

De la Biomédecine Quantitative à la Science des Données et à l'Intelligence Artificielle pour la Santé

Pr Roch Giorgi

 roch.giorgi@univ-amu.fr

En Guise d'Introduction

« Je ne connais pas d'être vivant, de cellule, tissu, organe, individu et peut-être même espèce, dont on ne puisse pas dire qu'il stocke de l'information, qu'il traite de l'information, qu'il émet et qu'il reçoit de l'information. »

Michel Serres

Recherches et Méthodes Quantitatives

- Déductives le plus souvent, testent des hypothèses
- Expérimentales ou non (observationnelles)
- Mesurent un évènement, un fait, un objet dont l'existence est déjà connue
- Objectif de représentativité, visée d'extrapolation
- Déployées sur un nombre variable d'unités d'analyse statistique
 - ✓ Petits à grands échantillons d'individus
- Estimation, inférence statistique

Processus de Recherche



Formulation d'une hypothèse de recherche



Mise en place d'un schéma d'étude adapté



Recueil des données de l'étude



Analyse des données de l'étude



Diffusion des résultats de l'étude

Le Fil d'Ariane

- ▶ Hypothèse de recherche
- ▶ Objectif principal
- ▶ Conduite de l'étude
 - ▶ Schéma d'étude
 - ▶ Population d'étude
 - ▶ Recueil des informations pertinentes
 - ▶ Analyse statistique
- ▶ Résultats de l'étude

Les Domaines des Questions Posées

Pronostic

Étiologie

Prévalence

Traitement

Incidence

Diagnostic/Dépistage

...



Les Données pour Répondre aux Questions

- Évolution au cours du temps
 - ✓ Nature des données recueillies
 - ✓ Nombre d'informations par individus
 - ✓ Nombre de répétitions de la mesure
 - ✓ ...



Socio-démographique

Clinique

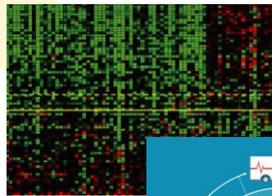
Imagerie

Biologique

Génétique

Textuelle

...



Des Questions pour une Médecine P...

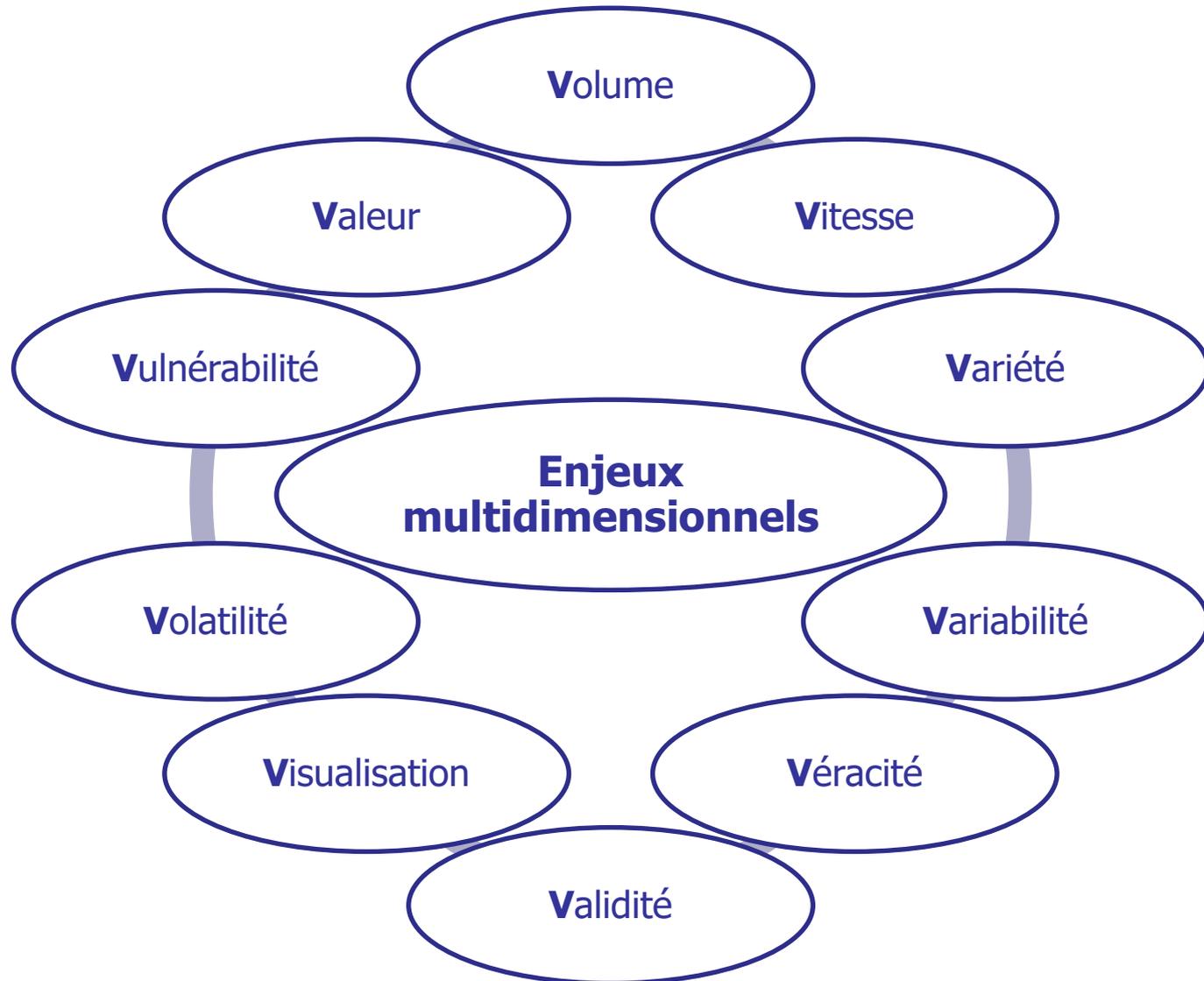
- **P**ersonnalisée
 - ✓ Profil personnel de l'individu
- **P**rédictive
 - ✓ Risque, traitement
- **P**réventive
 - ✓ Primaire / Secondaire / Tertiaire
- **P**articipative
 - ✓ Citoyens, patients acteurs de leur santé et de leurs soins
- **P**ertinente
 - ✓ Fondée sur un service médical rendu
- **P**arcours de soin
 - ✓ Adaptés à chacun

Données Massives (Big Data)

- En santé, mais pas uniquement
- Explosion de la quantité d'informations recueillies à chaque instant
 - ✓ Progrès technologiques
- Aubaine pour la recherche en santé
 - ✓ Source de nouvelles connaissances, indispensables à l'innovation et aux progrès médicaux
- Enjeux autour des données
 - ✓ Capture, partage, stockage, présentation, analyse

