Caixa d'eines

Anàlisi visual de dades: conceptes bàsics i eines


Julià Minguillón

Anàlisi visual de dades: conceptes bàsics i eines

Julià Minguillón
Estudis d'Informàtica, Multimèdia i Telecomunicació, Universitat Oberta de Catalunya
jminguillona@uoc.edu

Article rebut el novembre de 2017; revisat el desembre de 2017.

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.

Análisis visual de datos: conceptos básicos y herramientas

Resumen: Vivimos en un mundo físico que se proyecta sobre otro mundo, completamente digital, que consume datos como combustible principal, extrae y genera conocimiento. Actualmente, con estos datos se toman decisiones en todos los ámbitos, desde el personal hasta el de las grandes corporaciones multinacionales, así como el de la administración y el 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 clave: visualización de datos, análisis visual, herramientas, software.

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

Introducció

Els éssers humans som, principalment, visuals. El sistema de visió humana és una màquina sofisticada que permet capturar una gran quantitat d'informació de l'entorn i fer-la servir tant per a avaluar-lo com per a emprendre accions. De fet, es diu que el 90 % de la informació transmesa al cervell és visual i s'estima que les imatges es processen 60.000 cops més ràpidament que el text escrit, tot i que aquest fet no s'ha contrastat científicament.[1] La dita popular «una imatge val més que mil paraules», resumeix aquesta idea. Sigui com sigui, els experts en màrqueting saben que una imatge capta més l'atenció que no pas un text i aprofiten aquest coneixement per a transmetre idees i fets de manera efectiva i eficaç.[2]

Per tant, cada cop és més habitual l'ús d'imatges i vídeos per a transmetre informació visualitzant idees i fets en lloc de detallar-los de manera textual, no només en l'àmbit de la publicitat sinó també en qualsevol context on s'usin dades per a prendre decisions. Cal començar aclarint que actualment hi ha dos conceptes que de vegades s'usen indistintament i no són del tot equivalents, sinó més aviat complementaris. Ens referim al concepte de *visualització de dades* i a un altre que s'ha popularitzat més recentment anomenat *infografia*. Tots dos fan servir representacions gràfiques per a presentar les dades i relatar les històries que hi ha darrere, però amb objectius i procediments diferents.

Es pot definir *infografia* com una representació més visual que els mateixos textos, en la qual intervenen descripcions, narracions o interpretacions, presentades de manera gràfica normalment figurativa, que poden coincidir o no amb grafismes abstractes o sons. La infografia neix com un mitjà per a transmetre informació gràficament, que disposa d'un mètode per a representar la informació icònicament i textualment, de manera que l'usuari la pugui comprendre sense dificultat. En el procés de creació d'una infografia, que s'acostuma a fer emprant eines informàtiques, es recull un fet complex i s'explica de manera senzilla perquè es pugui interpretar amb un simple cop d'ull.

En comparació amb una infografia, la visualització de dades, també anomenada *visualització de la informació* (ja que posa les dades en el seu context), és l'estudi de la representació visual de dades abstractes (i potser interactives) per a reforçar la cognició humana, que inclou tant dades numèriques i no numèriques com text o informació geogràfica, per exemple. Per tant, es pot deduir que no hi ha, pràcticament, gaires diferències substancials entre tots dos conceptes, ja que hi ha una naturalesa comuna entre una infografia i una visualització, tal com explica molt bé Alberto Cairo,[3] que indica encertadament les subtils diferències conceptuals que hi ha:

> Alguns especialistes marquen una frontera entre les dues disciplines basada en el fet que, suposadament, la infografia consisteix a presentar informació per mitjà de gràfics estadístics, mapes i esquemes (exposició), mentre que la visualització es basa en la creació d'eines visuals (estàtiques o interactives) que un públic pugui fer servir per a explorar, analitzar i estudiar conjunts complexos de dades. Però pertanyen a un mateix continu en el qual cadascuna ocupa extrems oposats d'una línia. Aquesta és paral·lela a una altra els límits de la qual són definits per les paraules presentació i exploració.

En aquest article ens centrarem en les visualitzacions de dades com a mecanismes per a extreure'n coneixement, aprofitant les capacitats del sistema visual humà i les possibilitats que ofereix la mateixa visualització per a la manipulació de les dades, afegint-hi un cert grau d'interactivitat.

1. Jonathan Schwabish, «The 60,000 fallacy», En: *PolicyViz* [en línia]: *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. Antecedents històrics

La visualització de dades com a mecanisme de narració d'històries ha estat sempre present en el desenvolupament de la nostra societat al llarg del temps. Els éssers humans han dibuixat imatges per a comunicar-se des de fa milers d'anys, des de pictogrames a les parets d'una cova rupestre i els jeroglífics egipcis, fins a ideogrames i tota la iconografia moderna. L'ésser humà sempre ha fet servir imatges per a comunicar i explicar històries, perquè el cervell humà ha evolucionat d'aquesta manera i és molt eficient processant informació mitjançant el sistema visual.

Un dels millors reculls sobre els antecedents de la visualització de dades tal com l'entenem avui dia és el treball de Michael Friendly.[4] Es va publicar en un capítol d'un manual de visualització de dades que formava part d'una col·lecció de llibres d'estadística, la qual cosa mostra la importància de la visualització com a eina per a l'anàlisi de dades. L'article de Friendly està estructurat en una línia temporal, que inclou des de les primeres visualitzacions (mapes i diagrames) anteriors al segle XVII fins a l'actualitat (a partir de 1975), en què la tecnologia ha fet possible la creació massiva de visualitzacions. Una de les etapes més interessants destacades per Friendly és la segona meitat del segle XIX, quan es van desenvolupar moltes tècniques per a l'anàlisi estadística que s'aplicaven a tots els àmbits de la planificació social, la industrialització, el comerç i el transport. Això va provocar l'aparició de moltes innovacions en la visualització de dades, necessàries per a poder explicar les dades i els fenòmens, tan complexos, de la societat del moment. Un primer pas va ser la projecció d'elements en tres dimensions (3D) com a via d'escapament del pla, que fins aleshores limitava les possibilitats. Un altre exemple interessant va ser la combinació de mapes amb dades de cada regió, de manera que en una mateixa representació es combinaven dades espacials amb altres de temporals. Finalment, l'ús de gràfics per a l'anàlisi estadística (per exemple, la *correlació* era un concepte encara en desenvolupament) va permetre que Francis Galton i altres investigadors avancessin en la formalització de les observacions i les convertissin en tècniques estadístiques.

Un cèlebre cas del segle XIX és el del metge anglès John Snow, que es podria destacar com un dels iniciadors de la visualització de dades moderna. El mapa de Snow (figura 1) és considerat un dels primers exemples d'ús d'un mètode geogràfic per a descriure i localitzar els casos d'una epidèmia, així com l'origen més probable, cosa que mai s'havia fet abans. Això va permetre establir els mecanismes de transmissió de les malalties infeccioses. És un exemple de com una visualització de dades pot fer-se servir per a provar o refutar una hipòtesi, en aquest cas la transmissió del còlera per l'aigua, i convertir fets en dades que ajuden a prendre una decisió. D'aquesta manera, John Snow va demostrar que la causa dels casos de còlera a Londres era el consum d'aigua contaminada amb materials fecals. El 1854 va cartografiar en un plànol del Soho els pous d'aigua i els casos de còlera, així va localitzar el pou que n'era l'origen, que estava situat a Broad Street, al centre de l'epidèmia, fet pel qual va recomanar clausurar la bomba d'aigua que l'alimentava. Així va aconseguir disminuir la proliferació dels casos de còlera en aquesta zona de Londres.

