

Caja de herramientas

Análisis visual de datos: conceptos bàsicos y herramientas

Julià Minguillón

Julià Minguillón. «Anàlisi visual de dades: conceptes bàsics i eines». *Item: revista de biblioteconomia i documentació*, [en línea], 2017, vol. 2, n.º, 63, <https://www.raco.cat/index.php/Item/article/view/337008> [Consulta: 1-07-2020].

Traducción realizada por la UOC (2020)





Análisis visual de datos: conceptos básicos y herramientas

Julià MINGUILLÓN

Estudios de Informática, Multimedia y Telecomunicación, Universitat Oberta de Catalunya
jminguillona@uoc.edu

Resumen: Vivimos en un mundo físico que se proyecta sobre otro mundo, completamente digital, que consume datos como combustible principal, de los que extrae conocimiento y lo genera. Actualmente, con estos datos se toman decisiones en todos los ámbitos, desde el personal hasta el de las grandes corporaciones multinacionales, así como en la administración y en el sector académico. El uso de herramientas de inteligencia de negocio es cada vez más habitual para apoyar la toma de decisiones, pero en muchos casos estas herramientas funcionan de forma opaca, no permiten interpretar y entender una decisión basada en los datos. Es habitual, sin embargo, que estas herramientas proporcionen visualizaciones de los datos y los procesos subyacentes, un aspecto que puede permitir entender mejor la línea de razonamiento que se sigue en el momento de tomar decisiones. En este artículo se presentan los fundamentos del análisis visual de datos y algunos antecedentes históricos destacables, y se describen distintas herramientas para llevar a cabo estos análisis, con el objetivo de aprovechar las capacidades del sistema visual humano para detectar tendencias, patrones y anomalías, hacer comparaciones y establecer relaciones.

Palabras claves: visualización de datos, análisis visual, herramientas, software.

Anàlisi visual de dades: conceptes bàsics i eines

Resum: Vivim en un món físic que es projecta sobre un altre món, completament digital, que consumeix dades com a combustible principal, n'extreu coneixement i en genera. Actualment, amb aquestes dades es prenen decisions en tots els àmbits, des del personal fins al de les grans corporacions multinacionals, així com el de l'administració i l'acadèmic. L'ús d'eines d'intel·ligència de negoci és cada cop més habitual per a donar suport a la presa de decisions, però en molts casos aquestes eines funcionen de manera opaca, no permeten interpretar i entendre una decisió basada en les dades. És habitual, però, que aquestes eines proporcionin visualitzacions de les dades i dels processos subjacents, un aspecte que pot permetre entendre millor la línia de raonament que se segueix a l'hora de prendre decisions. En aquest article es presenten els fonaments de l'anàlisi visual de dades i alguns antecedents històrics destacables, i es descriuen diferents eines per a fer aquestes anàlisis, amb l'objectiu d'aprofitar les capacitats del sistema visual humà per a detectar tendències, patrons i anomalies, fer comparacions i establir relacions.

Paraules clau: visualització de dades, anàlisi visual, eines, programari.

Visual data analysis: basic concepts and tools

Abstract: We live in a physical world overlaid on another totally digital world, whose basic fuel is data, from which we extract and generate knowledge. These data are now used to take decisions at all levels, ranging from the purely personal through to major multinationals, as well as in the administration and academia. The use of business intelligence tools in providing decision-making support is becoming increasingly more common, although in many cases these tools operate "under the hood," leaving no room for interpreting and understanding the decisions taken on the basis of such data. Nevertheless, these tools usually visualize the underlying data and processes, thus helping to give us a better understanding of the rationale applied to the decision-making process. This article addresses the basics of visual data analysis, offering important historical background, while also describing the tools involved, with the ultimate aim of leveraging the human visual system's capacity to detect trends, patterns or anomalies, draw comparisons and establish relationships.

Keywords: data visualisation, visual analysis, tools, software

Introducción

Los seres humanos somos principalmente visuales. El sistema de visión humano es una máquina sofisticada que permite capturar una gran cantidad de información del entorno y usarla tanto para evaluarlo como para emprender acciones. De hecho, se cree que el 90 % de la información transmitida al cerebro es visual y se estima que las imágenes se procesan 60.000 veces más rápido que el texto escrito, aunque en realidad este hecho no ha sido aún contrastado científicamente.¹ El dicho popular «Una imagen vale más que mil palabras» resume a la perfección esta idea. Sea como fuere, los expertos en marketing saben bien que una imagen capta más la atención que un texto, y aprovechan dicho conocimiento para transmitir ideas y hechos de manera efectiva y eficaz.²

Cada vez es más habitual el uso de imágenes y vídeos para transmitir información mostrando ideas y hechos en lugar de detallarlos de manera textual, y no solo en el ámbito de la publicidad, sino también en cualquier contexto donde se usen datos para tomar decisiones. Conviene aclarar que actualmente hay dos conceptos que a veces se usan indistintamente y no son del todo equivalentes, sino más bien complementarios. Nos referimos al concepto de *visualización de datos* y a otro que se ha popularizado más recientemente denominado *infografía*. Los dos usan representaciones gráficas para presentar los datos y relatar las historias que hay detrás de estos, pero con objetivos y procedimientos diferentes.

Se puede definir *infografía* como una representación más visual que los mismos textos, en la cual intervienen descripciones, narraciones o interpretaciones, presentadas de manera gráfica, normalmente figurativa, que pueden coincidir o no con grafismos abstractos o sonidos. La infografía nace como un medio para transmitir información gráficamente, que dispone de un método para representar la información de manera icónica y textualmente, de

forma que el usuario la pueda comprender sin dificultad. En el proceso de creación de una infografía, que se acostumbra a hacer empleando herramientas informáticas, se recoge un hecho complejo y se explica de manera sencilla para que se pueda interpretar con un simple vistazo.

En comparación con una infografía, la visualización de datos, también llamada *visualización de la información* (ya que pone los datos en su contexto), es el estudio de la representación visual de datos abstractos (y quizás interactivos) para reforzar la cognición humana, que incluye tanto datos numéricos y no numéricos como texto o información geográfica, por ejemplo. Es decir, podemos deducir que prácticamente no hay muchas diferencias sustanciales entre los dos conceptos, puesto que existe una naturaleza común entre una infografía y una visualización, tal como explica muy bien Alberto Cairo,³ quien indica acertadamente las sutiles diferencias conceptuales entre ambos conceptos:

Algunos especialistas marcan una frontera entre las dos disciplinas basada en el hecho de que supuestamente la infografía consiste en presentar información por medio de gráficos estadísticos, mapas y esquemas (exposición), mientras que la visualización se basa en la creación de herramientas visuales (estáticas o interactivas) que un público pueda usar para explorar, analizar y estudiar conjuntos complejos de datos. Pero pertenecen a un mismo continuo en el que cada una de ellas ocupa extremos opuestos de una misma línea. Y esta es paralela a otra cuyos límites son definidos por las palabras presentación y exploración.

En este artículo nos centraremos en las visualizaciones de datos como mecanismos para extraer conocimiento, aprovechando las capacidades del sistema visual humano y las posibilidades que ofrece la misma visualización para la manipulación de los datos, y añadiendo un cierto grado de interactividad.

1. Jonathan Schwabish, «The 60,000 fallacy», En: *PolicyViz* [en línea]: *helping you do a better job processing, analyzing, sharing, and presenting your data*, September 17, 2015, <<https://policyviz.com/2015/09/17/the-60000-fallacy/>>, [Consulta: 4 nov. 2017].
2. Michel Wedel; Rik Pieters (ed.), *Visual marketing: From attention to action*. New York: Psychology Press, 2012.
3. Alberto Cairo, *El arte funcional: infografía y visualización de información*. Madrid: Alamut, 2011.

1. Antecedentes históricos

La visualización de datos como mecanismo de narración de historias ha estado presente en el desarrollo de nuestra sociedad a lo largo del tiempo. Los seres humanos han dibujado imágenes para comunicarse desde hace miles de años, desde pictogramas en las paredes de una cueva rupestre y los jeroglíficos egipcios, hasta ideogramas y toda la iconografía moderna. El ser humano siempre ha usado imágenes para comunicar y explicar historias, porque, dado el tipo de evolución que ha experimentado su cerebro, es más eficiente procesando información mediante el sistema visual.

