



Bonjour



Introduction à la visualisation de données

Plan :

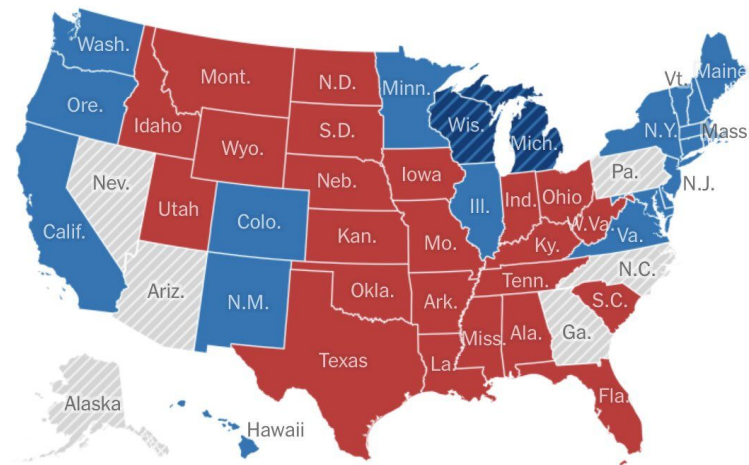


- Présentation du cours
- Critique
- Pourquoi visualiser ?
- Qu'est ce que la visualisation
- Type de données et de visualisation
- Variables graphiques
- Quelles solutions techniques

Analyse critique d'une visualisation

élections présidentielles aux états unis

- À qui s'adresse la visualisation ?
-> 1 proposition
- À quelle question la visualisation permet elle de répondre ?
-> 1 proposition
- Pourquoi (n')aimez vous (pas) cette visualisation ?
-> 2 raisons
- Quelles améliorations apporter ? -> 3 propositions





