

# Una Introducción a la Analítica

## Algunos casos de uso de Analítica Predictiva

Esta presentación describe, bajo un marco común, los conceptos fundamentales de Data Science, Analytics y Big Data y establece su similitudes y diferencias; se presentan ejemplos de casos prácticos de la aplicación de Machine Learning y Aprendizaje Estadístico.

Descargue la última versión de este documento de:  
<https://github.com/jdvelasq/data-science-docs/blob/master/sena.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

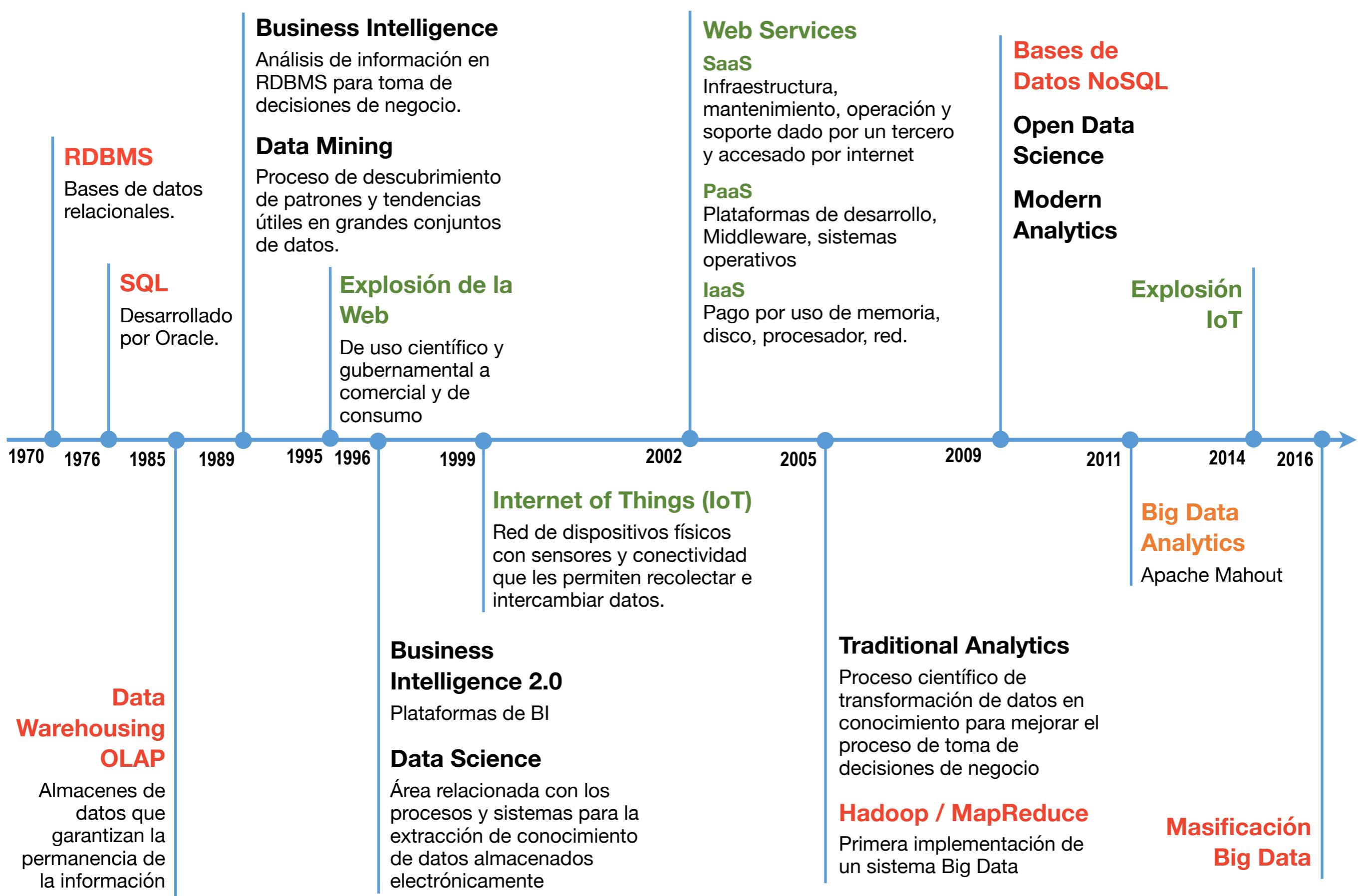
 jdvelasq@unal.edu.co

 @jdvelasquezh

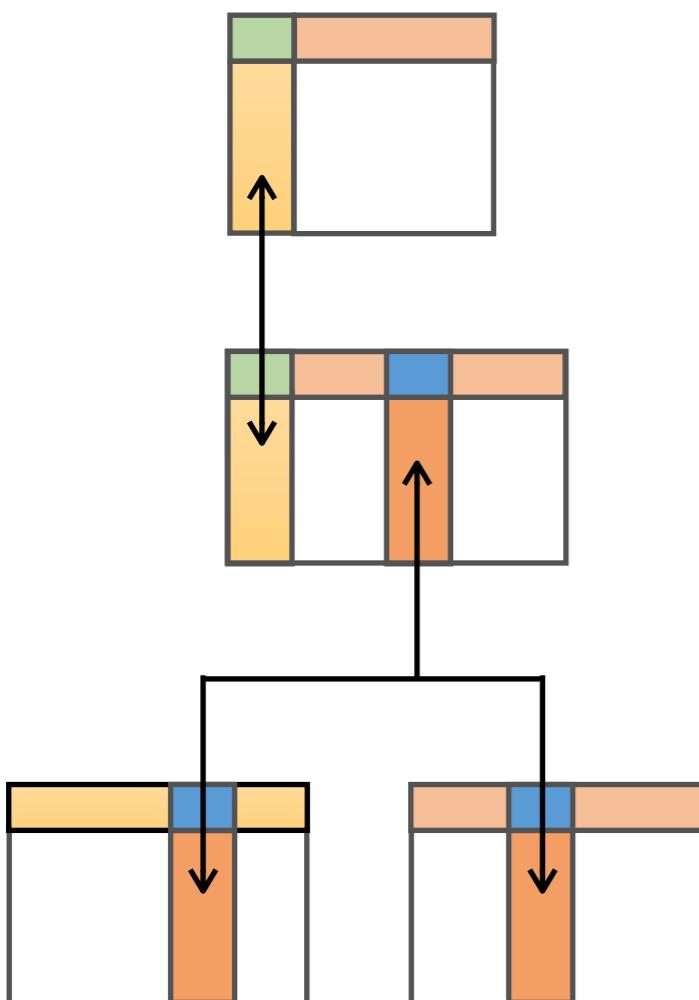
 <https://github.com/jdvelasq>

 <https://goo.gl/prkjAq>

 <https://goo.gl/vXH8jy>



# RDBMS – Relational Database Management System (1970)



## 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 (1976)

## 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);

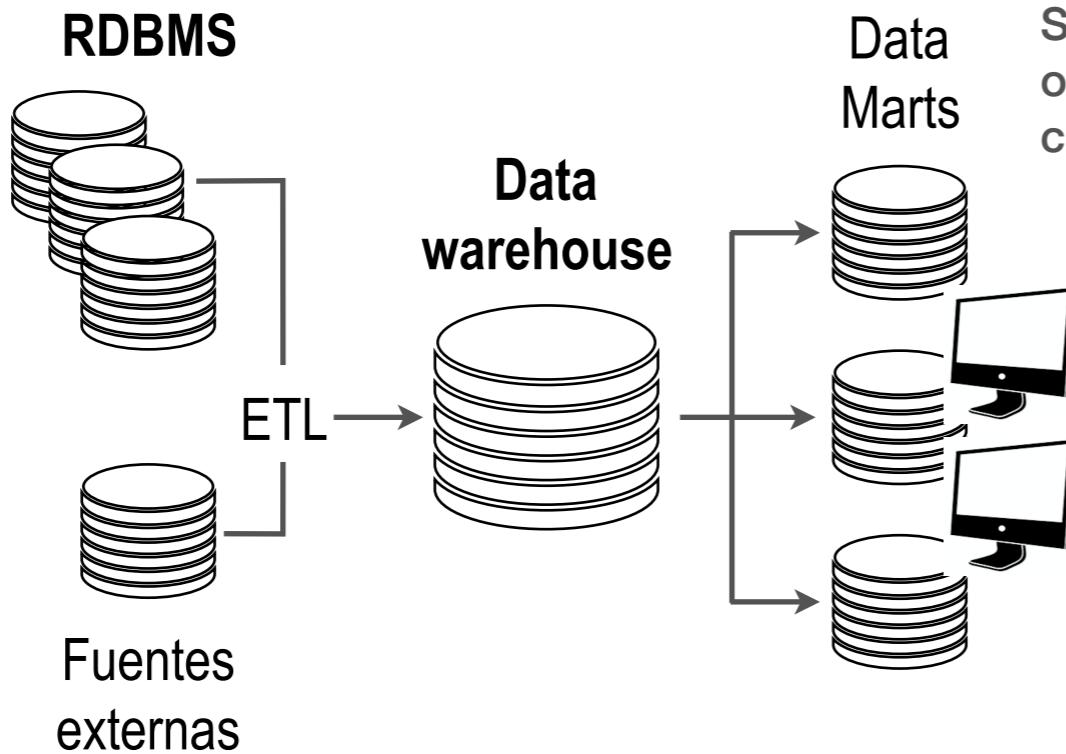
INSERT INTO phone_book2 ( [name], [phoneNumber] )
SELECT [name], [phoneNumber]
FROM phone_book
WHERE name IN ('John Doe', 'Peter Doe')

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

# Data Warehouse (1985)

## 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)

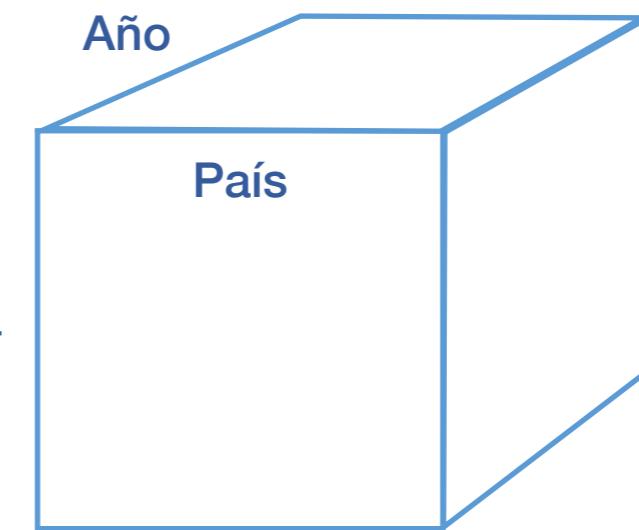
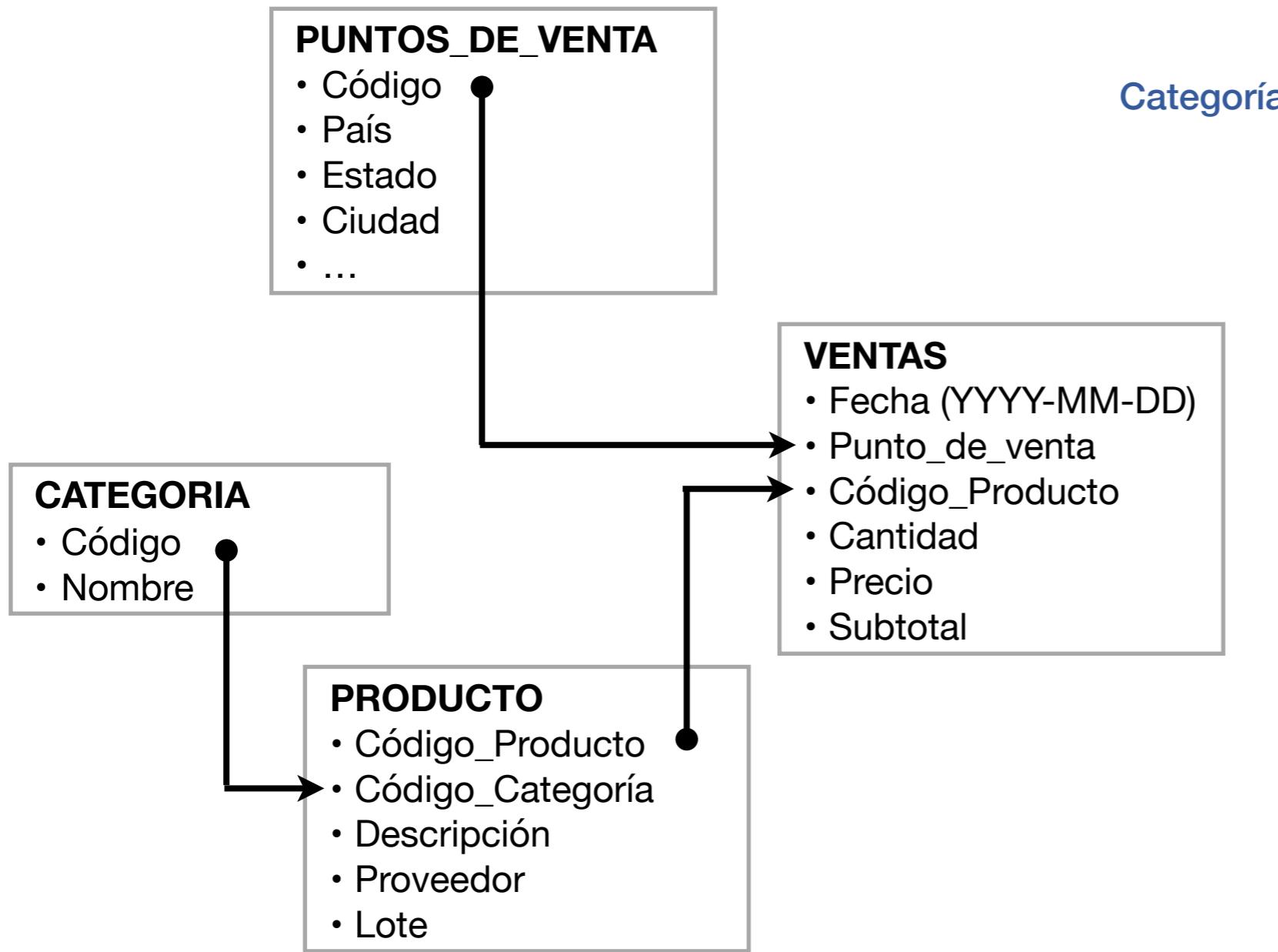
## Data Warehouse

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 (1985)

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



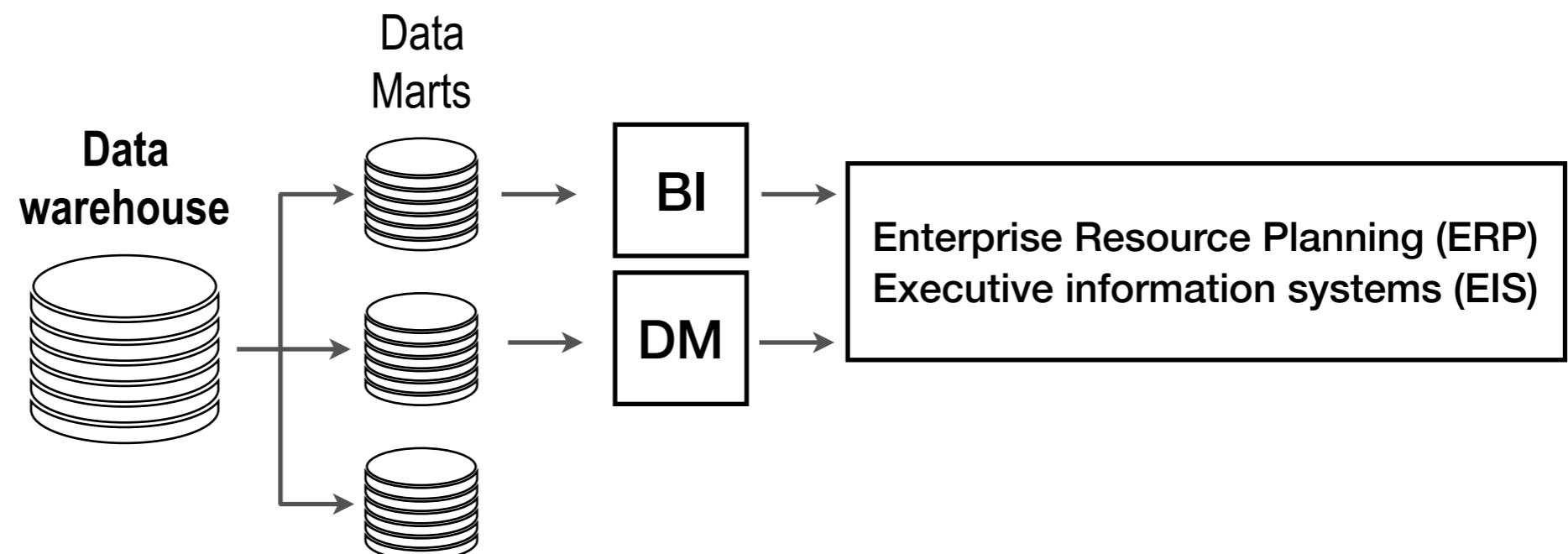
**Cubo OLAP**  
Arreglo multidimensional para consultar información

# Data Mining (1989)

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.