Una de las mejores compilaciones sobre los antecedentes de la visualización de datos tal como la entendemos hoy en día es el trabajo de Michael Friendly.⁴ Se publicó en un capítulo de un manual de visualización de datos que formaba parte de una colección de libros de estadística, lo cual muestra la importancia de la visualización como herramienta para el análisis de datos. El artículo de Friendly está estructurado en una línea temporal, que incluye desde las primeras visualizaciones (mapas y diagramas) anteriores al siglo XVII hasta la actualidad (a partir de 1975), donde la tecnología ha hecho posible la creación masiva de visualizaciones. Una de las etapas más interesantes destacadas por Friendly es la segunda mitad del siglo XIX, cuando se desarrollaron numerosas técnicas para el análisis estadístico que se aplicaban en todos los ámbitos de la planificación social, la industrialización, el comercio y el transporte. Esto provocó la aparición de muchas innovaciones en la visualización de datos, necesarias para poder explicar los datos y los complejos fenómenos de la sociedad del momento. Un primer paso fue la proyección de elementos en tres dimensiones (3D) como vía de escape del plano, que hasta entonces había limitado las posibilidades. Otro ejemplo interesante fue la combinación de mapas con datos de cada región, de forma que en una misma representación se combinaban datos espaciales con otros temporales. Finalmente, el uso de gráficos para el análisis estadístico (por ejemplo, la co-

rrelación era un concepto todavía en desarrollo) permitió que Francis Galton y otros investigadores avanzaran en la formalización de las observaciones y las convirtieran en técnicas estadísticas.

Un célebre caso que tuvo relevancia en el siglo XIX es el del médico inglés John Snow, quien bien podría ser destacado como uno de los iniciadores de la visualización moderna de datos. El mapa de Snow (figura 1) es considerado uno de los primeros ejemplos de uso de un método geográfico para describir y localizar los casos de una epidemia, así como su origen más probable, cosa que nunca se había hecho hasta ese momento. Esto permitió establecer los mecanismos de transmisión de las enfermedades infecciosas. Es un ejemplo de cómo una visualización de datos puede usarse para probar o refutar una hipótesis, en este caso la transmisión del cólera a través del agua, y convertir hechos en datos que ayudan a tomar una decisión. De esta manera, John Snow demostró que la causa de los casos de cólera en Londres era el consumo de agua contaminada con materiales fecales. El 1854 cartografió en un plano del Soho los pozos de agua y los casos de cólera, y así localizó el pozo donde se originó la transmisión de la enfermedad, situado en Broad Street, el centro de la epidemia, por lo que recomendó clausurar la bomba de agua que lo alimentaba. Así consiguió disminuir la proliferación de los casos de cólera en esta zona de Londres.

En la actualidad podríamos encontrar un estudio equivalente en diferentes análisis que se han hecho de fenómenos como AirBnB, para ver cómo nuevos actores pueden distorsionar y generar tensiones en un mercado tradicional, como es el caso de los pisos turísticos con licencia, o sin ella, y el incremento del alquiler de viviendas habitualmente dedicadas al alquiler de larga duración. En la mayoría de las ciudades, los alquileres privados de corta duración sin licencia turística son ilegales, especialmente por las diferencias con el modelo tradicional, en donde el sector se encuentra muy regulado y sometido a una gran cantidad de tasas e impuestos. Un ejemplo de este

4. Michael Friendly, «A brief history of data visualization». En: Chun-houh Chen; Wolfgang Härdle; Antony Unwin (ed.), *Handbook of data visualization*. Berlin; Heidelberg: Springer, 2008, p. 15-56.

crecimiento descontrolado se puede ver en una visualización creada por Kor Dwarshuis,⁵ en donde los datos que publica AirBnB se visualizan combinados en un mapa que geolocaliza tanto las ofertas de pisos como los alquileres, conjuntamente con un eje temporal que permite ver la diseminación de la oferta y la demanda a lo largo del tiempo, un aspecto que John Snow no pudo reflejar en su mapa por la tecnología del momento.

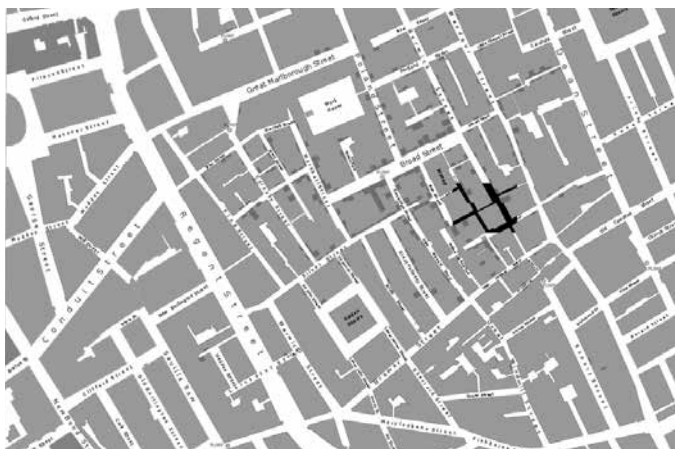


Figura 1. Mapa de la localización de los brotes de cólera en el Soho de Londres, por John Snow (1854). <https://ca.wikipedia.org/wiki/John_Snow#/media/File:Dr._John_Snow_Cholera_Map.svg>.

Estos ejemplos muestran que una visualización de los datos, en este caso superpuesta en un mapa, permite hacerse una idea rápida de la magnitud del fenómeno, así como facilitar la detección de elementos que por una razón u otra destacan sobre el resto, aportando conocimiento sobre el problema que se quiere resolver. Esta idea es la que proporciona soporte a lo que llamamos «análisis visual» (*visual analytics*), que se desarrolla a continuación.

2. Análisis visual

La utilización de visualizaciones de datos como mecanismo para analizarlos se fundamenta en las propiedades del sistema visual humano, que se define como una parte del sistema nervioso central que proporciona a los organismos vivos (en general) la habilidad de procesar visualmente, detectando e interpretando la luz visible, para entender el escenario que les rodea, creando lo que se entiende como *percepción*. Es importante destacar que el sistema visual humano lo forman no solo los ojos, sino también los nervios ópticos y áreas específicas del cerebro (el córtex visual), que combinan diferentes niveles de abstracción. De hecho, no vemos con los ojos, sino con el cerebro, puesto que la percepción es una combinación de diferentes procesos que realizan tareas diversas relacionadas con la visión. Tal como describen Card, MacKinlay y Shneiderman,⁶ la visualización de datos proporciona un procedimiento muy potente para permitir a los usuarios detectar e interpretar patrones en los datos.

El proceso de percepción permite a los humanos hacer tareas de manera eficiente, como, por ejemplo, discriminar colores, separar mediante el contraste, estimar distancias y medidas, determinar orientaciones y ángulos, y reconstruir el movimiento de los objetos que forman la escena, cosa que nos permite la navegación en el mundo físico tridimensional. Tal como describe Wandell,⁷ las propiedades del sistema visual humano y la codificación que resulta de la luz captada por la retina para realizar todas estas acciones también tienen implicaciones en el diseño de instrumentos que muestran información de manera visual, como, por ejemplo, una visualización de datos. El proceso de percepción de los humanos todavía no ha sido igualado por ninguna máquina o proceso artificial, a pesar de que es una línea de investigación muy activa desde hace ya unos cuantos años, la cual se ha visto impulsada últimamente por su auge en ámbitos como la inteligencia artificial y lo que se conoce como «aprendizaje profun-

5. <<https://dwarshuis.com/various/airbnb/barcelona/>>

6. Stuart K. Card, Jock MacKinlay, Ben Shneiderman (ed.), *Readings in information visualization: using vision to think*, San Diego; London; San Francisco: Academic Press / Morgan Kaufmann, 1999.

7. Brian A. Wandell, *Foundations of vision*, Sunderland: Sinauer Associates, 1995.

El proceso de percepción permite a los humanos hacer tareas de manera eficiente, como, por ejemplo, discriminar colores, separar mediante el contraste, estimar distancias y medidas, determinar orientaciones y ángulos, y reconstruir el movimiento de los objetos que forman la escena, lo cual nos permite la navegación en el mundo físico tridimensional.

do»,⁸ que aprovecha la disponibilidad de gran cantidad de datos y la elevada capacidad computacional necesaria para procesarlos.

