Visualización de datos

Tema 2: Básicos

Paulo Félix Lamas
Área de Ciencias da Computación e Intelixencia Artificial

Departamento de Electrónica e Computación

Una frase

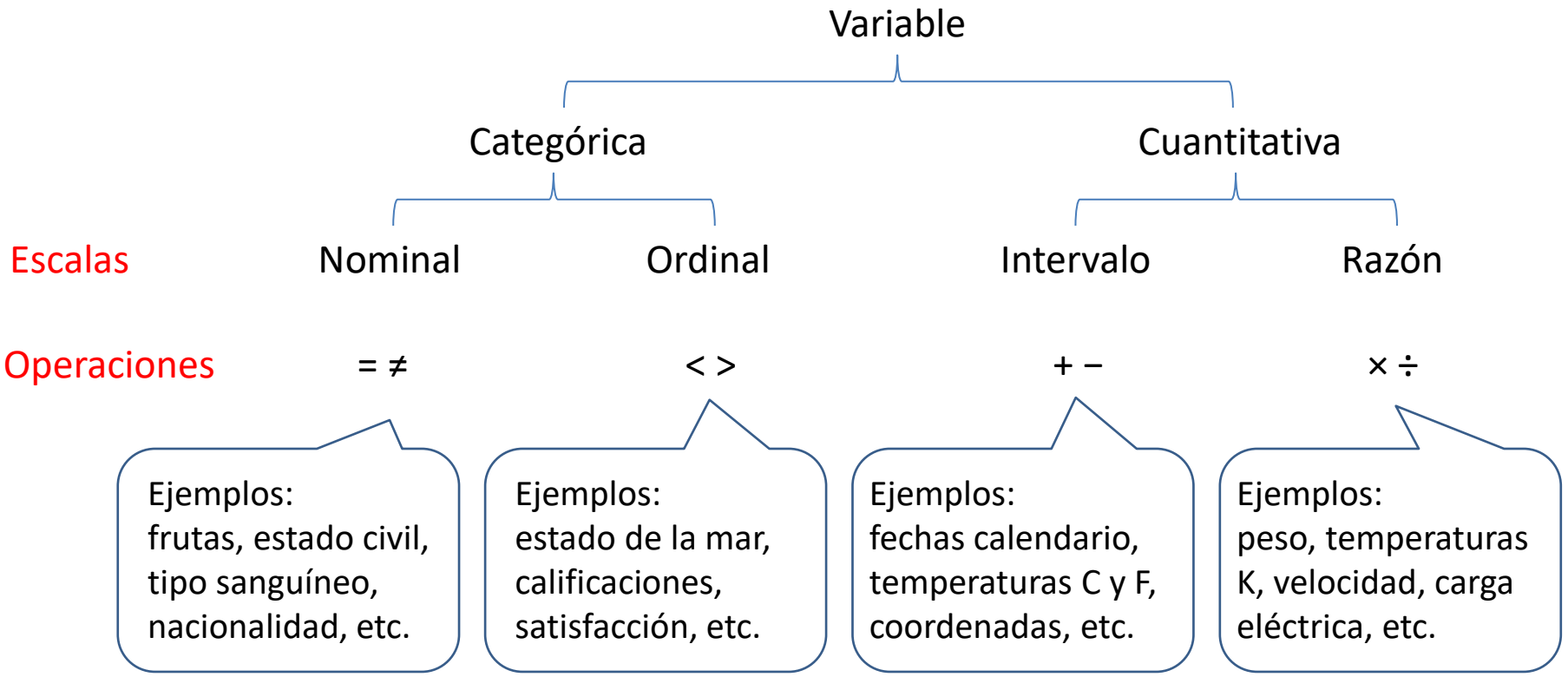
“The purpose of visualization is insight, not pictures.”
Ben Shneiderman

Tipos de medidas (Stevens, 1946)

TABLE 1

Scale	Basic Empirical Operations	Mathematical Group Structure	Permissible Statistics (invariantive)
NOMINAL	Determination of equality	*Permutation group* $x' = f(x)$ $f(x)$ means any one-to-one substitution	Number of cases Mode Contingency correlation
ORDINAL	Determination of greater or less	*Isotonic group* $x' = f(x)$ $f(x)$ means any monotonic increasing function	Median Percentiles
INTERVAL	Determination of equality of intervals or differences	*General linear group* $x' = ax + b$	Mean Standard deviation Rank-order correlation Product-moment correlation
RATIO	Determination of equality of ratios	*Similarity group* $x' = ax$	Coefficient of variation

Tipos de medidas (Stevens, 1946)



Podemos realizar conversiones entre escalas; por ejemplo, podemos convertir un intervalo en un ordinal o en un nominal: 80 grados C → muy caliente → quemado.

Tipos de datos y elementos visuales

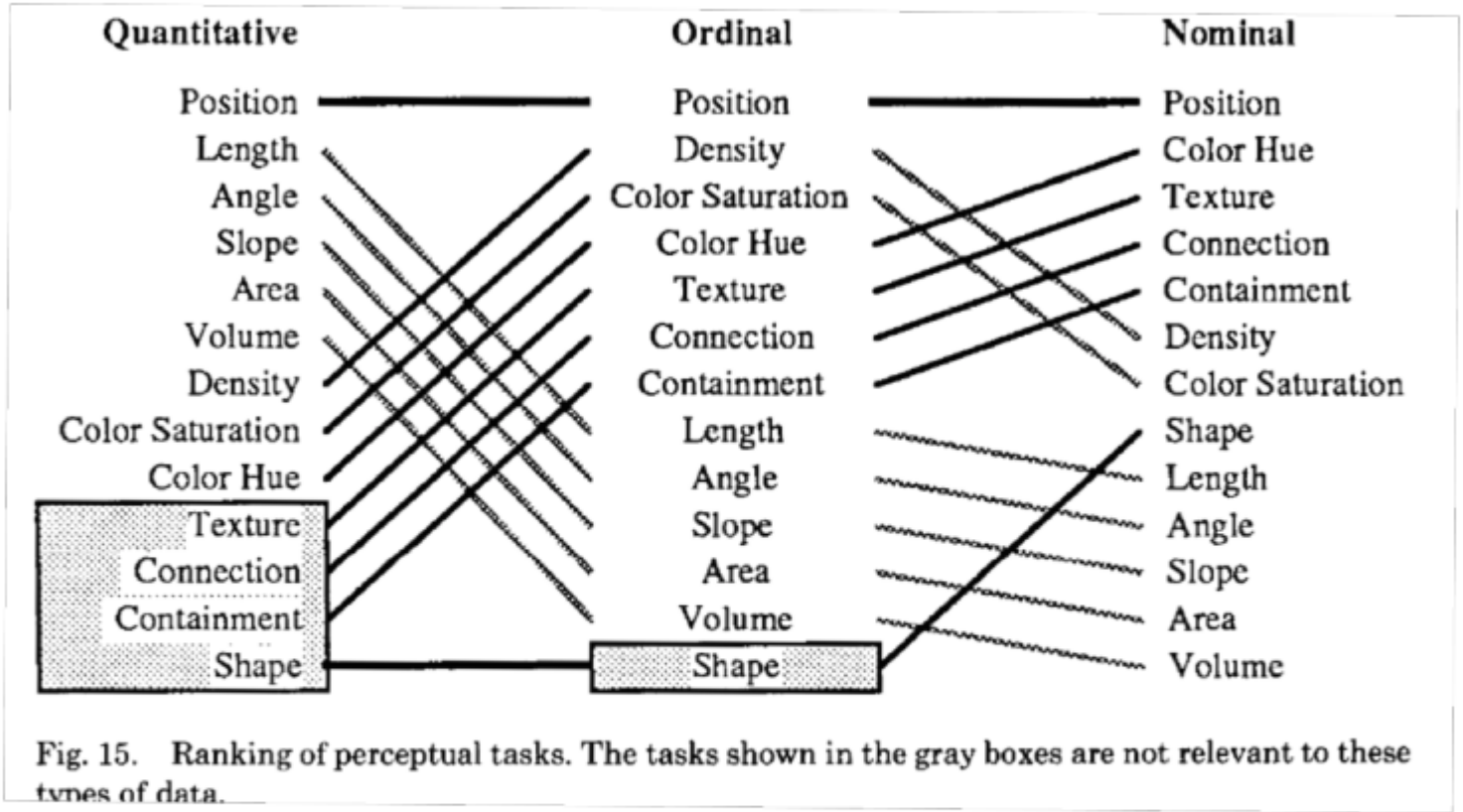
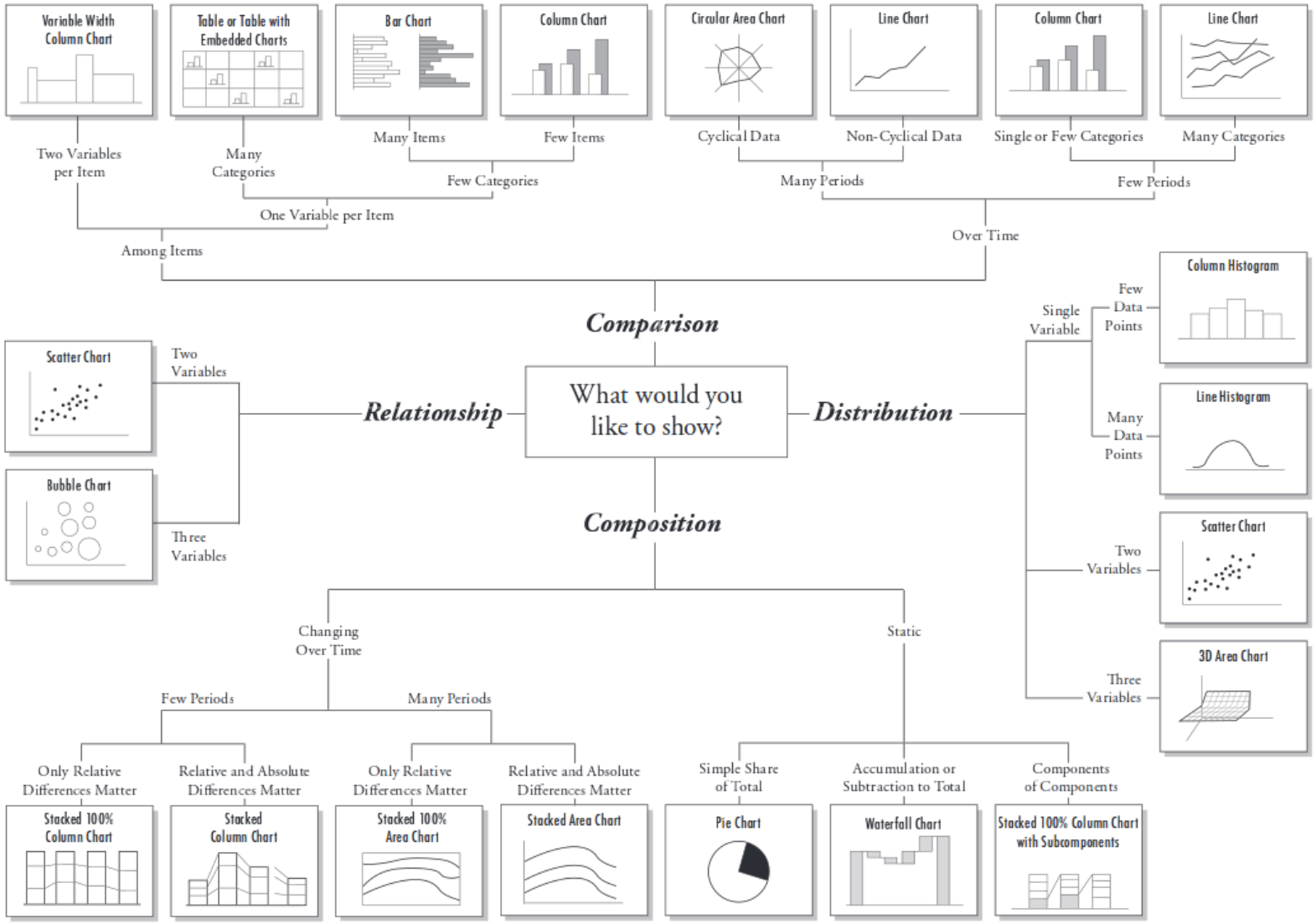


Fig. 15. Ranking of perceptual tasks. The tasks shown in the gray boxes are not relevant to these types of data.

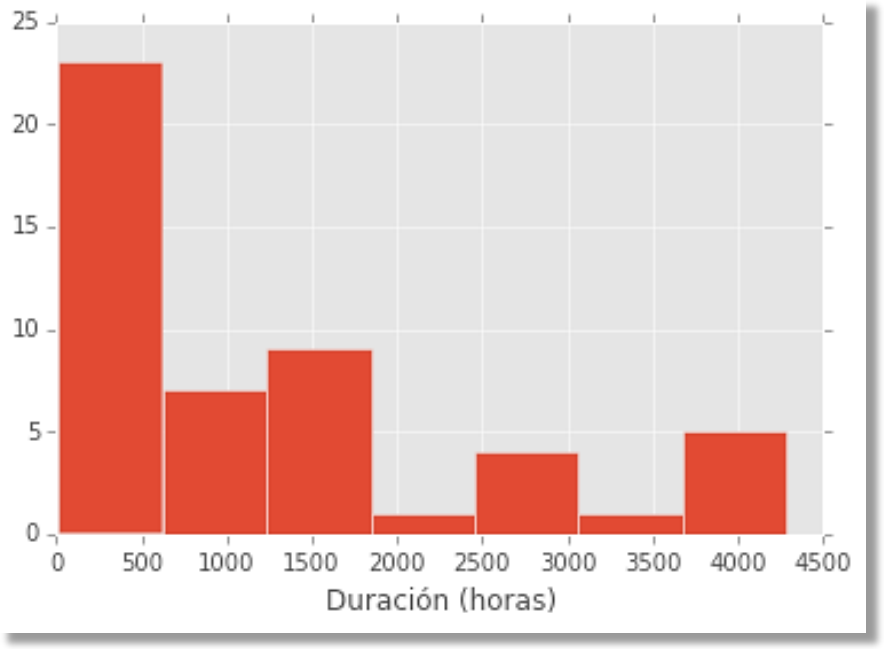
Mackinlay, J. (1986). Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics*, 5(2), 110-141. http://dx.doi.org.proxy.lib.duke.edu/10.1145/22949.22950

¿Qué tipo de gráfico escoger?



Histograma

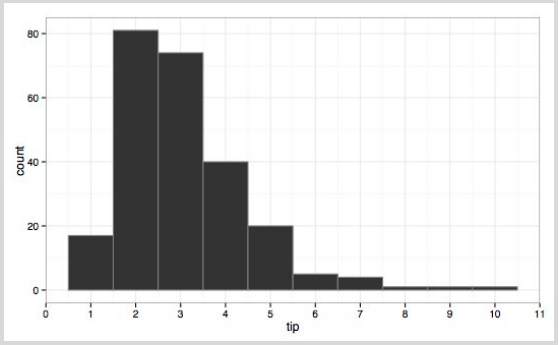
- Es una gráfica de barras verticales. Cada dato es parte de una sola barra, a la que también llamamos categoría. La anchura es habitualmente constante, y la altura es proporcional a la **frecuencia**, esto es, al número de datos en el intervalo. Si normalizamos a 1, obtenemos la **frecuencia relativa**.
- Un histograma es una representación **de la distribución de probabilidad** de una variable discreta. No es raro confundirlo con un diagrama de barras.

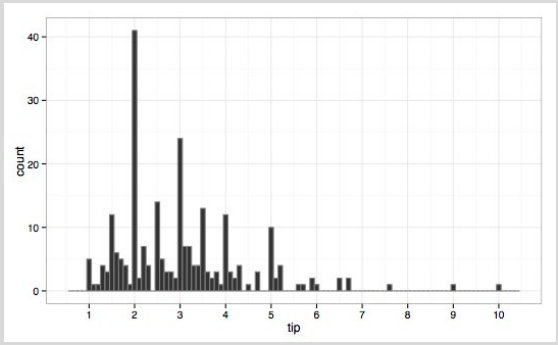


¿Qué defecto ves en esta figura?

