BIG DATA E INTELIGENCIA ARTIFICIAL PARA LAS CIENCIAS SOCIALES. **CONCEPTOS Y HERRAMIENTAS.**

JORDI MORALES I GRAS

eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

CONTENIDOS DEL TALLER

Introducción al paradigma de los Datos Masivos.

- De los retos técnicos (tecnológicos) a los retos interpretativos (científico-sociales).
- Características principales del nuevo datascape o escenario de datos.
- Los nuevos elementos de la caja de herramientas de la Ciencia Social.

Principales herramientas de la Ciencia Social computacional.

- Nuevo software y bases de datos SQL y NoSQL.
- Procesamiento del lenguaje natural (PLN).
- Machine Learning (ML) o aprendizaje automático aplicado a la Ciencia Social.
- Análisis de redes sociales (ARS).

Métodos y técnicas de investigación con datos de los Social Media.

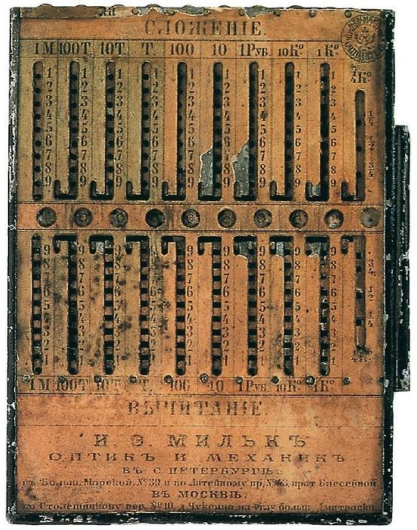
- Apis oficiales
- Web scrapping o raspado web.

BIG DATA: DE LOS RETOS TÉCNICOS A LOS RETOS INTERPRETATIVOS

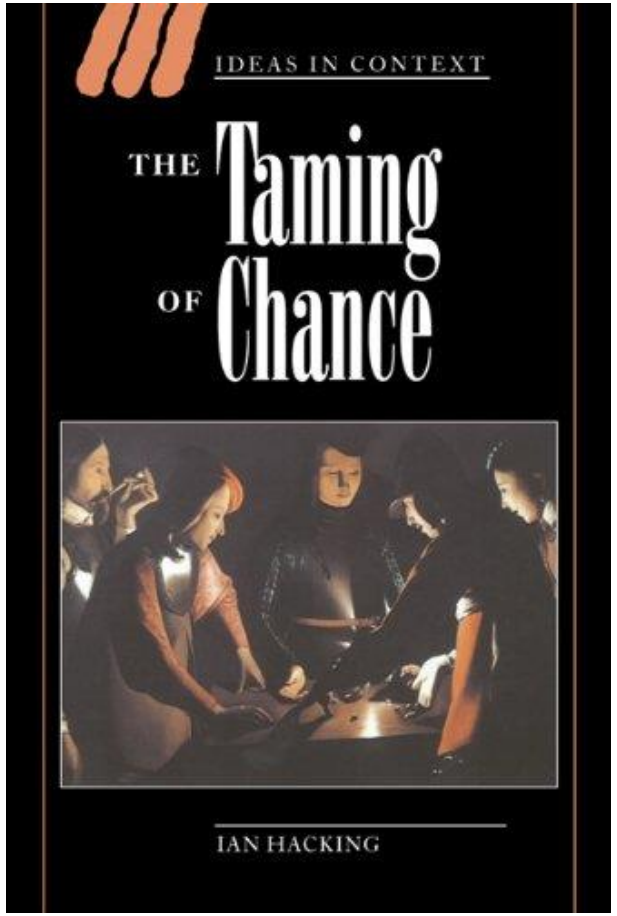
Alain Desrosières

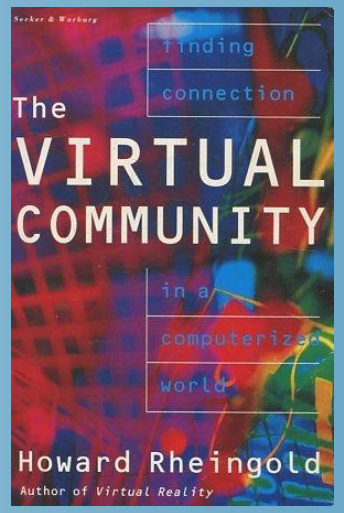
La politique des grands nombres

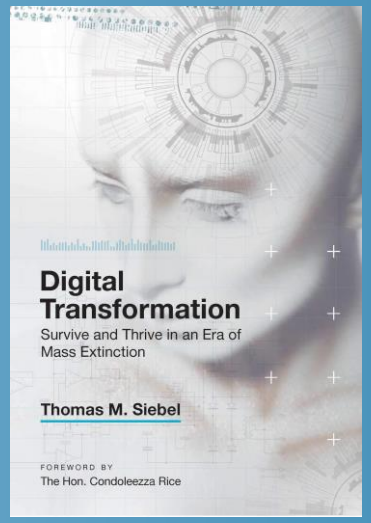
Histoire de la raison statistique



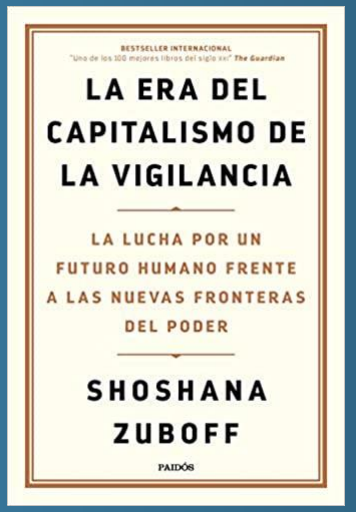
La Découverte/Poche

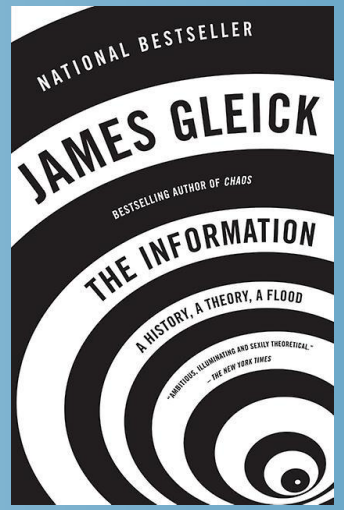


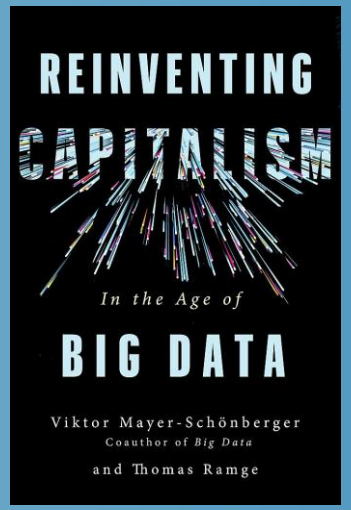

finding
connection
The
VIRTUAL
COMMUNITY
in a
computerized
world
Howard Rheingold
Author of Virtual Reality


Digital
Transformation
Survive and Thrive in an Era of
Mass Extinction
Thomas M. Siebel
FOREWORD BY
The Hon. Condoleezza Rice

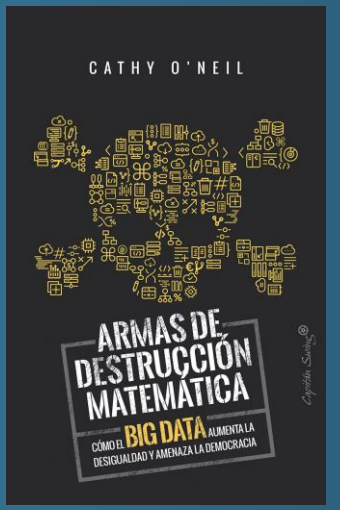

The Filter Bubble
What the
Internet
Is
Hiding
From
You
Eli Pariser


LA ERA DEL
CAPITALISMO DE
LA VIGILANCIA
LA LUCHA POR UN
FUTURO HUMANO FRENTE
A LAS NUEVAS FRONTERAS
DEL PODER
SHOSHANA
ZUBOFF
PAIDÓS

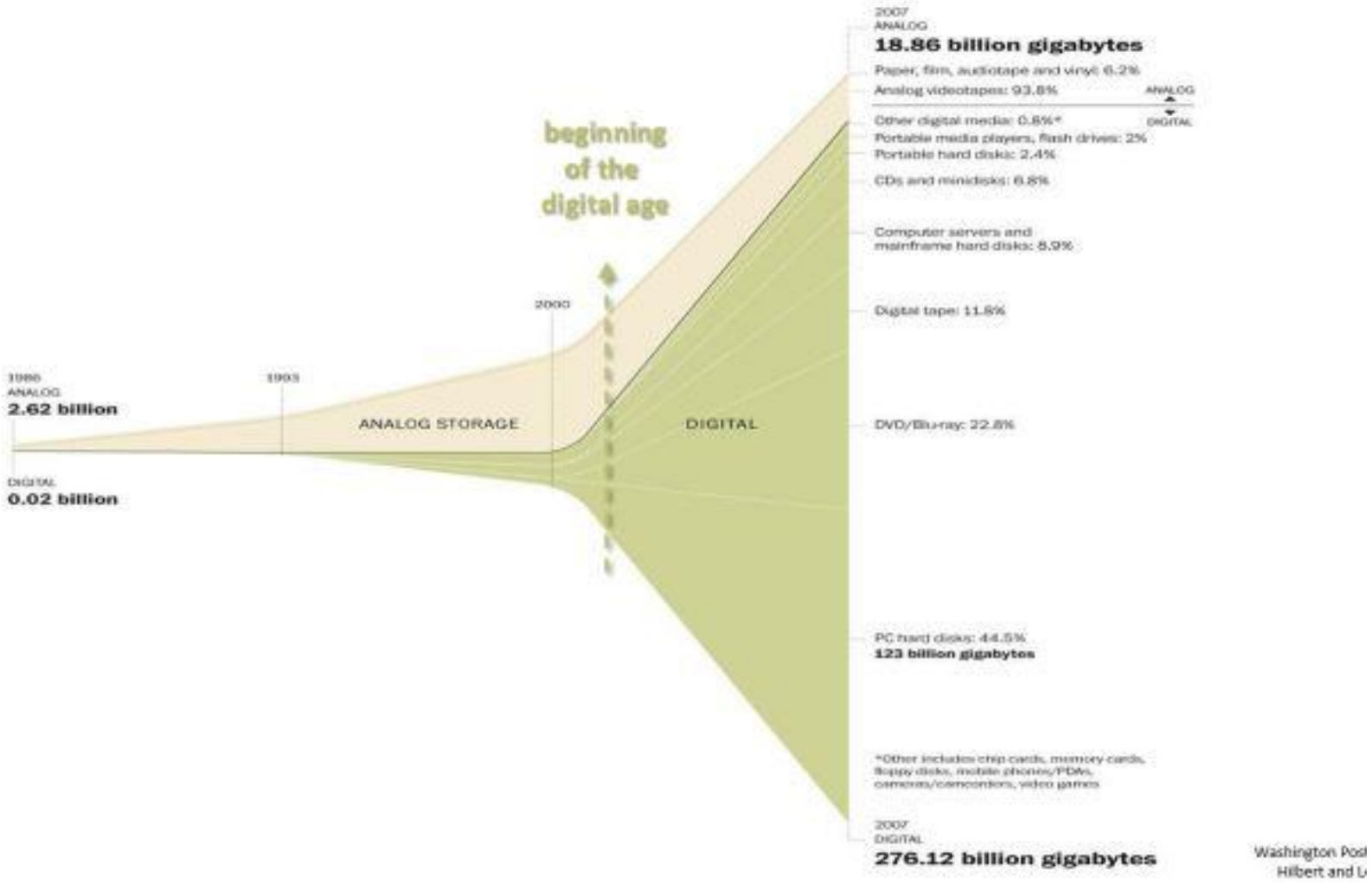

NATIONAL BESTSELLER
JAMES GLEICK
BESTSELLING AUTHOR OF CHAOS
THE INFORMATION
A HISTORY, A THEORY, A FLOOD


REINVENTING
CAPITALISM
In the Age of
BIG DATA
Viktor Mayer-Schönberger
Coauthor of Big Data
and Thomas Ramge

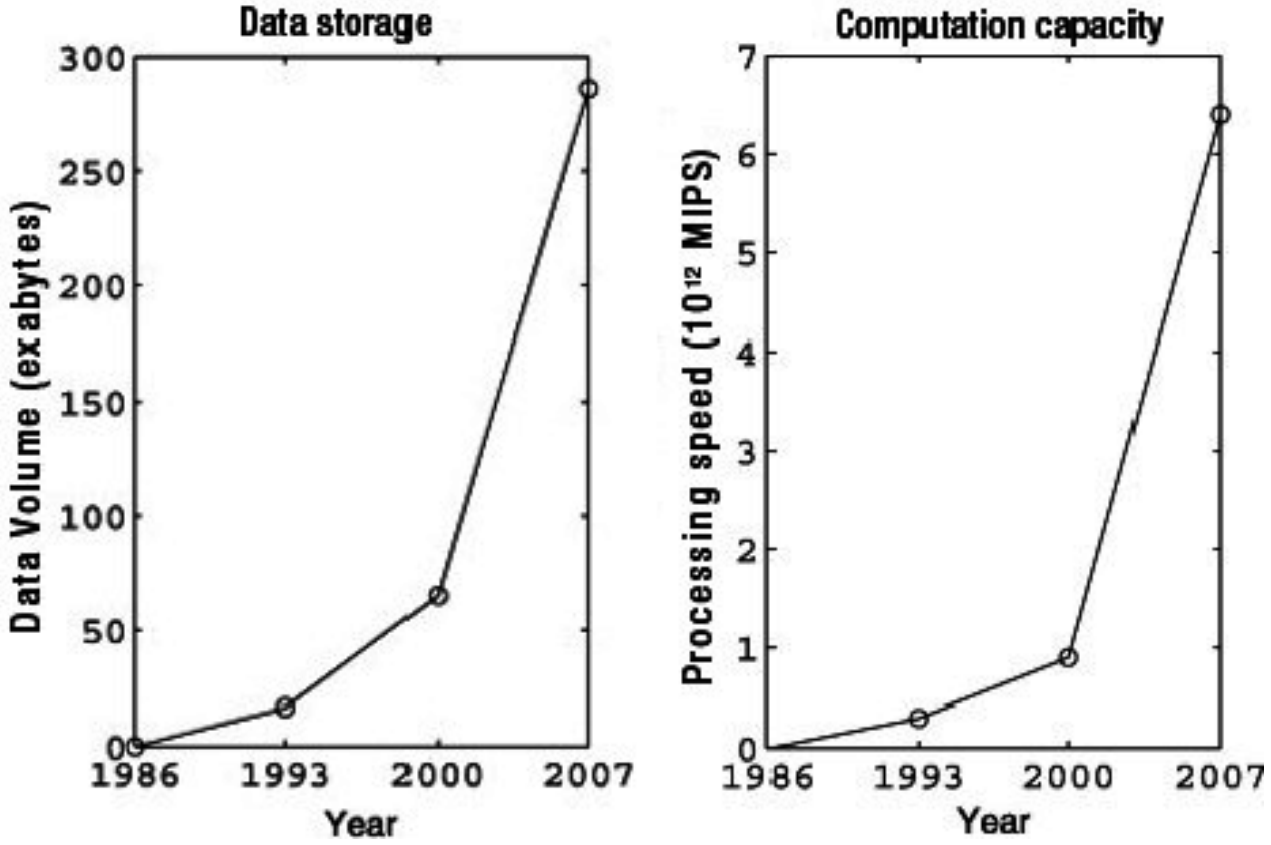

#digitalvertigo
how today's
online social revolution
is dividing, diminishing,
and disorienting us
@ajkeen
andrew keen, author of the cult of the amateur


CATHY O'NEIL
ARMAS DE
DESTRUCCIÓN
MATEMÁTICA
CÓMO EL BIG DATA AUMENTA LA
DESIGUALDAD Y AMENAZA LA DEMOCRACIA

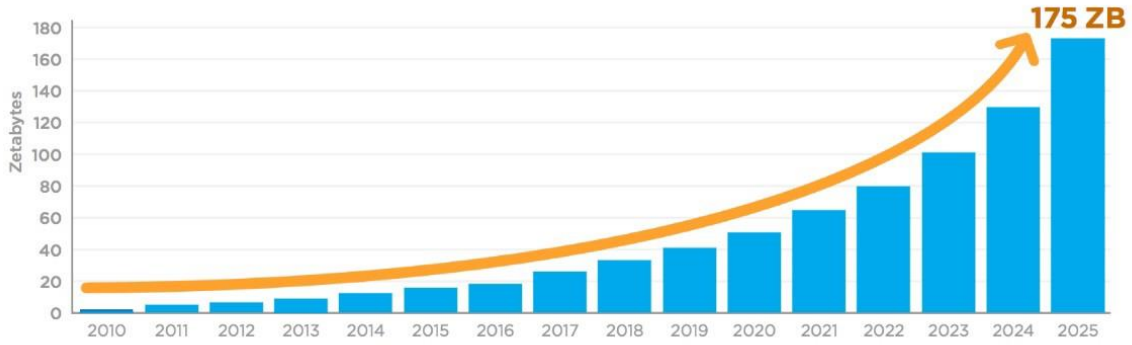

BIG DATA
with 8 V's
01
VOLUME
Can you find the information you are looking for?
02
VALUE
Can you find it when you most need it?
03
VERACITY
Are you dealing with information or disinformation?
04
VISUALISATION
Can you make sense at a glance? Does it trigger a decision?
05
VARIETY
Is a picture worth a thousand words in 70 languages? Is your information balanced?
06
VELOCITY
Information gains momentum and crises & opportunities evolve in real time. How is outlook for today?
07
VISCOSITY
Does it stick with you? Does it call for action?
08
VIRALITY
AHA to-go? Does it convey a message that can be pasted into a presentation or Instgrammed?



M. Hilbert and P. López, 2011. The world's technological capacity to store, communicate, and compute information.



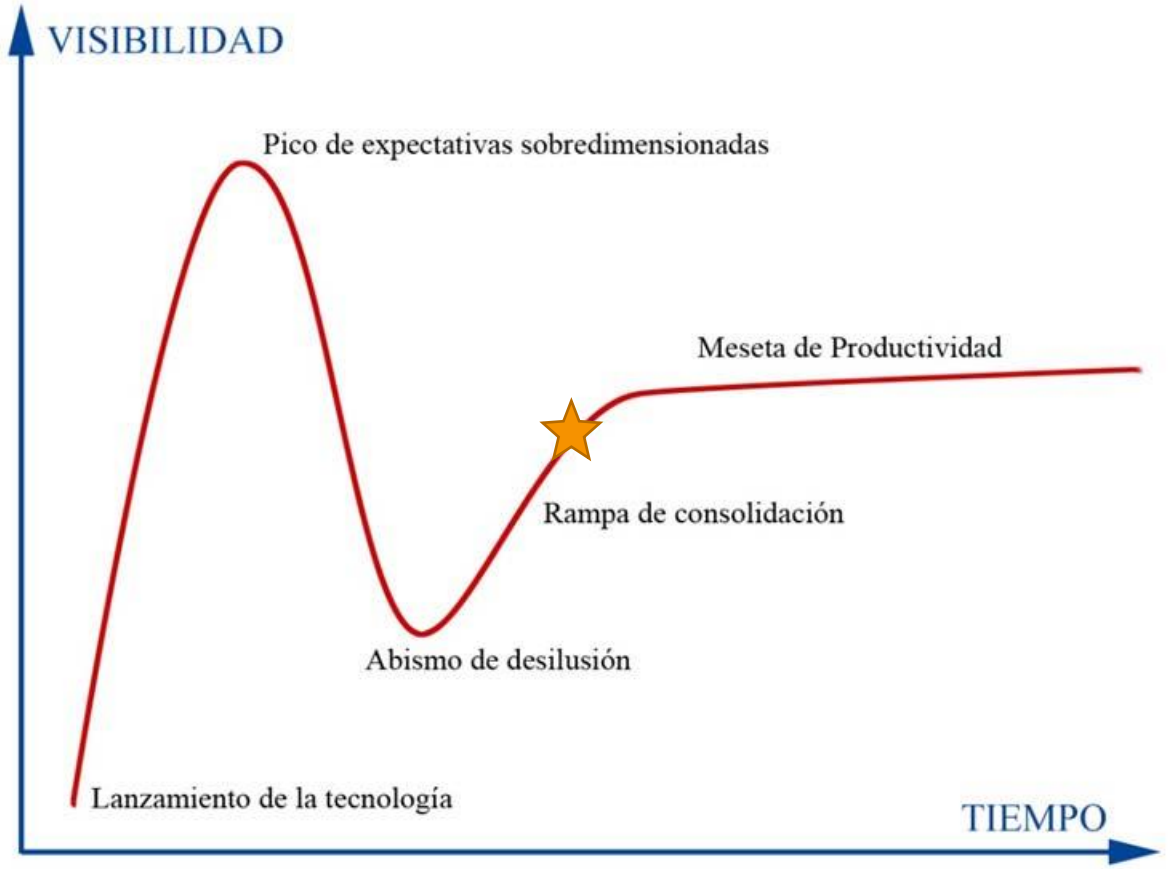
M. Hilbert and P. López, 2011. The world's technological capacity to store, communicate, and compute information.



Source: IDC Data Age 2025

WHAT IS A ZETTABYTE?

1,000,000,000,000 .....gigabyte
1,000,000,000,000 .....terabyte
1,000,000,000,000 .....petabyte
1,000,000,000,000 .....exabyte
1,000,000,000,000 .....zettabyte



Ciclo de sobreexplotación de les tecnologías de Gartner

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
- Crowdsource reporting for repairs to infrastructure
- Fulfill open records requests

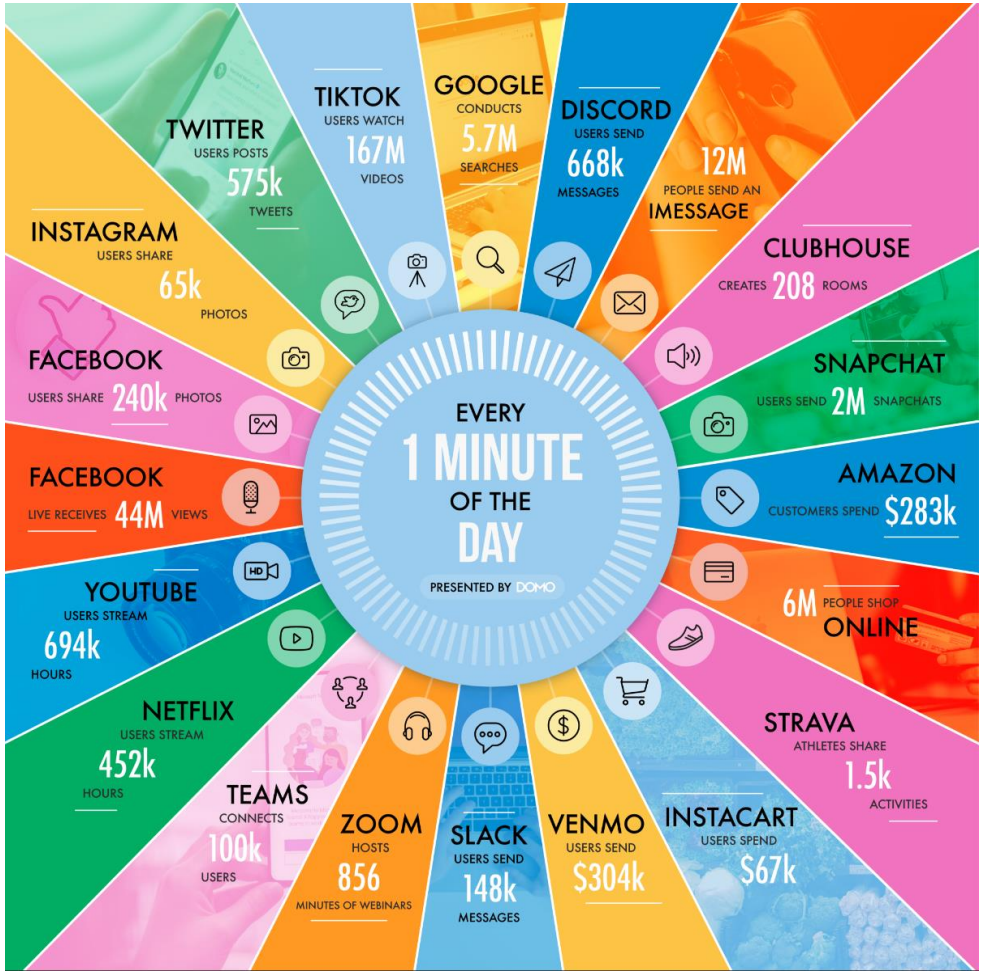
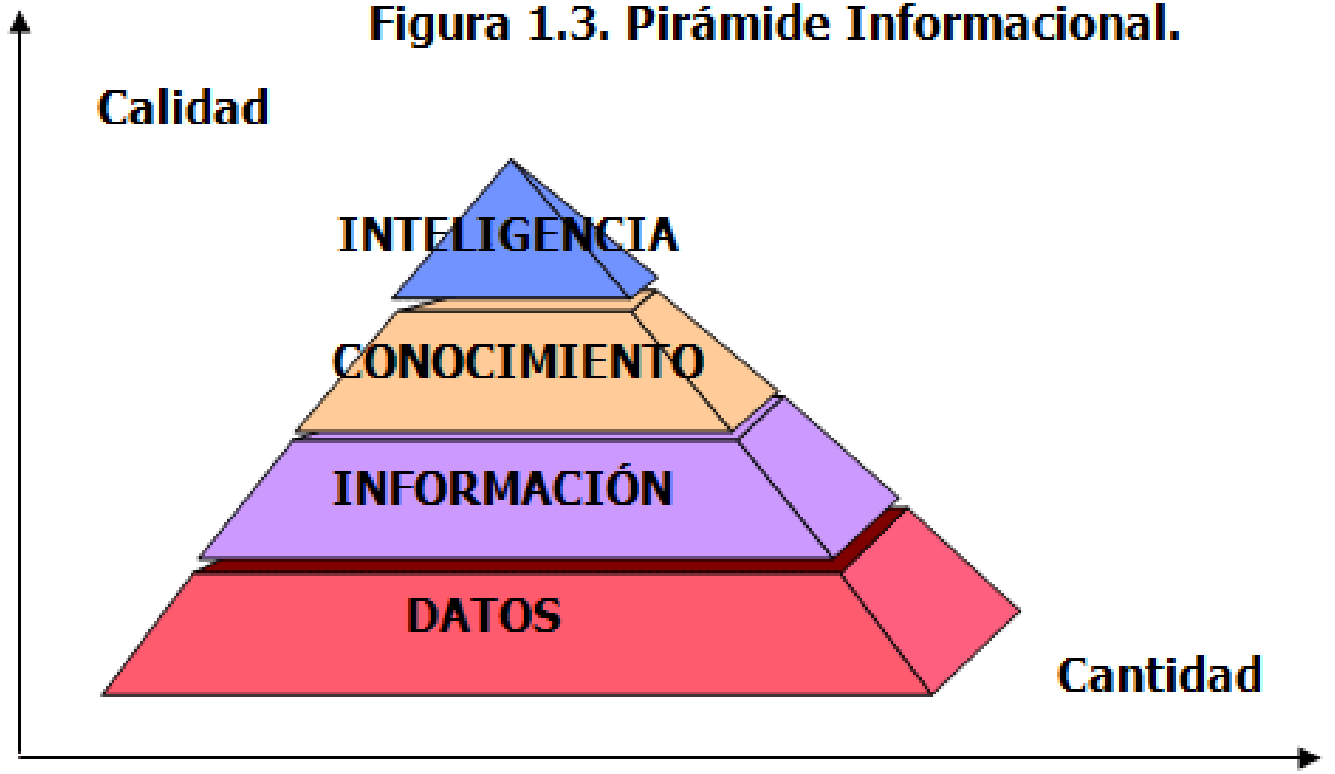
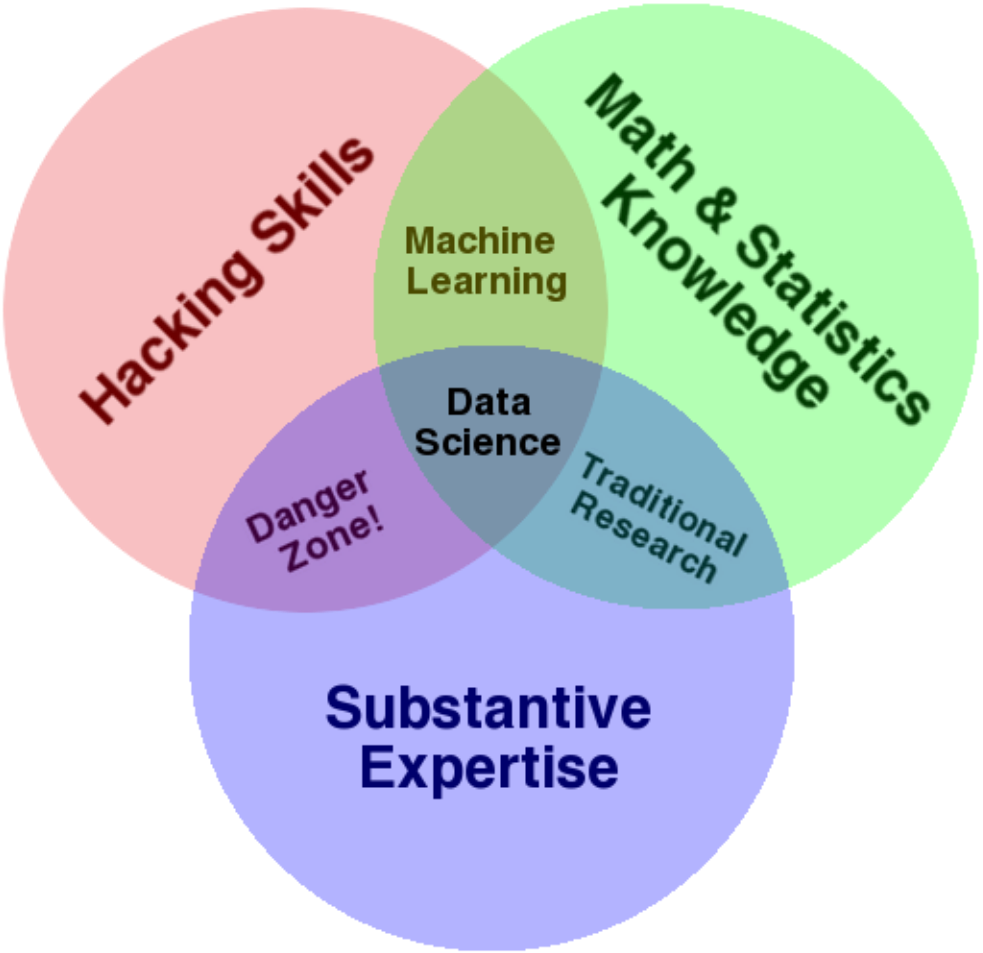

EVERY
1 MINUTE
OF THE
DAY
PRESENTED BY DOMO
TWITTER
USERS POSTS
575k
TWEETS
TIKTOK
USERS WATCH
167M
VIDEOS
GOOGLE
CONDUCTS
5.7M
SEARCHES
DISCORD
USERS SEND
668k
MESSAGES
12M
PEOPLE SEND AN
IMESSAGE
INSTAGRAM
USERS SHARE
65k
PHOTOS
CLUBHOUSE
CREATES 208 ROOMS
FACEBOOK
USERS SHARE 240k PHOTOS
SNAPCHAT
USERS SEND 2M SNAPCHATS
FACEBOOK
LIVE RECEIVES 44M VIEWS
AMAZON
CUSTOMERS SPEND $283k
YOUTUBE
USERS STREAM
694k
HOURS
6M PEOPLE SHOP
ONLINE
NETFLIX
USERS STREAM
452k
HOURS
STRAVA
ATHLETES SHARE
1.5k
ACTIVITIES
TEAMS
CONNECTS
100k
USERS
ZOOM
HOSTS
856
MINUTES OF WEBINARS
SLACK
USERS SEND
148k
MESSAGES
VENMO
USERS SEND
$304k
INSTACART
USERS SPEND
$67k

Figura 1.3. Pirámide Informacional.



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


Hacking Skills
Math & Statistics Knowledge
Substantive Expertise
Machine Learning
Data Science
Danger Zone!
Traditional Research

CARACTERÍSTICAS PRINCIPALES DEL NUEVO DATASCAPE O ESCENARIO DE DATOS.



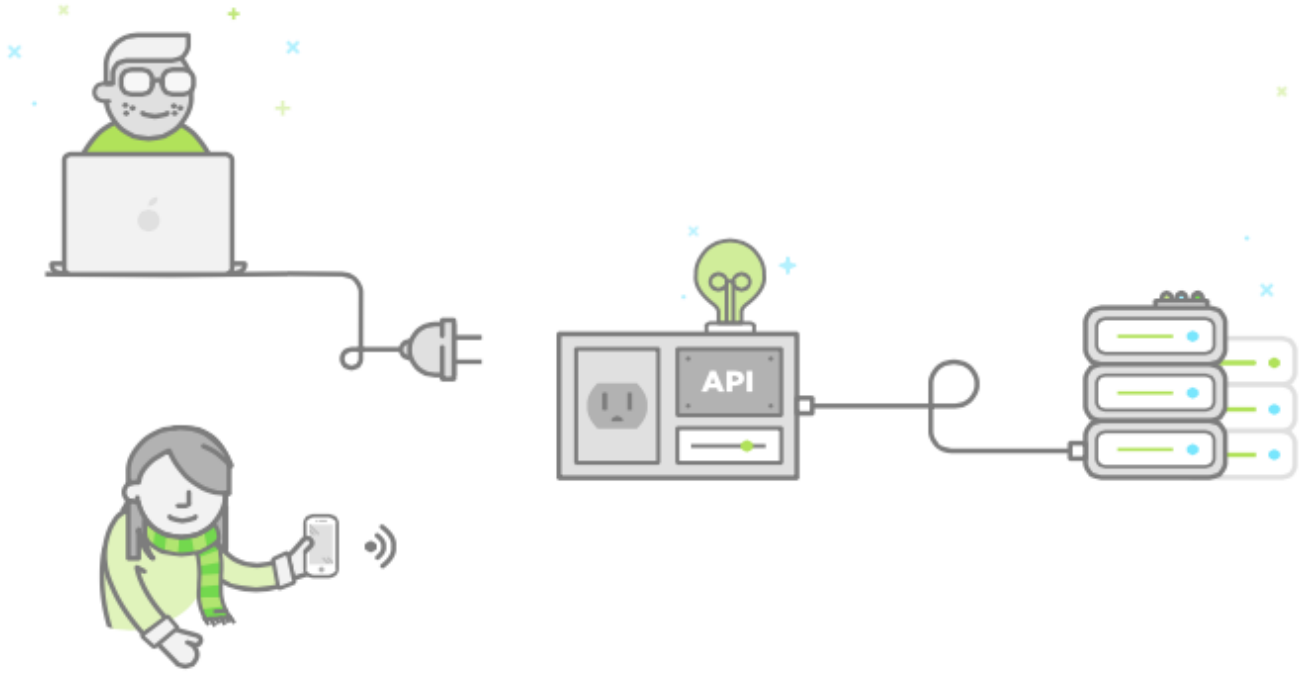
ESTRATEGIAS CLÁSICAS PARA LA OBTENCIÓN DE DATOS SOCIALES



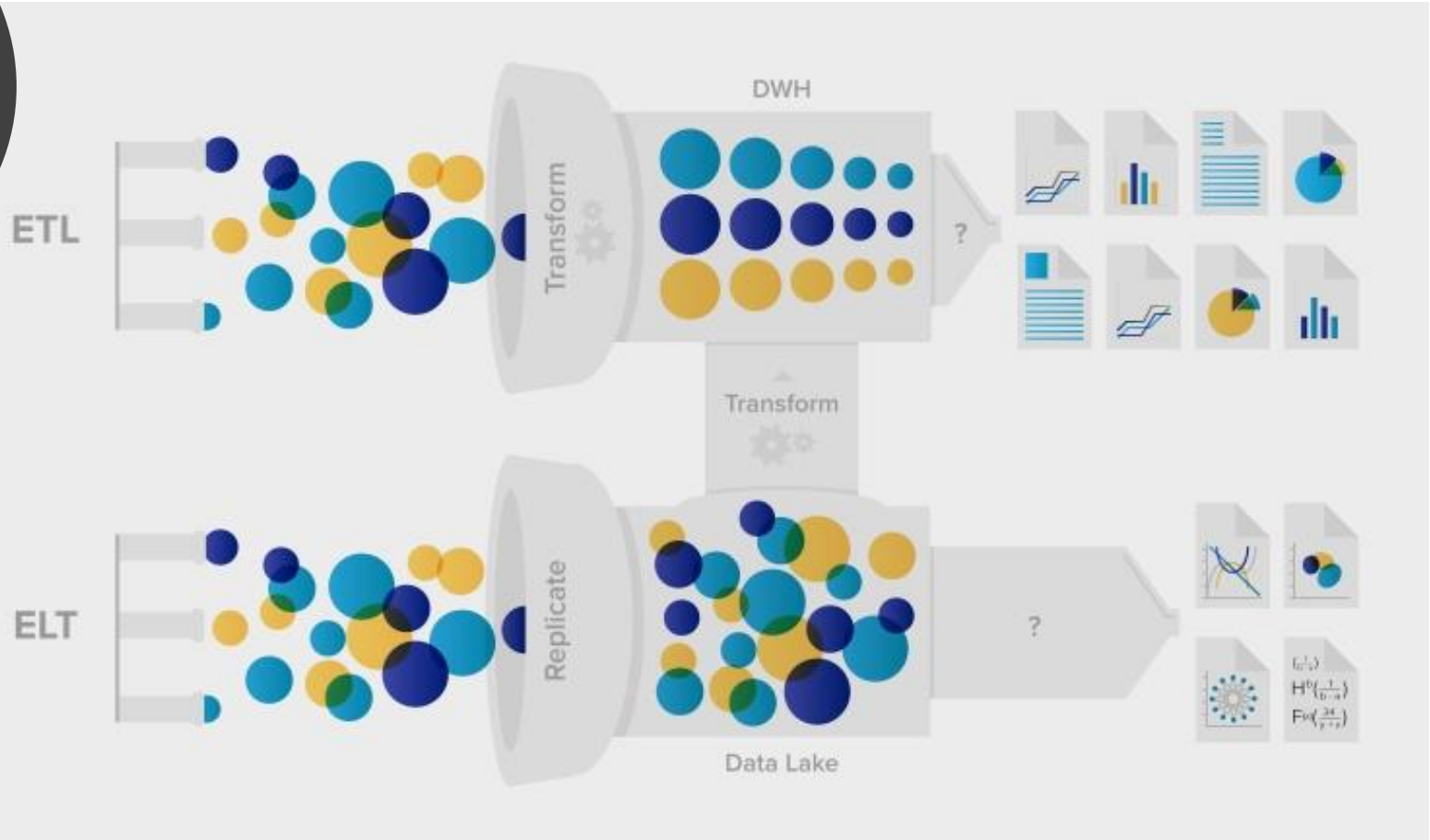
FUENTES DE DATOS MASIVOS

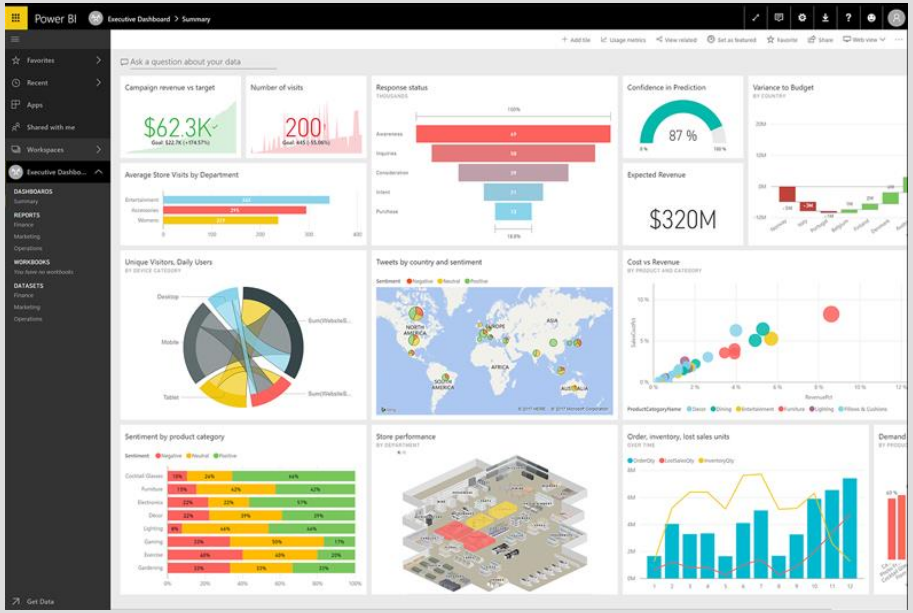
MEDIOS SOCIALES	Imagen, video, audio o texto de redes sociales virtuales
CLOUD	Público, privado o corporativo
WEB	Datos web, analytics
IoT	Sensores y dispositivos conectados
BASES DE DATOS	Datos secundarios tradicionales

