PROBLÈMES ET DÉFIS

APERÇU

1. Mauvaises données
2. Surapprentissage
3. Mégadonnées
4. Pertinence et portabilité
5. Biais, sophisme, interprétation
6. Mythes et erreurs
7. Avenir de la science des données, de l'intelligence artificielle et de l'apprentissage machine
8. Conclusion

OBJECTIFS D'APPRENTISSAGE

Être capable de décrire, à un niveau élevé, quelques-uns des problèmes et défis courants associés à l'analyse des données.

Comprendre la valeur d'un modèle approximatif.

Connaître la description des 5 V des mégadonnées.

Apprécier les utilisations appropriées des résultats de la science des données.

Prendre conscience de certains types courants de biais en science des données.

Prendre conscience de certains mythes et erreurs classiques en science des données.

MAUVAISES DONNÉES

PROBLÈMES ET DÉFIS

« Nous *disons* tous que nous aimons les données, mais ce n'est pas vrai. Ce que nous aimons, c'est obtenir des perspectives grâce aux données. Cela n'équivaut pas tout à fait à aimer les données. En fait, j'ose dire que je ne me soucie pas vraiment des données, et il semblerait que je ne suis pas le seul. »

(Q.E. McCallum, *Bad Data Handbook*)

MAUVAISES DONNÉES

L'ensemble de données répond-il aux **critères**?

- Entrées non valides, observations anormales, etc.

Données mises en forme en vue de la consommation humaine, pas de la lisibilité par les machines

Difficultés de **traitement de texte**

- Codage
- Caractères propres à l'application

MAUVAISES DONNÉES

Recueil de données **en ligne**

- Légalité de l'obtention des données
- Stockage des versions hors ligne

Détection des **mensonges** et des **inexactitudes**

- Signalisation des erreurs (mensonges ou inexactitudes)
- Utilisation d'un langage polarisant

Données et réalité

- Mauvaises données
- Mauvaise réalité?

MAUVAISES DONNÉES

Sources de **biais** et d'**erreurs**

- Biais d'imputation
- Codage supérieur ou inférieur (remplacement des valeurs extrêmes par des valeurs moyennes)
- Déclaration par procuration (chef du ménage pour le ménage)

Recherche de la **perfection**

- Données universitaires
- Données professionnelles
- Données gouvernementales
- Données relatives au service

MAUVAISES DONNÉES

Embûches de la science des données

- Analyse sans compréhension
- Utilisation d'un seul outil (par choix ou par décret)
- Analyse pour l'analyse
- Attentes irréalistes à l'égard de la science des données
- Selon le besoin de savoir, et vous n'avez pas besoin de savoir

Comparaison entre les bases de données, les fichiers et l'informatique en nuage

- Le nuage résoudra tous nos problèmes!

MAUVAISES DONNÉES

Quand est-ce **assez proche, assez bon**?

- Exhaustivité
- Cohérence
- Exactitude
- Responsabilité

DISCUSSION

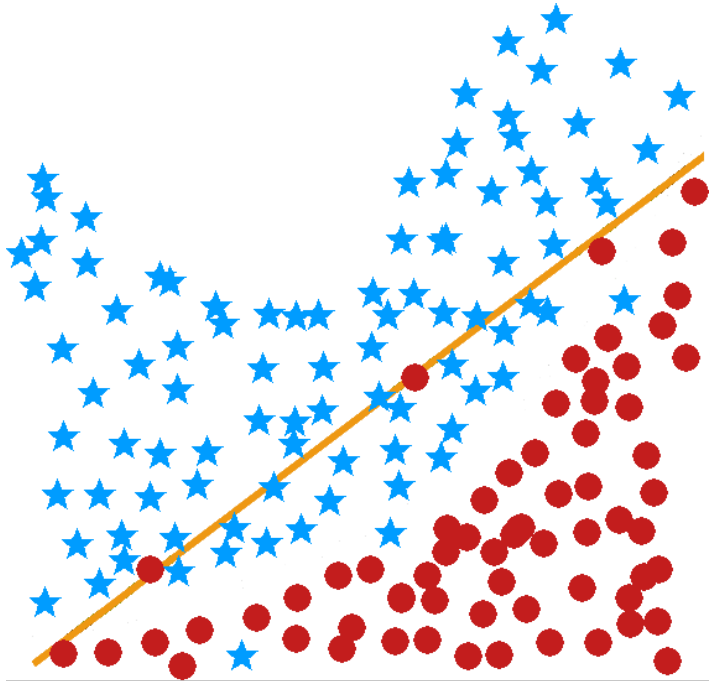
Selon la diction, les mauvaises informations sont synonymes de mauvaises conclusions. Quelles sont les conséquences pour les entreprises et les politiques publiques de la prise de décisions sur la base de mauvaises données?

