FOUILLE DE DONNÉES
ET AIDE À LA DECISION

Introduction à la data science et au machine learning.

M2 Informatique, Université Paris Cité

Fabien Viger

fabien.viger@gmail.com

TODOs

DATA SCIENCE & MACHINE LEARNING

UNE SCIENCE À LA MODE

Mais pourquoi ?

- ChatGPT, DeepFakes, AlphaZero: gros progrès récents des réseaux de neurones.
- Stockage et traitement des données : de moins en moins cher.
- “Big Data”: Volume + complexité des données, impossible de les comprendre "à la main" : SNCF, génétique, finance, …
- ***Data Scientist***, compétence **très** à la mode.
- Applications: pub, marketing, compréhension des clients → **optimisation**.

RENDRE LES ORDINATEURS INTELLIGENTS

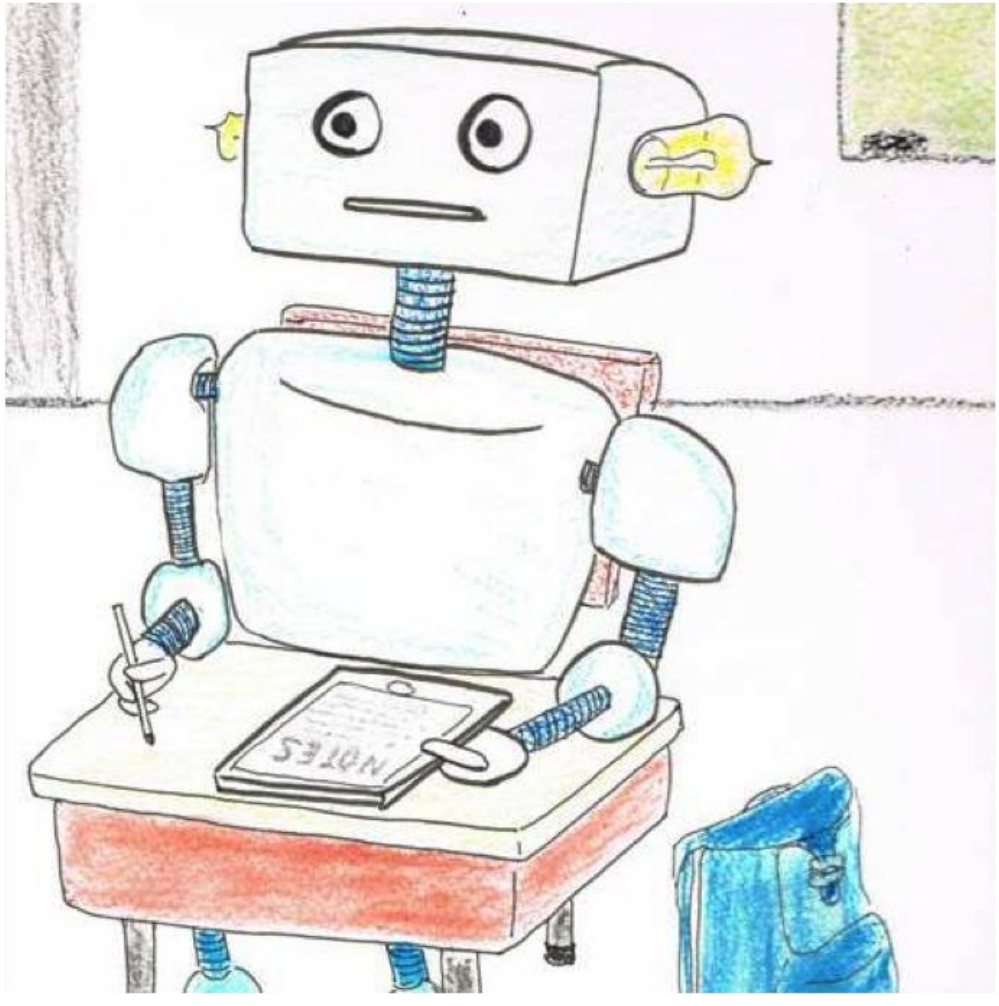


Figure: Tiré du blog du laboratoire "Computer and Cognition", NYU.

