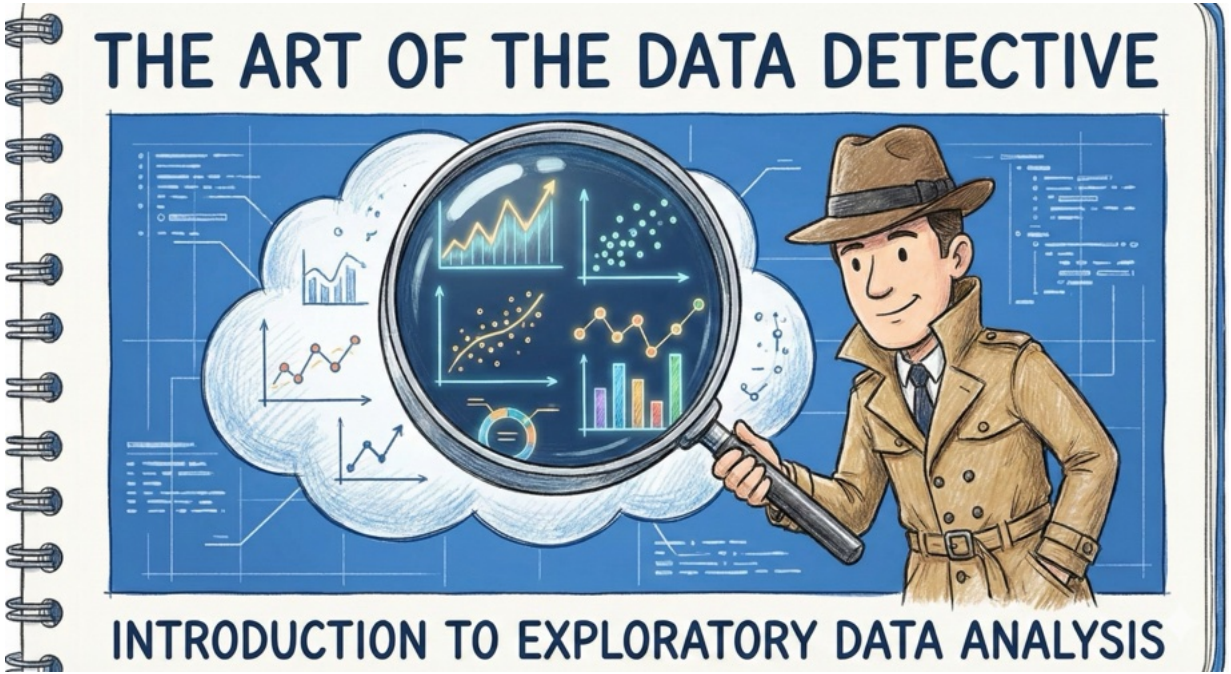
Julien VELCIN, Ecole Centrale de Lyon
https://velcin.github.io

BSc Data Science for Responsible Business
ECL - LIRIS - CNRS



Outline

- Why do we need EDA?
- Introduction and data types (briefing and crime scene)
- Data cleaning and preprocessing (securing the scene)
- Uni/bi-variate analysis, clustering (interrogating the suspects)
- Data storytellling and reporting (verdict)

Why do we need EDA?

- Impossibility to deal with the deluge of data
- Need automatic tools (which ones?)
- GIGO principle (Garbage In Garbage Out)
- Useful to discover new things, build hypotheses, make decisions
- Exploratory vs. Confirmatory analysis

Exploratory to confirmatory (John Tukey)

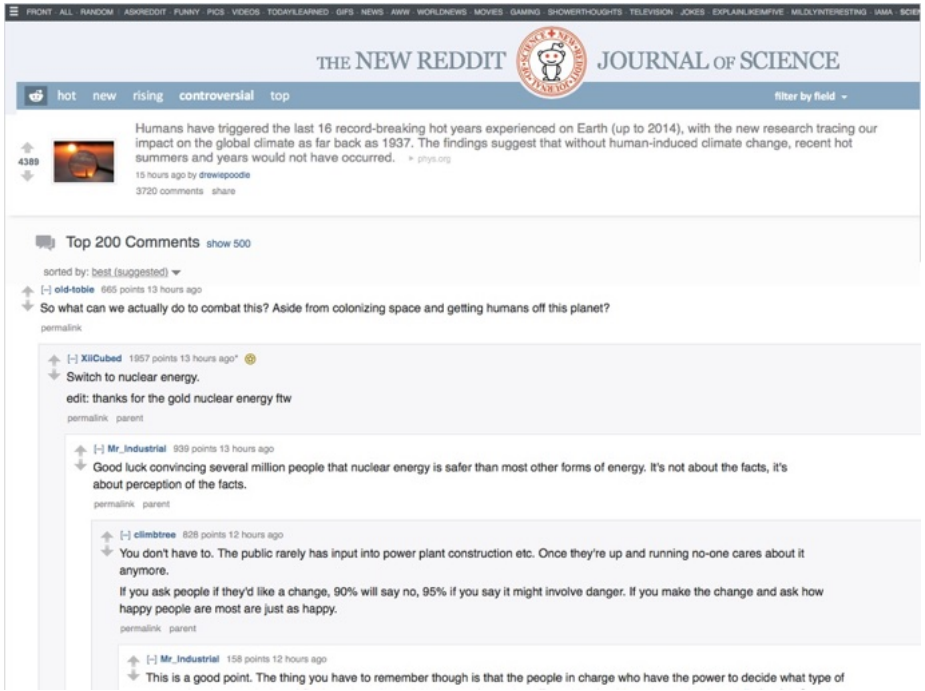
- EDA = analyze data to formulate sound hypotheses
- CDA = test those hypotheses
  - statistical tests, variance analysis…
  - machine learning approach (eg., regression)

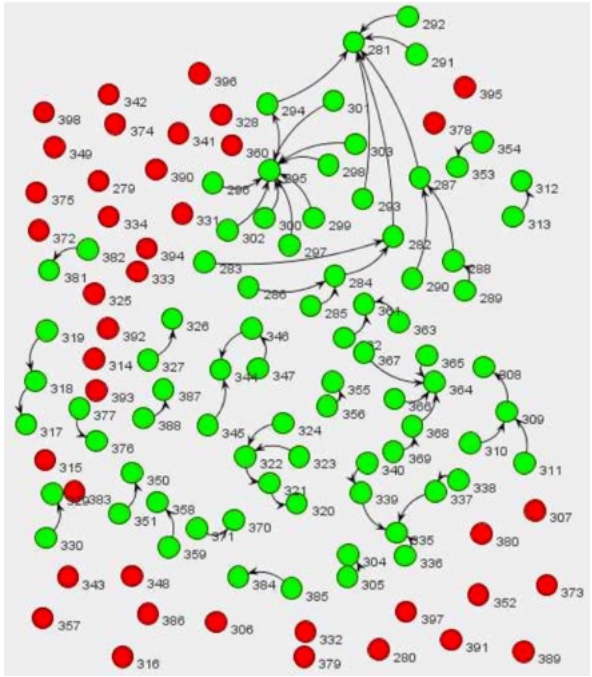
then:

- use models for prediction
- make sound decisions

We Need Both Exploratory and Confirmatory, J.W. Tukey, The American Statistician, Vol. 34, No. 1, pp. 23-25, 1980

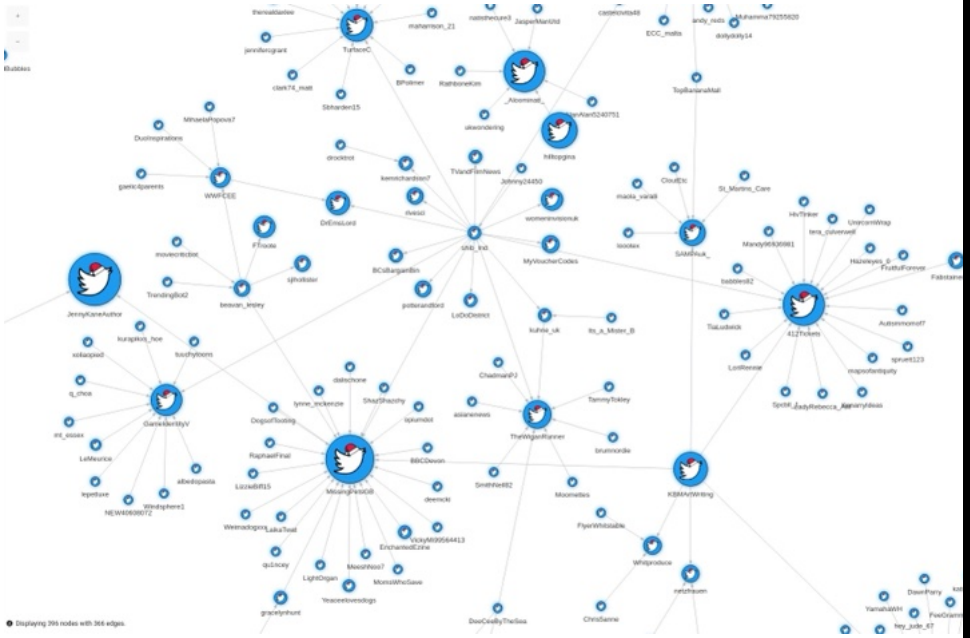
Illustration 1

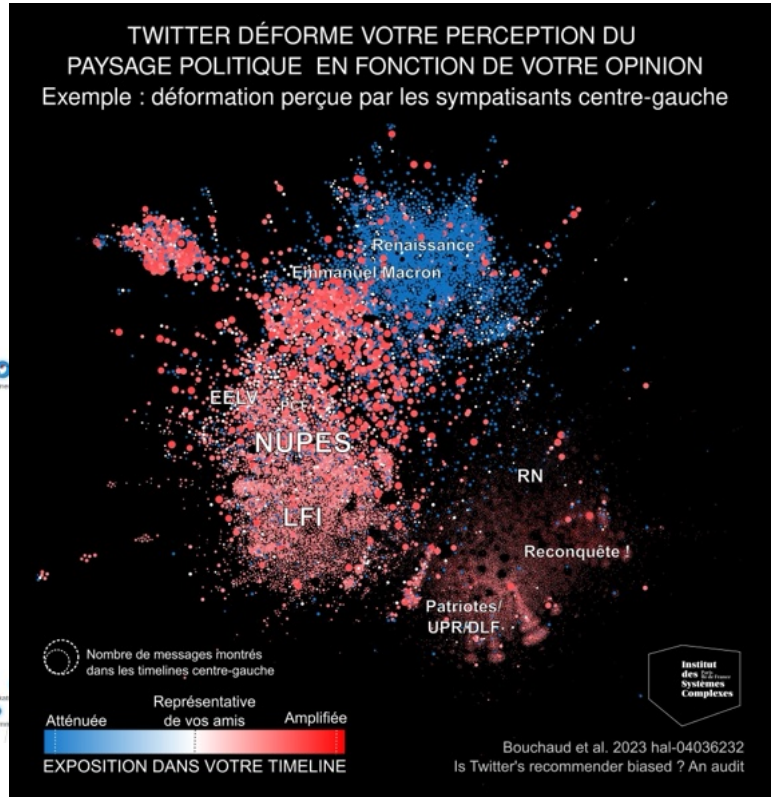




Work of Anna Stavrianou (PhD 2010) and Mathilde Forestier (PhD 2012)

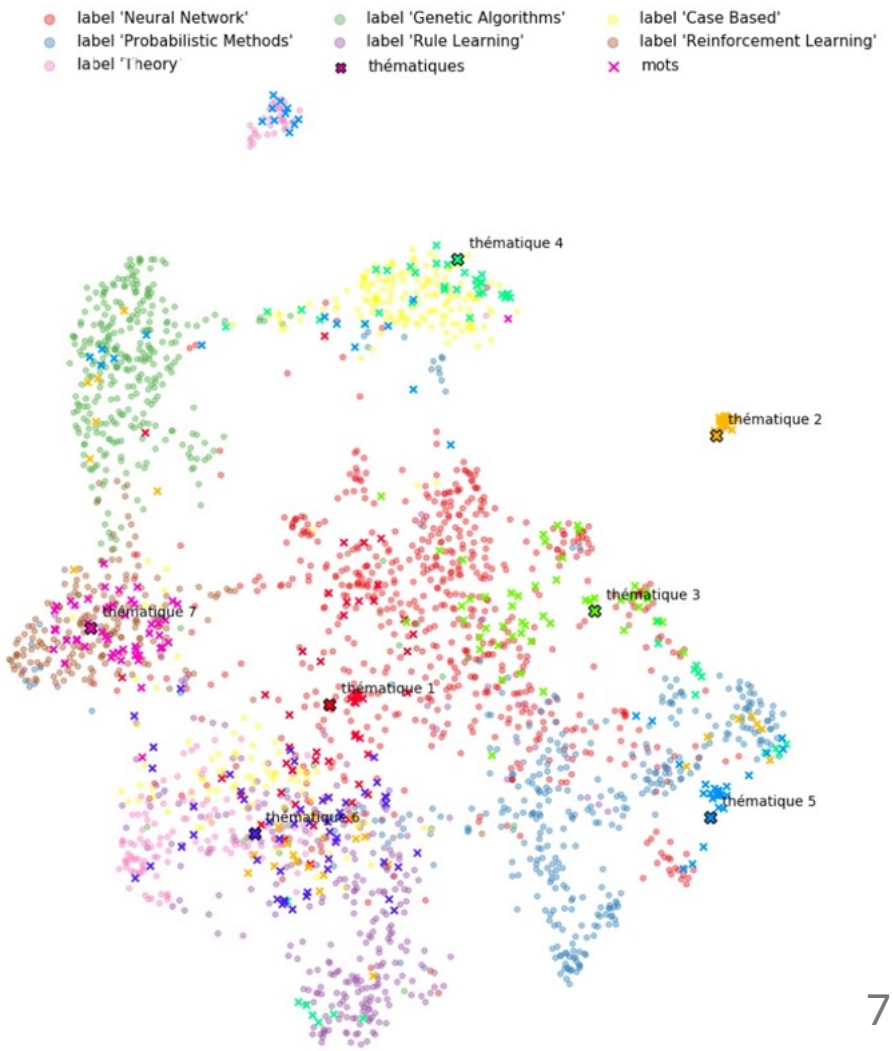
Illustration 2





Politoscope (https://politoscope.org)

Illustration 3



Work of Robin Brochier (PhD 2020)

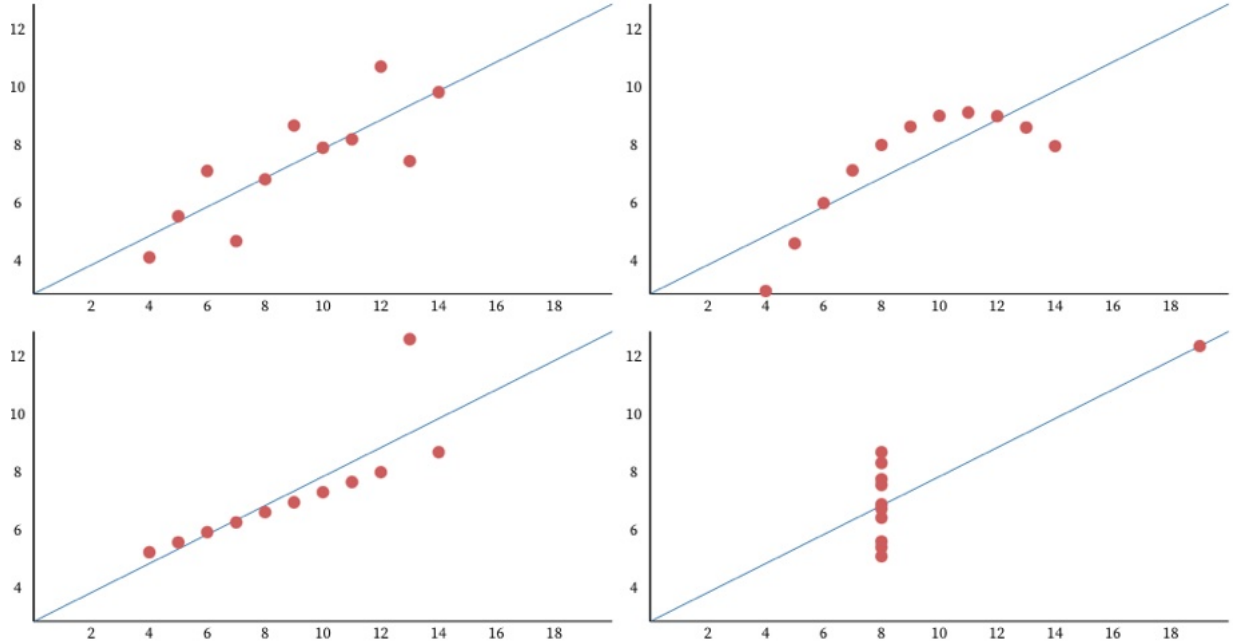
Illustration from Francis Anscombe (1973)

dataset 1 dataset 2 dataset 3 dataset 4

A		B		C		D	
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	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	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

