Hacia una visión unificada de Data Mining, Data Science, Analytics y Big Data

Esta presentación describe, bajo un marco común, los conceptos fundamentales de Data Science, Analytics y Big Data y establece su similitudes y diferencias.

Descargue la última versión de este documento de:
https://github.com/jdvelasq/data-science-docs/blob/master/ds-analytics-bigdata.pdf

JUAN DAVID VELÁSQUEZ HENAO, MSc, PhD
Profesor Titular
Departamento de Ciencias de la Computación y la Decisión
Facultad de Minas
Universidad Nacional de Colombia, Sede Medellín

jdvelasq@unal.edu.co
@jdvelasquezh
https://github.com/jdvelasq
https://goo.gl/prkjAq
https://goo.gl/vXH8jy

RDBMS

Bases de datos relacionales.

SQL

Desarrollado por Oracle.

Business Intelligence

Análisis de información en RDBMS para toma de decisiones de negocio.

Data Mining

Proceso de descubrimiento de patrones y tendencias útiles en grandes conjuntos de datos.

Explosión de la Web

De uso científico y gubernamental a comercial y de consumo

Web Services

SaaS

Infraestructura, mantenimiento, operación y soporte dado por un tercero y accesado por internet

PaaS

Plataformas de desarrollo, Middleware, sistemas operativos

IaaS

Pago por uso de memoria, disco, procesador, red.

Bases de Datos NoSQL

Open Data Science

Modern Analytics

Explosión IoT

1970 1976 1985 1989 1995 1996 1999 2002 2005 2009 2011 2014 2016

Internet of Things (IoT)

Red de dispositivos físicos con sensores y conectividad que les permiten recolectar e intercambiar datos.

Big Data Analytics

Apache Mahout

Data Warehousing OLAP

Almacenes de datos que garantizan la permanencia de la información

Business Intelligence 2.0

Plataformas de BI

Data Science

Área relacionada con los procesos y sistemas para la extracción de conocimiento de datos almacenados electrónicamente

Traditional Analytics

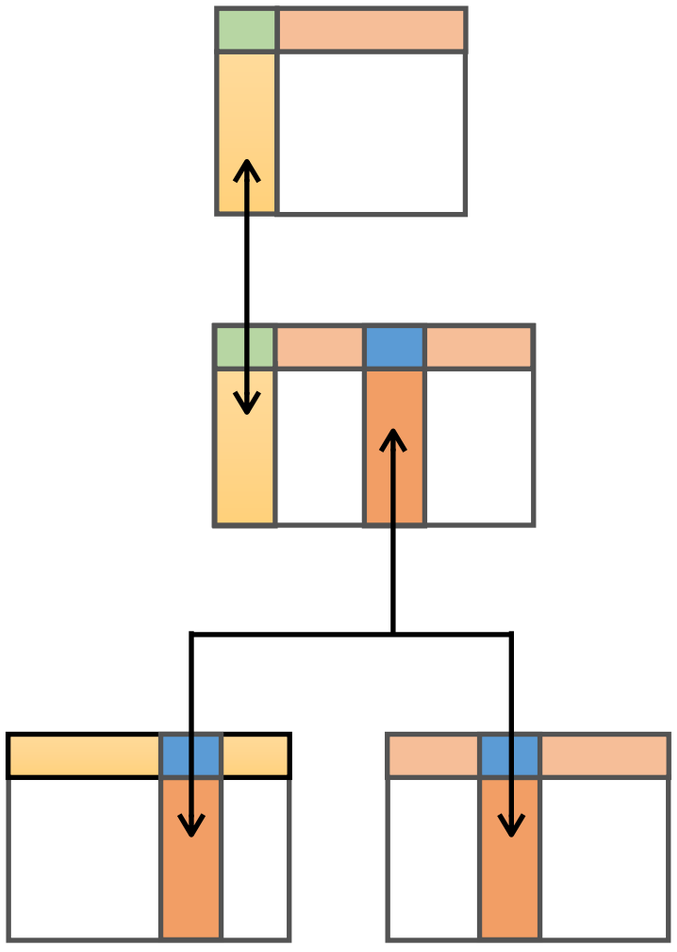
Proceso científico de transformación de datos en conocimiento para mejorar el proceso de toma de decisiones de negocio

Hadoop / MapReduce

Primera implementación de un sistema Big Data

Masificación Big Data

RDBMS — Relational Database Management System



Componentes

- Esquemas
- Tablas
- Consultas
- Reportes
- Vistas
- Otros elementos

Esquemas

- Definición de las tablas.
- Tipos de datos.
- Relaciones (uno a uno, uno a muchos, muchos a muchos).
- Campos clave.
- Reglas de negocio.

Funciones

- Definición.
- Manipulación (inserción, borrado, actualización, …)
- Seguridad e integridad.
- Recuperación y restauración.

Principales RDBMDS

- Oracle
- PostgressSQL
- Microsoft SQL server
- MySQL
- Microsfot Access
- DB2
- MariaDB
- Informix
- …

SQL — Structured Query Language

Data Definition Language (DDL)
- Create
- Alter
- Truncate
- Rename
- Drop

Data Manipulation Language (DML)
- Insert
- Update
- Delete
- Select

Data Control Language (DCL)
- Grant
- Revoke

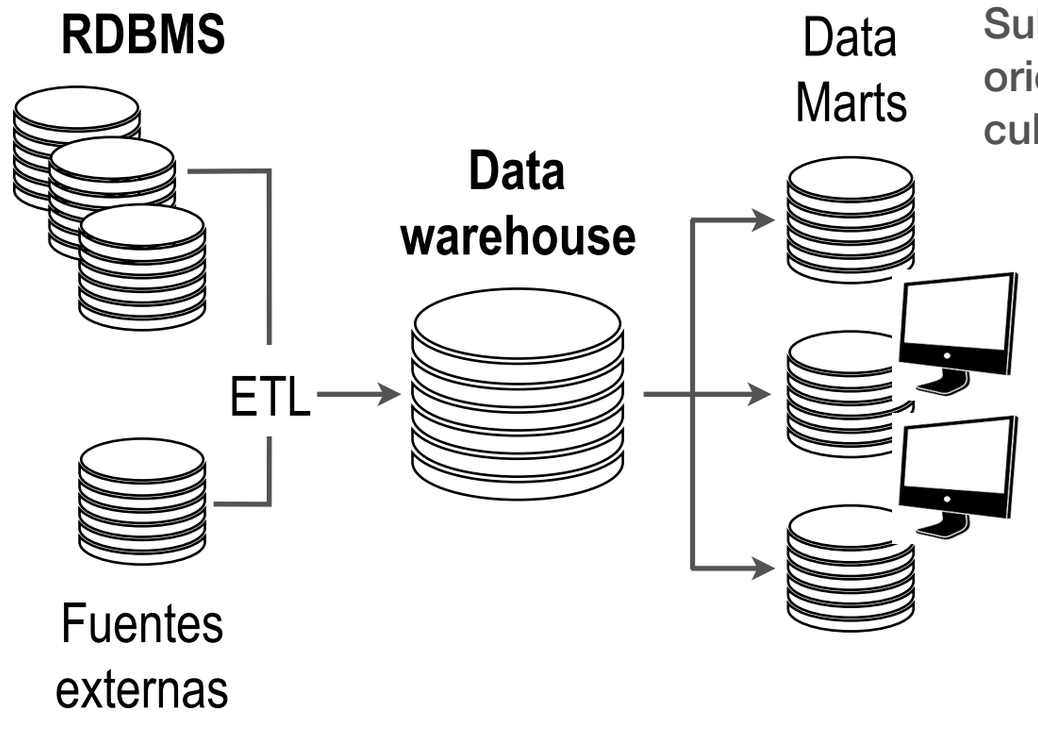
Transactions Control Language (TCL)
- Commit
- Rollback
- Savepoint

```
CREATE TABLE ‘CUSTOMERS';

ALTER TABLE 'ALUMNOS' ADD EDAD INT UNSIGNED;

DROP TABLE ‘ALUMNOS';

TRUNCATE TABLE ‘NOMBRE_TABLA';

SELECT * FROM Coches ORDER BY marca, modelo;

SELECT DISTINCT marca, modelo FROM coches;

INSERT INTO agenda_telefonica (nombre, numero)
VALUES ('Roberto Jeldrez', 4886850);

INSERT INTO phone_book2 ( [name], [phoneNumber] )
SELECT [name], [phoneNumber]
FROM phone_book
WHERE name IN ('John Doe', 'Peter Doe’)

DELETE FROM tabla WHERE columna1 = ‘valor1’;
```

Data Warehouse

ETL

- Extraction
- Transfomation
- Load



Data Mart

Subconjunto de datos de un Data Warehouse orientado a la la consulta. Es implementado usando cubos OLAP

Enterprise Resource Planning (ERP)
Executive information systems (EIS)

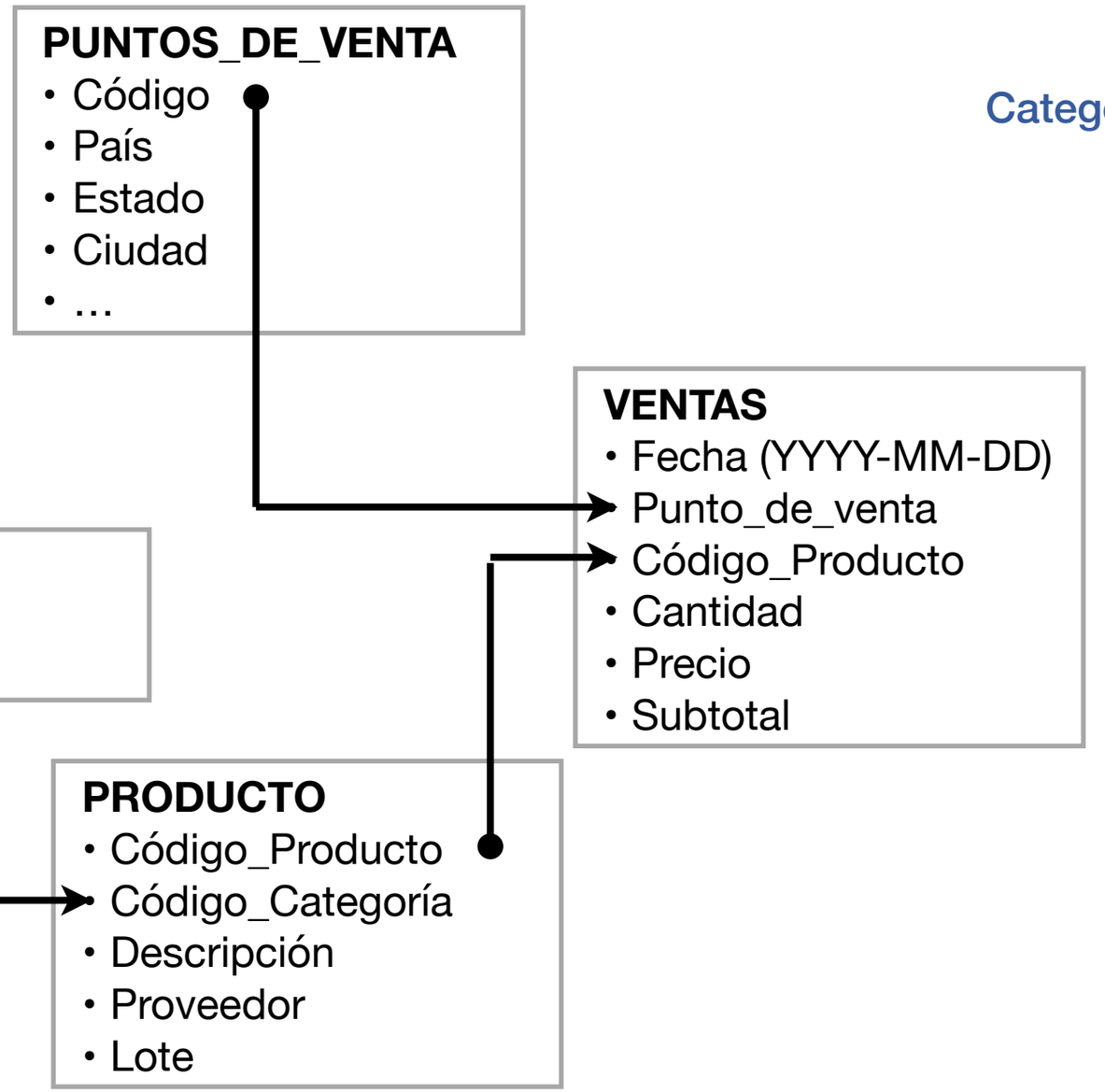
Data Warehouse

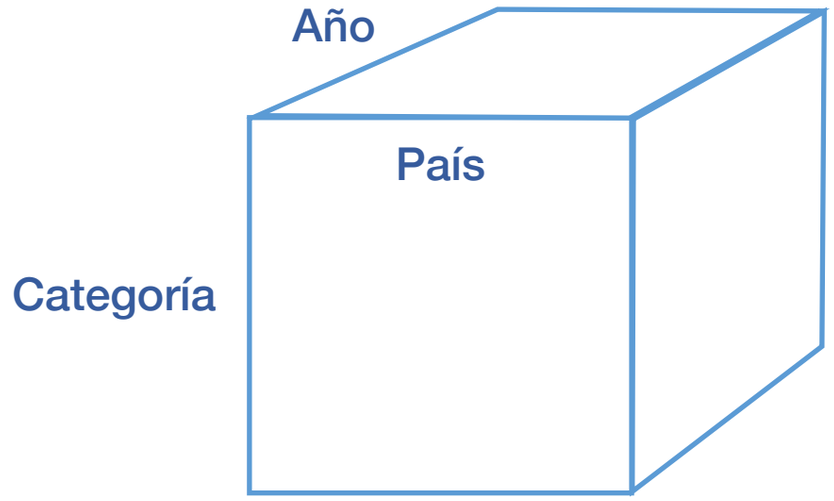
Bodegas de datos / Procesamiento analítico en línea

- Estructurado
- Orientado a temas
- Integrado (consistencia de los datos)
- No volátil (permanencia de la información, no se modifica ni se elimina)
- Variable en el tiempo
- Orientado al análisis y la divulgación de la información

OLAP — On-line Analytical Processing

Modelo para agilizar la consulta de grandes volúmenes de datos, mediante el almacenamiento de los datos en vectores multidimensionales





Cubo OLAP

Arreglo multidimensional para consultar información

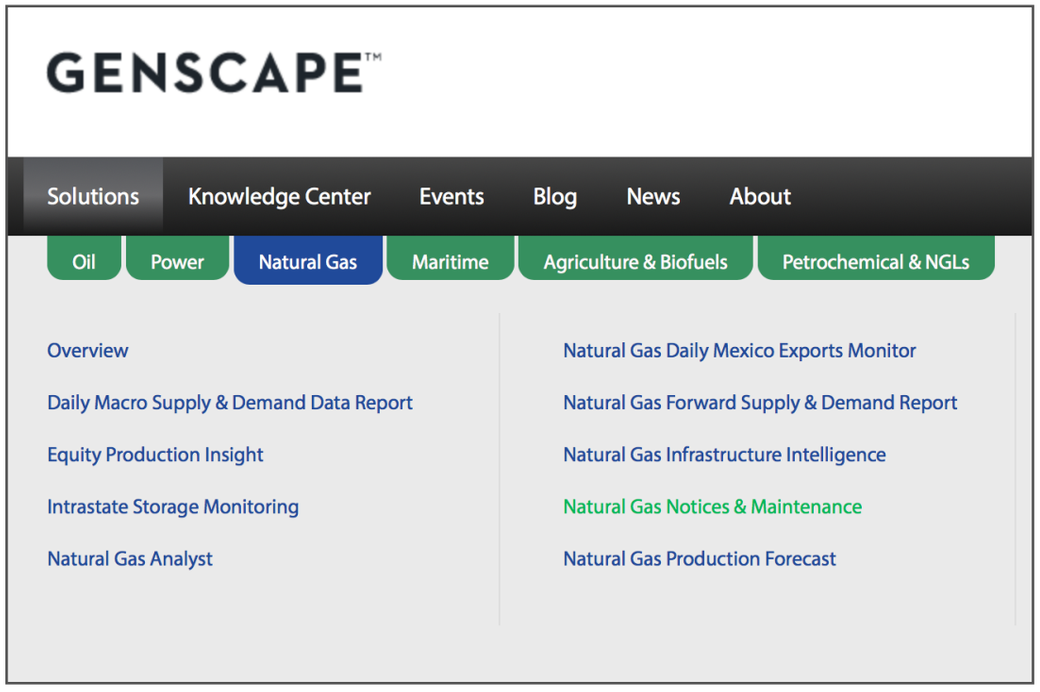
Business Intelligence

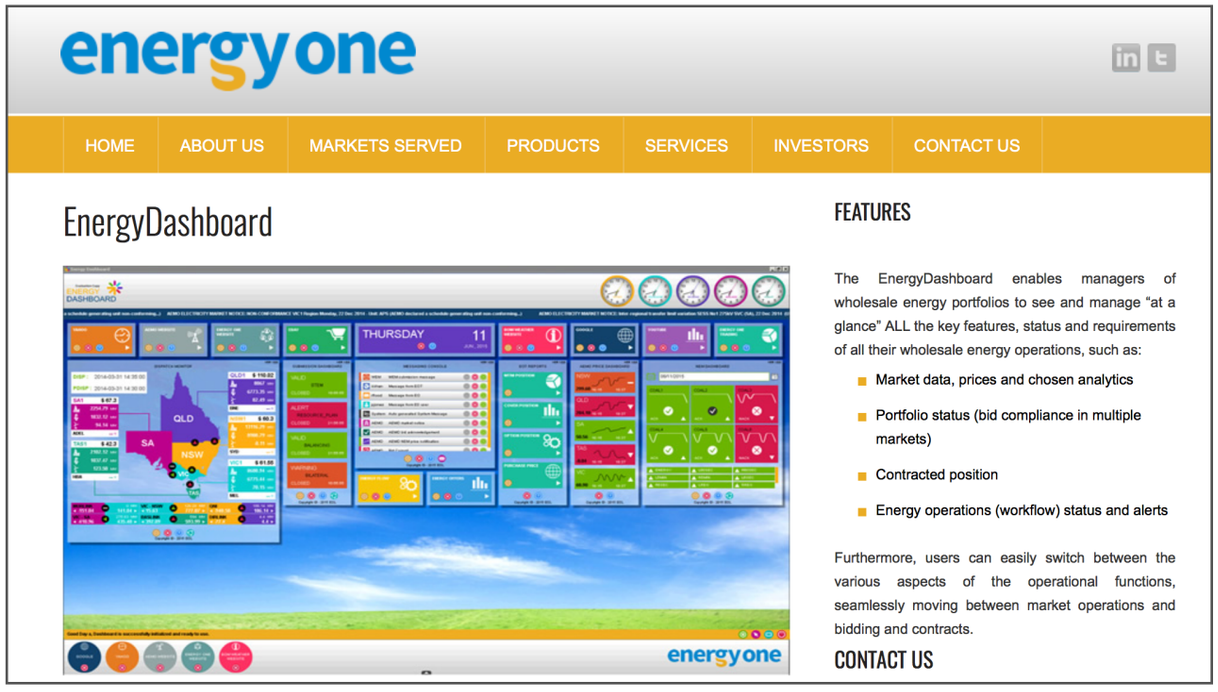
Software y servicios para analizar conjuntos de datos transaccionales y generar conocimiento para la toma de decisiones tácticas y estratégicas en organizaciones.

La BI se considera como parte de la Analítica Descriptiva (qué ocurrió en el pasado).

