BIG DATA for BUSINESS

2.5 Herramientas y software

Conecta Empleo

Contenido desarrollado por
Synergic Partners

Índice del módulo

2.5 HERRAMIENTAS Y SOFTWARE

- Lenguajes de Programación
- Herramientas de Control de Versiones
- Herramientas para Computación Distribuida
- Herramientas Deep Learning
- Herramientas Unificadas
- Herramientas de Visualización
- Comunidades

Lenguajes de Programación

Contenido desarrollado por **Synergic Partners**

Lenguajes de Programación

Existen múltiples lenguajes de programación, y sobre muchos pueden desarrollarse la ciencia de datos.

Nos centraremos en los dos lenguajes más usados y sobre los que se desarrolla de manera *open source* la mayoría de los procesos analíticos: **Python** y **R**.

Lenguajes de Programación

Python

Es un lenguaje multiparadigma con licencia de código abierto. Combina la programación orientada a objetos, la imperativa y la funcional.

Fue creado a finales de los ochenta por Guido van Rossum en el CWI (*Centrum Wiskunde & Informatica*) de los Países Bajos. El nombre viene de la afición de su creador por los Monty Python.

Este lenguaje es usado para sus analíticas por entidades tan importantes como Google o la NASA, entre otras muchas.

Lenguajes de Programación

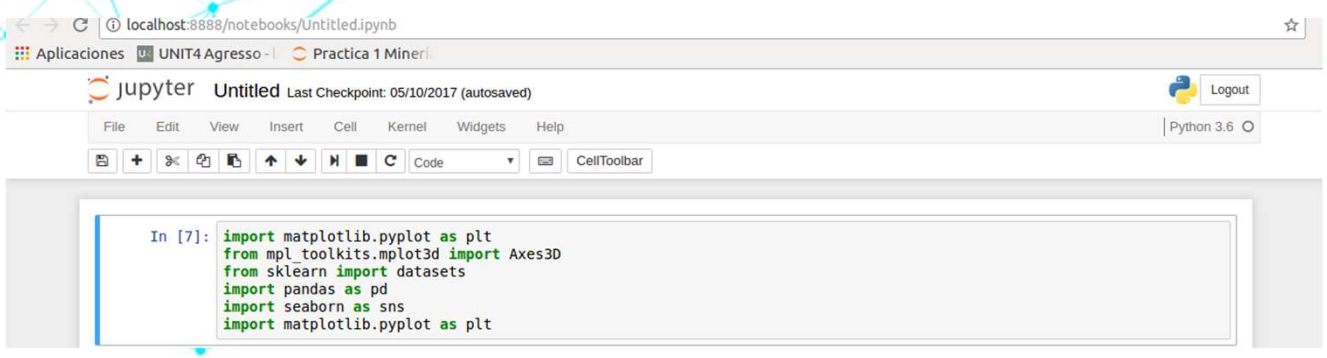
Python

Dónde descargarlo y encontrar la documentación: www.python.org

Podemos trabajar con Python desde un shell interactivo ejecutando scripts o a través de una interfaz (IDE).

Una de las interfaces más usadas es **Jupyter**, una aplicación web *open source* que permite crear y compartir documentos que contienen código, ecuaciones, visualizaciones o texto. Soporta más de 40 lenguajes de programación, entre ellos Python, además de permitir la integración con Spark.

http://jupyter.org/



Lenguajes de Programación

Python

Pandas

Es una biblioteca de Python para la gestión de datos a alto nivel a modo de tablas y de manera eficiente y sencilla.

```
In [94]: import pandas as pd

In [98]: dataset = pd.read_csv('iris.data.csv',names=['sepal length',
                                                      'sepal width',
                                                      'petal length',
                                                      'petal width'])

In [100]: dataset.head()
```

Out[100]:

	sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5.0	3.6	1.4	0.2	Iris-setosa

Lenguajes de Programación

Python

Scikit-Learn

Scikit-learn es una biblioteca de *machine learning* de software libre para Python. Cuenta con varios algoritmos de clasificación, regresión y clustering.

El proyecto scikit-learn comenzó como un proyecto Google Summer of Code de David Cournapeau.

http://scikit-learn.org/stable/index.html

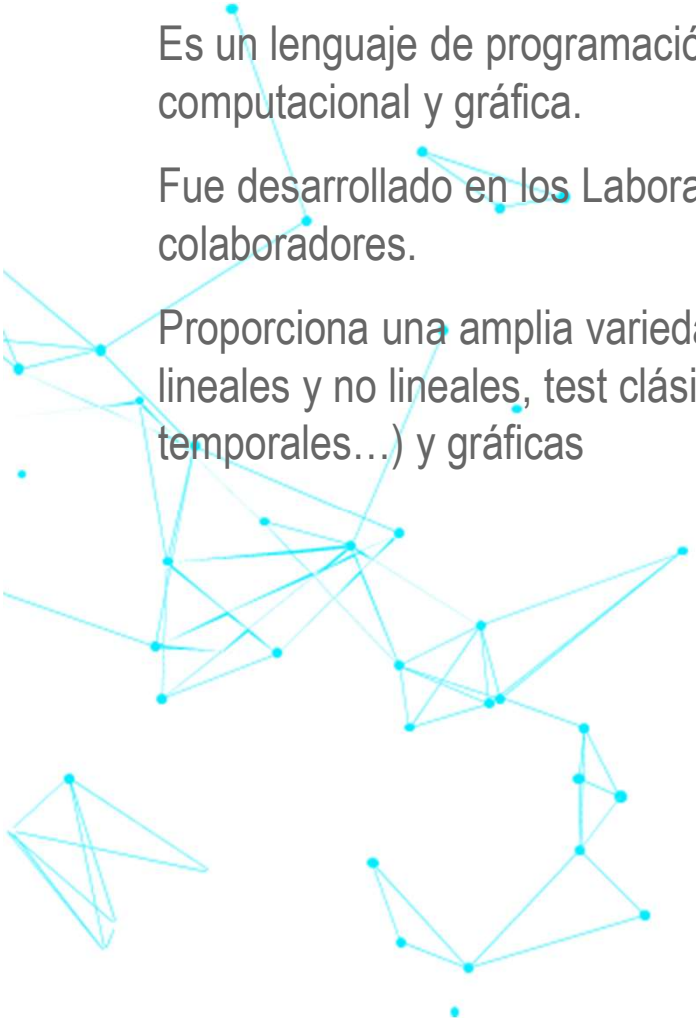
Lenguajes de Programación

R

Es un lenguaje de programación *open source* para la estadística computacional y gráfica.

Fue desarrollado en los Laboratorios Bell por John Chambers y sus colaboradores.

Proporciona una amplia variedad de técnicas estadísticas (modelos lineales y no lineales, test clásicos estadísticos,análisis de series temporales…) y gráficas



