Module 4 : Apprentissage Machine Non Supervisé et Extraction de Connaissances

Analyse de Données

Nicolas PASQUIER
Université Côte d'Azur
http://www.i3s.unice.fr/~pasquier

Co-financé par :

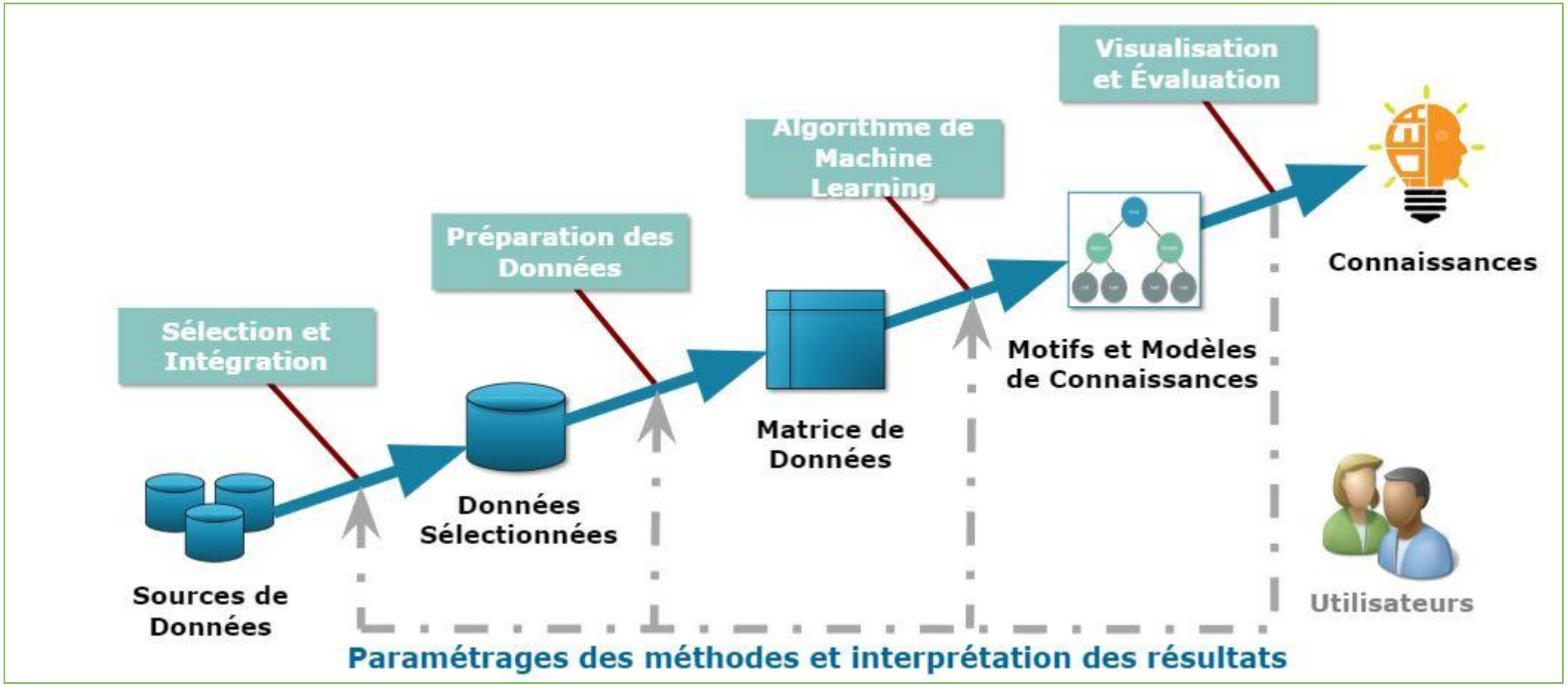
Use cases réalisés par les masters :

