

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

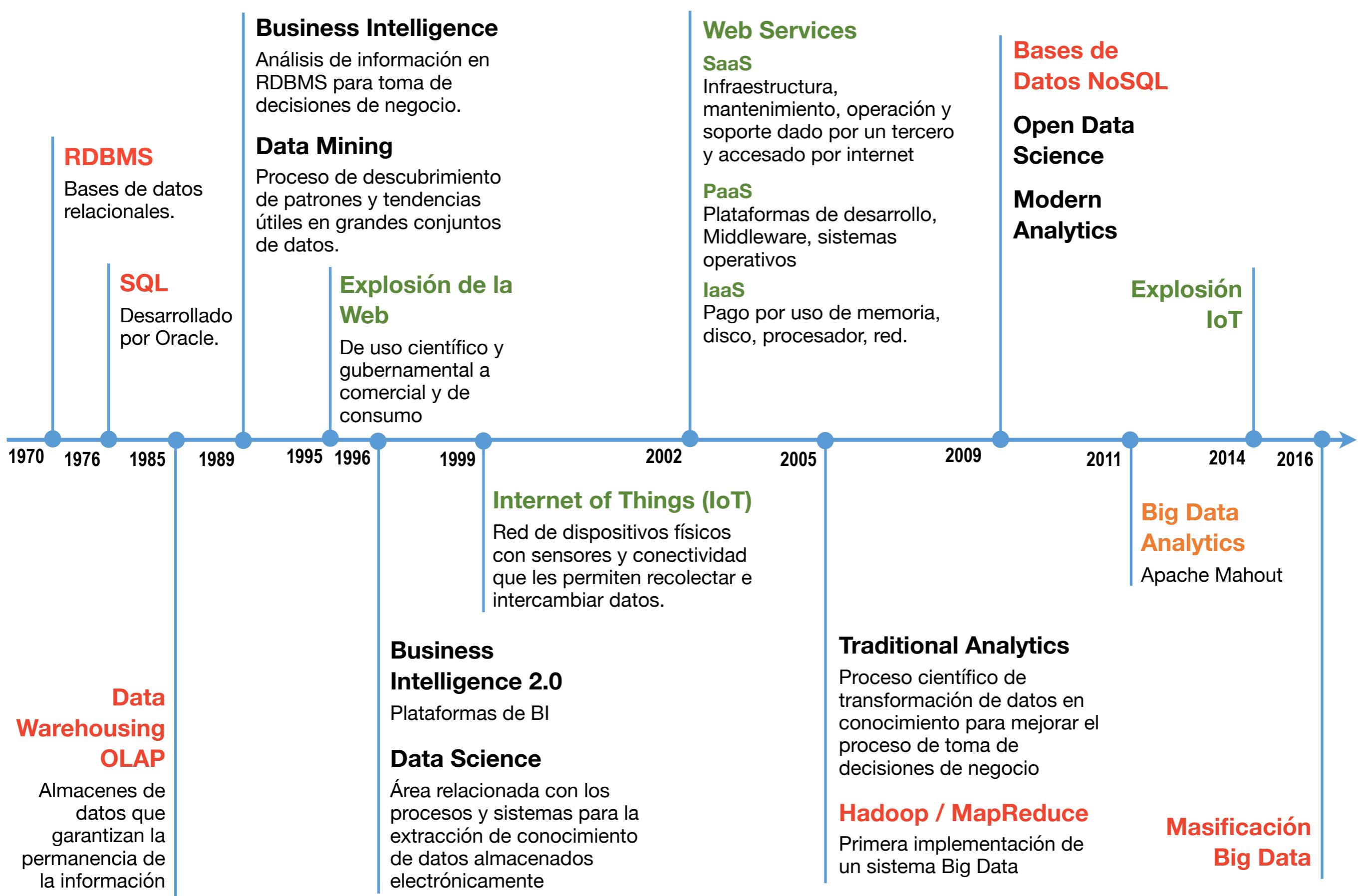
 jdvelasq@unal.edu.co

 @jdvelasquezh

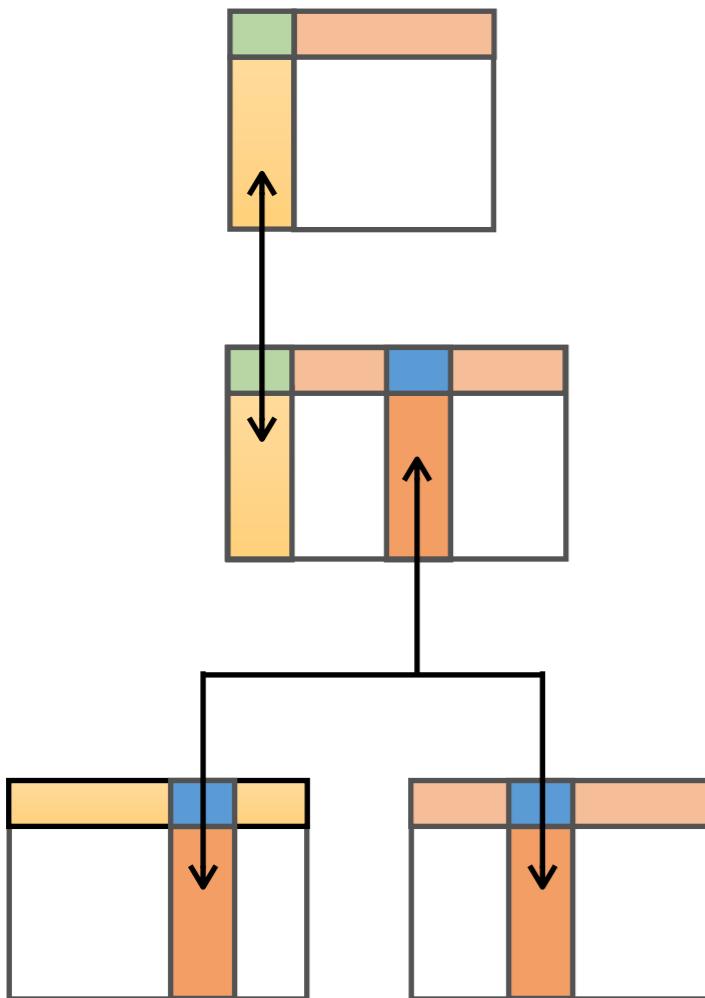
 <https://github.com/jdvelasq>

 <https://goo.gl/prkjAq>

 <https://goo.gl/vXH8jy>



RDBMS – 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
- Microsoft 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);

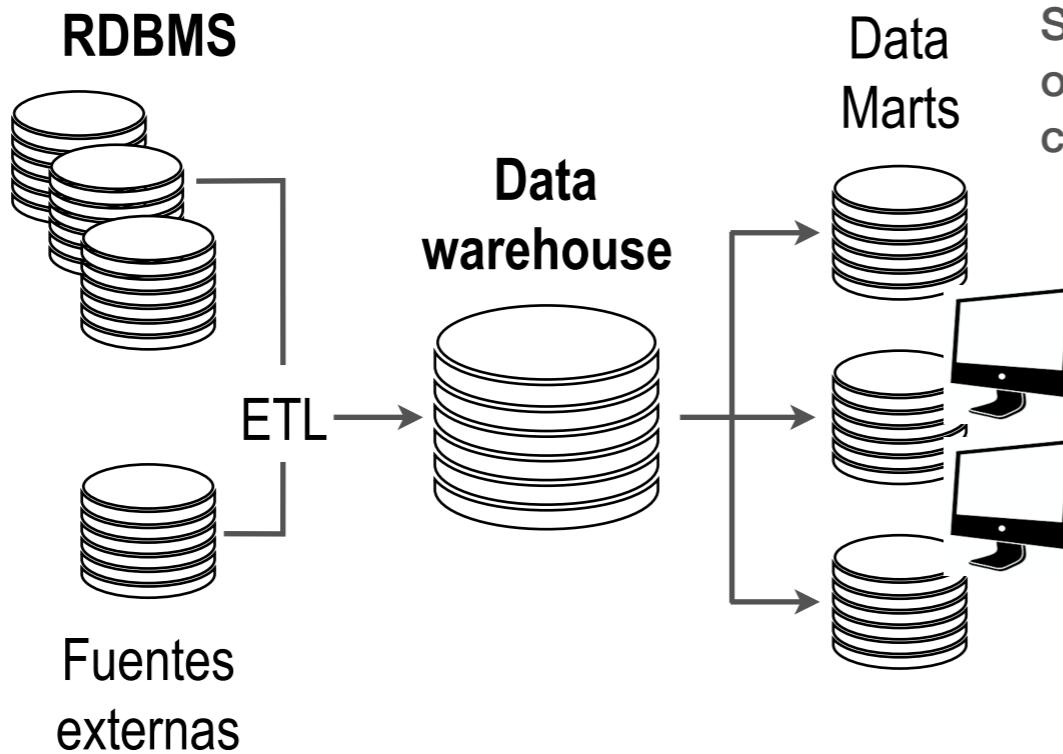
INSERT INTO phone_book2 ( [name], [phoneNumber] )
SELECT [name], [phoneNumber]
FROM phone_book
WHERE name IN ('John Doe', 'Peter Doe')