Joe Biden

225

68,868,787 votes

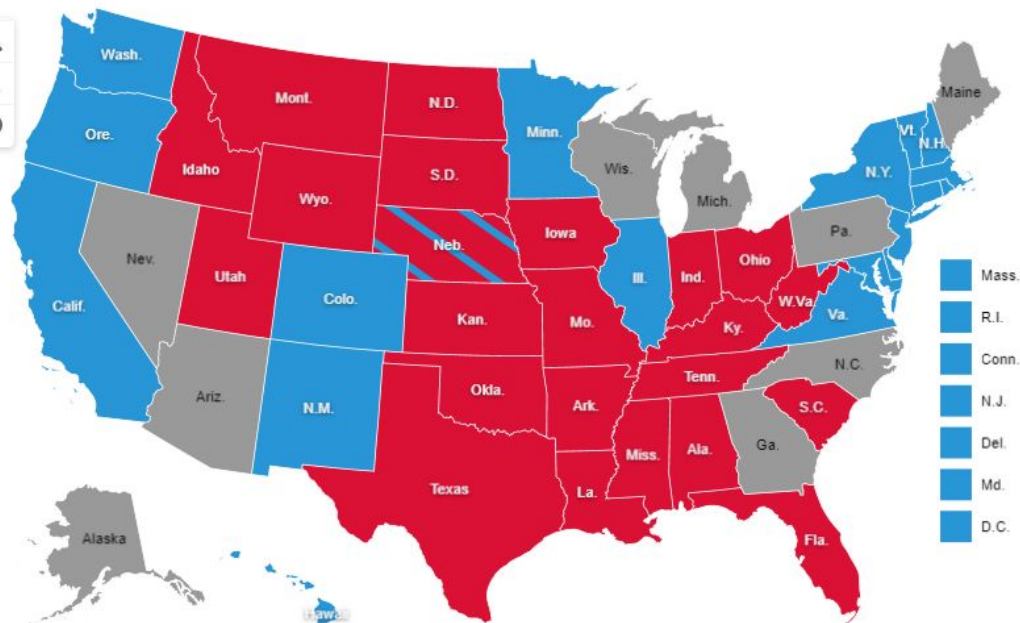
Donald Trump

213

66,198,569 votes



270 to win

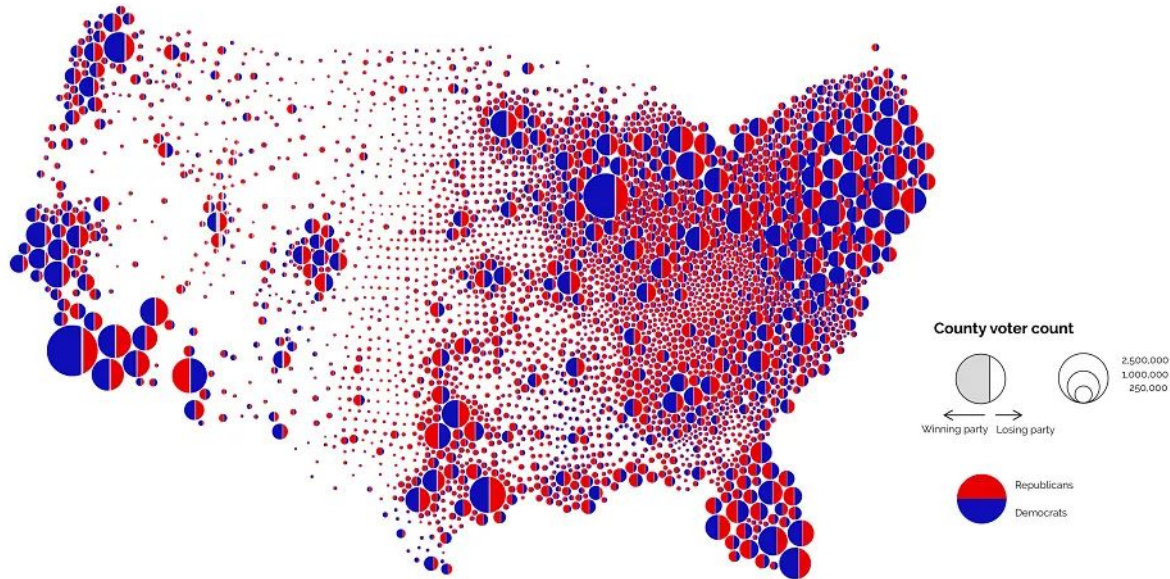


No Election No Results Counting Results

Win Gain Dem. Win Gain Rep. Win Gain Other



2016 U.S. presidential election





Dataviz

Pourquoi la visualisation de données



@code.know

Credit: @dataprofessor



Explosion de la quantité de données

- Comment faire sens des données ?
- Comment utiliser ces données dans les processus de décision ?
- Comment ne pas être surchargé ?

Défi: transformer les données en connaissance (découverte, compréhension) pour qu'elles deviennent utiles



Traiter les données: où les machines sont plus efficaces

Question bien définie, sur des données connues

- Quel est le taux de chômage sur l'année 2024 ?
- Quel gène mute fréquemment sur tel ensemble de patients ?

Décisions qui doivent être faites en un minimum de temps

- High-frequency trading
- Détection de fraudes



Traiter les données: où l'humain est plus efficace

Quand les questions ne sont pas bien définies (exploration)

- Quelle combinaison de gènes peut être associée à un cancer ?

Quand les résultats peuvent donner lieu à plusieurs interprétations

- Quelle est la relation entre l'emploi et la politique industrielle d'un pays ?
- Quel est l'impact du confinement sur la baisse des cas de covid-19 ?



Le classique : Pourquoi ne pas s'appuyer que sur l'analyse de données ?

Imaginons 4 datasets (observations) :

I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89



I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Statistiques

Moyenne

x: 9 y: 7.50

Variance

x: 11 y: 4.122

Corrélation

$x - y$: 0.816

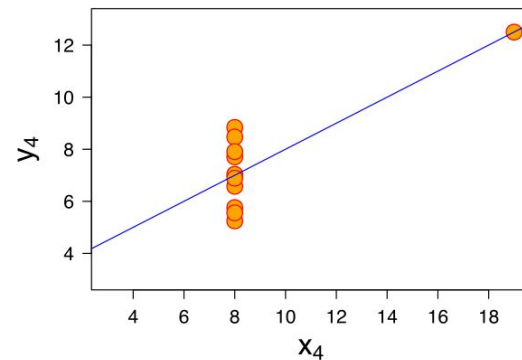
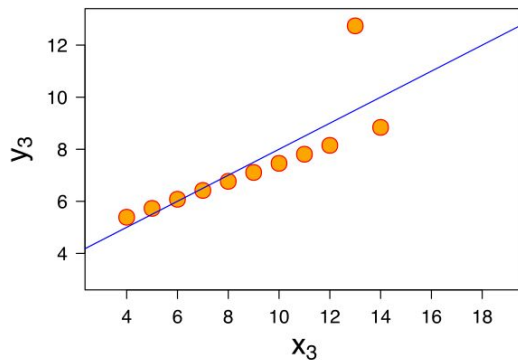
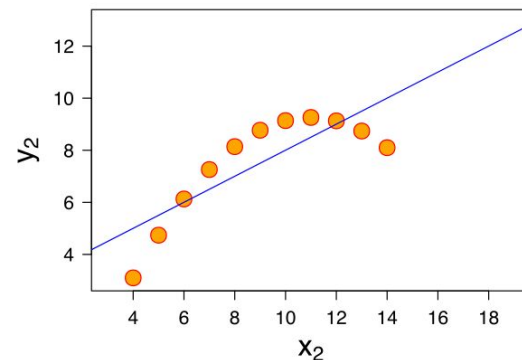
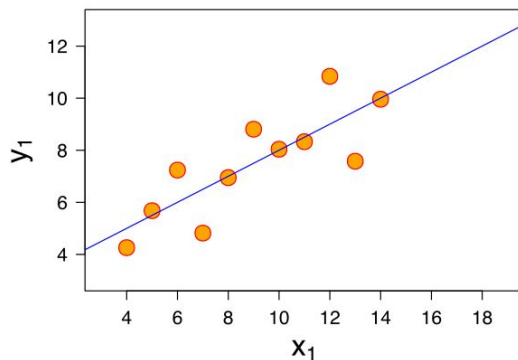
Régression linéaire:

$y = 3.00 + 0.500x$



Le classique : Pourquoi ne pas s'appuyer que sur l'analyse de données ?