LA RENCONTRE DE PLUSIEURS DISCIPLINES

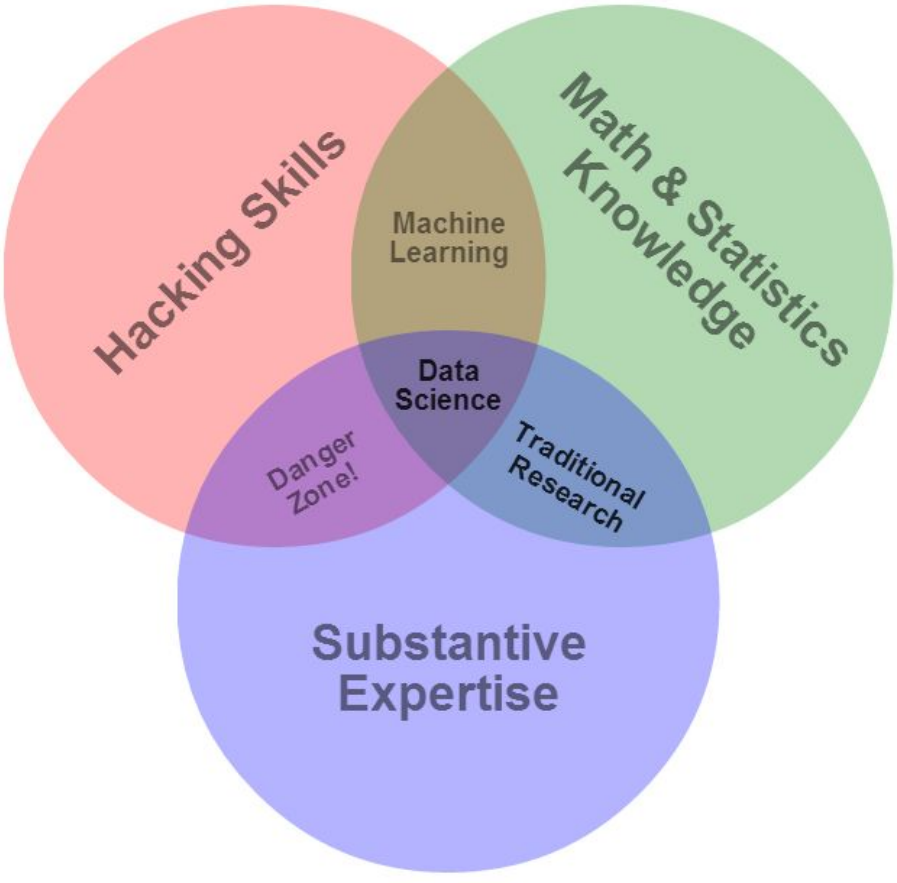


Figure: Tiré de econometricsense.blogspot.fr

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

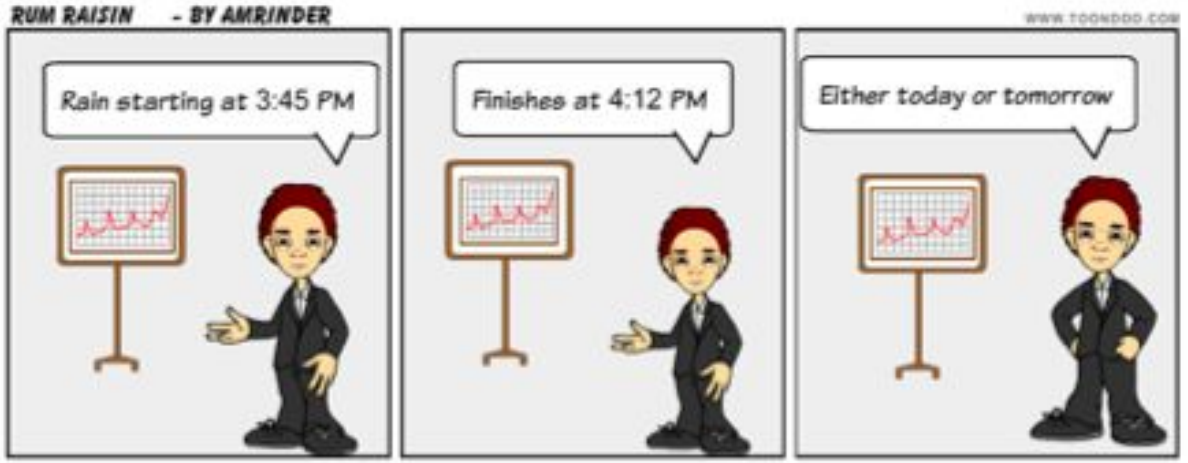
"Écris une introduction au cours de Data Science en français, 30 mots."

"Bienvenue au cours de Data Science ! Dans ce programme, nous explorerons les méthodes et les outils pour extraire des insights précieux à partir des données."

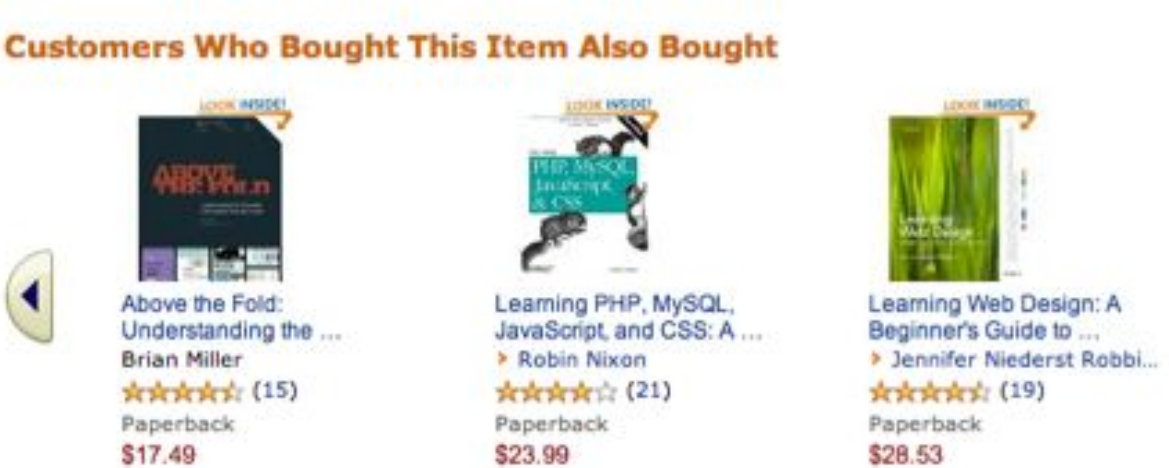
QUELQUES APPLICATIONS

"A fox flying in the sky, wearing a special backpack with four propellers rotating above like a helicopter"

QUELQUES APPLICATIONS



APPLICATIONS WEB



PROGRAMME DU COURS

- Décision **statistique**: statistiques 101, tests, intervalles de confiance.
- **Transformation des données**, encodage (aka ***Feature Engineering***)
- Apprentissage supervisé: Naive Bayes, Régression Linéaire, Régression Logistique, Arbres de Décision, Random Forests, SVM, Réseaux de Neurones.
- **Apprentissage** non supervisé: clustering, réduction de dimension.
- **Moteurs de recommandation**.

VALIDATION DU COURS

VALIDATION DU COURS

Note finale: **50% projet**, 30% TD, 20% QCM

Pas de rattrapage possible.

La présence au cours/TD n'est pas *obligatoire*, mais très fortement recommandée

Ne trichez pas! Ce n'est pas l'esprit du cours. Si c'est trop difficile, dites-le moi, on s'adapte. Je suis impitoyable avec la triche et le plagiat: tout copier/coller ou plagiat dans les TD, ou projet, ou QCM = 0 partout, sans rattrapage possible, + signalement à l'UFR.

SITE WEB

http://fabien.viger.free.fr/ml

PROJET: EXEMPLES

https://www.kaggle.com/datasets

Inspirez-vous, choisissez et validez avec moi.

PROJET: ORGANISATION

1 ou 2 personnes par projet. Pas de **touriste** !

Demi-page de “résumé des contribution personnelles” par écrit.

Soutenance: semaine avant exams, slides (.pdf): 15min, +15min de **questions** (10 + 10 si seul).

Les projets sont valorisables sur un CV !

Projet de A à Z, ou une bonne partie de:
Collecte des données + analyse + encodage [+ visualisation?] + méthodo + résultats.

PROJET: CODE

Du code (un minimum documenté, au moins commenté) doit accompagner le projet.

Il doit être envoyé avec les slides.

Langage: Python, Java, C++ OK; pour tout autre me demander.

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

PROJET: À RENDRE

Liste contributions perso: 1/2-page par étudiant

- Ne doit laisser aucun doute sur qui a fait quoi
- Soyez honnête! Il est normal de réutiliser des choses faites par d'autres. **Dites-le**.

Support de soutenance: ~10-15 pages (1 page par minute). Exemple de Plan:

- Intro
- Données [Analyse? Visualisation?]
- Méthodologie
- Résultats
- Discussion, Conclusion.
- Code: Contributions

PROJET: PLANNING

Semaines 2/3: Choix du projet et formation des équipes.
Semaines 4 à 9: Analyse et rédaction. RDV de suivi par groupe, + suivi par mail.
Semaine 11 (?) (examens): **soutenance**

Obtention des notes de projet rapidement (1 jour ou 2).

