

Statistiques, Probabilités, Analyse Spatiale

PRÉSENTATION DU MODULE

Paul Chapron¹ & Yann Ménéroux¹ & Juste Raimbault

2021-2022

¹IGN-ENSG-UGE



Introduction

«If you torture data long enough, it will confess»

Ronald Coase

«It is a capital mistake to theorize before one has data.»

Sherlock Holmes

«All models are wrong, but some are useful.»

George Box

«Data scientist (noun): Person who is better at statistics than any software engineer and better at software engineering than any statistician.»

Josh Wills

«There are three kinds of falsehoods: lies, damned lies and statistics»

Mark Twain

«Les statistiques sont une forme d'accomplissement de désir, tout comme les rêves.»

Jean Baudrillard

«Ce qui est simple est faux, ce qui est compliqué est inutile.»

Paul Valéry

«La statistique est la première des sciences inexactes.»

Édouard et Jules de Goncourt

- **Statistique** : Démarche scientifique visant à acquérir des connaissances sur l'état d'un objet difficilement perceptible sur le plan cognitif :
 - objet **massif** (population, big data...) : résumer, synthétiser, visualiser, échantillonner, compresser...
 - objet **incertain** (observations bruitées) : moyenner, borner, encadrer, compenser...

- **Statistique** : Démarche scientifique visant à acquérir des connaissances sur l'état d'un objet difficilement perceptible sur le plan cognitif :
 - objet **massif** (population, big data...) : résumer, synthétiser, visualiser, échantillonner, compresser...
 - objet **incertain** (observations bruitées) : moyenner, borner, encadrer, compenser...

En général, la statistique sert à **décrire** des états présents ou passés.

- **Probabilités** : théorie mathématique traitant de l'aléatoire, sans justification a priori (même si généralement fondée sur des besoins pratiques) et opérant dans un cadre contrôlé (information parfaite).

- **Probabilités** : théorie mathématique traitant de l'aléatoire, sans justification a priori (même si généralement fondée sur des besoins pratiques) et opérant dans un cadre contrôlé (information parfaite).

En général, les probabilités servent à **prédir** des états futurs.

- **Analyse Spatiale** : Étude de la répartition et de l'organisation d'objets localisés pour «déceler en quoi la localisation apporte un élément utile à la connaissances des objets étudiés et peut en expliquer les caractéristiques»

Pumain, Saint – Julien1997

- **Analyse Spatiale** : Étude de la répartition et de l'organisation d'objets localisés pour «déceler en quoi la localisation apporte un élément utile à la connaissances des objets étudiés et peut en expliquer les caractéristiques»

Pumain, Saint – Julien1997

Nous couvrirons principalement l'analyse spatiale **statistique** i.e. prenant en compte les coordonnées des unités statistiques et différentes distances.

Naissance au XVIIe ?



Siège de Platée (-430 av J.C.)

Naissance au XVIIe ?

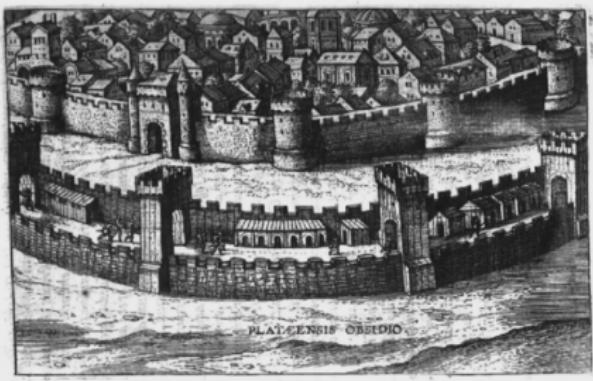


Siège de Platée (-430 av J.C.)

Naissance au XVIIe ?



Thucydide

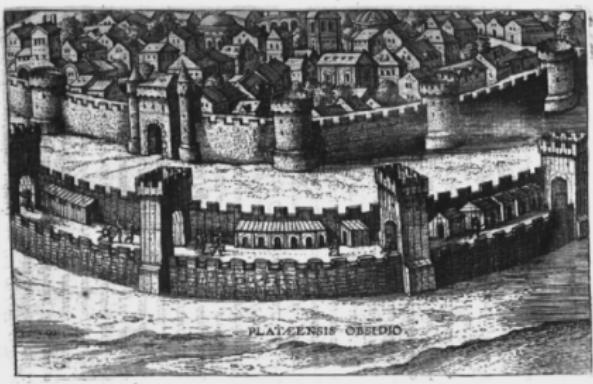


Siège de Platée (-430 av J.C.)

Naissance au XVIIe ?



Thucydide



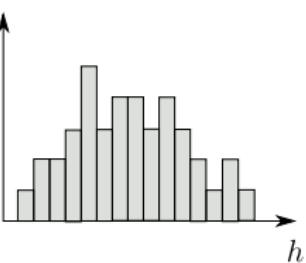
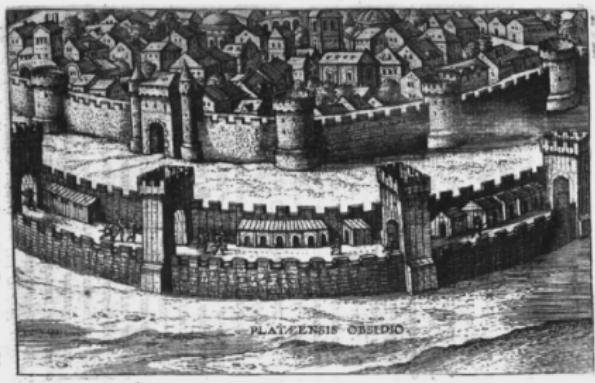
Siège de Platée (-430 av J.C.)

Naissance au XVIIe ?



Siège de Platée (-430 av J.C.)

Thucydide

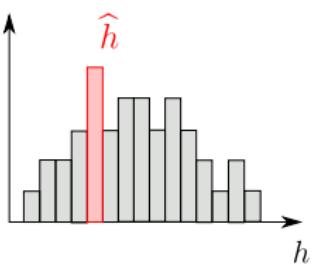
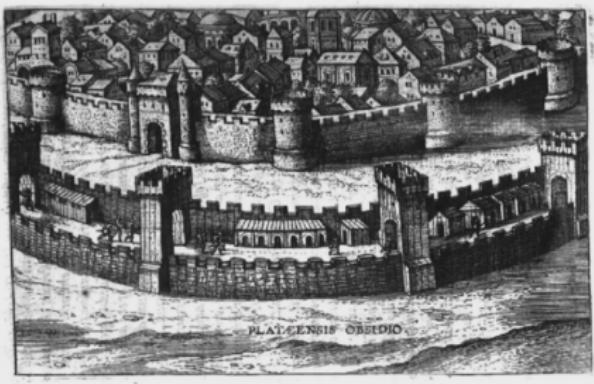


Naissance au XVIIe ?



Siège de Platée (-430 av J.C.)

Thucydide



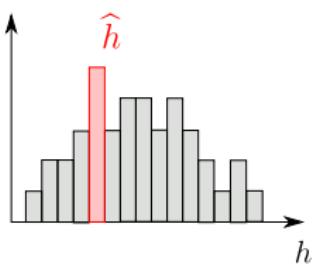
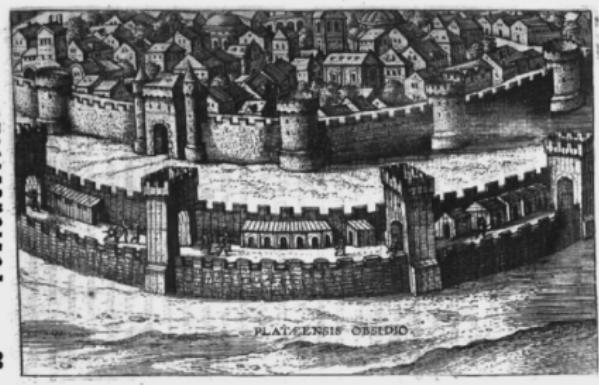
Naissance au XVIIe ?



Siège de Platée (-430 av J.C.)

$$L(h, x)$$

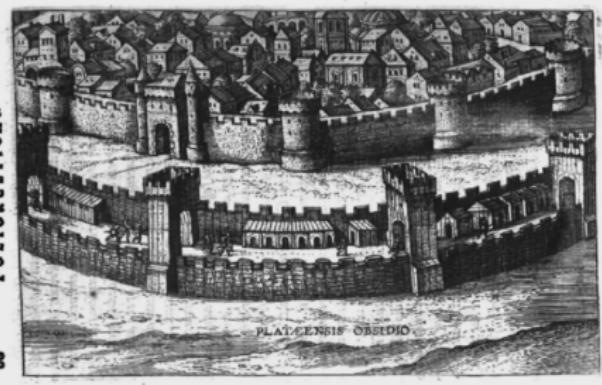
Thucydide



Naissance au XVIIe ?

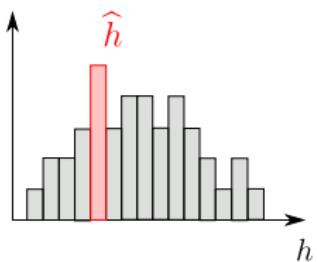


Thucydide



Siege de Platée (-430 av J.C.)

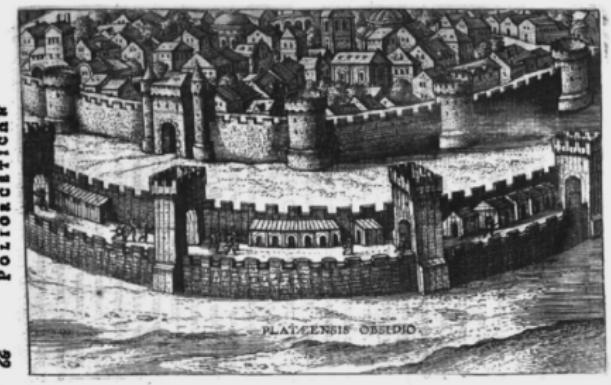
$$L(h, x) = 1 - \mathbf{1}_h(x) = \begin{cases} 0 & \text{si } x = h \\ 1 & \text{sinon.} \end{cases}$$



Naissance au XVIIe ?