En l'actualitat podríem trobar un estudi equivalent en diferents anàlisis que s'han fet de fenòmens com AirBnB, per veure com nous actors poden distorsionar i generar tensions en un mercat tradicional, com en el cas dels pisos turístics amb llicència o sense i l'increment del lloguer d'habitatges habitualment dedicats al lloguer de llarga durada. En la majoria de ciutats, els lloguers privats de curta durada sense llicència turística són il·legals, especialment per les diferències amb el model tradicional, en què el sector es troba molt regulat i sotmès a una gran quantitat de taxes i impostos. Un exemple d'aquest creixement descontrolat es pot veure en una visualització

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.

creada per Kor Dwarshuis,[5] en què les dades que publica AirBnB es visualitzen combinades en un mapa que geolocalitza tant les ofertes de pisos com els lloguers que hi ha hagut, conjuntament amb un eix temporal que permet veure la disseminació de l'oferta i la demanda al llarg del temps, un aspecte que John Snow no va poder reflectir en el seu mapa, per la tecnologia del moment.

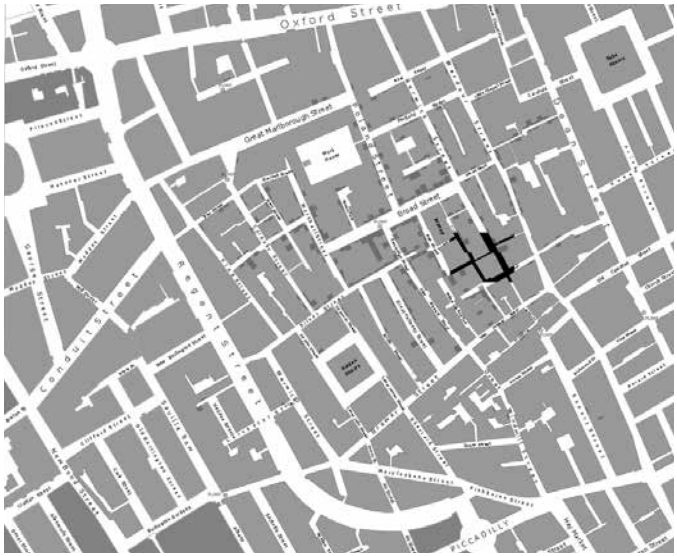

Figura 1. Mapa de la localització dels brots de còlera al Soho, Londres, per John Snow, 1854. <*https://ca.wikipedia.org/wiki/John_Snow#/media/File:Dr._John_Snow_Cholera_Map.svg*>.

Aquests exemples mostren que una visualització de les dades, en aquest cas superposada en un mapa, permet fer-se una idea ràpida de la magnitud del fenomen, així com facilitar la detecció d'elements que per una raó o altra destaquen sobre la resta, aportant coneixement sobre el problema a resoldre. És aquesta idea la que dona suport al que en diem «anàlisi visual» (*visual analytics*), i que es desenvolupa a continuació.

2. Anàlisi visual

La utilització de visualitzacions de dades com a mecanisme per a analitzar-les es fonamenta en les propietats del sistema visual humà, que es defineix com una part del sistema nerviós central que proporciona als organismes vius (en general) l'habilitat de processar visualment, detectant i interpretant la llum visible, per a entendre l'escenari que els envolta, creant el que s'entén com a percepció. És important destacar que el sistema visual humà el formen no només els ulls, sinó també els nervis òptics i àrees específiques del cervell (el còrtex visual), que combinen diferents nivells d'abstracció. De fet, no hi veiem amb els ulls, sinó amb el cervell, ja que la percepció és una combinació de diferents processos que fan tasques diverses relacionades amb la visió. Tal com descriuen Card, MacKinlay i Shneiderman,[6] la visualització de dades proporciona un procediment molt potent per a permetre als usuaris detectar i interpretar patrons en les dades.

Així, el procés de percepció permet als humans fer tasques de manera eficient, com ara discriminar colors, separar mitjançant el contrast, estimar distàncies i mides, determinar orientacions i angles, i reconstruir el moviment dels objectes que formen l'escena, cosa que ens permet la navegació en el món físic tridimensional. Tal com descriu Wandell,[7] les propietats del sistema visual humà i la codificació que resulta de la llum captada per la retina per fer totes aquestes accions també tenen implicacions en el disseny d'instruments que mostren informació de manera visual, com ara una visualització de dades. El procés de percepció dels humans encara no ha estat igualat per cap màquina o procés artificial, tot i que és una línia de recerca molt activa des de ja fa uns quants anys, la qual s'ha vist impulsada darrerament pel boom en àmbits com la intel·ligència artificial i el que es coneix com a «aprenentatge profund»,[8] aprofitant la disponibilitat de gran quantitat de dades i l'elevada capacitat computacional necessària per a processar-les.

5. <*http://www.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.
8. Yoshua Bengio, «Learning deep architectures for AI», *Foundations and trends® in Machine Learning*, vol. 2, no. 1 (2009), p. 1-127.

El procés de percepció permet als humans fer tasques de manera eficient, com ara discriminar colors, separar mitjançant el contrast, estimar distàncies i mides, determinar orientacions i angles, i reconstruir el moviment dels objectes que formen l'escena, cosa que ens permet la navegació en el món físic tridimensional.

Cal tenir en compte, però, que el sistema visual humà, tot i la potència i complexitat que té, també pot ser enganyat fàcilment mitjançant il·lusions òptiques,[9] mostrant deficiències que s'haurien d'evitar per no caure en un parany a l'hora de mostrar dades de manera gràfica. Per exemple, una perspectiva mal usada pot distorsionar les mides dels elements que es comparen. Un ús incorrecte de les saturacions i el contrast pot fer veure àrees de colors diferents com si fossin similars. A més, també caldrà tenir en compte que un percentatge gens menyspreable de la població mundial té algun tipus de deficiència visual pel que fa al processament del color, de manera que necessiten l'ús de codificacions alternatives (per exemple, escala de grisos i formes), o bé en els casos en què la visualització ha de ser impresa en blanc i negre. Finalment, també cal tenir en compte altres aspectes culturals que poden determinar la manera com s'interpreta una visualització de dades, ja que l'ordenació espacial dels elements que la componen i la seva comprensió poden dependre de si l'observador la llegeix d'esquerra a dreta i de dalt a baix, o bé al contrari. Per tant, cal tenir en compte tots aquests aspectes a l'hora de dissenyar una bona visualització de dades que permeti fer una anàlisi visual preliminar sense interferències causades per una mala decisió de codificació, aprofitant l'experiència d'estudis de l'àmbit de la publicitat en què, per exemple, s'ha demostrat que el color afecta directament la percepció del consumidor i que hi ha diferències culturals respecte a la percepció.[10] També des de l'àmbit del periodisme, en què els gràfics s'utilitzen per destacar o donar suport a una idea, hi ha hagut autors que han identificat els aspectes bàsics que determinen si una visualització de dades és una bona o una mala pràctica.[11]

No obstant això, a l'inici, el desenvolupament de l'anàlisi visual de dades no va tenir en compte els aspectes estètics, sinó que es va limitar a reduir les dades a un conjunt de mesures que en resumeixen les principals característiques, utilitzant gràfiques primitives i senzilles per a representar-les. Tal com descriu Friendly[12] en el seu projecte Milestones, el desenvolupament de l'ús de representacions gràfiques per a la descripció de dades quantitatives va anar lligat a la necessitat de resumir un concepte mesurable, normalment amb l'objectiu d'entendre'l millor.

