

# Jornada Data Science: Big Data y Machine Learning como herramienta de Futuro

---

# Sobre Mí (Manuel Garrido)

- Data Scientist Freelance
- UPV: Ingenieria Industrial
- IE: Master In Management
- Portugal → España → EEUU
- Consultor → Analista → Data Scientist

[manugarri.com/centic](http://manugarri.com/centic)

[hola@manugarri.com](mailto:hola@manugarri.com)

```
[~]$ whoami
```

# Data Science

# Por qué lo llaman Big Data cuando quieren decir Data Science?



TENDENCIAS: LOS NIÑOS GANAN PESO EN LAS DECISIONES DE COMPRA FAMILIAR

## Emprendedores

www.emprededores.es

# BIG DATA

PARA TU NEGOCIO

Lanzar productos, mejorar la gestión de clientes, realizar estudios de mercadoiables... y todas las **OPORTUNIDADES** de negocio del nuevo boom empresarial.

Promociones audaces que si funcionan

¿Por qué cierran más de los comercios? (Pág. 14)

ALQUILER POR HORAS: POR QUÉ TRIUNFA ESTE NUEVO DISTRIBUCIÓN COMERCIAL (Pág. 14)

Luis Martín Cabezas  
BUSINESS ANGEL

Nunca te premio a la altura de los emprendedores americanos. Son tipos diferentes.

Sólo el 14 por ciento de las empresas tiene una estrategia de Big Data. Pág. 18

53 mil 200 millones de dólares serán los ingresos del mercado de Big Data en 2016. Pág. 25

37 mil 427 fue el número de piezas publicadas en la edición especial de la revista de negocios Forbes. Pág. 29

Más de mil 330 millones de dólares son los ingresos que generó el Big Data en 2014. Pág. 43

mercadotecnia publicidad medios  
**Merca2.0**

Año 14 No.142, agosto 2015 \$65.00  
índices en mercadotecnia, publicidad y medios  
www.merca2.0.com

MERCADOTECNIA

## BIG DATA

¿EL FUTURO DEL MARKETING?

En la actualidad acceder a grandes cantidades de datos no es suficiente para competir. Es necesario analizar la data a favor de la marca. Pág. 25

PUBLICIDAD AGENDA DE LATAM (Pág. 43)

ESTUDIO ANUAL DE RADIO 2015 (Pág. 31)

PUBLICIDAD EN EL PRESENTE Y FUTURO DE LA PUBLICIDAD (Pág. 35)

INNOVACIÓN: ¿CON O SIN PROMOCIÓN? (Pág. 41)

MARKETING DIGITAL, IMPRESIÓN 3D Y MARKETING (Pág. 43)

... y Big Data llegó para quedarse

El uso de Big Data está creciendo en todo el mundo, incluyendo, cómo no, a las empresas españolas. Todos los estudios corroboran que es una tendencia creciente desde hace años y cuya proyección de futuro también muestra un comportamiento alcista.

# Entonces, que es Data Science?



Josh Wills  
@josh\_wills

Follow

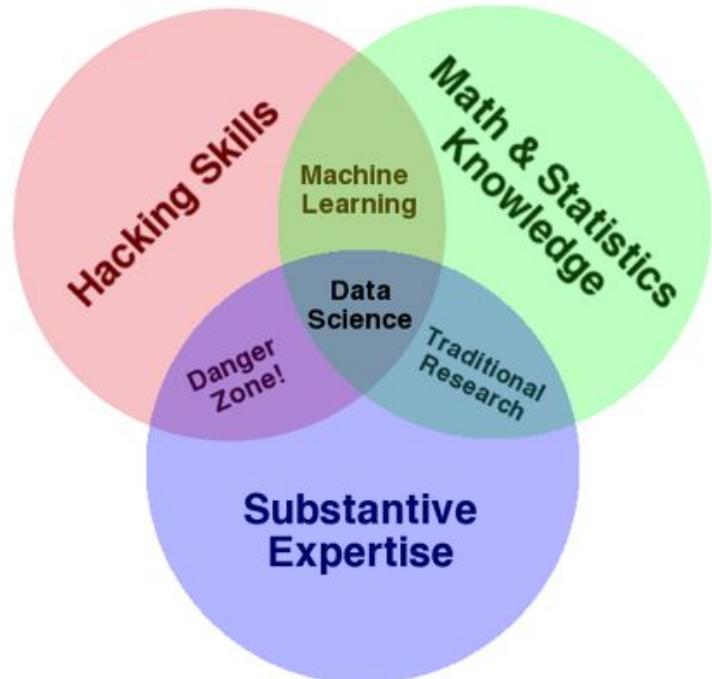
Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

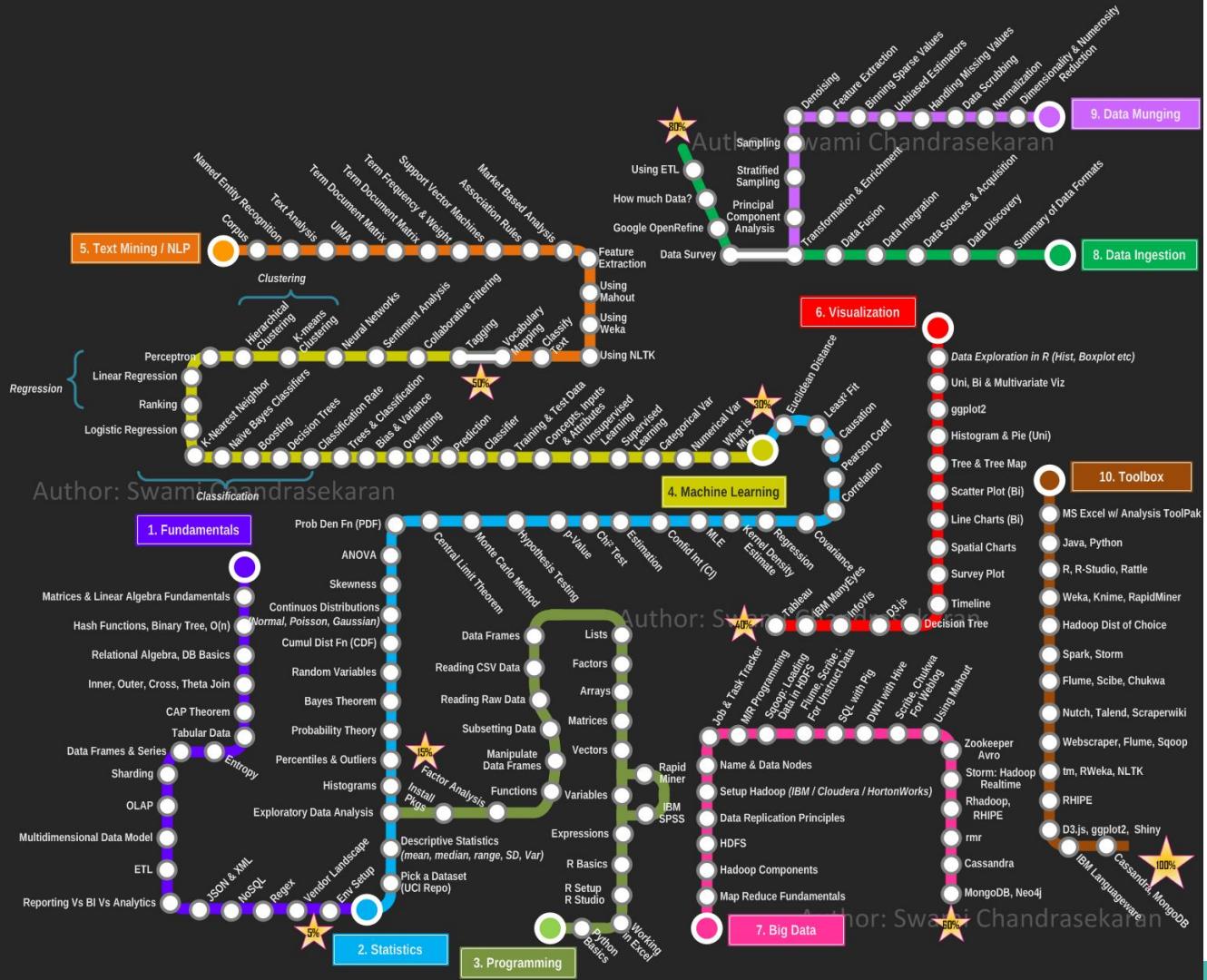
RETWEETS  
1,581 LIKES  
1,216



9:55 AM - 3 May 2012

52 1.6K 1.2K





# Data Science en el marco Europeo - Normativa

**Normativa Europea:** General Data Protection Regulation (GDPR) 2018



- Privacidad por defecto
- Derecho al olvido
- Notificación de pérdida de datos
- Portabilidad de datos
- Dificultad de sacar datos personales fuera de la UE

# Data Science en el marco Europeo - Open Data

**Impulso de los datos abiertos:** - European Open Data Portal



# Data Science en el marco Europeo - Conferencias

EuroScipy (Erlangen)

Data Science Summit Europe (29 Mayo Jerusalém)

