Data Tools Ecosystem for non-programmers

TABAKALERA

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

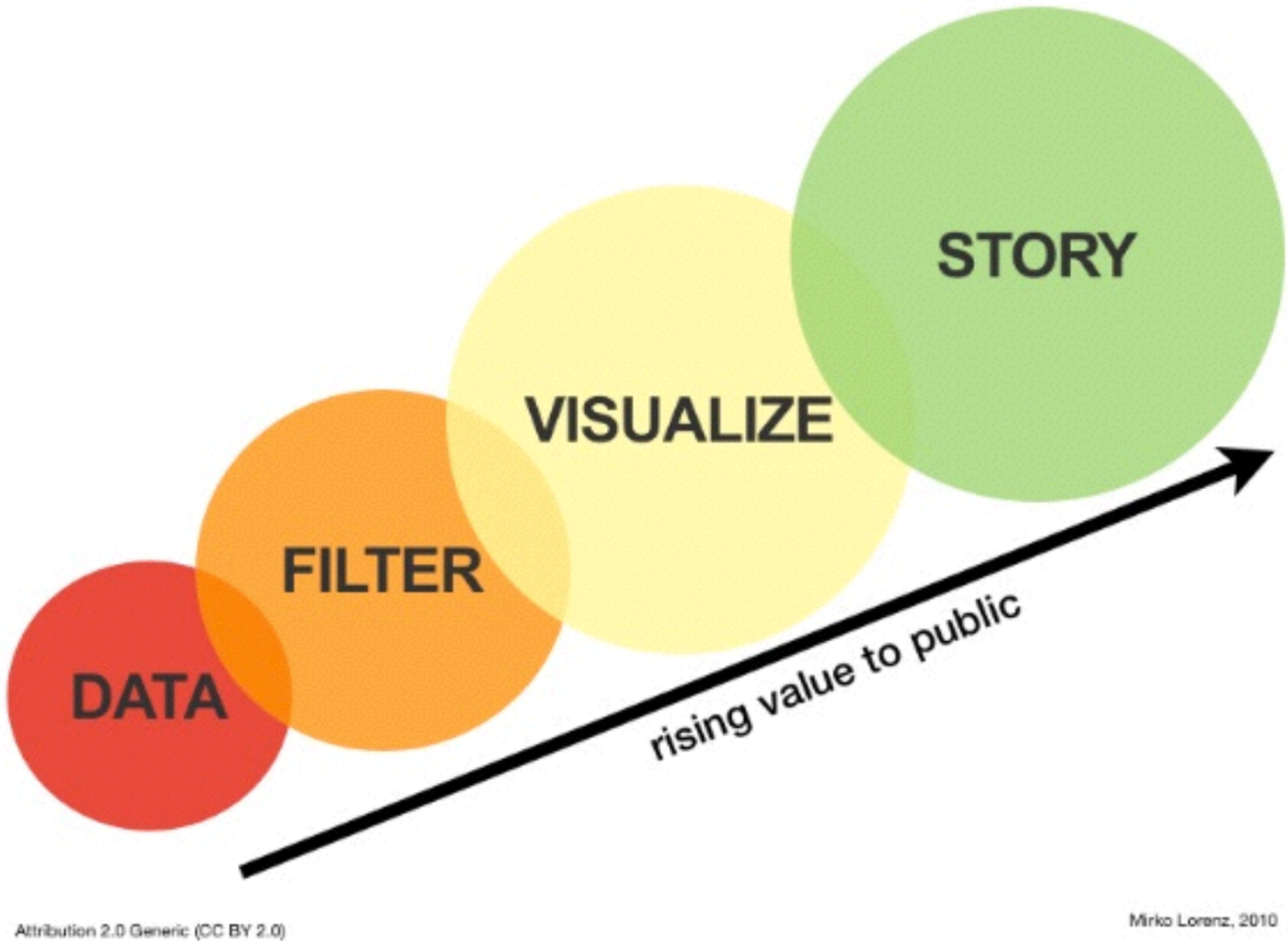
- Timeline.js
- Wordle
- Plot.ly
- Google Fusion Tables
- Gephi
- CartoDB
- MapBox
- Tilemill

Developers needed

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

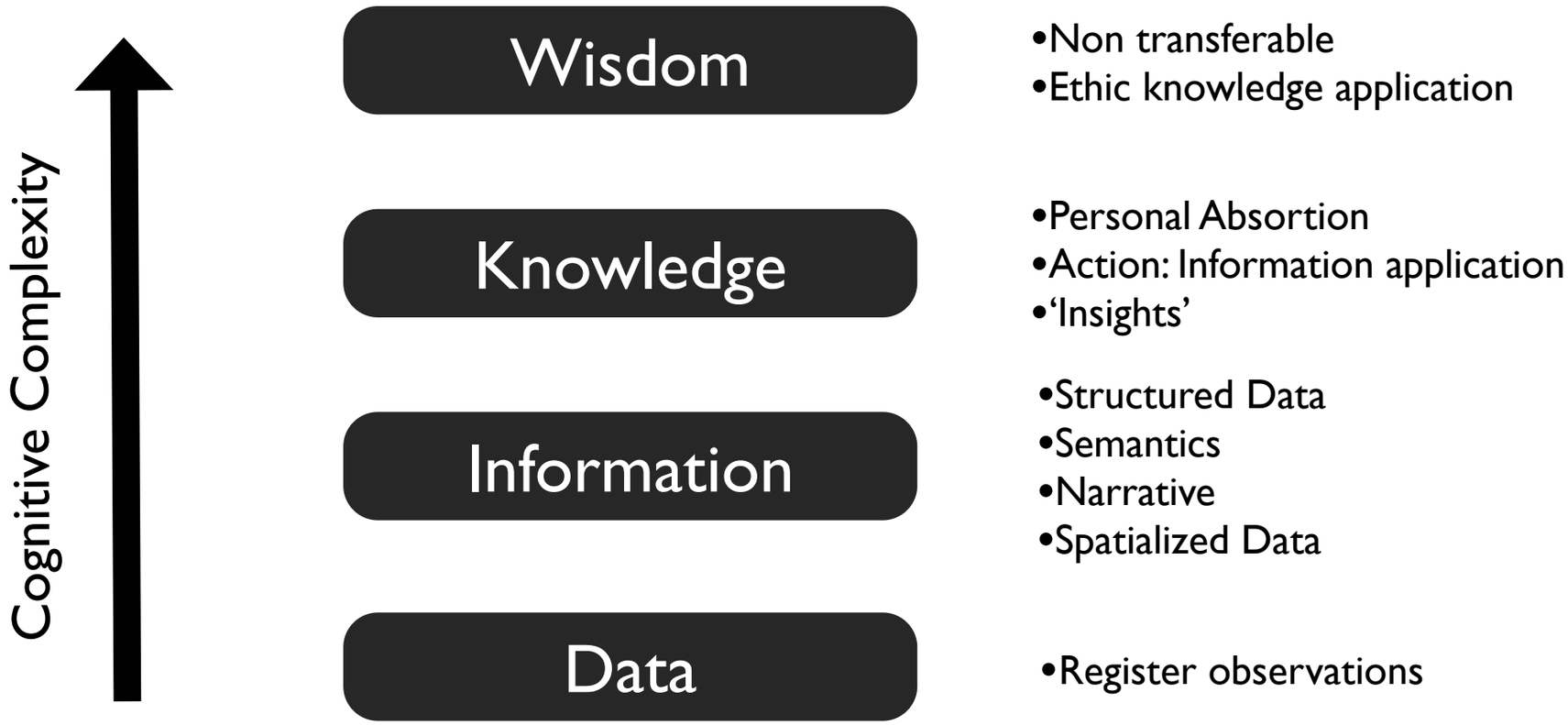
Resources

The Data Process

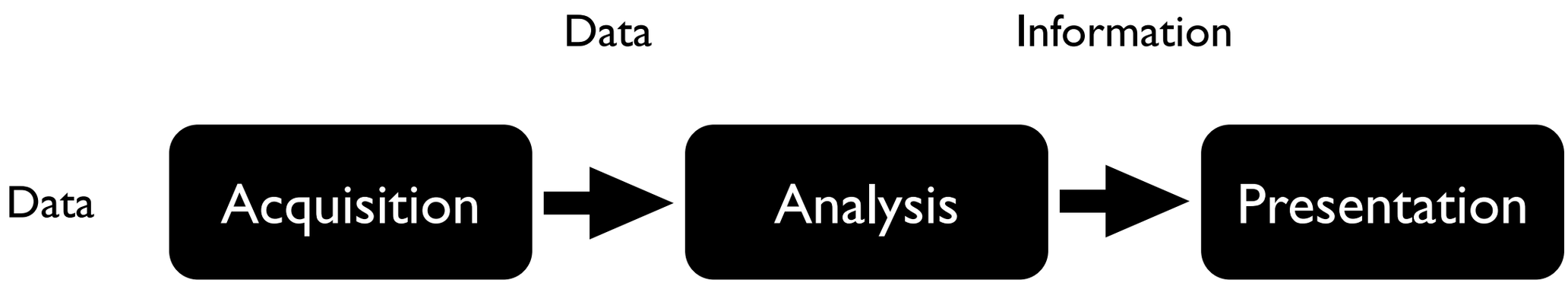


http://en.wikipedia.org/wiki/Data-driven_journalism

Why mess with data?



A Data Pipeline



- API's
- Scraping
- Files
- DB's
- Providers

- Text Mining
- Machine Learning
- Network Metrics
- Stats
- Filtering

- Visualization
- Story

Anscombe's quartet

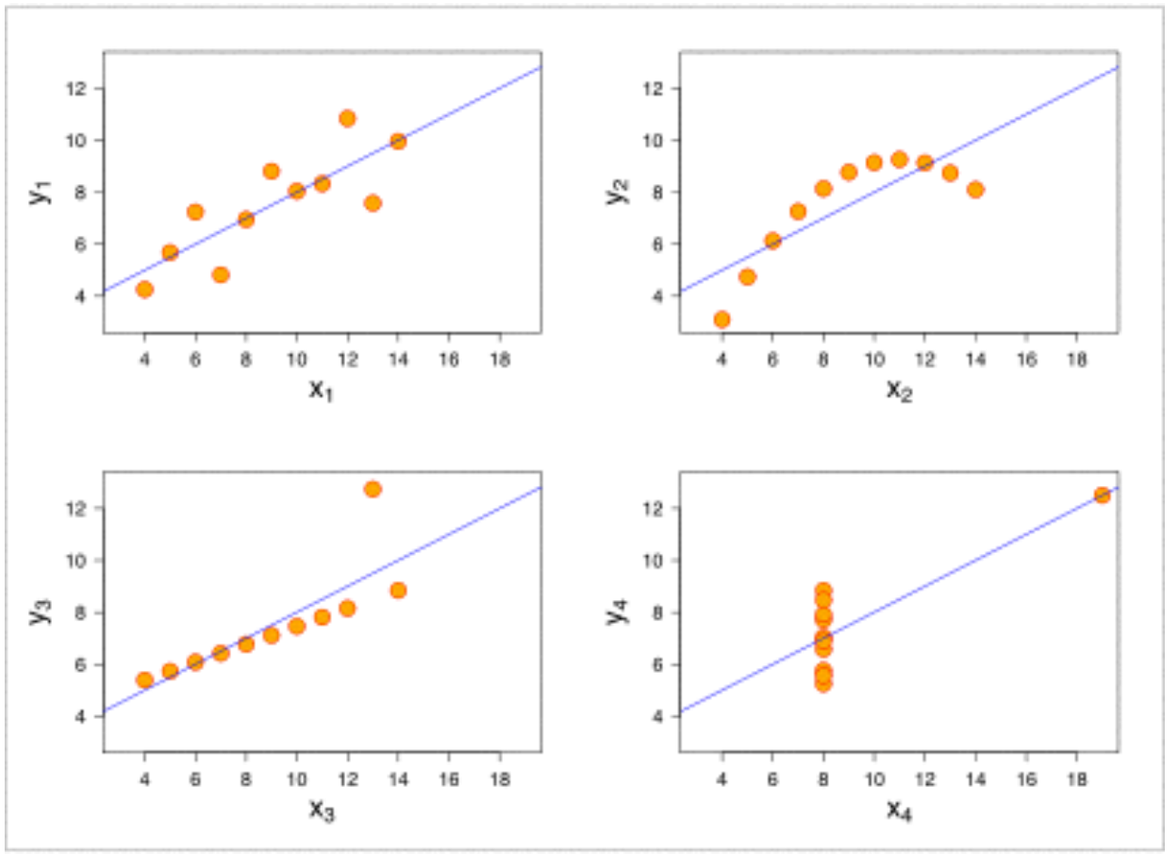
Anscombe's quartet

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)

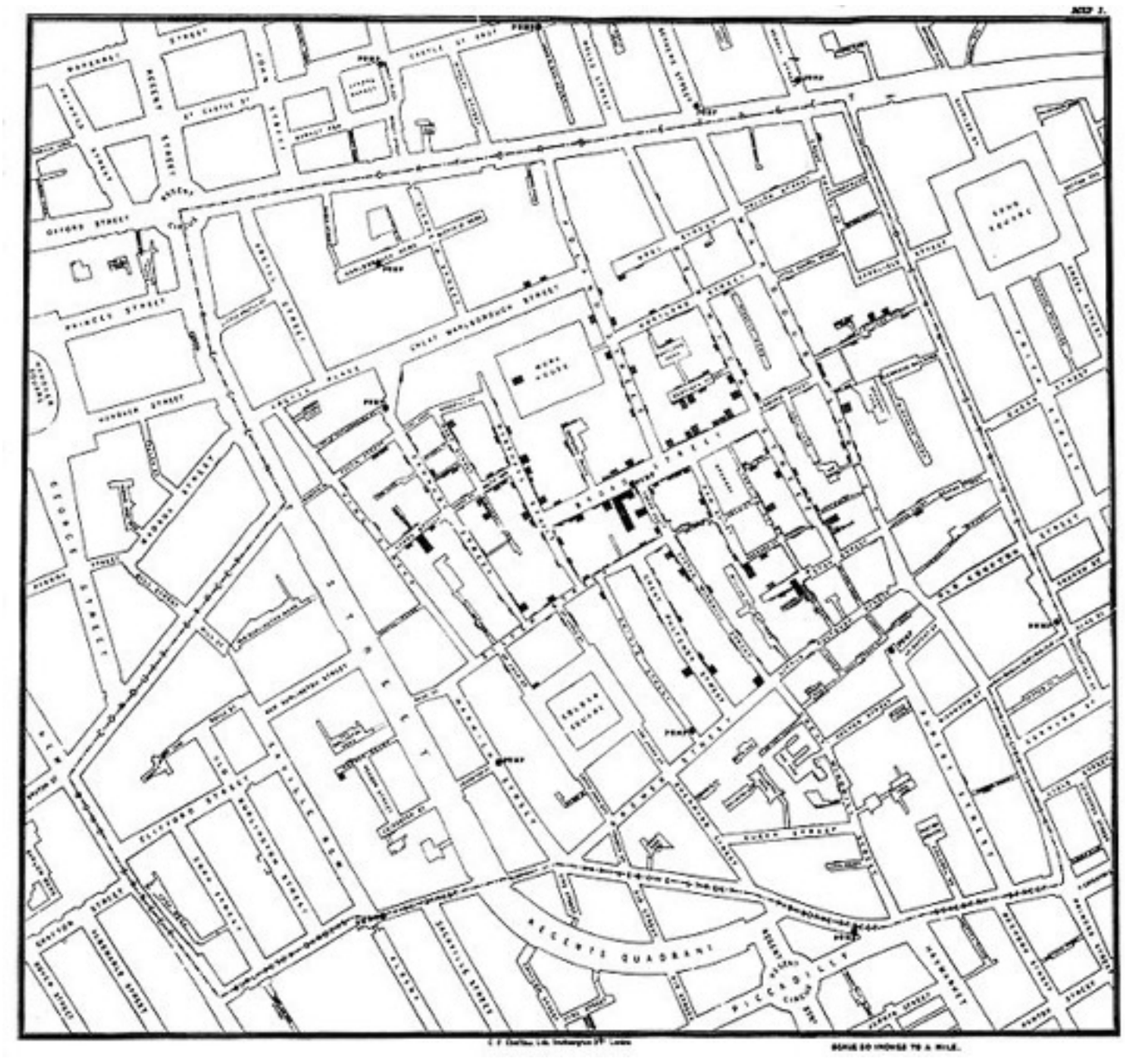
http://en.wikipedia.org/wiki/Anscombe's_quartet

