

## I. Introduction :

Le datamining a émergé comme une technologie signifiante pour gagner la connaissance à partir d'une grande quantité de données [9][10][11]. Cependant, il avait accru le souci d'utiliser cette technologie pour violer la privacy individuelle [12][13]. Ceci a mené au retour du gouvernement américain contre cette technologie. Par exemple, l'acte moratoire du datamining introduit par le sénat américain qui a interdit tout les programmes du datamining (y compris la recherche et le développement) par le département de défense américain [14].

L'objectif des approches du datamining [15] est de développer une connaissance généralisée, au lieu d'identifier une spécifique information sur les individus. Le plus grand problème n'est pas dans le datamining lui même, mais dans l'infrastructure utilisée pour le supporter. Plus les données sont complètes et précises, meilleurs seront les résultats du datamining. L'existence des ensembles de données complets, exhaustifs et précis soulève des questions de privacy indépendamment de leur usage prévu. Alors qu'une grande partie des données est déjà accessible, le fait que les données sont réparties entre plusieurs bases de données, chacune sous une autorité différente, fait que l'obtention des données pour une utilisation abusive est difficile.

Un autre problème est avec les résultats eux même. La communauté de recensement a depuis longtemps reconnu que les résumés de publication des données du recensement comportent des risques de violation de privacy. Les tableaux sommaires pour une petite région de recensement peuvent ne pas identifier un individu, mais en association (avec quelques connaissances sur l'individu, par exemple, le nombre d'enfants et le niveau d'éducation), elle peut rendre possible d'isoler un individu et de déterminer des informations privées. Les résultats du datamining représentent un nouveau type de « tableau sommaire ». Assurer la privacy signifie qu'il faut montrer que les résultats ne révèlent pas intrinsèquement les informations individuelles.

Les communautés du datamining et de la sécurité de l'information ont récemment commencé à aborder ces questions. En particulier, les techniques qui permettent d'explorer les données lorsqu'il ne nous est pas permis de les voir.

L'objectif de préservation de privacy en datamining est de permettre des situations où la connaissance présente dans les données est extraie pour utilisation, la privacy individuelle est protégée, et le porteur de données est protégé contre l'abus et la révélation des données.

Il y a de nombreux conducteurs menant à la demande accrue de la préservation de privacy en datamining. Du coté du datamining, la collecte accrue de données fournit de plus grandes opportunités de les analyser. En même temps, de plus en plus le monde concurrentiel soulève le coût de ne pas utiliser ces données.

Les conducteurs pour la préservation de privacy en datamining incluent :

- Les conditions légales de protection de données : Peut-être les mieux connus sont les règlements de la communauté européenne [12] et les règlements de healthcare de HIPAA aux États unis [13] mais beaucoup de juridictions développent de nouvelles lois de privacy souvent plus restrictives.
- Responsabilité du fait de la divulgation accidentelle de données : Même lorsque les protections juridiques n'empêchent pas le partage de données, les engagements contractuels exigent souvent la protection. Un exemple récent des Etats-unis d'un processeur de carte de crédit ayant 40 millions de numéros de cartes de crédit volés est un bon exemple : le processeur n'a pas été censé de gérer les données après fin du

traitement, mais conserve les anciennes données pour les analyser afin d'empêcher la fraude (c.-à-d., pour datamining).

- Les informations protégées constituent un compromis entre les gains d'efficacité possibles grâce au partage avec les fournisseurs, et le risque de mauvais usage de ces secrets commerciaux. L'optimisation d'une chaîne d'approvisionnement est un exemple, les entreprises font face à un compromis entre une plus grande efficacité de la chaîne d'approvisionnement, et la révélation des données sur les fournisseurs ou les clients qui peuvent compromettre la fixation des prix et les positions de négociation [16].
- Préoccupations antitrust de restreindre la capacité des concurrents à partager l'information : Comment les concurrents peuvent-ils partager des informations à des fins autorisées (par exemple, les recherches concertées sur les nouvelles technologies), mais encore prouver que l'information partagée ne permet pas de collusion dans la fixation des prix ?

Bien que ces exemples ne semblent pas être vraiment une question de privacy, la technologie de la préservation de privacy en datamining prend en charge tous ces besoins. Le but de la préservation de privacy en datamining est d'analyser les données tout en limitant la révélation de ces données à de nombreuses applications. Dans ce chapitre nous définissons ce domaine de recherche ainsi que ces principaux axes de recherche, et nous situons notre travail [1] qui concerne la préservation de privacy dans l'algorithme de clustering k-means dans cet axe de recherche.

