

Un panorama

Gilbert Saporta, CNAM

Où n'y a t-il pas de statistique?

- Partout où il y a de l'aléatoire
- Partout où il y a des données

Sciences « fondamentales »

- Mécanique newtonienne
 - Déterministe mais incertitudes de mesure
- Métrologie
 - Répétabilité (instrument), reproductibilité (opérateur)
 - Comparaisons inter laboratoires
 - Outils
 - Estimation, analyse de variance, tests de comparaisons
 - Moindres carrés



Sciences industrielles

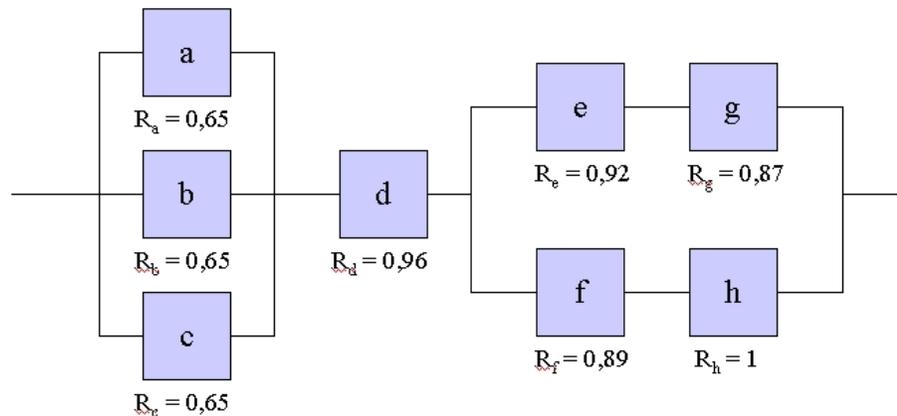
- Fiabilité des composants et des systèmes

- Lois de survie

- Sans vieillissement: loi exponentielle $P(T > t) = e^{-\frac{t}{\lambda}}$

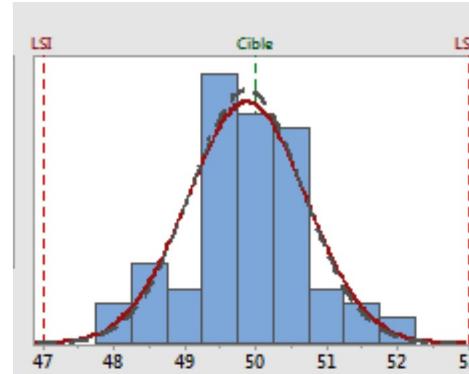
- Avec vieillissement: loi de Weibull $P(T > t) = e^{-\left(\frac{t}{\lambda}\right)^k}$