John Tukey va establir els principis de l'anàlisi de dades, definint-la com el conjunt de procediments per analitzar dades, les tècniques per interpretar-ne els resultats, les maneres de planificar la captura de dades per a facilitar-ne l'anàlisi i fer-la més precisa i acurada, i tot el maquinari i les estadístiques que cal aplicar per analitzar les dades.[13] Posteriorment, el mateix Tukey[14] va desenvolupar el concepte d'anàlisi exploratòria de dades, que té els objectius de suggerir hipòtesis sobre les causes d'un fenomen observat,

9. Richard L. Gregory, «Knowledge in perception and illusion», *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 352, no. 1358 (1997), p. 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, no. 4 (2000), p. 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 fur Klassifikation e. v., University of Dortmund, March 9-11, 2004*, Claus Weihs; Wolfgang Gaul (ed.), New York: Springer, 2005, p. 34-52.
13. John W. Tukey, «The future of data analysis». *The Annals of Mathematical Statistics*, vol. 33, no. 1 (1962), p. 1-67.
14. John W. Tukey, *Exploratory data analysis*. London: Sage, 1977.

avaluar assumpcions en les quals cal fonamentar la inferència estadística, donar suport a la selecció de les eines i tècniques estadístiques més apropiades i, finalment, proporcionar una base per a una recollida addicional de dades a través d'enquestes o experiments. Una primera visualització pot servir per copsar la naturalesa del problema i reduir el ventall de possibles aproximacions per resoldre'l.

Un exemple de la necessitat de representar gràficament les dades per a entendre'n la natura és l'anomenat *quartet d'Anscombe*, creat per Francis Anscombe el 1973, que comprèn quatre conjunts de dades que tenen les mateixes propietats estadístiques, però que evidentment són diferents quan se n'inspeccionen els gràfics respectius, mostrant les limitacions de l'ús de descriptors estadístics per a resumir un conjunt de dades (figura 2). Concretament, es tracta de quatre conjunts de dades d'onze punts en un pla (x, y), de manera que la mitjana i la variància de cada variable, com també la correlació entre totes dues i el coeficient de la recta de regressió òptima són idèntics per als quatre conjunts (o es podria pensar que són idèntics), mentre que són clarament diferenciables si s'utilitza una representació visual. Òbviament, cada conjunt representa el resultat de quatre processos diferents que podrien haver-lo generat. Així, es pot identificar una col·lecció de dades típica (figura superior esquerra), unes dades que segueixen una relació no lineal però ben clara (figura superior dreta), unes dades que segueixen una relació lineal tret d'una, identificant així una possible dada atípica (figura inferior esquerra) i, finalment, unes dades que mostren una relació no lineal entre les dues variables, però en què una simple dada atípica genera un coeficient de correlació elevat. Sense la visualització d'aquestes dades fent servir un simple gràfic de dispersió (x, y), és molt difícil fer-se a la idea de les quatre distri-

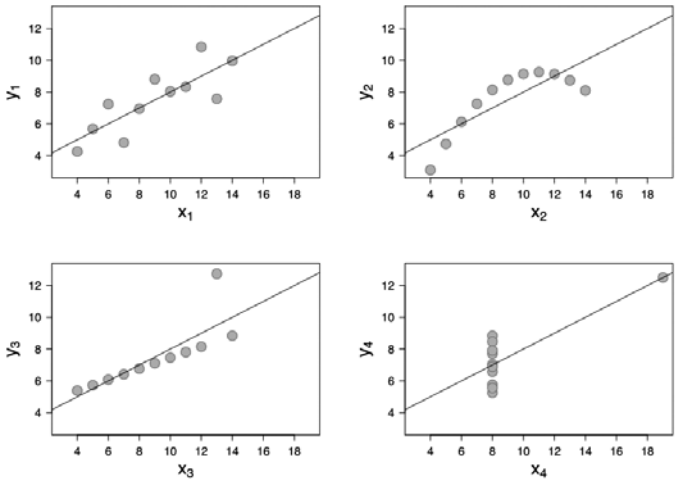

Figura 2. El quartet d'Anscombe
<https://en.wikipedia.org/wiki/Anscombe%27s_quartet#/media/File:Anscombe%27s_quartet_3.svg>.

Encara que es tracta d'un exemple sintètic, mostra de manera convincent les limitacions dels descriptors estadístics més habituals en els treballs de recerca i les possibilitats de la visualització com a eina d'anàlisi visual complementària. Recentment, Matejka i Fitzmaurice[15] han generat fins a una dotzena de conjunts de dades diferents[16] que mostren els mateixos descriptors estadístics, com a exemple de la necessitat d'usar l'anàlisi visual per a entendre millor les dades.

En la mateixa línia de raonament, aquest reduccionisme (en el sentit de reduir un conjunt de dades a un nombre reduït de descriptors estadístics) és descrit i criticat per Manovich,[17] que aposta per mostrar les dades originals als usuaris finals, totalment o parcialment, per tal de permetre'ls formar-se una idea de la seva natura, especial-

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], p. 1290-1294.
16. <*https://www.autodeskresearch.com/publications/samestats*>
17. Lev Manovich, «What is visualization?», *Poetess Archive Journal*, vol. 2, no 1 (2010), 32 p.

ment en dades que porten una càrrega semàntica molt important (p. ex., imatges i mapes). Segons Manovich, des de la segona meitat del segle XVIII fins avui hi ha hagut dos principis clau que han donat forma a la visualització d'informació. El primer és el principi de reducció, que consisteix en l'ús de gràfiques primitives (punts, línies, formes geomètriques simples…) per a la representació dels elements i les seves relacions, que revelen patrons i estructures subjacents sense necessitat de visualitzar les dades originals. Això ha comportat una pèrdua d'importància de les dades pel que fa a les representacions, massa esquemàtiques en alguns casos. El descriptor més senzill és la mitjana, acompanyat habitualment de la variància, que indica fins a quin punt les dades estan centrades al voltant de la mitjana, resumint tot un conjunt de dades a un valor o dos. El pas següent és fer servir diagrames de caixa per a descriure els quartils, mostrant la distribució de les dades i l'existència de possibles dades atípiques. Actualment s'utilitzen els diagrames de violí, que integren l'histograma de la distribució subjacent com a part de la visualització i afegeixen informació sobre la distribució real de les dades (figura 3), tot i que són una simplificació de la natura del conjunt de dades.