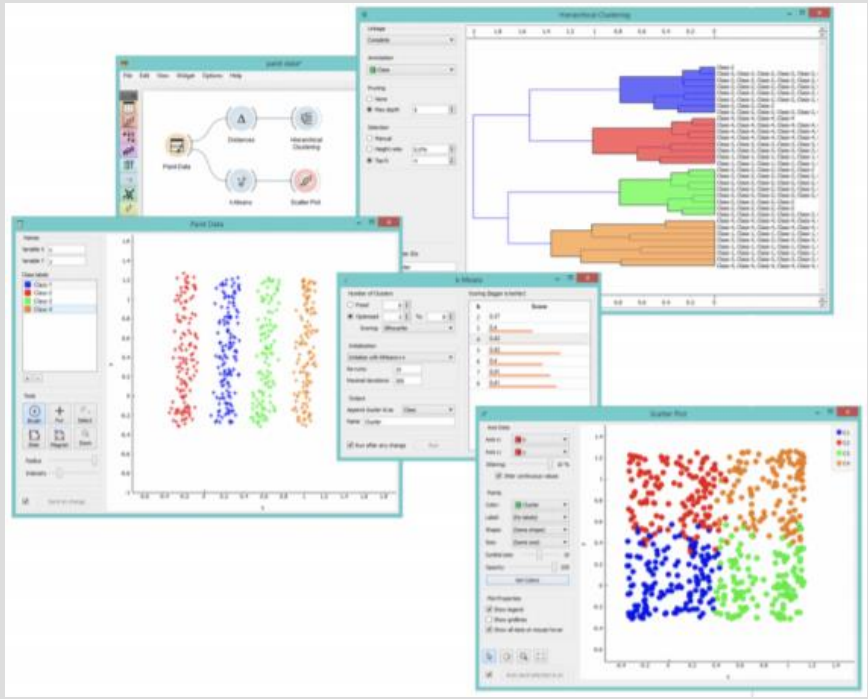
EL NUEVO ESCENARIO DE DATOS SOCIALES



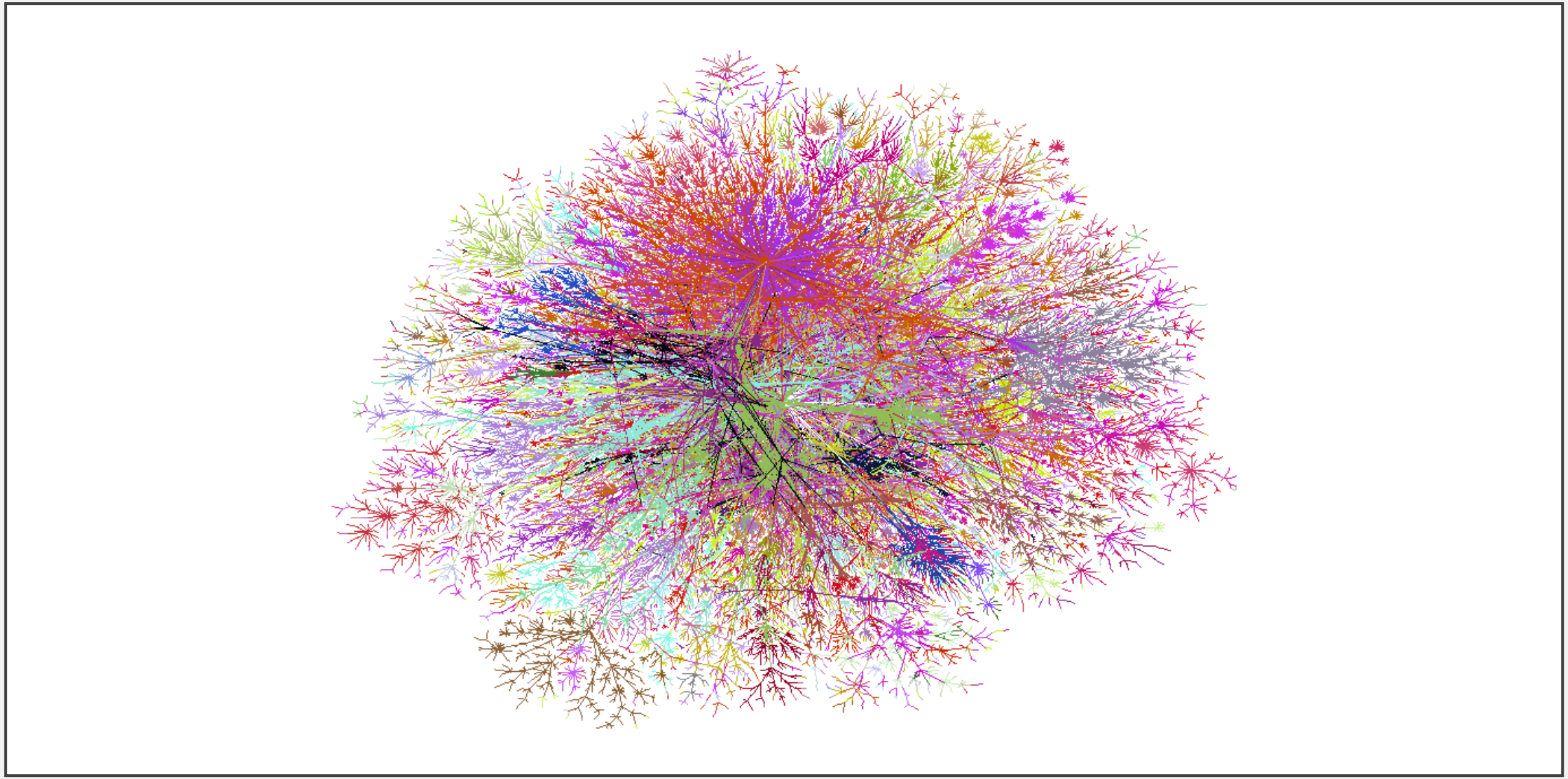
ETL VS ELT







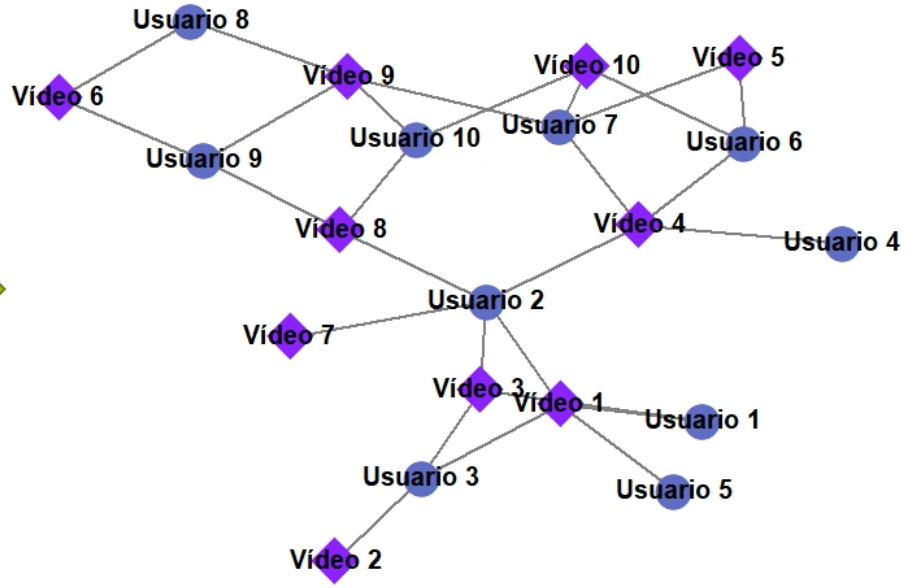
BI (ETL) VS. DATAMINING (ELT)



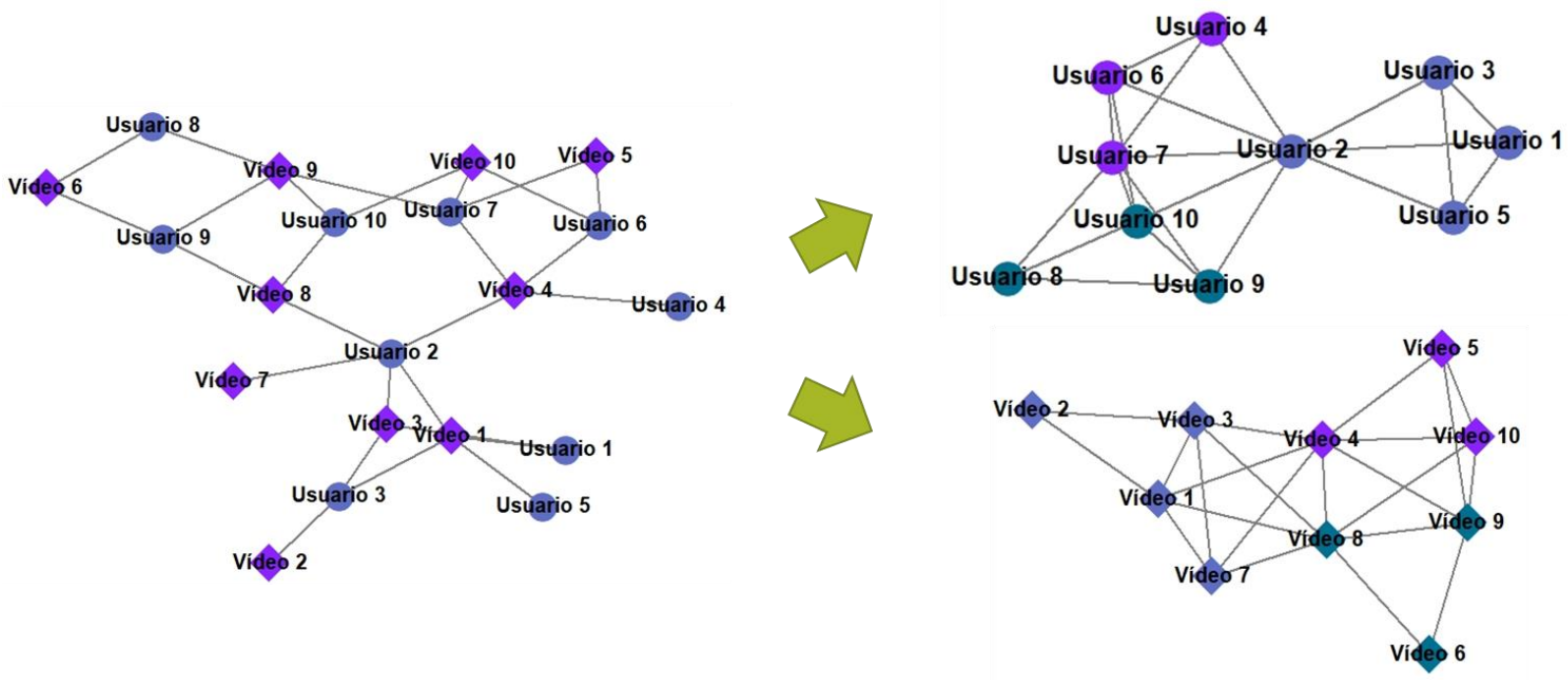
Vídeos

Usuaris

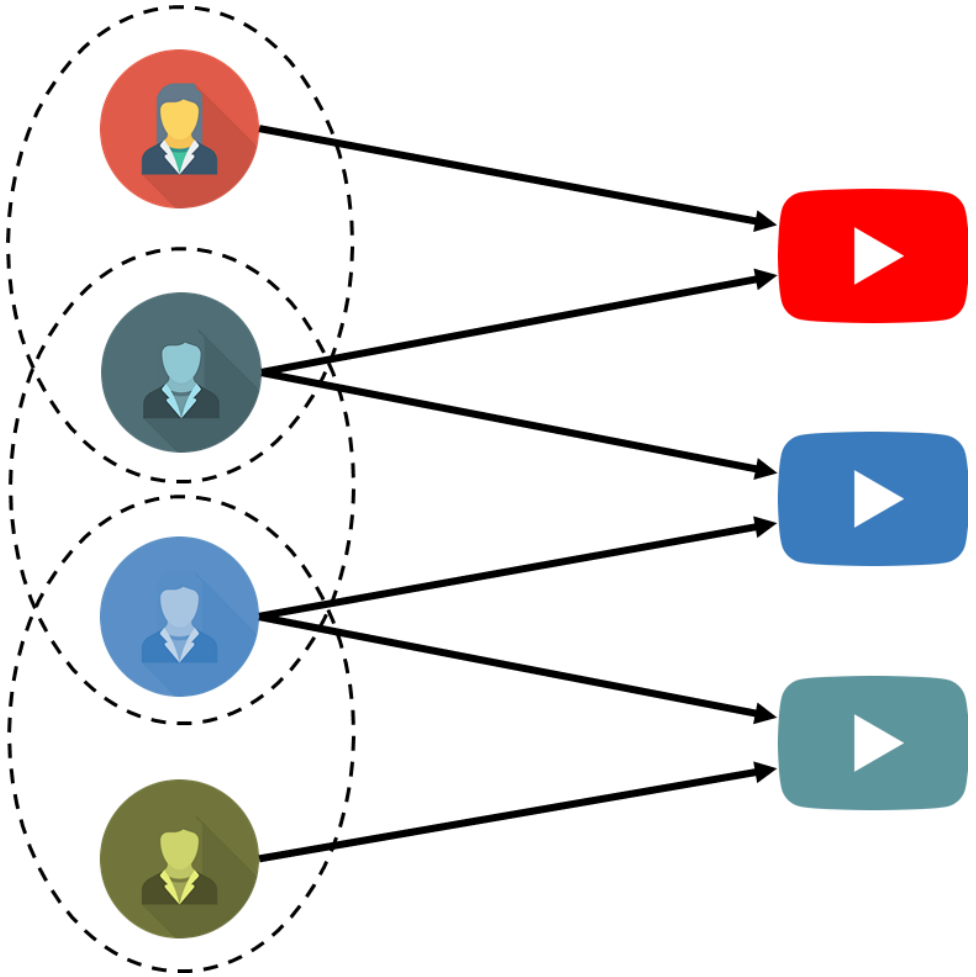
	1	2	3	4	5	6	7	8	9	10
1	x		x							
2	x		x	x			x	x		
3	x	x	x							
4				x						
5	x									
6				x	x					x
7				x	x				x	x
8						x			x	
9						x		x	x	
10								x	x	x



EJEMPLO: DE MATRIZ A RED



EJEMPLO: DE RED BIPARTITA A RED UNIPARTITA



NUEVAS HERRAMIENTAS PARA LA CIENCIA SOCIAL

Document 1

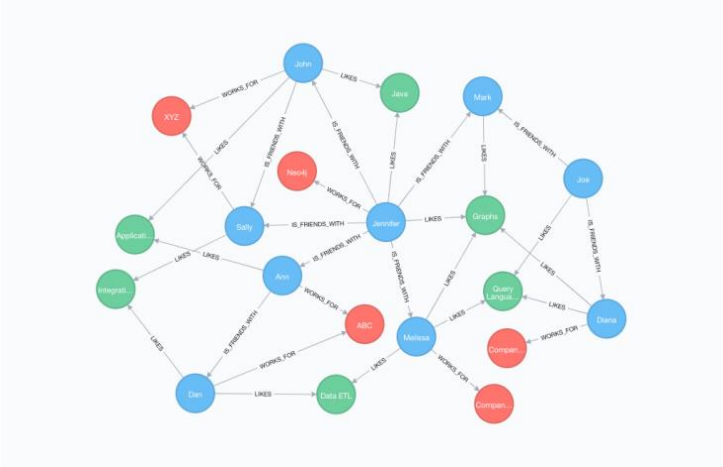
```
{
  "id": "1",
  "name": "John Smith",
  "isActive": true,
  "dob": "1964-30-08"
}
```

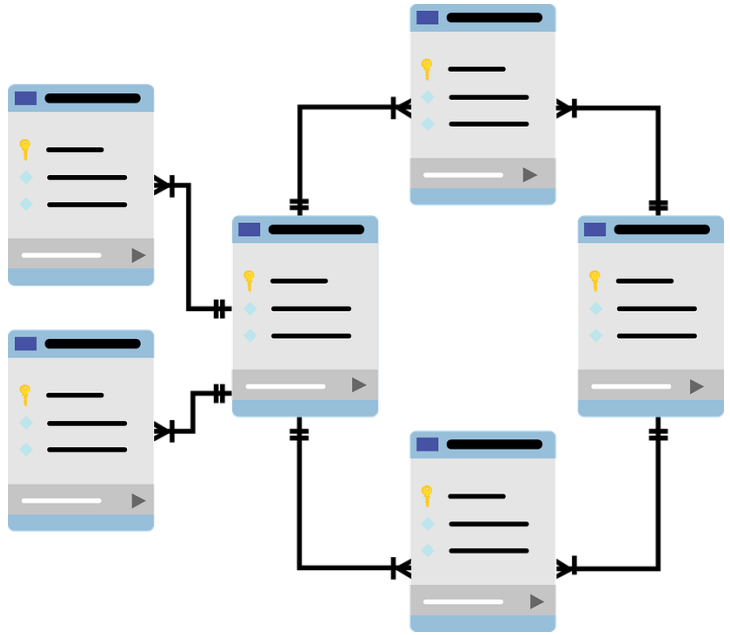
Document 2

```
{
  "id": "2",
  "fullName": "Sarah Jones",
  "isActive": false,
  "dob": "2002-02-18"
}
```

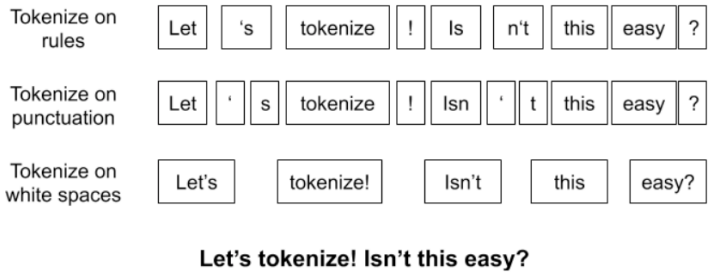
Document 3

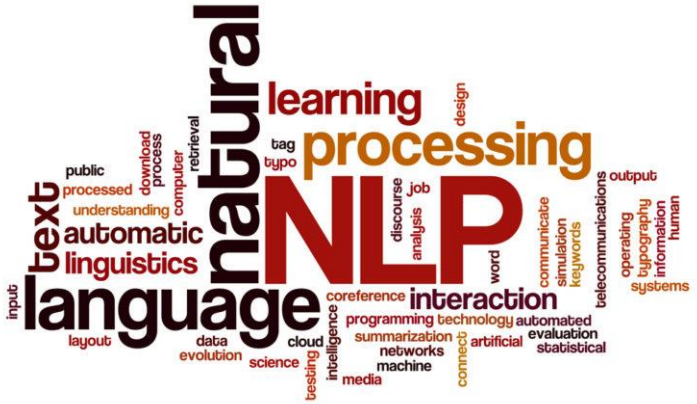
```
{
  "id": "3",
  "fullName":
  {
    "first": "Adam",
    "last": "Stark"
  },
  "isActive": true,
  "dob": "2015-04-19"
}
```

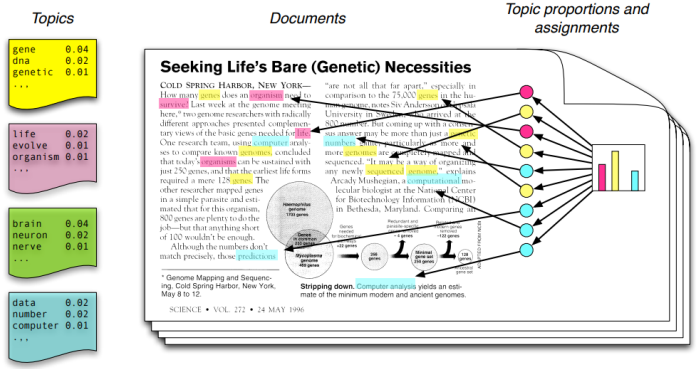




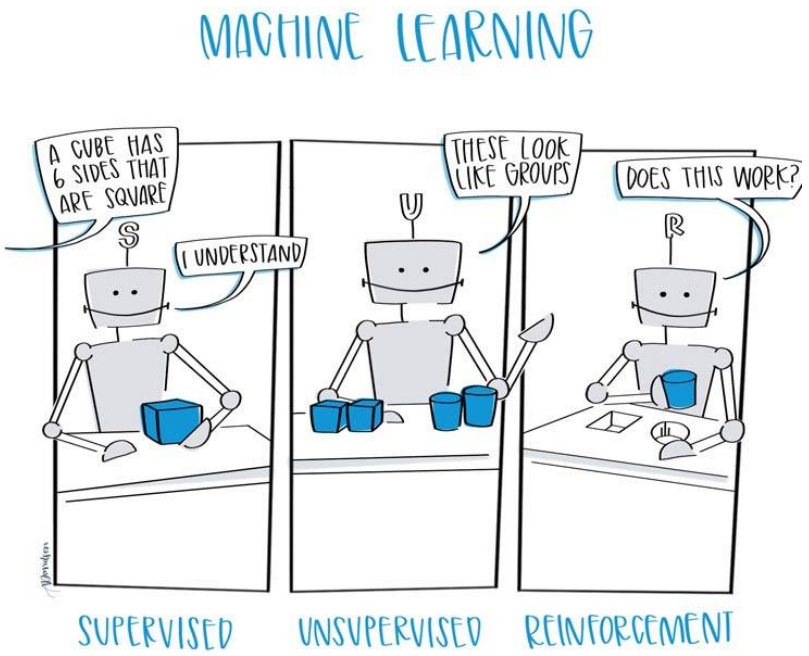
NUEVO SOFTWARE



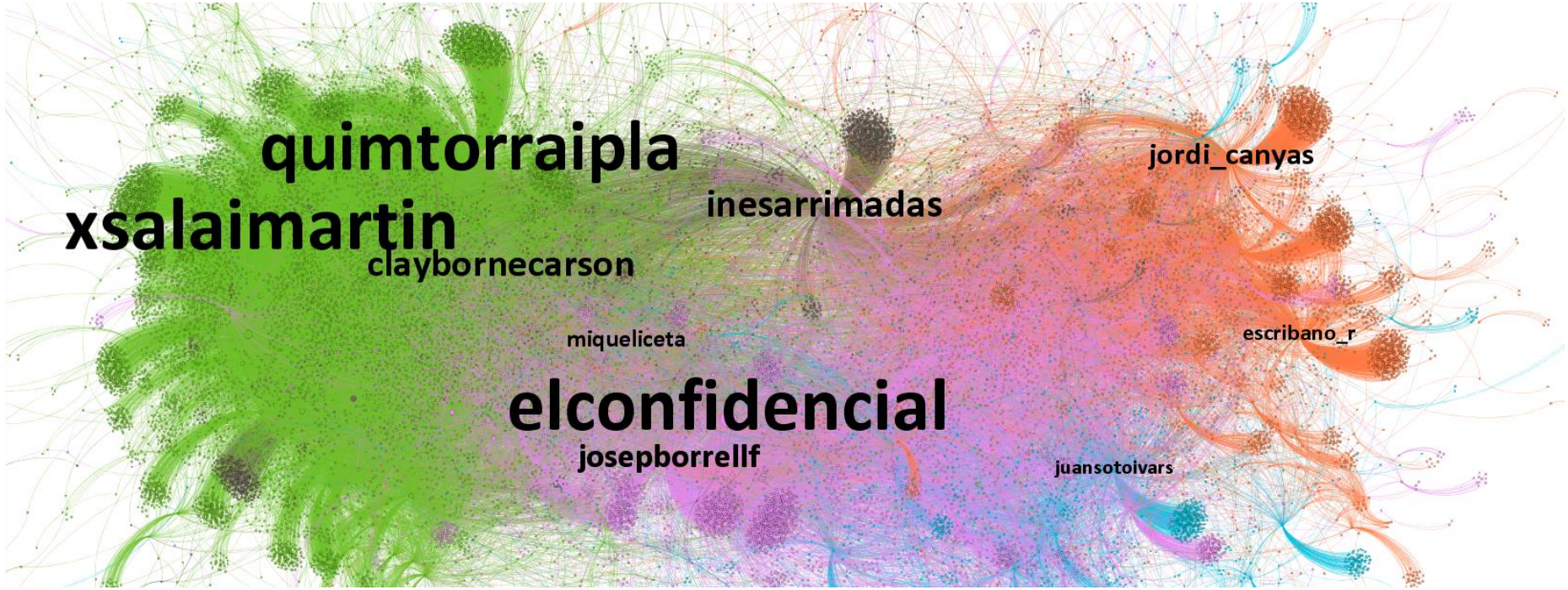




EL PROCESAMIENTO DEL LENGUAJE NATURAL



EL APRENDIZAJE AUTOMÁTICO O MACHINE LEARNING (ML)



EL ANÁLISIS DE REDES SOCIALES

NUEVO SOFTWARE Y BASES DE DATOS SQL Y NOSQL

COVID-19

Alarma en el Reino Unido por un fallo técnico en el recuento de contagios

• Unos 16.000 casos de coronavirus no fueron notificados a tiempo para rastrear los contactos

https://www.lavanguardia.com/internacional/20201005/483861404090/alarma-reino-unido-recuento-fallo-tecnico-contagios.html

SOFTWARE INADECUADO

↓↓↓

MALOS DATOS

↓↓↓

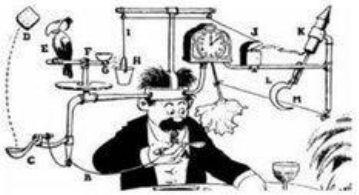
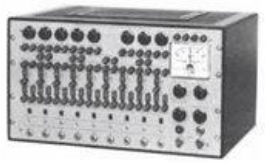
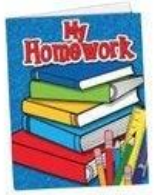
MALAS DECISIONES

Un **fallo técnico** que implicó que unos **16.000 casos de coronavirus** en el **Reino Unido** no fueran notificados a tiempo ha retrasado los esfuerzos del Gobierno británico para rastrear los contactos de esas personas que dieron positivo, informan este lunes los medios nacionales.

El error, según apuntan varios medios británicos, surgió desde un laboratorio que enviaba el reporte diario de los resultados de las PCR que realizaba en sus instalaciones en formato CSV (Coma Separated Values). Dicho formato es compatible con Excel, el programa que usa el PHE para indexar todos los casos. El archivo compartido con el departamento oficial incluía todo el histórico, no sólo los nuevos.

Día tras día, los responsables cargaban los nuevos datos al final del excel principal. Pero mientras que los archivos CSV pueden tener cualquier tamaño, los archivos de Microsoft Excel solo pueden tener una longitud de 1.048.576 filas. Cuando se abre un archivo CSV más largo que lo soportado en Excel, las filas inferiores se cortan y ya no se muestran, aunque el programa advierte al usuario que hay información que no cabe en el máximo establecido por Microsoft y, por ende, no será cargada en el documento.

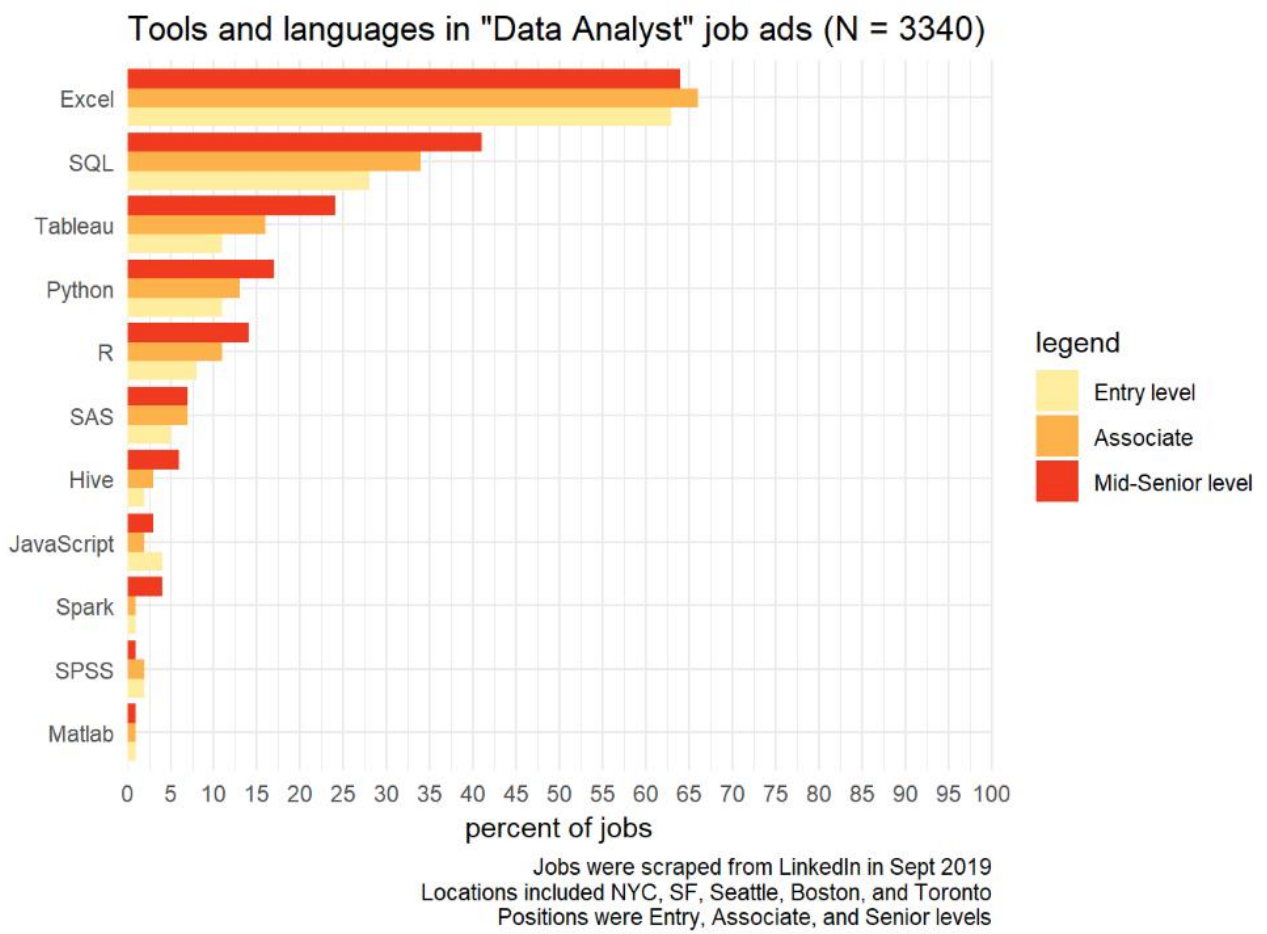
AS SEEN BY USERS OF …

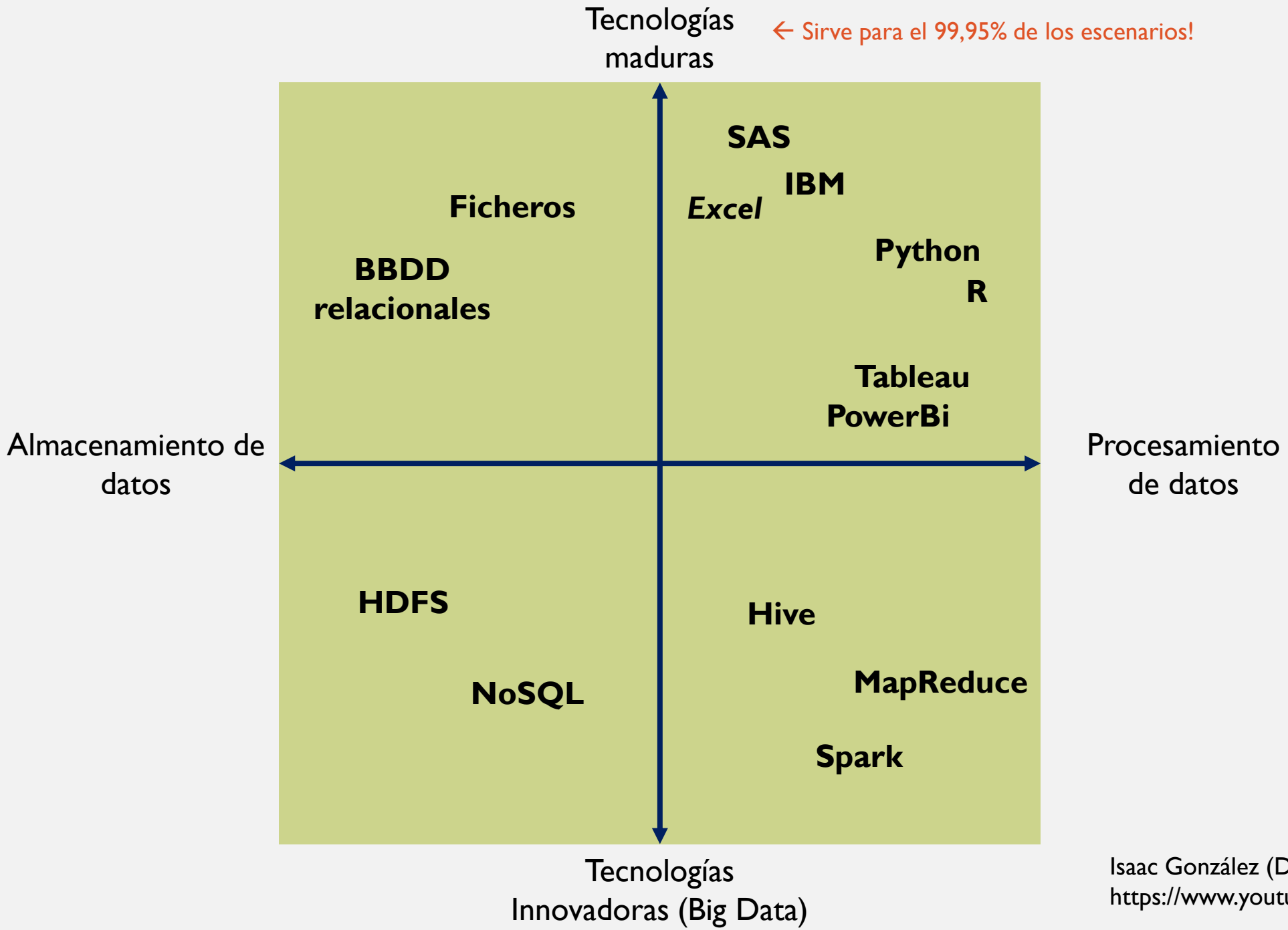
	STATA	R	SAS	PYTHON	SPSS
STATA					
R					
SAS					
PYTHON					
SPSS					

Data analyst jobs

3340 data analyst job ads were included in this analysis.

