Data Science 6 : sécurité des données – RGPD

J. Delpech – Cours : Data Science
Cursus M1 Data/IA – 2025/2026
Dernière mise à jour : février 2026

1. Introduction : pourquoi la sécurité des données ?

Objectif : Comprendre les enjeux de la sécurité des données et maîtriser les fondamentaux du RGPD dans un contexte Data Science

1.1 Un contexte de scandales et de fuites

- **Cambridge Analytica (2018)** : exploitation des données de 87 millions d'utilisateurs Facebook à des fins de manipulation électorale
- **Equifax (2017)** : fuite des données personnelles de 147 millions de personnes (numéros de sécurité sociale, permis de conduire...)
- **Yahoo (2013-2014)** : 3 milliards de comptes compromis
- **France Travail (2024)** : 43 millions de personnes potentiellement concernées

Ces incidents illustrent que la donnée est devenue un actif stratégique, mais aussi une cible privilégiée.

1.2 Les enjeux pour les organisations

Dimension	Enjeux
Économique	Valeur des données, coût des incidents (en moyenne 4,45 M$ par breach en 2023)
Réputationnel	Perte de confiance des clients et partenaires
Juridique	Sanctions, poursuites, responsabilité civile et pénale
Opérationnel	Interruption d'activité, perte de données critiques

1.3 La triade CIA

Les trois piliers fondamentaux de la sécurité de l'information :

- **Confidentialité** (*Confidentiality*) : seules les personnes autorisées accèdent aux données
- **Intégrité** (*Integrity*) : les données ne sont pas altérées de manière non autorisée
- **Disponibilité** (*Availability*) : les données sont accessibles quand on en a besoin

> Note: En Data Science, on ajoute souvent la **traçabilité** : pouvoir reconstituer l'historique des traitements effectués sur les données.

1.4 La responsabilité du Data Scientist

Le Data Scientist occupe une position particulière dans la chaîne de traitement :

- Accès à des volumes importants de données, souvent sensibles
- Capacité technique de croiser, enrichir, inférer des informations
- Création de modèles pouvant avoir des impacts sur les personnes (scoring, profilage...)
- Responsabilité éthique au-delà de la simple conformité légale

2. Le RGPD : principes fondamentaux

2.1 Qu'est-ce que le RGPD ?

Le **Règlement Général sur la Protection des Données** (RGPD / GDPR) est entré en application le 25 mai 2018. C'est un règlement européen directement applicable dans tous les États membres.

Champ d'application : toute organisation (européenne ou non) qui traite des données personnelles de résidents de l'UE.

> Note : Autres réglementations similaires : CCPA (Californie), LGPD (Brésil), PIPL (Chine)...

2.2 Données personnelles et données sensibles

Donnée personnelle : toute information permettant d'identifier directement ou indirectement une personne physique.

Directement identifiantes	Indirectement identifiantes
Nom, prénom	Adresse IP
Numéro de sécurité sociale	Identifiant client
Photo du visage	Données de géolocalisation
Email nominatif	Combinaison âge + code postal + profession

2.2 Données personnelles et données sensibles

Données sensibles (article 9) — traitement interdit par défaut :

- Origine raciale ou ethnique
- Opinions politiques, convictions religieuses ou philosophiques
- Appartenance syndicale
- Données génétiques ou biométriques
- Données de santé
- Vie sexuelle ou orientation sexuelle

Note : En Data Science, attention aux **inférences** : un modèle peut déduire des données sensibles à partir de données apparemment anodines.

2.3 Les 6 principes clés du RGPD

Principe	Description	Implication en Data Science
Licéité, loyauté, transparence	Base légale valide, information claire des personnes	Documenter et justifier chaque traitement
Limitation des finalités	Collecte pour des objectifs déterminés et légitimes	Pas de réutilisation sans vérification de compatibilité
Minimisation	Collecter uniquement les données nécessaires	Sélection rigoureuse des features
Exactitude	Données correctes et à jour	Processus de validation et mise à jour
Limitation de conservation	Durée proportionnée à la finalité	Politique de suppression, archivage
Intégrité et confidentialité	Mesures de sécurité appropriées	Chiffrement, contrôle d'accès, audit

2.4 Les bases légales du traitement

Pour traiter des données personnelles, il faut une base légale parmi les 6 suivantes :

1. **Consentement** : libre, spécifique, éclairé, univoque (et retirable)
2. **Contrat** : exécution d'un contrat avec la personne
3. **Obligation légale** : respect d'une obligation imposée par la loi
4. **Intérêts vitaux** : protection de la vie de la personne
5. **Mission d'intérêt public** : exécution d'une mission de service public
6. **Intérêt légitime** : intérêt du responsable de traitement (avec balance des intérêts)

Note : En recherche et Data Science, on s'appuie souvent sur l'**intérêt légitime** ou le **consentement**, mais l'intérêt légitime nécessite une analyse de balance des intérêts documentée.