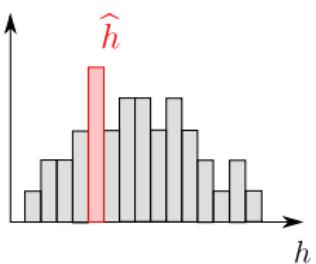
Thucydide



Siège de Platée (-430 av J.C.)

$$L(h, x) = 1 - \mathbf{1}_h(x) = \begin{cases} 0 & \text{si } x = h \\ 1 & \text{sinon.} \end{cases}$$

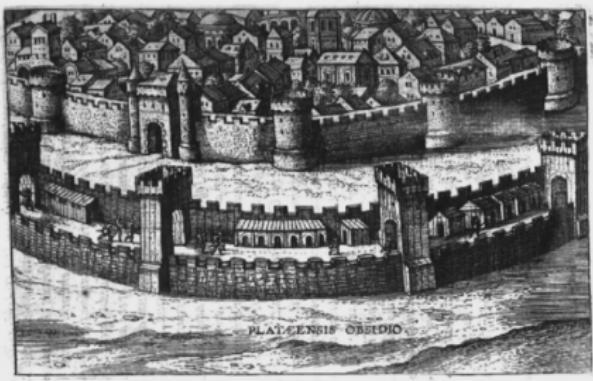
$$\hat{h} = \operatorname{argmin}_h \mathbb{E}[L(h, X)]$$



Naissance au XVIIe ?



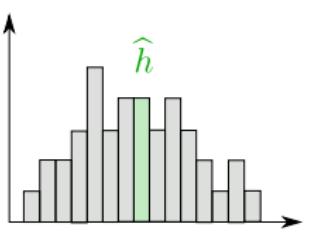
Thucydide



Siège de Platée (-430 av J.C.)

$$L(h, x) = |h - x| \quad (\text{Norme L1})$$

$$\hat{h} = \operatorname{argmin}_h \mathbb{E}[L(h, X)]$$



Naissance au XVIIe ?

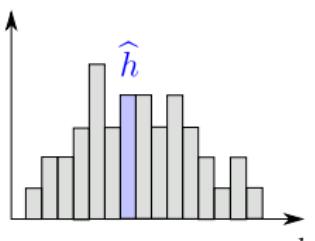
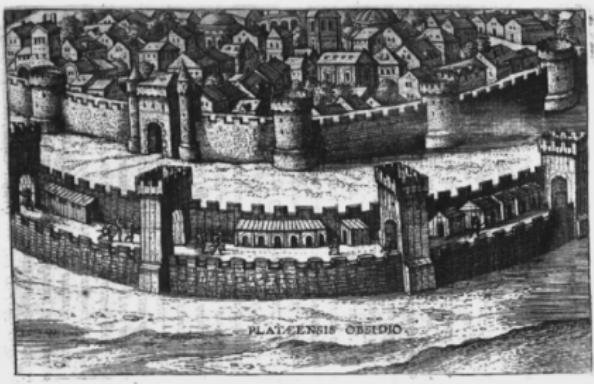


Siège de Platée (-430 av J.C.)

$$L(h, x) = (h - x)^2 \quad (\text{Norme L2})$$

$$\hat{h} = \operatorname{argmin}_h \mathbb{E}[L(h, X)]$$

Thucydide



Correspondance de Pascal et Fermat (1654) : *Deux joueurs déposent chacun une mise m pour jouer une partie en 3 manches, mais décident de se quitter après deux victoires de l'un et une victoire de l'autre. Comment devraient-ils se répartir la mise ?*

- probabilisme (XVI^{ème} siècle)
- Pas d'utilisation en dehors des jeux
- Leibniz : outil de connaissance et de raisonnement sur le réel.
La Statistique devient un instrument scientifique (1677)
- Paradoxe de Saint-Petersbourg (1713)
- Bayes : théorie des causes (1763)
- Fin XVIII^{ème} : Arithmétique politique (démographie, tables de mortalité, etc)

Le module

Support, TP, et assistance pour R et RStudio.

Programme du module

Semaine	Jour	Date	Commentaire / cours	Enseignant
semaine 43 du au 29/10/2022	lundi	24/10/2022	Introduction	Paul Chapron
	lundi	24/10/2022	Théorie des probabilités	Yann Ménervoux
	mardi	25/10/2022	Analyse univariée, bases de R	Paul Chapron
	mardi	25/10/2022	Analyse univariée, bases de R	Paul Chapron
	mercredi	26/10/2022	Analyse bivariée	Paul Chapron
	mercredi	26/10/2022	Théorie des probabilités	Yann Ménervoux
	jeudi	27/10/2022	Théorie des probabilités	Yann Ménervoux
	vendredi	28/10/2022	ACP	Paul Chapron
	vendredi	28/10/2022	Théorie des probabilités	Yann Ménervoux
	lundi	14/11/2022	Géostatistiques	Yann Ménervoux
semaine 46 du 26/11/2022	mardi	15/11/2022	Intro	Sébastien Mustière
	mardi	15/11/2022	Géostatistiques	Yann Ménervoux
	mercredi	16/11/2022	Statistiques spatiales	Juste Rimbault
	mercredi	16/11/2022	Statistiques spatiales	Juste Rimbault
	vendredi	18/11/2022	Intro	Sébastien Mustière
	lundi	21/11/2022	Statistiques spatiales	Juste Rimbault
	lundi	21/11/2022	Statistiques spatiales	Juste Rimbault
	mardi	22/11/2022	Lancement des projets	Y. Ménervoux / P. Chapron / J. Rimbault
	mercredi	23/11/2022	Apprentissage supervisé	Yann Ménervoux
	mercredi	23/11/2022	Apprentissage supervisé	Yann Ménervoux
semaine 49 du 05/12/2022	vendredi	25/11/2022	Réseaux spatiaux	Juste Rimbault
	vendredi	25/11/2022	Réseaux spatiaux	Juste Rimbault
	lundi	05/12/2022	Apprentissage non supervisé	Yann Ménervoux
	lundi	05/12/2022	Apprentissage non supervisé	Yann Ménervoux
	mardi	06/12/2022	projet en autonomie	
	mardi	06/12/2022	projet en autonomie	
	mercredi	07/12/2022	Méthodologie de la validation	Yann Ménervoux
	mercredi	07/12/2022	Méthodologie de la validation	Yann Ménervoux
	jeudi	08/12/2022	projet en autonomie	
	jeudi	08/12/2022	projet en autonomie	Y. Ménervoux / P. Chapron / J. Rimbault
10/12/2022	vendredi	09/12/2022	Deep Learning	
	vendredi	09/12/2022	Deep Learning	
	lundi	12/12/2022	projet en autonomie	
	lundi	12/12/2022	projet en autonomie	
	mardi	13/12/2022	projet en autonomie	
semaine 50 au	mardi	13/12/2022	Modèles de simulation	Juste Rimbault
	jeudi	15/12/2022	Interros "Stats / analyse spatiale / apprentissage"	
	vendredi	16/12/2022	Fin des projets	Y. Ménervoux / P. Chapron / J. Rimbault

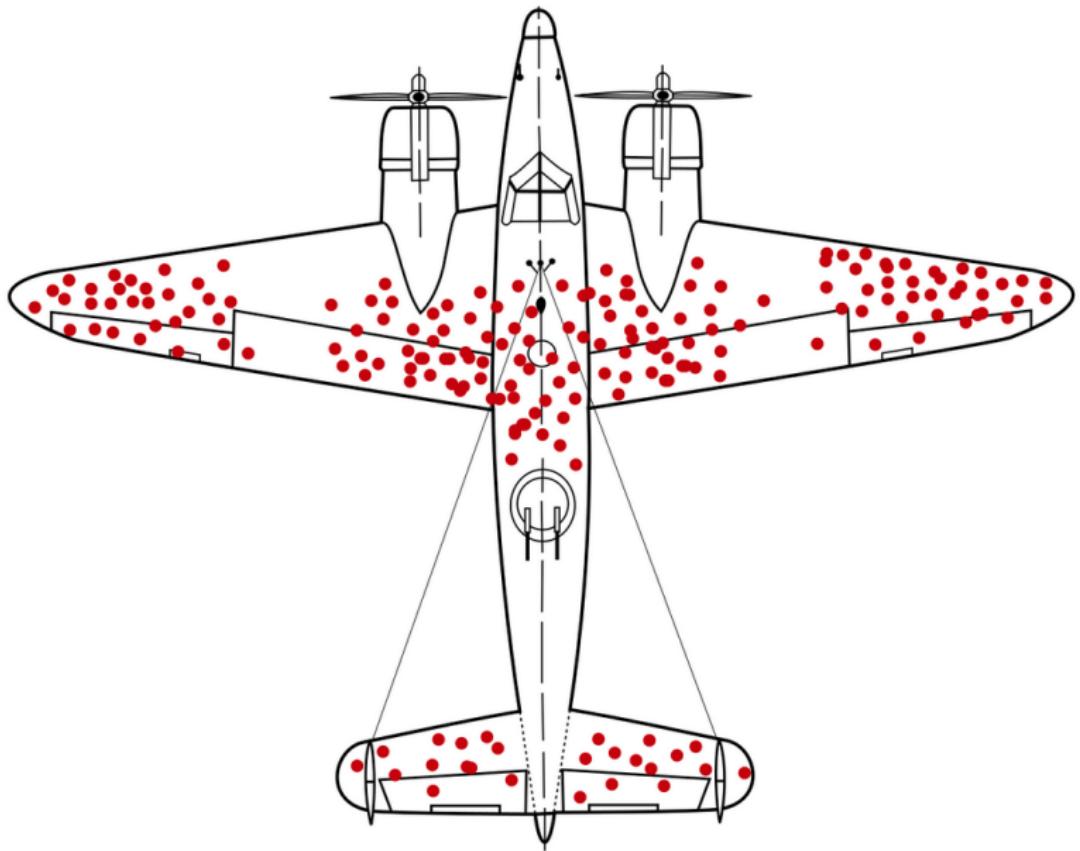
Pendant la WW2, la Royal Air Force voulait blinder ses avions contre la flotte et la DCA allemande. Blinder tout l'appareil est impossible car trop lourd.

