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

(con ejemplos en mercados de energía)

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

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

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

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

Definiciones

1985

Data Warehousing OLAP

Almacenes de datos que garantizan la permanencia de la información

1989

Business Intelligence

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

1995

Data Mining

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

1996

Open Data

Primer uso de este término

1999

Business Intelligence 2.0

Plataformas de BI

Data Science

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

2005

Internet of Things (IoT)

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

Traditional Analytics

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

Hadoop / MapReduce

Primera implementación de un sistema Big Data

2009

Bases de Datos NoSQL

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

Open Data Science

Modern Analytics

2011

Big Data Analytics

Apache Mahout

Definiciones

1973 Demand side management

1974 Primer sistema patentado de monitoreo con sensores remotos

1981 Privatización mercado eléctrico de Chile

1985 Control de la demanda basado en precio

1990 Privatización de la generación de UK

1998 Popularización del concepto de Smart Home

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

2005 Medidores inteligentes para electricidad y gas en Países Bajos

2006 Medidores inteligente en Nueva Zelanda

2007 Se usa el término Smart Grid por primera vez

2008 Medidores inteligentes en UK

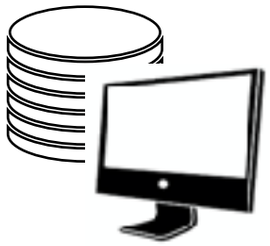
2010 Control de la demanda basado en tarifas en tiempo real

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

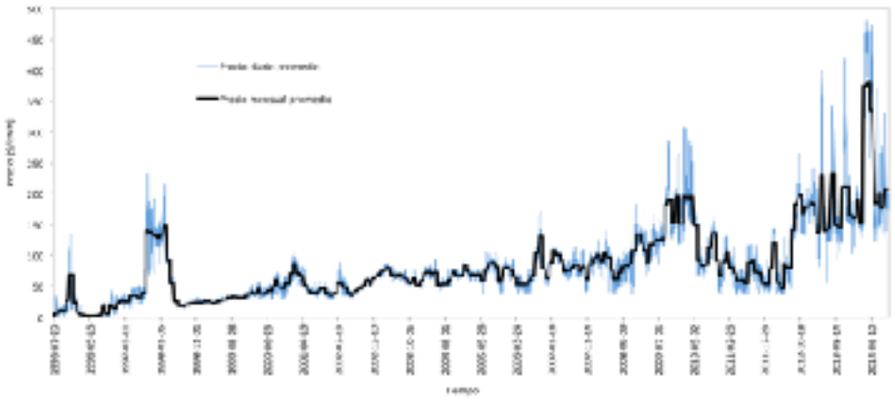
Modelos de decisión individual
Modelos de predicción del precio

Medidores inteligentes de consumo en Ontario

Definiciones



Técnicas de modelado

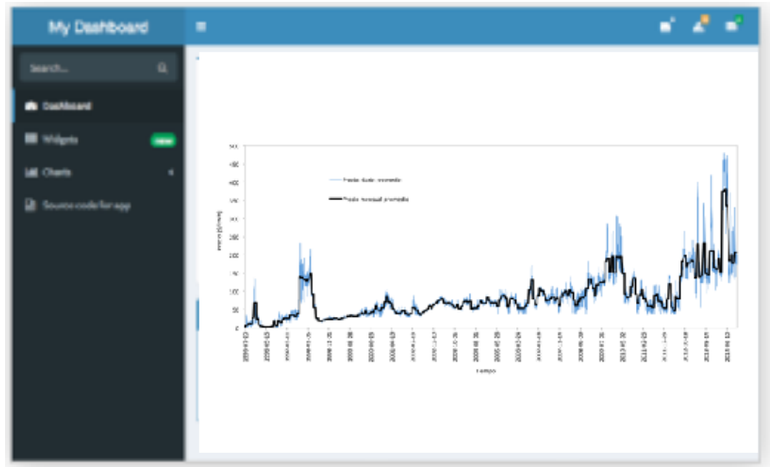


Producto de Datos

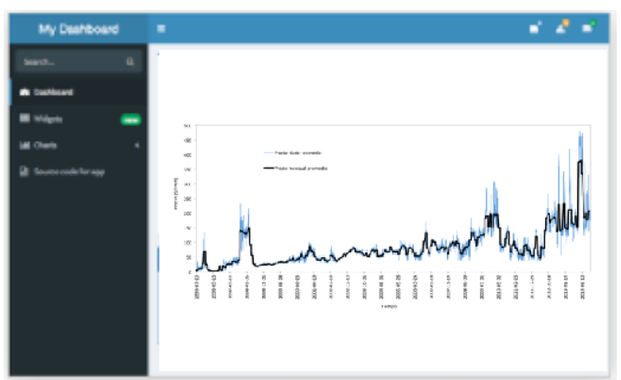
Aplicación que combina datos con algoritmos para inferencia, predicción u optimización para generar más datos e información valiosa.

