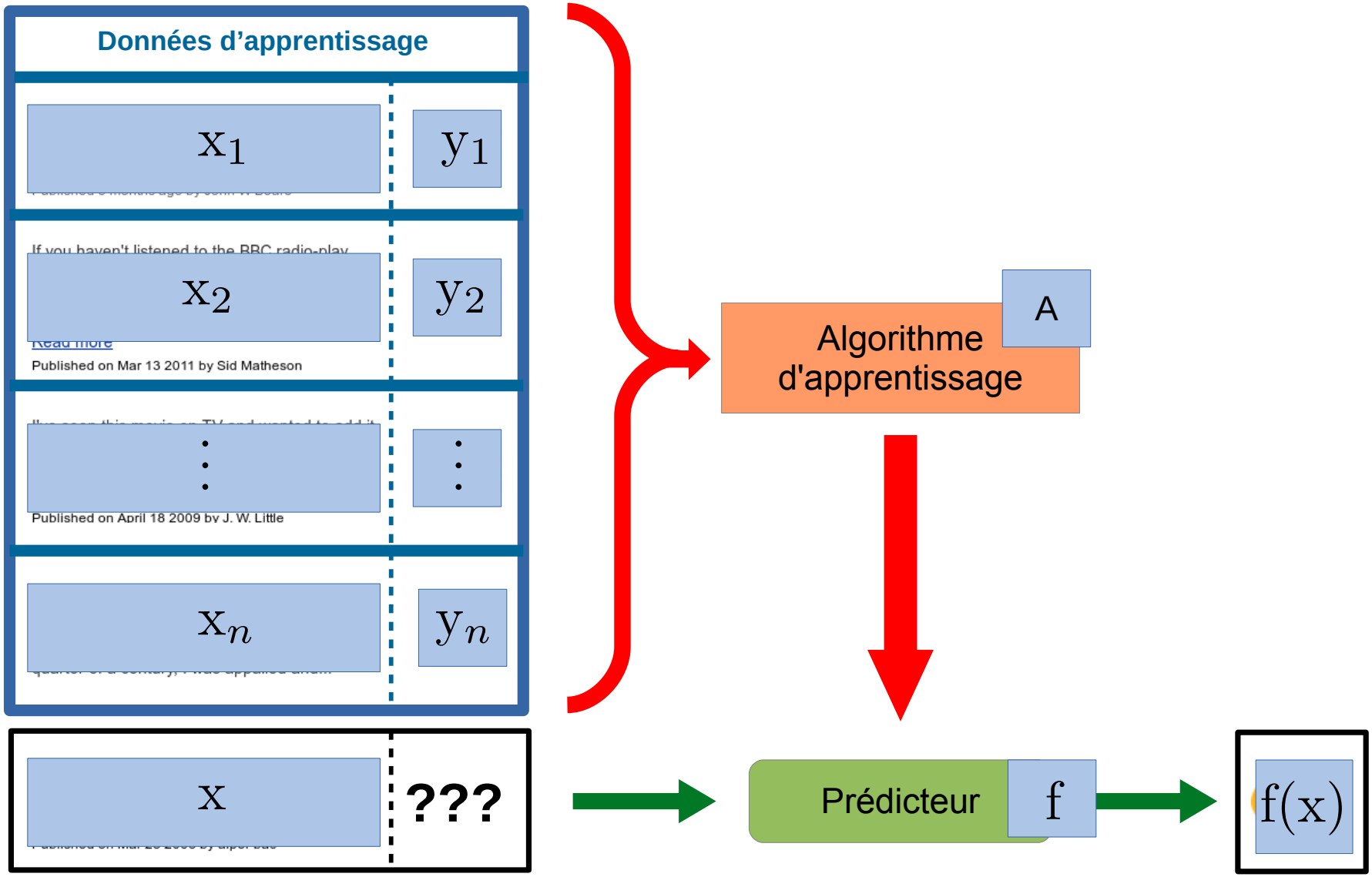
Apprentissage automatique: Exemple



Apprentissage automatique: Exemple



Abstraction mathématique



Définitions

Observation d'apprentissage

Une observation $(x, y) \in \mathcal{X} \times \mathcal{Y}$ est une paire **description-étiquette**.

Définitions

Observation d'apprentissage

Une observation $(x, y) \in \mathcal{X} \times \mathcal{Y}$ est une paire **description-étiquette**.

Ensemble d'apprentissage

$$S = \{ (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \}$$

Définitions

Observation d'apprentissage

Une observation $(x, y) \in \mathcal{X} \times \mathcal{Y}$ est une paire **description-étiquette**.

Ensemble d'apprentissage

$$S = \{ (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \}$$

Prédicteur

$$f : \mathcal{X} \to \mathcal{Y}, \quad f \in \mathcal{F}$$

Définitions

Observation d'apprentissage

Une observation $(x, y) \in \mathcal{X} \times \mathcal{Y}$ est une paire **description-étiquette**.

Ensemble d'apprentissage

$$S = \{ (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \}$$

Prédicteur

$$f : \mathcal{X} \to \mathcal{Y}, \quad f \in \mathcal{F}$$

Algorithme d'apprentissage

$$A(S) \longrightarrow f$$

Définitions

Observation d'apprentissage

Une observation $(x, y) \in \mathcal{X} \times \mathcal{Y}$ est une paire **description-étiquette**.

Ensemble d'apprentissage

$$S = \{ (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \}$$

Prédicteur

$$f : \mathcal{X} \to \mathcal{Y}, \quad f \in \mathcal{F}$$

Algorithme d'apprentissage

$$A(S) \longrightarrow f$$

Fonction de perte

$$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$$

Définitions

Observation d'apprentissage

Une observation $(x, y) \in \mathcal{X} \times \mathcal{Y}$ est une paire **description-étiquette**.

Ensemble d'apprentissage

$$S = \{ (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \}$$

Prédicteur

$$f : \mathcal{X} \to \mathcal{Y}, \quad f \in \mathcal{F}$$

Algorithme d'apprentissage

$$A(S) \longrightarrow f$$

Fonction de perte

$$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$$

Exemple: perte zéro-un

$$\ell_{01}(f(x), y) = \begin{cases} 0 \text{ si } f(x) = y \\ 1 \text{ sinon.} \end{cases}$$

Définitions

Observation d'apprentissage

Une observation $(x, y) \in \mathcal{X} \times \mathcal{Y}$ est une paire **description-étiquette**.

Ensemble d'apprentissage

$$S = \{ (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \}$$

Prédicteur

$$f : \mathcal{X} \to \mathcal{Y}, \quad f \in \mathcal{F}$$

Algorithme d'apprentissage

$$A(S) \longrightarrow f$$

Fonction de perte

$$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$$

Exemple: perte zéro-un

$$\ell_{01}(f(x), y) = \begin{cases} 0 \text{ si } f(x) = y \\ 1 \text{ sinon.} \end{cases}$$

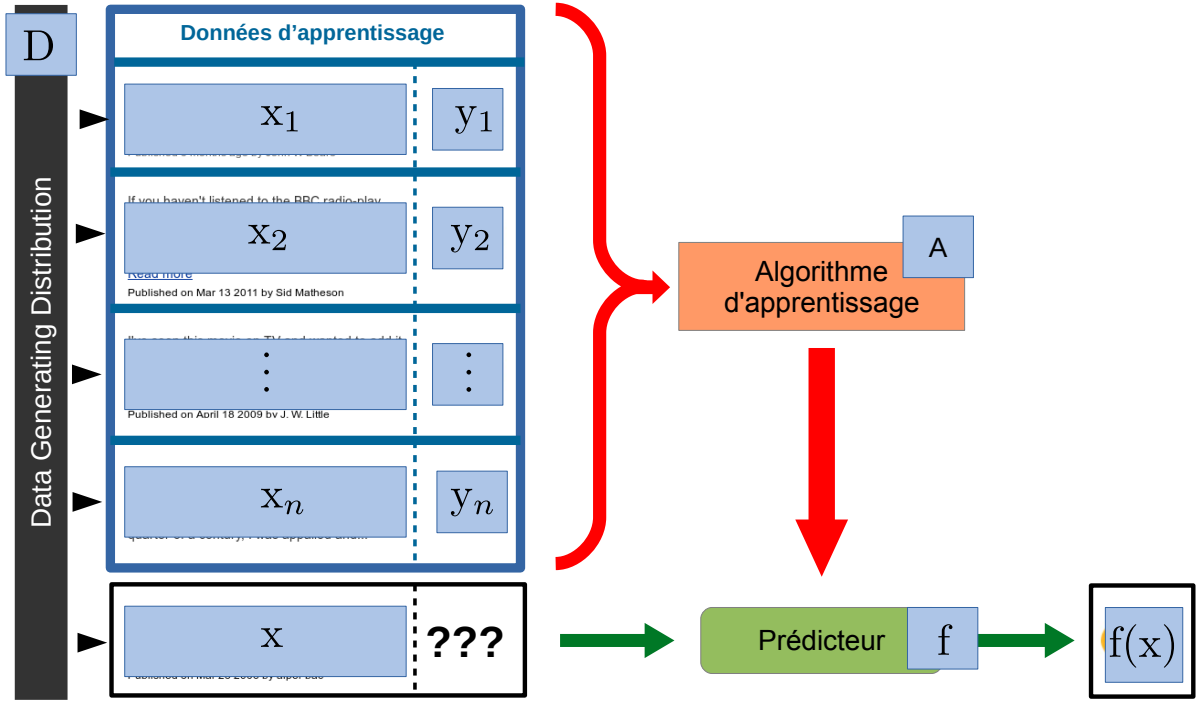
Perte empirique

$$\widehat{\mathcal{L}}_S(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(\mathbf{x}_i), y_i)$$