Pero hay que tener en cuenta que el sistema visual humano, pese a su potencia y complejidad, también puede ser engañado fácilmente mediante ilusiones ópticas,⁹ poniendo en evidencia deficiencias que se deberían evitar para no caer en una trampa a la hora de mostrar datos de manera gráfica. Por ejemplo, una perspectiva mal usada puede distorsionar las medidas de los elementos que se comparan. Un uso incorrecto de las saturaciones y el contraste puede hacer ver áreas de colores diferentes como si fueran similares. Además, también hay que tener en cuenta que un porcentaje nada desdeñable de la población mundial tiene algún tipo de deficiencia visual en cuanto al procesamiento del color, de forma que necesitan el uso de codificaciones alternativas (por ejemplo, escala de grises y formas), o bien los casos en los que la visualización tiene por objeto su impresión en blanco y negro. Finalmente, también cabe tener presentes otros aspectos culturales que pueden determinar la manera en

que se interpreta una visualización de datos, puesto que la ordenación espacial de los elementos que la componen y su comprensión pueden depender de si el observador la lee de izquierda a derecha y de arriba abajo, o bien al contrario. Es decir, hay que tener en cuenta todos estos aspectos a la hora de diseñar una buena visualización de datos que permita hacer un análisis visual preliminar sin interferencias causadas por una mala decisión de codificación, y aprovechar la experiencia de estudios del ámbito de la publicidad, donde, por ejemplo, se ha demostrado que el color afecta directamente a la percepción del consumidor y que hay diferencias culturales respecto a su percepción.¹⁰ O por poner otro ejemplo, en el sector del periodismo, donde los gráficos se utilizan para destacar o apoyar una idea, ha habido autores que han identificado los aspectos básicos que determinan si una visualización de datos es una buena o una mala práctica.¹¹

Sin embargo, en sus inicios, el desarrollo del análisis visual de datos no tuvo en cuenta los aspectos estéticos, sino que se limitó a reducir los datos a un conjunto de medidas que resumen las principales características, utilizando gráficos primitivos y sencillos para representarlas. Tal como describe Friendly¹² en su proyecto Milestones, el desarrollo del uso de representaciones gráficas para la descripción de datos cuantitativos estuvo ligado a la necesidad de resumir un concepto medible, normalmente con el objetivo de entenderlo mejor.

John Tukey estableció los principios del análisis de datos, definiéndolo como el conjunto de procedimientos para analizar datos, las técnicas para interpretar los resultados, las maneras de planificar la captura de datos para facilitar su análisis y hacerlo más preciso y cuidadoso, y todo el hardware y las estadísticas que hay que aplicar para ana-

8. Yoshua Bengio, «Learning deep architectures for AI», *Foundations and trends® in Machine Learning*, vol. 2, n.º 1 (2009), pp. 1-127.

9. Richard L. Gregory, «Knowledge in perception and illusion», *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 352, n.º 1358 (1997), pp. 1121-1127.

10. Thomas J. Madden; Kelly Hewett; Martin S. Roth, «Managing images in different cultures: a cross-national study of color meanings and preferences», *Journal of International Marketing*, vol. 8, n.º 4 (2000), pp. 90-107.

11. Dona M. Wong, *The Wall Street Journal Guide to information graphics: the dos and don'ts of presenting data, facts, and figures*, New York: Norton, 2010.

12. Michael Friendly, «Milestones in the history of data visualization: a case study in statistical historiography», En: Annual Conference of the Gesellschaft für Klassifikation (28a: 2004: Dortmund), *Classification: the ubiquitous challenge: proceedings of the 28th Annual Conference of the Gesellschaft für Klassifikation e. v., University of Dortmund, March 9-11, 2004*, Claus Weihs, Wolfgang Gaul (ed.), New York: Springer, 2005, pp. 34-52.

lizar los datos.¹³ Posteriormente, el mismo Tukey¹⁴ desarrolló el concepto de análisis exploratorio de datos, cuyos objetivos son sugerir hipótesis sobre las causas de un fenómeno observado, evaluar asunciones en las que hay que fundamentar la inferencia estadística, dar soporte a la selección de las herramientas y técnicas estadísticas más apropiadas y, finalmente, proporcionar una base para una recogida adicional de datos a través de encuestas o experimentos. Una primera visualización puede servir para conocer la naturaleza del problema y reducir el abanico de posibles aproximaciones para resolverlo.

Un ejemplo de la necesidad de representar gráficamente los datos para entender la naturaleza es el llamado *cuarteto de Anscombe*, creado por Francis Anscombe en 1973, que comprende cuatro conjuntos de datos que tienen las mismas propiedades estadísticas, pero que evidentemente son diferentes cuando se inspeccionan los gráficos respectivos, mostrando las limitaciones del uso de descriptores estadísticos para resumir un conjunto de datos (figura 2). Concretamente, se trata de cuatro conjuntos de datos de once puntos en un plano (x, y), de forma que la media y la varianza de cada variable, así como la correlación entre ambas y el coeficiente de la recta de regresión óptima, son idénticos para los cuatro conjuntos (o se podría pensar que son idénticos), mientras que son claramente diferenciables si se utiliza una representación visual. Obviamente, cada conjunto representa el resultado de cuatro procesos diferentes que podrían haberlo generado. Así, se puede identificar una colección de datos típica (figura superior izquierda), unos datos que siguen una relación no lineal pero bien clara (figura superior derecha), unos datos que siguen una relación lineal a excepción de uno, identificando así un posible dato atípico (figura inferior izquierda), y, finalmente, unos datos que muestran una relación no lineal entre las dos variables, en la que un simple dato atípico genera un coeficiente de correlación elevado. Sin la visualización de estos da-

tos usando un simple gráfico de dispersión (x, y) es muy difícil hacerse a la idea de las cuatro distribuciones subyacentes en cada caso.

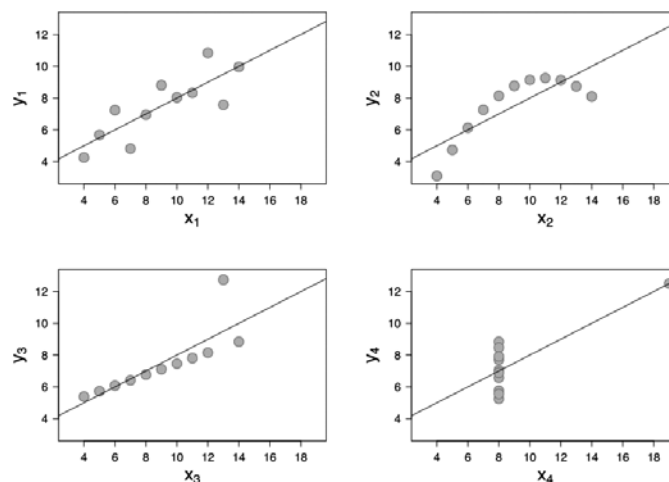


Figura 2. El cuarteto de Anscombe
<https://en.wikipedia.org/wiki/Anscombe%27s_quartet#/media/File:Anscombe%27s_quartet_3.svg>.

Aunque se trata de un ejemplo sintético, muestra de manera convincente las limitaciones de los descriptores estadísticos más habituales en los trabajos de investigación y las posibilidades de la visualización como herramienta de análisis visual complementario. Recientemente, Matejka y Fitzmaurice¹⁵ han generado hasta una docena de conjuntos de datos diferentes¹⁶ que muestran los mismos descriptores estadísticos como ejemplo de la necesidad de usar el análisis visual para entender mejor los datos.

En la misma línea de razonamiento, este reduccionismo (en el sentido de reducir un conjunto de datos a un número reducido de descriptores estadísticos) es descrito

13. John W. Tukey, «The future of data analysis». *The Annals of Mathematical Statistics*, vol. 33, n.º 1 (1962), pp. 1-67.

14. John W. Tukey, *Exploratory data analysis*, London: Sage, 1977.

15. Justin Matejka; George Fitzmaurice, «Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing», En: ACM CHI Conference on Human Factors in Computing Systems (2017: Denver), *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, [Denver: ACM, 2017], pp. 1290-1294.

16. <<https://www.autodeskresearch.com/publications/samestats>>

y criticado por Manovich,¹⁷ que apuesta por mostrar los datos originales a los usuarios finales, total o parcialmente, a fin de permitirles formarse una idea de su naturaleza, especialmente en datos que conllevan una carga semántica importante (por ejemplo, imágenes y mapas). Según Manovich, desde la segunda mitad del siglo XVIII hasta hoy, ha habido dos principios clave que han dado forma a la visualización de información. El primero es el principio de reducción, que consiste en el uso de gráficas primitivas (puntos, líneas, formas geométricas simples, etc.) para la representación de los elementos y sus relaciones, que revelan patrones y estructuras subyacentes sin necesidad de visualizar los datos originales. Esto ha comportado una pérdida de importancia de los datos en cuanto a las representaciones, demasiado esquemáticas en algunos casos. El descriptor más sencillo es la media, acompañado habitualmente de la varianza, que indica hasta qué punto los datos están centrados en torno a la media, resumiendo todo un conjunto de datos a un valor o dos. El paso siguiente es usar diagramas de caja para describir los cuartiles, mostrando la distribución de los datos y la existencia de posibles datos atípicos. Actualmente se utilizan los diagramas de violín, que integran el histograma de la distribución subyacente como parte de la visualización y añaden información sobre la distribución real de los datos (figura 3), aunque son una simplificación de la naturaleza del conjunto de datos.