DELETE FROM tabla WHERE columnal = 'valor1';
```

Data Warehouse

ETL

- Extraction
- Transformation
- Load



Data Mart

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

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

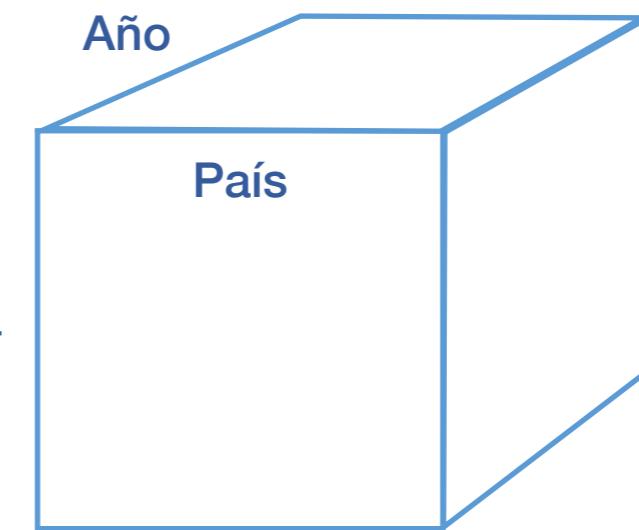
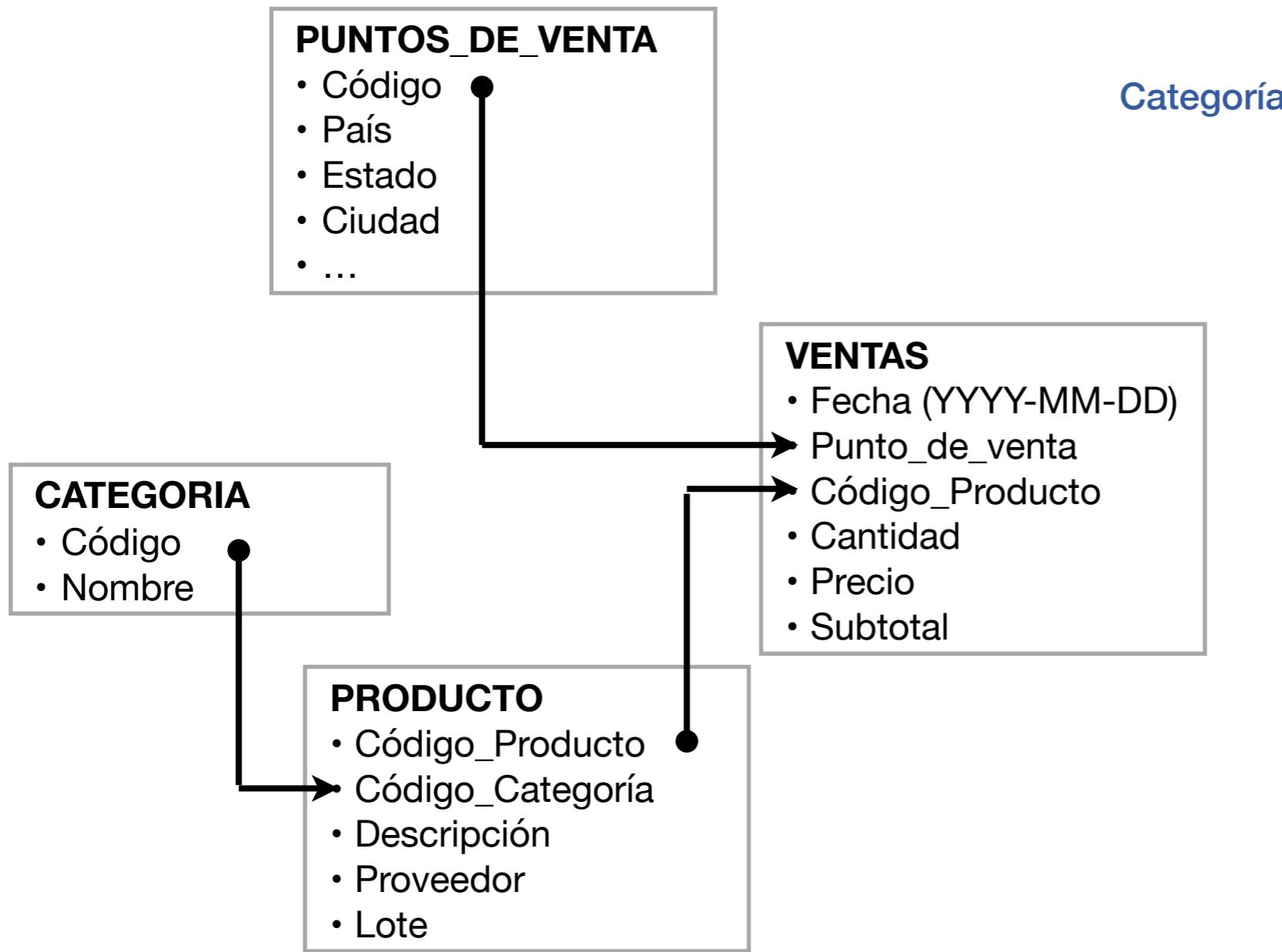
Data Warehouse

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



The screenshot shows the homepage of GENSCAPE. At the top, there is a navigation bar with links for Solutions, Knowledge Center, Events, Blog, News, and About. Below the navigation bar, there is a horizontal menu with colored buttons for Oil, Power, Natural Gas, Maritime, Agriculture & Biofuels, and Petrochemical & NGLs. The main content area is divided into two columns. The left column contains links for Overview, Daily Macro Supply & Demand Data Report, Equity Production Insight, Intrastate Storage Monitoring, and Natural Gas Analyst. The right column contains links for Natural Gas Daily Mexico Exports Monitor, Natural Gas Forward Supply & Demand Report, Natural Gas Infrastructure Intelligence, Natural Gas Notices & Maintenance, and Natural Gas Production Forecast.

The screenshot shows the Energy Dashboard of energyone. The top navigation bar includes links for HOME, ABOUT US, MARKETS SERVED, PRODUCTS, SERVICES, INVESTORS, and CONTACT US. To the right of the navigation bar, there is a 'FEATURES' section with a list of capabilities. Below the navigation bar, the main dashboard area displays a complex grid of data visualizations, including maps, charts, and graphs, representing various energy markets and operational metrics. The bottom right corner of the dashboard features the energyone logo.

FEATURES

The EnergyDashboard enables managers of wholesale energy portfolios to see and manage "at a glance" ALL the key features, status and requirements of all their wholesale energy operations, such as:

- Market data, prices and chosen analytics
- Portfolio status (bid compliance in multiple markets)
- Contracted position
- Energy operations (workflow) status and alerts

Furthermore, users can easily switch between the various aspects of the operational functions, seamlessly moving between market operations and bidding and contracts.

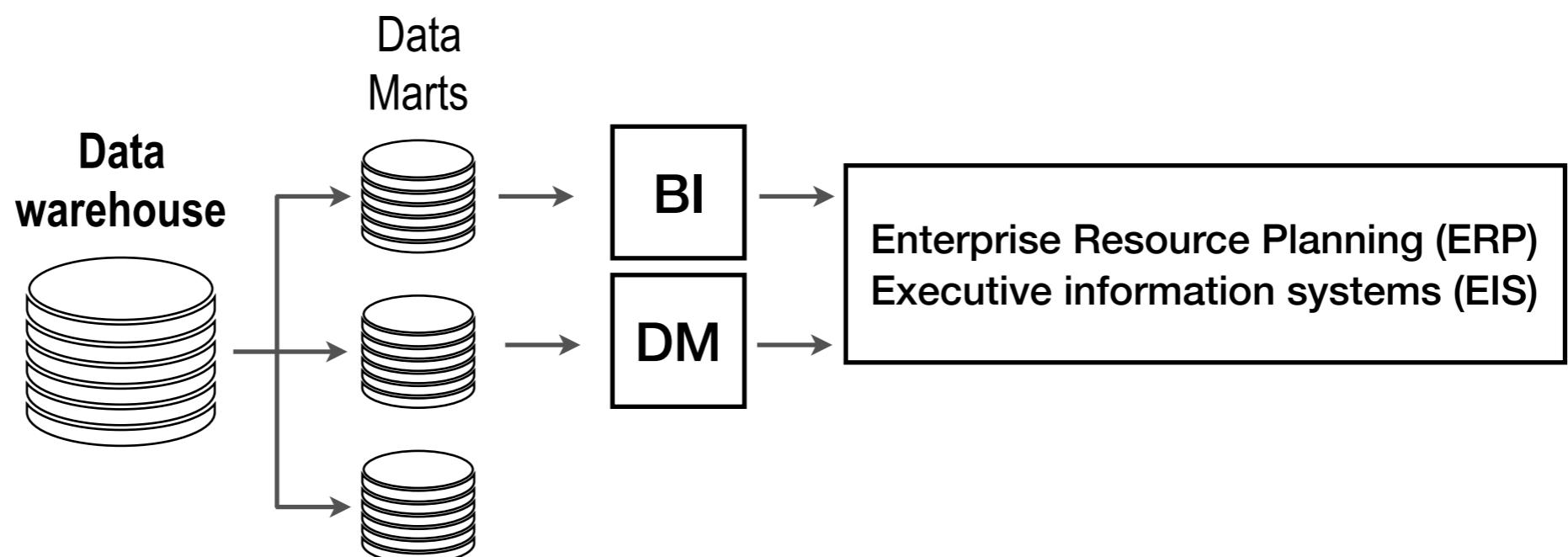
CONTACT US

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.

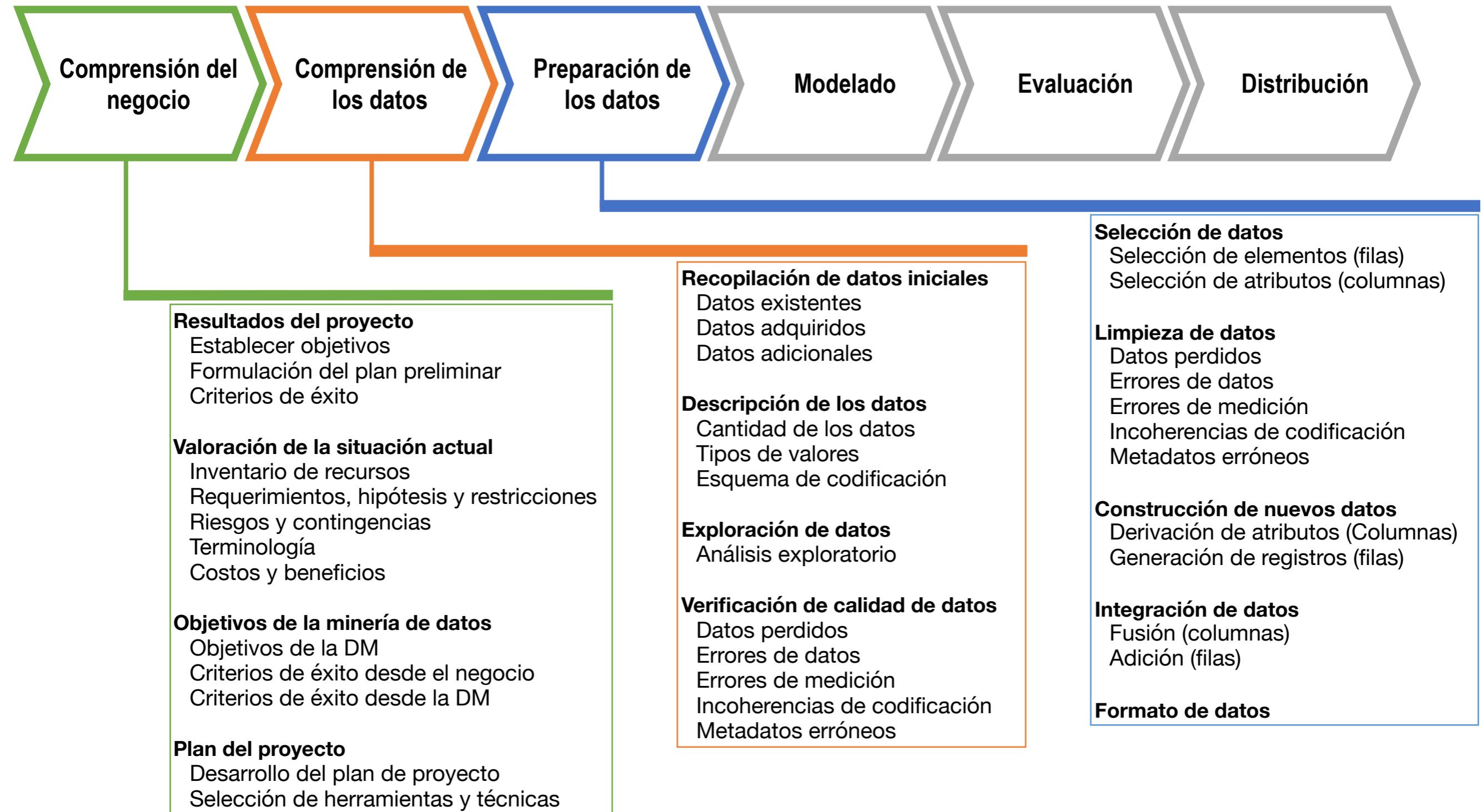
Tareas típicas:

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



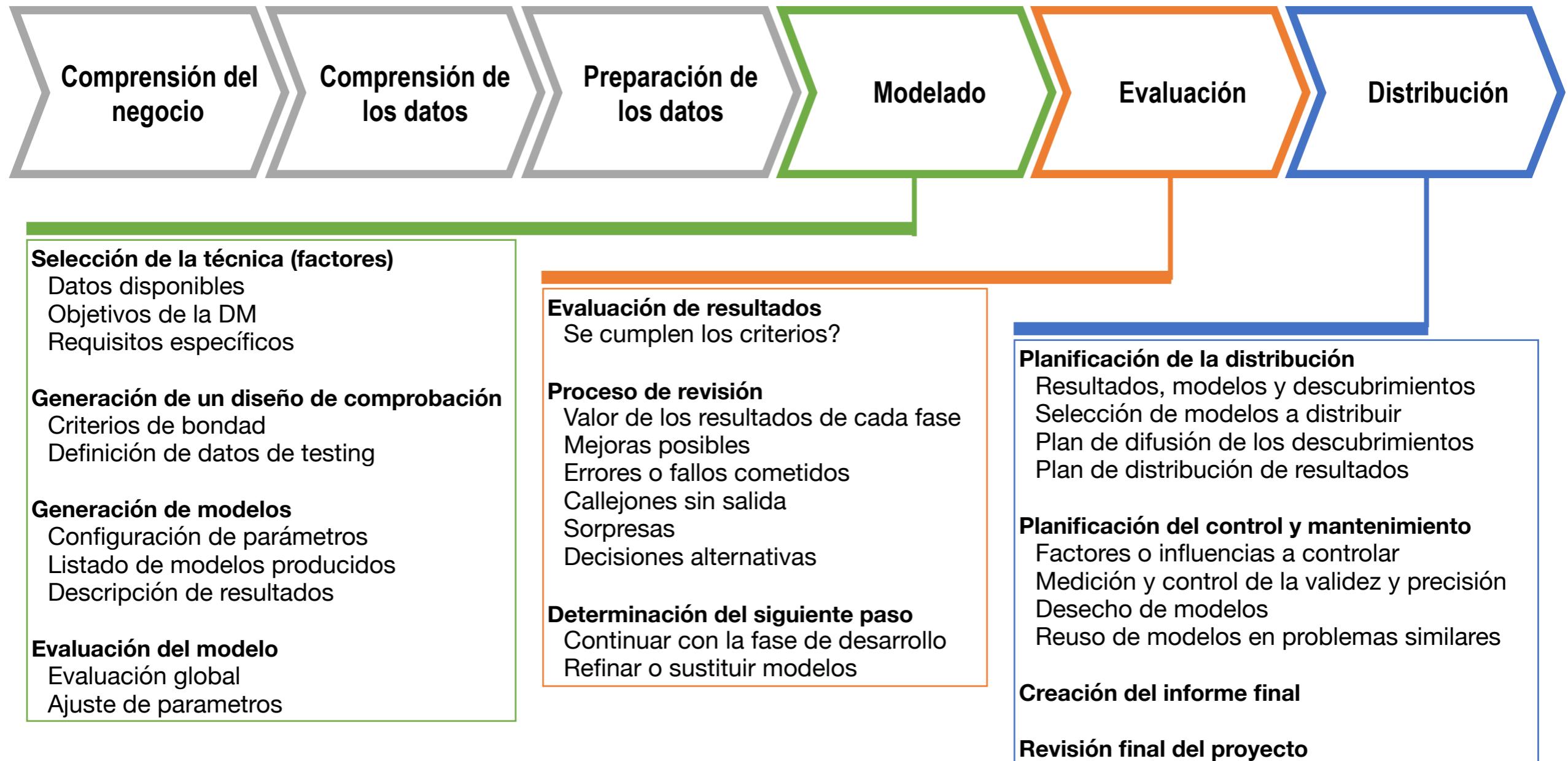
Data Mining

Metodología CRISP-DM



Data Mining

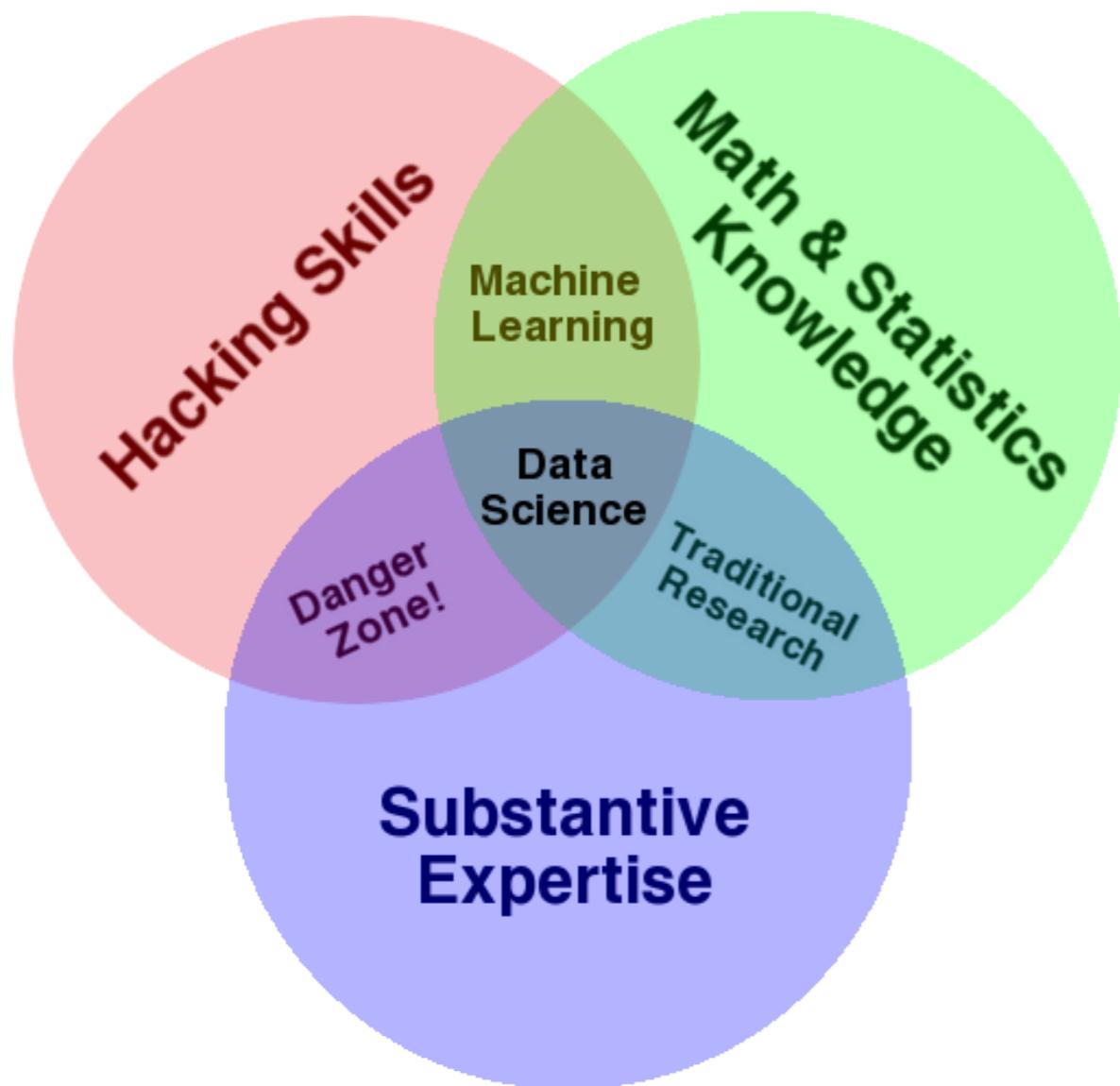
Metodología CRISP-DM



Data Mining

Disciplina	Tecnología	Habilidades	Foco
Análisis de datos	<ul style="list-style-type: none">Software para modelado de datosSoftware para diagramaciónSoftware para documentaciónSQLSoftware para perfilado de datos	<ul style="list-style-type: none">Modelado de datosAnálisis del negocioManipulación de datosEstadística básica	<ul style="list-style-type: none">Reglas de negocioDefinición de datosRelaciones entre datosAtributos de datosEstructuras de datosFuentes y usos de datosCalidad de datos
Inteligencia de Negocios	<ul style="list-style-type: none">ETL/SQLRDBMSReportesVisualización	<ul style="list-style-type: none">ProgramaciónAnálisis de datosModelado de datosDesarrollo de reportesEstadística BásicaAnálisis del negocio & EstrategiaPresentación oral	<ul style="list-style-type: none">Suministro de información y reporteVisualización de datosEstadísticos descriptivosIntegración de datos y consolidación
Data Mining	<ul style="list-style-type: none">Software estadísticoHerramientas de aprendizaje de máquinasLenguajes de programación	<ul style="list-style-type: none">ProgramaciónModelado de datosEstadística AvanzadaPresentación oral	<ul style="list-style-type: none">Análisis estadístico avanzadoManejo de grandes volúmenes de datosVisualización de datosModelos 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

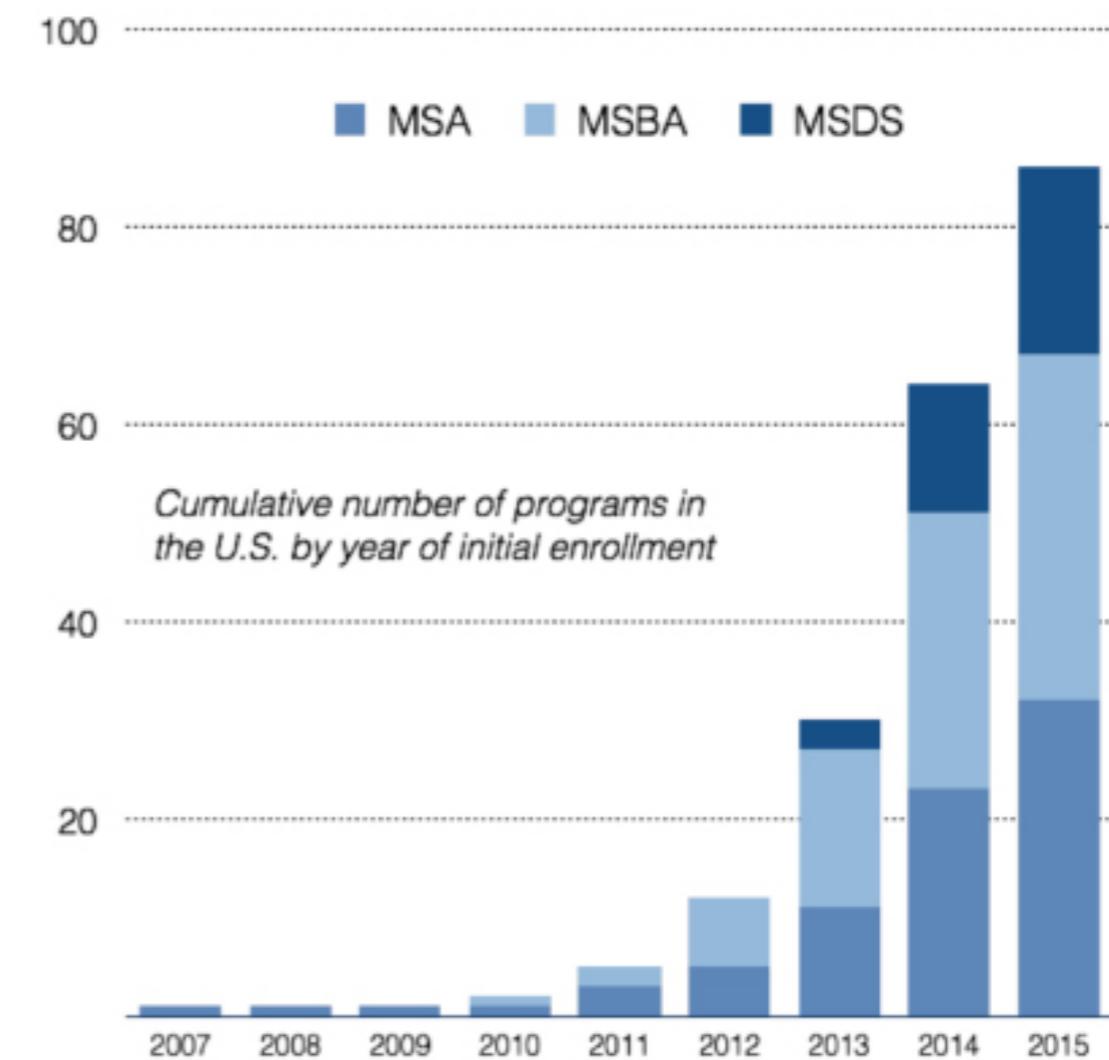
Best Job in America
for 2016

Sources: [25 Best Jobs in America](#) and [25 Highest Paying Jobs in America for 2016](#)

Data Science

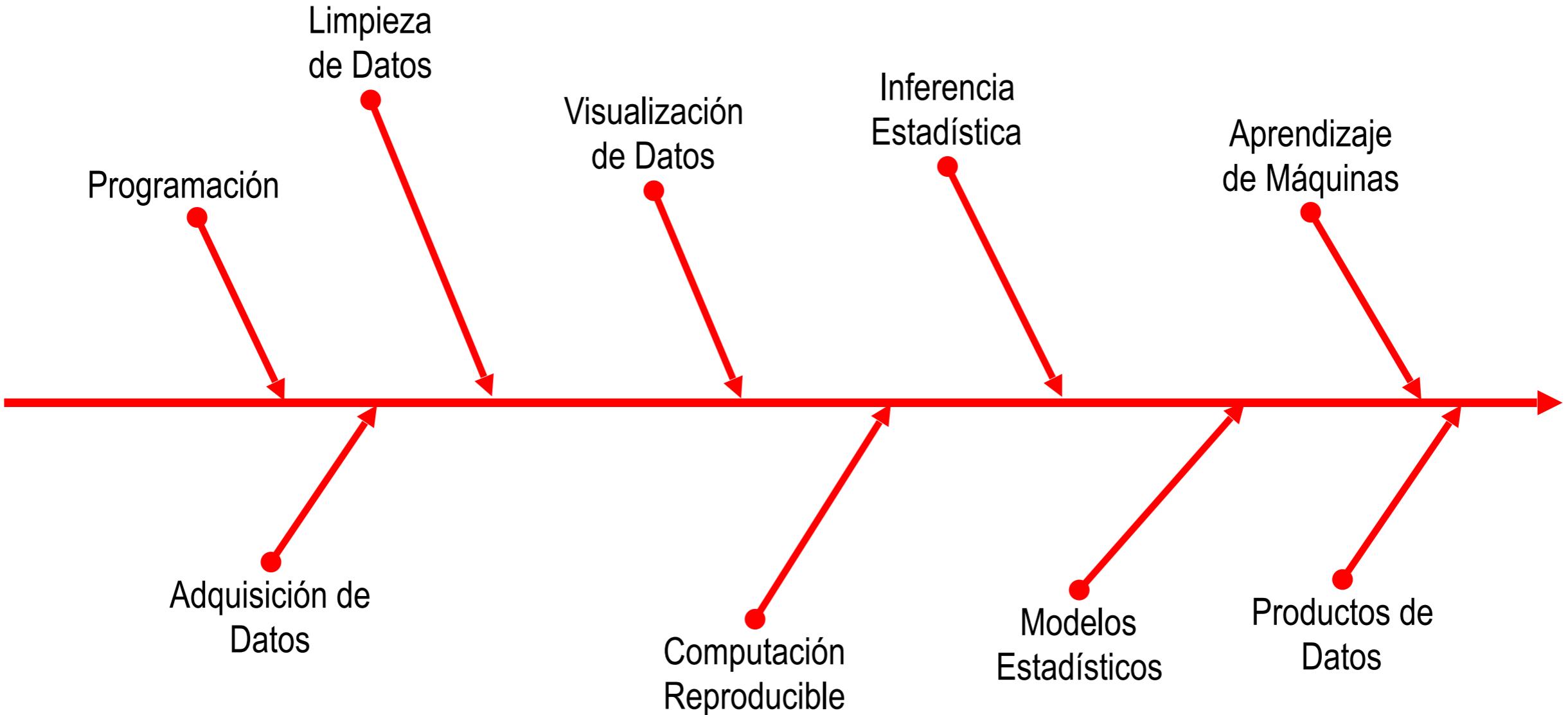


GROWTH OF MASTER'S DEGREE PROGRAMS IN ANALYTICS AND DATA SCIENCE



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

Data Science

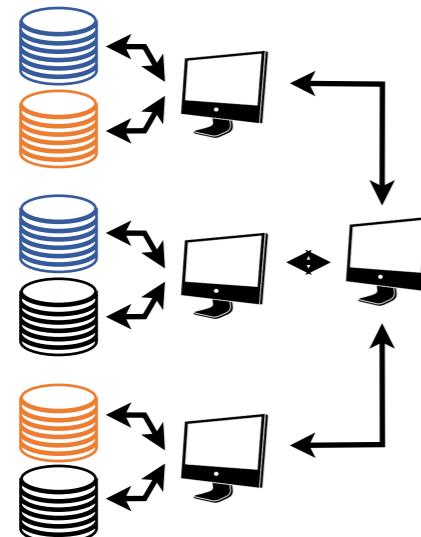


Data-driven decision making

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.

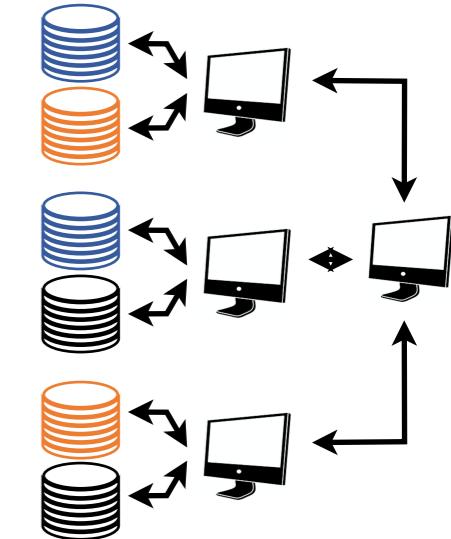
Platform as a Service (PaaS)

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.



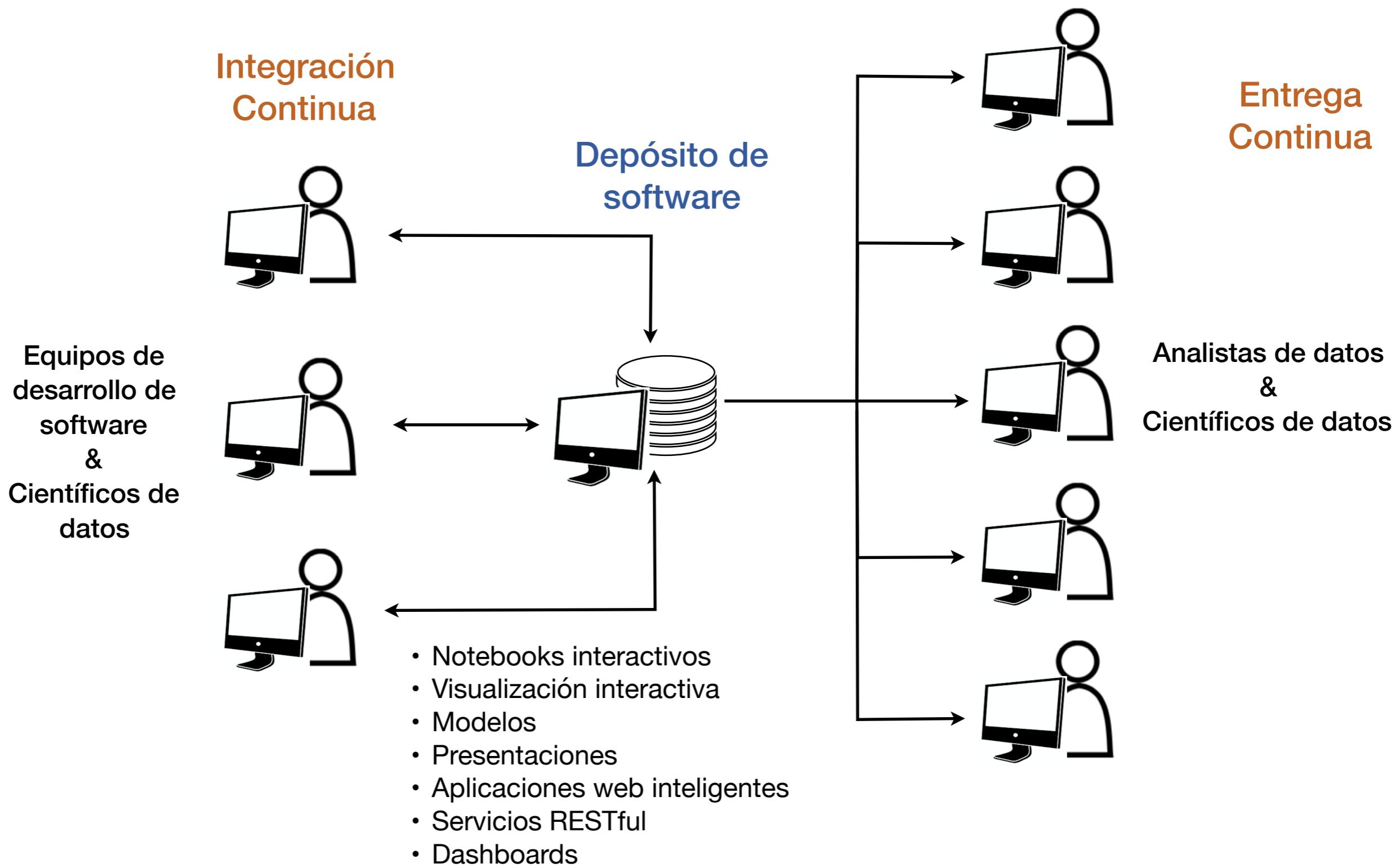
Nube

Internet

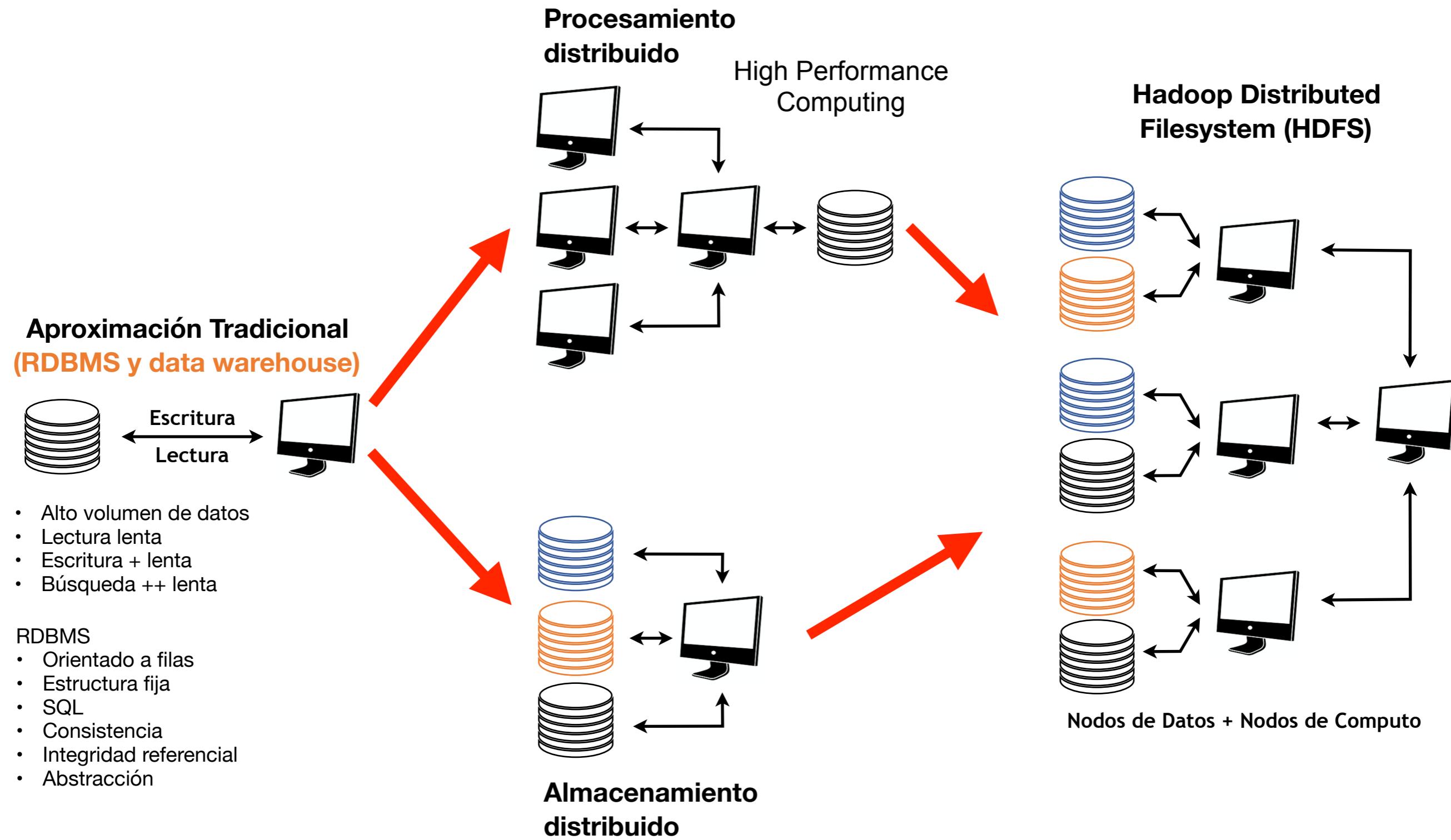


Máquina Local (Cliente)

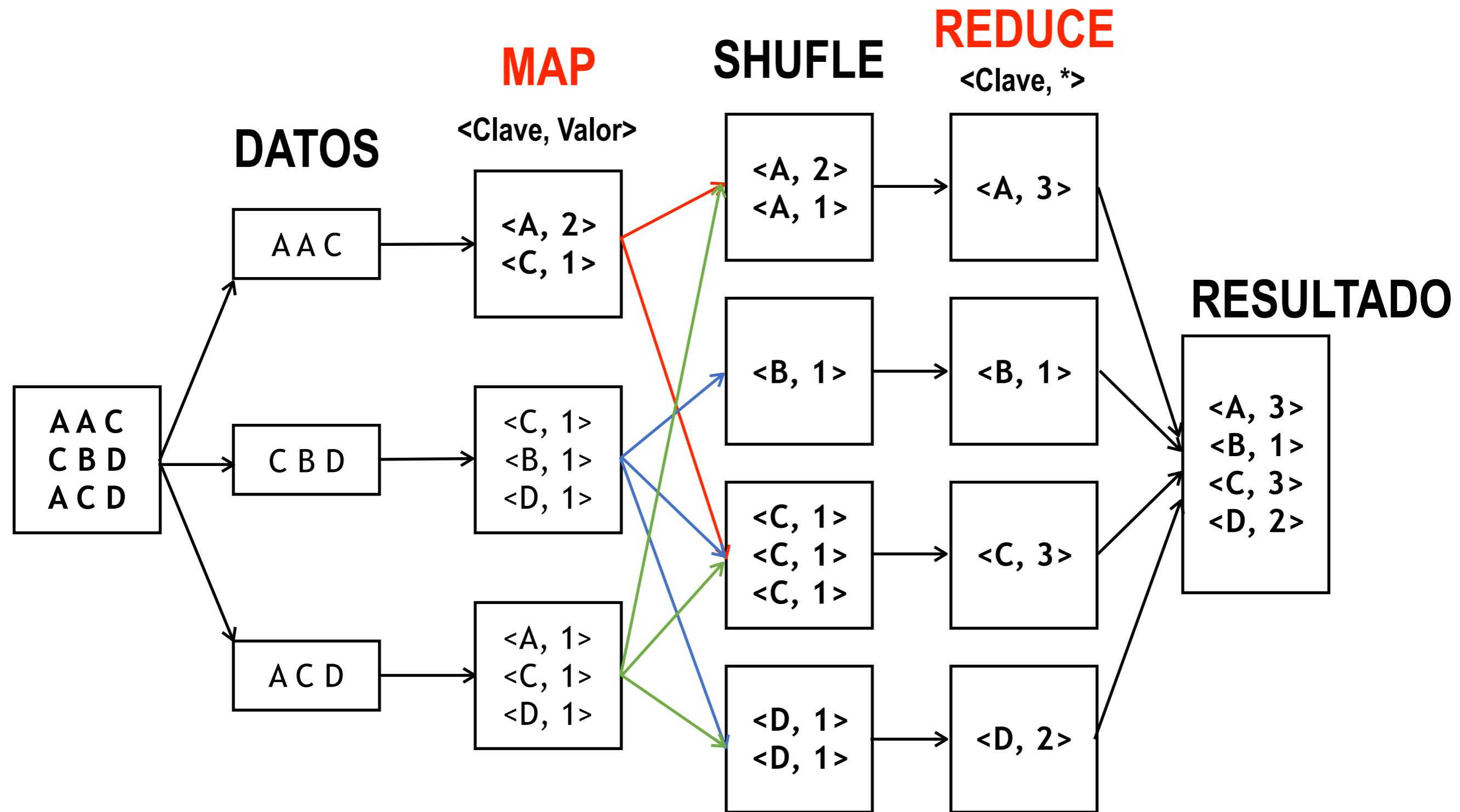
DevOps - Integración y Entrega continua de software



Hadoop / MapReduce

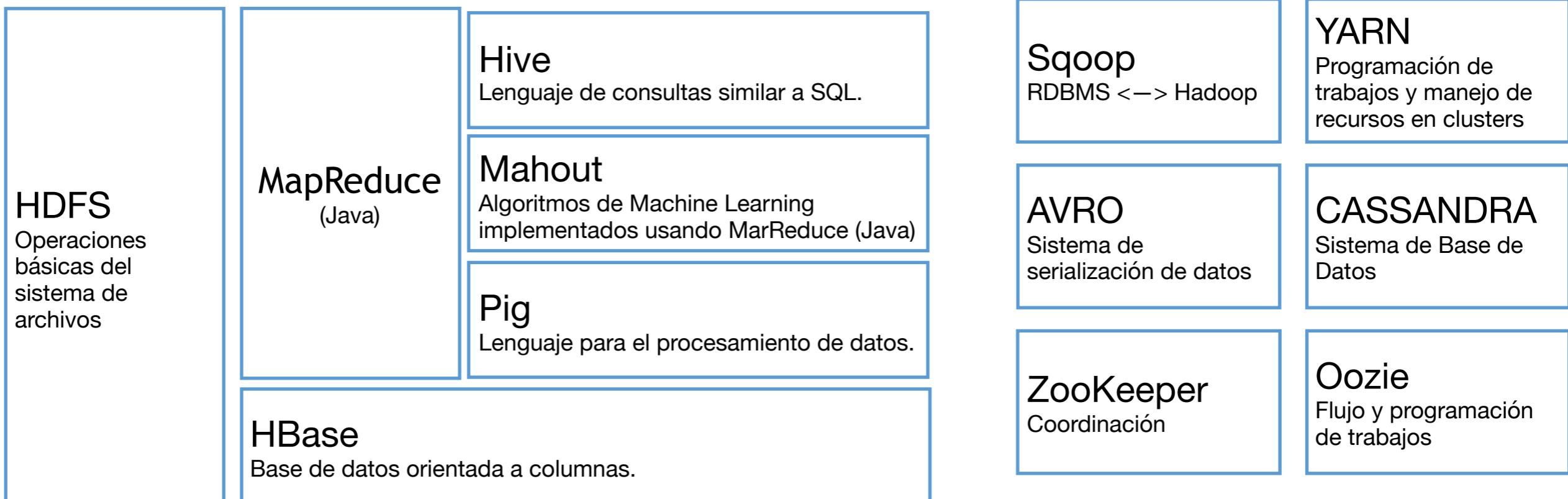


Hadoop / MapReduce



Hadoop / MapReduce

Ecosistema Apache Hadoop



Hadoop / MapReduce

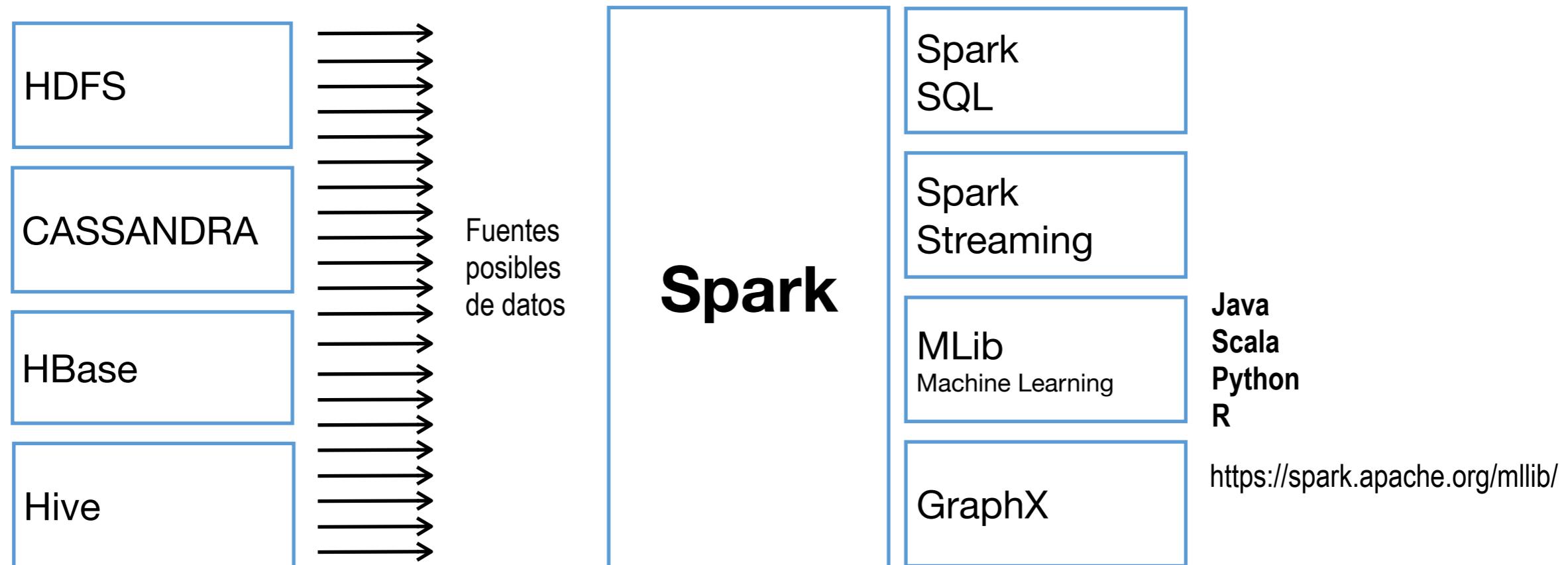
Ejemplo de Pig