Le Quartet d'Anscombe
https://en.wikipedia.org/wiki/Anscombe%27s_quartet



Don't (always) trust summary statistics



Alberto Cairo

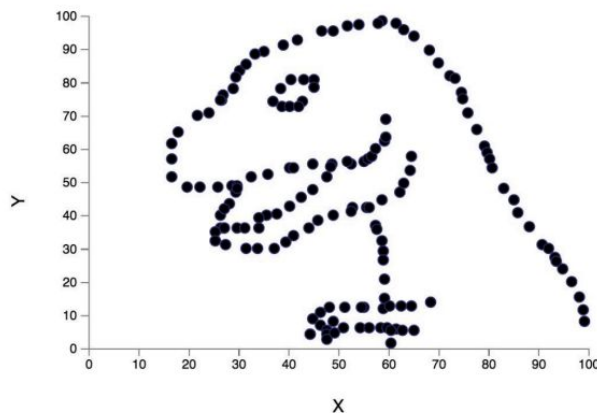
@albertocairo

Following

Don't trust summary statistics. Always
visualize your data first

robertgrantstats.co.uk/drawmydata.html

N = 157 ; X mean = 50.7333 ; X SD = 19.5661 ; Y mean = 46.495 ; Y SD = 27.2828 ;
Pearson correlation = -0.1772



5:47 AM - 15 Aug 2016



Don't (always) trust summary statistics

<https://jumpingrivers.github.io/datasauRus/>



Pourquoi visualiser ?

Brainstorming

Brainstorming sur l'utilité
de la visualisation de données.

10 minutes

5 raisons



Visualisation de données

Représenter graphiquement une information souvent abstraite et/ou à très grande volumétrie

Communiquer des chiffres ou des informations brutes en les transformant en objets visuels : points, barres, courbes, cartographies...

- Rendre compréhensible les données
- Détecter des erreurs, incohérences
- Faciliter le raisonnement et la prise de décision
- Transmettre de l'information
- Emettre des hypothèses



Les 3 grandes raisons

Enregistrer de l'information

- Plan, photo

Faciliter le raisonnement sur de l'information (analyser)

- Analyser et calculer
- Raisonner sur les données
- Feedback et interaction

Transmettre de l'information (présenter)

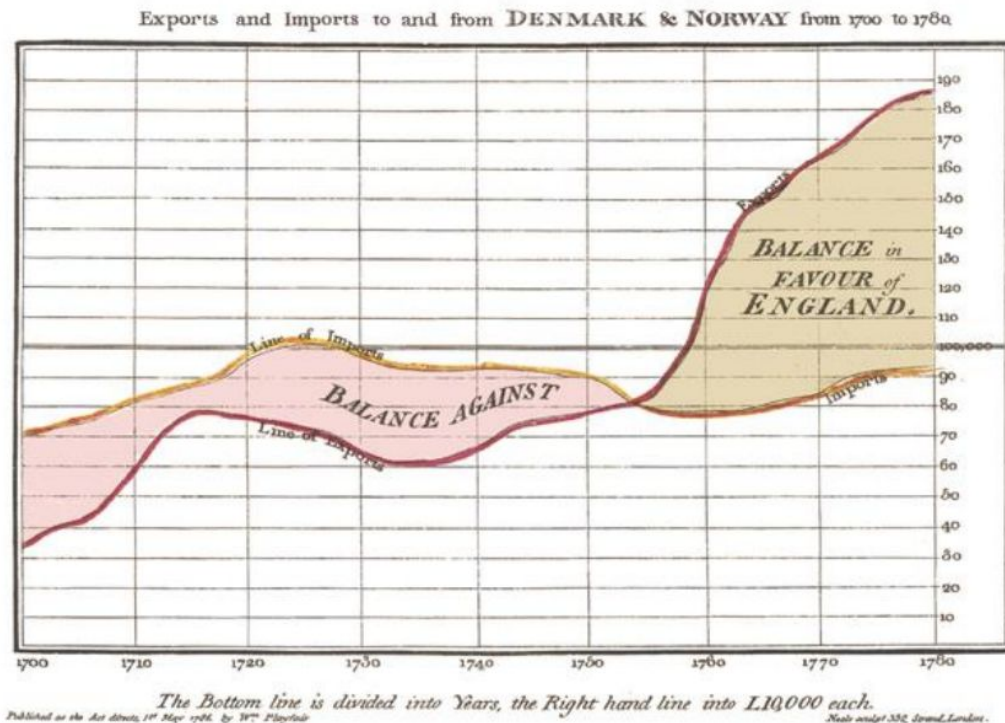
- Partager et persuader
- Collaborer et itérer
- Mettre en avant un aspect des données

Vous serez amené à travailler (collaborer) avec des non informaticiens



Historique

William Playfair (ingénieur et économiste écossais / 1759-1823) considéré comme l'inventeur des graphiques pour représenter les données : line plots, bar chart and pie chart



Historique

John Snow (médecin britannique / 1813–1858) connu pour avoir tracé (détecté) les sources de Choléra à Londres.