Graphs in Statistical Analysis, F.J. Anscombe, The American Statistician, Vol. 27, No. 1, pp. 17-21, 1973

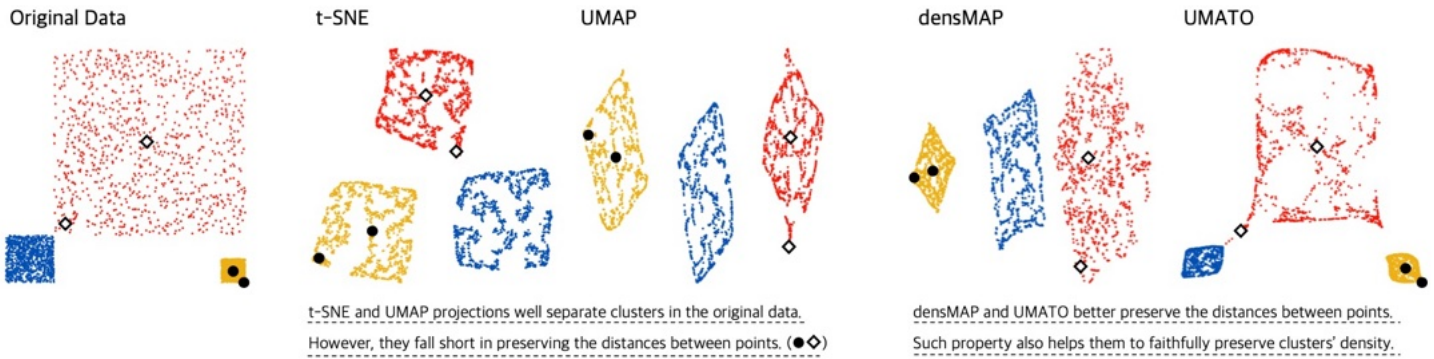
If we look at the datasets...



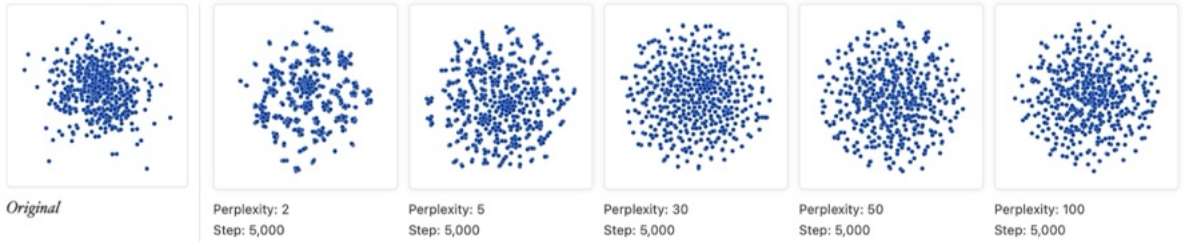
But...

	A	B	C	D
Moyenne des X	9	9	9	9
Variance des X	11	11	11	11
Moyenne des Y	7.5	7.5	7.5	7.5
Variance des Y	4.13	4.13	4.12	4.12
Coefficient de corrélation	0.82	0.82	0.82	0.82
Régression linéaire	y = 0.5x + 3	y = 0.5x + 3	y = 0.5x + 3	y = 0.5x + 3
Coefficient de détermination	0.67	0.67	0.67	0.67

Visualization fallacy (1)

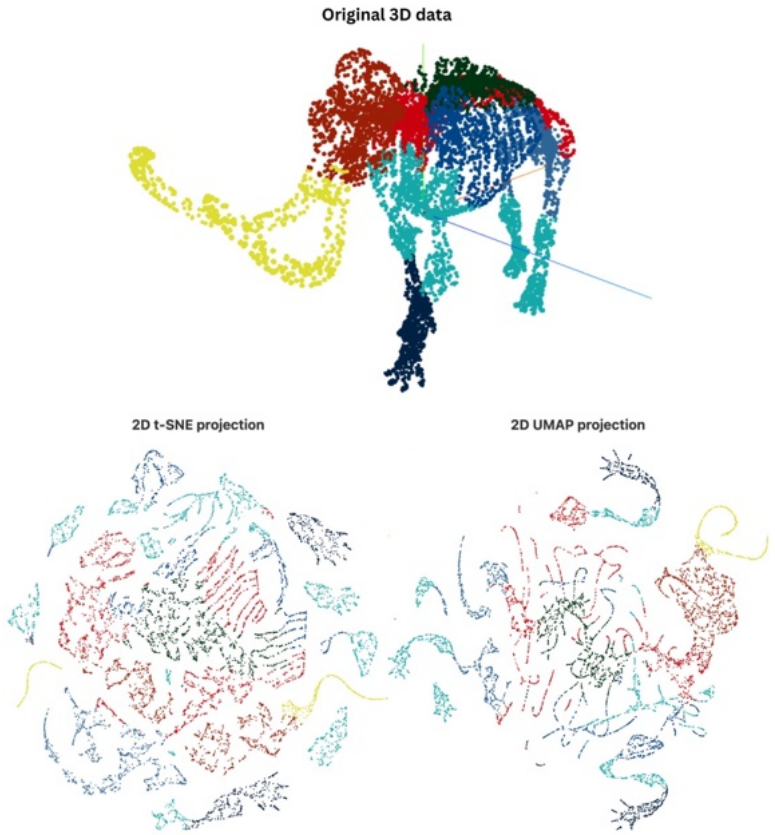


Stop Misusing t-SNE and UMAP for Visual Analytics (Jeon et al., 2025), https://arxiv.org/html/2506.08725v2

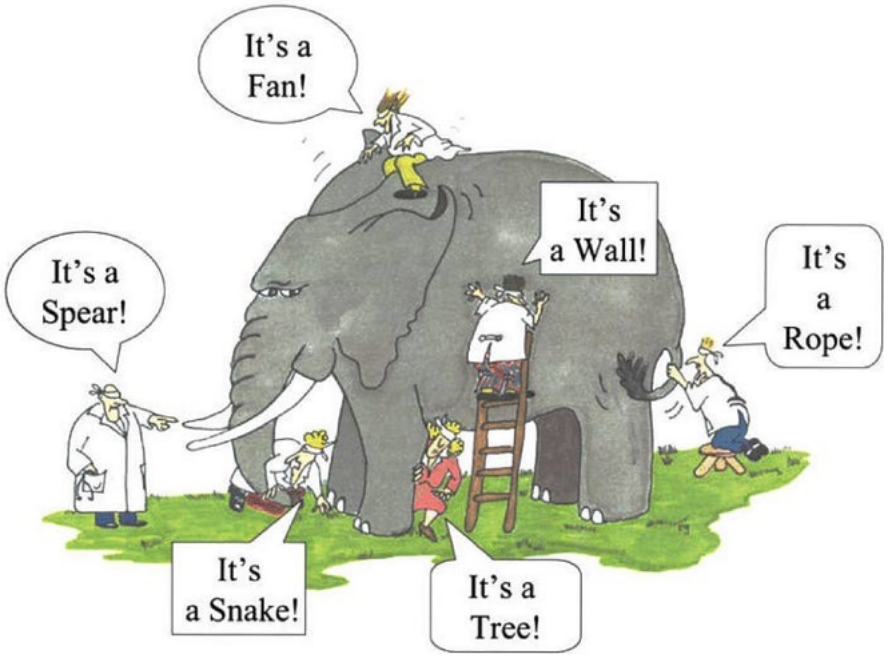


How to Use t-SNE Effectively, https://distill.pub/2016/misread-tsne/

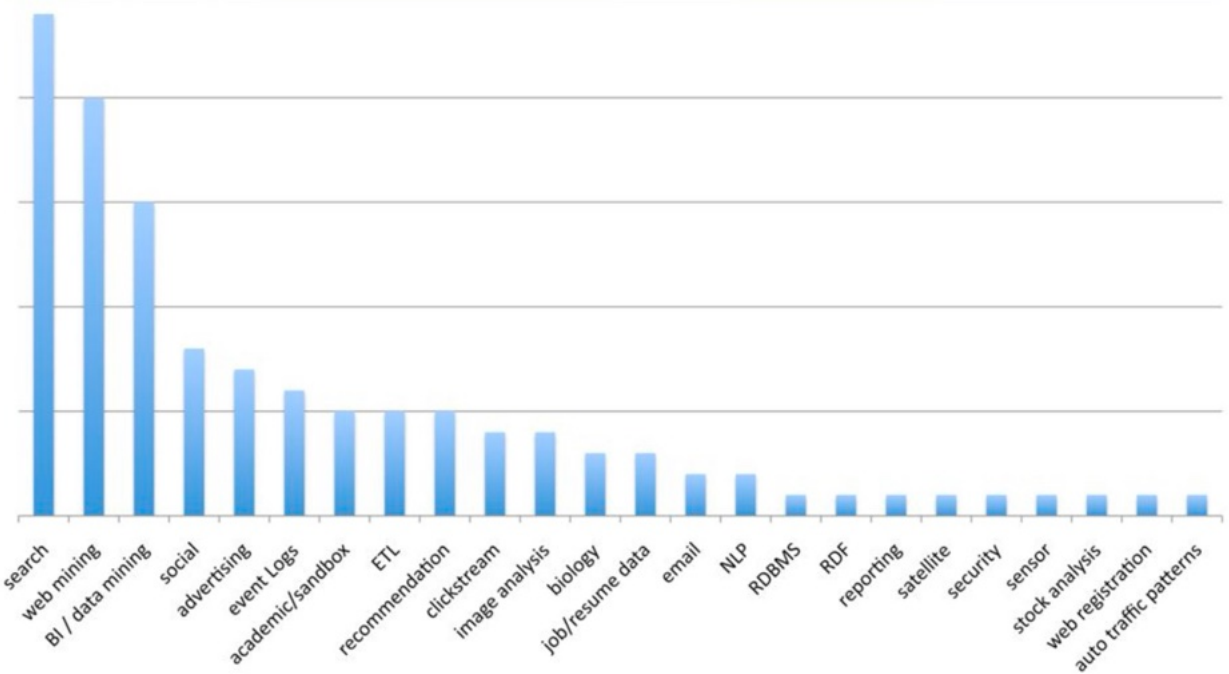
Visualization fallacy (2)



Always use various perspectives



Where do we need data science?



Application domains of data science (Corvelle Consulting)

Introduction and data types

briefing and crime scene

EDA as a police investigation

- Victim / client = business problem to be solved
- Crime scene = raw dataset
- Suspects = variables / features
- Investigation = cleaning, preprocessing, analysis (eg., clustering)
- Verdict = final storytelling

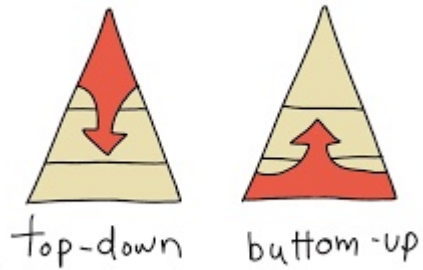
(idea from Tim Harford's book « The Data Detective » and… Gemini ;)

First of all: know your problem

- *Who* is asking?
- *What* the topic/domain?
- *Where* are the data located?
- *When* are the data produced?
- *Why* do we need EDA?

Different types of questions: Bottom-up vs top-down approches

- **Top-down** = answering precise questions
  - query in a database
  - information retrieval
  - supervised classification
- **Bottom-up** = finding interesting patterns
  - rules (relations between features)
  - clusters of similar objects
- Hybrid approches (cf. data processing pipeline)



Running example

- **What ?** Annual Database of Road Traffic Injury Accidents, published by the French « Observatoire national interministériel de la sécurité routière »
- **Who ?** Métropole de Lyon
- **Where ?** Annual Database of Road Traffic Injury Accidents (https://www.data.gouv.fr/datasets/)
- **When ?** 2018
- **Why ?** Compare the accidents by city, impact of the location (state of the ground…), role of lighting, is it an intersection?, etc.

Loading the crime scene

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Num_Acc	an	mois	jour	hrmn	lum	agg	int	atm	col	com	adr	gps	lat	long	dep
2	2.018E+11	18	1	24	1505	1	1	4	1	1	5	route des Ansereuilles	M	5055737	294992	590
3	2.018E+11	18	2	12	1015	1	2	7			11	Place du général de Gaul	M	5052936	293151	590
4	2.018E+11	18	3	4	1135	1	2	3			477	Rue nationale	M	5051243	291714	590
5	2.018E+11	18	5	5	1735	1	2	1	7	3	52	30 rue Jules Guesde	M	5051974	289123	590
6	2.018E+11	18	6	26	1605	1	2	1	1	3	477	72 rue Victor Hugo	M	5051607	290605	590
7	2.018E+11	18	9	23	630	2	2	1	2	6	52	D39	M	5052132	288837	590
8	2.018E+11	18	9	26	40	5	2	1	1	6	133	4 route de camphin	M	5052211	296652	590
9	2.018E+11	18	11	30	1715	5	2	1	1	6	11	rue saint exupéry	M	5053146	293875	590
10	2.018E+11	18	2	18	1557	1	1	1	1	6	550	rue de l'égalité	M	5053707	284896	590
11	2.018E+11	18	3	19	1530	1	2	2	1	1	51	face au 59 rue de Lille	M	5053639	281517	590
12	2.018E+11	18	5	28	1830	1	2	6	1	3	51	76 route de Lille	M	5054194	281987	590
13	2.018E+11	18	5	31	430	5	1	6	5	6	250	RN41-D 145	M	5058465	290868	590
14	2.018E+11	18	6	15	845	1	1	1	1	3	257	rue delval	M	5061613	283783	590
15	2.018E+11	18	7	19	1022	1	2	1	1	6	51	26 rue d'estaires	M	5053196	280251	590
16	2.018E+11	18	10	31	1945	3	1	1	1	6	303	rue chobourdin	M	5057886	285234	590
17	2.018E+11	18	1	15	725	2	2	2	1	6	398	Rue Nationale	M	5051098	312175	590
18	2.018E+11	18	2	9	1515	1	2	1	4	6	129	rue de l'égalité	M	5051293	316665	590
19	2.018E+11	18	4	7	1145	1	2	3	1	2	71	CD 917	M	5048030	313456	590

Doing your homework

Liste complète des champs avec le détail de leur contenu pour chaque fichier :

La rubrique CARACTERISTIQUES

Num_Acc
Numéro d'identifiant de l'accident

jour
Jour de l'accident

mois
Mois de l'accident

an
Année de l'accident

hrmn
Heure et minutes de l'accident

lum
Lumière : conditions d'éclairage dans lesquelles l'accident s'est produit
1 – Plein jour
2 – Crépuscule ou aube
3 – Nuit sans éclairage public
4 - Nuit avec éclairage public non allumé
5 – Nuit avec éclairage public allumé

dep
Département : Code INSEE (Institut National de la Statistique et des Etudes Economiques) du département suivi d'un 0 (201 Corse-du-Sud - 202 Haute-Corse)

