FOUILLE DE DONNÉES ET AIDE A LA DECISION

Introduction au machine learning.

Anne-Claire Haury

M2 Informatique
Université Denis Diderot

Premier semestre 2017-2018

LE MACHINE LEARNING

UNE SCIENCE À LA MODE

Pourquoi ?

Stockage et traitement des données : de moins en moins cher.

Impossible de les comprendre "à la main". Exemples : SNCF,

génétique, finance, réseaux sociaux, publicité...
Dépendance d'un grand nombre de facteurs.

Big Data: le mot magique (qui n'a pas toujours de sens)

Compétences recherchées par les entreprises (mots-clés) :
datamining, analyse de données, big data, traitement automatique

de texte, d'images, machine learning...
$$$$$

RENDRE LES ORDINATEURS INTELLIGENTS

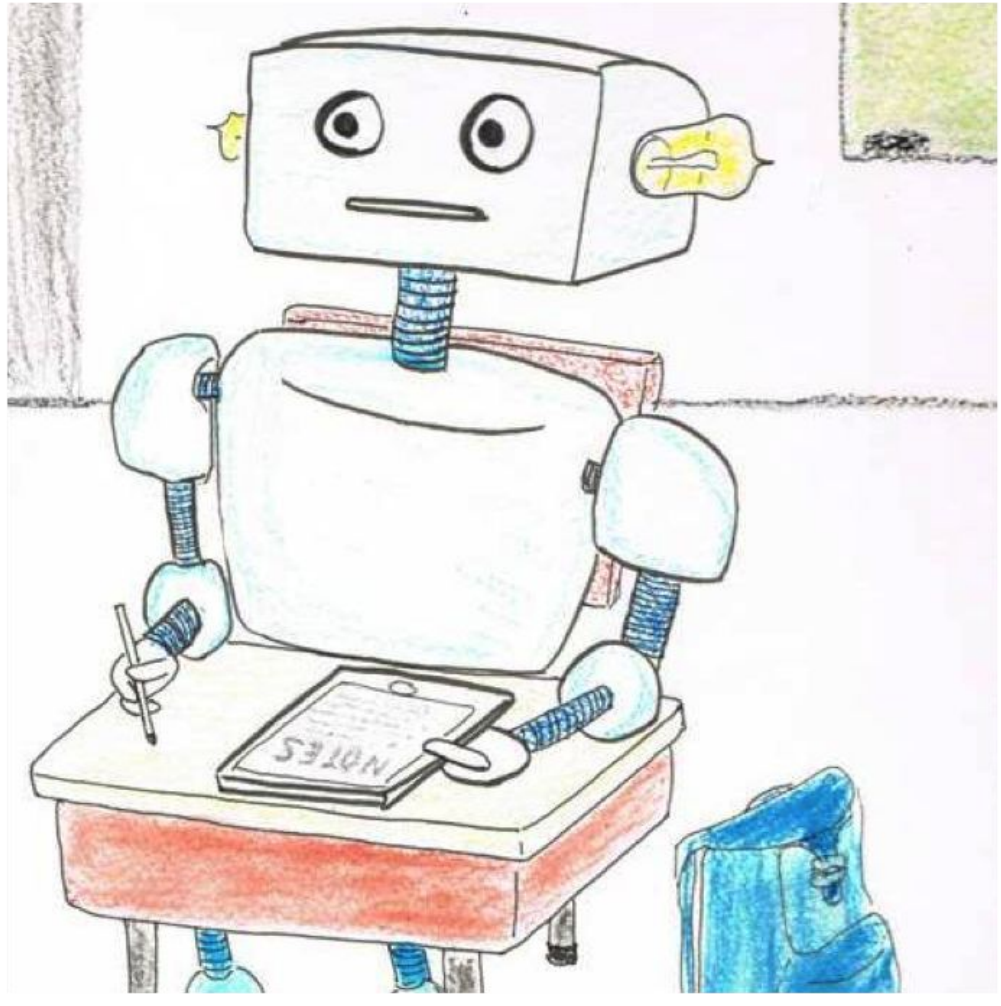


Figure: Tiré du blog du laboratoire "Computer and Cognition", NYU.