Défi de généralisation

Hypothèse *fréquentiste*

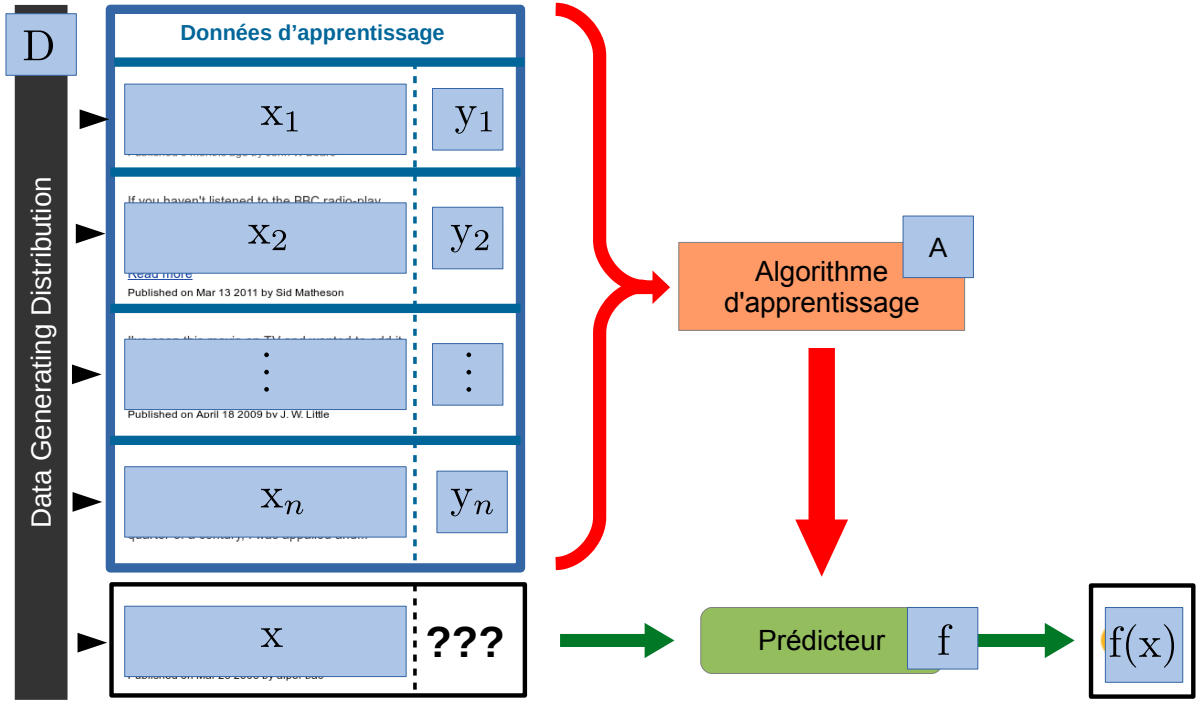
$$S = \{ (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \} \sim D^n$$



Défi de généralisation

Hypothèse *fréquentiste*

$$S = \{ (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \} \sim D^n$$



Objectif

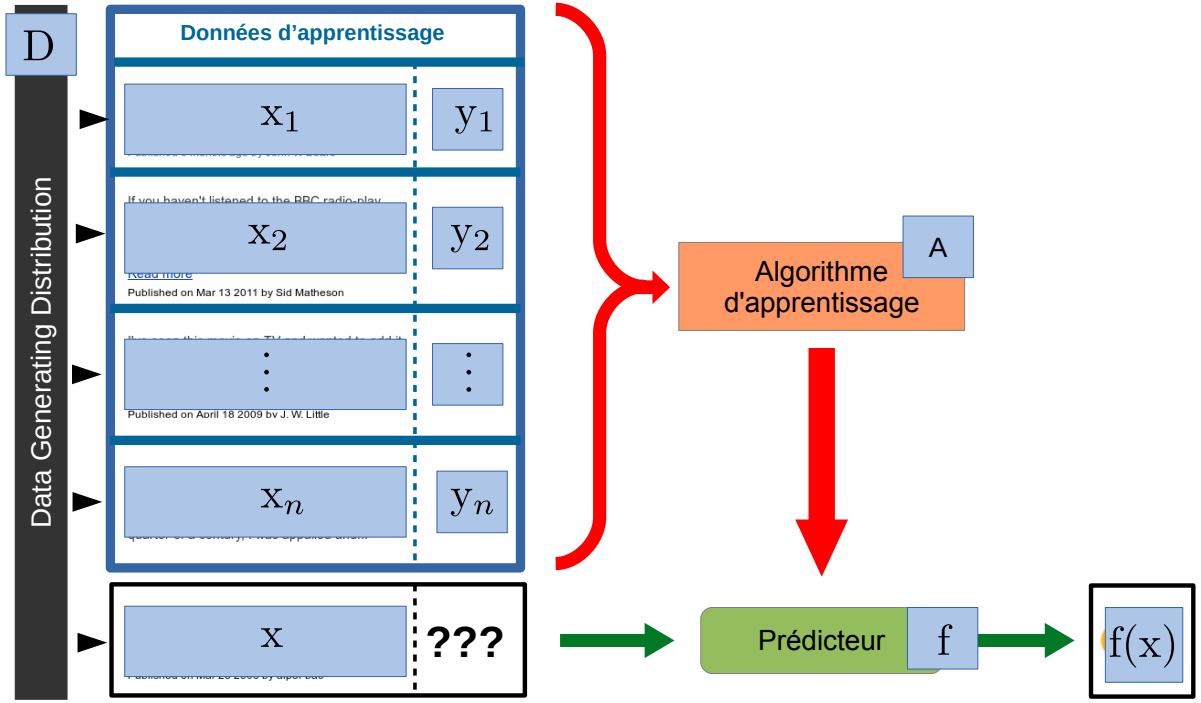
Minimiser la **perte en généralisation**

$$\mathcal{L}_D(f) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} \ell(f(\mathbf{x}), y)$$

Défi de généralisation

Hypothèse *fréquentiste*

$$S = \{ (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \} \sim D^n$$



Objectif

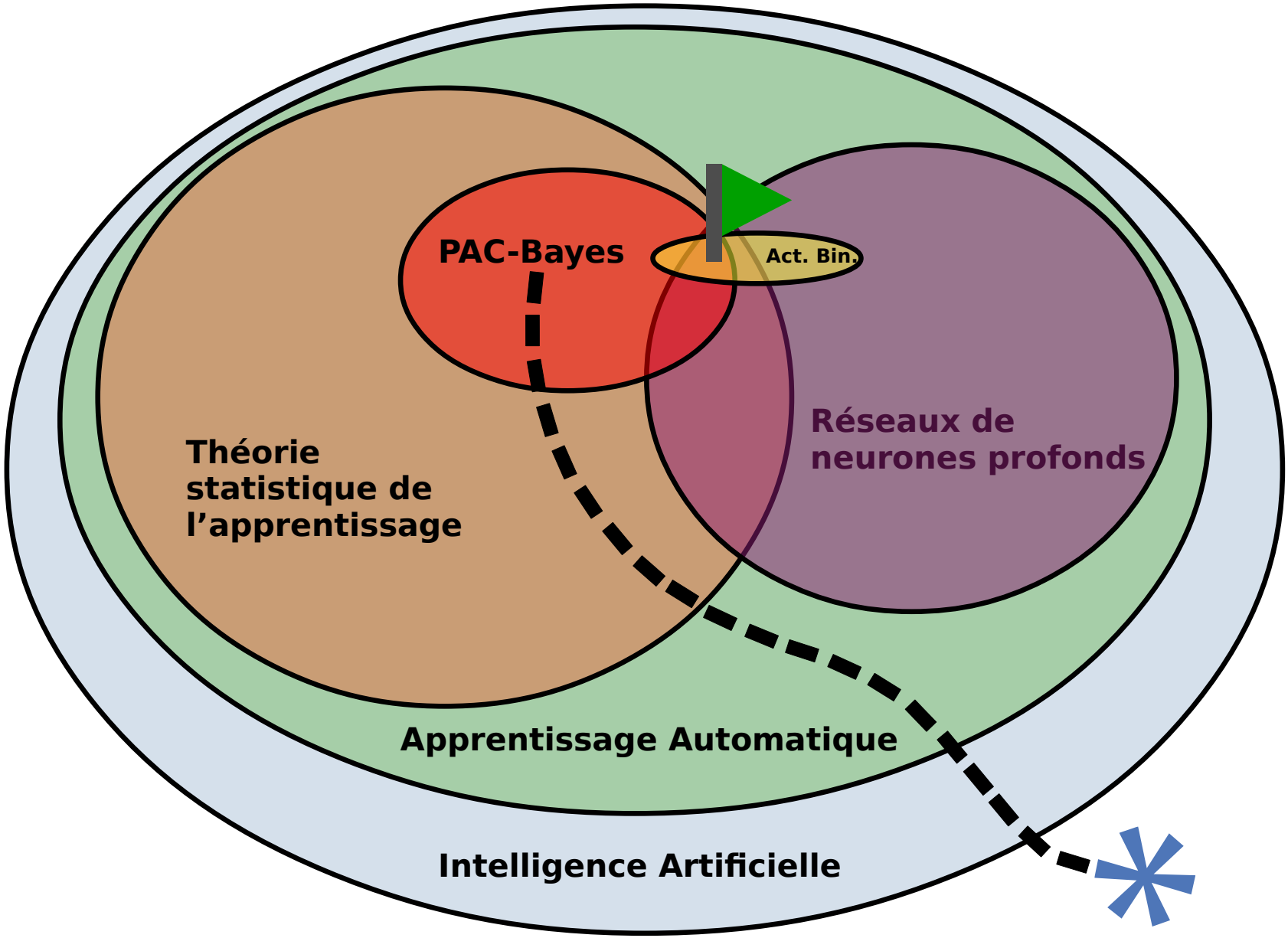
Minimiser la **perte en généralisation**

$$\mathcal{L}_D(f) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} \ell(f(\mathbf{x}), y)$$

On a accès *seulement* à la **perte empirique** sur S

$$\widehat{\mathcal{L}}_S(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(\mathbf{x}_i), y_i)$$

Carte


PAC-Bayes
Act. Bin.
Réseaux de
neurones profonds
Théorie
statistique de
l'apprentissage
Apprentissage Automatique
Intelligence Artificielle

Théorie PAC-Bayésienne (McAllester 1999)

La théorie PAC-Bayésienne est une théorie fréquentiste qui formule des garanties PAC sur des prédicteurs apparentés aux méthodes bayésiennes.

Théorie PAC-Bayésienne (McAllester 1999)

La théorie PAC-Bayésienne est une théorie fréquentiste qui formule des garanties PAC sur des prédicteurs apparentés aux méthodes bayésiennes.

Hypothèse fréquentiste

Chaque observation est générée *i.i.d.* par une **distribution de données** D.

Théorie PAC-Bayésienne (McAllester 1999)

La théorie PAC-Bayésienne est une théorie fréquentiste qui formule des garanties PAC sur des prédicteurs apparentés aux méthodes bayésiennes.

