Avancées en théorie PAC-Bayésienne

de bornes en généralisation à des algorithmes d'apprentissage supervisé et de transfert

Soutenance HDR — 7 avril 2025

Emilie MORVANT

Université Jean Monnet - Laboratoire Hubert Curien

Rapporteurs	**Marianne CLAUSEL**	Professeure	Université de Lorraine
	Colin DE LA HIGUERA	Professeur	Université de Nantes
	Liva RALAIVOLA	VP Recherche	Criteo AI Lab
		Professeur	Aix-Marseille Université
Examinateurs	**Stéphane CHRÉTIEN**	Professeur	Université Lyon 2
	Rémi EMONET	MCF HDR	Université Jean Monnet
Tuteur	**François JACQUENET**	Professeur	Université Jean Monnet

Laboratoire Hubert Curien
UMR • CNRS • 5516 • Saint-Étienne

Université Jean Monnet Saint-Étienne

cnrs

INSTITUT d'OPTIQUE GRADUATE SCHOOL
ParisTech

Parcours


Doctorat
Post-Doc
Maître de Conférences
PEDR
RIPEC
amU
ISTA
Université Jean Monnet Saint-Étienne
Laboratoire Hubert Curien
UMR • CNRS • 5516 • Saint-Étienne
Faculté des Sciences et Techniques
Saint-Étienne
TEMPS PARTIEL
2010 ..... 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025

Parcours

Responsable Licence 2 d'informatique

Responsable L1 d'info

Monitorat

Décharge jeune chercheur

Entre 250h et 300h par an

M1 Méthodologie de la Recherche
M1 Analyse de données
M2 Intro à l'IA

L1-L2 Programmation
L2 Systèmes d'exploitation

co-encadrements de stagiaires

Parcours

Webmaster équipe

Laboratoire Hubert Curien — Conseil du laboratoire - SST (20/24) - Aide à la communication du labo (17/22)

Membre de 15 comités de sélection (MCF)

Participation des comités de programmes, d'organisation, etc

SSFAM

CNU 27

MALIA

Responsable Licence 2 d'informatique

Responsable L1 d'info

Monitorat

Décharge jeune chercheur

Entre 250h et 300h par an

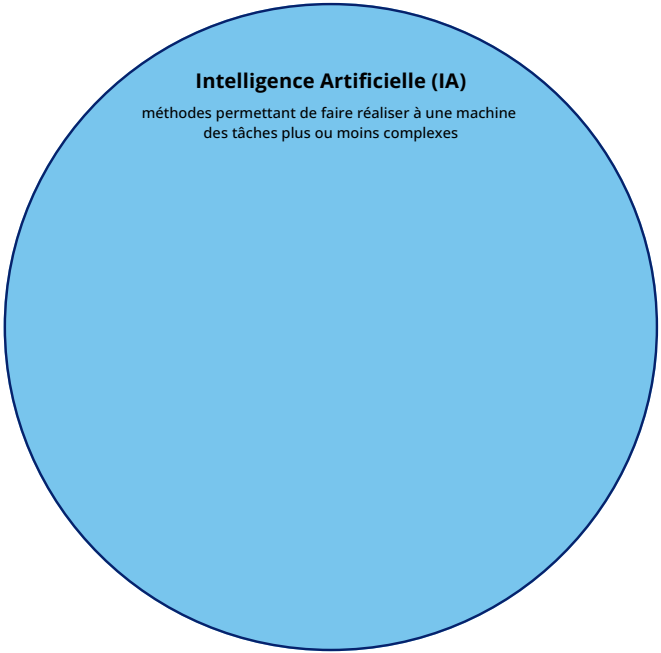
- M1 Méthodologie de la Recherche
- M1 Analyse de données
- M2 Intro à l'IA

