Seminario 4: Visualización para Análisis Exploratorio de Datos

Jorge Casillas
Universidad de Granada
http://decsai.ugr.es/~casillas

Inteligencia de Negocio, 2018-2019

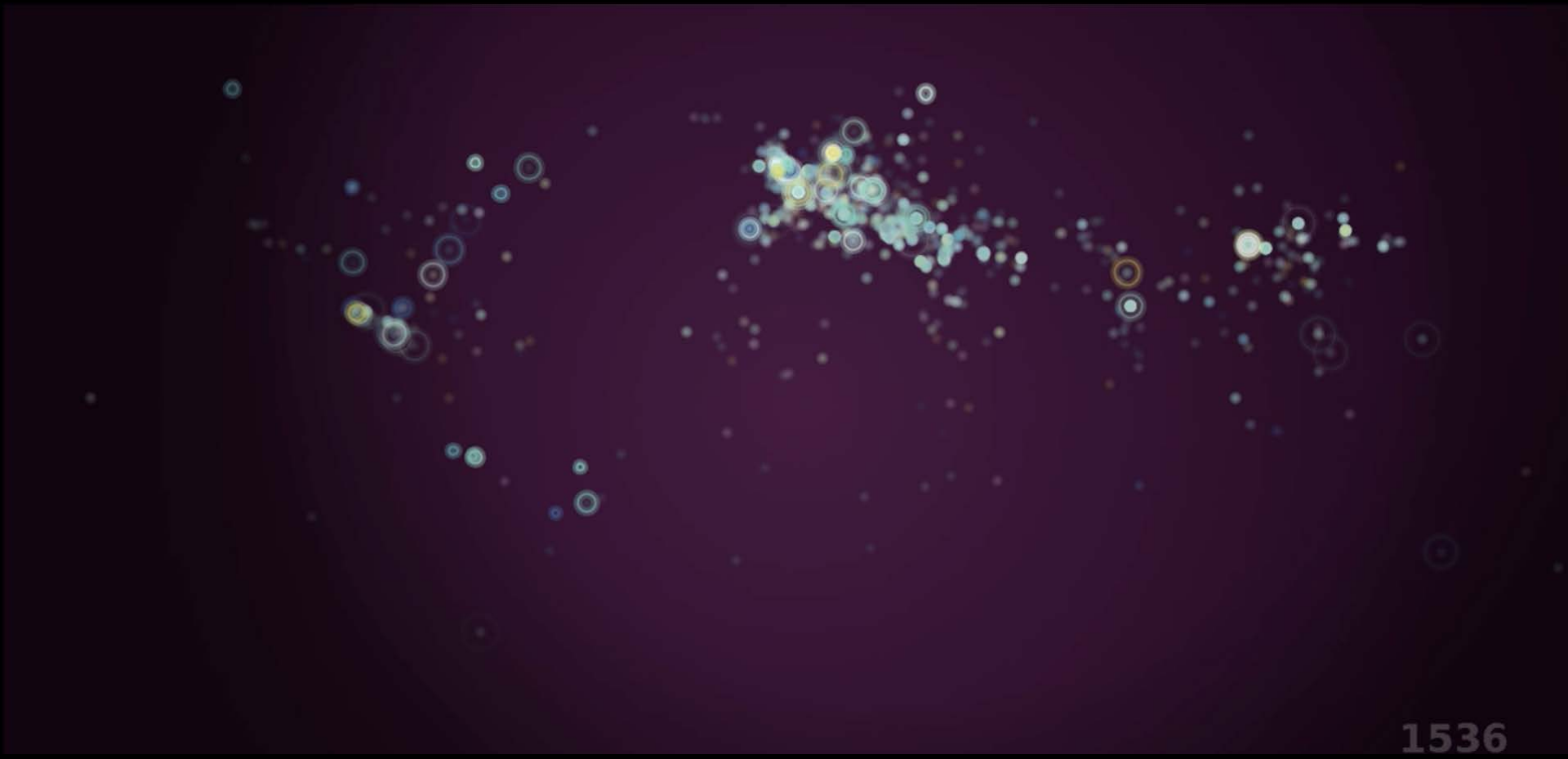
Motivación

- Presentar los datos de un modo visual distinto al habitual nos permite comprender tendencias, cambios, lagunas, relaciones, direcciones…
- Son aspectos difíciles (a veces imposibles) de tratar mediante aprendizaje automático pero que se pueden captar de modo muy contundente con técnicas de visualización avanzadas
- Estas técnicas no tienen generalmente una definición formal, ni poseen significancia estadística, quizás no expliquen las razones, quizás no den suficientes detalles… pero en cualquier caso ayudan a un ser humano a comprender de un vistazo una realidad compleja

Motivación

- Cuanto más compleja sea esa realidad, más necesario y útil resulta el análisis exploratorio de datos mediante visualización
- Cada vez disponemos de más y más datos, y cada vez tenemos sistemas de computación más potentes para procesarlos, de modo que cada vez está adquiriendo más importancia la visualización
- Una imagen es una poderosa herramienta para instalar conceptos en el proceso cognitivo, lo cual ayuda a comunicar mensajes y facilitar la toma de decisiones

¿Se puede resumir la historia de la humanidad en 100 segundos?

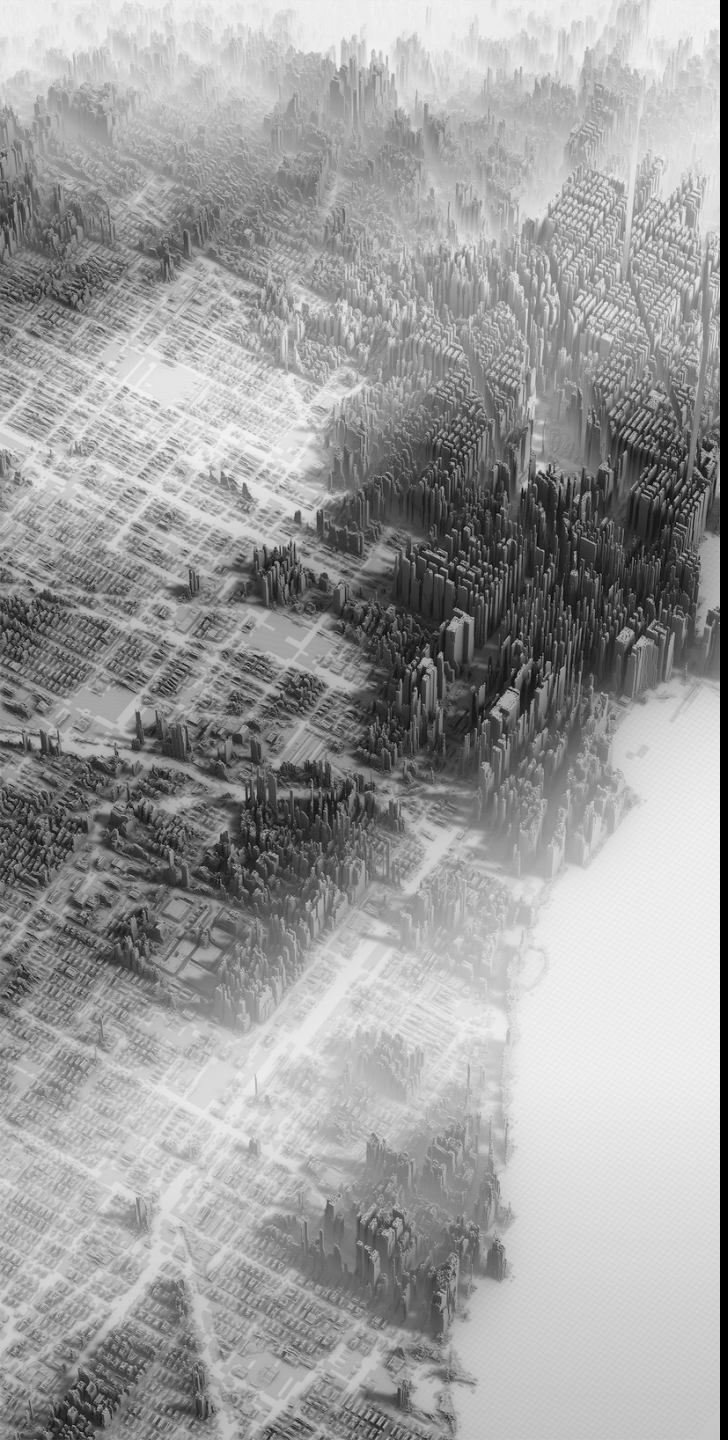


http://vimeo.com/19088241

Visualización

- Millones de datos pueden componer una imagen elocuente y de especial belleza

AZUL = lugares donde se usa Twitter, **ROJO** = lugares donde se tomaron fotos en Flickr, **BLANCO** = ambos



Chicago

Los Angeles

Herwig Scherabon

Desigualdad de ingresos

Visualización

- A veces, una forma ingeniosa de presentar datos simples aporta un gran impacto

The **True Size** of **Africa**

A small contribution in the fight against rampant *Immappancy*, by **Kai Krause**

Graphic layout for visualization only (some countries are cut and rotated)
But the conclusions are very accurate: refer to table below for exact data

COUNTRY	AREA x 1000 km²
China	9.597
USA	9.629
India	3.287
Mexico	1.964
Peru	1.285
France	633
Spain	506
Papua New Guinea	462
Sweden	441
Japan	378
Germany	357
Norway	324
Italy	301
New Zealand	270
United Kingdom	243
Nepal	147
Bangladesh	144
Greece	132
TOTAL	**30.102**
AFRICA	**30.221**

In addition to the well known social issues of *illiteracy* and *innumeracy*, there also should be such a concept as "*immappancy*", meaning *insufficient geographical knowledge.*

A survey with random American schoolkids let them guess the population and land area of their country. Not entirely unexpected, but still rather unsettling, the majority chose *"1-2 billion"* and *"largest in the world"*, respectively.

Even with Asian and European college students, geographical estimates were often off by factors of *2-3*. This is partly due to the highly distorted nature of the predominantly used mapping projections (such as *Mercator*).

A particularly extreme example is the worldwide misjudgement of the true size of *Africa*. This single image tries to embody the massive scale, which is larger than the *USA, China, India, Japan* and *all of Europe..... combined!*

Top 100 Countries

Area in square kilometers, Percentage of World Total
Sources: Britannica, Wikipedia, Almanac 2010

		AREA km²	%
1	Russia	17.098.242	11,50
2	Canada	9.984.670	6,70
3	China	9.596.961	6,40
4	United States	9.629.091	6,40
5	Brazil	8.514.877	5,70
6	Australia	7.692.024	5,20
7	India	3.287.263	2,30
8	Argentina	2.780.400	2,00
9	Kazakhstan	2.724.900	1,80
10	Sudan	2.505.813	1,70
11	Algeria	2.381.741	1,60
12	Congo	2.344.858	1,60
13	Greenland	2.166.086	1,50
14	Saudi Arabia	2.149.690	1,40
15	Mexico	1.964.375	1,30
16	Indonesia	1.860.360	1,30
17	Libya	1.759.540	1,20
18	Iran	1.628.750	1,10
19	Mongolia	1.564.100	1,10
20	Peru	1.285.216	0,86
21	Chad	1.284.000	0,86
22	Niger	1.267.000	0,85
23	Angola	1.246.700	0,85
24	Mali	1.240.192	0,83
25	South Africa	1.221.037	0,82
26	Colombia	1.141.748	0,76
27	Ethiopia	1.104.300	0,74
28	Bolivia	1.098.581	0,74
29	Mauritania	1.025.520	0,69
30	Egypt	1.002.000	0,67
31	Tanzania	945.087	0,63
32	Nigeria	923.768	0,62
33	Venezuela	912.050	0,61
34	Namibia	824.116	0,55
35	Mozambique	801.590	0,54
36	Pakistan	796.095	0,53
37	Turkey	783.562	0,53
38	Chile	756.102	0,51
39	Zambia	752.612	0,51
40	Myanmar	676.578	0,45
41	Afghanistan	652.090	0,44
42	Somalia	637.657	0,43
43	France	632.834	0,43
44	C. African Rep	622.984	0,42
45	Ukraine	603.500	0,41
46	Madagascar	587.041	0,39
47	Botswana	582.000	0,39
48	Kenya	580.367	0,39
49	Yemen	527.968	0,35
50	Thailand	513.120	0,34
51	Spain	505.992	0,34
52	Turkmenistan	488.100	0,33
53	Cameroon	475.442	0,32
54	Papua New Guinea	462.840	0,31
55	Uzbekistan	447.400	0,30
56	Morocco	446.550	0,30
57	Sweden	441.370	0,30
58	Iraq	438.317	0,29
59	Paraguay	406.752	0,27
60	Zimbabwe	390.757	0,26
61	Japan	377.930	0,25
62	Germany	357.114	0,24
63	Rep o.t. Congo	342.000	0,23
64	Finland	338.419	0,23
65	Vietnam	331.212	0,22
66	Malaysia	330.803	0,22
67	Norway	323.802	0,22
68	Côte d'Ivoire	322.463	0,22
69	Poland	312.685	0,21
70	Oman	309.500	0,21
71	Italy	301.336	0,20
72	Philippines	300.000	0,20
73	Burkina Faso	274.222	0,18
74	New Zealand	270.467	0,18
75	Gabon	267.668	0,18
76	Western Sahara	266.000	0,18
77	Ecuador	256.369	0,20
78	Guinea	245.857	0,17
79	United Kingdom	242.900	0,16
80	Uganda	241.038	0,16
81	Ghana	238.539	0,16
82	Romania	238.391	0,16
83	Laos	236.800	0,16
84	Guyana	214.969	0,14
85	Belarus	207.600	0,14
86	Kyrgyzstan	199.951	0,13
87	Senegal	196.722	0,13
88	Syria	185.180	0,12
89	Cambodia	181.035	0,12
90	Uruguay	176.215	0,12
91	Suriname	163.820	0,11
92	Tunisia	163.610	0,11
93	Nepal	147.181	0,10
94	Bangladesh	143.998	0,10
95	Tajikistan	143.100	0,10
96	Greece	131.957	0,09
97	Nicaragua	130.373	0,09
98	North Korea	120.538	0,08
99	Malawi	118.484	0,08
100	Eritrea	117.600	0,08
	TOP 100 TOTAL	132.632.524	89,34


United States


Europe


India

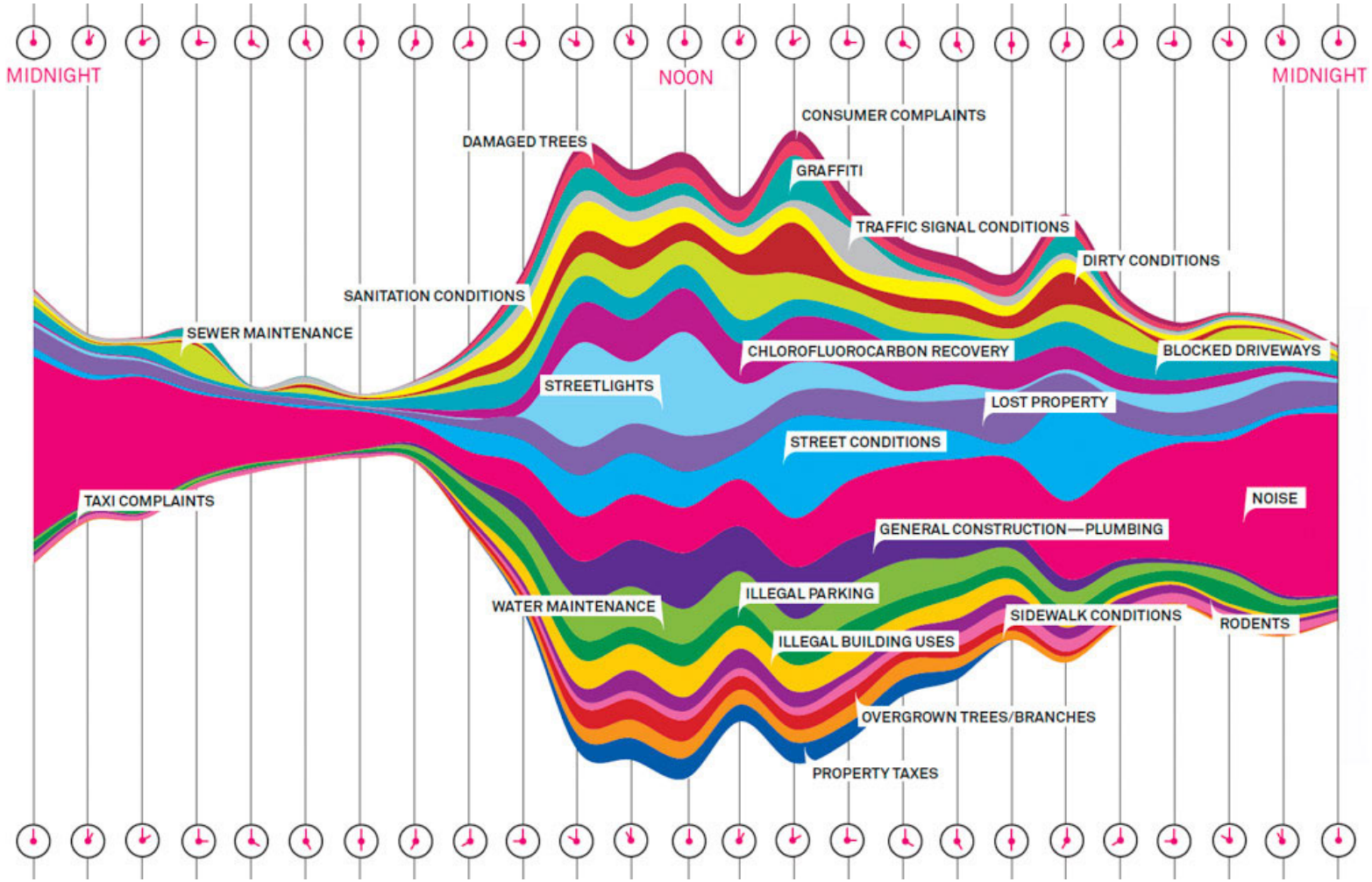

Japan


China

Visualización

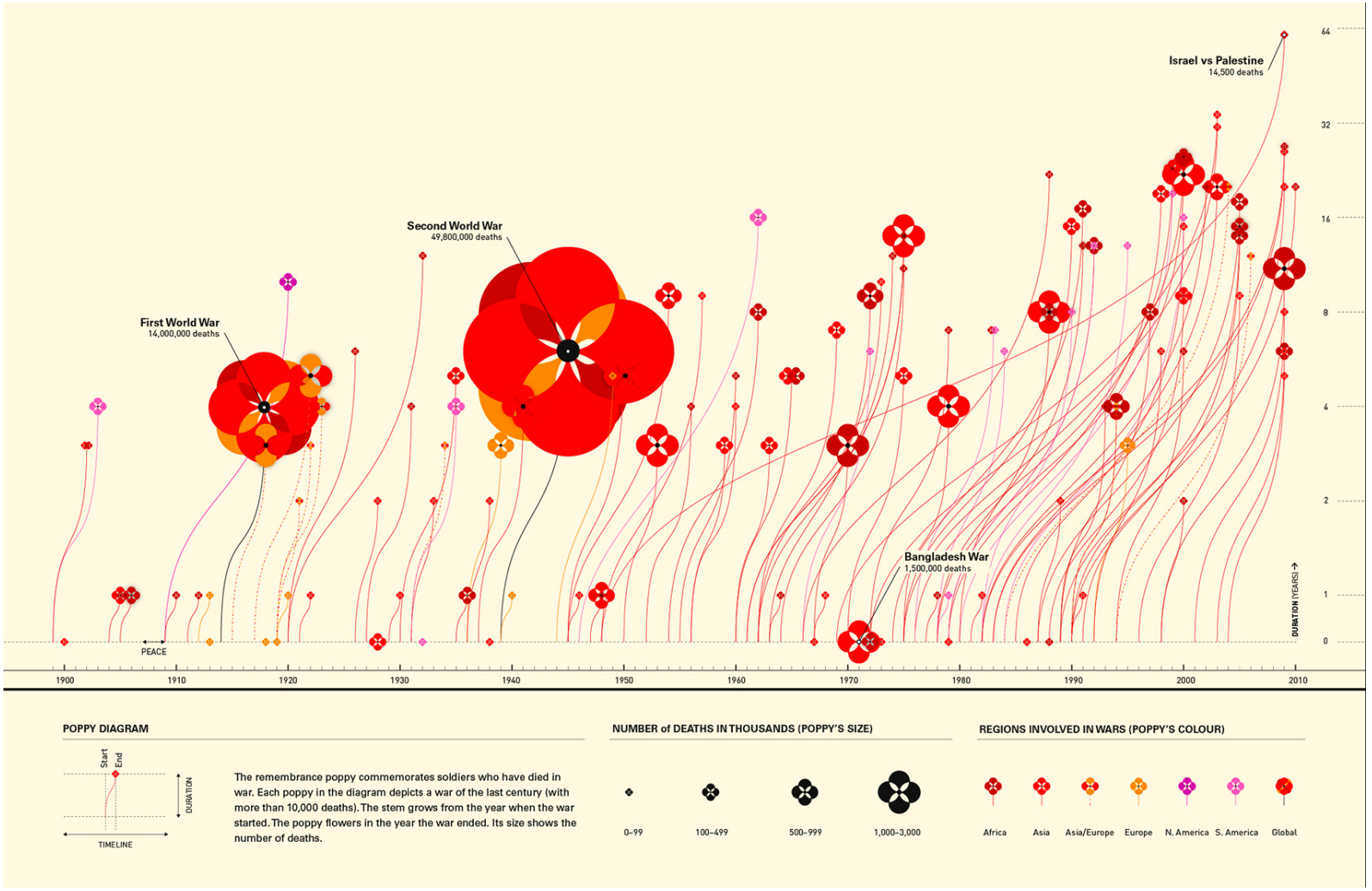
- Cada historia que quieras contar tiene un tipo de gráfico diferente

Llamadas de quejas en Nueva York



Themeriver o *streamgraph*: https://plotdb.com/chart/1038
http://bl.ocks.org/mbostock/4060954