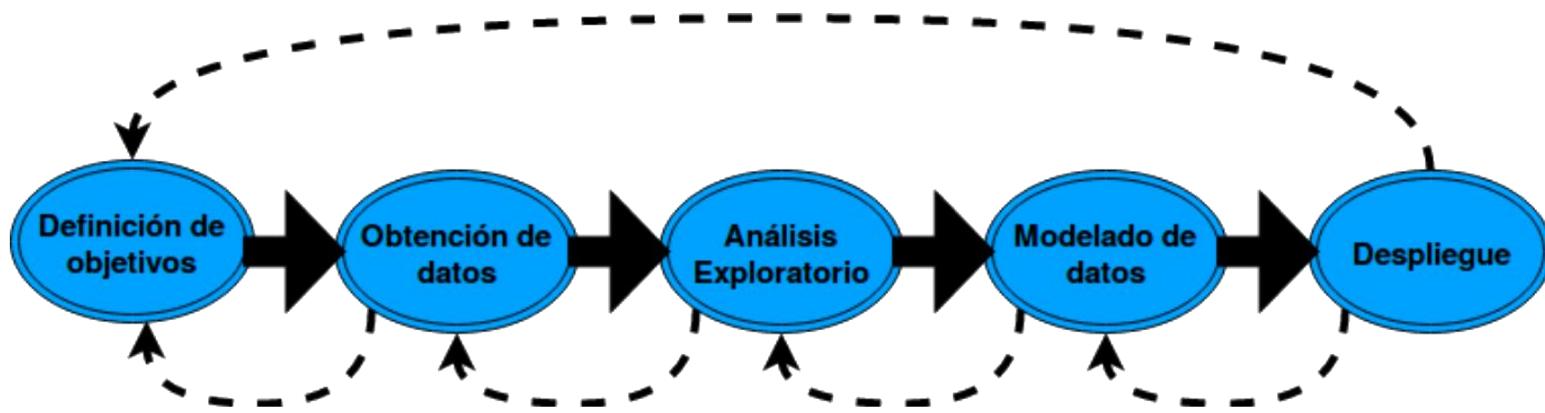
Spark Summit (24-26 Octubre Dublín)

Strata Data Conference (22-25 Mayo, Londres)

Data Conference (24-26 Julio Madrid)

Predictive Analytics World (11-12 Octubre Londres)

# El proceso de Data Science



# Como obtener datos

## Datos internos

### Repositorios de datos

- [Centro Regional de Estadística de Murcia](#)
- [Instituto Nacional de Estadística](#)
- [European Union Open Data Portal](#)
- [UNdata](#)
- [Open Data Inception](#)

### Brokers de datos

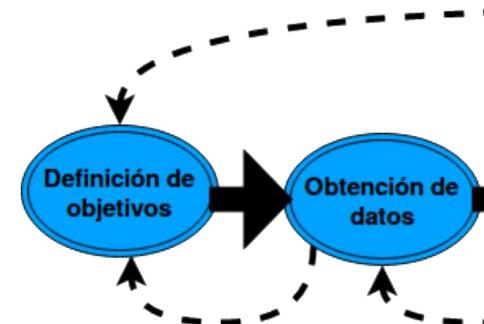
- [Experian](#)
- [Axiom](#)

### APIs

- [Yahoo finance](#)
- [OpenWeather](#)

### Web Scraping

- [Scrapy](#)
- [import.io](#)



# Tipos de datos - variables

<b>Continuas</b>	Edad, Altura, colores RGB
<b>Ordinales</b>	Ratings, Niveles educativos, Muy de acuerdo/De acuerdo/En desacuerdo
<b>Categóricas</b>	Hombre/Mujer, Apto/No Apto, días de la semana

# Tipos de datos - estructura

**Estructurados  
(<10%)**

Catálogo de biblioteca, bases de datos sql

**Semiestructurados  
(<10%)**

XML, JSON, CSV

**No Estructurados  
(>=80%)**

Emails, Fotos, PDF

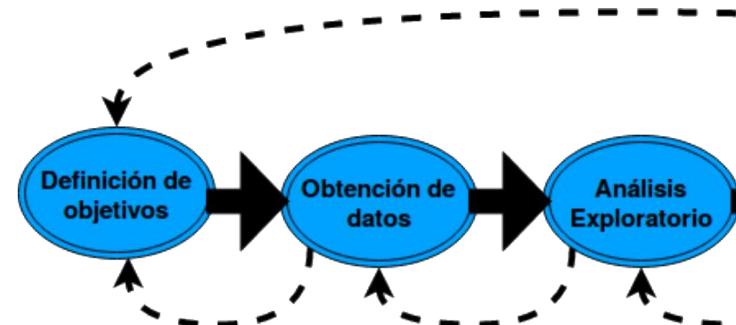
# Analytics

## 1) Métricas (y KPIs):

- a) Internas o Externas
- b) Accesibles, Accionables y Comparables

## Analytics

- c) Analizar y extraer valor de las métricas



Métricas	KPI	Analytics
# de peticiones a un servidor	Tiempo medio entre despliegue de versiones	¿Qué partes de la aplicación tarda más en responder?
Latencia media de una petición web		¿Qué canales de marketing funcionan mejor para conseguir clientes?
# de demos hechas a clientes en los últimos 30 días	Número de clientes cualificados en los últimos 90 días	¿Qué funcionalidades son las más reclamadas por nuestros potenciales clientes?
# de visitas a página web comercial		

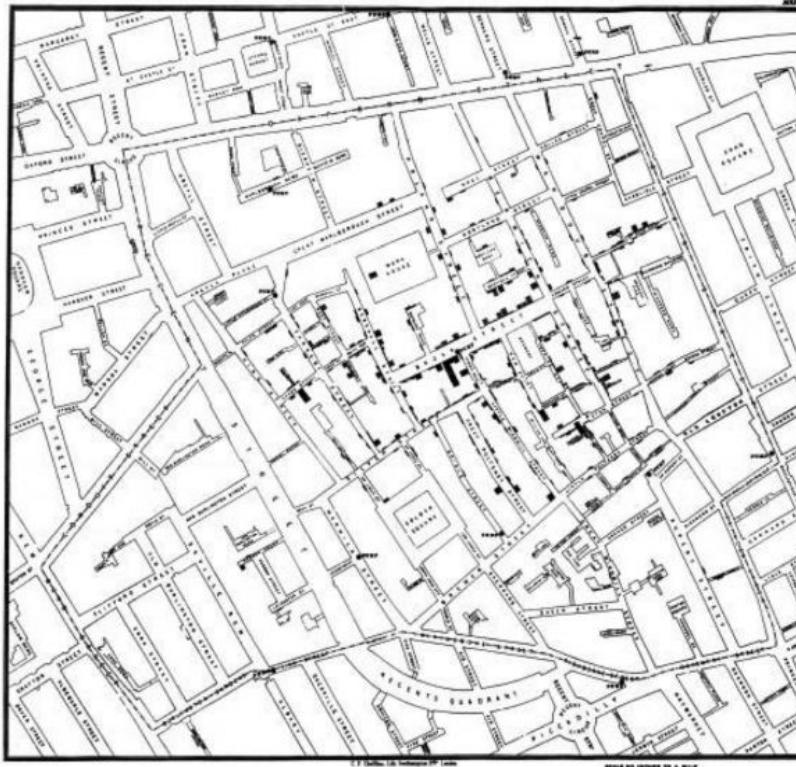
## Google Analytics

Browser	Visits	Visits
1. Firefox	843	53.02%
2. Internet Explorer	534	33.58%
3. Safari	103	6.48%
4. Chrome	54	3.40%
5. Mozilla	33	2.08%
6. Opera	16	1.01%
7. Mozilla Compatible Agent	4	0.25%
8. Camino	1	0.06%
-	-	-



# Visualización de Datos

Mapa de John Snow

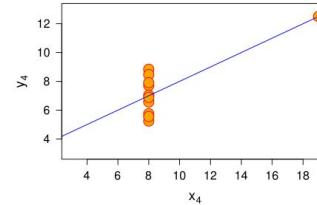
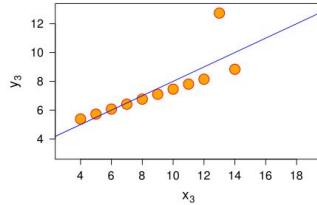
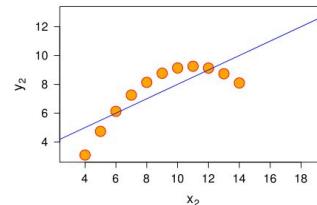
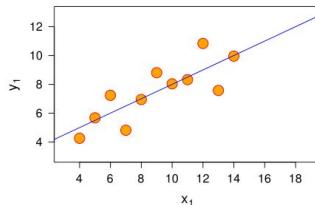


# Visualización de Datos

## Cuarteto de Anscombe

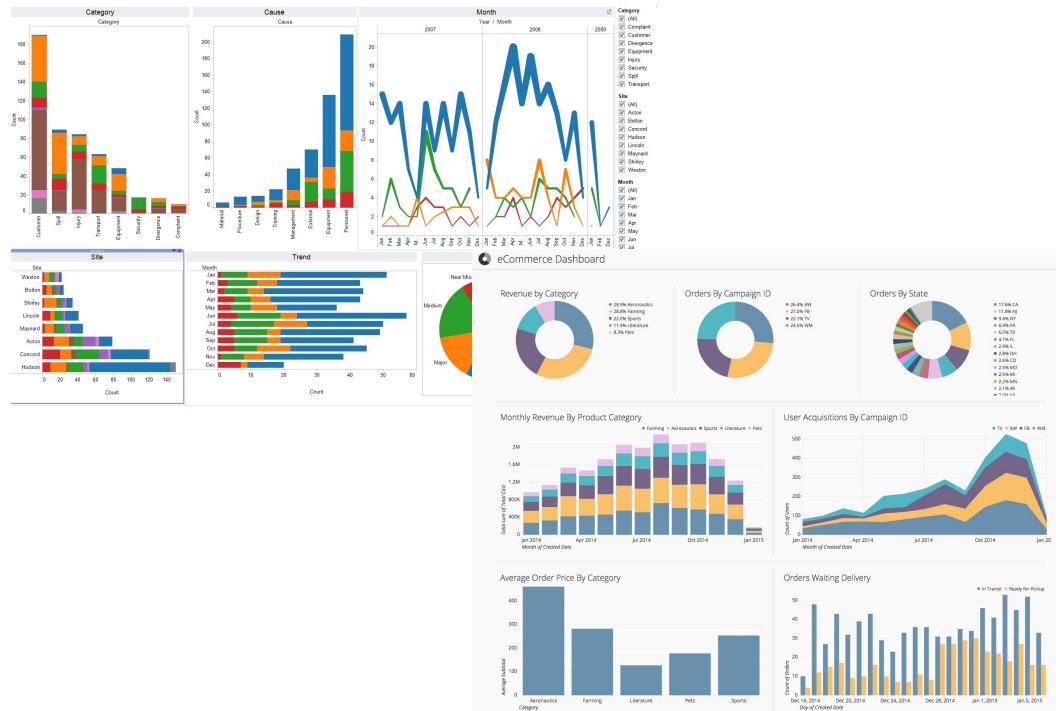
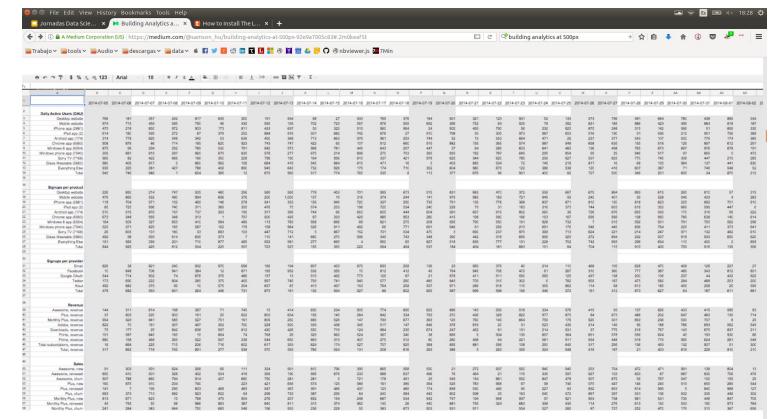
I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

