HEALTH DATA HUB

Données de synthèse

-

Méthodes de génération et limites

Sommaire

1. Introduction

Définition
Intérêt
Cas d’usage

1.1 Introduction: que sont les données synthétiques?

Définition

On appelle “données synthétiques” des données générées de façon aléatoire répliquant certaines propriétés d’une certaine base de données.

Propriétés réplicables

- Schéma (tables, clés de jointure, type des variables)
- Propriétés statistiques (distributions univariées, corrélations, …)
- Cohérence (règles métiers,…)

1.2 Pourquoi des données synthétiques?

“What I cannot create,

I do not understand.”

Motivations

- Restrictions d’accès
- Gain de temps (écrire les programmes sur les données synthétiques)
- Evaluer l’intérêt d’une base de données pour une étude
- Intérêt pédagogique

1.3 Cas d’usage

Cas n°1

Je ne sais pas si la base de données peut répondre à ma problématique

Cas n°2

Je souhaite me former à l’utilisation d’une certaine base de données, sans forcément attendre des résultats valables d’un point de vue statistique

Cas n°3

J’aurais bientôt accès à la base de données réelle et je souhaite commencer à écrire mes programmes

Cas n°4

Je n’ai pas le droit d’avoir accès aux données réelles mais je souhaite tout de même conduire une étude

2. Méthodes de génération

Sans accès à la base de données réelle
Confidentialité et avatarisation
Méthodes de deep learning

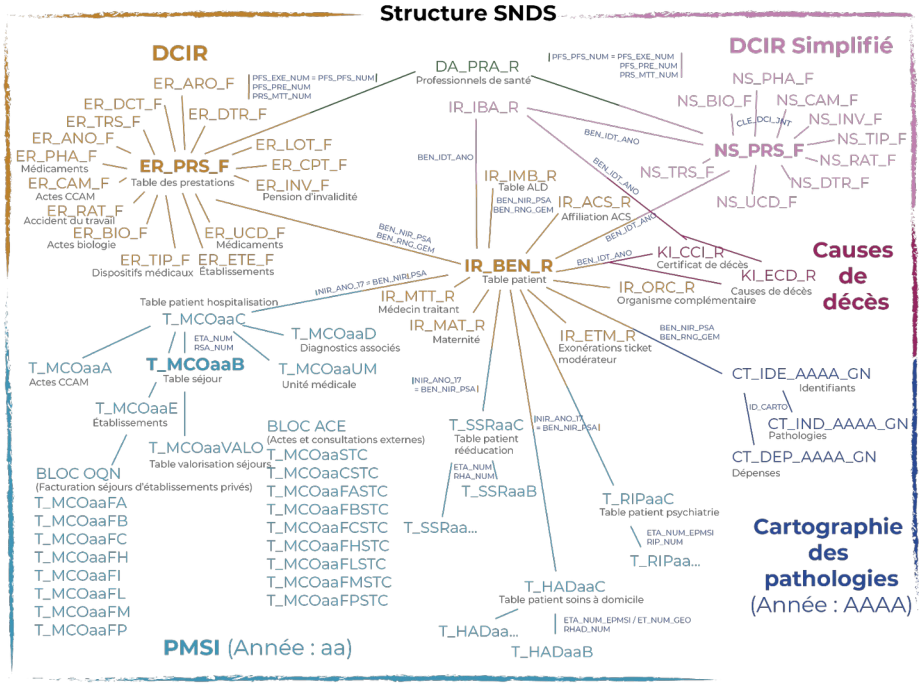
2. Méthodes de génération

Sans accès à la base de données réelle
Confidentialité et avatarisation
Méthodes de deep learning

2.1 Sans avoir accès à la base de données réelle

Un schéma

- liste des tables
- clés de jointure
- type des variables



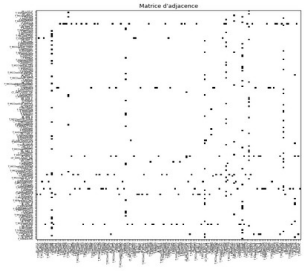
2.1 Sans avoir accès à la base de données réelle


```
{
    "fields": [
        {
            "name": "BEN_NIR_PSA",
            "description": "Identifiant anonyme du patient dans le SNIIRAM",
            "type": "string",
            "nomenclature": "-",
            "length": "17",
            "format": "default",
            "constraints": {
                "minLength": 17,
```