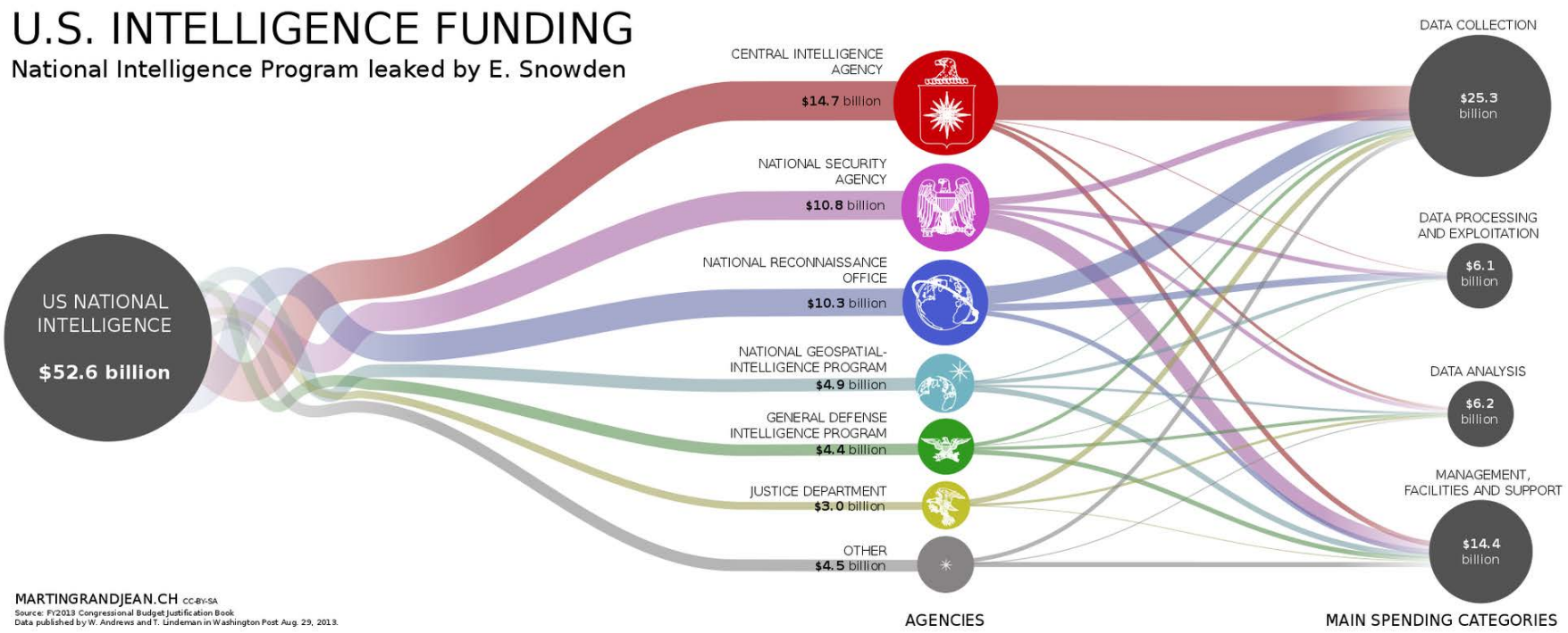
Las guerras del siglo XX y XXI



Fuente: http://www.poppyfield.org/

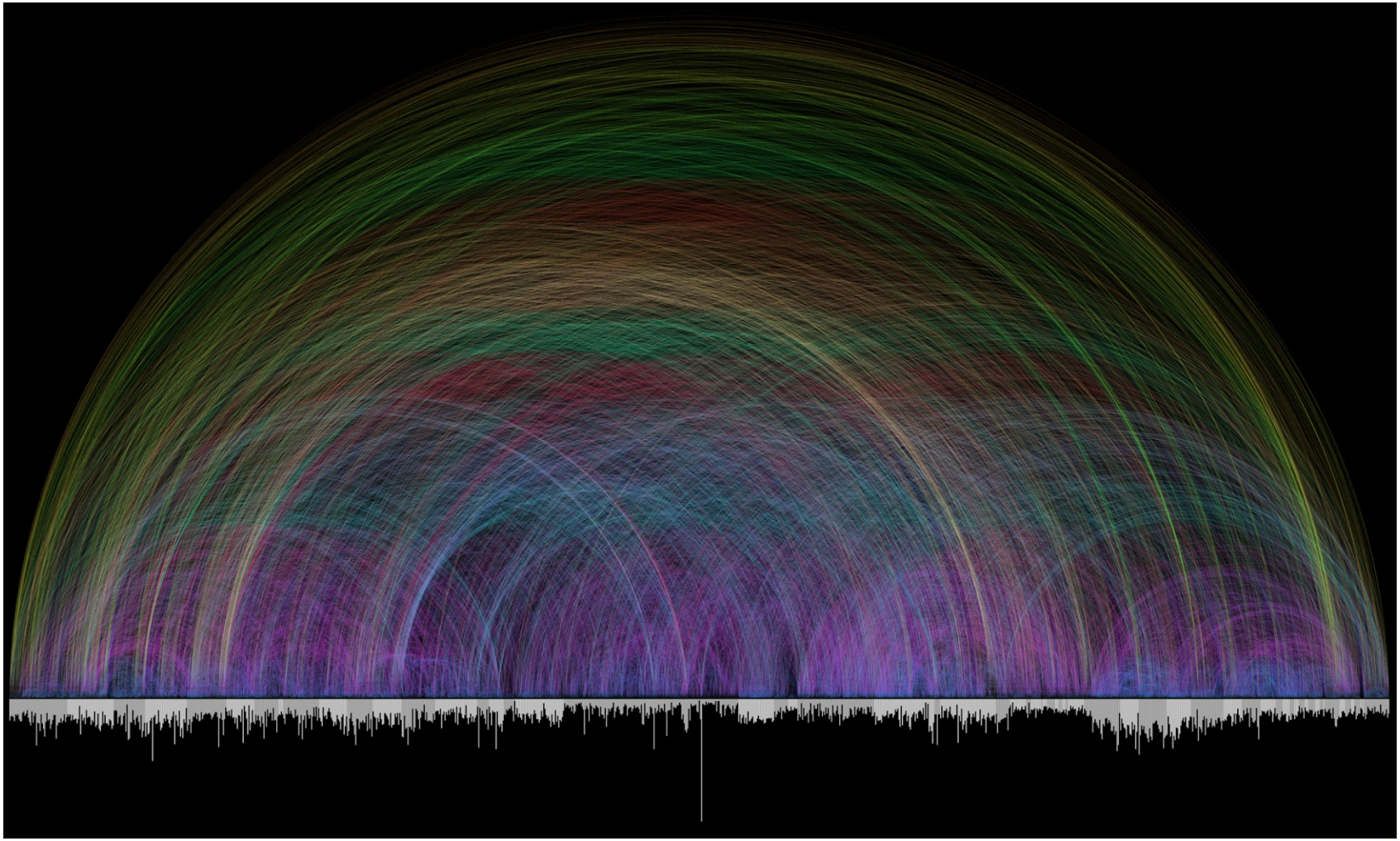
Line bubbles: https://plotdb.com/chart/1045/

Presupuestos del NIP en EE.UU.



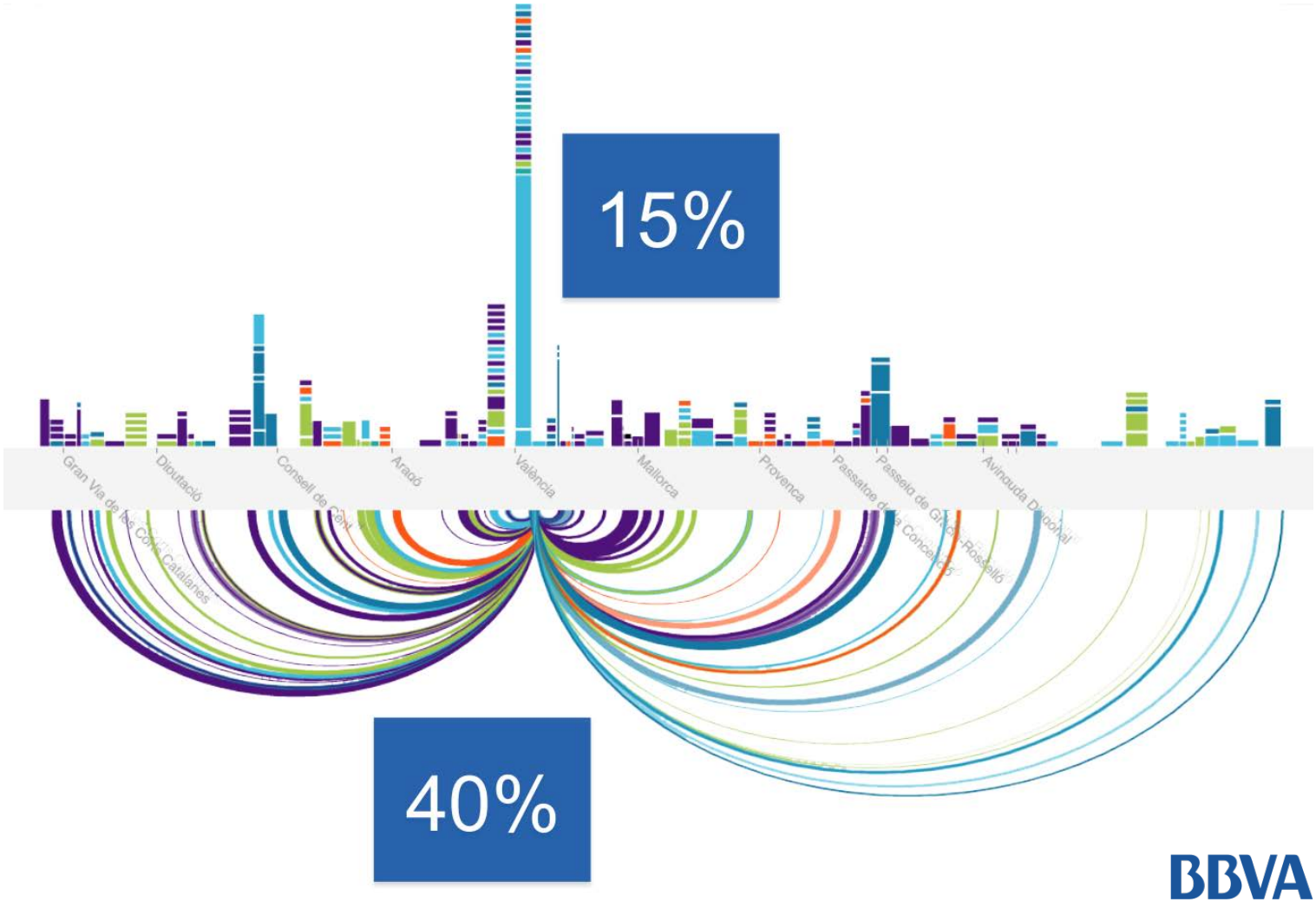
Sankey diagram: http://bost.ocks.org/mike/sankey/
http://sankey.csaladen.es/

Referencias de la Biblia



http://www.chrisharrison.net/index.php/Visualizations/BibleViz
http://bibviz.com/

¿Cuál es el comercio más influyente en Paseo de Gracia (Barcelona)?



BBVA

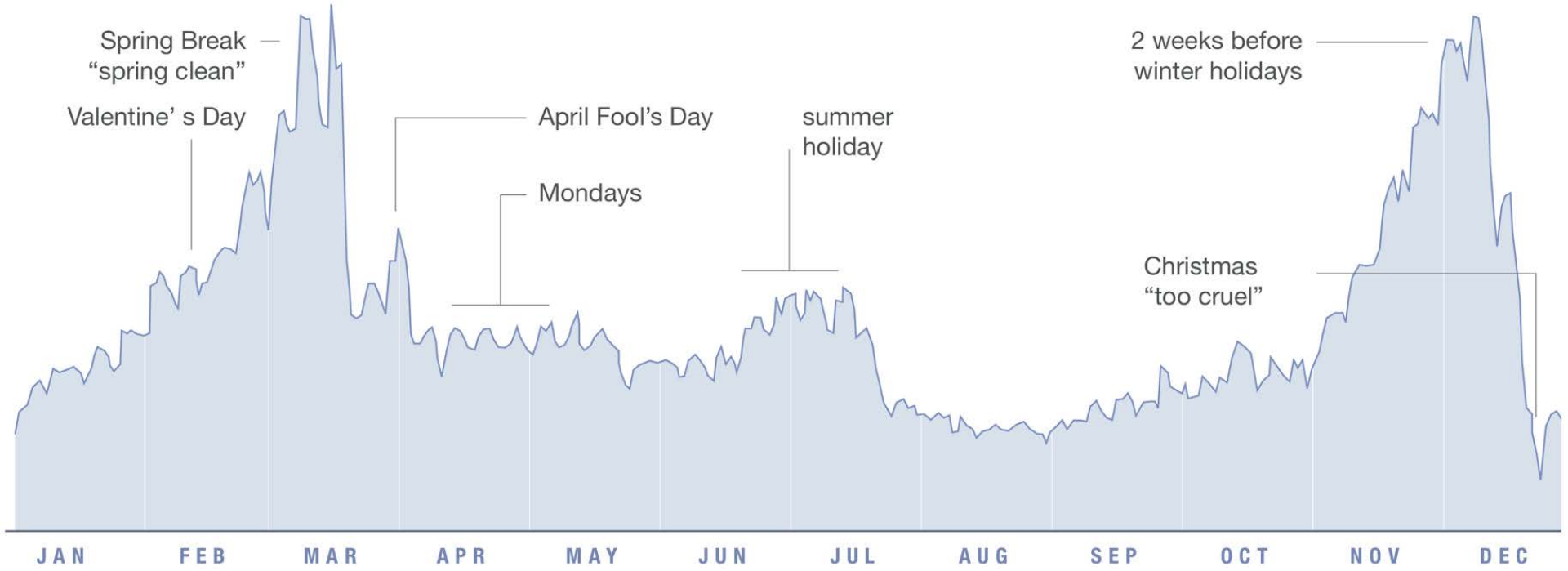
¿Cuál es el comercio más influyente en Paseo de Gracia (Barcelona)?

Visualización

- A veces la originalidad no está en la forma de representar los datos, sino en lo que se descubre al visualizarlos

¿Qué datos representan esta gráfica?

Momentos en los que se anuncian rupturas de pareja en Facebook



Resumen:

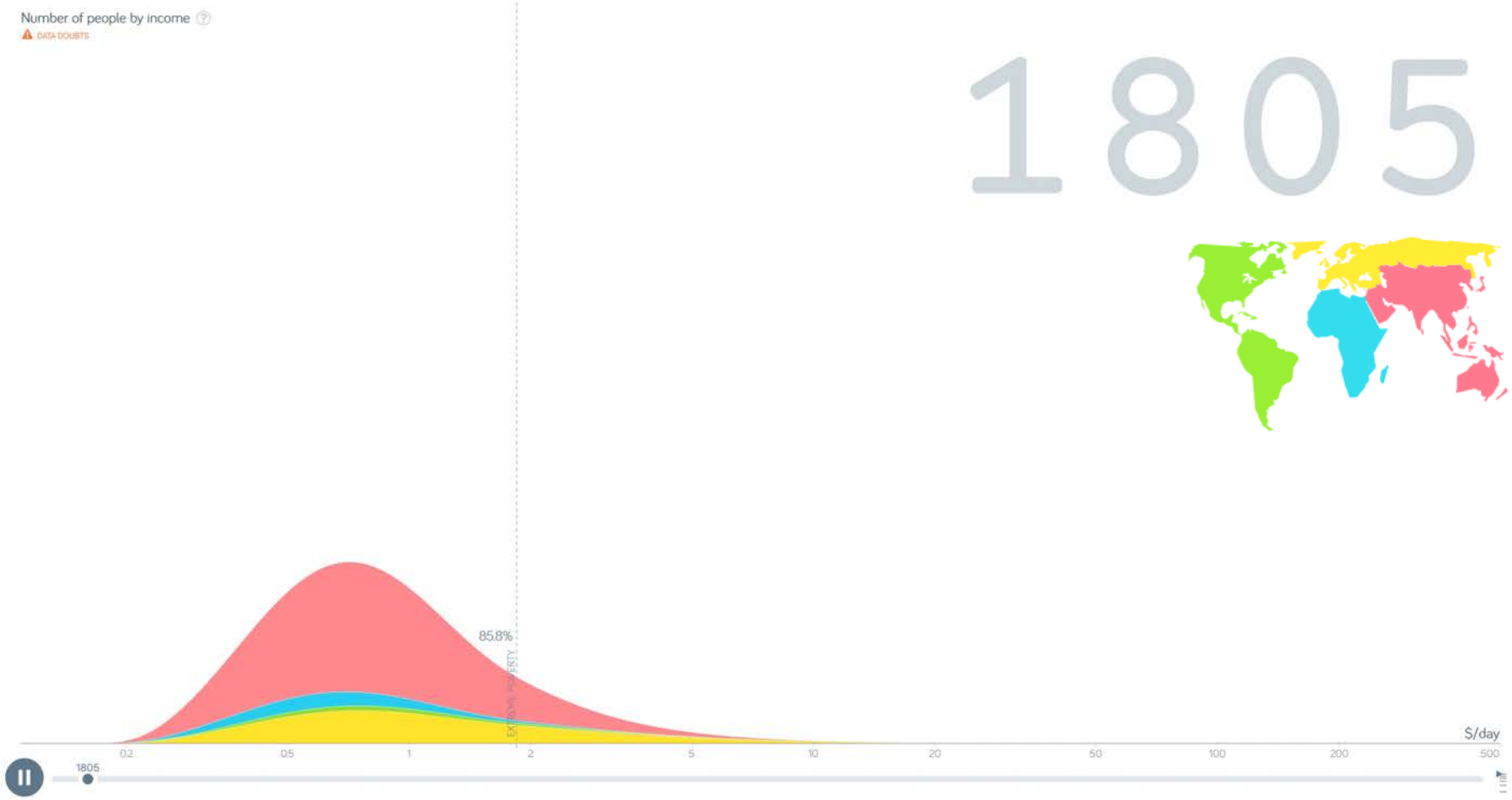
- Un gran pico justo antes de vacaciones de primavera
- La mayoría de rupturas se anuncian los lunes
- A la gente le gusta comenzar el verano estando soltero/a
- Un gran pico justo antes de Navidades
- El día con menos rupturas de todo el año es el día de Navidad (¡gracias a Dios!)

https://informationisbeautiful.net/

Visualización

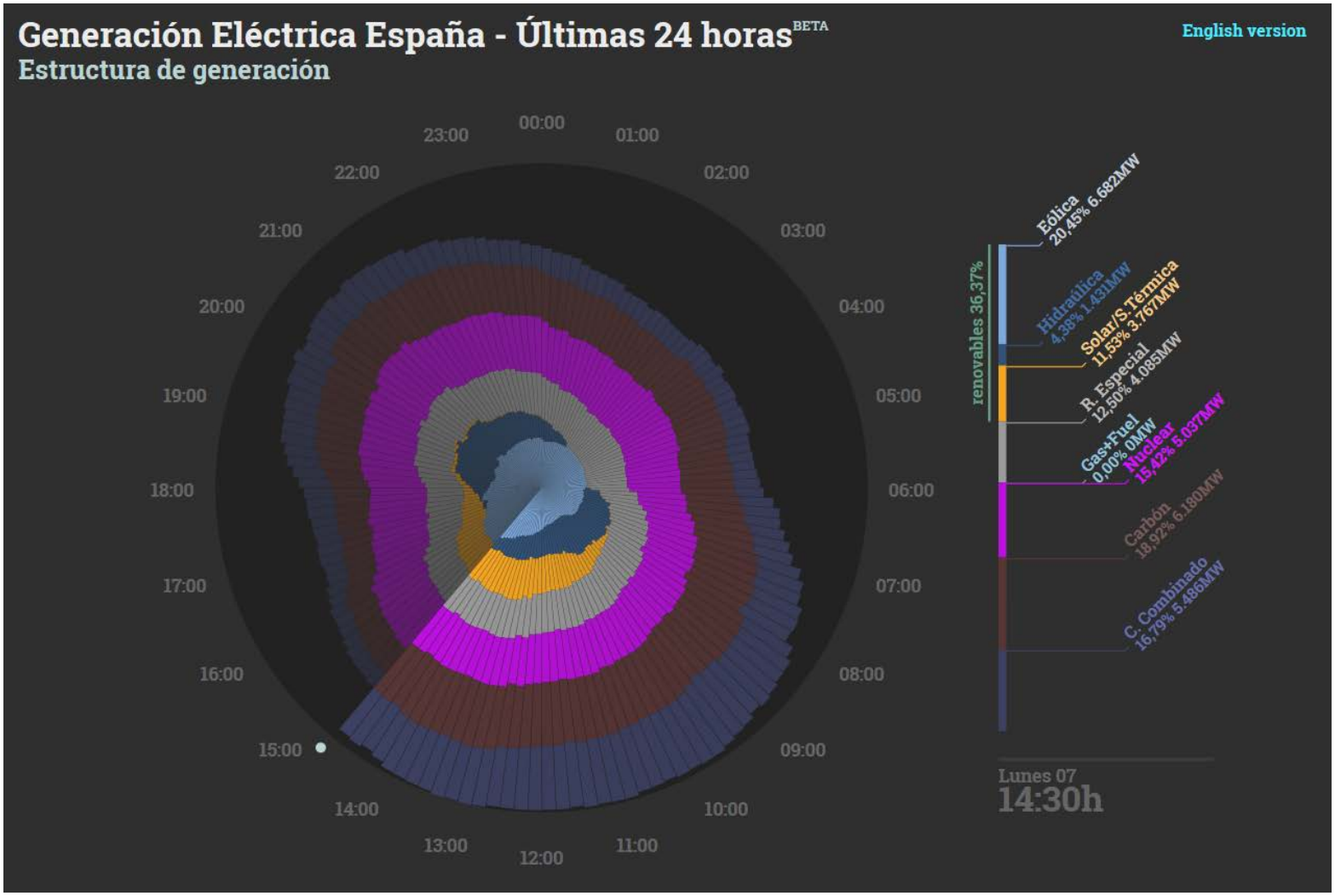
- Una gráfica sencilla pero que evolucione a lo largo del tiempo y/o sea interactiva también ayuda a comprender muchos problemas

Evolución del reparto de la riqueza



https://www.gapminder.org/tools

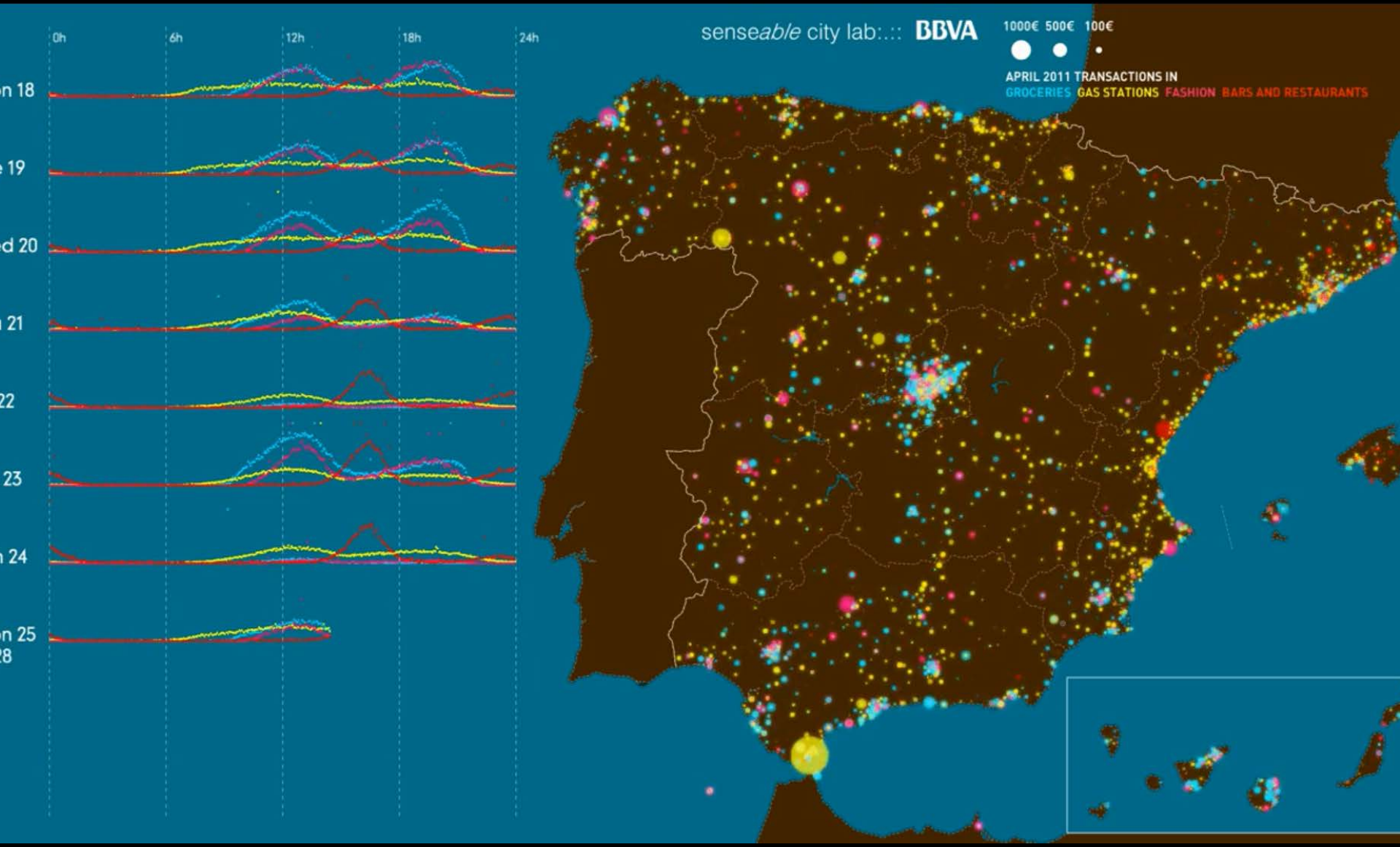
Generación Eléctrica en España (últimas 24h)