Où blinder les avions ?

On collecte des données : la localisation impacts de shrapnels et de balles des avions rentrés à la base en Angleterre.

Tropes de la statistique

L'avion



→ il faut blinder l'arrière et les ailes !

Mais ...

→ il faut blinder l'arrière et les ailes !

Mais ... il faut prendre en compte le *biais du survivant*

→ il faut blinder l'arrière et les ailes !

Mais ... il faut prendre en compte le *biais du survivant*

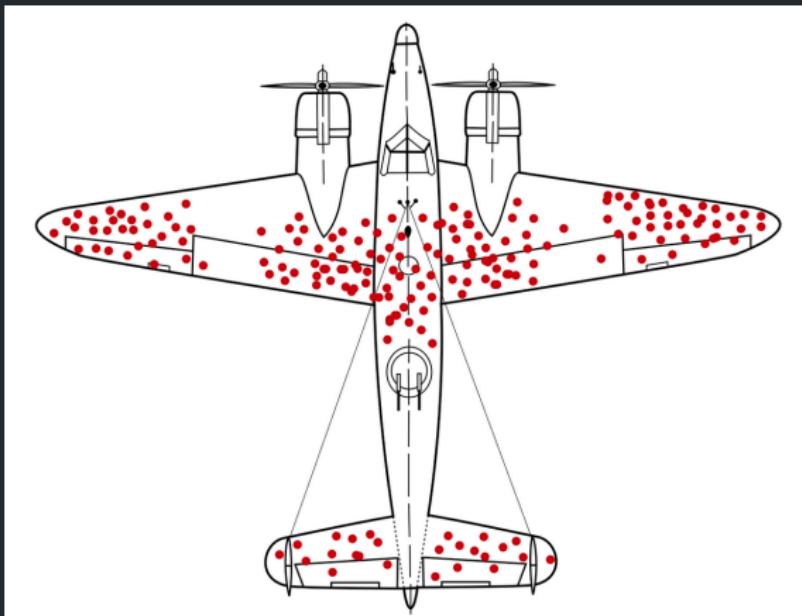
Les données proviennent d'avions qui sont rentrés à la base. Ceux qui ont été touchés gravement ont péri, leurs impacts ne sont pas dans les données.

Hypothèse : les impacts sont uniformes sur les avions, mais les avions rentrés le sont peut-être du fait que les impacts ont causé peu de dommages sur l'avion, leur permettant de rentrer.

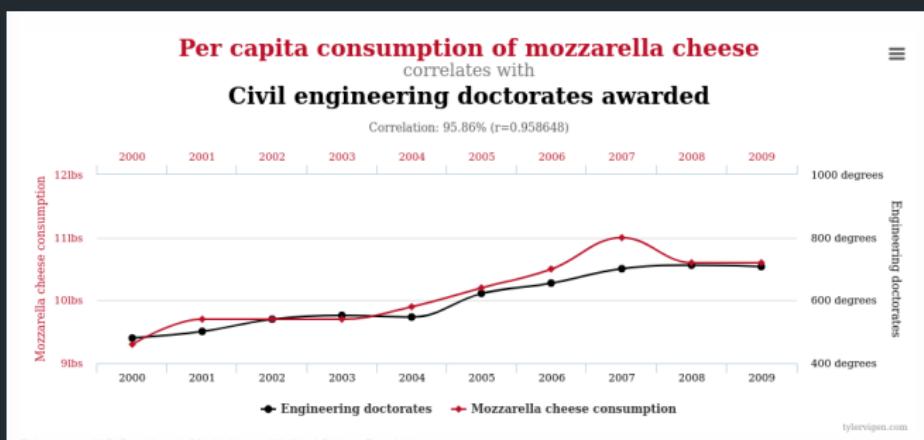
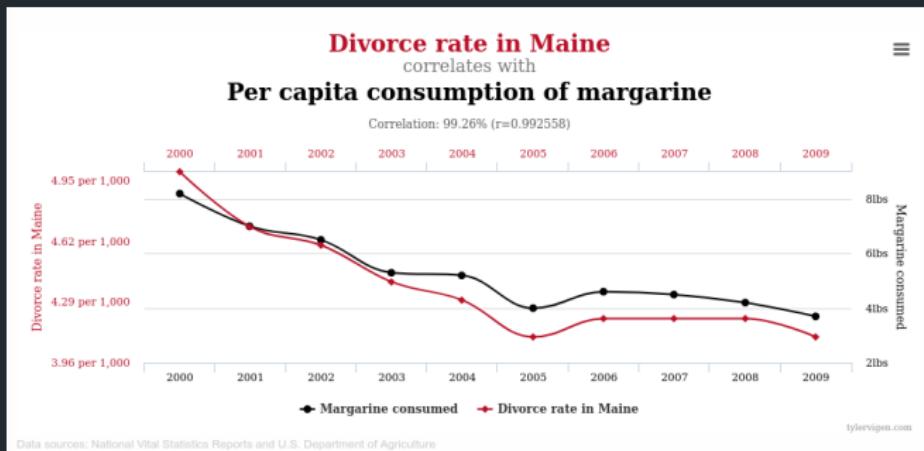
L'avion anglais de la WW2

→ il faut blinder le cockpit et les moteurs.

Ce qu'ils ont fait, avec succès (beaucoup moins de pertes).



Spurious correlations



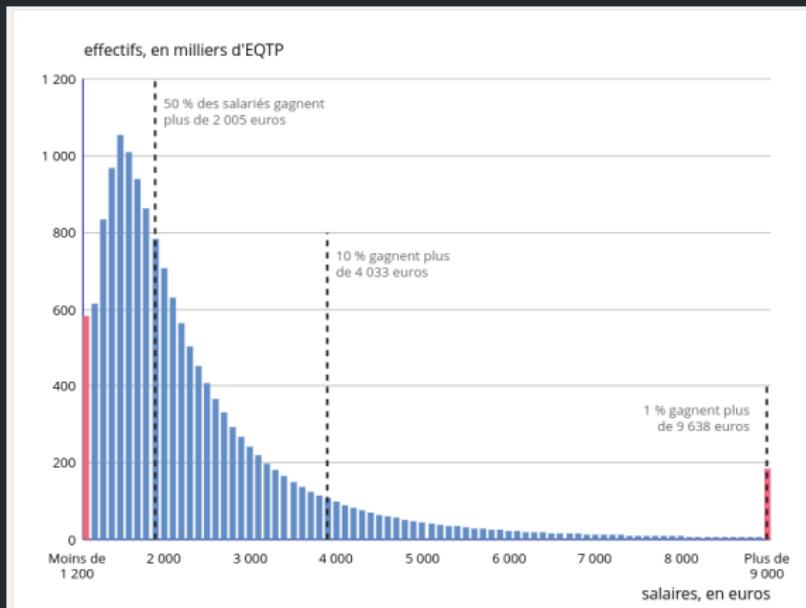
Sur les mêmes chiffres du bilan d'une entreprise en 2013 et 2014

- Mme. AAA : «tous les salaires ont baissé de 10%»
- M. BBB : «le salaire moyen a augmenté d'environ 18%»

	Ouvriers	Cadres
2013	effectif : 500	effectif : 100
	salaire : 1300	salaire : 2200
2014	effectif : 200	effectif : 400
	salaire : 1170	salaire : 1980

La fragilité de la moyenne

Salaire moyen en 2020 EQTP (hors Mayotte) : 2518€



Note : certains salaires en EQTP sont inférieurs au Smic ; ceci est en effet permis par certains statuts. Cependant, l'existence de rémunérations inférieures au Smic peut aussi provenir d'incohérences entre salaires et durées travaillées dans les déclarations administratives, qui ne peuvent être toutes redressées.

Lecture : en 2020, en EQTP, 50 % des salariés gagnent plus de 2 005 euros.

Champ : France hors Mayotte, salariés du privé et des entreprises publiques, y compris bénéficiaires de contrats aidés et de contrats de professionnalisation ; hors apprentis, stagiaires, salariés agricoles et salariés des particuliers employeurs.

Source : Insee, base Tous salariés 2020.

«Quand les distributions sont **unimodales et symétriques**, tout va bien..»

Sinon, agir avec précaution !

Représenter les données

Toujours «jeter un œil» aux données. A fortiori si elles sont spatiales !

- identification des tendances, des motifs
- complétude
- outliers

Visualiser **avant** : analyse visuelle exploratoire et

Visualiser **après** : restitution/diffusion des résultats des traitements

Toujours

- étiqueter les axes (unités)
- fournir une légende complète
- choisir des classes de valeurs et de couleurs sensées
- penser au support de diffusion

«Self explanatory» : Si on se demande quoi voir ou comment lire, c'est raté. Si c'est moche, trop petit, peu lisible, idem.