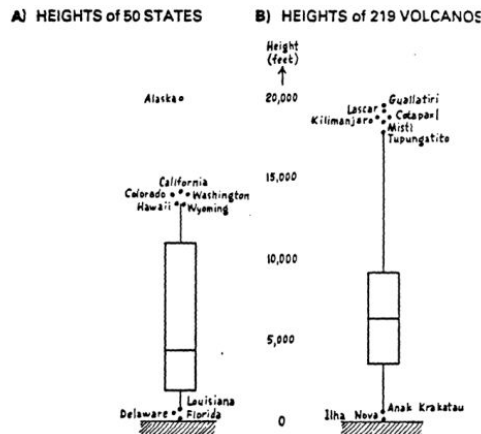
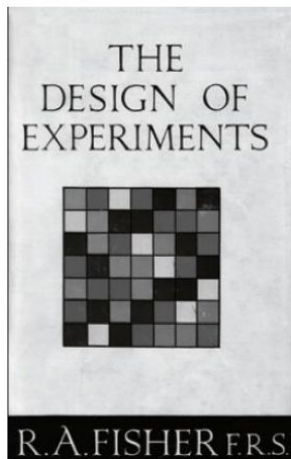


Historique

Ronald Fisher (1890-1962) et **John Tukey** (1915-2000) : méthodes graphiques avancées pour l'analyse des données.

Fisher : dessin des données pour comprendre les relations

Tukey : promotion de l'analyse de données exploratoires, il a créé en particulier le box plot, le stem plot et le leaf plot.



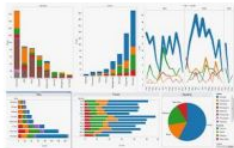
Et aujourd'hui ?



Amazon.fr - Data Visualisa...
amazon.fr



Datavisualisation : 3 outils pour créer de belles info...
archimag.com



Datavisualisation : les logiciels en vogue - w...
datavisualization.fr



Data visualisation : mieux communiquer...
http5000.com



12 solutions de datavisualisation
e-marketing.fr



Data visualisation charts...
vectorstock.com



Data Visualization : Top des eta...
leighta.fr



Data visualisation - Parfonse Crypto.fr
parfonsecrypto.fr



Introduction aux librairies de data...
blog.equency.com



Big Data Visualisation Colorée. Infographie Fut...
fr.123rf.com



Data Visualization : top des meilleurs outils de visua...
lebigdata.fr



Data Visualisation Design &MK47 | Aleasynspain
aleasynspain.com



Mes astuces pour réussir la data-visualisation | Hu...
news.humancoders.com



Dataviz

La **dataviz** (Data Visualisation) : désigne les techniques permettant de présenter des données sous forme visuelle afin d'en faciliter la compréhension et/ou l'analyse.



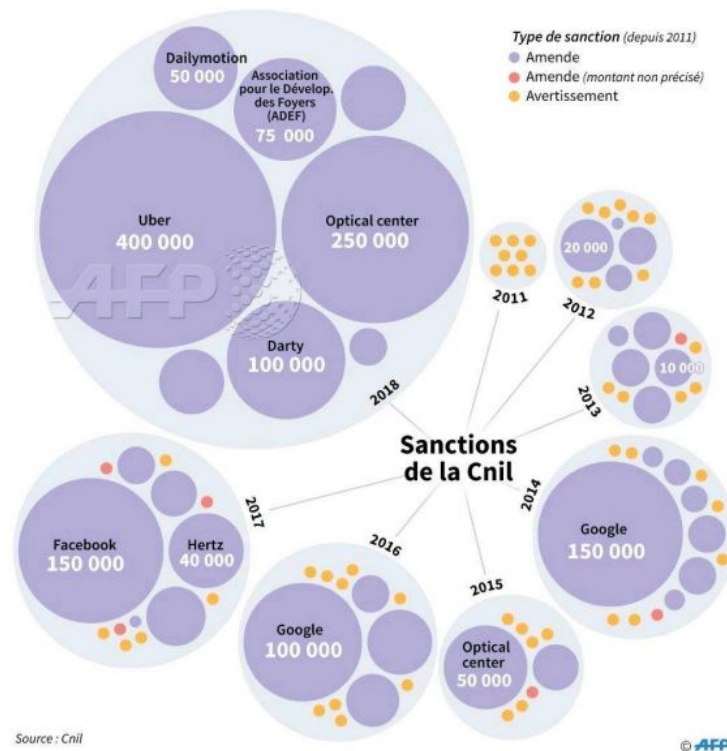
Dataviz “statiques”

