

Data Tools Ecosystem for non-programmers



Óscar Marín Miró
@outliers_es
www.outliers.es



Contents at a glance

The Data Process

- ▶ Why mess with Data?
- ▶ A Data Pipeline

Data Acquisition Tools

- ▶ Data Sources
- ▶ Google Forms & Google Docs
- ▶ Scraping tools
- ▶ Network acquisition tools
- ▶ PDF tools
- ▶ Open Refine

Data Analysis

- ▶ Excel capabilities
- ▶ Beyond excel

Data Visualization Tools

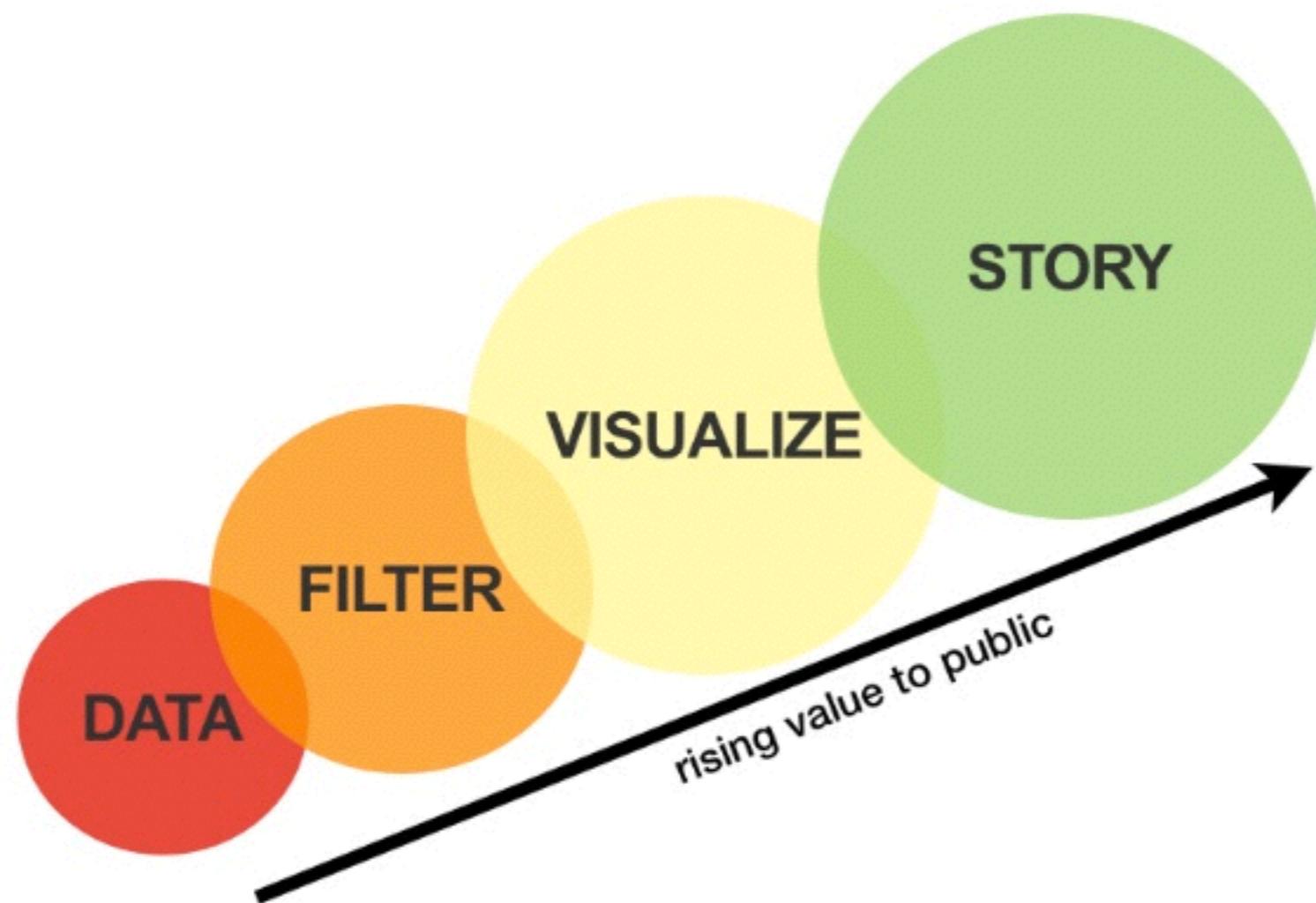
- ▶ Timeline.js
- ▶ Wordle
- ▶ Plot.ly
- ▶ Google Fusion Tables
- ▶ Gephi
- ▶ CartoDB
- ▶ MapBox
- ▶ Tilemill

Developers needed

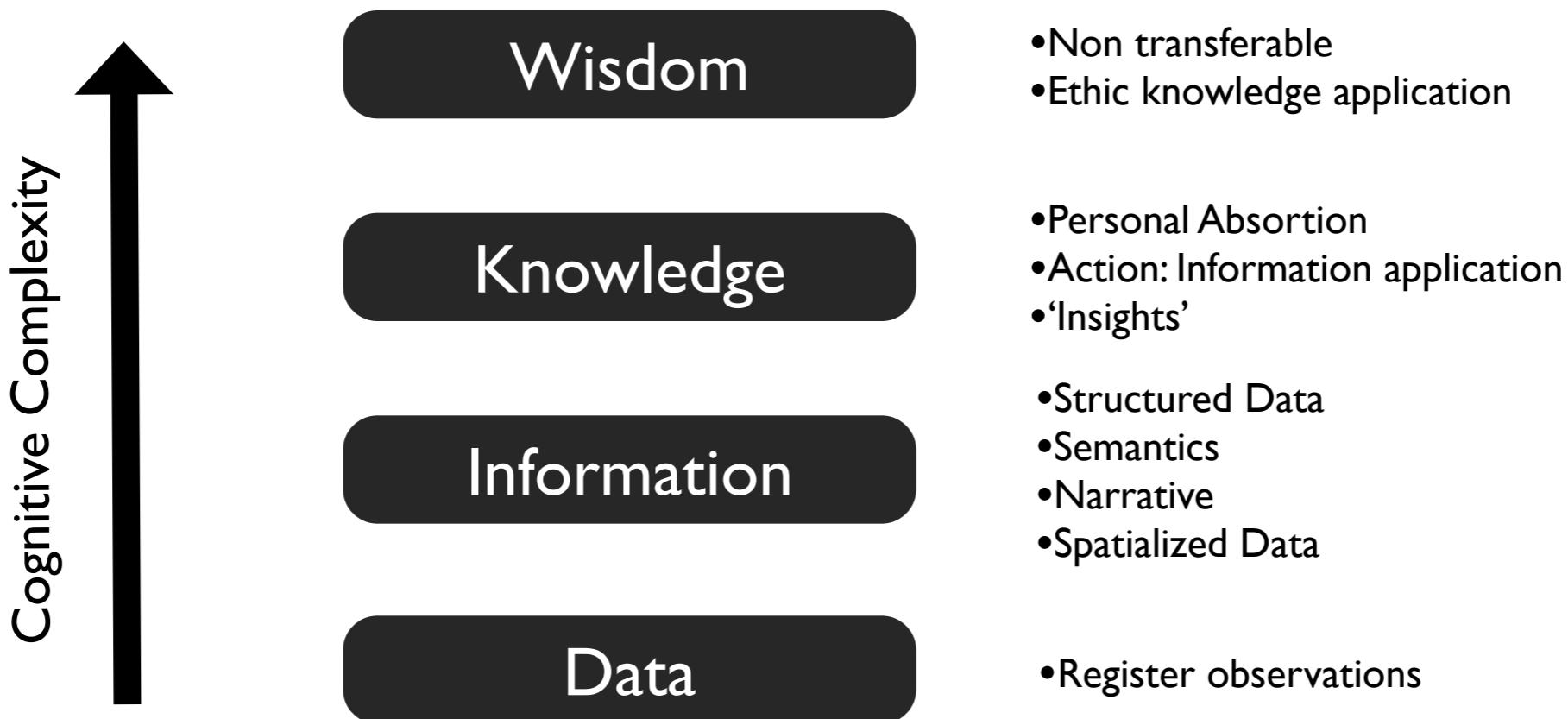
- ▶ Size
- ▶ Real-Time
- ▶ Analysis beyond Basics
- ▶ Interactivity*

Resources

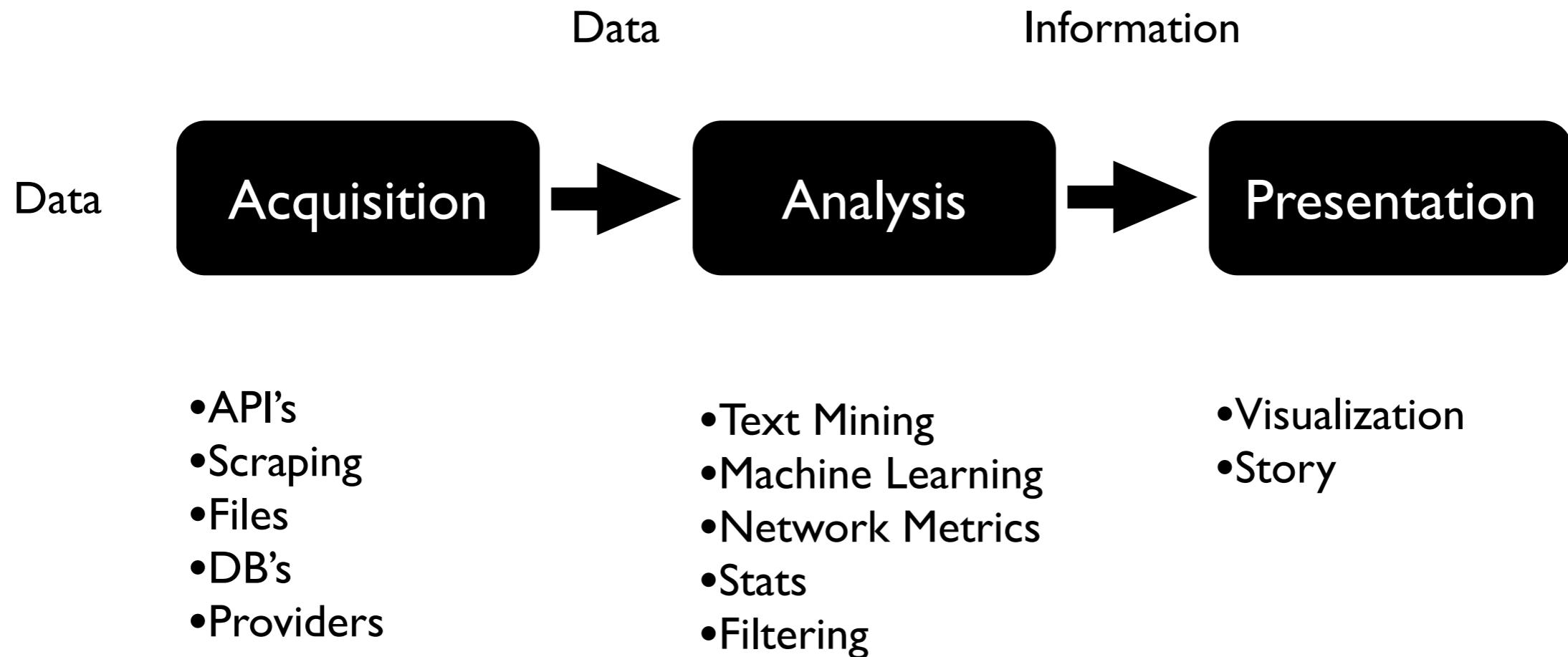
The Data Process



Why mess with data?