Lenguajes de Programación

R

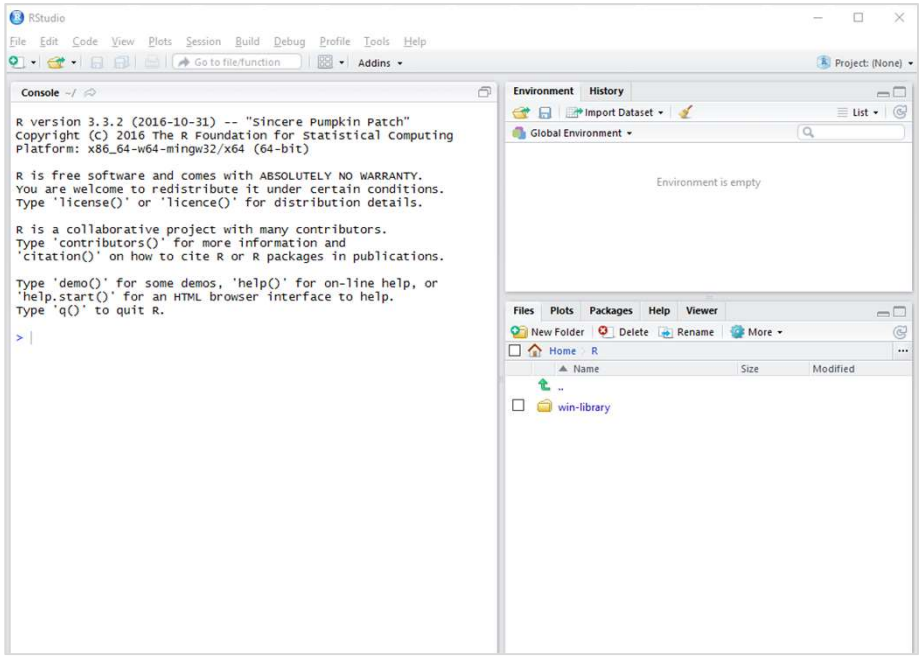
Podemos descargarlo y encontrar la documentación en:

https://cran.r-project.org/

Aunque R proporciona un intérprete de línea de comandos, existe un entorno de desarrollo que nos permite trabajar de manera más sencilla y visual, RStudio. Al igual que R es *open source*.

Podemos obtenerlo a través de:

www.rstudio.com



Lenguajes de Programación

R

R dispone de una librería machine learning similar Scikit-Learn denominada caret, creada por Max Kuhn

http://topepo.github.io/caret/index.html

Además, CRAN (The Comprehensive R Archive Network) gestiona y mantiene un repositorio de paquetes a las que se les exige unos requisitos de calidad y funcionalidad elevados para pertenecer a él.

Herramientas de Control de Versiones

Contenido desarrollado por **Synergic Partners**

Herramientas de control de versiones

> El control de versiones es un sistema que registra los cambios realizados sobre un archivo o conjunto de archivos a lo largo del tiempo, de modo que puedas recuperar versiones específicas más adelante.

Herramientas de control de versiones

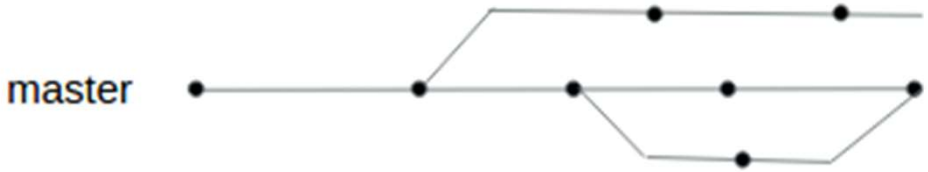
Git

Sistema de control de versiones distribuido *open source* desarrollado por los creadores de Linux. Almacena todo en repositorios locales de nuestro PC o en remoto y opera desde línea de comandos.

Todos los usuarios descargan una copia fiel y exacta del repositorio

Existe una rama master, durante el desarrollo se crean ramas secundarias que se fusionan a la master al finalizar.

https://git-scm.com/



Herramientas de control de versiones

Git

Son repositorios de Git basados en web de licencia de código abierto. Organizaciones como IBM, la NASA o el CERN los utilizan.

La mayor diferencia entre ambos es que GitHub no permite disponer de repositorios privados de forma gratuita mientras que GitLab si.

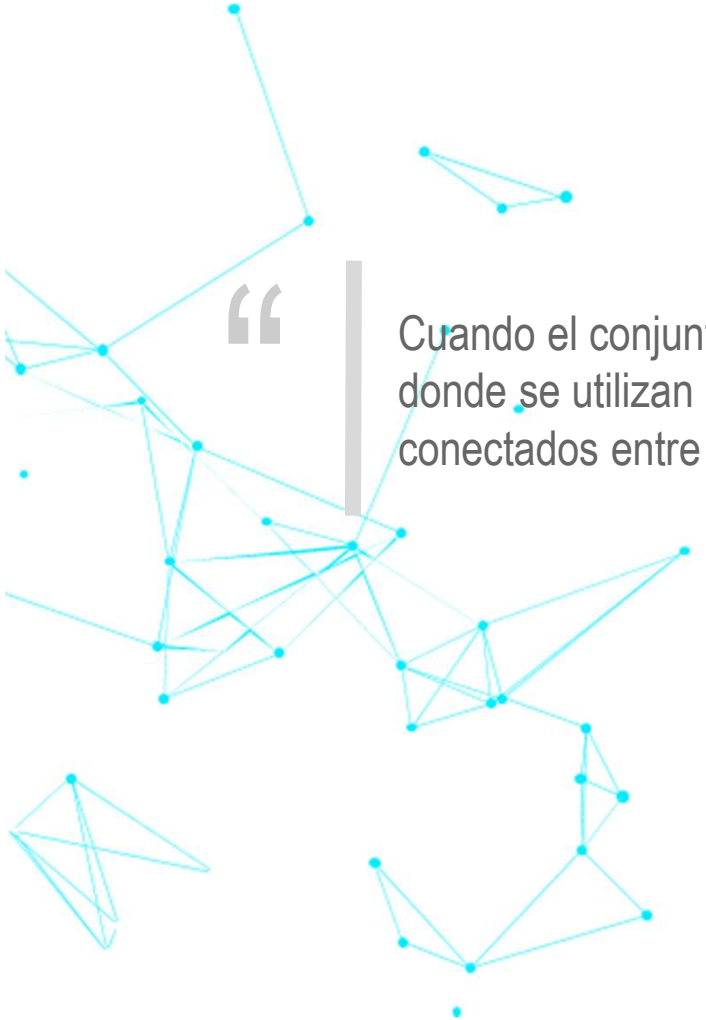
https://github.com/

https://about.gitlab.com/

Herramientas para Computación Distribuida

Contenido desarrollado por **Synergic Partners**

Herramientas para computación distribuida

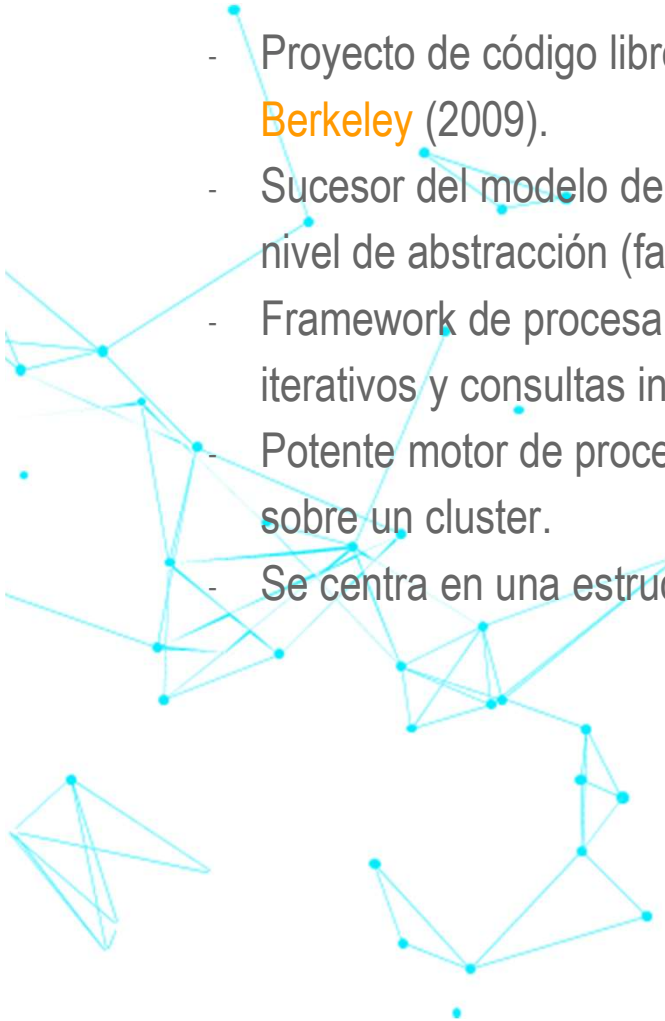


> Cuando el conjunto de datos es masivo se recurre a la computación distribuida donde se utilizan un gran número de ordenadores organizados en clusters, conectados entre sí como una red de comunicaciones.