Los hallazgos dan información detallada del negocio y son presentados como:

- Reportes
- Cuadros de mando
- Gráficos
- Mapas



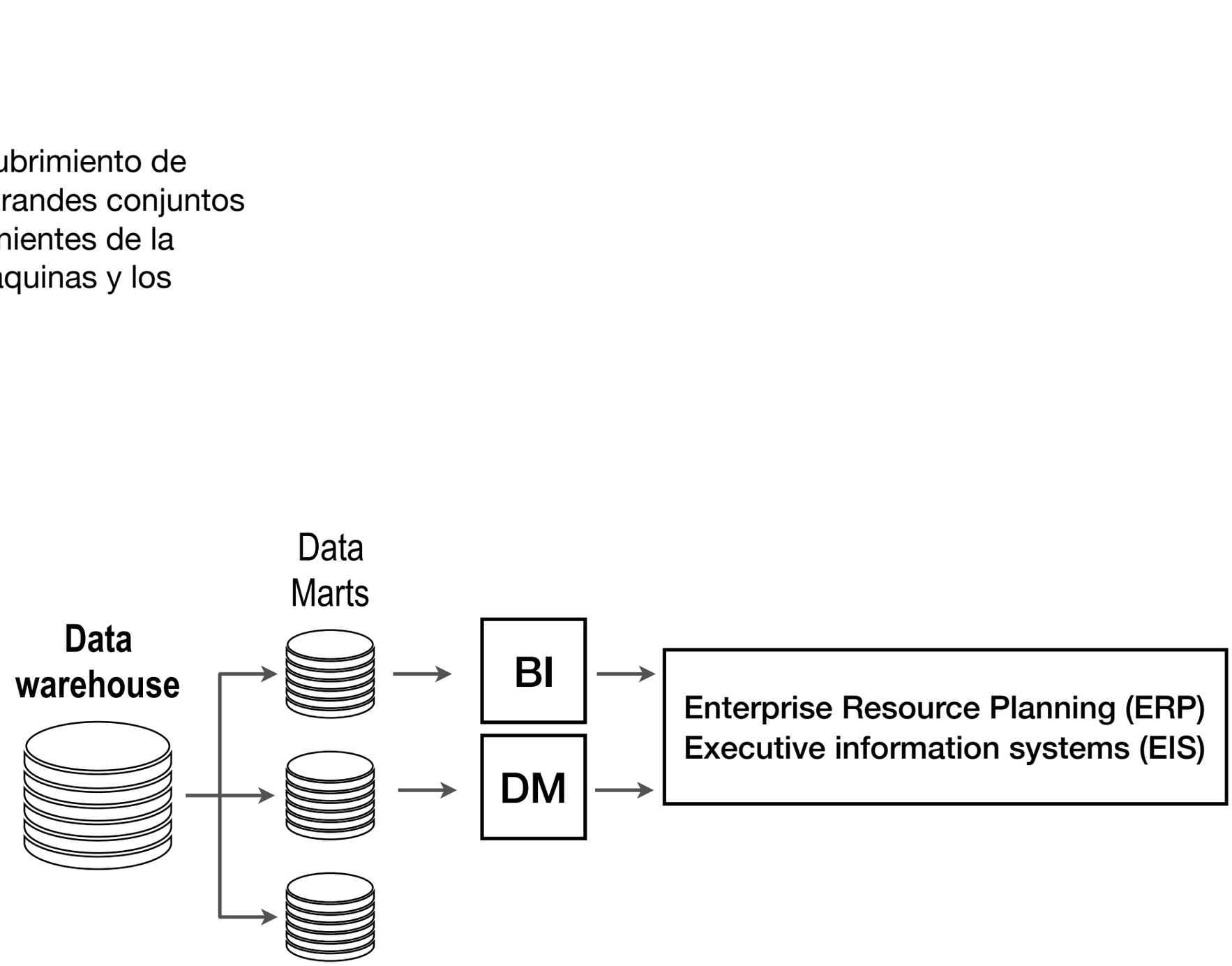


Data Mining

Proceso computacional de descubrimiento de patrones y tendencias útiles en grandes conjuntos de datos usando métodos provenientes de la Estadística, el Aprendizaje de Máquinas y los sistemas de bases de datos.

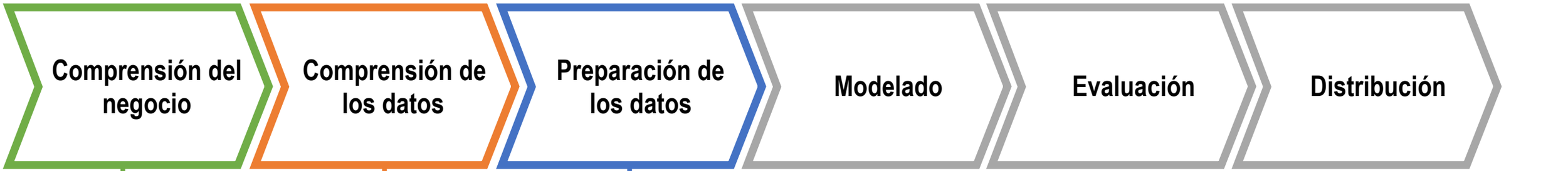
Tareas típicas:

- Detección de anomalías.
- Modelado de dependencias.
- Agrupamiento.
- Clasificación
- Regresión
- Resumen



Data Mining

Metodología CRISP-DM



Resultados del proyecto
- Establecer objetivos
- Formulación del plan preliminar
- Criterios de éxito

Valoración de la situación actual
- Inventario de recursos
- Requerimientos, hipótesis y restricciones
- Riesgos y contingencias
- Terminología
- Costos y beneficios

Objetivos de la minería de datos
- Objetivos de la DM
- Criterios de éxito desde el negocio
- Criterios de éxito desde la DM

Plan del proyecto
- Desarrollo del plan de proyecto
- Selección de herramientas y técnicas

Recopilación de datos iniciales
- Datos existentes
- Datos adquiridos
- Datos adicionales

Descripción de los datos
- Cantidad de los datos
- Tipos de valores
- Esquema de codificación

Exploración de datos
- Análisis exploratorio

Verificación de calidad de datos
- Datos perdidos
- Errores de datos
- Errores de medición
- Incoherencias de codificación
- Metadatos erróneos

Selección de datos
- Selección de elementos (filas)
- Selección de atributos (columnas)

Limpieza de datos
- Datos perdidos
- Errores de datos
- Errores de medición
- Incoherencias de codificación
- Metadatos erróneos

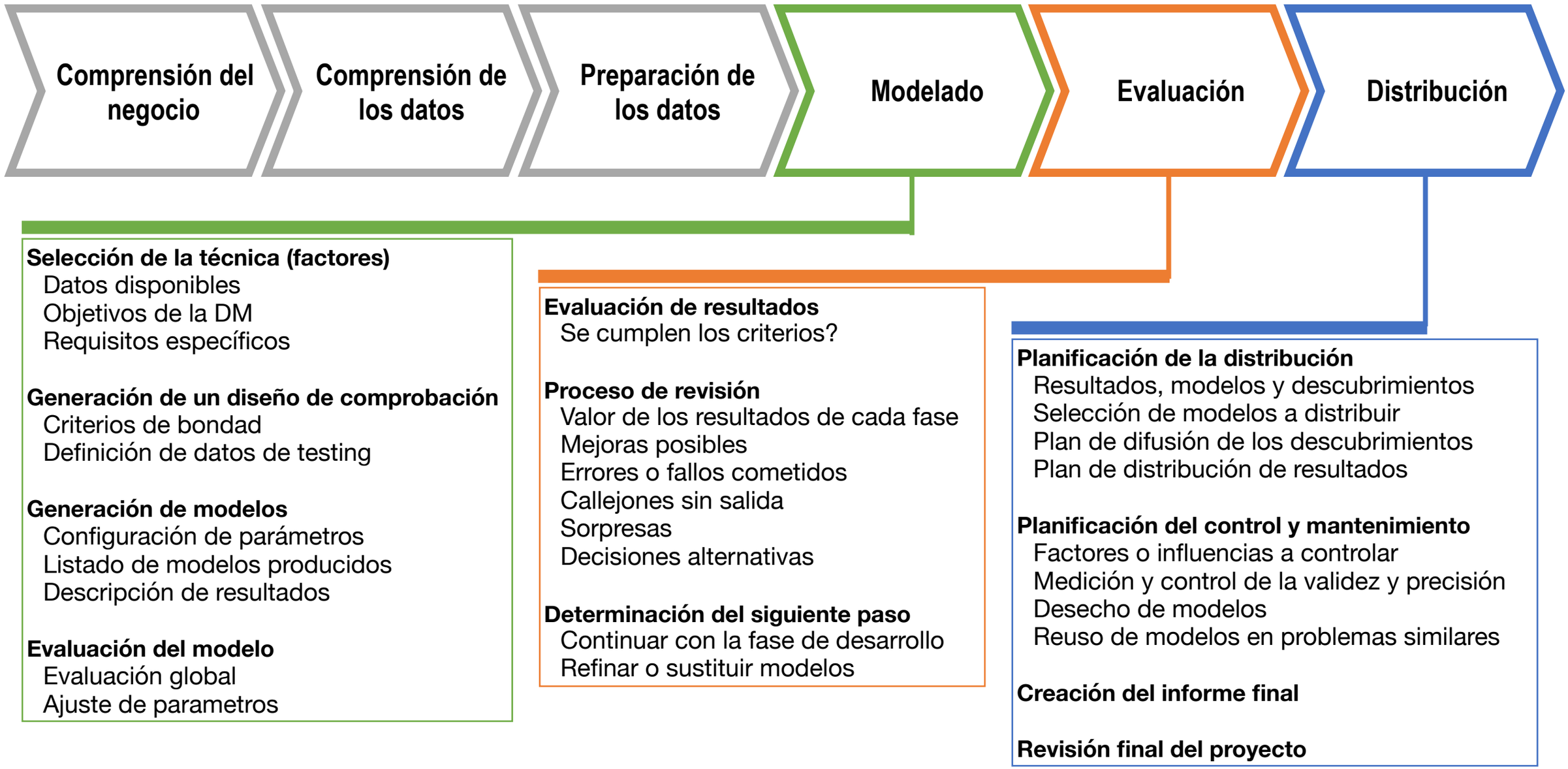
Construcción de nuevos datos
- Derivación de atributos (Columnas)
- Generación de registros (filas)

Integración de datos
- Fusión (columnas)
- Adición (filas)

Formato de datos

Data Mining

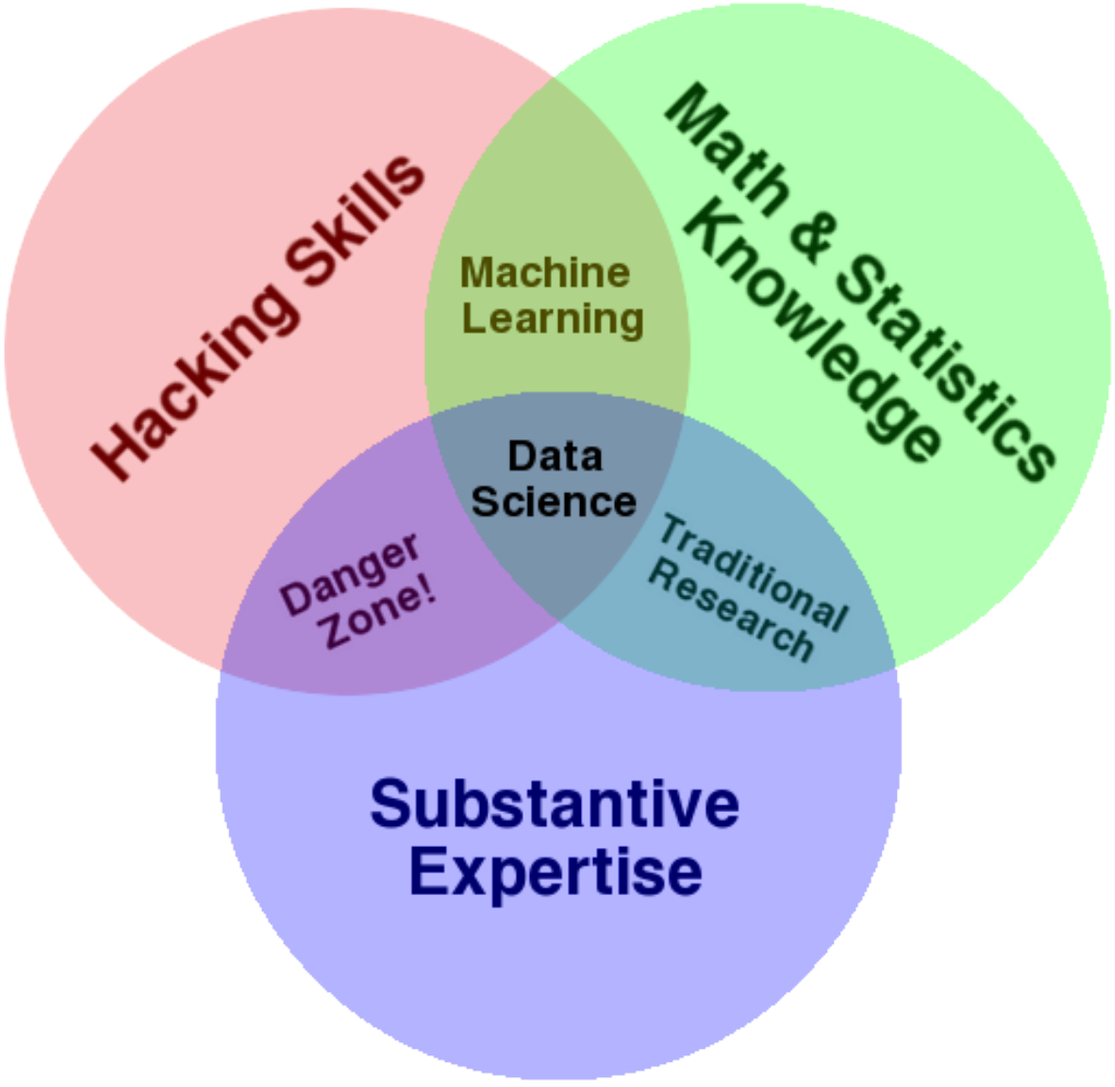
Metodología CRISP-DM



Data Mining

Disciplina	Tecnología	Habilidades	Foco
Análisis de datos	• Software para modelado de datos • Software para diagramación • Software para documentación • SQL • Software para perfilado de datos	• Modelado de datos • Análisis del negocio • Manipulación de datos • Estadística básica	• Reglas de negocio • Definición de datos • Relaciones entre datos • Atributos de datos • Estructuras de datos • Fuentes y usos de datos • Calidad de datos
Inteligencia de Negocios	• ETL/SQL • RDBMS • Reportes • Visualización	• Programación • Análisis de datos • Modelado de datos • Desarrollo de reportes • Estadística Básica • Análisis del negocio & Estrategia • Presentación oral	• Suministro de información y reporte • Visualización de datos • Estadísticos descriptivos • Integración de datos y consolidación
Data Mining	• Software estadístico • Herramientas de aprendizaje de máquinas • Lenguajes de programación	• Programación • Modelado de datos • Estadística Avanzada • Presentación oral	• Análisis estadístico avanzado • Manejo de grandes volúmenes de datos • Visualización de datos • Modelos de datos

Data Science



Data Science

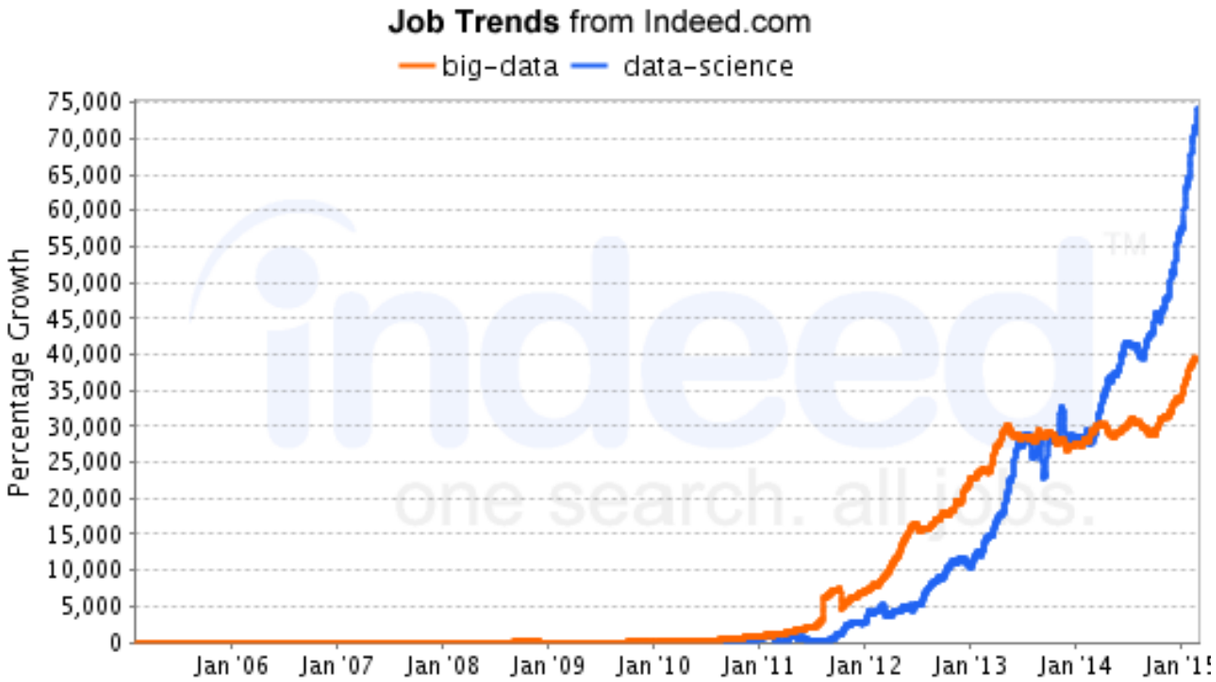
> A recent study by the McKinsey Global Institute concludes, "a shortage of the analytical and managerial talent necessary to make the most of Big Data is a significant and pressing challenge (for the U.S.)." The report estimates that there will be four to five million jobs in the U.S. requiring data analysis skills by 2018, and that large numbers of positions will only be filled through training or retraining. The authors also project a need for 1.5 million more managers and analysts with deep analytical and technical skills "who can ask the right questions and consume the results of analysis of big data effectively."