Données Massives (Big Data) et .V

(1)



Données Massives (Big Data) et .V (2)

- Enjeux multidimensionnels

- ✓ **V**olume

- En constante augmentation, qu'il faut stocker et traiter
- 43 trillions de gigabytes générés en 2020 (300 x plus qu'en 2002)

- ✓ **V**itesse

- De génération, production, actualisation des données
- 600 téraoctets de données entrantes par jour sur Facebook

- ✓ **V**ariété

- Données brutes, semi-structurées, non structurées (initialement)
- Sources multiples

- ✓ **V**ariabilité

- Incohérence dans les données (anomalies, valeurs aberrantes,...)
- Diversité de dimension des données (multiples types et sources)

Données Massives (Big Data) et .V (3)

- Enjeux multidimensionnels
 - ✓ **V**éracité
 - Confiance en ces données (provenance, fiabilité, contexte,...)
 - ✓ **V**alidité
 - Précision, correction des données pour l'usage auquel elles sont destinées
 - ✓ **V**isualisation
 - Des données et des résultats de l'analyse des données
 - Pour faciliter la compréhension, l'interprétation
 - ✓ **V**olatilité
 - Durée de pertinence, archivage, obsolescence des données

- Enjeux multidimensionnels
 - ✓ **V**ulnérabilité
 - Sécurisation des données
 - Les données sur la santé des personnes sont des données sensibles (CNIL)
 - ✓ **V**aleur
 - Commerciale, du fait de la valeur ajoutée de ces données

- Bases de données médico-administratives
Données objectives, très exhaustives à l'échelle de larges populations
 - ✓ Système National d'Information InterRégimes de l'Assurance Maladie (SNIIRAM)
 - Tous les remboursements effectués par l'Assurance maladie pour chaque cotisant, tout au long de leur vie (biologie, médicaments, ambulances, consultations avec dates et noms des professionnels de santé vus, codes du type de maladie dans certains cas...)
 - ✓ Système National des Données de Santé (SNDS)
 - SNIIRAM + PMSI + BCMD +...

- Cohortes

Groupe de personnes partageant un certain nombre de caractéristiques communes, suivies au cours du temps pour identifier la survenue d'événements de santé, des facteurs de risque ou de protection s'y rapportant

- ✓ Constances (<https://www.constances.fr>)

- Cohorte épidémiologique « généraliste » constituée d'un échantillon représentatif de 200 000 adultes âgés de 18 à 69 ans à l'inclusion, consultants des Centres d'examens de santé (CES) de la Sécurité sociale
- Un examen de santé tous les 5 ans, questionnaire tous les ans

- ✓ Observatoires de la santé en Europe

- Constance France + Allemagne +...

- Etudes cliniques

Travaux de recherche clinique, incluant des populations particulières de patients dont les profils de risque et les états de santé sont analysés

- ✓ Essai clinique DALIA (Vaccine Research Institute)

- Evaluation d'un vaccin thérapeutique contre le VIH
- Comptage de toutes les cellules immunitaires des patients
- Environ 800 mesures par patient et par visite
- Etude de l'expression génétique de nombreux marqueurs (47 000 sondes/patient/visite)
- Séquençage à haut débit du virus

- Objets de santé connectés

Objet équipé de capteur-s, mesurant un-des paramètre-s (glycémie, activité physique, fréquence cardiaque) enregistrés sur un serveur, analysés par un programme informatique pour informer ou alerter l'utilisateur ou le référent (médecin, proche), tracer des courbes de suivi, donner un score,...

