Unclassified / Non classifié

Les méthodes d'IA en santé publique : session 2

Lisa Lix, Ph. D.
Professeure, *College of Community and Global Health*,
Rady Faculty of Health Sciences, Université du Manitoba

Gemma Postill
Étudiante (M.D./Ph. D), *Clinical Epidemiology and Health Care Research*,
Institute of Health Policy, Management and Evaluation, University of Toronto

Affan Shoukat, Ph. D.
Professeur adjoint en science des données (*SGI*), *Department of Mathematics and Statistics*
Faculty of Science, University of Regina

AI4PH AI FOR PUBLIC HEALTH

Public Health Agency of Canada | Agence de la santé publique du Canada

Les méthodes d’IA en santé publique

01 **Première section : Lisa Lix**
Exploration des données de santé avec l’IA; choix : modèles statistiques c. modèles d’apprentissage automatique

02 **Deuxième section : Gemma Postill**
Modélisation prédictive et des risques

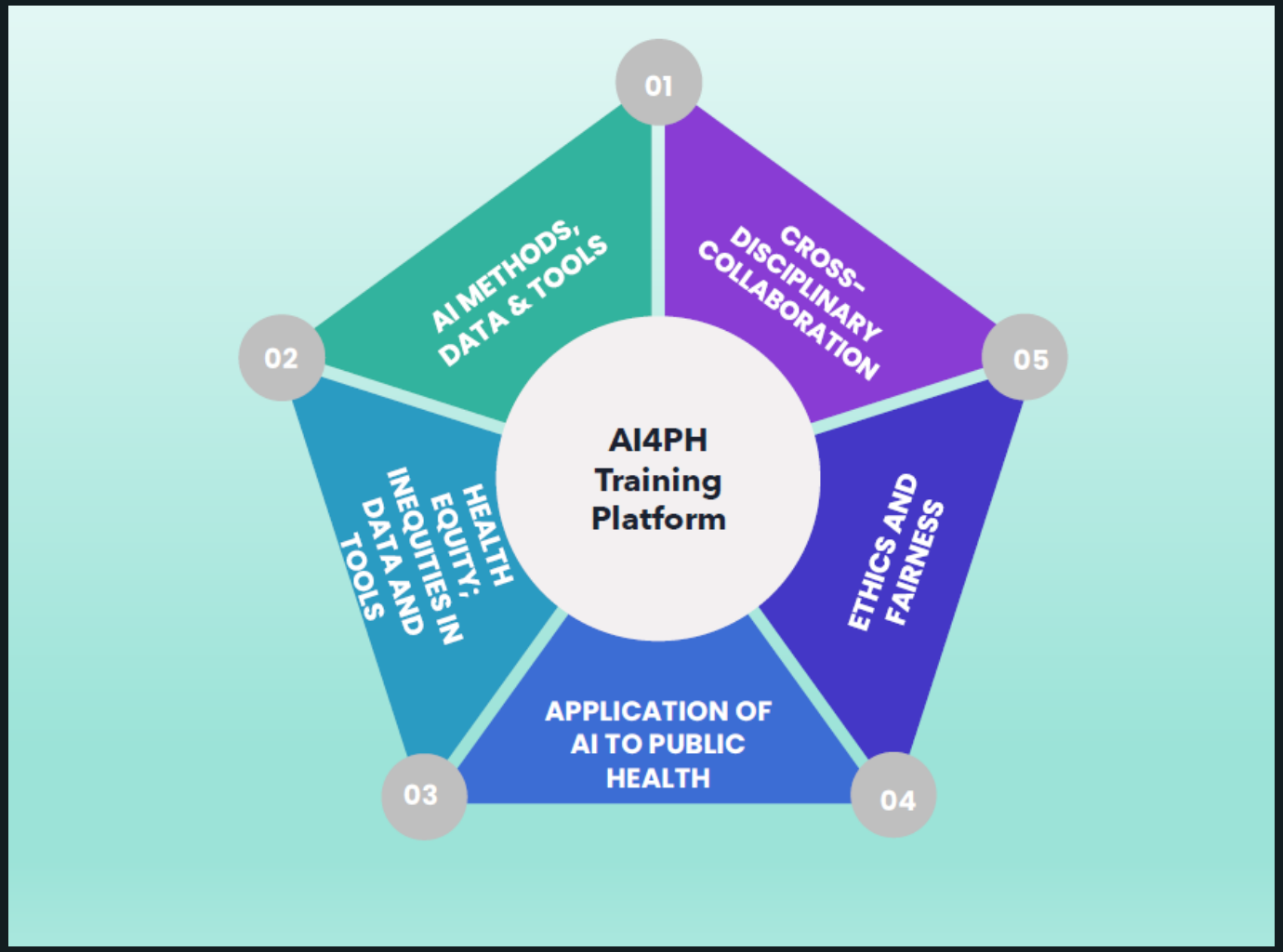
03 **Troisième section : Affan Shoukat**
Apprentissage automatique, prévision et explicabilité

