# Parcours et expérience d'une méthodologiste sénior de Statistique Canada : de l'Université de Montréal à l'application de méthodes statistiques pour œuvrer dans l'intérêt de la société canadienne

**Kenza Sallier, méthodologiste principale**

**Conférences en statistique pour étudiants de premier cycle**

**Département de Mathématiques et Statistique**

**Université de Montréal**

**15 septembre 2023**

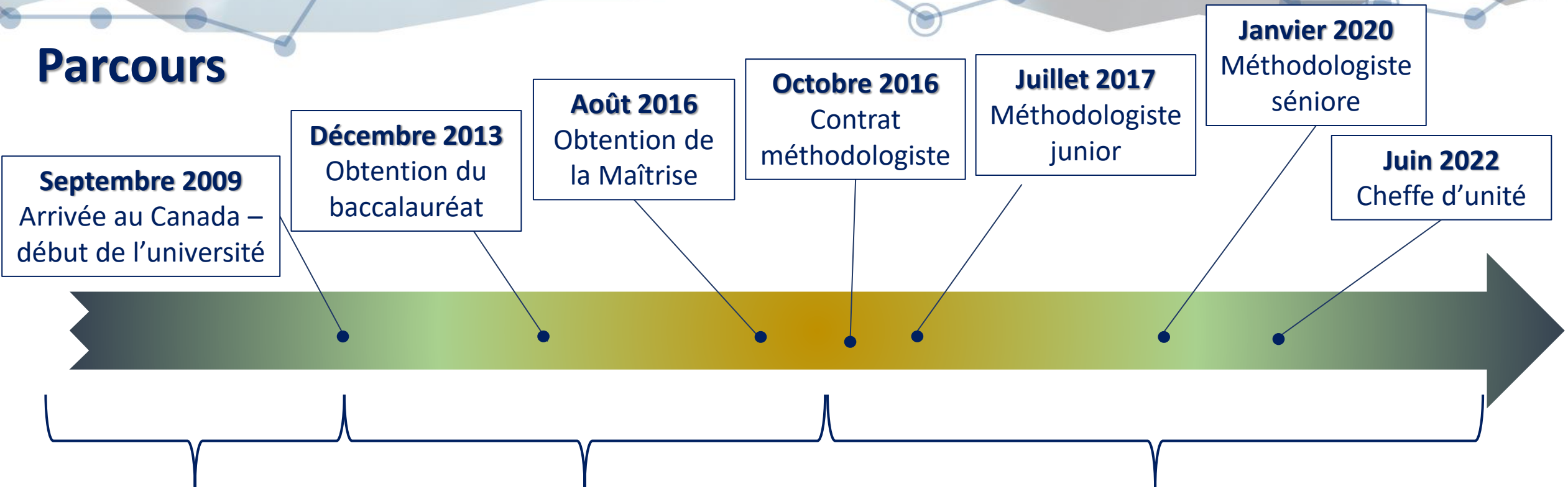
Éclairer grâce aux données, pour bâtir un Canada meilleur

# Sommaire

- Bref aperçu de mon parcours et de Statistique Canada
- **Exemple 1 :** Applications de méthodes statistiques pour le Recensement de la population canadienne
- **Exemple 2**: Recherche appliquée dans le cadre de l'innovation en confidentialité et accès aux données
- **Exemple 3:** Revue de littérature pour l'équité en science afin soutenir les prises de décisions informées
- Aller au-delà des compétences techniques
- Questions

# Parcours

# Parcours




**Système français :**
spécialisation en sciences et mathématiques

**Université de Montréal**

- Baccalauréat en mathématiques spécialité en statistique
- Maîtrise en statistique (séries chronologiques)

**Statistique Canada**

- Enquête sur l'Éducation
- Confidentialité et accès aux données
- Recensement de la population canadienne
- Plan d'action sur les données désagrégées