LA RENCONTRE DE PLUSIEURS DISCIPLINES

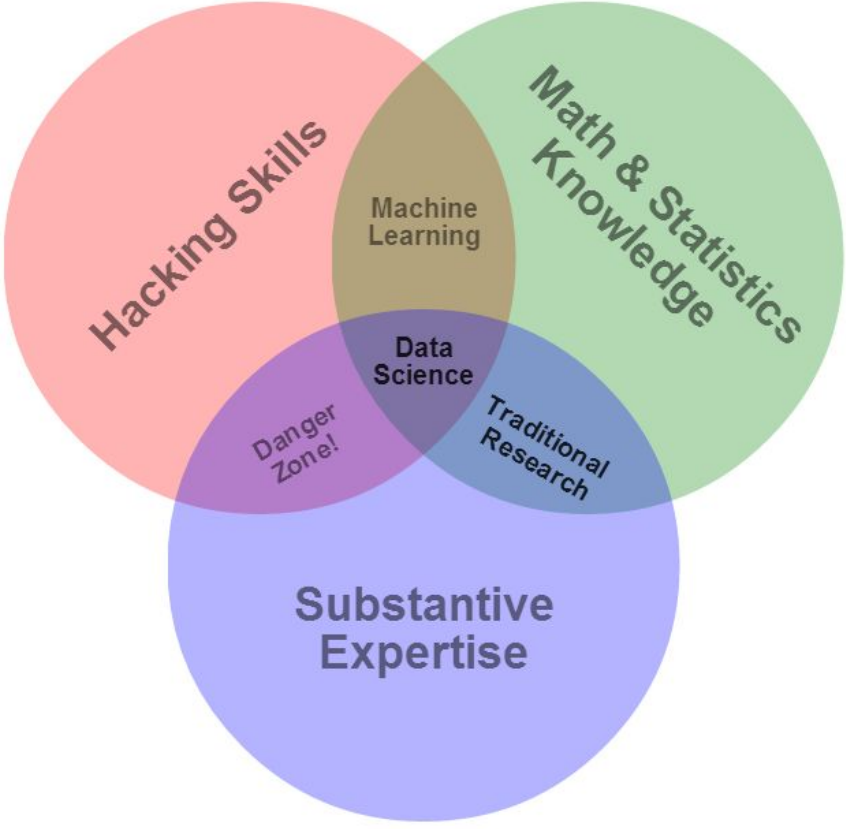


Figure: Tiré de econometricsense.blogspot.fr

Changement de cap de plus en plus observé :

des statistiques traditionnelles aux modèles algorithmiques.

Besoin de modélisation

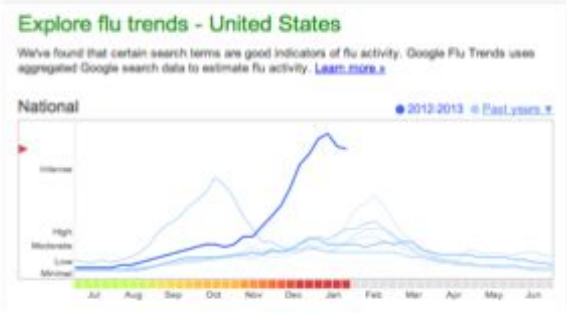
mais aussi de méthodes

rapides et généralisables à la

grande dimension.

QUELQUES APPLICATIONS

APPLICATIONS WEB



PROGRAMME DU COURS

- Transformation des données, encodage.
- Décision statistique: statistiques 101, tests, intervalles de confiance.
- Apprentissage non supervisé: clustering, réduction de dimension.
- Moteurs de recommandation.
- Apprentissage supervisé: régression, classification. Entre autres: Naive Bayes, régression linéaire, régression logistique, arbres de décision, random forests, SVM, réseaux de neurones.
- Si suffisamment de temps: séries temporelles/graphes.