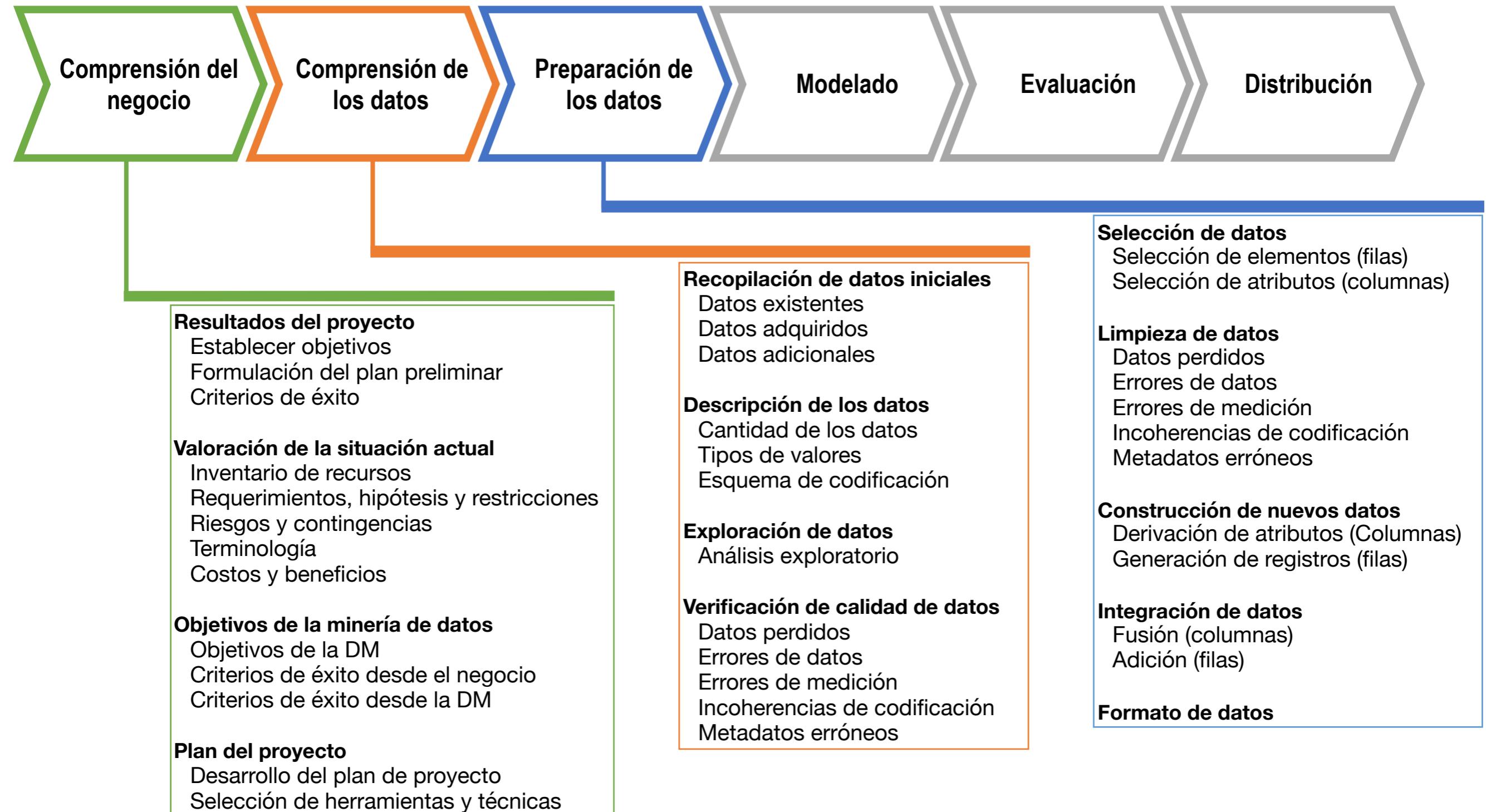
Tareas típicas:

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



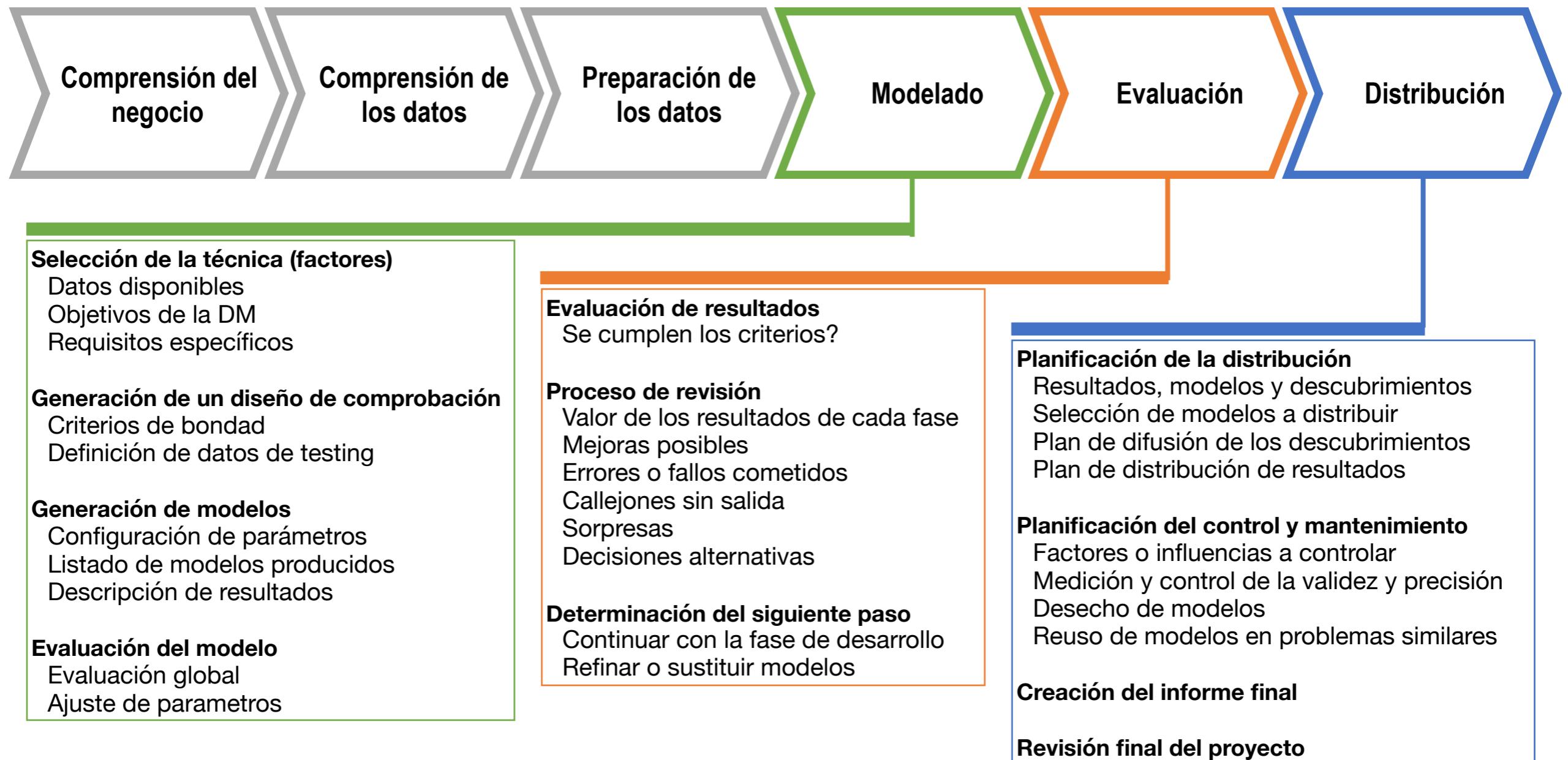
# CRISP-DM (1996)

Cross-industry standard process for data mining



# CRISP-DM (1996)

Cross-industry standard process for data mining



# Business Intelligence 2.0 (1996)

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 GENSCAPE website. The top navigation bar includes links for Solutions, Knowledge Center, Events, Blog, News, and About. Below this, there's a secondary navigation bar with 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 energyone Energy Dashboard. The top navigation bar includes links for HOME, ABOUT US, MARKETS SERVED, PRODUCTS, SERVICES, INVESTORS, and CONTACT US. Below this is a section titled 'FEATURES' with a list of capabilities. The main feature is the 'EnergyDashboard', which is shown as a complex grid of charts, graphs, and status indicators for various energy markets. The dashboard includes sections for market data, portfolio status, contracted positions, and energy operations. At the bottom, there's a 'CONTACT US' button.

**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

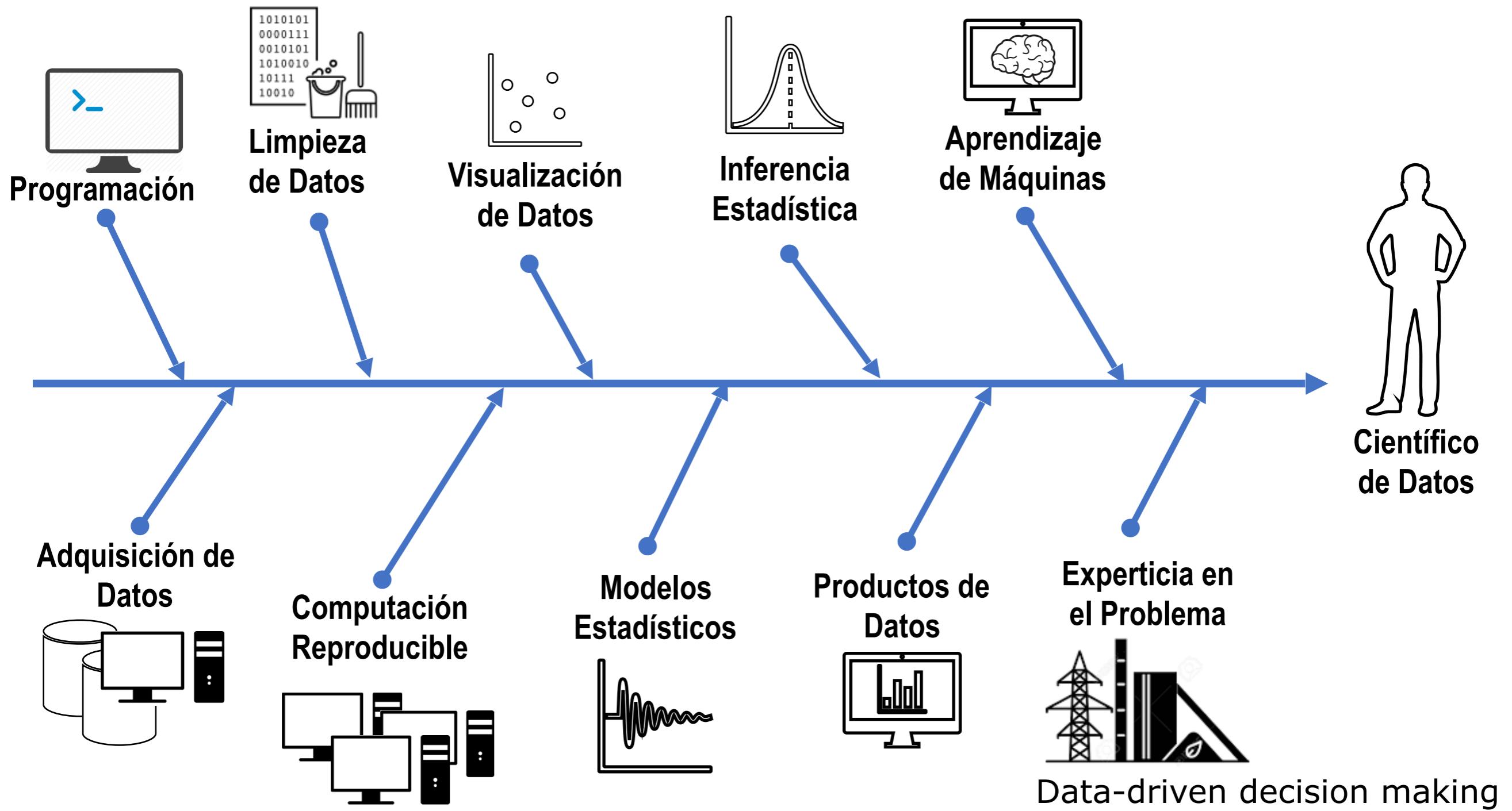
Furthermore, users can easily switch between the various aspects of the operational functions, seamlessly moving between market operations and bidding and contracts.

**CONTACT US**

# Business Intelligence 2.0 (1996)

Disciplina	Tecnología	Habilidades	Foco
Análisis de datos	<ul style="list-style-type: none"><li>Software para modelado de datos</li><li>Software para diagramación</li><li>Software para documentación</li><li>SQL</li><li>Software para perfilado de datos</li></ul>	<ul style="list-style-type: none"><li>Modelado de datos</li><li>Análisis del negocio</li><li>Manipulación de datos</li><li>Estadística básica</li></ul>	<ul style="list-style-type: none"><li>Reglas de negocio</li><li>Definición de datos</li><li>Relaciones entre datos</li><li>Atributos de datos</li><li>Estructuras de datos</li><li>Fuentes y usos de datos</li><li>Calidad de datos</li></ul>
Inteligencia de Negocios	<ul style="list-style-type: none"><li>ETL/SQL</li><li>RDBMS</li><li>Reportes</li><li>Visualización</li></ul>	<ul style="list-style-type: none"><li>Programación</li><li>Análisis de datos</li><li>Modelado de datos</li><li>Desarrollo de reportes</li><li>Estadística Básica</li><li>Análisis del negocio &amp; Estrategia</li><li>Presentación oral</li></ul>	<ul style="list-style-type: none"><li>Suministro de información y reporte</li><li>Visualización de datos</li><li>Estadísticos descriptivos</li><li>Integración de datos y consolidación</li></ul>
Data Mining	<ul style="list-style-type: none"><li>Software estadístico</li><li>Herramientas de aprendizaje de máquinas</li><li>Lenguajes de programación</li></ul>	<ul style="list-style-type: none"><li>Programación</li><li>Modelado de datos</li><li>Estadística Avanzada</li><li>Presentación oral</li></ul>	<ul style="list-style-type: none"><li>Análisis estadístico avanzado</li><li>Manejo de grandes volúmenes de datos</li><li>Visualización de datos</li><li>Modelos de datos</li></ul>

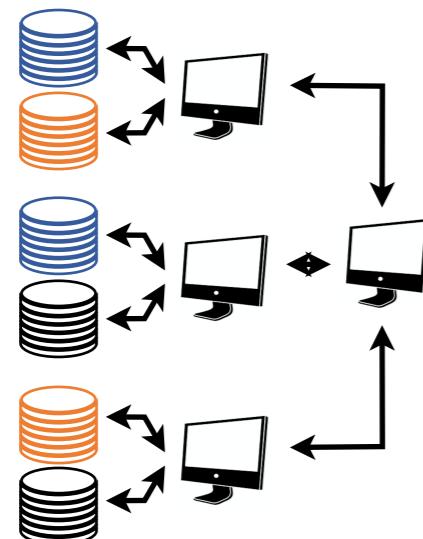
# Data Science (1996)



# Servicios Web (2002)

## 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.

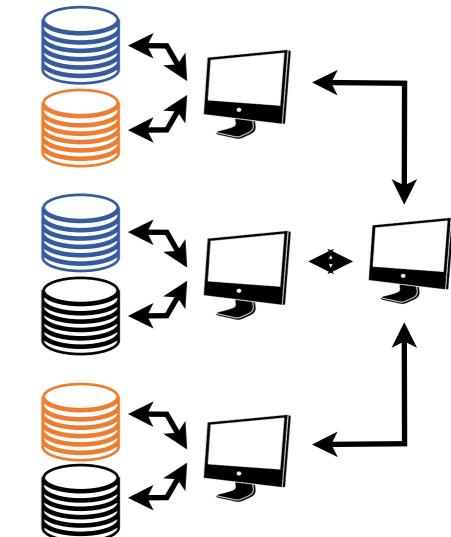
### Platform as a Service (PaaS)

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)

# Traditional Analytics (2005)



# Traditional Analytics (2005)

- Compresión de la historia.
- Pronóstico del futuro.
- Los datos están listos.
- Sólida fundamentación matemática.

## Aprendizaje Estadístico

Conjunto de herramientas fundamentadas en conceptos estadísticos para la comprensión de conjuntos de datos con el fin de modelar y predecir.

## Aprendizaje de Máquinas

Área de las ciencias de la computación orientada al desarrollo de sistemas inteligentes (solución de problemas como lo haría un experto humano).

- No se quiere comprender que pasó.
- Se desea mimificar la inteligencia.
- Pronóstico del futuro.
- Los datos están listos.
- Fundamentación matemática, pero sin el rigor de la estadística.

- No se quiere comprender que pasó.
- Pronóstico del futuro.
- Los datos están listos.
- Sólida fundamentación matemática.

## Modelado Predictivo

Área enfocada al uso de técnicas estadísticas utilizadas para pronosticar resultados (de un proceso).

## Minería de Datos

Descubrimiento de patrones y conocimiento de grandes conjuntos de datos usando técnicas estadísticas, aprendizaje de máquinas y herramientas de bases de datos.

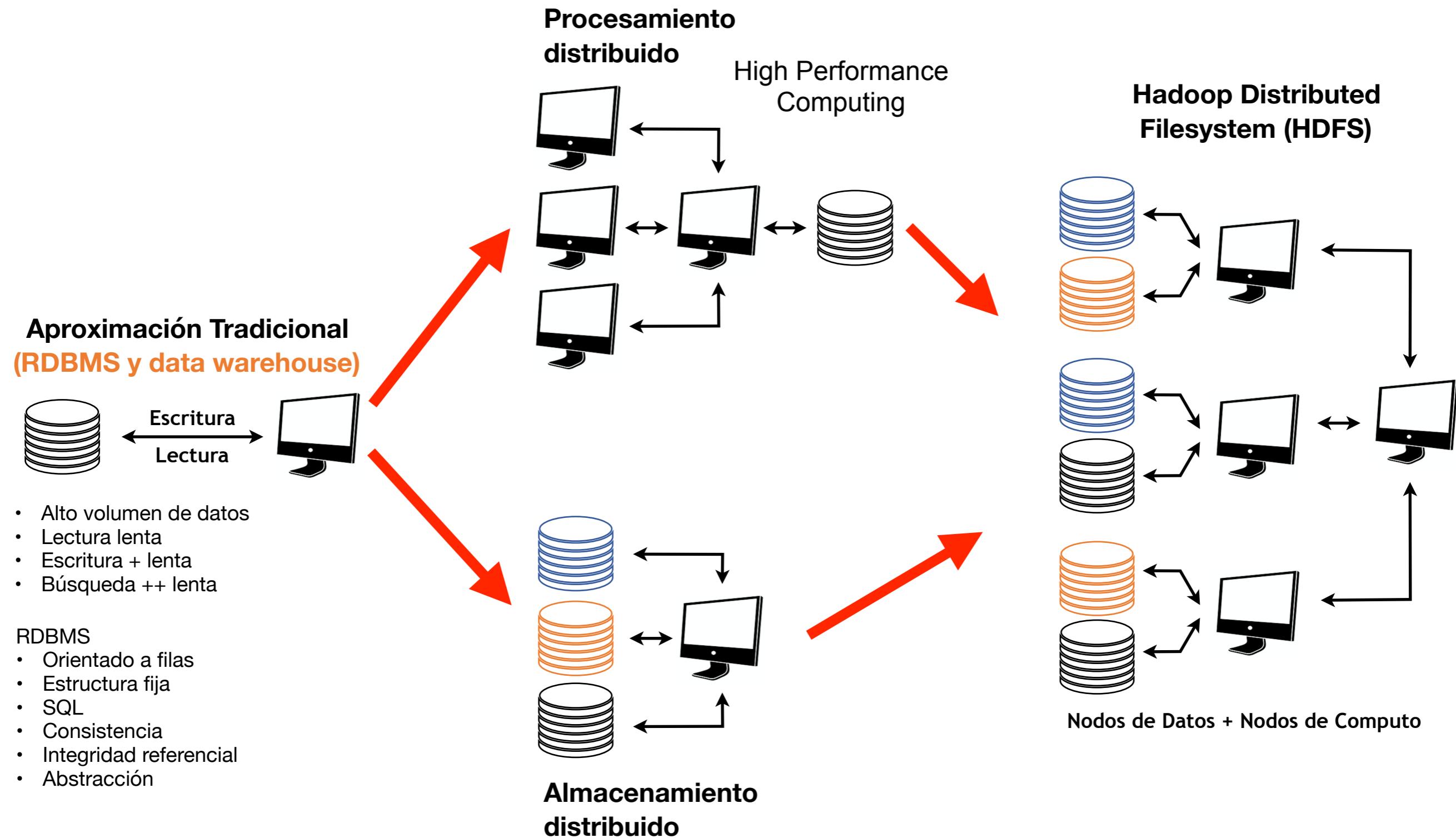
- No se quiere comprender que pasó.
- Pronóstico del futuro.
- Los datos están NO están listos.
- Sólida fundamentación matemática.

