

# Una Introducción a la Analítica

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/an-intro-to-analytics/>

**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

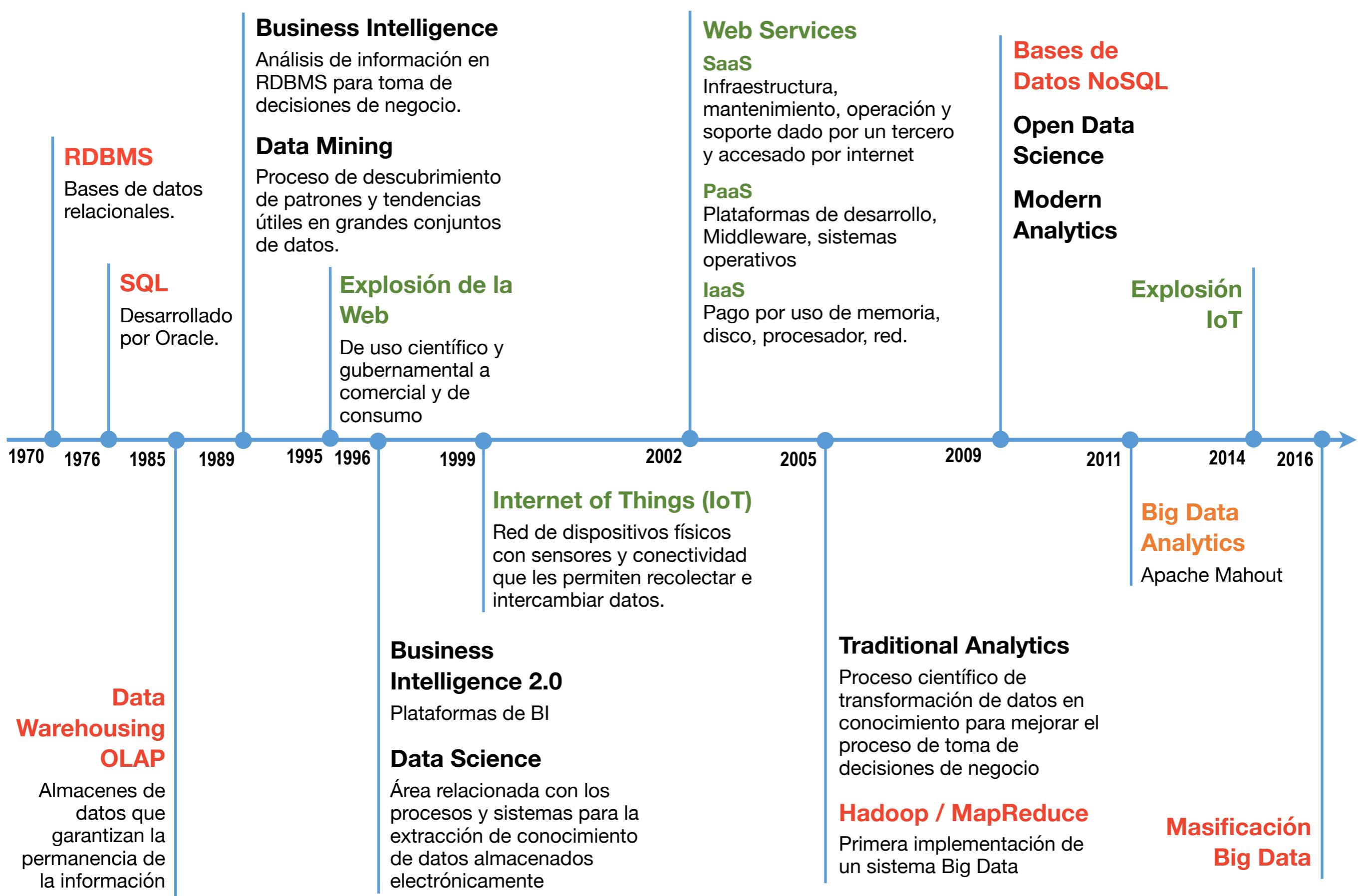
 jdvelasq@unal.edu.co

 @jdvelasquezh

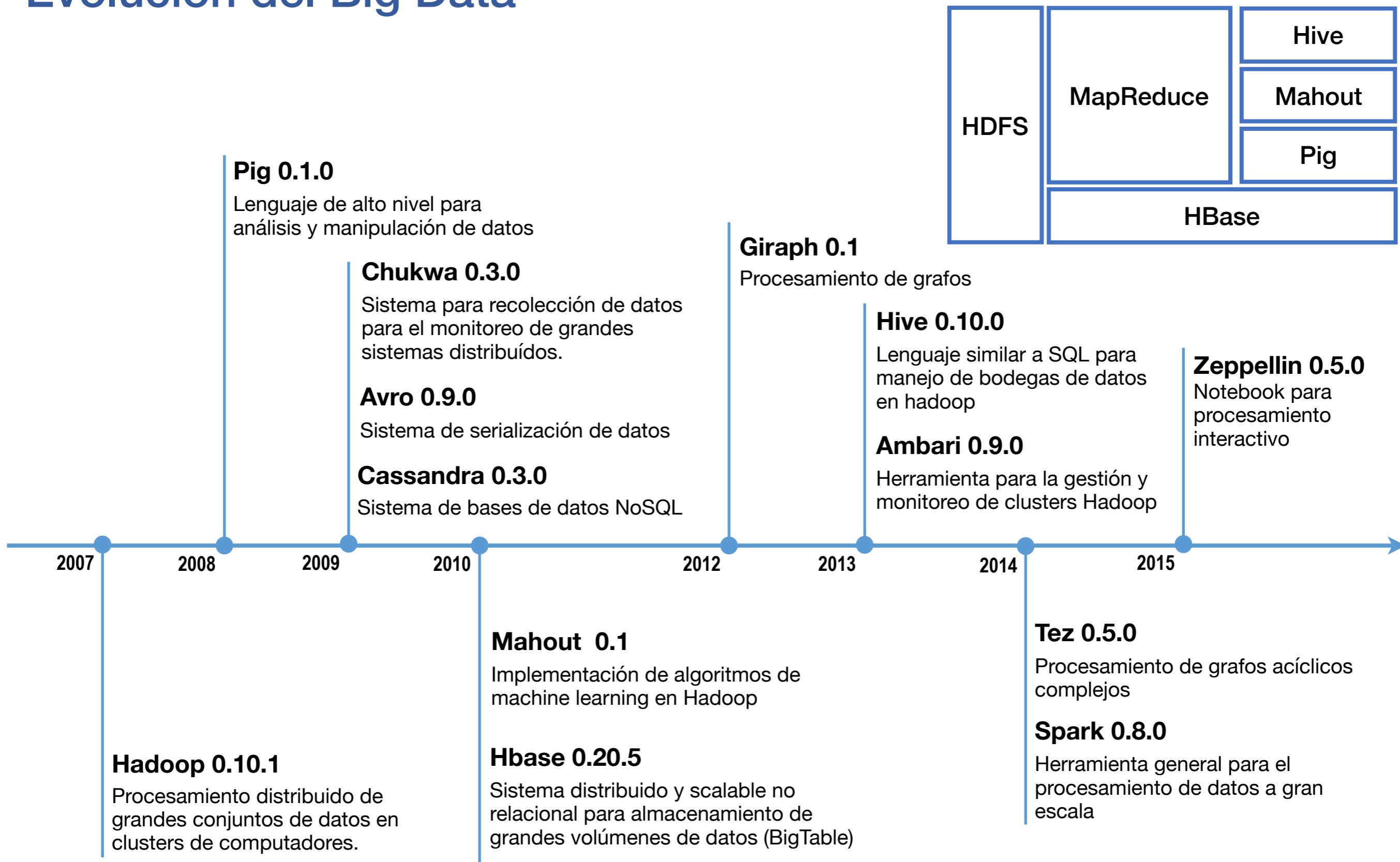
 <https://github.com/jdvelasq>

 <https://goo.gl/prkjAq>

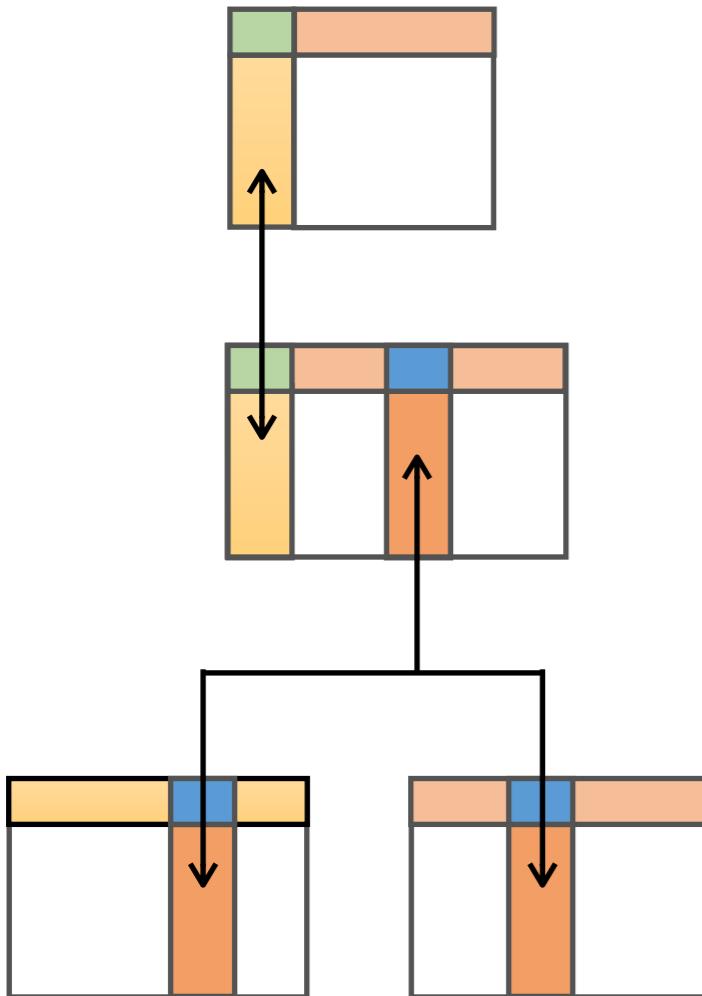
 <https://goo.gl/vXH8jy>



# Evolución del Big Data



# 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
- PostgreSQL
- 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;

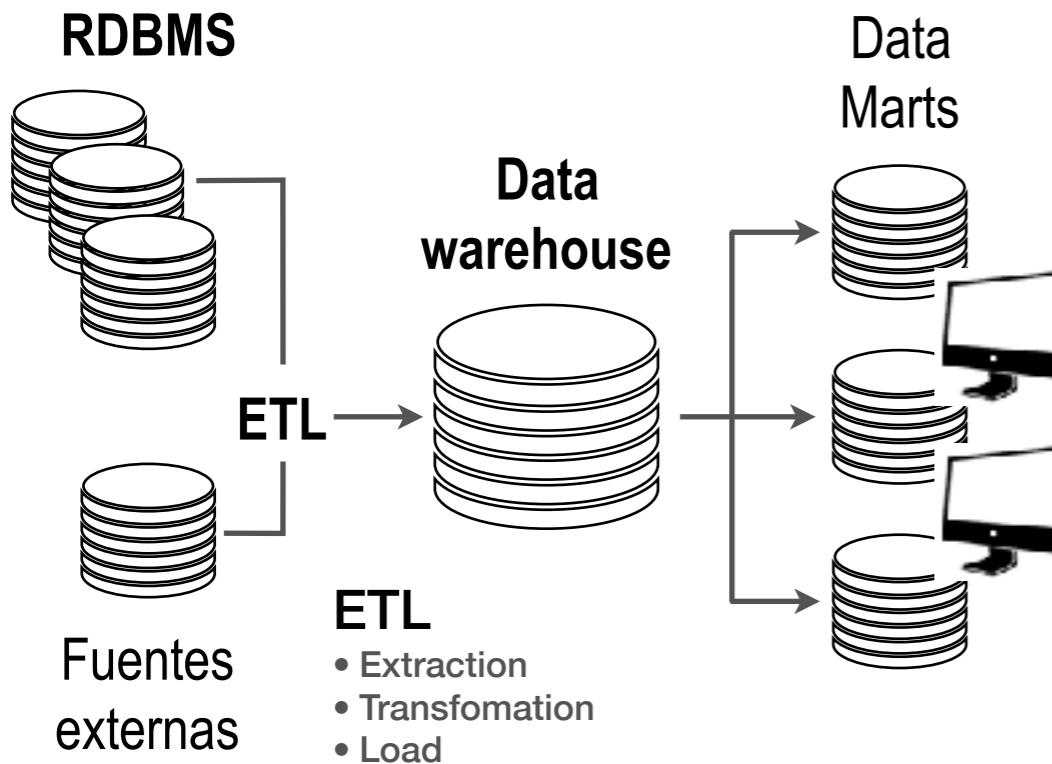
SELECT DISTINCT marca, modelo FROM coches;

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 y OLAP (1985)



## Data Mart

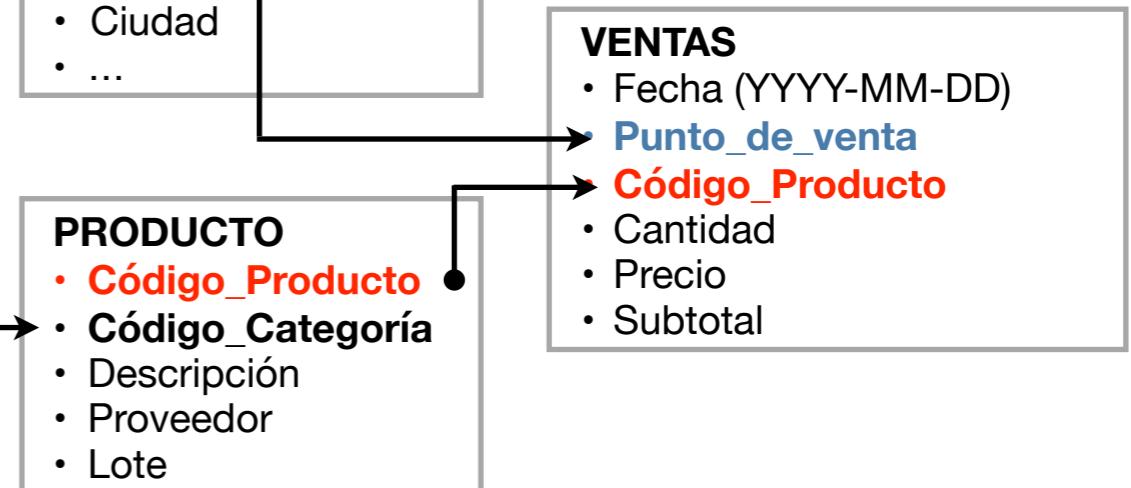
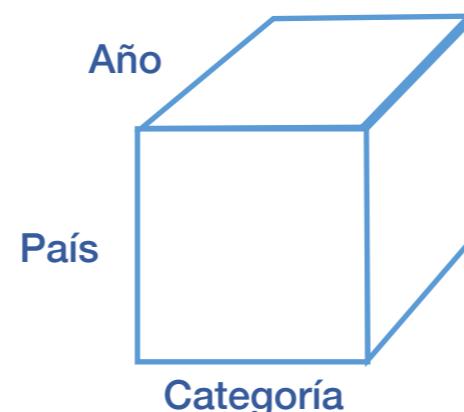
Subconjunto de datos de un Data Warehouse orientado a la consulta. Es implementado usando cubos OLAP

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

## Data Warehouse

### Bodegas de datos

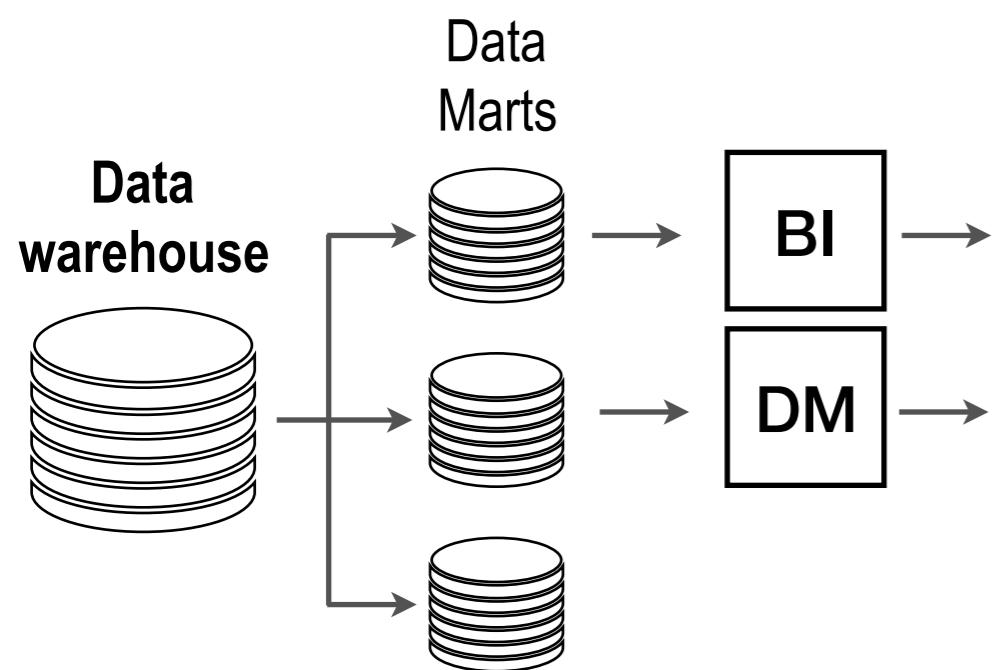
- Estructurado
- Orientado a temas
- Integrado (consistencia de los datos)
- No volátil (permanencia de la información, no se modifica ni se elimina)
- Variable en el tiempo
- Orientado al análisis y la divulgación de la información



## OLAP - On-line analytical processing

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

# Data Mining & Business Intelligence (1989)



## Inteligencia de Negocios

Conjunto de herramientas, productos y tecnologías para la creación y gestión del conocimiento del medio a partir de los datos disponibles en una organización.

Tareas típicas: visualización de datos, cálculo de indicadores, dashboards y reportes automáticos.

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

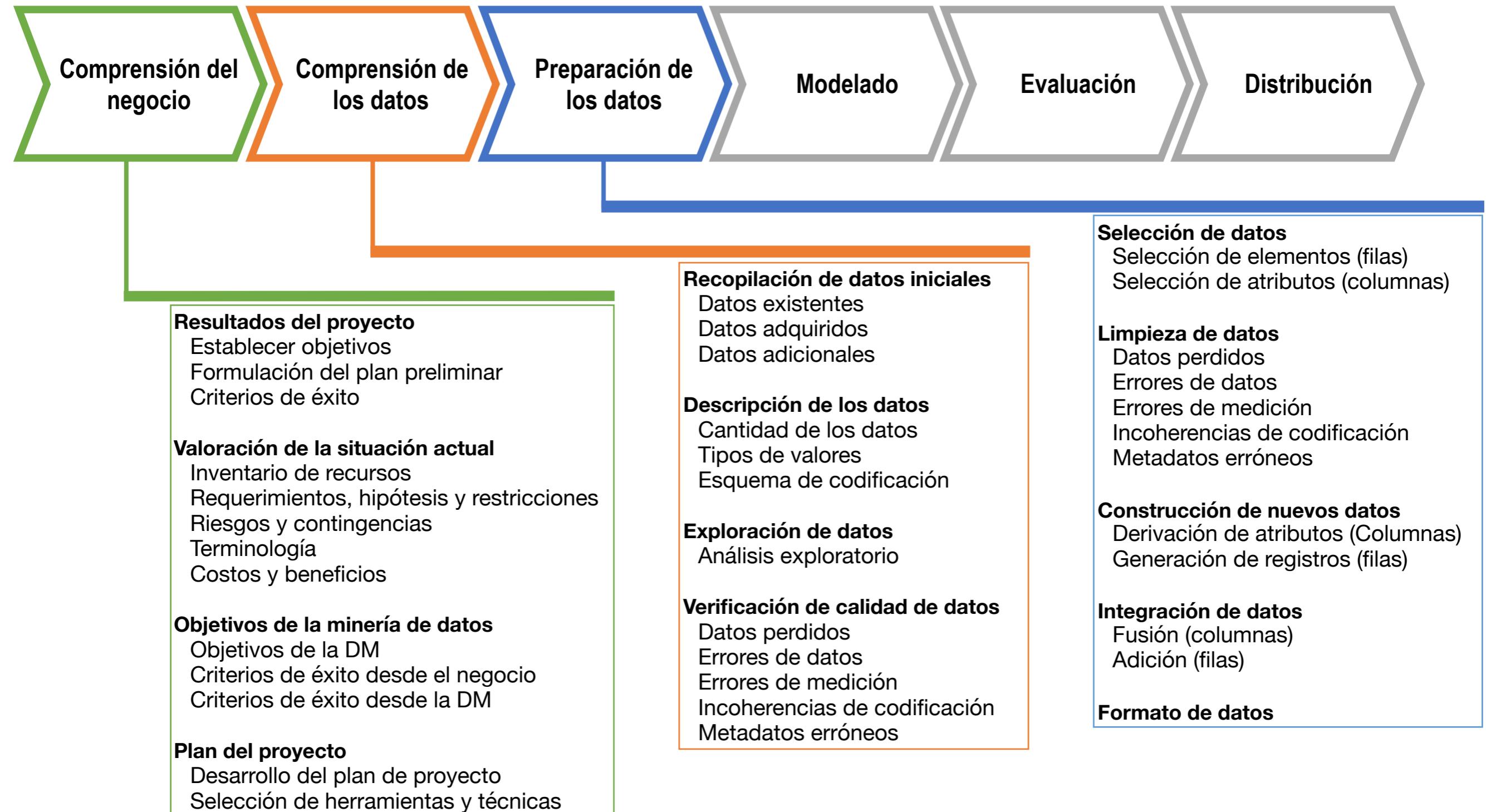
## Data Mining

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.

