



SAÉ 6 parcours C
:Organisation et protection
des données

Anonymisation d'une base
de données



Rappel des compétences

Compétence	Aprentissage Critique	
C4	AC1	Capturer et stocker des ensembles volumineux et complexes de données hétérogènes
	AC2	Préparer et extraire les données pour l'exploitation
	AC3	Appliquer des méthodes d'exploration et d'exploitation des données (apprentissage, informatique décisionnelle ou fouille de données)
	AC4	Mettre en production et optimiser le système de gestion de données de l'entreprise
C5	AC1	Mesurer les impacts économiques, sociétaux et technologiques d'un projet informatique
	AC2	Savoir intégrer un projet informatique dans le système d'information d'une organisation
	AC3	Savoir adapter un système d'information
C6	AC1	Organiser et partager une veille numérique
	AC2	Identifier les enjeux de l'économie de l'innovation numérique
	AC3	Guider la conduite du changement informatique au sein d'une organisation

Les compétences et ressources mobilisées

	Compétence 4	Compétence 5	Compétence 6
Ressource S5	AC4	AC2	AC1 AC3
R6.01 Entrepreneuriat			X X
R6.02 Droit du numérique et PI			X X
R6.03 Comm. : Information			X X
R6.C.05 Administration BD	X	X	
R6.C.06 Aide à la décision			

Projet : Anonymisation de données

Vous travaillez pour le service informatique de protection des données de l'université de Rennes.

Une professeure de mathématique veut donner à ses étudiants une base de données sur les JO afin que ses élèves procèdent à son analyse.

Elle doit procéder à l'anonymisation des données, pour ce faire elle demande à votre équipe de s'en occuper.

Attention : le jeu de données anonymisé devra conservé intégralement son sens en vue de l'analyse statistique.

Qu'est-ce que l'anonymisation ?

Qu'est-ce que l'anonymisation ?

Anonymisation = altération (irréversible) du jeu de données afin de rendre impossible l'identification d'une personne ou d'une entité.

/! Bien faire la différence avec la Pseudo-anonymisation

Qu'est-ce que l'anonymisation ?

Anonymisation = altération (**irréversible**) du jeu de données afin de rendre impossible l'identification d'une personne ou d'une entité.

/! Bien faire la différence avec la Pseudo-anonymisation

Pseudo-anonymisation = altération du jeu de données de afin de rendre impossible l'identification d'une personne ou d'une entité **sans informations supplémentaires.**

Différents types d'anonymisation

Anonymisation **STATIQUE** = Après anonymisation la donnée d'origine est perdue. Il est impossible d'accéder aux données de l'individu/l'entité.

Anonymisation **DYNAMIQUE** = Les données restent en base mais sont anonymisées en fonction de l'utilisateur. Par exemple faire différentes vues en fonction du profil utilisateur.

Pourquoi (pseudo-)anonymiser des données ?

- Permet l'exploitation de données à caractères personnelles dans le respect du droit à la vie privée
- Compromis déontologique, juridique et éthique entre la protection de l'individu et l'utilisation de ses données

Exigence pour l'anonymisation

- Transformer les données sensibles
- Conserver la pertinence des données
 - Préserver le sens des données
 - Maintenir le bon fonctionnement d'une application
- Eliminer les données à caractère personnel non-nécessaires et/ou facilitant l'identification
- Complexité :
 - Identifier les données strictement nécessaires à protéger
 - Définir un degré d'acceptabilité pour la conservation des données

Des méthodes

Randomisation : modification d'attributs dans la base de données tout en conservant la répartition globale.

Permet de protéger le jeu de données du risque d'inférence.

Exemple : permutation de données

Généralisation : Modification de l'échelle d'attributs de la base de données afin qu'ils soient communs à un ensemble de personnes.

Permet d'éviter l'individualisation et limite les corrélations.

Exemple : remplacement de la date de naissance par l'année uniquement.

Vérifier l'anonymisation

3 critères définies par les autorités de protection des données européennes permettent de s'assurer de l'anonymisation.

L'individualisation : il ne doit pas être possible d'isoler un individu dans le jeu de données

La corrélation : il ne doit pas être possible de relier entre eux des ensembles de données distincts concernant un même individu ;

L'inférence : il ne doit pas être possible de déduire, de façon quasi certaine, de nouvelles informations sur un individu.

Programme des 3 semaines

- Ressources 5 h
 - R6.02 Droit du numérique Bertrand DE VILLENEUVE 1 h
 - R6.03 Communication Elisabeth GENAIVRE 2 h
 - R6.C.05 Administration BD Quentin LEMESLE 2 h
- Projet 10 h (dont 1h TP noté)
- Autonomies 30 h
- Soutenance 1 h

Remarques : vous avez aussi des cours non associés à la SAÉ cette semaine



Évaluations

Ressources:

- Droit, Bertrand DE VILLENEUVE, coefficient 1
- Communication, Elisabeth GENAIVRE, coefficient 1
- BDD (évaluation individuelle), Constance THIERRY, coefficient 1.5

Projet : (coefficient 1.5)

- Rapport, rendu 06/06/2024 12 h
- Oral 15 min pour échanger sur le rapport

Livrables et évaluation

R6.C.05 Administration BD : TP noté (individuel) mercredi 05/06/2024
Trigger/Fonctions/vues pour l'anonymisation

Rapport et Oral

Rapport explicitant les changements effectués pour l'anonymisation et les procédures effectuées (max 20 pages)

Rappel : un rapport contient une introduction et une conclusion...

Schéma des bases comparés avant et après anonymisation

Critères d'évaluation du rapport

- Introduction / conclusion
- Présentation de l'existant, identification des données sensibles et définition des objectifs
- Démarches pour l'atteinte des objectifs (randomisation, généralisation)
- Résultats obtenus (vérification individualisation, corrélation, inférence)
- Répartition et organisation du travail
- Choix des illustrations et leur intégration au rapport
- Orthographe/Expression
- Réponses aux questions lors de la soutenance