L1-L2 Programmation

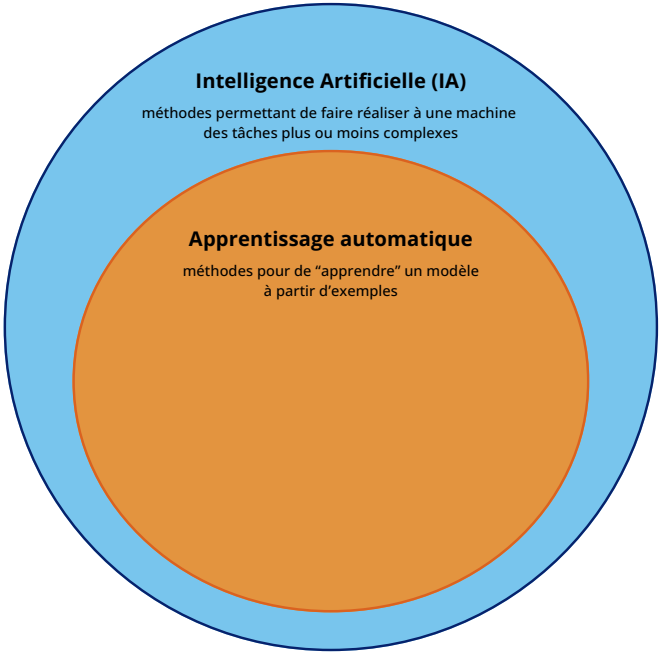
L2 Systèmes d'exploitation

co-encadrements de stagiaires

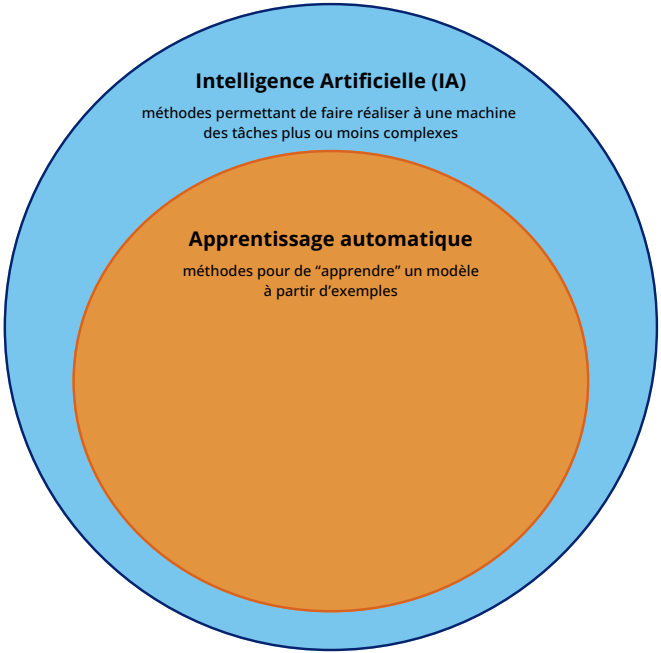
Ma trajectoire de recherche



Ma trajectoire de recherche

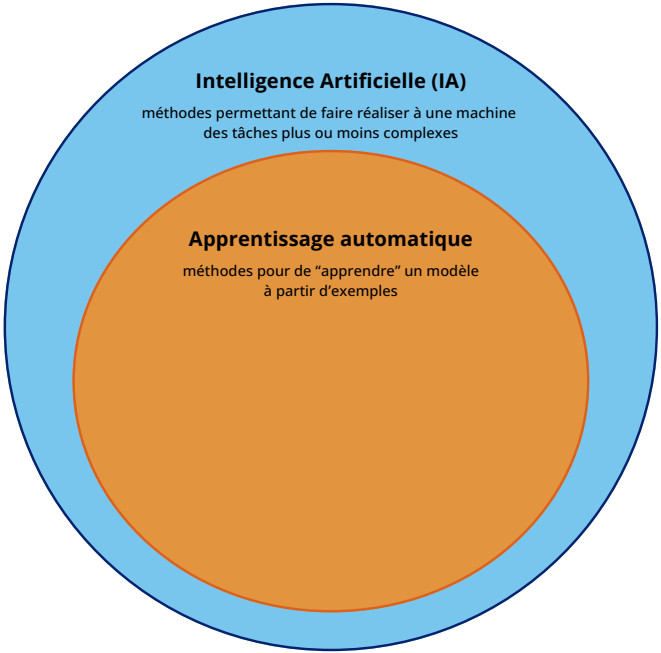


Ma trajectoire de recherche



ENJEUX SOCIÉTAUX EN IA

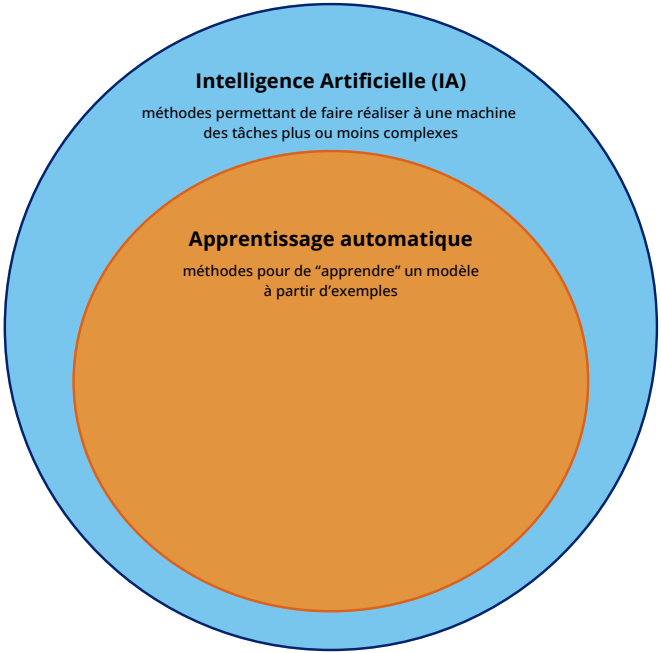
Ma trajectoire de recherche



ENJEUX SOCIÉTAUX EN IA

- **Confiance**

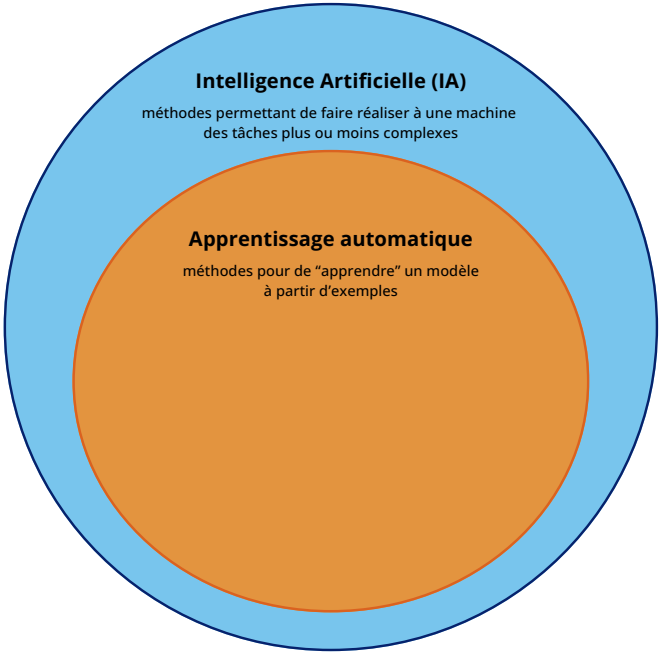
Ma trajectoire de recherche



ENJEUX SOCIÉTAUX EN IA

- **Confiance**
- **Explicabilité**

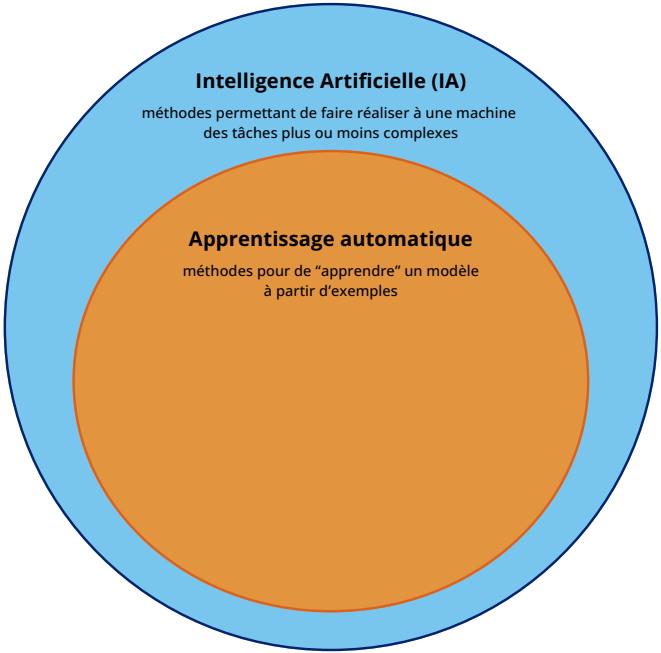
Ma trajectoire de recherche



ENJEUX SOCIÉTAUX EN IA

- **Confiance**
- **Explicabilité**
- **Adaptabilité**

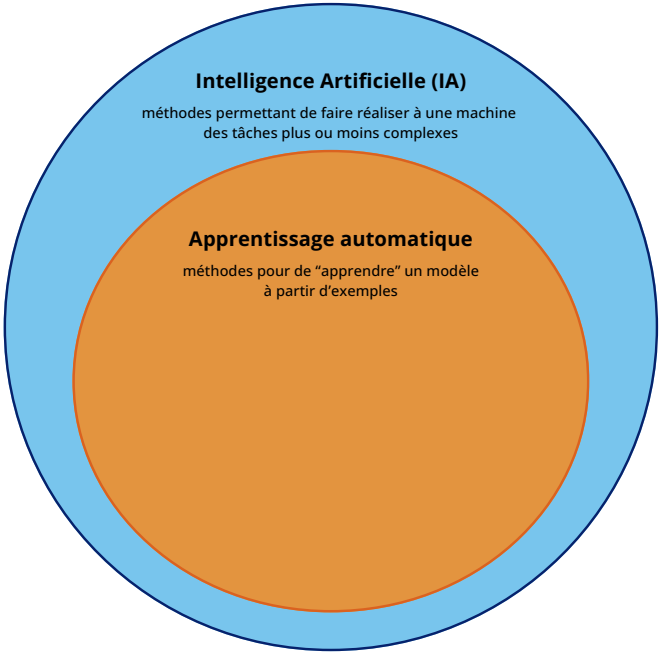
Ma trajectoire de recherche



ENJEUX SOCIÉTAUX EN IA

- **Confiance**
- **Explicabilité**
- **Adaptabilité**
- **Robustesse**

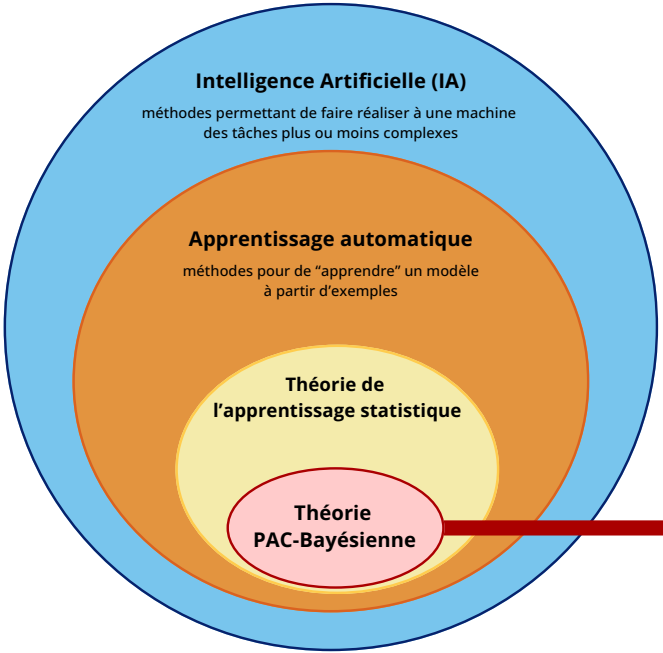
Ma trajectoire de recherche



ENJEUX SOCIÉTAUX EN IA

- **Confiance**
- **Explicabilité**
- **Adaptabilité**
- **Robustesse**
- **Équité**

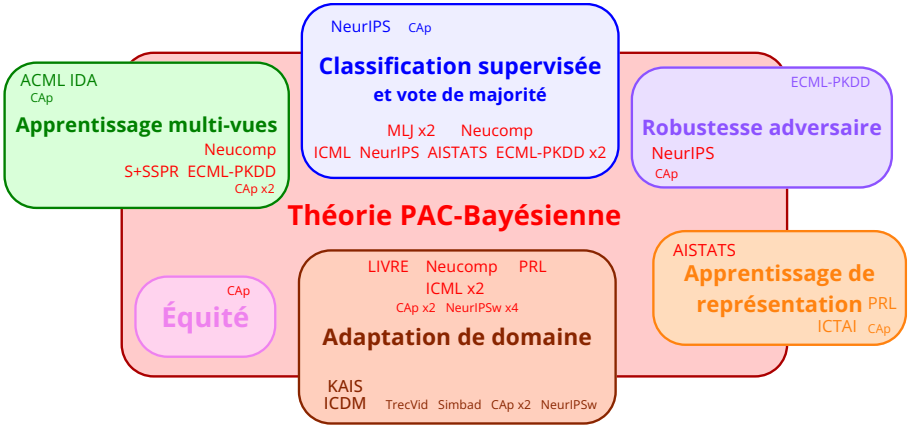
Ma trajectoire de recherche



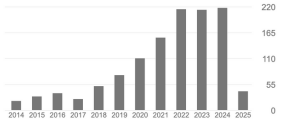
ENJEUX SOCIÉTAUX EN IA

- **Confiance**
- **Explicabilité**
- **Adaptabilité**
- **Robustesse**
- **Équité**