# Statistique Canada

# Statistique Canada

Logement

Recensement

- ✓ **Agence du Gouvernement du Canada: Statistique Canada est l'organisme national de statistique.**
- ✓ **Produit des données et de l'information de haute qualité, pertinente et actuelle décrivant la société canadienne**
  - ➢ Utilisées pour faire de la recherche
  - ➢ Utilisées pour développer des politiques à tous les niveaux du gouvernement
  - ➢ Aident les Canadiens à mieux comprendre leur pays
- ✓ **Bureau principal situé à Ottawa, mais depuis la pandémie, je travaille depuis Montréal et son bureau régional (centre-ville).**

Taux de criminalité

Inflation

COVID-19

Santé

Consommation de cannabis

PIB

Chômage

Origine ethnique

# Qu’est ce qu’un(e) méthodologiste?

**Méthodologiste = Mathématicien(ne) - Statisticien(ne)**

MÉTHODES

Les statisticien(ne)s-mathématicien(ne)s appliquent, adaptent et élaborent des méthodes mathématiques, statistiques ou d’apprentissage automatisé pour des problèmes pratiques

# Exemple 1 :

## Applications de méthodes statistiques pour le Recensement de la population canadienne

# Le Recensement de la population canadienne

- Enquête obligatoire menée tous les 5 ans
  - Dernier en 2021: mai à la mi-août
  - Prochain en 2026
- Objectifs: **dénombrer tous les individus et les logements au Canada et dresser le portrait socio-démographique de la population canadienne**
- Le Canada vient de dépasser le 40 million d'individus
- Nous visons un taux de réponse de **98%** nationalement mais également sur des petits niveaux géographiques

# Les Tests du Recensement de la population canadienne

- Enquête obligatoire menée tous les 5 ans aussi, mais à beaucoup plus petite échelle
- On tire un échantillon probabiliste
  - Derniers: mai-juin 2019
  - Prochains: mai-juin 2024
- Un des objectifs sur lesquels je travaille:
  - **Tester l'impact des changements du questionnaire sur les dénombrements et proportions obtenus pour la population générale et les sous-groupes d'intérêts visés par les questions**
- De façon simplifiée : On veut tester l'impact d'une nouvelle version du questionnaire en comparaison de la version de 2021

# Les Tests du Recensement de la population canadienne

## Grandes lignes du projet:

- ✓ **Déterminer les sous-populations d'intérêt (en plus de la population générale) impactées par la nouvelle version du questionnaire, en collaboration avec des experts des sujets abordés (démographie, santé, etc.)**
- ✓ **Établir la stratégie d'échantillonnage qui doit répondre aux contraintes suivantes (liste non-exhaustive) :**
  - Limite budgétaire qui se traduit en nombre total d'unité à échantillonner
  - Unités de l'ensemble du Canada et des deux langues officielles (stratification par Province/Territoires et langue)
  - Taux de réponse attendu en se basant sur les taux observés pour d'autres enquêtes (impact de la pandémie non-négligeable)
  - Détecter des différences d'auto-dénombrement entre les deux questionnaires de façon significative (tests d'hypothèse, puissance)
  - Assurer que l'allocation de la taille totale est faite de telle manière que les différentes tailles d'échantillon alloués permettent d'assurer la puissance voulue pour les tests d'hypothèses qui seront effectués
- ✓ **Collecter les données**
- ✓ **Traiter et pondérer les données en ajustant notamment pour la non-réponse**
- ✓ **Estimation de la variance**
- ✓ **Analyser les données et estimer l'impact du nouveau questionnaire**

# Les Tests du Recensement de la population canadienne

## Grandes lignes du projet:

- ✓ **Déterminer les sous-populations d'intérêt (en plus de la population générale) impactées par la nouvelle version du questionnaire, en collaboration avec des experts des sujets abordés (démographie, santé, etc.)**
- ✓ **Établir la stratégie d'échantillonnage qui doit répondre aux contraintes suivantes (liste non-exhaustive) :**
  - Limite budgétaire qui se traduit en nombre total d'unité à échantillonner
  - Unités de l'ensemble du Canada et des deux langues officielles (stratification par Province/Territoires et langue)
  - Taux de réponse attendu en se basant sur les taux observés pour d'autres enquêtes (impact de la pandémie non-négligeable)
  - Détecter des différences d'auto-dénombrement entre les deux questionnaires de façon significative (tests d'hypothèse, puissance)
  - Assurer que l'allocation de la taille totale est faite de telle manière que les différentes tailles d'échantillon alloués permettent d'assurer la puissance voulue pour les tests d'hypothèses qui seront effectués

