Création de données synthétiques pour la valorisation des données

Mouvement Desjardins
Adel Benlagra, Chef d'équipe
Antoine Langevin, Scientifique de données

Atelier de Résolution de Problèmes Industriels
21 août 2023

Desjardins

Structure de la présentation

1 Contexte de la problématique

2 Les données tabulaires synthétiques

3 Approches de génération de données tabulaires

4 La problématique

Annexe

Le Mouvement Desjardins

- **1er groupe financier coopératif** en Amérique du Nord avec une gamme complète de services financiers et d'assurance
- Institution financière la plus présente au Québec et est bien établie en Ontario offrant ses services à 121 villes et villages et près de **8 millions de membres et client(e)s**

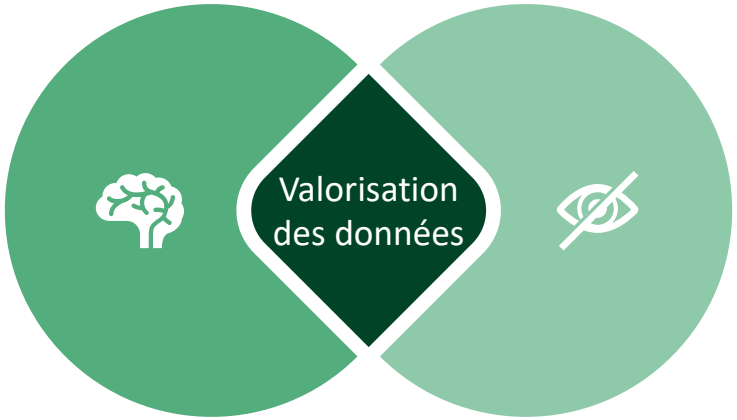
- Desjardins gère un **actif de 407,7 G$** au 31 décembre 2022 et a versé **518 M$ en ristourne aux membres et à la collectivité**
- Parmi les **100 meilleurs employeurs au Canada** en 2022 selon Mediacorp Canada
- Positionné au **4e rang mondial et au 1er rang au Canada** du classement des entreprises **favorables aux femmes** établi par le magazine *Forbes*

Contexte du projet

L'utilisation des données est au cœur de deux orientations stratégiques, en apparence difficile à concilier, du Mouvement

Création de valeur grâce à l'analytique avancée

Exploitation maximale de toutes les informations pertinentes



Engagement envers la conformité et la sécurité

L'utilisation responsable des données peut limiter l'accès à certaines données sensibles

Il existe plusieurs approches pour la protection des données sensibles

Dénominalisation

Retrait ou altération de certains identifiants directs ou indirects permettant, **prises en groupe**, de réidentifier une personne

Dépersonnalisation

Retrait ou altération de tout identifiant direct permettant d'identifier une personne.
L'identification n'est possible qu'en ayant recours à d'autres informations

Anonymisation

Retrait ou altération de tout identifiant direct et indirect permettant de réidentifier une personne
En principe, irréversible.

Techniques: Tokénisation, encryption, *masquage,* etc.