Hypothèse fréquentiste

Chaque observation est générée *i.i.d.* par une **distribution de données** D.

Bornes PAC (probablement approximativement correctes)

Avec probabilité $1-\delta$, la perte $\mathcal{L}_D$ du prédicteur f est inférieure à ϵ:

$$\Pr\left(\mathcal{L}_D(f) \leq \epsilon(\widehat{\mathcal{L}}_S(f), n, \delta, \ldots)\right) \geq 1-\delta$$

Théorie PAC-Bayésienne (McAllester 1999)

La théorie PAC-Bayésienne est une théorie fréquentiste qui formule des garanties PAC sur des prédicteurs apparentés aux méthodes bayésiennes.

Hypothèse fréquentiste

Chaque observation est générée *i.i.d.* par une **distribution de données** D.

Bornes PAC (probablement approximativement correctes)

Avec probabilité $1-\delta$, la perte $\mathcal{L}_D$ du prédicteur f est inférieure à ϵ:

$$\Pr\left(\mathcal{L}_D(f) \leq \epsilon(\widehat{\mathcal{L}}_S(f), n, \delta, \ldots)\right) \geq 1-\delta$$

Théorie PAC-Bayésienne (McAllester 1999)

La théorie PAC-Bayésienne est une théorie fréquentiste qui formule des garanties PAC sur des prédicteurs apparentés aux méthodes bayésiennes.

Hypothèse fréquentiste

Chaque observation est générée *i.i.d.* par une **distribution de données** D.

Bornes PAC (probablement approximativement correctes)

Avec probabilité $1-\delta$, la perte $\mathcal{L}_D$ du prédicteur f est inférieure à ϵ:

$$\Pr\left(\mathcal{L}_D(f) \leq \epsilon(\widehat{\mathcal{L}}_S(f), n, \delta, \ldots)\right) \geq 1-\delta$$

Théorie PAC-Bayésienne (McAllester 1999)

La théorie PAC-Bayésienne est une théorie fréquentiste qui formule des garanties PAC sur des prédicteurs apparentés aux méthodes bayésiennes.

Hypothèse fréquentiste

Chaque observation est générée *i.i.d.* par une **distribution de données** D.

Bornes PAC (probablement approximativement correctes)

Avec probabilité $1-\delta$, la perte $\mathcal{L}_D$ du prédicteur f est inférieure à ϵ:

$$\Pr\left(\mathcal{L}_D(f) \leq \epsilon(\widehat{\mathcal{L}}_S(f), n, \delta, \ldots)\right) \geq 1-\delta$$

Inspiration bayésienne

- Considère une **agrégation** de prédicteurs $\mathbf{E}_{f\sim Q}\,\mathcal{L}_D(f)$ (distribution Q)
- Incorpore des **connaissances a priori** sur le problème (distribution P)

Théorème PAC-Bayésien

Bornes PAC (probablement approximativement correctes)

Avec probabilité $1-\delta$, la perte $\mathcal{L}_D$ du prédicteur f est inférieure à ϵ:

$$\Pr\left(\mathcal{L}_D(f) \leq \epsilon(\widehat{\mathcal{L}}_S(f), n, \delta, \ldots)\right) \geq 1-\delta$$

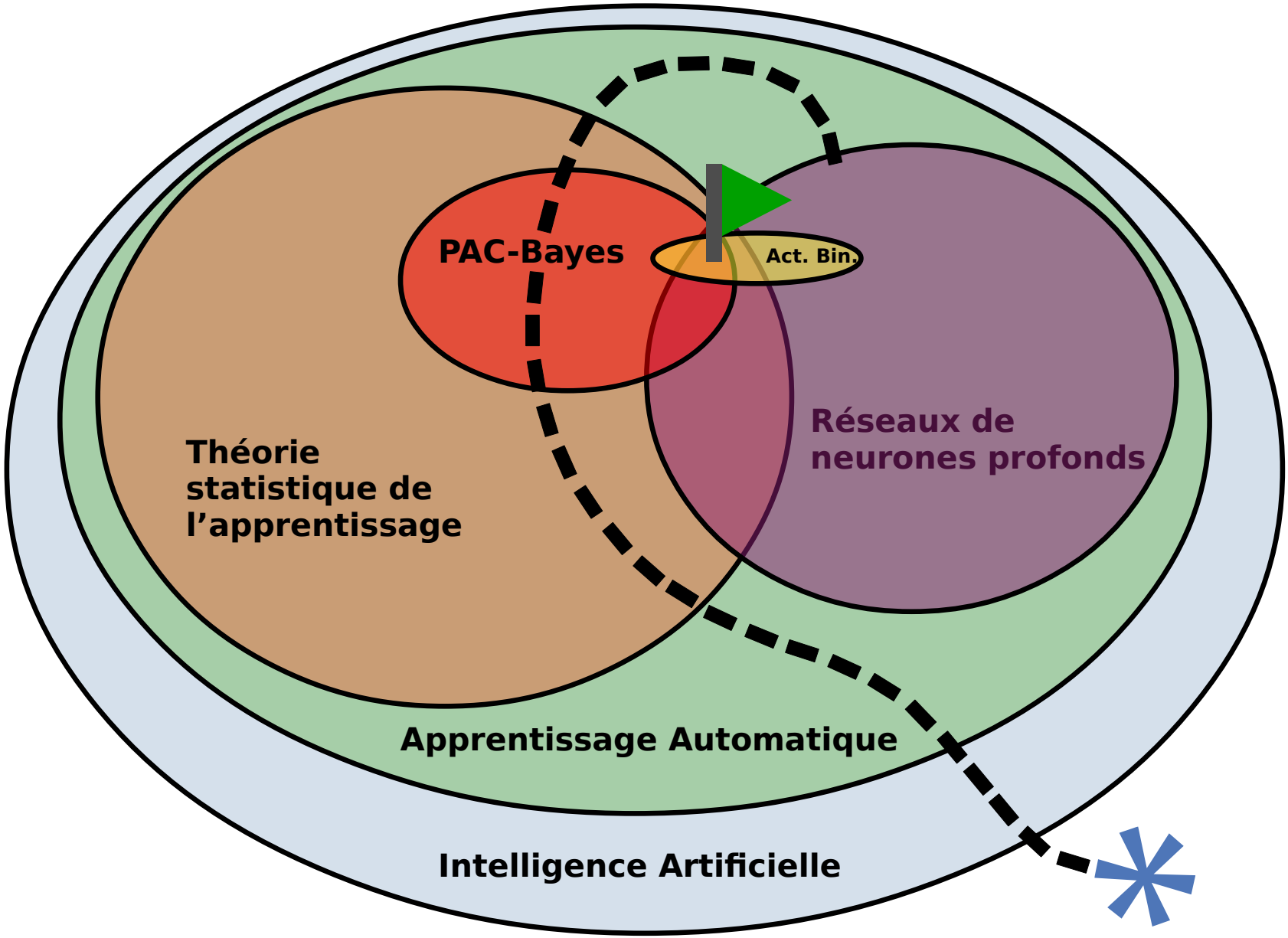
Borne PAC-Bayésienne

$$\Pr_{S\sim D^n}\left(\forall\, Q \text{ sur } \mathcal{F}, \forall C > 0 : \mathop{\mathbf{E}}_{f\sim Q} \mathcal{L}_D(f) \leq \epsilon\left(\mathop{\mathbf{E}}_{f\sim Q} \widehat{\mathcal{L}}_S(f), n, \delta, P, Q, C\right)\right) \geq 1-\delta$$

$$\epsilon(\ldots) = \frac{1}{1-e^{-C}}\left(1 - \exp\left(-C \mathop{\mathbf{E}}_{f\sim Q} \widehat{\mathcal{L}}_S(f) - \frac{1}{n}\left[\mathrm{KL}(Q\|P) + \ln\frac{2\sqrt{n}}{\delta}\right]\right)\right),$$

où $\mathrm{KL}(Q\|P) = \mathop{\mathbf{E}}_{h\sim Q} \ln\frac{Q(f)}{P(f)}$ est la **divergence de Kullback-Leibler**.

Carte


PAC-Bayes
Act. Bin.
Réseaux de
neurones profonds
Théorie
statistique de
l'apprentissage
Apprentissage Automatique
Intelligence Artificielle

Réseau de neurones

Problème de classification binaire:

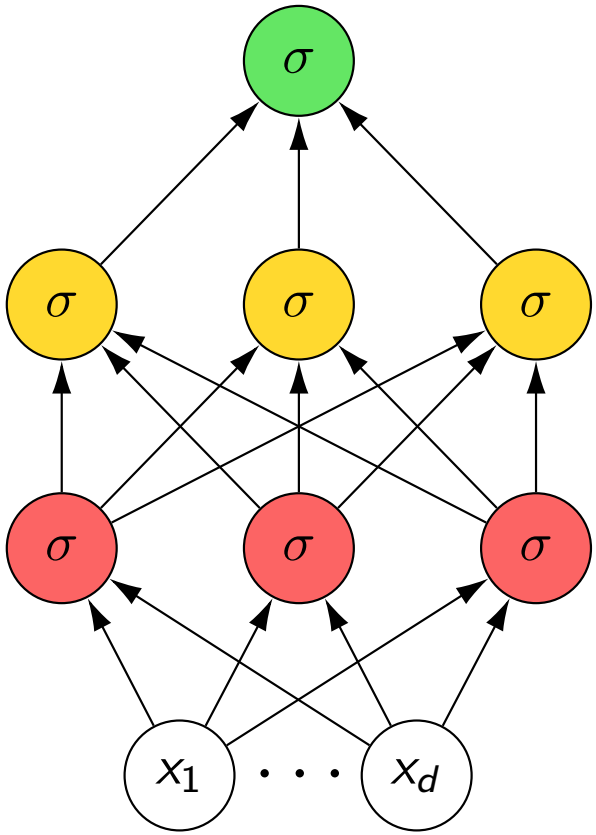
- $\mathbf{x} \in \mathbb{R}^d$
- $y \in \{-1, 1\}$

Architecture:

- L couches *pleinement connectées*
- La k-ième couche possède d_k neurones
- $\sigma : \mathbb{R} \to \mathbb{R}$ est une *fonction d'activation*

Paramètres:

- $\mathbf{W}_k \in \mathbb{R}^{d_k \times d_{k-1}}$ sont des matrices de poids.
- $\theta = \mathrm{vec}\left(\{\mathbf{W}_k\}_{k=1}^{L}\right) \in \mathbb{R}^D$