**Exercice d'allocation efficace d'unités qui vise à minimiser :**

- La variance
- Les coûts
- Fardeau du répondant sur les sous-populations d'intérêt

→

**Optimisation sous contraintes appliquée au contexte d'échantillonnage**

# Les Tests du Recensement de la population canadienne

- **Le Recensement est une priorité gouvernementale.** Les résultats permettent de prendre des décisions informées : il est primordial que le questionnaire puisse adresser les questions de la société et permettre d'obtenir un portrait exact et précis de ses constituants
- Les tests du Recensement sont une étape clé de cette priorité centrée sur la société canadienne
- **Cours particulièrement utiles pour ce projet :**
  - ✓ Concepts et méthodes en statistique
  - ✓ Échantillonnage (pour les concepts mais pour les études par simulation aussi)
  - ✓ Plan d'expérience
  - ✓ Tous les cours comprenant de l'optimisation
  - ✓ Laboratoire de statistique/consultation
  - ✓ Progiciels statistiques
  - ✓ Tous les cours prérequis

# Exemple 2 :

## Recherche appliquée dans le cadre de l'innovation en confidentialité et accès aux données

# Confidentialité et accès aux données

- Selon l'Acte sur la statistique, Statistique Canada se doit de collecter compiler et diffuser des données pour informer la société canadienne sur différents aspects qui la caractérise notamment les aspects socio-démographiques et économiques.

- Toutefois, il est primordial et obligatoire pour l'Agence de protéger la confidentialité des répondants

# Confidentialité et accès aux données

- **Les organismes nationaux de statistique (ONS) s'efforcent d'assurer une plus grande transparence et une plus grande ouverture :** nouvelle culture d'être ouvert par défaut

- Au Canada, cette culture s'étend à l'ensemble du **Gouvernement du Canada** :
  - Gouvernement ouvert | Gouvernement ouvert, Gouvernement du Canada

- **Être plus centrés sur l'utilisateur et faciliter l'accès aux données pertinentes aux utilisateurs externes**
  - Nécessité de diffuser plus d'ensembles de données de qualité pour utilisation externe à l'Agence

- Trouver des moyens de diffuser des données plus **désagrégées** : Plan d'action sur les données désagrégées : Pourquoi est-ce important pour vous (statcan.gc.ca)

- **Révolution des données :** les progrès technologiques et les capacités informatiques ont permis de mettre en œuvre des solutions innovantes

# Confidentialité et accès aux données

- **Les organismes nationaux de statistique (ONS) s'efforcent d'assurer une plus grande transparence et une plus grande ouverture :** nouvelle culture d'être ouvert par défaut

- Au Canada, cette culture s'étend à l'ensemble du **Gouvernement du Canada** :
  - Gouvernement ouvert | Gouvernement ouvert, Gouvernement du Canada

