

# Apprentissage automatique (Machine learning)

## Concepts de base

Visseho Adjiwanou, PhD.

Démographie, Département de Sociologie, UQAM

# Plan de présentation

- 1 Définition
- 2 Exemples
- 3 Types d'analyse
- 4 Formulation du problème
- 5 Features (Prédicteurs)
- 6 Conseils pratiques
- 7 Bénéfices
- 8 Sujets avancés

# Objectifs

- Comment “l'apprentissage automatique” diffère-t-il des statistiques traditionnelles ?
- Comment est-il utile pour vous, chercheurs en sciences sociales?
- Démystifier l'apprentissage automatique:
  - établir des liens avec les statistiques et l'analyse de données
  - approfondir certains des concepts et méthodes novateurs.

# Introduction

- Origine dans l'informatique, mais fortement influencé par:
  - les mathématiques et
  - les statistiques au cours des 15 à 20 dernières années.
- Nouveau nom de concepts anciens:
  - Régression logistique ==> Méthodes de classification, analyse de regroupement (une forme d'apprentissage non supervisé).
- Apprentissage de nouvelles méthodes :
  - Forêts aléatoires,
  - Machines à vecteurs de support et
  - les réseaux neuronaux.

# Introduction

- Tâches réalisées autrefois effectuées par des systèmes basés sur des règles,
- Demandant énormément de temps aux experts et d'efforts à développer et maintenir ces règles.
- Système rigide, non adaptatifs, difficiles à faire évoluer et coûteux à maintenir.
- Les systèmes de machine learning ont commencé à devenir populaires parce qu'ils pouvaient améliorer le système sur tous ces aspects.

# Introduction

- Introduction non-exhaustive au machine learning : se référer à (**Flach?**; **HastieTibshirani?**; **mitchell1997machine?**).
- Introduction courte et accessible au machine learning pour les scientifiques en sciences sociales
- Aperçu du processus global de machine learning,
- Fournir une introduction intuitive aux méthodes de machine learning,
- Donnero quelques conseils pratiques qui seront utiles pour utiliser ces méthodes, et
- Laisser une grande partie de la théorie statistique aux manuels de machine learning.

# Qu'est-ce que l'apprentissage automatique?

- Lorsque les humains améliorent leurs compétences avec l'expérience, on dit qu'ils **apprennent**. - Est-il également possible de programmer des ordinateurs pour qu'ils fassent de même ?
- Arthur Samuel a inventé le terme apprentissage automatique en 1959 (**samuel1959some?**),
  - Programmer l'ordinateur à jouer au jeu de dame
  - S'améliore après chaque partie
  - Est devenu meilleur que le programmeur humain.
- Autres appellations:
  - Apprentissage statistique,
  - Fouille de données, et
  - Reconnaissance des formes.

# Qu'est-ce que l'apprentissage automatique?

- L'apprentissage automatique a considérablement dépassé l'apprentissage du jeu de dames:
- Les systèmes d'apprentissage automatique ont appris
  - à conduire (et à garer) des voitures autonomes,
  - sont intégrés dans des robots,
  - peuvent recommander des livres,
  - des produits et des films qui nous intéressent (parfois),
  - identifier des médicaments, des protéines et des gènes à étudier pour guérir des maladies,
  - détecter le cancer et d'autres pathologies dans les radiographies et autres types d'imagerie médicale,
  - nous aider à comprendre comment le cerveau humain apprend le langage,
  - aider à identifier quels électeurs sont persuadables lors des élections,
  - détecter quels étudiants sont susceptibles de nécessiter un



# Qu'est-ce que l'apprentissage automatique?

- Au cours des 20 dernières années, AA est devenu un domaine interdisciplinaire:
  - l'informatique,
  - l'intelligence artificielle,
  - les bases de données et
  - les statistiques.
- Au cœur de l'apprentissage automatique, il s'agit de concevoir des systèmes informatiques qui s'améliorent avec le temps et l'expérience.

# Qu'est-ce que l'apprentissage automatique?

- Définition opérationnelle de Tom Mitchell : “Un programme informatique est dit **apprendre de l'expérience  $E$**  par rapport à une **certaine classe de tâches  $T$**  et une **mesure de performance  $P$** , si sa performance à des tâches de la classe  $T$ , mesurée par  $P$ , s'améliore avec l'expérience  $E$ ” (**mitchell1997machine?**).
- Définition intéressante car :
  - centrée sur la tâche
  - permet de considérer l'apprentissage automatique comme un outil utilisé dans un système plus large pour améliorer les résultats qui nous intéressent.

# Exemples d'apprentissage automatique commercial

## 1. *Reconnaissance vocale :*

- Les logiciels de reconnaissance vocale utilisent des algorithmes d'apprentissage automatique construits sur de grandes quantités de données d'entraînement initiales. L'apprentissage automatique permet à ces systèmes de s'adapter et de s'ajuster aux variations individuelles de la parole ainsi qu'aux différents domaines.

# Exemples d'apprentissage automatique commercial

## 2. *Voitures autonomes* :

- Le développement continu des voitures autonomes applique des techniques issues de l'apprentissage automatique. Un ordinateur embarqué analyse en continu les flux vidéo et les capteurs entrants pour surveiller les environs. Les données entrantes sont comparées à des images annotées pour reconnaître des objets tels que les piétons, les feux de signalisation et les nids-de-poule. Afin d'évaluer les différents objets, des ensembles de données d'entraînement énormes sont nécessaires, où des objets similaires ont déjà été identifiés. Cela permet à la voiture autonome de décider des actions à entreprendre ensuite.

# Exemples d'apprentissage automatique commercial

## 3. *Détection de fraude :*

- De nombreuses organisations publiques et privées sont confrontées au problème de la fraude et des abus. Les systèmes d'apprentissage automatique sont largement utilisés pour prendre des cas historiques de fraude et signaler les transactions frauduleuses au fur et à mesure qu'elles se produisent. Ces systèmes ont l'avantage d'être adaptatifs et de s'améliorer avec plus de données au fil du temps.

# Exemples d'apprentissage automatique commercial

## 4. *Publicités personnalisées :*

- De nombreuses boutiques en ligne proposent des recommandations personnalisées promouvant des produits susceptibles d'intéresser les utilisateurs. En se basant sur l'historique d'achat individuel et sur ce que d'autres utilisateurs similaires ont acheté dans le passé, le site prédit les produits qu'un utilisateur pourrait aimer et adapte ses recommandations. Netflix et Amazon sont deux exemples d'entreprises dont les logiciels de recommandation prédisent comment un client évaluerait un certain film ou produit et suggèrent ensuite les articles avec les évaluations prédites les plus élevées. Bien sûr, il y a quelques mises en garde ici, car ils ajustent ensuite les recommandations pour maximiser les profits.

# Exemples d'apprentissage automatique commercial

## 5. *Reconnaissance faciale* :

- Les systèmes de surveillance, les plateformes de réseaux sociaux et les logiciels d'imagerie utilisent tous la détection et la reconnaissance faciale pour d'abord détecter les visages dans les images (ou vidéos) et ensuite les identifier pour diverses tâches. Ces systèmes sont entraînés en fournissant des exemples de visages à un système d'apprentissage automatique qui apprend ensuite à détecter de nouveaux visages et à identifier les individus connus. Le chapitre sur les biais et l'équité mettra en évidence certaines préoccupations concernant ces types de systèmes.

# Exemples de machine learning en sciences sociales

## *1. Prédire les risques d'empoisonnement au plomb:*

- Potash et al (**Potash2015?**) ont collaboré avec le Département de la Santé Publique de Chicago et ont utilisé les forêts aléatoires (une méthode de classification en machine learning) pour prédire quels enfants sont à risque d'empoisonnement au plomb. Ce système d'alerte précoce a ensuite été utilisé pour prioriser les inspections des dangers liés au plomb afin de détecter et de remédier à ces dangers avant qu'ils n'aient un effet néfaste sur l'enfant.