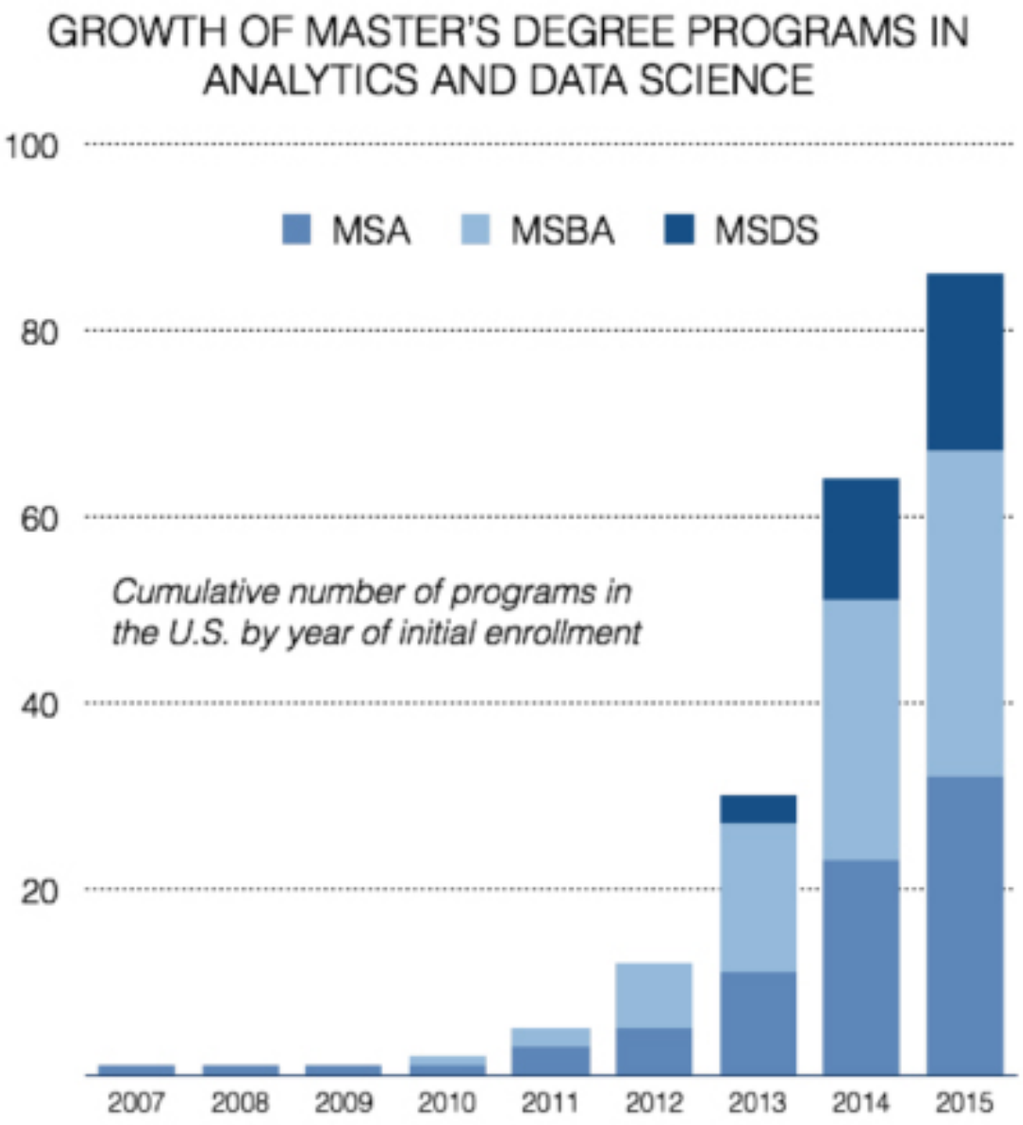
#16	3,433	$105,395	#1
Highest Paying Job in Demand	Number of Job Openings	Average Base Salary	Best Job in America for 2016

Sources: 25 Best Jobs in America and 25 Highest Paying Jobs in America for 2016

https://datascience.berkeley.edu/about/what-is-data-science/

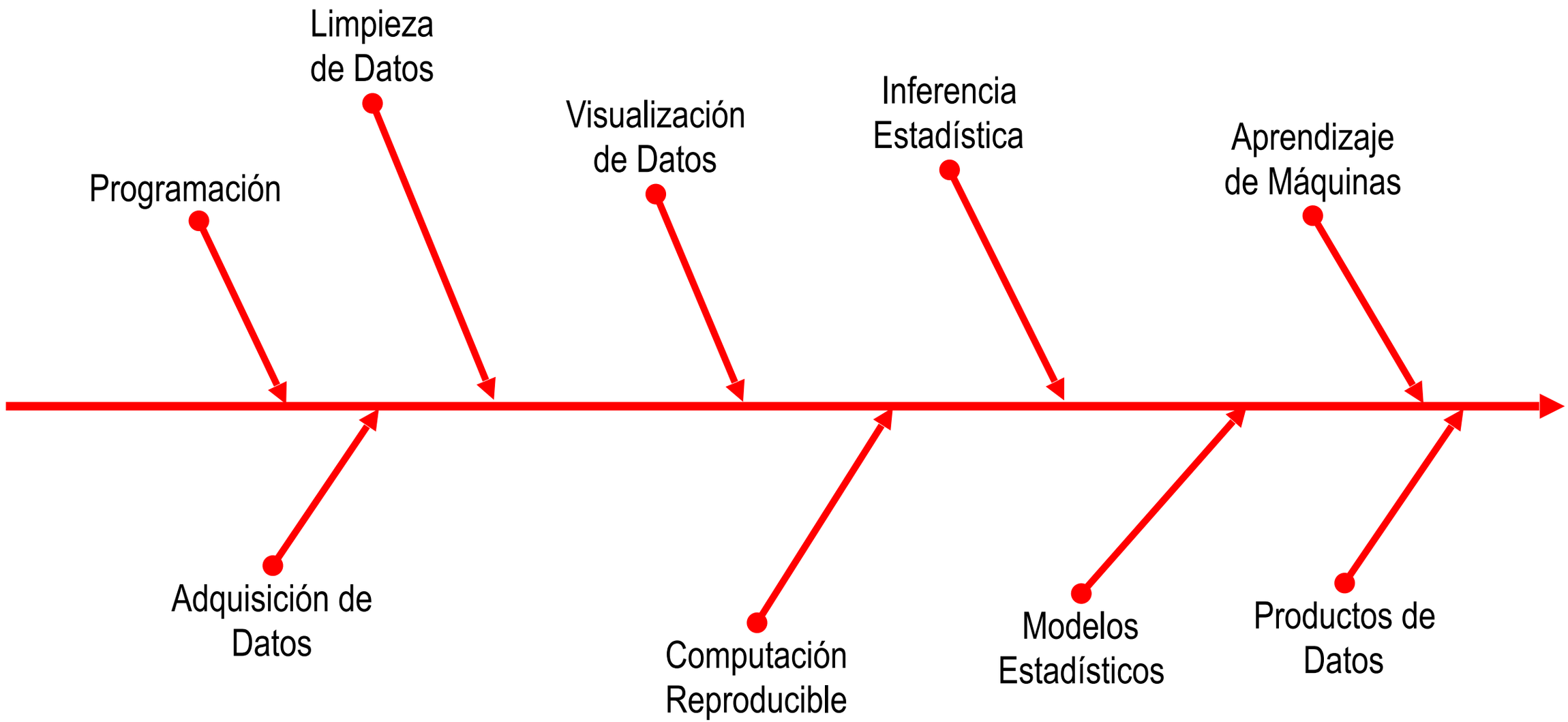
Data Science





http://analytics.ncsu.edu/?page_id=4184

Data Science

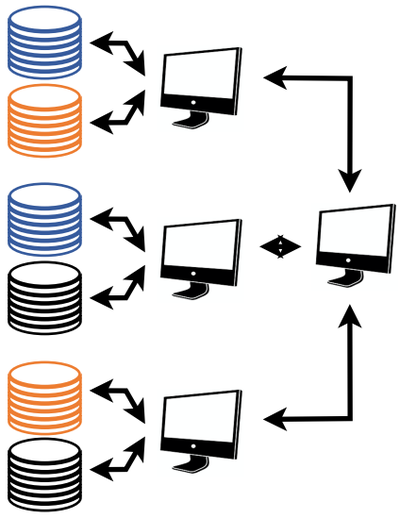


Data-driven decision making

Servicios Web

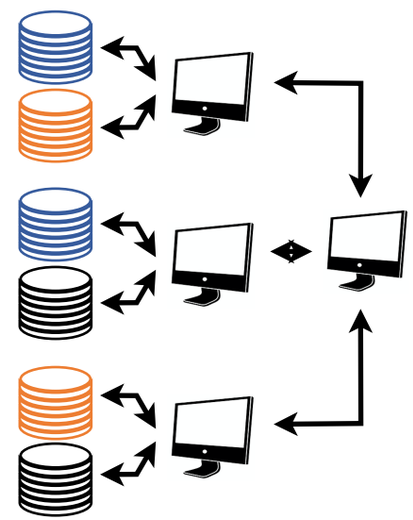
Computación local

Servidores + red + clientes



Cloud computing / utility computing

Servidores y almacenamiento en la nube + internet + clientes locales



Software as a Service (SaaS)

Software almacenado en máquinas suministradas por un tercero.

Aplicaciones accesadas vía un cliente o la Web.

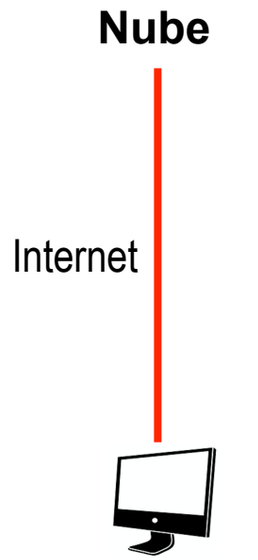
Orientado a aplicaciones de usuario final.

Platform as a Service (PaaS)

Orientado a desarrolladores.
Ambiente de desarrollo gestionado por un tercero.

Infrastructure as a Service (IaaS)

Bloques básicos para construcción de ambientes manejados por un tercero
Capacidad de procesamiento, almacenamiento, conectividad, seguridad, etc.



DevOps - Integración y Entrega continua de software

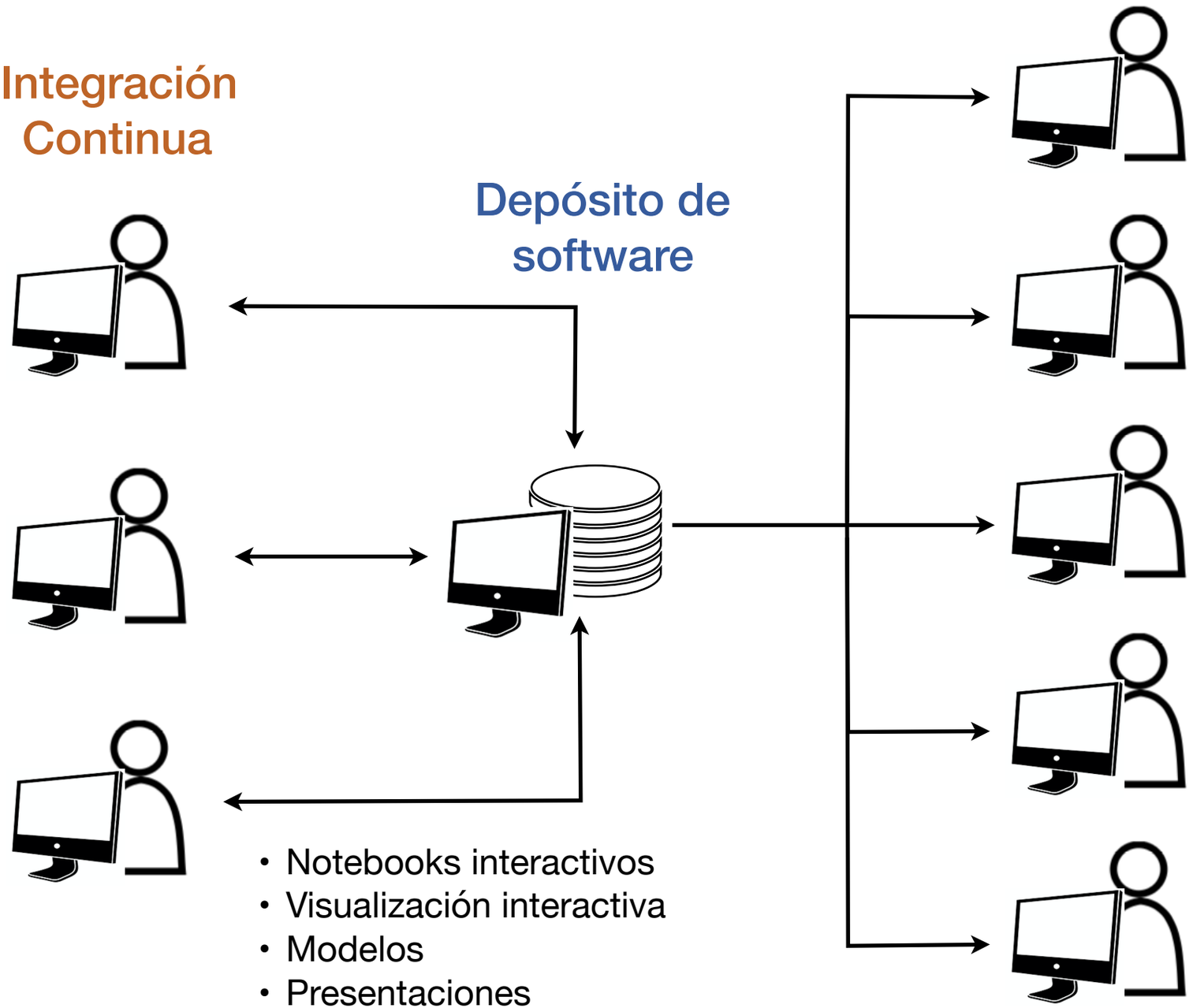
Integración Continua

Depósito de software

Entrega Continua

Equipos de desarrollo de software & Científicos de datos

Analistas de datos & Científicos de datos



- Notebooks interactivos
- Visualización interactiva
- Modelos
- Presentaciones
- Aplicaciones web inteligentes
- Servicios RESTful
- Dashboards

Hadoop / MapReduce

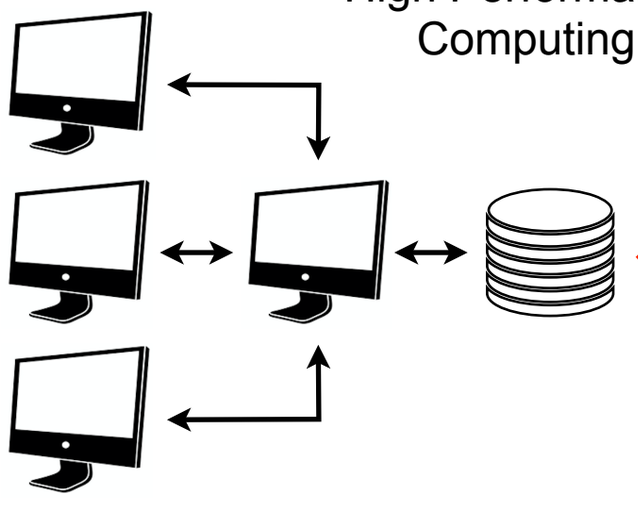
Aproximación Tradicional
(RDBMS y data warehouse)

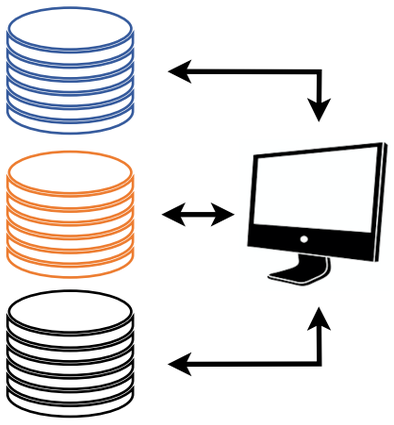
- Alto volumen de datos
- Lectura lenta
- Escritura + lenta
- Búsqueda ++ lenta

RDBMS
- Orientado a filas
- Estructura fija
- SQL
- Consistencia
- Integridad referencial
- Abstracción

Procesamiento distribuido

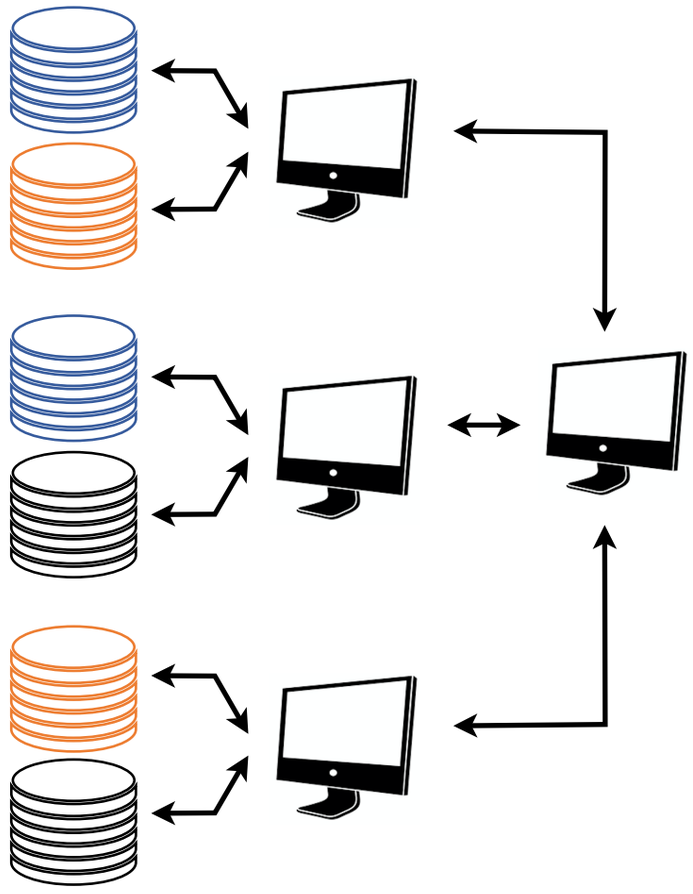
High Performance Computing





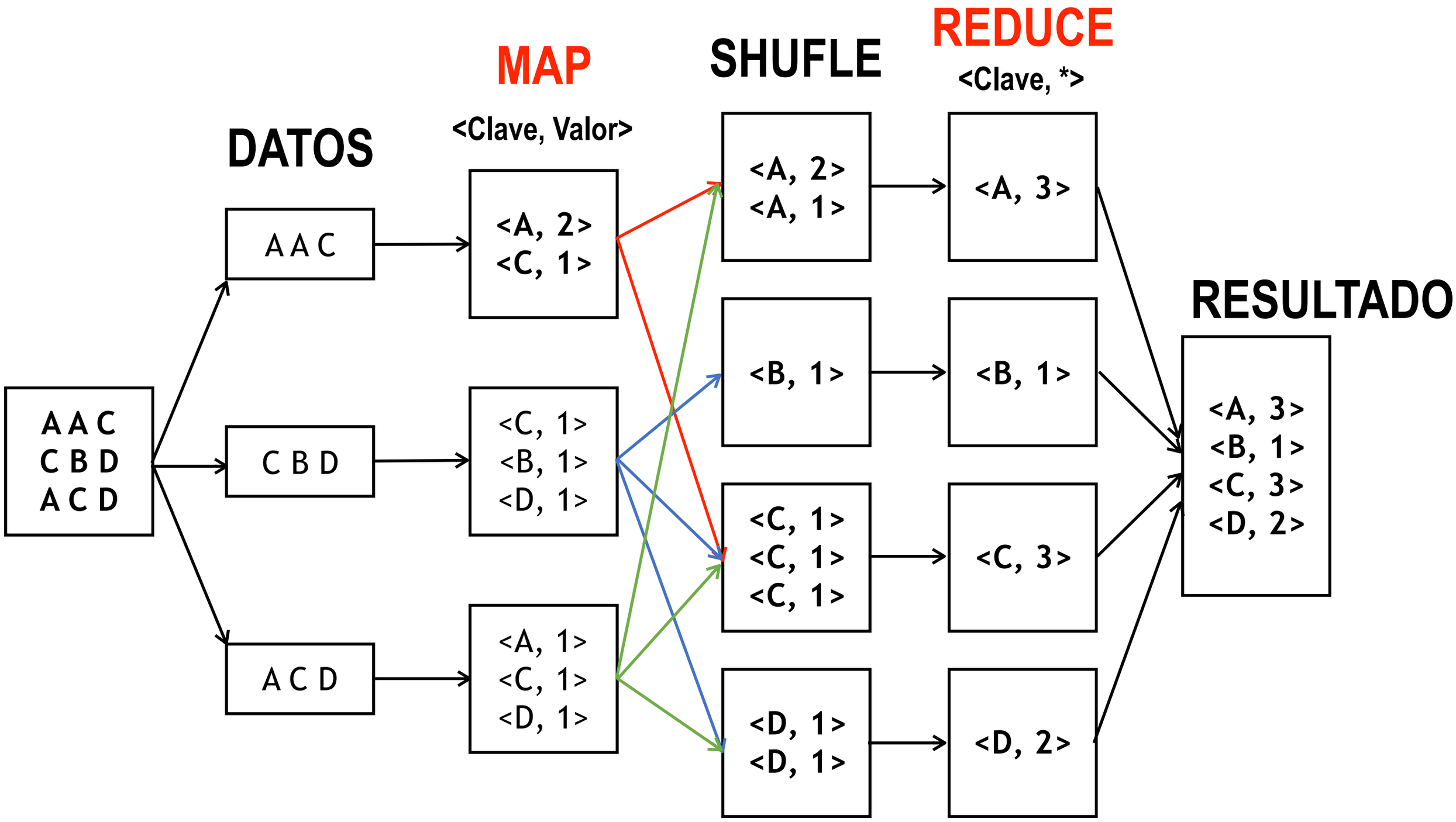
Almacenamiento distribuido

Hadoop Distributed Filesystem (HDFS)