2.5 Les droits des personnes

Droit	Description
Accès	Obtenir confirmation du traitement et copie des données
Rectification	Corriger des données inexactes
Effacement ("droit à l'oubli")	Suppression sous certaines conditions
Limitation	Gel du traitement en cas de contestation
Portabilité	Récupérer ses données dans un format réutilisable
Opposition	S'opposer au traitement (notamment profilage)

Note : Ces droits doivent pouvoir être exercés facilement. Le délai de réponse est d'un mois maximum.

2.6 Privacy by Design & Privacy by Default

Privacy by Design : intégrer la protection des données dès la conception d'un projet.

- Réfléchir aux données nécessaires avant de collecter
- Prévoir les mécanismes de suppression dès le départ
- Intégrer la sécurité dans l'architecture

Privacy by Default : par défaut, le niveau de protection doit être maximal.

- Pas de collecte excessive "au cas où"
- Paramètres les plus protecteurs activés par défaut
- Accès restreint au strict nécessaire

3. Acteurs et gouvernance

3.1 Les rôles clés

Acteur	Définition	Responsabilités
Responsable de traitement	Détermine les finalités et moyens du traitement	Conformité globale, choix des sous-traitants
Sous-traitant	Traite les données pour le compte du responsable	Respecter les instructions, garantir la sécurité
DPO (Délégué à la Protection des Données)	Référent interne sur la protection des données	Conseil, contrôle, interface avec la CNIL

Note : Le DPO est obligatoire pour les organismes publics et les traitements à grande échelle de données sensibles ou de surveillance systématique.

3.2 Obligations documentaires

Registre des traitements : document obligatoire recensant tous les traitements de données personnelles.

Contenu minimal :

- Nom et coordonnées du responsable
- Finalités du traitement
- Catégories de personnes et de données
- Destinataires
- Transferts hors UE éventuels
- Durées de conservation
- Mesures de sécurité

3.2 Obligations documentaires

Analyse d'Impact (DPIA) : obligatoire pour les traitements à risque élevé.

Critères déclencheurs (si 2 ou plus) :

- Évaluation/scoring
- Décision automatique avec effet juridique
- Surveillance systématique
- Données sensibles à grande échelle
- Croisement de données
- Personnes vulnérables
- Usage innovant de technologies
- Exclusion d'un droit/contrat

3.3 Notification des violations

En cas de violation de données (breach), le responsable de traitement doit :

1. **Notifier la CNIL** dans les **72 heures** (sauf risque improbable pour les droits des personnes)
2. **Informer les personnes concernées** si risque élevé pour leurs droits et libertés

Note : Une violation peut être une fuite, mais aussi une perte d'accès ou une altération non autorisée.

3.4 Sanctions

Amendes administratives :

- Jusqu'à **10 M€** ou 2% du CA mondial pour certains manquements
- Jusqu'à **20 M€** ou 4% du CA mondial pour les violations les plus graves

Exemples récents :

- Amazon (Luxembourg, 2021) : 746 M€
- Meta (Irlande, 2023) : 1,2 Md€
- Google (France, 2022) : 150 M€

Au-delà des amendes : sanctions pénales, actions de groupe, impact réputationnel majeur.

4. Risques et menaces en Data Science

4.1 Panorama des risques

Risques externes :

- **Phishing** : usurpation d'identité pour obtenir des accès
- **Ransomware** : chiffrement des données avec demande de rançon
- **Injection SQL** : exploitation de failles pour accéder aux bases
- **Man-in-the-middle** : interception des communications

Risques internes :

- **Erreur humaine** : envoi de fichiers au mauvais destinataire, mauvaise configuration
- **Négligence** : mots de passe faibles, données non chiffrées
- **Malveillance** : employé mécontent, espionnage industriel