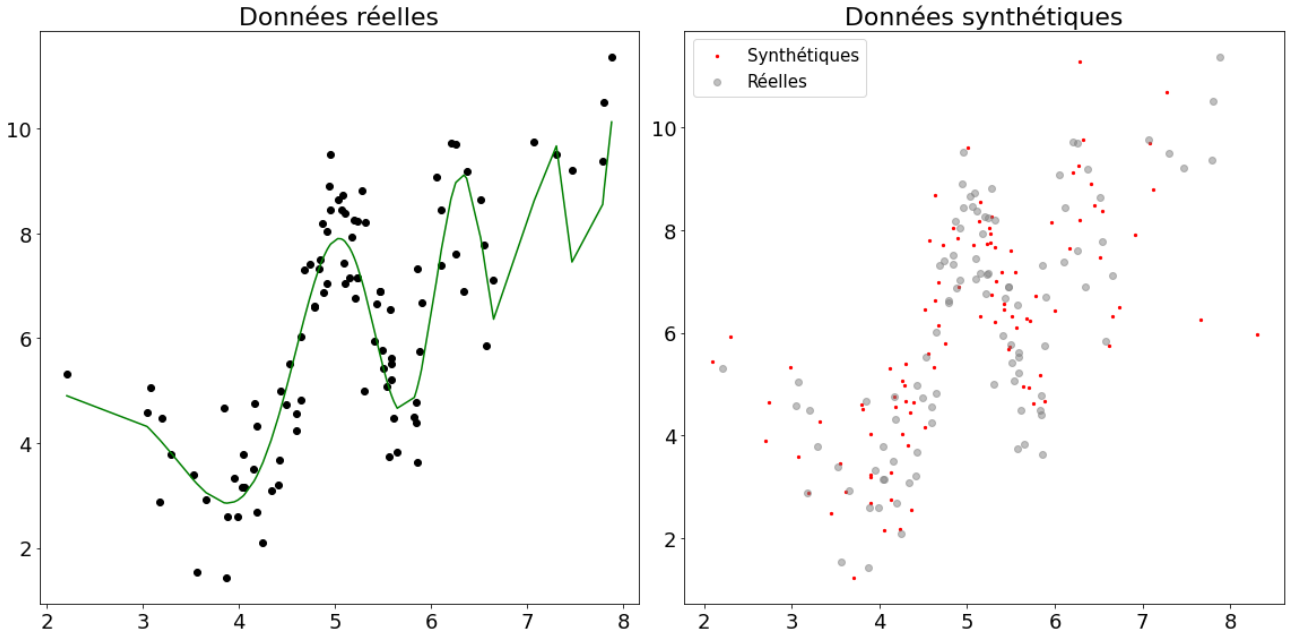
Données synthétiques

Génération d'un nouveau jeu de données artificielles avec les **mêmes propriétés statistiques** que les données originales sans possibilité, en théorie, de reconstruire des données sensibles

Qu’est-ce qu’une donnée synthétique ?

- Une donnée synthétique est une donnée **artificiellement générée** afin de simuler les propriétés d’une donnée réelle pour un besoin particulier

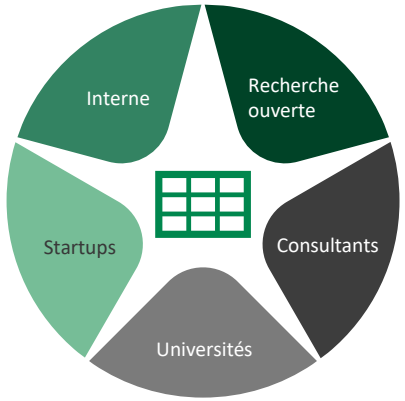
- Pour les besoins d’analytique, informative ou avancée, et la valorisation des données, il est impératif que les données synthétiques:
  - Aient les **mêmes propriétés statistiques**, univariés et multivariées, que les données réelles
  - Gardent les **dépendances**, statistiques et causales, entre des prédicteurs et une variable cible à prédire
  - Reproduire les **contraintes intrinsèques** dans les données réelles
  - Soient **similaires mais non identiques** pour préserver la confidentialité



Les données synthétiques doivent donc permettre de concevoir des analyses ou des modèles avec des résultats similaires à ceux obtenus à partir de données réelles

Avantages des données synthétiques

Faciliter et accélérer l'exploitation des données et **la collaboration** tout en préservant les informations individuelles sensibles **dans le respect des requis de gouvernance et de conformité**



Améliorer la performance des modèles en augmentant les données
(Ex: modélisation de la fraude ou la détection d'anomalies)