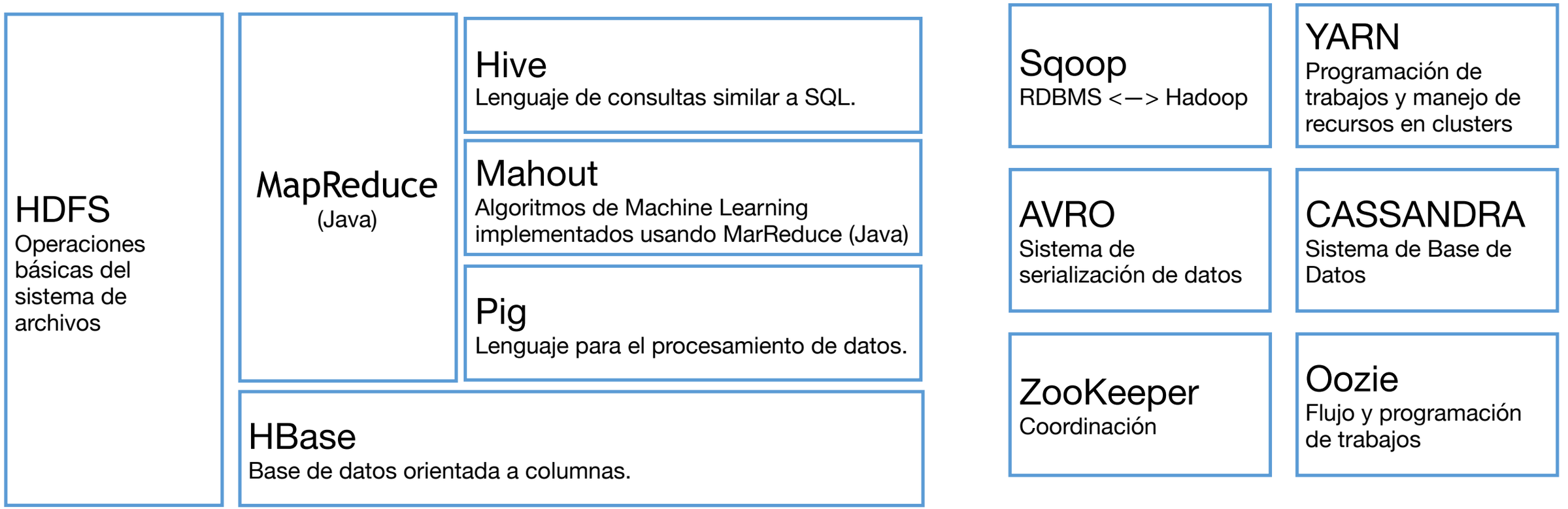
Nodos de Datos + Nodos de Computo

Hadoop / MapReduce



Hadoop / MapReduce

Ecosistema Apache Hadoop



Hadoop / MapReduce

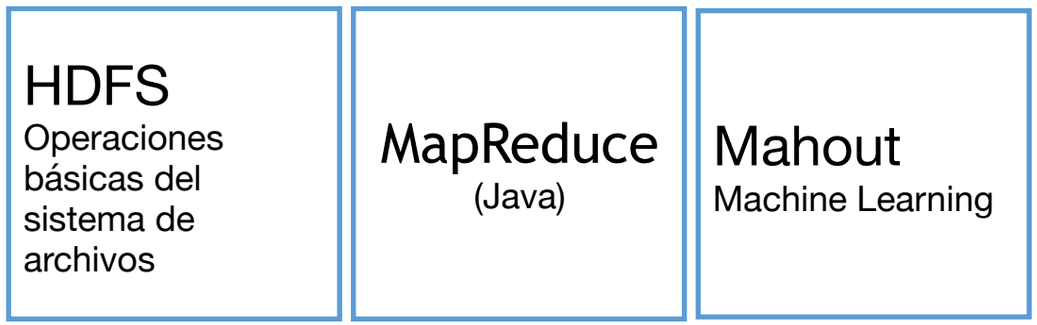
Ejemplo de Pig

```
records = LOAD 'sample.txt' AS (year:chararray, temperature:int, quality:int);
filtered_records = FILTER records BY temperature;
grouped_records = GROUP filtered_records BY year;
max_temp = FOREACH grouped_records GENERATE group, MAX(filtered_records.temperature);
DUMP max_temp;
```

Ejemplo de Hive

```
CREATE TABLE records (year STRING, temperature INT, quality INT)
  ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';
LOAD DATA LOCAL INPATH 'sample.txt' OVERWRITE INTO TABLE records;
SELECT year, MAX(temperature) FROM records GROUP BY year;
```

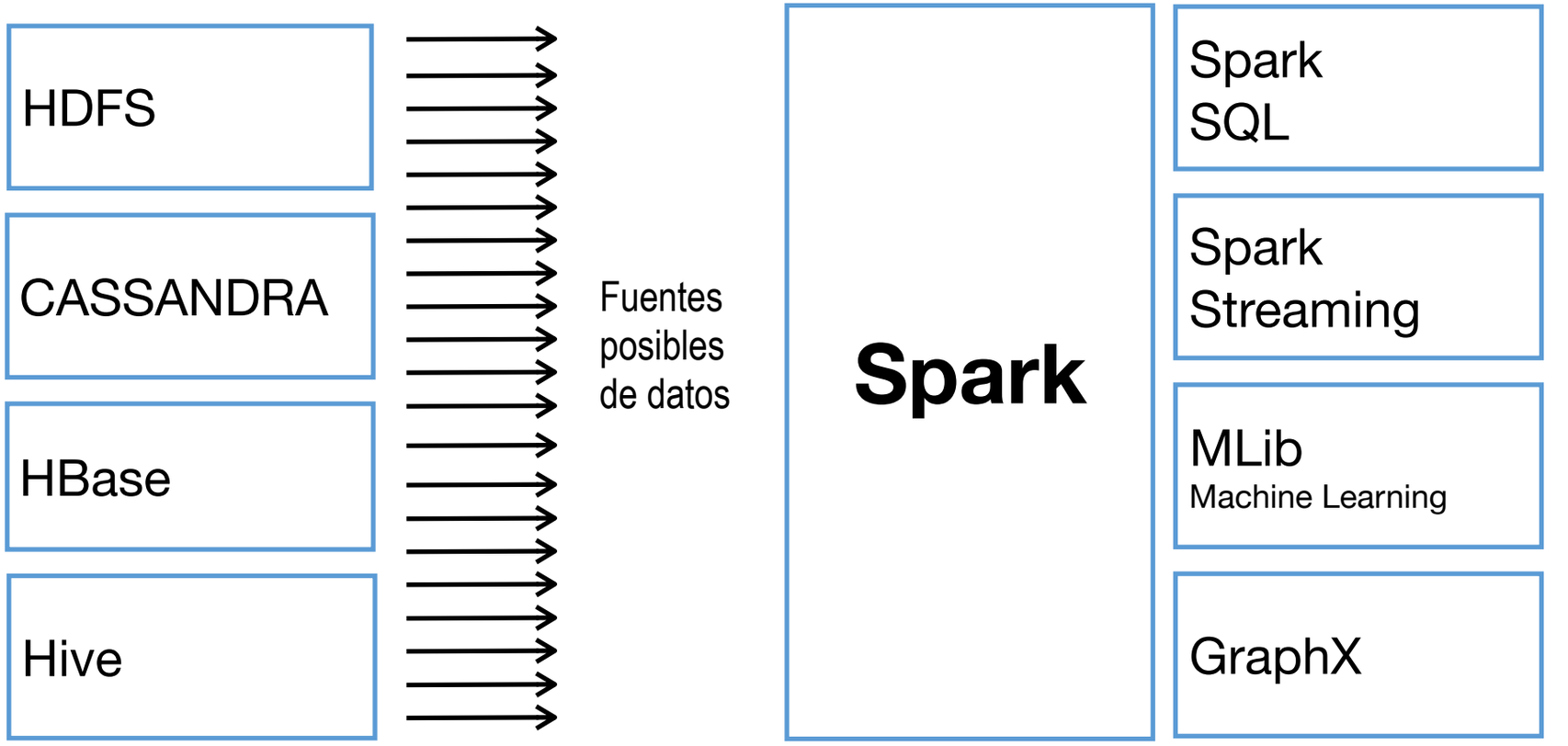
Hadoop / MapReduce



Hadoop / MapReduce

Regresión logística	**RHadoop**
Regresión lineal	rdfs
Clustering	rmr
Filtrado colaborativo	rhbase

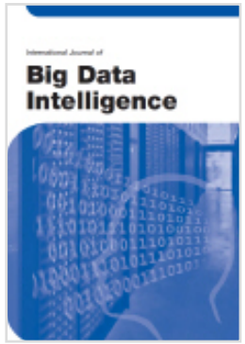
http://mahout.apache.org/users/basics/algorithms.html

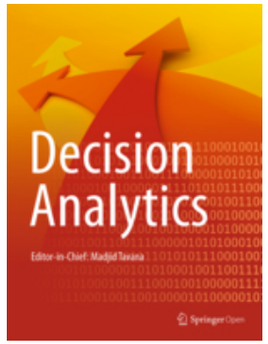


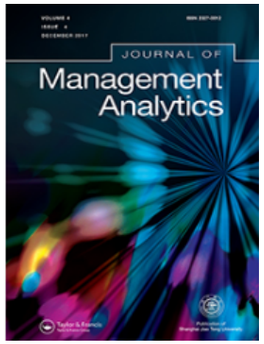
Java
Scala
Python
R

https://spark.apache.org/mllib/

Big Data / Data Science





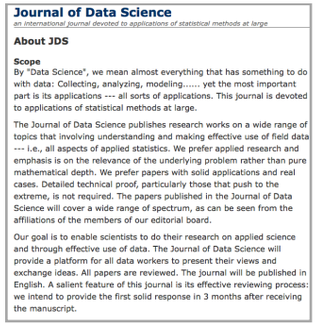


DATA SCIENCE JOURNAL

IEEE TRANSACTIONS ON
BIG DATA
IEEE computer society

2002 2003 2012 2013 2014 2015 2016

Journal of Data Science



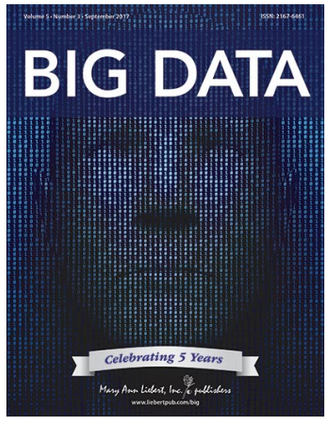
Journal of Data Science
an international journal devoted to applications of statistical methods at large

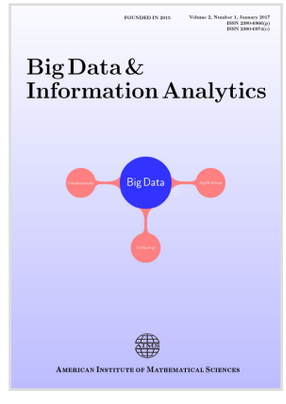
About JDS

Scope
By "Data Science", we mean almost everything that has something to do with data: Collecting, analyzing, modeling..... yet the most important part is its applications --- all sorts of applications. This journal is devoted to applications of statistical methods at large.

The Journal of Data Science publishes research works on a wide range of topics that involving understanding and making effective use of field data --- i.e., all aspects of applied statistics. We prefer applied research and emphasis is on the relevance of the underlying problem rather than pure mathematical depth. We prefer papers with solid applications and real cases. Detailed technical proof, particularly those that push to the extreme, is not required. The papers published in the Journal of Data Science will cover a wide range of spectrum, as can be seen from the affiliations of the members of our editorial board.

Our goal is to enable scientists to do their research on applied science and through effective use of data. The Journal of Data Science will provide a platform for all data workers to present their views and exchange ideas. All papers are reviewed. The journal will be published in English. A salient feature of this journal is its effective reviewing process: we intend to provide the first solid response in 3 months after receiving the manuscript.





NoSQL

Datos tabulares

KEY	Fecha	Planta	Generación
001	2017-10-01	Jaguas	100.2
002	2017-10-01	Playas	23.1
003	2017-10-01	Guatape	130.1

Document (JSON/XML)

```
[
 {
    Fecha:2017-10-01,
    Planta:Jaguas,
    Generación: 100.2
 },{
    Fecha:2017-10-01,
    Planta:Playas,
    Generación:23.1,
 },{
    Fecha:2017-10-01,
    Planta:Guatapé,
    Generación:130.1
  }
]
```

Pares <clave, valor>

```
Tabla001.Fecha=2017-10-01
Tabla001.Planta=Jaguas
Tabla001.Generación=100.2
Tabla002.Fecha=2017-10-01
Tabla002.Planta=Playas
Tabla002.Generación=23.1
Tabla003.Fecha=2017-10-01
Tabla003.Planta=Guatapé
Tabla003.Generación=130.1
```

Sistema orientado a filas

```
001:2017-10-01,Jaguas,100.2
002:2017-10-01,Playas,23.1
003:2017-10-01,Guatape,130.1
```

Column family database

```
001:{Fecha:2017-10-01, Planta:Jaguas, Generación:100.2}
002:{Fecha:2017-10-01, Planta:Playas, Generación:23.1}
003:{Fecha:2017-10-01, Planta:Guatapé, Generación:130.1}
```

Open Data Science

R

Python

Julia

C

Bash

KNIME

TensorFlow

Usuario

Analytics Solver (Frontline Systems)

Pentaho

IBM Watson explorer
IBM SPSS Modeler

SAS predictive Analytics

RapidMiner

Oracle Data Mining

TIBCO Analytics

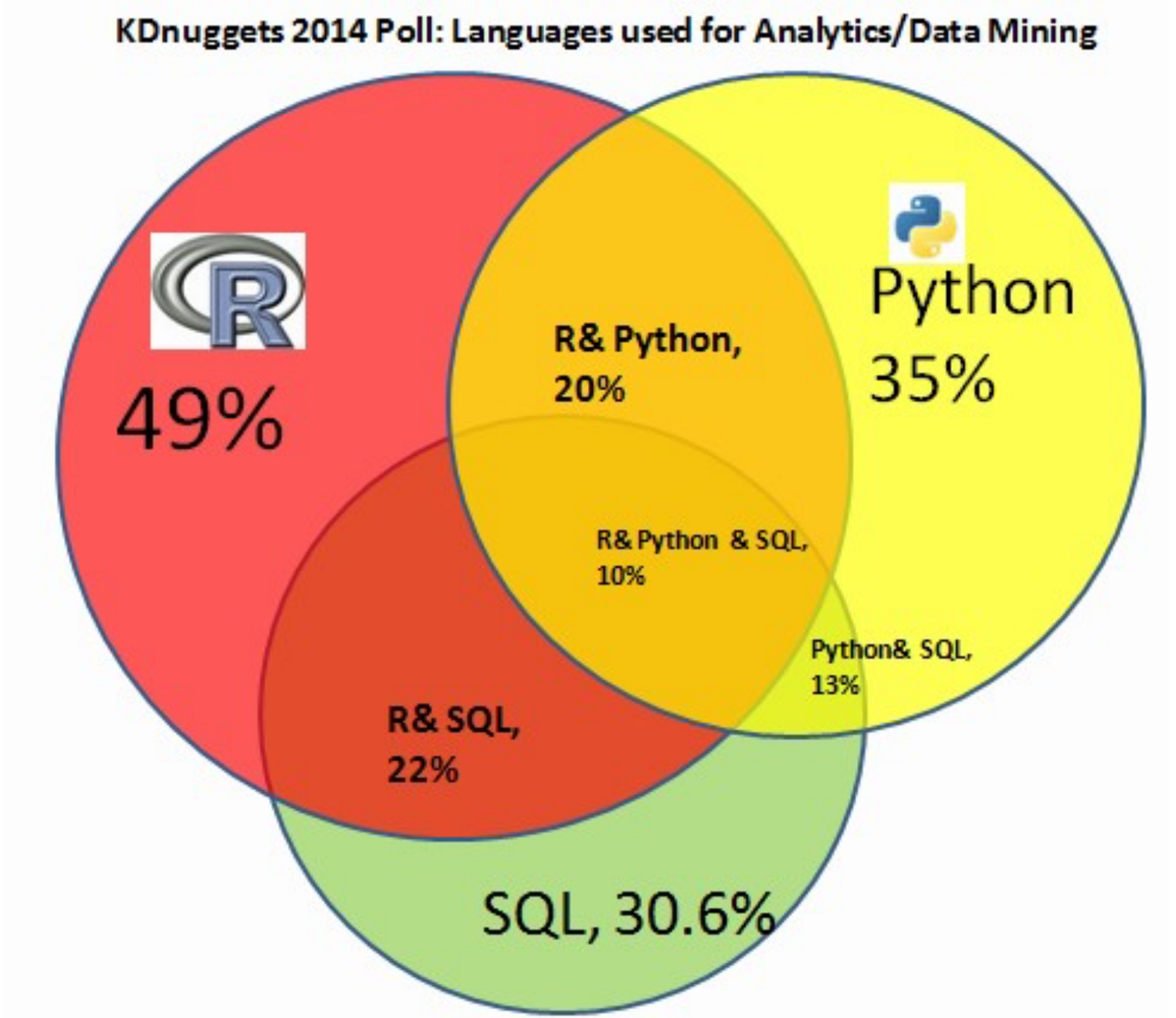
Salford Systems

Tableau

Qilk

Open Data Science

Popularidad de los lenguajes



http://www.kdnuggets.com/2014/08/four-main-languages-analytics-data-mining-data-science.html

Open Data Science

The **2015** Top Ten Programming Languages (IEEE Spectrum)

Language Rank	Types	Spectrum Ranking	Spectrum Ranking
1. Java		100.0	100.0
2. C		99.9	99.3
3. C++		99.4	95.5
4. Python		96.5	93.5
5. C#		91.3	92.4
6. R		84.8	84.8
7. PHP		84.5	84.5
8. JavaScript		83.0	78.9
9. Ruby		76.2	74.3
10. Matlab		72.4	72.8

http://spectrum.ieee.org/computing/software/the-2015-top-ten-programming-languages

Open Data Science

The **2016** Top Ten Programming Languages (IEEE Spectrum)

Language Rank	Types	Spectrum Ranking
1. C		100.0
2. Java		98.1
3. Python		98.0
4. C++		95.9
5. R		87.9
6. C#		86.7
7. PHP		82.8
8. JavaScript		82.2
9. Ruby		74.5
10. Go		71.9

http://spectrum.ieee.org/computing/software/the-2016-top-programming-languages

Open Data Science

The **2017** Top Ten Programming Languages (IEEE Spectrum)

Language Rank	Types	Spectrum Ranking
1. Python		100.0
2. C		99.7
3. Java		99.5
4. C++		97.1
5. C#		87.7
6. R		87.7
7. JavaScript		85.6
8. PHP		81.2
9. Go		75.1
10. Swift		73.7

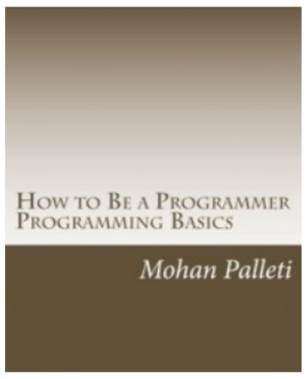
http://spectrum.ieee.org/computing/software/the-2017-top-programming-languages

Programación -- www.allitebooks.com

Algorithms Unlocked

By: Thomas H. Cormen

Have you ever wondered how your GPS can find the fastest way to your destination, selecting one route from seemingly countless possibilities in mere seconds? How your credit card account number is protected when you make a purchase over the Internet? The answer is algorithms. And how do…



How to Be a Programmer: Programming Basics

By: Mohan Palleti

A Self-help 97 pages book to learn the basics of programming using Microsoft Excel's VBA tools. Ideal resource for school teachers and educators wanting to teach programming basics.