EXAMEN(S) INDIVIDUEL: QCM

Examens sur machine ou sur table (sans documents), au cours du trimestre.

Nombre inconnu (au moins 1, peut-être 2)

Pendant les séances de TD (je préviens 1 semaine en avance)

TD/TP

Les TD sont sur machine, en **python**.

Ils doivent être rendus (Moodle), et seront **évalués** (automatiquement)
Rendu de TD: individuel, le jour même, à minuit.

Je fournirai des auto-tests dans la mesure du possible

Un corrigé de TD apparaîtra en ligne rapidement après le rendu.

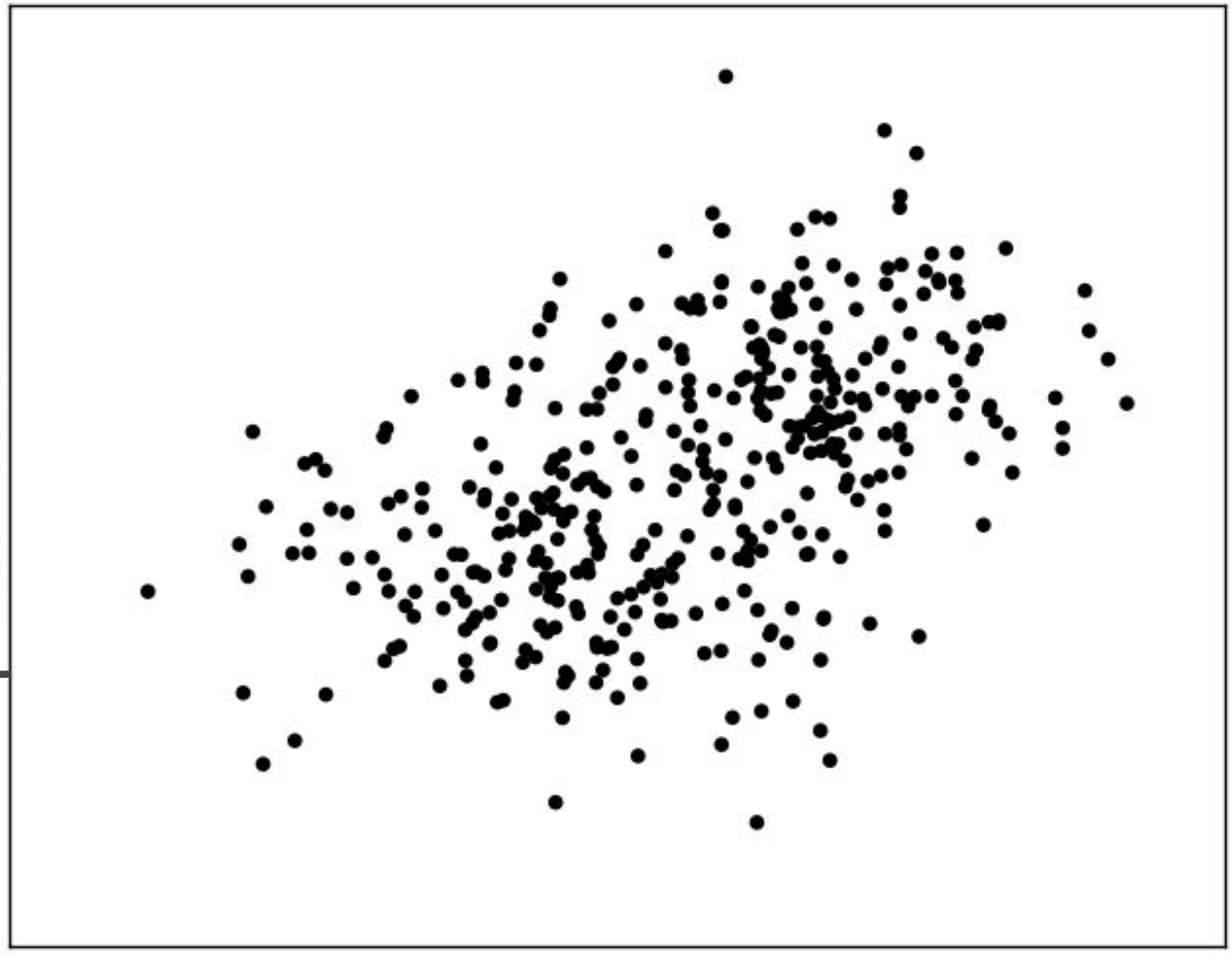
L'ESPRIT DU COURS

Interactif
Travail d'équipe
Appliqué
Toute proposition de thèmes à aborder est toujours la bienvenue.

QUESTIONS?