Tester des prototypes ou pilotes de produits sans recourir aux données réelles

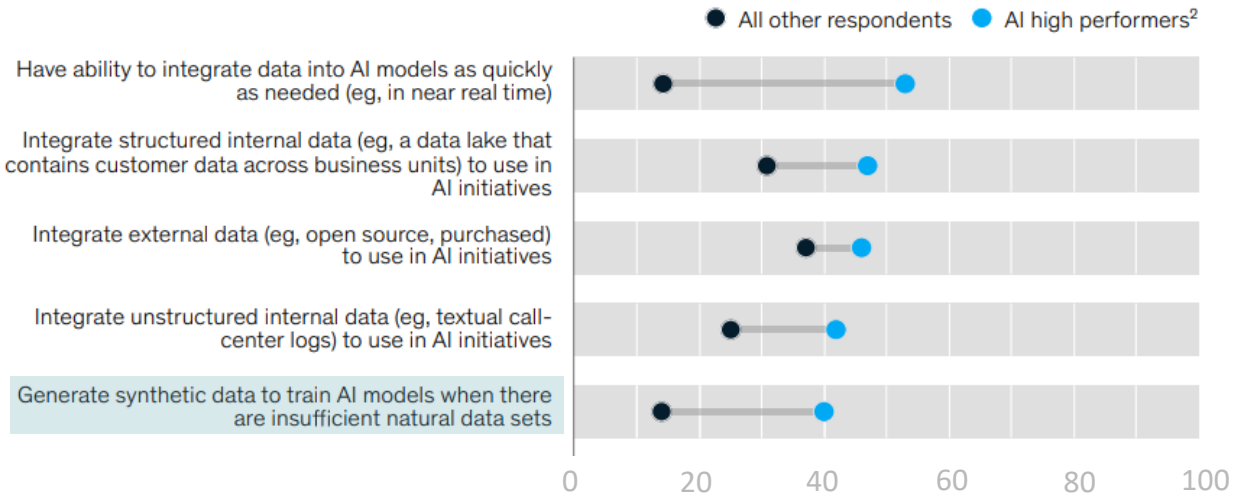
Faire des **tests de déploiements plus robustes** de systèmes ML

De nombreux avantages sont associés à l'utilisation des données synthétiques notamment pour la modélisation

L'essor du marché de la création de données synthétiques

- Les leaders en intelligence artificielle (IA) sont au moins **2 fois plus** susceptibles d'utiliser des données synthétiques dans leurs approches analytiques que le reste des entreprises

- Gartner prédit néanmoins que d'ici 2024, 60% des données utilisées pour l'analytique seront des données synthétiques générées par IA



