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

centic

info REGIÓN DE MURCIA

seimed centros empresa europa

enterprise europe network

M4.0 MURCIA INDUSTRIA4.0

Sobre Mí (Manuel Garrido)

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

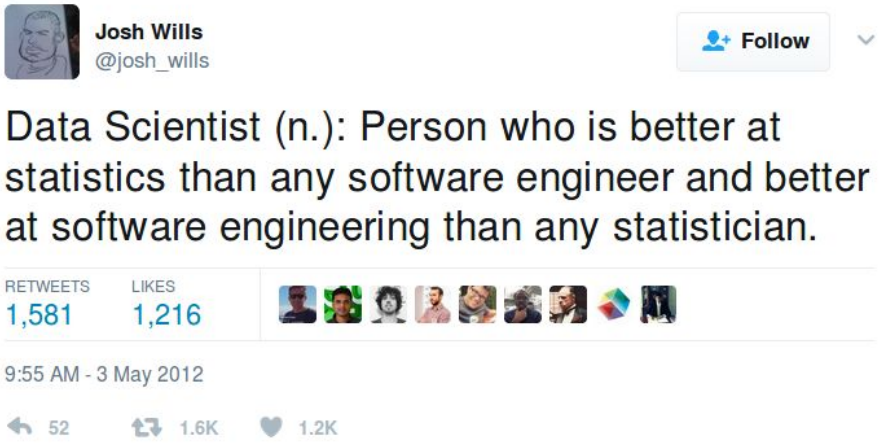
manugarri.com

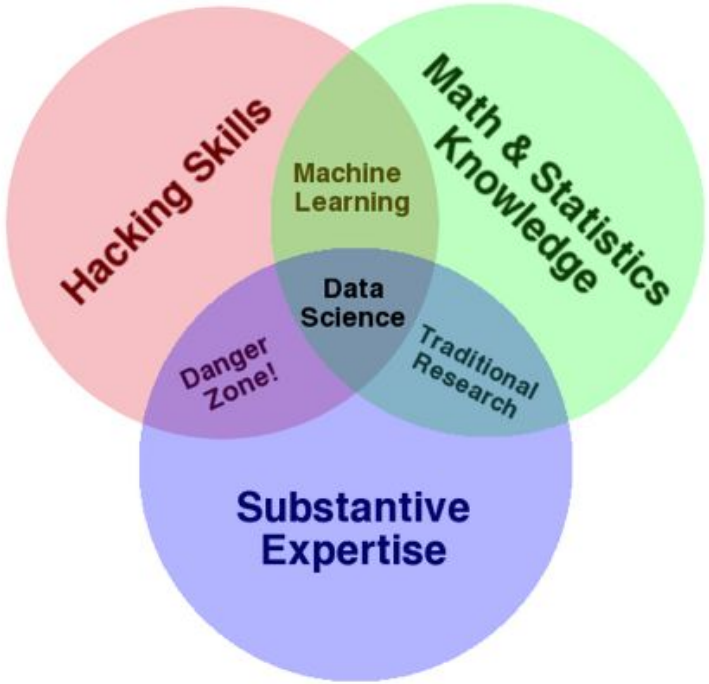
hola@manugarri.com


Data Science

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

Entonces, que es Data Science?






1. Fundamentals
Matrices & Linear Algebra Fundamentals
Hash Functions, Binary Tree, O(n)
Relational Algebra, DB Basics
Inner, Outer, Cross, Theta Join
CAP Theorem
Tabular Data
Data Frames & Series
Entropy
Sharding
OLAP
Multidimensional Data Model
ETL
Reporting Vs BI Vs Analytics
JSON & XML
NoSQL
Regex
Vendor Landscape
Env Setup
5%
2. Statistics
Prob Den Fn (PDF)
ANOVA
Skewness
Continuos Distributions (Normal, Poisson, Gaussian)
Cumul Dist Fn (CDF)
Random Variables
Bayes Theorem
Probability Theory
Percentiles & Outliers
Histograms
Exploratory Data Analysis
Central Limit Theorem
Monte Carlo Method
Hypothesis Testing
p-Value
Chi² Test
Estimation
Confid Int (CI)
MLE
Kernel Density Estimate
Regression
Covariance
15%
Factor Analysis
Install Pkgs
Descriptive Statistics (mean, median, range, SD, Var)
Pick a Dataset (UCI Repo)
3. Programming
Data Frames
Reading CSV Data
Reading Raw Data
Subsetting Data
Manipulate Data Frames
Functions
Lists
Factors
Arrays
Matrices
Vectors
Variables
Expressions
R Basics
R Setup
R Studio
Rapid Miner
IBM SPSS
Python Basics
Working in Excel
4. Machine Learning
Perceptron
Linear Regression
Ranking
Logistic Regression
Regression
K-Nearest Neighbor
Naive Bayes Classifiers
Boosting
Decision Trees
Classification Rate
Trees & Classification
Bias & Variance
Overfitting
Lift
Prediction
Classifier
Training & Test Data
Concepts, Inputs & Attributes
Unsupervised Learning
Supervised Learning
Categorical Var
Numerical Var
What is ML?
50%
Clustering
Hierarchical Clustering
K-means Clustering
Neural Networks
Sentiment Analysis
Collaborative Filtering
Tagging
Vocabulary Mapping
Classify Text
Using NLTK
Classification
Named Entity Recognition
Text Analysis
Corpus
Term Document Matrix
UIMA
Term Frequency & Weight
Support Vector Machines
Association Rules
Market Based Analysis
Feature Extraction
Using Mahout
Using Weka
5. Text Mining / NLP
6. Visualization
Euclidean Distance
Least Fit
Causation
Pearson Coeff
Correlation
30%
Data Exploration in R (Hist, Boxplot etc)
Uni, Bi & Multivariate Viz
ggplot2
Histogram & Pie (Uni)
Tree & Tree Map
Scatter Plot (Bi)
Line Charts (Bi)
Spatial Charts
Survey Plot
Timeline
Decision Tree
40%
Tableau
IBM ManyEyes
InfoVis
D3.js
7. Big Data
Job & Task Tracker
MR Programming
Sroop: Loading Data in HDFS
Flume, Scribe : For Unstruct Data
SQL with Pig
DWH with Hive
Scribe, Chukwa For Weblog
Using Mahout
Name & Data Nodes
Setup Hadoop (IBM / Cloudera / HortonWorks)
Data Replication Principles
HDFS
Hadoop Components
Map Reduce Fundamentals
Zookeeper Avro
Storm: Hadoop Realtime
Rhadoop, RHIPE
rmr
Cassandra
MongoDB, Neo4j
60%
8. Data Ingestion
Sampling
Stratified Sampling
Principal Component Analysis
Transformation & Enrichment
Data Survey
Using ETL
How much Data?
Google OpenRefine
80%
Data Fusion
Data Integration
Data Sources & Acquisition
Data Discovery
Summary of Data Formats
9. Data Munging
Denoising
Feature Extraction
Binning Sparse Values
Unbiased Estimators
Handling Missing Values
Data Scrubbing
Normalization
Dimensionality & Numerosity Reduction
10. Toolbox
MS Excel w/ Analysis ToolPak
Java, Python
R, R-Studio, Rattle
Weka, Knime, RapidMiner
Hadoop Dist of Choice
Spark, Storm
Flume, Scibe, Chukwa
Nutch, Talend, Scraperwiki
Webscraper, Flume, Sqoop
tm, RWeka, NLTK
RHIPE
D3.js, ggplot2, Shiny
IBM Languageware
Cassandra, MongoDB
100%
Author: Swami Chandrasekaran

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

datos.gob.es
reutiliza la información pública

DATOS ABIERTOS
EN LA REGIÓN DE MURCIA

Data Science en el marco Europeo - Conferencias

EuroScipy (Erlangen)

Data Science Summit Europe (29 Mayo Jerusalem)

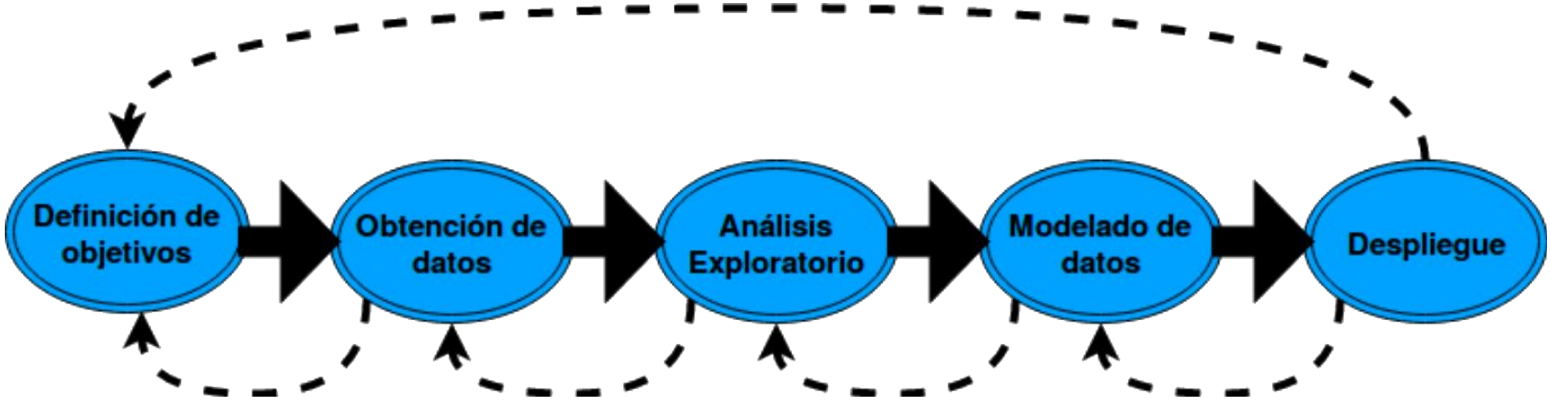
Spark Summit (24-26 Octubre Dublin)