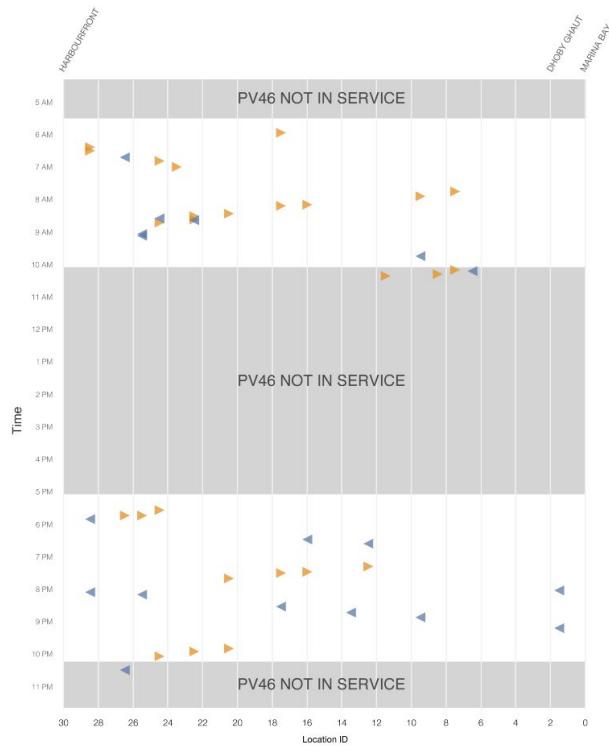
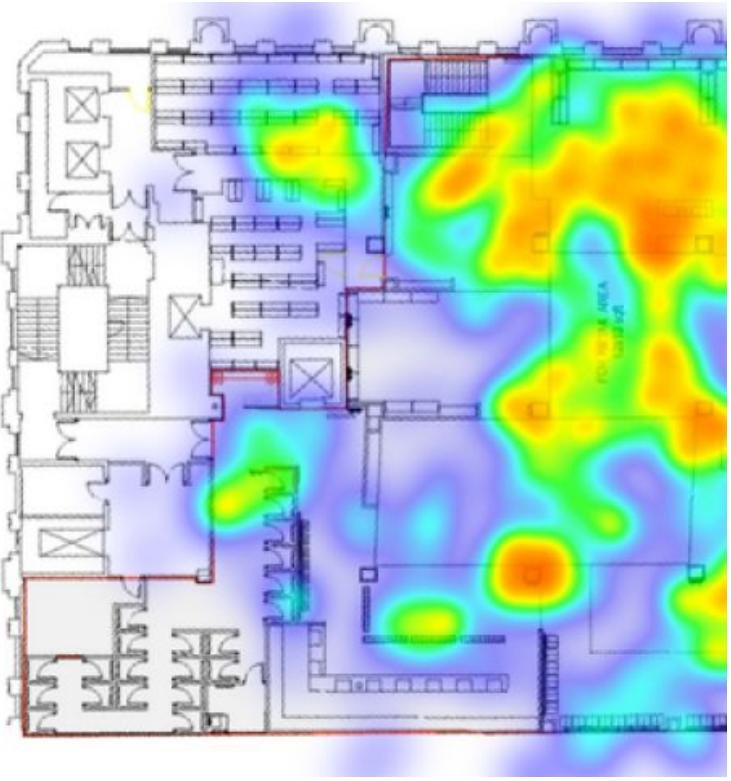
Propiedad	Valor
Media de cada una de las variables x	9.0
Varianza de cada una de las variables x	11.0
Media de cada una de las variables y	7.5
Varianza de cada una de las variables y	4.12
Correlación entre cada una de las variables x e y	0.816
Recta de regresión	$y = 3 + 0.5x$

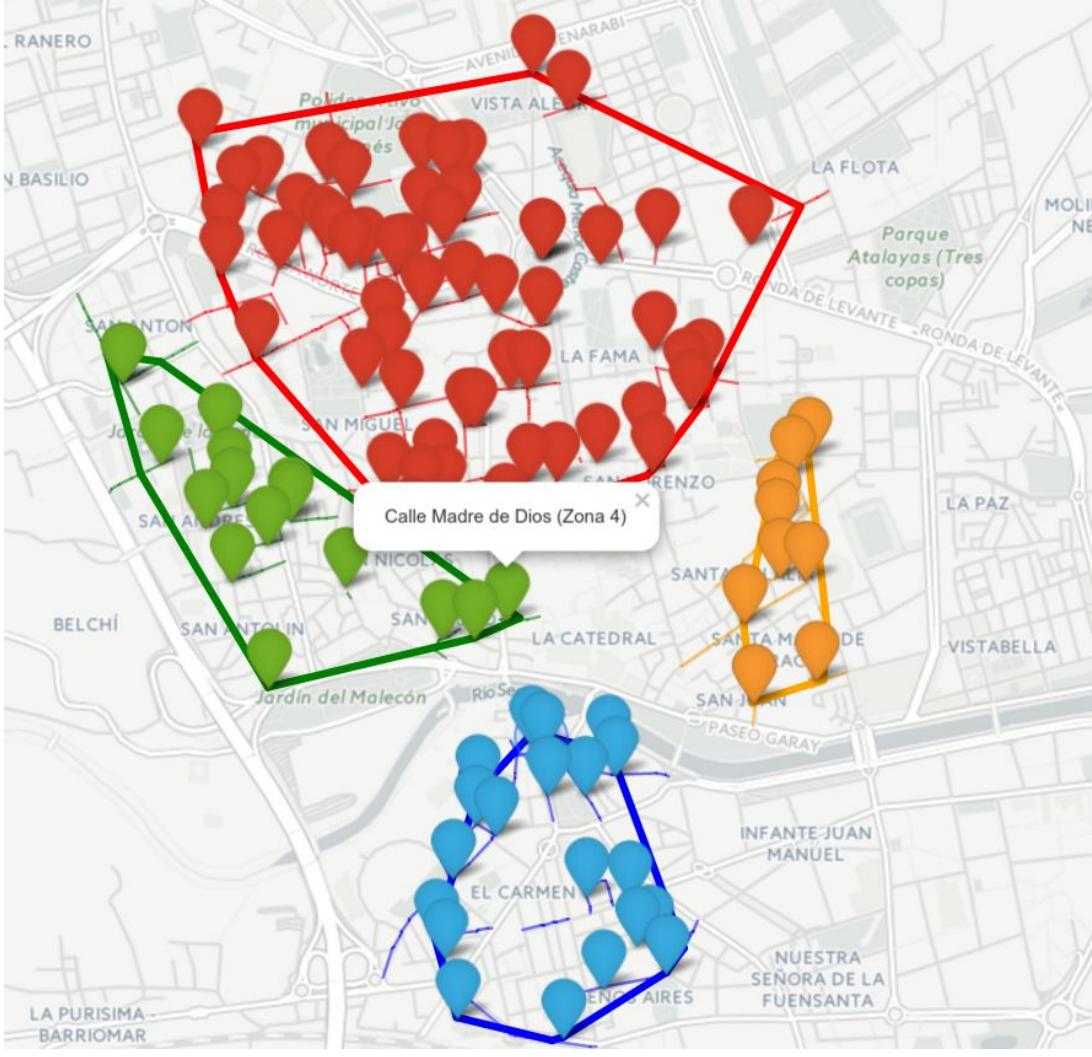


# Visualización de Datos

## Dashboard (cuadro de mando)







# Visualizacion de datos - Herramientas

## Dashboards (BI):

- [Tableau](#)
- [Chartio](#)
- [Grafana](#)