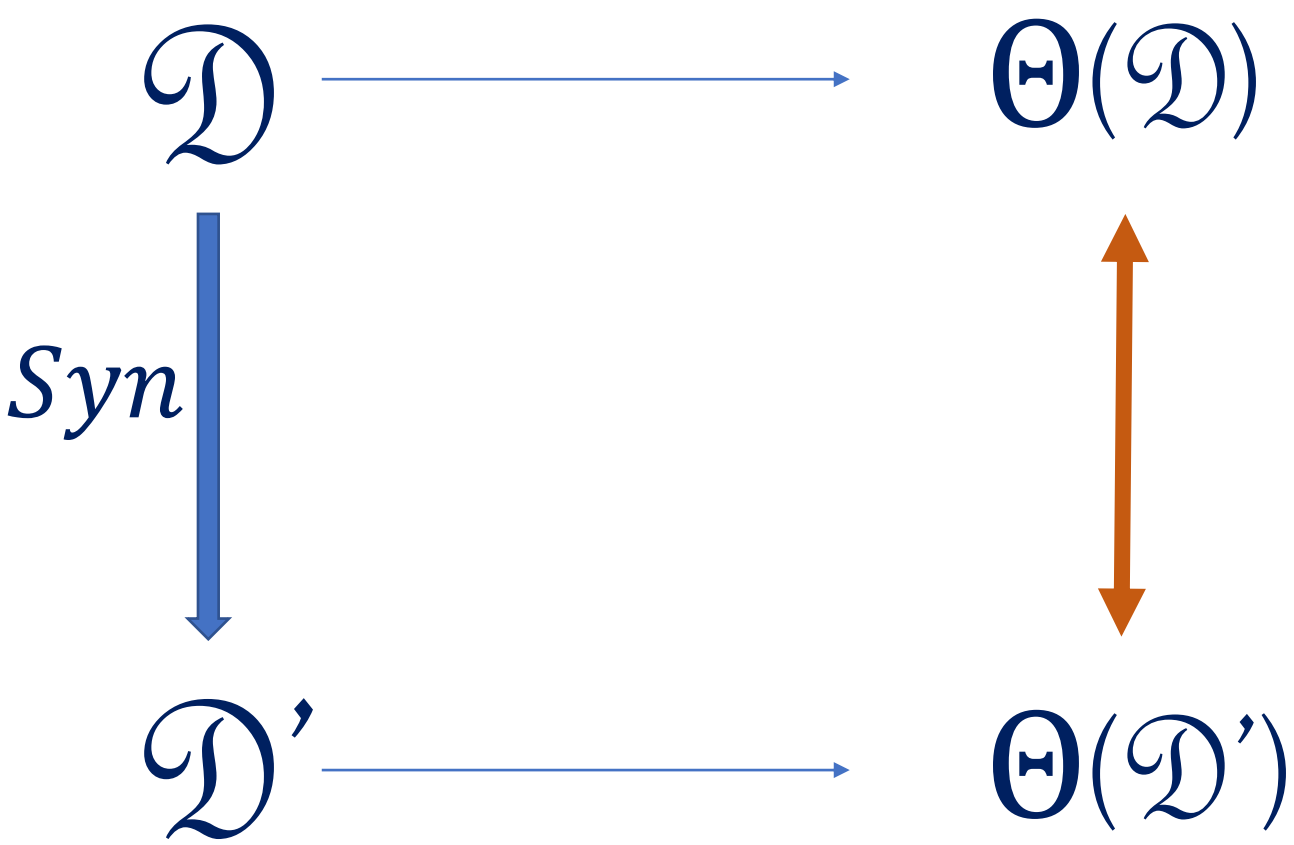
- **Être plus centrés sur l'utilisateur et faciliter l'accès aux données pertinentes aux utilisateurs externes**
  - Nécessité de diffuser plus d'ensembles de données de qualité pour utilisation externe à l'Agence

- Trouver des moyens de diffuser des données plus **désagrégées** : Plan d'action sur les données désagrégées : Pourquoi est-ce important pour vous (statcan.gc.ca)

- **Révolution des données :** les progrès technologiques et les capacités informatiques ont permis de mettre en œuvre des solutions innovantes

Les **données synthétiques** peuvent être une solution pour fournir des microdonnées analytiquement riches tout en respectant les impératifs d'intégrité et de confidentialité.

