Enjeux et défis

INTRODUCTION À L'APPRENTISSAGE AUTOMATIQUE

Nous disons tous que nous aimons les données, mais ce n'est pas le cas. Nous aimons obtenir des informations à partir des données. Ce n'est pas tout à fait la même chose que d'aimer les données elles-mêmes. En fait, j'ose dire que je n'aime pas les données, et il semble que je ne sois pas le seul. [Q.E. McCallum, *Bad Data Handbook*]

Les données, grandes ou petites, ne sont utiles que dans la mesure où les questions que vous leur posez le sont. [M. Jones, P. Silberzahn]

Rien n'est jamais absolument vrai. [Première loi de Sturgeon]

95% des choses sont des saletés. [Sturgeon's Maxim]

Il peut être tentant d'utiliser les données comme béquille dans la prise de décision : "Les données 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 prendre du recul, réfléchir et se demander si **ce que disent les données correspond bien** à ce que l'on sait et à ce que l'on attend du monde.

[Nicholas Diakopoulos, Harvard Business Review]

Session 4

Dear News Media,

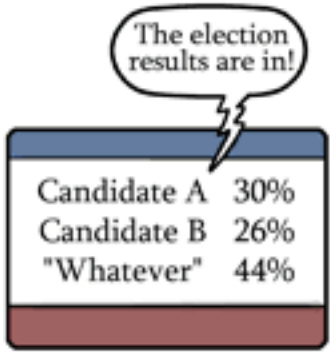
When reporting poll results, please keep in mind the following suggestions:

1. If two poll numbers differ by less than the margin of error, it's not a news story.

2. Scientific facts are not determined by public opinion polls.

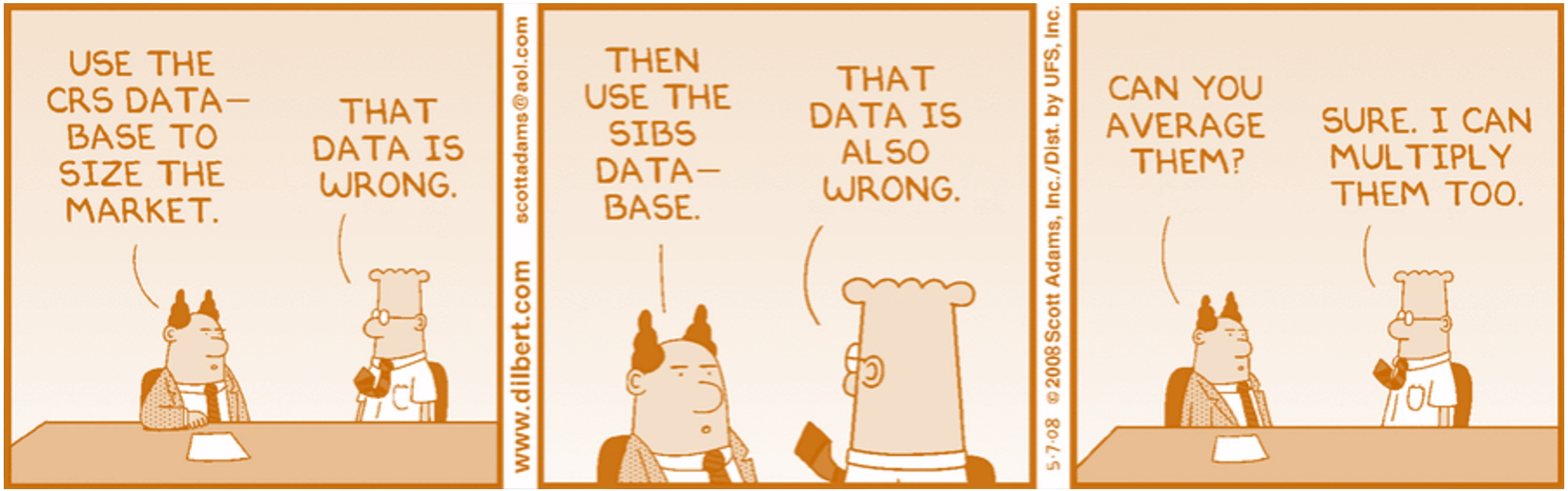
3. A poll taken of your viewers/internet users is not a scientific poll.

4. What if all polls included the option "Don't care"?



Signed,

-Someone who took a basic statistics course.



10. Données pourries et données massives

Mauvaises données

L'ensemble de données passe-t-il le **test de l'odeur** ?

- entrées non valides, observations anormales, etc.

Données formatées pour la consommation humaine et non pour la lecture par une machine

Difficultés de **traitement du texte**

- codage
- caractères spécifiques à l'application

Mauvaises données

Collecte de données **en ligne**

- légalité de l'obtention des données
- stockage des versions hors ligne

Détection des **mensonges** et des **erreurs**

- les erreurs de déclaration (mensonges ou erreurs)
- l'utilisation d'un langage polarisant

Données et réalité

- mauvaises données
- mauvaise réalité ?

Mauvaises données

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

- biais d'imputation
- remplacement des valeurs extrêmes par des valeurs moyennes
- déclaration par procuration (chef de ménage pour le ménage)

La recherche de la **perfection**

- données académiques
- données professionnelles
- données gouvernementales
- données de service

Mauvaises données

Les **pièges de** la science des données

- l'analyse sans la compréhension
- l'utilisation d'un seul outil (par choix ou par obligation)
- l'analyse pour le plaisir de l'analyse
- des attentes irréalistes à l'égard de la science des données
- sur la base du besoin de savoir ... et vous n'avez pas besoin de savoir

Bases de données, fichiers, et “cloud”

- le nuage/big data/apprentissage profond résoudra tous nos problèmes !

Mauvaises données

Assez près, assez bon (“close enough is good enough”) ?

- exhaustivité
- cohérence
- exactitude
- responsabilité

Données massives (mise en garde)

Pas une boule de cristal

- "Les performances passées ne garantissent pas les résultats futurs

Ne peut pas dicter les valeurs personnelles ou organisationnelles

- La bonne réponse (valeurs) peut être la mauvaise réponse (science des données)
- Les conclusions ne vivent pas dans le vide : le contexte est important
- L'obéissance aveugle à des résultats fondés sur des données est tout aussi dangereuse que le rejet fondé sur des réactions instinctives

Ne peut résoudre tous les problèmes

- "Quand on n'a qu'un marteau, tout ressemble à un clou."

Données massives vs. régulières

Quelle est la principale différence ?

- les ensembles de données sont **GRANDS**
- questions : collecte, capture, accès, stockage, analyse, visualisation

D'où proviennent les données ?

- les progrès technologiques permettent de dépasser les limites de la vitesse de traitement
- la détection d'informations, les appareils mobiles, les caméras et les réseaux sans fil

Quels sont les défis à relever ?

- la plupart des techniques ont été conçues pour de très petits ensembles de données.
- l'approche prend du temps... même pour les meilleurs analystes !

6. **variabilité**
7. **visualisation**

Le paradigme du 5V (7V ?)

1. **volume :** grandes quantités de données
2. **vélocité :** vitesse à laquelle les données sont créées, consultées, traitées
3. **variété :** différents types de données disponibles, qui ne peuvent pas toutes être sauvegardées dans des bases de données relationnelles (tableaux, images,...)
4. **véracité :** la qualité et l'exactitude des données masives sont plus difficiles à contrôler
5. **valeur :** transformer les données en quelque chose d'utile

Le problème des données massives

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

L'analyse de très grands ensembles de données en est un parfait exemple. L'analyse en *R* ou en *Python* avec des ensembles de données en augmentation constante entraîne des décalages informatiques. À terme, le temps nécessaire devient **impraticable**.

L'optimisation du code et l'utilisation d'un processeur plus rapide ne peuvent que soulager la situation.

C'est le **problème des données massives**.