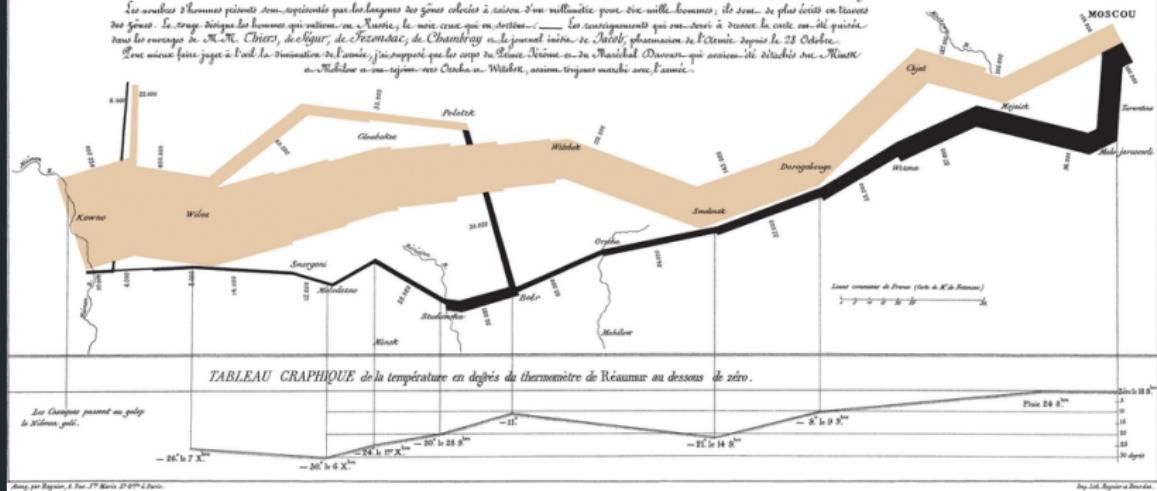
⇒ Faites simple

Exemples réussis

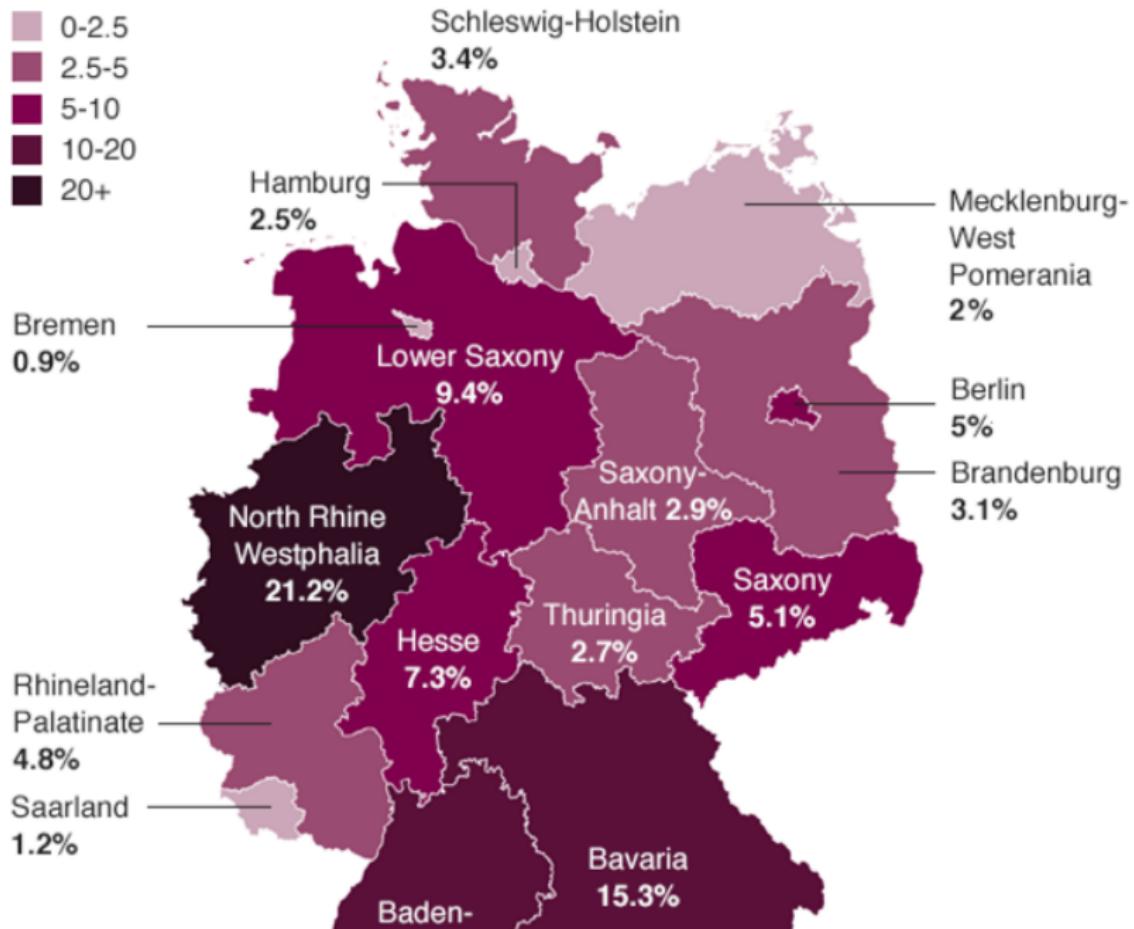
Carte Figurative des pecces successives en hommes de l'Armée Française dans la Campagne de Russie 1812-1813.

Dessiné par M. Minard, Ingénieur Général des Ponts et Chaussees en retraite. Paris, le 20 Novembre 1869.

Les auroches d'hommes présentent une... représentant que les bœufs des gares échoués à cause d'un... n'atteignent pour six mille hommes ; de sorte... de plus écrit en lettres dans les envois de M. M. Chatelet, de l'agréé, de Chambry... le journal intime de Napoléon, plusieurs de l'Armée depuis le 21 Octobre... Lors même faire jeter à l'est la dimension de l'armée, qui suppose que les corps de l'Armée... et de Marshal Davout, qui avaient été détachés sur... à Malibor n'avaient pas... Orléans à Wittenberg, assister... marché avec l'armée.