- Montages séries et parallèles

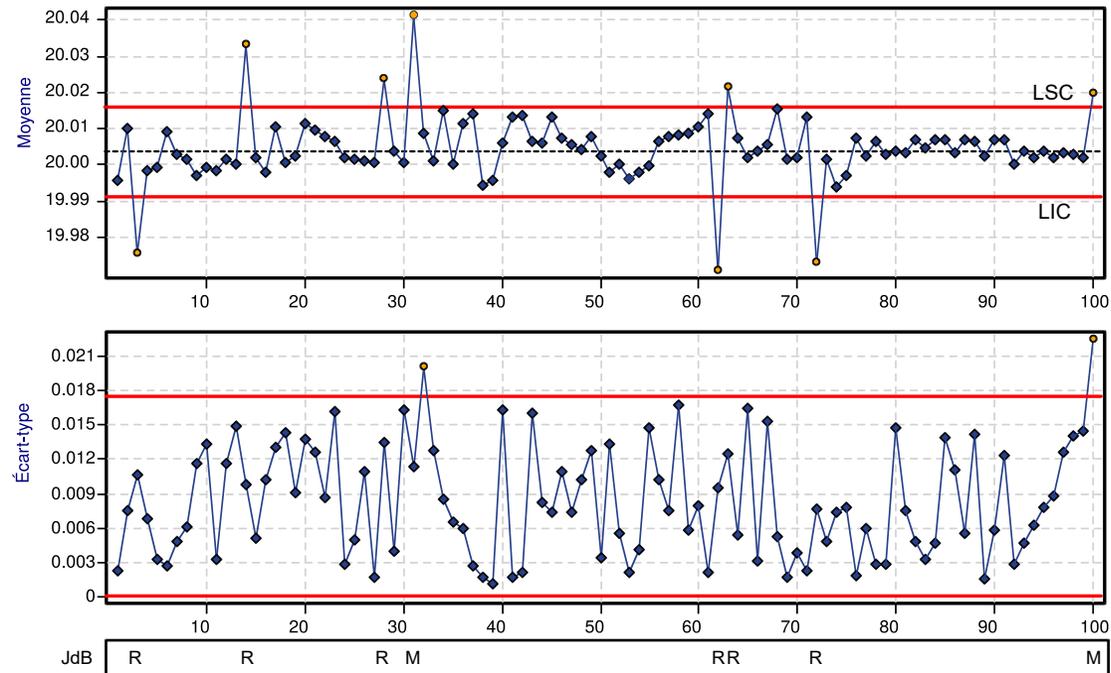


- Contrôle statistique des processus

- Aptitude ou **capabilité**



- **Cartes de contrôle**



Cosmétique

M. CHAVENT, C. GUINOT, Y. LECHEVALLIER, M. TENENHAUS

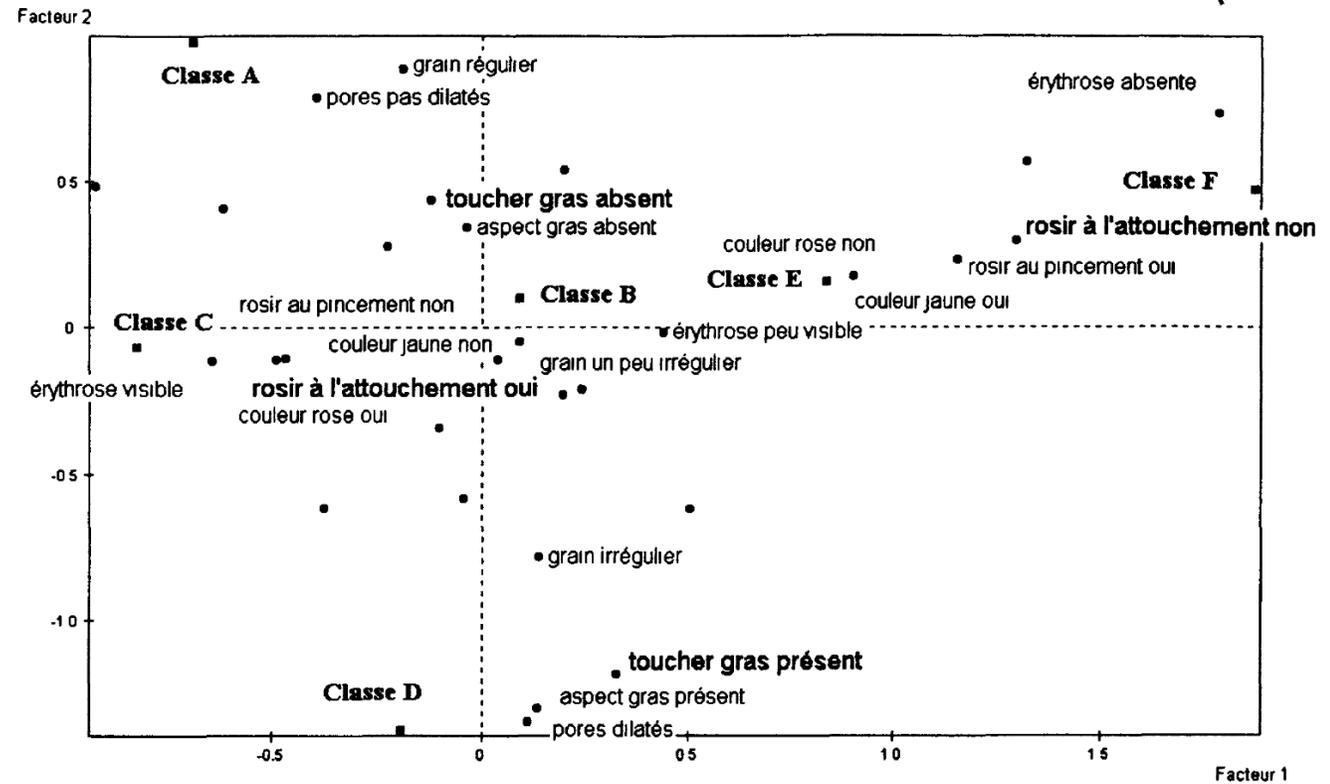
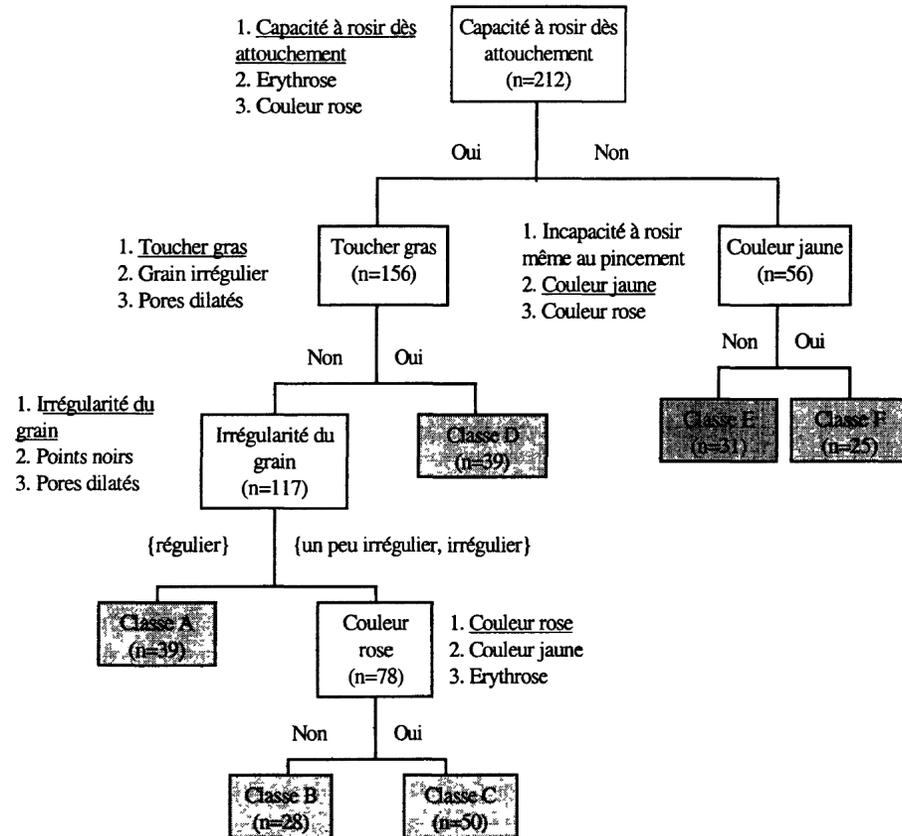
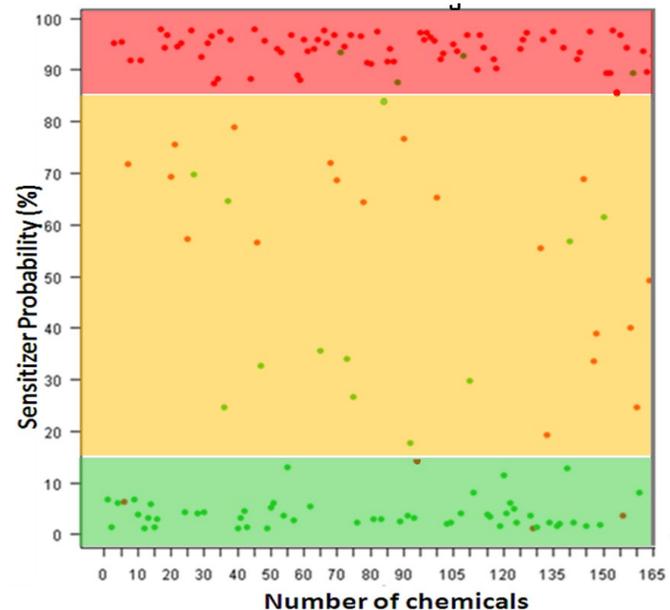


FIGURE 2
 Visualisation simultanée des modalités et des centres de gravité des 6 classes

- Modélisation des réactions cutanées

- Développement de stratégies alternatives suite à l'interdiction des expériences sur les animaux
- Un exemple: recherche de prédicteurs efficaces combinant 35 mesures de laboratoires



80.4 % sensibilité
85.5 % spécificité

(étude menée avec L'Oréal en 2012)

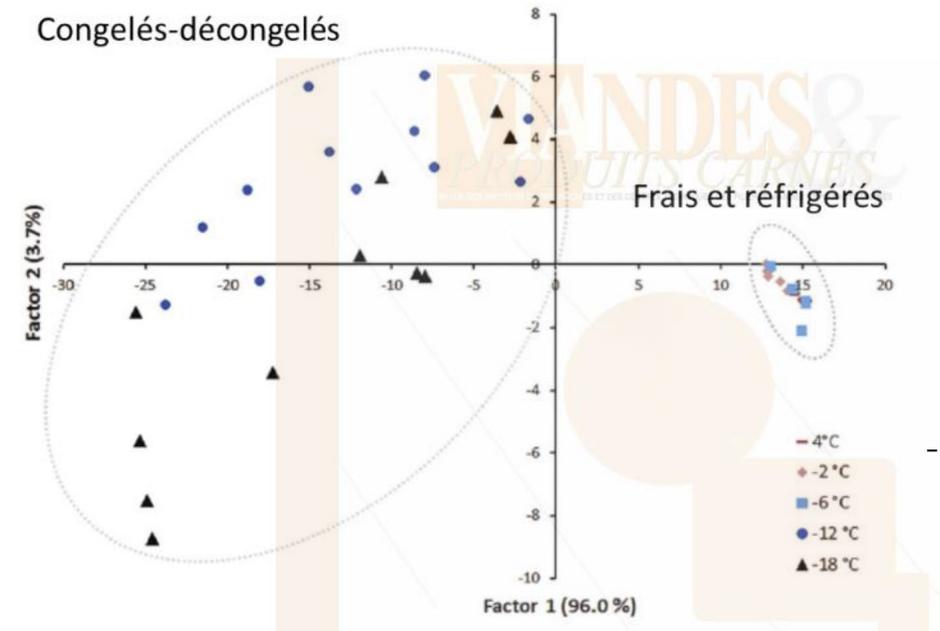
Classification supervisée

Chimiométrie

- « une branche de la chimie analytique qui utilise les outils mathématiques et statistiques dans le but d'obtenir le maximum d'informations à partir des mesures réalisées en laboratoire ou sur des procédés industriels. »
- Ex. Détection de fraudes, authentification de produits

visualisation, régressions, classifications

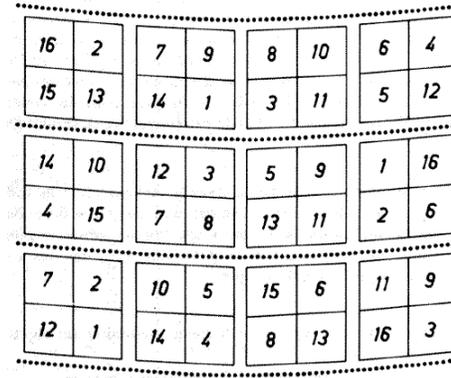
Figure 4 : Discrimination de filets de poulets frais réfrigérés ou congelés-décongelés grâce à une analyse en composantes principales de l'activité enzymatique de β -hydroxyacyl-CoA-déshydrogénase (données centrées et normalisées de la première minute de la réaction enzymatique ; Boerrigter-Eenling *et al.*, 2017)



Baeza, Authentification des viandes de volailles
Viandes & Produits Carnés – Mai 2020