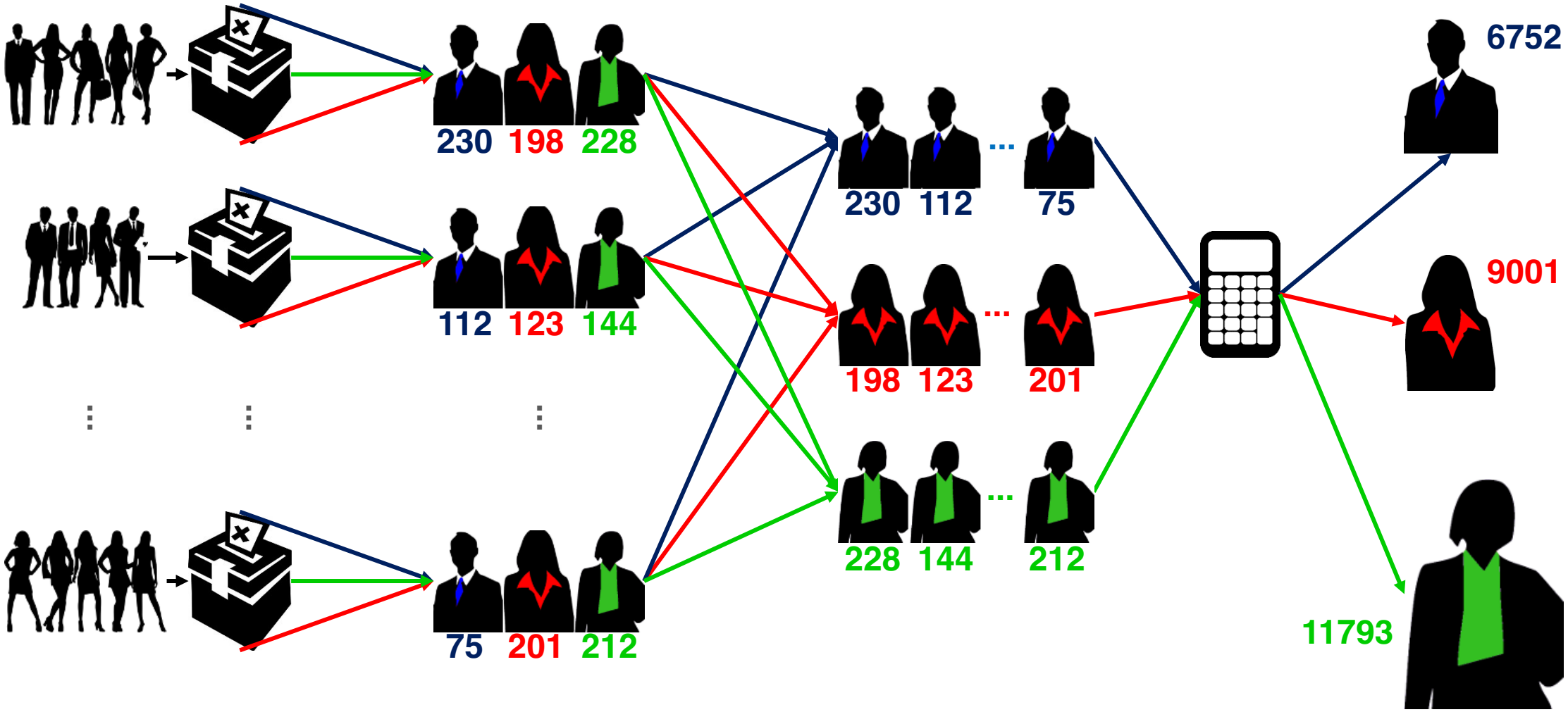
Informatique distribuée

La **répartition** des calculs entre plusieurs cœurs d'unité centrale/CPU peut diviser le temps de calcul par un facteur de 4, ou 32, ou 1000, etc. Cela permet aux algorithmes de s'exécuter sur les big data pour maintenir les analyses, les services intelligents et les recommandations mis à jour **quotidiennement**, **toutes les heures**, en temps **réel**.

Analogie **électorale** avec la parallélisation :

- dépouillement des votes dans les différents bureaux de scrutin d'une circonscription
- chaque bureau de scrutin compte simultanément ses propres votes et rapporte leur total
- les totaux de tous les bureaux de scrutin sont agrégés au siège des élections
- une seule personne comptant tous les bulletins de vote finirait par obtenir le même résultat, mais cela prendrait *trop de temps.*

Analogie : élection



Analogie : pizzéria

Les gains offerts par le **parallélisme** dépendent de la possibilité d'**adapter** les algorithmes en série à l'utilisation de **matériel parallèle (hardware).**

Analogie de la **pizzeria** pour les limites de la parallélisation :

- plusieurs cuisiniers peuvent préparer les garnitures en parallèle
- mais la cuisson de la croûte ne peut pas être parallélisée
- doubler la surface du four augmente le nombre de pizzas pouvant être préparées simultanément, mais n'accélére pas de manière substantielle la préparation d'une seule pizza
- parfois, des goulets d'étranglement empêchent tout gain de parallélisme : les gens font la queue des deux côtés d'une table pour obtenir de la soupe, mais il n'y a qu'une seule louche

Bonnes nouvelles

La plupart des tâches informatiques pratiques peuvent être parallélisées.

Les scientifiques des données modernes utilisent des cadres où l'informatique distribuée est déjà implémentée (Apache Spark et *MapReduce*, par exemple).

Prenez le temps de réfléchir à ce problème potentiel **avant de commencer** le processus de collecte et d'analyse des données – cela vous évitera des maux de tête à long terme !

Lectures conseillées

Données pourries et données massives

J. Leskovec, A. Rajamaran, J. D. Ullman, *Mining of Massive Datasets*. Cambridge Press, 2014.

Data Understanding, Data Analysis, Data Science
Volume 3: Spotlight on Machine Learning

19. Introduction to Machine Learning
    19.6 Issues and Challenges
    - Bad Data

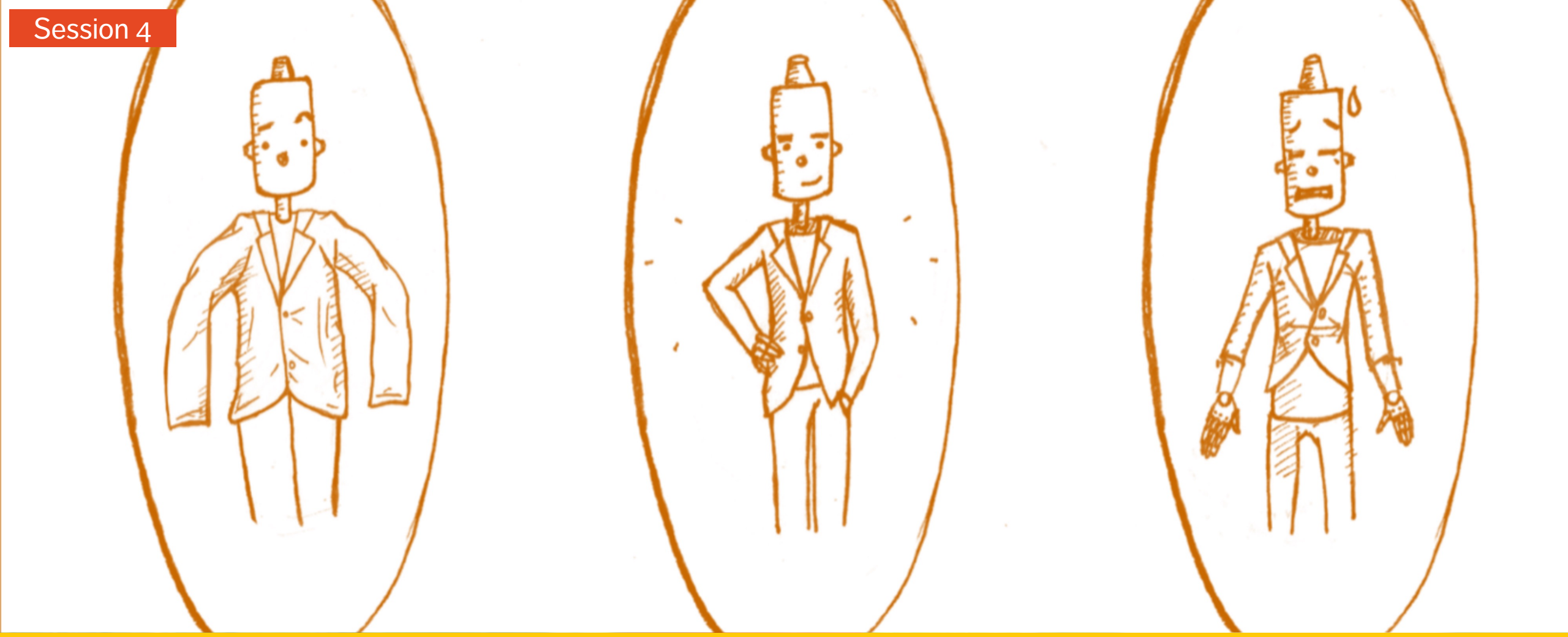
Exercices

Données pourries et données massives

1. Comme le dit le proverbe, "on ne peut rien faire sans rien faire". Quelles sont les conséquences analytiques, commerciales et de politique publique d'une prise de décision fondée sur de mauvaises données ?