VALIDATION DU COURS

VALIDATION DU COURS

Note finale = 50% projet - 50% examen.

Examen de rattrapage pour les étudiants ayant obtenu plus de 7/20.

La présence au cours/TP n'est pas obligatoire. En revanche pas d'ordinateur ni de téléphone portable pendant le CM. (C'est le moment de les ranger!)

Tout copier/coller ou plagiat dans le code ou le rapport de projet, ou triche à l'examen = 0 sans possibilité de rattrapage.

SITE WEB

https://sites.google.com/site/dataminingp7

EXEMPLES DE PROJETS

https://www.kaggle.com/datasets

Inspirez-vous, choisissez et validez avec Baptiste et/ou moi.

ORGANISATION

2 personnes maximum par projet.

Rapport écrit, programme et oral.

Les projets sont valorisables sur un CV.

Projet de A à Z: collecte des données + encodage + visualisation + méthodo + résultats.

CODE

Un programme (un minimum documenté, au moins commenté) doit accompagner le projet.

Langage: python de préférence, Java/C++ si justifié.

En fonction de votre projet: appli web, page html, exécutable, script... (Pas forcément d'interface.)

RAPPORT DE PROJET

Rapport à rendre avec le programme.

Une dizaine de pages (plus si nécessaire) comprenant:

- Présentation du projet/motivation.
- Description (visuelle et/ou tableau) des données.
- Méthodo utilisée.
- Résultats.
- Conclusion.

PLANNING

Semaines 2/3: Choix du projet et formation des équipes.

Semaines 4 à 9: Analyse et rédaction. 1 RDV de suivi par groupe et suivi par mail en permanence.

Semaine 9: rendu du rapport.

Semaine 10: oral/démo (pendant le dernier cours).

Avant le stage: obtention des notes (pas le plus important!)

L'ESPRIT DU COURS

Interactif
Travail d'équipe

Appliqué

Toute proposition de thèmes à aborder est toujours la bienvenue.

EXAMEN INDIVIDUEL

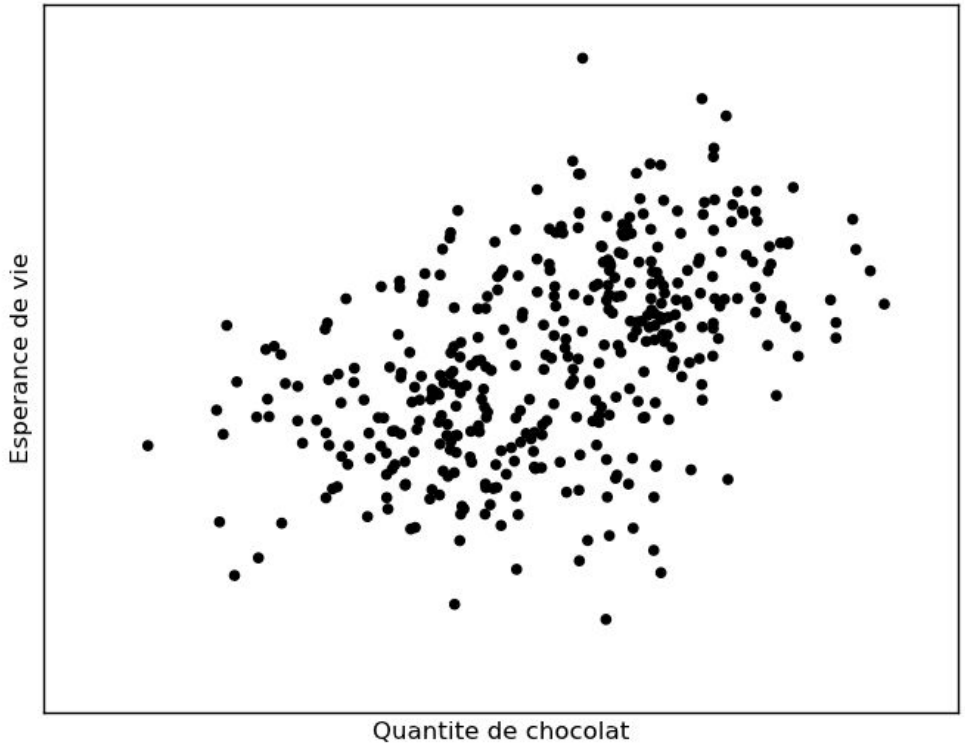
Devoir sur table sans documents, en milieu/fin de semestre.

