

# Técnicas y herramientas de extracción de datos

## Introducción a la extracción de datos

Programa Experto en Análisis, Investigación y Comunicación de Datos

Juan Sixto Cesteros

[jsixto@deusto.es](mailto:jsixto@deusto.es)

DeustoTech - Deusto Institute of Technology, University of Deusto

<http://www.morelab.deusto.es>

Abril 18, 2015



University of Deusto

[www.deusto.es](http://www.deusto.es)



## ► Juan Sixto Cesteros

- *email: [jsixto@deusto.es](mailto:jsixto@deusto.es)*
- Ingeniero en Informática por la Universidad de Deusto
- Máster de Desarrollo e Integración de Soluciones Software
- Realizando el doctorado en el área del **Procesamiento del lenguaje natural, minería de datos y análisis de sentimientos** aplicada a las **redes sociales**.
- Parte del equipo de investigación de DeustoTech - Internet.
- <http://morelab.deusto.es/>

# Resumen



Introducción

Fuentes de datos y cómo organizarse

Cómo realizar búsquedas

Introducción al scrapping de datos

Leyes sobre datos

Técnicas de geolocalización

Trabajar con datos en excel

Técnicas avanzadas de limpieza de datos

# Introducción



- ▶ La Pirámide Invertida
  - ▶ Título
  - ▶ Cuerpo
  - ▶ Apoyo (Citas o datos)
  - ▶ Información secundaria
- ▶ La pirámide invertida es una estructura que sugiere escribir organizando la **información** con los **datos** presentados de mayor a menor importancia.



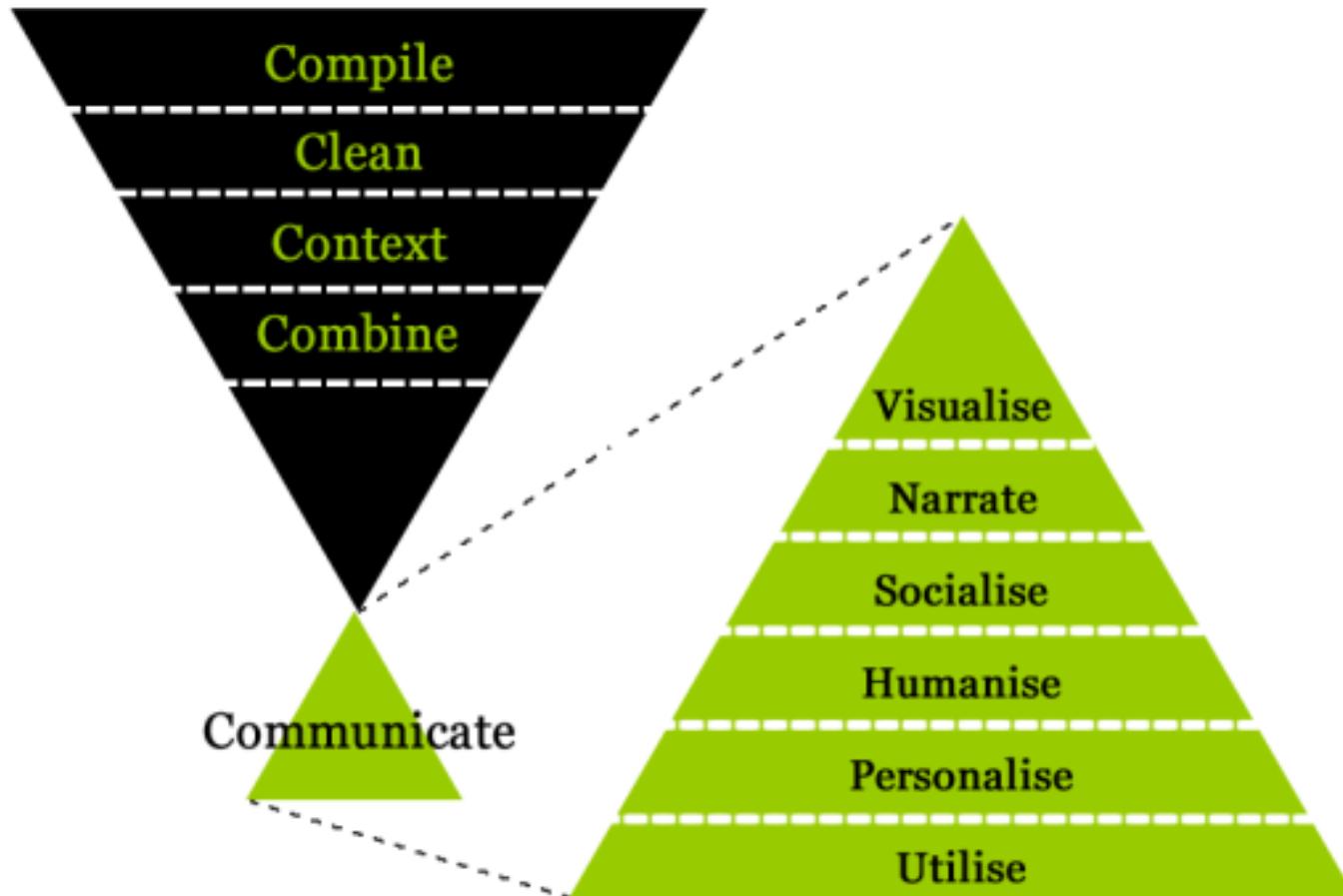
# Introducción



- ▶ La Pirámide Invertida
  - ▶ Título
  - ▶ Cuerpo
  - ▶ Apoyo (Citas o datos)
  - ▶ Información secundaria
- ▶ La pirámide invertida es una estructura que sugiere escribir organizando la **información** con los **datos** presentados de mayor a menor importancia.

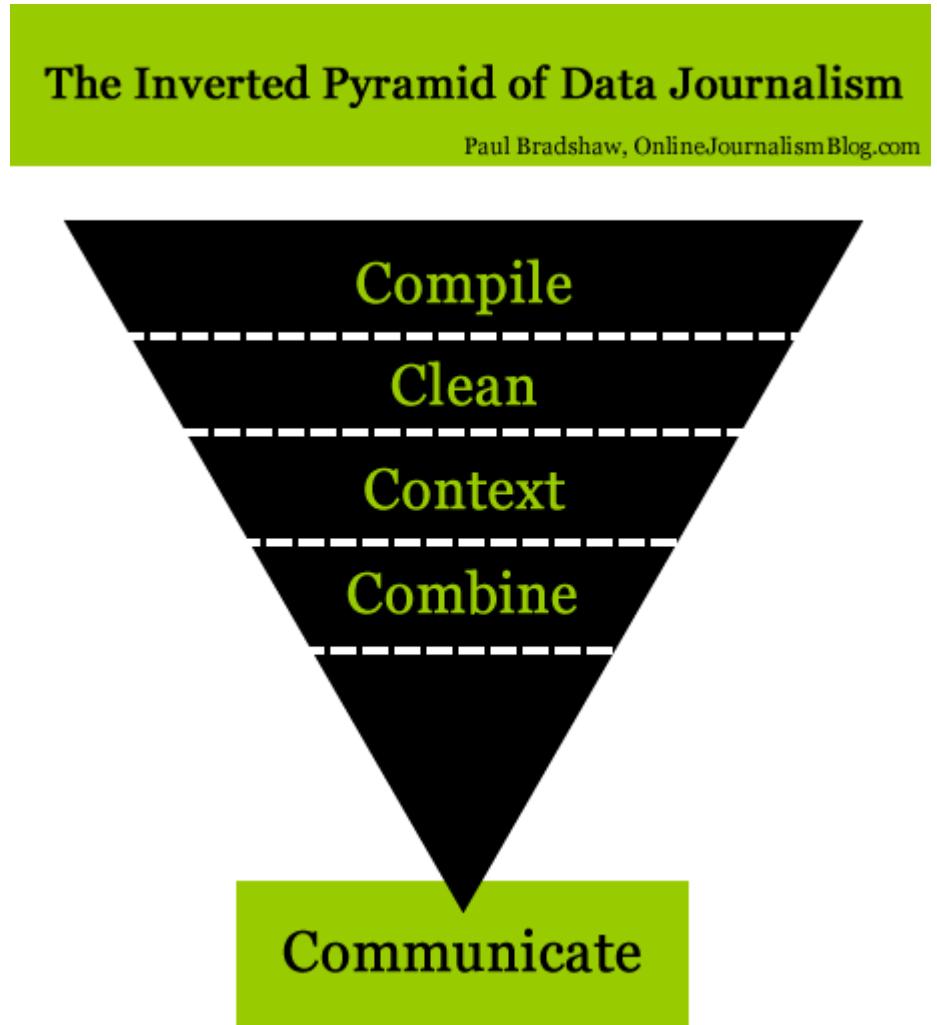


# Introducción





- ▶ La Pirámide Invertida del Periodismo de Datos
  - ▶ Paul Bradshaw





## ▶ Paul Bradshaw

- ▶ Fundador de *Online Journalism Blog*
- ▶ Profesor de periodismo online en la City University de Londres y en la Birmingham City University
- ▶ *The Online Journalism Handbook: Skills to survive and thrive in the digital age*