Anscombe's quartet



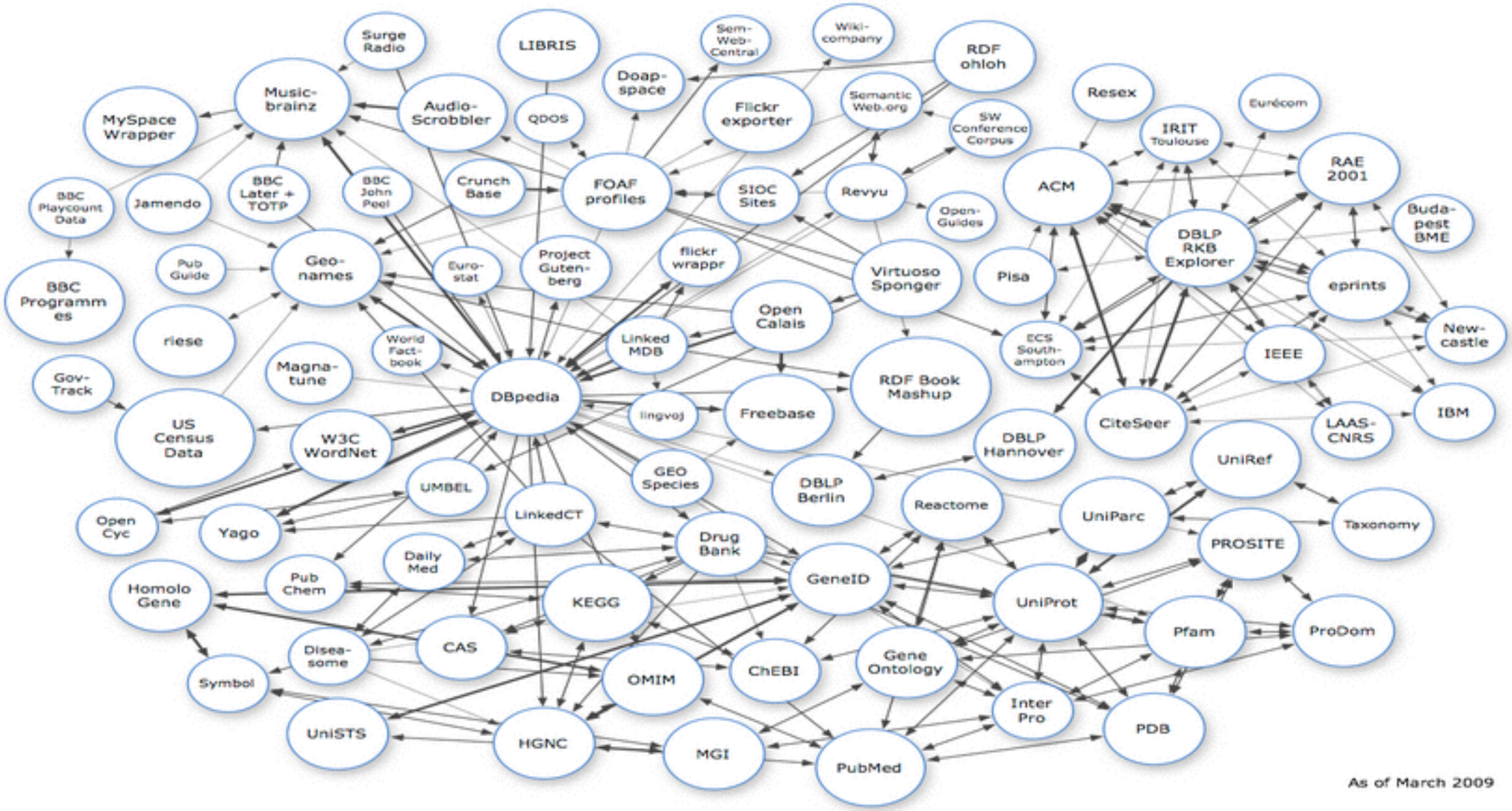
http://en.wikipedia.org/wiki/Anscombe's_quartet

John Snow



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

Data Acquisition Tools



As of March 2009

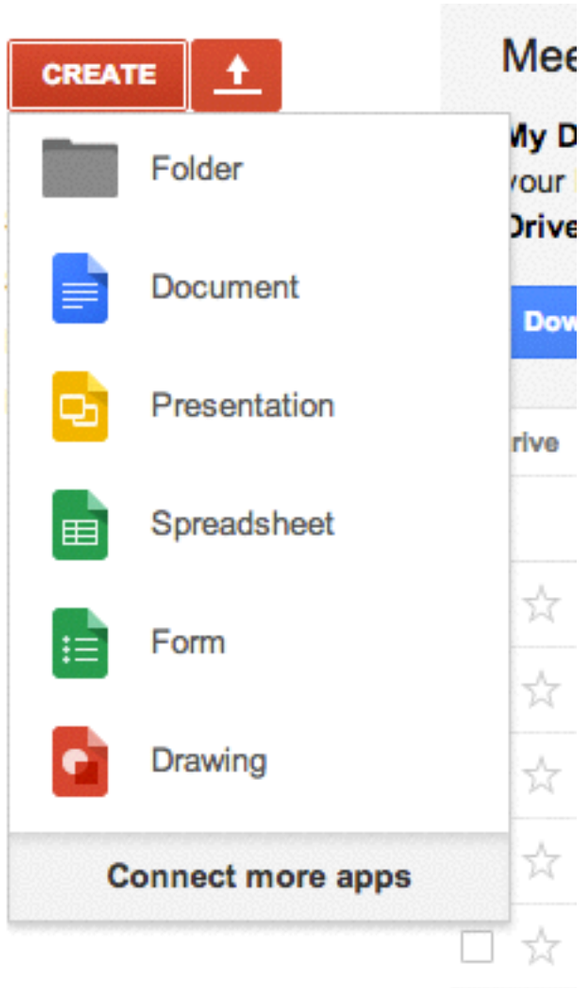
Data Formats

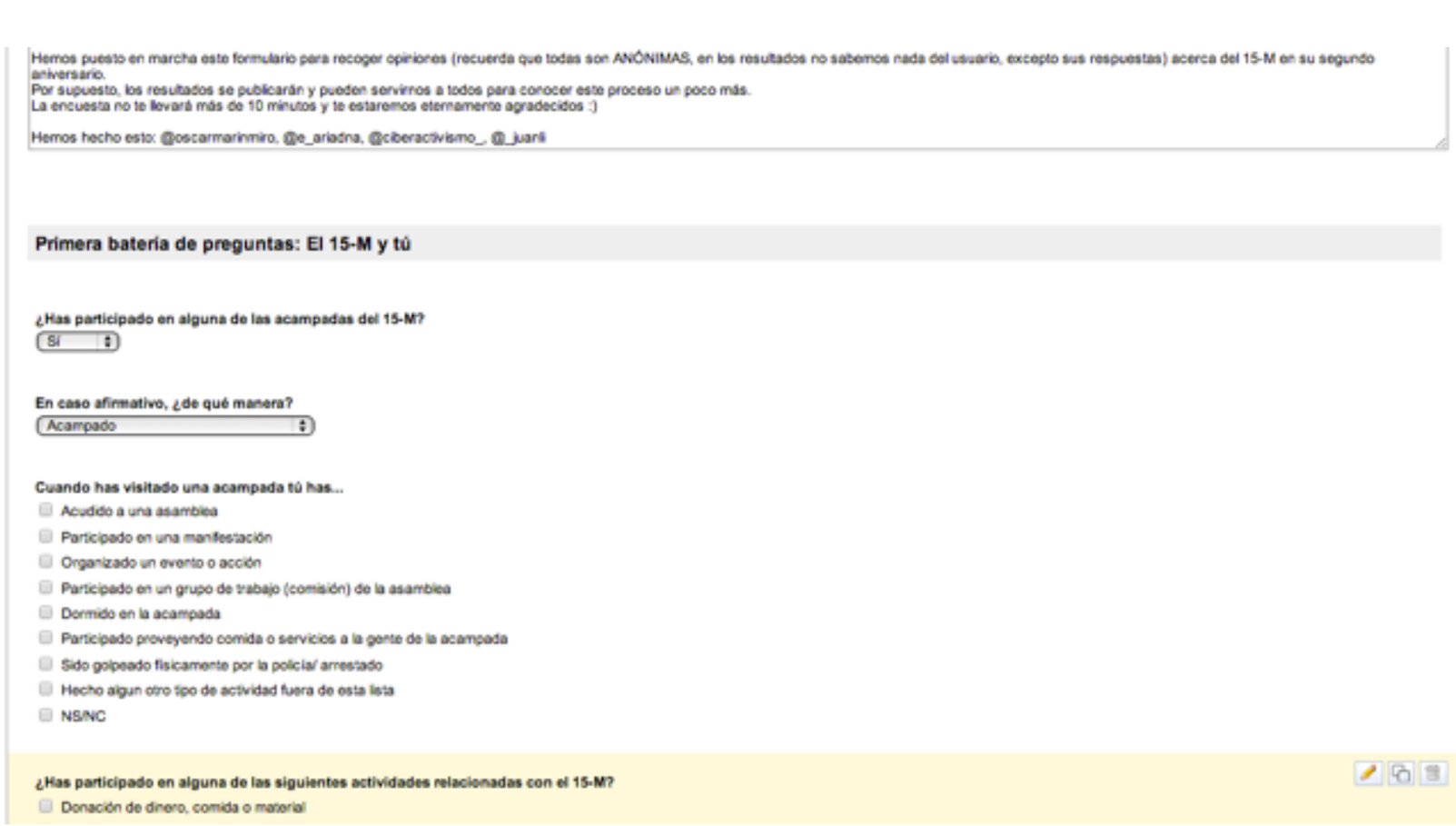
- JSON (Javascript Object Notation) - Dev
- CSV, XLS - Tabular
- XML - Old, but still used
- Web data

Data Acq Techniques

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

Google Forms




Hemos puesto en marcha este formulario para recoger opiniones (recuerda que todas son ANÓNIMAS, en los resultados no sabemos nada del usuario, excepto sus respuestas) acerca del 15-M en su segundo aniversario.
Por supuesto, los resultados se publicarán y pueden servirnos a todos para conocer este proceso un poco más.
La encuesta no te llevará más de 10 minutos y te estaremos eternamente agradecidos :)

Hemos hecho esto: @oscarmarinmiro, @e_ariadna, @ciberactivismo_, @_juanli

Primera batería de preguntas: El 15-M y tú

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

En caso afirmativo, ¿de qué manera?
Acampado

Cuando has visitado una acampada tú has...
- 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 algun otro tipo de actividad fuera de esta lista
- NS/NC

¿Has participado en alguna de las siguientes actividades relacionadas con el 15-M?
- Donación de dinero, comida o material

Google Forms

15MPoll ☆

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

oscar@outliers.es

Comments Share

$ % 123 Arial 10

fx Timestamp

	A	B	C	D	E	F	G	H	I	J	K
1	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 15-M:	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?	co
283	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		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	Una sem
								Firmar una petición en Change, Realizar un boicot a un producto o servicio por razones políticas, éticas				