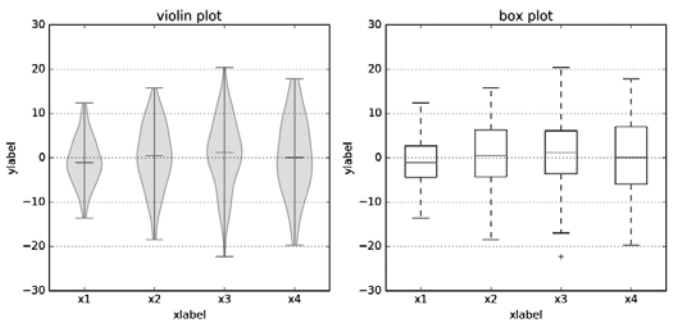

Figura 3. Ús del diagrama de violí (esquerra) com a evolució dels diagrames de caixa (dreta) *<https://matplotlib.org/1.5.1/examples/statistics/boxplot_vs_violin_demo.html>*.

Aquest reduccionisme, present en tots els àmbits de les ciències, proposa que el món pot analitzar-se a partir dels elements simples que el componen i de les regles que en regeixen les interaccions, de manera que es pugui comprendre la totalitat mitjançant una descripció simplificada o reduïda. Així, durant el segle XIX es van desenvolupar tots els gràfics típics per a representar aquestes dades «reduïdes» que permeten explicar aspectes socials, demogràfics, etc. De fet, va ser en aquesta època quan van aparèixer els gràfics de barres i de pastís, els histogrames, etc., tots conceptualitzats des d'aquesta visió reduccionista i fent servir els mateixos elements gràfics senzills (punts, línies, caixes, etc.).

D'altra banda, el segon principi que esmenta Manovich és l'ús de variables espacials (posició, grandària, forma, etc.) per a representar diferències en les dades i revelar, així, els patrons i les relacions existents més importants. En l'exemple (fictici) de la figura 3 es poden observar les diferències entre quatre conjunts de dades diferents pel que fa a la distribució d'una variable en cada conjunt. Manovich fa notar que la visualització d'informació privilegia les dimensions espacials sobre les altres i dona més importància a la topologia i a la geometria i menys a altres aspectes com ara el color, la saturació i la transparència. Així, per a representar un conjunt de dades, les dimensions més importants s'assignen a la disposició espacial (anomenada *layout*), mentre que la resta de dimensions es mapen habitualment en la resta de les variables visuals (color, etc.). En aquest cas, el color o la forma es fan servir només per a dividir els elements d'un conjunt de dades en diferents classes. Manovich esmenta com a possible raó la dificultat de reproduir representacions gràfiques mitjançant la tecnologia que hi ha en cada moment, la qual cosa limita l'ús del color, la transparència, etc. Els ordinadors han permès crear i manipular representacions més complexes, potenciant l'ús d'altres dimensions visuals.

Sigui com sigui, usant els instruments típics de l'estadística descriptiva, o bé mostrant les dades directament, l'ús de visualitzacions de dades pot tenir diferents objectius, tots relacionats amb l'anàlisi exploratòria descrita per Tukey:

- **Descriptiu**: resumeix de manera gràfica les propietats d'un conjunt de dades, mitjançant els seus descriptors estadístics bàsics: mitjana, variància, quartils, histogrames, etc.
- **Comparatiu**: combina els elements descrits en el punt anterior per mostrar semblances o diferències entre un o més conjunts de dades o d'acord amb una o més variables.
- **Detecció de tendències**: fa servir, habitualment, un eix temporal que permet veure ràpidament si un fenomen creix i/o decreix regularment.
- **Detecció de patrons**: la representació que es tria permet detectar agrupacions o repeticions en les dades en una, dues o, més rarament, tres dimensions.
- **Detecció d'anomalies** (dades atípiques): identifica elements que destaquen clarament de la resta per la seva posició o mida, entre altres atributs possibles.
- **Detecció de correlacions**: mostra el grau de relació entre dues variables. Cal tenir sempre present que correlació i causalitat són dos conceptes clarament diferents i que, de cap manera, el primer implica el segon. Cosa que pot resultar confusa en una visualització de dades.[18]
- **Relacional**: dona més èmfasi a les relacions entre dades que no pas a les dades en si, fa servir representacions en què la posició relativa dels elements és més important que l'absoluta, mentre que els atributs es mapen sobre les característiques de la representació (forma i color, principalment).
- **Jeràrquic**: mostra una estructura de relacions en què els elements s'agrupen d'acord amb una taxonomia i es poden visualitzar diferents nivells de detall.
- **Localitzat**: superposa les dades o els seus atributs en forma de capes sobre un mapa o esquema que aporta una semàntica molt més rica en forma de distàncies o posicions, tant relatives com absolutes.

Com era d'esperar, no hi ha una visualització de dades universal que en permeti copsar ràpidament la natura, bàsicament per dues raons: per una banda, la varietat de dades (especialment pel que fa al gran nombre d'atributs que poden usar-se per a descriure cada element del conjunt de dades) i, per l'altra, l'objectiu que es pretén assolir combinant un o més dels punts esmentats. Habitualment, cal fer més d'una exploració abans de poder formar-se una idea prou clara d'un conjunt de dades. En general, les dades són complexes i poden combinar diversos aspectes al mateix temps, entre d'altres se'n poden destacar els següents: són multidimensionals; van lligades a restriccions espaciotemporals, longitudinals (que evolucionen en el temps) o multimodals (combinen diferents fonts i orígens), i provenen de l'execució de múltiples processos paral·lels o models. Visualitzar dades inclou la gestió de tota aquesta complexitat per convertir-les en informació, és a dir, obtenir respostes a les preguntes o als objectius de la visualització que es pretén crear. De nou, l'anàlisi visual no substitueix l'estadística clàssica o la construcció de models de mineria de dades, sinó que aporta una perspectiva diferent basada en les capacitats del sistema visual humà, incorporant en l'objectiu de la visualització la possibilitat de fer diferents operacions amb les dades de manera més intuïtiva.

En aquest sentit, l'evolució de l'àmbit de la visualització de dades no s'ha centrat només en la capacitat de generar gràfics complexos amb més resolució en un breu lapse de temps, sinó que ha anat incorporant elements interactius en la mateixa visualització en forma d'operacions bàsiques (selecció, filtratge, etc.). D'acord amb el treball de Keim, *et al.*,[19] l'anàlisi visual de dades es fonamenta en un mantra que és una versió modificada del que va proposar Ben Shneiderman el 1996:

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

Així, el procés d'anàlisi visual consisteix en un cicle continu que s'inicia en les dades i les seves possibles transformacions, i que es bifurca en dues aproximacions com-

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, p. 154-175.

plementàries. La visualització i la construcció de models, entre les quals hi ha un diàleg amb l'objectiu d'extreure coneixement que pugui fer-se servir per a iterar el procés d'anàlisi visual amb més detall o complexitat (figura 4). La capacitat d'interacció ha de permetre fer a l'usuari de la visualització, almenys, les operacions bàsiques definides per Shneiderman (vista general, zoom, filtre i selecció).