```
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;
```

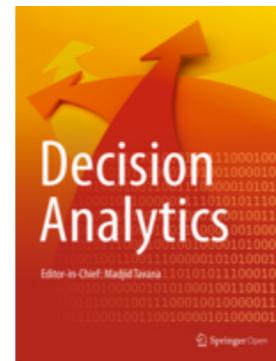
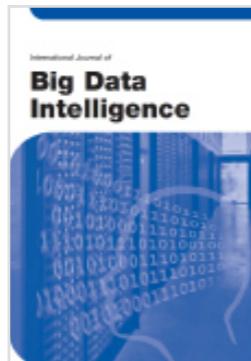
Ejemplo de Hive

```
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



Big Data / Data Science



DATA SCIENCE JOURNAL

2002

Journal of Data Science

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

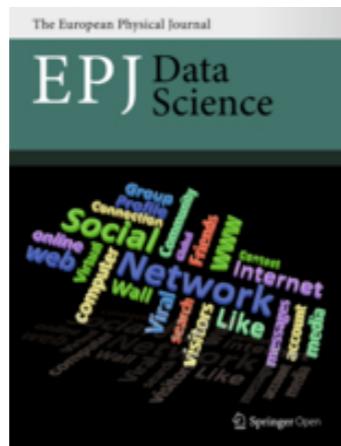
About JDS

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 mathematics. The journal is open to papers dealing with theory 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.