SURAPPRENTISSAGE

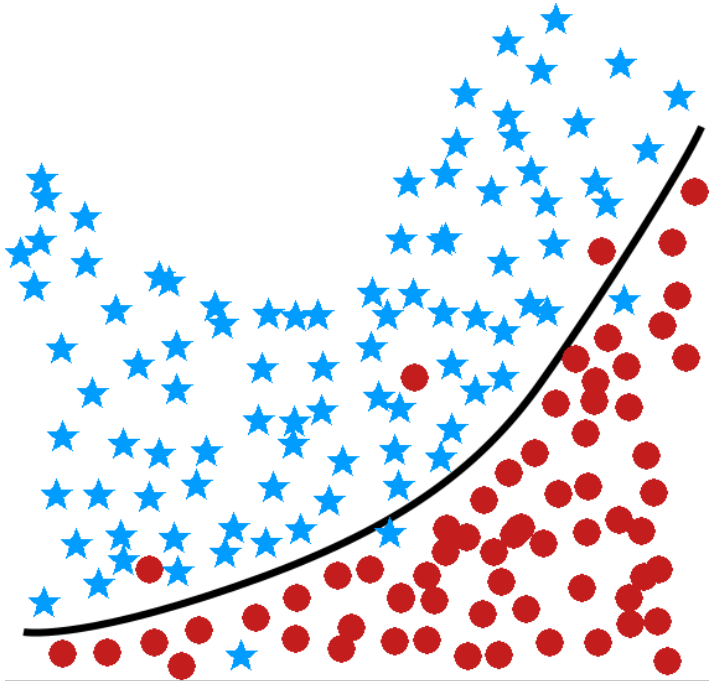
PROBLÈMES ET DÉFIS

(AMAR GONDALIYA, PINGAX [EN ANGLAIS SEULEMENT])

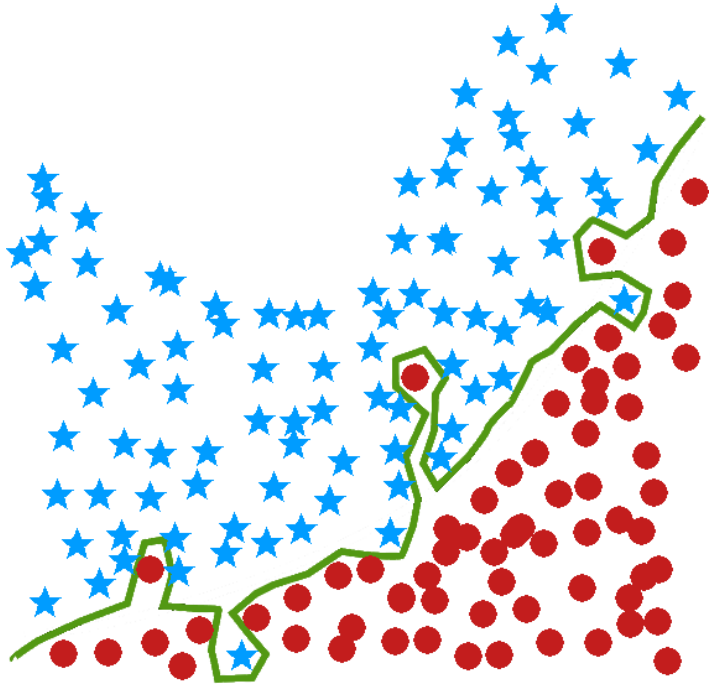
Boucle d'Or et les trois modèles



Sous-apprentissage



Bonne représentation



Surapprentissage

NOTIONS FONDAMENTALES

On espère que les règles ou modèles générés par n'importe quelle technique sur un **ensemble d'apprentissage** puissent être généralisés à de **nouvelles données** (ou **ensembles de validation/d'essai**).

Des problèmes surviennent lorsque les connaissances acquises grâce à un **apprentissage supervisé** ne se généralisent pas correctement aux données.

L'**apprentissage non supervisé** peut également être touché.

Ironiquement, cela peut se produire si les règles ou les modèles s'adaptent **trop bien** à l'ensemble d'apprentissage – les résultats sont **trop étroitement liés à l'ensemble d'apprentissage**.