2. Le fait qu'un ensemble de données soit considéré comme petit ou "grand" dépend non seulement de l'ensemble de données, mais aussi des outils disponibles.

   Générez des ensembles de données aléatoires de plus en plus grands (3 variables + 1 classe) pour les regrouper avec `kmeans()` et les classifier avec `rpart()`. Gardez en mémoire le temps d'exécution. Comment la durée d'exécution varie-t-elle en fonction du nombre d'observations ? À partir de quelle taille prévoyez-vous que les algorithmes seront trop lents et encombrants pour vos besoins ?



11. Ajustement et transférabilité

Principes de base

Pour être utiles, les règles ou les modèles générés par une technique sur un **ensemble de formation** doivent pouvoir être généralisés à de **nouvelles données** (ou à des **ensembles de validation/test**).

Des problèmes surviennent lorsque les connaissances acquises par l'**apprentissage supervisé** ne se généralisent pas correctement aux données.

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

Paradoxalement, cela peut se produire si les règles ou les modèles s'adaptent **trop bien** à l'ensemble d'apprentissage – les résultats sont **trop spécifiques à l'ensemble de formation**.

Exemple de règles

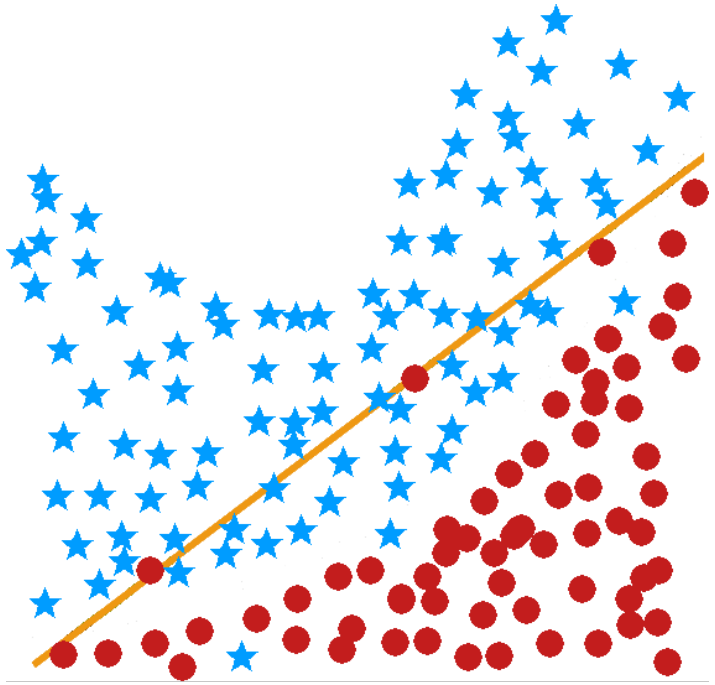
Règle I : sur la base d'une enquête menée auprès de 400 Allemands, nous déduisons que 43.75 % de la population mondiale a les cheveux noirs, 37.5 % les cheveux bruns, 9 % les cheveux blonds, 0.25 % les cheveux roux et 9.5 % les cheveux gris.

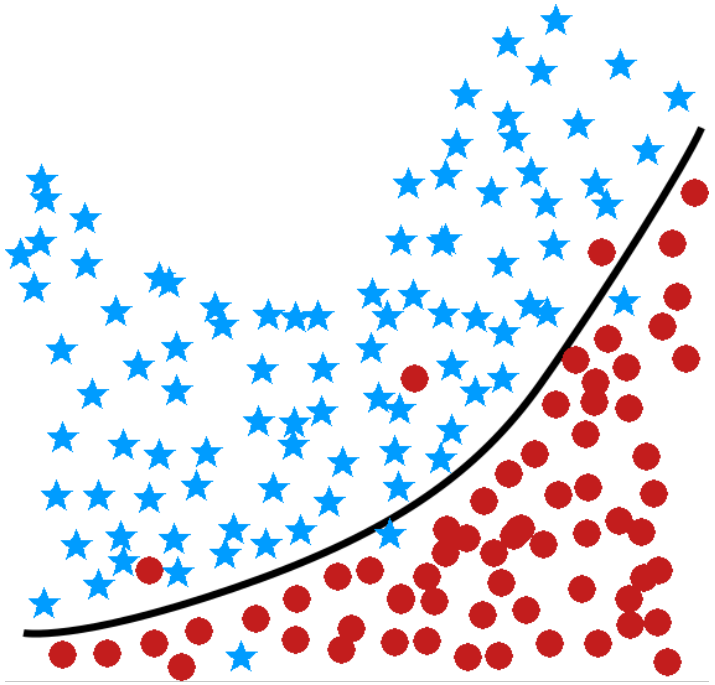
Règle II : les cheveux des humains sont soit noirs, soit bruns, soit blonds, soit roux, soit gris.

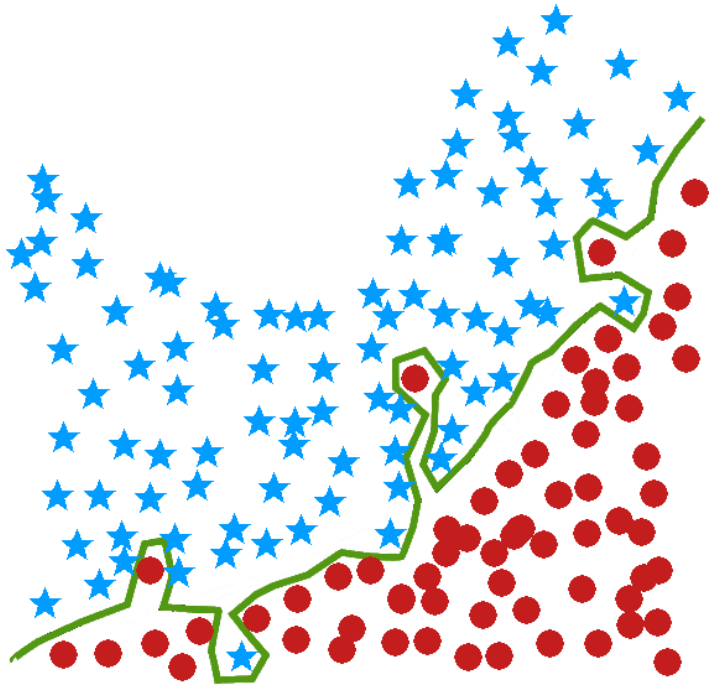
Règle III : environ 40 % des humains ont des cheveux noirs, 40 % des cheveux bruns, 5 % des cheveux blonds, 2 % des cheveux roux et 13 % des cheveux gris.