Agronomie

- Plans d'expériences

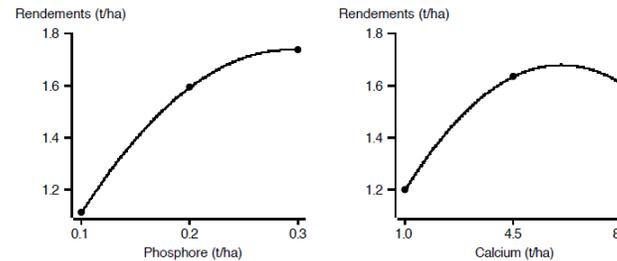


Planification



Réalisation

Développés par R.A.Fisher
autour de 1930



Interprétation

Sources de variation	Degrés de liberté	S. des carrés des écarts	Carrés moyens	F	P
Phosphore	2	6,2949	3,1474	39,9 ***	0,0000
Régression linéaire	1	5,7348	5,7348	72,8 ***	0,0000
Régression quadrat.	1	0,5601	0,5601	7,11 *	0,018
Calcium	2	3,5022	1,7511	22,2 ***	0,0000
Régression linéaire	1	2,4494	2,4494	31,1 ***	0,0001
Régression quadrat.	1	1,0528	1,0528	13,4 **	0,0023
Interaction	4	0,1525	0,0381	0,48	0,75
Blocs	2	1,9191	0,9596		
Variation résiduelle	15	1,1820	0,0788		
Totaux	25	13,0507			

Analyse

Dagnelie P. [2012]. Principes d'expérimentation : planification des expériences et analyse de leurs résultats. Edition électronique, <www.dagnelie.be>, 413 p.

Agro-alimentaire

- Evaluation sensorielle et cartographie des préférences

Juges x Produits x Attributs

Tableaux à 3 indices

Lahne, J., Abdi, H., Collins, T., & Heymann, H. (2019). [Bourbon and rye whiskeys are legally distinct but are not discriminated by sensory Descriptive Analysis.](#) *Journal of Food Science*, **84**, 629-639

11 juges, 24 produits, 24 descripteurs

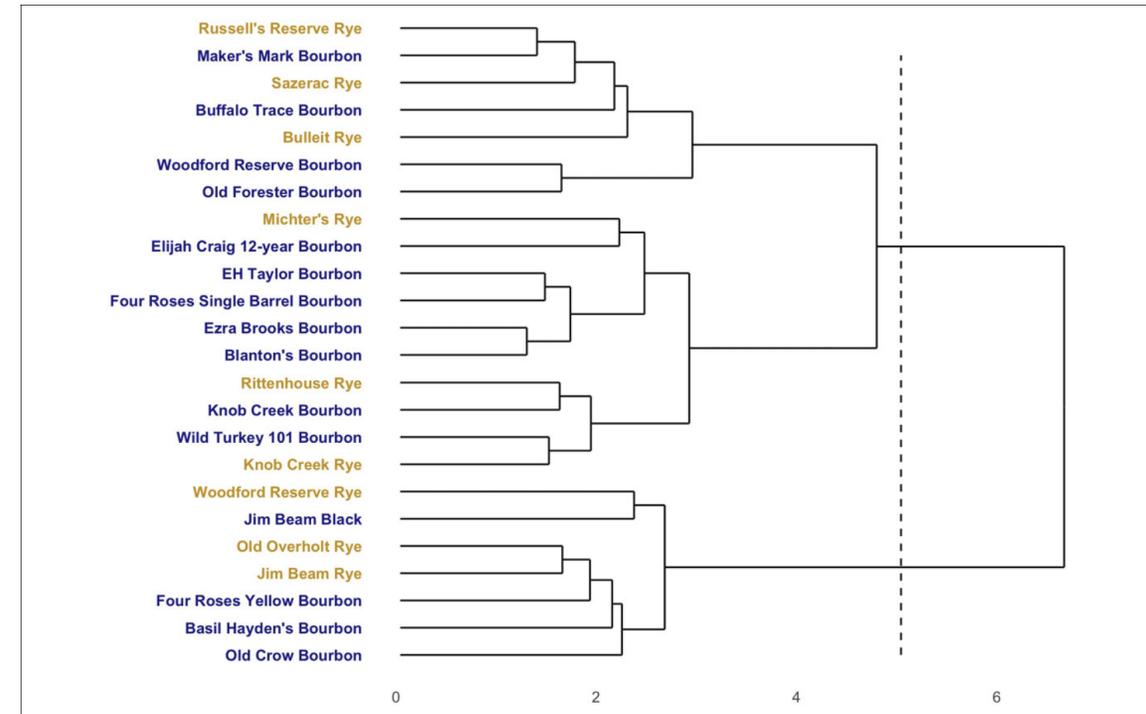
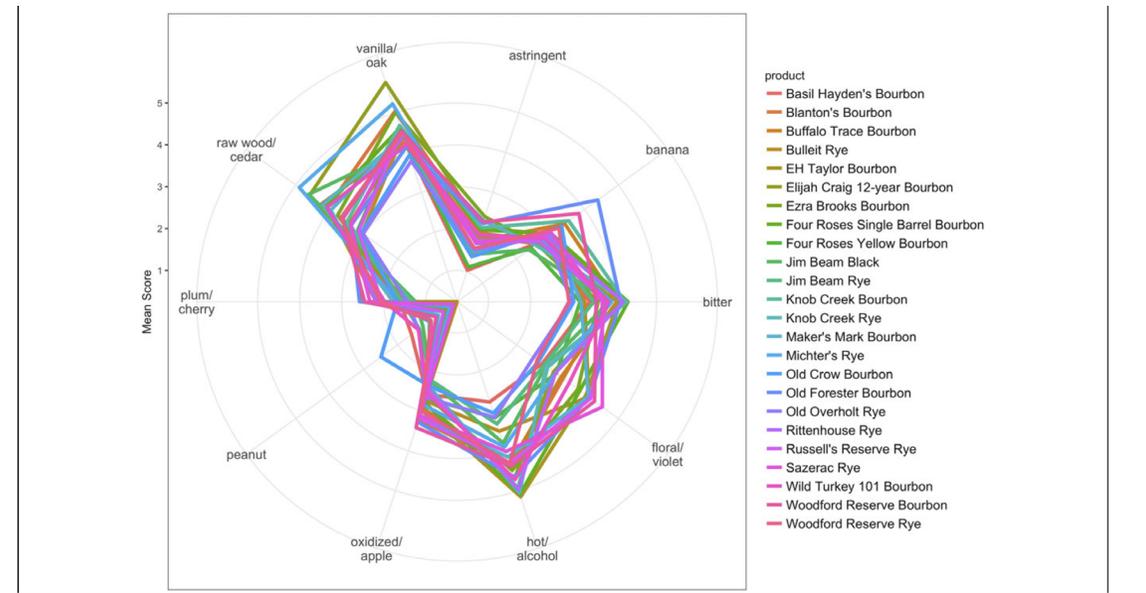


Figure 6–Hierarchical Cluster Analysis (HCA) of whiskey samples based on dissimilarities between mean sensory-attribute scores. Whiskeys are colored as in Figure 3 (bourbon is blue, rye is yellow).

Economie, démographie

- Population, emploi, revenus, entreprises, prix ...