2003



2012



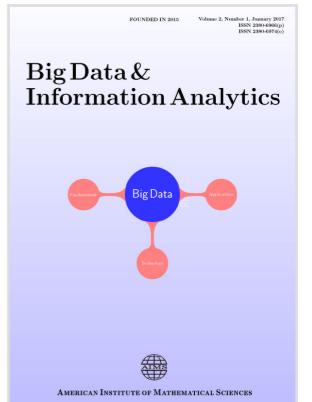
2013



2014



2015



2016

IEEE TRANSACTIONS ON
BIG DATA
IEEE computer society

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

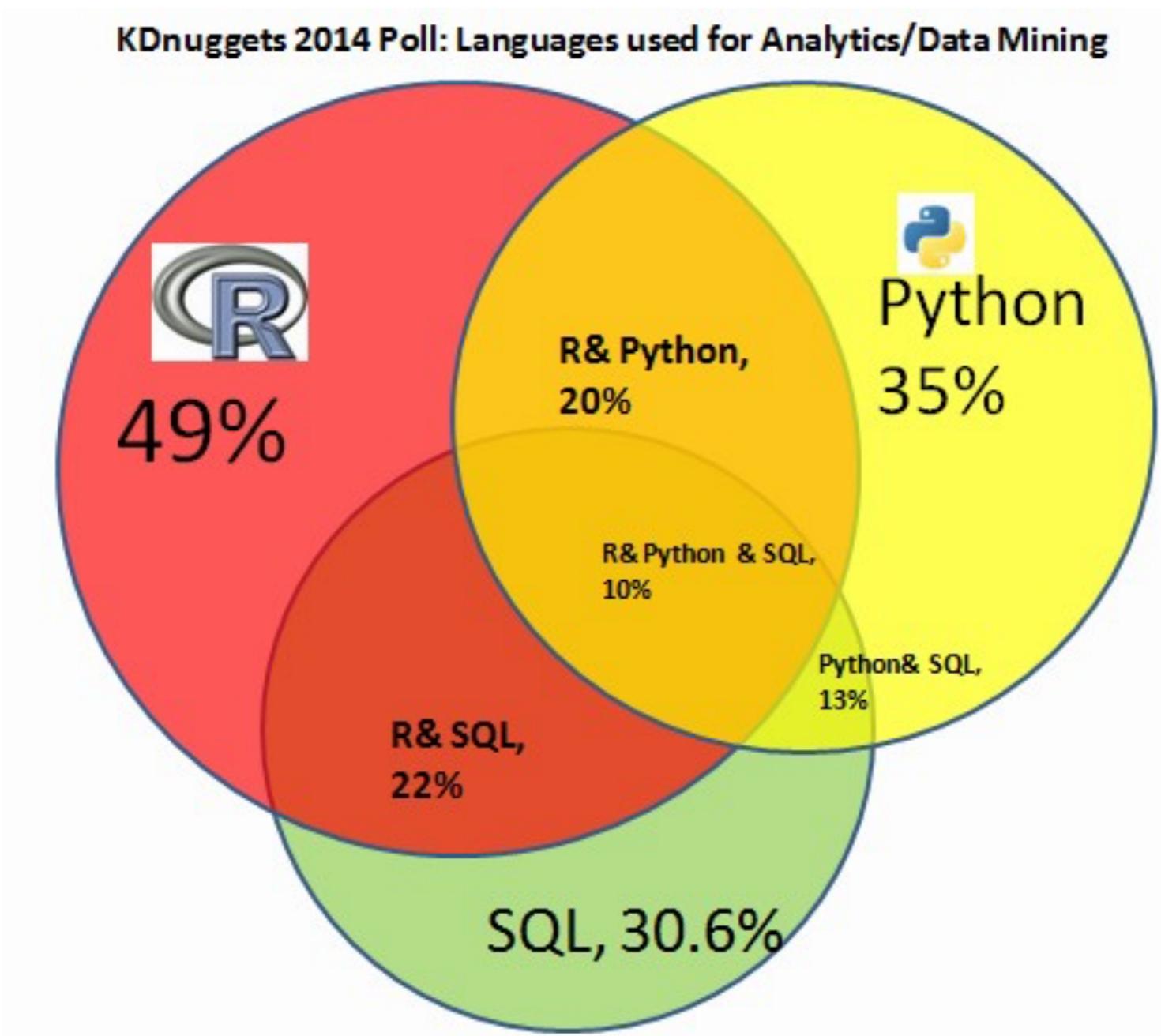
Salford Systems

Tableau

Qilk

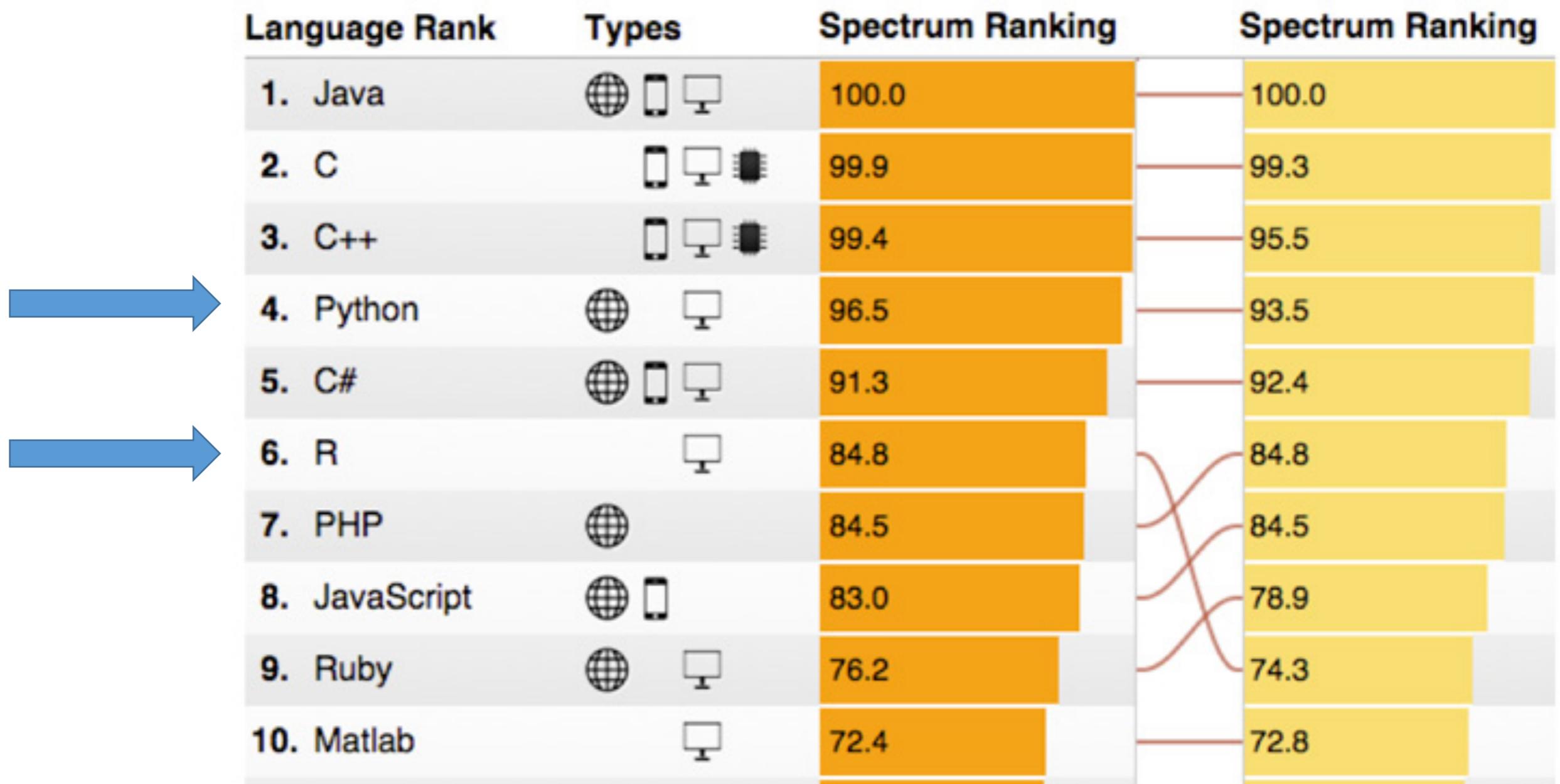
Open Data Science

Popularidad de los lenguajes



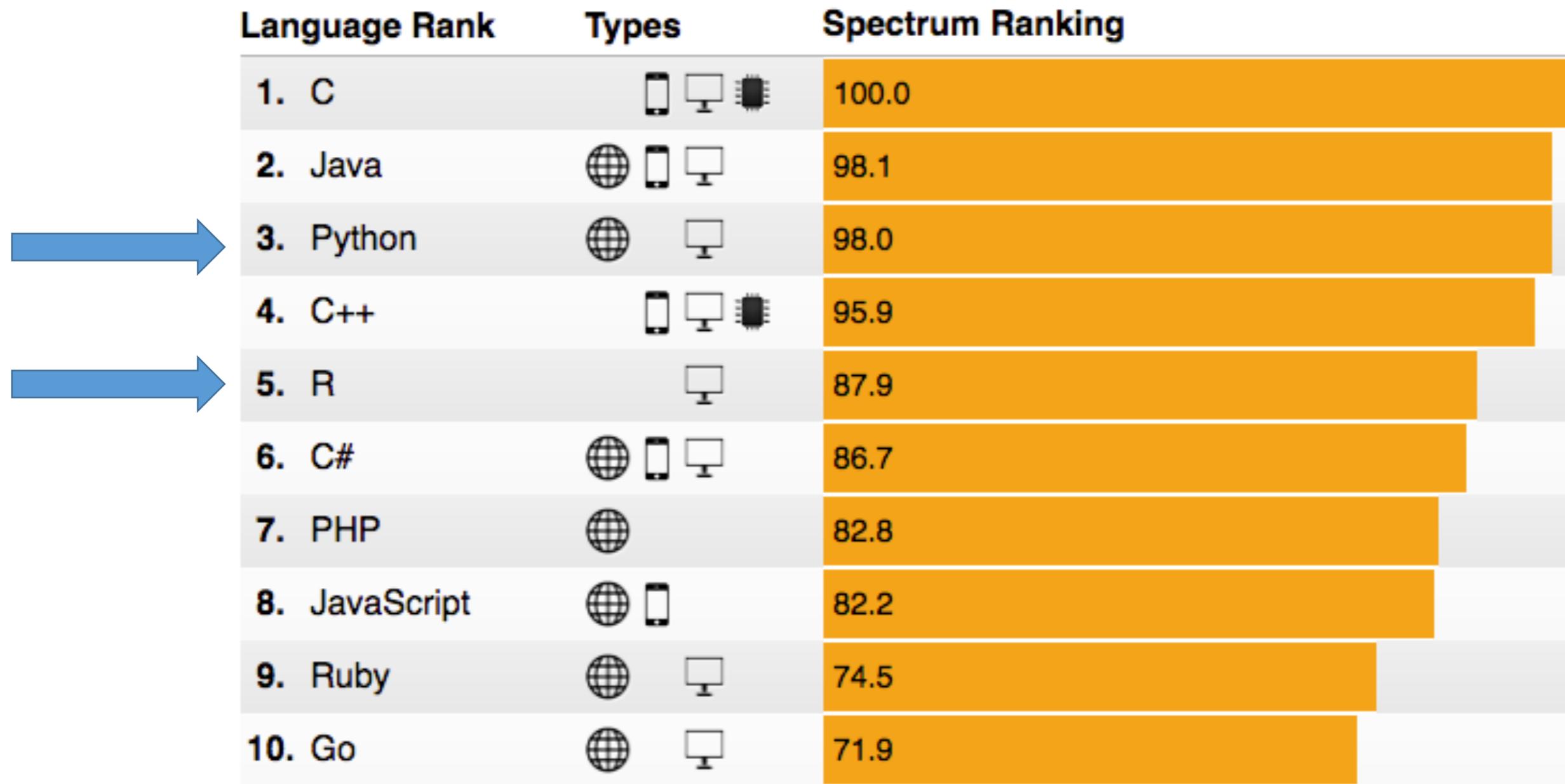
Open Data Science

The 2015 Top Ten Programming Languages (IEEE Spectrum)



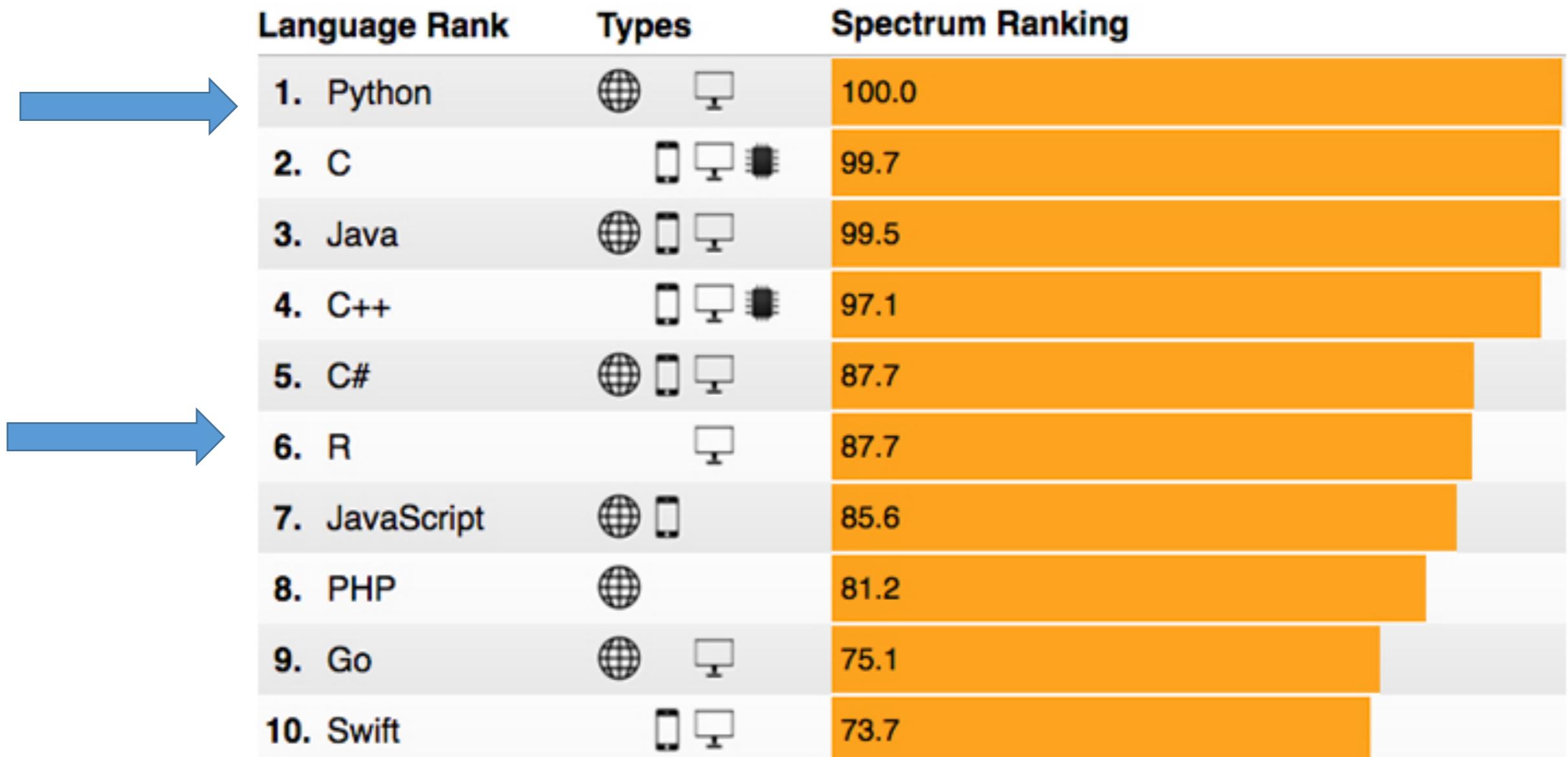
Open Data Science

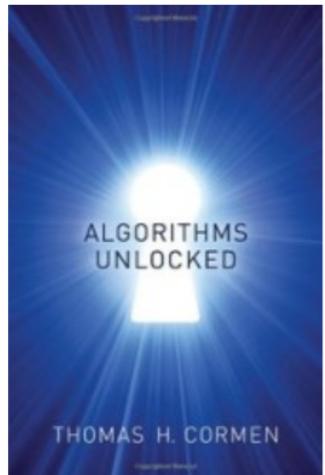
The **2016** Top Ten Programming Languages (IEEE Spectrum)



Open Data Science

The 2017 Top Ten Programming Languages (IEEE Spectrum)

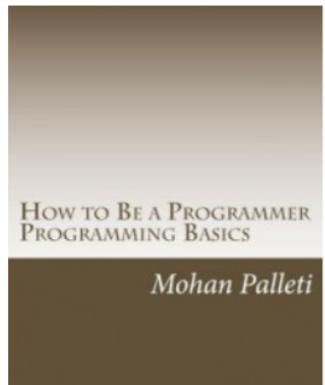




Algorithms Unlocked

By: Thomas H. Cormen

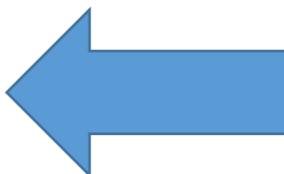
Have you ever wondered how your GPS can find the fastest way to your destination, selecting one route from seemingly countless possibilities in mere seconds? How your credit card account number is protected when you make a purchase over the Internet? The answer is algorithms. And how do...