# Exemples de machine learning en sciences sociales

- 2 identifier les policiers indésirables:

# Exemples de machine learning en sciences sociales

- ② identifier les policiers indésirables:
  - Carton et al (**Carton2016?**) ont utilisé une collection de méthodes de machine learning pour identifier les policiers à risque de comportements indésirables, tels que l'utilisation injustifiée de la force ou des tirs injustifiés ou des plaintes soutenues, afin de prioriser les interventions préventives telles que la formation et le conseil.

# Exemples de machine learning en sciences sociales

## 3. *Améliorer la réussite scolaire:*

- **Athey** et Wager (**athey2019?**) utilisent une modification des forêts aléatoires pour estimer les effets de traitement hétérogènes en utilisant un ensemble de données provenant de l'Étude nationale des mentalités d'apprentissage pour évaluer l'impact des interventions visant à améliorer la réussite des élèves.

# Exemples de machine learning en sciences sociales

## 4. *Comprendre le langage policier :*

- Voigt et al (**Voigt2017?**) utilisent des méthodes de machine learning pour analyser les séquences vidéo des caméras portées sur le corps et comprendre le respect du langage des policiers envers les membres des communautés blanche et noire lors des contrôles routiniers.

## Types d'analyse

## Types d'analyse

De nombreuses tâches d'analyse de données effectuées par les scientifiques sociaux peuvent être réparties en quatre types :

- ① Description :



## Types d'analyse

## 2 Détection :





# Types d'analyse

## 3 Prédiction :

- Le but ici est d'utiliser les mêmes données historiques que les méthodes de description et de détection, mais de les utiliser pour prédire des événements et des comportements futurs.

# Types d'analyse

- ④ Changement de comportement (ou Inférence causale) :

# Types d'analyse

- ④ Changement de comportement (ou Inférence causale) :
  - Le but ici est de comprendre les relations causales dans les données afin d'influencer les résultats qui nous importent.

# Le processus de Machine Learning

- 1 Comprendre le problème et l'objectif

# Le processus de Machine Learning

- 1 Comprendre le problème et l'objectif
- 2 Formuler le problème en termes de machine learning

# Le processus de Machine Learning

- 1 Comprendre le problème et l'objectif
- 2 Formuler le problème en termes de machine learning
- 3 Exploration et préparation des données



# Le processus de Machine Learning

- 1 Comprendre le problème et l'objectif
- 2 Formuler le problème en termes de machine learning
- 3 Exploration et préparation des données
- 4 Ingénierie des fonctionnalités (prédicteurs)

# Le processus de Machine Learning

- 1 Comprendre le problème et l'objectif
- 2 Formuler le problème en termes de machine learning
- 3 Exploration et préparation des données
- 4 Ingénierie des fonctionnalités (prédicteurs)
- 5 Modélisation :

# Le processus de Machine Learning

- 1 Comprendre le problème et l'objectif
- 2 Formuler le problème en termes de machine learning
- 3 Exploration et préparation des données
- 4 Ingénierie des fonctionnalités (prédicteurs)
- 5 Modélisation :
- 6 Interprétation du modèle :

# Le processus de Machine Learning

- 1 Comprendre le problème et l'objectif
- 2 Formuler le problème en termes de machine learning
- 3 Exploration et préparation des données
- 4 Ingénierie des fonctionnalités (prédicteurs)
- 5 Modélisation :
- 6 Interprétation du modèle :
- 7 Sélection du modèle :

# Le processus de Machine Learning

- 1 Comprendre le problème et l'objectif
- 2 Formuler le problème en termes de machine learning
- 3 Exploration et préparation des données
- 4 Ingénierie des fonctionnalités (prédicteurs)
- 5 Modélisation :
- 6 Interprétation du modèle :
- 7 Sélection du modèle :
- 8 Validation du modèle

# Le processus de Machine Learning

- 1 Comprendre le problème et l'objectif
- 2 Formuler le problème en termes de machine learning
- 3 Exploration et préparation des données
- 4 Ingénierie des fonctionnalités (prédicteurs)
- 5 Modélisation :
- 6 Interprétation du modèle :
- 7 Sélection du modèle :
- 8 Validation du modèle
- 9 Déploiement et surveillance

# Le processus de Machine Learning

Etapes à suivre pour un problème typique de machine learning :

- ① Comprendre le problème et l'objectif :
  - Cela peut sembler évident mais est souvent non trivial. Les problèmes commencent généralement par des descriptions vagues d'un objectif : améliorer les résultats de santé, augmenter les taux de diplomation, comprendre l'effet d'une variable  $X$  sur un résultat  $Y$ , etc. Il est très important de travailler avec des personnes qui comprennent le domaine étudié pour discuter et définir le problème de manière plus concrète. Quelle est la formulation analytique de la métrique que vous essayez d'améliorer ?

# Le processus de Machine Learning

Etapes à suivre pour un problème typique de machine learning :

- ① Comprendre le problème et l'objectif :
  - Cela peut sembler évident mais est souvent non trivial. Les problèmes commencent généralement par des descriptions vagues d'un objectif : améliorer les résultats de santé, augmenter les taux de diplomation, comprendre l'effet d'une variable X sur un résultat Y, etc. Il est très important de travailler avec des personnes qui comprennent le domaine étudié pour discuter et définir le problème de manière plus concrète. Quelle est la formulation analytique de la métrique que vous essayez d'améliorer ?
  - Le Data Science Project Scoping Guide  
[<http://www.datasciencepublicpolicy.org/home/resources/>



# Le processus de Machine Learning

- ② Formuler le problème en termes de machine learning :
  - S'agit-il d'un problème de classification ou de régression ?

# Le processus de Machine Learning

- ② Formuler le problème en termes de machine learning :
  - S'agit-il d'un problème de classification ou de régression ?
  - L'objectif est-il de construire un modèle qui génère une liste classée par risque d'un résultat, ou de détecter des anomalies à mesure que de nouvelles données arrivent ?

# Le processus de Machine Learning

- ② Formuler le problème en termes de machine learning :
  - S'agit-il d'un problème de classification ou de régression ?
  - L'objectif est-il de construire un modèle qui génère une liste classée par risque d'un résultat, ou de détecter des anomalies à mesure que de nouvelles données arrivent ?
  - Savoir quels types de tâches le machine learning peut résoudre vous permettra de mapper le problème sur lequel vous travaillez à un ou plusieurs cadres de machine learning et vous donnera accès à un ensemble de méthodes appropriées pour cette tâche.

# Le processus de Machine Learning

- ③ Exploration et préparation des données :
  - Ensuite, vous devez explorer soigneusement les données dont vous disposez.

# Le processus de Machine Learning

- ③ Exploration et préparation des données :
  - Ensuite, vous devez explorer soigneusement les données dont vous disposez.
  - Quelles données supplémentaires avez-vous besoin ou auxquelles avez-vous accès ?

# Le processus de Machine Learning

- ③ Exploration et préparation des données :
  - Ensuite, vous devez explorer soigneusement les données dont vous disposez.
  - Quelles données supplémentaires avez-vous besoin ou auxquelles avez-vous accès ?
  - Quelle variable utiliserez-vous pour faire correspondre les enregistrements pour intégrer différentes sources de données ?

# Le processus de Machine Learning

- ③ Exploration et préparation des données :
  - Ensuite, vous devez explorer soigneusement les données dont vous disposez.
  - Quelles données supplémentaires avez-vous besoin ou auxquelles avez-vous accès ?
  - Quelle variable utiliserez-vous pour faire correspondre les enregistrements pour intégrer différentes sources de données ?
  - Quelles variables existent dans le jeu de données ?

# Le processus de Machine Learning

- ③ Exploration et préparation des données :
  - Ensuite, vous devez explorer soigneusement les données dont vous disposez.
  - Quelles données supplémentaires avez-vous besoin ou auxquelles avez-vous accès ?
  - Quelle variable utiliserez-vous pour faire correspondre les enregistrements pour intégrer différentes sources de données ?
  - Quelles variables existent dans le jeu de données ?
  - Sont-elles continues ou catégorielles ?



# Le processus de Machine Learning

- ③ Exploration et préparation des données :
- Ensuite, vous devez explorer soigneusement les données dont vous disposez.
- Quelles données supplémentaires avez-vous besoin ou auxquelles avez-vous accès ?
- Quelle variable utiliserez-vous pour faire correspondre les enregistrements pour intégrer différentes sources de données ?
- Quelles variables existent dans le jeu de données ?
- Sont-elles continues ou catégorielles ?
- Qu'en est-il des valeurs manquantes ?