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

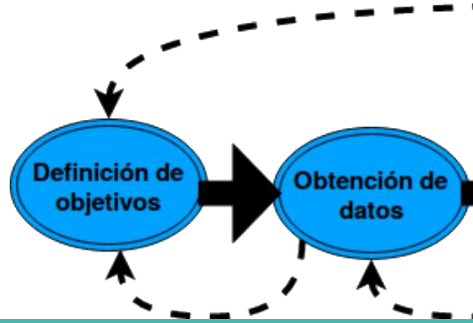
- Experian
- Acxiom

APIs

- Yahoo finance
- OpenWeather

Web Scraping

- Scrapy
- import.io



Tipos de datos - variables

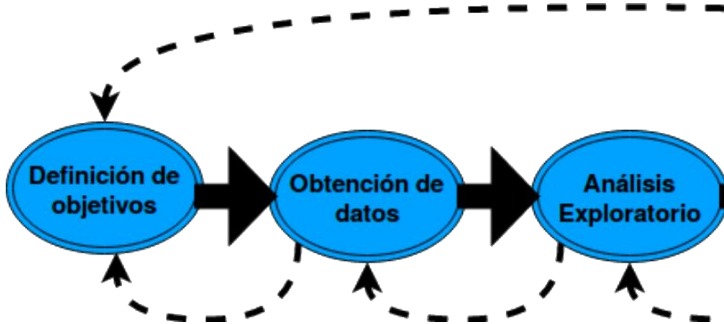
Continuas	Edad, Altura, colores RGB
Ordinales	Ratings, Niveles educativos, Muy de acuerdo/De acuerdo/En desacuerdo
Categóricas	Hombre/Mujer, Apto/No Apto, días de la semana

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) Accesibles, Accionables y Comparables
   b) Internas o Externas
2) **Analytics**
   a) Analizar y extraer valor de las métricas

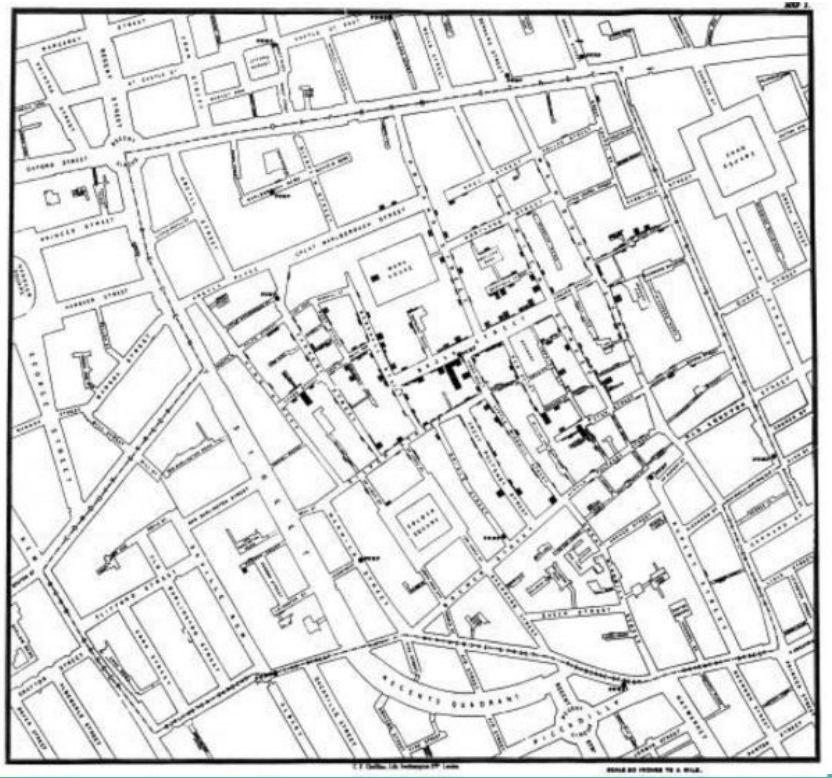


Analytics

Métricas	KPI	Analytics
# de peticiones a un servidor	Tiempo medio entre despliegues de versiones	¿Qué partes de la aplicación tardan mas en responder?
Latencia Media de una petición		
# de demos hechas a clientes en los últimos 30 dias	Número de clientes nuevos en los últimos 30 días	¿Qué canales de marketing funcionan mejor para conseguir clientes?
# de Visitas a página web por canal		¿Qué funcionalidades reclaman mas nuestros potenciales clientes?

Visualización de Datos

Mapa de John Snow

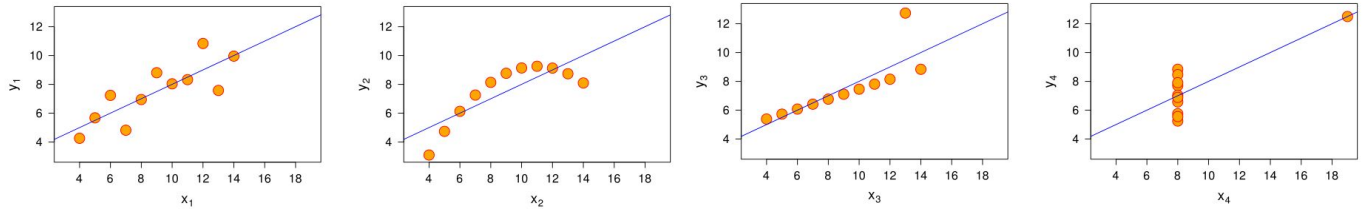


Visualización de Datos

Cuarteto de Anscombe

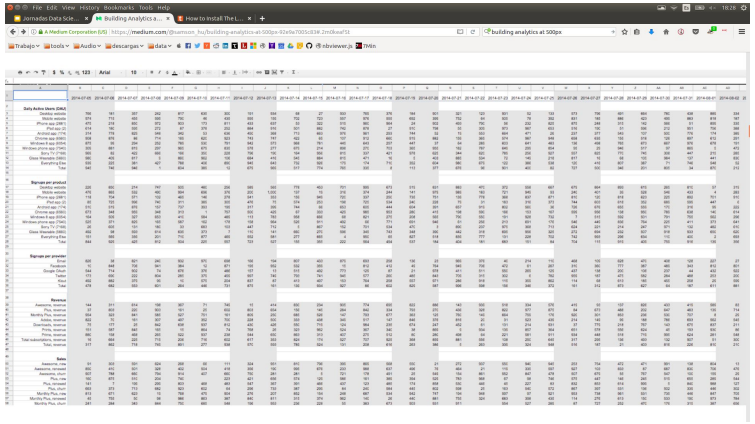
I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.5
12	10.84	12	9.13	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

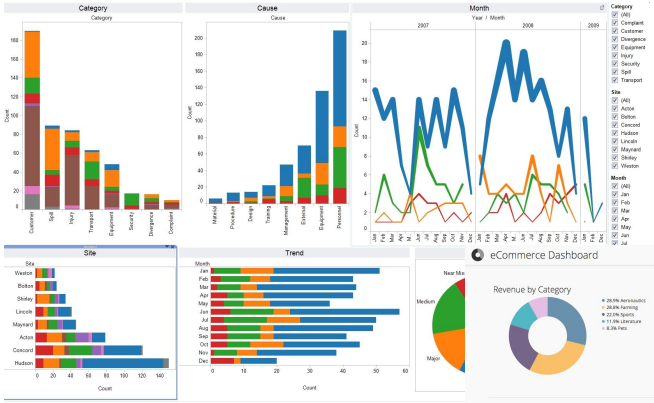
Propiedad	Valor
Media de cada una de las variables *x*	9.0
Varianza de cada una de las variables *x*	11.0
Media de cada una de las variables *y*	7.5
Varianza de cada una de las variables *y*	4.12
Correlación entre cada una de las variables *x* e *y*	0.816
Recta de regresión	$y = 3 + 0.5x$