Un point de départ: le schéma formel

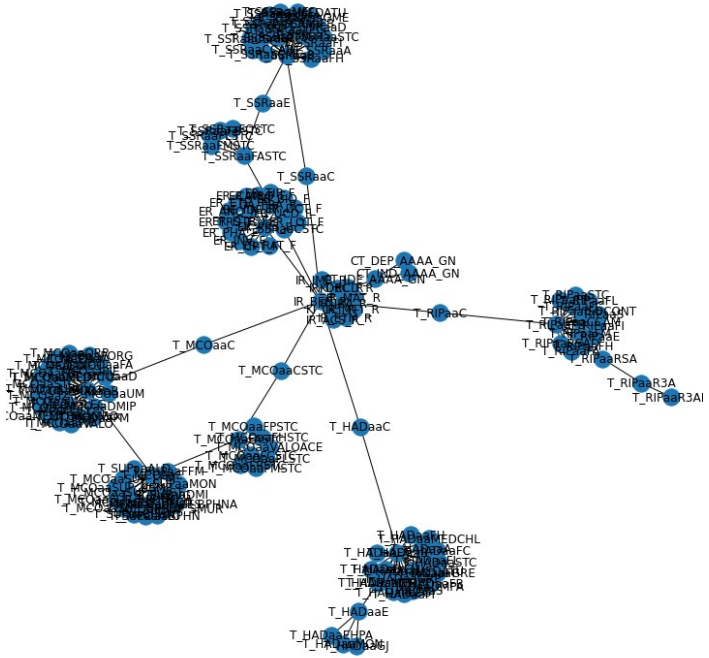
source	target	joint_var
T_HADaaFI	T_HADaaB	ETA_NUM_EPMSI + RHAD_NUM
T_HADaaFL	T_HADaaB	ETA_NUM_EPMSI + RHAD_NUM
T_HADaaFM	T_HADaaB	ETA_NUM_EPMSI + RHAD_NUM
T_HADaaFP	T_HADaaB	ETA_NUM_EPMSI + RHAD_NUM
T_HADaaGJ	T_HADaaE	ETA_NUM_EPMSI => ETA_NUM

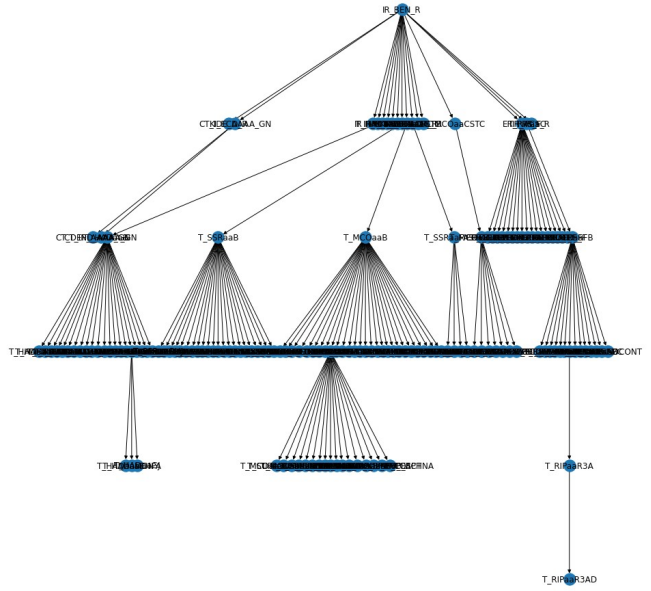
Du schéma formel aux liens entre les tables



