

# BIG DATA for BUSINESS

## 2.2 Fuentes y tipos de datos

---

# Conecta Empleo

Contenido desarrollado por  
**Synergic Partners**



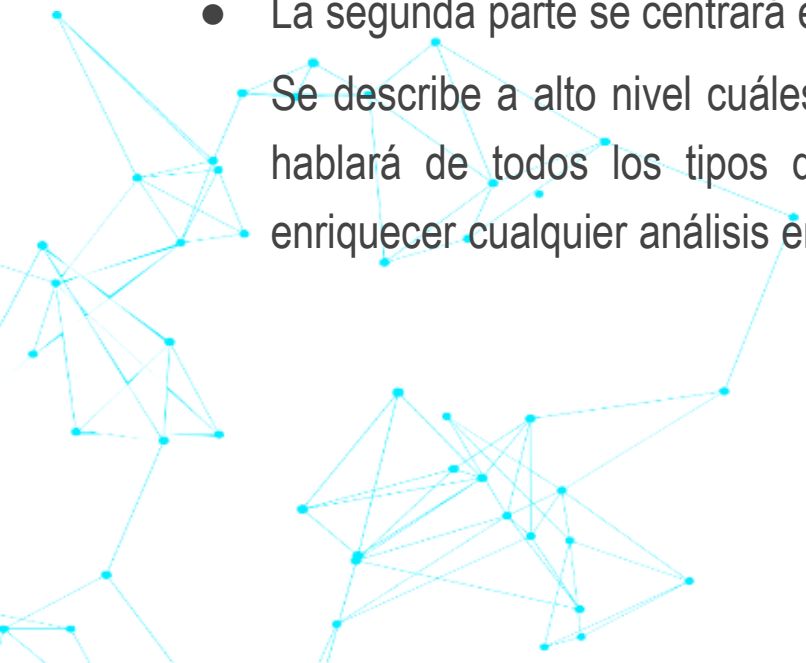
# Índice del módulo

## 2.2 FUENTES Y TIPOS DE DATOS

- Tipos de datos
- Fuentes de datos externas e internas
- Datos estructurados y no estructurados
- Concepto de API y Redes sociales
- Datos abiertos

El **módulo** se centra en dar una visión de los **TIPOS DE DATOS y FUENTES DE DATOS** existentes.

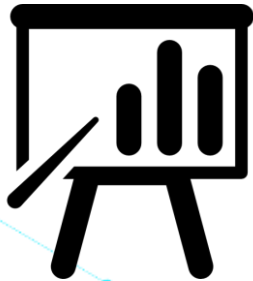
- La primera parte del módulo se dedicará a ver **qué tipos de datos hay**, cómo ha sido su evolución en los últimos años y el correspondiente efecto en la aparición del Big Data.
- La segunda parte se centrará en las **fuentes de datos**.  
Se describe a alto nivel cuáles son las fuentes de datos internas, y seguidamente se hablará de todos los tipos de fuentes externas existentes que pueden ayudar a enriquecer cualquier análisis empresarial.



# Tipos de datos



**Datos operacionales**



**Tendencias  
de mercado**



**Información  
demográfica del cliente**

# Tipos de datos

## CLIENTES

- Tarjetas de compra/clientes
- Oferta / respuesta
  - SMS
  - Clickstream
- Comportamiento en tienda
- Ratings y encuestas
- Geolocalización
  - Foros webs
  - Call centers
  - Sensores
  - Email

## REDES SOCIALES

- Twitter
- Facebook
- LinkedIn
- Pinterest
- Google +
- Youtube
- Blogs
- Wikis
- Yelp

## IoT

- Beacons
- Smart mirrors
- Digital Signage
  - RFID
- Smart packaging
- Smart price tags
- ...

## CADENA DE SUMINISTRO

- Órdenes de compra
  - Envíos
- Devoluciones
- Sensores
- Almacén
- Recibos
- Transportistas
- Información de producto
- Colocación en tienda
- Inteligencia de mercado

## SISTEMAS DE CAPTURA DE LA ORGANIZACIÓN

**BILLONES DE INTERACCIONES**

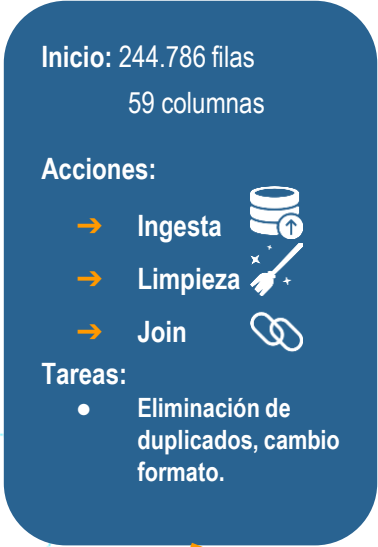
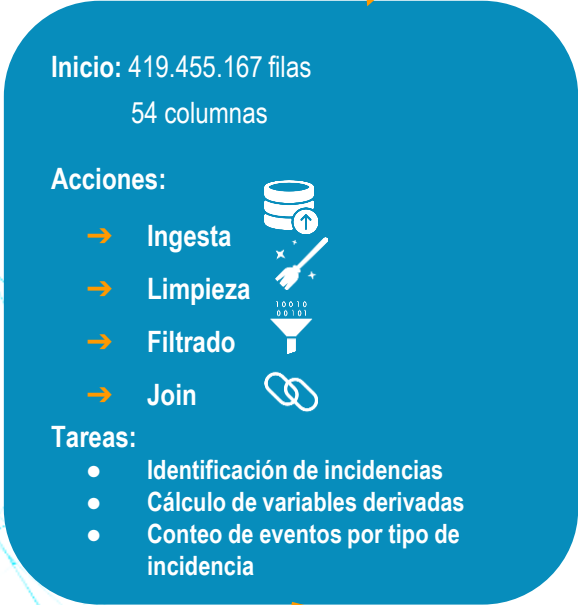
**MILLONES DE TRANSACCIONES**

## MERCADO

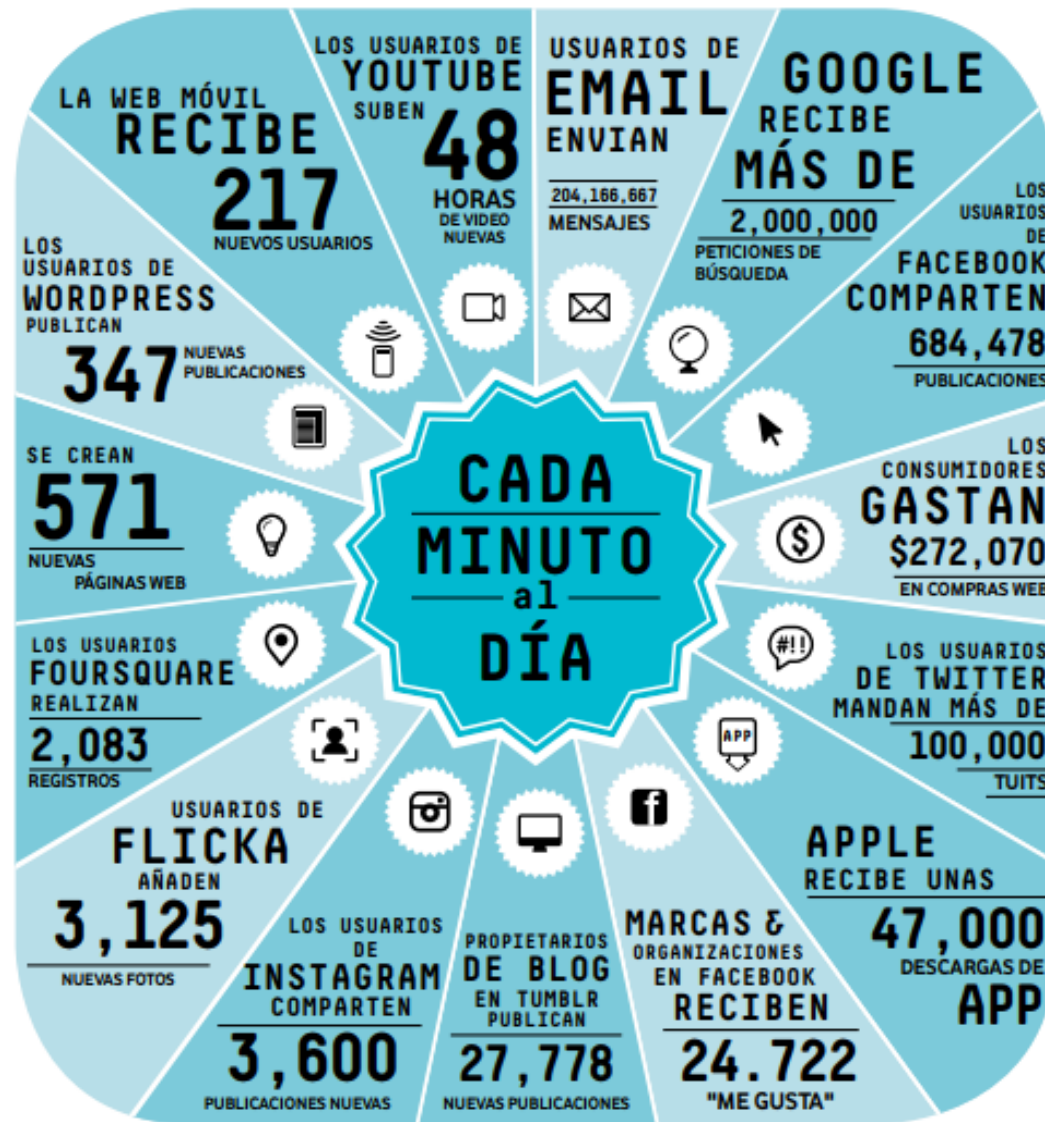
- Comercio
- Organización
- Demográfica
- Competencia
- Eventos
- Noticias del sector
- Situación económica
- Condiciones meteorológicas

# Tipos de datos

Quick win  
Securitas.  
Clasificación de  
alarmas.