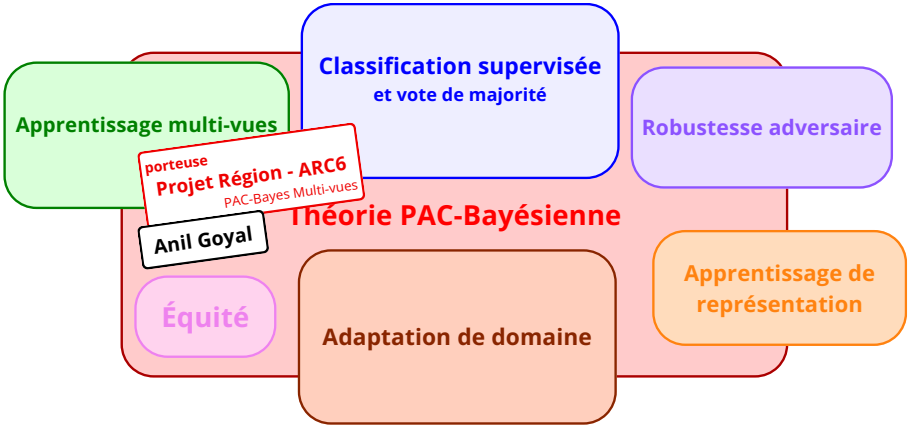
Thèmes principaux et publications

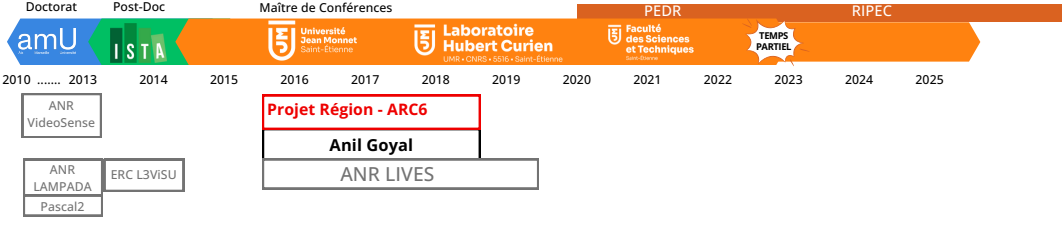


	Toutes	Depuis 2020
Citations	1213	952
indice h	16	13
indice i10	20	15

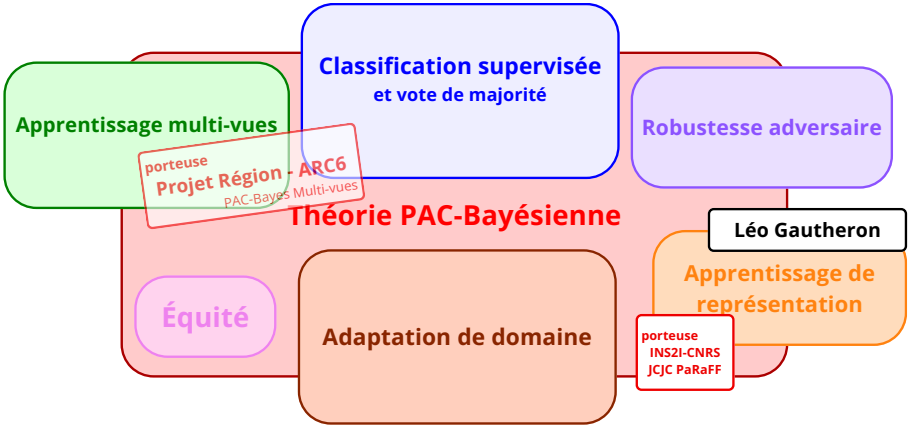


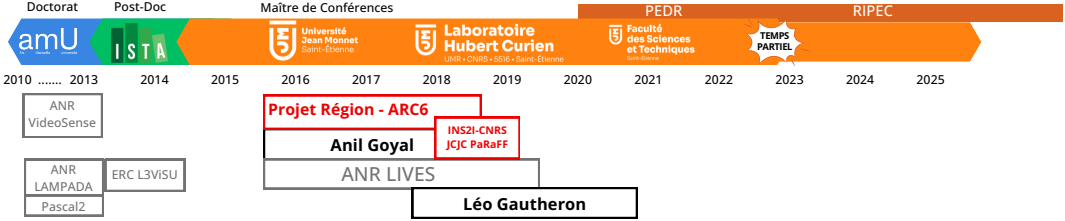
Thèmes principaux et projets



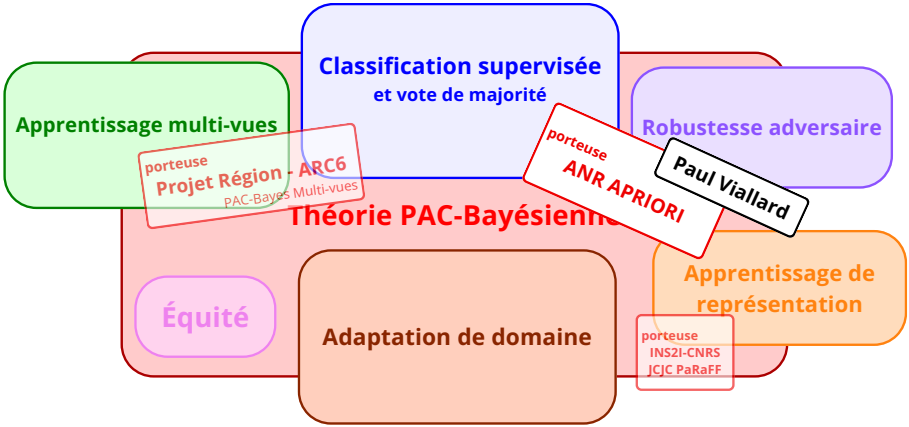


Thèmes principaux et projets

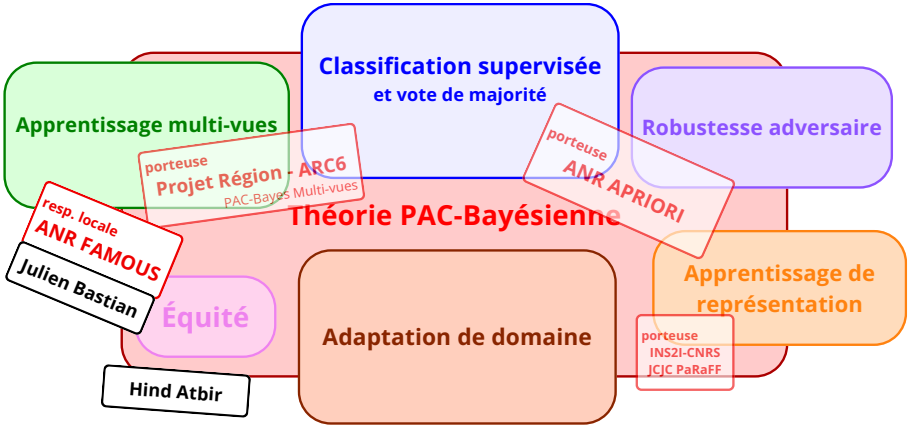


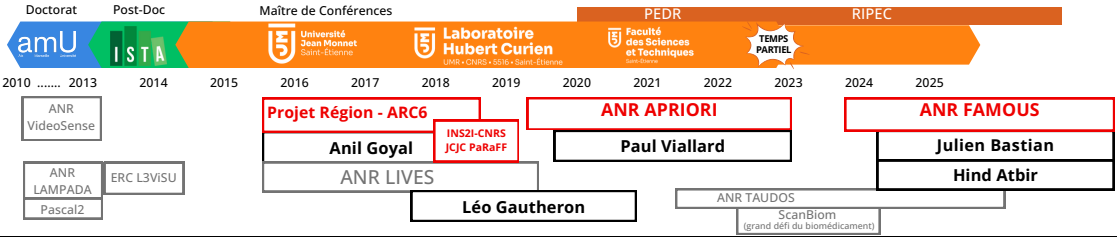


Thèmes principaux et projets

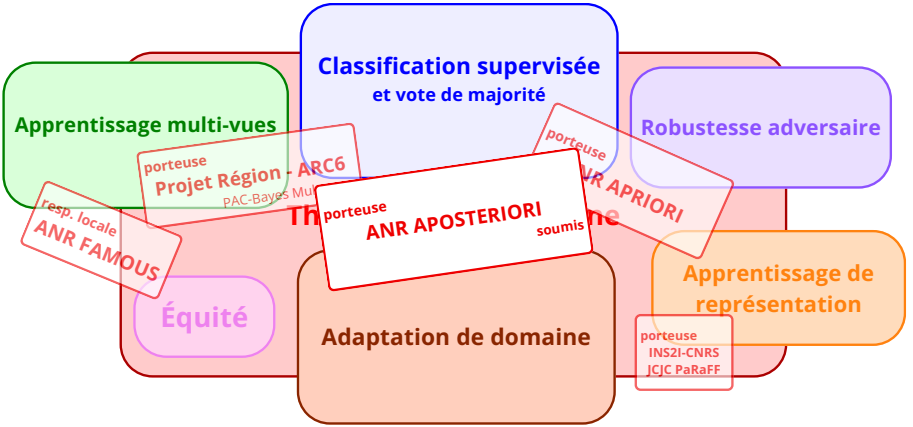


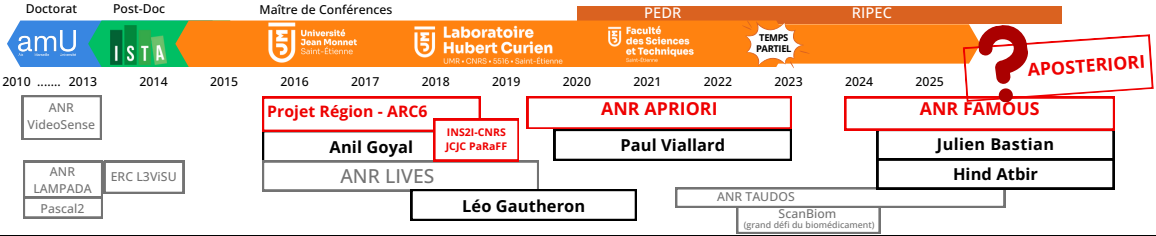
Thèmes principaux et projets





Thèmes principaux et projets



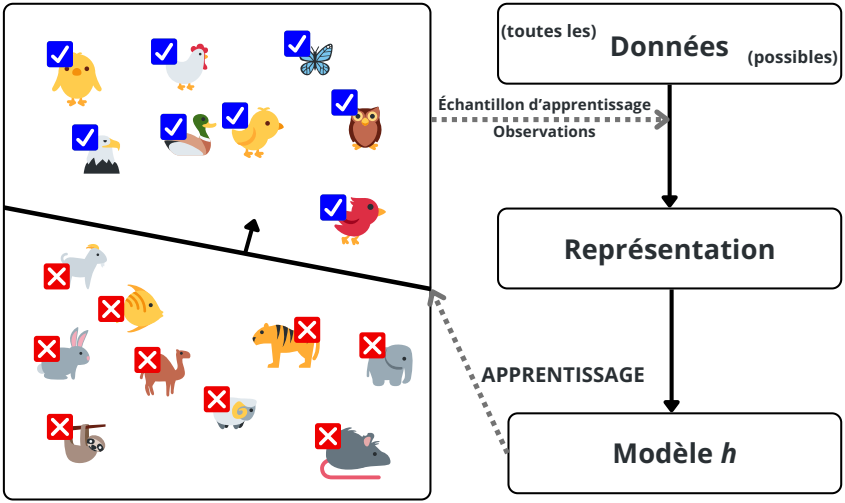


CONTEXTE SCIENTIFIQUE GÉNÉRAL

Classification supervisée

Tâche de **classification supervisée**

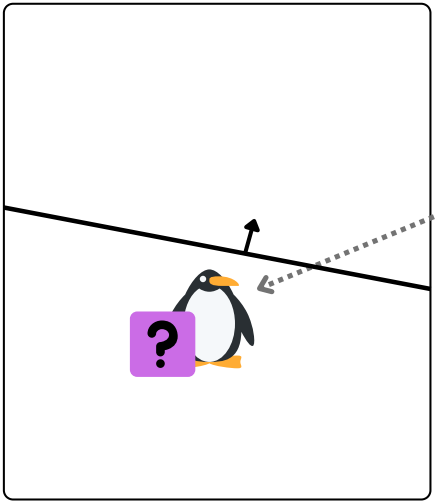
Classer les animaux en 2 catégories : ☑ avec des ailes ou ☒ sans ailes

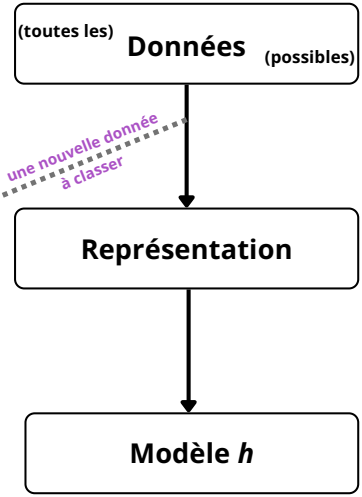


Classification supervisée

Tâche de **classification supervisée**

Classer les animaux en 2 catégories : ☑ avec des ailes ou ☒ sans ailes

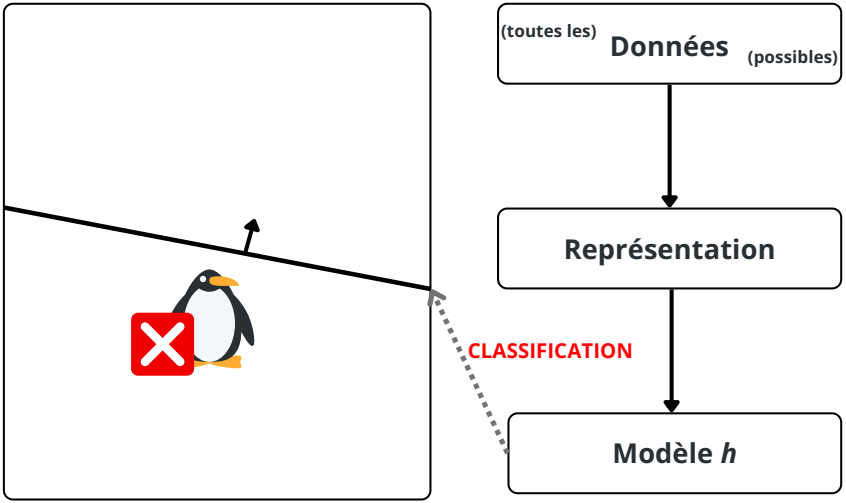




Classification supervisée

Tâche de **classification supervisée**

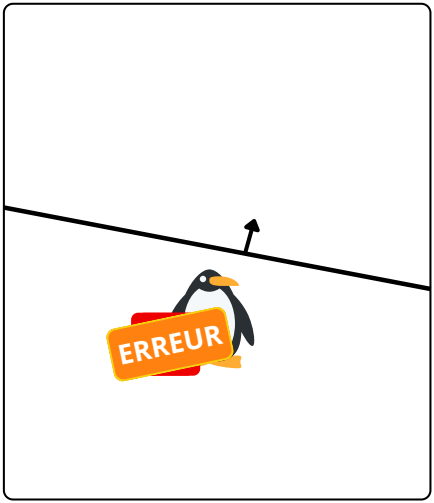
Classer les animaux en 2 catégories : ✅ avec des ailes ou ❌ sans ailes

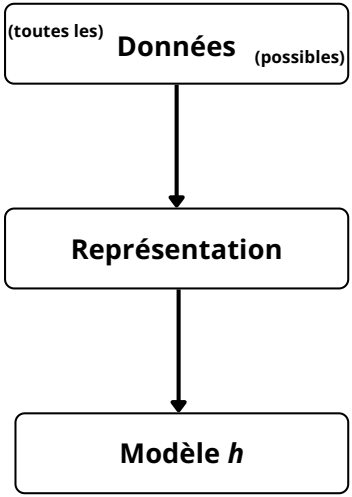


Classification supervisée

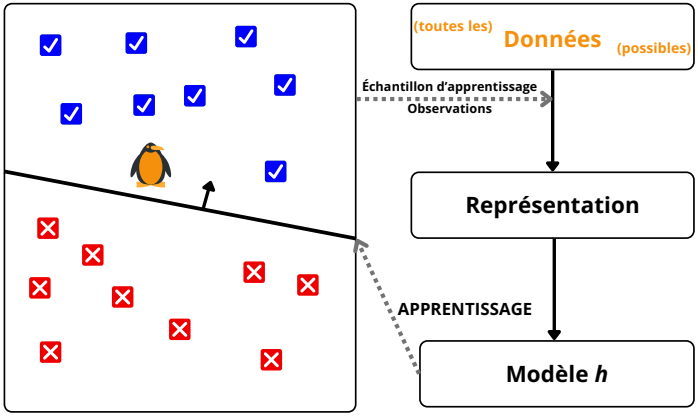
Tâche de **classification supervisée**

Classer les animaux en 2 catégories : ☑ avec des ailes ou ☒ sans ailes



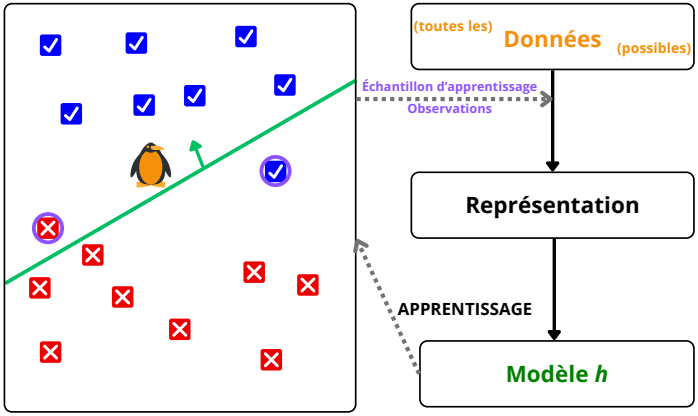


Bornes en généralisation



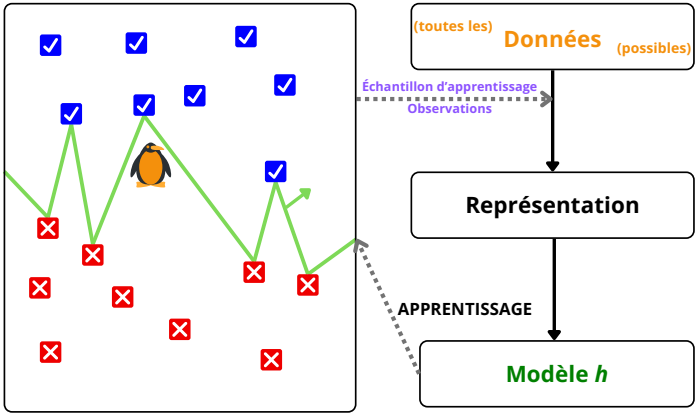
$$\underbrace{\text{erreur sur toutes les données}}_{\text{NON CALCULABLE}} \leq$$