Des liens à la matrice d'adjacence

2.1 Sans avoir accès à la base de données réelle

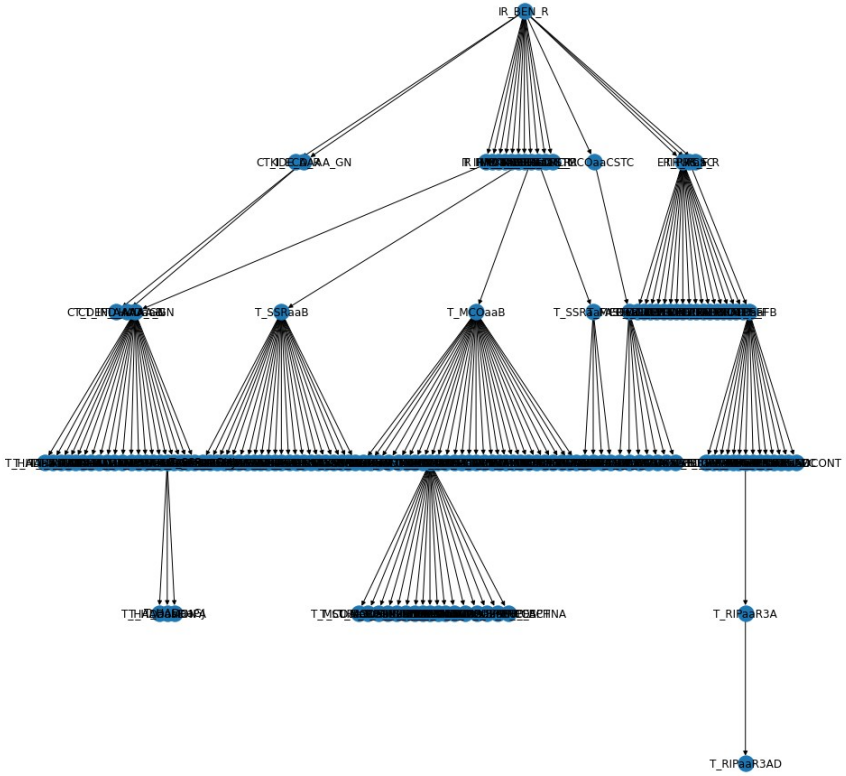




De la matrice d’adjacence, un graphe

Du graphe, un arbre

2.1 Sans avoir accès à la base de données réelle



Génération colonne par colonne

2.1 Sans avoir accès à la base de données réelle

ETA_NUM	RSA_NUM
9000325451	151242
9000646567	135868

Récupérés depuis T_MCOaaC (table parent)

ETA_NUM	RSA_NUM	AGE_ANN
9000325451	151242	65
9000646567	135868	42

Générés aléatoirement

ETA_NUM	RSA_NUM	AGE_ANN	...	DGN_PAL
9000325451	151242	65	...	X34018
9000646567	135868	42	...	S2200

Générés aléatoirement

2.1 Sans avoir accès à la base de données réelle

Premier levier de réalisme: La cohérence temporelle

On peut s'assurer que l'**ordre** des dates est cohérent, par exemple qu'une sortie d'hospitalisation a bien lieu après l'entrée correspondante.

→ comment trouver les couples de variables qui se correspondent?

Notre proposition: une recherche du **plus proche voisin**, avec une distance calculée à partir du nom de la variable et de sa description

Pour $x_1 = (var_1, desc_1)$ et $x_2 = (var_2, desc_2)$:

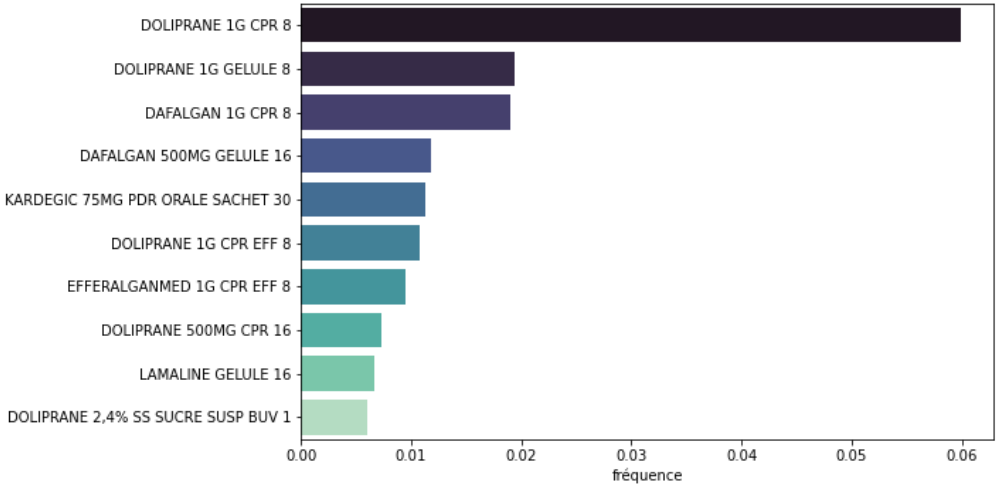
$$d(x_1, x_2) = \frac{1}{Z} Levenshtein(var_1, var_2) + 1 - \frac{<bow(desc_1), bow(desc_2)>}{||bow(desc_1)||_2 . ||bow(desc_2)||_2}$$

2.1 Sans avoir accès à la base de données réelle

Deuxième levier de réalisme: statistiques descriptives
Plutôt que de générer nos colonnes aléatoirement, on peut utiliser des données en open data pour se rapprocher des distributions réelles, par exemple des médicaments.

⚠ En revanche, introduire des corrélations est beaucoup plus difficile:

- ➢ les données de corrélation ne sont pas toujours en open data
- ➢ les corrélations sont souvent au niveau du parcours du patient, et pas au niveau tabulaire



2. Méthodes de génération

Sans accès à la base de données réelle
Confidentialité et avatarisation
Méthodes de deep learning

2.2.1 Risques liés à la confidentialité

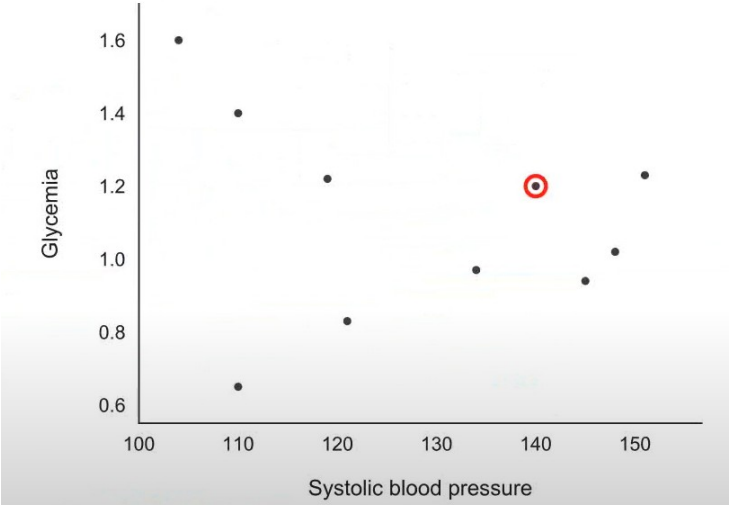
Nouveau paradigme: on a accès à la base de données réelle

- on peut exploiter de l'information contenue dans les données
  - *ex: perturbation de données existantes, deep learning, …*

- ⚠ on risque d'enfreindre les contraintes de confidentialité
  - *particulièrement risqué si le jeu de données de synthèse est destiné à l'open data*
  - *potentiellement difficile d'évaluer le niveau de risque représenté par un modèle, en particulier avec les modèles de deep learning (voir differential privacy)*
  - *arbitrage plausibilité/ confidentialité*