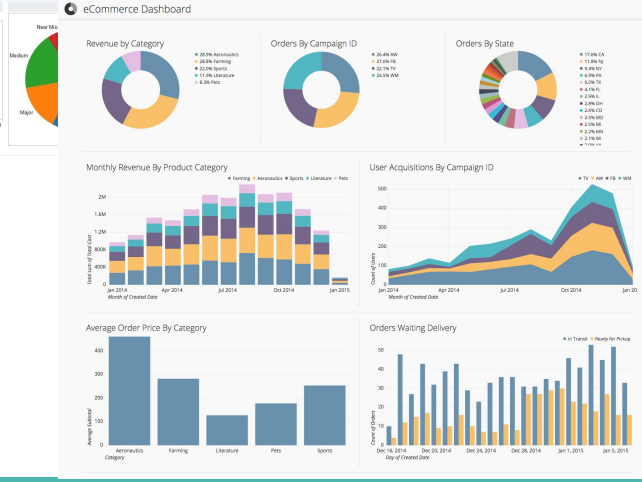


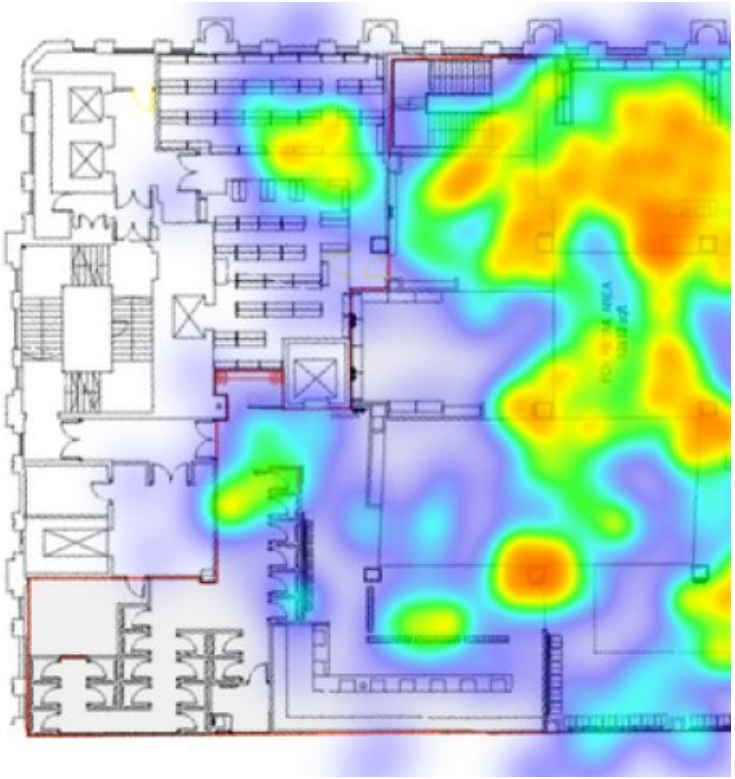
Visualización de Datos

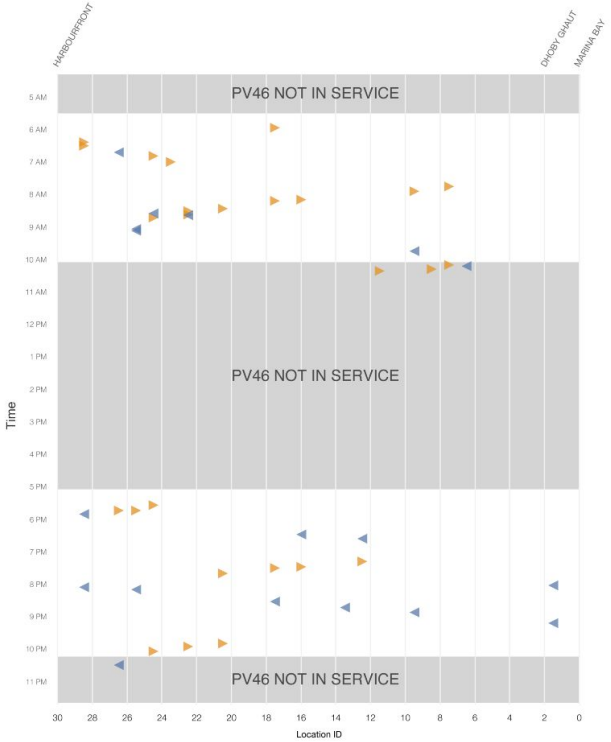
Dashboard (cuadro de mando)

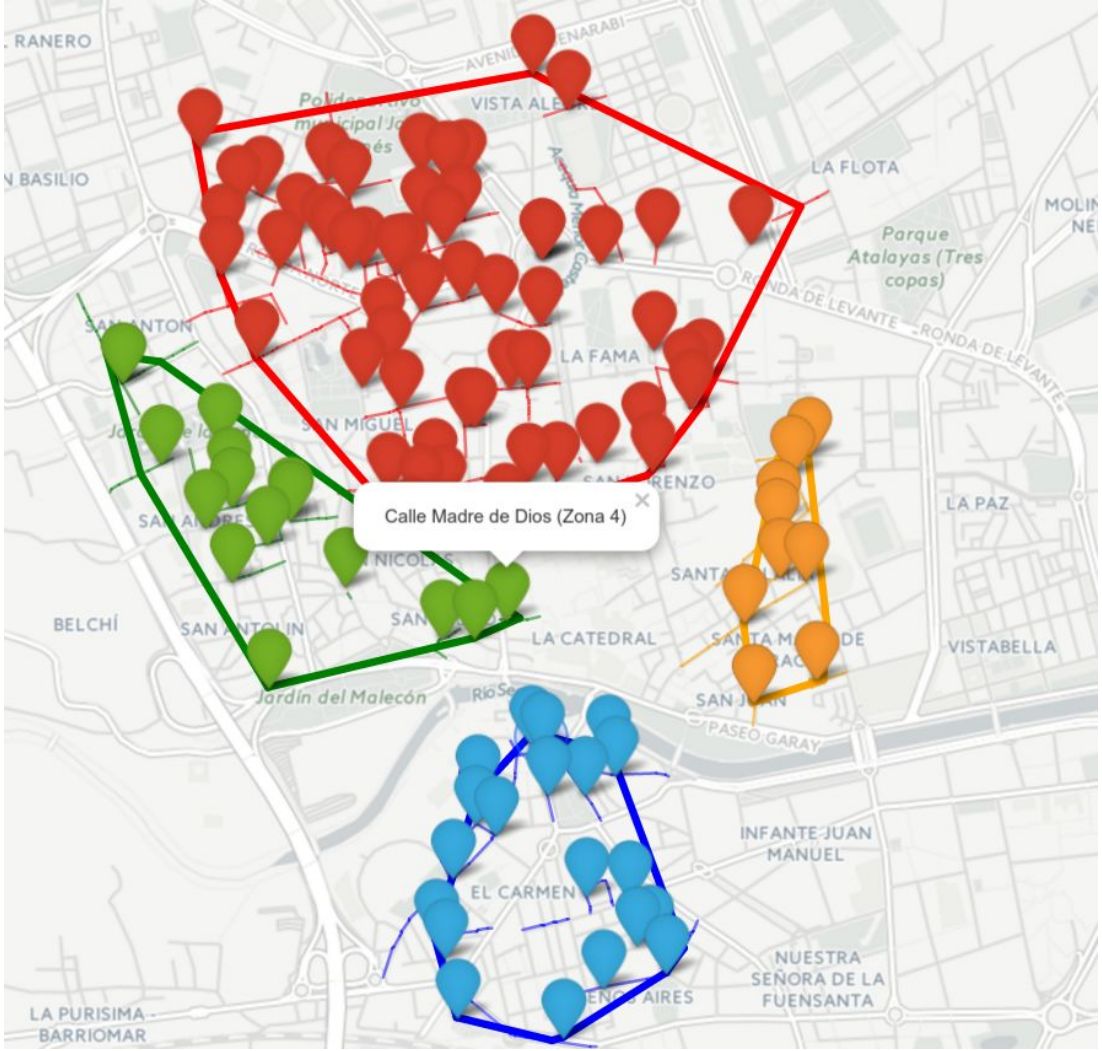









HARBOURFRONT
DHOBY GHAUT
MARINA BAY
PV46 NOT IN SERVICE
PV46 NOT IN SERVICE
PV46 NOT IN SERVICE
5 AM
6 AM
7 AM
8 AM
9 AM
10 AM
11 AM
12 PM
1 PM
2 PM
3 PM
4 PM
5 PM
6 PM
7 PM
8 PM
9 PM
10 PM
11 PM
Time
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
Location ID


Calle Madre de Dios (Zona 4)
VISTA ALEGRE
LA FLOTA
Parque Atalayas (Tres copas)
RONDA DE LEVANTE
LA FAMA
SAN MIGUEL
LA PAZ
BELCHÍ
LA CATEDRAL
VISTABELLA
Jardín del Malecón
PASEO GARAY
INFANTE JUAN MANUEL
EL CARMEN
NUESTRA SEÑORA DE LA FUENSANTA
LA PURISIMA - BARRIOMAR

Visualizacion de datos - Herramientas

Dashboards (BI):

-Tableau
- Chartio
- Grafana

Desarrollo web (JS):

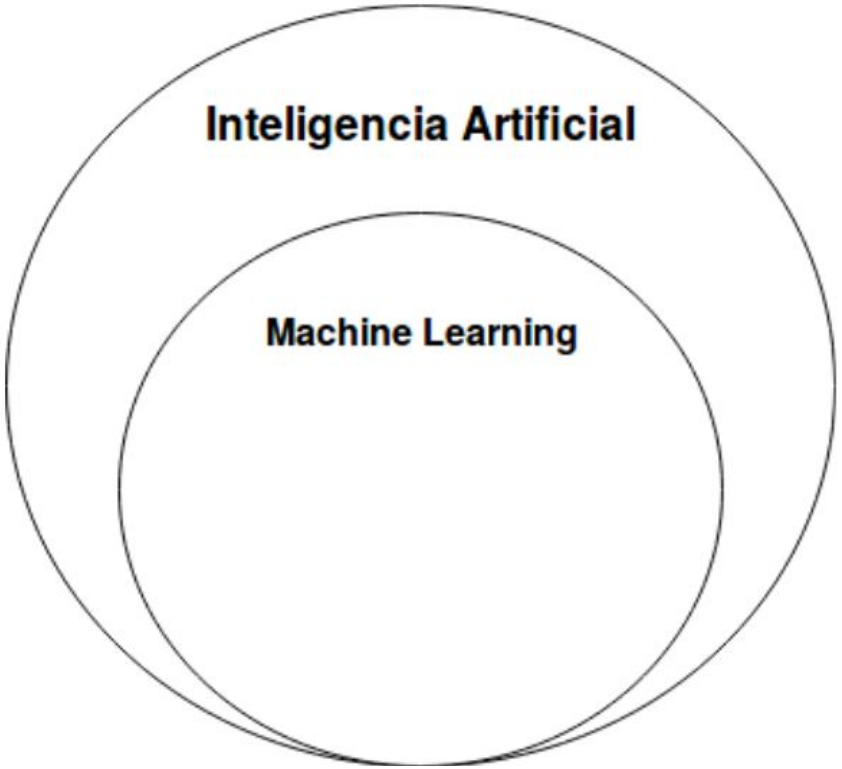
- D3.js
- ChartJS
- plotly

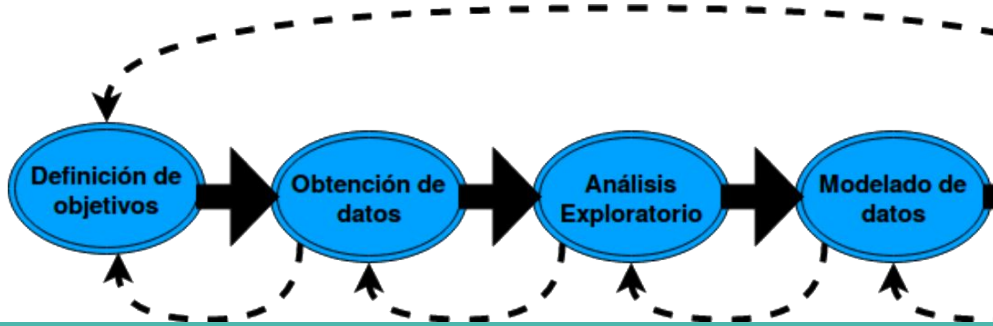
Graficos científicos:

- ggplot2
- Bokeh
- seaborn

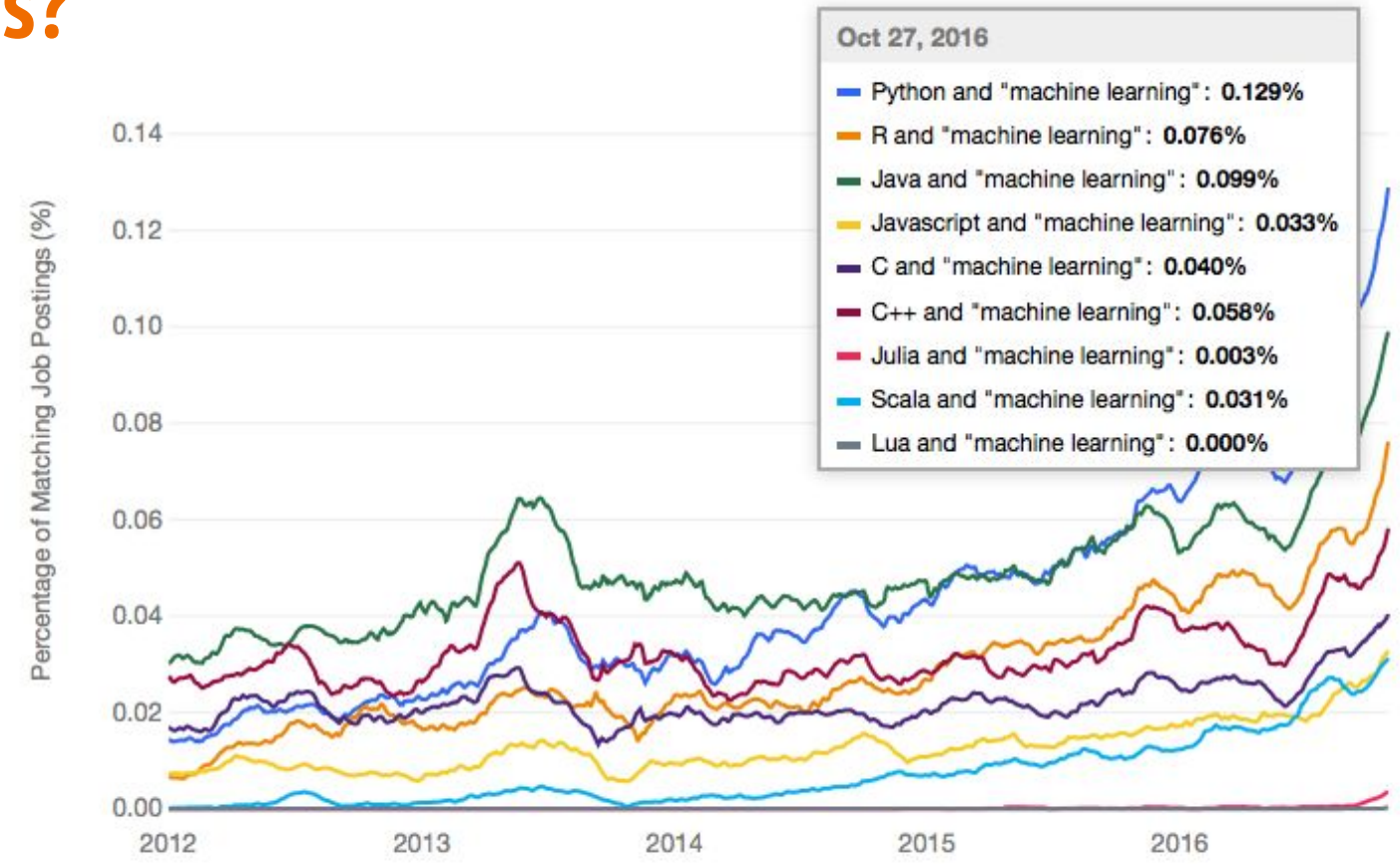

Machine Learning

Machine Learning





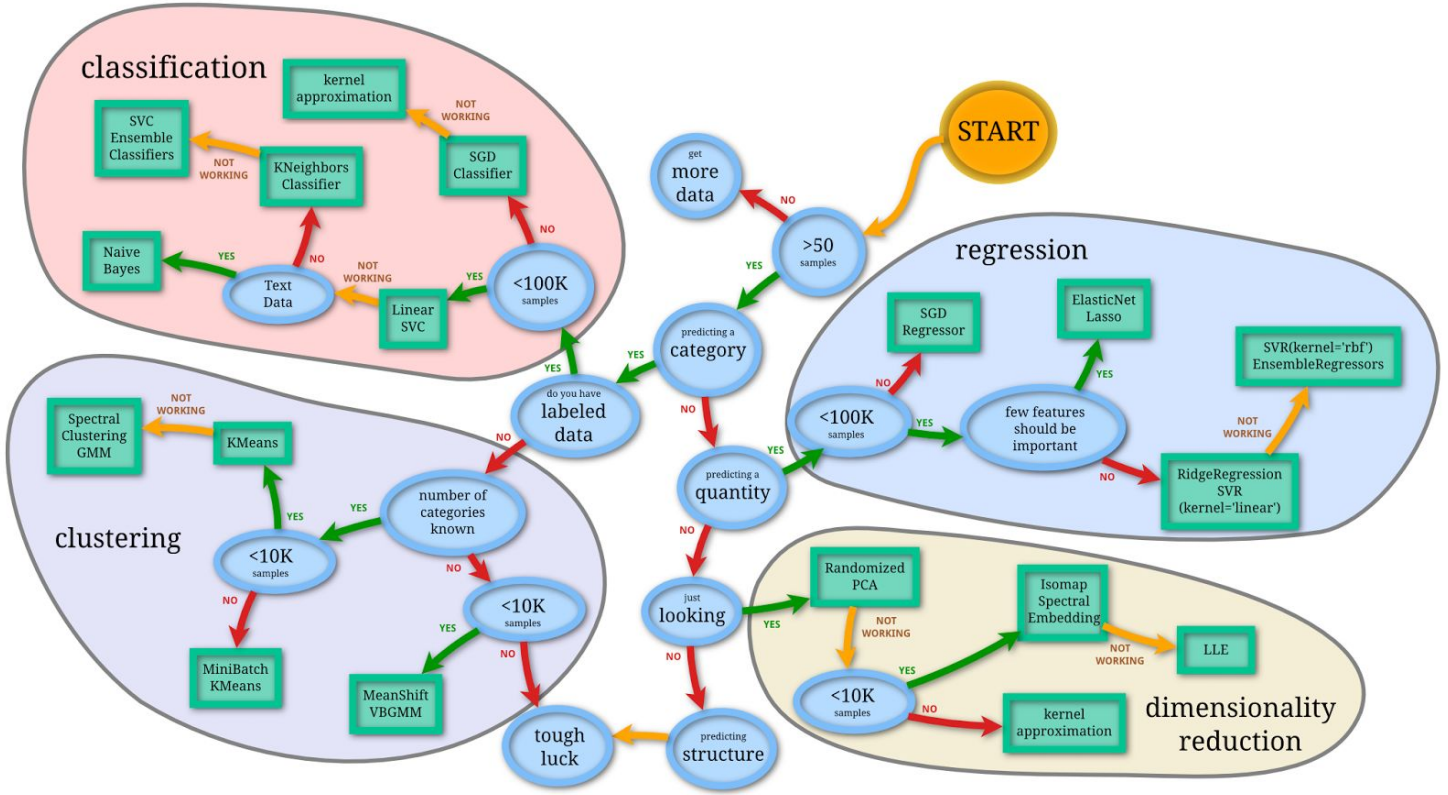
Lenguajes?