Bornes en généralisation



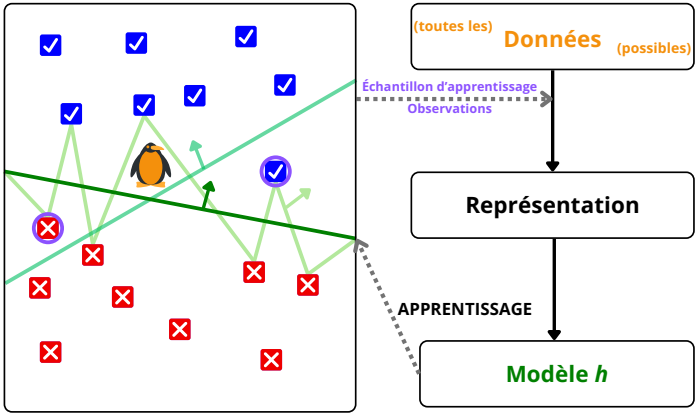
$$\underbrace{\text{erreur sur toutes les données}}_{\text{NON CALCULABLE}} \leq \underbrace{\text{erreur sur les données observées} + f\left(\text{nombre de données observées}\right)}_{\text{CALCULABLE (ou estimable)}}$$

Bornes en généralisation



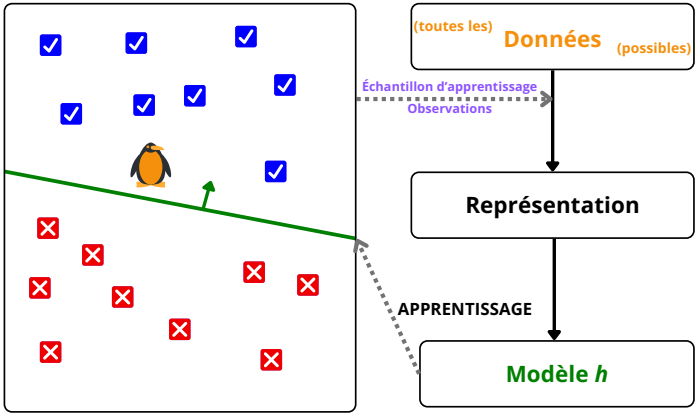
$$\underbrace{\text{erreur sur toutes les données}}_{\text{NON CALCULABLE}} \leq \underbrace{+ f\left(\text{complexité de } h, \right)}_{\text{CALCULABLE (ou estimable)}}$$

Bornes en généralisation



$$\underbrace{\text{erreur sur toutes les données}}_{\text{NON CALCULABLE}} \leq \underbrace{\text{erreur sur les données observées} + f\left(\text{complexité de } h, \text{nombre de données observées}\right)}_{\text{CALCULABLE (ou estimable)}}$$

Bornes en généralisation



$$\text{erreur sur toutes les données} \leq \text{erreur sur les données observées} + f\left(\text{complexité de } h, \frac{1}{\text{nombre de données observées}}\right)$$

Pour apprendre un modèle avec de bonnes garanties en généralisation, on peut minimiser le compromis entre erreur sur les données observées et complexité du modèle

Formalisation pour la classification supervisée

$\mathbb{X} \in \mathbb{R}^d$ espace d'entrée

$\mathbb{Y} = \{-1, +1\}$ espace de sortie

$\mathbb{H}$ ensemble de modèles tel que $\forall h \in \mathbb{H},\ h : \mathbb{X} \to \mathbb{Y}$

$\mathcal{D}$ distribution sur $\mathbb{X} \times \mathbb{Y}$

$\mathbb{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \overset{iid}{\sim} (\mathcal{D})^m$ ensemble d'apprentissage

Risque empirique associé : $\widehat{\mathrm{R}}_{\mathbb{S}}(h) = \dfrac{1}{m} \sum_{i=1}^{m} \mathrm{I}[\, h(\mathbf{x}) \neq y \,]$

Classification supervisée

Objectif : Trouver **le modèle** h de $\mathbb{H}$ qui minimise l'erreur sur toutes les données $\mathrm{R}_{\mathcal{D}}(h)$

$$\underbrace{\mathrm{R}_{\mathcal{D}}(h) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathrm{I}[\, h(\mathbf{x}) \neq y \,]}_{\text{Risque réel}}$$

Formes des bornes en généralisation PAC (Valiant, 1984)

Borne en généralisation dîtes Probably Approximately Correct

Étant donné une distribution $\mathcal{D}$ sur $\mathbb{X}\times\mathbb{Y}$, pour tout $\delta \in (0,1]$, on a

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\underbrace{R_{\mathcal{D}}(h)}_{\text{Risque réel}} \leq \underbrace{\widehat{R}_{\mathbb{S}}(h)}_{\text{Risque empirique}} + \underbrace{\Phi(h,\mathbb{S},\delta)}_{\text{Complexité}} \right] \geq 1-\delta$$

Avec une grande probabilité d'au moins $1-\delta$, le risque de h est plus petit que $\widehat{R}_{\mathbb{S}}(h)+\Phi(h,\mathbb{S},\delta)$

On veut que les bornes soient

- **Informatives** $\iff \widehat{R}_{\mathbb{S}}(h) + \Phi(h,\mathbb{S},\delta) \leq 1$
- **Précises** $\iff \left|R_{\mathcal{D}}(h) - \widehat{R}_{\mathbb{S}}(h)\right| \simeq \Phi(h,\mathbb{S},\delta)$
- **Fiables** $\iff$ un petit δ augmente la fiabilité de la borne, mais diminue sa précision
- **Calculables *(facilement)*** $\iff \widehat{R}_{\mathbb{S}}(h)+\Phi(h,\mathbb{S},\delta)$ est calculable/majorable avec $\mathbb{S}$

Formes des bornes en généralisation PAC (Valiant, 1984)

Borne en généralisation dîtes **Probably Approximately Correct**

Étant donné une distribution $\mathcal{D}$ sur $\mathbb{X}\times\mathbb{Y}$, pour tout $\delta \in (0,1]$, on a

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\underbrace{\mathrm{R}_{\mathcal{D}}(h)}_{\text{Risque réel}} \leq \underbrace{\widehat{\mathrm{R}}_{\mathbb{S}}(h)}_{\text{Risque empirique}} + \underbrace{\Phi(h,\mathbb{S},\delta)}_{\text{Complexité}} \right] \geq 1-\delta$$

Avec une grande probabilité d'au moins $1-\delta$, le risque de h est plus petit que $\widehat{\mathrm{R}}_{\mathbb{S}}(h)+\Phi(h,\mathbb{S},\delta)$

Trois grandes familles de bornes en généralisation PAC :

- Bornes en convergence uniforme (Vapnik et al., 1971 ; Bartlett et al., 2002)
- Bornes dépendantes d'un algorithme (Bousquet et al., 2002 ; Xu et al., 2012)
- Bornes PAC-Bayésiennes (Shawe-Taylor et al., 1997 ; McAllester, 1998)

Les trois familles de bornes

Borne en convergence uniforme pour un ensemble de modèles $\mathbb{H}$

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\Big[\ \forall h \in \mathbb{H},\ \ \mathrm{R}_{\mathcal{D}}(h) \ \leq\ \widehat{\mathrm{R}}_{\mathbb{S}}(h) \ +\ \Phi(\mathbb{H},\mathbb{S},\delta)\ \Big] \geq 1-\delta$$

Borne en pire cas — **Fixe pour tout $h \in \mathbb{H}$**

- Valable pour tous les modèles de $\mathbb{H}$

↪ Bornes en pire cas ⇝ souvent trop pessimistes

- $\Phi(h,\mathbb{S},\delta)$ fixe pour tous les modèles de $\mathbb{H}$, par exemple
  - Dimension de Vapnik-Chervonenkis (VC-dim) de $\mathbb{H}$
    - ✓ Peut être simple à interpréter (capacité d'apprentissage de $\mathbb{H}$)
    - ✗ Peut être très grande voir infinie ⇝ Borne non informative
  - Complexité Rademacher de $\mathbb{H}$
    - ✓ Plus précise que la VC-dim (prise en compte de $\mathcal{D}$)
    - ✗ Peut être difficile à calculer en pratique

Les trois familles de bornes

Borne dépendante d'un algorithme $\mathcal{A} : \mathbb{S} \mapsto h_{\mathbb{S}}$

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\mathrm{R}_{\mathcal{D}}(h_{\mathbb{S}}) \leq \widehat{\mathrm{R}}_{\mathbb{S}}(h_{\mathbb{S}}) + \Phi(\mathcal{A}, \mathbb{S}, \delta) \right] \geq 1 - \delta$$

Valide uniquement pour $h_{\mathbb{S}}$ appris avec $\mathcal{A}$ et $\mathbb{S}$

dépend de caractéristiques de l'algorithme $\mathcal{A}$

- Valable pour un seul modèle $h_{\mathbb{S}} \rightarrow$ celui appris $\mathcal{A}$
  - $\hookrightarrow$ Plus réaliste mais moins générale
- $\Phi(\mathcal{A}, \mathbb{S}, \delta)$ dépend de propriétés de $\mathcal{A}$, par exemple
  - Stabilité algorithmique
    - ✓ Capture la résistance à de faibles variations dans les données
    - ✗ Peut-être difficile à estimer
  - Robustesse algorithmique
    - ✓ Capture la résistance aux variations dans une même zone de l'espace
    - ✗ Peut être difficile à calculer en pratique

Les trois familles de bornes

Borne PAC-Bayésiennes

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\ \forall \rho,\ \mathop{\mathbb{E}}_{h\sim\rho} \mathsf{R}_{\mathcal{D}}(h) \ \leq\ \mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathsf{R}}_{\mathbb{S}}(h) \ +\ \Phi(\Delta(\rho,\pi),\mathbb{S},\delta)\ \right] \geq 1-\delta$$

Borne en espérance sur $\mathbb{H}$

Dépend d'une distance Δ entre ρ et un apriori π

- Borne en espérance sur $\mathbb{H}$
  $\hookrightarrow$ souvent plus précises que les bornes en pire cas
  $\hookrightarrow$ borne sur un risque stochastique sur $\mathbb{H}$
- Mais l'espérance sur $\mathbb{H}$ est étroitement reliée au vote de majorité sur $\mathbb{H}$
- Peut être facilement calculable (ou majorable)
  $\hookrightarrow$ et minimisable $\Rightarrow$ algorithme auto-certifié
- $\Phi(\Delta(\rho,\pi),\mathbb{S},\delta)$ dépend
  - du choix d'un *a priori* π sur l'espérance
  - d'une distance $\Delta(\rho,\pi)$ entre ρ et π (ex. KL-divergence)

Les trois familles de bornes

Borne en convergence uniforme pour un ensemble de modèles $\mathbb{H}$

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\ \forall h\in\mathbb{H},\ \ R_{\mathcal{D}}(h)\ \leq\ \hat{R}_{\mathbb{S}}(h)\ +\ \Phi(\mathbb{H},\mathbb{S},\delta)\ \right]\geq 1-\delta$$

