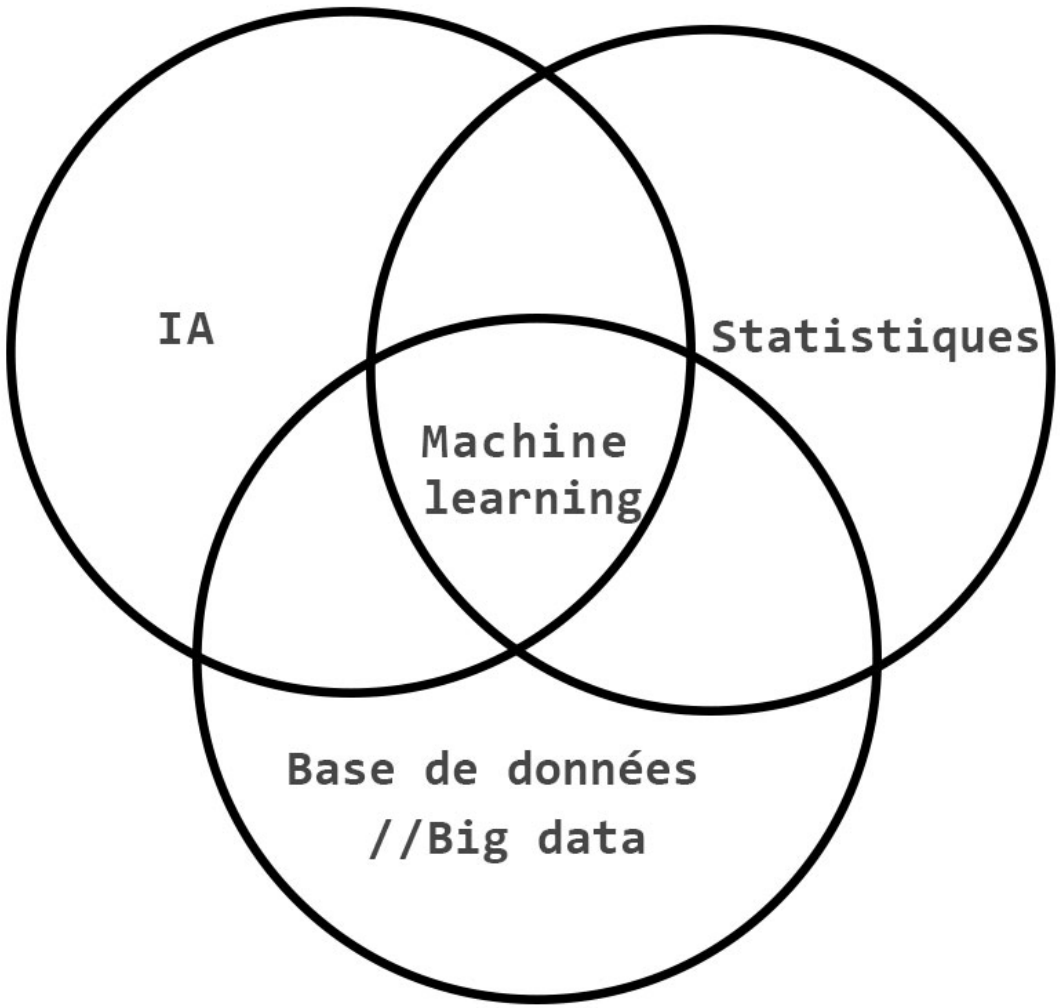
APPRENTISSAGE ARTIFICIEL : INTRODUCTION

M. Serrurier
IRIT, Toulouse, France

December 8,2015

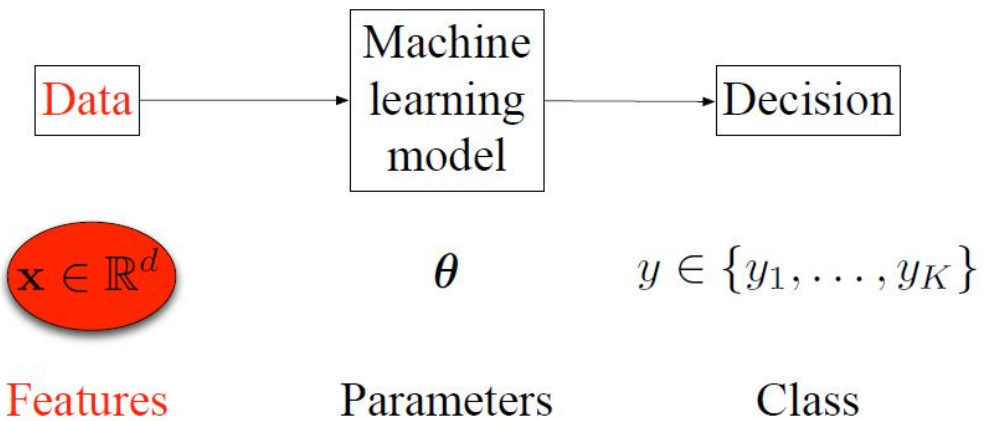
IA STATISTIQUES ET BIG DATA

// LA PLACE DU MACHINE LEARNING



APPRENTISSAGE ARTIFICIEL
// PRINCIPE

- Modélisation mathématique : Création en général **manuelle**, d'un modèle à partir de lois physiques ou mathématiques
- Apprentissage artificiel : Recherche automatique d'un modèle à partir de données pouvant être réutilisé dans un nouvel environnement ou une nouvelle situation (ex prédiction)



- Fouille de données : Extraction d'un maximum d'informations pertinentes

APPRENTISSAGE ARTIFICIEL
// QUIZZ

- Prévision météo
- Analyse de caddy
- Filtrage de spam
- Moteur de proposition (ex. amazon)
- Reconnaissance d'écriture
- Classification automatique de clients d'une assurance
- Création automatique de profil de clients
- Prédiction des numéros du loto

APPRENTISSAGE ARTIFICIEL
// APPLICATION

- Reconnaissance de visage
- Reconnaissance de la parole, de l'écriture
- Robots aspirateur
- Apprendre les préférences d'un utilisateur
- Apprendre à jouer (go, starcraft)
- Détection d'intrusion dans un système

Quel est le nombre a qui prolonge la séquence

$$1 \quad 2 \quad 3 \quad 5 \quad \ldots a?$$

- Solution(s). Quelques réponses valides :
  - $a = 6$. Argument : c'est la suite des entiers sauf 4.
  - $a = 7$. Argument : c'est la suite des nombres premiers.
  - $a = 8$. Argument : c'est la suite de Fibonacci
  - $a =$n'importe quel nombre réel supérieur ou égal à 5. Argument : la séquence présentée est la liste ordonnée des racines du polynôme :

    $$P = x^5 - (11+a)x^4 + (41+11a)x^3 - (61-41a)x^2 + (30+61a)x - 30a$$

    qui est le développement de :

    $$(x-1).(x-2).(x-3).(x-5).(x-a)$$

Généralisation

Il est facile de démontrer ainsi que n'importe quel nombre est une prolongation correcte de n'importe quelle suite de nombre