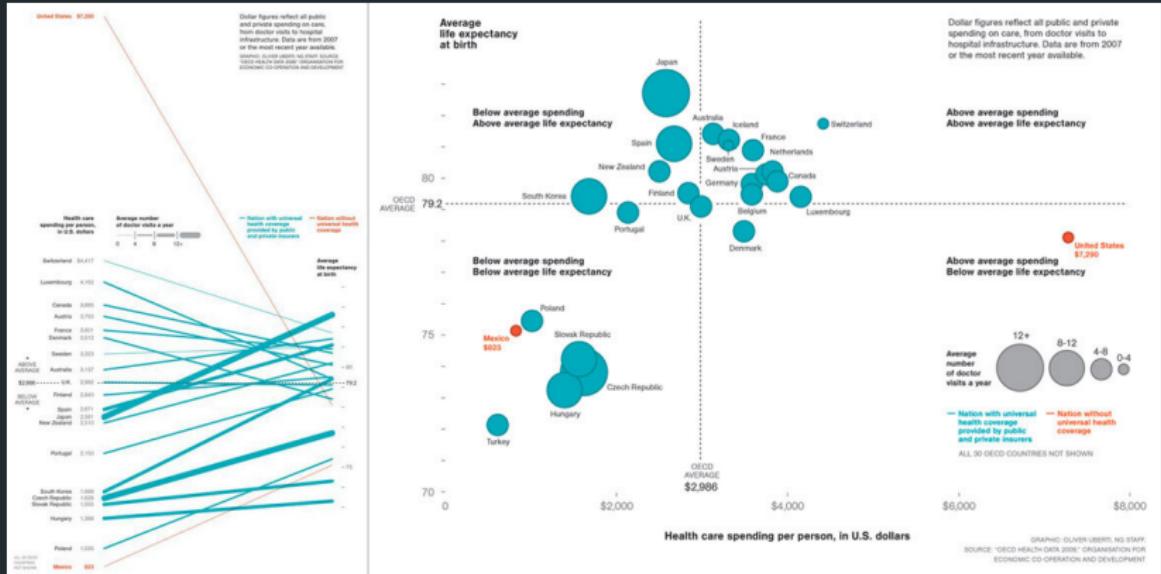


Distribution of Asylum Seekers in Germany



Migrant route to Germany





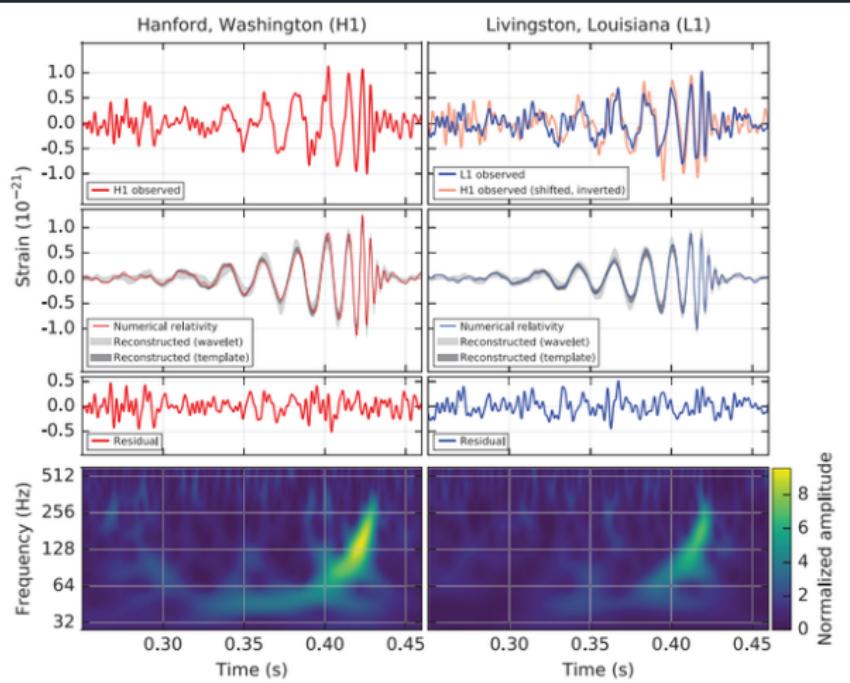
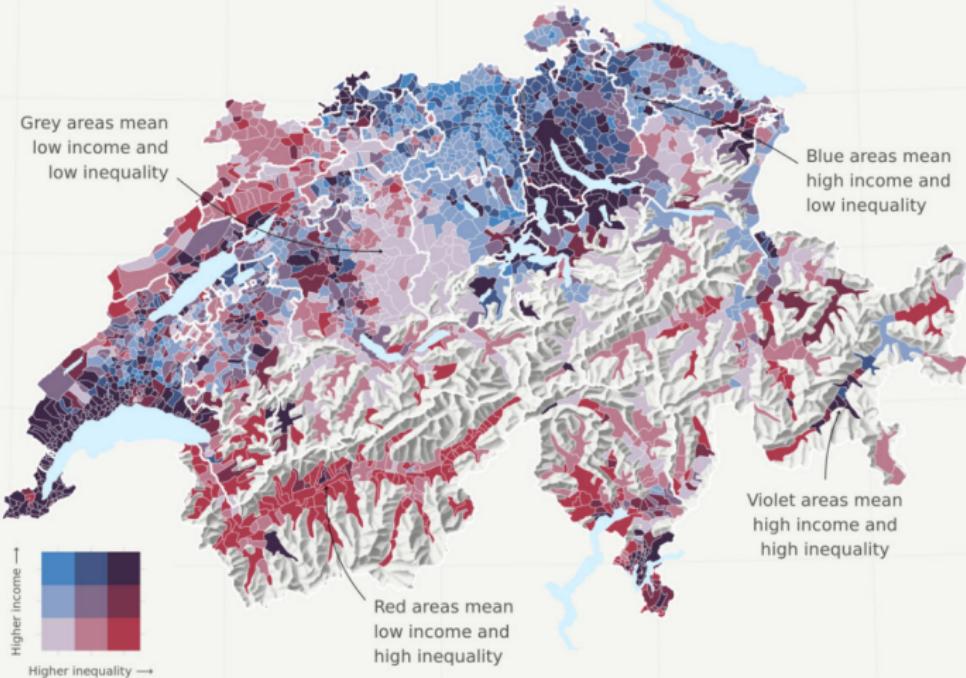


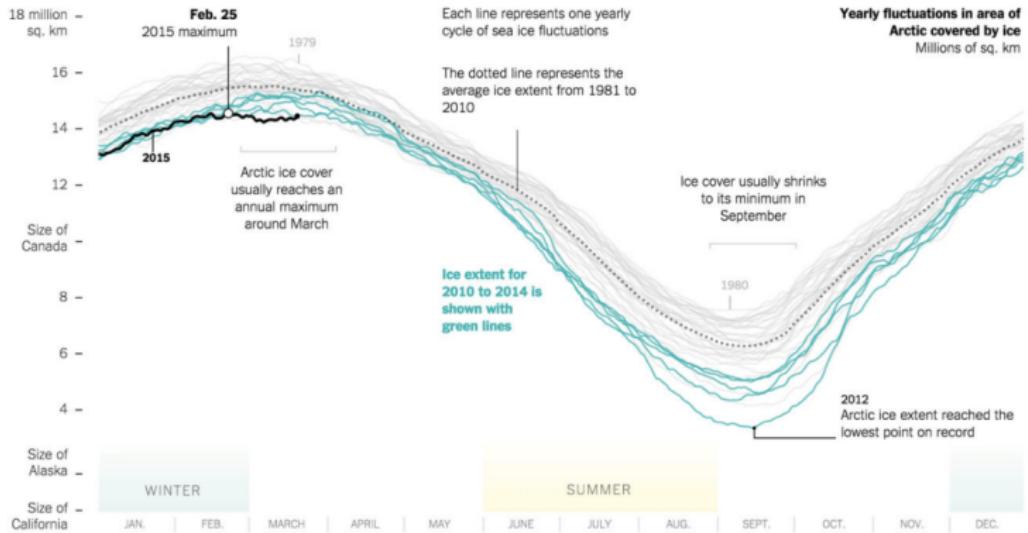
FIG. 1. The gravitational-wave event GW150914 observed by the LIGO Hanford (H1, left column panels) and Livingston (L1, right column panels) detectors. Times are shown relative to September 14, 2015 at 09:50:45 UTC. For visualization, all time series are filtered with a 35–350 Hz bandpass filter to suppress large fluctuations outside the detectors' most sensitive frequency band, and band-reject filters to remove the strong instrumental spectral lines seen in the Fig. 3 spectra. *Top row, left:* H1 strain. *Top row, right:* L1 strain. GW150914 arrived first at L1 and $6.9^{+0.5}_{-0.4}$ ms later at H1; for a visual comparison, the H1 data are also shown, shifted in time by this amount and inverted (to account for the detectors' relative orientations). *Second row:* Gravitational-wave strain projected onto each detector in the 35–350 Hz band. Solid lines show a numerical relativity waveform for a system with parameters consistent with those recovered from GW150914 [37,38] confirmed to 99.9% by an independent calculation based on [15]. Shaded areas show 90% credible regions for two independent waveform reconstructions. One (dark gray) models the signal using binary black hole template waveforms [39]. The other (light gray) does not use an astrophysical model, but instead calculates the strain signal as a linear combination of sine-Gaussian wavelets [40,41]. These reconstructions have a 94% overlap, as shown in [39]. *Third row:* Residuals after subtracting the filtered numerical relativity waveform from the filtered detector time series. *Bottom row:* A time-frequency representation [42] of the strain data, showing the signal frequency increasing over time.

Switzerland's regional income (in-)equality

Average yearly income and income (in-)equality in Swiss municipalities, 2015



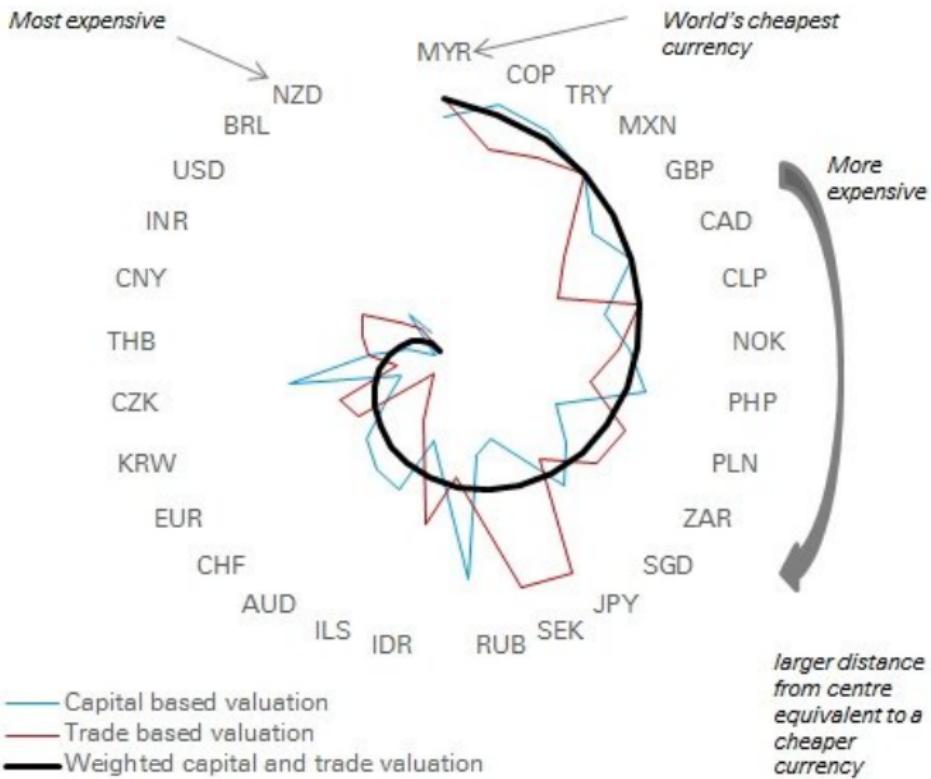
Map CC-BY-SA; Code: github.com/grssnbchr/bivariate-maps-ggplot2-sf
Authors: Timo Grossenbacher (@grssnbchr), Angelo Zehr (@angelozehr)
Geometries: ThemapKart BFS and swisstopo; Data: ESTV, 2015



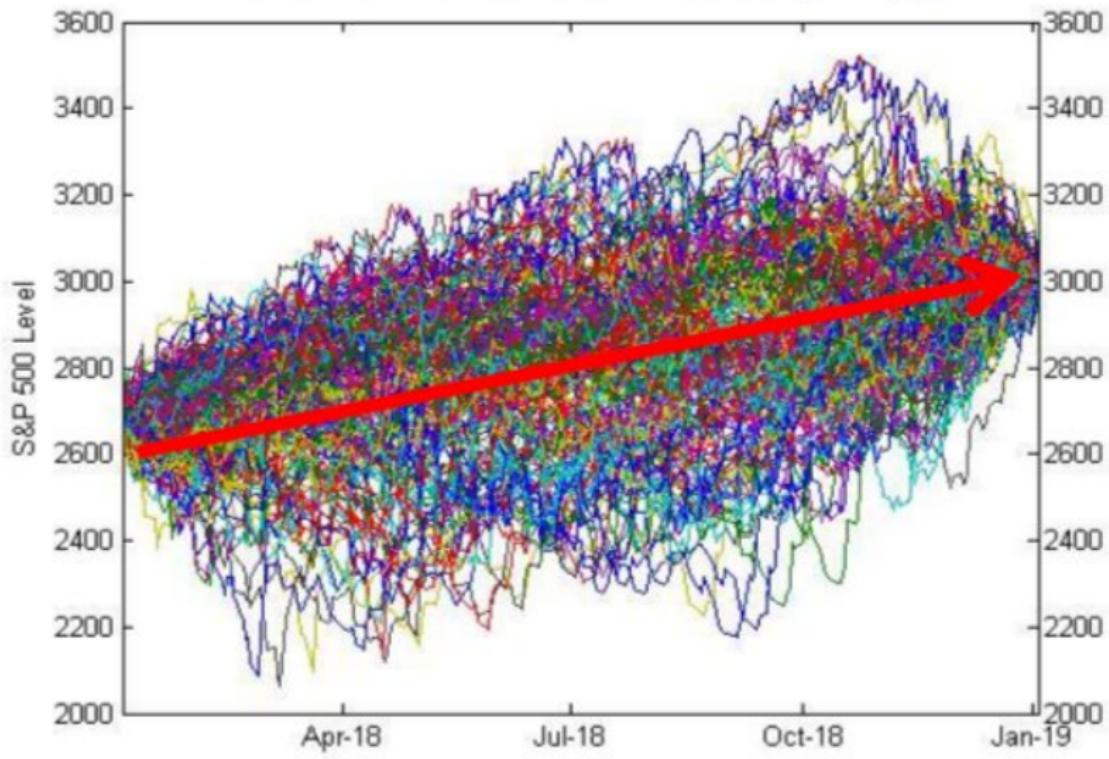
"Yearly Fluctuations in Area of Arctic Covered by Ice" by Derek Watkins (*New York Times*)