Programación -- ¿Usted sabe programar … / Es capaz de …?

¿Ordenar un vector de números?

¿Calcular la suma de los primeros 20 números primos?

¿Computar la inversa de una matriz?

Programación para ingeniería

Cómputo numérico.

Programación para Computer Sciences

Manipulación de texto.

Open Data Science

Explotación de HW moderno
- Servidores
- Clusters
- GPUs & Workstations

Storyboards
- Notebooks
- Exploración interactiva
- Programación visual
- Data IDE

Analytics
- Preparación de datos
- Estadística
- ML & Ensambles
- Deep learning
- Simulación y optimización
- Datos geoespaciales
- Texto y NLP
- Gráficos y redes
- Minería de audio, video e imágenes

SciPy
PyMC
StatsModels
Theano
Scikit-learn
NLTK
NetworkX
Theano
pycaffe
Pylearn2
R caret
R glmnet
R randomForest

SimPy
PyJMI
PyFMI
PyMC
Pyomo
CVXOPT
CVXPY
tao4py
pyopt
Pylpopt
PyGMO

Fuentes de datos modernas
- Big Data
- Spark
- NoSQL
- DW & SQL
- Archivos y servicios Web

Pandas
Blaze
GeoPandas
R plyr
R dplyr,
R tidyr
R reshape2
R sparklyr
R readr
R readXL
R lubridate
R stringr
R feather
R Tible
R ggpairs

Visualización
- Gráficos
- Visualización interactiva
- Big data
- Mapas & GIS
- 3D
- Streaming

Bokeh
Plot.ly
Seaborn
Geopandas
ggplot2

Aplicaciones modernas
- Notebooks
- Dashboards
- Aplicaciones visuales
- Servicios de datos

Open Data Science

Fuentes de datos

Archivos de datos y Web
- Archivos de texto delimitados
- JSON
- XML
- Archivos de Log
- Archivos específicos de aplicación

Data warehouse y SQL
- RDBMS
- Cubos de datos

Hadoop & Spark

Stream de datos

NoSQL
- Almacenes de documentos
- Bases de datos columnares
- Diccionarios (clave, valor)

(Data warehousing para gestión del mercado eléctrico)
(Sistemas de bases de datos en organizaciones)

Internet of Things

Red de dispositivos físicos con sensores y conectividad que les permiten recolectar e intercambiar datos.

Hogares inteligentes
Ciudades inteligentes
Vehículos eléctricos
Fuentes renovables de energía
Lineas de potencia
Perfil de la demanda
Respuesta de la demanda
Detección de fallos

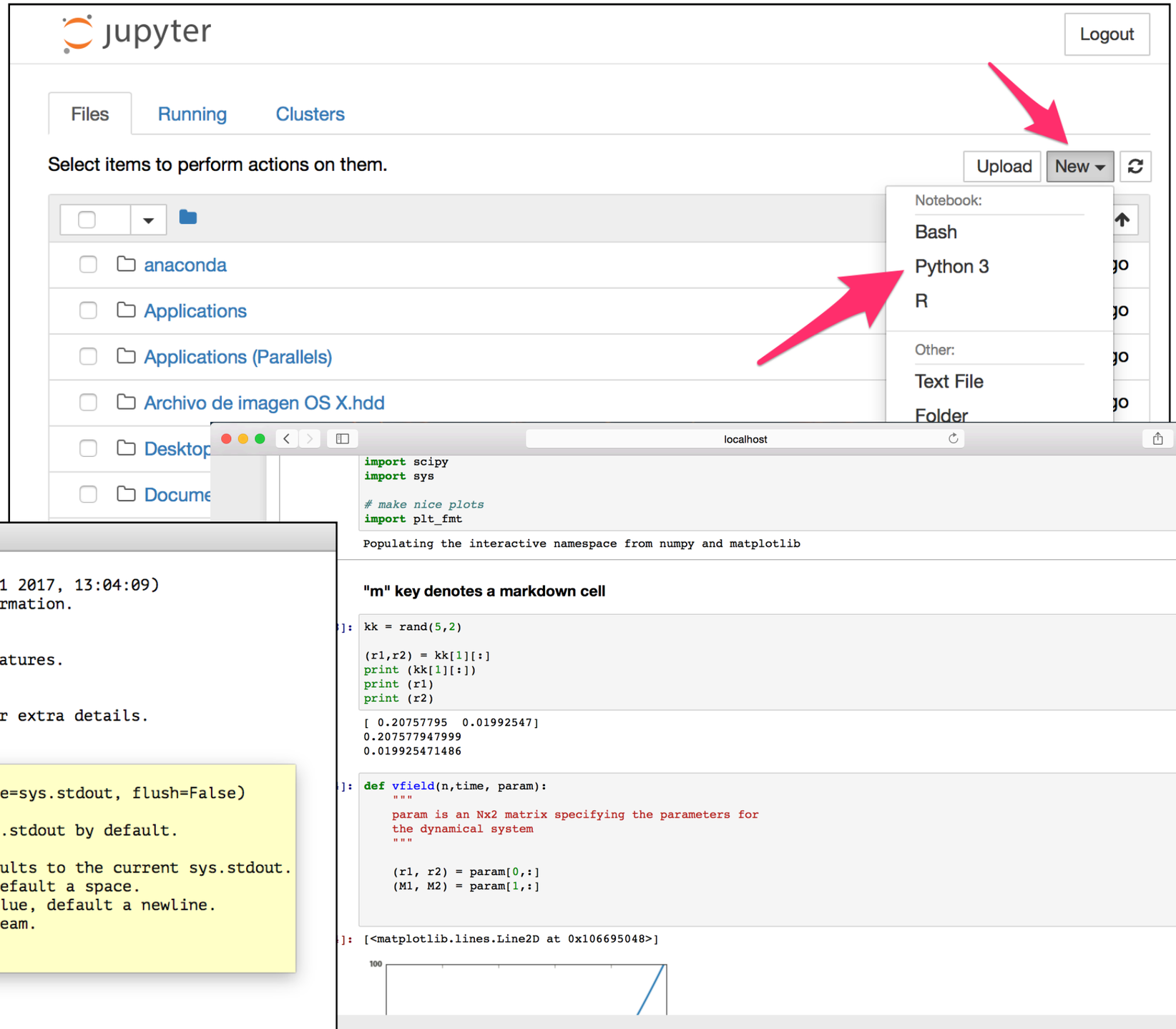
(Dispositivos usables, …)

Open Data Science

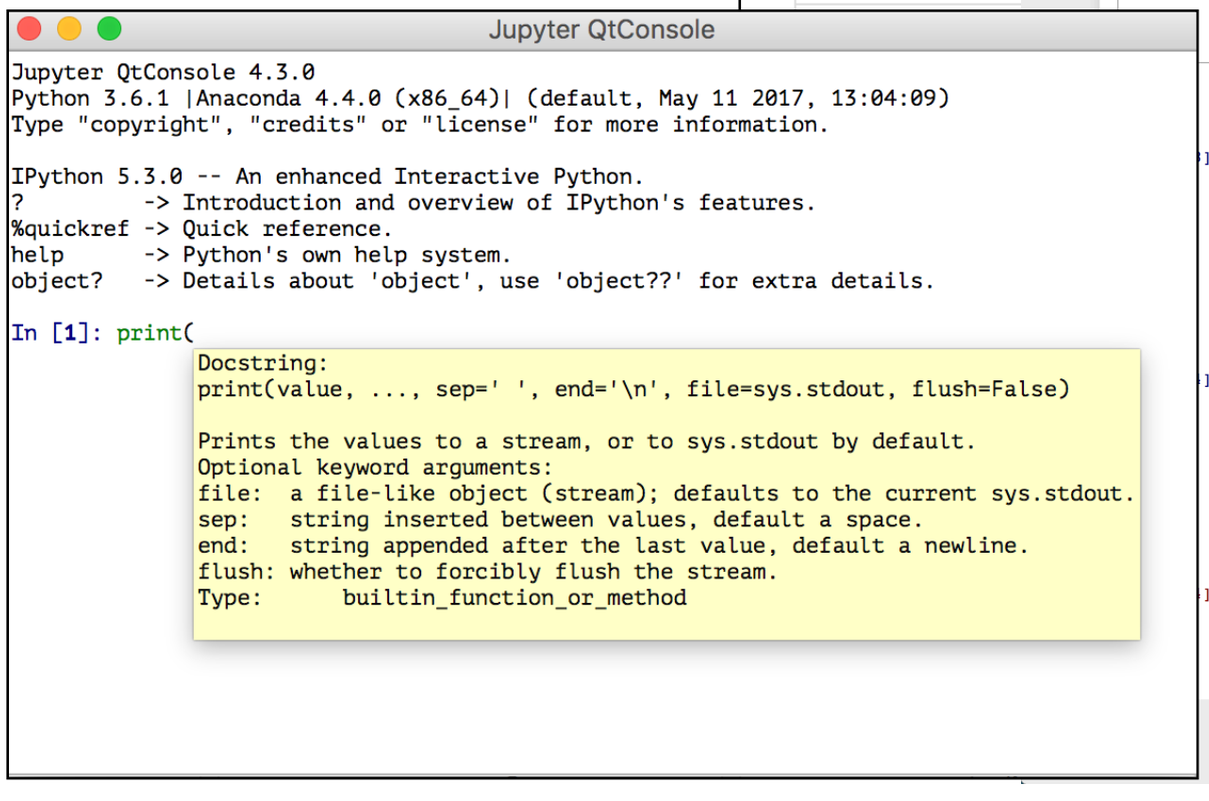
Storyboards
- Notebooks
- Exploración interactiva
- Programación visual
- Data IDE

Aplicaciones modernas
- Notebooks
- Dashboards
- Aplicaciones visuales
- Servicios de datos

Jupyter Notebook



Jupyter QtConsole

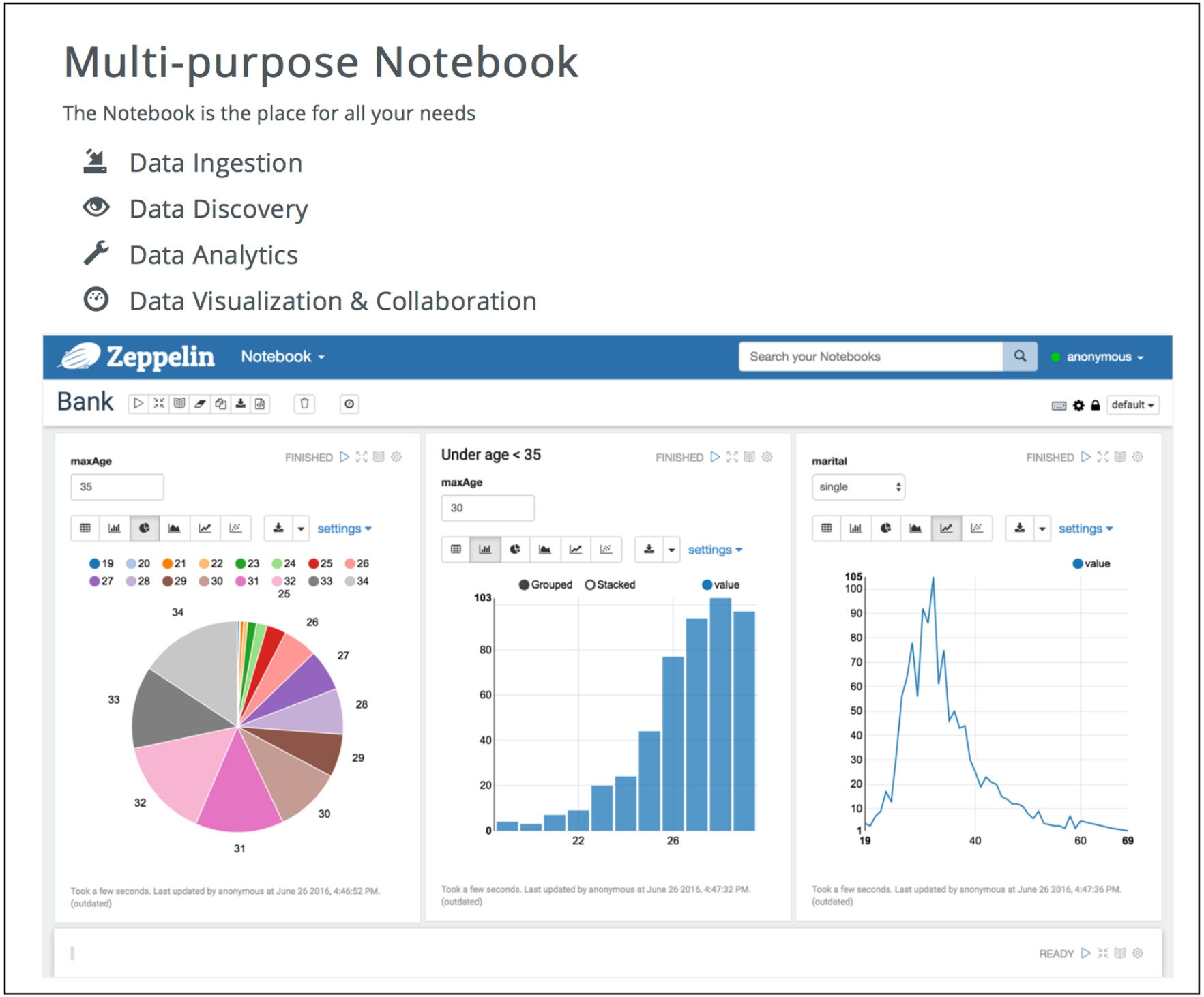


Open Data Science

Storyboards
- Notebooks
- Exploración interactiva
- Programación visual
- Data IDE

Aplicaciones modernas
- Notebooks
- Dashboards
- Aplicaciones visuales
- Servicios de datos

Apache Zepellin

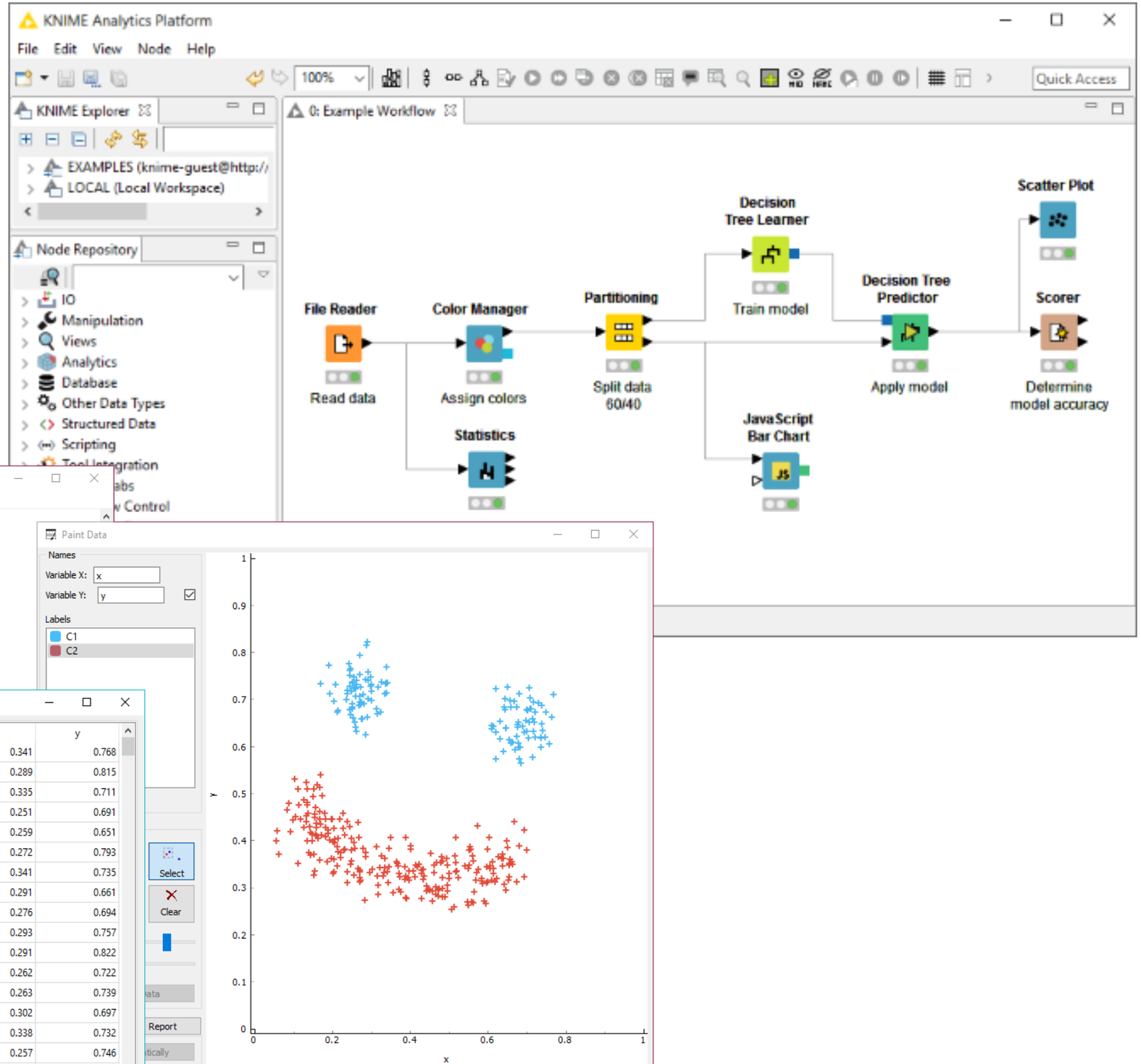


Open Data Science

Storyboards
- Notebooks
- Exploración interactiva
- Programación visual
- Data IDE

Aplicaciones modernas
- Notebooks
- Dashboards embebibles
- Aplicaciones visuales
- Servicios de datos