APPRENTISSAGE ARTIFICIEL
// MÉTHODES CONNUES

- Statistiques
  - Modèles linéaires, LASSO, SVM
  - *k* plus proche voisins, *k* means
  - Hidden Markov Model
  - Boosting
- Intelligence artificielle
  - Arbres de décisions
  - Réseaux de neurones
  - réseaux bayésiens/ Bayésien Naïf

APPRENTISSAGE ARTIFICIEL

// DIFFÉRENTS TYPES DE PROBLÈME

Apprentissage supervisé

Apprentissage non supervisé

Renforcement

APPRENTISSAGE SUPERVISÉ
// FORMALISATION

- Apprentissage supervisé :
  - À partir de l'échantillon d'apprentissage $S = \{(x_i, u_i)\}_{1,m}$
  - on cherche une loi de dépendance sous-jacente
- Par exemple une fonction h aussi proche possible de f (fonction cible) avec f telle que

$$u_i = f(x_i)$$

- Ou bien une distribution de probabilités $P(u_i/x_i)$
- But : afin de prédire l'avenir

APPRENTISSAGE SUPERVISÉ
// TYPE DE PROBLÈME

- Si u est une valeur continue
  - Régression
  - Estimation de densité
- Si u est une valeur discrète/nominale
  - Classification
- Si u est une valeur binaire
  - apprentissage de concept

APPRENTISSAGE SUPERVISÉ
// EXEMPLE

Ciel	Température	Humidité	vent	Jouer
Soleil	Chaud	Forte	faible	Non
Soleil	Chaud	Forte	Fort	Non
Couvert	Chaud	Forte	faible	Oui
Pluie	Doux	Forte	faible	Oui
Pluie	Frais	Normale	faible	Oui
Pluie	Frais	Normale	Fort	Non
Couvert	Frais	Normale	Fort	Oui
Soleil	Doux	Forte	faible	Non
Soleil	Frais	Normale	faible	Oui
Pluie	Doux	Normale	faible	Oui
Soleil	Doux	Normale	Fort	Oui
Couvert	Doux	Forte	Fort	Oui
Couvert	Chaud	Normale	faible	Oui
Pluie	Doux	Forte	Fort	Non

Question

Quelle classe attribuer à : (Soleil,Frais,Forte,Fort) ?

APPRENTISSAGE NON-SUPERVISÉ
// FORMALISATION

- Apprentissage non supervisé :
  - À partir de l'échantillon d'apprentissage $S = \{(x_i\}_{1,m}$
- on cherche des régularités sous-jacentes
  - Sous forme de sous-ensembles (Clustering)
  - Sous forme d'une densité (e.g. mixture de gaussiennes)
  - Sous forme d'un modèle complexe (e.g. réseau bayésien)
- Afin de résumer, détecter des régularités, comprendre

APPRENTISSAGE NON-SUPERVISÉ
// EXEMPLE

Ciel	Température	Humidité	vent	Jouer
Soleil	Chaud	Forte	faible	Non
Soleil	Chaud	Forte	Fort	Non
Couvert	Chaud	Forte	faible	Oui
Pluie	Doux	Forte	faible	Oui
Pluie	Frais	Normale	faible	Oui
Pluie	Frais	Normale	Fort	Non
Couvert	Frais	Normale	Fort	Oui
Soleil	Doux	Forte	faible	Non
Soleil	Frais	Normale	faible	Oui
Pluie	Doux	Normale	faible	Oui
Soleil	Doux	Normale	Fort	Oui
Couvert	Doux	Forte	Fort	Oui
Couvert	Chaud	Normale	faible	Oui
Pluie	Doux	Forte	Fort	Non

Question

Comment séparer les exemples en 3 classes homogènes ?

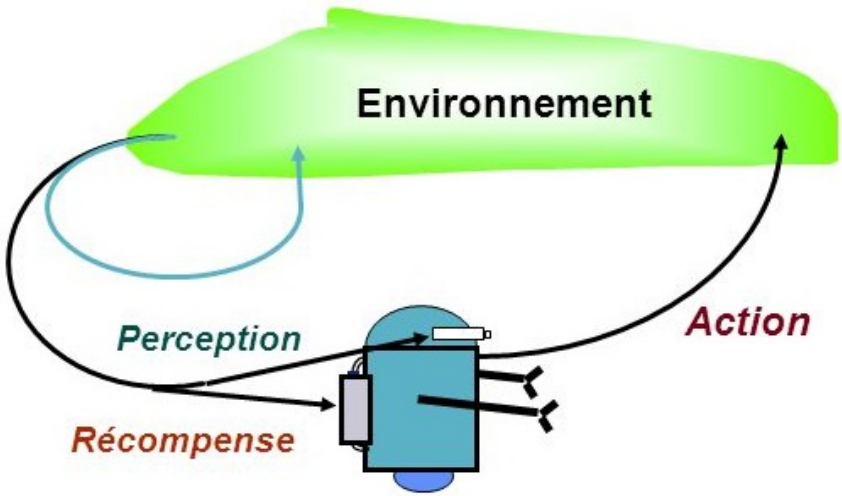
APPRENTISSAGE PAR RENFORCEMENT
// PRINCIPE

Données

- Une séquence de perceptions, d'actions et de récompenses
- Des renforcement *rt*
- *rt* peut sanctionner ou valider des actions

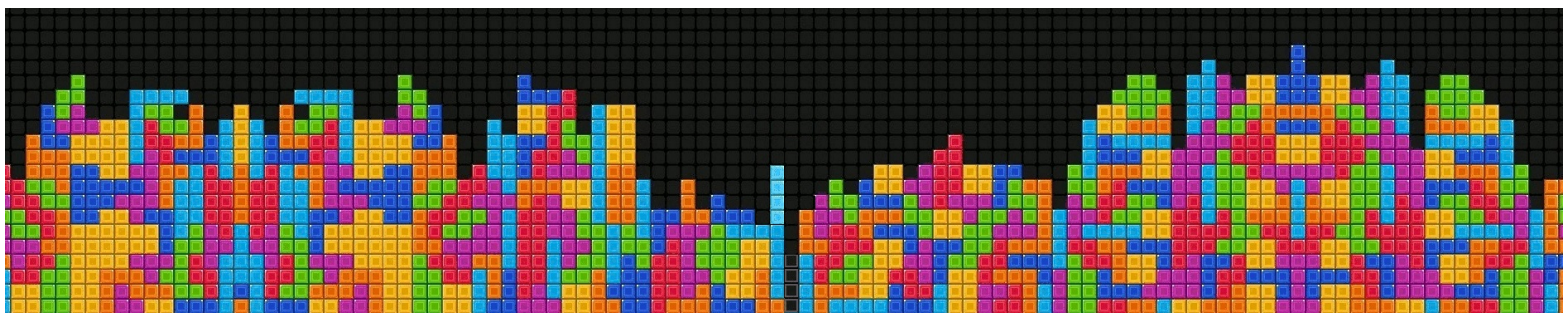
Le problème

Dans une situation donnée, choisir action afin de maximiser un gain sur le long terme



QUEL TYPE DE PROBLÈME ?