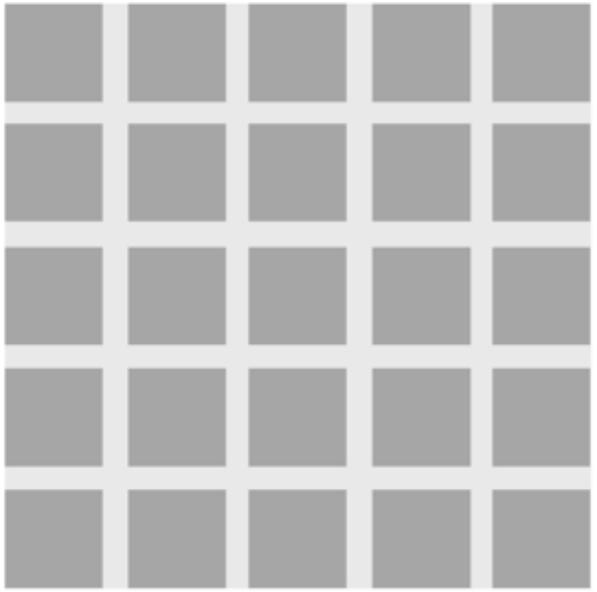
Tools / languages | Other skills | Degrees | Disciplines





Isaac González (DataScience ForBusiness)
https://www.youtube.com/watch?v=fyEBS2qpYEk

Structured data

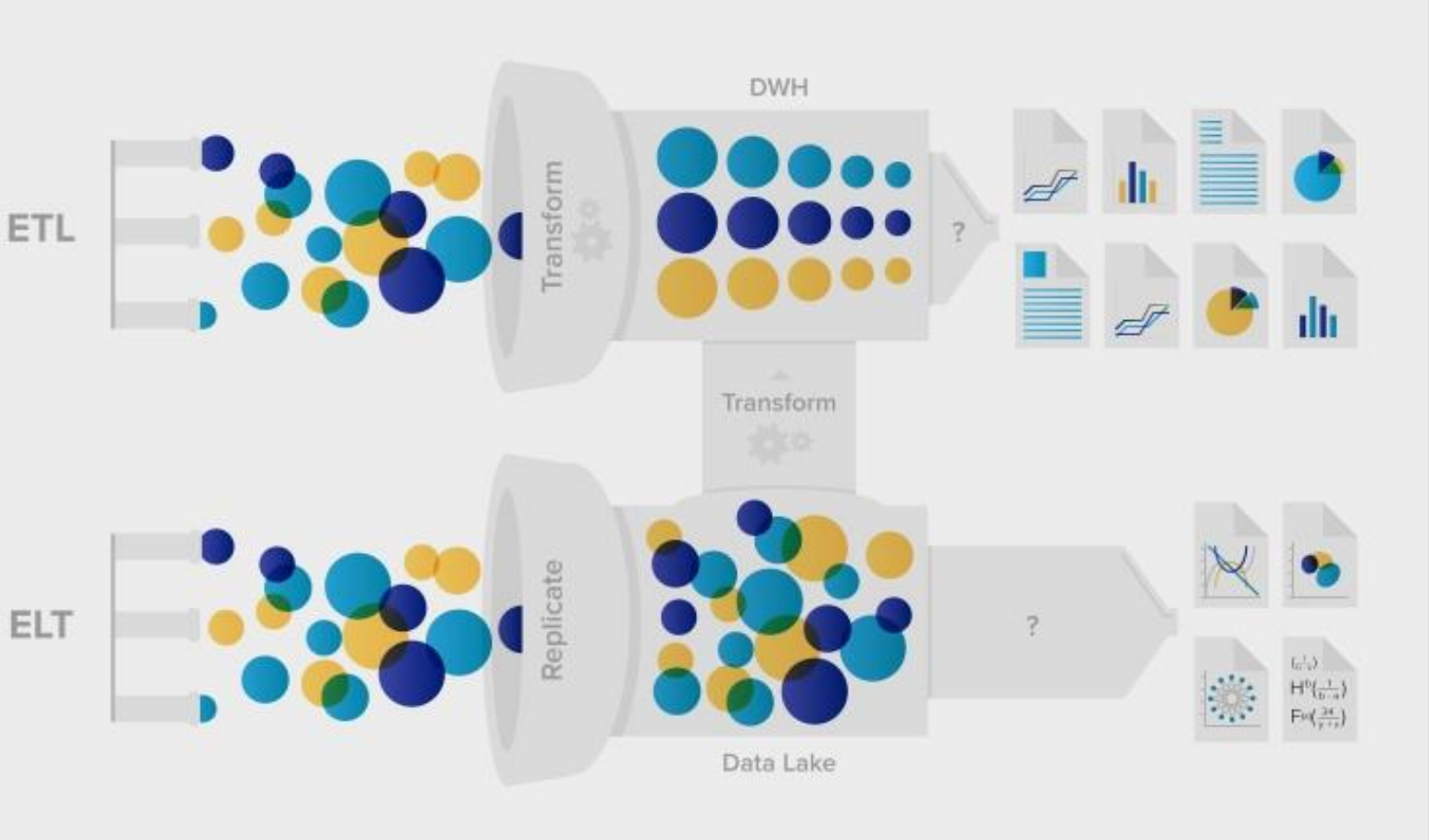


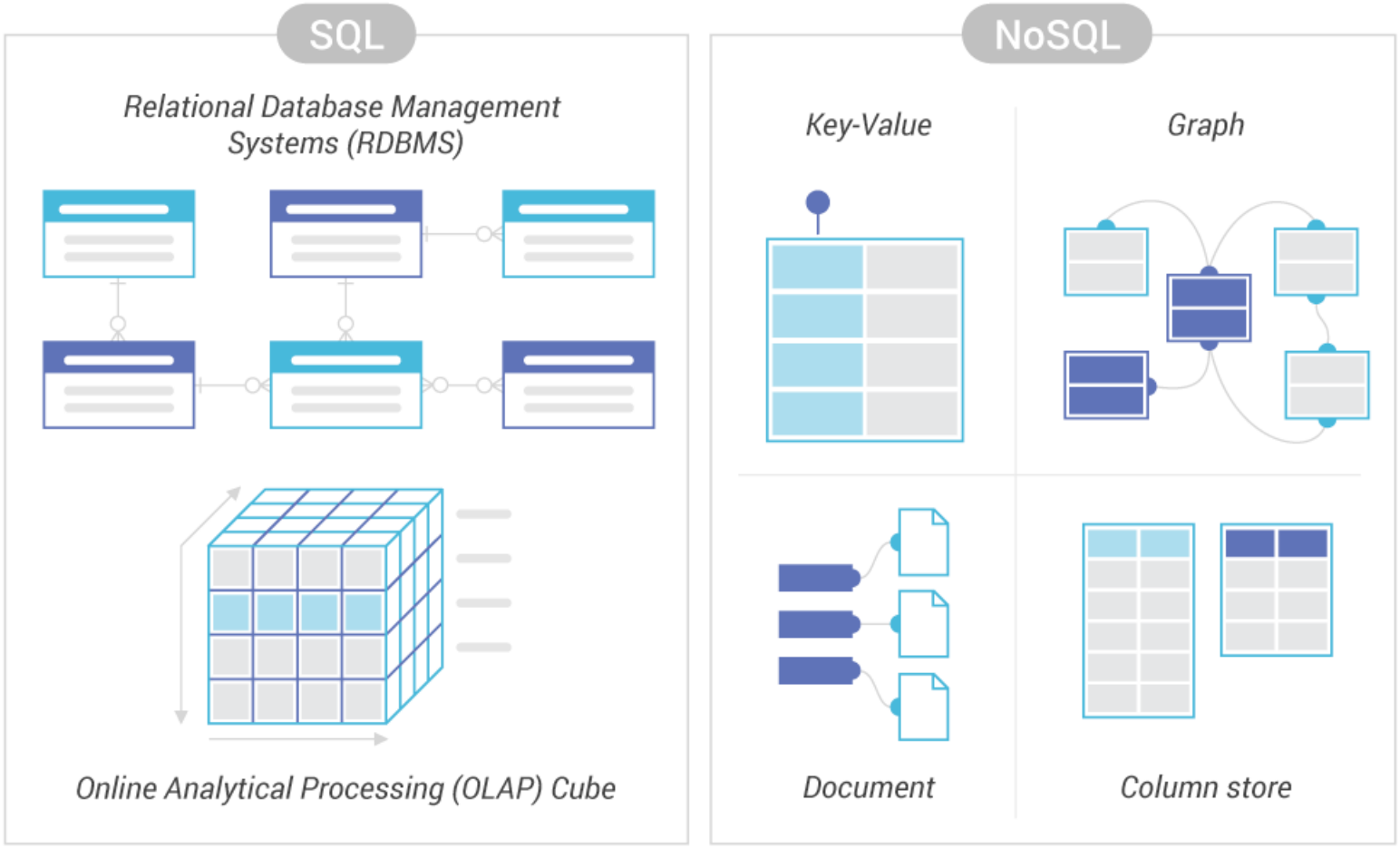
Database, CRM, ERP

Unstructured data

Text, audio, videos

ETL VS ELT



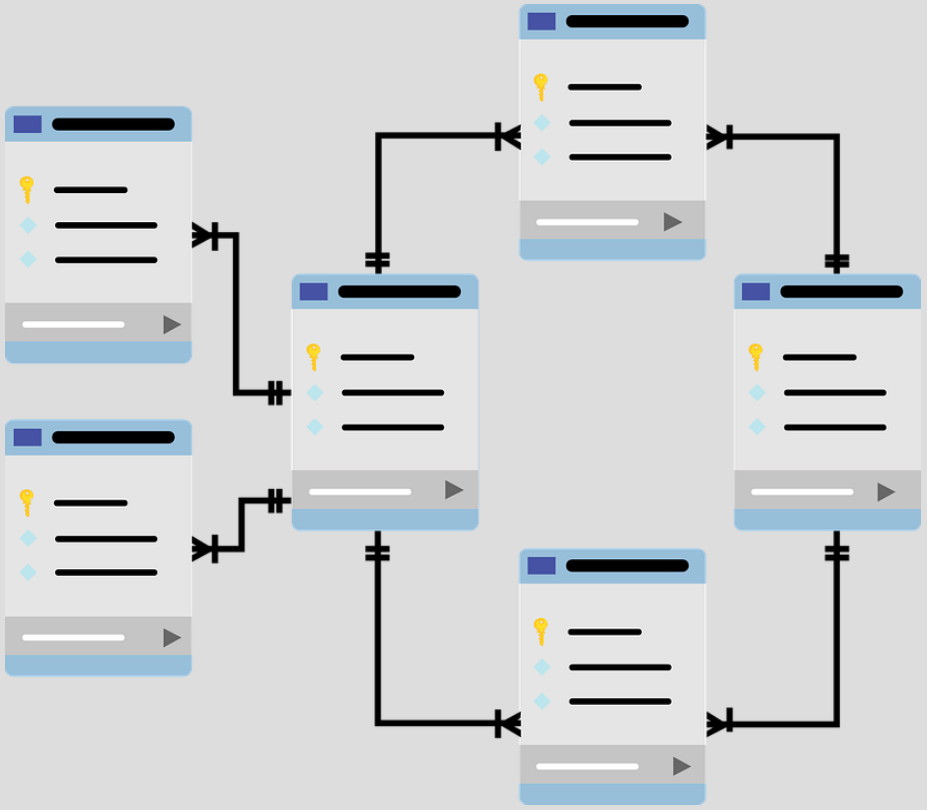

SQL
Relational Database Management Systems (RDBMS)
Online Analytical Processing (OLAP) Cube
NoSQL
Key-Value
Graph
Document
Column store

Quantitative

Qualitative

¿QUÉ ES UNA BASE DE DATOS?

BASES DE DATOS RELACIONALES



Microsoft SQL Server

ORACLE

MySQL

SQLite

PostgreSQL

CONSULTAS, FILAS Y COLUMNAS

```
SELECT EmployeeID,LastName,FirstName,BirthDate
FROM Employees
```

columna

EmployeeID	LastName	FirstName	BirthDate
1	Davolio	Nancy	1968-12-08
2	Fuller	Andrew	1952-02-19
3	Leverling	Janet	1963-08-30
4	Peacock	Margaret	1958-09-19
5	Buchanan	Steven	1955-03-04
6	Suyama	Michael	1963-07-02
7	King	Robert	1960-05-29
8	Callahan	Laura	1958-01-09
9	Dodsworth	Anne	1969-07-02
10	West	Adam	1928-09-19

fila

CLAVES

```
SELECT OrderID,OrderDate,EmployeeID
FROM Orders
ORDER BY OrderDate DESC
LIMIT 5
```

OrderID	OrderDate	EmployeeID
10443	1997-02-12	8
10442	1997-02-11	3
10440	1997-02-10	4
10441	1997-02-10	3
10439	1997-02-07	6

FirstName
Laura
Janet
Margaret
Janet
Michael

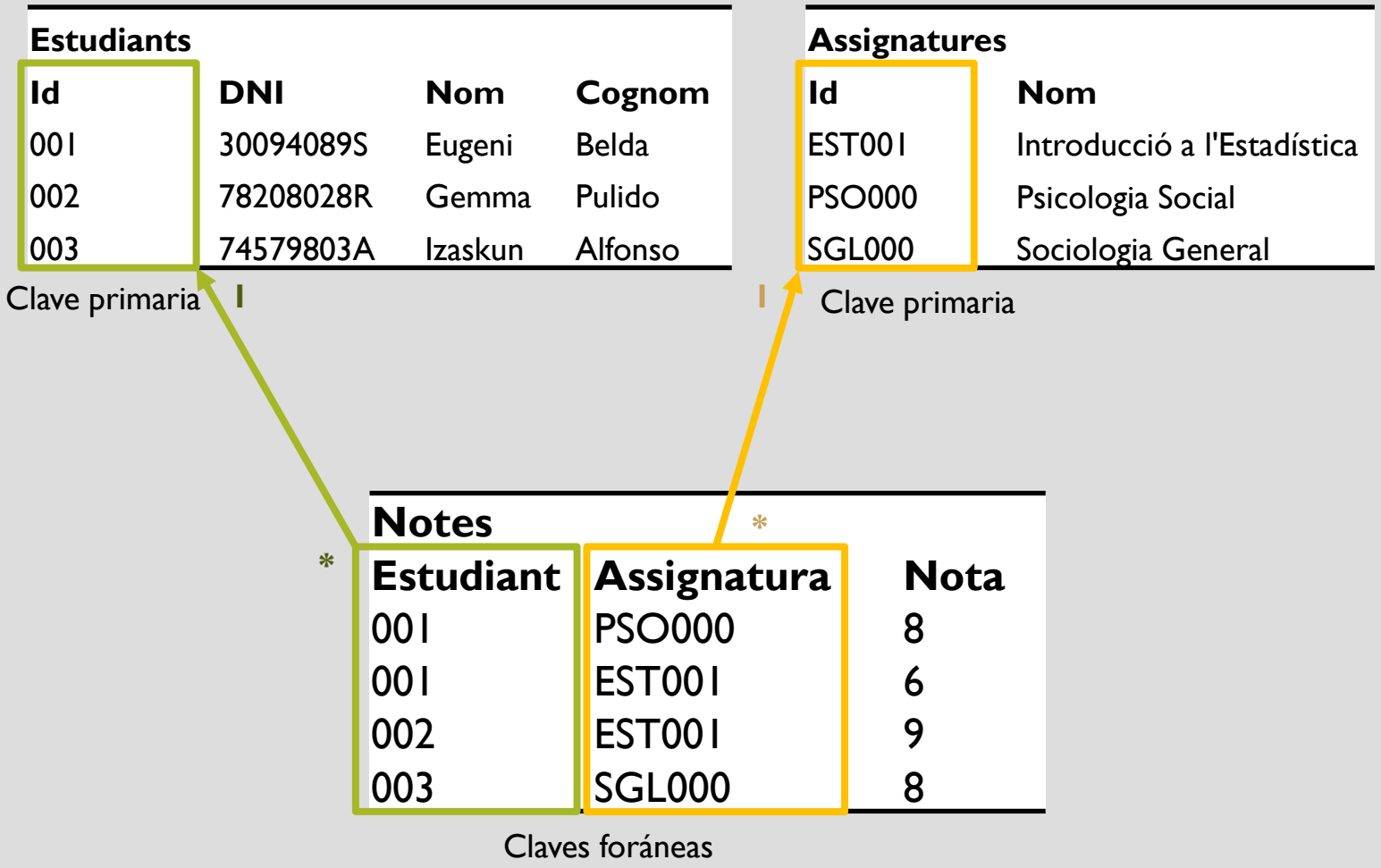
CLAVES NATURALES Y ARTIFICIALES

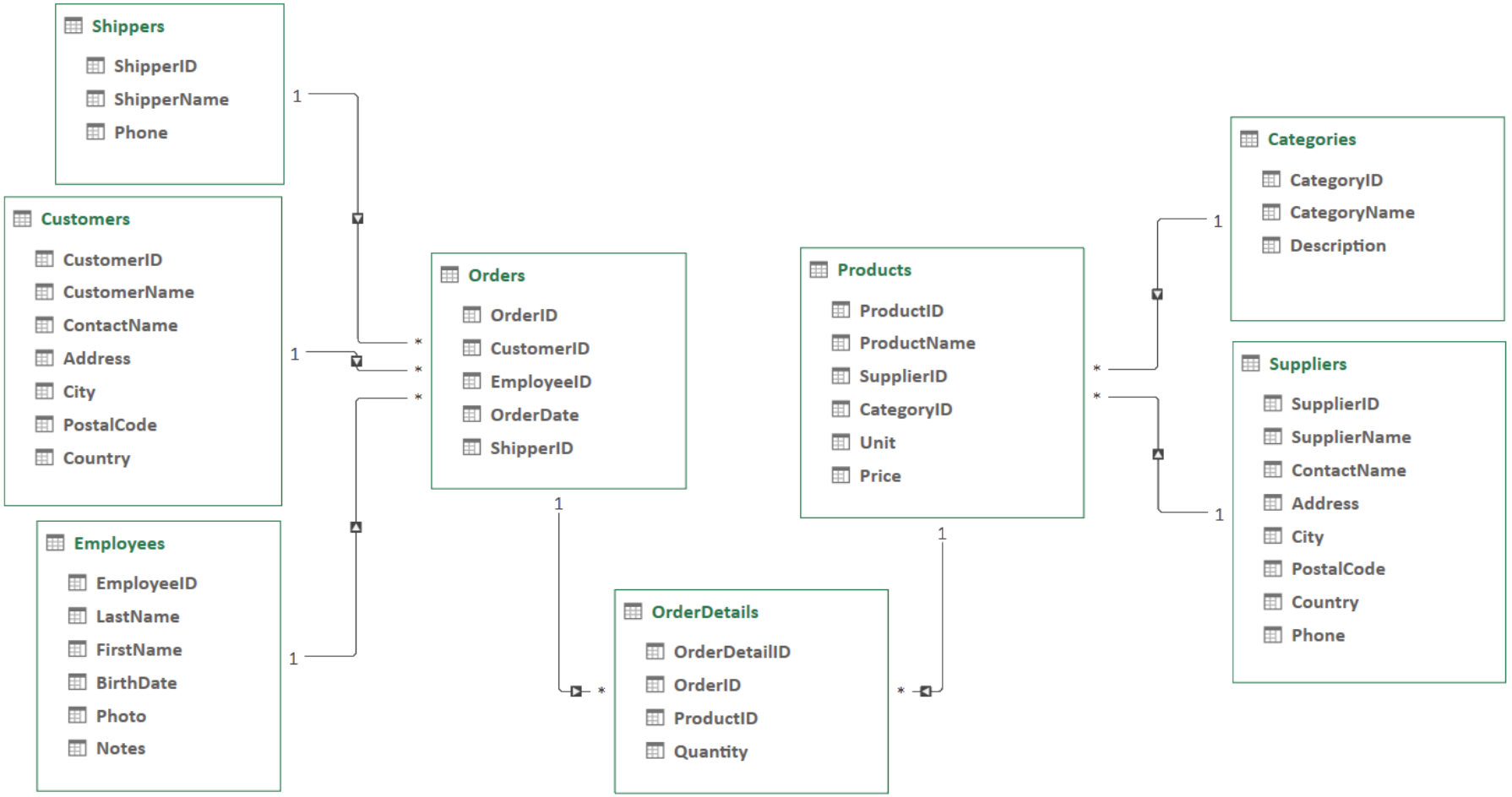
Estudiants

Id	DNI	Nom	Cognom
001	30094089S	Eugeni	Belda
002	78208028R	Gemma	Pulido
003	74579803A	Izaskun	Alfonso

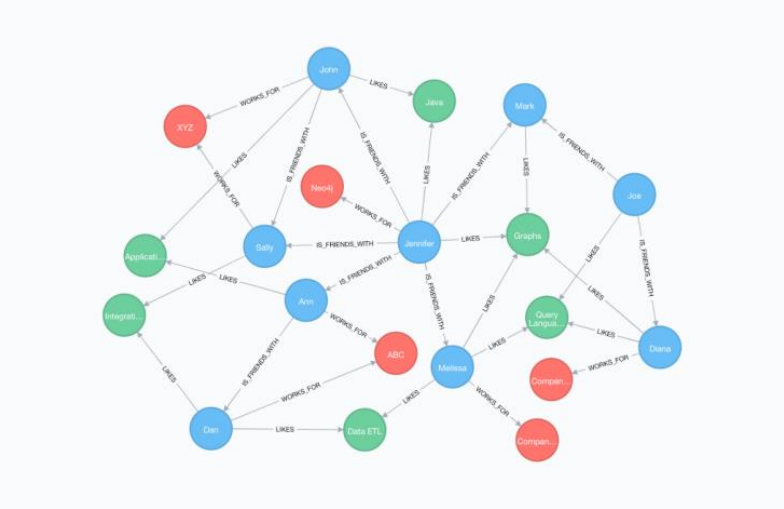
Clave artificial Clave natural

CLAVES PRIMARIAS Y FORÁNEAS




Shippers
ShipperID
ShipperName
Phone
Customers
CustomerID
CustomerName
ContactName
Address
City
PostalCode
Country
Employees
EmployeeID
LastName
FirstName
BirthDate
Photo
Notes
Orders
OrderID
CustomerID
EmployeeID
OrderDate
ShipperID
OrderDetails
OrderDetailID
OrderID
ProductID
Quantity
Products
ProductID
ProductName
SupplierID
CategoryID
Unit
Price
Categories
CategoryID
CategoryName
Description
Suppliers
SupplierID
SupplierName
ContactName
Address
City
PostalCode
Country
Phone
1
*

BASES DE DATOS NO RELACIONALES (NOSQL)



Document 1

```
{
  "id": "1",
  "name": "John Smith",
  "isActive": true,
  "dob": "1964-30-08"
}
```

Document 2

```
{
  "id": "2",
  "fullName": "Sarah Jones",
  "isActive": false,
  "dob": "2002-02-18"
}
```

Document 3

```
{
  "id": "3",
  "fullName":
  {
    "first": "Adam",
    "last": "Stark"
  },
  "isActive": true,
  "dob": "2015-04-19"
}
```

APACHE HBASE

Cassandra

riak

CouchDB relax

mongoDB

HYPERTABLE INC

Neo4j

redis

amazon DynamoDB

NOSQL VS. SQL

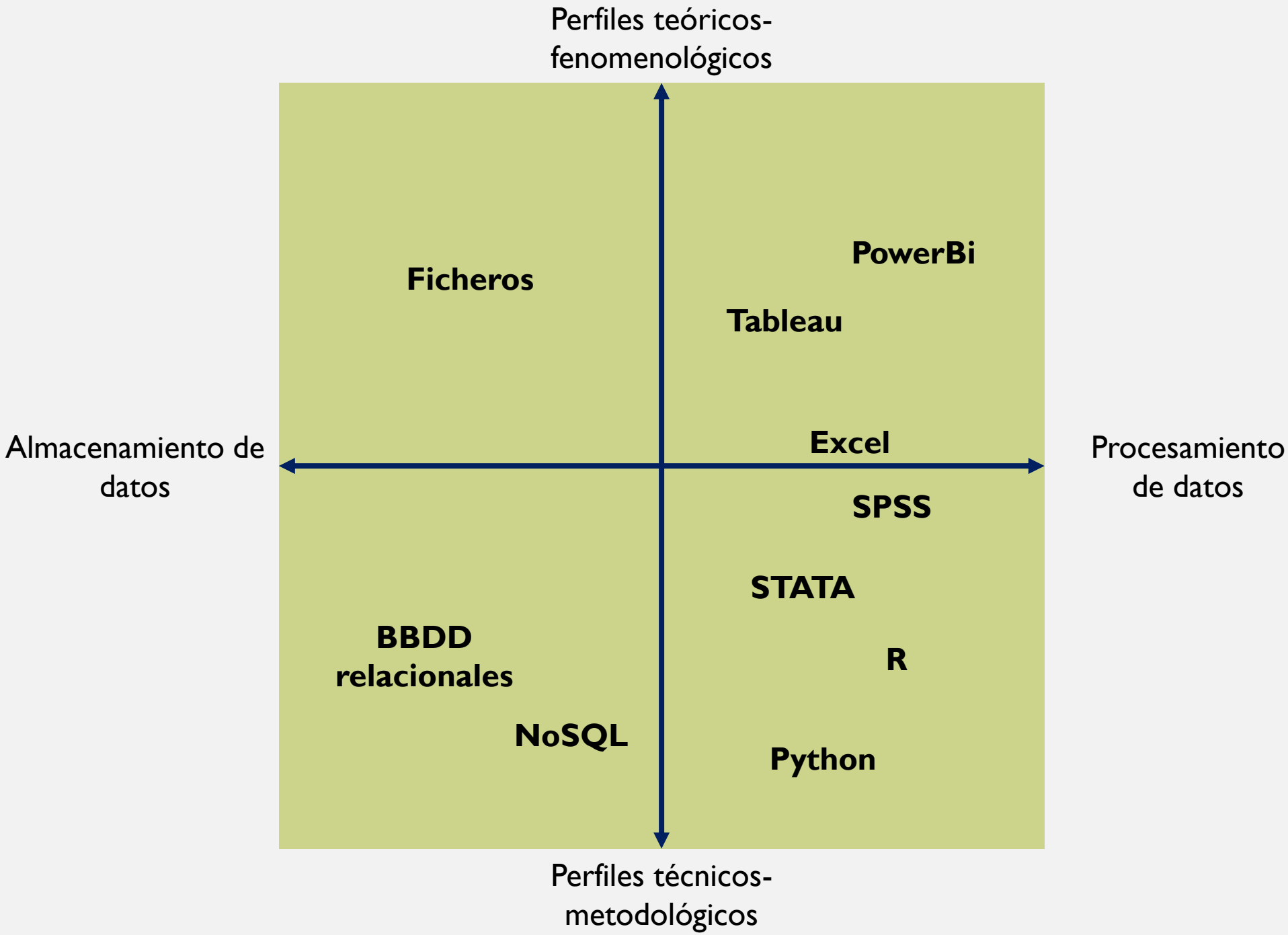
	NoSQL	SQL
Model	Non-relational	Relational
	Stores data in JSON documents, key/value pairs, wide column stores, or graphs	Stores data in a table
Data	Offers flexibility as not every record needs to store the same properties	Great for solutions where every record has the same properties
	New properties can be added on the fly	Adding a new property may require altering schemas or backfilling data
	Relationships are often captured by denormalizing data and presenting all data for an object in a single record	Relationships are often captured in normalized model using joins to resolve references across tables
	Good for semi-structured, complex, or nested data	Good for structured data
Schema	Dynamic or flexible schemas	Strict schema
	Database is schema-agnostic and the schema is dictated by the application. This allows for agility and highly iterative development	Schema must be maintained and kept in sync between application and database
Performance	Performance can be maximized by reducing consistency, if needed	Insert and update performance is dependent upon how fast a write is committed, as strong consistency is enforced. Performance can be maximized by using scaling up available resources and using in-memory structures.
	All information about an entity is typically in a single record, so an update can happen in one operation	Information about an entity may be spread across many tables or rows, requiring many joins to complete an update or a query
Scale	Scaling is typically achieved horizontally with data partitioned to span servers	Scaling is typically achieved vertically with more server resources

https://rmohan.com/?p=7562

NOSQL VS. SQL II

When to use different data management and analysis technologies
Text files, spreadsheets, and scripting language
• Your data are small
• Your analysis is simple
• You do not expect to repeat analyses over time
Statistical packages
• Your data are modest in size
• Your analysis maps well to your chosen statistical package
Relational database
• Your data are structured
• Your data are large
• You will be analyzing changed versions of your data over time
• You want to share your data and analyses with others
NoSQL database
• Your data are unstructured
• Your data are extremely large
• Your analysis will happen mostly outside the database in a programming language

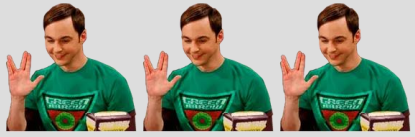
Foster, Ghani, Jarmin, Kreuter i Lane. 2020. Big Data and Social Science.


Perfiles teóricos-
fenomenológicos
Ficheros
PowerBi
Tableau
Excel
SPSS
STATA
R
Python
BBDD
relacionales
NoSQL
Almacenamiento de
datos
Procesamiento
de datos
Perfiles técnicos-
metodológicos


RStudio


ANACONDA


jamovi
Stats.
Open.
Now.

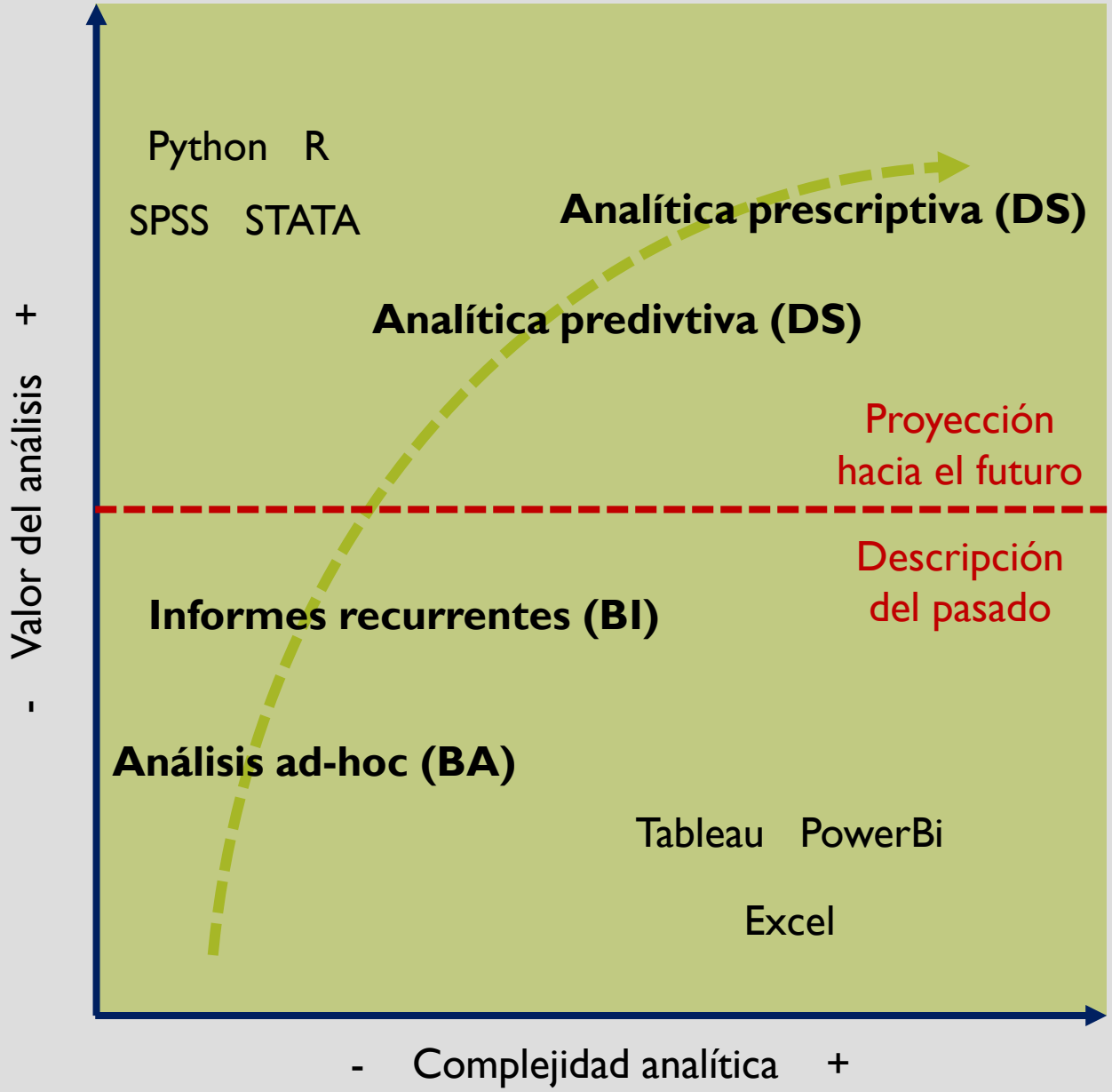



Grafana


Google
Data Studio


Power BI

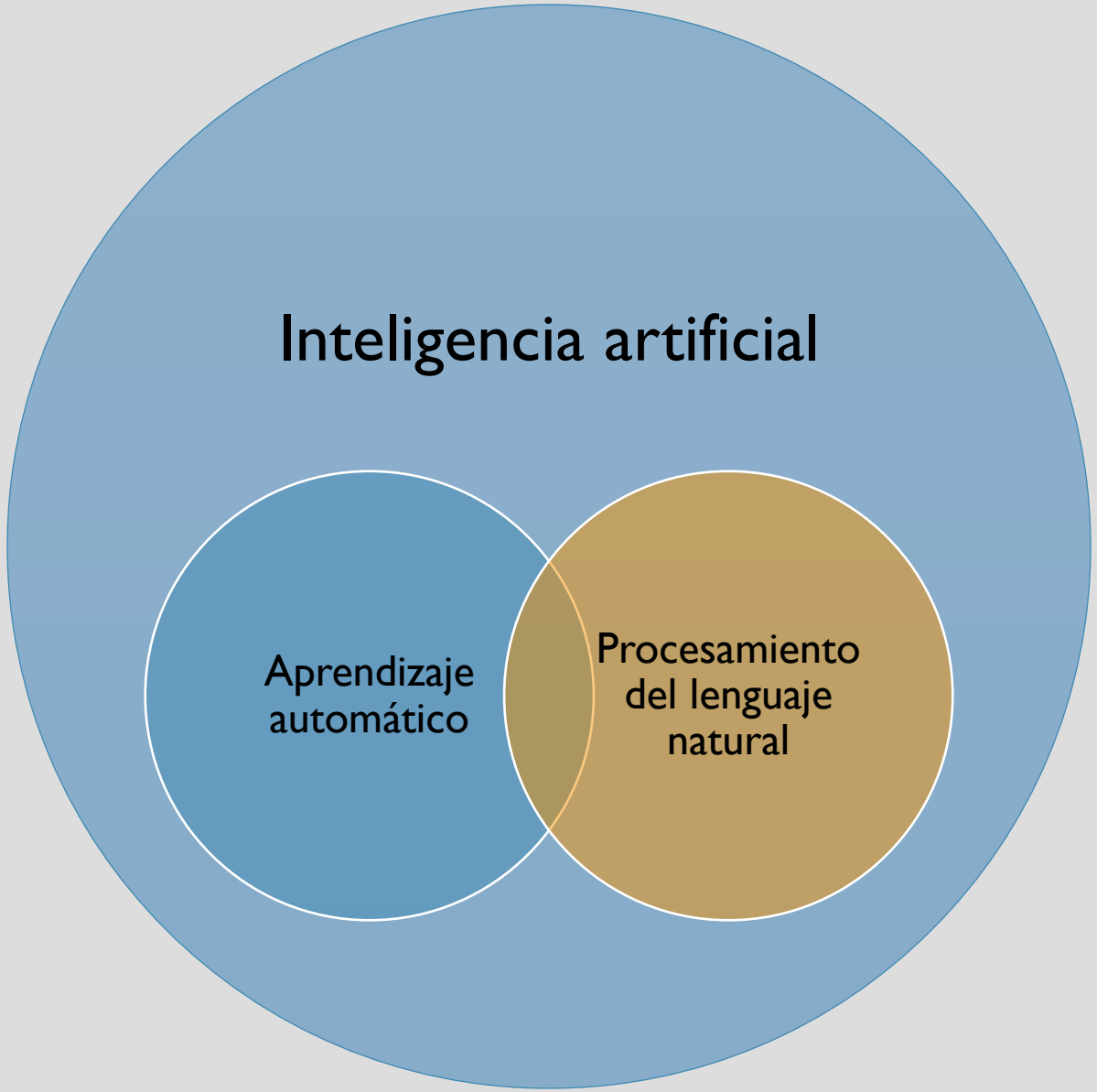

+ableau



Isaac González (DataScience ForBusiness)
https://www.youtube.com/watch?v=IQEMni-OOaA

Tom Davenport (2014) *Big Data at Work: Dispelling the Myths, Uncovering the Opportunities*. Harvard Business Review Press.

PROCESAMIENTO DEL LENGUAJE NATURAL


Inteligencia artificial
Aprendizaje automático
Procesamiento del lenguaje natural

ALGORITMOS DE REGLAS HEURÍSTICAS

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?

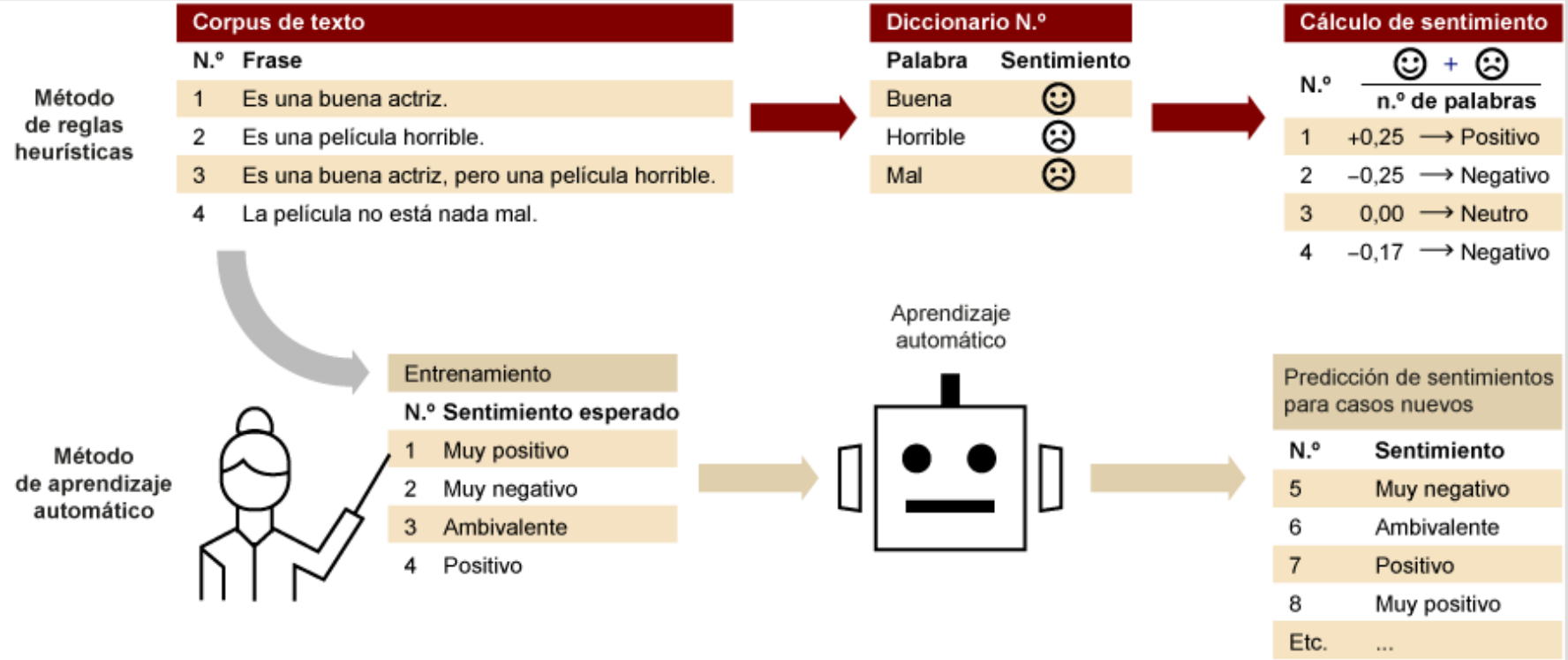
Let’s tokenize! Isn’t this easy?