A Data Pipeline



Anscombe's quartet

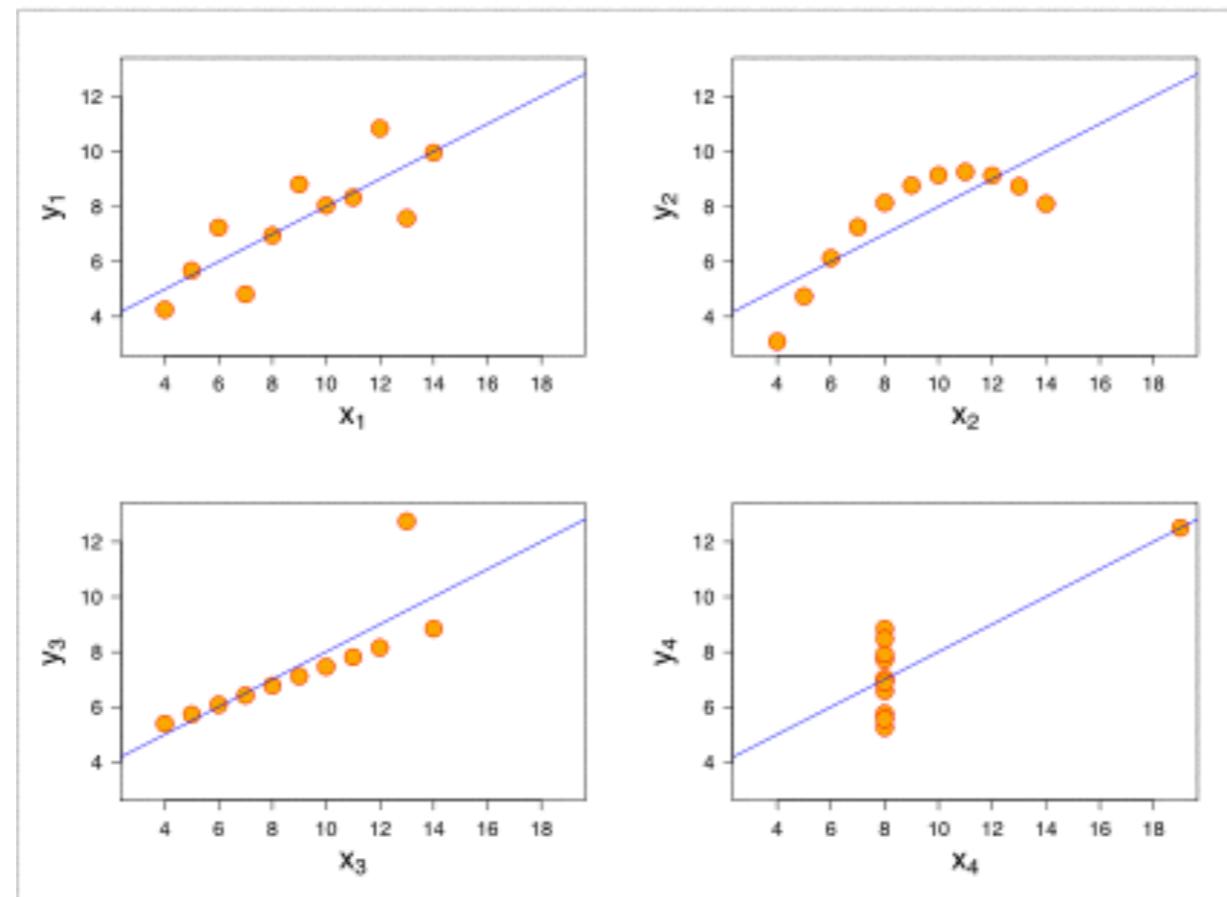
Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Property	Value
Mean of x in each case	9 (exact)
Variance of x in each case	11 (exact)
Mean of y in each case	7.50 (to 2 decimal places)
Variance of y in each case	4.122 or 4.127 (to 3 decimal places)
Correlation between x and y in each case	0.816 (to 3 decimal places)
Linear regression line in each case	$y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)

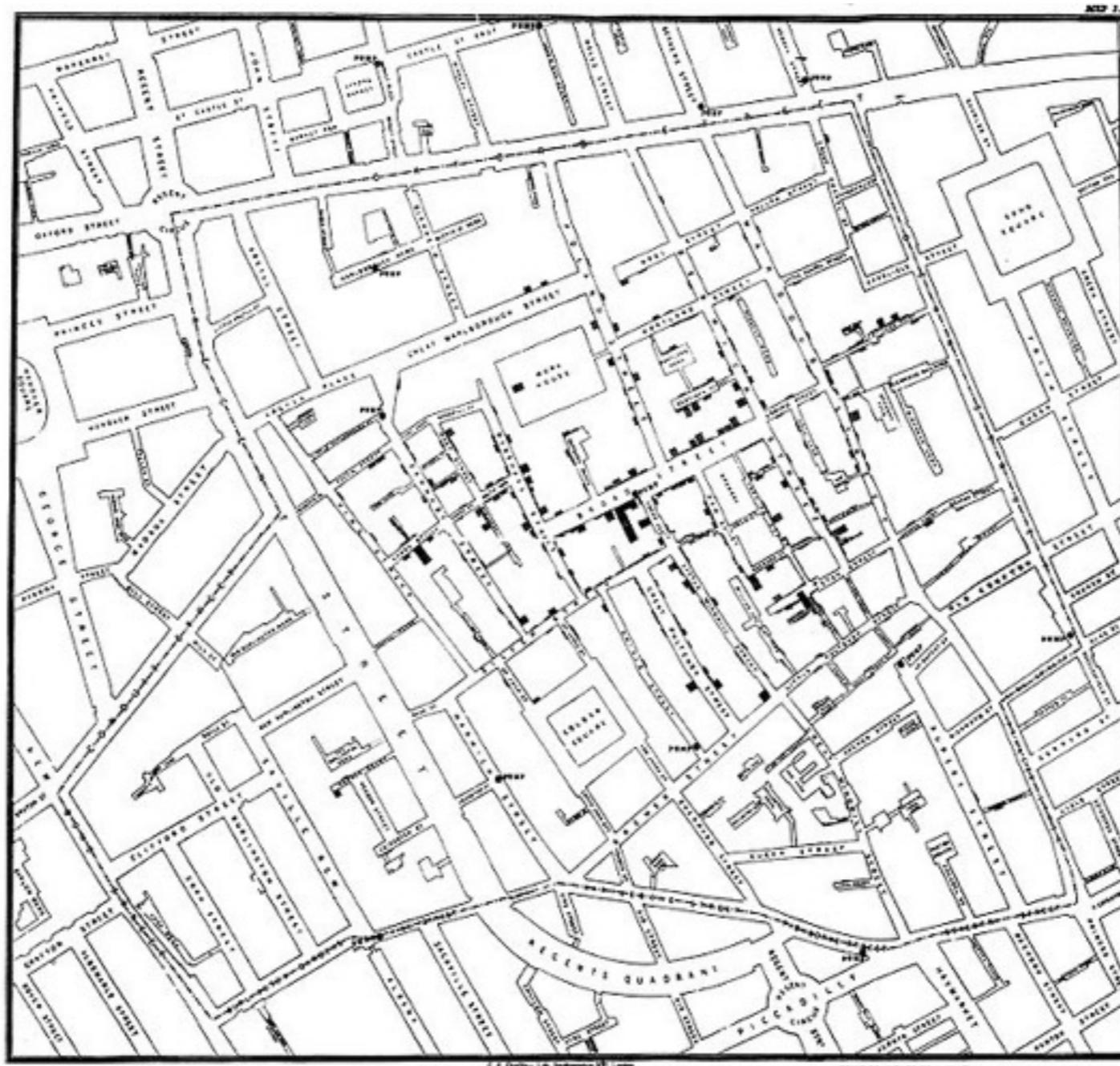
http://en.wikipedia.org/wiki/Anscombe's_quartet

Anscombe's quartet



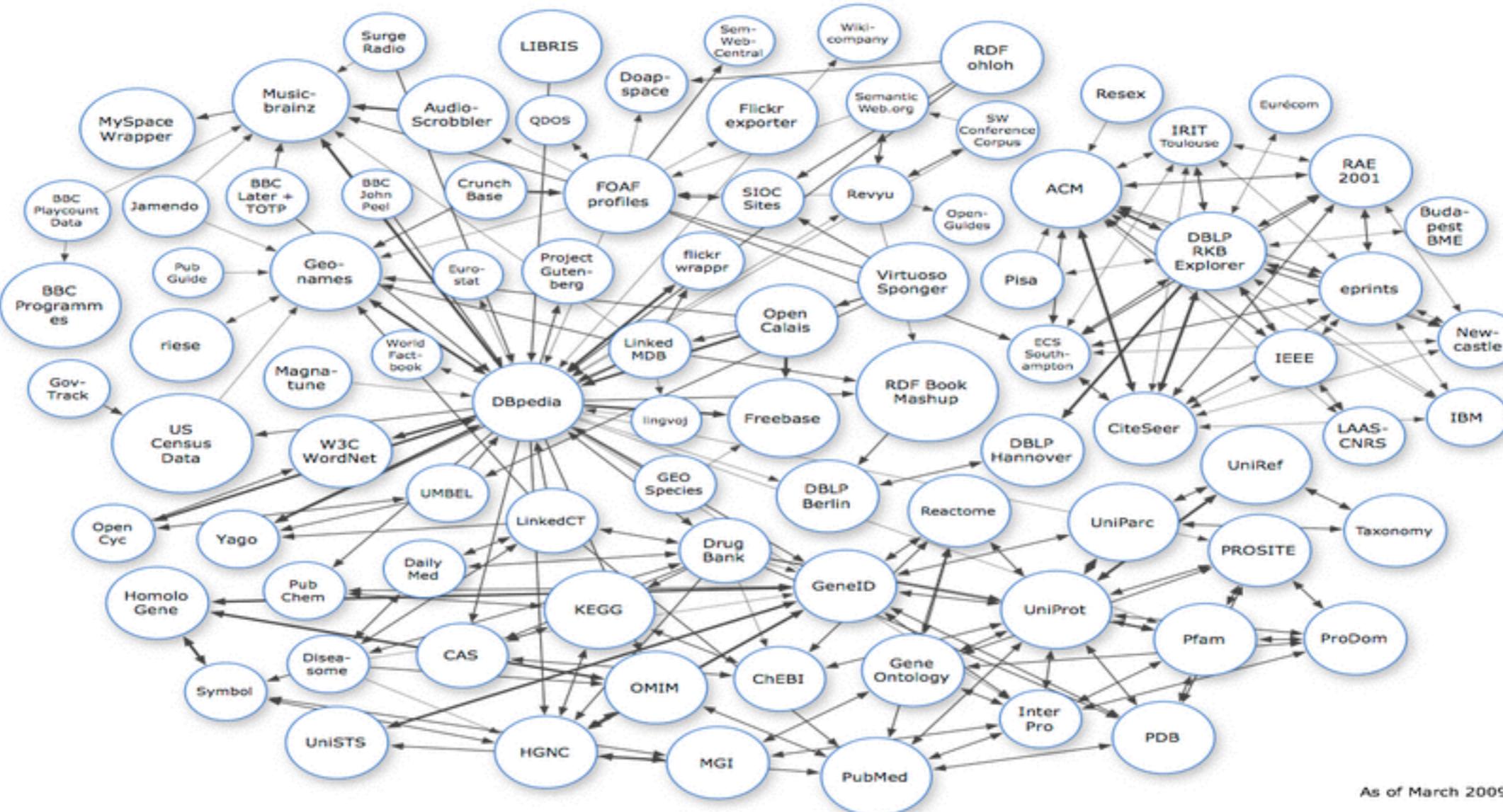
http://en.wikipedia.org/wiki/Anscombe's_quartet

John Snow



http://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak

Data Acquisition Tools



Data Formats

- JSON (Javascript Object Notation) - Dev
- CSV, XLS - Tabular
- XML - Old, but still used
- Web data

Data Acq Techniques

- Ask the user: Forms
- Download from Data Repositories
- Web data: Scrape
- Ask the provider: API access

Google Forms

The screenshot shows the Google Forms interface. On the left, there's a sidebar with a 'CREATE' button and icons for Folder, Document, Presentation, Spreadsheet, Form, and Drawing. Below these are 'My Drive' and 'Meet' sections. A 'Connect more apps' button is at the bottom of the sidebar.

The main area displays a survey titled 'Primera batería de preguntas: El 15-M y tú'. It includes the following questions:

- ¿Has participado en alguna de las acampadas del 15-M?
[Radio button options: Sí, No]
- En caso afirmativo, ¿de qué manera?
[Text input field: Acampado]
- Cuando has visitado una acampada tú has...
[List of radio buttons:
 - Acudido a una asamblea
 - Participado en una manifestación
 - Organizado un evento o acción
 - Participado en un grupo de trabajo (comisión) de la asamblea
 - Dormido en la acampada
 - Participado proveyendo comida o servicios a la gente de la acampada
 - Sido golpeado físicamente por la policía/ arrestado
 - Hecho algún otro tipo de actividad fuera de esta lista
 - NS/NC
- ¿Has participado en alguna de las siguientes actividades relacionadas con el 15-M?
[List of radio buttons:
 - Donación de dinero, comida o material

At the bottom right of the form area, there are three small icons: a pencil, a square, and a circle.

Google Forms

15MPoll star

File Edit View Insert Format Data Tools Form Help Last edit was 14 days ago

Comments Share

Timestamp	¿Has participado en alguna de las acampadas del 15-M?	En caso afirmativo, ¿de qué manera?	Cuando has visitado una acampada tú has...	¿Has participado en alguna de las siguientes actividades relacionadas con el 15-M?	¿Es el 15M el primer espacio donde participas?	En caso de haber contestado 'NO' a la pregunta anterior; ¿En qué tipo de espacio participabas antes del 15M?	Debajo hay una serie de diferentes formas de participación política y social. Indica cuál de estas has realizado antes del 15M:	En el último año, ¿a través de qué canales has obtenido información del 15M?	¿A través de qué pantalla has seguido las novedades del #15M?	En el último año, ¿qué canales de comunicación de las acampadas has consultado?
1	5/14/2013 11:54:25	No	Acampado	Participar en una manifestación, Participar en una marea ciudadana, Seguir la información de 15-M, Redifundir entre mis contactos (redes sociales, correo...) informaciones del/sobre el 15M	No me considero participante del 15M	Participar en una manifestación, Participar en un huelga	Firmar una petición en Change, Comprar o contratar un producto o servicio por razones políticas, éticas, ambientales.., Participar en una manifestación, Participar en un huelga	Facebook, Twitter, Prensa online, Televisión	Pantalla de televisión, Pantalla del ordenador de sobremesa	Streaming
283							Firmar una petición en Change, Realizar un boicot a un producto o servicio por razones políticas, éticas..			