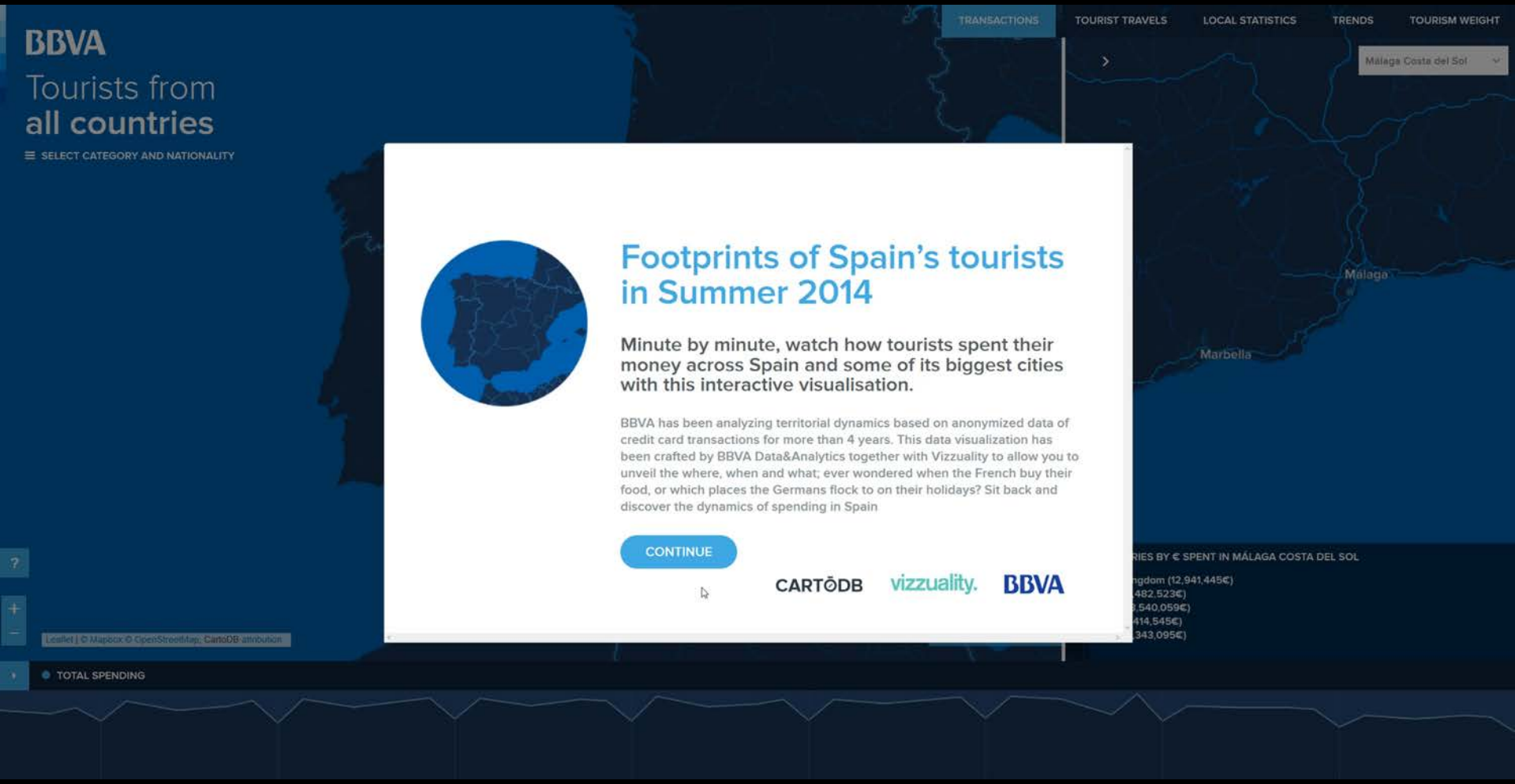
http://energia.ningunaparte.net/

Visualización

- Los vídeos sobre visualizaciones de datos ayudan a comprender dinámicas y tendencias

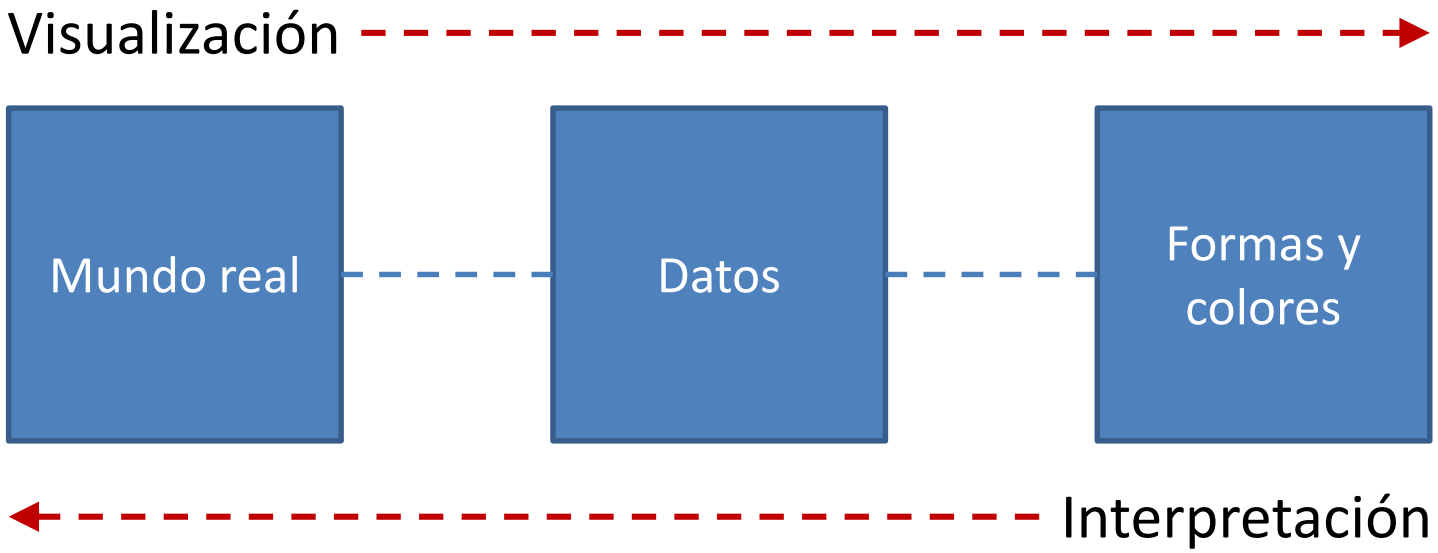


https://youtu.be/8J3T3UjHbrE

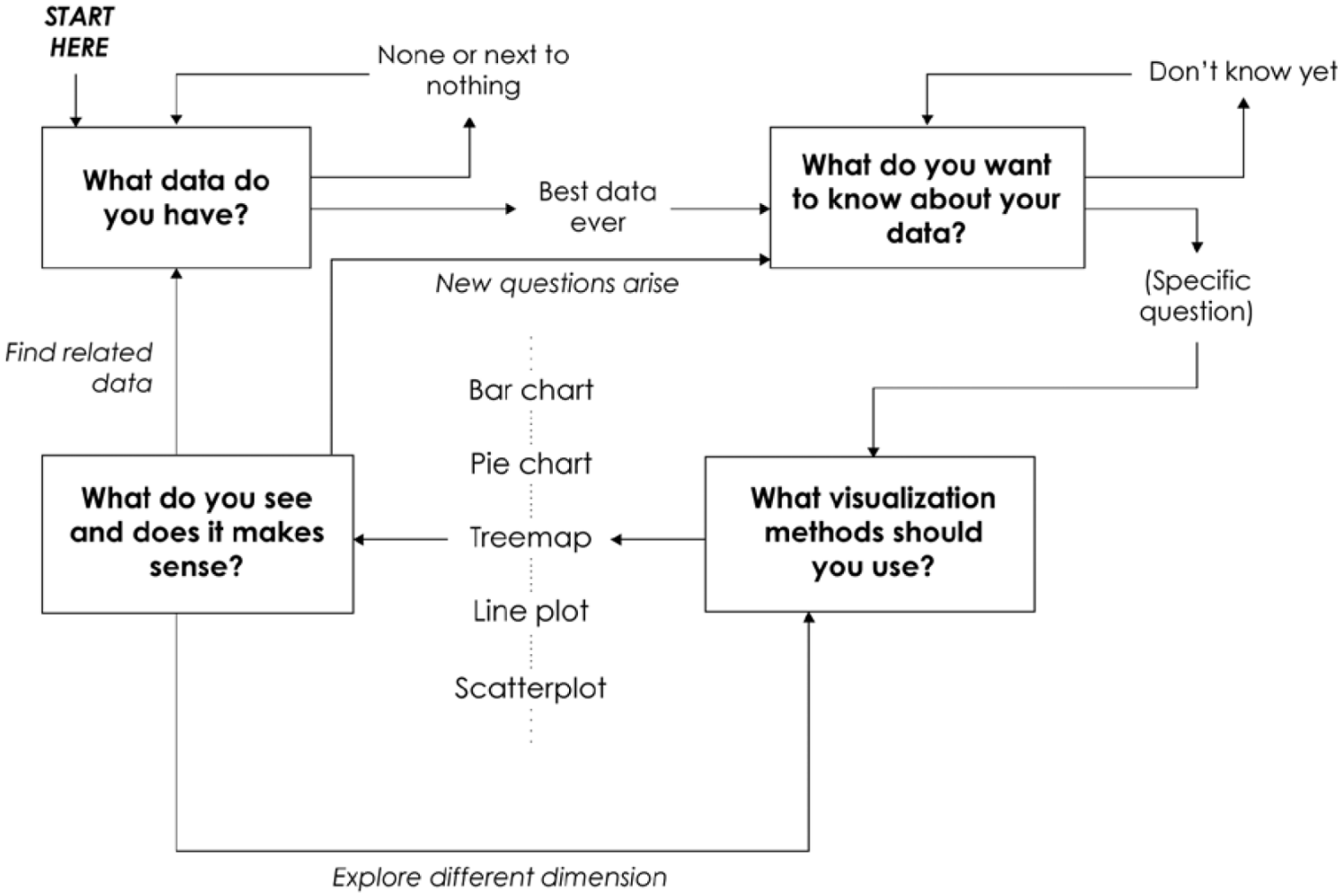


http://bbvatourism.vizzuality.com/

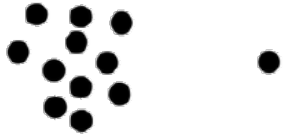
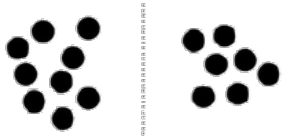
Visualización → Interpretación



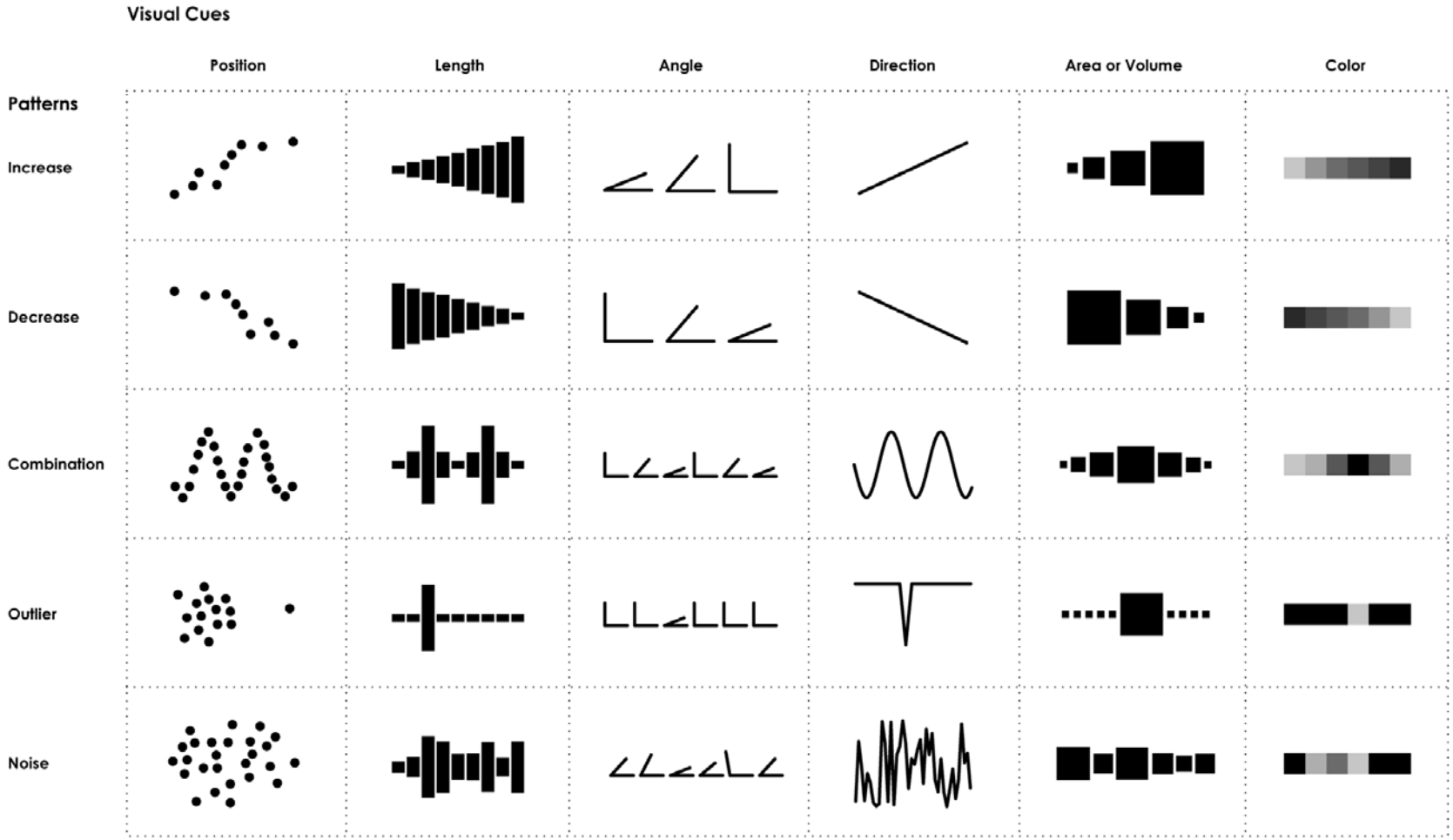
Proceso Iterativo de Exploración de Datos



¿Qué quiero representar y cómo hacerlo?

Possible questions Fill in the blanks	**Statistical concepts**	**Possible visuals**
What _____ is the best and worst?	Maximums and minimums	
How has _____ changed over time?	Temporal patterns	
What _____ stands out from the rest?	Outliers	
What makes _____ different from _____?	Clustering	
How are _____ and _____ related to each other?	Correlation	

¿Qué quiero representar y cómo hacerlo?



©Nathan Yau

Tipos de Datos ↔ Modos de Visualización

- Principales tipos de datos:
  - Datos categóricos
  - Series temporales
  - Datos espaciales
  - Múltiples variables
  - Distribuciones

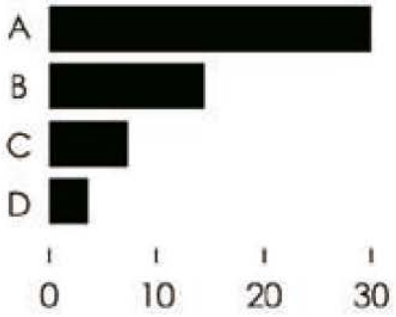
Datos Categóricos

Datos Categóricos

Categories

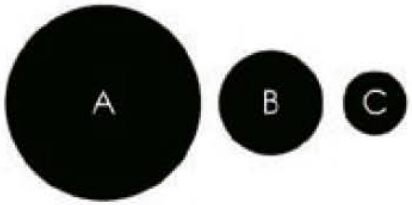
When your data is straightforward, with a value for each category, these are easy to read and create.

Bar graph



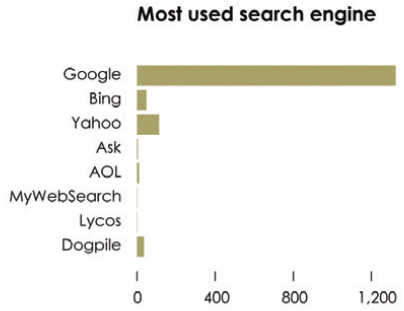
With length as visual cue, useful for straightforward comparisons

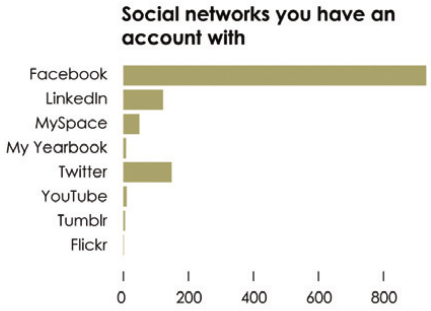
Symbol plot

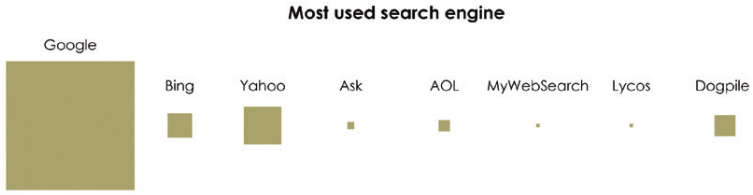


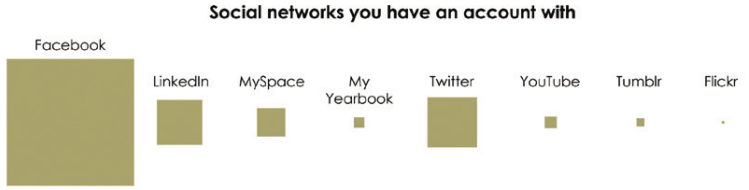
Can be used in place of bars, but can be hard to see small differences

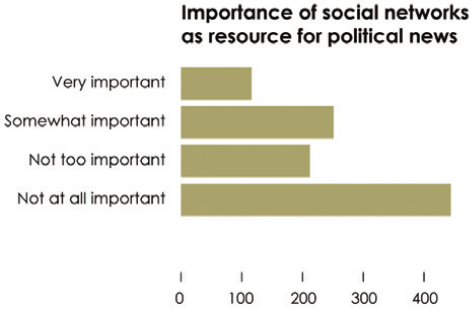
Datos Categóricos

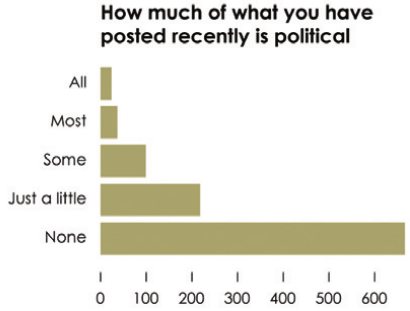


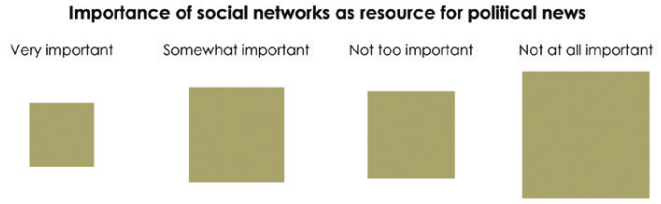


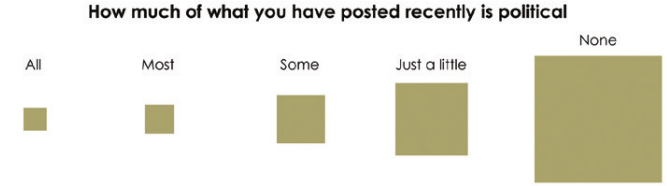










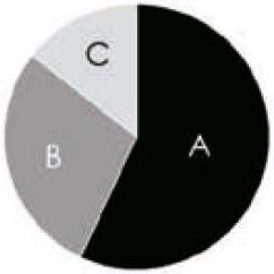


Datos Categóricos

Parts of a whole

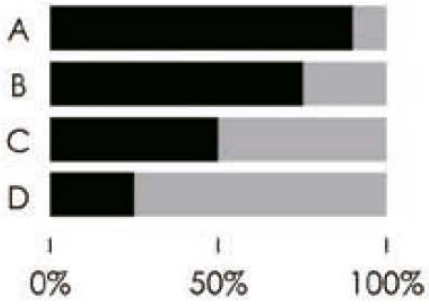
The categorical breakdown within a population can be interesting, and you might want to keep the groups together, although often not essential.