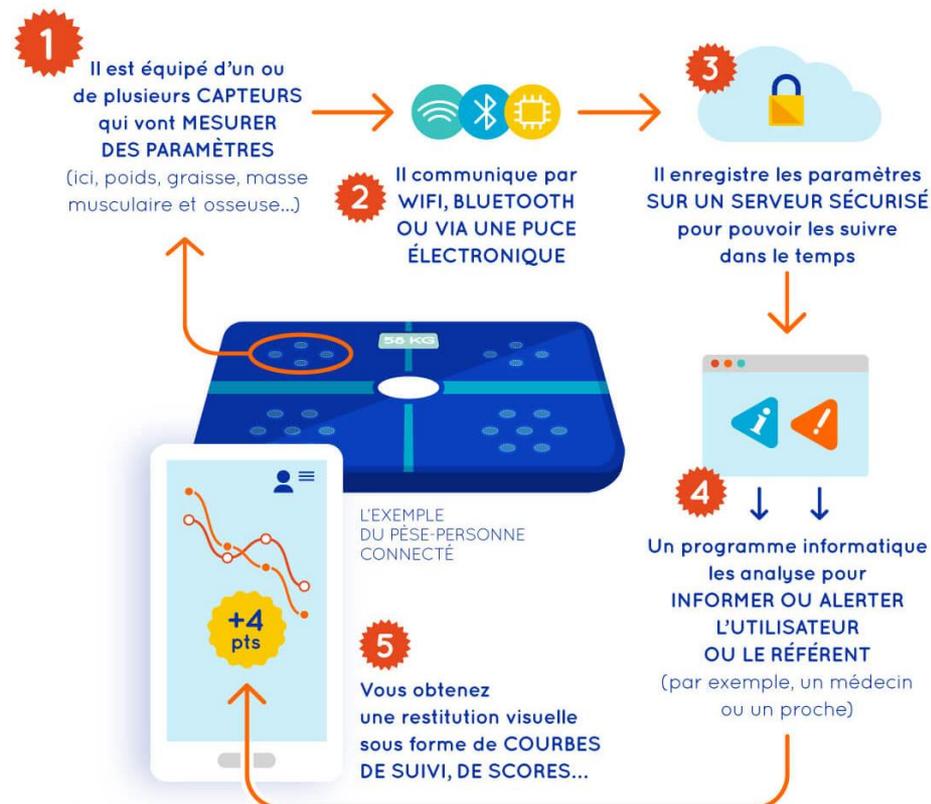
- ✓ De bien être et de loisirs

- Selon appétence individuelle

- ✓ Dispositifs médicaux à finalité médicale déclarée

- Contraintes réglementaires
- Service réel et évalué

- Objets de santé connectés
 - ✓ Un dispositif intégré, sécurisé

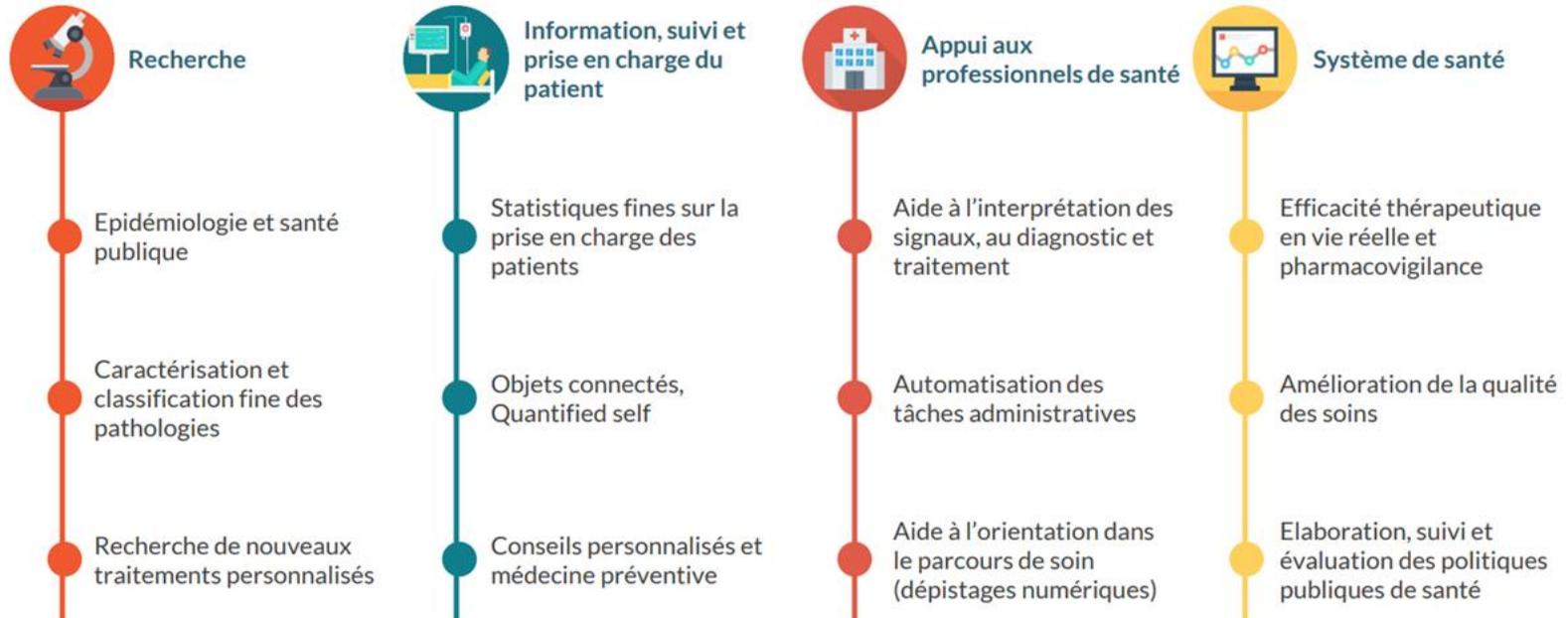


<https://www.assurance-prevention.fr/infographie-objets-connectes-de-quoi-parle-t-on>

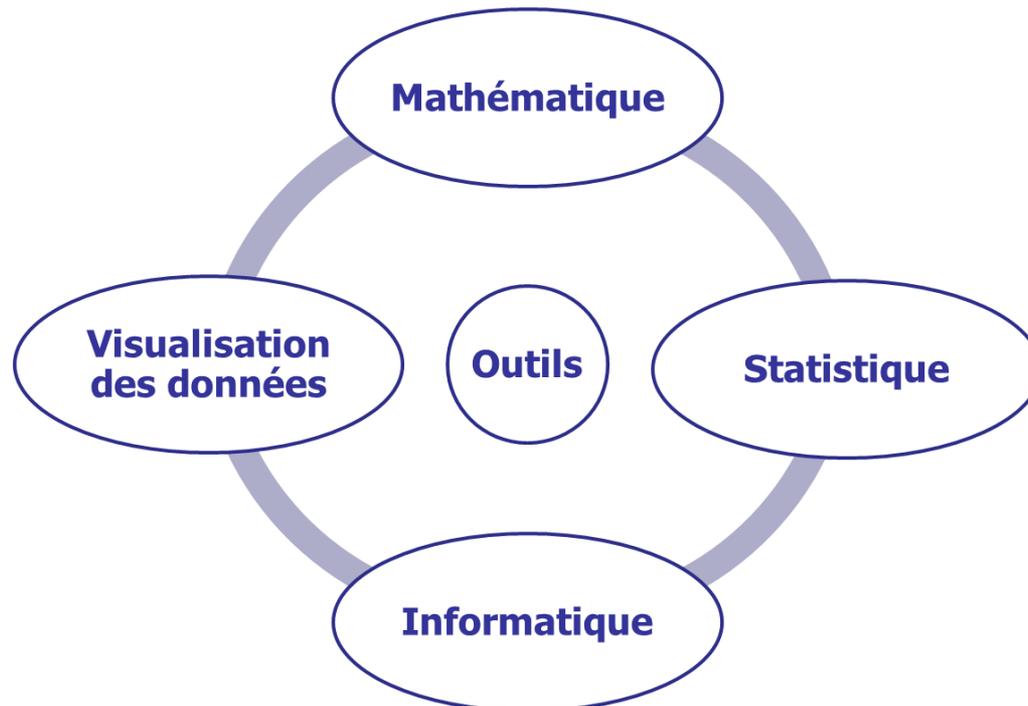
- Bases de données médico-administratives et d'autres sources

Plateforme des données de santé pour faciliter l'accès aux données de santé et leur ré-utilisation

✓ Health Data Hub (<https://www.health-data-hub.fr/>)



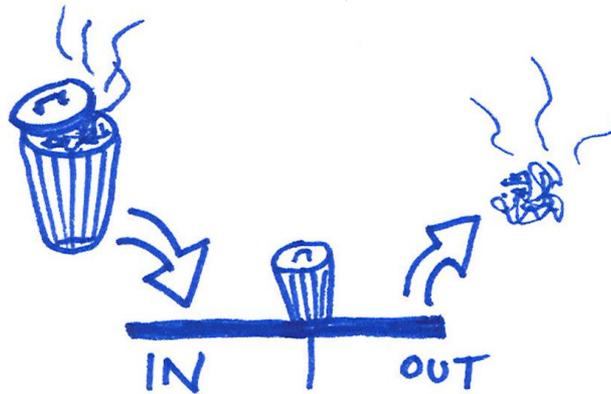
- Interdisciplinaire Informatique / Mathématique
- Extraction de connaissance d'ensembles de données
- Données hétérogènes, structurées ou non



- Méthodes
 - ✓ Fouilles de données (data mining)
 - ✓ Statistiques
 - ✓ Traitement du signal
 - ✓ Apprentissage automatique (machine learning)
 - ✓ Visualisation des données
 - ✓ ...
- Domaines
 - ✓ Nettoyage
 - ✓ Exploration
 - ✓ Classification
 - ✓ Analyse

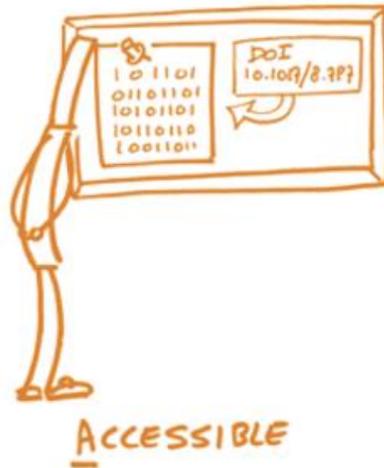
- Analyse nécessitant l'intégration, l'interprétation et l'élaboration de sens (connaissance du domaine)
- + Recherche guidée par les données
- Recherche basée sur des hypothèses

Mais...



