

ANÁLISIS DE REDES SOCIALES CON DATOS DE TWITTER

JORDI MORALES I GRAS

euskal
soziologia eta
zientzia
politikoaren
elkartea



asociación
vasca
de sociología
y ciencia política

CONTENIDOS DEL CURSO

SESIÓN 1

- Almacenamiento y tratamiento de datos masivos
- Nuevas herramientas de la Ciencia Social
- El Análisis de Redes Sociales (ARS): breve historia, fundamentos y ejemplos de aplicación
- Estrategias de obtención de datos tradicionales vs. computacionales
- [Captura de datos con T-hoarder](#)

SESIÓN 2

- Elementos básicos del ARS
- Algoritmos para el ARS y la visualización de datos masivos
- Estrategias de depuración de datos
- [Análisis de redes sociales con Gephi](#)
 - [Redes de menciones](#)
 - [Redes de seguidores](#)
 - [Redes semánticas](#)

SESIÓN 3

- [Análisis de redes sociales con Gephi](#)
- [Análisis de redes sociales con Pajek](#)
- Estrategias para el cruce de datos masivos
- [Cruce de datos con PowerBi](#)
- Cierre del curso y escenarios de futuro

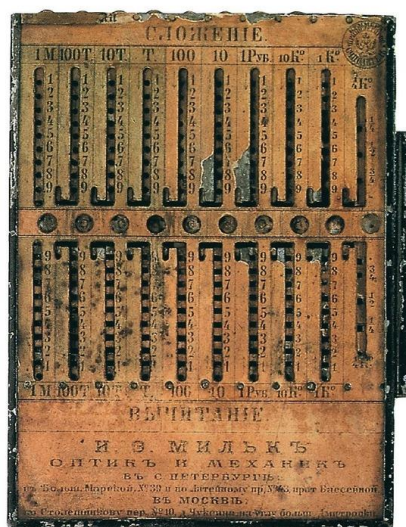
HORARIOS DEL CURSO

- Inicio: 16:00
- Descanso (breve): 17:30 – 17:35
- Finalización: 18:50

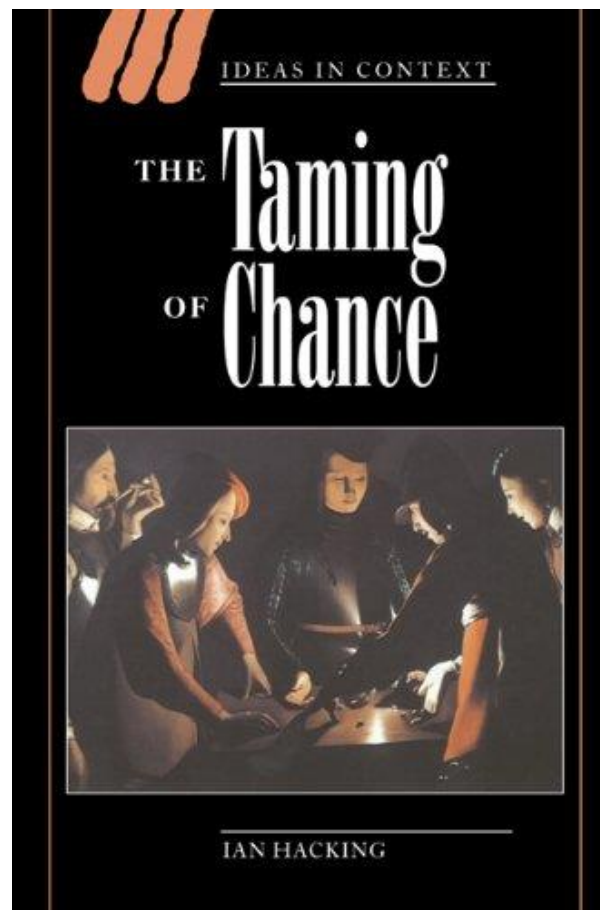
ALMACENAMIENTO Y TRATAMIENTO DE DATOS MASIVOS:
DE RETO TECNOLÓGICO A RETO INTERPRETATIVO

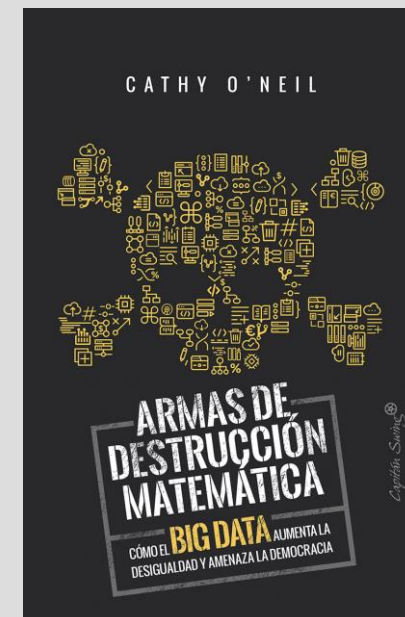
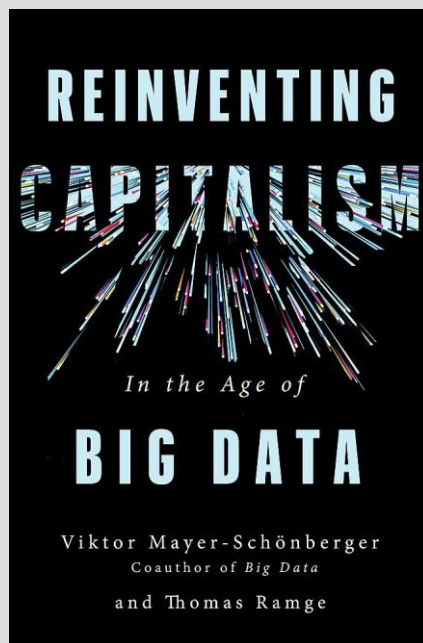
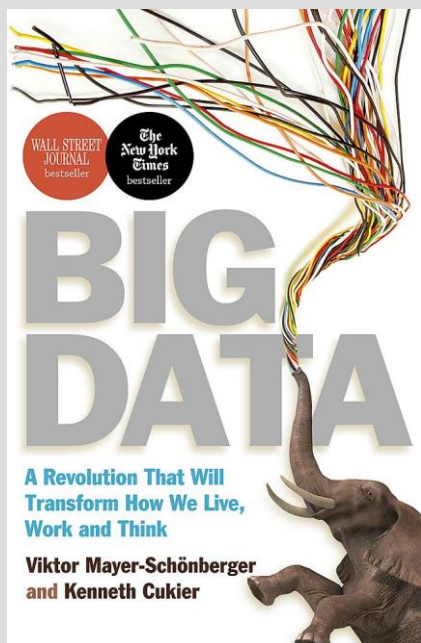
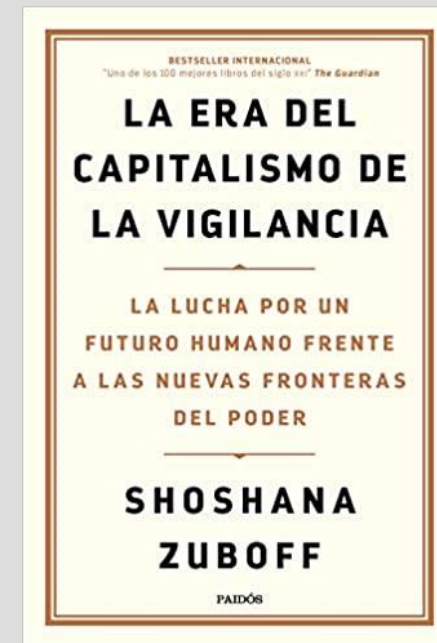
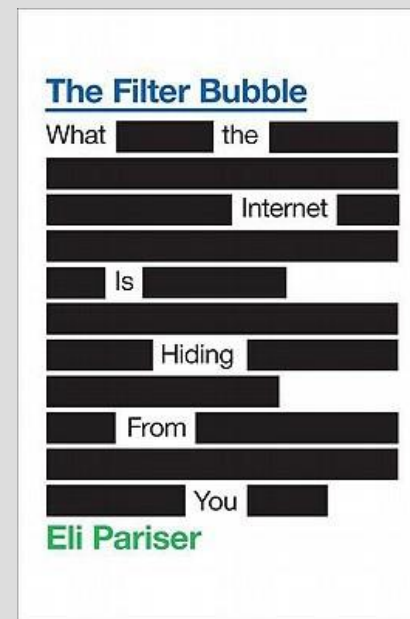
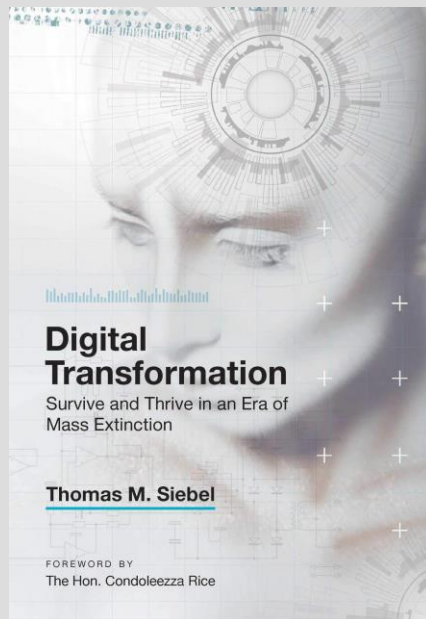
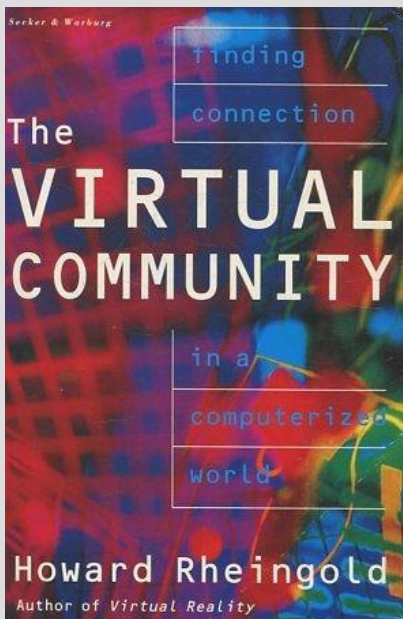
Alain Desrosières
La politique
des grands nombres

Histoire de la raison statistique



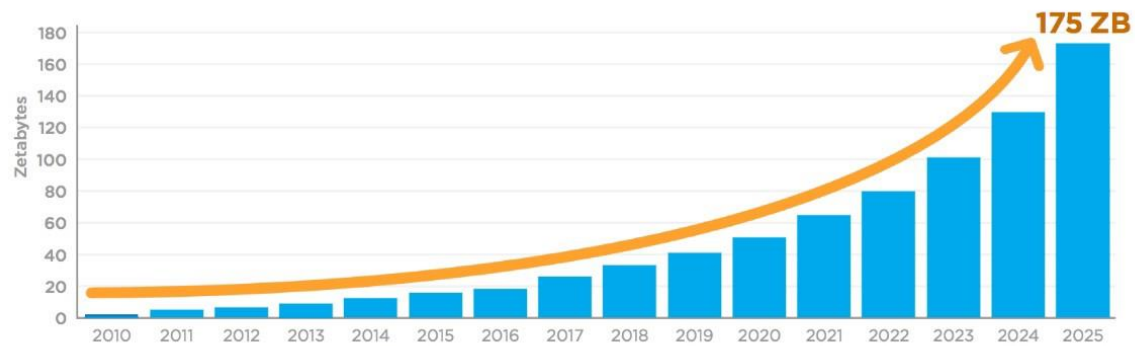
La Découverte/Poche











Source: *IDC Data Age 2025*

WHAT IS A ZETTABYTE?

1,000,000,000,000gigabyte

1,000,000,000,000terabyte

1,000,000,000,000petabyte

1,000,000,000,000exabyte

1,000,000,000,000zettabyte



Financial Services

- New Account Risk Screens
- Fraud Prevention
- Trading Risk
- Maximize Deposit Spread
- Insurance Underwriting
- Accelerate Loan Processing



Retail

- 360° View of the Customer
- Analyze Brand Sentiment
- Localized, Personalized Promotions
- Website Optimization
- Optimal Store Layout



Telecom

- Call Detail Records (CDRs)
- Infrastructure Investment
- Next Product to Buy (NPTB)
- Real-time Bandwidth Allocation
- New Product Development



Manufacturing

- Supplier Consolidation
- Supply Chain and Logistics
- Assembly Line Quality Assurance
- Proactive Maintenance
- Crowdsourced Quality Assurance



Healthcare

- Genomic data for medical trials
- Monitor patient vitals
- Reduce re-admittance rates
- Store medical research data
- Recruit cohorts for pharmaceutical trials



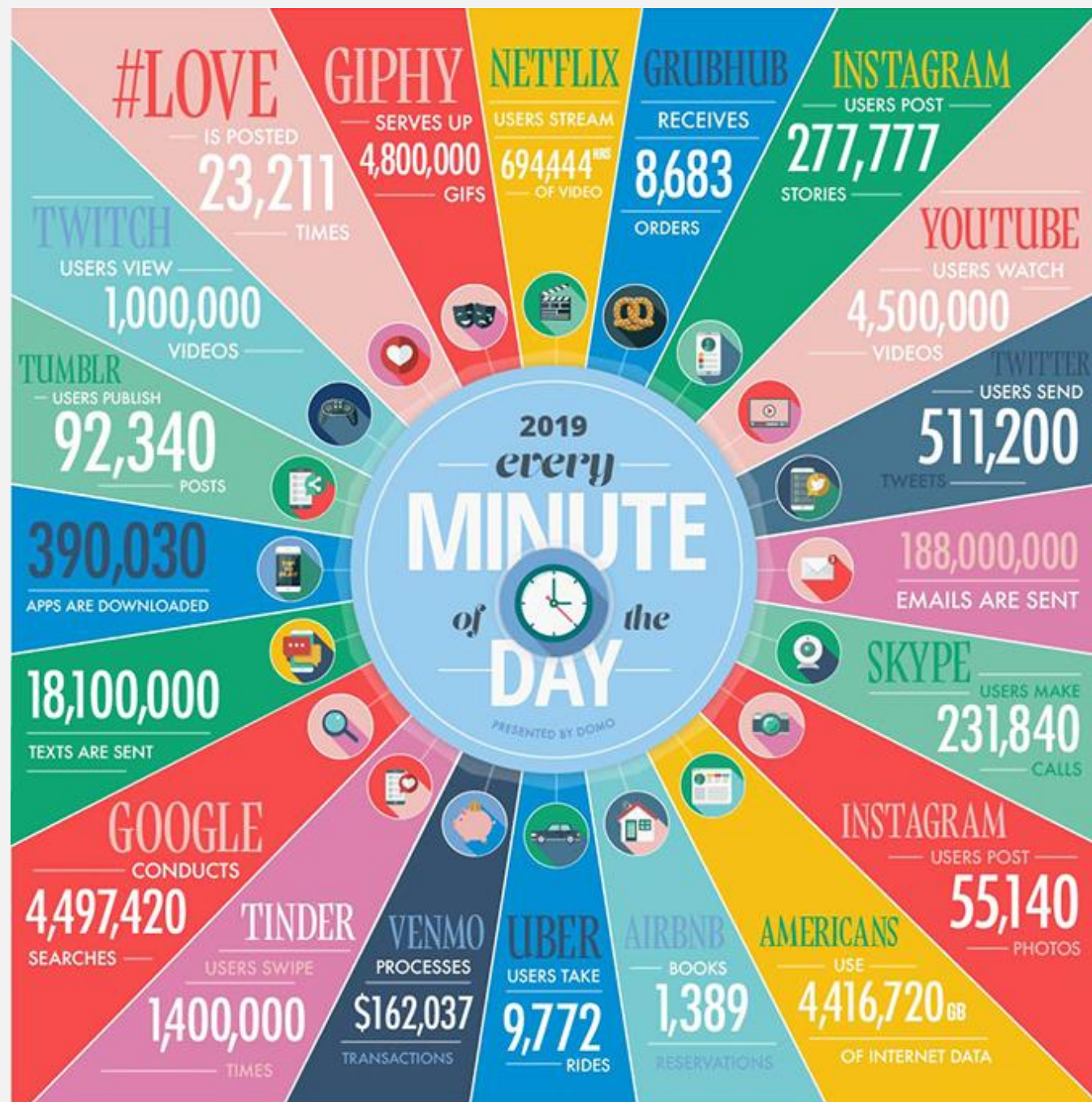
Utilities, Oil & Gas

- Smart meter stream analysis
- Slow oil well decline curves
- Optimize lease bidding
- Compliance reporting
- Proactive equipment repair
- Seismic image processing



Public Sector

- Analyze public sentiment
- Protect critical networks
- Prevent fraud and waste
- Crowdsourcing reporting for repairs to infrastructure
- Fulfill open records requests



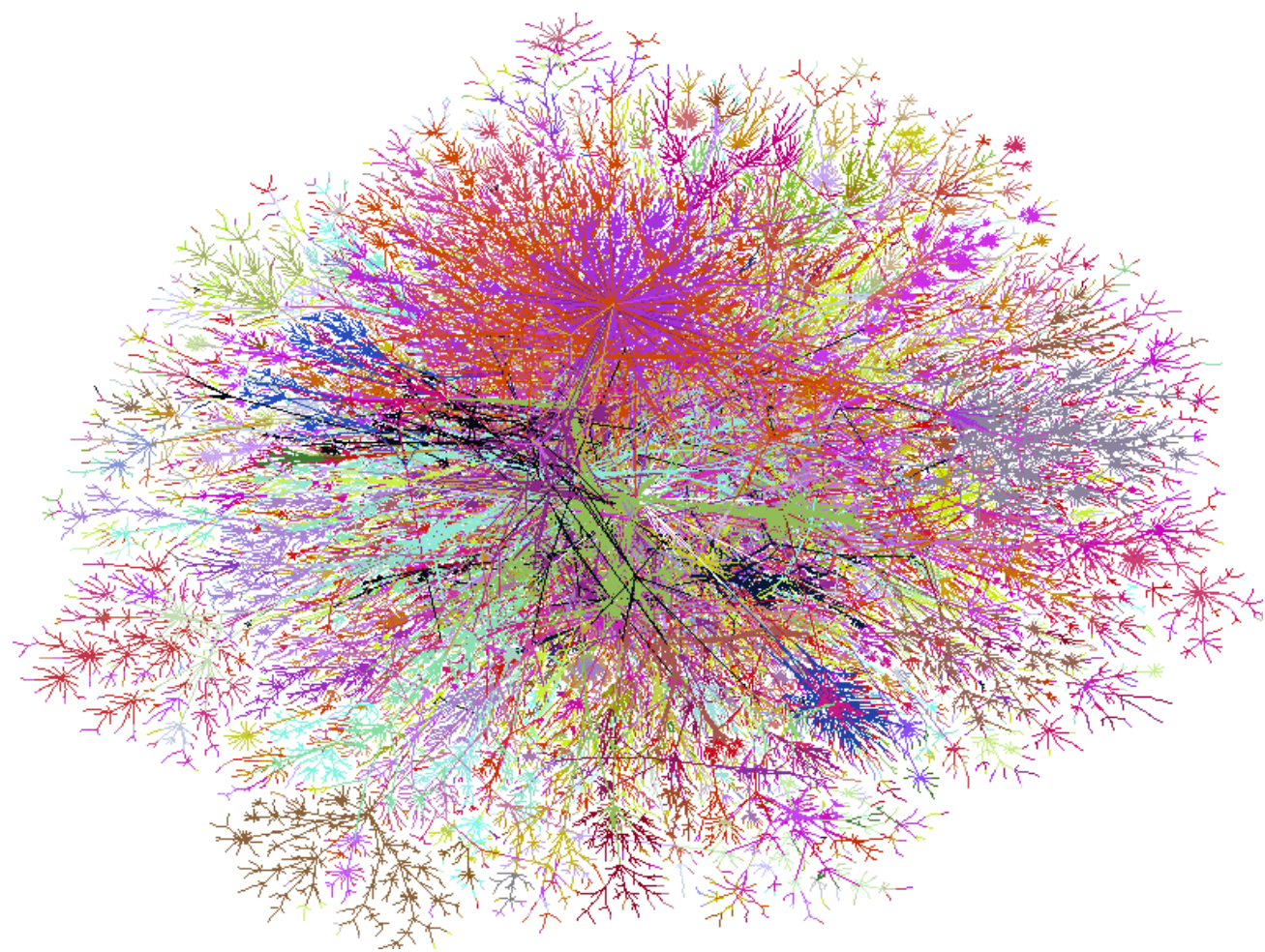
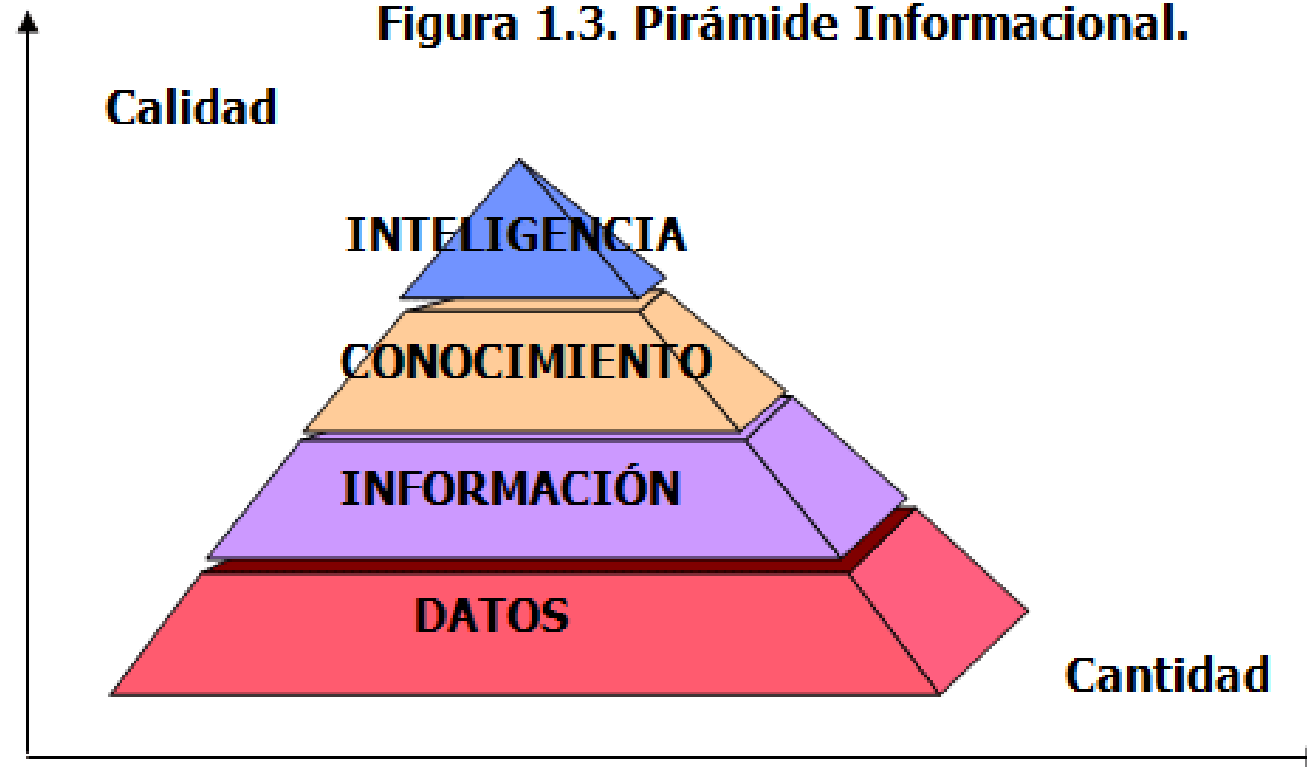
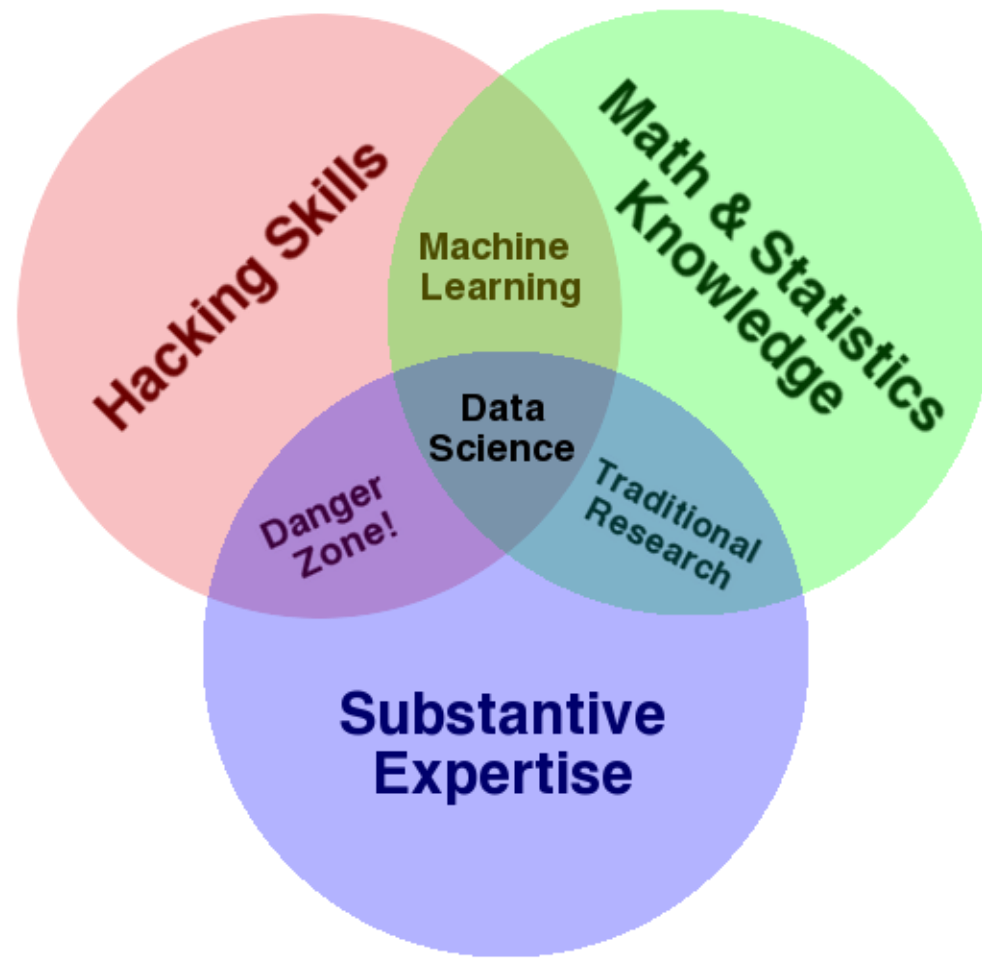


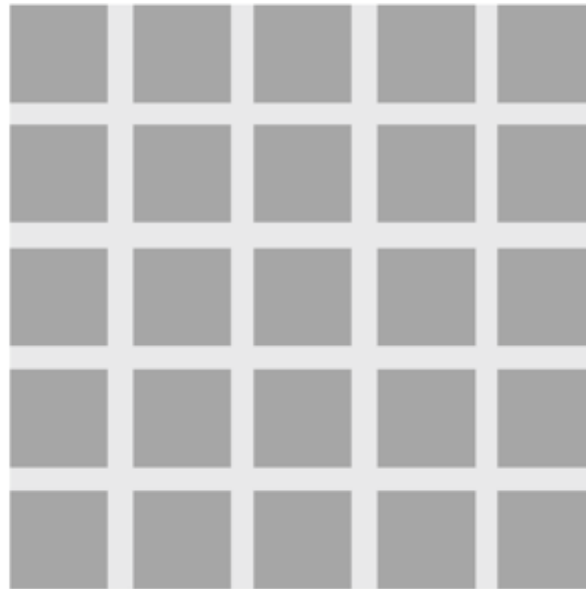
Figura 1.3. Pirámide Informacional.