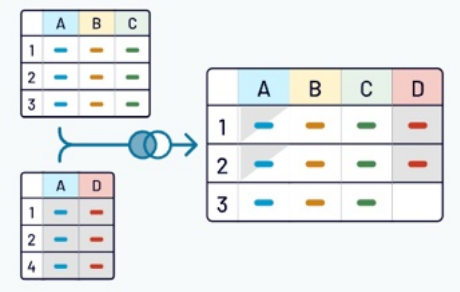
com

description of the data

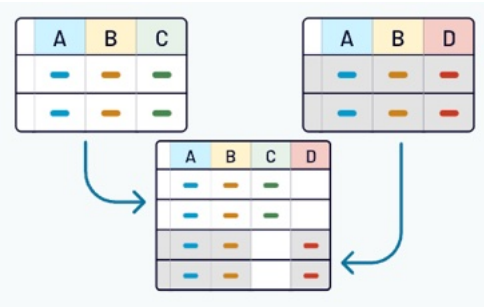
ACC-description-des-bases-de-donnees-onisr-annees-2005-a-2018.pdf

Gathering the evidence

Some examples:



append



join

Example using Pandas (1)

```
In [1]: df1 = pd.DataFrame(
   ...:     {
   ...:         "A": ["A0", "A1", "A2", "A3"],
   ...:         "B": ["B0", "B1", "B2", "B3"],
   ...:         "C": ["C0", "C1", "C2", "C3"],
   ...:         "D": ["D0", "D1", "D2", "D3"],
   ...:     },
   ...:     index=[0, 1, 2, 3],
   ...: )
   ...:

In [2]: df2 = pd.DataFrame(
   ...:     {
   ...:         "A": ["A4", "A5", "A6", "A7"],
   ...:         "B": ["B4", "B5", "B6", "B7"],
   ...:         "C": ["C4", "C5", "C6", "C7"],
   ...:         "D": ["D4", "D5", "D6", "D7"],
   ...:     },
   ...:     index=[4, 5, 6, 7],
   ...: )
   ...:

In [3]: df3 = pd.DataFrame(
   ...:     {
   ...:         "A": ["A8", "A9", "A10", "A11"],
   ...:         "B": ["B8", "B9", "B10", "B11"],
   ...:         "C": ["C8", "C9", "C10", "C11"],
   ...:         "D": ["D8", "D9", "D10", "D11"],
   ...:     },
   ...:     index=[8, 9, 10, 11],
   ...: )
   ...:

In [4]: frames = [df1, df2, df3]

In [5]: result = pd.concat(frames)

In [6]: result
Out[6]:
     A    B    C    D
0   A0   B0   C0   D0
1   A1   B1   C1   D1
2   A2   B2   C2   D2
3   A3   B3   C3   D3
4   A4   B4   C4   D4
5   A5   B5   C5   D5
6   A6   B6   C6   D6
7   A7   B7   C7   D7
8   A8   B8   C8   D8
9   A9   B9   C9   D9
10  A10  B10  C10  D10
11  A11  B11  C11  D11
```

df1

	A	B	C	D
0	A0	B0	C0	D0
1	A1	B1	C1	D1
2	A2	B2	C2	D2
3	A3	B3	C3	D3

df2

	A	B	C	D
4	A4	B4	C4	D4
5	A5	B5	C5	D5
6	A6	B6	C6	D6
7	A7	B7	C7	D7

df3

	A	B	C	D
8	A8	B8	C8	D8
9	A9	B9	C9	D9
10	A10	B10	C10	D10
11	A11	B11	C11	D11

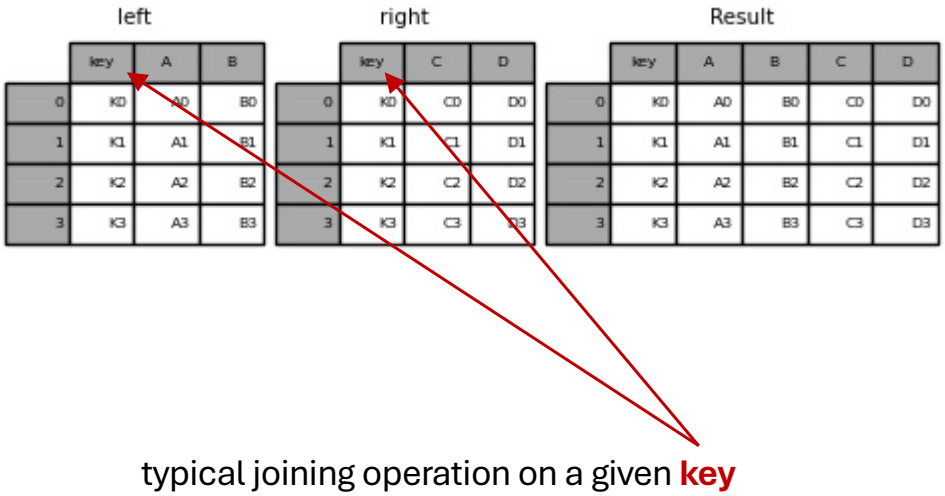
Result

	A	B	C	D
0	A0	B0	C0	D0
1	A1	B1	C1	D1
2	A2	B2	C2	D2
3	A3	B3	C3	D3
4	A4	B4	C4	D4
5	A5	B5	C5	D5
6	A6	B6	C6	D6
7	A7	B7	C7	D7
8	A8	B8	C8	D8
9	A9	B9	C9	D9
10	A10	B10	C10	D10
11	A11	B11	C11	D11

https://pandas.pydata.org/docs/user_guide/merging.html

Example using Pandas (2)

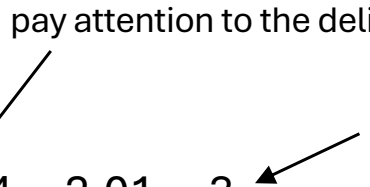
```
In [44]: left = pd.DataFrame(
    ....:     {
    ....:         "key": ["K0", "K1", "K2", "K3"],
    ....:         "A": ["A0", "A1", "A2", "A3"],
    ....:         "B": ["B0", "B1", "B2", "B3"],
    ....:     }
    ....: )
    ....:

In [45]: right = pd.DataFrame(
    ....:     {
    ....:         "key": ["K0", "K1", "K2", "K3"],
    ....:         "C": ["C0", "C1", "C2", "C3"],
    ....:         "D": ["D0", "D1", "D2", "D3"],
    ....:     }
    ....: )
    ....:

In [46]: result = pd.merge(left, right, on="key")

In [47]: result
Out[47]:
  key   A   B   C   D
0  K0  A0  B0  C0  D0
1  K1  A1  B1  C1  D1
2  K2  A2  B2  C2  D2
3  K3  A3  B3  C3  D3
```



typical joining operation on a given **key**

https://pandas.pydata.org/docs/user_guide/merging.html

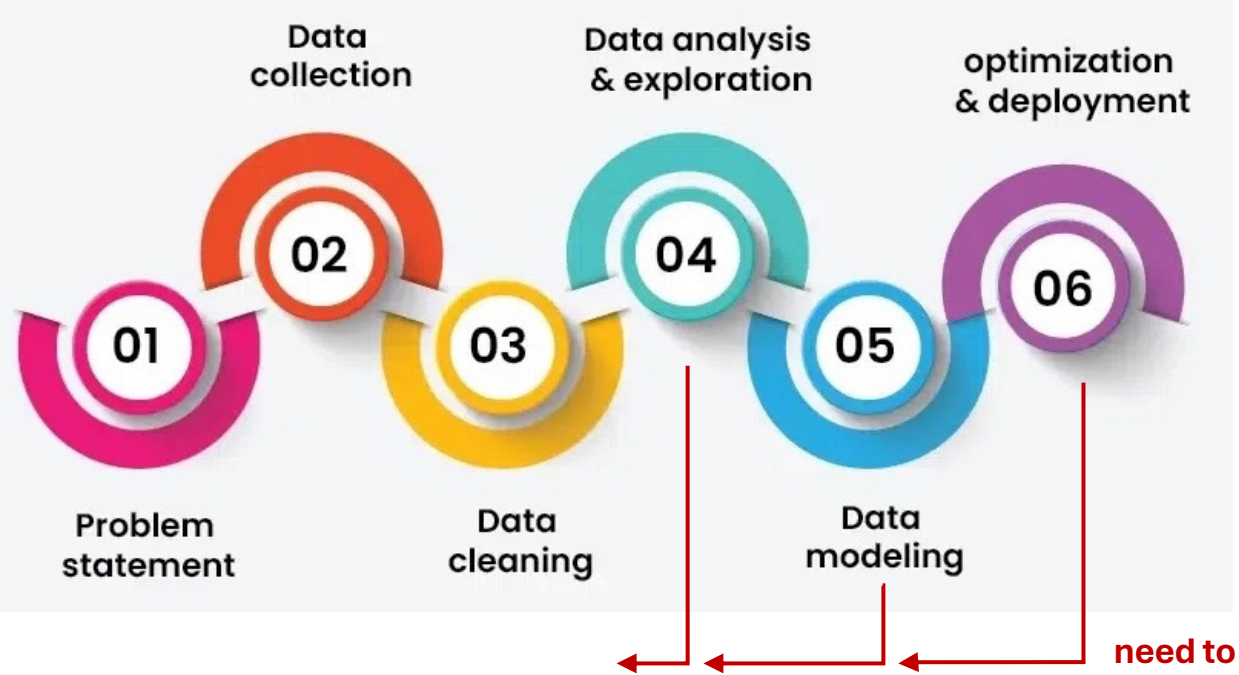
Variable types



- Numerical
  - continuous -10.24 2.01 3. $x \in \mathbb{R}$
  - discrete 1 2 10 -42 $x \in \{1, 2.., 6\}$
- Categorical
  - nominal rouge bleu vert color, job
  - ordinal un_peu beaucoup passionnément value scale
- Textual...

Each type needs its own processing and plots.

The data processing pipeline



Data cleaning and preprocessing

securing the scene

Data formatting

- Data can be stored in multiple formats:
  - simple text files .txt
  - tabular text files .csv .tsv
  - Excel files .xls .xlsx
  - document files .doc .docx .rtf .pdf
  - json files .json
  - images .jpg .jpeg .gif

  and so many others: .html .xml etc.

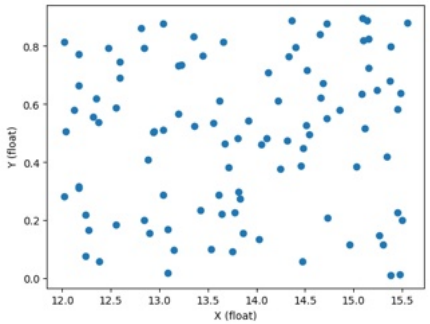
  (not even mentioning structured databases, see the lecture of Z. Bouyahya)
- Usually they can be easily loaded, but…
  - encoding issues (for text files, favor UTF-8 when possible)
  - size: really big files need specific processing (use the right library)
  - irregular: when your data don't really fit in a "natural" table

Illustration of « irregular » data

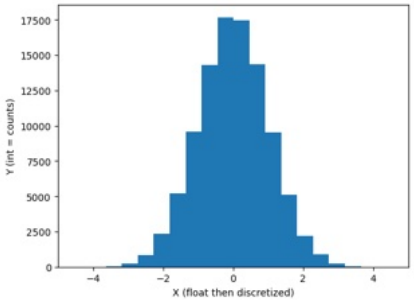
12223, University
12227, bridge, Sky
12828, Sunset
13801, Ground
14853, Tranceamerica
14854, San Francisco
15595, shibuya, Shrine
16126, fog, San Francisco
16520, California, ocean, summer, golden gate, beach, San Francisco

Fixing data types

- Pay attention to language conventions
  - 12,5 in FR means 12.5 in US
  - 14,380 in US means 14 380 in FR
- Cast int/long to float/double and the reverse way around
  - 13.0 → 13
  - 13.72 → 13 (floor) or 14 (ceiling) ?
  - 13 → 13.0
- Text or categorical data?
  - “firefighter” ≠ firefighter



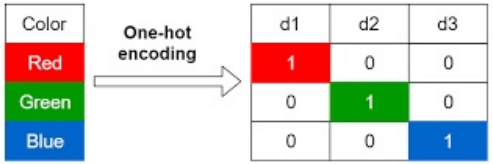
scatter plot



histogram

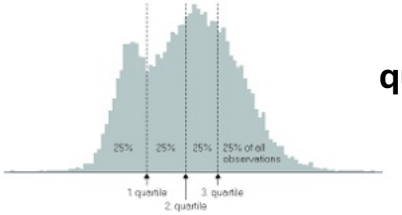
Make homogeneous features (1)

- Most analysis techniques need homogeneous data types (such as only textual information **xor** float variables)
- How to map categorical variables into numeric variables? →



One Hot Encoding (OHE)

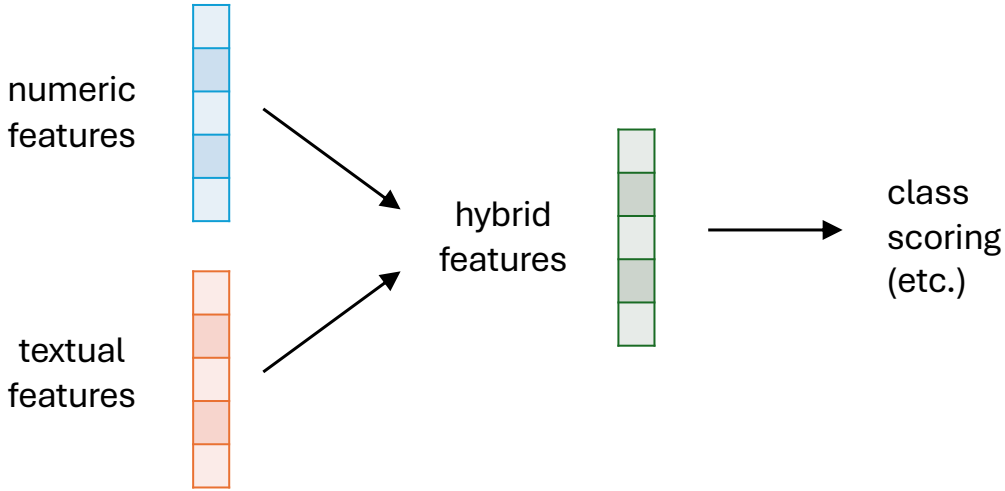
- How to map numeric variables into categorical variables? ←



quantile = interval with same population

Make homogeneous features (2)

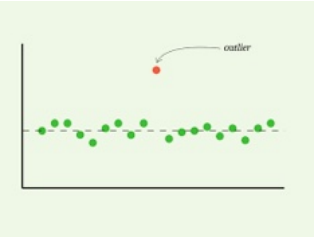
- Generally, it's all depend on the problem nature.
- For instance, currently in AI:



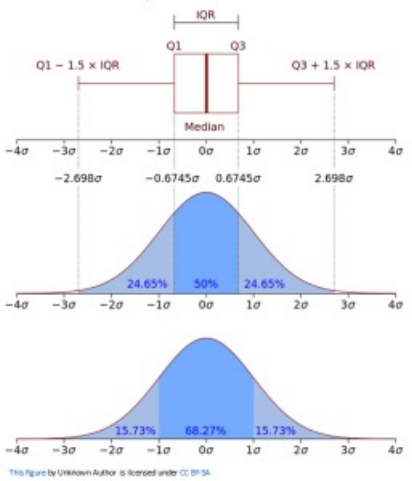
Missing values

- **Various reasons**: simple oversight, typos / data entry errors, change in data collection method, measurement impossibility…
- Finding the missing value is called **imputation** and can be done in different ways:
  - ‘missing’ / N/A as a **dedicated modality**
  - **delete** the corresponding feature for all objects, or delete the objects
  - **mean value** (continuous) / **most frequent value** or mode (categorical)
  - **prediction** based on other variables using a machine learning algorithm

Outliers / anomalies



- **Differenciate** outliers (rare values / patterns) from anomalies (abnormal behavior, error)
- -1 can mean ‘absent’ or the outlier can be caused by a data entry error (see “missing values”)
- **detection** can be easy by using simple statistic methods, such as IQR (IQR = Q3 - Q1): if the value is below Q1 - (1.5*IQR) or above Q3 + (1.5*IQR)
- careful **examination** of the detected cases
- specific treatment of the outliers or using **corrections** that are usually ad-hoc



Duplicates, inconsistencies

- **Repetition** of the same information usually impacts (biases) the analysis outcome

  You have to question if you consider that information is *twice* as important because it appears *twice* in your dataset

- **Inconstitencies** are related to the coherence / commensurability of your data:
  - numeric variable with different scales → same scale
  - textual data with different words for the same meaning → same token

  Usuall inconsistencies can be fixed by using **normalization**

Normalizing values

- **Objective:** to make the values comparable, and not favor a variable over another because of their variation range
- Different ways to achieve normalization:
  - simple normalization between 0 – 1 : $\frac{x_i - \min(X)}{\max(X) - \min(X)}$
  - Center and reduce so that μ = 0 and σ = 1 : $\frac{x_i - \mu}{\sigma}$

Handling textual data

This topic deserves a whole course on its own, but briefly we can deal with a string in two ways:

- value **as a categorical data**, in particular if you have a limited set of possible values (eg., color, job…)

→ you can use a simple OHE

limitation: no similarity/distance between values (exact match)

- value **as a numeric vector** that encodes semantic of the text

→ you can use a semantic encoder (eg., SentenceTransformer lib)

limitation: can be expensive to compute, how do you articulate this information with the other features?