INTRODUCTION

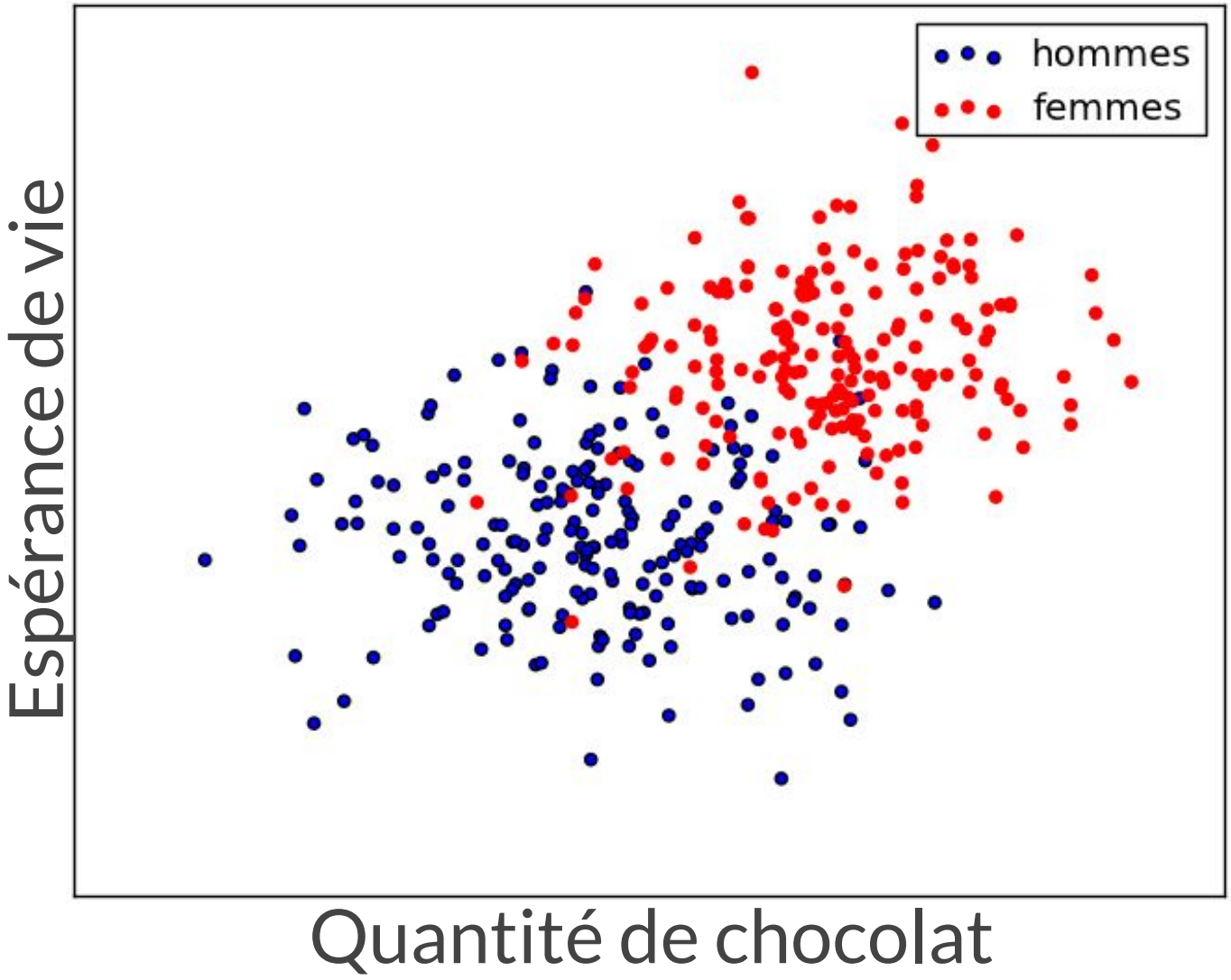
CHOCOLAT ET ESPÉRANCE DE VIE



Exemple fictif emprunté à Isabelle Guyon

Manger du chocolat augmente l'espérance de vie !

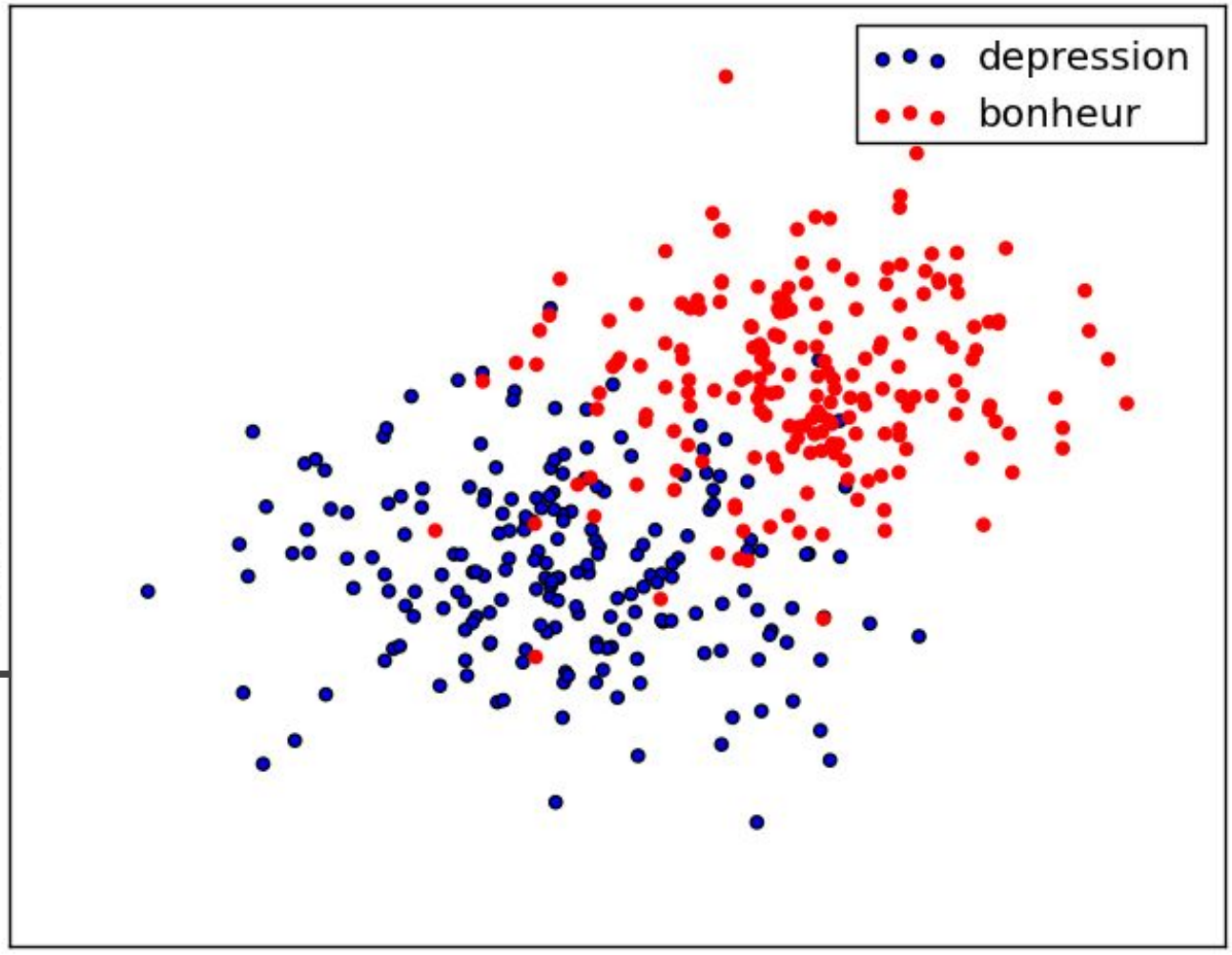
CHOCOLAT ET ESPÉRANCE DE VIE



Exemple fictif emprunté à Isabelle Guyon

Manger du chocolat n'augmente pas l'espérance de vie !