Fuente: Dante, Ponjuán, Gloria. Gestión de la información en las organizaciones. Principios, conceptos y aplicaciones. Santiago de Chile, 1998.

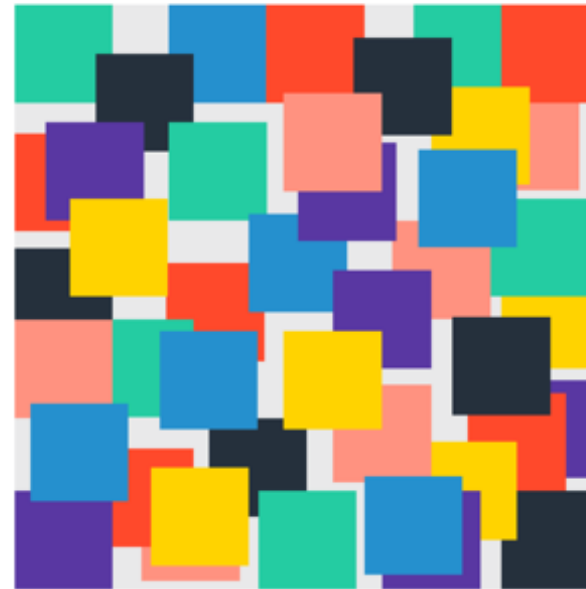


Structured data



Database, CRM, ERP

Unstructured data

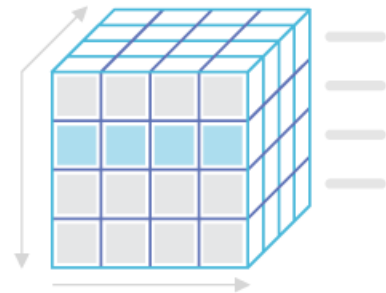
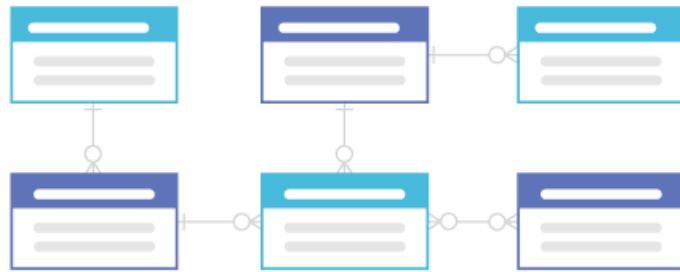


Text, audio, videos



SQL

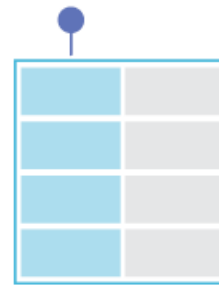
Relational Database Management Systems (RDBMS)



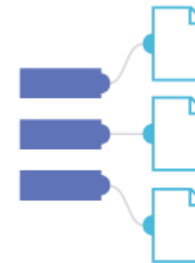
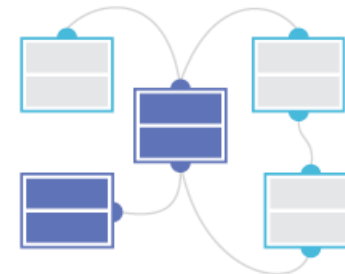
Online Analytical Processing (OLAP) Cube

NoSQL

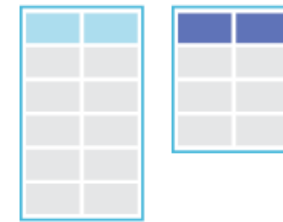
Key-Value



Graph



Document



Column store



Quantitative

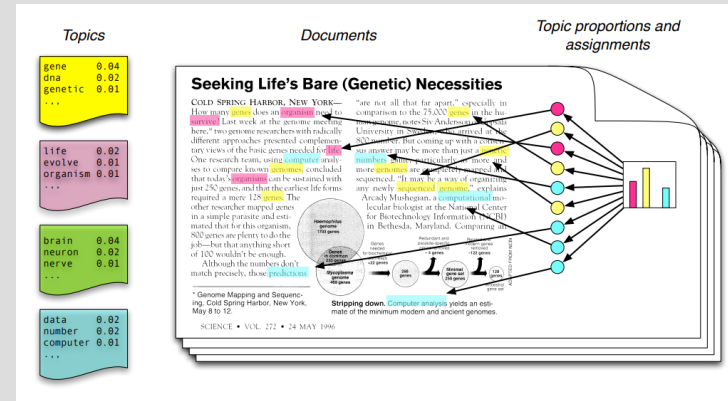
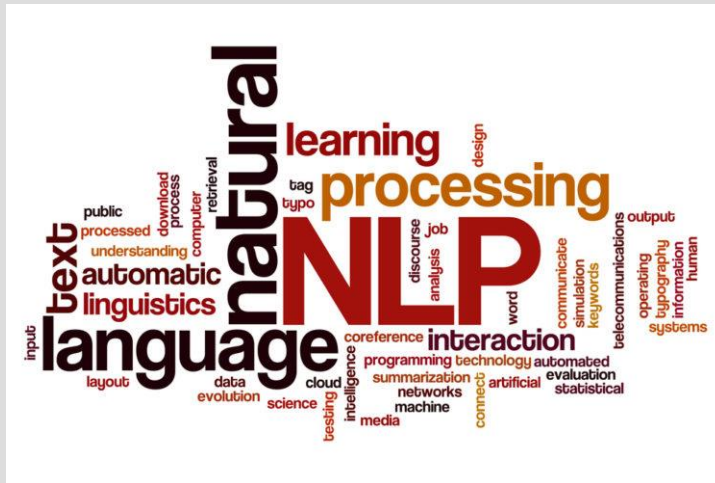


Qualitative

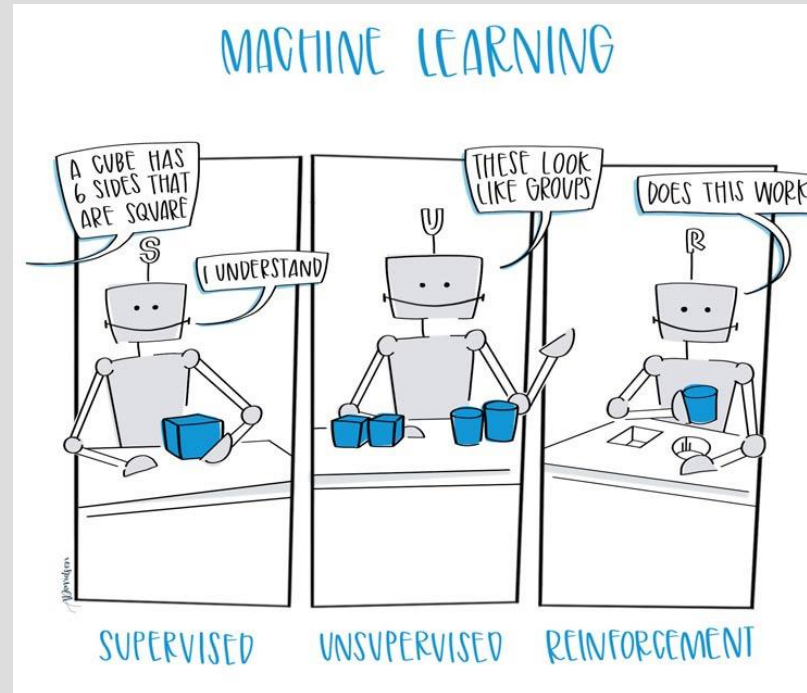
NUEVAS HERRAMIENTAS PARA LA CIENCIA SOCIAL

Tokenize on rules	Let	's	tokenize	!	Is	n't	this	easy	?		
Tokenize on punctuation	Let	'	s	tokenize	!	Isn	'	t	this	easy	?
Tokenize on white spaces	Let's	tokenize!	Isn't	this	easy?						

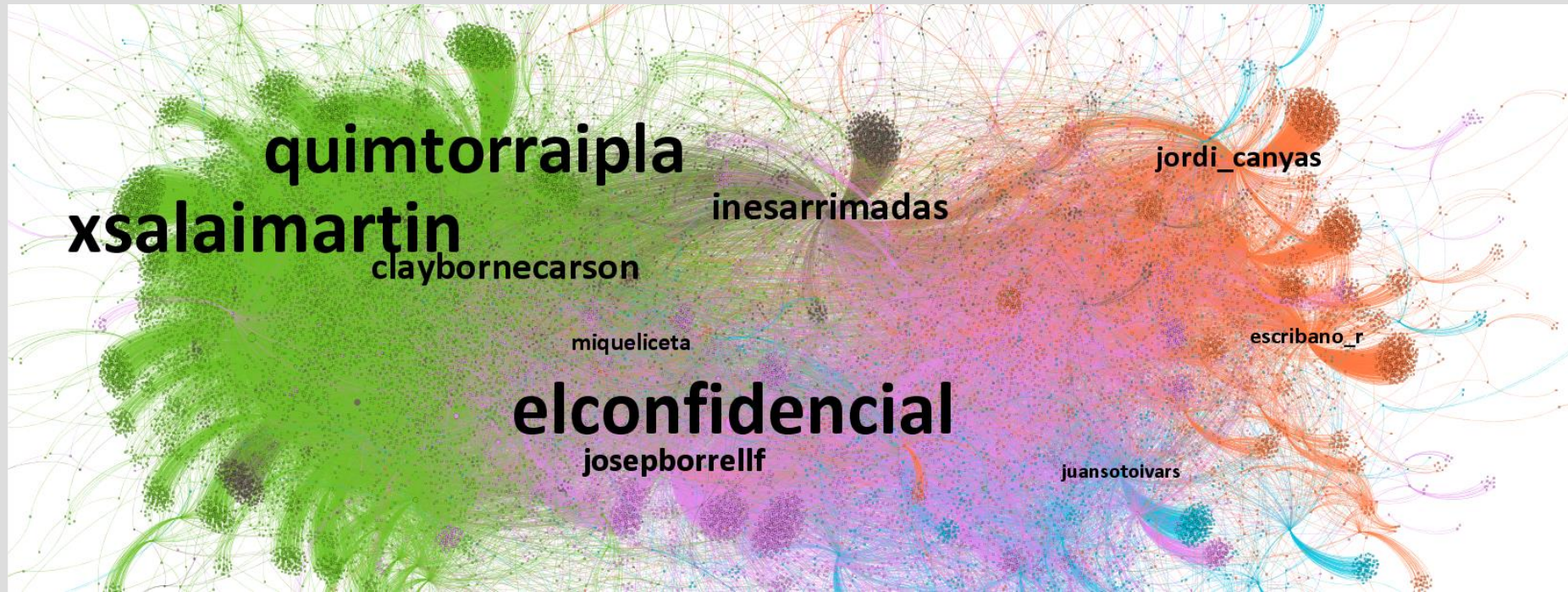
Let's tokenize! Isn't this easy?



EL PROCESAMIENTO DEL LENGUAJE NATURAL



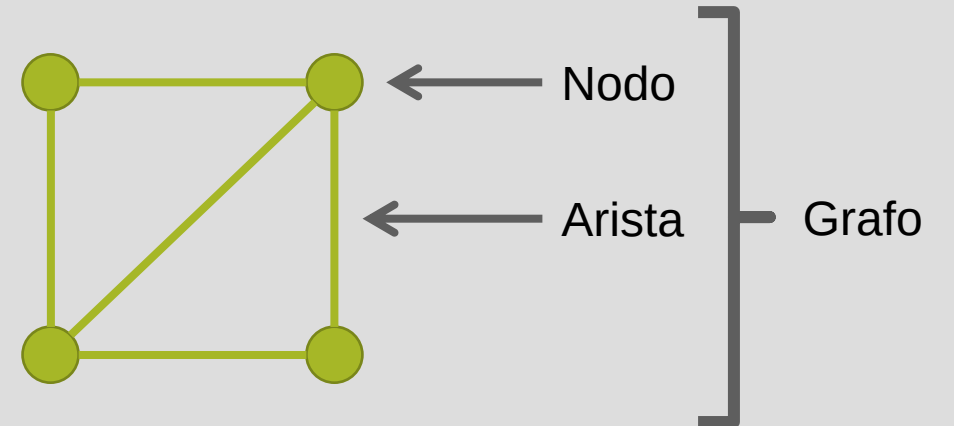
EL APRENDIZAJE AUTOMÁTICO O MACHINE LEARNING (ML)



EL ANÁLISIS DE REDES SOCIALES

EL ANÁLISIS DE REDES SOCIALES (ARS), BREVE HISTORIA,
FUNDAMENTOS Y EJEMPLOS DE APLICACIÓN.

El **ABC** del anàlisis de redes



Matrices, adyacencias, aristas y grafos

	1	2	3	4	5
1	0	1	0	0	1
2	1	0	1	1	1
3	0	1	0	1	0
4	0	1	1	0	1
5	1	1	0	1	0

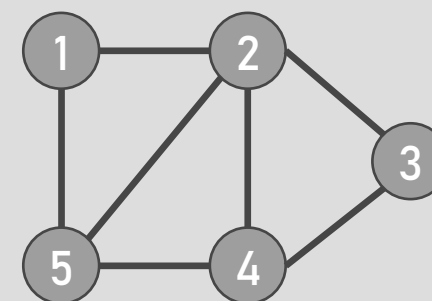
Matriz de
adyacencias

1	2	5	
2	3	4	5
3	4		
4	5		

Listado de
adyacencias

$V = \{1, 2, 3, 4, 5\}$
 $E: \{ (1, 2), (1, 5), (2, 3),$
 $(2, 4), (2, 5), (3, 4),$
 $(4, 5) \}$

Conjunto de
aristas

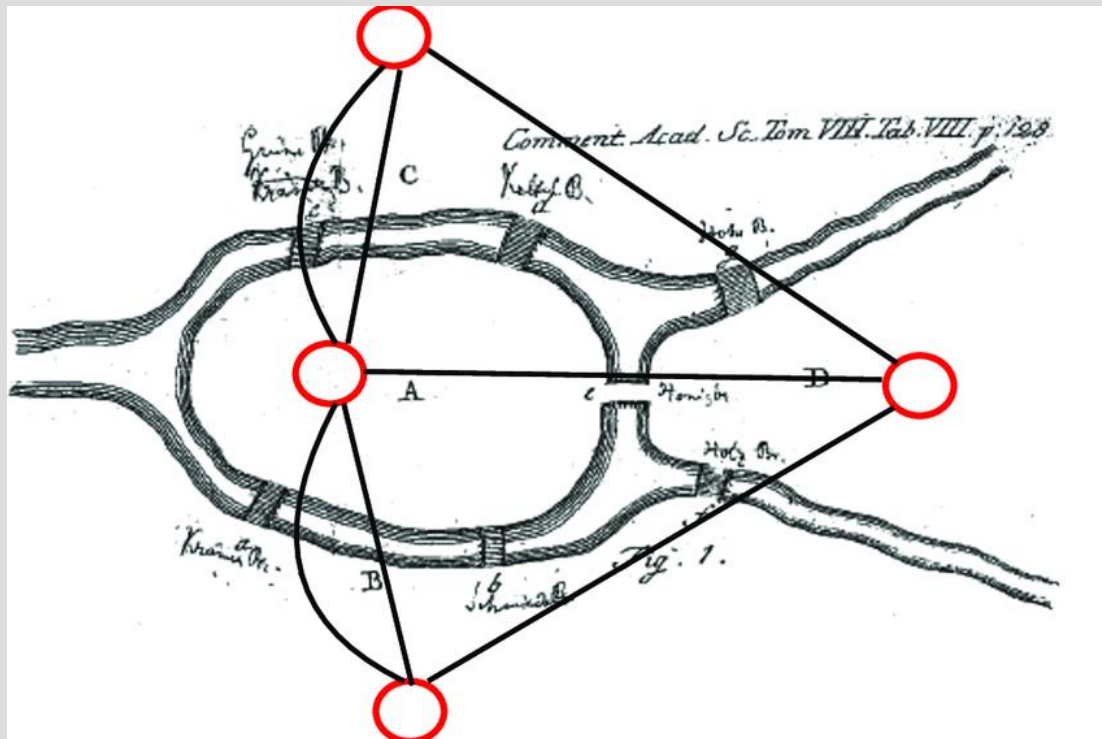
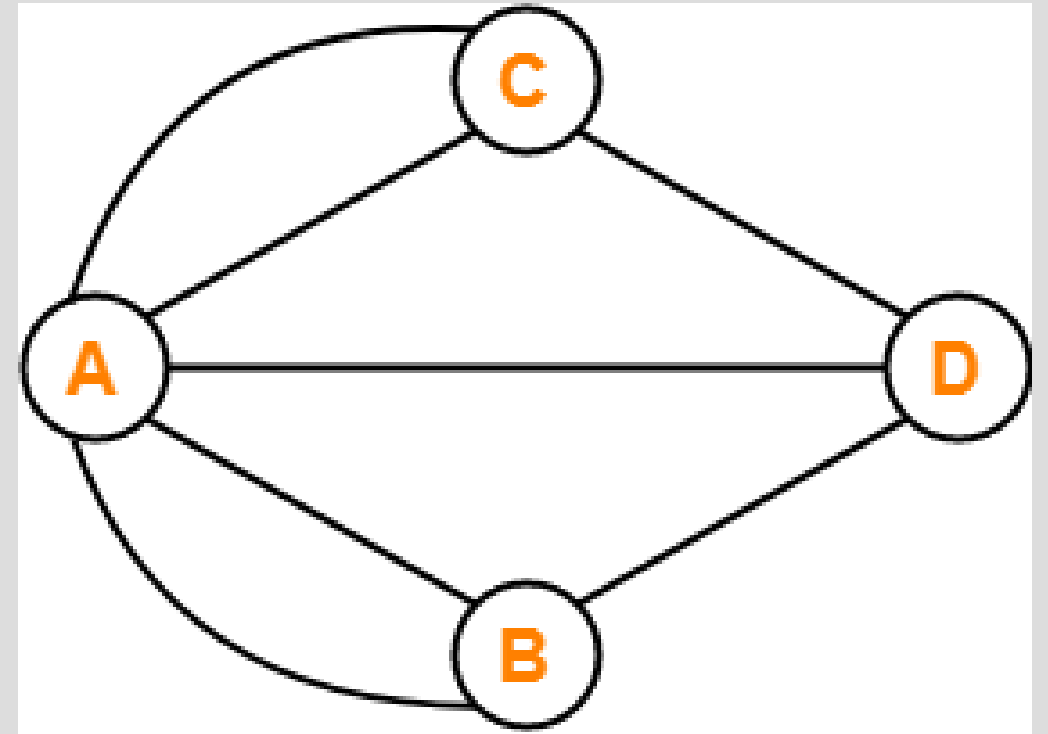
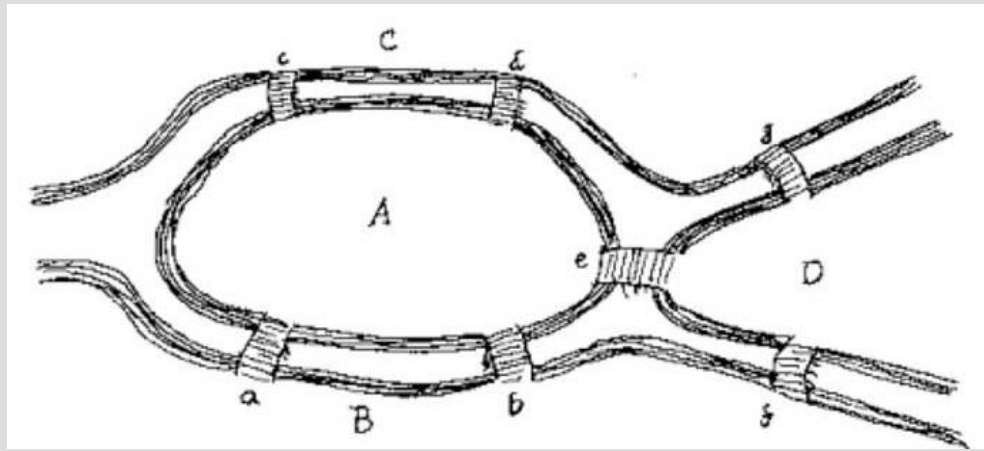


Grafo

La teoria de grafos



Euler



LA MIRADA RELACIONAL EN CIENCIA SOCIAL (I)

“Cada sistema debe estar compuesto por elementos de la misma naturaleza que el sistema mismo, el espíritu científico no nos permite observar a la sociedad como compuesta por individuos. El límite social verdadero es la familia—reducida, si fuera necesario, a su forma elemental, la pareja”

Comte, 1853

“Un grupo de seres humanos no deviene en sociedad porque cada uno de ellos tenga un objetivo determinado o un contenido vital subjetivo. Se convierte en una sociedad solo cuando la vitalidad de esos contenidos adquiere la forma de influencia recíproca; solo cuando un individuo tiene un efecto, inmediato o mediato, sobre otro es cuando una mera agregación espacial o sucesión temporal se transforma en sociedad. Si, por tanto, ha de existir una ciencia cuyo objeto sea la sociedad y nada más, deberá investigar de manera exclusiva esas interacciones, tipos y formas de asociación”

Simmel, 1908

LA MIRADA RELACIONAL EN CIENCIA SOCIAL (II)

“Algunas investigaciones sociales han centrado su atención de **manera consistente en las relaciones sociales que ligan a individuos, más que en los individuos mismos** (...) Los analistas de redes buscan descubrir varios tipos de entramado para tratar de determinar las condiciones bajo las cuales éstos emergen y descubrir sus consecuencias. **El enfoque estructural no se limita al estudio de relaciones sociales humanas y está presente en casi todos los campos de la ciencia.** Por ejemplo, los astrofísicos estudian la fuerza de atracción mutua de los planetas del sistema solar para explicar sus órbitas; los químicos moleculares examinan la manera en que diferentes tipos de átomos interactúan para formar distintos tipos de moléculas; los ingenieros eléctricos observan cómo las interacciones entre varios componentes electrónicos—como condensadores y resistencias—afectan el flujo de corriente en un circuito, y los biólogos estudian las maneras en que cada especie de un ecosistema interactúa con otras y las afecta.”