**Paul Bradshaw**



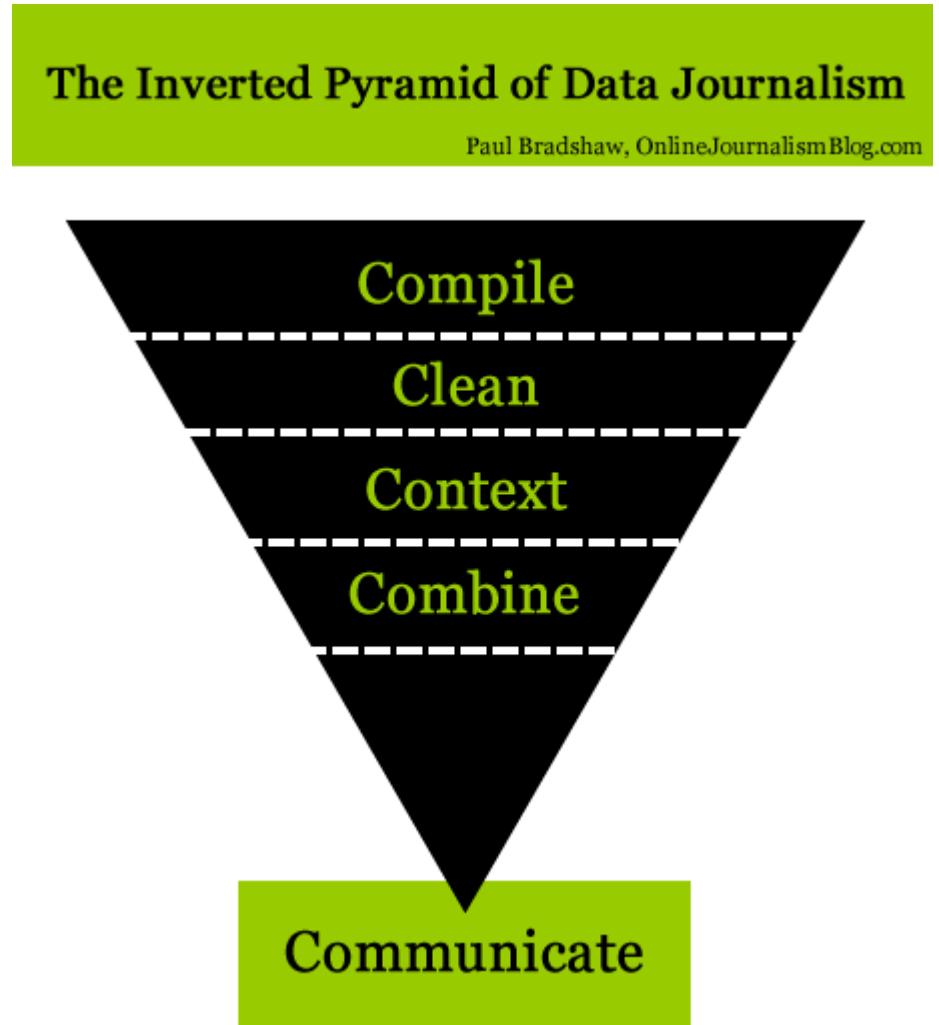
Bradshaw in January 2013

<b>Born</b>	Bolton
<b>Alma mater</b>	<a href="#">University of Central England</a>
<b>Occupation</b>	Journalist, Blogger Academic
<b>Employer</b>	Birmingham City University City University London
<b>Website</b>	
<a href="http://onlinejournalismblog.com">onlinejournalismblog.com</a>	

# Introducción



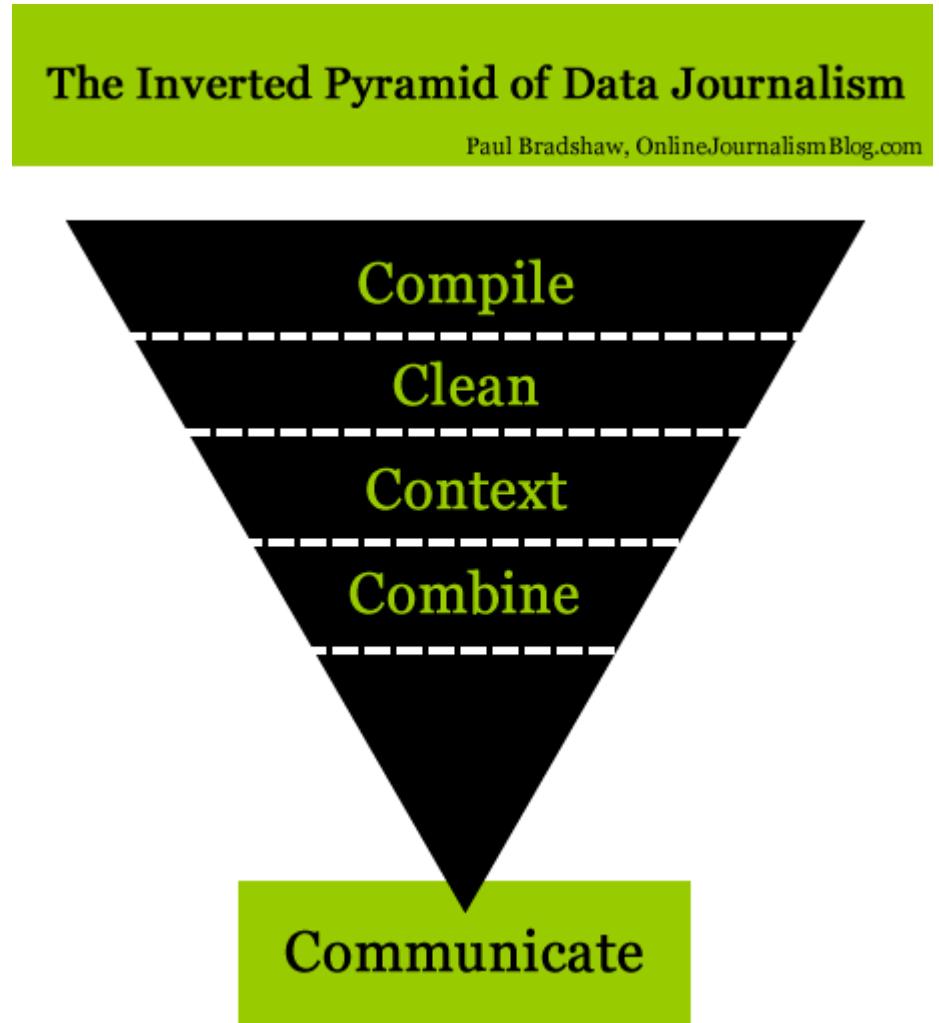
- ▶ La Pirámide Invertida del Periodismo de Datos
  - ▶ Paul Bradshaw
- ▶ Compilar
- ▶ Limpiar
- ▶ Contextualizar
- ▶ Combinar
- ▶ Comunicar



# Introducción

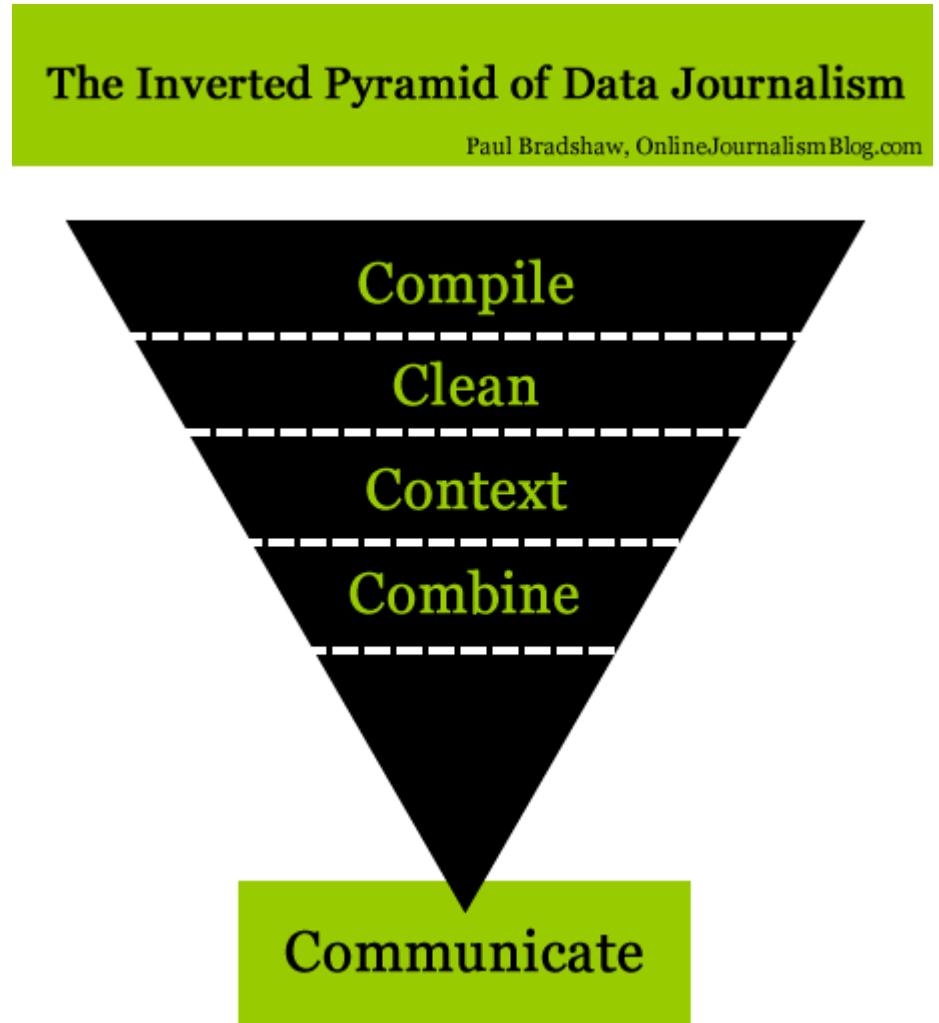


- ▶ La Pirámide Invertida del Periodismo de Datos
  - ▶ Paul Bradshaw
- ▶ Compilar
- ▶ Limpiar
- ▶ Contextualizar
- ▶ Combinar
- ▶ Comunicar





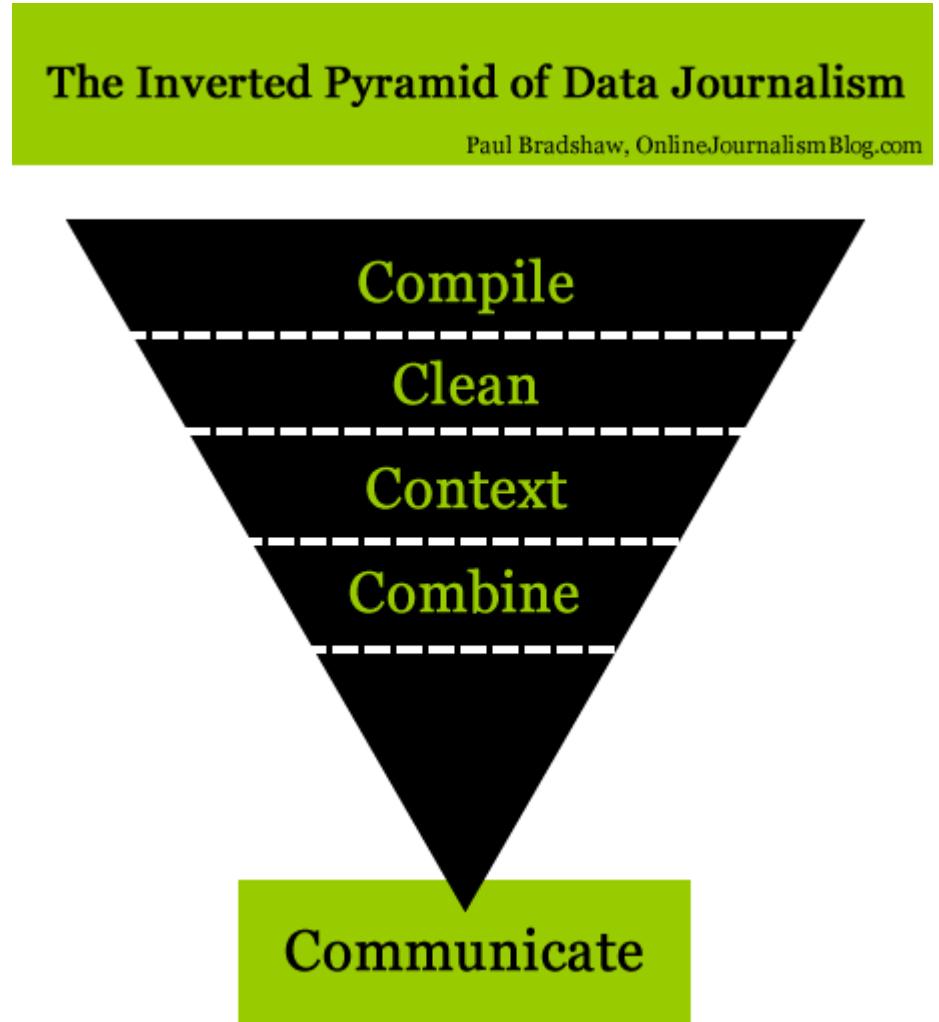
- ▶ **Compilar**
  - ▶ Análisis de Datos
  - ▶ Necesidad de Datos
    - ▶ Obtención
      - Solicitar
      - Búsqueda Web
      - Scraping
      - APIs
      - Formularios
      - Crowdsourcing



