# Nouvelles techniques pour coder dans une nomenclature : l'expérience de Statistique Canada

Présentation pour l'INSEE, janvier 2020

Yanick Beaucage, Harry François, Patrick Gallifa, Christie Sambell,

Éclairer grâce aux données, pour bâtir un Canada meilleur

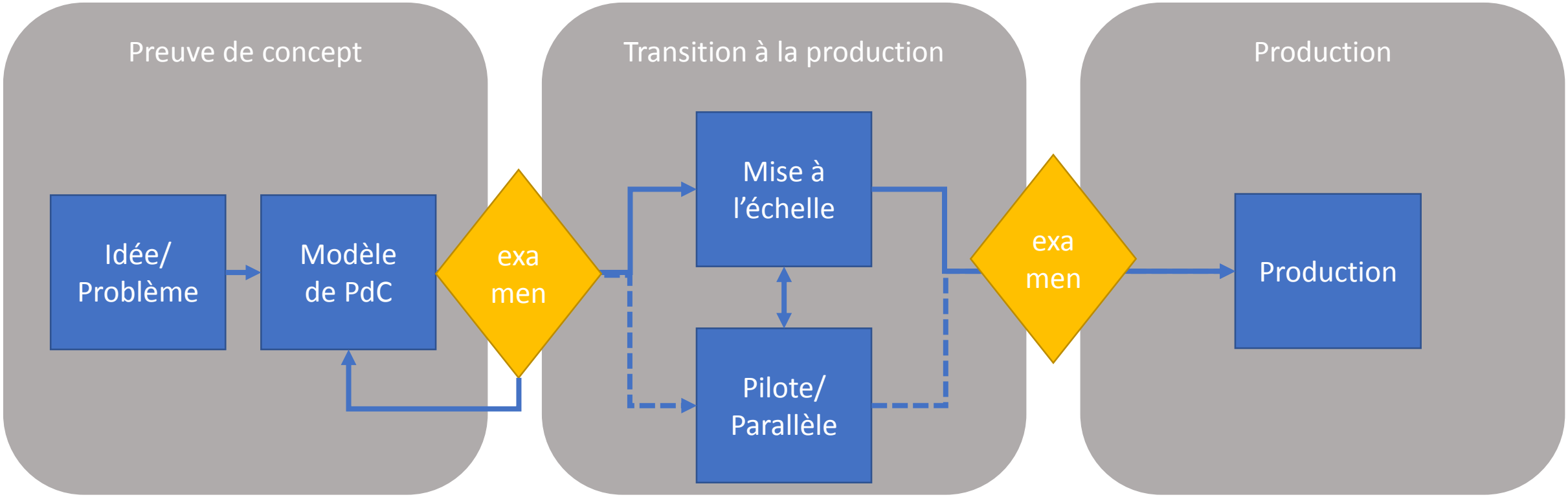
# Sommaire

- La Division de la science des données
- Fonctionnement
- Projets :
  - Survol de projets
  - Classification des données de caisse
    (Pause) Outil de codage généralisé
  - Classification pour l'Enquête sur les dépenses des ménages
- Conclusion

# La Division de la science des données

- Création récente
  - Direction des méthodes statistiques modernes et de la science des données
  - Directrice Sevgui Erman
- Mandat
  - Offrir une expertise en science de données appliquées
    - Conceptualiser, élaborer, déployer
  - Renforcer la capacité dans l'utilisation des mégadonnées et des données non-structurées
  - Offrir des services de consultation pour orienter l'ensemble de l'organisme
  - Établir un cadre de pour développer des applications en AA responsables
  - Travailler avec les communautés de praticiens en AA
- Modèle en étoile
  - La division est un carrefour rassembleur pour les activités de recherches et de développements
  - Des scientifiques de données (*data scientists*) locaux assurent la continuité
  - Favorise un changement de culture

# Fonctionnement



# Fonctionnement

- Discussions initiales
  - Obtention d'informations clés permettant la priorisation
    - Améliorations souhaitées
    - Possibilité de reproduction ailleurs dans l'organisme
    - Nécessite le traitement de données volumineuses ou non structurées
    - Considération éthique
  - Comité interne pour la priorisation de projets
- Processus de revue indépendant
  - Avant la mise en production
  - Validation des méthodes/algorithmes et de leur application
  - Construction d'un cadre pour l'apprentissage automatique responsable en cours
- Fermeture du projet
  - Mesures qui démontrent les gains obtenus
  - Mise en place d'un processus d'assurance de qualité