2.2.2 Avatarisation

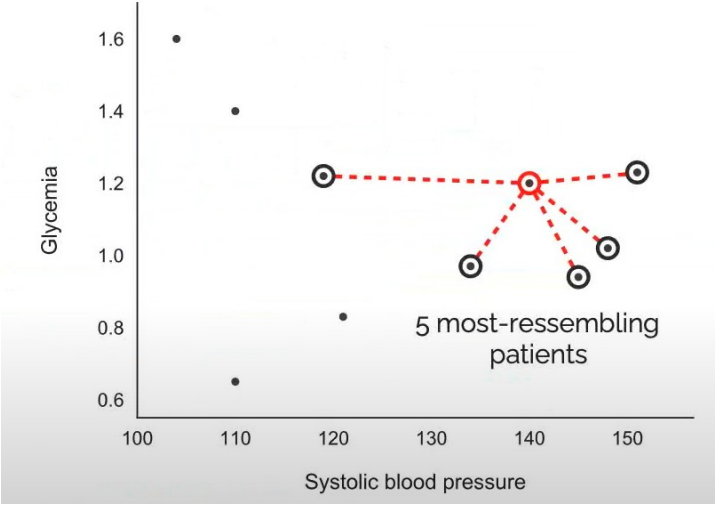
Technique développée et brevetée par la société **Octopize** - **Mimethik Data**



1ère étape - projection des données dans un espace euclidien (embedding)

2.2.2 Avatarisation

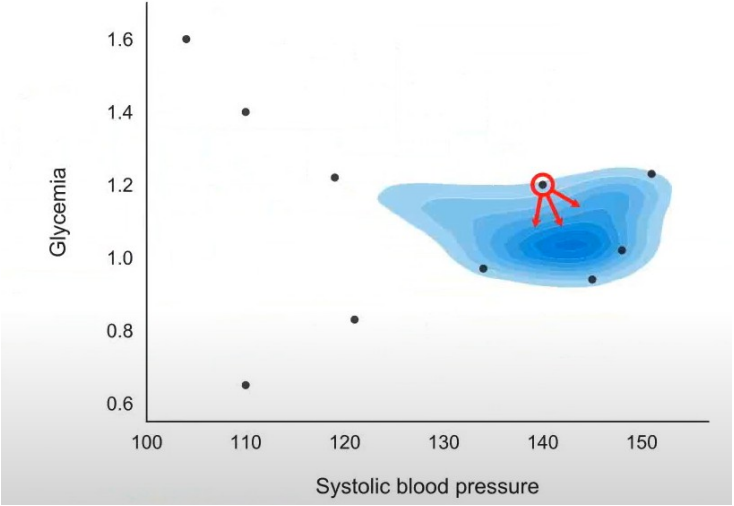
Technique développée et brevetée par la société **Octopize** - **Mimethik Data**



2ème étape - repérage des *n* plus proches voisins

2.2.2 Avatarisation

Technique développée et brevetée par la société **Octopize** - **Mimethik Data**



3ème étape - génération aléatoire de l'avatar

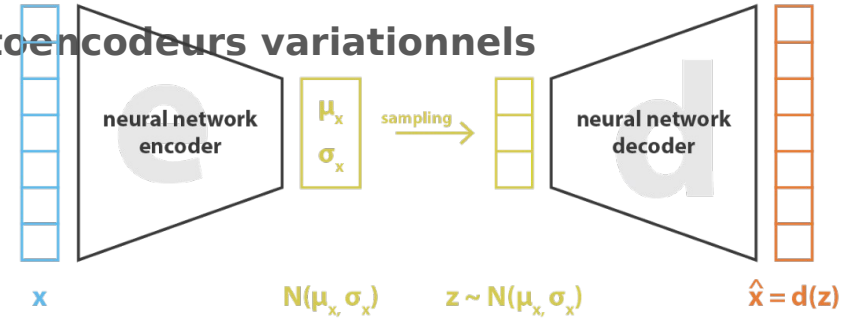
2. Méthodes de génération

Sans accès à la base de données réelle
Confidentialité et avatarisation
Méthodes de deep learning

2.3 Méthodes de deep learning

Paradigme: accès à la base de données réelle, pas de risques liés à la confidentialité

