Data Visualization

Welcome all

I am Jérôme Baton (jerome.baton@epita.fr)

I have been retrieving and presenting data for years, mostly a Java+backend specialist, Android, ...
Speaker

Now :
Proudfessor + enterTRainer
(several subjects in several places),
Author (« Learning Neo4j 3.x», Packt Publishing, 2017)

Data Visualization DATAVIZ

The major goals of this course are to understand how visual representations can help in the analysis and understanding of complex data, how to design effective visualizations, and how to create an interactive visualization using modern web-based frameworks.
It explores the underlying theory and practical concepts in creating visual representations of large amounts of data. It covers the core topics in data visualization: data representation, visualization toolkits, scientific visualization, medical visualization, information visualization, flow visualization, and volume rendering techniques

Apologies in advance for my accent
And bad ears

I might pronounce your names badly
or even the lectures content

Please bear with my puns and jokes

This is still serious matter

Today, we’ll start with theory, softwares

Other lectures will be about practice, projects and grading

Sessions details are in the Team folder

YOU

Are to take notes

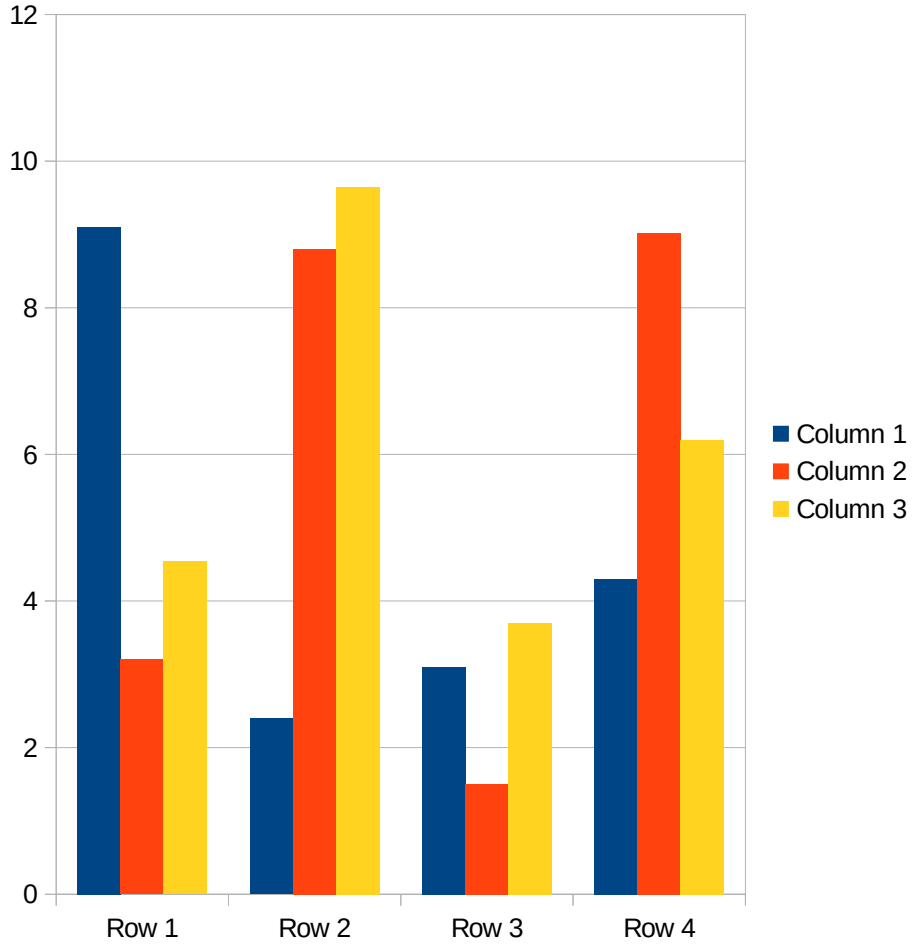
As everything is not in the slides

Are not allowed to take pictures/movies/streams/sound captations

What is dataviz ?

In short data vizualisation

Like this ---->

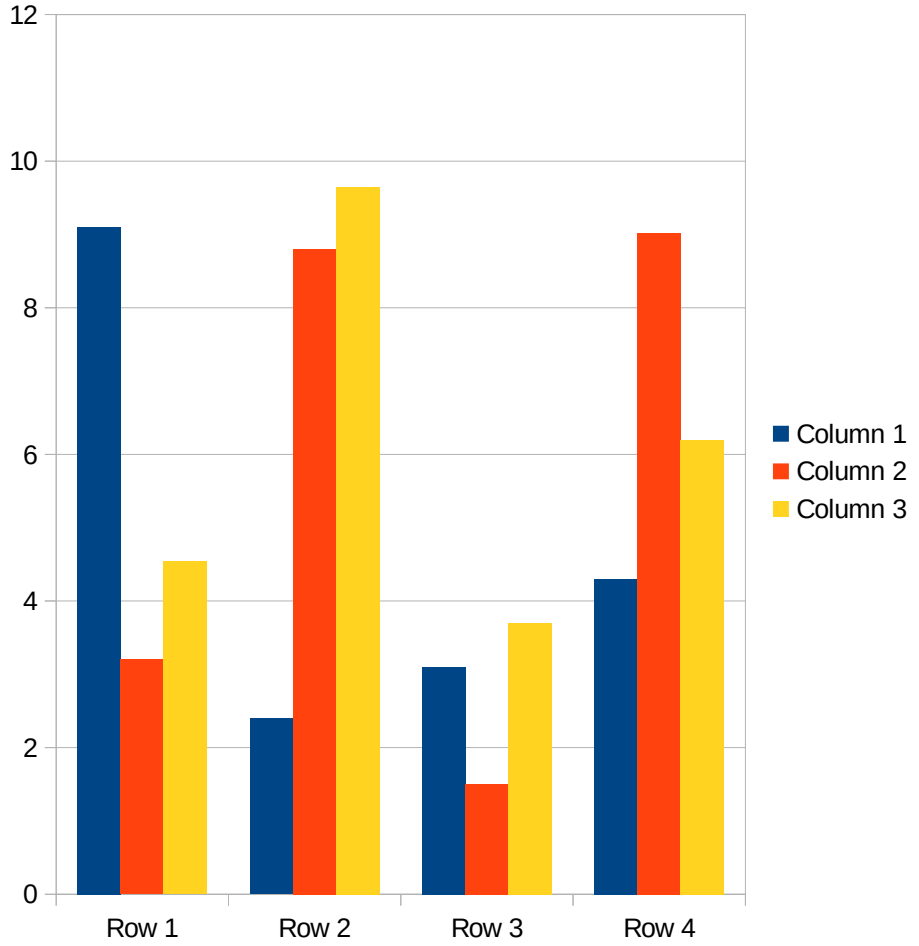


What for ?

Get a better understanding of the collected data

What is dataviz ?

In short data vizualisation



What for ?

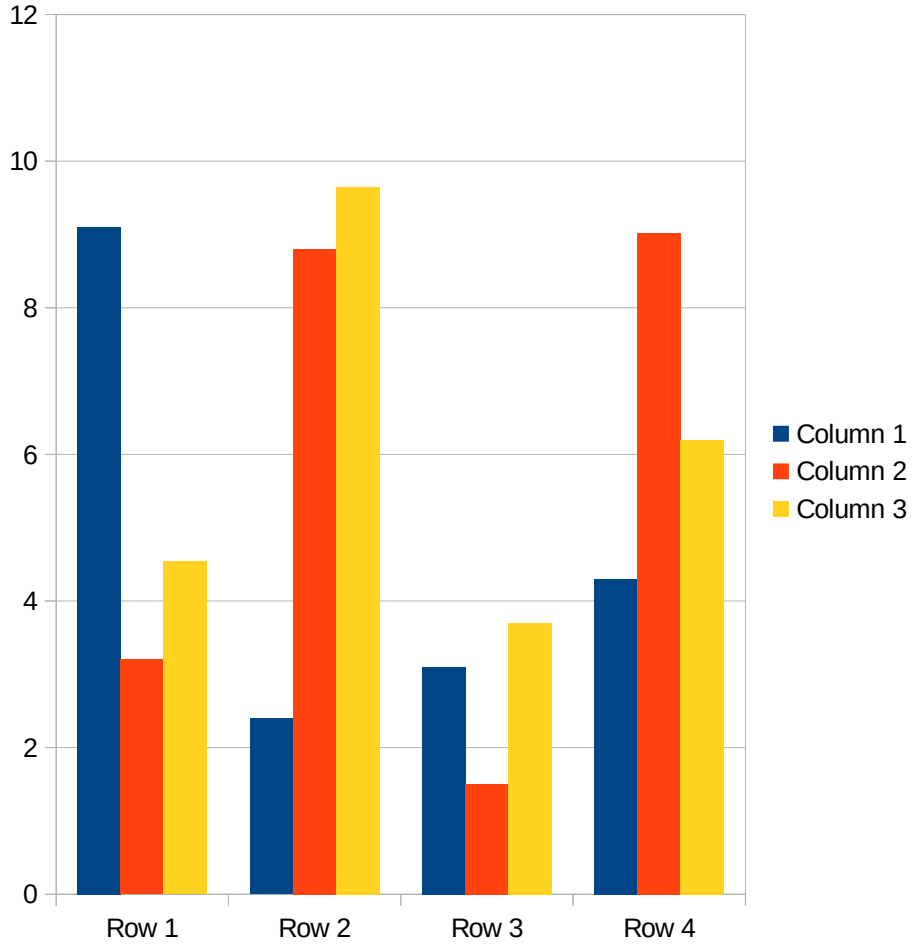
Get a better understanding of the collected data

What is dataviz ?

In short data vizualisation

What for ?

Get a better understanding of the collected data



What is this dataviz ?

In short data vizualisation

Could be
sales data,

could be IOT-generated data

Could be a random pic

What for ?