Derrière les chiffres : données de santé et statistiques

Lisa Lix
Professeure et titulaire d’une chaire de recherche du Canada
Université du Manitoba

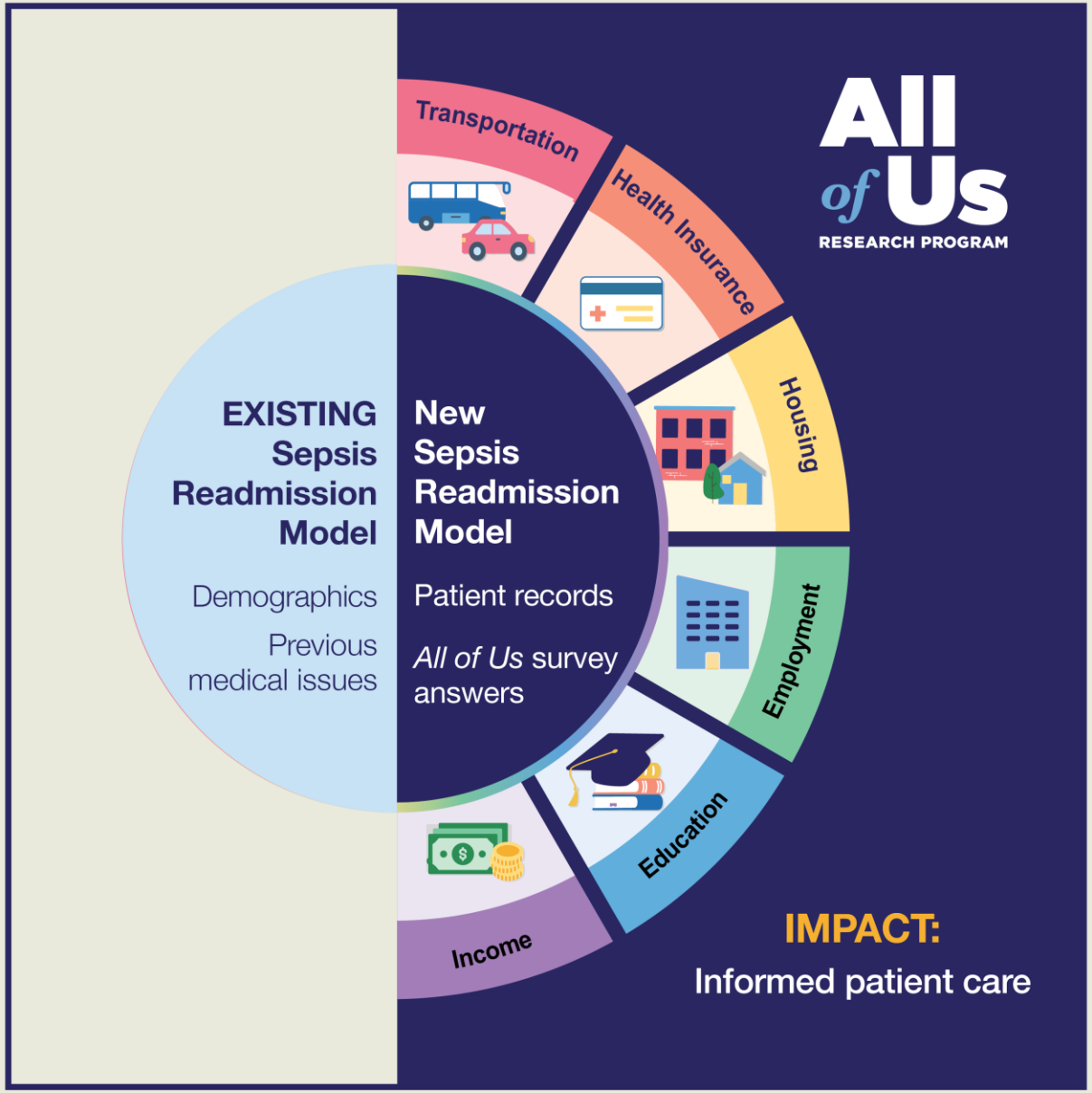


Plateforme de formation nationale d'*AI4PH*



- Formations courtes
- Bourses
- Stages
- Développement du sentiment d'appartenance
- Ateliers

*English only content / Contenu en anglais seulement