EXEMPLE

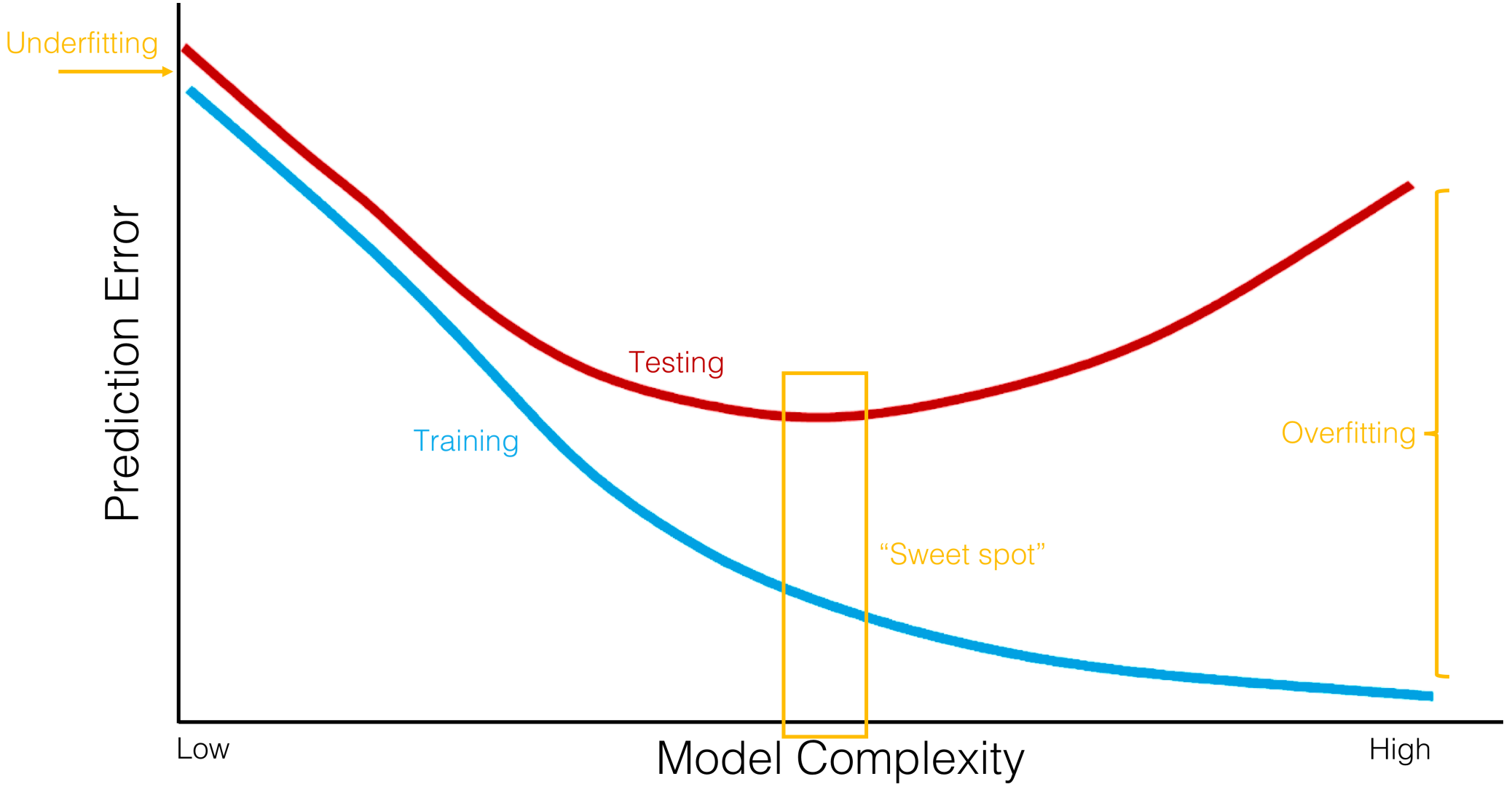
Règle I : D'après un sondage mené auprès de 400 Allemands, nous déduisons que 43,75 % de la population mondiale a les cheveux noirs, 37,5 % les cheveux bruns, 9 % les cheveux blonds, 0,25 % les cheveux roux et 9,5 % les cheveux gris.

Règle II : La couleur des cheveux est noire, brune, blonde, rousse ou grise.

Règle III : Env. 40 % des gens ont les cheveux noirs, 40 % les cheveux bruns, 5 % les cheveux blonds, 2 % les cheveux roux et 13 % les cheveux gris.

DISCUSSION

Laquelle des trois règles est la plus utile? La plus vague? Laquelle est trop spécifique?


Underfitting
Prediction Error
Testing
Training
Overfitting
"Sweet spot"
Low
Model Complexity
High

SURAPPRENTISSAGE

Il faut **TOUJOURS** évaluer les modèles sur des données pas encore examinées (d'essai).

SOLUTIONS POSSIBLES

On peut résoudre le surapprentissage de plusieurs façons :

- **Utilisation de nombreux ensembles d'apprentissage**
  Intersection autorisée (ou non : voir validation croisée)
- **Utilisation d'ensembles d'apprentissage plus grands**
  Répartition de 70 % - 30 % suggérée
- **Optimisation des données au lieu du modèle**
  La qualité des modèles est proportionnelle à celle des données

PROCÉDURES RECOMMANDÉES

Petits ensembles de données (moins de quelques centaines d'observations)

