Calculer avec des données chiffrées

Il est théoriquement possible de demander à un tiers de traiter des données même si votre requête et les données sont chiffrées. Le chiffrement complètement homomorphe permet ce tour de force. Les travaux dans ce domaine ont bien progressé.

Le principe est simple. Imaginons que vous chiffriez simplement vos données d_1, d_2, d_3 en les multipliant par une constante C (la clé de chiffrement). Vous envoyez les données chiffrées Cd_1, Cd_2 et Cd_3 sur le serveur tiers. Vous pouvez alors demander à l'opérateur d'effectuer certains calculs simples sur ces données, telle leur moyenne. Il suffit alors de diviser par C le résultat que l'opérateur vous renvoie pour obtenir la moyenne de vos données. Ce faisant, l'opérateur n'a pris connaissance ni des vraies données ni de leur moyenne.

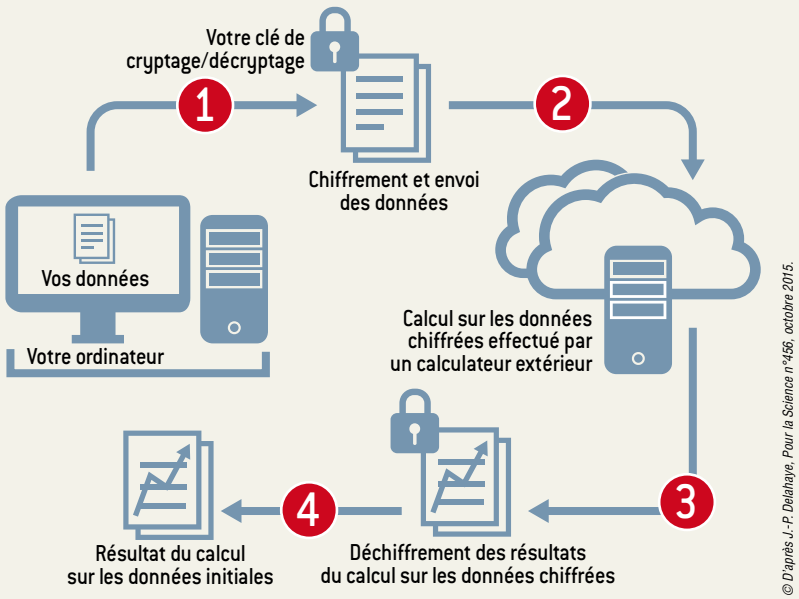
Certaines demandes peuvent être effectuées dans ces conditions, mais d'autres ne sont pas possibles sans révéler la clé de chiffrement.

Pour exécuter toutes sortes de calculs, le système de chiffrement doit être complètement homomorphe, c'est-à-dire qu'il doit présenter deux propriétés : $\mathscr{C}(d_1+d_2)=\mathscr{C}(d_1)\oplus\mathscr{C}(d_2)$ et $\mathscr{C}(d_1\times d_2)=\mathscr{C}(d_1)\otimes\mathscr{C}(d_2)$, où $\mathscr{C}$ est l'opération de chiffrement, d_1 et d_2 deux données, $\oplus$ et $\otimes$ deux opérations propres au chiffrement. L'opérateur tiers exécute les opérations $\oplus$ et $\otimes$ sur les données chiffrées, ce qui revient à additionner ou multiplier les données initiales.

Avec ces deux propriétés, toutes les opérations logiques booléennes (ET, OU, etc.) sont définissables et il est possible d'effectuer toutes les manipulations désirées sur les données.

Le système RSA est homomorphe pour la multiplication, mais pas pour l'addition. En 2009, Craig Gentry a proposé le premier système complètement homomorphe. Celui-ci était impossible à mettre en pratique : les calculs à effectuer au moment du chiffrement et du déchiffrement étaient trop volumineux. Et la taille des données à manipuler par l'opérateur tiers était des millions de fois plus importante que celle des données avant chiffrement.

Les travaux récents ont permis de réduire le surcoût en taille des données et en lourdeur de calcul. Le système est déjà utilisable dans certains cas simples.



© D'après J.-P. Delahaye, Pour la Science n°456, octobre 2015.

PLS

La recherche en cryptographie propose-t-elle des solutions alternatives ?

D. P. : Les systèmes tendent aujourd'hui vers un chiffrement de bout en bout par l'utilisateur. Ses données, sur le serveur d'un opérateur tel Google, seront alors chiffrées. L'utilisateur ne pourra plus utiliser les outils de l'opérateur pour effectuer des opérations sur ses données, telle une recherche par mots-clés. En effet, comment soumettre une requête à Google si celui-ci ne peut pas lire les données ? C'est théoriquement possible avec le « chiffrement complètement homomorphe », proposé en 2009 par Craig Gentry, un chercheur d'IBM. Et mieux, il est possible d'obtenir une réponse pertinente, que l'opérateur ne voit pas, à une question qu'il ne connaît pas ! Ainsi, avec une requête sur des données cryptées, Google pourra effectuer les opérations que vous lui demandez.