Word	***Lemmatization***	***Stemming***
was	be	wa
studies	study	studi
studying	study	study

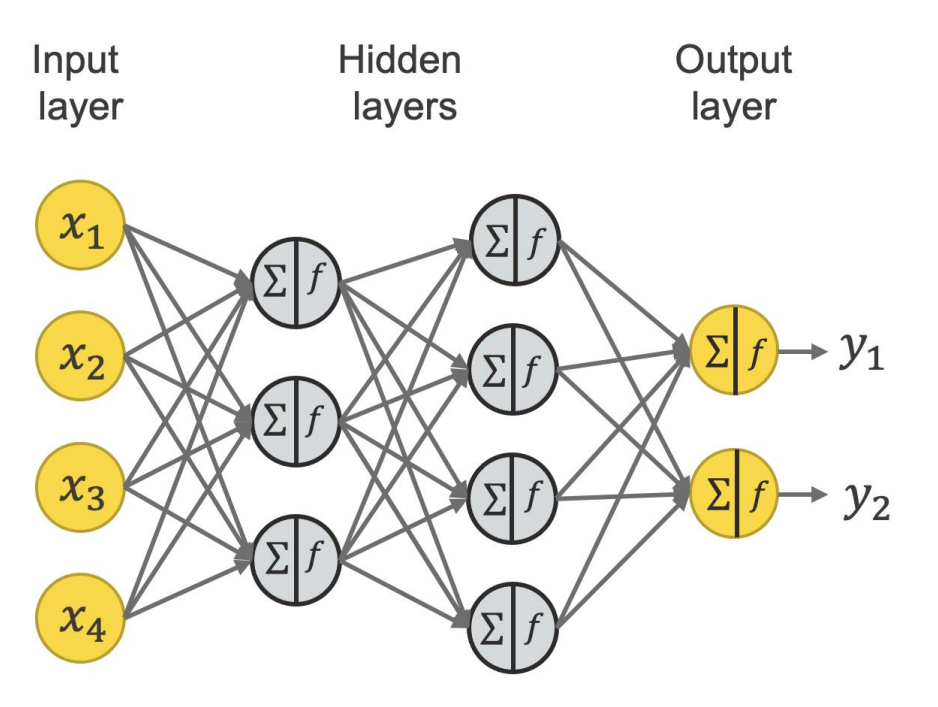
The Bag of Words Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…



DE LAS REGLAS HEURÍSTICAS AL APRENDIZAJE AUTOMÁTICO



itzuli
itzultzaile neuronala

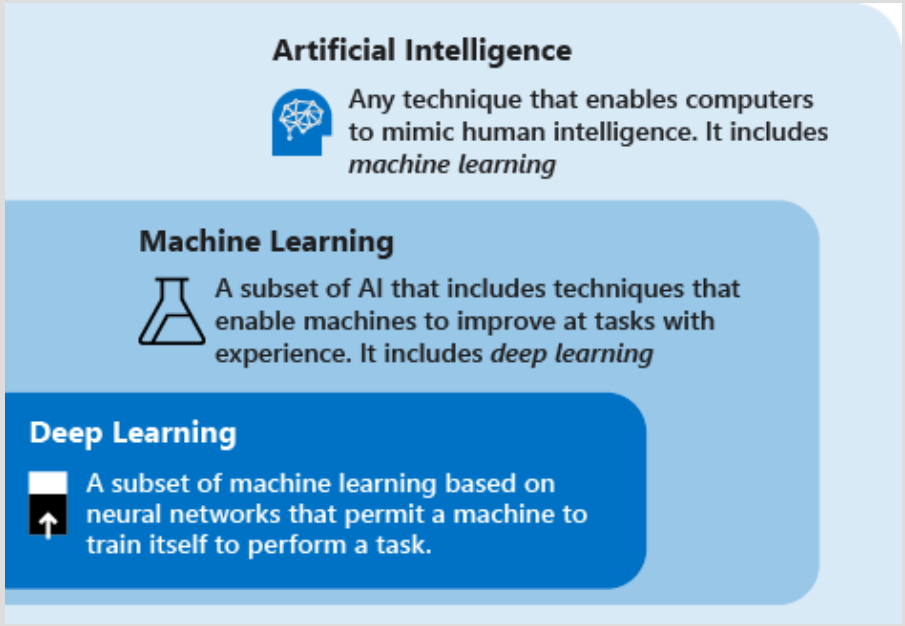
BETA batua

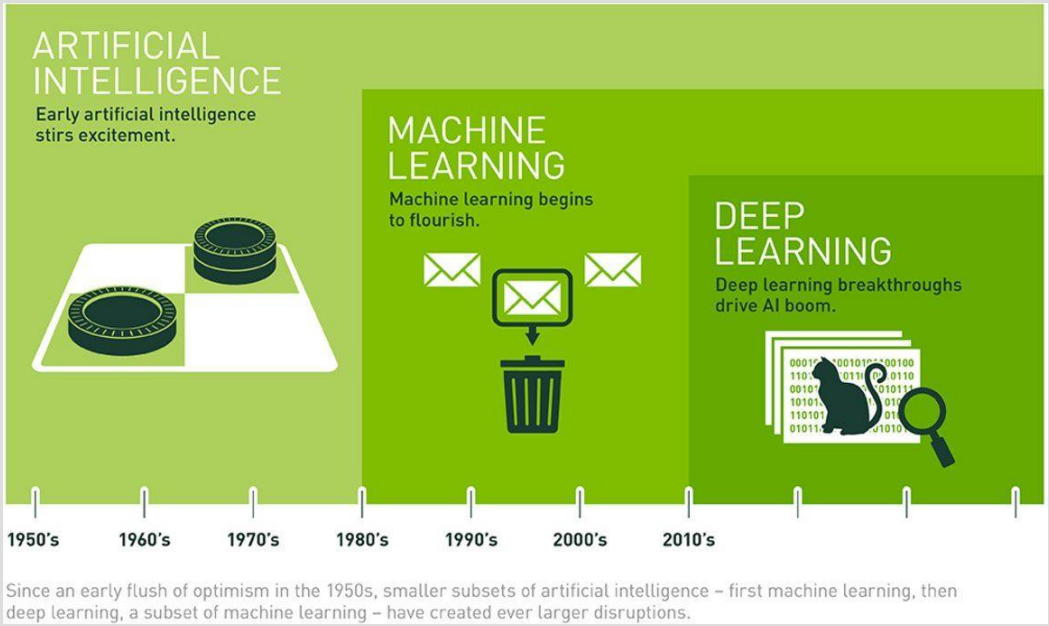
elia elhuyar

PLN + REDES NEURONALES

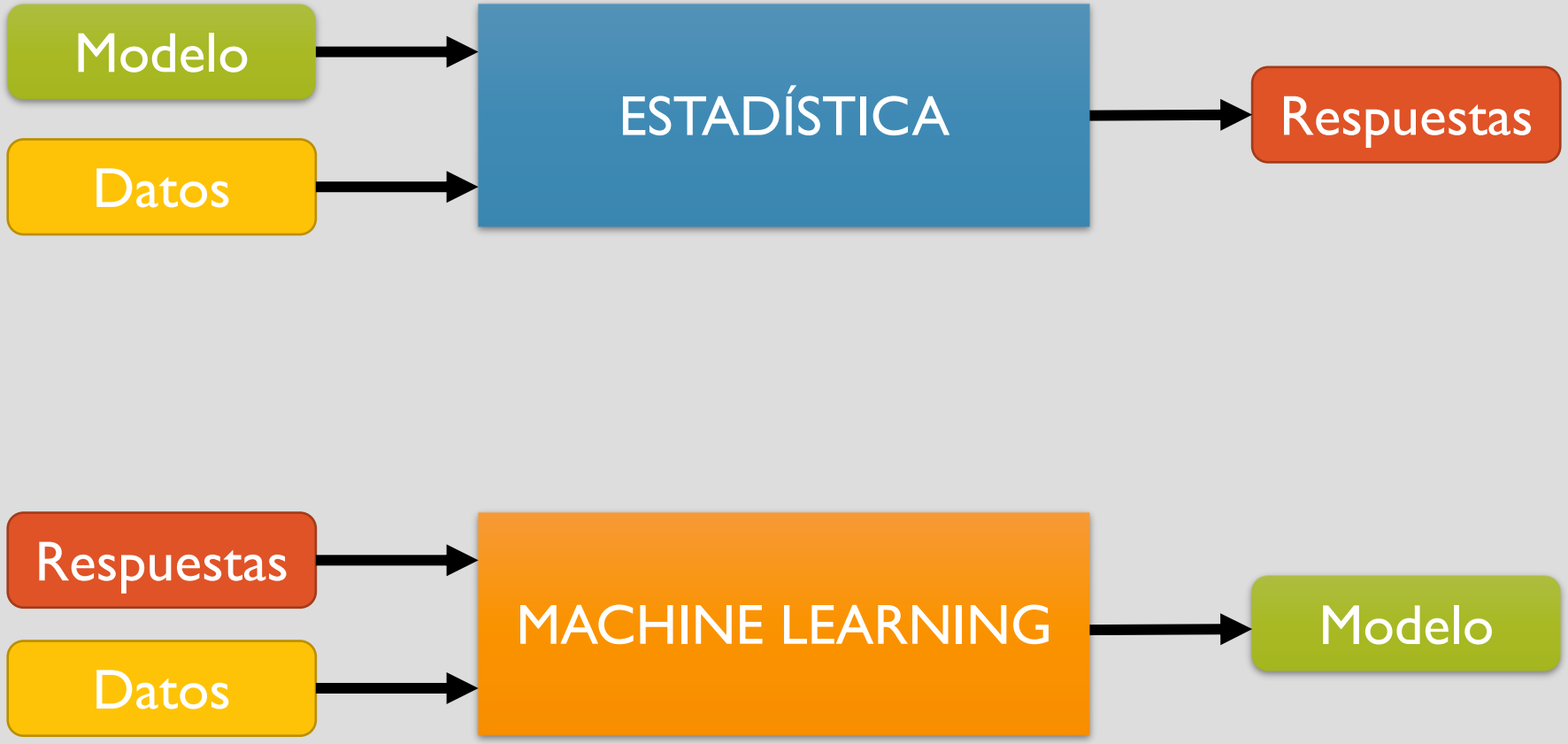
MACHINE LEARNING

MACHINE LEARNING, INTEL·LIGÈNCIA ARTIFICIAL Y DEEP LEARNING






STATISTICS
VS
MACHINE
LEARNING


Modelo
Datos
ESTADÍSTICA
Respuestas
Respuestas
Datos
MACHINE LEARNING
Modelo

ESTADÍSTICA (EJEMPLO)

Sexo o fenotipo (raza) ¿Qué afecta más al sueldo?

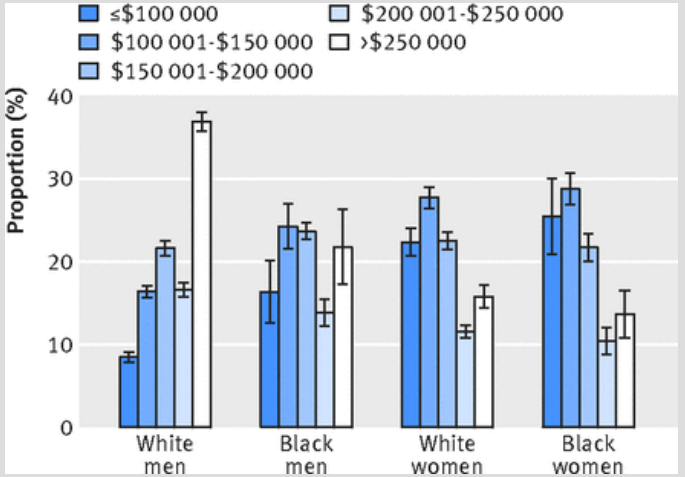
Differences in incomes of physicians in the United States by race and sex: observational study (2016)

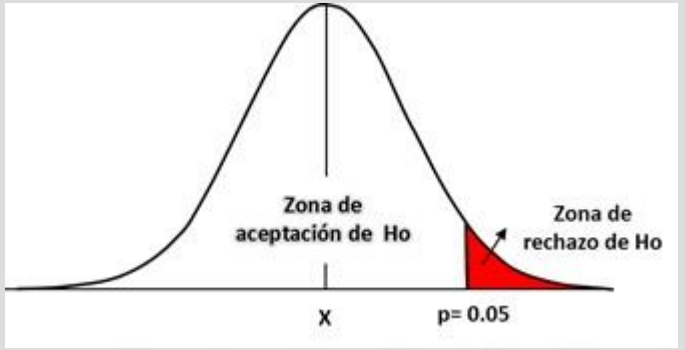
- 36.9% of **white male** physicians had adjusted annual income over $250 000 compared with 21.8% of **black male** physicians (P<0.001 for difference)
- 15.8% of **white female** physicians earned more than $250 000 compared with 13.6% of **black female** physicians (P=0.09 for difference)
- **White male** physicians earn substantially more than **black male** physicians, while **white and black female physicians** earn similar incomes to each other, but significantly less than their **male counterparts**

Modelo explicativo/inferencial
Sexo y Raza (Variables Independientes) → Sueldo (Variable dependiente)

Datos
2000-13 American Community Survey (61.327)
2000-08 Center for Studying Health System Change Survey (17.583)

Obtengo **Respuestas**





MACHINE LEARNING (EJEMPLO)

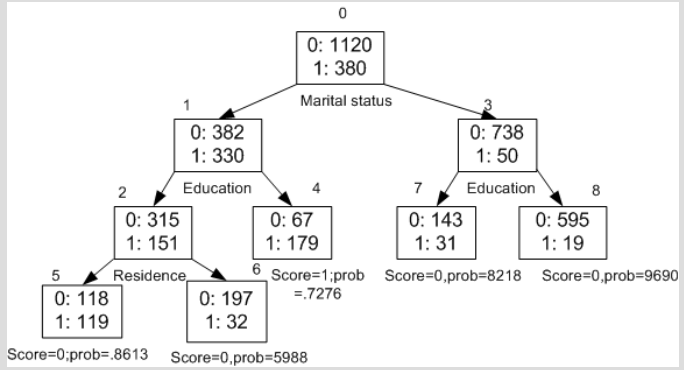
Quiero predecir el sueldo de un individuo (e.g. de un físico) en base a sus datos

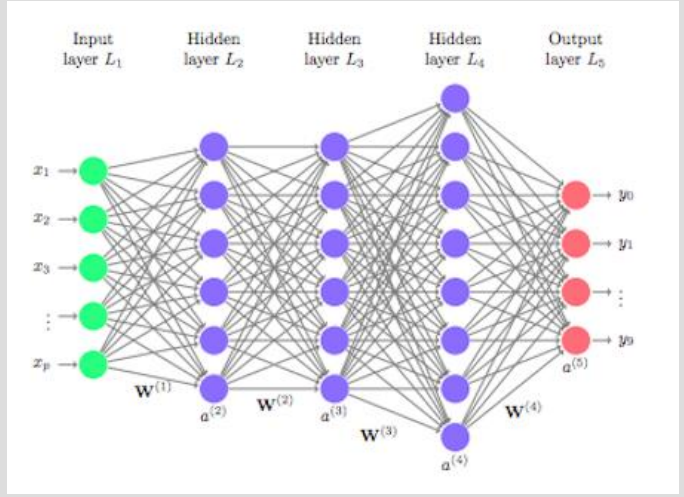
- Sexo, Raza y Sueldo… Edad, Universidad(es) en las que ha estudiado, Becas disfrutadas, Años en activo, Sector de empleo, Ciudad de residencia, Profesión de los progenitores, Idiomas, Estado civil, Nº de hijos…
- Amistades en Facebook y Twitter, Seguros contratados, Datos de consumo (tarjeta de crédito), Aficiones, Horas de navegación en Internet, Capital inmobiliario…
- Opción sexual, Creencias religiosas, Probabilidad de padecer cáncer…

Datos
Introduzco tantas variables (*features*) como quiera (o pueda), de tantas fuentes de datos como sea capaz de reunir

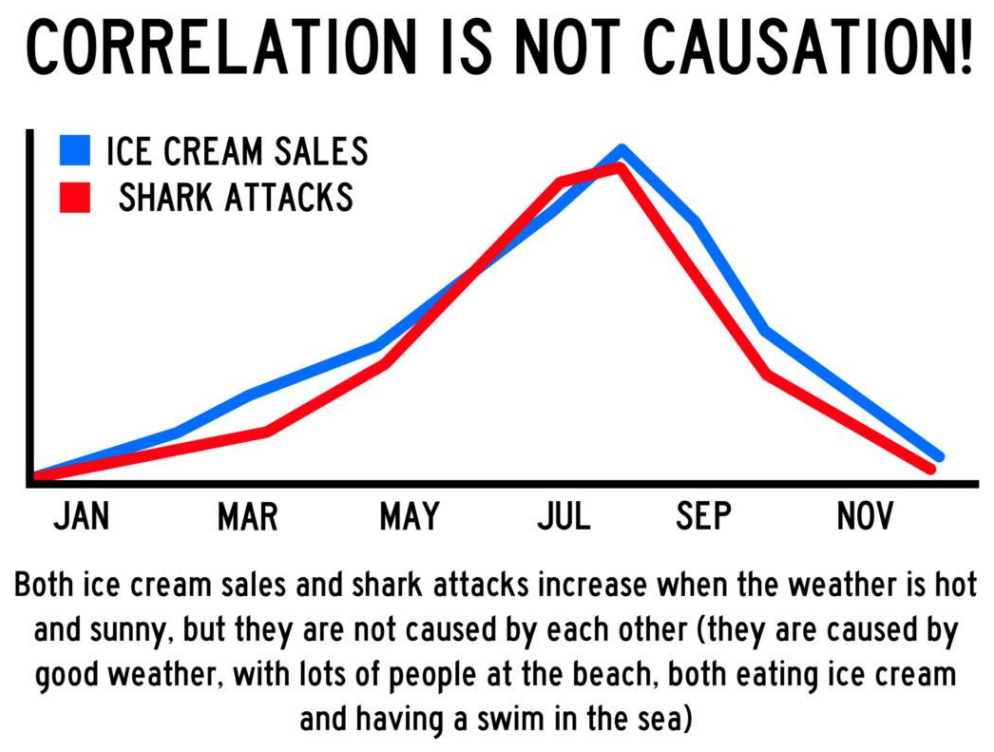
Respuestas
Entreno mi modelo para predecir la variable SUELDO (*target variable*) en base a casos conocidos y hago un test de validez

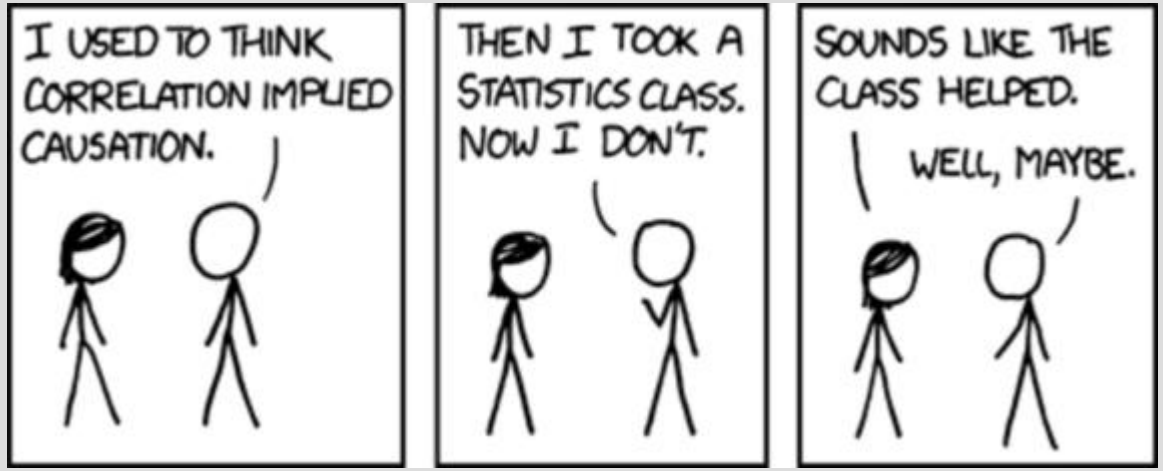
Obtengo un **modelo probabilístico-predictivo**

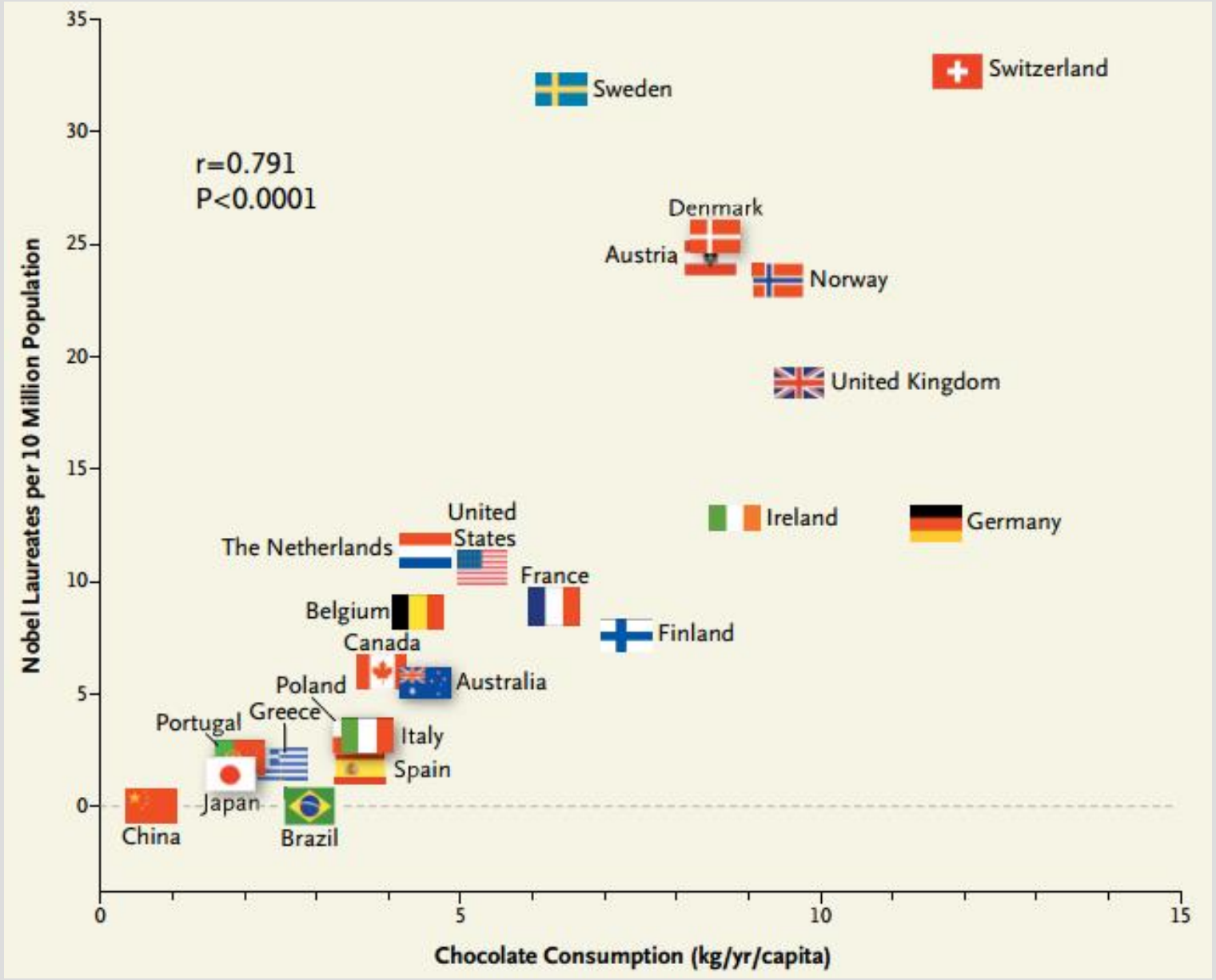


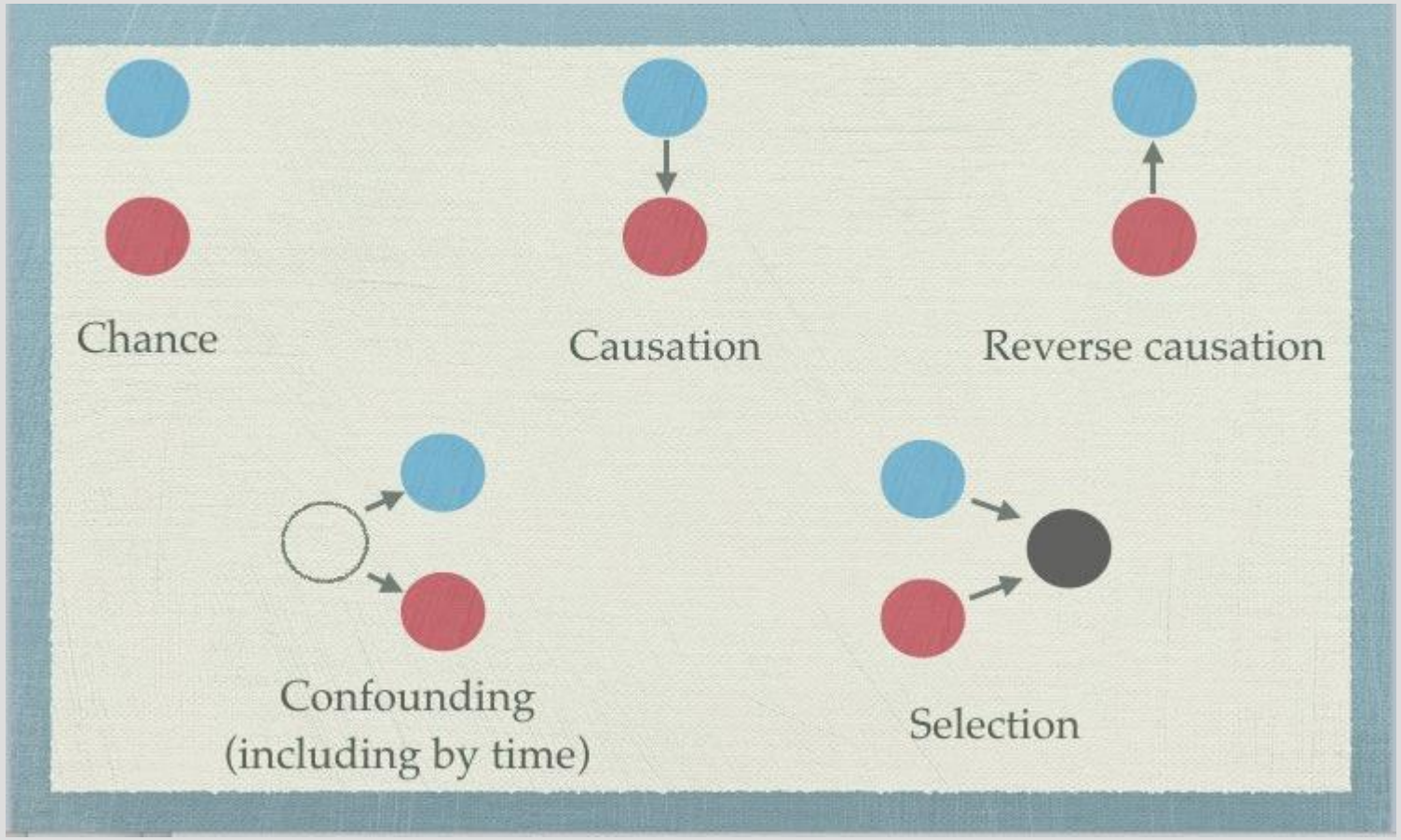


CORRELACIÓN Y CAUSALIDAD






r=0.791
P<0.0001
Nobel Laureates per 10 Million Population
Chocolate Consumption (kg/yr/capita)
0
5
10
15
20
25
30
35
Sweden
Switzerland
Denmark
Austria
Norway
United Kingdom
Ireland
Germany
United States
The Netherlands
France
Belgium
Finland
Canada
Australia
Poland
Greece
Portugal
Italy
Spain
Japan
China
Brazil



LA CAUSALIDAD ES **UN TIPO DE** CORRELACIÓN

ESTADÍSTICA VS. MACHINE LEARNING

ESTADÍSTICA

+ Explicación para datos disponibles (inferencia)
+ Orientada a la explicación de las relaciones entre variables (formalización, interpretación)
+ Principio de Parsimonia

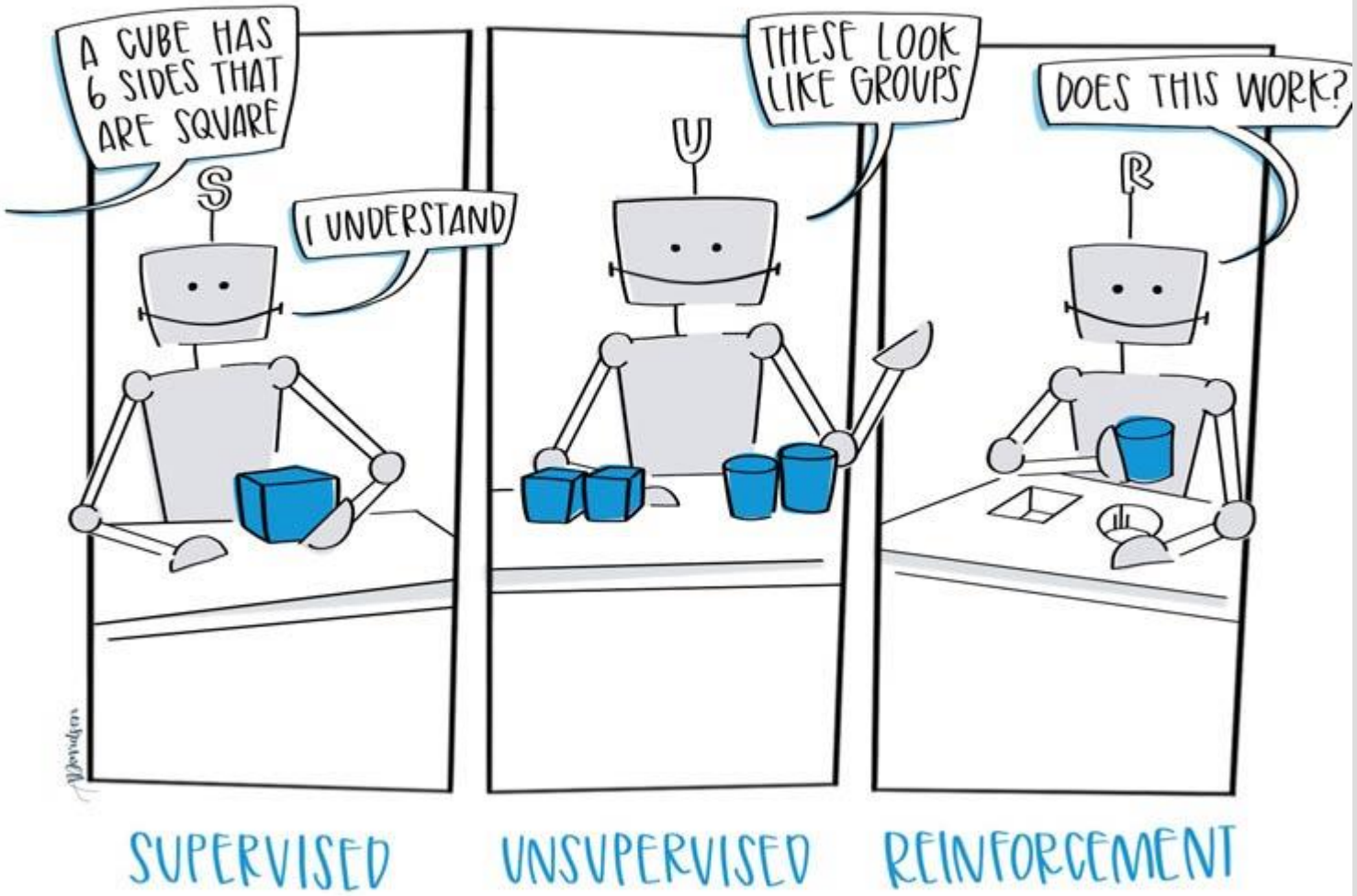
- Despreocupación (o incredulidad) respecto a la capacidad predictiva de los modelos de datos

MACHINE LEARNING

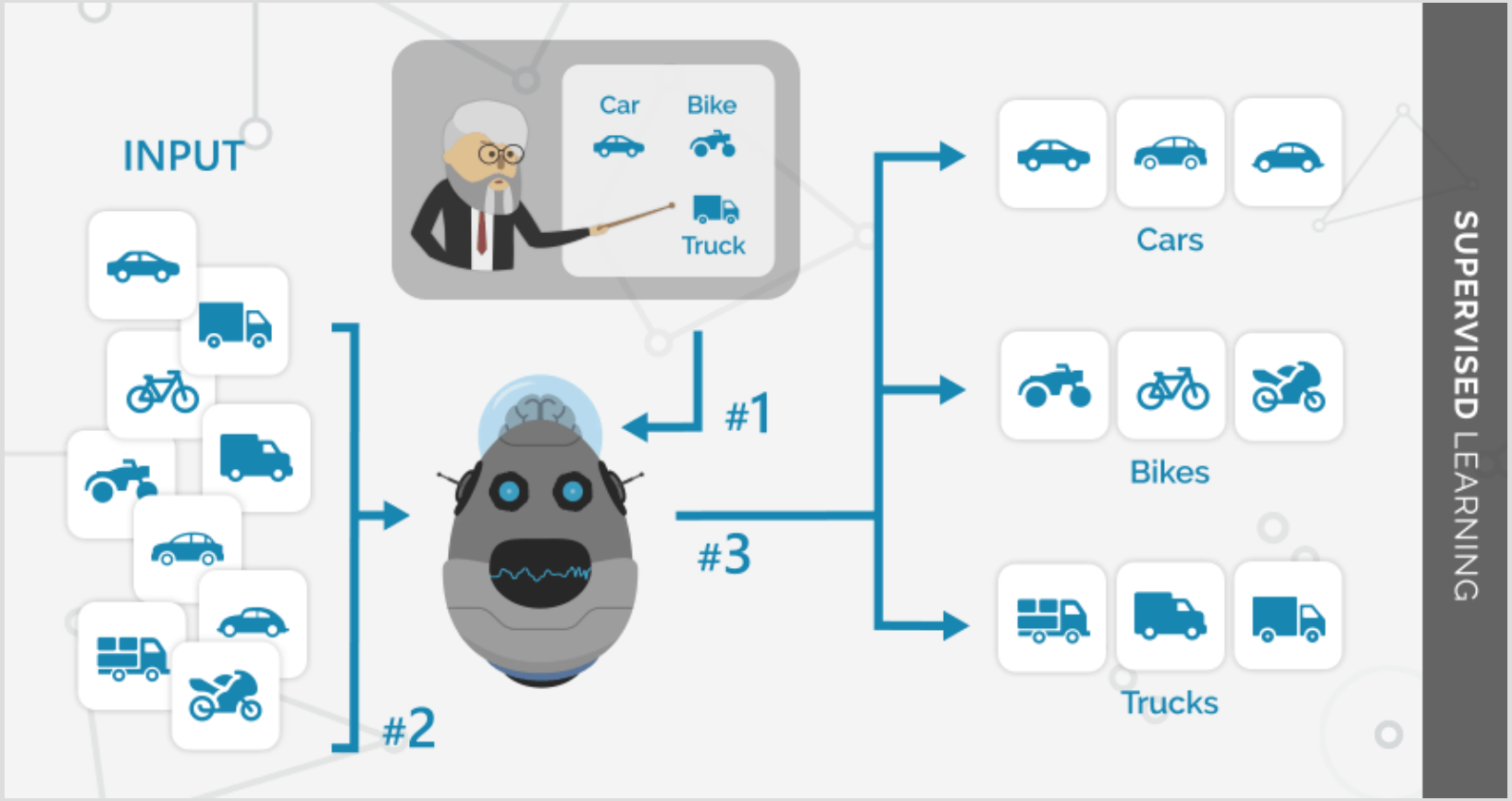
+ Entrenamiento y testeo para datos futuros (predicción)
+ Orientado a la obtención de resultados replicables
+ Capacidad de computación

- Desinterés por la interpretación de los modelos de datos

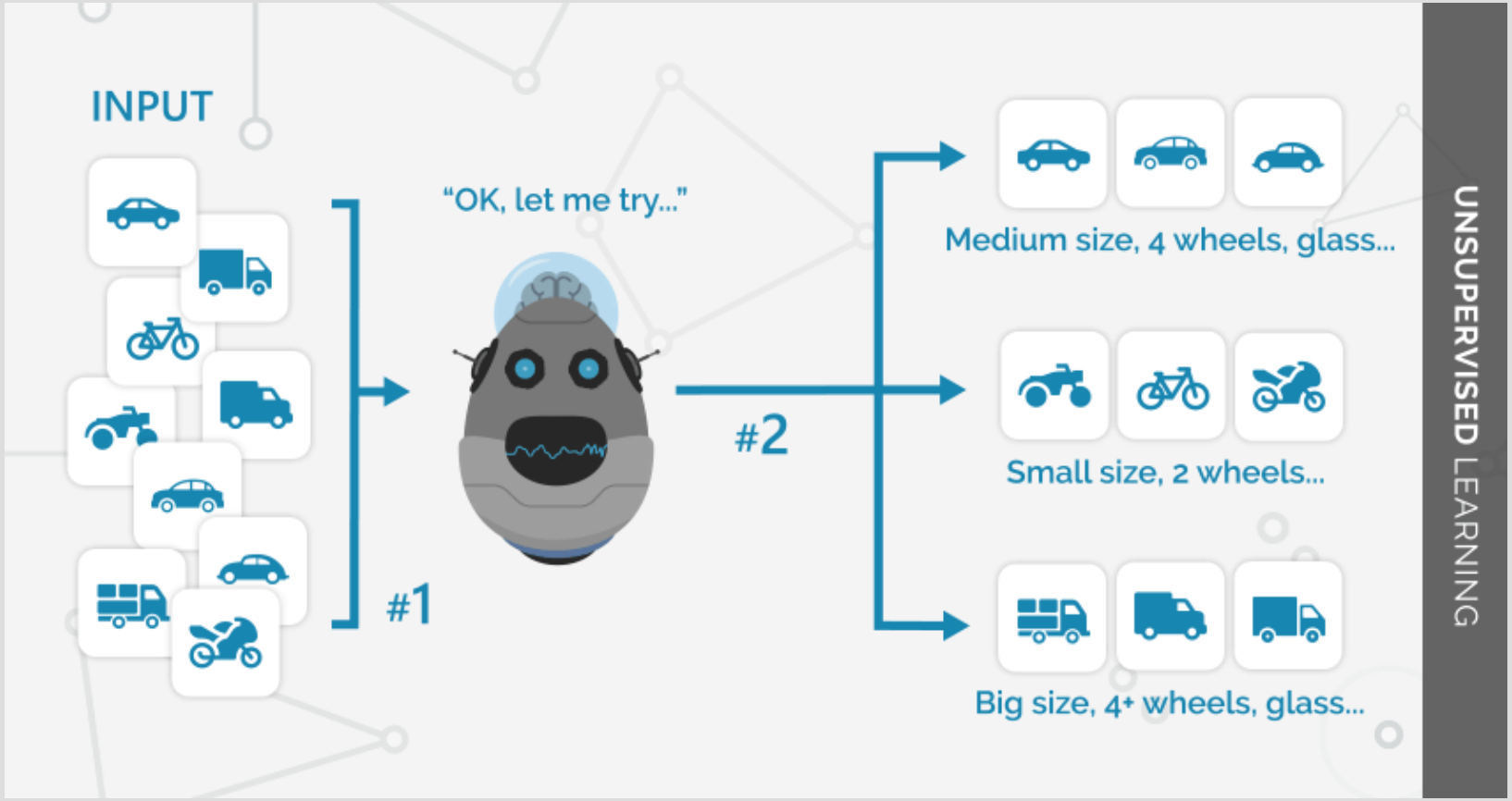
MACHINE LEARNING



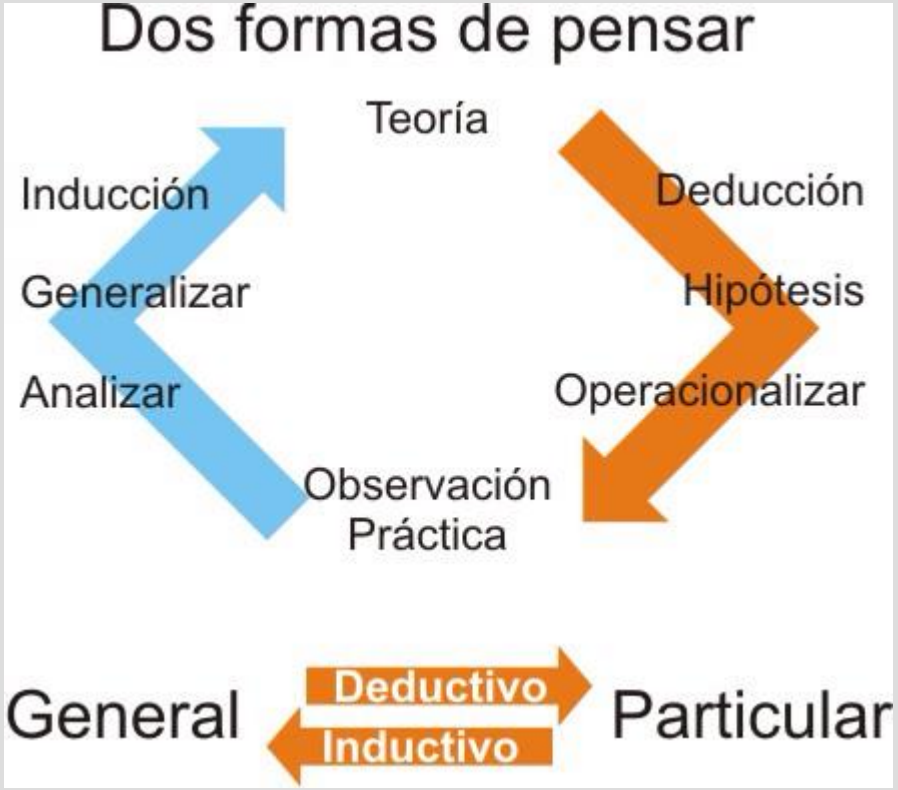
APRENDIZAJE SUPERVISADO



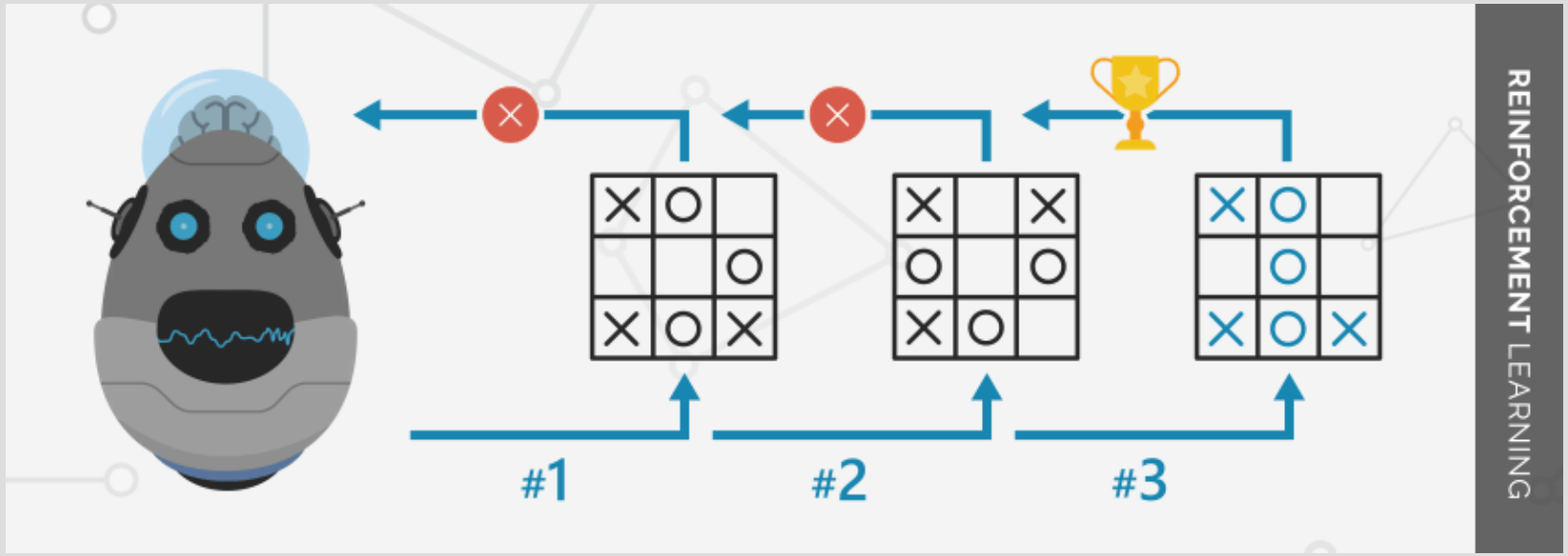
APRENDIZAJE NO SUPERVISADO

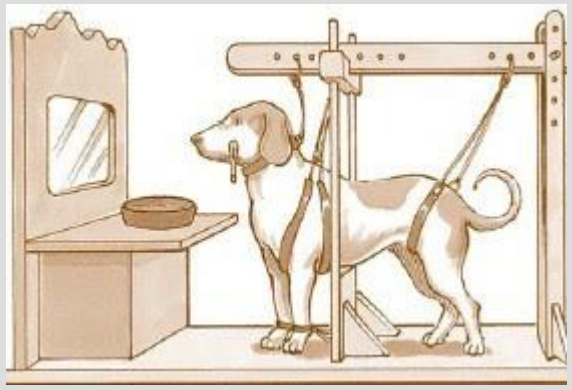


MÁQUINAS HECHAS A NUESTRA IMAGEN Y SEMEJANZA…



APRENDIZAJE POR REFUERZO





CLASSICAL MACHINE LEARNING

Data is pre-categorized or numerical

SUPERVISED

Predict a category

CLASSIFICATION

«Divide the socks by color»

