

REUTILIZACIÓN DE LA INFORMACIÓN DEL SECTOR PÚBLICO

USO DE HERRAMIENTAS BÁSICAS DE TRATAMIENTO DE DATOS



ÍNDICE

ÍNDICE.....	2
OBJETIVOS DIDÁCTICOS.....	3
INTRODUCCIÓN.....	3
CONTENIDOS.....	4
1. ¿QUÉ PUEDE HACERSE CON LOS DATOS PÚBLICOS?	5
1.1 Productos	5
1.2 Aplicaciones.....	6
1.3 Servicios.....	7
2. HERRAMIENTAS.....	8
TIPOS DE HERRAMIENTA.....	9
2.1 Scraping (extracción).....	9
2.1.1 Selección de herramientas: ParseHub, Import.io, PDFTables y Tabula	11
2.2 Tratamiento de datos.....	13
2.2.1 Selección de Herramientas: Open Refine y Data Wrangler	13
2.3 Análisis Estadístico	15
2.3.1 Selección de herramientas: RStudio y Matlab	16
2.4 Visualización.....	17
TIPOS DE VISUALIZACIONES	18
2.4.1 Selección de herramientas genéricas de visualización: Fusion Tables, Tableau Public, Datawrapper y RAW.....	22
2.4.2 Selección de herramientas de visualización geoespacial: CartoDB, Google MyMaps y Crowdmap.....	24
2.4.3 Selección de herramientas de visualización temporal: TimeFlow, TimelineJS	26
2.5 Generación de redes de datos relacionados.....	28
2.5.1 Selección de herramientas de redes de grafos: Fusion Tables y Gephi	29
RESUMEN.....	31

OBJETIVOS DIDÁCTICOS

Comenzamos presentando los **Objetivos didácticos**:

- ✓ Obtener una visión general de las diversas formas mediante las que se pueden aprovechar los conjuntos de datos utilizando diferentes técnicas de tratamiento de datos.
- ✓ Presentar algunas de las herramientas existentes en el mercado para las diferentes técnicas de tratamiento de datos.
- ✓ Conocer las principales características y utilidades de las herramientas de tratamiento de datos y explorar algunas de las posibilidades que ofrecen.



El mejor uso que pueda darse a tus datos se le ocurrirá a otra persona.

Rufus Pollock. Fundador de la Open Knowledge Foundation:
<http://rufuspollock.org/misc/>

INTRODUCCIÓN

La posibilidad de que tanto la ciudadanía como las empresas puedan acceder y utilizar cada vez mayor volumen de datos de una forma cada vez más amplia y variada, está facilitando el nacimiento de un nuevo mercado: el de la reutilización de la información pública.

Los **agentes infomediarios** toman los datos públicos y los reutilizan (recopilar y tratar dichos datos) para fines distintos de los de la Administración, generando con ellos productos y servicios de valor añadido.

Con Internet han surgido un gran número de herramientas destinadas al análisis, tratamiento y visualización de la información, lo que facilita la creación de dichos productos y servicios.

En esta unidad vamos a presentar una selección de herramientas de tratamiento, análisis y visualización de datos, para optimizar el uso de los datos abiertos gubernamentales y acercarnos a la actividad **infomediaria**.

CONTENIDOS

1. ¿QUÉ PUEDE HACERSE CON LOS DATOS PÚBLICOS?

Transformación de la información o datos públicos en forma de **productos, servicios o aplicaciones** por parte de los agentes infomediarios.

2. HERRAMIENTAS

Selección de algunas **herramientas de mercado** que se utilizan en las distintas fases de la elaboración de productos, servicios o aplicaciones que reutilizan datos públicos: **scraping, tratamiento de datos, análisis estadístico y visualización**.

1. ¿QUÉ PUEDE HACERSE CON LOS DATOS PÚBLICOS?

De acuerdo a la clasificación que presenta el **Estudio de Caracterización del Sector Infomediario en España (2014, ONTSI)**, los agentes infomediarios transforman la información o datos públicos en forma de **productos, servicios o aplicaciones**.

[!\[\]\(e2376d476d06eb31946dc01a69a4403a_img.jpg\) Estudio de Caracterización:
estudio_de_caracterizacion_del_sector_infomediario_en_espana_2014_parte_i_publica.pdf](#)

PRODUCTOS

Productos: Los productos basados en datos abiertos comprenden la reutilización de una o varias fuentes de información pública para generar utilidades de valor.

APLICACIONES

Las aplicaciones son herramientas informáticas (software) desarrolladas para una utilidad específica. Pueden estar diseñadas para uso general o bien, ser desarrolladas a medida para resolver un problema específico de un cliente.

SERVICIOS

A diferencia de los productos o aplicaciones, con los que se generan nuevas formas creativas para el uso de los datos, los servicios sobre datos abiertos se utilizan para consultoría sobre materias específicas.

Vamos a conocer a continuación los principales tipos de cada uno de ellos.

1.1 Productos

Los principales tipos de productos que son generados por los agentes infomediarios en España son:

DATOS TRATADOS

Son productos basados en el tratamiento de datos públicos para mejorar su calidad o dotarlos de un mayor contexto, cruzando fuentes de datos de distinto tipo. El 71% de las empresas generan productos de este tipo. El modelo de ingresos para productos sobre datos tratados suele ser el pago por suscripción.

MAPAS

Los mapas, y elementos gráficos afines, son elaborados por un 41% de las empresas del sector infomediario. Destaca que los productos sobre mapas son muy demandados por las Administraciones Públicas.

PUBLICACIONES

Las publicaciones son productos que ofrecen resúmenes o análisis sobre un área de información específica. Son elaboradas por un 28% de las empresas infomediarias. Suelen seguir un modelo de ingresos *freemium*: determinados contenidos parciales se ofrecen de forma gratuita debiendo realizarse un pago para obtener por ejemplo, la descarga completa de los contenidos.

DATOS EN BRUTO

Supone agregar diferentes conjuntos de datos gubernamentales publicados en distintas fuentes, y construir con ellos una base de datos que se ofrece tal cual. El 19% de las empresas generan productos de este tipo. La venta de datos en bruto se vincula a modelos de ingresos gratuitos sin restricciones, lo que significa que las empresas utilizarían el acceso a estos datos para captar clientes para otros productos diferentes (como serían los datos tratados), o para generar ingresos alternativos.