# Le processus de Machine Learning

- ③ Exploration et préparation des données :
- Ensuite, vous devez explorer soigneusement les données dont vous disposez.
- Quelles données supplémentaires avez-vous besoin ou auxquelles avez-vous accès ?
- Quelle variable utiliserez-vous pour faire correspondre les enregistrements pour intégrer différentes sources de données ?
- Quelles variables existent dans le jeu de données ?
- Sont-elles continues ou catégorielles ?
- Qu'en est-il des valeurs manquantes ?
- Pouvez-vous utiliser les variables dans leur forme originale ou devez-vous les modifier d'une manière ou d'une autre ?

# Le processus de Machine Learning

- ④ Ingénierie des fonctionnalités :
  - Dans le langage du machine learning, ce que vous pourriez connaître sous le nom de variables indépendantes ou prédicteurs ou facteurs ou covariables sont appelés “fonctionnalités”. Créer de bonnes fonctionnalités est probablement l'étape la plus importante dans le processus de machine learning. Cela implique de faire des transformations, de créer des termes d'interaction, ou d'agréger des points de données dans le temps et l'espace.

# Le processus de Machine Learning

## 5 Modélisation :

- Après avoir formulé le problème et créé vos fonctionnalités, vous avez maintenant un ensemble de méthodes parmi lesquelles choisir. Ce serait formidable s'il existait une méthode unique qui fonctionnait toujours le mieux pour un type de problème spécifique, mais cela rendrait les choses trop faciles. Chaque méthode fait une **hypothèse différente** sur la structure et la distribution des données et, avec de grandes quantités de données haute dimension (La dimensionnalité des données fait souvent référence au nombre de variables que nous avons dans les données), il est difficile de savoir a priori quelle hypothèse correspondra le mieux aux données dont nous disposons.

# Le processus de Machine Learning

## 5 Modélisation :

- Après avoir formulé le problème et créé vos fonctionnalités, vous avez maintenant un ensemble de méthodes parmi lesquelles choisir. Ce serait formidable s'il existait une méthode unique qui fonctionnait toujours le mieux pour un type de problème spécifique, mais cela rendrait les choses trop faciles. Chaque méthode fait une **hypothèse différente** sur la structure et la distribution des données et, avec de grandes quantités de données haute dimension (La dimensionnalité des données fait souvent référence au nombre de variables que nous avons dans les données), il est difficile de savoir a priori quelle hypothèse correspondra le mieux aux données dont nous disposons.

# Le processus de Machine Learning

- ⑥ Interprétation du modèle :
  - Une fois que nous avons construit les modèles de machine learning, nous voulons également comprendre ce qu'ils sont, quels prédicteurs ils ont jugé importants, et dans quelle mesure, quels types d'entités ils ont signalé comme étant à haut risque (et pourquoi), où ils ont commis des erreurs, etc. Tous ces aspects relèvent de l'interprétation, de l'interprétabilité et de l'explicabilité des modèles, qui sont actuellement des domaines de recherche actifs en machine learning.

# Le processus de Machine Learning

## 7 Sélection du modèle :

- Lorsque vous construisez un grand nombre de modèles possibles, vous avez besoin d'un moyen de sélectionner le modèle qui est le “meilleur”. Cette partie du chapitre couvrira la méthodologie pour tester les modèles sur des données historiques ainsi que discuter de divers métriques d'évaluation. Il est important de noter que parfois, la littérature sur le machine learning appellera cette étape “validation” en utilisant des données historiques, mais nous voulons la distinguer ici de la validation, qui est l'étape suivante.

# Le processus de Machine Learning

- 8 Validation du modèle :
- L'étape suivante, après la sélection du modèle (à l'aide de données historiques), est la validation. Validez sur de nouvelles données, ainsi qu'en concevant et en réalisant des essais sur le terrain ou des expériences.



# Le processus de Machine Learning

## 9. Déploiement et surveillance :

- Une fois que vous avez sélectionné le meilleur modèle et validé celui-ci en utilisant des données historiques ainsi qu'un essai sur le terrain, vous êtes prêt à mettre le modèle en pratique. Il est important de garder à l'esprit que de nouvelles données seront disponibles, que le monde changera et que le modèle pourrait également (ou devra) changer au fil du temps.

## Section 2

### Formulation du problème

# Formulation du problème : cartographier un problème aux méthodes d'apprentissage automatique

- Lorsque nous travaillons sur un nouveau problème, l'une des premières choses à faire est de le cartographier à une classe de méthodes d'apprentissage automatique.

# Formulation du problème : cartographier un problème aux méthodes d'apprentissage automatique

- Lorsque nous travaillons sur un nouveau problème, l'une des premières choses à faire est de le cartographier à une classe de méthodes d'apprentissage automatique.
- En général, les problèmes que nous aborderons, y compris les exemples ci-dessus, peuvent être regroupés en deux grandes catégories :

# Formulation du problème : cartographier un problème aux méthodes d'apprentissage automatique

- Lorsque nous travaillons sur un nouveau problème, l'une des premières choses à faire est de le cartographier à une classe de méthodes d'apprentissage automatique.
- En général, les problèmes que nous aborderons, y compris les exemples ci-dessus, peuvent être regroupés en deux grandes catégories :
- Apprentissage supervisé

# Formulation du problème : cartographier un problème aux méthodes d'apprentissage automatique

- Lorsque nous travaillons sur un nouveau problème, l'une des premières choses à faire est de le cartographier à une classe de méthodes d'apprentissage automatique.
- En général, les problèmes que nous aborderons, y compris les exemples ci-dessus, peuvent être regroupés en deux grandes catégories :
- Apprentissage supervisé
- Apprentissage non supervisé

# Apprentissage supervisé

- Ce sont des problèmes où il existe une variable **cible** (continue ou discrète) que nous voulons prédire ou classer.

# Apprentissage supervisé

- Ce sont des problèmes où il existe une variable **cible** (continue ou discrète) que nous voulons prédire ou classer.
- La classification, la prédiction et la régression entrent toutes dans cette catégorie.



# Apprentissage supervisé

- Ce sont des problèmes où il existe une variable **cible** (continue ou discrète) que nous voulons prédire ou classer.
- La classification, la prédiction et la régression entrent toutes dans cette catégorie.
- Les méthodes d'apprentissage supervisé prédisent une valeur  $Y$ , données des entrées  $X$  en apprenant (ou estimant, ajustant, entraînant) une fonction  $F$ , où  $F(X) = Y$ .

# Apprentissage supervisé

- X est l'ensemble des variables (connues sous le nom de features en apprentissage automatique, ou dans d'autres domaines comme prédicteurs) fournies en entrée et

# Apprentissage supervisé

- X est l'ensemble des variables (connues sous le nom de features en apprentissage automatique, ou dans d'autres domaines comme prédicteurs) fournies en entrée et
- Y est la variable cible/dépendante ou une étiquette (comme elle est connue en apprentissage automatique).

# Apprentissage supervisé

- X est l'ensemble des variables (connues sous le nom de features en apprentissage automatique, ou dans d'autres domaines comme prédicteurs) fournies en entrée et
- Y est la variable cible/dépendante ou une étiquette (comme elle est connue en apprentissage automatique).
- L'objectif des méthodes d'apprentissage supervisé est de rechercher cette fonction  $F$  qui estime ou prédit le mieux  $Y$

# Apprentissage supervisé

- $X$  est l'ensemble des variables (connues sous le nom de features en apprentissage automatique, ou dans d'autres domaines comme prédicteurs) fournies en entrée et
- $Y$  est la variable cible/dépendante ou une étiquette (comme elle est connue en apprentissage automatique).
- L'objectif des méthodes d'apprentissage supervisé est de rechercher cette fonction  $F$  qui estime ou prédit le mieux  $Y$
- Lorsque la sortie  $Y$  est catégorielle, cela s'appelle classification.