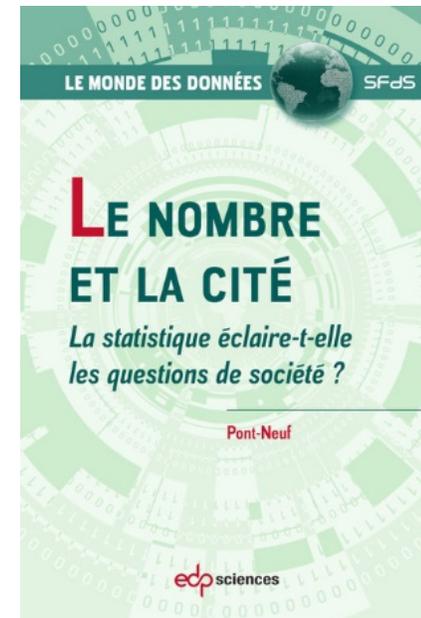


Institut national de la statistique
et des études économiques

Mesurer pour comprendre

- Modélisation, régression
- Lissage et prévisions de séries temporelle
- Visualisation

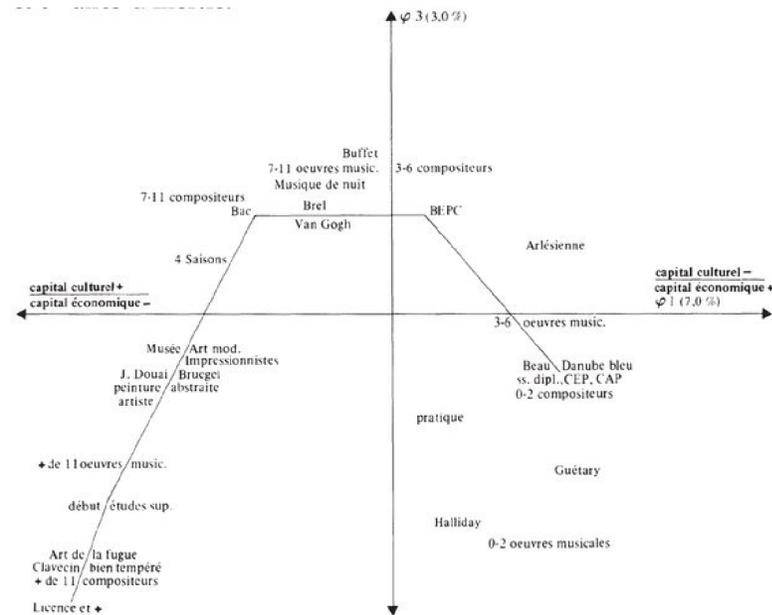
- Problèmes de société
 - Cafés de la statistique



Sociologie

- Enquêtes: tris à plat, tris croisés, analyses multivariées
- Usages extensifs de l'AFC par Pierre Bourdieu

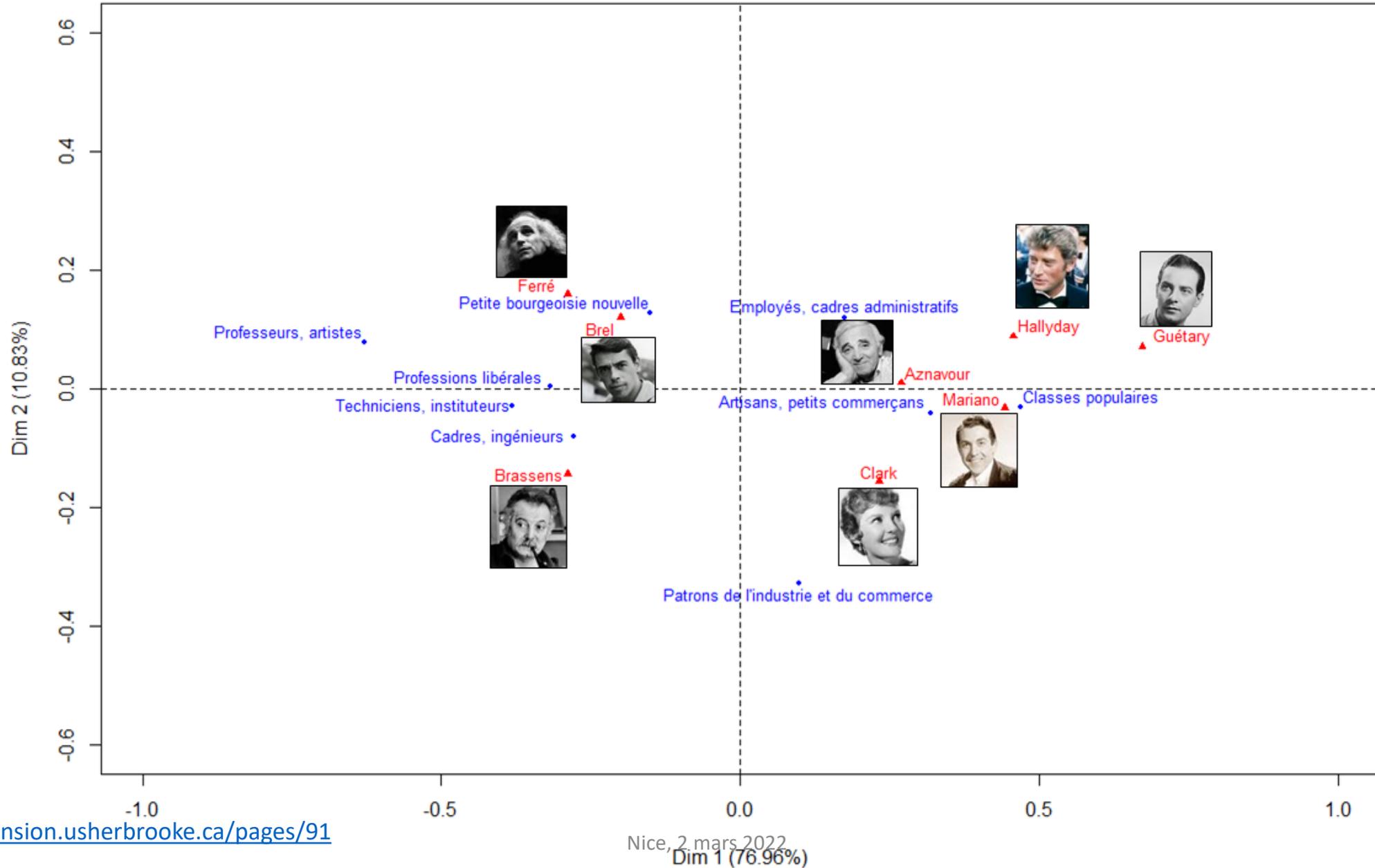
Graphique 17 – Variantes du goût petit-bourgeois.
Analyse des correspondances. Plan simplifié des 1^{er}
et 3^e axes d'inertie.



	Les chanteurs préférés							
	Guétary	Mariano	Clark	Aznavour	Hallyday	Brel	Brassens	Ferré
Classes populaires	52	31	49	87	29	40	69	33
Artisans, petits commerçants	22	20	30	36	2	31	32	20
Employés, cadres administratifs	61	35	71	135	33	138	116	86
Techniciens, instituteurs	3	4	15	15	2	43	56	27
Petite bourgeoisie nouvelle	11	13	19	35	6	49	59	51
Patrons de l'industrie et du commerce	13	10	27	34	6	23	62	10
Cadres, ingénieurs	6	16	53	65	12	98	163	85
Professions libérales	3	2	8	19	0	26	37	20
Professeurs, artistes	1	1	2	10	3	44	69	39