- Utiliser 100 à 200 répétitions d'une procédure d'**auto-amorçage**

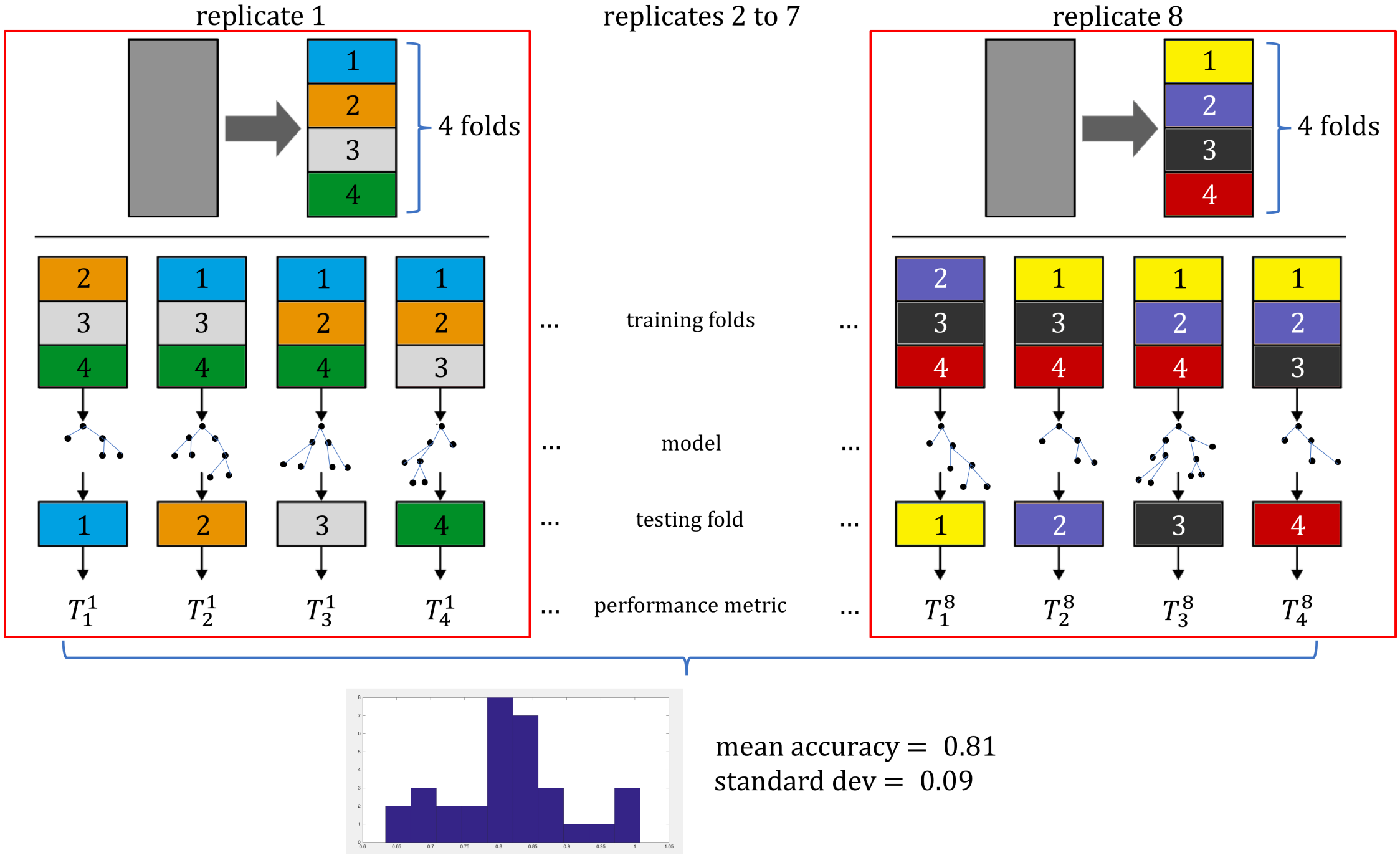
Ensembles de données de **taille moyenne** (moins de quelques milliers d'observations)

- Utiliser quelques répétitions d'une **validation croisée** découpée en 10 de l'ensemble d'apprentissage (voir la diapositive suivante)

Grands ensembles de données

- Utiliser quelques répétitions d'une répartition de **test** (70 %-30 %)

Remarque : Les limites de décision dépendent de la puissance de calcul et du nombre de tâches/flux de production.


replicate 1
replicates 2 to 7
replicate 8
4 folds
4 folds
training folds
model
testing fold
performance metric
T_1^1 T_2^1 T_3^1 T_4^1
T_1^8 T_2^8 T_3^8 T_4^8
mean accuracy = 0.81
standard dev = 0.09

MÉGADONNÉES

PROBLÈMES ET DÉFIS

« Les données, grandes ou petites, sont aussi utiles que les questions que vous leur posez. »
(Milo Jones et Philippe Silberzahn, Forbes Magazine [en anglais seulement])

UN MOT D'AVERTISSEMENT

Les mégadonnées ne sont pas une boule de cristal

- « Le rendement passé ne garantit pas les résultats futurs »

Les mégadonnées ne peuvent pas dicter des valeurs personnelles ou organisationnelles

- La bonne réponse sur le plan de la valeur peut être la mauvaise réponse sur le plan de la science des données
- Les conclusions basées sur les données n'existent pas en vase clos : le contexte compte
- L'obéissance aveugle à des résultats basés sur des données est aussi dangereuse qu'un rejet basé sur une réaction instinctive

Les mégadonnées ne peuvent pas résoudre tous les problèmes

- « Quand on n'a qu'un marteau, tout ressemble à un clou »

COMPARAISON ENTRE LES MÉGADONNÉES ET LES PETITES DONNÉES

Quelle est la différence principale?

- Les ensembles de données sont **VOLUMINEUX**
- Problèmes : collecte, capture, accès, stockage, analyse, visualisation

D'où viennent les données?

- Les progrès technologiques permettent de dépasser les limites de vitesse de traitement des données
- Détection de l'information, appareils mobiles, appareils photo et réseaux sans fil

Quels sont les défis?

- La plupart des techniques ont été élaborées pour de très petits ensembles de données
- La méthode directe laissera le meilleur analyste attendre les résultats pendant des années

PARADIGME DES 5 V

Volume : grandes quantités de données

Vélocité : vitesse à laquelle les données sont créées, consultées, traitées

Variété : différents types de données disponibles, ne peuvent pas tous être sauvegardés dans des bases de données relationnelles (tableaux, images,...)

Véracité : difficulté de contrôler la qualité et l'exactitude des mégadonnées

Valeur : transformation des données en quelque chose d'utile

Variabilité
Visualisation

PROBLÈME DES MÉGADONNÉES

De nombreux calculs sont effectués **instantanément**, d'autres prennent **beaucoup** de temps.

Le traitement de très grands ensembles de données en est un exemple parfait. L'analyse en R ou Python d'ensembles de données en croissance constante entraîne des décalages informatiques. Finalement, le temps nécessaire devient **« impossiblement » long**.

L'optimisation du code et l'utilisation d'un processeur plus rapide peuvent résoudre le problème dans une certaine mesure seulement.

C'est le **problème des mégadonnées**.