Données FAIR

FAIR DATA PRINCIPLES



<https://bibliotheques.u-bordeaux.fr/Soutien-a-la-recherche/Les-Donnees-de-recherche/FAIR-et-DMP-des-acronymes-a-connaître2>



<https://science-ouverte.inrae.fr/fr/les-donnees-et-le-numerique-scientifiques/produire-des-donnees-fair>

Intelligence Artificielle

- Naissance dans les années 1950
 - ✓ Réalisation de tâches humaines par des machines mimant l'activité du cerveau
- Développement dans les années 1980
 - ✓ Développement des techniques informatiques, augmentation de la puissance de calcul
- Nouvel essor depuis 2010
 - ✓ Massification des données
 - ✓ Nouvelle puissance de calcul

Concepts d'Intelligence Artificielle

- IA forte
 - ✓ Concevoir une machine capable de produire un comportement intelligent, ayant une conscience propre et une compréhension de ses propres raisonnements
- IA faible
 - ✓ Concevoir une machine capable d'agir comme si elle était intelligente, capable d'aider les êtres humains dans leurs tâches

Approche Symbolique de l'IA

- Paradigme dominant jusqu'au début des années 1990
- Recherche à expliquer les connaissances
 - ✓ Modélisation du raisonnement logique, représentation de la connaissance par des symboles formels
- Raisonnement en appliquant des règles logiques
 - ✓ Logique et raisonnement symbolique
 - ✓ Traitement automatique du langage naturel
 - ✓ Systèmes à base de règles, à base de connaissances ou systèmes-experts
 - ✓ Algorithmes de recherche dans les espaces d'états
- « Systèmes experts »

Exemple Raisonnement Symbolique

$$A \text{ et } B \rightarrow C$$

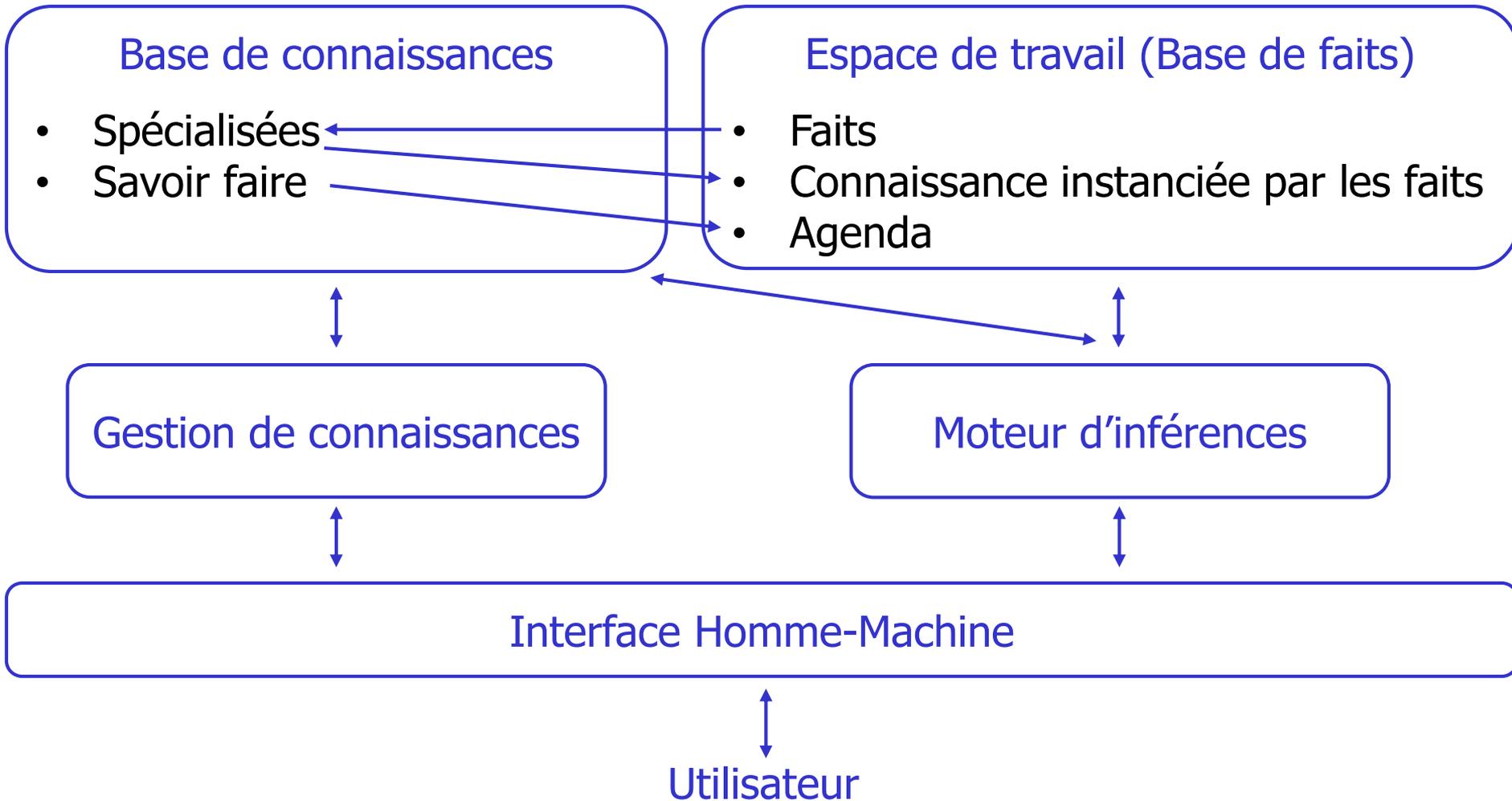
- Si on sait A et B alors C est vrai
- Si on cherche C alors essayer de savoir si A et B sont vrais
- Si A est vrai et C est Faux alors B est faux
- Si C est faux alors soit A soit B est faux