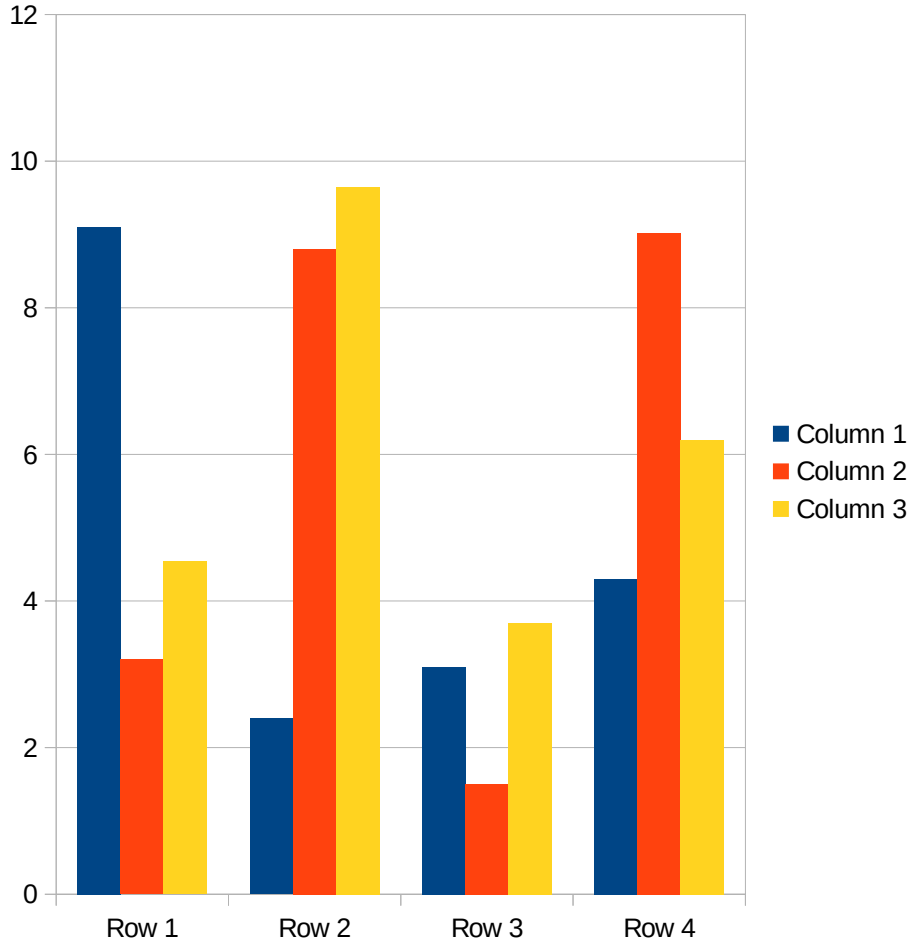
Get a better understanding of the collected data

What is this dataviz ?

This could be a chart of

the sales on three continents in four months

Or many different things



One could represent many different data

With the same chart

Beware not to use the same chart again and again (in your project … wink wink)

Understanding why we need data visualization

Out of all of our logical capabilities, we understand things best through the processing of visual information.

When data is represented visually, the probability of understanding complex builds and numbers increases

Mario Döbler and Tim Großmann

Understanding why we need data visualization

Visualizing data has many advantages, such as the following:

- **Complex data can be easily understood**
- **A simple visual representation of outliers, target audiences, and future markets can be created**
- **Storytelling can be done using dashboards and animations**
- **Data can be explored through interactive visualizations**

Understanding why we need data visualization

Sometimes the data is available but you still need the right tool to interpret it.

A chart is a tool

From « They live » (John Carpenter, 1988)

The power of charts

DEMO

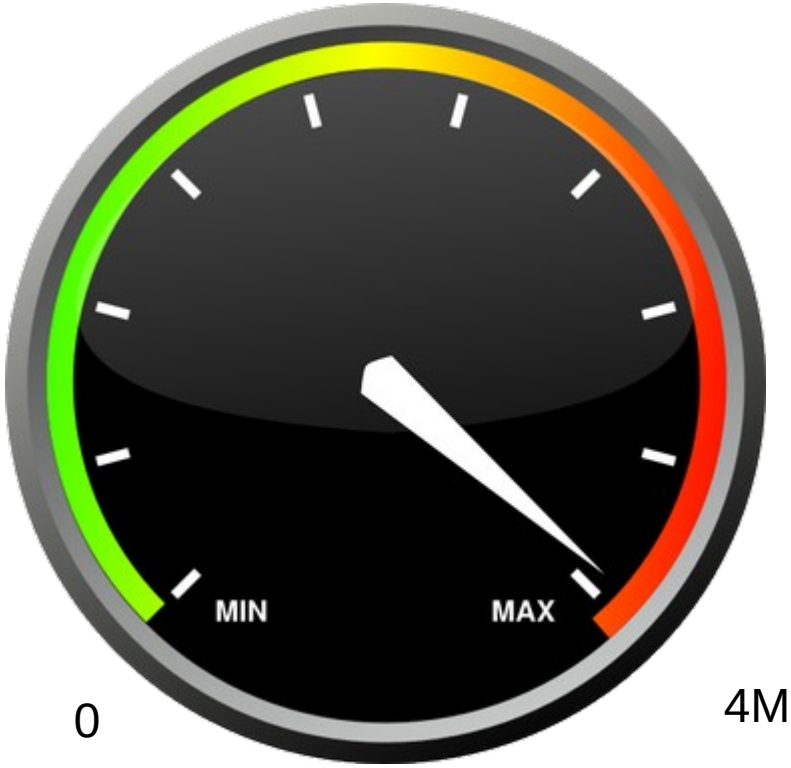
This ...

Value

4 M

Is different of

Even though they display the same value



The first conveys an information

(rough data)

The second conveys an information and eases interpretation

(interpreted data)

There is a difference between ‘nice blue’ and ‘OMG, ITS GONNA EXPLODE’

KEY SENTENCE

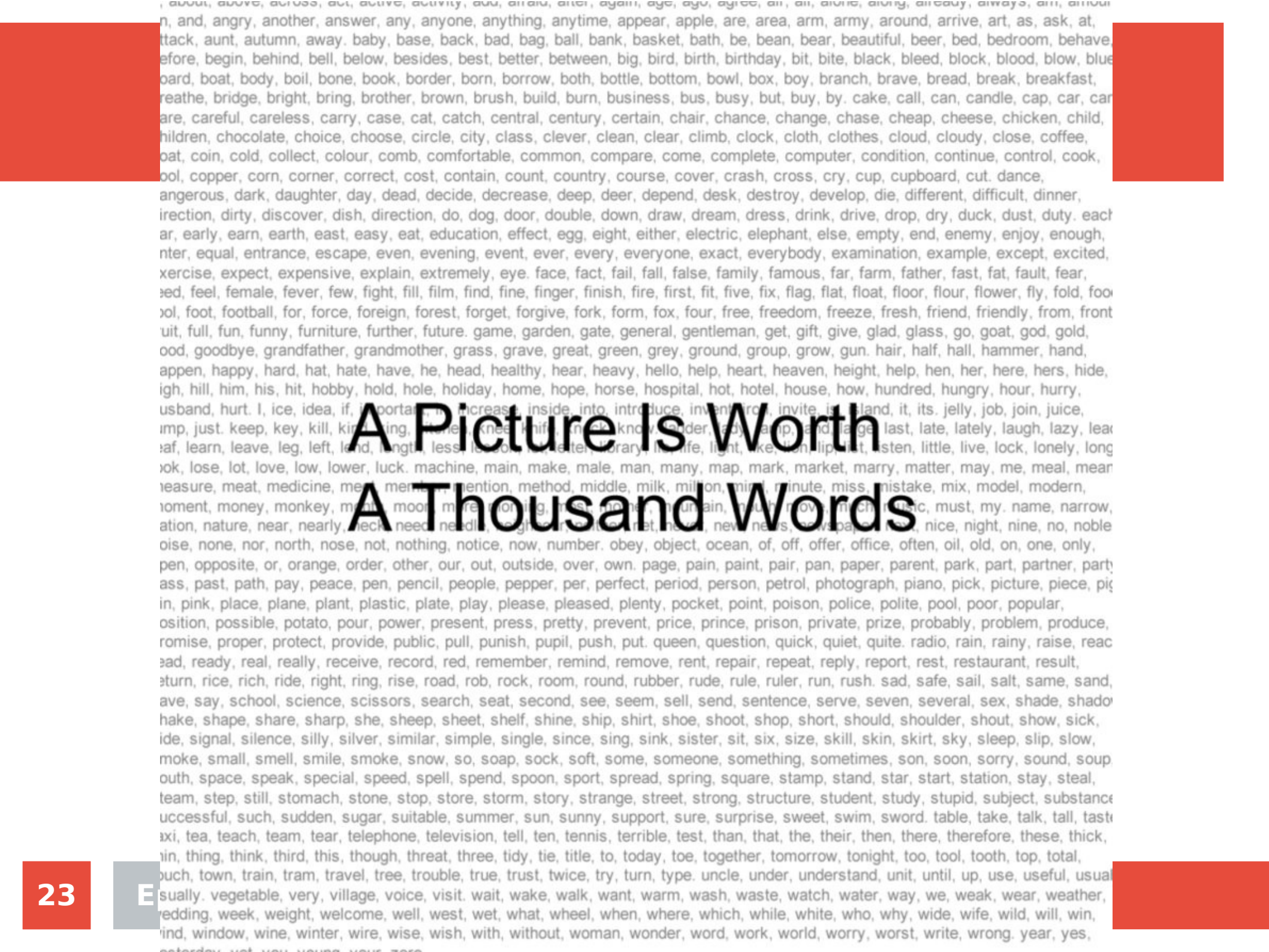
If there is one thing you should remember from this lecture in 20 years, it is this :