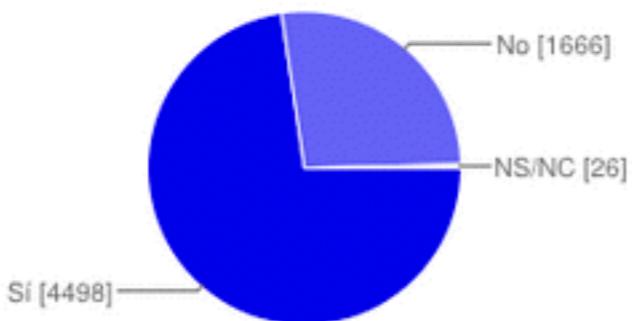
Google Forms

6194 [responses](#)

Summary [See complete responses](#)

Primera batería de preguntas: El 15-M y tú

¿Has participado en alguna de las acampadas del 15-M?



Sí	4498	73%
No	1666	27%
NS/NC	26	0%

En caso afirmativo, ¿de qué manera?



Acampado	2024	33%
Visitante	3196	52%
Miembro de alguna de las comisiones	485	8%
Otros	486	8%

Google Forms

PP ¿Con cuál de estos partidos te identificas más?

Motivaciones y preocupaciones

situacion critica de españa
desconfianza a los politicos
esperanza de cambio
preferentes
necesidad de un cambio

Colectivos asociados

comunistas
iu **pah**
perroflautas **dry**
abuso de poder de los politicos
izquierda unida y eta
asambleas locales psoe progres
acampada sol
anarquistas

Google Forms

CUP ¿Con cuál de estos partidos te identificas más?

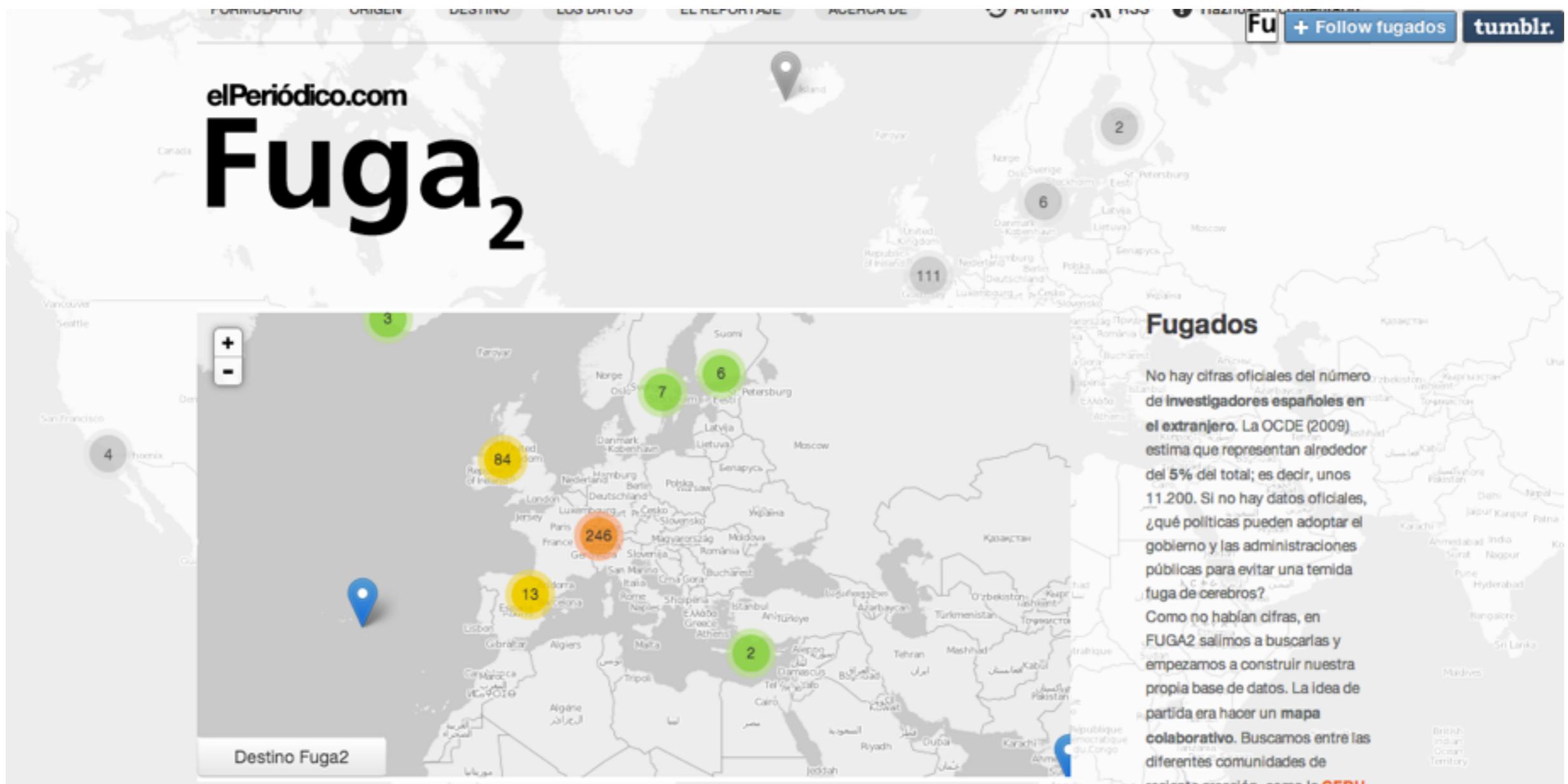
Motivaciones y preocupaciones

sanidad publica
educacion
solidaridad desigualdad social
crisis economica
futuro crisis politica crisis de valores sociales
corrupcion
los recortes anticapitalismo esperanza
democracia recortes
corrupcion politica **crisis** economia
participacion derechos
indignacion paro horizontalidad
capitalismo cambio
politica

Colectivos asociados

juventud sin futuro
escraches
15mparato marea roja
dry coordinadora 26s
iaioflautas pah
marea verde democracia real ya
iaioflautas Mareas ojo con tu ojo
educacion cup feministas yayoflautas
mareas ciudadanas marea blanca
pan acampadas asambleas de barrio
universitat indignada stop desahucios pacd

Google Forms



<http://data.elperiodico.com/>

Google Spreadsheet

A screenshot of a Google Sheets spreadsheet titled "Untitled spreadsheet". The formula bar shows the text "=import". A dropdown menu is open over cell A1, listing several functions related to importing data from URLs:

- ImportXML(url, query)
- ImportData(url)
- ImportFeed(url, query, headers, numItems)
- ImportHtml(url, query, index)** (This item is highlighted in the dropdown menu.)
- ImportRange(sspreadsheet_key, sheetrange)

The cell A1 contains the formula "=import". To the right of the dropdown menu, there is a tooltip for the "ImportHtml" function:

ImportHtml(URL, query, index)

The ImportHTML function imports the data in a particular table or list from an HTML page **Note:** The limit on the number of ImportHtml functions per spreadsheet is 50.

[View the complete list of functions.](#)

The spreadsheet has 31 rows labeled 1 through 31. The columns are labeled A through K. The top navigation bar includes File, Edit, View, Insert, Format, Data, Tools, Help, and a Share button.

<http://mashe.hawksey.info/2012/10/feeding-google-spreadsheets-exercises-in-import/>

Google Spreadsheet

en.wikipedia.org/wiki/Time_Person_of_the_Year

Apps | Ricardo Solé | https://drive.google.com | Django | Model field | http://www.colorpic.com | TP | datanalysis15m - Et | Amazon.com: So, Yo | popup | Other Bookmarks

Русский Simple English Slovenčina Српски / srpski Suomi Svenska ไทย Türkçe Українська Tiếng Việt 粵語 中文

Edit links

Filmmaker Michael Moore claims that director Mel Gibson cost him the opportunity to be Person of the Year alongside Gibson in 2004. Moore's controversial political documentary *Fahrenheit 9/11* became the highest-grossing documentary of all time the same year Gibson's *The Passion of the Christ* became a box-office success and also caused significant controversy. Moore said in an interview "I got a call right after the '04 election from an editor from Time Magazine. He said, 'Time Magazine has picked you and Mel Gibson to be Time's Person of the Year to put on the cover, Right and Left, Mel and Mike. The only thing you have to do is pose for a picture with each other. And do an interview together.' I said 'OK.' They call Mel up, he agrees. They set the date and time in LA. I'm to fly there. He's flying from Australia. Something happens when he gets home... Next thing, Mel calls up and says, 'I'm not doing it. I've thought it over and it is not the right thing to do.' So they put Bush on the cover."^[8] Another criticized^[citation needed] choice was the 2006 selection of "You", representing most if not all people for advancing the information age by using the Internet (via e.g. blogs, YouTube, MySpace and Wikipedia).^[9]

Persons of the Year [edit]

Year	Image	Choice	Lifetime	Notes
1927		Charles Lindbergh	USA	1902–1974 Lindbergh was, in May 1927, the first person to fly a plane non-stop from New York City, USA to Paris, France.
1928		Walter Chrysler	USA	1875–1940 In 1928, Chrysler oversaw a merger of his company with Dodge, and began work on his eponymous building.
1929		Owen D. Young	USA	1874–1962 Young chaired a committee which authored the Young Plan, a program for settlement of German reparations debts after World War I.
1930		Mahatma Gandhi	British Raj	1869–1948 Gandhi was the leader of the Indian independence movement. In 1930, he led the Salt Satyagraha, a 240-mile march to protest the imposition of taxes on salt by the British Raj. Gandhi was also the first non-American to win the honor.

Time_POTY ☆

File Edit View Insert Format Data Tools Help All changes saved in Drive

fx =importHTML("http://en.wikipedia.org/wiki/Time_Person_of_the_Year","table",1)

	A	B	C	D	E
1	=importHTML("http://en.wikipedia.org/wiki/Time_Person_of_the_Year","table",1)				
2					
3					
4					
5					
A					

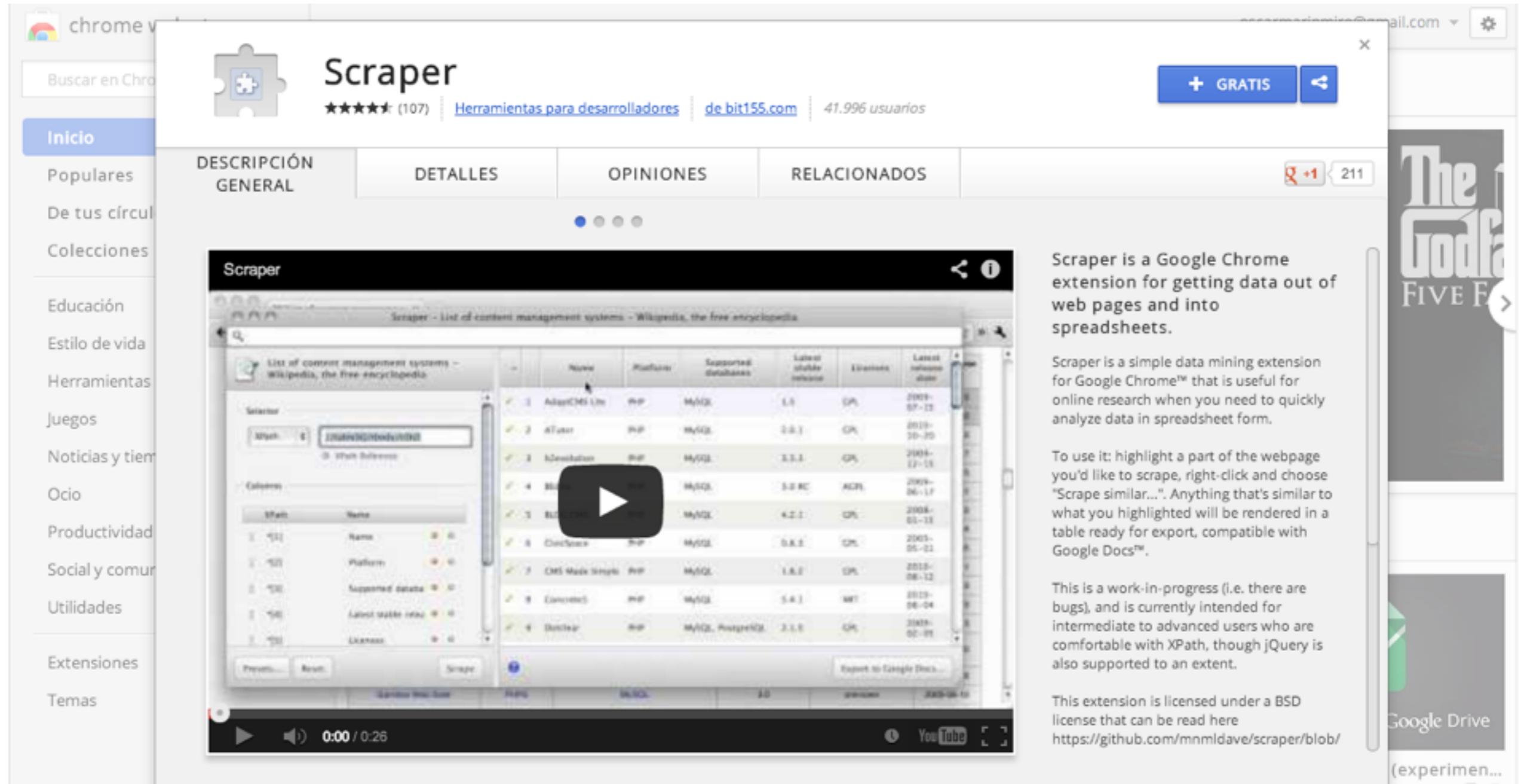
Time_POTY ☆

File Edit View Insert Format Data Tools Help All changes saved in Drive

Open Google Drive fx =CONTINUE(A1, 2, 1)