Borne en pire cas — **Fixe pour tout $h\in\mathbb{H}$**

Borne dépendante d'un algorithme $\mathcal{A}:\mathbb{S}\mapsto h_{\mathbb{S}}$

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\ R_{\mathcal{D}}(h_{\mathbb{S}})\ \leq\ \hat{R}_{\mathbb{S}}(h_{\mathbb{S}})\ +\ \Phi(\mathcal{A},\mathbb{S},\delta)\ \right]\geq 1-\delta$$

Valide uniquement pour $h_{\mathbb{S}}$ appris avec $\mathcal{A}$ et $\mathbb{S}$ — **dépend de caractéristiques de l'algorithme $\mathcal{A}$**

Borne PAC-Bayésienne

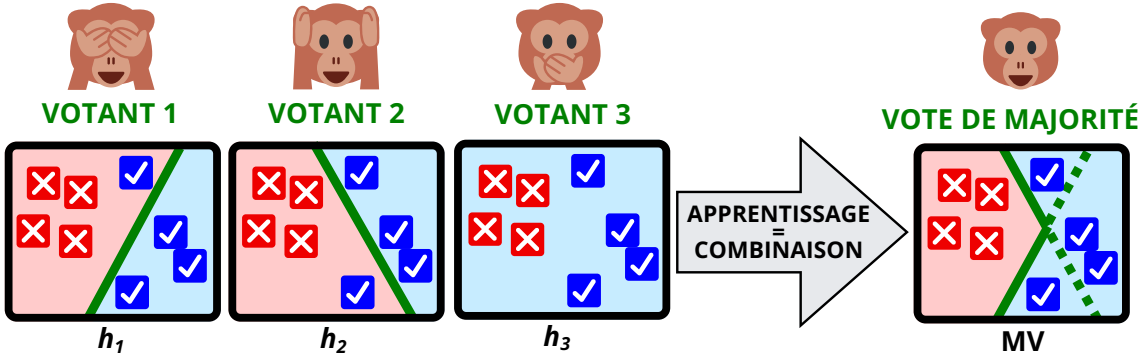
$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\ \forall\rho,\ \mathop{\mathbb{E}}_{h\sim\rho}R_{\mathcal{D}}(h)\ \leq\ \mathop{\mathbb{E}}_{h\sim\rho}\hat{R}_{\mathbb{S}}(h)\ +\ \Phi(\Delta(\rho,\pi),\mathbb{S},\delta)\ \right]\geq 1-\delta$$

Borne en espérance sur $\mathbb{H}$ — **Dépend d'une distance Δ entre ρ et un apriori π**

LA THÉORIE PAC-BAYÉSIENNE

Un peu plus en détails

La théorie PAC-Bayésienne — Le vote de majorité



$\mathbb{H}$: Ensemble de modèles simples (*i.e.*, de **votants**), ici 3 votants $\{h_1, h_2, h_3\}$

$\Longrightarrow$ Apprendre une combinaison pondérée des votants $\Longleftrightarrow$ Un vote de majorité MV

$$\mathrm{MV}(\cdot) = \mathrm{sign}\left[\sum_{h \in \mathbb{H}} \overbrace{\rho(h)}^{\text{poids de } h} \; h(\cdot) \right]$$

Classification supervisée PAC-Bayésienne

$\mathbb{X} \in \mathbb{R}^d$ espace d'entrée

$\mathbb{Y} = \{-1, +1\}$ espace de sortie

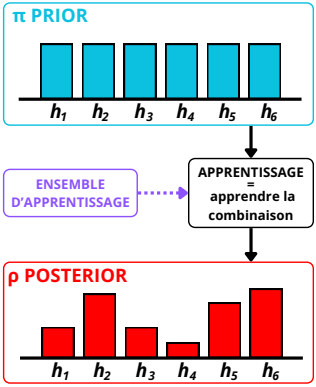
$\mathbb{H}$ ensemble de votants tel que $\forall h \in \mathbb{H},\ h : \mathbb{X} \to \mathbb{Y}$

$\mathcal{D}$ distribution sur $\mathbb{X} \times \mathbb{Y}$

$\mathbb{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \overset{iid}{\sim} \mathcal{D}^m$ ensemble d'apprentissage

π **distribution prior sur** $\mathbb{H}$

ρ **distribution posterior sur** $\mathbb{H}$



Classification supervisée **PAC-Bayésienne**

Objectif : Trouver le vote de majorité pondéré MV sur $\mathbb{H}$ qui minimise $\mathrm{R}_{\mathcal{D}}(\mathrm{MV})$

$$\underbrace{\mathrm{R}_{\mathcal{D}}(\mathrm{MV}) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathrm{I}[\,\mathrm{MV}(\mathbf{x}) \neq y\,]}_{\text{Risque réel}} \qquad \text{où} \quad \mathrm{MV}(\mathbf{x}) = \mathrm{sign}\left[\mathop{\mathbb{E}}_{h\sim\rho} h(\mathbf{x})\right]$$

Un point clé en PAC-Bayes : Le risque de Gibbs

Risque de Gibbs = espérance des risques individuels

$$\mathrm{R}_{\mathcal{D}}(\mathrm{MV}) \;\leq\; 2 \times \mathop{\mathbb{E}}_{h \sim \rho} \mathrm{R}_{\mathcal{D}}(h)$$

Avantages :

- Les garanties en **garanties en espérances** sur l'ensemble des votants $\mathbb{H}$
  $\Longrightarrow$ permettent d'avoir des garanties sur un modèle déterministe
- Permet de faire intervenir la **diversité** des votants

Important :

- D'autres bornes supérieures de $\mathrm{R}_{\mathcal{D}}(\mathrm{MV})$ existent

Les ingrédients d'une borne PAC-Bayésienne

Rappel : Forme d'une borne en généralisation PAC

$$\text{risque réel} \leq \text{risque empirique} + \; f\left(\text{complexité}(h)\;,\; \tfrac{1}{m}\right)$$

- Risque de Gibbs réel sur la distribution $\mathcal{D}$: $\mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h)$
- Risque empirique sur m exemples $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \sim (\mathcal{D})^m$:

$$\mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathrm{R}}_{\mathcal{S}}(h) = \frac{1}{m}\sum_{i=1}^{m} \mathop{\mathbb{E}}_{h\sim\rho} \mathrm{I}[\, h(\mathbf{x}_i) \neq y_i \,] \quad \text{(risque de Gibbs empirique)}$$

- Complexité (divergence de Kullback-Leiber) : $\mathrm{KL}(\rho\|\pi) = \mathop{\mathbb{E}}_{h\sim\rho} \ln \frac{\rho(h)}{\pi(h)}$
- Une mesure de déviation convexe entre les risques réel et empirique $D : [0,1]\times[0,1] \to \mathbb{R}$

Borne PAC-Bayésienne générale

avec $D(a,b) = (a-b)^2$ (borne de la forme de McAllester)

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\text{Pour toute distribution posterior } \rho \text{ sur } \mathbb{H},\ \underbrace{\mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h)}_{\text{risque réel}} \leq \underbrace{\mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathrm{R}}_{\mathbb{S}}(h)}_{\text{risque empirique}} + \sqrt{\frac{1}{2m}\left(\underbrace{\mathrm{KL}(\rho\|\pi)}_{\text{complexité}} + \ln\frac{2\sqrt{m}}{\delta}\right)}\right] \geq 1-\delta$$

Borne PAC-Bayésienne générale

avec $D(a,b) = (a-b)^2$ (borne de la forme de McAllester)

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\text{Pour toute distribution posterior } \rho \text{ sur } \mathbb{H},\ \underbrace{\mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h)}_{\text{risque réel}} \leq \underbrace{\mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathrm{R}}_{\mathbb{S}}(h)}_{\text{risque empirique}} + \sqrt{\frac{1}{2m}\left(\underbrace{\mathrm{KL}(\rho\|\pi)}_{\text{complexité}} + \ln\frac{2\sqrt{m}}{\delta}\right)}\right] \geq 1-\delta$$

Théorème PAC-Bayésien général

Pour toute distribution $\mathcal{D}$ on $\mathbb{X}\times\mathbb{Y}$, pour tout ensemble de votants $\mathbb{H}$, pour toute distribution prior π sur $\mathbb{H}$, pour tout $\delta \in (0,1]$, pour toute fonction convexe $D : [0,1]\times[0,1] \to \mathbb{R}$, on a

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\text{Pour toute distribution posterior } \rho \text{ sur } \mathbb{H},\ D\left(\mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h), \mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathrm{R}}_{\mathbb{S}}(h)\right) \leq \frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln\left(\frac{1}{\delta}\mathop{\mathbb{E}}_{\mathbb{S}\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim\pi} \mathrm{e}^{mD\left(\mathrm{R}_{\mathcal{D}}(h), \widehat{\mathrm{R}}_{\mathbb{S}}(h)\right)}\right)\right]\right] \geq 1-\delta$$

SUR LA FLEXIBILITÉ DU PAC-BAYES

Principe simplifié de mes contributions

- Dérivation de **nouvelles bornes** en généralisation PAC-Bayésiennes pour différents cadres
- Dérivation d'algorithmes d'**apprentissage supervisé**
  - Soit en s'inspirant des bornes
  - Soit en minimisant directement les bornes $\rightsquigarrow$ **Algorithmes auto-certifiés**

Fil rouge : une borne PAC-Bayésienne

Pour toute distribution $\mathcal{D}$ sur $\mathbb{X}\times\mathbb{Y}$, pour tout ensemble de votants $\mathbb{H}$,
pour toute distribution prior π sur $\mathbb{H}$, pour tout $\delta \in (0,1]$, on a

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left[\forall \text{ distribution posterior } \rho \text{ sur } h,\ \underbrace{\mathop{\mathbb{E}}_{h\sim\rho} R_{\mathcal{D}}(h)}_{\text{risque réel}} \le \underbrace{\mathop{\mathbb{E}}_{h\sim\rho} \widehat{R}_S(h)}_{\text{risque empirique}} + \sqrt{\frac{1}{2m}\left(\underbrace{\mathrm{KL}(\rho\|\pi)}_{\text{complexité}} + \ln\frac{2\sqrt{m}}{\delta}\right)}\right] \ge 1-\delta$$

Adaptation de domaine d'une tâche source vers une tâche cible

L'être humain est capable de s'adapter à une nouvelle tâche à partir de connaissances acquises

Adaptation de mouvements d'animaux au Manchuria Kung Fu

TACHE FACILE

Adaptation du Manchuria Kung Fu au ski

TACHE DIFFICILE

En apprentissage automatique

Se traduit par le besoin d'une mesure de capacité d'adaptation entre les tâches

Adaptation de domaine d'une tâche source vers une tâche cible

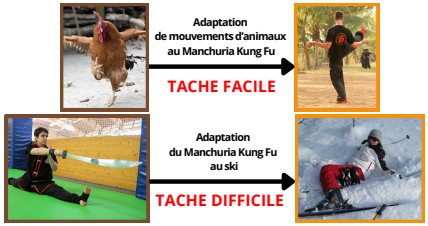
Deux distributions de données sur $\mathbb{X}\times\mathbb{Y}$

$\mathcal{S}$ domaine source $\quad\mathcal{T}$ domaine cible

Deux échantillons d'observations

$\mathbb{S}\sim\mathcal{S}$ étiqueté $\quad\mathbb{T}\sim\mathcal{T}_{\mathbb{X}}$ non étiqueté