# Introducción



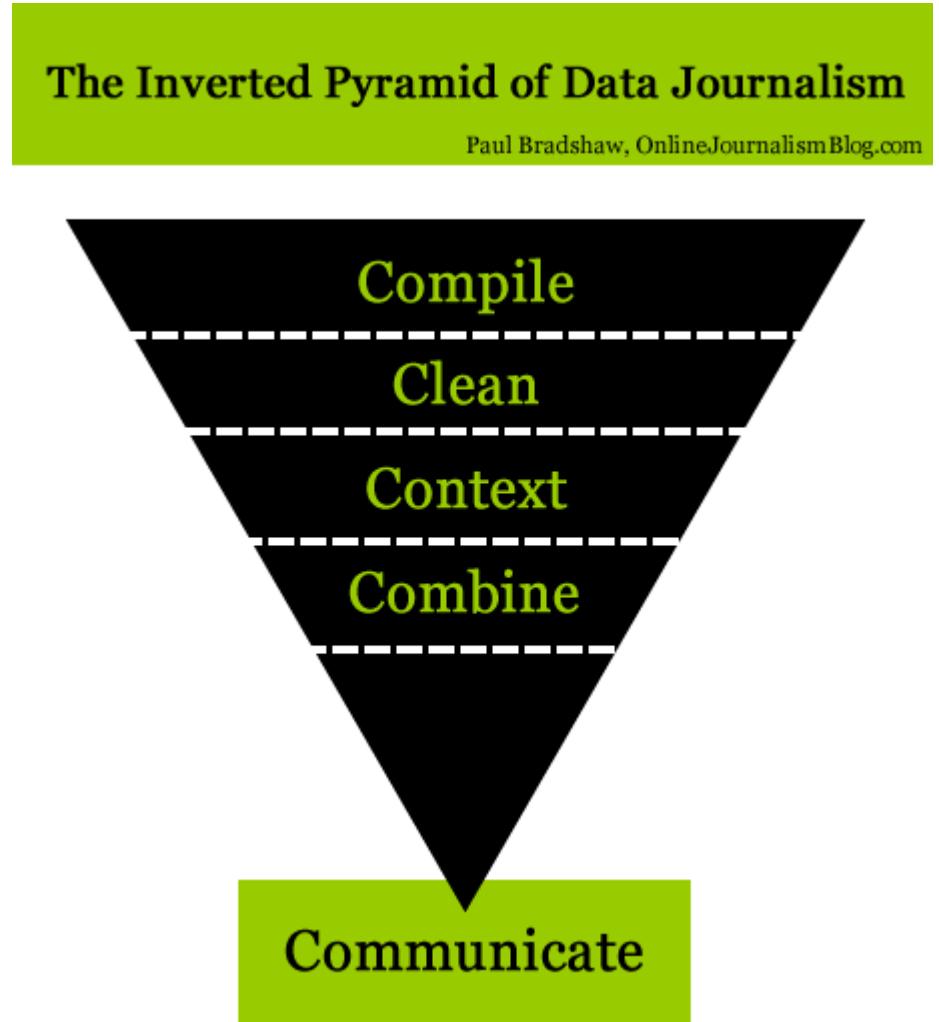
- ▶ Limpiar
  - ▶ Calidad de los Datos
  - ▶ Error humano
  - ▶ Formato
  - ▶ Duplicidades
  - ▶ Datos extraños



# Introducción

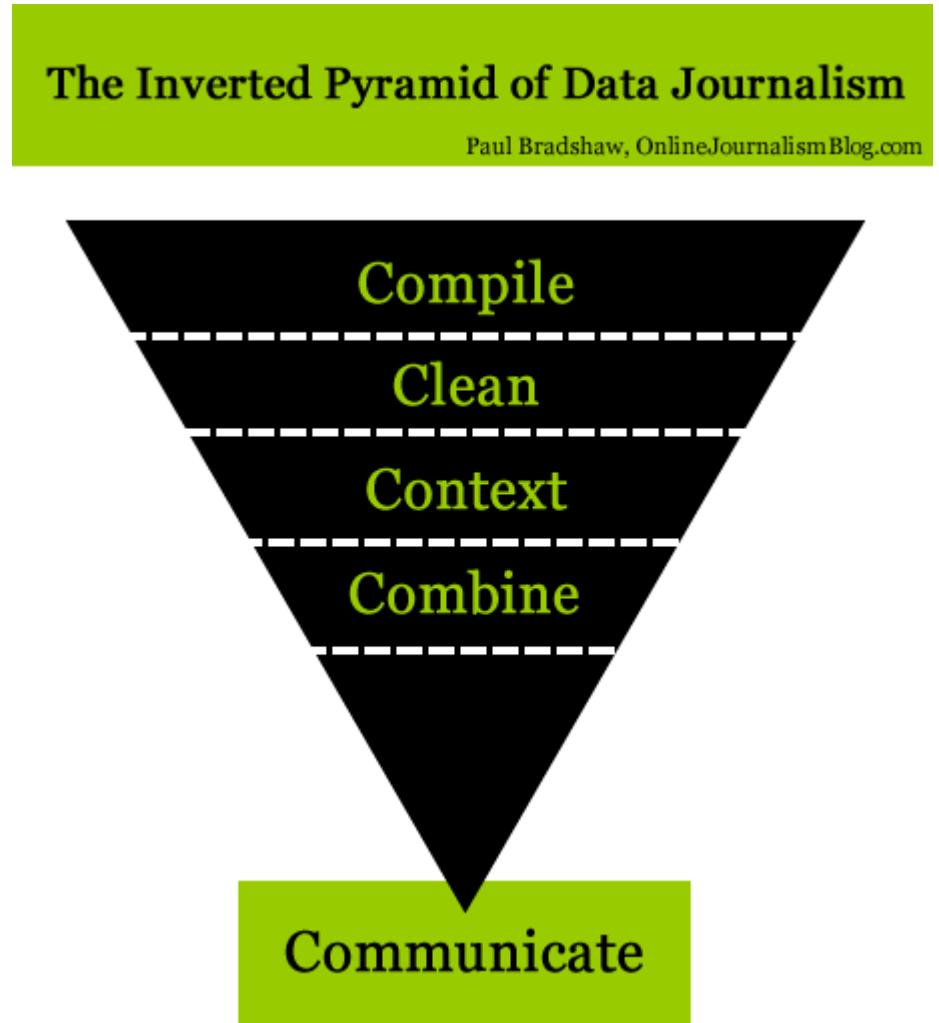


- ▶ Contextualizar
  - ▶ La información no puede siempre ser confiable
  - ▶ ¿Quién?
  - ▶ ¿Cuando?
  - ▶ ¿Por qué?
  - ▶ ¿Cómo? (Metodología)
  - ▶ Estadística





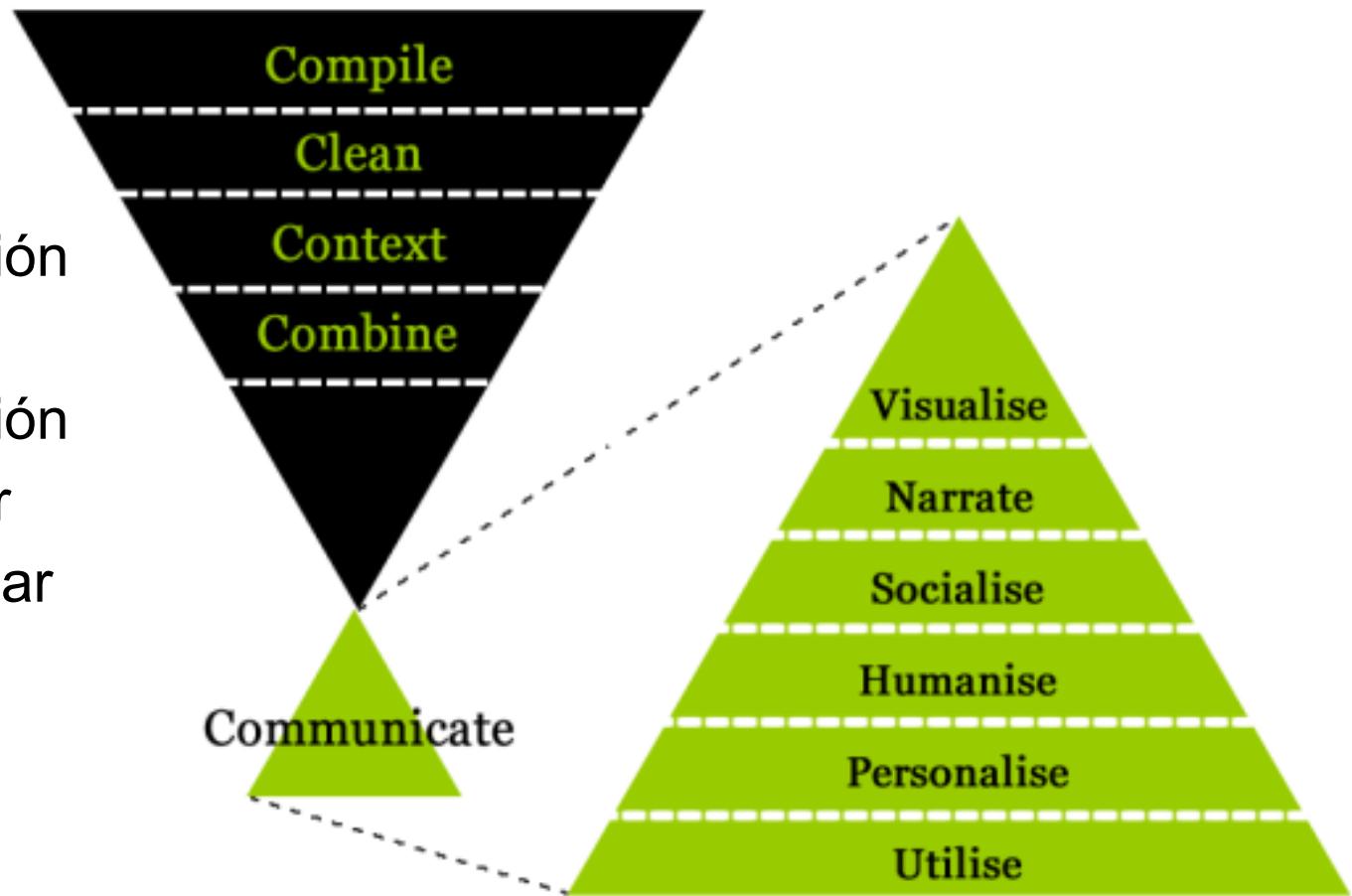
- ▶ Combinar
  - ▶ Múltiples datasets
  - ▶ Visualización
  - ▶ Extensión
  - ▶ ¿Ejemplos?
  - ▶ ¡Formato!