Des données riches peuvent permettre d'obtenir de meilleures prévisions

*English only content / Contenu en anglais seulement

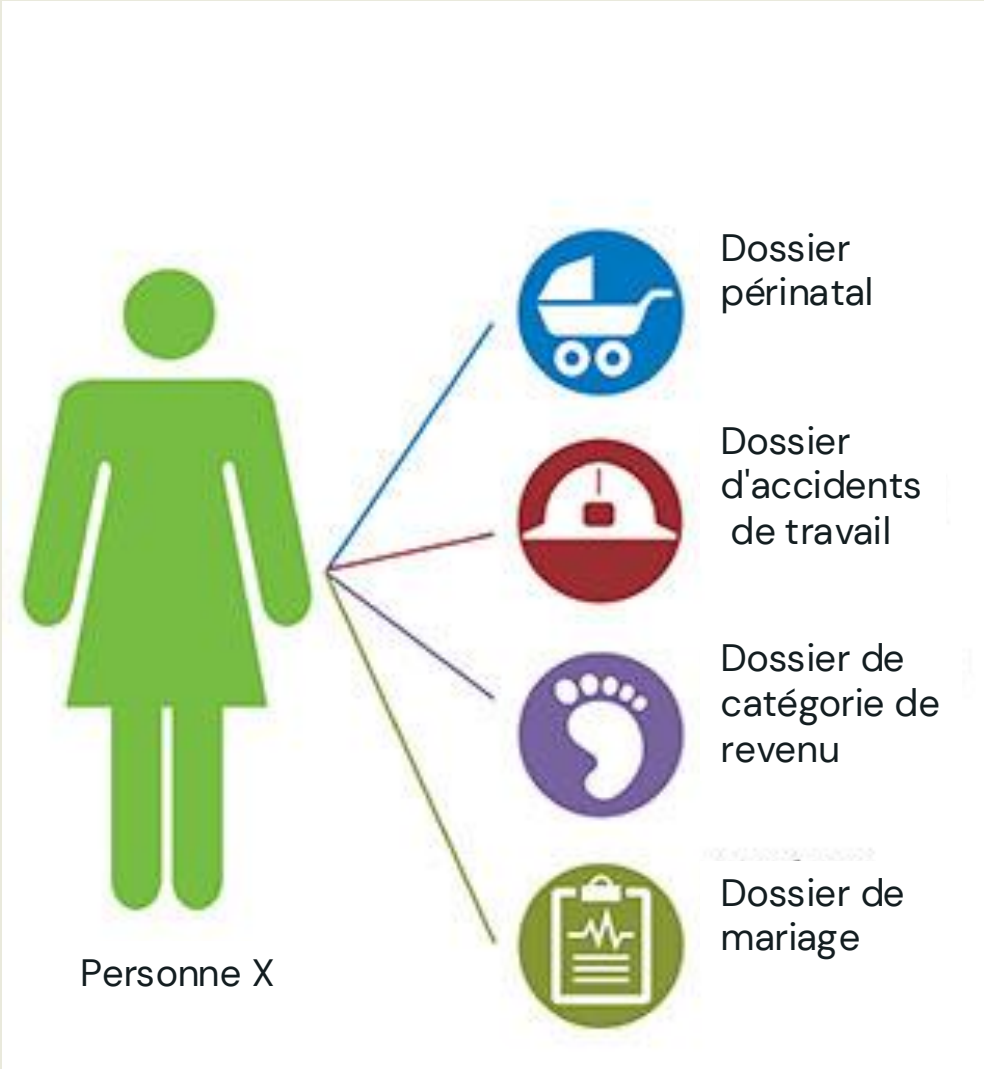
Reference: Amrollahi, F., et al. Inclusion of social determinants of health improves sepsis readmission prediction models. Journal of the American Medical Informatics Association, 2022; 29(7), 1263–1270.