## Analítica Predictiva

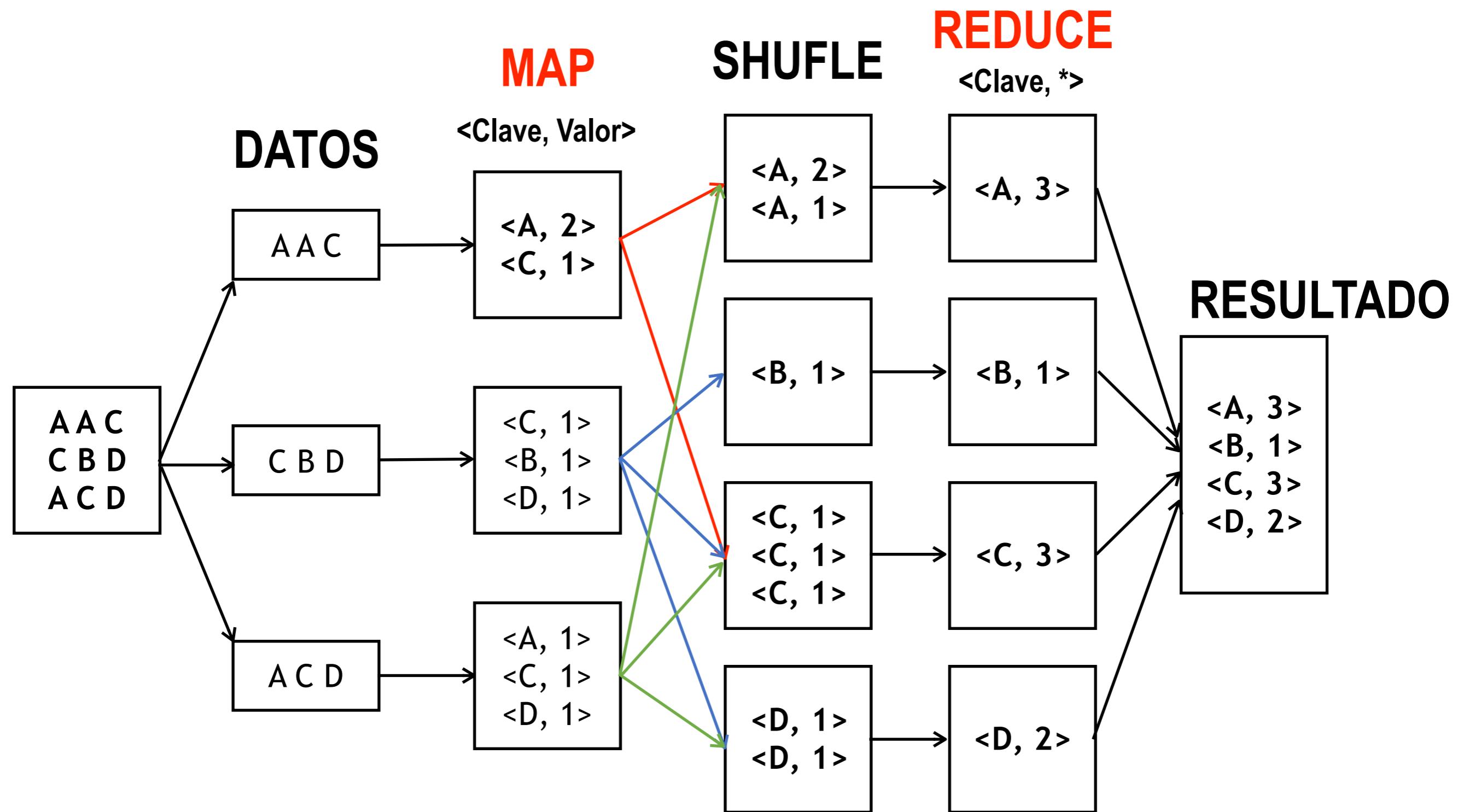
Área enfocada al uso de técnicas de modelado predictivo, aprendizaje de máquinas y minería de datos para pronosticar los resultados de un proceso en un contexto organizacional.

- Pronóstico del futuro.
- Los datos están NO están listos.
- Agrupa todas las anteriores.

# Hadoop / MapReduce (2005)



# Hadoop / MapReduce (2005)



# NoSQL (2009)

## 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

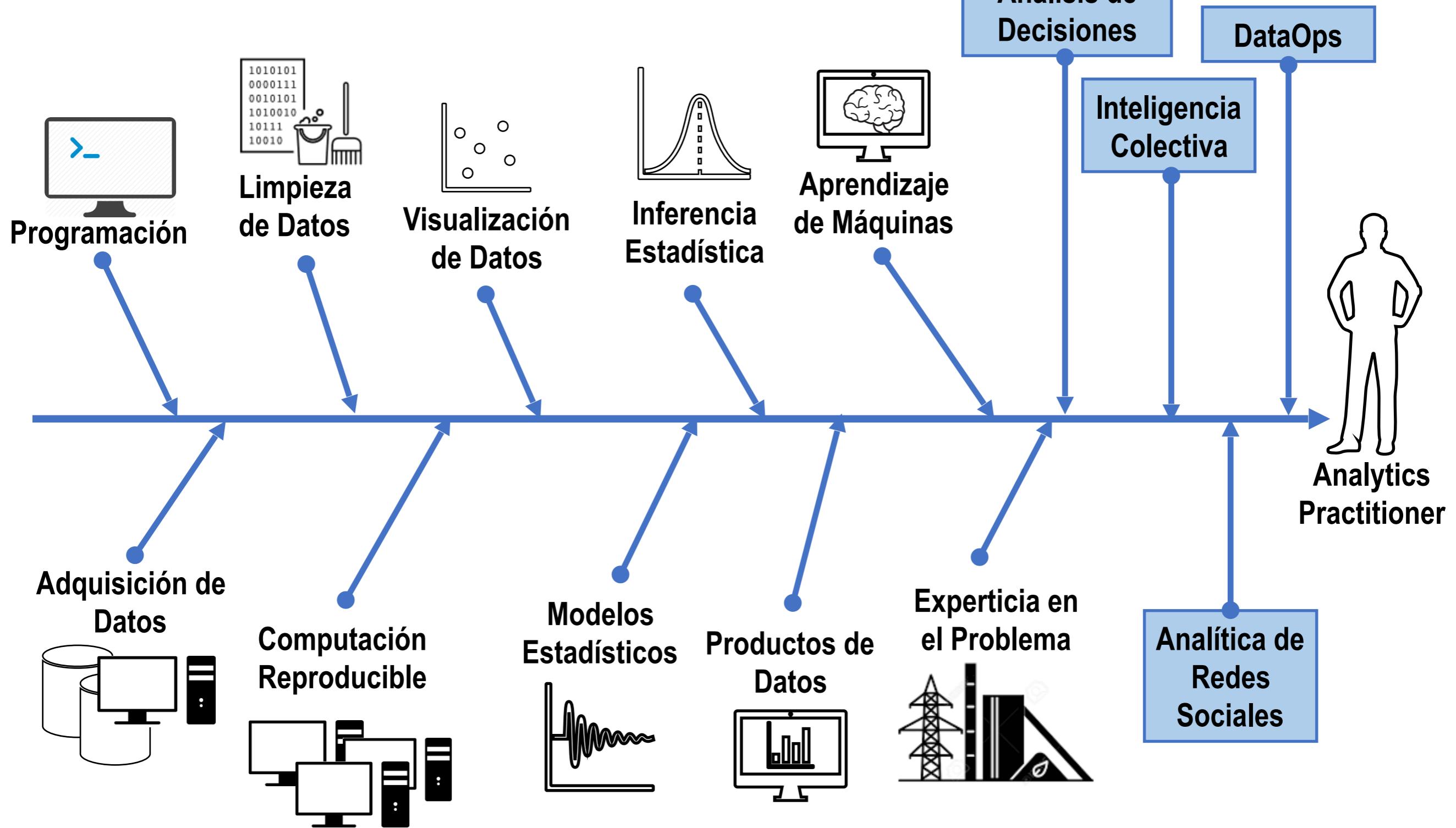
## Sistema orientado a filas

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

## Column family database

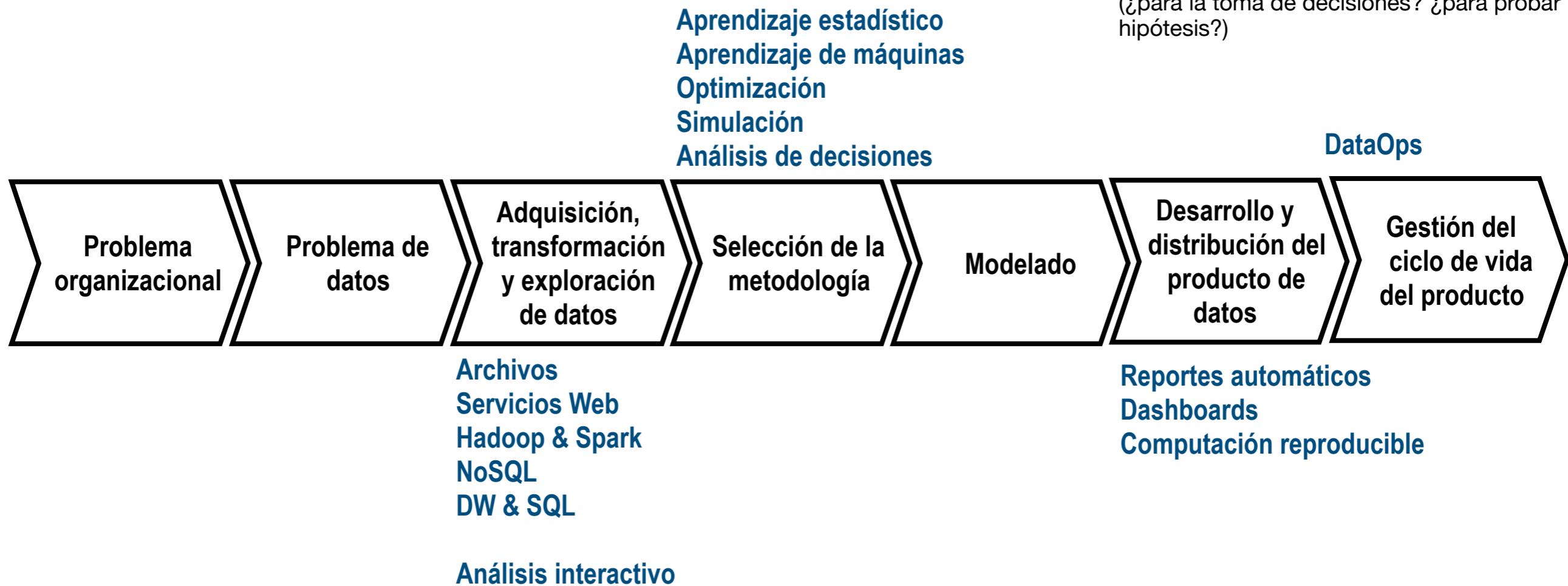
001:{Fecha:2017-10-01, Planta:Jaguas, Generación:100.2}  
002:{Fecha:2017-10-01, Planta:Playas, Generación:23.1}  
003:{Fecha:2017-10-01, Planta:Guatapé, Generación:130.1}

# Open Data Science & Modern Analytics (2009)



# Open Data Science & Modern Analytics (2009)

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

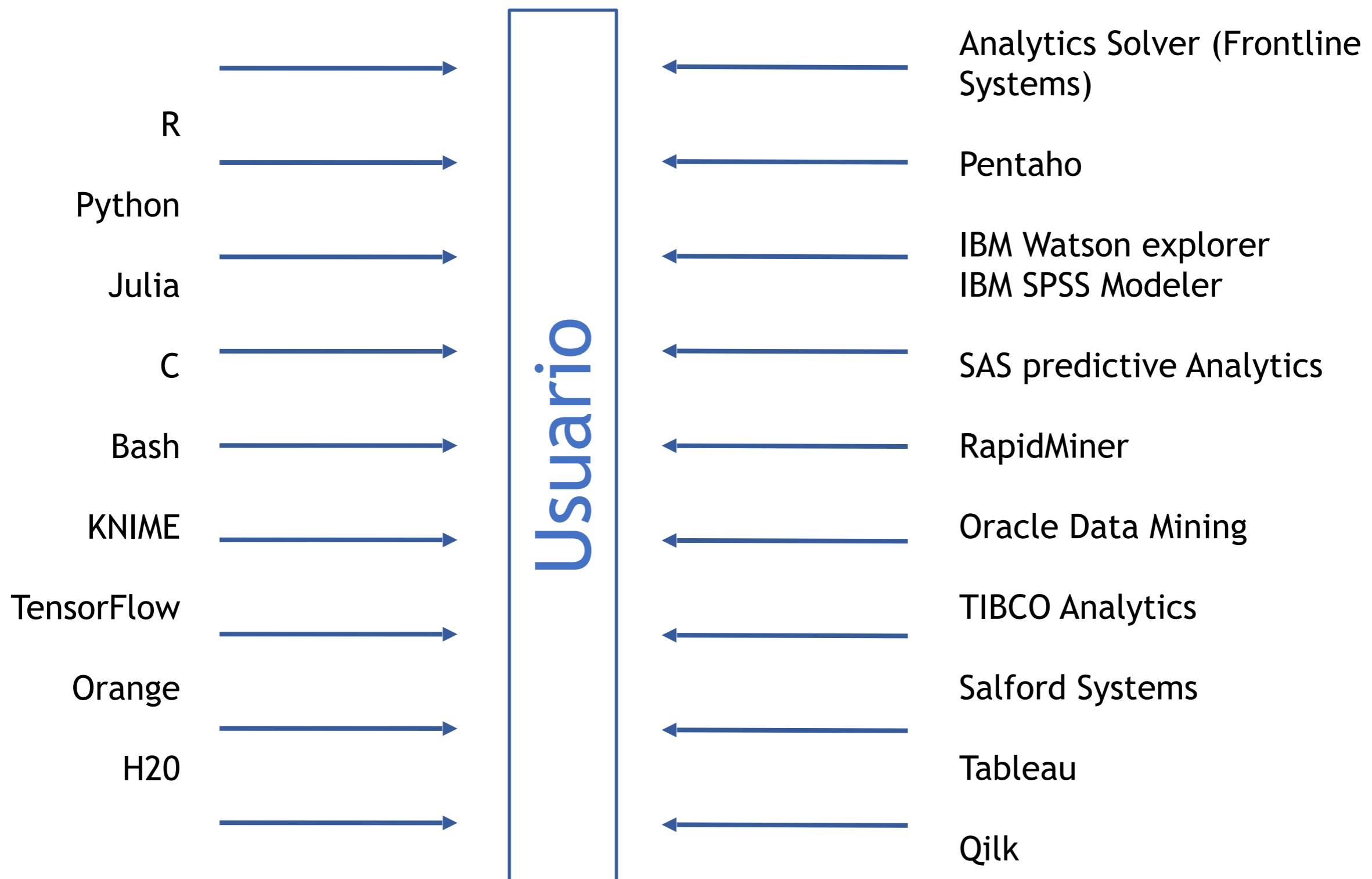
## Data Mining

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?)