Histogramas

- Si los intervalos no son de igual anchura, entonces es el área de cada barra la que resulta proporcional a la frecuencia de casos en el intervalo. Entonces la altura es la densidad de la frecuencia.
- Un histograma también proporciona una estimación de una **función de densidad de probabilidad** para una variable continua. Para ello se normaliza el área del histograma a 1.
- Uno de los elementos más importantes en el histograma es la elección de la anchura de las barras.

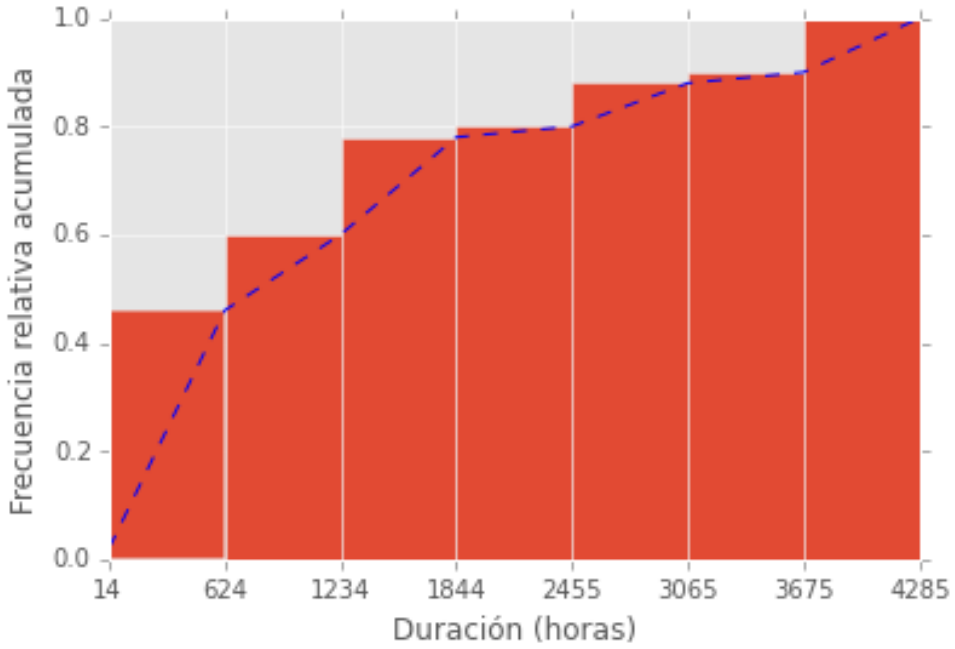




- **No hay un número de barras óptimo**. Una anchura mayor con una densidad pequeña reduce el ruido debido a la aleatoriedad del muestreo. Una anchura menor con una densidad alta añade precisión a la estimación de la densidad.
- Algunas fórmulas de cálculo: la **raíz cuadrada**:$\lceil\sqrt{n}\rceil$, **Sturges**: $\lceil\log_2 n\rceil$+1, **Rice**: $\lceil 2n^{1/3}\rceil$

Gráfica de distribución acumulada

- Las gráficas de distribución acumulada se obtienen a partir del histograma, donde cada barra acumula el porcentaje de datos correspondiente a cada categoría y a las anteriores.



- Permite responder a preguntas como: ¿Cuál es el porcentaje aproximado de baterías que fallará durante las primeras 1500 horas de operación? o ¿Qué representa una frecuencia relativa de 0.5?

Estadísticos robustos

- Medidas como la **media**, la **varianza** y la **desviación típica** se utilizan habitualmente para representar los datos, pero sufren una distorsión por la presencia de valores atípicos.
- Se proponen otros estadísticos, similares a los primeros, que sean robustos a la presencia de valores fuera de rango, o a pequeñas diferencias respecto a las típicas suposiciones basadas en modelos.
- La **mediana** es una medida robusta de tendencia central. Si los datos están ordenados de menor a mayor, la mediana es la observación central si el número de datos es impar, y es el promedio de los dos datos centrales si el número de datos es par.
- El **punto de ruptura** de la mediana, esto es, la proporción de valores incorrectos que puede manejar antes de proporcionar un resultado inexacto, es del 50%. El punto de ruptura de la media es del 0%.
- Algunas medidas robustas de dispersión son los **cuartiles**, los **deciles** y los **percentiles**:
  - Los cuartiles Q_1, Q_2 y Q_3 dividen la muestra en 4 partes iguales.
  - Los deciles $d_1,\ldots,d_9$ dividen la muestra en 10 partes iguales.
  - Los percentiles $p_1,\ldots,p_{99}$ dividen la muestra en 100 partes iguales.

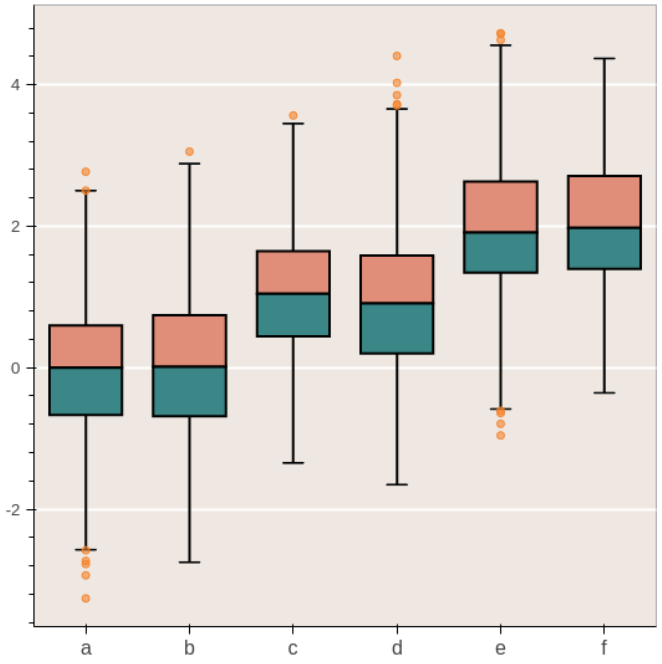
Gráficas de caja

- Los diagramas de caja se construyen a partir de las siguientes medidas:
  - El primer y el tercer cuartil, Q_1 y Q_3, que delimitan la caja central. La longitud de la caja recibe el nombre de **rango intercuartil**, $RI=Q_3-Q_1$.
  - Los **límites inferior y superior**, que se calculan como:

$$LI=\max\{\min\{x_i\}, Q_1-1.5\ RI\}$$

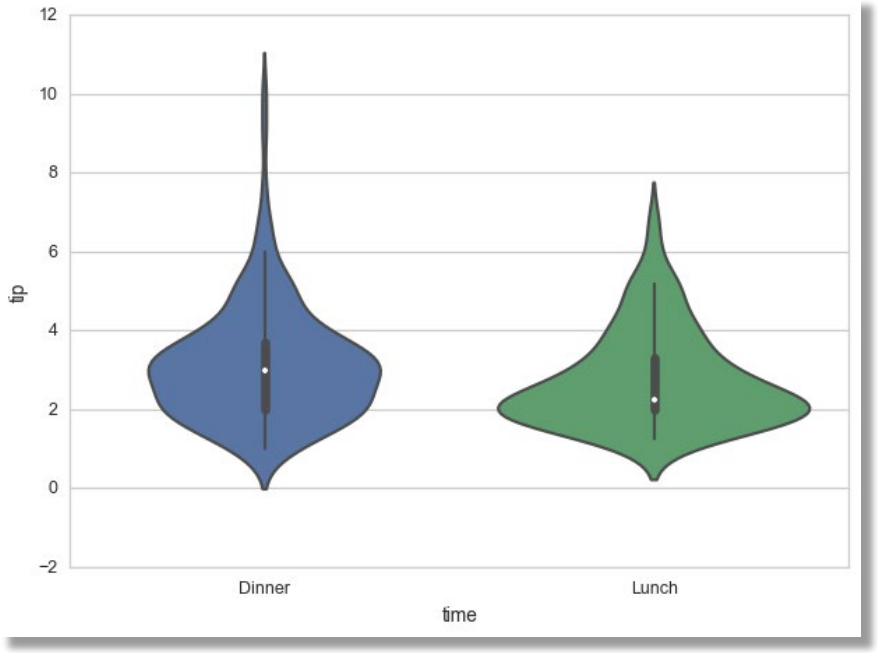
$$LS=\min\{\max\{x_i\}, Q_3+1.5\ RI\}$$

  - La **mediana**, o Q_2, representada mediante una línea que cruza la caja central.



Gráficos en violín

- Los gráficos en violín son una extensión de los diagramas de caja que muestran además la densidad de probabilidad de cada variable aleatoria contemplada en el experimento.
- La densidad de probabilidad se puede simplificar mediante su **histograma** o se puede estimar mediante algún **método no paramétrico basado en núcleos** *(kernel density estimation)*, funciones cuya integral es uno y su media es cero.



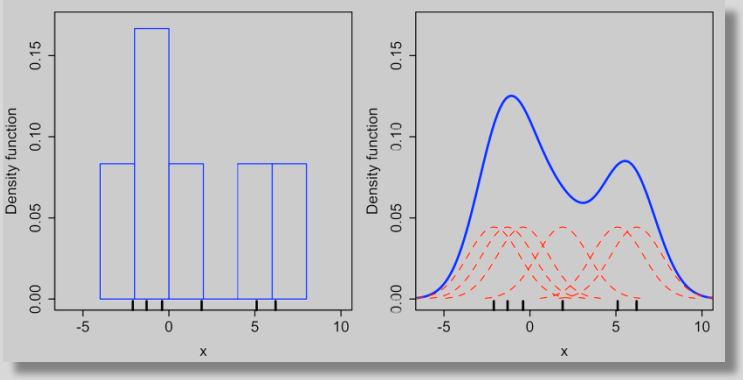
Métodos de estimación basados en núcleos

kernel density estimation (kde)

- Son un conjunto de métodos no paramétricos para estimar la **función de densidad de probabilidad** de una variable aleatoria.
- En la práctica nos permiten obtener una curva que suaviza el comportamiento del histograma a partir de una muestra.
- Sea (x_1,x_2,…,x_n) un conjunto de muestras independientes y uniformemente distribuidas obtenidas a partir de una distribución desconocida. Su estimación kde es una función:

$$\widehat{f_h}(\mathrm{x}) = \frac{1}{n}\sum_{k=1}^{n} K_h(x - x_k)$$

- La función núcleo K() es una función no negativa, cuya integral es uno y su media cero. El parámetro h realiza un suavizado de la función y se denomina **ancho de banda**. Se cumple que $K_h(x)=1/h\ K(x/h)$. Funciones núcleo son la triangular, la uniforme, la normal, etc.

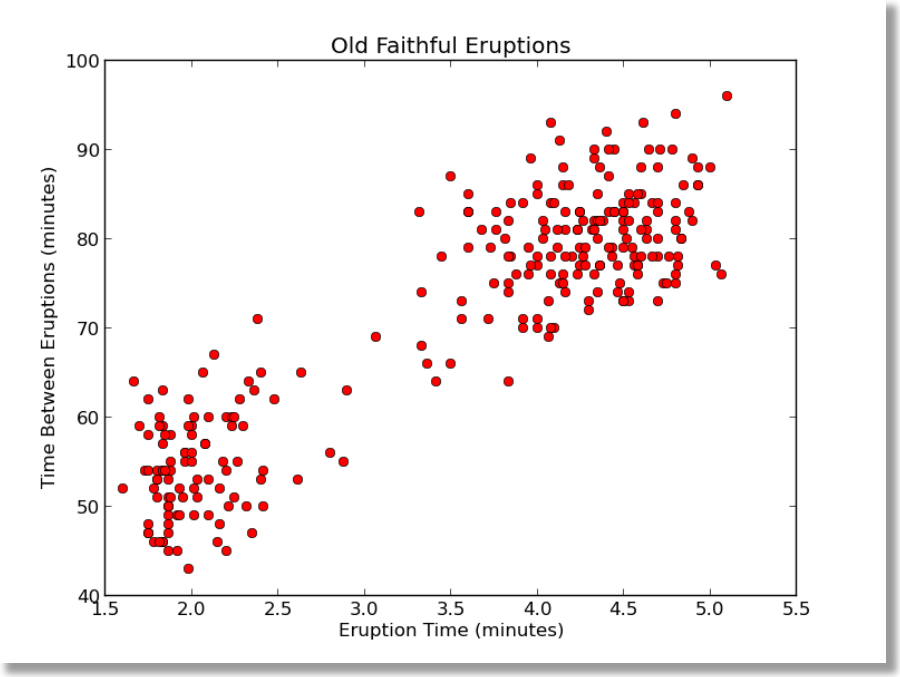


$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{1/5}$$

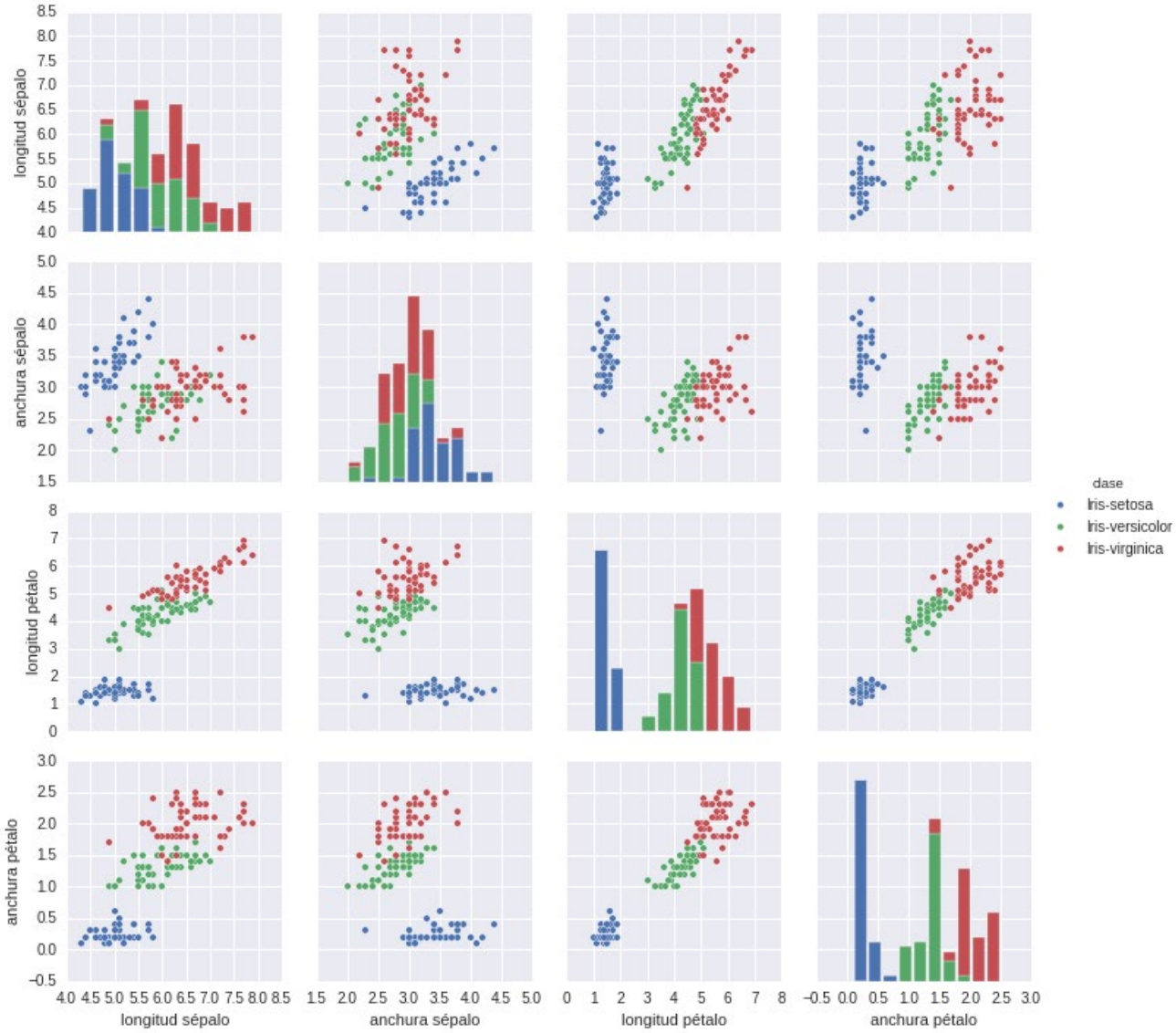
ancho de banda para núcleo normal

Diagramas de dispersión

- Los diagramas de dispersión permiten conocer la distribución de valores de los datos en espacios de múltiples dimensiones.
- Representan por parejas los distintos atributos de los datos.
- Son especialmente útiles como exploración previa en problemas de clasificación o agrupamiento.