KEY SENTENCE

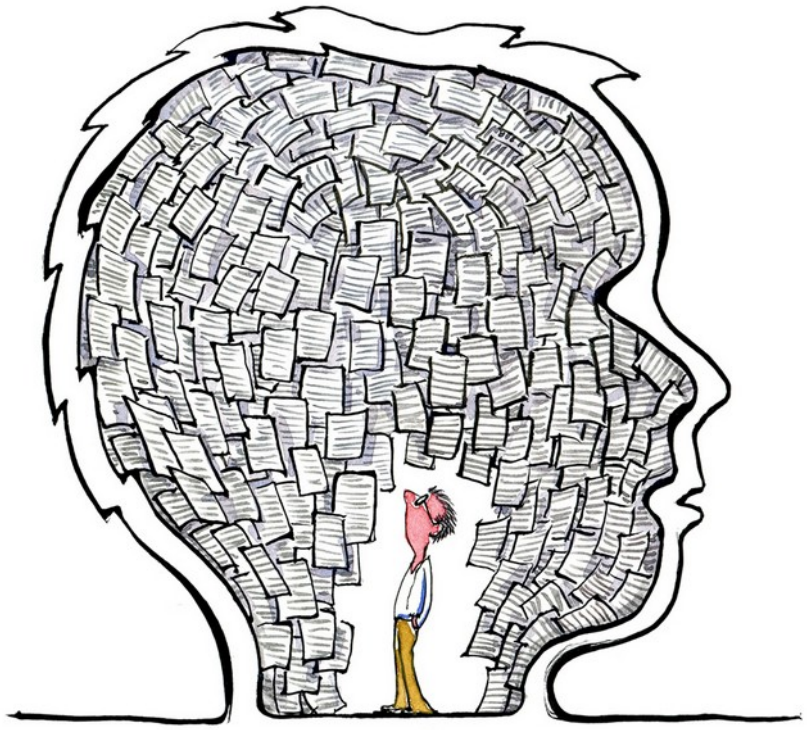
If there is one thing you should remember from this lecture in 20 years, it is this :

A PICTURE IS WORTH A THOUSAND WORDS !

(sorry I screamt)


A Picture Is Worth
A Thousand Words

Is it ?



Can a picture say more than a thousand words?

HikingArtist.com

1000 words =?

Two pages when the font type is Arial, the font size is 12pt, and it has single spacing.

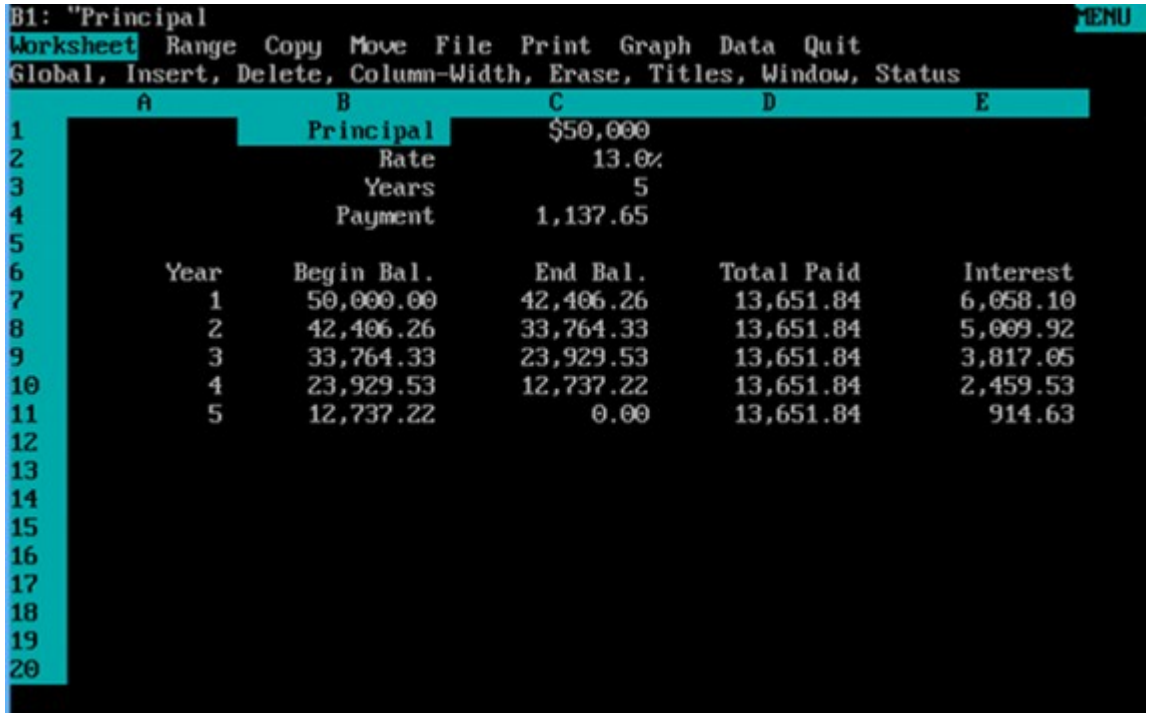
Four pages when the font type is Times New Roman,
the font size is 12pt, and it has double spacing.
These are the common requirements for both high school and college writing.

A good chart is always accompanied by a paragraph explaining the main informations conveyed by the chart.

Today’s DataViz is graphical

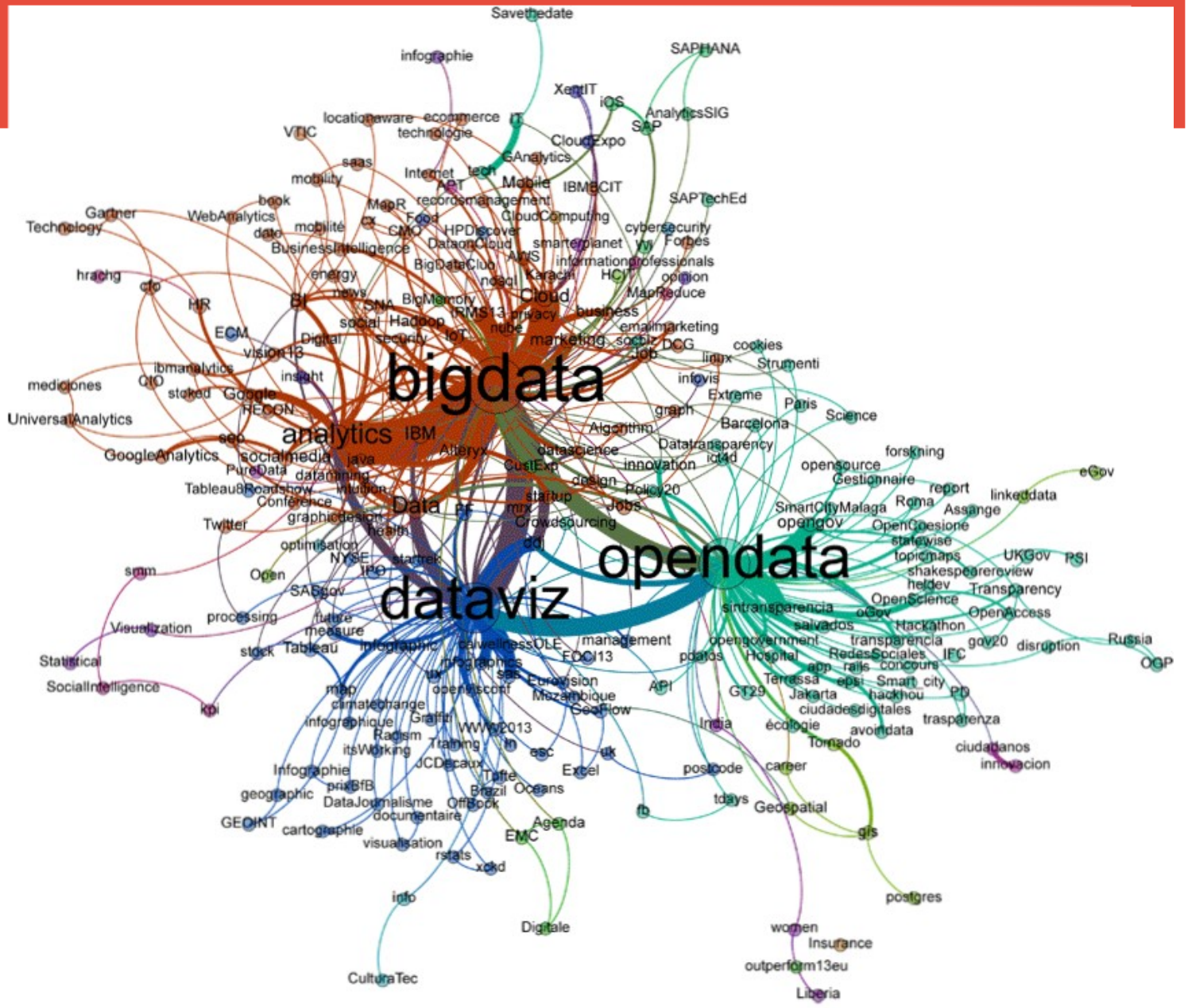
It has evoluted with technical capacities and its audience