- Génération de catégorie d'élève en fonction de leurs notes
- Prédiction du poids d'une personne en fonction de sa taille et de son âge
- IA apprenant à jouer à Tetris
- Prédiction de catégorie de mauvais payeurs dans une assurance
- Prédiction du taux échappement de CO_2 d'une faille en fonction de ses caractéristiques géologiques



APPRENTISSAGE
// QUESTIONS ESSENTIELLES

- Données et connaissances a priori
  - Quelles données sont disponibles ?
  - Que sait-on du problème ?
- Représentation
  - Comment représenter les exemples ?
  - Comment représenter les hypothèses ?
- Méthode et estimation
  - Quel est l'espace des hypothèses ?
  - Comment évaluer une hypothèse en fonction des exemples connus ?
- Évaluation de la performance après apprentissage ?
- Comment reconsidérer l'espace des hypothèses ?

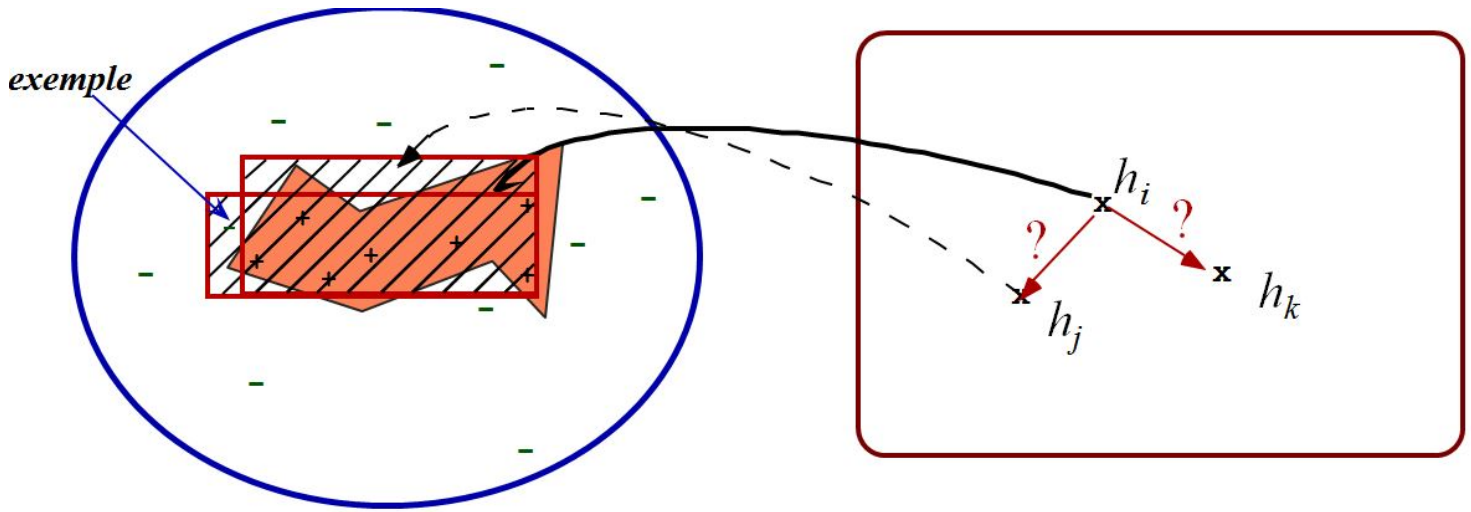
APPRENTISSAGE
// INGRÉDIENTS

- Espace des hypothèses H
- Espace des entrées X
- Protocole
  - Passif ou actif ? Incrémental (on-line) ou « tout ensemble » (off-line) ?
- Mesure de performance
- Algorithme d'exploration de H
  - algorithme local, global, descente de gradient, …

- Classification :
  - Taux d'erreur en classification
- Régression :
  - Erreur au carré
- Densité :
  - vraisemblance
- Idéal : faire ces mesures sur toute les données
- En pratique : mesures sur l'échantillon

CHOIX D'ESPACE DES HYPOTHÈSES

// ILLUSTRATION



Espace des exemples : **X** Espace des hypothèses : **H**

CHOIX D'ESPACE DES HYPOTHÈSES
// BORNE ERREUR RÉELLE

- Il faut contrôler l'expressivité de l'espace d'hypothèses
- Analyse statistique de l'induction [Vapnick]

borne sup

$$R_{rel}(h) \leq R_{Emp} + \sqrt{\frac{1}{m}(ln(G_H) - ln(\frac{\lambda}{4}))} + \frac{1}{m}$$

- G_H = dimension de Vapnick-Chervonenkis

CHOIX D'ESPACE DES HYPOTHÈSES
// DIMENSION DE VAPNICK-CHERVONENKIS

- Mesure la complexité de l'espace des hypothèse
- Un espace d'hypothèse pulvérise un ensemble si, pour tout étiquetage de cette ensemble, il existe une hypothèse qui ne fait pas d'erreur

VC dim

taille du plus grand ensemble pulvérisé par l'espace des hypothèses

CHOIX D'ESPACE DES HYPOTHÈSES

// EXERCICE

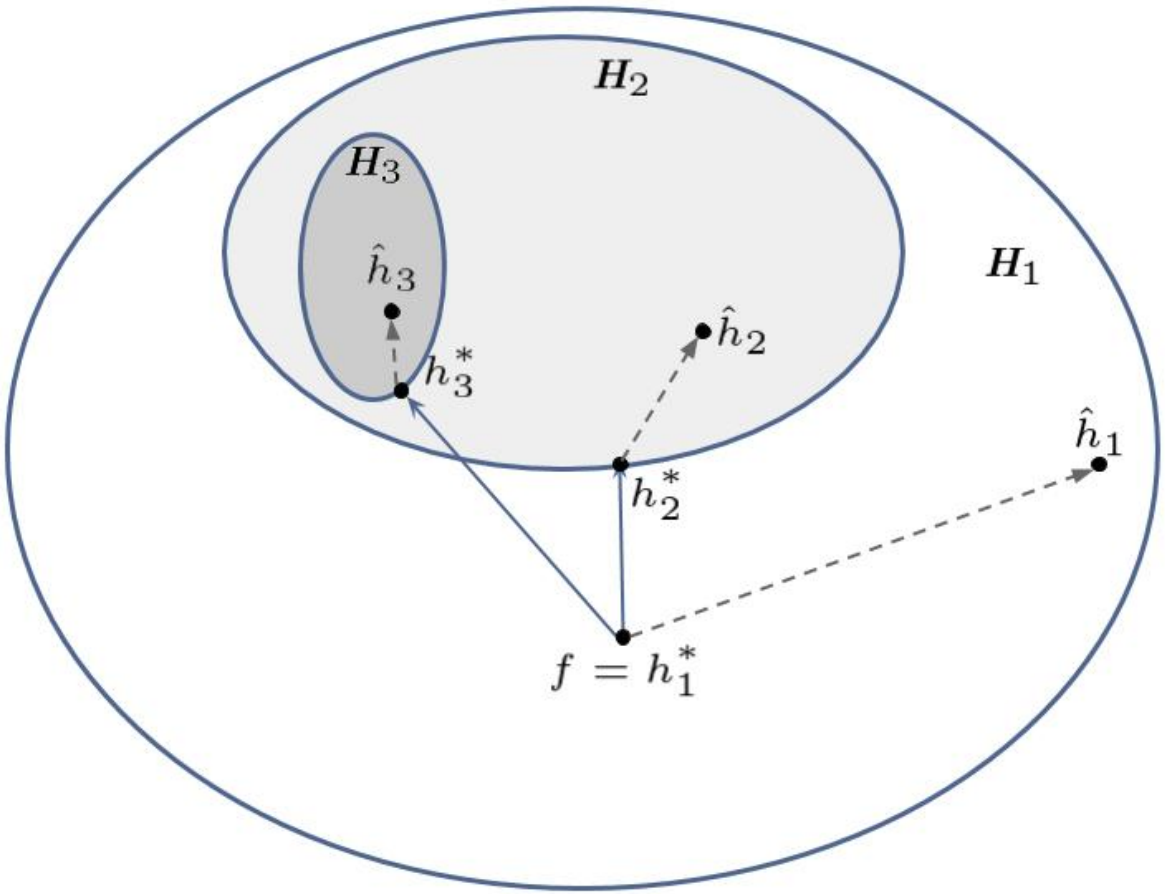
- Calculer la dimension de Vapnick de :
  - Signe de $x - b$ dans R
  - Droite dans R^2
  - Rectangle droit dans R^2
  - Rectangle quelconque dans R^2
  - Polygone dans R^2
  - Hyperplan dans R^d