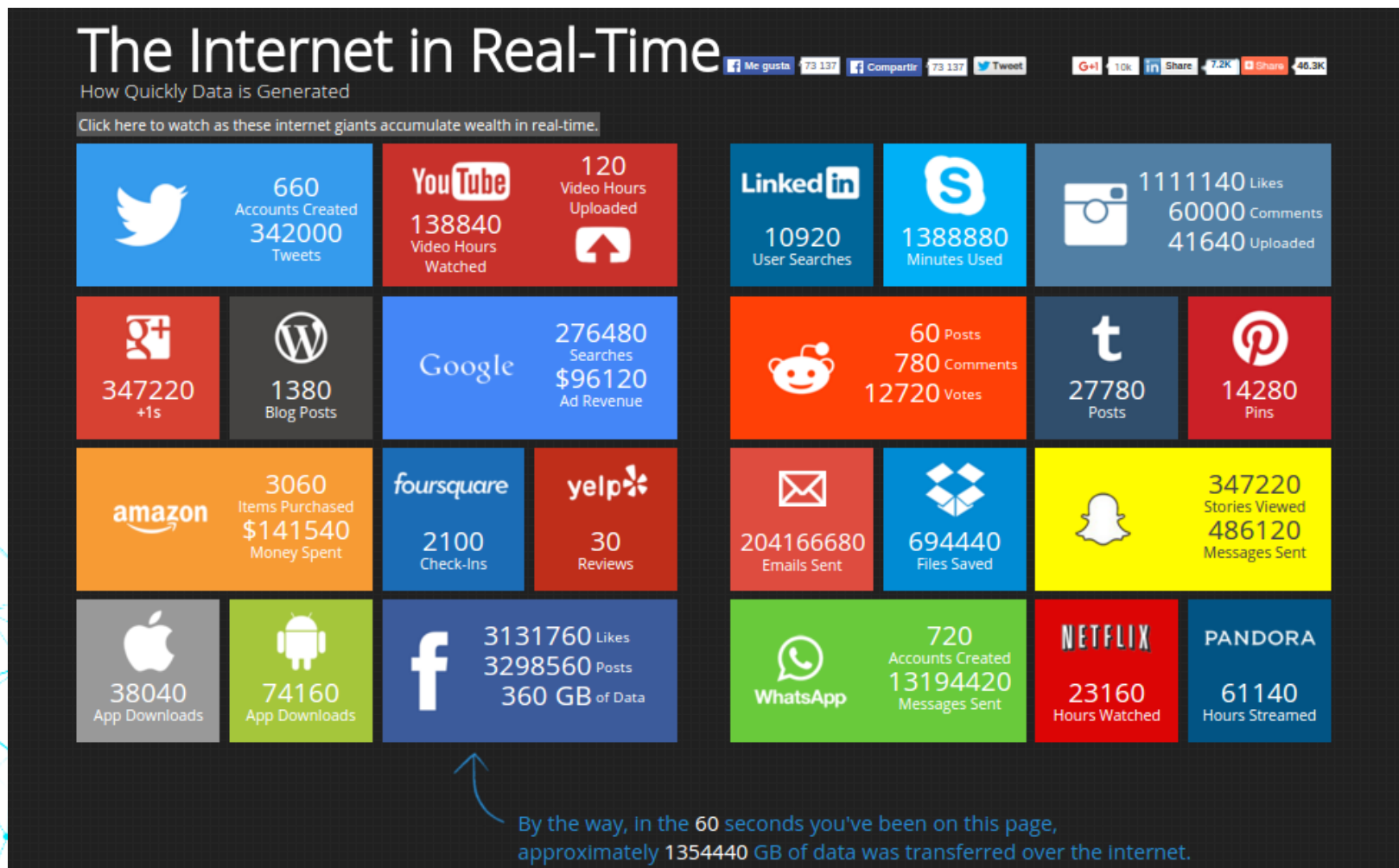


# Tipos de datos





# Tipos de datos

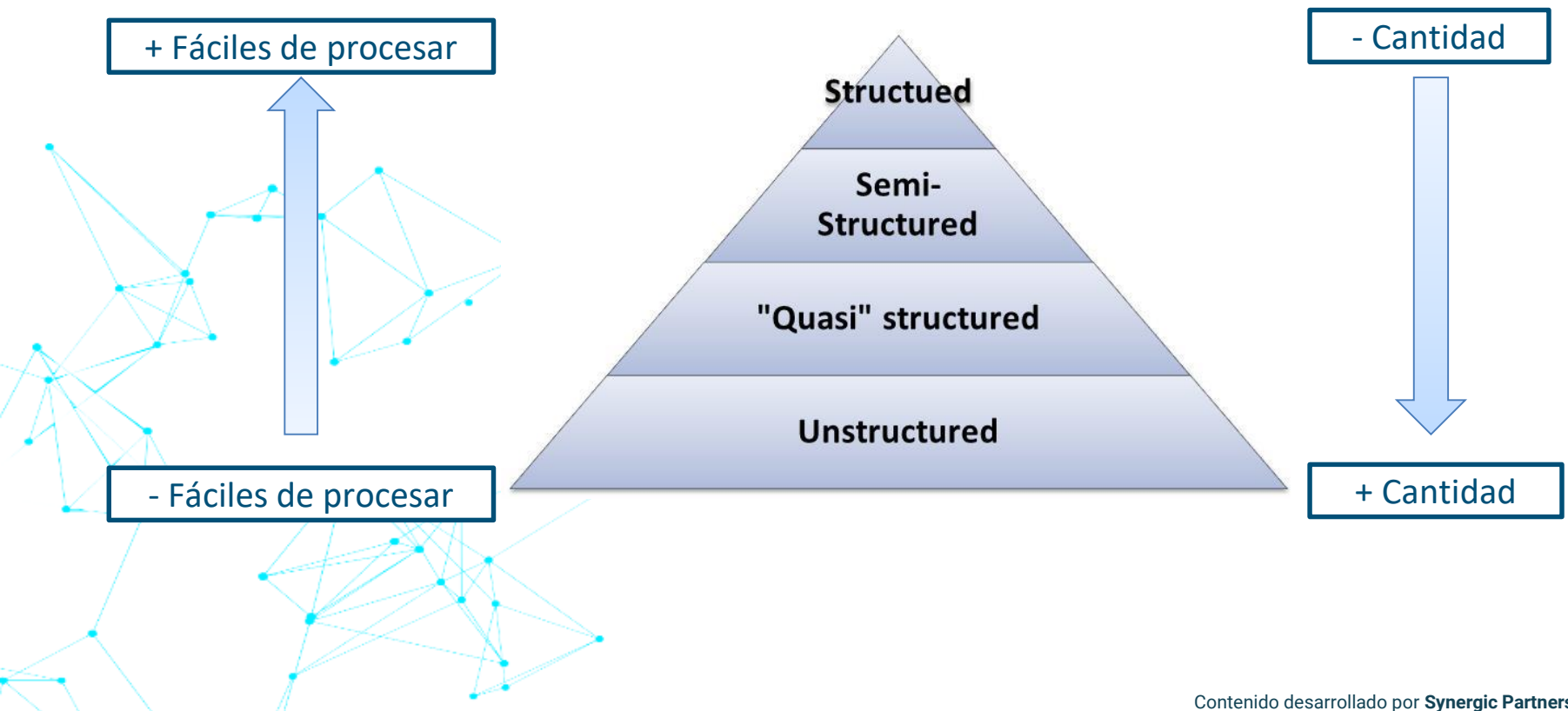


<https://visual.ly/community/infographic/how/internet-real-time>



# Tipos de datos I

A alto nivel, los datos se clasifican desde los más estructurados (Structured) que serían tablas con formato, hasta los más desestructurados (Unstructured), los cuales pueden ser archivos de texto, PDF's, imágenes, videos, etc.



# Tipos de datos II

## Structured

- Datos con un formato de dato establecido y estructura.
- Ejemplo: Datos transaccionales y OLAP.

## Semi-Structured

- Datos de texto con un patrón reconocible, el cual es apto para ser parseado (troceado).
- Ejemplo: Archivos XML que son definidos por un esquema XSD

## “Quasi”Structured

- Datos de texto con un patrón de datos difícil de identificar. Pueden ser formateados con esfuerzo, tiempo y herramientas específicas.
- Ejemplo: Registros de eventos o acciones en una web, logs.

## Unstructured

- Datos que no tienen ninguna coherencia ni patrón y usualmente están almacenados en distintos tipos de archivos
- Ejemplo: Archivos de texto, PDFs, Imágenes, Videos..