Cédric Scherer // rstudio::conf // July 2022

Musées des horreurs



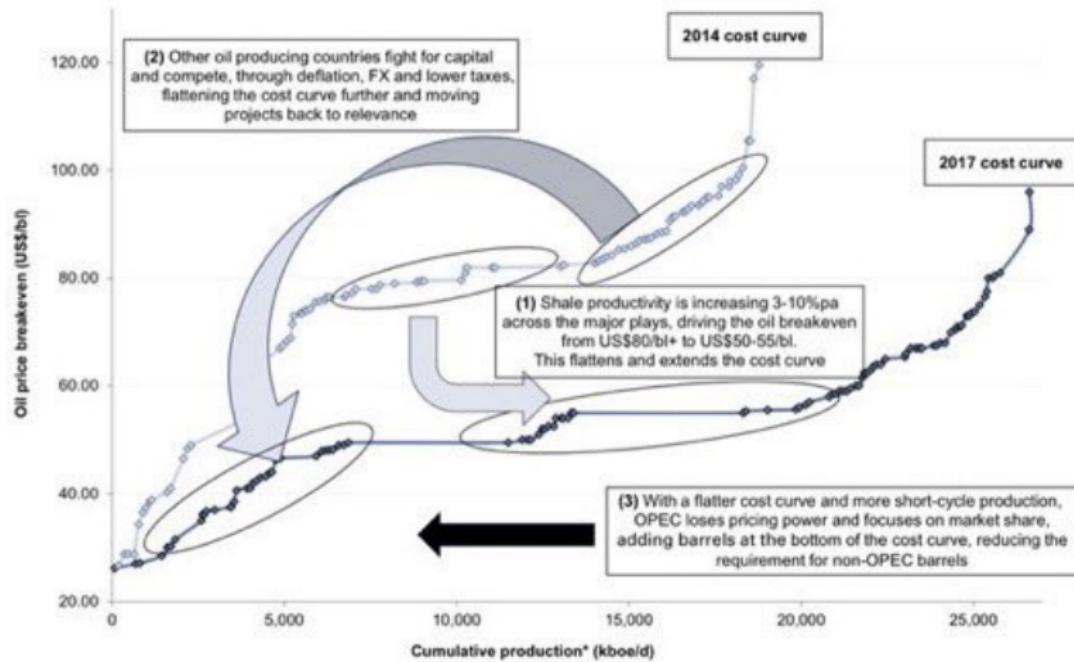
Potential Paths to SPX 3000 (+12 %)



Source: Credit Suisse Equity Derivatives Strategy

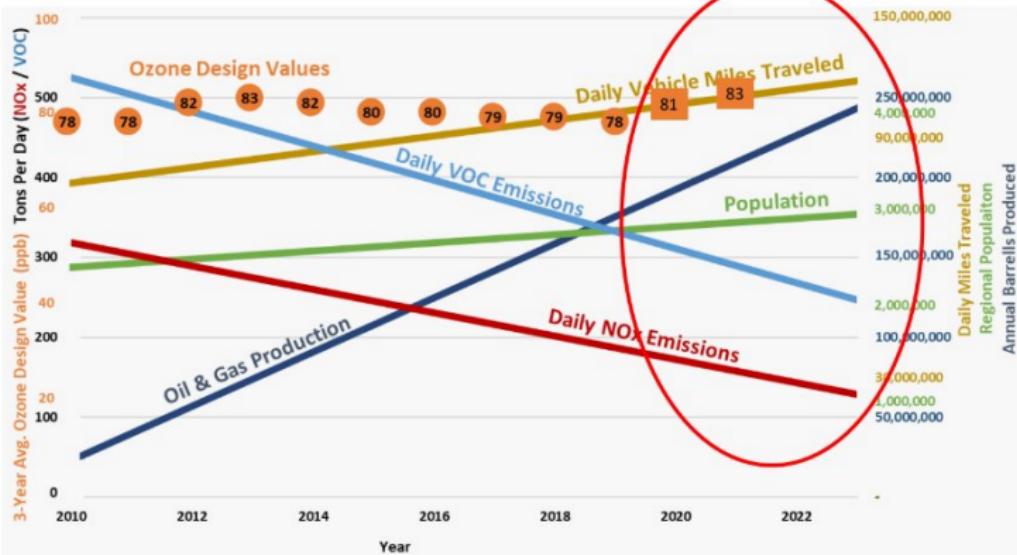
Exhibit 2: Short-cycle shale has engendered a structural deflationary cycle

Pre-sanction cost curve in 2017 vs. 2014 for non-OPEC from our Top Projects database



Source: Goldman Sachs Global Investment Research

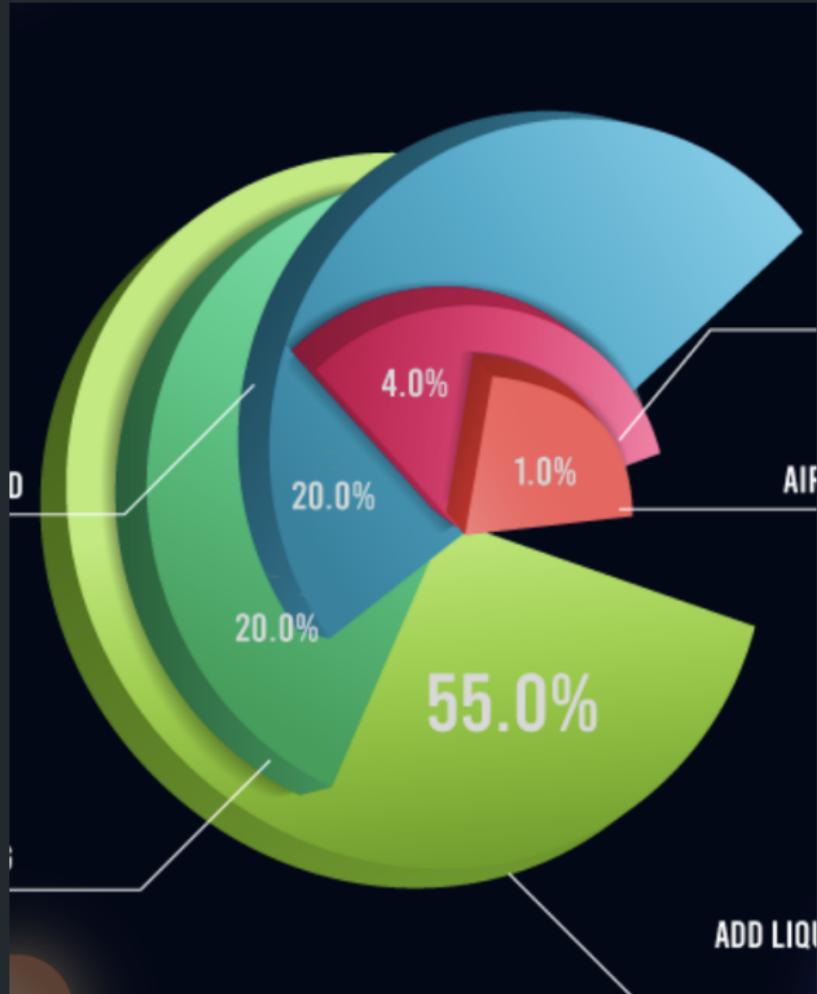
Regional Trends (2010-2023)



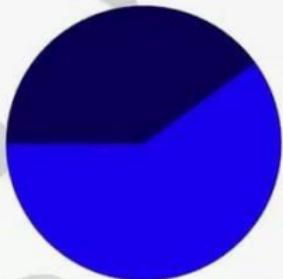
08/18/2022

AQCC Presentation SIP Planning Process

12



JEAN-LUC MÉLENCHON EST-IL ATTACHÉ AUX VALEURS
DÉMOCRATIQUES ?