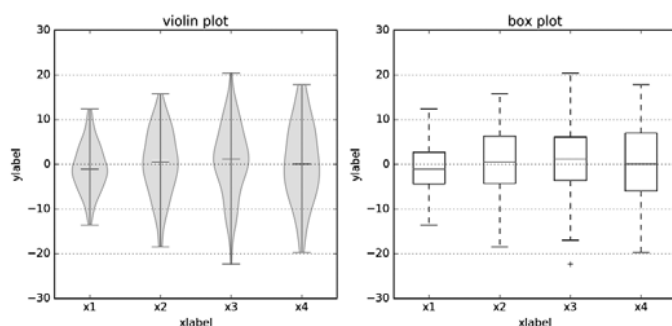


Figura 3. Uso del diagrama de violín (izquierda) como evolución de los diagramas de caja (derecha) <https://matplotlib.org/1.5.1/examples/statistics/boxplot_vs_violin_demo.html>.

Este reduccionismo, presente en todos los ámbitos de las ciencias, propone que el mundo puede analizarse a partir de los elementos simples que lo componen y de las reglas que rigen las interacciones, de forma que se pueda comprender la totalidad mediante una descripción simplificada o reducida. Así, durante el siglo XIX se desarrollaron todos los gráficos típicos para representar estos datos «reducidos» que permiten explicar aspectos sociales, demográficos, etc. De hecho, fue en esta época cuando aparecieron los gráficos de barras y de pastel, los histogramas, etc., todos conceptualizados desde esta visión reduccionista y usando los mismos elementos gráficos sencillos (puntos, líneas, cajas, etc.).

El segundo principio que menciona Manovich es el uso de variables espaciales (posición, tamaño, forma, etc.) para representar diferencias en los datos y revelar, así, los patrones y las relaciones existentes más importantes. En el ejemplo (ficticio) de la figura 3, se pueden observar las diferencias entre cuatro conjuntos de datos diferentes en cuanto a la distribución de una variable en cada conjunto. Manovich señala que la visualización de información privilegia las dimensiones espaciales sobre las otras y da más importancia a la topología y a la geometría, y menos a otros aspectos, como, por ejemplo, el color, la saturación y la transparencia. Así, para representar un conjunto de datos, las dimensiones más importantes se asignan a la disposición espacial (llamada *layout*), mientras que el resto de dimensiones se mapean habitualmente en el resto de las variables visuales (color, etc.). En este caso, el color o la forma se usan solo para dividir los elementos de un conjunto de datos en diferentes clases. Manovich menciona como posible razón la dificultad de reproducir representaciones gráficas mediante la tecnología que hay en cada momento, lo cual limita el uso del color, la transparencia, etc. Los ordenadores han permitido crear y manipular representaciones más complejas, potenciando el uso de otras dimensiones visuales.

Sea como fuere, mediante el uso de los instrumentos típicos de la estadística descriptiva, o bien mostrando los datos directamente, las visualizaciones de datos puede

17. Lev Manovich, «What is visualization?», *Poetess Archive Journal*, vol. 2, n.º 1 (2010), 32 p.

tener diferentes objetivos, todos relacionados con el análisis exploratorio descrito por Tukey:

- **Descriptivo:** resume de manera gráfica las propiedades de un conjunto de datos, mediante sus descriptores estadísticos básicos, como la media, la varianza, los cuartiles, los histogramas, etc.
- **Comparativo:** combina los elementos descritos en el punto anterior para mostrar parecidos o diferencias entre uno o más conjuntos de datos o de acuerdo con una o más variables.
- **Detección de tendencias:** habitualmente, usa un eje temporal que permite ver rápidamente si un fenómeno crece y/o decrece regularmente.
- **Detección de patrones:** la representación que se elige permite detectar agrupaciones o repeticiones en los datos en una, dos o, más raramente, tres dimensiones.
- **Detección de anomalías** (datos atípicos): identifica elementos que destacan claramente del resto por su posición o medida, entre otros atributos posibles.
- **Detección de correlaciones:** muestra el grado de relación entre dos variables. Hay que tener siempre presente que correlación y causalidad son dos conceptos claramente diferentes, y que de ninguna forma el primero implica el segundo, lo cual puede resultar confuso en una visualización de datos.¹⁸
- **Relacional:** da más énfasis a las relaciones entre datos que a los datos en sí, usa representaciones en las que la posición relativa de los elementos es más importante que la absoluta, mientras que los atributos se mapean sobre las características de la representación (forma y color, principalmente).
- **Jerárquico:** muestra una estructura de relaciones en la que los elementos se agrupan de acuerdo con una taxonomía y se pueden visualizar diferentes niveles de detalle.
- **Localizado:** superpone los datos o sus atributos en forma de capas sobre un mapa o esquema que aporta una semántica mucho más rica en forma de distancias o posiciones, tanto relativas como absolutas.

Como era de esperar, no existe una visualización de datos universal que permita conocer rápidamente su naturaleza, básicamente por dos razones: por un lado, la variedad de datos (especialmente en cuanto al gran número de atributos que pueden usarse para describir cada elemento del conjunto de datos) y, por el otro, el objetivo que se pretende lograr combinando uno o más de los puntos mencionados. Habitualmente, se necesita más de una exploración antes de poder formarse una idea lo suficiente clara de un conjunto de datos. En general, los datos son complejos y pueden combinar diversos aspectos al mismo tiempo, entre los que se pueden destacar los siguientes: son multidimensionales; van ligados a restricciones espaciotemporales, longitudinales (que evolucionan en el tiempo) o multimodales (combinan diferentes fuentes y orígenes), y provienen de la ejecución de múltiples procesos paralelos o modelos. Visualizar datos incluye la gestión de toda esta complejidad para convertirlos en información, es decir, obtener respuestas a las preguntas o a los objetivos de la visualización que se pretende crear. De nuevo, el análisis visual no sustituye a la estadística clásica o a la construcción de modelos de minería de datos, sino que aporta una perspectiva diferente basada en las capacidades del sistema visual humano, incorporando al objetivo de la visualización la posibilidad de hacer diferentes operaciones con los datos de manera más intuitiva.

En este sentido, la evolución del ámbito de la visualización de datos no se ha centrado solo en la capacidad de generar gráficos complejos con más resolución en un breve lapso de tiempo, sino que ha ido incorporando elementos interactivos en la misma visualización en forma de operaciones básicas (selección, filtraje, etc.). De acuerdo con el trabajo de Keim *et al.*,¹⁹ el análisis visual de datos se fundamenta en un mantra que es una versión modificada de lo que propuso Ben Shneiderman en 1996:

Analyse First. Show the Important. Zoom, Filter and Analyse Further. Details on Demand.

18. <<http://www.tylervigen.com/spurious-correlations>>

19. Daniel Keim et. al., «Visual analytics: Definition, process, and challenges», En: *Information visualization: human-centered issues and perspectives*, Berlin; Heidelberg: Springer, 2008, pp. 154-175.

El proceso de análisis visual consiste en un ciclo continuo que se inicia en los datos y sus posibles transformaciones, y que se bifurca en dos aproximaciones complementarias. Entre la visualización y la construcción de modelos hay un diálogo con el objetivo de extraer conocimiento que pueda usarse para iterar el proceso de análisis visual con más detalle o complejidad (figura 4). La capacidad de interacción tiene que permitir que el usuario de la visualización pueda, al menos, hacer las operaciones básicas definidas por Shneiderman (vista general, zoom, filtro y selección).

