BIG DATA ET SCIENCE DE DONNÉES

INTRODUCTION A LA SCIENCE DE DONNEES

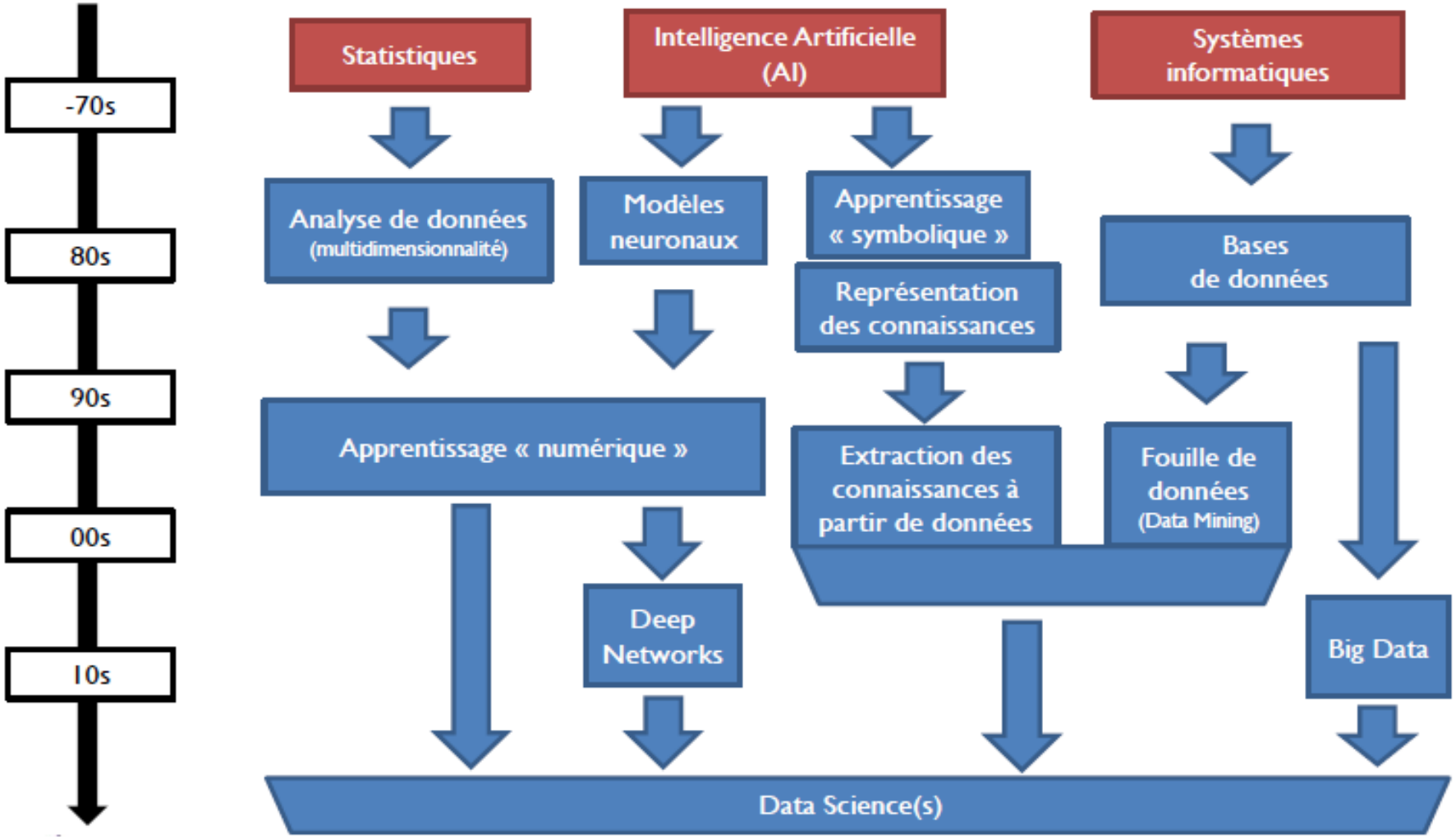
Master 1 Intelligence Artificielle

Université de M'sila, Département d'Informatique

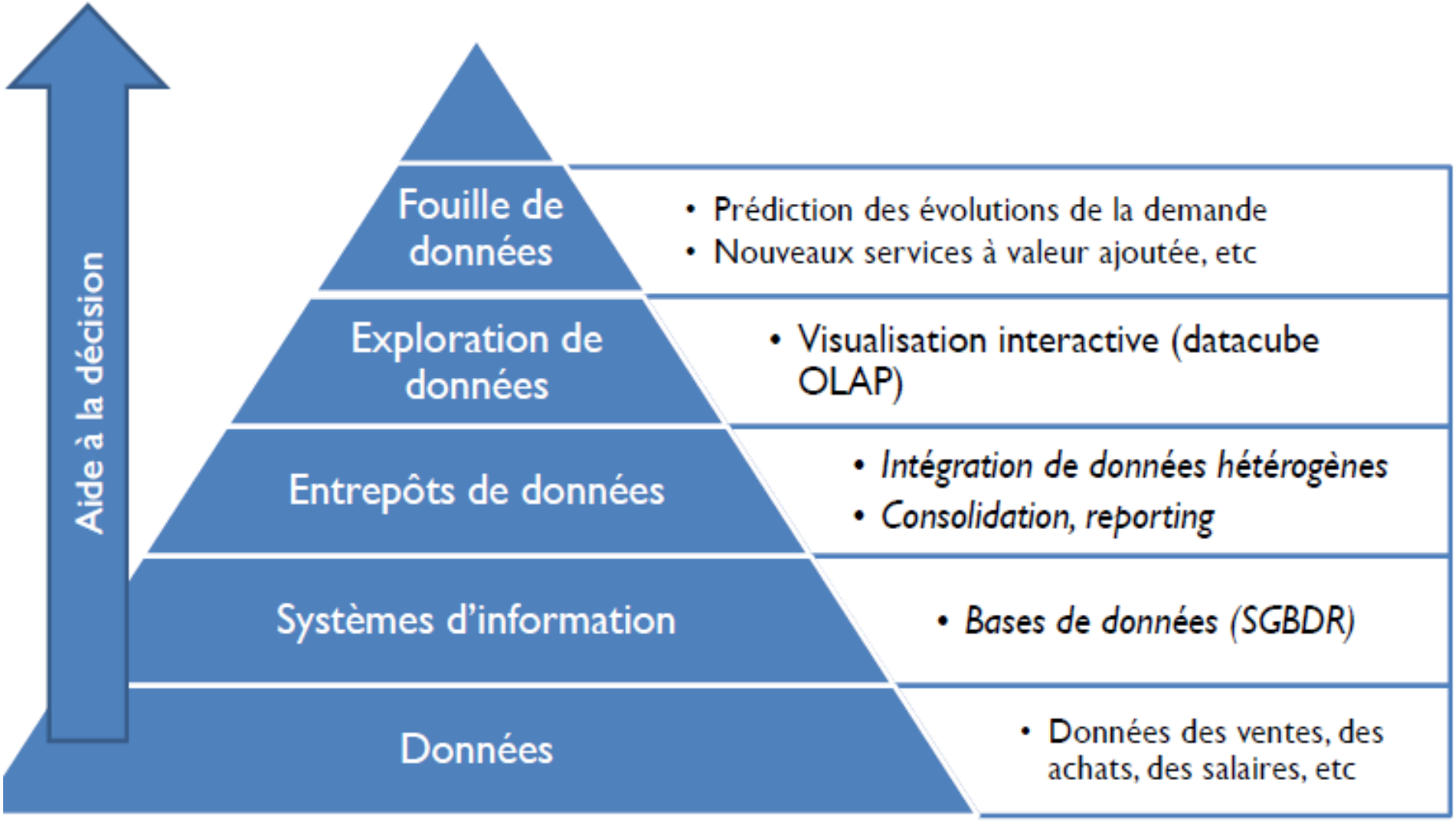
Dr Mehenni Tahar

2020-2021

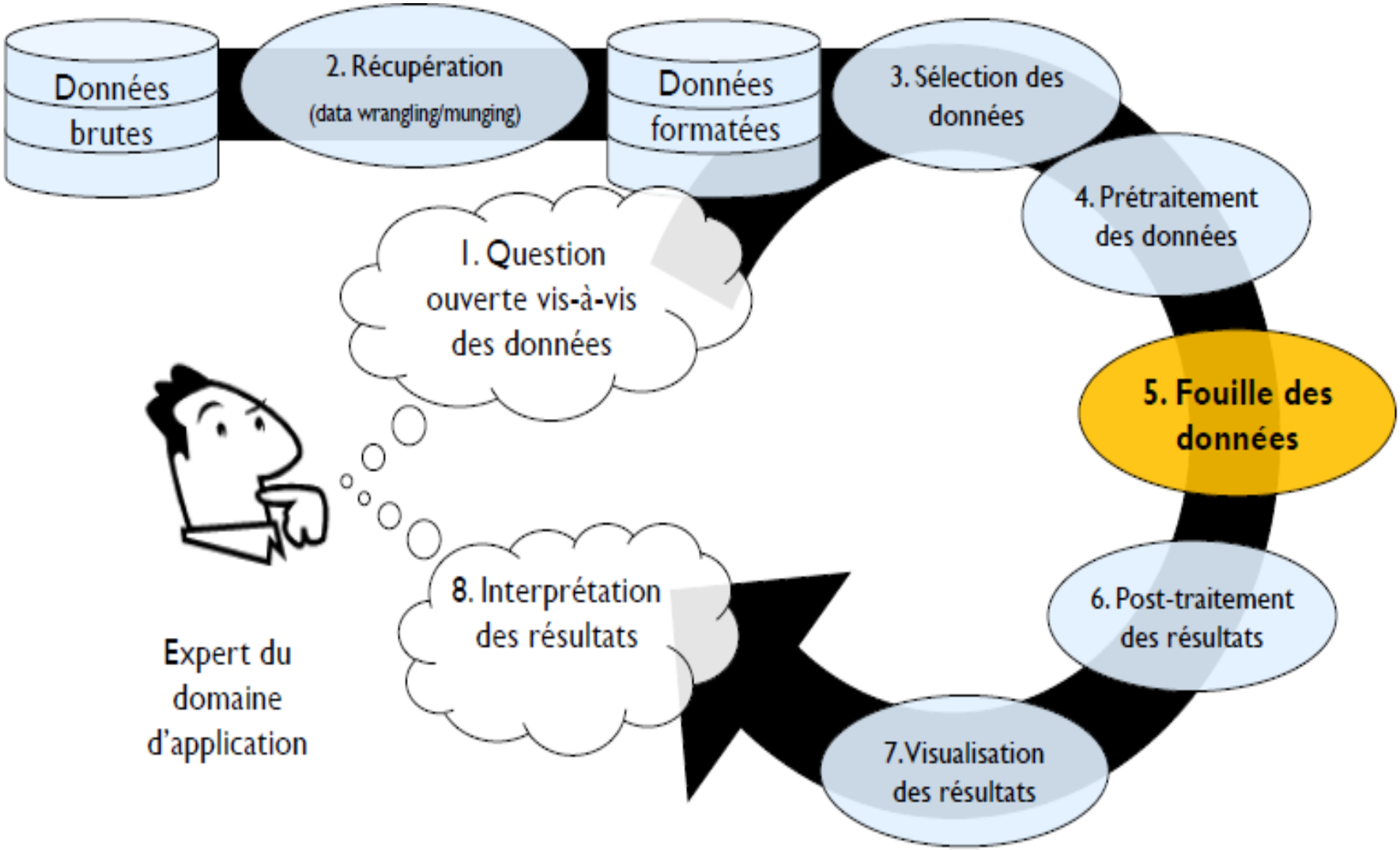
Les origines scientifiques de la science des données



L'aide à la décision etles bases de données décisionnelles (data warehouse)



Le processus d'extraction de connaissances (KDD: Knowledge Discovery from Data



Qu'est-ce que le machine learning ?

- Arthur Samuel, 1959: « le *machine learning est* le champ d'étude visant à donner la capacité à une machine d'apprendre sans être explicitement programmée ».
- Tom Mitchell, 1997: *« A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E ».*