Freeman, 2012

LA MIRADA RELACIONAL EN CIENCIA SOCIAL (III)

30s – 60s

- Moreno, Jennings y Lazarsfeld
- Warner et al. (Harvard)
- Lewin, Bavelas, Cartwright, Festinger, Harary y Katz (MIT-Iowa-Michigan)
- Strauss y Weil (Sorbonne)
- Rashevsky (Chicago)
- Radcliffe-Brown, Bott y Barnes (LSE)
- Merton, Lazarsfeld y Coleman (Columbia)

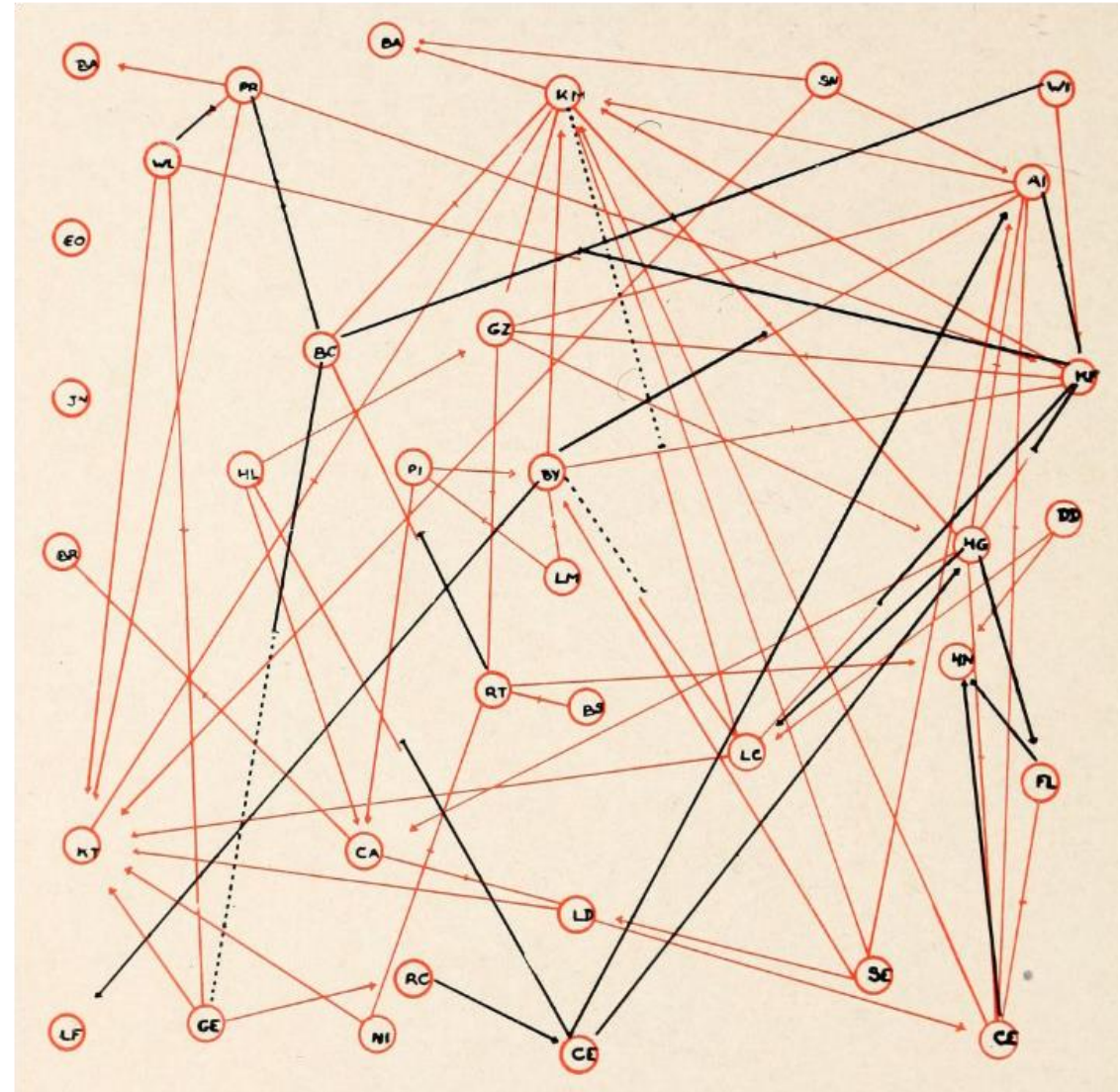
60s – 90s

- Kadushin y Blau (Columbia-Chicago)
- Deutsch, de Sola Pool y Kochen (MIT)
- Milgram (Stanford)
- White (MIT-Harvard-Columbia)
- Mokken, Anthonisse y Stokman (Amsterdam)
- Freeman – Sunshine (Syracuse)
- Granovetter (Harvard)
- Burt (Chicago)

90s – 2019

- White, Wellman, Azarian, Boorman, Breiger, Borgatti, Everett, Wasserman, Faust
- Requena Santos, Pizarro, Lozares, Molina, Maya Jariego
- Emirbayer, Goodwin, Mische, Tilly, Fluse
- Adamic, Brandes, De Nooy, Batagelj, Mrvar, Rotta, Noack, Yifan Hu, Jackomy...

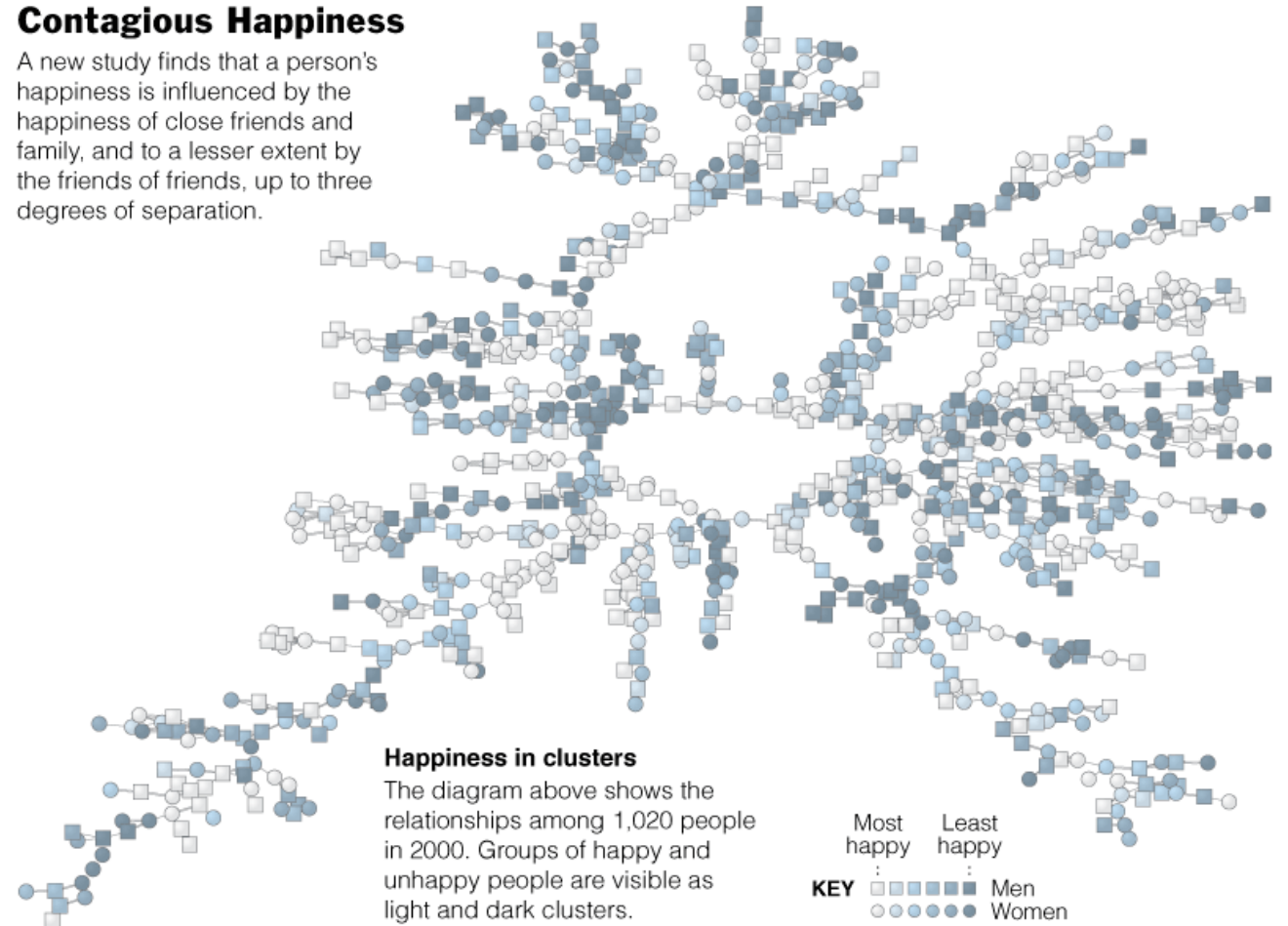
E.G. ATRACCIÓN Y
RECHAZO
(MORENO Y
JENNINGS, 1934)



E.G. LA DIFUSIÓN
DE LA FELICIDAD
(FOWLER Y
CHRISTAKIS, 2008)

Contagious Happiness

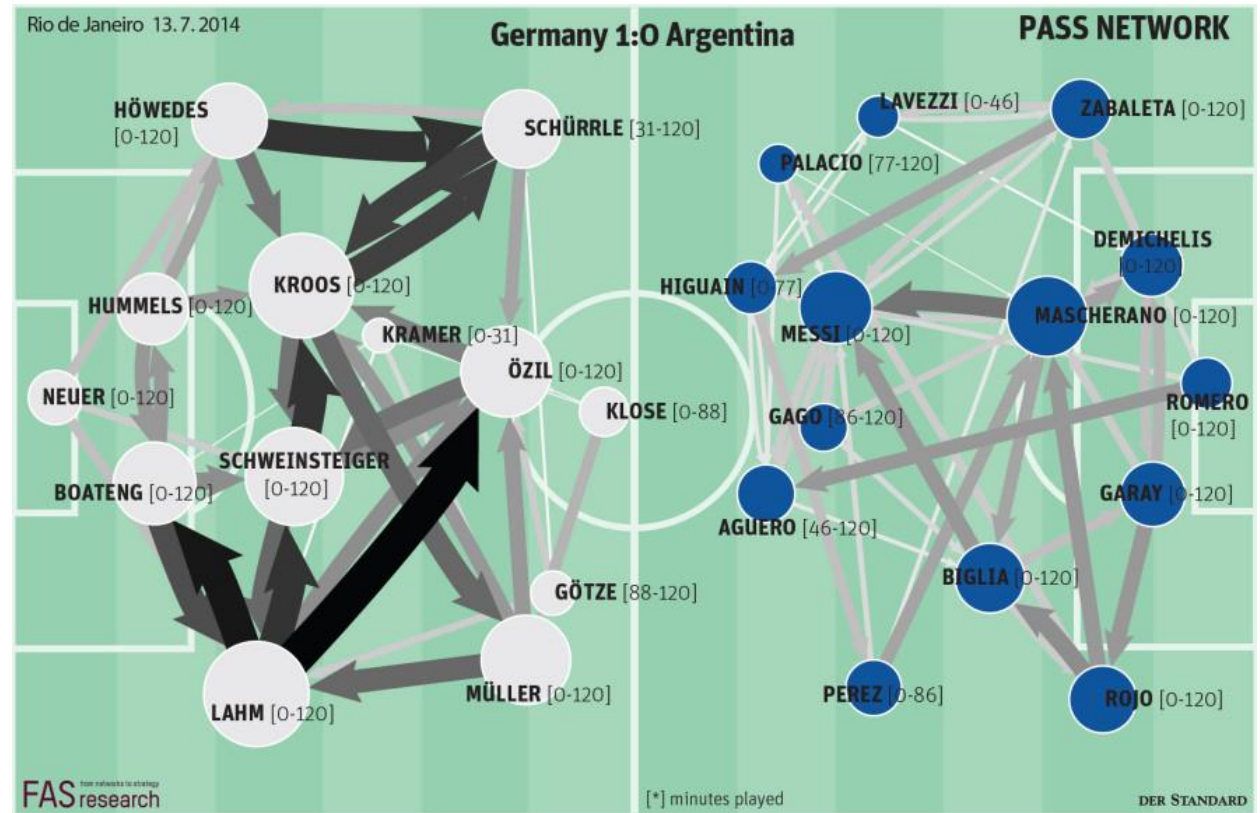
A new study finds that a person's happiness is influenced by the happiness of close friends and family, and to a lesser extent by the friends of friends, up to three degrees of separation.



Sources: James H. Fowler; Nicholas A. Christakis; BMJ

THE NEW YORK TIMES

E.G. PASES EN
PARTIDO DE
FUTBOL
(FAS RESEARCH,
2014)

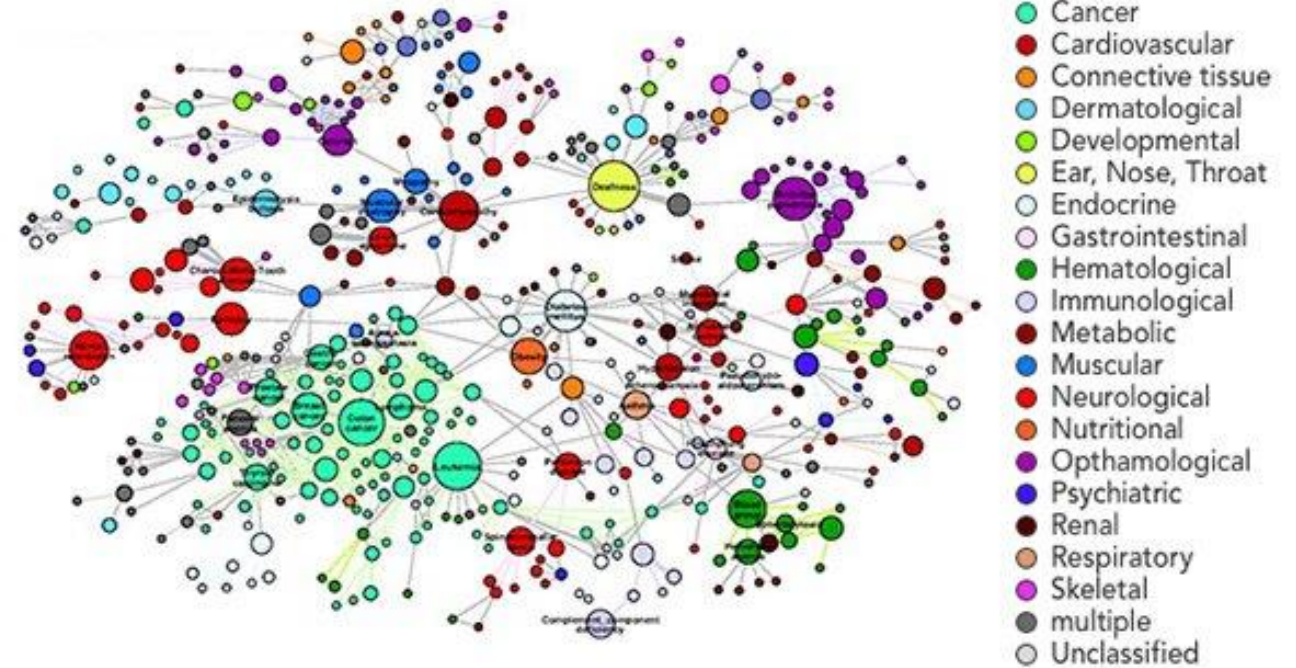


E.G. EL PADRINO II (MOVIEGALAXIES)



E.G.
ENFERMEDADES
HUMANAS
(THURNER, 2015)

Human Disease Network



E.G. MIGRACIONES
INTERNAS EN
CHINA
(BAIDU, 2015)



Rio De Janeiro

© 2014 peplemaps.org / @davetroy

E.G. LOS
INTERESES DE
L@S
CARIOCAS
(TROY Y
PEREIRA, 2014)



Edited by Marconi Pereira

ARS Y TWITTER: CASOS DE ESTUDIO PARA LAS
CIENCIAS SOCIALES.

CASO DE ANÁLISIS I LUTHER KING Y LA INDEPENDENCIA DE CATALUÑA



El Confidencial
@elconfidencial

Seguir

El Instituto Luther King de EEUU pide que Torra deje de usar su figura: "Es hipócrita"



El Confidencial

El Instituto Luther King de EEUU pide que Torra deje de usar su figura:
El director del centro de estudios se desmarca de las arengas del secesionismo catalán, que ha llegado a planear una marcha por los derechos civiles a semejan...

elconfidencial.com

22:44 - 13 sept. 2018

1.065 Retweets 1.389 Me gusta



177 1,1K 1,4K



Josep Borrell Fontelles
@JosepBorrellF

Seguir

Comparto con vosotros por su interés este artículo de @elconfidencial
El Instituto Luther King de EEUU pide que Torra deje de usar su figura: "Es hipócrita"



El Confidencial

El Instituto Luther King de EEUU pide que Torra deje de usar su figura:
El director del centro de estudios se desmarca de las arengas del secesionismo catalán, que ha llegado a planear una marcha por los derechos civiles a semejan...

elconfidencial.com

7:40 - 14 sept. 2018

1.569 Retweets 2.772 Me gusta



1,5K 1,6K 2,8K



Inés Arrimadas
@InesArrimadas

Seguir

El Instituto Luther King corrige al separatismo y pide que dejen de utilizar su nombre 🙅

"No es justo que usen su figura. No veo que formar parte de España sea una opresión ni que nadie les impida ejercer sus derechos"

El Confidencial @elconfidencial

El Instituto Luther King de EEUU pide que Torra deje de usar su figura: "Es hipócrita"
elconfidencial.com/espana/2018-09...

2:12 - 14 sept. 2018

2.226 Retweets 3.971 Me gusta



1,9K 2,2K 4,0K

I. EL IMPACTO INICIAL



Clayborne Carson
@ClayborneCarson

Seguir

On Martin Luther King, Jr., and the Catalan Independence Movement [#catalanrepublic claybornec.wordpress.com/2018/09/14/on-...](https://claybornec.wordpress.com/2018/09/14/on-...) via @ClayborneCarson

22:17 - 14 sept. 2018

2.642 Retweets 3.332 Me gusta



264 2,6K 3,3K



TONI SOLER
@soler_toni

Segueix

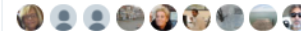
"I was shocked and disturbed today to discover that I had been misquoted in a Spanish newspaper that claimed I believe Martin Luther King, Jr. would have opposed the Catalan independence movement..."

claybornec.wordpress.com/2018/09/14/on-...

Tradueix el tuit

11:16 - 15 de set. de 2018

1.146 retuits 1.726 agradaments



13 1,1m 1,7m



Xavier Sala-i-Martin
@XSalaMartin

Seguir

Vaya! Resulta que aquello que el director del Instituto Luther King había pedido a @QuimTorraIPla que no utilizara el nombre de MLK!!! @elconfidencial manipuló sus palabras según dice él en su blog. No saben escribir sin mentir?

claybornec.wordpress.com/2018/09/14/on-...

2:07 - 15 sept. 2018

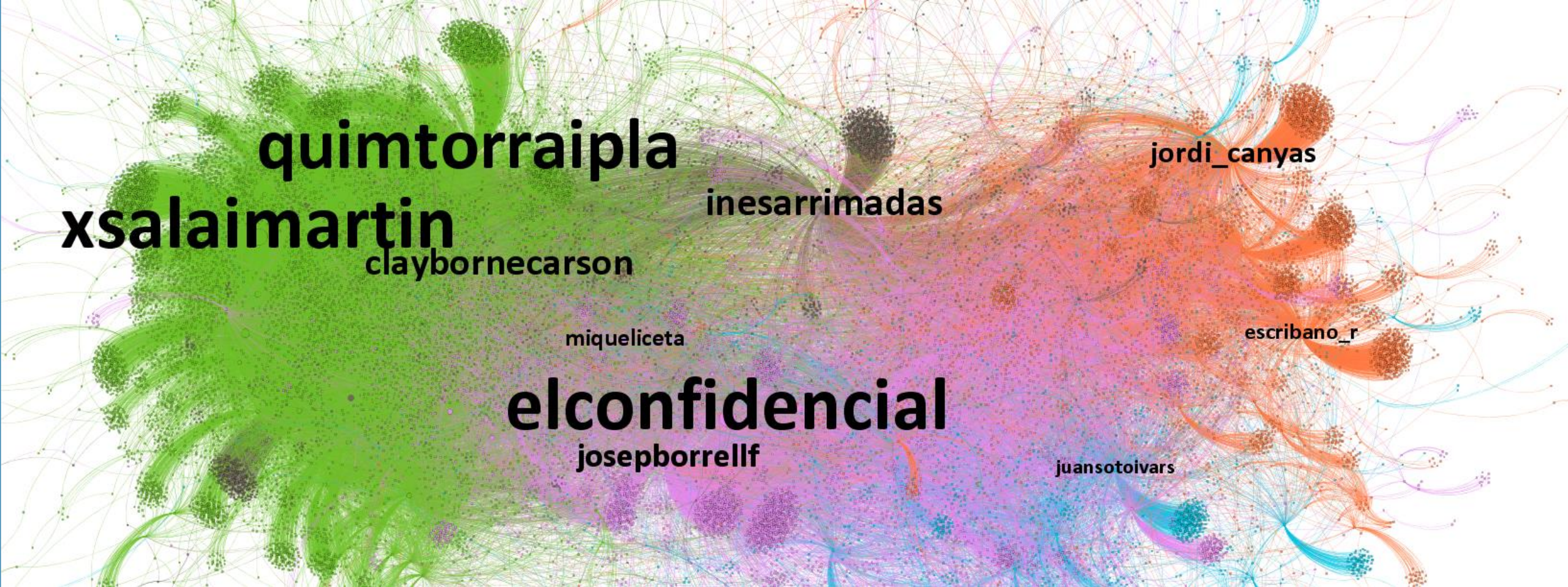
1.627 Retweets 2.739 Me gusta



74 1,6K 2,7K

2. LA DESINTOXICACIÓN

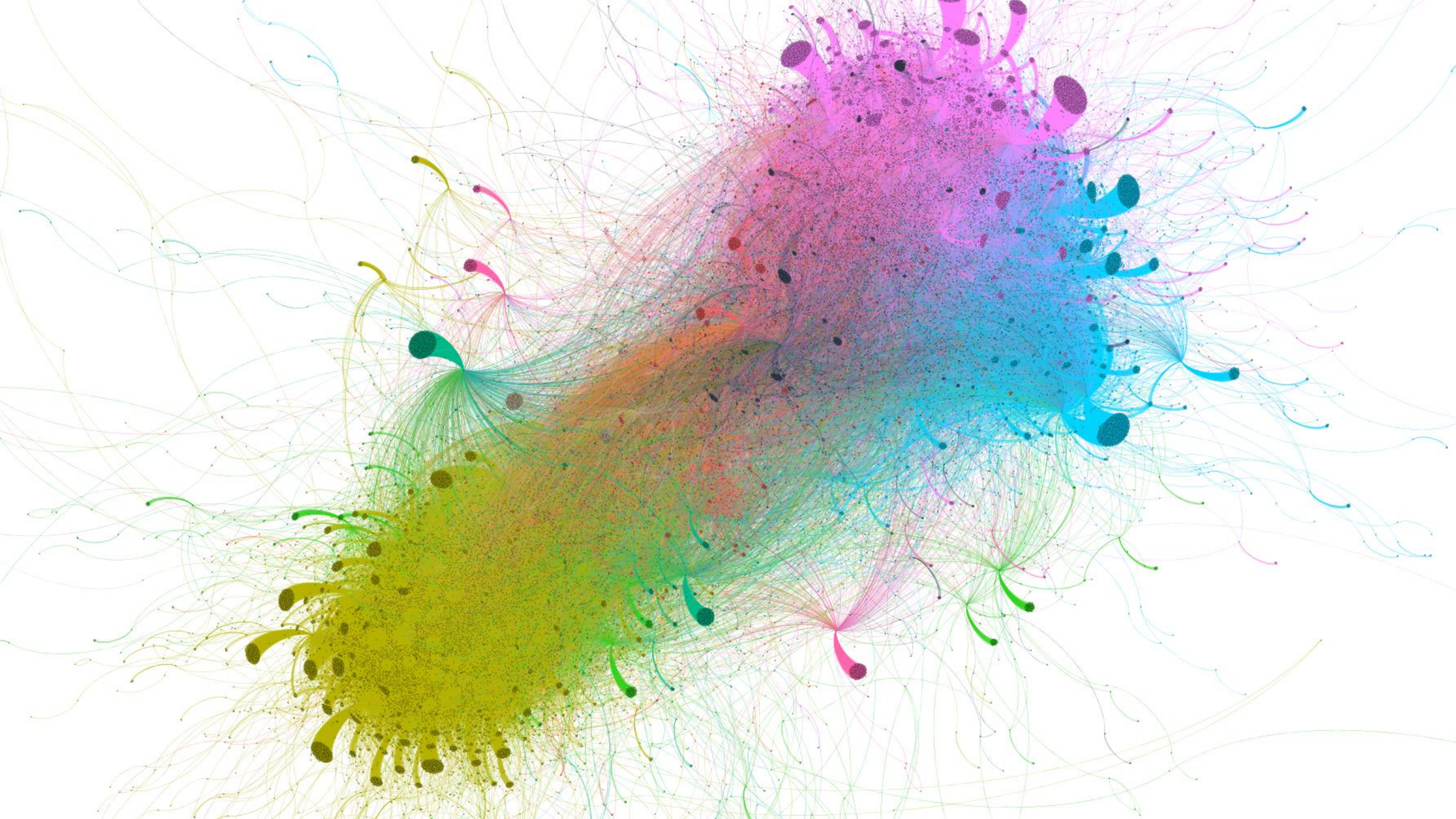
I was shocked and disturbed today to discover that I had been misquoted in a Spanish newspaper that claimed I believe Martin Luther King, Jr. would have opposed the Catalan independence movement. This distorts what I actually said in response to a Spanish reporter's questions about whether proponents of Catalan independence were justified in using King's name to support their struggle. In order to correct misunderstandings, I wish