# Algunos de los datos al que tienen acceso las Telco

## CDRs

(llamadas realizadas, número de origen y destino, duración de la llamada)

## Datos de conexiones inactivas

(descubrimiento del móvil por parte de la red, cambio de 3G a 4G, activación del móvil, cambio de cobertura-celda)

## Datos del CRM

(Demográficos, nacionalidad, DNI, información financiera - nivel de consumo)

## Datos de Roaming

(países visitados, nacionalidad de la tarjeta SIM)

## DPI

(qué apps se usan y cuáles conjuntamente, con qué frecuencia)

## Datos de la sesión Web

(interacción con webs, webs visitadas, tiempo medio de sesión)

## Datos de Red Fija

(análisis sobre la cantidad de uso sobre video bajo demanda)

## Catálogo de Celda

(mapa de localización del usuario acerca de datos CDR o conexiones inactivas)

## Datos del Call Center

## Datos WiFi de terminales como puntos de acceso

## Beacons

## Datos de Dispositivos IoT

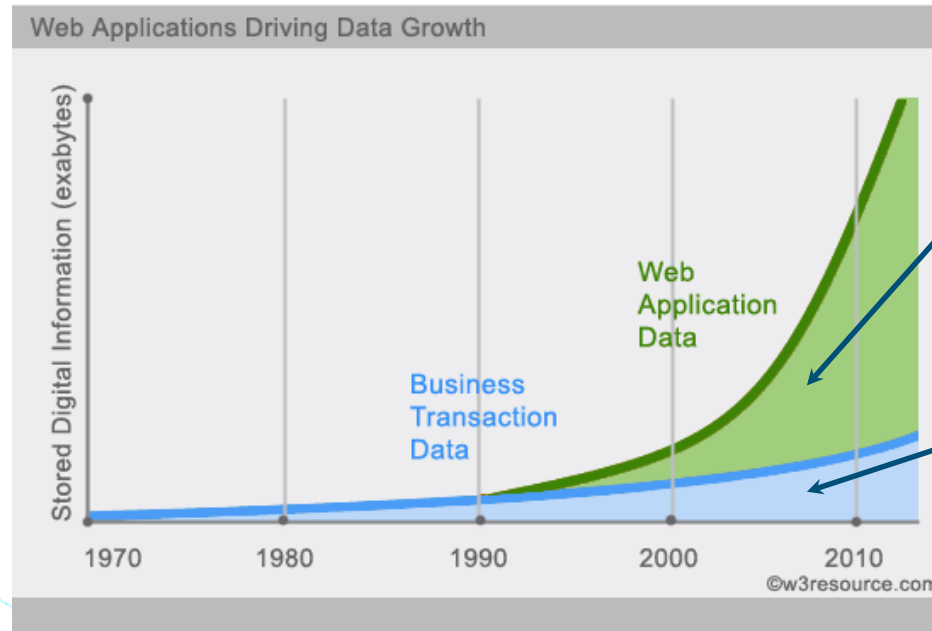
# Evolución de los datos I

Con la aparición de nuevas fuentes de datos (Redes Sociales, Smartphones, Sensores, etc.), así como la posibilidad de analizar datos que nunca antes habían sido explorados, la cantidad de información disponible está **creciendo de forma exponencial**:



Source: IDC's Digital Universe Study, sponsored by EMC, June 2011

## Evolución de los datos II



En algunos casos los datos son fuentes de datos tradicionales ya conocidas (datos de transacciones), o en otros, datos desestructurados (datos de interacciones), los cuales aumentan exponencialmente en los últimos años.

A partir de este momento donde el volumen y variedad de datos ha adquirido tal dimensión es donde se origina el **concepto de Big Data**.

An abstract network diagram consisting of numerous blue dots (nodes) connected by thin blue lines (edges). The nodes are scattered across the left side of the slide, with some forming small, dense clusters and others being isolated or part of larger, more complex structures. The overall shape of the network is irregular and organic.

## Fuentes de datos internas

# Bases de Datos Tradicionales

## Fuentes internas

Las fuentes internas de información están constituidas por los documentos internos que son las memorias o registros de las operaciones cotidianas de la empresa. Se generan diariamente a través de informes (en los diferentes departamentos áreas o unidades de negocio de la empresa), de documentos que sirven para realizar análisis para toma de decisiones, de estudios especializados realizados por terceros como estudios de mercado, diagnósticos, de manuales para organización de procedimientos, de productos, (muestran la dinámica, las características o otra variable que quiera investigar), de normas técnicas etc.

Algunos ejemplos de las fuentes de Información interna que se pueden consultar a través de:

- Informes en los diferentes departamentos áreas o unidades de negocio de la empresa.
- Documentos que sirven para realizar análisis para toma de decisiones, de estudios especializados realizados por terceros como estudios de mercado, diagnósticos, de manuales para organización.
- Procedimientos.
- Productos, (muestran la dinámica, las características o otra variable que quiera investigar).
- Normas técnicas.



# Fuentes de Datos Internos

- Son la fuente de información más importante de las empresas ya que recogen su actividad diaria y evolución. Para una empresa tipo serían los datos de clientes, productos, empleados, proveedores, etc.
- Tradicionalmente las empresas se han basado sólo en estos datos para la toma de decisiones.
- Es importante tenerlas todas bien identificadas para poder gestionar los datos interdepartamentalmente y extraer su valor al hacer análisis cruzando estos datos.
- Las disciplinas de Data Governance, Master Data Management y Data Quality que se verán en el siguiente módulo se ocupan de gestionar correctamente estos datos.

# Fuentes de Datos Internos - Bancos

A continuación se muestran algunos tipos de datos internos existentes en los bancos:

## Operacional

- Riesgos (solicitudes de crédito)
- Riesgos (resultado del préstamo )
- Información de Tarjetas
- Encuestas Satisfacción
- Reclamaciones
- Encuesta Calidad

## Plataforma Informativa

- Movimientos No Contables
- Datos de Tarjetas
- Recibos Domiciliados
- Transferencias
- TxC por Canal
- Datos Generales
- Clientes
- Intervinientes
- HIGECO – (Vinculación, Servicios, A&P)
- Rentabilidad
- Datos Adicionales
- Clientes
- Email
- Oficina Adquiriente
- Saldos personas
- Clientes (sección Censal, Plan Uno)
- Oficinas (cierre y geolocalización)
- Productos (alta y baja)

An abstract network diagram composed of numerous small blue dots (nodes) connected by thin, light blue lines (edges). The connections form various geometric shapes, including triangles and polygons, scattered across the left side of the slide. The overall effect is a complex, interconnected web of data points.

## Fuentes de datos externas

# Fuentes de Datos Externos - Introducción

Tradicionalmente las empresas solo se podían guiar por los datos que recogían internamente y tomar decisiones solo en base a estos.

Hoy en día, desde la aparición de internet y la proliferación de la idea de globalizar el mundo, **las empresas pueden enriquecer sus análisis con fuentes de datos externos que son públicos para cualquiera sin necesidad de pagar por ello y con fácil acceso.**



# Fuentes externas

Son las que proveen información generada fuera de la empresa, como en bibliotecas y centros de documentación, en entidades, públicas o privadas, en material impreso, en videos, en cintas de audio, en investigaciones de campo.

Algunos ejemplos de las fuentes de Información externa que se pueden consultar a través de

- Redes sociales
- Páginas web & IOT
- Publicaciones de entidades públicas como ministerios y empresas del Estado y gubernamentales.
- Publicaciones de organismos de desarrollo económico/ social. Publicaciones de asociaciones y cámaras de industria y/o comercio.
- Publicaciones de proveedores, distribuidores, sindicatos, etc...
- Catálogos de universidades/ centros de investigación.
- Directorios, paginas amarillas, bases de datos privadas
- Prensa nacional/ local

# Fuentes de Datos Externos - Introducción

## Open Data

- Filosofía y práctica surgida en los últimos años por las entidades públicas que consiste en poner su información al alcance de todos para fomentar su reutilización, sin restricciones de copyright, patentes u otros mecanismos de control.

## APIs

- Application Programming Interface: Permiten extraer los datos de portales web mediante funciones con las que el usuario puede personalizar y filtrar a su gusto.

## Redes Sociales

- Las redes sociales más importantes (Facebook, LinkedIn, Twitter, etc.) tienen diversas APIs que permiten acceder a información de valor para tener un conocimiento del cliente más detallado.

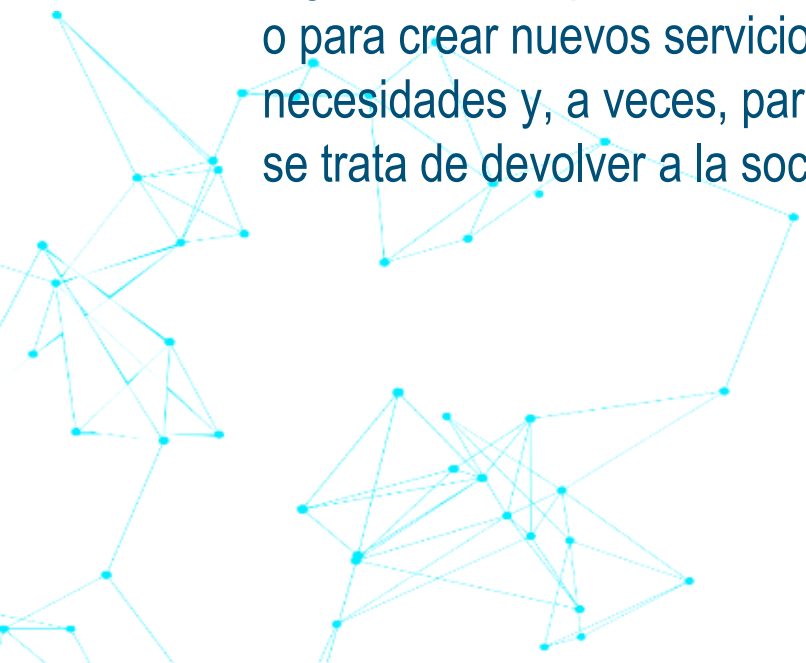
## Otros

- Otros sitios web que suministran Datasets de interés:
  - Instituto Nacional de Estadística
  - Catastro
  - Datos Públicos de Google

# Fuentes de Datos Externos - Open Data - Definición

**El Open Data es una filosofía y práctica surgida en los últimos años por las entidades públicas que consiste en poner la información al alcance de todos** para fomentar su reutilización, sin restricciones de copyright, patentes u otros mecanismos de control.