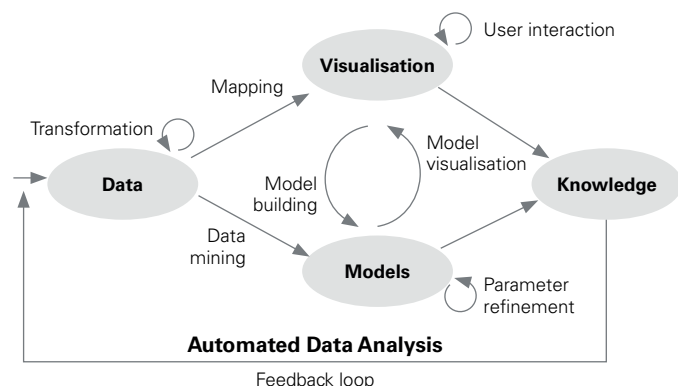


Figura 4. Proceso de análisis visual (Keim *et al.*, 2008).

Desde una perspectiva de análisis visual, las dos primeras etapas definidas en la figura 4 son la transformación (o adaptación) y la visualización de los datos, incluyendo la posible interacción para realizar operaciones con los datos. Por lo tanto, una vez establecido el objetivo del análisis visual de los datos, se trata de seleccionar un tipo de visualización más o menos interactiva que permita hacer la exploración preliminar. En una primera iteración, los datos pueden mostrarse tal como aparecen, sin ninguna transformación. El observador determina la visualización a partir del conocimiento que extrae, por ejemplo, visualizando el resultado de aplicar un análisis de componentes

principales si los datos muestran una cierta estructura entre variables que puede ser explotada.

Así, la construcción de una visualización de datos es un proceso que integra varios puntos de vista, desde el más cercano a la naturaleza de los datos, que involucra el uso de descriptores estadísticos y modelos de minería de datos para extraer conocimiento, hasta aspectos más ligados a la percepción del observador final, que incluye tanto elementos estéticos como otros culturales. En el próximo apartado se describen cuatro familias o categorías de herramientas para la creación de visualizaciones de datos en función de dos dimensiones: por un lado, el nivel de abstracción de la representación gráfica y, por el otro, el grado de interactividad permitido que determina la exploración posterior.

3. Herramientas para el análisis visual

En la actualidad hay muchas opciones para visualizar datos, sean del tipo que sean. De hecho, esta gran disponibilidad de herramientas y recursos ha sido, en parte, la causa de la popularización de las visualizaciones de datos en todos los ámbitos, no solo el científico.

Para clasificarlos se ha optado por agruparlos de acuerdo con las dos dimensiones mencionadas. La primera dimensión, el nivel de abstracción, hace referencia a la granularidad del tipo de objeto que se manipula para crear la visualización, y puede ir desde un píxel hasta paneles de control que combinan múltiples visualizaciones. La segunda dimensión, el grado de interactividad, hace referencia a las opciones que tiene el observador para manipular datos mediante la visualización, y va desde visualizaciones estáticas que no permiten ninguna interacción, hasta interfaces complejas que incluyen todas las operaciones básicas definidas por Shneiderman.²⁰ Algunas herramientas mencionadas en este apartado aparecen en el estudio de Cota *et al.*²¹ desde una perspectiva orientada a la visualización de grandes volúmenes de datos.

20. Ben Shneiderman, «The eyes have it: a task by data type taxonomy for information visualizations», En: IEEE Symposium on Visual Languages (1996: Washington). *VL'96: proceedings of the 1996 IEEE Symposium on Visual Languages*. Washington: IEEE Computer Society, 1996, pp. 336-343.

21. Manuel Pérez Cota *et al.* «Analysis of Current Visualization Techniques and Main Challenges for the Future». *Journal of Information Systems Engineering & Management*, vol. 2, n.º 3 (2017), art. 19.

El proceso de análisis visual consiste en un ciclo continuo que se inicia en los datos y sus posibles transformaciones, y se bifurca en dos aproximaciones complementarias.

3.1 Calculadoras gráficas

En esta primera categoría podemos encontrar herramientas de propósito general que permiten visualizar datos de manera sencilla, con gráficos predeterminados que, supuestamente, permiten medir o comparar diferentes valores entre sí, de acuerdo con el valor de uno o más atributos o entre conjuntos de datos diferentes, ligados a los descriptores estadísticos más habituales. Se trata, de hecho, de software con una clara orientación a la manipulación de datos con el objetivo de extraer conocimiento, sobre todo en forma de gráficos sencillos, aunque en algunos casos incluyen otras capacidades más avanzadas. Básicamente, estas herramientas convierten un conjunto de datos en una visualización predeterminada por el usuario, normalmente estática y sin ninguna interacción, con una capacidad de exploración limitada.

Si los datos se encuentran en forma tabular, la primera herramienta que se puede utilizar para visualizarlos es una hoja de cálculo, como Microsoft Excel, o también herramientas de software libre, como LibreOffice Calc,²² que mejora la importación de datos en formato .csv con un asistente para ajustar la codificación y estructura de los datos. Históricamente, las hojas de cálculo han sido las herramientas más usadas para generar los típicos gráficos que resumen los datos, como por ejemplo los gráficos de líneas, de barras o de pastel. La primera hoja de cálculo fue VisiCalc, desarrollada por la Apple II en 1979, seguida por la Lotus 1-2-3 para el IBM PC, en 1983, con mejores capacidades gráficas. Las dos se consideran las primeras verdaderas aplicaciones rupturistas (*killer-apps*) que impulsaron la venta de ordenadores personales. Las

herramientas actuales proporcionan una serie de opciones que convierten un conjunto de datos en un tipo de gráfico concreto, incluyendo la posibilidad de personalizar el aspecto mediante el uso de colores, tramas, opciones 2D y 3D, etc.

Otra opción más sencilla es usar un software específico para generar gráficos, como Gnuplot,²³ que permite visualizar datos de acuerdo con un repertorio de representaciones ligadas a descriptores estadísticos, como por ejemplo histogramas o diagramas de caja, entre otros. Gnuplot se orienta a la visualización de datos de una, dos o tres dimensiones, y se ha convertido en un motor gráfico que puede usarse desde diferentes lenguajes de programación y entornos gráficos, mediante un sencillo lenguaje script que genera gráficos a partir de datos y comandos sencillos.



Figura 5. Galería de visualizaciones creadas con R
<<https://www.r-graph-gallery.com/all-graphs.html>>.

El entorno de programación R da un paso más allá, ya que se hace extensible mediante el uso de paquetes específicos, todo en código abierto. R se ha convertido en el estándar *de facto* de la comunidad científica para el análisis y la visualización de datos, puesto que hay miles de paquetes disponibles de casi cualquier ámbito de conocimiento.²⁴ Además, existe una extensa comunidad de apoyo que va generando y manteniendo nuevos paquetes y documentación de manera continua, junto con espacios de difusión y comunicación entre usuarios.

22. <<https://www.libreoffice.org/discover/calc/>>

23. Jeff Racine, «Gnuplot 4.0: a portable interactive plotting utility», *Journal of Applied Econometrics*, vol. 21, n.º 1 (2006).

24. <https://cran.r-project.org/web/packages/available_packages_by_name.html>

La creación de gráficos con R sigue la filosofía descrita por Wilkinson,²⁵ en la que una visualización es la superposición de capas que añaden semántica de acuerdo con una sintaxis sencilla que define qué datos se quieren visualizar y cómo. Usando paquetes como ggplot2 se pueden crear gráficos de cualquier tipo (figura 5), como se puede ver en la galería de ejemplos,²⁶ mantenida también por la comunidad.

Por otro lado, es destacable la existencia de herramientas como por ejemplo Gephi,²⁷ un software que facilita la manipulación y visualización de grafos, estructuras matemáticas que permiten mostrar relaciones entre elementos, así como sus atributos (figura 6). Es un software muy usado en el entorno académico y por periodistas de investigación, puesto que permite analizar datos provenientes de redes sociales como Twitter y Facebook, identificar comunidades, los elementos más o menos importantes de la red o la densidad de las relaciones entre elementos de manera visual.

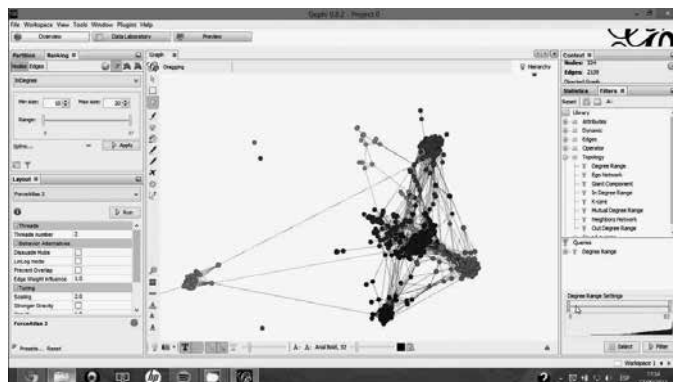


Figura 6. Uso de Gephi para la visualización de grafos.