Si le patient a le symptôme de la maladie
alors le patient a la maladie

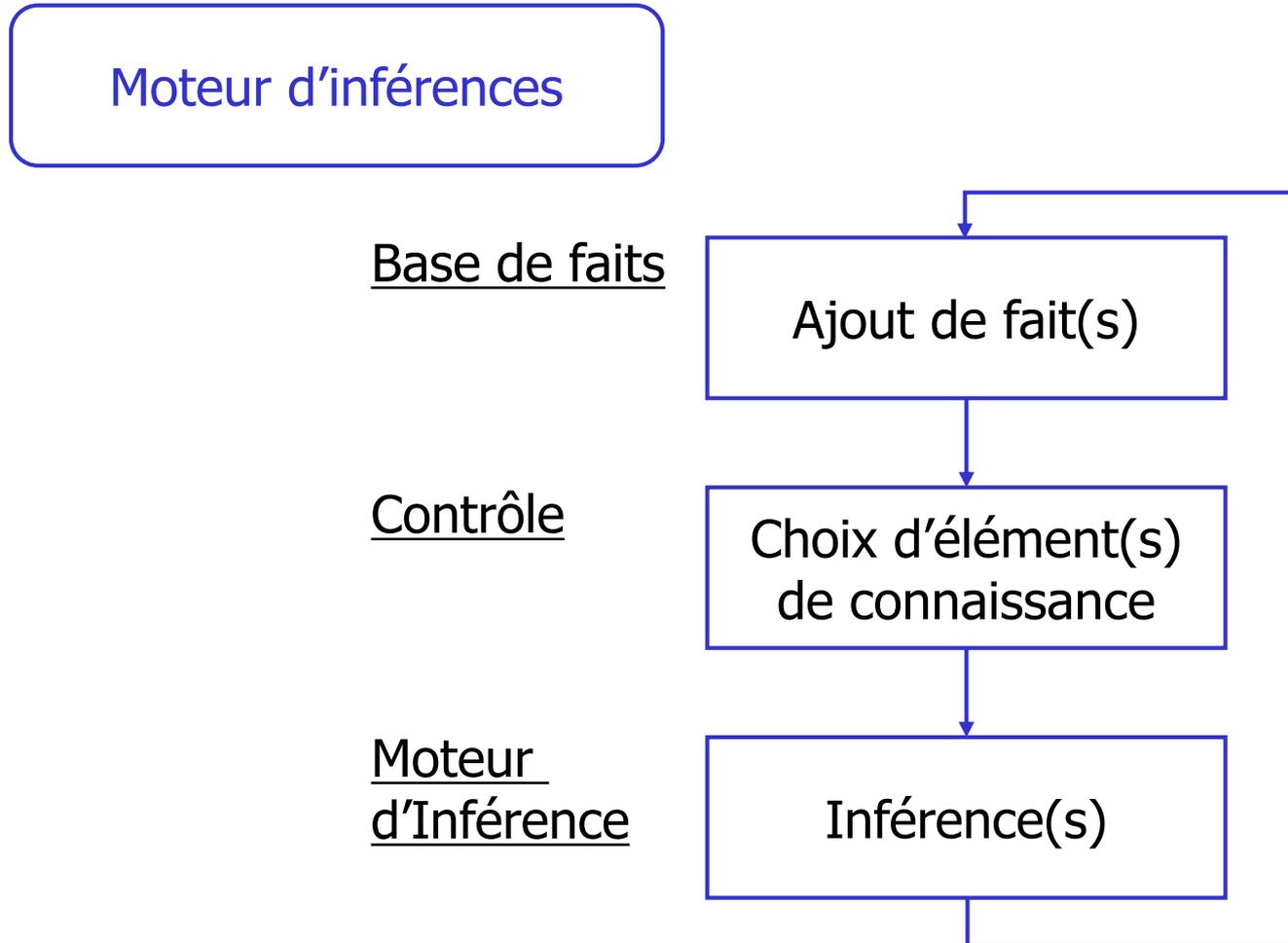
- Connaissance explicite avec des règles pour s'en servir
- Actions justifiées en faisant référence à ces règles
- Base de connaissance pouvant évoluer facilement
- Reposent sur des mécanismes d'inférence simples et robustes
- Règle de production
 - ✓ Si ... alors ...

Systemes Experts

(2)



Source : Pr Marius Fieschi, « Introduction à l'aide à la décision ».



Source : Pr Marius Fieschi, « Introduction à l'aide à la décision ».

Exemple : système MYCIN

(1)

- Système expert de diagnostic de maladies infectieuses du sang et à la prescription de médicaments (1972-73) (Shortliffe Edward H.)
- Moteur d'inférence, base de règles
- Traitement de l'incertitude
 - ✓ Règles affectées de coefficients de vraisemblance (facteurs de crédibilités)
 - ✓ Calculs de probabilités (sens bayésien)

Exemple : système MYCIN

(2)

- Exemple de règle de connaissance

Si

Coloration GRAM nég
et Morphologie : Bâtonnet
et Patient : hôte à risque

<prémisse-s>

Alors

On suggère : Pseudomonas

<conclusion>
<action>

- Généralisation sous forme de règle de méta connaissance

- Traitement de l'incertitude : facteur de crédibilité (FC)

$$A \rightarrow B \text{ (FC)}$$

- ✓ FC attaché à la règle
- ✓ Varie entre -1 et +1
- ✓ Les propositions (\bar{A}, \dots) sont également affectées d'un FC fourni par l'utilisateur ou déduit par des FC attachés à des règles
- ✓ Dans la base de connaissance (fournis par l'utilisateur)