Biais

toute connaissance qui restreint le champ des hypothèses que l'apprenant doit considérer à un instant donné.

- On ne peut pas apprendre sans biais
- Plus le biais est fort, plus l'apprentissage est facile
- différents biais :
  - biais de représentation
  - biais d'hypothèse
  - biais algorithmique

COMPROMIS BIAIS-VARIANCE
// ILLUSTRATION



- L'induction est une forme d'inférence faillible, il faut donc savoir évaluer sa qualité

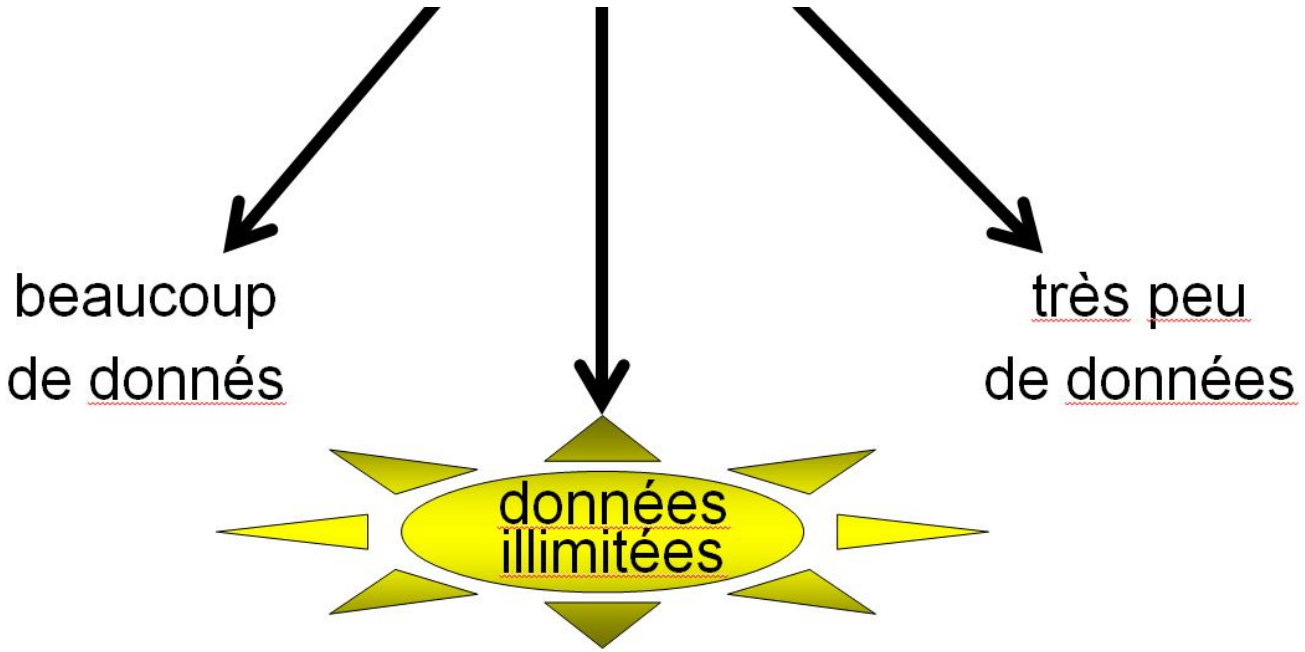
Questions

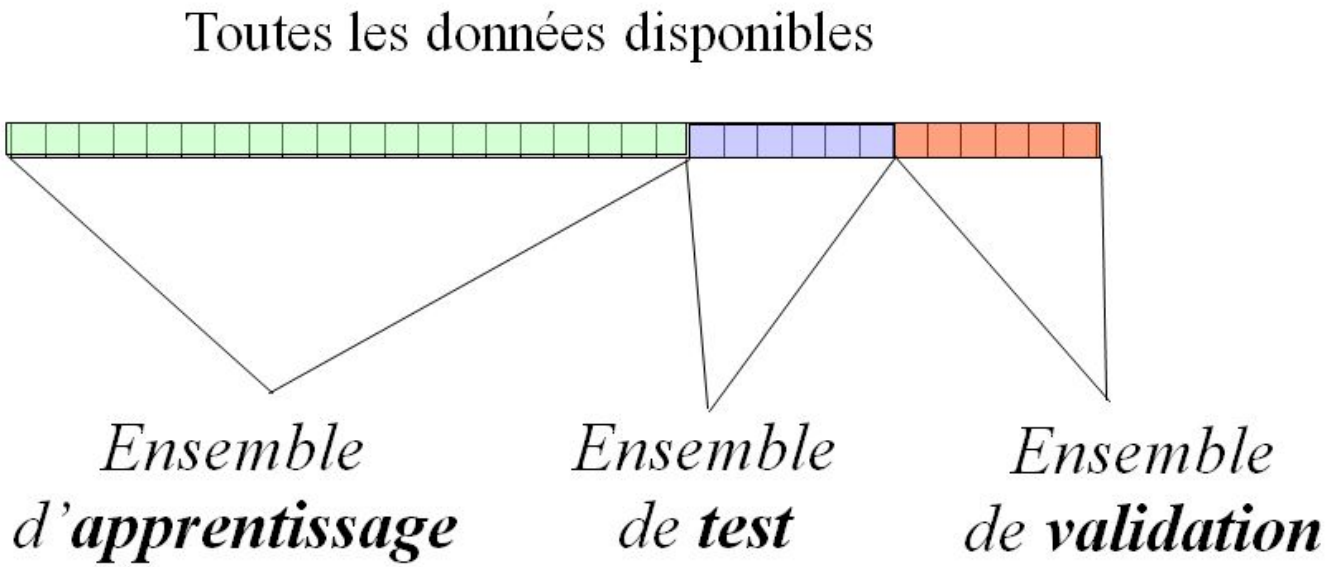
- Quelle est la performance d'un système sur un type de tâche ?
- Est-ce que mon système est meilleur que l'autre ?
- Comment dois-je régler mon système ?