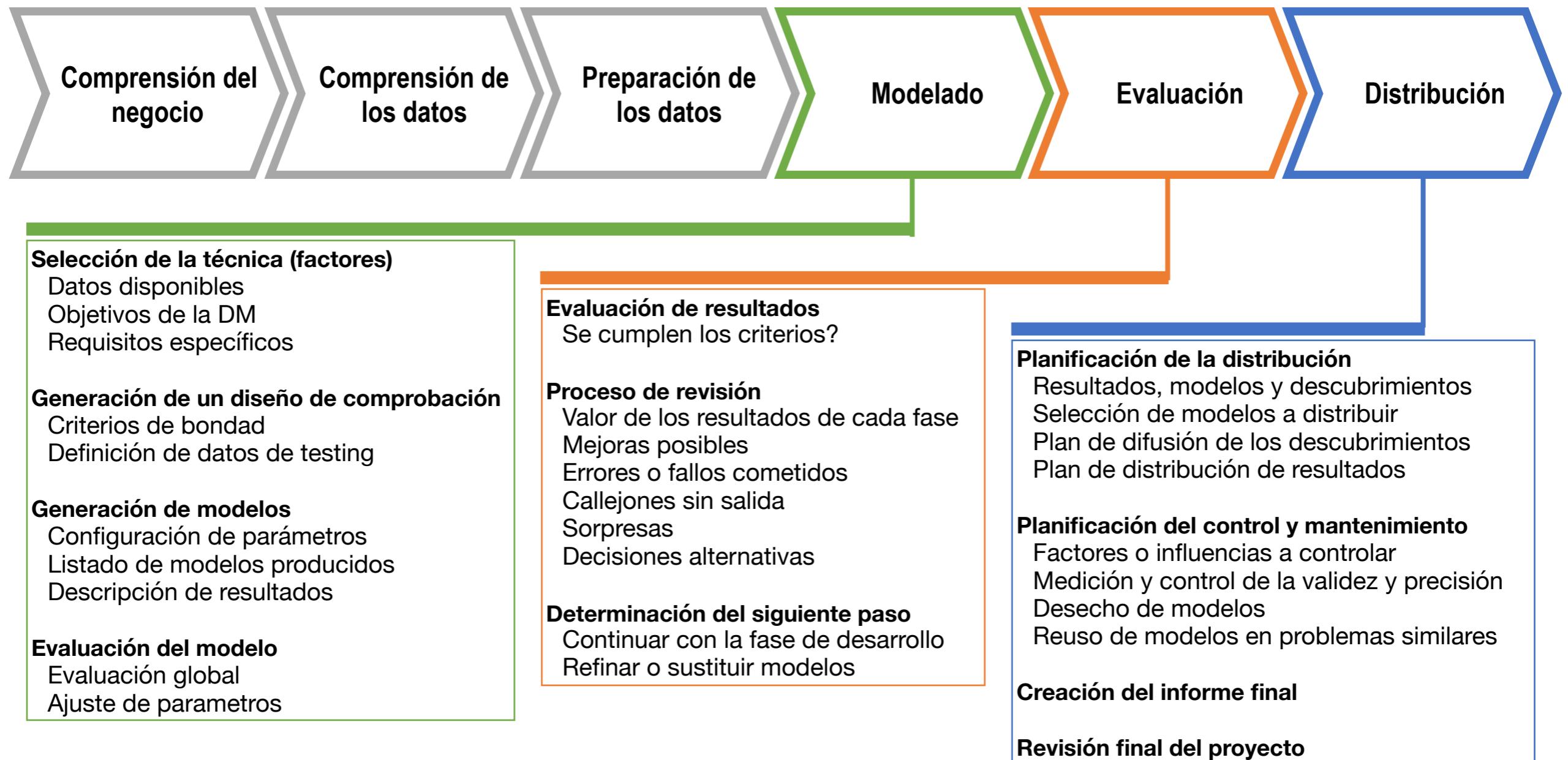
# CRISP-DM (1996)

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

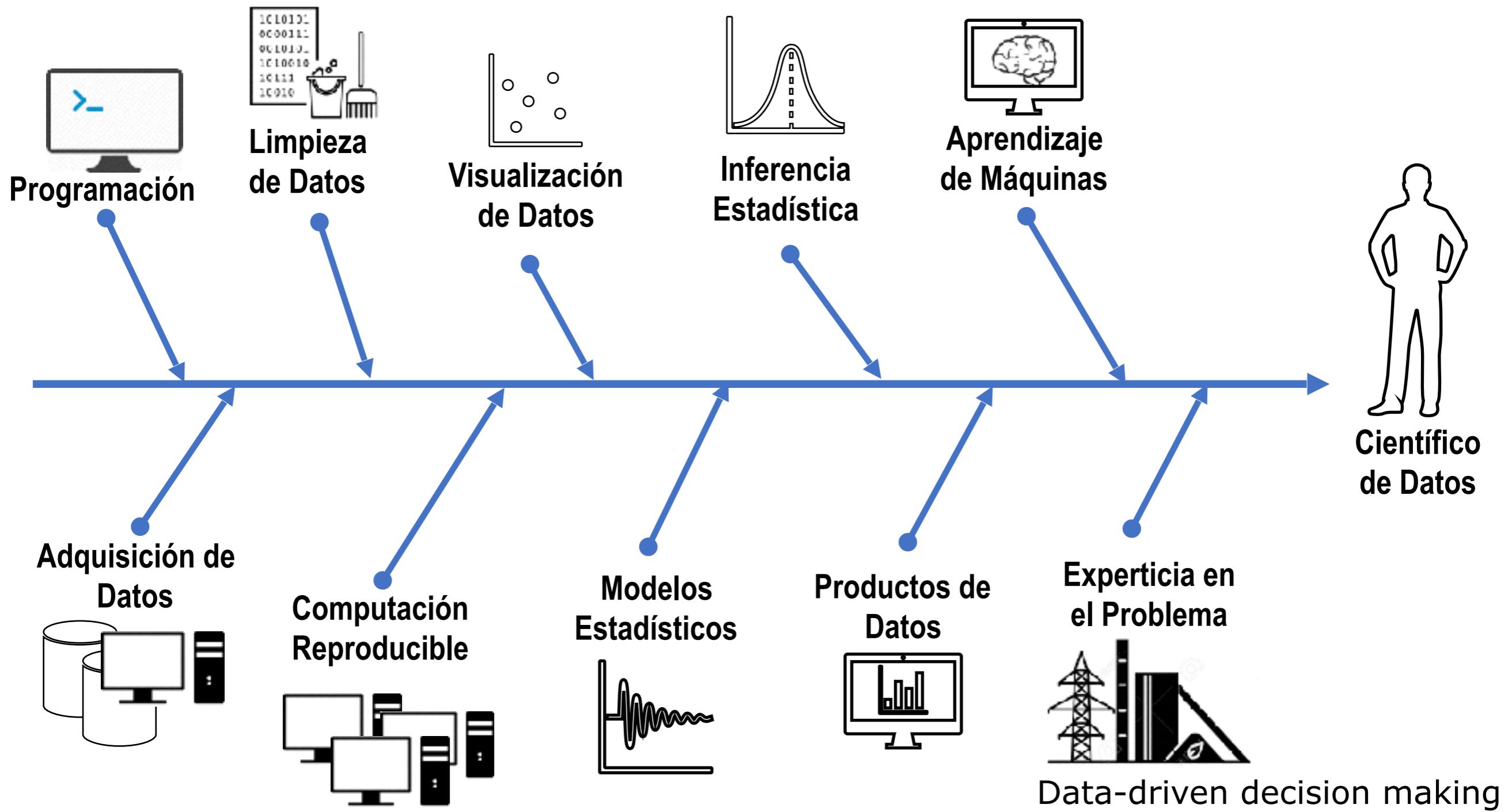
The image displays three distinct Business Intelligence (BI) dashboards:

- información INTELIGENTE**: A dashboard titled "Indicadores" showing various energy metrics. Key data points include:
  - Volumen Útil Diario: 12,810.78 GWh (↑1.14%)
  - Aportes: 198.01 GWh (↓6.36%)
  - Demanda Energía SIN Preliminar: 188.14 GWh (↑1.91%)
  - Generación Hidráulica: 155.15 GWh (↑14.34%)
  - Generación Térmica: 17.01 GWh (↑20.93%)
  - Importaciones Preliminar: 0.09 GWh (↓33.33%)
  - Precio Balsa Promedio: 110.06 \$/kWh
  - Transacciones en contratos: 32,521.85 \$M
  - Restricciones sin Alivios: 2,918.62 \$M
- GENSCAPE™**: A dashboard for energy market monitoring. It includes sections for "Overview", "Daily Macro Supply & Demand Data Report", "Equity Production Insight", "Intrastate Storage Monitoring", and "Natural Gas Analyst". Associated reports include "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".
- energyone**: A dashboard titled "EnergyDashboard" featuring a large map of North America and a grid of operational status boxes. The "FEATURES" section lists capabilities such as market data analysis, portfolio status monitoring, contracted position tracking, and energy operations status alerts. It also mentions the ability to switch between operational functions and market operations/bidding/contacts.

# 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>

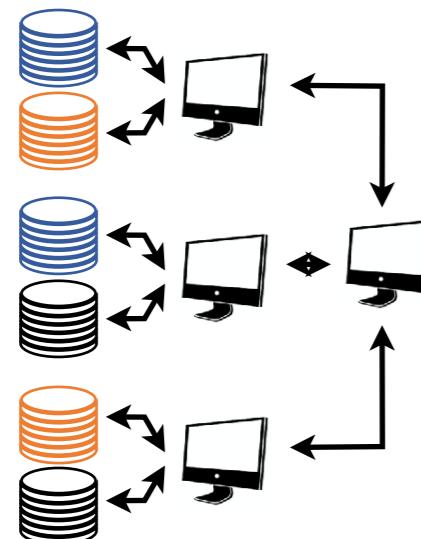
# Data Science (1996)



# 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.

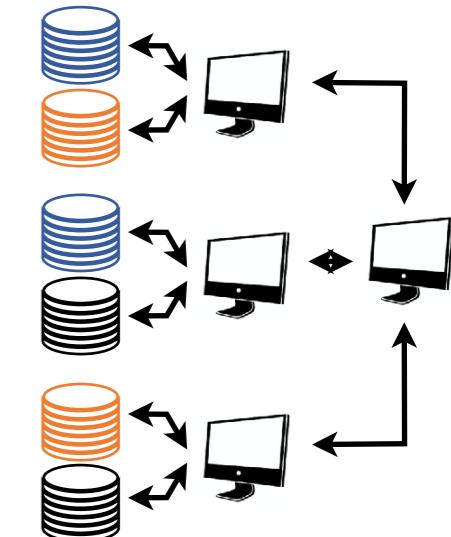
### Platform as a Service (PaaS)

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.



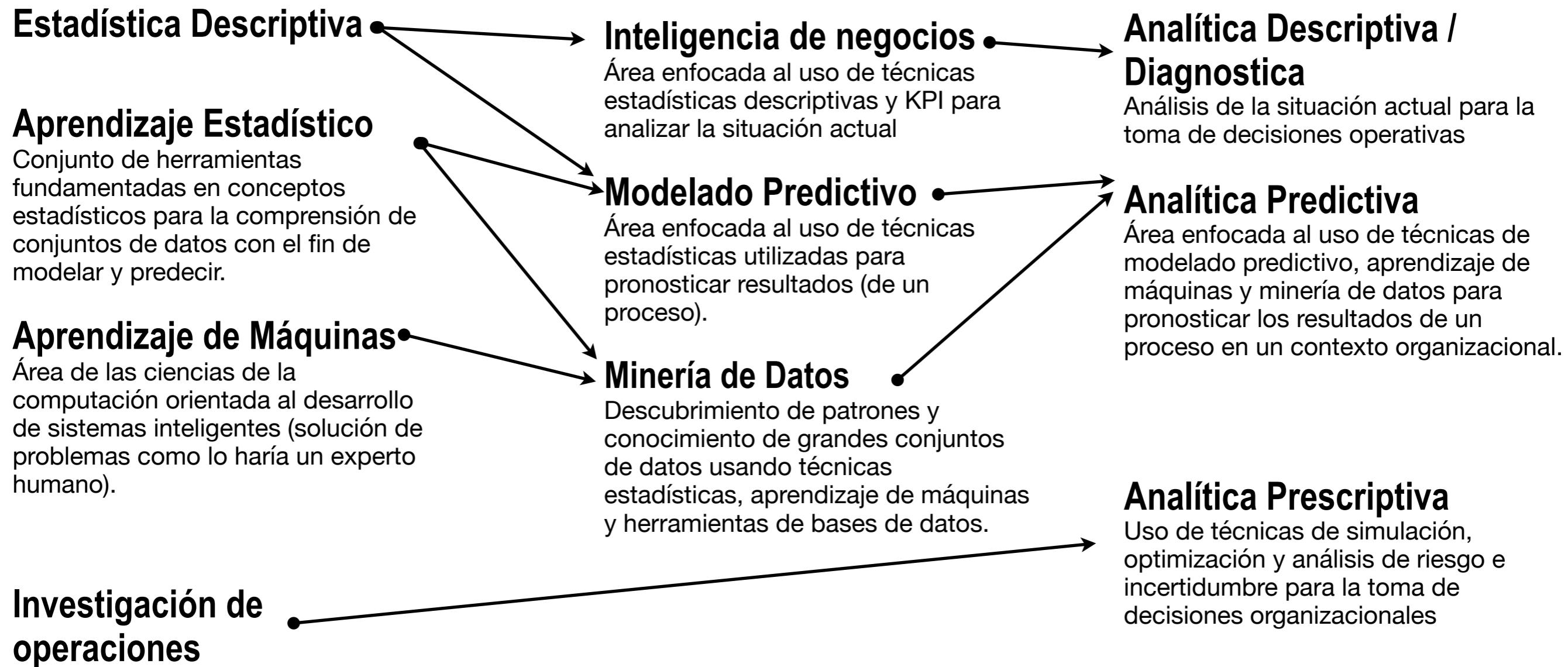
Nube

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.

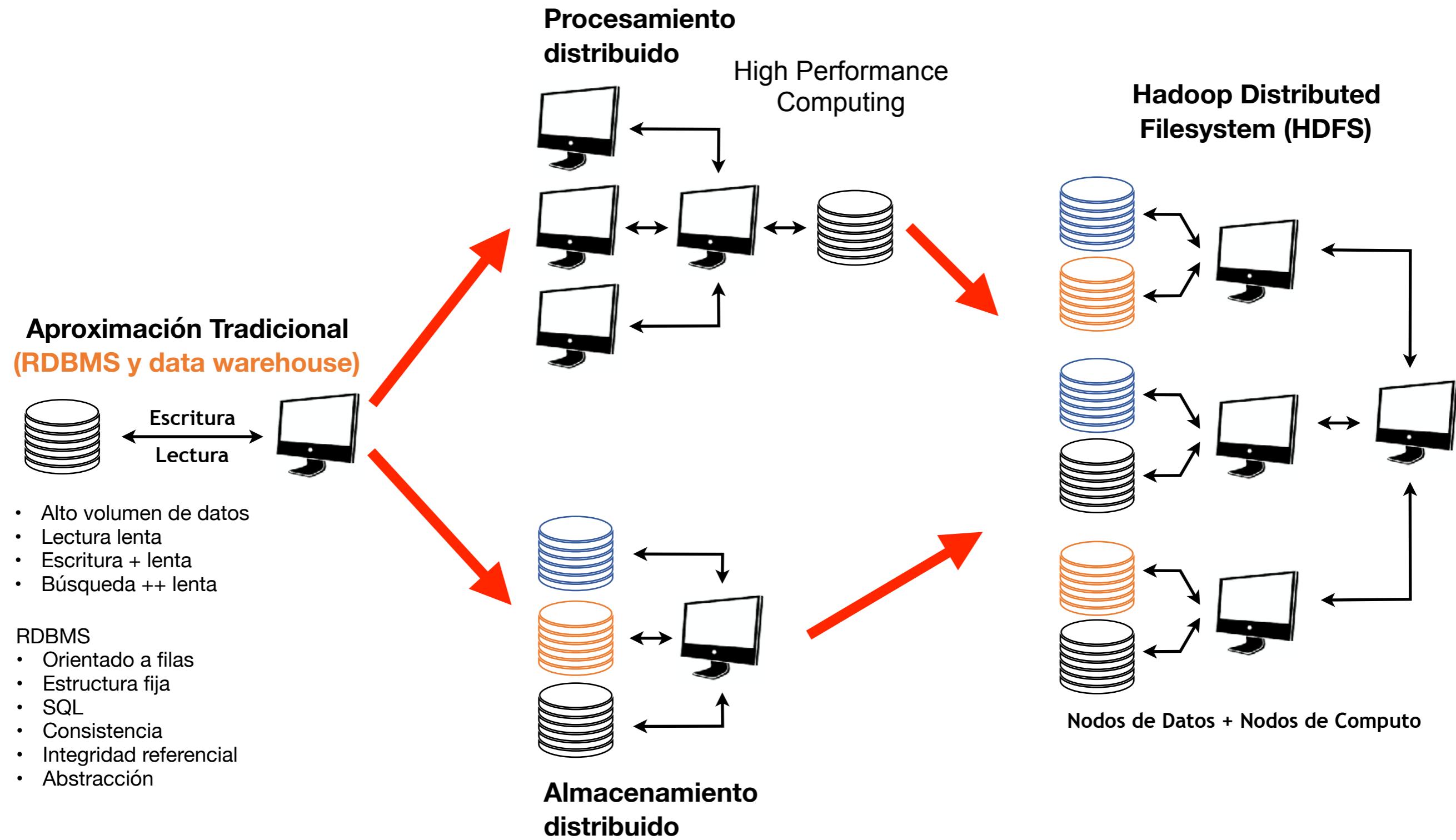
- No se quiere comprender que pasó.
- Pronóstico del futuro.
- Los datos están NO están listos.
- Sólida fundamentación matemática.

## Analítica Predictiva

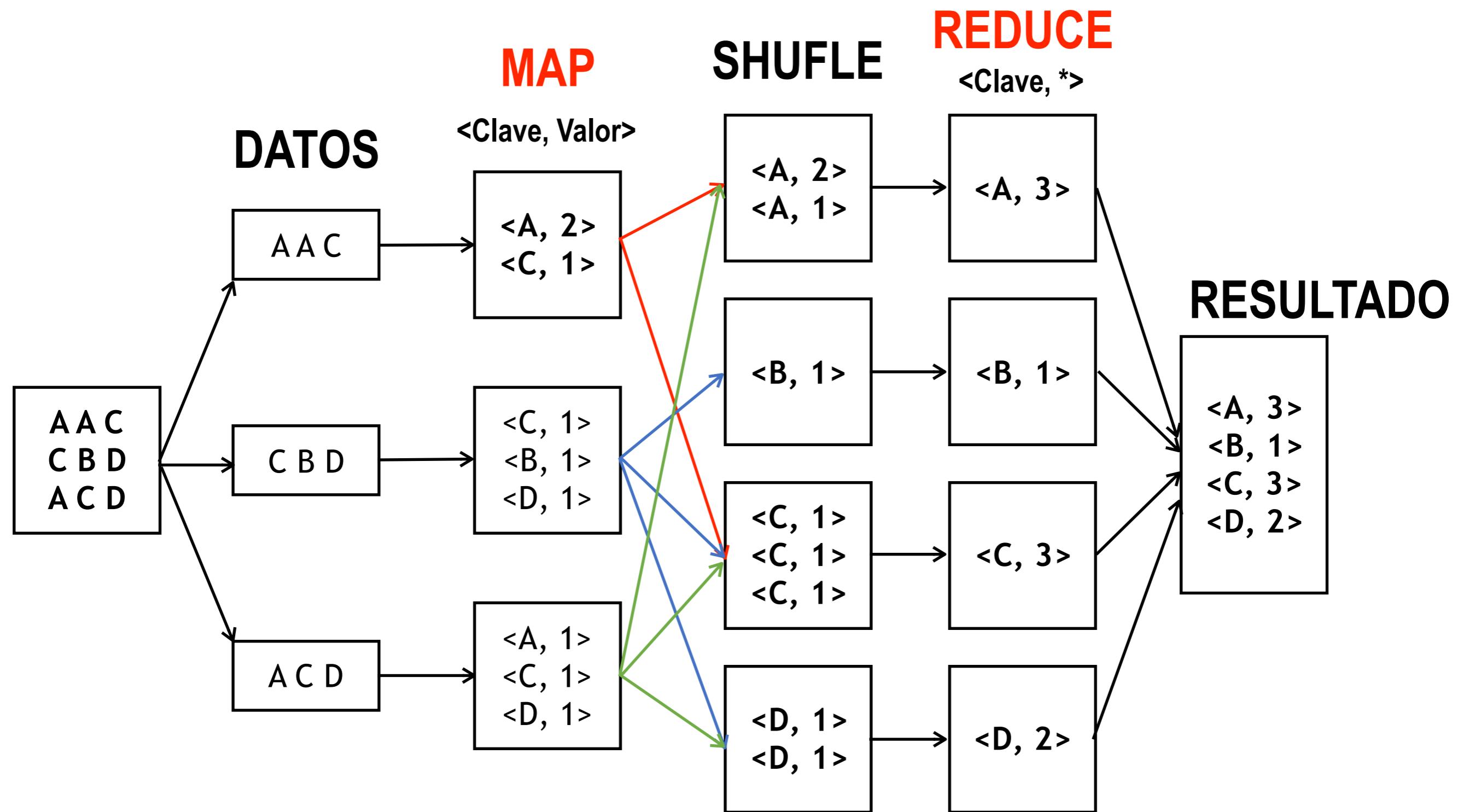
Área enfocada al uso de técnicas de modelado predictivo, aprendizaje de máquinas y minería de datos para pronosticar los resultados de un proceso en un contexto organizacional.