Solution n°1: Autoencodeurs variationnels



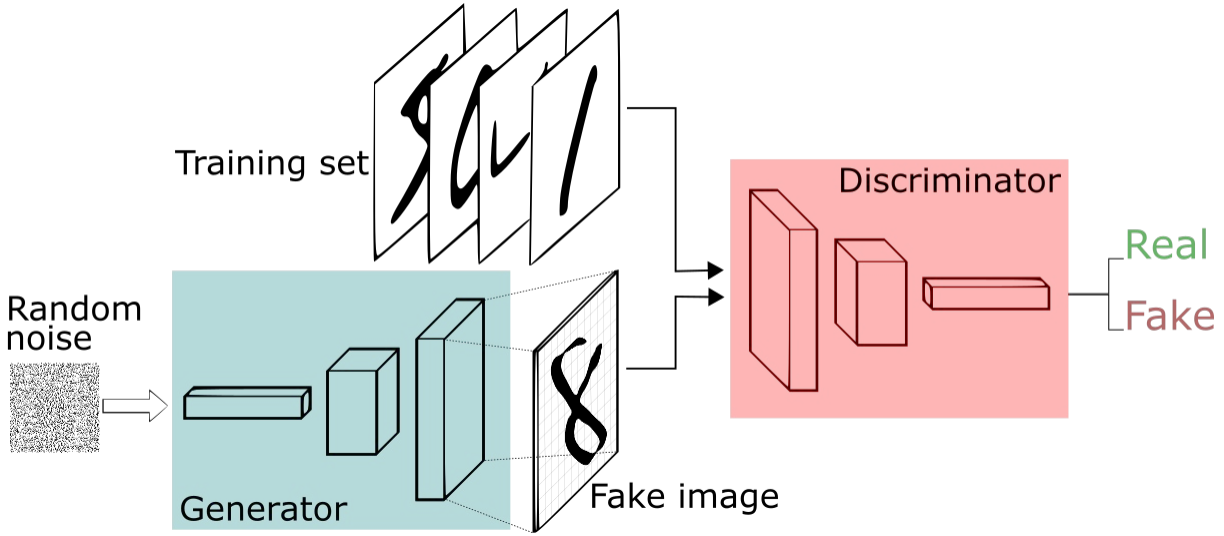
$$\text{loss} = \|x - \hat{x}\|^2 + KL[N(\mu_x, \sigma_x), N(0, I)] = \|x - d(z)\|^2 + KL[N(\mu_x, \sigma_x), N(0, I)]$$

Avantages

- Permet de reproduire toute la diversité des données d'entraînement
- La représentation obtenue peut être utilisée pour d'autres tâches

2.3 Méthodes de deep learning

Solution n°2: Réseaux adverses génératifs (GANs)



2.3 Méthodes de deep learning

Avantages

- Peut fournir d'excellents résultats selon les contextes
- Littérature en forte croissance (beaucoup d'*attention* portée à ces modèles depuis leur création en 2014)

Inconvénients

- Difficile à entraîner (potentiellement instable)
- *Mode collapse* → ne représente pas la diversité de la distribution réelle
- Règles métiers inférées (contraintes probabilistes)
- Typiquement un modèle où il est difficile d'évaluer le niveau de confidentialité