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

Boucle d'or et les trois modèles

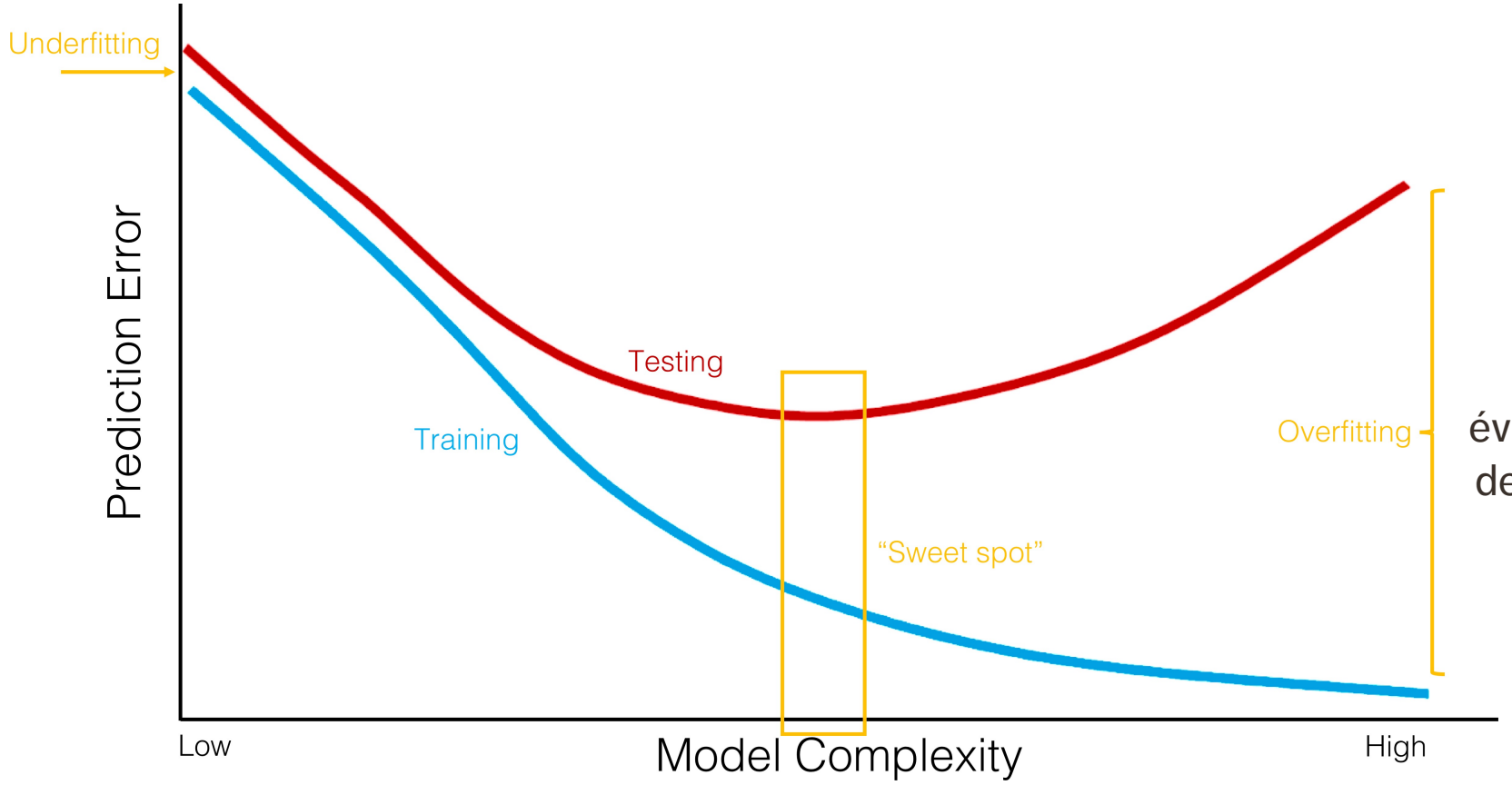






Mais attention à la **double descente** !

Compromis biais-variance



Il faut **TOUJOURS** évaluer les modèles sur des données nouvelles (données de test) !

Compromis biais-variance

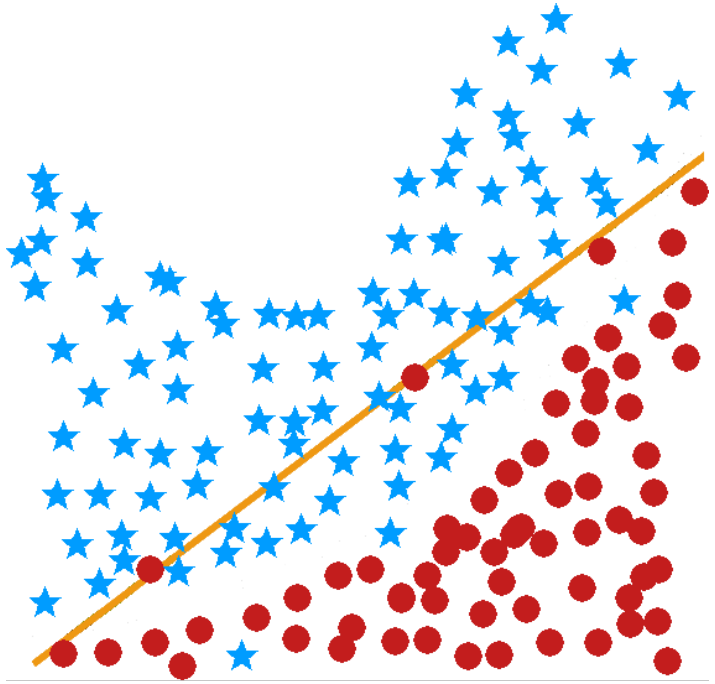
Nous **construisons** un modèle sur des **données historiques** et **évaluons** sa performance sur de **nouvelles données**.

Soit $\mathrm{Err}_{\mathrm{Te}}$ la performance du modèle sur les données d'essai :

$$\mathrm{Err}_{\mathrm{Te}} = \mathrm{Biais}^2_{\mathrm{Modèle}} + \mathrm{Variance}_{\mathrm{Modèle}}$$

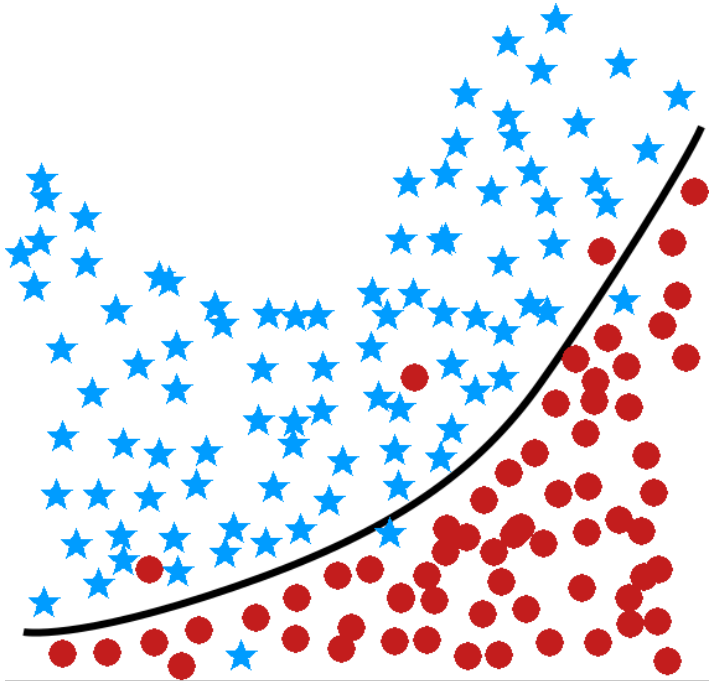
Le **biais** mesure la **précision** de la prédiction du modèle ; la **variance**, sa **sensibilité** aux (petits) changements dans les données.