# Apprentissage supervisé

- $X$  est l'ensemble des variables (connues sous le nom de features en apprentissage automatique, ou dans d'autres domaines comme prédicteurs) fournies en entrée et
- $Y$  est la variable cible/dépendante ou une étiquette (comme elle est connue en apprentissage automatique).
- L'objectif des méthodes d'apprentissage supervisé est de rechercher cette fonction  $F$  qui estime ou prédit le mieux  $Y$
- Lorsque la sortie  $Y$  est catégorielle, cela s'appelle classification.
- Lorsque  $Y$  est une valeur continue, cela s'appelle régression.

# Apprentissage supervisé

- Distinction clé en apprentissage automatique est que l'objectif n'est pas seulement de trouver la meilleure fonction  $F$  qui peut estimer ou prédire  $Y$  pour des résultats observés (les  $Y$  connus), mais de trouver celle qui **se généralise** le mieux aux nouvelles données non vues, souvent dans le futur.

# Apprentissage supervisé

- Distinction clé en apprentissage automatique est que l'objectif n'est pas seulement de trouver la meilleure fonction  $F$  qui peut estimer ou prédire  $Y$  pour des résultats observés (les  $Y$  connus), mais de trouver celle qui **se généralise** le mieux aux nouvelles données non vues, souvent dans le futur.
- Cette distinction rend les méthodes **plus axées sur la généralisation** et moins sur **le simple ajustement des données** que nous avons au mieux possible.



# Apprentissage supervisé

- Il est important de noter que vous faites cela implicitement lors de la réalisation de régressions en n'ajoutant pas de termes d'ordre supérieur pour obtenir de meilleures statistiques d'ajustement.

## Apprentissage supervisé

- Il est important de noter que vous faites cela implicitement lors de la réalisation de régressions en n'ajoutant pas de termes d'ordre supérieur pour obtenir de meilleures statistiques d'ajustement.
- En obtenant de meilleures statistiques d'ajustement, nous **sur-ajustons** les données et la performance sur les nouvelles données (non vues) diminue souvent.

## Apprentissage supervisé

- Il est important de noter que vous faites cela implicitement lors de la réalisation de régressions en n'ajoutant pas de termes d'ordre supérieur pour obtenir de meilleures statistiques d'ajustement.
- En obtenant de meilleures statistiques d'ajustement, nous **sur-ajustons** les données et la performance sur les nouvelles données (non vues) diminue souvent.
- Les méthodes comme le lasso (**tibshirani1996regression?**) pénalisent le modèle pour avoir trop de termes en réalisant ce que l'on appelle la **régularisation**. (En termes statistiques, la régularisation est une tentative d'éviter le sur-ajustement du modèle).

# Apprentissage non supervisé

- Ce sont des problèmes où il **n'existe pas** de variable cible que nous voulons prédire, mais nous voulons comprendre les regroupements ou les motifs “naturels” dans les données.

# Apprentissage non supervisé

- Ce sont des problèmes où il **n'existe pas** de variable cible que nous voulons prédire, mais nous voulons comprendre les regroupements ou les motifs “naturels” dans les données.
- Le clustering est l'exemple le plus courant de ce type d'analyse où vous avez  $X$  et voulez regrouper les  $X$  similaires ensemble.

## Apprentissage non supervisé

- Ce sont des problèmes où il **n'existe pas** de variable cible que nous voulons prédire, mais nous voulons comprendre les regroupements ou les motifs “naturels” dans les données.
- Le clustering est l'exemple le plus courant de ce type d'analyse où vous avez  $X$  et voulez regrouper les  $X$  similaires ensemble.
- Vous avez peut-être entendu parler de la “segmentation” utilisée dans le monde du marketing pour regrouper les clients similaires en utilisant des techniques de clustering.

# Apprentissage non supervisé

- Ce sont des problèmes où il **n'existe pas** de variable cible que nous voulons prédire, mais nous voulons comprendre les regroupements ou les motifs “naturels” dans les données.
- Le clustering est l'exemple le plus courant de ce type d'analyse où vous avez  $X$  et voulez regrouper les  $X$  similaires ensemble.
- Vous avez peut-être entendu parler de la “segmentation” utilisée dans le monde du marketing pour regrouper les clients similaires en utilisant des techniques de clustering.
- L'analyse en composantes principales (ACP) et les méthodes apparentées entrent également dans la catégorie de l'apprentissage non supervisé.

# Apprentissage supervisé ou non supervisé

- Entre les deux extrêmes de l'apprentissage supervisé et non supervisé, il existe un spectre de méthodes qui ont différents niveaux de supervision impliqués (Figure (ref?)(fig)).



# Apprentissage supervisé ou non supervisé

- Entre les deux extrêmes de l'apprentissage supervisé et non supervisé, il existe un spectre de méthodes qui ont différents niveaux de supervision impliqués (Figure (ref?)(fig)).
- La supervision dans ce cas est la présence de variables cibles (connues en apprentissage automatique comme étiquettes).



# Apprentissage supervisé ou non supervisé

- Entre les deux extrêmes de l'apprentissage supervisé et non supervisé, il existe un spectre de méthodes qui ont différents niveaux de supervision impliqués (Figure (ref?)(fig)).
- La supervision dans ce cas est la présence de variables cibles (connues en apprentissage automatique comme étiquettes).
- En apprentissage non supervisé, aucun des points de données n'a d'étiquettes.
- En apprentissage supervisé, tous les points de données ont des étiquettes.



## Section 3

### Features (Prédicteurs)

# Features

- Avant d'aborder les modèles et les méthodes, nous devons transformer nos données brutes en “features”.
- En sciences sociales, elles ne sont pas appelées features mais plutôt connues sous le nom de variables ou prédicteurs (ou covariables si vous faites de la régression).

# Features

- Avant d'aborder les modèles et les méthodes, nous devons transformer nos données brutes en "features".
- En sciences sociales, elles ne sont pas appelées features mais plutôt connues sous le nom de variables ou prédicteurs (ou covariables si vous faites de la régression).
- De bonnes features sont ce qui rend les systèmes d'apprentissage automatique efficaces.

# Features

- Avant d'aborder les modèles et les méthodes, nous devons transformer nos données brutes en "features".
- En sciences sociales, elles ne sont pas appelées features mais plutôt connues sous le nom de variables ou prédicteurs (ou covariables si vous faites de la régression).
- De bonnes features sont ce qui rend les systèmes d'apprentissage automatique efficaces.
- La génération de features (ou l'ingénierie des features, comme on l'appelle souvent) est là où une grande partie du temps est consacrée dans le processus d'apprentissage automatique.



# Features

- Avant d'aborder les modèles et les méthodes, nous devons transformer nos données brutes en "features".
- En sciences sociales, elles ne sont pas appelées features mais plutôt connues sous le nom de variables ou prédicteurs (ou covariables si vous faites de la régression).
- De bonnes features sont ce qui rend les systèmes d'apprentissage automatique efficaces.
- La génération de features (ou l'ingénierie des features, comme on l'appelle souvent) est là où une grande partie du temps est consacrée dans le processus d'apprentissage automatique.
- C'est aussi la phase où les recherches et les enseignements antérieurs du domaine abordé peuvent être intégrés dans le processus d'apprentissage automatique.

# Features

- En tant que chercheurs ou praticiens en sciences sociales, vous avez passé beaucoup de temps à construire des features, en utilisant des transformations, des variables binaires et des termes d'interaction.

# Features

- En tant que chercheurs ou praticiens en sciences sociales, vous avez passé beaucoup de temps à construire des features, en utilisant des transformations, des variables binaires et des termes d'interaction.
- Tout cela est toujours nécessaire et crucial dans le cadre de l'apprentissage automatique.



# Features

- Il est courant d'avoir des modèles entraînés sur des milliers de features.

# Features

- Il est courant d'avoir des modèles entraînés sur des milliers de features.
- Bien sûr, il est important de garder à l'esprit qu'augmenter le nombre de features nécessite d'avoir suffisamment de données pour éviter le sur-ajustement.

# Les approches utilisées pour créer des features

- 1 Transformations, telles que le logarithme, le carré et la racine carrée.