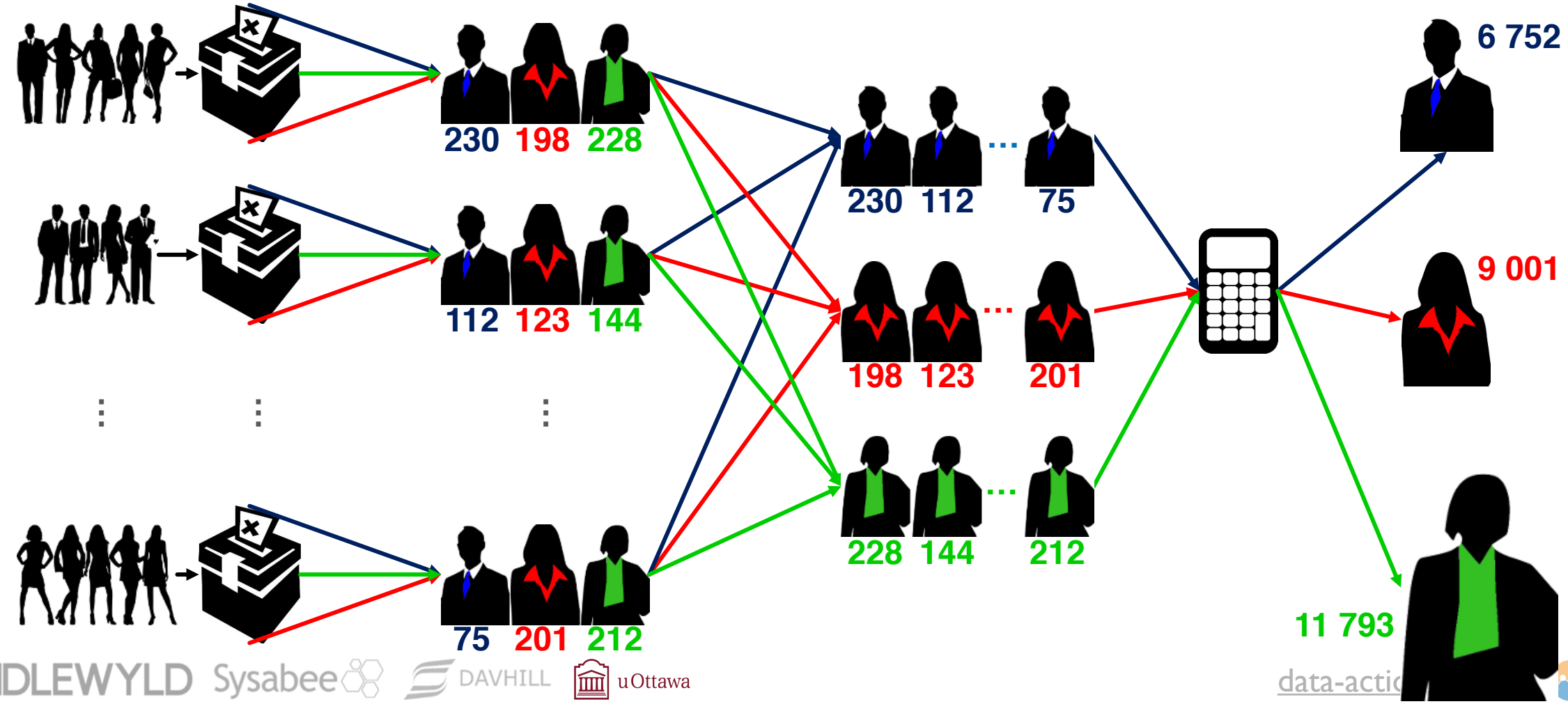
INFORMATIQUE RÉPARTIE

La **répartition** des calculs entre plusieurs cœurs de processeur/processeurs peut diviser le temps de calcul par un facteur de 4, 32 ou 1 000. Cela permet aux algorithmes de s'exécuter sur les mégadonnées et de mettre à jour les analyses, les services intelligents et les recommandations **chaque jour**, **toutes les heures** et **en temps réel**.

Analogie de l'**élection** pour décrire la parallélisation :

- Dépouillement du scrutin dans les différents bureaux de vote d'une circonscription
- Chaque bureau dépouille simultanément ses bulletins et rapporte son total
- Les totaux de tous les bureaux de vote sont agrégés au siège des élections
- Si une seule personne comptait tous les bulletins de vote, on finirait par obtenir le même résultat, mais cela prendrait *trop de temps*

ANALOGIE : ÉLECTIONS



ANALOGIE : PIZZERIA

Les avantages du parallélisme dépendent de la possibilité d'adapter les algorithmes de série pour utiliser un matériel parallèle.

Analogie de la **pizzeria** pour les limitations de la parallélisation/goulot d'étranglement :

- Plusieurs cuisiniers peuvent préparer les garnitures en parallèle
- Mais la cuisson de la croûte ne peut pas être parallélisée
- Le doublement du volume du four augmentera le nombre de pizzas que l'on peut préparer simultanément, mais n'accélère pas substantiellement la durée de préparation d'une pizza en particulier
- Parfois, les goulots d'étranglement préviennent tout gain de parallélisme : les gens font la queue des deux côtés d'une table pour aller chercher de la soupe, mais il n'y a qu'une louche

NOUVELLE NOUVELLE

La plupart des tâches de calcul pratiques peuvent être parallélisées et le sont. Les spécialistes en science des données modernes utilisent des cadres dans lesquels l'informatique répartie est déjà implémentée (par exemple, Apache Spark implémente MapReduce).