Les données riches proviennent souvent du couplage de sources diverses

Le couplage permet de relier les renseignements sur une personne provenant d'une source à ceux concernant cette même personne provenant d'une autre source

- Couplage déterministe
- Couplage probabiliste

Le couplage est un moyen économique et efficace de recueillir des informations



Source : https://www.popdata.bc.ca/datalinkage

Types de données de santé utilisées dans les applications d'IA



Données d'enquête : instruments d'auto-évaluation transversaux ou longitudinaux qui recueillent des renseignements sur l'état de santé, le recours aux soins de santé et les déterminants de la santé

> Exemples : Enquête sur la santé dans les collectivités canadiennes, Enquête canadienne sur les mesures de la santé

Données issues des médias sociaux : recensement des comportements des utilisateurs et utilisatrices des médias sociaux; comprend les contenus consultés, les interactions et les publications

> Exemples : Twitter, Facebook, Reddit

Types de données de santé utilisées dans les applications d'IA – 2



Données administratives : gestion du système de santé

Exemples : dossier d'hospitalisation, médicaments sur ordonnance

Données de registres : suivi des soins et des traitements prodigués aux personnes atteintes de certains problèmes de santé

Exemples : registres des prothèses de hanche, de cancers

Données des dossiers médicaux électroniques (DME) : utilisation par les médecins pour tenir à jour les renseignements médicaux des personnes hospitalisées ou externes

Bon nombre de ces sources de données constituent des données secondaires, l'utilisateur n'a donc aucun contrôle sur la façon dont elles sont recueillies

Biais dans les données de santé secondaires

- **Biais liés aux données**
  - Valeurs manquantes ou incomplètes
  - Erreurs de saisie
- **Biais d'origine humaine**
  - Biais implicites des prestataires de soins de santé
  - Biais liés à l'orientation et à l'admission : sous-représentation des personnes ayant un accès insuffisant aux soins médicaux
  - Biais de diagnostic
  - Biais liés aux disparités de risque
- **Biais liés aux machines**
  - Biais liés à l'équipement



Repérez les biais

Données issues des médias sociaux	Données issues des registres cliniques
• Biais liés à la population • Biais liés aux méthodes (p. ex. Choix des mots-clés et des mot-clics) • Biais analytiques (p. ex. choix des méthodes de traitement automatique du langage naturel)	• Biais liés à la population • Exhaustivité de la saisie des données • Biais liés aux machines utilisées pour l'enregistrement des mesures cliniques

ELSEVIER

International Journal of Medical Informatics

Volume 164, August 2022, 104804



Review article

Biases in using social media data for public health surveillance: A scoping review

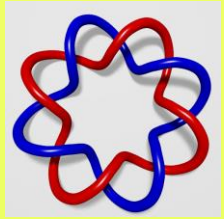
Yunpeng Zhao [a 1] ✉, Xing He [a 1] ✉, Zheng Feng [a] ✉, Sarah Bost [a], Mattia Prosperi [b] ✉, Yonghui Wu [a] ✉, Yi Guo [a] ✉, Jiang Bian [a] ✉