- Des visualisation de données figées, représentant les données ciblées
- Cette catégorie a l'avantage d'être complètement adaptée aux rapports et aux présentations



DataViz “statiques”

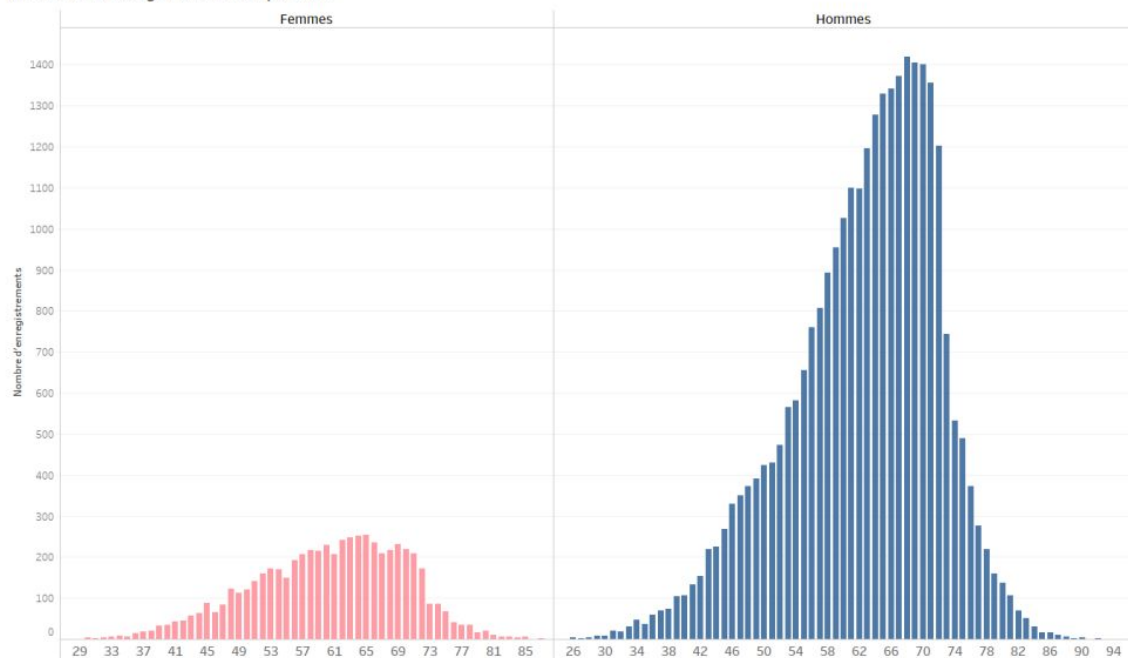
Répartition



DataViz “statiques”

Distribution

Distribution des ages des mairies par sexe



DataViz Interactive

Exemples :

<https://square.github.io/crossfilter/>

<https://dataviz.rennesmetropole.fr/quisommesnous/>



Type de données

Les données

À la base de toute visualisation

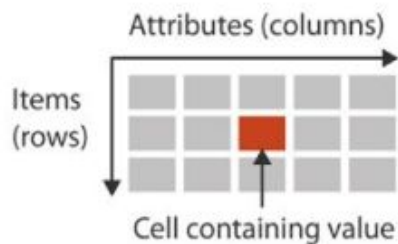
Un bon designer de visualisation doit connaître :

- Les propriétés des données
- Les méta-données associées
- Ce que les gens veulent tirer des données

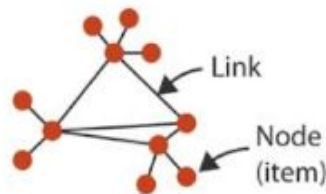


Types de jeux de données

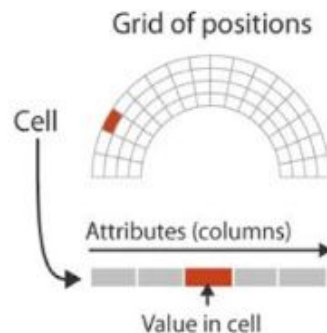
→ Tables



→ Networks



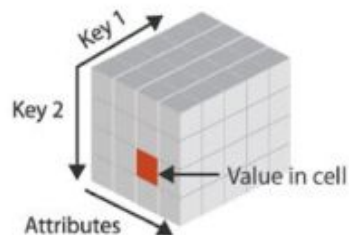
→ Fields (Continuous)



→ Geometry (Spatial)