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?

Programación para
ingeniería
Cómputo numérico.

¿Calcular la suma de los primeros 20 números primos?

¿Computar la inversa de una matriz?

Programación para
Computer Sciences

Manipulación de texto.

Open Data Science

Explotación de HW

moderno

- Servidores
- Clusters
- GPUs & Workstations

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 Tibble

R ggpairs

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

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

Jupyter Notebook

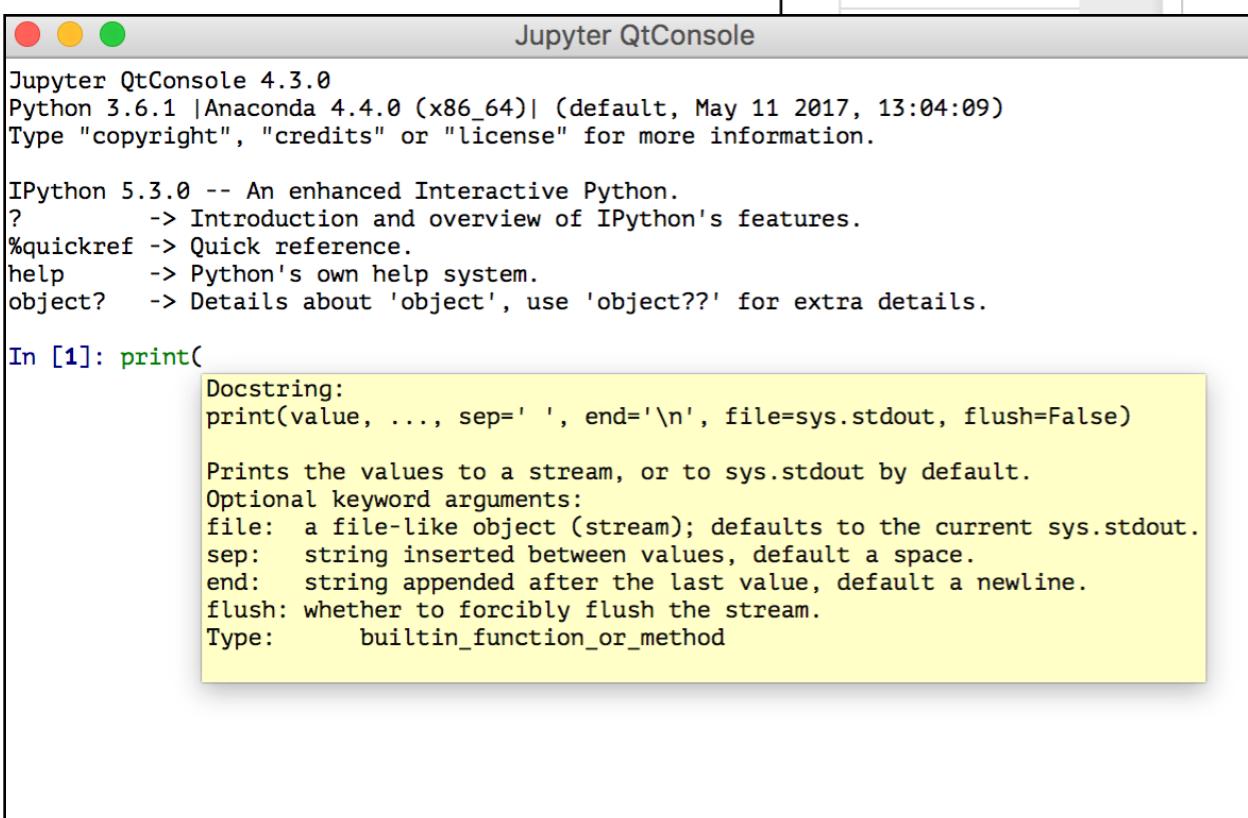
Storyboards

- Notebooks
- Exploración interactiva
- Programación visual
- Data IDE

Aplicaciones modernas

- Notebooks
- Dashboards
- Aplicaciones visuales
- Servicios de datos

Jupyter QtConsole



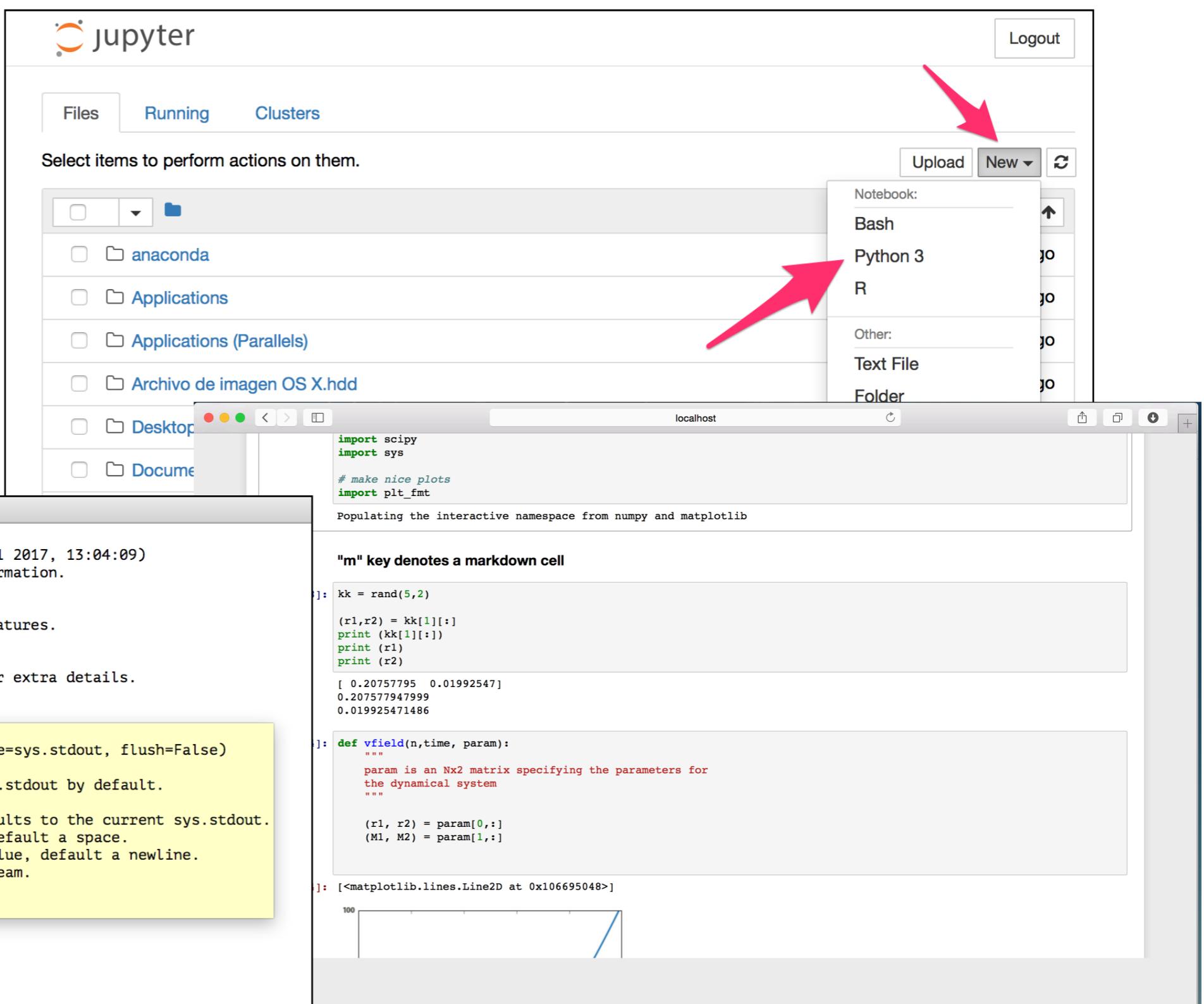
Jupyter QtConsole 4.3.0
Python 3.6.1 |Anaconda 4.4.0 (x86_64)| (default, May 11 2017, 13:04:09)
Type "copyright", "credits" or "license" for more information.

IPython 5.3.0 -- An enhanced Interactive Python.
? -> Introduction and overview of IPython's features.
%quickref -> Quick reference.
help -> Python's own help system.
object? -> Details about 'object', use 'object??' for extra details.

In [1]: `print(`

Docstring:
`print(value, ..., sep=' ', end='\n', file=sys.stdout, flush=False)`

Prints the values to a stream, or to sys.stdout by default.
Optional keyword arguments:
file: a file-like object (stream); defaults to the current sys.stdout.
sep: string inserted between values, default a space.
end: string appended after the last value, default a newline.
flush: whether to forcibly flush the stream.
Type: builtin_function_or_method



jupyter

Logout

Files Running Clusters

Select items to perform actions on them.

Upload New ↗

Notebook: Bash Python 3 R Other: Text File Folder

localhost

```
import scipy
import sys

# make nice plots
import plt_fmt

Populating the interactive namespace from numpy and matplotlib
```

"m" key denotes a markdown cell

```
:> kk = rand(5,2)
(r1,r2) = kk[1][:]
print (kk[1][:])
print (r1)
print (r2)
[ 0.20757795  0.01992547]
0.207577947999
0.019925471486
```

```
:> def vfield(n,time,param):
    """
    param is an Nx2 matrix specifying the parameters for
    the dynamical system
    """

    (r1, r2) = param[0,:]
    (M1, M2) = param[1,:]
```

```
:> [<matplotlib.lines.Line2D at 0x106695048>]
```

A red arrow points to the "New" button in the top right corner of the Jupyter interface. Another red arrow points to the "m" key in the text "m key denotes a markdown cell".

Open Data Science

Storyboards

- Notebooks
- Exploración interactiva
- Programación visual
- Data IDE

Aplicaciones modernas

- Notebooks
- Dashboards
- Aplicaciones visuales
- Servicios de datos

Apache Zeppelin

Multi-purpose Notebook

The Notebook is the place for all your needs

- >Data Ingestion
- Data Discovery
- Data Analytics
- Data Visualization & Collaboration

The screenshot shows the Apache Zeppelin web interface with three open notebooks:

- Bank**: A pie chart titled "maxAge" with values ranging from 19 to 34. The largest segments are 33 (grey) and 32 (pink). Below the chart is the note: "Took a few seconds. Last updated by anonymous at June 26 2016, 4:46:52 PM. (outdated)".
- Under age < 35**: A bar chart titled "maxAge" with values from 19 to 30. The bars are grouped. The chart shows a sharp increase in values starting around age 22. Below the chart is the note: "Took a few seconds. Last updated by anonymous at June 26 2016, 4:47:32 PM. (outdated)".
- marital**: A line chart titled "single" showing the value of marital status over time. The Y-axis ranges from 1 to 105. The line starts at 1, peaks at approximately 100 around time 35, and then declines. Below the chart is the note: "Took a few seconds. Last updated by anonymous at June 26 2016, 4:47:36 PM. (outdated)".

At the top, there is a search bar "Search your Notebooks" and a user dropdown "anonymous". The bottom right corner has a "READY" button.

Open Data Science

KNIME

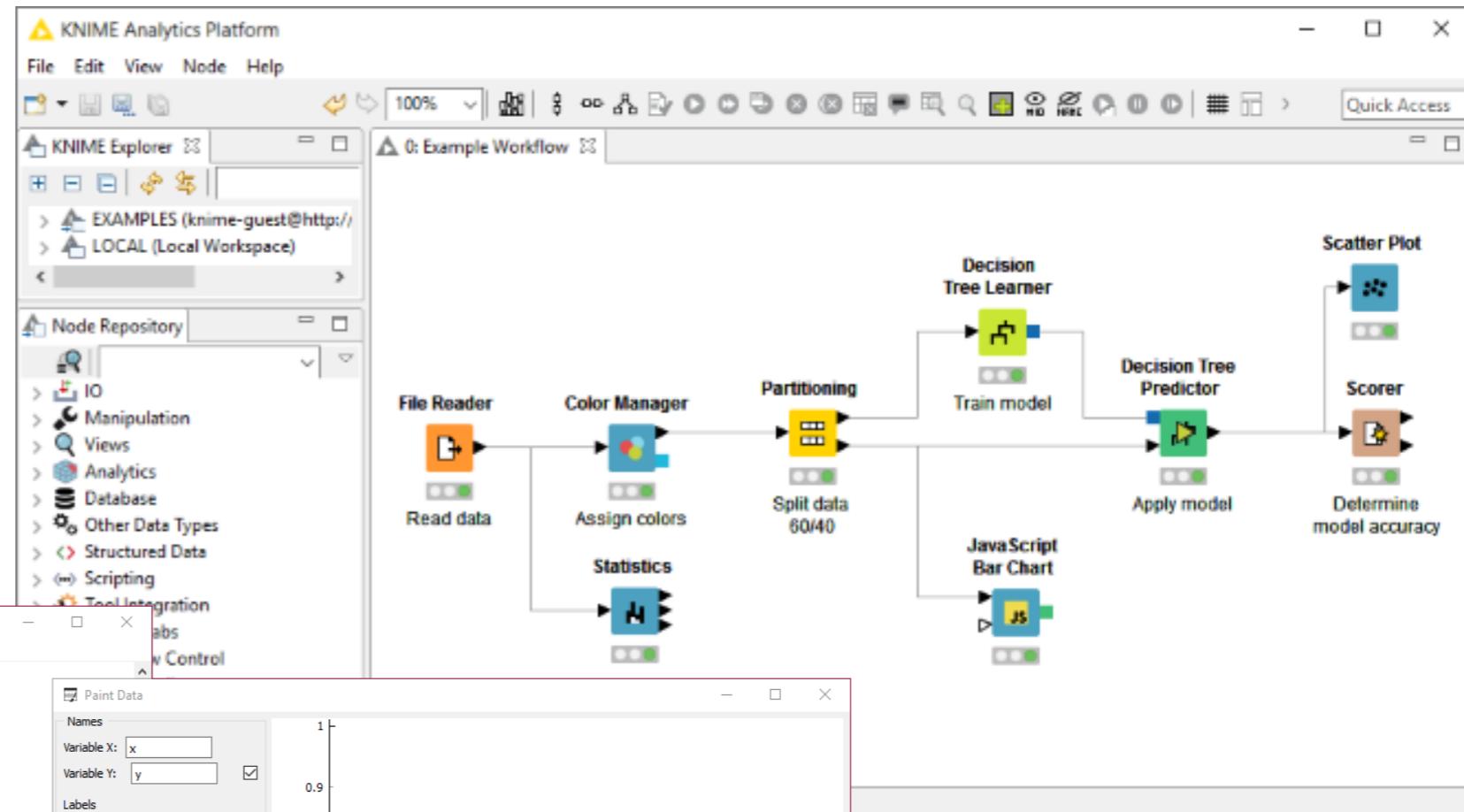
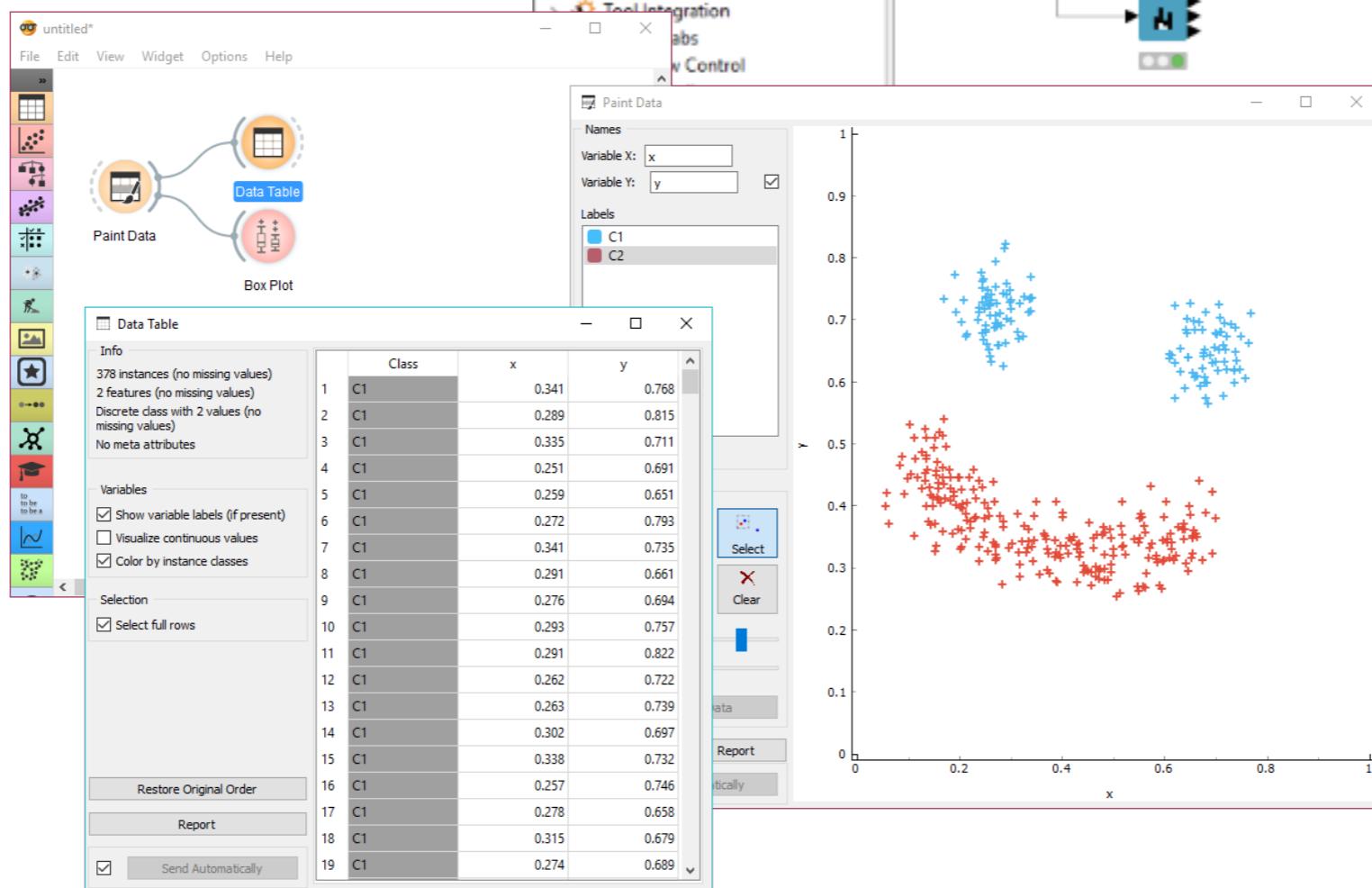
Storyboards

- Notebooks
- Exploración interactiva
- Programación visual
- Data IDE

Aplicaciones modernas

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

Orange



Open Data Science

Aplicaciones modernas

- Notebooks
 - Dashboards
 - Aplicaciones visuales
 - Servicios de datos

R Markdown

The figure shows a screenshot of the RStudio interface. The top menu bar includes 'File', 'Edit', 'File', 'Project: (None)', 'Addins', 'Environment', 'History', 'Files', 'Plots', 'Packages', 'Help', and 'Viewer'. The left pane displays an R Markdown script with code and a plot. The right pane shows the resulting Shiny application.