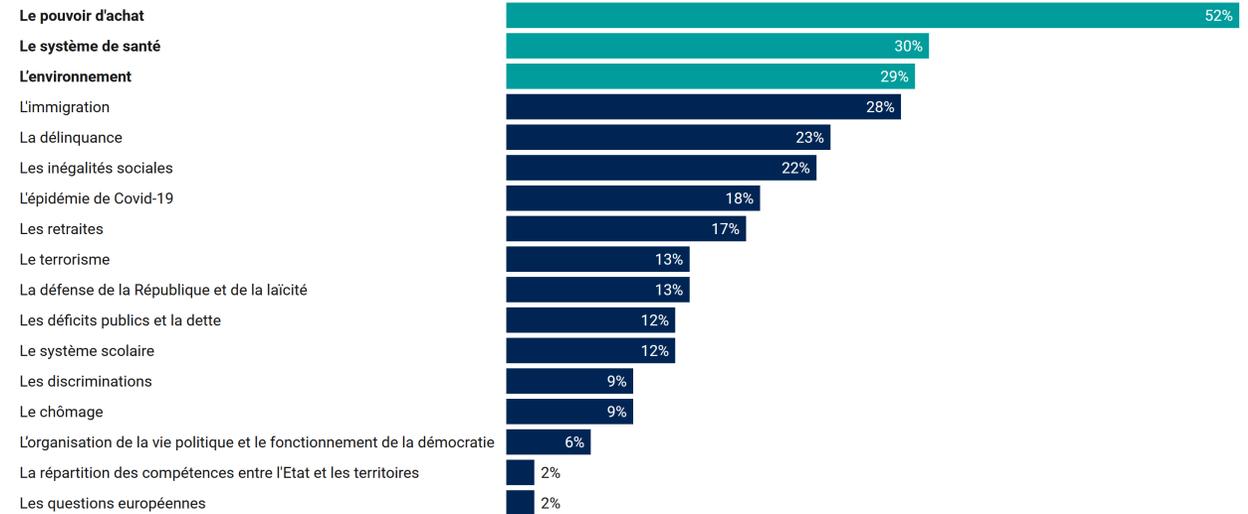
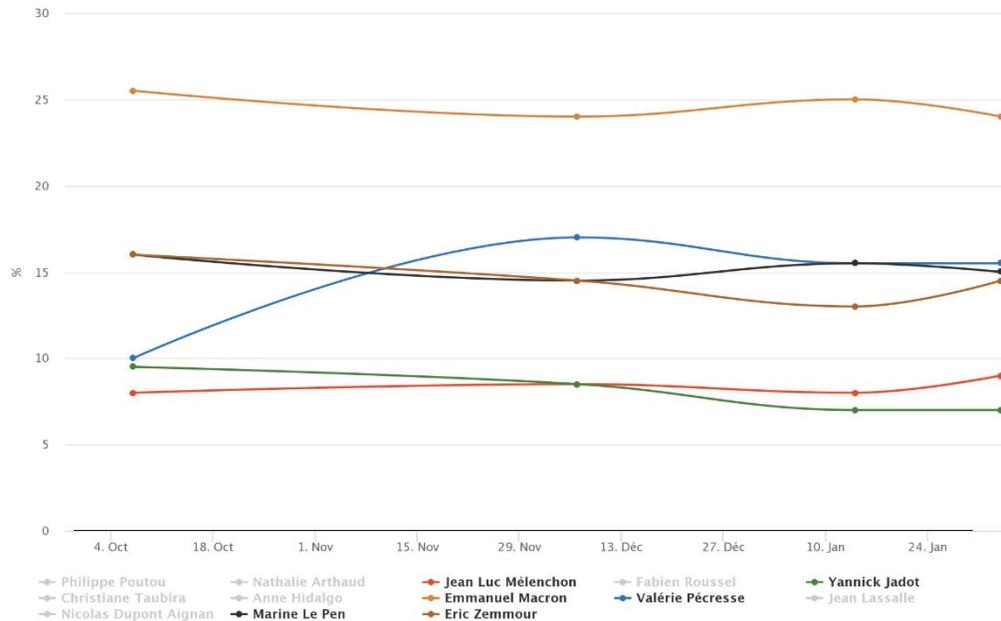
Analyse des correspondances

Exemple de Pierre Bourdieu (1979)



Opinion et sciences politiques

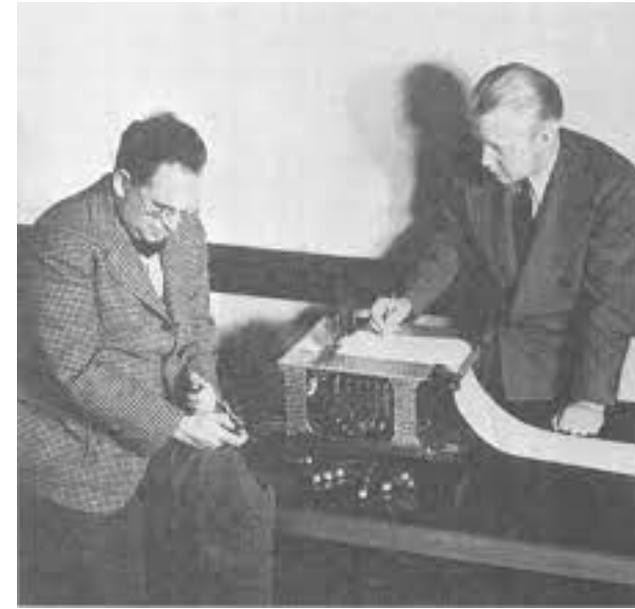
- Enquêtes par sondages



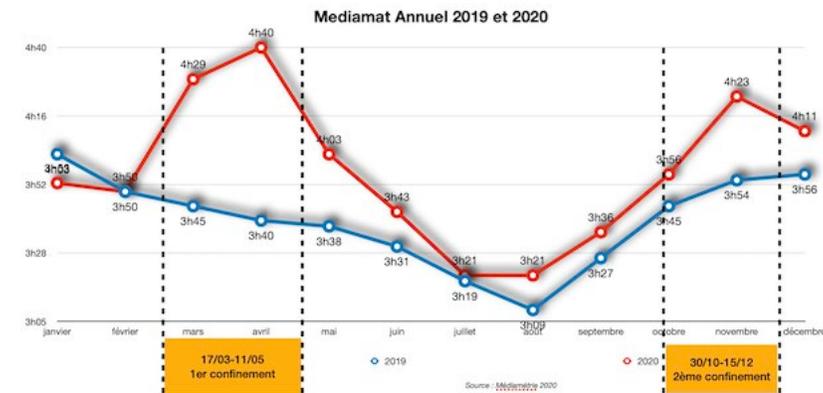
Base : ensemble de l'échantillon
 Source: Ipsos-Sopra Steria pour le Cevipof, Le Monde et la Fondation Jean Jaurès - 11 février 2022

Médias

- De l'analyseur de Lazarsfeld-Stanton (1938)



- à Médiamétrie



En ces temps de campagne présidentielle, une équipe de Sciences Po Paris et de l'Institut national de l'audiovisuel (INA) apporte une pièce originale au débat en proposant de mesurer l'inclination politique des médias audiovisuels et d'en identifier les causes. D'éventuels biais seraient-ils propres à une chaîne? Ou liés à des changements de propriétaire? Ou encore aux propres choix des journalistes?

« Nous ne sommes pas les premiers à nous intéresser aux biais politiques dans les médias, mais souvent cela est fait par l'analyse du contenu à l'aide d'algorithmes, assez bons pour des pays politiquement bipartisés comme les États-Unis, mais moins adaptés à des pays plus divers politiquement comme la France », explique Nicolas Hervé, de l'INA. « D'où notre choix de le faire en recensant les temps de parole et la couleur politique des invités dans diverses émissions télévisées », précise Julia Cagé, professeure d'économie à Sciences Po (également présidente de la Société des lecteurs du Monde).

L'équipe, complétée par Camille Urvoy, de l'université de Mannheim (Allemagne), et Moritz Hengel, doctorant à Sciences Po, a donc puisé dans les archives de l'INA pour collecter les noms de plus de 309 000 invités distincts dans plus de 6 millions d'émissions de toute nature, diffusées par quatorze chaînes de télé et huit radios entre 2002 et 2020. Plus de 67 000 hôtes invités ont également été recensés, dont 25 000 journalistes environ (les animateurs d'émission pouvant être également sportifs, artistes...). Ce corpus audiovisuel a été choisi parce que les chaînes sont soumises à des règles d'équilibre en matière de temps de parole, contrairement par exemple à la presse écrite.

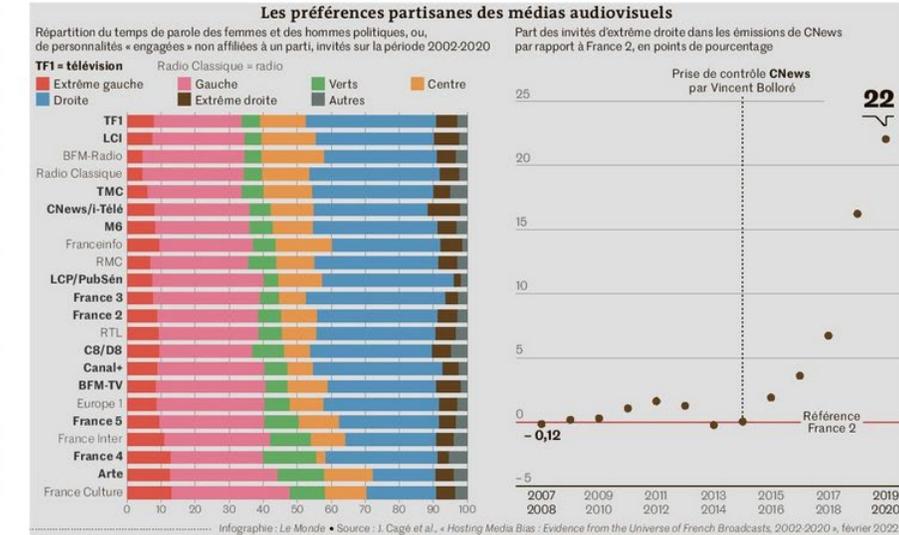
Les personnalités politiques ont ensuite été rangées dans l'une des six catégories utilisées par les spécialistes (extrême gauche, gauche, écologistes, libéraux ou centristes, droite, extrême droite). En outre, l'équipe a attribué une étiquette politique aux invités dits « engagés », qui échappent aux comptes des régulateurs comme ceux du Conseil supérieur de l'audiovisuel (CSA, devenu l'autorité de régulation de la communication audiovisuelle et numérique, en janvier) : 21 300 invités ont ainsi été classés politiquement en fonction de leur soutien à des politiciens par des textes, des participations à des universités d'été ou à des cercles de réflexion.