De este modo, tanto los ciudadanos, como empresas y otras organizaciones pueden acceder a estos datos fácilmente para informarse o para crear nuevos servicios que ayuden a las personas a resolver sus necesidades y, a veces, para explotar un nuevo tipo de negocio. Por tanto, se trata de devolver a la sociedad sus datos para extraer nuevo valor.





# Fuentes de Datos Externos - Open Data - Licencias

Para que el **Open Data** se materialice, son necesarias licencias de uso que, a través del concepto de copyright y la legislación del país correspondiente, regule el acceso a los datos y los derechos de uso.

La Open Source Initiative (OSI) mantiene una definición de **open source** y una lista de licencias que la cumplen.



<https://opensource.org/licenses>

En concreto, para contenidos textuales (i.e. no software) se utilizan mucho las licencias Creative Commons (CC).

<https://creativecommons.org/>



# Fuentes de Datos Externos - Open Data

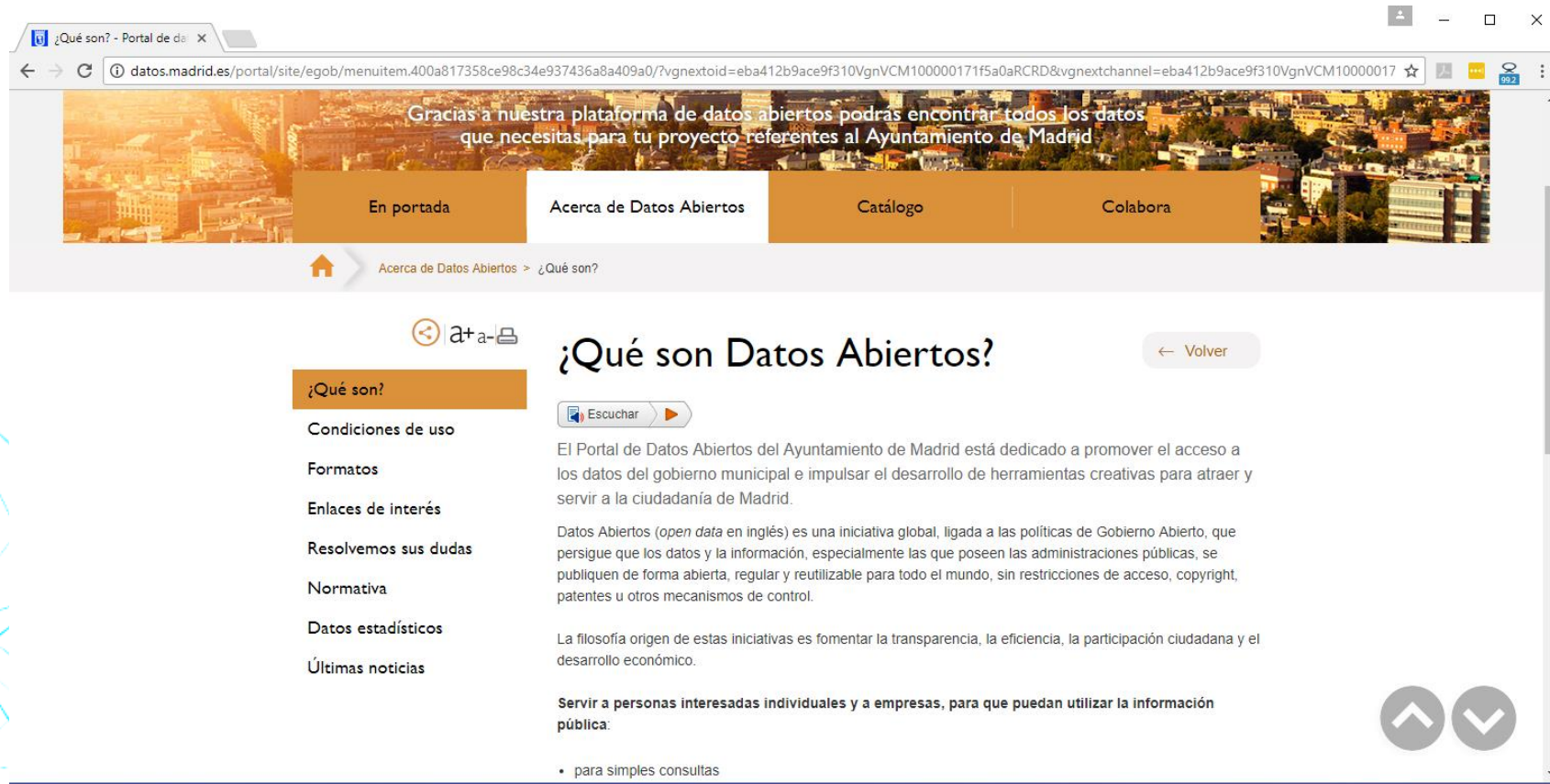
## Objetivos

En el caso del sector público, los objetivos principales de la apertura de datos son:

- **Aumentar la transparencia de la administración pública** a partir de poner los datos públicos al alcance de toda la sociedad, ya sean ciudadanos u organizaciones.
- **La explotación de los datos por parte de los colectivos antes mencionados a fin de crear nuevos servicios y aplicaciones** que mejoren la calidad de vida de las personas.
- **Determinar cuáles son las necesidades de la sociedad** en cuanto a datos públicos abiertos para que se puedan desarrollar estos nuevos servicios y aplicaciones.
- **Promover la reutilización** para poder extraer todo su valor.
- **Promoción económica del territorio** mediante las iniciativas *Open Data*.

# Fuentes de Datos Externos - Open Data

## Open Data Madrid



The screenshot shows the Open Data Madrid website. The browser address bar displays the URL: [datos.madrid.es/portal/site/egob/menueitem.400a817358ce98c34e937436a8a409a0/?vgnextoid=eba412b9ace9f310VgnVCM100000171f5a0aRCRD&vgnnextchannel=eba412b9ace9f310VgnVCM10000017](http://datos.madrid.es/portal/site/egob/menueitem.400a817358ce98c34e937436a8a409a0/?vgnextoid=eba412b9ace9f310VgnVCM100000171f5a0aRCRD&vgnnextchannel=eba412b9ace9f310VgnVCM10000017). The main banner features a cityscape image with the text: "Gracias a nuestra plataforma de datos abiertos podrás encontrar todos los datos que necesitas para tu proyecto referentes al Ayuntamiento de Madrid". Below the banner is a navigation bar with four buttons: "En portada", "Acerca de Datos Abiertos" (which is highlighted), "Catálogo", and "Colabora". Below the navigation bar is a breadcrumb trail: "Acerca de Datos Abiertos > ¿Qué son?". The main content area is titled "¿Qué son Datos Abiertos?" and includes a "Volver" button. A sidebar on the left contains a list of links: "¿Qué son?", "Condiciones de uso", "Formatos", "Enlaces de interés", "Resolvemos sus dudas", "Normativa", "Datos estadísticos", and "Últimas noticias". The main text explains that the portal is dedicated to promoting access to municipal data and developing creative tools. It also mentions that Open Data is a global initiative linked to Open Government, aiming to make data and information, especially from public administrations, available in an open, regular, and reusable manner. The text concludes with the philosophy of these initiatives: to foster transparency, efficiency, citizen participation, and economic development. At the bottom, it states the goal is to serve individuals and companies who can use public information for simple queries.

¿Qué son?

Condiciones de uso

Formatos

Enlaces de interés

Resolvemos sus dudas

Normativa

Datos estadísticos

Últimas noticias

### ¿Qué son Datos Abiertos?

← Volver

Escuchar

El Portal de Datos Abiertos del Ayuntamiento de Madrid está dedicado a promover el acceso a los datos del gobierno municipal e impulsar el desarrollo de herramientas creativas para atraer y servir a la ciudadanía de Madrid.

Datos Abiertos (*open data* en inglés) es una iniciativa global, ligada a las políticas de Gobierno Abierto, que persigue que los datos y la información, especialmente las que poseen las administraciones públicas, se publiquen de forma abierta, regular y reutilizable para todo el mundo, sin restricciones de acceso, copyright, patentes u otros mecanismos de control.

La filosofía origen de estas iniciativas es fomentar la transparencia, la eficiencia, la participación ciudadana y el desarrollo económico.

Servir a personas interesadas individuales y a empresas, para que puedan utilizar la información pública:

- para simples consultas

<http://datos.madrid.es/portal/site/egob>

# Fuentes de Datos Externos - Otras

Otras fuentes de datos abiertos más globales:

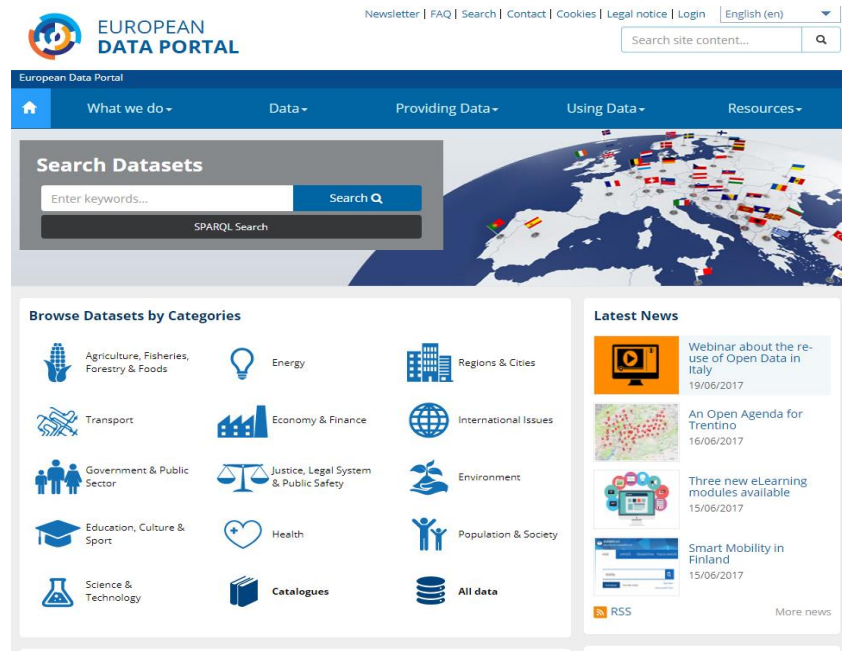
**datos.gob.es**  
reutiliza la información pública

- Web principal => <http://datos.gob.es>
- Catálogo => <http://datos.gob.es/es/catalogo>
- API => <http://datos.gob.es/es/accessible-apidata>
- SPARQL => <http://datos.gob.es/es/sparql>



EUROPEAN  
DATA PORTAL

- <https://www.europeandataportal.eu/>



# Fuentes de Datos Externos - Otras

Otras fuentes que pueden ser de interés para enriquecer los análisis de las organizaciones son:

- **World Bank Data:** El Banco Mundial proporciona ciertos indicadores macroeconómicos para sus objetivos de desarrollo:



<http://data.worldbank.org/>

- **Datos Públicos de Google:** Google ofrece estadísticas sobre el Desarrollo Mundial, Salarios, Desempleo, Deuda gubernamental en Europa, etc.