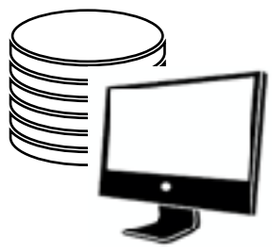
- Aprendizaje a partir de los datos.
- Auto-adaptación
- Ampliamente aplicable.

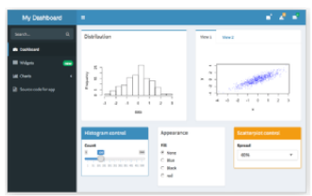
Técnicas de modelado
+ Reportes

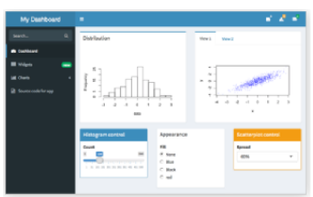


Inteligencia de Negocios









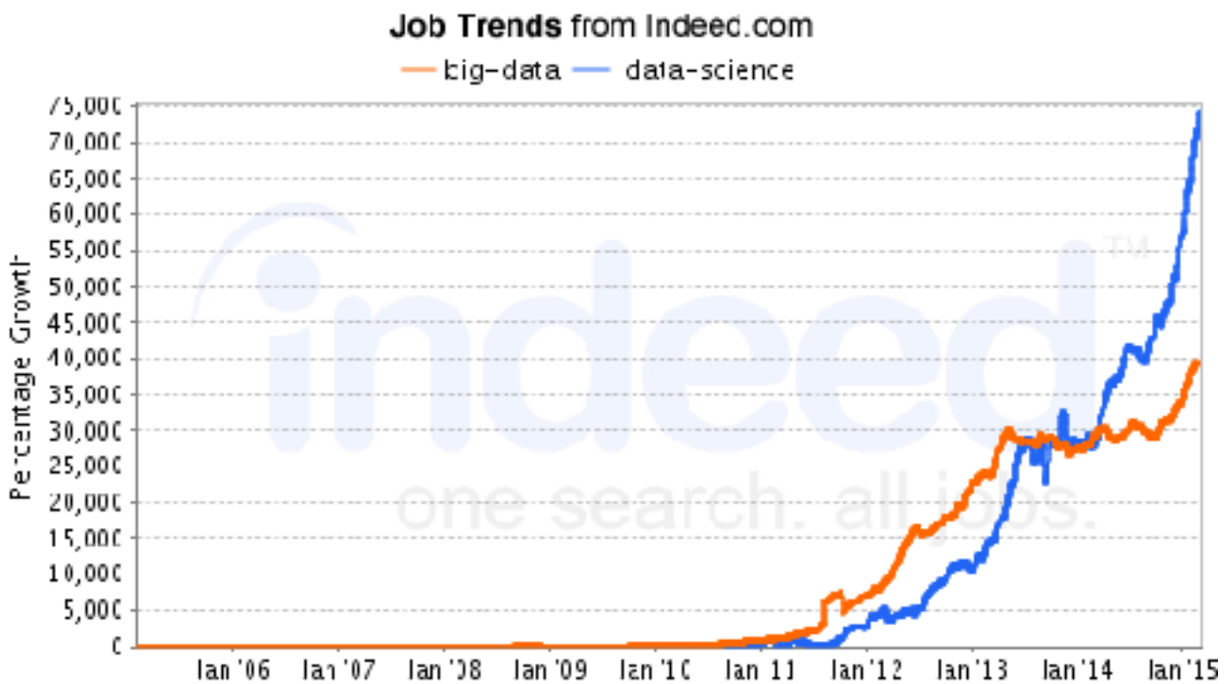
Analytics

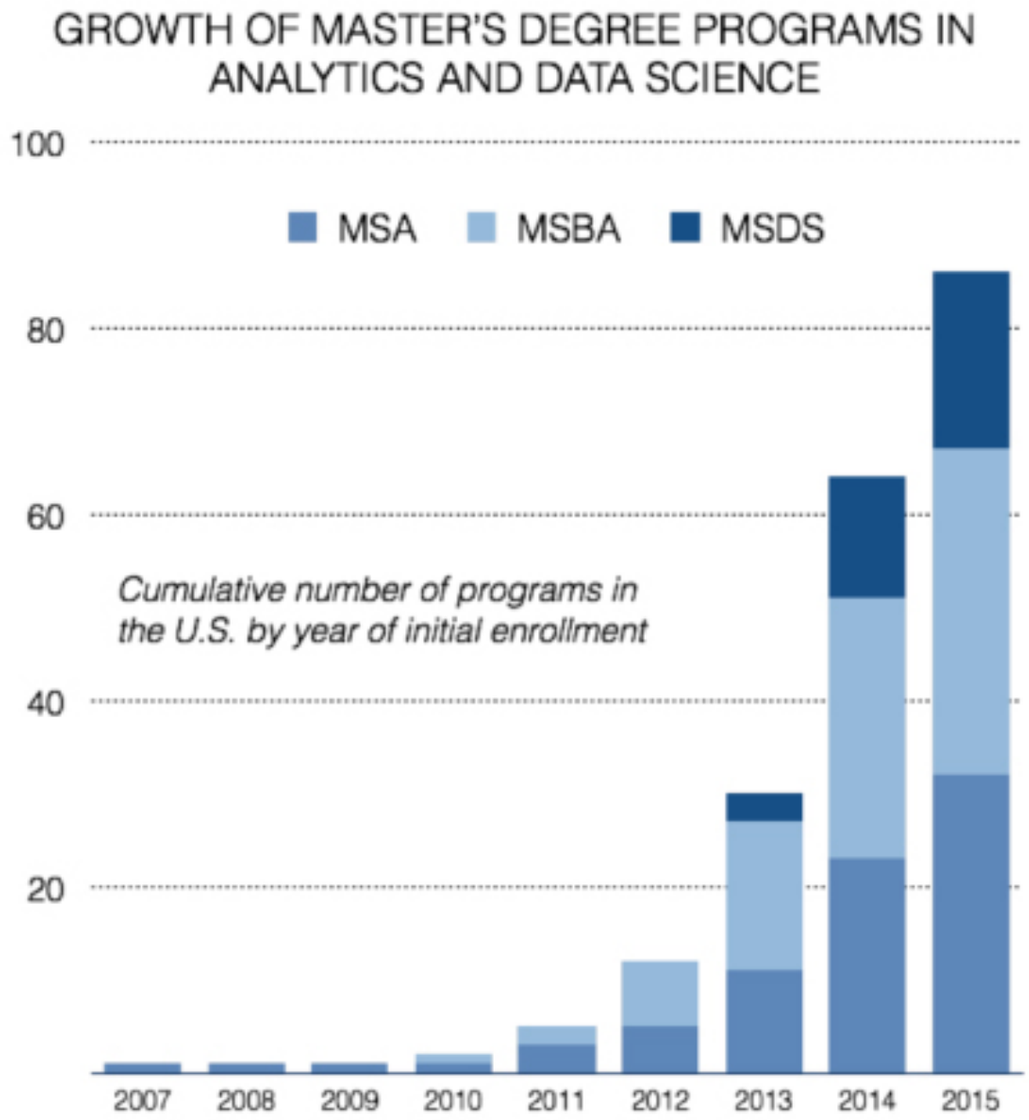
Perspectiva

> 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."

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

Evolución de empleos/educación en Big Data & Data Science





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

Data Science Analytics

Big Data

1941 - Explosión de la información
1966 - Primeros sistemas de cómputo centralizado
1970 - Bases de datos relacionales

Concise survey of computer methods - **1974**

1976 - Lenguaje SQL (Oracle)

International Association for Statistical Computing - **1977**

1980 - “Data expands to fill the space available”
1985 - Data warehousing

1st Knowledge discovery in Databases (KDD) workshop - **1989**

1989 - Inteligencia de negocios (concepto)
1992 - Primer sistema para reporte de bases de datos (Crystal Report)
1995 - Explosión de la Web

Se usa por primera vez el término Data Science - **1996**

1996 - Primeras plataformas de inteligencia de negocios

Journal Data Mining and Knowledge Discovery - **1997**

1997 - 1r uso del término Big Data en un articulo
1998 - Crisis de la BI por el volumen de datos
1999 - Internet of Things (IoT), Predictive Analytics
2000 - Se reconoce la explosión de datos potencialmente relevantes

Data Science como una disciplina independiente - **2001**

2001 - Software as a Service (SaaS), Data Volume, Velocity, and Variety

Data Science Journal - **2002**

2002 - Web Services

Journal of Data Science - **2003**

2004 - Google’s MapReduce

Se usa por primera vez el término Analytics - **2005**

2005 - Yahoo desarrolla Hadoop basado en MapReduce de Google

Se reconoce el término Data Scientist - **2008**

2008 - Google procesa 20 Petabytes en un día.

“… sexy job in the next ten years will be statisticians” - **2009**

2009 - Aparece el término NoSQL.
2010 - Cloud
2011 - Aparece el término Data Lakes

Data Scientist: The Sexiest Job of the 21st Century - **2012**

2013 - Cloud
2014 - Explosión de IoT

Deep Learning y explosion de Machine Learning - **2015**

2015 - Smart cities
2016 - Las organizaciones implementan Big Data de forma masiva