Définition de l'Extraction de Connaissances pour l'Analyse de Données

- Objectif : extraire des motifs et modèles de connaissances à partir de très grands ensembles de données.
- Difficultés algorithmiques
  - Taille des données : volume impose leur stockage en mémoire(s) secondaire(s) dont les temps d'accès (≈ms) nécessitent de minimiser le nombre de lectures.
  - Taille de l'espace de recherche : nombre de motifs ou modèles de connaissances potentiellement valides exponentiel dans la taille des données.
- Extraire des connaissances à partir des données pour :
  - Identifier des relations entre les données (e.g. liens entre valeurs des variables).
  - Comprendre les structures de l'espace des données (e.g. groupes d'instances similaires).
  - Apprendre des modèles prédictifs (e.g. prédictions de valeurs de variables).
- Différents types de motifs et modèles, différentes représentations
  - Règles, partitionnements, séquences de valeurs, fonctions, etc.
  - Indicateurs statistiques de pertinence (précision) et de portée (importance) associés à chaque motif ou modèle.

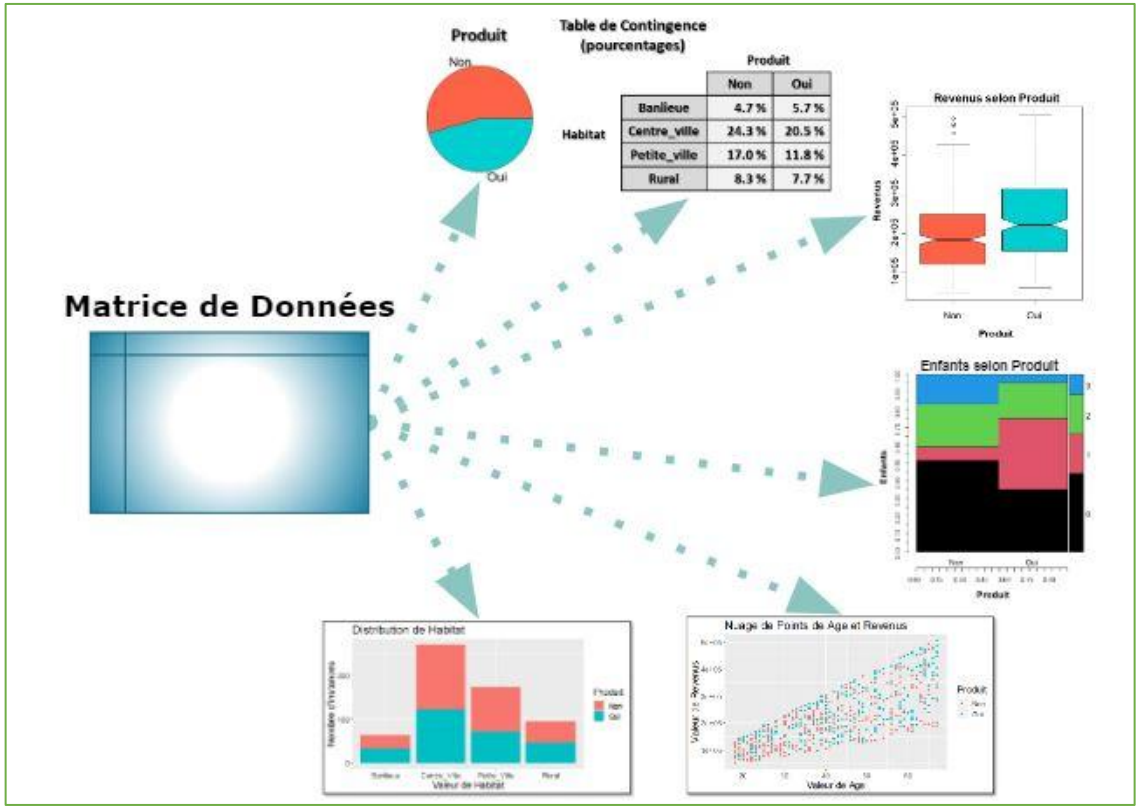
Processus d'Extraction de Connaissances à Partir des Données

- Processus interactif et itératif



Matrice de Données en Entrée des Algorithmes

- Chaque ligne est une instance (exemple, observation, tuple)
  - Exemples : un client, un film.
  - Définie par une valeur pour chaque colonne.
- Chaque colonne est une variable (attribut, dimension, champ)
  - Propriété ou caractéristique des instances .
  - Exemples : âge du client, durée du film.
- Dictionnaire des données
  - Type, taille, signification et domaine de valeurs de chaque variable.
- Première étape : vérifier la qualité des données
  - Valeurs erronées, manquantes, mal codées.
  - Outils : visualisations, comptages, statistiques descriptives.