1.2 Aplicaciones

Los principales tipos de aplicaciones que crean los agentes infomediarios reutilizando información pública son:

SOFTWARE CLIENTE

Utilidades que consumen datos o información pública mediante una conexión directa al servidor de datos o información de la Administración Pública. Es el principal formato de comercialización de aplicaciones, y son generadas por el 28% de las empresas. El pago por uso es el modelo de ingresos más frecuente, aunque también se generan ingresos por servicios de soporte o mantenimiento.

APLICACIONES DISPOSITIVOS MÓVILES

Un 20% de las empresas realizan aplicaciones de este tipo. Son frecuentes los modelos de uso gratuito sin restricciones.

INFORMACIÓN GPS

Las empresas que ofertan información GPS suponen un 17% del total de las empresas infomediarias que desarrollan aplicaciones.

ALERTA SMS/CORREO ELECTRÓNICO

Notifican a suscriptores con recordatorios específicos sobre la disponibilidad de determinados contenidos (Boletines oficiales, censos y directorios, normas y jurisprudencia). El 16% de las empresas realizan este tipo de aplicaciones que reutilizan datos o información pública.

1.3 Servicios

A diferencia de los productos o aplicaciones, con los que se generan nuevas formas creativas para el uso de los datos, los servicios sobre datos abiertos se utilizan para consultoría sobre materias específicas.

Los principales tipos de servicios basados en datos abiertos que ofrecen los agentes infomediarios en España son:

INFORMES PERSONALIZADOS

Reutilización de información sobre la base de censos y directorios, informes o boletines oficiales para ofrecer información personalizada, por ejemplo informes de solvencia crediticia o empresarial. El 55% de las empresas generan servicios infomediarios de este tipo, donde el modelo de ingresos suele ser de pago por informe.

ASESORAMIENTO

Servicios de asesoría, investigación y estudios de mercado que se desarrollan llave en mano para los clientes, basados en el tratamiento de la información y orientados a dar soporte en los procesos de toma de decisiones. Son elaborados por un 44% de las empresas y la forma de provisión de servicios suele ser a través del pago por cada servicio.

COMPARATIVAS

Servicios de comparación de precios en internet de distintos productos o servicios como hoteles, seguros, coches, etc. La información reutilizada es esencialmente privada. Un 43% de las empresas realiza productos de este tipo.

CLIPPING O RECORTE

El *clipping* sobre datos públicos se realiza recopilando (recortes) información o datos públicos de forma sistemática para su posterior oferta, o a petición de un cliente. Un 12% de las empresas realiza servicios sobre información pública son de este tipo.

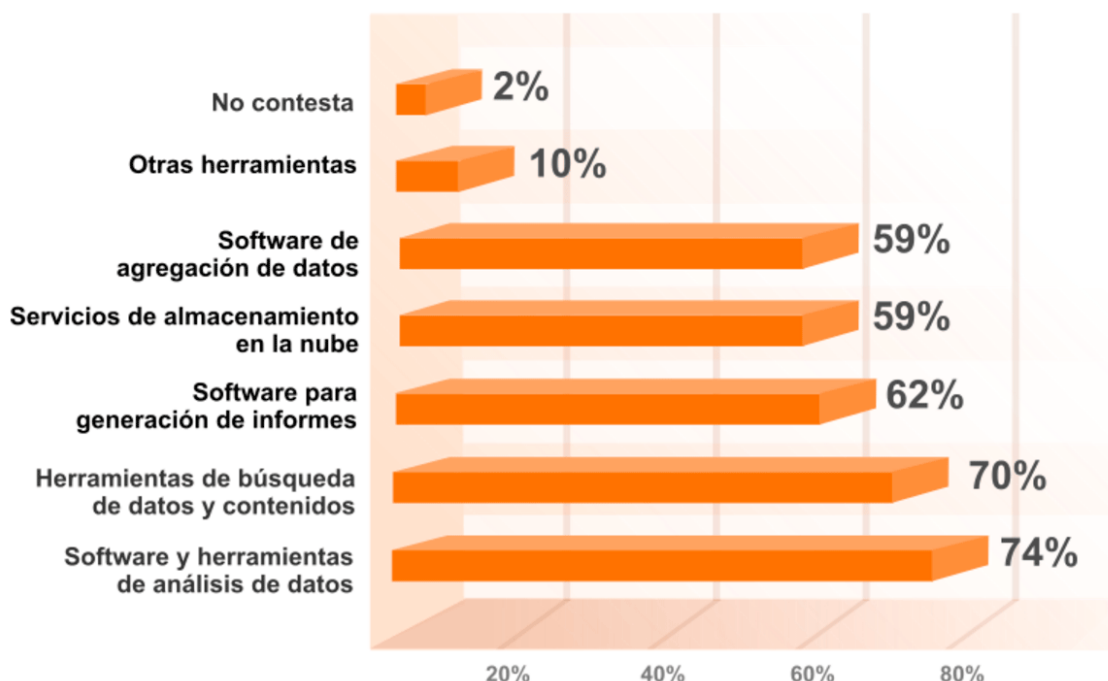
2. HERRAMIENTAS

Existe todo un ecosistema de herramientas que los agentes infomediarios emplean para la elaboración de productos, servicios o aplicaciones basados en datos abiertos.

El **Estudio de Caracterización del Sector Infomediario en España** destaca que:

- ✓ El 74% de ellos utilizan software y herramientas de análisis de datos.
- ✓ El 70% herramientas de búsqueda de datos/contenidos.
- ✓ Un 62% herramientas de generación de informes, entre otros.

HERRAMIENTAS UTILIZADAS EN LA GENERACIÓN DE PRODUCTOS



+ MÁS INFORMACIÓN

El Estudio de Caracterización del Sector Infomediario en España (2014, ONTSI) destaca que los agentes infomediarios emplean para la elaboración de sus productos, servicios o aplicaciones, tanto herramientas de mercado como *herramientas ad hoc orientadas al tratamiento de los datos públicos*.

Fuente:

[estudio de caracterizacion del sector infomediario en espana 2014 parte i publica.pdf](#)

En esta unidad vamos a presentar una selección de herramientas de mercado que se pueden utilizar en las distintas fases de la elaboración de productos, servicios o aplicaciones que reutilizan datos públicos.

TIPOS DE HERRAMIENTA

SCRAPING

Objeto: Extracción

Extracción de datos desde fuentes de información en formatos no apropiados para la reutilización.

TRATAMIENTO

Objeto: Transformación

Convierten la estructura de un conjunto de datos a otra diferente que nos resulte más apta para el aprovechamiento posterior en forma de productos, servicios o aplicaciones.