Obviamente hay otras muchas herramientas para generar gráficos a partir de datos, normalmente en forma tabular, pero también de una expresión matemática. De una larga lista²⁸ se puede destacar: Matlab, un software propietario muy usado en el ámbito del análisis de datos matriciales y de señales, así como su versión en abierto, denominada GNU Octave; Mathematica, que incorpora capacidades algebraicas simbólicas; y finalmente, Orange,²⁹ una opción nueva y muy interesante para la exploración visual de datos mediante programación visual.

3.2 Lenguajes de programación para crear gráficos

La utilización de lenguajes de programación específicos para la creación de gráficos representa un más más allá, porque permite tener un control total de la visualización creada, pero, obviamente, con el coste de tener que programar todos los detalles. El nivel de abstracción es el píxel, aunque dispone también de gráficas básicas primitivas, como por ejemplo la línea y el polígono. Por otro lado, el grado de interacción es potencialmente muy elevado, aunque resulta muy complicado tener que programar todas las posibilidades.

En la actualidad, el lenguaje más habitual para generar gráficos con ordenador es el llamado *Processing*.³⁰ Este lenguaje permite la manipulación de un espacio virtual (lienzo o *canvas*) que se puede trasladar a la pantalla de manera total o parcial, permitiendo la creación de imágenes y el control total de la interacción con otros dispositivos de entrada o salida. Processing se ha convertido en un estándar *de facto* para toda una comunidad de creadores con perfil no tecnológico, pero que desean usar el ordenador como herramienta creativa más allá de la uti-

25. Leland Wilkinson, *The grammar of graphics*, New York: Springer, 2006.

26. <<https://www.r-graph-gallery.com/>>

27. Mathieu Bastian; Sebastian Heymann; Mathieu Jacomy, «Gephi: an open source software for exploring and manipulating networks», En: International AAAI Conference on Weblogs and Social Media (3rd: 2009: San Jose), *Proceedings of the Third International AAAI Conference on Weblogs and Social Media: 17-20 May 2009, San Jose, California, USA*, Menlo Park: AAAI Press 2009, pp. 361-362.

28. <https://en.wikipedia.org/wiki/List_of_information_graphics_software>

29. Janez Demšar et al., «Orange: data mining toolbox in Python», *Journal of Machine Learning Research*, vol. 14, n.º 1 (2013), pp. 2349-2353.

30. Casey Reas; Ben Fry, *Processing: a programming handbook for visual designers and artists*, Cambridge: MIT Press, cop. 2007.

lización de herramientas cerradas.³¹ Processing también se ha usado para visualizar datos. Son destacables los ejemplos de Aaron Koblin, que usan datos de los vuelos comerciales (figura 7), o de Brendan Dawes, con el su proyecto Cinema Redux, que usa los principios de Manovich³² y muestra los datos directamente.

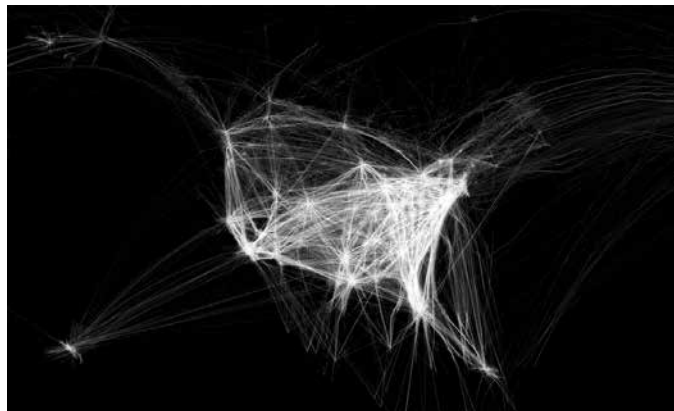


Figura 7. Visualización de los vuelos sobre el territorio de Estados Unidos
<<http://www.aaronkoblin.com/work/flightpatterns/>>.

Lo que resulta interesante de la visualización de Aaron Koblin es que con solo usar la posición de los aviones comerciales a lo largo del tiempo mientras siguen una ruta entre dos ciudades (de hecho, entre dos aeropuertos), se puede identificar la forma del país (en este caso, Estados Unidos), las zonas con más densidad de población y también zonas oscuras donde no sobrevuela ningún avión, sea por la razón que fuere. Todos los aeropuertos cercanos en las grandes ciudades estadounidenses también se ven claramente. En este caso la interacción se limita a poder reproducir el movimiento de los aviones en el tiempo y a poder hacer zoom para mostrar el detalle alrededor de los aeropuertos usando la codificación de colores que ha elegido Koblin: blanco cuando el avión está en tierra y de colores (en función de la compañía, o bien azul en general) cuando vuela. Como se puede ver en el vídeo³³

creado por el mismo Koblin, los datos permiten deducir la estructura de las pistas de aterrizaje y despegue de cada aeropuerto.

El concepto de *lienzo* no es exclusivo del Processing. En el mundo de la programación web también se pueden generar gráficos en un lienzo incrustado en una página web que se visualiza mediante un navegador, sin necesidad de ningún software específico. Se trata de una región definida como un elemento HTML en la que es posible dibujar mediante gráficos primitivos básicos (puntos, líneas, etc.), pero también con elementos más complejos, como por ejemplo polígonos, texto e imágenes (por ejemplo, iconos).

De hecho, las dos opciones están convergiendo, puesto que Processing también se encuentra disponible para ejecutarse en línea, como una página web. Al principio se ejecutaba con `processing.js`, una traducción de código Processing para ejecutarse como código JavaScript, y más recientemente con lo que se conoce como `p5.js`, una librería JavaScript para generar gráficos dentro de un documento HTML, como una nueva reinterpretación de la idea original propuesta por Processing. En el mundo de los lenguajes de programación, la elección de una opción u otra depende de muchos factores, aunque el uso intensivo de las posibilidades del lenguaje HTML hace de `p5.js` una opción más interesante para los programadores web, mientras que Processing es una plataforma excelente para iniciarse en la programación y la creación de gráficos.

A pesar de las posibilidades que ofrece poder manipular una visualización de datos a escala del píxel, la necesidad de tener que describir la visualización como un proceso que se ejecuta paso a paso hace que esta opción quede reservada a usuarios con un cierto nivel de conocimientos de programación, normalmente cuando ninguna de las otras opciones es suficiente o, sencillamente, se quieren explorar visualizaciones radicalmente diferentes de las tradicionales. Como muchas veces se trata de una progra-

31. Hartmut Bohnacker *et al.*, *Generative design: visualize, program, and create with processing*, New York: Princeton Architectural Press, 2012.

32. Lev Manovich, *op. cit.*

33. <<http://www.aaronkoblin.com/work/flightpatterns/>>

mación *ad hoc* para un proyecto de visualización concreto, generalmente esta opción es la que se usa menos.

3.3 Librerías

Siguiendo con la programación web, una opción intermedia entre las calculadoras gráficas y el uso de lenguajes de programación es la utilización de librerías JavaScript que proporcionan una capa de abstracción entre el usuario (y sus datos) y el lienzo donde se mostrará la visualización, de forma que no se trabaja a escala de píxel o de gráfica primitiva, sino a un nivel superior, manipulando conceptos sencillos, como por ejemplo diagramas, y configuraciones más complejas (*layouts*). La interacción también es más sencilla que en el caso de los lenguajes de programación, puesto que la mayoría de librerías incorporan una serie de opciones para capturar los acontecimientos de alto nivel producidos por el usuario, como por ejemplo el uso del ratón y el teclado, y la búsqueda de contenidos.

Igual que en el software para crear visualizaciones, existen una gran cantidad de librerías JavaScript que permiten incorporar gráficos en una página web para visualizar datos. Entre otros, se pueden destacar: InfoVis³⁴ (hoy en día obsoleta y superada); Raphaël,³⁵ orientada a crear diagramas sencillos; sigma.js,³⁶ orientada a la visualización de grafos; o Leaflet,³⁷ que permite crear mapas y superponer datos. Las librerías han ido evolucionando junto con el lenguaje HTML, las hojas de estilo CSS, el modelo de objetos de documento DOM y el lenguaje de gráficos vectoriales SVG, integrando las diferentes capas que componen una página web, pero manteniendo el acceso individual a los elementos que la componen, lo cual facilita la creación de visualizaciones dinámicas.