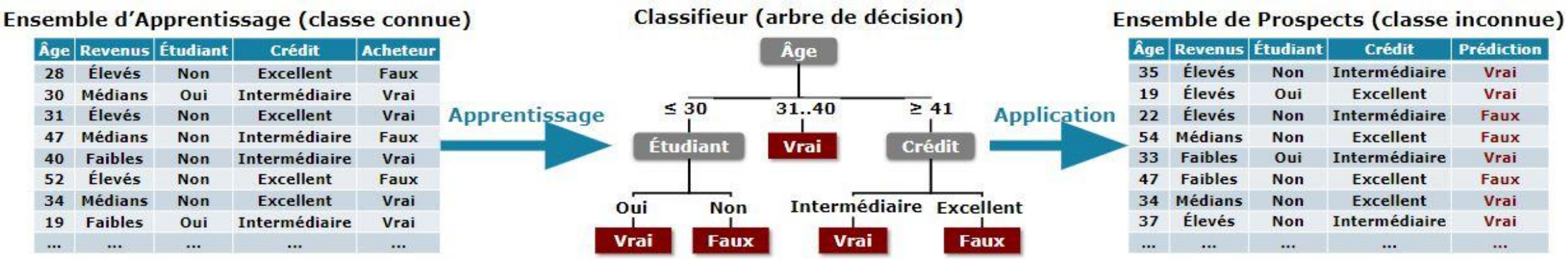
Types de Motifs et Modèles de Connaissances

- **Modèles prédictifs**
  - Objectif : apprendre un modèle de prédiction de la valeur d'une variable à partir des exemples décrivant les expériences passées.
  - **Apprentissage supervisé :**
    - Les variables n'ont pas toutes le même rôle : la variable dont la valeur est à prédire est la **variable cible**, les variables testées pour sa prédiction sont les **variables prédictives**.
  - Classification : prédiction de la valeur d'une variable catégorielle, i.e. non numérique.
  - Régression : prédiction de la valeur d'une variable numérique.

- **Motifs descriptifs**
  - Objectif : extraire des connaissances sur les données pour comprendre leurs relations entre elles
  - **Apprentissage non-supervisé :**
    - Toutes les variables ont le même rôle, i.e. sont traitées de la même manière par l'algorithme.
  - Analyse de liens : relations entre valeurs des variables (colonnes) de la matrice.
  - Analyse de similarités : relations de similarités entre instances (lignes) de la matrice.