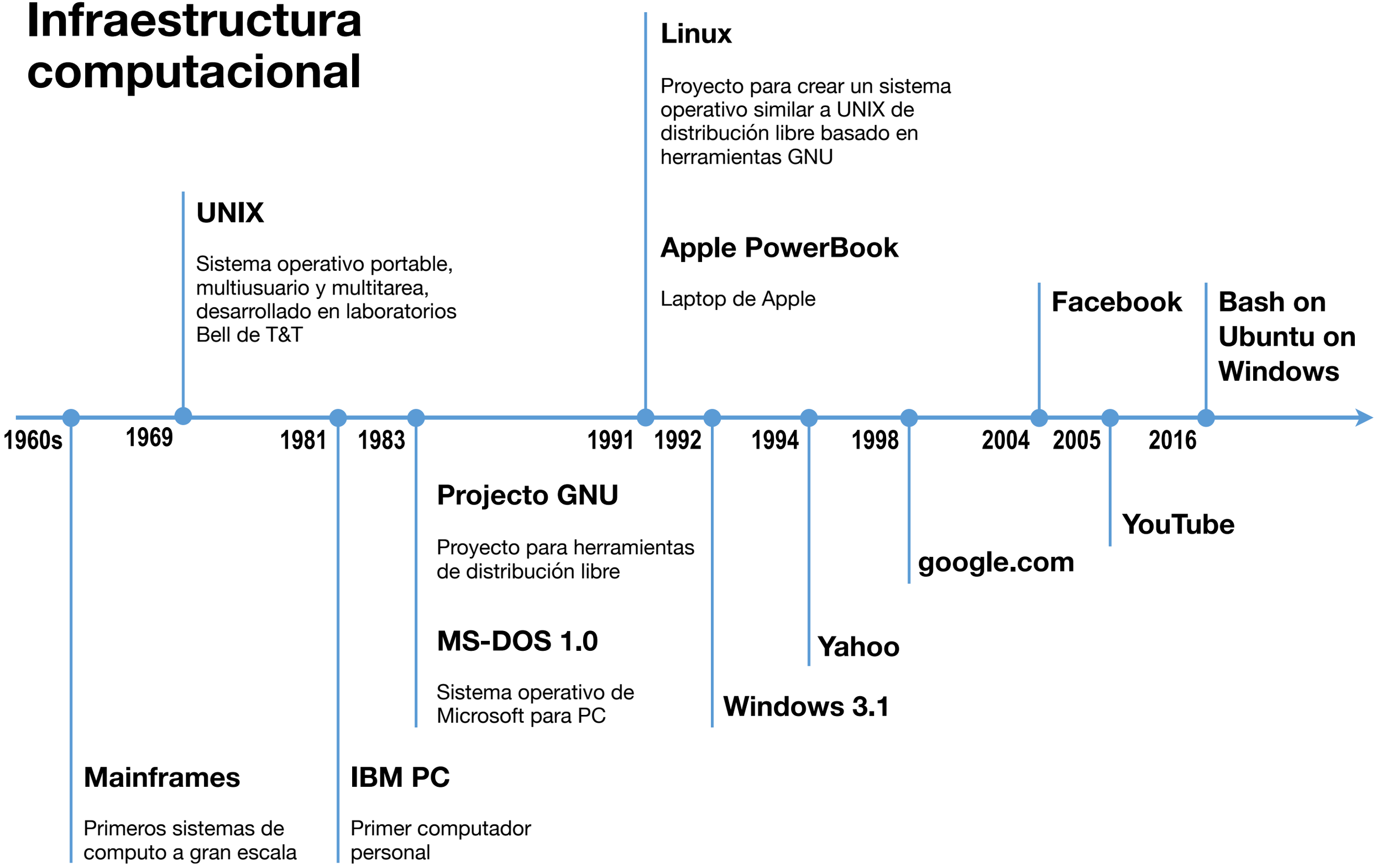
A Very Short History Of Data Science (Forbes)
https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#5f40380955cf

Elementos constitutivos

- Infraestructura computacional
- Datos
- Programas (Algoritmos, técnicas y metodologías)

Infraestructura computacional

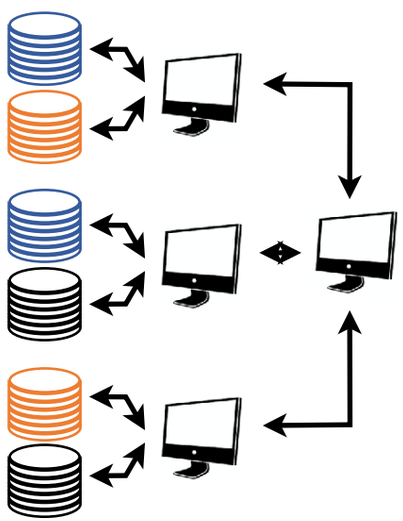
Infraestructura computacional



Infraestructura computacional

Computación local

Servidores + red + clientes



Cloud computing / utility computing

Servidores y almacenamiento en la nube + internet + clientes locales

Software as a Service (SaaS)

Software almacenado en máquinas suministradas por un tercero.

Aplicaciones accesadas vía un cliente o la Web.

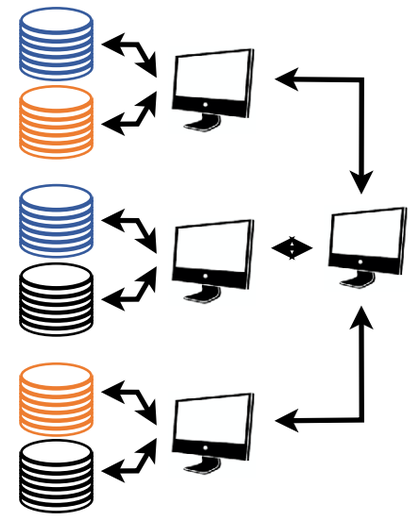
Orientado a aplicaciones de usuario final.