From this

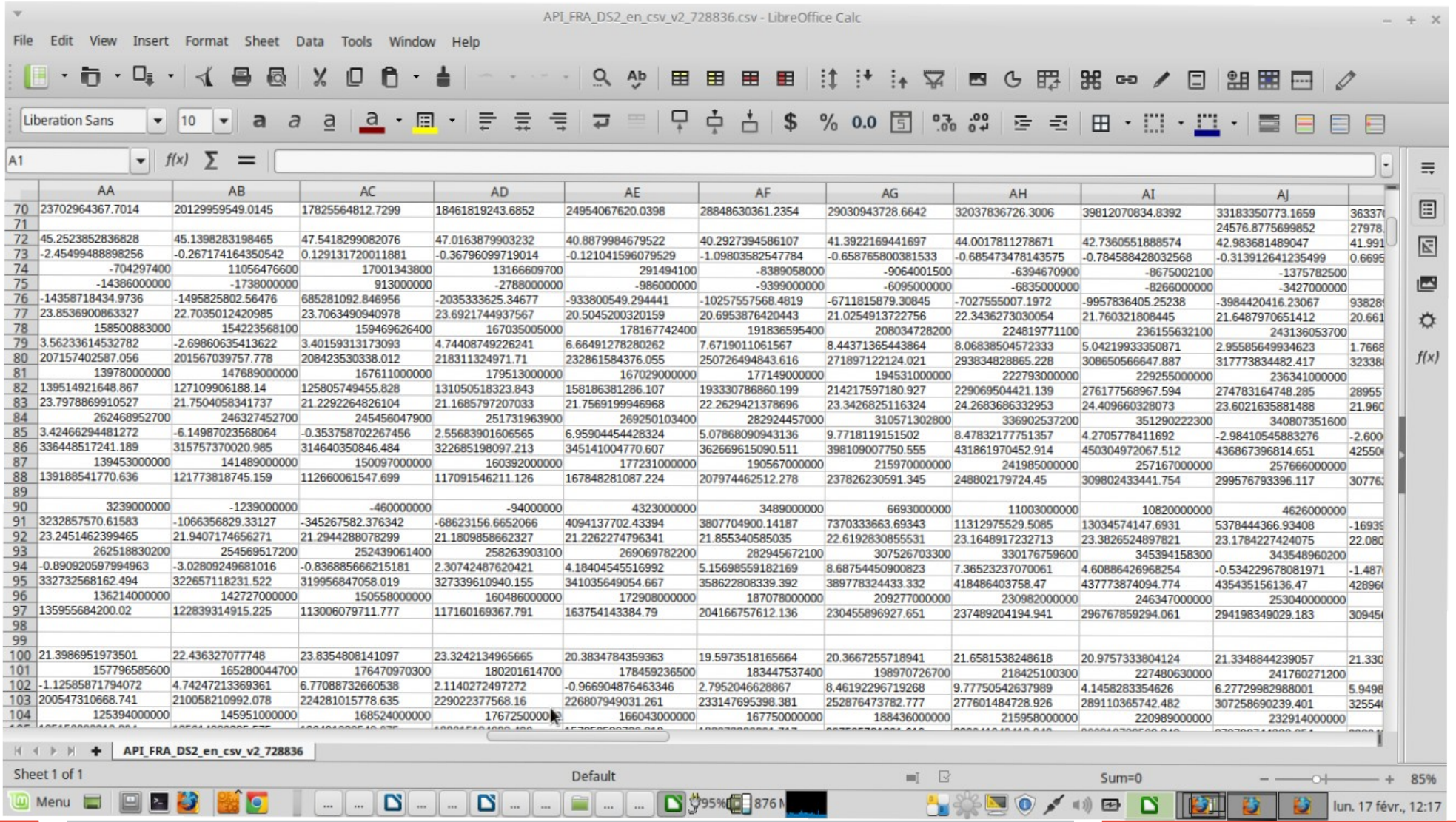


This is high-end software in the 80’s

To this


bigdata
opendata
dataviz
analytics
Data
Cloud
IBM
Savethedate
infographie
SAPHANA
XenIT
iOS
AnalyticsSIG
SAP
CloudExpo
GAnalytics
locationaware
ecommerce
technologie
VTIC
saas
mobility
Internet
tech
Mobile
IBMCIT
SAPTechEd
recordsmanagement
CloudComputing
cybersecurity
Forbes
smarterplanet
informationprofessionals
opinion
MapReduce
Gartner
Technology
WebAnalytics
book
MapR
Food
CMO
HPDiscover
DataonCloud
BusinessIntelligence
BigDataCluo
AWS
Karachi
HCIT
nosql
energy
news
BigMemory
SNA
Hadoop
IRMS13
privacy
business
emailmarketing
socbiz
DCG
cookies
linux
Strumenti
infovis
Extreme
Paris
Science
Barcelona
Algorithm
Datatransparency
graph
forskning
opensource
Gestionnaire
eGov
report
linkeddata
Roma
Assange
SmartCityMalaga
opengov
OpenCoesione
statewise
topicmaps
UKGov
PSI
shakespearereview
heldev
Transparency
OpenScience
OpenAccess
sintransparencia
oGov
Hackathon
salvados
opengovernment
transparencia
gov20
disruption
Russia
OGP
RedesSociales
IFC
Hospital
app
rails
concours
Terrassa
epsi
Smart_city
hackhou
PD
GT29
Jakarta
ciudadesdigitales
trasparenza
avoindata
Tornado
ciudadanos
innovacion
India
écologie
career
postcode
Geospatial
gis
fb
tdays
Agenda
EMC
Digitale
women
Insurance
outperform13eu
Liberia
CulturaTec
info
xckd
rstats
visualisation
cartographie
GEOINT
documentaire
OffBook
DataJournalisme
geographic
prixBfB
Oceans
Brazil
Taffe
Excel
JCDecaux
itsWorking
Training
WWW2013
Graffiti
SeaFlow
Mozambique
operngscont
climatechange
map
API
Eurovision
FOCI13
management
Tableau
stock
Statistical
SocialIntelligence
kpi
Visualization
processing
measure
Infographic
Radism
SAGgov
Open
smm
optimisation
NYSE
IPO
startrek
Twitter
graphicdesign
Conference
Tableau8Roadshow
PureData
datamining
socialmedia
GoogleAnalytics
seo
java
Attleryx
CustExp
datascience
innovation
design
Policy20
startup
Jobs
CrowdSourcing
mix
RF
health
ddj
UniversalAnalytics
mediciones
CIO
stocked
Google
RECON
ibmanalytics
insight
vision13
Digital
security
IoT
marketing
Job
ECM
HR
BI
social
Hadoop
nube
hrachg
cto
Tableau
OLE
Infographics
ux
sas
ssc
uk
asc
in
Training
Radism
WWW2013
Graffiti
SeaFlow
Mozambique
operngscont
climatechange
map
Eurovision
FOCI13
management
Infographie
prixBfB
DataJournalisme
geographic
documentaire
OffBook
Oceans
Brazil
Taffe
Excel
JCDecaux
itsWorking
Agenda
EMC
cartographie
GEOINT
visualisation
rstats
xckd
info
Digitale
CulturaTec
postgres

Still available, but nobody likes this



	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ
70	23702964367.7014	20129959549.0145	17825564812.7299	18461819243.6852	24954067620.0398	28848630361.2354	29030943728.6642	32037836726.3006	39812070834.8392	33183350773.1659
71										24576.8775699852
72	45.2523852836828	45.1398283198465	47.5418299082076	47.0163879903232	40.8879984679522	40.2927394586107	41.3922169441697	44.0017811278671	42.7360551888574	42.983681489047
73	-2.45499488898256	-0.267174164350542	0.129131720011881	-0.36796099719014	-0.121041596079529	-1.09803582547784	-0.658765800381533	-0.685473478143575	-0.784588428032568	-0.313912641235499
74	-704297400	11056476600	17001343800	13166609700	291494100	-8389058000	-9064001500	-6394670900	-8675002100	-1375782500
75	-14386000000	-1738000000	913000000	-2788000000	-986000000	-9399000000	-6095000000	-6835000000	-8266000000	-3427000000
76	-14358718434.9736	-1495825802.56476	685281092.846956	-2035333625.34677	-933800549.294441	-10257557568.4819	-6711815879.30845	-7027555007.1972	-9957836405.25238	-3984420416.23067
77	23.8536900863327	22.7035012420985	23.7063490940978	23.6921744937567	20.5045200320159	20.6953876420443	21.0254913722756	22.3436273030054	21.760321808445	21.6487970651412
78	158500883000	154223568100	159469626400	167035005000	178167742400	191836595400	208034728200	224819771100	236155632100	243136053700
79	3.56233614532782	-2.69860635413622	3.40159313173093	4.74408749226241	6.6649127828026 2	7.6719011061567	8.44371365443864	8.06838504572333	5.04219933350871	2.95585649934623
80	207157402587.056	201567039757.778	208423530338.012	218311324971.71	232861584376.055	250726494843.616	271897122124.021	293834828865.228	308650566647.887	317773834482.417
81	139780000000	147689000000	167611000000	179513000000	167029000000	177149000000	194531000000	222793000000	229255000000	236341000000
82	139514921648.867	127109906188.14	125805749455.828	131050518323.843	158186381286.107	193330786860.199	214217597180.927	229069504421.139	276177568967.594	274783164748.285
83	23.7978869910527	21.7504058341737	21.2292264826104	21.1685797207033	21.7569199946968	22.2629421378696	23.3426825116324	24.2683686332953	24.409660328073	23.6021635881488
84	262468952700	246327452700	245456047900	251731963900	269250103400	282924457000	310571302800	336902537200	351290222300	340807351600
85	3.42466294481272	-6.14987023568064	-0.353758702267456	2.55683901606565	6.95904454428324	5.07868090943136	9.7718119151502	8.47832177751357	4.2705778411692	-2.98410545883276
86	336448517241.189	315757370020.985	314640350846.484	322685198097.213	345141004770.607	362669615090.511	398109007750.555	431861970452.914	450304972067.512	436867396814.651
87	139453000000	141489000000	150097000000	160392000000	177231000000	190567000000	215970000000	241985000000	257167000000	257666000000
88	139188541770.636	121773818745.159	112660061547.699	117091546211.126	167848281087.224	207974462512.278	237826230591.345	248802179724.45	309802433441.754	299576793396.117
89										
90	3239000000	-1239000000	-460000000	-94000000	4323000000	3489000000	6693000000	11003000000	10820000000	4626000000
91	3232857570.61583	-1066356829.33127	-345267582.376342	-68623156.6652066	4094137702.43394	3807704900.14187	7370333663.69343	11312975529.5085	13034574147.6931	5378444366.93408
92	23.2451462399465	21.9407174656271	21.2944288078299	21.1809858662327	21.2262274796341	21.855340585035	22.6192830855531	23.1648917232713	23.3826524897821	23.1784227424075
93	262518830200	254569517200	252439061400	258263903100	269069782200	282945672100	307526703300	330176759600	345394158300	343548960200
94	-0.890920597994963	-3.02809249681016	-0.836885666215181	2.30742487620421	4.18404545516992	5.15698559182169	8.68754450900823	7.36523237070061	4.60886426968254	-0.534229678081971
95	332732568162.494	322657118231.522	319956847058.019	327339610940.155	341035649054.667	358622808339.392	389778324433.332	418486403758.47	437773874094.774	435435156136.47
96	136214000000	142727000000	150558000000	160486000000	172908000000	187078000000	209277000000	230982000000	246347000000	253040000000
97	135955684200.02	122839314915.225	113006079711.777	117160169367.791	163754143384.79	204166757612.136	230455896927.651	237489204194.941	296767859294.061	294198349029.183
98										
99										
100	21.3986951973501	22.436327077748	23.8354808141097	23.3242134965665	20.3834784359363	19.5973518165664	20.3667255718941	21.6581538248618	20.9757333804124	21.3348844239057
101	157796585600	165280044700	176470970300	180201614700	178459236500	183447537400	198970726700	218425100300	227480630000	241760271200
102	-1.12585871794072	4.74247213369361	6.77088732660538	2.1140272497272	-0.966904876463346	2.7952046628867	8.46192296719268	9.77750542637989	4.1458283354626	6.27729982988001
103	200547310668.741	210058210992.078	224281015778.635	229022377568.16	226807949031.261	233147695398.381	252876473782.777	277601484728.926	289110365742.482	307258690239.401
104	125394000000	145951000000	168524000000	176725000000	166043000000	167750000000	188436000000	215958000000	220989000000	232914000000