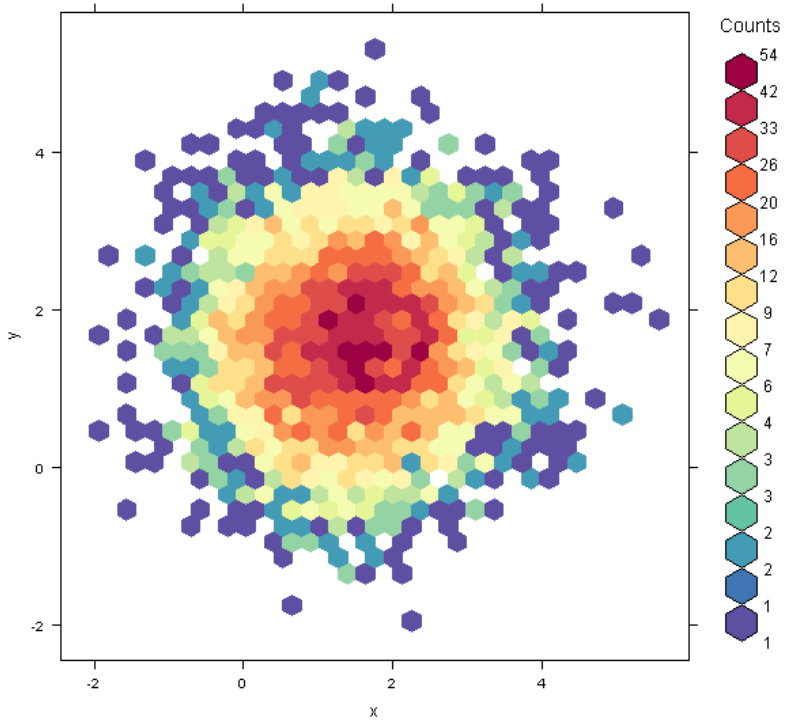


Diagramas de dispersión



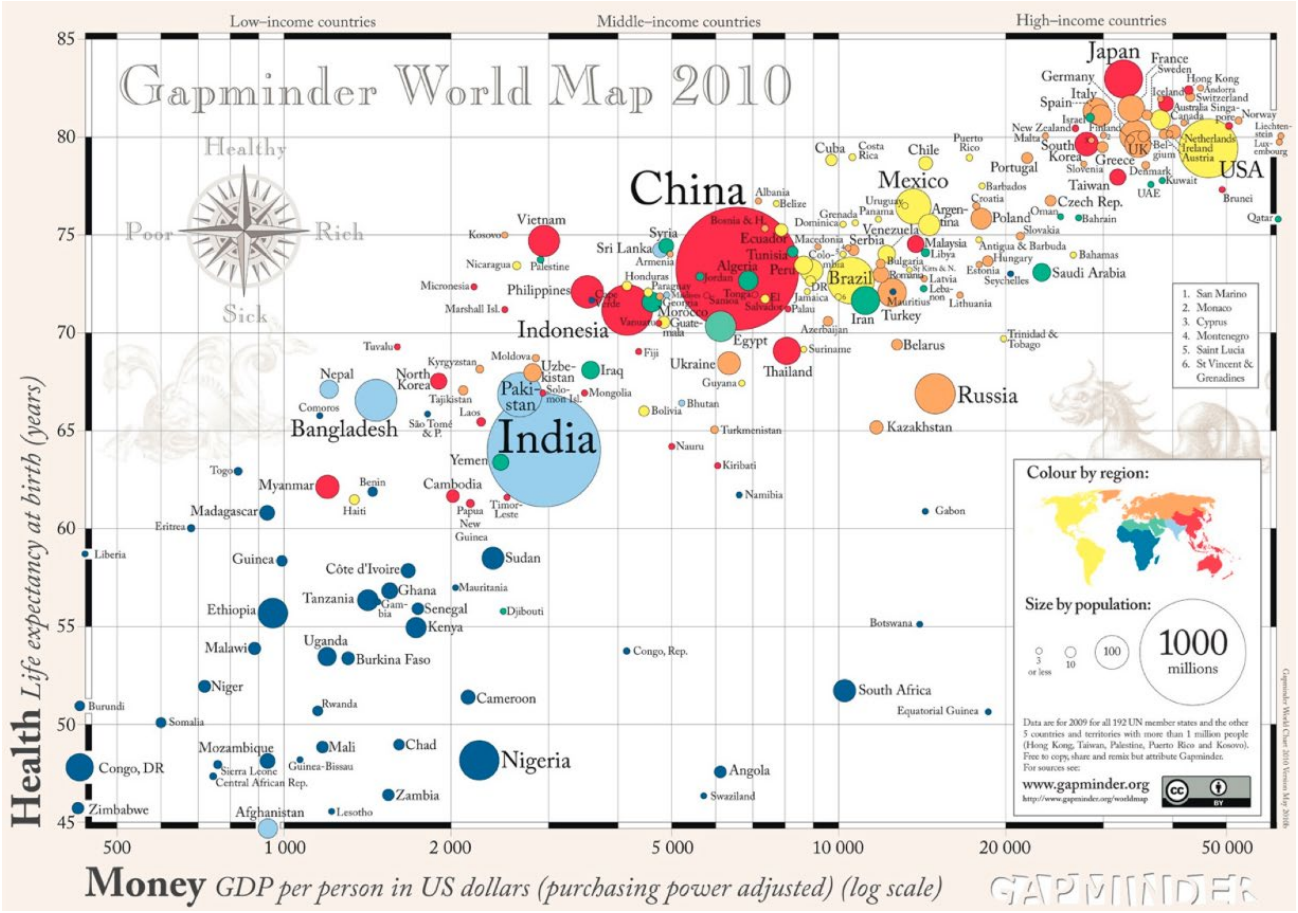
Gráficos de agrupamiento hexagonal

- Los gráficos de agrupamiento hexagonal son una forma de histograma en dos dimensiones donde se realiza un conteo del número de muestras que se encuentran en cada uno de los hexágonos que permite teselar el plano. Suele ser más informativo que un diagrama de dispersión.



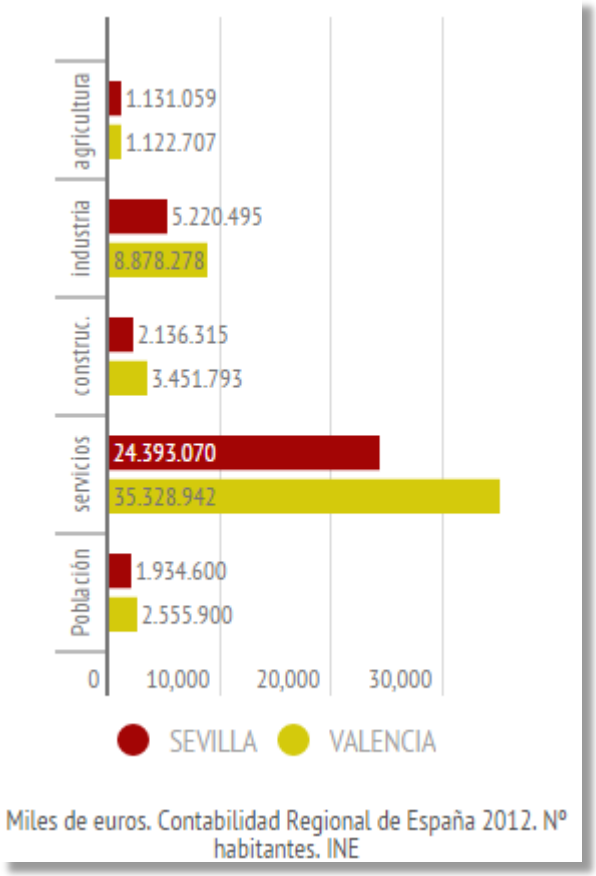
Gráficos de burbujas

- Los diagramas de burbujas son un tipo de gráficos de dispersión en los que aparece una tercera componente de los datos, en forma de radio de una burbuja centrada en las coordenadas de las dos primeras componentes del diagrama de dispersión.



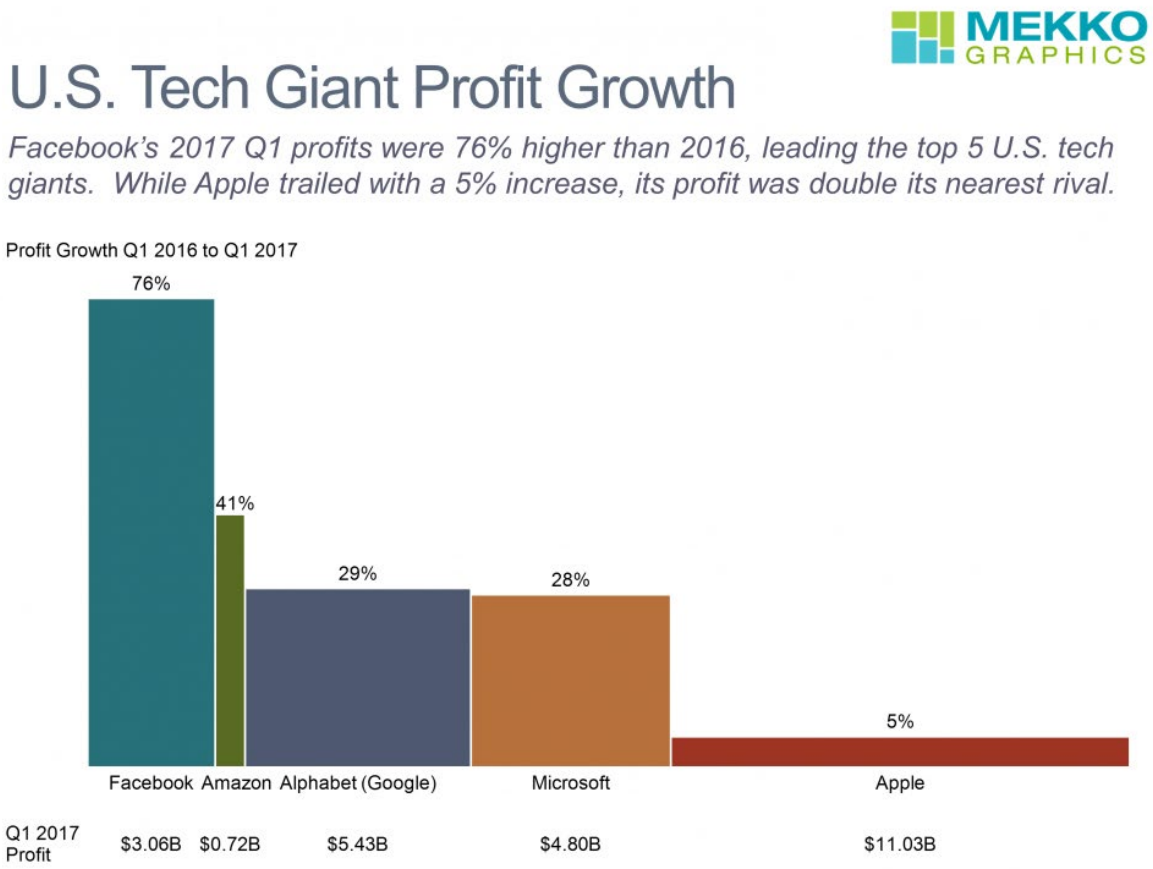
Gráficos de barras

- Permiten comparar múltiples entidades entre sí o múltiples categorías.



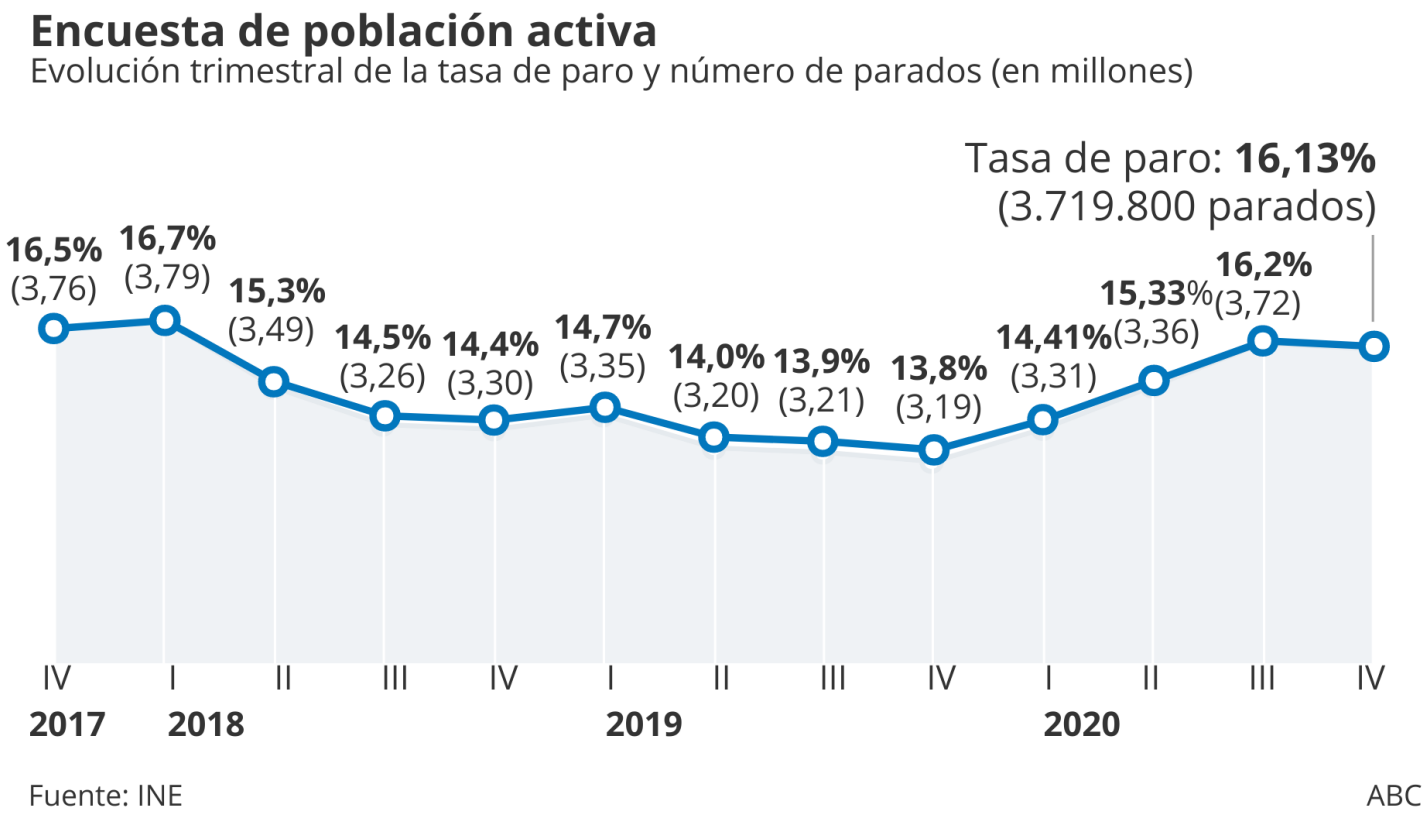
Gráficos de barras con ancho variable

- Son representaciones bidimensionales, conteniendo información en ambos ejes. Admiten apilamiento en cada categoría.