	A	B	C	D	E	F
1	Year	Image	Choice	Lifetime	Notes	
2	1927		Charles Lindbergh	USA	1902–1974	Lindbergh was, in May 1927, the first person to fly a plane non-stop from New York City, USA to Paris, France.
3	1928		Walter Chrysler	USA	1875–1940	In 1928, Chrysler oversaw a merger of his company with Dodge, and began work on his eponymous building.
4	1929		Owen D. Young	USA	1874–1962	Young chaired a committee which authored the Young Plan, a program for settlement of German reparations debts after World War I.
						Gandhi was the leader of the Indian independence movement. In 1930, he led the Salt

Scrape w/ Chrome Scraper



The screenshot shows the Scrape extension page on the Google Chrome Web Store. The extension icon is a puzzle piece, and the title is "Scrape". It has a rating of 4.5 stars from 107 reviews. The developer is listed as "Herramientas para desarrolladores" and "de bit155.com". It has 41,996 users. A "GRATIS" button is visible. The main content area displays a screenshot of the extension's interface, which is a spreadsheet-like tool for extracting data from web pages. A video player is overlaid on the screenshot, showing a play button. To the right of the screenshot, there is descriptive text about the extension's purpose and usage.

Scrape is a Google Chrome extension for getting data out of web pages and into spreadsheets.

Scrape is a simple data mining extension for Google Chrome™ that is useful for online research when you need to quickly analyze data in spreadsheet form.

To use it: highlight a part of the webpage you'd like to scrape, right-click and choose "Scrape similar...". Anything that's similar to what you highlighted will be rendered in a table ready for export, compatible with Google Docs™.

This is a work-in-progress (i.e. there are bugs), and is currently intended for intermediate to advanced users who are comfortable with XPath, though jQuery is also supported to an extent.

This extension is licensed under a BSD license that can be read here <https://github.com/mnmlDave/scrapers/blob/>

<https://chrome.google.com/webstore/detail/scrapers/mbigbapnjcgaffohmbkdlcacepngjd>

Scrape w/ Chrome Scraper

The screenshot shows a web browser window with the URL www.imdb.com/chart/top?ref_=nv_ch_250_4. The main content is the "IMDb Charts" section titled "Top 250". It displays the top 250 movies voted by regular IMDb users, showing their rank, title, year, IMDb rating, user rating, and an "Add to Watchlist" button. The table includes rows for "The Shawshank Redemption" (1994), "The Godfather" (1972), "The Godfather: Part II" (1974), "Pulp Fiction" (1994), "The Good, the Bad and the Ugly" (1966), "The Dark Knight" (2008), "12 Angry Men" (1957), and "Schindler's List" (1993). To the right of the chart, there is an advertisement for the IMDb app and a sidebar with links to "IMDb Charts", "US Box Office", and "Top Movies by Genre".

Rank & Title	IMDb Rating	Your Rating	Action
1. The Shawshank Redemption (1994)	9.2	RATE	Add to Watchlist
2. The Godfather (1972)	9.2	RATE	Add to Watchlist
3. The Godfather: Part II (1974)	9.0	RATE	Add to Watchlist
4. Pulp Fiction (1994)	8.9	RATE	Add to Watchlist
5. The Good, the Bad and the Ugly (1966)	8.9	RATE	Add to Watchlist
6. The Dark Knight (2008)	8.9	RATE	Add to Watchlist
7. 12 Angry Men (1957)	8.9	RATE	Add to Watchlist
8. Schindler's List (1993)	8.9	RATE	Add to Watchlist

IMDb Charts
Top 250
As voted by regular IMDb users

Showing 250 Titles

Sort by: **IMDb Rating**

READ NEWS, WATCH VIDEOS AND MORE WITH THE IMDB APP

Download Now ▶

ad feedback

IMDb Charts

US Box Office

IMDb Top 250

IMDb Bottom 100

Top Movies by Genre

- Action
- Adventure
- Animation
- Biography
- Comedy
- Crime
- Documentary
- Drama

Scrape w/ Chrome Scraper

The screenshot shows the IMDb Top 250 chart page. The first movie, "The Shawshank Redemption" (1994), is selected, highlighted with a blue background. A context menu is open over this entry, listing options: Copy, Search Google for '1. The Shawshank Redemption (1994) 9.2 RATE...', Print..., Buffer Selected Text, and Scrape similar... (which is highlighted with a blue bar). Other menu items include Inspect Element, Look Up in Dictionary, and Speech. To the right of the menu, there's an advertisement for the IMDb app with a smartphone icon and the text "READ NEWS, WATCH VIDEOS AND MORE WITH THE IMDb APP". Below the menu, there's a "Top Movies by Genre" sidebar with links to Action, Adventure, Animation, Biography, Comedy, Crime, Documentary, and Drama.

Rank & Title	IMDb Rating	Your Rating
1. The Shawshank Redemption (1994)	9.2	9.2
2. The Godfather (1972)	9.1	9.1
3. The Godfather: Part II (1974)	9.1	9.1
4. Pulp Fiction (1994)	8.9	8.9
5. The Good, the Bad and the Ugly (1966)	8.9	8.9
6. The Dark Knight (2008)	8.9	8.9
7. 12 Angry Men (1957)	8.9	8.9
8. Schindler's List (1993)	8.9	8.9

Scrape w/ Chrome Scraper

The screenshot shows the Google Chrome Scraper extension interface. The main window title is "Scraper - IMDb Top 250 - IMDb". On the left, there's a sidebar titled "IMDb Charts" with a "IMDb Top 250 - IMDb" icon. Below it, under "Selector", is an "XPath" dropdown set to "/div[3]/div[1]/div/table/tbody/tr" and an "XPath Reference" link. Under "Columns", there's a table mapping XPath expressions to column names: *[1] to "Name", *[2] to "Rank & Title", *[3] to "IMDb Rating", and *[4] to "Your Rating". At the bottom of the sidebar are "Presets...", "Reset", and "Scrape" buttons. The main content area displays the top 250 movies from IMDb. The first 8 rows of the table are:

	*	Rank & Title	IMDb Rating	Your Rating	*[5]
	[1]				
1		1. The Shawshank Redemption (1994)	9.2	RATE 1 2 3 4 5 6 7 8 9 10 9.3/10 X	Add to Watchlist
2		2. The Godfather (1972)	9.2	RATE 1 2 3 4 5 6 7 8 9 10 9.2/10 X	Add to Watchlist
3		3. The Godfather: Part II (1974)	9.0	RATE 1 2 3 4 5 6 7 8 9 10 9/10 X	Add to Watchlist
4		4. Pulp Fiction (1994)	8.9	RATE 1 2 3 4 5 6 7 8 9 10 9/10 X	Add to Watchlist
5		5. The Good, the Bad and the Ugly (1966)	8.9	RATE 1 2 3 4 5 6 7 8 9 10 9/10 X	Add to Watchlist
6		6. The Dark Knight (2008)	8.9	RATE 1 2 3 4 5 6 7 8 9 10 9/10 X	Add to Watchlist

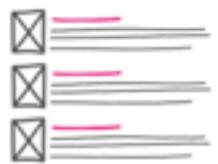
On the right side of the main window, there's a sidebar with movie thumbnails for "The Dark Knight", "12 Angry Men", and "Schindler's List", followed by their titles and ratings (8.9 each) and an "Add to Watchlist" button. Below this is a list of genres: Adventure, Animation, Biography, Comedy, Crime, Documentary, Drama.

<http://dataist.wordpress.com/2012/10/12/get-started-with-screenscraping-using-google-chromes-scraper-extension/>

Scrape w/ import.io

import  io

Our browser is quite new and we're still working out the kinks. [Click here](#) for a list of our known issues.



Create Data Set

+ New Data Source

Create a new source to bring data into import.io



Connector
Use a search box to retrieve one or more pages of results



Extractor
Take a single page and extract all of the data from it



Crawler
Get data from multiple similar pages on the same site



My Data

<http://import.io/>

PDF Extraction: Tabula



Tabula

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

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

Download for
Windows

Download for
Mac

View source on
GitHub

Current Version: [0.9.1 \(archive\)](#)



Using Tabula

1. Upload a file with tables you would like to copy.
2. Draw a box around the area of the table you would like to copy.
(Note: currently, Tabula can't select tables over multiple pages)
3. You will be given the option to copy the table as a CSV (comma-separated values) file or download the CSV or TSV (tab separated values). If you notice any errors in the table, you can make text edits to the selected text before copying or

<http://tabula.nerdpower.org/>

PDF Extraction: Tabula

Tabula is experimental software

[Home](#)

[About](#)

Example. Mr. and Mrs. Brown are filing a joint return. Their taxable income on Form 1040, line 43, is \$25,300. First, they find the \$25,300-\$25,300 taxable income line. Next, they find the column for married filing jointly and read down the column. The amount shown where the taxable income line and filing status column meet is \$2,921. This is the tax amount they should enter on Form 1040, line 44.

Line 43 (taxable income) is—		And you are—			
All less than	Single	Married filing jointly	Married filing separa- tely	Head of a house- hold	
0	0	0	0	0	
5	5	1	1	1	
10	10	2	2	2	
15	15	3	3	3	
20	20	4	4	4	
25	25	5	5	5	
30	30	6	6	6	
35	35	7	7	7	
40	40	8	8	8	
45	45	9	9	9	
50	50	10	10	10	
55	55	11	11	11	
60	60	12	12	12	
65	65	13	13	13	
70	70	14	14	14	
75	75	15	15	15	
80	80	16	16	16	
85	85	17	17	17	
90	90	18	18	18	
95	95	19	19	19	
100	100	20	20	20	
105	105	21	21	21	
110	110	22	22	22	
115	115	23	23	23	
120	120	24	24	24	
125	125	25	25	25	
130	130	26	26	26	
135	135	27	27	27	
140	140	28	28	28	
145	145	29	29	29	
150	150	30	30	30	
155	155	31	31	31	
160	160	32	32	32	
165	165	33	33	33	
170	170	34	34	34	
175	175	35	35	35	
180	180	36	36	36	
185	185	37	37	37	
190	190	38	38	38	
195	195	39	39	39	
200	200	40	40	40	
205	205	41	41	41	
210	210	42	42	42	
215	215	43	43	43	
220	220	44	44	44	
225	225	45	45	45	
230	230	46	46	46	
235	235	47	47	47	
240	240	48	48	48	
245	245	49	49	49	
250	250	50	50	50	
255	255	51	51	51	
260	260	52	52	52	
265	265	53	53	53	
270	270	54	54	54	
275	275	55	55	55	
280	280	56	56	56	
285	285	57	57	57	
290	290	58	58	58	
295	295	59	59	59	
300	300	60	60	60	
305	305	61	61	61	
310	310	62	62	62	
315	315	63	63	63	
320	320	64	64	64	
325	325	65	65	65	
330	330	66	66	66	
335	335	67	67	67	
340	340	68	68	68	
345	345	69	69	69	
350	350	70	70	70	
355	355	71	71	71	
360	360	72	72	72	
365	365	73	73	73	
370	370	74	74	74	
375	375	75	75	75	
380	380	76	76	76	
385	385	77	77	77	
390	390	78	78	78	
395	395	79	79	79	
400	400	80	80	80	
405	405	81	81	81	
410	410	82	82	82	
415	415	83	83	83	
420	420	84	84	84	
425	425	85	85	85	
430	430	86	86	86	
435	435	87	87	87	
440	440	88	88	88	
445	445	89	89	89	
450	450	90	90	90	
455	455	91	91	91	
460	460	92	92	92	
465	465	93	93	93	
470	470	94	94	94	
475	475	95	95	95	
480	480	96	96	96	
485	485	97	97	97	
490	490	98	98	98	
495	495	99	99	99	
500	500	100	100	100	

(Continued)

Page 1

Page 2

Page 3

Loading...