EVALUATION
// APPROCHES

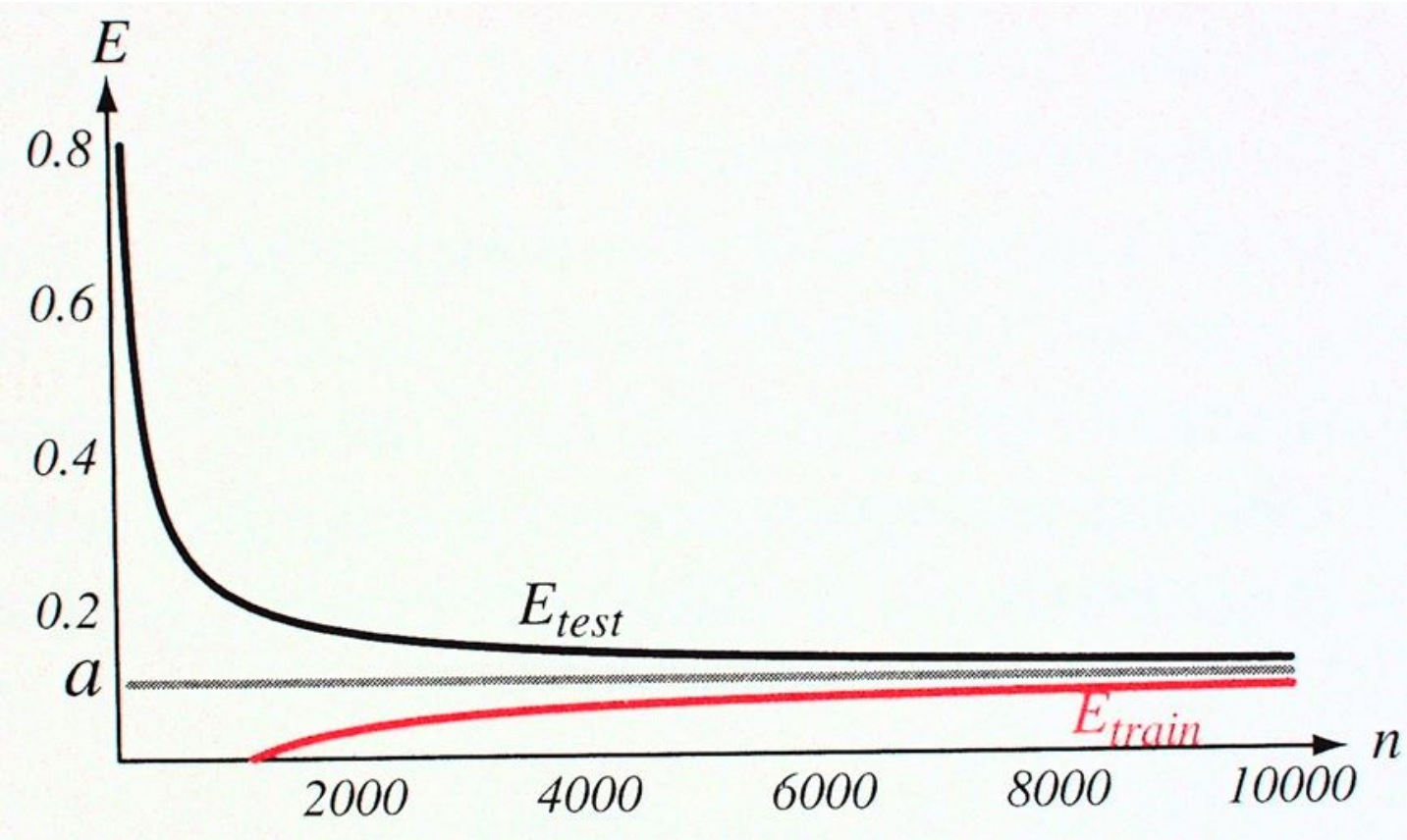
- Évaluation théorique a priori
  - Dimension de Vapnik-Chervonenkis
  - Critères sur la complexité des modèles : MDL / AIC / BIC
  - Estimer l'optimisme de la méthode et ajouter ce terme au taux d'erreur
- Évaluation empirique
  - E.g. taux d'erreur : (dans le cas d'un classifieurs binaire avec une fonction de coût lié au nombre d'erreurs)

EVALUATION
// CAS IDÉAL

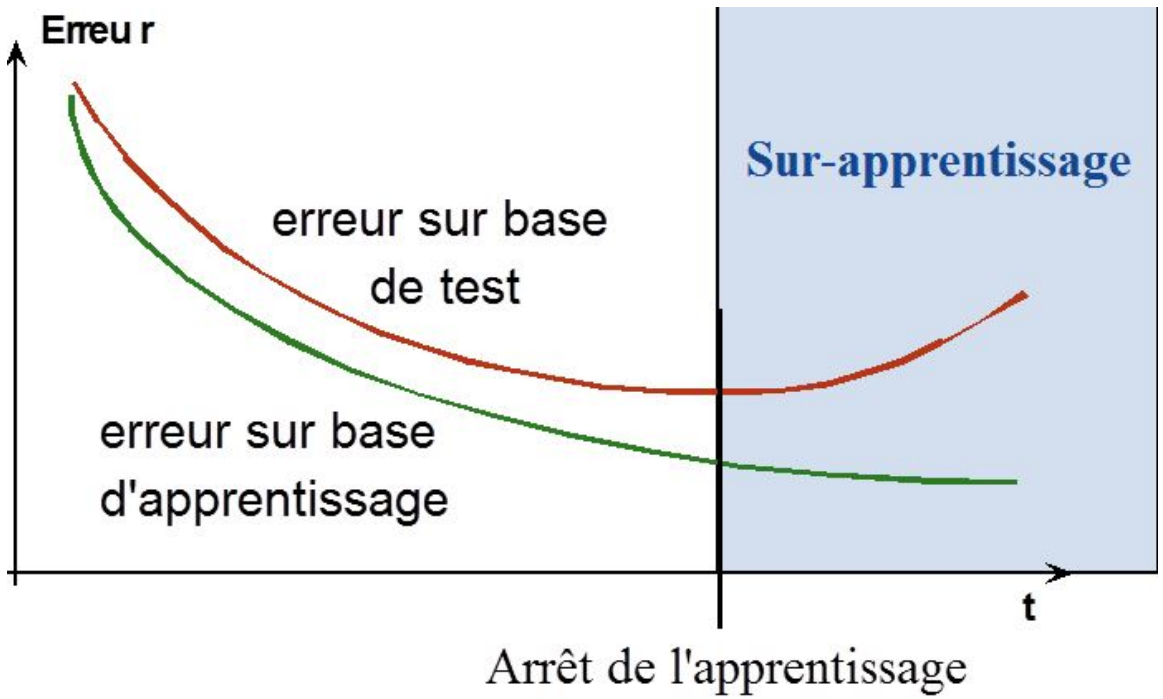



Toutes les données disponibles
Ensemble d'apprentissage
Ensemble de test
Ensemble de validation