Google Forms

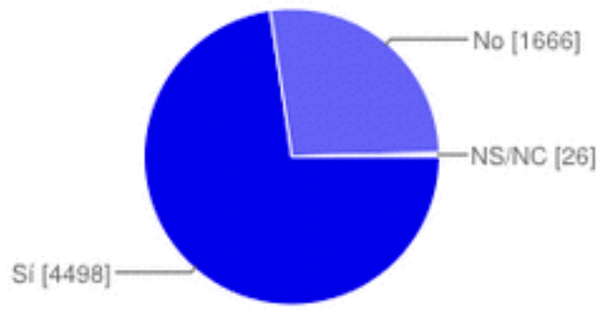
6194 responses

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

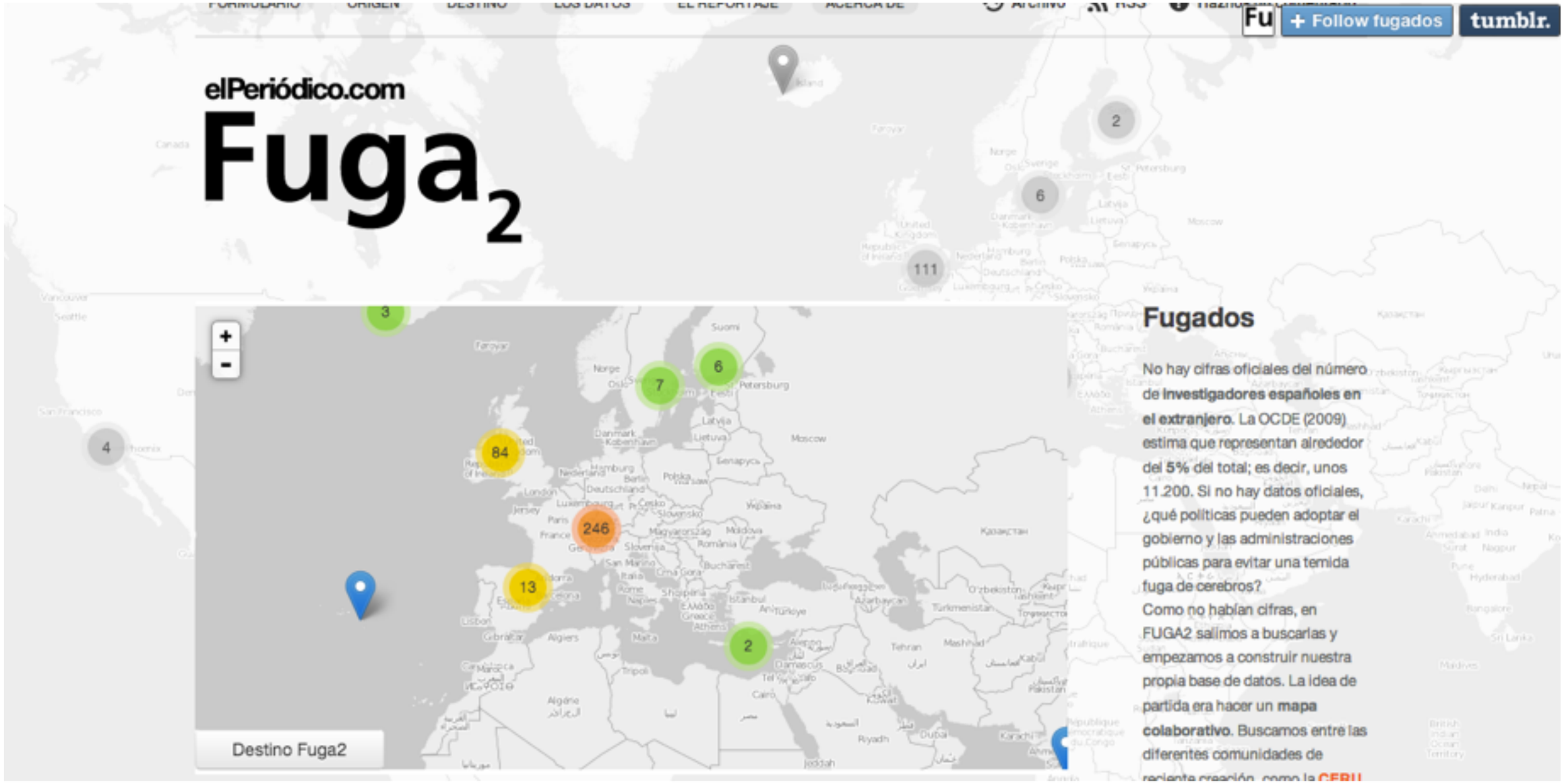
Google Forms

CUP ¿Con cuál de estos partidos te identificas más?

Motivaciones y preocupaciones

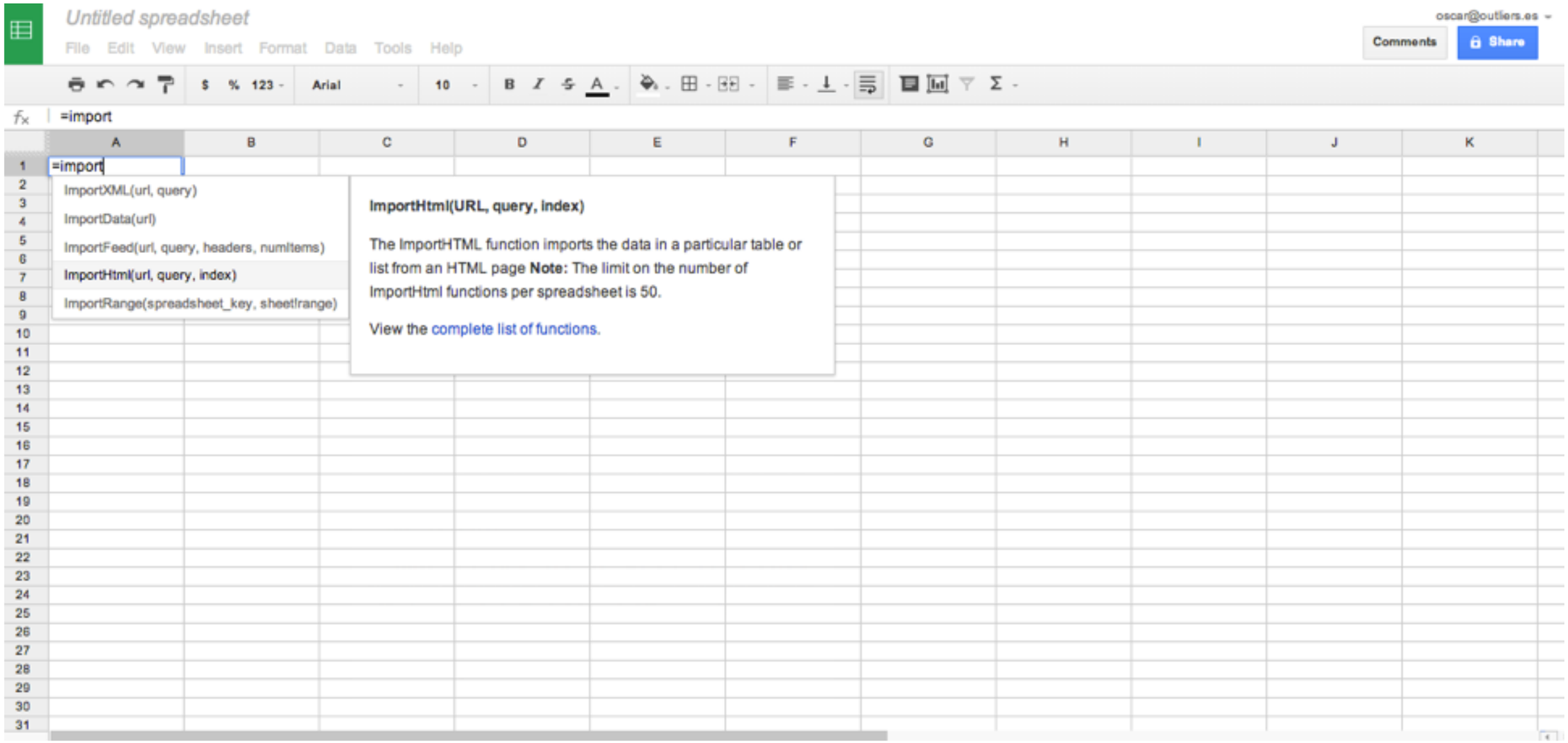
Colectivos asociados

Google Forms



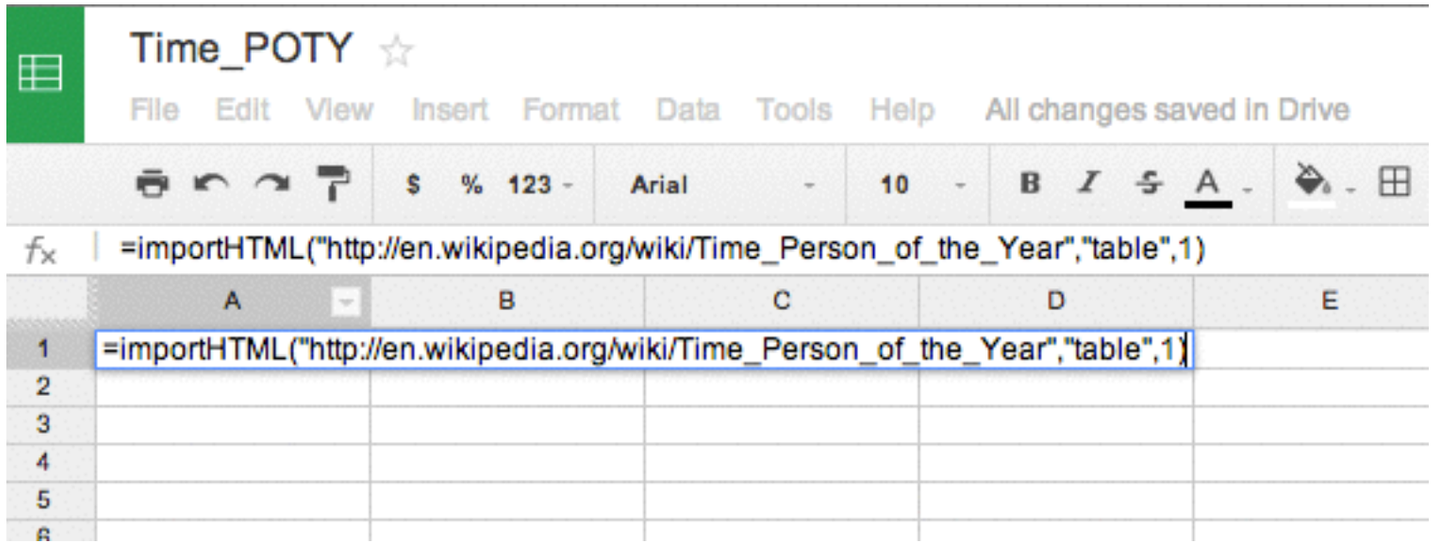
http://data.elperiodico.com/

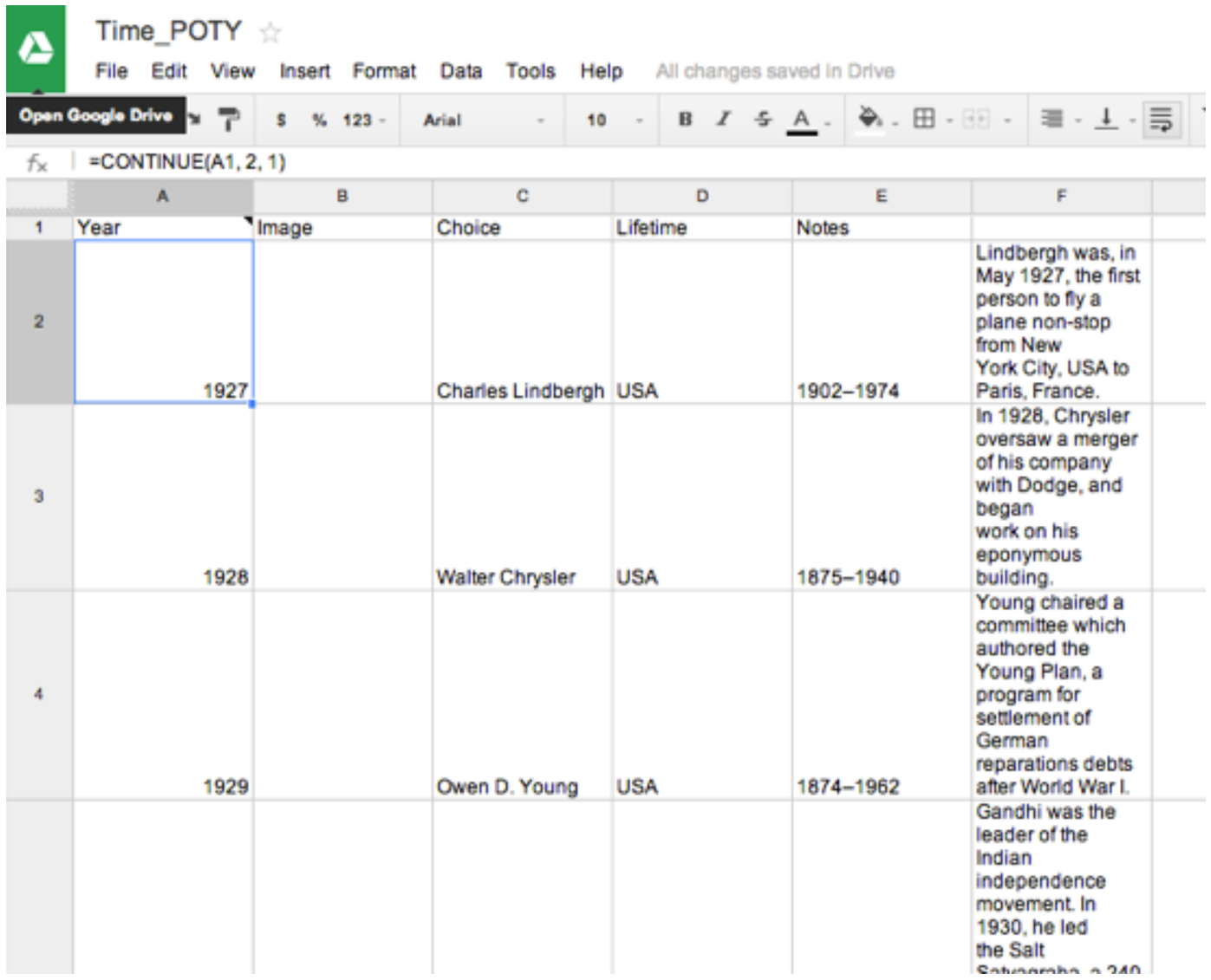
Google Spreadsheet



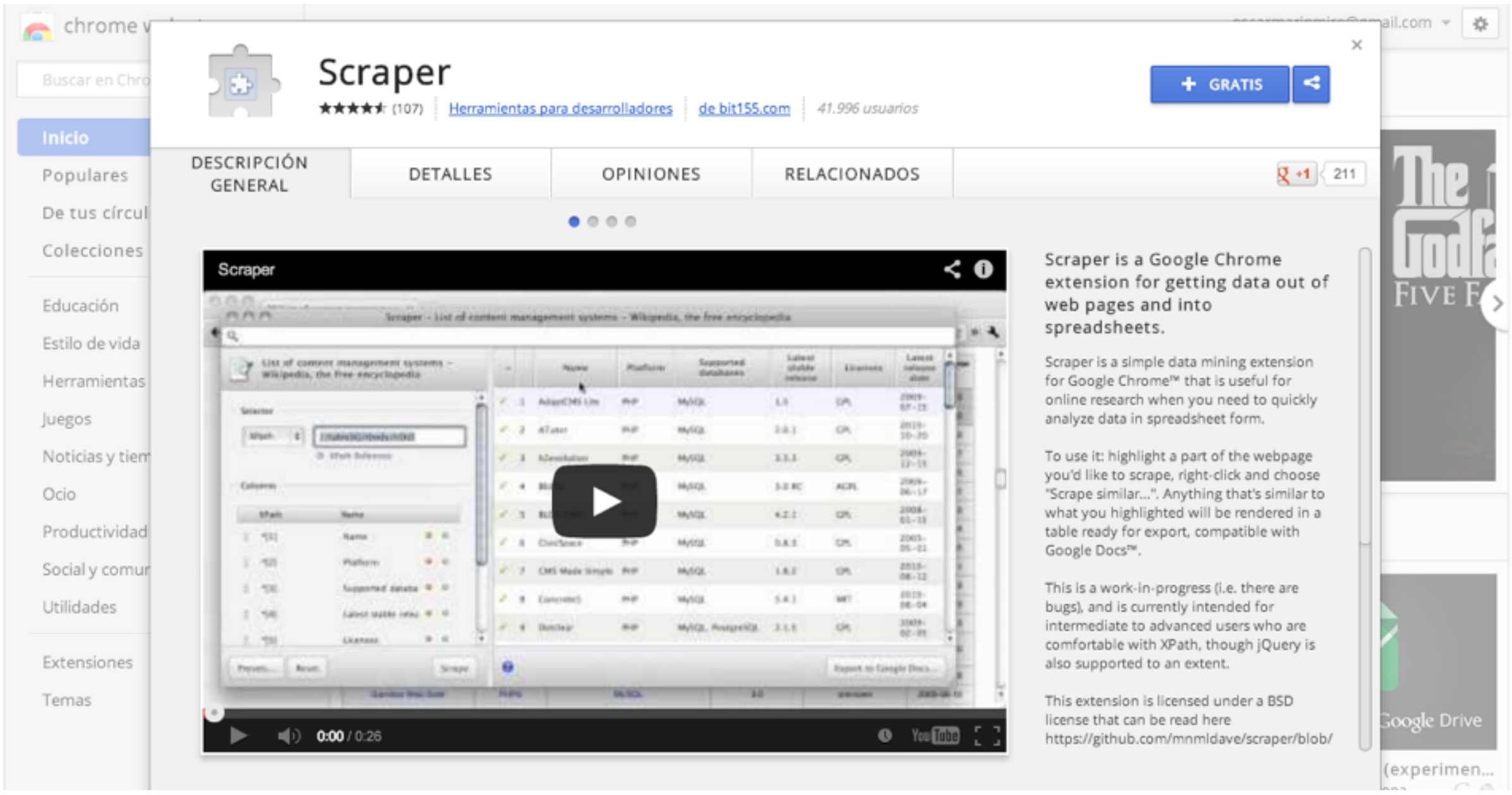
http://mashe.hawksey.info/2012/10/feeding-google-spreadsheets-exercises-in-import/