Boucle d'or et les trois modèles

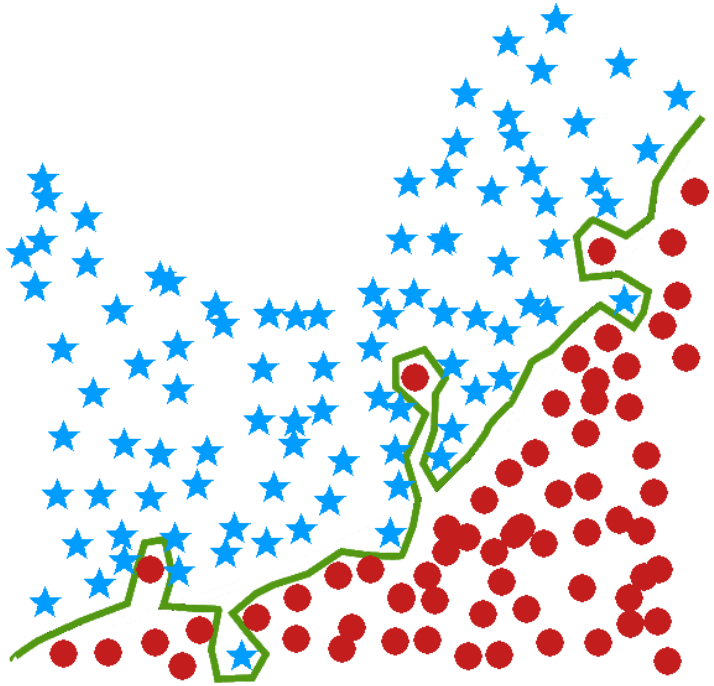


sous-ajustement
biais = élevé
variance = faible
erreur = **élevée**

les prévisions ne sont pas très précises



juste ce qu'il faut
biais = moyen
variance = moyenne
erreur = **moyenne**



sur-ajustement
biais = faible
variance = élevée
erreur = **élevée**

le modèle est trop spécifique aux données

Solutions possibles

Le sous-ajustement peut être surmonté en considérant des modèles plus complexes.

L'adaptation excessive peut être surmontée de plusieurs façons :

- **utilisation de plusieurs ensembles de formation**
  le chevauchement est autorisé (ou non : voir validation croisée)
- **utiliser des ensembles de formation plus grands**
  on suggère une répartition 70% - 30%
- **optimiser les données plutôt que le modèle**
  les modèles ne valent que ce que valent les données

Procédures recommandées

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

- utiliser 100 à 200 répétitions d'une procédure **bootstrap**

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

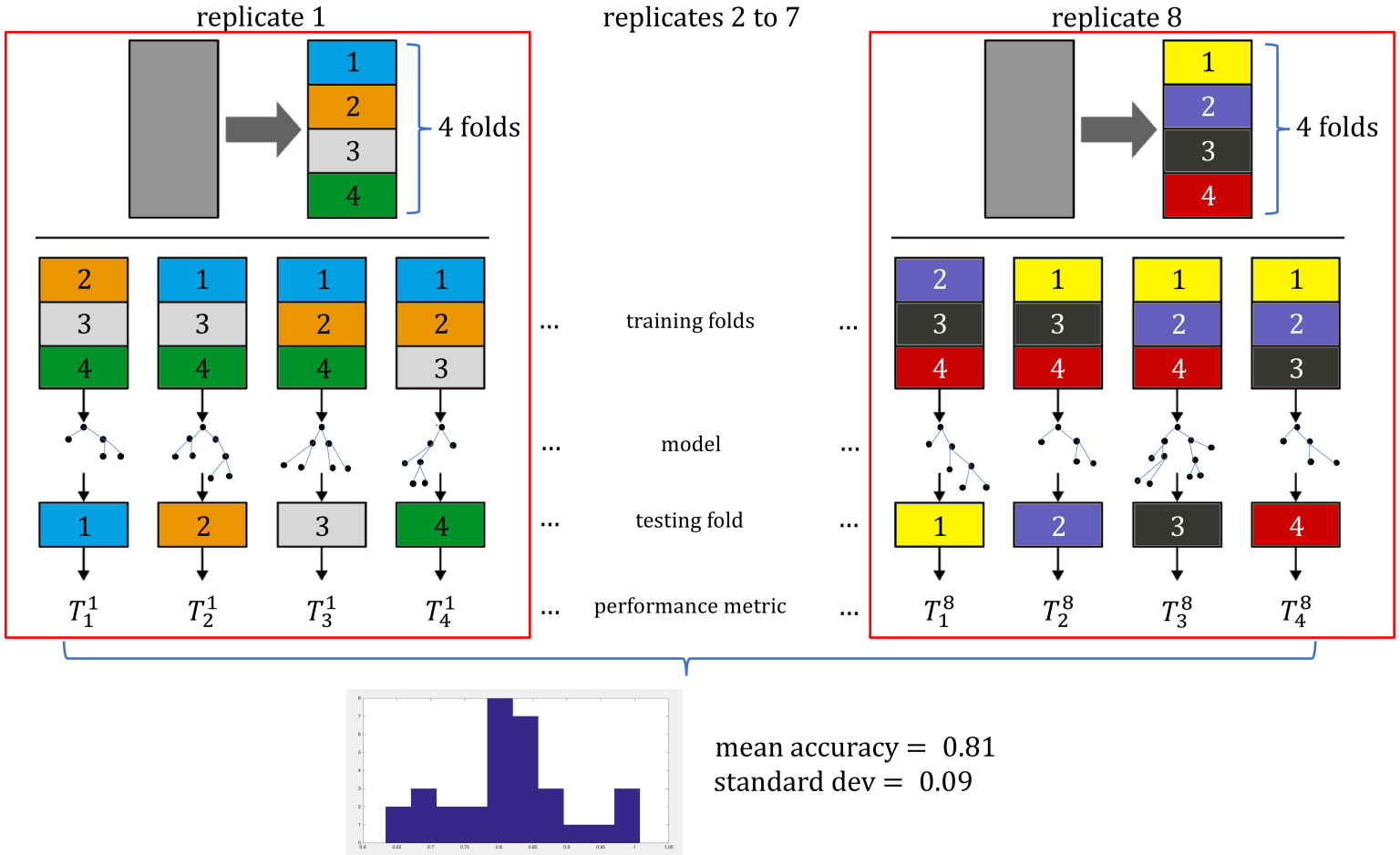
- utiliser quelques répétitions de la **validation croisée** à 10 plis sur l'ensemble de formation (voir diapositive suivante)

Grands ensembles de données

- faire quelques répétitions de avec 70%-30%

Les **frontières de décision** dépendent de la puissance de calcul et du nombre de tâches dans les flux de travail.

Validation croisée



mean accuracy = 0.81
standard dev = 0.09

Pertinence et transférabilité

La science des données et les modèles d'apprentissage automatique continueront d'être fortement utilisés dans les années à venir.

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

Les méthodes DS/ML **ne sont pas appropriées** :

- Si vous devez absolument utiliser un ensemble de données existant (**“legacy”**) au lieu d'un ensemble de données **idéal** ("ce sont les meilleures données dont nous disposons !")

Pertinence et transférabilité

Les méthodes DS/ML **ne** sont **pas appropriées** (suite) :

- si l'ensemble de données possède des attributs qui permettent de prédire une valeur intéressante, mais qui **ne seront jamais disponibles** lorsqu'une prédiction est nécessaire

  Exemple : le temps total passé sur un site web peut permettre de prédire les achats futurs d'un visiteur, mais la prédiction doit être faite avant que le temps total passé sur le site web ne soit connu.

- si vous tentez de prédire l'**appartenance à une classe** à l'aide de regroupement

  Exemple : le regroupement de données sur les défaillances de prêts peut aboutir à un grappe contenant plusieurs défaillants. Si de nouvelles instances sont ajoutées à cette grappe, doivent-elles être automatiquement considérées comme défaillants ? (non)

Hypothèses non transférables

Chaque modèle émet certaines hypothèses sur ce qui est/ n'est pas **pertinent** pour son fonctionnement, mais il y a une tendance à ne collecter que des données **supposées** pertinentes pour une situation donnée.

Si les données sont utilisées dans d'autres contextes, ou pour faire des prédictions en fonction d'attributs sans données, comment valider les résultats ?

- **Exemple :** peut-on utiliser un modèle qui prédit les défauts de paiement des prêts hypothécaires pour prédire également les défauts de paiement des prêts automobiles ? Une voiture n'est pas une maison : elles jouent des rôles différents dans la vie d'un individu ; les valeurs sont d'ampleurs différentes ; etc.
- Cela dit, n'y a-t-il vraiment aucun lien entre les deux ?