EVALUATION
// IDÉAL



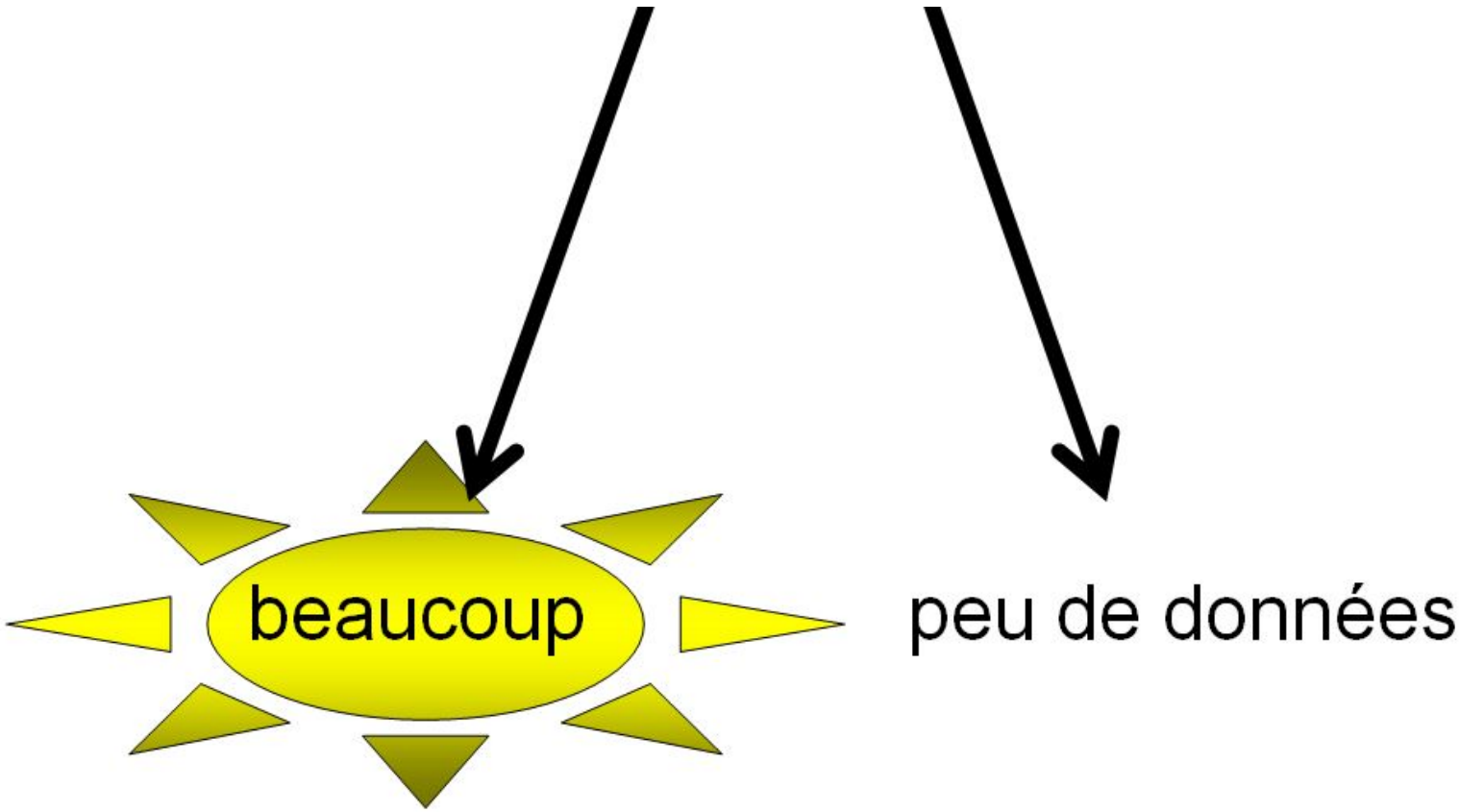
EVALUATION
// SUR-APPRENTISSAGE



EVALUATION
// UTILISATION DE L'ENSEMBLE DE VALIDATION

- On règle les paramètres de l'algorithme d'apprentissage sur la base d'évaluation
  - Nombre de couches cachées d'un réseaux de neurones
  - Nombre de voisin put le *kpp* ...
- En essayant de réduire l'erreur de test
- Pour avoir une estimation non optimiste de l'erreur, il faut recourir à une base d'exemples non encore vus :
  - la base de validation