Déséquilibres nets

Au bilan, les enseignements sont nombreux au fil des cinquante-trois pages, et autant en annexes, du préprint mis en ligne le 16 février. D'abord, les politiques représentent environ 25 % de tous les invités de ces émissions, et les personnes « engagées » 2,6 %.

Ensuite, globalement, il existe des déséquilibres nets concernant les temps de parole. L'extrême gauche, l'extrême droite et les écologistes sont sous-représentés par rapport aux autres catégories, avec moins de 10 % des temps de parole pour chacun sur la période. Le déséquilibre est accentué si l'on tient compte de leur popularité mesurée dans les sondages. La gauche et la droite – auxquelles s'ajoute La République en marche (LRM), parti classé parmi les centristes, sur la dernière période – ne bénéficient pas des mêmes temps de parole, l'un supplantant l'autre au gré des alternances des quinquennats, y compris en prenant en compte la règle du temps de parole « obligatoire » dévolu au gouvernement (un tiers du total).

Autre conclusion, les médias n'ont pas la même orientation et sont soumis à des biais en faveur de l'un ou l'autre camp, comme le montre une comparaison des parts d'invités de chaque étiquette, avec celles de France 2, utilisées comme référence. Ainsi, LCP-Public Sénat a environ



BIG DATA

Les biais politiques dans l'audiovisuel passés à la moulinette statistique

L'analyse de millions d'émissions diffusées entre 2002 et 2020 affine une typologie des préférences partisanes exercées par les chaînes et certains journalistes

LES EXTRÊMES ET LES VERTS SONT SOUS-REPRÉSENTÉS PAR RAPPORT AUX AUTRES, AVEC MOINS DE 10 % DES TEMPS DE PAROLE POUR CHACUN SUR LA PÉRIODE

12 points de pourcentage de plus d'invités de droite comme de gauche que France 2. France Inter « aime » autant la gauche que France 2, mais moins la droite (9 points de pourcentage en moins). Arte est la chaîne qui a le plus grand pourcentage d'invités classés extrême gauche ou écologistes (presque 5 points de plus que France 2). C'est sur les radios (France Culture, RTL, Radio Classique), en plus d'Arte, que l'orientation est la plus marquée pour l'extrême gauche. Mais Radio Classique, RMC et BFM-Radio (devenue BFM Business en 2010) ont une préférence plus prononcée pour LRM que les autres médias audiovisuels sur la période récente. Les chaînes d'info en continu se distinguent avec BFM-TV, plutôt plus que les autres pour la gauche et la droite. LCP pour LRM, et CNews (anciennement i-Télé) pour l'extrême droite (5 points de plus que France 2). Pour CNews, l'étude montre l'évolution depuis sa prise de contrôle par Vincent Bolloré

en 2014. Alors qu'elle était orientée comme France 2 avant cette période, l'écart grandit jusqu'à plus de 20 points de pourcentage en faveur de l'extrême droite sur la période 2019-2020. « Nos calculs montrent qu'il y a un effet causal entre le rachat et ce biais, qu'on ne trouve pas par exemple chez BFM, pourtant rachetée [par Patrick Drahi] durant la période étudiée », relève Julia Cagé.

« Faiblesses de la régulation »

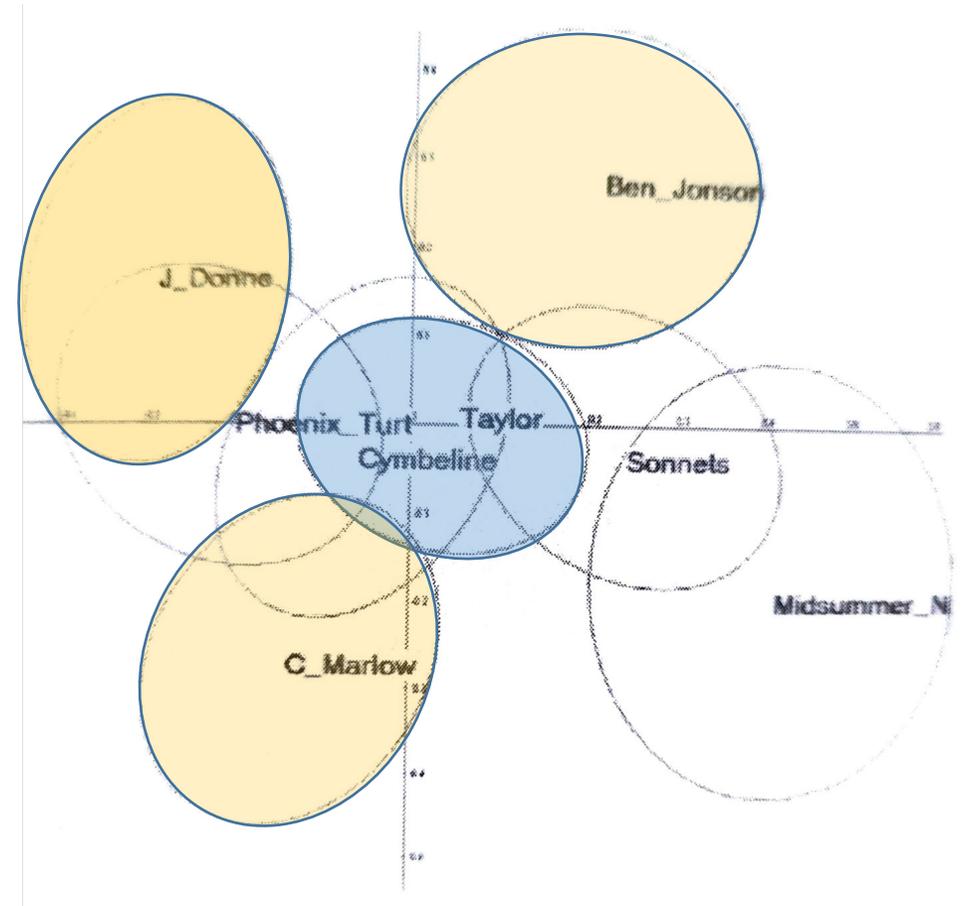
Un graphique montre ainsi qu'il existe des journalistes et des animateurs qui – une fois prises en compte les caractéristiques des médias pour lesquels ils travaillent, par exemple les préférences de l'actionnaire ou des téléspectateurs – reçoivent nettement plus d'invités de gauche que de droite, ou l'inverse, mais, surtout, que certains ont une propension plus grande à privilégier les invités d'extrême droite que d'autres. « Nous ne pouvons pas publier les noms, mais nous les connaissons et n'avons pas été surpris avec ce que nous connaissons du monde médiatique », précise Julia Cagé.

L'équipe a l'intention de poursuivre son travail pour en corriger certaines limites, en estimant mieux les temps de parole (obtenus pour l'instant en divisant la durée de l'émission par le nombre d'invités) et en analysant le contenu des interventions pour mieux apprécier les thèmes discutés et leur coloration politique éventuelle. « Ce travail montre aussi les faiblesses de la régulation de ces chaînes, notamment parce qu'elle ne fait pas un suivi systématique aussi fin des invités et ne comptabilise pas comme "politique" les invités non membres d'un parti », souligne Sylvain Parasie. « Globalement, la prise en compte de ces personnalités engagées ne change pas nos principales conclusions, mais elle amplifie certains effets, notamment pour les chaînes d'information en continu » ajoute Julia Cagé.