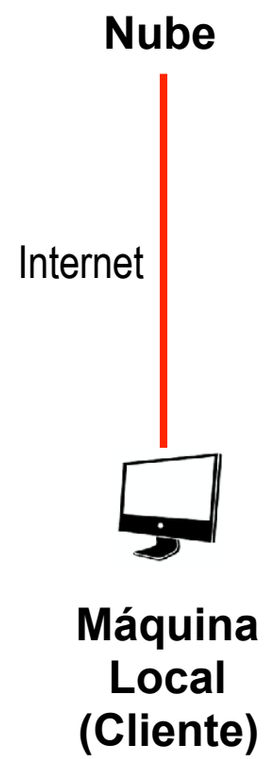
Platform as a Service (PaaS)

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

Infrastructure as a Service (IaaS)

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



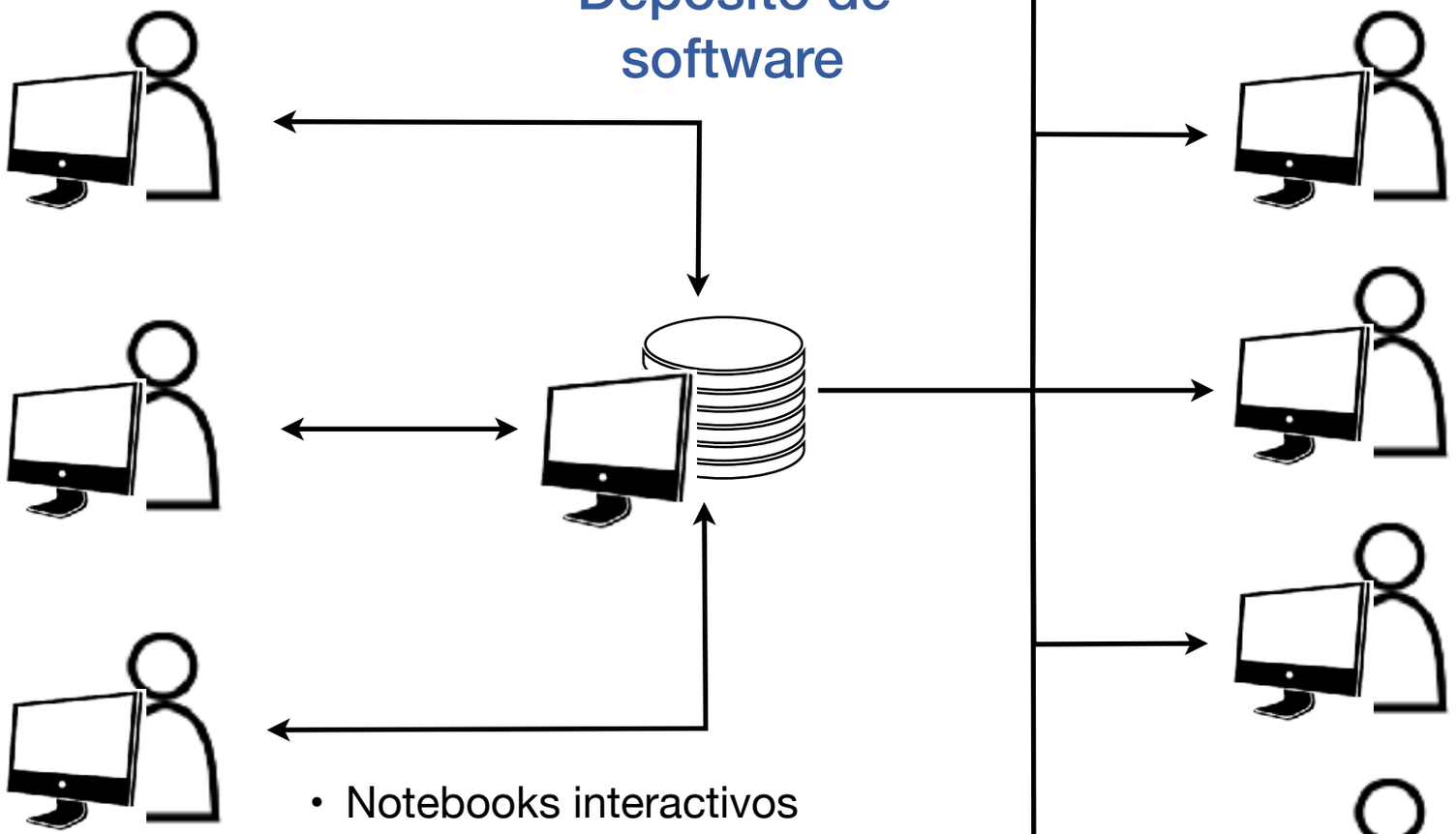


DevOps

Integración Continua

Entrega Continua

Depósito de software



Equipos de desarrollo de software & Científicos de datos

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

Analistas de datos & Científicos de datos

Datos

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 columnar
- 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, …)