# Open Data Science & Modern Analytics (2009)



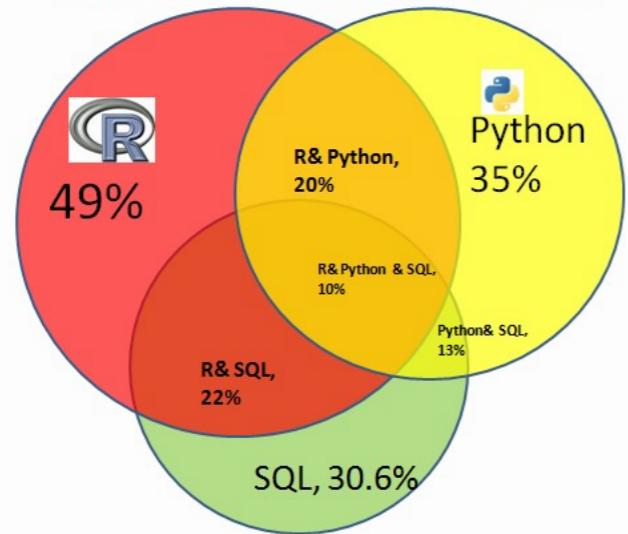
# Open Data Science & Modern Analytics (2009)

The Top Ten Programming Languages  
(IEEE Spectrum)

**2015**

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

KDNuggets 2014 Poll: Languages used for Analytics/Data Mining



**2016**

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

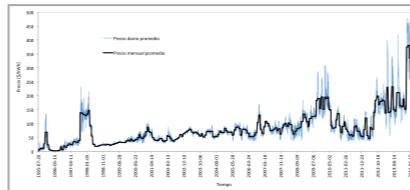
**2017**

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

# Open Data Science & Modern Analytics (2009)

Estadística y  
aprendizaje de  
máquinas

Los datos  
están listos



Modelado de  
datos

Inteligencia  
de Negocios

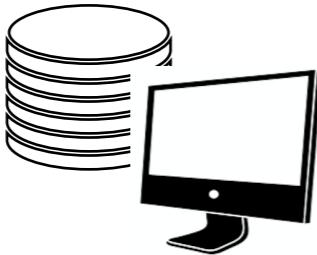
DW / OLAP



Generación,  
agregación, análisis  
y visualización de  
datos del negocio

Minería de  
Datos

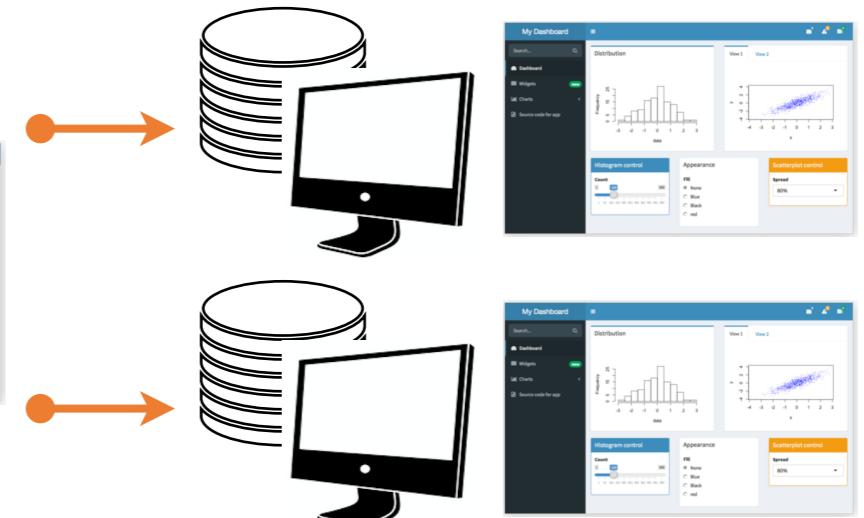
DW / OLAP



Descubrimiento de  
patrones y tendencias  
clave

Analytics

DW / OLAP  
Hadoop & Spark  
NoSQL ...

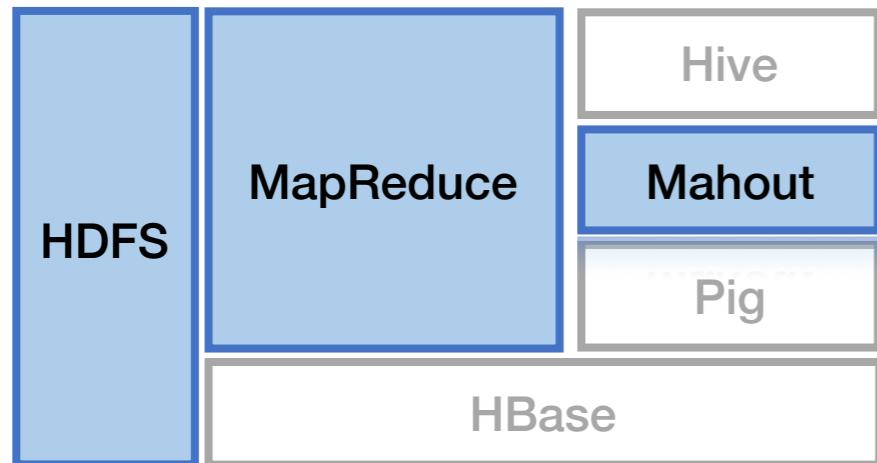


## Producto de Datos

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 (2011)



## Apache Mahout

Implementación en Map/Reduce (Java y otros) de los algoritmos de aprendizaje estadístico y aprendizaje de máquinas



## Spark's MLlib

Implementación en Spark de los algoritmos de aprendizaje estadístico y aprendizaje de máquinas

{  
Java  
Scala  
Python  
R

Estadística básica

Clasificación y regresión

Filtrado colaborativo

Agrupamiento

Reducción de dimensiones

Extracción de características

Minería de patrones frecuentes

Métricas de evaluación

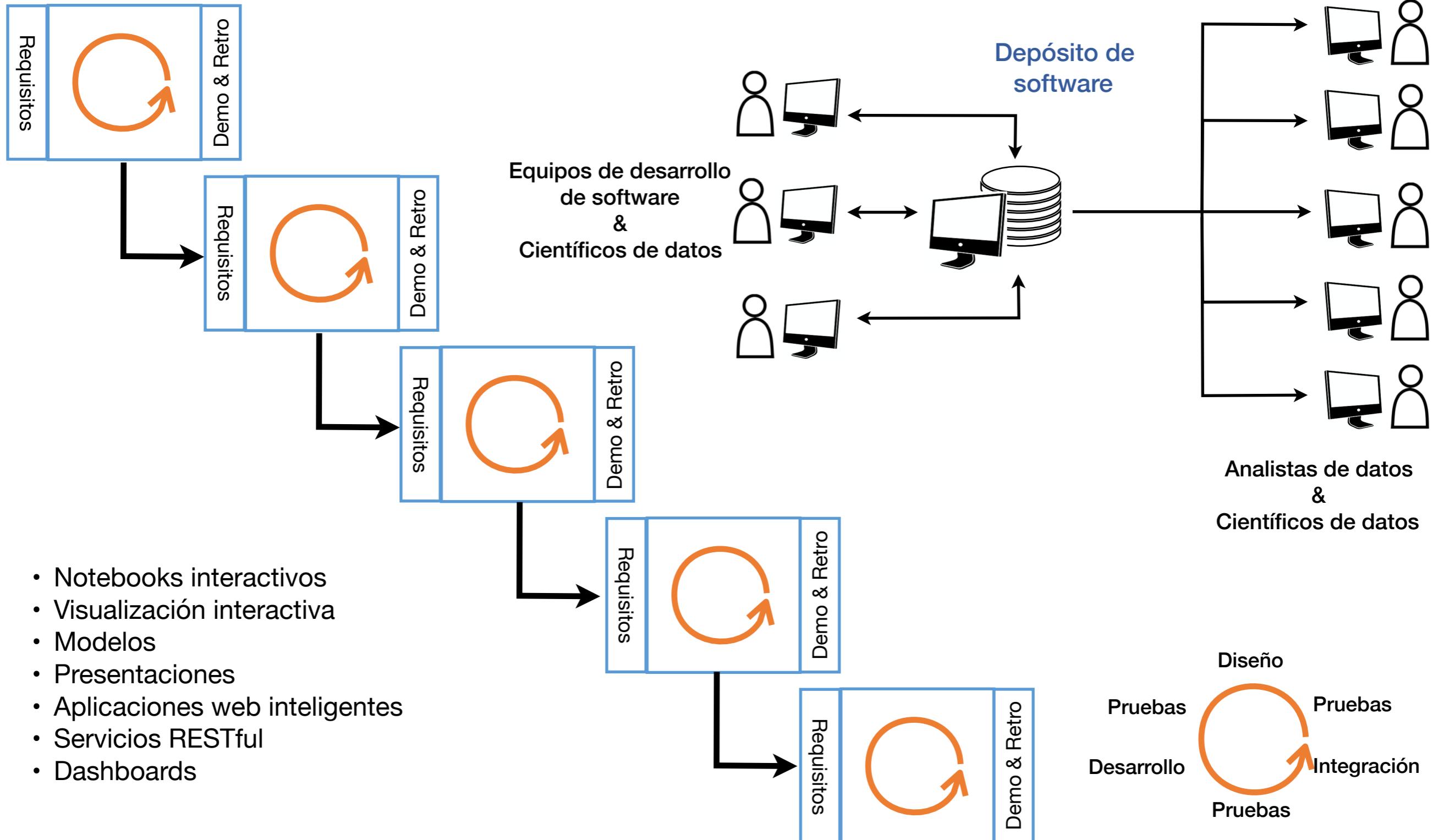
Exportación de modelos

Optimización

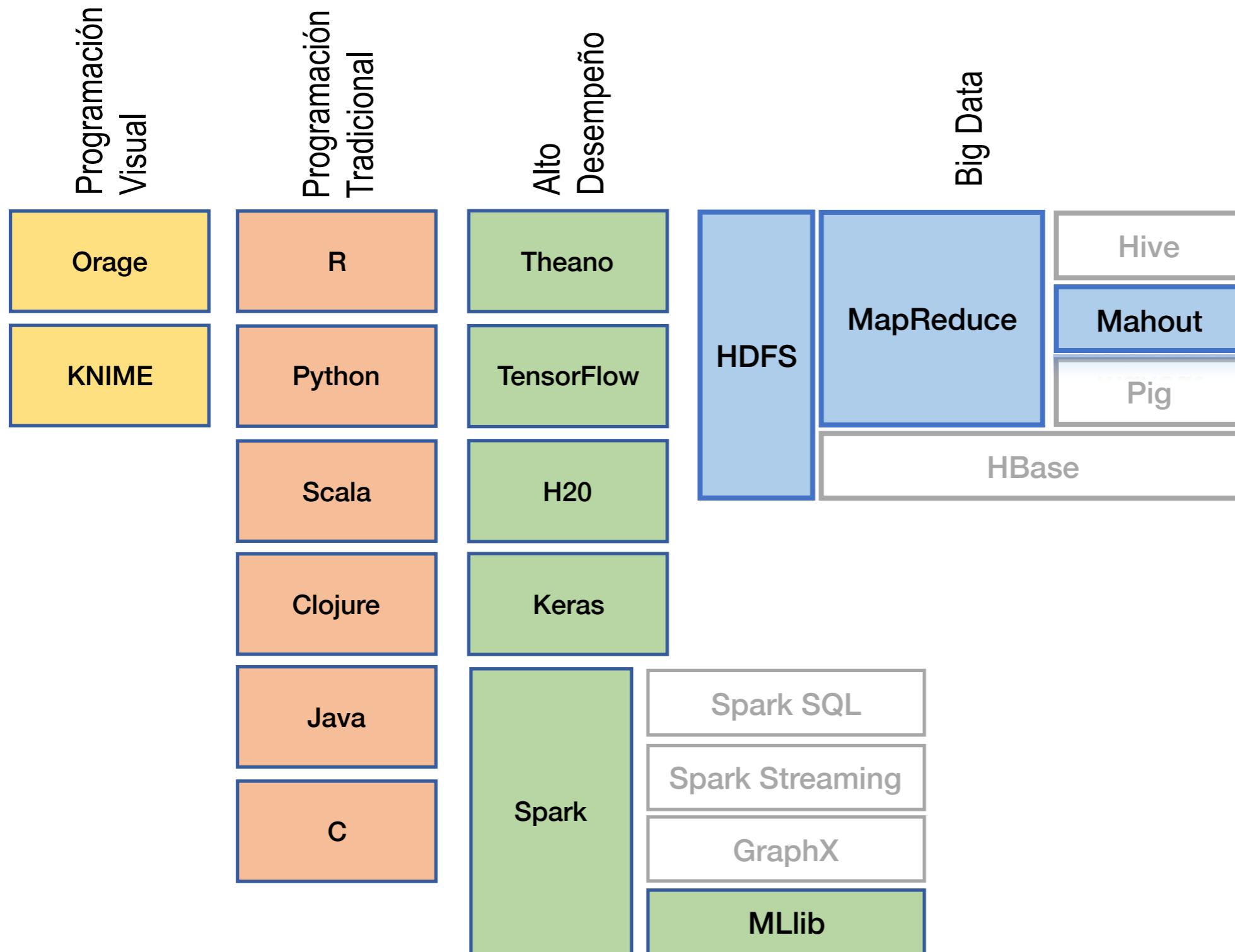


{  
Computación de alto desempeño  
Deep Learning

# DataOps (2015)



# Machine Learning & Predictive Analytics



## Agrupamiento

K-means  
k-NN  
Canopy clustering  
Fuzzy K-means  
Streaming K-means  
Text clustering

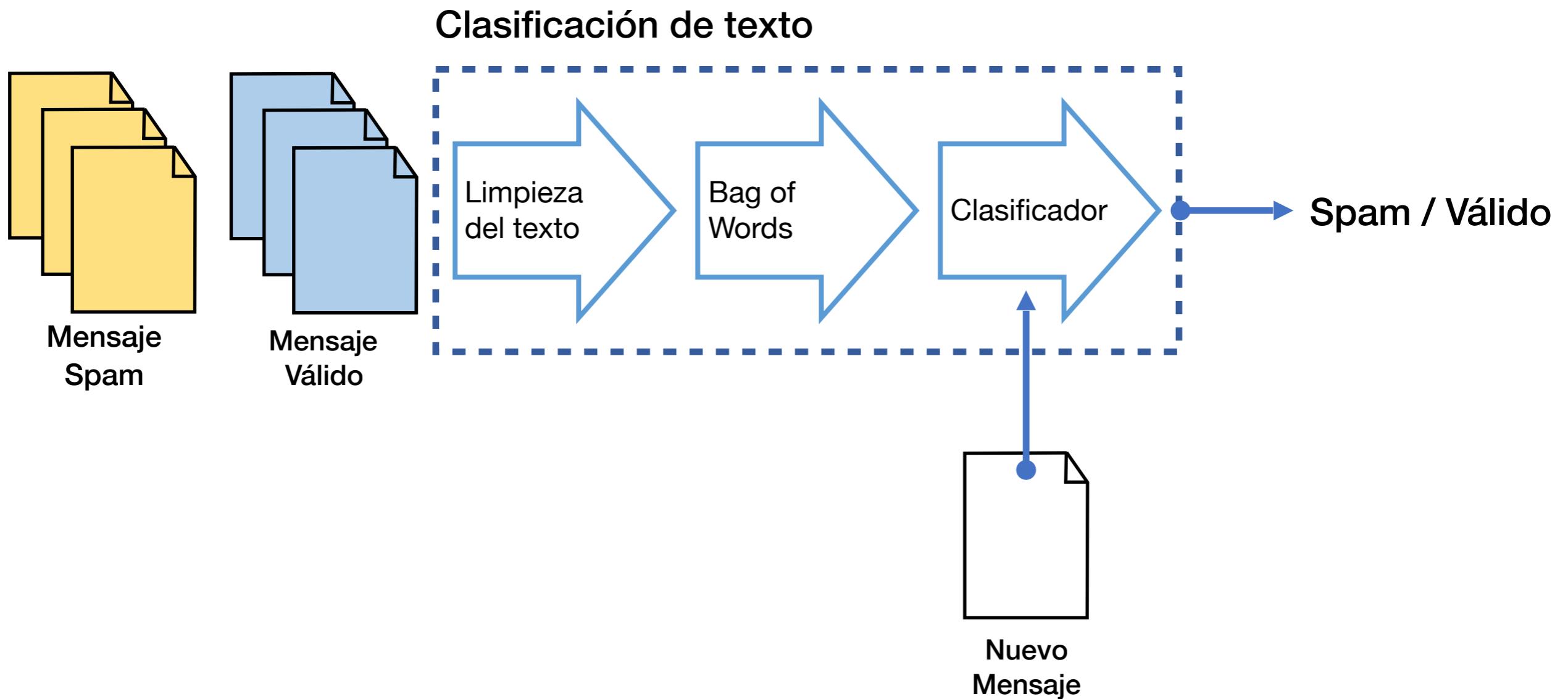
## Regresión

Regresión lineal y extensiones (ridge, LASSO, ElasticNet)  
Regresión no lineal  
Redes neuronales artificiales  
Aprendizaje Profundo

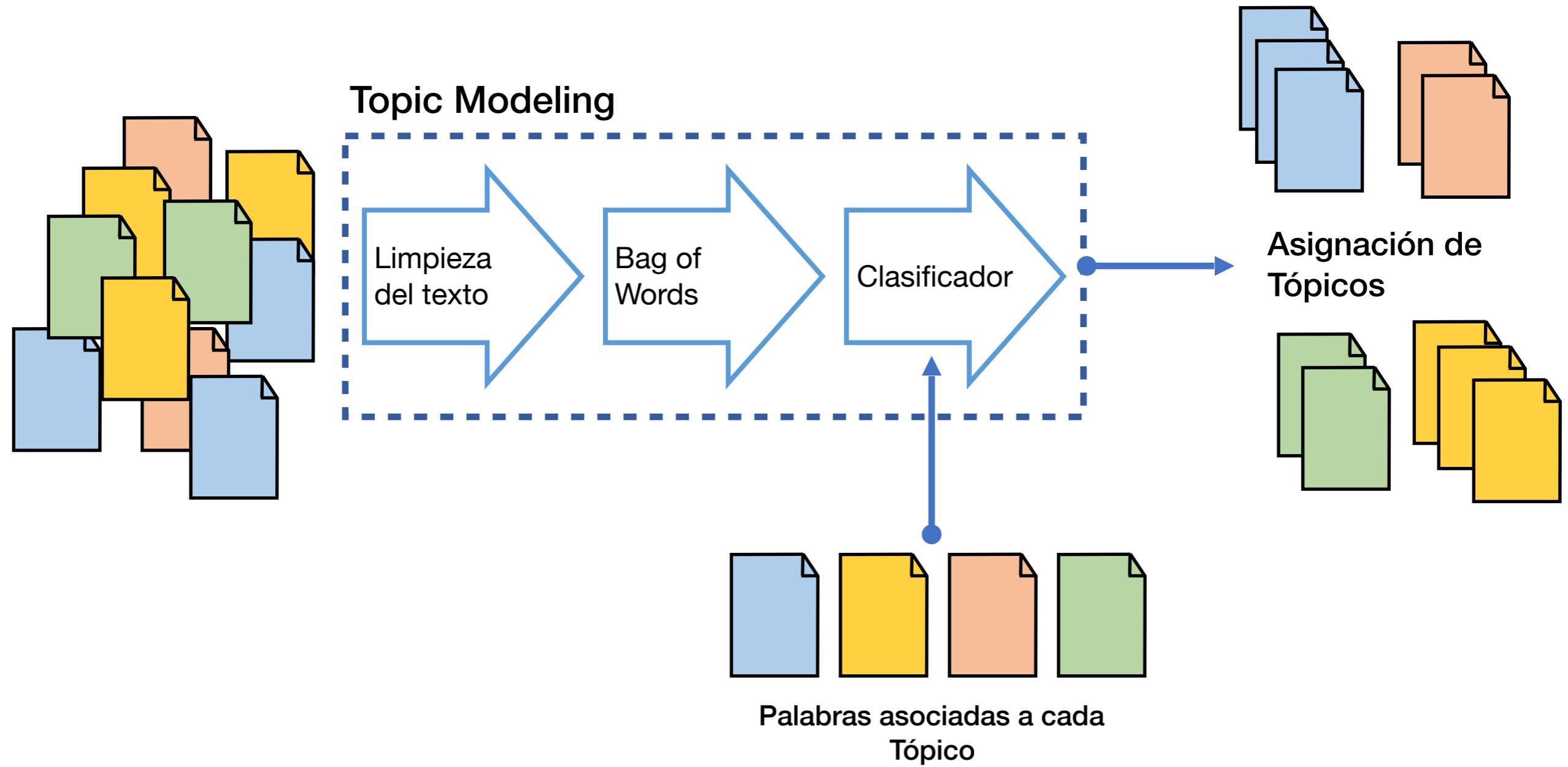
## Clasificación

K nearest neighbors  
Naive Bayes  
Rule-based systems  
Tree-based classification  
Logistic & Logit regression  
Ensembles