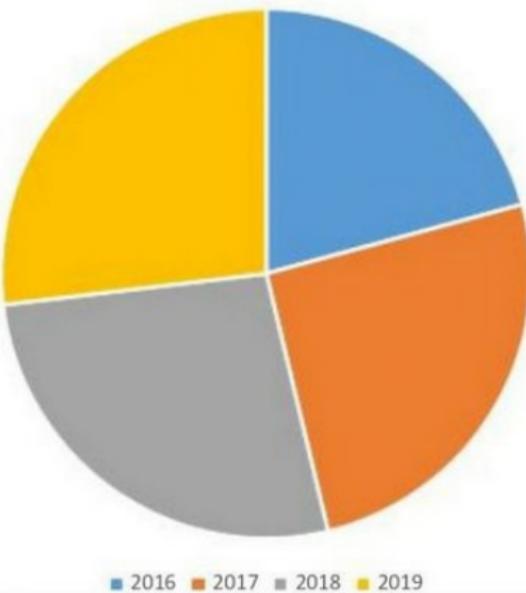
44%
OUI
66%
NON

SOURCE : SONDEMENT IFOP POUR SUD RADIO

"Si nous n'agissons pas, on pourrait se retrouver dans une situation proche de celle de mars, ça pourrait dire reconfinement" (Castex/France 2)

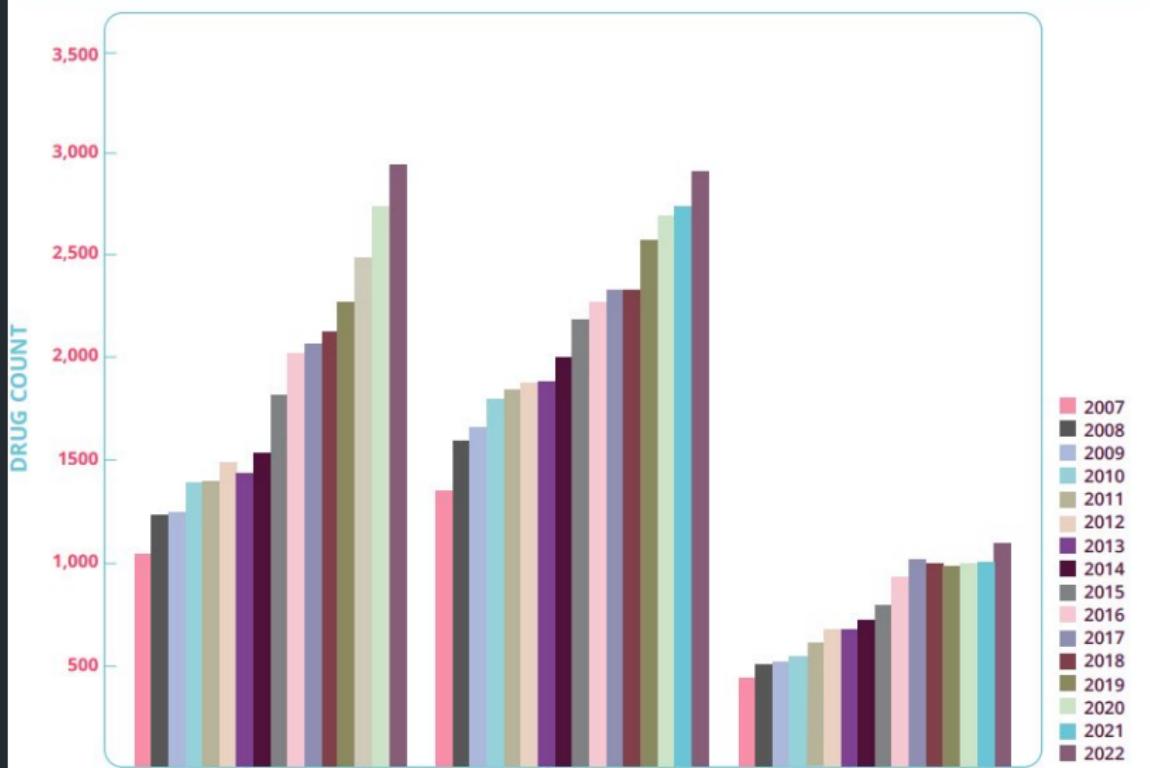


Chart showing shortage of vehicles for conveying inmates to court

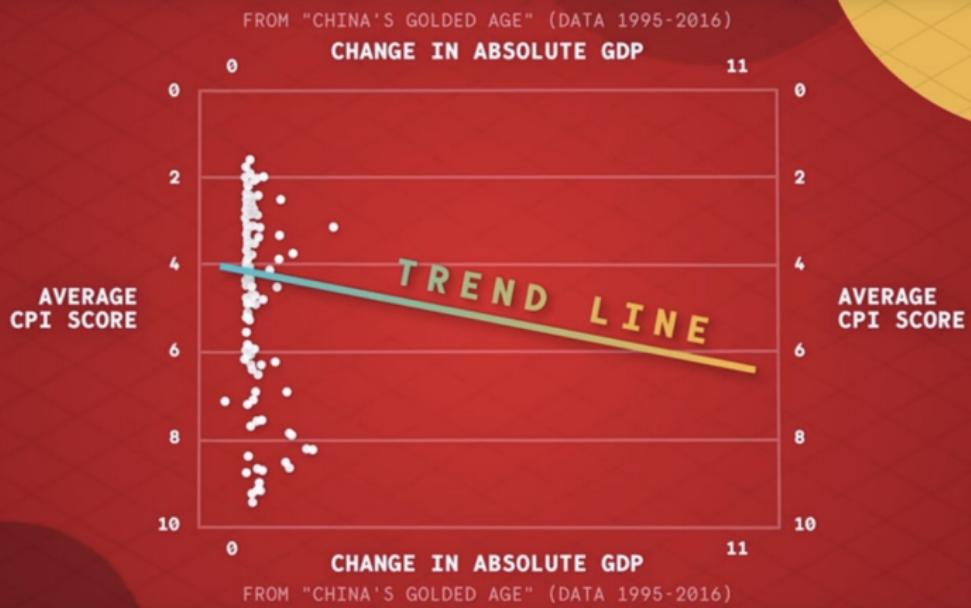


Before 2016, the Nigerian prisons said it had only two vehicles for conveying an average of 400 people to court, daily.

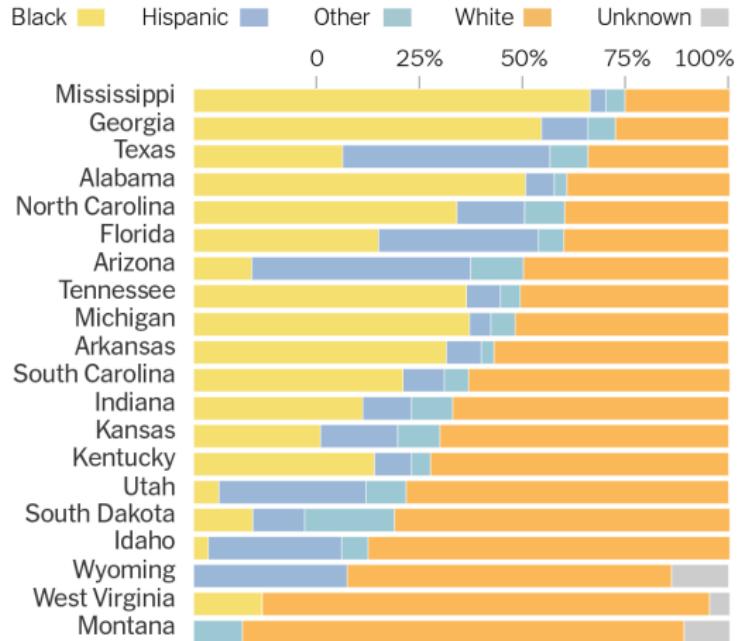
FIGURE 3:
Clinical phase trends, 2007-22



Source: Pharmanprojects® January 2022



Race and ethnicity of abortion patients in states that could ban or restrict the procedure

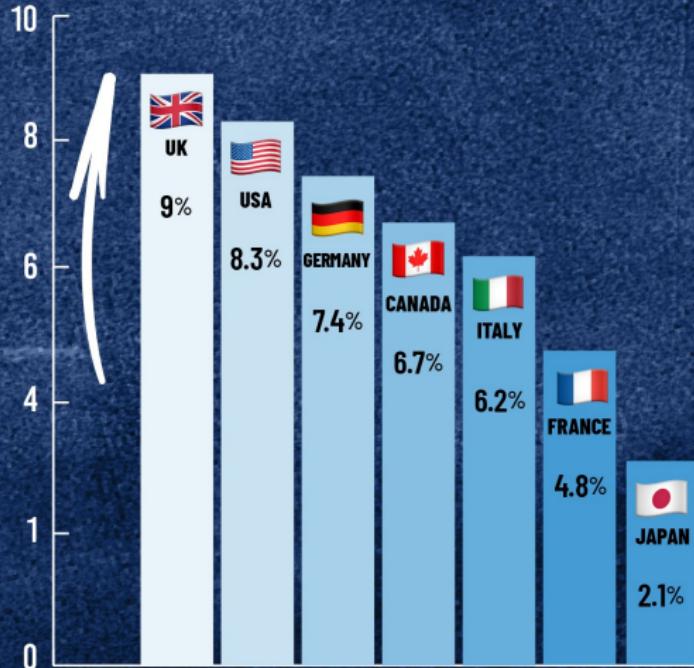


Note: Data for each group reflect only a single race or ethnicity.
Data is not available for Iowa, Louisiana, Missouri, Nebraska,
North Dakota, Ohio, Oklahoma or Wisconsin. Data in small
values is suppressed for confidentiality reasons.