ANÁLISIS ESTADÍSTICO

Objeto: Análisis

Proceso de transformación de la información en conocimiento mediante análisis descriptivo, inferencial o predictivo, entre otros.

VISUALIZACIÓN

Objeto: Presentación

Presentar los productos, servicios o aplicaciones en formas aprovechables para el usuario final como mapas, gráficos o líneas del tiempo, entre otros.

2.1 Scraping (extracción)

El *data scraping*, *raspado de datos* o *escrapeo de datos* es un conjunto de técnicas de programación con las que es posible extraer la información (o los datos), sea cual sea la presentación, de un documento mediante ingeniería inversa. Fuente: https://es.wikipedia.org/wiki/Web_scraping

Estas técnicas simulan la exploración humana de un documento o sitio web, **extrayendo la información que contiene y presentándola en un formato**



reutilizable¹.

Las técnicas de *scraping* tienen algunas limitaciones, por lo que se utilizan como **herramientas de último recurso**, es decir, cuando la fuente de información no proporciona los datos en formatos reutilizables, y aplicar estas técnicas es la única alternativa.

Los **inconvenientes** más habituales son:

FORMATO

Cuando los datos que se presentan en el documento origen no son listados o tablas, el resultado tras el *scraping* queda poco estructurado, complicando el trabajo de procesado de los datos. Para estos casos, los asistentes de *scraping* como los que presentamos en el siguiente punto, no son del todo útiles, al requerir habilidades en desarrollo software para el uso de las librerías de programación de *scraping*.

PAGINACIÓN

Cuando los resultados en el documento origen están repartidos en diversas páginas, el *scraping* resulta complejo y requiere en ocasiones de conocimientos de programación para poder realizar la extracción.

CODIFICACIÓN DE CARACTERES

Son necesarias herramientas (o una combinación de ellas) capaces de manejar un juego de caracteres amplio (tildes, eñes, caracteres extraños), pues en caso contrario, la codificación de los mismos en el documento origen puede dar lugar a extracciones con errores.

+ MÁS INFORMACIÓN

Las técnicas de *data scraping* incluyen las herramientas de **Web Scraping**, con las que se obtienen datos estructurados a partir de páginas web, y las herramientas de **PDF Scraping**, que extraen datos a partir de tablas en documentos PDF.

¹ Antes de poder aplicar técnicas de *scraping* para obtener datos de un sitio web es necesario conocer cuáles son las licencias o condiciones de uso aplicables en dicho sitio.

2.1.1 Selección de herramientas: ParseHub, Import.io, PDFTables y Tabula

Se han seleccionado una serie de herramientas para realizar *Web Scraping* y *PDF Scraping* en base a criterios de sencillez de uso (sin necesidad de programar), tipo de aplicaciones (web o multiplataforma) y si son de licencia libre o si permiten un uso gratuito de alguna de sus versiones.

Parsehub

parsehub.com | [Tutoriales: quickstart](#) | [Vídeos: extract data from lists](#)

TIPO Web Scraping

TECNOLOGÍA Aplicación web y API.

CONDICIONES DE USO Privativo. Freemium.

AUTOR Debuggex, Inc.

Asistente visual para obtener los datos presentes en web de terceros en forma de datos estructurados y como API de consulta. Su plan FREE permite hasta 5 proyectos de scraping.



Import.io

import.io | [Galería: showcase](#) | [Tutoriales: knowledgebase](#)

TIPO Web Scraping

TECNOLOGÍA Aplicación web,

escritorio (Java) y API.

CONDICIONES DE USO De uso libre.

AUTOR Import.io Ltd.

Conversor automático de páginas web en datos estructurados. Tiene una versión web y otra de escritorio multiplataforma. La aplicación de escritorio permite realizar *scrapings* de forma ilimitada.



PDFTables

pdftables.com | [Tutoriales: convert-pdf-to-excel](#)

TIPO PDF Scraping

TECNOLOGÍA Aplicación web.

CONDICIONES DE USO Privativo. Freemium.

AUTOR ScraperWiki.

Conversor automático de ficheros PDF a datos estructurados. El plan gratuito tan solo soporta *scraping* para 50 páginas o 2 documentos. Los planes de pago proveen de una API para automatizar la conversión de ficheros.



Tabula

tabula.technology | [Videos: www.youtube.com/watch?v=yEyQqrJvSHk](https://www.youtube.com/watch?v=yEyQqrJvSHk)

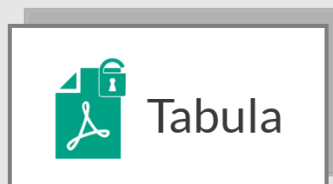
TIPO PDF Scraping

TECNOLOGÍA Aplicación de escritorio (Java).

CONDICIONES DE USO Libre (MIT License).

AUTOR Manuel Aristarán

Asistente para convertir tablas en ficheros PDF, a datos estructurados. Es una aplicación de escritorio multiplataforma de licencia libre.



2.2 Tratamiento de datos

En ocasiones, los conjuntos de datos representados en tablas no son perfectos, están expresados en diferentes formas, utilizan abreviaturas, contiene errores de codificación, etc., y corregirlos de forma manual no es viable.

Se utilizan entonces unas **herramientas de transformación** de datos para mejorar la calidad de la información y hacerlos **completamente reutilizables**², pues ayudan a filtrar, mejorar y clasificar la información. Fuente: “Decálogo del Reutilizador de Datos del Sector Público” (2014, datos.gob.es)

Para este tratamiento de los datos existe un conjunto de técnicas:

DATA CLEANSING (limpieza de datos)

Corrigen errores en los datos que afectan a la calidad, por ejemplo eliminan caracteres extraños que pueden dificultar la búsqueda o normalizar nombres de ciudades o códigos postales a fin de que todas sigan la misma nomenclatura. Fuente: “Data Cleansing” en Wikipedia.

DATA WRANGLING (transformación de datos)

Convierten la estructura de los datos en otra diferente, más apta para su utilización posterior en forma de visualización o análisis estadístico. Fuente: “Data Wrangling” en Wikipedia

RECORD LINKAGE (enlace de datos)

Vinculan registros entre conjuntos de datos para relacionarlos. Por ejemplo, relacionar las adjudicaciones (conjunto de datos de contratación pública) con las partidas presupuestarias correspondientes (conjunto de datos de presupuestos). Fuente: “Record Linkage” en Wikipedia

El uso de estas herramientas se realiza como paso previo al **aprovechamiento de la información (en forma de, por ejemplo, análisis o visualizaciones)**, ya que es necesario que dicha información se encuentre previamente en formatos reutilizables y con cierta estructura, como una tabla en un fichero HTML, un fichero de texto separado por comas o tabuladores (CSV, TSV), o una hoja de cálculo (XLS, ODS).