# Introducción



- ▶ Comunicar
  - ▶ Visualización
  - ▶ Narración
  - ▶ Socialización
  - ▶ Humanizar
  - ▶ Personalizar
  - ▶ Utilizar



# Fuentes de datos y cómo organizarse



- ▶ Fuentes de datos



- ▶ Fuentes de datos
  - ▶ Portales de datos oficiales
  - ▶ Organizaciones de Consumidores y Usuarios
  - ▶ Instituciones científicas o académicas
  - ▶ Motores de Búsqueda
  - ▶ Datos públicos
  - ▶ Datos en directo

# Fuentes de datos y cómo organizarse



- ▶ Formatos de datos
  - ▶ Data vs Machine-Readable Data
    - ▶ Preparados para ser procesados por una máquina
    - ▶ No preparados para ser mostrados a usuarios finales



- ▶ Formatos de datos
  - ▶ Data vs Machine-Readable Data
    - ▶ Preparados para ser procesados por una máquina
    - ▶ No preparados para ser mostrados a usuarios finales
  - ▶ Datos para visualización
    - ▶ HTML (Página Web)
    - ▶ Documento Word (Texto Formateado)
    - ▶ Documento PDF (Texto Maquetado)
    - ▶ JPEG (Imágenes)



- ▶ Formatos de datos
  - ▶ Data vs Machine-Readable Data
    - ▶ Preparados para ser procesados por una máquina
    - ▶ No preparados para ser mostrados a usuarios finales
  - ▶ Datos para visualización
    - ▶ HTML (Página Web)
    - ▶ Documento Word (Texto Formateado)
    - ▶ Documento PDF (Texto Maquetado)
    - ▶ JPEG (Imágenes)
  - ▶ Datos para procesar
    - ▶ XLS (Excel)
    - ▶ CSV (Abierto y sencillo)
    - ▶ XML (Lenguaje de marcado)
    - ▶ SQL (Bases de datos)



# Fuentes de datos y cómo organizarse

- ▶ Formatos de datos
  - ▶ Hojas de cálculo
    - ▶ Microsoft Excel
    - ▶ Apache OpenOffice
    - ▶ LibreOffice
    - ▶ Google Spreadsheets

The screenshot shows a Google Sheets document titled "Library Statistics". The spreadsheet contains data about equipment types and their counts from 2009 to 2013. The columns are labeled A through F, representing the year. The data includes items like Computers, Laptops, Scanners, Printers, Photocopiers, Carrels, Open tables, Study rooms, Large screen monitors, Projectors, and Kindles.

	A	B	C	D	E	F
1	Equipment Type	2009	2010	2011	2012	2013
2	Computers	4	4	5	5	6
3	Laptops	3	2	5	7	7
4	Scanners	0	1	1	1	1
5	Printers	1	1	1	1	1
6	Photocopiers	1	1	2	2	2
7	Carrels	45	45	30	30	30
8	Open tables	4	4	6	7	8
9	Study rooms	1	1	1	2	2
10	Large screen monitors	0	1	1	2	2
11	Projectors	1	1	1	1	1
12	Kindles	0	0	0	0	10
13						
14						
15						
16						



- ▶ Formatos de datos
  - ▶ CSV (*comma-separated values*)
    - ▶ Formato abierto y sencillo
    - ▶ En forma de tablas
    - ▶ Puede abrirse como Hojas de Cálculo

```
Title,Author,ISBN13,Pages
1984,George Orwell,978-0451524935,268
Animal Farm,George Orwell,978-0451526342,144
Brave New World,Aldous Huxley,978-0060929879,288
Fahrenheit 451,Ray Bradbury,978-0345342966,208
Jane Eyre,Charlotte Brontë,978-0142437209,532
Wuthering Heights,Emily Brontë,978-0141439556,416
Agnes Grey,Anne Brontë,978-1593083236,256
Walden,Henry David Thoreau,978-1420922615,156
Walden Two,B. F. Skinner,978-0872207783,301
"Eats, Shoots & Leaves",Lynne Truss,978-1592400874,209
```

# Fuentes de datos y cómo organizarse



- ▶ Formatos de datos
  - ▶ Lenguajes de Marcado
    - ▶ XML
    - ▶ HTML
  - ▶ Desarrollado por el World Wide Web Consortium (W3C)
  - ▶ Utilizado para almacenar datos en forma legible

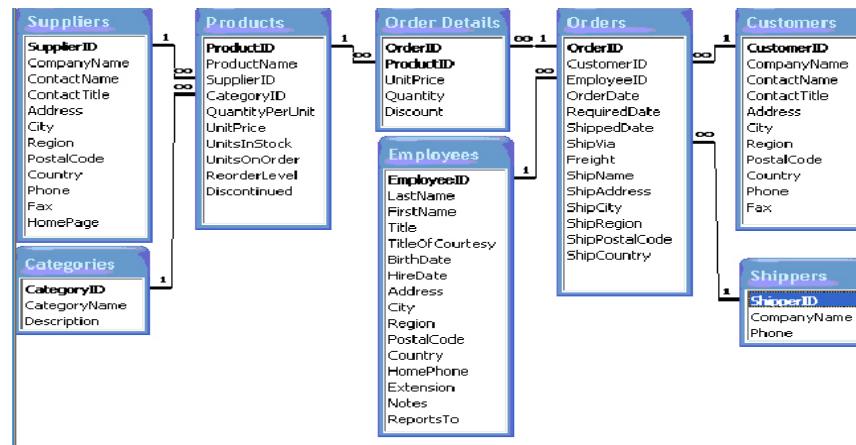
```
<?xml version="1.0"?>
<quiz>
  <qanda seq="1">
    <question>
      Who was the forty-second
      president of the U.S.A.?
    </question>
    <answer>
      William Jefferson Clinton
    </answer>
  </qanda>
  <!-- Note: We need to add
       more questions later.-->
</quiz>
```

**XML**

# Fuentes de datos y cómo organizarse



- ▶ Formatos de datos
  - ▶ Bases de datos
    - ▶ Colección de información organizada de forma que un programa de ordenador pueda seleccionar rápidamente los fragmentos de datos que necesite.
    - ▶ Se organizan por campos, registros y archivos.



# Cómo realizar búsquedas



- ▶ Muchos conjuntos de datos se encuentran indexados por los motores de búsqueda.

# Cómo realizar búsquedas



- ▶ Muchos conjuntos de datos se encuentran indexados por los motores de búsqueda.
- ▶ Algunos motores de búsqueda nos permiten buscar ficheros por tipo.
- ▶ Por ejemplo Google:
  - ▶ spreadsheets : *filetype:XLS* o *filetype:CSV*
  - ▶ bases de datos : *filetype:DB*
  - ▶ PDFs : *filetype:PDF*

# Cómo realizar búsquedas



- ▶ También podemos buscar por URLs o dominios concretos:
  - ▶ *Ejemplo: inurl:<download>*
  - ▶ *Ejemplo: site:opendata.euskadi.eus*

# Introducción al scraping de datos



- ▶ Web scraping es una técnica utilizada mediante programas de software para extraer información de sitios web



# Introducción al scraping de datos

- ▶ Web scraping es una técnica utilizada mediante programas de software para extraer información de sitios web
- ▶ Scraping es un método que te permite extraer datos escondidos en un documento, como páginas web y PDF, y los hace útiles para usarlos después.



# Introducción al scraping de datos

- ▶ Web scraping es una técnica utilizada mediante programas de software para extraer información de sitios web
- ▶ Scraping es un método que te permite extraer datos escondidos en un documento, como páginas web y PDF, y los hace útiles para usarlos después.
- ▶ Data scraping o ‘Raspado de Datos’
- ▶ Existen incontables herramientas de Scraping
  - ▶ Google Spreadsheets
  - ▶ Tábula



# Introducción al scraping de datos

- ▶ Google Spreadsheets
  - ▶ Servicio vía web de hojas de cálculo
  - ▶ Similar a Microsoft Excel o LibreOffice Calc
  - ▶ Maneja ficheros .xls .ods y .csv
  - ▶ Servicio Web integrado en Google Drive
  - ▶ Método sencillo para el scraping de datos

# Introducción al scraping de datos



- ▶ Google Spreadsheets
  - ▶ ImportHTML
  - ▶ <https://support.google.com/docs/answer/3093339>
  - ▶ Importar tablas y listas directamente desde la Web
  - ▶ Ejemplo:
    - ▶ IMPORTHTML("http://en.wikipedia.org/wiki/Demographics\_of\_India";"table";4)
- ▶ Ejercicio:
  - ▶ ¿Población de la C.A. de Euskadi por año de nacimiento, según el territorio histórico y el sexo en 2014?