Ejemplo de un reporte del IDEAM (Fuentes de datos)

I D E A M - INSTITUTO DE HIDROLOGIA, METEOROLOGIA Y ESTUDIOS AMBIENTALES

SISTEMA DE INFORMACION HIDROMETEOROLOGICA

EVALUACION HORARIA DE LA DIRECCION (D) Y VELOCIDAD (V) EM M/S DEL VIENTO EN SUPERFICIE

-METEORO 3- MAR 1987

PROCESO Jul 4-2017 *

ESTACION : 23125060 ALBANIA

LATITUD 0545 N	DEPARTAMENTO SANTANDER	TIPO EST CO	SUBZONA HIDR. QDA CANUTILLO
LONGITUD 7354 W		ENTIDAD 1 IDEAM	ZONA HIDROGR. RIO MIRA
ELEVACION 1690 m.s.n.m	MUNICIPIO ALBANIA	REGIONAL 8 SANTANDERES-ARA	AREA HIDROGR. M/LENA-CAUCA

C U A D R O NO 1 (1 PARTE)

- HORA	00-01		01-02		02-03		03-04		04-05		05-06		06-07		07-08		08-09		09-10		10-11		11-12		12-13		13-14	
DIA -	D	V	D	V	D	V	D	V	D	V	D	V	D	V	D	V	D	V	D	V	D	V	D	V	D	V	D	V
1	NE	1.6	NE	1.0	NE	2.3	NE	1.3	NE	1.3	NE	1.6	NE	1.6	NE	1.0	SW	1.2	SW	1.6	SW	1.3	SW	2.0	SW	1.5	NW	1.2
2	NE	1.3	NE	1.3	SW	1.6	NE	1.6	NE	1.5	NE	2.2	SE	1.2	SE	.6	NW	1.0	SW	1.5	SW	1.8	SW	2.0	SW	2.9	SW	2.3
3	NE	.9	NE	.6	NE	.9	NE	1.2	NE	1.3	NE	1.0	NE	1.2	SW	.9	SW	1.0	SW	1.8	SW	2.2	SW	2.0	SW	2.0	SW	2.6
4	SE	1.0	SE	.6	NE	.9	NE	.9	SE	1.0	NE	.9	NE	.9	NE	.6	SW	.9	SW	1.0	SW	1.5	SW	1.3	SW	1.6	SW	2.0
5	NW	.6	NW	.6	NE	1.3	NE	1.0	SE	.6	SW	.6	NE	.6	SW	1.0	SW	1.0	SW	1.2	SW	1.2	SW	2.2	SW	2.2	SW	2.6
6	NE	1.0	NE	1.3	NE	1.0	NE	1.0	NE	.9	NE	.9	NE	.6	SW	.6	SW	1.2	SW	1.2	SW	2.0	SW	2.0	SW	2.0	SW	1.6
7	NE	.6	SE	.5	VR	.6	NE	.5	NE	.6	SW	.5	SE	.6	SW	.9	SW	1.2	SW	1.5	SW	1.2	SW	2.2	SW	1.2	SW	1.2
8	NE	1.3	NE	1.0	NE	1.2	NE	1.5	NE	1.2	SE	.6	NE	.9	NW	.6	SW	1.3	SW	1.8	NW	.9	SW	1.6	SW	2.4	SW	2.4
9	NE	1.5	NE	1.5	NE	1.6	NE	1.3	NE	.6	NE	.6	SW	.6	SE	.5	SE	.6	SW	.6	SE	.9	SW	1.6	SW	1.5	SW	1.3
10	SE	.6	SW	.6	SE	.6	NE	.6	NE	.9	NE	1.2	NE	.6	SW	.6	SW	1.3	SW	1.3	SW	1.5	SW	2.0	SW	1.0	SW	2.0
11	NE	1.8	NE	1.5	NE	1.5	NE	1.0	NE	.9	NE	.9	E	.6	NW	.6	SW	1.0	SW	1.6	NW	1.3	SW	2.0	SW	1.8	SW	2.6
12	NE	1.2	NE	1.3	NE	1.5	NE	1.6	NE	1.5	NE	1.2	NE	1.3	SW	.9	SW	1.5	SW	1.8	SW	1.8	SW	2.6	SW	1.6	SW	2.2
13	NE	.9	NE	.9	NE	1.2	NE	1.0	NE	.6	NE	.6	VR	.6	SW	.6	W	1.0	SW	1.5	SW	1.0	SW	1.0	SW	.6	SW	.6
14	NE	.9	NE	.9	NE	.9	NE	.9	NE	.9	NE	.9	NE	.6	SW	.9	SW	1.0	SW	1.5	SW	1.5	SW	2.0	SW	2.2	SW	2.0
15	SE	.5	NE	.5	NE	.6	SE	.6	NE	1.0	NE	.9	NE	.5	SW	.6	SW	1.0	SW	1.0	SW	.9	SW	1.2	SE	1.2	SW	1.0
16	NE	.6	NE	.6	NE	.6	NE	.6	NE	.6	E	.5	NE	.6	NW	.5	SW	.6	SW	.9	SW	1.8	SW	.9	NW	.9	SW	2.0
17	NE	1.3	NE	1.3	NE	1.5	NE	1.0	NE	1.6	NE	1.5	NE	.6	SW	.9	SW	1.2	SW	1.6	SW	2.0	SW	2.6	SW	2.6	SW	2.0
18	NE	1.5	NE	2.0	NE	1.2	NE	1.5	NE	2.2	NE	1.3	NE	.9	SW	.6	SW	1.3	SW	2.0	SW	1.5	SW	2.2	SW	2.4	SW	2.3
19	NE	.9	NE	1.2	NE	.9	NE	1.0	NE	1.0	NE	1.0	NE	1.2	SW	.9	SW	1.5	SW	1.3	SW	1.8	SW	2.0	SW	2.0	SW	2.0
20	NE	1.8	NE	1.6	NE	1.6	NE	2.2	NE	2.3	NE	2.0	NE	1.8	NE	.6	SW	.9	SW	1.2	SW	1.5	SW	1.5	SW	2.4	SW	2.3
21	NE	1.5	NE	1.5	NE	2.2	NE	2.0	NE	1.5	NE	1.5	NE	.6	SW	1.3	SW	1.5	SW	.9	SW	1.2	SW	1.2	SW	.9	SW	1.6