# Synthèse de données




$\mathcal{D}$ jeu de données original

$\mathcal{D}'$ jeu de données synthétiques

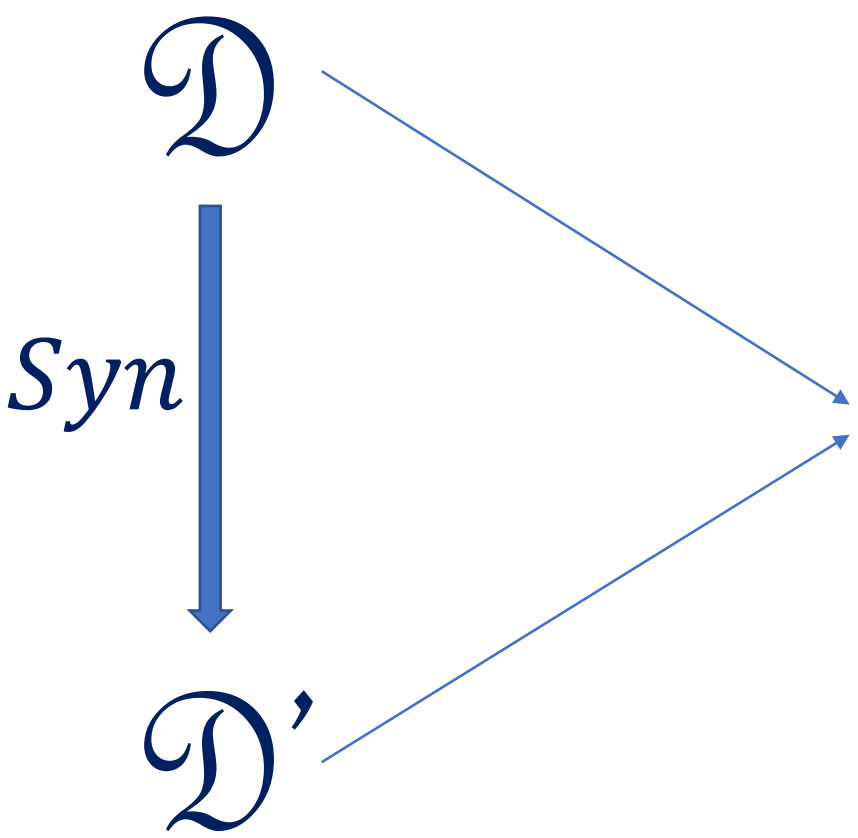
$Syn$ le synthétiseur

$\Theta$ résultats d'analyses effectuées par les utilisateurs

$U$ valeur analytique

$U(\mathcal{D}')$ fonction de la distance

# Synthèse de données




$\mathcal{D}$ jeu de données original

$\mathcal{D}'$ jeu de données synthétiques

$Syn$ le synthétiseur

$\Theta$ résultats d'analyses effectuées par les utilisateurs

$U$ valeur analytique

Distance minimale, valeur analytique maximale

# Que sous-tend Θ?

## I. Statistiques et tests univariés

- ✓Moyenne
- ✓Variance
- ✓Quantiles
- ✓Tests-t
- ✓Tests de Kolmogorov-Smirnov
- ✓Etc.

Le but est d'obtenir les mêmes **conclusions statistiques***

## II. Statistiques et tests multivariés

- ✓Covariance
- ✓Corrélation
- ✓Modèle linéaire général
  - ✓Régression linéaire
  - ✓Régression de poisson
  - ✓Régression logistique
  - ✓Probit/Tobit
  - ✓Etc.
- ✓Analyse de survie
- ✓Apprentissage automatisé
- ✓Etc.

*Sans savoir à l'avance quels tests et analyses vont être effectués.

# Que sous-tend Θ?

## I. Statistiques et tests univariés

- ✓Moyenne
- ✓Variance
- ✓Quantiles
- ✓Tests-t
- ✓Tests de Kolmogorov-Smirnov
- ✓Etc.

Qu'ont-ils tous en commun ?

## II. Statistiques et tests multivariés

- ✓Covariance
- ✓Corrélation
- ✓Modèle linéaire général
  - ✓Régression linéaire
  - ✓Régression de poisson
  - ✓Régression logistique
  - ✓Probit/Tobit
  - ✓Etc.
- ✓Analyse de survie
- ✓Apprentissage automatisé
- ✓Etc.

# Que sous-tend $\Theta$?

## I. Statistiques et tests univariés

- ✓Moyenne
- ✓Variance
- ✓Quantiles
- ✓Tests-t
- ✓Tests de Kolmogorov-Smirnov
- ✓Etc.