Cas d'application	Checkers learning	Handwriting recognition	Robot driving learning
Tasks T	Playing checkers	Recognizing and classifying handwritten words within images	Driving on public four-lane highways using vision sensors
Performance measure P	Percent of games won against opponents	Percent of words correctly classified	Average distance traveled before an error (as judged by human overseer)
Training experience E	Playing practice games againts itself	A database of handwritten words with given classifications	A sequence of images and steering commands recorded while observing a human driver

Algorithmes supervisés et non supervisés

- **Les algorithmes supervisés:** Extraient de la connaissance à partir d'un ensemble de données contenant des couples entrée-sortie.
  - Ces couples sont déjà « connus », dans le sens où les sorties sont définies a priori.
  - La valeur de sortie peut être une indication fournie par un expert : par exemple, des valeurs de vérité de type OUI/NON ou MALADE/SAIN.
  - Ces algorithmes cherchent à définir une représentation compacte des associations entrée-sortie, par l'intermédiaire d'une fonction de prédiction.
- **les algorithmes non supervisés:** n'intègrent pas la notion d'entrée-sortie.
  - Toutes les données sont équivalentes (on pourrait dire qu'il n'y a que des entrées).
  - les algorithmes cherchent à organiser les données en groupes.
  - Chaque groupe doit comprendre des données similaires et les données différentes doivent se retrouver dans des groupes distincts.
  - L'apprentissage ne se fait plus à partir d'une indication qui peut être préalablement fournie par un expert, mais uniquement à partir des fluctuations observables dans les données.