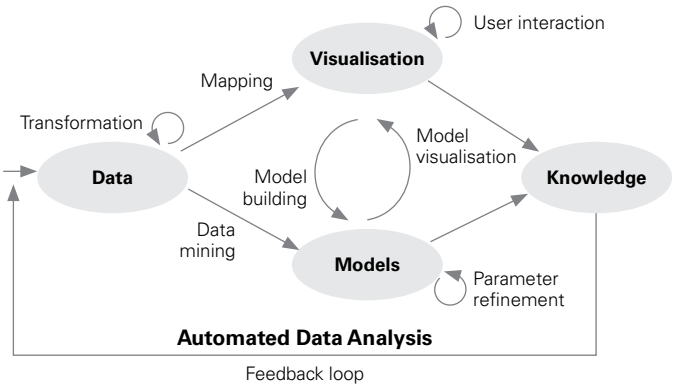

Figura 4. Procés d'anàlisi visual (Keim, *et al.*, 2008).

Des d'una perspectiva d'anàlisi visual, les dues primeres etapes definides a la figura 4 són la transformació (o adaptació) i la visualització de les dades, incloent-hi la possible interacció per fer operacions amb les dades. Per tant, un cop establert l'objectiu de l'anàlisi visual de les dades, es tracta de seleccionar un tipus de visualització més o menys interactiva que permeti fer l'exploració preliminar. En una primera iteració, les dades poden mostrar-se tal com apareixen, sense cap transformació. L'observador determina la visualització a partir del coneixement que en treu, per exemple, visualitzant el resultat d'aplicar una anàlisi de components principals si les dades mostren una certa estructura entre variables que pot ser explotada.

Així, la construcció d'una visualització de dades és un procés que integra diversos punts de vista, des del més proper a la natura de les dades, que involucra l'ús de descriptors estadístics i models de mineria de dades per treure'n coneixement, fins a aspectes més lligats a la percepció de l'observador final, que inclou tant elements estètics com altres de culturals. En el proper apartat es descriuen quatre famílies o categories d'eines per a la creació de visualitzacions de dades en funció de dues dimensions: per una banda, el nivell d'abstracció de la representació gràfica, i per l'altra, el grau d'interactivitat permès que determina l'exploració posterior.

3. Eines per a l'anàlisi visual

En l'actualitat hi ha moltes opcions per a visualitzar dades, siguin de la mena que siguin. De fet, aquesta gran disponibilitat d'eines i recursos ha estat, en part, la causa de la popularització de les visualitzacions de dades en tots els àmbits, no només el científic.

Per a classificar-les s'ha optat per agrupar-les d'acord amb les dues dimensions esmentades. La primera dimensió, el nivell d'abstracció, fa referència a la granularitat del tipus d'objecte que es manipula per a crear la visualització, i pot anar des d'un píxel fins a panells de control que combinen múltiples visualitzacions. La segona dimensió, el grau d'interactivitat, fa referència a les opcions que té l'observador per manipular dades mitjançant la visualització, i va des de visualitzacions estàtiques que no permeten cap interacció fins a interfícies complexes que inclouen totes les operacions bàsiques definides per Shneiderman.[20] Algunes eines esmentades en aquest apartat apareixen en l'estudi de Cota, *et al.*[21] des d'una perspectiva orientada a la visualització de grans volums de dades.

20. Ben Shneiderman, «The eyes have it: a task by data type taxonomy for information visualizations». En: IEEE Symposium on Visula Languages (1996: Washington). *VL'96: proceedings of the 1996 IEEE Symposium on Visual Languages*. Washington: IEEE Computer Society, 1996, p. 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, no. 3 (2017), art. no. 19.

El procés d'anàlisi visual consisteix en un cicle continu que s'inicia en les dades i les seves possibles transformacions, i que es bifurca en dues aproximacions complementàries.

3.1 Calculadores gràfiques

En aquesta primera categoria podem trobar eines de propòsit general que permeten visualitzar dades de manera senzilla, amb gràfics predeterminats que, suposadament, permeten mesurar o comparar diferents valors entre si, d'acord amb el valor d'un o més atributs o entre conjunts de dades diferents, lligats als descriptors estadístics més habituals. Es tracta, de fet, de programari amb una clara orientació a la manipulació de dades amb l'objectiu d'extreure'n coneixement, sobretot, en forma de gràfics senzills, tot i que en alguns casos inclouen altres capacitats més avançades. Bàsicament, aquestes eines converteixen un conjunt de dades en una visualització predeterminada per l'usuari, normalment estàtica i sense cap interacció, amb una capacitat d'exploració limitada.

Si les dades es troben en forma tabular, la primera eina que es pot utilitzar per a visualitzar-les és un full de càlcul, com ara Microsoft Excel o, també, eines de programari lliure com LibreOffice Calc,[22] que millora la importació de dades en format .csv amb un assistent per a ajustar la codificació i estructura de les dades. Històricament, els fulls de càlcul han estat les eines més usades per a generar els típics gràfics que resumeixen les dades, com ara els gràfics de línies, de barres o de pastís. El primer full de càlcul va ser VisiCalc, desenvolupat per l'Apple II l'any 1979, seguit de Lotus 1-2-3 per a l'IBM PC, l'any 1983, amb capacitats gràfiques més bones. Totes dues es consideren les primeres veritables aplicacions rupturistes (*killer-apps*) que van impulsar la venda d'ordinadors personals. Les eines actuals proporcionen un seguit d'opcions que converteixen un conjunt de dades en un tipus de gràfic concret, incloent-hi la possibilitat de personalitzar l'aspecte, mitjançant l'ús de colors, trames, opcions 2D i 3D, etc.

Una altra opció més senzilla és fer servir un programari específic per a generar gràfics, com ara Gnuplot,[23] que permet visualitzar dades d'acord amb un repertori de representacions lligades a descriptors estadístics, com ara histogrames o diagrames de caixa, entre d'altres. Gnuplot s'orienta a la visualització de dades d'una, dues o tres dimensions, i s'ha convertit en un motor gràfic que pot fer-se servir des de diferents llenguatges de programació i entorns gràfics, mitjançant un senzill llenguatge script que genera gràfics a partir de dades i comandes senzilles.

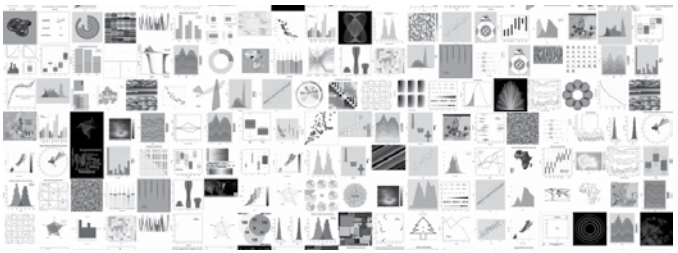

Figura 5. Galeria de visualitzacions creades amb R <*www.r-graph-gallery.com/2016/08/02/the-r-graph-gallery/*>.

Un pas més enllà el fa l'entorn de programació R, que és extensible mitjançant l'ús de paquets específics, tot en codi obert. R s'ha convertit en l'estàndard *de facto* de la comunitat científica per a l'anàlisi i la visualització de dades, ja que hi ha milers de paquets disponibles de gairebé qualsevol àmbit de coneixement.[24] A més, hi ha una extensa comunitat de suport que va generant i mantenint nous paquets i documentació de manera contínua, juntament amb espais de difusió i comunicació entre usuaris.