Pie chart



Parts add to 100 percent, typically sorted clockwise for readability

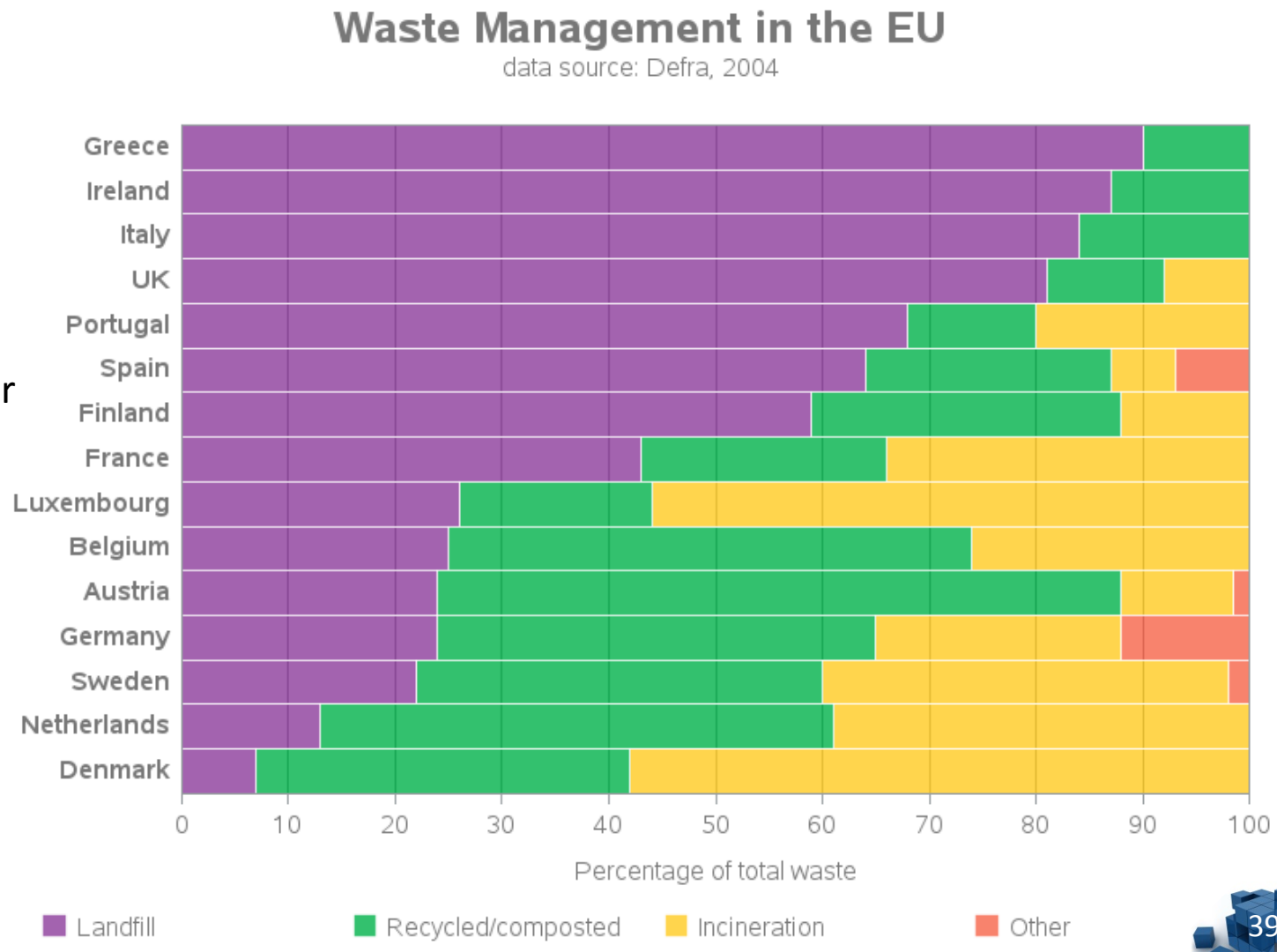
Stacked bar chart



Often used to show poll results and can also be used for raw counts

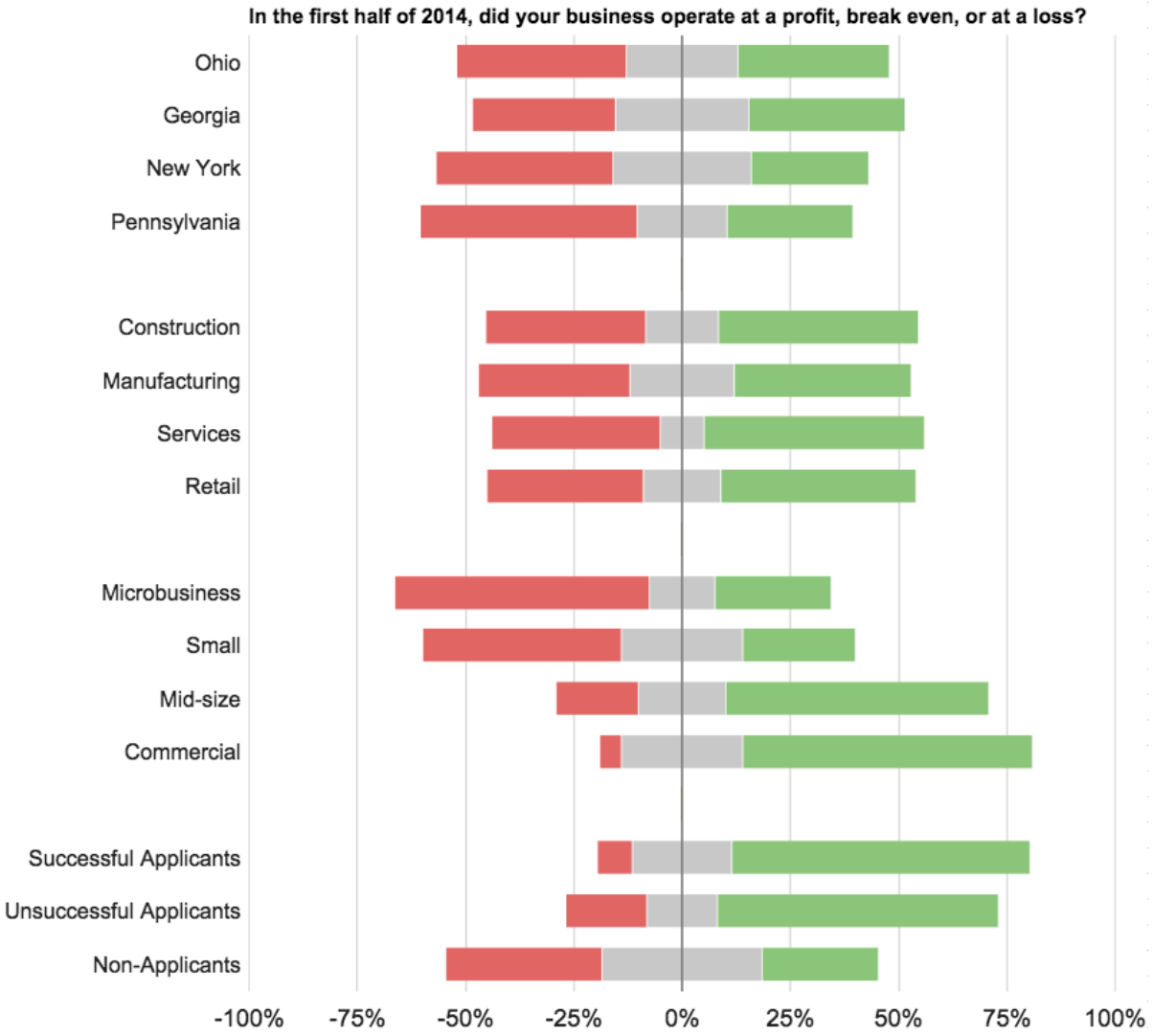
Datos Categóricos

Una gráfica de barras normalizadas ayuda a enfatizar el reparto entre categorías en lugar de la magnitud

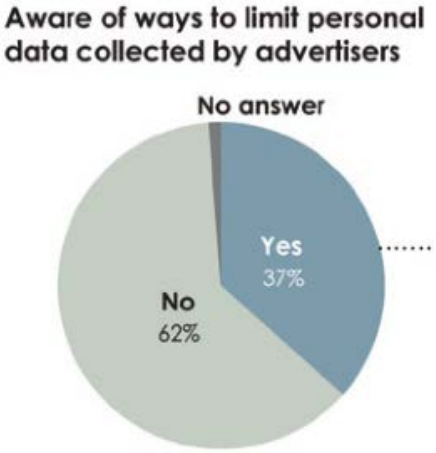


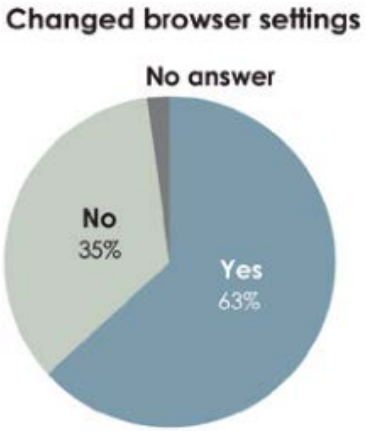
Datos Categóricos

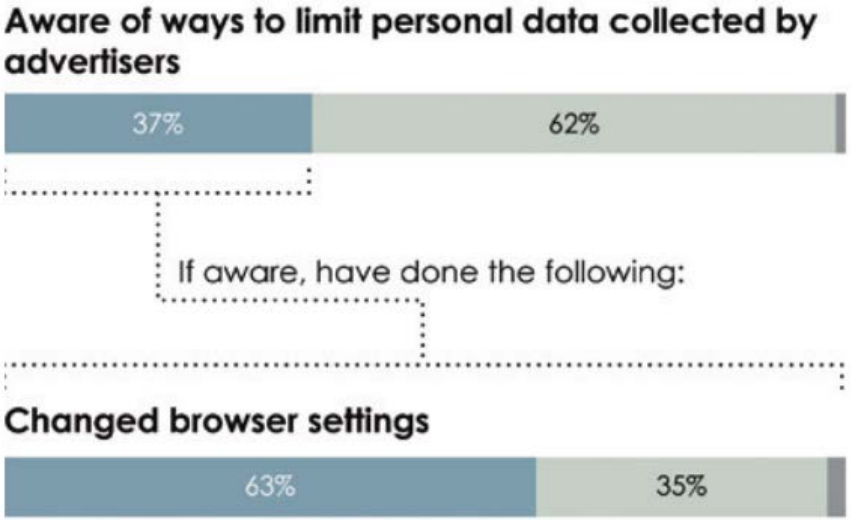
Los resultados de un cuestionario se pueden reflejar bien con barras horizontales, colores para cada respuesta y centrados en la respuesta media



Datos Categóricos





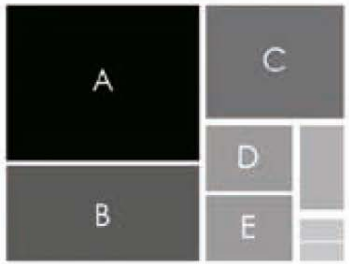


Datos Categóricos

Subcategories

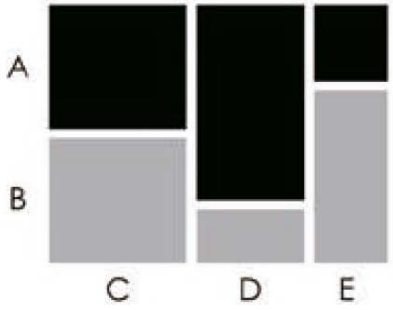
Data can have a hierarchical structure, which can be important in data interpretation and it often allows for different points of view.

Treemap



Shows hierarchical structure in a compact space, area often combined with color

Mosaic plot



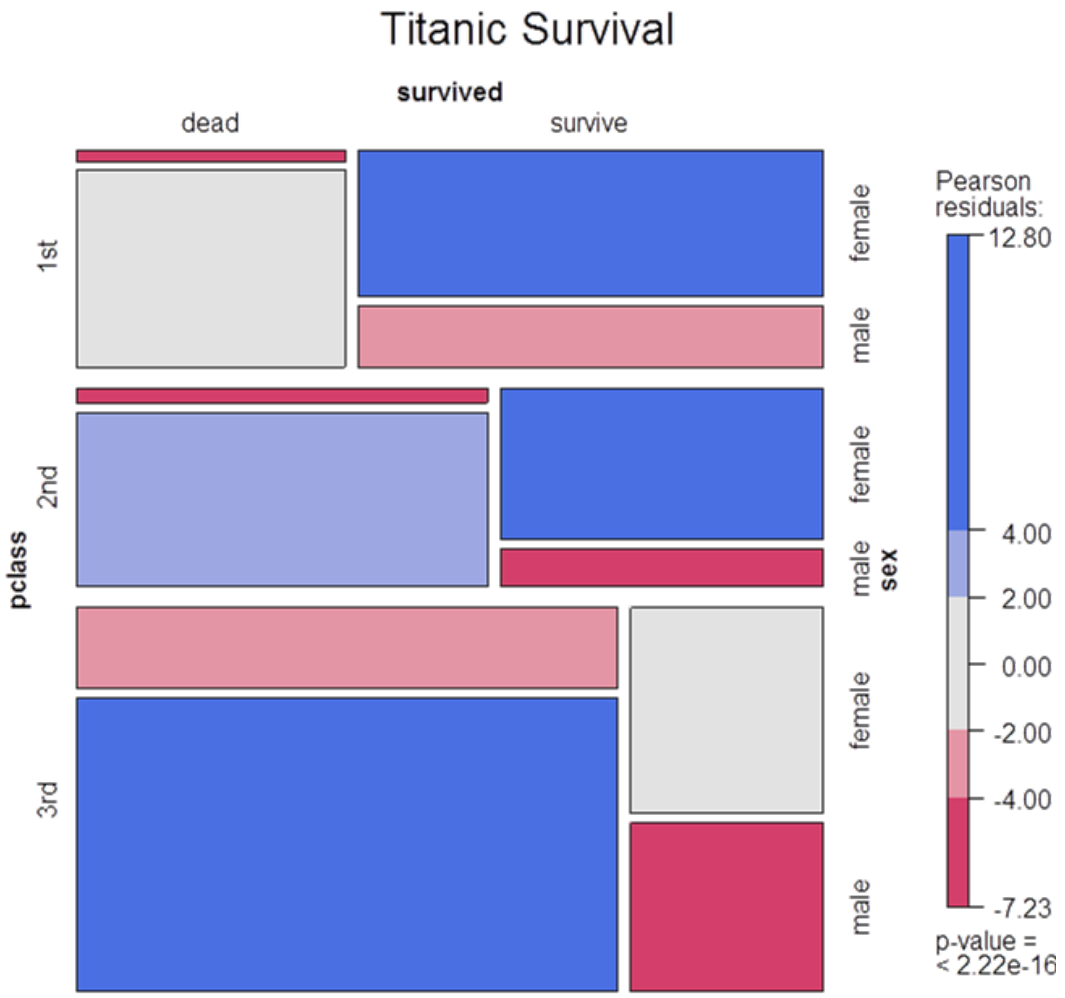
Allows comparisor across multiple categories in one view

Treemap: http://bl.ocks.org/mbostock/4063582

Datos Categóricos

Sunburst: https://www.aculocity.com/labs/sunburst-chart

Datos Categóricos

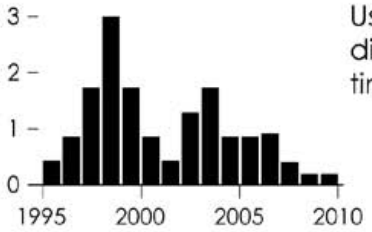


Mosaic plot: http://www.datavis.ca/online/mosaics/about.html

Series Temporales

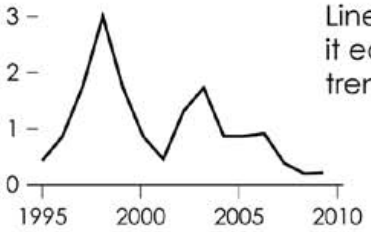
Series Temporales

Bar graph



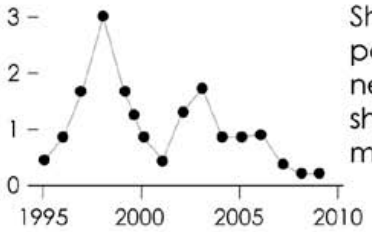
Useful for discrete points in time

Line chart



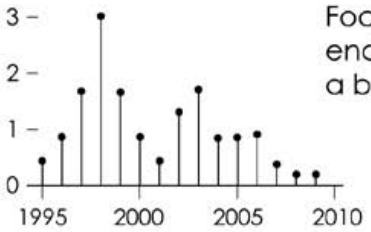
Lines can make it easier to see trends

Dot plot



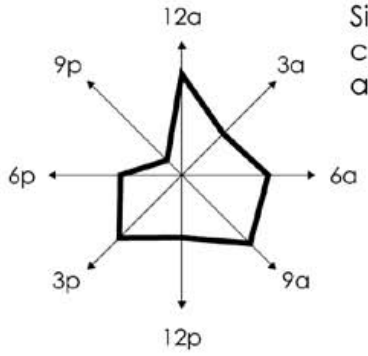
Shows distinct points but might need line to show trend if not much data

Dot-bar graph



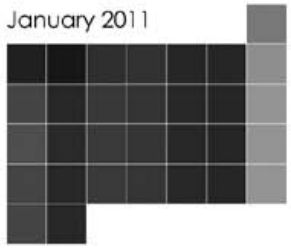
Focuses more on endpoints than a bar graph

Radial plot



Similar to line chart but wraps around

Calendar

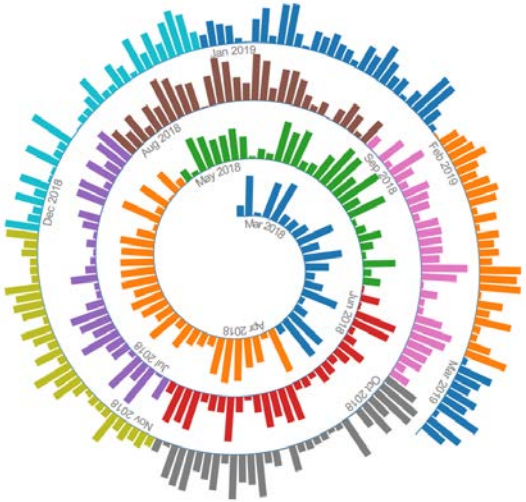


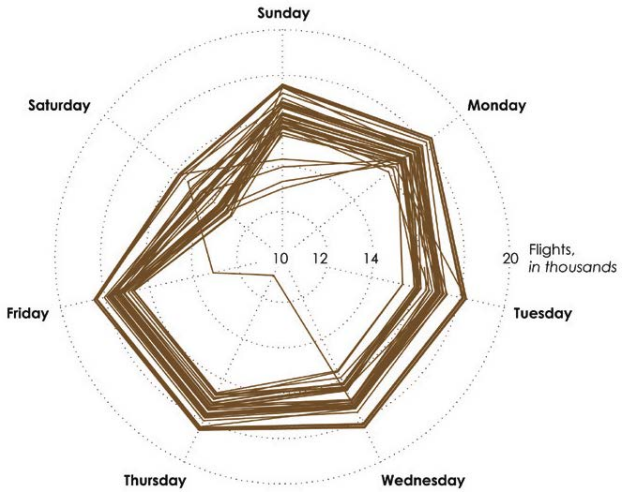
Patterns for days of week seen more easily than views above

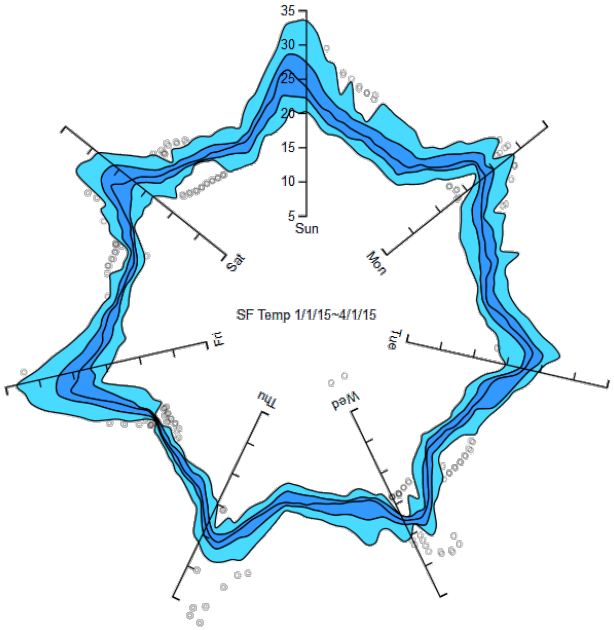
Más ideas:
https://uxmag.com/articles/its-about-time

Series Temporales

Se pueden emplear colores por periodo (por ejemplo, por año) y representar una serie larga en espiral (izquierda). Si la serie es periódica, el patrón se refuerza con un gráfico radial (izquierda). Incluso se puede usar un boxplot radial (derecha) para resumir la distribución de probabilidad





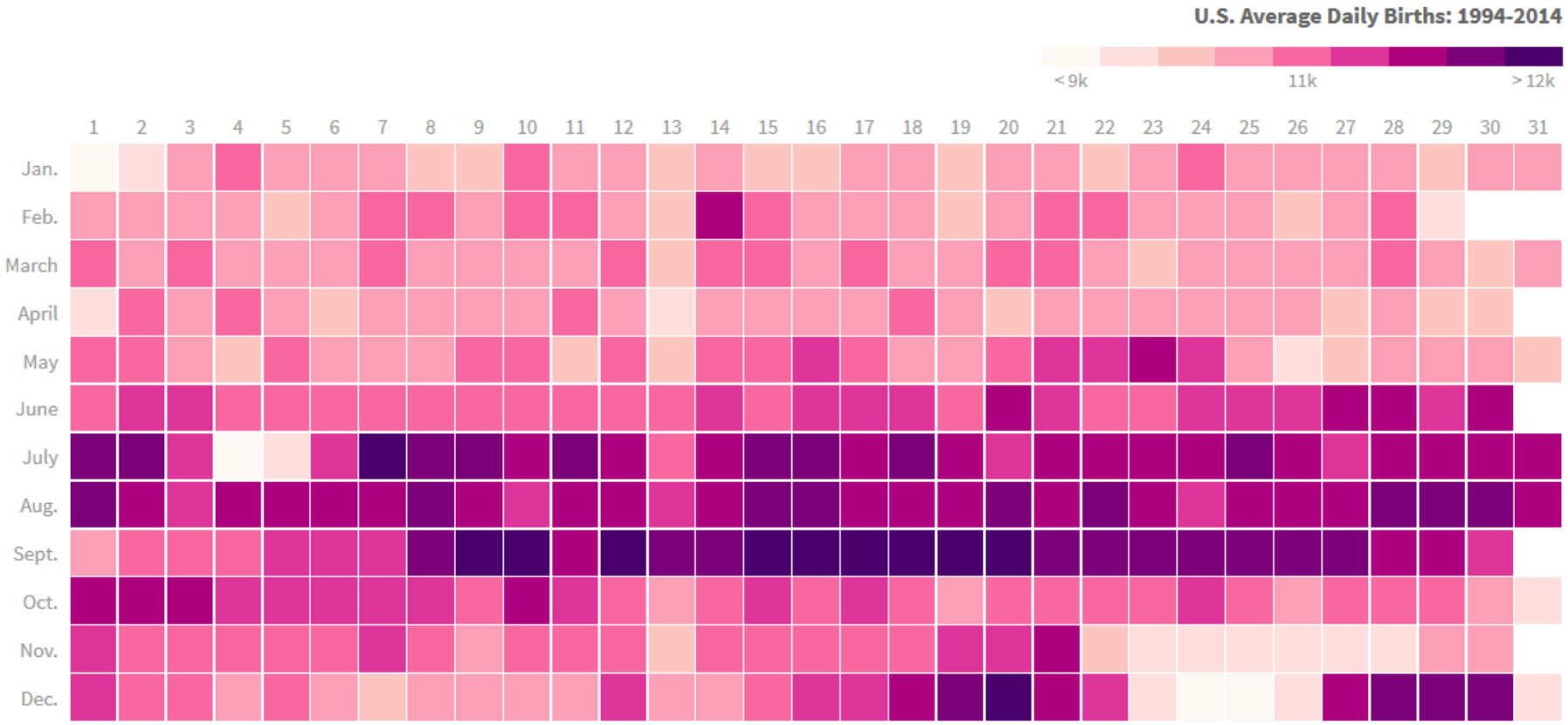


Spiral plot:
https://bl.ocks.org/arpitnarechania/027e163073864ef2ac4ceb5c2c0bf616

Radial boxplot:
http://bl.ocks.org/davidwclin/ad5d13db260caeffe9b3

Series Temporales

- Frecuencia de nacimientos por día en EE.UU.



http://thedailyviz.com/2016/09/17/how-common-is-your-birthday-dailyviz/

Calendar view: http://bl.ocks.org/mbostock/4063318

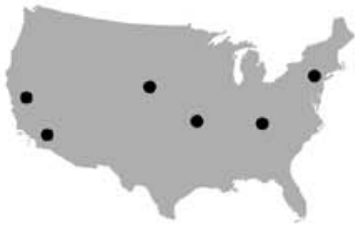
Datos Espaciales

Datos Espaciales

Locations

A direct translation of latitude and longitude to two-dimensional space is straightforward and intuitive, but can pose challenges when there are a lot of locations.

Location map



Points represent locations and can be scaled by metric

Connections

Points can be connected to show relationships between locations

Regions

Oftentimes the the density of individual points across regions is more informative than points on a map that can overlap.