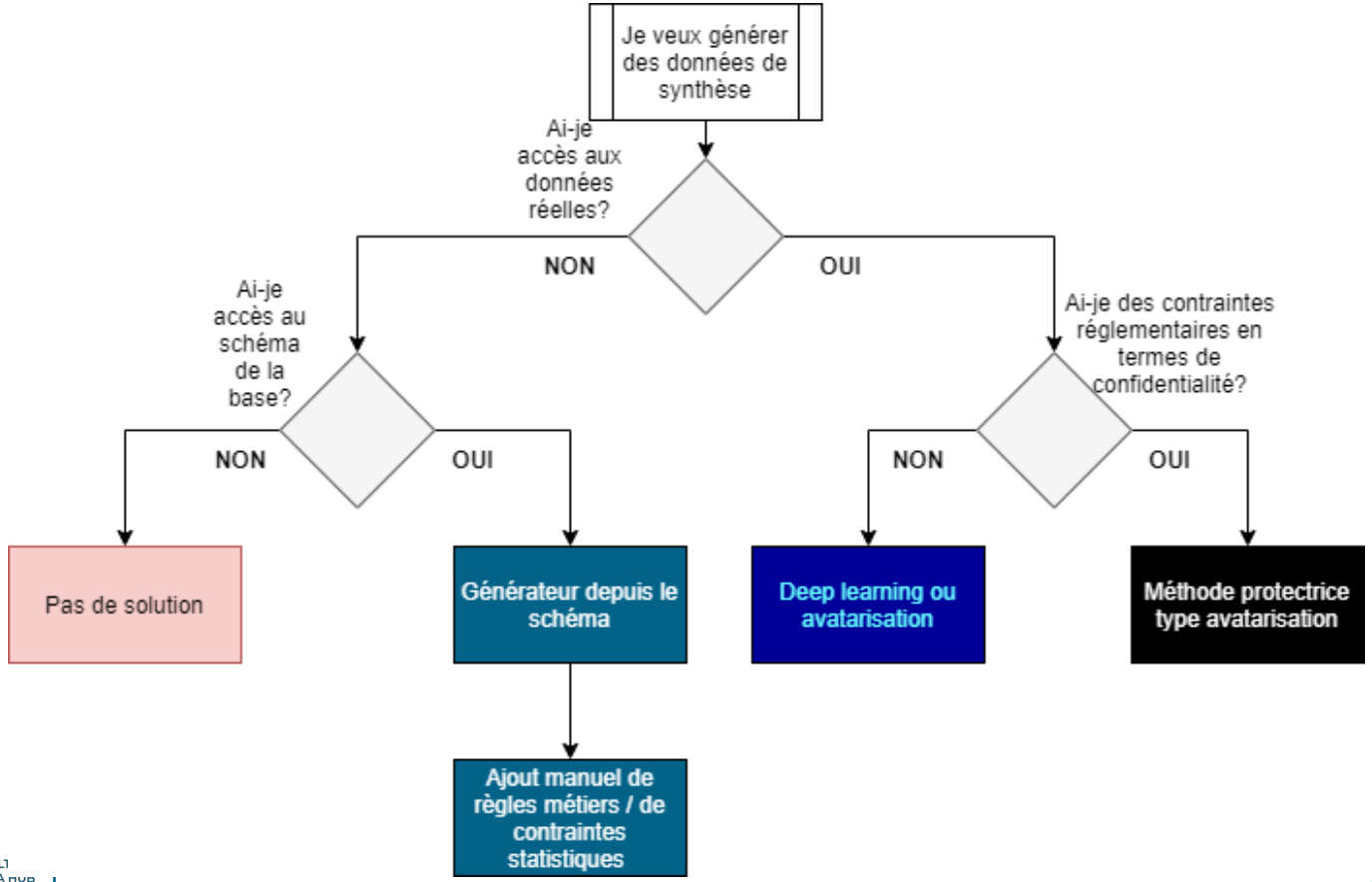
Pour aller plus loin

- Certaines méthodes permettent de combattre l'effet *mode collapse.* Voir notamment la littérature sur les Wasserstein GAN
- Certaines techniques permettent d'assurer un certain niveau de confidentialité. Voir la littérature sur la differential privacy

3. Conclusion

Bien choisir sa méthode de génération et en évaluer les limites

3. Bien choisir sa méthode de génération



3. Limites des différentes méthodes

Génération depuis le schéma

Pas ou peu de vraisemblance statistique

Avatarisation

Nécessite d’avoir accès aux données réelles

Nécessite un wrapper pour revenir au schéma de la base

La phase d’embedding peut être compliquée (eg parcours patients)

Deep learning

Nécessite d’avoir accès aux données réelles

Contraintes de confidentialité

Nécessite un wrapper pour revenir au schéma de la base

Bibliographie

- Goodfellow et al., **Generative adversarial networks**
- Kingma et al. **Auto-encoding variational Bayes**
- Arjovksy et al. **Wasserstein GAN**
- Dwork, **The Algorithmic Foundations of Differential privacy**
- Choi et al., **Generating Multi-label Discrete Patient Records using Generative Adversarial Networks**
- Xu et al. , **Modeling tabular data with conditional GAN**
- Yale et al, **Privacy preserving Synthetic Health Data**
- Van der Schaar et al, **Time-Series generative adversarial networks**
- Tout le travail de Mihaela van der Schaar (voir notamment le challenge *Hide-and-Seek* NeurIPS 2020)
- ...