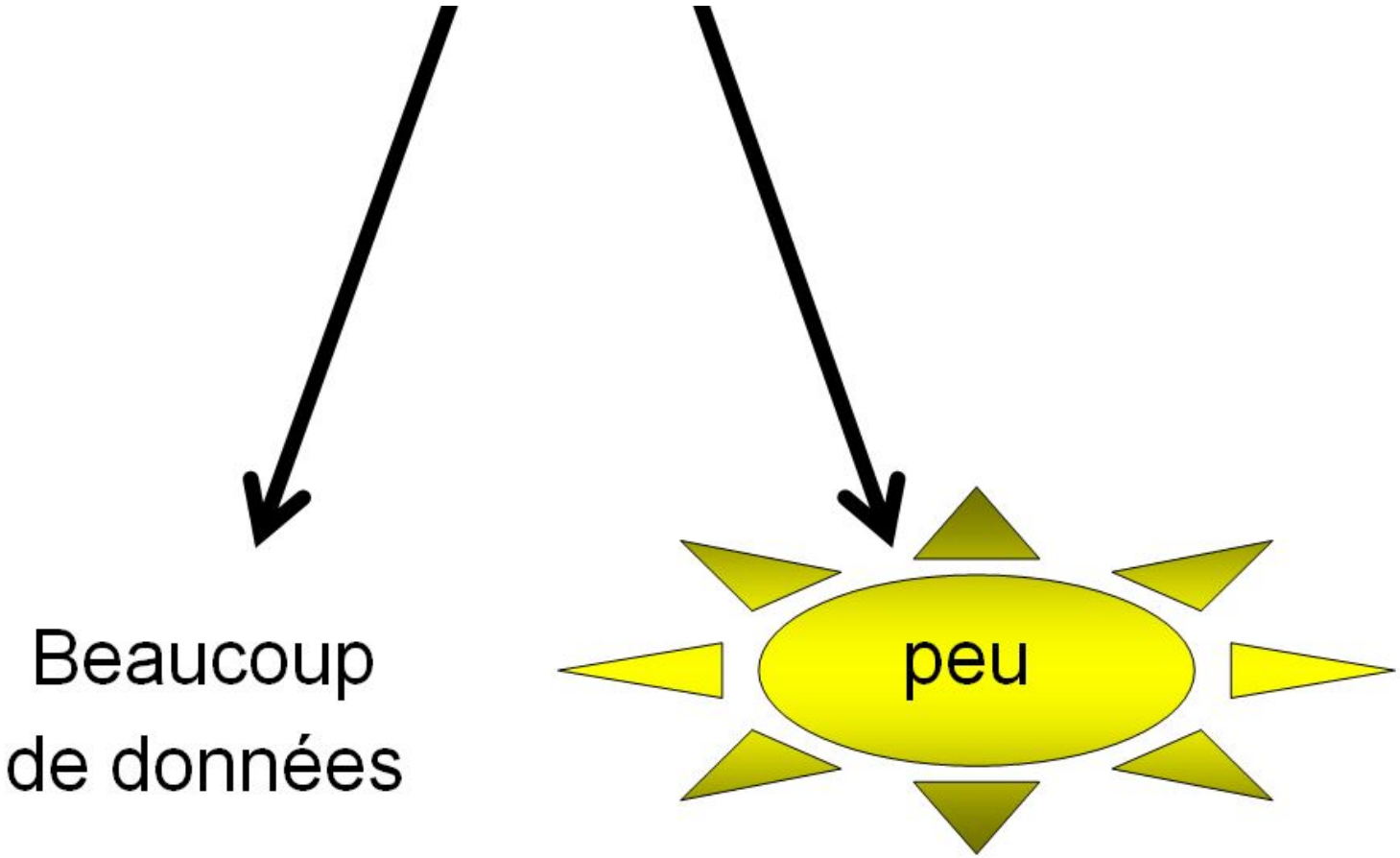
EVALUATION
// CAS FAVORABLE



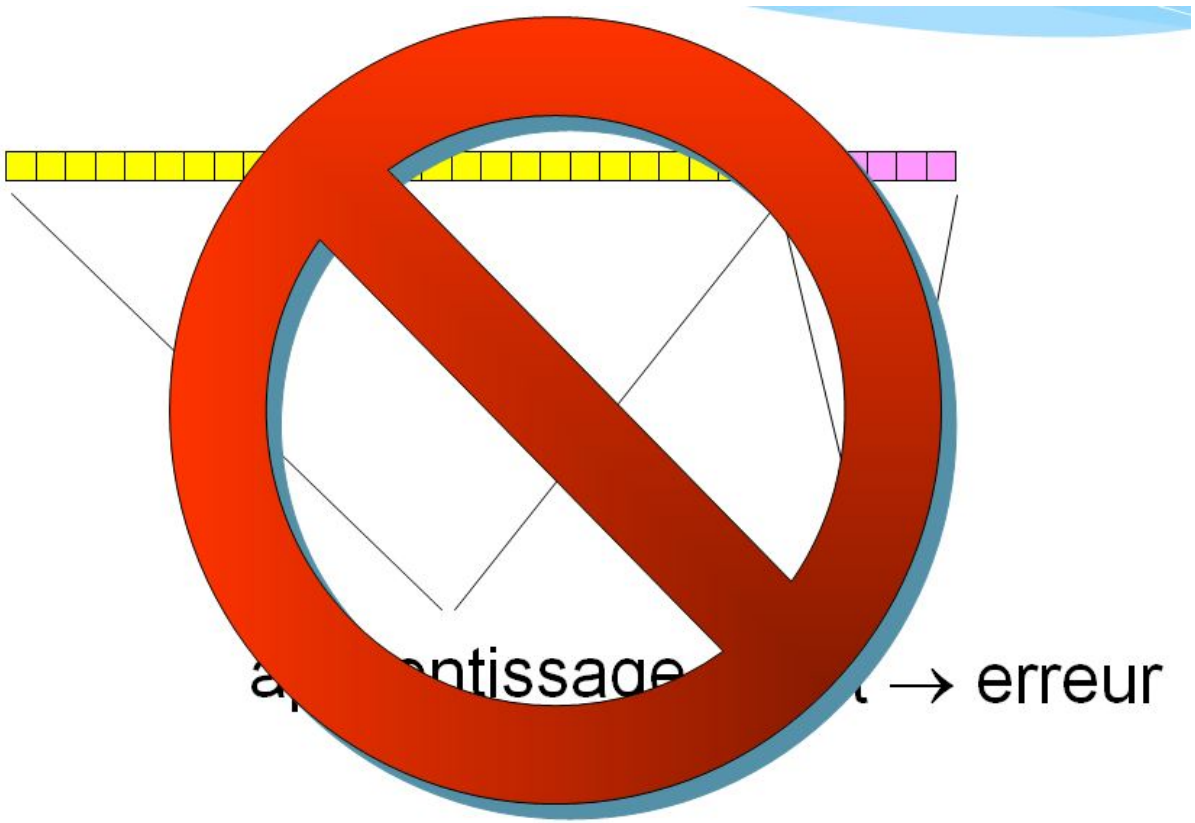
- Comme précédemment + intervalles de confiances
- Exemple :
  - X contient m exemples tirés indépendamment
  - m > 30
- Avec une probabilité de 95%, l'erreur vraie E_{vrai} est dans intervalle :