Show more

*English only content / Contenu en anglais seulement

Considérations lors de l’utilisation de données de santé secondaires

Liens



Propriété



Qualité



Accès

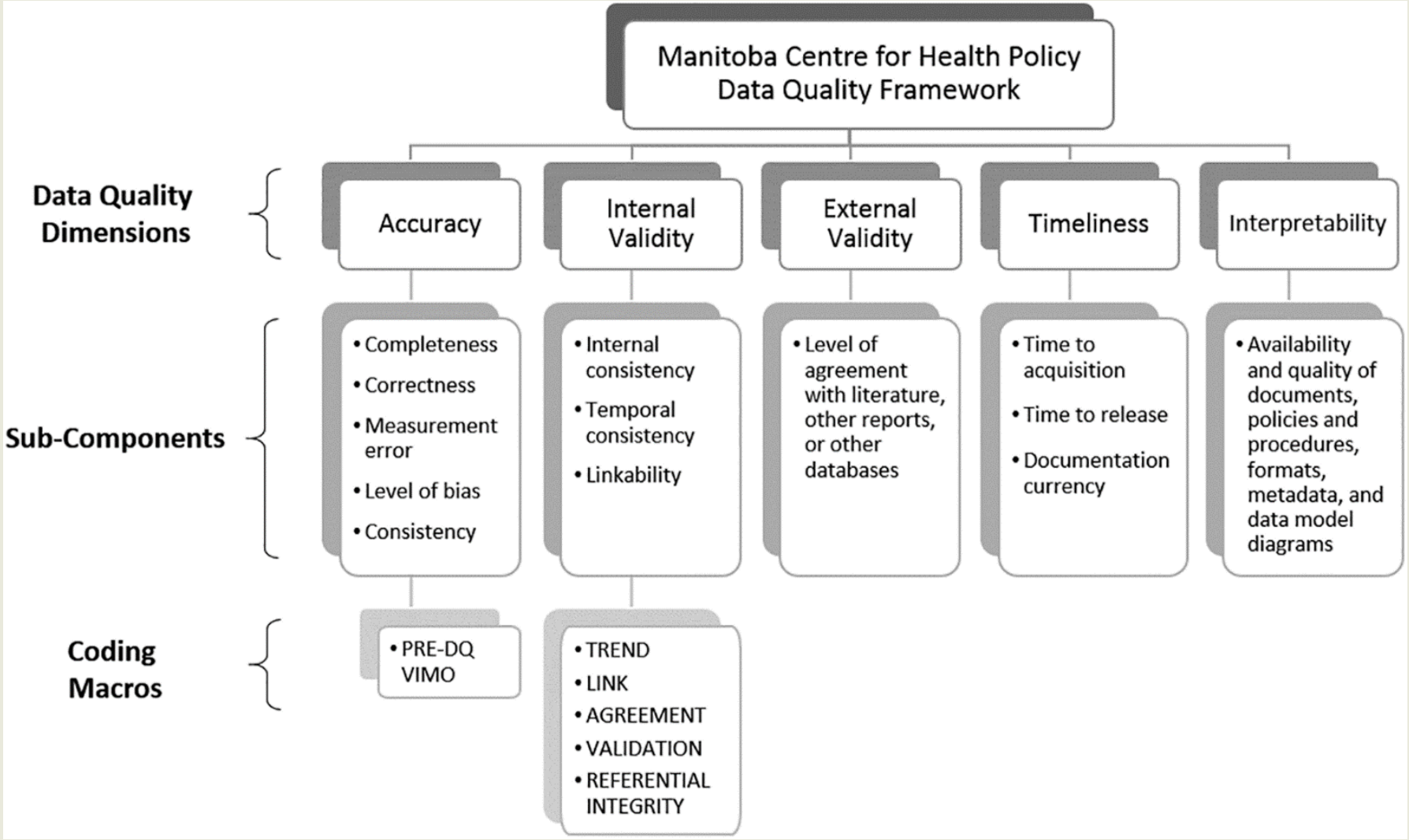


Consentement



Sécurité et confidentialité





Source : *Assessing the quality of administrative data for research: a framework from the Manitoba Centre for Health Policy*. J Am Med Inform Assoc. 2017;25(3):224-229. doi:10.1093/jamia/ocx078

*English only content / Contenu en anglais seulement

Données synthétiques pour protéger la vie privée (et plus)



Définition des données synthétiques (*Alan Turing Institute*, RU) :

- Données générées à l'aide d'un modèle mathématique ou d'un algorithme spécialement conçu à cet effet, dans le but de résoudre une ou plusieurs taches liées à la science des données

Il est possible de créer de nombreux types de données synthétiques :

- Tableaux
- Séries chronologiques
- Texte
- Images
- Vidéo
- Audio

À quoi servent les données synthétiques?