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

# Hadoop / MapReduce (2005)



# Hadoop / MapReduce (2005)



# 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;
```

# 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

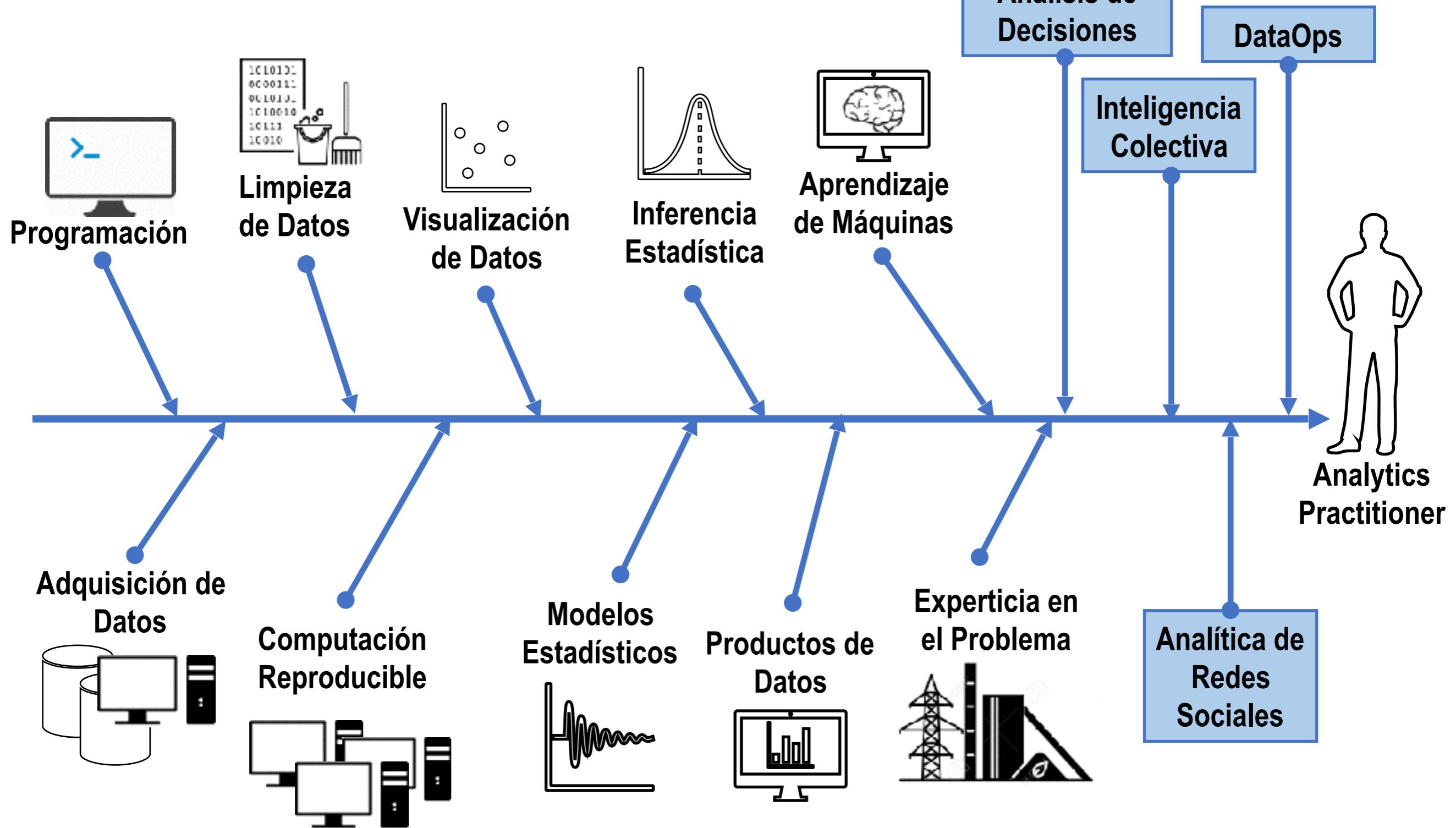
## Sistema orientado a filas

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

## Column family database

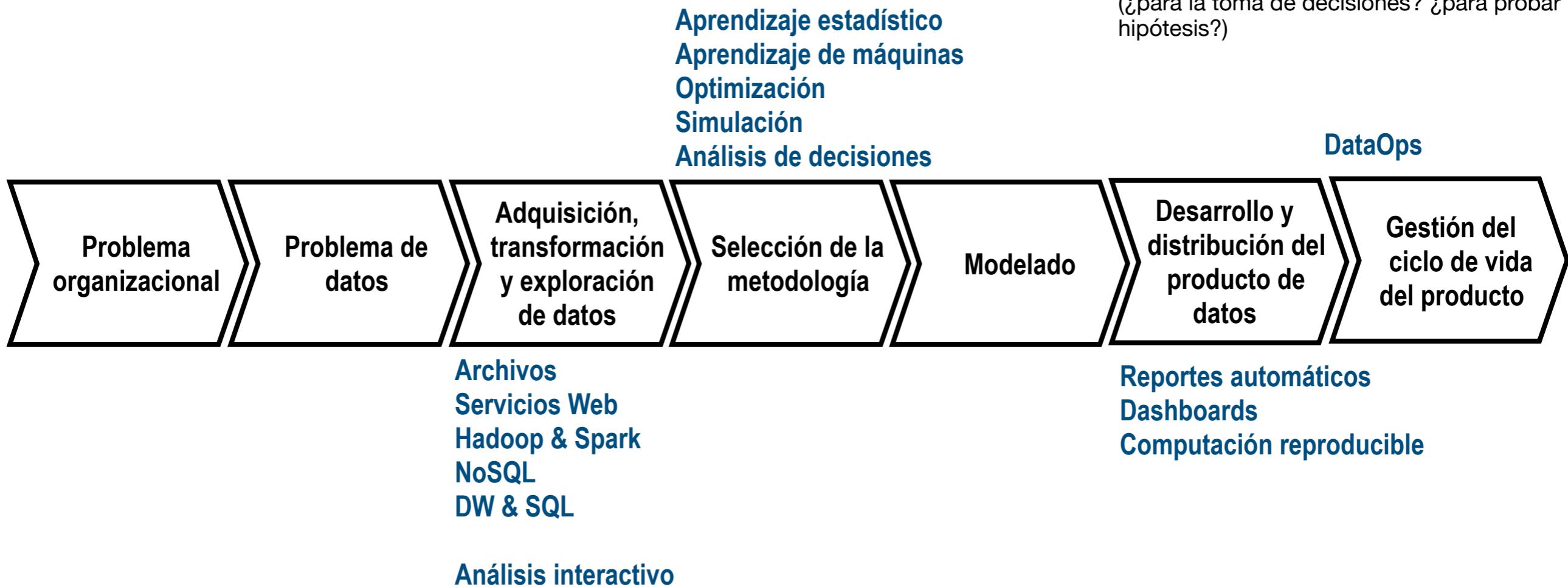
001:{Fecha:2017-10-01, Planta:Jaguas, Generación:100.2}  
002:{Fecha:2017-10-01, Planta:Playas, Generación:23.1}  
003:{Fecha:2017-10-01, Planta:Guatapé, Generación:130.1}

# 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?)

# Modern Analytics (2009)

## FASES / DIMENSIONES

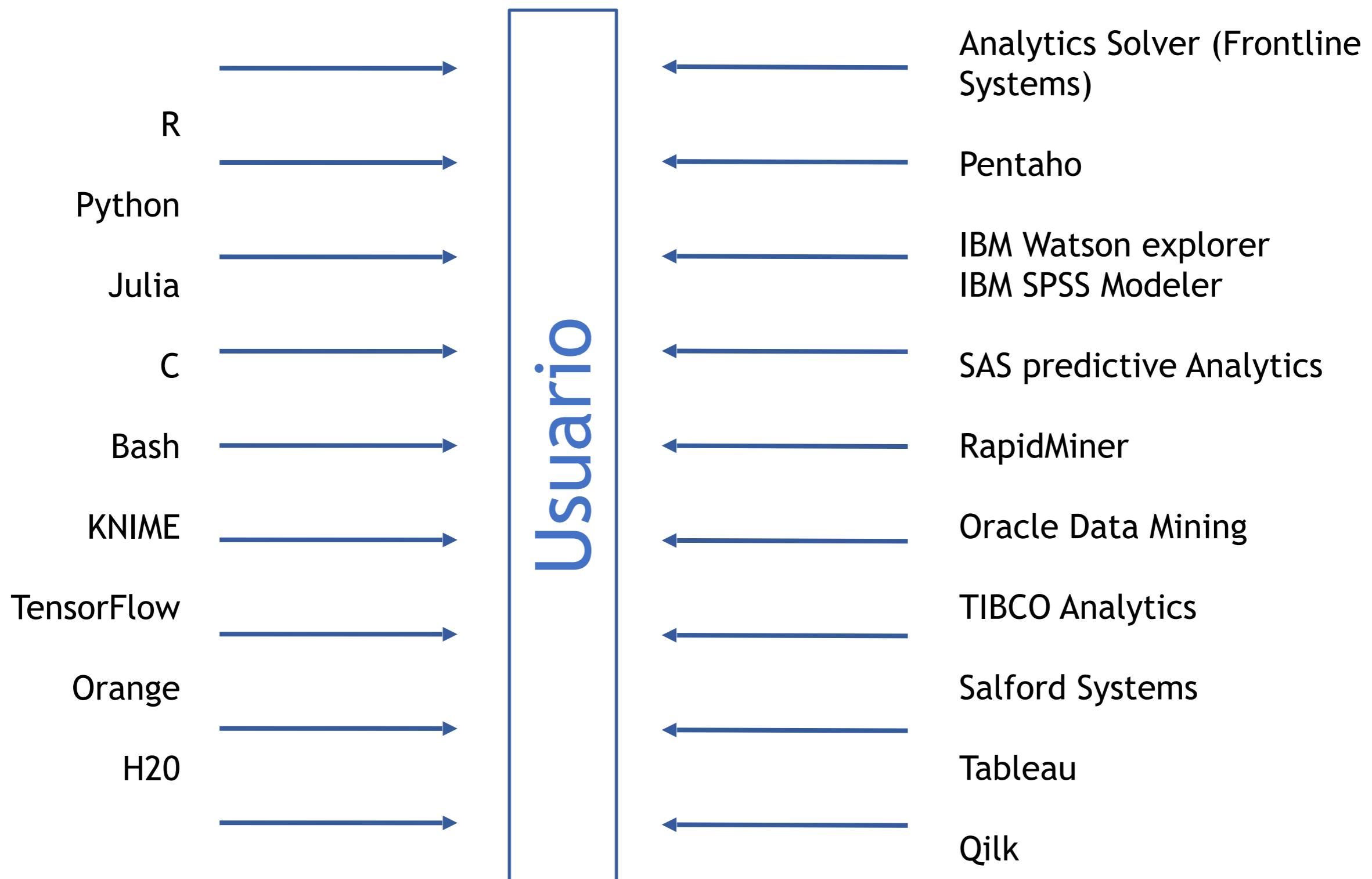
Definición del problema de negocio.	Definición del problema de analítica.	Datos
<p>La habilidad de entender un problema de negocios y determinar cuando el problema es solucionable mediante la analítica.</p> <ul style="list-style-type: none"><li>• Obtener o recibir la definición del problema y los requisitos de usabilidad.</li><li>• Identificar stakeholders.</li><li>• Determinar si el problema es solucionable mediante analítica.</li><li>• Refinar la definición del problema y definir restricciones.</li><li>• Definir el conjunto inicial de beneficios y costos para el negocio.</li><li>• Obtener consenso entre los stakeholders sobre la definición del problema.</li></ul>	<p>Habilidad para reformular un problema de negocio en un problema de datos con una solución analítica potencial.</p> <ul style="list-style-type: none"><li>• Reformular el problema en términos de los datos.</li><li>• Desarrollar un conjunto propuesto de variables explicativas y relaciones con las salidas.</li><li>• Establecer el conjunto de supuestos relacionados con el problema.</li><li>• Definir los criterios de éxito.</li><li>• Realizar el inventario de recursos para la ejecución del proyecto.</li><li>• Obtener consenso entre los stakeholders.</li></ul>	<p>Habilidad de trabajar efectivamente con datos para identificar relaciones potenciales entre variables que ayudarán a refinar la formulación del problema en términos del negocio y de la analítica.</p> <ul style="list-style-type: none"><li>• Identificar, priorizar los datos necesarios y sus fuentes.</li><li>• Adquirir los datos.</li><li>• Armonizar, escalar, limpiar y compartir datos.</li><li>• Construir, integrar y formatear los datos.</li><li>• Analizar los datos e identificar relaciones entre los datos.</li><li>• Verificar la calidad de los datos.</li><li>• Documentar y reportar hallazgos.</li><li>• Refinar la formulación del problema en términos del negocio y de la analítica.</li></ul>

# Modern Analytics (2009)

## FASES / DIMENSIONES

<b>Selección de la metodología.</b>	<b>Desarrollo del modelo</b>	<b>Despliegue del modelo</b>	<b>Gestión del ciclo de vida del modelo</b>
<p>Habilidad para identificar y seleccionar aproximaciones potenciales para resolver el problema de negocio</p> <ul style="list-style-type: none"><li>• Identificar las aproximaciones disponibles para la solución del problema.</li><li>• Seleccionar las herramientas de software.</li><li>• Definir los métodos para probar los modelos.</li><li>• Refinar el análisis de costos y beneficios</li><li>• Desarrollar un plan inicial del proyecto.</li><li>• Seleccionar las aproximaciones a utilizar.</li></ul>	<p>Habilidad para identificar y construir modelos efectivos para ayudar a resolver el problema de negocio.</p> <ul style="list-style-type: none"><li>• Identificar las estructuras del o los modelos.</li><li>• Correr y evaluar modelos.</li><li>• Calibrar modelos y datos.</li><li>• Integrar modelos.</li><li>• Documentar y reportar hallazgos.</li></ul>	<p>Habilidad para realizar el despliegue del modelo que ayuda a solucionar el problema de negocio.</p> <ul style="list-style-type: none"><li>• Evaluación del modelo en términos del negocio.</li><li>• Publicar reportes con hallazgos.</li><li>• Desarrollar los requerimientos del modelo, del sistema y de usabilidad para producción.</li><li>• Despliegue del modelo/ sistema en producción.</li><li>• Soportar el desarrollo</li></ul>	<p>Habilidad para gestionar el ciclo de vida con el fin de evaluar los beneficios del modelo para el negocio sobre el tiempo.</p> <ul style="list-style-type: none"><li>• Documentar la estructura del modelo.</li><li>• Evaluar permanentemente la calidad del modelo</li><li>• Recalibrar y mantener el modelo.</li><li>• Desarrollar actividades de entrenamiento.</li><li>• Evaluar periodicamente los beneficios del modelo.</li></ul>

# Open Data Science & Modern Analytics (2009)

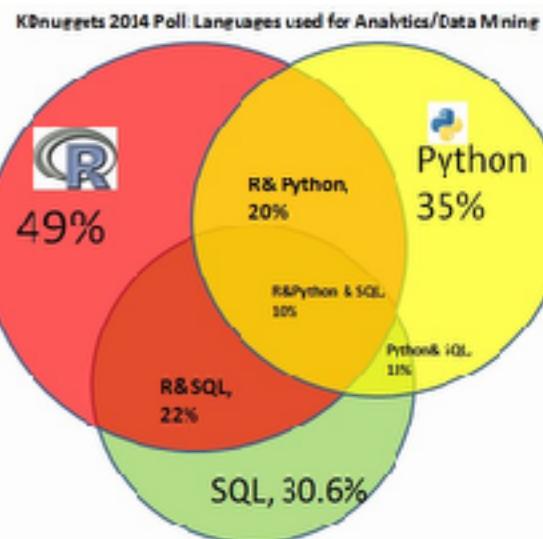


# Open Data Science & Modern Analytics (2009)

## The Top Ten Programming Languages (IEEE Spectrum)

Language Rank	Types	Spectrum Ranking	Spectrum Ranking
1. Java	🌐💻🖥️	100.0	100.0
2. C	💻🖥️⚙️	99.9	89.3
3. C++	💻🖥️⚙️	99.4	95.5
4. Python	🌐💻	98.5	93.5
5. C#	🌐💻	91.9	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

Language Rank	Types	Spectrum Ranking
1. Python	🌐💻	100.0
2. C	💻🖥️⚙️	99.7
3. Java	🌐💻🖥️	98.5
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



2015

2016

2017

2018

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	🌐	82.6
8. JavaScript	🌐💻	82.2
9. Ruby	🌐💻	74.5
10. Go	🌐💻	71.9

Language Rank	Types	Spectrum Ranking
1. Python	🌐💻⚙️	100.0
2. C++	💻🖥️⚙️	98.4
3. C	💻🖥️⚙️	98.2
4. Java	🌐💻	97.5
5. C#	🌐💻	89.8
6. PHP	🌐	85.4
7. R	💻	83.3
8. JavaScript	🌐💻	82.6
9. Go	🌐💻	76.7
10. Assembly	⚙️	74.5

# 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)

```
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
```

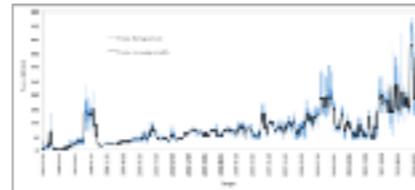
ESTACION;FECHA;HORA;DIRECCION;VELOCIDAD
AQUITANIA;2005-04-16;11:10:00;135;6,3
AQUITANIA;2005-04-16;11:20:00;135;5,1
AQUITANIA;2005-04-16;11:30:00;135;6,3
AQUITANIA;2005-04-16;11:40:00;113;6,1
AQUITANIA;2005-04-16;11:50:00;135;4,1
AQUITANIA;2005-04-16;12:00:00;135;5,5
AQUITANIA;2005-04-16;12:10:00;135;5,4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

# 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

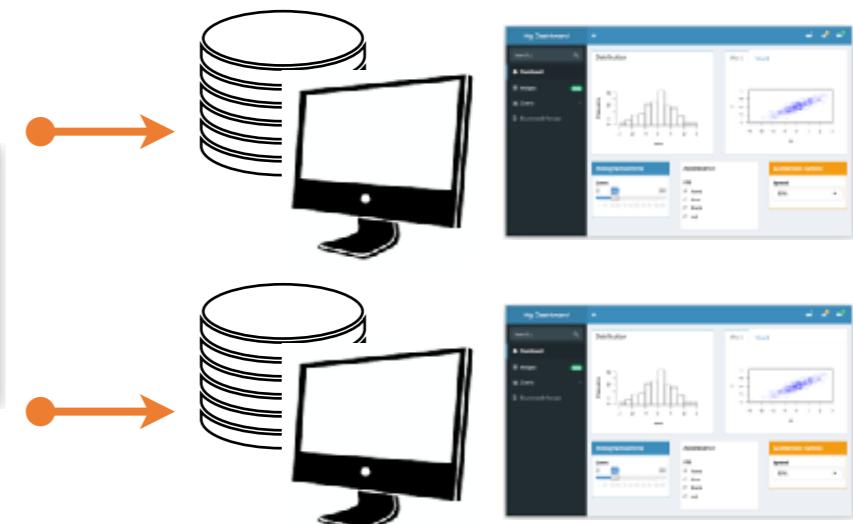
Minería de  
Datos

DW / OLAP



Descubrimiento de  
patrones y tendencias  
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.

# Open Data Science & Modern Analytics (2009)

## Aplicaciones modernas

- Notebooks
- Dashboards
- Aplicaciones visuales
- Servicios de datos

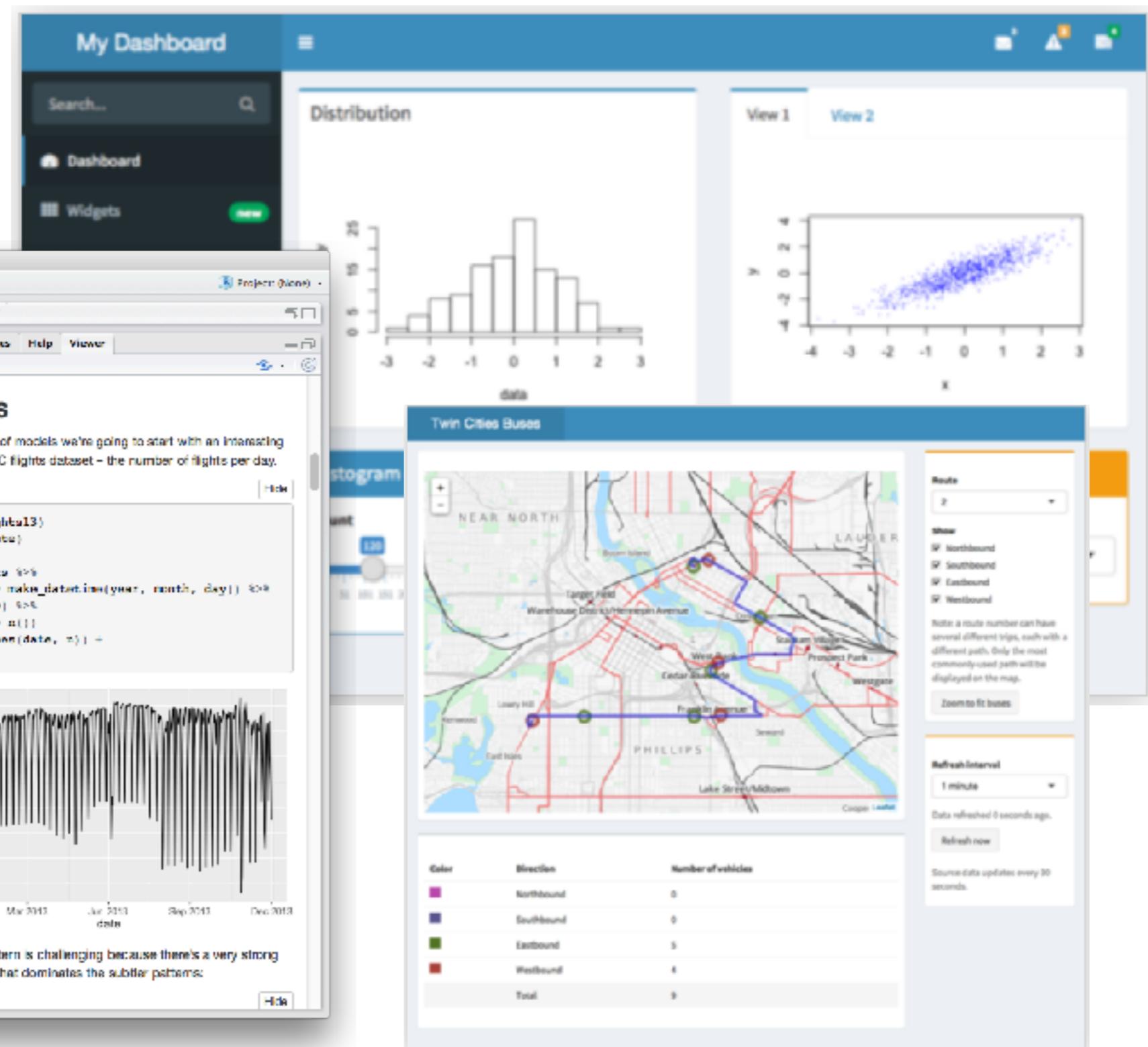
## R Dashboards

## R Markdown

The screenshot shows the RStudio interface. On the left, the code editor displays R Markdown code for generating a histogram and a line plot. The right pane contains two plots: a histogram of flight residuals and a line plot showing flight counts over time.