# Introducción al scraping de datos



## ► Ejercicio:

- ▶ ¿Población de la C.A. de Euskadi por año de nacimiento, según el territorio histórico y el sexo en 2014?
- ▶ [www.eustat.es](http://www.eustat.es) (Euskal Estatistika Erakundea - Instituto Vasco de Estadística)
- ▶ =IMPORTHTML("http://www.eustat.es/elementos/ele0011400/ti\_Poblacion\_de\_la\_CA\_de\_Euskadi\_por\_ao\_de\_nacimiento\_segun\_el\_teritorio\_histrico\_y\_el\_sexo\_2014/tbl0011424\_c.html";"table";1)

# Introducción al scraping de datos



## ► Ejercicio:

- ▶ ¿Comunidades y ciudades autónomas de España por Densidad y Superficie?
- ▶ [http://es.wikipedia.org/wiki/Anexo:Comunidades\\_y\\_ciudades\\_aut%C3%B3nomas\\_de\\_Espa%C3%A1a](http://es.wikipedia.org/wiki/Anexo:Comunidades_y_ciudades_aut%C3%B3nomas_de_Espa%C3%A1a)
- ▶ =IMPORTHTML("http://es.wikipedia.org/wiki/Anexo:Comunidades\_y\_ciudades\_aut%C3%B3nomas\_de\_Espa%C3%A1a";"table";1)



# Introducción al scraping de datos

- ▶ Google Spreadsheets
  - ▶ ImportFEED
  - ▶ <https://support.google.com/docs/answer/3093337>
  - ▶ Extraer datos de RSS o ATOM
  - ▶ Sindicar o compartir contenido en la web.
  - ▶ Se utiliza para difundir información actualizada frecuentemente a usuarios que se han suscrito a la fuente de contenidos
  - ▶ Ejemplo:
    - ▶ =IMPORTFEED("http://news.google.com/?output=atom")

# Introducción al scraping de datos



- ▶ Google Spreadsheets
  - ▶ ImportFEED
  - ▶ Ejercicio:
    - ▶ Crear Feed Rss con noticias de un periódico cualquiera.



# Introducción al scraping de datos



- ▶ Google Spreadsheets
  - ▶ ImportXML
  - ▶ <https://support.google.com/docs/answer/3093342>
  - ▶ Importa datos en formato XML, HTML, CSV, TSV, RSS y ATOM XML feeds.
  - ▶ Utiliza XPath
  - ▶ <http://www.w3schools.com/xpath/>



# Introducción al scraping de datos

- ▶ Tábula
  - ▶ Escraper de datos para PDFs
  - ▶ <http://tabula.technology/>
  - ▶ Instalar Tábula

# Introducción al scraping de datos



## Tabula



Tabula is a tool for liberating data tables locked inside PDF files.

[View the Project on GitHub](#)  
tabulapdf/tabula

[Download for Windows](#)

[Download for Mac](#)

[View source on GitHub](#)

**Current Version:** 0.9.7

**Other Versions:** [pre-releases & archives](#)

**Need help?** Open an issue on [Github](#).

We'd love to hear from you! Say hi on Twitter at [@TabulaPDF](#)



## How Can Tabula Help Me?

If you've ever tried to do anything with data provided to you in PDFs, you know how painful it is — there's no easy way to copy-and-paste rows of data out of PDF files. Tabula allows you to extract that data into a CSV or Microsoft Excel spreadsheet using a simple, easy-to-use interface. Tabula works on Mac, Windows and Linux.

## Who Uses Tabula?

Tabula is used to power investigative reporting at news organizations of all sizes, including [ProPublica](#), [The Times of London](#), [Foreign Policy](#), [La Nación \(Argentina\)](#) and the [St. Paul \(MN\) Pioneer Press](#).

Grassroots organizations like [SchoolCuts.org](#) rely on Tabula to turn clunky documents into human-friendly public resources.

And researchers of all kinds use Tabula to turn PDF reports into Excel spreadsheets, CSVs, and JSON files for use in analysis and database applications.

## How to Install Tabula

Note: You'll need a copy of Java installed. You can [download Java here](#).

1. Download the version of Tabula for your operating system:
  - **Windows:** [tabula-win.zip](#)
  - **Mac OS X:** [tabula-mac.zip](#)
    - OSX 10.8+ users: if you have issues opening the app, [see the two notes at bottom of this page](#)
  - **Linux/Other:** [tabula-jar.zip](#) (view README.txt inside for instructions)
2. Extract the zip file. (Instructions: [Windows](#), [Mac](#))
3. Go into the folder you just extracted. Run the "Tabula" program inside.
4. A web browser will open. If it doesn't, open your web browser, and go to <http://localhost:8080>. There's Tabula!

# Introducción al scraping de datos



Fork 129    Star 1,221



## Tabula

Liberate data tables trapped inside PDF files

Upload a PDF

Auto-Detect Tables

Table auto-detection can be time-consuming, especially for large PDFs.

Choose File

No file chosen

Uploaded files

[Es3057mar\\_A.pdf](#) (2015-04-17 10:51)

[Es3057mar\\_A.pdf](#) (2015-04-13 17:22)

Tabula was created by [Manuel Aristarán](#) with the help of [ProPublica](#), [La Nación DATA](#) and [Knight-Mozilla OpenNews](#)



# Introducción al scraping de datos

Tabula is experimental software    Home    About

Page 1



Page 2



Page 3



Page 4



CIS  
Estudio n°3057. BARÓMETRO DE MARZO 2015

Marzo 2015

## Pregunta 7A

¿Cuál es, a su juicio, el principal problema que existe actualmente en España? ¿Y el segundo? ¿Y el tercero?  
(MULTIRRESPUESTA).

El paro	80,3
Las drogas	0,3
La inseguridad ciudadana	2,2
El terrorismo, ETA	0,2
Las infraestructuras	0,2
La sanidad	11,8
La vivienda	1,9
Los problemas de índole económica	24,9
Los problemas relacionados con la calidad del empleo	3,4
Los problemas de la agricultura, ganadería y pesca	0,1
La corrupción y el fraude	50,8
Los perones	2,2
Losas políticas en general, los partidos y la política	20,0
La Administración de Justicia	1,6
Los problemas de índole social	10,4
El racismo	0,0
La inmigración	1,9
La violencia contra la mujer	0,2
Los problemas relacionados con la juventud	2,7
La crisis de valores	2,4
La educación	9,2
Los problemas medioambientales	0,2
El Gobierno y partidos o políticos concretos	2,4
El funcionamiento de los servicios públicos	0,6
Los nacionalismos	0,7
Los problemas relacionados con la mujer	0,2
El terrorismo internacional	0,8
Las preocupaciones y situaciones personales	0,1
Estatutos de autonomía	0,2
Las negociaciones con ETA	0,0
Reforma Laboral	0,1
"Los recortes"	3,9
Los bancos	1,2
La subida del IVA	1,1
Los desahucios	2,4
El fraude fiscal	0,2
Las hipotecas	0,0
La Monarquía	0,1
Subida de tarifas energéticas	0,1
Otras respuestas	4,6
Ninguno	0,1
N.S.	0,6
N.C.	0,1
(N)	(2,476)

Download All Data

Clear All Selections

Preview Data Automatically? [?](#)

[Help](#)

Pág 4

CIS  
Estudio n°3057. BARÓMETRO DE MARZO 2015

Marzo 2015



# Introducción al scraping de datos

## Extracted tabular data

El paro	80,3
Las drogas	0,3
La inseguridad ciudadana	2,2
El terrorismo, ETA	0,2
Las infraestructuras	0,2
La sanidad	11,8
La vivienda	1,9
Los problemas de índole económica	24,9
Los problemas relacionados con la calidad del empleo	3,4
Los problemas de la agricultura, ganadería y pesca	0,1
La corrupción y el fraude	50,8
Las pensiones	2,2
Los/as políticos/as en general, los partidos y la política	20,0
La Administración de Justicia	1,6

[Copy to Clipboard](#)[Download CSV](#)[Close](#)[x Advanced Options:](#)

Extraction Method:

[Original](#)[Spreadsheet](#)[Download Data As...](#)

# Introducción al scraping de datos



## ► Ejercicios

- Barómetro de Febrero 2015 del CIS: Principales problemas de España
- CIS Enero 2015: Valoración de Líderes Políticos



- ▶ Derechos sobre los Datos
  - ▶ Consideraciones generales
    - ▶ Antes de publicar, ¿son ya públicos los datos?.



- ▶ Derechos sobre los Datos
  - ▶ Consideraciones generales
    - ▶ Antes de publicar, ¿son ya públicos los datos?.
    - ▶ Solicitudes formales
      - Suelen llevar tiempo. Cuanto antes mejor.
      - Consultar tiempos mínimos y máximos de solicitud.



- ▶ Derechos sobre los Datos
  - ▶ Consideraciones generales
    - ▶ Antes de publicar, ¿son ya públicos los datos?.
    - ▶ Solicituds formales
      - Suelen llevar tiempo. Cuanto antes mejor.
      - Consultar tiempos mínimos y máximos de solicitud.
  - ▶ Conocer los derechos sobre los datos
    - Obligación de contestar.



# Leyes sobre datos

- ▶ Derechos sobre los Datos
  - ▶ Consideraciones generales
    - ▶ Antes de publicar, ¿son ya públicos los datos?.
    - ▶ Solicitudes formales
      - Suelen llevar tiempo. Cuanto antes mejor.
      - Consultar tiempos mínimos y máximos de solicitud.
  - ▶ Conocer los derechos sobre los datos
    - Obligación de contestar.
  - ▶ Sencillez
    - Mejor solicitar consultas sencillas sobre los datos



# Leyes sobre datos

- ▶ Derechos sobre los Datos
  - ▶ Consideraciones generales
    - ▶ Antes de publicar, ¿son ya públicos los datos?.
    - ▶ Solicitudes formales
      - Suelen llevar tiempo. Cuanto antes mejor.
      - Consultar tiempos mínimos y máximos de solicitud.
  - ▶ Conocer los derechos sobre los datos
    - Obligación de contestar.
  - ▶ Sencillez
    - Mejor solicitar consultas sencillas sobre los datos
  - ▶ Solicitudes Internacionales
    - Internet nos permite solicitar datos a todas partes



# Leyes sobre datos

- ▶ Derechos sobre los Datos en España
  - ▶ Ley de Transparencia, Acceso a la Información Pública y Buen Gobierno
  - ▶ Leyes de prensa
  - ▶ Ley Orgánica de Protección de Datos



# Leyes sobre datos

- ▶ Peligros sobre los datos
  - ▶ Origen de los datos
  - ▶ Solicitar siempre datos en crudo
  - ▶ Casos de Copyright, uso y liberación de datos
    - ▶ Licencias
      - Dominio público
      - Licencias permisivas o sólo de atribución
      - Licencias copyleft, recíprocas o de compartir por igual
  - ▶ Otras leyes.