Ejemplo de un reporte de XM

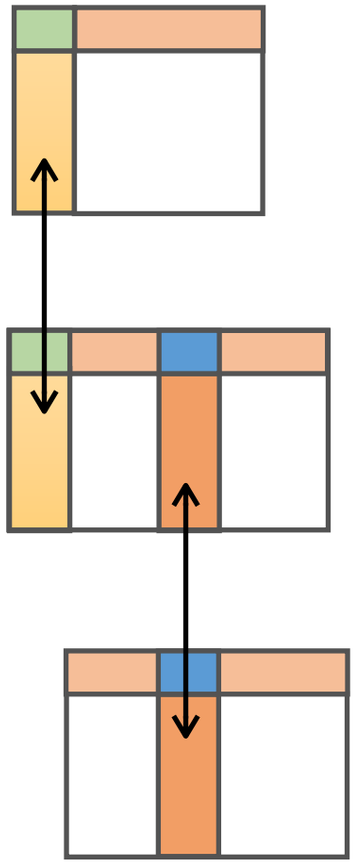
Demanda desagregada a nivel horario.
Registros > 244.000 por año

	A	B	C	D	E	F	G	H	I	J
1	FECHA	Cll	Tip	P01	P02	P03	P04	P05	P06	P07
2	2016-01-01	111	R	22,54	21,79	24,04	22,40	22,27	17,57	21,74
3	2016-01-01	111	N	1.499,15	1.378,47	1.398,94	1.353,87	1.299,21	1.261,15	1.165,96
4	2016-01-01	112	R	230,93	228,83	230,83	230,08	227,26	224,94	224,07
5	2016-01-01	112	N	1.052,35	1.066,18	1.067,42	932,78	848,43	821,37	828,92
6	2016-01-01	113	N	829,95	781,96	862,27	765,32	847,33	799,84	972,68
7	2016-01-01	119	N	35,33	37,97	35,04	43,43	46,28	34,90	44,09
8	2016-01-01	121	N	509,30	502,91	505,36	497,51	483,40	674,93	608,67
9	2016-01-01	121	R	13,02	13,12	12,66	12,52	9,90	12,36	11,93
10	2016-01-01	122	R	191,14	191,58	192,88	190,65	181,18	137,20	165,85
11	2016-01-01	122	N	1.902,79	1.867,25	1.837,59	1.812,63	1.789,16	1.765,61	1.699,16
12	2016-01-01	123	N	136,76	136,52	132,77	133,80	131,66	131,97	127,93
13	2016-01-01	123	R	14,38	15,02	14,92	13,82	13,63	14,82	17,69
14	2016-01-01	124	N	2.358,04	2.264,04	2.236,77	2.208,11	2.003,35	1.931,69	1.842,89
15	2016-01-01	124	R	465,83	468,99	415,05	370,19	386,83	474,15	548,28
16	2016-01-01	125	N	13.937,40	14.101,58	13.783,46	13.362,54	12.159,18	10.116,37	9.907,91