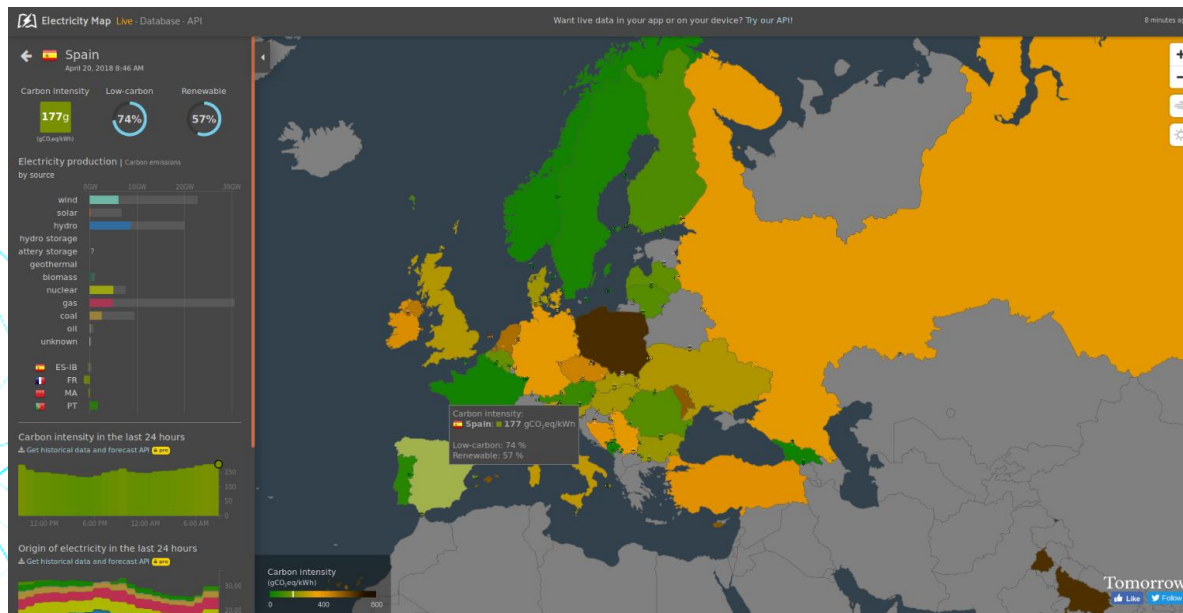
<https://www.google.com/publicdata/directory>



Public Data

# Fuentes de Datos Externos - Otras

Electricity Map es una aplicación web que usa datos públicos de consumo eléctrico para crear un mapa interactivo de generación de CO2 en tiempo real.



Para el mercado europeo, su fuente de información principal es ENTSO-E, una iniciativa de los países europeos para aumentar la transparencia de los mercados eléctricos.



# Redes Sociales



- Las **redes sociales** son una de las **fuentes** de información **externas** más interesantes para las organizaciones.
- De ellas se pueden obtener una gran **variedad** de **datos**, tanto de las **personas** individuales como de las **empresas**.

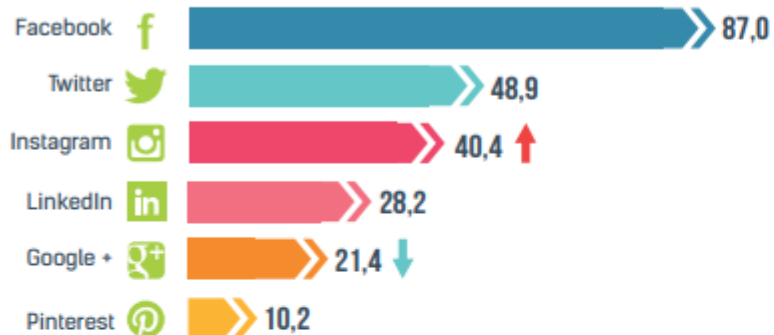
El acceso a la información en redes sociales se realiza a partir de:

- **Crawling** o reproducción de búsquedas humanas mediante un programa.
- **APIs** que las propias redes sociales ponen al alcance de los desarrolladores.

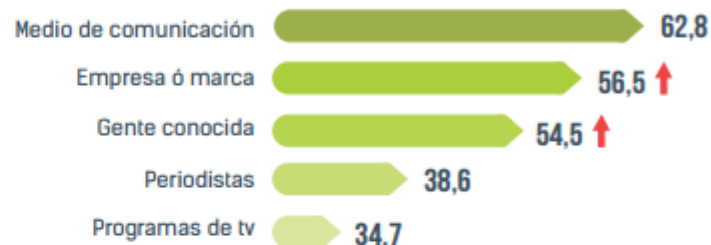


# Redes Sociales - Uso en España

## REDES SOCIALES UTILIZADAS (ÚLTIMOS 30 DÍAS) (%)



## SEGUIMIENTO EN REDES SOCIALES (ÚLTIMOS 30 DÍAS) (%)



[ Base: Acceden a redes sociales ]

## SEGUIMIENTO DE YOUTUBERS (%)

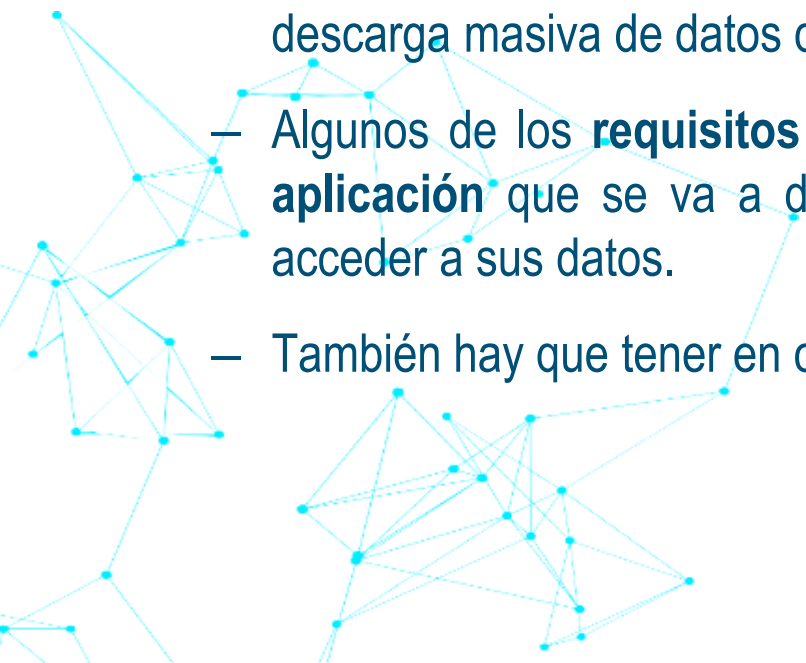


<http://www.aimc.es/otros-estudios-trabajos/navegantes-la-red/infografia-resumen-19-navegantes-la-red/>

# Redes Sociales

Para la **utilización** de las **APIs** se tienen que tener en cuenta diversos aspectos:

- Normalmente las redes sociales disponen **más de una API** en función de la información a la que se quiere acceder. Por este motivo, hay que informarse para conocer cuál proporciona los datos deseados.
- La utilización de la API no suele ser libre, sino que está sometida a los **términos de uso** de la red social. Es muy importante conocer estos términos para estar dentro del **marco legal**. La mayoría de ellas no permiten la descarga masiva de datos de usuarios.
- Algunos de los **requisitos** para su uso más comunes son el **registro** de la **aplicación** que se va a desarrollar y el **consentimiento** del **usuario** para acceder a sus datos.
- También hay que tener en cuenta los **límites de uso**.