Réseau de neurones

Problème de classification binaire:

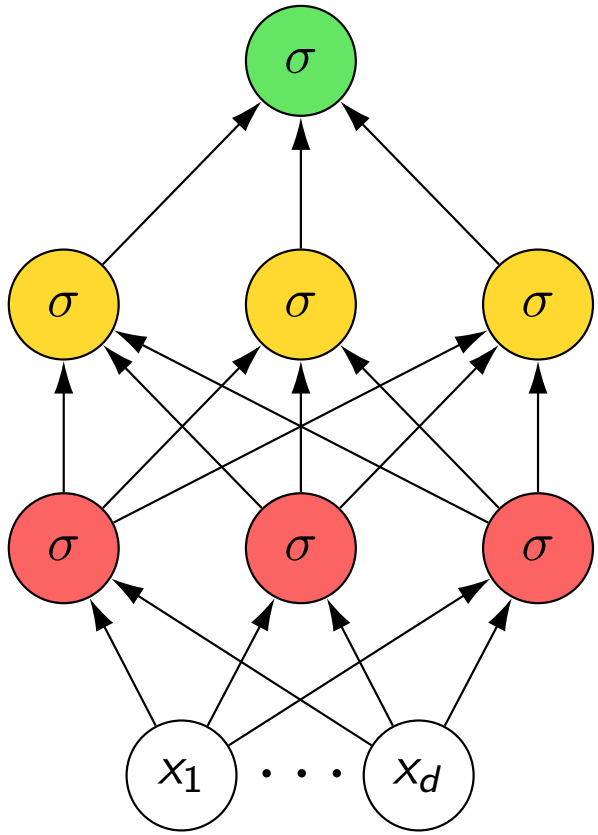
- $\mathbf{x} \in \mathbb{R}^d$
- $y \in \{-1, 1\}$

Architecture:

- L couches *pleinement connectées*
- La k-ième couche possède d_k neurones
- $\sigma : \mathbb{R} \to \mathbb{R}$ est une *fonction d'activation*

Paramètres:

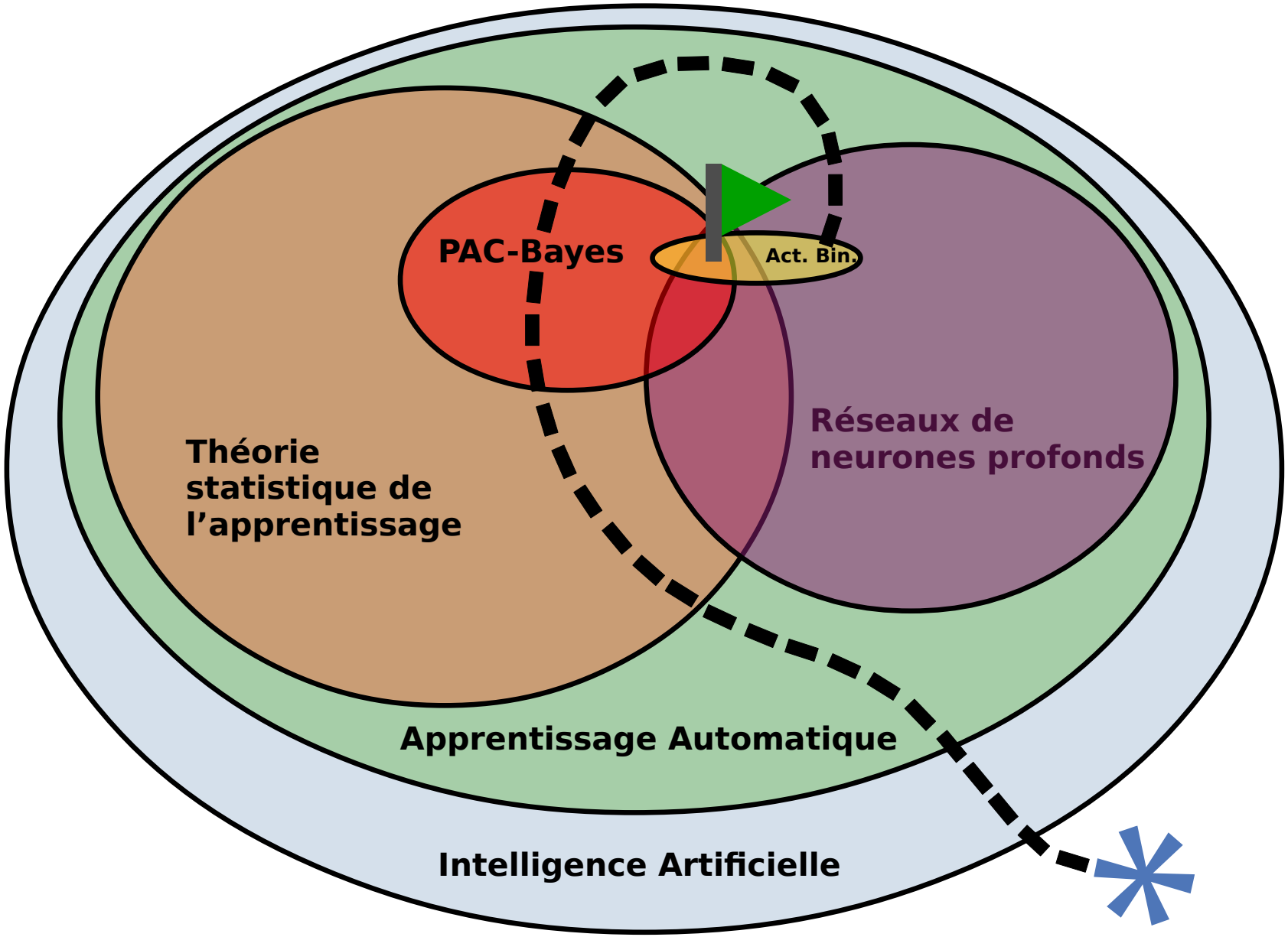
- $\mathbf{W}_k \in \mathbb{R}^{d_k \times d_{k-1}}$ sont des matrices de poids.
- $\theta = \text{vec}\left(\{\mathbf{W}_k\}_{k=1}^{L}\right) \in \mathbb{R}^D$



Prédiction

$$f_\theta(\mathbf{x}) = \sigma\left(\mathbf{w}_L \sigma\left(\mathbf{W}_{L-1}\sigma\left(\ldots\sigma\left(\mathbf{W}_1\mathbf{x}\right)\right)\right)\right).$$

Carte


PAC-Bayes
Act. Bin.
Réseaux de
neurones profonds
Théorie
statistique de
l'apprentissage
Apprentissage Automatique
Intelligence Artificielle

Réseau de neurones à *activations binaires*

Problème de classification binaire:

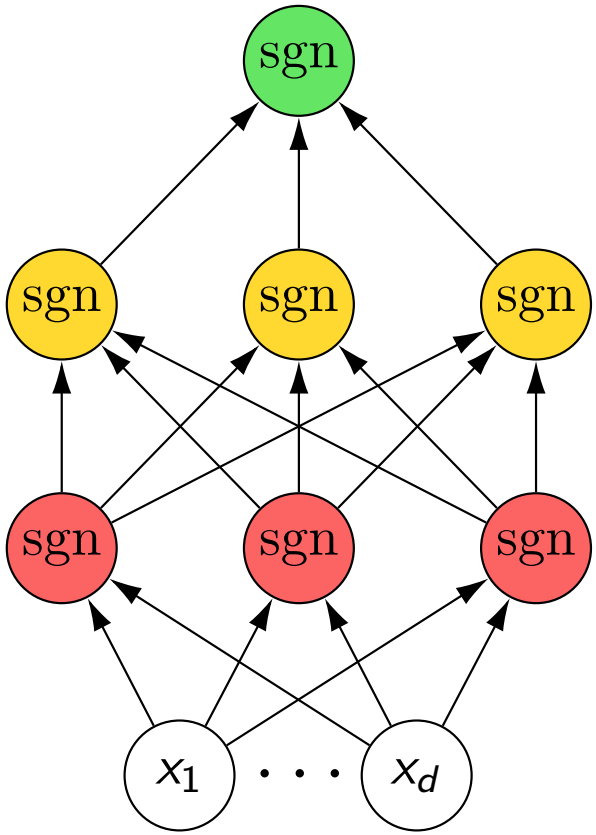
- $\mathbf{x} \in \mathbb{R}^d$
- $y \in \{-1, 1\}$

Architecture:

- L couches *pleinement connectées*
- La k-ième couche possède d_k neurones
- $\text{sgn}(a) = 1$ si $a > 0$ et $\text{sgn}(a) = -1$ sinon

Paramètres:

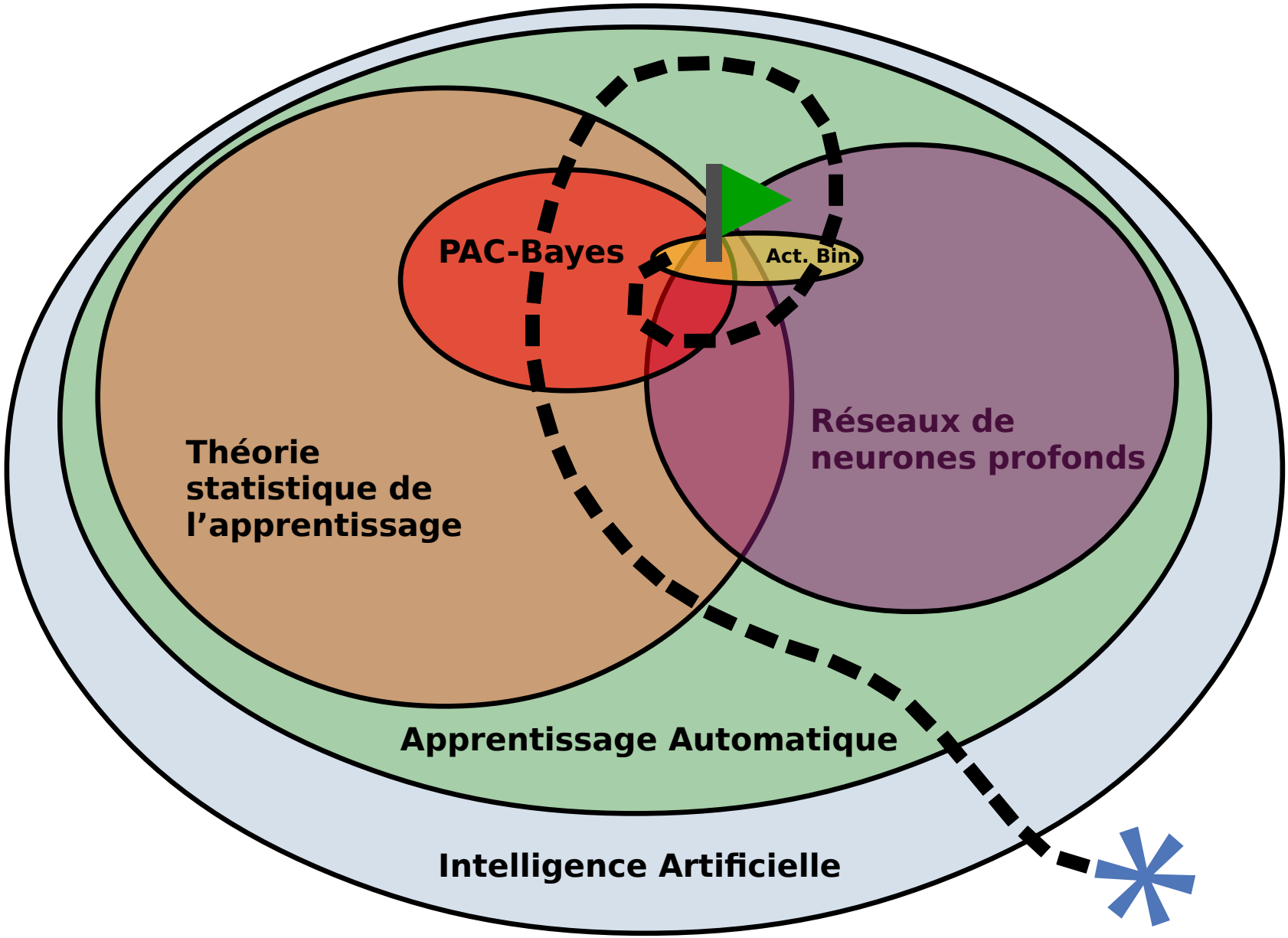
- $\mathbf{W}_k \in \mathbb{R}^{d_k \times d_{k-1}}$ sont des matrices de poids.
- $\theta = \text{vec}\left(\{\mathbf{W}_k\}_{k=1}^{L}\right) \in \mathbb{R}^D$



Prédiction

$$f_\theta(\mathbf{x}) = \text{sgn}(\mathbf{w}_L \text{sgn}(\mathbf{W}_{L-1} \text{sgn}(\ldots \text{sgn}(\mathbf{W}_1 \mathbf{x})))) .$$

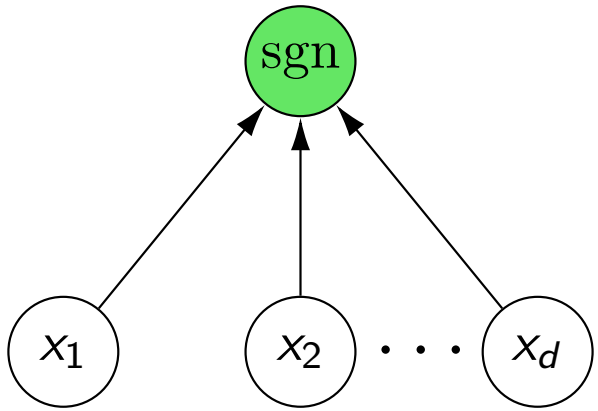
Carte


PAC-Bayes
Act. Bin.
Réseaux de
neurones profonds
Théorie
statistique de
l'apprentissage
Apprentissage Automatique
Intelligence Artificielle

Une couche

"PAC-Bayesian Learning of Linear Classifiers" (Germain et al., 2009)

$f_{\mathbf{w}}(\mathbf{x}) = \operatorname{sgn}(\mathbf{w} \cdot \mathbf{x})$, avec $\mathbf{w} \in \mathbb{R}^d$.



Une couche

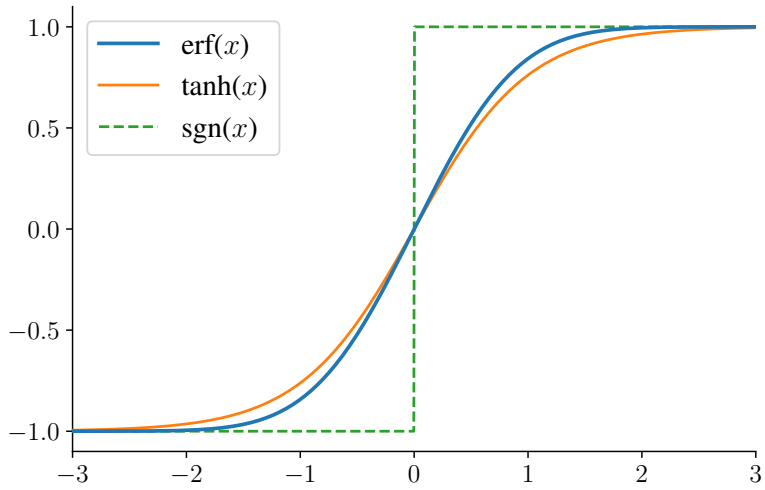
"PAC-Bayesian Learning of Linear Classifiers" (Germain et al., 2009)

$f_{\mathbf{w}}(\mathbf{x}) = \operatorname{sgn}(\mathbf{w} \cdot \mathbf{x})$, avec $\mathbf{w} \in \mathbb{R}^d$.

Analyse PAC-Bayésienne:

- Espace des prédicteurs $\mathcal{F}_d = \{f_{\mathbf{v}} | \mathbf{v} \in \mathbb{R}^d\}$
- Prior gaussien $P_{\mathbf{w}_0} = \mathcal{N}(\mathbf{w}_0, I_d)$ sur $\mathcal{F}_d$
- Posterior gaussien $Q_{\mathbf{w}} = \mathcal{N}(\mathbf{w}, I_d)$ sur $\mathcal{F}_d$
- Prédicteur

$$F_{\mathbf{w}}(\mathbf{x}) = \mathop{\mathbf{E}}_{\mathbf{v} \sim Q_{\mathbf{w}}} f_{\mathbf{v}}(\mathbf{x}) = \operatorname{erf}\left(\frac{\mathbf{w} \cdot \mathbf{x}}{\sqrt{2}\|\mathbf{x}\|}\right)$$



Une couche

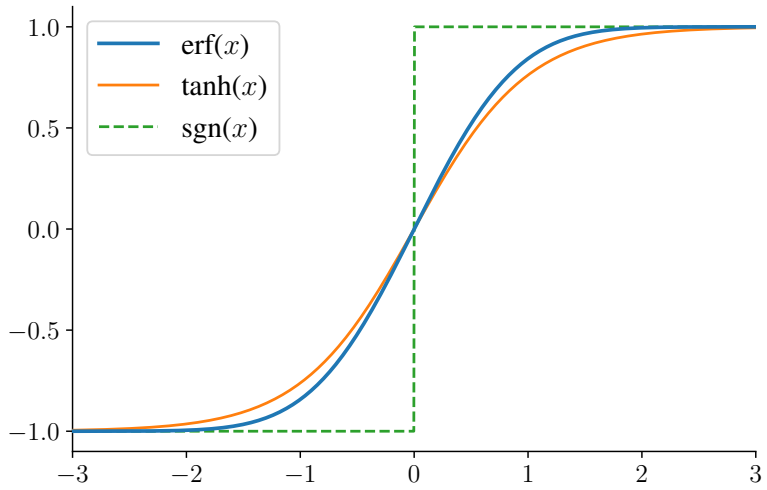
"PAC-Bayesian Learning of Linear Classifiers" (Germain et al., 2009)

$$f_{\mathbf{w}}(\mathbf{x}) = \operatorname{sgn}(\mathbf{w} \cdot \mathbf{x}), \text{ avec } \mathbf{w} \in \mathbb{R}^d.$$

Analyse PAC-Bayésienne:

- Espace des prédicteurs $\mathcal{F}_d = \{f_{\mathbf{v}} | \mathbf{v} \in \mathbb{R}^d\}$
- Prior gaussien $P_{\mathbf{w}_0} = \mathcal{N}(\mathbf{w}_0, I_d)$ sur $\mathcal{F}_d$
- Posterior gaussien $Q_{\mathbf{w}} = \mathcal{N}(\mathbf{w}, I_d)$ sur $\mathcal{F}_d$
- Prédicteur

$$F_{\mathbf{w}}(\mathbf{x}) = \mathop{\mathbf{E}}_{\mathbf{v} \sim Q_{\mathbf{w}}} f_{\mathbf{v}}(\mathbf{x}) = \operatorname{erf}\left(\frac{\mathbf{w} \cdot \mathbf{x}}{\sqrt{2}\|\mathbf{x}\|}\right)$$