## II. Difficulté de définir la privacy :

L'analyse de ce que signifie le droit à la privacy est un vrai problème, tel que si la privacy constitue un droit fondamental, ou si les individus ont et/ou devraient être concernés par elle. Plusieurs définitions de privacy ont été données, et elles changent selon le contexte, la culture, et l'environnement. Par exemple, dans le papier de Warren & Brandeis [17] publié en 1890, les auteurs définissent la privacy comme « le droit d'être seul ». Plus tard, dans le papier publié en 1967 [18], Westin définit la privacy comme « le désir des peuples à choisir librement dans quelles circonstances et dans quelle mesure ils s'exposent eux, leur attitude et leur comportement aux autres ». Schoeman [19] définit la privacy comme "Le droit de déterminer quels sont les renseignements personnels communiqués aux autres". Plus récemment, Garfinkel [20] déclare que la privacy est sur la possession de soi, l'autonomie et l'intégrité. Dans les définitions ci-dessus, la privacy est vue comme un concept social ou culturel. Cependant, avec l'omniprésence des ordinateurs et de l'émergence du Web, la privacy est également devenue un problème numérique [21]. Avec l'évolution du Web et l'émergence du datamining. Les soucis de la privacy ont posé des défis techniques fondamentalement différents de celles qui ont eu lieu avant l'ère de l'information. Des standards ont été établis pour définir clairement la privacy, comme dans le dictionnaire relatif aux données [22] où on définit la privacy comme « l'absence d'une intrusion non autorisée ». En considérant la préservation de privacy en datamining, ceci fournit certaines indications : Si les utilisateurs ont eu l'autorisation d'utiliser les données pour une tâche particulière du datamining. Alors, il n'y a pas un problème de "privacy". Cependant, la seconde part est la plus difficile. Si l'utilisation des données n'est pas autorisée, quelle utilisation peut être considérée comme "intrusion".

Un autre standard commun parmi la plupart des lois de "privacy" (ex: European community guidelines [12] ou US. healthcare laws [13]) est que la privacy s'applique seulement aux "données individuellement identifiables".

La combinaison entre "intrusion" et "individuellement identifiable" mène à un standard pour juger la préservation de privacy en datamining. Les techniques de préservation de privacy en datamining doivent assurer que n'importe quelle information révélée:

1. Ne peut pas être tracé à un individu (résultat du datamining); ou
2. Ne constitue pas une intrusion (processus du datamining).

Dans l'ère des technologies de l'information, se réfère à la privacy le droit des utilisateurs de cacher leurs renseignements personnels et ont un certain degré de contrôle sur l'utilisation des renseignements personnels divulgués à des tiers [23][24][25].

Clairement, le concept de privacy est souvent plus complexe que réalisé. En particulier en datamining, la définition de préservation de privacy n'est toujours pas claire. Une notable exception est dans le travail présenté dans [26], dans lequel la Préservation de Privacy en DataMining (PPDM) est définit comme obtenir des résultats valides du datamining sans avoir à apprendre les valeurs des données sous-jacentes. Cependant à ce point, chaque technique existante du PPDM a sa propre définition de privacy. Le souci primaire sur PPDM est que les algorithmes sont analysés pour les effets secondaires qu'ils s'engagent dans la privacy des données. Par conséquent, la définition pour PPDM est proche des définitions dans [19][26]. PPDM englobe le double objectif de la réunion des exigences de la confidentialité et de fournir des résultats valables du datamining. Dans ce qui suit, la définition met l'accent sur le dilemme de l'équilibre entre la préservation de la privacy et de la divulgation des connaissances.

### **III. Préservation de la privacy en datamining (PPDM) [27]:**

En général, la préservation de la privacy apparaît dans deux dimensions majeures: l'information personnelle des utilisateurs et l'information concernant leur activité collective. On se réfère à la première comme préservation de la privacy individuelle et à la seconde comme préservation de la privacy collective, qui est liée à la privacy d'entreprise dans [26].

#### **III.1 Préservation de la privacy individuelle:**

L'objectif principal de la privacy des données est la protection des renseignements personnels identifiables. En général, l'information est considéré comme personnellement identifiable si elle peut être liée, directement ou indirectement, à une personne physique. Ainsi, lorsque les données personnelles sont soumises à l'exploitation, les valeurs des attributs associées à des personnes sont privées et doivent être protégés contre la divulgation. Les algorithmes du datamining sont alors en mesure d'apprendre à partir des paramètres globaux plutôt qu'à partir des caractéristiques d'un individu en particulier.