22. <*https://www.libreoffice.org/discover/calc/*>
23. Jeff Racine, «Gnuplot 4.0: a portable interactive plotting utility». *Journal of Applied Econometrics*, vol. 21, no. 1 (2006).
24. <*https://cran.r-project.org/web/packages/available_packages_by_name.html*>

La creació de gràfics amb R segueix la filosofia descrita per Wilkinson,[25] en què una visualització és la superposició de capes que afegeixen semàntica d'acord amb una sintaxi senzilla que defineix quines dades es volen visualitzar i com. Usant paquets com ggplot2 es poden crear gràfics de qualsevol mena (figura 5), com es pot veure en la galeria d'exemples[26] que també manté la comunitat.

D'altra banda, és destacable l'existència d'eines com ara Gephi,[27] un programari que facilita la manipulació i visualització de grafs, estructures matemàtiques que permeten mostrar relacions entre elements, així com els seus atributs (figura 6). És un programari molt usat en l'entorn acadèmic i per periodistes d'investigació, ja que permet analitzar dades provinents de xarxes socials com Twitter i Facebook, identificar comunitats, els elements més o menys importants de la xarxa, o la densitat de les relacions entre elements de manera visual.

Figura 6. Ús de Gephi per a la visualització de grafs.

Òbviament hi ha moltes altres eines per a generar gràfics a partir de dades, normalment en forma tabular, però també d'una expressió matemàtica. D'una llarga llista[28] se'n pot destacar: Matlab, un programari propietari molt usat en l'àmbit de l'anàlisi de dades matricials i de senyals, així com la seva versió oberta, anomenada GNU Octave; Mathematica, que incorpora capacitats algebraiques simbòliques; i finalment, Orange,[29] una opció nova i molt interessant per a l'exploració visual de dades mitjançant programació visual.

3.2 Llenguatges de programació per a crear gràfics

Un pas més enllà el representa la utilització de llenguatges de programació específics per a la creació de gràfics. Això permet tenir un control total de la visualització creada, però, òbviament, amb el cost d'haver de programar-ne tots els detalls. El nivell d'abstracció és el píxel, disposant també de gràfiques bàsiques primitives com ara la línia i el polígon. D'altra banda, el grau d'interacció és potencialment molt elevat, tot i que resulta molt complicat haver de programar totes les possibilitats.

En l'actualitat, el llenguatge més habitual per a generar gràfics amb ordinador és l'anomenat *Processing*.[30] Aquest llenguatge permet la manipulació d'un espai virtual (llenç o *canvas*) que es pot traslladar a la pantalla de manera total o parcial, permetent la creació d'imatges i el control total de la interacció amb altres dispositius d'entrada o sortida. Processing ha esdevingut un estàndard *de facto* per a tota una comunitat de creadors amb perfil no tecnològic, però que desitgen usar l'ordinador com a eina creativa més enllà de la utilització d'eines tanca-

25. Leland Wilkinson, *The grammar of graphics*, New York: Springer, 2006.
26. <*http://www.r-graph-gallery.com/*>
27. Mathieu Bastian; Sebastien 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, p. 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, no. 1 (2013), p. 2349-2353.
30. Casey Reas; Ben Fry, *Processing: a programming handbook for visual designers and artists*, Cambridge: MIT Press, cop. 2007.

des.[31] Processing també s'ha fet servir per a visualitzar dades. Són destacables els exemples d'Aaron Koblin, que fan servir dades dels vols comercials (figura 7), o de Brendan Dawes, amb el seu projecte Cinema Redux, que fa servir els principis de Manovich[32] i mostra les dades directament.

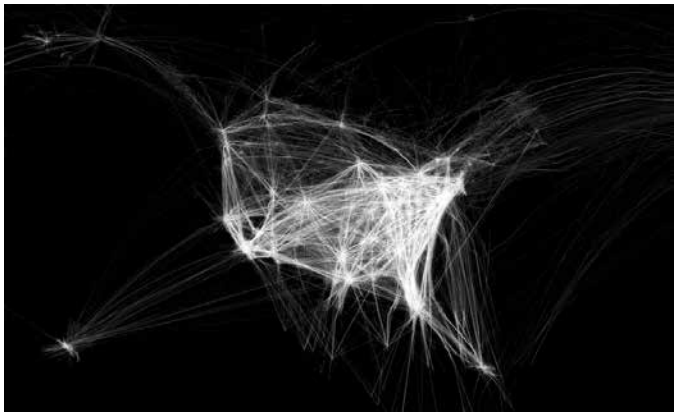

Figura 7. Visualització dels vols sobre el territori dels Estats Units <*http://www.aaronkoblin.com/work/flightpatterns/*>.

El que resulta interessant de la visualització d'Aaron Koblin és que, només fent servir la posició dels avions comercials al llarg del temps mentre segueixen una ruta entre dues ciutats (de fet, entre dos aeroports), es pot identificar la forma del país (en aquest cas, els Estats Units), les zones amb més densitat de població i també zones fosques on no sobrevola cap avió, sigui per la raó que sigui. Tots els aeroports propers a les grans ciutats nord-americanes també es veuen clarament. En aquest cas la interacció es limita a poder reproduir el moviment dels avions en el temps i a poder fer zoom per a mostrar el detall al voltant dels aeroports usant la codificació de colors que ha triat Koblin: blanc quan l'avió és a terra i de colors (en funció de la companyia, o bé blau en general) quan vola. Com es pot veure en el vídeo[33] creat pel mateix Koblin, les dades permeten deduir l'estructura de pistes d'aterratge i enlairament de cada aeroport.

El concepte de *llenç* no és exclusiu del Processing. En el món de la programació web també es poden generar gràfics en un llenç incrustat en una pàgina web que es visualitza mitjançant un navegador, sense necessitat de cap programari específic. Es tracta d'una regió definida com un element HTML en què és possible dibuixar mitjançant gràfiques primitives bàsiques (punts, línies, etc.), però també amb elements més complexos com ara polígons, text i imatges (per exemple, icones).

De fet, totes dues opcions estan convergint, ja que Processing també es troba disponible per a executar-se en línia, com una pàgina web. Al principi s'executava amb processing.js, que és una traducció de codi Processing per a executar-se com a codi JavaScript, i més recentment amb el que es coneix com a p5.js, una llibreria JavaScript per a generar gràfics dins d'un document HTML, com una nova reinterpretació de la idea original proposada per Processing. En el món dels llenguatges de programació, l'elecció d'una opció o l'altra depèn de molts factors, tot i que l'ús intensiu de les possibilitats del llenguatge HTML fa de p5.js una opció més interessant per als programadors web, mentre que Processing és una plataforma excel·lent per a iniciar-se en la programació i la creació de gràfics.

Tot i les possibilitats que ofereix poder manipular una visualització de dades a escala de píxel, la necessitat d'haver de descriure la visualització com un procés que s'executa pas a pas fa que aquesta opció quedi reservada a usuaris amb un cert nivell de coneixements de programació, normalment quan cap de les altres opcions és suficient o, senzillament, es volen explorar visualitzacions

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/*>

radicalment diferents de les tradicionals. Com que molts cops es tracta d'una programació *ad hoc* per a un projecte de visualització concret, generalment aquesta opció és la que es fa servir menys.