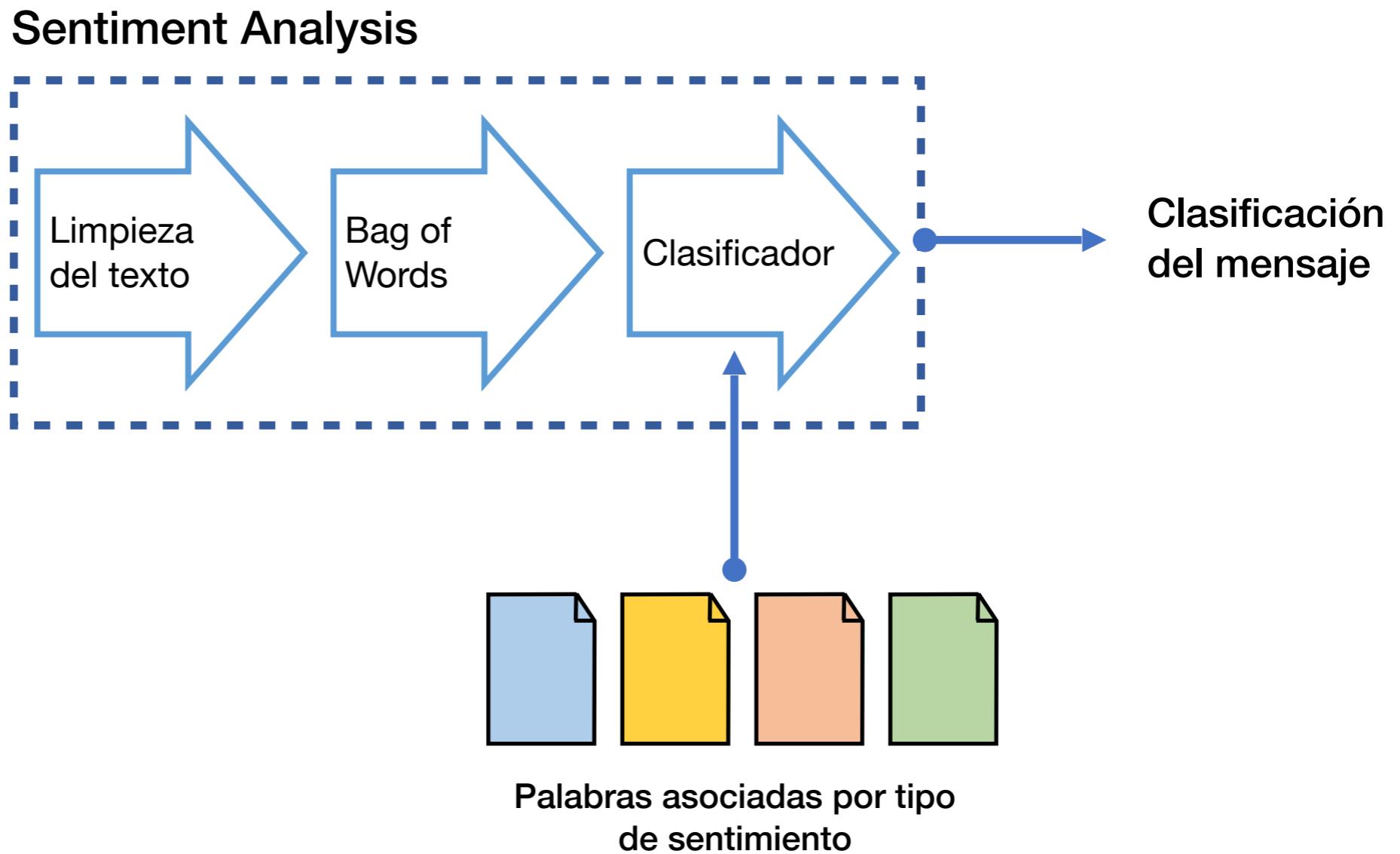
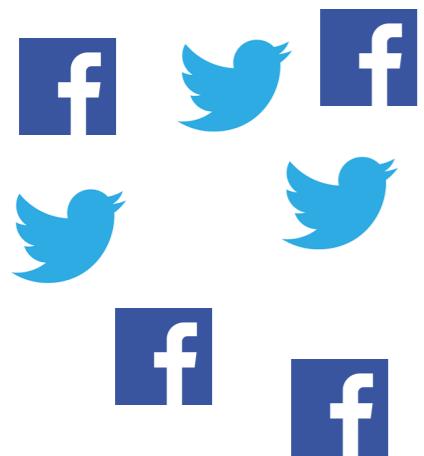
# Machine Learning & Predictive Analytics



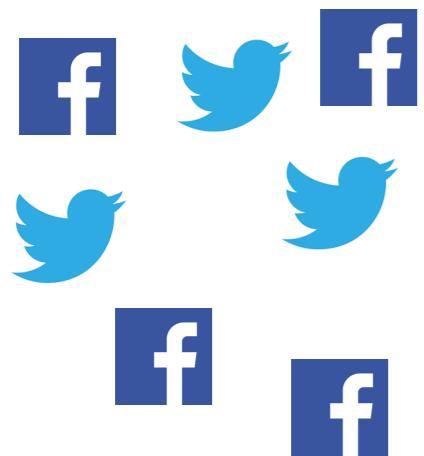
# Machine Learning & Predictive Analytics



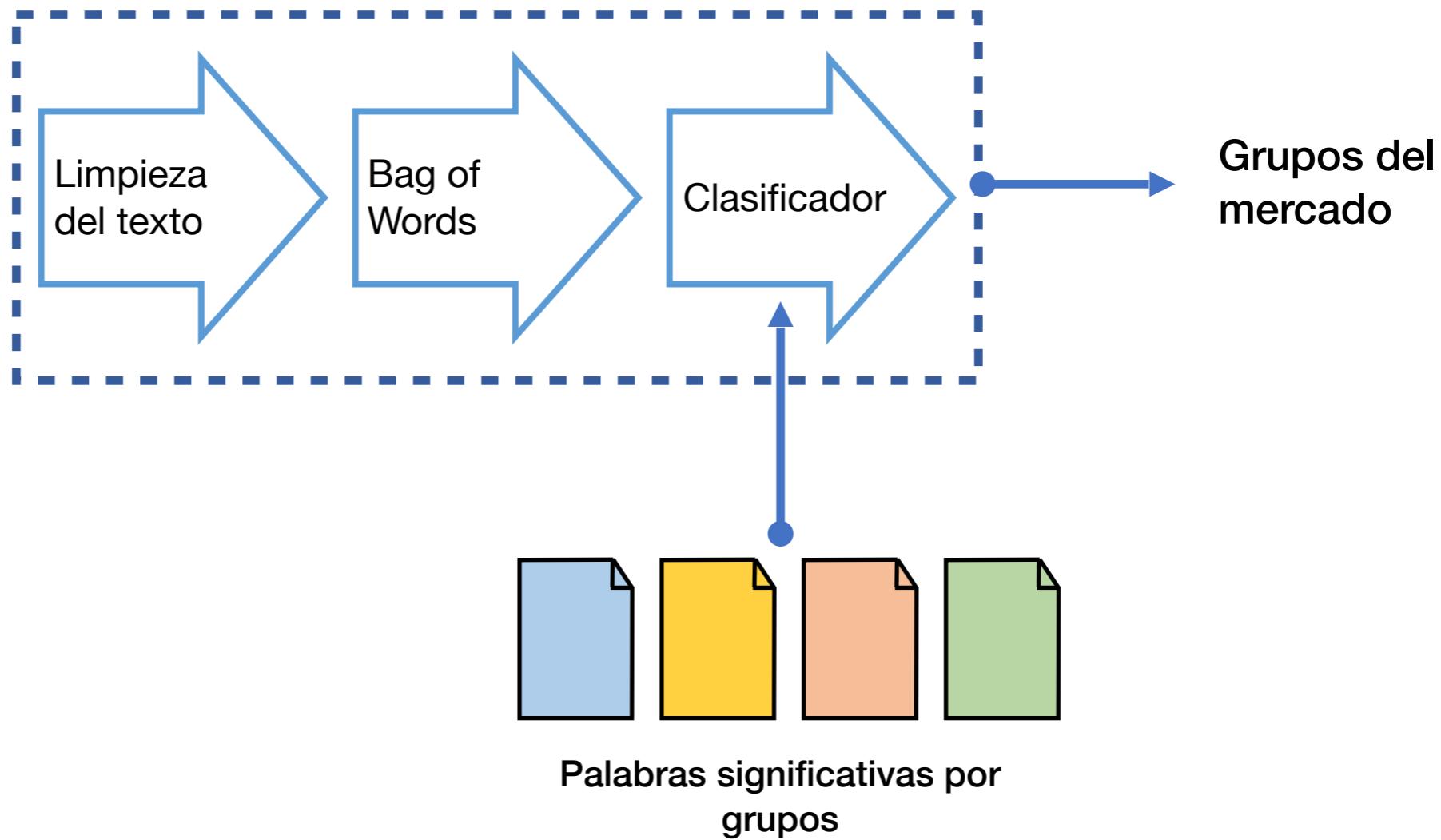
# Machine Learning & Predictive Analytics