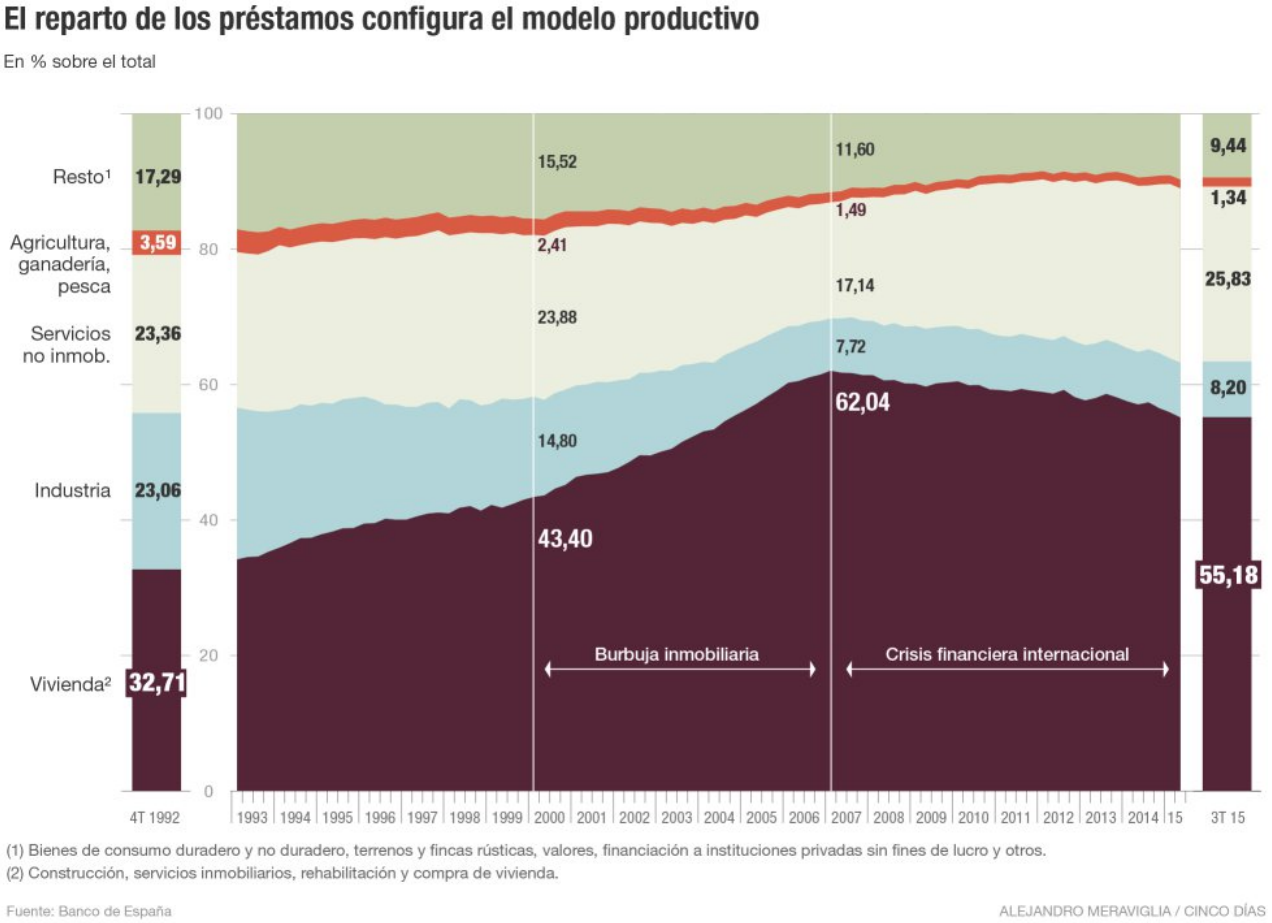
Gráficos de barras a lo largo del tiempo

- Se busca realizar comparaciones entre variables tomando como base de representación el tiempo.



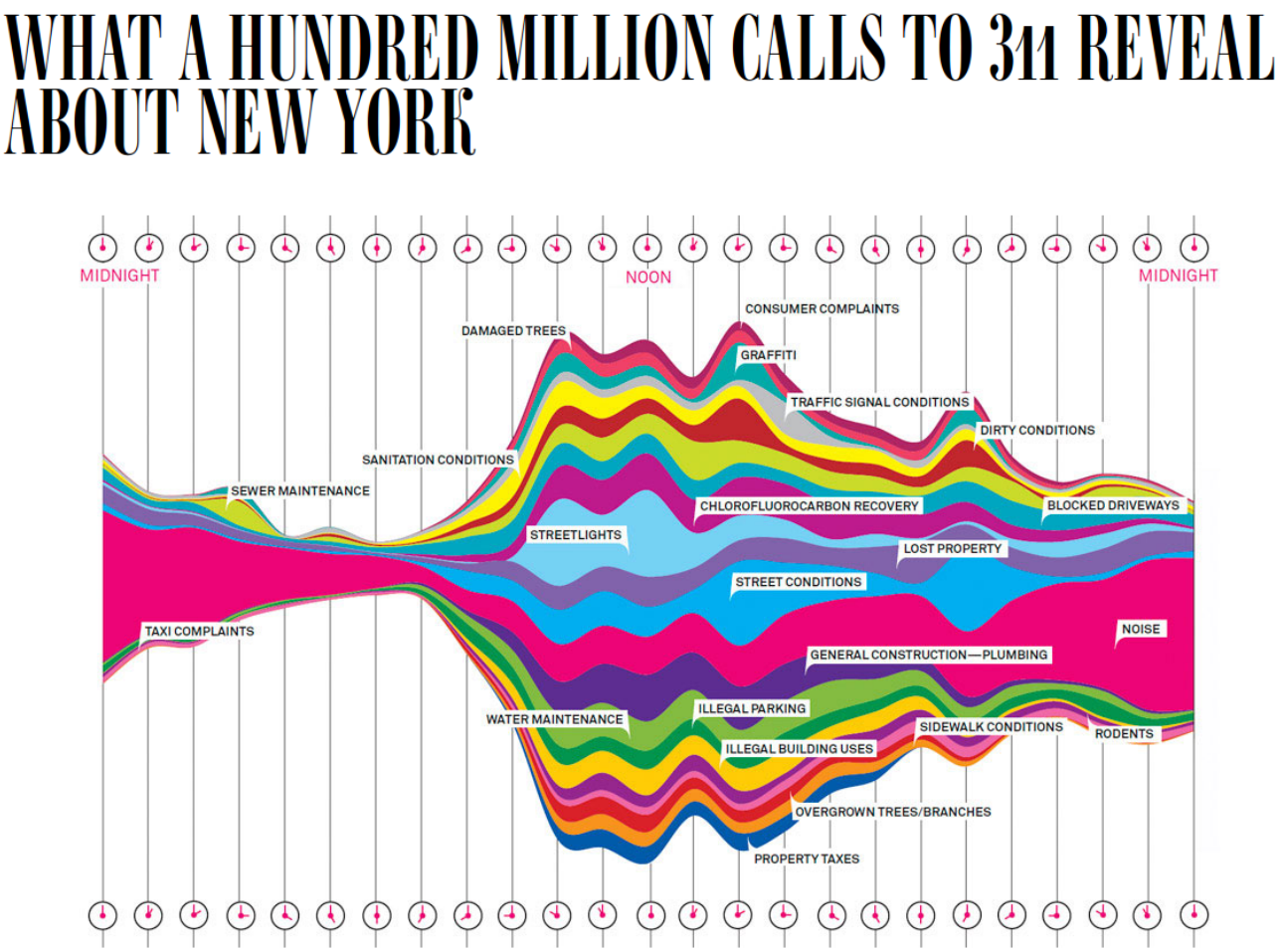
Gráficos de área

- A veces es particularmente interesante apilar los valores de las categorías, cuando interesa mostrar incrementos, o cuando representan una parte de un total.



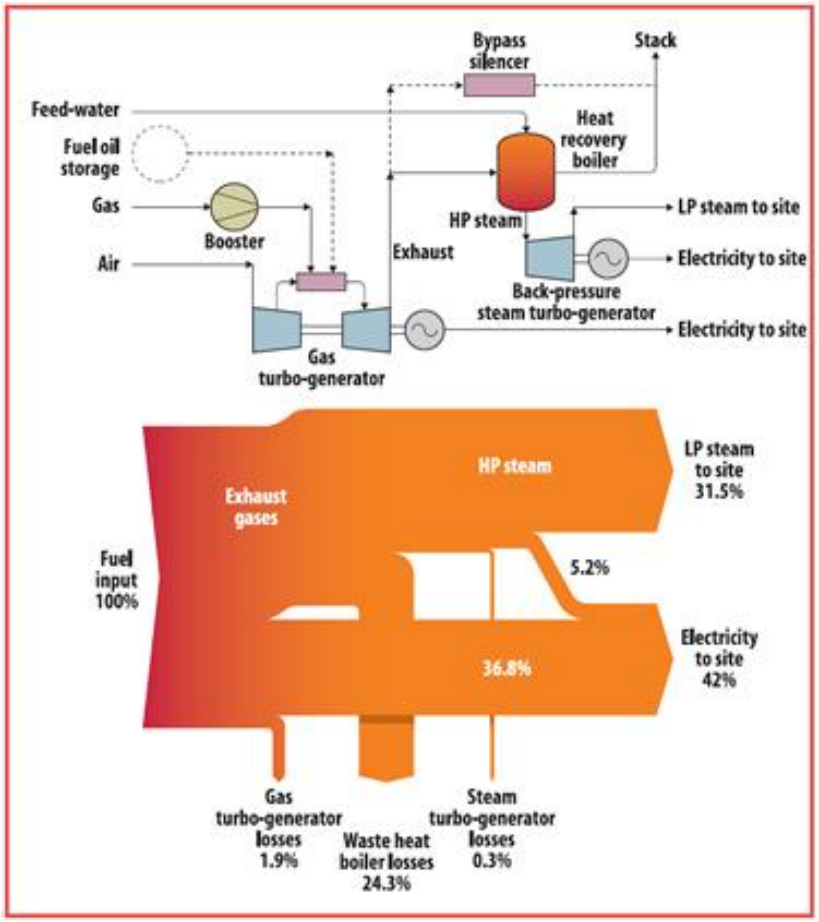
Gráficos de corriente

- A partir de los gráficos de barras o de líneas apilados se construye este tipo de gráficos donde la información está contenida en el área y se centra la figura en un eje horizontal.



Gráficos Sankey

- Es un tipo de diagrama de flujo donde el ancho de las flechas es proporcional al tamaño del flujo.

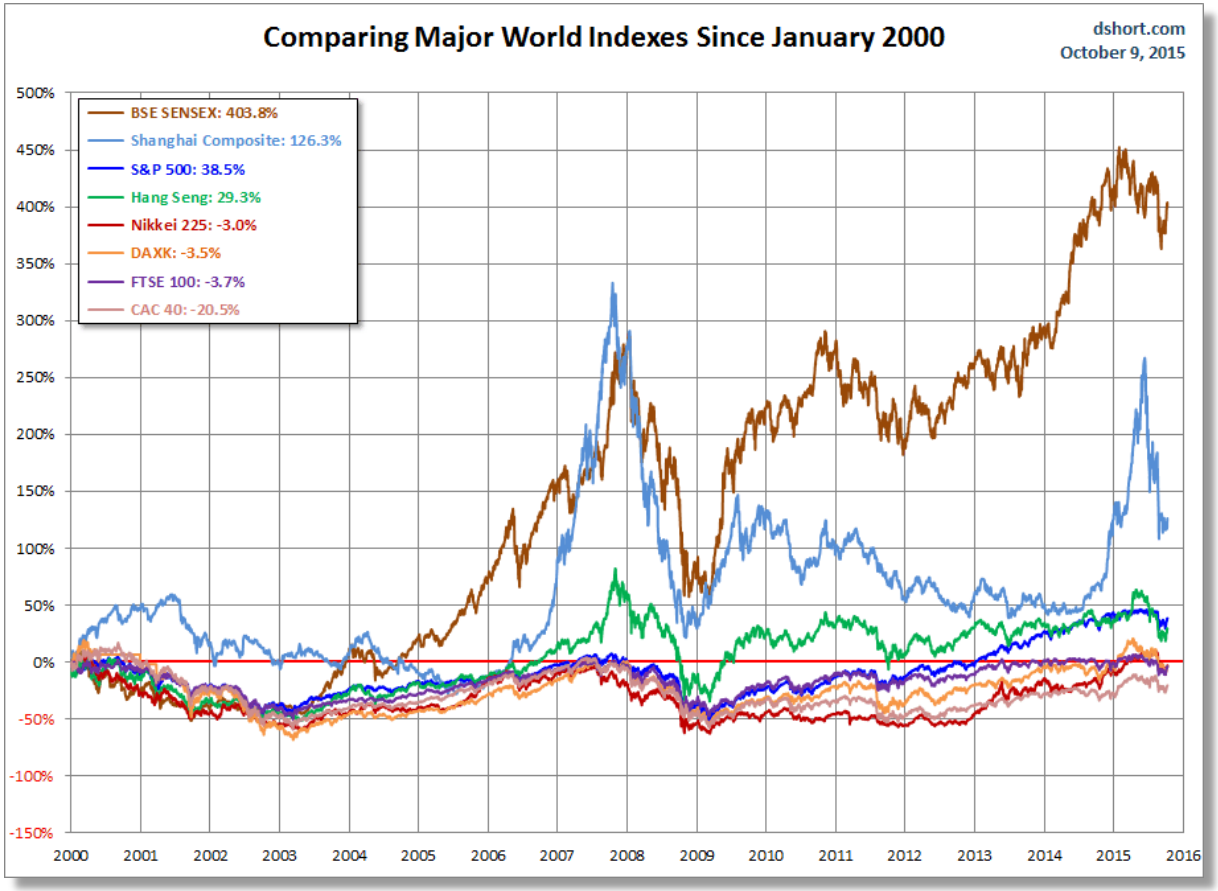


Diagramas de líneas

- En algunos casos se puede usar un diagrama de líneas cuando se desean analizar tendencias y se desprecia el error de interpolación.

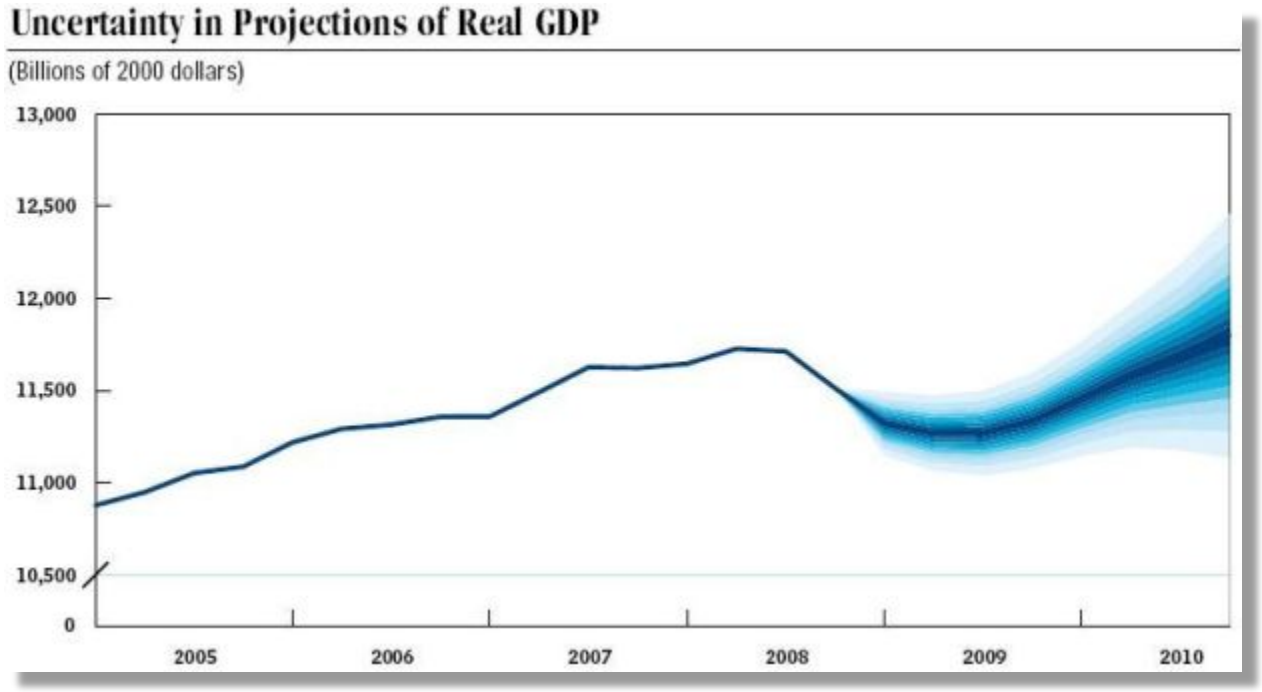
Diagramas de líneas

- Particularmente útiles para comparar la evolución de múltiples parámetros en el tiempo.