(back in the days) Dataviz was for company boards

It used to be reports

Litterally piles of paper

Then reports with charts

Still paper piles

Or just charts in a presentation

(also invention of ‘death by powerpoint’)

Dashboards

On demand

Now : all this plus animated charts, 3D

Tomorrow charts in VR ?

We are viewers, not readers anymore

Lets not compare time spent reading books now and before

Mainstream TV happened.

We want information quick

The Hurry Generations

Because of living with remote controls.

Attention span : size matters

From

From an ocean of data (Internet)

Or data lakes :) (company zone)

(back in the days) Dataviz was for company boards

It used to be reports

Litterally piles of paper

Then reports with charts

Still paper piles

Or just charts in a presentation

(also invention of ‘death by powerpoint’)

Dashboards

On demand

Now : all this plus animated charts, 3D

Tomorrow charts in VR ?

From https://en.wikipedia.org/wiki/Data_lake

A data lake is a system or repository of data stored in its natural/raw format, [1] usually object blobs or files. A data lake is usually a single store of all enterprise data including raw copies of source system data and transformed data used for tasks such as reporting, visualization, advanced analytics and machine learning. A data lake can include structured data from relational databases (rows and columns), semi-structured data (CSV, logs, XML, JSON), unstructured data (emails, documents, PDFs) and binary data (images, audio, video). [2]

A data swamp is a deteriorated and unmanaged data lake that is either inaccessible to its intended users or is providing little value.[3]

To which I'll append 'data dump'

Dataviz conveys a message

If you produce a chart

- You want it to illustrate a point

However …. its difficult to get understood

- You need to make it crystal clear
- Many traps on the way

DataViz must be adapted to the audience

- The how
  - Is audience reading LTR or RTL ?
  - Some colors convey a different meaning in different cultures !
- The what
  - Dont show board’s salaries to a company salarymen

Dataviz must be adapted to the era

- Dont represent booming business numbers with explosions.
- Avoid anything that can be stamped as not politically correct

However ...

Some rules stay valid

- Show ordered data
- Use a legend
- Create a title
- Keep user’s attention

And to maintain attention, you must choose the colors wisely

The RED color

Red – Fire: Red is an auspicious as well as a popular color in Chinese culture.
It symbolizes luck, happiness and joy. It also represents summer season and the Southern Direction. It is the color worn by brides, since it is believed to be an auspicious color for warding off evil.
Red also represents vitality, celebration and fertility in traditional Chinese color symbolism.

Today, westerners take red to be a representation of China's Revolution or political system.

However, this is definitely not the case since ancient Chinese roots certainly did not factor in the revolution for it. If you are doing business in China, make use of bit of red in marketing, but not too much.

I like red too – Jérôme, France (hence the slides)

Red is the enemy : CCCP / USSR – says every world map representation in the 80s

Colors meanings (west world only ?)

Colors

So, choose wisely !

A beautiful palette

- is not friendly to the color blind
- 4 % of World population
  - /!\ there is several kinds of color blindness /!\
- See colourblindness simulators

BIG ISSUE

Not easy to satisfy everybody when you have a global audience (web site).

Because people tend to see things

But not mandatorily what you want to show.

- Things they love
- Things they hate

Not mandatorily what you show, but what they interpret

Communication.
Example : this course

(any lecture)

What I know
What I thought of
What I wrote
What you read
What you take away
How you will explain it to someone else
What that person will remember
...

All these can be different from one another
Hopefully not

Communication is difficult

Charts are a way to communicate,

to give an information disguised

as a scientific proof

https://en.wikipedia.org/wiki/Milgram_experiment

ANTIPATTERN : ONE DATAVIZ TO RULE THEM ALL !

Predicate :

ONE DATAVIZ CAN EXPRESS ANY DATA !

It is Camembert viz (Eng : pie chart)

Or so seem to think many, many journalists.

FYI : Camembert is from Normandy

Real camembert is made from unpasteurized milk

ONE DATAVIZ TO RULE THEM ALL !

Predicate :

ONE DATAVIZ CAN EXPRESS ANY DATA !
It is Camembert viz (Eng : pie chart) !!

Experience replies :

No, not at all

Same data

Can be represented in many ways

So many chart types to choose from

Same data

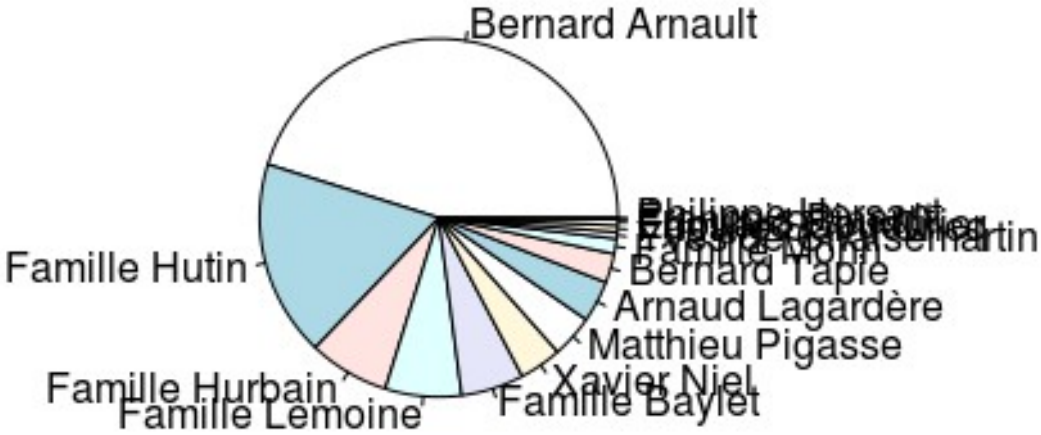
Can be represented in many ways

So many chart types to choose from

The role of charts is to make data show a meaning within a glance

Pie – Most known & badly used

We will see later what this data is



Your comments ?

Source : https://github.com/wadael/presse_fr

Did I hear ‘unreadable’ ?