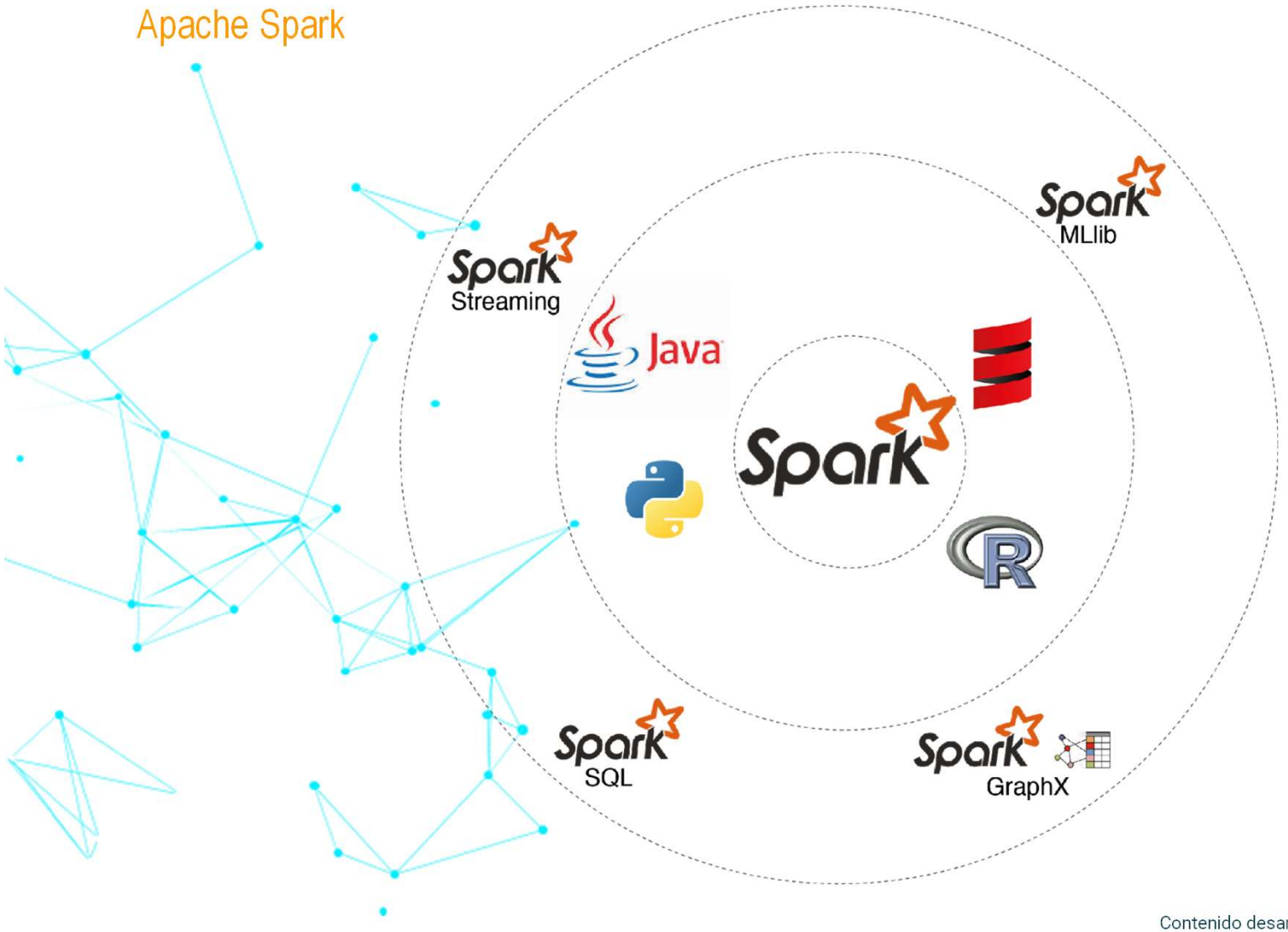
Herramientas para computación distribuida

Apache Spark

- Proyecto de código libre (*open-source*) creado en el AMPLAB de la Universidad de Berkeley (2009).
- Sucesor del modelo de programación MapReduce. Más rápido (x100), y con un mayor nivel de abstracción (facilidad de desarrollo)
- Framework de procesamiento unificado: procesamiento batch y streaming, algoritmos iterativos y consultas interactivas.
- Potente motor de procesamiento de datos masivos en memoria (in-memory processing) sobre un cluster.
- Se centra en una estructura de datos denominada RDD (*resilient distributed dataset).*



Herramientas para computación distribuida



Herramientas para computación distribuida

SparkML

Paquete introducido a partir de la versión 1.2 de Spark con algoritmos de *machine learning*. Incluye algoritmos de clasificación, regresión, clustering, recomendación, entre otras utilidades.

Podemos encontrar dos paquetes, Spark ML y Spark MLlib, el primero está construido sobre *data frames* para *pipelines* de ML, mientras que el segundo está construido sobre RDDs.

Encontramos toda la documentación en el siguiente enlace:

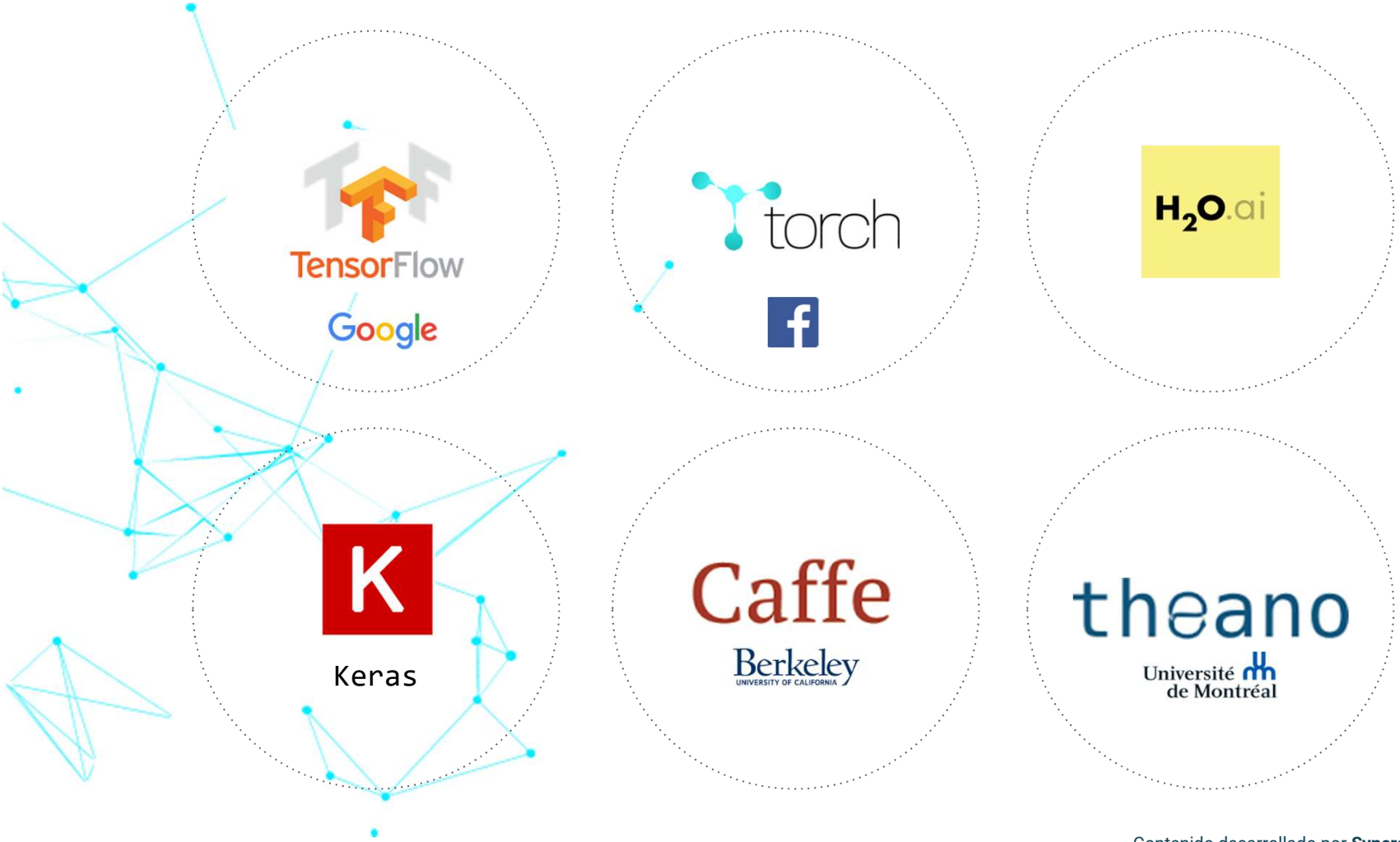
https://spark.apache.org/docs/latest/ml-guide.html