## Desarrollo web (JS):

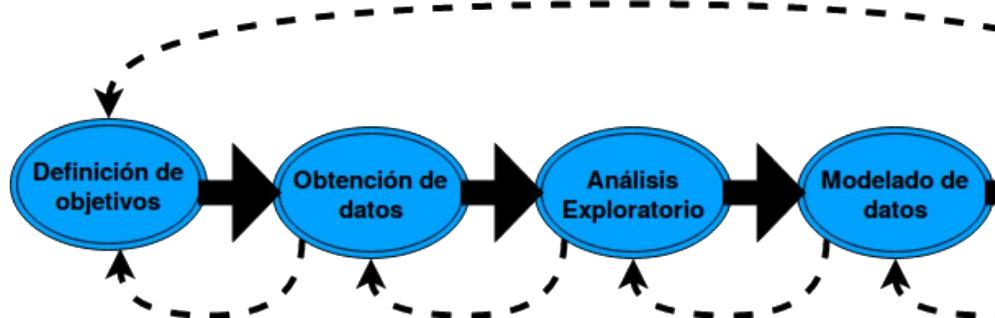
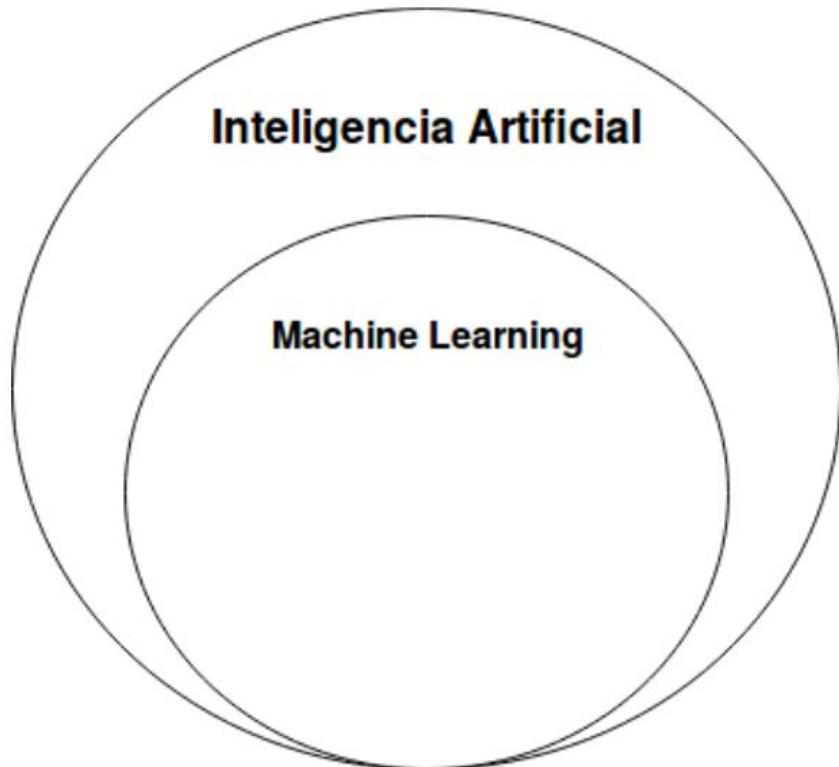
- [D3.js](#)
- [ChartJS](#)
- [plotly](#)

## Graficos científicos:

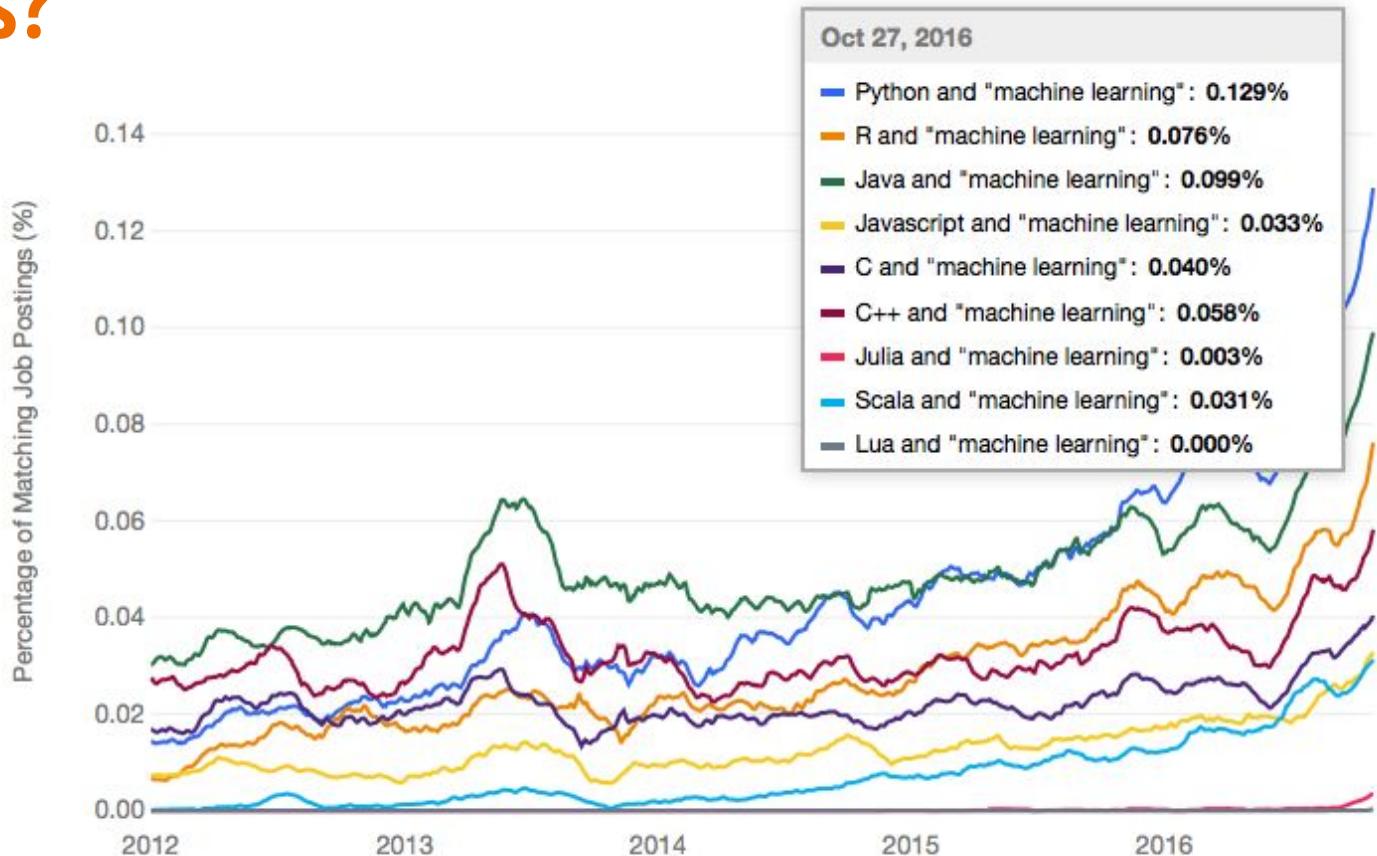
- [ggplot2](#)
- [Bokeh](#)
- [seaborn](#)

# Machine Learning

# Machine Learning



# Lenguajes?



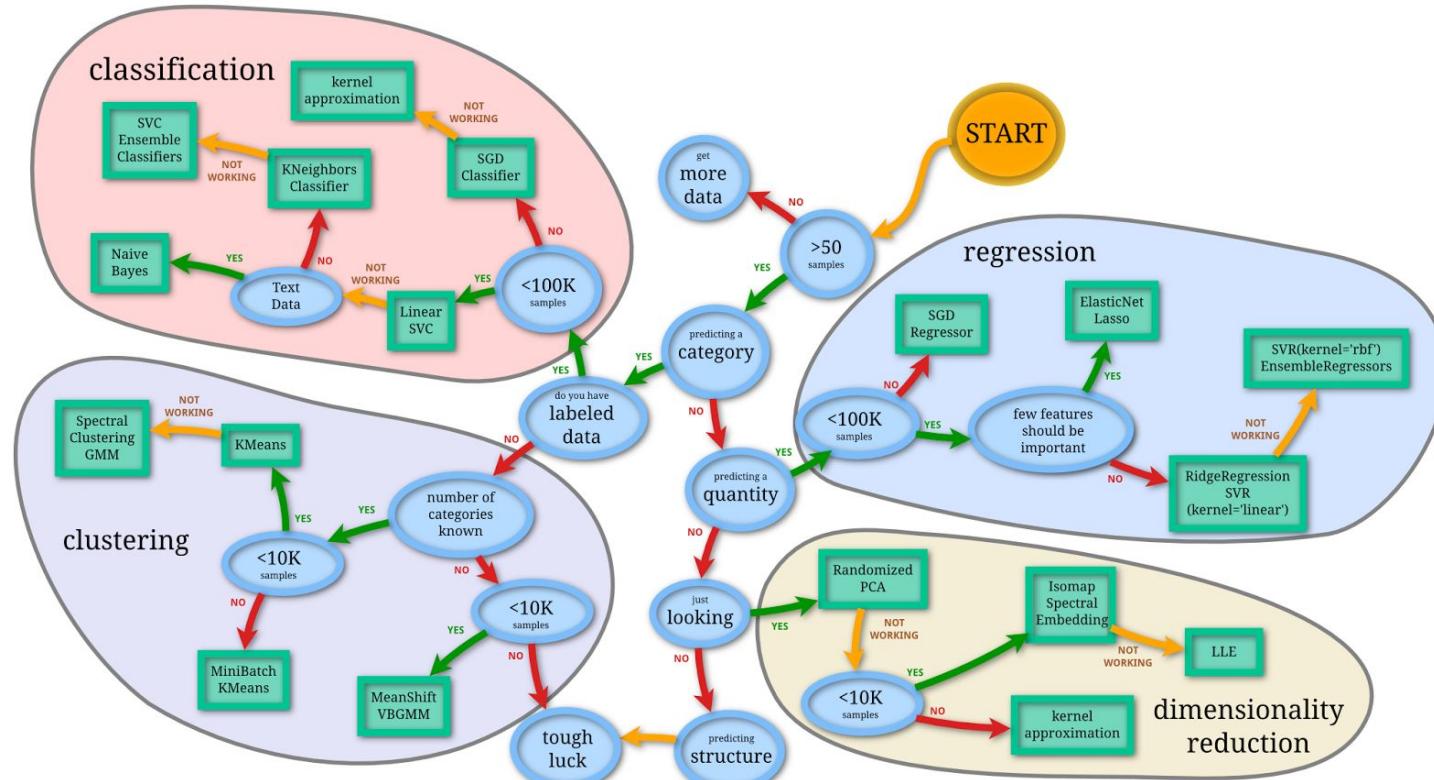
# Machine Learning - Tipos de problemas

	Variable Continua	Variable categórica
Supervisado	Regresión	Clasificación
No Supervisado	Reducción de la dimensionalidad	Clustering