- Protéger la confidentialité des données
- Développer et entraîner les modèles
- Enrichir les ensembles de données existants
- Soutenir les activités d’apprentissage

Avantages des données synthétiques



Protection de la vie privée

- Réduction du risque de divulgation de renseignements réels sur la santé des personnes
- Représentation des caractéristiques des personnes réelles seulement

Amélioration de la disponibilité et de l'accès en matière de données

- Ensemble de données pouvant être librement partagés et utilisés sans restrictions de confidentialité, contrairement aux données de santé réelles

Données améliorées pour l'entraînement des modèles

- Création d'ensembles de données à grande échelle reflétant des populations diverses, des maladies rares ou des scénarios précis qui peuvent être peu représentés dans les données du monde réel

Avantages des données synthétiques – 2



Amélioration de l'entraînement et de la validation des modèles

- Ressource pour l'entraînement et la validation des modèles d'apprentissage automatique, particulièrement pour les maladies rares ou les scénarios complexes où les données du monde réel sont limitées

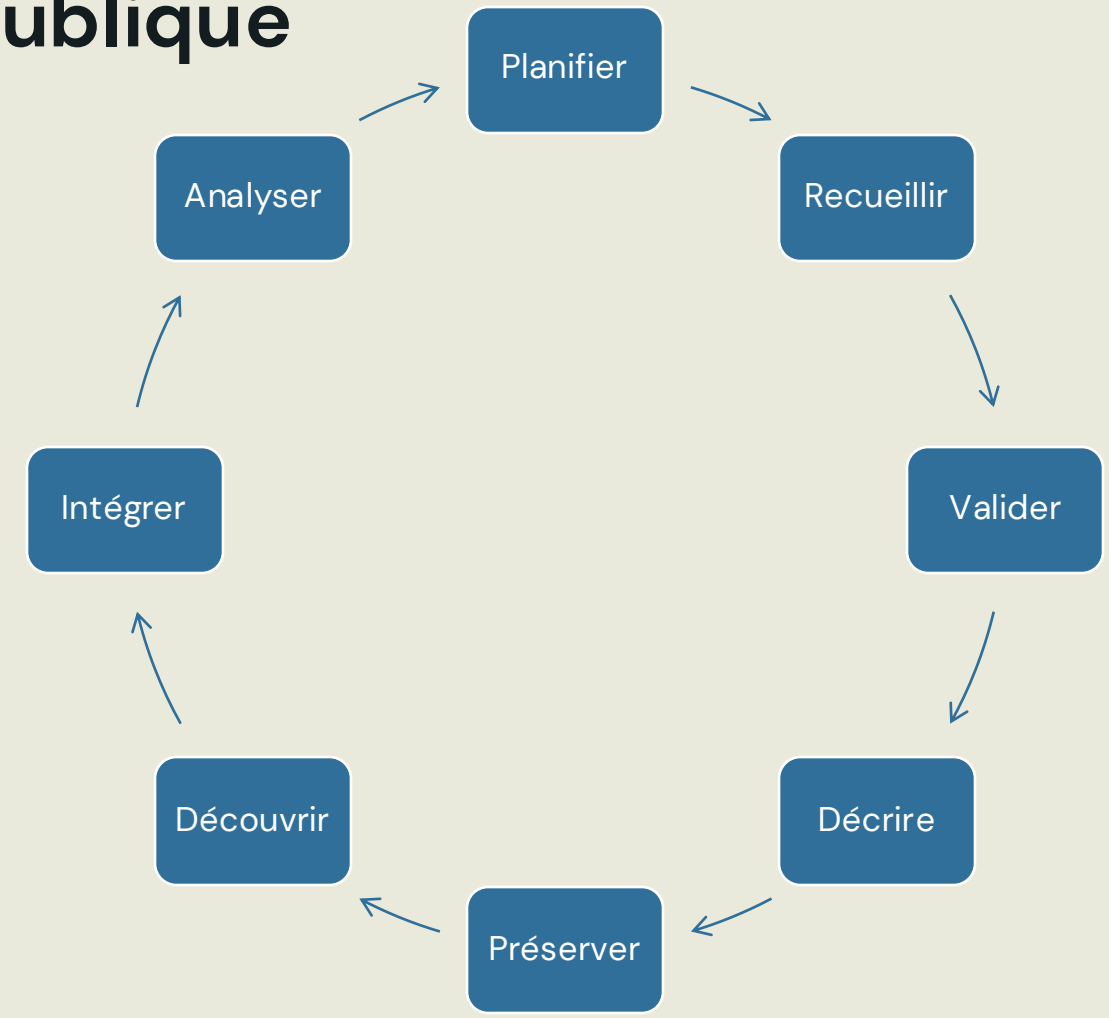
Résolution des biais dans les données

- Création de données synthétiques de qualité possible afin d'atténuer les biais présents dans les données du monde réel, ce qui peut conduire à des algorithmes et des applications plus justes et équitables

Innovation accélérée

- Accès plus rapide aux données pouvant accélérer la recherche, la création de logiciels et l'essai de nouveaux modèles d'IA

Cycle de vie des données : gestion des données de santé publique



Ce cadre décrit la séquence des étapes à suivre pour

- recueillir de nouvelles données ou extraire des données d'un dépôt existant,
- évaluer la qualité des données,
- garantir que les métadonnées et la documentation sont bien saisies, et
- préparer les données en vue de leur analyse (p. ex. étapes de prétraitement)

Source : https://www.dataone.org/data-life-cycle (en anglais seulement)

Modèles statistiques c. apprentissage automatique : comment choisir?

À PRENDRE EN COMPTE

Complexité des données

- Les données non structurées ou de grande dimension peuvent mieux convenir aux modèles d'apprentissage automatique