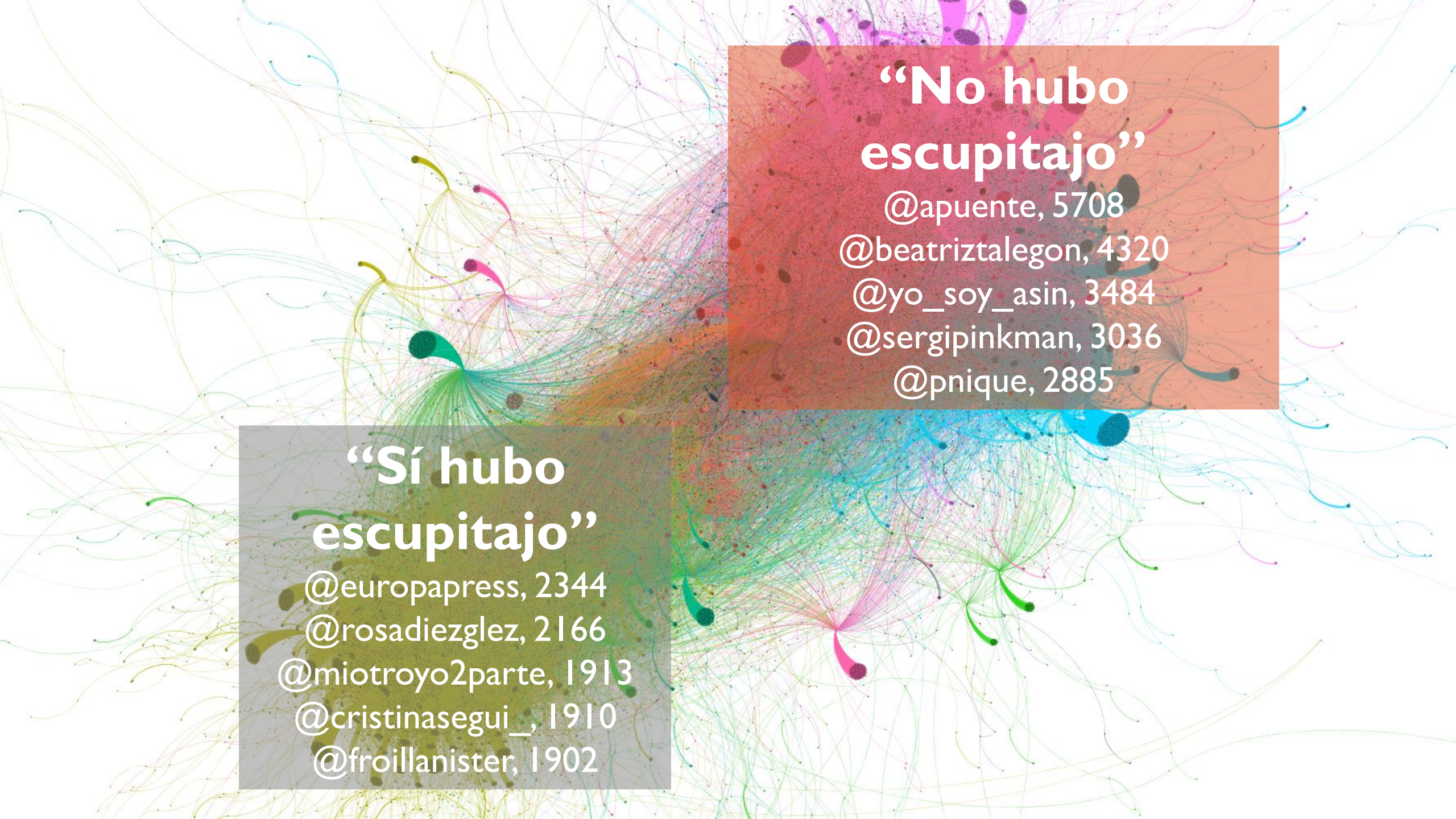


3. VIRALIZACIÓN DIFERENCIADA

CASO DE ANÁLISIS 2

EL ESCUPITAJO A BORRELL





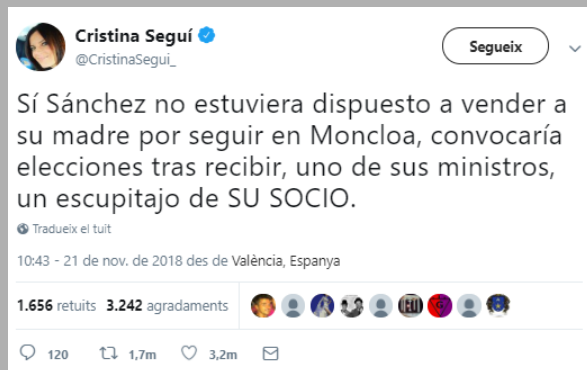
“No hubo escupitajo”

@apunte, 5708
@beatriztalegon, 4320
@yo_soy_asin, 3484
@sergipinkman, 3036
@pnique, 2885

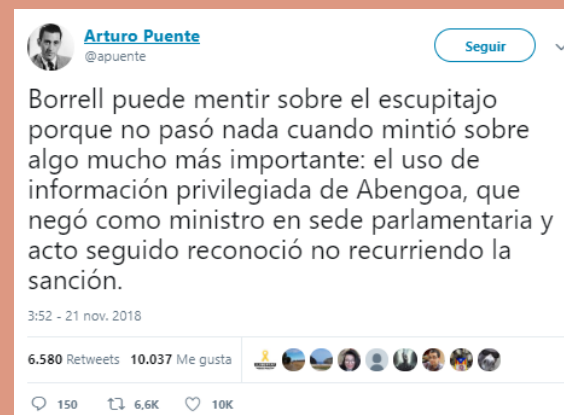
“Sí hubo escupitajo”

@europapress, 2344
@rosadiezglez, 2166
@miotroyo2parte, 1913
@cristinasegui_, 1910
@froillanister, 1902

“Sí hubo escupitajo”

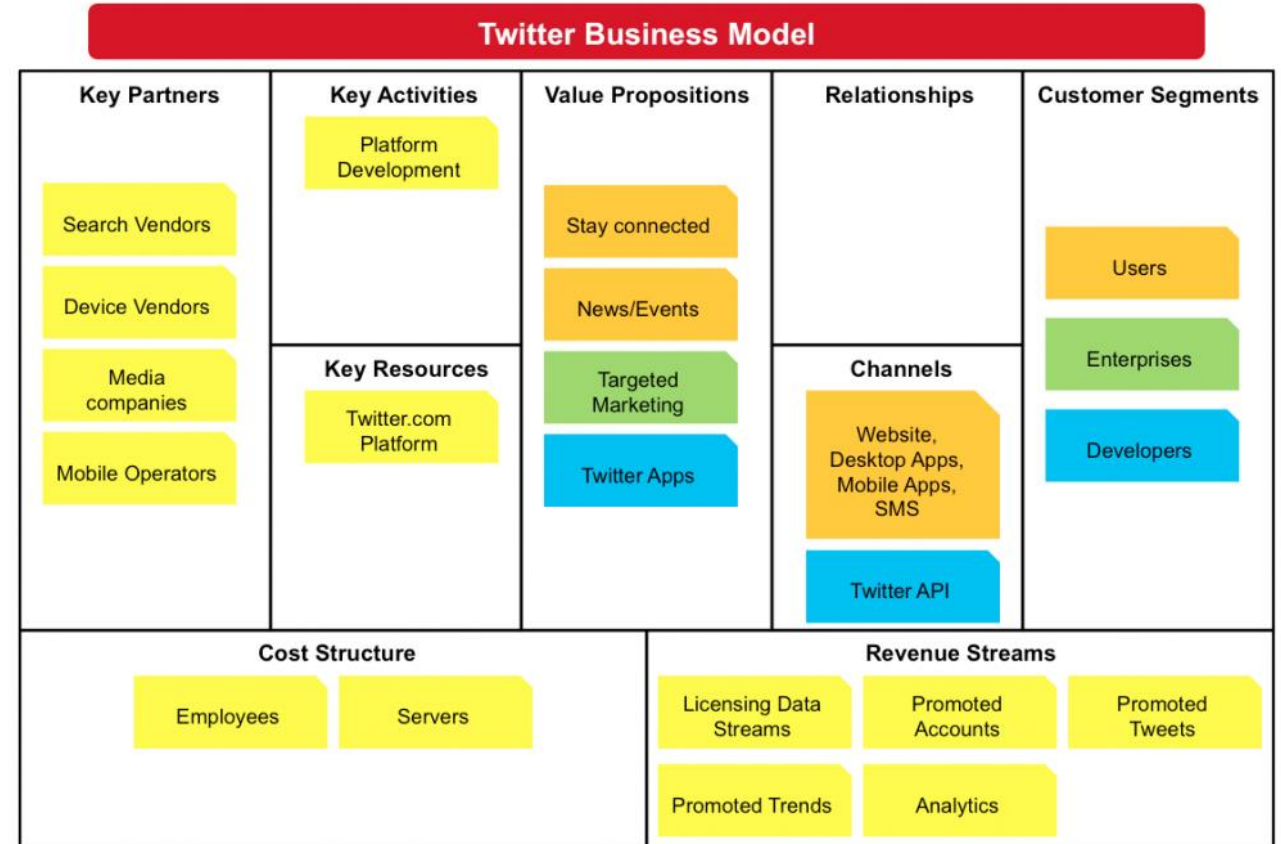


“No hubo escupitajo”



LA API DE TWITTER

¿POR QUÉ
TWITTER
TIENE API?



www.businessmodelgeneration.com

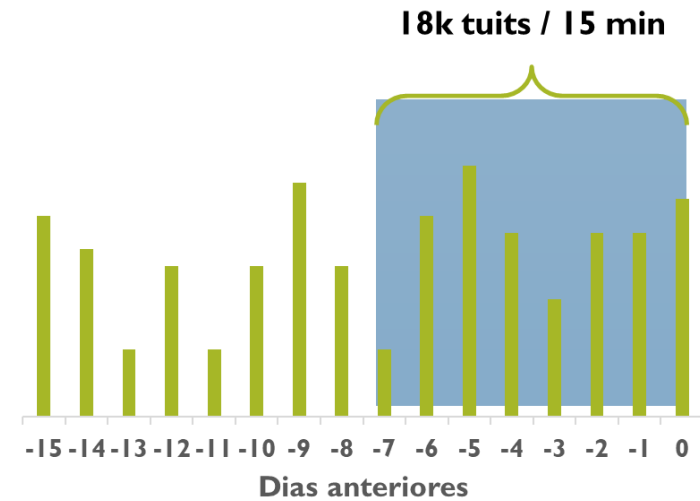
LAS APIS DE TWITTER

Feature summary

Category	Product name	Supported history
Standard	Standard Search API	7 days
Premium	Search Tweets: 30-day endpoint	30 days
Premium	Search Tweets: Full-archive endpoint	Tweets from as early as 2006
Enterprise	30-day Search API	30 days
Enterprise	Full-archive Search API	Tweets from as early as 2006

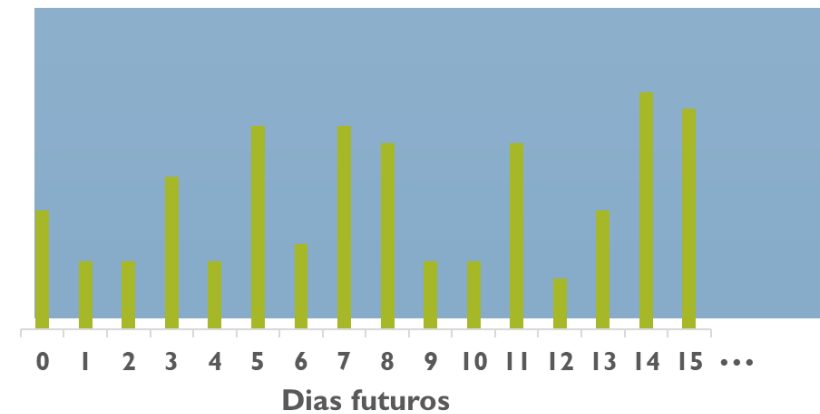
CONSULTAS EN TWITTER

Consultas retroactivas

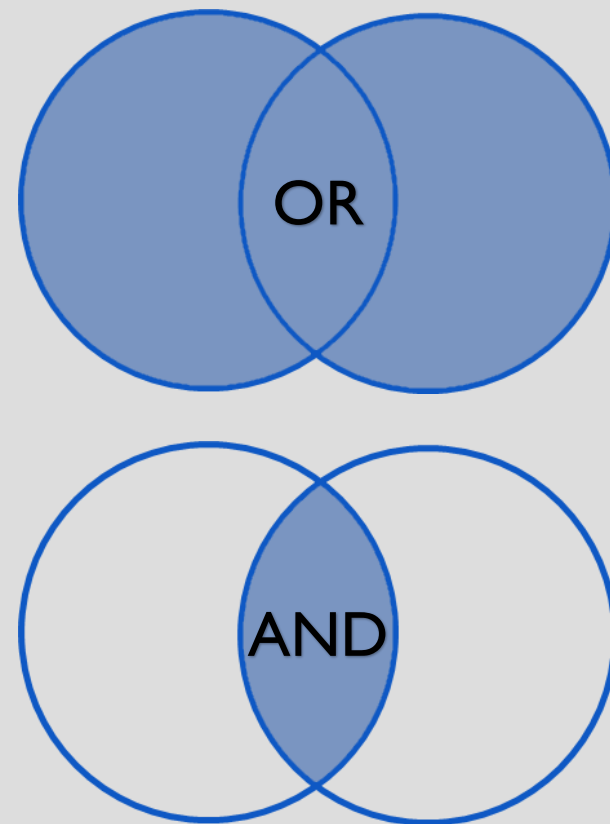


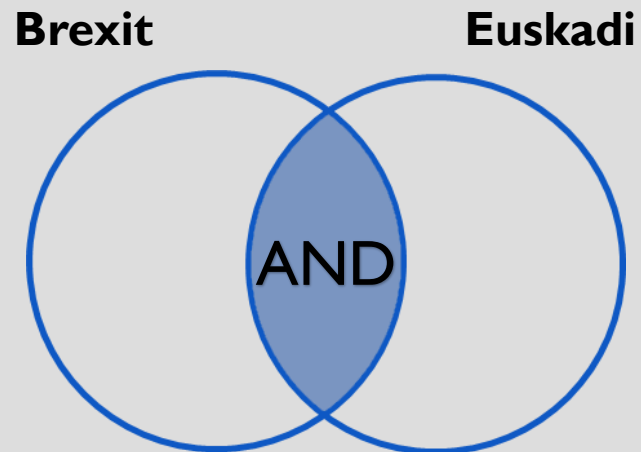
Consultas en tiempo real

→ → → Max 1% tuits



El lenguaje de la **API**



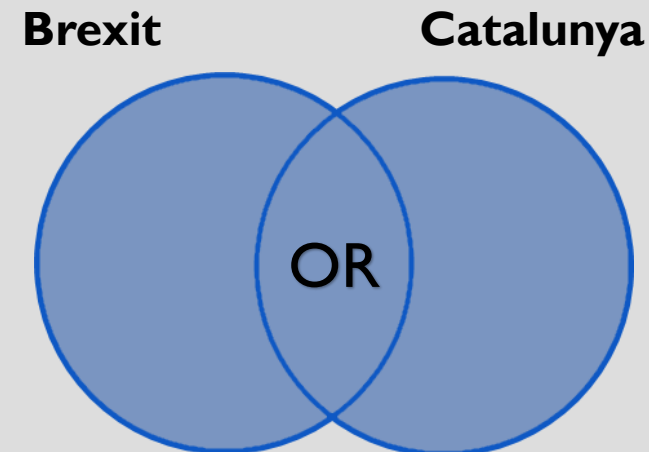


“Me inquieta como afectará el **Brexit** en **Euskadi**”

“Nos conocimos en **Euskadi** mucho antes del **Brexit**”

“A mi el **Brexit** no me asusta”

“Lloverá en **Euskadi**, da igual cuando leas esto”

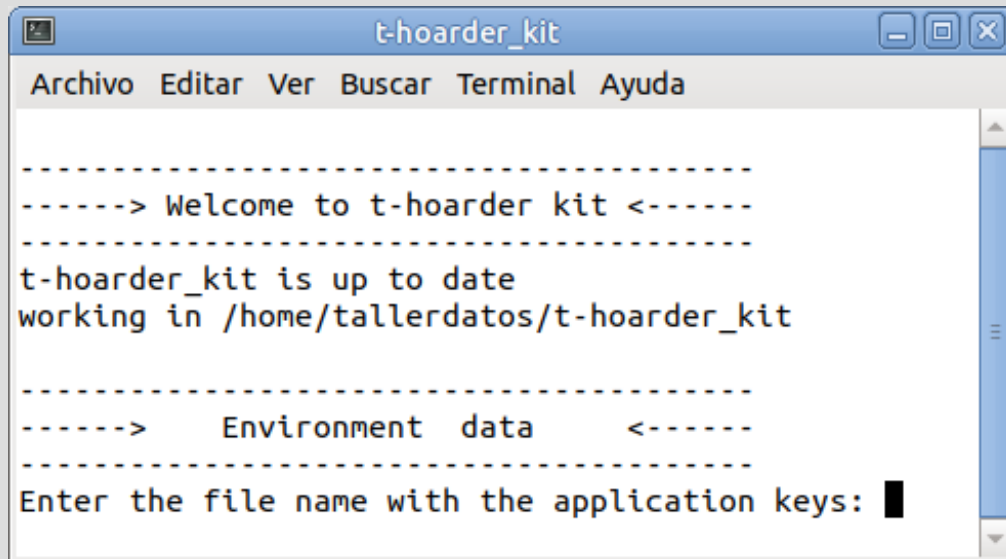


“Me inquieta como afectará el **Brexit** en **Euskadi**”

“Nos conocimos en **Euskadi** mucho antes del **Brexit**”

“A mi el **Brexit** no me asusta”

“Lloverá en **Euskadi**, da igual cuando leas esto”



```
t-hoarder_kit
Archivo  Editar  Ver  Buscar  Terminal  Ayuda

-----> Welcome to t-hoarder kit <-----
----->
t-hoarder_kit is up to date
working in /home/tallerdatos/t-hoarder_kit

-----> Environment data <-----
----->
Enter the file name with the application keys: █
```

Brexit AND Europa lang:es

Brexit AND “Unión Europea” lang:es

“La **Unión** de Comerciantes de Bilbao analizan los efectos de **Brexit** en el turismo de procedencia **europea**”

(Brexit AND Europa) OR (Brexit AND “Uninó
Europea”) lang:es

CAPTURA DE DATOS CON T-HOARDER

[HTTPS://GITHUB.COM/CONGOSTO](https://github.com/congosto)

T-HOARDER KIT

```
-----  
What function do you want to run?  
-----
```

1. Get a user token access
2. Get users information (profile | followers | following | relations | tweets | role)
3. Make a query on Twitter
4. Get tweets on real time
5. Generate the declared relations graph (followers or following or both)
6. Generate the dynamic relations graph (RT | reply | mentions)
7. Processing tweets (entities| classify| users | spread)
8. utils (sort| user-cards| add-communities |spread-by-communities| get_photos-community)
9. Exit

```
--> Enter option: █
```