# Fonctionnement

- Infrastructure
  - L'infonuagique
  - Accès à des ordinateurs de type GPU pour le traitement d'image
  - Accès aux logiciels de type open source
- Recrutement
  - Définition de l'ensemble des compétences requises
  - Participation au recrutement d'employés et d'étudiants spécialisés
- Formation
  - Participation active au sein de différentes communautés de pratique
  - Séminaire et présentation
- Collaboration

## Fonctionnement

# COLLABORATION EXTERNE

# Survol de projets

- Classification d'images satellites
  - Identification des grandes cultures, serres, panneaux solaires, mises en chantier
- Classification de texte
  - Marchandises achetées, vendues, transportées, importées, exportées
  - Regroupements de commentaires
- Classification d'information provenant de journaux
  - Détection d'événements économiques
  - Indicateur de sentiment envers l'économie en « temps réel »

# Survol de projets

- Extraction d'information en utilisant la théorie des graphes
  - Information économique à partir de tableaux provenant de fichiers PDF
  - Information sur le mouvement de bétails
- Estimation du rendement des cultures à partir de données satellitaires
- Application de méthodes d'apprentissage automatique dans la recherche
- Recherches
  - Méthodes d'encryptage homomorphe
  - Estimation à partir d'un échantillon non-probabiliste

# Motivation

- Objectif : obtenir les totaux pour les produits vendus par région et par code du Système de classification des produits de l'Amérique du Nord (SCPAN) pour les détaillants.
- Nous menons actuellement une enquête pour obtenir ces totaux.
- Les données des lecteurs optiques sur le commerce de détail contiennent des renseignements au sujet de chaque produit vendu.
- L'utilisation des données de lecteurs optiques
  - Permet d'obtenir des niveaux plus détaillés
  - Réduit le fardeau de réponse.

# Données de caisse sur le commerce de détail

- Disponibles pour un premier grand détaillant
  - Toutes les semaines
  - Inclut chaque vente de produit ainsi que plusieurs variables
- Très volumineux : des millions d'enregistrements et plusieurs giga-octets de données par semaine.
- Exemples :

| Nom du produit | Rayon du magasin | Code interne du détaillant | Marque du magasin | Code postal | … |
|---|---|---|---|---|---|
| Pomme verte fraîche | Données générales sur les aliments | 25 | Marque1 | ABCDEF | |
| Poulet congelé ARTICLE 1234 | Produits réfrigérés et congelés | 72 | Marque2 | FEDCBA | |
| Ordinateur portable doté d'un processeur 4 cœurs et de 16 Go de RAM | Électronique | 86 | Marque1 | ABCABC | |

# Traitement des données

- Un tableau de concordance entre les SCPAN recherchés et les codes internes du détaillant a été élaboré par nos analystes
  - On peut donc assigner un code SCPAN à chaque entrée
  - Puis faire le total par région ou encore par regroupement de SCPAN
- Le traitement de ce grand volume de données constituait un défi
  - La mémoire vive étant inférieure à la quantité de données hebdomadaires.
- Une approche *MapReduce* a été adoptée
  - Les données sont d'abord divisées en blocs.
  - Puis, chaque bloc est traité indépendamment (on calcule les totaux désirés)
  - On calcule les totaux globaux en sommant les agrégats de chacun des blocs.
- Permet de traiter en quelques heures un mois de données
  - La majorité du temps étant passée à simplement charger les fichiers

# Difficultés avec cette approche

- Ce ne sont pas tous les détaillants qui auront un code interne
- Dans certains cas, il y a plusieurs codes du SCPAN pour le même code interne du détaillant.
- On doit continuellement mettre à jour le fichier de concordance
- On pourrait prédire le code du SCPAN à partir des champs de texte associés à chaque produit.
  - En utilisant l'apprentissage automatique
  - En se servant des données codées comme données d'apprentissage
  - L'algorithme apprend à associer à chaque entrée de texte le bon code
  - On peut alors prédire le code SCPAN de nouvelles entrées de texte
    - Sans se servir du code interne

# Modèle du sac de mots (*bag of words*)

- Dans le modèle du sac de mots, nous recherchons la présence d'un mot n'importe où dans le document.
- Par exemple, dans la chaîne de caractères « lait au chocolat délicieux », les mots « lait », « chocolat » et « délicieux » sont présents et ils seront utilisés comme éléments dans le modèle.
- Le contexte et l'ordre des mots est ignoré.
- Une matrice (appelée matrice des termes du document) est créée
  - les lignes représentent les produits et
  - les colonnes représentent les mots.
  - Chaque cellule contient un indicateur (0 ou 1) selon la présence du mot dans le texte correspondant au produit
    - On peut également utiliser la fréquence au lieu d'un indicateur binaire, mais les résultats sont semblables