# Machine Learning - Tipos de algoritmos

<b>Regresión</b> <i>Numérico, Supervisado</i>	Precio de una vivienda
<b>Clasificación</b> <i>Categórico, Supervisado</i>	Cliente compra/no compra
<b>Clustering</b> <i>Categórico, No Supervisado</i>	Segmentación de clientes
<b>Reducción de Dimensionalidad</b> <i>Numérico, No Supervisado</i>	Reducción de imágenes para clasificación

# Machine Learning - Tipos de algoritmos



# Machine Learning - Herramientas

## Programación:

- [scikit-learn](#)
- [R](#)
- [Wowpal wabbit](#)

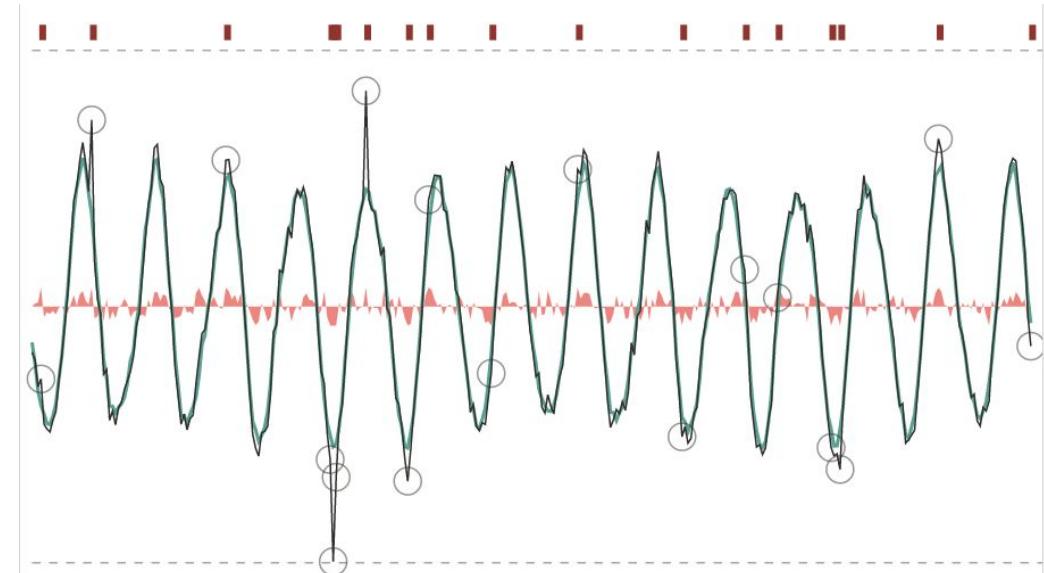
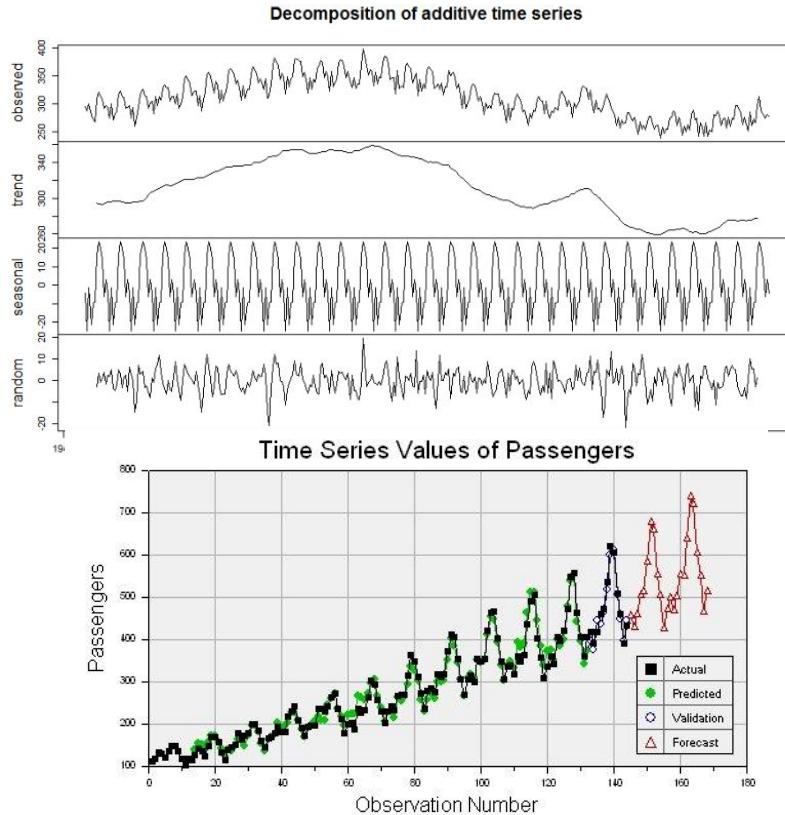
## UI:

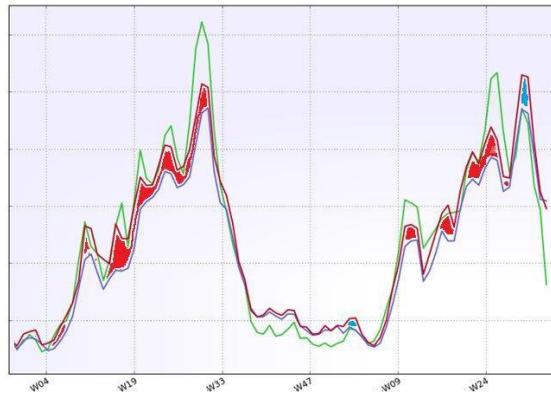
- [Weka](#)
- [Rapidminer](#)
- [SAS/ SPSS](#)

## Servicio:

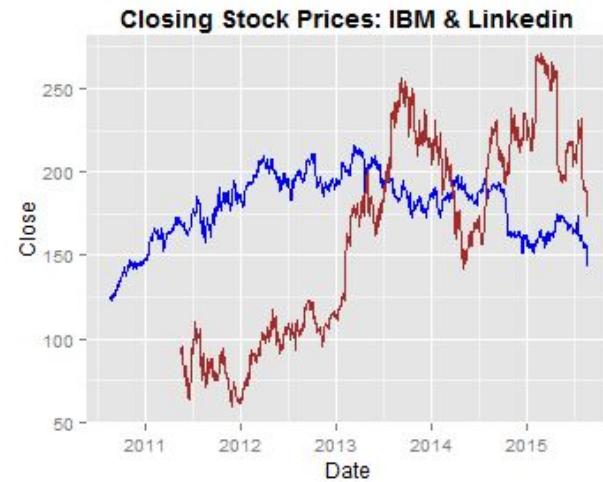
- [IBM Watson](#)
- [Google Cloud](#)
- [Windows Azure \(Cortana Intelligence\)](#)

# Análisis de Series Temporales

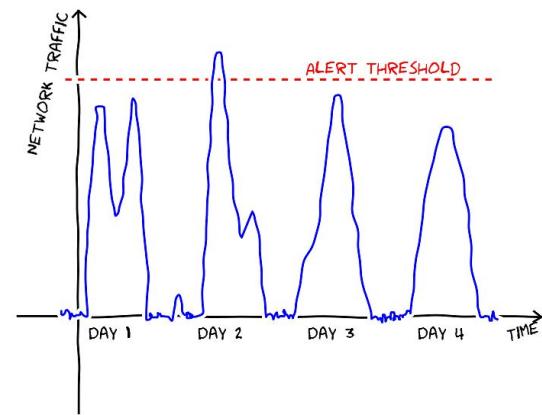




— Demand  
— Old Forecast  
— New Forecast



— ibm  
— Inkd



# Análisis de Series Temporales - Herramientas

**Python:**

- statsmodels.tsa

- Prophet

**R:**

- forecast

- CausalImpact

- Prophet

# Procesado de Lenguaje Natural (NLP)

Language: English

[Copy permalink](#)

[Show API url](#)

[Show me the response](#)

The Mona Lisa is a 16th century oil painting created by Leonardo. It's held at the Louvre in Paris.