KNIME



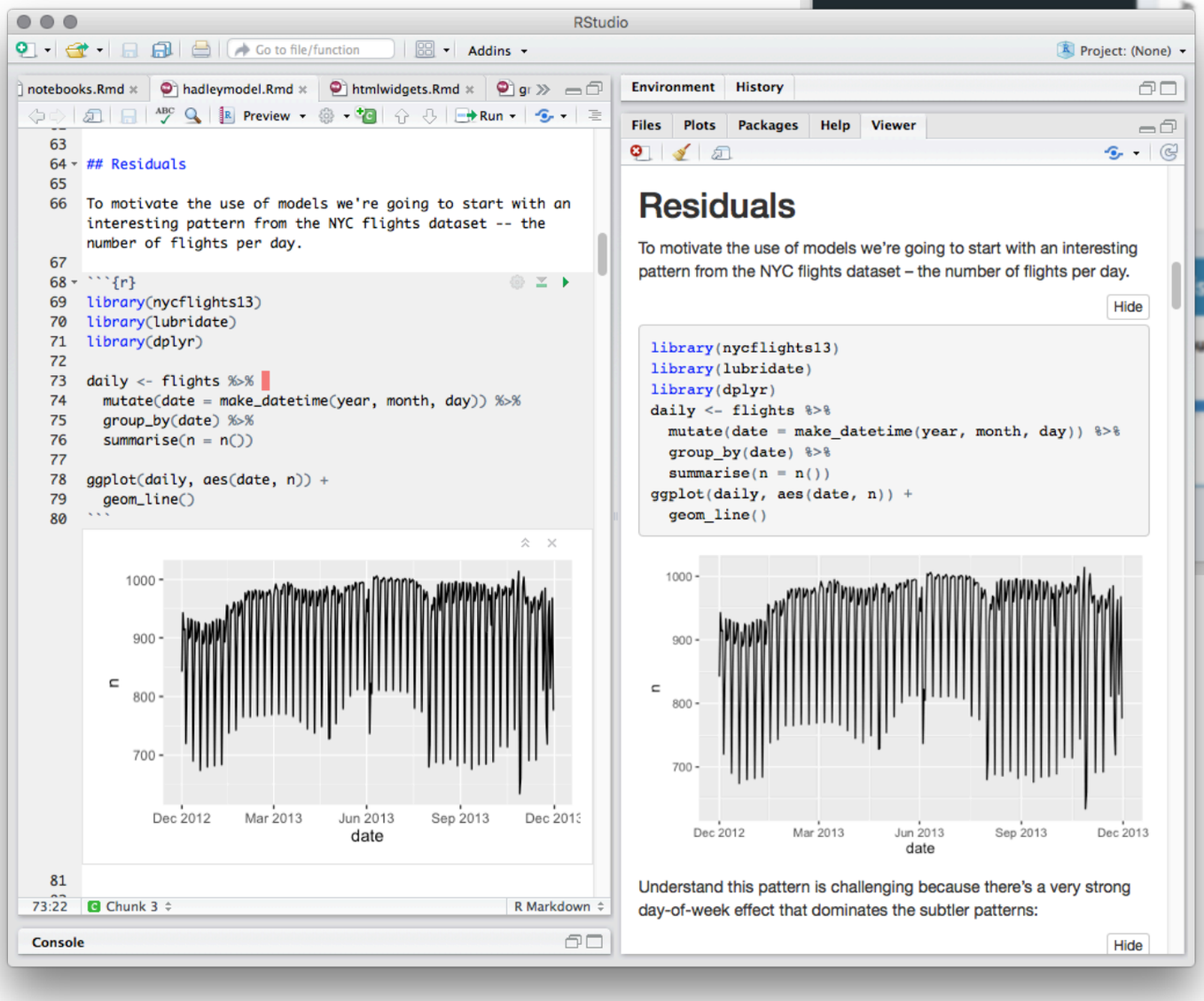
Orange

Open Data Science

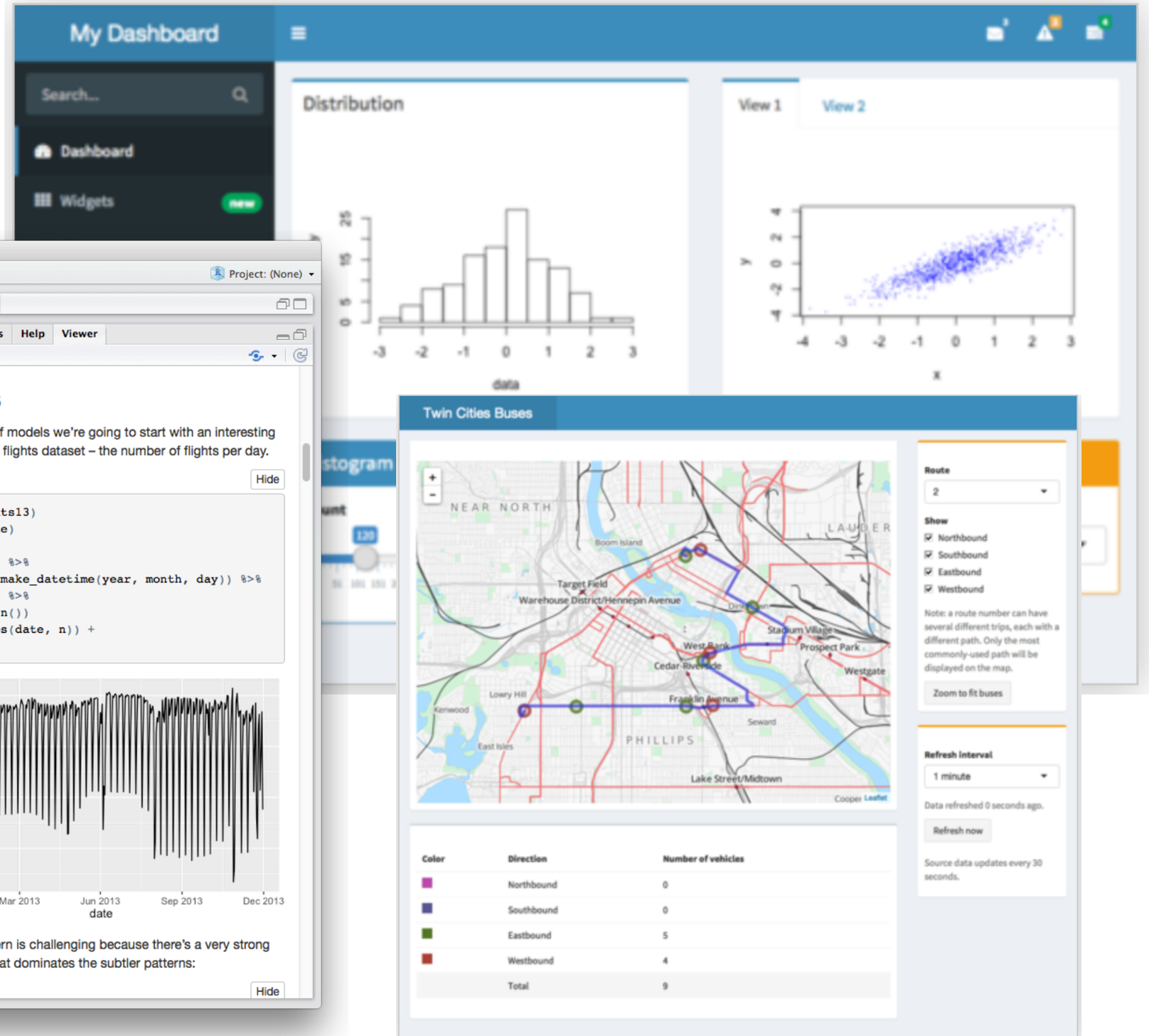
Aplicaciones modernas

- Notebooks
- Dashboards
- Aplicaciones visuales
- Servicios de datos

R Markdown



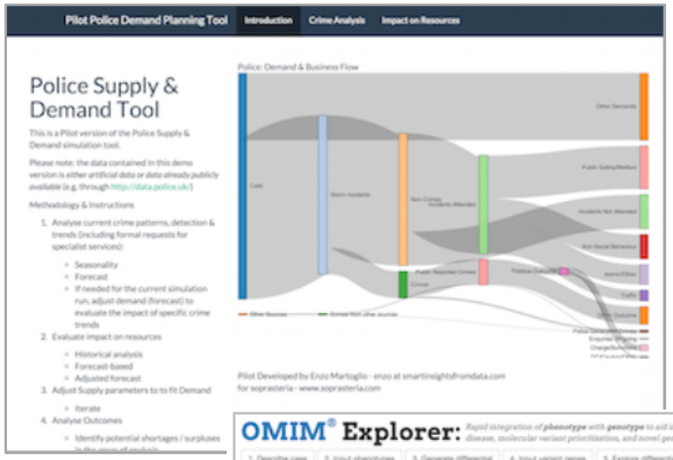
R Dashboards

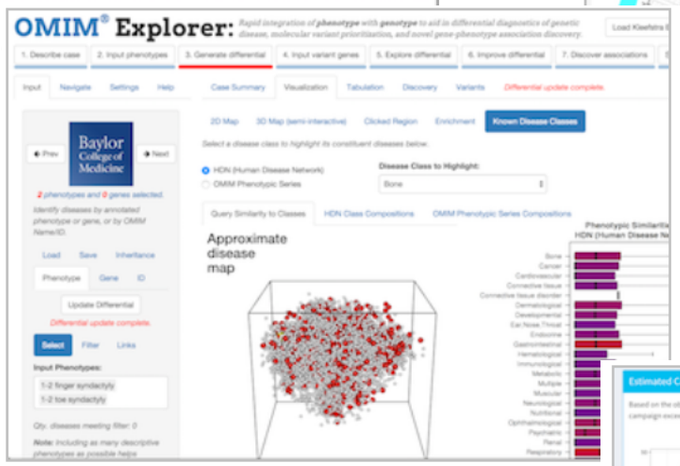


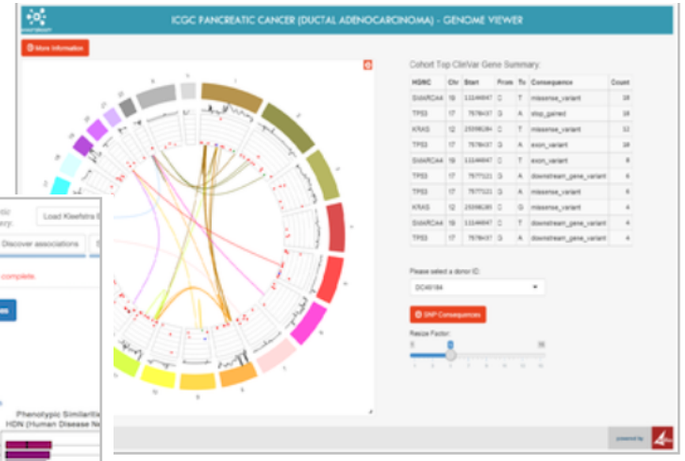
Open Data Science

Aplicaciones modernas

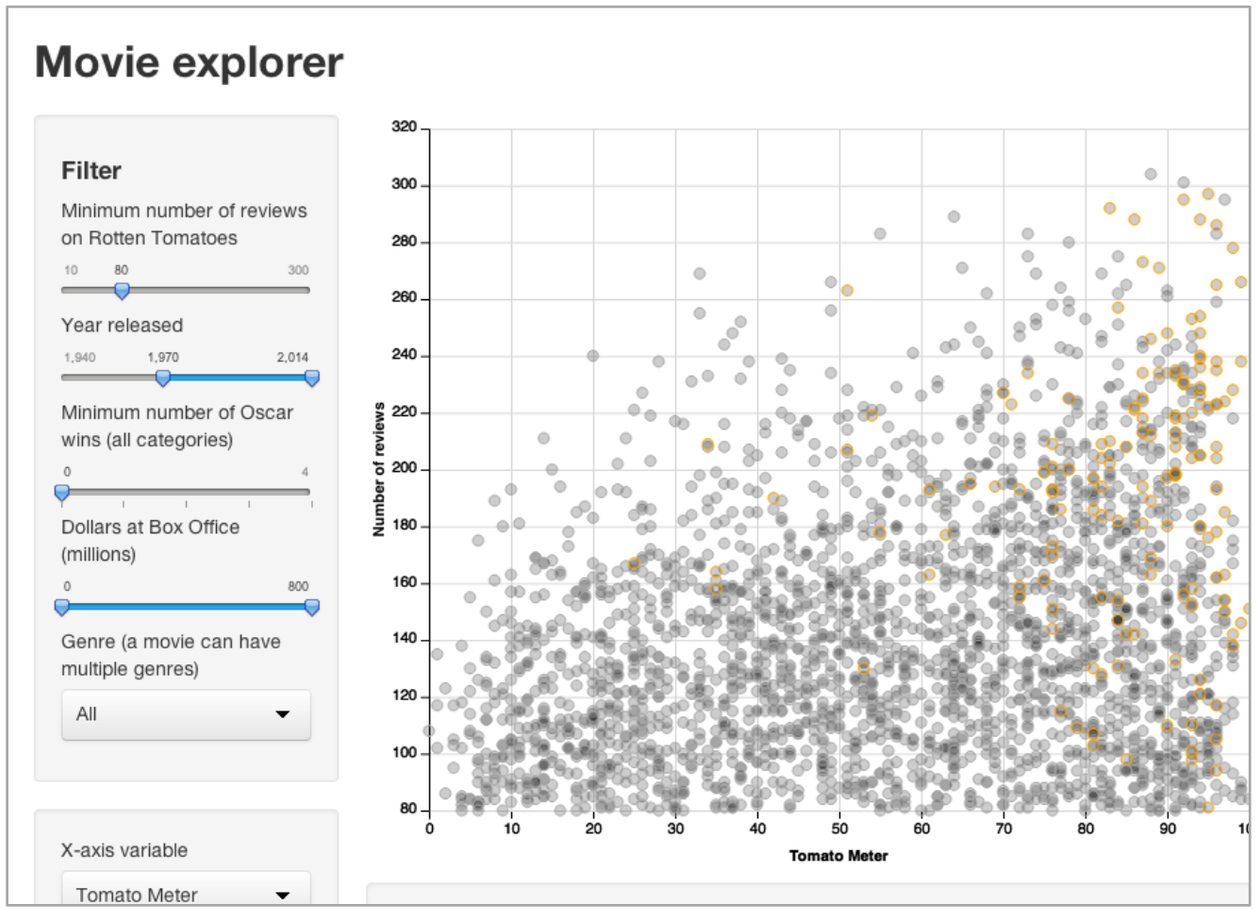
- Notebooks
- Dashboards
- Aplicaciones visuales
- Servicios de datos

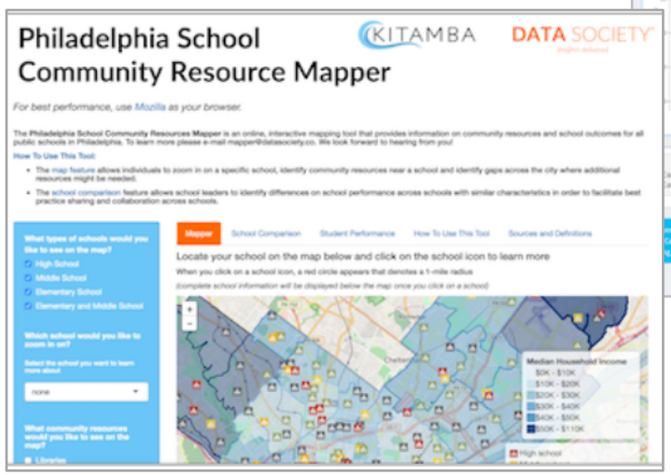






R Shiny





Open Data Science

Visualización

- Gráficos
- Visualización interactiva
- Big data
- Mapas & GIS
- 3D
- Streaming

Bokeh

BeakerX

BeakerX: Beaker extensions for Jupyter

build passing | chat on gitter | JitPack 0.1.1 | npm package 0.0.6 | pypi package 0.2.4.dev0

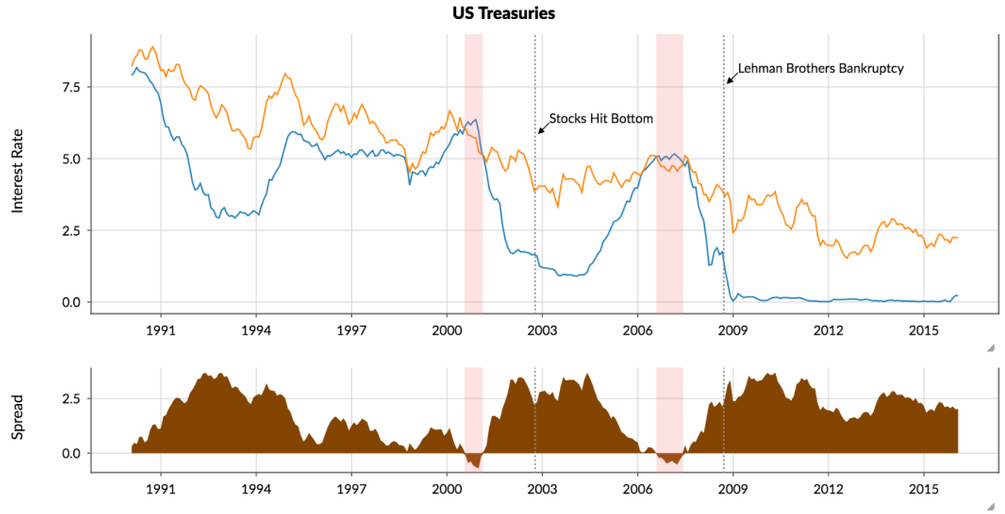
BeakerX is a collection of JVM kernels with widgets, plotting, tables, autotranslation, and other extensions to the Jupyter Notebook and
changes with

The documer

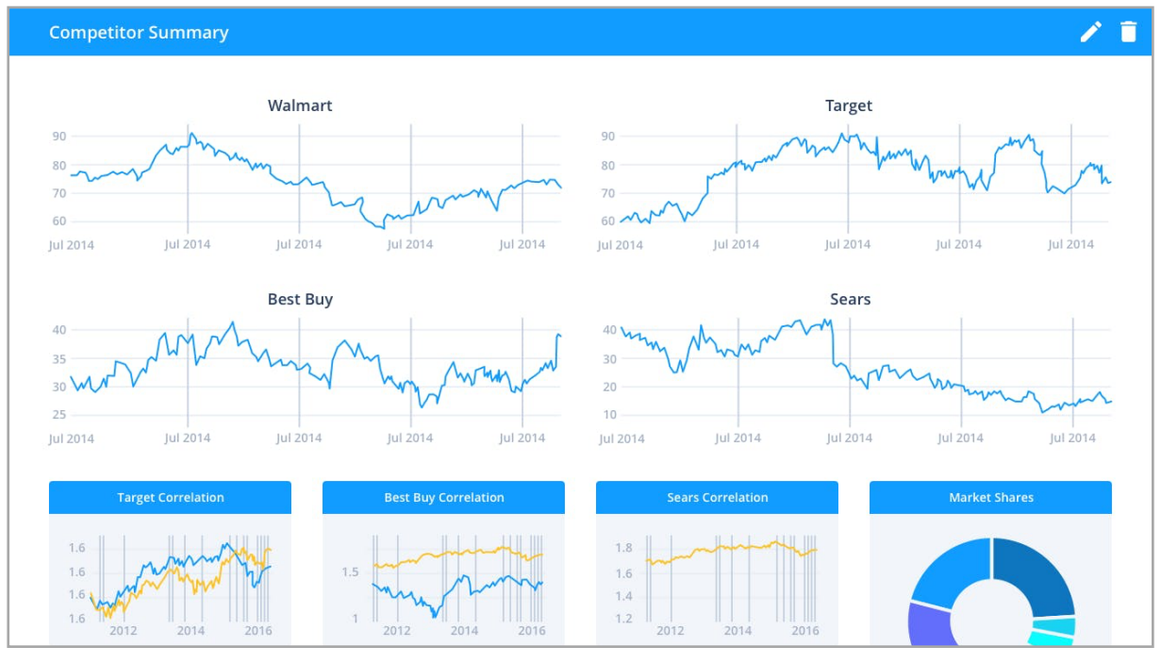
BeakerX is th
are hiring.