# Les approches utilisées pour créer des features

- 1 Transformations, telles que le logarithme, le carré et la racine carrée.
- 2 Variables binaires (indicatrices) : Cela se fait souvent en prenant des variables catégorielles (comme la ville) et en créant une variable binaire pour chaque valeur (une variable pour chaque ville dans les données). Celles-ci sont également appelées variables indicatrices.



# Les approches utilisées pour créer des features

- 1 Transformations, telles que le logarithme, le carré et la racine carrée.
- 2 Variables binaires (indicatrices) : Cela se fait souvent en prenant des variables catégorielles (comme la ville) et en créant une variable binaire pour chaque valeur (une variable pour chaque ville dans les données). Celles-ci sont également appelées variables indicatrices.
- 3 Discrétisation : Plusieurs méthodes nécessitent que les features soient discrètes plutôt que continues. Plusieurs approches existent pour convertir des variables continues en variables discrètes, la plus courante étant le binning à intervalles égaux.

## Les approches utilisées pour créer des features

- 4 Agrégation : Les features agrégées constituent souvent la majorité des features pour un problème donné. Ces agrégations utilisent différentes fonctions d'agrégation (compte, minimum, maximum, moyenne, écart type, etc.), souvent sur des fenêtres de temps et d'espace variées. Par exemple, données urbaines, nous voudrions calculer le nombre (et minimum, maximum, moyenne, variance) de crimes dans un rayon de  $m$  miles d'une adresse dans les  $t$  mois passés pour des valeurs variables de  $m$  et  $t$ , et ensuite utiliser toutes ces valeurs comme features dans un problème de classification. Les features d'agrégation spatio-temporelles vont être extrêmement importantes lorsque vous construirez des modèles d'apprentissage automatique.



# Les approches utilisées pour créer des features

- Garder le modèle simple le rend plus rapide à entraîner et plus facile à comprendre et à expliquer.

# Vocabulaire de l'apprentissage automatique

- Apprentissage : En apprentissage automatique, vous remarquerez le terme apprentissage qui sera utilisé dans le contexte de “l'apprentissage” d'un modèle. C'est ce que vous connaissez probablement sous le nom d'ajustement ou d'estimation d'une fonction, ou encore d'entraînement ou de construction d'un modèle. Ces termes sont tous des synonymes et sont utilisés de manière interchangeable dans la littérature de l'apprentissage automatique.
  - Exemples : Ce sont des points de données, des lignes, des observations ou des instances.

# Vocabulaire de l'apprentissage automatique

- Features : Ce sont les variables indépendantes, les attributs, les variables prédictives et les variables explicatives.
- Étiquettes : Cela inclut la variable de réponse, la variable dépendante, la variable cible ou les résultats.
- Sous-ajustement : Cela se produit lorsqu'un modèle est trop simple et ne capture pas suffisamment bien la structure des données.

# Vocabulaire de l'apprentissage automatique

- Sur-ajustement : Cela se produit lorsqu'un modèle est possiblement trop complexe et modélise le bruit dans les données, ce qui peut entraîner une mauvaise performance de généralisation. Utiliser des mesures in-sample pour la sélection de modèles peut en être la cause.
- Régularisation : C'est une méthode générale pour éviter le sur-ajustement en appliquant des contraintes supplémentaires au modèle appris. Par exemple, dans la construction de modèles de régression logistique, une approche courante est de s'assurer que les poids du modèle (coefficients) sont, en moyenne, de faible amplitude.

# Vocabulaire de l'apprentissage automatique

- Deux régularisations courantes sont :
  - la régularisation  $L1$  (utilisée par le lasso), qui a un terme de pénalité encourageant la somme des valeurs absolues des paramètres à être petite ; et
  - la régularisation  $L2$ , qui encourage la somme des carrés des paramètres à être petite.



# Interprétabilité des modèles

- En tant que scientifiques sociaux (ou bons praticiens de l'apprentissage automatique), nous ne nous contentons pas de construire des modèles d'apprentissage automatique, mais nous voulons aussi comprendre ce que les modèles ont “appris” et comment les utiliser pour faire des inférences et des décisions.
- Comprendre ou interpréter les modèles d'apprentissage automatique est une exigence clé pour la plupart des problèmes de sciences sociales et de politiques publiques.

# Interprétabilité des modèles

- Il y a plusieurs raisons à cela, notamment :
  - Fournir des informations pouvant aider à déboguer et à améliorer les modèles
  - Augmenter la confiance dans les modèles et par conséquent accroître leur adoption par les décideurs
  - Améliorer les décisions prises en utilisant les modèles en renforçant les prédictions correctes et en aidant le décideur à ignorer les mauvaises prédictions.
  - Aider à sélectionner les interventions appropriées basées sur les explications
  - Offrir des recours juridiques aux personnes affectées par les décisions prises en utilisant les modèles

# Explications globales versus au niveau individuel

Lorsque nous pensons à l'interprétabilité des modèles, il existe deux types d'interprétabilité :

- Globale : Au niveau global du modèle
- Individuelle : Expliquer une classification/prédiction individuelle faite par un modèle
- Les deux sont importantes pour différentes raisons.

## Explications globales versus au niveau individuel

- Nous avons besoin d'interprétabilité globale pour aider à comprendre le modèle dans son ensemble, mais nous avons aussi besoin d'explications pour les classifications individuelles lorsque ces modèles aident une personne à prendre des décisions sur des cas individuels.
- Un travailleur social identifiant le risque pour un client de retourner dans un refuge pour sans-abris et déterminant les interventions appropriées pour réduire ce risque, ou un conseiller dans une agence pour l'emploi déterminant la probabilité pour une personne d'être au chômage de longue durée et la connectant avec des programmes de formation ou des opportunités d'emploi appropriés, ont besoin d'explications au niveau individuel des prédictions/recommandations générées par le modèle d'apprentissage automatique.

# Interprétabilité globale

- Chaque méthode aboutit à un modèle qui doit être interprété de manière appropriée pour cette méthode.
- Par exemple, pour un arbre de décision, nous pouvons vouloir visualiser l'arbre pour comprendre quels types de classifications sont faites.
- Cela peut bien sûr devenir fastidieux et difficile si l'arbre est extrêmement grand.

# Interprétabilité globale

- Pour les modèles de régression logistique, nous pouvons examiner les coefficients et les rapports de cotes, mais il est souvent difficile de prendre en compte mentalement les différentes variables se contrôlant mutuellement.
- En général, les modèles discutés ci-dessus ont différentes manières d'exposer leur "importance des features" et nous utilisons souvent cela comme proxy pour l'interprétabilité globale du modèle.
- Une autre façon d'interpréter un modèle est de comprendre comment le modèle note les points de données individuels.

# Interprétabilité globale

- Nous pouvons prendre l'ensemble des entités notées par le modèle, et générer des tableaux croisés qui mettent en évidence comment le top  $x\%$  des entités notées/prédites diffèrent du reste des entités.
- Cette approche nous permet d'avoir une idée de ce que fait le modèle, non pas en général, mais sur les entités qui nous intéressent, et rend l'interprétabilité un peu plus intuitive et généralisable à travers différents types de modèles.
- Une approche différente adoptée par certains dans ce domaine a été d'utiliser des modèles sparsifiés (Utilisant un petit nombre de features/prédicteurs), les rendant plus faciles à interpréter.

# Interprétabilité globale

- La motivation derrière ces modèles simples et sparsifiés est qu'ils sont intrinsèquement interprétables et ne nécessitent pas l'utilisation d'analyses supplémentaires pour que les humains les comprennent.
- Ces modèles peuvent ne pas bien fonctionner pour chaque tâche, il est donc important pour nous d'explorer la gamme de modèles en termes de performance et de complexité, et de décider quel niveau et type d'interprétabilité nous avons besoin et comment équilibrer cela avec la précision (Nous utilisons la précision comme proxy pour différentes mesures de performance basées sur la matrice de confusion telles que la précision, le rappel, etc.) de ces modèles.



## Explications au niveau individuel

- Bien qu'il soit important de comprendre les modèles que nous construisons à un niveau global, dans de nombreuses applications des sciences sociales, nous voulons obtenir une explication de pourquoi un point de données a été classé/prédit d'une certaine manière par le modèle.
- Beaucoup de travaux récents ont porté sur les méthodes permettant de générer des explications au niveau individuel pour les prédictions faites par les modèles d'apprentissage automatique.