#### **III.2 Préservation de la privacy collective:**

La protection des données personnelles peut ne pas suffire. Parfois, on peut avoir besoin de se protéger contre l'apprentissage des connaissances sensibles représentants les activités d'un groupe. On se réfère à la protection des connaissances sensibles comme la préservation de la vie privée collective. Dans le cas de la préservation de la privacy collective, les organisations doivent faire face à certains conflits intéressants. Par exemple, lorsque des

renseignements personnels sont soumis à des processus d'analyse qui produisent des faits nouveaux sur les habitudes d'achat des utilisateurs, les loisirs, ou les préférences, ces faits pourraient être utilisés dans les recommandations du système pour prévoir ou d'influer sur leurs comportements futurs d'achats. En général, ce scénario est bénéfique tant pour les utilisateurs et les organisations. Toutefois, lorsque des organisations partagent les données dans un projet collaboratif, le but n'est pas seulement de protéger les informations personnelles identifiables, mais aussi quelques paramètres stratégiques. Dans le monde du commerce, de tels paramètres sont décrits comme des connaissances qui peuvent fournir des avantages concurrentiels, et doivent donc être protégées [28]. Plus difficile est de protéger les connaissances découvertes à partir des informations confidentielles (par exemple, médicale, financière et des renseignements criminels). L'absence de mesures de privacy peut également compromettre la privacy des individus. Alors que la violation de la privacy est claire, la violation de la privacy collective peut conduire à la violation de la privacy individuelle.

#### IV. Caractérisation des scénarios en PPDM [27]:

Avant de décrire les paramètres généraux pour la caractérisation des scénarios en PPDM, prenons deux exemples de la vie réelle lorsque PPDM pose différentes contraintes:

- **Scénario 1:** Un hôpital partage quelques données pour des fins de recherches (exemple, concernant un groupe de patients des maladies similaires). Les administrateurs de sécurité de l'hôpital peuvent supprimer certains identifiants (par exemple nom, adresse, numéro de téléphone, ... etc.) à partir des dossiers des patients pour répondre aux exigences de privacy. Toutefois, les données publiées peuvent ne pas être pleinement protégées. Un enregistrement d'un patient peut contenir d'autres informations qui peuvent être liés à d'autres ensembles de données qui ré identifient les individus ou les entités [29]. Comment peut-on identifier des groupes de patients atteints d'une maladie similaire, sans révéler les valeurs des attributs associés avec eux ?
- **Scénario 2:** Deux ou plusieurs compagnies ont un très large ensemble de données ou d'enregistrement sur les activités d'achat de leurs clients. Ces compagnies décident de mener en collaboration une exploration des règles d'association sur leurs ensembles de données pour leur intérêt réciproque puisque cette collaboration leur apporte un avantage sur les autres concurrents. Toutefois, certaines de ces compagnies peuvent ne pas vouloir partager certains paramètres stratégiques cachés au sein de leurs propres données (appelée aussi les règles d'association restrictives) avec les autres parties. Ils tiennent à transformer leurs données de telle façon que ces règles d'association restrictives peuvent être découvertes, mais d'autres ne peuvent pas l'être. Est-il possible pour ces entreprises de bénéficier d'une telle collaboration, en partageant leurs données tout en préservant certaines règles restrictives d'association ?

Notons que les scénarios ci-dessus décrivent différents problèmes de préservation de la privacy. Chaque scénario présente un ensemble de défis. Par exemple, le scénario 1 est un exemple typique de la préservation de la privacy, tandis que le scénario 2 se réfère à la préservation de la privacy collective. Comment peut-on caractériser les scénarios dans PPDM? Une alternative est de les décrire en termes de paramètres généraux. Dans [30], certains paramètres sont suggérés:

- **Le résultat:** Se réfère aux résultats du datamining souhaité. Par exemple, quelqu'un peut rechercher les règles d'association qui peuvent identifier les relations parmi les attributs ou les relations parmi les comportements des clients acheteurs comme dans le scénario 2, ou même peuvent souhaiter de grouper les données comme dans le scénario 1.
- **La distribution des données:** Comment les données sont disponibles pour le datamining: sont-elles centralisées ou distribuées à travers plusieurs sites? Dans le cas de données distribuées à travers de nombreux sites, les entités sont décrites avec le même schéma dans tous les sites (distributions horizontales) ou différents sites contiennent des attributs différents pour une seule entité (parfois verticales) ?
- **Préservation de la privacy:** Quelles sont les exigences de la préservation de privacy ? Si la préoccupation est uniquement celle des valeurs associées à une entité individuelle de ne pas être révélés (par exemple, des renseignements personnels), les techniques doivent être axées sur la protection de ces informations. Dans d'autres cas, la notion de ce qu'est ce la "connaissance sensible" ? peut ne pas être connue à l'avance. Cela conduirait à une évaluation humaine des résultats intermédiaires avant de rendre les données disponibles pour le datamining.