Groovy with Interactive Plotting and Tables:

```
// Then use a CombinedPlot to get stacked plots with linked X axis.
def c = new CombinedPlot(title: "US Treasuries", initWidth: 1000)

// add both plots to the combined plot, and including their relative heights.
c.add(p1, 3)
c.add(p2, 1)
```



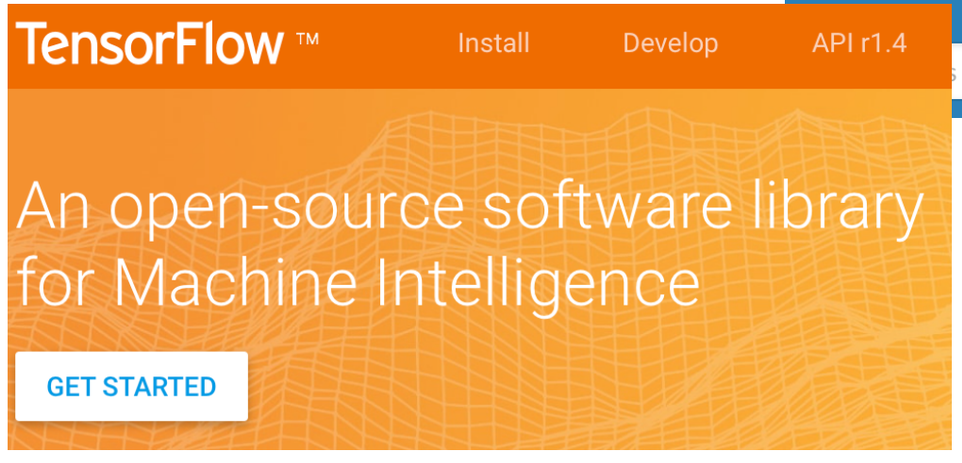
plot.ly



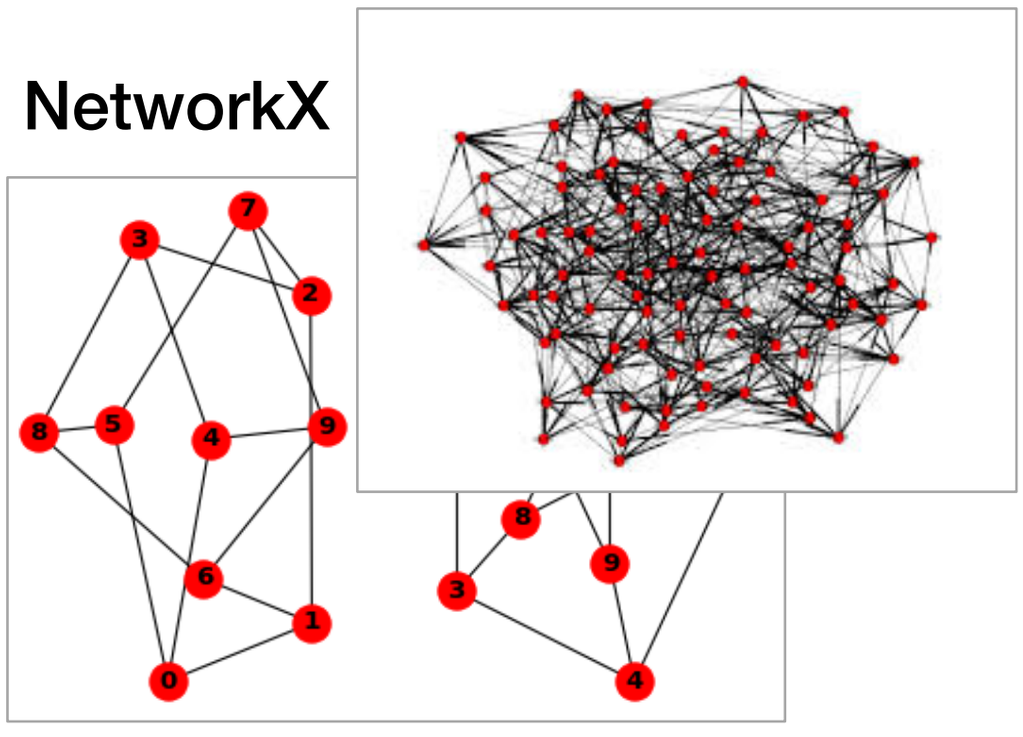
Open Data Science

Analytics

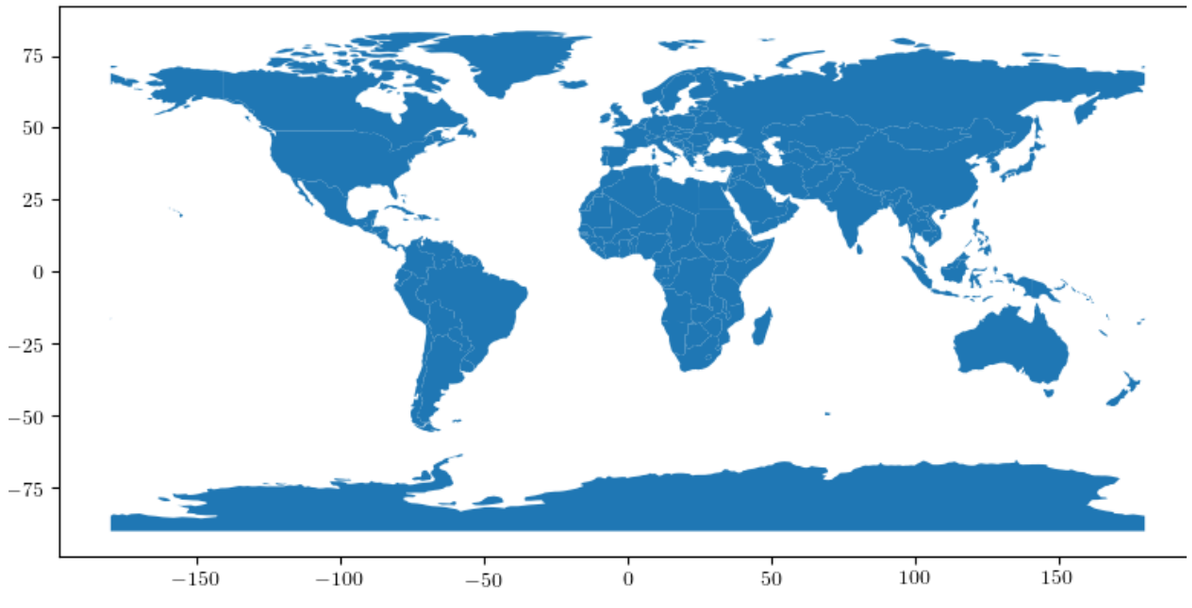
- Preparación de datos
- Estadística
- ML & Ensambles
- Deep learning
- Simulación y optimización
- Datos geoespaciales
- Texto y NLP
- Gráficos y redes
- Minería de audio, video e imágenes



NetworkX



GeoPandas



Open Data Science

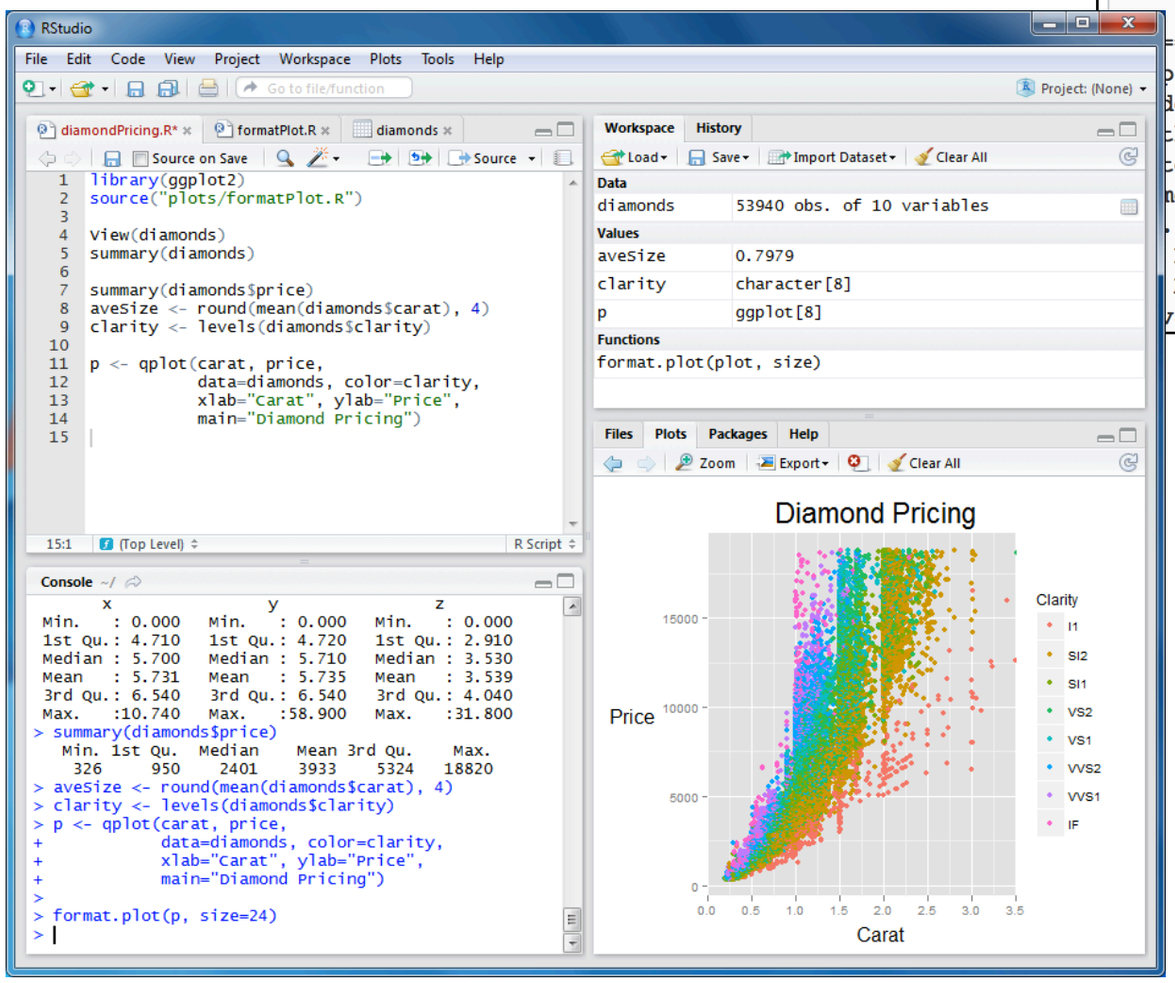
Analytics

- Preparación de datos
- Estadística
- ML & Ensambles
- Deep learning
- Simulación y optimización
- Datos geoespaciales
- Texto y NLP
- Gráficos y redes
- Minería de audio, video e imágenes

Python StatsModels

```
In [1]: import numpy as np

In [2]: import statsmodels.api as sm

In [3]: import statsmodels.formula.api as smf

# Load data
In [4]: dat = sm.datasets.get_rdataset("Guerry", "HistData").data

# Fit regression model (using the natural log of one of the regressors)
In [5]: results = smf.ols('Lottery ~ Literacy + np.log(Pop1831)', data=dat).fit()

# Inspect the results
In [6]: print(results.summary())
                            OLS Regression Results
==============================================================================
Dep. Variable:                Lottery   R-squared:                       0.348
Model:                            OLS   Adj. R-squared:                  0.333
Method:                 Least Squares   F-statistic:                     22.20
Date:                Tue, 28 Feb 2017   Prob (F-statistic):           1.90e-08
Time:                        21:38:05   Log-Likelihood:                -379.82
No. Observations:                  86   AIC:                             765.6
Df Residuals:                      83   BIC:                             773.0
Df Model:                           2
Covariance Type:            nonrobust
```

RStudio



Open Data Science

Analytics

- Preparación de datos
- Estadística
- ML & Ensambles
- Deep learning
- Simulación y optimización
- Datos geoespaciales
- Texto y NLP
- Gráficos y redes
- Minería de audio, video e imágenes

PYOMO

HOME / ABOUT / DOWNLOAD / DOCUMENTATION / BLOG

Documentation

Online Documentation

Pyomo Online Documentation (html, pdf, epub)
PySP Online Documentation (pdf)
Pyomo Wikipedia Page (html)

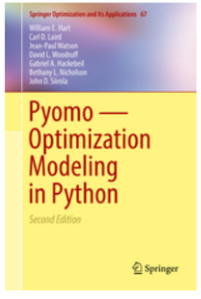
Examples

Pyomo Gallery (browse)
Online examples from the Pyomo software repository: (browse) (zipfile)

Citation

If you use Pyomo for your work, please cite the Pyomo book (bibtex) and the Pyomo paper (bibtex).
If you use PySP for your work, please cite the PySP paper (bibtex).

The Pyomo Book



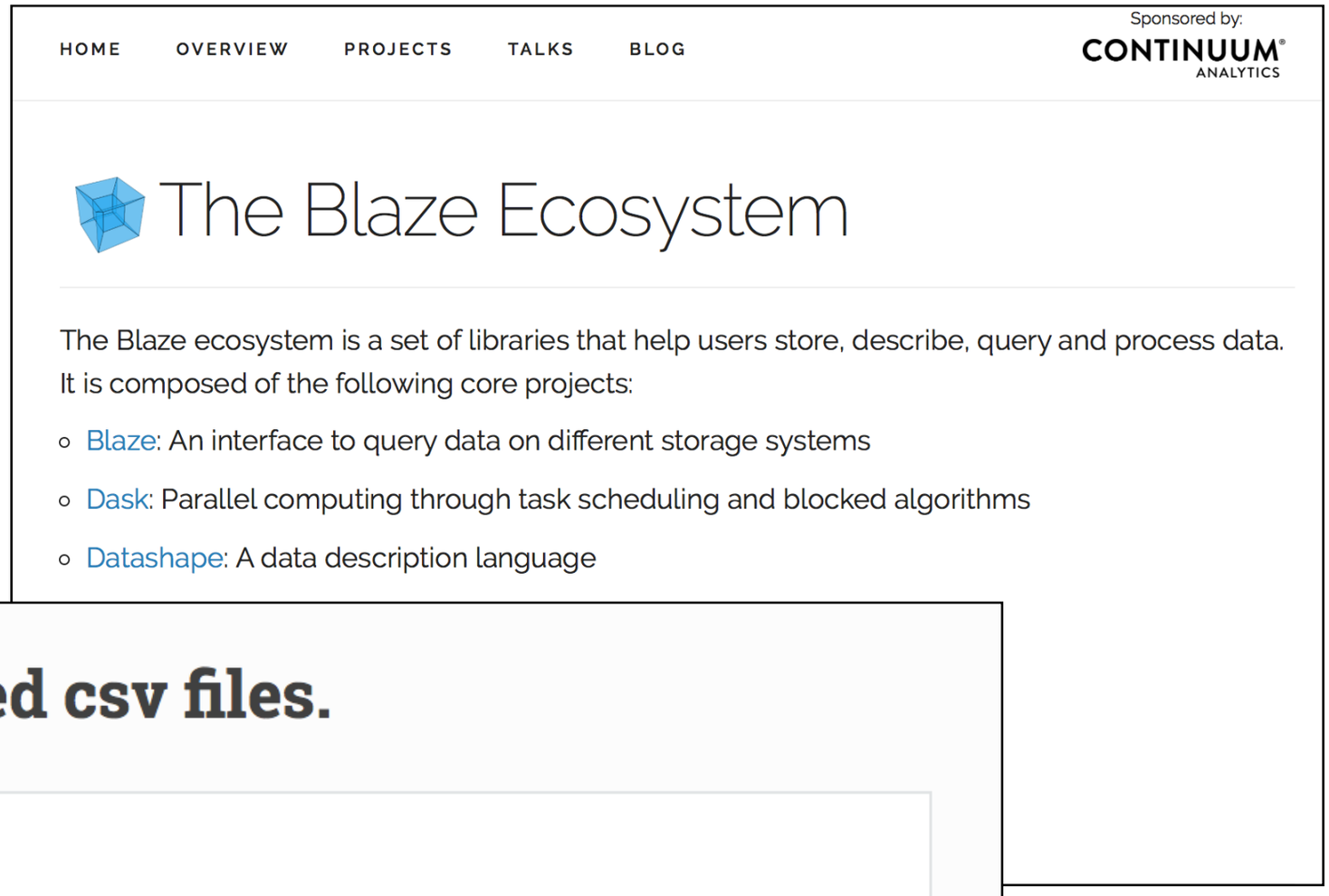
Hart, William E., Carl D. Laird, Jean-Paul Watson, David L. Woodruff, Gabriel A. Hackebeil, Bethany L. Nicholson, and John D. Siirola. *Pyomo – Optimization Modeling in Python*. Second Edition. Vol. 67. Springer, 2017.

The Second Edition of the book describes capabilities in the Pyomo 5.x series. The First Edition (2012) describes the capabilities from the Coopr 3.1 release. Some changes beginning in the Pyomo 4.0 release are not backwards compatible with the First Edition.