Choropleth map

Defined regions colored by data and meaning can change based on scale

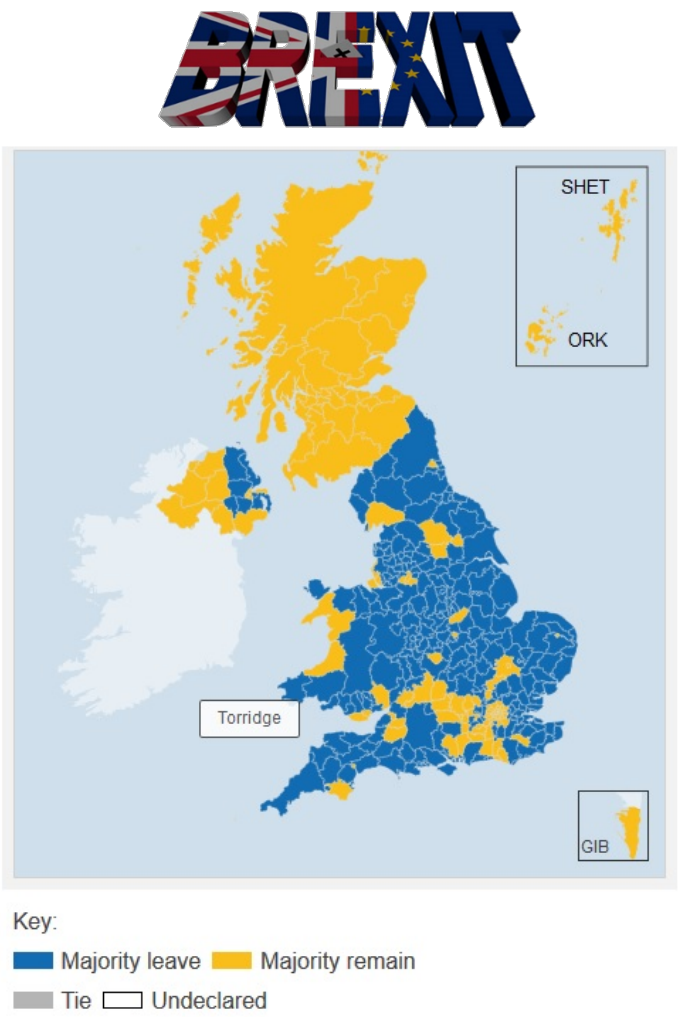
Countour map

Lines show data continuously over geography, using density

Datos Espaciales

http://bbvatourism.vizzuality.com/

Datos Espaciales



England

Leave **53.4%** 15,188,406 VOTES — Remain **46.6%** 13,266,996 VOTES

Counting complete — Turnout: 73.0%

Northern Ireland

Leave **44.2%** 349,442 VOTES — Remain **55.8%** 440,437 VOTES

Counting complete — Turnout: 62.9%

Scotland

Leave **38.0%** 1,018,322 VOTES — Remain **62.0%** 1,661,191 VOTES

Counting complete — Turnout: 67.2%

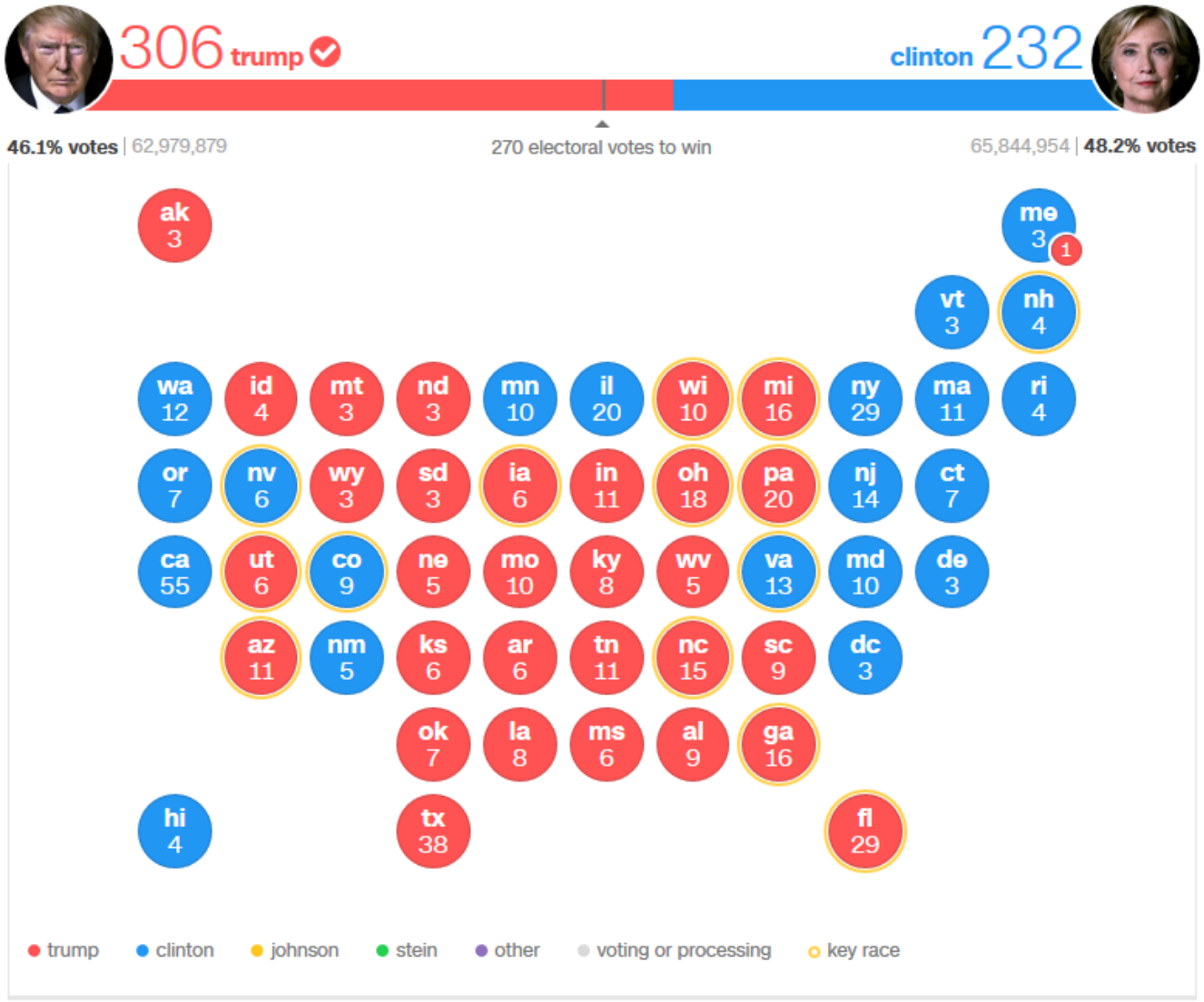
Wales

Leave **52.5%** 854,572 VOTES — Remain **47.5%** 772,347 VOTES

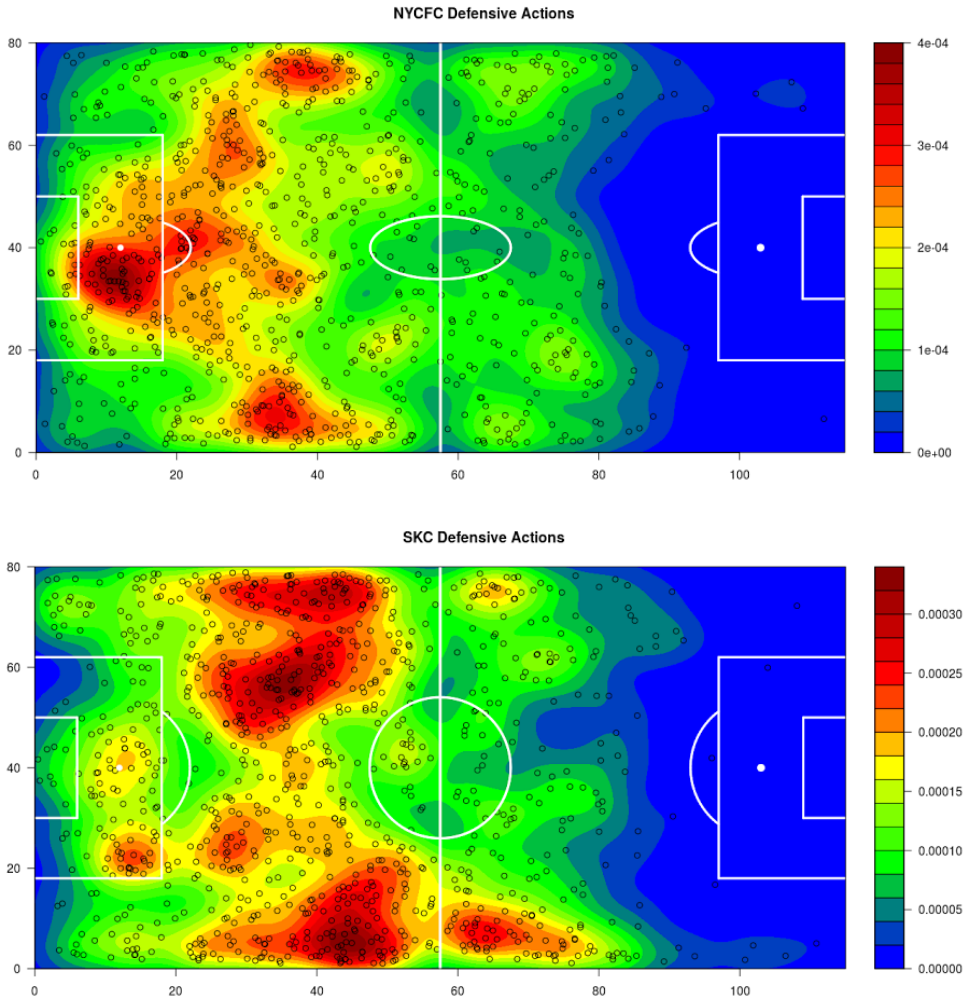
Counting complete — Turnout: 71.7%

Choropleth: http://bl.ocks.org/fhernand/be1e9c9fdb0473292abf

Datos Espaciales



Datos Espaciales

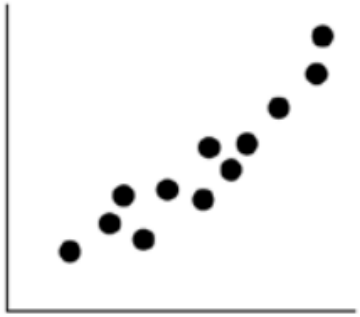


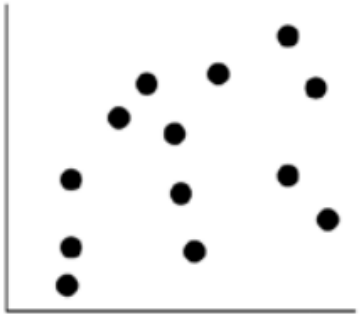
http://www.americansocceranalysis.com/home/2015/8/17/shot-limiting-bringing-the-heat-maps

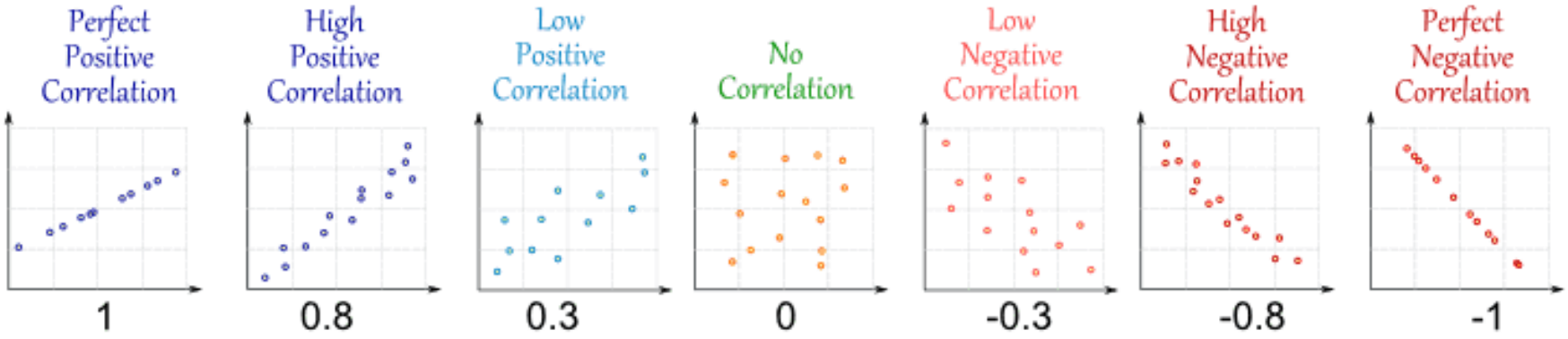
Múltiples Variables

Múltiples Variables

correlación fuerte ----------→ correlación débil

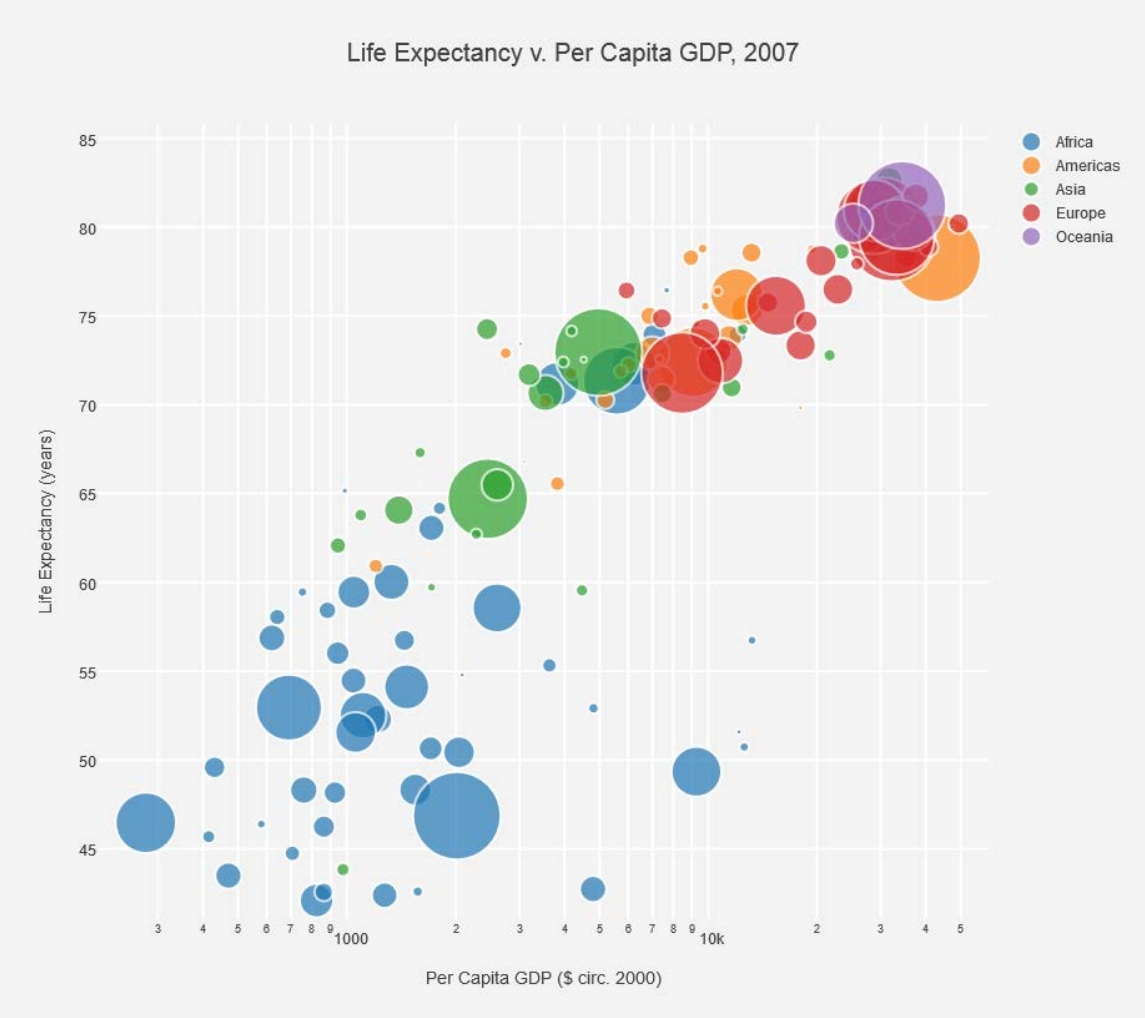






Múltiples Variables

Con un gráfico de burbujas podemos visualizar cuatro variables (x, y, tamaño del punto y color)



Bubble chart: http://help.plot.ly/make-a-bubble-chart/

Múltiples Variables

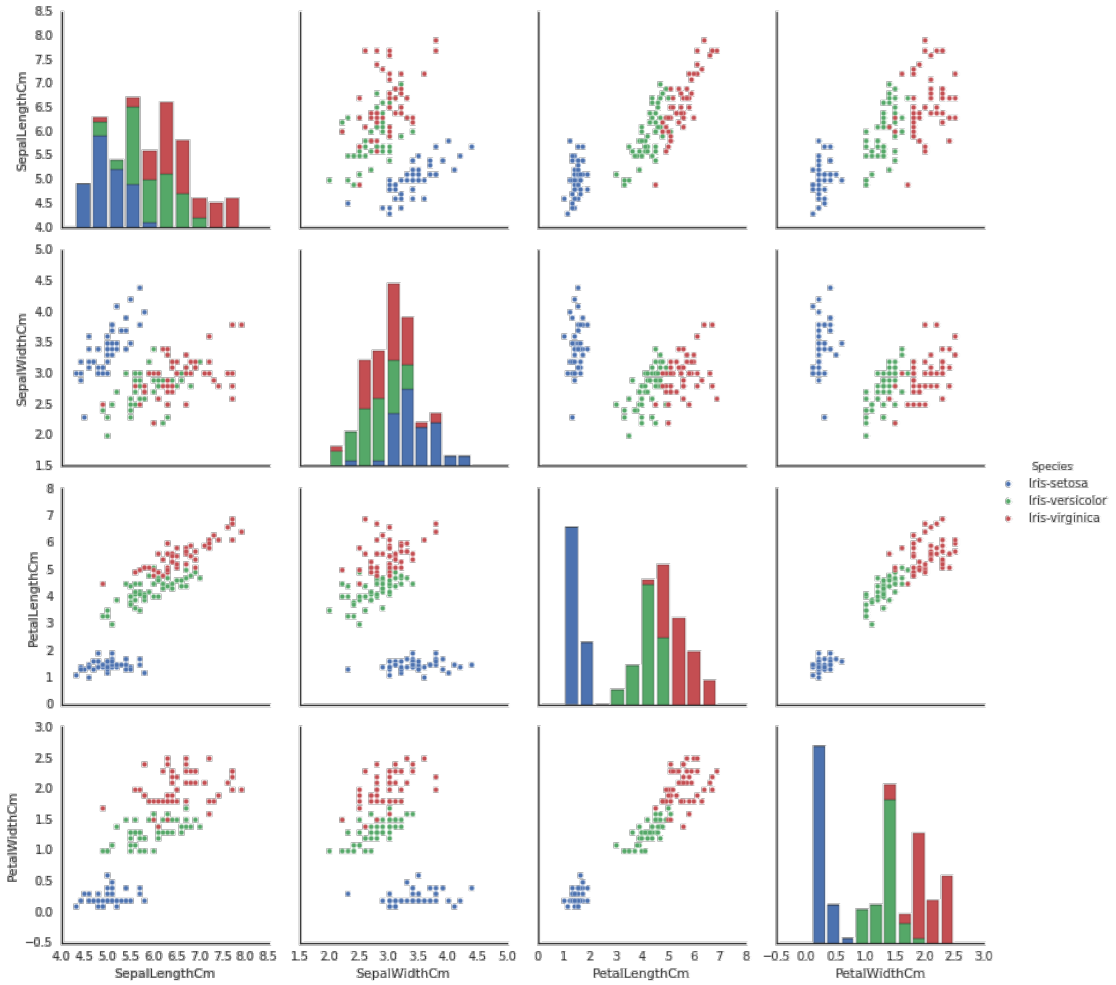
Se puede hacer dinámico para comprender la evolución de cada punto

https://www.gapminder.org/tools

Múltiples Variables

Con un scatter matrix podemos comprender de un vistazo la naturaleza de los datos y posibles correlaciones entre variables.

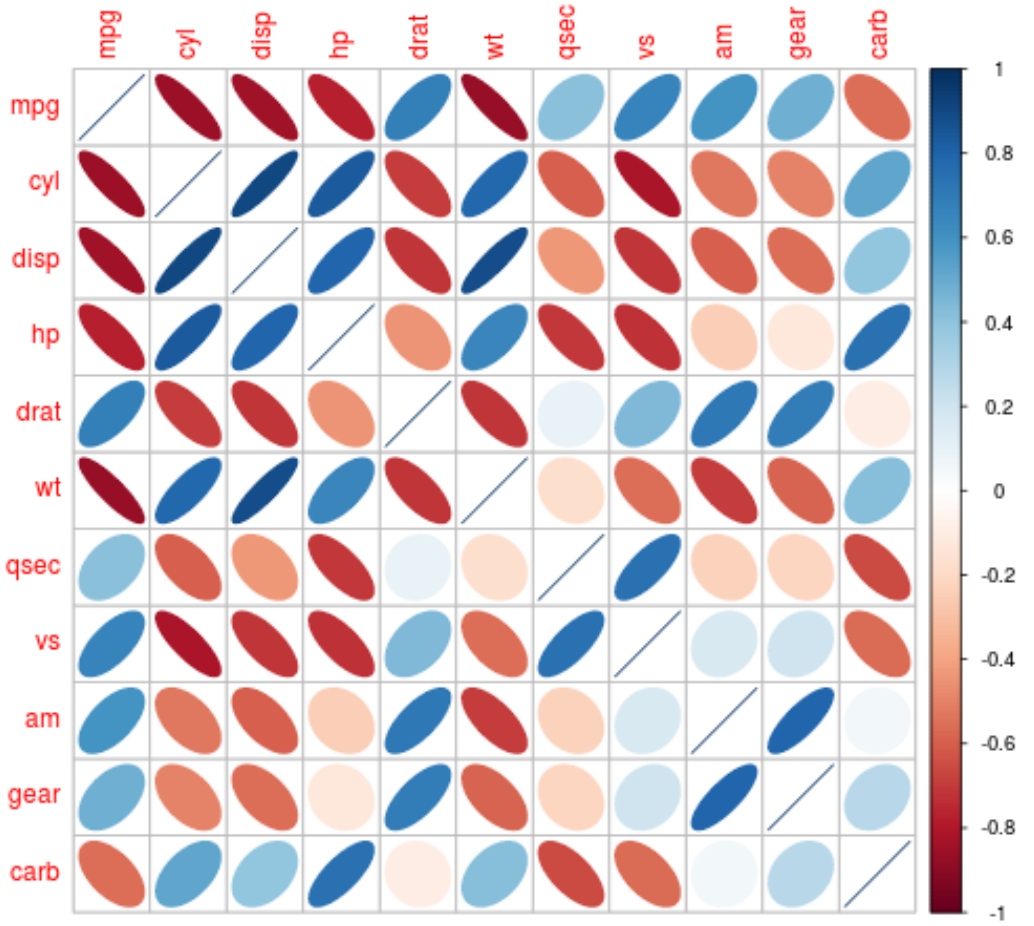
Aprovechamos la diagonal para mostrar la distribución de cada variable



Scatter matrix: http://benjiec.github.io/scatter-matrix/demo/demo.html

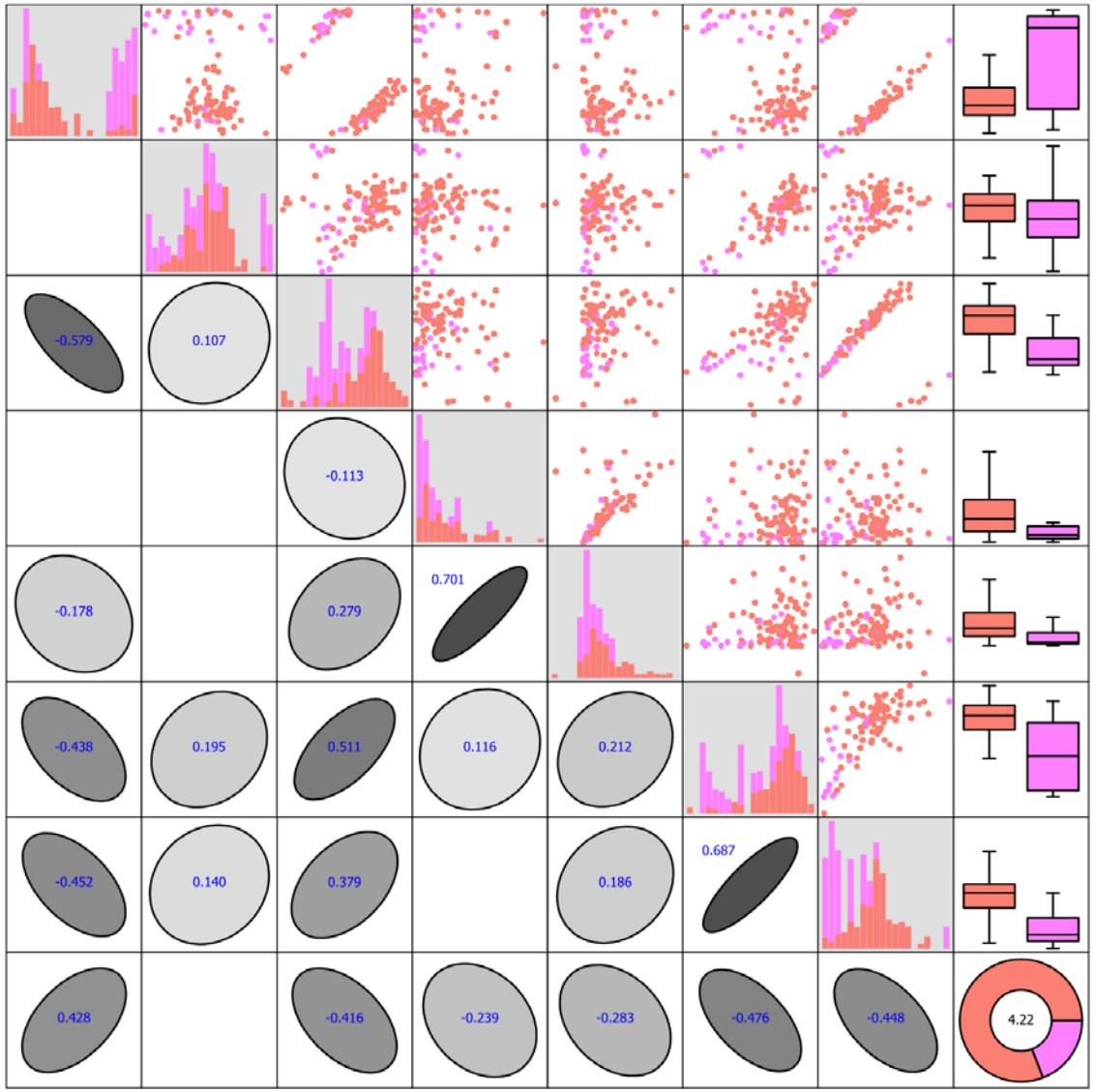
Múltiples Variables

Cuando son muchas variables podemos representar, en su lugar, el coeficiente de correlación de Pearson (0 si no hay correlación, -1 si hay correlación lineal negativa perfecta o 1 para correlación positiva) mediante elipses y colores



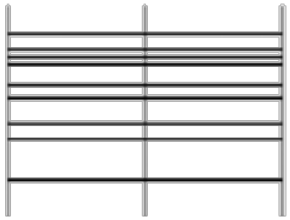
Múltiples Variables

Y, por qué no,
¡todo en uno!