La distribution conjointe

## II. Statistiques et tests multivariés

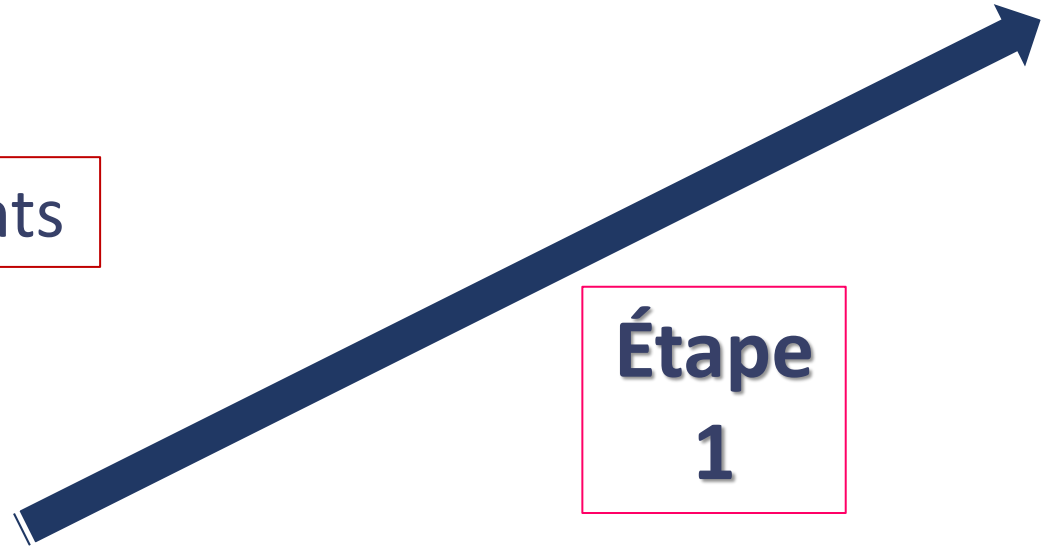
- ✓Covariance
- ✓Corrélation
- ✓Modèle linéaire général
  - ✓Régression linéaire
  - ✓Régression de poisson
  - ✓Régression logistique
  - ✓Probit/Tobit
  - ✓Etc.
- ✓Analyse de survie
- ✓Apprentissage automatisé
- ✓Etc.