Open Data Science

Fuentes de datos modernas

- Big Data
- Spark
- NoSQL
- DW & SQL
- Archivos y servicios Web



Blaze / Odo

Combining separate, gzipped csv files.

```
>>> from blaze import odo
>>> from pandas import DataFrame
>>> odo(example('accounts_*.csv.gz'), DataFrame)
   id      name  amount
0   1     Alice     100
1   2       Bob     200
2   3   Charlie     300
3   4       Dan     400
4   5     Edith     500
```

Open Data Science

Fuentes de datos modernas

- Big Data
- Spark
- NoSQL
- DW & SQL
- Archivos y servicios Web

```python
import sqlite3
conn = sqlite3.connect('example.db')

c = conn.cursor()
c.execute('''
          CREATE TABLE person
          (id INTEGER PRIMARY KEY ASC, name varchar(250) NOT NULL)
          ''')
c.execute('''
          CREATE TABLE address
          (id INTEGER PRIMARY KEY ASC, street_name varchar(250), street_number varchar(250),
           post_code varchar(250) NOT NULL, person_id INTEGER NOT NULL,
           FOREIGN KEY(person_id) REFERENCES person(id))
          ''')

c.execute('''
          INSERT INTO person VALUES(1, 'pythoncentral')
          ''')
c.execute('''
          INSERT INTO address VALUES(1, 'python road', '1', '00000', 1)
          ''')

conn.commit()
```

SQLite

```python
import sqlite3
conn = sqlite3.connect('example.db')

c = conn.cursor()
c.execute('SELECT * FROM person')
print c.fetchall()
c.execute('SELECT * FROM address')
print c.fetchall()
conn.close()
```

Open Data Science

Explotación de HW moderno

- Servidores
- Clusters
- GPUs & Workstations

Numba — https://numba.pydata.org

ipyparallel — https://github.com/ipython/ipyparallel

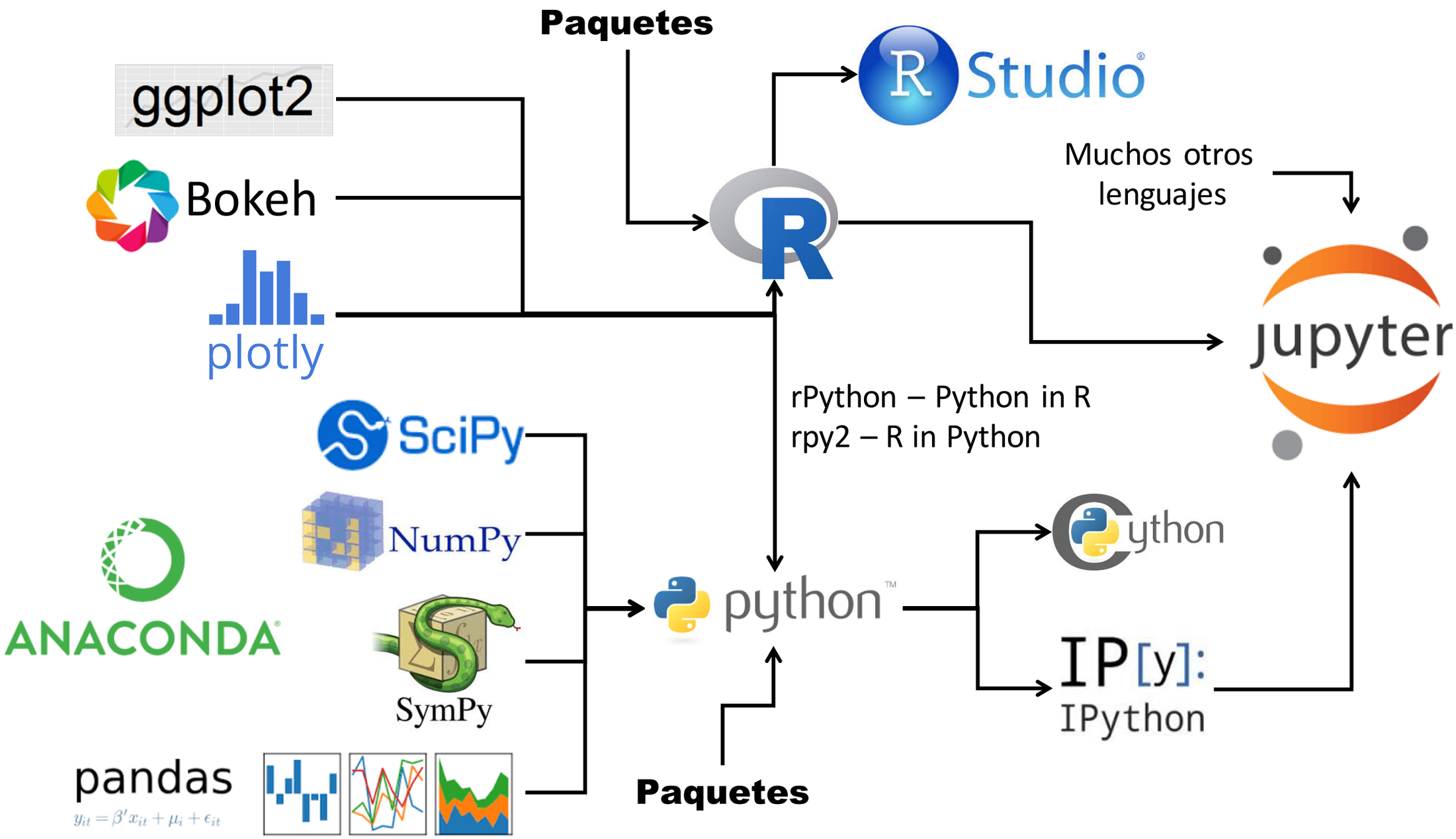
mpi4py — http://pythonhosted.org/mpi4py/

Theano — http://deeplearning.net/software/theano/

pyCUDA —https://mathema.tician.de/software/pycuda/

```python
from numba import jit
from numpy import arange

# jit decorator tells Numba to compile this function.
# The argument types will be inferred by Numba when function is called.
@jit
def sum2d(arr):
    M, N = arr.shape
    result = 0.0
    for i in range(M):
        for j in range(N):
            result += arr[i,j]
    return result

a = arange(9).reshape(3,3)
print(sum2d(a))
```

Open Data Science



Open Data Science

```
echo "ESTACION;FECHA;ANO;MES;DIA;HORA;HHMMSS;DIRECCION;VELOCIDAD" > datos
tail +2 AQUITANIA.csv >> datos

## Elimina lineas vacias
sed -e '/^$/d' datos > out.1

## borra lineas en blanco
sed -e '/;;;;/d' out.1 > datos

## llena las horas vacías
sed -e 's/;;/;00:00:00;/g' datos > out.1

## etcetera …
```

```
## promedio para cada hora
csvsql --query  "select ESTACION, FECHA, ANO, MES,
   DIA, HORA, DIRECCION, avg(VELOCIDAD) as VELOCIDAD from 'out'
   group by ESTACION, FECHA, HORA" out.5 > out.6
```

```
ESTACION;FECHA;HORA;DIRECCION;VELOCIDAD
AQUITANIA;2005-04-16;11:10:00;135;6,3
AQUITANIA;2005-04-16;11:20:00;135;5,1
AQUITANIA;2005-04-16;11:30:00;135;6,3
AQUITANIA;2005-04-16;11:40:00;113;6,1
AQUITANIA;2005-04-16;11:50:00;135;4,1
AQUITANIA;2005-04-16;12:00:00;135;5,5
AQUITANIA;2005-04-16;12:10:00;135;5,4
AQUITANIA;2005-04-16;12:20:00;135;5,5
AQUITANIA;2005-04-16;12:30:00;90;4,6
AQUITANIA;2005-04-16;12:40:00;90;6,7
```

```
ESTACION,FECHA,ANO,MES,DIA,HORA,DIRECCION,VELOCIDAD
AQUITANIA,2005-04-16,2005,4,16,11,135,5.58
AQUITANIA,2005-04-16,2005,4,16,12,90,5.45
AQUITANIA,2005-04-16,2005,4,16,13,135,4.866666666666667
AQUITANIA,2005-04-16,2005,4,16,14,135,3.6666666666666665
AQUITANIA,2005-04-16,2005,4,16,15,135,3.466666666666667
AQUITANIA,2005-04-16,2005,4,16,16,135,3.6999999999999993
AQUITANIA,2005-04-16,2005,4,16,17,135,4.833333333333333
AQUITANIA,2005-04-16,2005,4,16,18,135,4.766666666666667
AQUITANIA,2005-04-16,2005,4,16,19,135,4.350000000000005
AQUITANIA,2005-04-16,2005,4,16,20,135,2.683333333333333
AQUITANIA,2005-04-16,2005,4,16,21,135,3.1999999999999997
```

Analytics

Proceso científico de transformación de datos en conocimiento para mejorar el proceso de toma de decisiones [Informs].

Data Mining

Proceso de descubrimiento de patrones y tendencias útiles en grandes conjuntos de datos.

Data Science

Área relacionada con los procesos y sistemas para la extracción de conocimiento de datos almacenados electrónicamente (¿para la toma de decisiones? ¿para probar hipótesis?)

Aprendizaje estadístico
Aprendizaje de máquinas
Optimización
Simulación
Análisis de decisiones

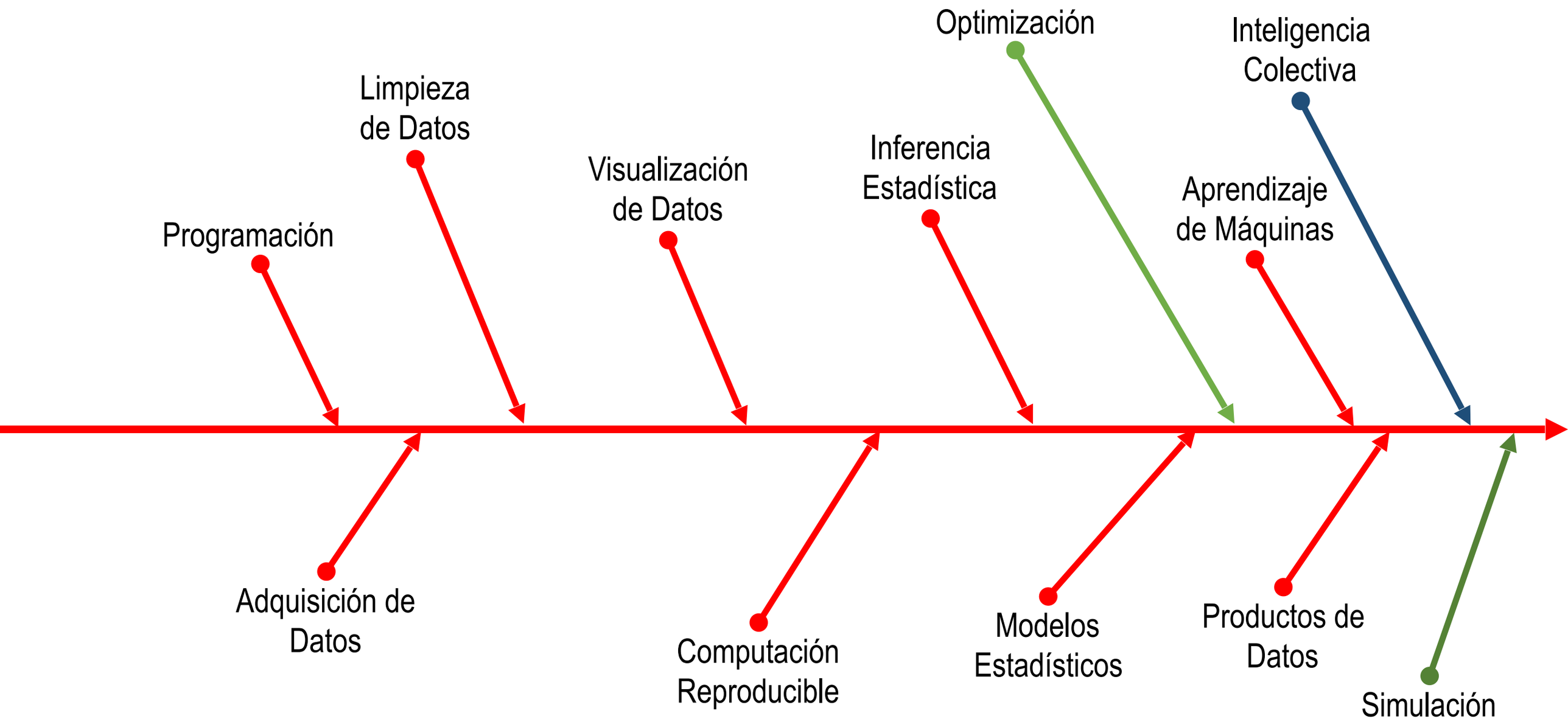
Archivos
Servicios Web
Hadoop & Spark
NoSQL
DW & SQL

Análisis interactivo

Reportes automáticos
Dashboards
Computación reproducible

Infraestructura computacional
- **Un procesador**
  - **Computación en máquinas locales**
- **Muchos procesadores**
  - **Computación en la nube**

Analytics

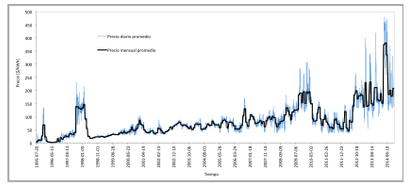


Data-driven decision making!

Analytics

Estadística y aprendizaje de máquinas

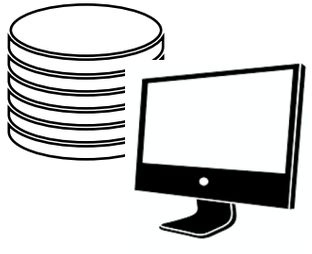
Los datos están listos



Modelado de datos

Inteligencia de Negocios

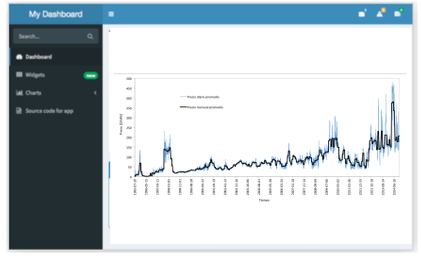
DW / OLAP



Generación, agregación, análisis y visualización de datos del negocio

Minería de Datos

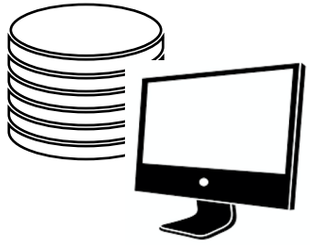
DW / OLAP

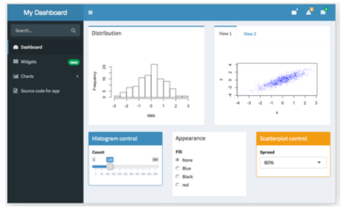


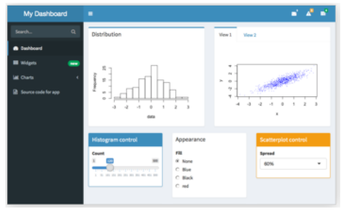
Descubrimiento de patrones y tendencias claves

Analytics

DW / OLAP
Hadoop & Spark
NoSQL …







Producto de Datos

Aplicación que combina datos con algoritmos para inferencia, predicción u optimización para generar más datos e información valiosa.

- Aprendizaje a partir de los datos.
- Auto-adaptación
- Ampliamente aplicable.

Big Data Analytics

MLib
Machine Learning

Mahout vs MLib

Collaborative Filtering *with CLI drivers*

User-Based Collaborative Filtering

Item-Based Collaborative Filtering

Matrix Factorization with ALS

Matrix Factorization with ALS on Implicit Feedback

Weighted Matrix Factorization, SVD++

Classification *with CLI drivers*

Logistic Regression - trained via SGD

Naive Bayes / Complementary Naive Bayes

Hidden Markov Models

Clustering *with CLI drivers*

Canopy Clustering

k-Means Clustering

Fuzzy k-Means

Streaming k-Means

Spectral Clustering

Dimensionality Reduction *note: most scala reduction algorithms are available through th* Core Library for all engines

Singular Value Decomposition

Lanczos Algorithm

Stochastic SVD

PCA (via Stochastic SVD)

QR Decomposition

Topic Models

Latent Dirichlet Allocation

Miscellaneous

RowSimilarityJob

Collocations

Sparse TF-IDF Vectors from Text

XML Parsing

Email Archive Parsing

Evolutionary Processes

Algorithms

MLlib contains many algorithms and utilities.

ML algorithms include:

- Classification: logistic regression, naive Bayes,...
- Regression: generalized linear regression, survival regression,...
- Decision trees, random forests, and gradient-boosted trees
- Recommendation: alternating least squares (ALS)
- Clustering: K-means, Gaussian mixtures (GMMs),...
- Topic modeling: latent Dirichlet allocation (LDA)
- Frequent itemsets, association rules, and sequential pattern mining

ML workflow utilities include:

- Feature transformations: standardization, normalization, hashing,...
- ML Pipeline construction
- Model evaluation and hyper-parameter tuning
- ML persistence: saving and loading models and Pipelines

Other utilities include:

- Distributed linear algebra: SVD, PCA,...
- Statistics: summary statistics, hypothesis testing,...

Hacia una visión unificada de Data Science, Analytics y Big Data

Esta presentación describe, bajo un marco común, los conceptos fundamentales de Data Science, Analytics y Big Data y establece su similitudes y diferencias.

Descargue la última versión de este documento de:
https://github.com/jdvelasq/data-science-docs/blob/master/ds-analytics-bigdata.pdf

JUAN DAVID VELÁSQUEZ HENAO, MSc, PhD
Profesor Titular
Departamento de Ciencias de la Computación y la Decisión
Facultad de Minas
Universidad Nacional de Colombia, Sede Medellín

jdvelasq@unal.edu.co
@jdvelasquezh
https://github.com/jdvelasq
https://goo.gl/prkjAq
https://goo.gl/vXH8jy

Definiciones

1973 Demand side management

1974 Primer sistema patentado de monitoreo con sensores remotos

1981 Privatización mercado eléctrico de Chile

1985 Control de la demanda basado en precio

1990 Privatización de la generación de UK

1998 Popularización del concepto de Smart Home

2004 Medidores de consumo de electricidad por intervalos en el marcado Australiano

2005 Medidores inteligentes para electricidad y gas en Países Bajos

2006 Medidores inteligente en Nueva Zelanda

2007 Se usa el término Smart Grid por primera vez

2008 Medidores inteligentes en UK

2010 Control de la demanda basado en tarifas en tiempo real

Modelos de planificación centralizada del mercado
Modelos de predicción de la demanda

Modelos de decisión individual
Modelos de predicción del precio

2006 Medidores inteligentes de consumo en Ontario

Disciplina	Tecnología	Habilidades	Foco
Inteligencia de Negocios	• ETL/SQL • RDBMS • Reportes • Visualización	• Programación • Análisis de datos • Modelado de datos • Desarrollo de reportes • Estadística Básica • Análisis del negocio & Estrategia • Presentación oral	• Suministro de información y reporte • Visualización de datos • Estadísticos descriptivos • Integración de datos y consolidación
Análisis de datos	• Software para modelado de datos • Software para diagramación • Software para documentación • SQL • Software para perfilado de datos	• Modelado de datos • Análisis del negocio • Manipulación de datos • Estadística básica	• Reglas de negocio • Definición de datos • Relaciones entre datos • Atributos de datos • Estructuras de datos • Fuentes y usos de datos • Calidad de datos
Ciencia de los Datos (Analytics)	• Software estadístico • Datos columnares • Map-Reduce • NoSQL • Lenguajes de programación • Software para graficación • Software para optimización, simulación, predicción y análisis de decisiones	• Estadística avanzada • Programación • Análisis del negocio • Arquitecturas y tecnologías modernas para el manejo de datos • Desarrollo de productos de datos • Simulación de sistemas • Optimización • Predicción	• Modelado predictivo • Análisis estadístico avanzado • Minería de datos • Manejo de datos no estructurados • Manejo de grandes volúmenes de datos • I+D • Análisis de decisiones

Infraestructura computacional

1960s

Mainframes

Primeros sistemas de computo a gran escala

1969

UNIX

Sistema operativo portable, multiusuario y multitarea, desarrollado en laboratorios Bell de T&T

1981

IBM PC

Primer computador personal

1983

MS-DOS 1.0

Sistema operativo de Microsoft para PC

1991

Linux

Proyecto para crear un sistema operativo similar a UNIX de distribución libre basado en herramientas GNU

1992

Windows 3.1

Apple PowerBook

Laptop de Apple

1994

Yahoo

1998

google.com

2004

Facebook

2005

YouTube

2016

Bash on Ubuntu on Windows

Projecto GNU

Proyecto para herramientas de distribución libre

Data Science and Data Scientists: What's in a Name?
Saunders, 2013

Data Architect
Data Engineer
Diseño y estructura de las bases de datos.

Data Manager
Gestiona la creación y mantenimiento de las bases de datos.

ETL Developer
Gestiona la extracción, transformación y carga de los datos a las bases de datos.

Data Analyst
Fuentes y usos de los datos.

Business Intelligence Practitioner
Combinación de negocios + tecnología con el fin de proveer información a las unidades de negocios para toma de decisiones

Data Scientist
Habilidades en la programación de computadores para manejo de datos y modelado predictivo (estadística, aprendizaje de máquinas, minería de datos, etc.).

Analytics Practitioner
Data Science + Optimización + Simulación