# Explications au niveau individuel

- Ces méthodes se divisent en deux catégories :
  - Méthodes spécifiques aux modèles : Elles sont utilisées pour générer des explications pour les prédictions faites par une classe spécifique de méthodes, telles que les réseaux neuronaux ou les forêts aléatoires.

# Explications au niveau individuel

- Ces méthodes se divisent en deux catégories :
  - Méthodes spécifiques aux modèles : Elles sont utilisées pour générer des explications pour les prédictions faites par une classe spécifique de méthodes, telles que les réseaux neuronaux ou les forêts aléatoires.
  - Méthodes agnostiques aux modèles : Elles peuvent être utilisées pour générer des explications pour des prédictions individuelles faites par n'importe quel type de modèle.

## Explications au niveau individuel

- Il est important de garder à l'esprit que ces “explications” ne sont généralement pas causales et se limitent souvent à être une liste classée de “features”.

## Explications au niveau individuel

- Il est important de garder à l'esprit que ces “explications” ne sont généralement pas causales et se limitent souvent à être une liste classée de “features”.
- Une manière de les envisager est que ces features étaient les plus importantes pour attribuer à ce point de données le score qu'il a reçu par un modèle particulier.

## Explications au niveau individuel

- Il est important de garder à l'esprit que ces “explications” ne sont généralement pas causales et se limitent souvent à être une liste classée de “features”.
- Une manière de les envisager est que ces features étaient les plus importantes pour attribuer à ce point de données le score qu'il a reçu par un modèle particulier.
- C'est actuellement un domaine de recherche actif en apprentissage automatique et on espère qu'il mûrira en un ensemble de méthodes et d'outils utiles pour les scientifiques sociaux utilisant l'apprentissage automatique pour résoudre des problèmes nécessitant une compréhension meilleure et plus profonde de leurs prédictions.

## Section 4

### Conseils pratiques

# Conseils pratiques

- 1 Éviter les fuites de données



# Conseils pratiques

- 1 Éviter les fuites de données
- 2 Utiliser un proxy pour la variable de résultat (étiquette) en tant que caractéristique

## Conseils pratiques

- 1 Éviter les fuites de données
- 2 Utiliser un proxy pour la variable de résultat (étiquette) en tant que caractéristique
- 3 Utiliser l'ensemble des données pour les imputations.

# Conseils pratiques

- 1 Éviter les fuites de données
- 2 Utiliser un proxy pour la variable de résultat (étiquette) en tant que caractéristique
- 3 Utiliser l'ensemble des données pour les imputations.
- 4 Utiliser l'ensemble des données pour les discrétisations ou les normalisations

## Conseils pratiques

- 1 Éviter les fuites de données
- 2 Utiliser un proxy pour la variable de résultat (étiquette) en tant que caractéristique
- 3 Utiliser l'ensemble des données pour les imputations.
- 4 Utiliser l'ensemble des données pour les discrétisations ou les normalisations
- 5 Utiliser l'ensemble des données pour la sélection des caractéristiques

## Conseils pratiques

- 1 Éviter les fuites de données
- 2 Utiliser un proxy pour la variable de résultat (étiquette) en tant que caractéristique
- 3 Utiliser l'ensemble des données pour les imputations.
- 4 Utiliser l'ensemble des données pour les discrétisations ou les normalisations
- 5 Utiliser l'ensemble des données pour la sélection des caractéristiques
- 6 Utiliser des informations provenant de l'avenir

# Conseils pratiques

- 1 Éviter les fuites de données
- 2 Utiliser un proxy pour la variable de résultat (étiquette) en tant que caractéristique
- 3 Utiliser l'ensemble des données pour les imputations.
- 4 Utiliser l'ensemble des données pour les discrétisations ou les normalisations
- 5 Utiliser l'ensemble des données pour la sélection des caractéristiques
- 6 Utiliser des informations provenant de l'avenir
- 7 Utiliser (des proxys/transformation de) les résultats futurs comme caractéristiques

## Conseils pratiques

- 8 Effectuer une validation croisée k-fold standard lorsque vous avez des données temporelles.

## Conseils pratiques

- 8 Effectuer une validation croisée k-fold standard lorsque vous avez des données temporelles.
- 9 Utiliser des données (comme des caractéristiques) qui se sont produites avant l'heure de l'entraînement du modèle mais qui ne seront disponibles que plus tard.



## Conseils pratiques

- 8 Effectuer une validation croisée k-fold standard lorsque vous avez des données temporelles.
- 9 Utiliser des données (comme des caractéristiques) qui se sont produites avant l'heure de l'entraînement du modèle mais qui ne seront disponibles que plus tard.
- 10 Utiliser des données (en tant que lignes) dans l'ensemble d'entraînement en fonction d'informations provenant de l'avenir.

## Conseils pratiques

- 8 Effectuer une validation croisée k-fold standard lorsque vous avez des données temporelles.
- 9 Utiliser des données (comme des caractéristiques) qui se sont produites avant l'heure de l'entraînement du modèle mais qui ne seront disponibles que plus tard.
- 10 Utiliser des données (en tant que lignes) dans l'ensemble d'entraînement en fonction d'informations provenant de l'avenir.
- 11 Sélectionner certains modèles, caractéristiques et autres choix de conception basés sur les humains (développeurs ML, experts du domaine) sachant ce qui s'est passé dans le futur

## Conseils pratiques

- ⑧ Effectuer une validation croisée k-fold standard lorsque vous avez des données temporelles.
- ⑨ Utiliser des données (comme des caractéristiques) qui se sont produites avant l'heure de l'entraînement du modèle mais qui ne seront disponibles que plus tard.
- ⑩ Utiliser des données (en tant que lignes) dans l'ensemble d'entraînement en fonction d'informations provenant de l'avenir.
- ⑪ Sélectionner certains modèles, caractéristiques et autres choix de conception basés sur les humains (développeurs ML, experts du domaine) sachant ce qui s'est passé dans le futur
- ⑫ En règle générale, si vous rencontrez un modèle d'apprentissage automatique qui fonctionne vraiment bien, c'est probablement parce que vous avez commis une erreur

## Conseils pratiques

Conseils pratiques utiles lors de l'utilisation de l'apprentissage automatique.

- 1 Éviter les fuites de données

# Conseils pratiques

Conseils pratiques utiles lors de l'utilisation de l'apprentissage automatique.

- 1 Éviter les fuites de données
  - Les fuites de données se produisent lorsque votre modèle a accès à des données au moment de l'entraînement/construction qu'il n'aurait pas au moment du test/déploiement/prédiction. Le résultat est un modèle trop optimiste qui fonctionne beaucoup moins bien lorsqu'il est déployé.

## Conseils pratiques

Conseils pratiques utiles lors de l'utilisation de l'apprentissage automatique.

- ① Éviter les fuites de données
  - Les fuites de données se produisent lorsque votre modèle a accès à des données au moment de l'entraînement/construction qu'il n'aurait pas au moment du test/déploiement/prédiction. Le résultat est un modèle trop optimiste qui fonctionne beaucoup moins bien lorsqu'il est déployé.
  - Les formes les plus courantes de fuites se produisent en raison de problèmes temporels - inclure des données futures dans votre modèle parce que vous les avez lors de la sélection du modèle, mais il existe de nombreuses autres façons

## Conseils pratiques

- 2 La plus grande (et évidente) erreur : Utiliser un proxy pour la variable de résultat (étiquette) en tant que caractéristique. Celle-ci est souvent facile à détecter car vous obtenez une performance parfaite, mais elle est plus nuancée lorsque le proxy est une approximation de l'étiquette/variable de résultat et que l'augmentation de la performance est plus subtile à détecter facilement. Effectuer toute transformation ou inférence en utilisant l'ensemble des données

## Conseils pratiques

- ② La plus grande (et évidente) erreur : Utiliser un proxy pour la variable de résultat (étiquette) en tant que caractéristique. Celle-ci est souvent facile à détecter car vous obtenez une performance parfaite, mais elle est plus nuancée lorsque le proxy est une approximation de l'étiquette/variable de résultat et que l'augmentation de la performance est plus subtile à détecter facilement. Effectuer toute transformation ou inférence en utilisant l'ensemble des données
- ③ Utiliser l'ensemble des données pour les imputations. Faites toujours l'imputation basée uniquement sur votre ensemble d'entraînement, pour chaque ensemble d'entraînement. Inclure l'ensemble de test permet aux informations de s'infiltrer dans vos modèles, surtout dans les cas où le monde change à l'avenir.