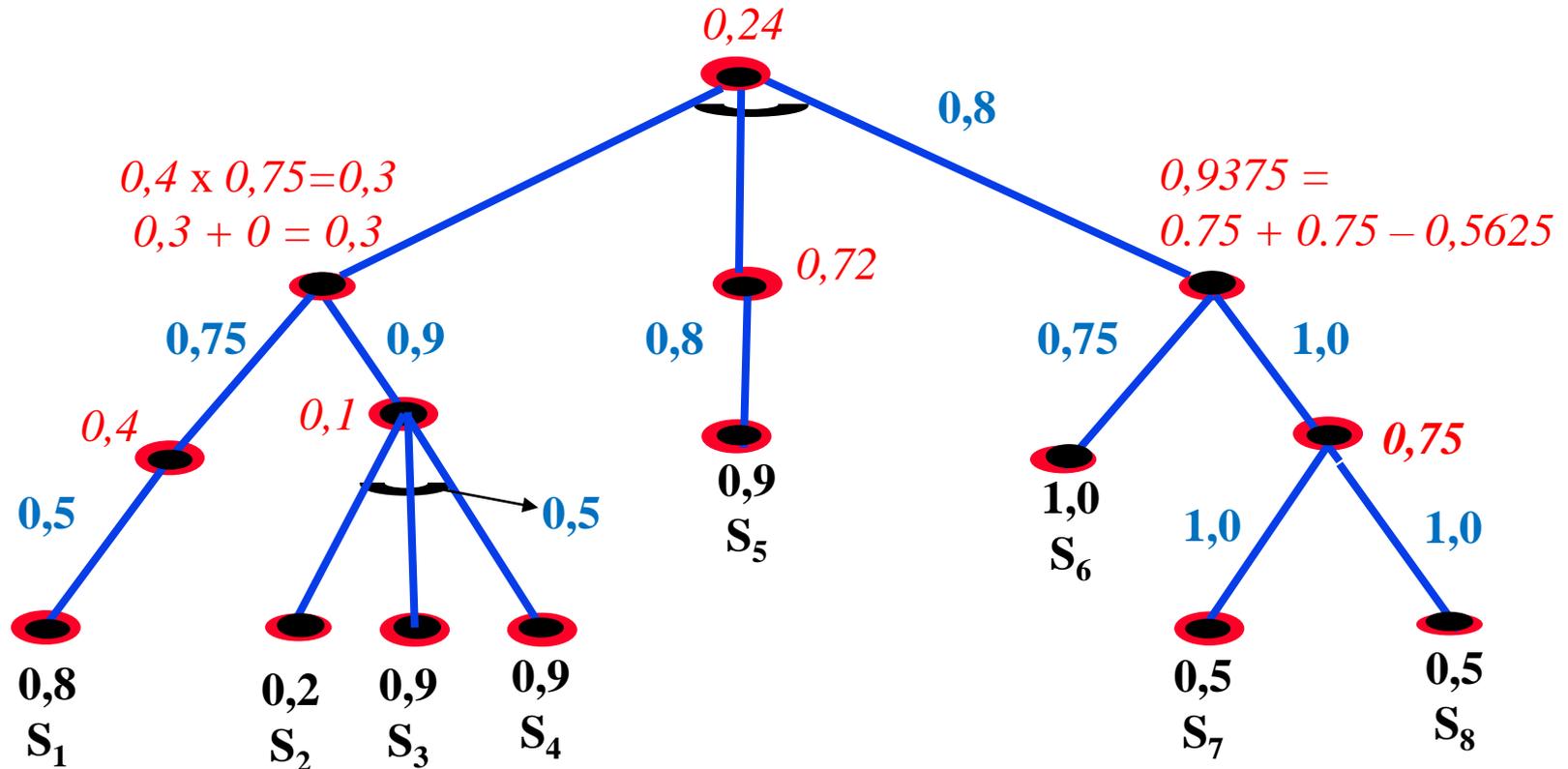
- Crédibilité des prémisses et opérations
 - ✓ $FC(S_1 \text{ ou } S_2) = \text{Max}\{FC(S_1), FC(S_2)\}$
 - ✓ $FC(S_1 \text{ et } S_2) = \text{Min}\{FC(S_1), FC(S_2)\}$
 - ✓ $FC(\neg S) = - FC(S)$
- Règles calculatoires entre les FC des règles et des prémisses (FC calculés)
- Règles avec les prémisses, les conclusions/actions, FC fournis/calculés pouvant être représentées sous la forme d'un arbre de décision

Exemple : système MYCIN

(5)

Nœuds ET : $CF(S_1 \wedge S_2) = A \cdot \min(CF(S_1), CF(S_2))$

Nœuds OU : $CF_{1,2}(S) = CF_1(S) + CF_2(S) - CF_1(S) \cdot CF_2(S)$



Les chiffres en bleu sont les FC des règles de la base de connaissance.

Les chiffres en rouge sont les FC calculés.

Les chiffres en noir sont fournis par l'utilisateur.

Source : Pr Marius Fieschi, « Introduction à l'aide à la décision ».

Approche Numérique de l'IA

- Développée grâce à l'augmentation de la puissance de calcul, la massification des données
- Recherche à extraire les connaissances implicitement
- Raisonne sur les données, expérience accumulée
- Recherche des régularités dans les données disponibles pour extraire des connaissances, sans modèle préétabli
- Apprentissage automatique (machine learning)
 - ✓ Améliorer les performances à résoudre des tâches sans être explicitement programmé pour chacune, par essais et erreurs, avant de se débrouiller tout seul

Apprentissage Machine (Automatique)

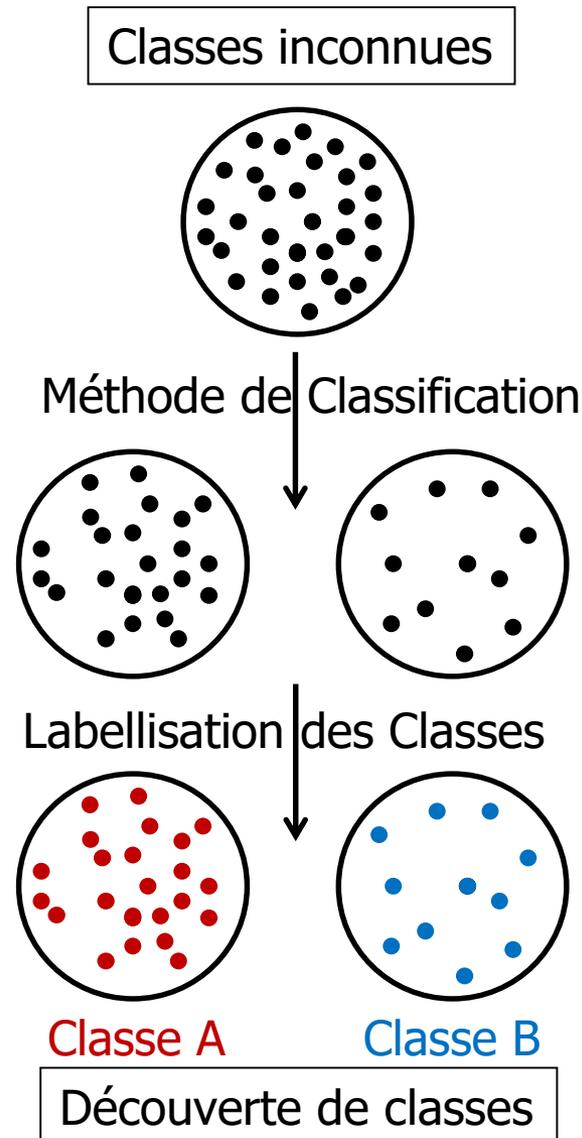
"Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's?"