# Técnicas de geolocalización

- ▶ Geolocalización
  - ▶ ¿Dónde?



# Técnicas de geolocalización

- ▶ Geolocalización
  - ▶ ¿Donde?
  - ▶ “El poder de las redes sociales: Geolocalización para noticias”



# Técnicas de geolocalización

- ▶ Geolocalización
  - ▶ ¿Donde?
  - ▶ “El poder de las redes sociales: Geolocalización para noticias”
  - ▶ Nos permite limitar el alcance de las noticias



# Técnicas de geolocalización

- ▶ Geolocalización
  - ▶ ¿Donde?
  - ▶ “El poder de las redes sociales: Geolocalización para noticias”
  - ▶ Nos permite limitar el alcance de las noticias
  - ▶ Conocer su origen



# Técnicas de geolocalización

- ▶ Geolocalización
  - ▶ ¿Donde?
  - ▶ “El poder de las redes sociales: Geolocalización para noticias”
  - ▶ Nos permite limitar el alcance de las noticias
  - ▶ Conocer su origen
  - ▶ Validar la localización para evitar engaños

# Técnicas de geolocalización





# Técnicas de geolocalización





# Técnicas de geolocalización

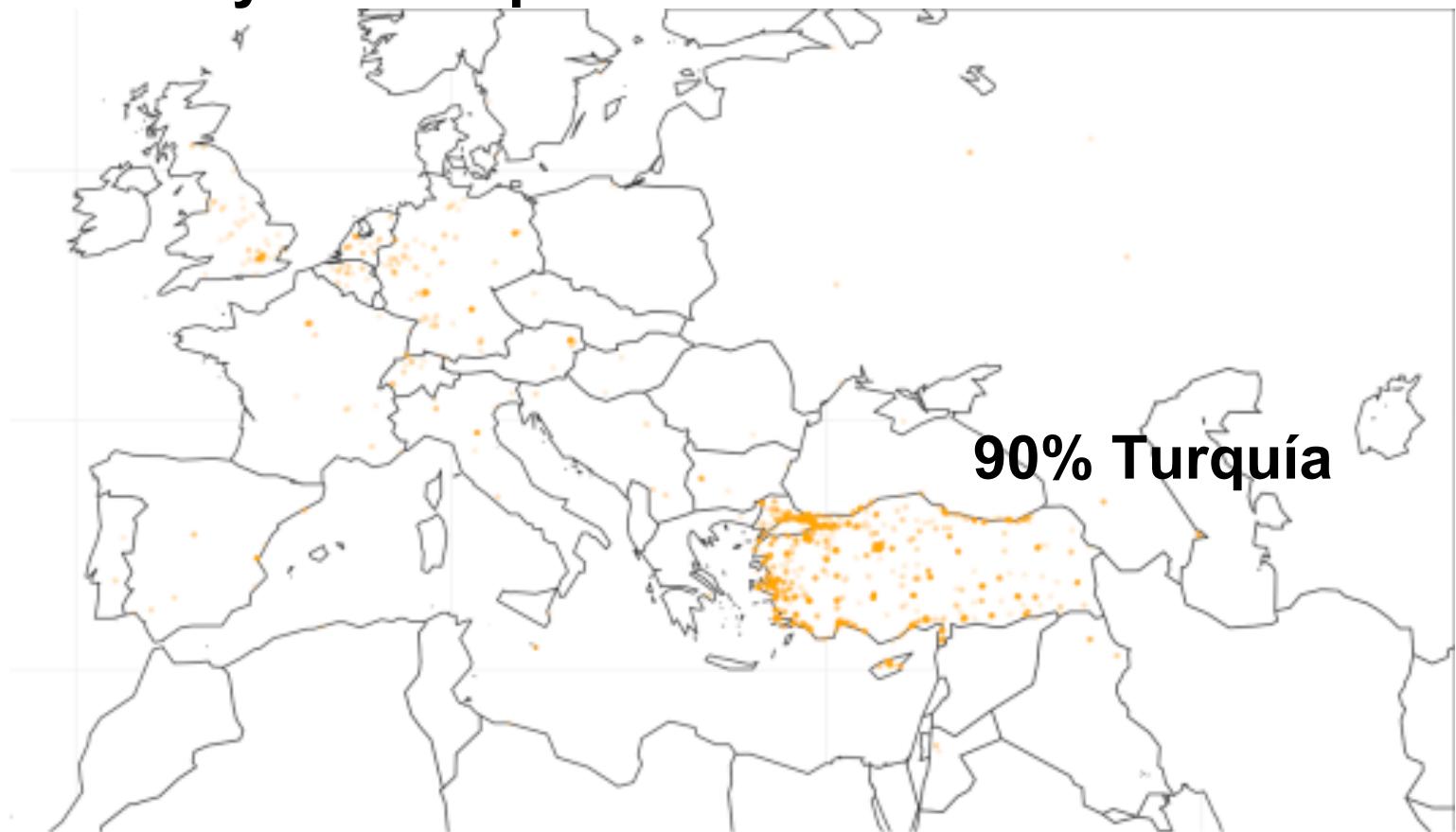
#BugünTelevizyonlarıKapat





# Técnicas de geolocalización

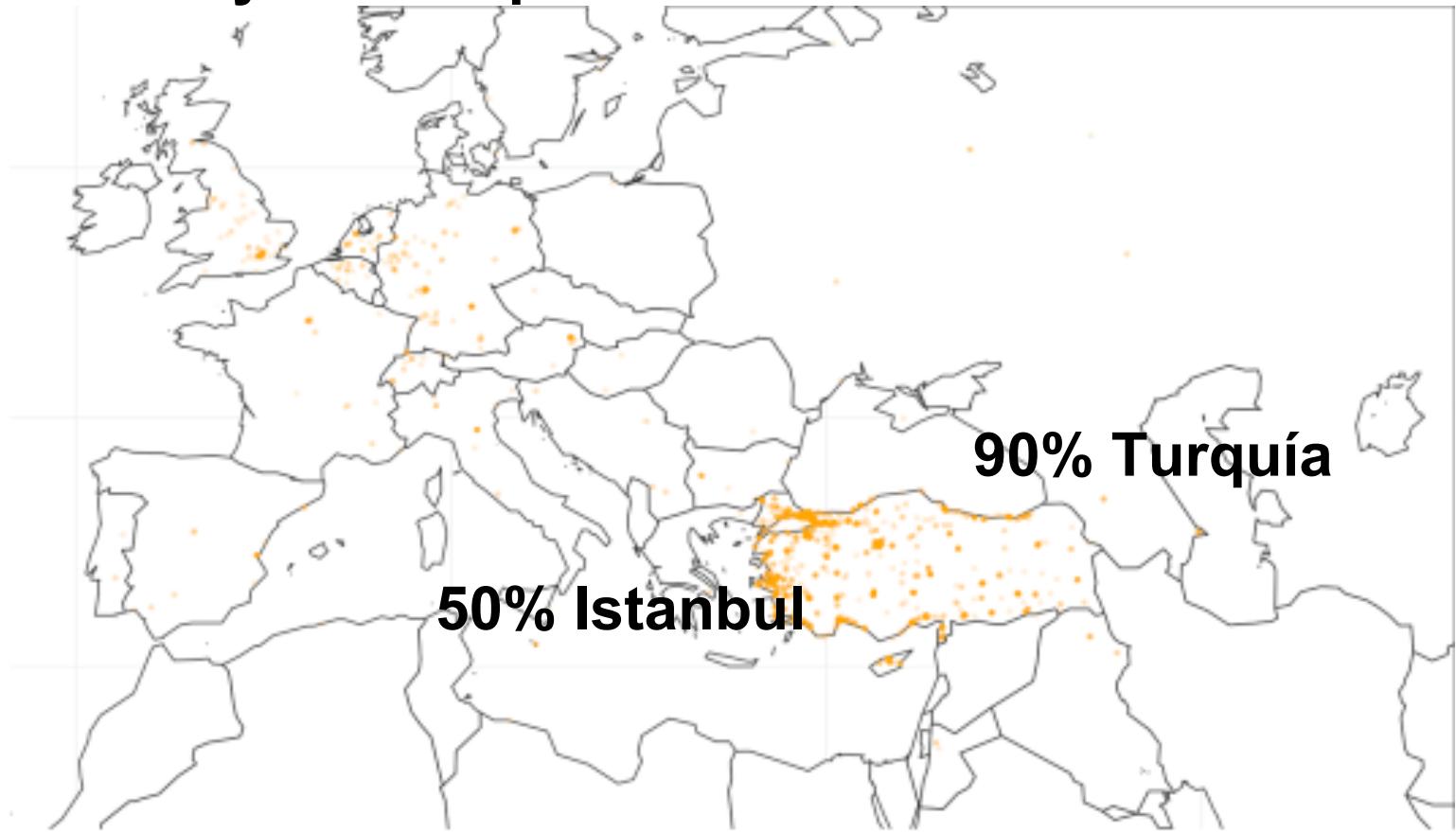
#BugünTelevizyonlarıKapat





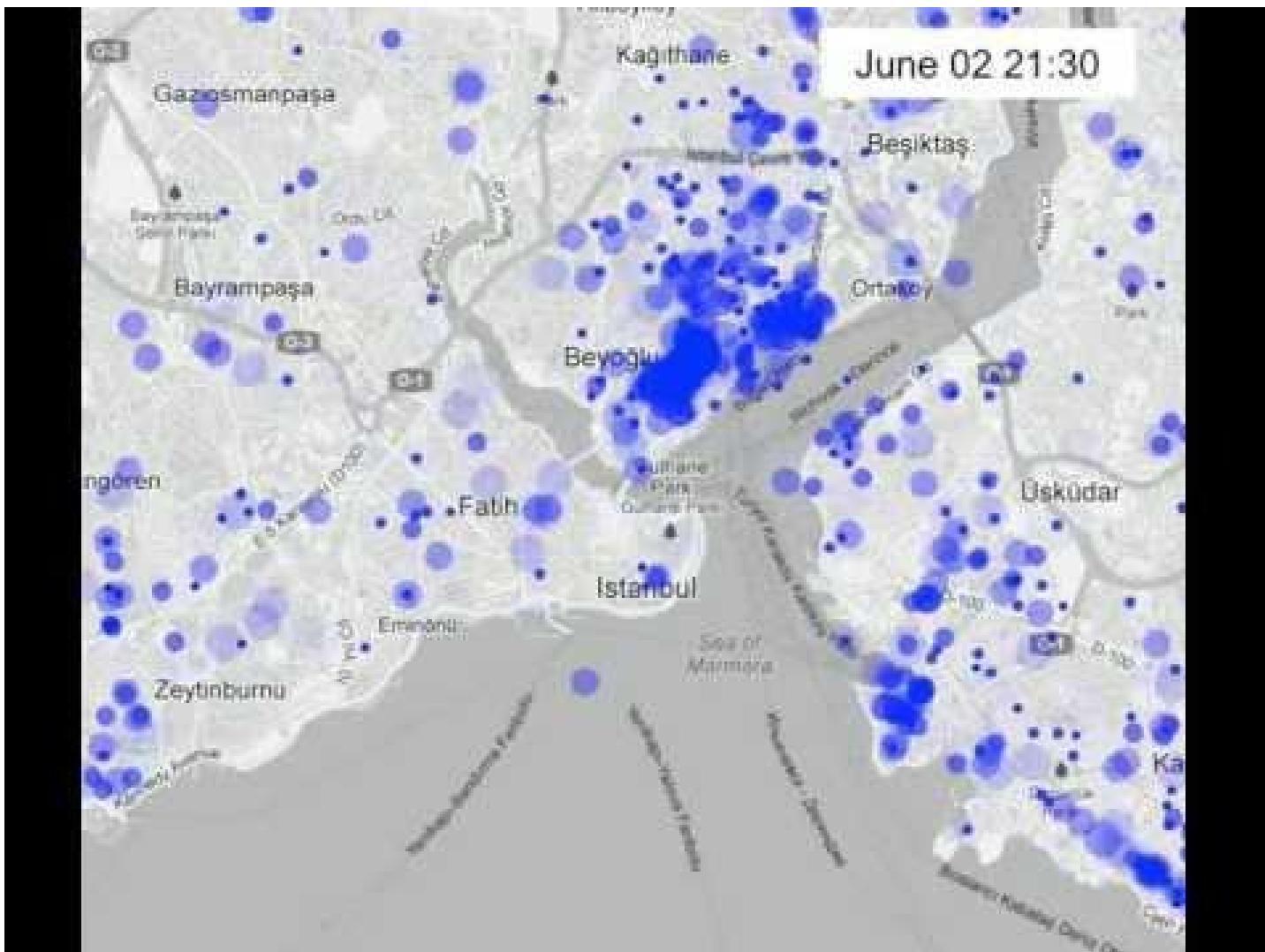
# Técnicas de geolocalización

#BugünTelevizyonlarıKapat





# Técnicas de geolocalización





# Técnicas de geolocalización

- **Sólo el 1% de los usuarios genera alrededor del 80 por ciento de todos los retweets**
- **Tres cuartas partes de los usuarios que hablan de las protestas no dieron ningún retweets en absoluto.**



# Técnicas de geolocalización

- ▶ Caso de Estudio - Bellingcat
  - ▶ <https://www.bellingcat.com/>



# Técnicas de geolocalización

- ▶ Caso de Estudio - Bellingcat
  - ▶ <https://www.bellingcat.com/>
  - ▶ Open source investigations tools and techniques



# Técnicas de geolocalización

- ▶ Caso de Estudio - Bellingcat
  - ▶ <https://www.bellingcat.com/>
  - ▶ Open source investigations tools and techniques
  - ▶ Confirmar la localización de un vídeo sin salir de casa



# Técnicas de geolocalización

- ▶ Caso de Estudio - Bellingcat
  - ▶ <https://www.bellingcat.com/>
  - ▶ Open source investigations tools and techniques
  - ▶ Confirmar la localización de un vídeo sin salir de casa
  - ▶ <https://www.bellingcat.com/resources/how-tos/2014/07/09/a-beginners-guide-to-geolocation/>



- ▶ Caso de Estudio - Bellingcat