Google Spreadsheet


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

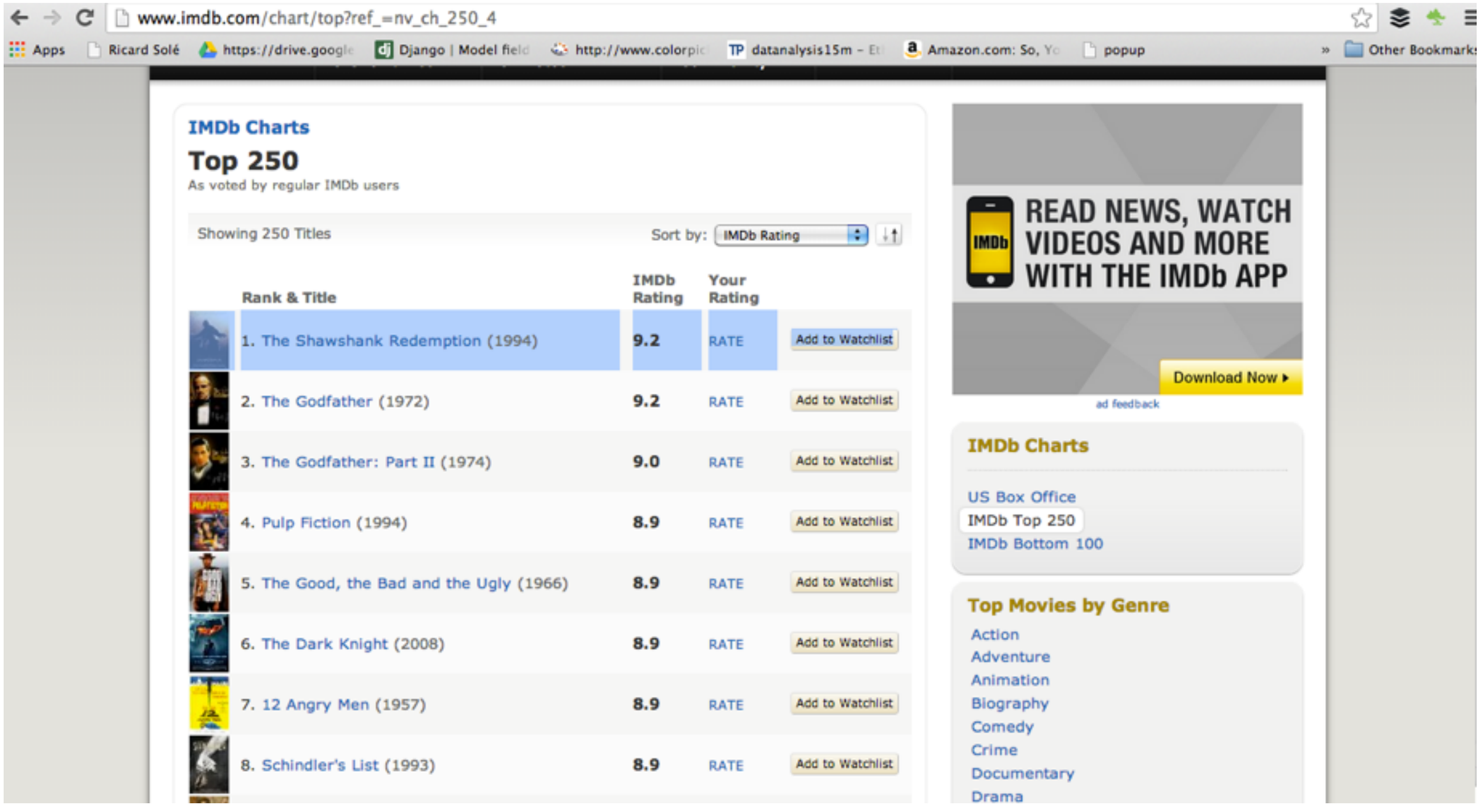

Time_POTY
File Edit View Insert Format Data Tools Help All changes saved in Drive
Open Google Drive
=CONTINUE(A1, 2, 1)
A B C D E F
1 Year Image Choice Lifetime Notes
2 1927 Charles Lindbergh USA 1902–1974 Lindbergh was, in May 1927, the first person to fly a plane non-stop from New York City, USA to 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

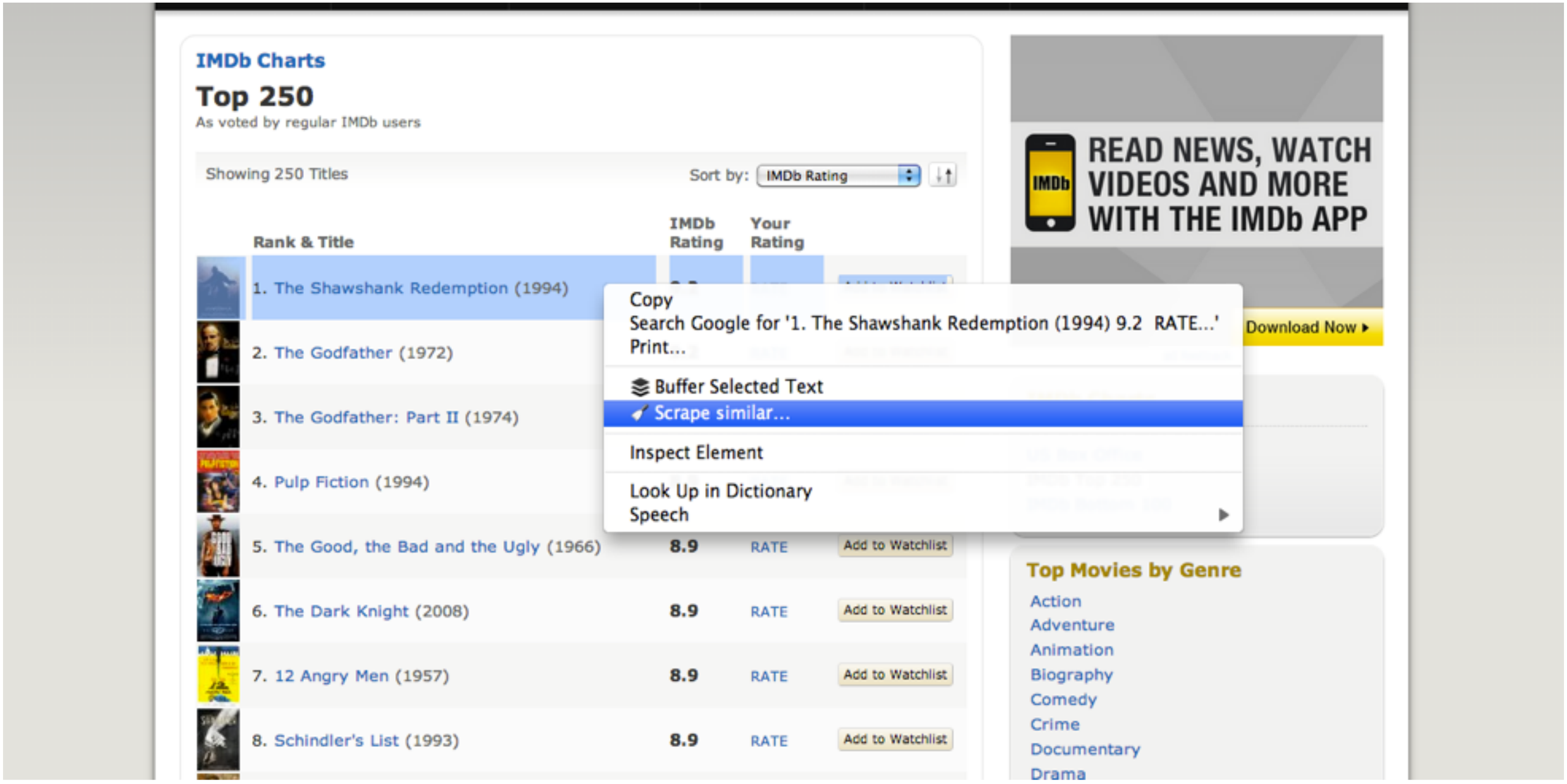


https://chrome.google.com/webstore/detail/scraper/mbigbapnjcgaffohmbkdlecaccepngjd