3.3 Llibreries

Seguint amb la programació web, una opció intermèdia entre les calculadores gràfiques i l'ús de llenguatges de programació és la utilització de llibreries JavaScript que proporcionen una capa d'abstracció entre l'usuari (i les seves dades) i el llenç on es mostrarà la visualització, de manera que no es treballa a escala de píxel o de gràfica primitiva, sinó a un nivell superior, manipulant conceptes senzills com ara diagrames i configuracions més complexes (*layouts*). La interacció també és més senzilla que en el cas dels llenguatges de programació, ja que la majoria de llibreries incorporen un seguit d'opcions per a capturar els esdeveniments d'alt nivell produïts per l'usuari, com ara l'ús del ratolí i el teclat, i la cerca de continguts.

Igual que en el programari per a crear visualitzacions, hi ha una gran quantitat de llibreries JavaScript que permeten incorporar gràfics en una pàgina web per a visualitzar dades. Entre d'altres, es poden destacar: InfoVis[34] (avui dia obsoleta i superada); Raphaël,[35] orientada a crear diagrames senzills; sigmajs,[36] orientada a la visualització de grafs; o Leaflet,[37] que permet crear mapes i superposar-hi dades. Les llibreries han anat evolucionant juntament amb el llenguatge HTML, els fulls d'estil CSS, el model d'objectes de document DOM i el llenguatge de gràfics vectorials SVG, integrant les diferents capes que componen una pàgina web, però mantenint l'accés individual als elements que la componen, cosa que facilita la creació de visualitzacions dinàmiques.

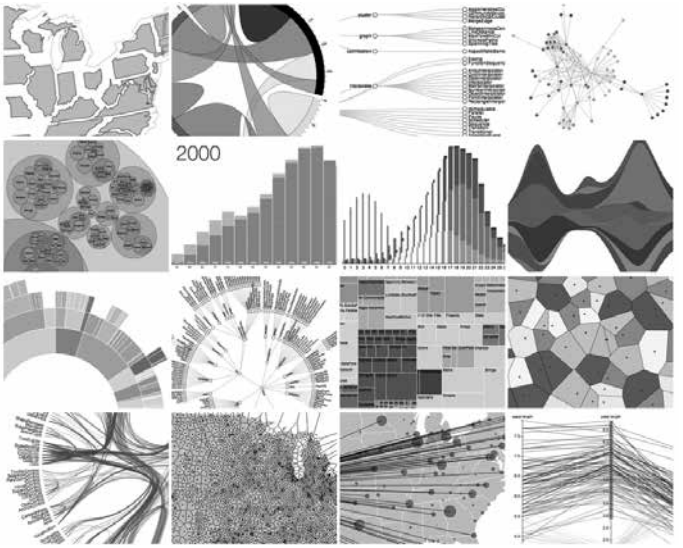

Figura 8. Galeria de configuracions (*layouts*) creades amb D3.

Fruit d'aquesta evolució, l'any 2011 es va donar a conèixer la primera versió de D3 (acrònim de Data-Driven Documents), una llibreria que també s'ha escrit en JavaScript per a proporcionar un control total sobre tots els elements que componen una pàgina web i el seu lligam amb les dades que es volen visualitzar.[38] D3 combina una estètica molt bona i altament configurable amb un elevat potencial de funcionalitats, oferint múltiples configuracions predeterminades en forma de *layouts* (figura 8). D3 es va popularitzar quan Mick Bostock va fitxar pel diari *The New York Times* amb l'objectiu de crear visualitzacions noves i interactives per explicar històries mitjançant dades. Una de les més destacables és la cobertura i l'anàlisi dels resultats de les eleccions als Estats Units del 2014, com ara les

34. <*https://philogb.github.io/jit/*>
35. <*http://dmitrybaranovskiy.github.io/raphael/*>
36. <*http://sigmajs.org/*>
37. <*http://leafletjs.com/*>
38. Michael Bostock; Vadim Ogievetsky; Jeffrey Heer, «D³ data-driven documents», *IEEE transactions on visualization and computer graphics*, vol. 17, no. 12 (2011), p. 2301-2309.

paraules usades per cada candidat en els discursos,[39] combinant gràfics amb textos i descriptors estadístics (figura 9).

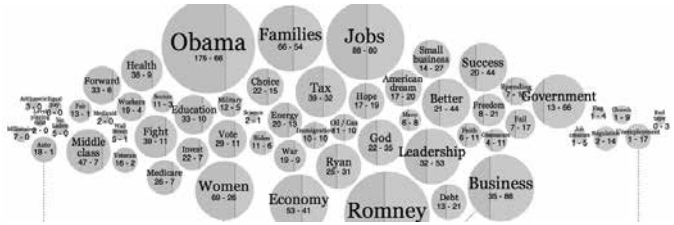

Figura 9. Paraules usades en els discursos de les eleccions als Estats Units del 2014, ordenades per mida i color (blau, demòcrata; vermell, republicà) (Bostock; Ogievetsky; Heer, 2011).

El principal avantatge de D3 és la interactivitat integrada en la visualització, cosa que permet fer les operacions bàsiques definides per Shneiderman[40] de manera senzilla. Aquesta interacció, combinada amb la gran varietat de configuracions disponibles per a visualitzar dades, facilita la manipulació de conjunts de dades un cop s'ha decidit quina configuració o combinació de configuracions és la més adequada. En aquest sentit, tot i la complexitat interna i les dificultats que comporta crear visualitzacions de dades en D3, és possible utilitzar-lo com si fos una caixa negra,[41] de manera que si les dades es troben en un format concret, es pot reaprofitar una visualització existent per mostrar-les, i només cal fer petits canvis relatius en el codi i, especialment, en els fulls d'estil que determinen els aspectes estètics de la visualització. La galeria[42] d'exemples de visualitzacions de dades creades en D3 és un molt bon punt de partida per a comprovar les possibilitats que ofereix aquesta llibreria.

3.4 Entorns gràfics

Finalment, la darrera opció correspon a la sofisticació del concepte de calculadora gràfica, que proporciona veritables entorns gràfics que donen suport a tot el procés de manipulació, preprocessament i visualització de dades. Darrerament han aparegut moltes opcions comercials que competeixen per a proporcionar una solució completa, no només per a la visualització de dades, sinó també per a la presa de decisions en un entorn de negoci, i la visualització de dades és un aspecte més integrat en l'eina. La consultora tecnològica nord-americana Gartner fa un estudi[43] anual de les millors eines d'intel·ligència de negoci i analítica del mercat. D'ençà de ja fa uns quants anys, Tableau està posicionada com una de les millors eines en aquest sector, i aquest darrer any s'ha situat entre les tres líders, juntament amb Qlik i amb els serveis oferts per Microsoft Cloud (entre els quals destaca especialment Power BI).

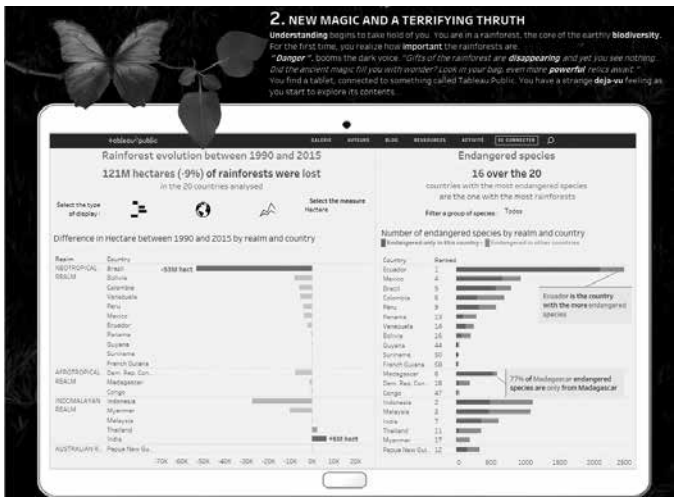

Figura 10. Exemple de narració visual amb Tableau <*https://public.tableau.com/en-us/s/gallery/tale-rainforest*>.