ESTRATEGIAS DE OBTENCIÓN DE DATOS TRADICIONALES VS. COMPUTACIONALES



ESTRATEGIAS CLÁSICAS PARA LA OBTENCIÓN DE DATOS SOCIALES



FUENTES DE DATOS MASIVOS



MEDIOS SOCIALES

Imagen, video, audio o texto de redes sociales virtuales



CLOUD

Público, privado o corporativo



WEB

Datos web, analytics



IoT

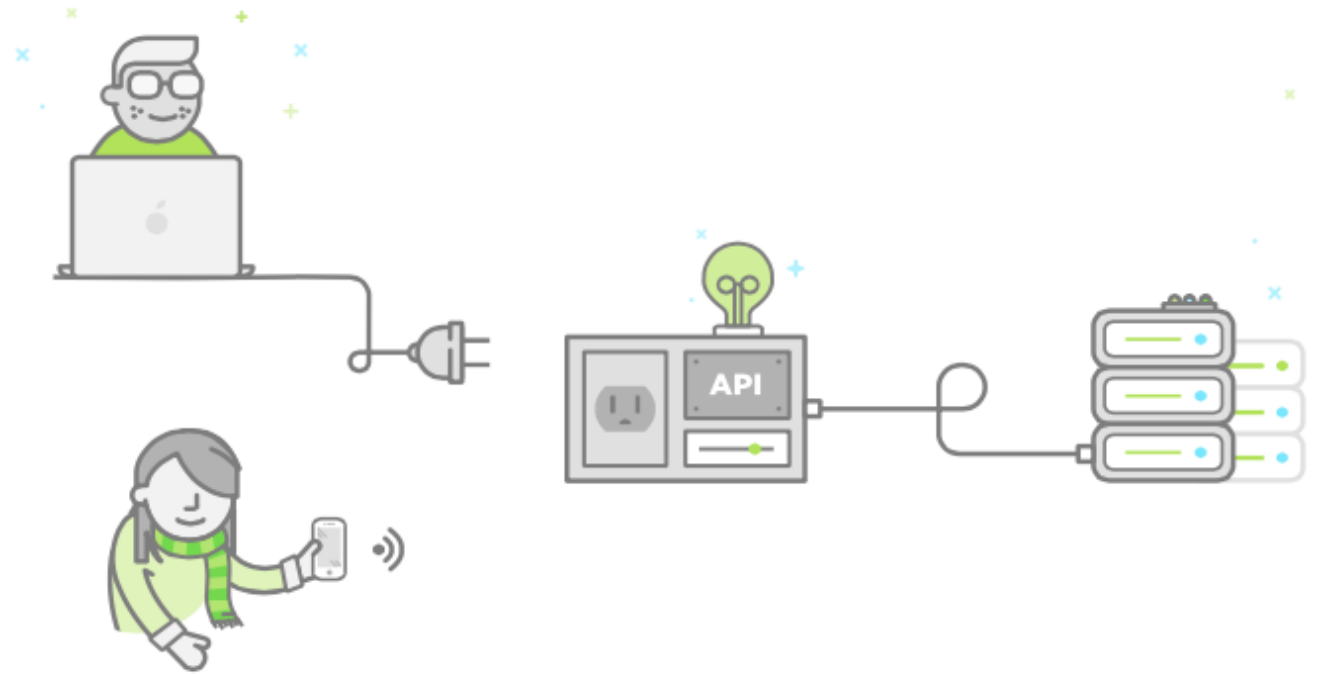
Sensores y dispositivos conectados



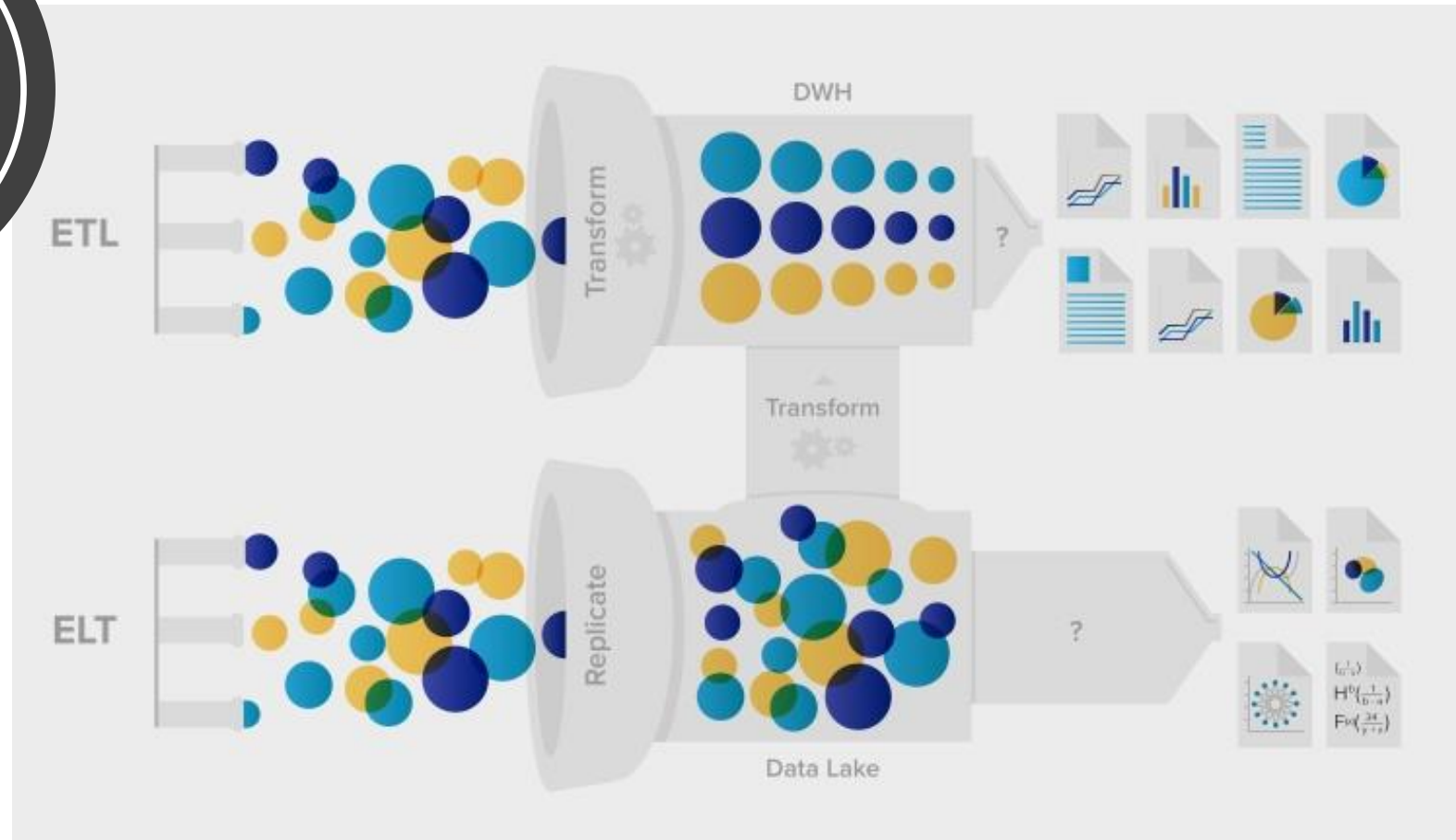
BASES DE DATOS

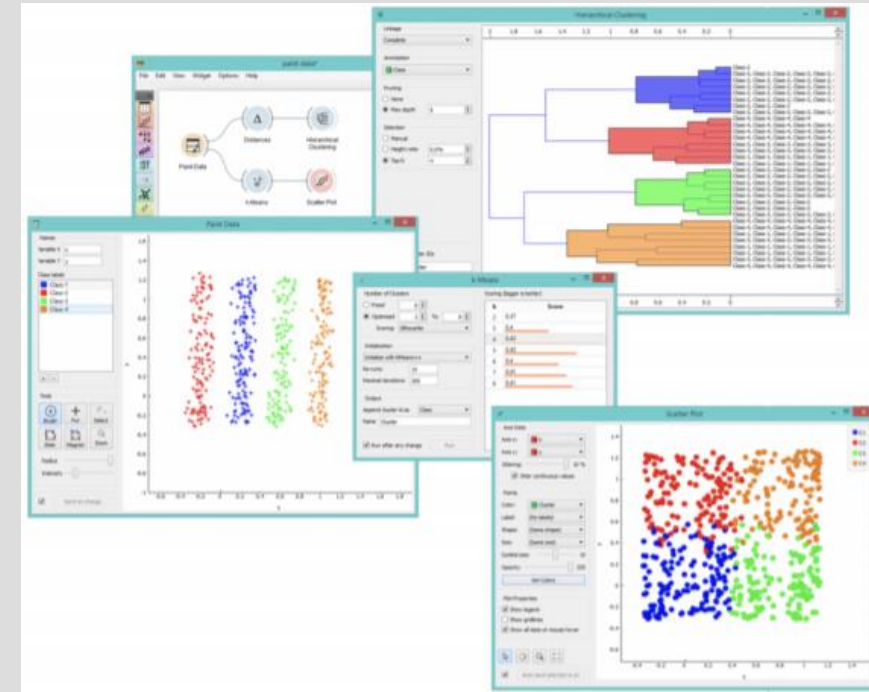
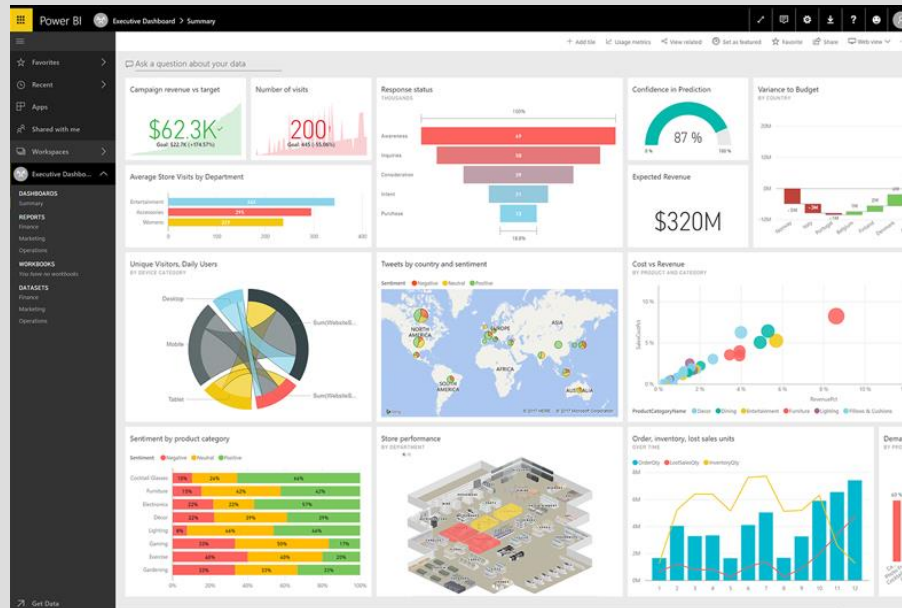
Datos secundarios tradicionales

EL NUEVO CIRCUITO DEL DATO

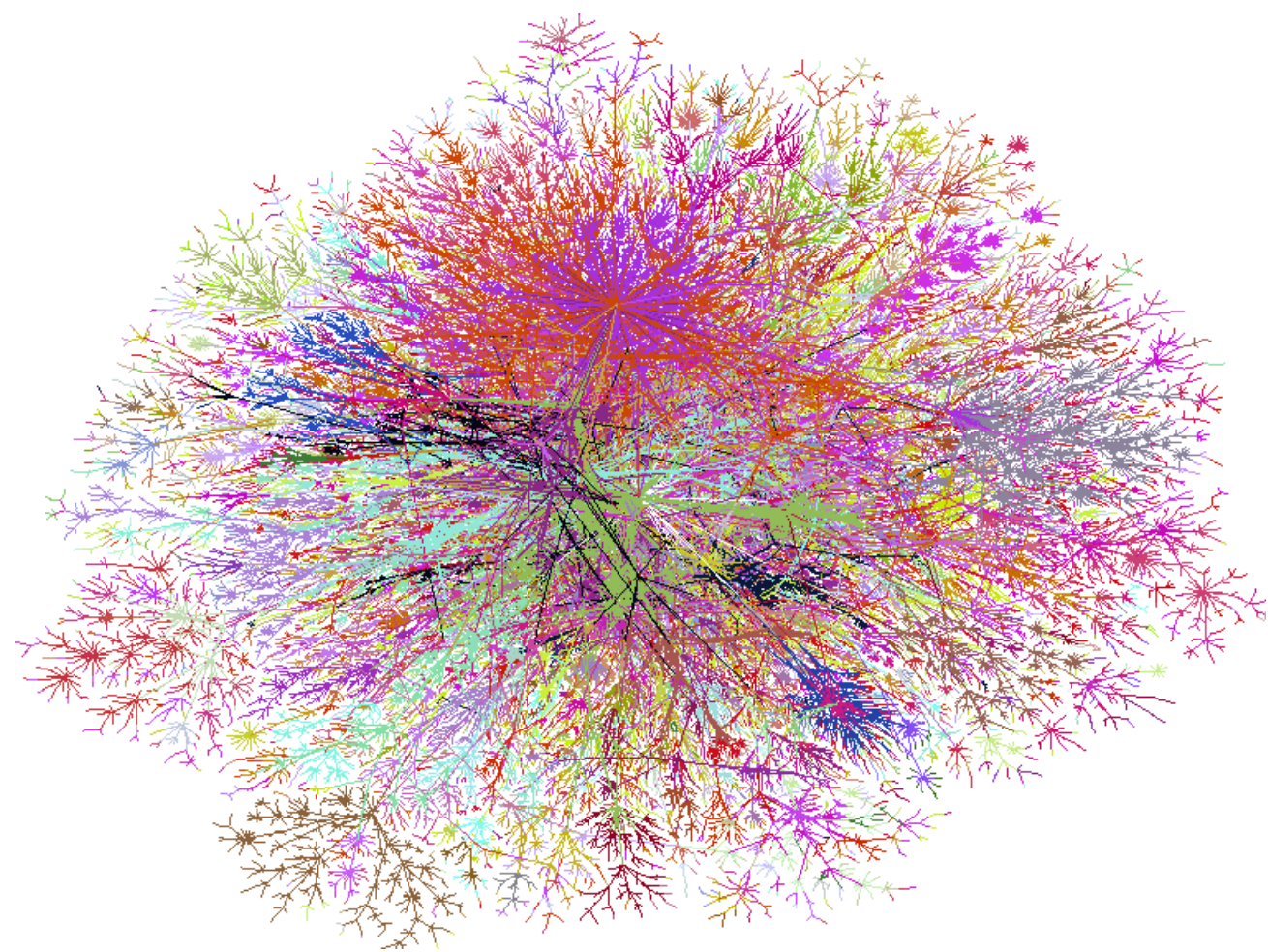


ETL VS ELT

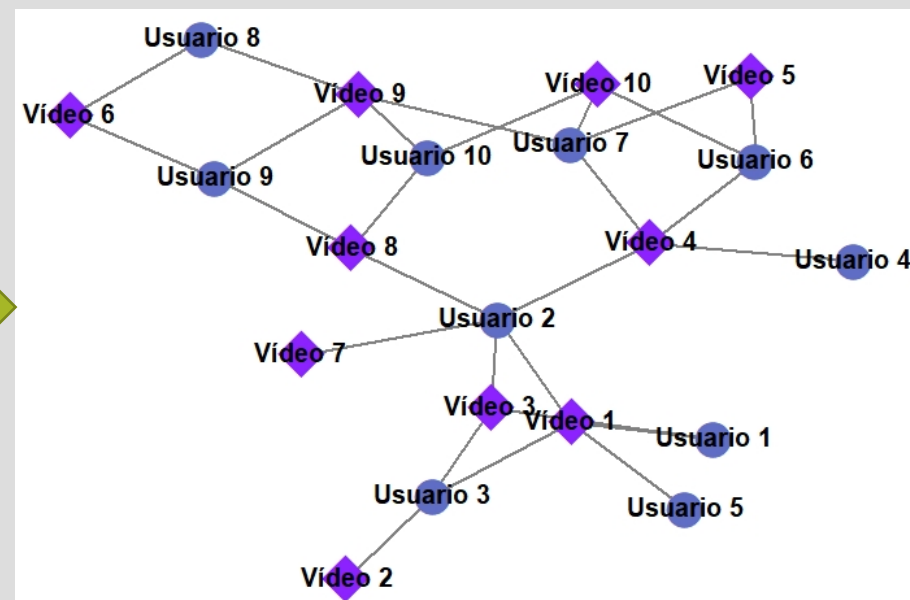




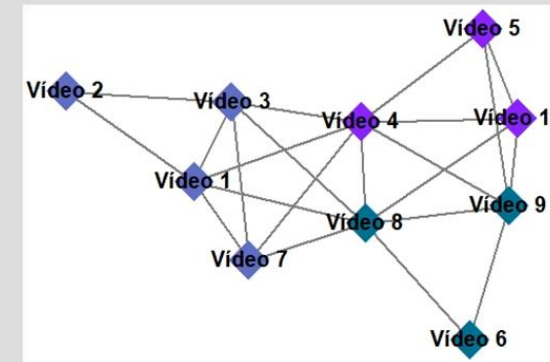
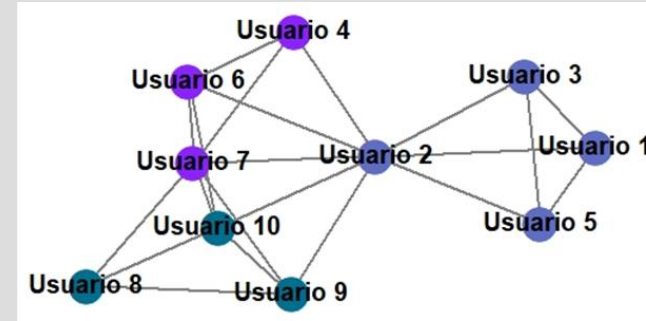
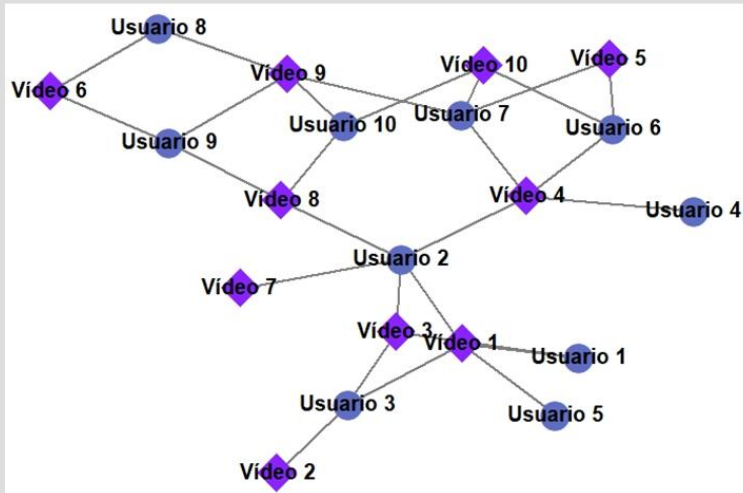
BI (ETL) VS. DATAMINING (ELT)



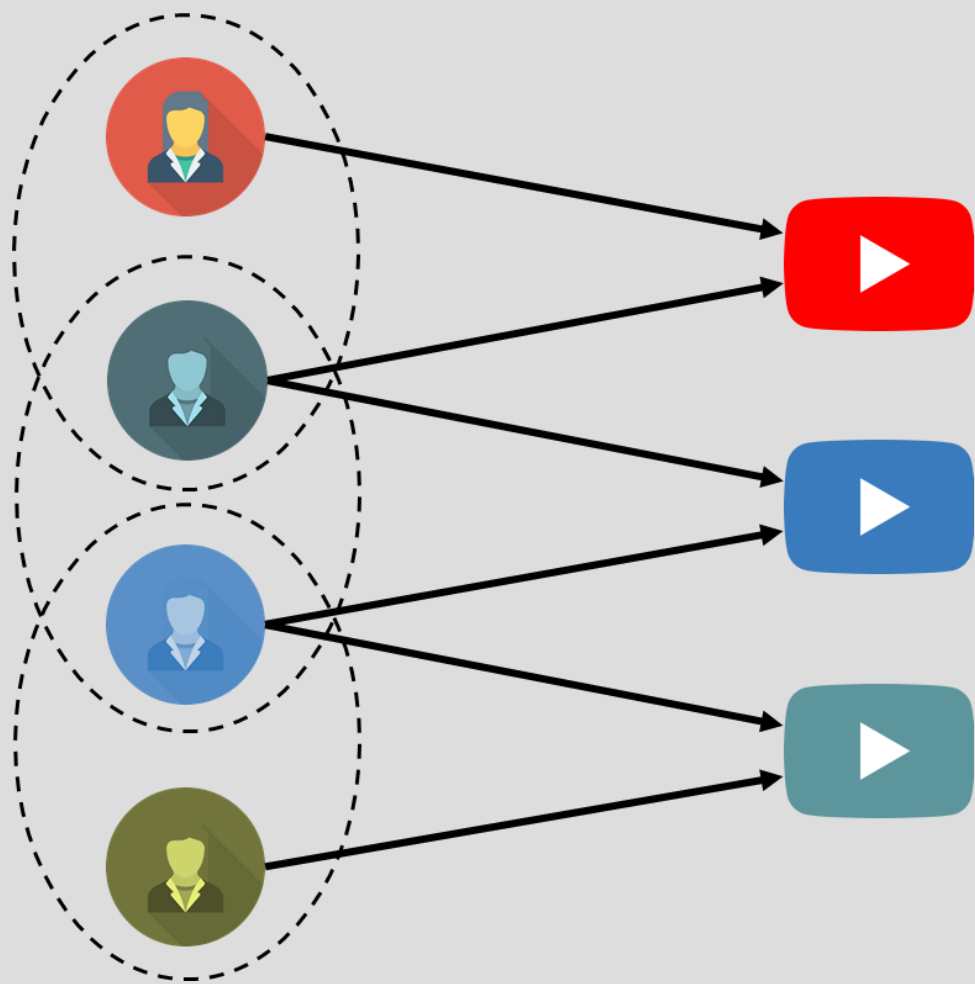
	Vídeos									
	1	2	3	4	5	6	7	8	9	10
Usuarios	1	x		x						
	2	x		x	x			x	x	
	3	x	x	x						
	4				x					
	5	x								
	6				x	x				x
	7				x	x			x	x
	8						x		x	
	9					x		x	x	
	10							x	x	x



EJEMPLO: DE MATRIZ A RED



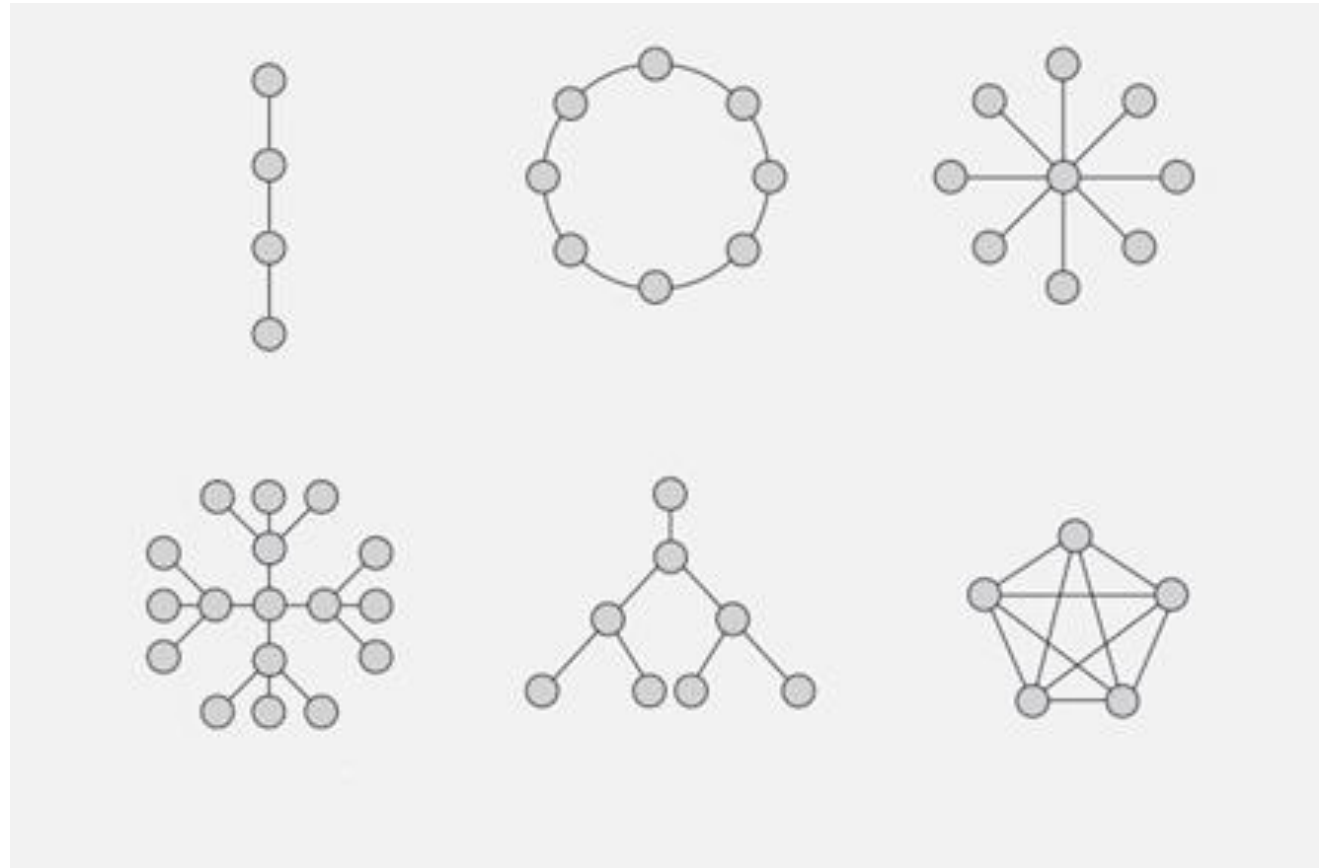
EJEMPLO: DE RED BIPARTITA A RED UNIPARTITA



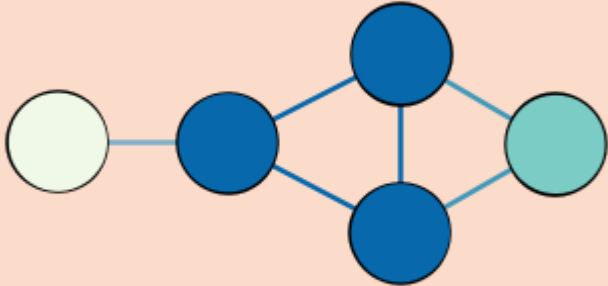
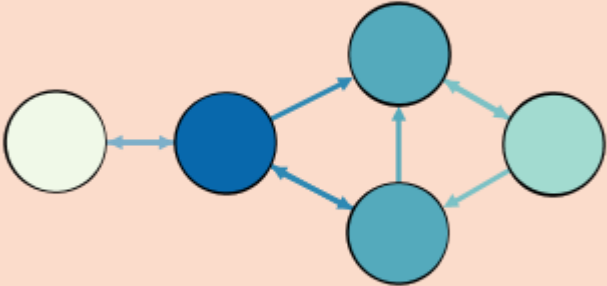
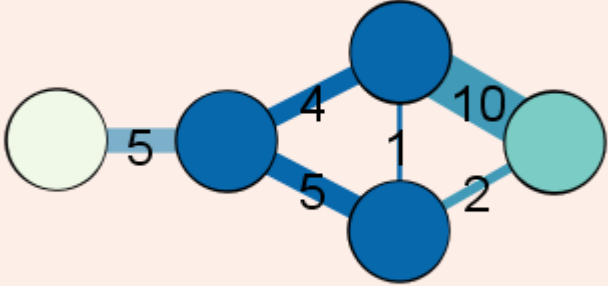
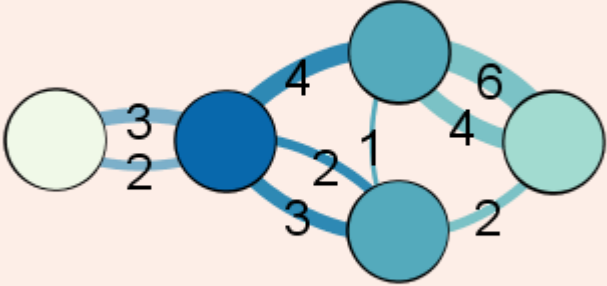
INDICADORES DEL ARS EN DATOS MASIVOS

CONCEPTOS BÁSICOS

- Nodos (personas, empresas, instituciones, **cuentas de Twitter...**)
- Enlaces (referencialidad, relaciones, transacciones, **menciones...**)
- Grafos de red:



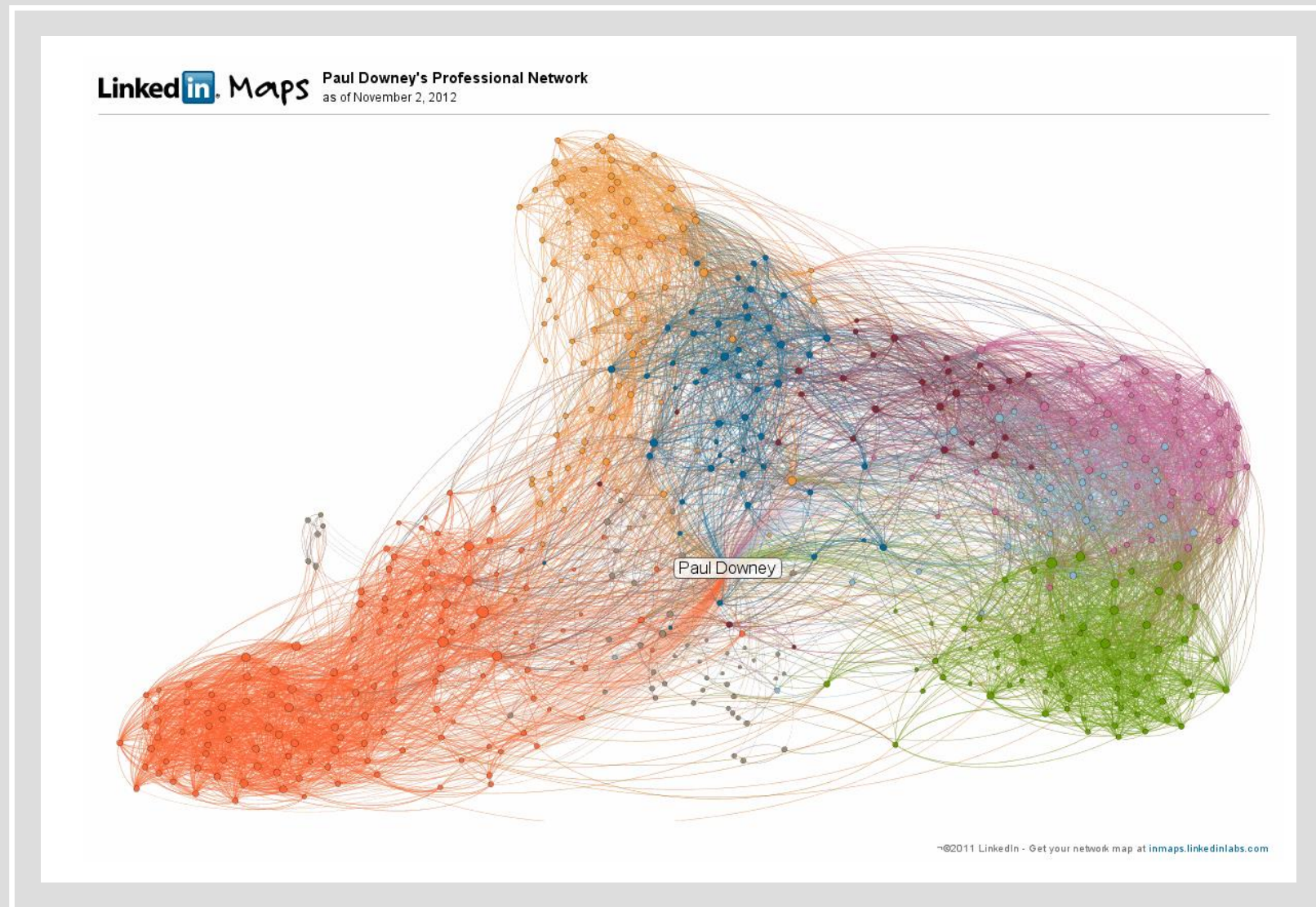
TIPOS DE RED (DISTINCIÓN FORMAL)

¿Las aristas tienen dirección? → ¿Las aristas tienen peso? ↴	NO (RED NO DIRIGIDA)	SI (RED DIRIGIDA)
NO (RED NO PONDERADA)	 A diagram of an undirected, unweighted graph. It consists of 6 nodes: a light green node on the far left, a dark blue node next to it, two dark blue nodes in the center, and a light green node on the far right. The edges are: a single edge between the leftmost light green node and the first dark blue node; two edges connecting the first dark blue node to the two central dark blue nodes; and three edges connecting the two central dark blue nodes to the rightmost light green node.	 A diagram of a directed, unweighted graph with the same structure as the previous one. The edges are directed: from the leftmost light green node to the first dark blue node; from the first dark blue node to both central dark blue nodes; and from both central dark blue nodes to the rightmost light green node.
SI (RED PONDERADA)	 A diagram of an undirected, weighted graph with the same structure. The edges are labeled with weights: 5 for the edge between the leftmost light green node and the first dark blue node; 4 for the edge between the first dark blue node and the top central dark blue node; 5 for the edge between the first dark blue node and the bottom central dark blue node; 1 for the edge between the two central dark blue nodes; 10 for the edge between the top central dark blue node and the rightmost light green node; 2 for the edge between the bottom central dark blue node and the rightmost light green node.	 A diagram of a directed, weighted graph with the same structure. The edges are directed and labeled with weights: 3 for the edge from the leftmost light green node to the first dark blue node; 4 for the edge from the first dark blue node to the top central dark blue node; 2 for the edge from the first dark blue node to the bottom central dark blue node; 1 for the edge from the top central dark blue node to the bottom central dark blue node; 6 for the edge from the top central dark blue node to the rightmost light green node; 2 for the edge from the bottom central dark blue node to the rightmost light green node; and 3 for the edge from the bottom central dark blue node back to the first dark blue node.

RED DE
AMISTADES
EN
FACEBOOK

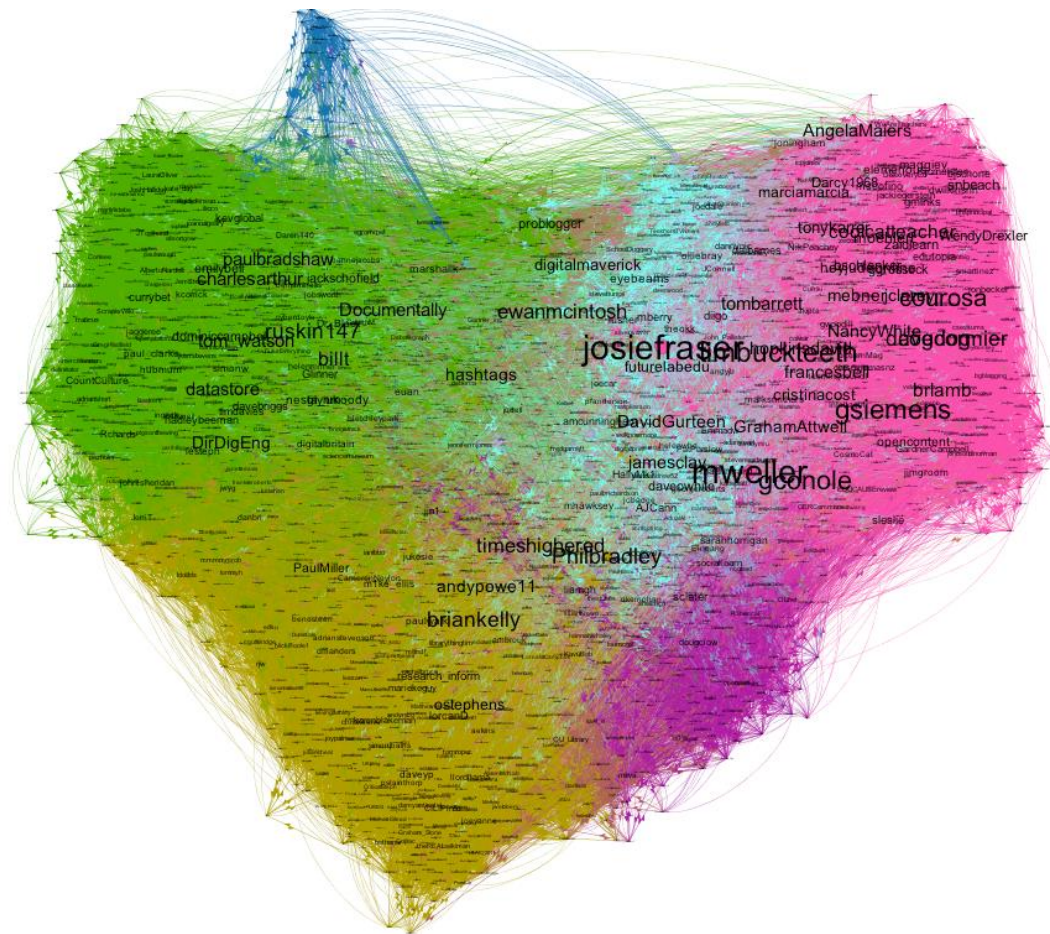


Funcionalidad obsoleta

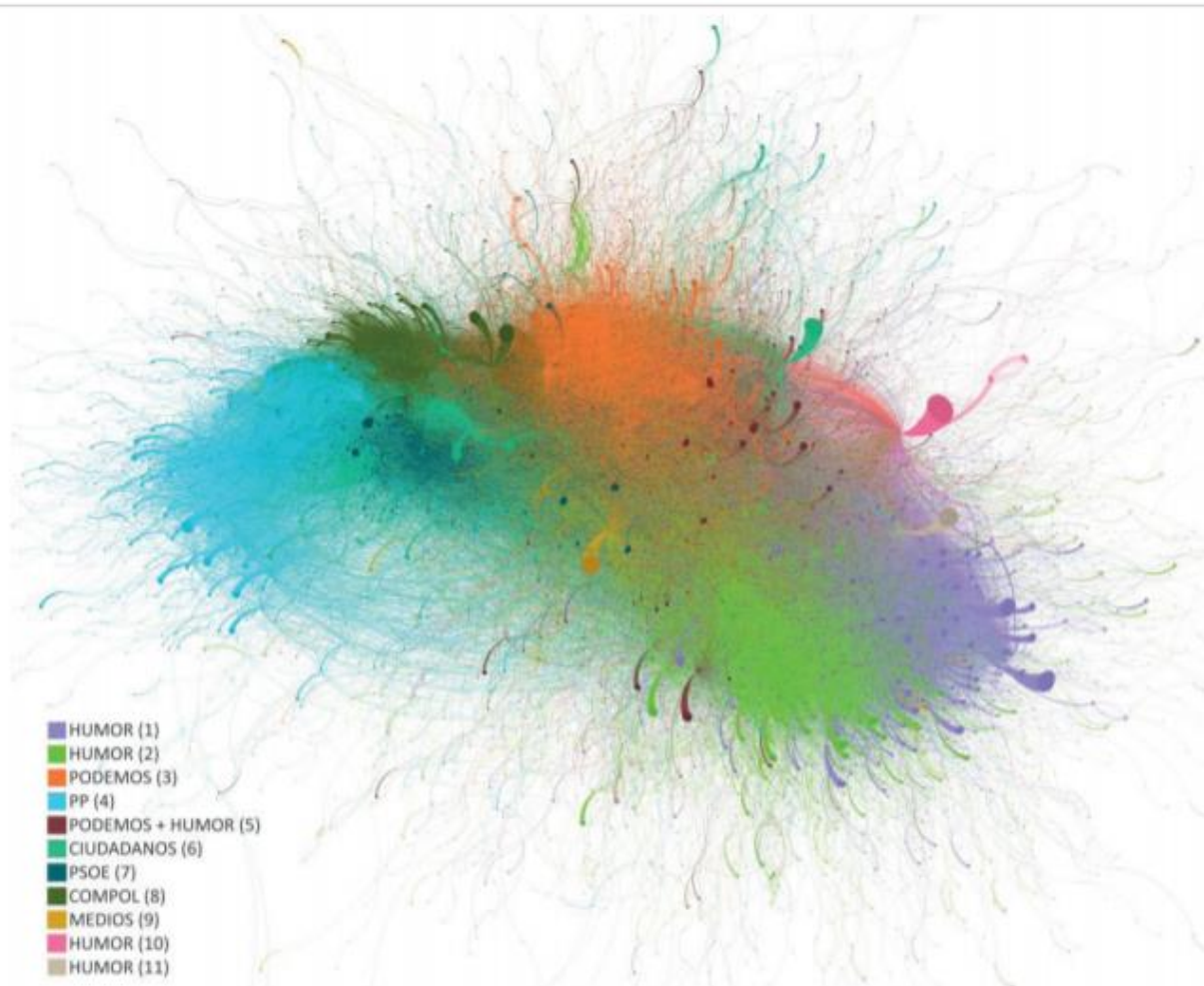


Funcionalidad obsoleta

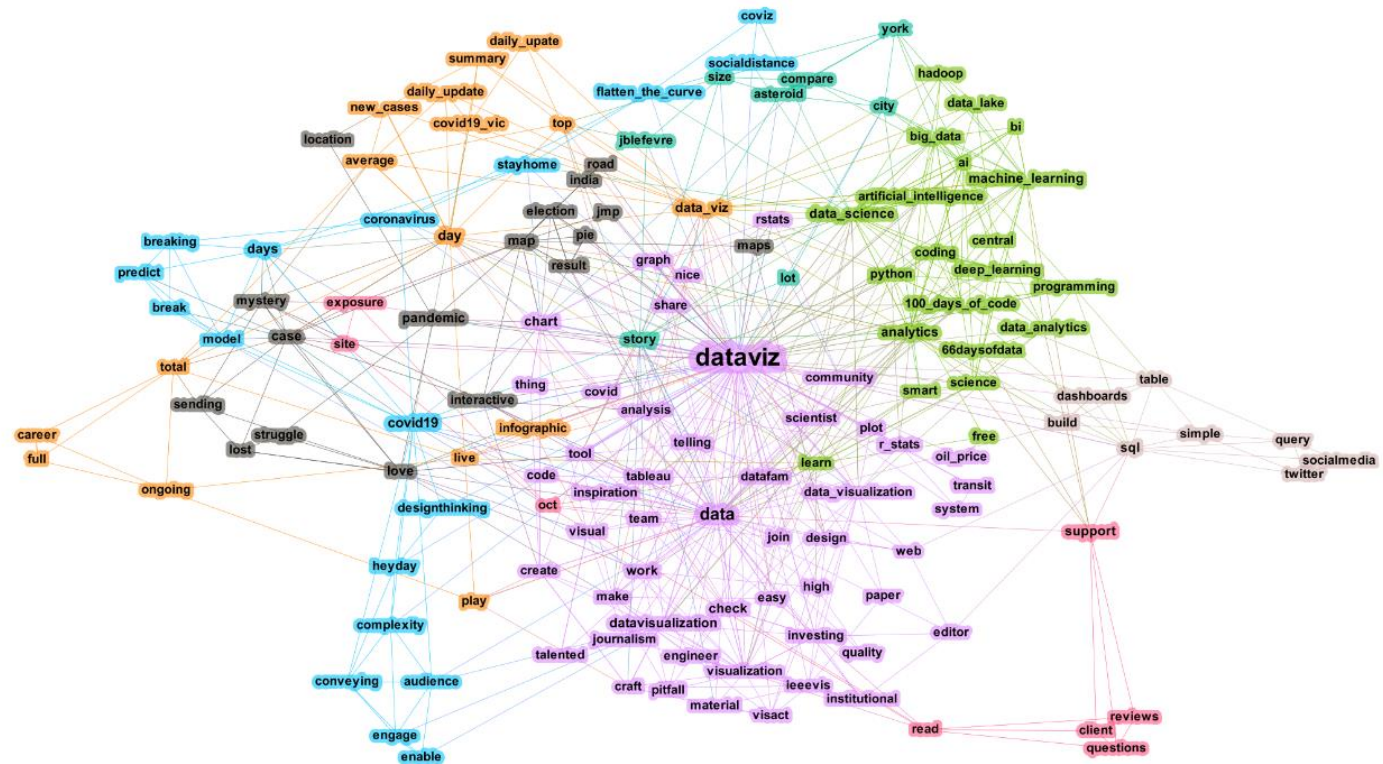
RED DE
SEGUIDORES
TWITTER



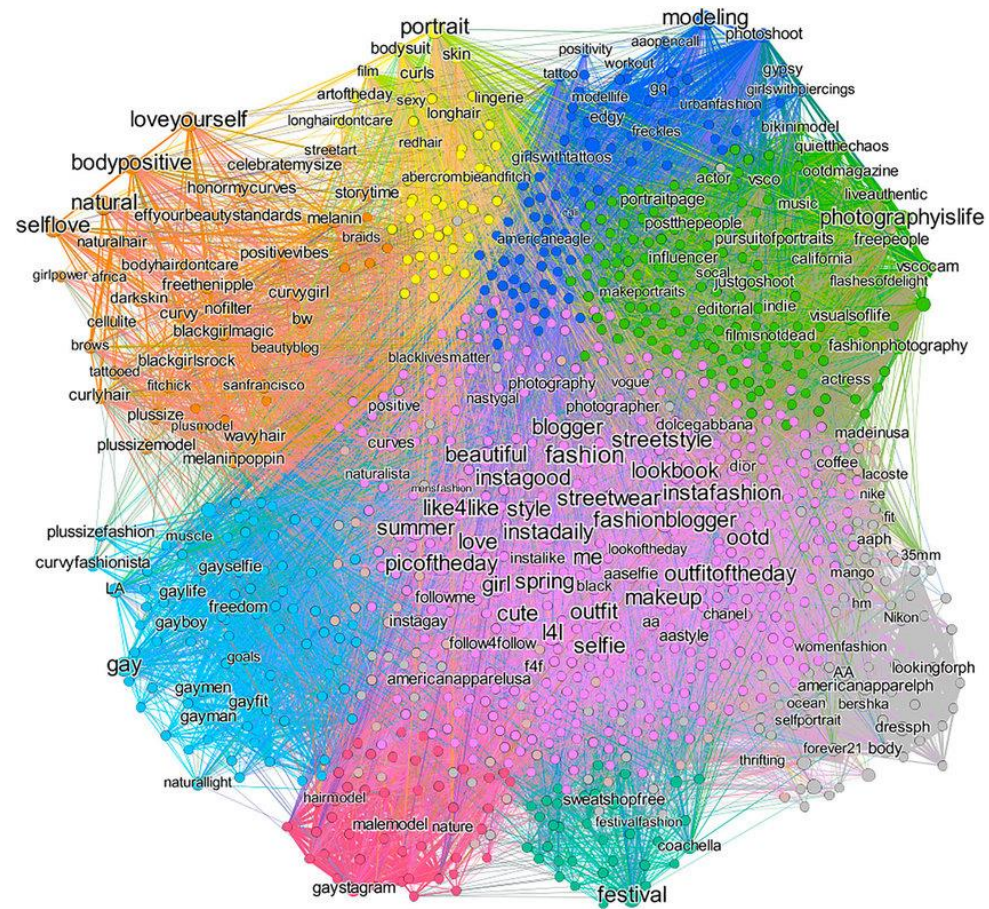
RED DE MENCIONES EN TWITTER



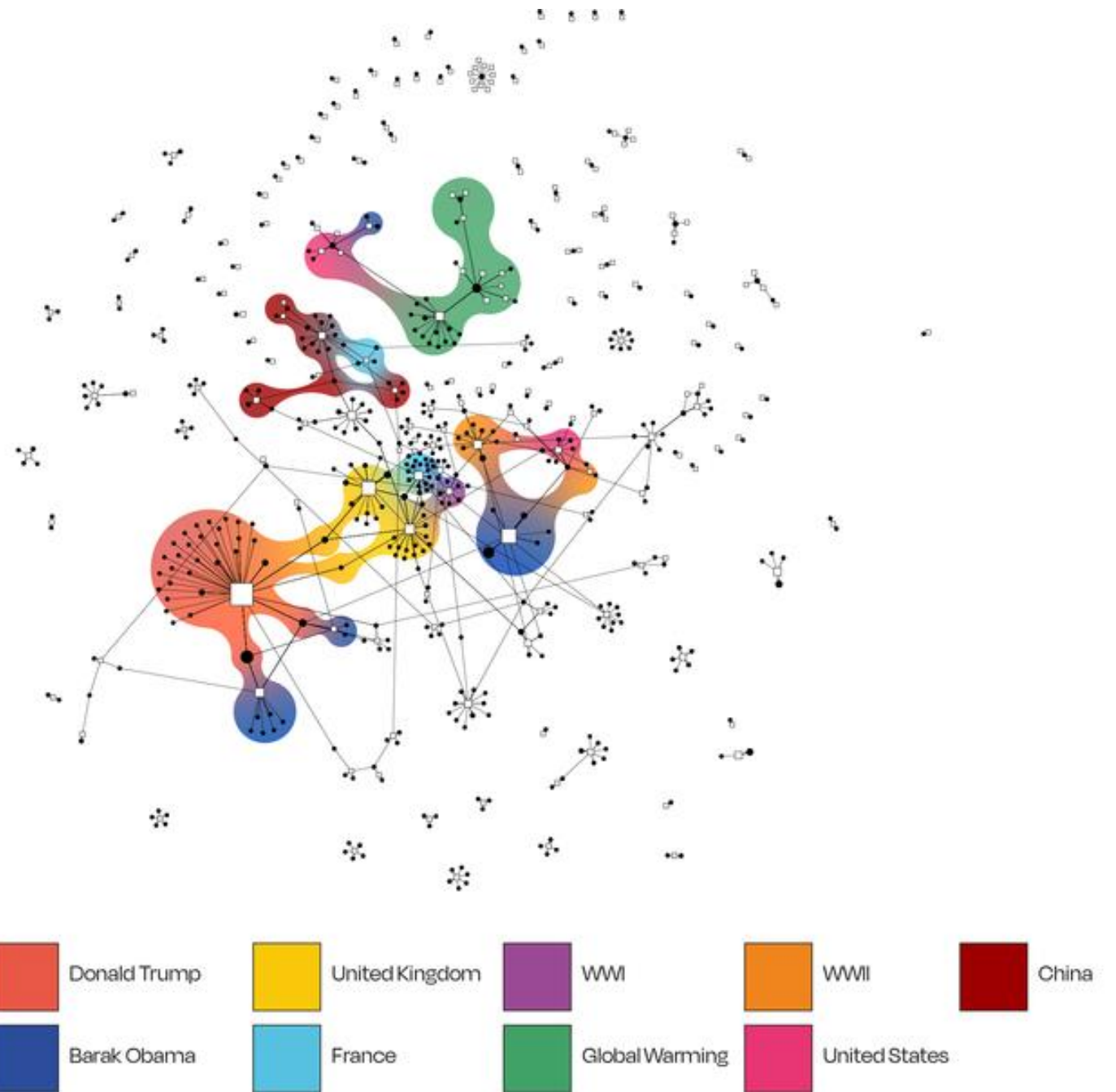
RED DE CONCEPTOS (TWITTER)



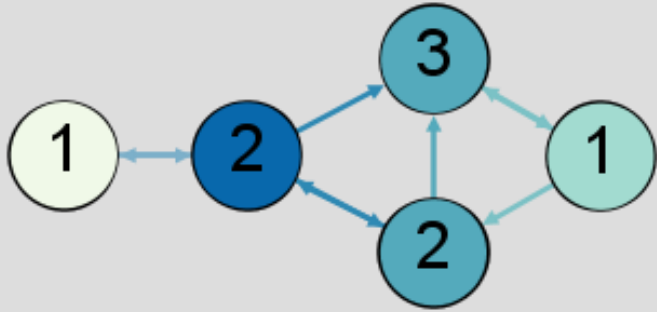
RED DE HASHTAGS (INSTAGRAM)



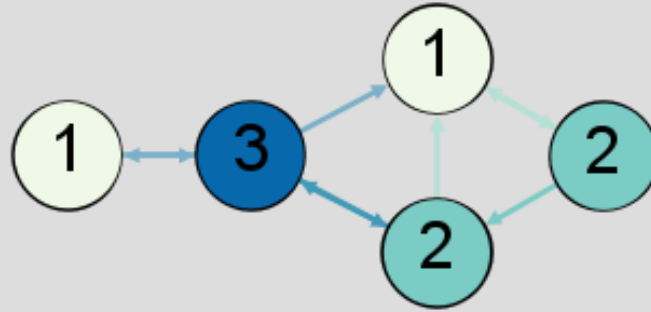
RED DE
AUTORES E
IMÁGENES
(WIKIPEDIA)