PERTINENCE ET PORTABILITÉ

PROBLÈMES ET DÉFIS

« Il peut être tentant d'utiliser les données comme une béquille dans la prise de décision : " Ce sont les données qui le disent! " Mais **parfois, les données nous déçoivent**, et la corrélation intéressante que vous avez trouvée n'est qu'un sous-produit d'un échantillon désordonné et biaisé. [...] Les sceptiques avisés peuvent vous aider à prendre du recul, à réfléchir et à vous demander si **ce que disent les données correspond réellement** à ce que vous savez et attendez du monde. »

PERTINENCE ET PORTABILITÉ

Les modèles de science des données seront largement utilisés dans les années à venir (cela a déjà commencé).

Nous avons discuté des avantages et des inconvénients de certaines applications pour des raisons éthiques et non techniques, mais il existe également des **défis techniques**.

Les méthodes de la science des données ne sont **pas** appropriées si :

- vous devez absolument utiliser des ensembles de données existants (**hérités**) au lieu d'un ensemble de données idéal (« ce sont les meilleures données dont nous disposons! »)

PERTINENCE ET PORTABILITÉ

Les méthodes de la science des données ne sont **pas** appropriées si : (suite)

- l'ensemble de données a des attributs qui prédisent utilement une valeur d'intérêt, mais qui ne sont pas disponibles lorsqu'une prédiction est requise

  Exemple : Le temps total passé sur un site Web peut prédire les futurs achats d'un visiteur, mais cette prévision doit être faite avant que le temps total passé sur le site Web soit connu...

- vous voulez prédire l'appartenance à une classe en utilisant un algorithme d'apprentissage non supervisé

  Exemple : Le regroupement des données sur les prêts en défaut peut conduire à une grappe contenant de nombreux emprunteurs en défaut. Si de nouvelles instances sont ajoutées à cette grappe, faut-il les considérer comme des emprunteurs en défaut?

HYPOTHÈSES NON TRANSFÉRABLES

Chaque modèle émet certaines hypothèses sur ce qui est ou non **pertinent** pour son fonctionnement, mais on a tendance à ne recueillir que des données **censées** être pertinentes pour une situation donnée.

Si les données sont utilisées dans d'autres contextes ou pour effectuer des prédictions en fonction d'attributs sans données, la validation des résultats est impossible.

- **Exemple :** Pouvons-nous utiliser un modèle qui prédit les emprunteurs hypothécaires en défaut pour prévoir également les détenteurs d'un prêt auto en défaut?

DISCUSSION

N’y a-t-il vraiment aucun lien entre les défauts de paiement hypothécaire et les défauts de paiement de prêt auto?

BIAIS, SOPHISMES ET INTERPRÉTATION

PROBLÈMES ET DÉFIS

« Si l'écart entre deux résultats de sondage est inférieur à la marge d'erreur, il n'y a rien à reporter. Les faits scientifiques ne sont pas déterminés par des sondages d'opinion. Un sondage effectué auprès de vos téléspectateurs/internautes n'est pas un sondage scientifique.
Qu'arriverait-il si tous les sondages incluaient l'option « Je n'en ai rien à faire »?

(Jorge Chan, Piled Higher and Deeper [en anglais seulement])

BIAIS, SOPHISMES ET INTERPRÉTATION

Lorsque vous consultez (ou menez) des études, vous devriez essayer de déterminer comment les biais suivants auraient pu entrer en jeu :

- **Biais de sélection** (quelles données a-t-on incluses, comment les a-t-on sélectionnées?)
- **Biais d'omission de variable** (a-t-on ignoré des variables pertinentes?)
- **Biais de détection** (des connaissances antérieures ont-elles influé sur les résultats?)
- **Biais de financement** (qui paie ceci?)
- **Biais de publication** (qu'est-ce qui n'est pas publié?)
- **Biais de surveillance des données** (fait-on trop d'efforts?)
- **Biais analytique** (le choix de la méthode donnée a-t-il influé sur les résultats?)
- **Biais d'exclusion** (exclue-t-on certaines observations/unités bien précises?)

BIAIS, SOPHISMES ET INTERPRÉTATION

La corrélation n'est pas un lien de causalité (mais c'est un indice!)

Les tendances extrêmes peuvent induire en erreur.

Il faut rester dans les limites d'une étude.

Gardez le taux de base à l'esprit.

Des résultats étranges se produisent parfois (paradoxe de Simpson).

BIAIS, SOPHISMES ET INTERPRÉTATION

Le hasard joue un rôle.

Toute activité analytique comporte une composante humaine.

De petits effets peuvent tout de même être (statistiquement) significatifs.

Méfiez-vous des statistiques sacro-saintes (valeur p, etc.).

DISCUSSION

La présence d'un biais invalide-t-elle nécessairement les résultats?

MYTHES ET ERREURS

PROBLÈMES ET DÉFIS

« Rien n'est toujours absolument ainsi. »

(Première loi de Sturgeon)

« Quatre-vingt-dix pour cent de toute chose est du déchet. »

(maxime de Sturgeon)

[A.K. Maheshwari, *Business Intelligence and Data Mining*]

MYTHES ET ERREURS DE LA SCIENCE DES DONNÉES

Mythe nº 1 – La science des données a trait aux algorithmes.

Mythe nº 2 – La science des données a trait à l'exactitude prédictive.

Mythe nº 3 – La science des données nécessite un entrepôt de données.