$\mathrm{dist}(\mathcal{S},\mathcal{T})$: Distance entre domaines



La distance entre les domaines permet de relier les deux domaines

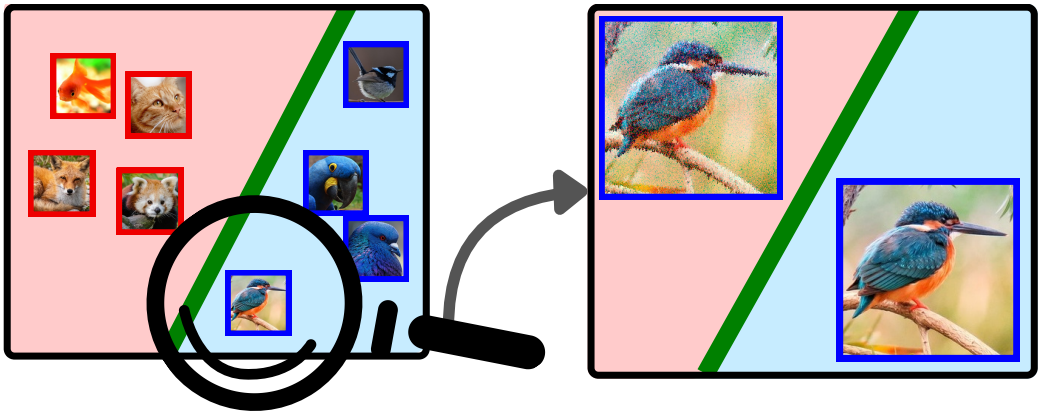
$\Longleftrightarrow$ À quel point peut-on utiliser les étiquettes sources pour minimiser le risque cible ?

$$\mathop{\mathbb{P}}_{\substack{\mathbb{S}\sim\mathcal{S}\\ \mathbb{T}\sim\mathcal{T}_{\mathbb{X}}}}\left[\forall\text{ posterior }\rho\text{ sur }\mathbb{H},\ \mathop{\mathbb{E}}_{h\sim\rho}R_{\mathcal{T}}(h)\le \underbrace{\mathop{\mathbb{E}}_{h,h'\sim\rho}\widehat{R}_{\mathbb{T}}(h,h')}_{\text{désaccord (sans étiquette)}}+\frac{1}{\mathrm{dist}(\mathcal{S},\mathcal{T})}\times\underbrace{\mathop{\mathbb{E}}_{h,h'\sim\rho}\widehat{R}_{\mathbb{S}}(h,h')}_{\text{erreur jointe (avec étiquettes)}}+\Phi(\mathrm{KL}(\rho\|\pi),(\mathbb{S}\cup\mathbb{T}),\delta)\right]\ge 1-\delta$$

$$\widehat{R}_{\mathbb{T}}(h,h')=\sum_{\mathbf{x}\in\mathbb{T}}\tfrac{1}{|\mathbb{T}|}\,\mathrm{I}[h(\mathbf{x})\neq h'(\mathbf{x})]\qquad \widehat{R}_{\mathbb{S}}(h,h')=\sum_{(\mathbf{x},y)\in\mathbb{S}}\tfrac{1}{|\mathbb{S}|}\,\mathrm{I}[h(\mathbf{x})\neq y]\,\mathrm{I}[h'(\mathbf{x})\neq y]$$

PAC-Bayes pour la robustesse adversaire

Comment garantir que le modèle sera résistant à des attaques malveillantes imperceptibles ?



On veut garantir que le modèle **se trompe le moins possible sur des données bruitées**

PAC-Bayes pour la robustesse adversaire

Espace des données $\times$ Espace des bruits

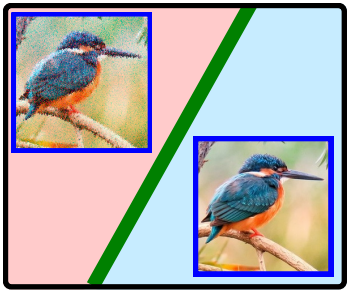
$$\underbrace{(\mathbb{X}\times\mathbb{Y}) \times \mathbb{B}}$$

$\mathcal{E}$ distribution sur $(\mathbb{X}\times\mathbb{Y})\times\mathbb{B}$

Ensemble d'apprentissage perturbé/bruité

$$\widehat{\mathbb{S}} = \left\{\left((\mathbf{x}_i, y_i), \{\epsilon_j^i\}_{j=1}^n\right)\right\}_{i=1}^m \sim (\mathcal{E}^n)^m$$

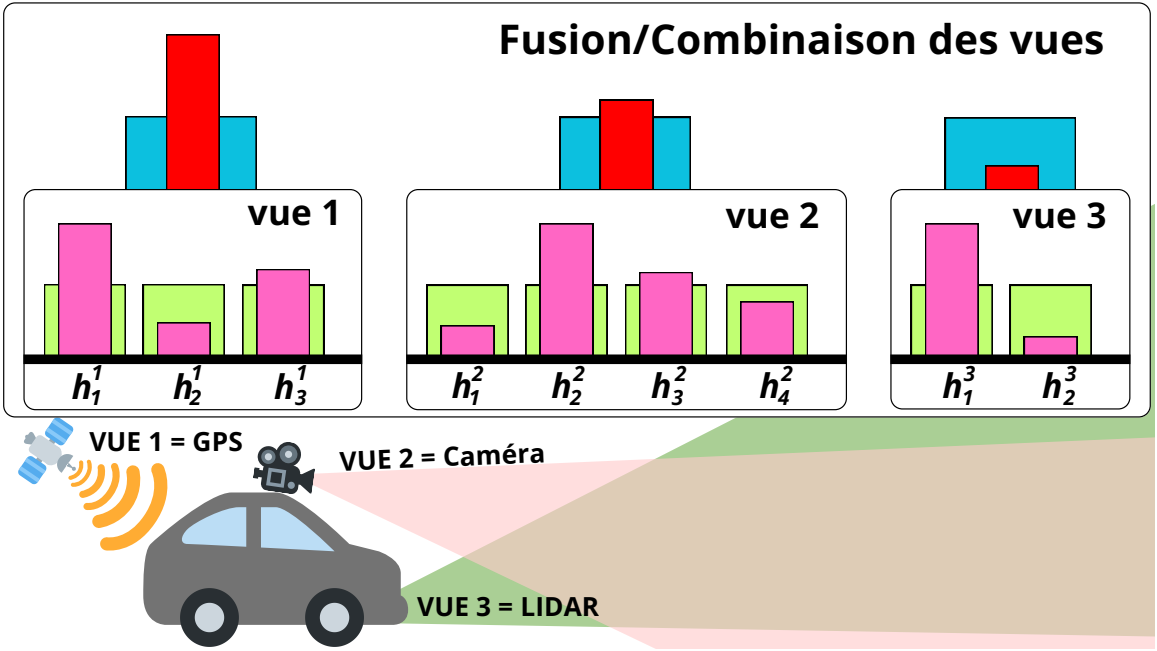
$\Leftrightarrow$ Chaque exemple $(\mathbf{x}_i, y_i)$ est perturbé avec n bruits



$$\mathop{\mathbb{P}}_{\widehat{\mathbb{S}}\sim(\mathcal{E}^n)^m}\left[\ \forall \text{ posterior } \rho \text{ sur } \mathbb{H},\ \mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{E}}(h) \ \le\ \mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathrm{R}}_{\widehat{\mathbb{S}}}(h) \ +\ \sqrt{\frac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{m+1}{\delta}\right]}\ \right] \ge 1-\delta$$

où $\mathrm{R}_{\mathcal{E}}(h) = \mathop{\mathbb{E}}_{((\mathbf{x},y),\epsilon)\sim\mathcal{E}} \mathrm{I}[h(\mathbf{x}+\epsilon)\neq y]$ et $\widehat{\mathrm{R}}_{\widehat{\mathbb{S}}}(h) = \frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n \mathrm{I}\left[h(\mathbf{x}_i+\epsilon_j^i)\neq y_i\right]$

Apprentissage multi-vues PAC-Bayésien



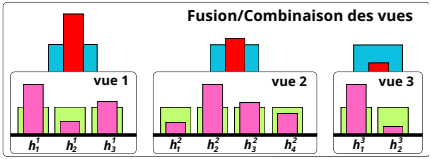
Apprentissage multi-vues PAC-Bayésien

$\mathbf{V}$ ensemble de V vues

$\mathbb{X} = \mathbb{X}^1 \times \cdots \times \mathbb{X}^V$ espace d'entrée multi-vues

$\forall v \in \mathbf{V},\ \mathbb{H}^v$ un ensemble de votants par vue

$$\forall h \in \mathbb{H}^v,\ h : \mathbb{X}^v \to \mathbb{R}$$



Hiérarchie de distributions sur les votants

π distribution hyper-prior sur $\mathbf{V}$

ρ distribution hyper-posterior sur $\mathbf{V}$

$\forall v \in \mathbf{V},\ P^v$ distribution prior sur $\mathbb{H}^v$

$\forall v \in \mathbf{V},\ Q^v$ distribution posterior sur $\mathbb{H}^v$

$$\underset{\mathbb{S}\sim\mathcal{D}^m}{\mathbb{P}}\left[\ \forall v \in \mathbf{V}, \forall \text{ posterior } Q^v \text{ sur } \mathbb{H}^v,\ \forall \text{ hyper-posterior } \rho \text{ sur } \mathbf{V},\right.$$
$$\left.\underset{v\sim\rho}{\mathbb{E}}\ \underset{h\sim Q^v}{\mathbb{E}} \mathrm{R}_{\mathcal{D}}(h) \ \le\ \underset{v\sim\rho}{\mathbb{E}}\ \underset{h\sim Q^v}{\mathbb{E}} \widehat{\mathrm{R}}_{\mathbb{S}}(h) + \sqrt{\frac{1}{2m}\left(\underset{v\sim\rho}{\mathbb{E}} \mathrm{KL}(Q^v\|P^v) + \mathrm{KL}(\rho\|\pi) + \ln\frac{2\sqrt{m}}{\delta}\right)}\ \right] \ge 1-\delta$$

Le vote de majorité stochastique PAC-Bayésien

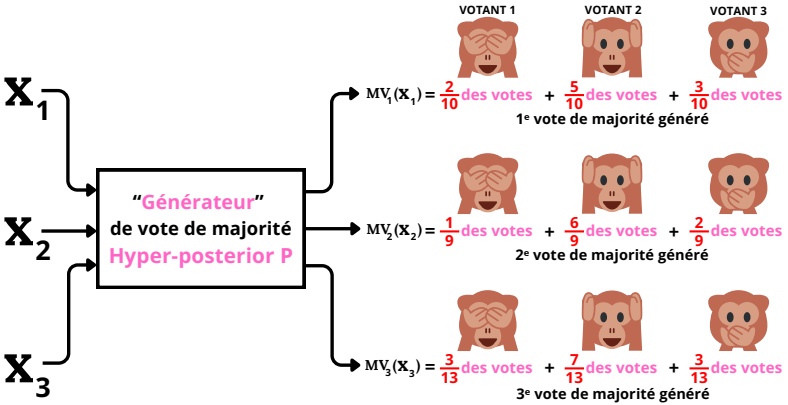
Rappel

$\mathop{\mathbb{E}}_{h\sim\rho} R(h)$ $\rightsquigarrow$ Les bornes PAC-Bayésiennes sont probabilistes sur $\mathbb{H}$

$\rightsquigarrow$ Elles portent sur un modèle stochastique, plutôt qu'un modèle déterministe

MAIS elles permettent de borner le risque du vote de majorité $\mathrm{MV}(\cdot) = \mathrm{sign}\left[\mathop{\mathbb{E}}_{h\sim\rho} h(\cdot)\right]$

$\hookrightarrow$ Via des relaxations *plus ou moins* précises et pas nécessairement facile à minimiser



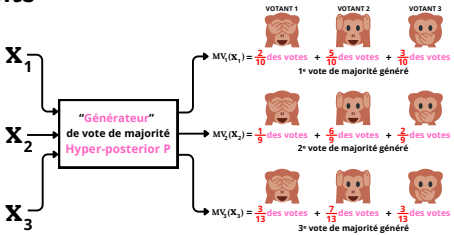
Le vote de majorité stochastique PAC-Bayésien

Une sorte de hiérarchie de distributions sur les votants

Π hyper-prior sur $\mathbb{H}$ (distribution de dirichlet)