# Modèle des n-caractères (n-grams)

- Au lieu de rechercher la présence de mots, nous recherchons la présence de sous-chaînes de caractères de longueur n.
- Par exemple, pour « **lait au chocolat délicieux** », les sous-chaînes de 3 à 6 caractères incluent entre autres :
  - « délici »
  - « dél »
  - « choco »
  - « at dél »
  - « au cho »
  - « lait »
- Offre un bon moyen de régler plusieurs problèmes, comme les doublons, le pluriel et les variantes, les mots tronqués, etc.
  - Permet de tenir compte de l'ordre des mots

# Algorithme d'apprentissage automatique

- Une fois que la matrice est créée (à partir des mots ou des n-caractères)
  - On utilise l'apprentissage automatique pour déterminer quels mots ou n-caractères sont associés à chaque code du SCPAN.
- Il existe de nombreux algorithmes d'apprentissage automatique différents
  - Dans le cas de ce projet, nous utilisons l'algorithme XGBoost jumelé à des apprenants linéaires en raison de la très grande dimension de la matrice de termes du document.
- La régularisation est utilisée pour favoriser les modèles simples
  - L'option de régularisation L2 dans XGBoost est utilisée ici.

# Modèle linéaire utilisé

- Pour chaque classe c possible du code du SCPAN, un vecteur de pondération $\mathbf{w}_c$ et un terme de biais $b_c$ sont créés.
- On obtient alors la cote prédite $p_c$ à partir de l'équation linéaire: $p_c = b_c + \boldsymbol{w}_c \cdot \boldsymbol{x} = b_c + \sum_{i=0}^{n} w_{ci} x_i$
- Pour chaque enregistrement, on choisit comme prédiction
  - Classe (code du SCPAN) ayant obtenu la cote prédite $p_c$ la plus élevée
- Ici, le vecteur d'entrée **x** est le vecteur indicateur pour chaque mot/n-caractère.
- Nous pouvons convertir les cotes en probabilités, ce qui est utile pour certaines tâches :

$$\boldsymbol{p}_i' = \frac{\exp(\boldsymbol{p}_i)}{\sum_{j=1}^{n} \exp(p_{i,j})}$$

- XGBoost offre un moyen très efficace et très flexible de former un tel modèle.
  - Détermine les $\mathbf{w}_c$ et les $b_c$ qui s'harmonisent le mieux aux données d'apprentissage.
  - La régularisation L2 favorise le maintien de facteurs de pondération faibles dans la mesure du possible.

# Résultats obtenus

- Un taux d'exactitude très élevé (plus de 99 %) a été observé pour les entrées de ce détaillant qui n'étaient pas présentes dans les données d'apprentissage.
- Pas surprenant
  - les données sont très riches en caractéristiques et étiquetées de façon systématique.
- Malheureusement, il pourrait subsister des erreurs par rapport à la réalité
  - Par exemple, si tous les sacs à ordures ont été mal étiquetés comme étant des sacs de vêtements, on obtiendra une exactitude élevée malgré tout
  - On doit donc valider, effectuer des corrections puis entraîner le modèle de nouveau.
- Un deuxième détaillant a été testé
  - On a étiqueté manuellement un échantillon de produits
  - En comparant aux prédictions, un taux d'exactitude de 86 % a été obtenu.
- Nous avons décidé de passer le modèle en production.

# Cadre de contrôle de la qualité

- Un cadre de contrôle de la qualité a été élaboré pour évaluer la qualité du modèle une fois entré en production.
- Lorsqu'on obtient de nouvelles données du détaillant
  - On prédit le code SCPAN à l'aide du modèle
  - On tire un échantillon stratifié par type de produit (existant ou nouveau) et par niveau de probabilité
    - Échantillon choisi de façon systématique et ordonné selon le code du SCPAN et les ventes.
  - Un analyste étiquette manuellement cet échantillon sans voir les prédictions du code du SCPAN, afin d'éviter tout biais.
- Les prédictions sont ensuite comparées avec le codage manuel et on estime le niveau d'exactitude obtenu.
  - Cette estimation peut être basée sur les comptes ou pondérée économiquement.