Herramientas Deep Learning

Contenido desarrollado por **Synergic Partners**

Herramientas de Deep Learning

Las siguientes herramientas son las utilizadas en la programación de *Deep Learning*

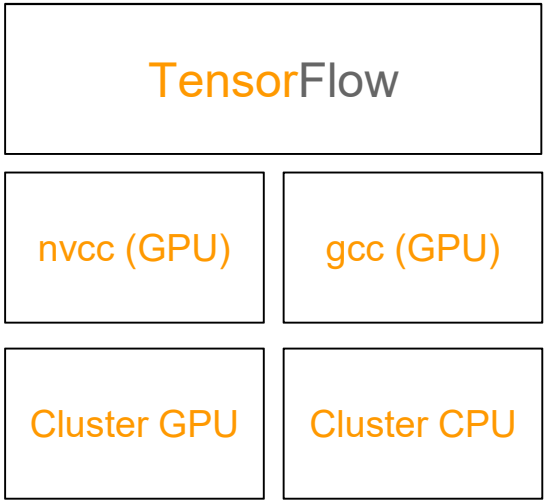


Herramientas de Deep Learning

Tensorflow

Es una biblioteca para realizar modelos deep learning.

- Desarrollado por Google y liberada (open source) recientemente.
- La mayor ventaja es que permite la abstracción de la aplicación de funciones a tensores, realizando la computación de forma automática.
  - Un tensor es un tipo de array multidimensional
- Está pensada para conectar de forma sencilla con dataflow que es una herramienta de Google para ETL de datos.



Herramientas de Deep Learning

H2O

Es una plataforma de código abierto, distribuida, rápida y escalable para machine learning y análisis predictivo.

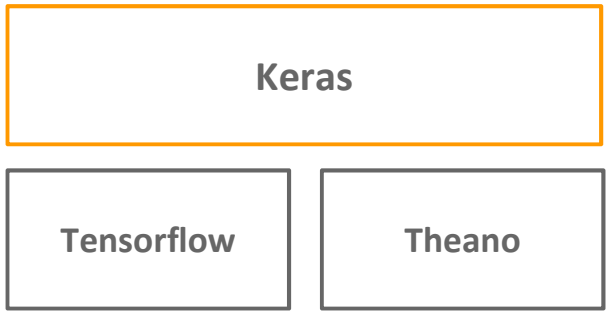
- Agrupa un conjunto de modelos pre-construidos configurables para abstraer al usuario de los detalles de implementación.
- Tiene un manejo muy eficiente de la memoria, ya que usa compresión en memoria, lo que le permite manejar grandes datasets incluso en clusters pequeños.
- Ofrece APIs para entornos de programación R, Python, Scala, Java, etc.
- También cuenta con una interfaz gráfica web que facilita la implementación y análisis de los modelos.
- Incluye algoritmos de clasificación, clustering, modelos lineales generales, análisis estadístico, herramientas de optimización, preprocesamiento de datos, y deep learning.
- Ofrece integración con R y RStudio, así como también con Spark y SparkML mediante Sparkling Water.

Herramientas de Deep Learning

Keras

Es una capa de abstracción a Tensorflow y Theano.

- Permite computar en GPU, pero no en cluster CPU.
- Tiene una programación sintetizada frente a Tensorflow y Theano, por lo tanto si no es necesario computar en un cluster GPU Keras es un buen primer paso.
- Keras es capaz de incorporar código de Tensorflow y Theano
- Permite utilizar el **Tensorboard**.



Herramientas Unificadas

Contenido desarrollado por **Synergic Partners**

Herramientas Unificadas

Existen un conjunto de herramientas que engloban los principales algoritmos de *machine learning* y nos permiten importar, limpiar datos y realizar analíticas sobre los mismos. Estos *softwares* disponen de una interfaz que permite al usuario sin conocimientos de programación, hacer análisis y extraer insights a partir de un conjunto de datos crudos.