2.2.1 Selección de Herramientas: Open Refine y Data Wrangler

A continuación se han seleccionado dos herramientas que permiten realizar tratamiento de datos. Los **criterios de selección** aplicados responden a: uso sencillo sin necesidad de programar, aplicaciones web o multiplataforma, y que sean de licencia libre o que permitan un modo gratuito de uso de alguna de sus versiones.

² Traducción desde “Data processing and visualization tools” (2013, datos.gob.es para EPSI Platform).

Open Refine

[open refine](#) | [Documentación: wiki](#) | [Tutoriales: External-Resources](#) | [Vídeos: www.youtube.com/user/GoogleRefine](#)

TECNOLOGÍA Aplicación de escritorio

Multiplataforma.

CONDICIONES DE USO Libre (BSD License).

AUTOR MetaWeb Technologies.



Facilita limpiar y transformar datos desordenados o completarlos haciendo uso de servicios web externos como DBPedia. Permite realizar operaciones de data cleansing (limpieza), data wrangling (transformación), record linkage (enlazado de datos). Open Refine es la versión open source de Google Refine. Google abandonó el desarrollo de Google Refine y lo liberó como software libre. Desde entonces, el proyecto Open Refine mantiene una versión evolucionada de Google Refine.

Data Wrangler

[data wrangler](#) | [Vídeo: https://vimeo.com/19185801](https://vimeo.com/19185801)

TECNOLOGÍA Aplicación web (HTML).

CONDICIONES DE USO Freeware / Propietario.

AUTOR Stanford Visualization Group.



Esta aplicación web es un Asistente visual para la limpieza y transformación de datos. Capaz de exportar en tablas de análisis de datos que aplicaciones como Excel, R o Tableau necesitan. Permite realizar operaciones de *data cleansing* (limpieza), *data wrangling* (transformación).

2.3 Análisis Estadístico

Las herramientas de análisis estadístico permiten realizar exploraciones avanzadas de conjuntos de datos grandes y formular modelos que permitan correlacionar variables o realizar predicciones. Serán, por tanto, de gran valor a la hora de **sacar el máximo partido de los datos realizando análisis complejos**. Fuente: “Decálogo del Reutilizador de Datos del Sector Público” (2014, datos.gob.es)

Actividades que típicamente se realizan con las herramientas de análisis estadístico son:

CLUSTERING (análisis de agrupamiento)

El análisis de agrupamiento o clustering consiste en **agrupar un conjunto de objetos**, de forma que los objetos en el mismo grupo (llamado un cluster) son más similares entre sí (en un sentido u otro) que respecto a los de otros grupos (grupos).

Por ejemplo, agrupar las paradas de autobuses según se encuentren a 0.5 km, 1 km o más de 1 km de distancia entre ellas para analizar qué barrios tienen mejor comunicación que otros. Fuente: “Cluster Analysis” en Wikipedia.

ANÁLISIS DE REGRESIÓN

Proceso estadístico para estimar las relaciones entre variables. Este análisis ayuda a entender cómo cambia el valor típico de una variable dependiente cuando una de las variables independientes se cambia, mientras que el resto se mantiene fijo.

Por ejemplo, si se cuenta con el conjunto de datos adecuado, se podría realizar un análisis de regresión para intentar conocer si lo que más afecta a los retrasos de los autobuses es la existencia de semáforos en su ruta, si el camino tiene pendientes, o la circulación en hora punta. Fuente: “Regression Analysis” en Wikipedia.

ANÁLISIS PREDICTIVO

Se utiliza para analizar datos actuales e históricos y **poder realizar predicciones** acerca del futuro, o acontecimientos no conocidos.

Por ejemplo, en caso de contar con el conjunto de datos adecuado, podría estimarse el retraso que en distintas rutas de autobuses introduciría un nuevo semáforo. Fuente: “Predictive Analysis” en Wikipedia.

El uso de estas herramientas está recomendado cuando se desea diseñar un modelo matemático o estadístico para la realización de simulaciones, predicciones o sistemas de recomendación, por ejemplo. El verdadero aprovechamiento de estas herramientas requiere de conocimientos medios/avanzados en estadística inferencial y probabilidad.

2.3.1 Selección de herramientas: RStudio y Matlab

A continuación mostramos dos herramientas de análisis de datos que han sido seleccionadas por su amplia aceptación por parte de profesionales de la estadística en el mundo académico y empresarial, y por tratarse de aplicaciones de escritorio de fácil instalación.

RStudio

[rstudio](#) | [Galería: gallery](#) | [Tutoriales: online-learning](#)

TECNOLOGÍA Aplicación de escritorio

Multiplataforma.

CONDICIONES DE USO Libre (AGPL v3)

AUTOR RStudio.

RStudio es un entorno para utilizar el lenguaje de programación R, que dispone de todas las funcionalidades necesarias para realizar estadística descriptiva (incluyendo gráficos), inferencial y probabilística. Posee funcionalidades específicas para realizar análisis de agrupamiento, regresión y predictivo.

Matlab

[matlab](#) | [Galería](#) | [Tutoriales](#)

TECNOLOGÍA Aplicación de escritorio

Multiplataforma.

CONDICIONES DE USO Privativa. Trial disponible

AUTOR The MathWorks, Inc.

MatLab es un entorno para utilizar el lenguaje de programación matlab. Permite la realización de cálculo numérico, análisis y visualización de datos, programación y desarrollo de algoritmos. Su funcionalidad es más amplia que R aunque ofrece características similares. Posee funcionalidades específicas para realizar análisis de agrupamiento, regresión y predictivo.

+ MÁS INFORMACIÓN

Cabe destacar que, con el desarrollo de las *tecnologías Big Data*, los análisis de agrupamiento, regresión o análisis y las potenciales aplicaciones como simulaciones, predicciones o recomendaciones descritos en el apartado anterior, se realizan utilizando otro tipo de herramientas, siendo **Hadoop** una de las más populares. El aprovechamiento de estas herramientas requiere de un perfil profesional especializado en *data-science* y una infraestructura de computación adecuada. Las suites de escritorio RStudio o MatLab no han sido diseñadas para trabajar con fuentes de datos *Big Data*, esto es, fuentes de datos de gran volumen y variedad que se actualicen a gran velocidad.