PDF Extraction: Tabula

Extracted tabular data

x

3,000					
3,000	3,0503,100	303308	303308	303308	303308
3,1003,150	3,1503,200	313318	313318	313318	313318
3,2003,250	3,2503,300	323328	323328	323328	323328
3,3003,350	3,3503,400	333338	333338	333338	333338
3,4003,450	3,4503,500	343348	343348	343348	343348
3,5003,550	3,5503,600	353358	353358	353358	353358
3,6003,650	3,6503,700	363368	363368	363368	363368
3,7003,750	3,7503,800	373378	373378	373378	373378
3,8003,850	3,8503,900	383388	383388	383388	383388
3,900	3,950	393	393	393	393

Use row/columns separators ?

[Close](#)

[Copy to clipboard as CSV](#)

[Download data ▾](#)

Network Data: Flocker

FLOCKER A Twitter real-time monitor

What?

FLOCKER is a Twitter real-time retweets networks builder.

Why?

Twitter is nowadays the fastest way to access and spread information. There are tools and services offering the possibility to monitor Twitter's stream. There are also tools offering the possibility to build networks based on retweets and mentions from a given dataset. But we haven't found any tool combining both functionalities (except Gephi's plugin Retweet Monitor).

Some of us worked in the mentioned plugin for Gephi and abandoned it. Gephi, although very useful and complete, is a complicate tool for both users and developers. Based on our experience we are trying to provide FLOCKER with the features most requested/used in Gephi by people analyzing Twitter.

Who?

FLOCKER is a project developed by Outliers.

Current status

Currently, FLOCKER is under development. At this moment you can:

- Login using your Twitter's account
- Filter the stream using terms, hashtags or Twitter's usernames
- See how the retweets network is dinamically built
- Explore the data using the *data laboratory*
- Change the colors used to display nodes and edges
- Export the generated graph as GEXF
- Export the generated graph as PNG
- Export the generated graph as SVG

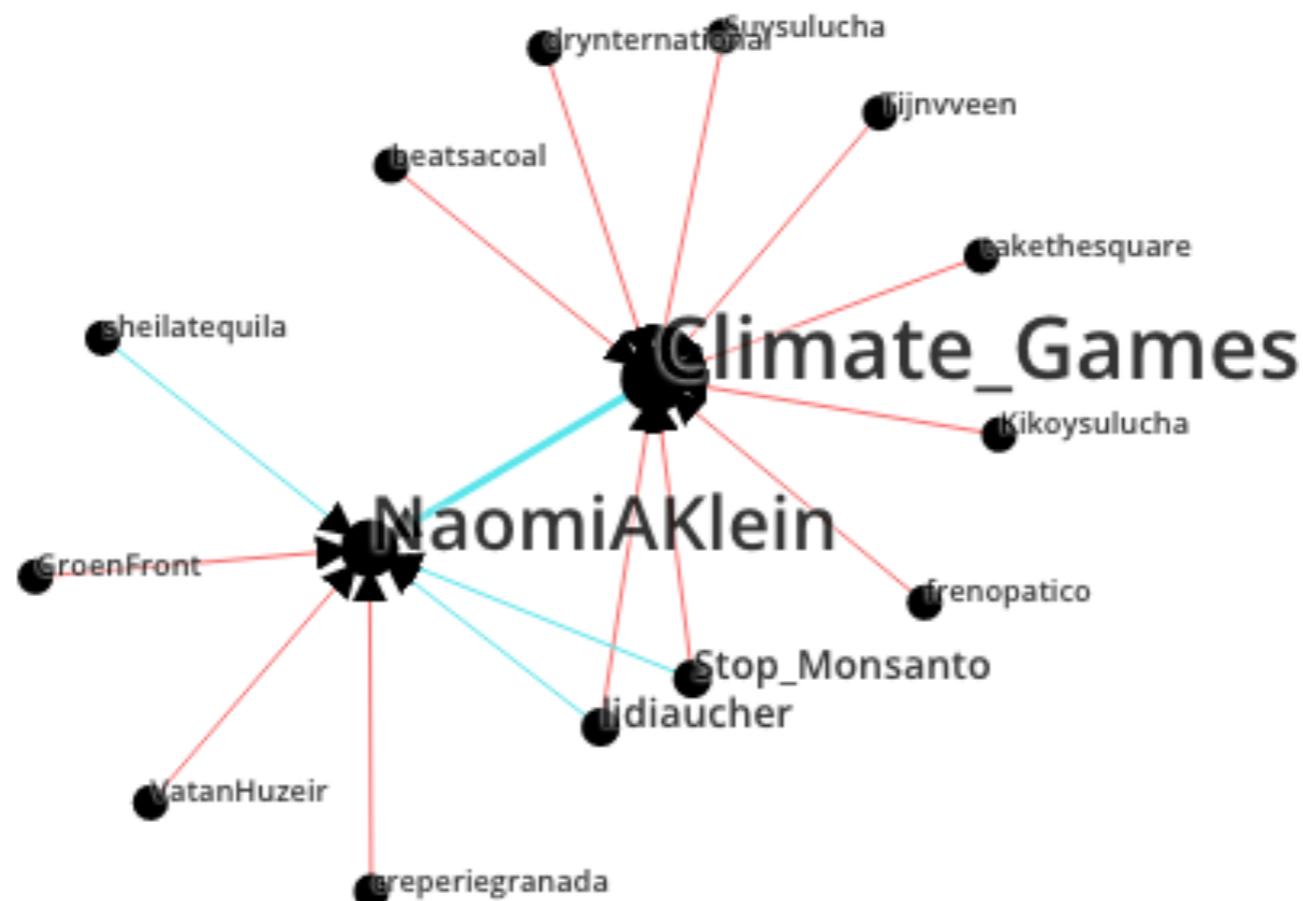
The chart on the right shows the percentage of features we have currently developed.



► Start using FLOCKER!