Predict a number

REGRESSION

«Divide the ties by length»

Data is not labeled in any way

UNSUPERVISED

Divide by similarity

CLUSTERING

«Split up similar clothing into stacks»

DIMENSION REDUCTION (generalization)

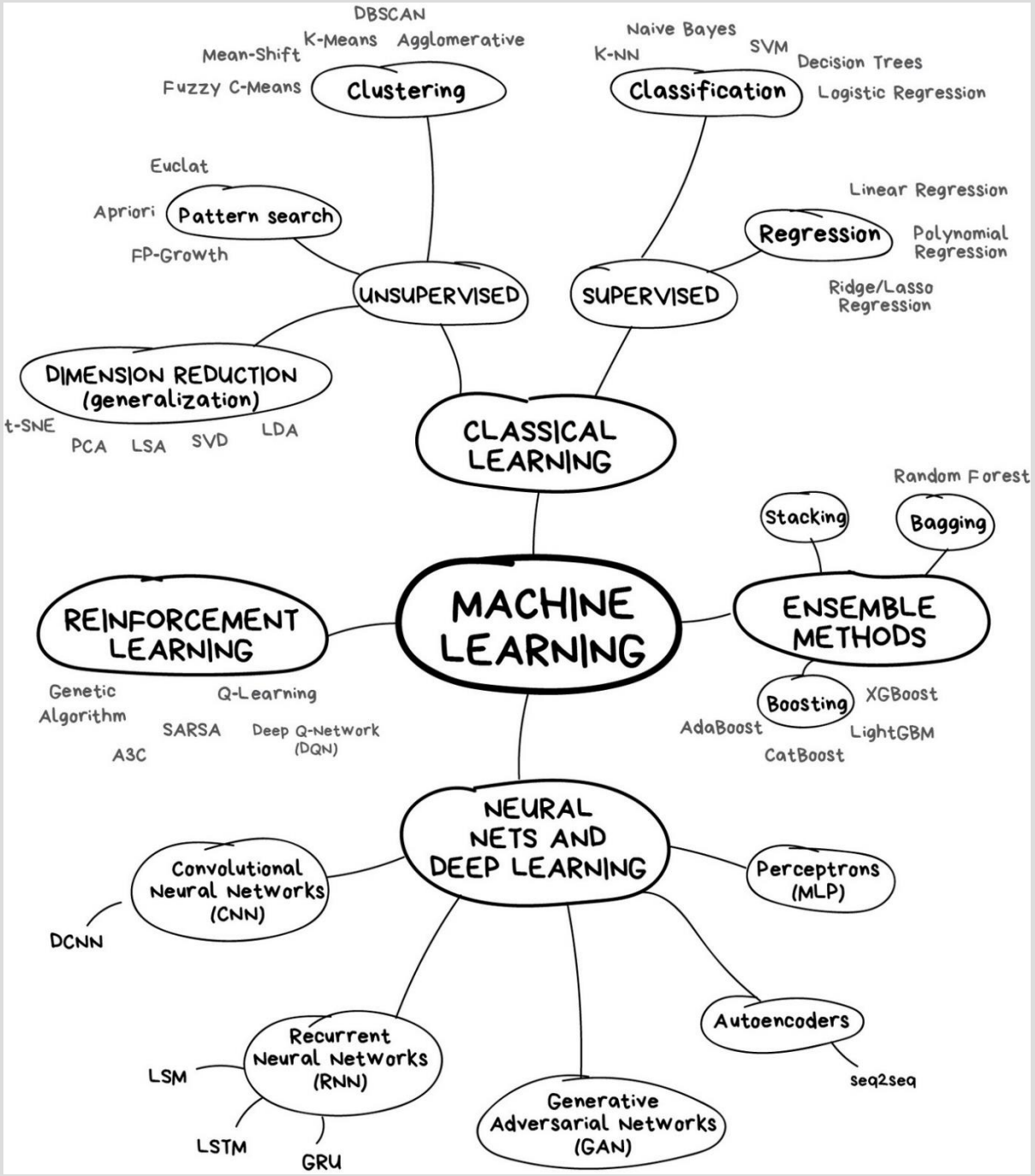
«Make the best outfits from the given clothes»

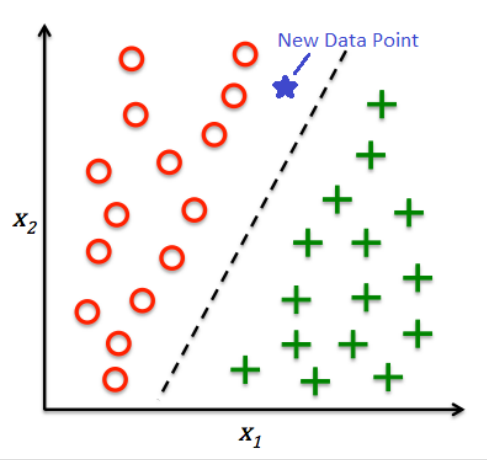
Find hidden dependencies

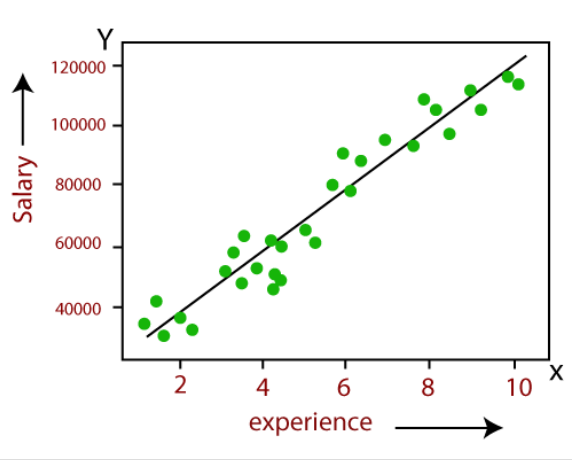
Identify sequences

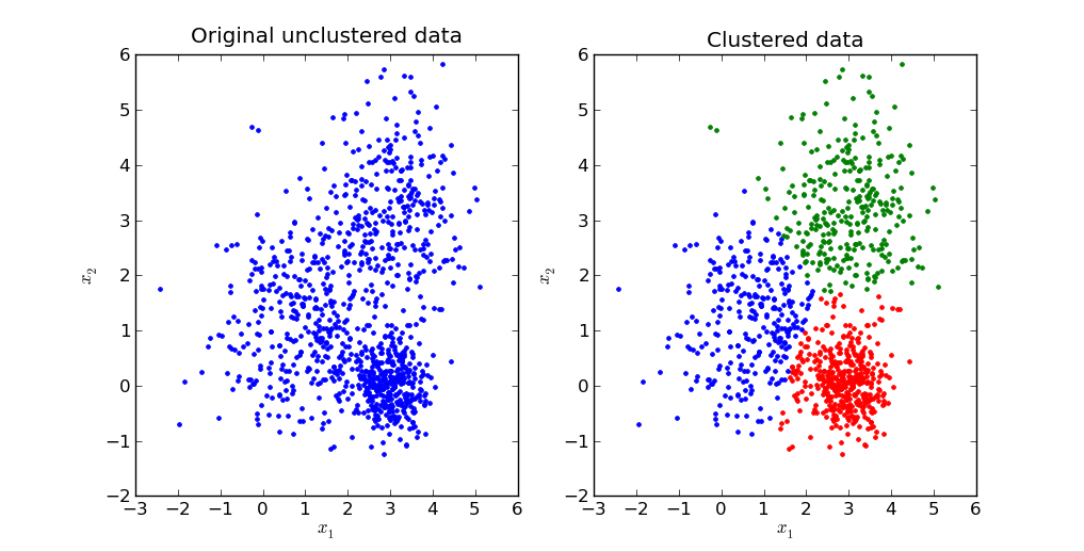
ASSOCIATION

«Find what clothes I often wear together»


MACHINE LEARNING
CLASSICAL LEARNING
UNSUPERVISED
Clustering
DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Pattern search
Euclat
Apriori
FP-Growth
DIMENSION REDUCTION (generalization)
t-SNE
PCA
LSA
SVD
LDA
SUPERVISED
Classification
Naive Bayes
K-NN
SVM
Decision Trees
Logistic Regression
Regression
Linear Regression
Polynomial Regression
Ridge/Lasso Regression
REINFORCEMENT LEARNING
Genetic Algorithm
Q-Learning
SARSA
Deep Q-Network (DQN)
A3C
ENSEMBLE METHODS
Stacking
Bagging
Random Forest
Boosting
XGBoost
AdaBoost
LightGBM
CatBoost
NEURAL NETS AND DEEP LEARNING
Convolutional Neural Networks (CNN)
DCNN
Perceptrons (MLP)
Recurrent Neural Networks (RNN)
LSM
LSTM
GRU
Generative Adversarial Networks (GAN)
Autoencoders
seq2seq







Aprendizaje supervisado

Objetivo principal: Categorizar nuevos datos en base a características conocidas (i.e. clasificar).

Algoritmos habituales: Regresión, Árbol de Decisión, Random Forest, SVM, AdaBoost, Redes Neuronales, Sentiment Analysis

Aprendizaje no supervisado

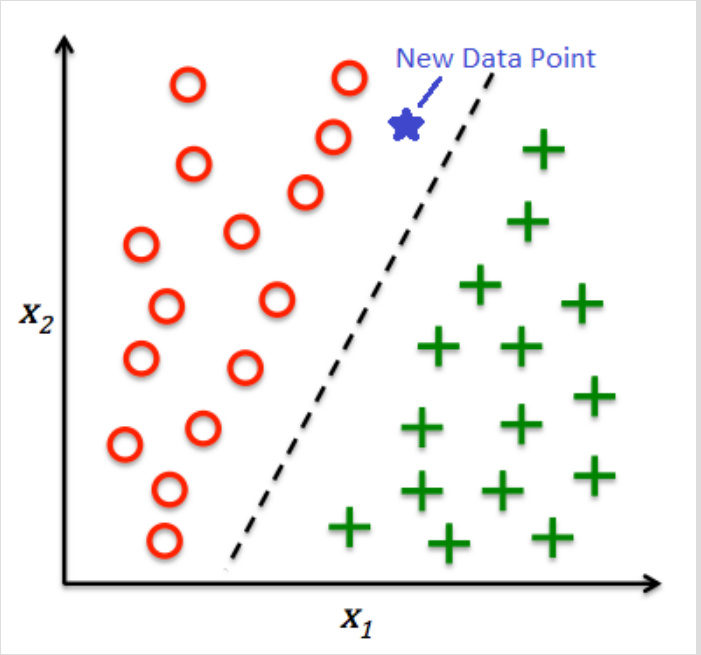
Objetivo principal: Categorizar datos en base a características endógenas subyacentes (i.e. clusterizar).

Algoritmos habituales: K-Means, Componentes Principales, Modularidad, Topic Modeling

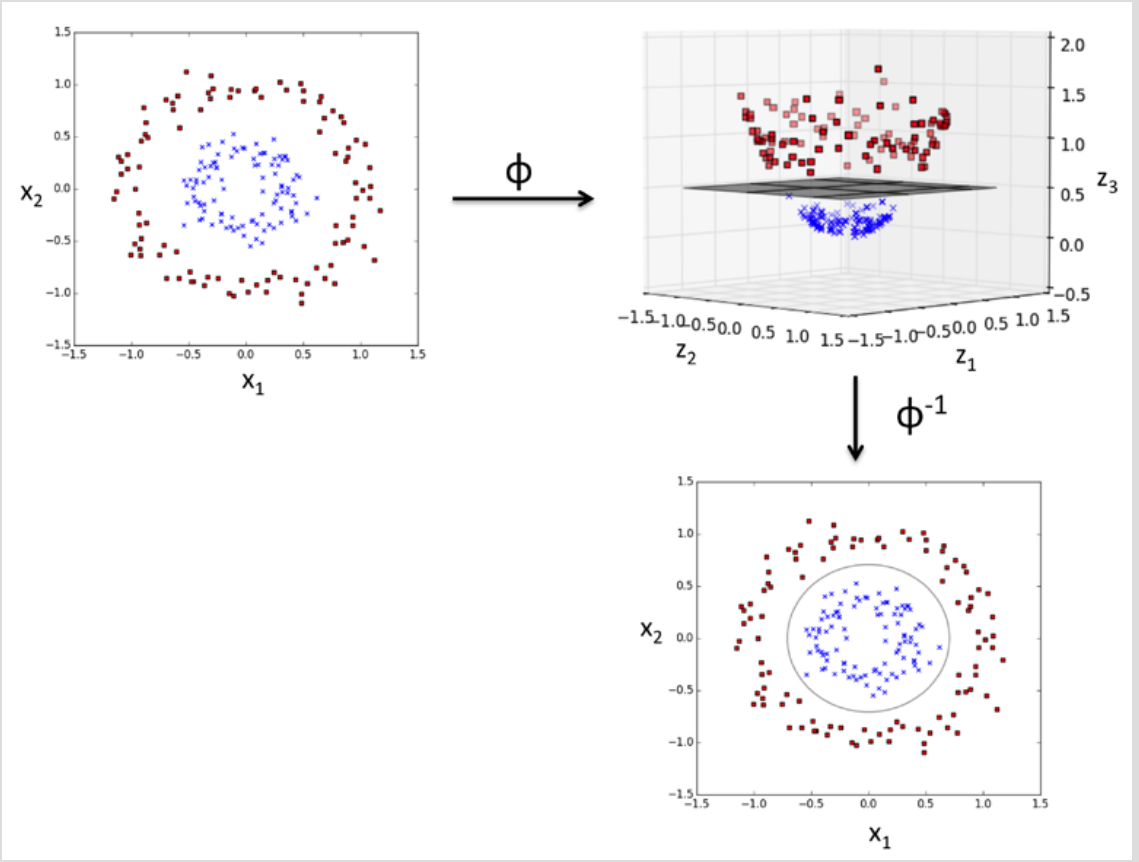
CLASIFICACIÓN (SUPERVISADA)

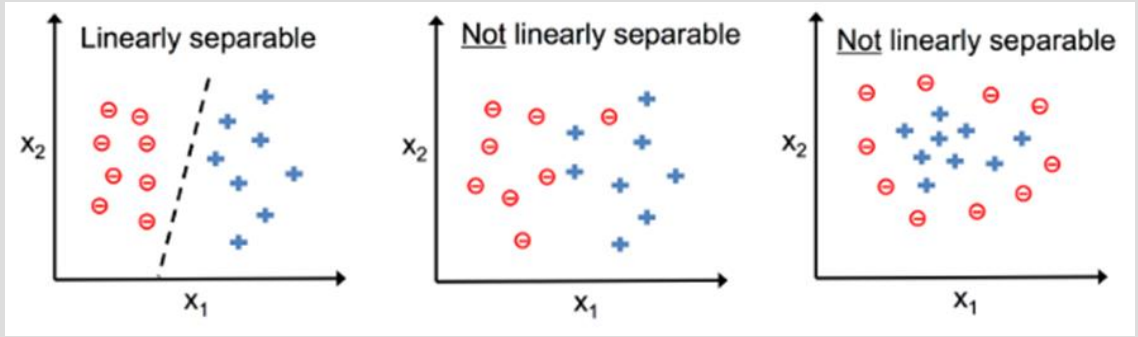
- Imputación de grupos en base a características conocidas o aprendidas
  - Filtros de SPAM en el correo
  - Concesión o no de una hipoteca
  - Detección de lenguas en textos
  - Análisis de sentimiento
  - Identificación de discursos de odio

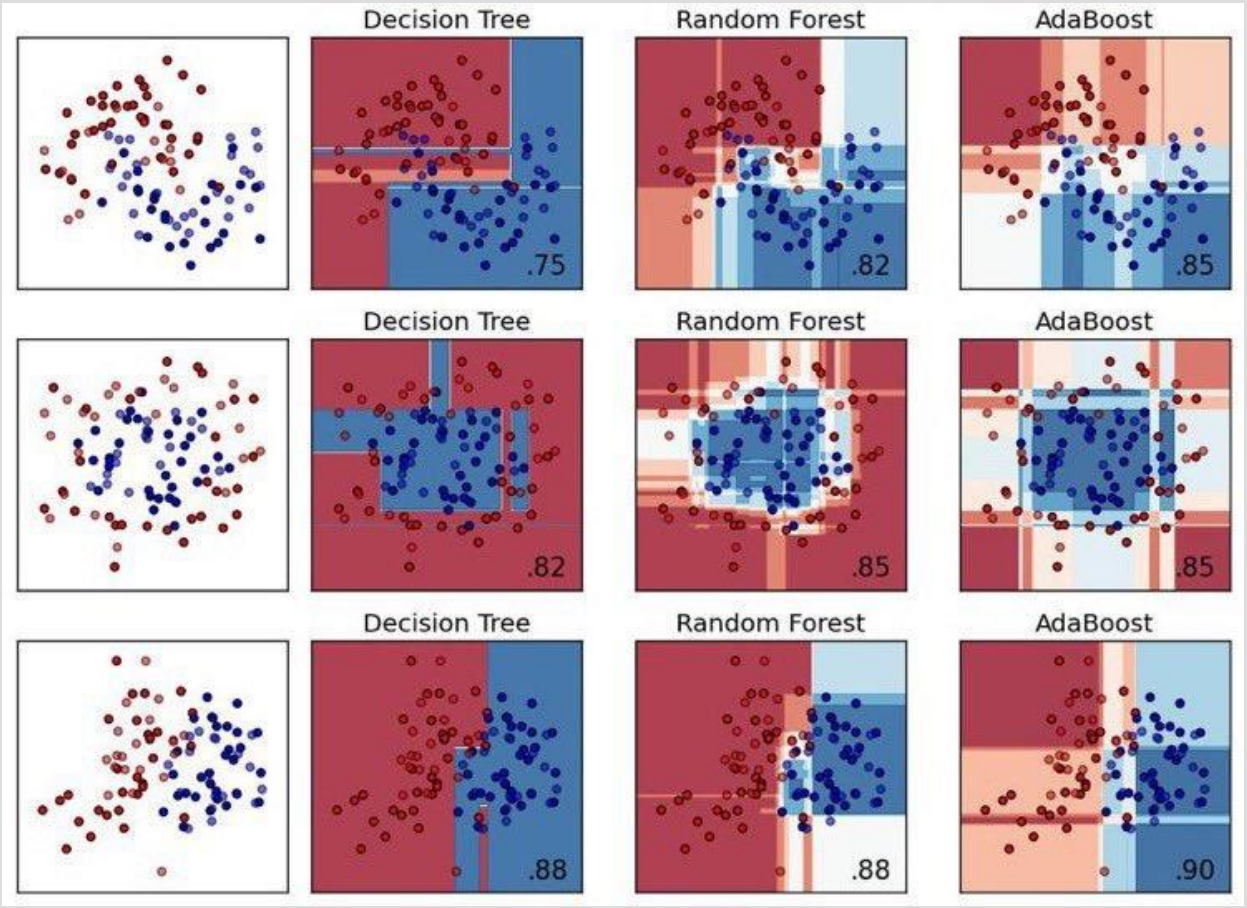
El target de un algoritmo clasificatorio va a ser normalmente*** una variable discreta o categórica




GAME OF THRONES VI
BATTLE OF THE BASTARDS: SHIELD WALL

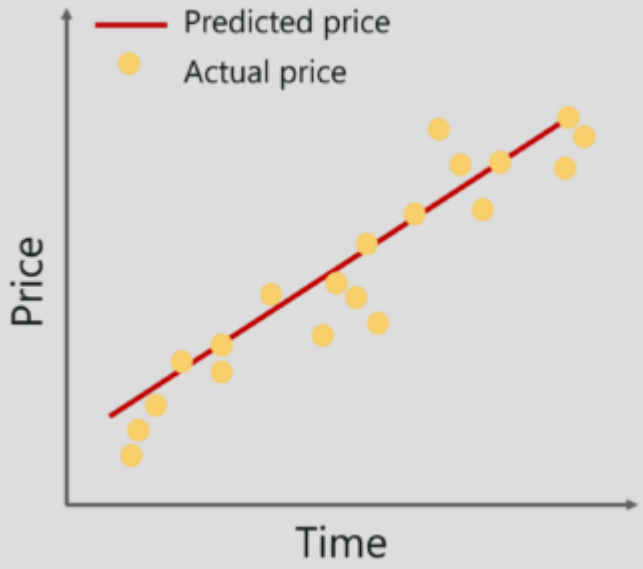

x_2
x_1
ϕ
z_3
z_2
z_1
ϕ^{-1}
x_2
x_1


Linearly separable
Not linearly separable
Not linearly separable
x2
x1
x2
x1
x2
x1

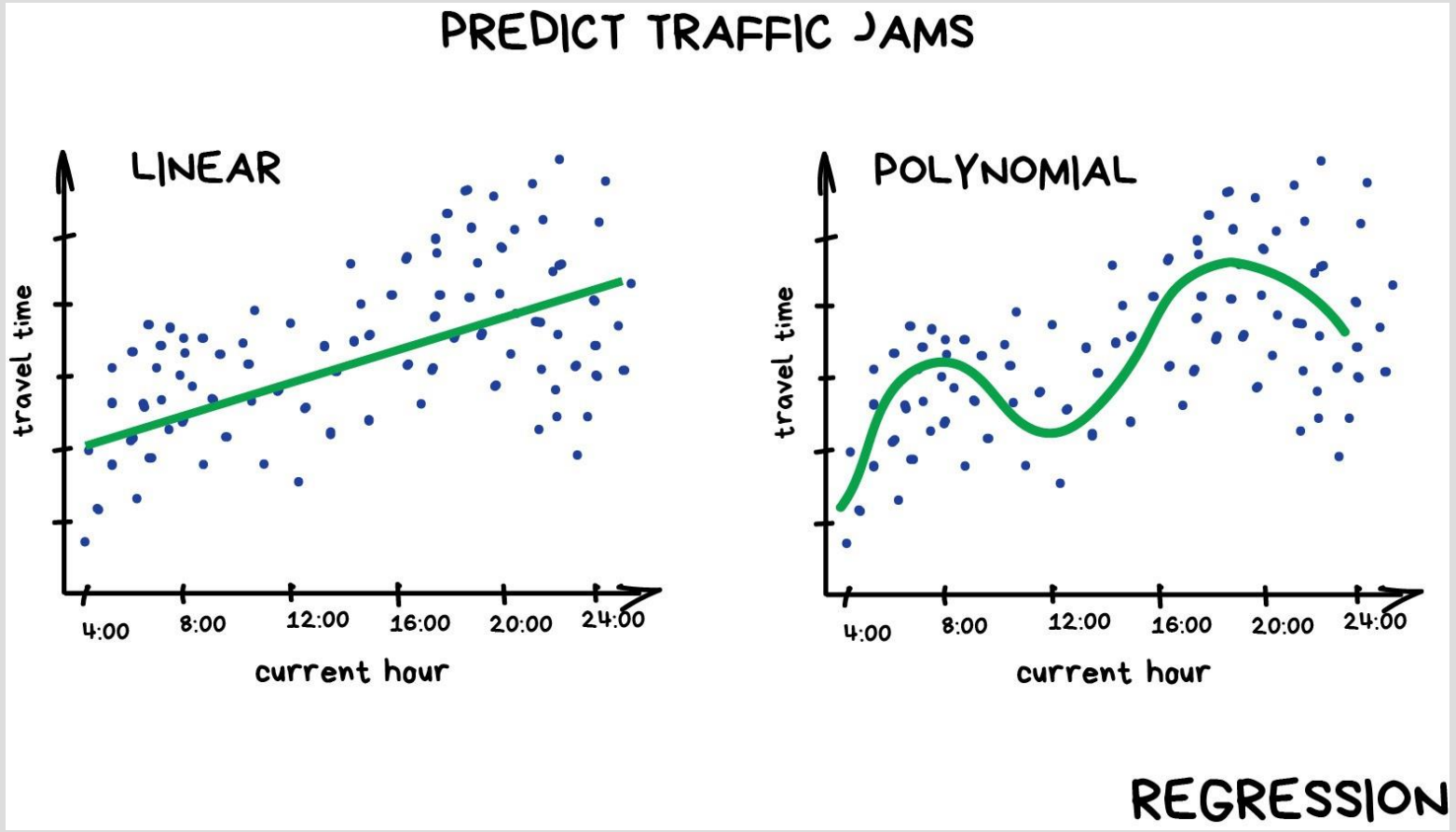

Decision Tree
Random Forest
AdaBoost
.75
.82
.85
Decision Tree
Random Forest
AdaBoost
.82
.85
.85
Decision Tree
Random Forest
AdaBoost
.88
.88
.90

REGRESIÓN (SUPERVISADA)

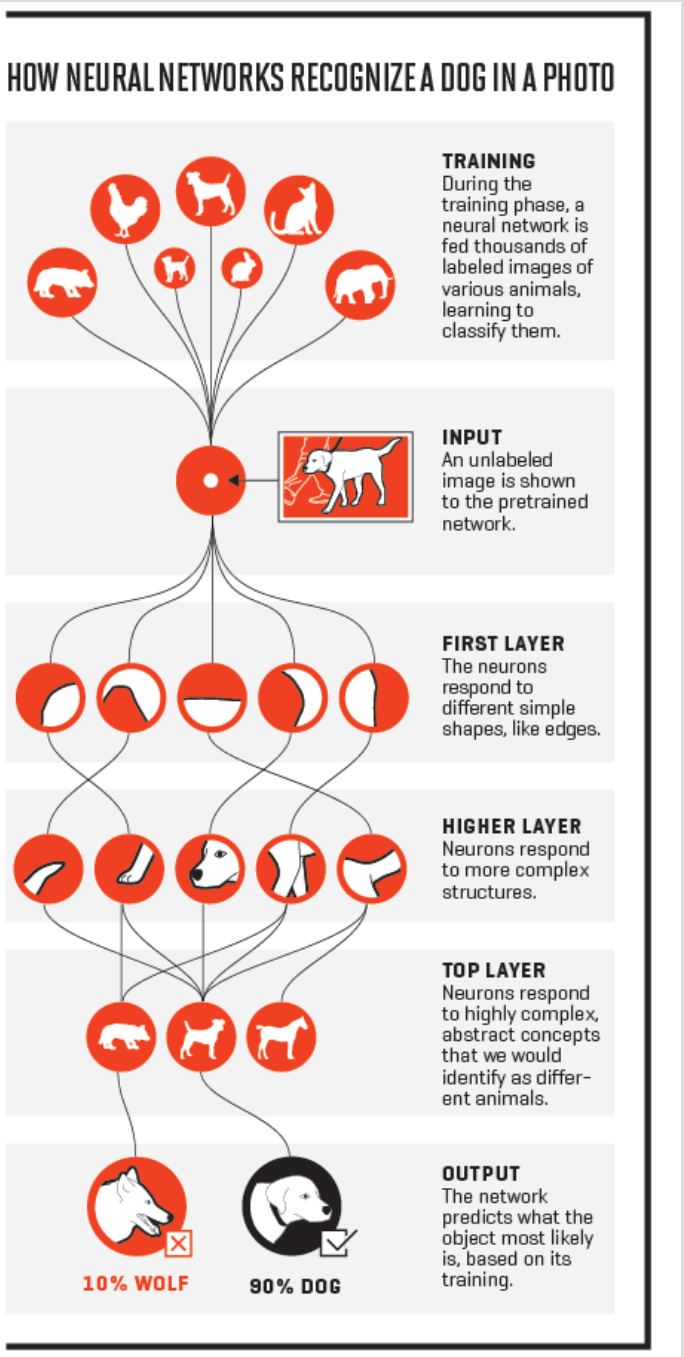
- Predicción de valores numéricos en base a patrones conocidos
  - Predicción bursátil
  - Análisis de tendencias de consumo
  - Prospectiva comercial o financiara
  - Diagnóstico médico



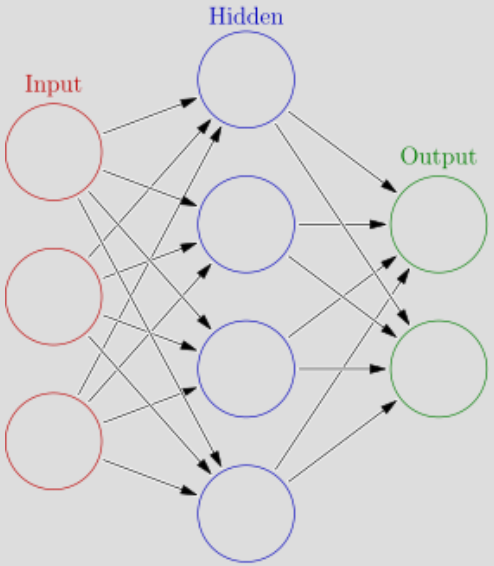
El target de un algoritmo de regresión va a ser siempre una variable numérica y continua

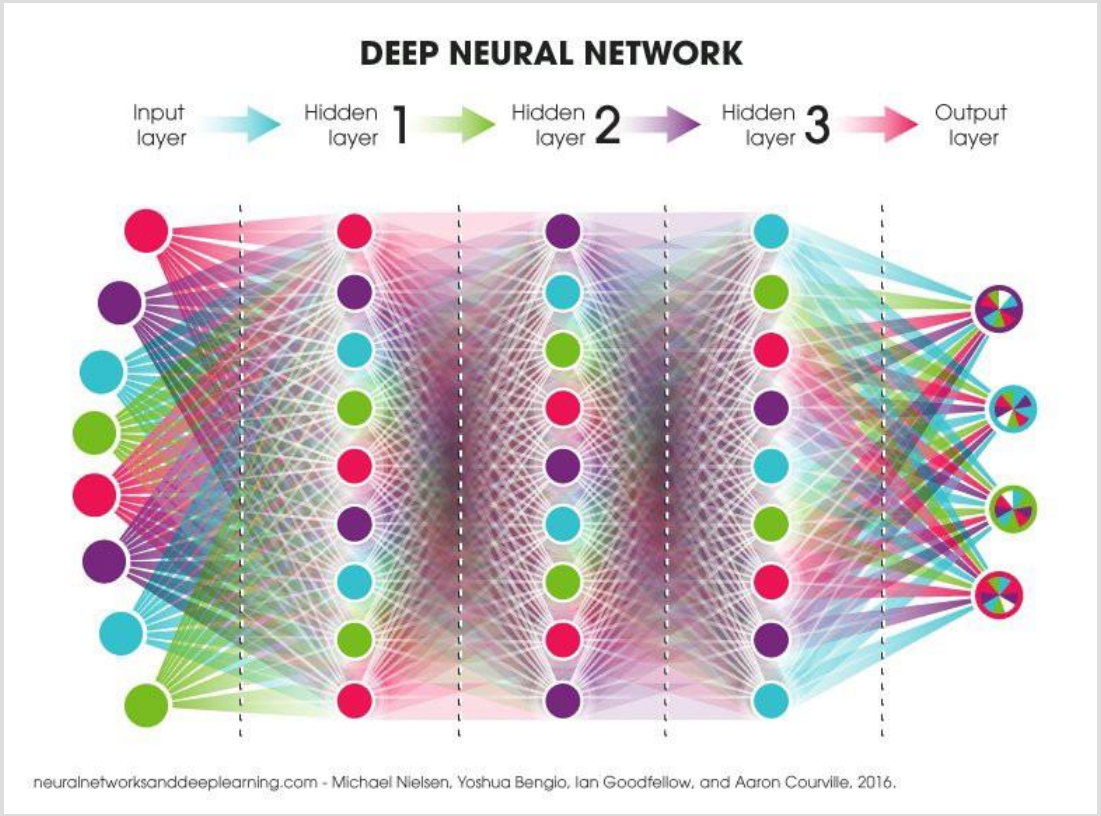

PREDICT TRAFFIC JAMS
LINEAR
travel time
4:00
8:00
12:00
16:00
20:00
24:00
current hour
POLYNOMIAL
travel time
4:00
8:00
12:00
16:00
20:00
24:00
current hour
REGRESSION