1 person

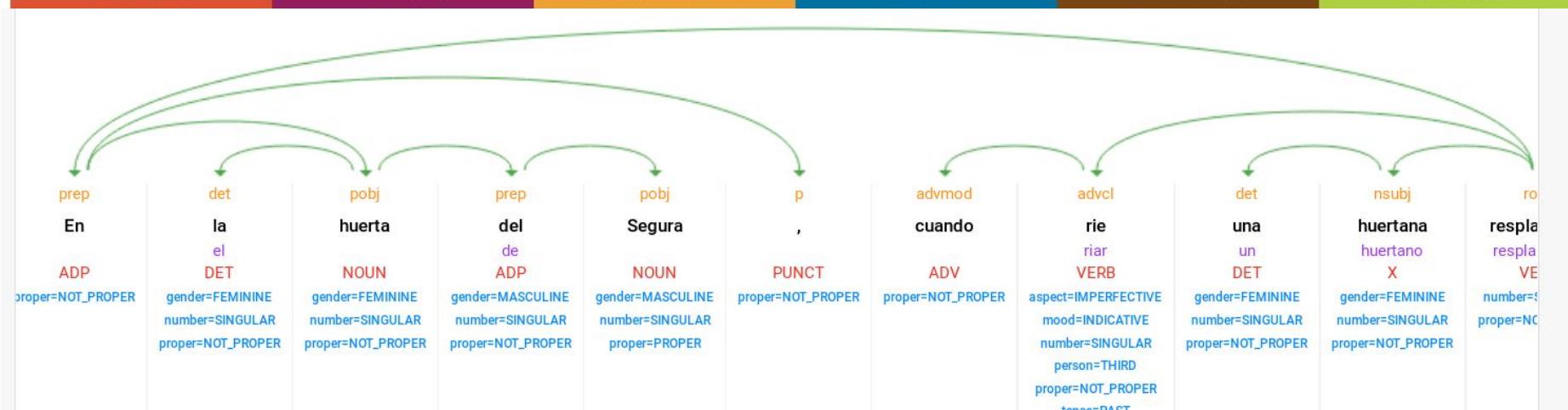
1 work

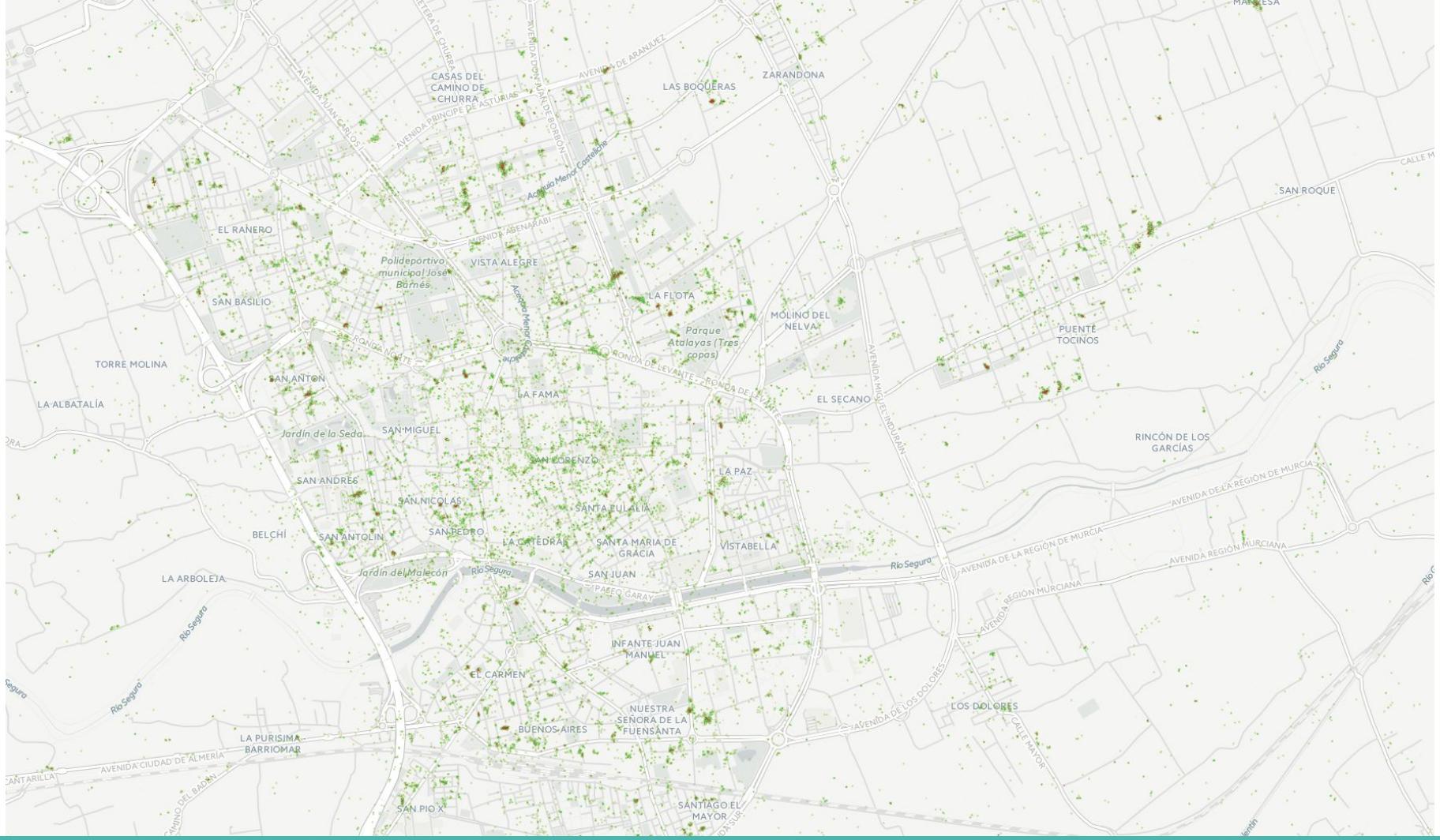
0 organisations

2 places

0 events

1 concept





The New York Times

# Cooking

TRUTH: QT UN OT  
4 tablespoons melted nonhydrogenated

NA OT  
margarine , melted coconut oil or canola oil

GUESS: QT UN OT  
4 tablespoons melted nonhydrogenated

NA CO NA OT NA  
margarine , melted coconut oil or canola oil



Google

in:spam

Gmail ▾

Compose

Inbox

Important

Sent Mail

Drafts

All Mail

Spam (5)

SoftMaker Software GmbH Vintage font

no\_reply PCWinSoft

no\_reply PCWinSoft

Abelssoft Unser Tipp:

1-abc.net News We have th

Delete forever Not spam

A red arrow points from the "Not spam" button to the "Not spam" link in the list of emails.

# Procesado de Lenguaje Natural - Herramientas

## Python:

- [NLTK](#)
- [textblob](#)
- [gensim](#)
- [spacy](#)

## R:

- [tm](#)

## Java:

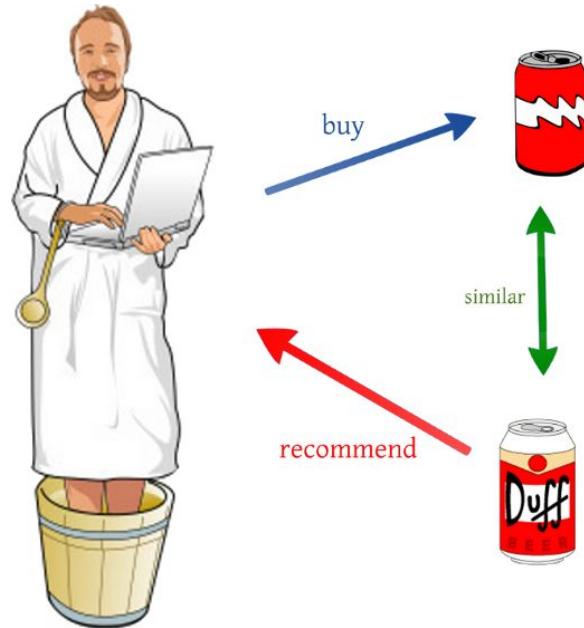
- [Stanford CoreNLP](#)

## Servicios:

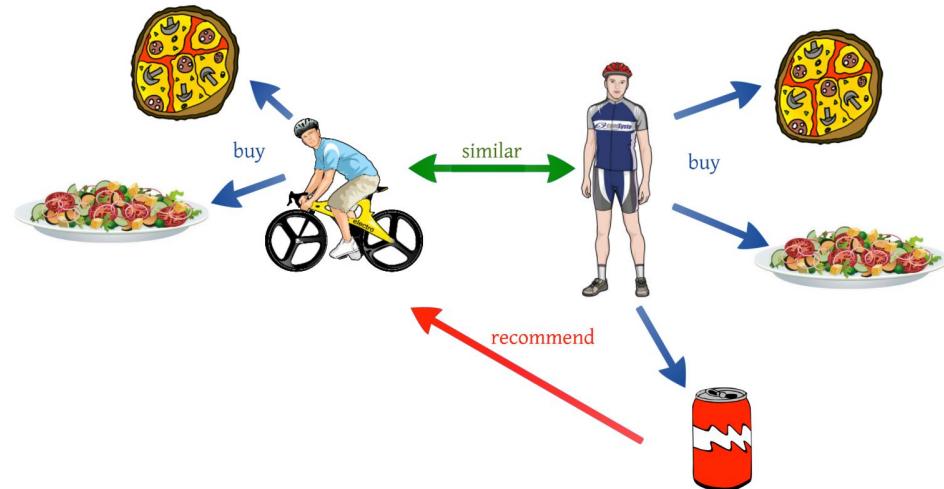
- [MonkeyLearn](#)

# Sistemas de recomendación

Filtrado de Contenidos



Filtrado Colaborativo





Busca en PcComponentes...