Lectures conseillées

Sous-ajustement/sur-ajustement et transférabilité

Data Understanding, Data Analysis, Data Science
Volume 3: Spotlight on Machine Learning

19. Introduction to Machine Learning
    19.6 Issues and Challenges
    - Overfitting/Underfitting
    - Transferability

20. Regression and Value Estimation
    20.1 Statistical Learning
    - Model Evaluation
    - Bias-Variance Trade-Off

    20.3 Resampling Methods
    - Cross-Validation

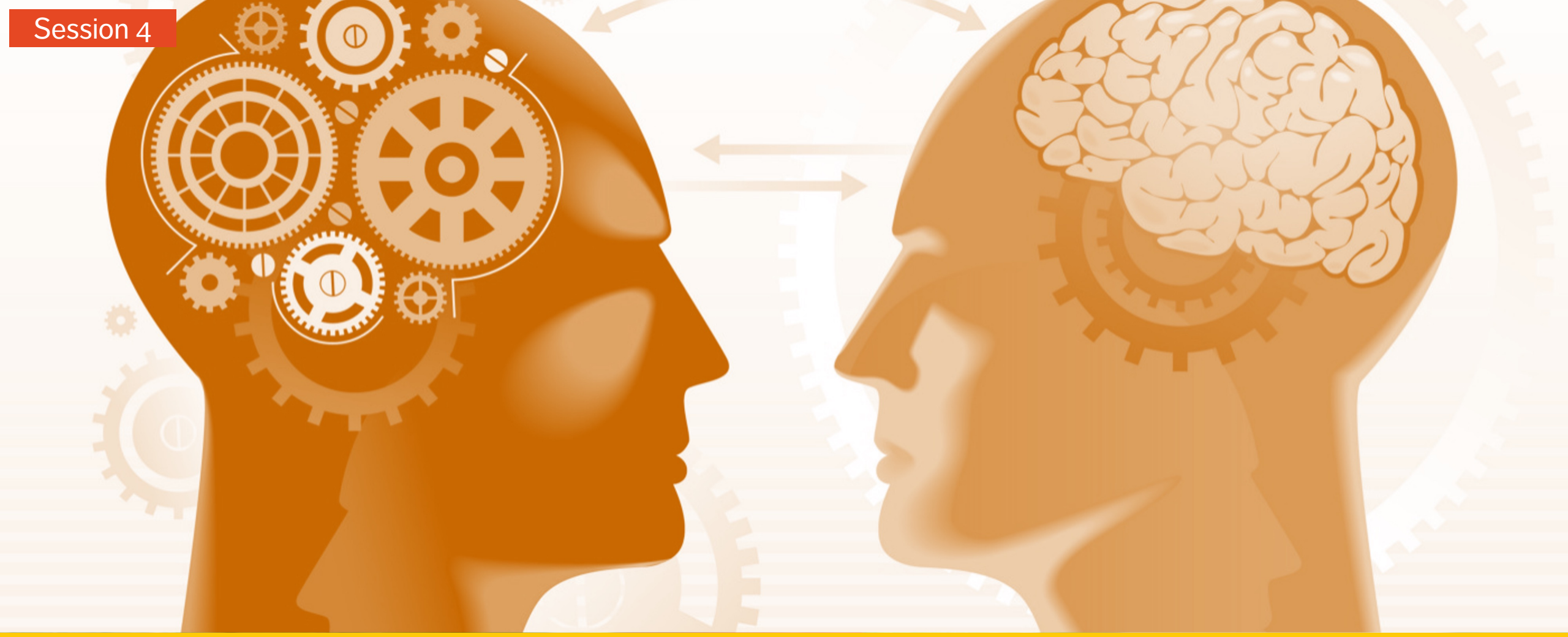
    20.4 Model Selection
    - Optimal Models

Exercices

Sous-ajustement/sur-ajustement et transférabilité

1. Cet exercice illustre le sur- et le sous-ajustement.
   a. Générez aléatoirement $n = 150$ valeurs de x dans $[1,10]$.
   b. Générez aléatoirement $n = 150$ valeurs de y selon selon $y = 10 + x - 2x^2 + 17x^3 + \varepsilon$, où ε est un terme d'erreur aléatoire de votre choix.
   c. Ajustez aux données un modèle linéaire, un modèle quadratique, un modèle cubique, et un modèle polynomial de degré 10.
   d. Ajoutez 3 observations aux données comme dans les étapes a. et b. Répétez l'étape c. Les modèles changent-ils beaucoup ?
   e. À quel(s) modèle(s) feriez-vous confiance pour faire des prédictions sur de nouvelles données ?

2. Modifiez l'exemple Gapminder de DUDADS (Cross-Validation) afin de sélectionner un modèle dans la question précédente.



12. Faits divers

Biais, erreurs, et interprétation

Lors de la consultation (ou de la réalisation) d'études, il faut se méfier de **biais** :

- **sélection** (quelles données ont été incluses, comment ont-elles été sélectionnées ?)
- **variables omises** (des variables pertinentes ont-elles été ignorées ?)
- **détection** (les connaissances préalables ont-elles influencé les résultats ?)
- **financement** (qui paie pour tout cela ?)
- **publication** (qu'est-ce qui n'est pas publié ?)
- **“snooping” des données** (trop d'efforts ?)
- **analytique** (le choix d'une méthode spécifique a-t-il influencé les résultats ?)
- **exclusion** (certaines observations/unités sont-elles exclues ?)

Mais la présence de biais invalide-t-elle **nécessairement** les résultats ?

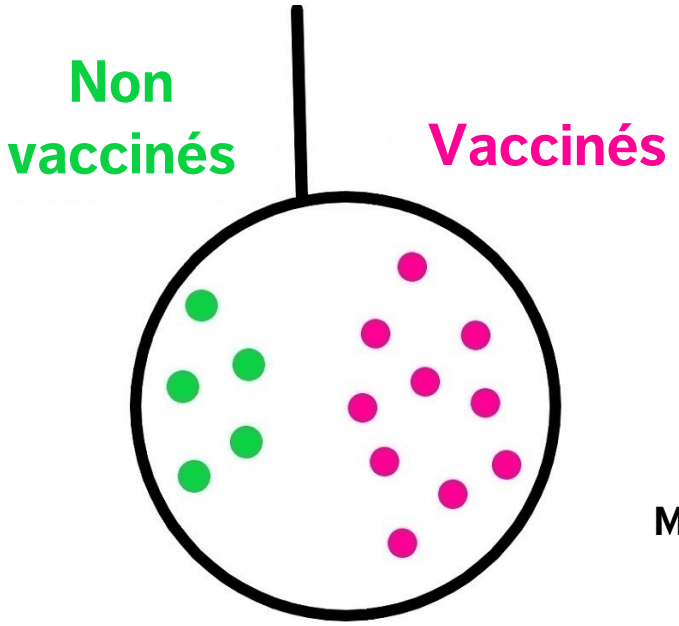
Biais, erreurs, et interprétation

Rappelez-vous :

- corrélation n'est pas causalité (mais c'est un indice !)
- les tendances extrêmes peuvent induire en erreur
- restez dans le domaine d'analyse d'une étude
- gardez le taux de base à l'esprit
- les résultats contre-intuitifs ne sont pas toujours faux (paradoxe de Simpson, loi de Benford, etc.)
- le hasard joue un rôle
- toute activité analytique comporte une composante humaine
- de petits effets peuvent encore être (statistiquement) significatifs
- il faut se méfier des statistiques sacro-saintes (valeur p, etc.)