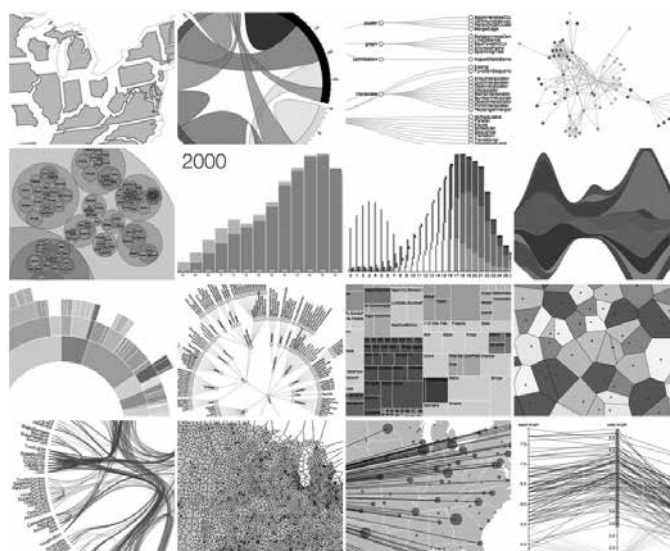


Figura 8. Galería de configuraciones (*layouts*) creadas con D3.

Fruto de esta evolución, en 2011 se dio a conocer la primera versión de D3 (acrónimo de Data-Driven Documents), una librería que también se ha escrito en JavaScript para proporcionar un control total sobre todos los elementos que componen una página web y su vínculo con los datos que se quieren visualizar.³⁸ D3 combina una estética muy buena y altamente configurable con un elevado potencial de funcionalidades, ofreciendo múltiples configuraciones predeterminadas en forma de *layouts* (figura 8). D3 se popularizó cuando Mick Bostock fichó por el diario *The New York Times* con el objetivo de crear visualizaciones nuevas e interactivas para explicar historias mediante datos. Una de las más destacables fue la cobertura y el análisis de los resultados de las elecciones del 2014 en Estados Unidos, como por ejemplo las palabras usadas por cada candidato en los discursos,³⁹ combinando gráficos con textos y descriptores estadísticos (figura 9).

34. <<https://philogb.github.io/jit/>>

35. <<http://dmitrybaranovskiy.github.io/raphael/>>

36. <<http://sigmajs.org/>>

37. <<https://leafletjs.com/>>

38. Michael Bostock; Vadim Ogievetsky; Jeffrey Heer, «D³ data-driven documents», *IEEE transactions on visualization and computer graphics*, vol. 17, n.º 12 (2011), pp. 2301-2309.

39. <<https://archive.nytimes.com/www.nytimes.com/interactive/2012/09/06/us/politics/convention-word-counts.html>>

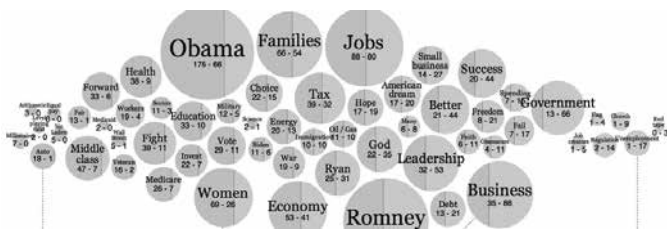


Figura 9. Palabras usadas en los discursos de las elecciones del 2014 en Estados Unidos, ordenadas por medida y color (izquierda, demócrata; derecha, republicano) (Bostock; Ogievetsky; Heer, 2011).

La principal ventaja de D3 es la interactividad integrada en la visualización, lo que permite realizar las operaciones básicas definidas por Shneiderman⁴⁰ de manera sencilla. Esta interacción, combinada con la gran variedad de configuraciones disponibles para visualizar datos, facilita la manipulación de conjuntos de datos una vez se ha decidido qué configuración o combinación de configuraciones es la más adecuada. En este sentido, pese a la complejidad interna y las dificultades que comporta crear visualizaciones de datos en D3, es posible utilizarlo como si fuera una caja negra,⁴¹ de forma que si los datos se encuentran en un formato concreto, se puede aprovechar una visualización existente para mostrarlos, y solo hay que hacer pequeños cambios relativos en el código y, especialmente, en las hojas de estilo que determinan los aspectos estéticos de la visualización. La galería⁴² de ejemplos de visualizaciones de datos creados en D3 es un muy buen punto de partida para comprobar las posibilidades que ofrece esta librería.

proceso de manipulación, preprocesamiento y visualización de datos. Últimamente han aparecido muchas opciones comerciales que compiten para proporcionar una solución completa, no solo para la visualización de datos, sino también para la toma de decisiones en un entorno de negocio, y la visualización de datos es solo un aspecto más integrado en la herramienta. La consultora tecnológica estadounidense Gartner realiza un estudio⁴³ anual de las mejores herramientas de inteligencia de negocio y analítica del mercado. Desde hace ya unos cuantos años, Tableau está posicionada como una de las mejores herramientas en este sector, y este último año se ha situado entre las tres líderes, junto con Qlik y los servicios ofrecidos por Microsoft Cloud (entre los que destaca especialmente Power BI).

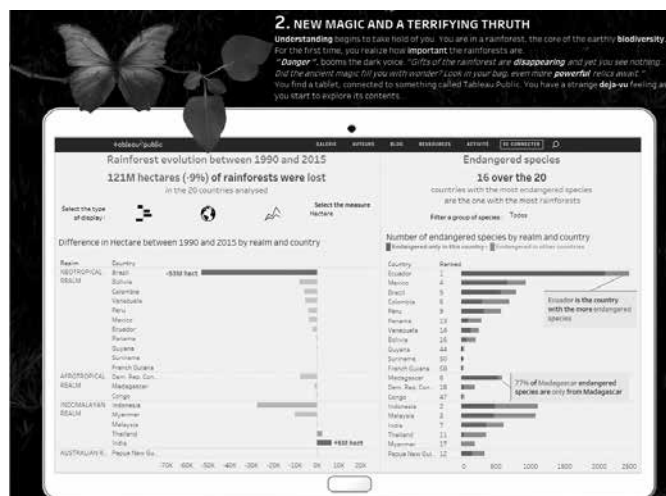


Figura 10. Ejemplo de narración visual con Tableau
<<https://public.tableau.com/es-es/gallery/tale-rainforest>>.

3.4 Entornos gráficos

Finalmente, la última opción corresponde a la sofisticación del concepto de calculadora gráfica, que proporciona verdaderos entornos gráficos que dan soporte a todo el

De hecho, si solo se tienen en cuenta las funcionalidades para crear visualizaciones de datos, una de las opciones más populares hoy en día es Tableau, una herramienta creada a partir de la investigación en el Departamento de

40. Ben Shneiderman, *op. cit.*

41. <<http://personal.uoc.edu/VIS/D3/ES/>>

42. <https://github.com/d3/d3/wiki/gallery>

43. <<https://www.analyticsvidhya.com/blog/2020/02/gartners-2020-magic-quadrant-bi-analytics-tools/>>

Informática de la Universidad de Stanford. Tableau permite crear diagramas a partir de uno o más conjuntos de datos, añadir capas con interactividad para hacer las operaciones típicas de filtraje, selección, etc., y construir verdaderos paneles de control (también llamados cuadros de mando o *dash-boards*) que sirven para narrar historias a partir de los datos (figura 10). Tableau permite, por medio de simples interacciones, tener diferentes perspectivas sobre un conjunto de datos. Además, nos ofrece numerosas técnicas de análisis y elementos estadísticos que pueden incorporarse a las visualizaciones creadas para entender mejor los datos y extraer conocimiento a partir del uso de representaciones gráficas.

Otra opción similar a Tableau es Quadrigram, que intenta combinar lo mejor de dos mundos. Por un lado, un entorno potente para la creación de visualizaciones de datos y, por el otro, la opción de generar código abierto para poder exportar la visualización como si fuera una página web. La filosofía de Quadrigram se basa en tres principios: el de no linealidad, puesto que las ideas que dan soporte a la visualización de datos se añaden a medida que se van creando y desarrollando; el principio de iteración, ya que el conocimiento que aporta la visualización se construye repitiendo el esquema definido por Keim *et al.*⁴⁴ y, finalmente, el principio que considera los datos como un material vivo, que evoluciona constantemente, un aspecto que la visualización tiene que poder capturar.