Cette méthode de chiffrement est un grand axe de recherche actuel *(voir l'encadré ci-contre)*. Des progrès ont été réalisés pour réduire les temps de calcul. Par exemple, un calcul test par un système complètement homomorphe qui prenait plusieurs heures se fait maintenant en quelques secondes. On reste loin d'une mise en place à grande échelle, mais ce système sera la prochaine étape dans la sécurisation des données.

PLS

L'ordinateur quantique, s'il devient réalité, menacera-t-il les systèmes de cryptage ?

D. P. : Cette menace reste théorique, mais nous y réfléchissons déjà. L'ordinateur quantique exploite des propriétés de la physique quantique pour accroître les capacités de calcul. Ainsi, un algorithme adapté pourrait effectuer certaines opérations plus vite qu'avec un ordinateur classique. En particulier, nous avons déjà un algorithme quantique de factorisation. Le jour où un ordinateur quantique pourra l'utiliser, le système RSA sera obsolète. Nous étudions donc des systèmes de chiffrement classiques, dits postquantiques, dont le principe repose sur des problèmes difficiles que même les algorithmes quantiques ne peuvent résoudre efficacement – pour l'instant... ■

Propos recueillis par Sean BAILLY

La Biostatistique a son instrument

Stata fournit les outils statistiques, graphiques et de traitement de données nécessaires à toute recherche impliquant de la biostatistique.

- Analyse de survie multi-niveaux
- Estimateurs pour les données de panel
- Calcul de puissance pour les tables de contingence
- Méthodes de rééchantillonnage et de simulation
- Imputation Multiple
- Convient aux mesures longitudinales ou transversales

Programmable

Précis

Exhaustif

ritme.com/stata

+33 (0) 1 42 46 00 42

Distributeur officiel en France, Suisse, Belgique et Luxembourg

RITME - 34 Bd Haussmann, 75009 Paris, France

RITME SCIENTIFIC SOLUTIONS

STATA® 14 release

CABINET DE CURIOSITÉS SOCIOLOGIQUES par Gérald Bronner

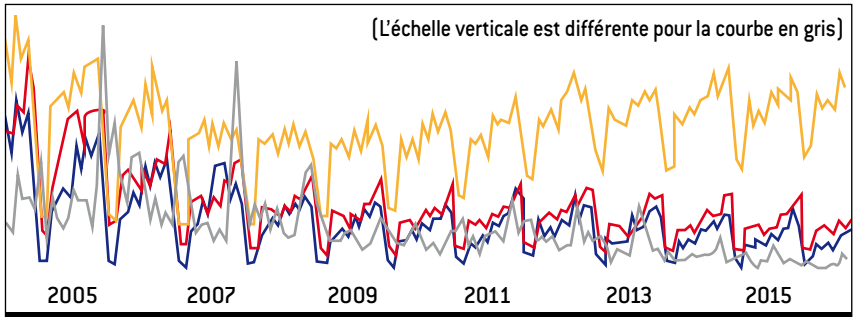
Les classiques à la lumière des données massives

L'examen des recherches faites par les internautes offre un critère pour déterminer si un auteur est ou non un classique.

Qu'est-ce qui fait qu'un auteur est considéré comme un classique de la littérature ou de la philosophie ? La question est redoutable, car il est difficile d'identifier les mécanismes qui feront qu'une œuvre sera durablement remarquée. Rappelons le célèbre mot de saint Augustin sur la définition du temps : « Si personne ne m'interroge, je le sais ; si je veux répondre à cette demande, je l'ignore. » Il en va de même avec les classiques : nous savons bien ce qu'ils sont, mais nous ne sommes plus d'accord sur rien dès qu'il s'agit d'avancer vers une définition.

Je n'ai ainsi aucun doute sur le fait que Victor Hugo soit un classique et Marc Lévy non. Et je sais aussi que ce n'est pas parce que le premier est mort et le second encore vivant. On est donc tenté d'admettre une définition minimaliste : un auteur serait un classique parce qu'un certain nombre d'acteurs sociaux (universitaires, journalistes et commentateurs de toutes sortes) le considèrent comme tel. Dès lors, est-il possible d'identifier grâce aux données massives (ou *big data*) les conséquences de ce consensus social autour d'un auteur qui nous permettrait d'identifier sa présence au Panthéon des humanités au-delà de la seule impression que nous pouvons en avoir ?