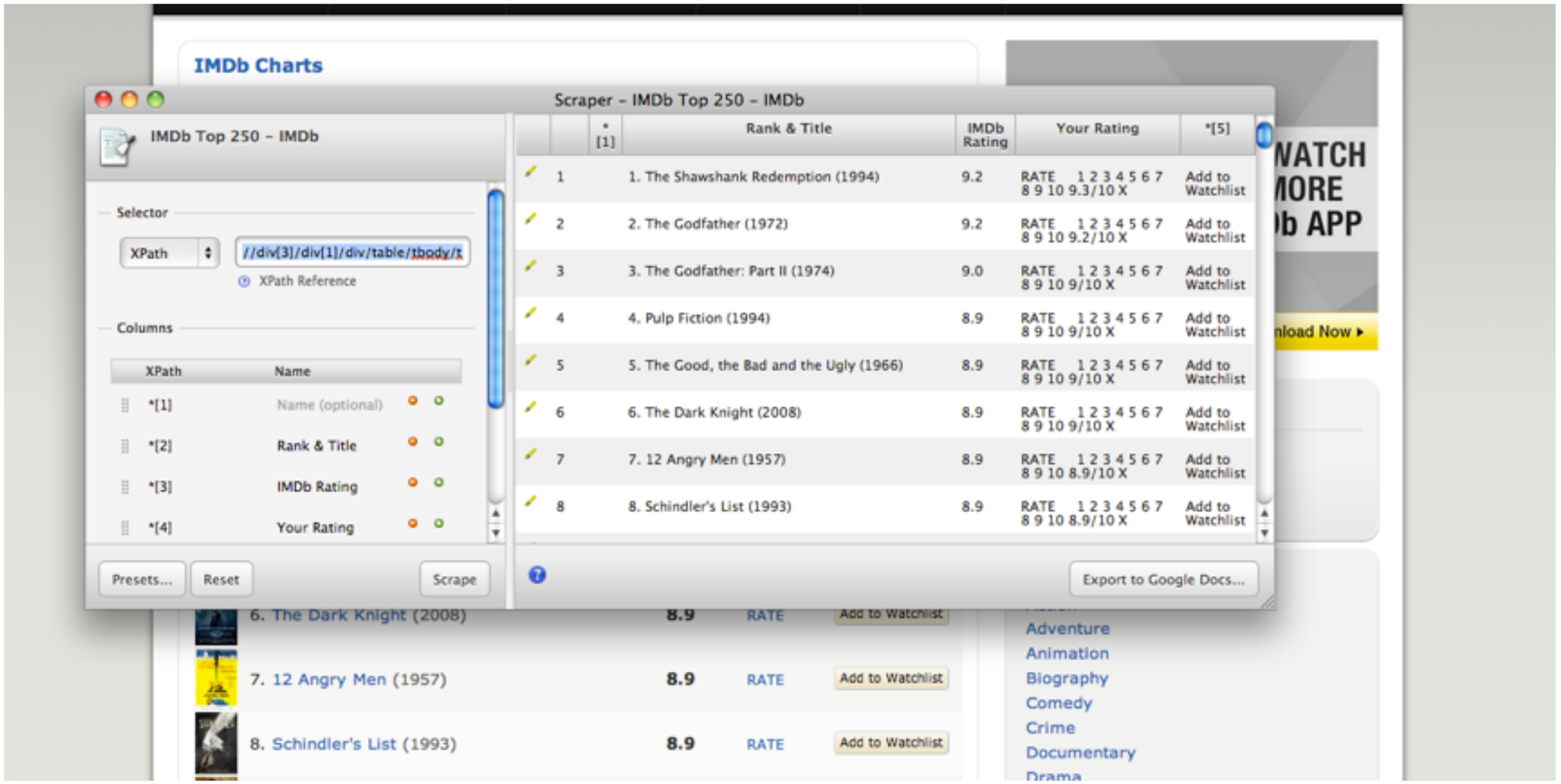
Scrape w/ Chrome Scraper



Scrape w/ Chrome Scraper

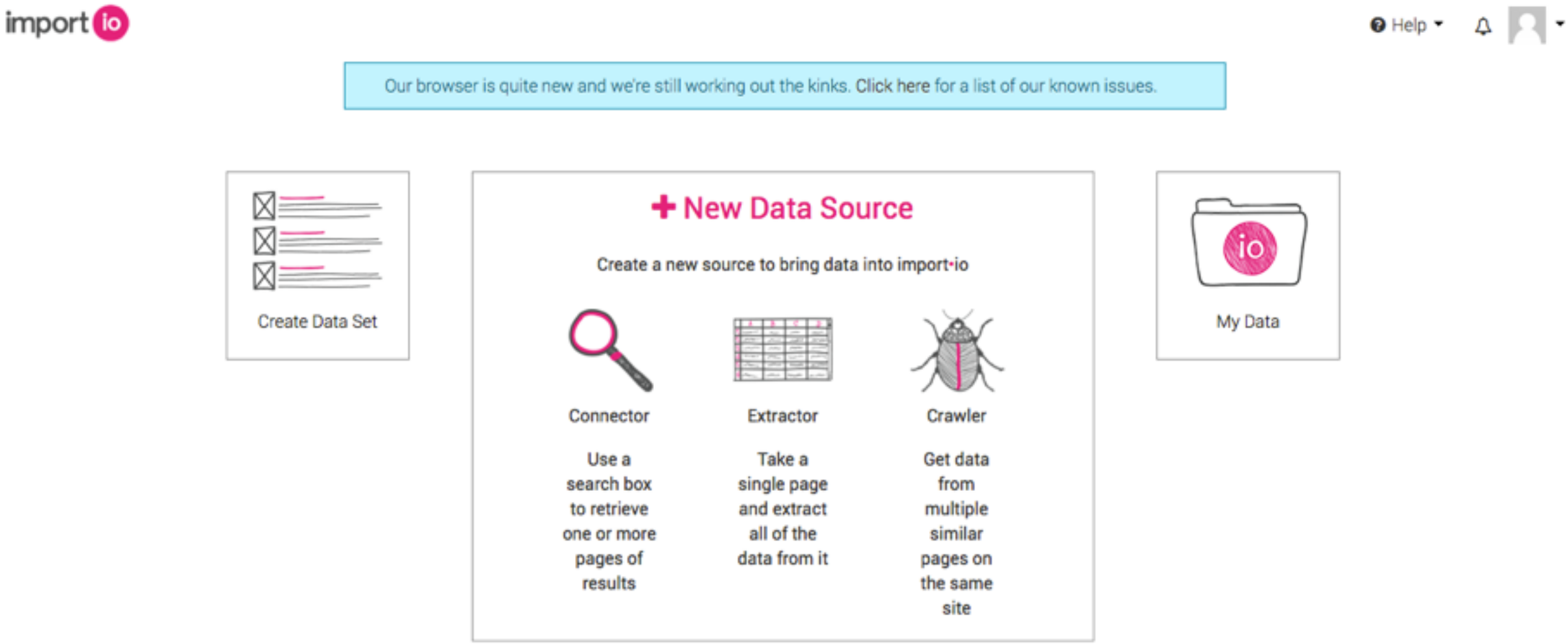


Scrape w/ Chrome Scraper



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

Scrape w/ import.io

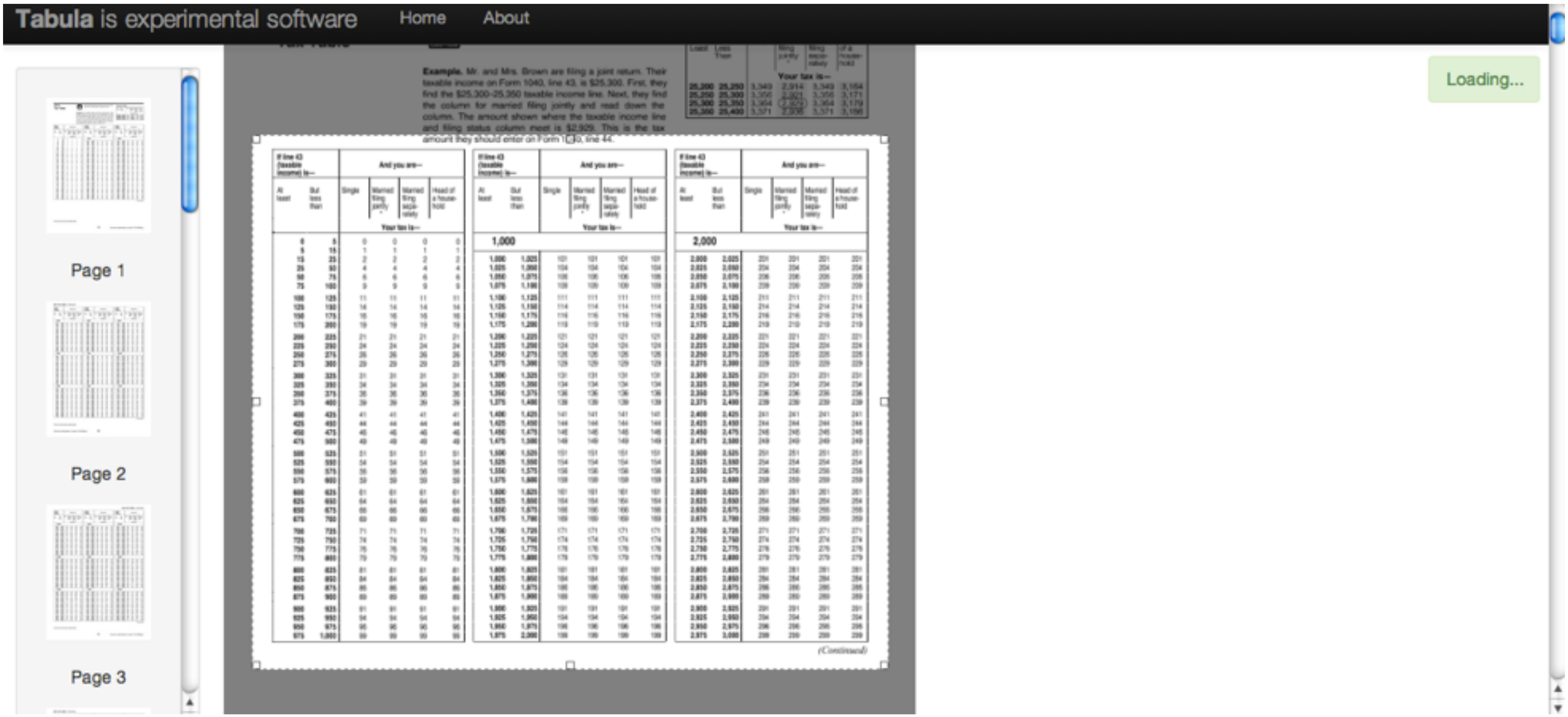


http://import.io/

PDF Extraction: Tabula

http://tabula.nerdpower.org/

PDF Extraction: Tabula



http://tabula.nerdpower.org/

PDF Extraction: Tabula

Extracted tabular data

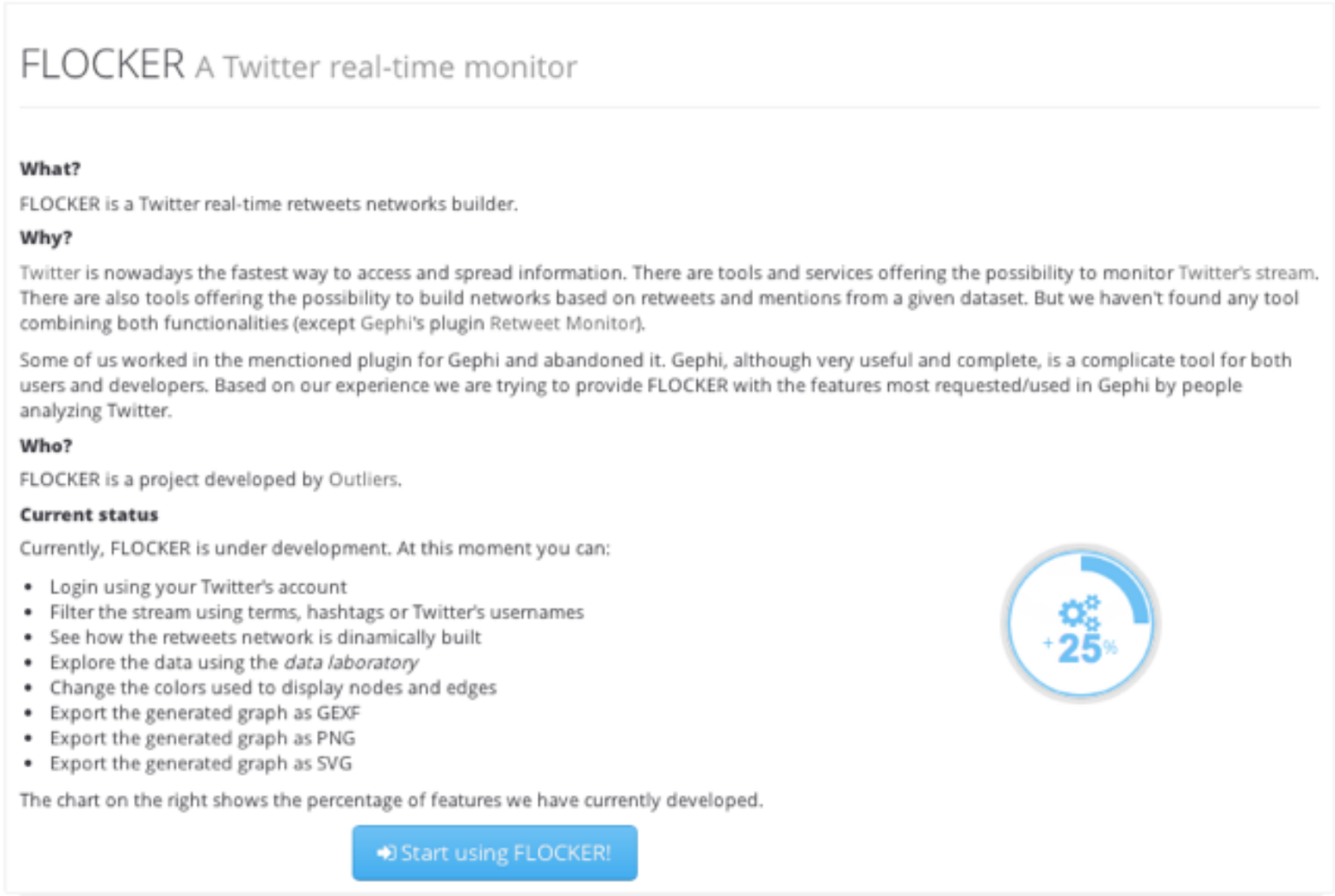
3,000					
3,0003,050	3,0503,100	303308	303308	303308	303308
3,1003,150	3,1503,200	313318	313318	313318	313318
3,2003,250	3,2503,300	323328	323328	323328	323328
3,3003,350	3,3503,400	333338	333338	333338	333338
3,4003,450	3,4503,500	343348	343348	343348	343348
3,5003,550	3,5503,600	353358	353358	353358	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