QUESTIONS?

INTRODUCTION

CHOCOLAT ET ESPÉRANCE DE VIE

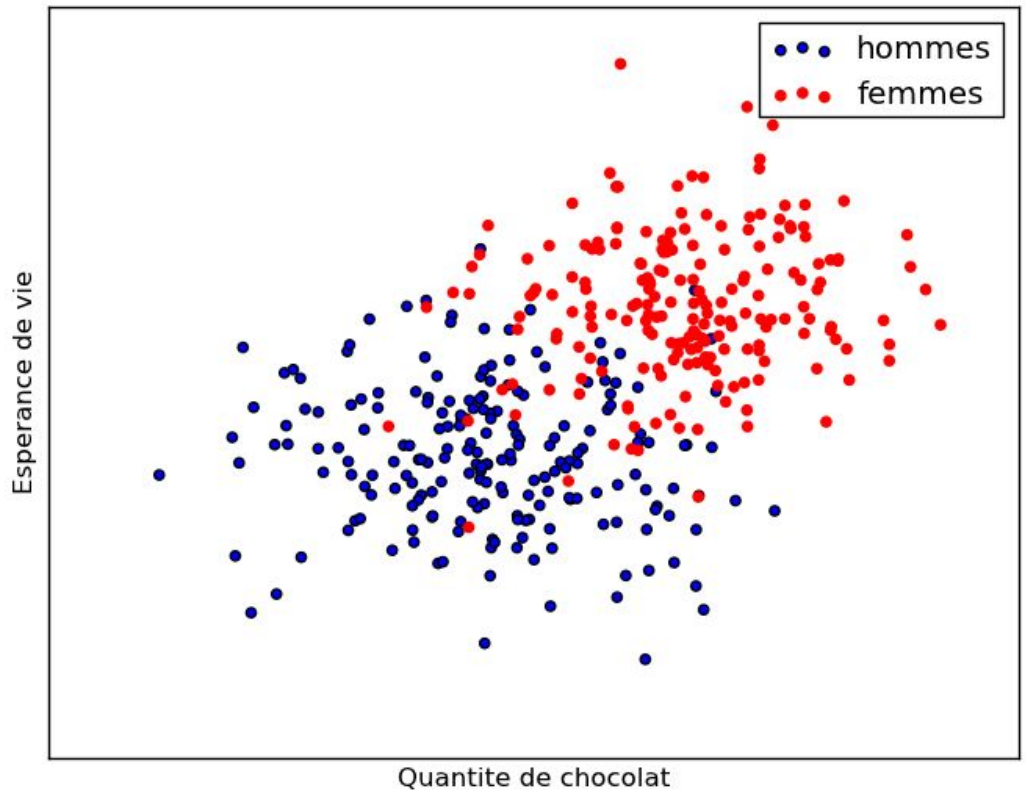
Exemple emprunté à Isabelle Guyon



Manger du chocolat augmente l'espérance de vie.