  - ▶ <https://www.bellingcat.com/>
  - ▶ Open source investigations tools and techniques
  - ▶ Confirmar la localización de un vídeo sin salir de casa
  - ▶ <https://www.bellingcat.com/resources/how-tos/2014/07/09/a-beginners-guide-to-geolocation/>
  - ▶ El 15 de agosto de 2011, la oposición Libia afirmó haber capturado la pequeña localidad de Tiji, publicando el siguiente video en youtube como prueba.



# Técnicas de geolocalización





# Técnicas de geolocalización





# Técnicas de geolocalización

- ▶ ¿Que necesitamos?
  - ▶ Mapa
    - ▶ Google Maps
    - ▶ Bing Maps
    - ▶ Yahoo! Maps
    - ▶ Wikimapia
      - [wikimapia.org](http://wikimapia.org)
      - Buscar: Tiji (cuidado Nombres Árabes)



# Técnicas de geolocalización





# Técnicas de geolocalización





# Técnicas de geolocalización



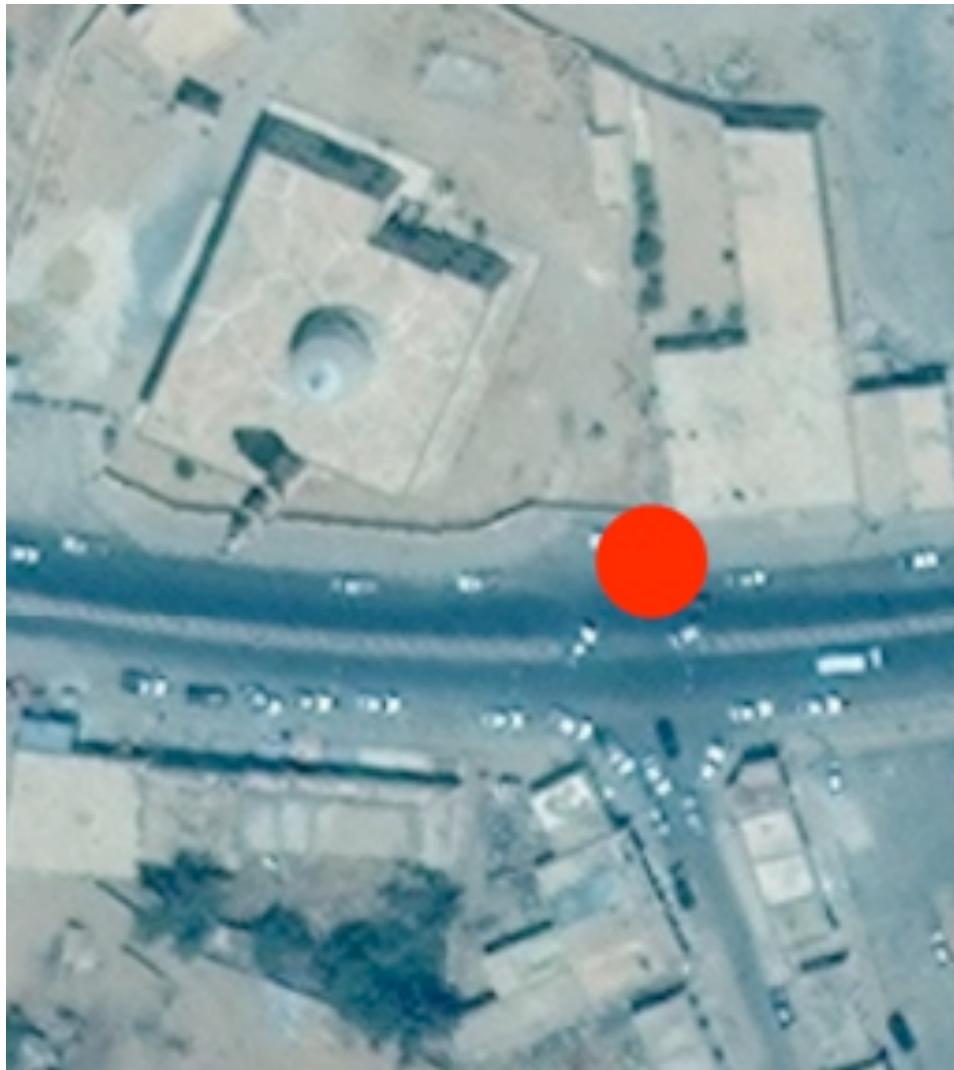


# Técnicas de geolocalización





# Técnicas de geolocalización





# Técnicas de geolocalización





# Técnicas de geolocalización

- ▶ Caso de Estudio - Bellingcat



- ▶ Caso de Estudio - Bellingcat
  - ▶ Verification and Geolocation Tricks and Tips with Google Earth



- ▶ Caso de Estudio - Bellingcat
  - ▶ Verification and Geolocation Tricks and Tips with Google Earth
  - ▶ Google Earth



- ▶ Caso de Estudio - Bellingcat
  - ▶ Verification and Geolocation Tricks and Tips with Google Earth
  - ▶ Google Earth
    - ▶ es un programa informático que muestra un globo virtual que permite visualizar múltiple cartografía, con base en la fotografía satelital.



- ▶ Caso de Estudio - Bellingcat
  - ▶ Verification and Geolocation Tricks and Tips with Google Earth
  - ▶ Google Earth
    - ▶ es un programa informático que muestra un globo virtual que permite visualizar múltiple cartografía, con base en la fotografía satelital.
    - ▶ Una herramienta muy poderosa para verificar y localizar imágenes por todo el mundo.



- ▶ Caso de Estudio - Bellingcat
  - ▶ Verification and Geolocation Tricks and Tips with Google Earth
  - ▶ Google Earth
    - ▶ es un programa informático que muestra un globo virtual que permite visualizar múltiple cartografía, con base en la fotografía satelital.
    - ▶ Una herramienta muy poderosa para verificar y localizar imágenes por todo el mundo.
    - ▶ Especialmente útil para la geolocalización en zonas de conflicto.