# Mise en œuvre

- Initialement, on tirait un échantillon de 1 000 enregistrements par semaine pour le CQ.
- Deux analystes effectuaient l'étiquetage de l'échantillon de façon indépendante
  - Tout désaccord était soumis au superviseur qui déciderait du code final.
- Le taux de précision obtenu était supérieur à 80 % au cours de plusieurs semaines d'essais.
  - L'exactitude est restée relativement stable après plusieurs essais de ce genre
  - On a décidé de réduire la fréquence d'échantillonnage à une fois par mois.
- La précision selon le poids économique était supérieure, variant de 92 % à 94 %.
- On a constaté une bonne corrélation (négative) entre la probabilité prédite de la classe du code et le taux d'erreur.
- Dans l'ensemble, les analystes sont satisfaits de ces résultats.
  - L'apprentissage automatique est beaucoup plus cohérent que l'étiquetage manuel initial.
  - L'approche est utilisée en production depuis maintenant un an

# Prochaines étapes

- Les analystes
  - corrigent les étiquettes dans les données d'apprentissage,
  - ajoutent de nouvelles descriptions de produits.
- On créera un nouveau modèle avec l'ensemble de données mis à jour.
  - La stratégie de contrôle de qualité sera ensuite appliquée et nous verrons quel en est l'effet sur la précision du modèle.
- On voudra augmenter le nombre de détaillants inclus dans le travail
- Nous étudierons les techniques d'apprentissage par transfert et d'apprentissage actif
  - Afin de réduire au minimum la quantité de données étiquetées requises pour les nouveaux détaillants.
  - D'autres algorithmes de classification de texte (y compris les réseaux neuronaux) seront également mis à l'essai.
- G-Code est maintenant disponible

# G-Code

## Système généralisé de codage développé par Statistique Canada

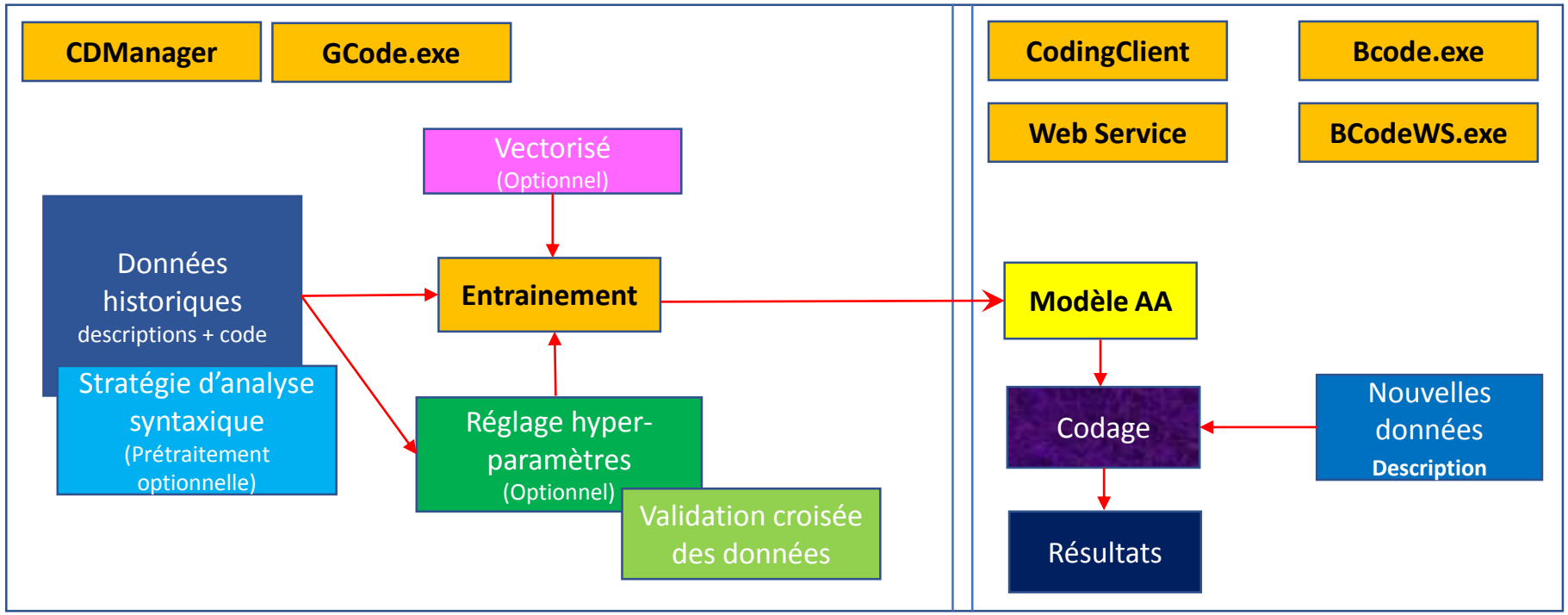
- Peut coder une gamme de variables: industrie, profession, éducation, origine, etc.;
- Langue de la description peut-être soit le français, l'anglais ou d'autres langues;
- La description peut comporter un ou plusieurs mots;
- Le codage peut-être réalisé par lots ou de manière interactive pour des millions de descriptions;
- Offre une flexibilité :
  - dans le choix des paramètres de codage; et
  - dans le choix des méthodes de codage.
- Fournit également une interface de programmation d'applications (API) à laquelle on accède à partir du service Web;
- La dernière version (V3.2) comporte un module de codage par apprentissage automatique (AA).