<http://flocker.outliers.es/>

Network Data: Flocker



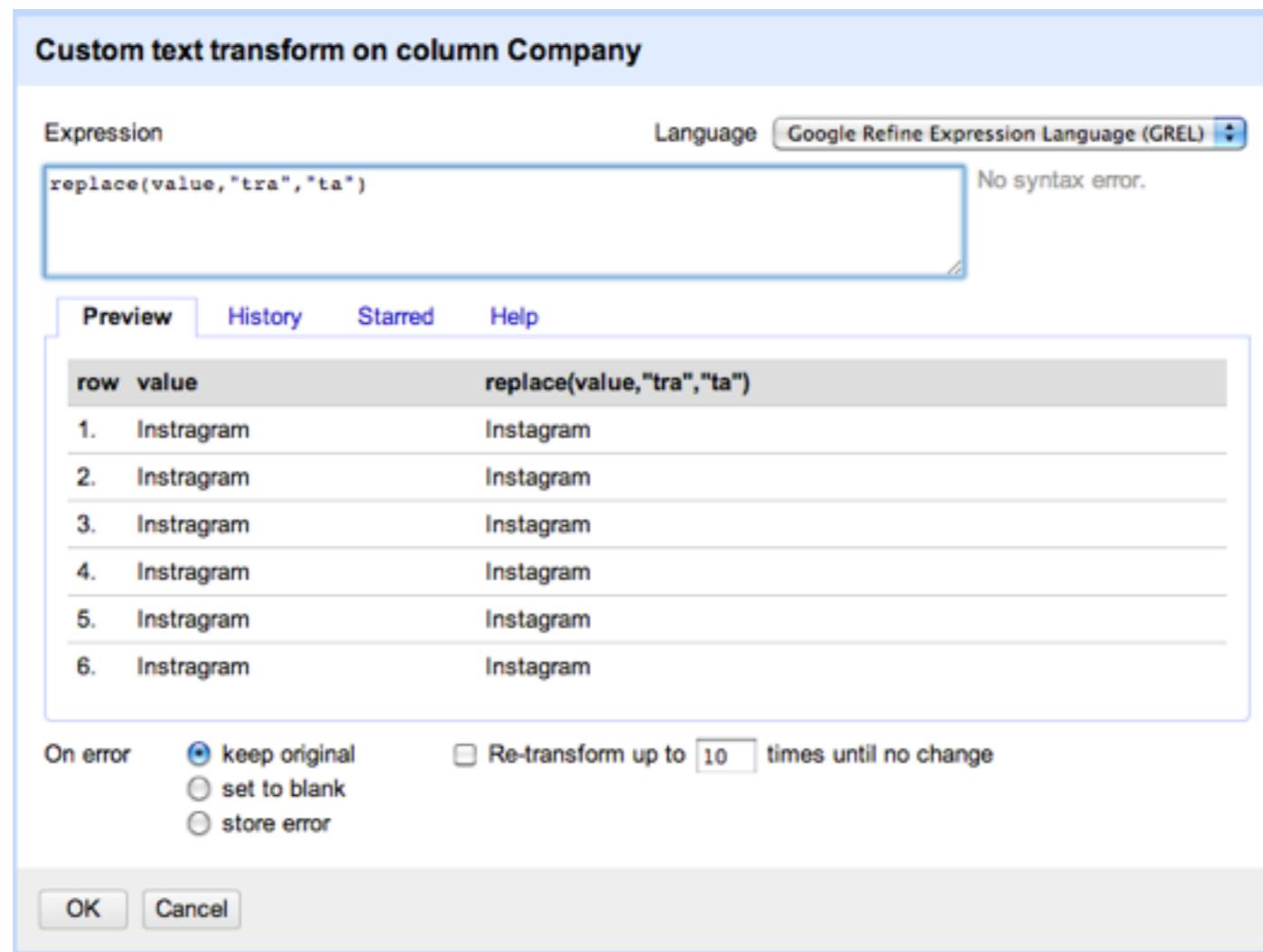
Open Refine: Cleaning Data



<https://github.com/OpenRefine>

<https://code.google.com/p/google-refine/downloads/list>

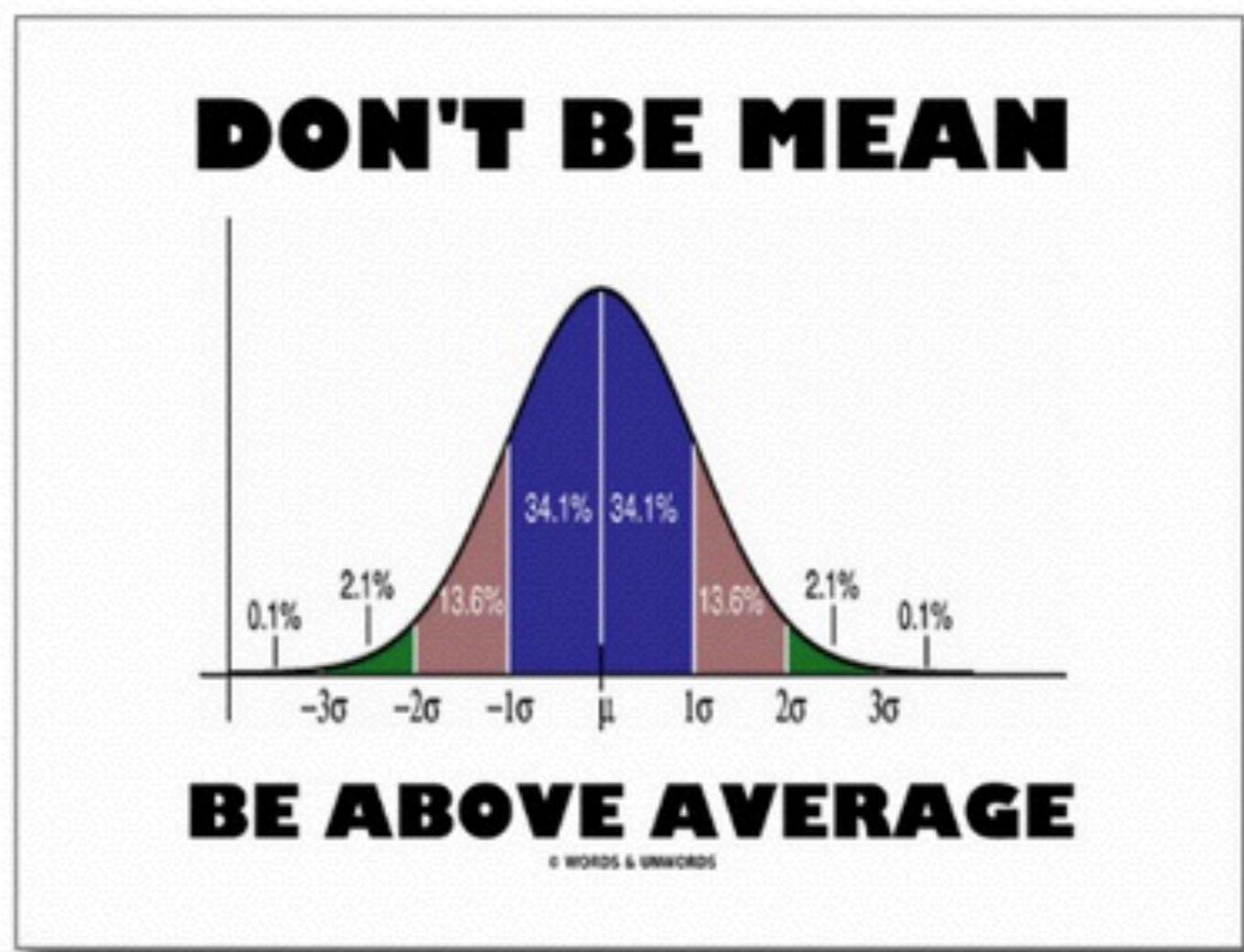
Open Refine: Cleaning Data



<https://github.com/OpenRefine/OpenRefine/wiki/GREL-Functions>

<https://github.com/OpenRefine/OpenRefine/wiki/GREL-String-Functions>

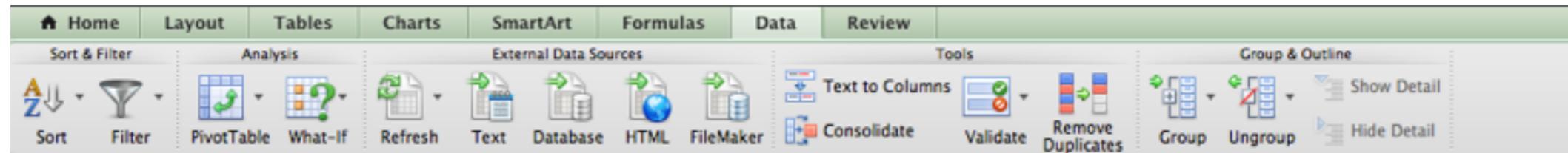
Data Analysis Tools



Data Analysis Tools

- Numeric analysis
- Network analysis
- Language analysis

Data Analysis w/ MS Excel

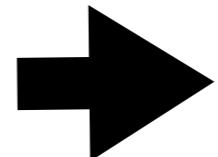


- Sort
- Basic Stats: Mean, Median, Sum...
- Filter
- Charts
- Pivot tables
- What-if
- Data Analysis Tool

Data Analysis: MS Excel

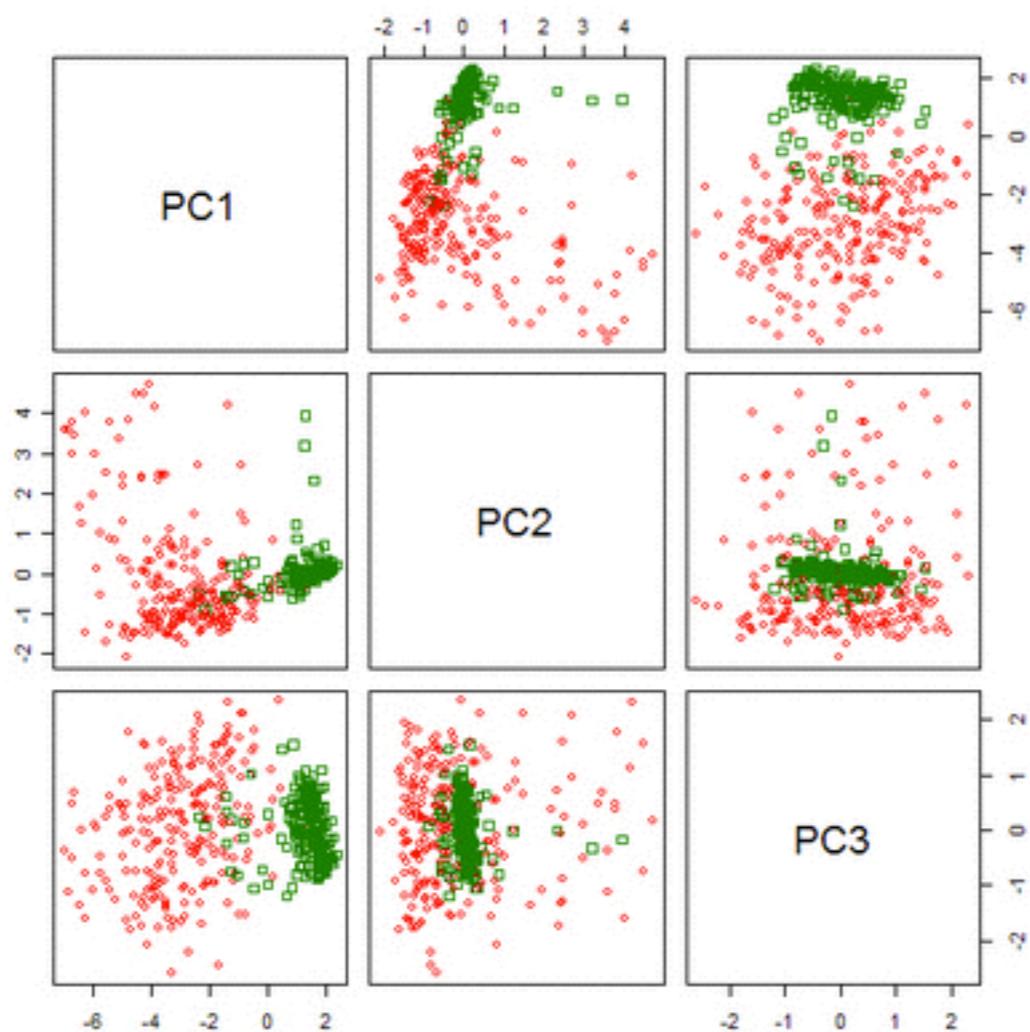
‘Pivot Tables’

A2	fx	1						
1	Order ID	Product	Category	Amount	Date	Country		
2	1	Carrots	Vegetables	\$4,270	1/6/2012	United States		
3	2	Broccoli	Vegetables	\$8,239	1/7/2012	United Kingdom		
4	3	Banana	Fruit	\$617	1/8/2012	United States		
5	4	Banana	Fruit	\$8,384	1/10/2012	Canada		
6	5	Beans	Vegetables	\$2,626	1/10/2012	Germany		
7	6	Orange	Fruit	\$3,610	1/11/2012	United States		
8	7	Broccoli	Vegetables	\$9,062	1/11/2012	Australia		
9	8	Banana	Fruit	\$6,906	1/16/2012	New Zealand		
10	9	Apple	Fruit	\$2,417	1/16/2012	France		
11	10	Apple	Fruit	\$7,421	1/16/2012	Canada		

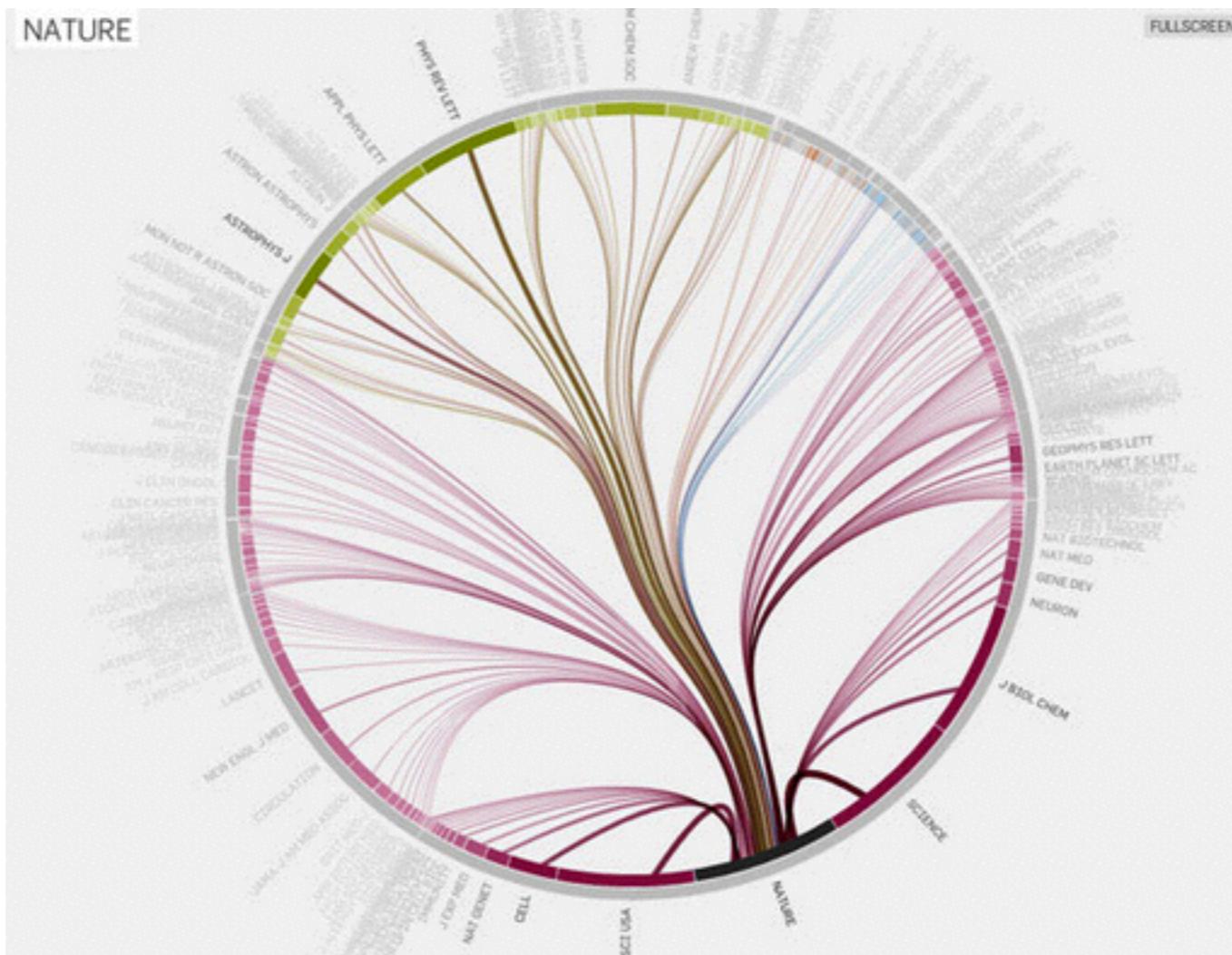


A3	fx	Sum of Amount		
1	A	B	C	D
2	Country	(All)		
3	Sum of Amount			
4	Product	Total		
5	Apple	191257		
6	Banana	340295		
7	Beans	57281		
8	Broccoli	142439		
9	Carrots	136945		
10	Mango	57079		
11	Orange	104438		
12	Grand Total	1029734		

Data Analysis: R



Data Visualization Tools



<http://well-formed.eigenfactor.org/>

Data Visualization Tools

- Timelines
- Words
- Chart tools
- Networks
- Maps

Data Visualization Tools

Timeline JS

The screenshot shows the Timeline JS interface. At the top, there's a navigation bar with tabs: Timeline JS (selected), Overview, Description, Examples, and Help. To the right of the navigation is the Knight Lab logo, which includes the text "NORTHWESTERN UNIVERSITY" and "knight lab".