Oct 27, 2016
Python and "machine learning": 0.129%
R and "machine learning": 0.076%
Java and "machine learning": 0.099%
Javascript and "machine learning": 0.033%
C and "machine learning": 0.040%
C++ and "machine learning": 0.058%
Julia and "machine learning": 0.003%
Scala and "machine learning": 0.031%
Lua and "machine learning": 0.000%
Percentage of Matching Job Postings (%)
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
2012
2013
2014
2015
2016

Machine Learning - Tipos de problemas

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

Machine Learning - Tipos de algoritmos



Machine Learning - Tipos de algoritmos

Regresión *Numérico, Supervisado*	Precio de una vivienda
Clasificacion *Categórico, Supervisado*	Cliente compra/no compra
Clustering *Categórico, No Supervisado*	Segmentación de clientes
Reducción de Dimensionalidad *Numérico, No Supervisado*	Reducción de imágenes para clasificacion

Machine Learning - Herramientas

Programación:

- scikit-learn
- R
- Wowpal wabbit

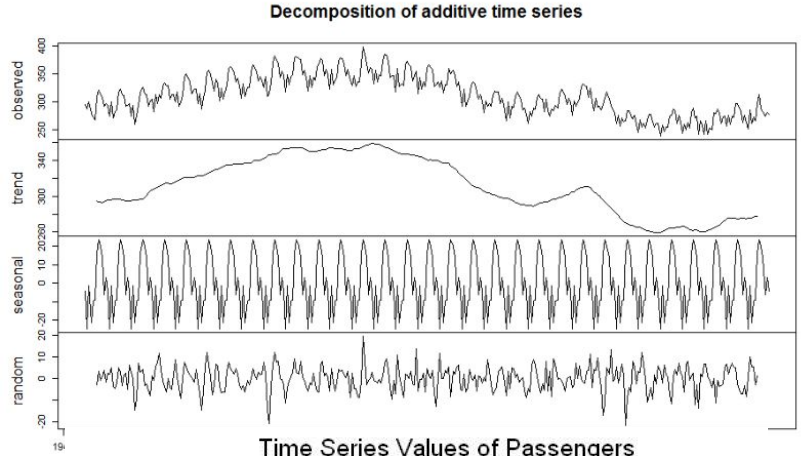
UI:

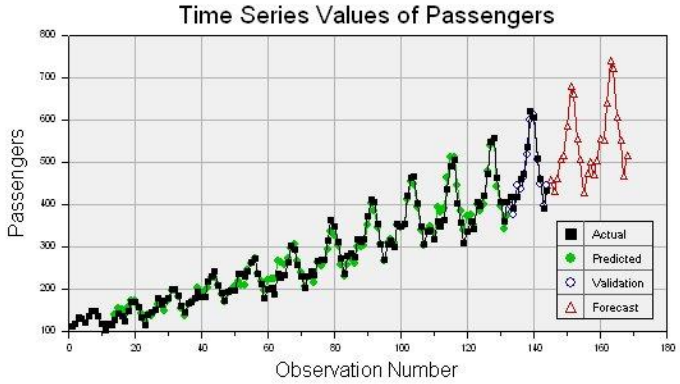
- Weka
- Rapidminer
- SAS/ SPSS

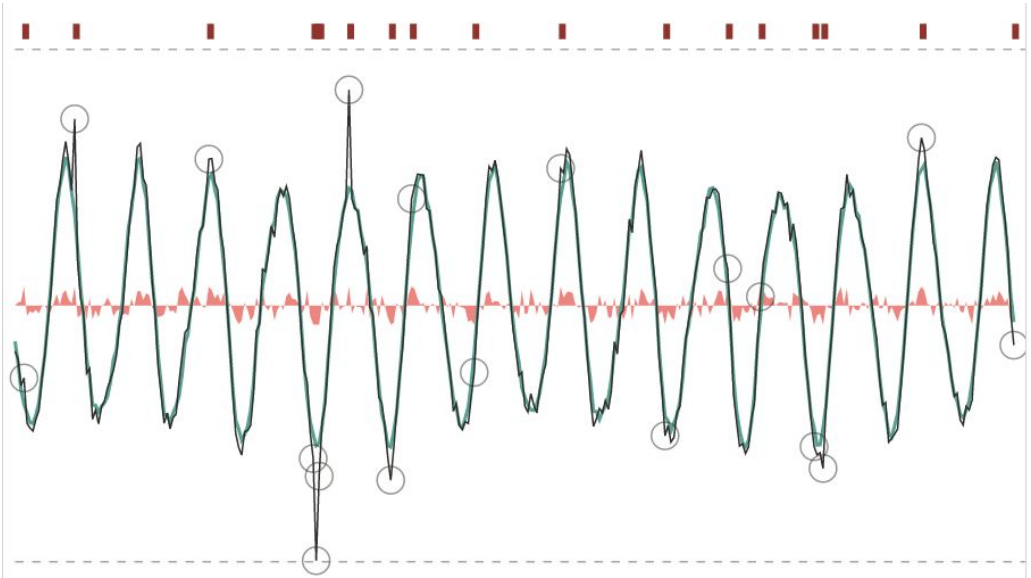
Servicio:

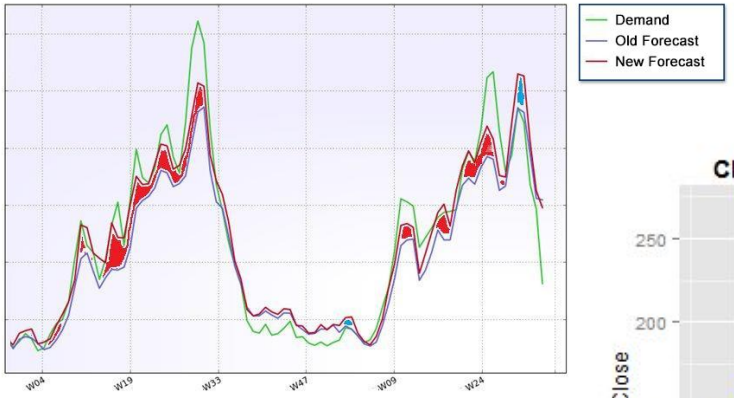
- IBM Watson
- Google Cloud
- Windows Azure (Cortana Intelligence)

Análisis de Series Temporales

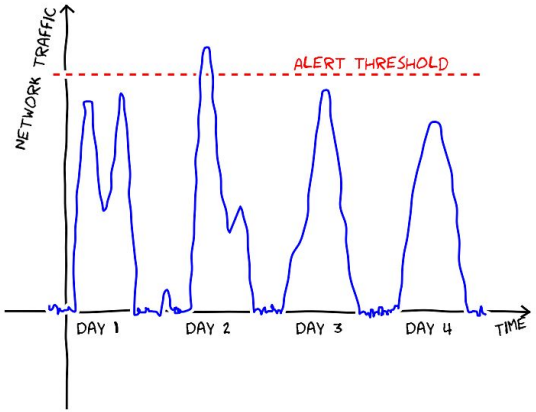







Demand
Old Forecast
New Forecast
W04
W19
W33
W47
W09
W24


Closing Stock Prices: IBM & Linkedin
Close
250
200
150
100
50
2011
2012
2013
2014
2015
Date
ibm
lnkd


NETWORK TRAFFIC
ALERT THRESHOLD
DAY 1
DAY 2
DAY 3
DAY 4
TIME

Análisis de Series Temporales - Herramientas

Python:

- statsmodels.tsa

- Prophet

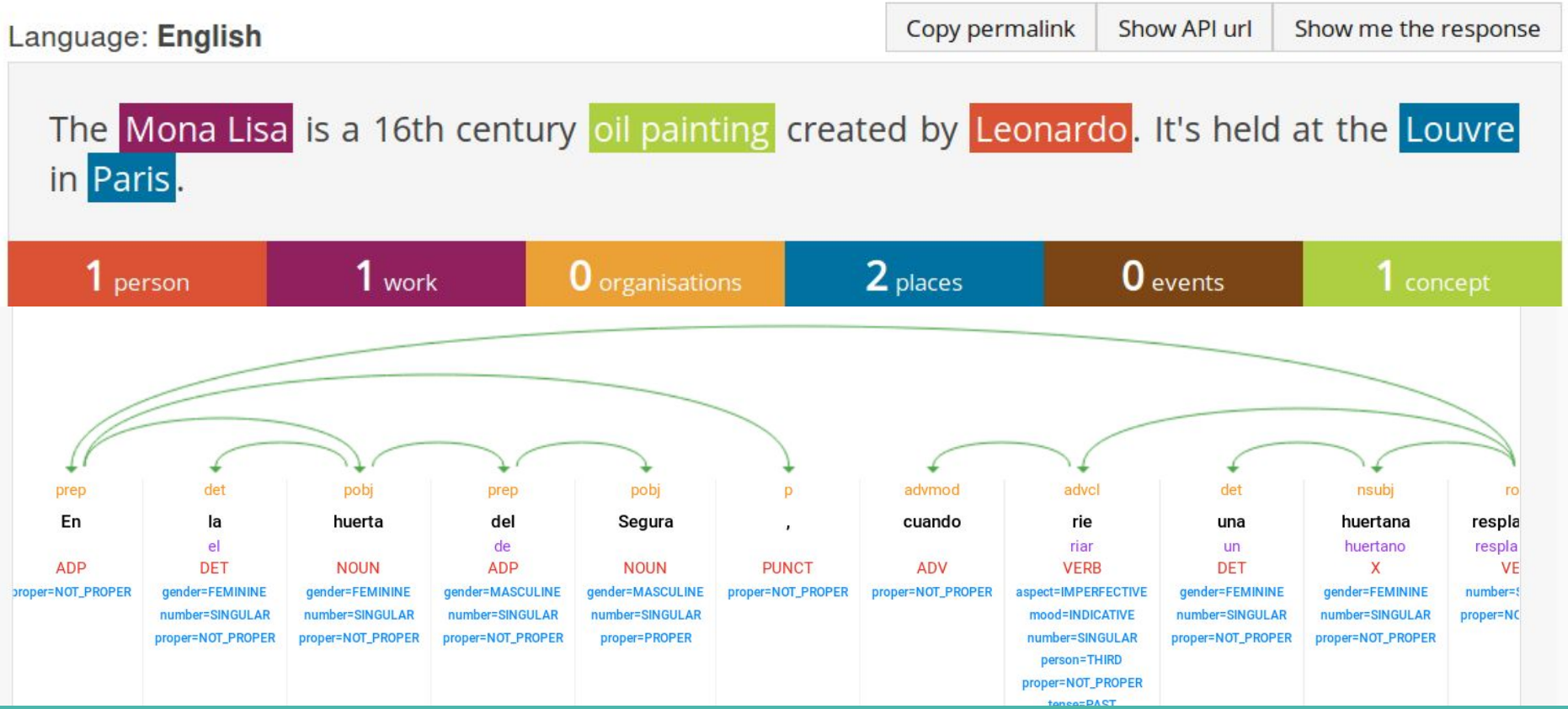
R:

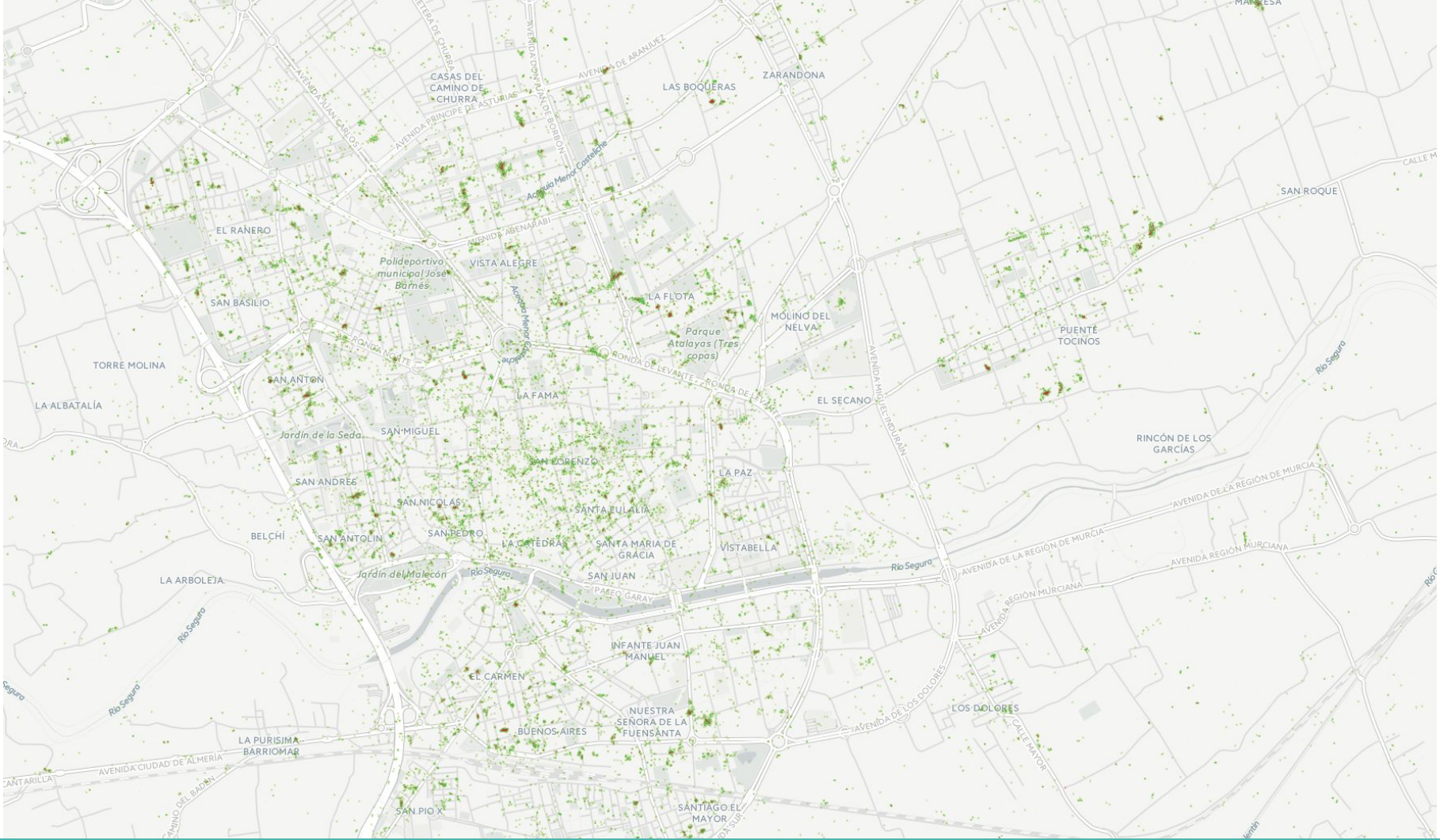
- forecast

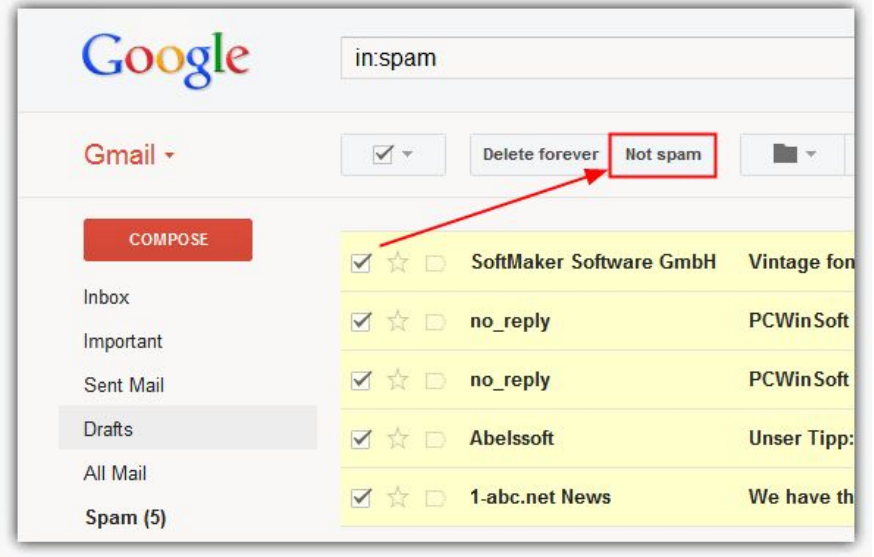
- CausalImpact

- Prophet

Procesado de Lenguaje Natural (NLP)




CASAS DEL CAMINO DE CHURRA
LAS BOQUERAS
ZARANDONA
SAN ROQUE
EL RANERO
Polideportivo municipal José Barnés
VISTA ALEGRE
SAN BASILIO
LA FLOTA
MOLINO DEL NELVA
Parque Atalayas (Tres copas)
PUENTE TOCINOS
TORRE MOLINA
SAN ANTON
LA ALBATALÍA
LA FAMA
EL SECANO
Jardín de la Seda
SAN MIGUEL
RINCÓN DE LOS GARCÍAS
SAN LORENZO
LA PAZ
SAN ANDRES
SAN NICOLAS
SANTA EULALIA
BELCHÍ
SAN ANTOLIN
SAN PEDRO
LA CATEDRA
SANTA MARIA DE GRACIA
VISTABELLA
LA ARBOLEJA
Jardín del Malecón
SAN JUAN
INFANTE JUAN MANUEL
EL CARMEN
NUESTRA SEÑORA DE LA FUENSANTA
LOS DOLORES
BUENOS AIRES
LA PURISIMA BARRIOMAR
SAN PIO X
SANTIAGO EL MAYOR
Río Segura
AVENIDA DE ARANJUEZ
AVENIDA PRINCIPE DE ASTURIAS
AVENIDA JUAN CARLOS I
AVENIDA DON JUAN DE BORBON
Acequia Menor Castellón
AVENIDA BENAHABIL
RONDA DE LEVANTE
RONDA DE LEVANTE
AVENIDA MIGUEL INDURAIN
AVENIDA DE LA REGIÓN DE MURCIA
AVENIDA REGIÓN MURCIANA
PASEO GARAY
AVENIDA DE LOS DOLORES
CALLE MAYOR
AVENIDA CIUDAD DE ALMERIA
CAMINO DEL BADEN



The New York Times

Cooking

TRUTH: QT 4 UN tablespoons OT melted nonhydrogenated NA margarine OT , melted coconut oil or canola oil

GUESS: QT 4 UN tablespoons OT melted nonhydrogenated NA margarine CO , melted NA coconut oil OT or canola NA oil

Procesado de Lenguaje Natural - Herramientas

Python:

- NLTK
- textblob
- gensim
- spacy

R:

- tm

Java:

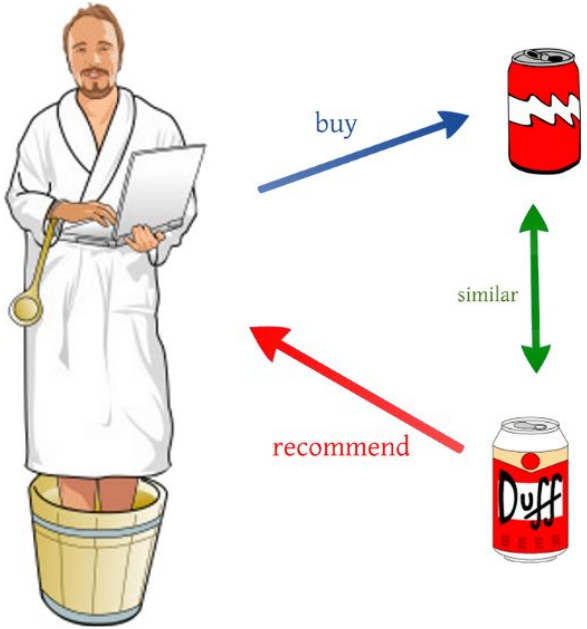
- Stanford CoreNLP

Servicios:

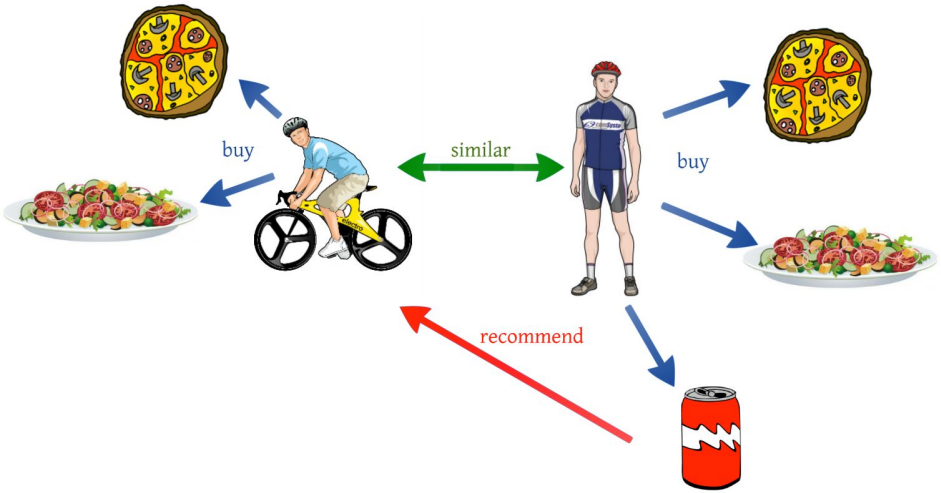
- MonkeyLearn

Sistemas de recomendación

Filtrado de Contenidos



Filtrado Colaborativo



PC COMPONENTES

Busca en PcComponentes...