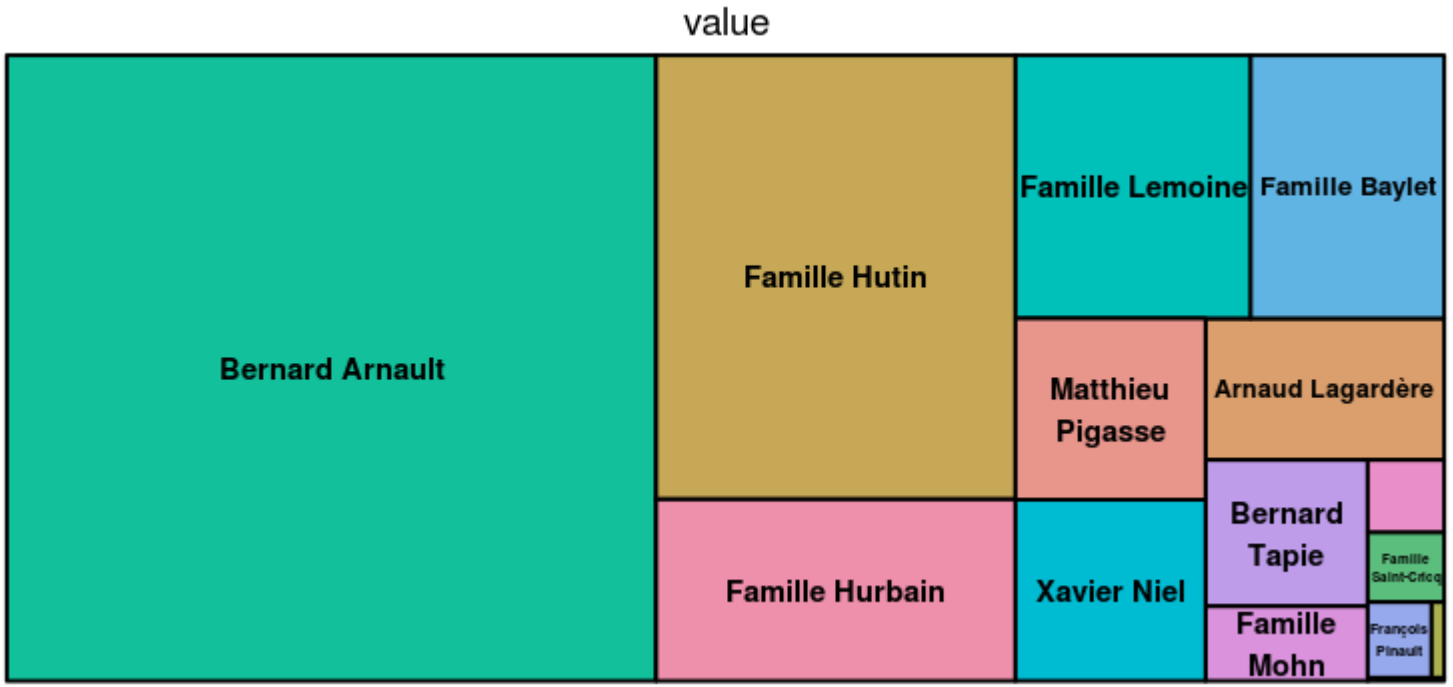
Indeed, unreadable

5 points for Gryffindor

‘Too much data’ issue

This chart is better

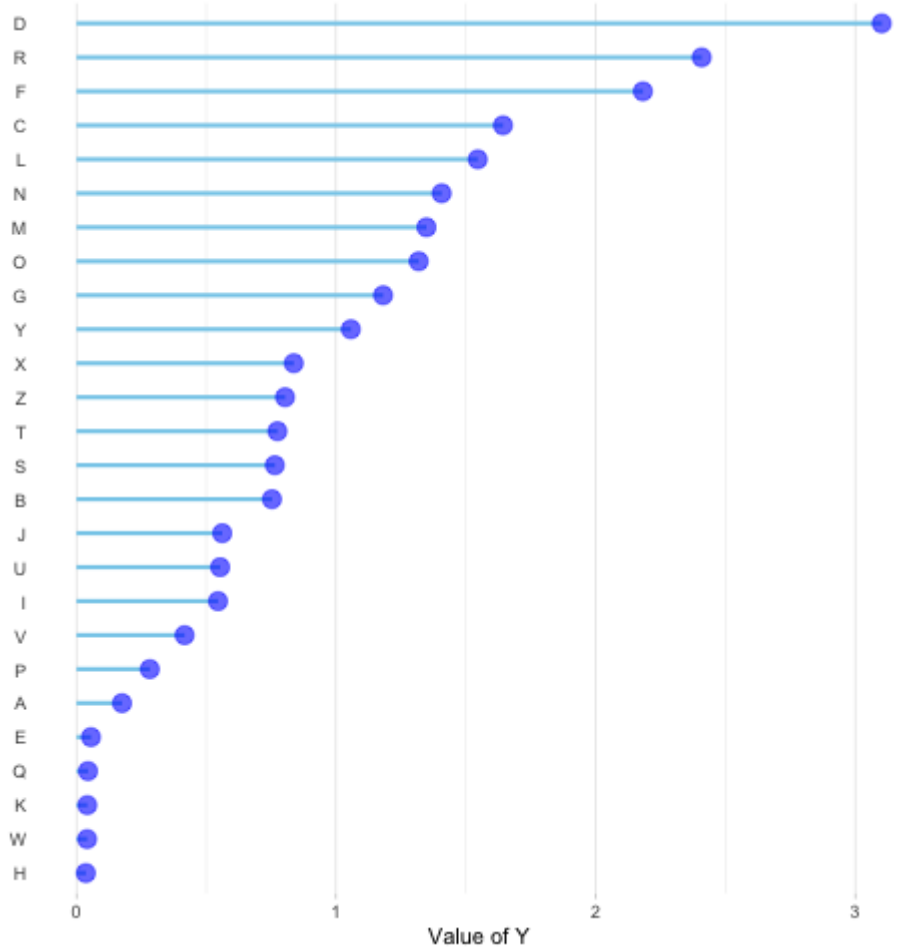
(treemap)



Lollipop

Gives a better insight

Than pie



! Do not forget to sort !

Plenty of fish in the ocean

To find the good dataviz, you need to select a few factors

*** nature of information (numeric, map ...)**

*** how many parameters is there ?**

1 : business per year

2 : business per year per business line

3 : business per year per business line per continent

*** how many dimensions ?**

/!\ Color scheme

Is all data shown ?

Charts are used to fool people

They are displayed as proofs of anything

And just like numbers can say anything and its contrary, charts are the same

How to fool people

Scales

- Sometimes scale is given and not respected
- Linear and logarithmic produce different results
- No origin in the chart

A scale must start at zero and be linear !

- Because this is implied for most readers

Units

- Different units on the same axis

Examples of this

On every election day in France, I see lying charts

Like 40 % being more than the half of a half-disc representing the new chamber of Parliament repartition.

With the % written on it …..

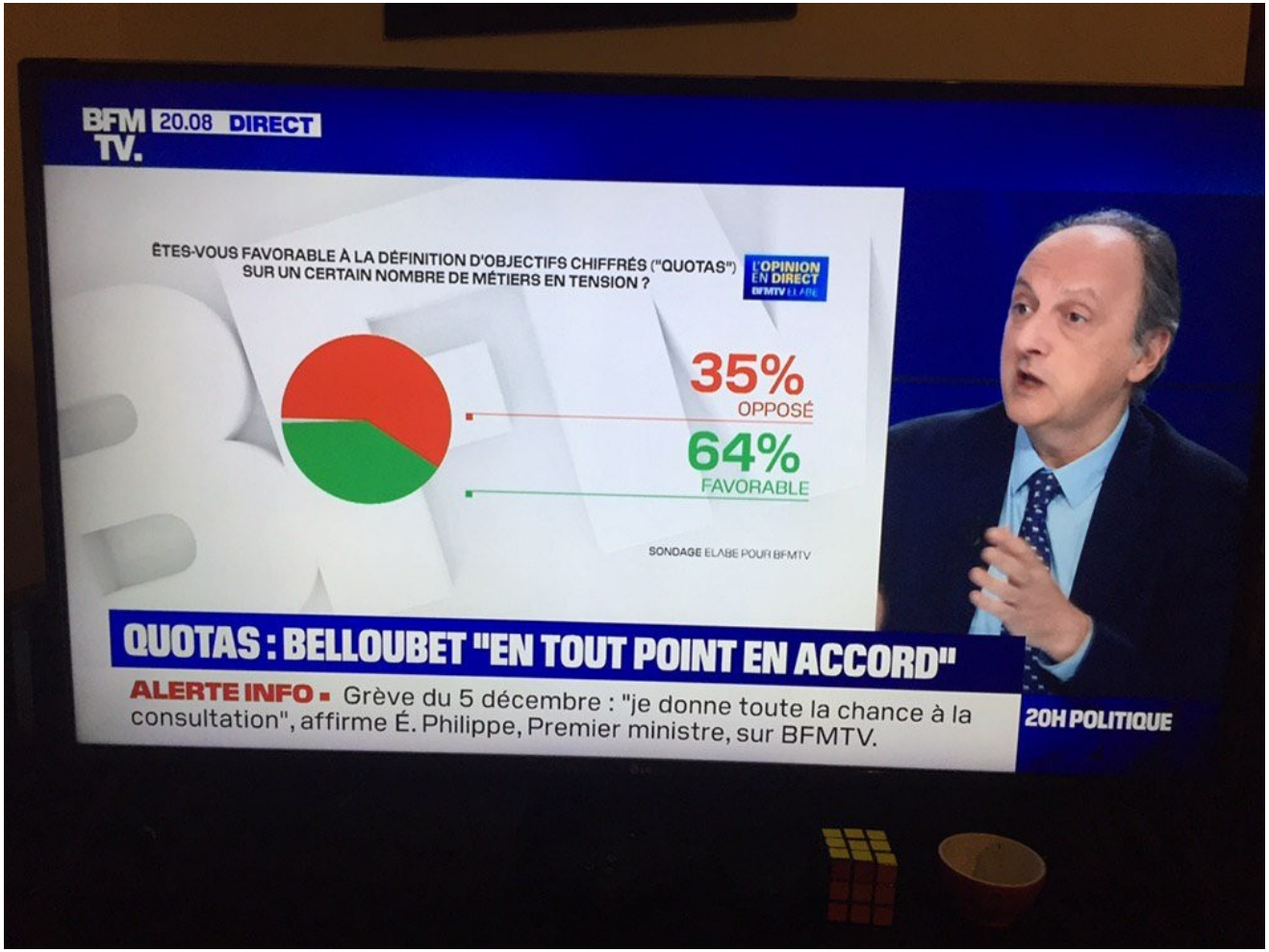
BFM TV
17:01
ÊTES-VOUS FAVORABLE À LA DÉFINITION D'OBJECTIFS CHIFFRÉS ("QUOTAS") SUR UN CERTAIN NOMBRE DE MÉTIERS EN TENSION ?
L'OPINION EN DIRECT
35%
OPPOSÉ
64%
FAVORABLE
SONDAGE ELABE POUR BFMTV
ALERTE INFO • Pour 59% des Français, l'immigration et l'asile sont "des sujets majeurs" qui "doivent être évoqués" (Elabe/BFMTV)
BFM STORY

Local FoxNews at it again !