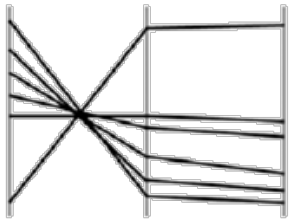


Múltiples Variables

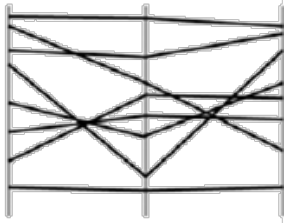
Positive correlation
Lines run parallel

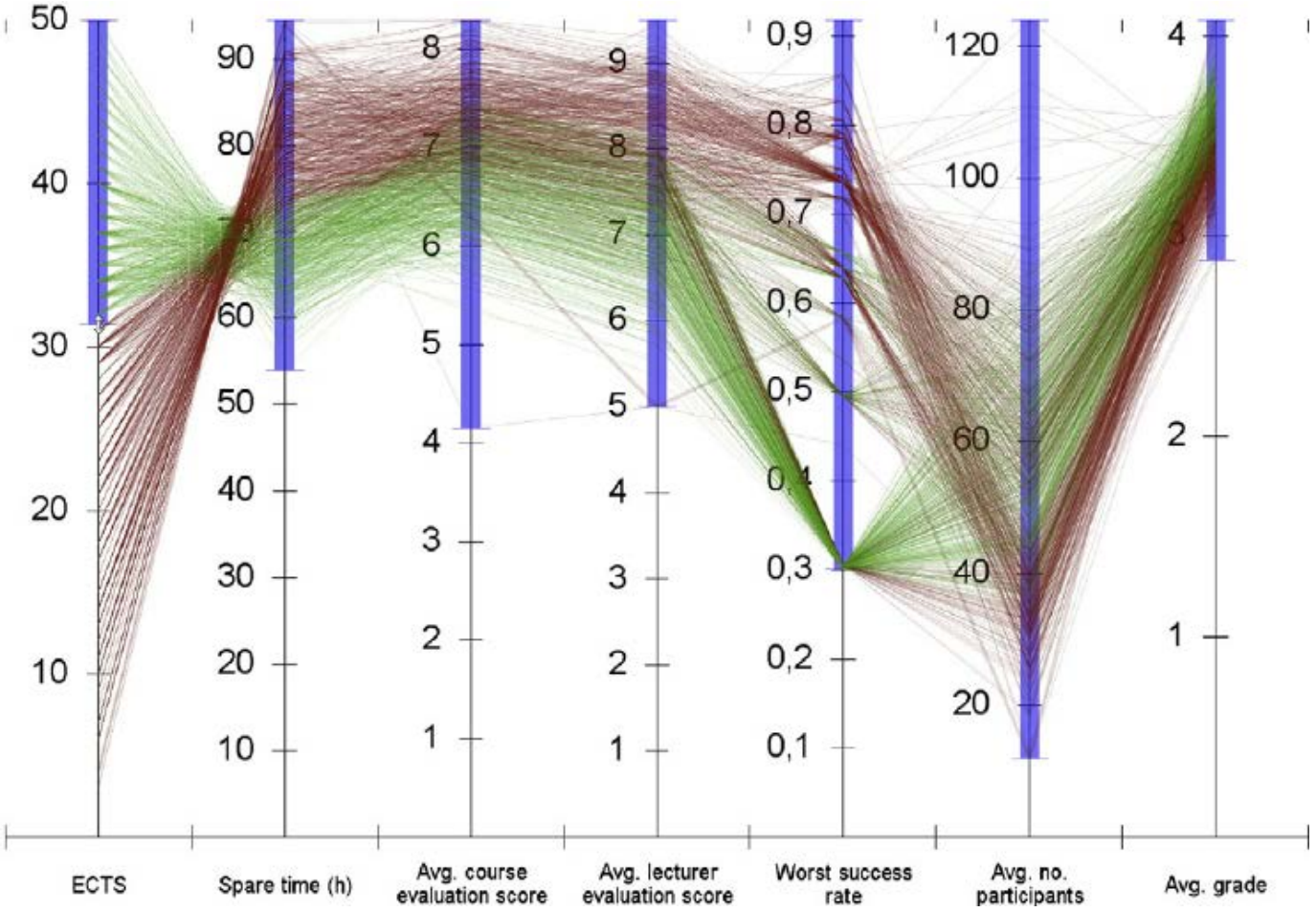


Negative correlation
Lines cross consistently



Weak correlation
No clear direction

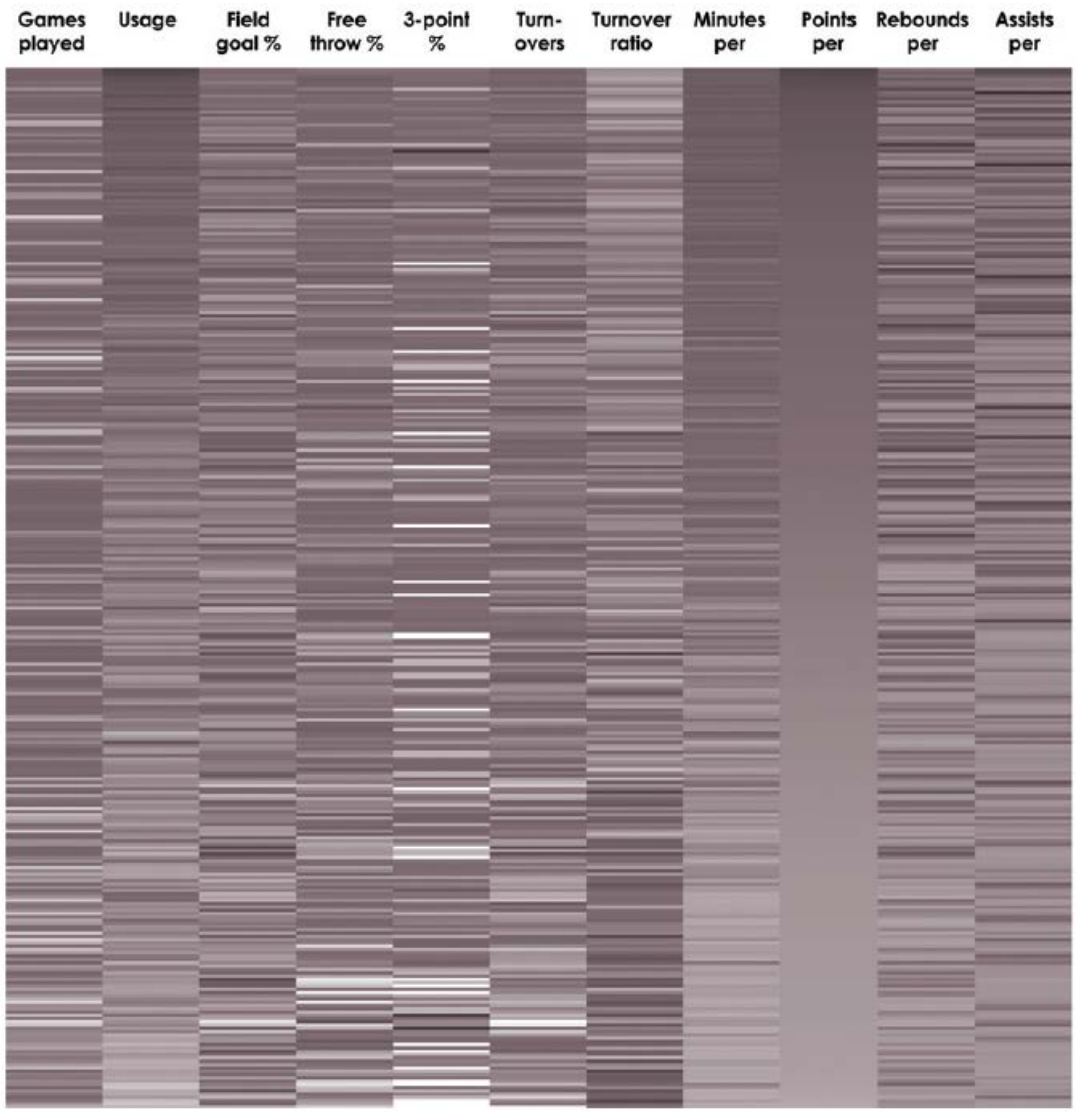




Parallel coordinates: http://bl.ocks.org/mbostock/7586334

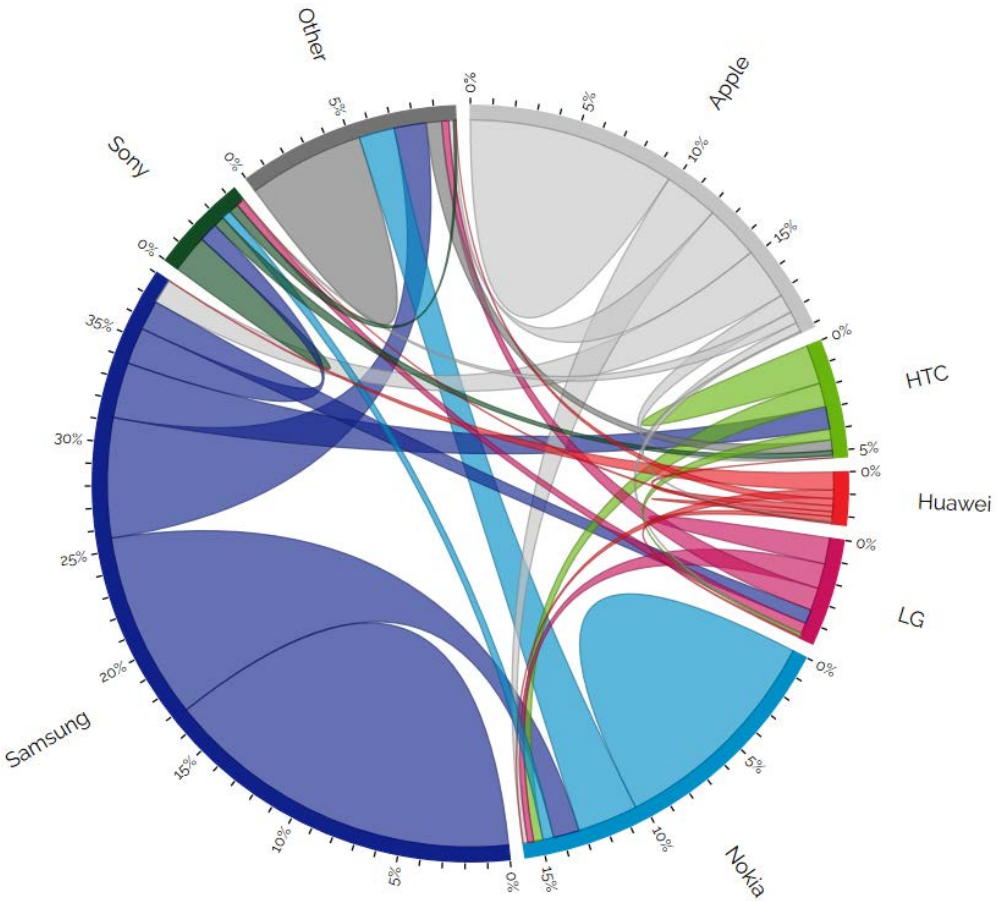
Múltiples Variables

En gráficos *heat map* es recomendable ordenar una columna (en el ejemplo, "*Points per*") para ver mejor las correlaciones y tendencias



Múltiples Variables

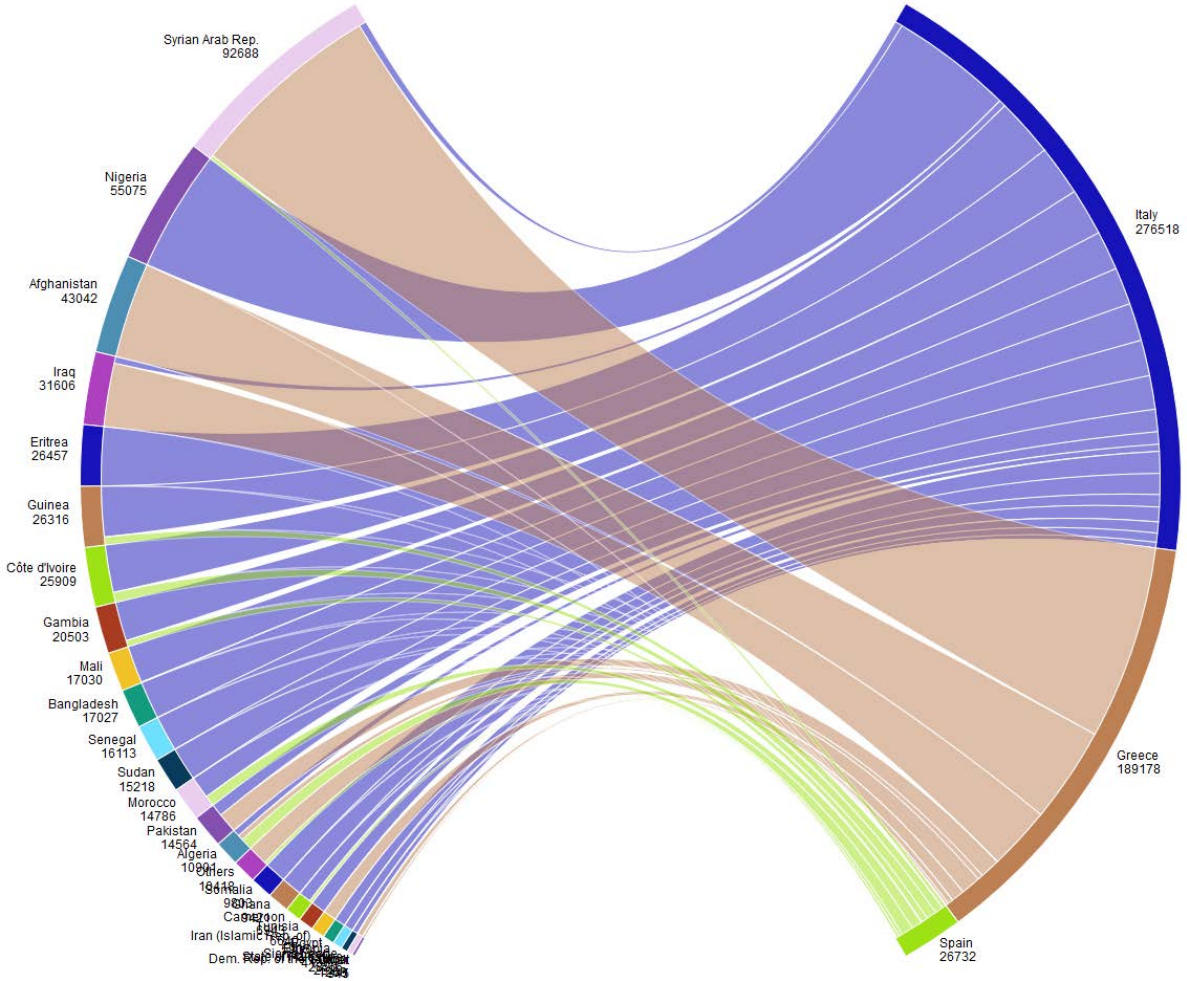
La rueda de dependencia o *chord* ayuda a representar cómo se relacionan las variables por pares



http://www.visualcinnamon.com/2014/12/using-data-storytelling-with-chord.html
http://www.redotheweb.com/DependencyWheel/
Dependency wheel: http://circos.ca/