# Processus derrière l'utilisation de l'apprentissage automatique (AA)

| | Préparation des données | | Production |
|---|---|---|---|
| | Étape 1 (Optionnelle) | Étape 2 (Obligatoire) | Étape 3 |
| | **Réglage des hyper-paramètres / Validation croisée** | **Entrainement** | **Prédiction (Codage)** |
| **Objectif** | • Réglage des hyper-paramètres sert à optimiser le modèle<br>• Validation croisée sert à évaluer la précision du modèle | Créer un modèle en utilisant un algorithme d'apprentissage automatique et des données à partir desquelles des leçons sont tirées | Prédire y étant donnée x en utilisant un modèle AA<br>Y: chaîne de caractères<br>X: texte non-structuré |
| **Exigences / Bénéfices à tirer** | • Architecture de traitement puissante<br>• L'analyse des données<br>• Expertise en AA | • Architecture de traitement puissante<br>• L'analyse des données<br>• Expertise en AA<br>• Réglage des hyper-paramètres / Validation croisée | • Un modèle optimisé (impacts sur les résultats)<br>• Processus généralisé |

# G-Code

## Processus de codage utilisant l'apprentissage automatique



# L'option d'apprentissage automatique dans G-Code

I. **FastText: Réseaux de neurones**
   1) Complètement intégré dans G-Code;
   2) Accessible via les processus G-Code Batch et le service Web;
   3) Réseaux de neurones fournissant une bonne précision tout en étant rapide;
   4) Prêt à être utilisé en production dans un Recensement ou dans toute autre enquête comme celles sur la population active(EPA), les dépenses des ménages (EDM), etc..

II. **XGBoost: eXtreme gradient boosted decision trees (arbres de décision)**
   1) Accessible via les processus G-Code Batch et le service Web;
   2) Option pour injecter dans le processus, du code de programmation défini par l'utilisateur ;
   3) Prêt à être utilisé en production, en particulier, pour l'enquête annuelle sur le fret et le camionnage.

# Les algorithmes de réglage des hyper-paramètres

## Quatre algorithmes y sont inclus:

1) Recherche par grille: recherche de toutes les combinaisons possibles;
2) Recherche aléatoire: (par défaut) combinaison de recherches aléatoires;
3) Optimisation aléatoire: recherche de combinaisons au hasard proche de la meilleure valeur actuelle afin de trouver la valeur optimale;
4) Optimisation évolutive: générer une «génération» initiale de modèles, puis les reproduire pour créer progressivement de meilleurs combinaisons.

# L'Enquête sur les dépenses des ménages (EDM)

- Modes de collecte : entrevue et journal
- Pour le journal, les répondants de l'EDM doivent fournir de l'information sur chacune de leurs dépenses pendant une période fixe de temps
- Chaque dépense doit être assignée à un code en se basant sur sa description
- En moyenne, il faut coder ~250,000 items par année
- Pour le cycle 2019, il y a 684 codes EDM au total
- On aimerait donc automatiser le codage
  - Données de 2017 sont les plus récentes

**GOODS AND SERVICES INCLUDING FOOD FROM STORES**

| Item # | Date of expense<br>dd/mm<br>Example: 21/06 | Description of item<br>Write one item per line. Please print.<br>See page 6 in the Diary Guide for help with this section.<br>Reminder: Please enter snacks, beverages and meals purchased from restaurants or fast-food outlets in the section that begins on page 7. | Cost<br>Do not include taxes.<br>$ ¢ |
|---|---|---|---|
| | 21/06 | NO SPENDING | |
| | 22/06 | GAS | 36.00 |
| 1 | 15/04 | Pita bread | 1.79 |
| 2 | 15/04 | Red bell pepper | 3.08 |
| 3 | 15/04 | Money Gift | 500.00 |
| 4 | 16/04 | Lottery tickets | 5.00 |
| 5 | 17/04 | NO SPENDING | |

Par exemple, des codes doivent être assignés aux descriptions *Pita bread, Red bell pepper, Money Gift,* etc..