## Productos Relacionados



Asus B150-PLUS



MSI B150M Mortar Arctic



Gigabyte GA-H270M-DS3H



MSI B350M Mortar Arctic



MSI H170M-A PRO



101 €



101 €



105 €



105 €



105 €



87 articles viewed recently

alexander.spangher  
All Recommendations

1. THE UPSHOT  
What Ted Cruz's Early Fund-Raising Means, and Doesn't



2. Lawyers Chosen to Present Case for Gay Marriage



3. How Not to Catch a Cold on a Plane



4. LETTER  
Ted Cruz's Imaginings



5. Japan: Weeks After a Birthday, the World's Oldest Person Dies



6. VIDEO  
The Martyr's Daughter



7. Campaign Finance Complaints Filed Against 4 Presidential Hopefuls



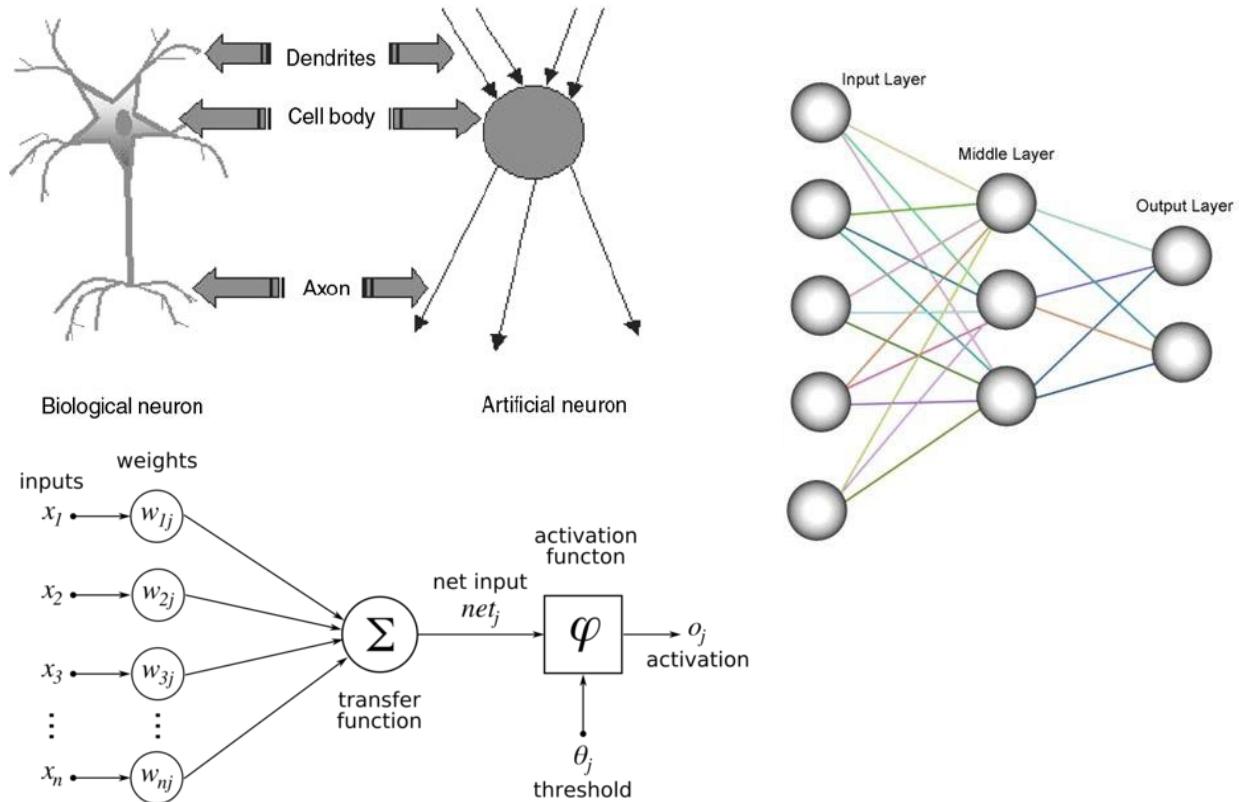
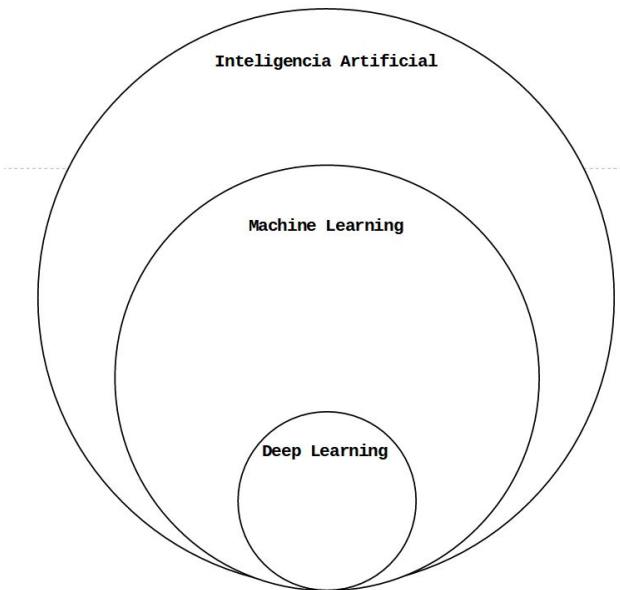
8. GAIL COLLINS  
Indiana Loses Its Game



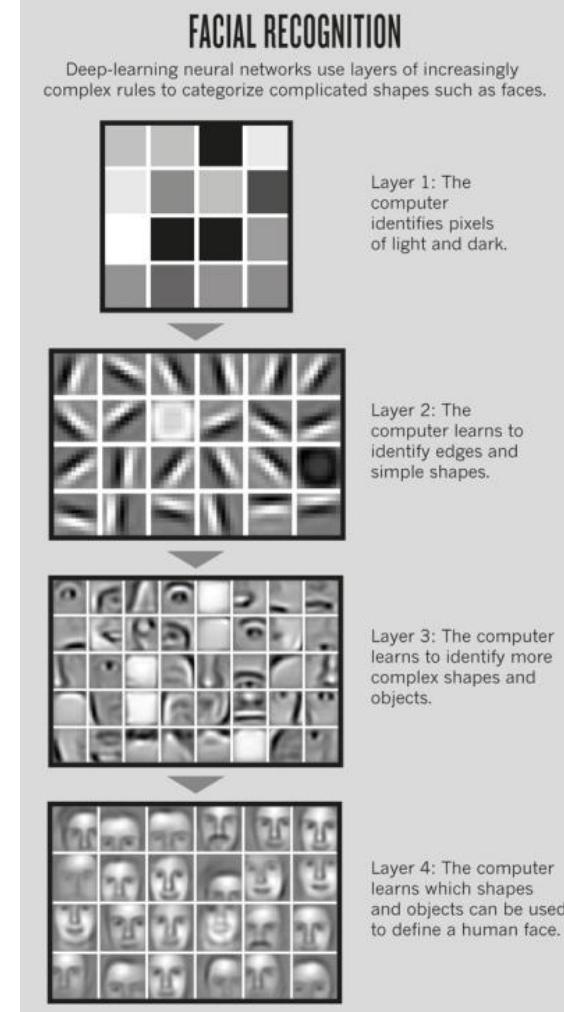
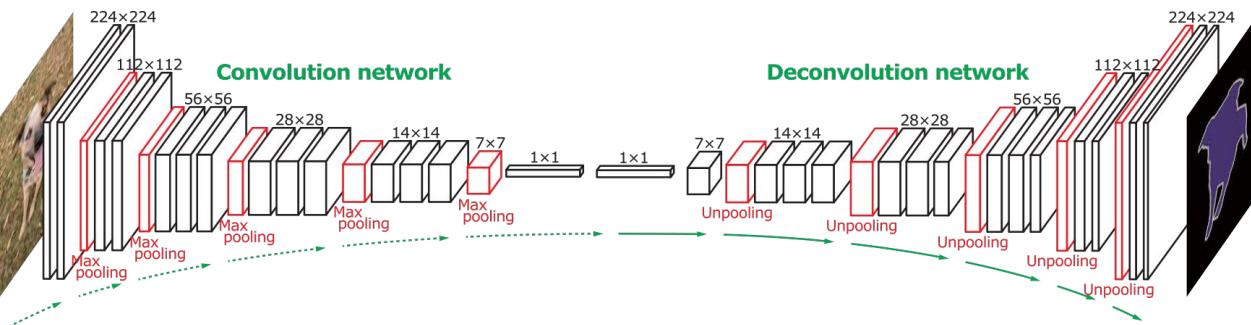
9. Eroding Freedom in the Name of Freedom



# Deep Learning



# Deep Learning



# STYLE SAGE

Input  
(Image Data)



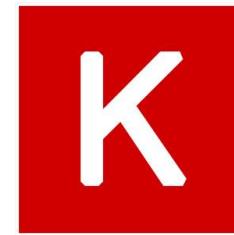
Output  
(Probability Vector)

- Dress : **94.8%**
- Skirt: **4.1%**
- Jacket: **1.2%**
- Pant: **0.1%**
- Socks: **0.01%**
- ...