# Machine Learning & Predictive Analytics

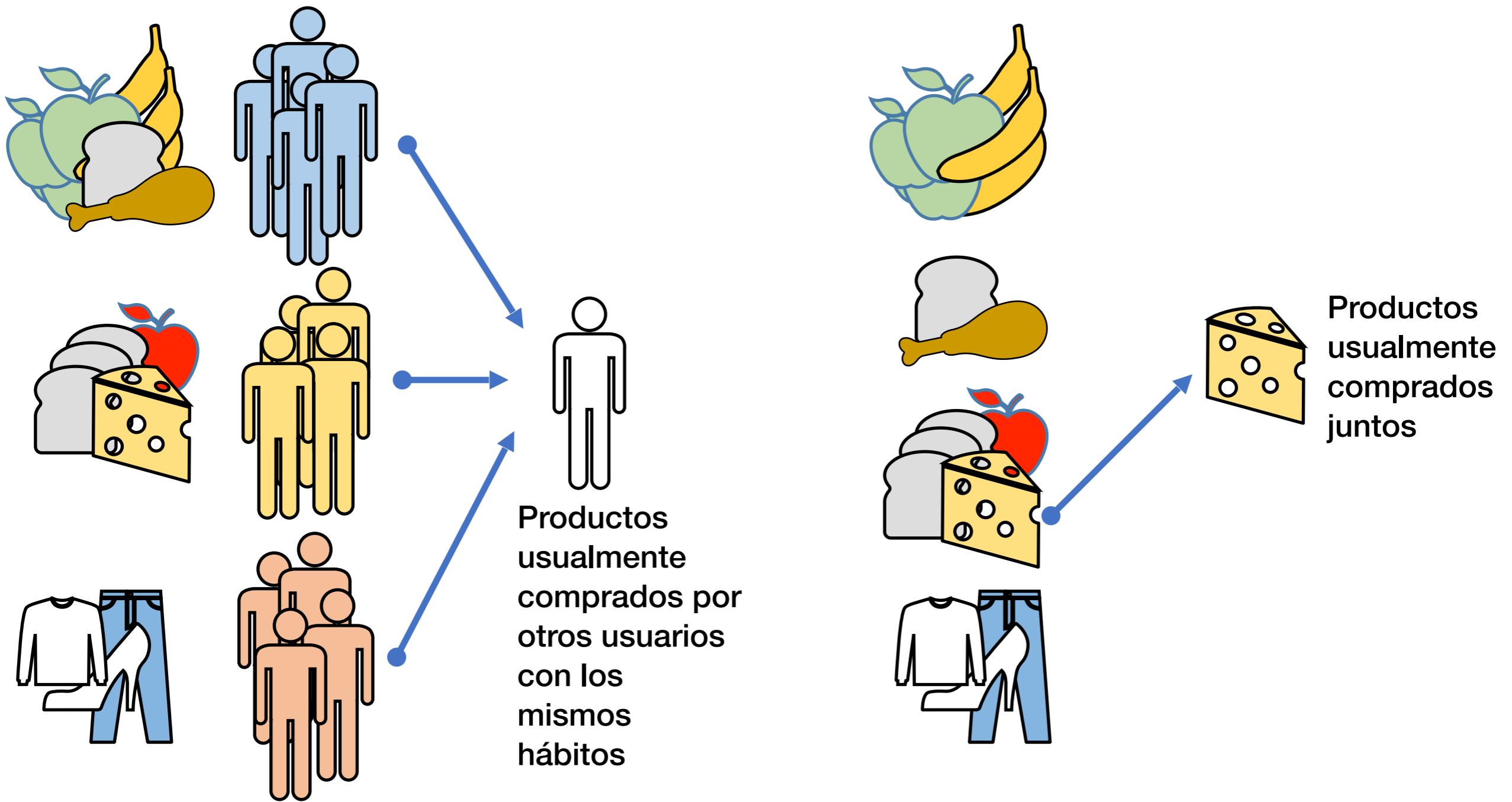


## Clasificación del mercado

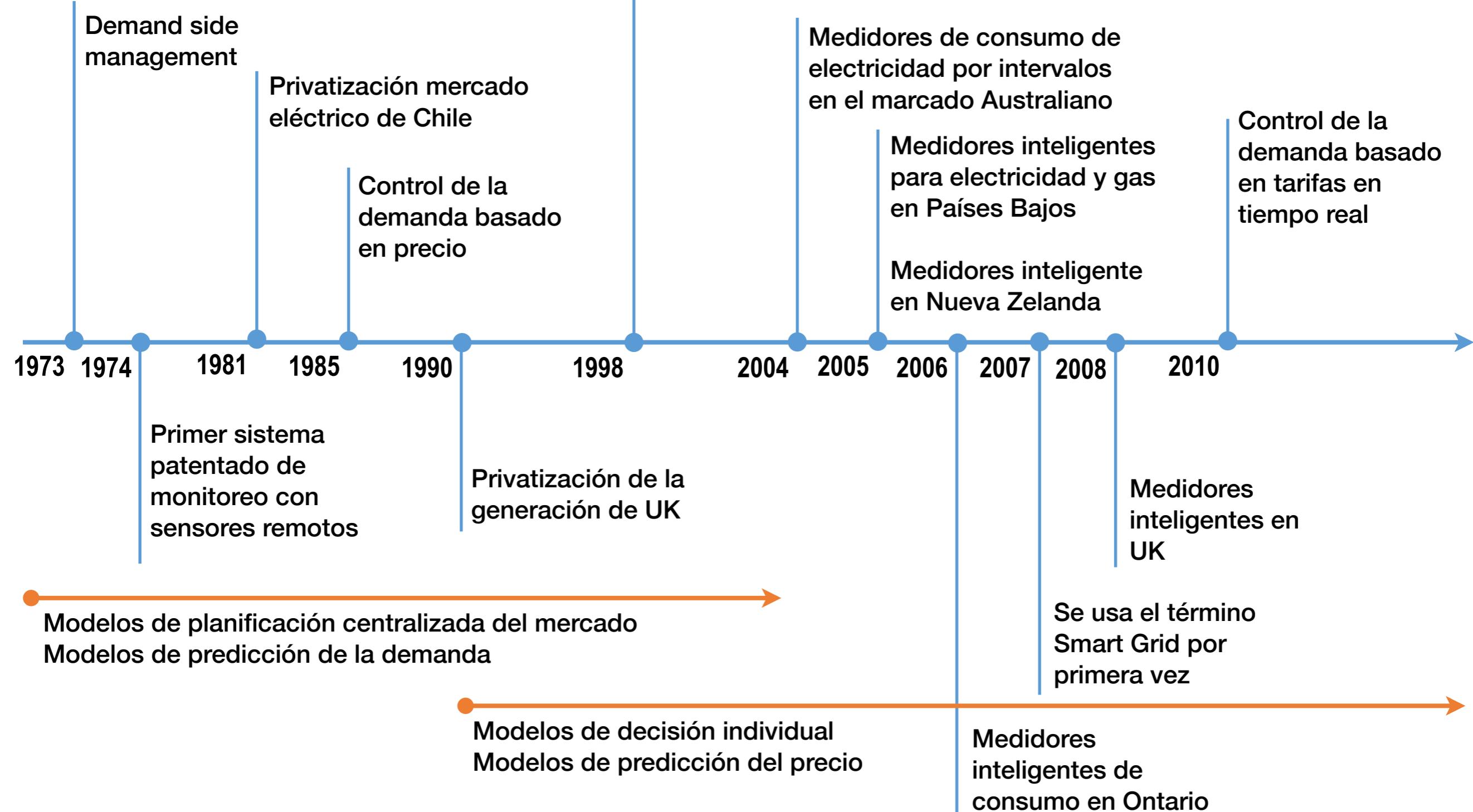


# Machine Learning & Predictive Analytics

## Association Rules & Recommender Systems



# Industria de la Energía Eléctrica



# Open Data Science & Modern Analytics

# Open Data Science & Modern Analytics (2009)

## 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

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

## 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 & Modern Analytics (2009)

## Jupyter Notebook

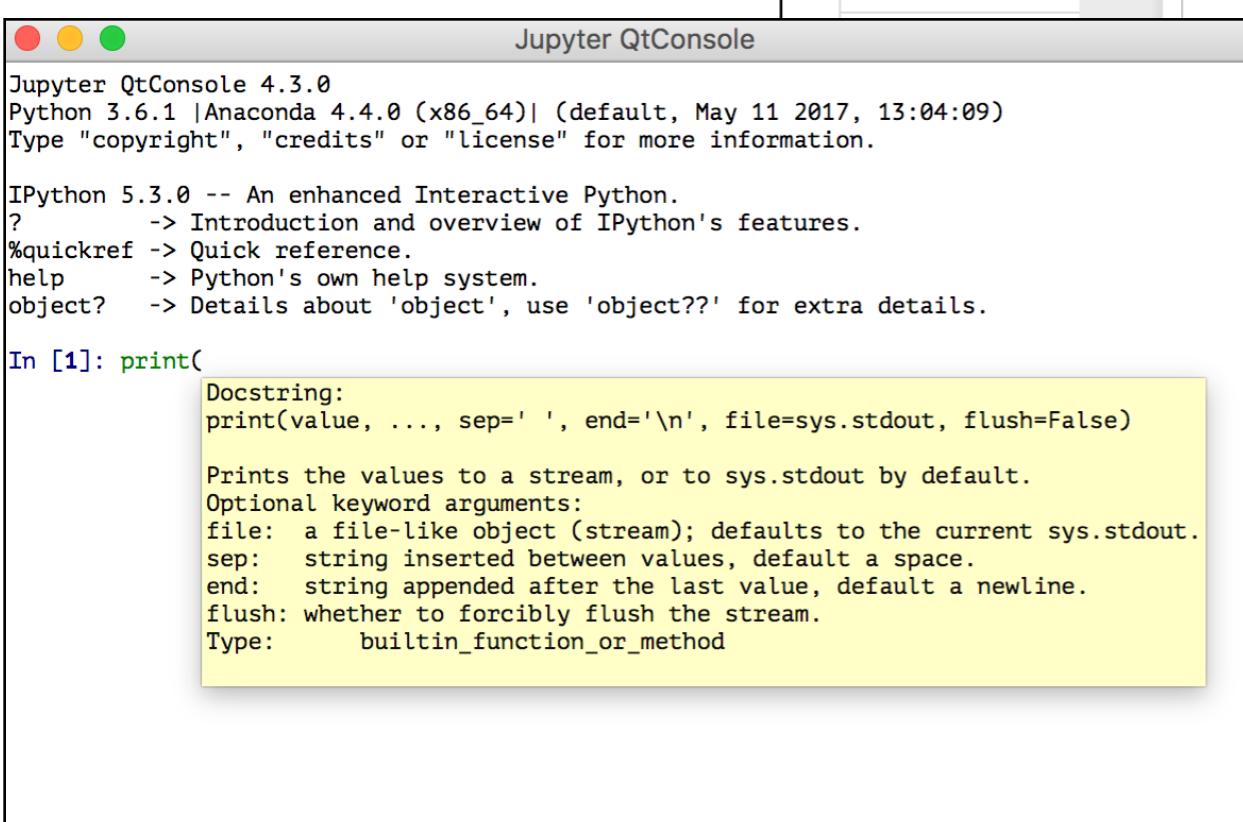
### Storyboards

- Notebooks
- Exploración interactiva
- Programación visual
- Data IDE

### Aplicaciones modernas

- Notebooks
- Dashboards
- Aplicaciones visuales
- Servicios de datos

### Jupyter QtConsole

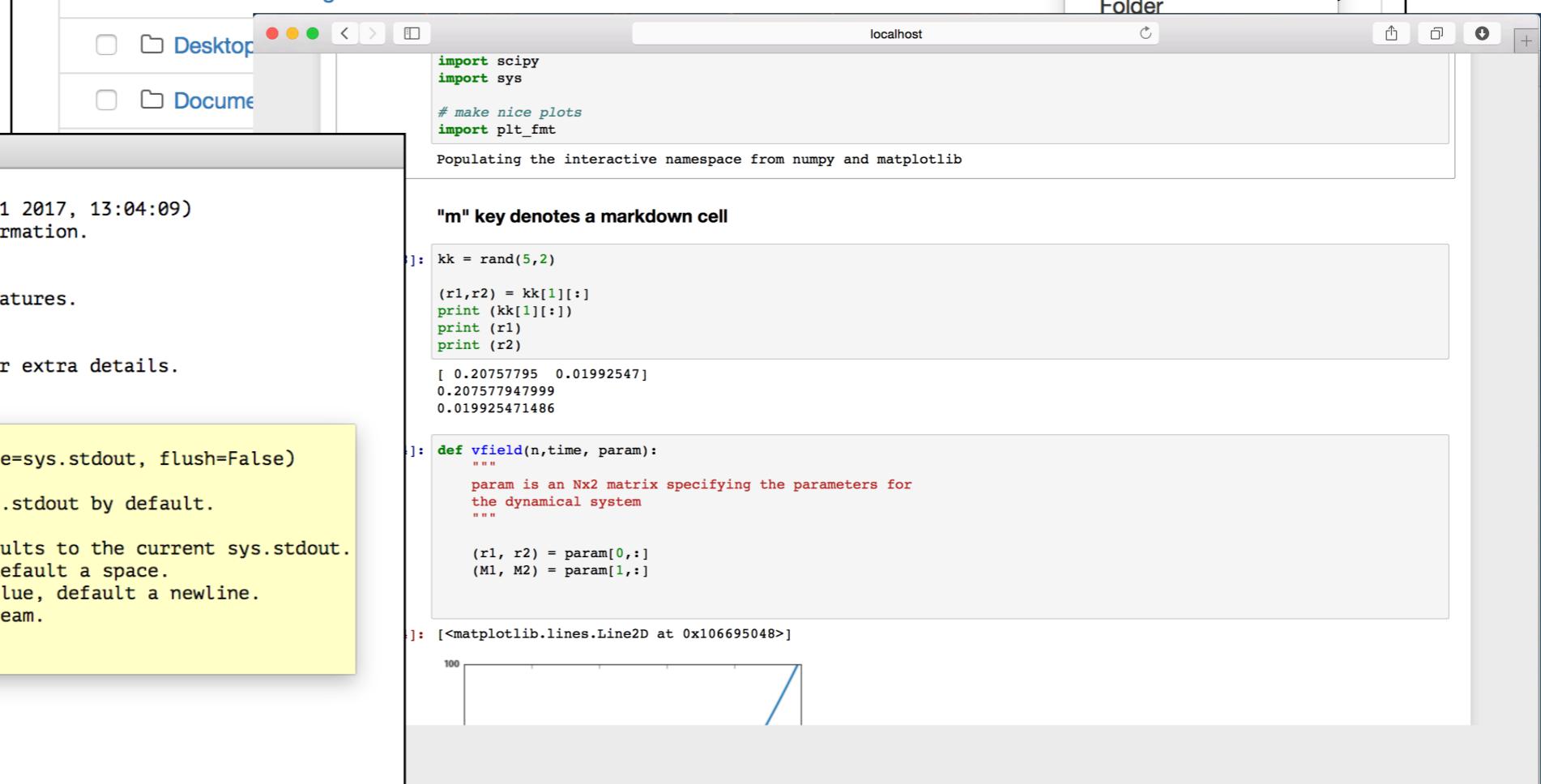
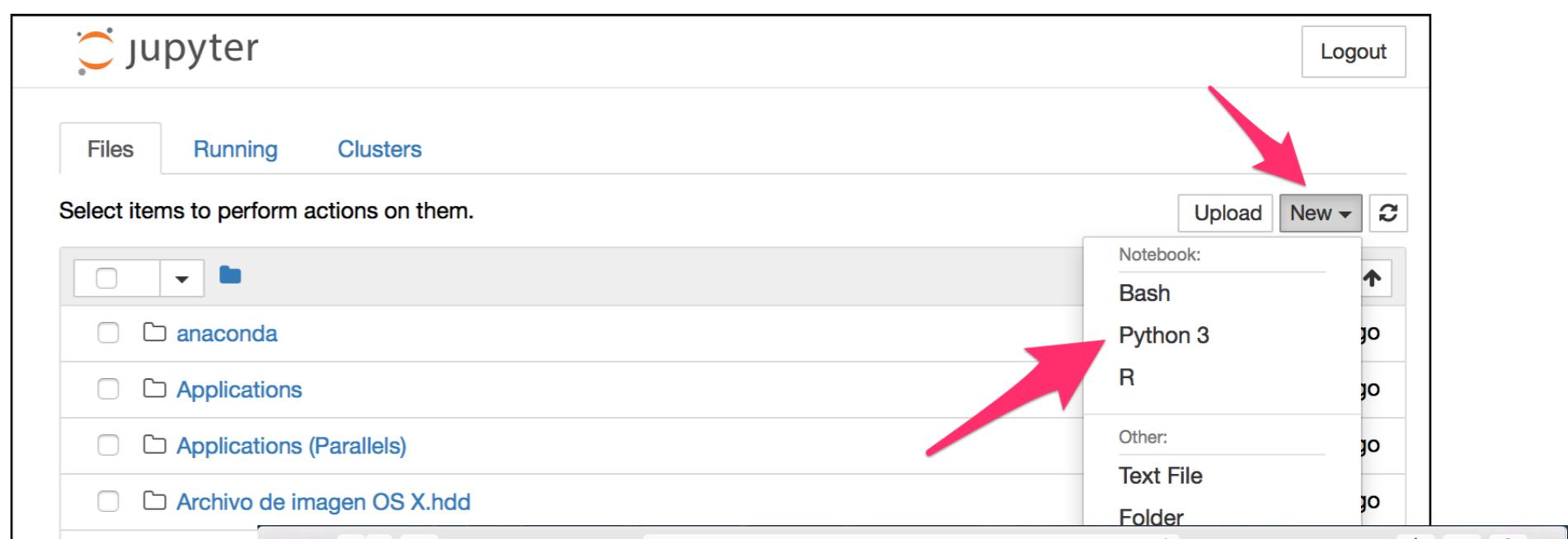


```
Jupyter QtConsole 4.3.0
Python 3.6.1 |Anaconda 4.4.0 (x86_64)| (default, May 11 2017, 13:04:09)
Type "copyright", "credits" or "license" for more information.

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
```



```
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,:]

]: [
```

# Open Data Science & Modern Analytics (2009)

KNIME

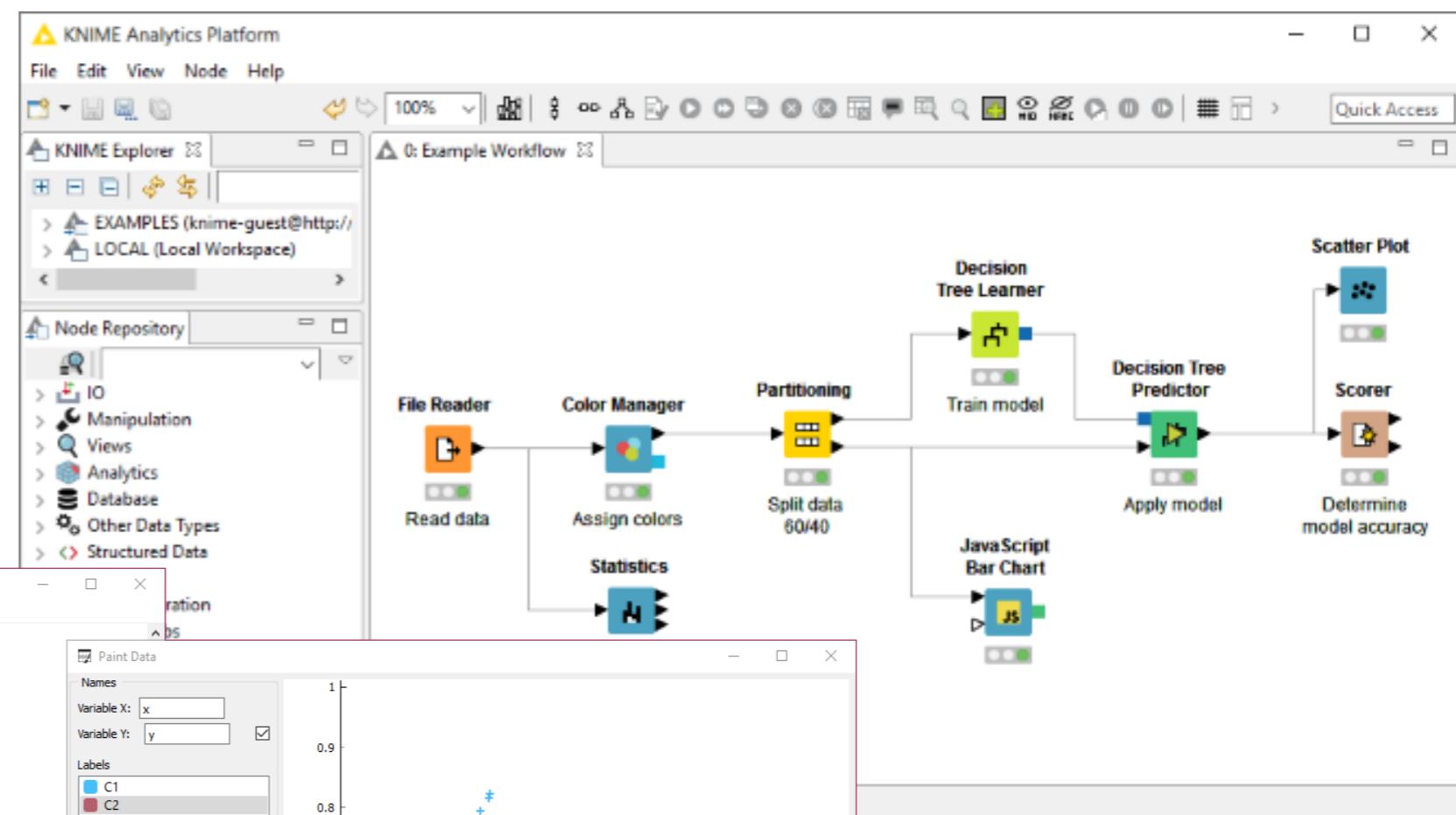
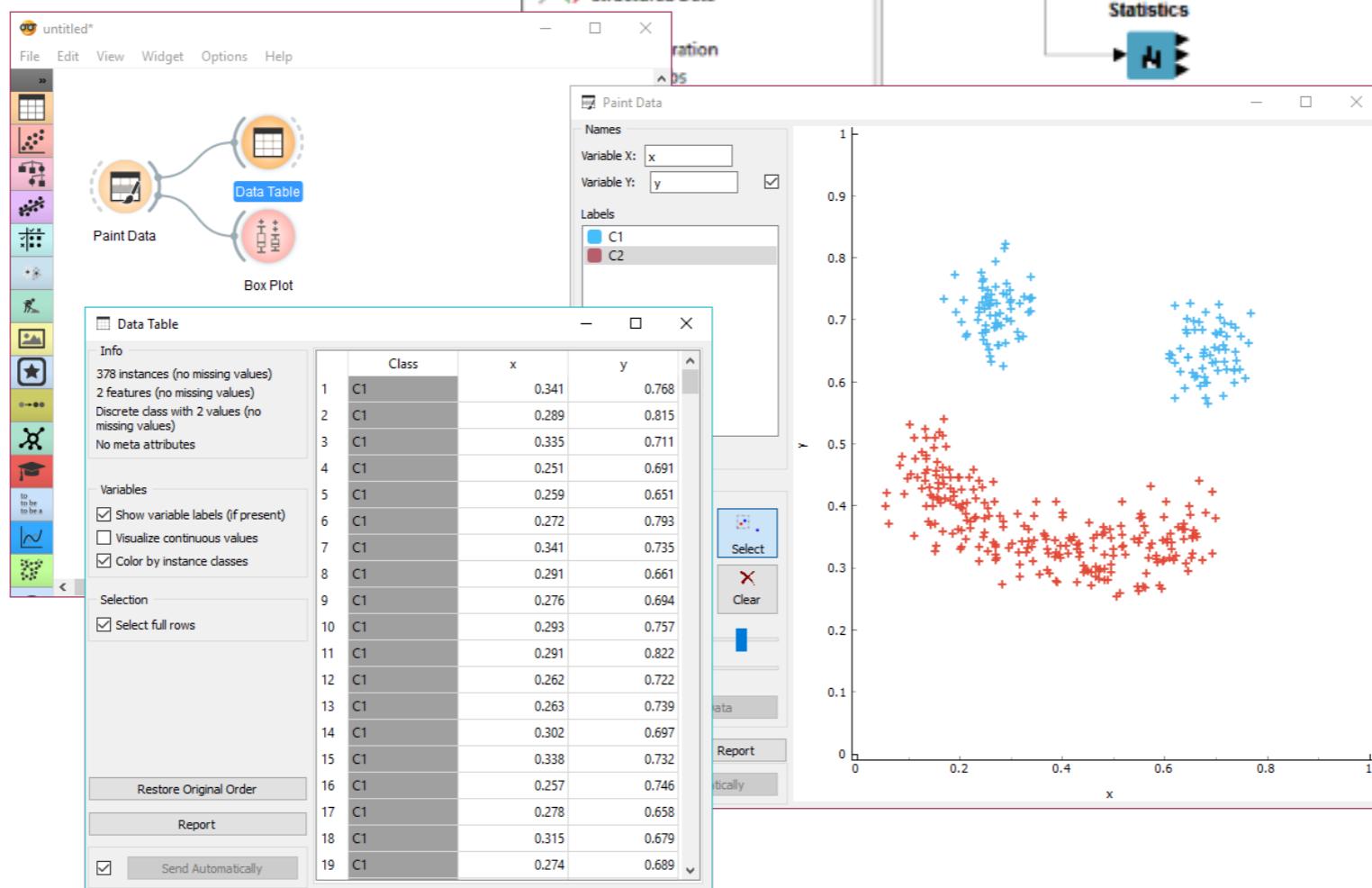
## Storyboards