Data analysis

interrogating the suspects

Univariate statistics

For numeric variables:

df.describe()

mean

standard deviation

median

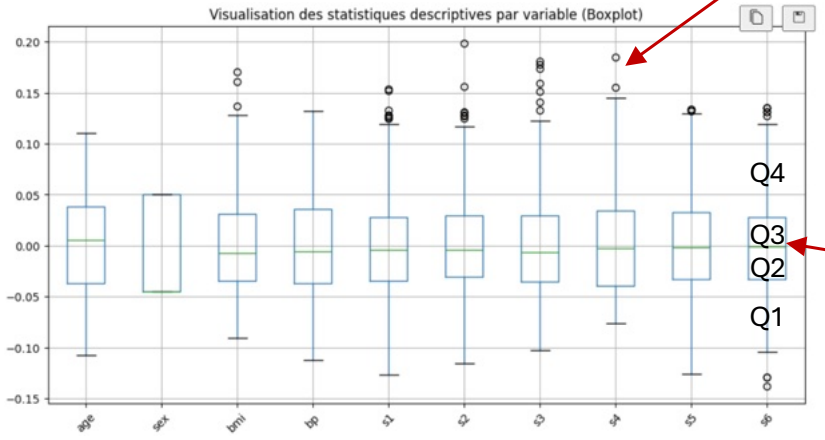
Statistiques descriptives :

	age	sex	bmi	bp	s1
count	4.420000e+02	4.420000e+02	4.420000e+02	4.420000e+02	4.420000e+02
mean	-2.511817e-19	1.230790e-17	-2.245564e-16	-4.797570e-17	-1.381499e-17
std	4.761905e-02	4.761905e-02	4.761905e-02	4.761905e-02	4.761905e-02
min	-1.072256e-01	-4.464164e-02	-9.027530e-02	-1.123988e-01	-1.267807e-01
25%	-3.729927e-02	-4.464164e-02	-3.422907e-02	-3.665608e-02	-3.424784e-02
50%	5.383060e-03	-4.464164e-02	-7.283766e-03	-5.670422e-03	-4.320866e-03
75%	3.807591e-02	5.068012e-02	3.124802e-02	3.564379e-02	2.835801e-02
max	1.107267e-01	5.068012e-02	1.705552e-01	1.320436e-01	1.539137e-01

	s2	s3	s4	s5	s6
count	4.420000e+02	4.420000e+02	4.420000e+02	4.420000e+02	4.420000e+02
mean	3.918434e-17	-5.777179e-18	-9.042540e-18	9.293722e-17	1.130318e-17
std	4.761905e-02	4.761905e-02	4.761905e-02	4.761905e-02	4.761905e-02
min	-1.156131e-01	-1.023071e-01	-7.639450e-02	-1.260971e-01	-1.377672e-01
25%	-3.035840e-02	-3.511716e-02	-3.949338e-02	-3.324559e-02	-3.317903e-02
50%	-3.819065e-03	-6.584468e-03	-2.592262e-03	-1.947171e-03	-1.077698e-03
75%	2.984439e-02	2.931150e-02	3.430886e-02	3.243232e-02	2.791705e-02
max	1.987880e-01	1.811791e-01	1.852344e-01	1.335973e-01	1.356118e-01

Visualizing univariate statistics (1)

```
df.boxplot(column=["Col1", "Col2", "Col3"])
```



boxplot (whisker plot)

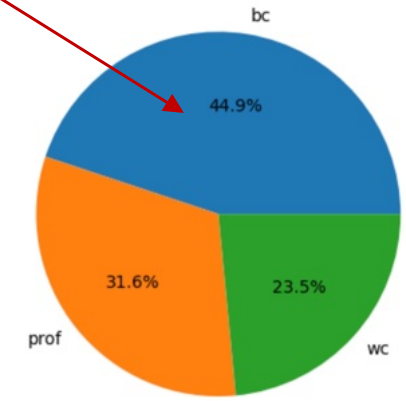
Visualizing univariate statistics (2)

```
df["Col1"].value_counts()
```

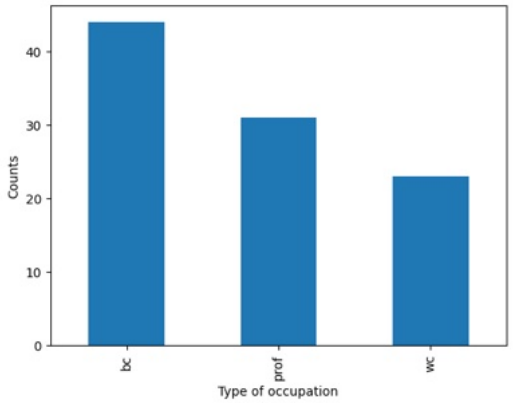
```
type
bc      44
prof    31
wc      23
Name: count, dtype: int64
```

mode

```
df["Col1"].value_counts().plot.pie(autopct='%1.1f%%')
....plot.bar(xlabel='Type of occupation', ylabel='Counts')
```



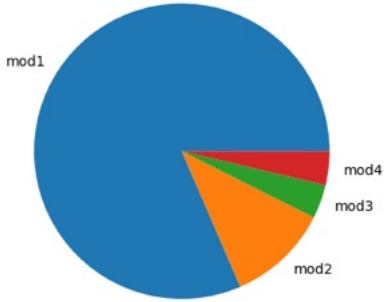
pie chart

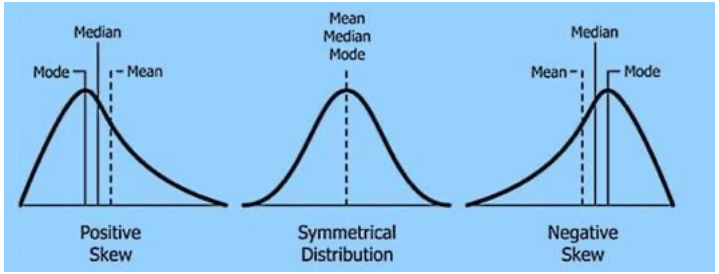


bar chart

Lessons we can draw

- Balance / skewness of the distribution

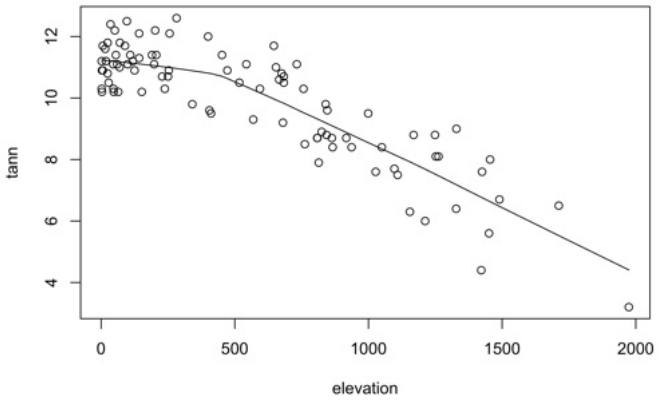




- What can you do?
  - data balancing (eg. SMOTE), data augmentation
  - data transformation
    (see https://medium.com/@lamunozs/dealing-with-high-skewed-data-a-practical-guide-part-iii-19fc38a10a7c)

Bivariate statistics (1)

- For numeric variables:

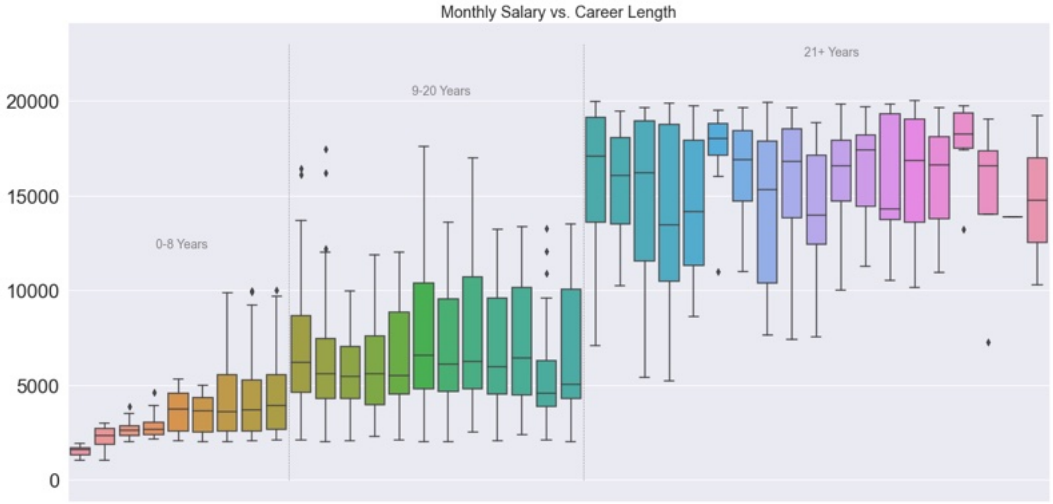


scatter plots

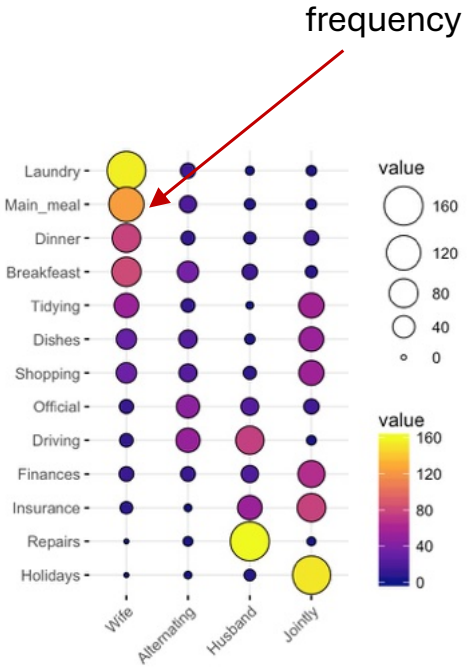
correlation heatmaps
(pearson, spearman...)

Bivariate statistics (2)

- For mixed variables:



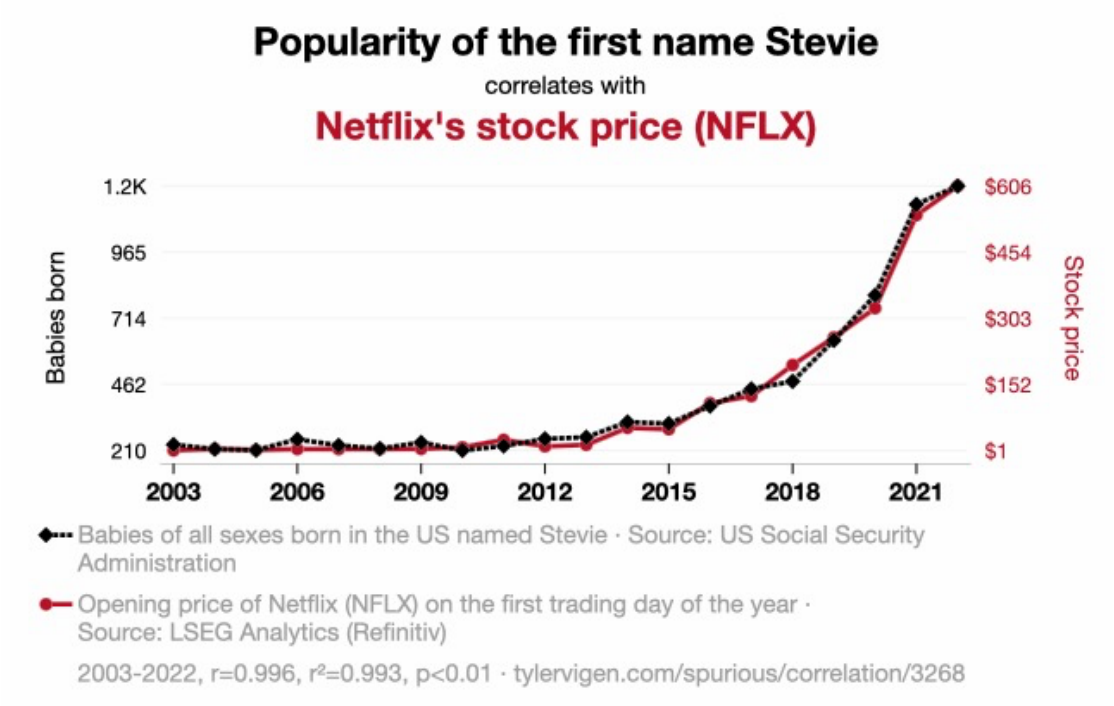
different modalities of the same variable (here, ordinal)



based on contingency tables

Beware spurious correlations

“Correlation doesn’t imply causation”



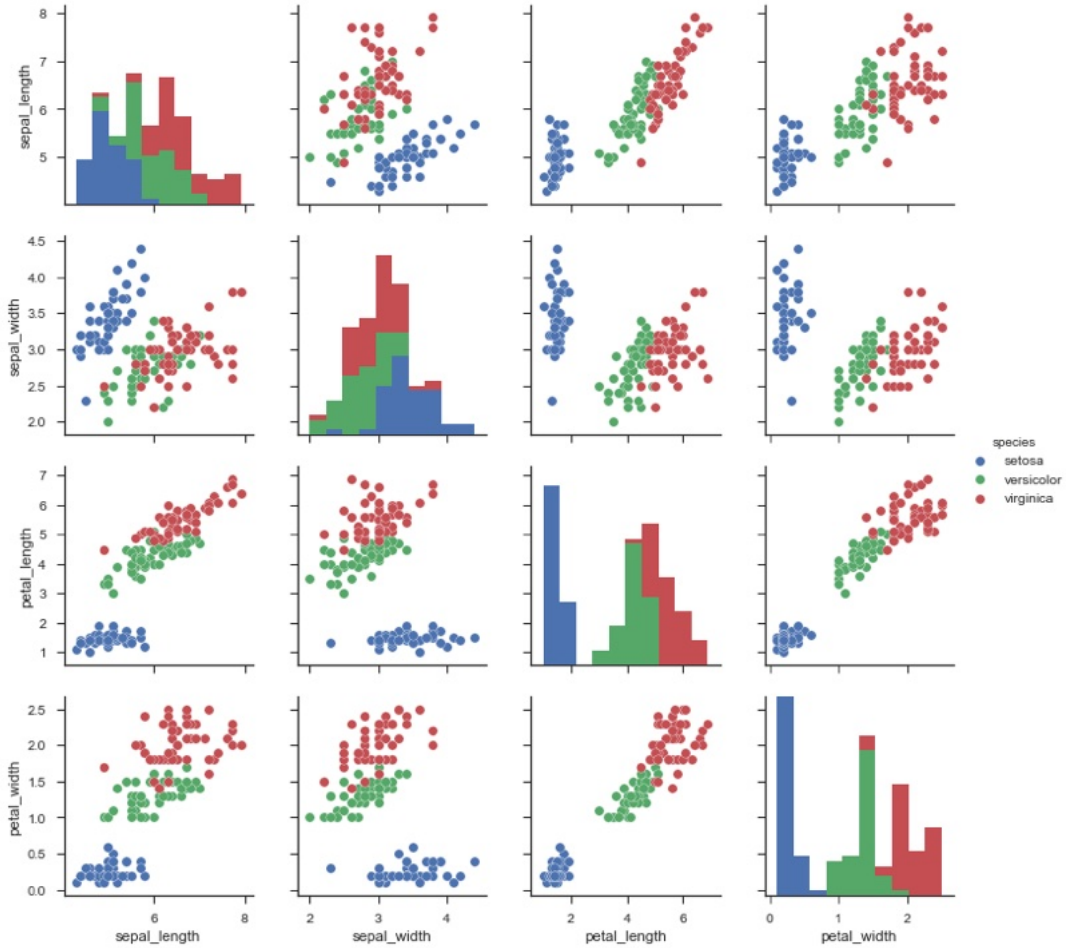
https://www.tylervigen.com/spurious-correlations