Voyons justement comment se comporte l'algorithme Google Tendances, qui indique le volume de recherches menées par les internautes sur tel ou tel sujet durant une période donnée et dans un pays considéré. J'ai fait cette expérience pour « Marc Lévy ». La courbe de requêtes en France, entre 2004 et 2016, ne présente pas de régularité particulière. De façon assez logique, ses pics correspondent à la sortie d'un nouveau livre avant l'été. En revanche, la structure de demande concernant les auteurs considérés comme classiques est au premier abord étrange et déconcertante. On aurait pu croire qu'être classique, c'est recueillir un volume de recherches constant et peu affecté par les événements.



REQUÊTES EFFECTUÉES SUR GOOGLE entre 2004 et 2015 en France portant sur « Victor Hugo » *(en jaune)*, « Baudelaire » *(en rouge)*, « Molière » *(en bleu)* et « Marc Lévy » *(en gris)*.

G. Bronner

Il n'en est rien, comme l'indiquent les courbes associées à « Victor Hugo », « Molière » et « Baudelaire » *(ci-dessus)*. Le volume de demande n'est pas le même pour chacun de ces auteurs, mais la structure de la demande est étonnamment semblable, avec un cycle identifiable à l'œil nu (avec des pics et des creux en phase).

Comment expliquer cette structure ? Le mystère n'est pas bien grand. Il suffit de remarquer que les creux correspondent systématiquement à la période juillet-août et les pics au reste de l'année. Ainsi, le consensus autour d'un auteur – ce qui permet de le qualifier de classique – se traduit en activité sociale (ledit auteur fait partie de la bibliographie scolaire, universitaire etc.) fortement influencée par un calendrier préétabli. C'est sans doute là le meilleur critère à retenir pour identifier méthodiquement les auteurs classiques.

À partir de là, on pourrait mettre en place un nouveau type d'études pour faire des comparaisons internationales : Molière est-il aussi un classique en Chine ou en Allemagne ? On pourrait aussi identifier l'élévation au rang de classique de tel ou tel auteur durant une période donnée. La courbe de Marguerite Duras, par exemple, a une structure bien moins typique que celle de Gustave Flaubert. Ce critère me paraît bien plus fiable que l'intronisation d'un auteur dans la collection de la Pléiade ou toute autre forme de reconnaissance sociale. L'agrégation des recherches collectives – plutôt que le Panthéon – est une forme solide de reconnaissance.

Gérald BRONNER est professeur de sociologie à l'université Paris-Diderot.

Abonnez-vous à POUR LA SCIENCE

Plus de 24% de réduction

OFFRE LIBERTÉ

4,90€ PAR MOIS SEULEMENT

BULLETIN D'ABONNEMENT

À renvoyer accompagné de votre règlement à : Pour la Science - Service abonnements - 19 rue de l'industrie - BP 90 053 - 67 402 Illkirch cedex

PAS463

❑ **OUI**, je m'abonne à *Pour la Science* **formule Découverte.**
Je règle par prélèvement automatique de 4,90€ par mois et je complète l'autorisation ci-contre. **J'économise 24% par mois.**
(IPV4E90)

MES COORDONNÉES

Nom : _____

Prénom : _____

Adresse : _____

_____

Code postal |_|_|_|_|_|

Ville : _____

Tél. Pour le suivi client (facultatif) : |_|_|_|_|_|_|_|_|_|_|

E-mail obligatoire : _____

@ _____

J'accepte de recevoir les informations de *Pour la Science* ❑ OUI ❑ NON
et de ses partenaires ❑ OUI ❑ NON

Délai de livraison: dans le mois suivant l'enregistrement de votre règlement. Offre réservée aux nouveaux abonnés, valable jusqu'au 31/05/16 en France métropolitaine uniquement. Pour un abonnement à l'étranger, merci de consulter notre site www.pourlascience.fr. Conformément à la loi "Informatique et libertés" du 6 janvier 1978, vous disposez d'un droit d'accès et de rectification aux données vous concernant en adressant un courrier à *Pour la Science*. Votre abonnement en prélèvement est reconduit automatiquement et peut être interrompu par simple lettre.

MANDAT DE PRÉLÈVEMENT SEPA En signant ce mandat SEPA, j'autorise Pour la Science à transmettre des instructions à ma banque pour le prélèvement de mon abonnement dès réception de mon bulletin. Je bénéficie d'un droit de rétractation dans la limite de 8 semaines suivant le premier prélèvement. Plus d'informations auprès de mon établissement bancaire.