- Notebooks
- Exploración interactiva
- Programación visual
- Data IDE

## Aplicaciones modernas

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

## Orange



# Open Data Science & Modern Analytics (2009)

## Aplicaciones modernas

- Notebooks
- Dashboards
- Aplicaciones visuales
- Servicios de datos

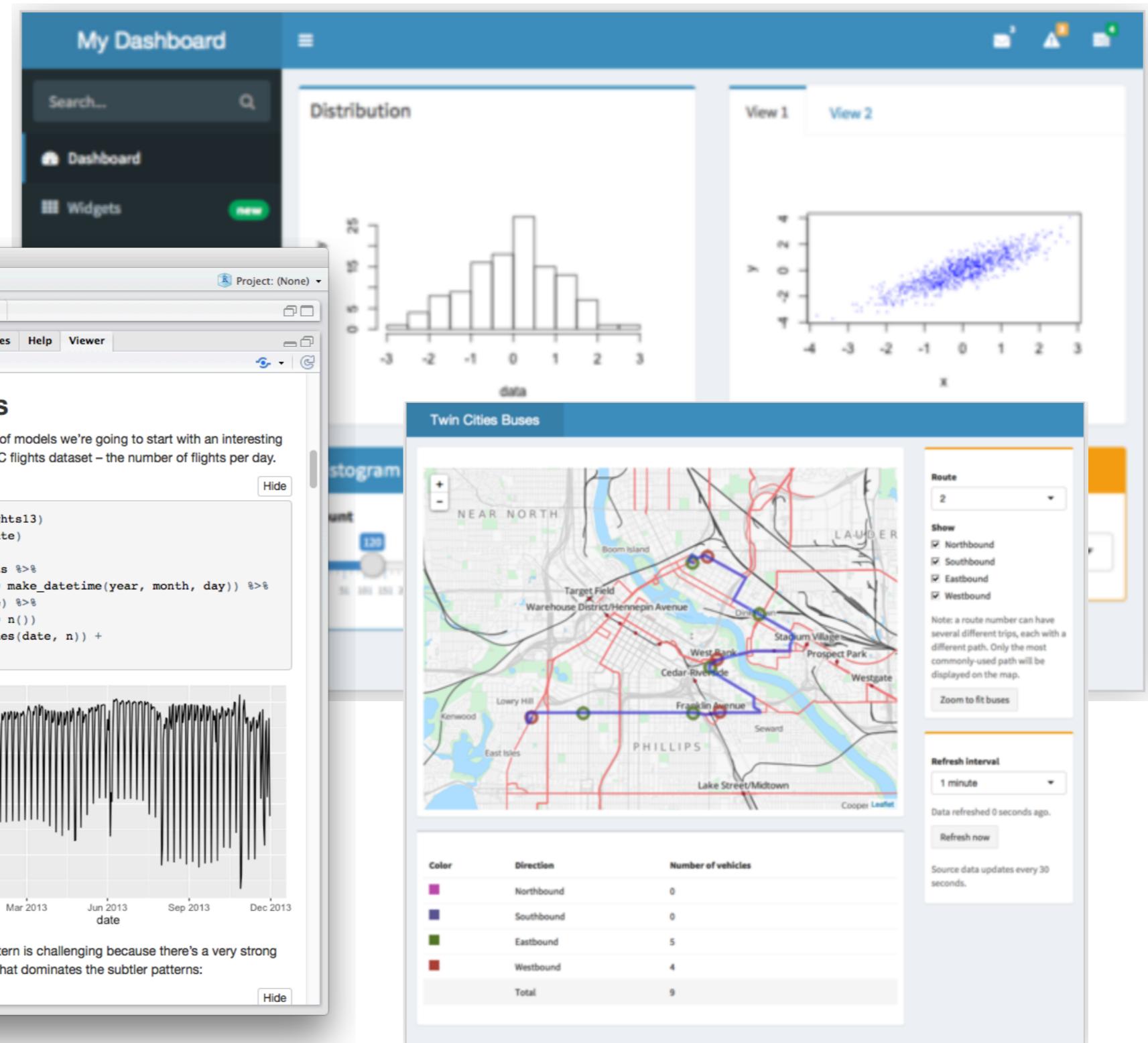
## R Dashboards

## R Markdown

The screenshot shows the RStudio interface. On the left, the code editor displays R Markdown code for generating a plot of flight residuals. The code includes imports for `nycflights13`, `lubridate`, and `dplyr`, and uses `ggplot2` to create a line plot of daily flight counts over time. Two line plots are shown side-by-side, illustrating the seasonal pattern of flights.

```
53
54 ## Residuals
55
56 To motivate the use of models we're going to start with an
57 interesting pattern from the NYC flights dataset -- the
58 number of flights per day.
59
60
61 ````{r}
62 library(nycflights13)
63 library(lubridate)
64 library(dplyr)
65
66 daily <- flights %>%
67   mutate(date = make_datetime(year, month, day)) %>%
68   group_by(date) %>%
69   summarise(n = n())
70
71 ggplot(daily, aes(date, n)) +
72   geom_line()
73 ````
```

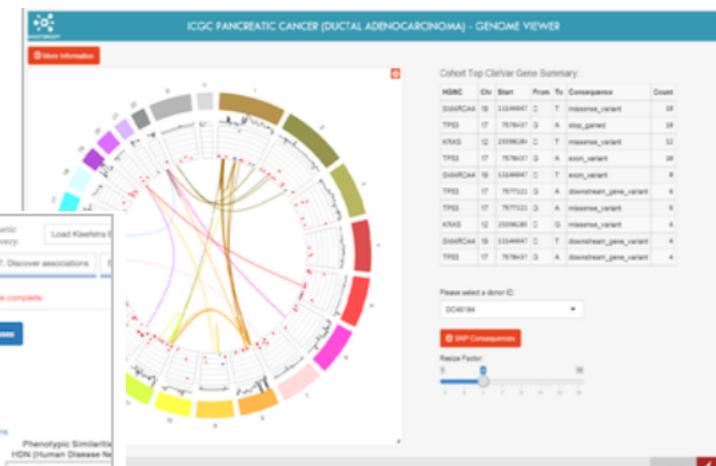
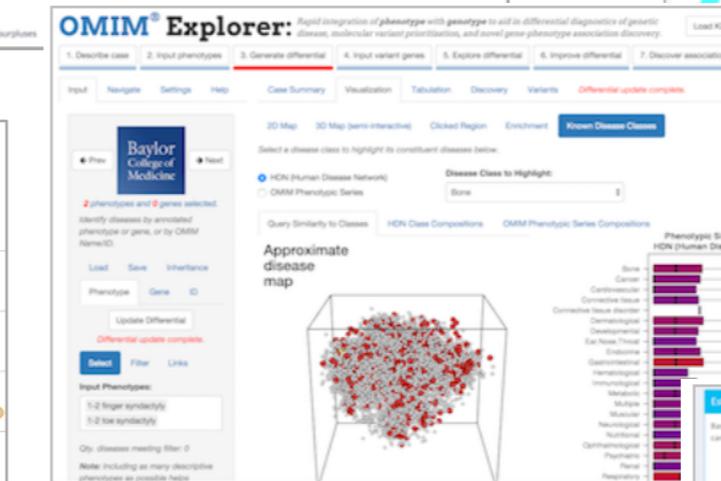
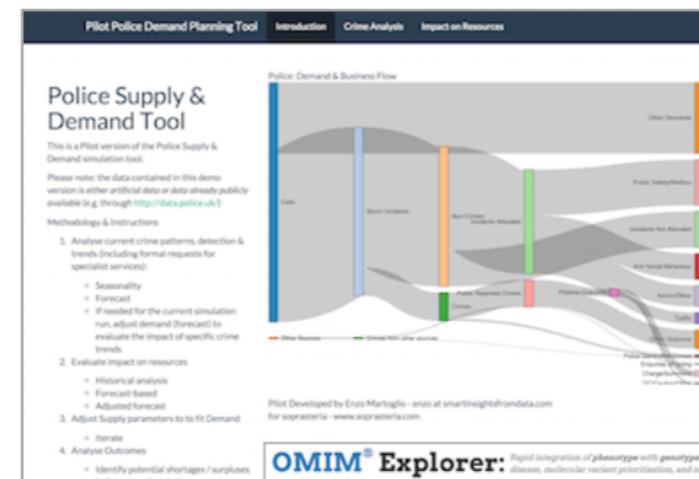
Understand this pattern is challenging because there's a very strong day-of-week effect that dominates the subtler patterns:



# Open Data Science & Modern Analytics (2009)

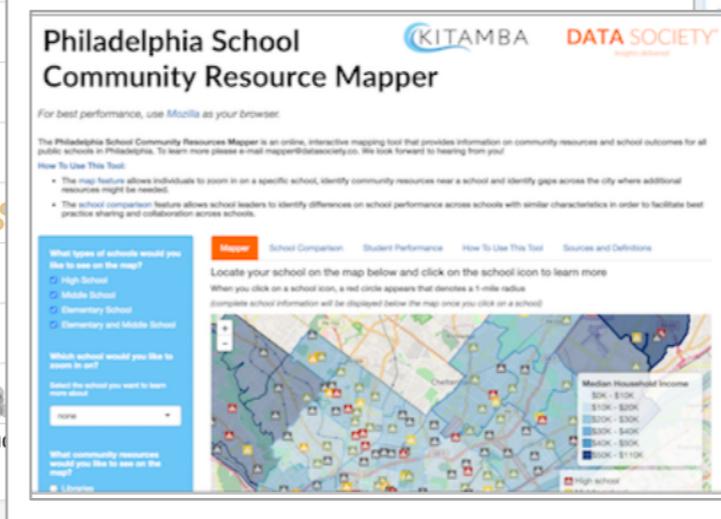
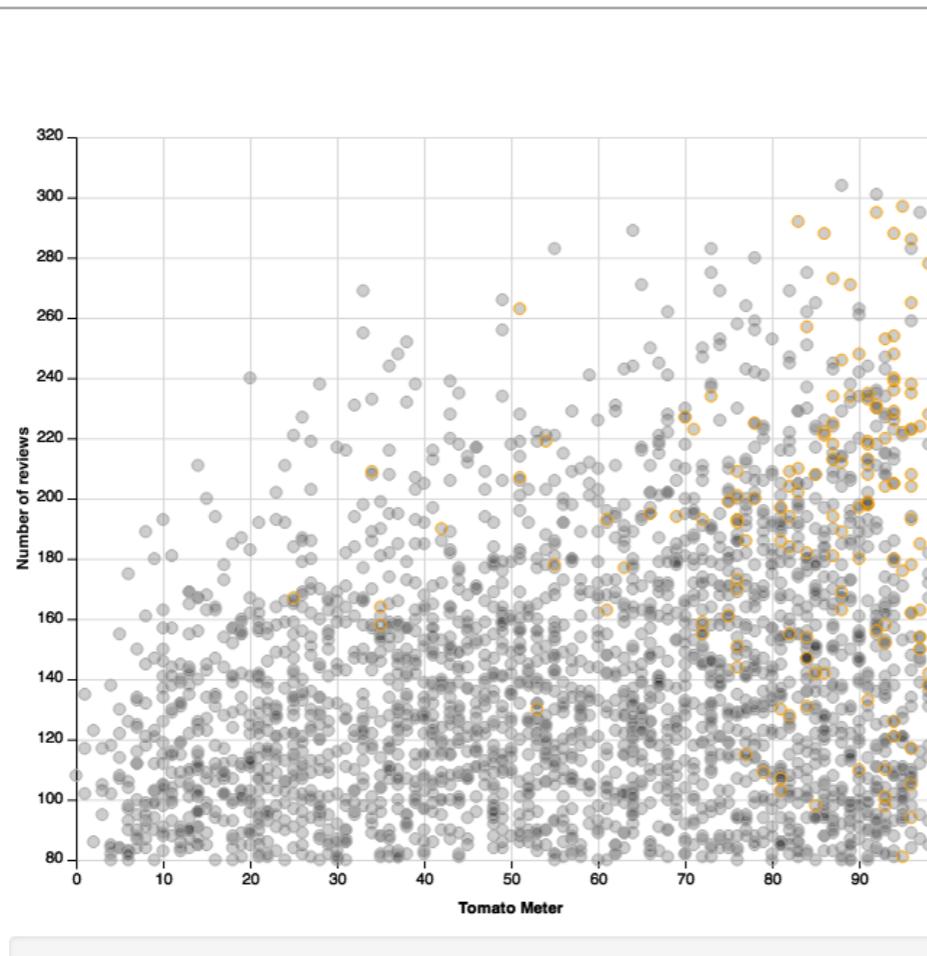
# Aplicaciones modernas

- Notebooks
  - Dashboards
  - Aplicaciones visuales
  - Servicios de datos



# R Shiny

## Movie explorer



# Open Data Science & Modern Analytics (2009)

## 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

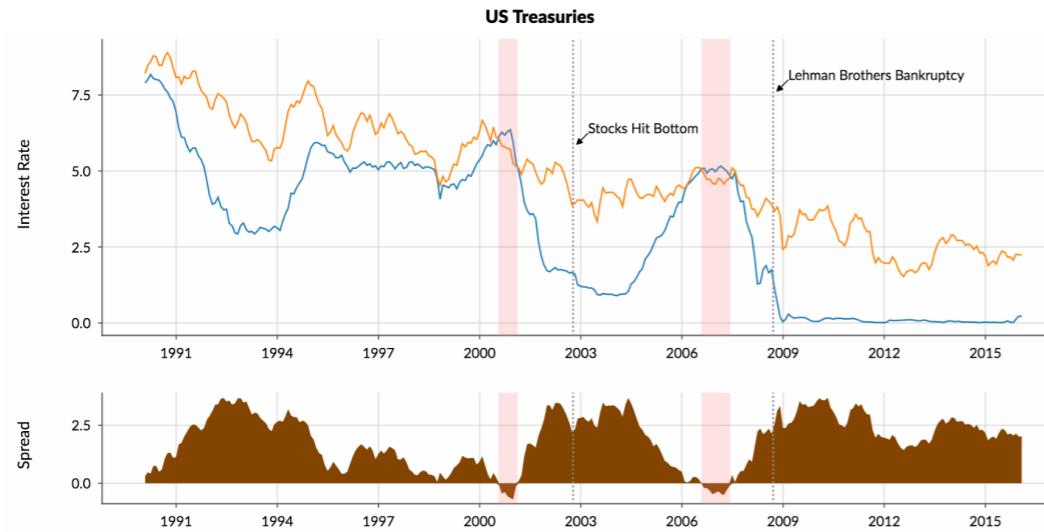
BeakerX is a collection of JVM kernels with widgets, plotting, tables, autotranslation, and other extensions to the Jupyter Notebook and changes with

The document

BeakerX is th  
are hiring.

#### Groovy with Interactive Plotting and Tables:

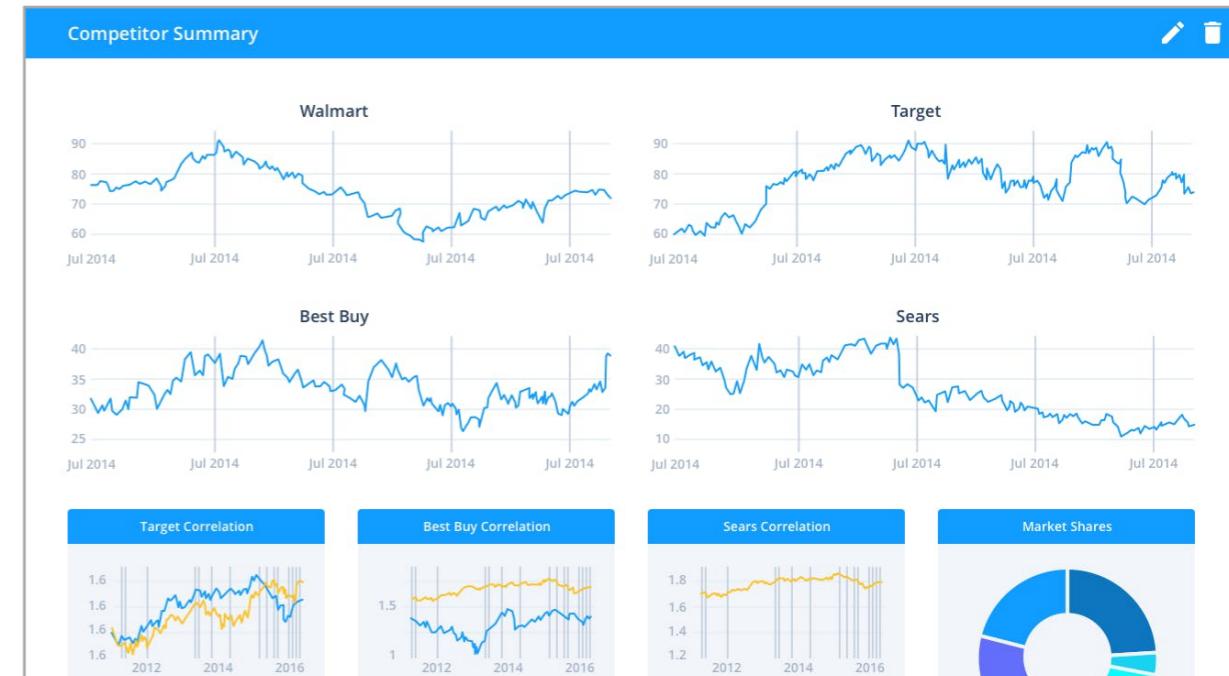
```
// Then use a CombinedPlot to get stacked plots with linked X axis.  
def c = new CombinedPlot(title: "US Treasuries", initWidth: 1000)  
  
// add both plots to the combined plot, and including their relative heights.  
c.add(p1, 3)  
c.add(p2, 1)
```