# La synthèse de données

**Distribution conjointe**

Répondants

$\mathcal{D}$



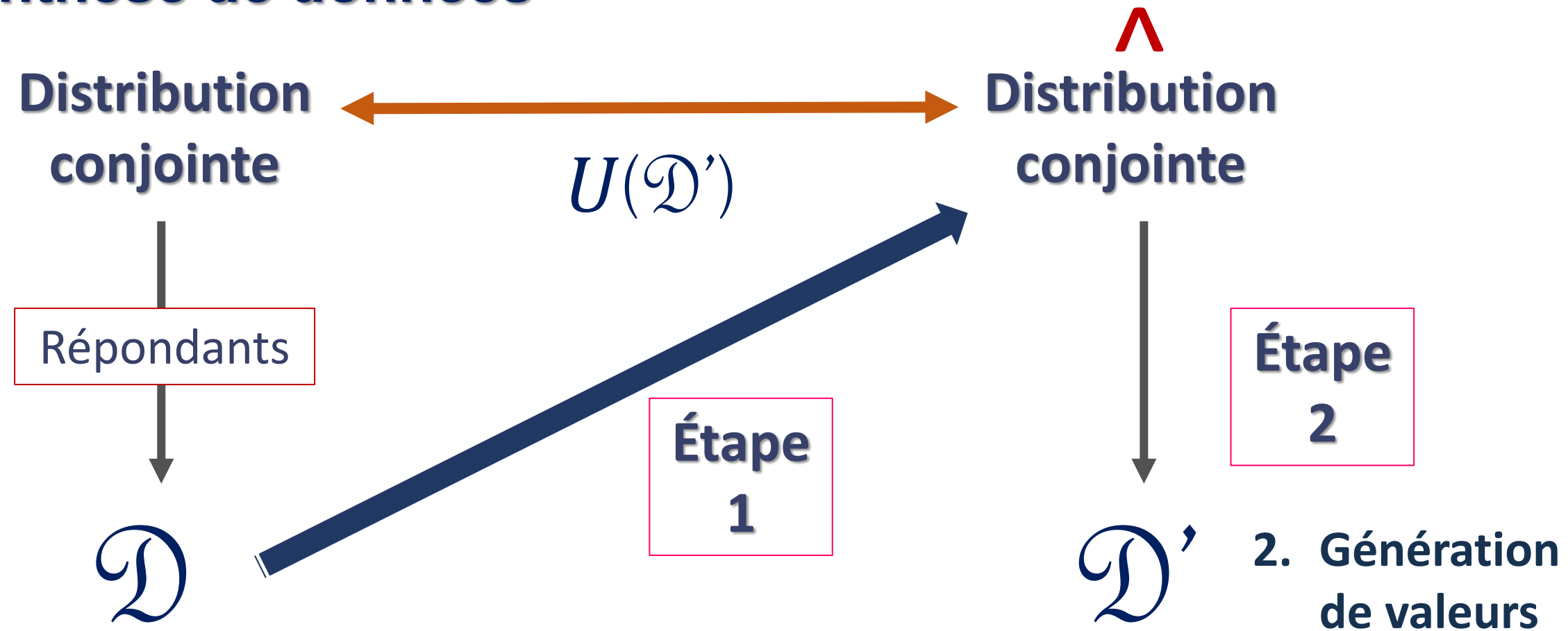
**Étape 1**

$\Lambda$

**Distribution conjointe**

1. **Modélisation** pour capturer la **distribution conjointe** des données originales

# La synthèse de données



# Synthèse de données et imputation

- La synthèse de données est une forme d'imputation massive : on supprime intentionnellement toutes les données pour les remplacer

- C'est en fait Rubin qui a proposé d'utiliser cette approche dans le cadre de protection de la confidentialité

- Plusieurs techniques existent pour approcher la distribution conjointe et générer des données synthétiques à partir de proxy: Fully Conditional Specification, Neural Networks, Pseudo Likelihood, etc.

# Synthèse de données à Statistique Canada

- Statistique Canada a rendu public des versions synthétiques de deux ensembles de données différents liés aux informations du recensement, de la mortalité et du cancer (2018 et 2019). Produits pour des hackathons
- Mêmes conclusions statistiques que celles trouvées dans les données originales
- Dans les deux cas, les données synthétiques ont été générées à l'aide du package R synthpop
- Un ensemble de données actuellement utilisé par l'École de la fonction publique du Canada

# Synthèse de données à Statistique Canada

- Même si l'idée, en théorie a été proposée il y a longtemps en **pratique** c'est une innovation: j'ai été impliquée dans beaucoup d'initiatives internationales (conférences, articles, groupes de travail)
- Statistique Canada a mené la scène internationale sur le sujet, notamment en termes de publication : https://unece.org/statistics/publications/synthetic-data-official-statistics-starter-guide (collaboration avec l'ONU)
- **Cours particulièrement utiles pour ce projet :**
  - ✓ Concepts et méthodes en statistique
  - ✓ Analyse multivariée appliquée
  - ✓ Fondements théoriques en science des données
  - ✓ Méthodes de prévision
  - ✓ Échantillonnage
  - ✓ Laboratoire de statistique/consultation
  - ✓ Progiciels statistiques
  - ✓ Tous les cours prérequis

# Exemple 3 :

Revue de littérature pour l’équité en science afin de soutenir les prises de décisions informées

# Plan d’action sur les données désagrégées (PADD)

- La diffusion de données pertinentes et actuelles, sous la forme la plus désagrégée possible, a toujours été un principe fondamental de Statistique Canada.
- Les mouvements sociaux pour les droits autochtones, la justice raciale et l'équité économique sont désormais à l'avant-garde, exigeant des données pour guider les processus décisionnels.
- En réponse, le gouvernement du Canada a financé la phase de développement de cinq ans de l’Agence pour le PADD commençant en 2021
- Et à mon tour, j'ai participé l'année dernière à un projet de recherche financé par le PADD.

# Plan d’action sur les données désagrégées (PADD)

- Le PADD mènera à la production de renseignements statistiques détaillés qui mettront en lumière les expériences par des groupes de population particuliers, dont les femmes, les peuples autochtones, les populations racisées et les personnes ayant une incapacité

Nous parlons de groupes de la population ayant des caractéristiques rendant inefficace les méthodes d’échantillonnage traditionnelles parce que :

- Populations rares (très faible prévalence dans la population)
- Populations cachées (manque d’identification au préalable)

# Plan d’action sur les données désagrégées (PADD)

- De nombreuses méthodes existent : laquelle utiliser ?
- Comment évaluer la faisabilité de l'échantillonnage ?
- Comment réduire le suréchantillonnage ?
- Comment être plus conscient du fardeau des répondants ?