CHOCOLAT ET ESPÉRANCE DE VIE

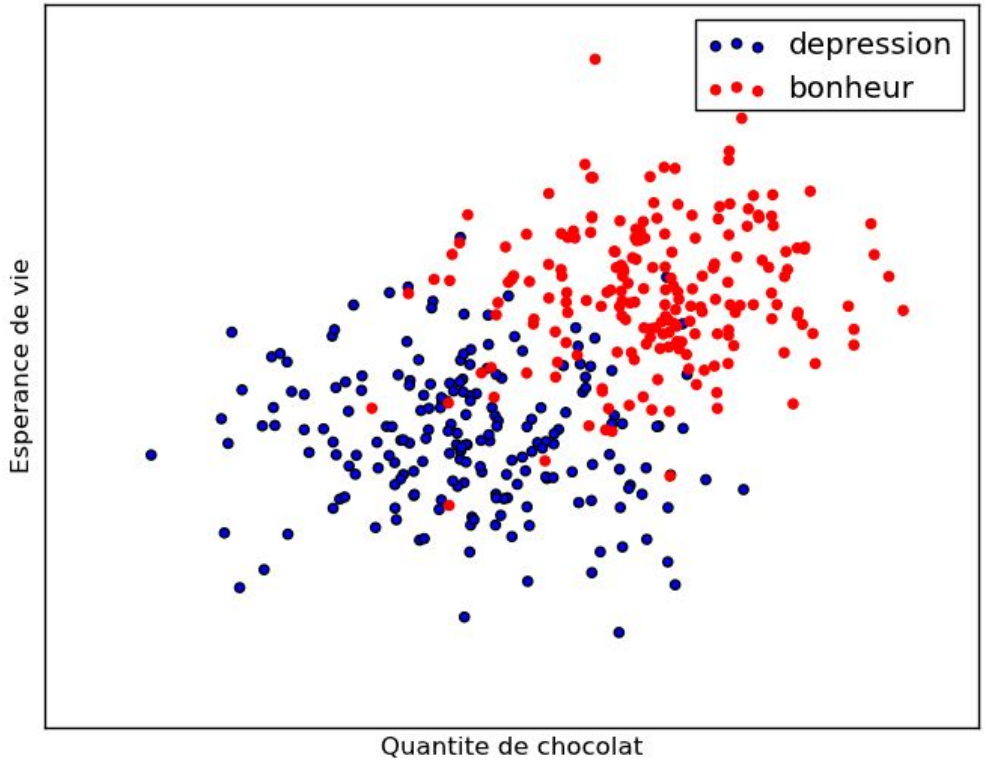
Exemple emprunté à Isabelle Guyon



Manger du chocolat n´augmente pas l´espérance de vie.