39. <*http://www.nytimes.com/interactive/2012/09/06/us/politics/convention-word-counts.html*>
40. Ben Shneiderman, *op. cit.*
41. <*http://oer.uoc.edu/VIS/D3/*>
42. <*https://github.com/d3/d3/wiki/gallery*>
43. <*https://www.gartner.com/doc/reprints?id=1-3RTAT4N&ct=170124&st=sb*>

De fet, si només es tenen en compte les funcionalitats per a crear visualitzacions de dades, una de les opcions més populars avui dia és Tableau, una eina creada a partir de la recerca en el departament d'informàtica de la Universitat de Stanford. Tableau permet crear diagrames a partir d'un o més conjunts de dades, afegir capes amb interactivitat per a fer les operacions típiques de filtratge, selecció, etc., i construir veritables panells de control (també anomenats quadres de comandament o *dashboards*) que serveixen per a narrar històries a partir de les dades (figura 10). Tableau permet, per mitjà de simples interaccions, tenir diferents perspectives sobre un conjunt de dades. A més, ens ofereix nombroses tècniques d'anàlisi i elements estadístics que poden incorporar-se a les visualitzacions creades per a entendre millor les dades i extreure coneixement a partir de l'ús de representacions gràfiques.

Una altra opció similar a Tableau és Quadrigram, que intenta combinar el millor de dos mons. Per una banda, un entorn potent per a la creació de visualitzacions de dades i, per l'altra, l'opció de generar codi obert per poder exportar la visualització com si fos una pàgina web. La filosofia de Quadrigram es basa en tres principis: el de no linealitat, ja que les idees que donen suport a la visualització de dades s'afegeixen a mesura que es van creant i desenvolupant, el principi d'iteració, ja que el coneixement que aporta la visualització es construeix repetint l'esquema definit per Keim, et al.[44] i, finalment, el principi que considera les dades com un material viu, que evoluciona constantment, un aspecte que la visualització ha de poder capturar.

4. Conclusions

La visualització de dades és un àmbit que darrerament s'ha vist impulsat per la gran quantitat de dades que esperen ser analitzades, la capacitat computacional disponible i també per la disponibilitat d'eines per a la seva manipulació, deixant enrere la idea de les visualitzacions de dades com a simples resums gràfics estàtics. En l'actualitat, una visualització de dades pot incorporar un elevat grau d'interactivitat de manera que permeti manipular les dades directament, integrant operacions bàsiques com ara la selecció i el filtratge, així com d'altres que permetin extreure'n coneixement mitjançant una primera inspecció visual.

Sense substituir l'anàlisi estadística clàssica de qualsevol estudi que faci servir dades per a comprendre la realitat i prendre decisions, la visualització de dades pot aportar coneixement sobre el problema que cal resoldre d'una manera senzilla i eficient alhora.

En aquest sentit, l'anàlisi visual de dades explota les característiques del sistema visual humà, que és molt eficient per a detectar característiques bàsiques de les dades representades gràficament en forma de tendències, patrons, anomalies, etc. Sense substituir l'anàlisi estadística clàssica de qualsevol estudi que faci servir dades per a comprendre la realitat i prendre decisions, la visualització de dades pot aportar coneixement sobre el problema que cal resoldre d'una manera senzilla i eficient alhora.

Per a crear visualitzacions de dades hi ha quatre aproximacions diferents, però complementàries, definides en funció de dues dimensions: per una banda, el nivell d'abstracció dels elements que componen la visualització, que va des del píxel fins al concepte de panell de control, i per l'altra, el grau d'interactivitat, que pot anar des de visualitzacions estàtiques fins a veritables interfícies dinàmiques en què l'observador participa plenament. No hi ha cap eina universal per a visualitzar un conjunt de dades qualsevol, sinó que en funció dels objectius de la visualització serà necessari triar entre

44. Daniel Keim, *op. cit.*

solucions *ad hoc* programades des de l'inici, o bé optar per la creació de visualitzacions estàndard mitjançant eines més o menys complexes i amb més o menys grau d'interactivitat. Sigui com sigui, l'ús de la web com a plataforma per a la creació i difusió d'aplicacions amb un gran component visual i amb possibilitats d'accedir a grans volums de dades i analitzar-les fa que cada cop sigui una opció més interessant i amb més potencial, i cada vegada hi ha més eines que permeten crear visualitzacions tant de manera local com per a compartir-les després mitjançant la web.

El futur de l'anàlisi visual de dades implica fer avançar el model definit per Keim, *et al.*,[45] introduint mecanismes automàtics per al reconeixement de la natura de les dades, igual que en altres àmbits lligats a la intel·ligència artificial, en què tendències com ara l'aprenentatge profund (*deep learning*) s'estan imposant per a extreure coneixement de les dades sense haver de pressuposar un model amb anterioritat. La gran quantitat de dades i la creixent capacitat de càlcul disponible per a analitzar-les fa pensar que la visualització de dades serà un àmbit de coneixement en plena expansió els pròxims anys.

Bibliografia

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, p. 361-362.

Bengio, Yoshua. «Learning deep architectures for AI». *Foundations and trends® in Machine Learning*, vol. 2, no. 1 (2009), p. 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. «D^3 data-driven documents», *IEEE transactions on visualization and computer graphics*, vol. 17, no. 12 (2011), p. 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, no. 1 (2013), p. 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, p. 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 fur Klassifikation e. v., University of Dortmund, March 9-11, 2004*. New York: Springer, 2005, p. 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, no. 1358 (1997), p. 1121-1127.

45. *Íbid.*

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

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, no. 4 (2000), p. 90-107.

MANOVICH, Lev. «What is visualization?», *Poetess Archive Journal*, vol. 2, no 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], p. 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, no. 3 (2017), art. no. 19.

RACINE, Jeff. «Gnuplot 4.0: a portable interactive plotting utility». *Journal of Applied Econometrics*, vol. 21, no. 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: *PolicyViz* [en línia]: *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 data type taxonomy for information visualizations». En: IEEE Symposium on Visula Languages (1996: Washinton). *VL'96: proceedings of the 1996 IEEE Symposium on Visual Languages*. Washington: IEEE Computer Society, 1996, p. 336-343.

SUN, Guo-Dao, *et al*. «A survey of visual analytics techniques and applications: state-of-the-art research and future challenges». *Journal of Computer Science and Technology*, vol. 28, no. 5 (2013), p. 852-867.

THOMAS, James J.; COOK, Kristin A. «A visual analytics agenda». *IEEE computer graphics and applications*, vol. 26, no. 1 (2006), p. 10-13.

TUKEY, John W. «The future of data analysis». *The Annals of Mathematical Statistics*, vol. 33, no. 1 (1962), p. 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. ■