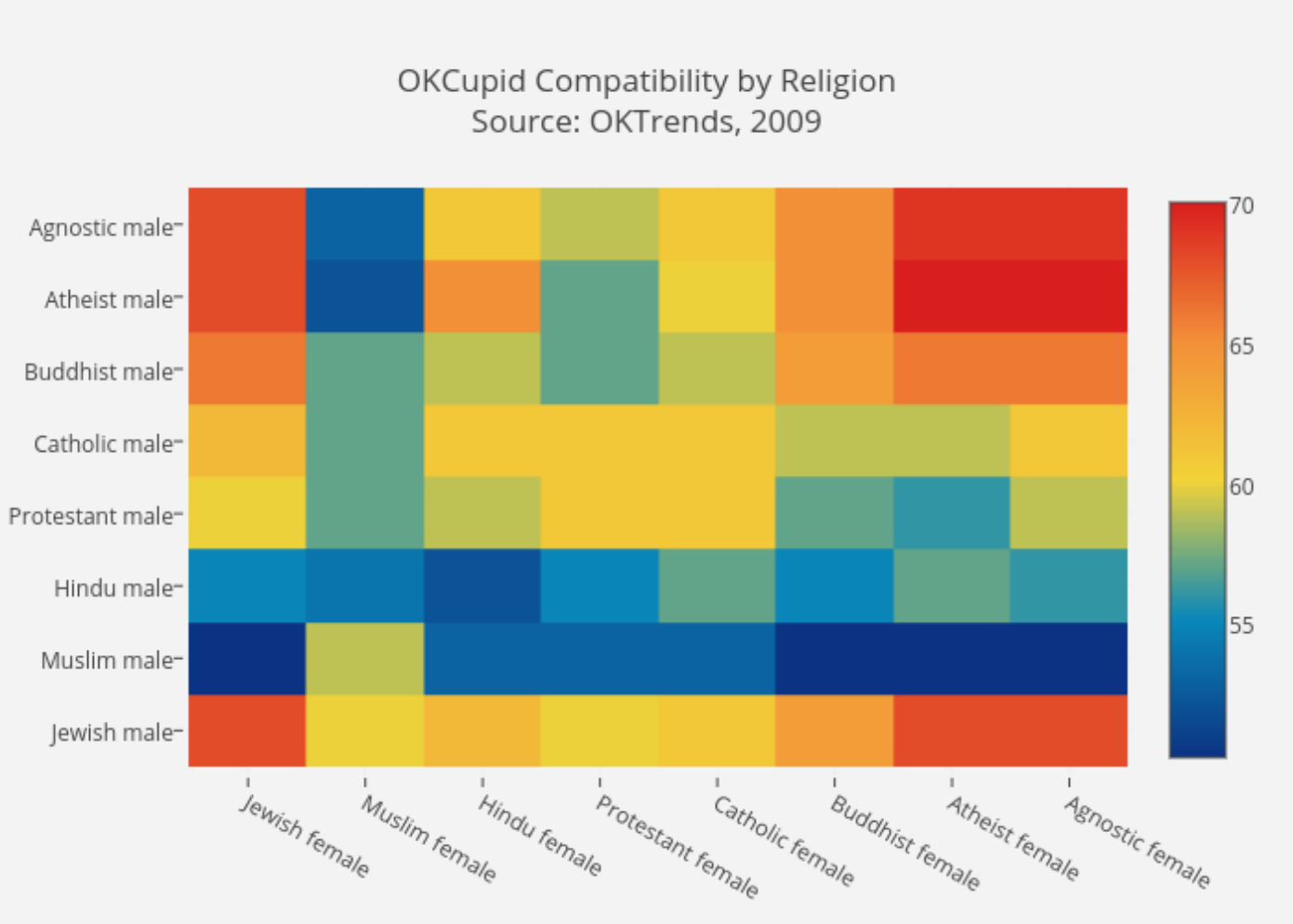
Múltiples Variables

Un híbrido entre Sankey y Rueda de Dependencia



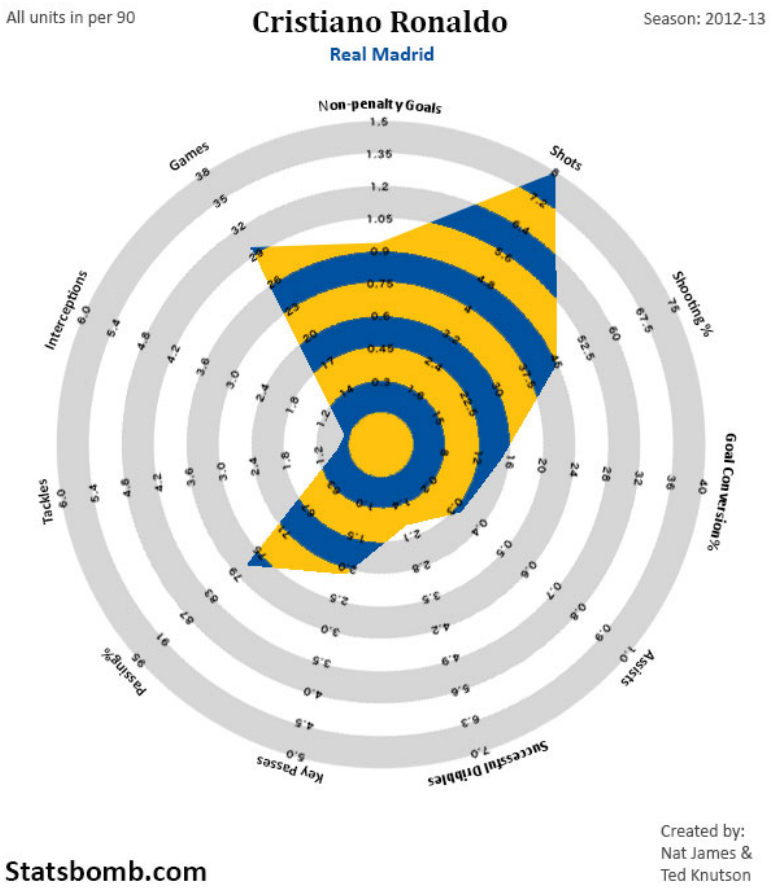
https://plotdb.com/chart/1039/

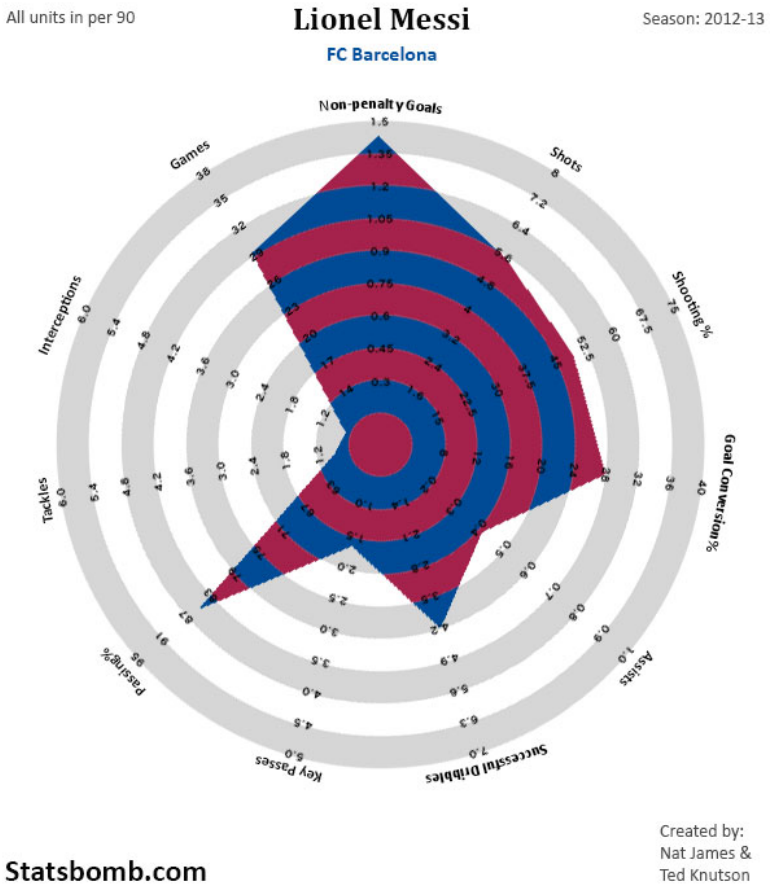
Múltiples Variables



Heatmap: http://help.plot.ly/make-a-heatmap/

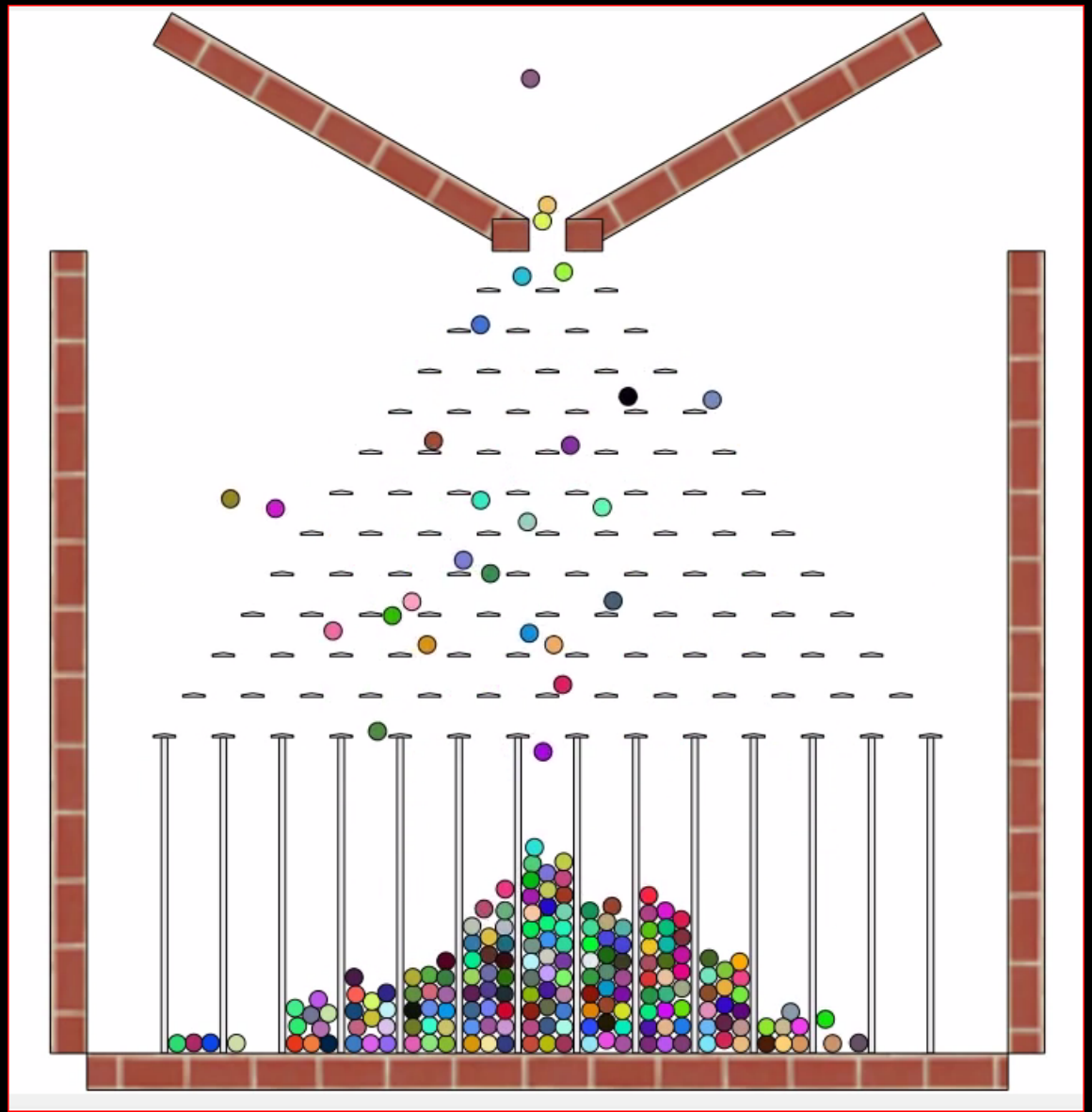
Múltiples Variables





Radar (spider) chart: http://bl.ocks.org/nbremer/6506614

Distribuciones

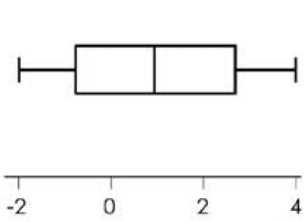


Distribuciones

Distribution Summary

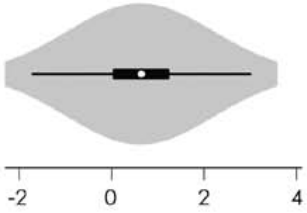
You can visualize data at different granularities with the charts above. These show key values for a less specific view of distributions.

Box plot

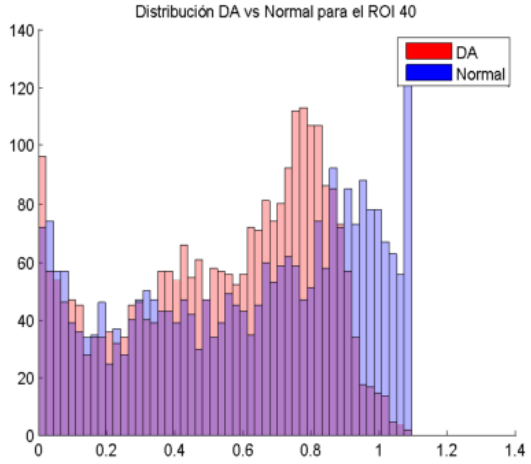


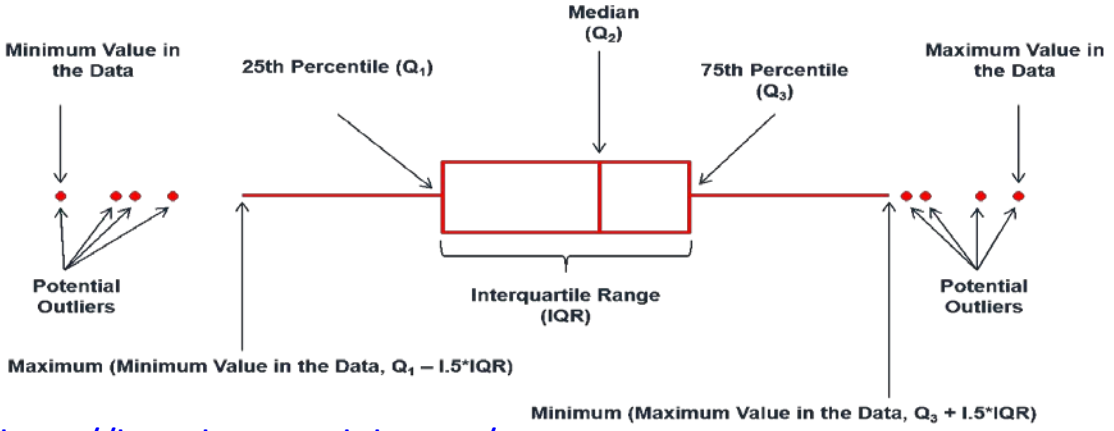
Shows range, median and quartiles

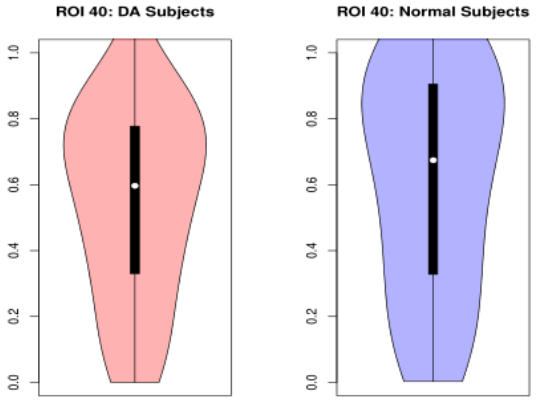
Violin plot



A combination of a box plot and density plot







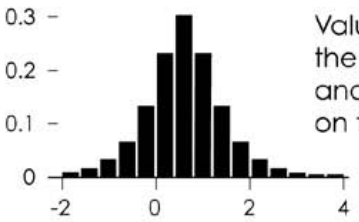
http://boxplot.tyerslab.com/

Distribuciones

Distribution of one variable

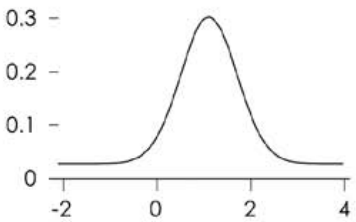
You can see where data is clustered and see any outliers by keeping track of where they sit on a value axis.

Histogram

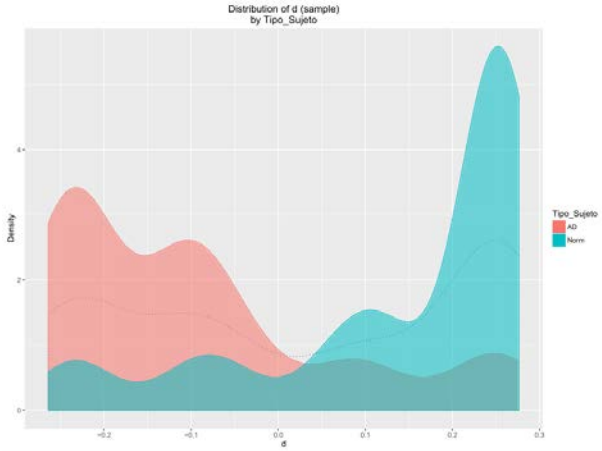


Value axis on the horizontal and density on the vertical

Density plot



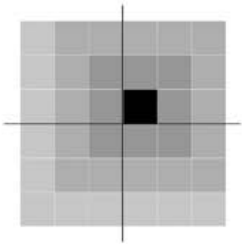
Like histo-gram but continuous instead of bins



Distribution of multiple variables

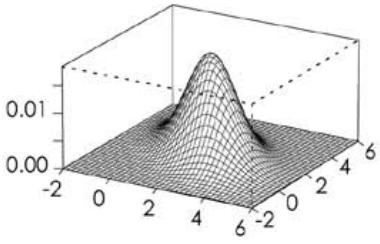
Sometimes values come as pairs, and it makes sense to show both values at the same time.

Heat map



Density on a a 2-D plane, using color as visual cue

Surface plot

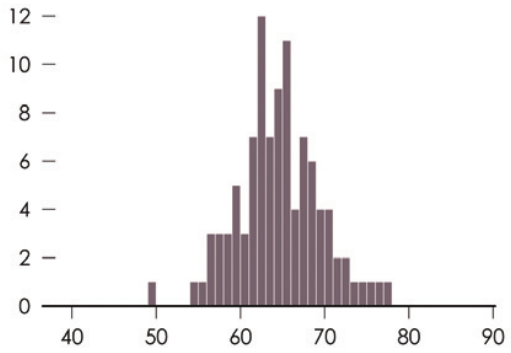


Shows same patterns as heat map, but uses height instead of color

Distribuciones

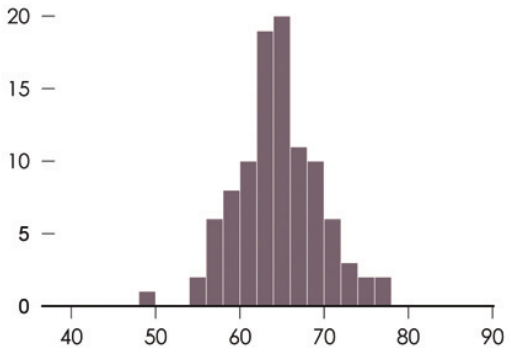
One-inch bins

Small bins shows variations at higher granularity.



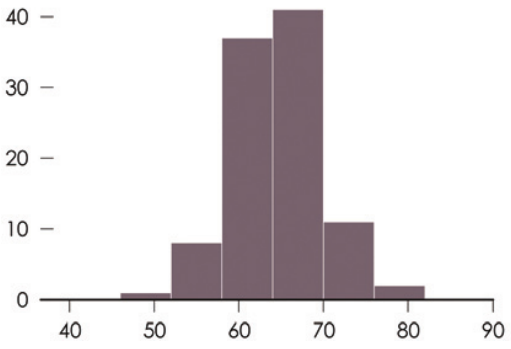
Two-inch bins

You see less variation, but the distribution around the median is more obvious.



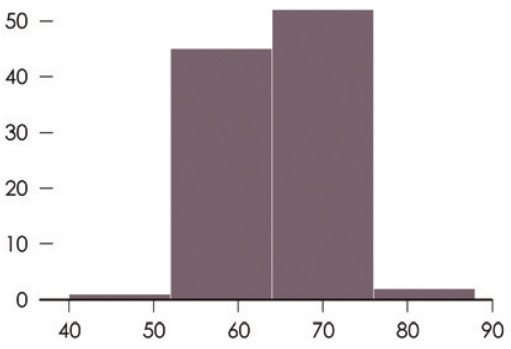
Half-foot bins

You can see distribution around the median, but you can only see some variation.



One-foot bins

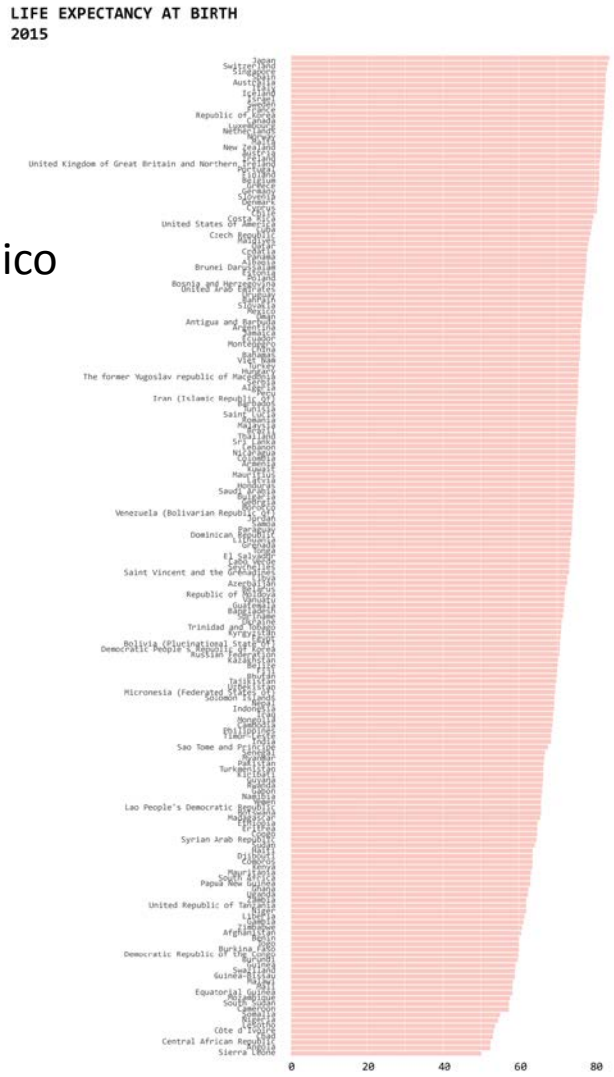
The spread of the data isn't as obvious. because the larger bins show less detail.



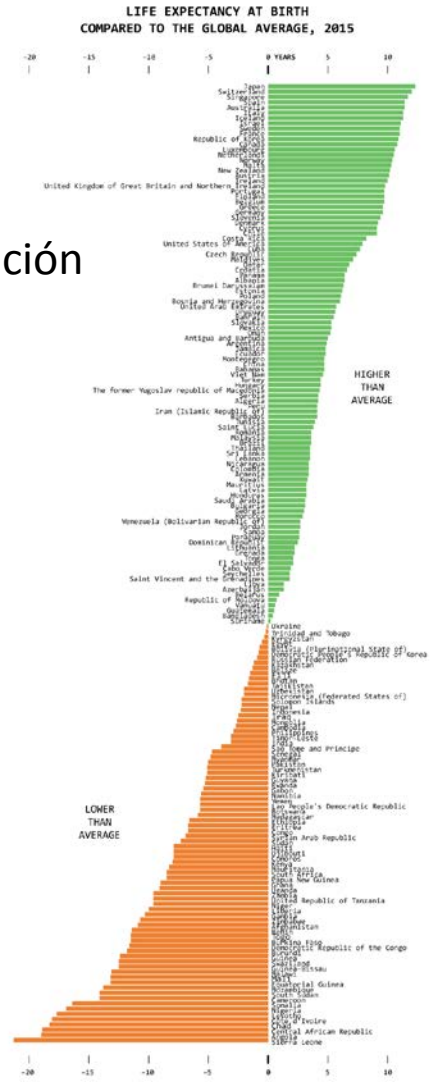

SUMMARY

Todo se puede ver de diversas formas: Piensa cómo quieres contarlo

Un clásico



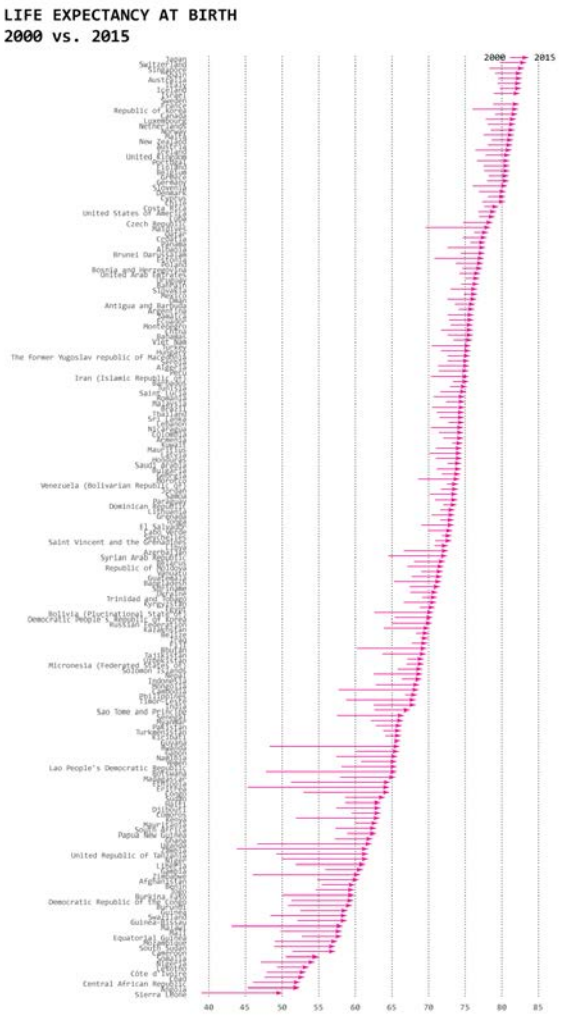
Comparación relativa



https://flowingdata.com/2017/01/24/one-dataset-visualized-25-ways/

Todo se puede ver de diversas formas: Piensa cómo quieres contarlo

O bien haciendo énfasis en el progreso en los últimos años

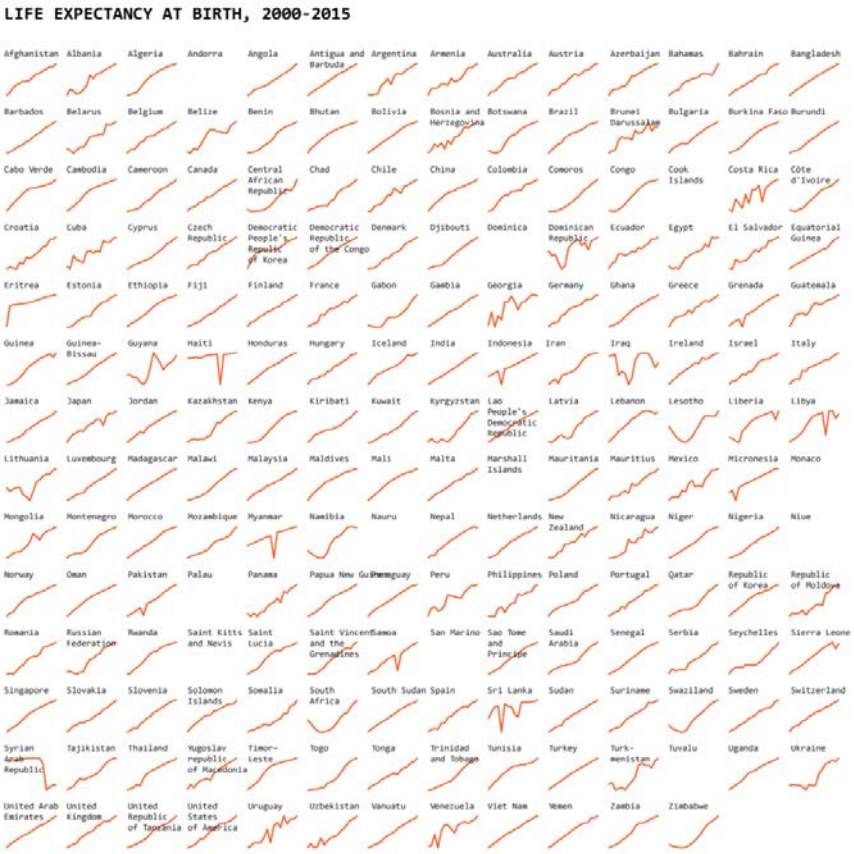


También animado

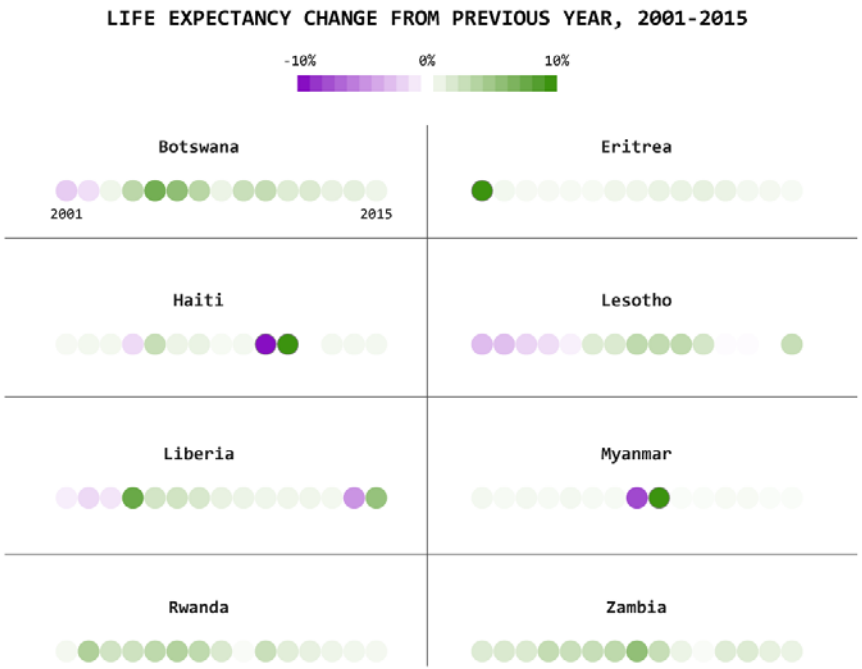
https://flowingdata.com/2017/01/24/one-dataset-visualized-25-ways/