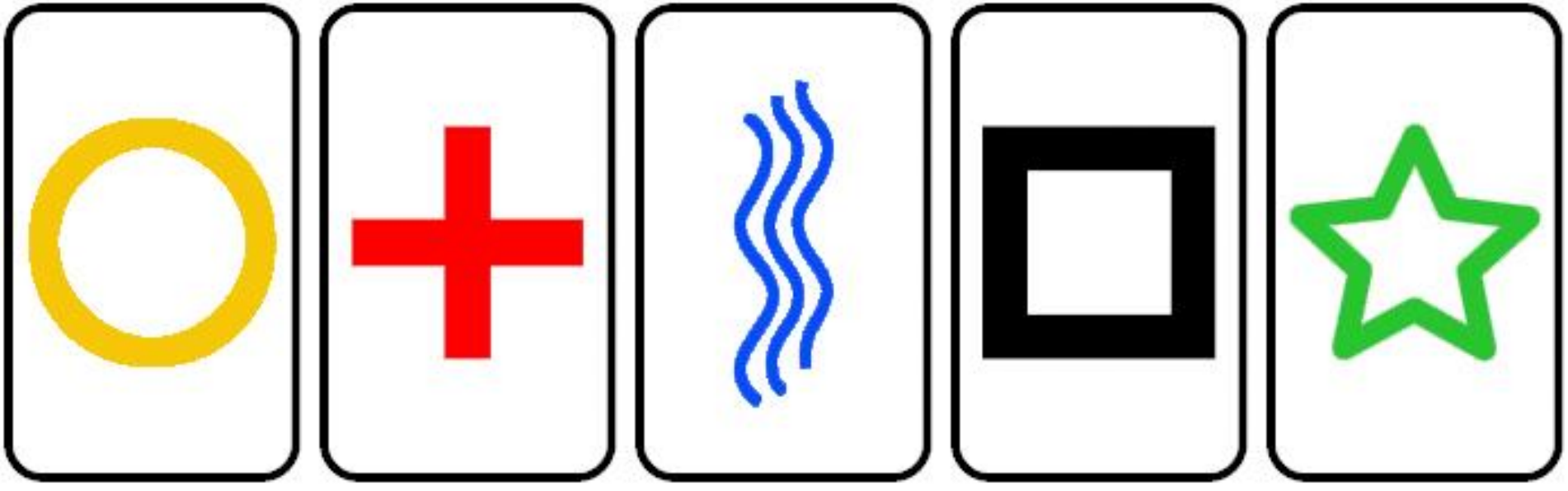
CHOCOLAT ET ESPÉRANCE DE VIE

Exemple emprunté à Isabelle Guyon



Manger du chocolat augmente peut-être l'espérance de vie.

LES EXPÉRIENCES DE RHINE



Source: Wikipedia

PILE OU FACE?

PILE OU FACE?

Conclusion: porter un t-shirt rouge augmente les chances de tirer des faces...

VÊTEMENTS ET FÉCONDITÉ

Women Are More Likely to Wear Red or Pink at Peak Fertility

Alec T. Beall

Jessica L. Tracy

University of British Columbia

Alec T. Beall, Department of Psychology, University of British Columbia, 2136 West Mall, Vancouver, British Columbia V6T 1Z4, Canada E-mail: alec@psych.ubc.ca

Author Contributions Both authors contributed to the study design. Data collection, analyses, and interpretations were performed by A. T. Beall under the supervision of J. L. Tracy. Both authors contributed to the composition of the manuscript, with A. T. Beall composing initial drafts. Both authors approved the final version of the manuscript for submission.

Abstract

Although females of many species closely related to humans signal their fertile window in an observable manner, often involving red or pink coloration, no such display has been found for humans. Building on evidence that men are sexually attracted to women wearing or surrounded by red, we tested whether women show a behavioral tendency toward wearing reddish clothing when at peak fertility. Across two samples (N = 124), women at high conception risk were more than 3 times more likely to wear a red or pink shirt than were women at low conception risk, and 77% of women who wore red or pink were found to be at high, rather than low, risk. Conception risk had no effect on the prevalence of any other shirt color. Our results thus suggest that red and pink adornment in women is reliably associated with fertility and that female ovulation, long assumed to be hidden, is associated with a salient visual cue.

VÊTEMENTS ET FÉCONDITÉ

Women Are More Likely to Wear Red or Pink at Peak Fertility

Alec T. Beall

Jessica L. Tracy

University of British Columbia

Alec T. Beall, Department of Psychology, University of British Columbia, 2136 West Mall, Vancouver, British Columbia V6T 1Z4, Canada E-mail: alec@psych.ubc.ca

Author Contributions Both authors contributed to the study design. Data collection, analyses, and interpretations were performed by A. T. Beall under the supervision of J. L. Tracy. Both authors contributed to the composition of the manuscript, with A. T. Beall composing initial drafts. Both authors approved the final version of the manuscript for submission.

Abstract

Although females of many species closely related to humans signal their fertile window in an observable manner, often involving red or pink coloration, no such display has been found for humans. Building on evidence that men are sexually attracted to women wearing or surrounded by red, we tested whether women show a behavioral tendency toward wearing reddish clothing when at peak fertility. Across two samples (N = 124), women at high conception risk were more than 3 times more likely to wear a red or pink shirt than were women at low conception risk, and 77% of women who wore red or pink were found to be at high, rather than low, risk. Conception risk had no effect on the prevalence of any other shirt color. Our results thus suggest that red and pink adornment in women is reliably associated with fertility and that female ovulation, long assumed to be hidden, is associated with a salient visual cue.