**Nécessité de centraliser les informations, bonnes pratiques, méthodes, références pour les gestionnaires d’enquête impliqués dans les initiatives du PADD**

# Plan d’action sur les données désagrégées (PADD)

## Projet de recherche ayant deux résultats principaux:

1. **Document de principes directeurs : revue de la littérature et considérations pratiques**
   - ✓ Pas un manuel exhaustif: un point de départ pour les processus d’échantillonnage et de collecte de données DDAP
   - ✓ Objectif global : centraliser les informations disponibles sur les méthodes et les meilleures pratiques
   - ✓ Distribué pour l’ensemble des employés de Statistique Canada et servant de base pour des recommandations faites à d’autres partenaires tels que les Instituts de recherche en santé du Canada

2. **Tableaux de dénombrements : Évaluation de la prévalence des groupes d'intérêt DDAP**
   - ✓ Utilisation du Recensement de la population de 2021
   - ✓ Permet d'évaluer la faisabilité de l'échantillonnage
   - ✓ Disponible sur demande justifiée aux gestionnaires d’enquête

# Plan d’action sur les données désagrégées (PADD)

**Cours particulièrement utiles pour ce projet :**

- ✓ Concepts et méthodes en statistique
- ✓ Échantillonnage
- ✓ Laboratoire de statistique/consultation
- ✓ Tous les cours prérequis

# Aller au-delà des compétences techniques

# Aller au-delà des compétences techniques

- Dans le monde du travail il est important de réaliser que d'autres compétences sont primordiales car il est rare de travailler seul
- Notamment :
  - Comment savoir communiquer et vulgariser des concepts techniques
  - Collaborer avec autrui et mobiliser les forces de chacun
- Pour ceux intéressés : compétences clés en leadership

# Pourquoi ce choix de carrière?

- J'ai toujours aimé les mathématiques
- Les statistiques semblaient être un moyen de **pouvoir appliquer mes capacités en mathématiques de façon plus concrète**
- **Statistique Canada me permet d'œuvrer pour la société sur des aspects qui sont importants à mes yeux** :
  - L'équité et la diversité
  - La facilitation de la recherche via la démocratisation de l'accès aux données
- Il existe d'autres domaines à l'Agence: l'économie, l'environnement, la santé, l'accès à l'habitation, etc.
- Grand nombre de personnes très compétentes (beaucoup d'experts internationaux)
- **Les valeurs humaines sont une priorité ainsi que la santé mentale et le développement continu (formation, opportunités d'apprentissage, conférences, etc.)**

# Et les études ?

- Ce n’était pas facile (☺) mais excellente formation!
- Apprendre pour comprendre et pas juste pour passer permet une longévité des connaissances: la rigueur est primordiale
- Aidez-vous les uns les autres, et poser des questions !
- Je connais aussi plusieurs autres anciens étudiants qui ont eu des parcours de carrière intéressants et inspirants!

# Projets de grande envergure : Apprentissage automatisé en action

## Couplage d'enregistrements

- Utiliser de l'apprentissage non-supervisé pour trouver un moyen de surmonter les erreurs ou variantes dans les noms utilisés lors du couple d'appariements

## Codage

- Classifier les dépenses provenant du journal de l'Enquête sur les dépenses des ménages

## Modélisation

- Modéliser le statut d'occupation des logements privés pour le Recensement de 2021

## Imputation/Estimation

- Évaluer l'utilisation d'apprentissage de réseaux de neurones pour remplacer l'imputation par donneur pour les variables d'impôts sur le revenu

## Contrôle de divulgation

- Générer des fichiers synthétiques avec une grande valeur analytique

**Apprentissage automatisé utilisé dans une variété de projets**

# MERCI !
# THANK YOU!

kenza.sallier@statcan.gc.ca