```
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
70 ##(r)
71 library(nycflights)
72 library(lubridate)
73 library(dplyr)
74
75 daily <- flights %>%
76   mutate(date = make_datetime(year, month, day)) %>%
77   group_by(date) %>%
78   summarise(n = n())
79
80 ggplot(daily, aes(date, n)) +
81   geom_line()
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
```

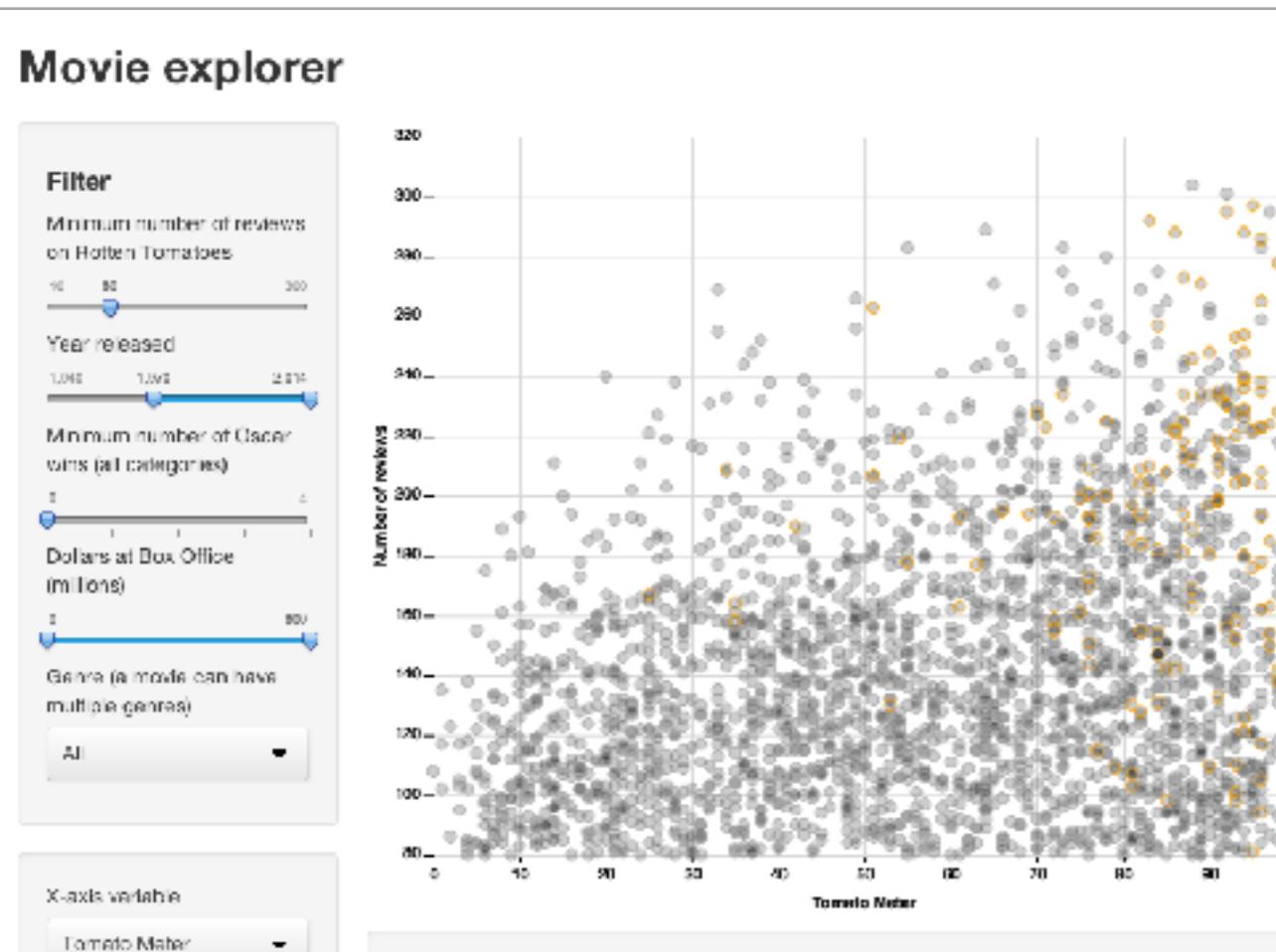
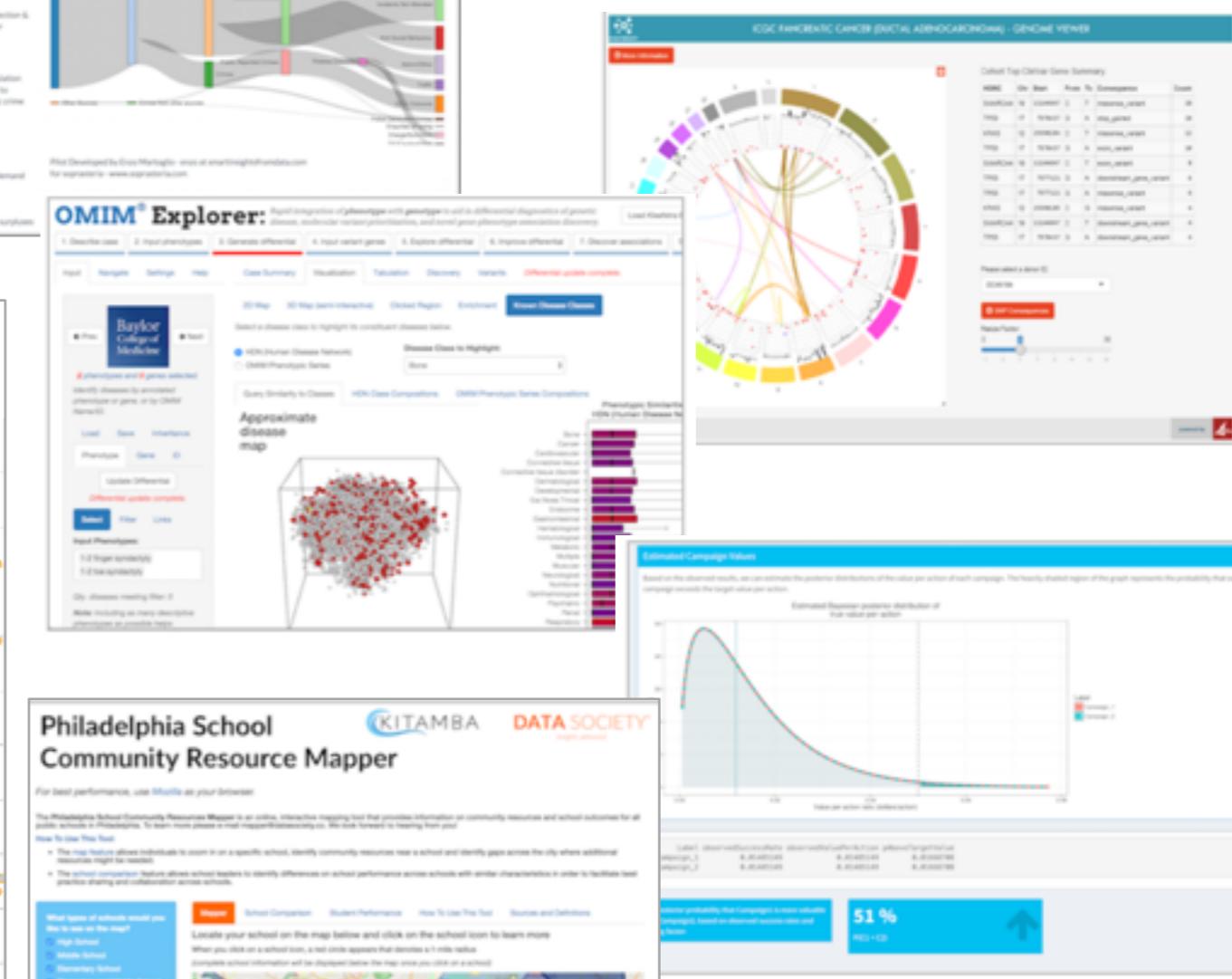
Understand this pattern is challenging because there's a very strong day-of-week effect that dominates the subtler patterns.



# Open Data Science & Modern Analytics (2009)

# Aplicaciones modernas

- Notebooks
  - Dashboards
  - Aplicaciones visuales
  - Servicios de datos



# Open Data Science & Modern Analytics (2009)

## Visualización

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

## BeakerX

### BeakerX: Beaker extensions for Jupyter

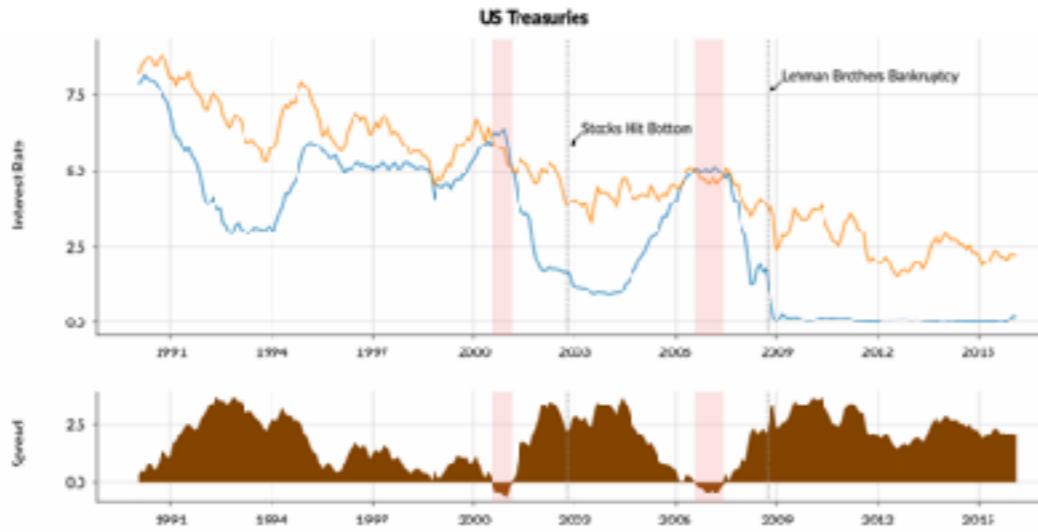
[build passing](#) [chat on gitter](#) [JPack 0.1.1](#) [npm package 0.0.6](#) [pypl package 0.2.4.dev0](#)

BeakerX is a collection of JVM kernels with widgets, plotting, tables, autotranslation, and other extensions to the Jupyter Notebook and IPython. It is built on top of the Beaker framework, which provides a rich set of tools for scientific computing and data visualization.

#### 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)

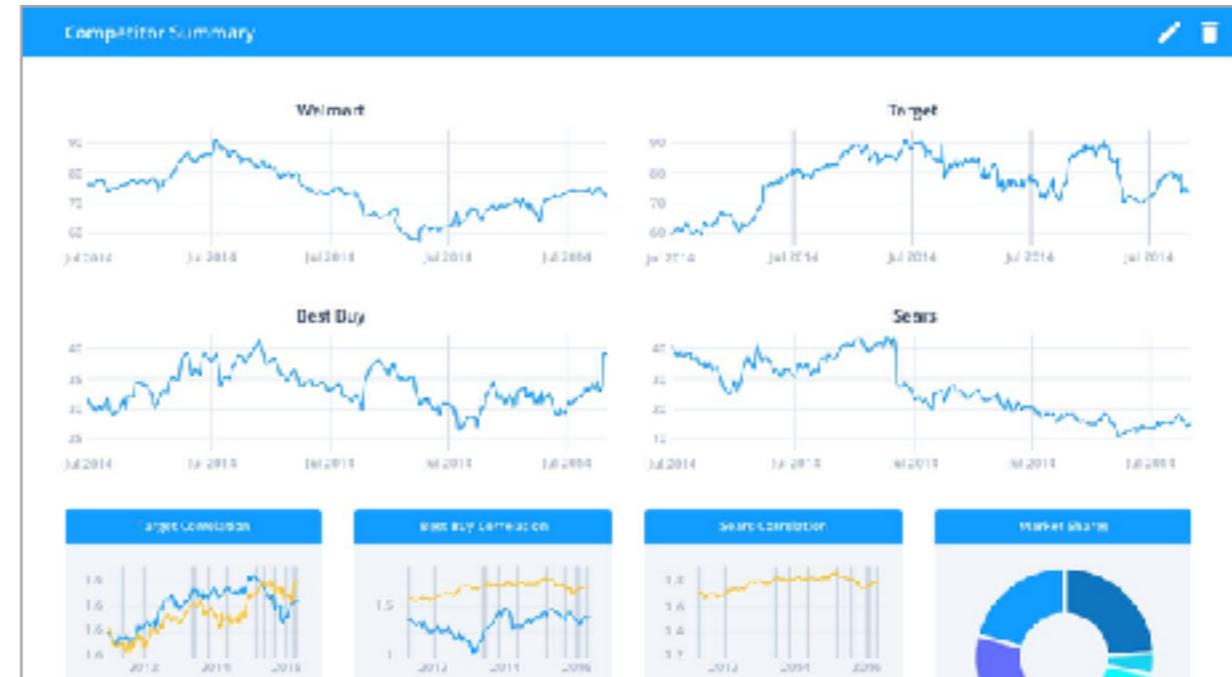
// add both plots to the combined plot, and including their relative heights.
c.add(p1, 3)
c.add(p2, 1)
```



## Bokeh



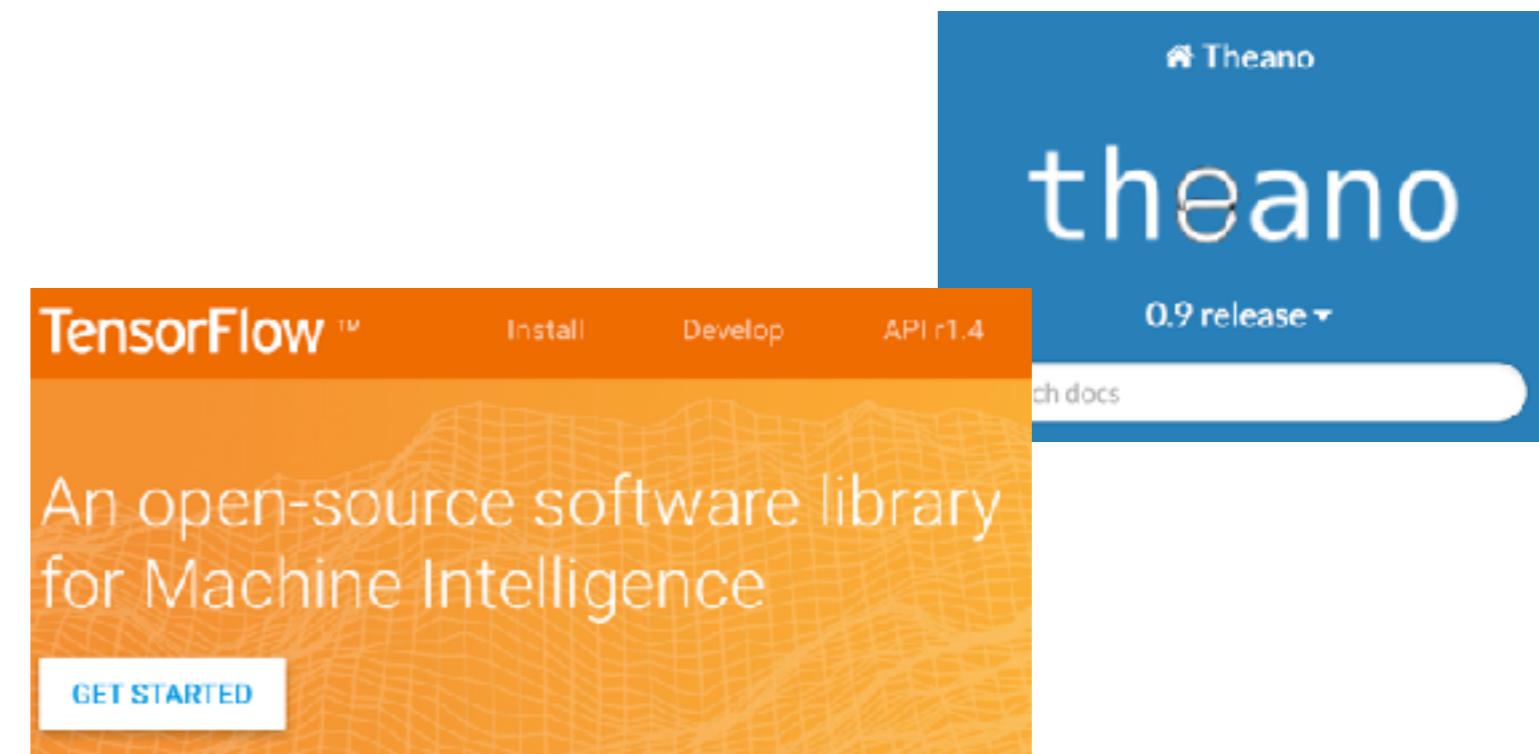
## [plot.ly](#)



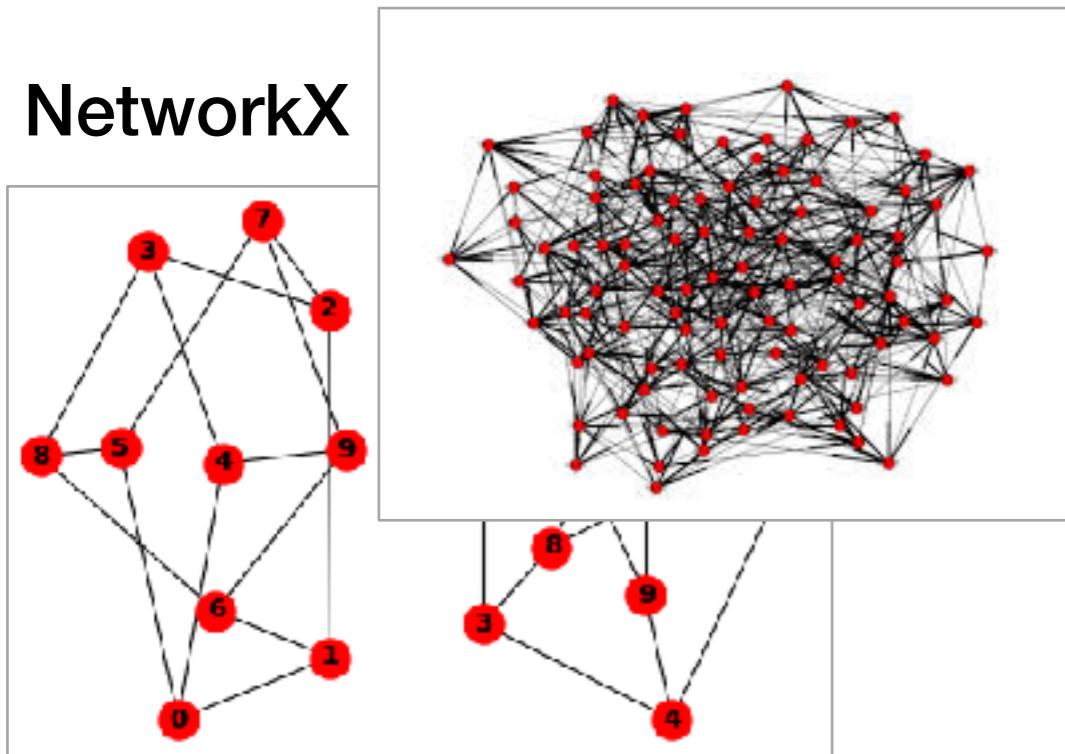
# Open Data Science & Modern Analytics (2009)

## Analytics

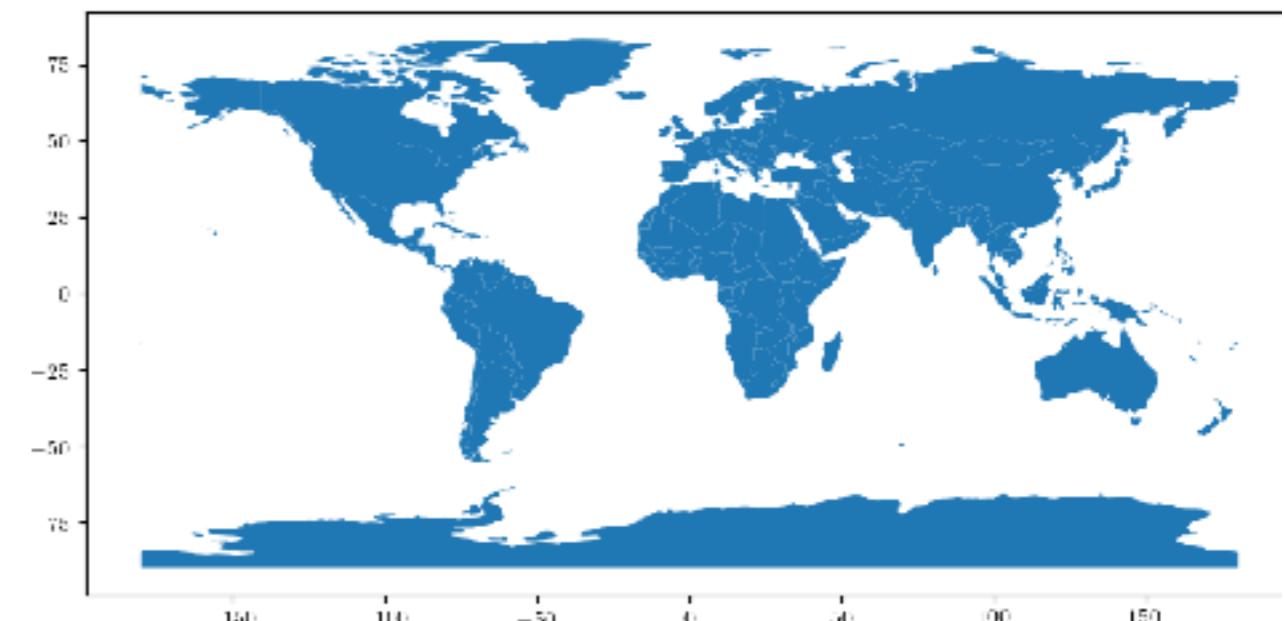
- Preparación de datos
- Estadística
- ML & Ensambles
- Deep learning
- Simulación y optimización
- Datos geoespaciales
- Texto y NLP
- Gráficos y redes
- Minería de audio, video e imágenes



## NetworkX



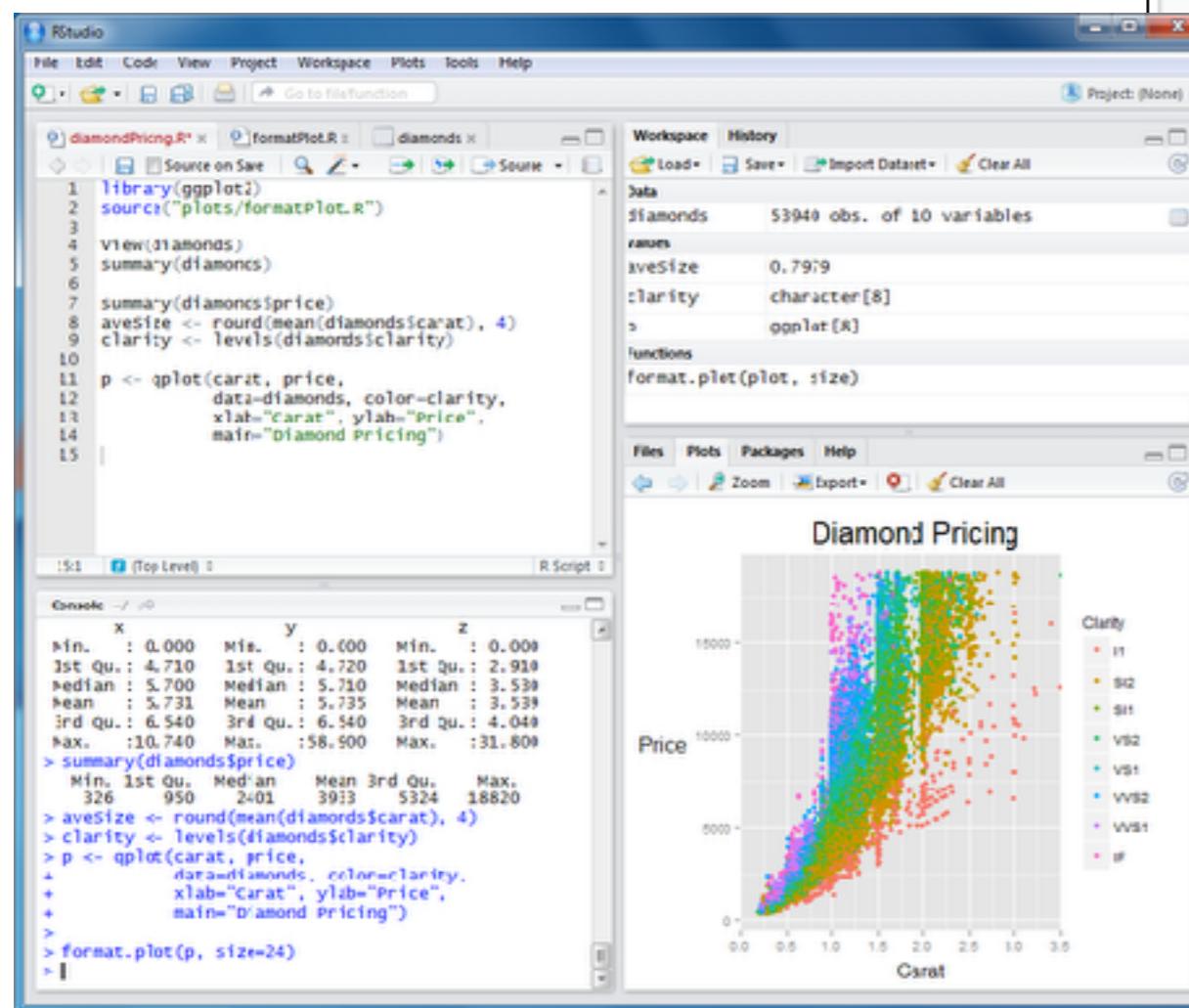
## GeoPandas



# Open Data Science & Modern Analytics (2009)

## 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



```
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
```

P. Variable:	Lottery	R-squared:	0.348
del:	OLS	Adj. R-squared:	0.333
thod:	Least Squares	F-statistic:	22.20
te:	Tue, 28 Feb 2017	Prob (F-statistic):	1.90e-08
ne:	21:38:05	Log-Likelihood:	-379.82
. 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('SELECT * FROM person')
6 print c.fetchall()
7 c.execute('SELECT * FROM address')
8 print c.fetchall()
9 conn.close()
```

```
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()
```

# 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/>

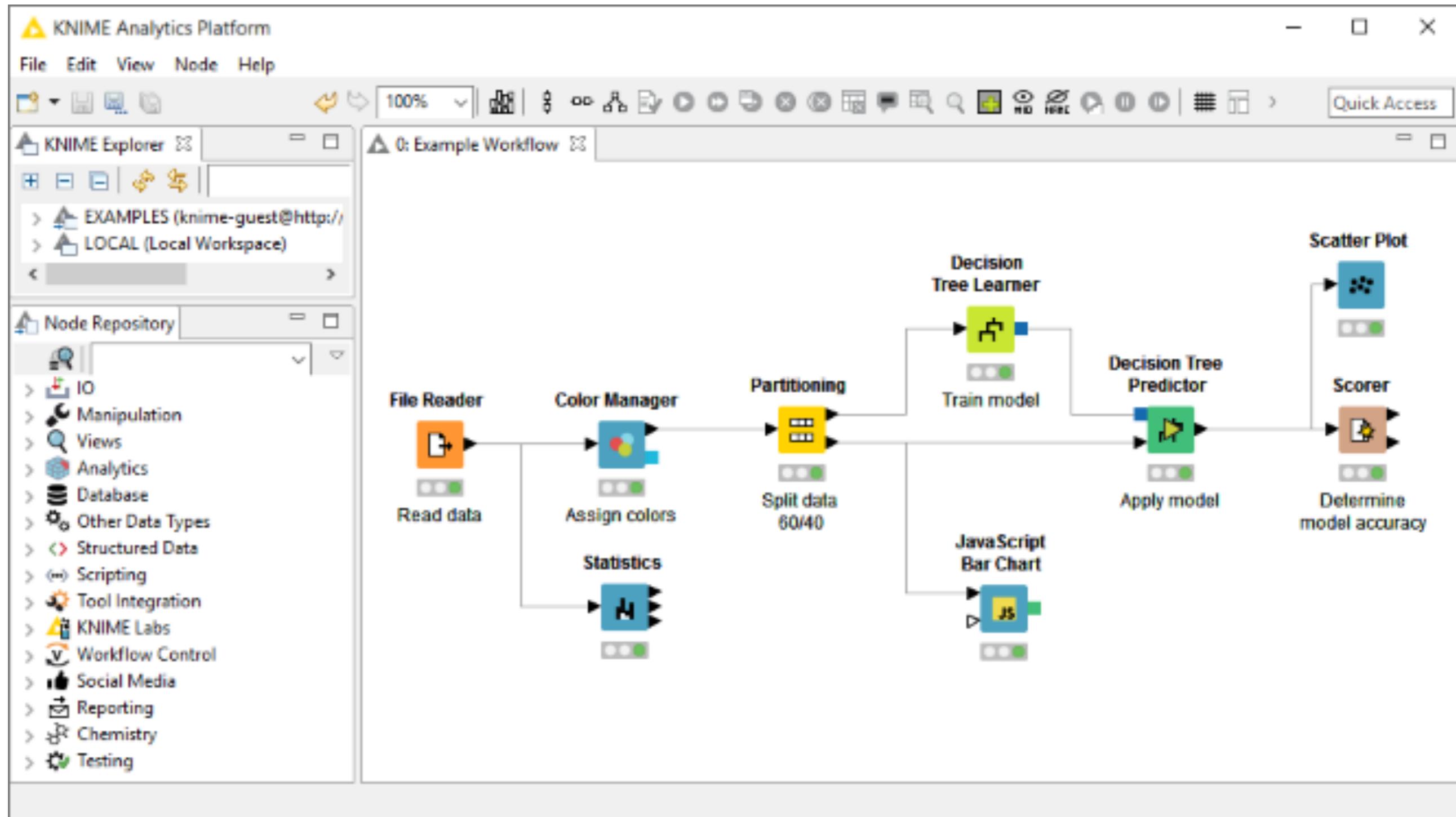
pyCUDA – <https://mathematician.de/software/pycuda/>

```
from numba import jit
from numpy import arange

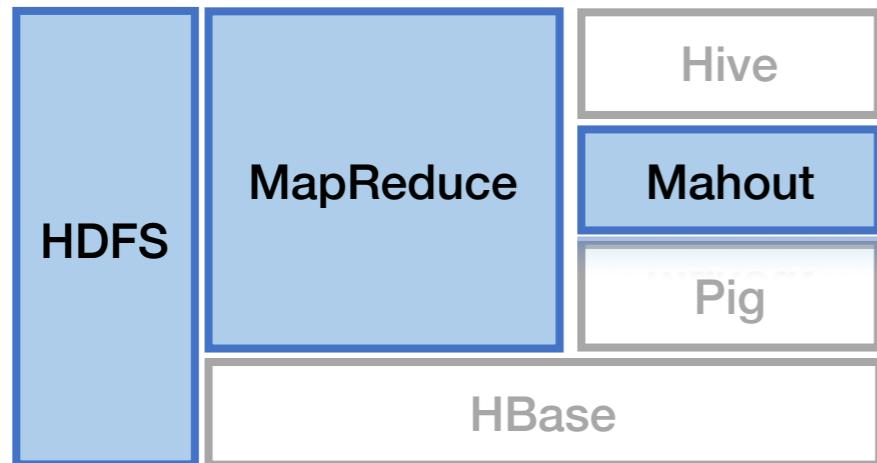
# jit decorator tells Numba to compile this function.
# The argument types will be inferred by Numba when function is called.
@jit
def sum2d(arr):
    M, N = arr.shape
    result = 0.0
    for i in range(M):
        for j in range(N):
            result += arr[i,j]
    return result

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

# KNIME (Lanzada desde 2006)

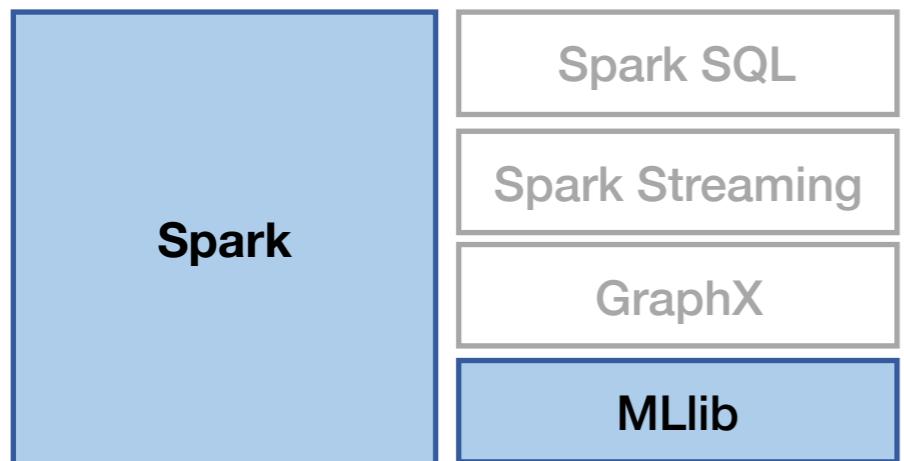


# Big Data Analytics (2011)



## 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

Theano

H2O

Keras

TensorFlow

# Pyomo (2012)



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](#))

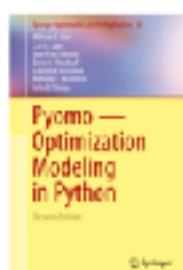
Online examples from the Pyomo software repository: ([browse](#)) ([zipfile](#))

### 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-Bethany L. Nicholson, and John D. Floudas. Pyomo — Optimization Modeling in Python, Second Edition. Vol. 67. Springer, 2012.