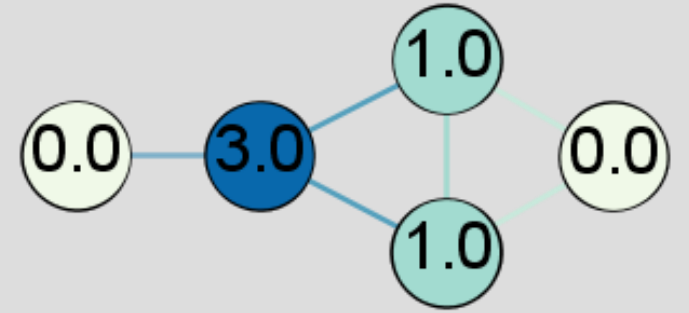
PROPIEDADES BÁSICAS DE LOS NODOS



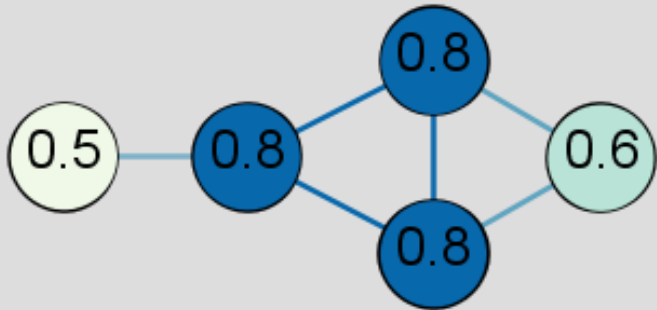
GRADO DE ENTRADA



GRADO DE SALIDA



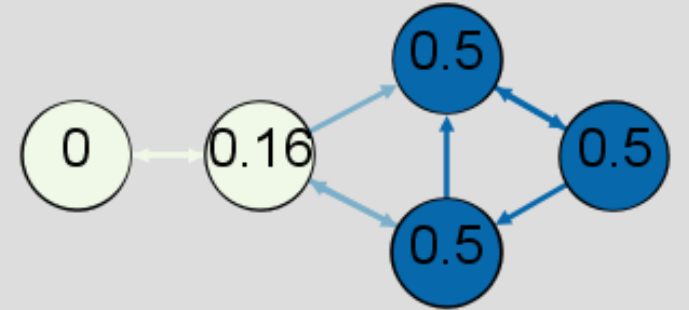
INTERMEDIACIÓN



CERCANÍA

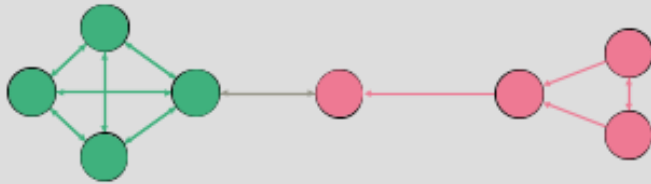


PAGERANK

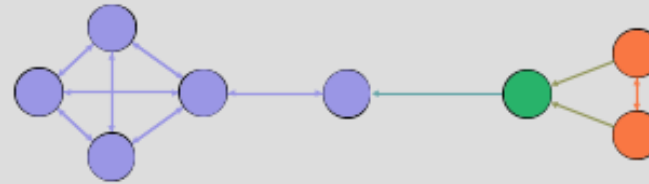


COEF.AGRUPACIÓN

MÉTODOS DE CLUSTERING

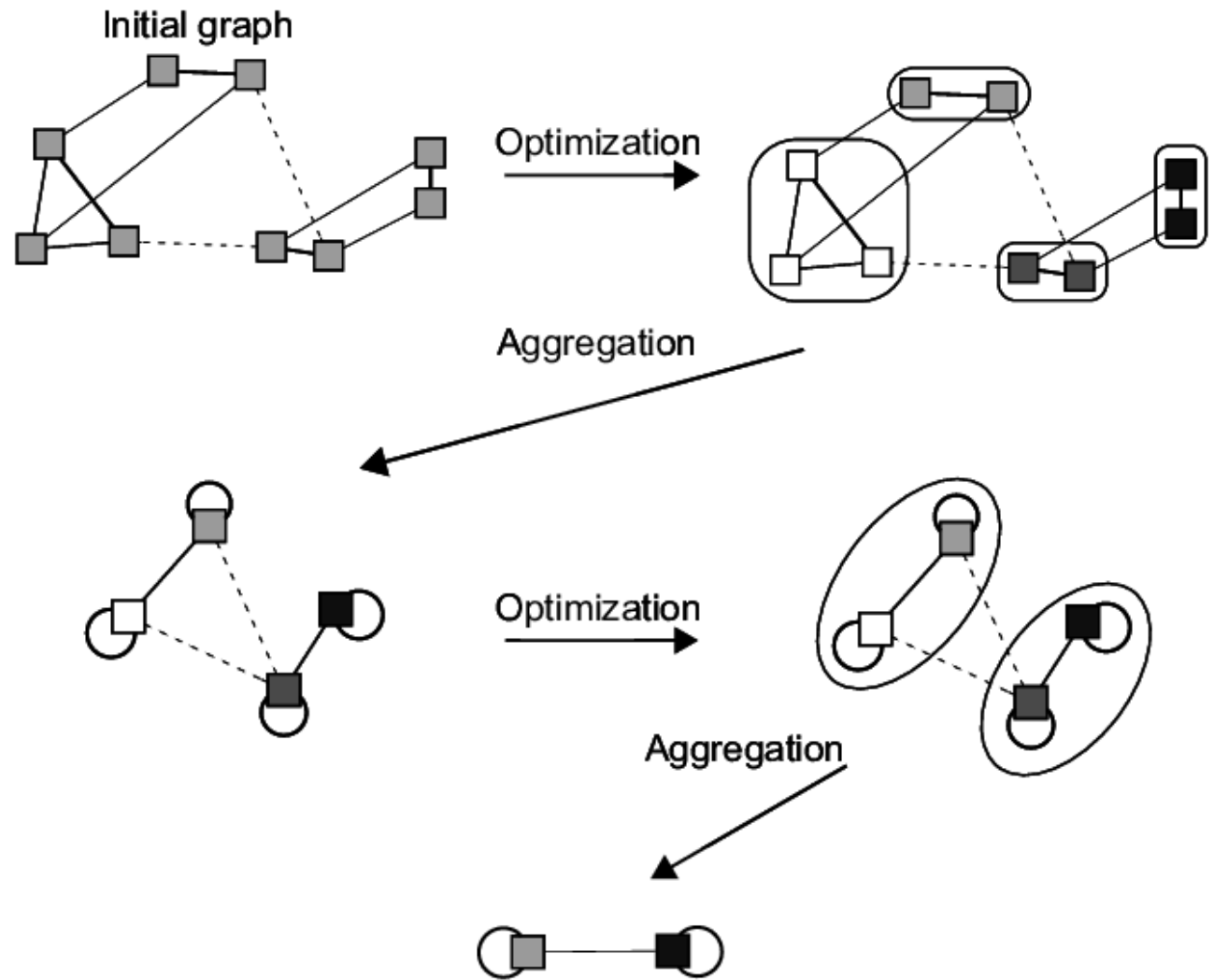


MODULARIDAD
($Q = 0.393$)



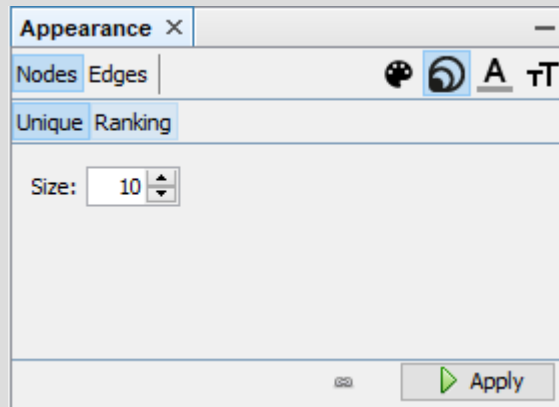
COMPONENTES
FUERTES

MÉTODO LOUVAIN (MODULARIDAD)

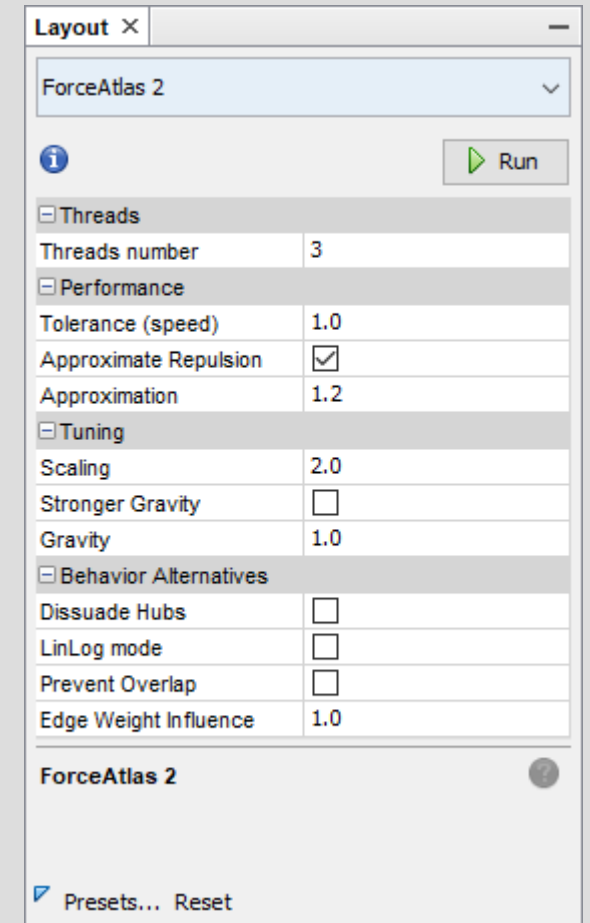


ALGORITMOS DE VISUALIZACIÓN DE DATOS

SINTETIZACIÓN DE GRAFOS MASIVOS (**GEPHI**)



Para determinar **color, tamaño, y opciones de texto de nodos y aristas** con parámetros **únicos o escalables**.

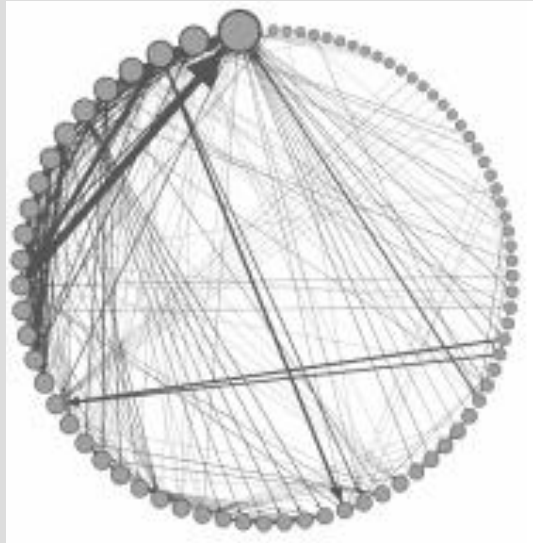


Para determinar el **posicionamiento de los nodos**:

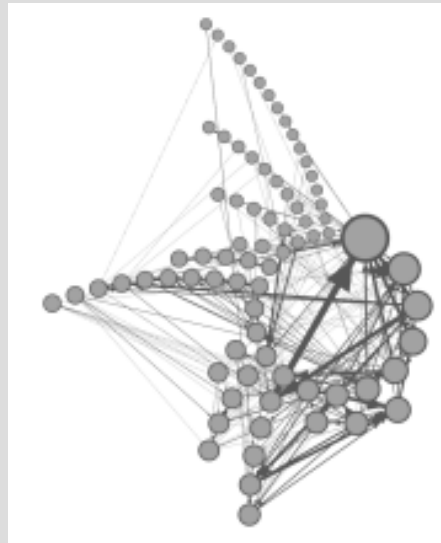
Algoritmos conceptuales: Circular, Radial, GeoLocation

Algoritmos de fuerza bruta: Force Atlas 2, Fruchterman-Reingold, Yifan-Hu, Yifan-Hu Multinivel, Open Ord

ALGORITMOS CONCEPTUALES



Circular

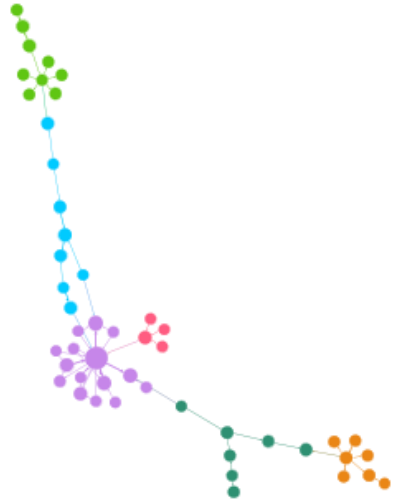


Radial



GeoLayout

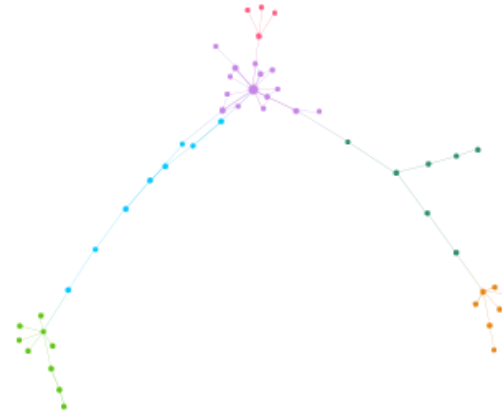
ALGORITMOS DE FUERZA BRUTA



Force Atlas 2



Fruchterman-
Reingold



Yifan Hu



Open Ord +
Noverlap

ANÁLISIS DE DATOS CON GEPHI

[HTTPS://GEPHI.ORG/](https://gephi.org/)

PRUEBAS ESTADÍSTICAS (MÁS) AVANZADAS PARA EL ARS EN DATOS MASIVOS

PRUEBAS ESTADÍSTICAS (**PAJEK**)

- Elementos más importantes
 - Networks: Archivos de red (.net)
 - Partitions: Propiedades de los nodos representables en números reales (.clu)
 - Vectors: Propiedades de los nodos representables en números reales e irreales (.vec)
- Los elementos se pueden combinar entre sí (OPERATIONS) para descubrir propiedades de las redes (e.g. homofilia-heterofilia)
- Dos archivos de un mismo elemento pueden combinarse entre sí (PARTITIONS, VECTORS) para la evaluación de aspectos de la red (e.g. V de Crammer entre particiones)

ALGORITMOS CLAVE EN ARS MASIVO (COHESIÓN)

- ~~Densidad (n° de aristas como proporción respecto al máximo posible)~~

- Centralidad de grado (de entrada y de salida)

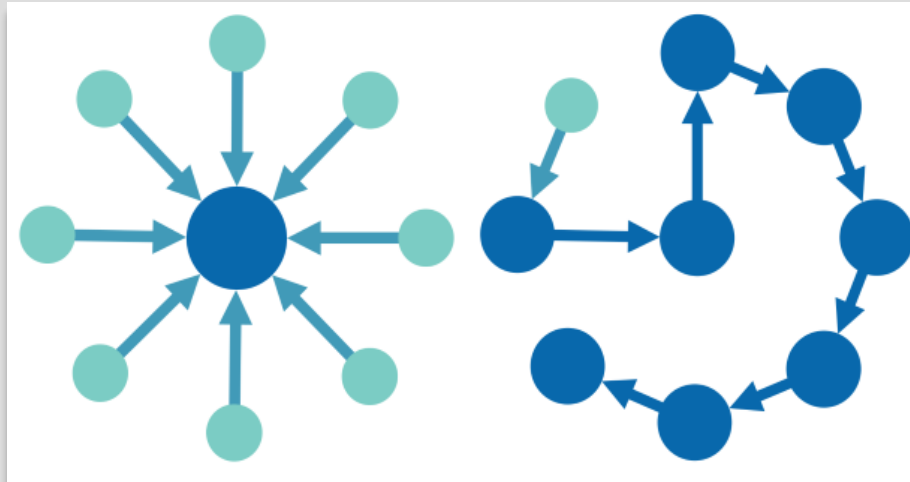
- Centralización de grado (de entrada y de salida)

$$C_D = \frac{\sum_{i=1}^n [C_D(p^*) - C_D(p_i)]}{[(P - 1)(P - 2)]}$$

- Componentes fuertes y débiles (semitutas y rutas)

- Núcleos-k (subgrafos máximamente conectados)

EL CASO DE LA CENTRALIZACIÓN DE GRADO



9 ACTORES
8 CONEXIONES

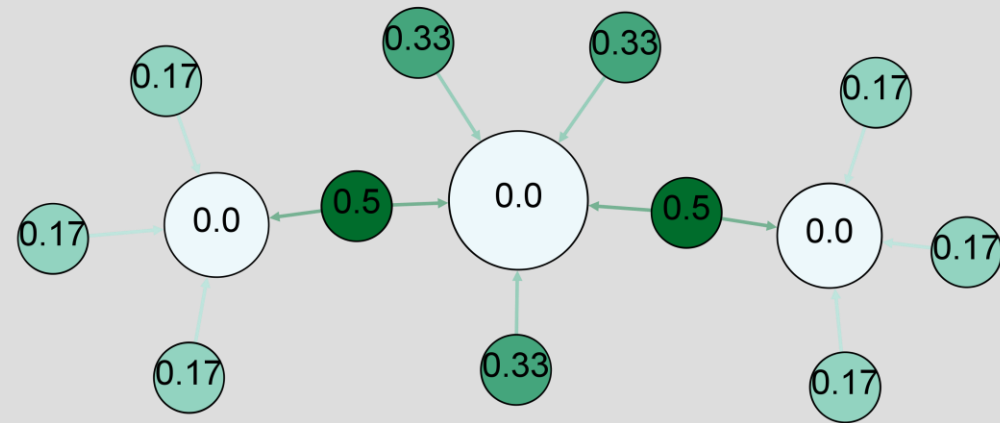
CENTRALIZACIÓN DE
GRADO DE ENTRADA
100%

9 ACTORES
8 CONEXIONES

CENTRALIZACIÓN DE
GRADO DE ENTRADA
2%

EL CASO DE LA MEDIACIÓN

- ~~Distribución de distancias~~
- ~~Centralidad de Intermediación~~
- HITS (Autoridades y Hubs)



ESTRATEGIAS DE DEPURACIÓN TRADICIONALES, SEMÁNTICAS Y ALGORÍTMICAS

DEPURACIÓN TRADICIONAL

Recode into Different Variables: Old and New Values

Old Value

☒ Value:

☐ System-missing

☐ System- or user-missing

☐ Range: through

☐ Range, LOWEST through value:

☒ Range, value through HIGHEST:

☐ All other values

New Value

☒ Value:

☐ System-missing

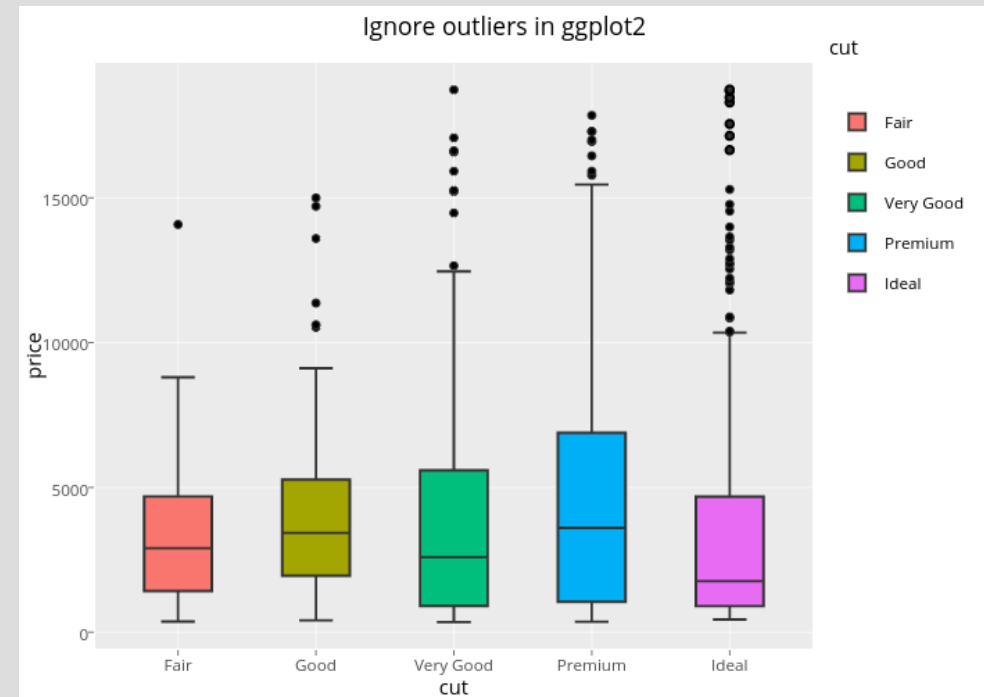
☐ Copy old value(s)

Old --> New:

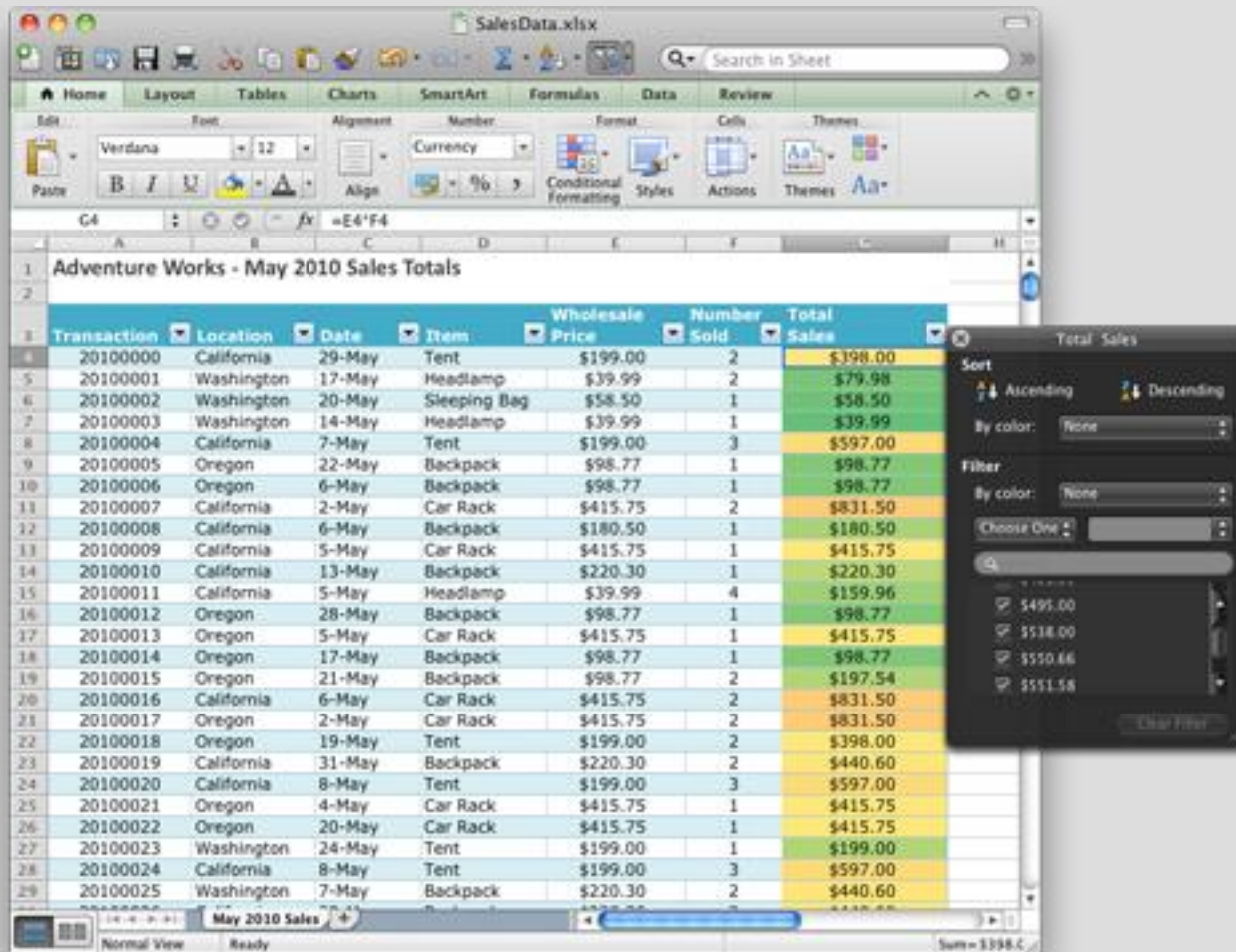
Lowest thru 5 --> 'Dissatisfactory'

☒ Output variables are strings Width:

☐ Convert numeric strings to numbers ('5' -> 5)