## Conseils pratiques

- 4 Utiliser l'ensemble des données pour les discrétisations ou les normalisations/mises à l'échelle ou de nombreuses autres transformations basées sur les données. Même raison que #2. La gamme d'une variable (comme l'âge par exemple) peut changer à l'avenir et le savoir fera paraître vos modèles meilleurs qu'ils ne le sont réellement.

## Conseils pratiques

- ④ Utiliser l'ensemble des données pour les discrétisations ou les normalisations/mises à l'échelle ou de nombreuses autres transformations basées sur les données. Même raison que #2. La gamme d'une variable (comme l'âge par exemple) peut changer à l'avenir et le savoir fera paraître vos modèles meilleurs qu'ils ne le sont réellement.
- ⑤ Utiliser l'ensemble des données pour la sélection des caractéristiques. Mêmes raisons que #2 et #3. Pour jouer la sécurité, d'abord divisez en ensembles d'entraînement et de test, puis faites tout ce que vous avez à faire en utilisant ces données.

## Conseils pratiques

- 6 Utiliser des informations provenant de l'avenir (qui ne seront pas disponibles au moment de l'entraînement ou de la prédiction)

## Conseils pratiques

- 6 Utiliser des informations provenant de l'avenir (qui ne seront pas disponibles au moment de l'entraînement ou de la prédiction)
- 7 Utiliser (des proxys/transformation de) les résultats futurs comme caractéristiques : similaire à #1

## Conseils pratiques

- ⑧ Effectuer une validation croisée k-fold standard lorsque vous avez des données temporelles. Si vous avez des données temporelles (qui ne sont pas stationnaires), la validation croisée k-fold va mélanger les données et un ensemble d'entraînement contiendra (probablement) des données futures et un ensemble de test contiendra (probablement) des données passées.

## Conseils pratiques

- 8 Effectuer une validation croisée k-fold standard lorsque vous avez des données temporelles. Si vous avez des données temporelles (qui ne sont pas stationnaires), la validation croisée k-fold va mélanger les données et un ensemble d'entraînement contiendra (probablement) des données futures et un ensemble de test contiendra (probablement) des données passées.
- 9 Utiliser des données (comme des caractéristiques) qui se sont produites avant l'heure de l'entraînement du modèle mais qui ne seront disponibles que plus tard. C'est assez courant dans les cas où il y a un décalage/délai dans la collecte ou l'accès aux données. Un événement peut se produire aujourd'hui, mais il n'apparaît dans la base de données que plus tard, une semaine, un mois ou un an plus tard, et bien qu'il soit

## Conseils pratiques

- 10 Utiliser des données (en tant que lignes) dans l'ensemble d'entraînement en fonction d'informations provenant de l'avenir. Inclure des lignes correspondant à certains critères (à l'avenir) dans l'ensemble d'entraînement, comme tout le monde qui a reçu un service social dans les 3 prochains mois, fuit des informations à votre modèle via un ensemble d'entraînement biaisé.

## Conseils pratiques

- 10 Utiliser des données (en tant que lignes) dans l'ensemble d'entraînement en fonction d'informations provenant de l'avenir. Inclure des lignes correspondant à certains critères (à l'avenir) dans l'ensemble d'entraînement, comme tout le monde qui a reçu un service social dans les 3 prochains mois, fuit des informations à votre modèle via un ensemble d'entraînement biaisé.
- 11 Sélectionner certains modèles, caractéristiques et autres choix de conception basés sur les humains (développeurs ML, experts du domaine) sachant ce qui s'est passé dans le futur. Il s'agit d'une zone grise - nous voulons utiliser toutes nos connaissances du domaine pour construire des systèmes plus efficaces, mais parfois cela peut ne pas se généraliser à l'avenir et entraîner des modèles surajustés/trop optimistes au



## Conseils pratiques

- 12 En règle générale, si vous rencontrez un modèle d'apprentissage automatique qui fonctionne vraiment bien, c'est probablement parce que vous avez commis une erreur qui entraîne une fuite de données. Une façon d'approfondir est de regarder les importances des caractéristiques de votre modèle pour voir si la ou les caractéristiques les plus importantes peuvent être à l'origine de cette fuite.

# Pipeline d'apprentissage automatique

- Lorsque vous travaillez sur des projets d'apprentissage automatique, il est judicieux de structurer votre code comme un pipeline modulaire afin de pouvoir essayer facilement différentes approches et méthodes sans avoir à effectuer de réorganisation majeure.
- Un bon pipeline contiendra des modules pour :
  - l'importation des données,
  - l'exploration,
  - la génération de caractéristiques,
  - la classification, et
  - l'évaluation.
- Vous pouvez ensuite instancier un flux de travail spécifique en combinant ces modules.

# Pipeline d'apprentissage automatique

- Un composant important du pipeline d'apprentissage automatique est la comparaison des différentes méthodes.
- Avec toutes les méthodes disponibles et tous les hyperparamètres qui les accompagnent, comment savons-nous quel modèle utiliser et quels hyperparamètres sélectionner ?
- Et que se passe-t-il lorsque nous ajoutons de nouvelles caractéristiques au modèle ou lorsque les données présentent une “dérive temporelle” et changent avec le temps ?

## Pipeline d'apprentissage automatique

- Une approche simple consiste à avoir un ensemble imbriqué de boucles qui parcourent toutes les méthodes auxquelles vous avez accès, puis énumèrent tous les hyperparamètres pour cette méthode, créent un produit cartésien et les parcourent tous, les comparant à travers différents critères d'évaluation et sélectionnant le meilleur à utiliser par la suite.
- Vous pouvez même ajouter différents sous-ensembles de caractéristiques et tranches temporelles à cette boucle
- Triage (<http://github.com/dssg/triage>) est un bon exemple de pipeline d'apprentissage automatique conçu pour résoudre de nombreux problèmes de politique publique.

## Section 5

### Bénéfices

# Bénéfices pour les scientifiques sociaux

## *-1. Utilisation de meilleures méthodes et méthodologies de prédiction*

- >- Génération de contrefactuels
- >- Appariement

## *-2. Erreur de spécification du modèle*

- 3. Meilleure analyse de texte :      -4. Sondages adaptatifs*
- 5. Estimation des effets de traitement hétérogènes*
- 6. Sélection de variables*

# Bénéfices pour les scientifiques sociaux

## 1. Utilisation de meilleures méthodes et méthodologies de prédiction :

- Les statistiques traditionnelles et les sciences sociales n'ont pas beaucoup mis l'accent sur les méthodes de prédiction.

# Bénéfices pour les scientifiques sociaux

## 1. Utilisation de meilleures méthodes et méthodologies de prédiction :

- Les statistiques traditionnelles et les sciences sociales n'ont pas beaucoup mis l'accent sur les méthodes de prédiction.
- Les chercheurs en apprentissage automatique ont passé les 30 dernières années à développer et à adapter des méthodes axées sur cette tâche.



# Bénéfices pour les scientifiques sociaux

## 1. Utilisation de meilleures méthodes et méthodologies de prédiction :

- Les statistiques traditionnelles et les sciences sociales n'ont pas beaucoup mis l'accent sur les méthodes de prédiction.
- Les chercheurs en apprentissage automatique ont passé les 30 dernières années à développer et à adapter des méthodes axées sur cette tâche.
- Nous croyons qu'il y a beaucoup de valeur pour les chercheurs et les praticiens en sciences sociales à en apprendre davantage sur ces méthodes, à les appliquer, voire à les augmenter (**Kleinberg2015?**).