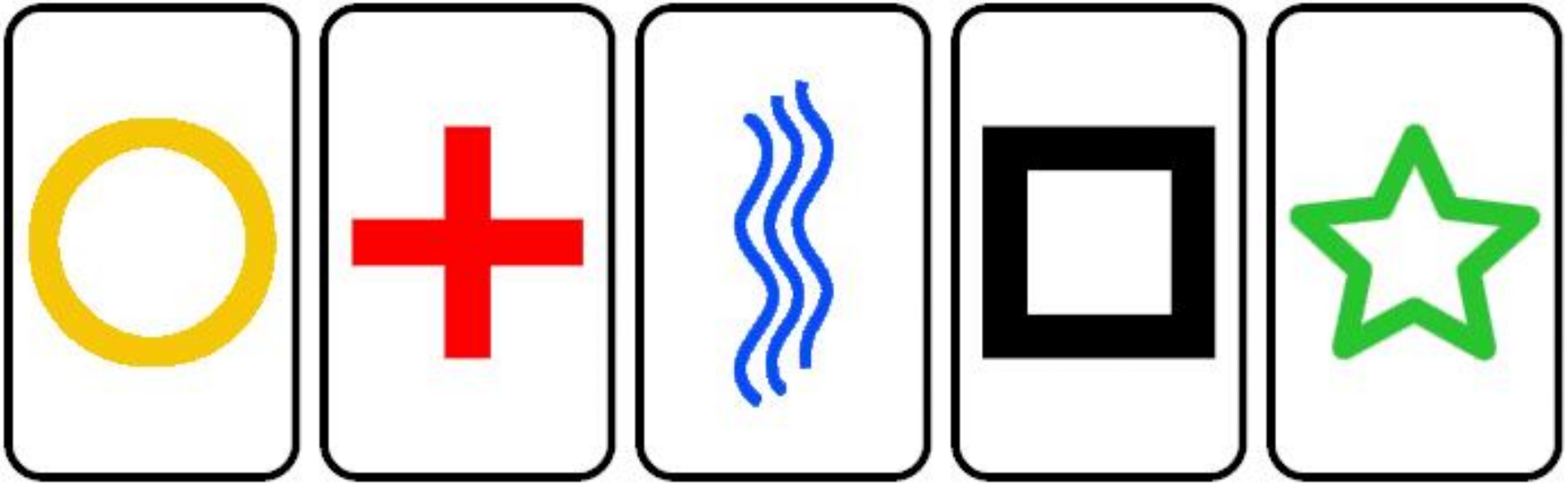
CHOCOLAT ET ESPÉRANCE DE VIE



Exemple fictif emprunté à Isabelle Guyon

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

LES EXPÉRIENCES DE RHINE



Source: Wikipedia

VOUS AVEZ UNE MAUVAISE INTUITION STATISTIQUE (SI SI!)

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

[Données fictives !]	Nombre de chômeurs	Nb. 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 pts.
- 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 avec queue de cheval : 10%
- Femme blanche aux cheveux blonds : 33%
- Voiture en partie jaune : 10%
- Couple “inter-racial” dans une voiture : 0.1%

Conclusion: probabilité que les Collins soient innocents < 1/12 **millions**. Ils sont **condamnés**.

La cour d’appel **annule** cela. **Pourquoi ?**

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 les probabilités entre elles, comme les événements ne sont **pas** indépendants, il faut considérer les probabilités **conditionnelles**.

Exemple: la probabilité (à l'époque) d'avoir une moustache sachant que l'on a une barbe est très élevée, disons 80%. Donc la probabilité d'avoir une barbe ET une moustache devient 10% x 80% au lieu de 10% x 25%. Idem pour les autres évènements.

PARADOXE DE SIMPSON

100 étudiants (50 IMPAIRS et 50 DATA) choisissent chacun entre 2 cours : **Fouille** et **Systèmes**. Pourcentages de validation [fictif!] :

Fouilles de données		**Systèmes avancés**	
IMPAIRS	DATA	IMPAIRS	DATA
90%	84.5%	70%	60%

Les IMPAIRS réussissent mieux chacun des cours…. Donc ils ont ***forcément*** un meilleur taux de réussite globale (sur les 2 cours), non?

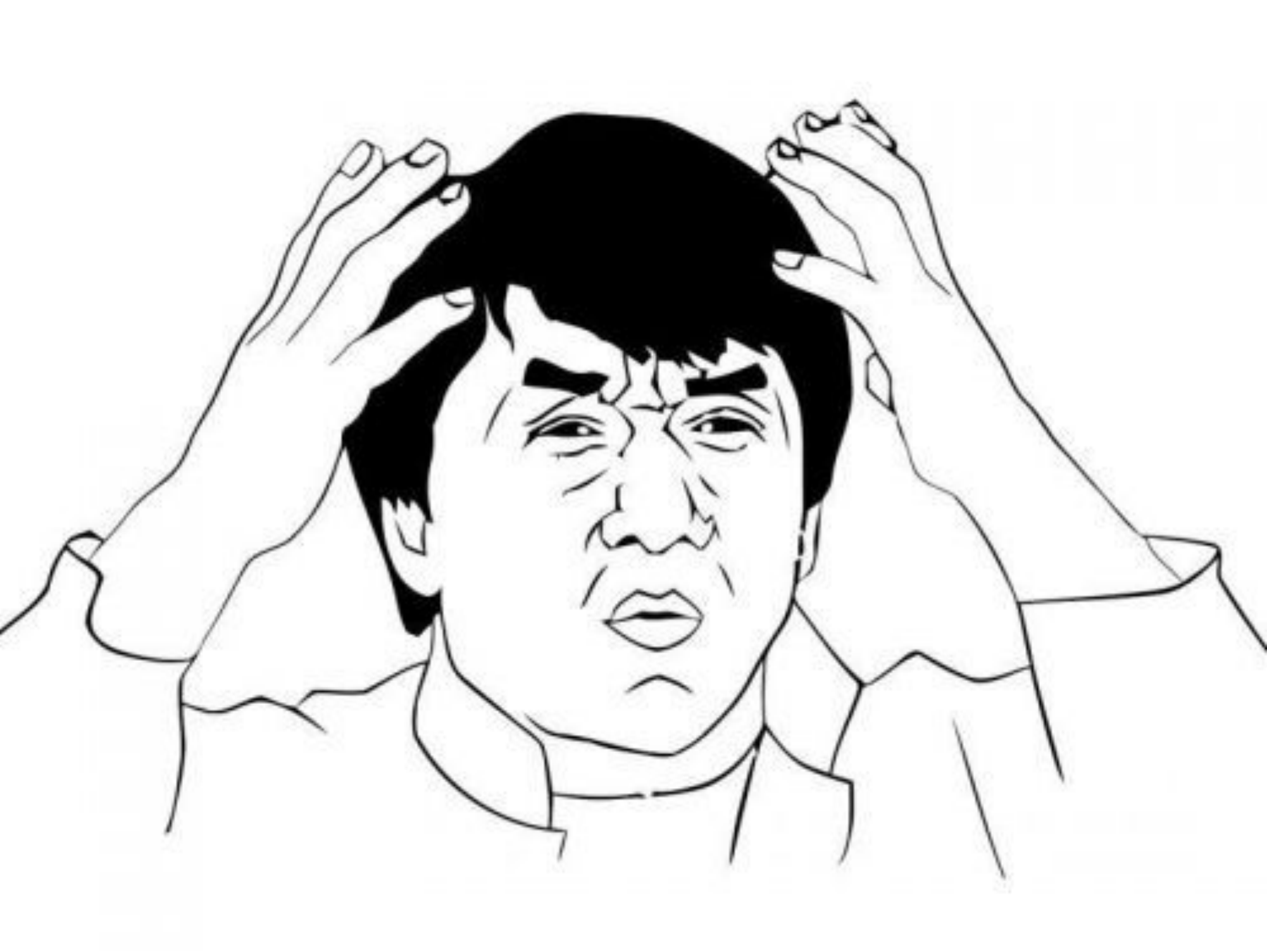
PARADOXE DE SIMPSON

100 étudiants (50 IMPAIRS et 50 DATA) choisissent chacun entre 2 cours : **Fouille** et **Systèmes**. Pourcentages de validation [fictif!] :