→ *Multidimensional Table*



→ *Trees*



-> *Ce qu'on veut visualiser*

Type de données de base

Unités fondamentales

Constituent les jeux de données

- Item / élément
- Lien
- Attribut
- Position
- Grille



Exemple item/attribut

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box		7/17/07
32	7/16/07	2-High	Medium Box		7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69		5 4-Not Specified	Small Pack	0.44	6/6/05
69		5 4-Not Specified	Wrap Bag	0.6	6/6/05
70	12/18/06	5-Low	Small Box	0.59	12/23/06
70	12/18/06	5-Low	Wrap Bag	0.82	12/23/06
96	4/17/05	2-High	Small Box	0.55	4/19/05
97	1/29/06	3-Medium	Small Box	0.38	1/30/06
129	11/19/08	5-Low	Small Box	0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

attribute

item

cell



Élément et attribut

Élément :

- Entité individuelle, discrète.
- Ex: un patient, une voiture

Attribut :

- Propriété mesurée ou observée
- Ex: taille, pression sanguine (patient), vitesse (voiture)



Lien, Position et Grille

Lien

- Relation entre deux éléments
- Ex : “amitié sur Facebook”

Position

- Données spatiales (en 2D ou 3D)
- Ex : latitude/longitude

Grille

- Stratégie d'échantillonnage pour données continues
- Ex: positions de stations météo



Type d'échelles

Nominale (catégoriel)

- Fruits: pommes, oranges, ...

Ordinale (ordonné)

- Consommation énergétique : A, B, C,...
- Peut être compté et ordonné mais pas mesuré

Intervalle (zéro arbitraire)

- Dates, longitude, latitude

Ratio (zéro fixé)

- Le zéro a un sens (rien)
- Mesure physique : poids, longueur, ...



Type d'échelles (Exemples et opérations)

Nominale (catégoriel)

- Opérations : $=$, \neq

Ordinale (ordonné)

- Opérations : $=$, \neq , $>$, $<$

Intervalle (zéro arbitraire)

- Opérations : $=$, \neq , $>$, $<$, $+$, $-$

Ratio (zéro fixé)

- Le zéro a un sens (rien)
- Mesure physique : poids, longueur, ...





variables graphiques

Marques simples

Munzner, 2014,
Visualization Analysis and Design.