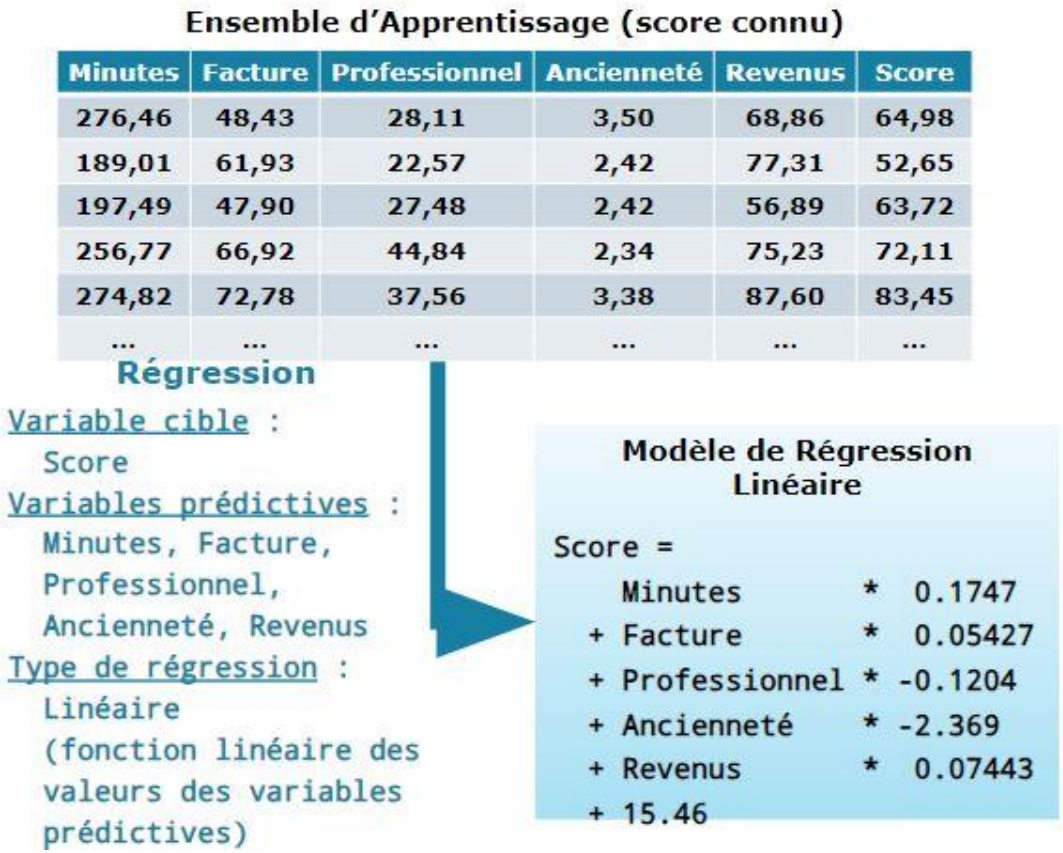
Apprentissage Supervisé : Classification

- Objectif : apprendre un modèle de prédiction de la valeur d'une **variable catégorielle** (i.e. discrète) en fonction des valeurs des autres variables pour l'appliquer ensuite à de nouvelles données
  - Apprentissage du **classifieur** (modèle de prédiction) à partir de l'ensemble d'apprentissage constitué d'instances dont la **classe** (valeur de la variable cible) est connue.
  - Application du classifieur pour prédire la classe de nouvelles instances de classe inconnue.
- Différents types de classifieurs
  - Arbres de décision, réseaux de neurones, forêts d'arbres de décision aléatoires, machines à vecteurs de support, méthodes bayésiennes, apprentissage par analogie.



Apprentissage Supervisé : Régression

- Objectif : apprendre un modèle de prédiction de la valeur d'une **variable numérique continue** en fonction des valeurs des autres variables.
- Estimer la relation entre :
  - Une variable dépendante, appelée **cible**.
  - Les variables explicatives, appelées **prédictives**.
- Différents types de régression :
  - Linéaire simple ou multiple, polynomiale, logistique, non paramétrique, etc.
  - Défini la nature de la fonction résultat.
  - Type adéquat pour l'application dépend des propriété de la matrice de données.



Analyse de Liens : Algorithme Apriori

- **Motifs Fréquents :** motifs de connaissances décrivant les cooccurrences fréquentes de valeurs de variables parmi les instances de la matrice.
  a. **Itemsets Fréquents :** concurrences fréquentes de valeurs de variables (items).
  b. **Règles d'Association :** règles d'implication conditionnelles fréquentes entre valeurs de variables.