OVERFITTING



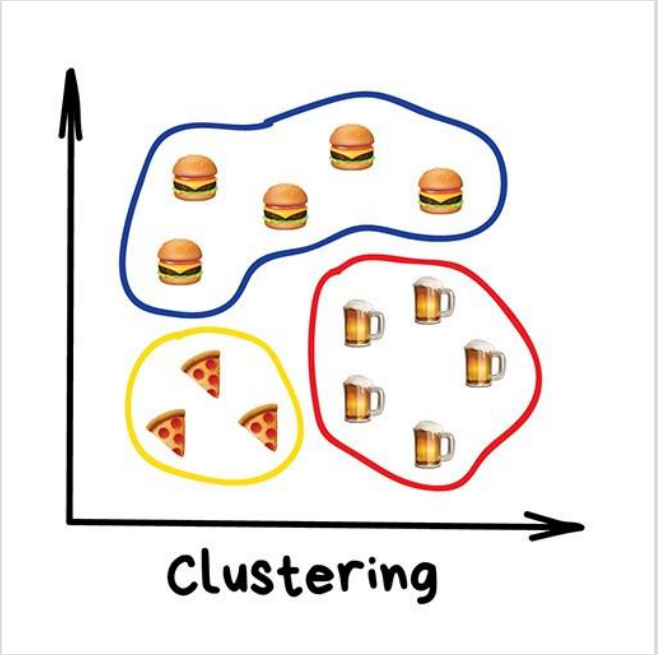
REDES NEURONALES Y DEEP LEARNING

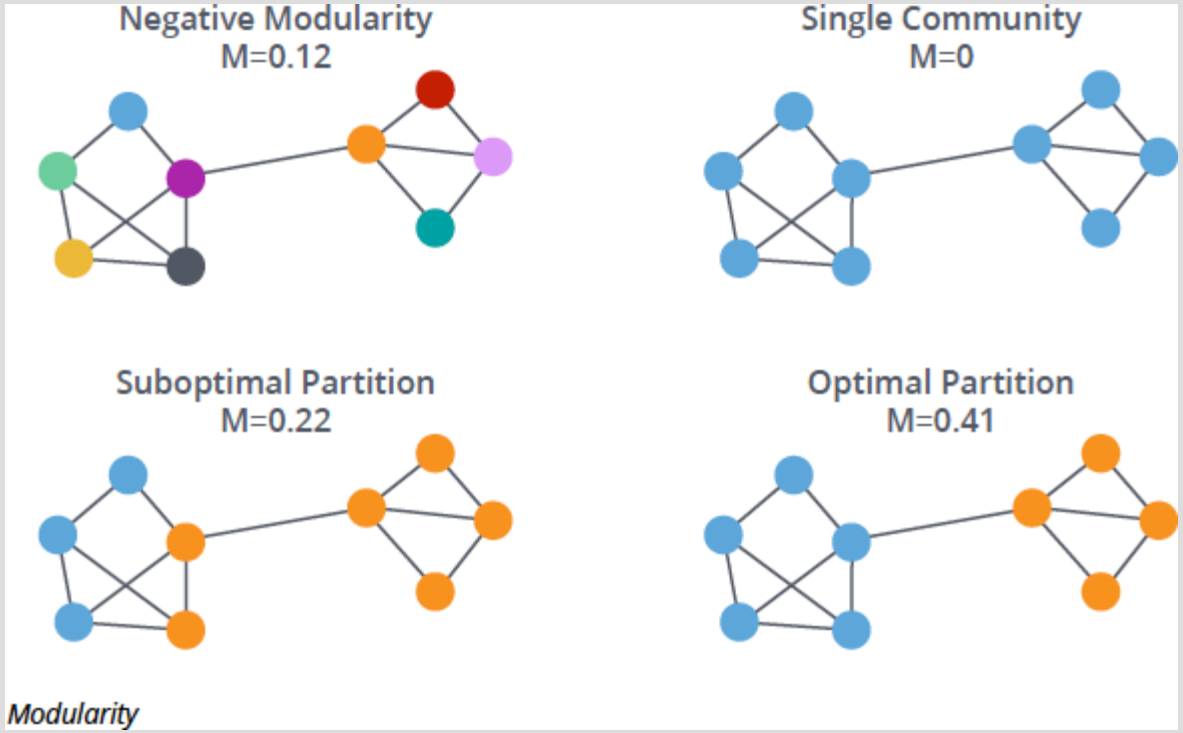




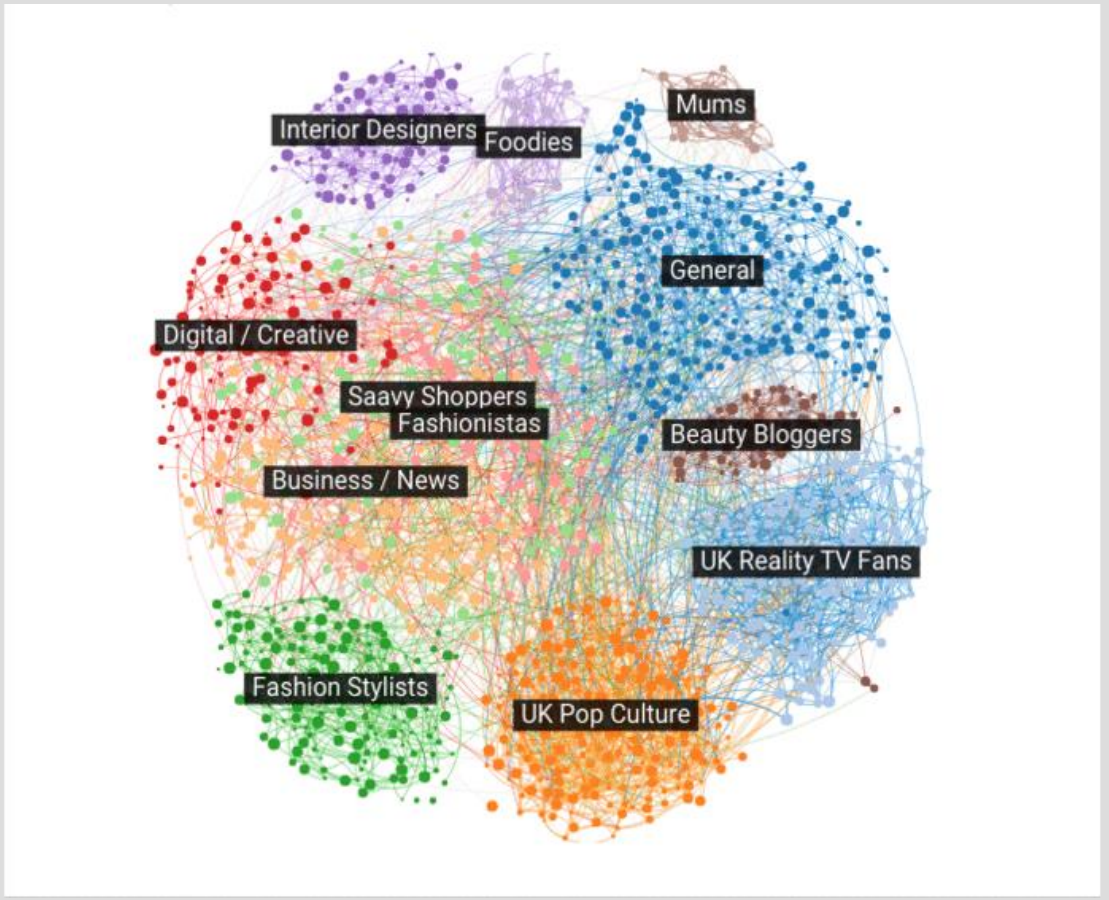
CLUSTERING (NO SUPERVISADO)

- Divide los casos sin basarse en características conocidas
  - Segmentación de mercados
  - Agrupación de textos (Topic Modeling)
  - Agrupación de imágenes (Image Embeddings)
  - Descubrimiento de patrones "ocultos" en los datos




Negative Modularity
M=0.12
Single Community
M=0
Suboptimal Partition
M=0.22
Optimal Partition
M=0.41

Modularity


Interior Designers
Foodies
Mums
General
Digital / Creative
Saavy Shoppers
Fashionistas
Beauty Bloggers
Business / News
UK Reality TV Fans
Fashion Stylists
UK Pop Culture

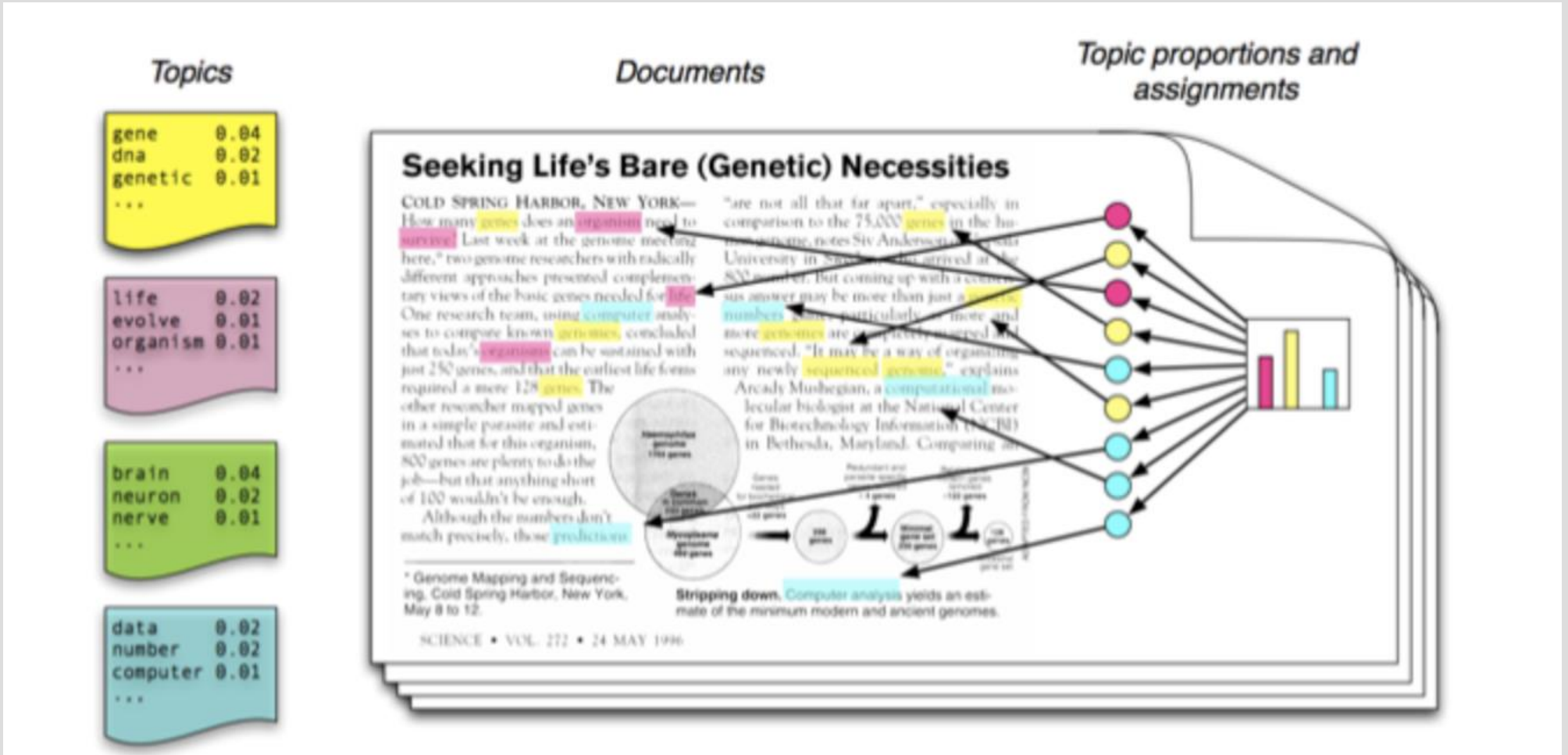


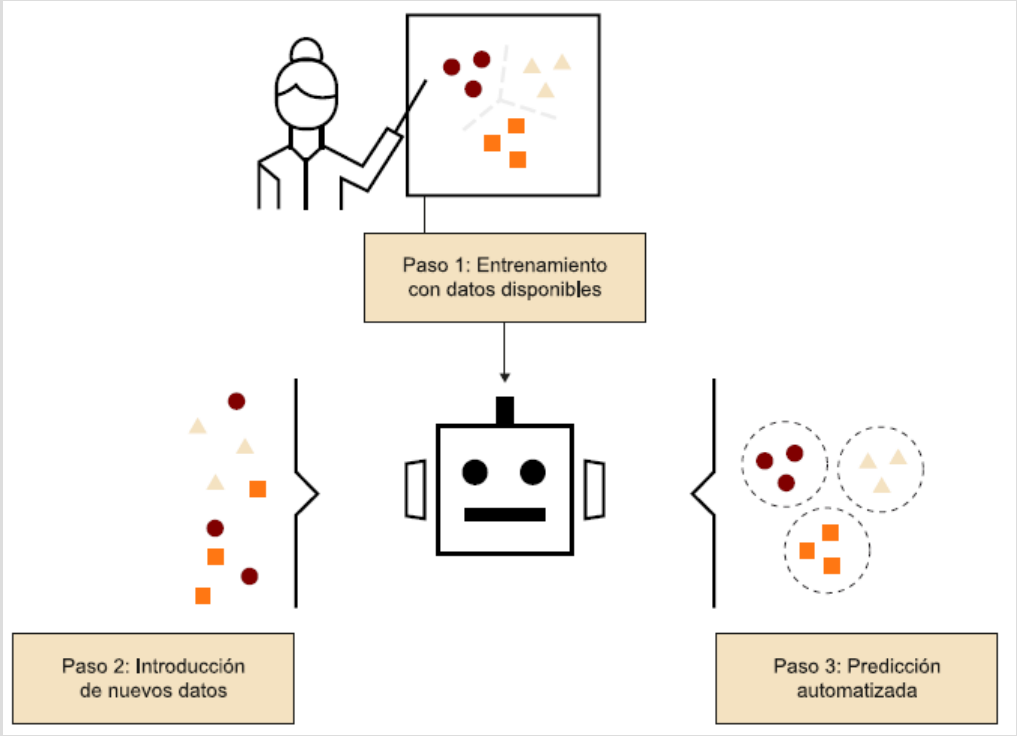
Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

FAIR DATA TREATMENT / JUSTICIA DE DATOS

Los algoritmos no solamente pueden aprender de nuestros propios sesgos y de los sesgos contenidos en los datos, sino que son capaces de amplificarlos y magnificarlos.

Para corregirlos podemos implementar…

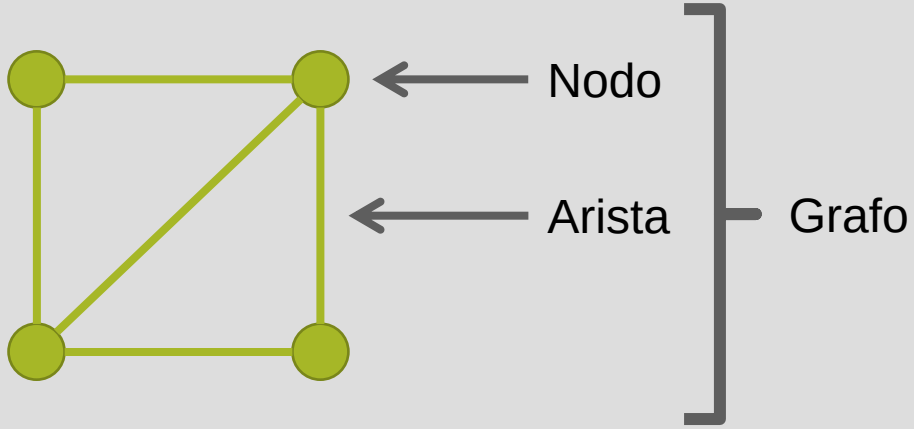
1. **Intervenciones sobre los datos de entrenamiento.**
2. **Intervenciones sobre el algoritmo inductor**.
3. **Intervenciones sobre los resultados del algoritmo**.



Edizel, Bora; Bonchi, Francesco; Hajian, Sara; Panisson, André; Tassa, Tamir (2020). «FaiRecSys: Mitigating algorithmic bias in recommender systems». *International Journal of Data Science and Analytics* (vol. 9, núm. 2, págs. 197-213). Nueva York: Springer.

EL ANÁLISIS DE REDES SOCIALES

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

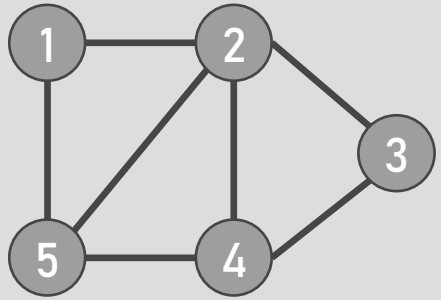
Matriz de adyacencias

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) }

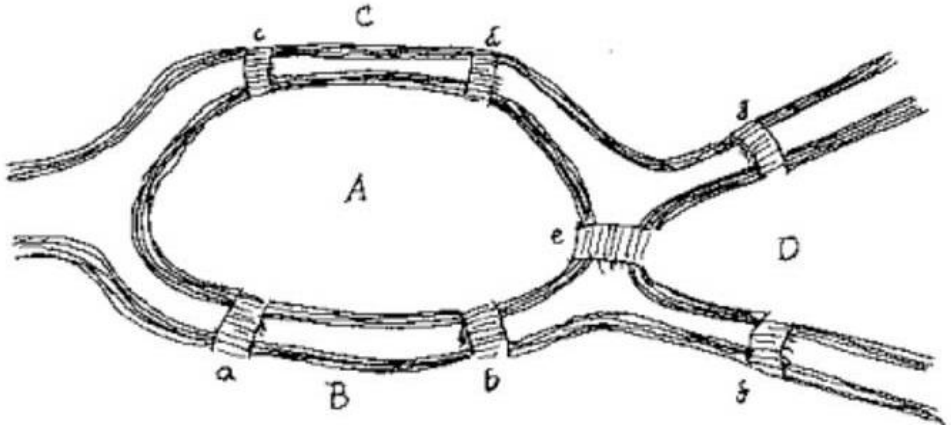
Conjunto de aristas

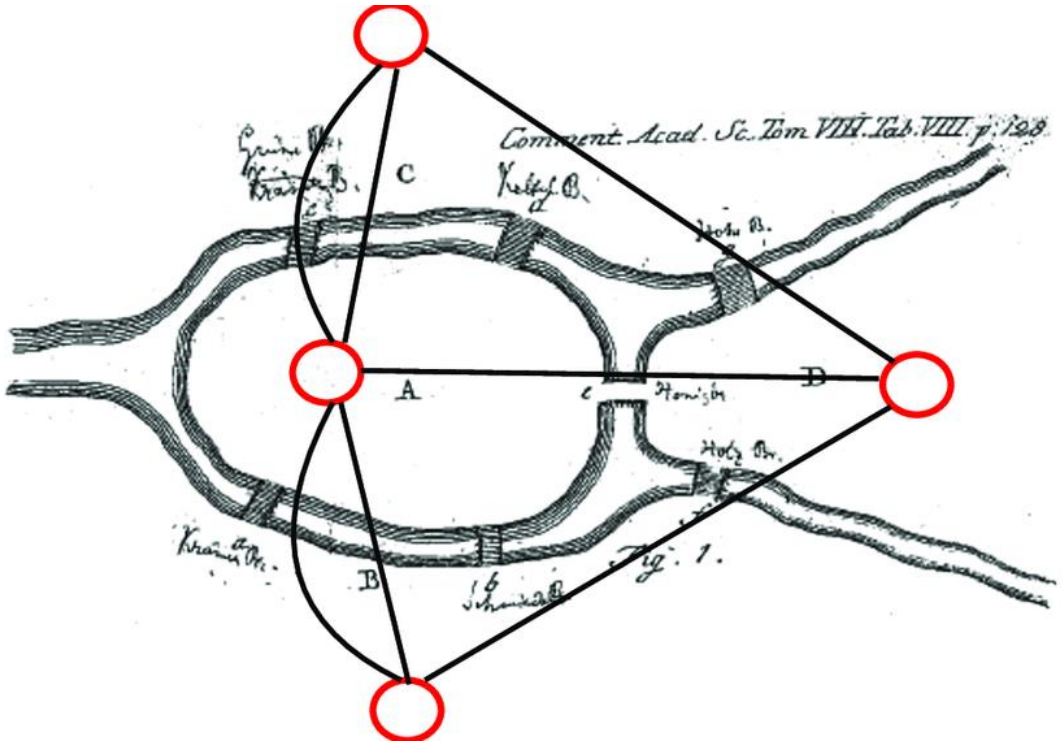


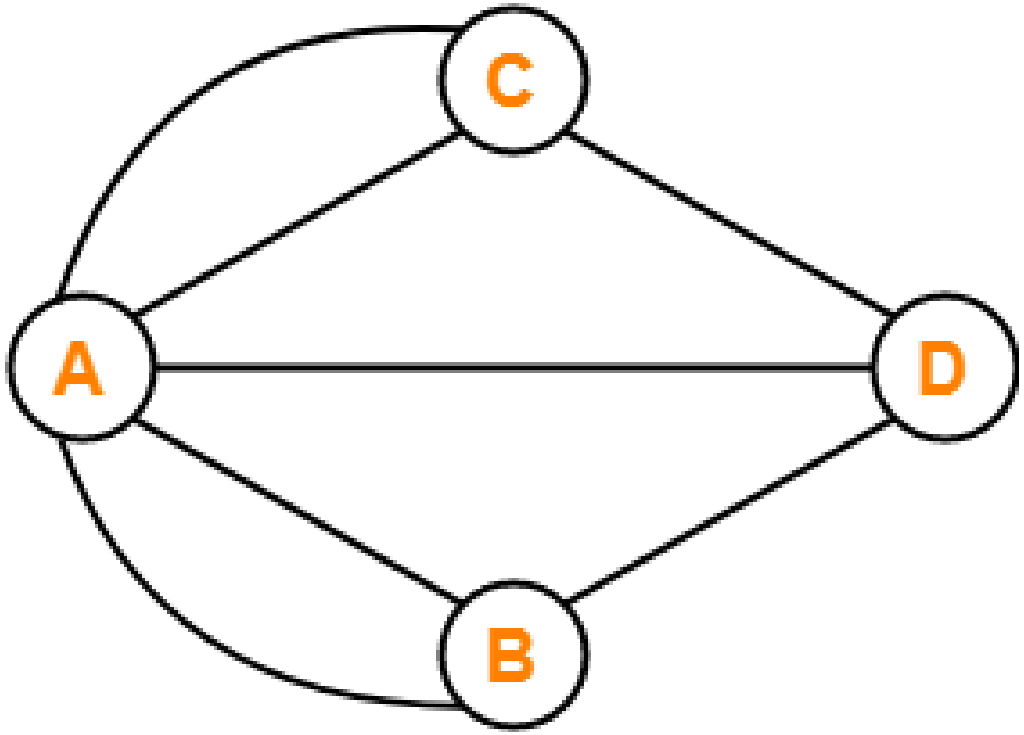
Grafo

La **teoria de grafos**

Euler


c
C
f
g
A
e
D
a
B
b
h


Comment. Acad. Sc. Tom. VIII. Tab. VIII. p. 128
C
A
c
D
B
Fig. 1.


C
A
D
B

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)

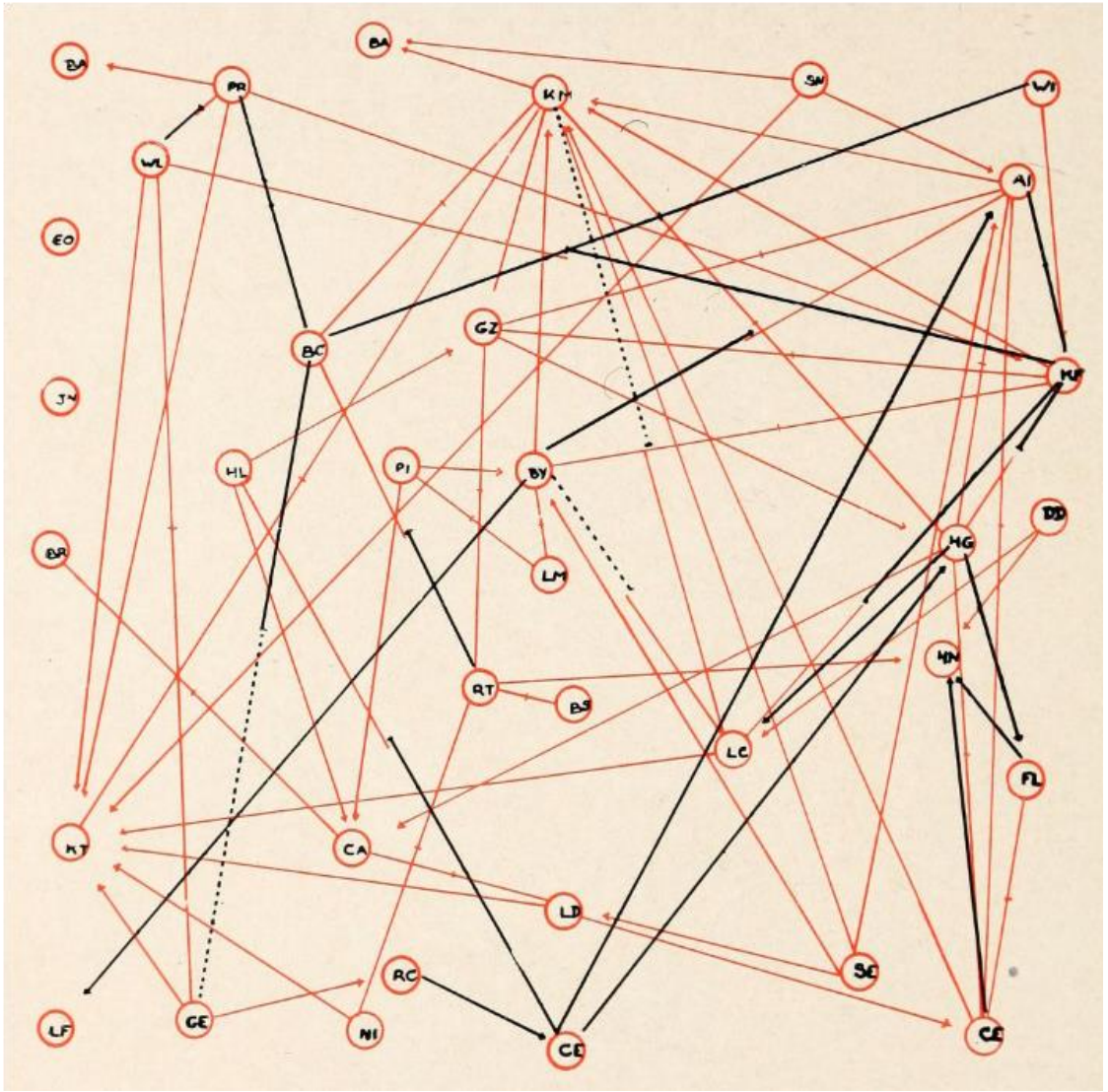
60s – 90s

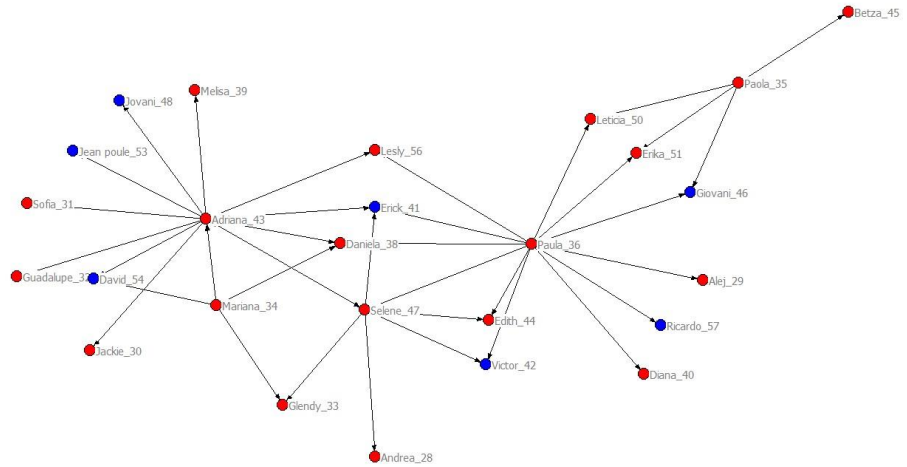
- Kadushin y Blau (Columbia-Chicago)
- Deutsch, de Sola Pool y Kochen (MIT)
- Milgram (Stanford)
- White (MIT-Harvard-Columbia)
- Mokken ,Anthonisse y Stokman (Amsterdam)
- Freeman – Sunshine (Syracuse)
- Granovetter (Harvard)
- Burt (Chicago)

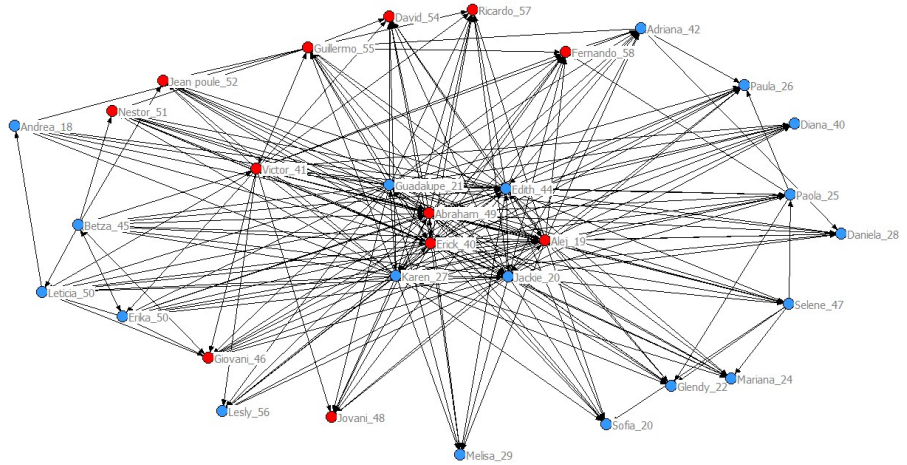
90s – 2019

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

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

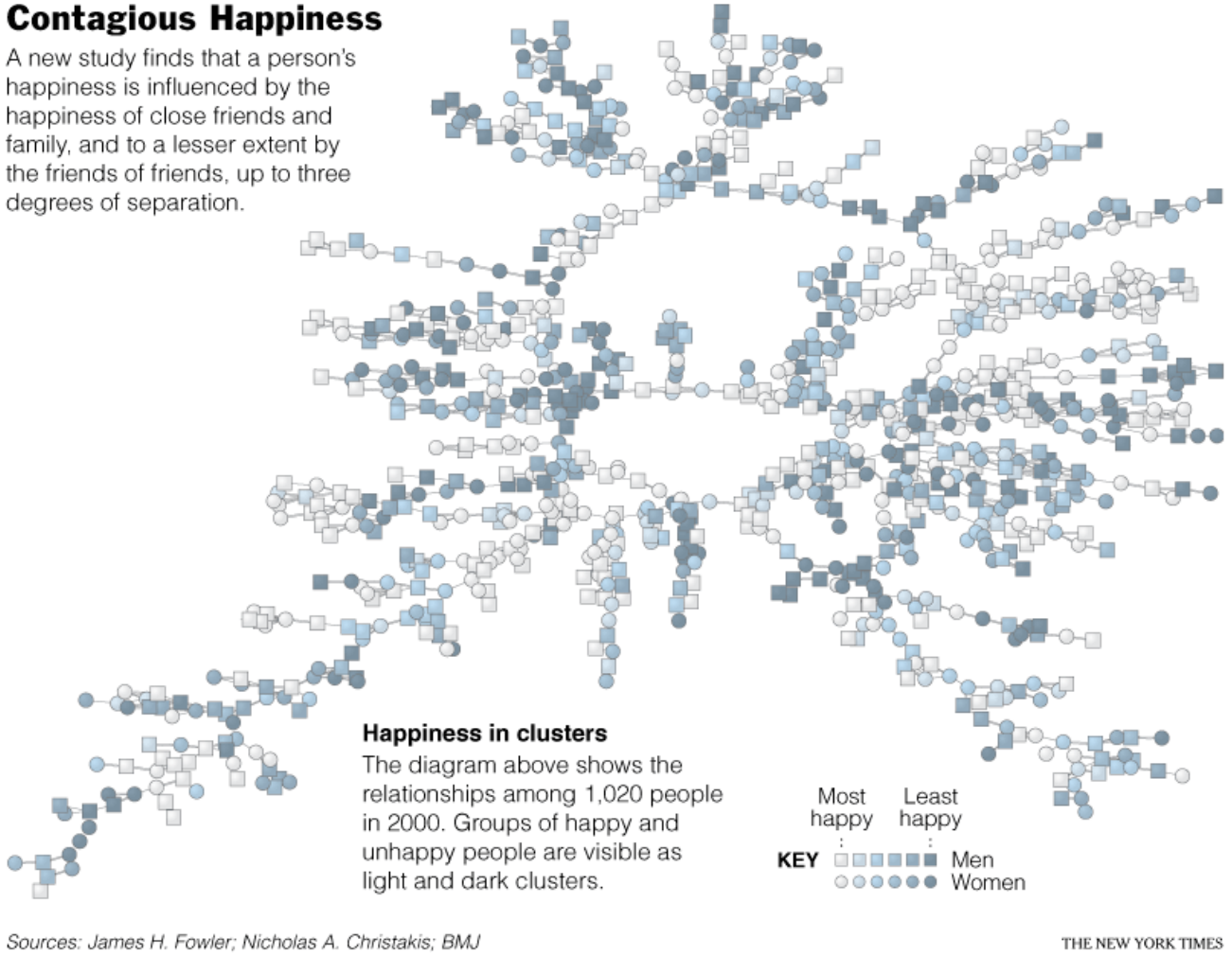




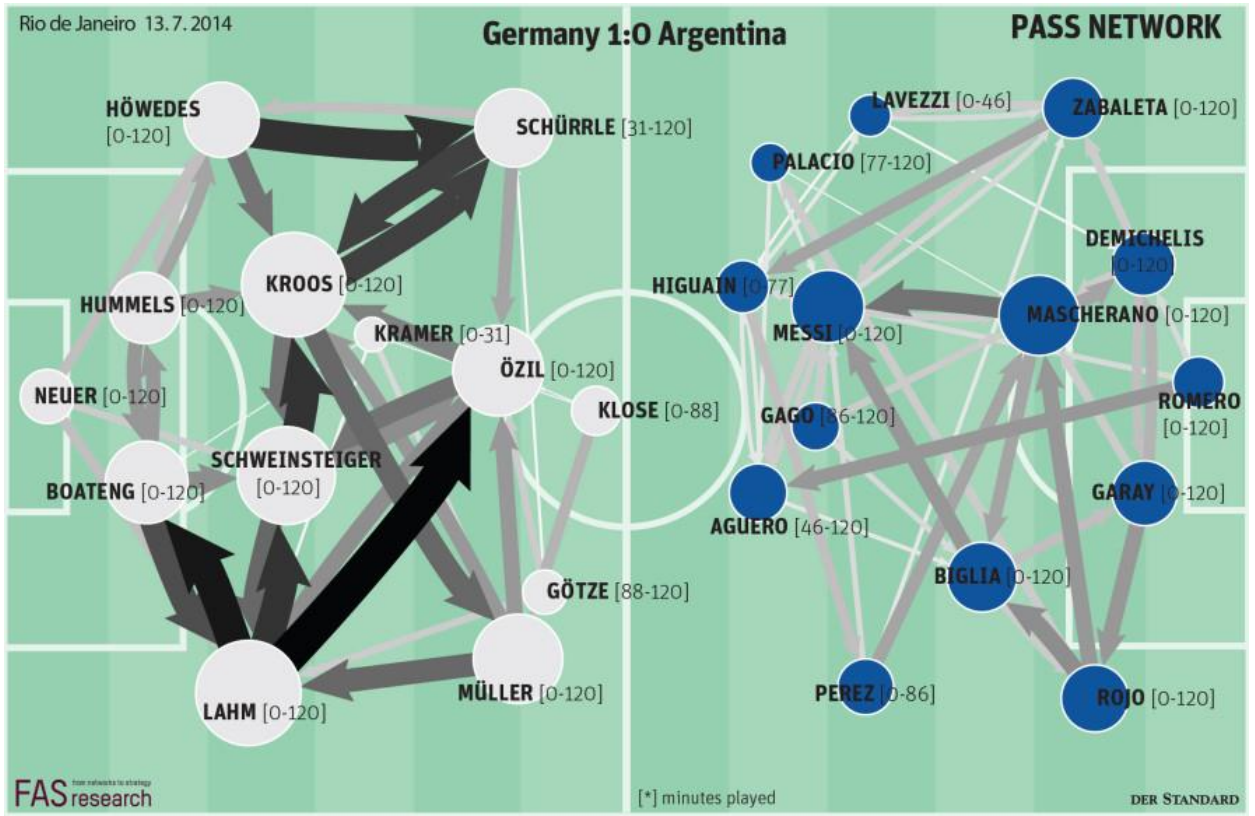


E.G. POPULARIDAD VS. ANTIPATÍA (MAYA JARIEGO Y VIDAL RAMOS, 2014)

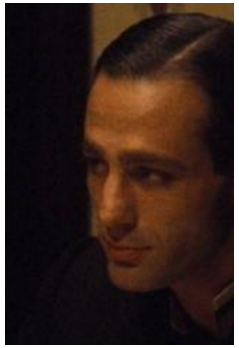
E.G. LA DIFUSIÓN DE LA FELICIDAD (FOWLER Y CHRISTAKIS, 2008)



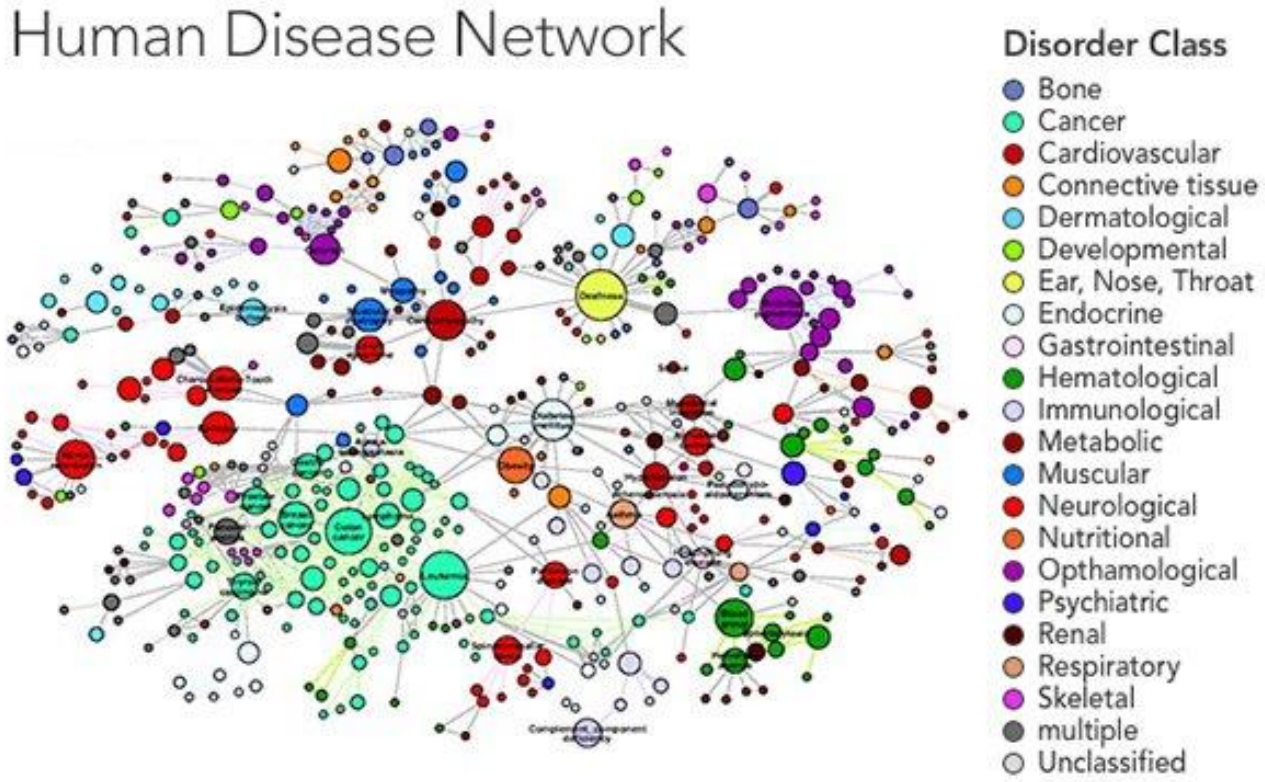
E.G. PASES EN PARTIDO DE FUTBOL (FAS RESEARCH, 2014)



E.G. EL PADRINO II (MOVIEGALAXIES)



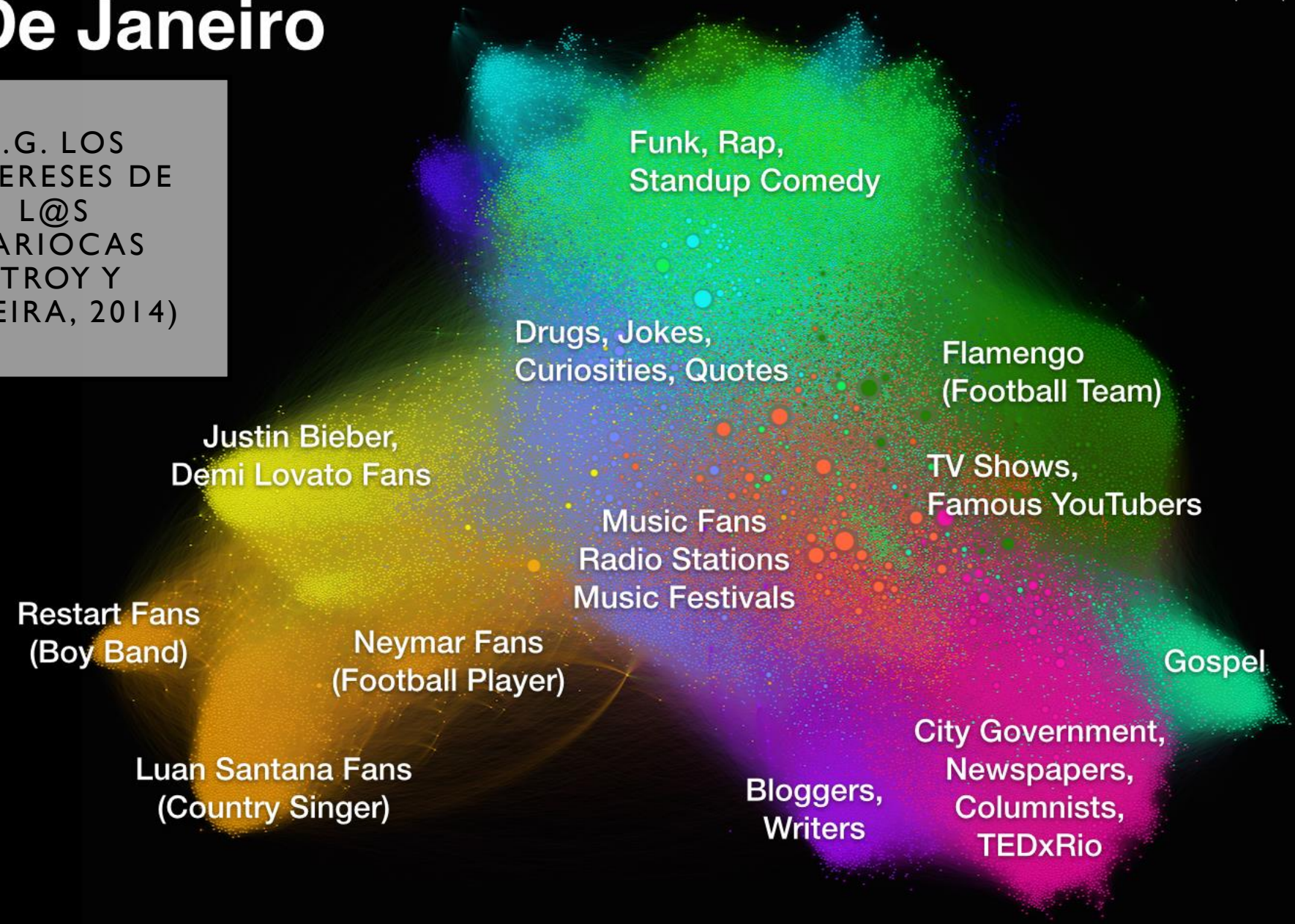
E.G. ENFERMEDADES HUMANAS (THURNER, 2015)