Code in R Markdown:

```
notebooks.Rmd x hadleymodel.Rmd x htmlwidgets.Rmd x gr > Run > Environment History
```

```
63
64 ## Residuals
65
66 To motivate the use of models we're going to start with an
67 interesting pattern from the NYC flights dataset -- the
68 number of flights per day.
69 ````{r}
70 library(nycflights13)
71 library(lubridate)
72 library(dplyr)
73 daily <- flights %>%
74   mutate(date = make_datetime(year, month, day)) %>%
75   group_by(date) %>%
76   summarise(n = n())
77
78 ggplot(daily, aes(date, n)) +
79   geom_line()
80 ````
```

Plot in R Markdown:

A line plot showing the number of flights (n) over time (date). The y-axis ranges from 700 to 1000, and the x-axis shows dates from December 2012 to December 2013. The plot displays a clear seasonal trend with weekly fluctuations.

Shiny Application Content:

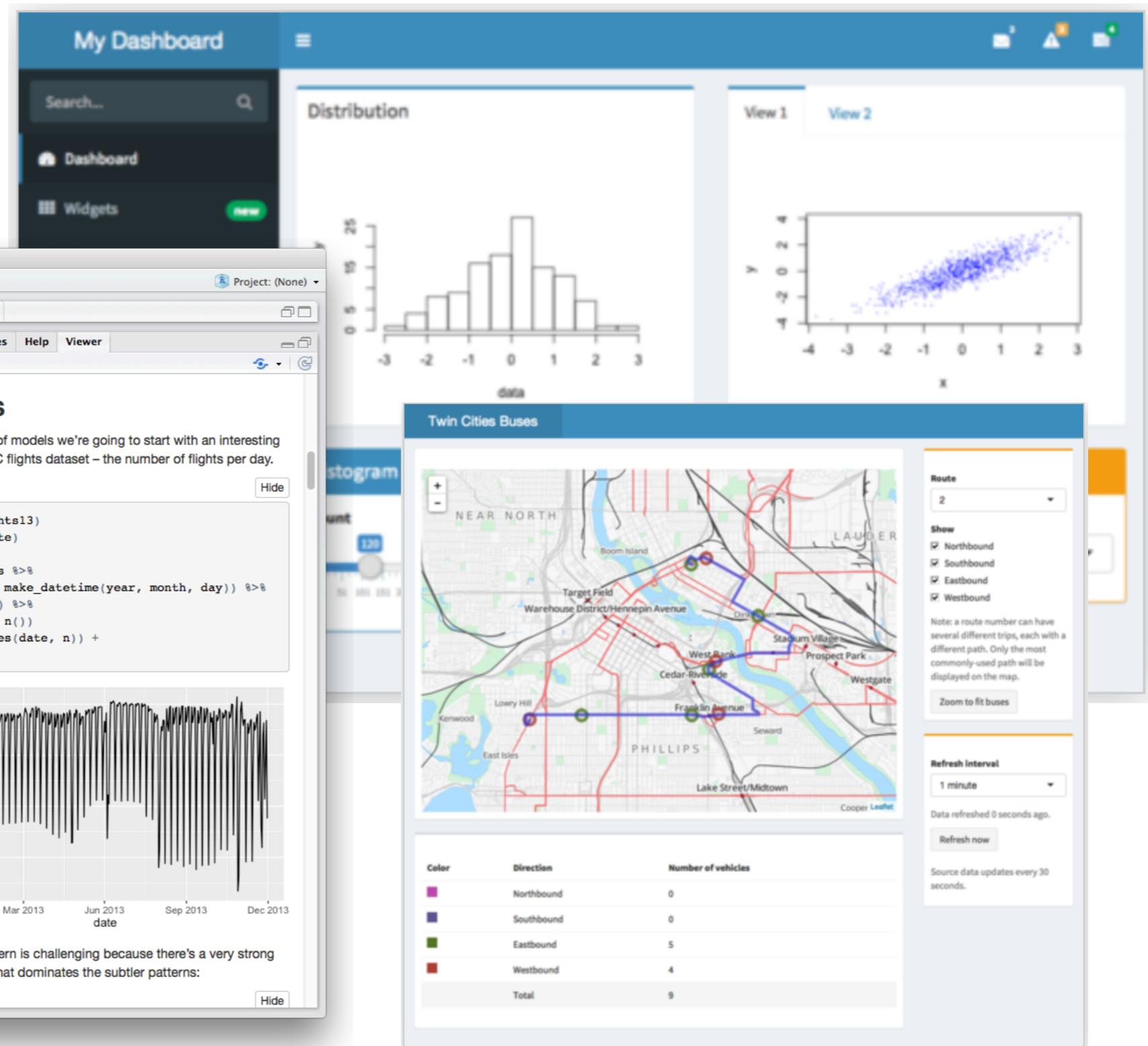
Residuals

To motivate the use of models we're going to start with an interesting pattern from the NYC flights dataset – the number of flights per day.

```
library(nycflights13)
library(lubridate)
library(dplyr)
daily <- flights %>%
  mutate(date = make_datetime(year, month, day)) %>%
  group_by(date) %>%
  summarise(n = n())
ggplot(daily, aes(date, n)) +
  geom_line()
```

Shiny Application Plot:

A line plot showing the number of flights (n) over time (date). The y-axis ranges from 700 to 1000, and the x-axis shows dates from December 2012 to December 2013. The plot displays a clear seasonal trend with weekly fluctuations, identical to the one in the R Markdown code.



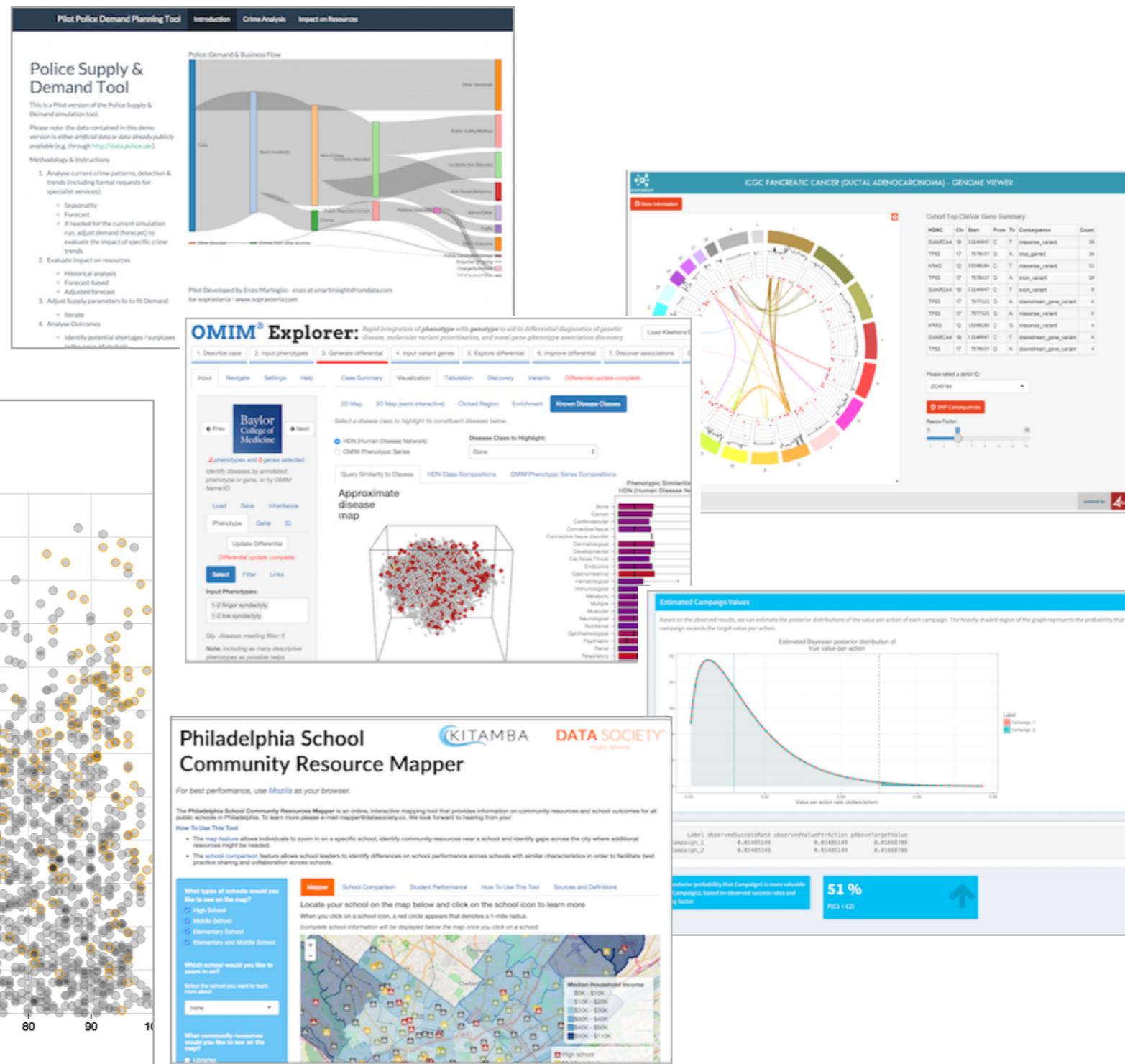
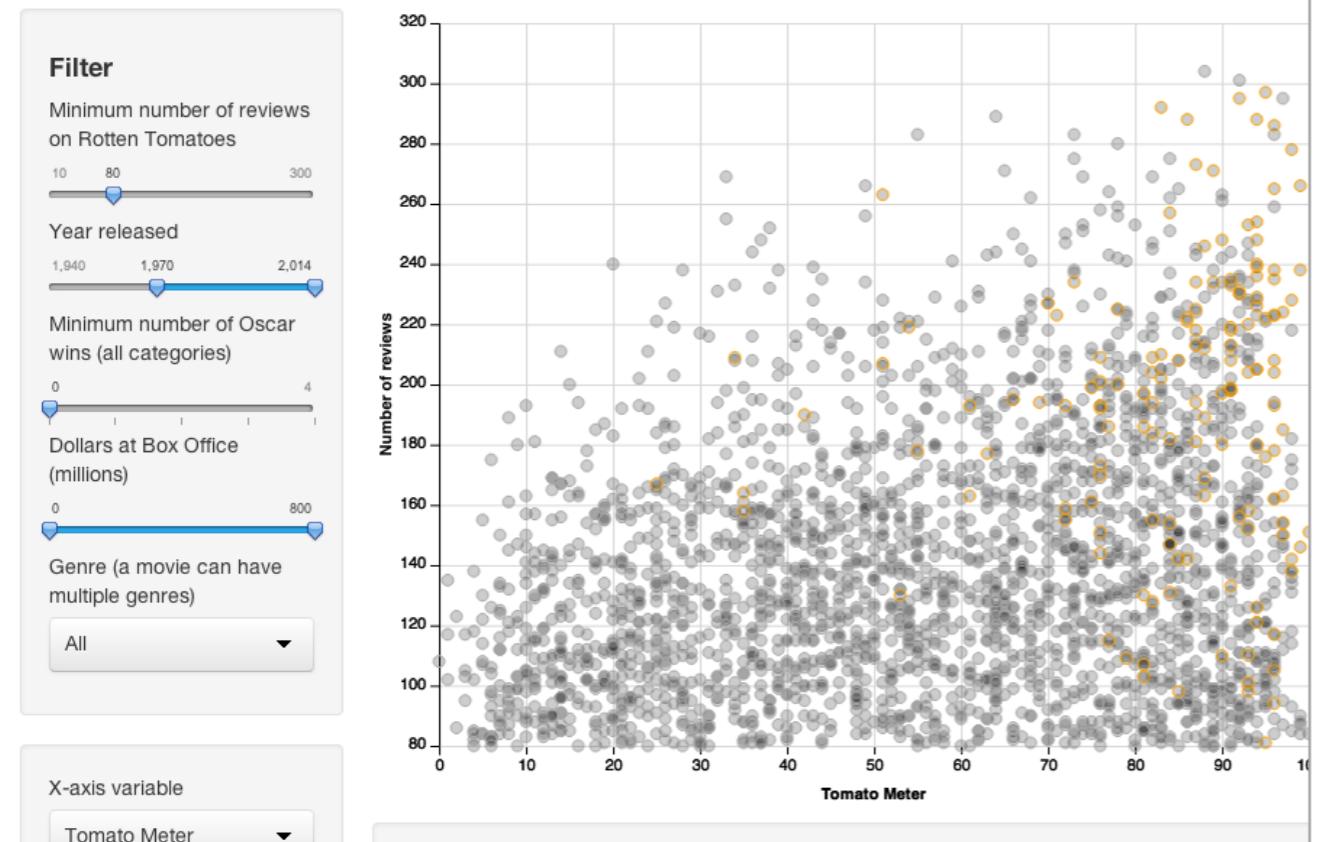
Open Data Science

Aplicaciones modernas

- Notebooks
- Dashboards
- Aplicaciones visuales
- Servicios de datos

R Shiny

Movie explorer



Open Data Science

Visualización

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

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

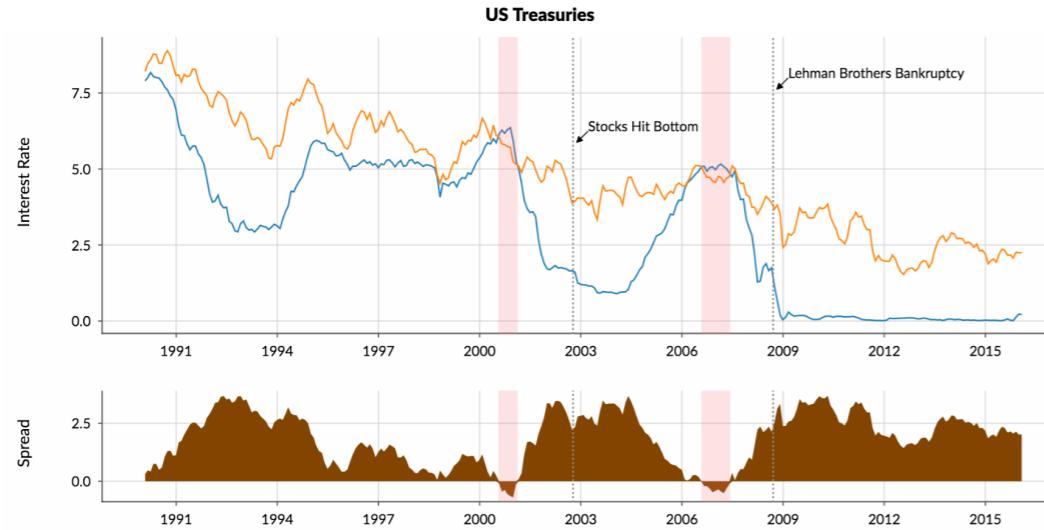
BeakerX is a collection of JVM kernels with widgets, plotting, tables, autotranslation, and other extensions to the Jupyter Notebook and changes with

The document

BeakerX is th
are hiring.