Todo se puede ver de diversas formas: Piensa cómo quieres contarlo

Distribución global según año



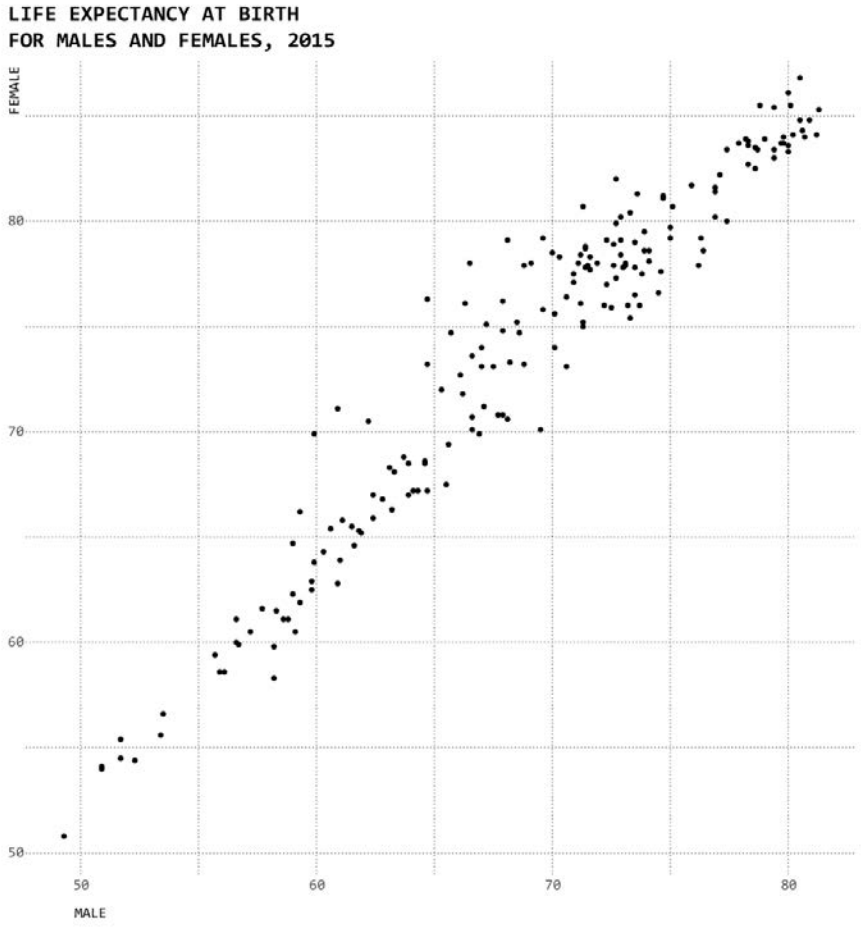
Énfasis en países con grandes fluctuaciones

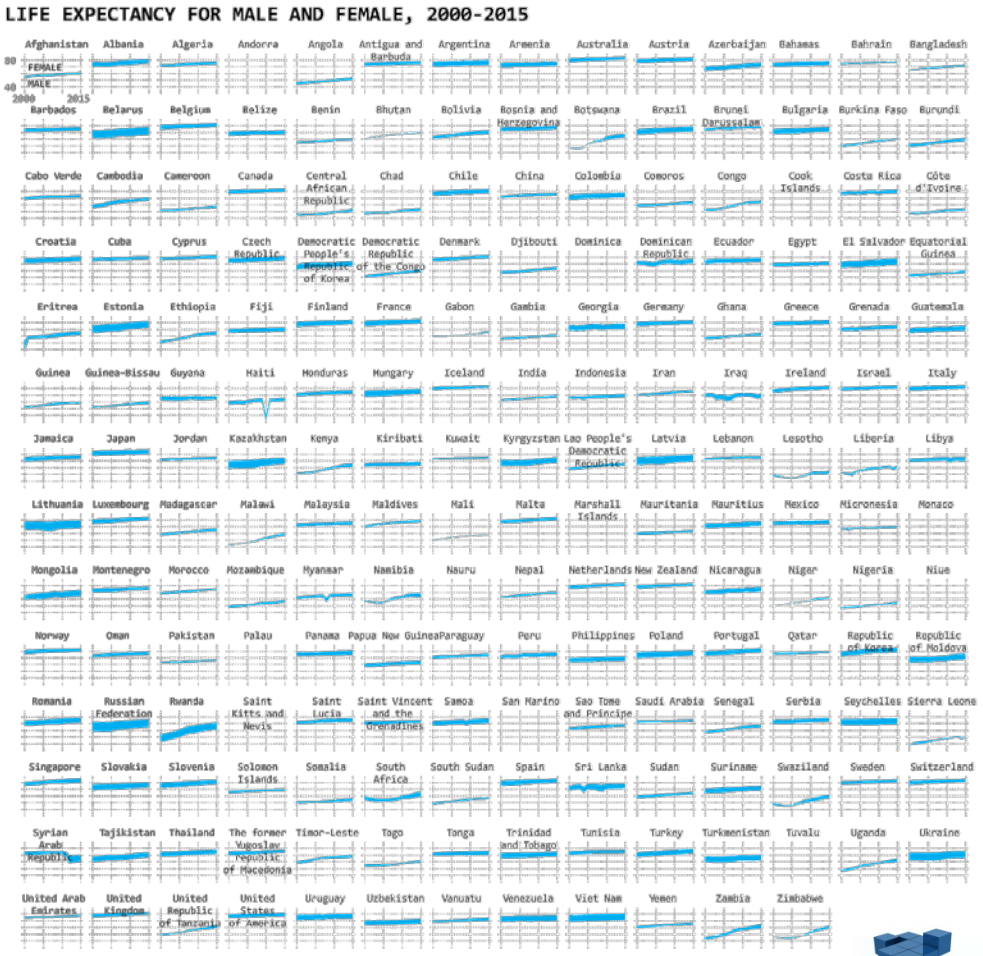


https://flowingdata.com/2017/01/24/one-dataset-visualized-25-ways/

Todo se puede ver de diversas formas: Piensa cómo quieres contarlo

Según género

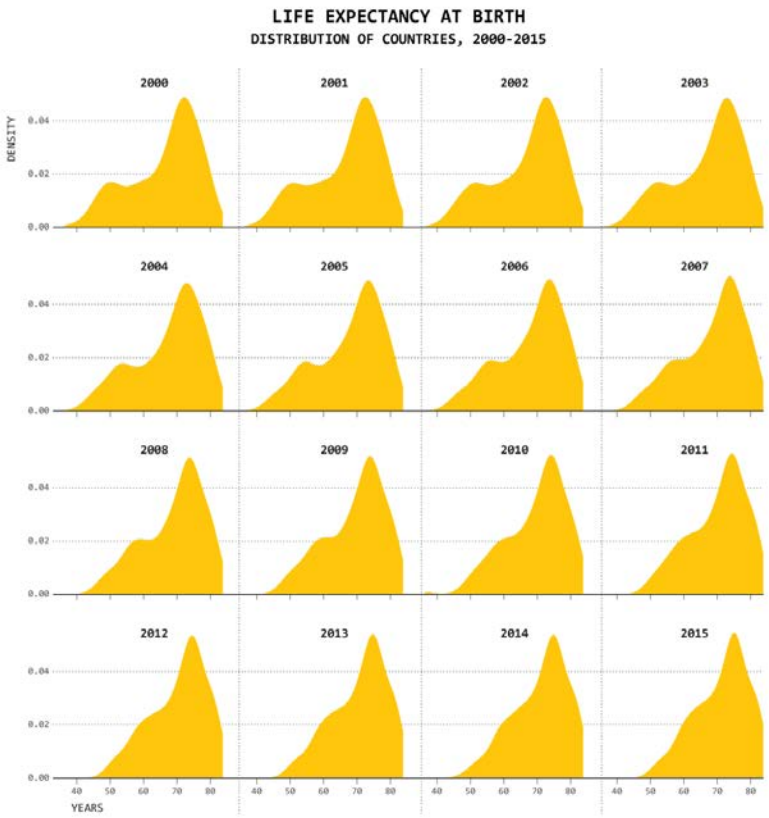




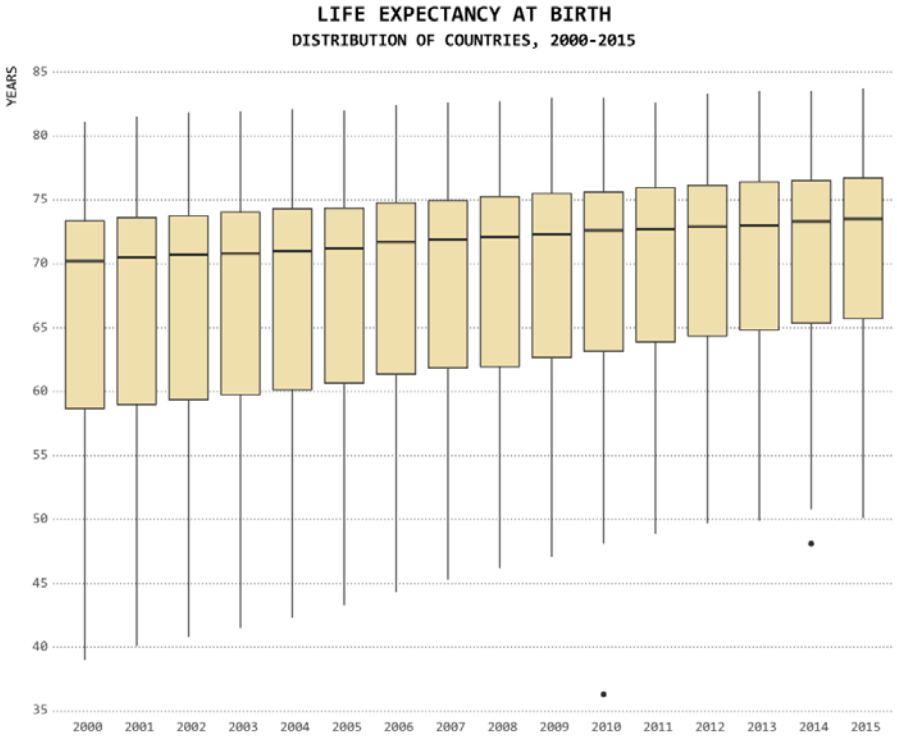
https://flowingdata.com/2017/01/24/one-dataset-visualized-25-ways/

Todo se puede ver de diversas formas: Piensa cómo quieres contarlo

Distribución global según año



Distribución pero más informativo



https://flowingdata.com/2017/01/24/one-dataset-visualized-25-ways/


SOFTWARE

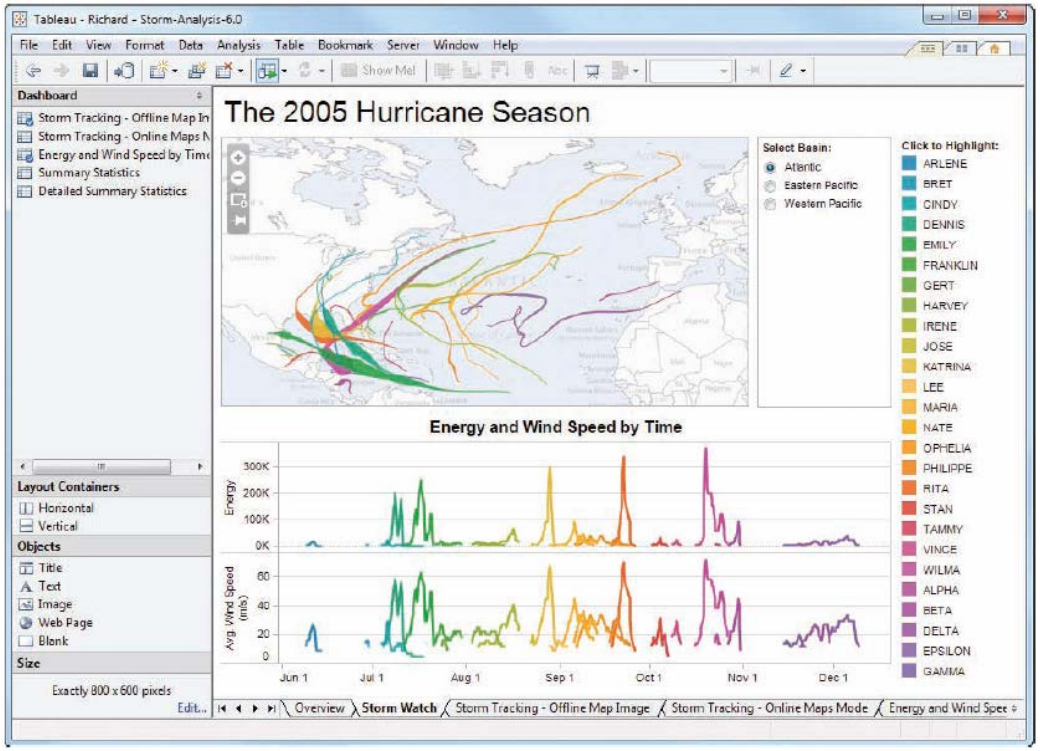
Software

- Herramientas de visualización: Microsoft Excel, Google Spreadsheets, Tableau Software, QlikView, Spotfire, EDEN, R, Python, JavaScript…
- Específicos:
  - Gephi: grafos y redes
  - ImagePlot: imágenes
  - TreeMap
  - TileMill, indiemapper, GeoCommons, ArcGIS: datos espaciales
- El software de visualización tiende a especializarse en determinados tipos de datos o características

Tableau Software

http://www.tableausoftware.com

- Orientado a Inteligencia de Negocio
- Muy completo
- De propósito general
- Fácil diseño, interactividad, cuidada estética…
- De pago



R

http://www.r-project.org/

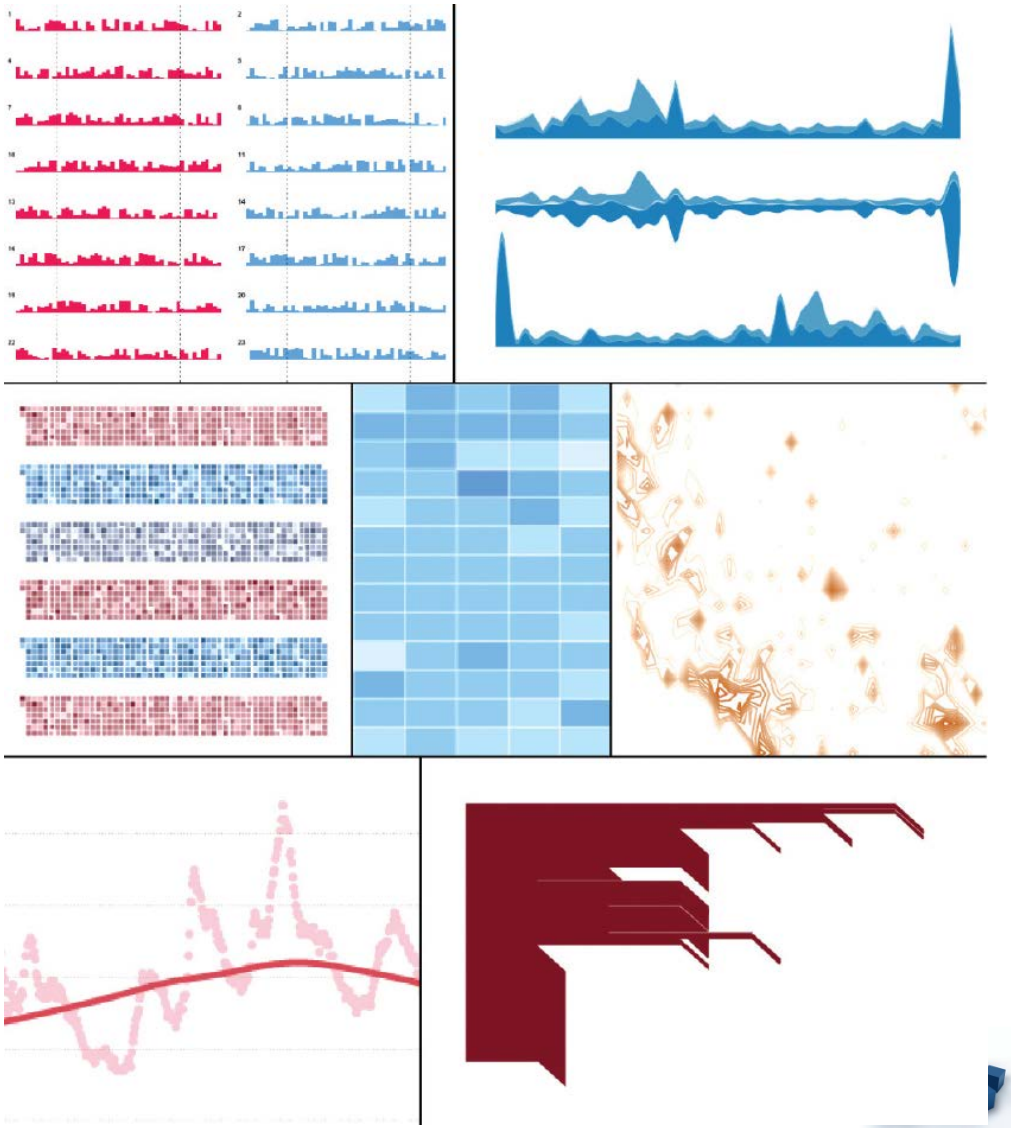
Python

https://matplotlib.org/

https://seaborn.pydata.org/

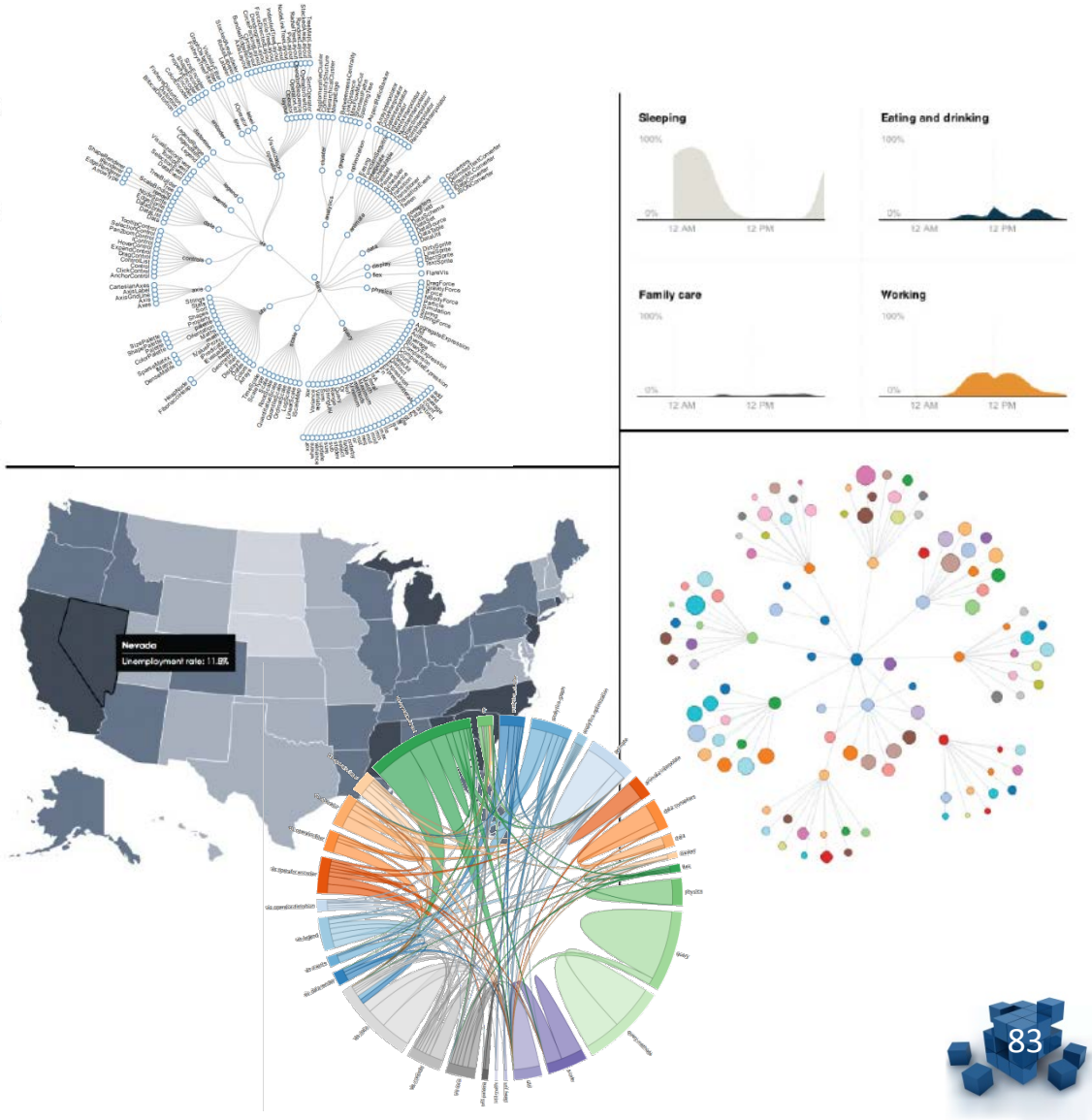
- Lenguajes de programación
- Código abierto
- Python es especialmente potente para integración de software y soluciones web

- Extensiones en R: ggplot2, network, ggmaps, animation, portfolio… https://github.com/fcharte/ExploraVisualizaconR
- Bibliotecas en Python: matplotlib, seaborn… https://blog.modeanalytics.com/python-data-visualization-libraries/



JavaScript

- Necesidad de programación
- Ideal para Web y gráficos interactivos
- Bibliotecas:
  - Data-Driven Documents (D3): http://d3js.org
  - Raphaël
  - JavaScript InfoVis Toolkit…
- Para empezar:
  https://my.infocaptor.com/free_data_visualization.php
- Se puede usar en KNIME:
  https://www.knime.org/blog/from-d3-example-to-interactive-knime-view-in-10-minutes



Fuentes de Datos y Software para Extraerlos

- Datos:
  - https://data.world
  - https://data.humdata.org/
  - http://www.kdnuggets.com/datasets/index.html
  - http://blog.bigml.com/2013/02/28/data-data-data-thousands-of-public-data-sources/
  - http://blog.visual.ly/data-sources/
  - http://www.datawrangling.org/some-datasets-available-on-the-web/
  - https://github.com/caesar0301/awesome-public-datasets
  - http://mldata.org/repository
  - http://homepage.data-planet.com/
  - http://getthedata.org
  - http://census.okfn.org
  - http://datamarket.azure.com/browse/data
  - http://aws.amazon.com/es/datasets/
  - https://www.quandl.com
  - https://datamarket.com/data/list/?q=all
  - http://kaggle.com
  - https://open-data.europa.eu/es
  - http://datos.gob.es
  - http://www.red.es/redes/datosabiertos/
  - http://granadaendatos.granadaimedia.com/
  - http://transparente.ugr.es
- Software:
  - https://scraperwiki.com/
  - http://sourceforge.net/projects/gimagereader/
  - http://schoolofdata.org/

Referencias

https://datavizcatalogue.com/

- Ben Fry, Visualizing Data. Exploring and Explaining Data with Processing Environment, O'Reilly, 2007. ISBN: 978-0-596-51455-6
- Nathan Yau, Visualize This: The FlowingData Guide to Design, Visualization, and Statistics, Wiley, 2011. ISBN: 978-0470944882. http://flowingdata.com/
- Nathan Yau, Data Points: Visualization that Means Something, Wiley, 2013. ISBN: 978-1118462195
- Marjorie M.K. Hlava, Visualization for Data Analysis: A New Way to Look at Content (Julio 2012) http://youtu.be/cjz8I-eWAW8
- The beauty of data visualization - David McCandless: https://www.youtube.com/watch?v=5Zg-C8AAIGg
- http://www.informationisbeautiful.net
- http://www.r2d3.us/visual-intro-to-machine-learning-part-1/
- http://gothamknowledge.tumblr.com/post/66252844859/circle-of-life-the-beautiful-new-way-to-visualize
- http://www.smartdatacollective.com/jagadish/410273/why-we-love-these-5-data-visualization-tools-and-you-should-too
- https://public.tableau.com/s/gallery
- https://data2.unhcr.org/en/dataviz
- https://plotdb.com/
- https://www.gapminder.org/tools