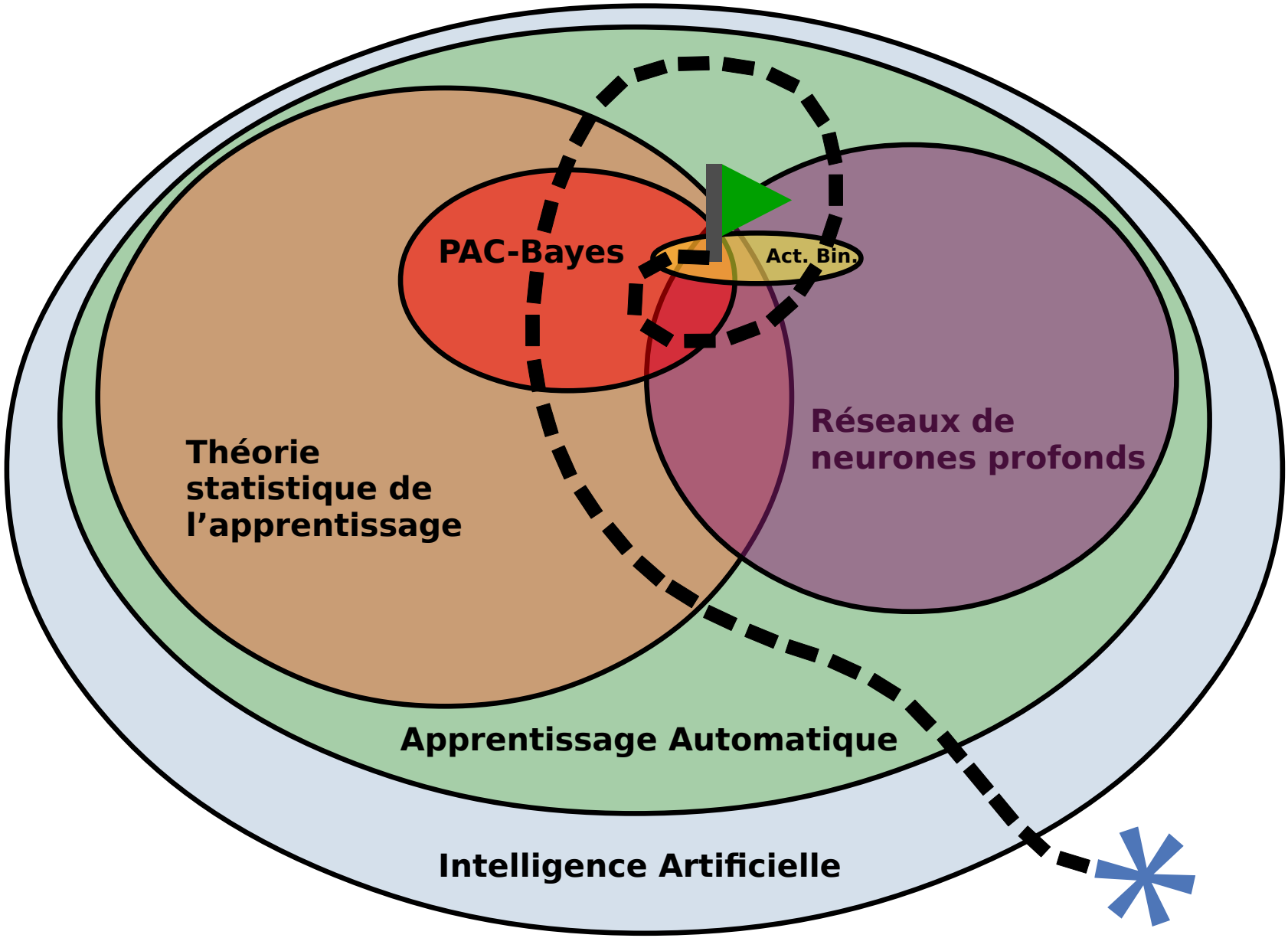


Minimisation de la borne

(perte linéaire $\ell(y, y') \stackrel{\text{def}}{=} \frac{1}{2}(1 - yy')$)

$$C\, n\, \widehat{\mathcal{L}}_S(F_{\mathbf{w}}) + \mathrm{KL}(Q_{\mathbf{w}} \| P_{\mathbf{w}_0}) = C \sum_{i=1}^{n} \operatorname{erf}\left(-y_i \frac{\mathbf{w} \cdot \mathbf{x}_i}{\sqrt{2}\|\mathbf{x}_i\|}\right) + \frac{1}{2}\|\mathbf{w} - \mathbf{w}_0\|^2 .$$

Carte

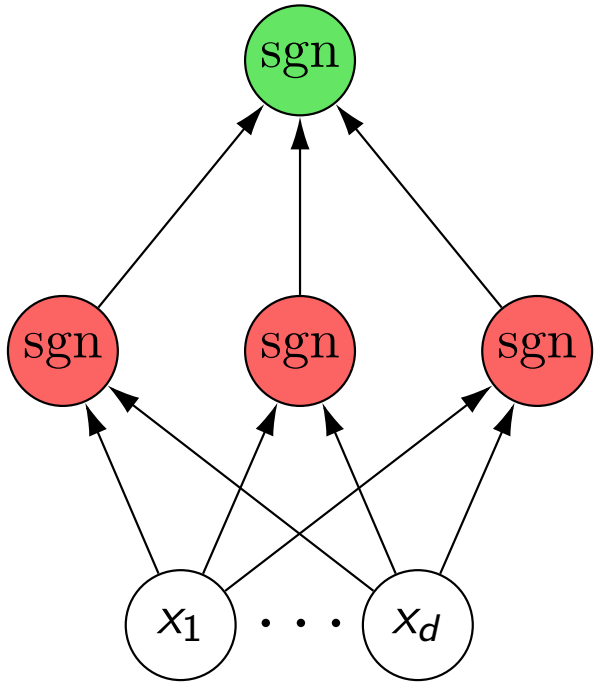

PAC-Bayes
Act. Bin.
Réseaux de
neurones profonds
Théorie
statistique de
l'apprentissage
Apprentissage Automatique
Intelligence Artificielle

Deux couches

Posterior $Q_\theta = \mathcal{N}(\theta, I_D)$ sur l'ensemble des réseaux $\mathcal{F}_D = \{f_{\tilde{\theta}} \,|\, \tilde{\theta} \in \mathbb{R}^D\}$, avec

$$f_\theta(\mathbf{x}) = \operatorname{sgn}\big(\mathbf{w}_2 \cdot \operatorname{sgn}(\mathbf{W}_1 \mathbf{x})\big)\,.$$

$$\begin{aligned}
F_\theta(\mathbf{x}) &= \mathop{\mathbf{E}}_{\tilde{\theta}\sim Q_\theta} f_{\tilde{\theta}(\mathbf{x})} \\
&= \int_{\mathbb{R}^{d_1 \times d_0}} Q_1(\mathbf{V}_1) \int_{\mathbb{R}^{d_1}} Q_2(\mathbf{v}_2) \operatorname{sgn}(\mathbf{v}_2 \cdot \operatorname{sgn}(\mathbf{V}_1\mathbf{x})) d\mathbf{v}_2 d\mathbf{V}_1 \\
&= \sum_{\mathbf{s}\in\{-1,1\}^{d_1}} \operatorname{erf}\left(\tfrac{\mathbf{w}_2\cdot\mathbf{s}}{\sqrt{2d_1}}\right) \int_{\mathbb{R}^{d_1\times d_0}} \mathbb{1}[\mathbf{s} = \operatorname{sgn}(\mathbf{V}_1\mathbf{x})] Q_1(\mathbf{V}_1)\, d\mathbf{V}_1 \\
&= \sum_{\mathbf{s}\in\{-1,1\}^{d_1}} \underbrace{\operatorname{erf}\left(\frac{\mathbf{w}_2\cdot\mathbf{s}}{\sqrt{2d_1}}\right)}_{F_{\mathbf{w}_2}(\mathbf{s})} \underbrace{\prod_{i=1}^{d_1}\left[\frac{1}{2} + \frac{s_i}{2}\operatorname{erf}\left(\frac{\mathbf{w}_1^i\cdot\mathbf{x}}{\sqrt{2}\,\|\mathbf{x}\|}\right)\right]}_{\Pr(\mathbf{s}|\mathbf{x},\mathbf{W}_1)}.
\end{aligned}$$



Ingrédients de la borne PAC-Bayes

Prédicteur

$$F_\theta(\mathbf{x}) = \mathop{\mathbf{E}}_{\tilde{\theta}\sim Q_\theta} f_{\tilde{\theta}}(\mathbf{x}) = \sum_{\mathbf{s}\in\{-1,1\}^{d_1}} \operatorname{erf}\left(\frac{\mathbf{w}_2\cdot\mathbf{s}}{\sqrt{2d_1}}\right) \prod_{i=1}^{d_1}\left[\frac{1}{2} + \frac{s_i}{2}\operatorname{erf}\left(\frac{\mathbf{w}_1^i\cdot\mathbf{x}}{\sqrt{2}\,\|\mathbf{x}\|}\right)\right].$$

Ingrédients de la borne PAC-Bayes

Prédicteur

$$F_\theta(\mathbf{x}) = \mathop{\mathbf{E}}_{\tilde{\theta}\sim Q_\theta} f_{\tilde{\theta}(\mathbf{x})} = \sum_{\mathbf{s}\in\{-1,1\}^{d_1}} \operatorname{erf}\left(\frac{\mathbf{w}_2\cdot\mathbf{s}}{\sqrt{2d_1}}\right) \prod_{i=1}^{d_1}\left[\frac{1}{2} + \frac{s_i}{2}\operatorname{erf}\left(\frac{\mathbf{w}_1^i\cdot\mathbf{x}}{\sqrt{2}\,\|\mathbf{x}\|}\right)\right].$$

Perte empirique

$$\widehat{\mathcal{L}}_S(F_\theta) = \mathop{\mathbf{E}}_{\tilde{\theta}\sim Q_\theta} \widehat{\mathcal{L}}_S(f_{\tilde{\theta}}) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{2} - \frac{1}{2}y_i F_\theta(\mathbf{x}_i)\right].$$

Ingrédients de la borne PAC-Bayes

Prédicteur

$$F_\theta(\mathbf{x}) = \mathop{\mathbf{E}}_{\tilde{\theta}\sim Q_\theta} f_{\tilde{\theta}(\mathbf{x})} = \sum_{\mathbf{s}\in\{-1,1\}^{d_1}} \operatorname{erf}\left(\frac{\mathbf{w}_2\cdot\mathbf{s}}{\sqrt{2d_1}}\right) \prod_{i=1}^{d_1}\left[\frac{1}{2} + \frac{s_i}{2}\operatorname{erf}\left(\frac{\mathbf{w}_1^i\cdot\mathbf{x}}{\sqrt{2}\,\|\mathbf{x}\|}\right)\right].$$

Perte empirique

$$\widehat{\mathcal{L}}_S(F_\theta) = \mathop{\mathbf{E}}_{\tilde{\theta}\sim Q_\theta} \widehat{\mathcal{L}}_S(f_{\tilde{\theta}}) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{2} - \frac{1}{2}y_i F_\theta(\mathbf{x}_i)\right].$$

Terme de complexité

$$\mathrm{KL}(Q_\theta\|P_{\theta_0}) = \frac{1}{2}\|\theta-\theta_0\|^2.$$

Ingrédients de la borne PAC-Bayes

Prédicteur

$$F_\theta(\mathbf{x}) = \mathop{\mathbf{E}}_{\tilde{\theta}\sim Q_\theta} f_{\tilde{\theta}(\mathbf{x})} = \sum_{\mathbf{s}\in\{-1,1\}^{d_1}} \operatorname{erf}\left(\frac{\mathbf{w}_2\cdot\mathbf{s}}{\sqrt{2d_1}}\right) \prod_{i=1}^{d_1}\left[\frac{1}{2} + \frac{s_i}{2}\operatorname{erf}\left(\frac{\mathbf{w}_1^i\cdot\mathbf{x}}{\sqrt{2}\,\|\mathbf{x}\|}\right)\right].$$