4.2 Vulnérabilités spécifiques à la Data Science

Vulnérabilité	Exemple	Mitigation
Datasets publics avec données sensibles	Datasets Kaggle mal anonymisés	Vérifier l'origine et la conformité
Credentials exposés	API keys dans notebooks Git	Utiliser des secrets managers, .gitignore
Notebooks partagés	Jupyter avec données en clair	Anonymiser avant partage
Model inversion attacks	Reconstruction de données d'entraînement	Differential privacy, audits
Absence de traçabilité	Impossible de savoir qui a accédé à quoi	Logging systématique

Exemple : usage de `dotenv`

1. Créer un ficher `.env` et le faire figurer dans le `.gitignore`
2. Définir la clef API dans `.env` :

```
API_KEY = a25!ldk*skfj$23f34
```

3. Charger la clef avec `dotenv` :

```
from dotenv import load_dotenv

load_dotenv()
api_key = os.environ["API_KEY"]
client = MyAPI(api_key=api_key)
client.get(request)
```

4.3 Le risque de ré-identification

Des données apparemment anonymes peuvent permettre de ré-identifier des personnes par croisement.

Cas célèbre :

- En 2006, AOL publie 20 millions de requêtes de recherche "anonymisées". Des journalistes identifient rapidement l'utilisateur n°4417749 comme Thelma Arnold, 62 ans, grâce à ses recherches sur son nom, sa ville, ses problèmes de santé…
- DEFCON 17 : Svea Eckert et Andreas Dewes analysent les données de connexion (publiques) issues de 10 extensions Chrome pour 3M d'allemands et on pu relier des informations très précises à des individus (juge, homme politique, enquête de police…)

Quasi-identifiants : combinaison de variables qui, ensemble, identifient une personne.

Voir :

- « fingerprinting »,
- cours sur ACP (plus le nombre de dimension augmente, plus les points sont singuliers

4.4 Biais et discrimination

Les modèles de ML peuvent amplifier ou créer des discriminations :

- Données d'entraînement biaisées historiquement
- Variables proxy pour des caractéristiques protégées (code postal → origine ethnique)
- Manque de représentativité dans les datasets

Exemple : algorithme de recrutement d'Amazon pénalisant les candidatures féminines (entraîné sur 10 ans de CV... majoritairement masculins).

5. Mesures de protection

5.1 Mesures techniques

Chiffrement :

- Au repos (données stockées) : AES-256
- En transit (données transmises) : TLS 1.3
- De bout en bout pour les données très sensibles

Contrôle d'accès :

- Principe du moindre privilège
- Authentification forte (MFA)
- Gestion des identités (IAM)
- Revue régulière des droits

Journalisation :

- Qui a accédé à quelles données, quand, pourquoi
- Conservation sécurisée des logs
- Alertes sur comportements anormaux

5.2 Anonymisation vs Pseudonymisation

K-anonymat : chaque enregistrement est indistinguable d'au moins k-1 autres pour les quasi-identifiants.

Exemple avec k=2 :

Âge	Code postal	Maladie
30-40	75*	Diabète
30-40	75*	Grippe
30-40	75*	Diabète

L-diversité : au sein de chaque groupe k-anonyme, au moins l valeurs différentes pour l'attribut sensible.

T-closeness : la distribution de l'attribut sensible dans chaque groupe est proche de la distribution globale.

Differential Privacy : ajout de bruit calibré pour garantir mathématiquement qu'une requête ne révèle pas si un individu est dans le dataset.

5.4 Techniques de pseudonymisation

Hachage (hashing) :

```
import hashlib

def pseudonymiser_email(email):
    return hashlib.sha256(email.encode()).hexdigest()[:16]

# "jean.dupont@email.fr" → "a7b3c9d2e5f1..."
```

Le hachage simple est vulnérable aux attaques par dictionnaire. Utiliser la méthode du **salage (salt)** (valeur secrète ajoutée).

Tokenisation : remplacement par un jeton aléatoire avec table de correspondance sécurisée.

5.4 Techniques de pseudonymisation

Tokenisation : remplace une donnée sensible par un **jeton aléatoire** (token) sans lien mathématique avec la donnée originale.

La correspondance est stockée dans une **table sécurisée** séparée.