TYPE DE PAIEMENT : PAIEMENT RÉCURRENT

Titulaire du compte

Nom : _____

Adresse : _____

Code postal |_|_|_|_|_| Ville : _____

Désignation du compte à débiter

BIC (Identification internationale de la banque) |_|_|_|_|_|_|_|_|_|_|_|

IBAN |_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|
(Numéro d'identification international du compte bancaire)

Établissement teneur du compte

Nom : _____

Adresse : _____

Code postal |_|_|_|_|_| Ville : _____

Date et signature

Joindre un RIB

Organisme Créancier
Pour la Science - 8 rue Férou - 75006 Paris
N° ICS FR92ZZZ426900
N° de référence unique de mandat (RUM)

Partie réservée au service abonnement. Ne rien inscrire

HOMO SAPIENS INFORMATICUS chronique de Gilles Dowek

Quel est le prix de notre anonymat ?

Les dossiers médicaux sont une source d'informations pour la recherche, mais il faut protéger l'identité des personnes.

Les hôpitaux, écoles, administrations, etc. accumulent des données qui sont, pour les chercheurs, une mine d'or. Par exemple, les données collectées au contact des malades par les médecins qui les soignent ont une énorme valeur pour les chercheurs en médecine. Mais diffuser des dossiers médicaux, même en direction de chercheurs, soulève des problèmes éthiques, en particulier parce que ces chercheurs, même s'ils ne les publient pas directement, publieront des informations qui en sont issues. Comment garantir que ces informations ne permettent pas, à une personne mal intentionnée, de remonter aux données initiales et d'en déduire l'état de santé d'une personne particulière ? Outre que cela constituerait une violation de son intimité, ces informations, utilisées par son employeur ou son assureur, pourraient nuire concrètement à cette personne.

Une solution simple est de demander à l'hôpital d'anonymiser ces dossiers, c'est-à-dire d'y supprimer le nom propre des patients, qui n'est pas utile aux chercheurs. Cette méthode est toutefois souvent qualifiée de « pseudo-anonymisation », car elle n'est pas très efficace.

Par exemple, supprimer le nom du compositeur Wolfgang Amadeus ***, né à Salzbourg le 27 janvier 1756 et mort à Vienne le 5 décembre 1791, n'empêche pas de l'identifier. Certes, tout le monde n'est pas aussi célèbre que l'auteur de *La Flûte enchantée*, mais une étude a montré que 87 % des personnes vivant aux États-Unis étaient identifiables par leur code postal, leur sexe et leur date de naissance. La conjonction de ces trois valeurs constitue un « quasi-identificateur » – et il en existe bien d'autres. Ce résultat n'est, en fait, pas si surprenant, quand on prend conscience qu'il y a 6 milliards de combinaisons possibles pour ces valeurs et seulement 300 millions de personnes aux États-Unis. Or supprimer l'une de ces informations interdirait aux chercheurs de faire des statistiques sur l'influence du lieu de résidence, du sexe ou de l'âge des patients, sur, par exemple, la fréquence d'une maladie.

santé-pratique-paris.fr

ENTRE DONNÉES EXPLOITABLES et anonymat complet, quel est le meilleur compromis ?

Les chercheurs en informatique proposent aujourd'hui des procédures plus complexes de protection de la vie privée des patients, fondées sur une détérioration – appauvrissement ou bruitage – des données. Il existe de nombreuses manières de détériorer des données, par exemple ne garder que l'année de naissance des patients ou modifier leur date de naissance de quelques jours, afin de rendre leur identification plus difficile. La question qui se pose désormais est celle du bon degré de détérioration des données : il faut suffisamment les appauvrir ou les bruiter pour rendre l'identification difficile, mais pas trop afin qu'elles gardent une pertinence.

Ce dilemme n'est en fait que le symptôme d'un autre, plus fondamental : dans l'état actuel de nos connaissances – qui peuvent certes encore progresser – nous ne savons pas garantir simultanément le parfait anonymat des patients et la meilleure efficacité de la recherche. Nous devons donc sacrifier une petite partie de l'un au profit de l'autre.

Quel compromis souhaitons-nous ? Naguère, les chercheurs avaient accès à très peu de données sur les patients. Le choix était donc fait, par l'état des techniques, en faveur de l'anonymat, au détriment de l'efficacité de la recherche. Mais aujourd'hui, du fait du développement des techniques de collecte des données, nous avons davantage le choix.

Continuer comme avant, ce qui reviendrait à interdire la communication des dossiers médicaux aux chercheurs, ou à les détériorer complètement, est possible. Mais ce compromis autrefois imposé par l'état des techniques n'est pas nécessairement celui que nous devons préférer. Nous pouvons aussi abandonner une petite partie de notre anonymat au profit d'une recherche plus efficace : il nous suffit, pour cela, de décider de moins détériorer les données. ■

Gilles DOWEK est chercheur chez Inria et membre du conseil scientifique de la Société informatique de France.