Perte empirique

$$\widehat{\mathcal{L}}_S(F_\theta) = \mathop{\mathbf{E}}_{\tilde{\theta}\sim Q_\theta} \widehat{\mathcal{L}}_S(f_{\tilde{\theta}}) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{2} - \frac{1}{2}y_i F_\theta(\mathbf{x}_i)\right].$$

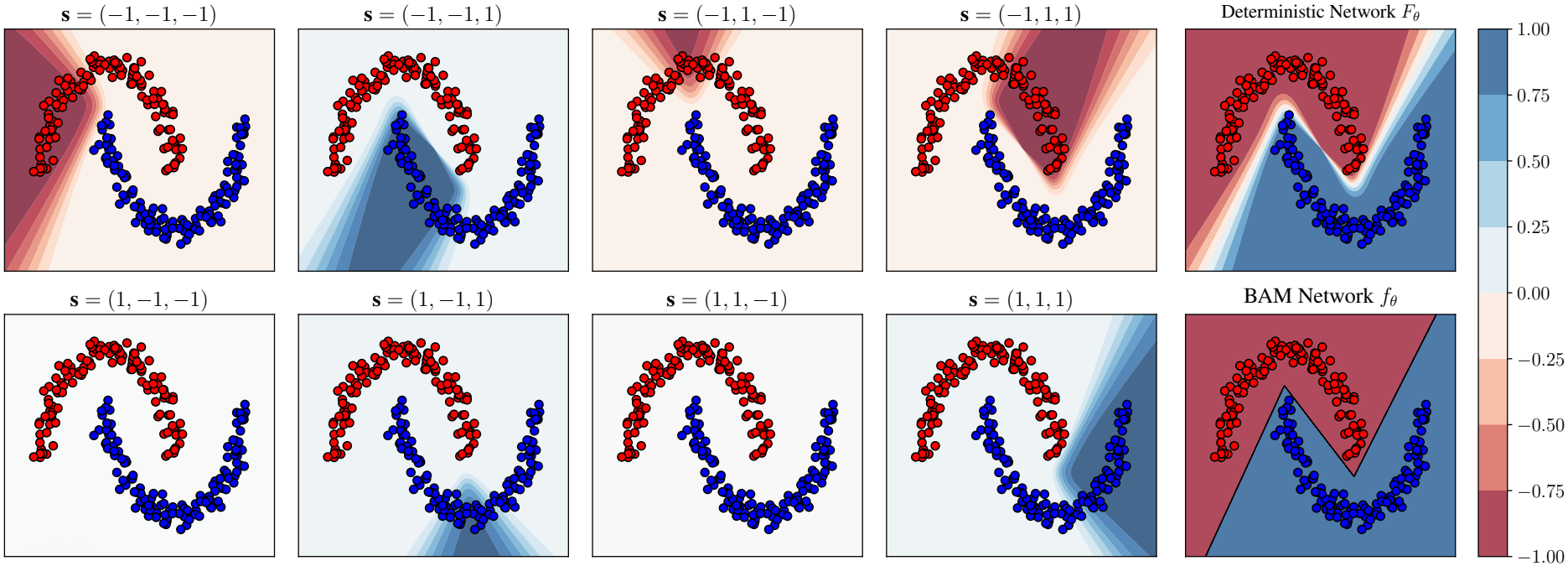
Terme de complexité

$$\mathrm{KL}(Q_\theta\|P_{\theta_0}) = \frac{1}{2}\|\theta-\theta_0\|^2.$$

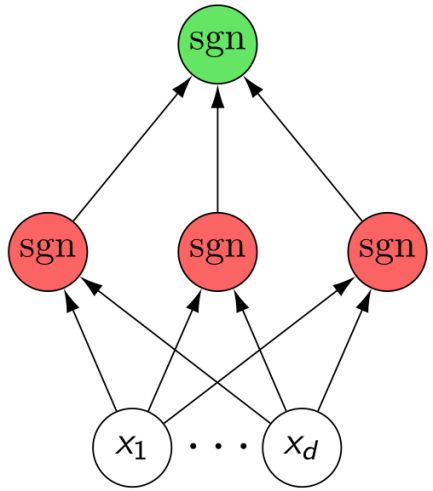
Garantie de généralisation

$$\frac{1}{1-e^{-C}}\left(1-\exp\left(-C\,\widehat{\mathcal{L}}_S(F_\theta) - \frac{1}{n}[\mathrm{KL}(Q_\theta\|P_\mu) + \ln\tfrac{2\sqrt{n}}{\delta}]\right)\right)$$

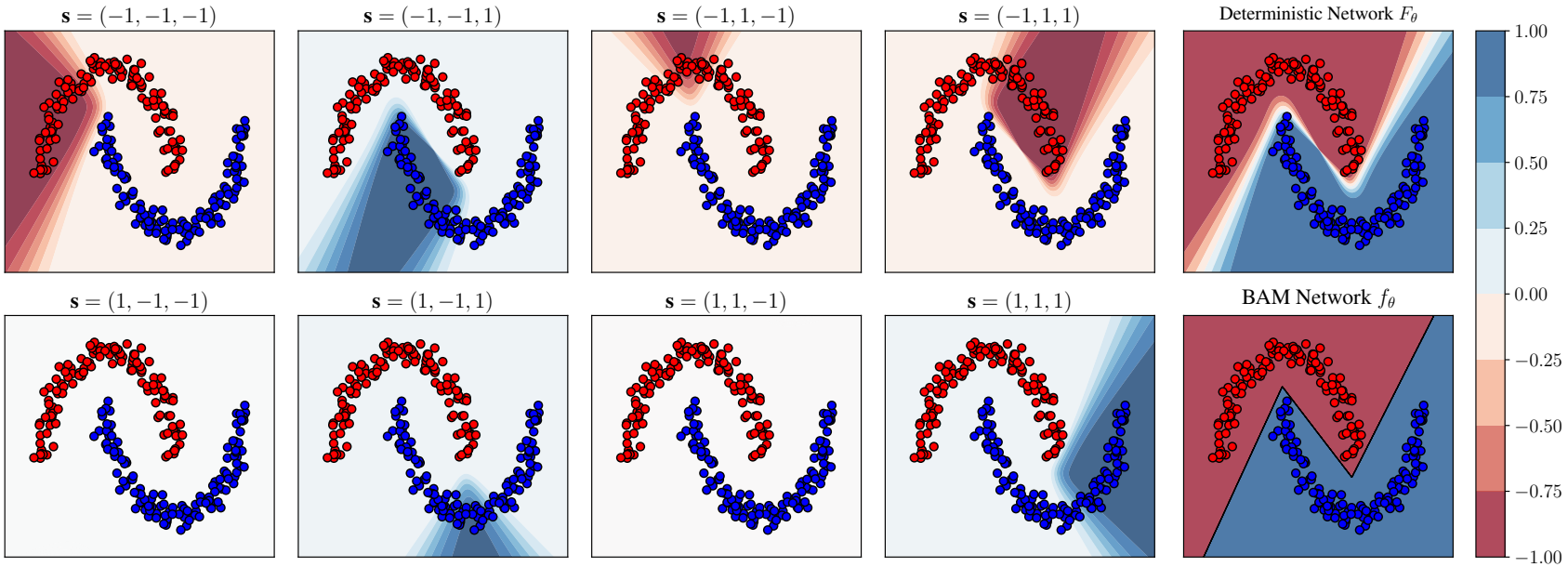
Visualisation



$$F_\theta(\mathbf{x}) = \sum_{\mathbf{s}\in\{-1,1\}^{d_1}} \underbrace{\operatorname{erf}\left(\frac{\mathbf{w}_2 \cdot \mathbf{s}}{\sqrt{2d_1}}\right)}_{F_{\mathbf{w}_2}(\mathbf{s})} \underbrace{\prod_{i=1}^{d_1}\left[\frac{1}{2} + \frac{s_i}{2}\operatorname{erf}\left(\frac{\mathbf{w}_1^i \cdot \mathbf{x}}{\sqrt{2}\,\|\mathbf{x}\|}\right)\right]}_{\Pr(\mathbf{s}|\mathbf{x},\mathbf{W}_1)}.$$

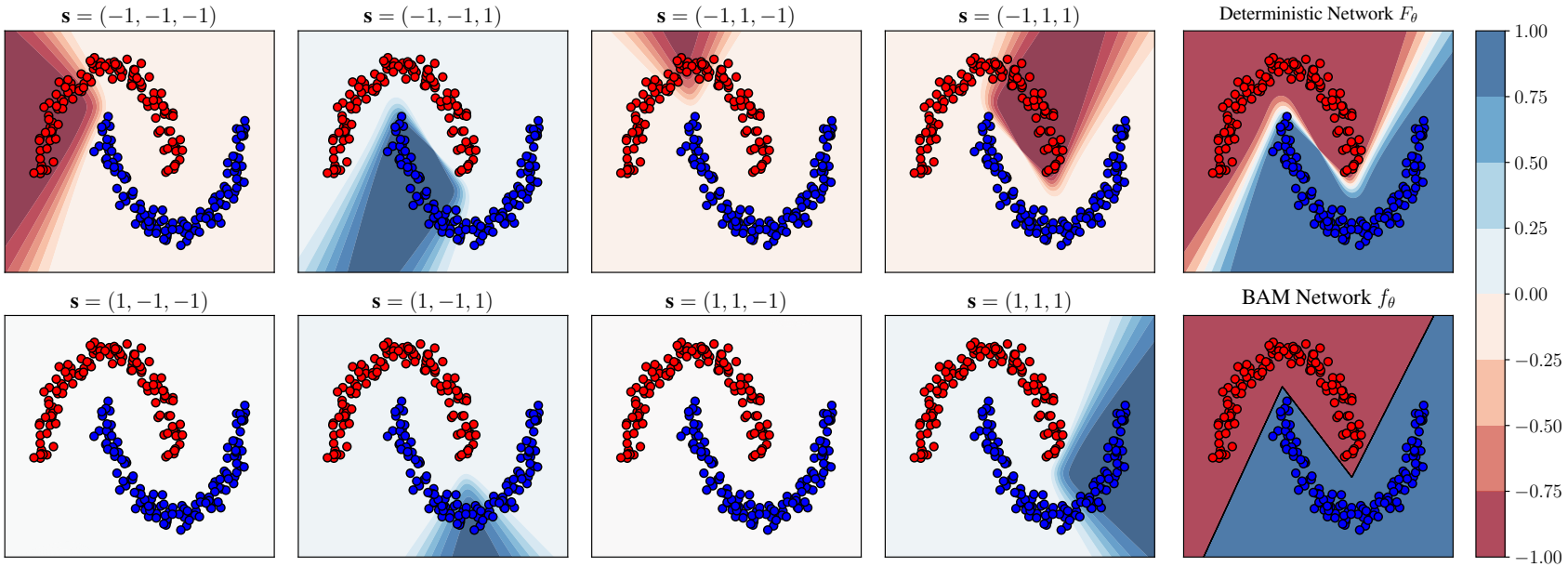


Approximation stochastique



$$F_\theta(\mathbf{x}) = \sum_{\mathbf{s} \in \{-1,1\}^{d_1}} F_{\mathbf{w}_2}(\mathbf{s}) \Pr(\mathbf{s}|\mathbf{x}, \mathbf{W}_1)$$