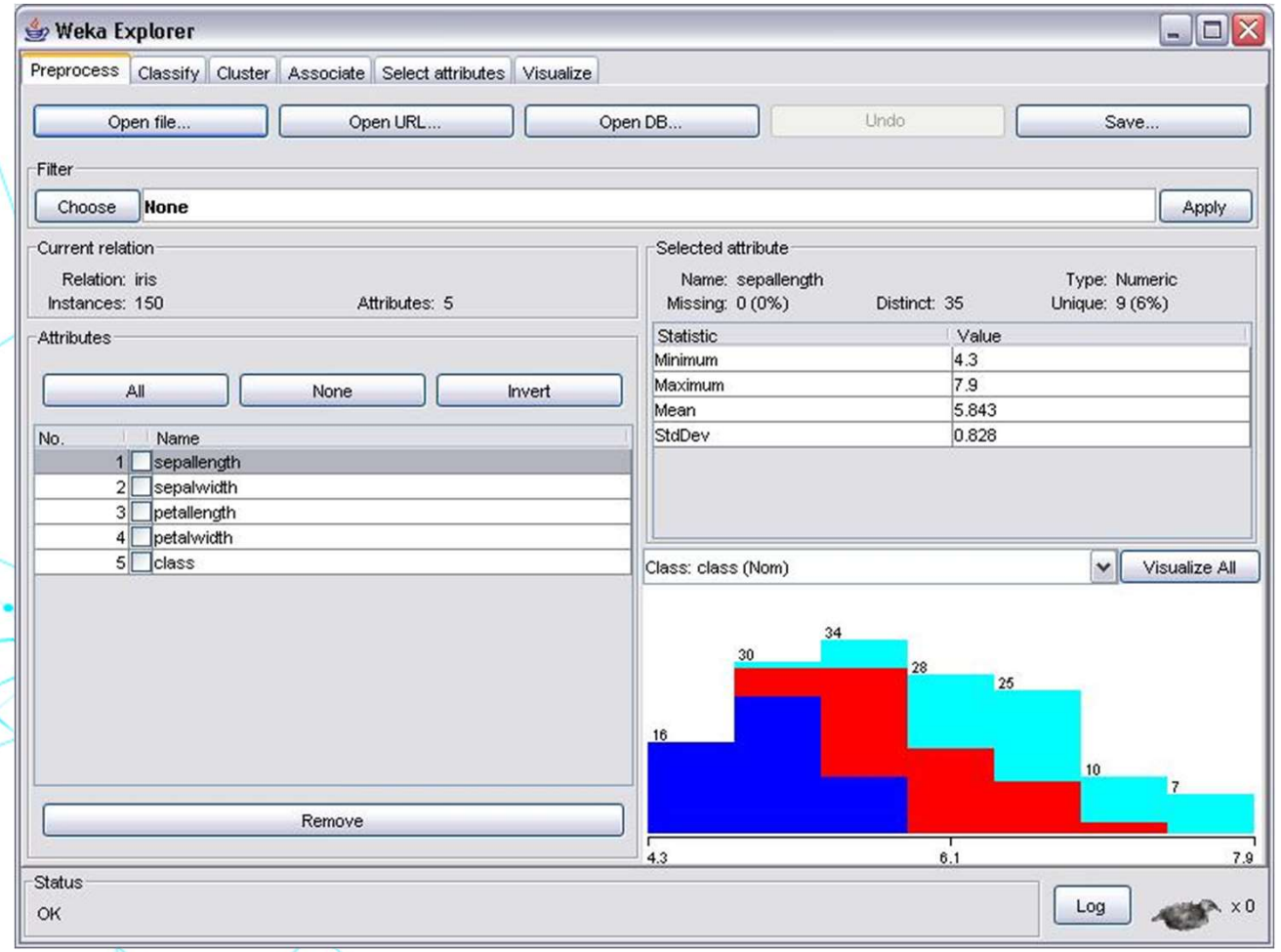
Herramientas Unificadas

Weka

Weka es un *software* libre escrito en Java y desarrollado en la universidad de Waikato, Nueva Zelanda. Contiene herramientas para pre-procesamiento de datos, clasificación, regresión, clustering, reglas de asociación y visualización.
Los algoritmos pueden aplicarse directamente sobre el dataset o ejecutarse desde línea de comandos.
Podemos obtenerlo:

http://www.cs.waikato.ac.nz/ml/weka/

Herramientas Unificadas

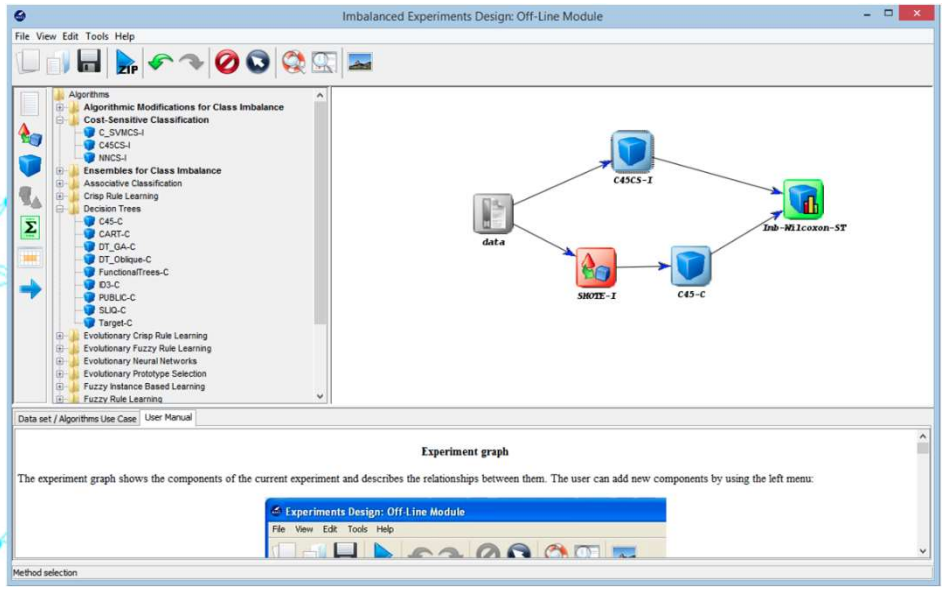


Herramientas Unificadas

KEEL

Es una herramienta escrita en Java similar a Weka y desarrollada por la UGR. Dispone de algoritmos de preprocesamiento, clasificación y regresión, así como de distintos métodos de validación de modelos. Disponen además de un repositorio propio de conjuntos de datos de libre acceso con el fin de disponer de un benchmark unificado.

http://www.keel.es/



KEEL-dataset
Data set repository

Herramientas Unificadas

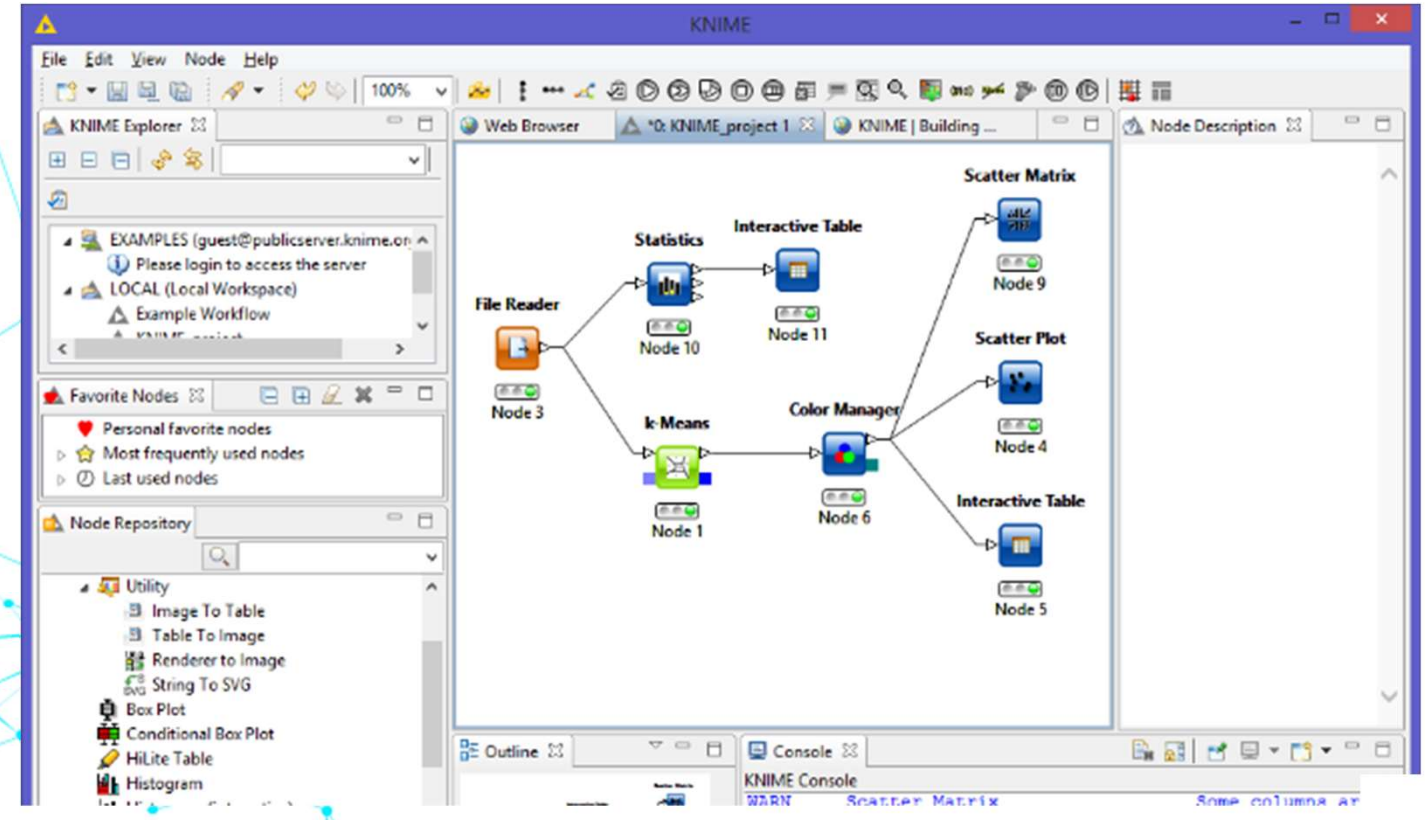
Knime

Knime (*Konstanz Information Miner*) es una plataforma de minería de datos , construida sobre Eclipse, que permite el desarrollo de modelos en un entorno visual.
Fue desarrollada por el departamento de bioinformática de la Universidad de Constanza, Alemania.
Es una herramienta gráfica que dispone de una serie de nodos, los cuales contienen diferentes algoritmos. Estos nodos se conectan mediante flechas que representan el flujo de datos.

https://www.knime.org/

Las funciones de Weka están implementadas dentro de Knime.