## Bokeh



## plot.ly



# Open Data Science & Modern Analytics (2009)

## 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

Theano

TensorFlow™

Install Develop API r1.4

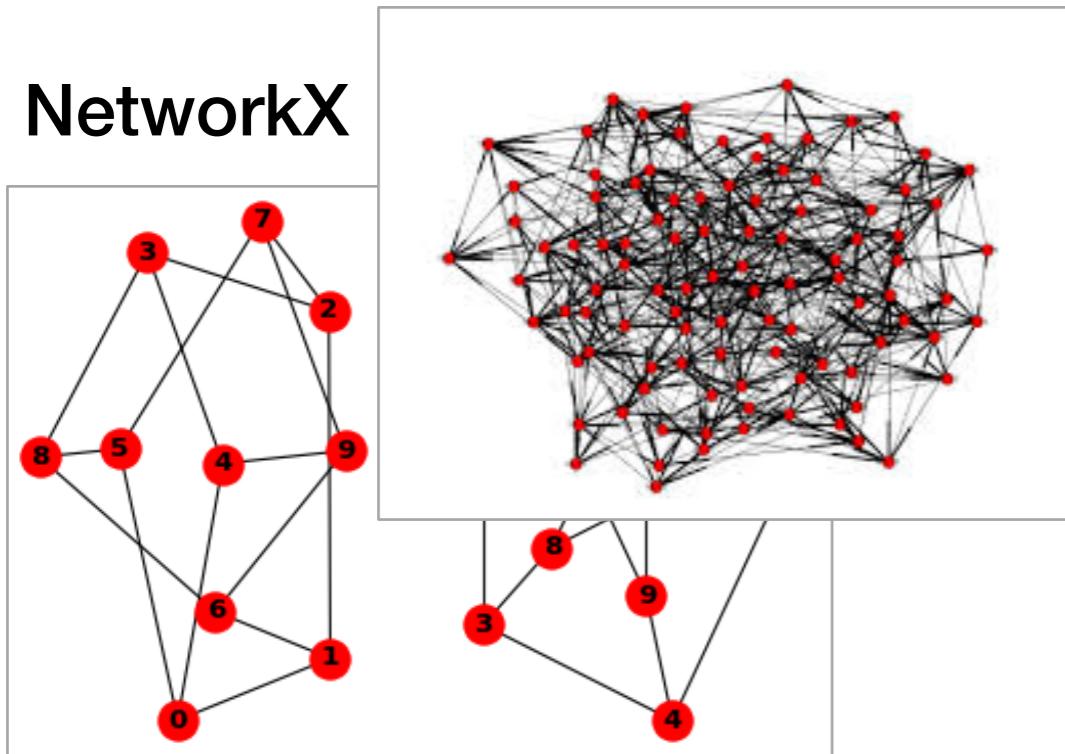
An open-source software library for Machine Intelligence

GET STARTED

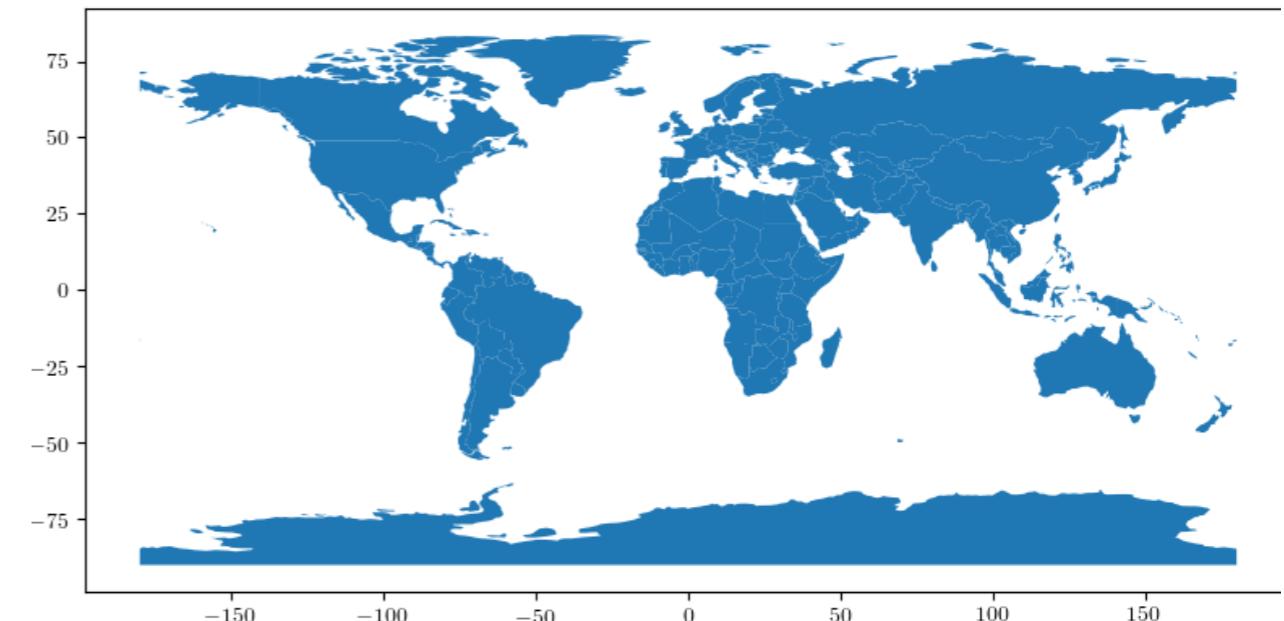
0.9 release ▾

ch docs

## NetworkX



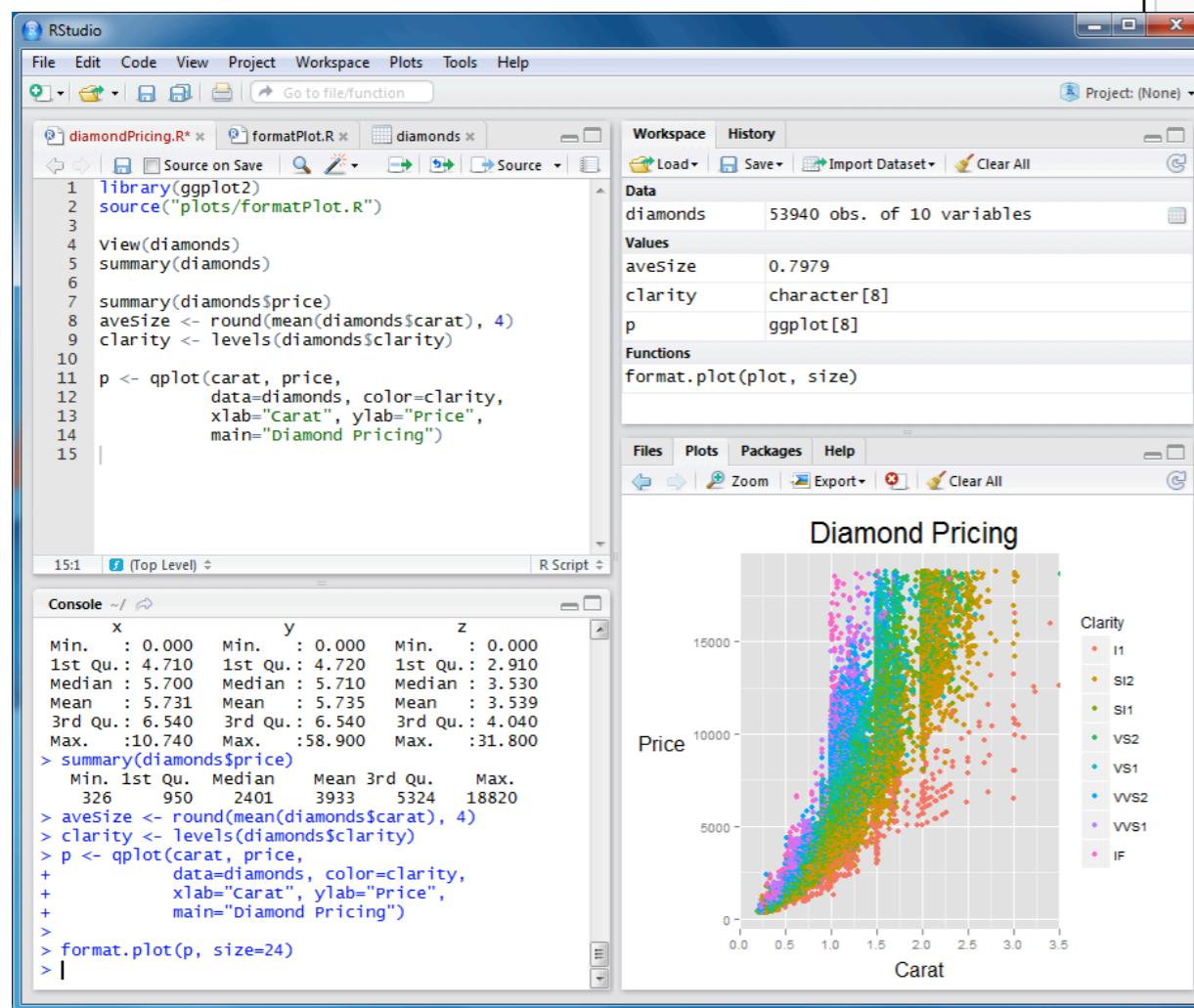
## GeoPandas



# Open Data Science & Modern Analytics (2009)

## 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



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

RStudio

## Python StatsModels

# Open Data Science & Modern Analytics (2009)

## Fuentes de datos

### modernas

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

## Blaze / Odo

Sponsored by:  
**CONTINUUM<sup>®</sup>**  
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

## 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 & Modern Analytics (2009)

## 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 & Modern Analytics (2009)

## 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 & Modern Analytics (2009)

```
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,2005-04-16,2005,4,16,13,135,4.8666666666666667
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,2005-04-16,2005,4,16,17,135,4.8333333333333333
AQUITANIA,2005-04-16,2005,4,16,18,135,4.7666666666666667
AQUITANIA,2005-04-16,2005,4,16,19,135,4.3500000000000005
AQUITANIA,2005-04-16,2005,4,16,20,135,2.6833333333333333
AQUITANIA,2005-04-16,2005,4,16,21,135,3.1999999999999997

# Modern Analytics (2009)

## 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](#))

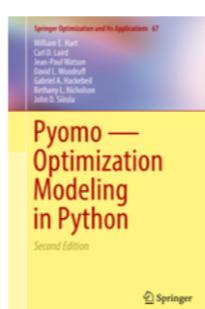
### Examples

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

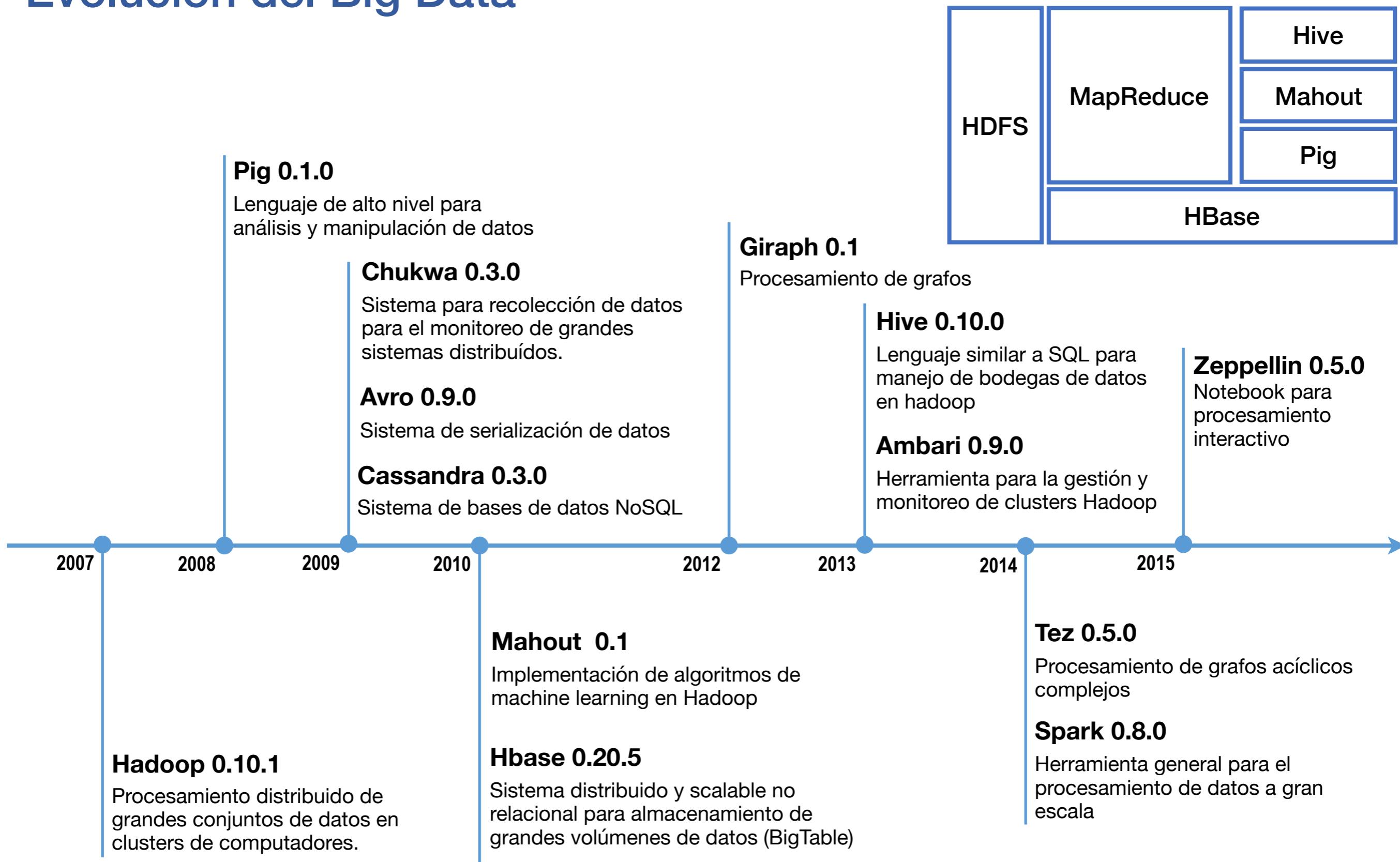


Hart, William E., Carl D. Laird, Jean-Paul Watson, David L. Woodruff, Gabriel A. Hackebeil, Bethany L. Nicholson, and John D. Siirola. *Pyomo – Optimization Modeling in Python*. Second Edition. Vol. 67. Springer, 2017.

*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.*

# Big Data Analytics

# Evolución del Big Data



# Pig Latin (2008)

```
CROSS  
EXPLAIN  
FILTER  
FOREACH  
GENERATE  
GROUP  
ILLUSTRATE  
JOIN  
LIMIT  
LOAD  
ORDER  
STREAM  
SPLIT  
STORE  
SET  
QUIT
```

Lenguaje similar al SQL para el análisis de grandes volúmenes de datos en Hadoop representados como flujos de datos.

## 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;
```

# Apache Hive (2013)

## Ejemplo de Hive

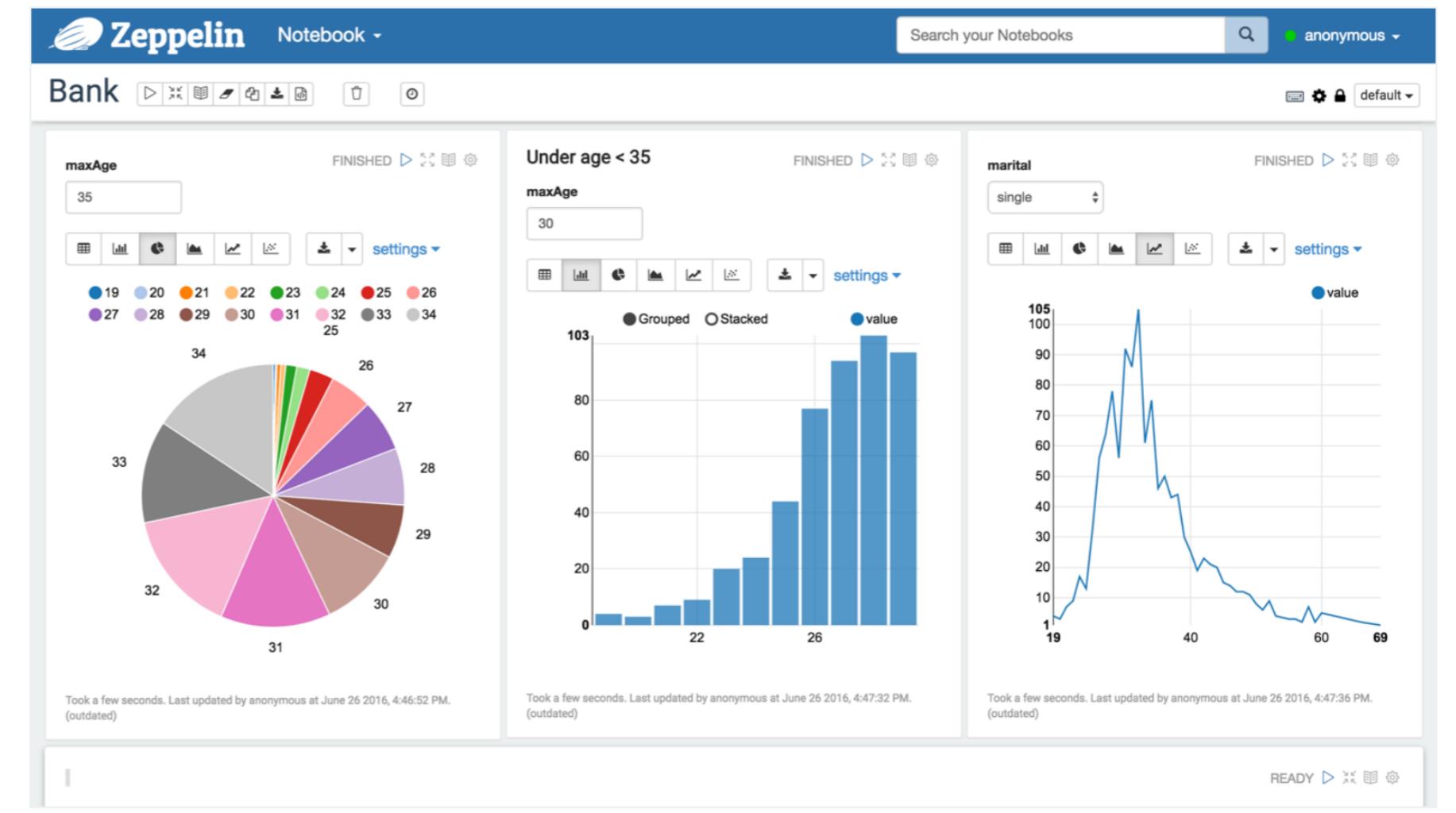
```
CREATE TABLE records (year STRING, temperature INT, quality INT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';
LOAD DATA LOCAL INPATH 'sample.txt' OVERWRITE INTO TABLE records;
SELECT year, MAX(temperature) FROM records GROUP BY year;
```

# Apache Zeppelin (2015)

# Multi-purpose Notebook

The Notebook is the place for all your needs

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



# Una Introducción a la Analítica

## Algunos casos de uso de Machine Learning

Esta presentación describe, bajo un marco común, los conceptos fundamentales de Data Science, Analytics y Big Data y establece su similitudes y diferencias; se presentan ejemplos de casos prácticos de la aplicación de Machine Learning y Aprendizaje Estadístico.

Descargue la última versión de este documento de:  
<https://github.com/jdvelasq/data-science-docs/blob/master/sena.pdf>