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

Rio De Janeiro

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

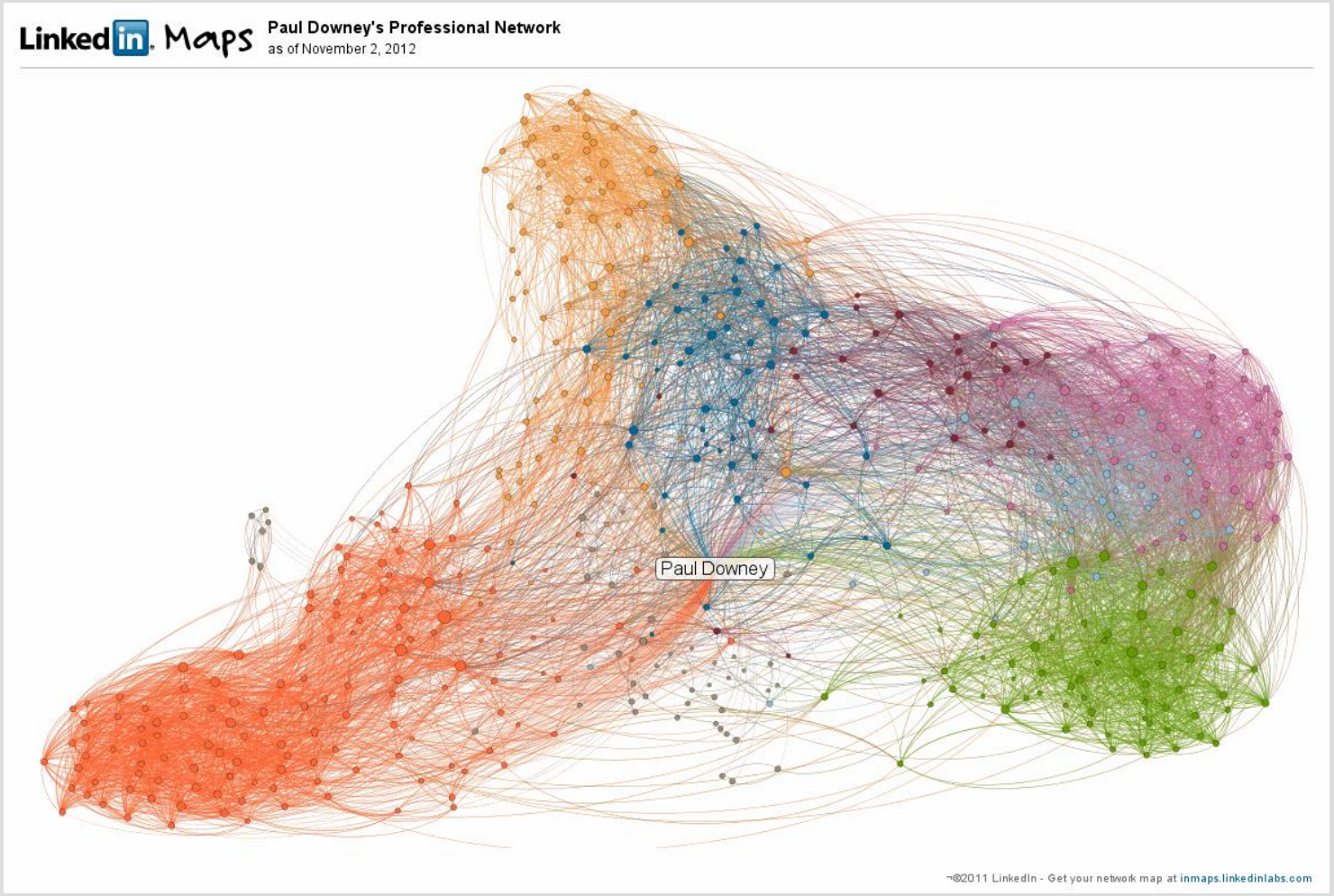


Edited by Marconi Pereira

RED DE AMISTADES EN FACEBOOK

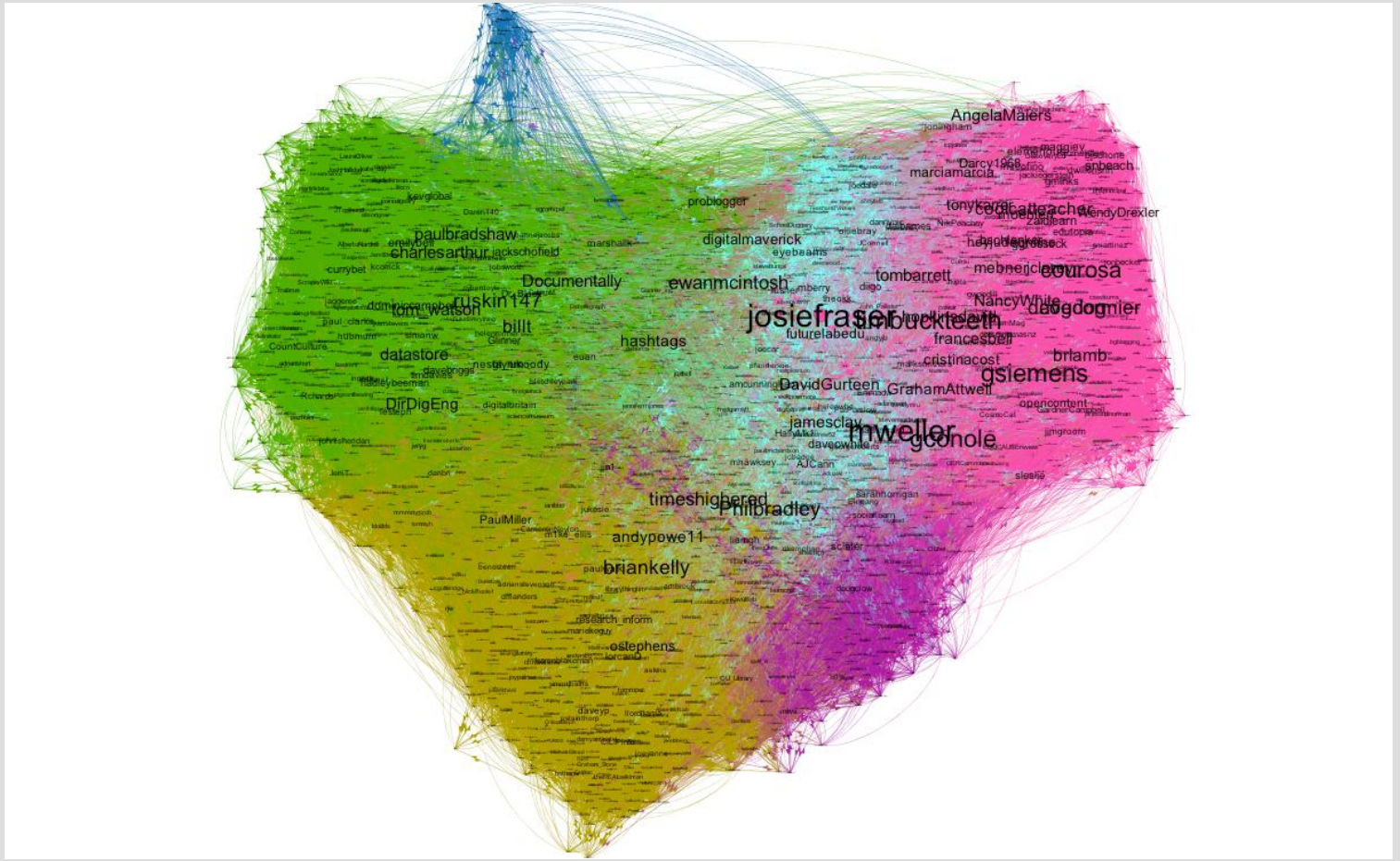
Funcionalidad obsoleta

RED DE AMISTADES EN LINKEDIN

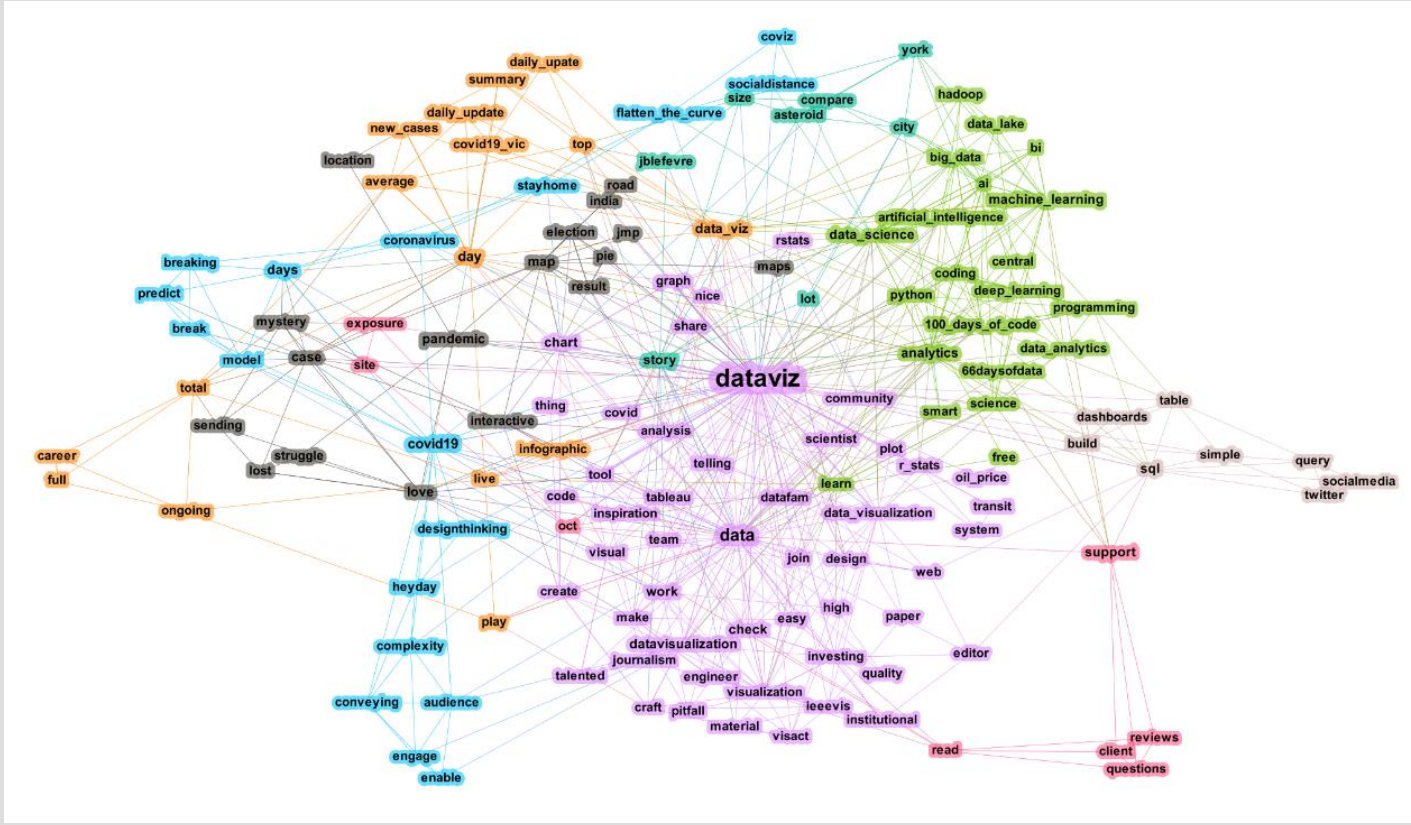


Funcionalidad obsoleta

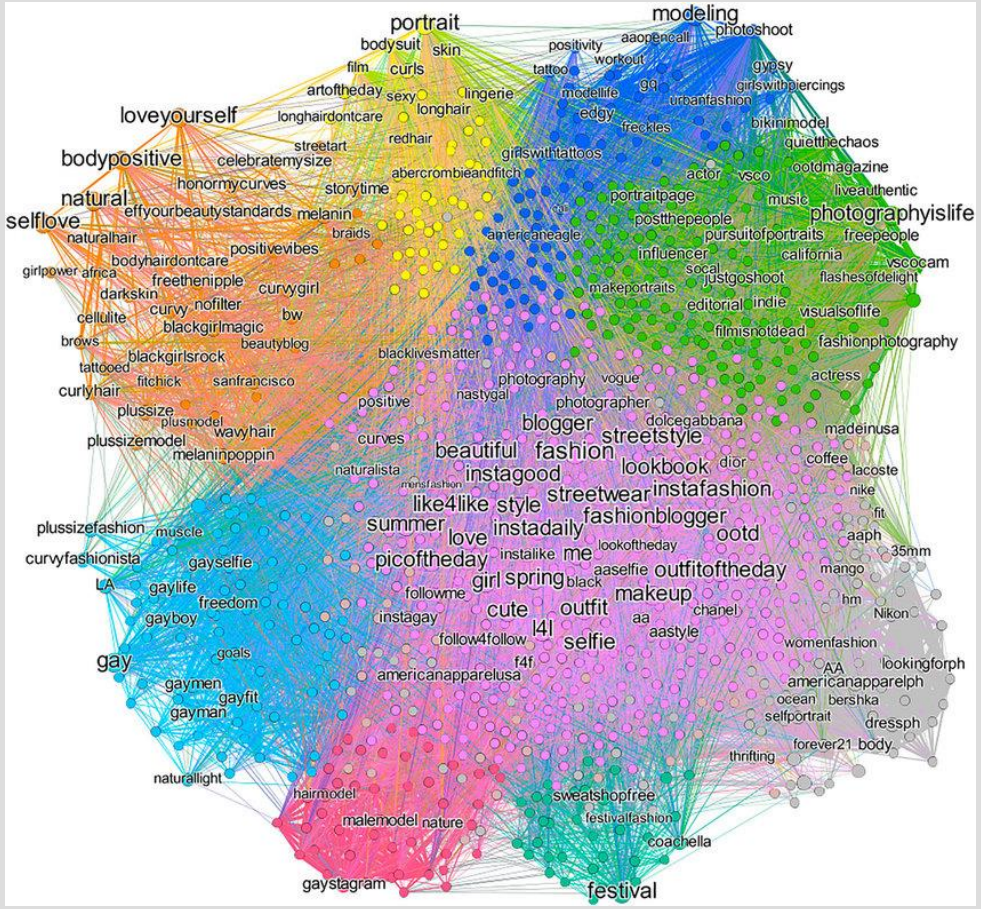
RED DE SEGUIDORES TWITTER



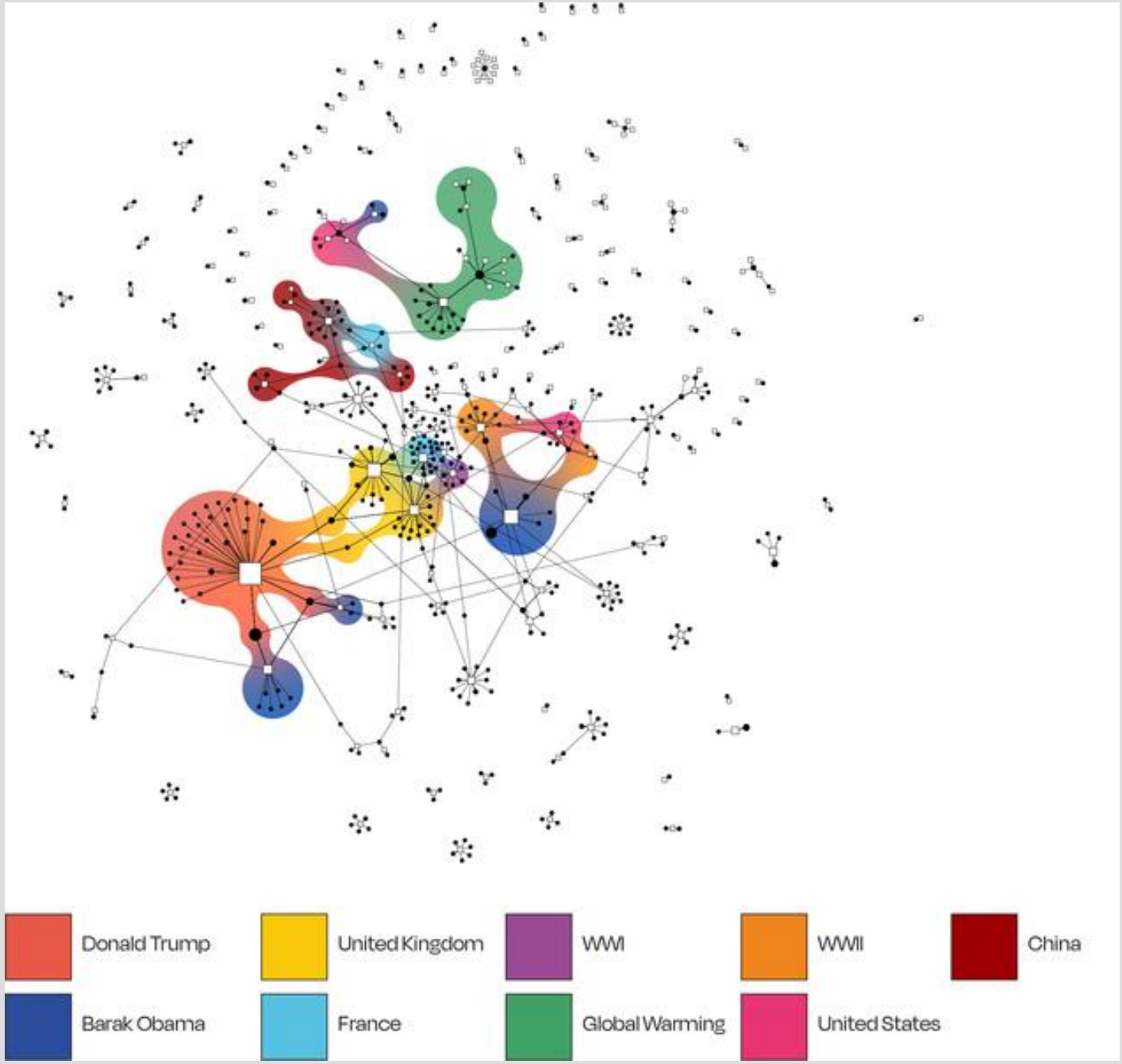
RED DE CONCEPTOS (TWITTER)



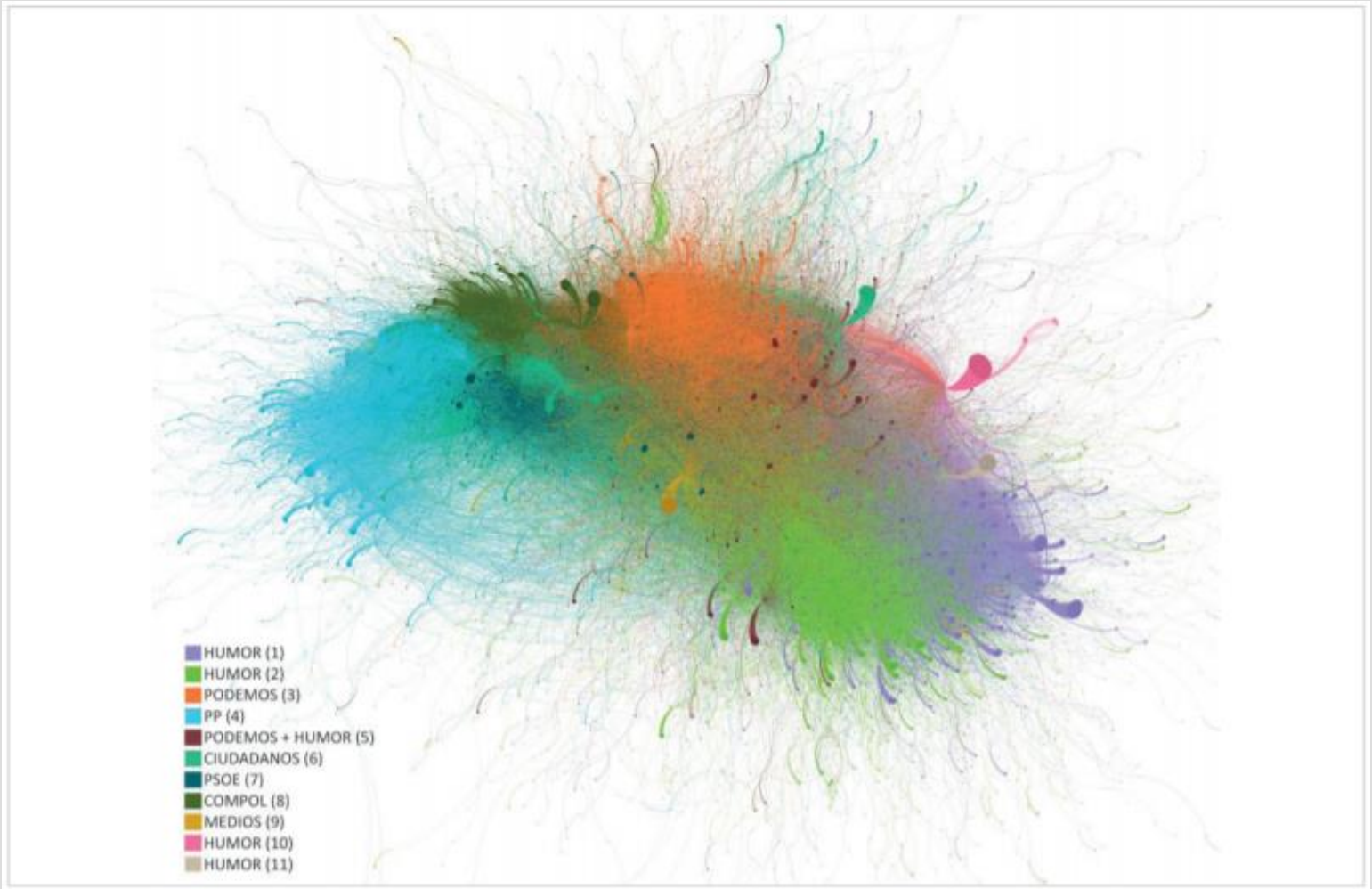
RED DE HASHTAGS (INSTAGRAM)



RED DE AUTORES E IMÁGENES (WIKIPEDIA)



RED DE MENCIONES EN TWITTER

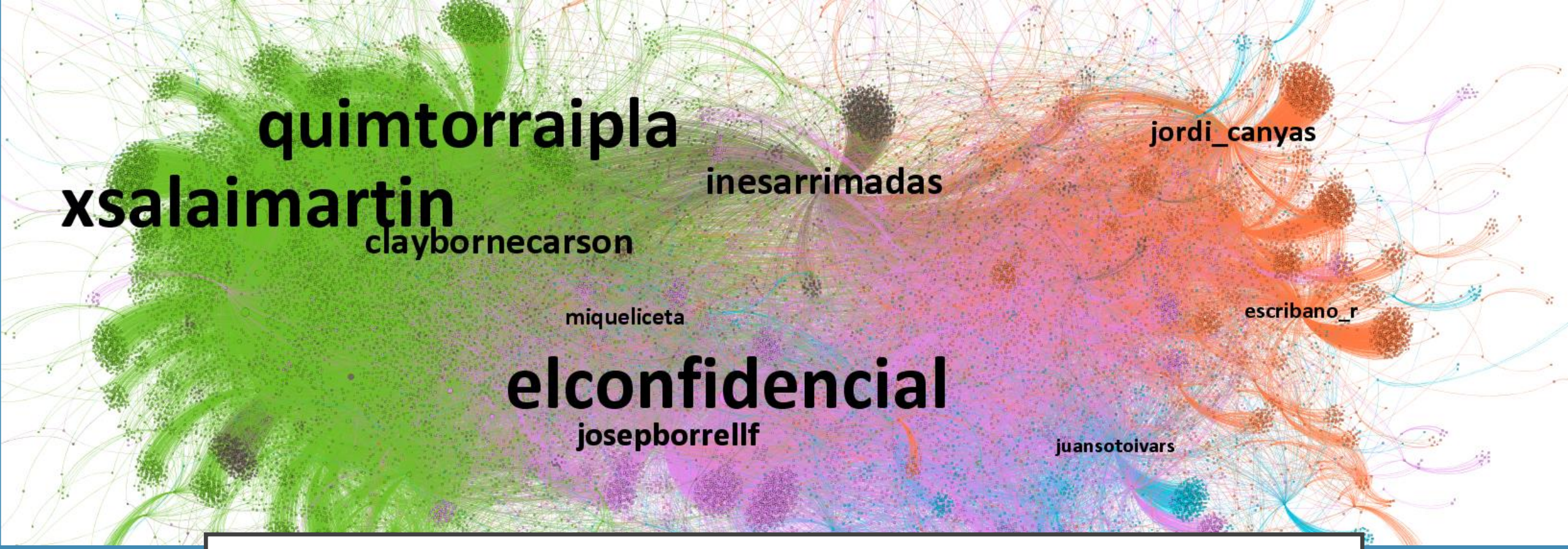


CASO DE ANÁLISIS I
LUTHER KING Y LA INDEPENDENCIA DE CATALUÑA

1. EL IMPACTO INICIAL

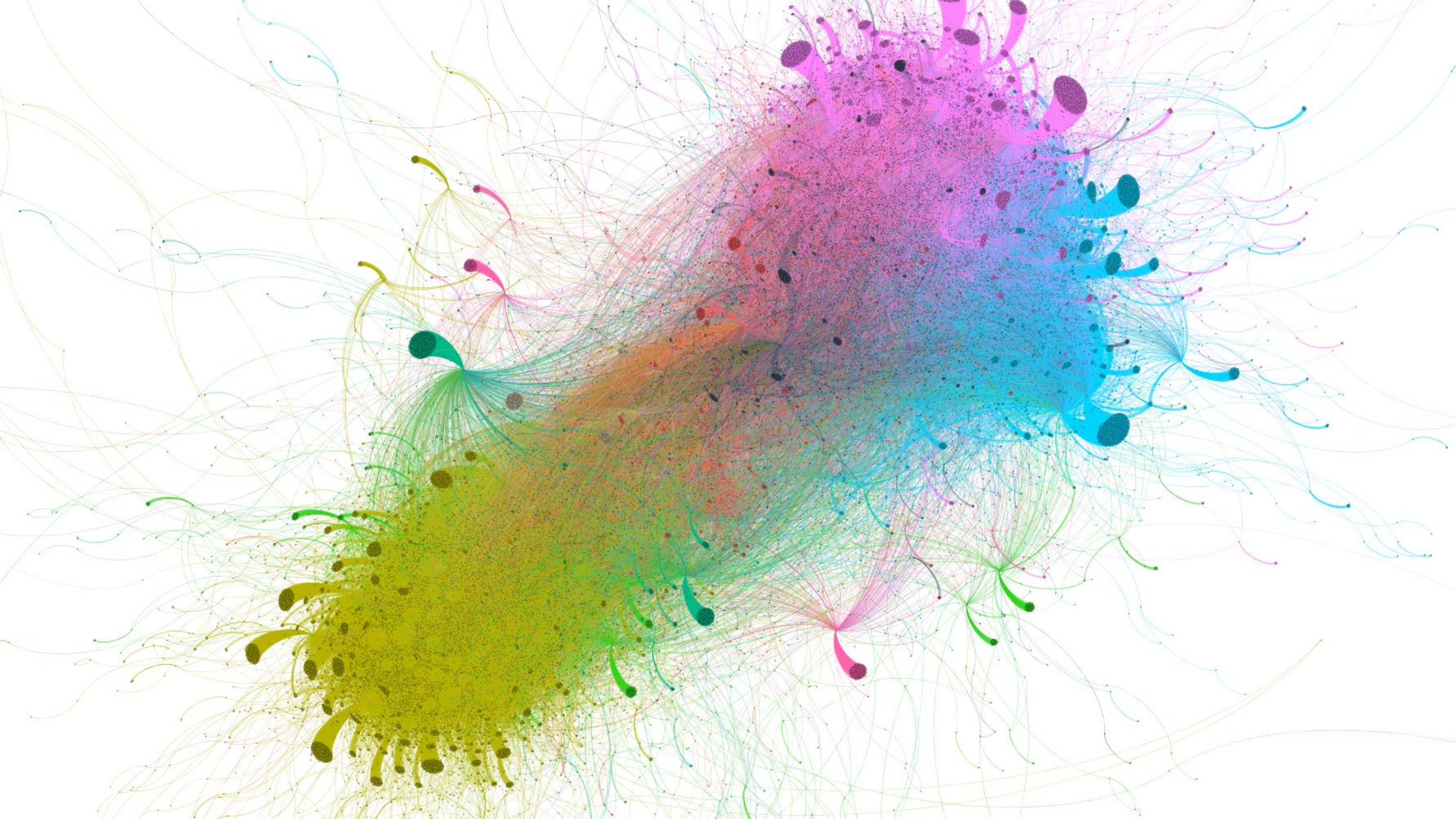
2. LA DESINTOXICACIÓN

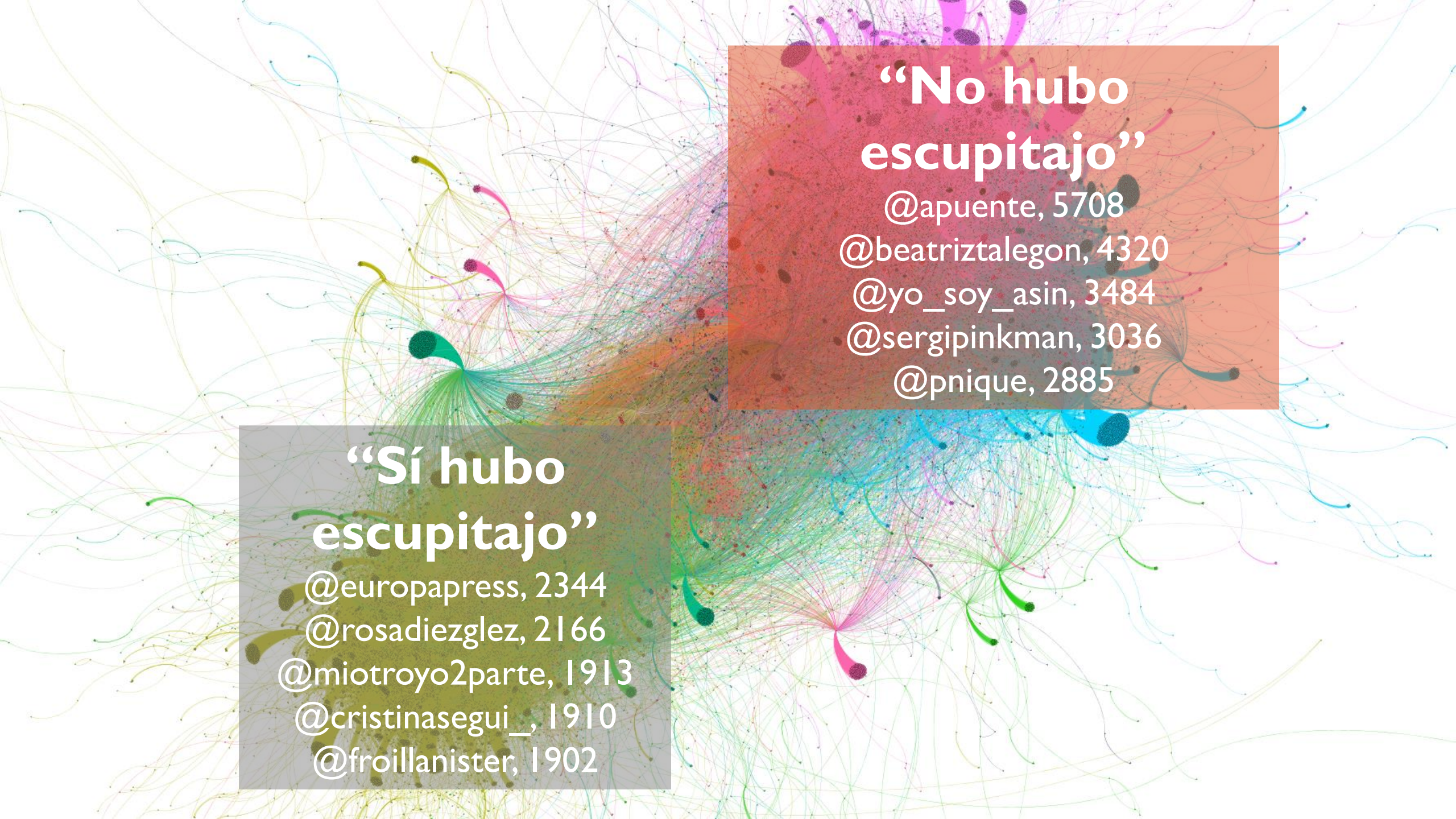
I was shocked and disturbed today to discover that I had been misquoted in a Spanish newspaper that claimed I believe Martin Luther King, Jr. would have opposed the Catalan independence movement. This distorts what I actually said in response to a Spanish reporter's questions about whether proponents 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”
@apuente, 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”

Europa Press @europapress · 21 nov.
La dirección del PSOE no vio el escupitajo de ERC a Borrell y achaca al PP la crispación bit.ly/2Aa6iVP

555 724 561

Chino de China
@unchinodechina
Seguir

En respuesta a @europapress

No ven un golpe de estado, van a ver un escupitajo...

5:19 - 21 nov. 2018

1.668 Retweets 3.774 Me gusta

151 1,7K 3,8K

Rosa Díez
@rosadiezgiez
Seguir

La Ministra Delgado destituye al abogado del Estado que defendido la rebelión en la causa por el 1-O. Esto si que es un escupitajo a la independencia profesional y al propio sistema democrático. Esto también es corrupción, y de las mas graves.

El Gobierno destituye al abogado del Estado que defendió la rebelión en el 'p...
El Ministerio de Justicia que encabeza Dolores Delgado ha destituido al abogado del Estado, Edmundo Bal, después de que se opusiese a firmar el escrito de acusación
elindependiente.com

10:51 - 21 nov. 2018

3.012 Retweets 4.122 Me gusta

Cristina Seguí
@CristinaSegui_
Segueix

Sí Sánchez no estuviera dispuesto a vender a su madre por seguir en Moncloa, convocaría elecciones tras recibir, uno de sus ministros, un escupitajo de SU SOCIO.

Tradueix el tuit

10:43 - 21 de nov. de 2018 des de València, Espanya

1.656 retuits 3.242 agradaments

120 1,7m 3,2m

“No hubo escupitajo”

Bea Talegón
@BeatrizTalegon
Seguir

Tan cierto es el escupitajo de #Borrell como la violencia de los #CDR, como el Golpe de Estado en Cataluña, como la Malversación de la Generalitat para el 1 de octubre, como la Rebelión, como la Sedición, como TODAS las PUÑETERAS MENTIRAS contra los soberanistas catalanes.

9:14 - 21 nov. 2018

3.997 Retweets 8.908 Me gusta

729 4,0K 8,9K

Arturo Puente
@apuente
Seguir

Borrell puede mentir sobre el escupitajo porque no pasó nada cuando mintió sobre algo mucho más importante: el uso de información privilegiada de Abengoa, que negó como ministro en sede parlamentaria y acto seguido reconoció no recurriendo la sanción.

3:52 - 21 nov. 2018

6.580 Retweets 10.037 Me gusta

150 6,6K 10K

Pablo Echenique
@pnique
Seguir

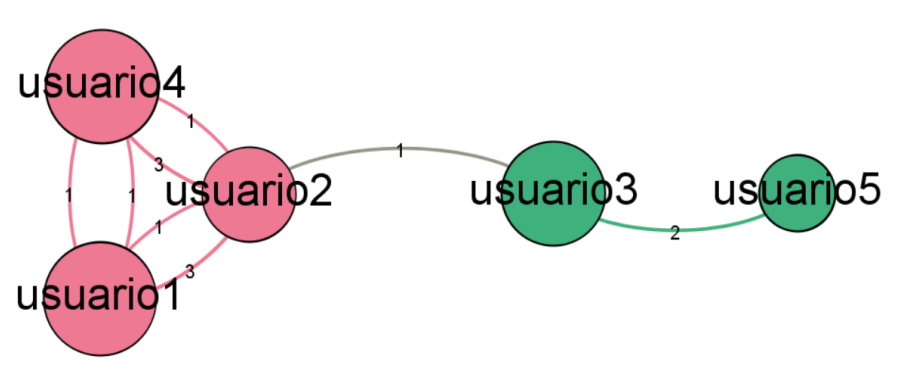
A lo mejor, entre los gritos de "fascistas" y "golpistas" y el escupitajo fantasma a Borrell, se les olvida contarte que @Pablo_Iglesias_ ha propuesto a @sanchezcastejon que los 22.000 millones que nos costó Bankia sirvan para tener una banca pública como en Alemania u Holanda.

2:24 - 21 nov. 2018

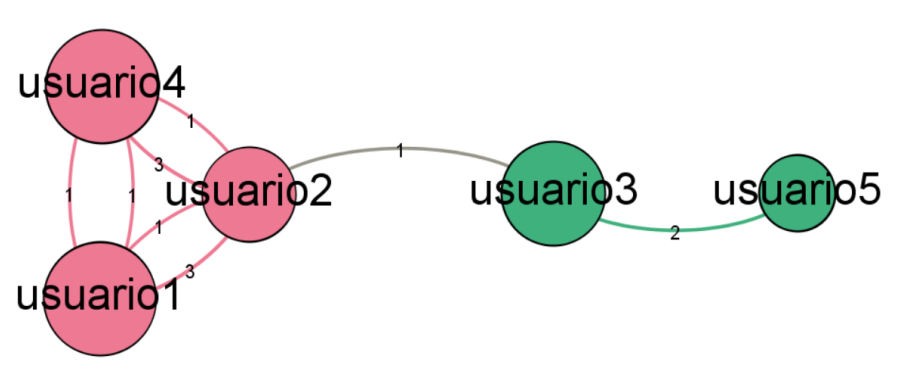
3.135 Retweets 5.069 Me gusta

902 3,1K 5,1K

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	.@usuario 1 deberíamos ir a surfear juntos un día de estos, a ver si se apunta @usuario2
usuario2	@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!!!