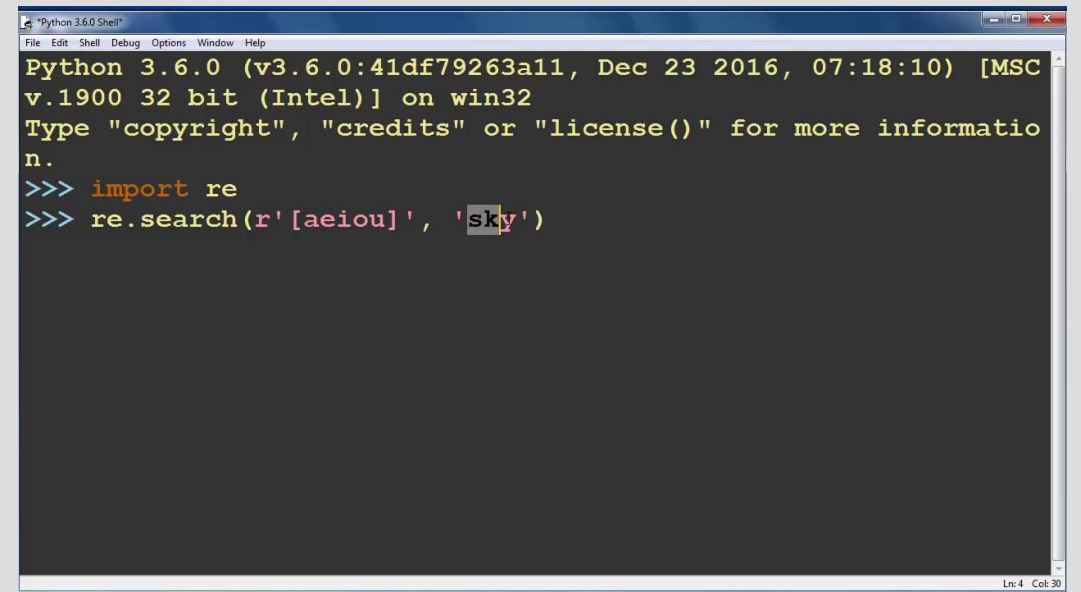


DEPURACIÓN SEMÁNTICA



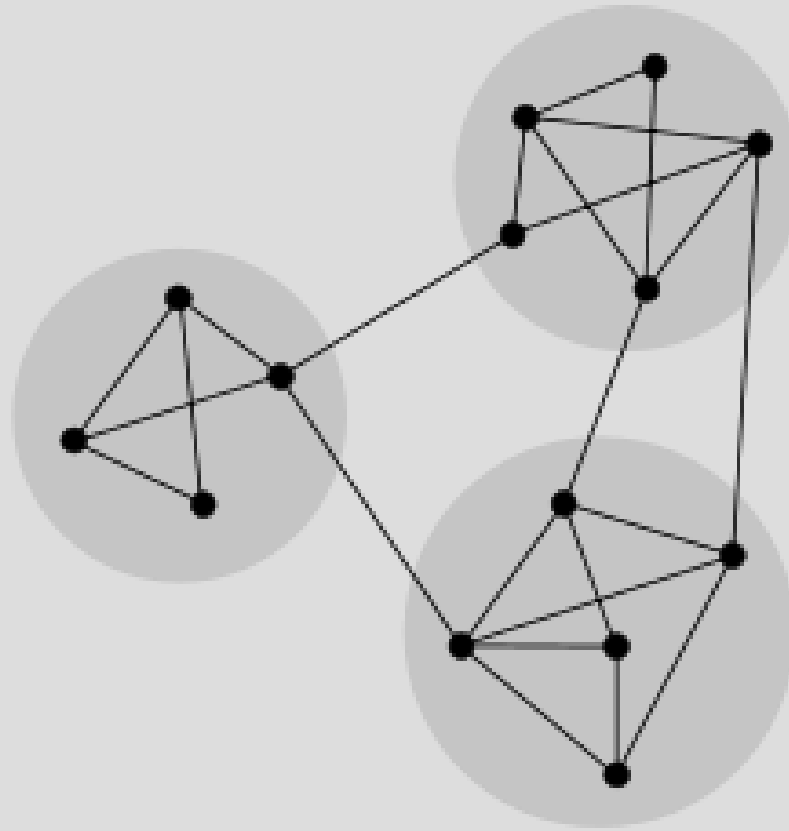
The screenshot shows an Excel spreadsheet titled "SalesData.xlsx" with a table of sales data. A sorting dialog box is open, showing the "Total Sales" column selected for sorting in descending order. The table has columns: Transaction, Location, Date, Item, Wholesale Price, Number Sold, and Total Sales. The data is sorted by Total Sales in descending order.

Transaction	Location	Date	Item	Wholesale Price	Number Sold	Total Sales
20100000	California	29-May	Tent	\$199.00	2	\$398.00
20100001	Washington	17-May	Headlamp	\$39.99	2	\$79.98
20100002	Washington	20-May	Sleeping Bag	\$58.50	1	\$58.50
20100003	Washington	14-May	Headlamp	\$39.99	1	\$39.99
20100004	California	7-May	Tent	\$199.00	3	\$597.00
20100005	Oregon	22-May	Backpack	\$98.77	1	\$98.77
20100006	Oregon	6-May	Backpack	\$98.77	1	\$98.77
20100007	California	2-May	Car Rack	\$415.75	2	\$831.50
20100008	California	6-May	Backpack	\$180.50	1	\$180.50
20100009	California	5-May	Car Rack	\$415.75	1	\$415.75
20100010	California	13-May	Backpack	\$220.30	1	\$220.30
20100011	California	5-May	Headlamp	\$39.99	4	\$159.96
20100012	Oregon	28-May	Backpack	\$98.77	1	\$98.77
20100013	Oregon	5-May	Car Rack	\$415.75	1	\$415.75
20100014	Oregon	17-May	Backpack	\$98.77	1	\$98.77
20100015	Oregon	21-May	Backpack	\$98.77	2	\$197.54
20100016	California	6-May	Car Rack	\$415.75	2	\$831.50
20100017	Oregon	2-May	Car Rack	\$415.75	2	\$831.50
20100018	Oregon	19-May	Tent	\$199.00	2	\$398.00
20100019	California	31-May	Backpack	\$220.30	2	\$440.60
20100020	California	8-May	Tent	\$199.00	3	\$597.00
20100021	Oregon	4-May	Car Rack	\$415.75	1	\$415.75
20100022	Oregon	20-May	Car Rack	\$415.75	1	\$415.75
20100023	Washington	24-May	Tent	\$199.00	1	\$199.00
20100024	California	8-May	Tent	\$199.00	3	\$597.00
20100025	Washington	7-May	Backpack	\$220.30	2	\$440.60

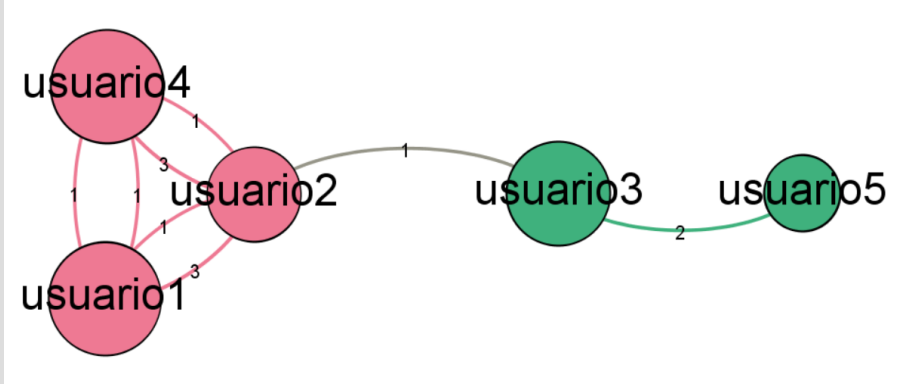


```
Python 3.6.0 (v3.6.0:41df79263a11, Dec 23 2016, 07:18:10) [MSC
v.1900 32 bit (Intel)] on win32
Type "copyright", "credits" or "license()" for more informatio
n.
>>> import re
>>> re.search(r'[aeiou]', 'sky')
```


DEPURACIÓN ALGORÍTMICA



CRUCE Y MODELAJE DE DATOS MASIVOS



USUARIO	GRADO DE ENTRADA	GRADO DE SALIDA	MODULARIDAD	BETWEENNES S
usuario1	4	2	0	0
usuario2	2	7	0	0,16
usuario3	3	0	1	0
usuario4	4	2	0	0
usuario5	0	2	1	0

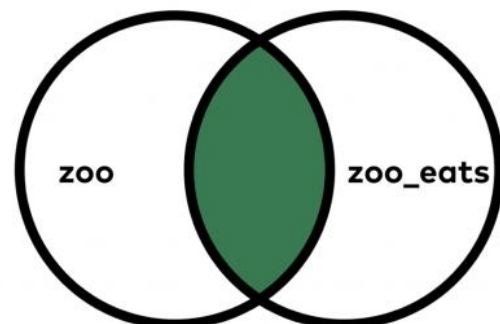


USUARIO	TUIT
usuario1	Me encanta el surf!
usuario1	En serio, me encanta el surf!!!!
usuario1	#Surflovers
usuario2	RT: @usuario1 Me encanta el surf!
usuario2	Oye @usuario4 dice @usuario1 que le gusta el surf, como a tí!
usuario3	Surfo de anginas y no he podido ir a trabajar
usuario1	RT: @usuario2 Oye @usuario4 dice @usuario1 que le gusta el surf, como a tí!
usuario5	RT: @usuario3 Surfo de anginas y no he podido ir a trabajar
usuario3	Perdón, no quise decir *surfo sino SUFRO
usuario4	.@usuario1 deberíamos ir a surfear juntos un día de estos, a ver si se apunta @usuario2
usuario4	@usuario4 a mí ni me metáis en vuestras movidas surferas... que estoy muy bien casa
usuario5	RT @usuario3: Perdón, no quise decir *surfo sino SUFRO
usuario2	@usuario4 @usuario1 iros a surfear con @usuario3 que dice que surfea hasta anginas!!!

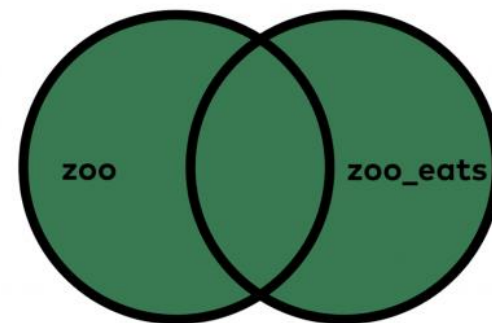


GRADO DE ENTRADA	GRADO DE SALIDA	MODULARIDAD	BETWEENNES S
4	2	0	0
4	2	0	0
4	2	0	0
2	7	0	0,16
2	7	0	0,16
3	0	1	0
4	2	0	0
0	2	1	0
3	0	1	0
4	2	0	0
2	7	0	0,16
0	2	1	0
2	7	0	0,16

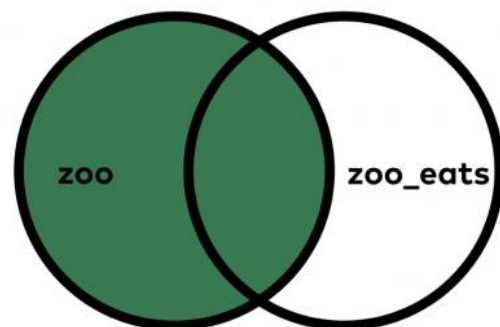
INNER



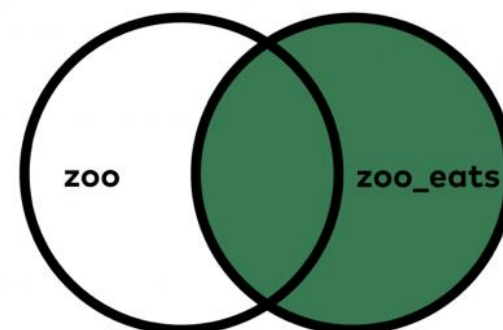
OUTER



LEFT



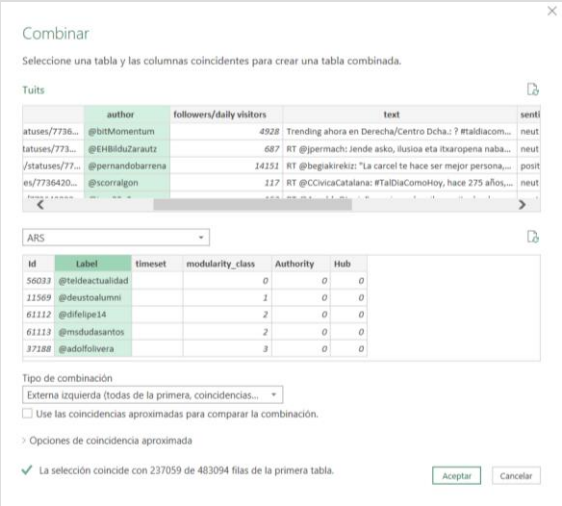
RIGHT





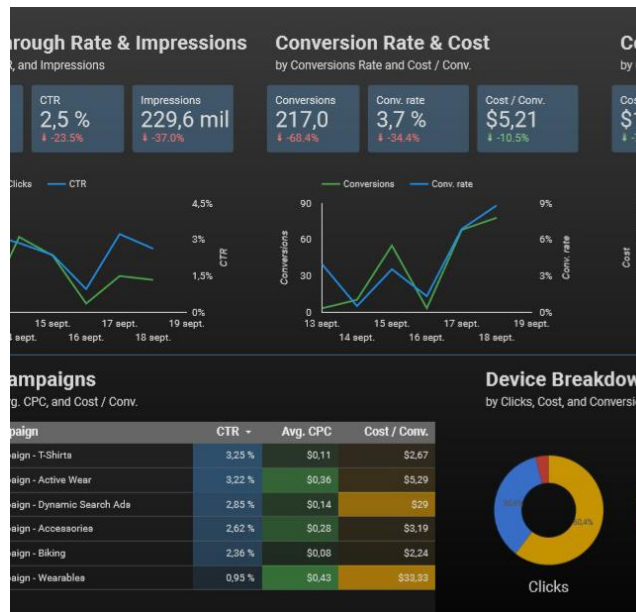
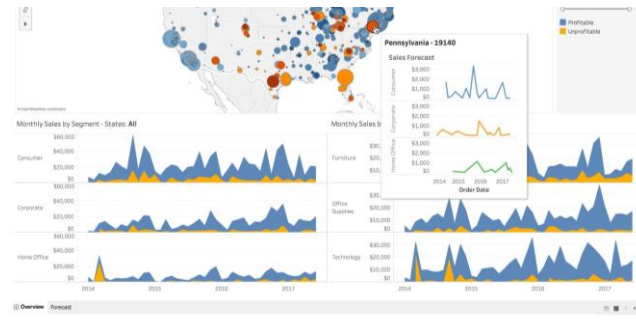
```
SELECT tweet_id, autor, seguidores, texto
FROM BD_tuits
INNER JOIN BD_ARS ON BD_1.autor = BD_2.autor;
```

USUARIO	TUIT	GRADO DE ENTRADA	GRADO DE SALIDA	MODULARIDAD	BETWEENNES S
usuario1	Me encanta el surf!	4	2	0	0
usuario1	En serio, me encanta el surf!!!!	4	2	0	0
usuario1	#Surflovers	4	2	0	0
usuario2	RT: @usuario1 Me encanta el surf!	2	7	0	0,16
usuario2	Oye @usuario4 dice @usuario1 que le gusta el surf, como a tí!	2	7	0	0,16
usuario3	Surfo de anginas y no he podido ir a trabajar	3	0	1	0
usuario1	RT: @usuario2 Oye @usuario4 dice @usuario1 que le gusta el surf, como a tí!	4	2	0	0
usuario5	RT: @usuario3 Surfo de anginas y no he podido ir a trabajar	0	2	1	0
usuario3	Perdón, no quise decir *surfo sino SUFRO	3	0	1	0
usuario4	.@usuario1 deberíamos ir a surfear juntos un día de estos, a ver si se apunta @usuario2	4	2	0	0
usuario2	@usuario4 a mí ni me metáis en vuestras movidas surferas... que estoy muy bien casa	2	7	0	0,16
usuario5	RT @usuario3: Perdón, no quise decir *surfo sino SUFRO	0	2	1	0
usuario2	@usuario4 @usuario1 iros a surfear con @usuario3 que dice que surfea hasta anginas!!!	2	7	0	0,16



USUARIO	TUIT	GRADO DE ENTRADA	GRADO DE SALIDA	MODULARIDAD	BETWEENNES S
usuario1	Me encanta el surf!	4	2	0	0
usuario1	En serio, me encanta el surf!!!!	4	2	0	0
usuario1	#Surflovers	4	2	0	0
usuario2	RT: @usuario1 Me encanta el surf!	2	7	0	0,16
usuario2	Oye @usuario4 dice @usuario1 que le gusta el surf, como a tí!	2	7	0	0,16
usuario3	Surfo de anginas y no he podido ir a trabajar	3	0	1	0
usuario1	RT: @usuario2 Oye @usuario4 dice @usuario1 que le gusta el surf, como a tí!	4	2	0	0
usuario5	RT: @usuario3 Surfo de anginas y no he podido ir a trabajar	0	2	1	0
usuario3	Perdón, no quise decir *surfo sino SUFRO	3	0	1	0
usuario4	.@usuario1 deberíamos ir a surfear juntos un día de estos, a ver si se apunta @usuario2	4	2	0	0
usuario2	@usuario4 a mí ni me metáis en vuestras movidas surferas... que estoy muy bien casa	2	7	0	0,16
usuario5	RT @usuario3: Perdón, no quise decir *surfo sino SUFRO	0	2	1	0
usuario2	@usuario4 @usuario1 iros a surfear con @usuario3 que dice que surfea hasta anginas!!!	2	7	0	0,16

BIG DATA MÁS ALLÁ DEL ARS

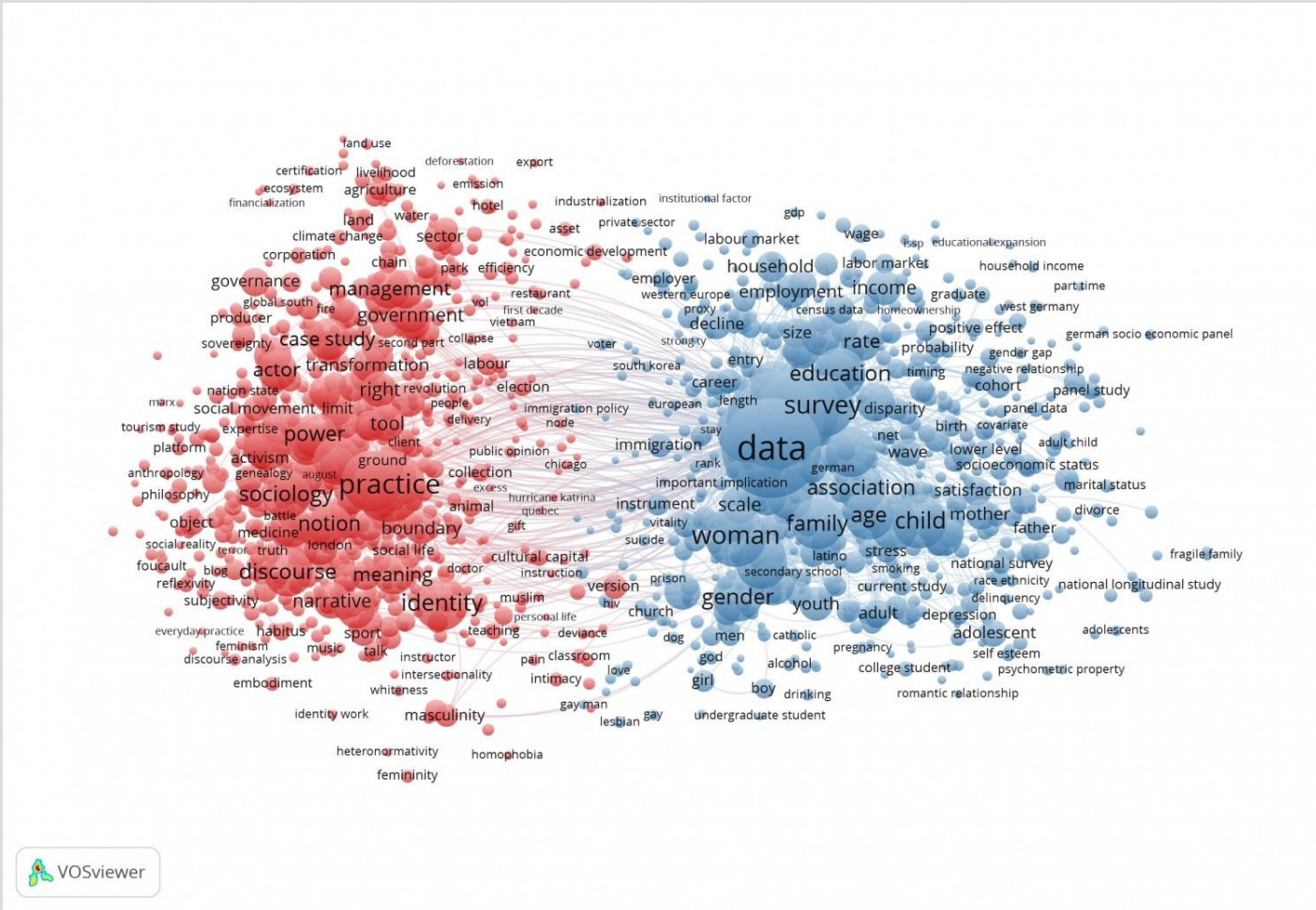


UNA VISUALIZACIÓN VALE MÁS QUE 1.000.000 CIFRAS

¿Y si la estadística clásica no fuera la perspectiva que exige el siglo XXI?



¿Y si la diferencia entre investigación CUANTI y CUALI ya no fuera tan importante?...
¿Y si, en realidad, no existiera?



¿Qué
conocimientos
aplicados en
Análisis de
Datos se os
pedirán
mañana?

Data analyst jobs

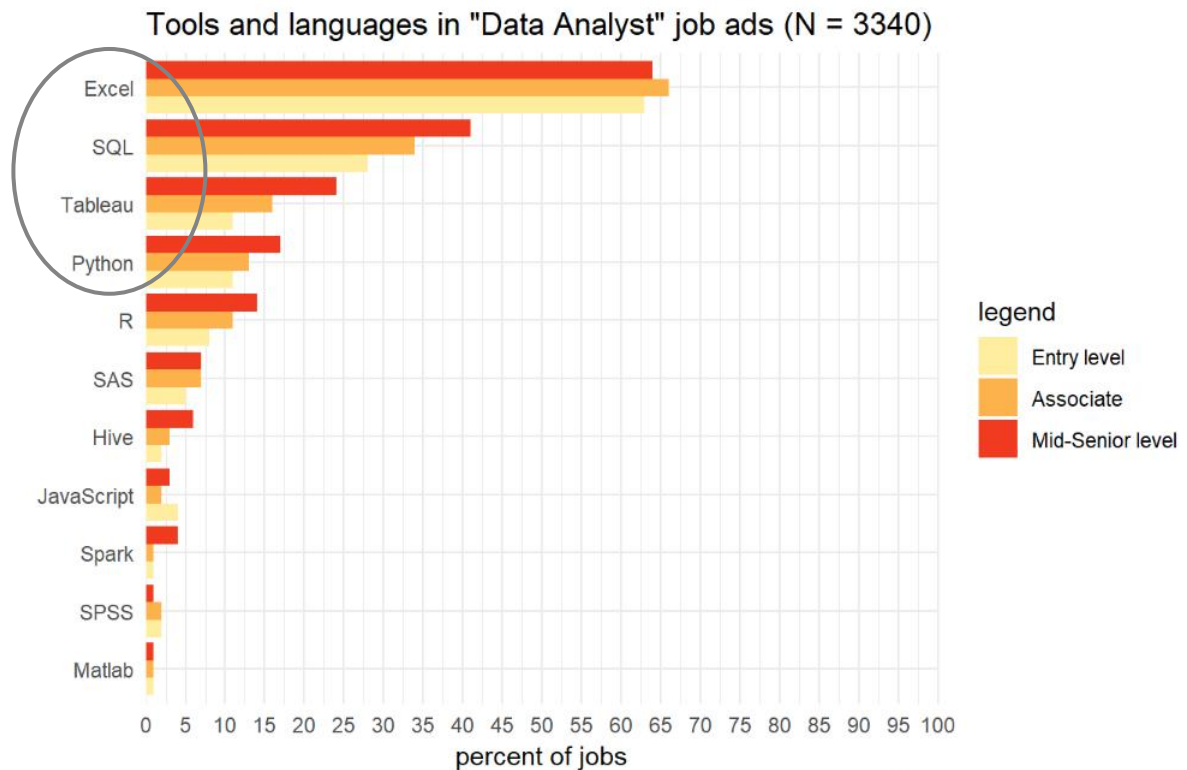
3340 data analyst job ads were included in this analysis.

Tools / languages

Other skills

Degrees

Disciplines



Jobs were scraped from LinkedIn in Sept 2019
Locations included NYC, SF, Seattle, Boston, and Toronto
Positions were Entry, Associate, and Senior levels