Clearly 48 % is more than a half :)

Also, color choice is never innocent

(green is support government)


BFM TV.
20.08 DIRECT
ÊTES-VOUS FAVORABLE À LA DÉFINITION D'OBJECTIFS CHIFFRÉS ("QUOTAS") SUR UN CERTAIN NOMBRE DE MÉTIERS EN TENSION ?
L'OPINION EN DIRECT
35%
OPPOSÉ
64%
FAVORABLE
SONDAGE ELABE POUR BFMTV
QUOTAS : BELLOUBET "EN TOUT POINT EN ACCORD"
ALERTE INFO • Grève du 5 décembre : "je donne toute la chance à la consultation", affirme É. Philippe, Premier ministre, sur BFMTV.
20H POLITIQUE

Same old song

The year was 2005

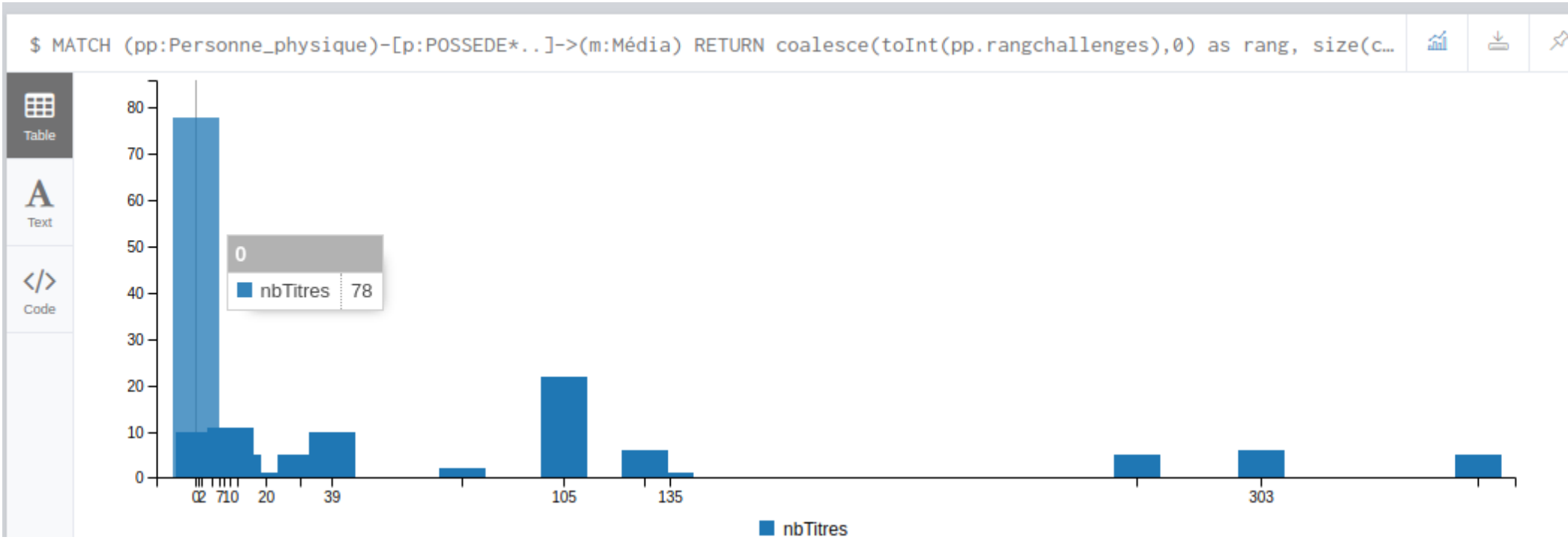
Final result :

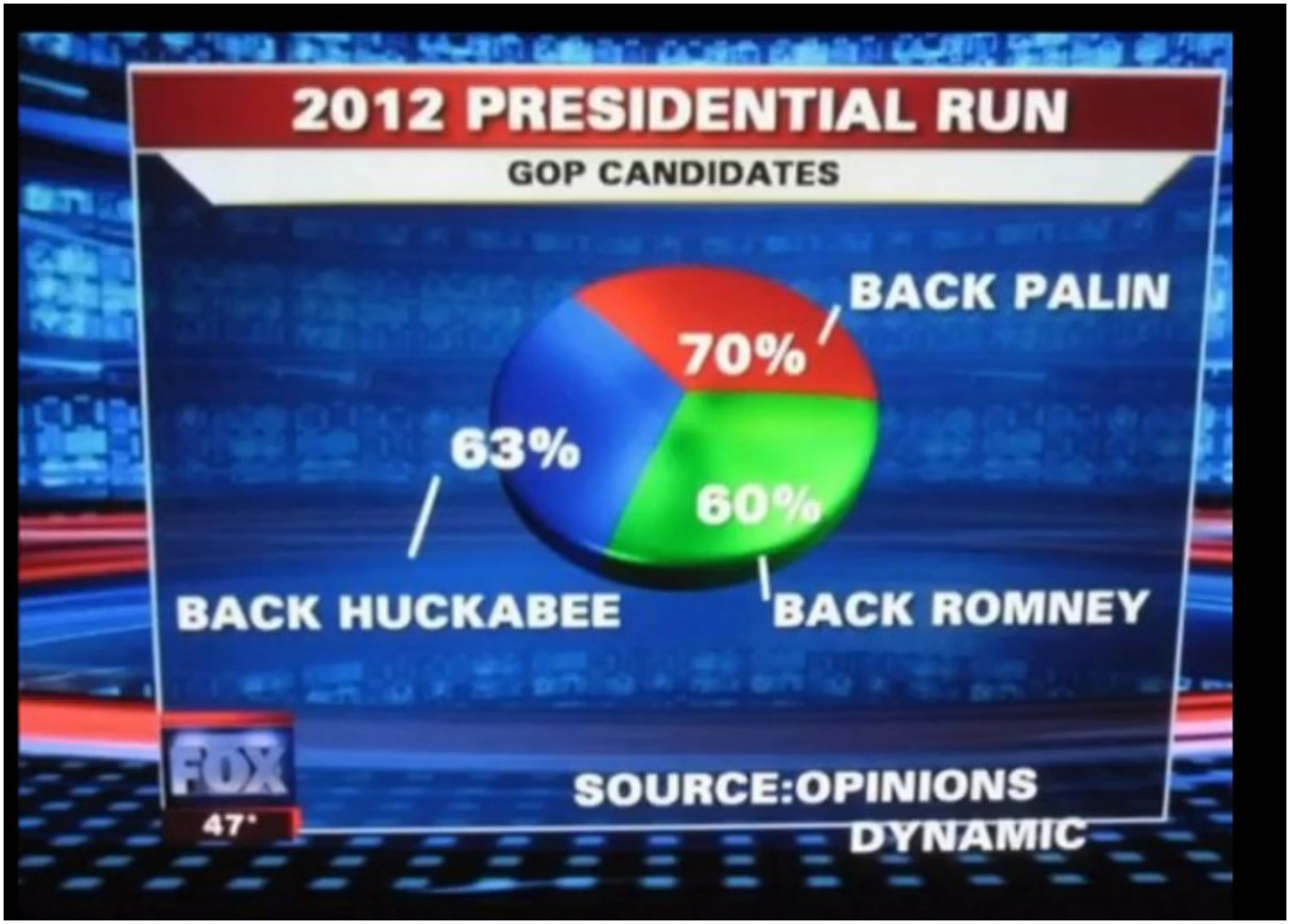
No 54,68 %

The chart is mainly red, this is an issue

One for which I can only blame myself

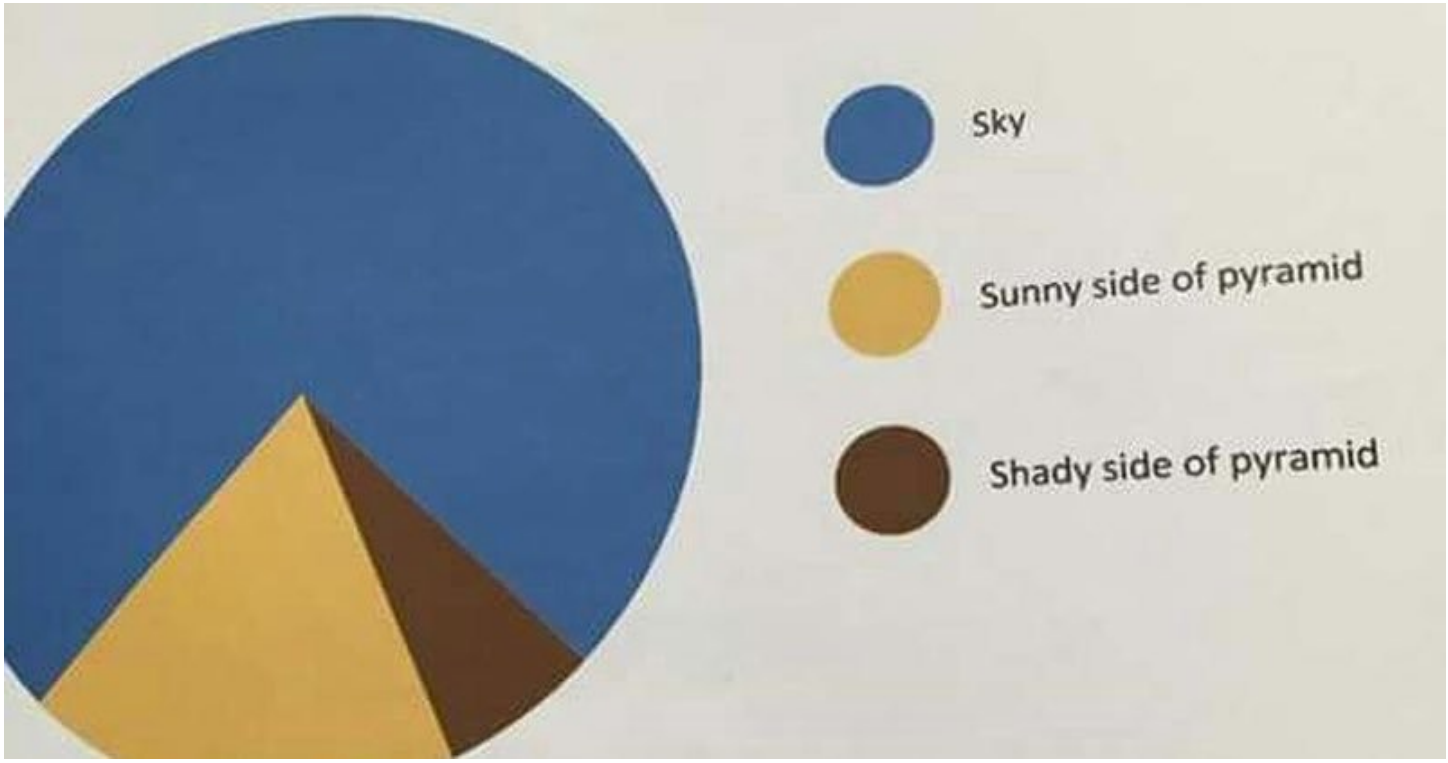
This is an ugly, useless bar chart




2012 PRESIDENTIAL RUN
GOP CANDIDATES
BACK PALIN
70%
63%
60%
BACK HUCKABEE
BACK ROMNEY
FOX
47'
SOURCE:OPINIONS
DYNAMIC

U.S humor (American pie)