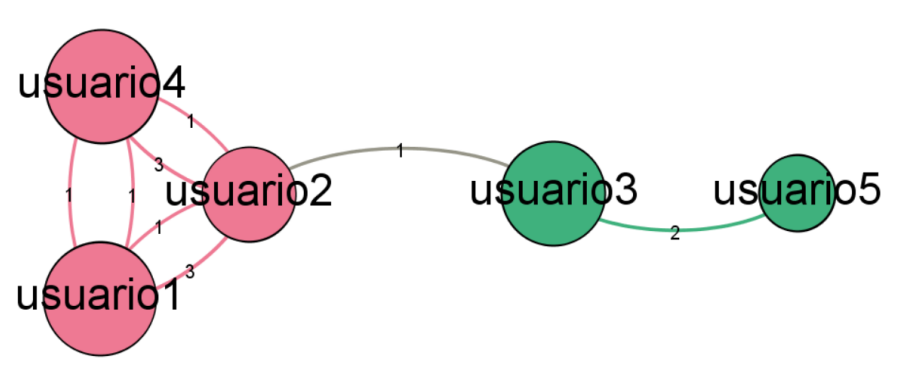


USUARIO	TUIT
usuario1	Me encanta el surf!
usuario1	En serio, me encanta el surf!!!!!
usuario1	#Surflovers
usuario2	RT: @usuario1 Me encanta el surf!
usuario2	Oye @usuario4 dice @usuario1 que le gusta el surf, como a tí!
usuario3	Surfo de anginas y no he podido ir a trabajar
usuario1	RT: @usuario2 Oye @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	.@usuario 1 deberíamos ir a surfear juntos un día de estos, a ver si se apunta @usuario2
usuario2	@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!!!



USUARIO	GRADO DE ENTRADA	GRADO DE SALIDA	MODULARIDAD	BETWEENNESS
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

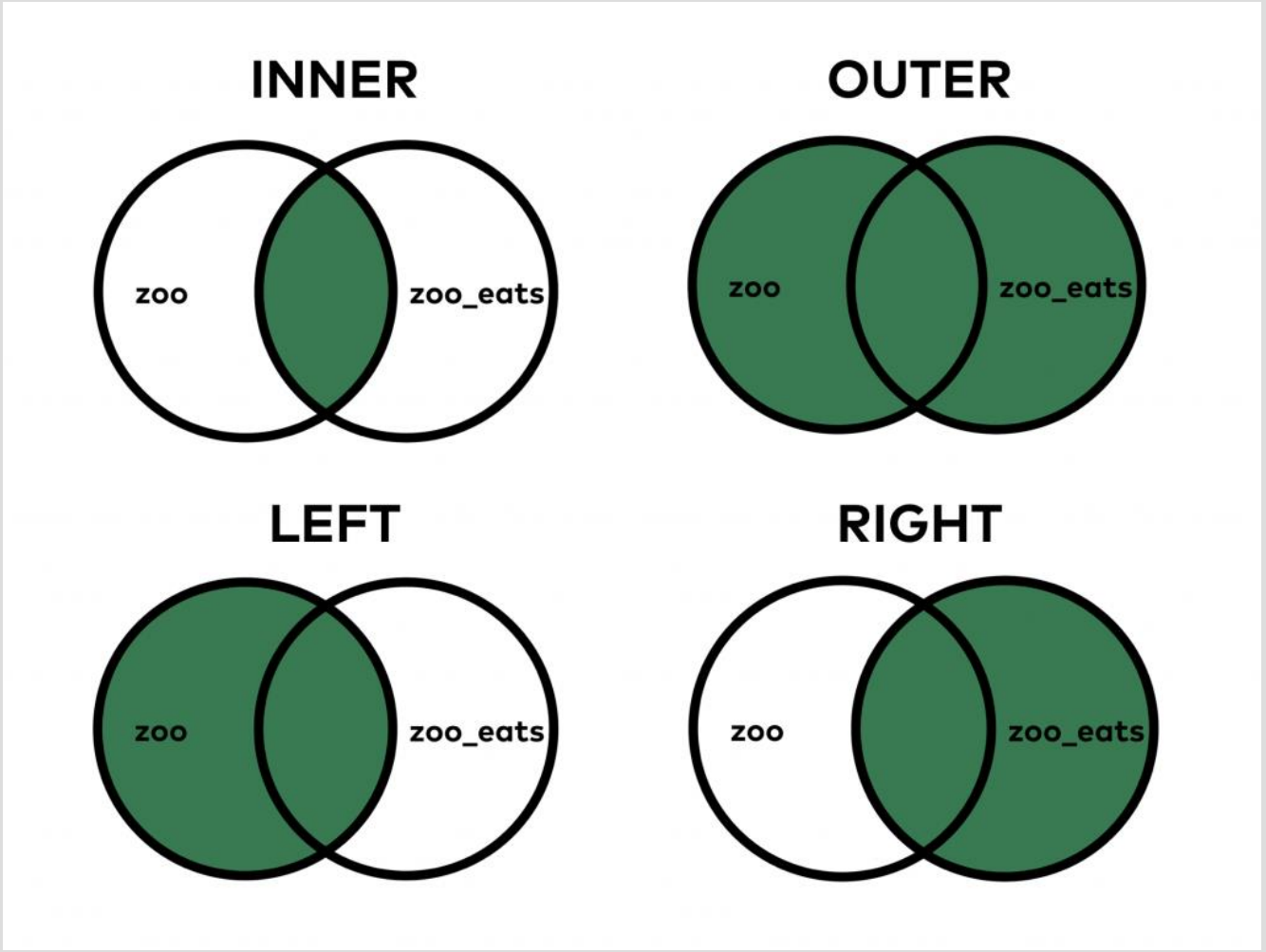
USUARIO	TUIT
usuario1	Me encanta el surf!
usuario1	En serio, me encanta el surf!!!!!
usuario1	#Surflovers
usuario2	RT: @usuario1 Me encanta el surf!
usuario2	Oye @usuario4 dice @usuario1 que le gusta el surf, como a tí!
usuario3	Surfo de anginas y no he podido ir a trabajar
usuario1	RT: @usuario2 Oye @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	.@usuario 1 deberíamos ir a surfear juntos un día de estos, a ver si se apunta @usuario2
usuario2	@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!!!



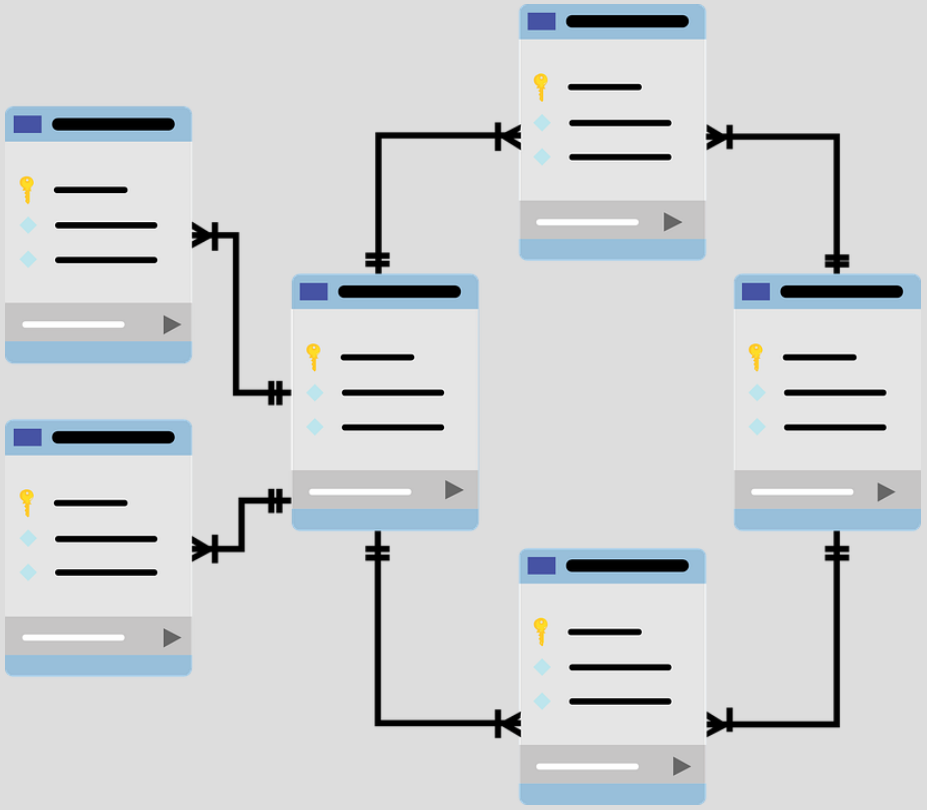
USUARIO	GRADO DE ENTRADA	GRADO DE SALIDA	MODULARIDAD	BETWEENNESS
usuario1	4	2	0	0
usuario2	2	7	0	0,16
usuario3	3	0	1	0
usuario4	4	2	0	0
usuario5	0	2	1	0

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	.@usuario 1 deberíamos ir a surfear juntos un día de estos, a ver si se apunta @usuario2
usuario2	@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	BETWEENNESS
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
zoo
zoo_eats
OUTER
zoo
zoo_eats
LEFT
zoo
zoo_eats
RIGHT
zoo
zoo_eats

BASES DE DATOS RELACIONALES



Microsoft SQL Server

ORACLE

MySQL

SQLite

PostgreSQL


Grafana

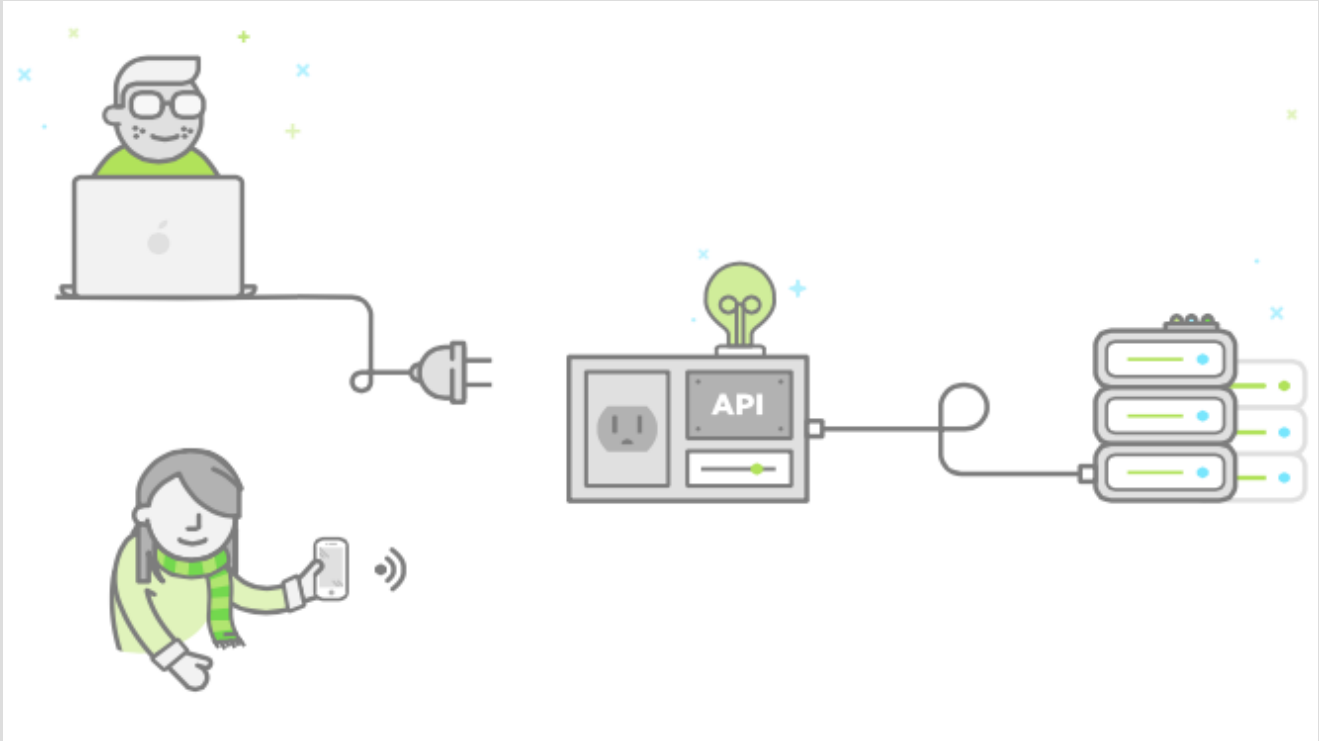

Google
Data Studio

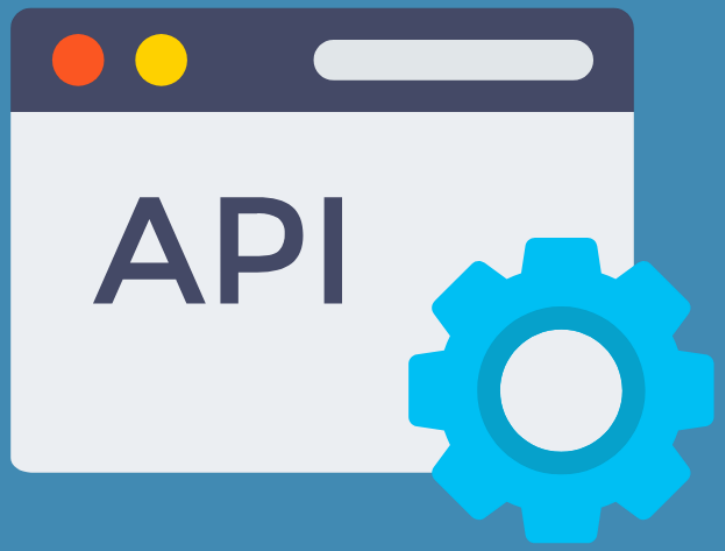

Power BI

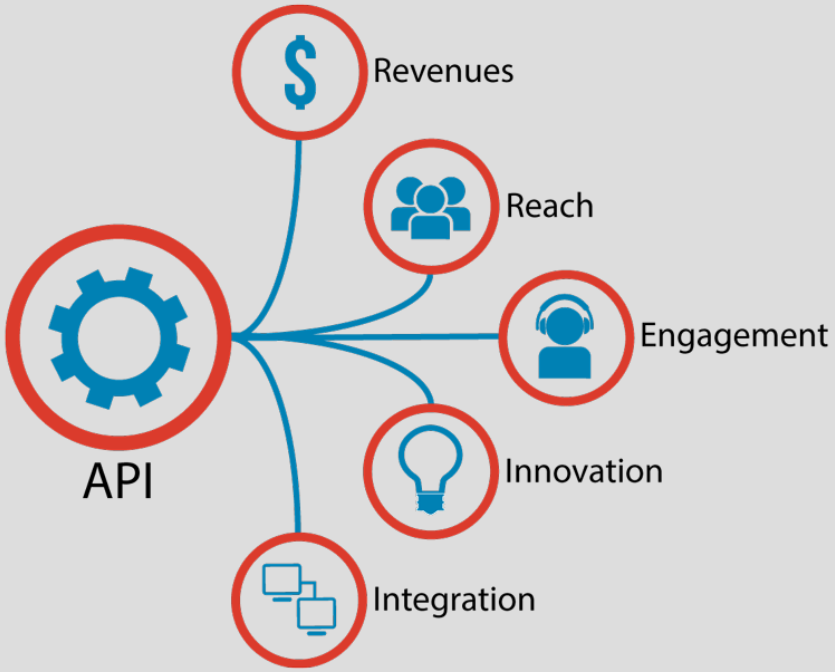

+ableau

ADQUISICIÓN DE LOS DATOS: APIS OFICIALES VS. WEB SCRAPPING O RASPADO WEB

EL NUEVO DATASCAPE

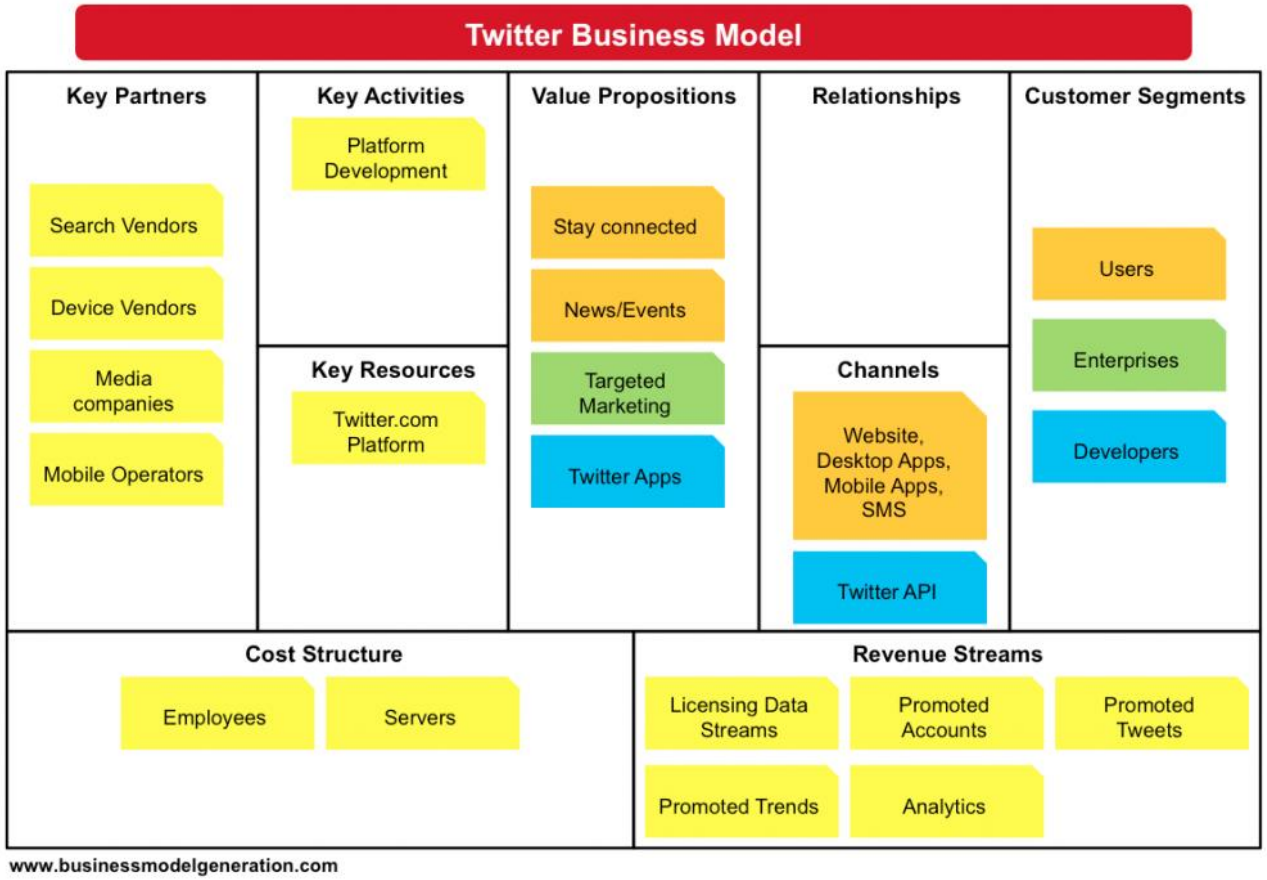



API


Revenues
Reach
Engagement
API
Innovation
Integration

LAS APIS

¿POR QUÉ TWITTER TIENE API?



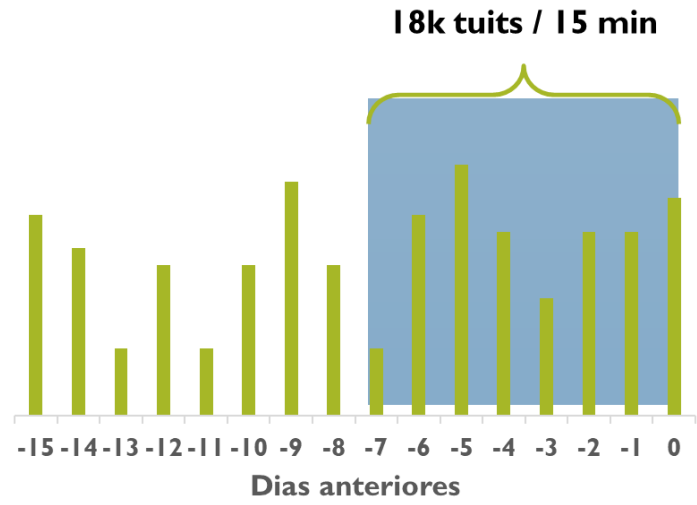
LAS APIS DE TWITTER

Feature summary

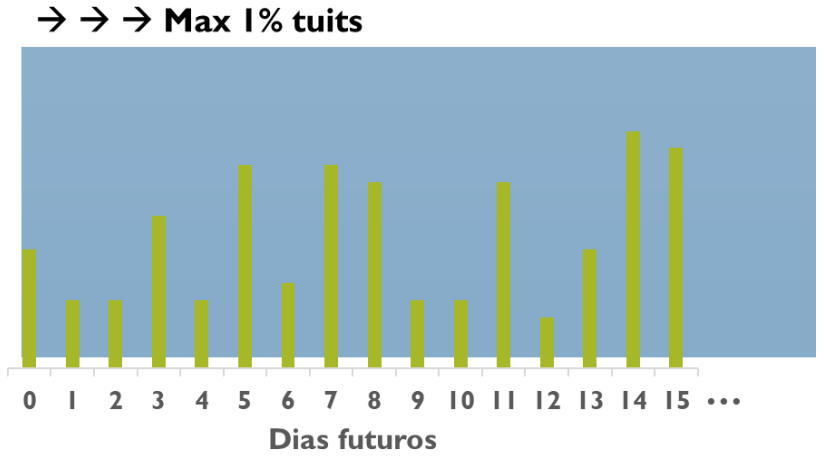
Category	Product name	Supported history
Standard	Standard Search API	7 days
Premium	Search Tweets: 30-day endpoint	30 days
Premium	Search Tweets: Full-archive endpoint	Tweets from as early as 2006
Enterprise	30-day Search API	30 days
Enterprise	Full-archive Search API	Tweets from as early as 2006

CONSULTAS EN TWITTER

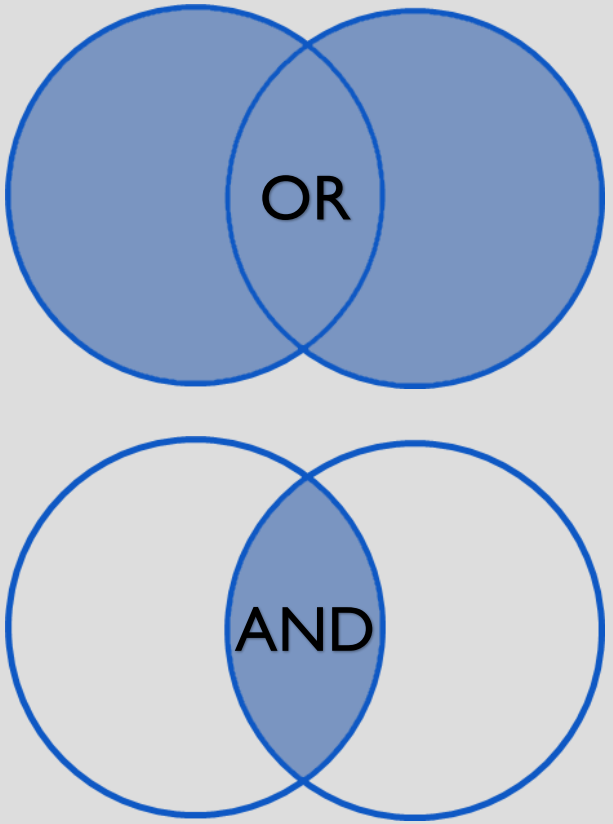
Consultas retroactivas

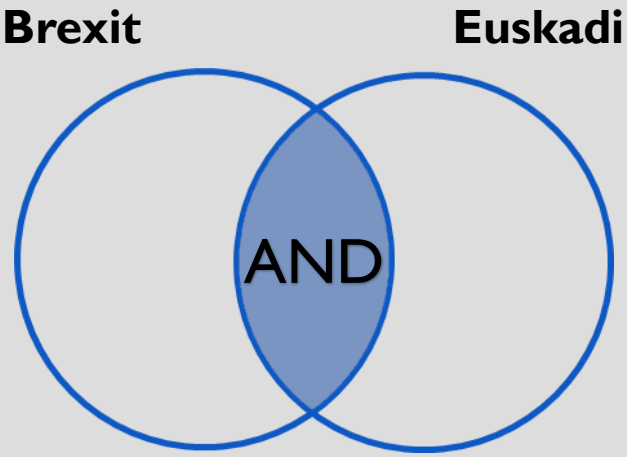


Consultas en tiempo real



El lenguaje de la **API**



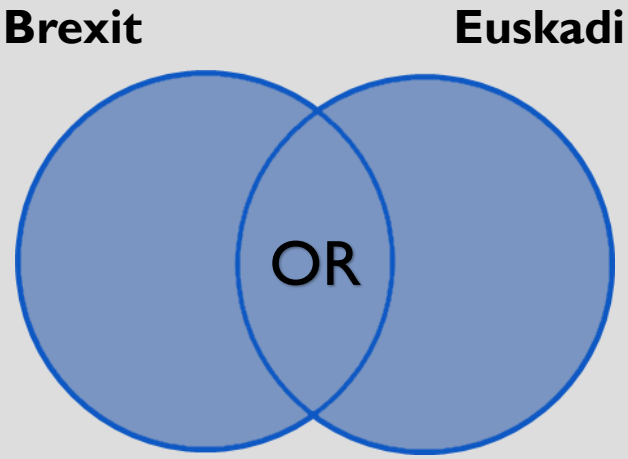


"Me inquieta como afectará el **Brexit** en **Euskadi**"

"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**"

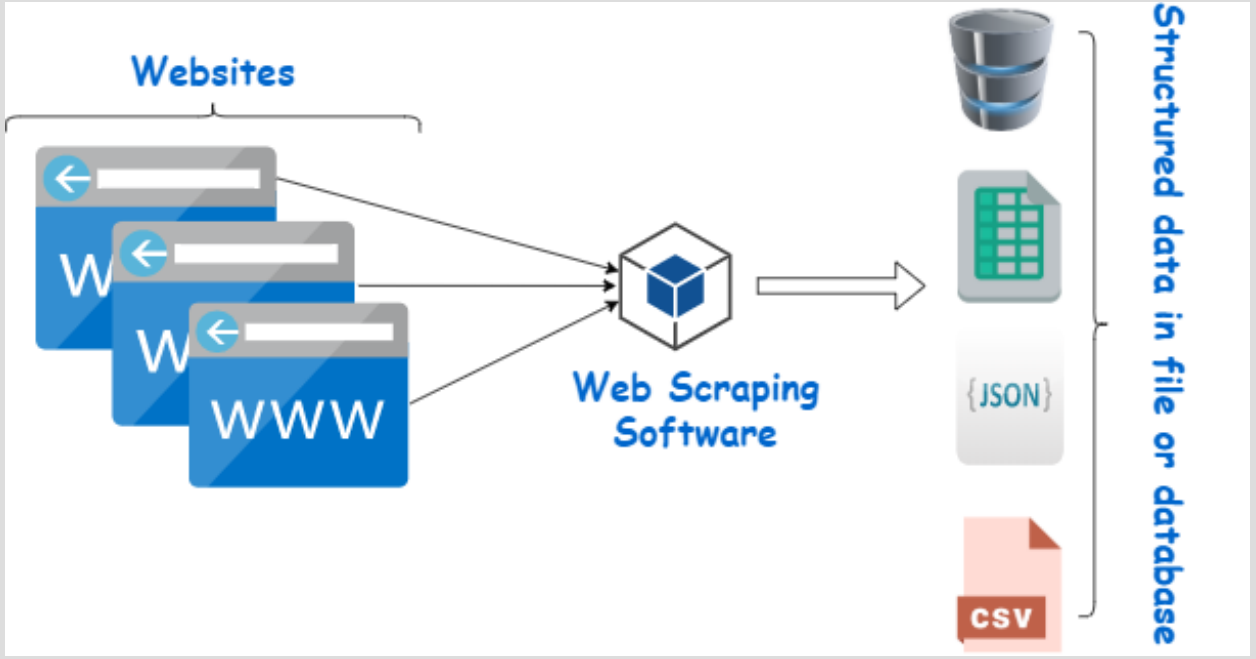
"A mi el **Brexit** no me asusta"

"Lloverá en **Euskadi**, da igual cuando leas esto"

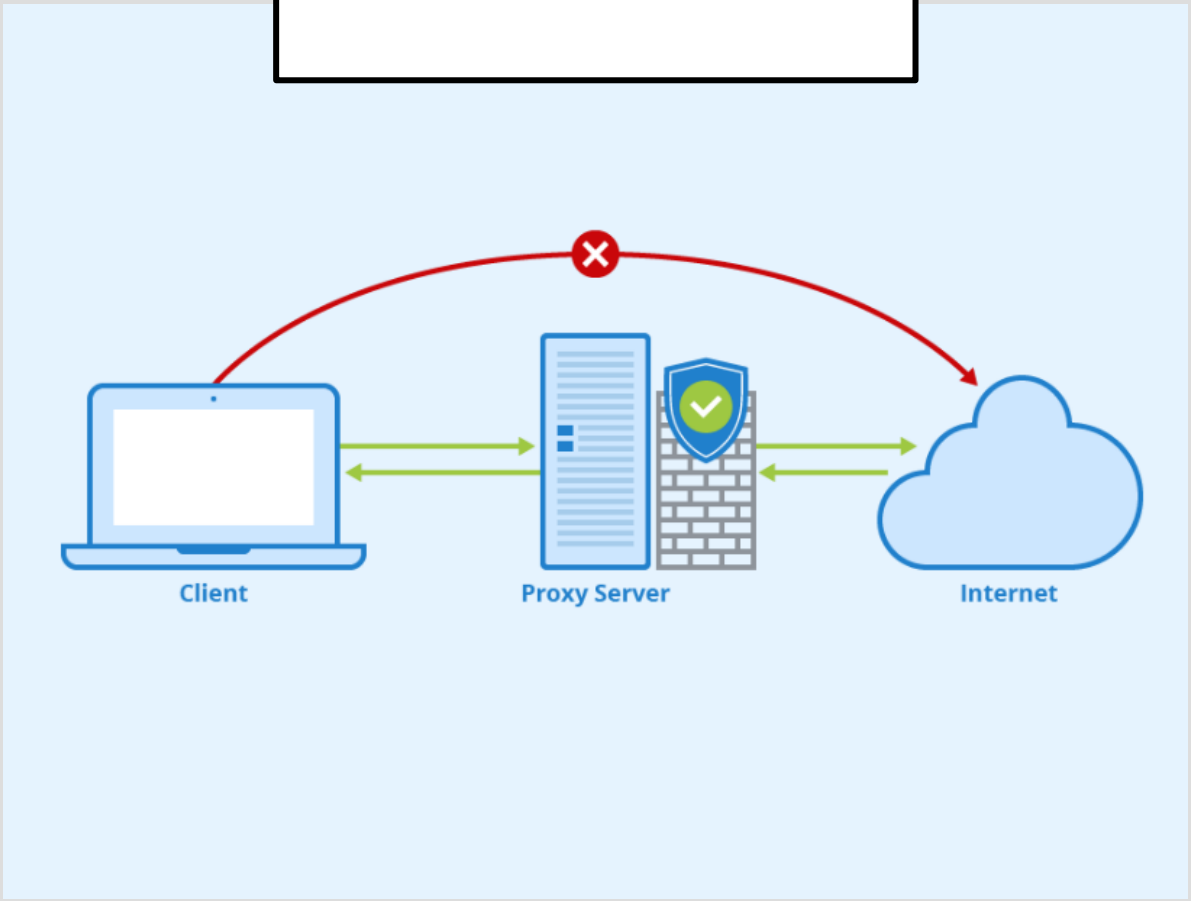
Plataforma	Endpoints gratuitos	Endpoints de pago	Dificultad de uso	Conocimientos necesarios
Twitter	Histórico 7 días Tiempo real Academic track	Histórico completo	Media	ETL / Gestión big data *Servidores *Programación
Facebook	Información páginas Histórico páginas		Alta *Software intermedio	ETL / Gestión big data *Servidores
Instagram	Datos propios		Alta	ETL / Gestión big data
Youtube	Contenidos Comentarios Datos de usuarios		Baja	Manejo de Excel

LAS APIS DE LOS SOCIAL MEDIA

RASPADO WEB


Websites
WWW
Web Scraping Software
{JSON}
csv
Structured data in file or database

PROXYS Y CRAWLERS



Luminati

proxycrawl

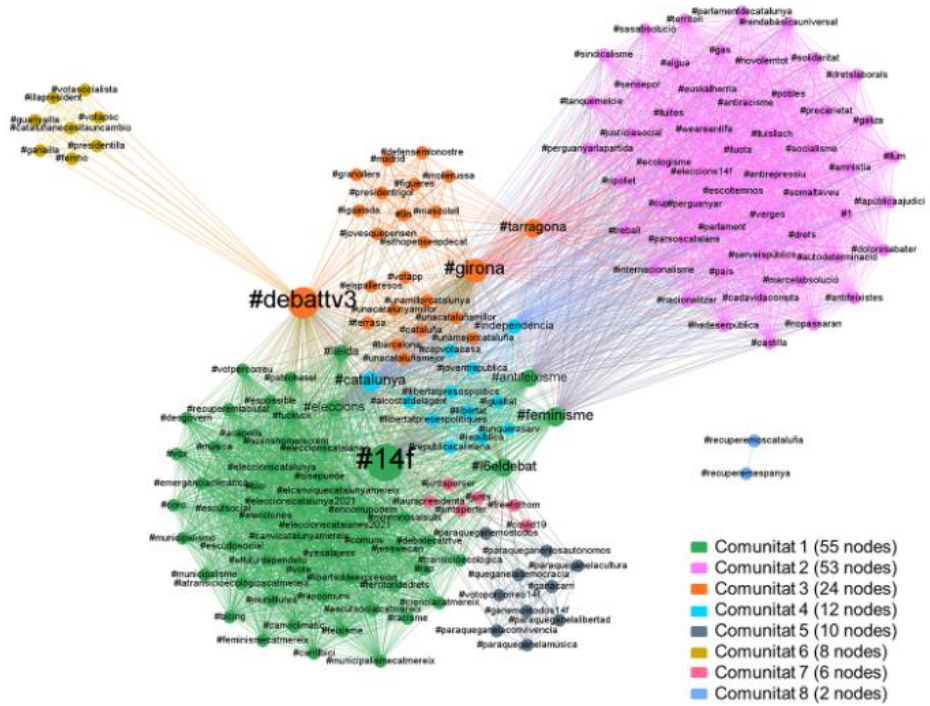
PHANTOM BUSTER //

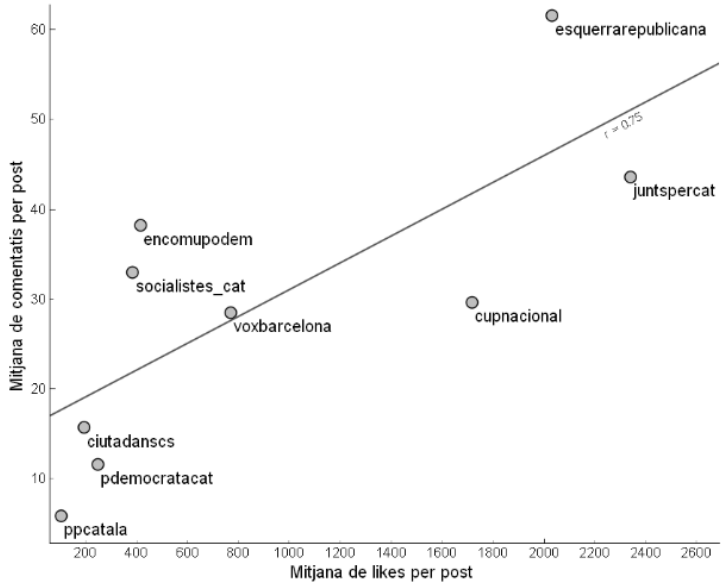
ISSN 2013-9004 (digital); ISSN 0210-2862 (paper)
https://doi.org/10.5565/rev/papers.2967

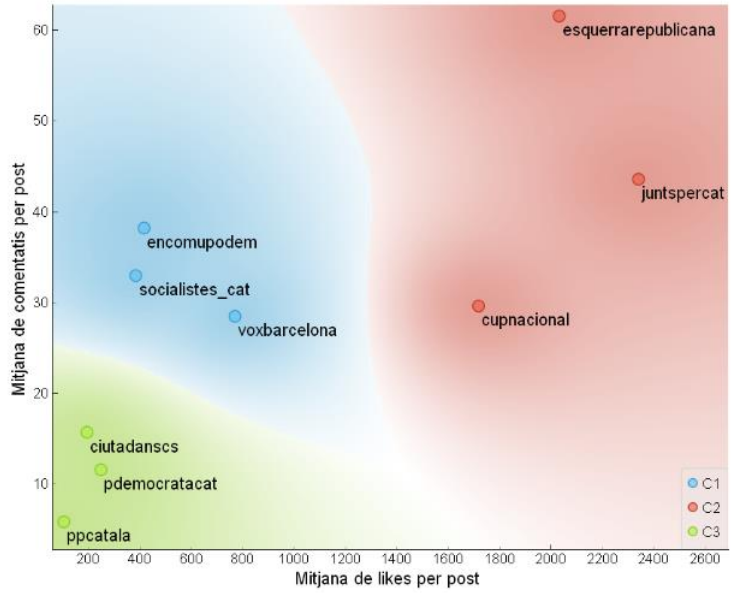
Papers 2022, 107/1
147-174

El 14F a Instagram. Una proposta d'articulació de tècniques de raspat web i anàlisi de xarxes

Jordi Morales-i-Gras
Oriol Sánchez-i-Vallès
Network Outsight
morales.jordi@gmail.com; sanchezv.oriol@gmail.com







EJEMPLO DE DATOS DE WEBSCRAPPING


```
[
  {
    "date": "Thursday, November 12, 2020 at 10:23 AM",
    "post_text": "Billie Eilish - “Therefore I Am”  Out
now. https://smarturl.it/ThereforeIAm Watch the music
video for “Therefore I Am”, directed by Billie.
https://youtu.be/RUQl6YcMalg",
    "likes": 305000,
    "comments": 7800,
    "top_comments": [
      {
        "text": "If you could read this message and
support me, Billie Eilishit will honestly make my day. I
am a guitarist and pianist, songwriter and I know people
write this all the time but I believe I’ll be the one
you’ll be happy to listened to. I do all my songs…See
More",
        "author_name": "WhileWild",
        "created": "Thursday, November 12, 2020 at 1:26
PM"
      }
    ],
    "shares": 34000
  }
]
```

APIS OFICIALES VS. WEBSCRAPING

APIS OFICIALES

+ Robustez, rapidez y gratuidad hasta cierto punto
+ Abundancia de metadatos
+ Costes elevados en las opciones de pago

- Limitaciones abundantes en las opciones gratuitas: temporales y de volumen
- Muestreos automáticos y mal documentados
- Encriptaciones innecesarias

WEBSRAPING

+ Acceso a datos históricos y sin limitaciones de volumen
+ Acceso a los datos públicos en su totalidad, sin muestreos ni encriptaciones
+ Intermediarios muy económicos

- Falta de estándares y referencias en la literatura científica
- Procesos inestables y a menudo bloqueados
- Metadatos escasos

NOTAS FINALES

I si l'estadística "clàssica" no és la perspectiva quantitativa que demanden les societats del segle XXI?

¿LA CIENCIA SOCIAL COMPUTACIONAL PODRÁ SUPERAR LA BRECHA ENTRE LA INVESTIGACIÓN CUALITATIVA Y CUANTITATIVA?