- Les modèles statistiques traditionnels peuvent convenir à des données plus simples

Volume et corrélation des données

- La régression Lasso et les modèles d'apprentissage automatique basés sur des arbres sont utiles pour identifier les variables les plus « importantes » dans un vaste ensemble de variables potentiellement redondantes

Distribution des données

- Les modèles statistiques supposent que les données présentent des caractéristiques précises, telles qu'une distribution normale, l'indépendance, une variance constante et des relations linéaires
- Les modèles d'apprentissage automatique formulant un minimum d'hypothèses sur la distribution des données, mais nécessitent de grandes quantités de données pour apprendre les modèles et les relations entre celles-ci

Choix du modèle : principales caractéristiques

Caractéristiques	Modèles statistiques	Modèles d'apprentissage automatique (non supervisés)
Question centrale	« Pourquoi cela se produit-il? » (Inférence)	« Que se passera-t-il ensuite? » (Prédiction)
Causes profondes	Mathématiques appliquées / théorie des probabilités	Science informatique / intelligence artificielle
Interprétabilité	Élevée (coefficients, p-values)	Variable ou faible (« boîte noire »)
Exigences en matière de données	Ensembles de données plus petits et plus propres	Données volumineuses, complexes ou non structurées
Outils courants	R, SAS, logiciels de statistiques classiques	Python, TensorFlow, scikit-learn

MESSAGES CLÉS

Les biais présents dans les données secondaires peuvent influencer la précision et la généralisation des modèles d'apprentissage automatique, et, au final, la valeur de ces modèles pour la prise de décision.

Les modèles d'apprentissage automatique ne reposent pas sur des hypothèses rigoureuses concernant les données, mais ne constituent pas toujours le choix optimal (ou nécessaire). Il convient de tenir compte de la question de recherche, ainsi que de la distribution des données.

Merci de votre participation

POUR RESTER EN CONTACT


ai4ph.dlsph@utoronto.ca


@ai4ph.bsky.social


/ai4ph-health-research-training-platform


Espace de travail Slack d'*AI4PH*

AI4PH est financé par les
Instituts de recherche en santé du Canada (IRSC).