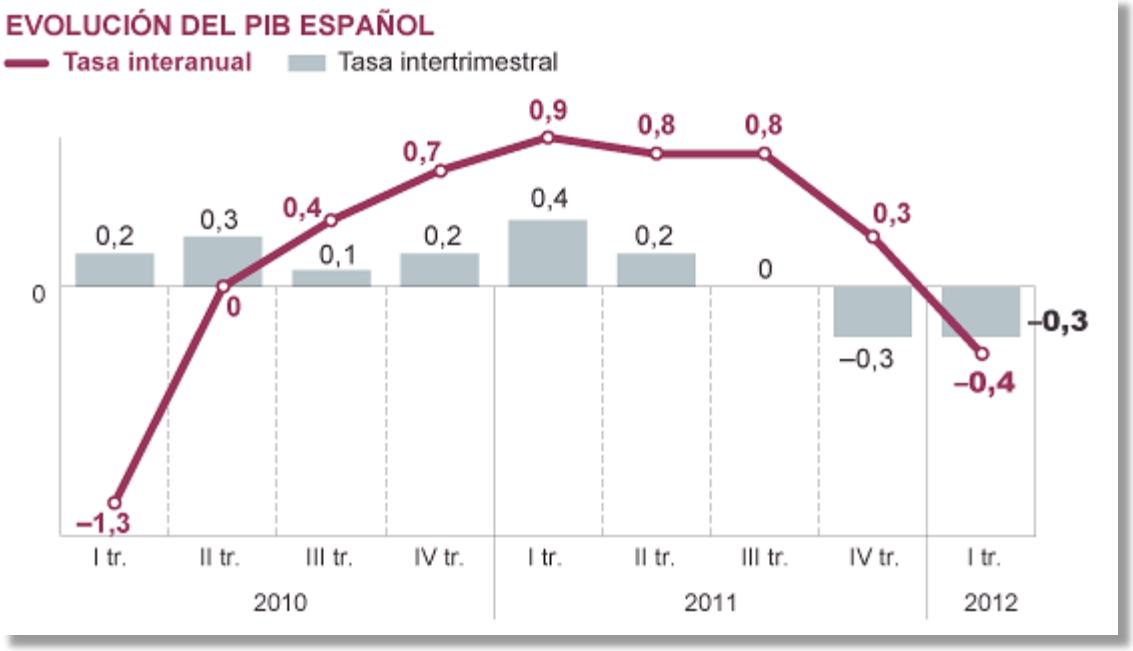
Diagramas de proyección

- Son una extensión de los diagramas de líneas en los que se dispone de un modelo de predicción con incertidumbre. Se proyectan en el futuro distintas predicciones donde su probabilidad se traduce en un esquema de colores.



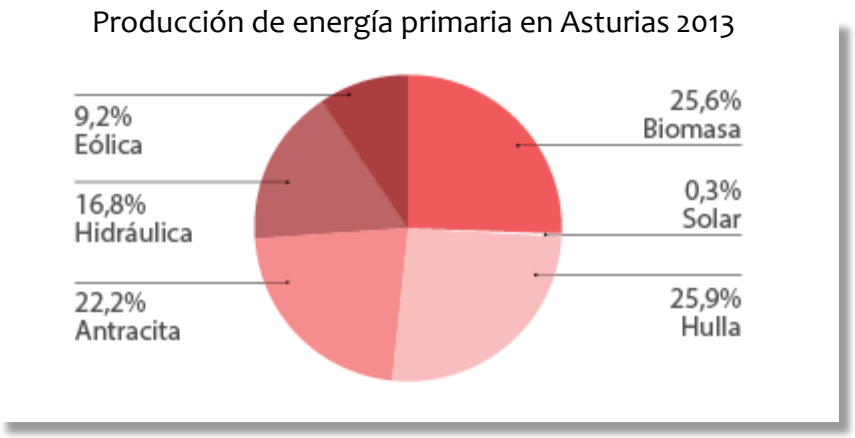
Diagramas combinados

- Es importante valorar la utilización de múltiples esquemas de representación para mejorar la calidad de la información proporcionada.



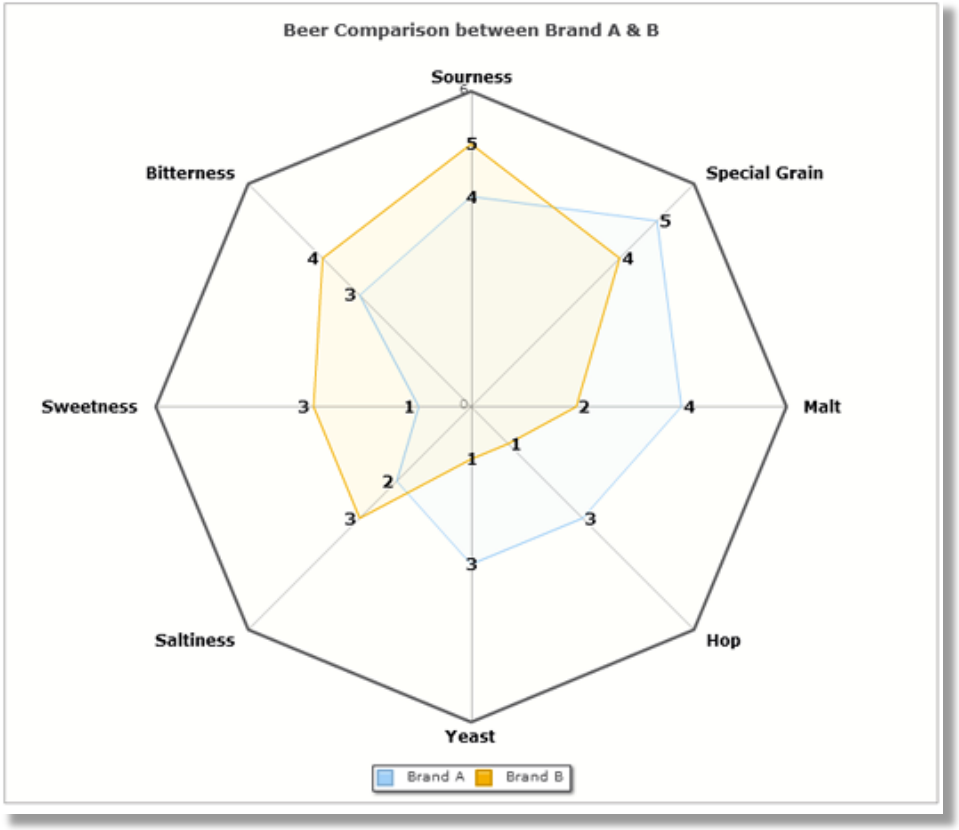
Diagramas de sectores

- Se utilizan para representaciones de cantidades fraccionarias.

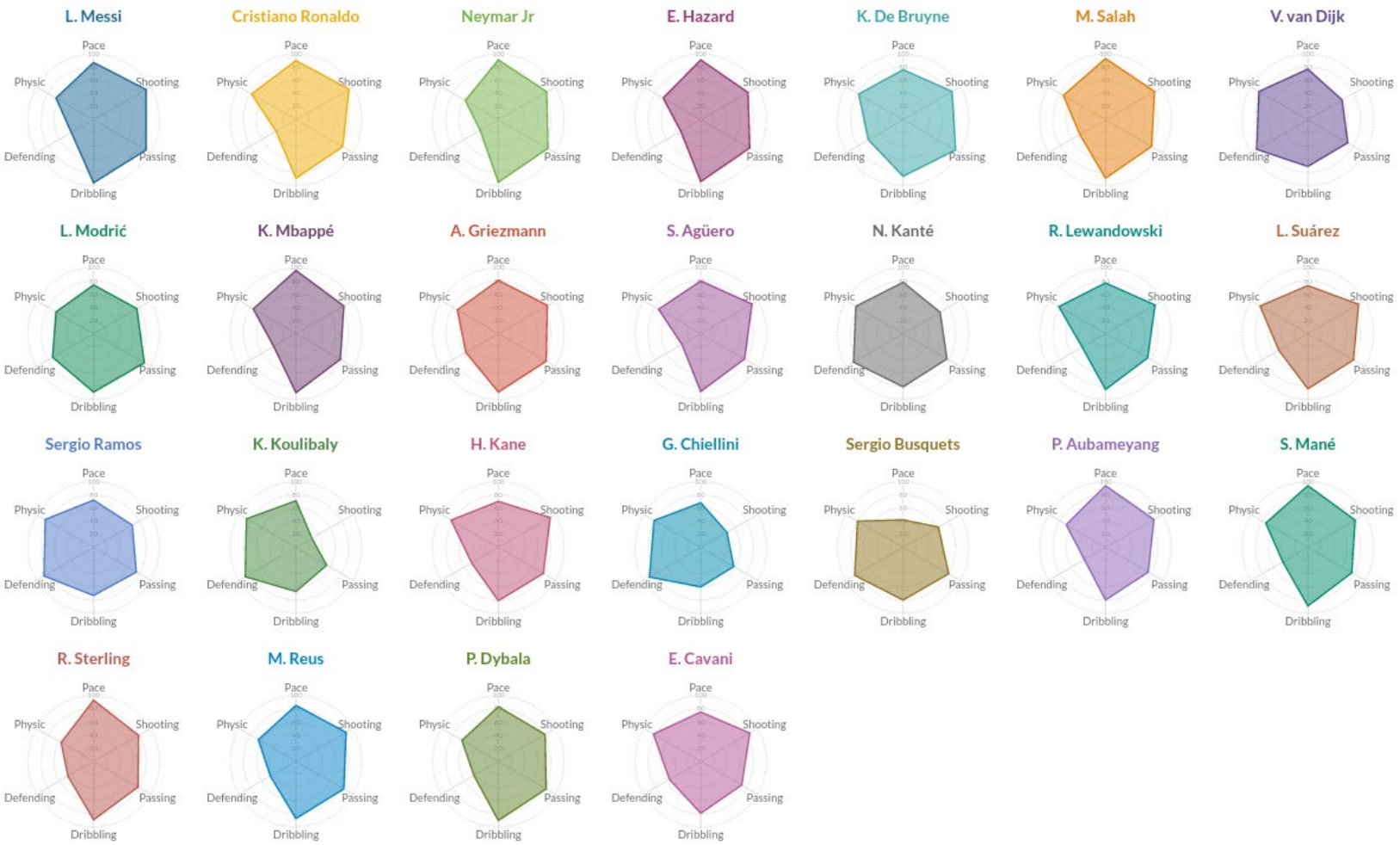


Gráficos en tela de araña

- También llamados gráficos de radar, gráficos poligonales, o gráficos web, entre otros. Permiten visualizar múltiples dimensiones con distintas escalas en un mismo gráfico en el plano. Es necesario seleccionar de modo muy cuidadoso el orden de representación de los ejes y sus escalas para poder visualizar patrones relevantes.

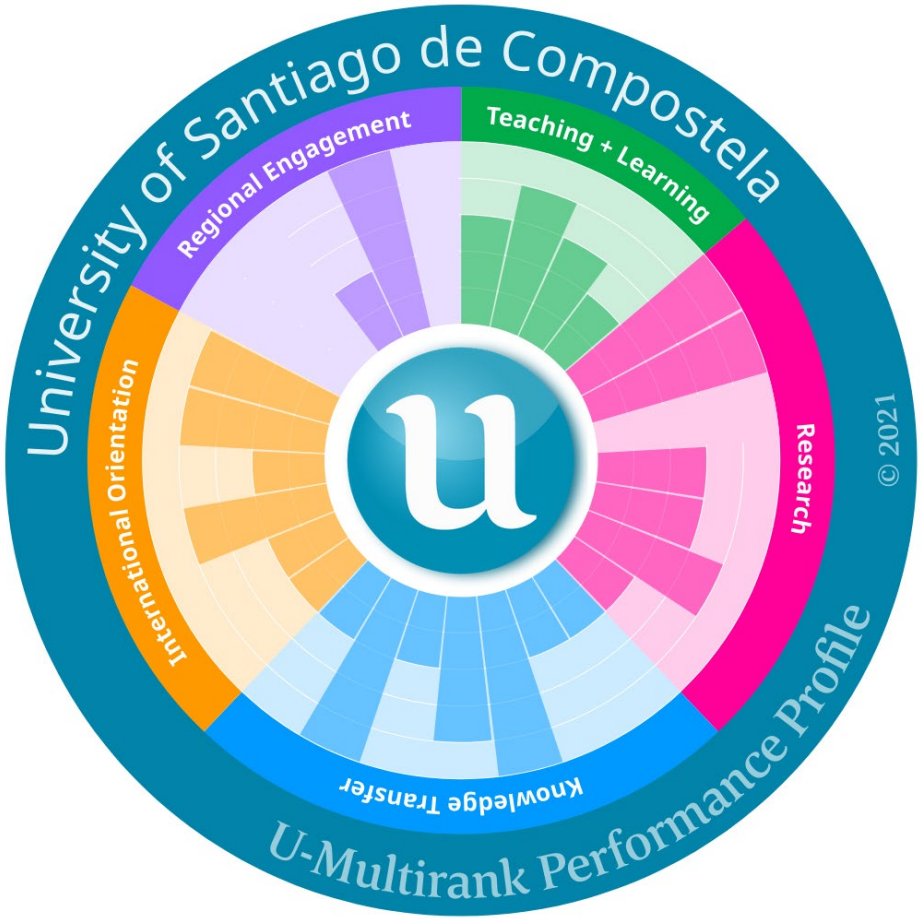


Gráficos en tela de araña



Gráficos de áreas polares

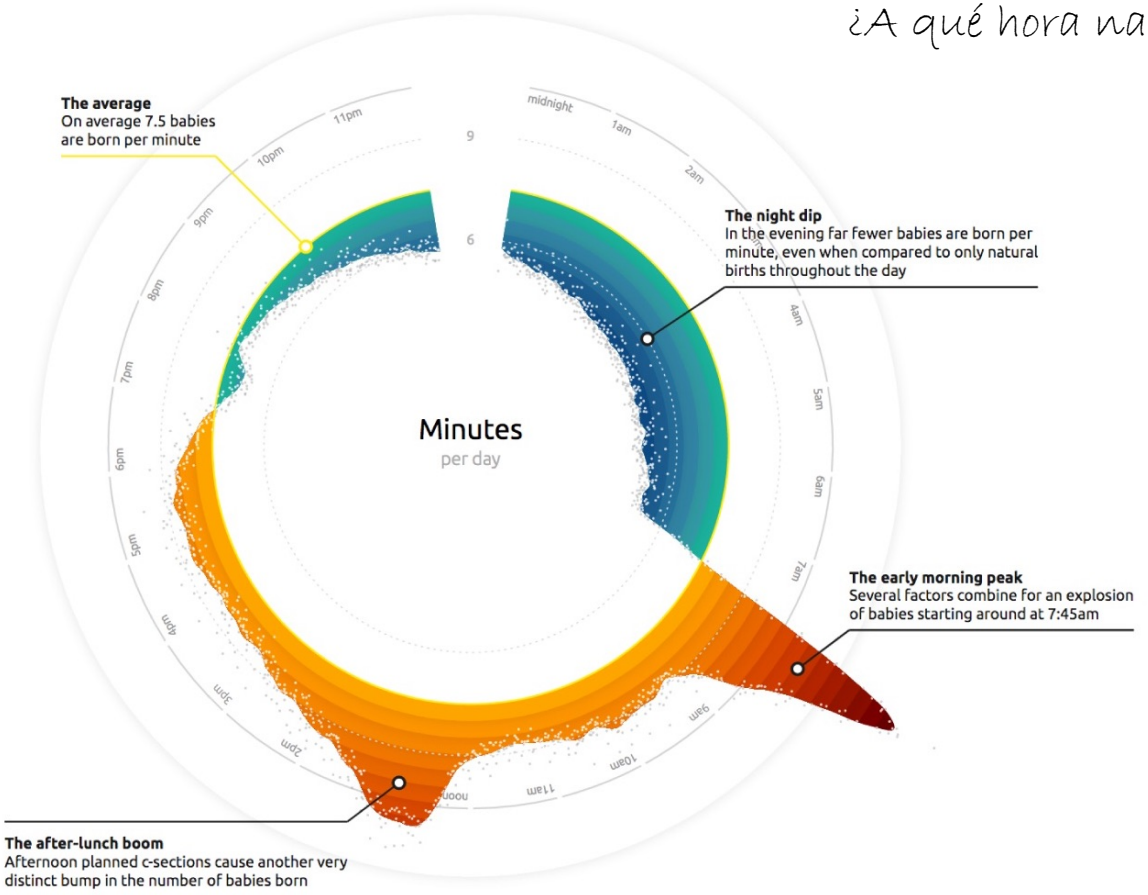
- También permiten visualizar múltiples dimensiones con distintas escalas en un mismo gráfico en el plano. El valor de cada observación es proporcional a la distancia de cada sector respecto al centro del círculo.