Productos Relacionados

Asus B150-PLUS

101 €

MSI B150M Mortar Artic

101 €

Gigabyte GA-H270M-DS3H

105 €

NOVEDAD PROMOCIÓN

MSI B350M Mortar Arctic

105 €

MSI H170M-A PRO

105 €

MOST EMAILED | MOST VIEWED | RECOMMENDED FOR YOU

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

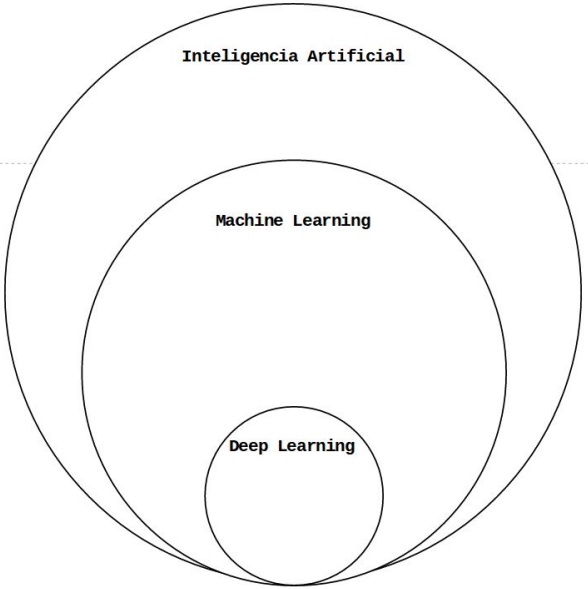
6. VIDEO
The Martyr's Daughter

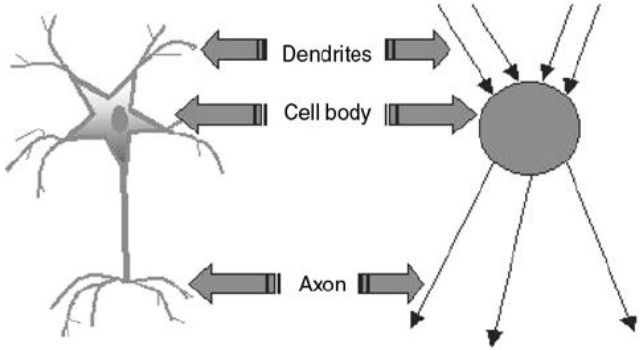
7. Campaign Finance Complaints Filed Against 4 Presidential Hopefuls

8. GAIL COLLINS
Indiana Loses Its Game

9. Eroding Freedom in the Name of Freedom

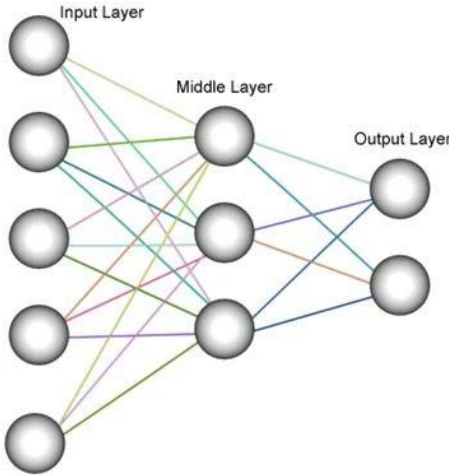
Deep Learning

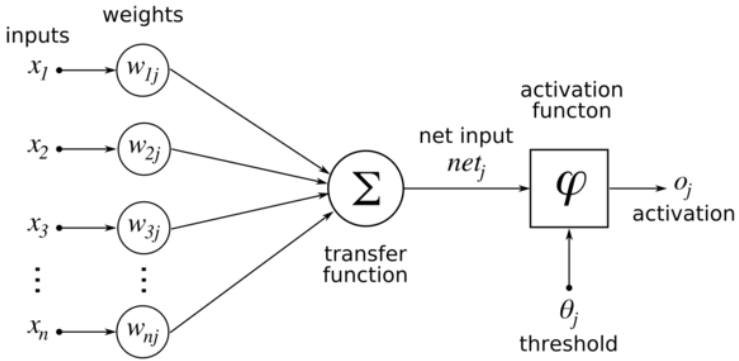




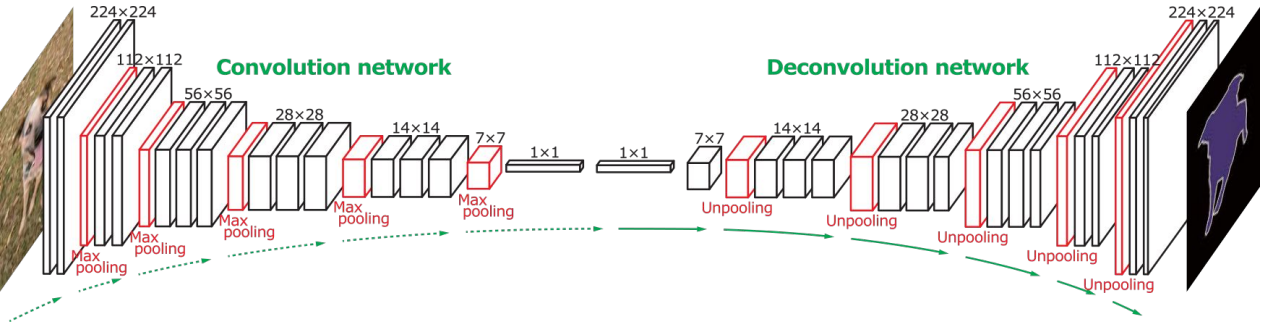
Biological neuron

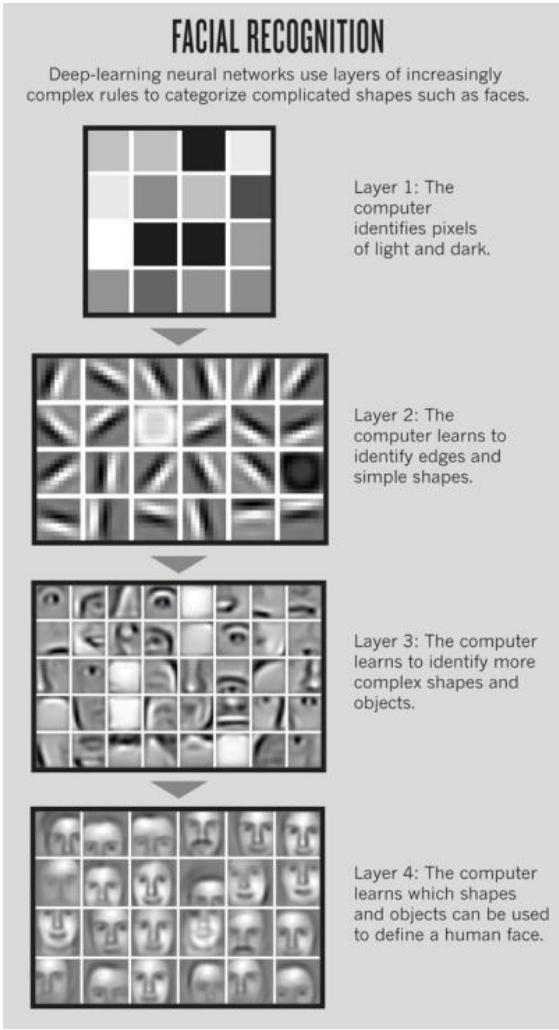
Artificial neuron





Deep Learning



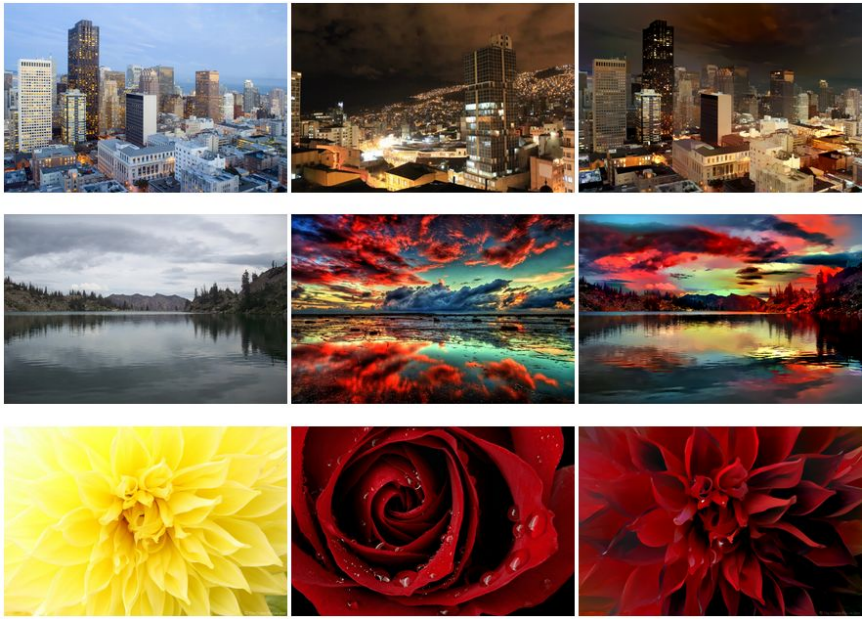


STYLESAGE

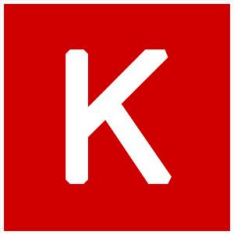
Input
(Image Data)