# Técnicas de geolocalización

- ▶ Caso de Estudio - Bellingcat
  - ▶ Verification and Geolocation Tricks and Tips with Google Earth
  - ▶ Imágenes históricas
  - ▶ Ejemplo:
    - ▶ La **Batalla de Damasco** corresponde a un enfrentamiento acontecido entre el **15 de julio** y el **4 de agosto de 2012** entre las **Fuerzas Armadas de Siria** y distintos grupos sublevados en la ciudad capital de **Damasco**. Todo en el marco de la **guerra civil que azota al país**.
    - ▶  $33^{\circ}32'30.09''\text{N}$        $36^{\circ}20'43.69''\text{E}$



# Técnicas de geolocalización



23/05/2012

# Técnicas de geolocalización



19/08/2012



- ▶ Caso de Estudio - Bellingcat
  - ▶ Verification and Geolocation Tricks and Tips with Google Earth
  - ▶ Otros Ejemplos:
    - ▶ 47.848926, 39.750993 - Paso Fronterizo Ukraine - Russia
    - ▶ 48° 3'51.59"N 39°49'52.48"E - Paso Fronterizo Ukraine - Russia
    - ▶ 47°19'4.36"N 39°42'6.36"E - Cementerio Rostov
    - ▶ 39°21'25.53"N 141°54'47.92"E - Otsuchi, Japan

# Técnicas de geolocalización



- ▶ Caso de Estudio - Bellingcat
  - ▶ Cambiando el ángulo



# Técnicas de geolocalización



- ▶ Caso de Estudio - Bellingcat
  - ▶ Desarrollo de proyectos





# Técnicas de geolocalización

- ▶ Caso de Estudio - Bellingcat
  - ▶ Terrenos 3D
  - ▶ Muy útil para verificar vídeos y fotos
  - ▶





# Técnicas de geolocalización

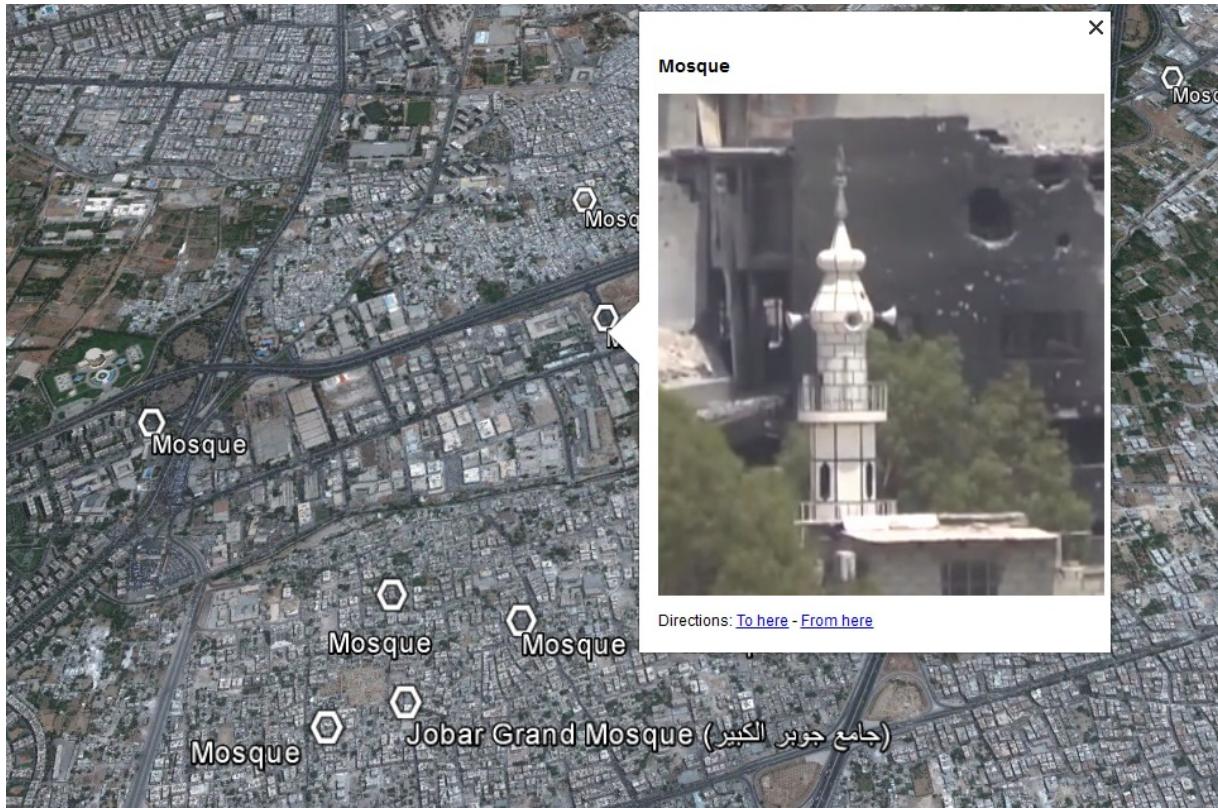
- ▶ Caso de Estudio - Bellingcat
  - ▶ Edificios 3D





# Técnicas de geolocalización

- ▶ Caso de Estudio - Bellingcat
  - ▶ Marcas y búsquedas





# Trabajar con datos en excel

- ▶ Primeros pasos
  - ▶ Crear un nuevo libro
  - ▶ Una o varias Hojas de Cálculo
  - ▶ Celdas y numeración
  - ▶ Valores y Funciones
  - ▶ Formularios



# Trabajar con datos en excel

- ▶ Funciones y fórmulas más comunes
  - ▶ Suma/Resta
    - ▶ SUM()
  - ▶ Multiplicar/Dividir
    - ▶ MULTIPLY()/DIVIDE()
    - ▶ “\*” y “/”
  - ▶ Potencia
    - ▶ POWER()
    - ▶ SQRT()
  - ▶ Ejercicio
    - ▶ Comprobar totales de documentos previos



# Trabajar con datos en excel

- ▶ Operaciones más comunes:
  - ▶ Símbolo de sumatorio
    - ▶ SUM()
    - ▶ AVERAGE()
      - Ofrece el valor promedio numérico de un conjunto de datos, sin tener en cuenta el texto
  - ▶ COUNT()
    - Ofrece el recuento de valores numéricos de un conjunto de datos.
  - ▶ MAX()
    - Ofrece el valor máximo de un conjunto de datos numérico.
  - ▶ MIN()
    - Ofrece el valor mínimo de un conjunto de datos numérico.
  - ▶ Otras...
    - <https://support.google.com/docs/table/25273>



# Trabajar con datos en excel

- ▶ Funciones rápidas:
  - ▶ Igual =
  - ▶ Suma +
  - ▶ Resta -
  - ▶ Producto \*
  - ▶ División /
  - ▶ Porcentaje %
  - ▶ Exponencial ^
  - ▶ Mayor >
  - ▶ Mayor o igual >=
  - ▶ Menor <
  - ▶ Menor o igual a <=
  - ▶ Distinto <>
  - ▶ Referencias a celdas \$



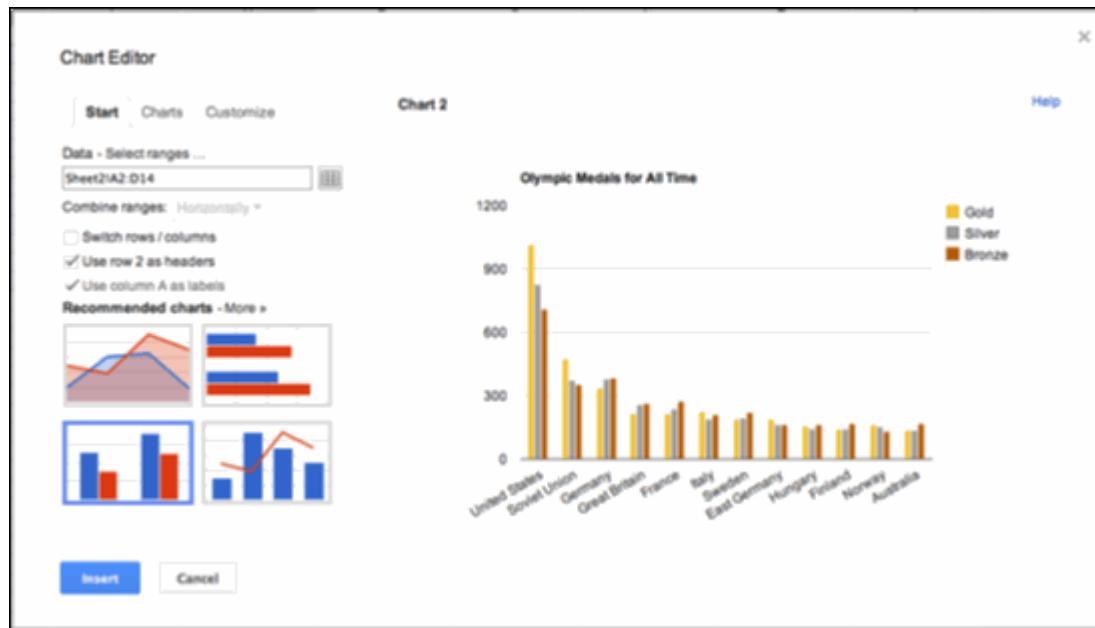
# Trabajar con datos en excel

- ▶ Funciones Lógicas
  - ▶ IF
    - ▶ IF(expresión\_lógica; valor\_si\_true; valor\_si\_false)
  - ▶ AND
    - ▶ AND(expresión\_lógica1; [expresión\_lógica2; ...])
  - ▶ OR
    - ▶ OR(expresión\_lógica1; [expresión\_lógica2; ...])
- ▶ SUMIF
  - ▶ SUMIF(intervalo; criterio; [intervalo\_suma])



# Trabajar con datos en excel

- ▶ Gráficos
  - ▶ Datos - Seleccionar intervalos ...
  - ▶ Manera rápida y sencilla de visualizar datos





- ▶ Necesidad de limpiar los datos
  - ▶ Problemas que podemos encontrar en los datos
    - ▶ Pobre diseño del esquema de datos
    - ▶ Entradas erróneas
      - Faltas de ortografía
      - Redundancias y duplicados
      - Valores contradictorios
    - ▶ Singularidad
    - ▶ Nombres incorrectos
    - ▶ División de datos



- ▶ Herramientas para limpiar los datos
  - ▶ Excel
    - ▶ Para pequeños fallos
    - ▶ De facil solución
  - ▶ Openrefine
    - ▶ <http://openrefine.org/>
    - ▶ Herramienta potente para limpiar, modificar y formatear los datos.



# Bibliografía

- ▶ <http://datajournalismhandbook.org/>
- ▶ <http://schoolofdata.org/>
- ▶ <http://onlinejournalismblog.com/>
- ▶ <http://www.theatlantic.com/international/archive/2013/06/these-charts-show-how-crucial-twitter-is-for-the-turkey-protesters/276798/>
- ▶ <https://www.bellingcat.com>
- ▶ <http://blogs.ianacion.com.ar/data/datos-abiertos/como-usar-google-refine-para-trabajar-una-base-de-datos/>
- ▶ <http://openrefine.org/>
- ▶ <http://tabula.technology/>
- ▶ [wikipedia.org](http://wikipedia.org)
- ▶ <http://tejiendo-redes.com/>
- ▶



All rights of images are reserved by **the original owners**, the rest of the content is licensed under a **Creative Commons by-sa 3.0** license.



# **Técnicas y herramientas de extracción de datos**

## **Introducción a la extracción de datos**

Juan Sixto

*{jsixto@deusto.es}*

DeustoTech - Deusto Institute of Technology, University of Deusto

<http://www.morelab.deusto.es>