Patients hospitalisés avec la COVID

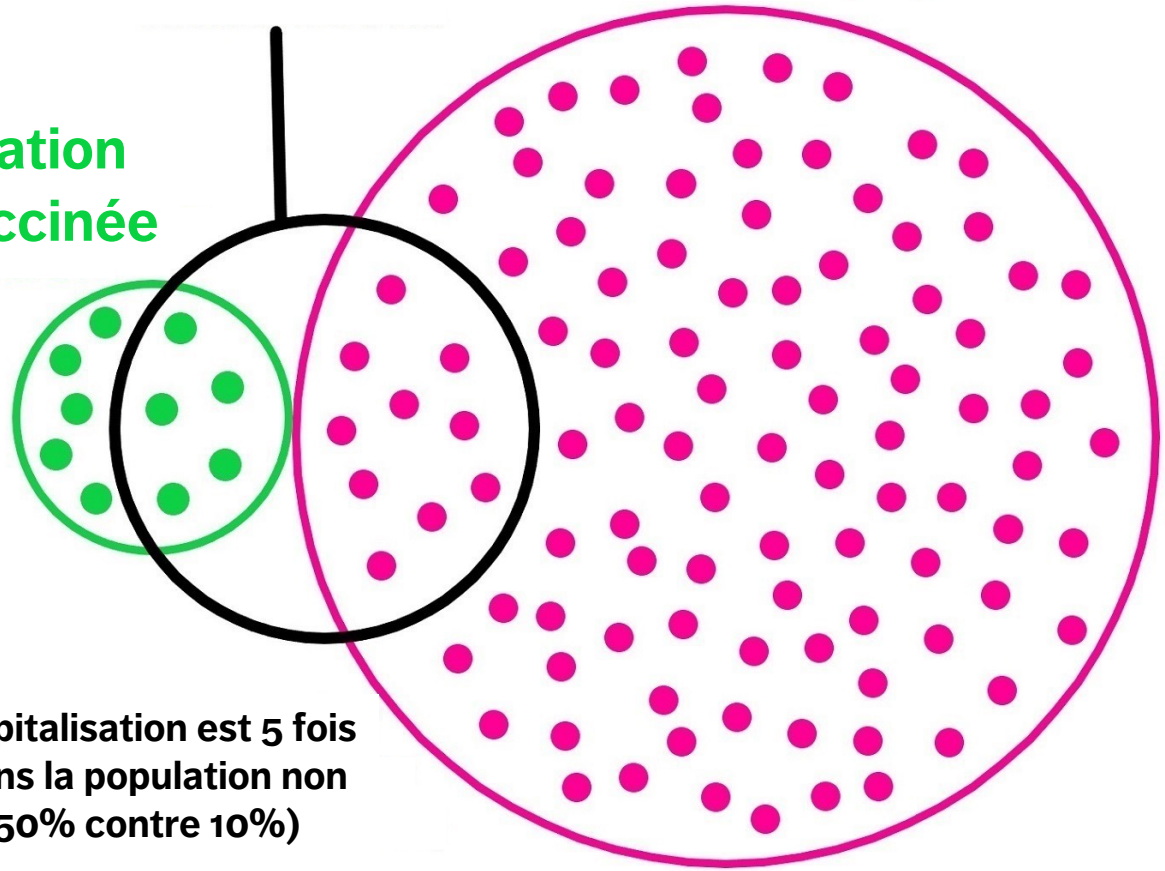
Source: Twitter.com/MarcRummy



More vaccinated than unvaccinated people in the hospital

Mais constatez le taux de chaque groupe dans la population totale

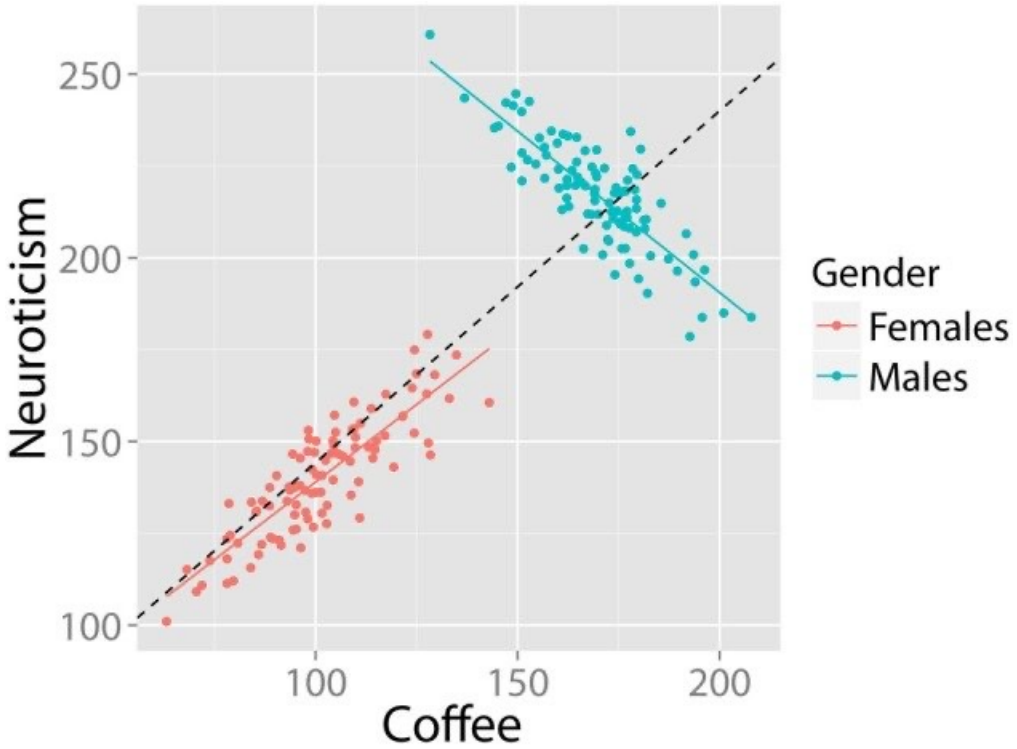
Patients hospitalisés avec la COVID

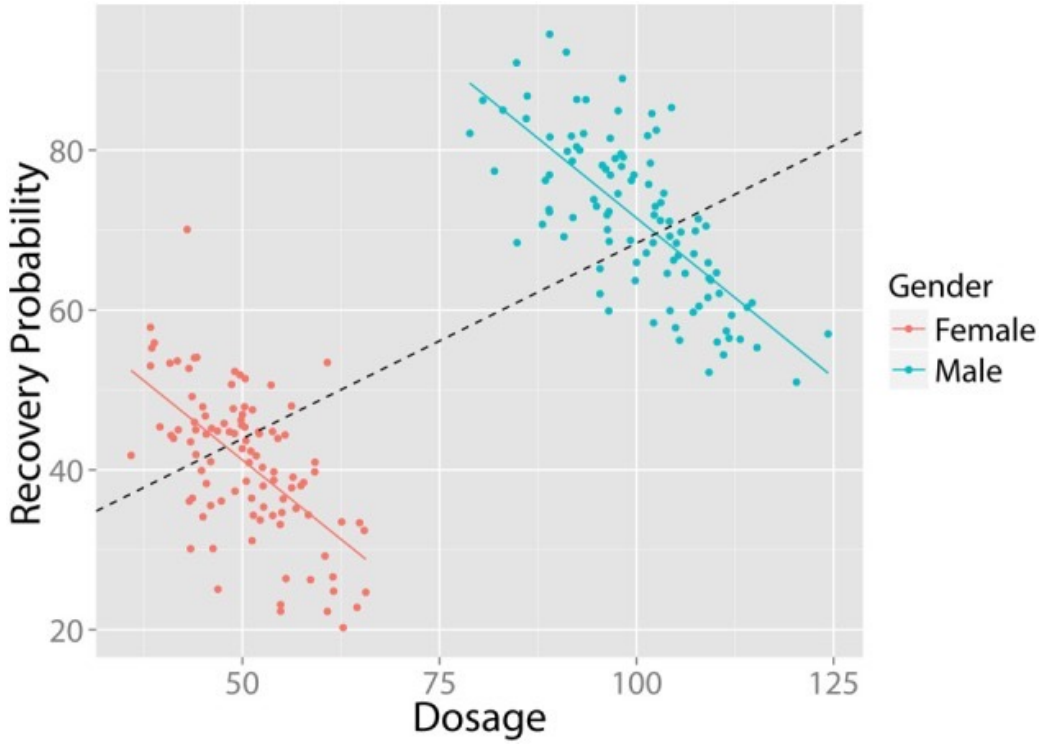


Le taux d'hospitalisation est 5 fois plus élevé dans la population non vaccinée (50% contre 10%)

Note : les taux présentés servent à illustrer le concept de l'erreur du taux de base lorsque le taux de vaccination est élevé

[Kievit et al., *Simpson's Paradox in Psychological Science: A Practical Guide*]





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

Mythes et erreurs sur les DS/ML

Les mythes :

- DS/ML concerne les algorithmes
- DS/ML est une question de précision prédictive
- DS/ML nécessite des entrepôts de données et une infrastructure fantaisiste
- DS/ML nécessite une grande quantité de données
- DS/ML nécessite des experts techniques (?)