Multivariate analysis

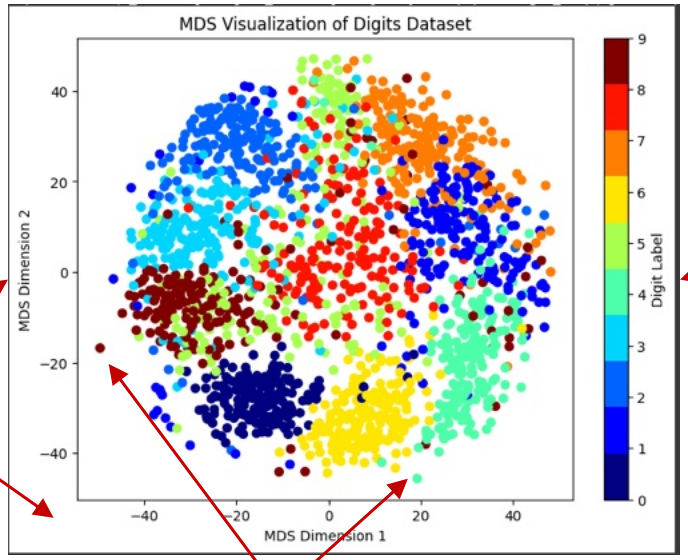
Beyond 2 variables, it's often difficult to give a simple visualization of the relations between variables

Possible solutions :

- projection techniques
- clustering algorithms



Projection to 2D (or even 3D) spaces

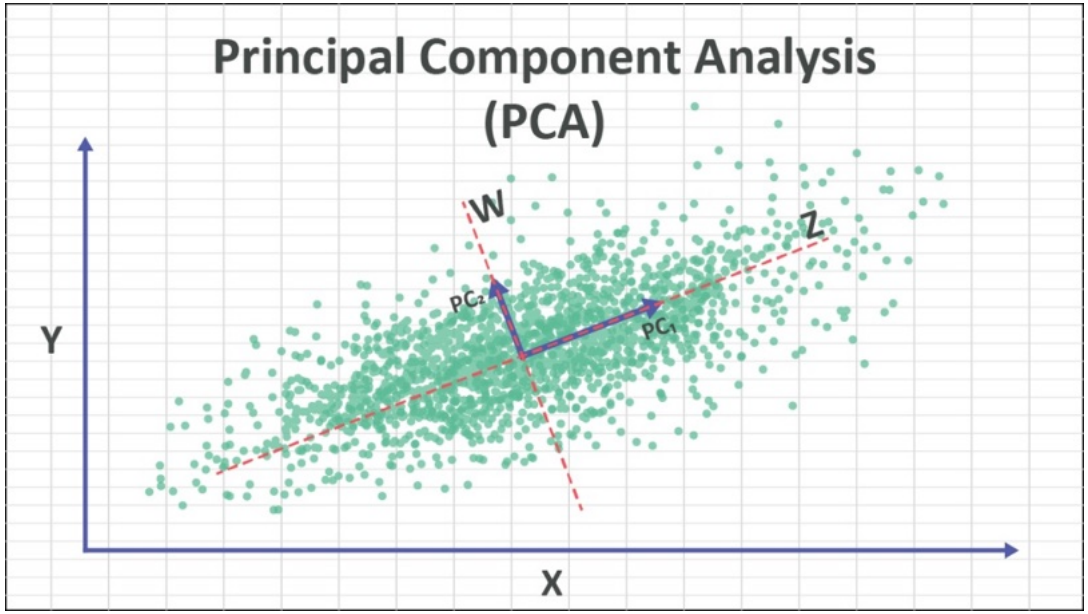


usually difficult to interprete The axes

meaning of the labels?

we can spot outliers

Principal Component Analysis (1)



Principal Component Analysis (2)

Specifically designed for numeric variables, PCA is based on a **spectral decomposition** (eigenvectors induced space) that best "explain" the variance observed in the data

weight of each variable

proportion of the variance explained

Principal component analysis					
	PC(1)	PC(2)	PC(3)	PC(4)	PC(5)
Variance	2.76	1.65	0.30	0.19	0.09
Propotion	55.2%	33.1%	6.1%	3.9%	1.8%
Cum. Propotion	55.2%	88.3%	94.3%	98.2%	100.0%

variables

Loadings	PC(1)	PC(2)	PC(3)	PC(4)	PC(5)
population	0.227	-0.657	-0.640	0.308	-0.109
median school yrs	0.503	0.324	-0.383	-0.605	-0.359
total employment	0.339	-0.587	0.426	-0.499	0.331
misc professional services	0.560	0.014	0.488	0.455	-0.491
median house value	0.516	0.344	-0.153	0.287	0.714

new space generated by PC(1), PC(2)… PC(5)

Values	PC(1)	PC(2)	PC(3)	PC(4)	PC(5)
	1.795	0.902	0.467	0.323	0.075
	-2.259	1.642	0.447	-0.477	-0.478
	-2.664	0.460	0.569	0.299	-0.131
	0.995	1.853	-0.166	-0.358	0.334
	0.747	1.707	-0.066	-0.013	0.467
	-1.480	-1.311	0.368	0.650	0.371
	-1.072	0.473	-1.545	0.511	-0.220
	-0.100	-1.137	-0.342	-0.589	0.095
	1.233	-0.933	-0.270	-0.216	-0.184
	3.282	-0.318	0.352	0.444	-0.486
	-0.689	-1.718	0.115	0.123	0.152
	0.211	-1.620	0.070	-0.698	0.004

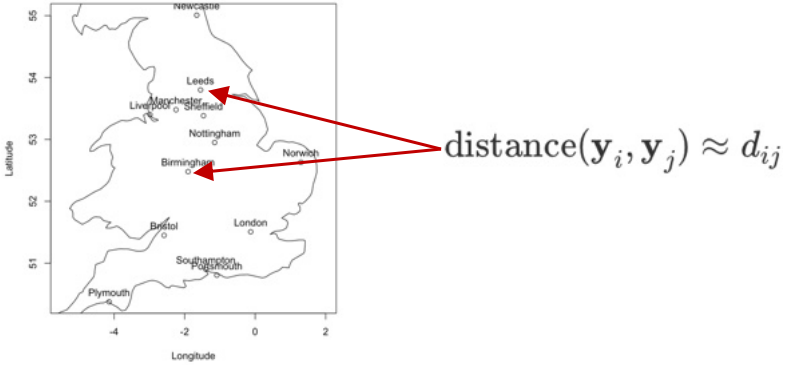
coordinates of example 1 in this new space

Lessons we draw

- High correlation between variables → reduce redundancy
- Scaling issues → rescale / normalize the variables
- Nonlinear relations btw variables → create new variables (eg., ^2)
- See preliminary clustering effects → calibrate further analyses

Multidimensional scaling (MDS)

- MDS compute 2D coordinates such that the **distance** in this new space reflects the distance in the original space



$$\text{distance}(\mathbf{y}_i, \mathbf{y}_j) \approx d_{ij}$$

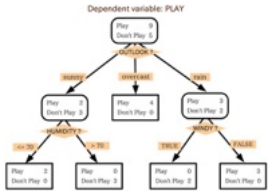
	London	Birmingham	Manchester	Leeds	Newcastle	Liverpool	Portsmouth	Southampton	Nottingham	Bristol	Sheffield	Norwich
London	0	163	262	273	403	286	103	112	175	170	228	159
Birmingham	163	0	114	99	282	125	195	179	73	124	105	217
Manchester	262	114	0	58	174	50	308	293	94	227	53	255
Leeds	273	149	58	0	135	105	335	323	98	271	47	230
Newcastle	403	282	174	135	0	199	469	458	231	401	181	328
Liverpool	286	125	50	105	199	0	317	299	132	219	101	299
Portsmouth	103	195	308	335	469	317	0	24	239	127	288	261
Southampton	112	179	293	323	458	299	24	0	229	103	276	268
Nottingham	175	73	94	98	231	132	239	229	0	193	53	169
Bristol	170	124	227	271	401	219	127	103	193	0	228	296
Sheffield	228	105	53	47	181	101	288	276	53	228	0	203
Norwich	159	217	255	230	328	299	261	268	169	296	203	0
Plymouth	308	281	369	420	542	346	221	202	353	162	382	453

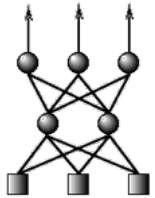
- Two famous algorithms: T-SNE et UMAP

One step further: modeling

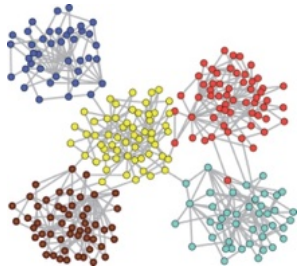
- All the exploration techniques can help in building a **model** of your data that answer your questions in a better way (eg. more robust)
- A model is a way to **simplify** the reality of your problem by linking the variables in a mathematical way

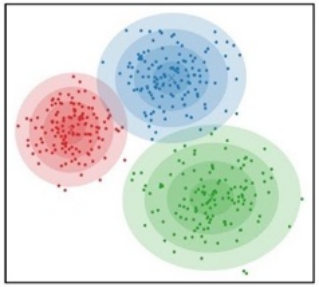
Supervised machine learning





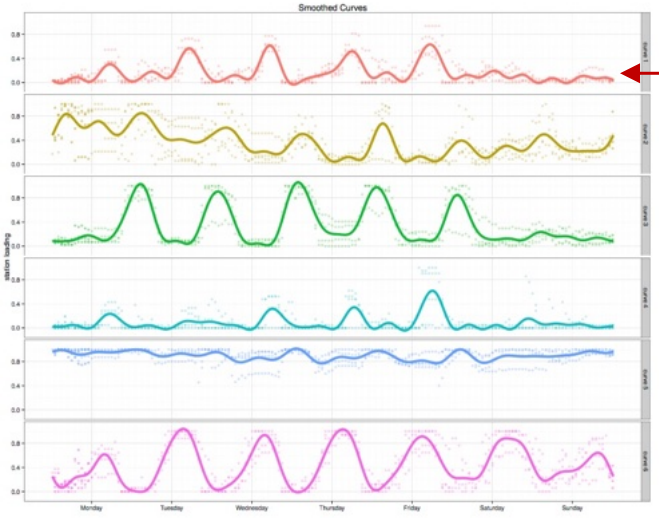
Clustering



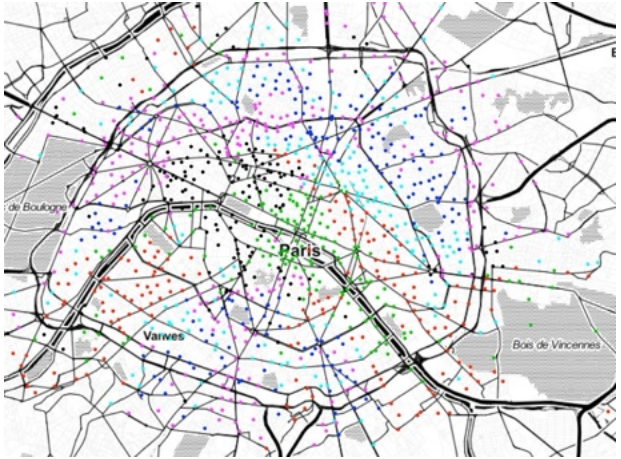


Clustering?

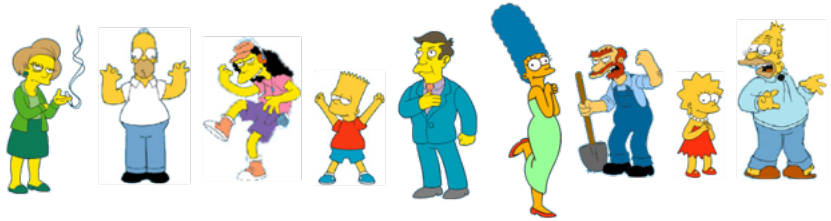
- Learn regularities (**patterns**) for building groups of similar objects