2.4 Visualización

La visualización engloba aquellas técnicas que se utilizan para crear imágenes, diagramas o animaciones para comunicar un mensaje. [Fuente: “Visualization \(computer graphics\)” en Wikipedia](#)

Su objetivo es *la creación de una representación visual que ayuda a comprender un fenómeno complejo*, transformando lo simbólico en geométrico. [Fuente: “Visualization Handbook: Introduction \(preface XIV\)” en Google Books](#).

En el libro *The Visual Display of Quantitative Information* (1983, Edward Tufte), se define la efectividad de las visualizaciones (*graphical displays*) *como ideas complejas que son comunicadas con claridad, precisión y eficiencia, permitiendo el análisis visual, la comparativa y la causalidad*. [Fuente: “Characteristics of effective graphical displays” en Wikipedia](#).

El autor destacaba que, entre otras propiedades, las visualizaciones deben:

- ✓ Mostrar los datos evitando distorsionar lo que tienen que decir.
- ✓ Retar al ojo a comparar.
- ✓ Mostrar datos en distintos niveles de más agregados a más detallados.

+ MÁS INFORMACIÓN

En su página web *Milestones in the history of thematic cartography, statistical graphics and data visualization*, **Michael Friendly** señala lo siguiente:

“El nacimiento del pensamiento estadístico fue también acompañado del ascenso del pensamiento visual [...] varias formas gráficas fueron inventadas para hacer las propiedades numéricas (tendencias, distribuciones, etc.) más fácilmente comunicables y accesibles por medio de la inspección visual.”

Michael Lewis Friendly (1945, Nueva York): profesor de Psicología en la Universidad de York en Ontario, Canadá, y Director del Servicio de Consultoría Estadística, reconocido por sus contribuciones a la historia de datos y visualización de la información.

Edward Rolf Tufte (1942, Kansas City, Misuri, EE. UU.): profesor emérito de la Universidad de Yale, en la que dictó cursos sobre evidencia estadística y diseño de información y de interfaces. Es autor de varios libros sobre visualización de información cuantitativa.

TIPOS DE VISUALIZACIONES

En función del mensaje a comunicar

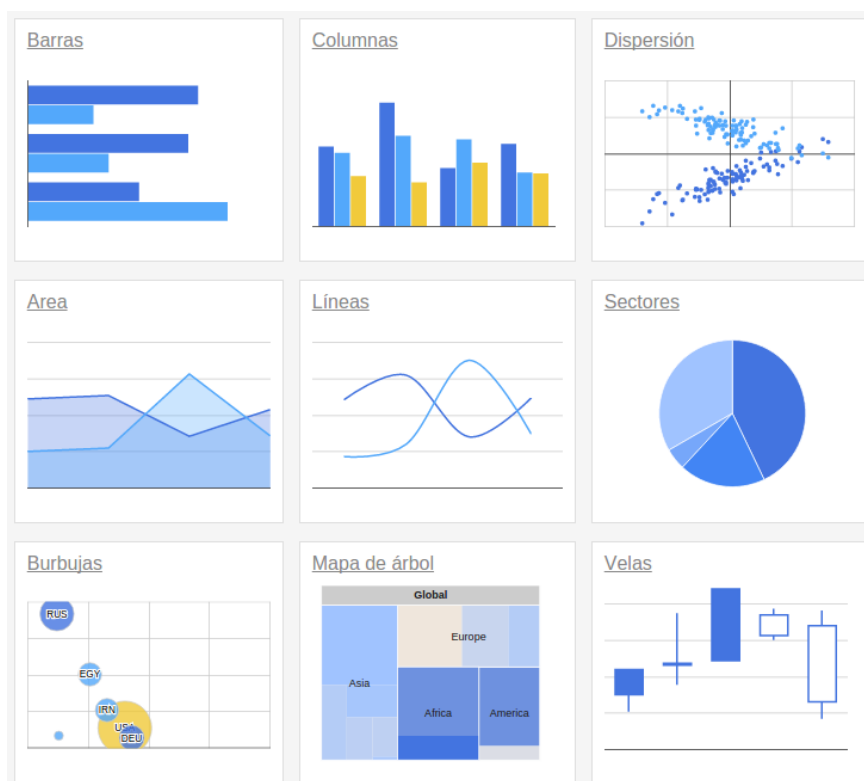
Fuente: “Programa Especializado de Ciencia de Datos”. John Hopkins University, en Coursera.

- **Visualizaciones Exploratorias:** Se construyen sobre un análisis descriptivo de datos y el objetivo es comunicar propiedades (relevancia, relaciones), patrones (tendencias, correlaciones, etc.) o estrategias de modelado de los datos visualizados.
- **Visualizaciones Expositivas:** Se pretende transmitir como mensaje los resultados de un análisis o investigación. El objetivo es comunicar de forma visual la información descubierta como resultado de la investigación.

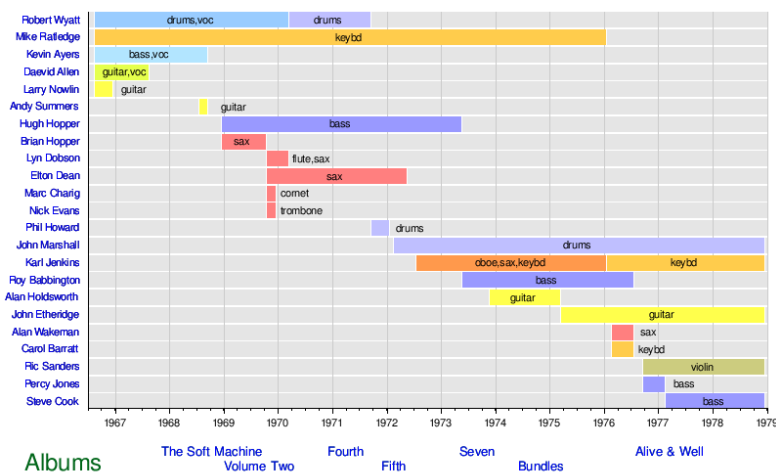
En función del tipo de elemento geométrico para la visualización

Fuente: Clasificación según Display Services en “Data Processing and Visualization Tools” (2013, datos.gob.es para EPSI Platform).