➔ Points



➔ Lines



➔ Areas



Canaux visuels

➔ Position

➔ Horizontal



➔ Vertical



➔ Both



➔ Color



➔ Shape



➔ Tilt



➔ Size

➔ Length



➔ Area

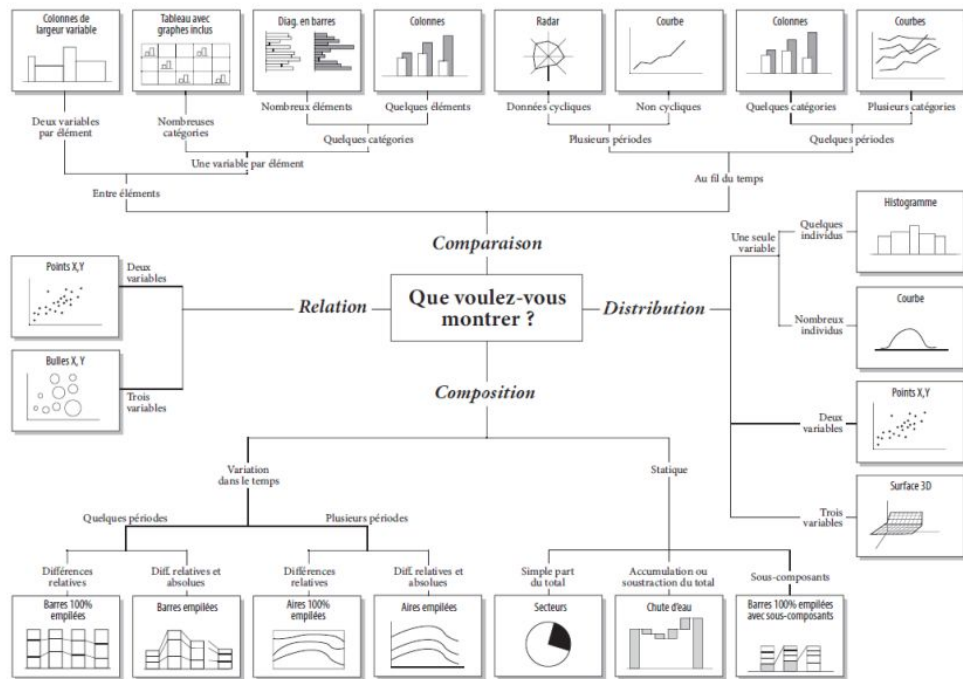


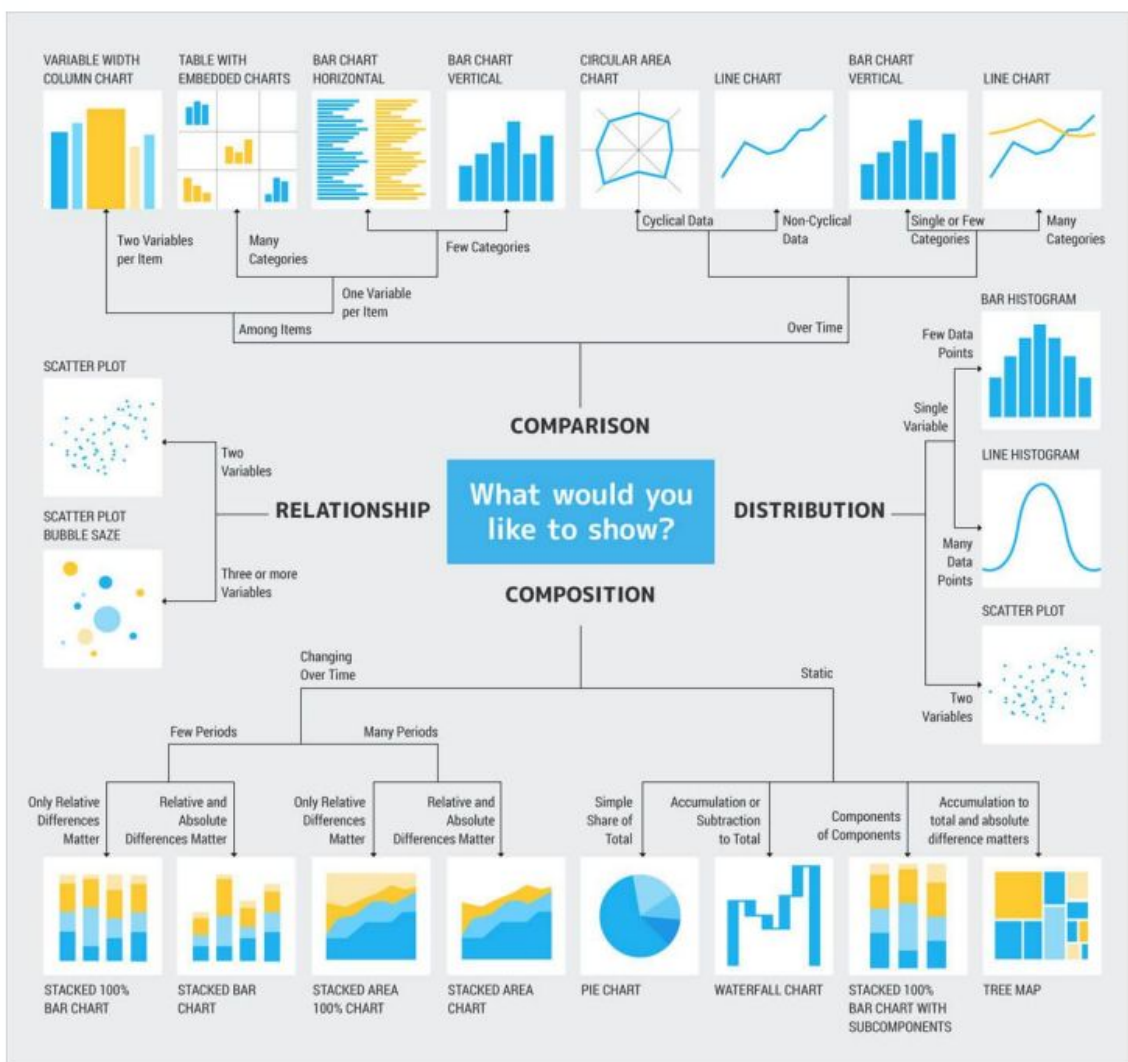
➔ Volume



Quels types de dataviz

Arbre de décision des graphiques





Des outils pour choisir

<https://datavizcatalogue.com/search.html>



Exercice (choisir la bonne visualisation)

- Comparer les performances de ventes de différents produits d'un supermarché au cours du temps
- Visualiser les réponses d'une enquête pour connaître l'opinion d'une population sur l'autorisation de l'euthanasie (pour, contre, sans avis) en fonction de l'âge des répondants
- Montrer géographiquement la distribution de cas de covid sur les différentes régions en France.
- Comparer les pourcentages de vente des différentes marques de smartphones
- Étudier la relation existante entre deux variables continues. Par exemple, la température et le taux de vente de glaces