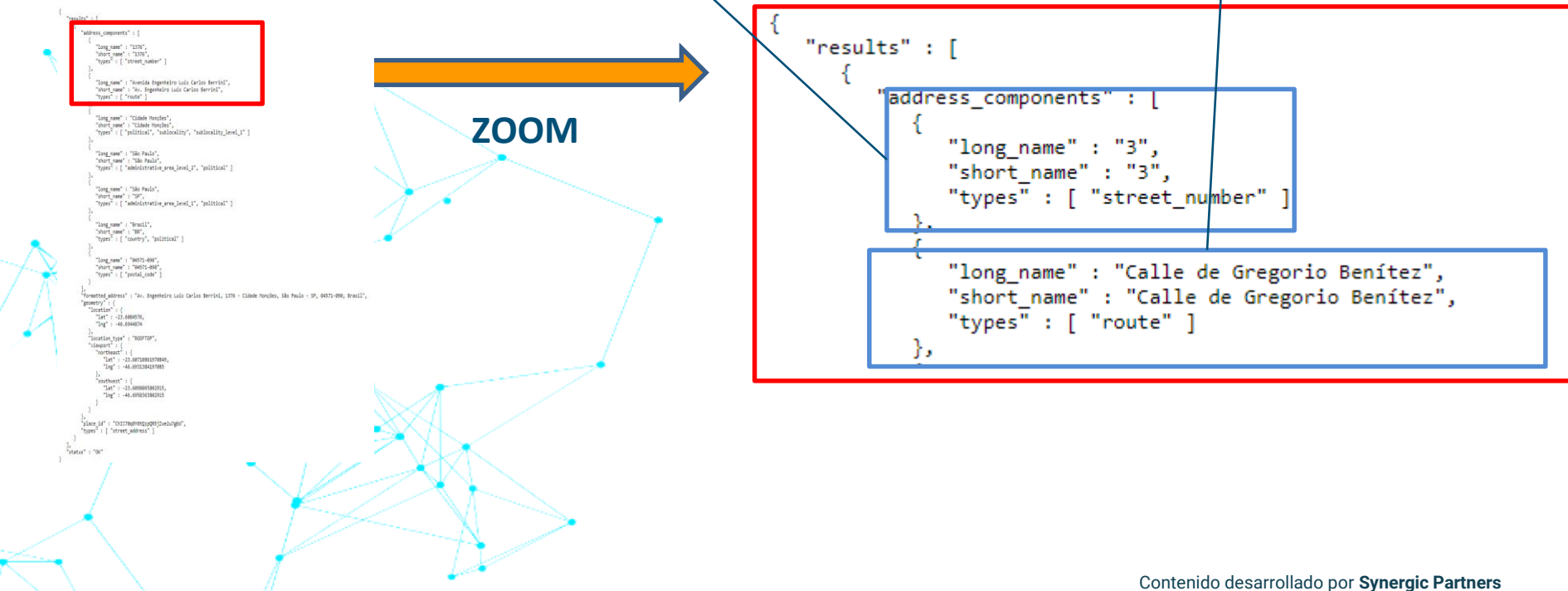
# Redes Sociales - APIs – Ejemplo de utilización I

Para tener una idea básica del funcionamiento de una API, a continuación se expone un ejemplo de la **API de google Maps** para direcciones:

**Con la URL:**

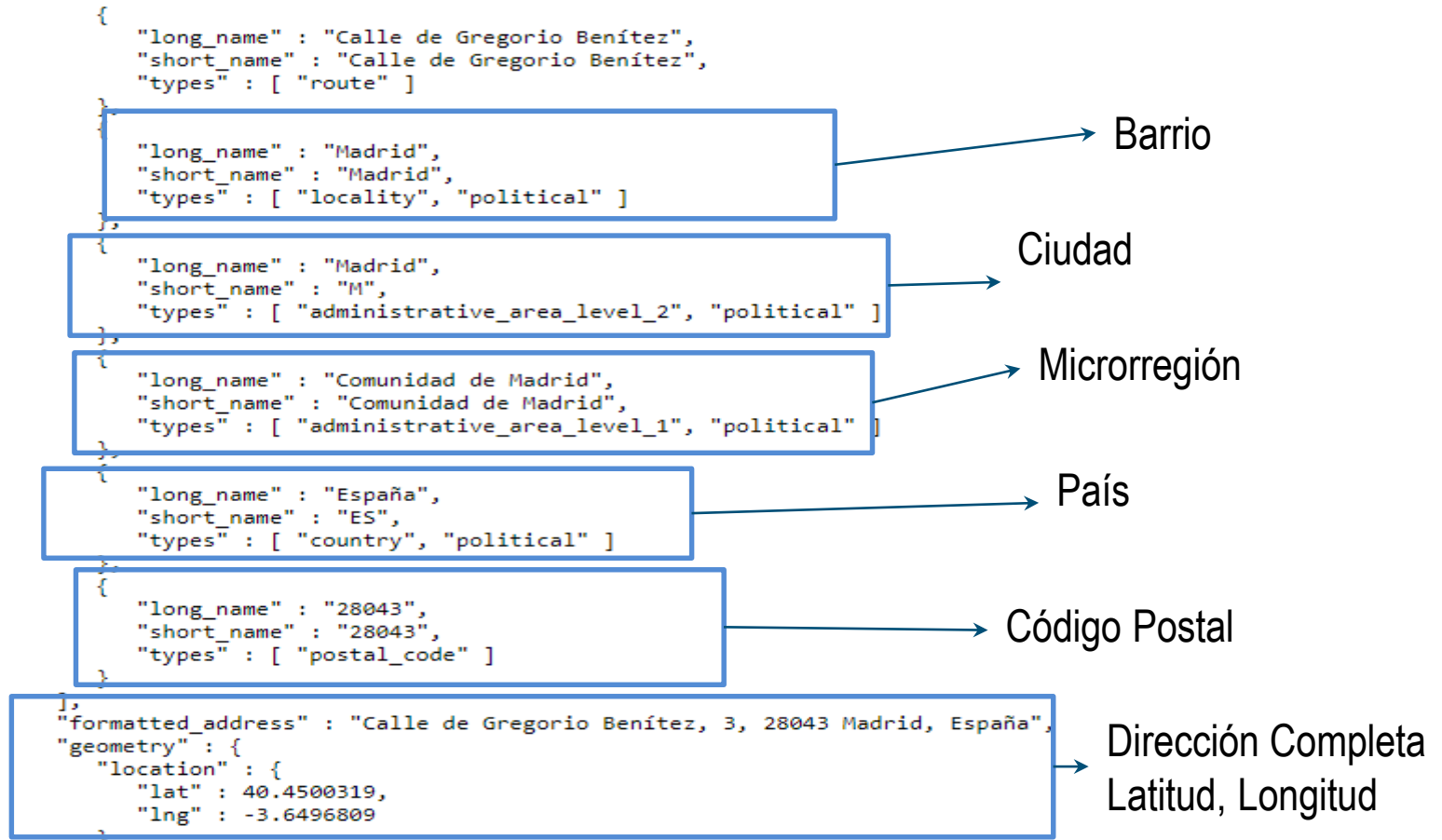
<http://maps.googleapis.com/maps/api/geocode/json?address=Calle de Gregorio Benítez, 3, 28043 Madrid>

**El navegador devuelve en formato JSON:**



# Redes Sociales - APIs – Ejemplo de utilización II

Continuación:



Debidamente parseados los datos se pueden trasladar en una tabla, añadiendo valor para la empresa al tener direcciones más completas y poder hacer así mejores análisis.

## APIs

**Graph API:** permite leer y escribir en facebook. Está formada por:

**Keyword Insight API:** buscar comentarios sobre algún tema.

**Public feed API:** proporciona los estados actualizados de usuarios y páginas.

**Chat API:** sirve para utilizar el chat de facebook en webs y aplicaciones.

**Ads API:** se utiliza para crear anuncios en facebook.

## Requisitos

Algunas están vetadas. Hay opciones que requieren el **permiso** del **usuario**.

## Límites de uso

**Depende** de la información solicitada y el uso.

## Información que se obtiene

**Perfil público** del usuario y su lista de **amigos**. El resto de información depende de las opciones de **privacidad** definidas por el usuario.

## APIs

**Data API:** incorpora funcionalidades de youtube. Se pueden obtener características de un video.

**Analytics API:** estadísticas, popularidad e información demográfica de los vídeos, canales de youtube o usuarios. Permite filtrar los datos.

**Live Streaming API:** gestión de eventos en tiempo real en youtube.

## Requisitos

Es necesario registrarse.

## Límites de uso

**Depende** de la información solicitada y la acción que se realice.

## Información que se obtiene

**Análisis** de los **vídeos**: número de reproducciones, **popularidad**, **valoración** de los usuarios, **localización** de los usuarios, etc.

## APIs

**Profile API:** devuelve el perfil de un usuario.

**Connections API:** devuelve los contactos de un usuario.

**People Search:** permite realizar búsquedas en LinkedIn.

**Company Look API y Company Search:** búsqueda y obtención de datos básicos de empresas.

**Job Lookup API y Job Search:** búsqueda e información sobre ofertas de trabajo.

## Requisitos

Es necesario **registrar la aplicación**. Algunas requieren de un **permiso especial**.

## Límites de uso

**Dependen** de la información que se solicita, y varían por aplicación, usuario y desarrollador.

## Información que se obtiene

Perfil de usuarios, información básica de empresas y red de contactos de usuarios.



## APIs

**REST API:** API general de twitter, hay diversas opciones: timeline, tweets, búsqueda, streaming (últimos tweets), mensajes privados, amigos y seguidores, usuarios, usuarios sugeridos, favoritos, listas, búsquedas guardadas, localización de los tweets, tendencias y report de spam.

**Streaming API:** seguimiento en tiempo real de tweets.

## Requisitos

Es necesario darse de **alta**. Algunas opciones requieren el **permiso** del **usuario**.

## Límites de uso

**Depende** de la información solicitada y el uso, **aprox.180 peticiones** por **usuario/15 min**. Histórico de 7 días. Existen empresas de pago que proporcionan mayor histórico

## Información que se obtiene

**Tweets, información básica** de los **usuarios**, **red de amigos y seguidores**, **localización** de los tweets, tendencias, etc.