Paramètre : *minsupport* = 50%

Paramètres : *minconfiance* = 75%
taille_conséquence = 1

Transactions de Ventes

TID	Articles achetés
1	lait, céréales, thé
2	lait, café, céréales, sucre
3	café, céréales, sucre
4	café, sucre
5	lait, café, céréales, sucre
6	café, céréales, sucre

Itemsets Fréquents

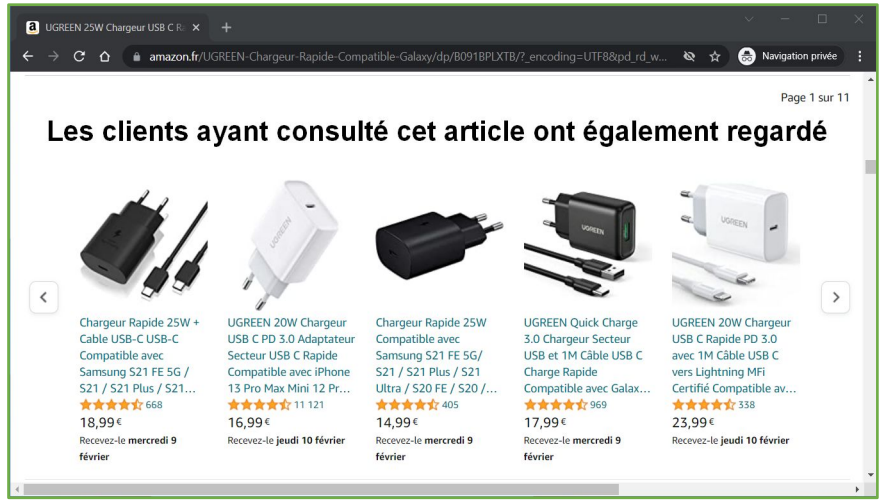
Itemset	TID	Support
café, sucre	2,3,4,5,6	83% (5)
café, céréales	2,3,5,6	66% (4)
céréales, sucre	2,3,5,6	66% (4)
café, céréales, sucre	2,3,5,6	66% (4)
lait, céréales	1,2,5	50% (3)

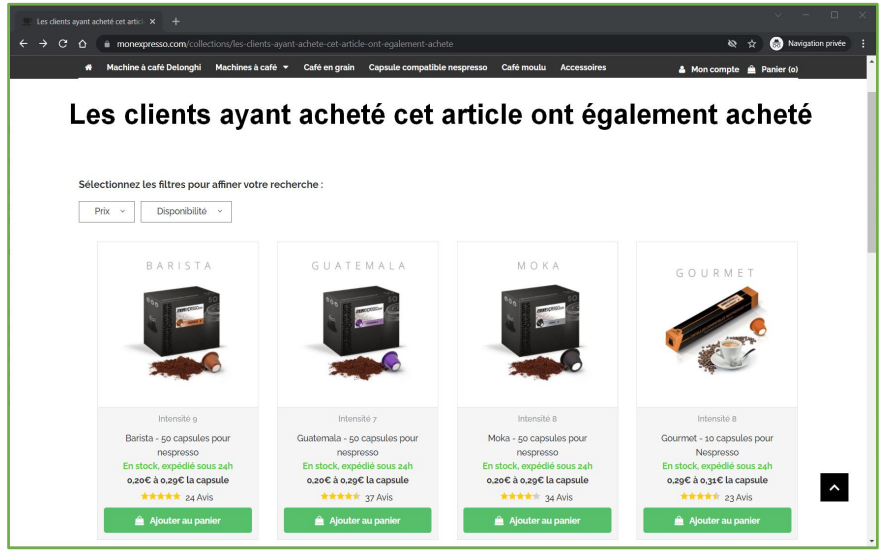
Règles d'Association

Antécédent		Conséquence	Support	Confiance	Lift
café	→	sucre	83% (5)	100%	1,20
sucre	→	café	83% (5)	100%	1,20
café	→	céréales	66% (4)	80%	0,96
céréales	→	café	66% (4)	80%	0,96
sucre	→	céréales	66% (4)	80%	0,96
céréales	→	sucre	66% (4)	80%	0,96
café, sucre	→	céréales	66% (4)	80%	0,96
café, céréales	→	sucre	66% (4)	100%	1,20
céréales, sucre	→	café	66% (4)	100%	1,20
lait	→	céréales	50% (3)	100%	1,20