Conclusion: les femmes atteignant leur pic de fécondité portent 3 fois plus de vêtements rouges que les autres...

ESPRIT STATISTIQUEMENT CRITIQUE

Les absurdités et manipulations à base de chiffres sont partout : politique, presse, et même recherche.

Les chiffres ont, pour la plupart des gens, une autorité intrinsèque ("c'est scientifique").

Les conclusions ne sont que le fruit de l'interprétation. Il faut dissocier résultats et conclusion.

On ne fait rien dire du tout aux chiffres, mais on peut les utiliser pour faire passer ses opinions.

Un des objectifs de ce cours : ne plus se faire manipuler !

VOUS AVEZ UNE MAUVAISE INTUITION STATISTIQUE (SI SI!)

Un dîner presque parfait

PEUT-ON FAIRE DIRE AUX CHIFFRES CE QUE L'ON VEUT ?

Données fictives !	Nombre de chômeurs	Nombre de travailleurs potentiels
Année 1	1.000.000	10.000.000
Année 2	1.010.000	11.000.000

VRAI
OU
FAUX ?

"Le chômage a augmenté de 1%."

"Le taux de chômage a baissé de 0.9 points."

"Il y a 10.000 chômeurs de plus."

"Le taux de chômage a baissé de 10%."

PEOPLE VS COLLINS

1964. Un vol. Les témoins affirment avoir vu un homme noir barbu et moustachu et une femme blonde avec une queue de cheval s'enfuir dans une voiture jaune. Malcolm et Janet Collins correspondent à la description...

PEOPLE VS COLLINS

Raisonnement du procureur :

- Homme noir portant une barbe : 10%
- Homme noir portant une moustache : 25%
- Femme blanche portant une queue de cheval : 10%
- Femme blanche ayant des cheveux blonds : 33%
- Voiture en partie jaune : 10%
- Couple “inter-racial” dans une voiture : 0.1%

Ils en concluent que la probabilité que les Collins soient innocents est de 1/12 millions. Ils sont donc condamnés.

La cour d’appel annule la condamnation. Quelle était l’erreur du jury lors du procès en première instance ?

EXPLICATION

Admettons que les probabilités, bien qu'estimées sans doute arbitrairement, soient justes.

L'erreur principale est d'avoir ignoré les dépendances entre les événements.

Au lieu de multiplier toutes les probabilités entre elles, il faut prendre en compte le fait que les événements ne sont pas indépendants et considérer les probabilités conditionnelles.

Par exemple, la probabilité d'avoir une moustache sachant que l'on a une barbe est très élevée, disons 90%. Donc la probabilité d'avoir une barbe ET une moustache devient 10% x 90% au lieu de 10% x 25%. De même pour les autres événements.

PARADOXE DE SIMPSON

100 étudiants (50 hommes et 50 femmes) sont répartis sur 2 cours : fouille de données et systèmes avancés. Voici leurs pourcentages de validation des cours (exemple fictif!).

Fouilles de données		Systèmes avancés	
Hommes	Femmes	Hommes	Femmes
90%	84.5%	70%	60%

Les hommes réussissent mieux chacun des cours.

PARADOXE DE SIMPSON

100 étudiants (50 hommes et 50 femmes) sont répartis sur 2 cours : fouille de données et systèmes avancés. Voici leurs pourcentages de validation des cours (exemple fictif!).

Fouilles de données		**Systèmes avancés**	
Hommes	Femmes	Hommes	Femmes
90%	84.5%	70%	60%

Réussite globale:

Hommes	Femmes
74%	82%

EXPLICATION

Les femmes sont plus nombreuses dans le cours où elles réussissent le mieux. Dans le cours où elles réussissent mieux, elles font un meilleur score que les hommes dans le cours où ils réussissent mieux. C'est donc une question de répartition.