The Second Edition of the book describes the capabilities of Pyomo 4.0. The features described in the book that are not yet available in Pyomo 4.0 are not included in this version of the book.

$$\begin{aligned} \text{min } & 2x_1 + 3x_2 \\ \text{s. t. } & 3x_1 + 4x_2 \geq 1 \\ & x_1, x_2 \geq 0 \end{aligned}$$

```
from __future__ import division
from pyomo.environ import *

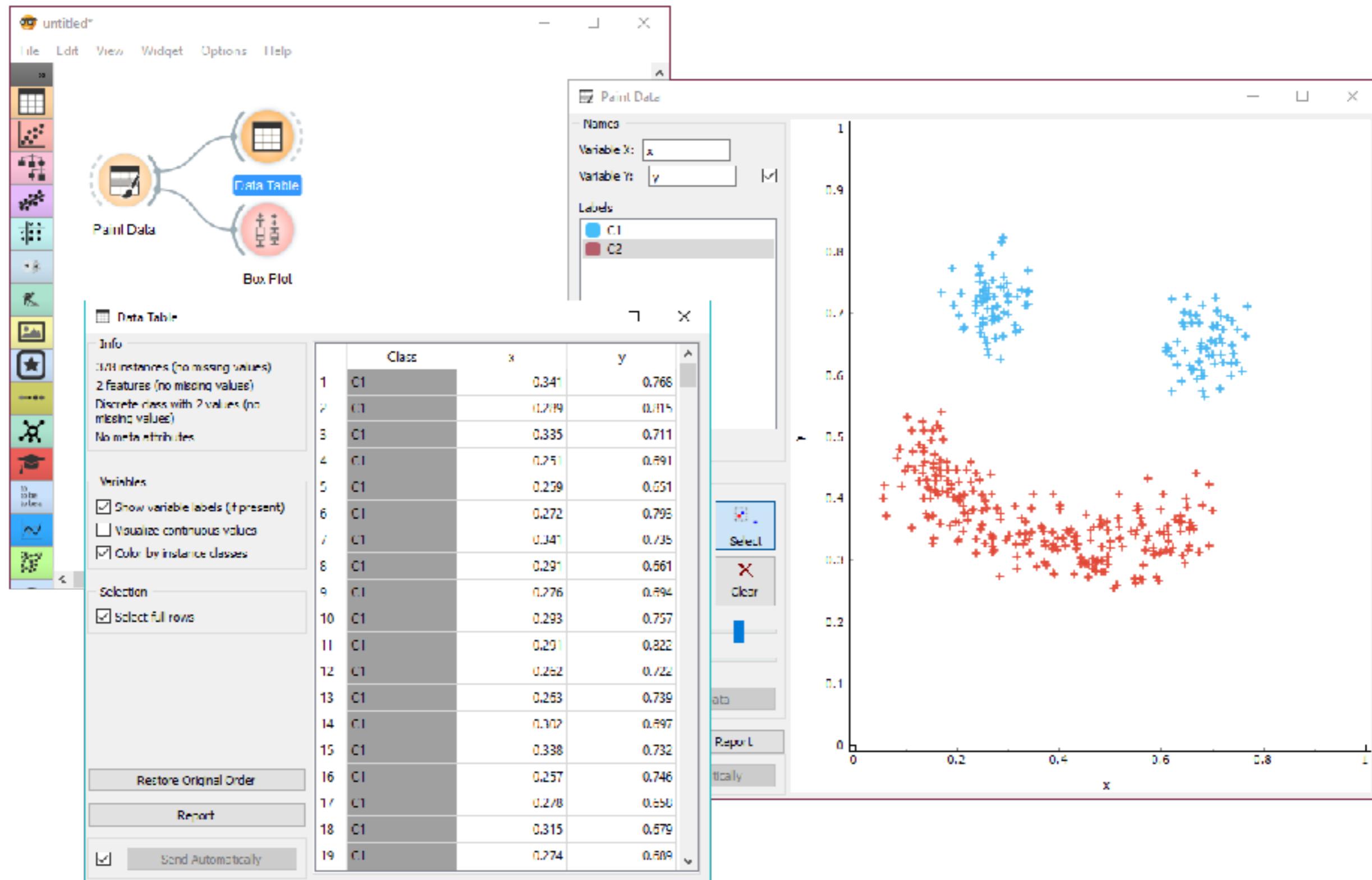
model = ConcreteModel()

model.x = Var([1,2], domain=NonNegativeReals)

model.OBJ = Objective(expr = 2*model.x[1] + 3*model.x[2])

model.Constraint1 = Constraint(expr = 3*model.x[1] + 4*model.x[2] >= 1)
```