Fouilles de données		**Systèmes avancés**	
IMPAIRS	DATA	IMPAIRS	DATA
90%	84.5%	70%	60%

Réussite **globale**: (sur l'ensemble des 2 cours)

IMPAIRS	DATA
74%	82%



EXPLICATION

Les DATA préfèrent le cours où ils réussissent bien. Leur faible réussite en Systèmes ne compte presque pas: ils sont peu. Mot-clé: Répartition.

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

Réussite **globale**: (sur l'ensemble des 2 cours)

IMPAIRS	DATA
74%	82%

PARADOXE DES ANNIVERSAIRES

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

N / 365 ?

PARADOXE DES ANNIVERSAIRES

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

> 50% si N ≥ 23
> 80% si N ≥ 35
> 90% si N ≥ 41
> 95% si N ≥ 47
> 99% si N ≥ 58
> 99.9% N ≥ 69 (et dans ce cas, 1 chance sur 3 d'avoir 3 qui ont la même)

EXEMPLE AVEC 50 PERSONNES

Probabilité que 50 personnes aient des anniversaires **tous différents:**

$$p = \frac{365}{365} \times \frac{364}{365} \times \cdots \times \frac{365 - 50 + 1}{365}$$

$$= \frac{365 \times 364 \times \cdots \times 316}{365^{50}}$$

$$= 0.0296$$

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

INTUITION

Il y a $N \times (N-1) / 2 = O(N^2)$ **paires** de personnes dans un groupe de N personnes.

On peut imaginer que la probabilité d'avoir les anniversaires tous différents est similaire à la probabilité que **chaque paire**, de manière indépendante (en première approximation), ait un anniversaire différent: $P \approx (364/365)^{(N^2/2)}$

$P \approx (1 - 1/365)^{(N^2/2)}$

Quand $N^2/2 \approx 365$, i.e. N=27, ça ressemble à $\lim_{N \to \infty} (1-1/N)^N = \mathbf{e}^{-1} \approx 0.368$ (vraie: 0.373..)

TEST: AU POKER (TEXAS HOLD'EM)

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

TEST: AU POKER (TEXAS HOLD'EM)

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 (parfois *drastiquement*) la donne!

PARADOXE DES TROIS PORTES (MONTY HALL)

Jeu télé. Le candidat est devant 3 portes. 1 porte cache une voiture, les 2 autres **rien**.

- Il choisit une porte.
- L'animateur ouvre **systématiquement** 1 des 2 autres, **qui ne cache pas la voiture.**
- L'animateur propose au candidat de changer son choix de porte s'il le désire

Le candidat doit-il changer de porte ?

PARADOXE DES TROIS PORTES (MONTY HALL)

Jeu télé. Le candidat est devant 3 portes. 1 porte cache une voiture, les 2 autres **rien**.

- Il choisit une porte.
- L'animateur ouvre **systématiquement** 1 des 2 autres, **qui ne cache pas la voiture.**
- L'animateur propose au candidat de changer son choix de porte s'il le désire

Le candidat doit-il changer de porte ? **OUI**

EXPLICATION

Regardons les probabilités :

- Au départ, le candidat a 1 chance sur 3 de choisir la bonne porte. Donc 2/3 que la bonne porte soit parmi les 2 autres.
- Lorsque le présentateur en ouvre une autre qui ne contient pas la voiture, il apporte une information : il élimine une des 2 portes restantes. Donc l'autre a 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.

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 sont 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 !

Préparation au TD(s) à venir

Les TDs sont en Python. **Sondage**!

A) Je suis un assez bon codeur en python

B) J'ai déjà codé en python, je m'en souviens

C) Jamais

D) C'est quoi un python?

Python a une **excellente** documentation en ligne. Exemple: "python sort" sur Google

STATISTICS 101

TODOs

NOUS PARLERONS DE

Variables aléatoires

Loi de probabilité

Echantillons

Significativité et intervalles de confiance

Tests statistiques

EXEMPLE 1

On cherche à répondre à ce type de questions:

Sur 1000 personnes (500 hommes, 500 femmes), on observe que les femmes gagnent en moyenne 2000 euros et les hommes 2300.

Peut-on en déduire que les femmes gagnent moins que les hommes?

Même question avec 10 femmes, 10 hommes.
Idem avec 100 femmes, 100 hommes.

EXEMPLE 2

On parle de **A**/**B** testing.

Sur une page web, on peut choisir 2 modèles pour le même bouton: un rouge et un bleu.

- Sur 1000 personnes l'ayant vu, 23 ont cliqué sur le rouge.
- Sur 500 personnes l'ayant vu, 17 ont cliqué sur le bleu.

Peut-on dire : "le bouton bleu marche mieux?"

EXEMPLE 3

Peut-on affirmer qu'il y a plus de garçons que de filles en informatique ?

LA SIGNIFICATIVITÉ

On se pose donc la question de la **fiabilité** d'un chiffre et de la validité des conclusions.

DÉFINITION

Un résultat est **statistiquement significatif** à 5% (ou 95% selon les conventions de langage) si la probabilité qu'on l'observe par "chance" (eg. par hasard) est **≤ 5%.**

Exemple: "La pièce est truquée: plus souvent face que pile" avec (3 face, 1 pile), avec (10 face, 0 pile), avec (103 face, 57 piles).

DÉFINITION

Un résultat est **statistiquement significatif** à 5% (ou 95% selon les conventions de langage) si la probabilité qu'on l'observe par "chance" (eg. par hasard) est **≤ 5%.**

Exemple: "La pièce est truquée: plus souvent face que pile" avec (3 face, 1 pile), avec (10 face, 0 pile), avec (103 face, 57 piles).

Intuitivement, cela a un rapport avec:

- **La taille de l'échantillon**
- **L'hétérogénéité de l'échantillon.**

VARIABLES ALÉATOIRES

Définition (pas très mathématique)

Une variable aléatoire (v.a.) est une application définissant l'ensemble des résultats possibles pour une expérience donnée.

Une variable aléatoire est donc la chose que l'on **observe** et qui nous intéresse.