Selon ces paramètres, la démarche à entreprendre pour préserver la privacy en datamining change. Plusieurs algorithmes et méthodes ont été développés dans le domaine, chacune de ces méthodes traite une tâche particulière du datamining et répond à un besoin particulier de la préservation de privacy. Ceci a conduit à plusieurs travaux de recherche qui développent des classifications sur ces méthodes et algorithmes afin d'établir des standards et de chercher une sorte de généralisation de solution.

## **V. Préservation de la privacy en datamining : Modèles et Algorithmes.**

Les méthodes et techniques de préservation de privacy en datamining ont fait l'objet de plusieurs discussions [31][32][34][35]. Le problème est de déterminer les critères sur lesquels on peut classer ces méthodes et techniques. Nous détaillons dans ce qui suit les classifications qui ont été proposées, ainsi que les critères adoptés dans ces classifications.

### **V.1 La classification de V. S. Verykios, E. Bertino [31]:**

Dans [31], les auteurs proposent une classification se basant sur les dimensions suivantes:

- La distribution des données
- La modification des données
- L'algorithme du datamining.
- Les règles ou les données à cacher
- La préservation de privacy.

La première dimension se réfère à la distribution des données. Certaines des approches de préservation de privacy en datamining ont été développées pour les données centralisées, alors que d'autres se réfèrent à un scénario de données distribuées.

La deuxième dimension concerne les schémas de modification de données. La modification des données est basée sur l'idée de ne pas fournir les données réelles à l'algorithme du datamining. En général, la modification des données est utilisée pour modifier les valeurs

originales de la base de données qu'on veut révéler au public et cette méthode assure une préservation élevée de la privacy. Les méthodes de modification incluent : la perturbation qui est accomplie par l'altération de la valeur d'un attribut par un nouvel attribut en ajoutant du bruit, le blocage, qui est le remplacement d'un attribut existant par « ? » ou l'échantillonnage, qui se rapporte à libérer des données pour seulement un groupe d'une population.

La troisième dimension se réfère à l'algorithme du datamining, pour lequel la modification des données prend place. Pour l'instant, les divers algorithmes du datamining ont été considérés en isolation l'un par rapport à l'autre. Parmi eux, les idées les plus importantes qui ont été développées pour la classification des algorithmes du datamining comme les inducteurs d'arbre de décision, les algorithmes de fouille des règles d'association, les algorithmes de clustering, les ensembles approximatifs et les réseaux bayesiens.

Le dernier aspect qui est le plus important, se réfère à la technique de préservation de la privacy utilisée pour la modification choisie des données. La modification choisie est nécessaire afin de parvenir à une utilité supérieure pour les données modifiées étant donné que la privacy n'est pas compromise. Les techniques qui ont été appliquées dans ce sens, sont les suivantes:

- a. Les techniques à base d'heuristique comme la modification adaptative qui ne modifie que les valeurs sélectionnées, ce qui minimise la perte d'utilité plutôt que toutes les valeurs disponibles.
- b. Les techniques à base de cryptographie : comme le calcul multi-partie sécurisé où le calcul est sécurisé si à la fin, aucune partie n'apprend rien sauf sa propre entrée et les résultats. Ceci est le modèle de privacy étudié dans notre travail [1].
- c. Les techniques à base de reconstruction : où la distribution originale des données est reconstruite à partir des données aléatoires.

## **V.2 La classification de C. Aggarwal et S.YU [32]:**

Dans [32], Les auteurs propose une autre classification selon la façon de préservation de privacy, en dressant un aperçu de recherche sur un grand nombre d'algorithmes et de méthodes dans le domaine, que nous résumons dans ce qui suit :

### **V.2.1 La méthode de perturbation des données:**

La perturbation des données est une technique de préservation de privacy en datamining dans laquelle un bruit est ajouté aux données afin de masquer les valeurs d'un attribut au fil des enregistrements [36][37]. Le bruit ajouté est suffisamment large que les valeurs individuelles d'un enregistrement ne peuvent pas être repêchées. Par conséquent, des techniques sont conçues pour dériver des distributions globales à partir des enregistrements perturbés. Par la suite, les techniques de datamining peuvent être développées de telle sorte qu'elles travaillent avec ces distributions globales.