# Bénéfices pour les scientifiques sociaux

## 1. Utilisation de meilleures méthodes et méthodologies de prédiction :

- Deux tâches courantes qui peuvent être améliorées en utilisant de meilleures méthodes de prédiction sont :
  - la génération de contrefactuels (essentiellement un problème de prédiction) et
  - l'appariement.
- De plus, les ensembles de rétention et la validation croisée peuvent être utilisés comme méthodologie de sélection de modèle avec toutes les méthodes de régression et de classification existantes, ce qui entraîne une amélioration de la sélection de modèle et des estimations d'erreur.

# Bénéfices pour les scientifiques sociaux

## 2. Erreur de spécification du modèle :

- Les régressions linéaires et logistiques sont des techniques courantes d'analyse de données en sciences sociales.
- Une hypothèse fondamentale dans les deux cas est qu'ils sont additifs sur les paramètres.
- L'apprentissage automatique fournit des outils lorsque cette hypothèse est trop limitante.

# Bénéfices pour les scientifiques sociaux

## 2. Erreur de spécification du modèle :

- Hainmueller et Hazlett (**hainmueller2014kernel?**), par exemple, réanalysent des données qui ont été initialement analysées avec une régression logistique et parviennent à des conclusions substantiellement différentes.
- Ils soutiennent que leur analyse, qui est plus flexible et basée sur une méthodologie d'apprentissage supervisé, fournit trois insights supplémentaires par rapport au modèle original.

# Bénéfices pour les scientifiques sociaux

## 2. Erreur de spécification du modèle :

- Premièrement, les performances prédictives sont similaires
- Deuxièmement, leur modèle leur permet de calculer des effets
- Enfin, la réanalyse découvre également des interactions

# Bénéfices pour les scientifiques sociaux

## 3. Meilleure analyse de texte :

- Le texte est partout, mais malheureusement les humains sont lents et coûteux dans l'analyse des données textuelles.
- Ainsi, les ordinateurs sont nécessaires pour analyser de grandes collections de texte.
- Les méthodes d'apprentissage automatique peuvent aider à rendre ce processus plus efficace.

# Bénéfices pour les scientifiques sociaux

## 3. Meilleure analyse de texte :

- Feldman et Sanger (**FeldmanSanger?**) donnent un aperçu de différentes méthodes automatiques pour l'analyse de texte.
- Grimmer et Stewart (**grimmer2013text?**) donnent des exemples plus spécifiques pour les scientifiques sociaux.

# Bénéfices pour les scientifiques sociaux

## 4. Sondages adaptatifs :

- Certaines questions de sondage ont un grand nombre de catégories de réponses possibles.
- Par exemple, les classifications professionnelles internationales décrivent plus de 500 catégories professionnelles, et il est prohibitif de poser toutes les catégories lors du sondage.
- Au lieu de cela, les répondants répondent à une question ouverte sur leur travail et les algorithmes d'apprentissage automatique peuvent utiliser les réponses textuelles pour suggérer de petits ensembles d'options de réponse plausibles.
- Les répondants peuvent ensuite sélectionner quelle option décrit le mieux leur occupation, ce qui permet d'économiser les coûts de codage après l'entrevue (**Schierholz2018?**).



# Bénéfices pour les scientifiques sociaux

## 5. Estimation des effets de traitement hétérogènes :

- Une approche standard de l'inférence causale est l'attribution de différents traitements (par exemple, des médicaments) aux unités d'intérêt (par exemple, les patients).
- Les chercheurs calculent ensuite généralement l'effet moyen du traitement - la différence moyenne des résultats pour les deux groupes.

# Bénéfices pour les scientifiques sociaux

## 5. Estimation des effets de traitement hétérogènes :

- Il est également intéressant de savoir si les effets de traitement diffèrent pour différents sous-groupes (par exemple, un médicament est-il plus efficace pour les jeunes ?).
- L'analyse de sous-groupes traditionnelle a été critiquée et remise en question par diverses techniques d'apprentissage automatique (**green2012modeling?**; **imai2013estimating?**).

# Bénéfices pour les scientifiques sociaux

## 6. Sélection de variables :

- Bien qu'il existe de nombreuses méthodes de sélection de variables, les méthodes régularisées telles que le lasso sont très efficaces et efficientes lorsqu'elles sont confrontées à de grandes quantités de données.
- Varian (**varian2014big?**) entre dans plus de détails et donne d'autres méthodes d'apprentissage automatique qui peuvent être utiles pour la sélection de variables.
- Nous pouvons également trouver des interactions entre paires de variables (à alimenter dans d'autres modèles) en utilisant des forêts aléatoires, en examinant les variables qui co-occurrent dans le même arbre, et en calculant la force de l'interaction en fonction du nombre d'arbres dans lesquels elles co-occurrent, de leur occurrence dans les arbres et de leur

## Section 6

### Sujets avancés

## Sujets avancés

- ① L'apprentissage semi-supervisé, où une combinaison de données étiquetées et non étiquetées est utilisée pour l'entraînement, en fonction d'un ensemble d'hypothèses.
- De telles méthodes sont utiles lorsque l'étiquetage des données est coûteux et lorsque des données non étiquetées peuvent aider à améliorer les modèles d'apprentissage automatique. Consultez le volume édité de MIT Press (**Chapelle2006?**) pour des explications et des exemples.

## Sujets avancés

- ② Les systèmes de recommandation :
  - Ceux-ci sont couramment utilisés par les services audio et vidéo comme YouTube pour générer des listes de lecture ou par les boutiques en ligne comme Amazon pour suggérer des produits supplémentaires qu'un client pourrait souhaiter acheter. Plus généralement, les systèmes de recommandation visent à prédire les préférences qu'un utilisateur pourrait avoir. Une stratégie consiste à recommander des produits qui ont des caractéristiques similaires à ceux déjà sélectionnés par le même utilisateur (indépendamment des autres). Une autre stratégie recommande un produit si d'autres personnes ayant un profil similaire ont sélectionné le même produit dans le passé.

# Sujets avancés

- ③ L'apprentissage actif
  - Un ensemble d'algorithmes d'apprentissage automatique qui interrogent l'utilisateur ou une autre source d'information pour obtenir des étiquettes pour les points de données qui sont les plus bénéfiques pour les modèles d'apprentissage automatique.

# Sujets avancés

- ③ L'apprentissage actif
  - Un ensemble d'algorithmes d'apprentissage automatique qui interrogent l'utilisateur ou une autre source d'information pour obtenir des étiquettes pour les points de données qui sont les plus bénéfiques pour les modèles d'apprentissage automatique.
  - Cela s'oppose au processus standard d'apprentissage automatique où nous sélectionnons souvent aléatoirement les points de données à étiqueter. Les approches d'apprentissage actif pour la sélection de points de données à étiqueter ont montré qu'elles réduisaient l'effort nécessaire pour entraîner les modèles d'apprentissage automatique.



## Sujets avancés

- ④ L'apprentissage par renforcement :
- Les méthodes d'apprentissage automatique “supervisées” sont “à un coup” et prennent des points de données et des étiquettes en entrée.

## Sujets avancés

- ④ L'apprentissage par renforcement :
  - Les méthodes d'apprentissage automatique “supervisées” sont “à un coup” et prennent des points de données et des étiquettes en entrée.
  - L'apprentissage par renforcement est un paradigme d'apprentissage automatique différent où le programme d'apprentissage automatique prend une série d'actions/décisions et reçoit un retour d'information différé (récompense ou pénalité) lors de l'exécution d'une tâche.

## Sujets avancés

- ④ L'apprentissage par renforcement :
  - Les méthodes d'apprentissage automatique “supervisées” sont “à un coup” et prennent des points de données et des étiquettes en entrée.
  - L'apprentissage par renforcement est un paradigme d'apprentissage automatique différent où le programme d'apprentissage automatique prend une série d'actions/décisions et reçoit un retour d'information différé (récompense ou pénalité) lors de l'exécution d'une tâche.
  - L'objectif de l'apprentissage par renforcement est de déterminer la prochaine meilleure action à prendre afin de maximiser les performances à long terme. Cela a été appliqué à des scénarios tels que les jeux (jeu de dames, échecs,

# Conclusion

- Les méthodes que nous allons voir deviendront bientôt si courantes que vous ne pouvez plus vous limiter à la régression linéaire ou logistique dans vos analyses.

## Ressources