# Orange3 (2013)

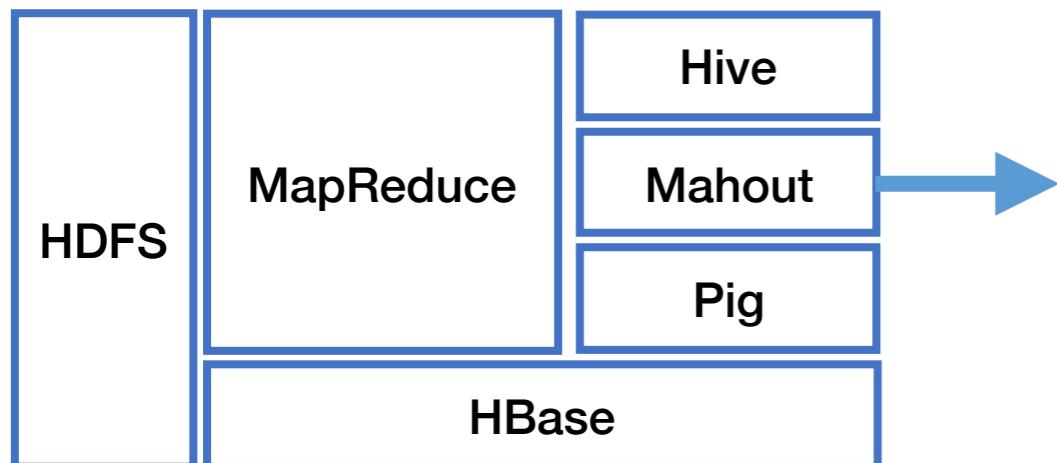


# Apache Hive (2013)

## 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;
```

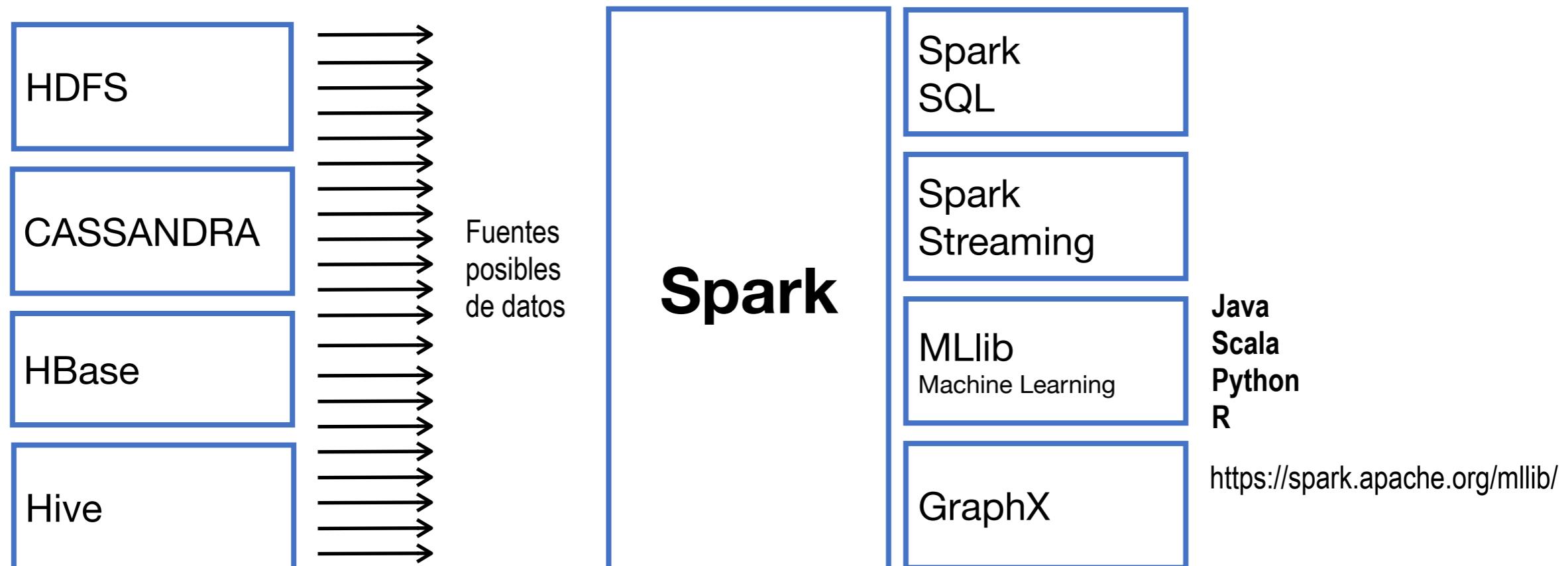
# Apache Spark (2014)



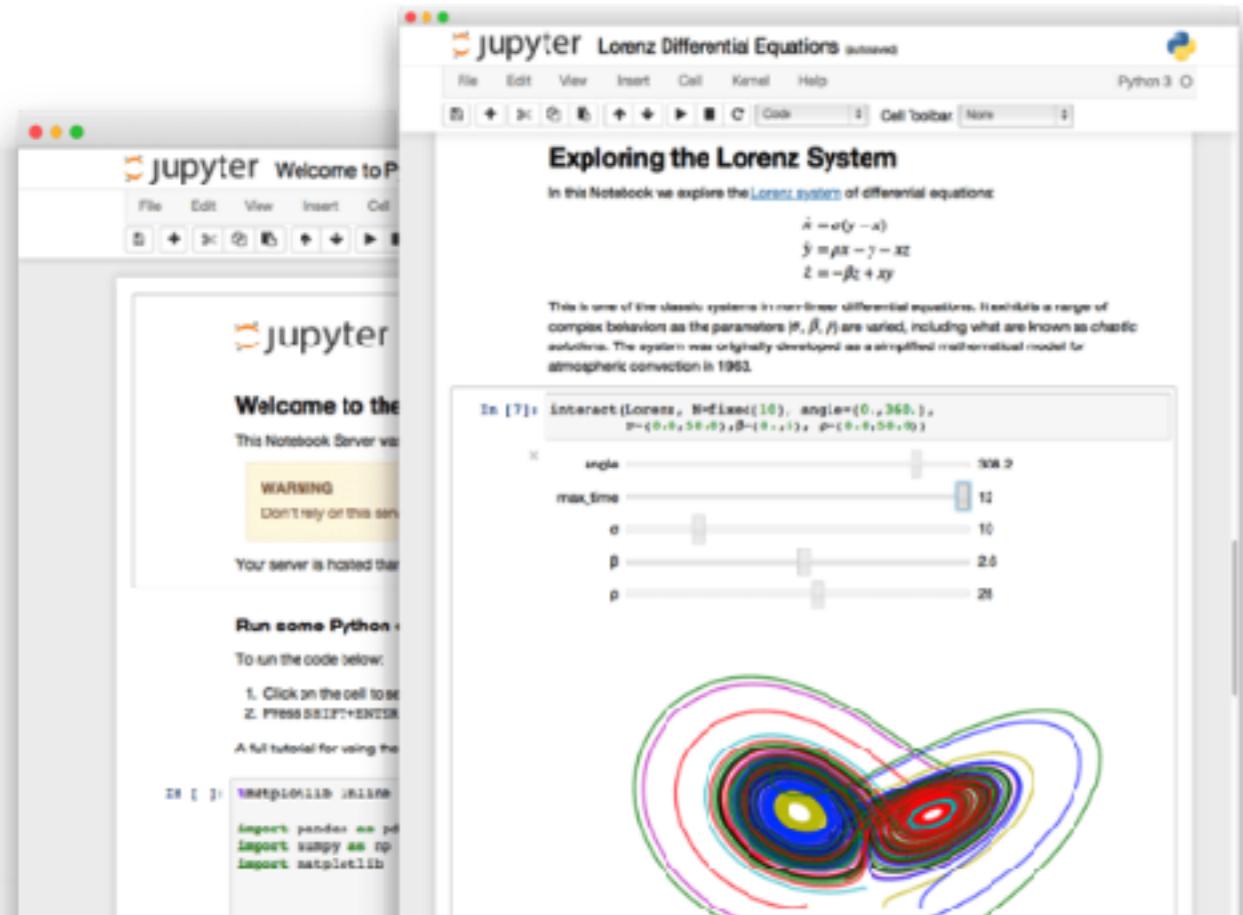
Hadoop / MapReduce

Regresión logística  
Regresión lineal  
Clustering  
Filtrado colaborativo  
<http://mahout.apache.org/users/basics/algorithms.html>

RHadoop  
rdfs  
rnr  
rhbase



# Jupyter Notebook (2015)



## Jupyter QtConsole

# The Jupyter Notebook

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

Try it in your browser

Install the Notebook

A screenshot of the Jupyter QtConsole. The title bar says 'Jupyter QtConsole'. The console window shows a Python session. In cell [1], the command `print?` is run, displaying the docstring and help information for the `print` function. The docstring is: `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`.

# Apache Zeppelin (2015)

## Multi-purpose Notebook

The Notebook is the place for all your needs

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

 **Zeppelin** Notebook -   anonymous

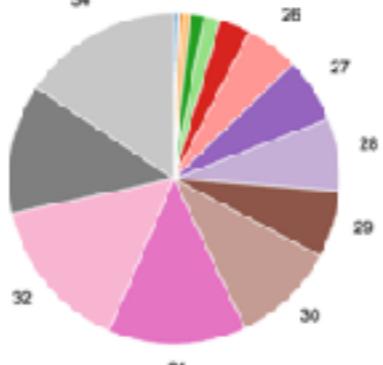
**Bank** 

**marital**  
finished

**age**  
35

**settings**

19	20	21	22	23	24	25	26
27	28	29	30	31	32	33	34



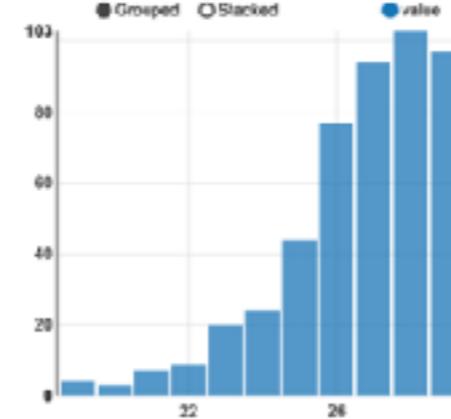
Took a few seconds. Last updated by anonymous at June 26 2014, 4:46:52 PM.  
[outdated]

**Under age < 35**  
finished

**age**  
30

**settings**

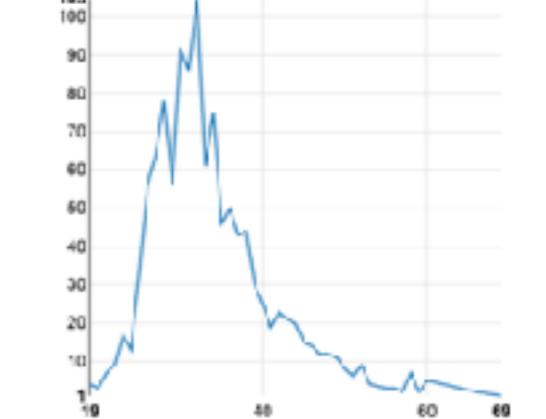
Grouped  Slacked **value**



Took a few seconds. Last updated by anonymous at June 26 2014, 4:47:32 PM.  
[outdated]

**marital**  
single

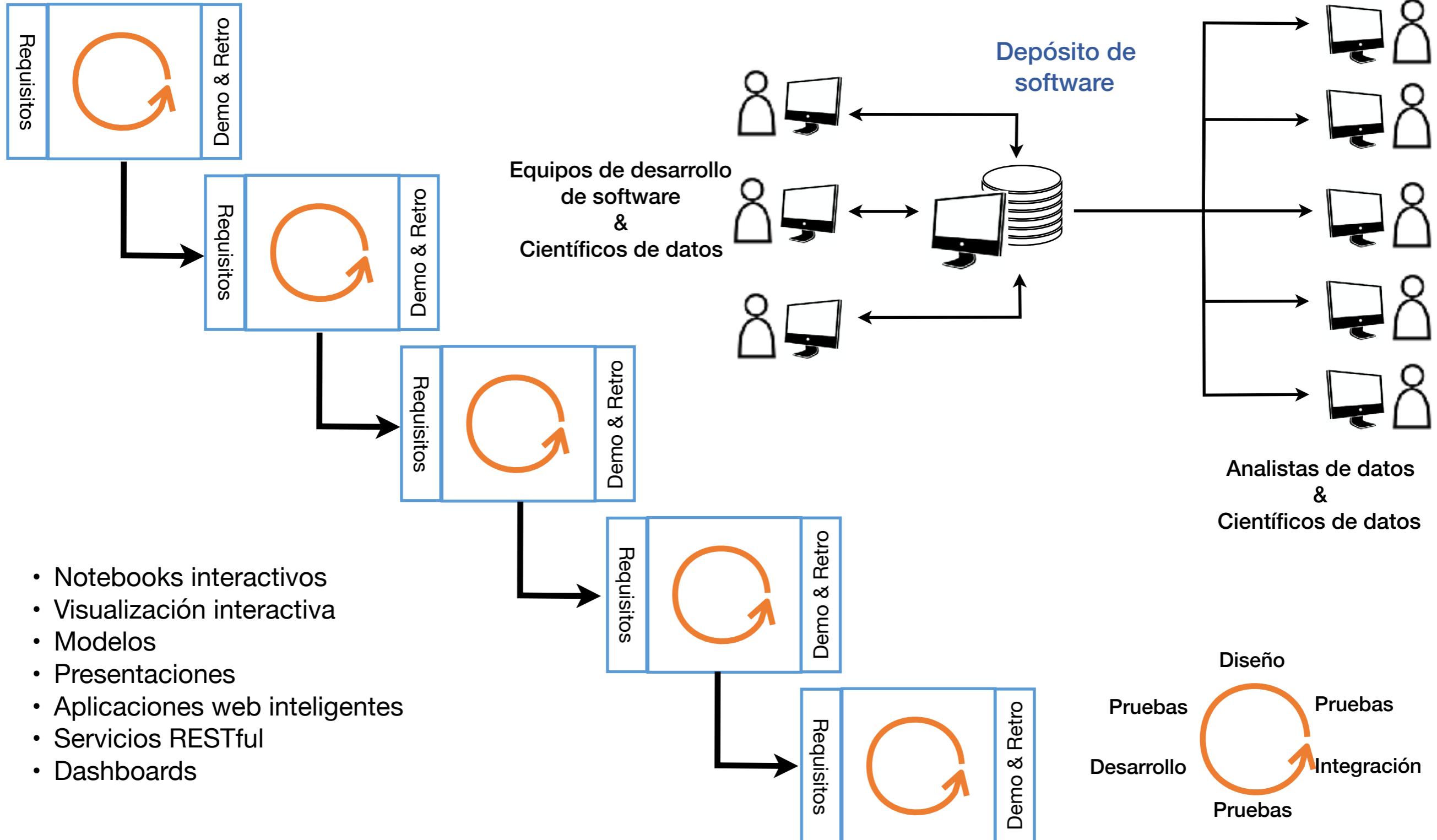
**settings**



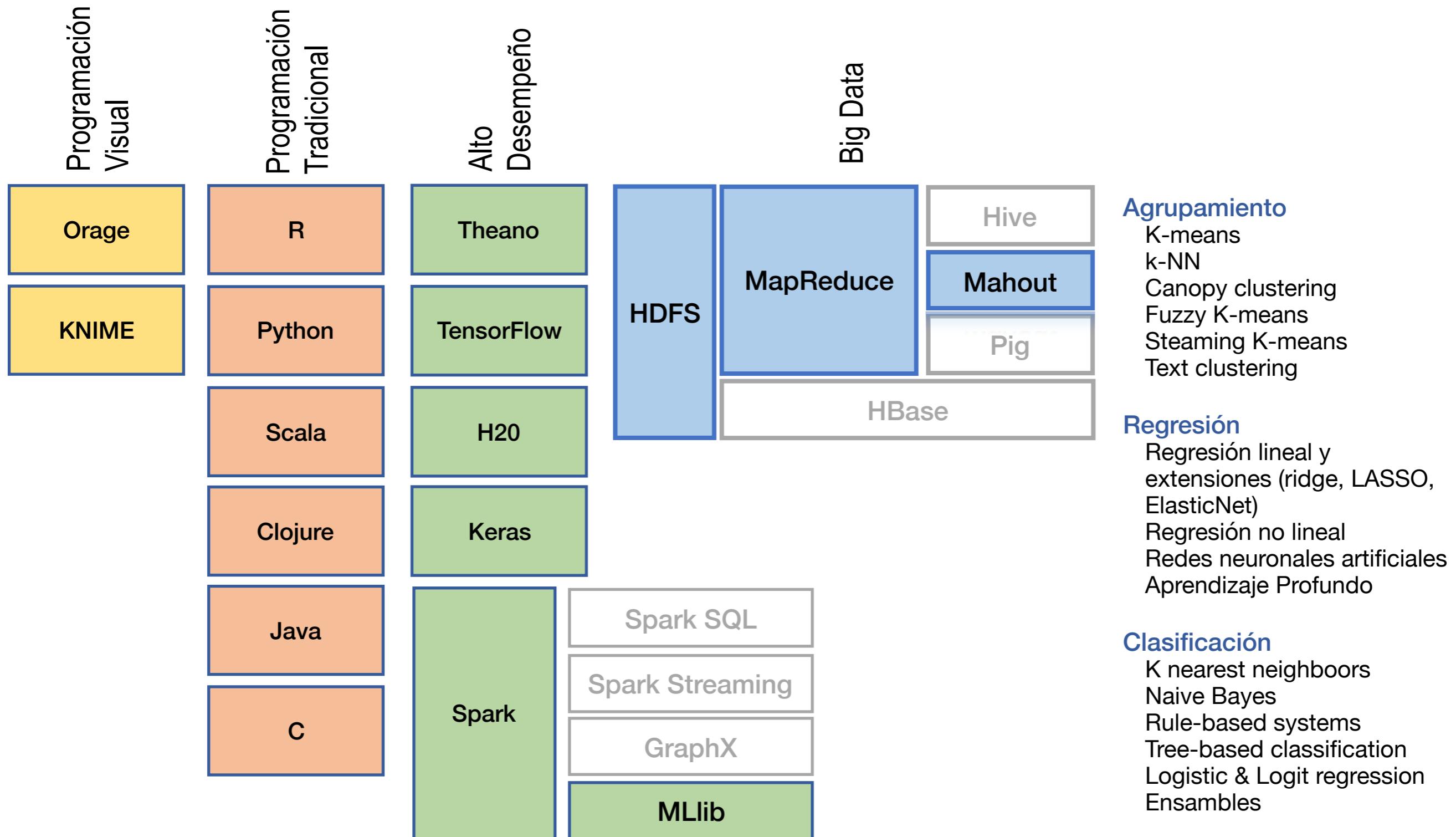
Took a few seconds. Last updated by anonymous at June 26 2014, 4:47:34 PM.  
[outdated]

**READY** 

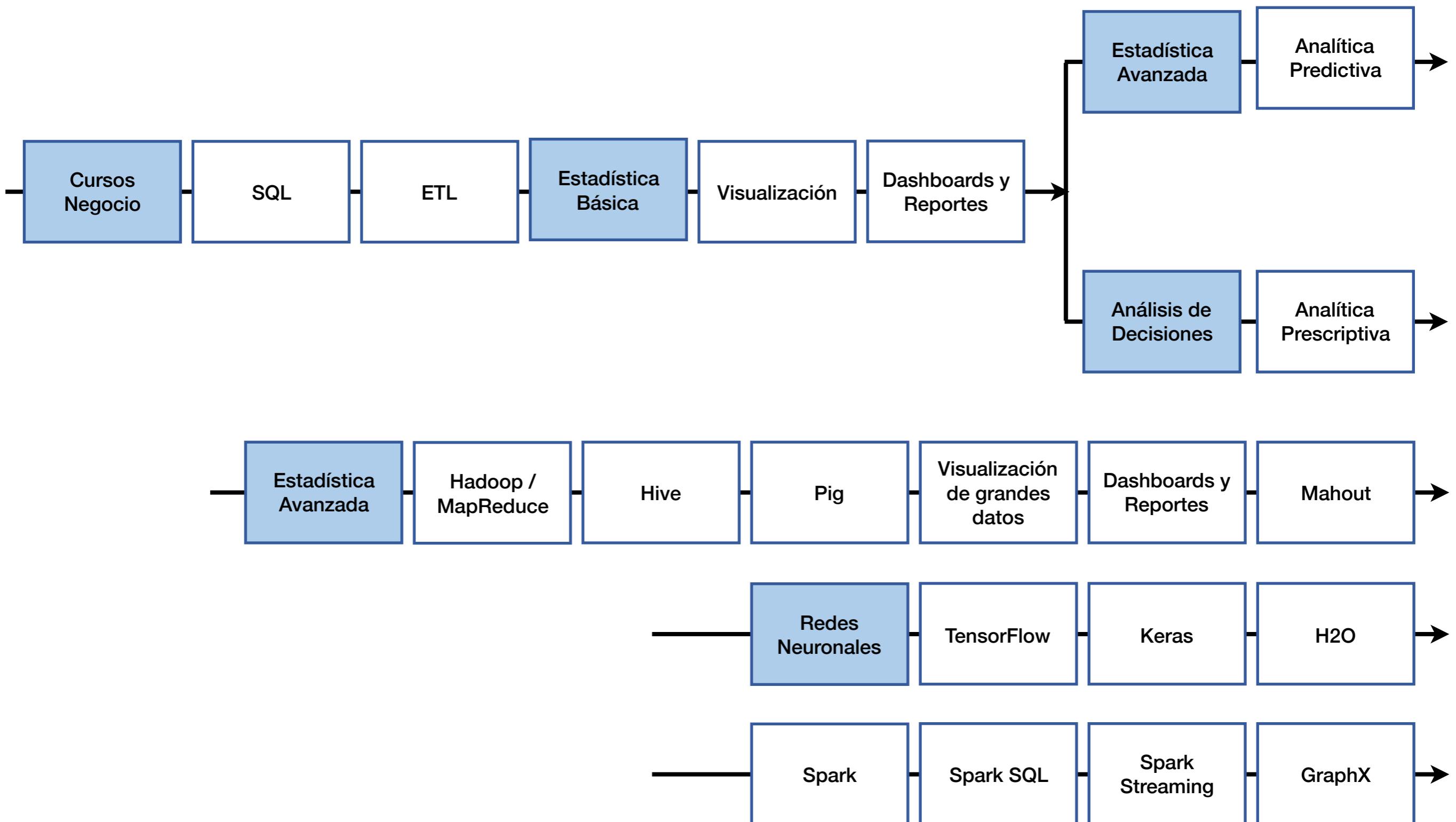
# DataOps (2015)



# Open Data Science & Modern Analytics (2018)



# Open Data Science & Modern Analytics (2018)



# Una Introducción a la Analítica

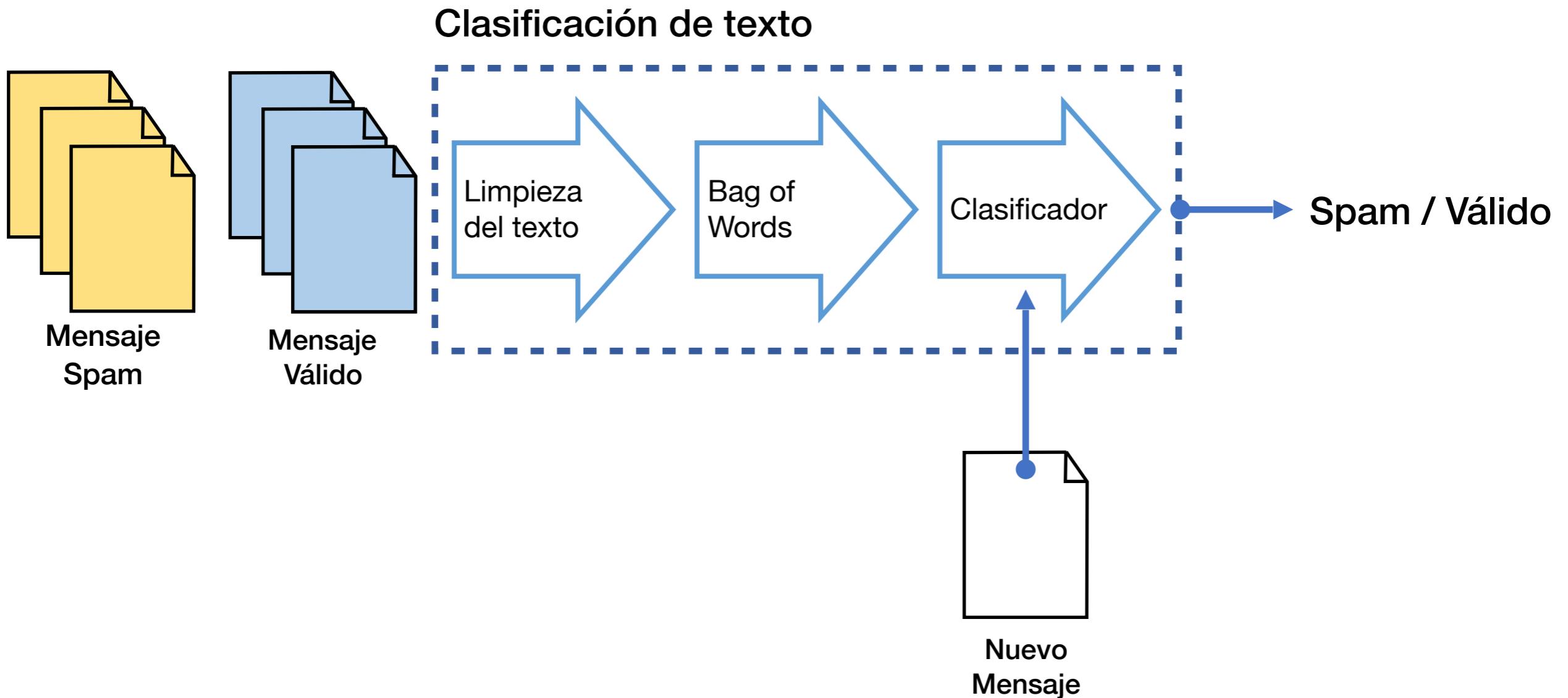
## Algunos casos de uso de Machine Learning

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.

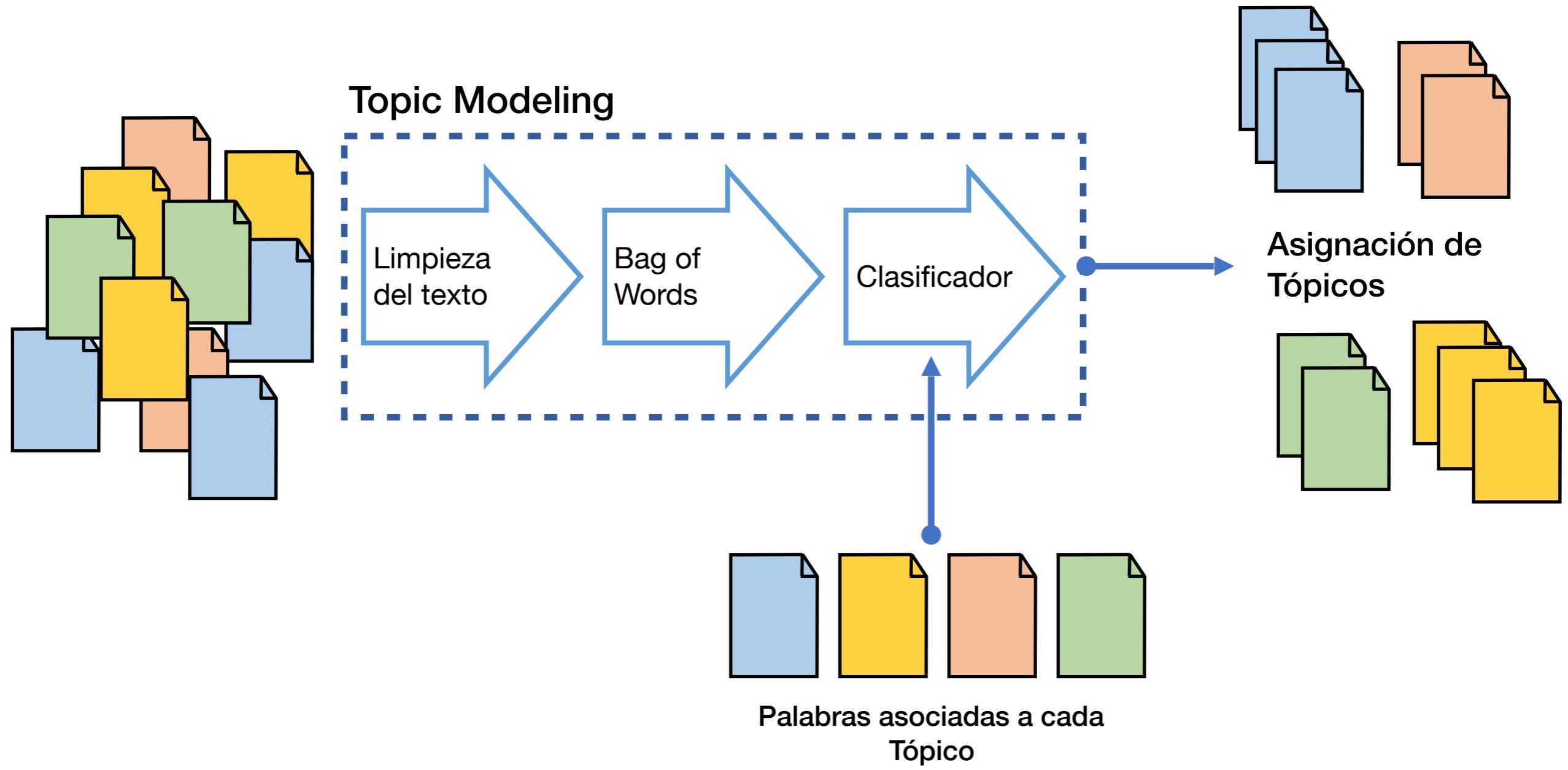