### **V.2.2 Les modèles *k-anonymité* et *l-diversité* [38][39]:**

Le modèle *k-anonymité* [38] est développé à cause de la possibilité de l'identification indirecte des enregistrements à partir des bases de données publiques. C'est parce que les combinaisons d'attributs des enregistrements peuvent être utilisées pour identifier exactement les enregistrements individuels.

Dans la méthode k-anonymité, on réduit la granularité de la représentation des données en utilisant les techniques comme la généralisation et la suppression. Cette granularité est réduite suffisamment que n'importe quel enregistrement donné trace sur au moins  $k$  autre enregistrements dans les données. Le modèle l-diversité [39] est conçu pour traiter certaines faiblesses dans le modèle k-anonymité, puisque protéger les identités au niveau k-individuels n'est pas comme protéger les valeurs sensibles correspondantes, en particulier, lorsqu'il y a une homogénéité des valeurs sensibles dans un groupe. Pour faire ainsi, le concept de la diversité intra-groupe des valeurs sensibles est favorisé dans les schémas d'anonymisation.

### **V.2.3 Préservation distribuée de la privacy :**

Dans plusieurs cas, des entités individuelles peuvent souhaiter dériver des résultats composés à partir d'ensembles de données qui sont distribués à travers ces entités. Une telle distribution peut être horizontale (lorsque les enregistrements sont distribués à travers ces entités multiples) ou verticale (lorsque les attributs sont distribués à travers plusieurs entités). Tandis que les différentes entités peuvent ne pas désirer partager leur ensemble de données entières, elles peuvent consentir au partage de l'information limité en utilisant une variété de protocoles. L'effet global de telles méthodes doit maintenir la privacy pour chaque entité individuelle, tout en dérivant des résultats composés sur les données entières. Ceci est l'objectif de notre travail pour l'algorithme de clustering k-means [1].

### **V.2.4 Rétrogradation de l'efficacité de l'application :**

Dans plusieurs cas, même si les données peuvent ne pas être disponibles, le résultat des applications telles que l'exploration des règles d'association, la classification ou le traitement des requêtes peut déboucher sur des violations de privacy. Ceci à mener à la recherche de rétrograder l'efficacité des applications par la modification des données ou des applications. Quelques exemples de telles techniques comprennent la protection des règles d'association [40], rétrogradation des classificateurs [41] et l'audit des requêtes [42].

### **V.3 Autres classifications [33][34][35]:**

Dans [33], les auteurs donnent un aperçu de recherche et une classification des méthodes et techniques de préservation de privacy en datamining dans le cas où les algorithmes du datamining s'exécutent sur un ensemble de données distribué sur plusieurs sites (parties), et ils distinguent entre deux approches celle de la perturbation des données et celle basé sur le calcul multi-partie sécurisé ou la cryptographie. L'objectif ici est protéger les données individuelles ou sensibles lorsqu'on n'est pas autorisé à les voir. Il s'agit d'une préservation de privacy au niveau du processus du datamining lors de la communication des parties. Les auteurs privilégient les méthodes basées sur le calcul multi-partie sécurisé par rapport aux méthodes basées sur la perturbation ou la randomisation car le défi de cet axe de recherche est dans la façon d'obtenir des résultats valides à partir des données perturbées. Aussi, il n'existe pas de garanties formelles de préservation de la privacy. Le second axe, est basé sur la séparation des autorités : Les données sont présumées être contrôlées par des entités différentes, et l'objectif est que ces entités doivent coopérer pour obtenir des résultats valides du datamining sans avoir à révéler leurs propres données.

La classification donnée dans [34], développe un aperçu de recherche sur tous les algorithmes de préservation de privacy mais seulement pour les données distribuées verticalement, du fait que ce modèle de partitionnement de données est très important, et il se répète souvent dans le monde réel. Dans [35] l'auteur discute les méthodes de préservation de privacy basées sur la

cryptographie, il décrit les résultats et discute leur efficacité, il démontre aussi la correspondance des méthodes de cryptographie à la préservation de privacy en datamining.

Le besoin est aussi crucial pour fournir une vue complète sur les métriques qui évaluent les différentes méthodes et algorithmes de préservation de privacy en datamining existants afin de gagner en perspicacité dans la conception des mesures efficaces et des algorithmes. E. Bertino et al, présentent dans l'un de leurs travaux récents un aperçu de la quantification des algorithmes de préservation de privacy en datamining [43].