Exemples d’Applications d’Analyse de Liens

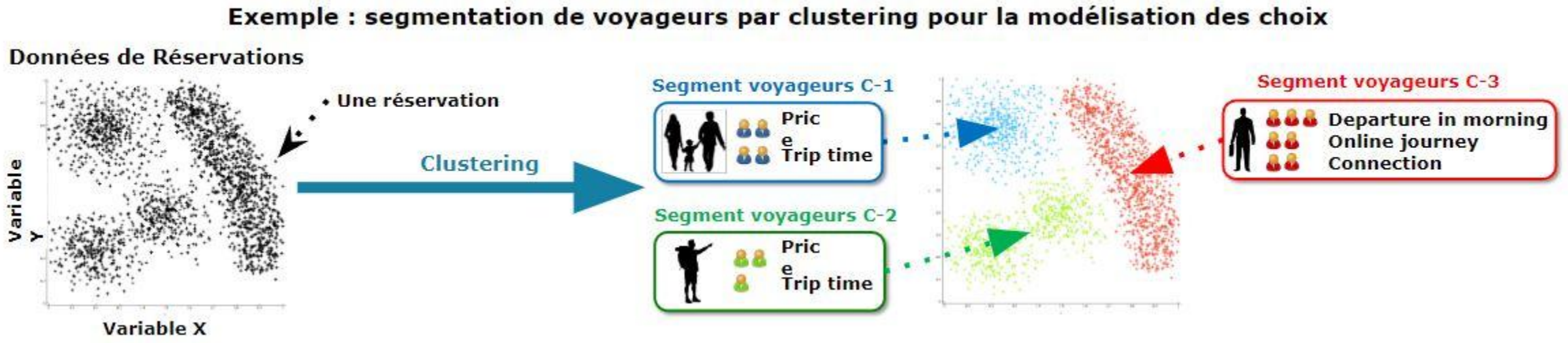
- Analyse des historiques d’activités sur les sites Internet
- Données sources
  - Utilisateurs identifiés : historique des consultations
  - Utilisateurs anonymes : articles consultés durant la session
- Exemple d’**itemset fréquent** (combinaison d’articles consultés fréquemment)
  - {chargeur USB-C 100W, chargeur USB-C 60W, câble USB-C 5A}
- Analyse des historiques d’achats sur les sites Internet
  - Données sources
  - Utilisateurs identifiés : historique des achats
  - Utilisateurs anonymes : articles achetés simultanément
- Exemple de règle d’association (liens conditionnels entre achats)
  - Achat[Nespresso Colombia] → Achat[Nespresso Costa Rica]





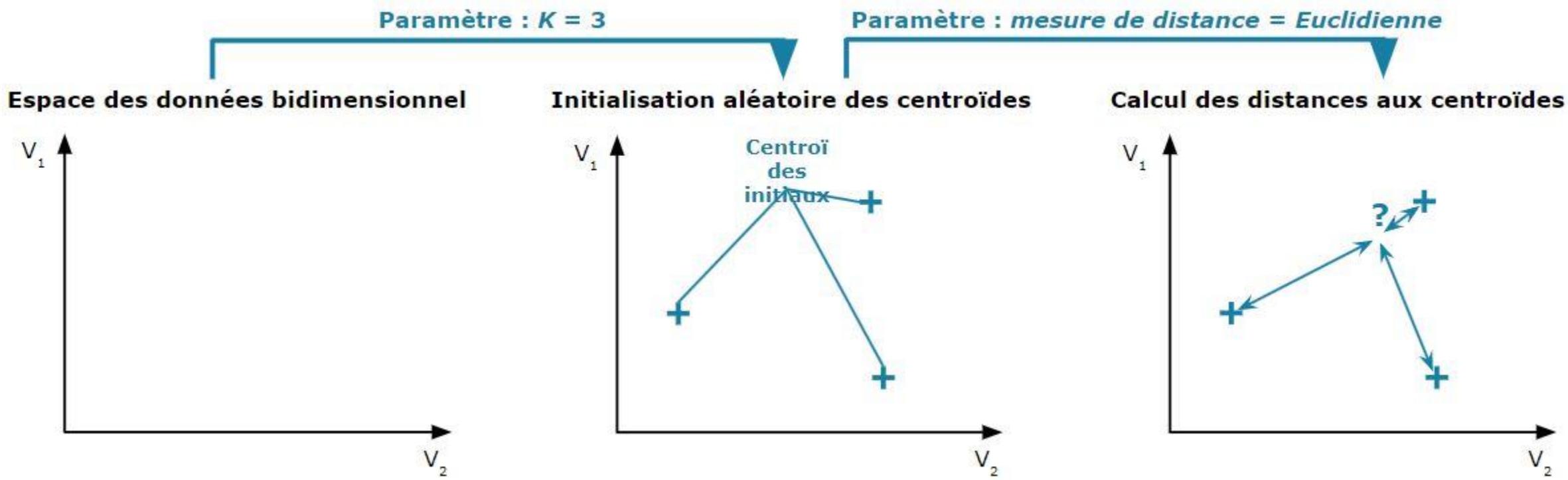
Analyse de Similarités : Clustering et Détection d'Exceptions