DAVID LAROUSSE

Littérature et langage

- Attribution d'un poème à Shakespeare (G.Taylor, 1985) 429 mots
 - Comparé à des œuvres de John Donne, Ben Jonson et Christopher Marlowe
 - À partir des fréquences d'emploi des mots dans l'œuvre de Shakespeare
 - Analyse multivariée, domaines de confiance



Lebart, L., Pincemin, B., & Poudat, C. (2019). *Analyse des données textuelles*. PUQ.

Gestion

- Etudes de marché et conception de nouveaux produits
- Connaissance client, fidélisation
 - Risque de résiliation
- Notation des risques (crédit bancaire)

- Plans d'expériences, régression logistique

If these were the options available to you when buying a new compact photo camera, which would you choose?

Task (6 of 12)

Brand			
Resolution	18 MP	8 MP	5 MP
Optical zoom	10x	4x	8x
Battery life	<200 photos	300-400 photos	>400 photos
Image stabilizer	Yes	Yes	No
Price	€ 299	€ 149	€ 249

Assurances, actuariat

- Analyse des risques individuels et collectifs
 - Prime pure ou **espérance** de coût
 - Tables de mortalité
- Catastrophes naturelles et **valeurs extrêmes**
 - Processus de **Poisson** pour les événements
 - Intensité X
 - Loi de la plus grande de n valeurs

$$P\left(\sup(X_1, \dots, X_n) < x\right) = \left(P(X < x)\right)^n$$

Médecine, pharmacie

- Essais cliniques
 - Essais contrôlés randomisés
 - Tests d'équivalence
 - Pharmacovigilance
- Epidémiologie
 - Modèles stochastiques de diffusion

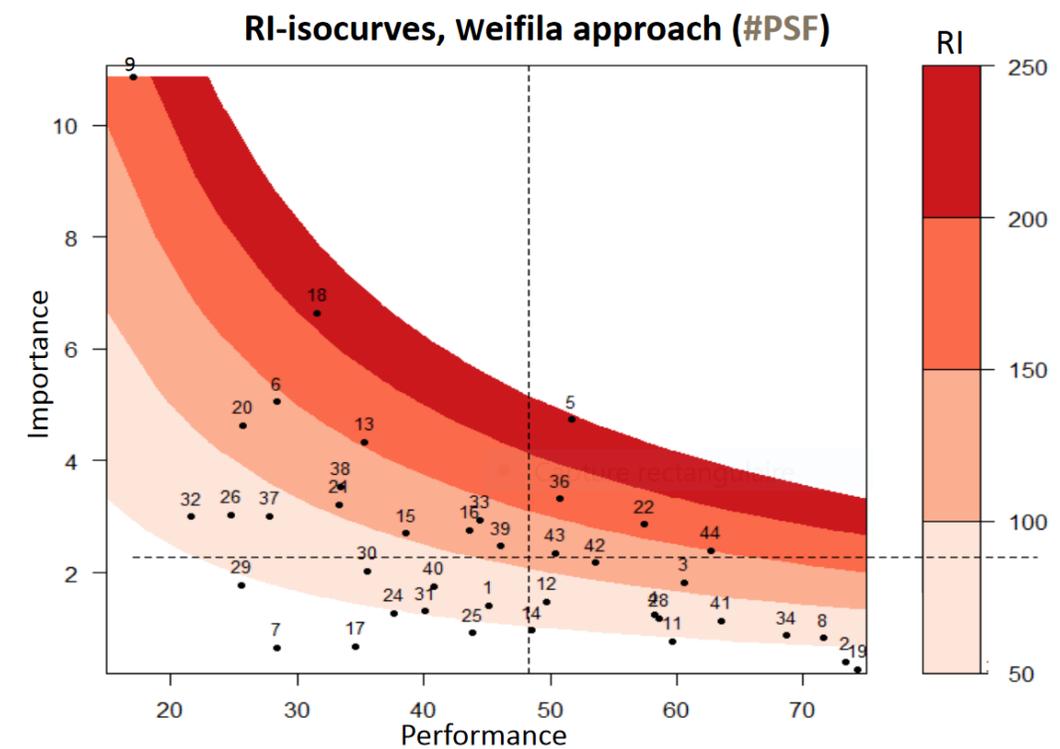
Sensibilité, spécificité, faux positifs etc.

Santé publique

Une application aux facteurs psychosociaux (FPS) liés au travail et ayant un impact sur la santé mentale. (Daouda, Temime, Saporta, Hocine 2019)

- **Échantillon** représentatif de 3200 travailleurs français.
- État de santé mentale mesuré par GHQ 28 «General Health Questionnaire with 28 items».
- 44 Facteurs psycho-sociaux
- Mesure de la performance : **prévalence** de l'exposition à chaque FPS.
- Mesure de l'importance : force de l'association entre la santé mentale et le FPS dans un modèle de **régression**

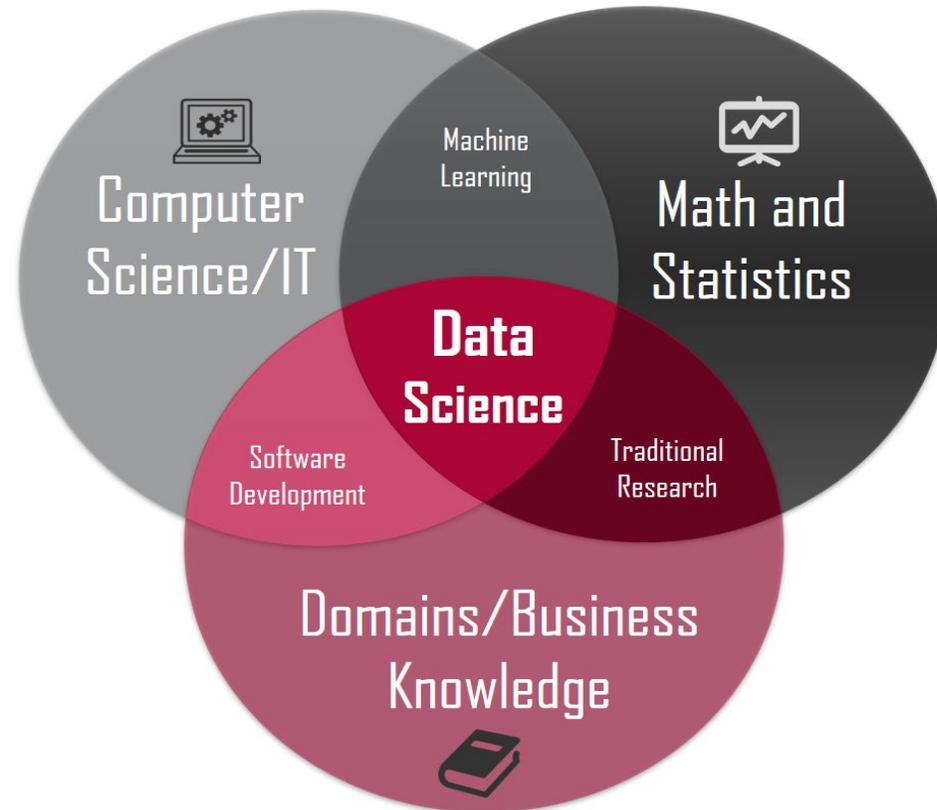
Données recueillies en 2018 par IPSOS-France pour la Fondation Pierre Deniker pour la Santé Mentale.