## **VI. Les métriques d'évaluation des algorithmes de préservation de privacy en datamining :**

Dans [43], les auteurs définissent les principaux objectifs que doit imposer un algorithme PPDM :

- Un algorithme PPDM doit empêcher la découverte de données sensibles.
- Il doit être résistant aux différentes techniques du datamining
- Il ne doit pas compromettre l'accès et l'utilisation de données non sensibles.
- Il ne doit pas avoir une complexité calculatoire exponentielle.

Egalement ; on définit l'ensemble suivant des critères sur lesquels un algorithme PPDM peut être évalué :

- Le niveau de privacy qui est offert par la technique de préservation, qui indique de prés comment l'information sensible, celle qui a été cachée, peut encore être estimée.
- L'échec de cacher l'information ; c'est la partie des informations sensibles qui ne sont pas cachés par l'application de la préservation de privacy.
- La qualité des données ; après l'exécution de la technique de préservation de privacy, il faut considérer à la fois la qualité des données elles-mêmes et la qualité des résultats du datamining.
- La complexité, c'est la capacité d'un algorithme de préservation de privacy à s'exécuter avec une bonne performance en termes de toutes les ressources impliquées par l'algorithme.

En considérant le cas des algorithmes de datamining qui s'exécutent sur des ensembles de données distribuées ou les techniques de préservation de privacy qui entrent dans la catégorie des données distribuées, on distingue comme sus cité deux approches différentes celles de la perturbation des données et celles basées sur la cryptographie. Les trois premiers critères d'évaluation des algorithmes de PPDM concernent en particulier la première approche, où il n'existe pas de garanties formelles de préservation privacy, et le risque de perdre en qualité des données et des résultats à cause du processus de randomisation des données. Tandis que la deuxième approche est concernée particulièrement par la complexité des algorithmes puisqu'on utilise les techniques avancées de la cryptographie.

Dans notre travail, nous nous intéressons à la préservation de privacy basée sur la cryptographie en utilisant le modèle du calcul multi – partie sécurisé, le scénario est que l'ensemble de données est distribué sur plusieurs parties ou sites, ce qui peut être vu comme



le datamining distribué. La plupart des outils opèrent sur le principe de récolter tout les données dans un site central, et exécuter ensuite l'algorithme du datamining. Il existe un certain nombre d'applications qui ne sont pas faisable sous une telle méthodologie, d'où la nécessité du datamining distribué. Il existe plusieurs situations où ce besoin s'impose:

**1. Connectivité:** Transmission de grandes quantités de données vers un site central peut ne pas être faisable.

**2. Hétérogénéité des sources:** Est-il plus facile de combiner les résultats ou de combiner les sources?

**3. Confidentialité des sources:** Les organisations peuvent être prêtes à partager les résultats du datamining et non pas les données.

Plusieurs algorithmes ont été proposés pour le datamining distribué. Cheung et al. [44] ont proposé une méthode pour une distribution horizontale des données, et une classification distribuée a été proposée. Un travail récent a adressé une classification sur les données distribuées verticalement [45]. Cependant, aucun de ces travaux n'a soulevé les soucis de privacy. Nous résumons dans ce qui suit les manières dont lesquelles les données peuvent être distribuées.

## VII. Les modèles de distribution de données:

Il est nécessaire d'abord de présenter les différentes façons dont lesquelles les données sont distribuées dans le monde réel. Il existe deux modèles de base de distribution des données: partitionnement horizontal (distribution homogène) et partitionnement vertical (distribution hétérogène). On définit une base de données  $D$  en termes d'entités pour lesquelles les données sont collectées et l'information qui est collectée pour chaque entité. Ainsi,  $D \equiv (E, I)$ , où  $E$  est l'ensemble des entités et  $I$  l'ensemble des attributs qui sont collectés. On suppose qu'il existe  $t$  différents sites (parties),  $P_1, \dots, P_t$  rassemblant les bases de données  $D_1 \equiv (E_1, I_1), \dots, D_t \equiv (E_t, I_t)$  respectivement.