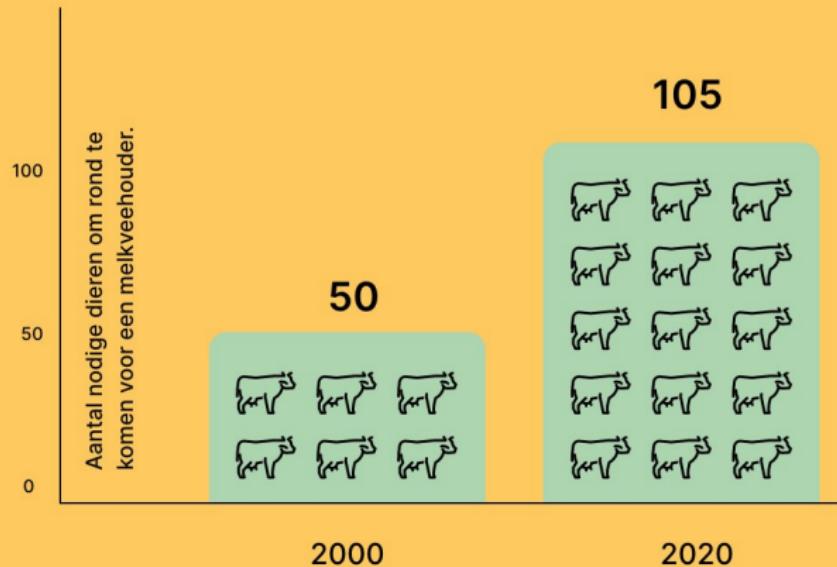


UNDER THE CONSERVATIVES

INFLATION IN THE UK IS HIGHER THAN ANY OTHER G7 COUNTRY

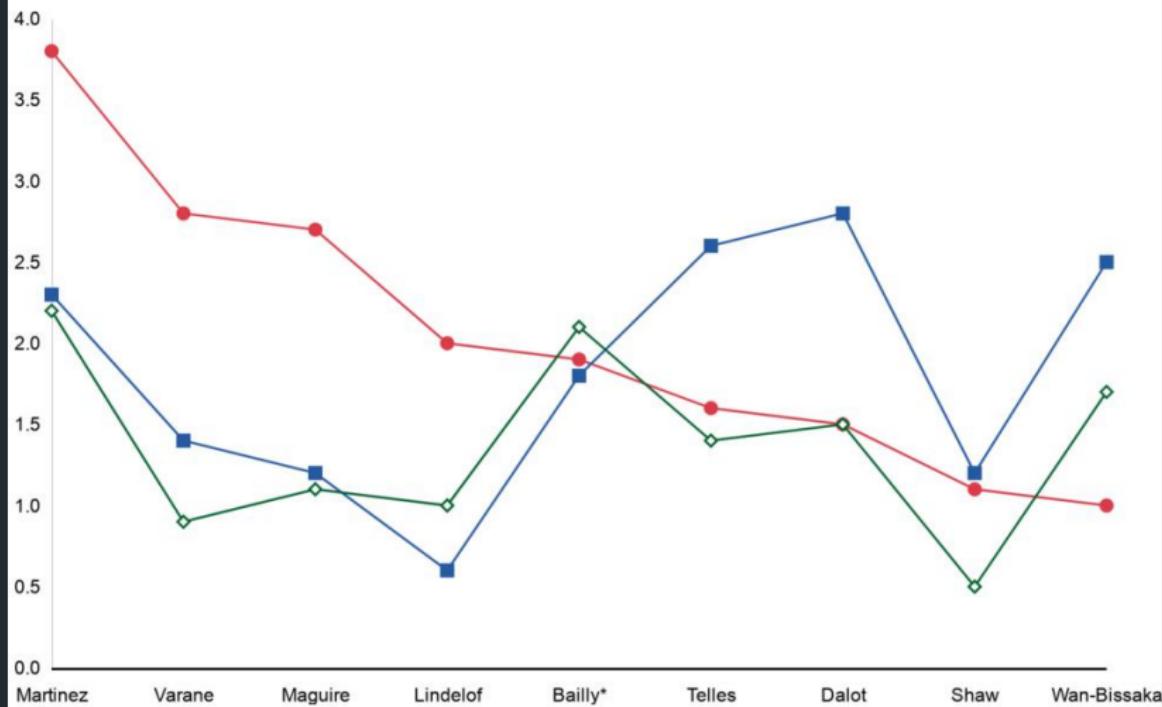


Boeren moeten steeds meer dieren houden om rond te komen.



Stats per 90 minutes in the league last season

Aerial duels won Tackles Interceptions



* Bailly's stats are for his whole time at Man Utd in the Premier League

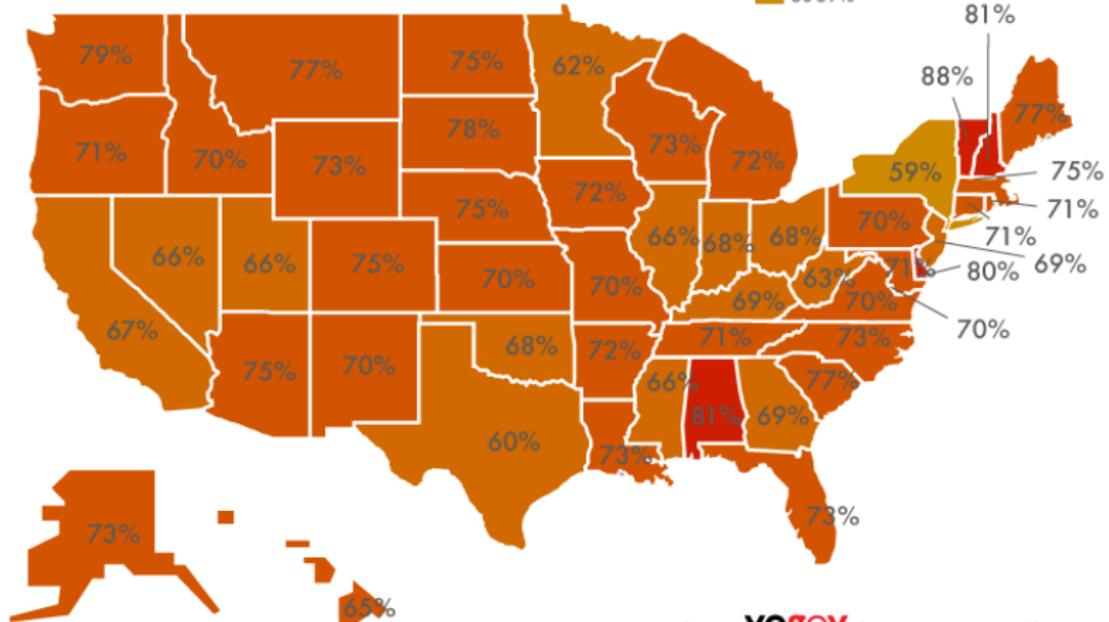
Source: Opta

BBC

Licensed Drivers Per Capita

Map of All Fifty States

- 80%+
- 70-79%
- 60-69%
- 50-59%



copyright 2017 **yogov** learn more at yogov.org/blog

**VOTO
2013**
El pueblo decide

ELECCIÓN PRESIDENCIAL 2013

Venezolana
de Televisión

PORCENTAJES



% 49,07

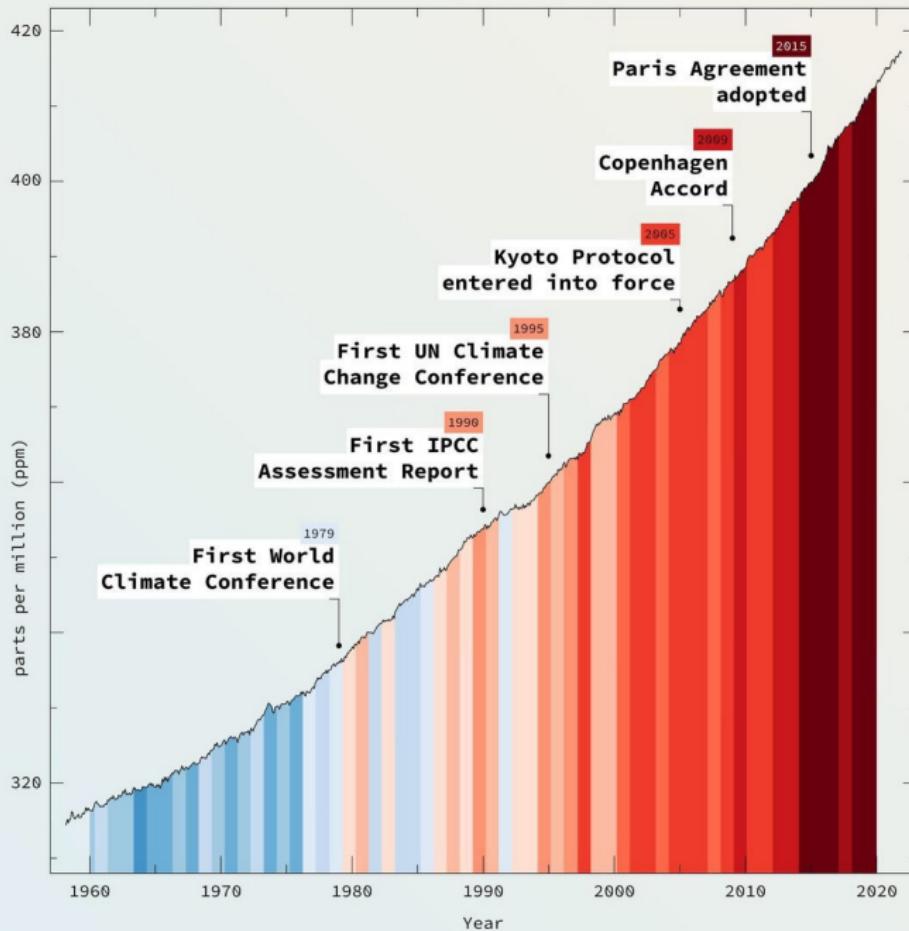


Nicolás Maduro Moros



Henrique Capriles Rad

Trends in Atmospheric CO₂ vs Global Temperature Change



Composite Graph of: Atmospheric CO₂ at Mauna Loa Observatory | December 2021 – Scripps Institution of Oceanography & NOAA Global Monitoring Laboratory | #ShowYourStripes – Graphics & Lead scientist: Ed Hawkins, National Centre for Atmospheric Science, University of Reading; Data: UK Met Office | Design by: sustentio [PG] | Licence: CC-BY

① Ad Info

\$

2014
\$102 BILLION

2020
\$187
BILLION

DRUG REBATES BY YEAR