Herramientas Unificadas

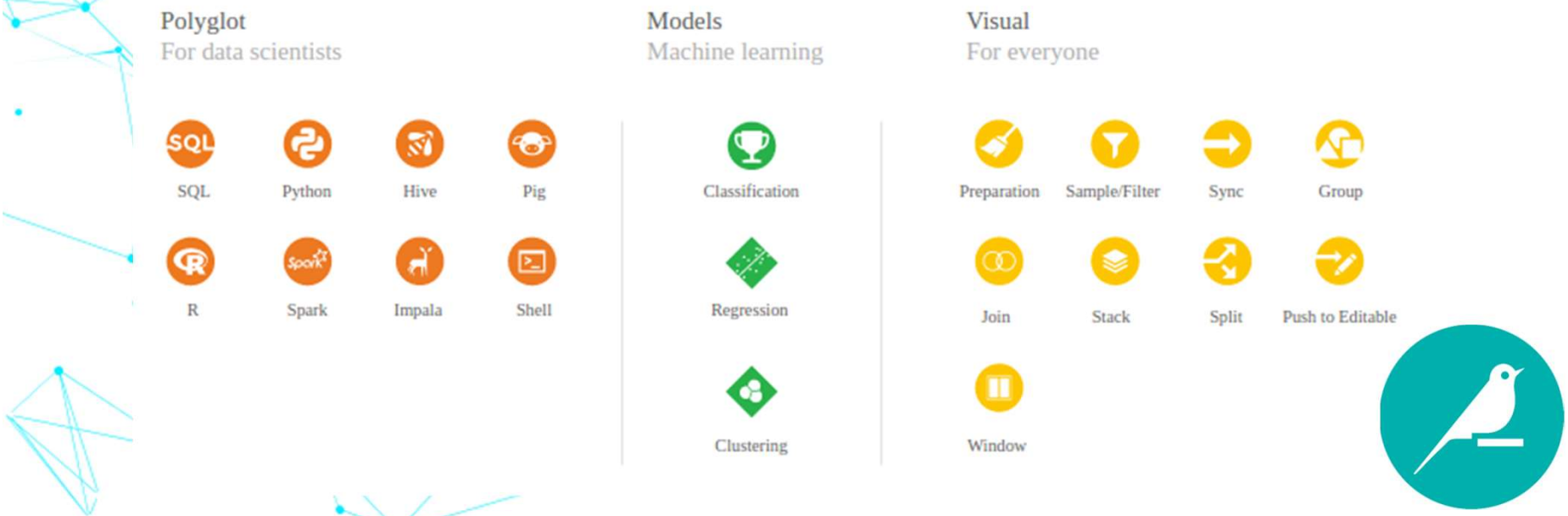


Herramientas Unificadas

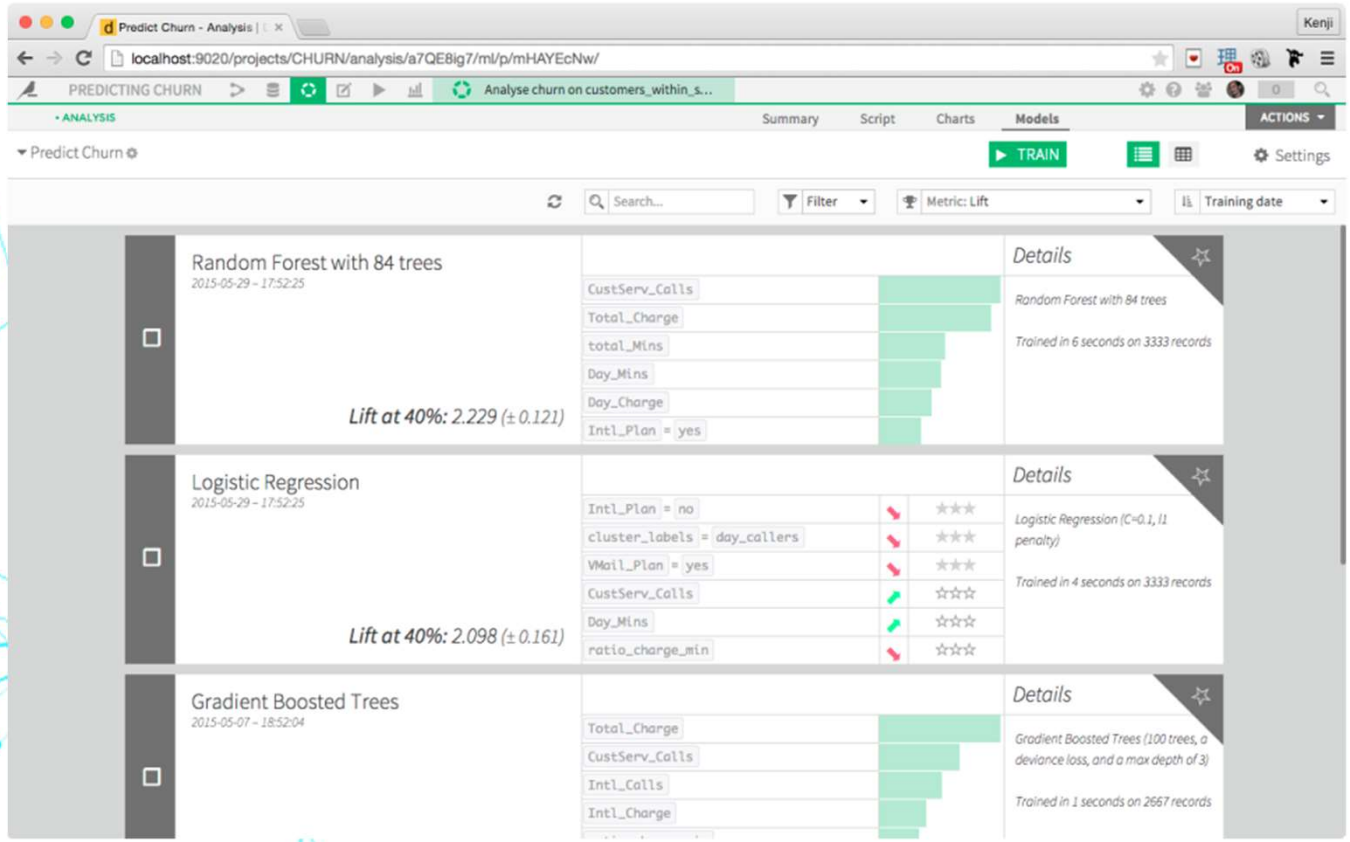
Dataiku

Es un software colaborativo para la ciencia de datos y big data.
Permite el uso de notebooks de Python, R, SQL, Spark, entre otros. Tiene incorporados modelos de clasificación, regresión y clustering, además de herramientas de preprocesado de datos.

https://www.dataiku.com/



Herramientas Unificadas



Herramientas Unificadas

Rapidminer

Herramienta *open source* para el análisis y minería de datos que permite el desarrollo de procesos de analítica mediante el encadenamiento de operadores a través de un entorno gráfico.
Está desarrollada en Java y permite un flujo de trabajo desde el preprocesado, pasando por el modelado para obtener valores de negocio.

https://rapidminer.com/

Podemos visualizar una demo a través del siguiente enlace:
https://rapidminer.com/resource/fraud-demo/

Herramientas de Visualización

Contenido desarrollado por **Synergic Partners**

Herramientas de Visualización

> En los procesos de minería de datos, uno de los pasos más importantes es la visualización. Esta puede ser tanto a nivel exploratorio de datos como para extraer los insights de los resultados.