Algorithmes de régression et de classification

- La distinction régression/classification se fait au sujet des algorithmes supervisés.
- **Dans un problème de régression**: *Y* peut prendre une infinité de valeurs dans l'ensemble continu des réels (noté $Y \in$). Ce peut être des températures, des tailles, des PIB, des taux de chômage, ou tout autre type de mesure n'ayant pas de valeurs finies a priori.
- **Dans un problème de classification**: *Y* prend un nombre fini *k* de valeurs ($Y = \{1, \ldots, k\}$). On parle alors d'étiquettes attribuées aux valeurs d'entrée. C'est le cas des valeurs de vérité de type OUI/NON ou MALADE/SAIN évoqués précédemment.

Algorithmes du Data Scientist

Algorithme	Mode d'apprentissage	Type de problème à traiter
Régression linéaire univariée	Supervisé	Régression
Régression linéaire multivariée	Supervisé	Régression
Régression polynomiale	Supervisé	Régression
Régression régularisée	Supervisé	Régression
Naive Bayes	Supervisé	Classification
Régression logistique	Supervisé	Classification
Clustering hiérarchique	Non supervisé	-
Clustering non hiérarchique	Non supervisé	-
Arbres de décision	Supervisé	Régression ou classification
Random forest	Supervisé	Régression ou classification
Gradient boosting	Supervisé	Régression ou classification
Support Vector Machine	Supervisé	Régression ou classification
Analyse en composantes principales	Non supervisé	-

La validation croisée

- Créer un échantillon d'entraînement, sur lequel on va constituer le modèle, et un échantillon de test, sur lequel on va tester le modèle.
- le *data scientist* choisit généralement de diviser ses données en trois :
  - un jeu d'entraînement, bien sûr;
  - un jeu dit de validation : celui-ci va être utilisé pour tester les différents modèles paramétrés sur le jeu d'e*ntraînement*
  - et un vrai jeu de test, qu'on garde de côté et qui ne sera utilisé que tout à la fin du processus de modélisation, afin de tester le plus honnêtement possible la capacité de généralisation du modèle retenu.

Principales Méthodes de validation croisée

- **La méthode LOOV** (*leave-one-out cross-validation*): Sortir une observation *i* de l'ensemble du jeu de données et à calculer le modèle sur les *m*-1 données restantes. On utilise ce modèle pour prédire *i* et on calcule l'erreur de prévision. On répète ce processus pour toutes les valeurs de $i = 1, \ldots, m$. Les *m* erreurs de prévision peuvent alors être utilisés pour évaluer la performance du modèle en validation croisée.
- **La méthode LKOV** (*leave-k-out cross-validation*): même principe que LOOV, sauf que l'on sort non pas une, mais *k* observations à prédire à chaque étape (donc LOOV est équivalent à LKOV pour $k = 1$). Le processus est répété de façon à avoir réalisé tous les découpages possibles en données de modélisation/de prévision.
- **La méthode *k-fold cross-validation:*** les données sont aléatoirement divisées en *k* sous-échantillons de tailles égales, dont l'un est utilisé pour la prévision et les *k*-1 restants pour l'estimation du modèle. Le processus n'est répété que *k* fois.

Choix de la métrique de performance

- **Pour les problèmes de régression**
  - Nombreuses sont les mesures disponibles pour évaluer la qualité d'un modèle de régression.