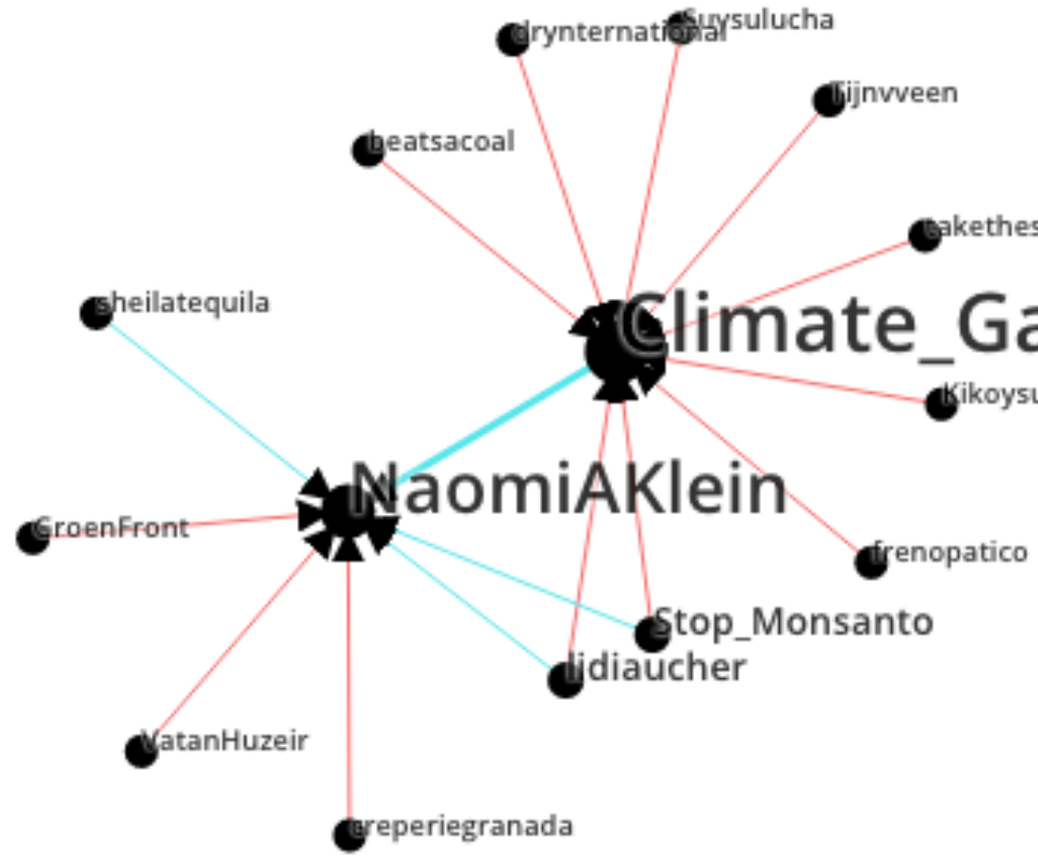
http://tabula.nerdpower.org/

Network Data: Flocker



http://flocker.outliers.es/

Network Data: Flocker



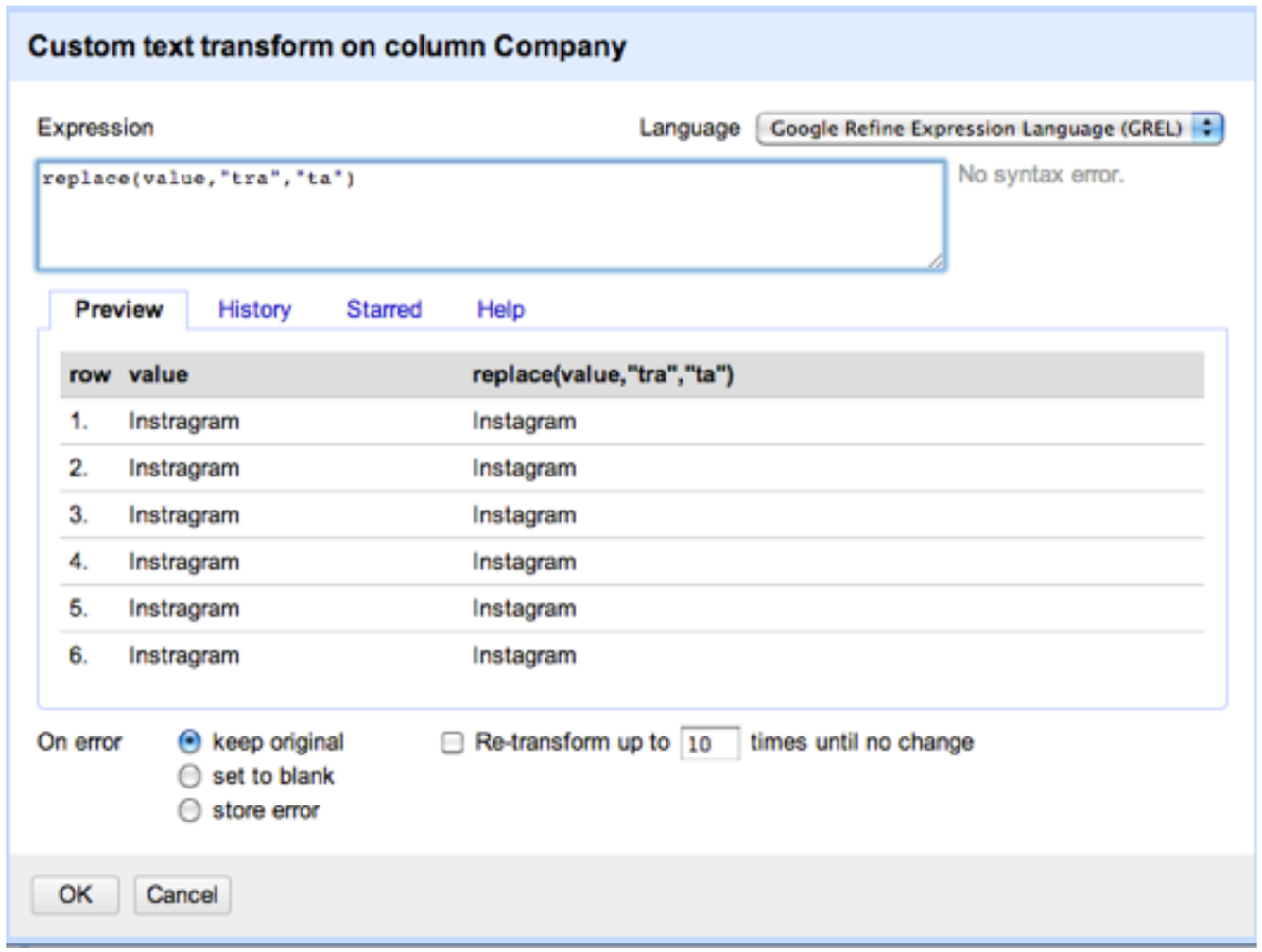
http://flocker.outliers.es/

Open Refine: Cleaning Data

https://github.com/OpenRefine

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

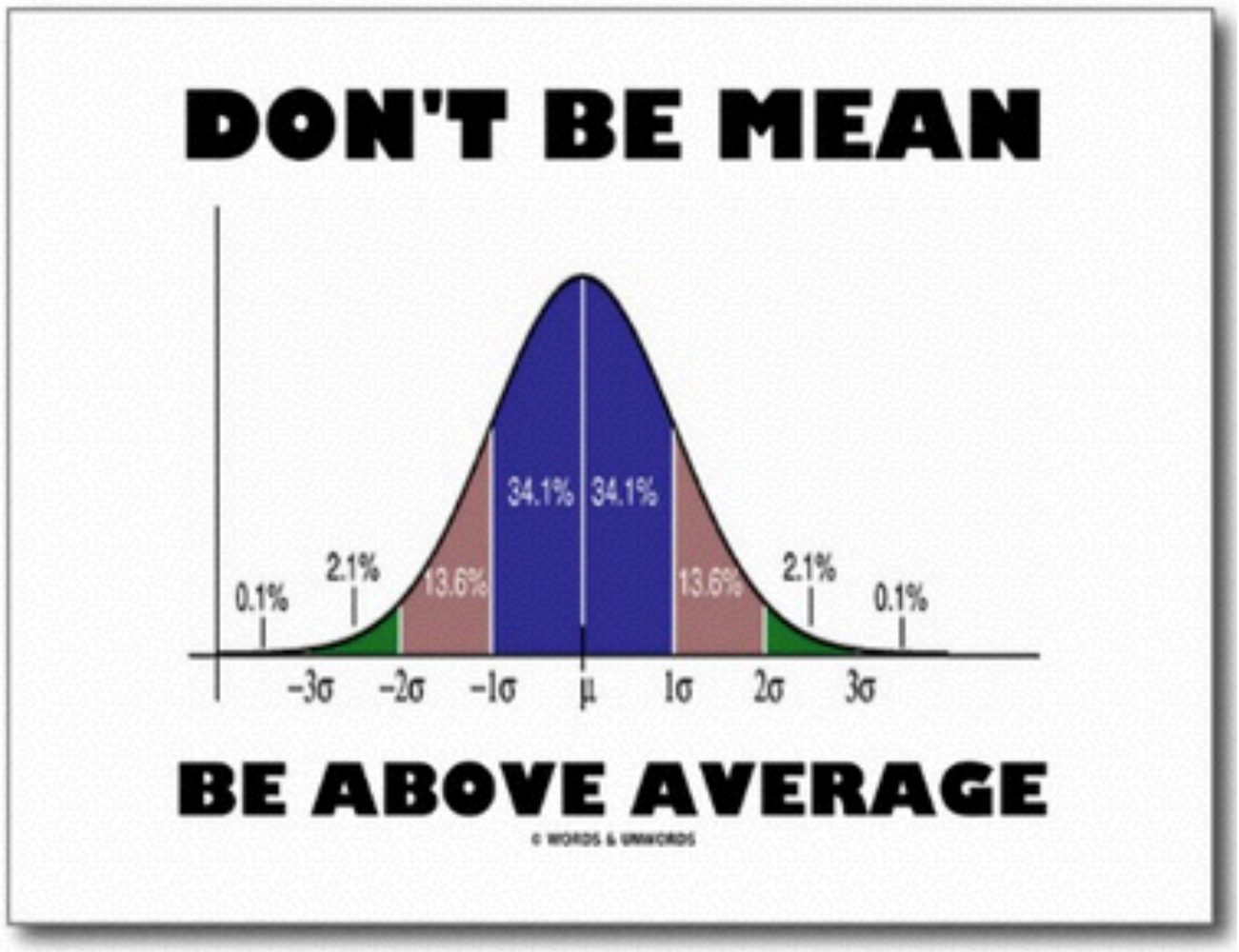
Open Refine: Cleaning Data



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

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

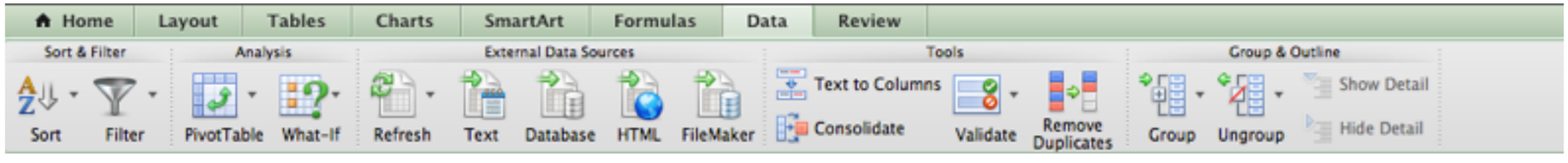
Data Analysis Tools



Data Analysis Tools

- Numeric analysis
- Network analysis
- Language analysis

Data Analysis w/ MS Excel

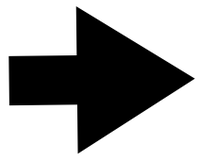


- Sort
- Basic Stats: Mean, Median, Sum...
- Filter
- Charts
- Pivot tables
- What-if
- Data Analysis Tool

Data Analysis:MS Excel ‘Pivot Tables’

A2 fx 1

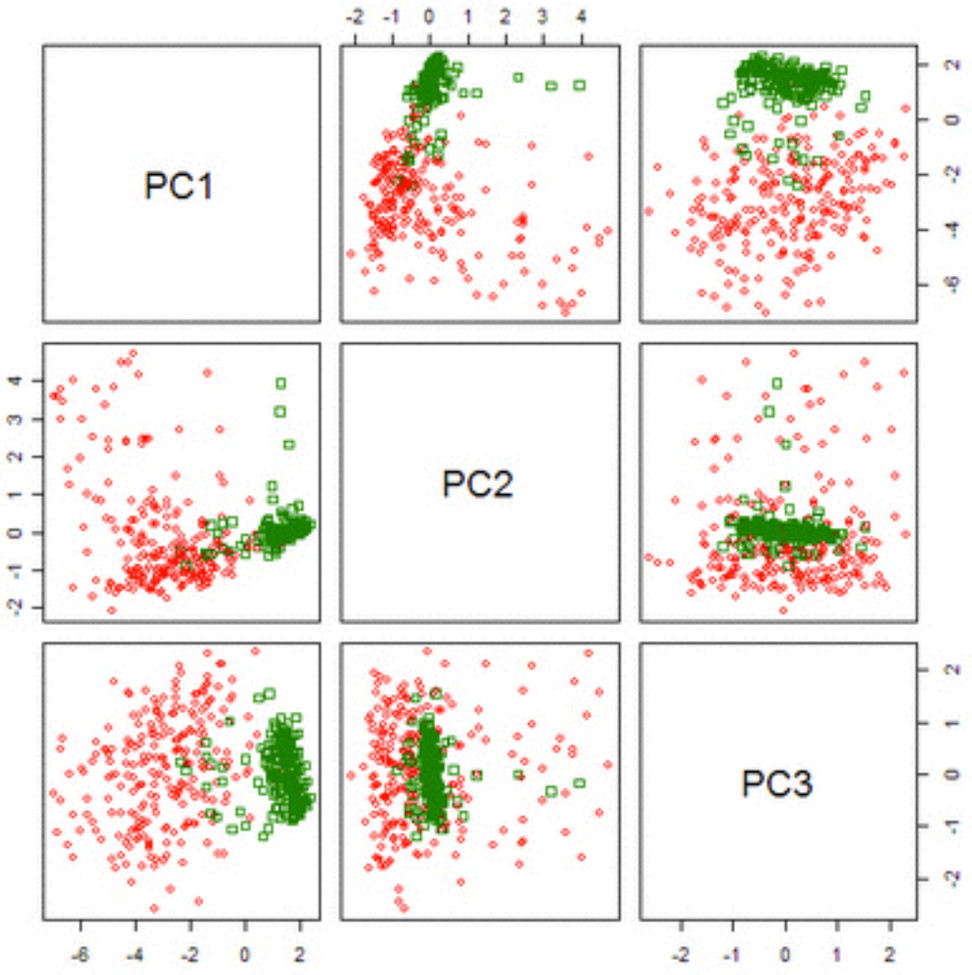
	A	B	C	D	E	F	G	H
1	Order ID	Product	Category	Amount	Date	Country		
2	1	Carrots	Vegetables	$4,270	1/6/2012	United States		
3	2	Broccoli	Vegetables	$8,239	1/7/2012	United Kingdom		
4	3	Banana	Fruit	$617	1/8/2012	United States		
5	4	Banana	Fruit	$8,384	1/10/2012	Canada		
6	5	Beans	Vegetables	$2,626	1/10/2012	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	$7,431	1/16/2012	Canada		



A3 fx Sum of Amount

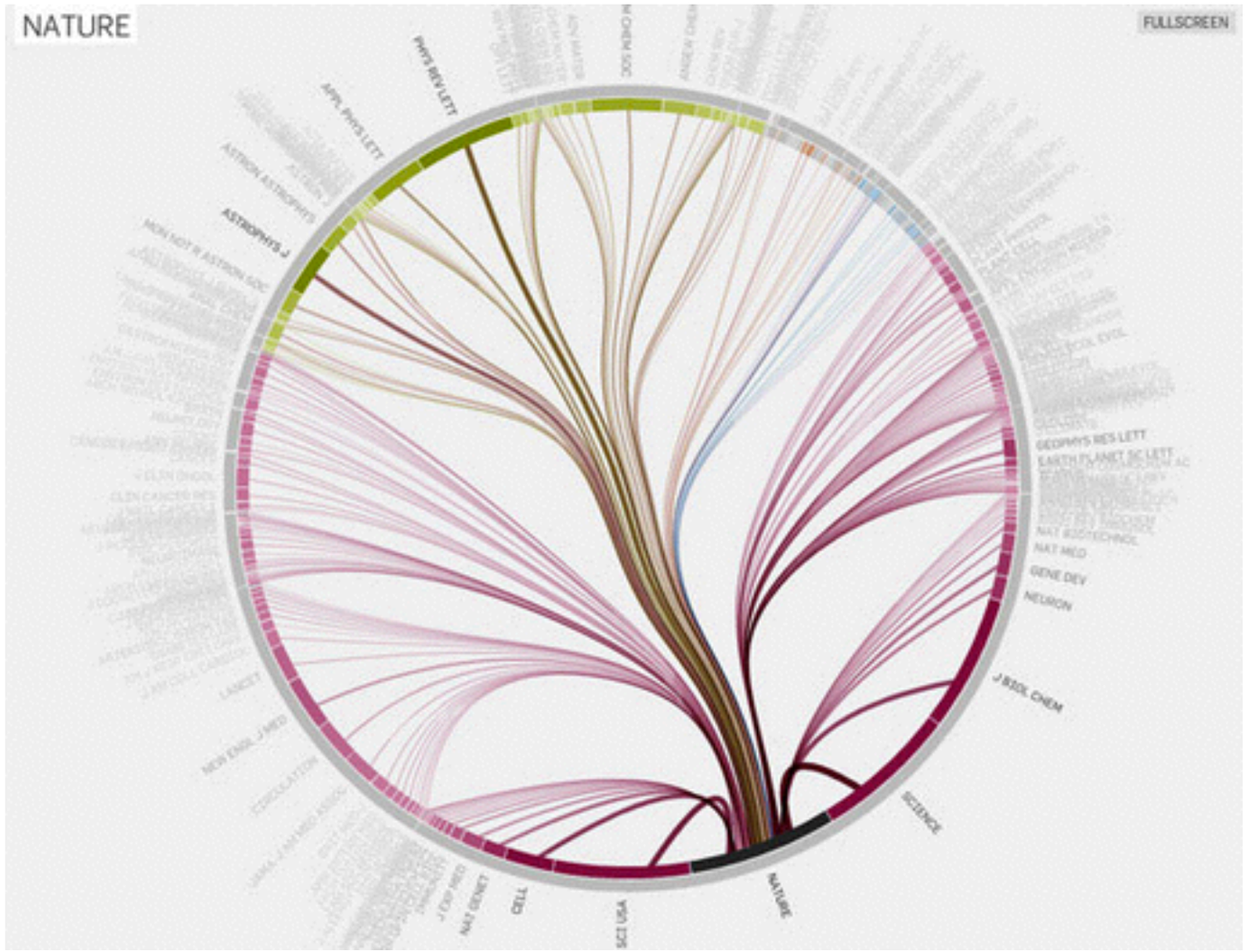
	A	B	C	D
1	Country	(All)		
2				
3	Sum of Amount			
4	Product	Total		
5	Apple	191257		
6	Banana	340295		
7	Beans	57281		
8	Broccoli	142439		
9	Carrots	136945		
10	Mango	57079		
11	Orange	104438		
12	Grand Total	1029734		