$$E_{Emp} \pm 1.96 * \sqrt{\frac{E_{Emp} * (1 - E_{Emp})}{m}}$$

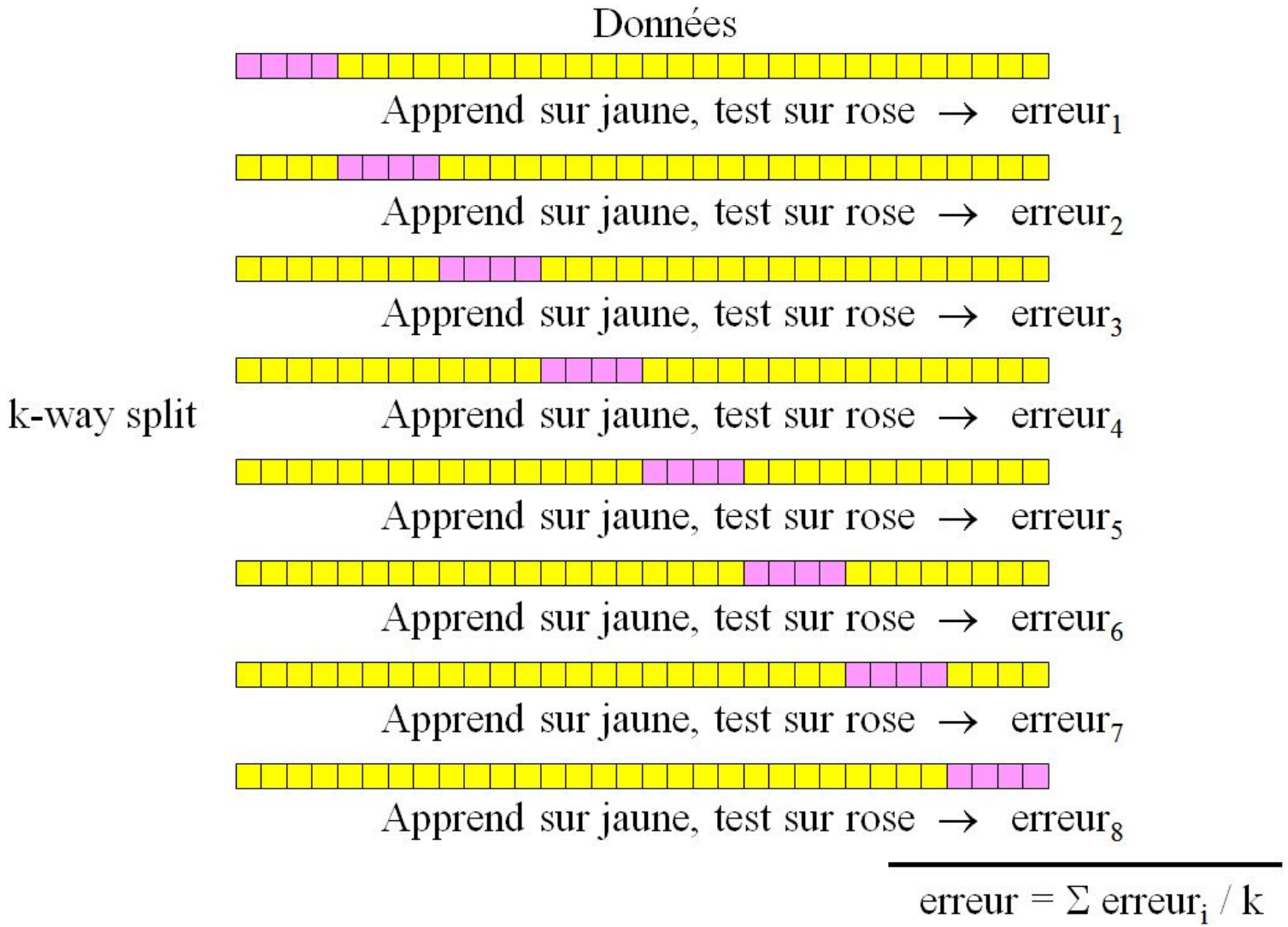
EVALUATION
// CAS GÉNÉRAL



EVALUATION
// ENSEMBLES



EVALUATION
// k VALIDATION CROISÉE



- Attention à votre fonction de coût :
  - qu'est-ce qui importe pour la mesure de performance (ex. distribution déséquilibrée) ?
- Données en nombre fini :
  - calculez les intervalles de confiance
- Données rares :
  - Attention à la répartition entre données d'apprentissage et données test. Validation croisée.

Attention !!

N'oubliez pas l'ensemble de validation

APPRENTISSAGE : LES LIMITES
// POURQUOI ÇA MARCHE POUR GOOGLE ET PAS POUR MOI

No free lunch theorem

Il n'existe pas de modèle qui bat tous les autres sur tous les problèmes d'apprentissage

- Apprendre nécessite des données fiables
  - Variables suffisantes pour construire le modèle
  - Données propre, bien labellisée, sans biais
  - Suffisamment de données : plus le modèle est compliqué plus il faut de données
- De l'expertise
  - Choix des algorithmes/paramètres
  - Sélection de variables
- Du temps de calcul
  - Deep learning pour la traduction : 3 semaines de calculs intensifs

APPRENTISSAGE : LES LIMITES
// RESSOURCES

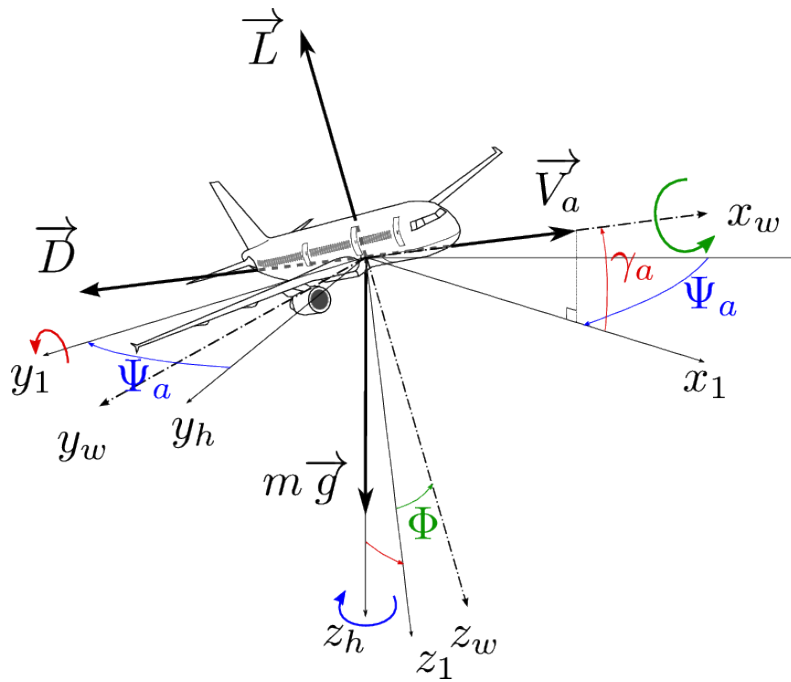
- Données libres
  - UCI
  - Imagenet
- Calcul possible sur carte graphique (GPU)
- Bibliothèque disponible dans plusieurs langages
  - R
  - Python scikit-learn
  - OpenAI

APPRENTISSAGE : LES LIMITES

// PREUVE THÉORIQUE

- Garanties mathématiques
  - Convergence/optimalité
  - Bornes sur l'erreur

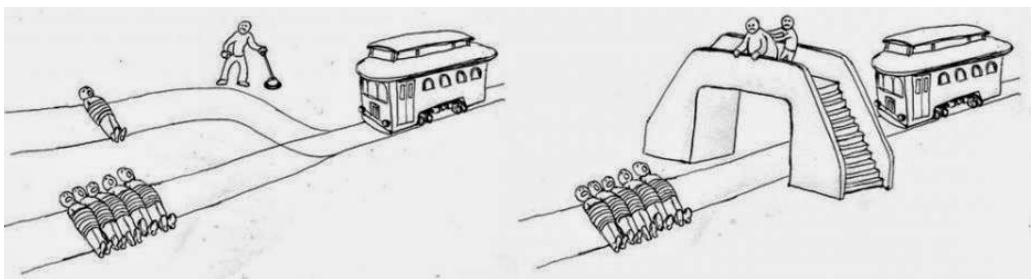
- Problème pour les applications critiques ex :
  - Prédiction de trajectoire d'avion
  - Détection d'intrusions dans un système



APPRENTISSAGE : LES LIMITES
// INTERPRÉTABILITÉ DES RÉSULTATS

- Réseau de neurones = boîtes noires
  - Modèle non interprétable
  - Impossible d'expliquer les prédictions

- Problèmes éthiques et juridiques
  - Choix post bac
  - Prédiction criminalité
  - Problème du tramway : voiture autonome



Intimité différentielle

Problème lié aux informations que l'on veut cacher, mais qui peuvent être trouvé çà l'aide d'algorithme d'apprentissage

- Deux problèmes avec le machine learning et les données publiques
  - Vie privée des utilisateurs
  - Secret industriel

- Exemple
  - Challenge Netflix
  - Poids d'un avions
  - Facebook et le divorce

APPRENTISSAGE ARTIFICIEL
// CONCLUSIONS

- Apprentissage artificiel présent partout :
  - Systèmes de recommandation (amazon, netflix, ...)
  - Publicité en ligne
  - Voiture sans chauffeur
- Apprentissage artificiel au carrefour de nombreux domaines :
  - Statistiques
  - Intelligence artificielle
  - Optimisation
- Pose de nouveaux problèmes Éthiques/philosophique/juridique