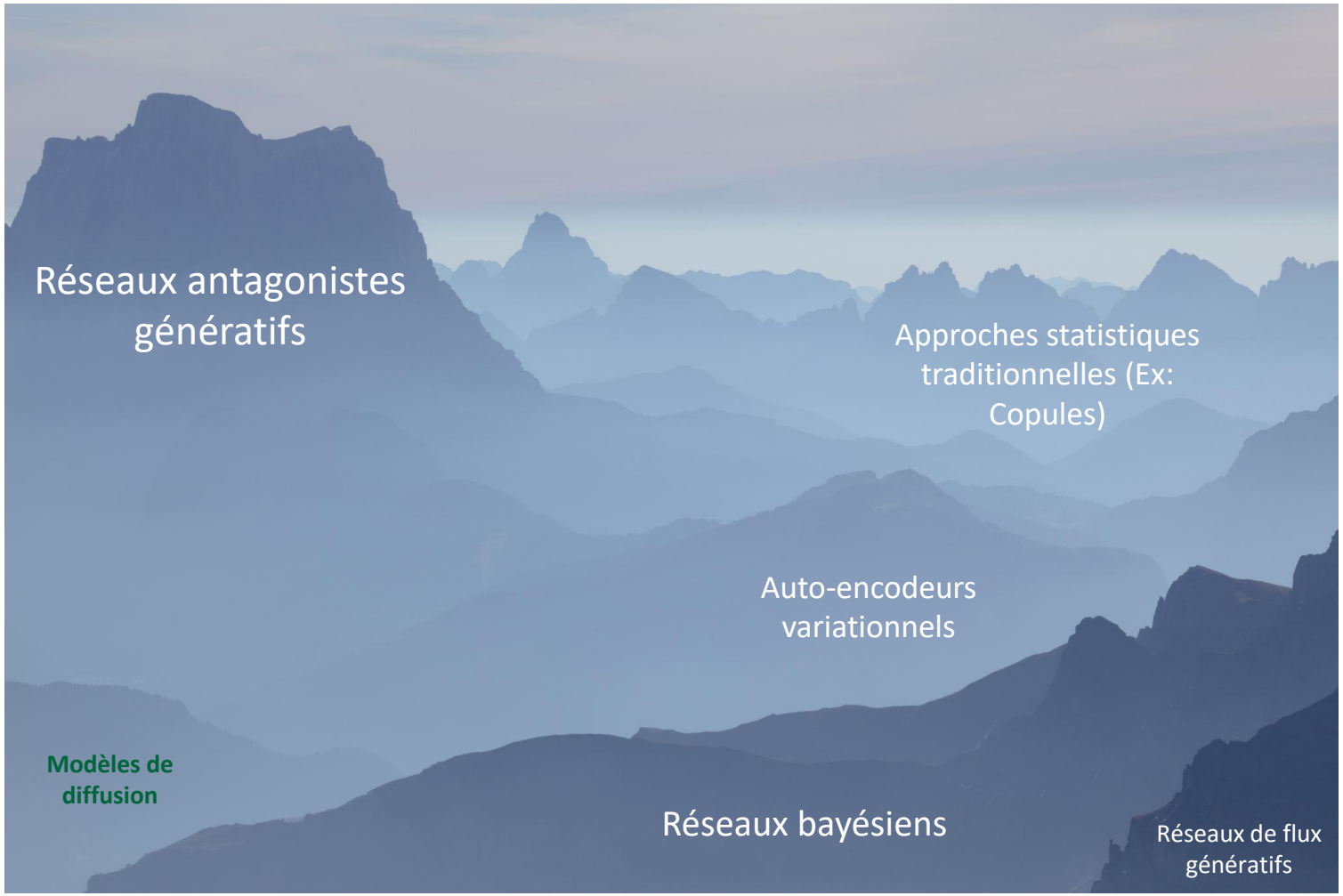
The state of AI in 2022—and a half decade in review | McKinsey

- Plusieurs produits commerciaux† existent sur le marché pour la génération de données synthétiques

MOSTLY·AI · Diveplane · TONIC THE FAKE DATA COMPANY · sarus · gretel™

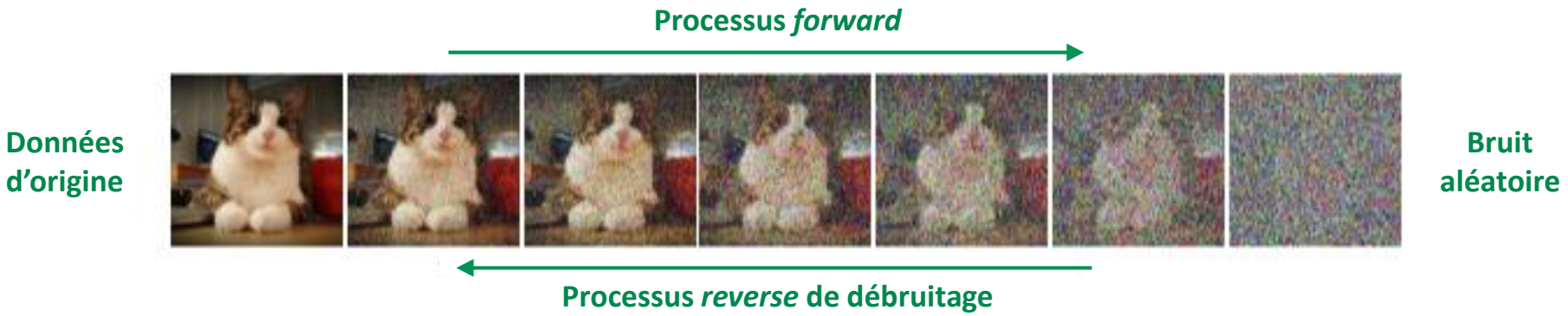
Le développement d'une expertise en données synthétiques est un avantage compétitif pour la valorisation des données