P hyper-posterior $\mathbb{H}$ $\quad$ ρ posterior tirée selon P

$$\hookrightarrow \mathrm{MV}(\cdot) = \operatorname{sign}\left[\mathop{\mathbb{E}}_{h\sim\rho} h(\cdot)\right]$$



$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\forall \text{ hyper-posterior } \mathrm{P} \text{ sur } \mathbb{H},\ \mathop{\mathbb{E}}_{\rho\sim\mathrm{P}} \mathrm{R}_{\mathcal{D}}(\mathrm{MV}) \leq \mathop{\mathbb{E}}_{\rho\sim\mathrm{P}} \widehat{\mathrm{R}}_{\mathbb{S}}(\mathrm{MV}) + \sqrt{\frac{1}{2m}\left(\mathrm{KL}(\mathrm{P}\|\Pi) + \ln\frac{2\sqrt{m}}{\delta}\right)}\right] \geq 1-\delta$$

Désintégration des bornes PAC-Bayésiennes

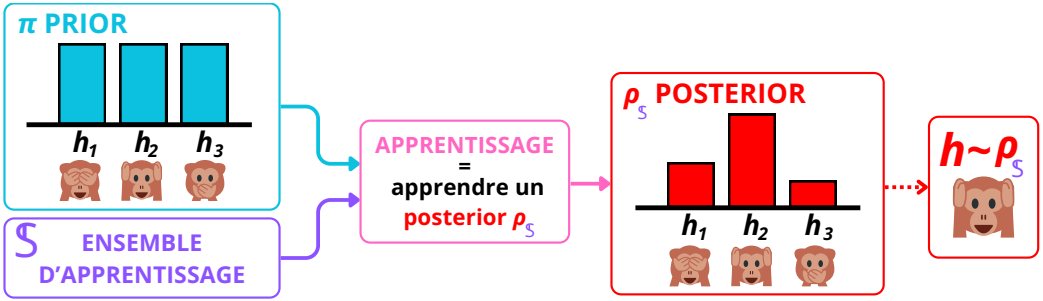
Rappel

$\mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}(h)$ $\rightsquigarrow$ Les bornes PAC-Bayésiennes sont probabilistes sur $\mathbb{H}$

$\rightsquigarrow$ Elles portent sur un modèle stochastique, plutôt qu'un modèle déterministe

$\Downarrow$

Désintégrer les bornes permet de ne les faire porter que sur un **seul** modèle



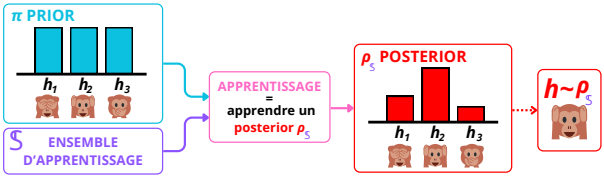
Désintégration des bornes PAC-Bayésiennes

$\mathbb{S} \sim \mathcal{D}^m$ ensemble d'apprentissage

π distribution prior sur $\mathbb{H}$

A algorithme déterministe

$\hookrightarrow A(\mathbb{S}, \pi) = \rho_{\mathbb{S}}$ distribution posterior sur $\mathbb{H}$



$$\mathop{\mathbb{P}}_{\substack{\mathbb{S}\sim\mathcal{D}^m \\ h\sim\rho_{\mathbb{S}}}} \left[R_{\mathcal{D}}(h) \leq \widehat{R}_{\mathbb{S}}(h) + \sqrt{\frac{1}{2m}\left[\underbrace{D_\alpha(\rho_{\mathbb{S}}\|\pi)}_{\text{Divergence de Rényi}} + \frac{2\alpha-1}{\alpha-1}\ln\frac{1}{\alpha} + \ln(2\sqrt{m}) \right]} \right] \geq 1-\delta$$

$$D_\alpha(\rho_{\mathbb{S}}\|\pi) = \frac{1}{\alpha-1}\ln\left[\mathop{\mathbb{E}}_{h\sim\pi}\left[\frac{\rho_{\mathbb{S}}(h)}{\pi(h)} \right]^\alpha \right]$$

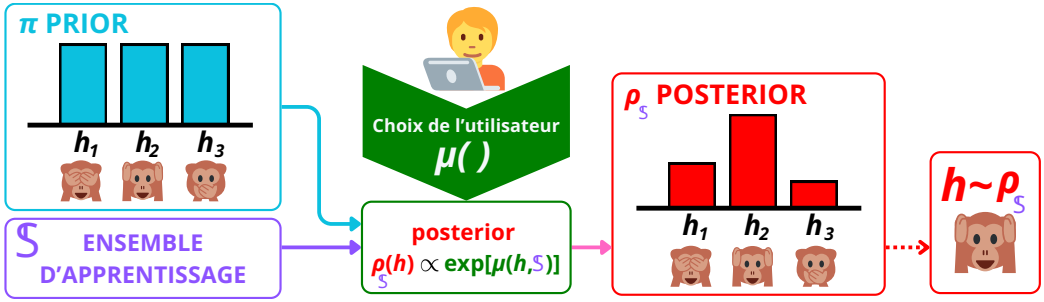
Complexités arbitraires

Une remarque importante

Le terme de complexité dans les bornes en généralisation dépend du cadre considéré

⇓

Comment rendre les bornes plus flexibles ?



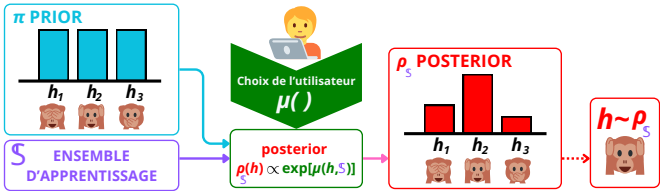
Complexités arbitraires

On considère une **fonction personnalisable** $\mu : \mathbb{H} \times (\mathbb{X} \times \mathbb{Y})^m \to \mathbb{R}$ (*e.g.*, une régularisation)

$\mathbb{S} \sim \mathcal{D}^m$ ensemble d'apprentissage

π distribution prior sur $\mathbb{H}$

$\rho_{\mathbb{S}}$ distribution posterior sur $\mathbb{H}$
telle que $\rho_{\mathbb{S}}(h) \propto e^{\mu(h,\mathbb{S})}$



Forme générale : $\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m, h\sim\rho_{\mathbb{S}}} \left[\mathrm{R}_{\mathcal{D}}(h) \leq \widehat{\mathrm{R}}_{\mathbb{S}}(h) + \Phi_{\mu}(h, \mathbb{S}, \delta) \right] \geq 1 - \delta$

$$\mathop{\mathbb{P}}_{\substack{h'\sim\pi,\\ \mathbb{S}\sim\mathcal{D}^m,\, h\sim\rho_{\mathbb{S}}}} \left[\mathrm{R}_{\mathcal{D}}(h) \;\leq\; \widehat{\mathrm{R}}_{\mathbb{S}}(h) \;+\; \sqrt{\frac{1}{2m}\left[\underbrace{\mu(h', \mathbb{S})}_{\text{Prior}} - \underbrace{\mu(h, \mathbb{S})}_{\text{Posterior}} + \frac{8\sqrt{m}}{\delta^2} \right]_+} \right] \geq 1 - \delta$$

Propriétés clés des bornes PAC-Bayésiennes

- **Flexibles** et adaptables à différents scénarios
  - Compatibles avec différents types de modèles
    - Modèle stochastique par nature
    - Modèle déterministe par nature via les votes de majorité
    - Modèle déterministe via la désintégration
  - (Hiérarchie de) distributions sur des ensembles (de modèles ou d'autres objets)

- **Intermédiaires** entre bornes en convergence uniforme et dépendante d'un algorithme
  - Les complexités arbitraires permettent même de retrouver ces deux types de bornes

- **Précises** et informatives
  - Particulièrement pour affiner un modèle *a priori*

- **Pratiques** et exploitables $\rightsquigarrow$ Souvent calculables ou majorables
  - Minimisables directement via un algorithme auto-certifié
  - Meilleure explicabilité et confiance des modèles par nature de l'algorithme

Forme générale des résultats

Forme générale des bornes PAC-Bayésiennes

Borne en espérance sur $\mathbb{H}$

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\ \forall \rho,\ D\Big(\mathop{\mathbb{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h),\ \mathop{\mathbb{E}}_{h\sim\rho} \widehat{\mathrm{R}}_{\mathbb{S}}(h)\Big) \ \leq\ \Phi(\Delta(\rho,\pi),\mathbb{S},\delta)\ \right] \geq 1-\delta$$

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m,\, h\sim\rho_{\mathbb{S}}}\left[\ D\big(\ \mathrm{R}_{\mathcal{D}}(h)\ ,\ \widehat{\mathrm{R}}_{\mathbb{S}}(h)\ \big) \ \leq\ \Phi(\Delta(\rho_{\mathbb{S}},\pi),\mathbb{S},\delta)\ \right] \geq 1-\delta$$

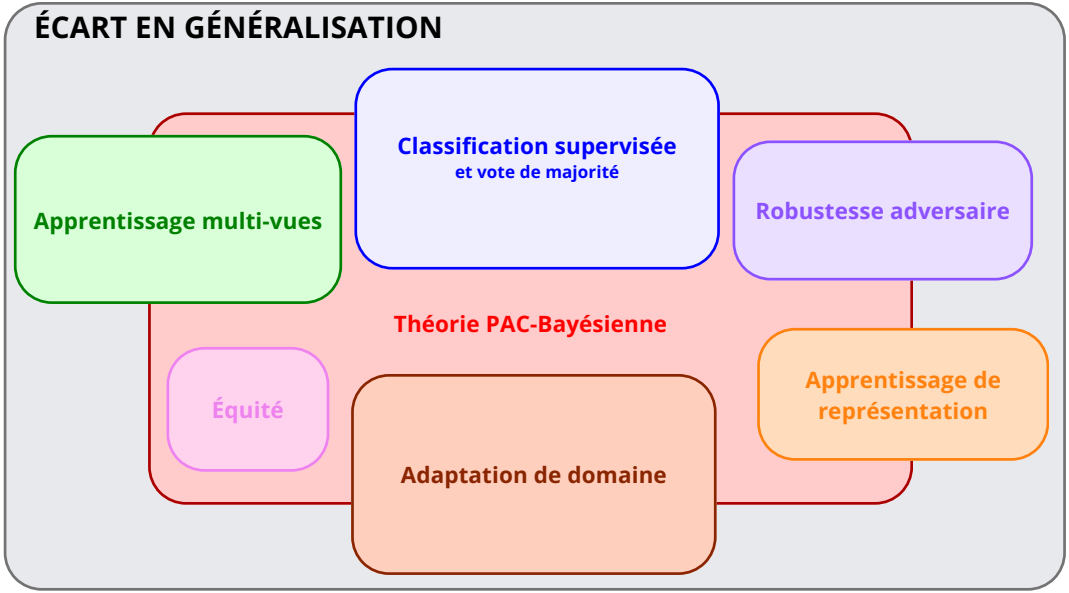
Borne pour un unique $h\sim\rho_{\mathbb{S}}$ — OU Peut être personnalisable

$$\mathop{\mathbb{P}}_{\mathbb{S}\sim\mathcal{D}^m}\left[\ \underbrace{D\Big(\mathrm{R}_{\mathcal{D}}(\text{modèle}),\widehat{\mathrm{R}}_{\mathbb{S}}(\text{modèle})\Big)}_{\textbf{écart en généralisation}} \leq \text{Borne}\ \right] \geq 1-\delta$$

écart en généralisation

distance entre risque réel et empirique

En fait. . .