Comportamientos periódicos

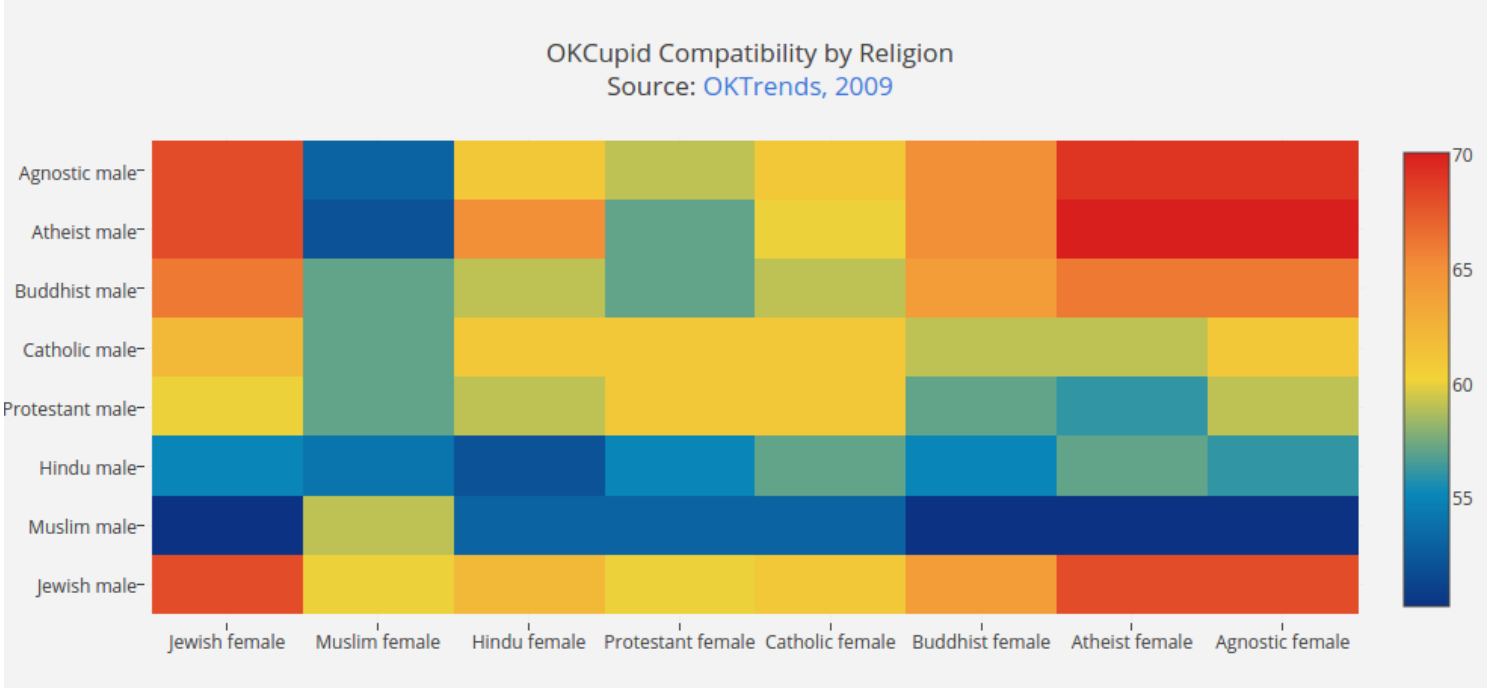
- Los gráficos circulares se utilizan en ocasiones para visualizar comportamientos periódicos o estacionales, sobre todo cuando en cada ciclo destacan en el tiempo determinados eventos.

¿A qué hora nacen los niños?



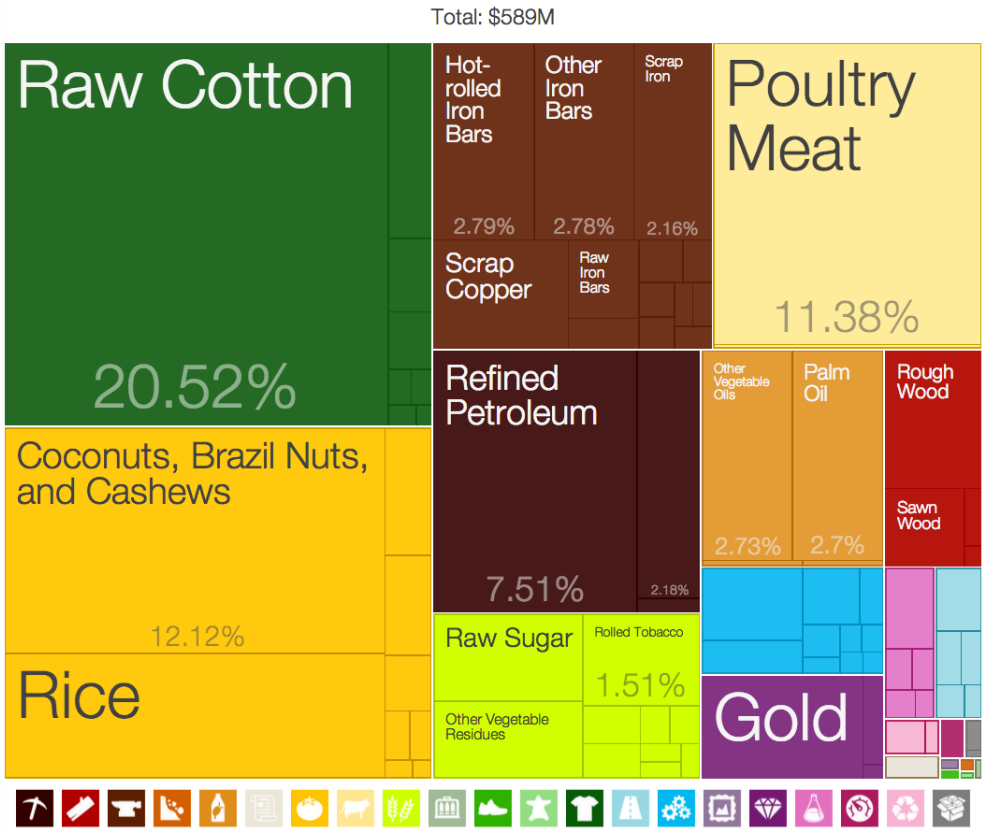
Mapas de calor

- Son representaciones gráficas de una matriz donde los valores numéricos son transformados en códigos de colores. Facilitan tareas de agrupamiento visual si podemos realizar cambios en la ordenación de filas y columnas.



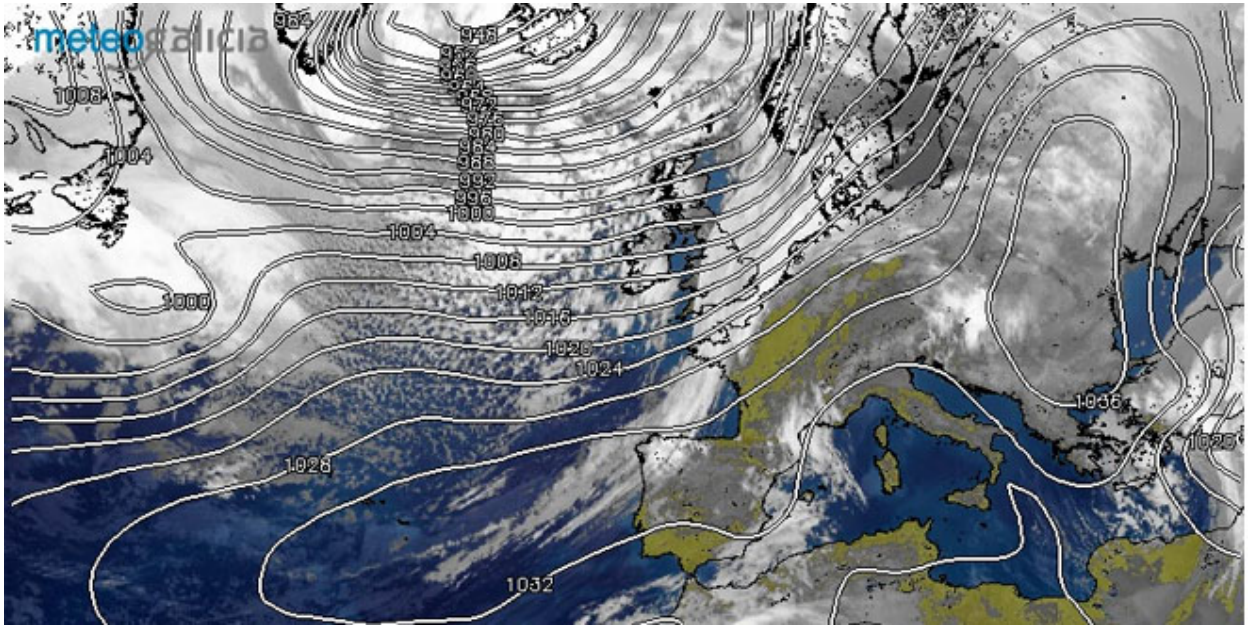
Mapas jerárquicos

- Son representaciones gráficas que proceden de la clasificación jerárquica realizada a partir del resultado de un árbol de decisión (*tree maps*). A cada rama del árbol le corresponde un rectángulo, que se subdivide en tantos rectángulos como sub-ramas.



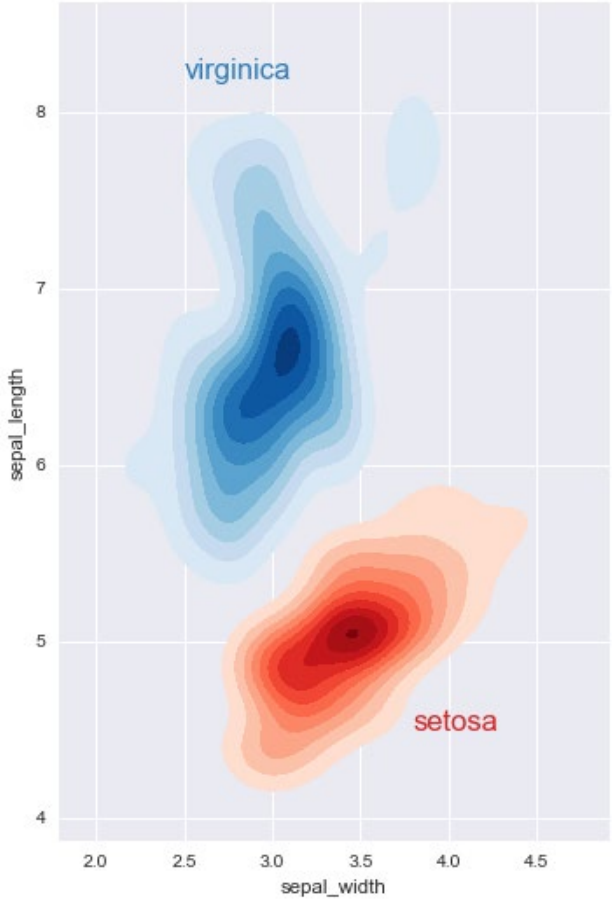
Mapas de contorno

- Son representaciones gráficas en las que se utiliza una curva para unir puntos que comparten un mismo valor para una función determinada. Son habituales en meteorología, para representar isobaras, o en representaciones cartográficas.



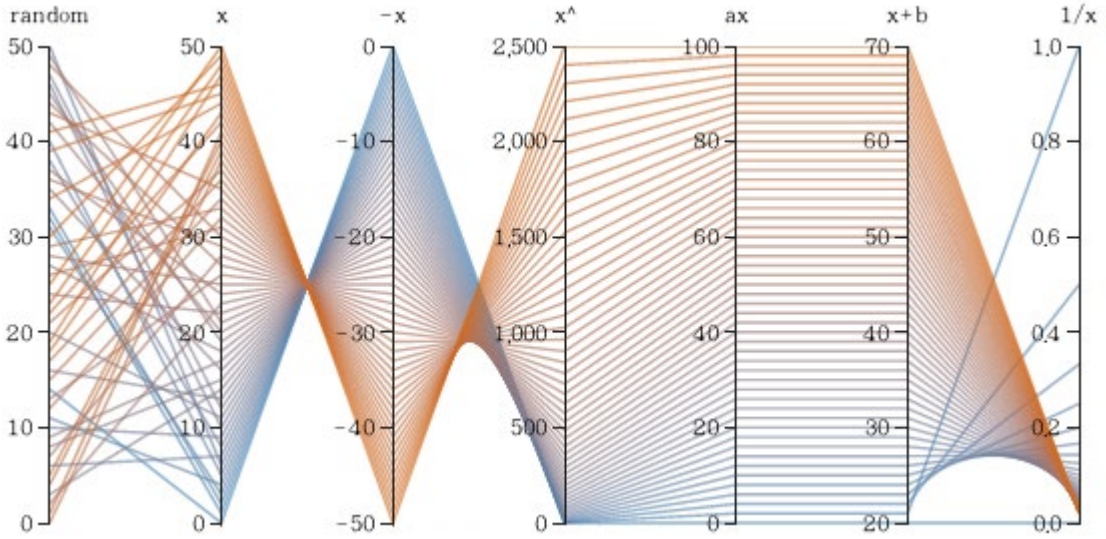
Mapas de contorno

- El uso de mapas de contorno se ha extendido a la representación de histogramas y funciones de densidad de probabilidad (*kernel density estimation*).



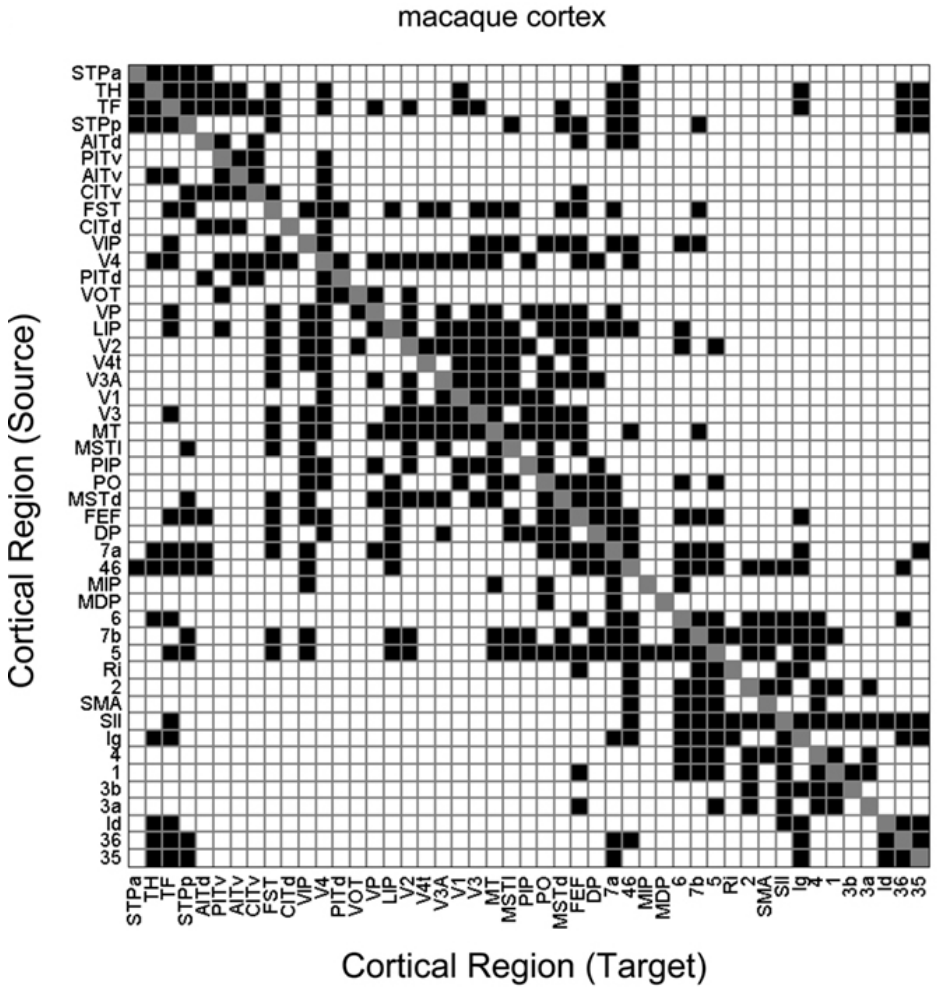
Diagramas de coordenadas paralelas

- Se utilizan para visualizar datos en espacios de múltiples dimensiones. Se representa el conjunto de dimensiones como ejes paralelos, de modo que un punto se representa mediante una línea quebrada que une los valores que toma en cada una de las coordenadas. Permite el reconocimiento de patrones.