Alan Turing, 1950

Apprentissage Automatique

- Données nombreuses → difficulté d'étudier toutes les combinaisons possibles
- Besoin d'ajuster un modèle au vue des données
 - ✓ Simplification de la complexité
 - ✓ Opérationnalité
- Apprentissage
 - ✓ Supervisé
 - Nature des données d'entraînement inconnue
 - ✓ Non supervisé
 - Nature des données d'entraînement connue
 - ✓ Par renforcement, semi-supervisé, actif

Apprentissage Non Supervisé



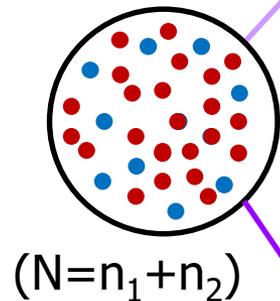
Apprentissage Supervisé-Validation Interne

Echantillon d'apprentissage
($n_1=2/3$)

Définition de la règle de
prédiction/classification

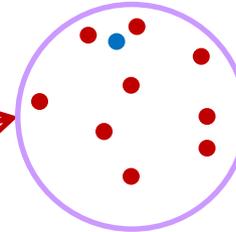
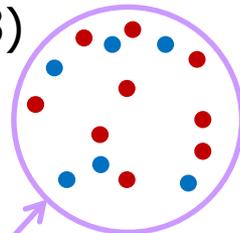
$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \quad \sqrt[n]{a} = \sqrt[n]{a}$$
$$|x| = \begin{cases} -x, & x < 0 \\ x, & x > 0 \end{cases}$$
$$(x-y)(x+y) = x^2 - y^2$$
$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd}$$
$$(x+a)^n = \sum_{k=0}^n \binom{n}{k} x^k a^{n-k}$$

Classes connues

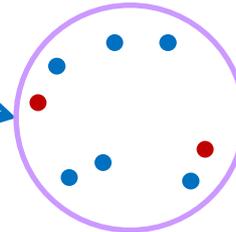


($N=n_1+n_2$)

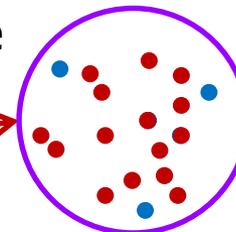
Echantillon de validation
($n_2=1/3$)



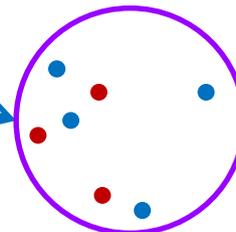
% classification
correcte



Application de la règle de
prédiction/classification

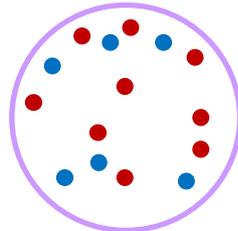


% classification
correcte ?



Apprentissage Supervisé-Validation Externe

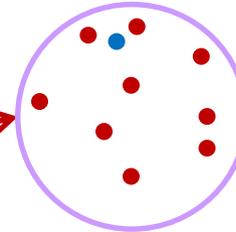
Echantillon d'apprentissage



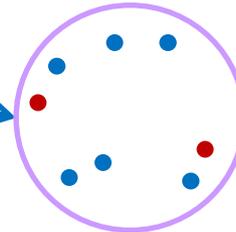
Classes connues

Définition de la règle de prédiction/classification

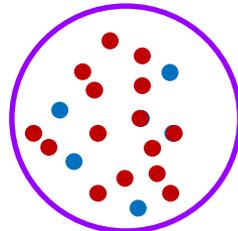
$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \quad \sqrt[n]{a} = \sqrt[n]{a}$$
$$|x| = \begin{cases} -x, & x < 0 \\ x, & x > 0 \end{cases}$$
$$(x-y)(x+y) = x^2 - y^2$$
$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd}$$
$$(x+a)^n = \sum_{k=0}^n \binom{n}{k} x^k a^{n-k}$$



% classification correcte

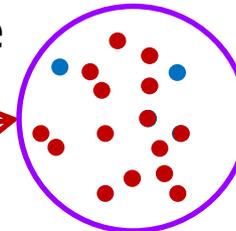


Application de la règle de prédiction/classification

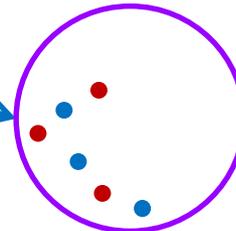


Classes connues

Echantillon externe de validation



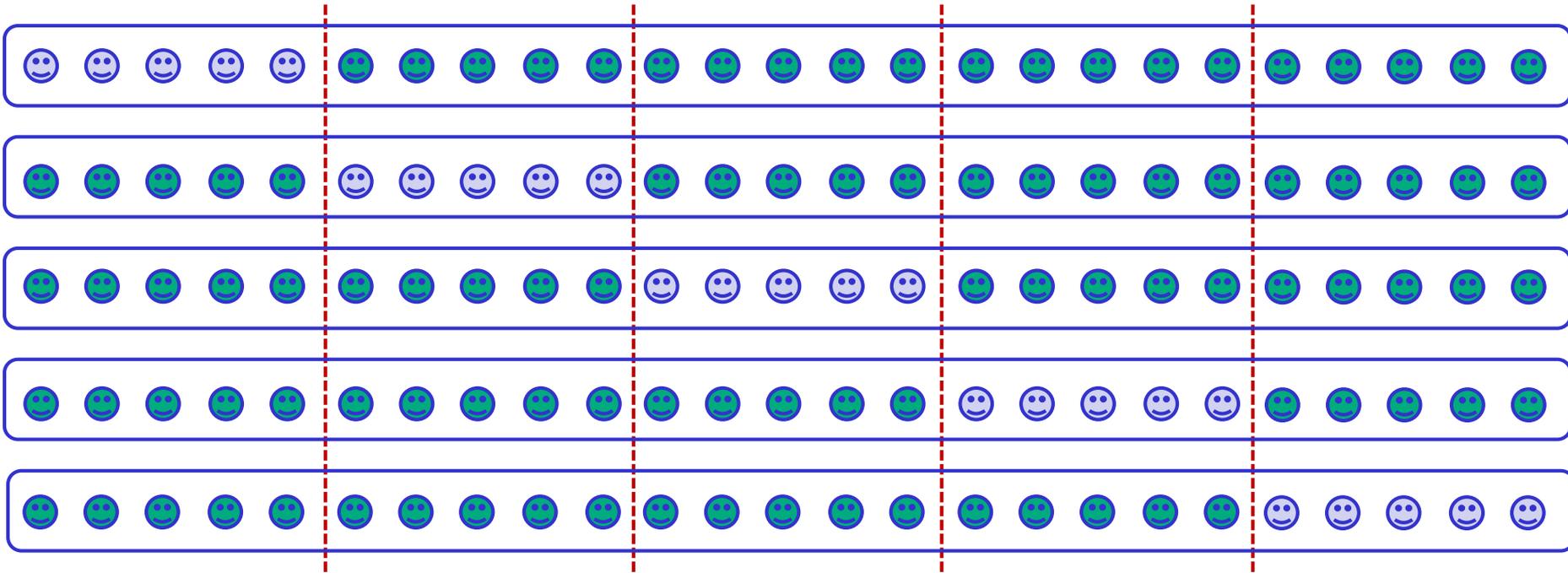
% classification correcte ?