Fouilles de données		**Systèmes avancés**	
Hommes	Femmes	Hommes	Femmes
90% (9/10)	84.5% (38/45)	70% (28/40)	60% (3/5)

Réussite globale:

Hommes	Femmes
74% (37/50)	82% (41/50)

PARADOXE DES ANNIVERSAIRES

Quelle est la probabilité que deux personnes parmi vous aient la même date d'anniversaire ?

https://goo.gl/forms/ZiD5U83M7ZPyHElq2

ou

https://sites.google.com/site/dataminingp7/formulaires

PARADOXE DES ANNIVERSAIRES

Quelle est la probabilité que deux personnes parmi vous aient la même date d'anniversaire ?

> 50% si vous êtes plus de 23

> 80% si vous êtes plus de 35

> 90% si vous êtes plus de 41

> 95% si vous êtes plus de 47

> 99% si vous êtes plus de 58

EXPLICATION

Il serait **très improbable** que vous ayez tous une date différente d'anniversaire. Itérons:

- La première personne choisit sa date parmi 365 dates. Il reste 364 choix pour la seconde.
- La seconde choisit sa date. Il reste 363 choix.
- La n-ème personne a (365 − n + 1) choix.

Si on transforme cela en probabilités, on obtient :

$$p = \frac{365}{365} \times \frac{364}{365} \times \cdots \times \frac{365 - n + 1}{365}$$

p est la probabilité que les n personnes aient des anniversaires différents. Très rapidement, cette probabilité devient **infime** (on ne multiplie que des nombres < 1). La probabilité que deux personnes **au moins** partage la même date est donc 1 − p.

EXEMPLE AVEC 50 PERSONNES

Probabilité d'avoir des anniversaires différents:

$$
\begin{aligned}
p &= \frac{365}{365} \times \frac{364}{365} \times \cdots \times \frac{365 - 50 + 1}{365} \\
&= \frac{365 \times 364 \times \cdots \times 316}{365^{50}} \\
&= 0.0296
\end{aligned}
$$

Il y a donc 97% de chances qu'au moins 2 personnes aient le même anniversaire.

ET AU POKER ?

Sur un jeu de 52 cartes, quelle est la probabilité que j'aie une paire d'As ?

A - 2/52
B - 1/52
C - 1/221
D - 3/1225
E - 1/2652

ET AU POKER ?

Sachant que j'ai As/Roi dans la main, quelle est la probabilité que mon adversaire ait une paire d'As?

A - 2/52
B - 1/52
C - 1/221
D - 3/1225
E - 1/2652

CE QU'ON EN CONCLUT

Avoir de l'information change drastiquement la donne!

PARADOXE DES TROIS PORTES (MONTY HALL)

Un candidat à un jeu télévisé se trouve devant 3 portes. Derrière 2 portes, il n'y a rien. Derrière 1 des portes, une voiture.

- Il choisit une porte.
- L'animateur ouvre l'une des deux autres **qui ne cache pas la voiture.**
- Il reste donc 1 porte choisie au départ et une autre porte fermée.
- L'animateur propose au candidat de changer de porte

Le candidat a-t-il intérêt à changer de porte ?

PARADOXE DES TROIS PORTES (MONTY HALL)

Un candidat à un jeu télévisé se trouve devant 3 portes. Derrière 2 portes, il n'y a rien. Derrière 1 des portes, une voiture.

- Il choisit une porte.
- L'animateur ouvre l'une des deux autres **qui ne cache pas la voiture.**
- Il reste donc 1 porte choisie au départ et une autre porte fermée.
- L'animateur propose au candidat de changer de porte

Le candidat a-t-il intérêt à changer de porte ? **OUI**

EXPLICATION

Regardons les probabilités :

- Au départ, le candidat a 1 chance sur 3 de choisir la bonne porte
- Lorsque le présentateur en ouvre une autre qui ne contient pas la voiture, il apporte une information supplémentaire : la porte restante a donc 2 chances sur 3 de contenir la voiture.

Le candidat doit donc changer de porte, passant sa probabilité de gagner de 1/3 à 2/3.

anneclaire.haury@gmail.com