# Approche

- Codage par appariement direct
  - Jusqu'ici, un fichier de référence est utilisé pour autocoder par appariement direct (taux d'autocodage d'environ 40% pour l'EDM 2017).
  - Un fichier de référence "élargi" récemment développé permet d'atteindre un taux d'autocodage plus élevé (environ 51% pour l'EDM 2017)
  - Clients peu intéressés à maintenir un fichier de référence au cours des années
- Motivation
  - Bons résultats obtenus sur l'autocodage de données de caisse
  - Pourrait-on faire mieux qu'un taux d'autocodage de 50%, en conservant un bon taux d'exactitude?
  - L'apprentissage automatique semble prometteur.

# Apprentissage automatique à l’aide de G-Code

- G-Code
  - Intégration récente de l’algorithme FastText
  - Initialement conçu pour être principalement utilisé avec l’interface graphique
    - Inconvénient:  approche laborieuse et inefficace.
  - Automatiser l’utilisation de G-Code en SAS (application autonome (*standalone*))
    - Macros SAS développées ‘autour’ des procédures de G-Code
      - automatise l’entraînement et le test du modèle,
      - l’optimisation des hyper paramètres et
      - l’évaluation de la qualité des classifications (calculant exactitude, précision et sensibilité )
    - Ces macros sont autonomes et font toutes les conversions de fichier nécessaires

# Développement du modèle

- Préparation des fichiers
  - Création du fichier Maître (standardisation des codes),
  - Des données d’entraînement
    - Plusieurs sous-ensemble des données 2010 à 2016 ont été évalués
    - 2016 seulement, 2015 et 2016, 2014 à 2016, etc…
  - Des données de test (2017)
- Utilisation de G-Code (FastText) pour la modélisation
  - Retourne un score associé à chaque prédiction,
    - entre 0 (aucune confiance) et 10 (confiance absolue)
  - Réglage des hyper paramètres
    - Optimisation pour chacun des ensembles de données d’entraînement

# Développement du modèle

- Création du modèle final
  - basé sur les données de 2014 à 2017
  - Algorithme stable d’un point de vue global et local
  - Résultats aussi validés par l’utilisation d’un algorithme d’apprentissage automatique indépendant
    - Utilisant un réseau de neurones
  - Exactitude  d’environ 94% avec un seuil de score de 7.5+
    - On exclut les codes dont la précision historique est de moins 50%
  - Taux final d’autocodage d’environ 80%
    - Suffisamment bons pour la prochaine production (2021)

# Prochaines étapes

- Intégrer l’utilisation de G-Code dans le système de production
- Évaluer les résultats de l’autocodage sur les données finales de 2019
- Continuer le travail pour améliorer le modèle (par-exemple, inclure le coût de l’item dans le modèle d’apprentissage automatique)
- Entraîner un nouveau modèle après chaque année d’enquête tout en ajoutant les nouveaux items codés manuellement dans le jeu de données d’entraînement
- Développer une approche de contrôle de la qualité pour la production

- Éventuellement : récolter les reçus, les lire et les coder automatiquement

# Conclusion

- La science des données est en pleine expansion à Statistique Canada
  - De façon générale et en particulier pour la classification
  - On cherche à automatiser des processus, à profiter de nouvelles sources de données
  - On veut produire plus rapidement, à un niveau plus détaillé et en minimisant le fardeau de réponse, des statistiques d'aussi bonne qualité
- Nous devons encadrer ces nouvelles initiatives
  - Nous assurer que les modèles et algorithmes sont appropriés et utilisés de façon responsable
  - Mettre en place des processus de contrôle de qualité pour relancer la modélisation
- Nous devons standardiser nos processus et profiter de G-Code pour ce faire
- Nous devons trouver des façons de vulgariser et de faire accepter ces nouvelles méthodes par nos utilisateurs et le public en général
- Nous devons fournir des mesures de qualité reliée à l'utilisation de ces nouveaux outils et suivre leur évolution dans le temps