More Pie humor

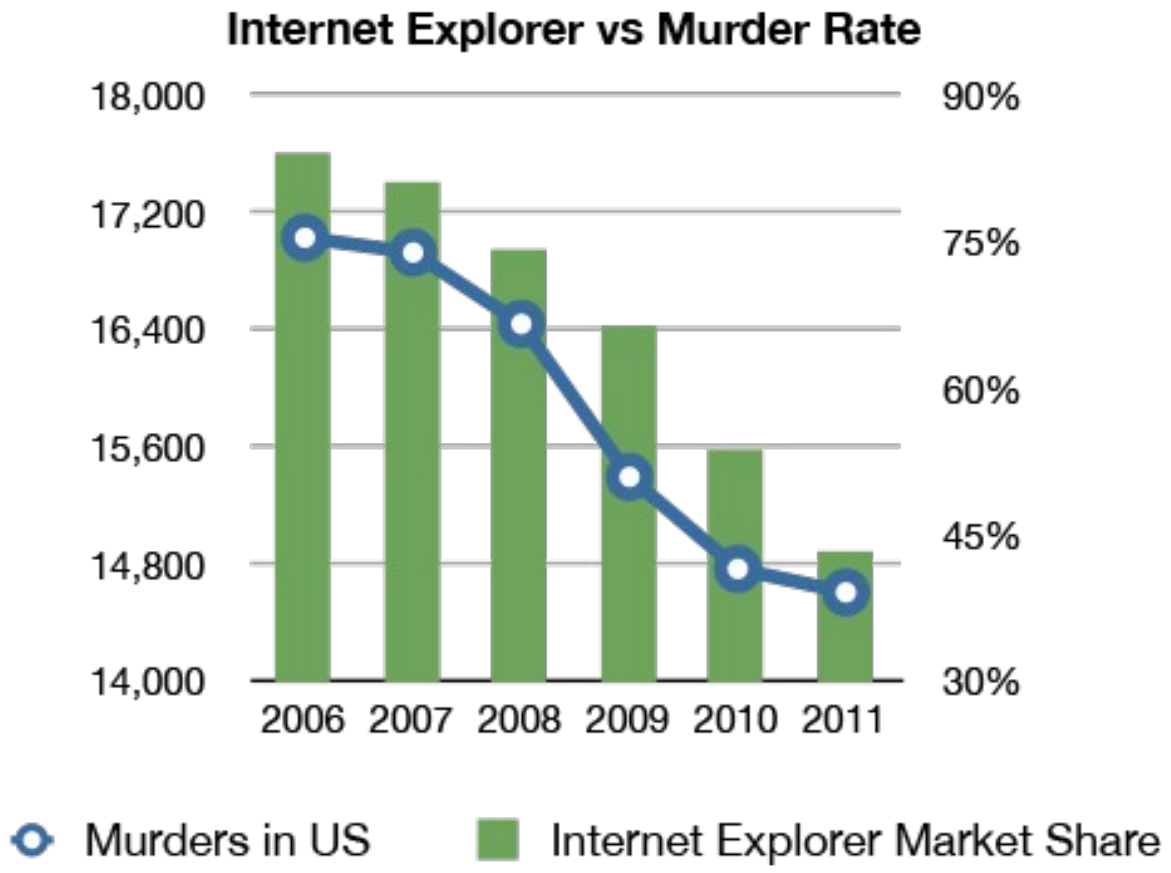


Perfect pie usage

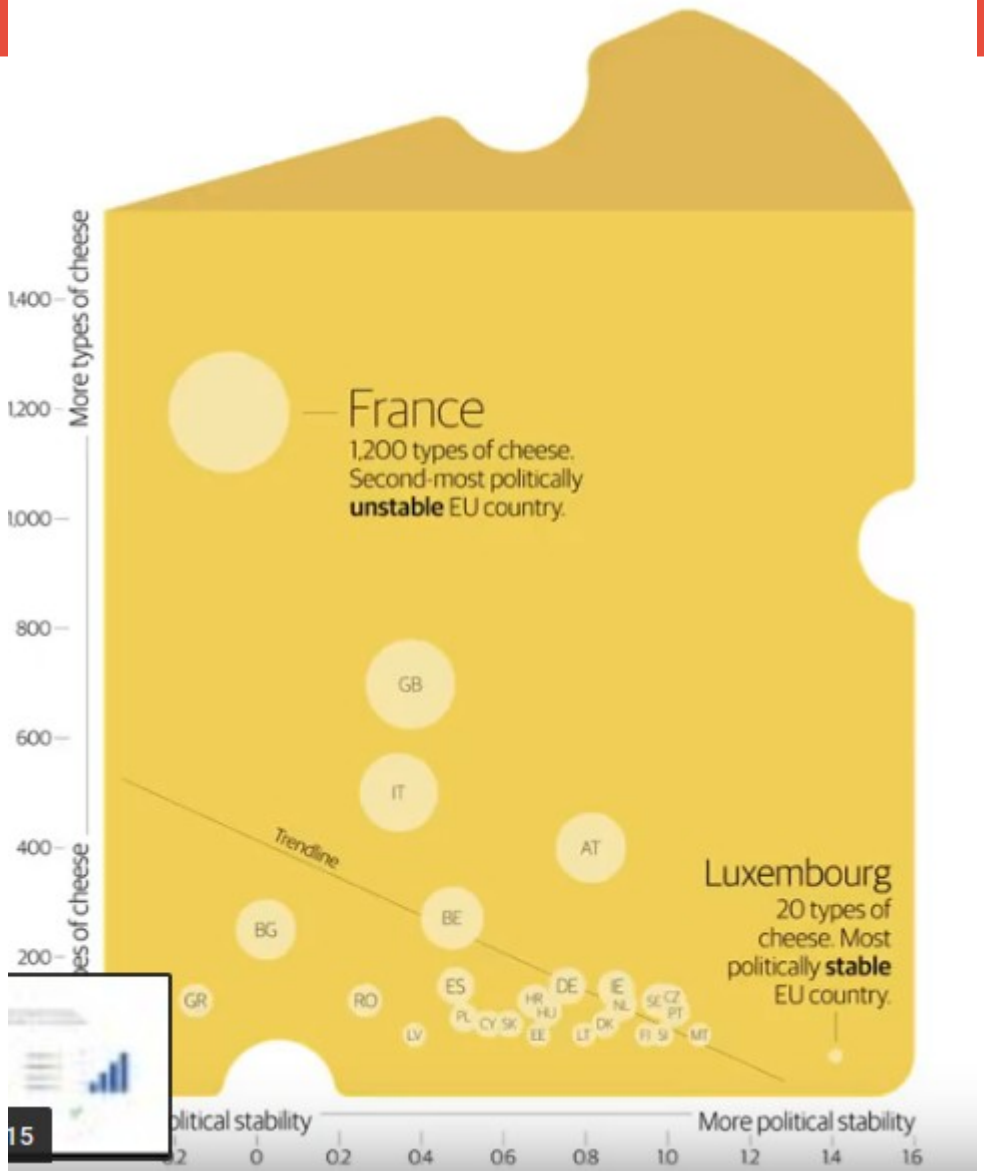
One more faulty chart

Numbers

can be used to prove anything



Not bad, just too rich



Lets have a new look to the first slide.

It was BAAAD

find why

Data Visualization DATAVIZ

The major goals of this course are to understand how visual representations can help in the analysis and understanding of complex data, how to design effective visualizations, and how to create an interactive visualization using modern web-based frameworks.
It explores the underlying theory and practical concepts in creating visual representations of large amounts of data. It covers the core topics in data visualization: data representation, visualization toolkits, scientific visualization, medical visualization, information visualization, flow visualization, and volume rendering techniques

- **Title repetition**
- **Unreadable block of text**
  - Not easy to read on screen
  - No spacing
  - Same term repeated ad nauseam
  - leads to → no idea what was on the slide as soon as slide changes
- **Logo in an unexpected place**
  - Top Left is more common

That was just a warning

Presenting efficiently is another skill

However some items are shared between charting and powerpointing.

(this is outside of course’ scope)

Rules of thumb

A few

(we’ll add more along the way)

Overloading

DO NOT OVERLOAD the chart

Ask yourself

« Do you really think this needs 5 lines ? »

« Where is my area ? » (hidden)

« how to read a bubble-filled bubble chart ? »

« should I compare life expectancy in the last 50 years for all the known countries in the same chart ? »

Organize logically

Choose how you sort data.

Do not go with defaults
(in python, defaults set by the chart library)

Help comparisons

Why having two line charts looking alike side by side

if you can draw two lines in one chart ?

No overplotting

Keep it readable

Remember the ballerinas

Also, think of a bubble chart will all circles grouped

Find a good title

Use another pair of eyes

Ask someone who did not work on your viz

how (s)he understands your viz.

Ask this person the basic questions you wanted to answer with your chart.

That must match

Plenty of fish in the ocean

Some people on the Internet have made a terrific work in classifying the existing data visualizations.

They made the flowchart to decide which chart to use.

Lets have a look at a

https://datavizcatalogue.com/

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

https://www.r-graph-gallery.com/all-graphs/

Bookmark and explore

https://datavizcatalogue.com/

And

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

Are very useful sites to decide what chart to use.

Bookmark them

Explore them

Credits

http://pre15.deviantart.net/f493/th/pre/i/2012/142/1/9/a_picture_is_worth_a_thousand_words_by_shadowxomega5-d50oo8p.png

https://hikingartist.files.wordpress.com/2012/07/a-picture-is-worth-a-thousand-words-3.jpg

https://www.outsidethebeltway.com/farewell-to-lotus-1-2-3/

https://mathix.org/linux/archives/12851

https://assets.lls.fr/pages/7363852/HIS.2de.AP.graphique_greve_.jpg

Recommended book

Data Visualization with Python

Mario Döbler and Tim Großmann - Packt Publishing

Exercises

List the steps to create the best chart possible

Get DASH

Install

Get sample project

Adapt it (or get Sakila DB)

Lets pick a few dataviz and for each one, lets find

A case when it is perfect (explain why)

A case when it is a bad choice (explain differently)

2 dataviz per student, please synch yourselves

Will be shared with group later