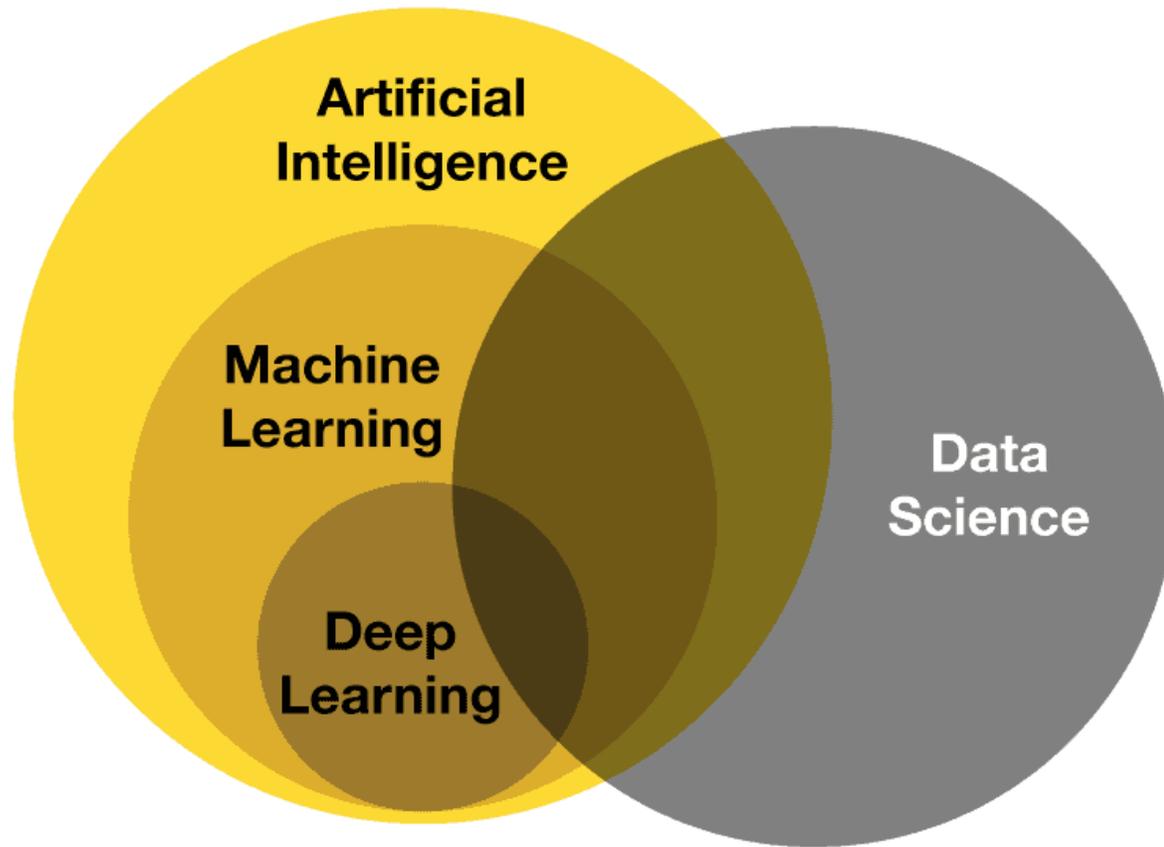


Rang	Item	Libellé
1	5	La communication et les échanges d'information au sein de mon entreprise sont satisfaisants
2	18	Je sais que je peux m'appuyer sur les gens avec qui je travaille
3	9	J'arrive à mener de front ma vie professionnelle et ma vie personnelle
4	36	Mon travail me met dans des situations émotionnellement éprouvantes
5	22	J'ai une bonne visibilité sur mes possibilités d'évolution professionnelle au sein de l'entreprise
6	13	Je me sens reconnu et apprécié dans mon travail

Informatique et Data Science

- Intelligence artificielle, apprentissage, science des données





<https://deviq.io/resources/articles/ai-vs-ml-vs-data-science/>

Risque de réidentification de données anonymisées

- Combinaison de code postal, age, sexe sur deux bases de données américaines (fichier électoral et données de santé) : risque de 0.97
- Calcul de la probabilité du risque de réidentification par une attaque à base de quasi-identifiants (3 parmi p). La méthode QaR calcule un **risque extrême**, (ex : 99 %) d'un **échantillon** de probabilités de réidentification.

AFNOR-SPEC Z90-030

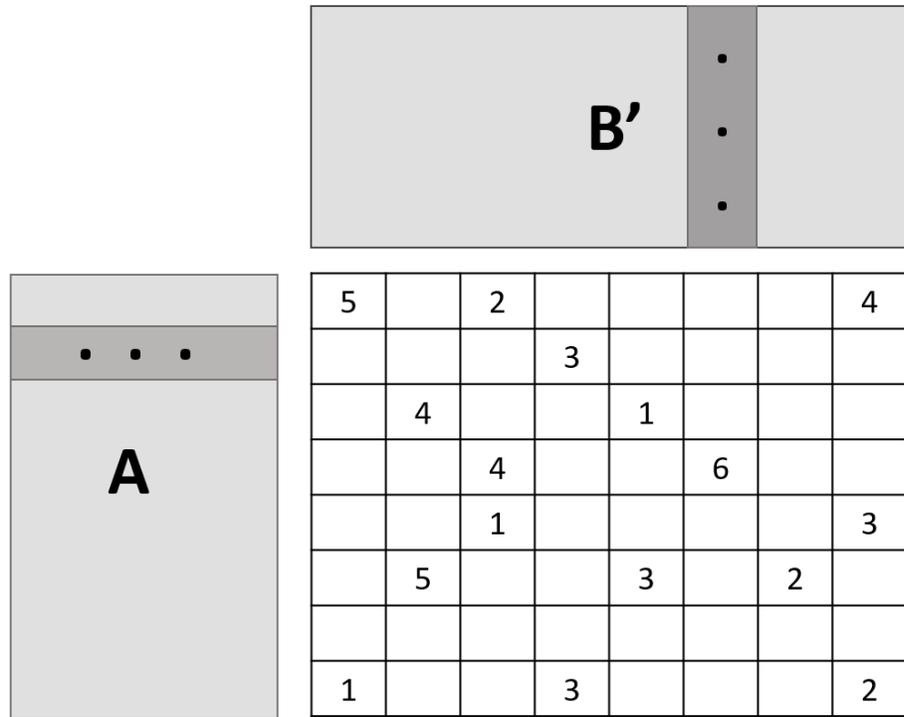
Systemes de recommandation

Recommendation systems	Categories	Recommendation systems	Categories
	E-commerce		E-commerce
	E-commerce		Music
	Music		Music
	Video		Video
	Movie		Movie
	Social network		Social network
	News		News

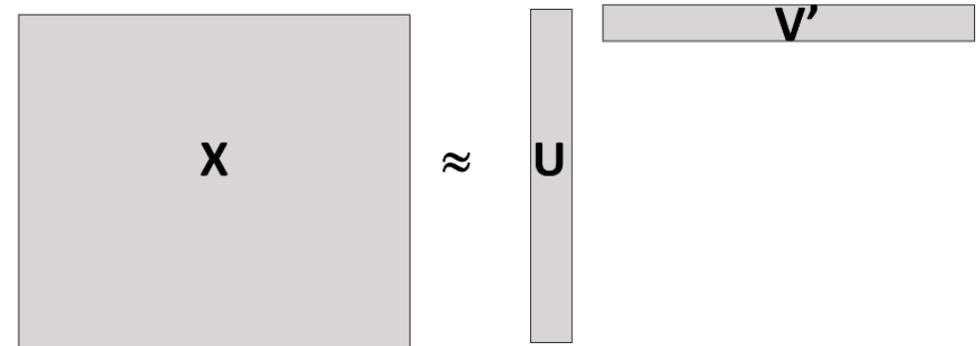
	Vampyres	Wild Wild Planet	Xanadu	Yor	Zardo
Annie	4	5	1	?	2
Bob	2	4	2	5	4
Cédric	2	1	4	2	4

Age	Sexe	Province	Genre
23	F	Québec	SF
24	M	Québec	SF
18	M	Alberta	Fantaisie

Genre	Horreur	SF	Fantaisie	SF	Fantaisie
Année	1974	1965	1980	1983	1973
Acteurs connus?	Non	Non	Oui	Non	Oui



Prédire les notes inconnues.
Complétion de matrice avec ou sans des informations sur les utilisateurs et les items.



Et bien d'autres domaines...

- Finance
- Génétique
- Biologie
- Imagerie
- Environnement et écologie
- Sport
- Droit et justice
- Transports, énergie

Pour conclure

- Au-delà des modèles :

un état d'esprit, un mode de pensée