Panorama des approches de génération de données tabulaires synthétiques



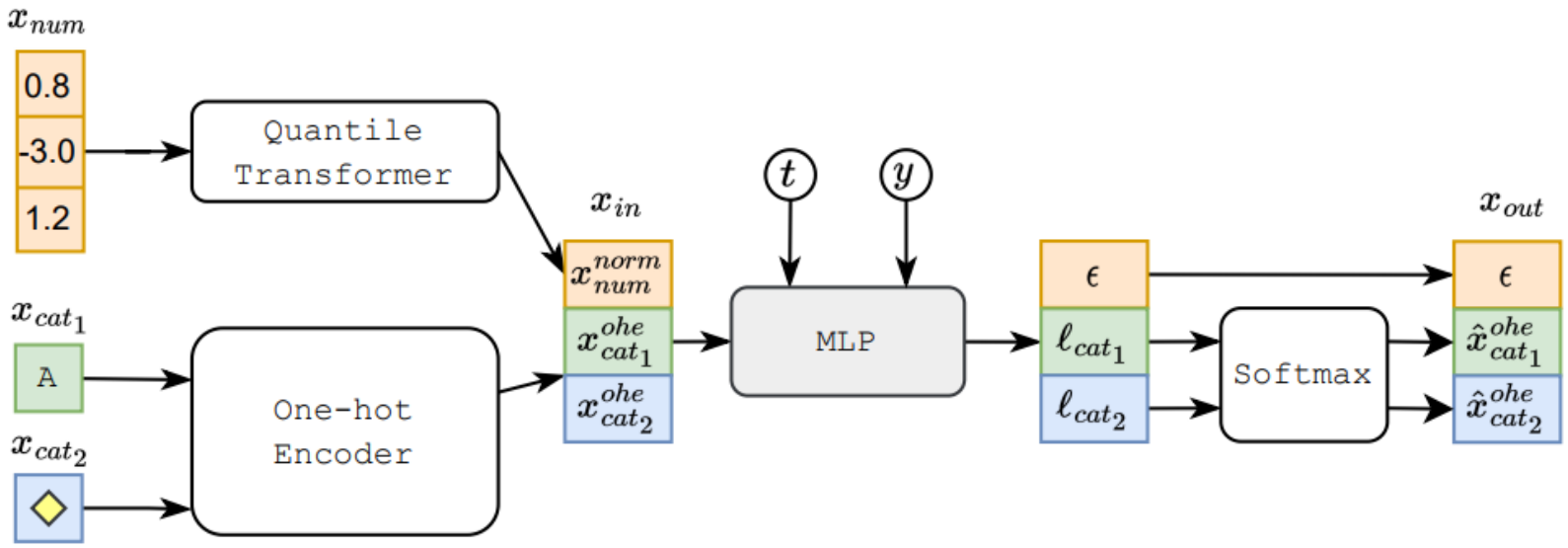
Les modèles de diffusion

- Un modèle de diffusion est constitué de deux processus de diffusion markoviens:
  - Un processus dit « *forward* » qui consiste à ajouter, de manière graduelle, du bruit à la donnée d'origine jusqu'à l'obtention d'un signal complètement aléatoire
  - Un processus dit « *reverse* » qui regénère la donnée d'origine en débruitant le signal aléatoire



Génération de données tabulaires avec un modèle de diffusion: TabDDPM

- Le modèle tabDPPM est un modèle de diffusion « open source » pour générer des données tabulaires en utilisant:
  - Un réseau de neurones type perceptron multicouche (MLP)
  - Une diffusion gaussienne pour les variables numériques prétraitées avec une transformation quantile
  - Une diffusion multinomiale pour les variables catégoriques encodées avec *one hot encoding*