Descargue la última versión de este documento de:  
<https://github.com/jdvelasq/data-science-docs/blob/master/sena.pdf>

Disciplina	Tecnología	Habilidades	Foco
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>
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>
Ciencia de los Datos (Analytics)	<ul style="list-style-type: none"><li>Software estadístico</li><li>Datos columnares</li><li>Map-Reduce</li><li>NoSQL</li><li>Lenguajes de programación</li><li>Software para graficación</li><li>Software para optimización, simulación, predicción y análisis de decisiones</li></ul>	<ul style="list-style-type: none"><li>Estadística avanzada</li><li>Programación</li><li>Análisis del negocio</li><li>Arquitecturas y tecnologías modernas para el manejo de datos</li><li>Desarrollo de productos de datos</li><li>Simulación de sistemas</li><li>Optimización</li><li>Predicción</li></ul>	<ul style="list-style-type: none"><li>Modelado predictivo</li><li>Análisis estadístico avanzado</li><li>Minería de datos</li><li>Manejo de datos no estructurados</li><li>Manejo de grandes volúmenes de datos</li><li>I+D</li><li>Análisis de decisiones</li></ul>

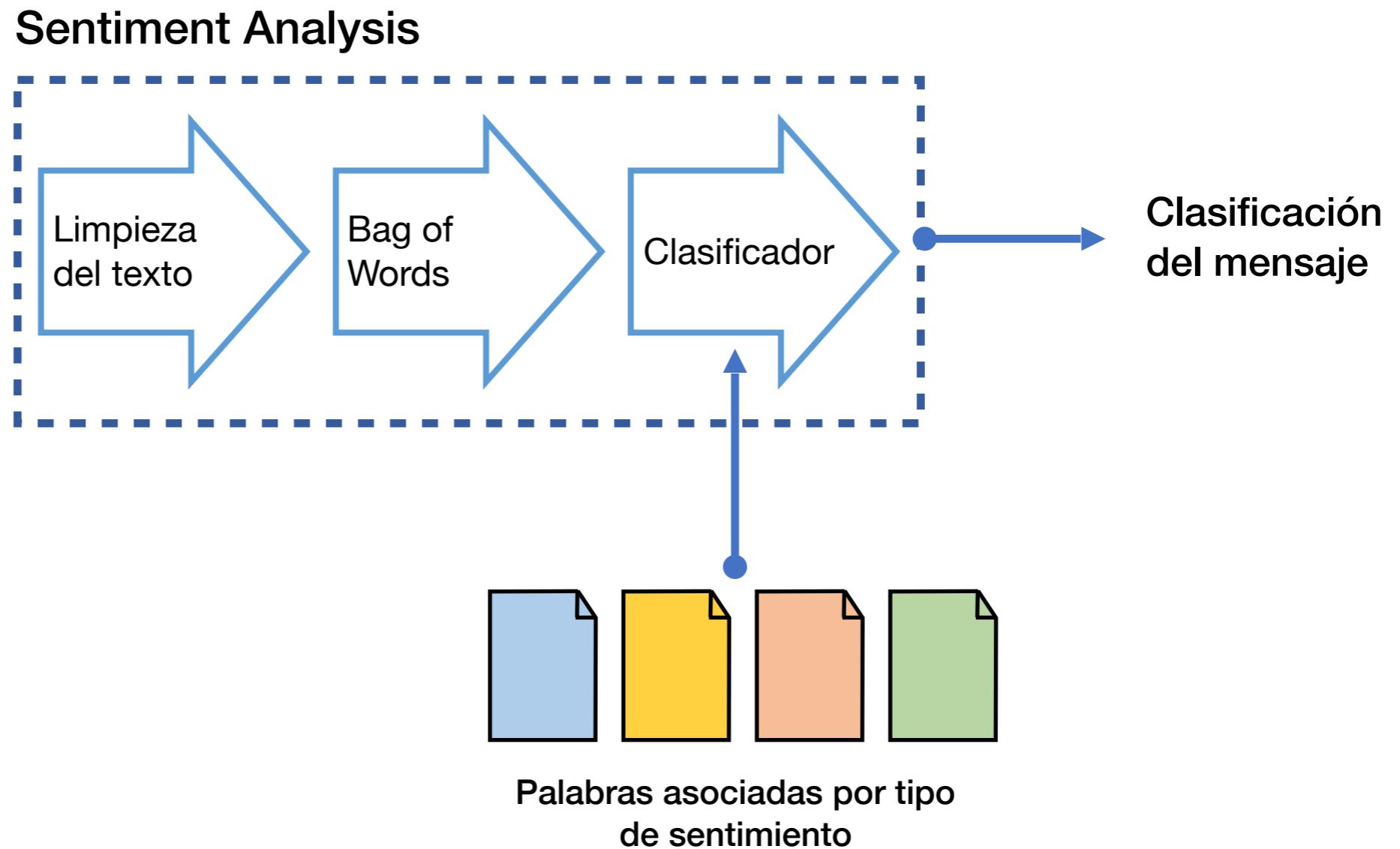
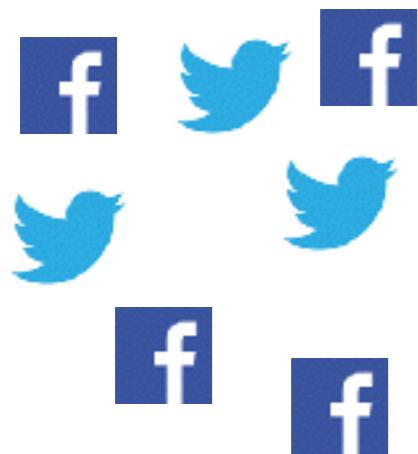
# Machine Learning & Predictive Analytics



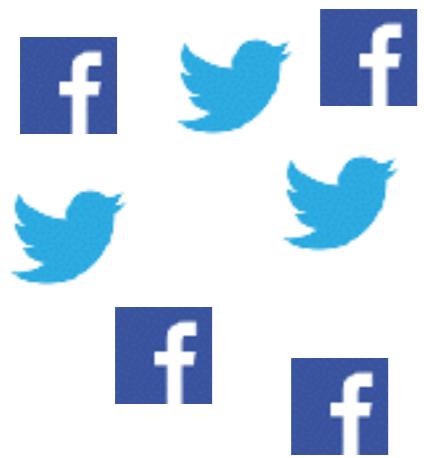
# Machine Learning & Predictive Analytics



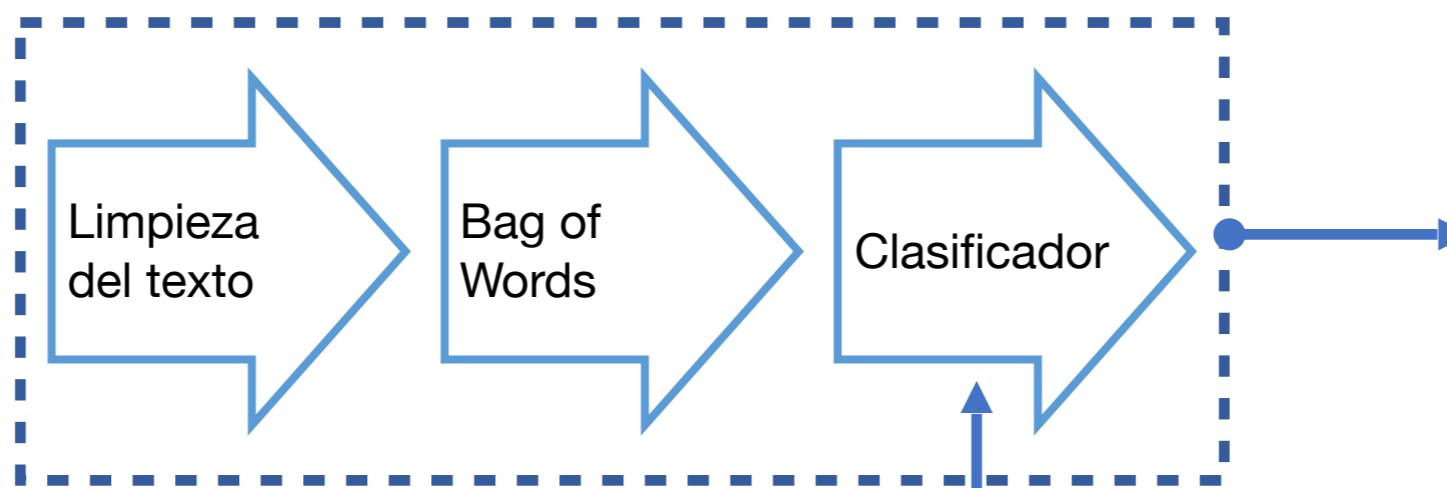
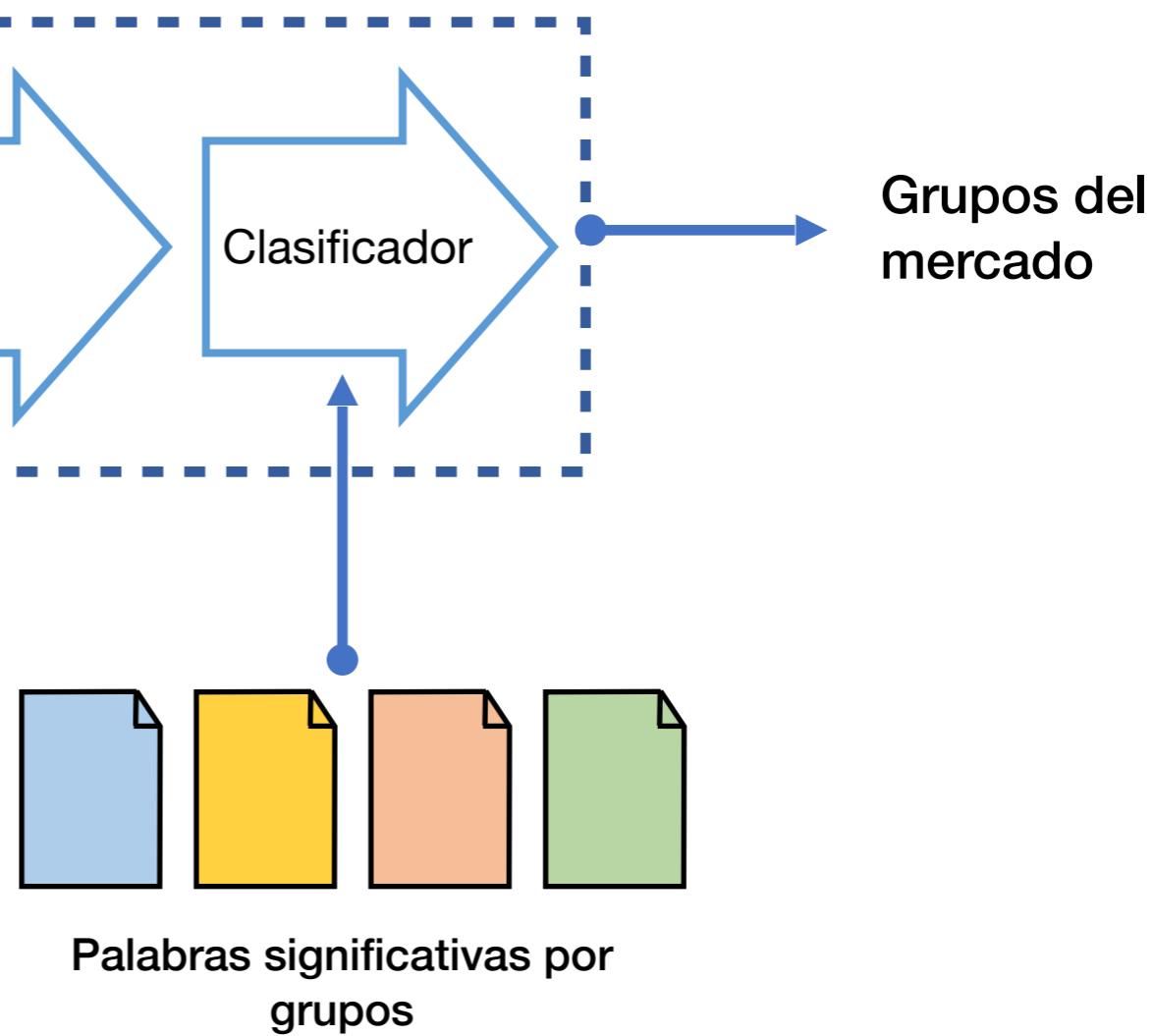
# Machine Learning & Predictive Analytics



# Machine Learning & Predictive Analytics

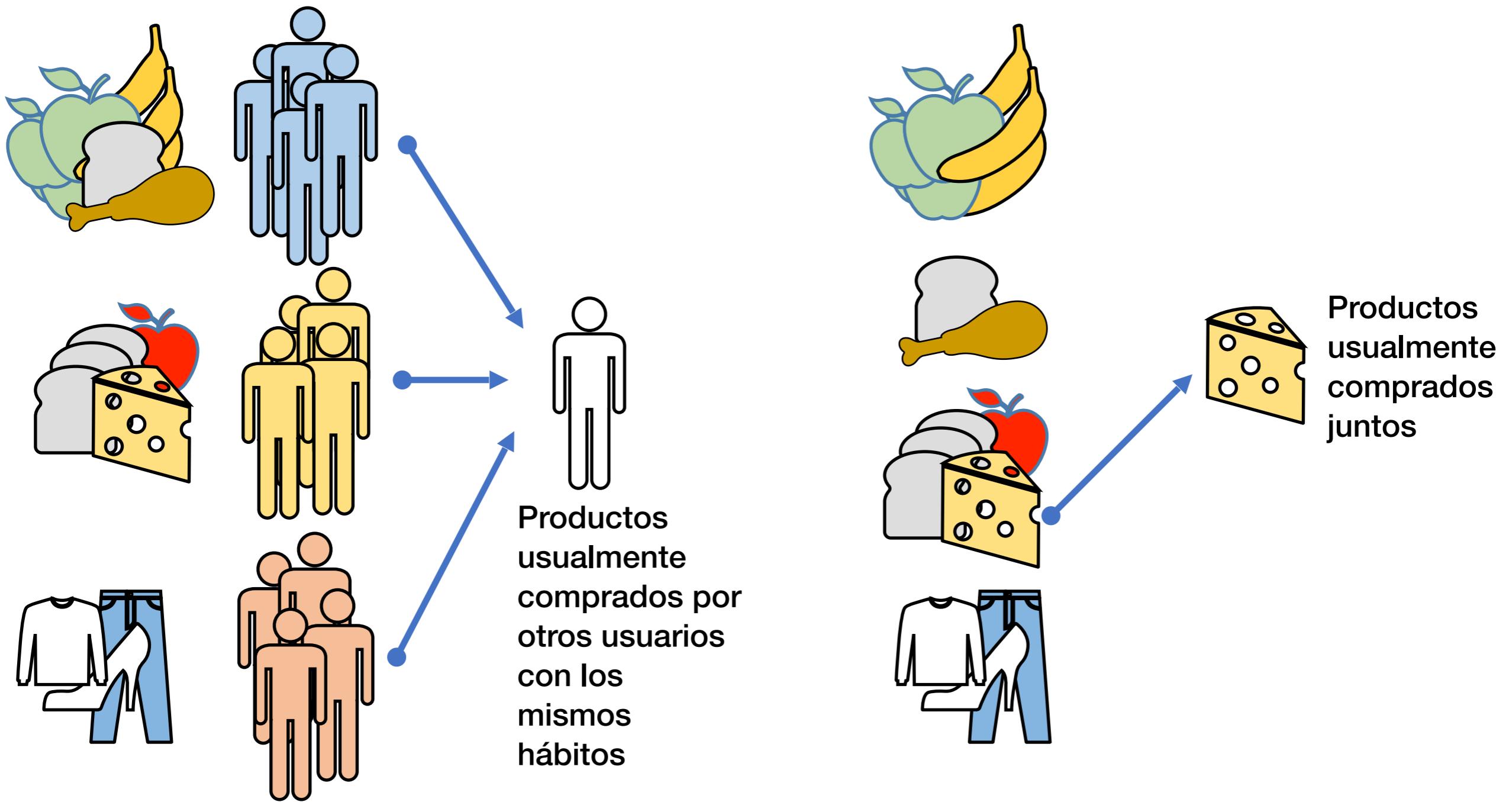


## Clasificación del mercado



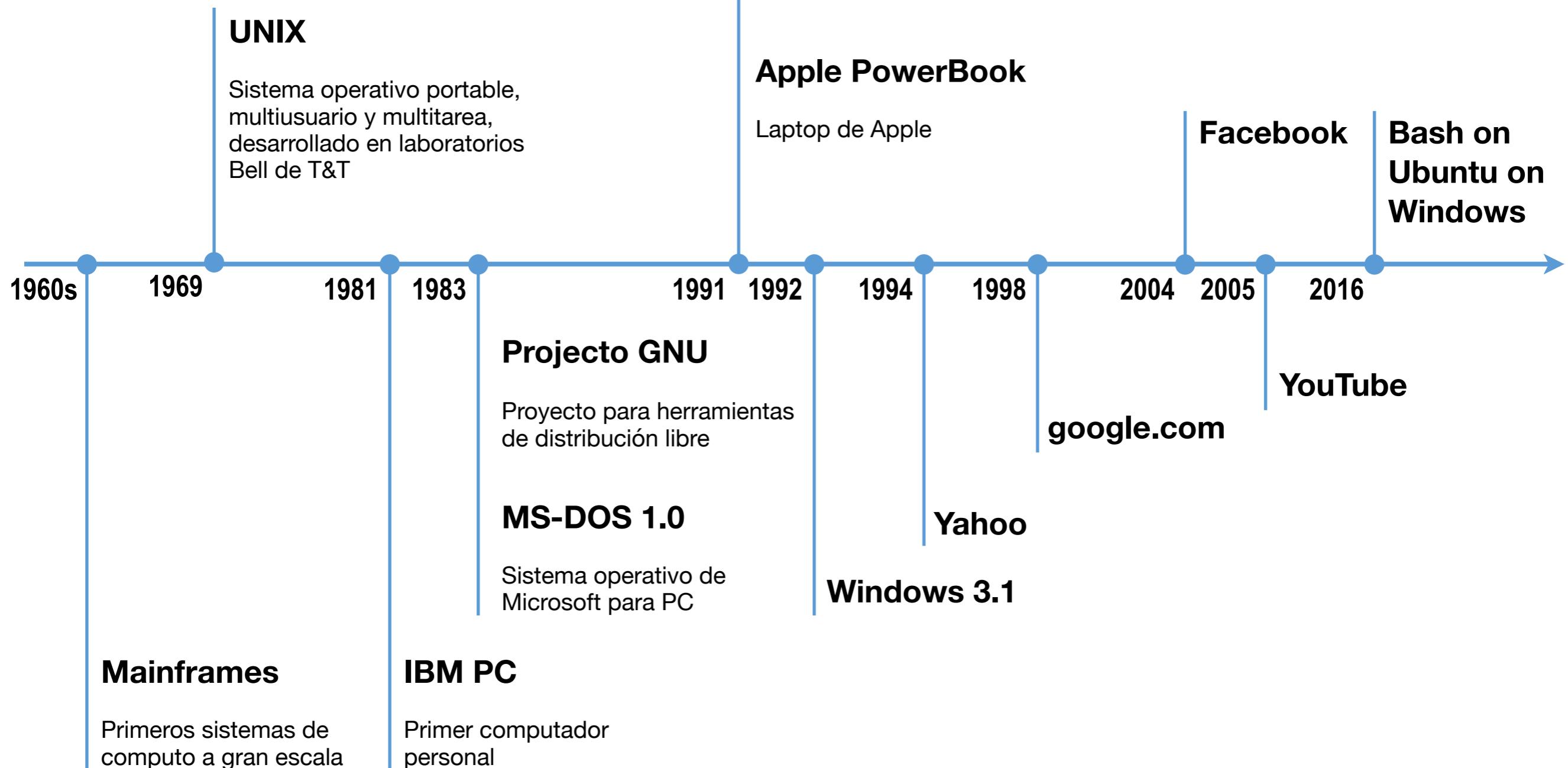
# Machine Learning & Predictive Analytics

## Association Rules & Recommender Systems



# Open Data Science & Modern Analytics

# 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

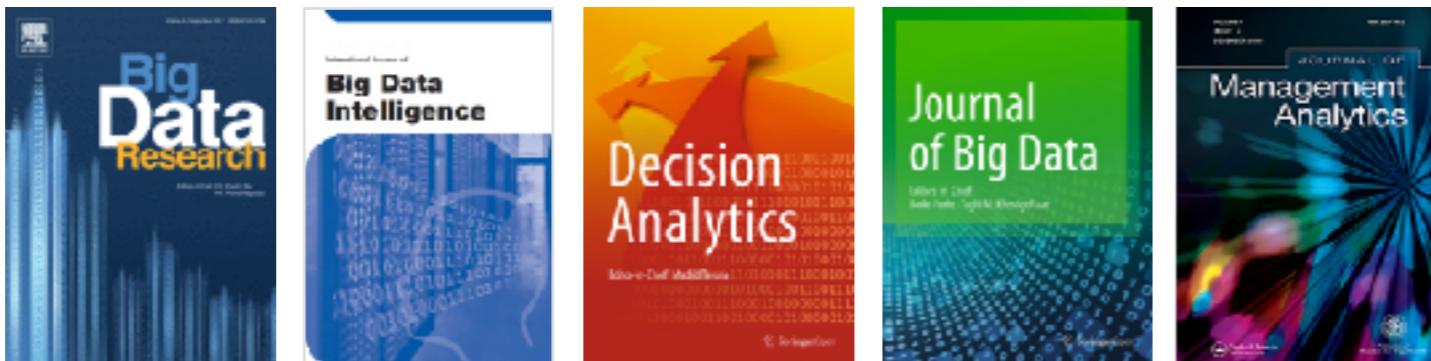
## **Data Scientist**

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

# Big Data / Data Science



## DATA SCIENCE JOURNAL

2002

2003

## Journal of Data Science



2012



2013



2014



2015



2016



# Data Science (1996)

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

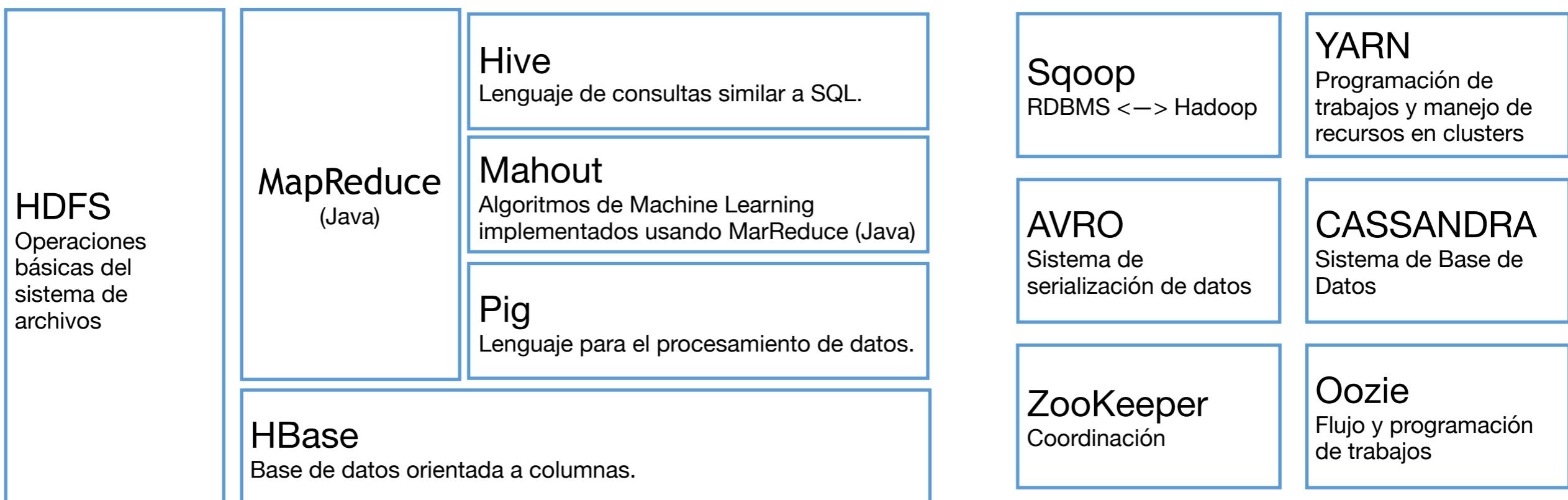
Average Base Salary

Best Job in America  