• l'erreur moyenne absolue (MAE, *Mean Absolute Error*) :

$$\frac{1}{m}\sum_{i=1}^{m}|y_i-\hat{y}_i|$$

• la racine carrée de la moyenne du carré des erreurs (RMSE, *Root Mean Squared Error*) :

$$\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i-\hat{y}_i\right)^2}$$

Choix de la métrique de performance

- **Pour les problèmes de classification**
  - L'évaluation d'un problème de classification se base sur une matrice de confusion, qui met en regard des données prédites et des données observées

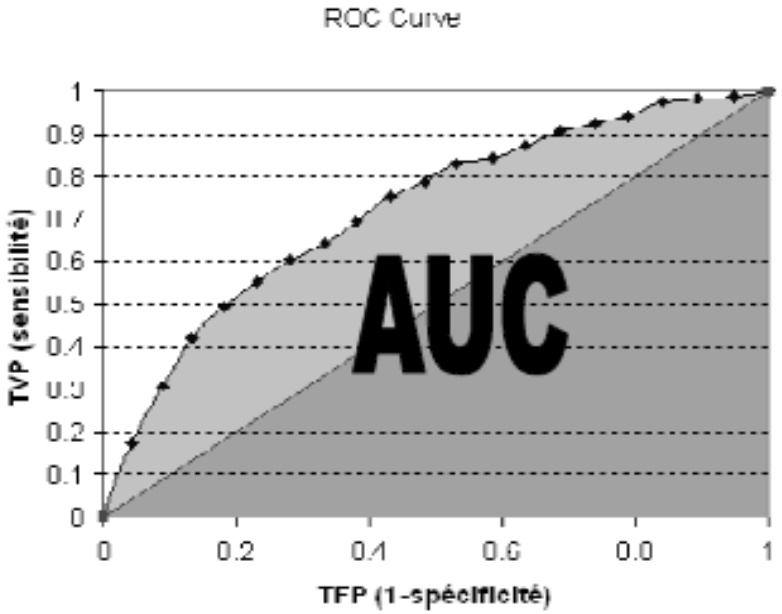
		Observations		
		+	-	Total
Prédictions	+	Vrais positifs (VP)	Faux positifs (FP)	Total des positifs prédits (VP + FP)
	-	Faux négatifs (FN)	Vrais négatifs (VN)	Total des négatifs prédits (FN + VN)
	Total	Total des vrais positifs observés (VP + FN)	Total des vrais négatifs observés (FP + VN)	Taille totale de l'échantillon (N)

- **Recall:** VP/(VP + FN)
- **Précision:** VP/(VP + FP)
- **F1 score:** 2*(Recall*Precision)/(Recall+Precision)

La courbe ROC

- ROC signifie *Receiver Operating Characteristic.*
- La courbe ROC est tracée dans un espace de deux dimensions définies par *α (*taux de Vrais Positifs VP) en ordonnée et 1-*β (β* taux de Vrais Négatifs VN) en abscisse.

- À (0, 0) le classificateur déclare toujours 'négatif'
- À (1, 1) le classificateur déclare toujours 'positif'
- Un classificateur aléatoire tracera une droite allant de (0, 0) à (1, 1).
- À (0, 1) le classificateur n'a aucun faux positif ni aucun faux négatif, et est par conséquent parfaitement exact, ne se trompant jamais.
- À (1, 0) le classificateur n'a aucun vrai négatif ni aucun vrai positif.



- AUC indique la probabilité pour que la fonction SCORE place un positif devant un négatif (dans le meilleur des cas AUC = 1).
- Si SCORE classe au hasard les individus (c.-à-d. le modèle de prédiction ne sert à rien), AUC = 0.5 (droite en diagonal: au-dessus: meilleur que l'aléatoire, au-dessous: moins bine que l'aléatoire)

Les espaces de grande dimension

- La dimension d'un problème de *machine learning* correspond au nombre *n* de variables de la matrice *X*.
- Un *n* grand peut poser de nombreux pièges que le *data scientist* doit savoir surmonter.
- En général, on cherche à sélectionner judicieusement un sous-espace pertinent, pour améliorer la modélisation.
- L'objectif est de conserver le maximum de l'information contenue dans les données, avec un minimum de variables.
- Deux approches sont envisageables:
  - Sélectionner un nombre restreint des variables les plus importantes (Exemple: feature selection).
  - Créer des variables « synthétiques » à partir des variables initiales, ensuite reconstruire un nouvel espace de dimension réduite (Exemple: Analyse en Composantes Principales).

Les valeurs manquantes

- **Deux grandes catégories de données manquantes:**
  - Lorsque l'on ne dispose d'aucune information sur un individu (c'est-à-dire une ligne complètement vide dans une matrice).
  - Lorsque l'on dispose d'une information incomplète sur un individu (une ligne partiellement renseignée).
- **Traitement des valeurs manquantes:**
  - Imputation par règle : si on la connaît, on peut définir une règle métier qui permet de calculer la valeur manquante à partir des autres données disponibles ;
  - Replacement des valeurs manquantes par la valeur moyenne de l'ensemble des réponses
  - Imputation par régression.
  - Méthode du plus proche voisin