- **Similarité** entre deux instances :
  - Évaluée par comparaison des valeurs des variables pour les deux instances.
  - Mesure de distance entre les instances (points) dans l'espace des données .
- **Clustering** : regroupement des instances similaires en clusters (groupes).
- **Détection d'exceptions** : identification des instances hors normes, aux propriétés inhabituelles.
  - Instances dont la distance aux autres points dans l'espace des données est importante (i.e. isolées).



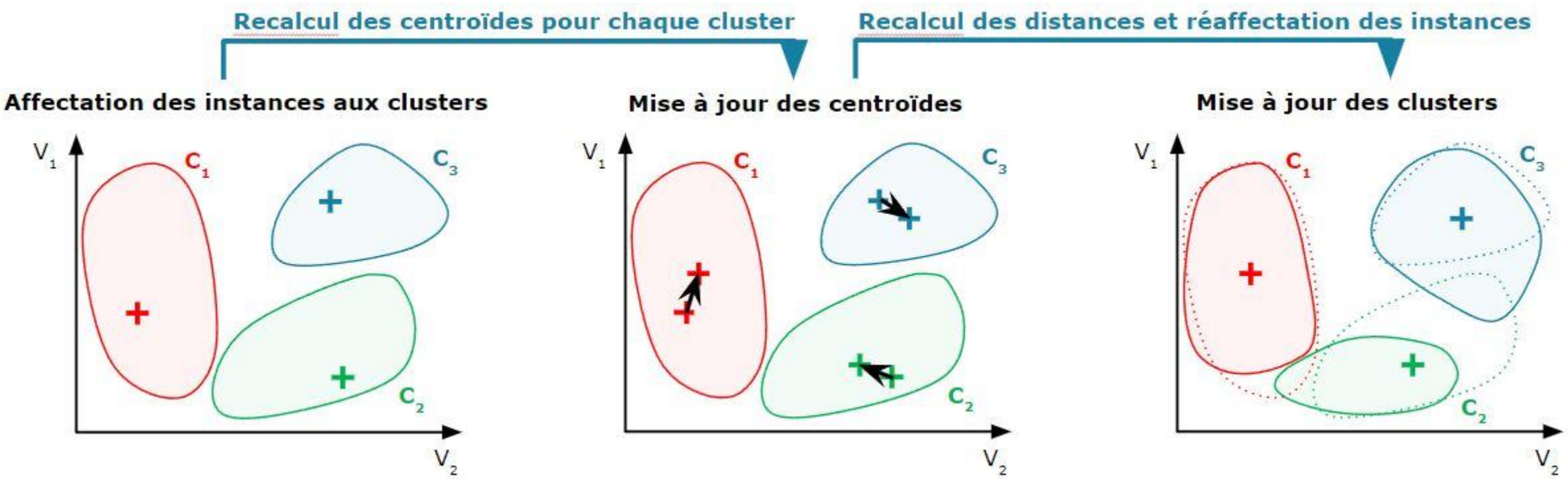
Clustering : Algorithme des K-means

- Objectif : matrice de données bidimensionnelle à partitionner en 3 clusters (paramètre K = 3).
- Dimensions : variables V_1 et V_2 .