Herramientas de Visualización

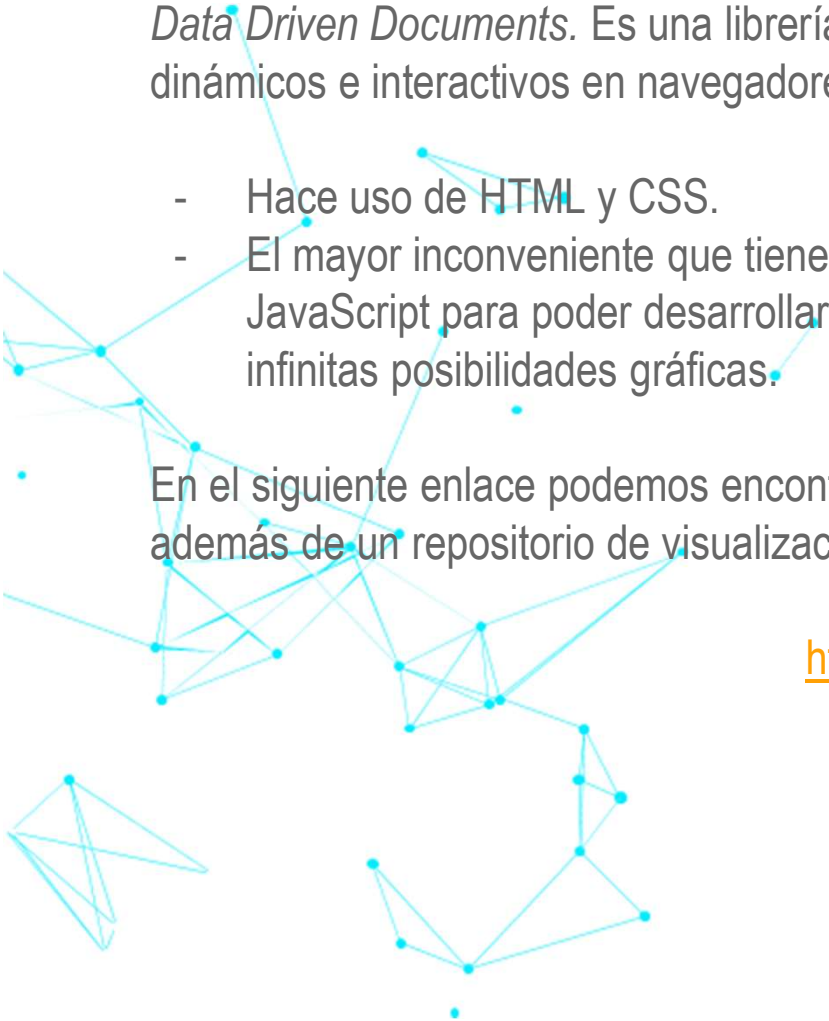
D3

Data Driven Documents. Es una librería de JavaScript para producir infogramas dinámicos e interactivos en navegadores web.

- Hace uso de HTML y CSS.
- El mayor inconveniente que tiene es que precisa conocimientos de programación en JavaScript para poder desarrollar las visualizaciones. Esto lleva, sin embargo, a infinitas posibilidades gráficas.

En el siguiente enlace podemos encontrar toda la documentación sobre esta librería además de un repositorio de visualizaciones:

https://d3js.org/

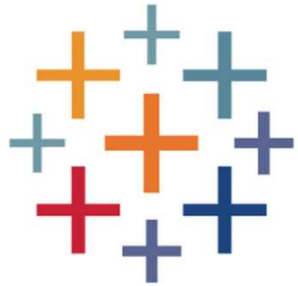


Herramientas de Visualización

Tableau

- Software de análisis de datos mediante visualización interactiva.
- Dispone de una interfaz que facilita su uso para cualquier tipo de público.
- Tableau dispone de tres productos: un software de escritorio, un servidor para la difusión de dashboards y una plataforma hospedada en la nube que permite publicar y dar lugar a visualizaciones colaborativas.

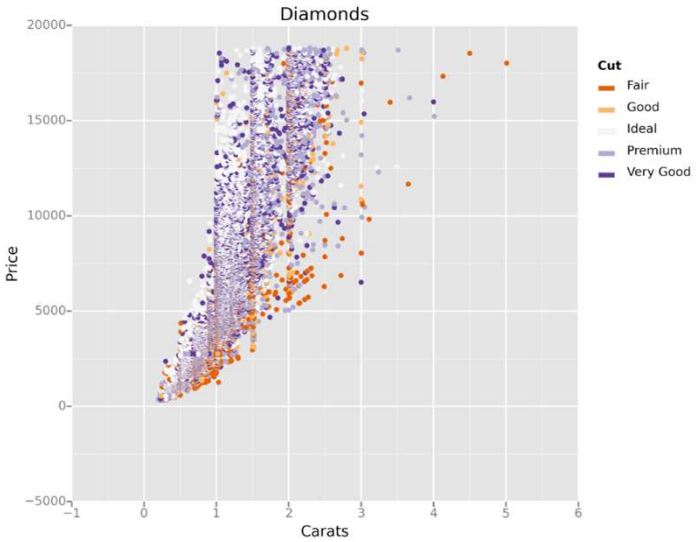
https://www.tableau.com/es-es



Herramientas de Visualización

ggplot

- Librería de R para la visualización de datos creada por Hadley Wickham.
- Permite crear gráficos que representan datos numéricos y categóricos univariados y multivariados de una manera directa. Puede agrupar datos y representarlos por color, símbolo, tamaño y transparencia.
- Precisa de conocimientos de programación en R.



Herramientas de Visualización

Shiny

- Es un paquete de código abierto que proporciona un marco web para construir aplicaciones web usando R.
- Permite interactuar con datos sin tener que manipular el código.
- Podemos obtener más información en:

https://shiny.rstudio.com/

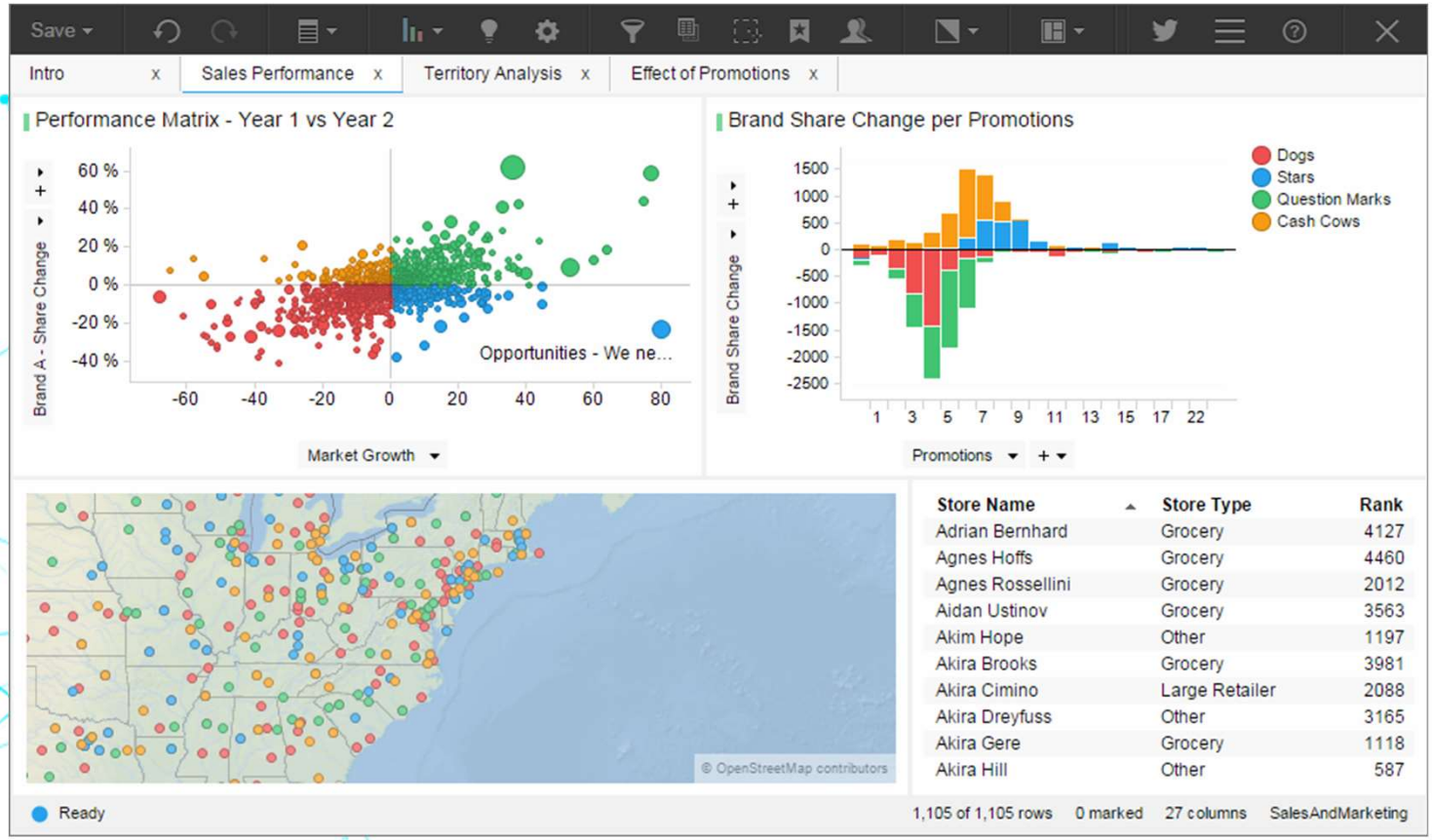
Herramientas de Visualización

TIBCO Spotfire

- Es un software de análisis que permite extraer información de forma rápida.
- Es una herramienta similar a Tableau que permite la creación de dashboards interactivos de una manera sencilla sin la necesidad de manipular código.
- Spotfire fue creado en los 90 pero no se dio realmente a conocer hasta 2007 cuando fue adquirido por la compañía TIBCO.
- Al igual que Tableau, no tiene licencia libre.

http://spotfire.tibco.com

Herramientas de Visualización



Spotfire®

Herramientas de Visualización

Kibana

Herramienta *open source* perteneciente a Elastic, que permite visualizar y explorar datos que se encuentran indexados en ElasticSearch.

Tiene una sencilla interfaz basada en navegador web que permite crear y compartir rápidamente dashboards dinámicos que muestran cambios sobre queries ejecutadas en Elasticsearch en tiempo real.

https://www.elastic.co/products/kibana

Comunidades

Contenido desarrollado por **Synergic Partners**

Comunidades

> El campo de la ciencia de datos está en continua evolución por lo que el uso de comunidades y blogs web donde compartir conocimientos y novedades suponen una herramienta muy útil para un *Data Scientist*.

Comunidades

Kaggle

Plataforma fundada en 2010 para competiciones analíticas donde las compañías e investigadores publican sus datos y problemas analíticos para que gente de todo el mundo ayude a solucionarlos y obtener los mejores resultados.
Este recurso es usado por empresas para reclutar posibles *data scientist* para las mismas.

Netflix publica aquí una competición llamada **Netflix Prize** donde el objetivo es mejorar el resultado de un algoritmo de filtrado colaborativo para predecir los ratios por usuario de películas.

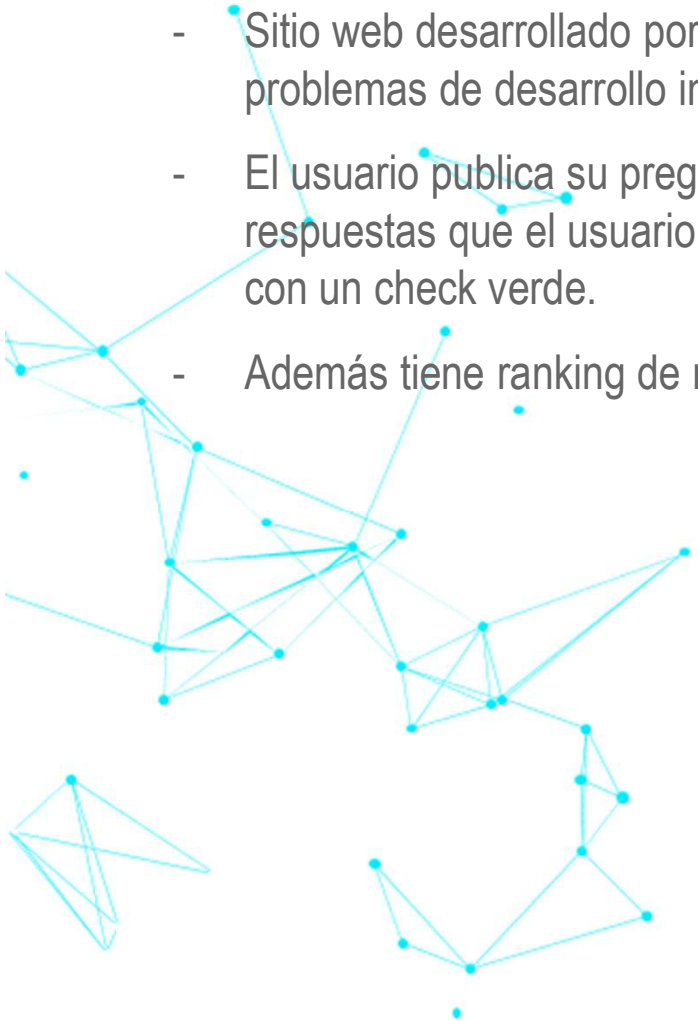
https://www.kaggle.com/

Comunidades

Stack Overflow

- Sitio web desarrollado por Jeff Attwood para encontrar soluciones a diferentes problemas de desarrollo informático.
- El usuario publica su pregunta y el resto de usuarios dan respuesta. Aquellas respuestas que el usuario verifica que dan resultado a su problema son marcadas con un check verde.
- Además tiene ranking de reputación, para preguntas y respuestas.

http://stackoverflow.com/



Comunidades

Cross Validated

- Sitio web también de formato pregunta-respuesta enfocado a personas interesadas en estadística, *machine learning*, minería y análisis de datos y visualización.
- Es similar a Stack Overflow, solo que este está enfocado a desarrolladores y Cross Validated focaliza más sobre los temas descritos anteriormente.

https://stats.stackexchange.com/

Comunidades

Formación online

Existen además multitud de plataformas para la formación online, a continuación se enumeran algunas de ellas:

- **Code Academy:** Cursos de programación gratuitos. https://www.codecademy.com/es
- **Big Data University:** https://bigdatauniversity.com/
- **Data Camp:** Python, R y Data Science. https://www.datacamp.com/
- **Coursera:** Cursos online de destacadas Universidades de todo el mundo. https://es.coursera.org/
- **Udacity:** Organización educativa que ofrece cursos abiertos. https://www.udacity.com/
- **Edx:** Plataforma de cursos gratuitos fundada por el Instituto Tecnológico de Massachusetts y Harvard. https://www.edx.org/

Contenido desarrollado por **Synergic Partners**