The main area displays a horizontal timeline with three media items:

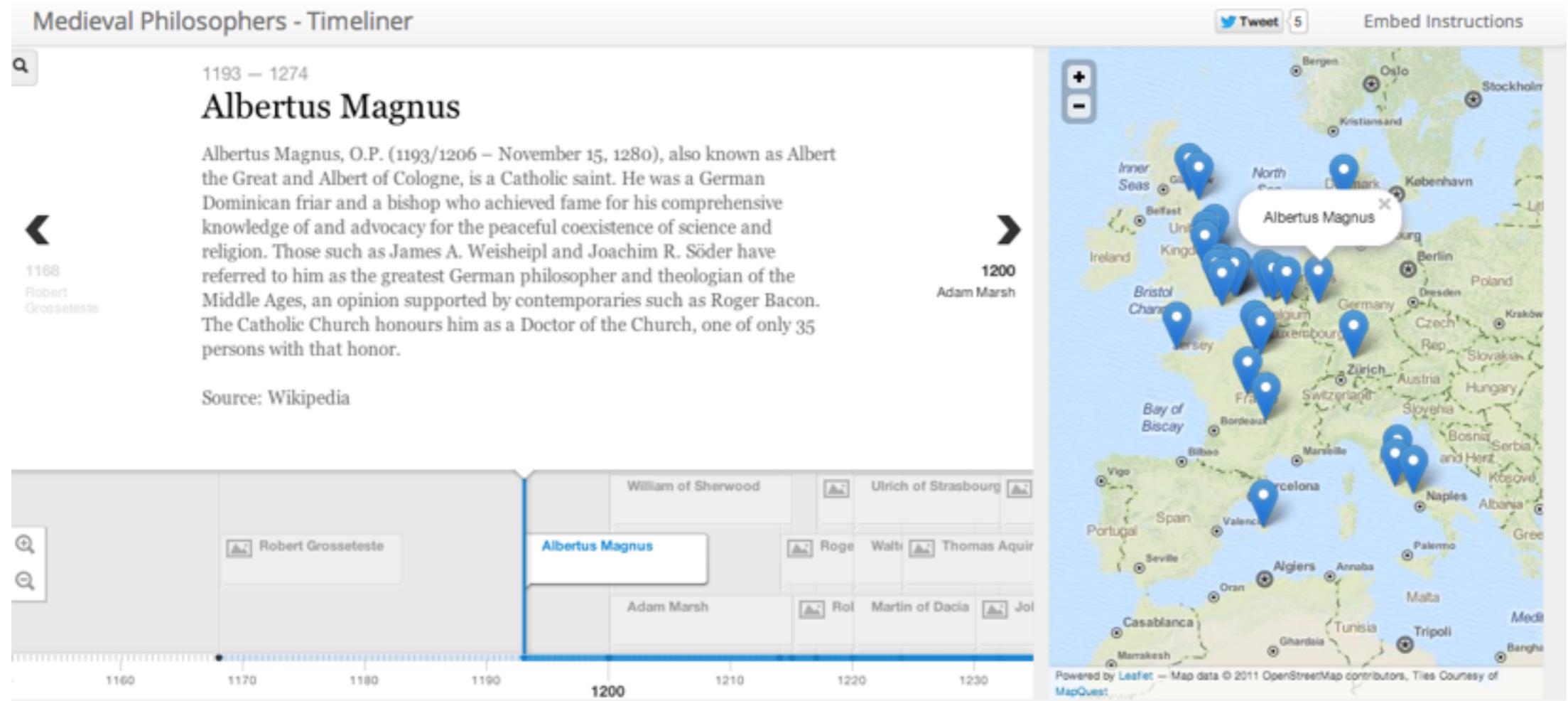
- 1871**: A thumbnail of a portrait labeled "Le portrait mystérieux". Below it, a caption reads: "Illustrate your Timeline with photos, videos, tweets and more."
- 1899**: A thumbnail of a YouTube video by Georges Méliès. The video player shows a play button and the duration "0:00 / 1:09". Below it, the caption reads: "YouTube video". A detailed description follows: "To add a YouTube video, just add a link to it in the media field. No embedding necessary." and provides the URL <http://youtu.be/lIvftGgps24>.
- 1902**: A thumbnail of a YouTube video with no text. Below it, the caption reads: "YouTube with no text".

At the bottom, there's a toolbar with icons for back, forward, search, and other functions. A sidebar on the left contains links like "It's Easy to Make Your Own Timeline" and "Illustrate your Timeline with photos, videos, tweets and more.". On the right, there are buttons for "Wikipedia", "YouTube video", "YouTube with no text", and "Blockquote".

<http://timeline.knightlab.com/>

Data Visualization Tools

Timeliner (OKFN)



Data Visualization Tools

Wordle

Wordle™ [Home](#) [Create](#) [Gallery](#) [Credits](#) [News](#) [Forum](#) [FAQ](#) [Advanced](#)

Paste in a bunch of text:

[Go](#)

OR

Enter the URL of any blog, blog feed, or any other web page that has an Atom or RSS feed.

[Submit](#)

© 2013 [Jonathan Feinberg](#)

[Terms of Use](#)

build #1411

<http://www.wordle.net/create>

Data Visualization Tools



Data Visualization Tools

Quadrigram

New geography of the *haute cuisine*

2005 - 2015

As a result of all those changes, the “geography” of the *haute cuisine* has mutated. If in 2005 Europe gathered 70% of the restaurants included in the list, in 2015 it's just 50%. Asia, South America and they concentrate the majority of the new succesfull proposals.

Select a year in the slider to visualize the number of restaurants in the top50 by continent.