DONNÉES OPÉRATIONNELLES (accessible aux équipes)

Token	Autres données
TKN_8f4e2a	Achat: 150€
TKN_3c7d9e	Achat: 89€

TABLE DE CORRESPONDANCE (vault sécurisé, accès restreint)

Token	Valeur originale
TKN_8f4e2a	4532-XXXX-XXXX-1234
TKN_3c7d9e	4916-XXXX-XXXX-5678

Tokenisation vs. Hachage

Aspect	Hachage (« salé » ou non)	Tokenisation
Réversibilité	Impossible (one-way)	Possible via la table
Lien mathématique	Oui (déterministe)	Non (aléatoire)
Cas d'usage	Pseudonymisation définitive	Besoin de retrouver la donnée
Exemple	Anonymiser pour analytics	Paiements (besoin de débiter la carte)

5.5 Mesures organisationnelles

- **Politique de sécurité** formalisée et communiquée
- **Formation** régulière des équipes
- **Clauses contractuelles** avec les sous-traitants (DPA)
- **Procédures de gestion des incidents** testées
- **Audits** et tests d'intrusion périodiques
- **Gouvernance** claire avec rôles et responsabilités définis

6. Démonstration pratique -> notebook

7. Méthodologie projet : intégrer la conformité

7.1 Check-list RGPD pour un projet Data Science

Phase de cadrage :

- ☐ Quelle est la finalité précise du traitement ?
- ☐ Quelle base légale s'applique ?
- ☐ Quelles données sont réellement nécessaires (minimisation) ?
- ☐ Existe-t-il des données sensibles ?
- ☐ Une DPIA est-elle requise ?

Phase de collecte/acquisition :

- ☐ Les personnes sont-elles informées ?
- ☐ Le consentement est-il valide (si applicable) ?
- ☐ Les données proviennent-elles de sources licites ?
- ☐ Les contrats avec les fournisseurs incluent-ils les clauses RGPD ?

7.1 Check-list RGPD pour un projet Data Science

Phase de traitement :

- ☐ L'accès aux données est-il limité au strict nécessaire ?
- ☐ Les données sont-elles pseudonymisées/anonymisées si possible ?
- ☐ Les traitements sont-ils tracés ?
- ☐ Les mesures de sécurité sont-elles en place ?

Phase de conservation/suppression :

- ☐ La durée de conservation est-elle définie et justifiée ?
- ☐ Un processus de suppression est-il en place ?
- ☐ Les droits des personnes peuvent-ils être exercés ?

7.2 Modèle de fiche de traitement simplifiée

Cf. fichier pdf

7.3 Matrice d'analyse des risques

Risque	Probabilité (1-4)	Impact (1-4)	Criticité	Mesures
Fuite de données			P × I	
Accès non autorisé				
Ré-identification				
Non-conformité				
Biais discriminatoire				

Échelle :

- Probabilité : 1 (rare) → 4 (très probable)
- Impact : 1 (mineur) → 4 (catastrophique)
- Criticité : 1-4 (faible), 5-8 (modéré), 9-12 (élevé), 13-16 (critique)

8. Ressources et références

Sites officiels

- **CNIL** : cnil.fr — Guides, modèles, actualités
- **RGPD texte officiel** : eur-lex.europa.eu
- **EDPB** (Comité européen) : edpb.europa.eu — Lignes directrices

Outils

- **PIA** (Privacy Impact Assessment) : outil CNIL pour les analyses d'impact
- **ARX** : outil open source d'anonymisation
- **hashlib** : bibliothèque Python standard pour le hachage

Pour aller plus loin

- Guide CNIL "RGPD : guide du développeur"
- Cours en ligne MOOC CNIL "L'atelier RGPD"
- La bibliothèque du DPO (AFCDP)

9. Exercice : prédiction de risque de crédit

Contexte : Une banque souhaite développer un modèle de ML pour prédire le risque de défaut de paiement des demandeurs de crédit.

Données disponibles :

- Identité (nom, prénom, date de naissance)
- Coordonnées (adresse, téléphone, email)
- Situation professionnelle (employeur, revenus, ancienneté)
- Historique bancaire (solde moyen, incidents de paiement)
- Situation familiale (situation matrimoniale, personnes à charge)

Questions :

1. **Identification** : Quelles sont les données personnelles ? Les données sensibles potentielles ?
2. **Base légale** : Quelle base légale semble la plus appropriée ?
3. **Risques** : Identifiez au moins 3 risques majeurs (sécurité, discrimination, conformité).
4. **Mesures** : Proposez des mesures techniques et organisationnelles adaptées.
5. **Anonymisation** : Si vous deviez publier une étude statistique, comment anonymiseriez-vous les données ?
6. **Droits** : Comment garantiriez-vous le droit d'accès et le droit d'opposition ?
7. **Éthique** : Quels biais potentiels pourrait contenir le modèle ? Comment les atténuer ?