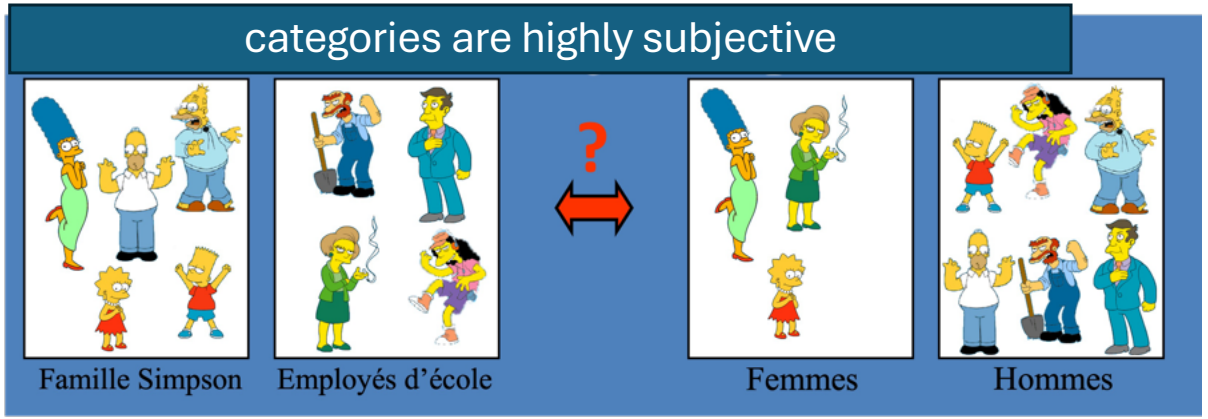


Here, example of a group of velib stations with a similar usage



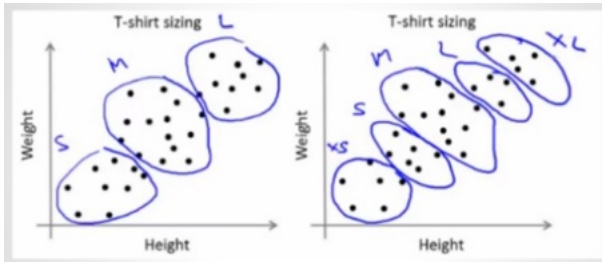
Difficulty of clustering objects



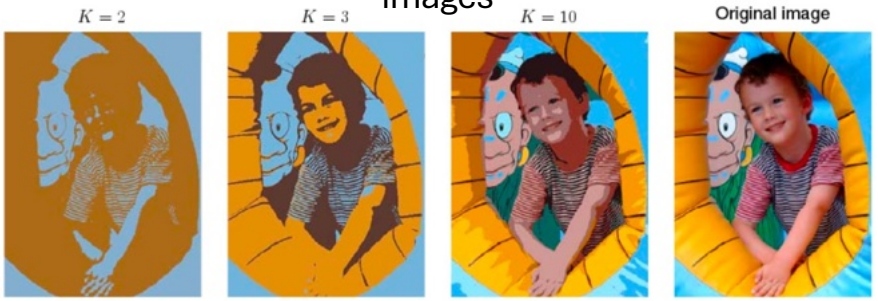


Useful for diverse kinds of data

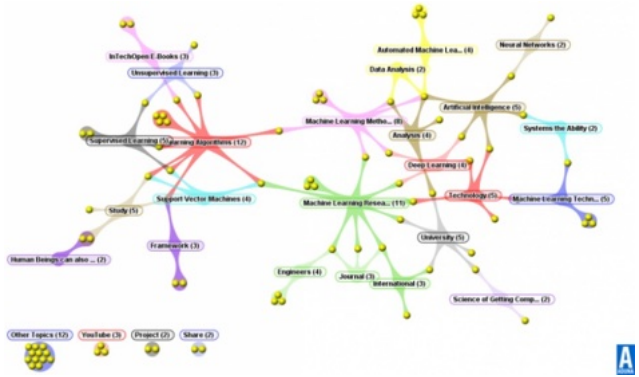
numeric data



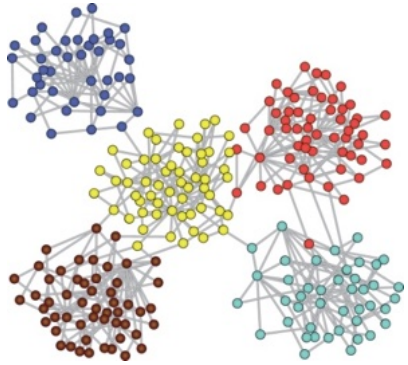
images



textual data



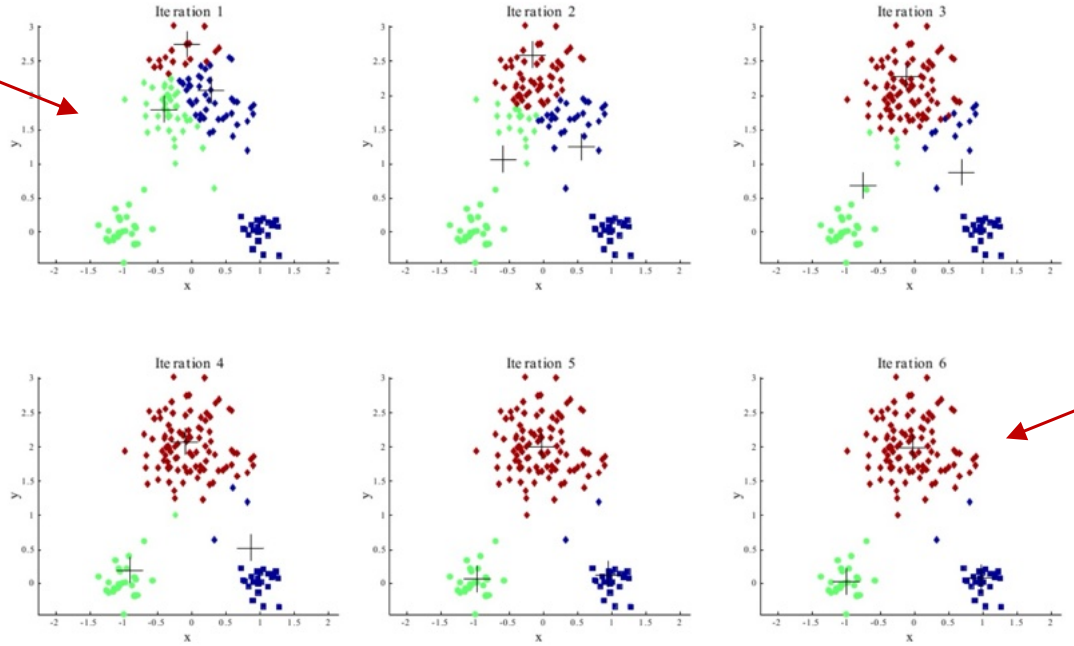
structured data



K-means algorithm

each step = compute the means and then reaffect the points

initial state (here, choose K=3)



get a stable partition at the end

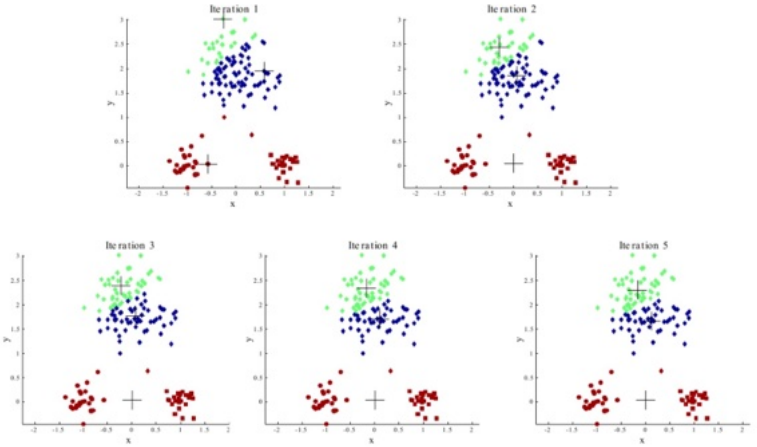
Pros and cons

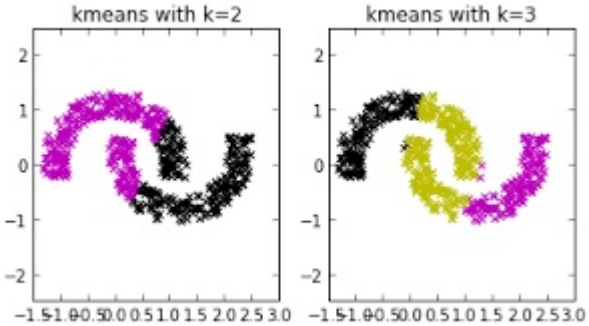
Pros:

- Simple idea
- Fast processing
- Many variants

Cons:

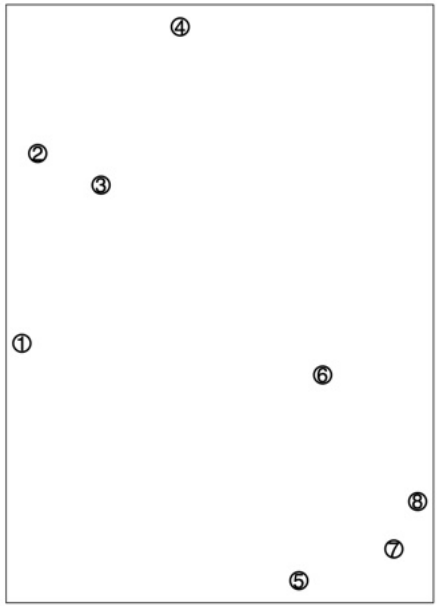
- Sensitive to initialization
- Restricted to convex (spherical) shapes

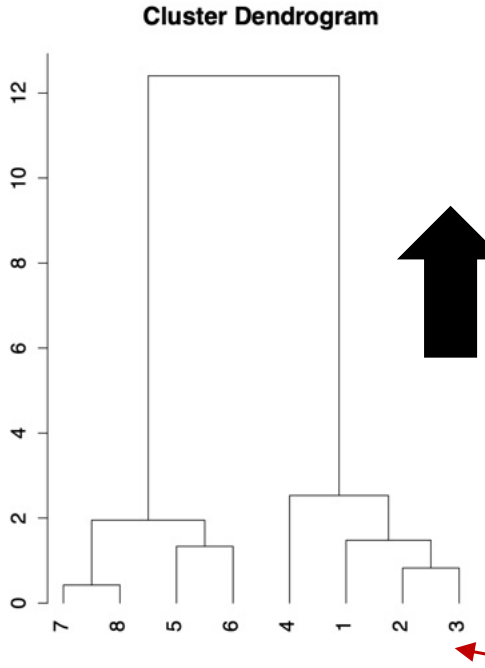




Hierarchical Agglomerative Clustering (HAC)

your data
(here, in 2D)





we build the clusters
by **merging** the points
one step at a time
(different ways to merge)

data points

Pros and cons

Pros:

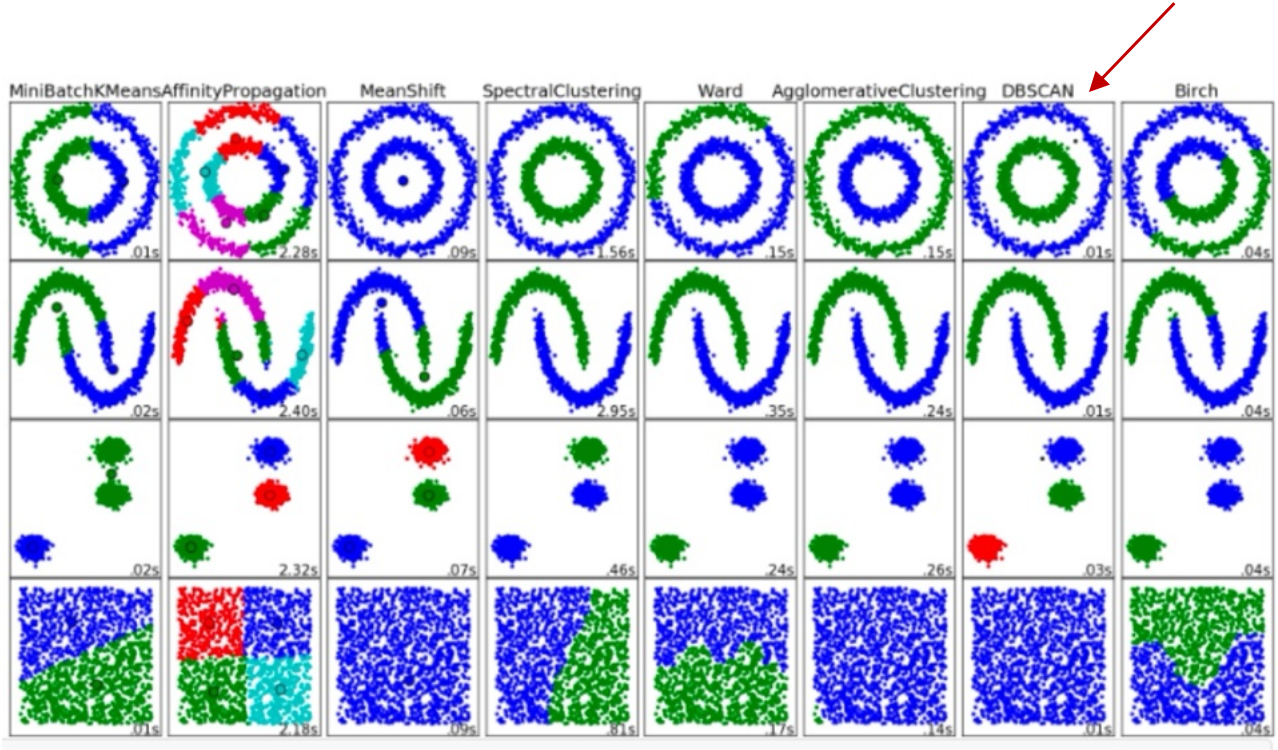
- Structuration in a hierarchy
- You can entail a partition at a given level

Cons:

- Choosing the heuristic for merging (single-link, max, average…)
- Quadratic cost (but you can initialize with KMeans)

DBScan

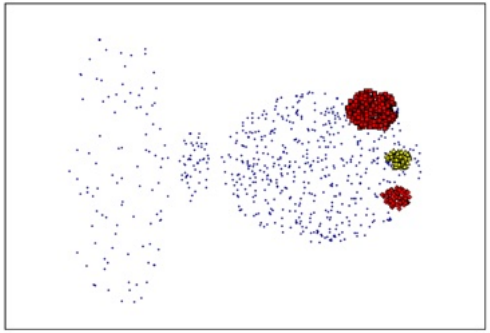
based on



Pros and cons

Pros:

- Simple idea
- Able to find complex shapes



(MinPts=4, Eps=9.75)

Cons:

- Difficulty to handle clusters of various densities
- Highly sensitive to its parameters (MinPts, Eps)

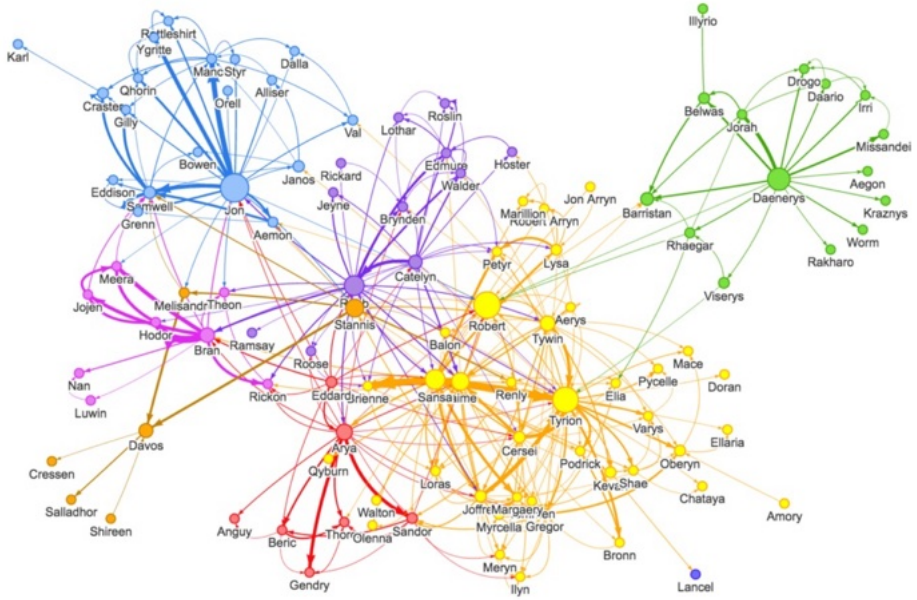
Difficulties in applying clustering

- Usually difficult to **set the parameters**, which can dramatically impact the final result (eg., metric)
- Find the “best” **number of clusters** (K)
- How do you **evaluate** the structure you get?

Lessons we draw

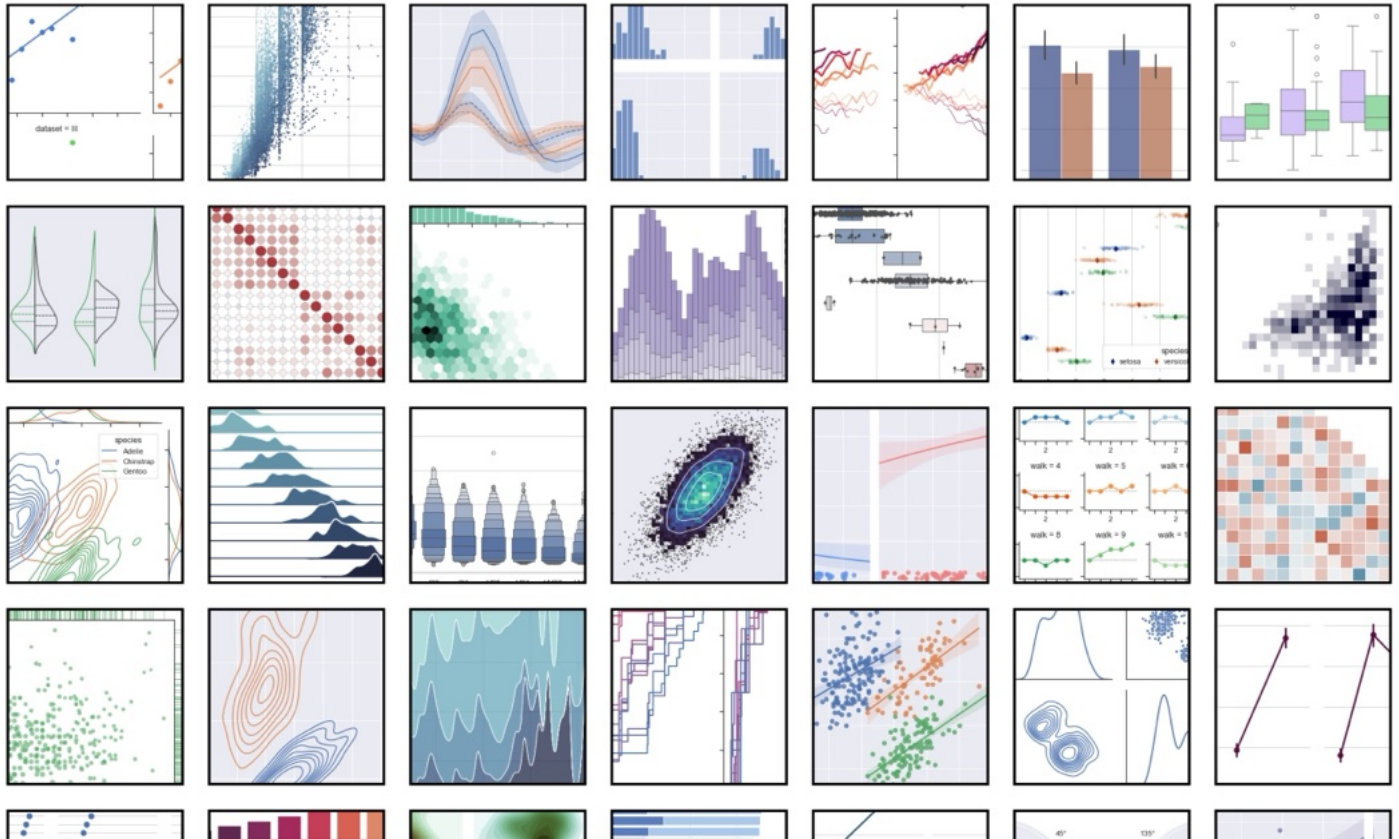
- Clustering gives us **new insight** on the data
- Clustering is a way to **summarize** the whole data, usually coupled with visualization tools
- Clustering can help us to **build a model** (eg., for prediction or for outlier detection)