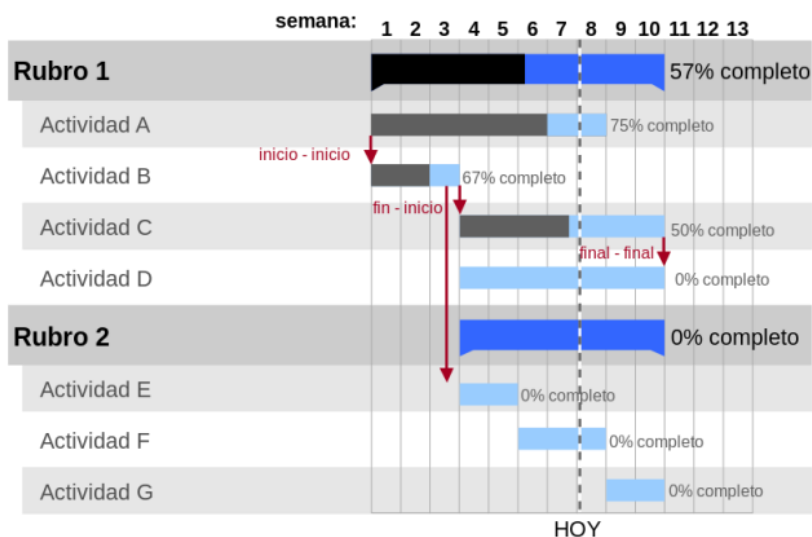
- **Aplicaciones genéricas:** Asisten a la representación visual de datos empleando gráficas típicamente empleadas en la estadística descriptiva tales como: gráficos de barras, de columnas, dispersión, área, líneas o velas.



- **Visualización Temporal:** Visualizaciones especialmente diseñadas para la representación temporal.
 - **Línea del tiempo:** Son representaciones visuales de sucesos ocurridos a lo largo del tiempo, normalmente presentados de forma cronológica. Fuente: [Timeline Soft Machine](#), [Wikipedia](#)



- **Diagrama de Gantt:** Representación de la dedicación prevista para diferentes tareas o actividades a lo largo de un tiempo total determinado. [Fuente: Diagrama de Gantt, Wikipedia](#)

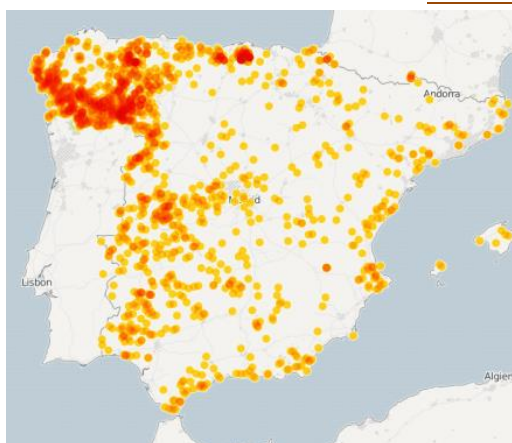


○ **Visualización Geoespacial:** Representación de datos sobre mapas. En función de la geometría empleada para distribuir los datos a lo largo del mapa destacamos los siguientes tipos.

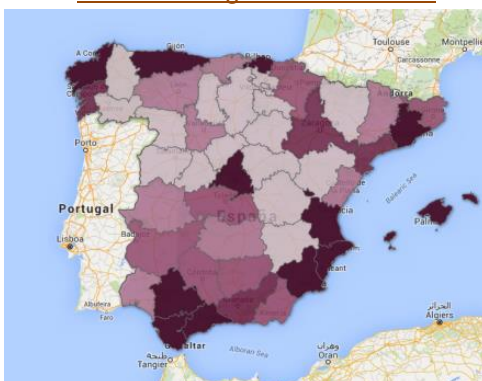
- **Mapa de distribución de puntos (o de densidad de puntos, o de “pines”):** Se basan en una dispersión visual de puntos para mostrar un patrón espacial. Los puntos pueden, además, emplear diferentes símbolos, tamaños y colores para presentar la distribución de otras variables. [Fuente: adoptaunaplaya.org](http://adoptaunaplaya.org)



- **Mapa de calor (isopletras, heat map):** Presenta los valores individuales agrupados en áreas; según la proximidad asigna un color, representando las diferentes proporciones de una variable estadística. Fuente: espanaenllamas.es



- **Mapa de coropletas:** Se emplean áreas sombreadas para presentar las diferentes proporciones de una variable estadística, a menudo coincidiendo con divisiones administrativas. Fuente: [Google Fusion Tables](https://www.google.com/fusiontables/)



2.4.1 Selección de herramientas genéricas de visualización: Fusion Tables, Tableau Public, Datawrapper y RAW.

A continuación mostramos cuatro herramientas para realizar análisis de datos. La selección se basa en la facilidad de manejo (no es necesario programar), la disponibilidad de versiones del software de libre uso y la posibilidad de insertar los resultados en websites externos, por ejemplo, un blog o página personal.

Google Fusion Tables, DataWrapper y RAW ofrecen un amplio abanico de posibilidades para la visualización de datos; desde los empleados habitualmente en la estadística descriptiva, hasta otros más novedosos que se han popularizado con el uso de herramientas en Internet. **DataWrapper y Tableau Public** son ampliamente utilizados por los profesionales del **periodismo de datos**.

Google Fusion Tables

[fusiontables](#) | [Galería](#) | [Vídeos](#) | [Tutoriales](#)

TECNOLOGÍA Aplicación web
(HTML5, Javascript).

CONDICIONES DE USO Freeware / Propietario.


AUTOR Google.

Provee de gráficos de barras, columnas, líneas, sectores, de red, dispersión y mapas, además de varios resúmenes estadísticos. Permite, además, operaciones de enlace de datos (record linkage) y análisis descriptivos. Los gráficos generados pueden ser insertados en páginas web externas. Fusion Tables utiliza la tecnología de Google para asistir en la visualización de conjuntos de datos muy grandes y cuenta con una API para desarrolladores.

Tableau Public

[tableau public](#) | [Galería](#) | [Vídeos](#) | [Tutoriales](#)

TECNOLOGÍA Aplicación de escritorio Windows/Mac.



+tableau+public

CONDICIONES DE USO Freeware / Propietario.

AUTOR Tableau Software Inc.

Permite realizar gráficos de barras, columnas, líneas, dispersión, velas, mapa árbol, burbujas, gantt y mapas, entre otros. Permite, además, operaciones de enlace de datos (*record linkage*) y análisis descriptivo. Los gráficos generados pueden ser insertados en páginas web externas.