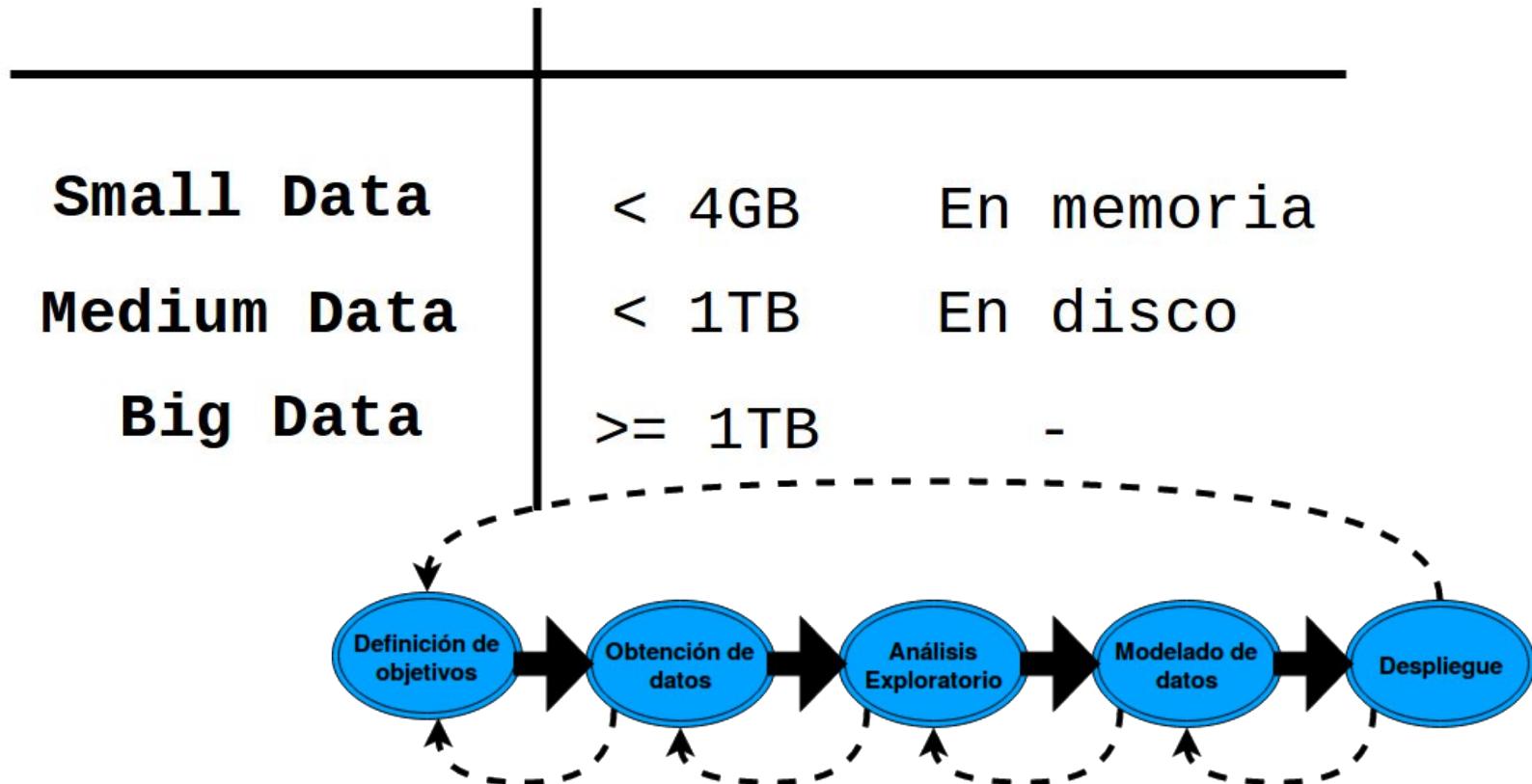
# Deep Learning - Herramientas

Caffe



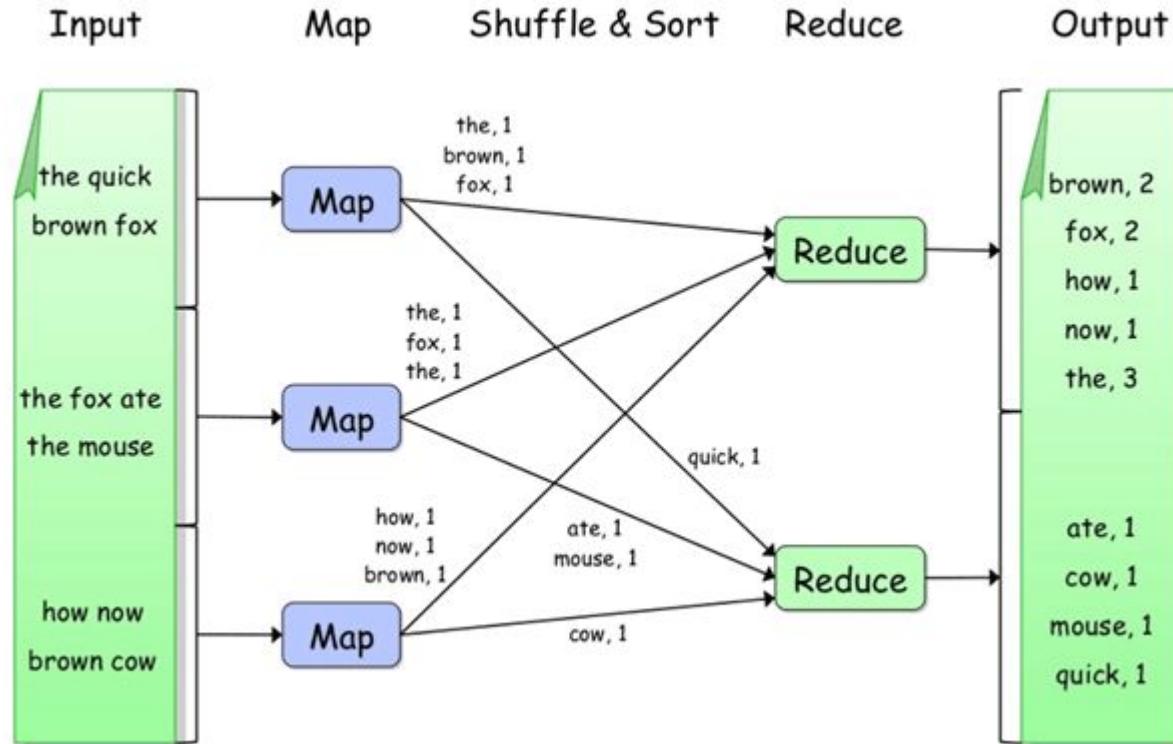
# Big Data

# Big Data

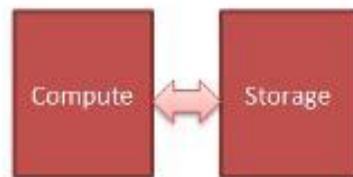


# Big Data

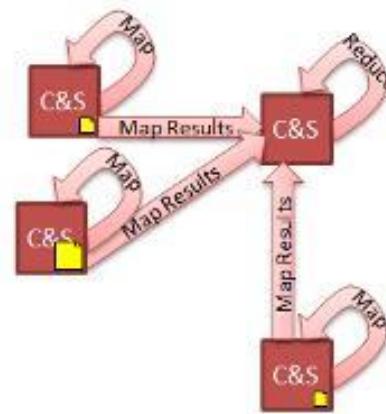
## MapReduce



# Big Data



HPC



Hadoop or  
Spark 1.1

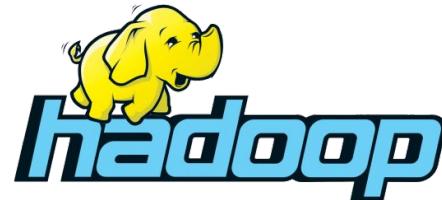


# Big Data - Herramientas

**Batch**  
(Procesamiento offline, por lotes)

**Streaming**  
(Procesamiento en tiempo real)

**Batch + Streaming**



# Big Data - Herramientas (cont.)



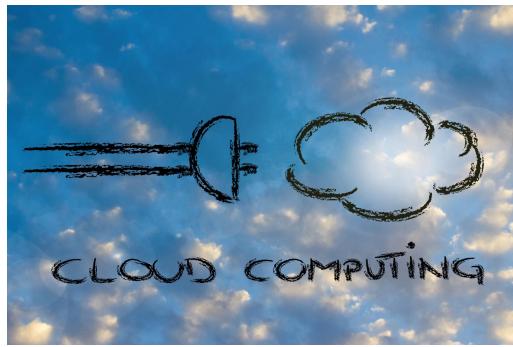
Google BigQuery



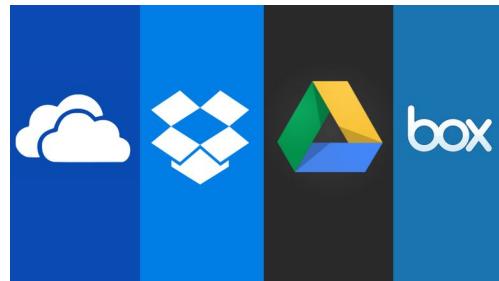
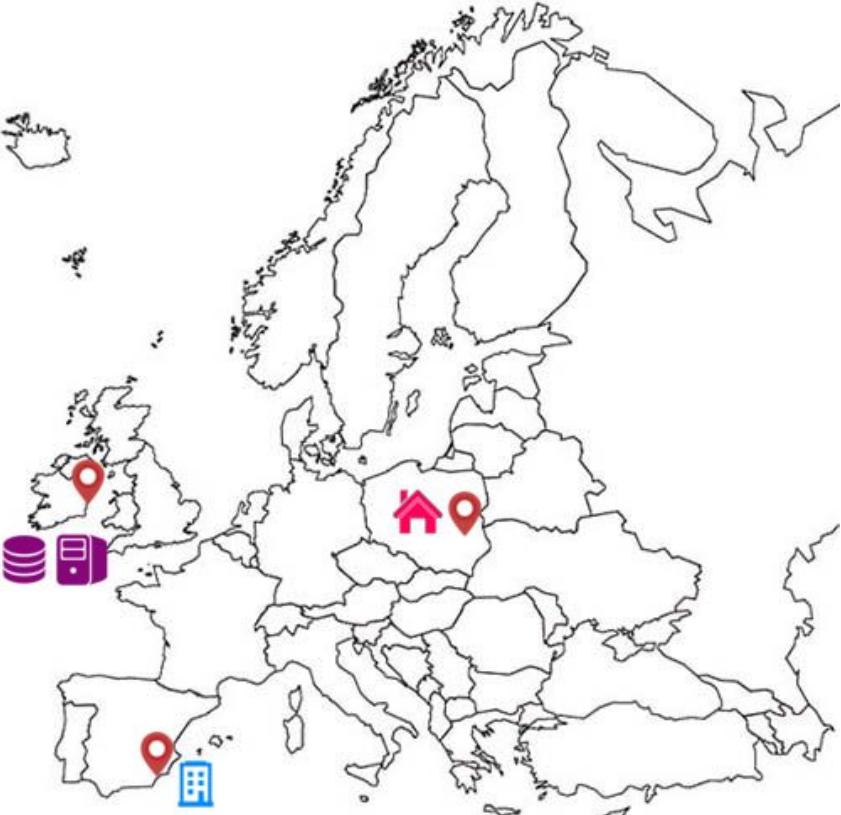
**AMAZON  
REDSHIFT**



# La nube



# Cloud Computing (la nube)



# Cloud Computing - Proveedores



# Cómo unirse a todo esto?

- 1) Mentalidad Data-Céntrica
- 2) Infraestructura
- 3) Data Scientist ( ó conocimientos de Data Science)

# Gracias!



Centro Tecnológico de la Información y las Comunicaciones

[centic@centic.es](mailto:centic@centic.es)  
[www.centic.es](http://www.centic.es)



**Manuel Garrido Peña**

[hola@manugarri.com](mailto:hola@manugarri.com)