Le partitionnement horizontal des données suppose que les différentes parties collectent la même sorte d'information sur différentes entités. Donc, dans le partitionnement horizontal:

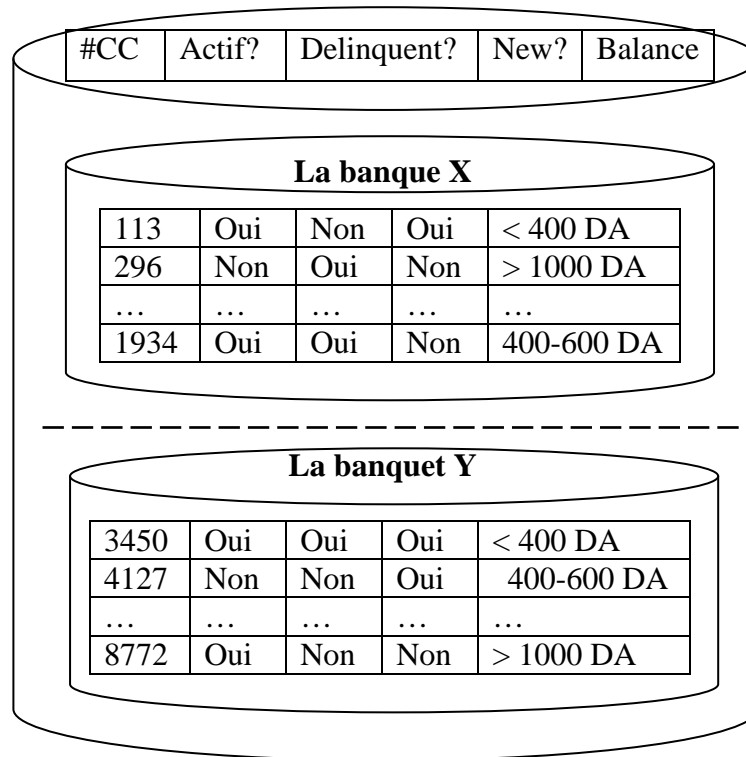
$$E_G = \cup_i E_i = E_1 \cup \dots \cup E_t, \text{ et } I_G = \cap_i I_i = I_1 \cap \dots \cap I_t.$$

Plusieurs cas de cette situation existent dans la vie réelle. Par exemple, toutes les banques collectent plusieurs informations similaires, mais la base des clients pour chaque banque tend à être toute à fait différente. La figure 1 démontre le partitionnement horizontal de données. La figure montre deux banques : la banque X et la banque Y, chacune d'elles collecte les informations sur les cartes de crédits de leurs clients respectifs. Attributs comme la balance du compte, si le compte est nouveau ou pas, actif, délinquant sont collectées par les deux. Le fusionnement des deux bases de données devrait mener à des modèles prédictifs plus précis utilisés pour des activités comme la détection de fraude.

D'autre part, le partitionnement vertical des données suppose que différentes parties collecte différents ensembles d'attributs pour un même ensemble d'entités. Ainsi, dans le partitionnement vertical de données:

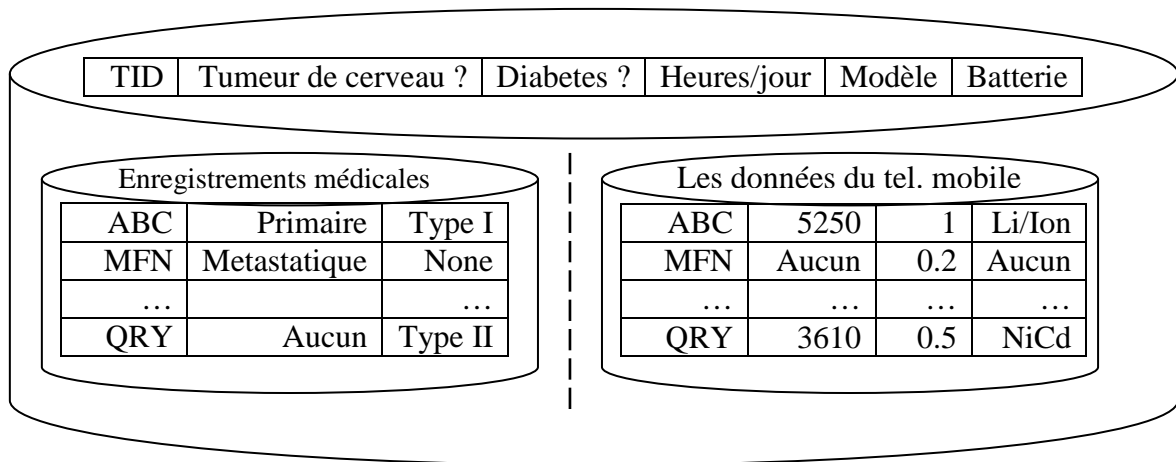
$$E_G = \cap_i E_i = E_1 \cap \dots \cap E_k, \text{ et } I_G = \cup_i I_i = I_1 \cup \dots \cup I_k$$