Es la **versión gratuita** del software de **Tableau**: solo permite trabajar con datos en formato **hoja de cálculo** y cualquier trabajo requiere ser salvado en la galería pública online.

Datawrapper

[datawrapper](#) | [Galería](#) | [Tutoriales](#)

TECNOLOGÍA Aplicación web (HTML5, Javascript).



ABZV Datawrapper

CONDICIONES DE USO Freeware / Libre (MIT)

AUTOR Journalism++ Cologne GmbH

Permite realizar gráficos de barras, columnas, líneas, sectores y mapas. Facilita, además, ciertas operaciones de limpieza de datos. Los gráficos generados pueden ser insertados en páginas web externas.

La versión web de Datawrapper **no almacena online** los gráficos generados por sus usuarios: para ello cuenta con **planes de pago**. Al ser *software* libre puede ser descargado e instalado en un hosting externo para mayor control.

RAW

[raw](#) | [Vídeo: //vimeo.com/75866661](https://vimeo.com/75866661)

TECNOLOGÍA Aplicación web
(HTML5, JS, d3.js).

CONDICIONES DE USO Freeware / Libre
(LGPL).

AUTOR Density Design.

Es un asistente web que facilita a usuarios sin conocimientos de programación la realización de visualizaciones [d3.js](#) (<http://raw.densitydesign.org/>): dendrogramas, burbujas, mapa árbol, mapa mental, coordenadas paralelas, etc. La librería de programación **d3.js** permite a desarrolladores de software la realización de todo tipo de visualizaciones interactivas, pero **RAW** facilita, al menos, la realización de algunas de estas visualizaciones a usuarios no-técnicos. Los gráficos generados pueden ser insertados en páginas web externas.



2.4.2 Selección de herramientas de visualización geoespacial: CartoDB, Google MyMaps y Crowdmap

Se han seleccionado en base a su facilidad de uso (al alcance del usuario con capacitación tecnológica media) y a la posibilidad de almacenar los mapas online para compartirlos o insertarlos en websites externos, como blog o página personal.

Tanto **CartoDB** como **Google Maps** son herramientas que cuentan con una larga trayectoria y amplia aceptación para proyectos comerciales y no comerciales. **Crowdmap (Ushahidi)** es más específica pues está orientada a la creación de aplicaciones crowdsourcing de mapas³. [Fuente: www.ushahidi.com/about](http://www.ushahidi.com/about)

³ Ushahidi nace como herramienta para el reporte ciudadano de los disturbios raciales de Kenia de 2008 y es posteriormente usado para ofrecer reportes ciudadanos geo-referenciados durante el terremoto de Haití de 2010 a los servicios de emergencia y protección civil.

CartoDB

[cartodb](#) | [Galería](#) | [Tutoriales](#)

TECNOLOGÍA Aplicación web
(Javascript, PostGIS).

CONDICIONES DE USO Freeware / Libre.

AUTOR Vizzuality.



Permite realizar mapas de puntos, de calor y coropletas. Facilita, además, ciertas operaciones de limpieza de datos y *record linkage*. Los mapas generados pueden ser insertados en páginas web externas. La versión *online* de **CartoDB** almacena los mapas generados hasta 250 MB (para más espacio cuenta con planes de pago). Al ser software libre puede ser descargado e instalado en un hosting externo para mayor control.

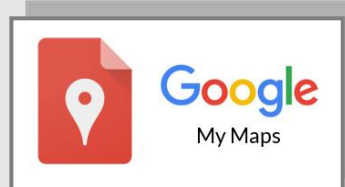
GoogleMyMaps

[my maps](#) | [Galería](#) | [Vídeos](#) | [Tutoriales](#)

TECNOLOGÍA Aplicación web.

CONDICIONES DE USO Freeware / Propietario.

AUTOR Google.



Permite realizar mapas de puntos, de calor y coropletas. Facilita, además, ciertas operaciones de *data cleansing* (limpieza de datos) y *record linkage* (enlace de datos). Los mapas generados pueden ser insertados en páginas web externas.

Crowdmap

[crowdmap \(ushahidi\)](#) | [Galería](#) | [Vídeos](#)

TECNOLOGÍA Aplicación web.

CONDICIONES DE USO Freeware / Libre (LGPLv3).

AUTOR Ushahidi.

Permite realizar mapas de puntos en forma de aplicaciones web / móvil crowdsourcing, esto es, aplicaciones abiertas a que cualquier usuario pueda contribuir con puntos sobre el mapa o con información adicional sobre un mapa de puntos. Al ser software libre puede ser descargado e instalado en un hosting externo para mayor control. Los mapas generados pueden ser insertados en páginas web externas.



2.4.3 Selección de herramientas de visualización temporal: TimeFlow, TimelineJS

Además de **Tableau** y **RAW**, que permiten realizar visualizaciones temporales, presentamos a continuación dos herramientas específicas: **Timeflow** y **TimelineJS**. Las ventajas que presentan estas herramientas son la facilidad de uso (al alcance del usuario con capacitación tecnológica media), que cuenten con licencia libre y que permitan insertar los resultados en websites externos como un blog o página personal. No hemos encontrado herramientas que cumplan estas condiciones para la generación de **diagramas de Gantt**, aunque **Tableau Public** permite realizarlas.

TimeFlow

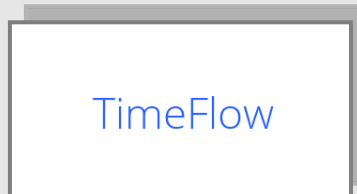
[timeflow](#) | [Galería](#) | [Tutoriales](#)

TECNOLOGÍA Aplicación de Escritorio multiplataforma.

CONDICIONES DE USO Libre.

AUTOR Flowing Media, Inc & Duke University.

Permite realizar distintos tipos de visualizaciones temporales: línea del tiempo, calendario, tabla y listado a partir de hojas de cálculo. Los gráficos generados pueden ser exportados en formato HTML.



TimelineJS

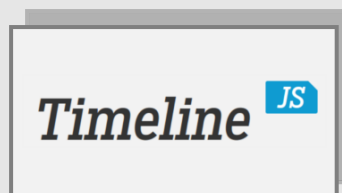
[timelinejs](#) | [Galería](#) | [Tutoriales](#)

TECNOLOGÍA Aplicación web (HTML5, Javascript).

CONDICIONES DE USO Libre (Mozilla Public License v2).