LES VARIABLES DISCRÈTES

Définition

Une variable aléatoire (v.a.) est dite **discrète** lorsqu'elle peut prendre un nombre **fini** (ou infini dénombrable) de valeurs.

LES VARIABLES DISCRÈTES

Exemples:

- X variable aléatoire liée à l'expérience "pile ou face" : les valeurs possibles de X sont 0 et 1.
- X variable aléatoire liée à l'expérience "réponse à un sondage": $X \in$ {"pas satisfait", "plutôt satisfait", "très satisfait"}.
- X variable aléatoire liée à l'expérience "nombre de personnes dans un supermarché":
  $X \in \mathbb{N}$ c'est-à-dire $X \in \{0, 1, 2, \ldots, \}$

LES VARIABLES CONTINUES

Définition

Une variable aléatoire (v.a.) est dite **continue** lorsqu'elle peut prendre un nombre **infini indénombrable** de valeurs.

LES VARIABLES CONTINUES

Exemples:

- X est la variable aléatoire liée à l'expérience "heure d'arrivée d'un ami": les valeurs possibles de X se trouvent dans [15h, 16h].
- X représente la variable aléatoire liée à l'expérience "mon poids": les valeurs possibles de X se trouvent entre 0 et $+\infty$.
- X représente la variable aléatoire liée à l'expérience "Note moyenne en M2": les valeurs possibles de X sont entre 0 et 20.

LOI DE PROBABILITÉ

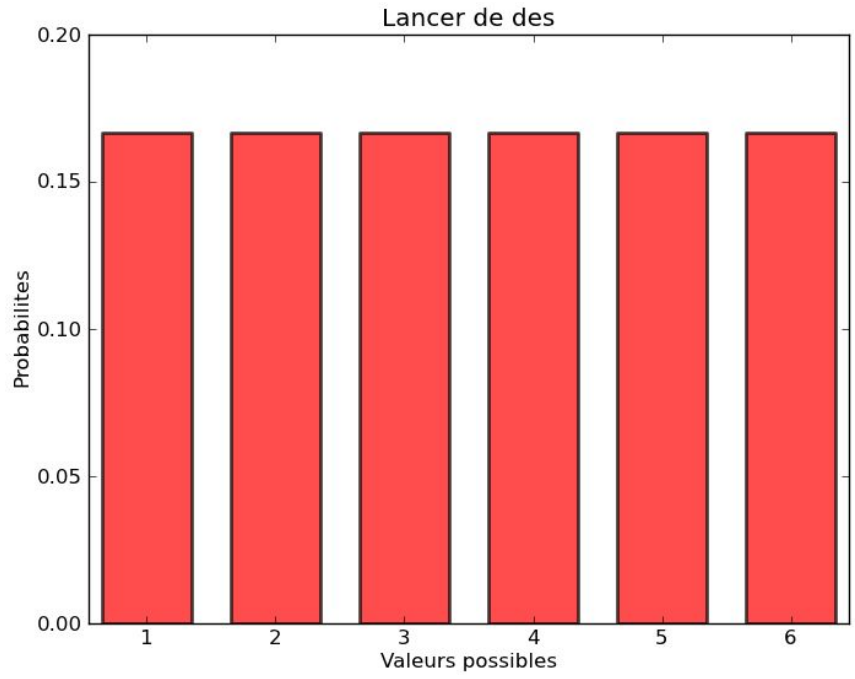
Définition (pas mathématique du tout!)

Une loi (ou distribution) de probabilité définit le comportement d'une variable aléatoire.

Autrement dit, la loi définit les probabilités qu'une variable aléatoire prenne une valeur ou un ensemble de valeurs. Attention : elle ne dit pas quelle valeur va être prise !

EX 1: LOI UNIFORME DISCRETE

X représente la variable aléatoire liée à l'expérience **"lancer de dé"**: X prend ses valeurs entre 1 et 6.



$P(X=1) = P(X=2) = P(X=3)$
$= P(X=4) = P(X=5) = P(X=6)$
$= \mathbf{1/6}$. La somme vaut **1**.

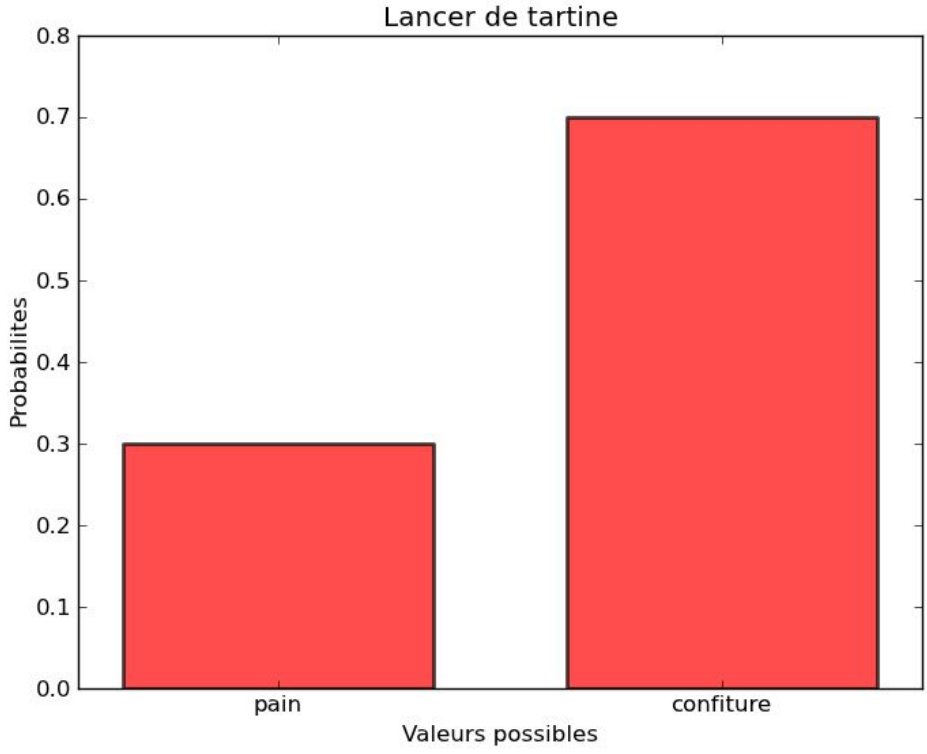
Loi loi est dite **uniforme**

Car toutes les probabilités sont les mêmes.

On parle aussi d'**équiprobabilité**.

EX 2: LOI DE BERNOULLI

X représente la variable aléatoire liée à l'expérience **"lancer de tartine"**: X prend les valeurs 0 ou 1.



P(X=côté pain) = 0.3

P(X=côté confiture) = 0.7

La somme vaut **1**.