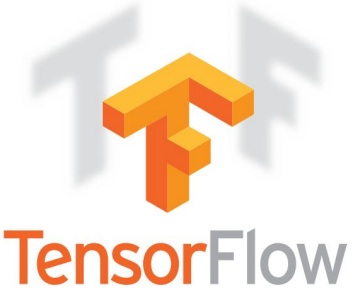
Output
(Probability Vector)

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



Deep Learning - Herramientas

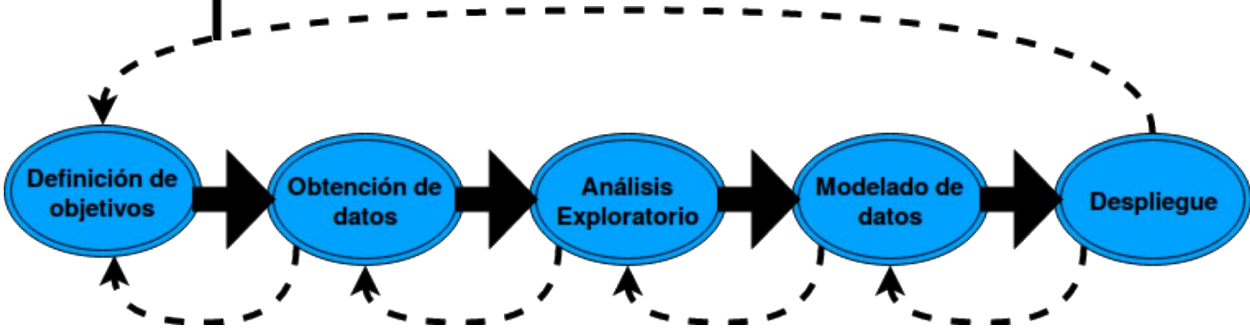





Big Data

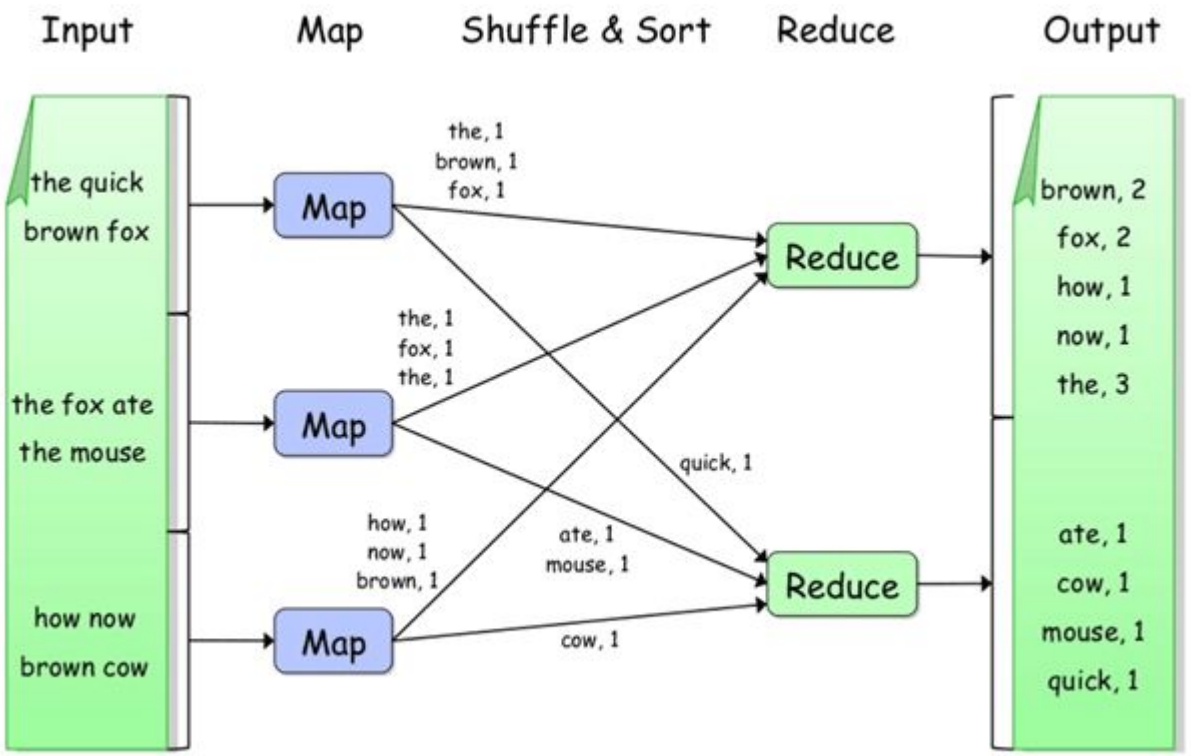
Big Data

Small Data	< 4GB	En memoria
Medium Data	< 1TB	En disco
Big Data	>= 1TB	-

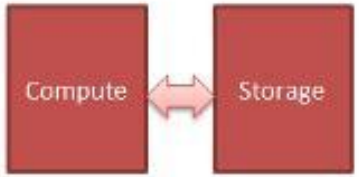


Big Data

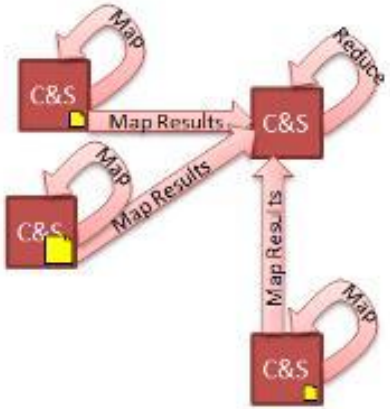
MapReduce



Big Data



HPC



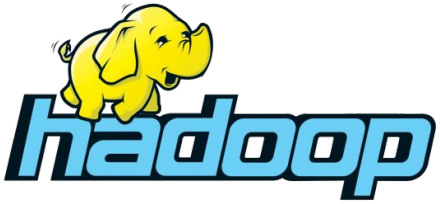
Hadoop or
Spark 1.1

Big Data - Herramientas

Batch
(Procesamiento offline, por lotes)

Streaming
(Procesamiento en tiempo real)

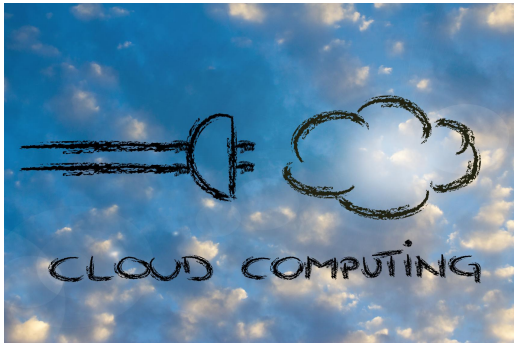
Batch + Streaming



Big Data - Herramientas (cont.)

Google BigQuery

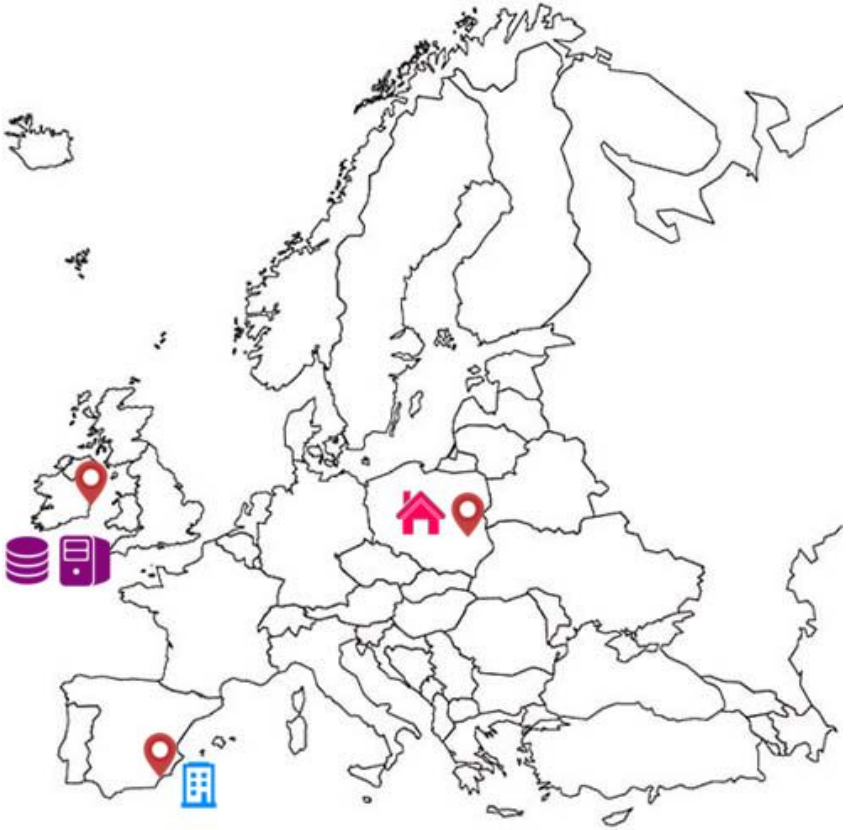
La nube

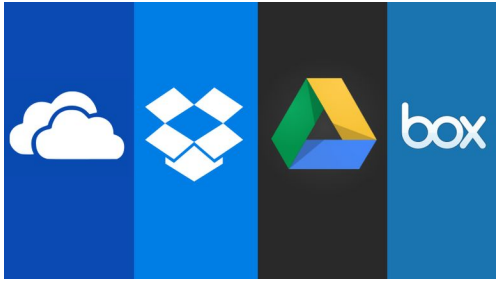

CLOUD COMPUTING


La "Nube"
Ventajas y Desventajas
ADrive


Cloud

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!

www.centic.es

hola@manugarri.com