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

Mythes et erreurs sur les DS/ML

Les erreurs :

- choisir le mauvais problème
- être enseveli sous des tonnes de données sans compréhension des métadonnées
- ne pas planifier le processus d'analyse des données
- avoir une connaissance insuffisante de l'entreprise et du domaine
- utiliser des outils d'analyse de données incompatibles
- utiliser des outils trop spécifiques
- ignorer les prédictions/enregistrements individuels au profit de résultats agrégés
- manquer de temps
- mesurer les résultats différemment du sponsor/des parties prenantes
- croire naïvement ce que l'on nous dit sur les données

L'avenir de la DS/ML/AI

Ce dont nous n'avons pas parlé :

- des tonnes d'algorithmes de classification et de regroupement
- les systèmes de recommandation
- le flux de données
- l'analyse bayésienne des données
- le traitement du langage naturel et l'exploration de textes
- la sélection des caractéristiques et réduction des dimensions (fléau de la dimensionnalité)
- l'ingénierie des données
- ... et bien d'autres choses encore !

L'avenir de la DS/ML/AI

Tâches futures :

- véhicules sans conducteurs
- la traduction automatique et la compréhension du language (GPT?)
- la détection et la prévention des perturbations du climat et des écosystèmes
- la science des données automatisée (?!)
- la détection et la prévention des événements astronomiques catastrophiques
- l'intelligence artificielle explicable

L'avenir de la DS/ML/AI

Tendances futures :

- nouvelles questions
- nouveaux outils
- nouvelles sources de données
- la science des données comme composante de l'emploi
- intelligence augmentée/enchevêtrée

En conclusion

La DS/ML est une activité d'équipe.

Les considérations éthiques sont cruciales.

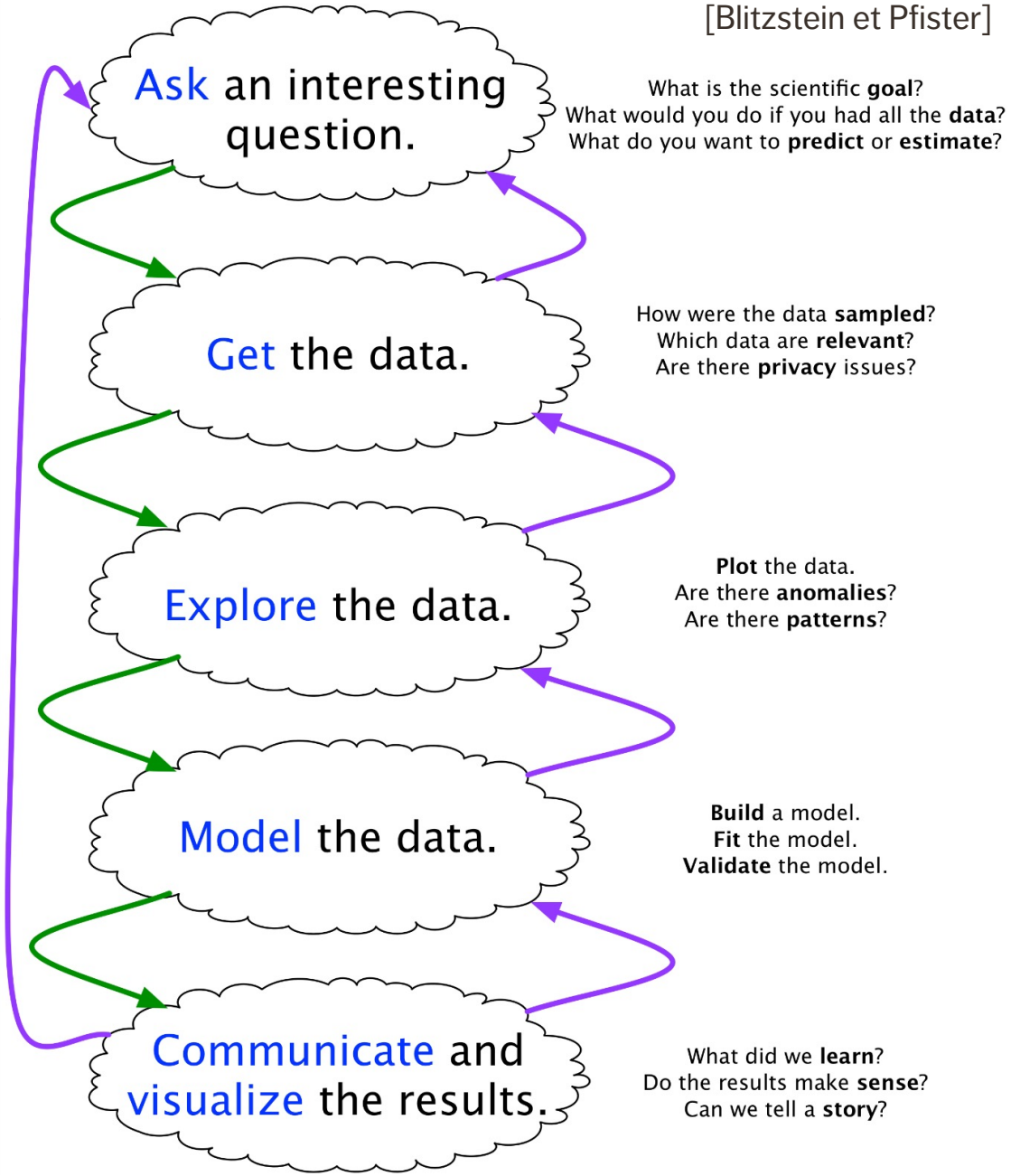
Laissez parler les données.

Recherchez des informations exploitables.

Supervisé vs. non supervisé.

Nettoyer, préparer et visualiser les données.

[Blitzstein et Pfister]



Exercices

Faits divers

1. Quelle est votre approche préférée ? "éprouvée, testée et vraie" ou "science des données perturbatrice" ? Que faudrait-il pour que vous considériez l'envers de la médaille ?

2. Vrai ou faux ?
   a. La performance prédictive d'un modèle supervisé est évaluée sur l'ensemble de formation.
   b. La validation croisée peut être utilisée pour réduire le risque de surajustement d'un modèle prédictif.
   c. Il est toujours préférable d'utiliser autant de variables que possible dans un modèle.
   d. Si les observations comportant des valeurs manquantes sont supprimées, cela peut entraîner des biais et des erreurs.
   e. Nous pouvons utiliser un algorithme de regroupement pour prédire l'appartenance à une classe.

Exercices

Faits divers

2. Vrai ou faux ? (suite)
   f. Si toutes les méthodes ne donnent pas le même résultat, c'est la preuve qu'il est impossible de répondre à la question.
   g. La connaissance du domaine n'est nécessaire que lorsque l'on travaille avec des données anciennes.
   h. Les sponsors et les clients doivent connaître tous les détails de l'analyse.
   i. Il est impossible de planifier le processus d'analyse des données avant de savoir à quoi elles ressemblent.
   j. Les données disponibles ne sont pas toujours appropriées ou représentatives de la situation que nous modélisons.

3. De quelle manière voyez-vous la DS/ML devenir un élément crucial de votre travail ? Cette évolution est-elle bien accueillie ? Comment souhaitez-vous être impliqué ?