Kotelnikov et al., TabDDPM: Modelling tabular data with diffusion models, PMLR 202:17564-17579, 2023

La problématique (1/2)

- La problématique proposée pour l'atelier vise à mettre à l'épreuve l'approche de création de données tabulaires et évaluer sa performance

- **Question 1: Est-il possible de discriminer une observation réelle d'une observation synthétique ?**
  - Un modèle de classification devra être entrainé sur un jeu de données avec des observations réelles et synthétiques étiquetées
  - Une analyse des éléments d'explicabilité permettant de comprendre ce qui discrimine les observations synthétiques des observations réelles devra également être menée
  - (Bonus) Toute mesure (Ex: modification du modèle de génération de données synthétiques) qui permette de mitiger le potentiel de discrimination entre les observations réelles et synthétiques.
  - La métrique d'évaluation est le score AUC (*Area under the ROC curve*)

- Nous avons mis en place une page sur la plateforme Kaggle qui vous permettra de:
  - En savoir plus sur le jeu de données utilisé
  - Soumettre vos prédictions sur un ensemble de test 1 fois par jour afin de vous permettre de vos ajuster avant la proposition finale à la fin de l'atelier.

La problématique (2/2)

- La problématique proposée pour l'atelier vise à mettre à l'épreuve l'approche de création de données tabulaires et évaluer sa performance

- **Question 2: Est-il possible de reconstruire les attributs sensibles d'une observation réelle à partir d'observations synthétiques ?**

niveau emploi	montant	qualification	…	sexe	etranger
junior	1963		…	?	?
		qualifié	…	?	?

+ **Jeu de données synthétiques** ?? →

sexe	etranger
Homme	Oui
Femme	Non

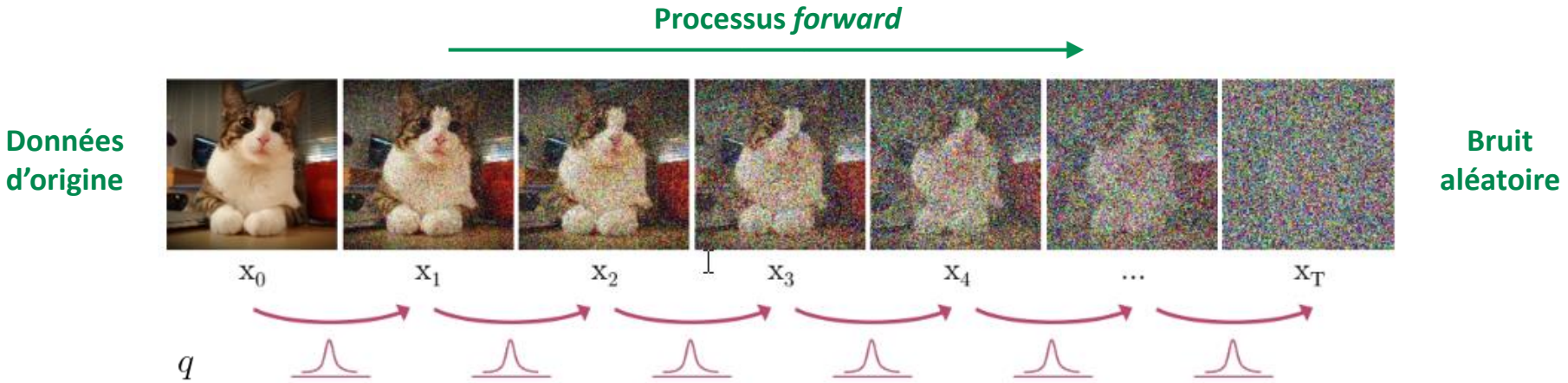
  - Un modèle devra reconstruire plusieurs informations sensibles à partir d'informations parcellaires d'une observation réelle et un ensemble de données synthétiques
  - Une analyse sur le nombre minimal de variables ou d'observations synthétiques utilisées serait intéressante
  - La métrique d'évaluation est l'exactitude (*Accuracy*) moyenne sur les attributs sensibles à prédire