ÉCART EN GÉNÉRALISATION
Classification supervisée
et vote de majorité
Apprentissage multi-vues
Robustesse adversaire
Théorie PAC-Bayésienne
Équité
Apprentissage de
représentation
Adaptation de domaine

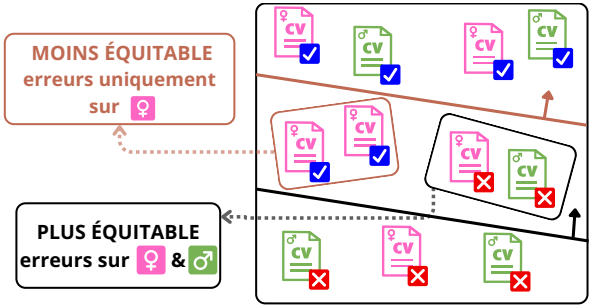
AU-DELÀ DE L'ÉCART EN GÉNÉRALISATION

Projet de recherche : Au-delà de l'écart en généralisation

Tâche de classification supervisée équitable

Classer des CVs en 2 catégories :
☑ pertinent ☒ non pertinent

Objectif : Trouver un modèle équitable et garantir de bonnes performances
⇒ Ne pas apprendre les biais !

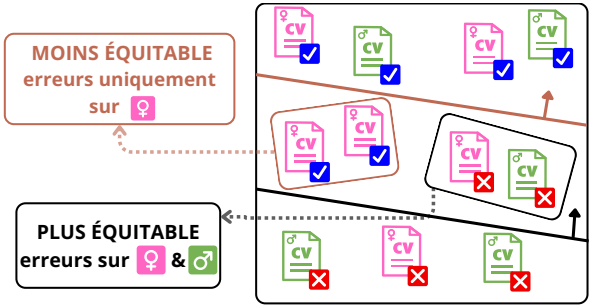


Projet de recherche : Au-delà de l'écart en généralisation

Tâche de classification supervisée équitable

Classer des CVs en 2 catégories :
☑ pertinent ☒ non pertinent

Objectif : Trouver un modèle équitable et garantir de bonnes performances
⇒ Ne pas apprendre les biais !



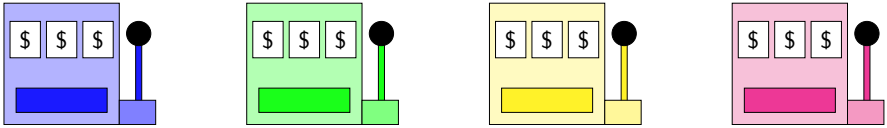
Une solution : Considérer une contrainte/mesure d'équité
↪ **Équité** = capacité à bien classer malgré les biais dans les données

- Algorithme auto-certifié PAC-Bayésien pour garantir un compromis performance/équité
  - Contrôle de l'écart entre les **biais observés** et l'**équité espérée**
- Construire des mesures d'équité robustes aux biais

Projet de recherche : Au-delà de l'écart en généralisation

Les bandits manchots

Bras 1 – $r_t(1) \sim \mathcal{D}_1$ Bras 2 – $r_t(2) \sim \mathcal{D}_2$ Bras 3 – $r_t(3) \sim \mathcal{D}_3$ Bras 4 – $r_t(4) \sim \mathcal{D}_4$



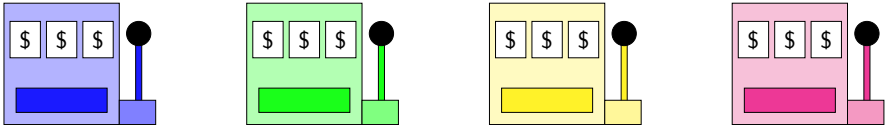
Quel bras/machine choisir à l'étape t pour maximiser le gain cumulé *(reward)*

Objectif : Trouver la meilleure stratégie *(policy)* pour maximiser le gain cumulé

Projet de recherche : Au-delà de l'écart en généralisation

Les bandits manchots

Bras 1 – $r_t(1) \sim \mathcal{D}_1$ Bras 2 – $r_t(2) \sim \mathcal{D}_2$ Bras 3 – $r_t(3) \sim \mathcal{D}_3$ Bras 4 – $r_t(4) \sim \mathcal{D}_4$



Quel bras/machine choisir à l'étape t pour maximiser le gain cumulé *(reward)*

Objectif : Trouver la meilleure stratégie *(policy)* pour maximiser le gain cumulé

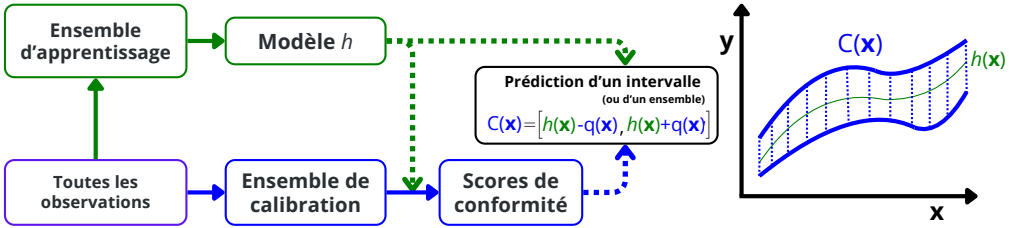
Une solution : Minimiser le regret au cours du temps

↪ **Regret** = Écart entre **gain cumulé actuel** et **gain optimal** (si le meilleur bras était tjrs choisi)

- Algorithme auto-certifié PAC-Bayésien pour garantir un regret faible
- Analyse dynamique du regret au cours du temps
  - Étudier comment le PAC-Bayes peut équilibrer exploration et exploitation

Projet de recherche : Au-delà de l'écart en généralisation

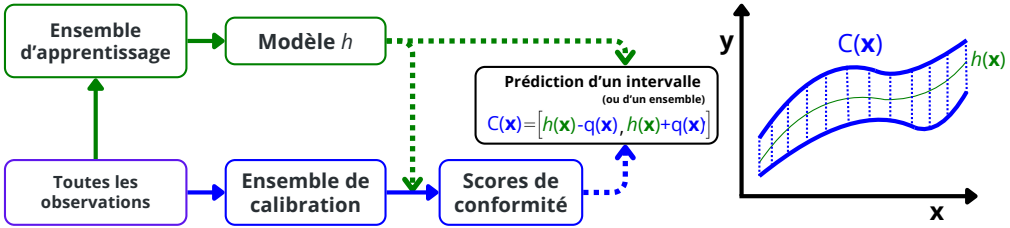
la prédiction conforme : quantification de l'incertitude



Objectif : Comment s'assurer qu'un modèle prédit des intervalles fiables, pas trop grands ?

Projet de recherche : Au-delà de l'écart en généralisation

la prédiction conforme : quantification de l'incertitude



Objectif : Comment s'assurer qu'un modèle prédit des intervalles fiables, pas trop grands ?

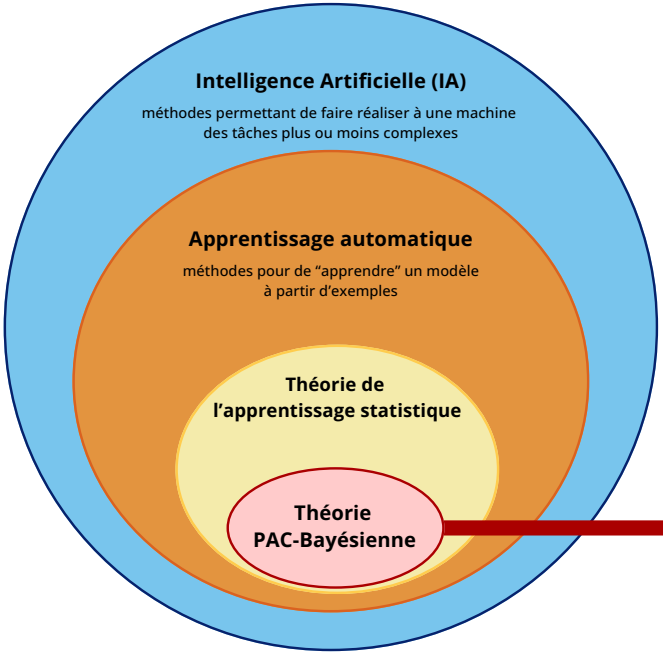
Une solution : Contrôler un compromis entre couverture et efficacité

↪ **couverture** = l'intervalle doit contenir la vraie valeur dans $X\%$ des cas

↪ **efficacité** = l'intervalle doit être le plus petit possible

- Algorithme auto-certifié pour garantir un compromis couverture/efficacité
- Ajustement de la largeur des intervalles via le PAC-Bayes

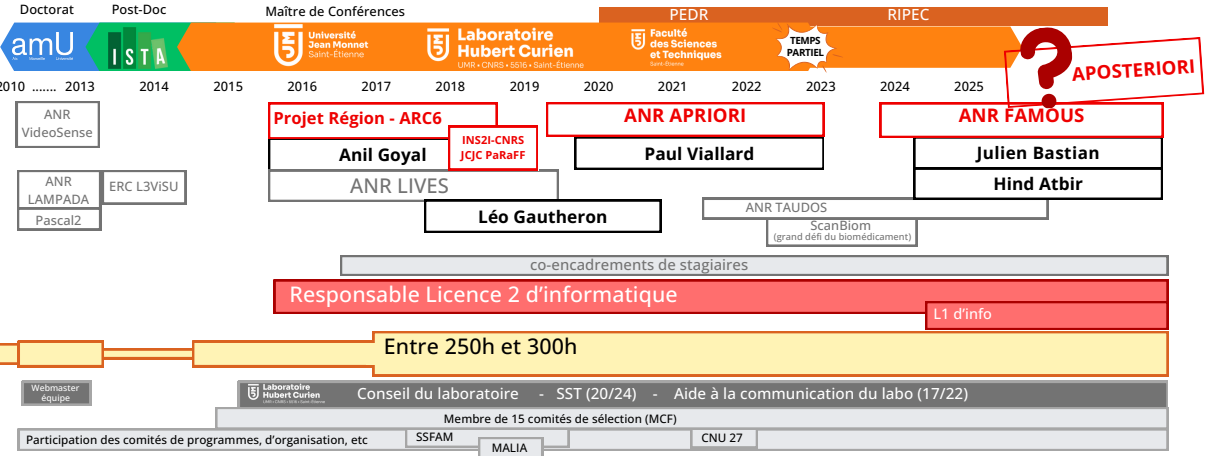
Projet de recherche : Au-delà de l'écart en généralisation



ENJEUX SOCIÉTAUX EN IA

- **Confiance**
- **Explicabilité**
- **Adaptabilité**
- **Robustesse**
- **Équité**

MERCI POUR VOTRE ATTENTION



Références I

Vladimir Vapnik et Alexey Chervonenkis. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory of Probability and its Applications*. (1971).

Leslie Valiant. A Theory of the Learnable. *Communications of the ACM*. (1984).

John Shawe-Taylor et Robert Williamson. A PAC Analysis of a Bayesian Estimator. *Annual Conference on Learning Theory*. (1997).

David McAllester. Some PAC-Bayesian Theorems. *Annual Conference on Learning Theory*. (1998).

Peter Bartlett et Shahar Mendelson. Rademacher and Gaussian Complexities : Risk Bounds and Structural Results. *Journal of Machine Learning Research*. (2002).

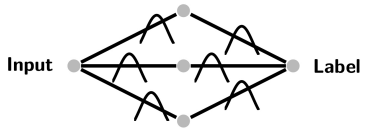
Olivier Bousquet et André Elisseeff. Stability and Generalization. *Journal of Machine Learning Research*. (2002).

Huan Xu et Shie Mannor. Robustness and Generalization. *Machine Learning*. (2012).

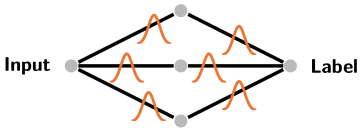
Algorithme de minimisation de la borne désintégrée

Apprentissage d'un réseau de neurone avec une borne PAC-Bayésienne désintégrée

1. Prior/posterior Gaussian distributions associated with the weights of the neural network



Prior distributions



Posterior distribution

2. Learn T priors $\mathbb{P} = \{\pi_t\}_{t=1}^{T}$ in T epochs with $\mathbb{S}'$

3. Learn the distribution $\rho_{\mathbb{S}}$ with $\mathbb{S}$ from a prior π_t selected with $\mathbb{S}$
4. Sample the neural network $h \sim \rho_{\mathbb{S}}$