Validation Croisée

- Jeux de données découpé en k parties (k folds)

☺ Validation 🟢 Apprentissage



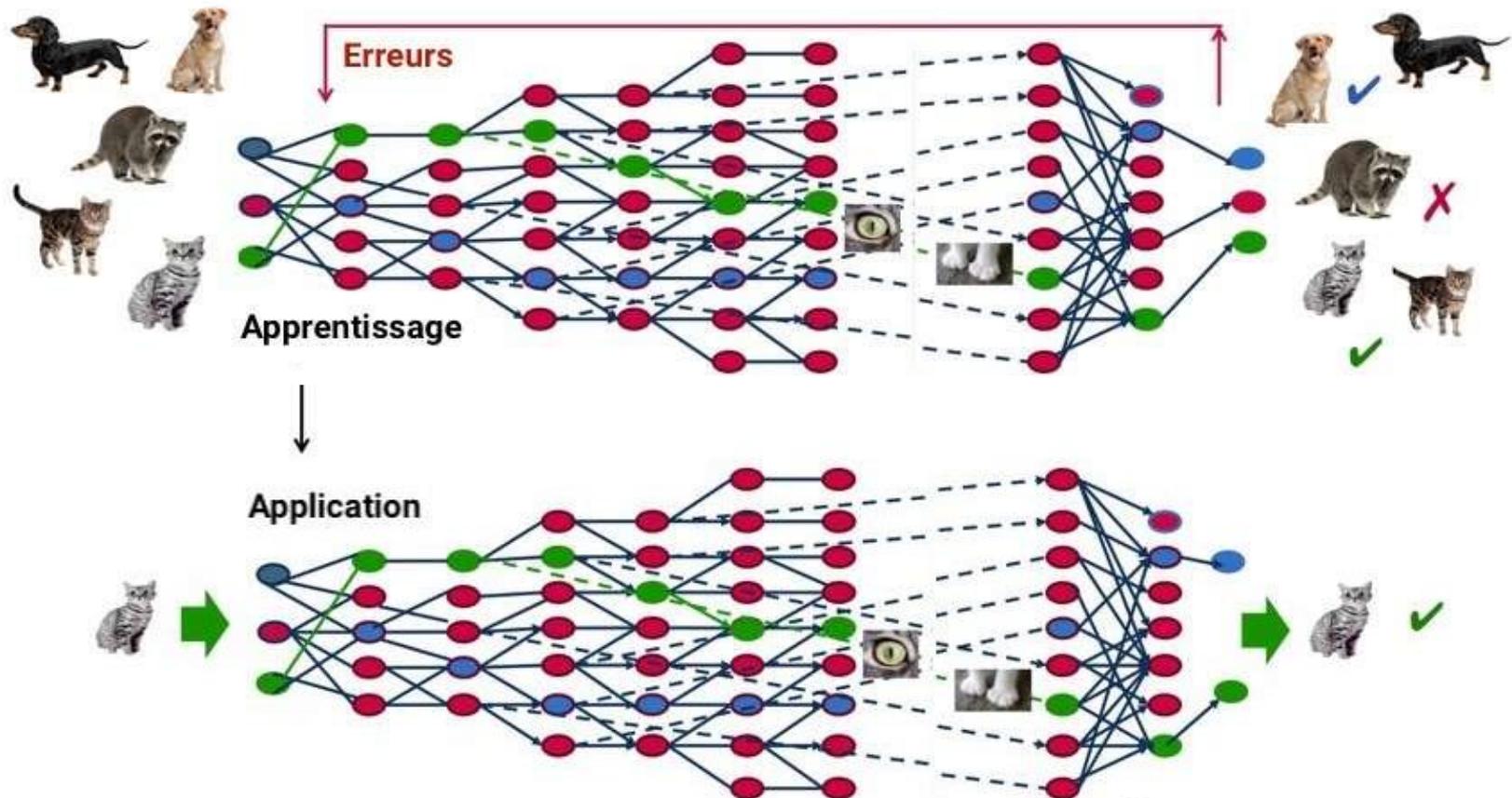
- Performance du modèle de prédiction/classification
 - ✓ Sur l'ensemble des données
 - ✓ En moyennant les performances obtenues sur les k parties, avec erreur type

Algorithmes

- Machines à vecteur de support (support vector machine, SVM)
- Boosting
- Réseaux de neurones
- K plus proches voisins
- Forêts aléatoires
- Modèles de régression linéaire, logistique, discriminante, de mélange
- Algorithmes génétiques
- ...

Exemple Analyse d'Images

- Machine learning
 - ✓ Deep learning – Réseaux de neurones



IA en/pour la Santé ?

- Des résultats contrastés
 - ✓ Des succès
 - ✓ Des échecs
- Performances des outils, techniques et méthodes développés
- Pratiques courantes
 - ✓ Adéquation ?
 - ✓ Déploiement en milieu clinique ?
 - ✓ Impact sur l'organisation des soins cliniques ?

Des domaines d'Applications et des Enjeux (1)

- Médecine prédictive
 - ✓ Prédiction de la survenue d'une maladie, de son évolution
- Médecine de précision
 - ✓ Recommandation de traitement « personnalisé », ciblé
- Aide à la décision
 - ✓ Diagnostique, thérapeutique
- Robots compagnons, domotique
 - ✓ Silver économie

Source : Thiebault R. (2016). Big data en santé. <https://www.inserm.fr/information-en-sante/dossiers-information/big-data-en-sante> (Janvier 2021)

Des domaines d'Applications et des Enjeux

(2)

- Chirurgie assistée par ordinateur
 - ✓ Améliorer la précision des gestes, opération à distance, contrôle en temps réel du déroulement d'une opération
- Prévention en population
 - ✓ Pharmacovigilance
 - ✓ Anticipation d'une épidémie

Source : Thiebault R. (2016). Big data en santé. <https://www.inserm.fr/information-en-sante/dossiers-information/big-data-en-sante> (Janvier 2021)

Des domaines d'Applications et des Enjeux

(3)

- Les données et leur partage
- Apporter une aide réelle pour la santé, dans la pratique médicale
- Confiance dans le résultat, la décision
 - ✓ IA explicable
- Validation de l'algorithme, du dispositif, du logiciel
- Enjeux sociétaux et éthiques