Data Analysis: R



http://www.r-project.org/

Data Visualization Tools

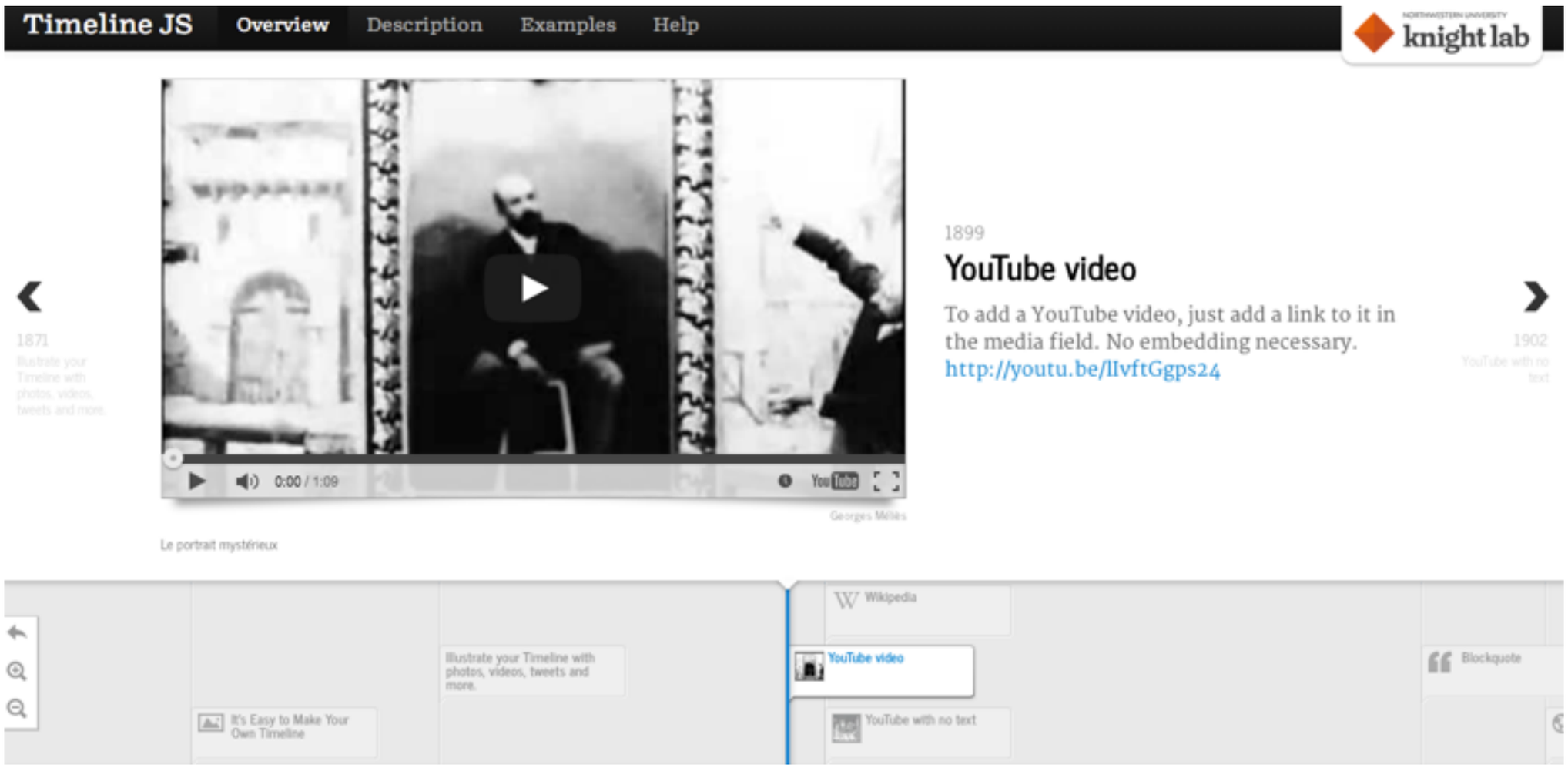


http://well-formed.eigenfactor.org/

Data Visualization Tools

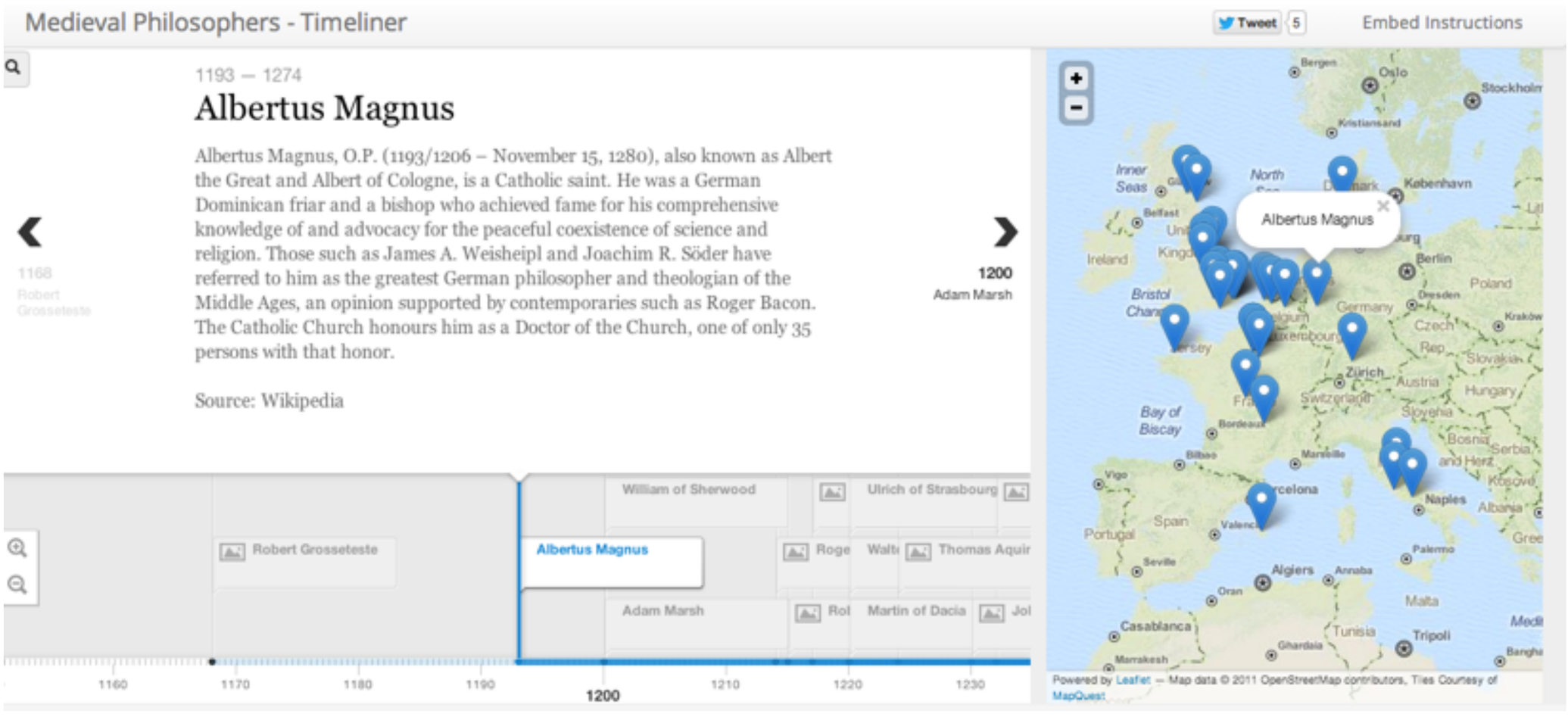
- Timelines
- Words
- Chart tools
- Networks
- Maps

Data Visualization Tools
Timeline JS



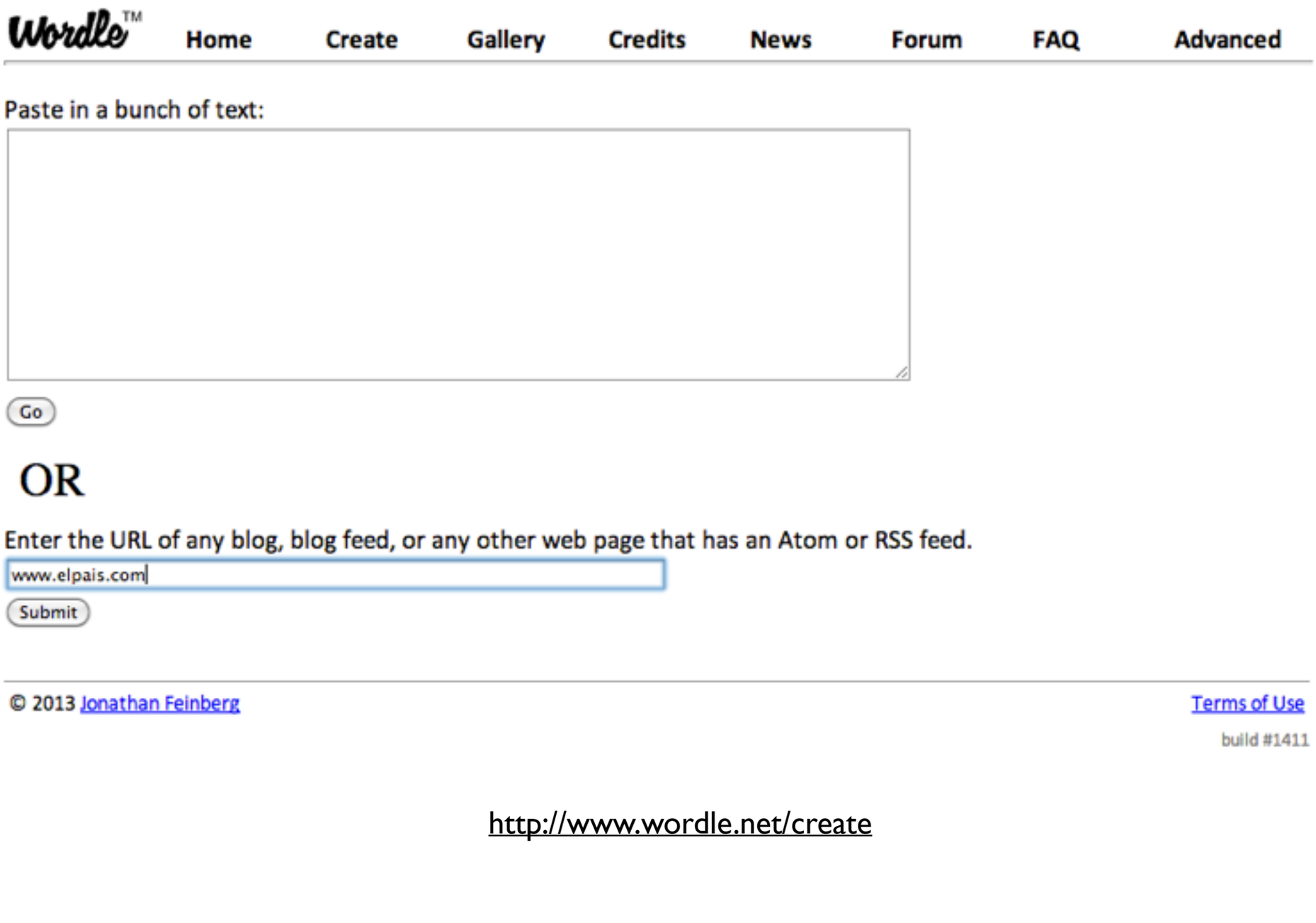
http://timeline.knightlab.com/

Data Visualization Tools
Timeliner (OKFN)



http://timeliner.okfnlabs.org

Data Visualization Tools
Wordle



http://www.wordle.net/create

Data Visualization Tools

http://www.wordle.net/create

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.