4. Conclusiones

La visualización de datos es un ámbito que últimamente se ha visto impulsado por la gran cantidad de datos que esperan ser analizados, la capacidad computacional disponible y también por la disponibilidad de herramientas para su manipulación, dejando atrás la idea de las visualizaciones de datos como simples resúmenes gráficos estáticos. En la actualidad, una visualización de datos puede incorporar un elevado grado de interactividad, de forma que permita manipular los datos directamente, in-

Sin sustituir el análisis estadístico clásico de cualquier estudio que use datos para comprender la realidad y tomar decisiones, la visualización de datos puede aportar conocimiento sobre el problema que hay que resolver de una manera sencilla y eficiente a la vez.

tegrando operaciones básicas, como por ejemplo la selección y el filtraje, así como otras que permitan extraer conocimiento mediante una primera inspección visual.

En este sentido, el análisis visual de datos explota las características del sistema visual humano, que es muy eficiente para detectar características básicas de los datos representados gráficamente en forma de tendencias, patrones, anomalías, etc. Sin sustituir el análisis estadístico clásico de cualquier estudio que use datos para comprender la realidad y tomar decisiones, la visualización de datos puede aportar conocimiento sobre el problema que hay que resolver de una manera sencilla y eficiente a la vez.

Para crear visualizaciones de datos hay cuatro aproximaciones diferentes, pero complementarias, definidas en función de dos dimensiones: por un lado, el nivel de abstracción de los elementos que componen la visualización, que va desde el píxel hasta el concepto de panel de control, y por el otro, el grado de interactividad, que puede ir desde visualizaciones estáticas hasta verdaderas interfaces dinámicas en las que el observador participa plenamente. No hay ninguna herramienta universal para visualizar un conjunto de datos cualquiera, sino que en función de los objetivos de la visualización, será necesario elegir entre soluciones *ad hoc* programadas desde el inicio o bien optar por la creación de visualizaciones estándar mediante herramientas más o menos complejas, y con más o menos grado de interactividad. Sea como fuere, el uso de la web como plataforma para la creación y difusión de aplicaciones con un gran componente visual

44. Daniel Keim, *op. cit.*

y con posibilidades de acceder a grandes volúmenes de datos y analizarlos hace que cada vez sea una opción más interesante y con más potencial, y cada vez hay más herramientas que permiten crear visualizaciones tanto de manera local como para compartirlas después mediante la web.

El futuro del análisis visual de datos implica hacer avanzar el modelo definido por Keim *et al.*,⁴⁵ introduciendo mecanismos automáticos para el reconocimiento de la naturaleza de los datos, igual que en otros ámbitos ligados a la inteligencia artificial, donde tendencias como el aprendizaje profundo (*deep learning*) se están imponiendo para extraer conocimiento de los datos sin tener que presuponer un modelo con anterioridad. La gran cantidad de datos y la creciente capacidad de cálculo disponible para analizarlos hace pensar que la visualización de datos será un ámbito de conocimiento en pleno auge durante los próximos años.

Bibliografía

BASTIAN, Mathieu; HEYMANN, Sebastien; JACOMY, Mathieu. «Gephi: an open source software for exploring and manipulating networks». En: International AAAI Conference on Weblogs and Social Media (3rd: 2009: San Jose), *Proceedings of the Third International AAAI Conference on Weblogs and Social Media: 17-20 May 2009, San Jose, California, USA*, Menlo Park: AAAI Press 2009, pp. 361-362.

BENGIO, Yoshua. «Learning deep architectures for AI», *Foundations and trends® in Machine Learning*, vol. 2, n.º 1 (2009), pp. 1-127.

BOHNACKER, Hartmut, *et al.* *Generative design: visualize, program, and create with processing*. New York: Princeton Architectural Press, 2012.

BOSTOCK, Michael; OGIEVETSKY, Vadim; HEER, Jeffrey. «D³ data-driven documents», *IEEE transactions on visualization*

and computer graphics, vol. 17, n.º 12 (2011), pp. 2301-2309.

CARD, Stuart K.; MACKINLAY, Jock; SHNEIDERMAN, Ben (ed.). *Readings in information visualization: using vision to think*. San Diego; London; San Francisco: Academic Press: Morgan Kaufmann, 1999.

CAIRO, Alberto. *El arte funcional: infografía y visualización de información*. Madrid: Alamut, 2011.

DEMŠAR, Janez, *et al.* «Orange: data mining toolbox in Python», *Journal of Machine Learning Research*, vol. 14, n.º 1 (2013), pp. 2349-2353.

FRIENDLY, Michael. «A brief history of data visualization». En: Chun-houh Chen; Wolfgang Härdle; Antony Unwin (ed.). *Handbook of data visualization*. Berlin; Heidelberg: Springer, 2008, pp. 15-56.

— «Milestones in the history of data visualization: a case study in statistical historiography». En: Annual Conference of the Gesellschaft für Klassifikation (28a: 2004: Dortmund), *Classification: the ubiquitous challenge: proceedings of the 28th Annual Conference of the Gesellschaft für Klassifikation e. v., University of Dortmund, March 9-11, 2004*. New York: Springer, 2005, pp. 34-52.

FRY, Ben. *Visualizing data: Exploring and explaining data with the processing environment*. Sebastopol: O'Reilly Media, 2007.

GREGORY, Richard L. «Knowledge in perception and illusion», *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 352, n.º 1358 (1997), pp. 1121-1127.

KEIM, Daniel, *et al.* «Visual analytics: Definition, process, and challenges». En: *Information visualization: human-centered issues and perspectives*, Berlin; Heidelberg: Springer, 2008, pp. 154-175.

45. *Ibid.*

MADDEN, Thomas J.; HEWETT, Kelly; ROTH, Martin S. «Managing images in different cultures: a cross-national study of color meanings and preferences», *Journal of International Marketing*, vol. 8, n.º 4 (2000), pp. 90-107.

MANOVICH, Lev. «What is visualization?», *Poetess Archive Journal*, vol. 2, n.º 1 (2010), 32 p.

MATEJKA, Justin; FITZMAURICE, George. «Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing». En: ACM CHI Conference on Human Factors in Computing Systems (2017: Denver), *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. [Denver: ACM, 2017], pp. 1290-1294.

PÉREZ COTA, Manuel, *et al.* «Analysis of Current Visualization Techniques and Main Challenges for the Future». *Journal of Information Systems Engineering & Management*, vol. 2, n.º 3 (2017), art. 19.

RACINE, Jeff. «Gnuplot 4.0: a portable interactive plotting utility». *Journal of Applied Econometrics*, vol. 21, n.º 1 (2006).

REAS, Casey; FRY, Ben. *Processing: a programming handbook for visual designers and artists*. Cambridge: MIT Press, cop. 2007.

SCHWABISH, Jonathan. «The 60,000 fallacy». En: *Policy-Viz [en línea]: helping you do a better job processing, analyzing, sharing, and presenting your data*, September 17, 2015. <<https://policyviz.com/2015/09/17/the-60000-fallacy/>>. [Consulta: 4 nov. 2017].

SHNEIDERMAN, Ben. «The eyes have it: a task by fecha type taxonomy for information visualizations». En: IEEE Symposium on Visual Languages (1996: Washinton). *VL'96: proceedings of the 1996 IEEE Symposium on Visual Languages*. Washington: IEEE Computer Society, 1996, pp. 336-343.

SUN, Guo-Dao, *et al.* «A survey of visual analytics techniques and applications: state-of-the-art research and futu-

re challenges». *Journal of Computer Science and Technology*, vol. 28, n.º 5 (2013), pp. 852-867.

THOMAS, James J.; COOK, Kristin A. «A visual analytics agenda». *IEEE computer graphics and applications*, vol. 26, n.º 1 (2006), pp. 10-13.

TUKEY, John W. «The future of data analysis». *The Annals of Mathematical Statistics*, vol. 33, n.º 1 (1962), pp. 1-67.

— *Exploratory data analysis*. London: Sage, 1977.

WANDELL, Brian A. *Foundations of vision*. Sunderland: Sinauer Associates, 1995.

WEDEL, Michel; PIETERS, Rik (ed.). *Visual marketing: From attention to action*. New York: Psychology Press, 2012.

WILKINSON, Leland. *The grammar of graphics*. New York: Springer, 2006.

WONG, Dona M. *The Wall Street Journal guide to information graphics: the dos and don'ts of presenting data, facts, and figures*. New York: Norton, 2010. ■