Mapas de adyacencia

- Se utiliza para representar grafos complejos mediante una matriz que colorea los nodos que están conectados.

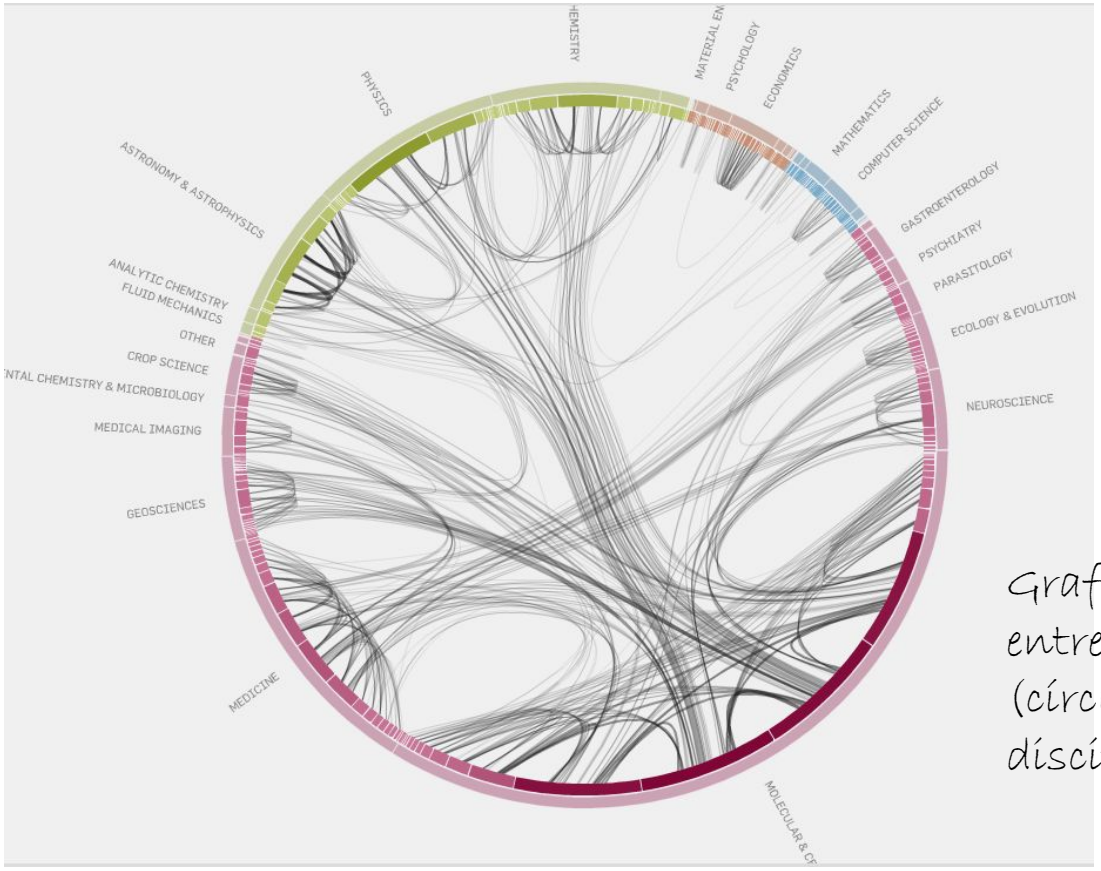


Una vez que se dispone de la matriz de adyacencia esta se reordena para separar aquellos subgrupos formados por nodos interconectados entre sí, realizando una permutación de filas (y consiguientemente, columnas)

E. Forsyth and L. Katz, 1946

Gráficos circulares

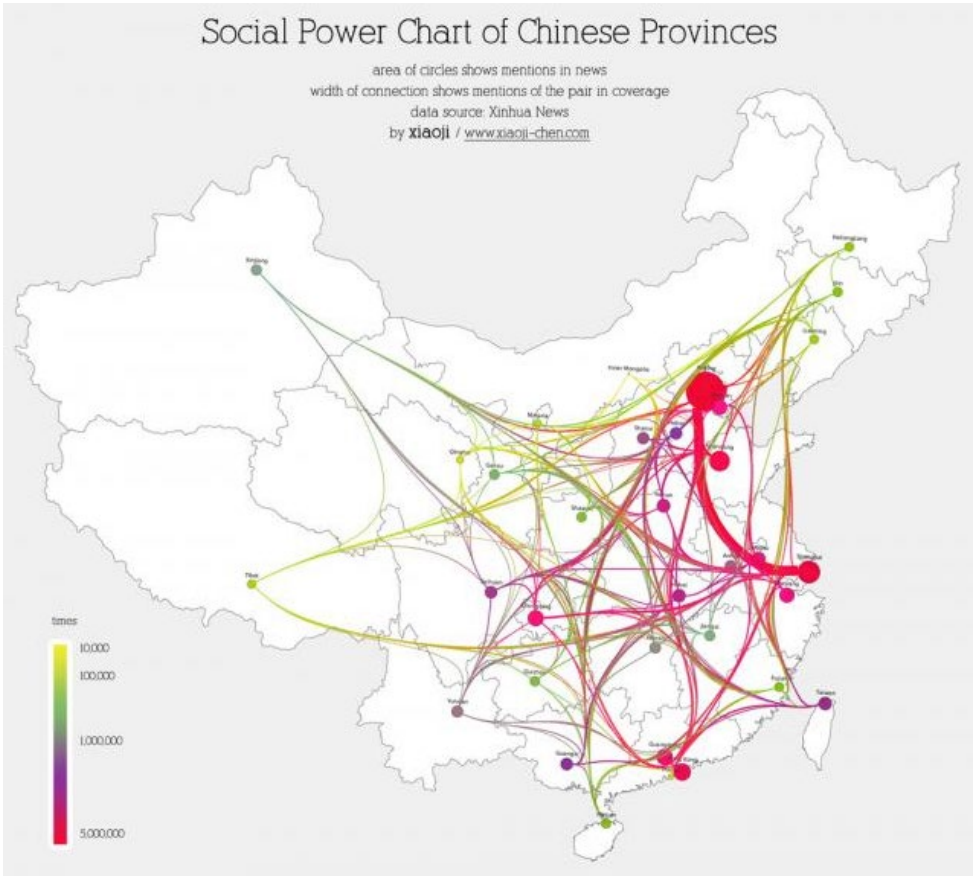
- Son una representación alternativa a un grafo no dirigido, útil cuando el número de nodos y arcos resulta muy alto. Puede ser útil para la identificación de patrones a partir de una acumulación de arcos.



Grafo que representa citas entre distintas publicaciones (círculo interior) y distintas disciplinas (círculo exterior).

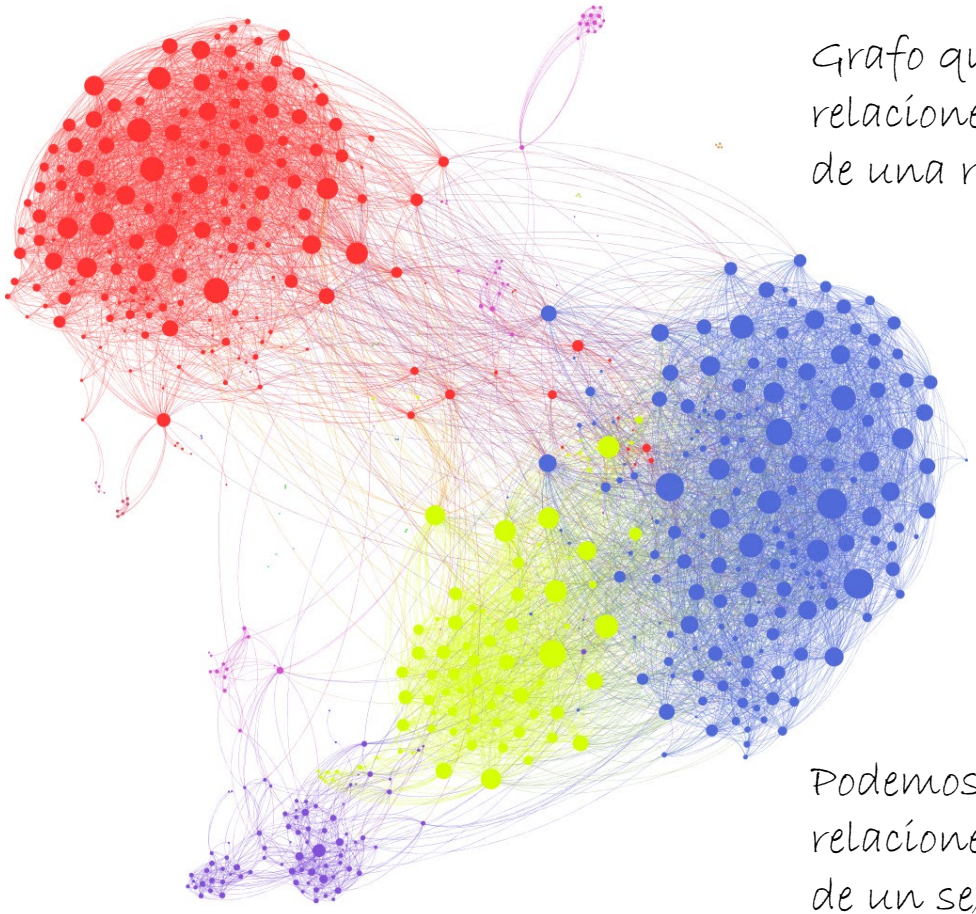
Grafos extendidos

- Son una representación extendida de un grafo, al que se añade más información parametrizando los elementos de la visualización (tamaño de los nodos, escalas de tonalidad, grueso de los arcos, etc.).



Redes

- La representación de relaciones en red es útil para realizar un gran número de preguntas, o realizar operaciones de agrupamiento o clasificación, de un modo muy intuitivo.

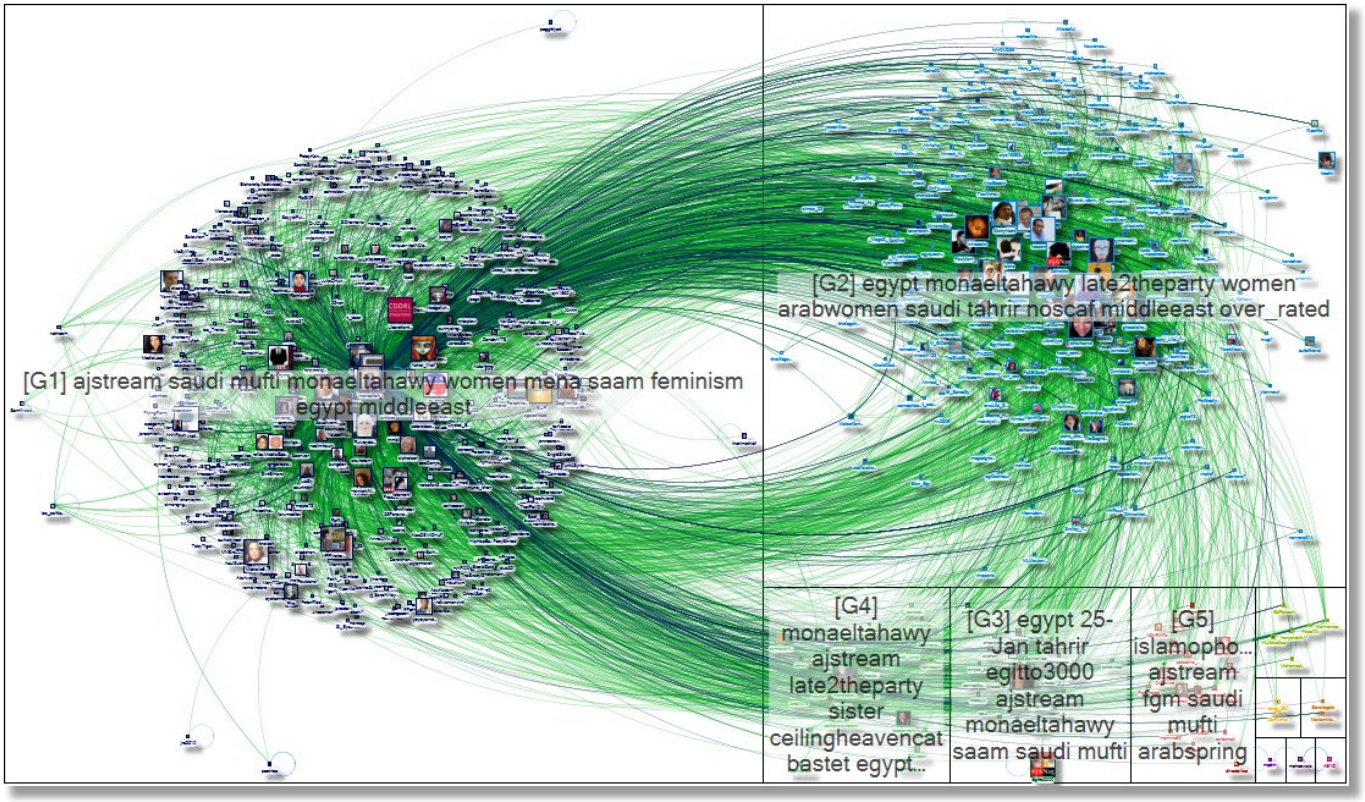


Grafo que representa relaciones entre individuos de una red social

Podemos preguntar por las relaciones entre individuos de un sexo, de una edad, etc.

Redes

- Es importante la búsqueda de patrones a partir de la visualización de redes, en particular, en redes sociales.



Se muestra un ejemplo de patrón de polaridad en la red social Twitter ante un artículo sobre la represión de la mujer en Oriente Medio.