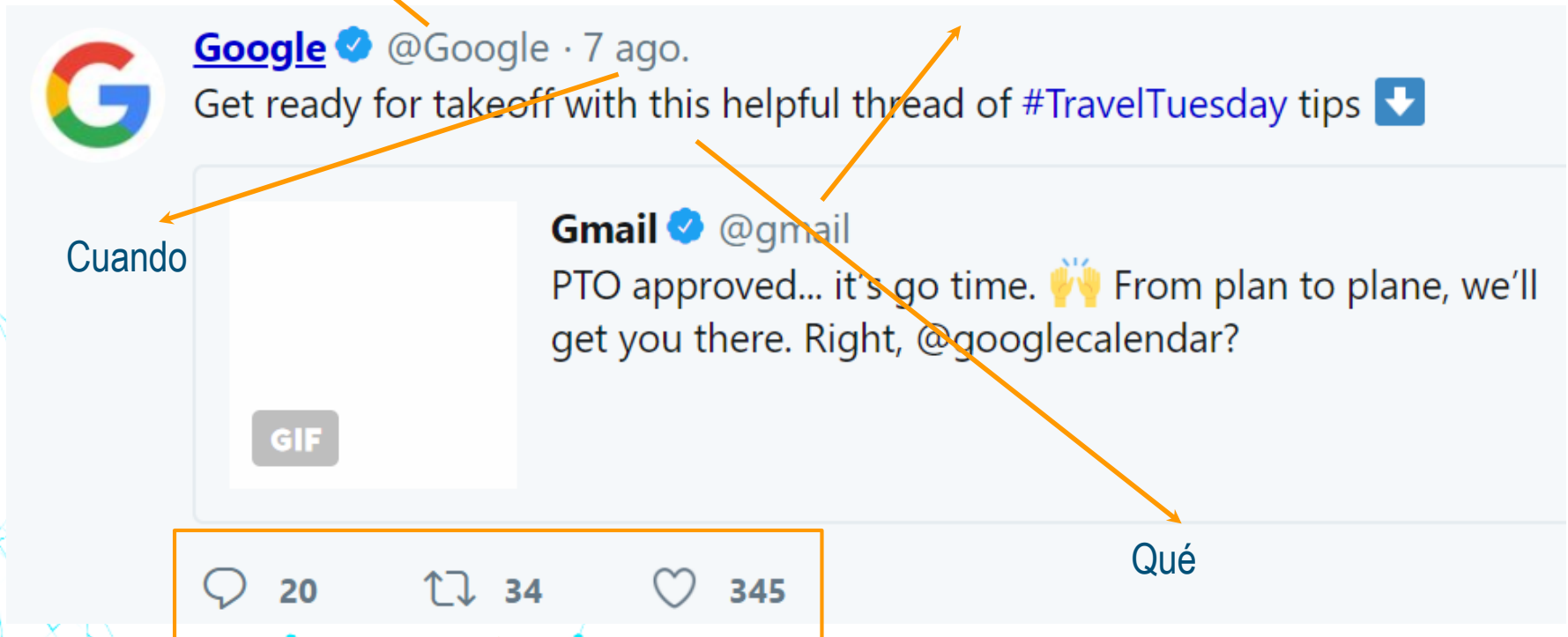
Quién

Con quién

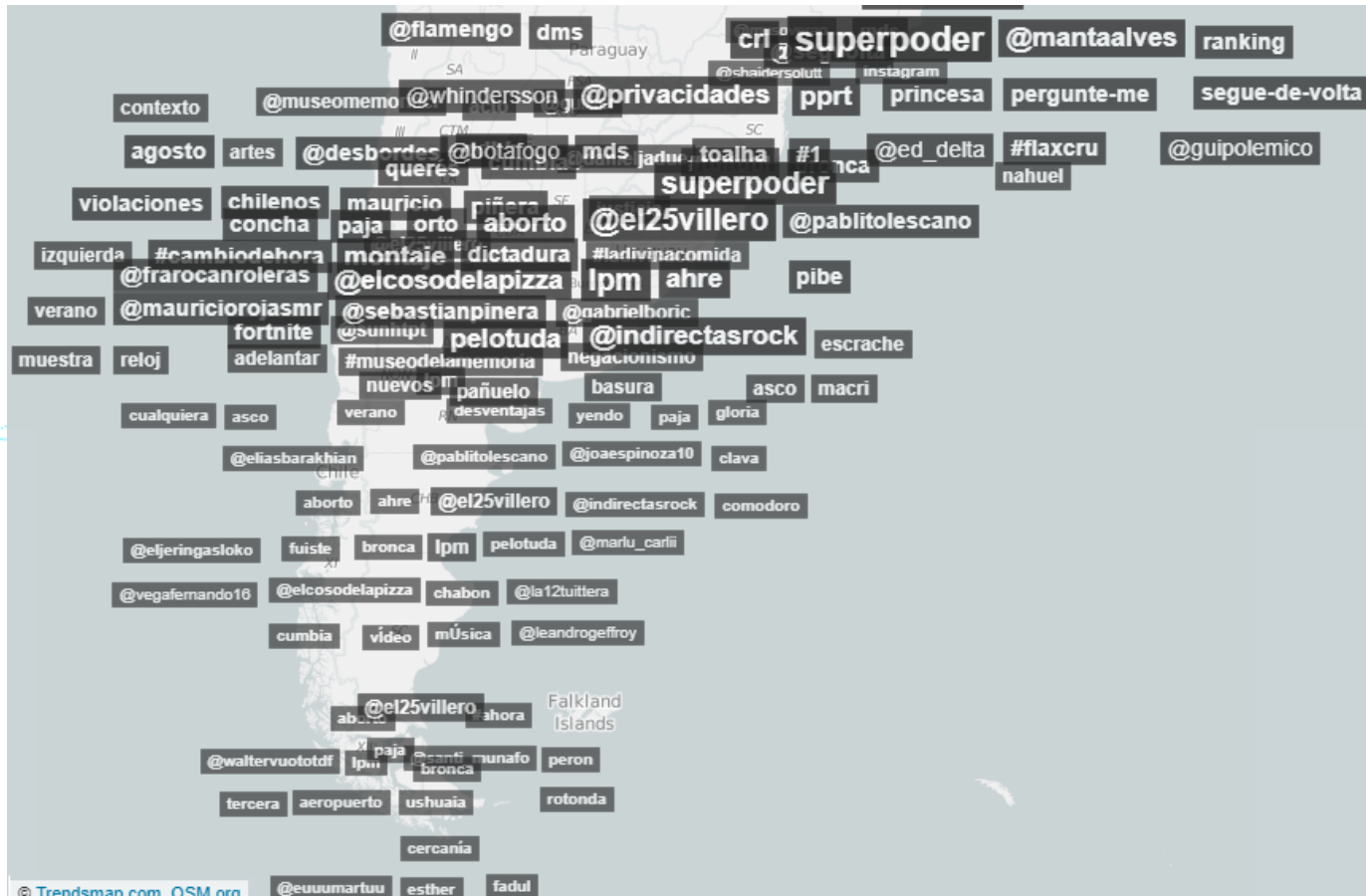
Cuando

Qué

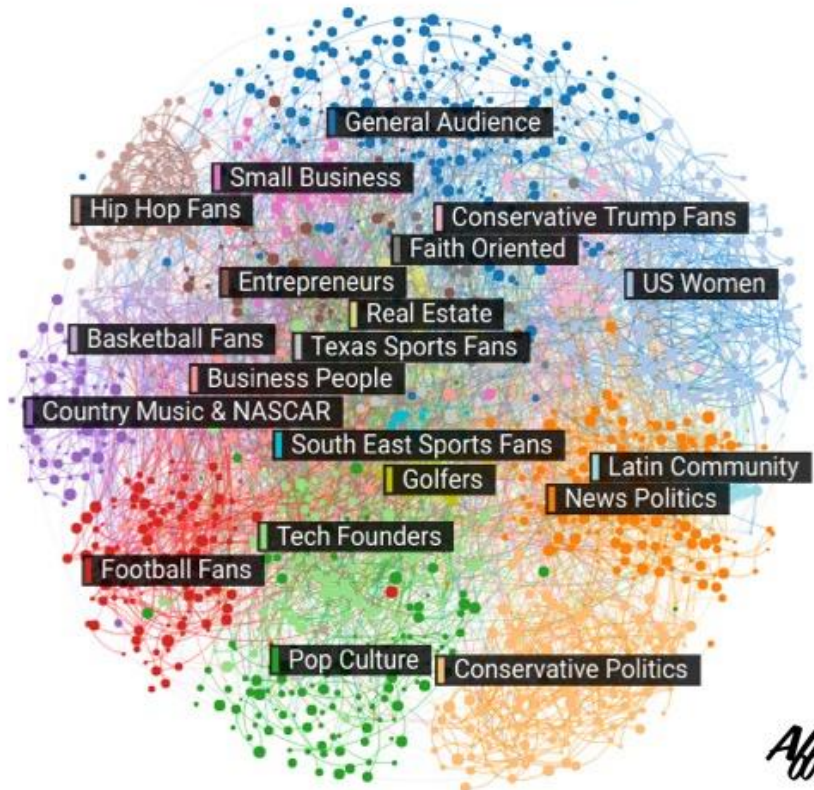
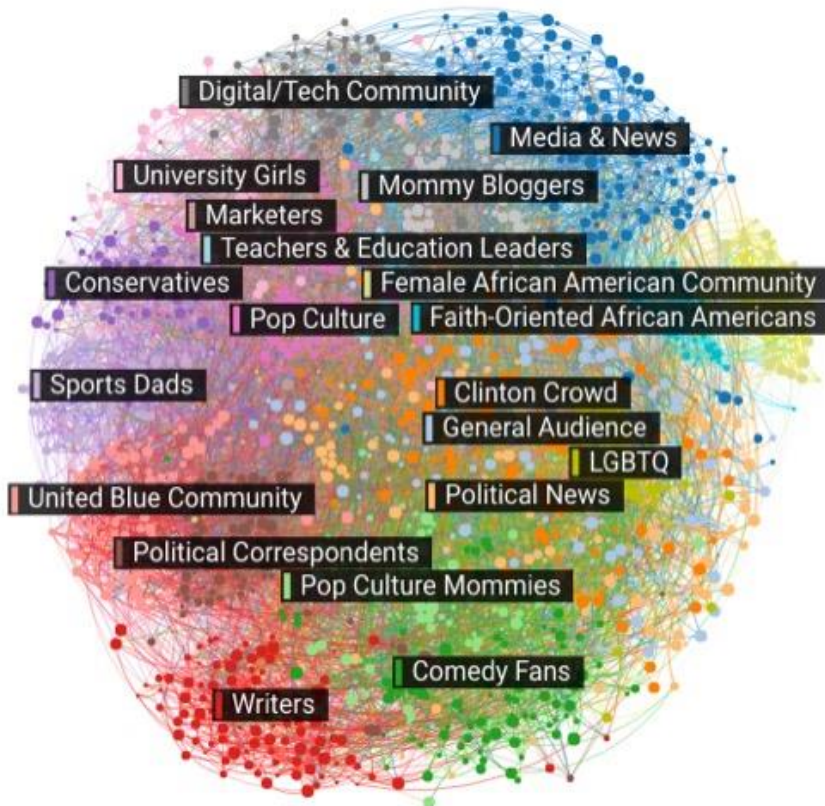
Repercusión