Mythe nº 4 – La science des données nécessite de grandes quantités de données.

Mythe nº 5 – La science des données nécessite des experts techniques.

MYTHES ET ERREURS DE LA SCIENCE DES DONNÉES

Erreur nº 1 – Sélectionner le mauvais problème.

Erreur nº 2 – Se retrouver enseveli sous des tonnes de données sans aucune compréhension des métadonnées.

Erreur nº 3 – Ne pas planifier le processus d'analyse des données.

Erreur nº 4 – Avoir des connaissances insuffisantes sur l'entreprise et le domaine.

Erreur nº 5 – Utiliser des outils d'analyse des données incompatibles.

MYTHES ET ERREURS DE LA SCIENCE DES DONNÉES

Erreur nº 6 – Utiliser des outils trop spécifiques.

Erreur nº 7 – Ignorer les prédictions/enregistrements individuels en faveur de résultats regroupés.

Erreur nº 8 – Manquer de temps.

Erreur nº 9 – Mesurer les résultats différemment du promoteur.

Erreur nº 10 – Croire naïvement ce qu'on nous dit au sujet des données.

DISCUSSION

La science des données consiste à poser les bonnes questions et à accepter des solutions imaginatives.

Dans la bataille entre la « science des données éprouvée et vérifiée » et la « science des données perturbatrice », dans quel camp vous rangez-vous?

EXERCICE – QUESTIONS DE TYPE VRAI OU FAUX

1. La performance prédictive d'un modèle supervisé est évaluée sur l'ensemble d'apprentissage.
2. On peut utiliser la validation croisée pour réduire le risque de surapprentissage d'un modèle prédictif.
3. Il est toujours préférable d'utiliser autant de variables que possible dans un modèle.
4. Si l'on supprime des observations dont les valeurs sont manquantes, cela peut entraîner des biais et des erreurs.
5. Nous pouvons utiliser un algorithme de groupement pour prédire l'appartenance à une classe.

EXERCICE – QUESTIONS DE TYPE VRAI OU FAUX

6. Si toutes les méthodes ne donnent pas le même résultat, cela prouve qu'il est impossible de répondre à la question.
7. Les connaissances sur l'entreprise et le domaine ne sont nécessaires que lorsque vous travaillez avec d'anciennes données.
8. Les promoteurs et les clients doivent être informés de tous les détails analytiques.
9. Il est impossible de planifier le processus d'analyse des données avant de savoir à quoi les données ressemblent.
10. Les données disponibles ne sont pas toujours appropriées ou ne représentent pas toujours la situation que nous modélisons.

AVENIR DE LA SCIENCE DES DONNÉES, DE L'INTELLIGENCE ARTIFICIELLE ET DE L'APPRENTISSAGE MACHINE

PROBLÈMES ET DÉFIS

CE DONT NOUS N'AVONS PAS PARLÉ

Des tonnes d'autres algorithmes de classification et de groupement

Systèmes de recommandation

Flux de données

Analyse bayésienne des données

Traitement du langage naturel (approfondi)

Extraction de traits et réduction de la dimension (fléau de la dimension)

Ingénierie des données

... et bien plus encore!

TÂCHES DE L'AVENIR

Véhicules sans chauffeur

Traduction automatique et compréhension des langues

Détection et prévention des perturbations du climat et des écosystèmes

Science des données automatisée (?!)

Détection et prévention des événements catastrophiques astronomiques

L'intelligence artificielle explicable

TENDANCES DE L'AVENIR

Nouvelles questions

Nouveaux outils

Nouvelles sources de données

Science des données comme composante de travail

Intelligence amplifiée/en essaim

CONCLUSION

PROBLÈMES ET DÉFIS

[Joe Blitzstein et Hanspeter Pfister, http://cs109.org (en anglais seulement)]

La science des données est une activité d'équipe à laquelle participent des experts en la matière.