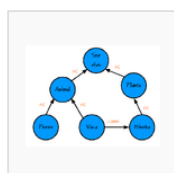
AUTOR Knight Lab, Northwestern University.

Permite realizar líneas del tiempo a partir de hojas de cálculo que provengan de **Google Drive**. La representación visual final es una colección cronológica de diapositivas a la que puede añadirse imagen o vídeo. Las líneas del tiempo generadas pueden ser insertados en páginas web externas. Al ser *software* libre puede ser descargado e instalado en un *hosting* externo para mayor control, por ejemplo construir líneas del tiempo desde datos en otro tipo de ficheros o bases de datos.



2.5 Generación de redes de datos relacionados

Los **gráficos de redes** permiten descubrir redes y patrones existentes en una serie de datos relacionados. Visualmente se componen de **nodos** y **aristas**. Una arista que une dos nodos representa una relación existente entre ambos. Fuente: "Decálogo del Reutilizador de Datos del Sector Público" (2014, datos.gob.es)



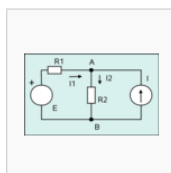
Mapas conceptuales



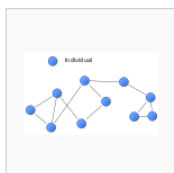
Plano de estaciones del metro.



Plano de autopistas.



Circuito eléctrico



Sociograma de una red social



Topología de red de computadores



Organigramas



Arquitectura de redes de telefonía móvil

Los gráficos de redes (o **grafos**) son objeto de estudio de la matemática discreta, y se han empleado para **modelar y resolver problemas**, como encontrar la distancia más corta entre dos nodos, el mínimo recorrido que pase por todos los nodos o el agrupamiento por similitud o proximidad. Se trata, por tanto, de herramientas que

simultáneamente **permiten el análisis y la visualización** de información.

Las redes de datos se utilizan tanto para la exploración visual de las **relaciones entre nodos** como para el análisis matemático y estadístico de estas relaciones. Si el objeto de la visualización o el análisis no son las relaciones entre nodos, **se recomienda el uso de otras herramientas** de visualización.

Para poder utilizar herramientas de visualización o análisis de redes de datos es requisito previo contar con un conjunto de datos que contenga tanto una **lista de nodos, como de aristas** (relaciones entre nodos). Fuente: [Aplicaciones de la teoría de grafos. Wikipedia.](http://es.wikipedia.org/wiki/Teoría_de_grafos)

EJEMPLO

El modelado mediante grafos se emplea para aplicaciones destinadas a encontrar el recorrido óptimo por autopista, agrupar usuarios de **twitter** en función de cuánto se retuitean entre ellos, visualizar el recorrido que los visitantes hacen por nuestra página web, o encontrar patrones de adjudicación de contratos a empresas desde los diferentes organismos públicos.

2.5.1 Selección de herramientas de redes de grafos: Fusion Tables y Gephi

Presentamos dos herramientas para realizar la visualización de redes de datos. La primera es **Fusion Tables**, herramienta que ya hemos presentado, y que permite una visualización muy básica de redes de datos de forma muy sencilla. La segunda es **Gephi**, una suite específica de visualización y análisis de redes de datos. Se ha seleccionado en base a la facilidad de uso (al alcance del usuario con capacitación tecnológica media) y por ser de licencia libre.

Fusion Tables

[fusiontables](#) | [Galería](#) | [Vídeos](#) | [Tutoriales](#)

TECNOLOGÍA Aplicación web
(HTML5, Javascript).

CONDICIONES DE USO Freeware / Propietario.

AUTOR Google.

Permite realizar gráficos de red de tipo jerárquico, esto es, los distintos valores de propiedad común a todos los nodos se convierten a su vez en otra categoría de nodos (y de otro color), y las relaciones entre estos se realizan automáticamente. Se pueden ponderar los nodos empleando una variable numérica, y esto determina el tamaño de los nodos. Los gráficos de red interactivos generados pueden ser insertados en páginas web externas.



Gephi

[gephi](#) | [Galería](#) | [Tutoriales](#)

TECNOLOGÍA Aplicación Escritorio
multiplataforma.

CONDICIONES DE USO Libre (GPL-CDDL).

AUTOR Gephi Consortium.

Es una suite interactiva de creación, visualización y análisis para todo tipo de redes de datos. Permite importar datos desde hoja de cálculo y conexión directa a bases de datos. Es capaz de trabajar con decenas de miles de nodos y realizar cualquier tipo de análisis típico de redes: agrupamiento, caminos mínimos, etc. Es software libre multiplataforma y las visualizaciones generadas pueden ser exportadas en formato imagen y vectorial para ser incrustadas en sitios externos.



RESUMEN

Hemos finalizado la unidad, y repasamos a continuación los **puntos principales**:

- ✓ Los **agentes infomediarios** transforman la información o datos públicos en productos, servicios o aplicaciones, y para ello utilizan sobre todo herramientas de análisis de datos, de búsquedas y de generación de informes.
- ✓ Existe todo un ecosistema de **herramientas de libre uso**, disponibles en Internet, para realizar todas las etapas clave en la reutilización de datos abiertos gubernamentales e información pública: **extracción, transformación, análisis, presentación**. Muchas de estas herramientas permiten guardar nuestro trabajo *online* y compartirlo o insertarlo en otros websites.
- ✓ Cuando las fuentes de información pública no proveen de datos reutilizables (datos en ficheros PDF o web), muchos agentes infomediarios recurren a **técnicas de scraping** para extraer los datos en formas reutilizables. Existen herramientas que ayudan en la realización de técnicas de scraping, pero no son efectivas en todos los casos.
- ✓ Para preparar un conjunto de datos en los formatos/estructuras requeridos por las herramientas de análisis o visualización, se emplean **técnicas de tratamiento de datos** como la limpieza, transformación o enlace de datos.
- ✓ Las **herramientas de análisis estadístico** nos permiten transformar la información de un conjunto de datos en conocimiento: en forma de análisis descriptivos, inferenciales, modelos predictivos o sistemas de recomendación.
- ✓ Las **herramientas de visualización** ayudan a explorar un conjunto de datos de forma interactiva (visualizaciones exploratorias), a presentar los resultados de un análisis de datos o a comunicar un mensaje (visualizaciones expositivas). El desarrollo de las tecnologías de la comunicación y la información ha multiplicado los tipos de visualizaciones, que se han añadido a los clásicos gráficos empleados en estadística o los mapas de cartografía.