Par exemple, Ford collectent des informations sur les véhicules manufacturés. Fire-stone collecte des informations sur les pneus manufacturés. Les véhicules peuvent liés aux pneus. Cette liaison peut être utilisée pour joindre les bases de données. La base de données globale peut être alors explorée pour révéler des informations utiles. La figure 2 montre le partitionnement vertical de données. D'abord, nous voyons une compagnie d'assurance hypothétique d'un hôpital rassemblant des enregistrements médicaux tels que le type de tumeur du cerveau et de diabète. D'autre part, un fournisseur de téléphonie mobile pourrait



**Fig.1 Distribution horizontale / Distribution homogène des données**

### Vue globale de la base de données



**Fig.2 Distribution verticale / hétérogène des données**

rassembler d'autres informations telles que, la quantité approximative du temps antenne utilisée chaque jour, le modèle de téléphone portable et le genre de batterie utilisé. Le fusionnement de cette information pour des clients communs et l'exécution d'algorithmes de datamining pourrait donner des corrélations complètement inattendues (par exemple, une personne avec le type de diabète I utilisant un téléphone mobile avec des batteries de Li/Ion pour plus qu'une heure par jour est très susceptibles de souffrir des tumeurs cérébrales primaires). Il serait impossible d'obtenir une telle information en considérant l'une des bases de données toute seule.

Il existe encore d'autres modèles de distribution de données, comme par exemple le modèle de partitionnement arbitraire de données [46] qui généralise les deux cas (partitionnement vertical et horizontal), où différents attributs pour différentes entités pour un même ensemble de données peuvent être possédés par plusieurs parties.

Le besoin de préserver la privacy dans le cas du datamining distribué est plus clair que celui lorsqu'il est centralisé, car la préservation de privacy ici permet d'explorer les données lorsqu'il ne nous est pas permis de les voir. Le deuxième problème, le potentiel des résultats du datamining de révéler des informations privées, a reçu moins d'attention. C'est en grande partie parce que les concepts de la privacy ne sont pas bien définis - et sans une définition formelle, il est difficile de dire si la privacy a été violée. Plusieurs travaux de préservation de privacy sur des ensembles de données distribués ont été développés pour plusieurs tâches du datamining (la classification, la découverte des règles d'association, le clustering, ...). Les premiers travaux de préservation de privacy en datamining sur un ensemble de données distribué ont été donnés par [36] [47] sur l'algorithme ID3 concernant la classification par les arbres de décisions [48][49]. Chaque approche applique un modèle de privacy différent. Dans [36], la privacy est protégée en perturbant l'ensemble original de données à l'entrée de l'algorithme. La question est de savoir comment s'assurer du résultat de l'algorithme sur l'ensemble de données perturbé à partir de l'ensemble de données original et comment garantir formellement la préservation de privacy. Dans [47], un modèle plus rigoureux est utilisé, celui du calcul multi-partie sécurisé qui est basé sur les techniques de cryptographie. Notre travail [1] entre dans le cadre de la préservation de privacy en appliquant le modèle du calcul multi-partie sécurisé dans l'algorithme de clustering k-means qui est une technique

exploratoire très utilisée du datamining. La préservation de privacy individuelle ou même collective est réalisée en protégeant les items (les points de données) de la révélation lorsqu'ils sont distribués sur plusieurs parties.

### **VIII. Conclusion:**

Dans ce chapitre, nous avons défini le domaine de préservation de privacy en datamining, et nous avons présenté les grands axes de recherches dans ce domaine, et aussi défini les métriques d'évaluation des algorithmes de préservation de privacy en datamining et les modèles de distribution de données. Le clustering est aussi une tâche plus exploratoire du datamining et qui pose le même problème de privacy lorsque les données proviennent de sources différentes. La plupart des travaux de préservation de privacy en clustering sont élaborés sur l'algorithme de k-means en appliquant le modèle du calcul multi partie sécurisé. Plusieurs parties porteuses de données participent dans l'exécution de l'algorithme k-means sans qu'une partie ne prenne connaissance des données des autres parties. Dans le chapitre suivant, nous définissons le clustering et l'algorithme k-means dont nous sommes intéressés par la préservation de privacy lorsqu'il est exécuté sur un ensemble de données distribué.