Groovy with Interactive Plotting and Tables:

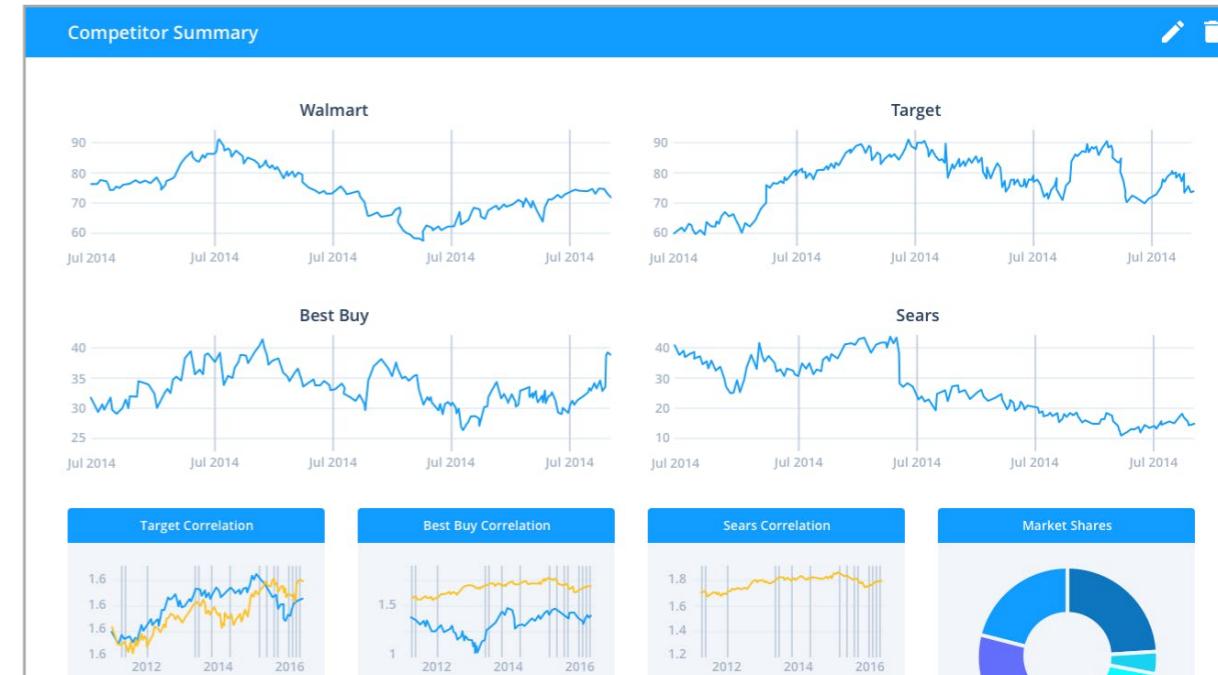
```
// Then use a CombinedPlot to get stacked plots with linked X axis.  
def c = new CombinedPlot(title: "US Treasuries", initWidth: 1000)  
  
// add both plots to the combined plot, and including their relative heights.  
c.add(p1, 3)  
c.add(p2, 1)
```



Bokeh



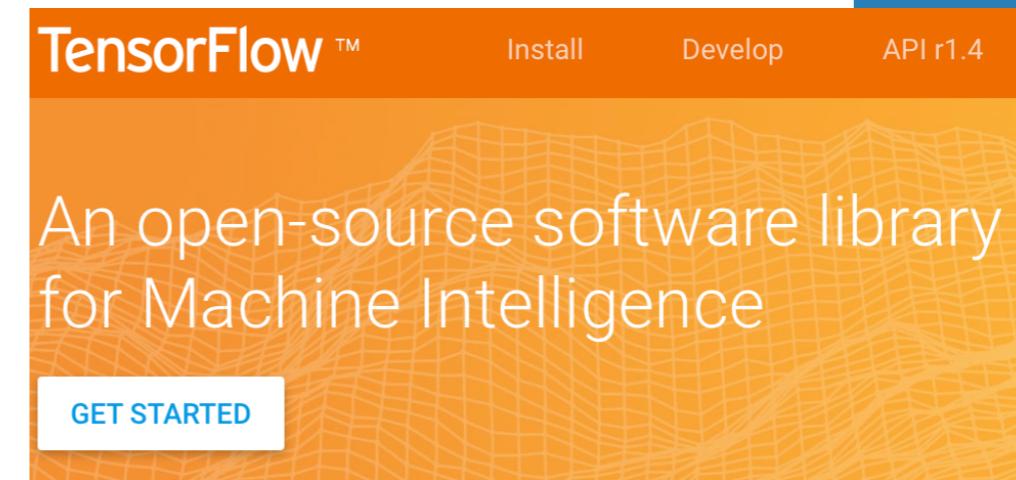
plot.ly



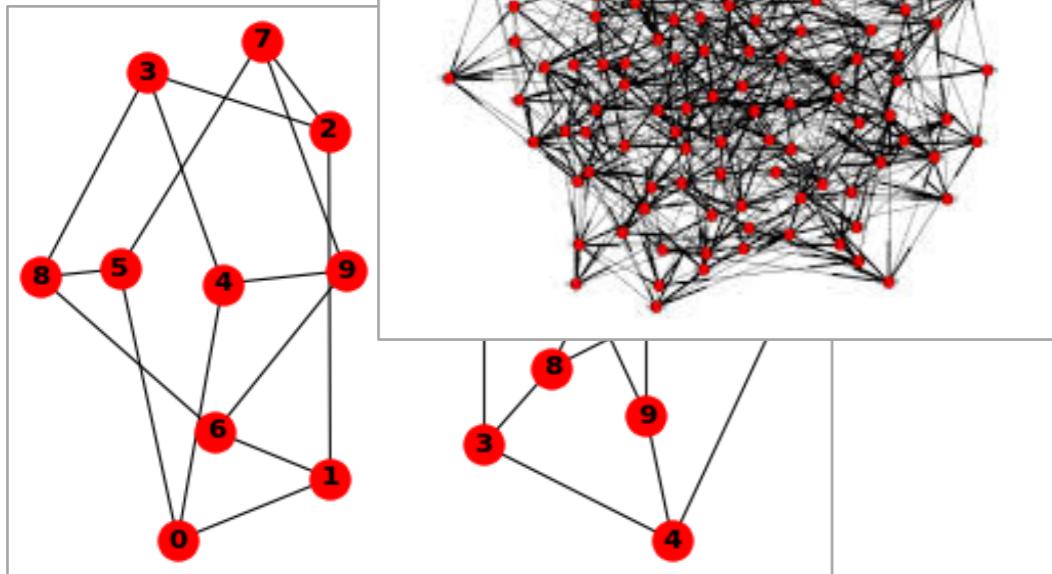
Open Data Science

Analytics

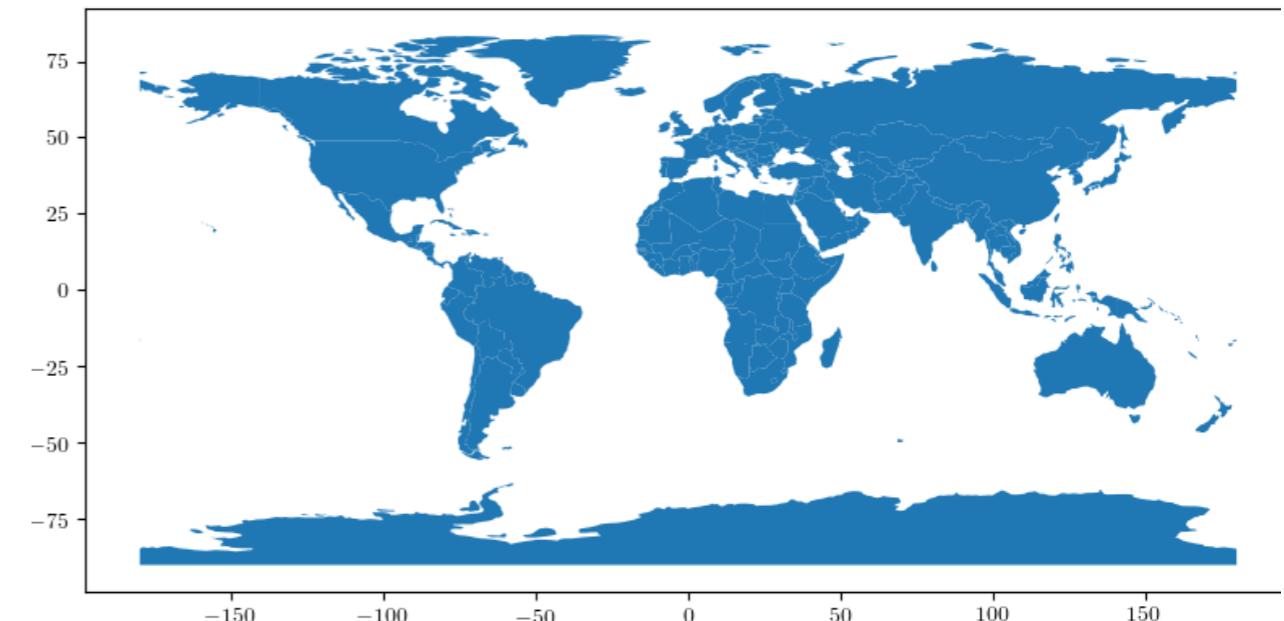
- Preparación de datos
- Estadística
- ML & Ensambls
- Deep learning
- Simulación y optimización
- Datos geoespaciales
- Texto y NLP
- Gráficos y redes
- Minería de audio, video e imágenes



NetworkX



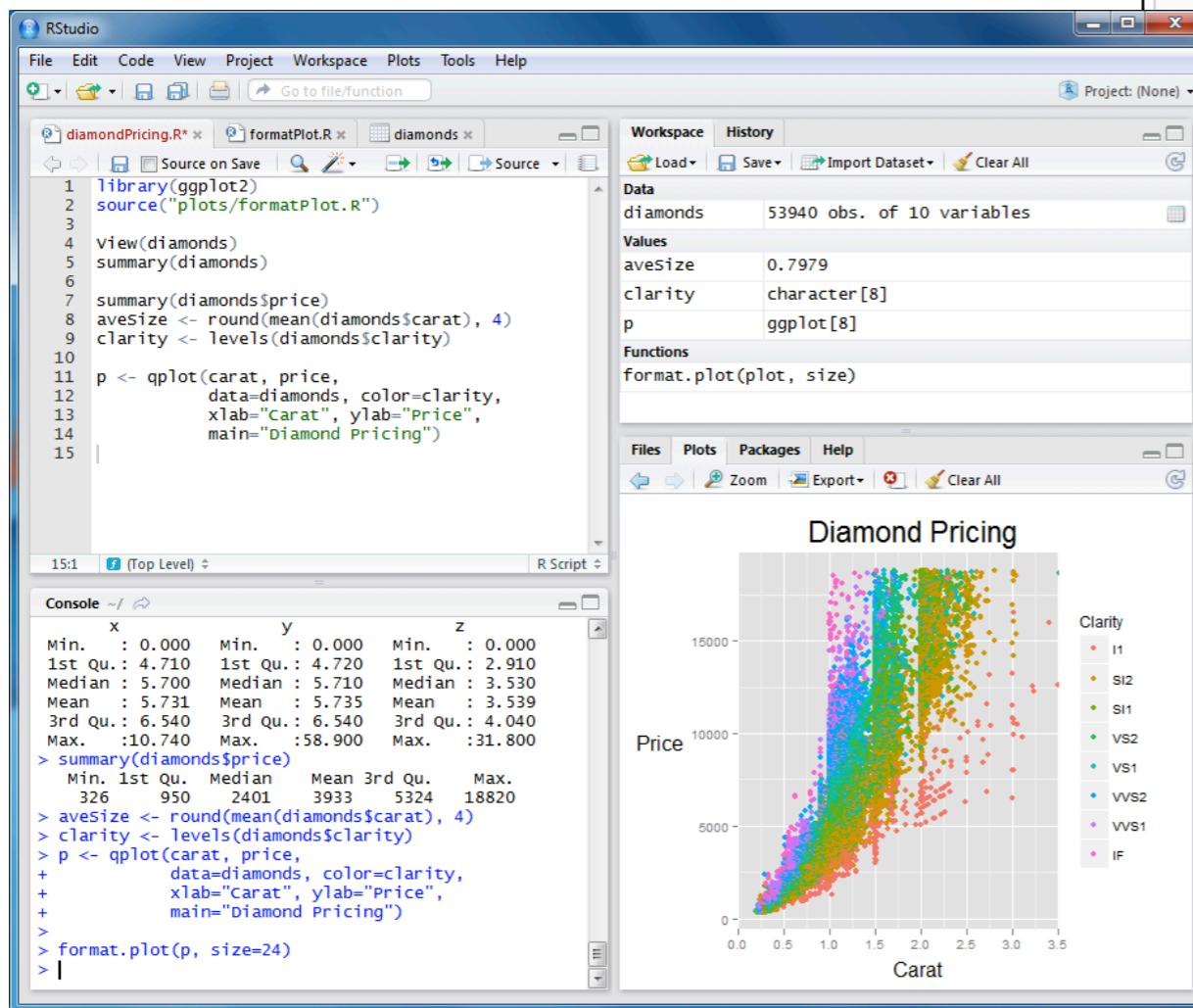
GeoPandas



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
Observations:      86   AIC:             765.6
Residuals:       83   BIC:             773.0
Model:                   2
Variance Type: nonrobust
```

RStudio

Open Data Science

Analytics

- Preparación de datos
- Estadística
- ML & Ensambls
- 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](#))

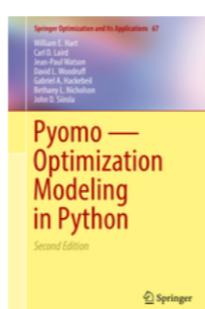
Examples

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
```

Sponsored by:
CONTINUUM[®]
ANALYTICS

HOME OVERVIEW PROJECTS TALKS BLOG

The Blaze Ecosystem

The Blaze ecosystem is a set of libraries that help users store, describe, query and process data. It is composed of the following core projects:

- [Blaze](#): An interface to query data on different storage systems
- [Dask](#): Parallel computing through task scheduling and blocked algorithms
- [Datashape](#): A data description language

Open Data Science

Fuentes de datos

modernas

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

SQLite

```
1 import sqlite3
2 conn = sqlite3.connect('example.db')
3
4 c = conn.cursor()
5 c.execute('')
6     CREATE TABLE person
7     (id INTEGER PRIMARY KEY ASC, name varchar(250) NOT NULL)
8     ''')
9 c.execute('')
10    CREATE TABLE address
11    (id INTEGER PRIMARY KEY ASC, street_name varchar(250), street_number varchar(
12    250),
13    post_code varchar(250) NOT NULL, person_id INTEGER NOT NULL,
14    FOREIGN KEY(person_id) REFERENCES person(id))
15    ''')
16 c.execute('')
17     INSERT INTO person VALUES(1, 'pythoncentral')
18     ''')
19 c.execute('')
20     INSERT INTO address VALUES(1, 'python road', '1', '00000', 1)
21     ''')
22 conn.commit()
```

```
1 import sqlite3
2 conn = sqlite3.connect('example.db')
3
4 c = conn.cursor()
5 c.execute('SELECT * FROM person')
6 print c.fetchall()
7 c.execute('SELECT * FROM address')
8 print c.fetchall()
9 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/>

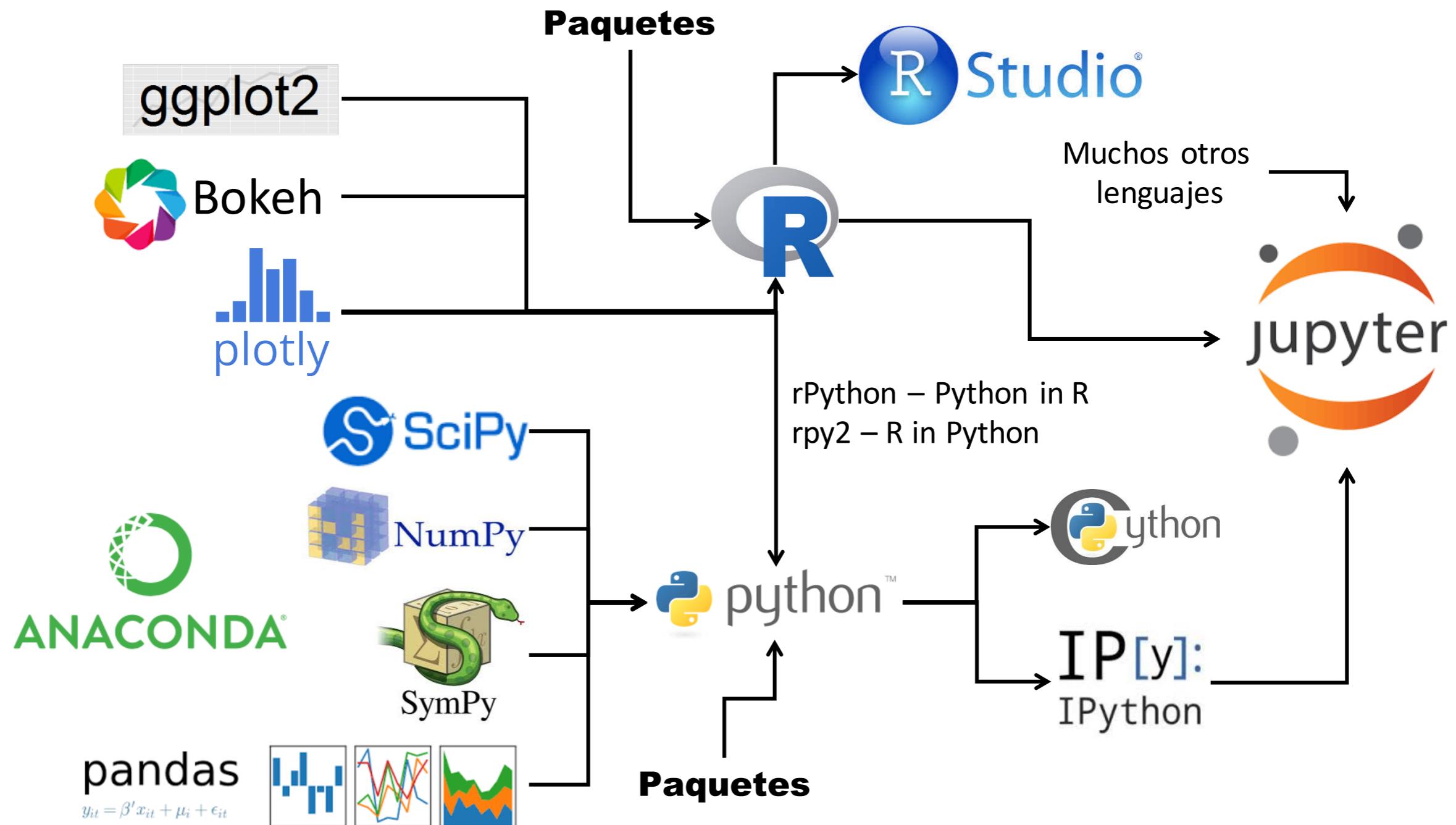
pyCUDA – <https://mathematician.de/software/pycuda/>

```
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 vacias
sed -e 's/;;;/00:00:00/g' datos > out.1