- Nous avons mis en place une page sur la plateforme Kaggle qui vous permettra de:
  - En savoir plus sur le jeu de données et la métrique utilisés
  - Soumettre vos prédictions sur un ensemble de test 1 fois par jour afin de vous permettre de vos ajuster avant la proposition finale à la fin de l'atelier.

Annexe

Les modèles de diffusion: processus forward

- Le processus « *forward* » consiste à ajouter, de manière graduelle, du bruit à la donnée d'origine jusqu'à l'obtention d'un signal complètement aléatoire



- À chaque itération, le signal x_t est échantillonné conditionnellement à x_{t-1}. Exemple pour un bruit gaussien

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x_t}; \sqrt{1-\beta_t}\mathbf{x_{t-1}}, \beta_t\mathbf{I})$$

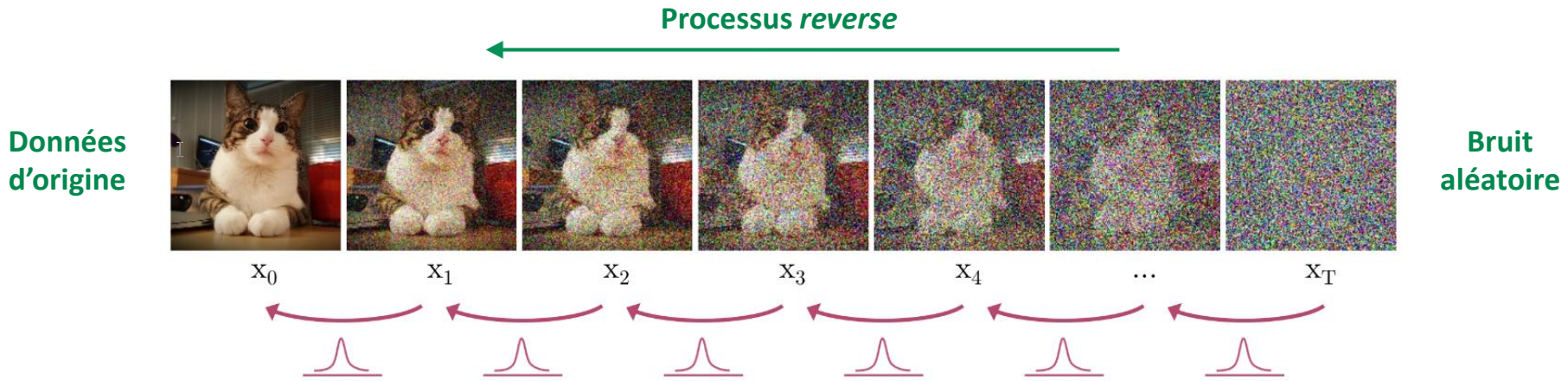
- Les paramètres β_t sont choisi de sorte à ce que le signal final soit distribué selon

$$q(\mathbf{x}_T|\mathbf{x}_0) \approx \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I}))$$

Ho et al., Denoising Diffusion Probabilistic Models, NeurIPS 2020

Les modèles de diffusion: processus reverse

- Le processus « *reverse* » regénère la donnée d'origine en débruitant le signal aléatoire



- Le débruitage commence à partir du signal aléatoire distribué selon (nous continuons avec l'exemple gaussien)

$$p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$$

- Le signal débruité itérativement au temps t-1 se génère par échantillonnage, conditionnel au temps t, selon la distribution

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t^2\mathbf{I})$$

où la fonction $\mu_\theta(\mathbf{x}_t, t)$ est apprise par un réseau de neurones.

Ho et al., Denoising Diffusion Probabilistic Models, NeurIPS 2020