Many possible visualization tools



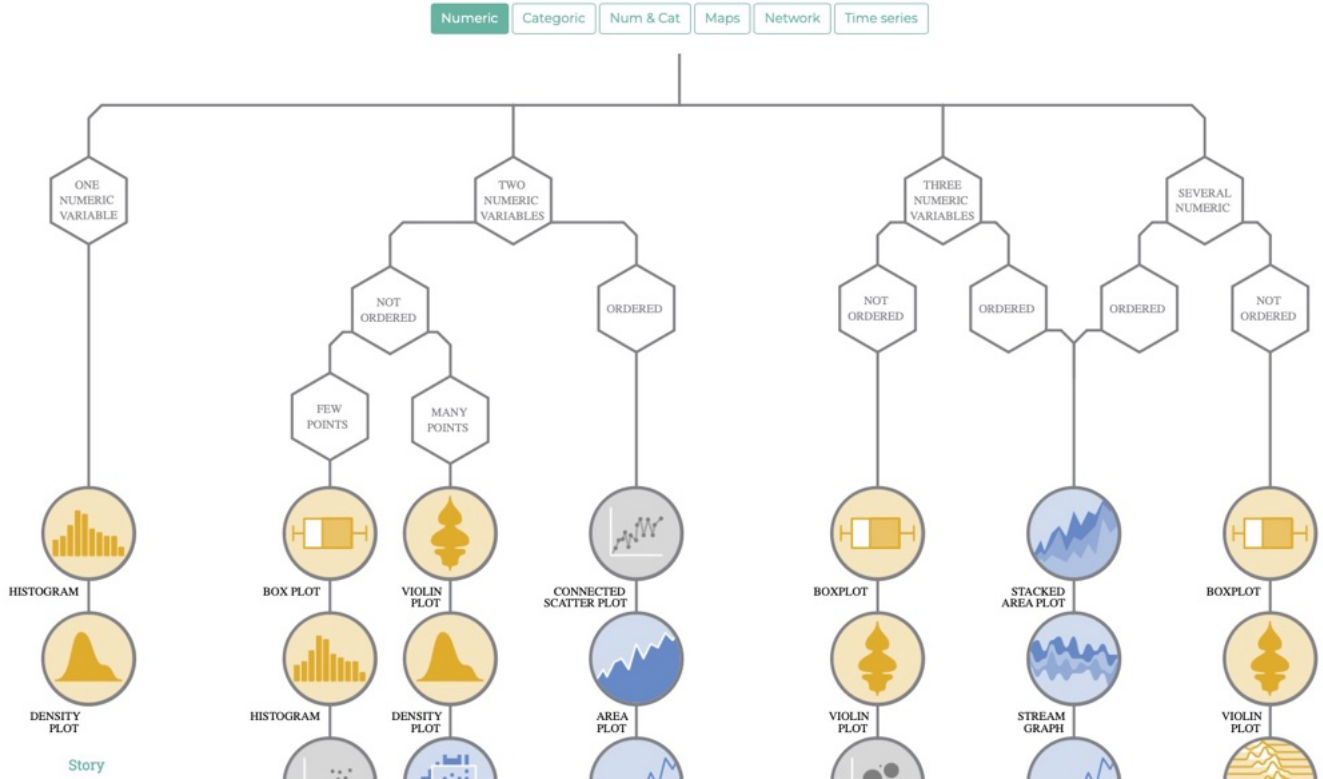
Many, many…

https://seaborn.pydata.org/examples/index.html

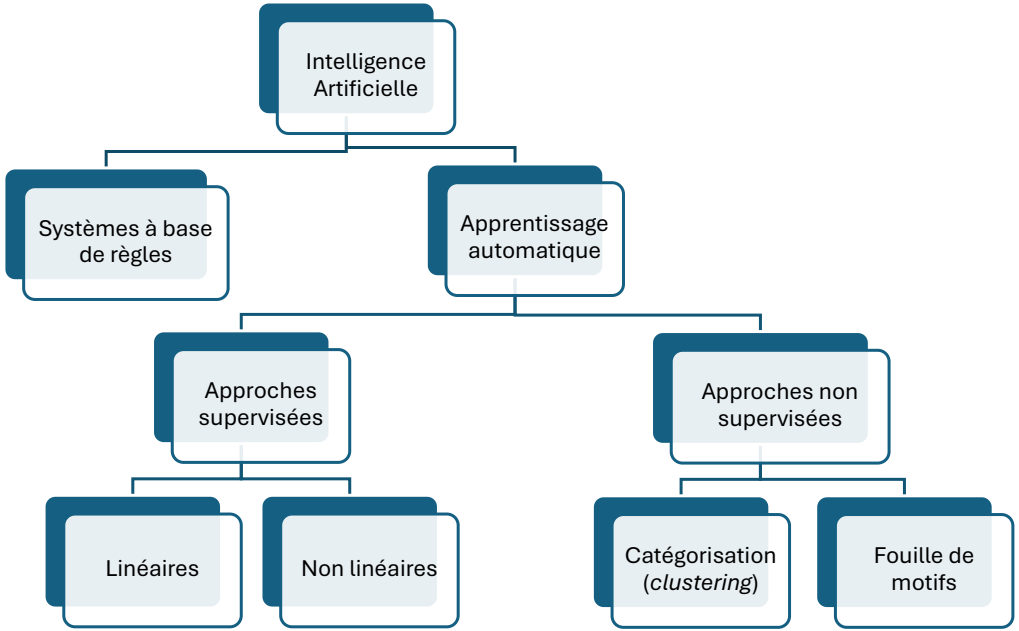


Some guides can help you

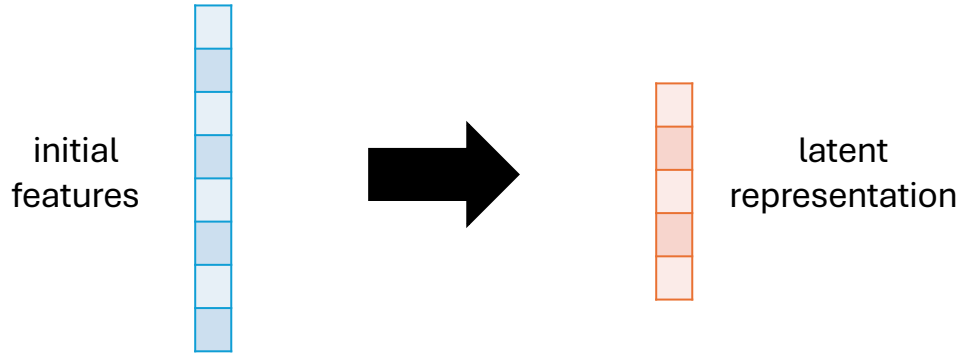
https://www.data-to-viz.com/



A few words on AI and machine learning

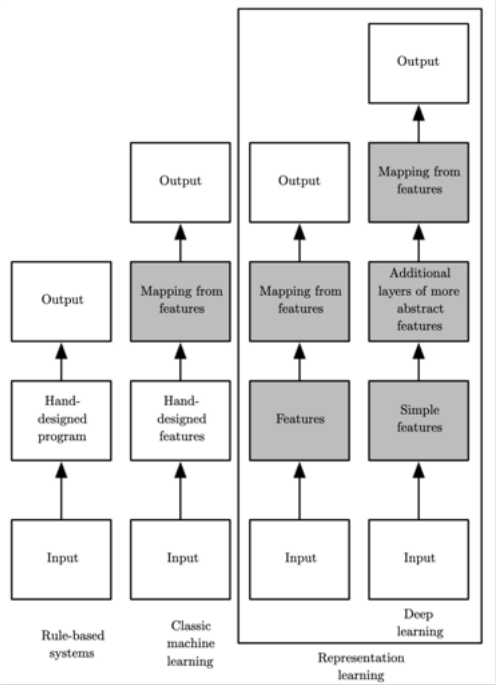


Notion of latent space



- Various way to get this **new representation**:
  - projection based on internal correlations / patterns (eg., clustering)
  - transformation guided by a downstream task (eg., classification)
- Link with AI and representation learning

Representation learning

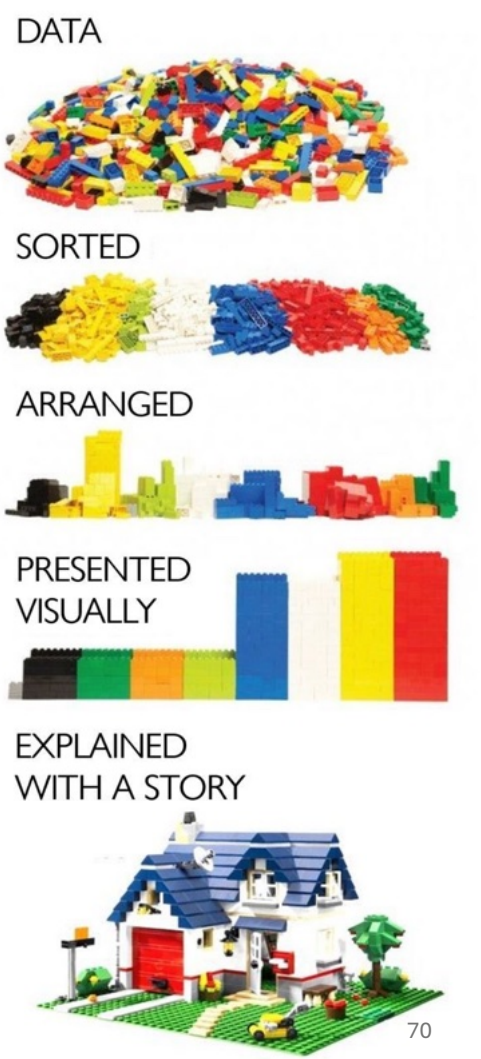


	matricul	departem	ptvente	sexe	age	sitfamil	anciente	csp	codeqlt
9	941007	31	1	Shom	45	Fmar	393	Pouv	A
10	981681	31	1	Shom	38	Fcel	63	Pcad	D
11	1231602	31	1	Shom	60	Fmar	321	Pcad	A
12	1248922	65	2	Shom	34	Fcel	362	Pemp	B
13	1252395	31	2	Shom	56	Fmar	187	Pcad	A
14	1343806	31	1	Shom	49	Fmar	209	Pcad	A
15	1544241	65	1	Sfem	61	Fsep	185	Psan	A
16	1740892	31	4	Sfem	41	Fdiv	268	Psan	A
17	1970765	31	1	Shom	50	Fmar	94	Pcad	A
18	2647067	24	1	Shom	43	Fcel	85	Part	B
19	2849823	31	1	Shom	26	Fcel	75	Psan	B
20	3258308	31	1	Shom	52	Fmar	38	Pcad	B
21	3346139	31	1	Shom	32	Fcel	147	Pcad	A
22	3901483	32	2	Shom	33	Fcel	162	Pcad	D
23	5023873	82	5	Sfem	43	Fmar	18	Pemp	C

Deep Learning, by I. Goodfellow, Y. Bengio and A. Courville, MIT Press, 2016: http://www.deeplearningbook.org

Data storytellling and reporting

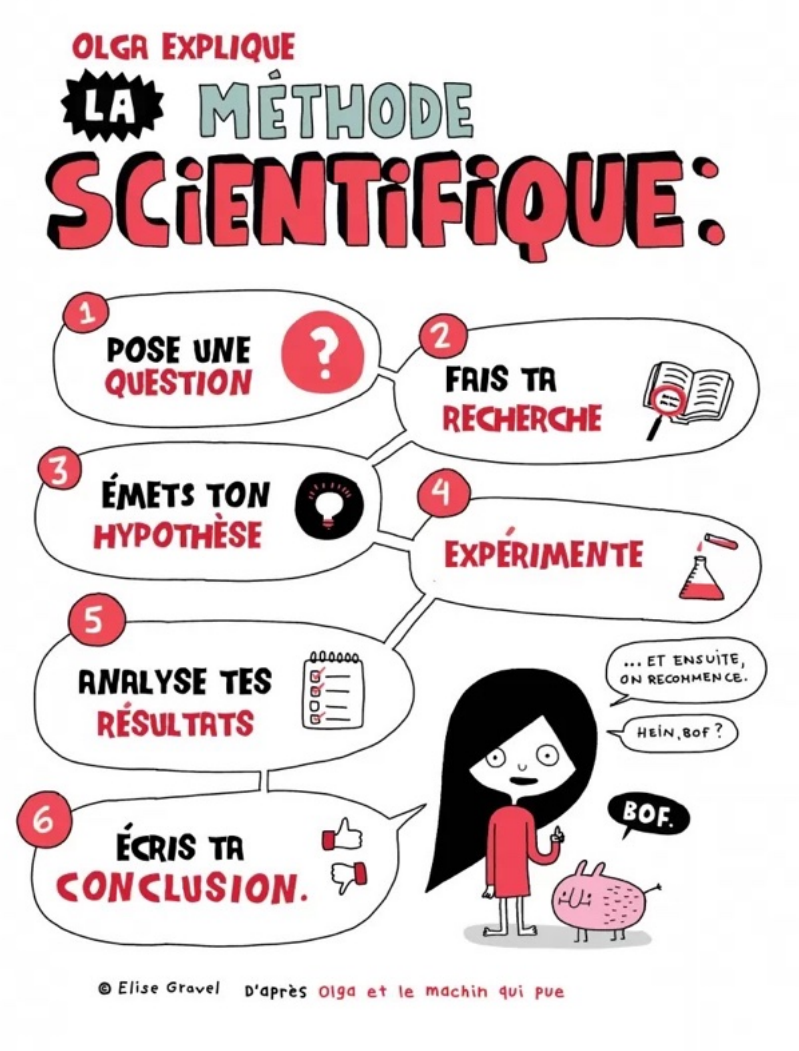
verdict



From Explorer to Guide

- We go from multiple rounds of **exploration** to an **explanation** to others : know your audience (« who »)
- Your investigations should lead to interesting **findings** that are good candidates as assumptions
- Don’t hesitate to strenghten those findings by computing and crossing **quantitative measures** (your evidences), such as correlation score, statistical tests, cluster inertia, etc.
- **Articulate** the different elements you’re showing altogether

Link with the scientific process



Simplify, simplify, simplify

- Your audience is prone to information **overload**
- Keep it **simple**:
  - remove chart junks, useless lines, 3D with no point…
  - choose wisely a **limited number** of charts that support your claim
  - avoid overloaded notebooks (or use modules)
- Prioritize information
  - distinguish what's **really important**
- Use **visual cues**
  - play with font, size, color
  - add figures and illustrations

https://medium.com/@sepidy/what-is-cognitive-overload-in-ux-3-tips-to-minimize-it-in-your-designs-f56caba66fe5

Build a narrative arc

- Presentation as a **story** you tell:
  - context = problem, what we already know
  - conflict = what you have found in the data, the problems, the surprises which install a form of **tension** in your audience
  - resolution = recommandation, insight, what we have learnt
- **Lessons** you've learnt and how to go beyond
  - did you answer to your initial question?
  - if *yes*: What are the evidence? Is it possible to improve the improve?
  - if *not*: What's missing? Is it possible to find another way?