2005



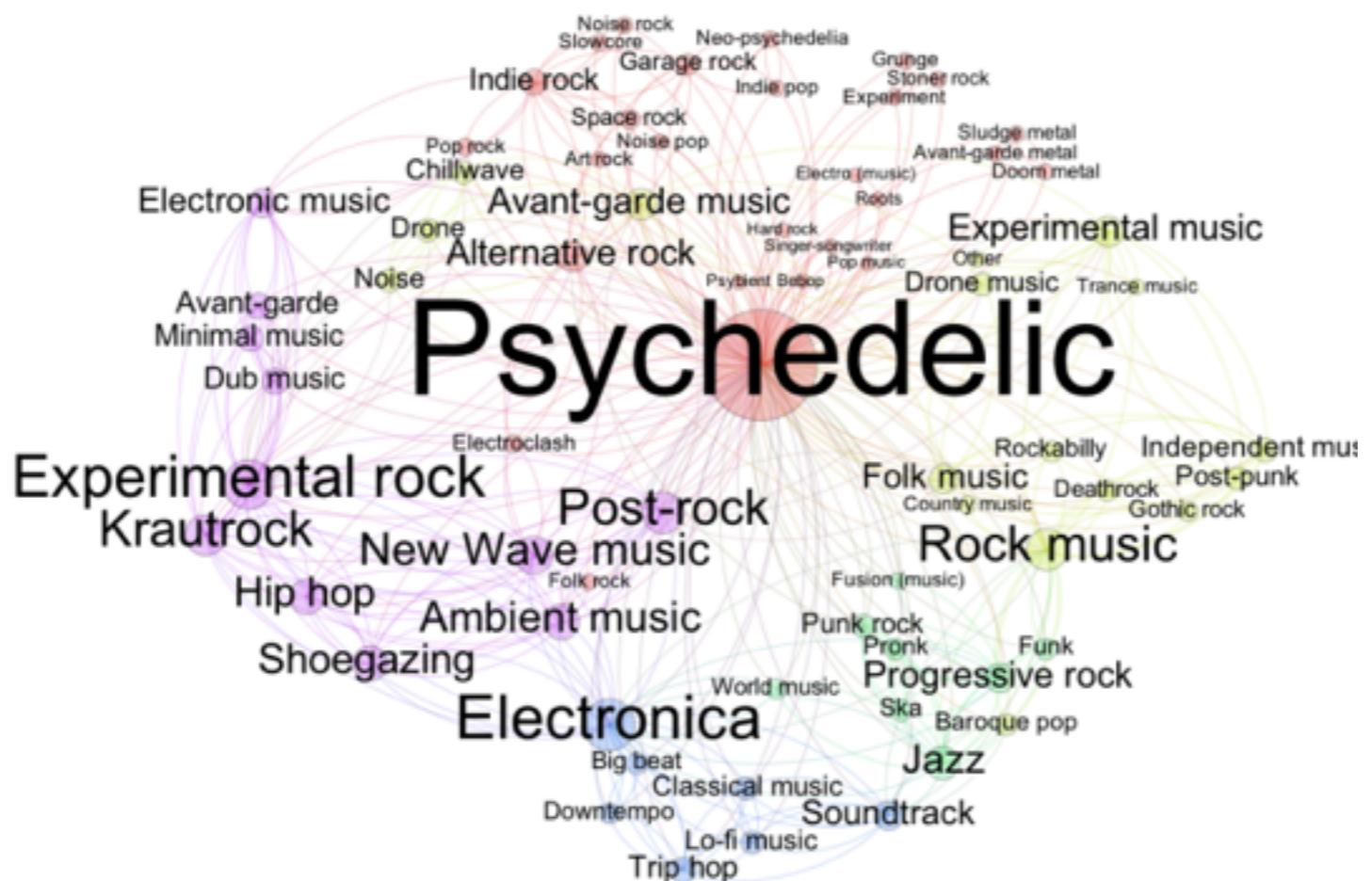
Data Visualization Tools



<http://www.tableausoftware.com/public/>

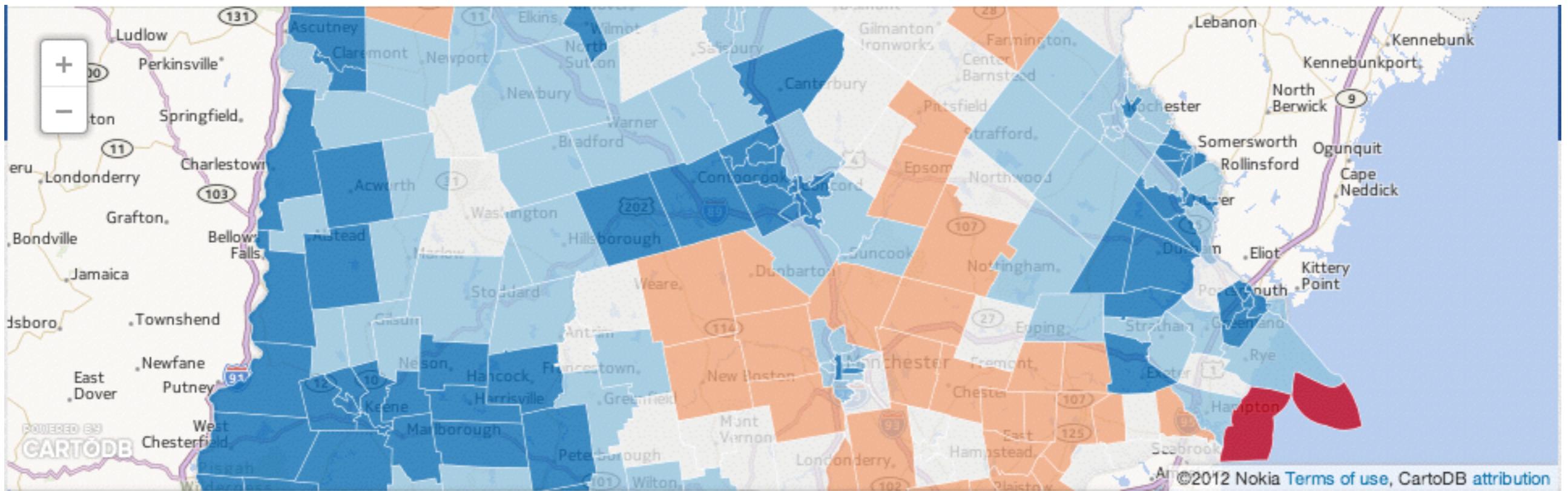
Data Visualization Tools

Gephi



Data Visualization Tools

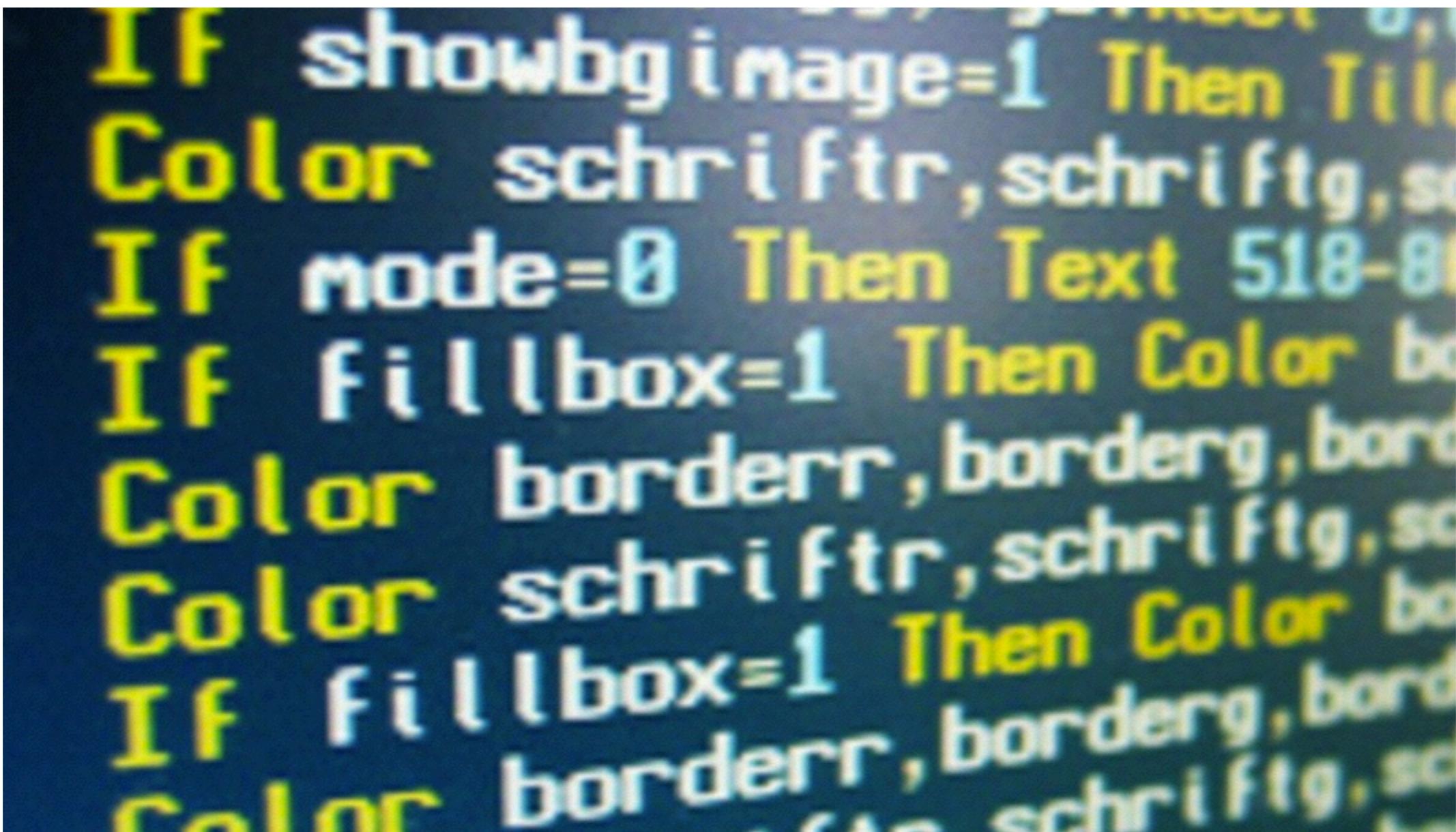
CartoDB



http://developers.cartodb.com/tutorials/electoral_map.html

<http://cartodb.com/>

Developers needed



```
If showbgimage=1 Then TitleColor = schriftfarbe
If mode=0 Then Text 518-8
If fillbox=1 Then Color = borderfarbe
Color = borderr, borderg, borderb
Color = schriftfarbe, schriftfarbe, schriftfarbe
Color = schriftfarbe, schriftfarbe, schriftfarbe
If fillbox=1 Then Color = borderfarbe
If fillbox=1 Then Color = borderfarbe
Color = borderr, borderg, borderb
Color = schriftfarbe, schriftfarbe, schriftfarbe
```

When you'll need a developer

- ▶ Size
- ▶ Real-Time
- ▶ Analysis beyond Basics
- ▶ Interactivity
- ▶ API access

References



OKFN

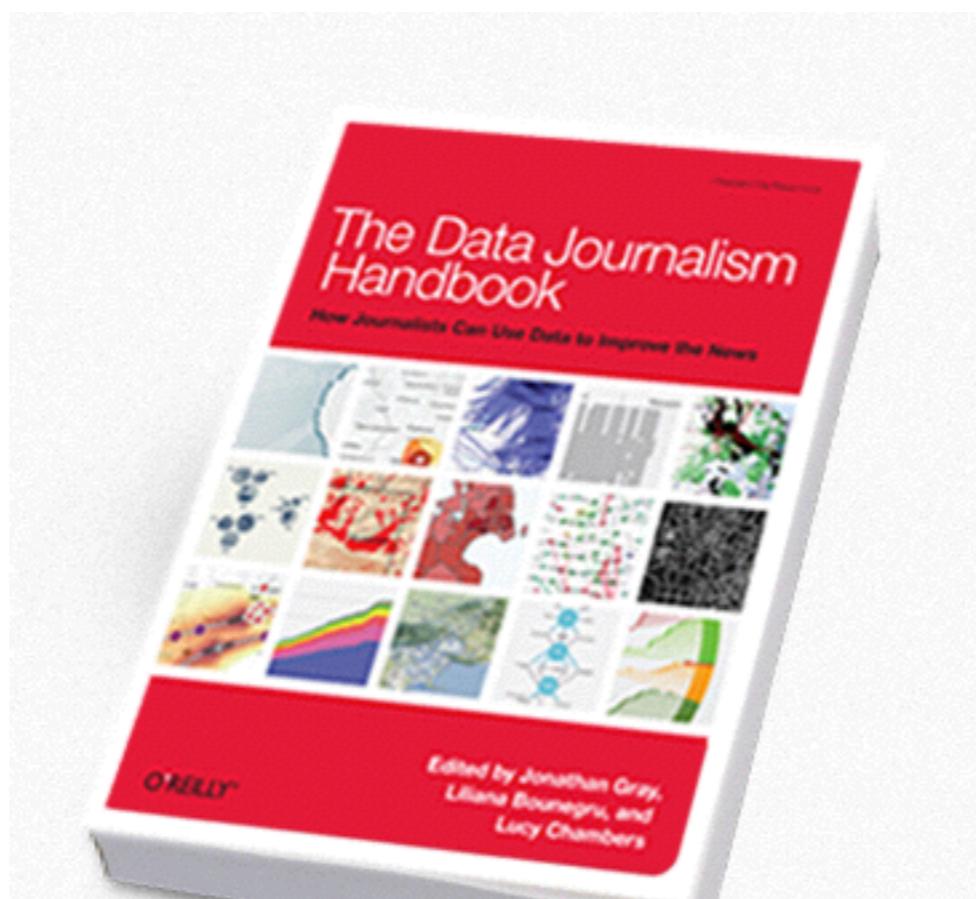
<http://okfn.org/>

<http://schoolofdata.org/>

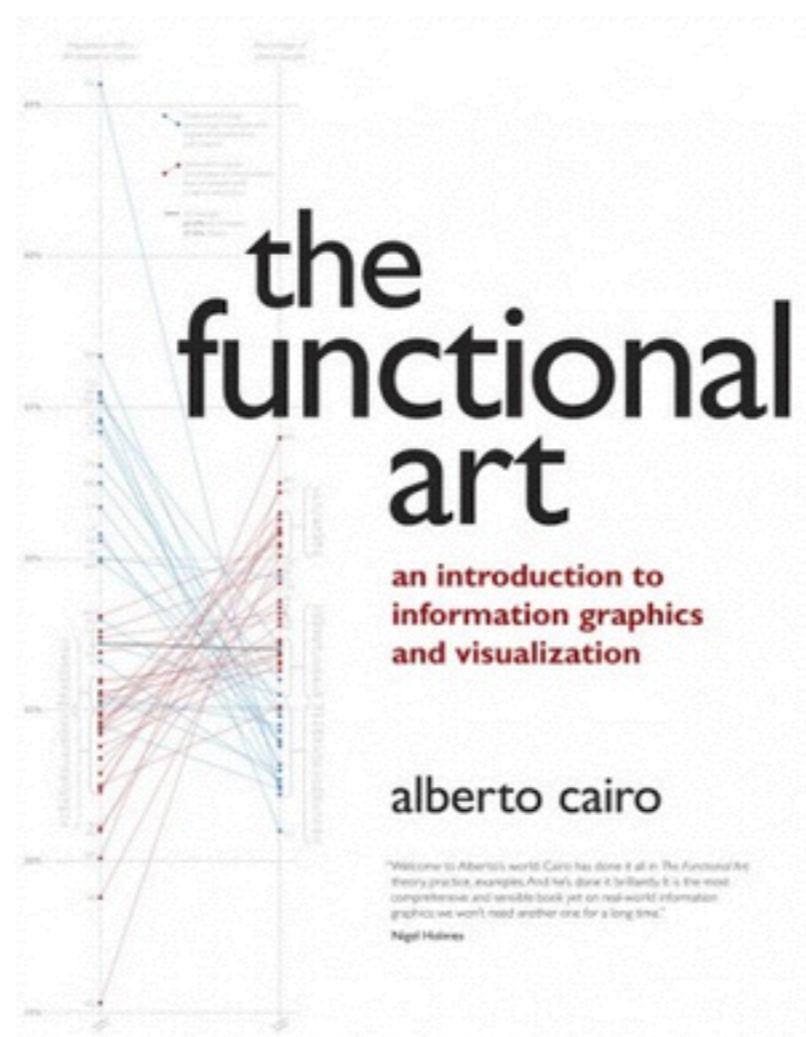
<http://ckan.org/>

<http://2014.okfestival.org/>

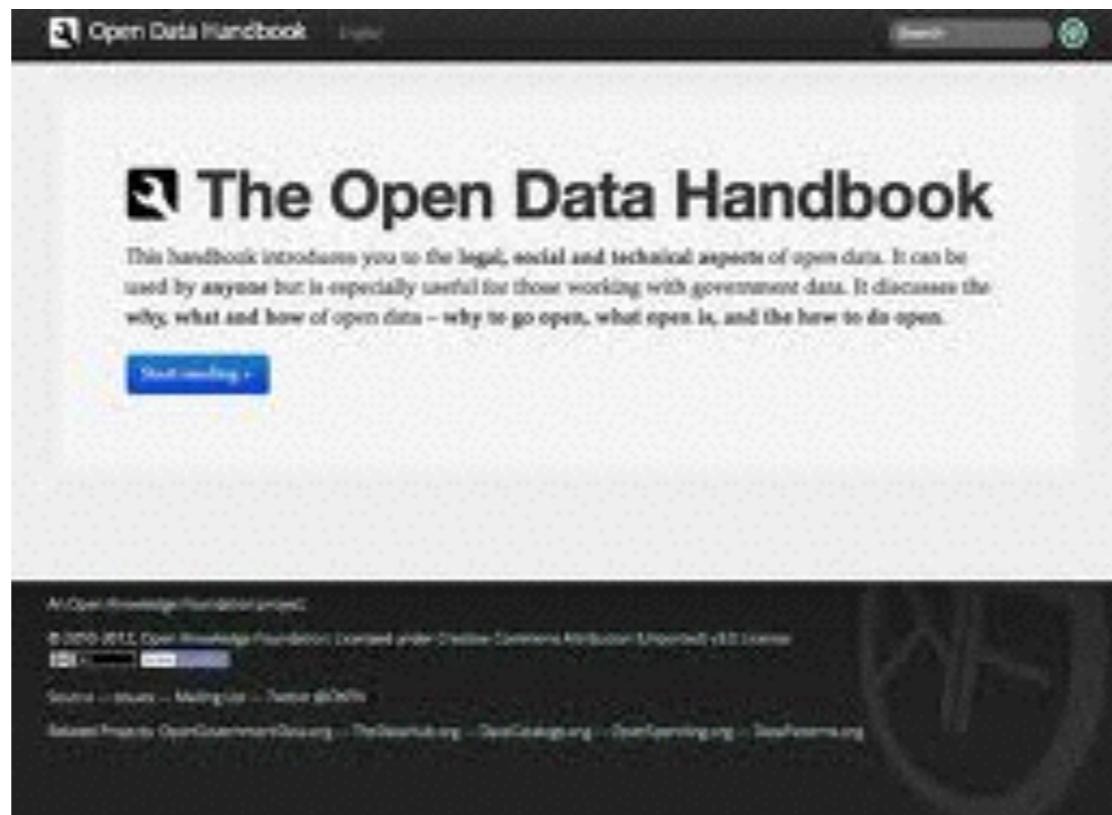
Books



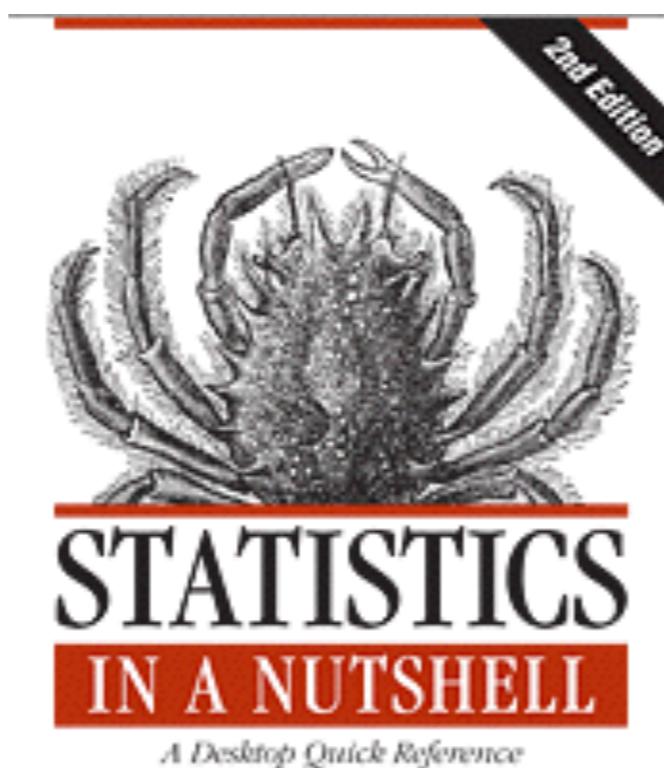
Books



Books



Books



O'REILLY®

Sarah Boslaugh

Data Sources

<http://ckan.org/>

<http://www.theguardian.com/data>

<http://www.google.com/publicdata/>

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

<http://datahub.io/>

<http://www.quora.com/Data/Where-can-I-find-large-datasets-open-to-the-public>

Any questions?
Thanks for your attention!