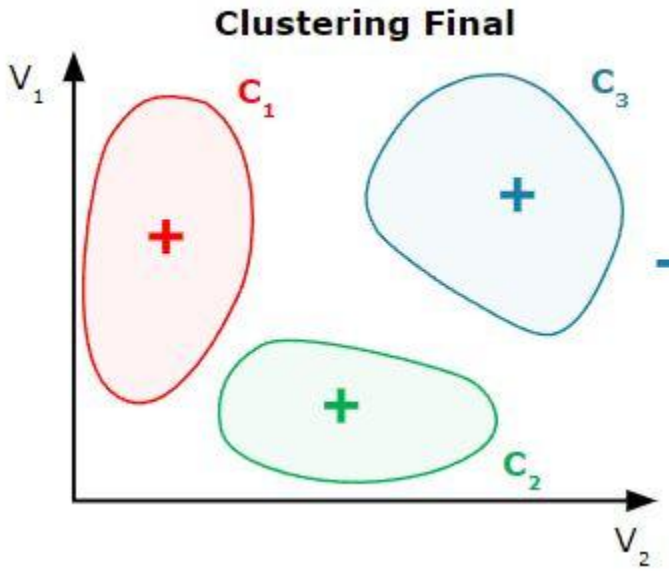
Algorithme des K-means : Itérations

- Itérations : re calcul des centroïdes (moyennes des variables par cluster), recalcul des distances des instances aux centroïdes et réaffectation des instances au centroïde le plus proche si nécessaire.



Algorithme des K-means : Résultat

- Arrêt des itérations lorsque la stabilité est atteinte : pas de réaffectation d'instance.
- Résultat de l'algorithme :
  - Identifiant du cluster d'affectation pour chaque instance.
  - Distance de chaque instance au centroïde de son cluster d'affectation .
  - Description des centroïdes des clusters : distribution des valeurs de chaque variable pour les instances du cluster (moyenne, écart-type, etc.) .



Résultat

Matrice de Données avec Clusters

ID	V_1	V_2	Cluster	Distance
1	10.50	0.20	1	0.007
2	1.70	1.18	2	0.013
3	10.90	2.51	3	0.019
4	3.27	4.98	2	0.043
5	5.21	0.39	1	0.025
...	...	...	...	...

Centroïdes des Clusters

Cluster	mean(V_1)	mean(V_2)
1	10.02	0.37
2	3.36	1.21
3	10.34	2.79

Références Internet et Bibliographie

- Sites Internet
  - KDNuggets : site de référence en Artificial Intelligence, Business Analytics, Big Data, Data Mining, Data Science et Machine Learning. https://www.kdnuggets.com/
  - SIGKDD: The community for data mining, data science and analytics. https://kdd.org/

- Bibliographie
  - Data Mining - The Textbook; Charu C. Aggarwal, Springer, 2015.
  - Data Science : Fondamentaux et Études de Cas – Machine Learning avec Python et R. Éric Biernat, Michel Lutz & Yann LeCun, Eyrolles, 2015.

Logiciel R et Langage Python

- Logiciel R
  - R and Data Mining - Examples and Case Studies, Yanchang Zhao, Academic Press, Elsevier, 2012. Documents, exemples, tutoriels et principales librairies R pour l'extraction de connaissances. http://www.rdatamining.com
  - CRAN Task Views : librairies et fonctions centrales par types d'applications https://cran.r-project.org/web/views/MachineLearning.html https://cran.r-project.org/web/views/Cluster.html
  - R Interface to Keras Deep Learning Library (Tensorflow) https://keras.rstudio.com/

- Langage Python
  - Librairie Scikit-Learn : méthodes d'apprentissage supervisé et non-supervisé https://scikit-learn.org/stable/supervised_learning.html
  - Librairie Mlxtend : Machine Learning Extensions http://rasbt.github.io/mlxtend/
  - Keras : The Python Deep Learning Library (Tensorflow) https://keras.io/