Use visual features

- **Highlight** the main points by using colors, arrows… to **guide** your audience eyes
- **Less is more**: use visual cues sparingly, they must be strategically selected, not juste decoratively (“christmas tree effect”)

Success and failure of dataviz

- Sources:

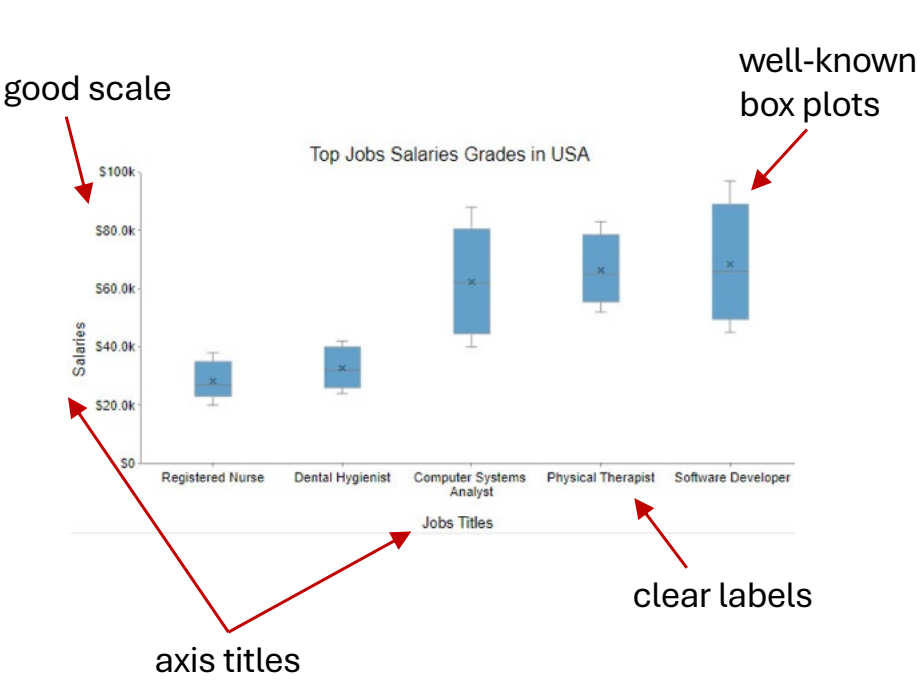
  https://medium.com/@Ana_kin/graphs-gone-wrong-misleading-data-visualizations-d4805d1c4700

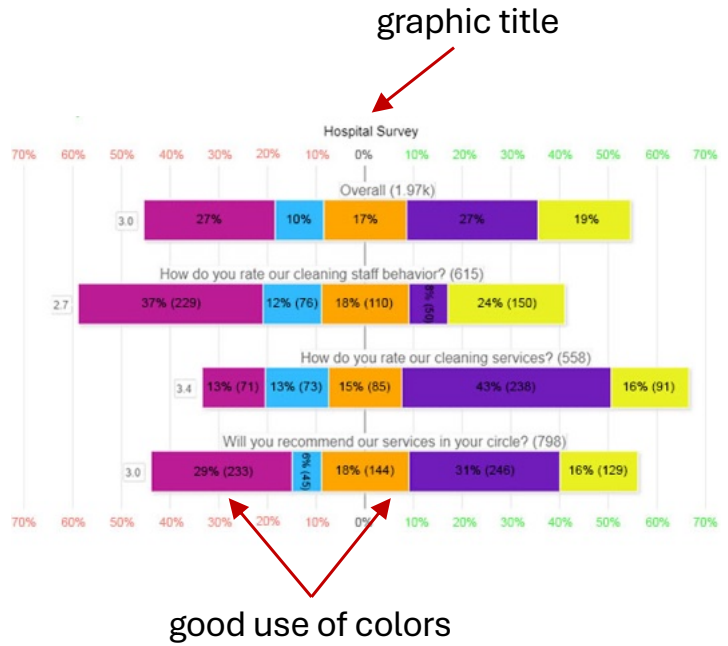
  https://www.reddit.com/r/dataisugly/

- Various reasons of bad visualization:
  - bad scales
  - visual distortions
  - inconsistencies and factual errors
  - aesthetical choices over clarity
  - wrong interpolation

  etc.

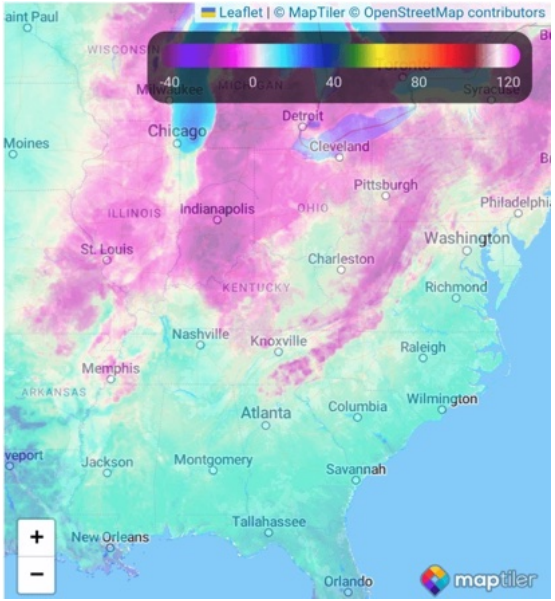
Some good visualizations





Bad scales


The WHITE HOUSE
U.S. STEEL PRODUCTION INCREASES
Total U.S. Steel Production (Mt) 2024 v. 2025
81.8
81.6
81.4
81.2
81.0
80.8
80.6
80.4
80.2
2024 Total Steel Production (Mt)
Total Steel Production 2025 (Mt)


Leaflet | © MapTiler © OpenStreetMap contributors
-40
0
40
80
120
Chicago
Detroit
Cleveland
Pittsburgh
Indianapolis
Washington
Charleston
Richmond
Nashville
Knoxville
Raleigh
Memphis
Atlanta
Columbia
Wilmington
Jackson
Montgomery
Savannah
Tallahassee
New Orleans
Orlando
maptiler

Distortion

- This chart has been created by FoxNews to show the percent share of the supporters of the three candidates Palin, Romney, and Huckabee during the 2012 presidential run.

- This 3D pie chart is difficult to read and doesn't sum to 1!

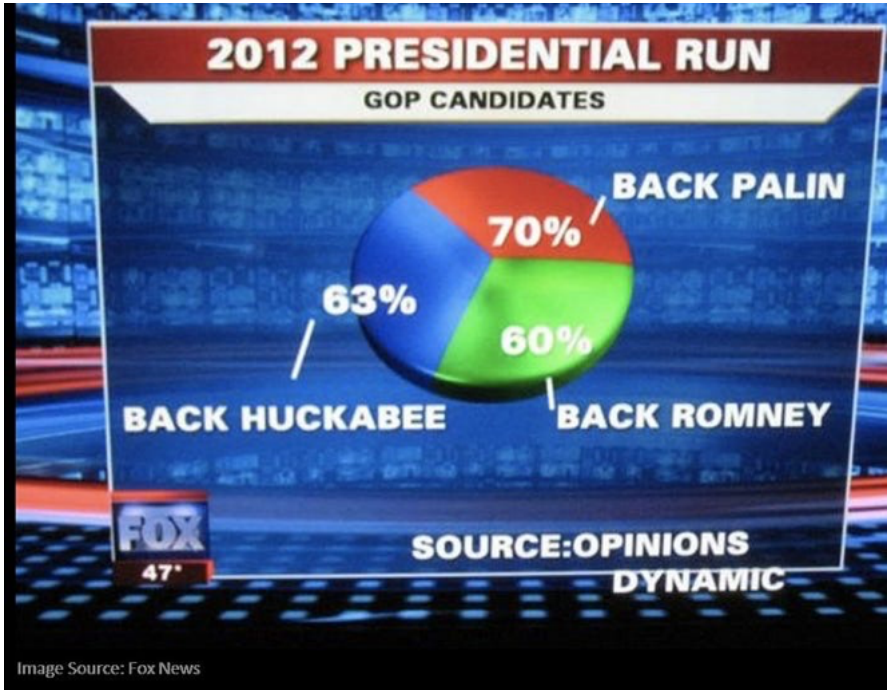
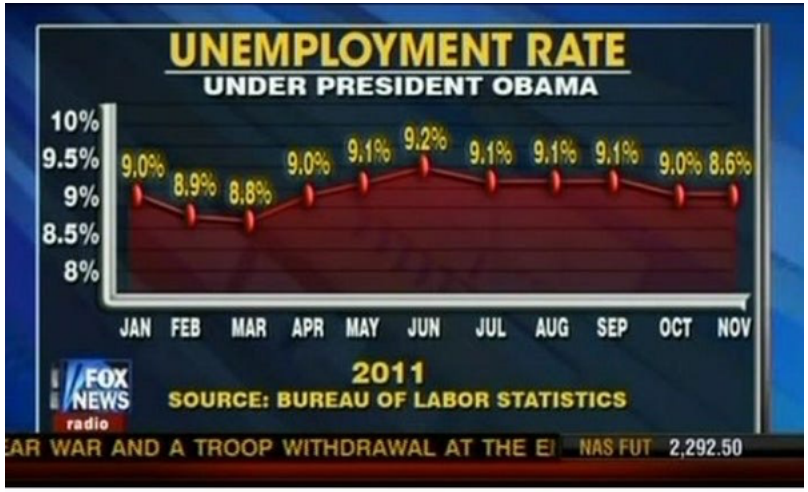


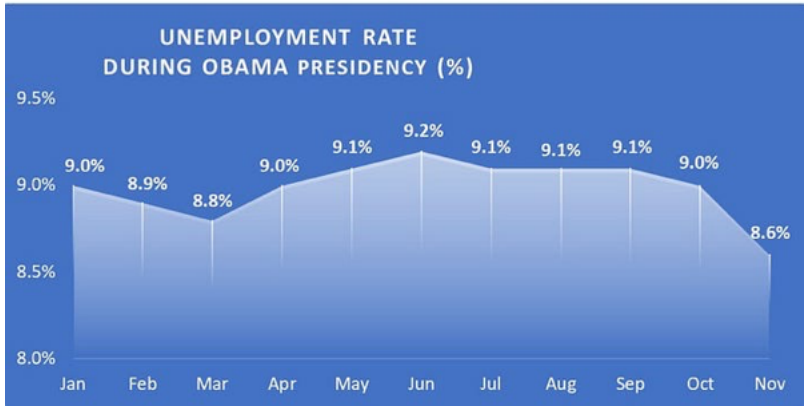
Image Source: Fox News

Inconsistencies

- This chart has been created by Fox News to showcase changes in the unemployment rate during Obama's presidency.

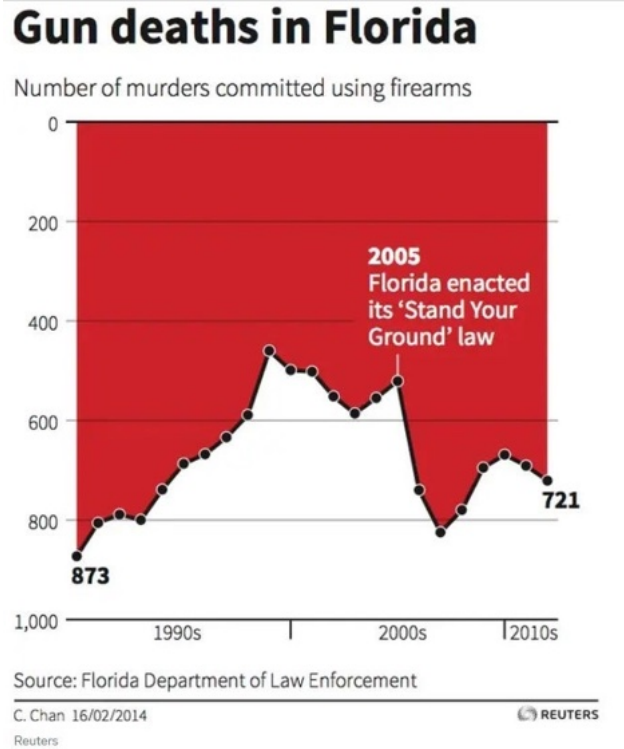
- In addition to a questionable choice for the y axis, we can observe a factual **error** for November rate.



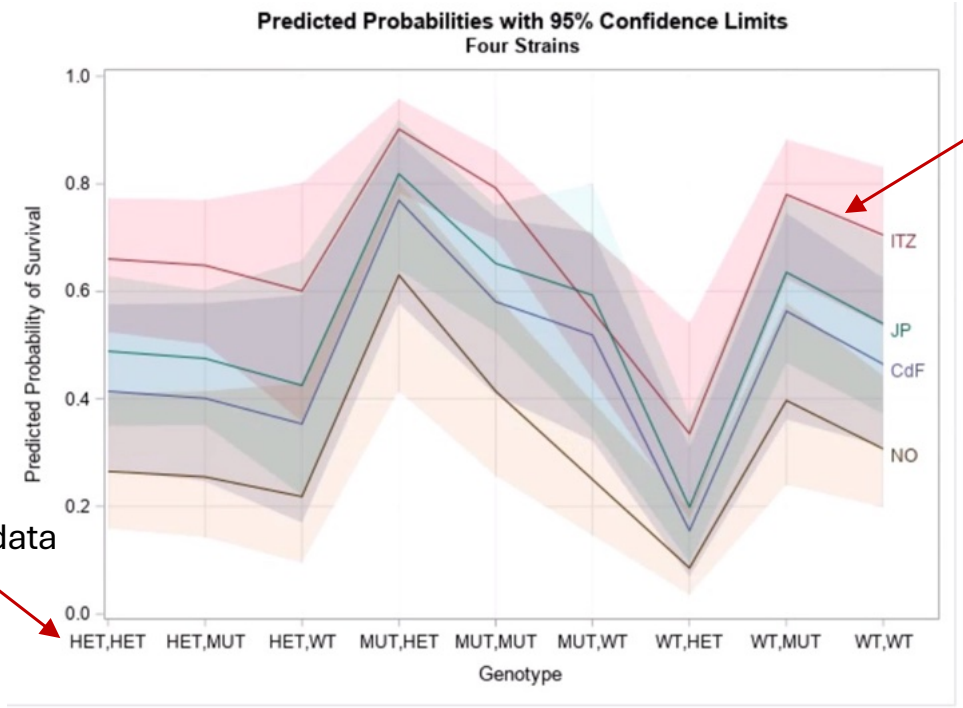


Aesthetical choices over clarity

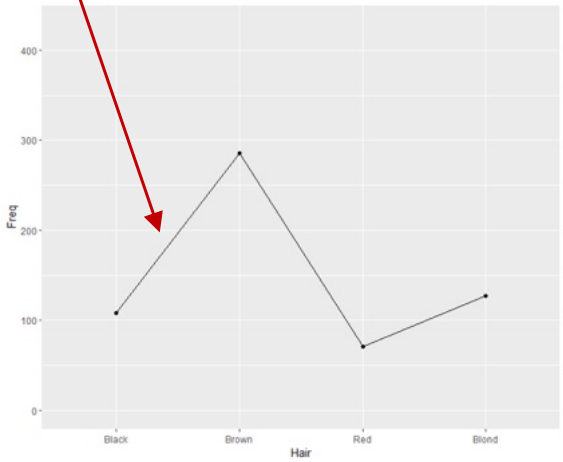
- **Context:** in 2005, Florida enacted a “stand your ground” law = allowing individuals to “hold their ground” and use deadly force to protect or defend against imminent threat of death
- This chart was created by Reuters and published in 2014



Wrong interpolation



what does it even mean?



categorical data

Conclusion

- You always need a **conclusion** that summarize your investigation
- Don’t hesitate to include **your own viewpoint**
- Be cautious **not to over-interprete** (but don’t be too shy)
- Ask other people (or AI) to **proofread** your report
- For an oral presentation: **rehearse, rehearse, rehearse**