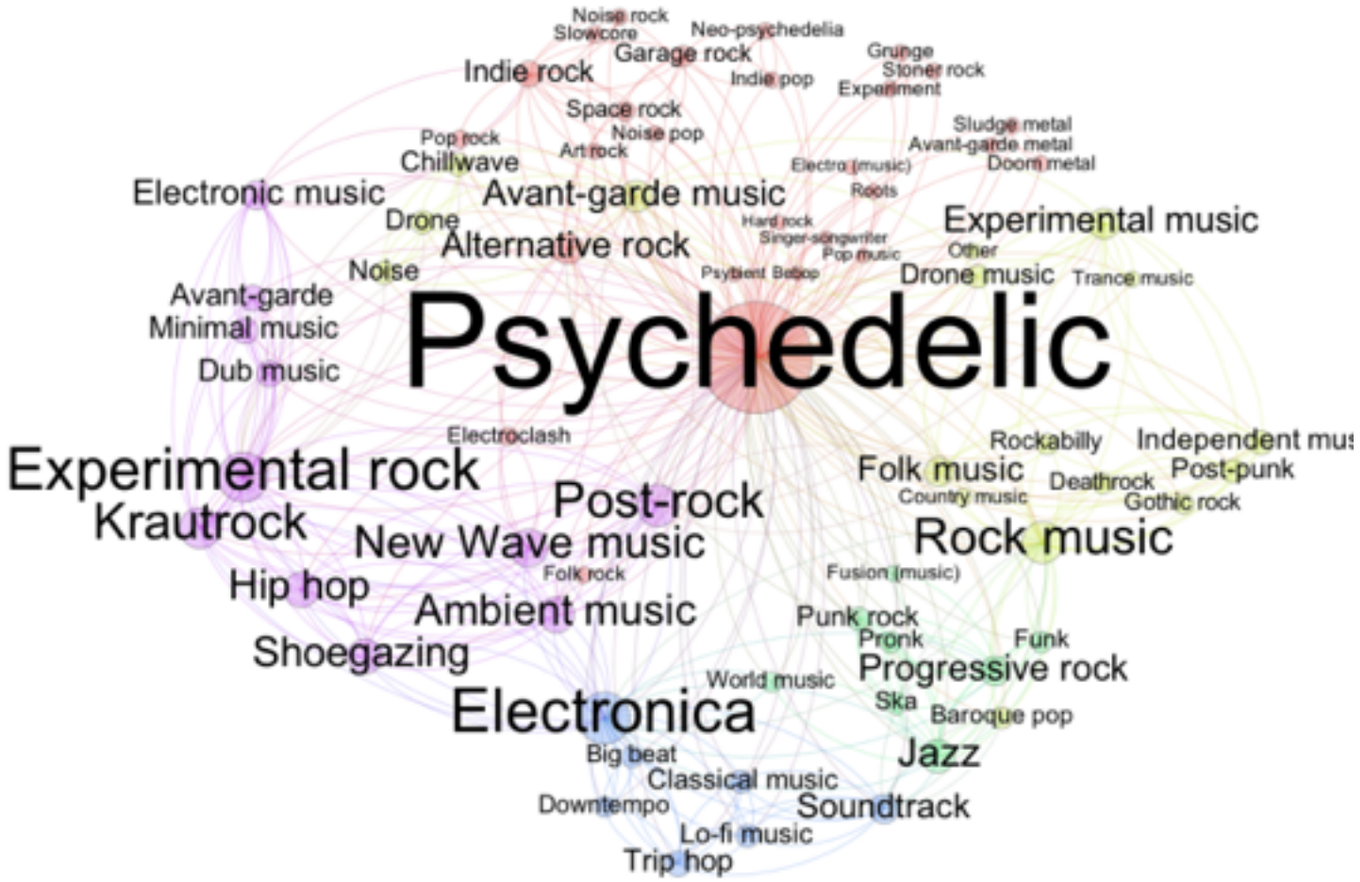
2005

http://www.quadrigram.com/

Data Visualization Tools

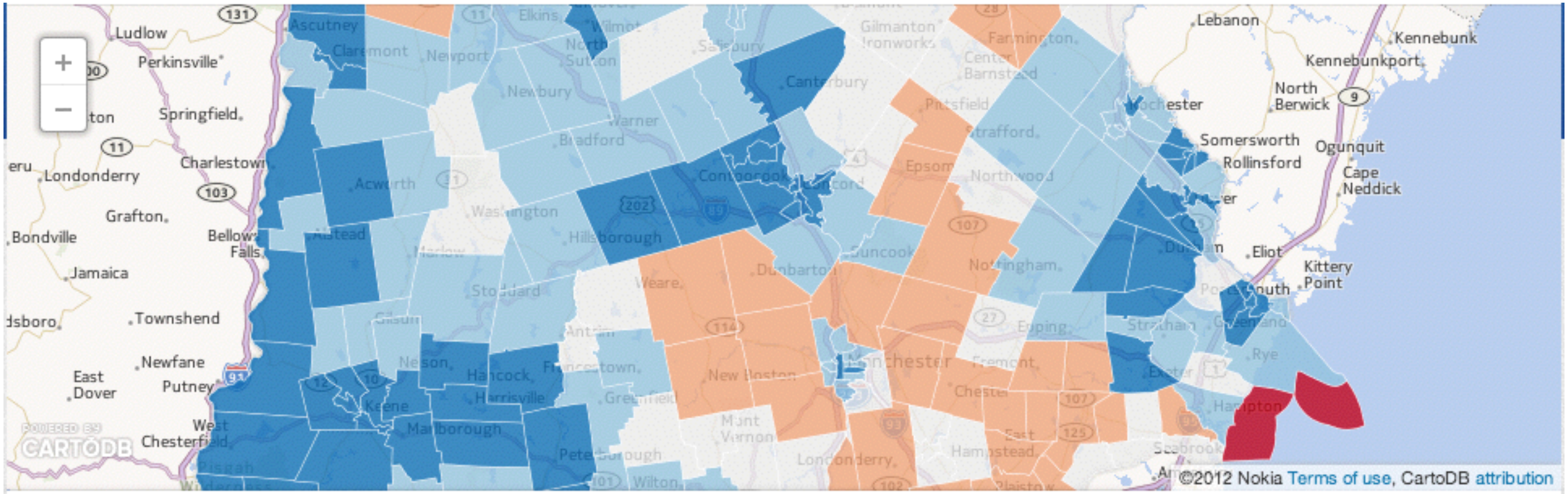
http://www.tableausoftware.com/public/

Data Visualization Tools
Gephi



http://gephi.org/

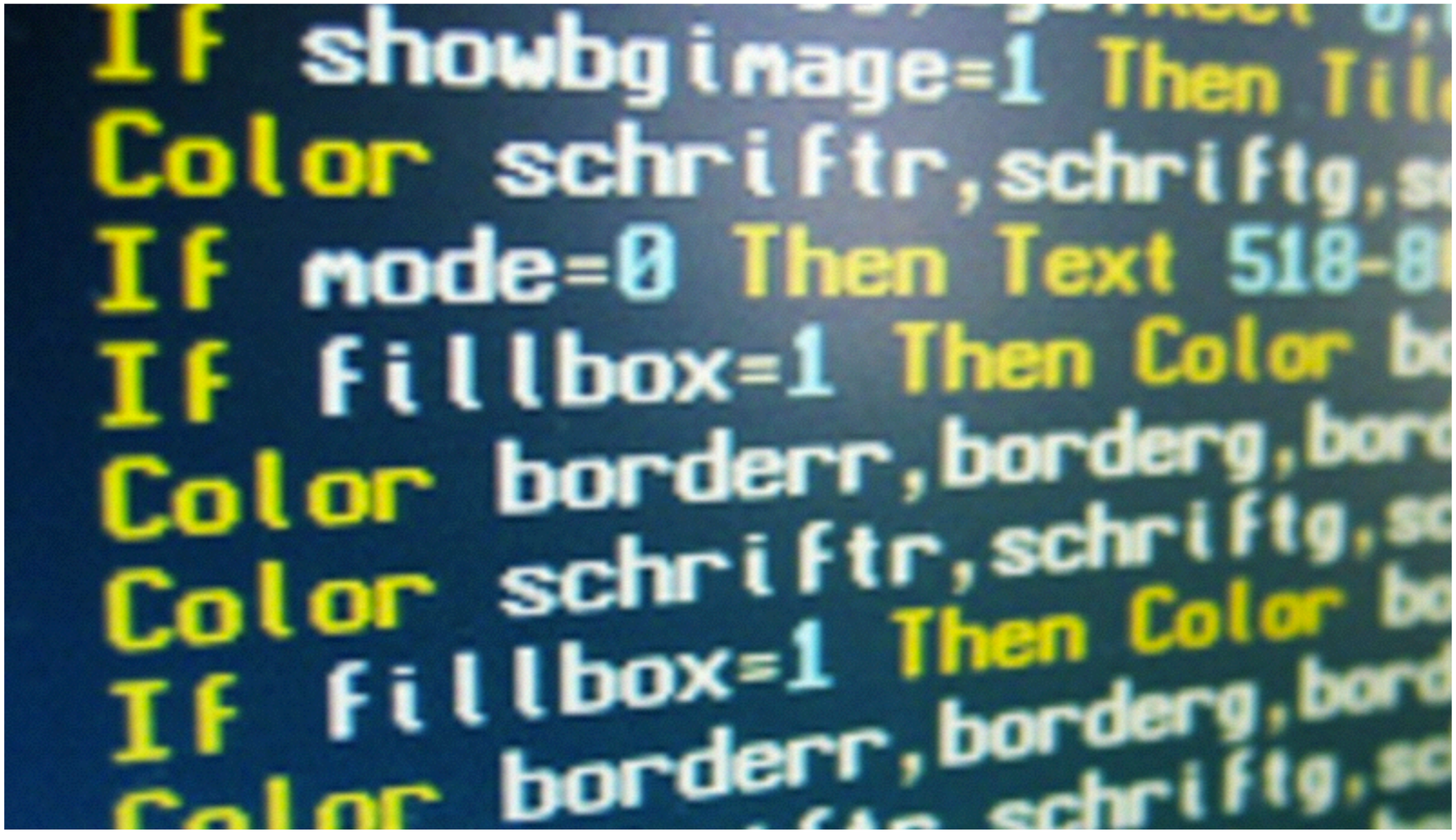
Data Visualization Tools
CartoDB



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

http://cartodb.com/

Developers needed



When you'll need a developer

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

References


matter
SYN.
ref·er·ence
referred,
person, commit
connection:
directing of atten
4. a)
work

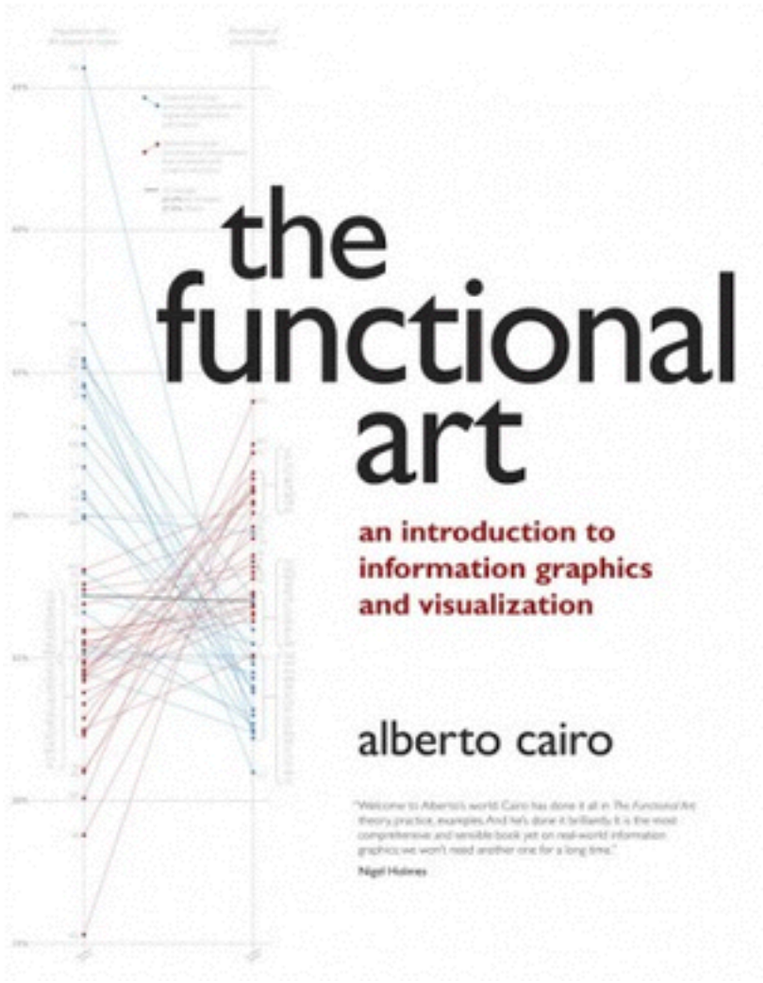
OKFN

http://okfn.org/
http://schoolofdata.org/
http://ckan.org/
http://2014.okfestival.org/

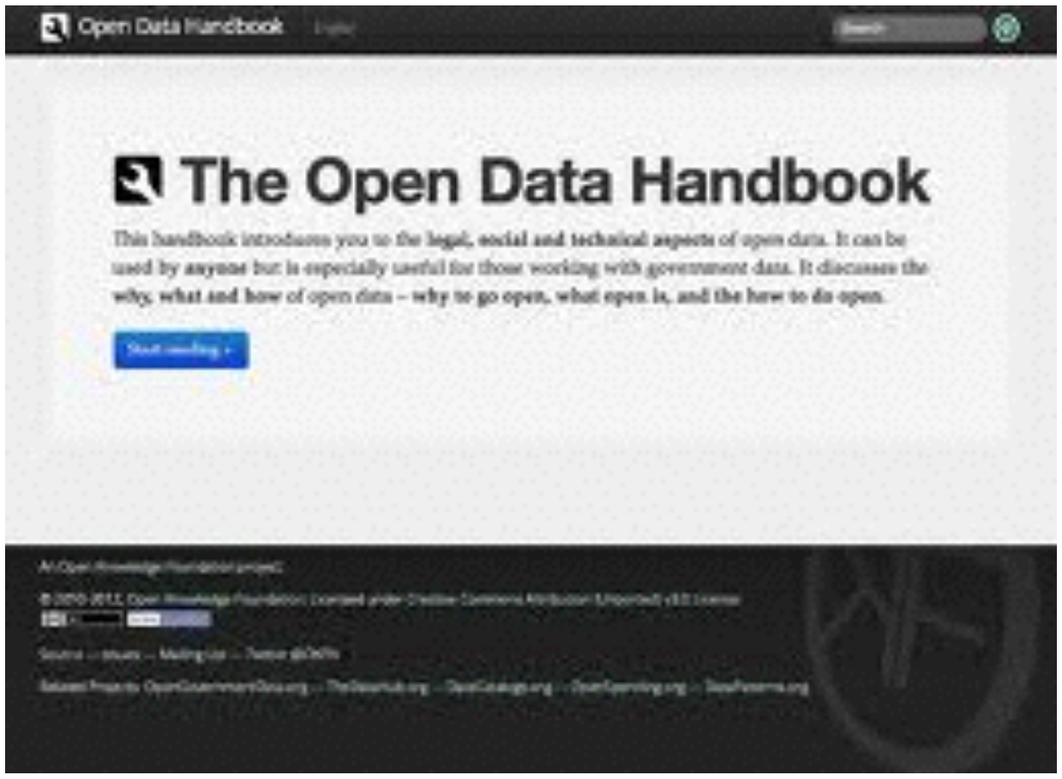
Books


The Data Journalism Handbook
How Journalists Can Use Data to Improve the News
O'REILLY
Edited by Jonathan Gray, Liliana Bounegru, and Lucy Chambers

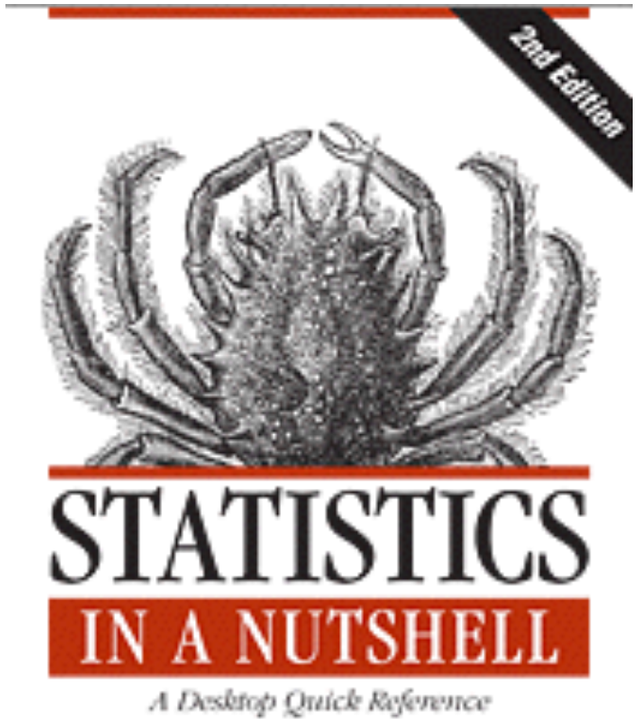
Books


the
functional
art
an introduction to
information graphics
and visualization
alberto cairo
"Welcome to Alberto's world. Cairo has done it all in The Functional Art:
theory, practice, examples. And he's done it brilliantly. It is the most
comprehensive and sensible book yet on real-world information
graphics; we won't need another one for a long time."
Nigel Holmes

Books


Open Data Handbook
The Open Data Handbook
This handbook introduces you to the legal, social and technical aspects of open data. It can be used by anyone but is especially useful for those working with government data. It discusses the why, what and how of open data – why to go open, what open is, and the how to do open.

Books


2nd Edition
STATISTICS
IN A NUTSHELL
A Desktop Quick Reference
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!

TABAKALERA