Approximation stochastique



$$F_\theta(\mathbf{x}) = \sum_{\mathbf{s}\in\{-1,1\}^{d_1}} F_{\mathbf{w}_2}(\mathbf{s}) \Pr(\mathbf{s}|\mathbf{x}, \mathbf{W}_1)$$

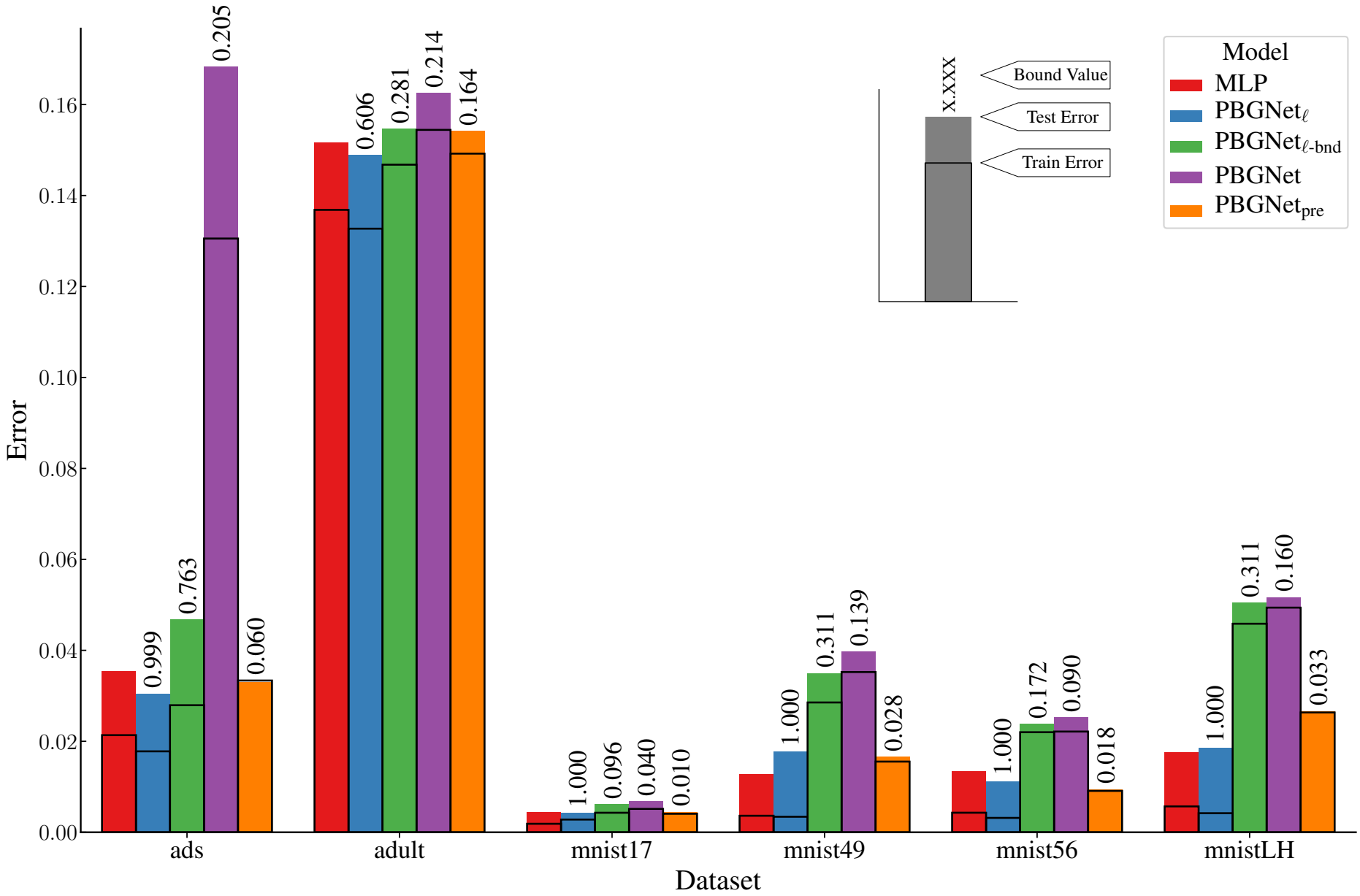
Échantillonnage de Monte Carlo

On génère T vecteurs binaires aléatoires $\{\mathbf{s}^t\}_{t=1}^T$ selon $\Pr(\mathbf{s}|\mathbf{x}, \mathbf{W}_1)$

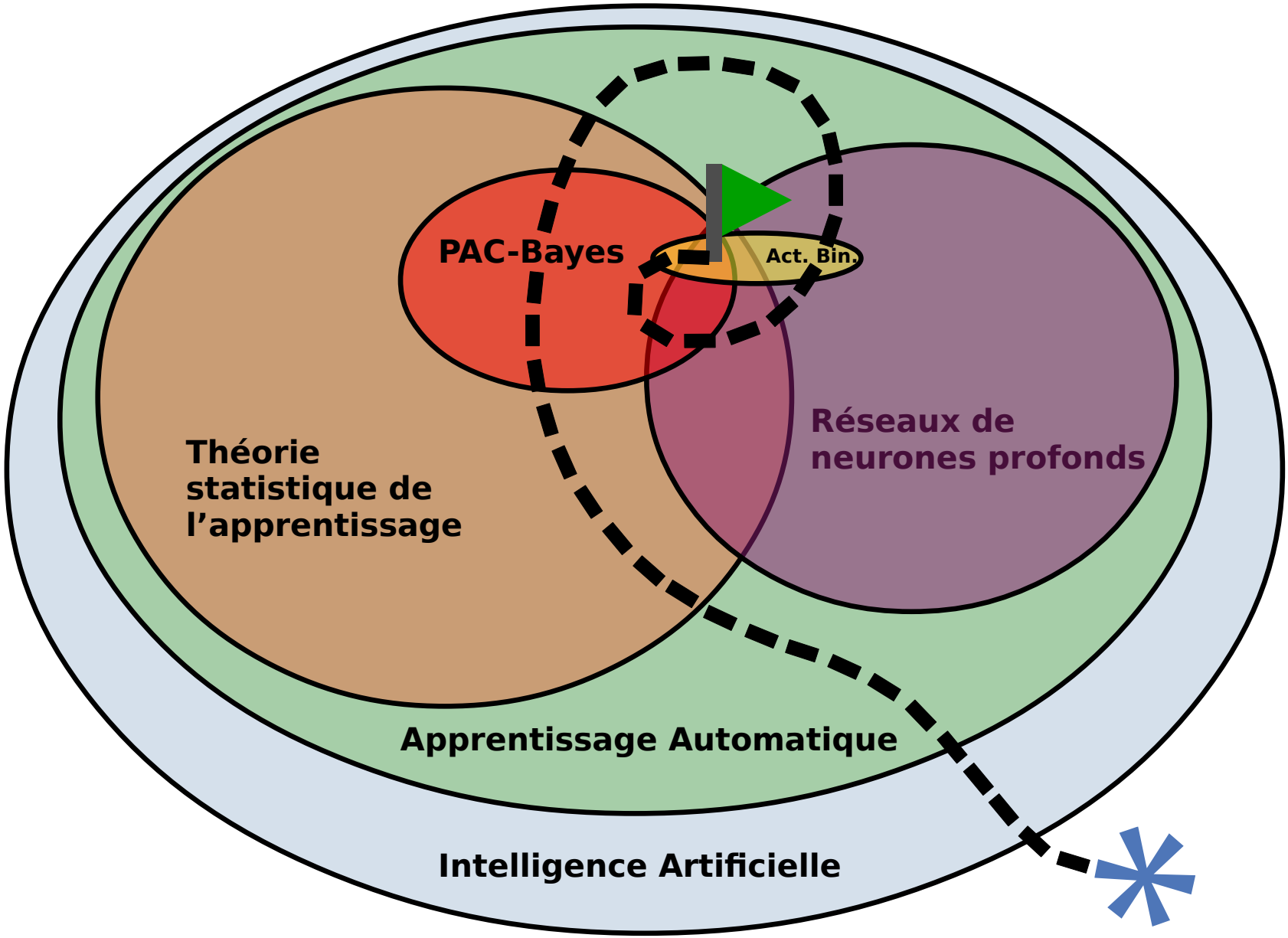
Prédiction.

$$F_\theta(\mathbf{x}) \approx \frac{1}{T} \sum_{t=1}^{T} F_{\mathbf{w}_2}(\mathbf{s}^t) .$$

Expérimentation: Résultats



Carte


PAC-Bayes
Act. Bin.
Réseaux de
neurones profonds
Théorie
statistique de
l'apprentissage
Apprentissage Automatique
Intelligence Artificielle

Perspectives

Étendre la méthode:

- Multiclasse
- Réseaux à convolutions
- Apprentissage par transfert

Analyser les propriétés:

- Robustesse
- Interprétabilité

Perspectives

Étendre la méthode:

- Multiclasse
- Réseaux à convolutions
- Apprentissage par transfert

Analyser les propriétés:

- Robustesse
- Interprétabilité

Opportunités de stages!