Le **paramètre** de la loi est **p**. Ici, p = 0.7. Dans le cas d'un pile ou face avec une pièce équilibrée, p = ½ = 0.5.

EX 3: LOI UNIFORME CONTINUE

X représente la variable aléatoire liée à l'expérience **"heure d'arrivée d'un ami"**: X prend ses valeurs entre 15h et 17h.

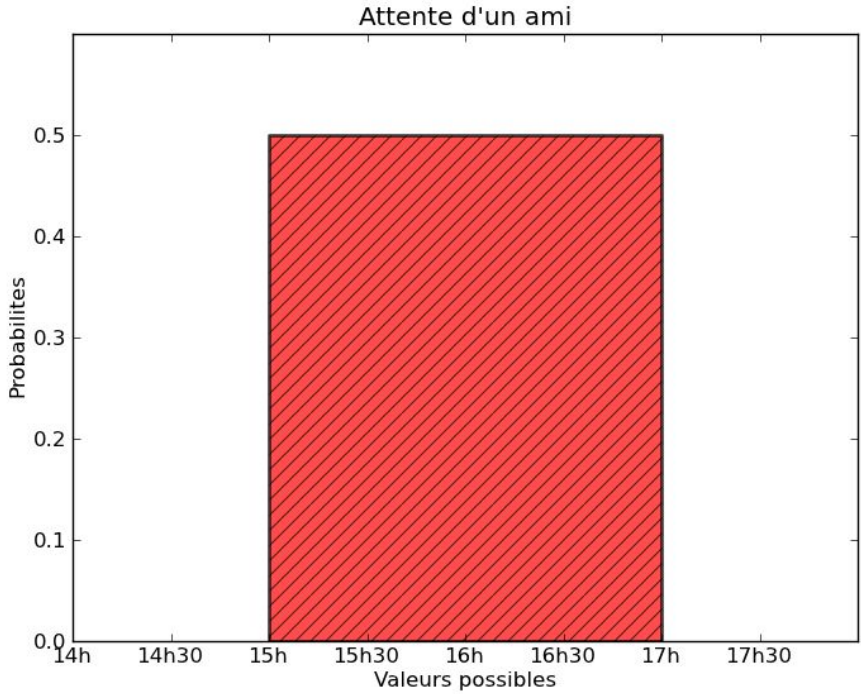
$P(X = 15:00) = 0$

$P(X = 15:30) = 0$

$P(X \in [15:10, 15:40])$

$= 1/4 = 0.25$

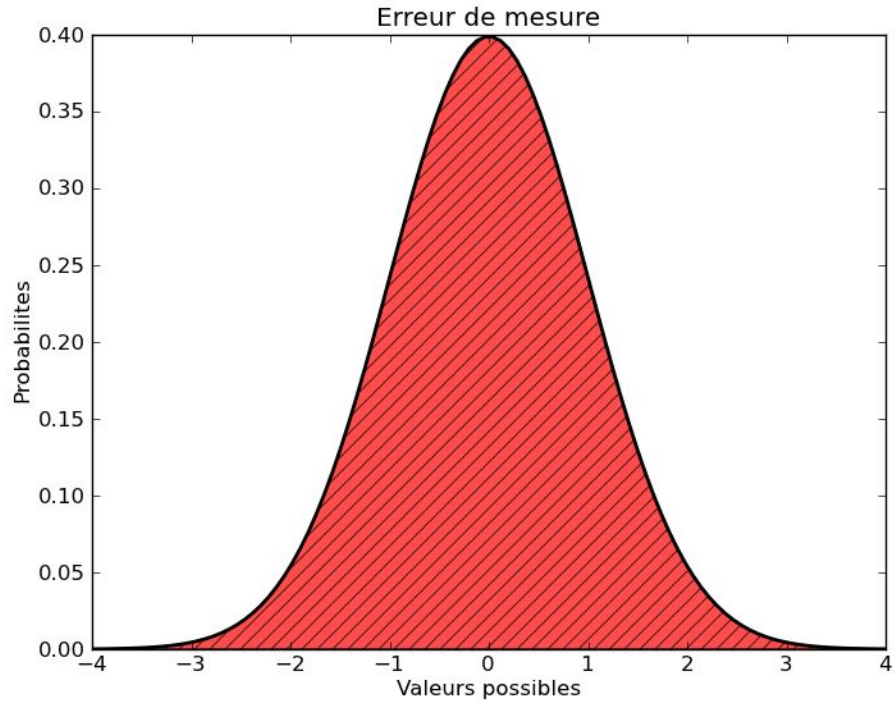
$P(X \geq 16:15) = 3/8 = 0.375$



La "somme" des probabilités vaut 1. Ici, il s'agit de l' aire sous la courbe.

EX 3: LOI UNIFORME CONTINUE

X représente la variable aléatoire liée à l'expérience **"erreur de mesure"**: X prend ses valeurs entre -∞ et +∞.



$P(X = 0) = 0$

$P(X = 3) = 0$

$P(X \in [-2, 2]) \approx 0.95$

$P(X > 2.33) = 0.01$

"Somme" des proba = 1 (aire sous la courbe).

Ici, les deux **paramètres** de la loi sont : moyenne = 0, écart-type = 1 (Nous y reviendrons)

MOMENTS D'UNE V.A.

L'**espérance** correspond à la moyenne *théorique* d'une variable aléatoire. On la note **E(X)**.

$$E(X) = \sum_{i \in valeurs\ possibles} P(X = i) * i$$

La **variance**, notée **V(X)** représente la *dispersion* d'une variable aléatoire autour de sa moyenne.
Variance grande ⇒ variable dispersée, imprévisible.
L'**écart-type** est la racine carrée de la variance.

$$Variance(X) = E((X - E(X))^2) = \sum_{i \in valeurs\ possibles} P(X = i) * (i - E(X))^2$$

MOMENTS D'UNE V.A.

(exemple au tableau)

PROBABILITÉS VS STATISTIQUES

Les **probabilités** permettent de représenter un état théorique des choses.

Les **statistiques** utilisent des données (souvent une grande quantité), expérimentales (ou simulées, dans le cadre d'un cours), et permettent de comprendre, d'**estimer les valeurs théoriques**.

ÉCHANTILLON STATISTIQUE

(X1,…,Xn) est appelé **échantillon** si les variables aléatoires X1,...,Xn sont **indépendantes** et suivent la même loi. On dit qu'elles sont **i.i.d.** : **indépendantes et identiquement distribuées.**

Exemples:

- X1...Xn sont n lancers de pile ou face. Ils sont indépendants, et ont tous la même loi de probabilité.
- X1...Xn représentent le QI de n personnes. Ces personnes sont indépendantes et leur QI suit la même distribution (assimilée à loi normale).

fabien.viger@gmail.com