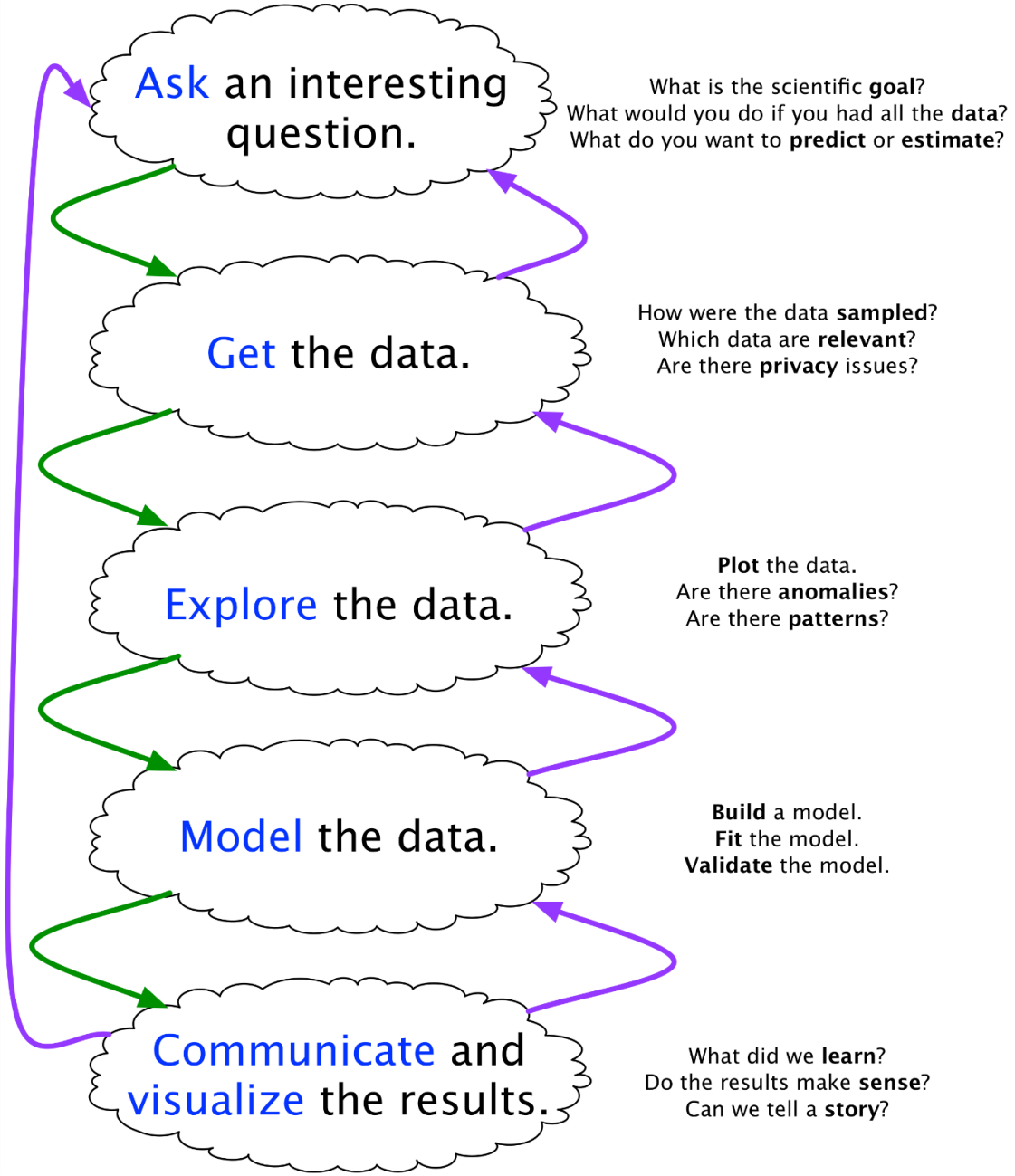
Les considérations d'ordre éthique sont primordiales et ne doivent pas nécessairement être en conflit avec la rentabilité.

Laissez parler les données.

Soyez à l'affût des idées exploitables!

Processus supervisé ou non supervisé

Une grande partie du temps d'analyse est consacrée à la préparation des données.



IDLEWYLD Sysabee DAVHILL uOttawa

RÉFÉRENCES

PROBLÈMES ET DÉFIS

RÉFÉRENCES

Aggarwal, C.C., éd. *Data Classification: Algorithms and Applications*, CRC Press, 2015.

Aggarwal, C.C., et C.K. Reddy, éd. *Data Custering: Algorithms and Applications*, CRC Press, 2014.

Torgo, L. *Data Mining with R: Learning with Case Studies*, 2e éd., CRC Press, 2017.

McCallum, Q.E. *Bad Data Handbook*, O'Reilly, 2013.

Maheshwari, A.K. *Business Intelligence and Data Mining*, Business Expert Press, 2015.

Provost, F., et T. Fawcett. *Data science pour l'entreprise*, Eyrolles, 2018.

Frank, E., et I.H. Witten. *Data Mining: Practical Machine Learning Tools and Techniques*, 2e éd., Elsevier, 2005.

RÉFÉRENCES

Sur Internet : https://hbr.org/2013/07/how-google-flu-trends-is-getting-to-the-bottom.

Sur Internet : *9 types of research bias and how to avoid them*.

Entrée Bias dans Wikipedia. Sur Internet : https://en.wikipedia.org/wiki/Bias.

Entrée Selection Bias dans Wikipedia. Sur Internet : https://en.wikipedia.org/wiki/Selection_bias.

Cochrane Methods. Sur Internet : « Assessing Risk of Bias in Included Studies ».

Quantshare. Sur Internet : « Data Snooping Bias ».

Entrée Bias (statistics) dans Wikipedia. Sur Internet : https://en.wikipedia.org/wiki/Bias_(statistics).

Entrée Benford's Law dans Wikipedia. Sur Internet : https://en.wikipedia.org/wiki/Benford%27s_law.

RÉFÉRENCES

Silver, N. *The Signal and the Noise: Why So Many Predictions Fail – But Some Don't*, Penguin Press, New York, 2012.

Lewis, M. *Moneyball: The Art of Winning an Unfair Game, Norton*, New York, 2003.

Uri Simonshon. Sur Internet : http://opim.wharton.upenn.edu/~uws/.

Sur Internet : https://en.wikipedia.org/wiki/Data_analysis_techniques_for_fraud_detection.

Flaherty, D. « The Vaccine-Autism Connection: A Public Health Crisis Caused by Unethical Medical Practices and Fraudulent Science », *Annals of Pharmacotherapy*, vol. 45, nº 10 (octobre 2011), p. 1302-1304.

Reinhart, A. Sur Internet : *Statistics Done Wrong*.

RÉFÉRENCES

Sur Internet : https://www.datacamp.com/community/blog/data-science-past-present-future.

Kargupta, H., J. Han, P.S. Yu, R. Motwani et V. Kumar, éd. *Next Generation of Data Mining*, CRC/Chapman & Hall, 2009.