Representación de grafos mediante fuerzas

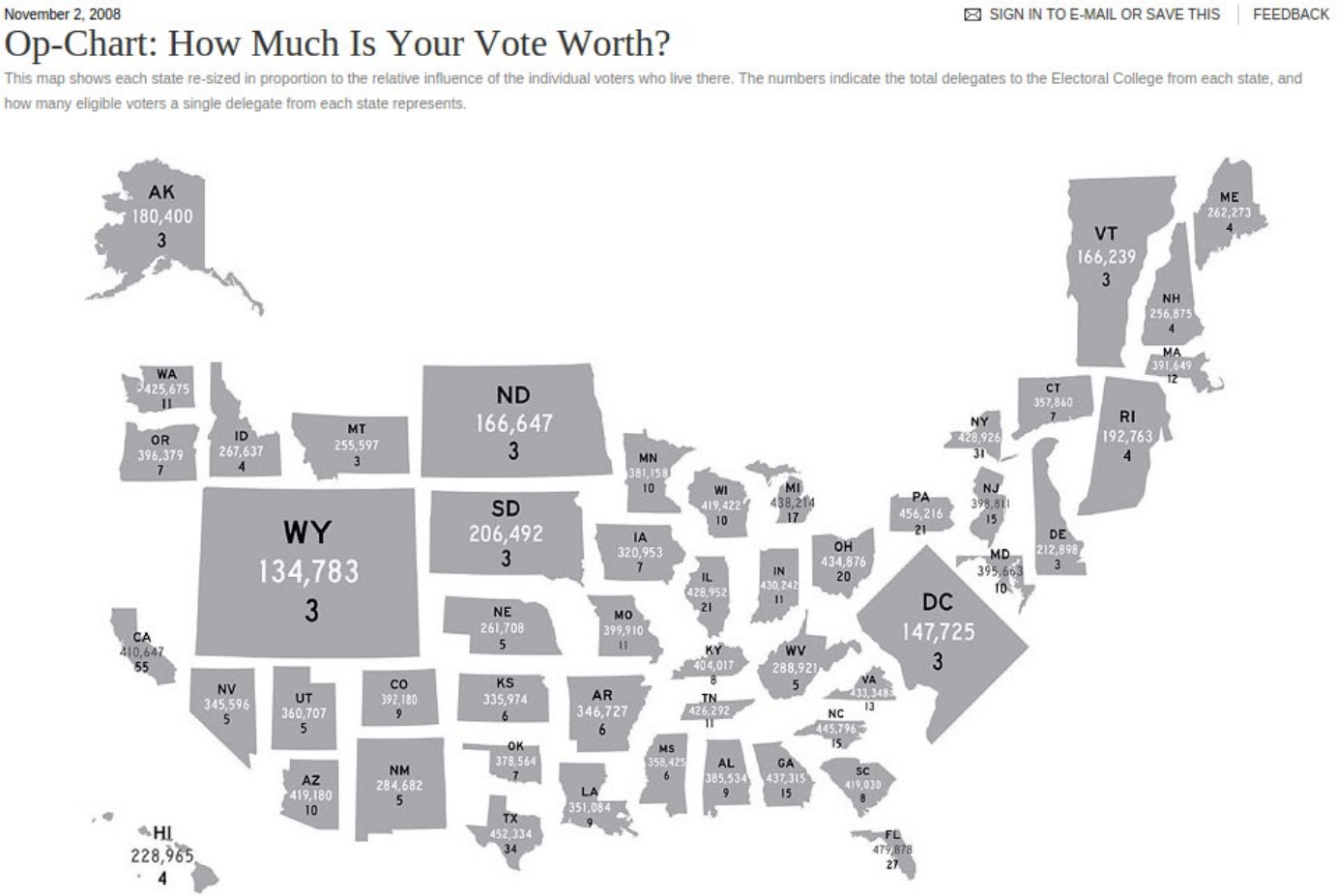
- Un problema común en la representación de grafos con un gran número de nodos es encontrar una disposición de los nodos que permita una lectura comprensiva del grafo.
- La representación de grafos mediante fuerzas supone utilizar nociones propias de **la física** para distribuir el conjunto de nodos en un determinado espacio vectorial. Se busca minimizar el número de cruces entre los arcos del grafo.
- La idea es atribuir algún tipo de fuerza entre los distintos nodos:
  - Para aquellos **nodos conectados** mediante algún arco, se asume que están unidos mediante algún resorte que obedece la ley de Hooke, y que ejerce una **fuerza de atracción** elástica entre ambos nodos.
  - Para todos los nodos del grafo se asume una **fuerza de repulsión** electrostática entre ellos. También es posible utilizar otras ecuaciones de repulsión.
- Se pueden añadir fuerzas de gravitación para ligar entre sí algunos nodos no relacionados mediante arcos.
- La disposición final del grafo es el resultado de simular el sistema físico resultante y esperar a que llegue a una **configuración de equilibrio**, lo que sucede en un mínimo de la energía total del sistema. Los arcos tienden a tener la misma longitud, y los nodos no conectados tienden a alejarse entre ellos.
- **Software**: Tulip, Graphviz, Gephi, Cytoscape

Herramientas para grafos

- D3
  https://d3js.org
- Gephi
  http://gephi.org
- Cytoscape
  http://cytoscape.org
- Network Workbench
  http://nwb.cns.iu.edu
- Sci2
  https://sci2.cns.iu.edu
- VOS viewer
  http://www.vosviewer.com
- GUESS
  http://graphexploration.cond.org
- R
- SigmaJS
  http://sigmajs.org
- Circos
  http://circos.ca

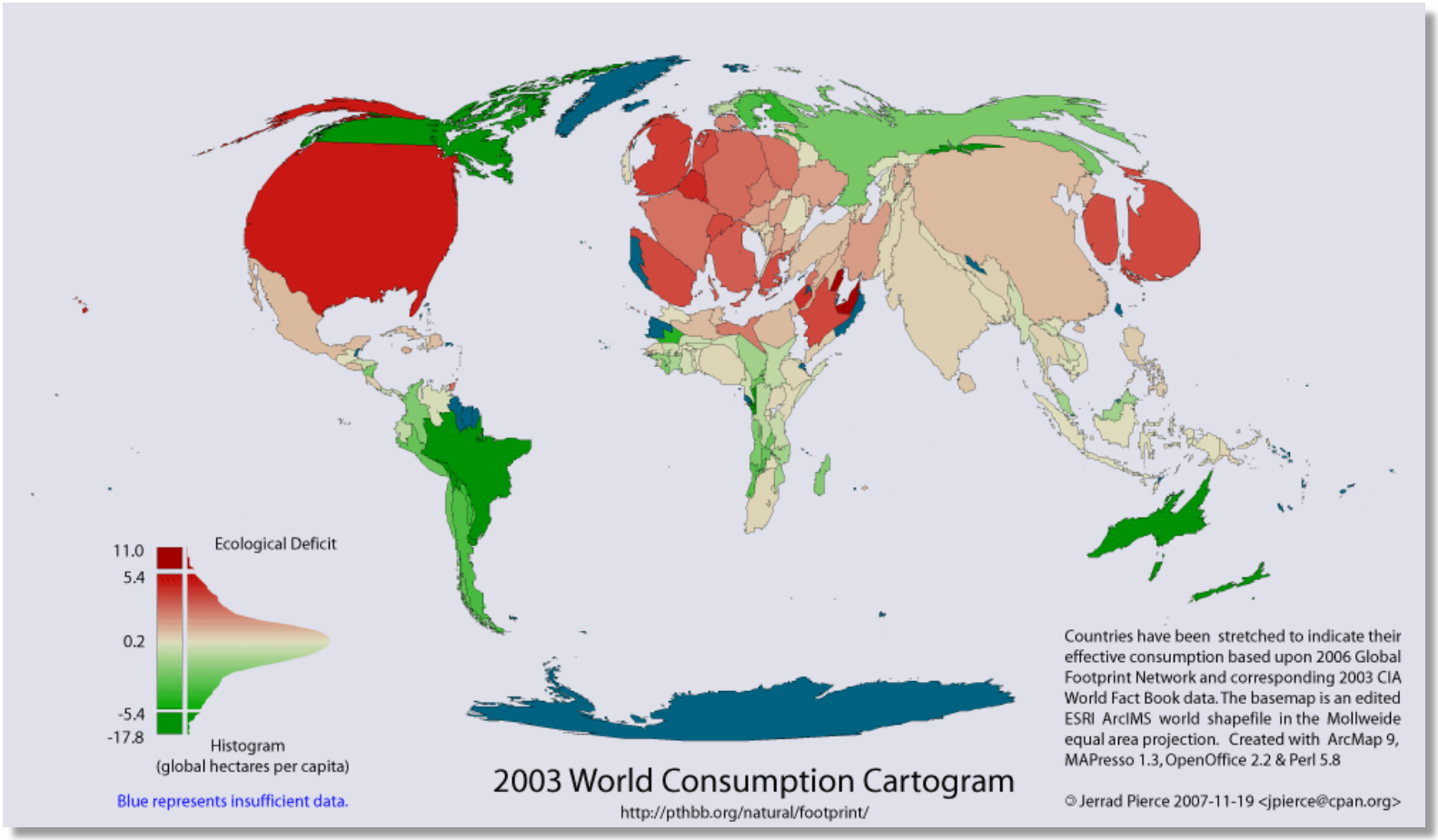
Cartogramas

- Muestran un mapa con algún esquema de representación de alguna variable relevante en términos de área o distancia. En algunos casos la representación del mapa resulta proporcionalmente distorsionada.



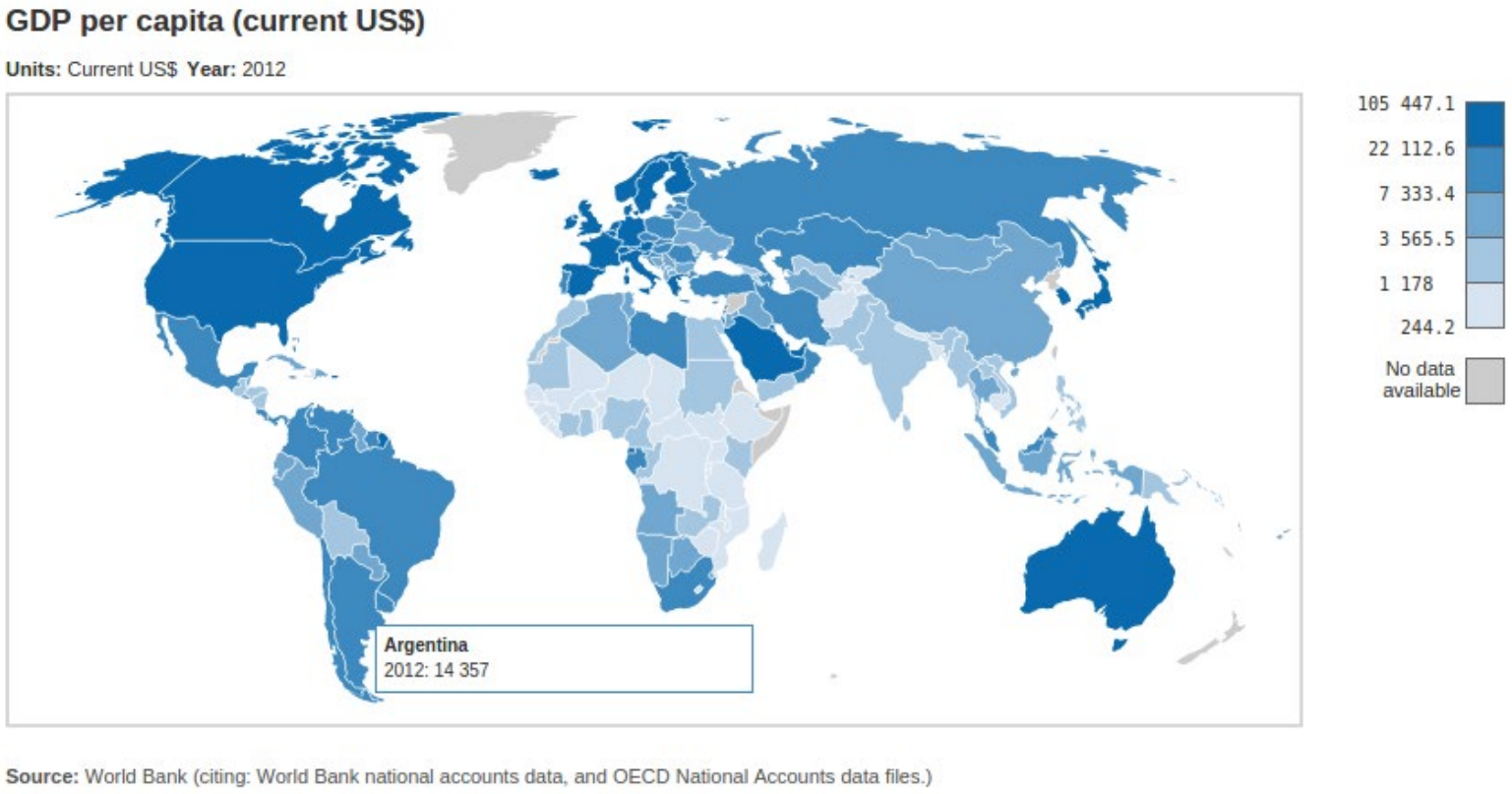
Cartogramas

- Muestran un mapa con algún esquema de representación de alguna variable relevante en términos de área o distancia. En algunos casos la representación del mapa resulta proporcionalmente distorsionada.



Mapas choropleth

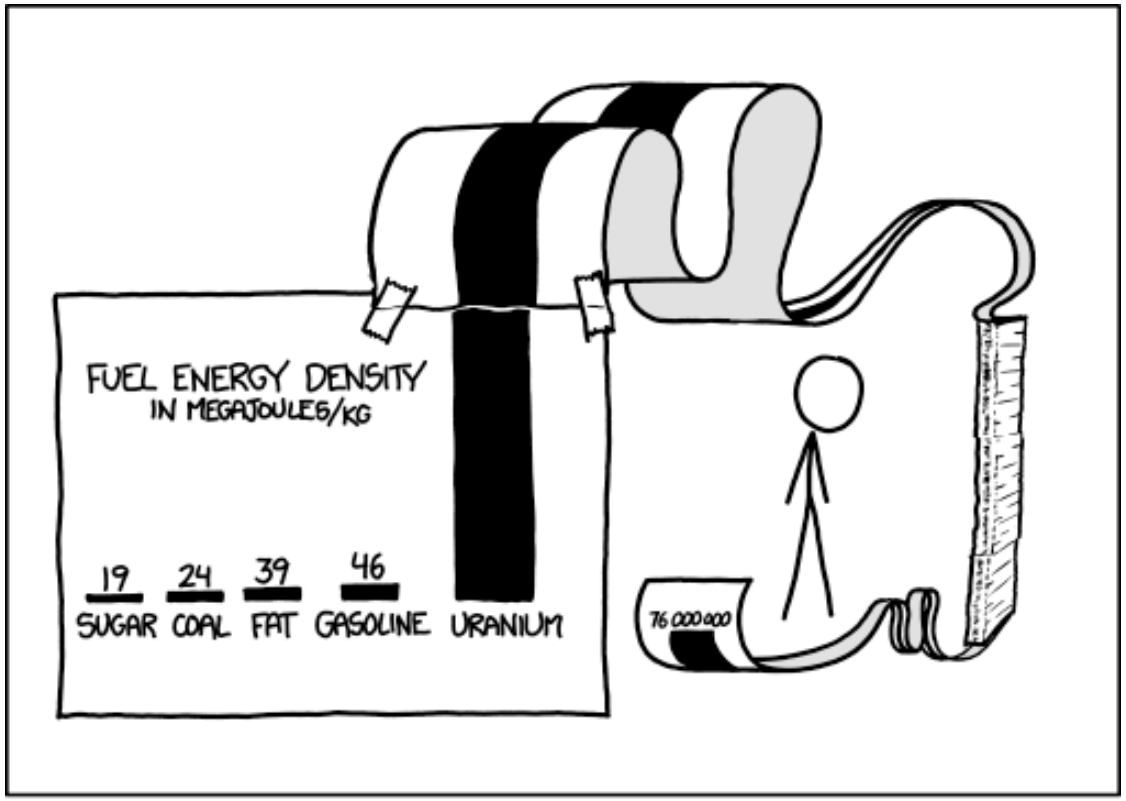
- Son cartogramas donde se colorean las regiones en función de alguna variable de interés.



Herramientas para mapas

- ArcGIS
- QGIS
- Tableau Public
- CartoDB
- Google Fusion Tables
- Google Earth
- GeoCommons
- JavaScript
  - D3
    http://d3js.org
  - Leaflet
    http://leafletjs.com
  - Kartograph
    http://kartograph.org
  - Polymaps
    http://polymaps.org
  - Google Maps API
- Muy básicos
  - Google Spreadsheets
  - BatchGeo
  - OpenHeatMap

Escala lineal vs logarítmica



SCIENCE TIP: LOG SCALES ARE FOR QUITTERS WHO CAN'T FIND ENOUGH PAPER TO MAKE THEIR POINT *PROPERLY.*

Aplicaciones software

- Python
- JMP
- Tableau
- Gephi
- D3.js
- R
- Kibana

Bibliografía

- E. Forsyth and L. Katz, A matrix approach to the analysis of sociometric data: preliminary report, *Sociometry*, 9(4): 340-347, 1946.
- J. Mackinlay, Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics*, 5(2):110-141, 1986.
- S.S. Stevens, On the theory of scales of measurement. *Science*, 103(2684), 1946.