Trendsmap (entre otros) ofrece información sobre las palabras clave de los últimos tweets en una zona geográfica concreta.



# Redes Sociales – API Twitter



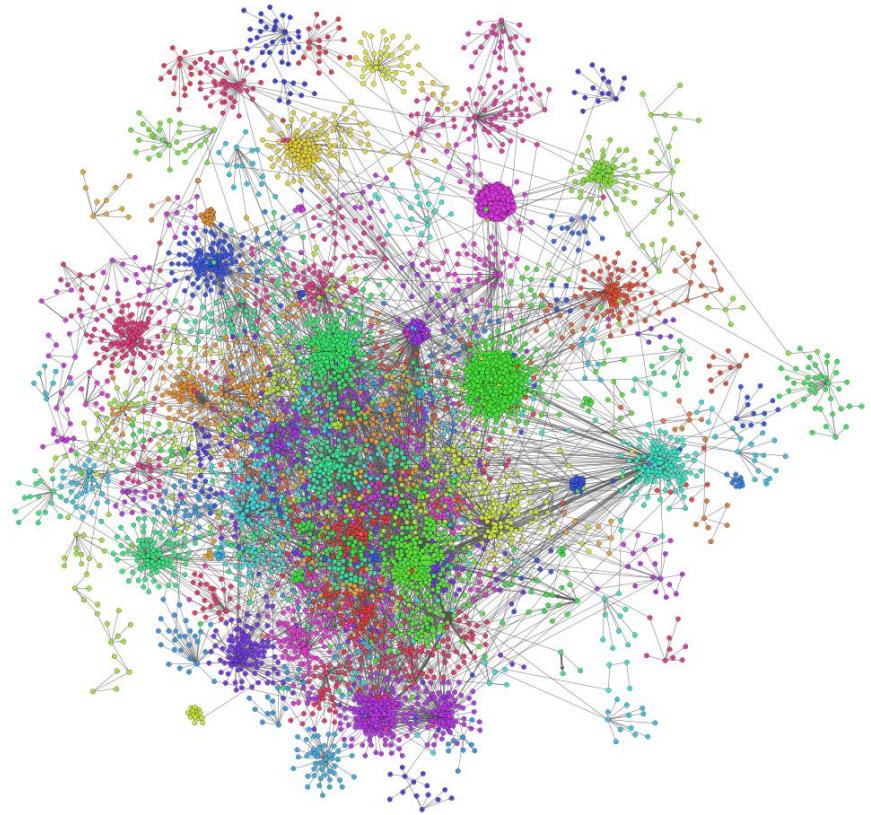
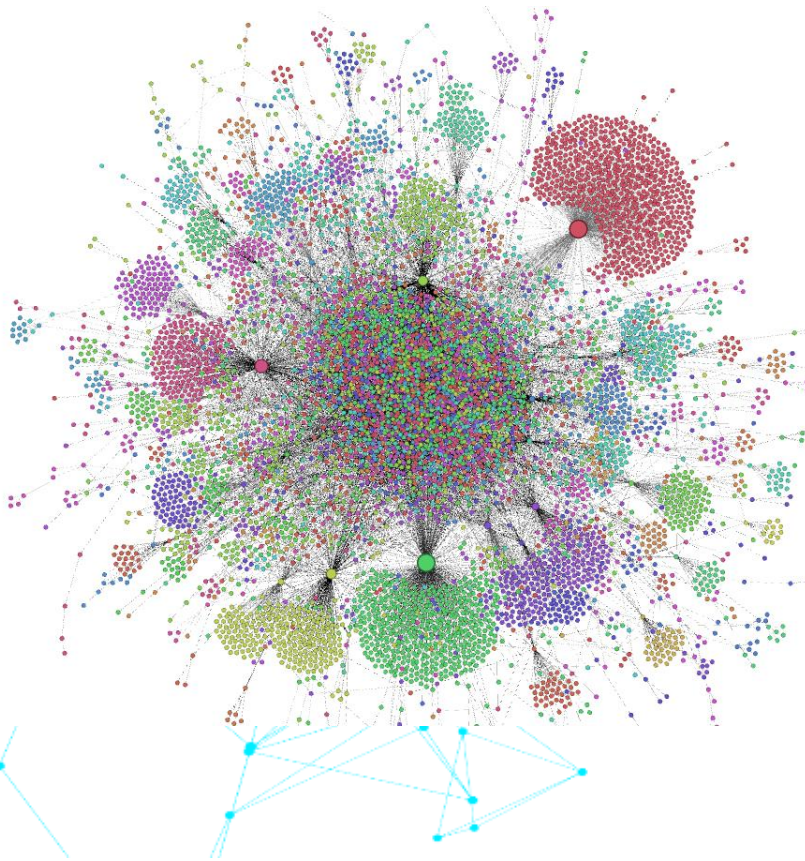
*Affinio*



# Redes Sociales – Datos de redes sociales

## Marketing Viral

¿Cómo debo seleccionar el target de personas para optimizar mi campaña de marketing viral?



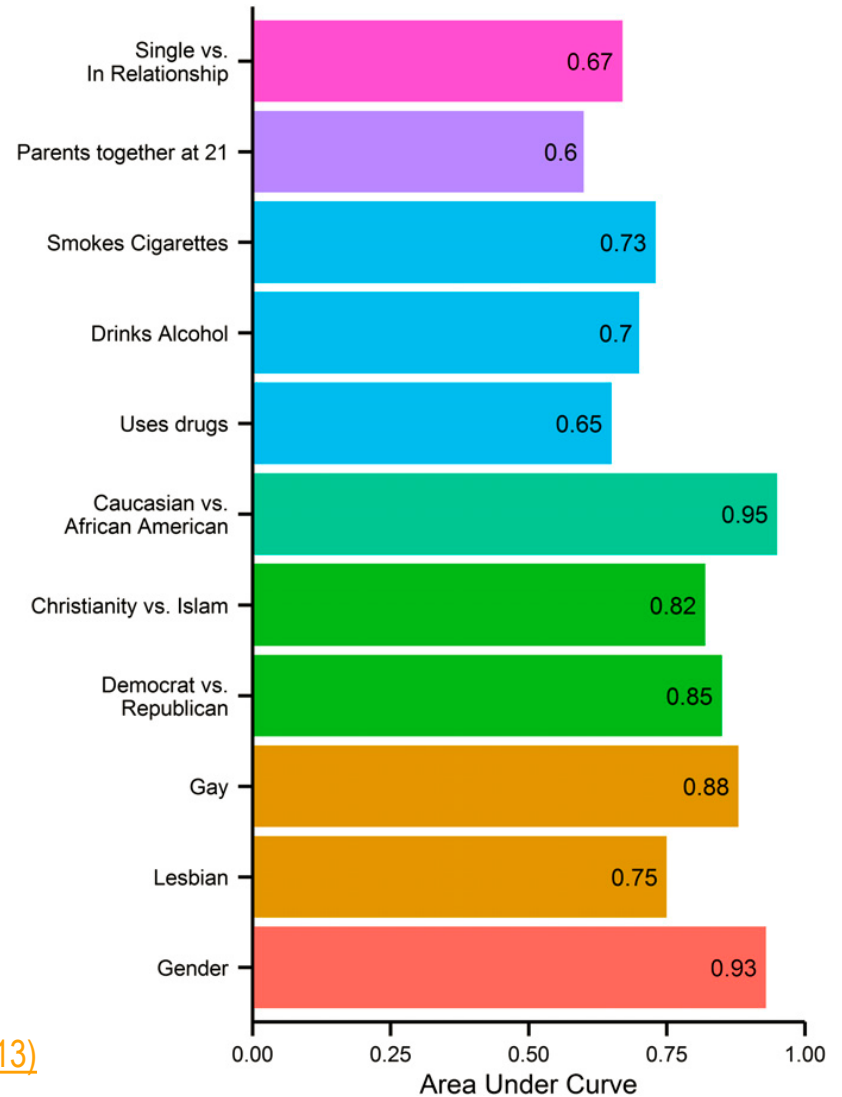
# Redes Sociales – Datos de redes sociales

**Es posible:**

Conocer la edad, sexo, orientación política (y mucho más con la actividad en la red social).

- ¿Orientación sexual?
- ¿Bebe?
- ¿Fuma?

M. Kosinski et al. (2013)



# Conecta Empleo