for 2016

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

# Hadoop / MapReduce (2005)

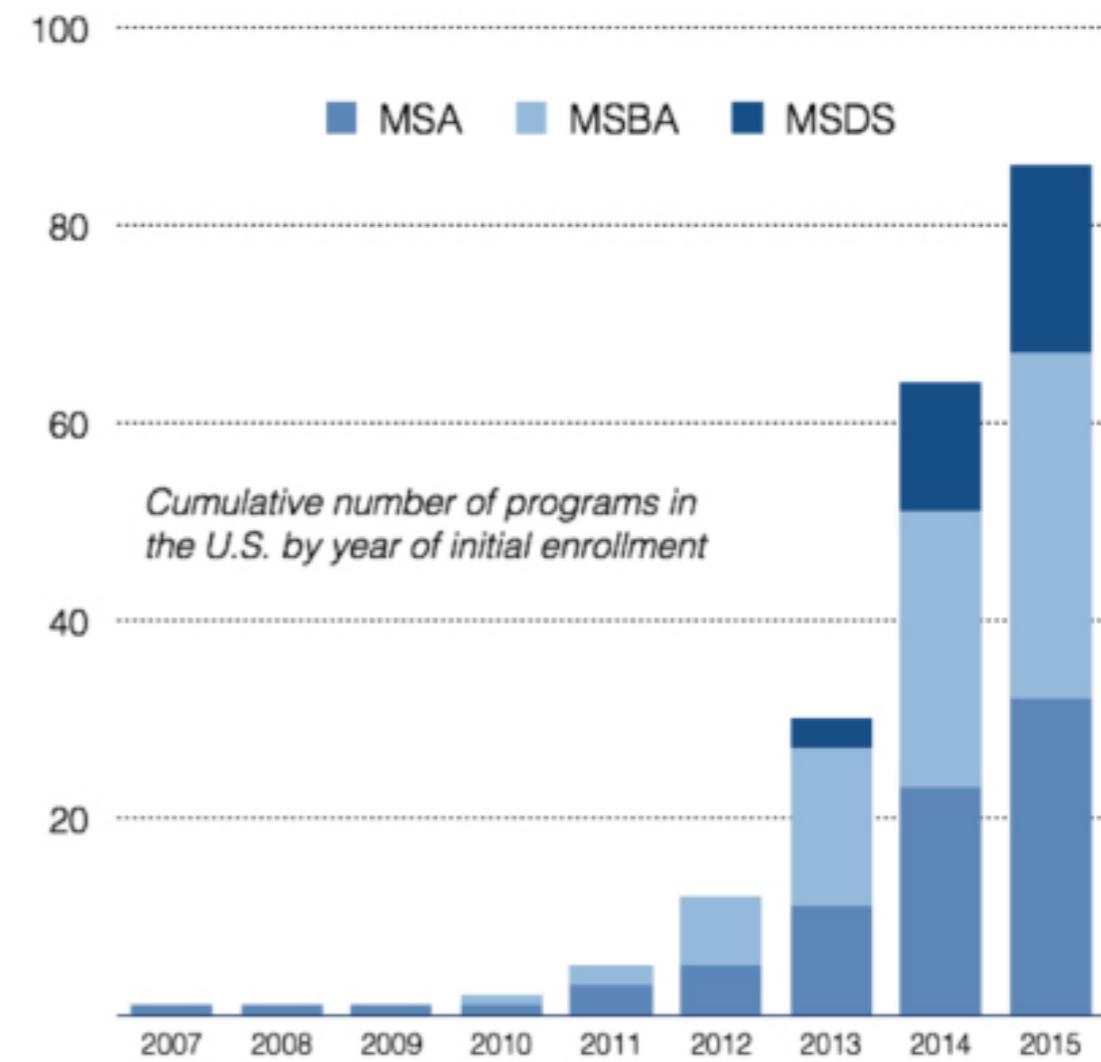
## Ecosistema Apache Hadoop



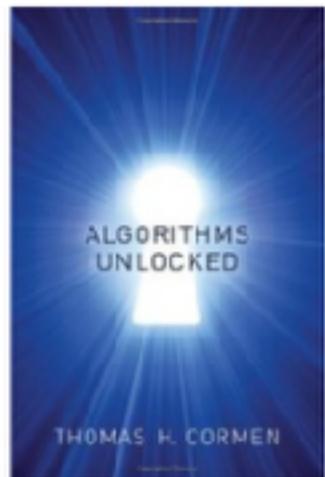
# Data Science (1996)



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



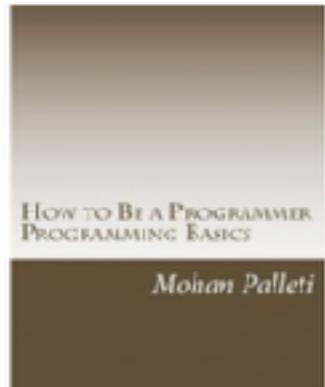
[http://analytics.ncsu.edu/?page\\_id=4184](http://analytics.ncsu.edu/?page_id=4184)



## Algorithms Unlocked

By: Thomas H. Cormen

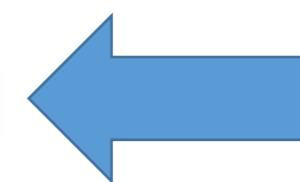
Have you ever wondered how your GPS can find the fastest way to your destination, selecting one route from seemingly countless possibilities in mere seconds? How your credit card account number is protected when you make a purchase over the Internet? The answer is algorithms. And how do...



## 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 & Modern Analytics (2009)

## 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

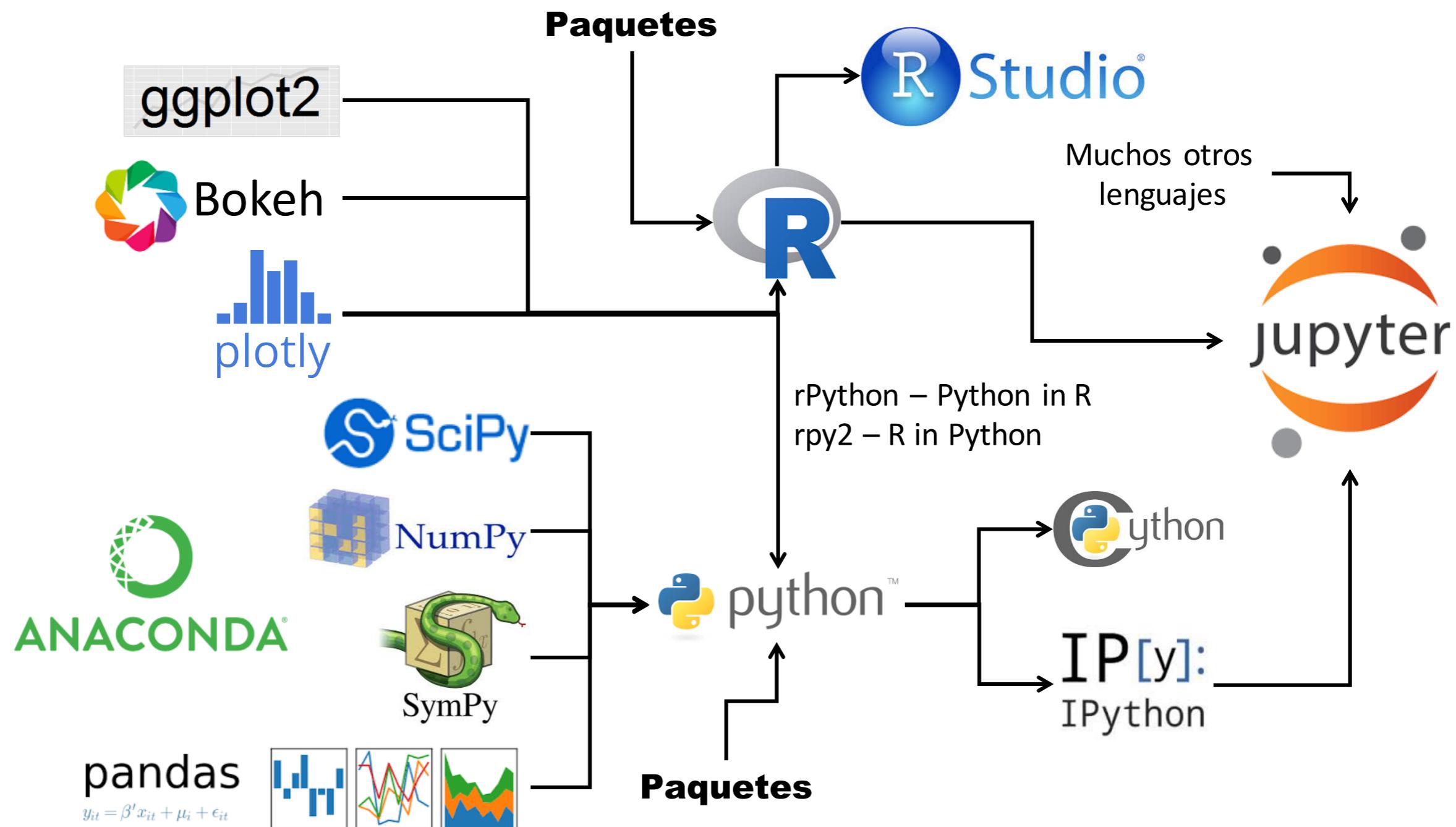
Perfil de la demanda

Respuesta de la demanda

Detección de fallos

(Dispositivos usables, ...)

# Open Data Science & Modern Analytics (2009)



# Una Introducción a la Analítica

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/an-intro-to-analytics/>

**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

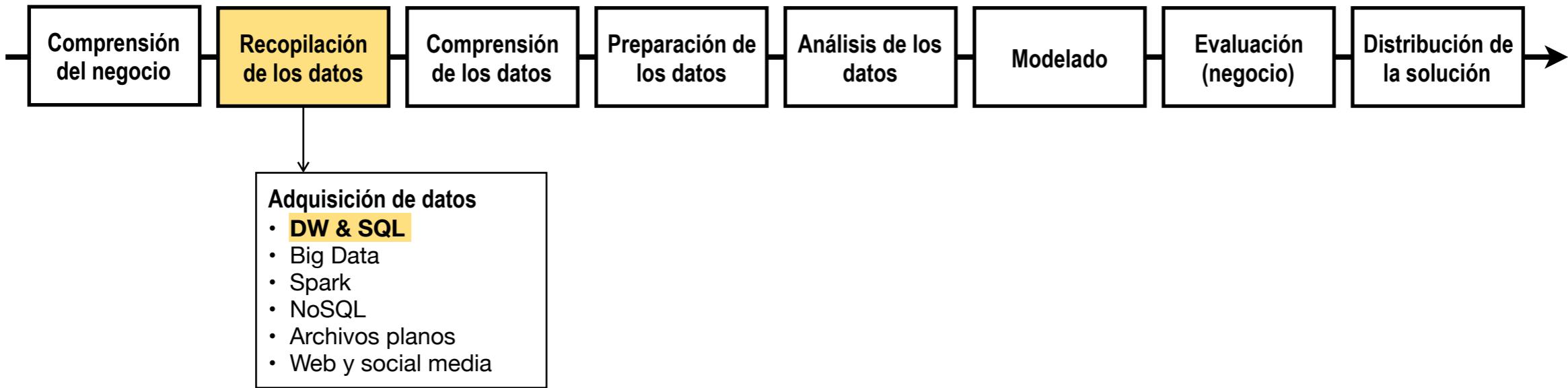
 @jdvelasquezh

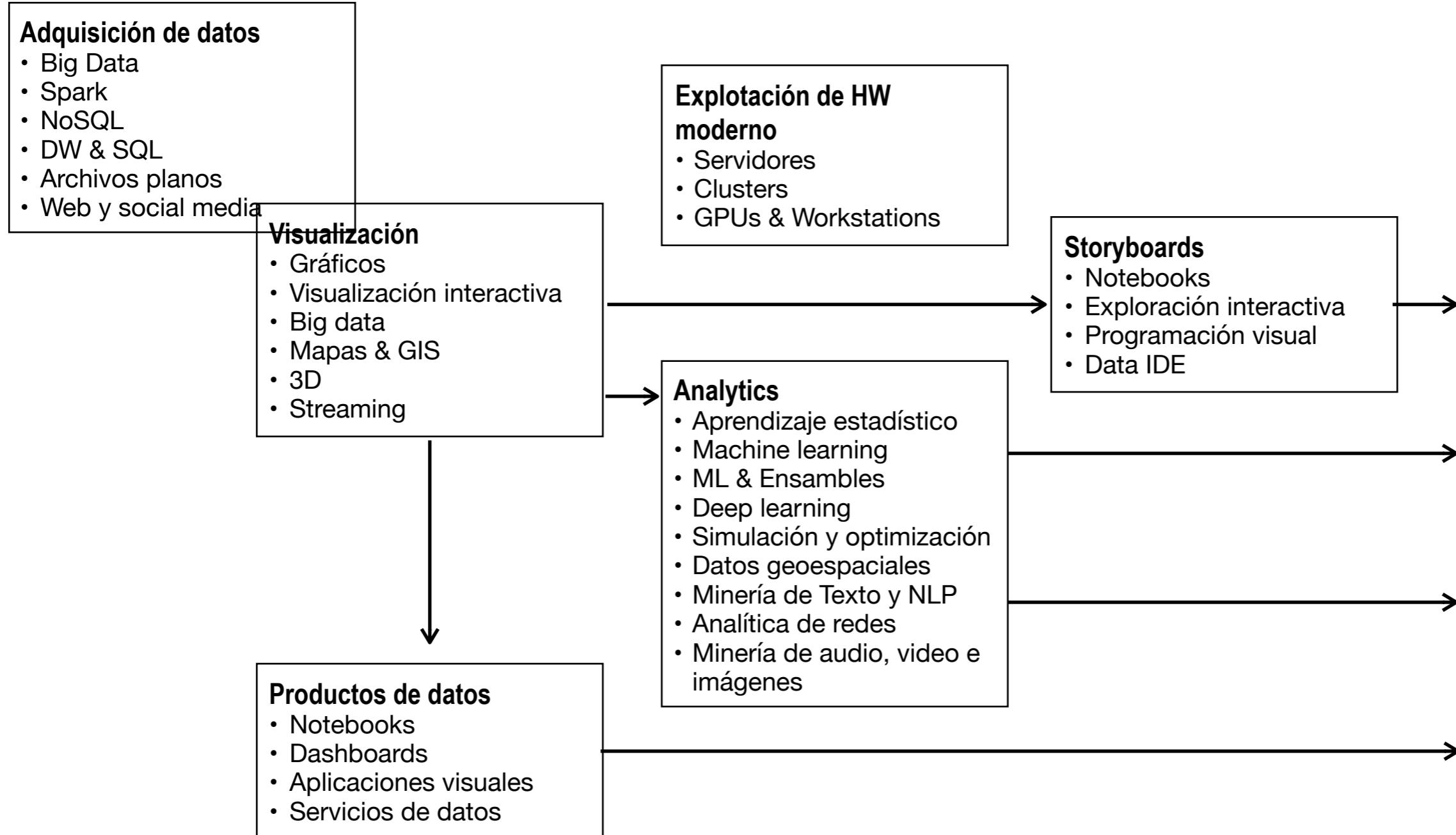
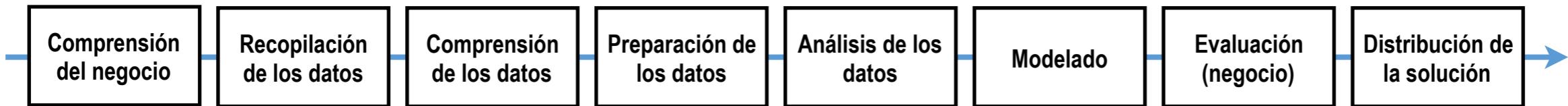
 <https://github.com/jdvelasq>

 <https://goo.gl/prkjAq>

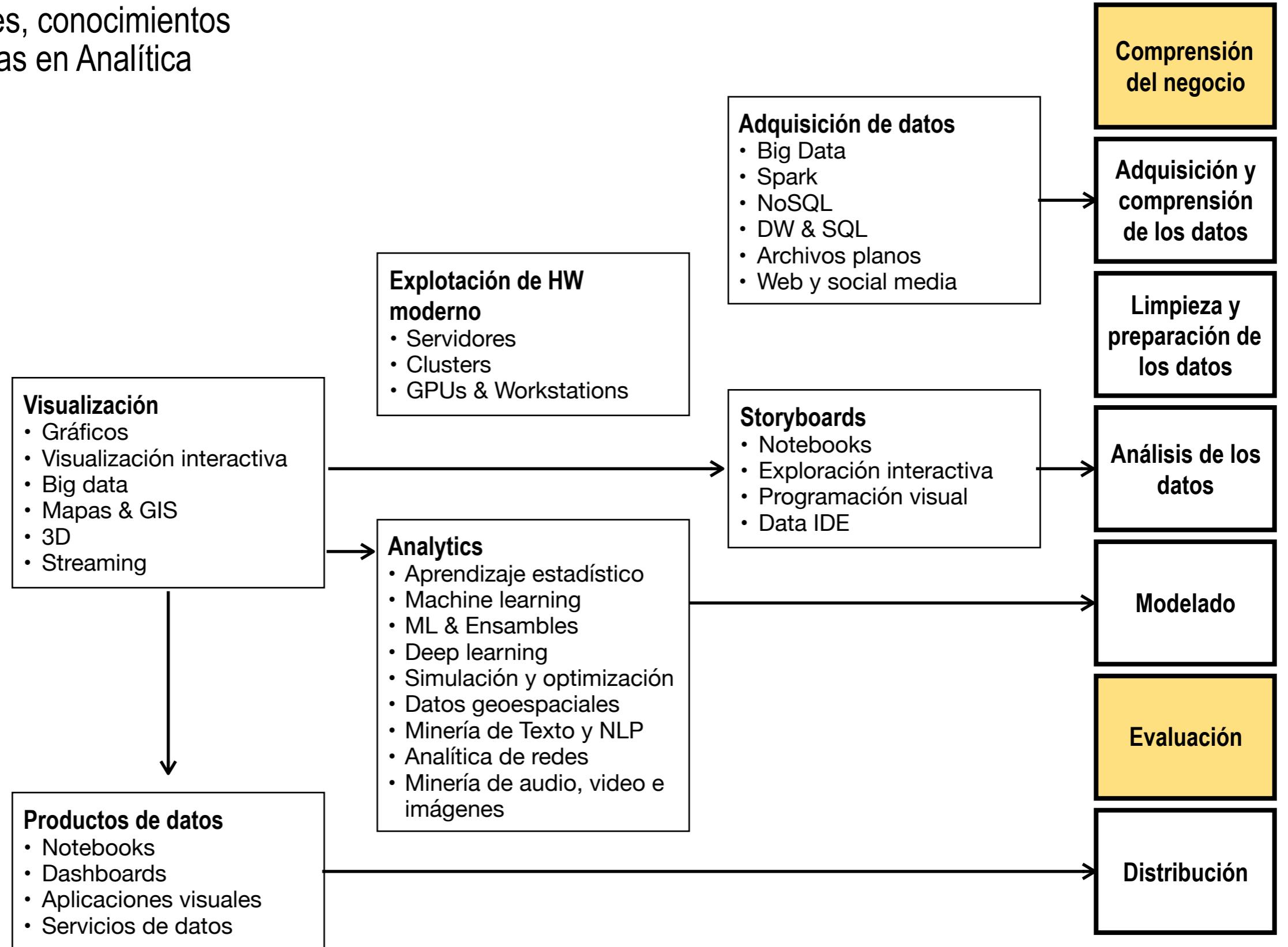
 <https://goo.gl/vXH8jy>

# Separadores

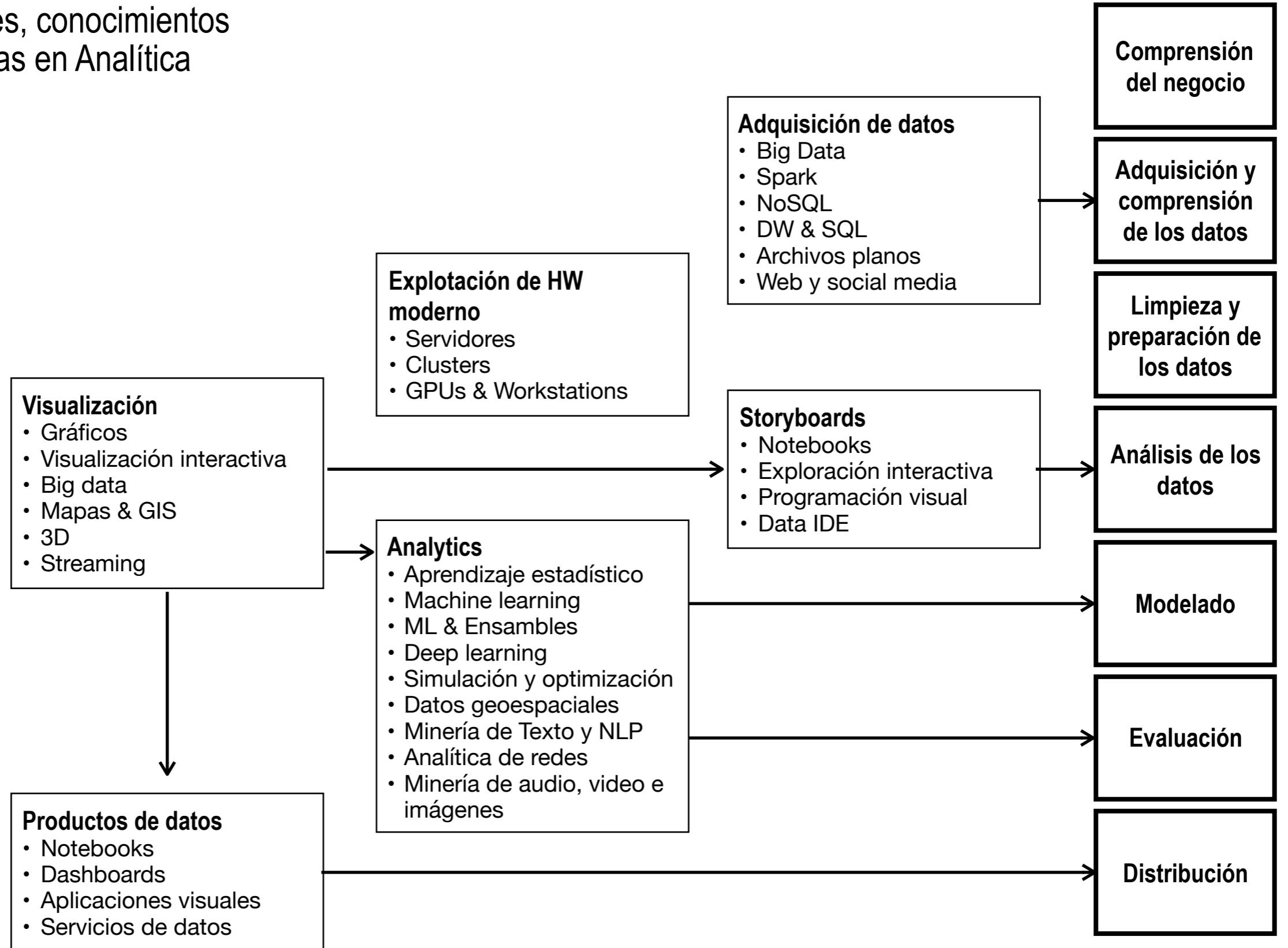


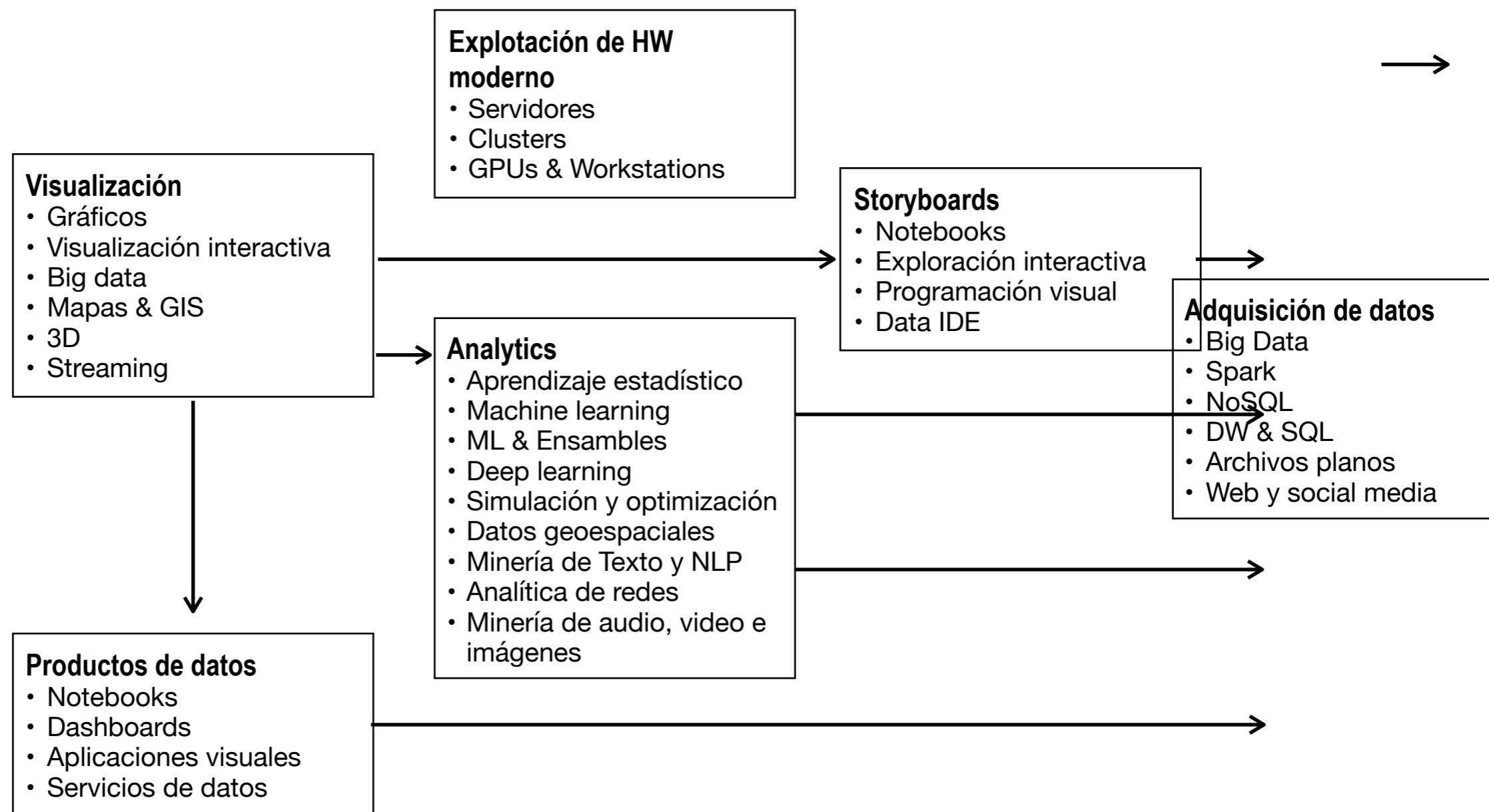
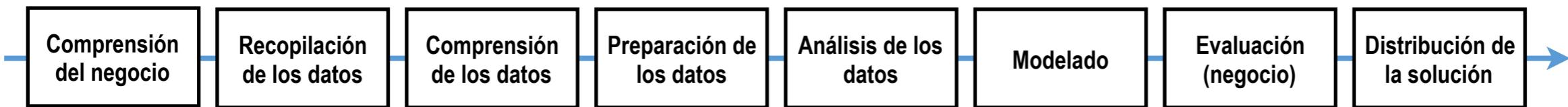


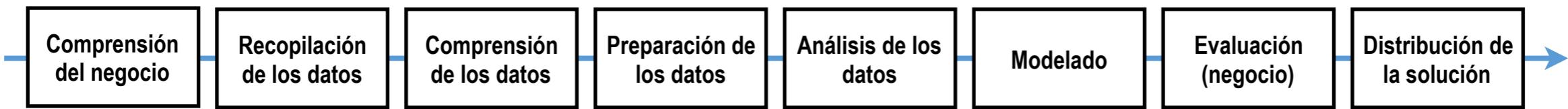
# Mapa de fases, conocimientos y metodologías en Analítica



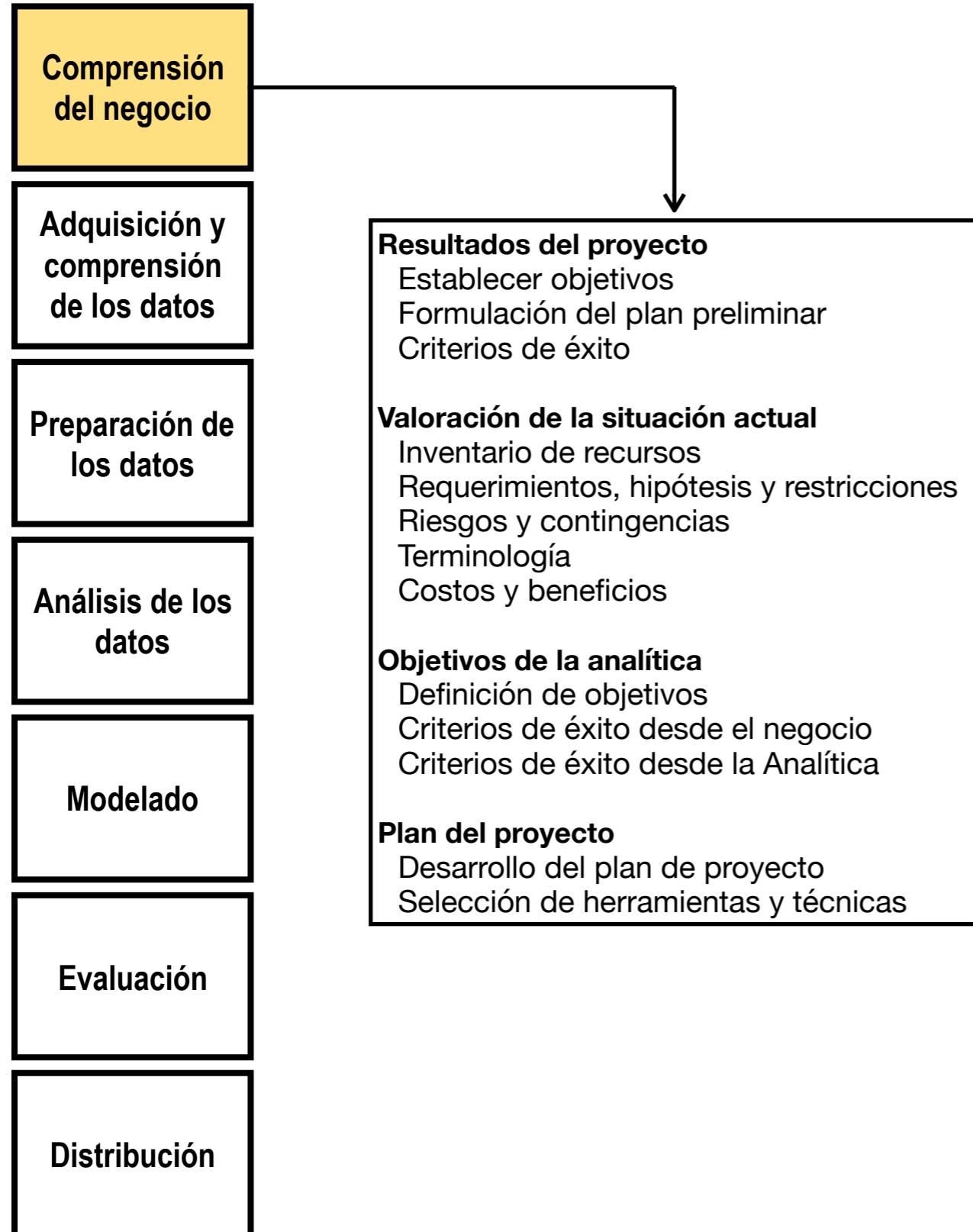
# Mapa de fases, conocimientos y metodologías en Analítica







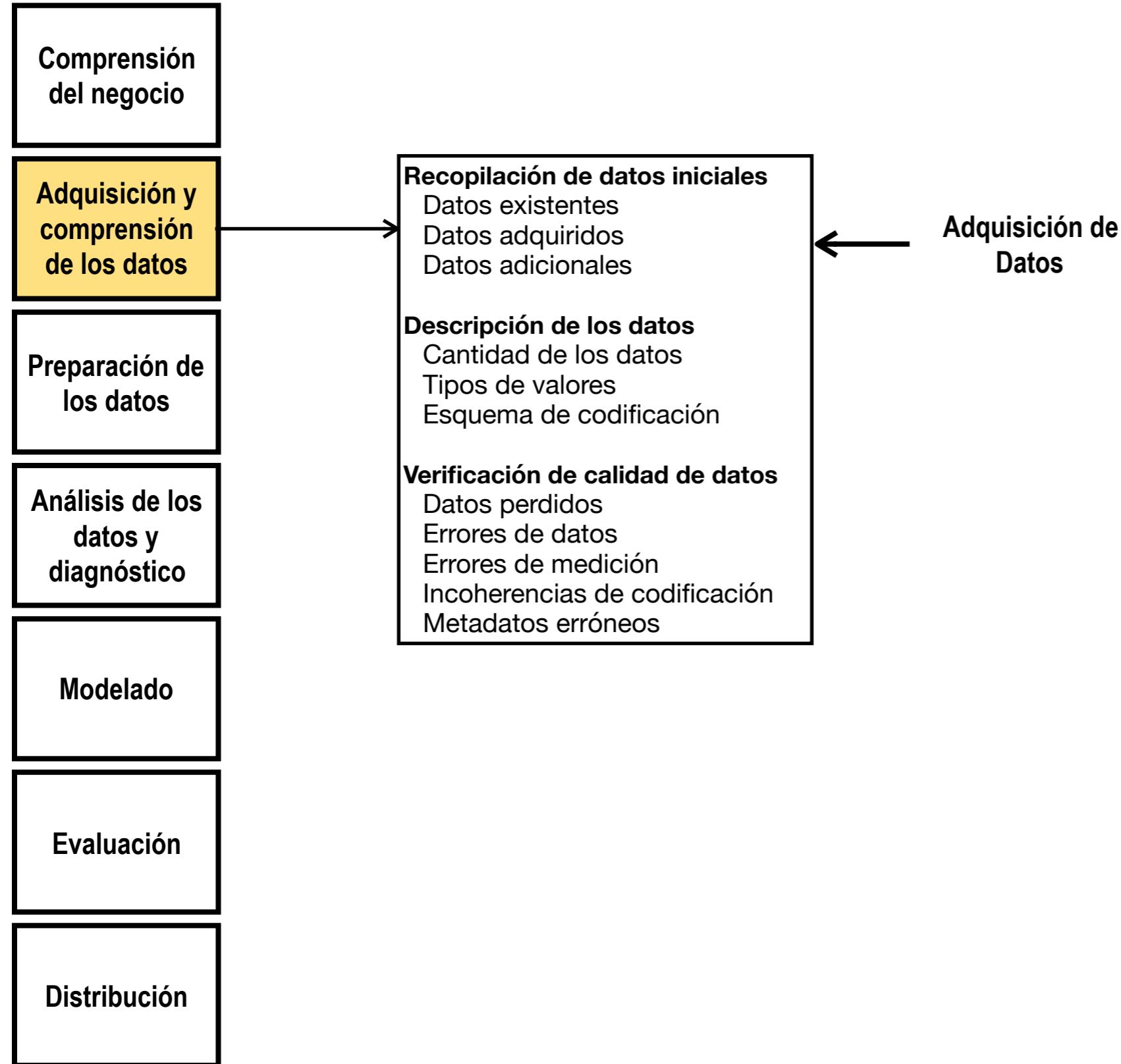
# Mapa de fases, conocimientos y metodologías en Analítica



Conocimiento  
Experto



# Mapa de fases, conocimientos y metodologías en Analítica



RDBMS &  
Datawarehouses



Archivos planos



Internet y  
Social Media



Hadoop



# Industria de la Energía Eléctrica