## etcetera ...

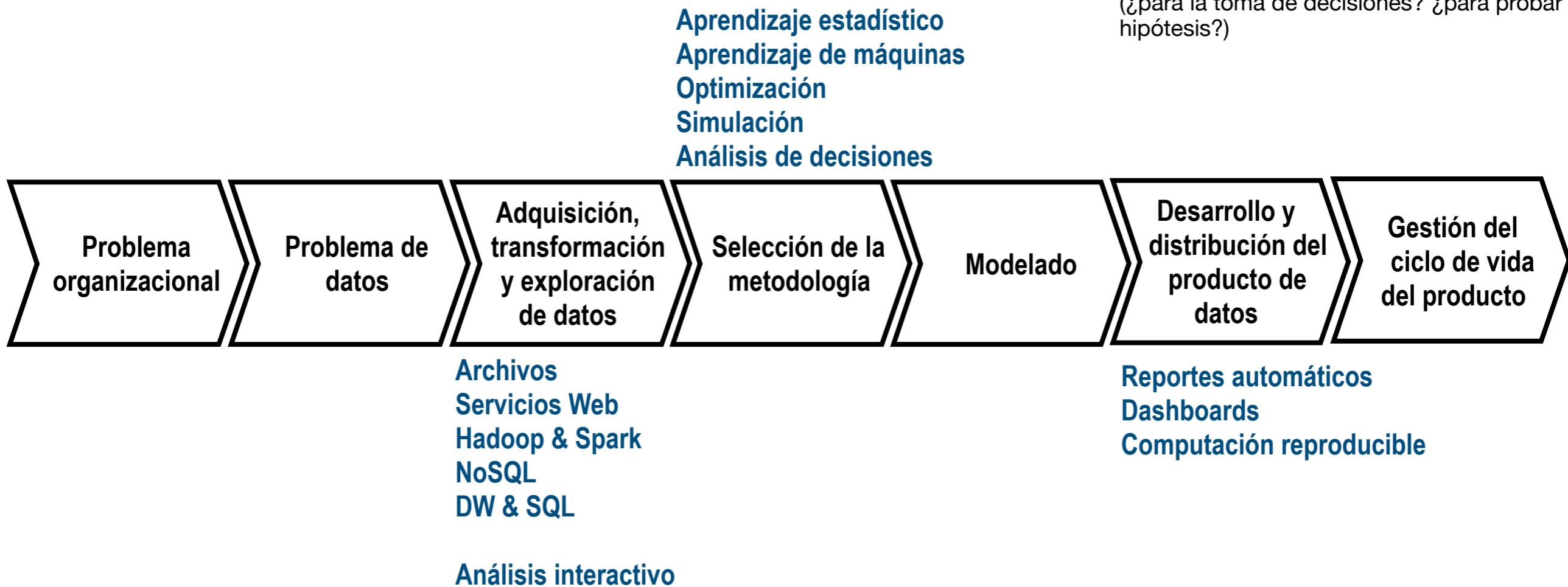
## promedio para cada hora
csvsql --query "select ESTACION, FECHA, ANO, MES,
  DIA, HORA, DIRECCION, avg(VELOCIDAD) as VELOCIDAD from 'out'
  group by ESTACION, FECHA, HORA" out.5 > out.6
```

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,2005-04-16,2005,4,16,15,135,3.4666666666666667
AQUITANIA,2005-04-16,2005,4,16,16,135,3.6999999999999993
AQUITANIA,2005-04-16,2005,4,16,17,135,4.833333333333333
AQUITANIA,2005-04-16,2005,4,16,18,135,4.7666666666666667
AQUITANIA,2005-04-16,2005,4,16,19,135,4.3500000000000005
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].



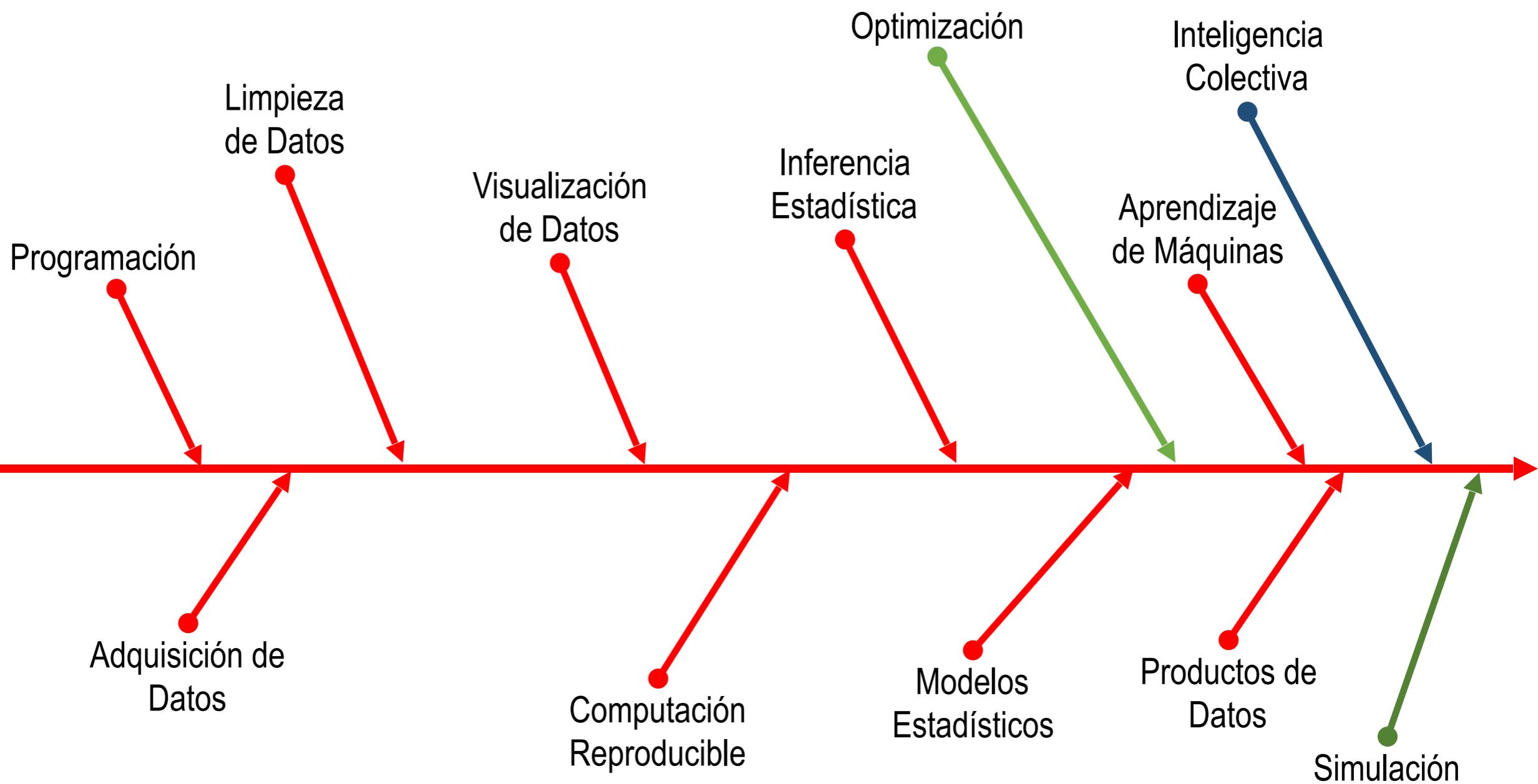
Infraestructura computacional

{ Un procesador

Muchos procesadores

{ Computación en máquinas locales
Computación en la nube

Analytics

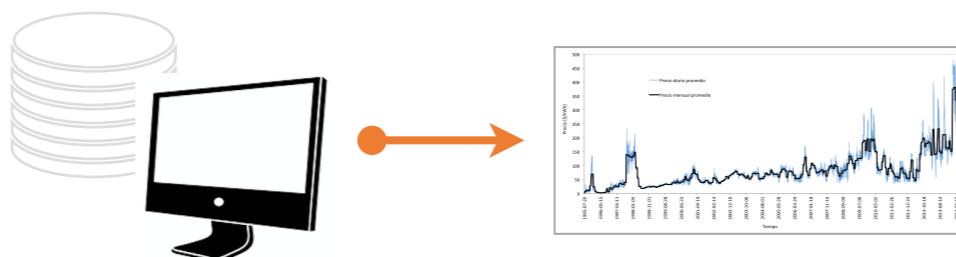


Data-driven decision making!

Analytics

Estadística y
aprendizaje de
máquinas

Los datos
están listos



Modelado de
datos

Inteligencia
de Negocios

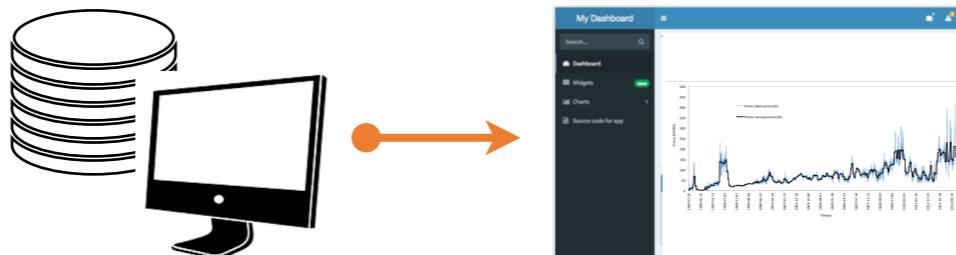
DW / OLAP



Generación,
agregación, análisis
y visualización de
datos del negocio

Minería de
Datos

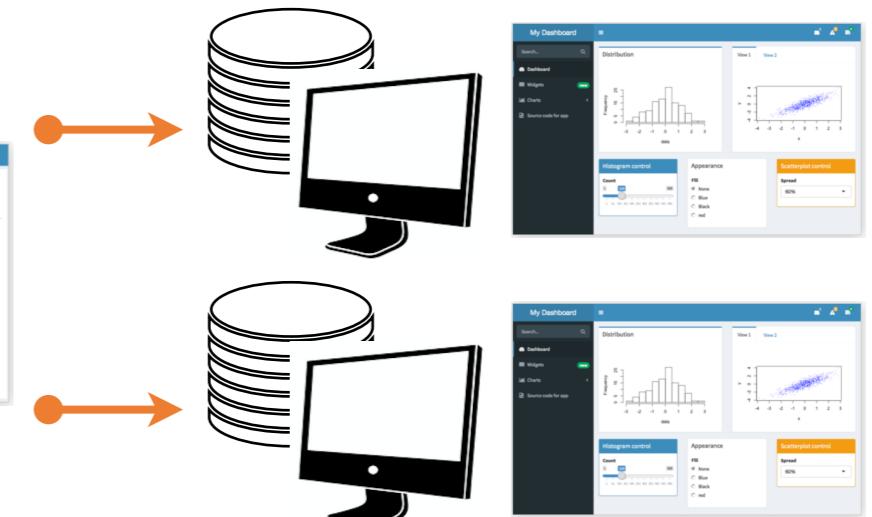
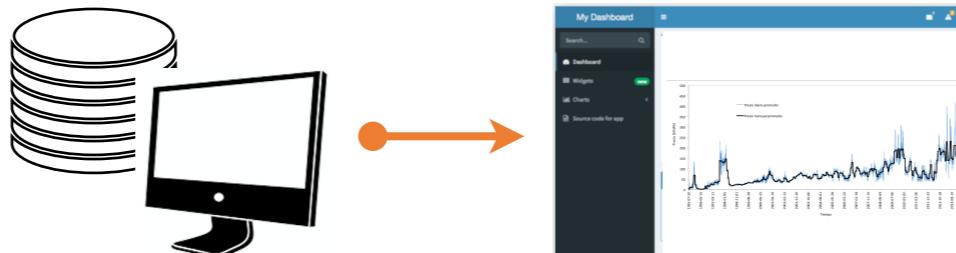
DW / OLAP



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

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 scale reduction algorithms are available through the 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

MLib

Machine Learning

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

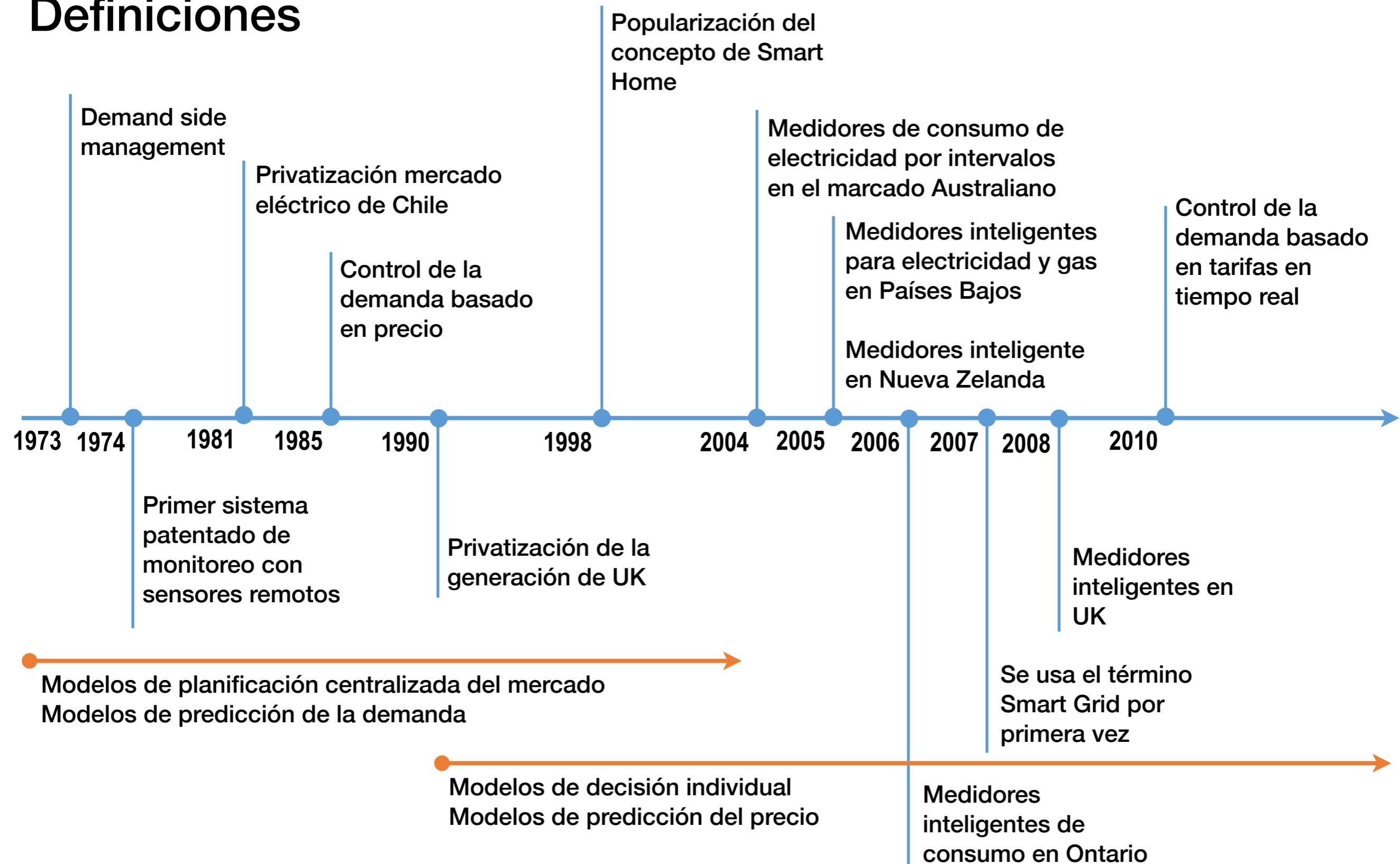
 @jdvelasquezh

 <https://github.com/jdvelasq>

 <https://goo.gl/prkjAq>

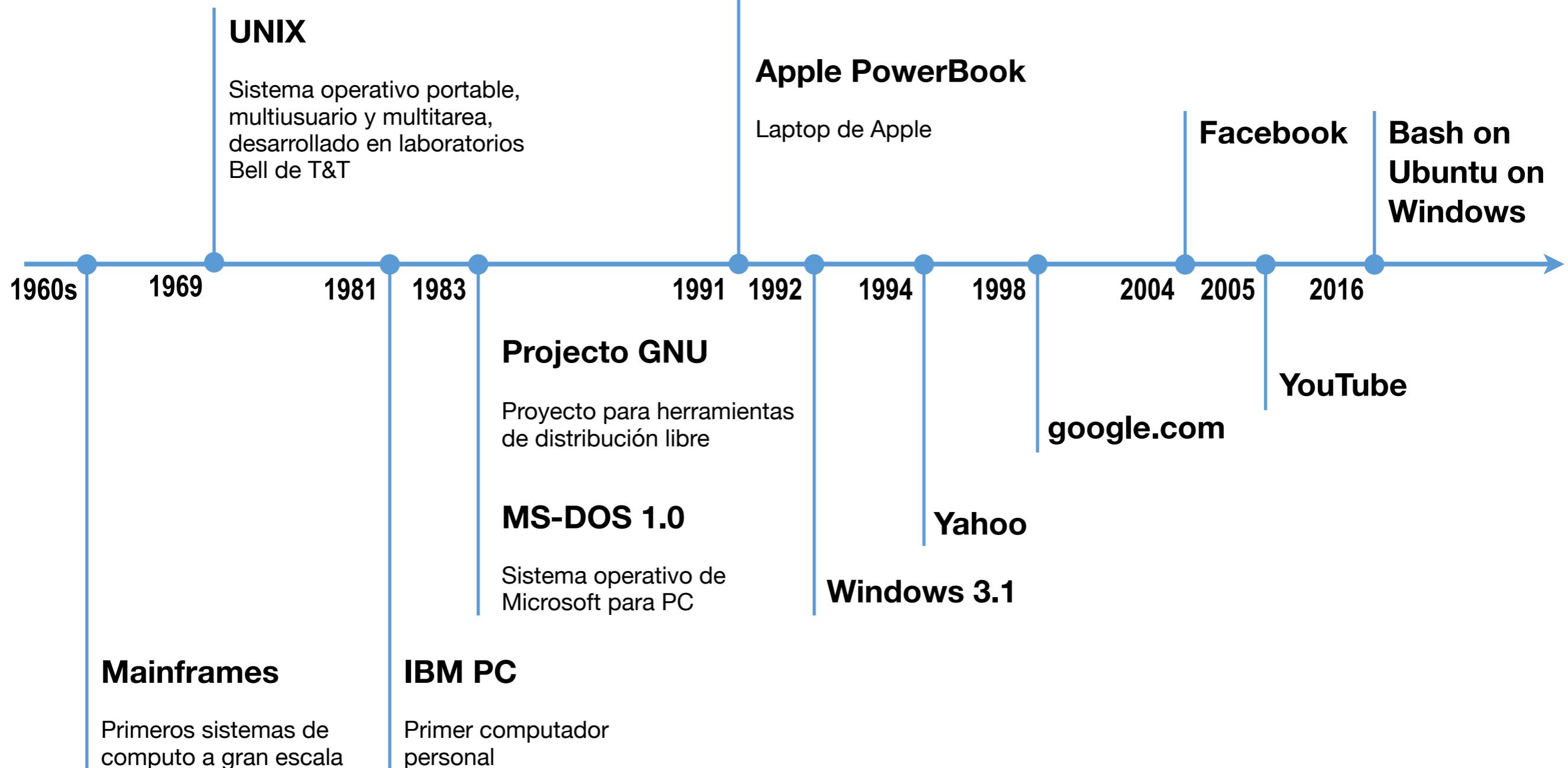
 <https://goo.gl/vXH8jy>

Definiciones



Disciplina	Tecnología	Habilidades	Foco
Inteligencia de Negocios	<ul style="list-style-type: none">ETL/SQLRDBMSReportesVisualización	<ul style="list-style-type: none">ProgramaciónAnálisis de datosModelado de datosDesarrollo de reportesEstadística BásicaAnálisis del negocio & EstrategiaPresentación oral	<ul style="list-style-type: none">Suministro de información y reporteVisualización de datosEstadísticos descriptivosIntegración de datos y consolidación
Análisis de datos	<ul style="list-style-type: none">Software para modelado de datosSoftware para diagramaciónSoftware para documentaciónSQLSoftware para perfilado de datos	<ul style="list-style-type: none">Modelado de datosAnálisis del negocioManipulación de datosEstadística básica	<ul style="list-style-type: none">Reglas de negocioDefinición de datosRelaciones entre datosAtributos de datosEstructuras de datosFuentes y usos de datosCalidad de datos
Ciencia de los Datos (Analytics)	<ul style="list-style-type: none">Software estadísticoDatos columnaresMap-ReduceNoSQLLenguajes de programaciónSoftware para graficaciónSoftware para optimización, simulación, predicción y análisis de decisiones	<ul style="list-style-type: none">Estadística avanzadaProgramaciónAnálisis del negocioArquitecturas y tecnologías modernas para el manejo de datosDesarrollo de productos de datosSimulación de sistemasOptimizaciónPredicción	<ul style="list-style-type: none">Modelado predictivoAnálisis estadístico avanzadoMinería de datosManejo de datos no estructuradosManejo de grandes volúmenes de datosI+DAnálisis de decisiones

Infraestructura computacional



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