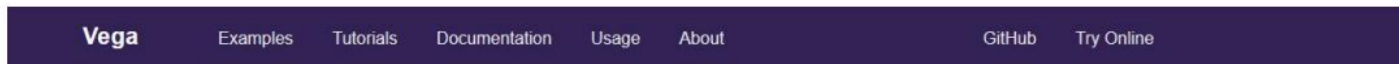


Les clés pour une bonne visualisation

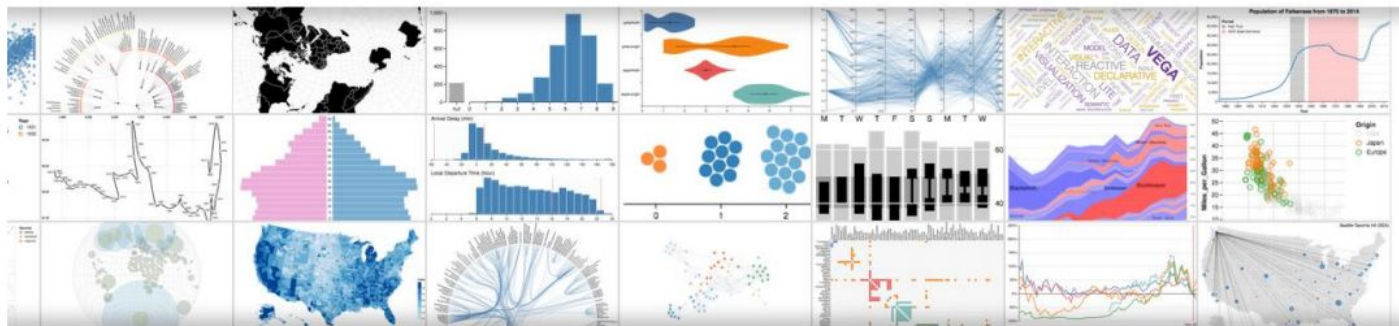
- Le choix des sources de données (primordial)
- Fiabilité de la source des données
- Bien structurer les données : Variables qualitatives et/ou quantitatives
- La construction d'indicateurs pertinents
- Mobilisation de méthodes statistiques (descriptives, corrélation ,...)
- Un mode de visualisation adapté au message et aux données
- Plusieurs type de visualisation
- Des couleurs attrayantes
- Le bon médium de diffusion (présentation, rapport, site Web, blog, réseaux sociaux,...)



Quelles solutions techniques



Vega – A Visualization Grammar



Vega is a *visualization grammar*, a declarative language for creating, saving, and sharing interactive visualization designs. With Vega, you can describe the visual appearance and interactive behavior of a visualization in a JSON format, and generate web-based views using Canvas or SVG.

Vega provides basic building blocks for a wide variety of visualization designs: **data loading** and **transformation**, **scales**, **map projections**, **axes**, **legends**, and **graphical marks** such as rectangles, lines, plotting symbols, etc. Interaction techniques can be specified using **reactive signals** that dynamically modify a visualization in response to **input event streams**.

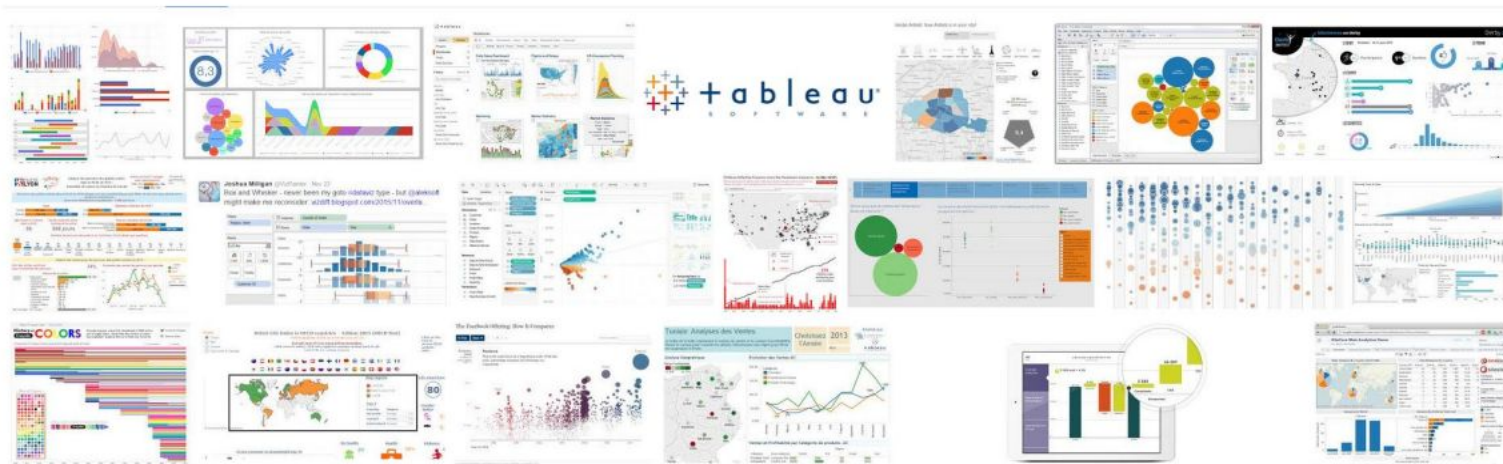
Version 4.4.0

A Vega specification defines an interactive visualization in a JSON format. Specifications are parsed by Vega's JavaScript runtime to generate both static images or interactive web-based views. Vega provides a convenient representation for computational generation of visualizations, and can serve as a foundation for new APIs and client analysis tools.



Quelles solutions techniques

Aujourd'hui. Tableau : Le leader du marché, très intuitif et efficace





EPITA

ÉCOLE D'INGENIEURS EN INFORMATIQUE