Se desea obtener la curva típica de consumo para el día

Almacenamiento de los datos

RDBMS

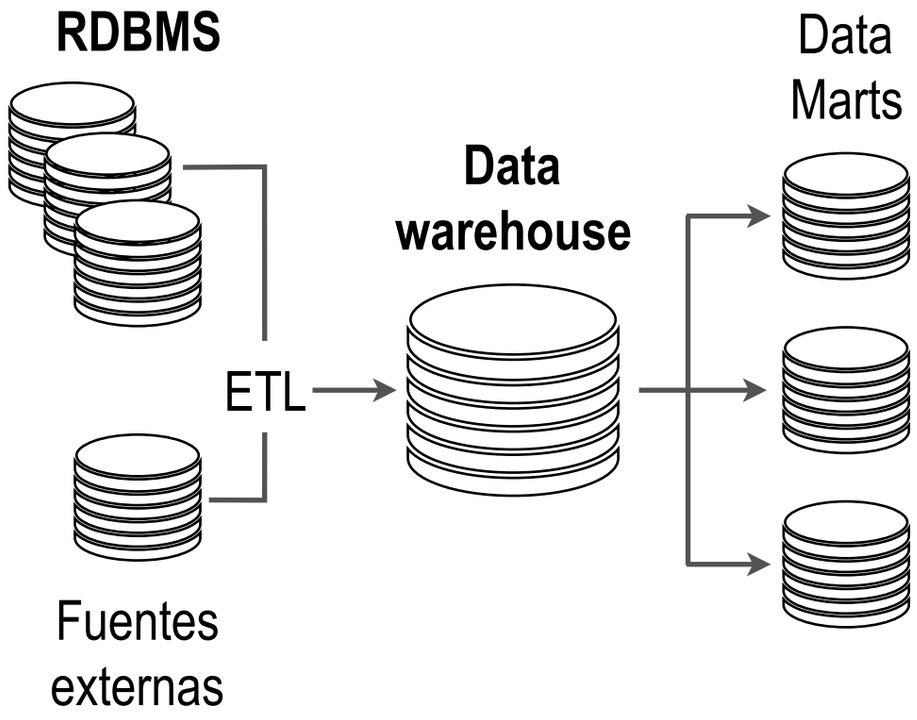


Componentes

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

Principales RDBMDS

- Oracle
- PostgressSQL
- Microsoft SQL
- Reportes
- Vistas
- Otros elementos



Business Intelligence

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

Data warehouse / OLAP

Bodegas de datos / Procesamiento analítico en línea

- Estructurado
- Integrado
- No volátil (permanencia de la información)
- Variable en el tiempo
- Orientado al análisis y la divulgación de la información
- Cubos multidimensionales

Inteligencia de Negocios

GENSCAPE™

Solutions | Knowledge Center | Events | Blog | News | About

Oil | Power | Natural

Overview

Daily Macro Supply & Demand D

Equity Production Insight

Intrastate Storage Monitoring

Natural Gas Analyst

GENSCAPE™

Solutions | Knowledge Center | Events | Blog | News | About

The Industry's **Only** Proprietary Natural Gas Data

LEARN MORE

Inteligencia de Negocios

información INTELIGENTE

Sign In | Buscar Metadata

Inicio | Demanda | Hidrología | Intercambios | Oferta | Transacciones y Prec

Información Inteligente

Indicadores

12,810.78 GWh
Volumen Útil Diario
Valor Anterior: 12,793.21 GWh
⬆.14%
2017-10-03 Ver Más

198.01 GWh
Aportes
Valor Anterior: 210.60 GWh
⬇6.36%
2017-10-03 Ver Más

188.14 GWh
Demanda Energía SIN Preliminar
Valor Anterior: 184.55 GWh
⬆1.91%
2017-10-03 Ver Más

155.15 GWh
Generación Hidráulica
Valor Anterior: 132.90 GWh
⬆14.34%
2017-10-02 Ver Más

17.01 GWh
Generación Térmica
Valor Anterior: 13.45 GWh
⬆20.93%
2017-10-02 Ver Más

0.09 GWh
Importaciones Preliminar
Valor Anterior: 0.12 GWh
⬇33.33%
2017-10-03 Ver Más

110.06 $/kWh
Precio Bolsa Promedio Aritmético TX1
Valor Anterior: 100.25 $/kWh
⬆8.91%
2017-10-02 Ver Más

32,521.85 $M
Transacciones en contratos
Valor Anterior: 32,521.85 $M
●.00%
2017-09-30 Ver Más

2,918.62 $M
Restricciones sin Alivios
Valor Anterior: 2,618.43 $M
⬆10.29%
2017-10-02 Ver Más

Inteligencia de Negocios

Inicio | Demanda | Hidrología | Intercambios | Oferta | Transacciones y P

Información Inteligente > Transacciones > Histórico Transacciones

- Precios
- AGC
- Bolsa
- Cargo por Confiabilidad
- Cargos por Uso
- Contratos
- Disponibilidad
- Desviaciones
- Impuestos
- Reconciliaciones

Tipo	Nombre	Descripción Contenido	Tamaño de archivo	Modificado
	AGC			25/01/2013 9:47
	Bolsa			20/11/2012 8:20
	Contratos			19/11/2012 16:33
	Desviaciones			19/11/2012 16:38
	Precios			19/11/2012 16:33
	Reconciliaciones			19/11/2012 16:29
	Servicios			08/08/2012 16:18
	Compras_Bolsa_Internacional_(kWh)_2016		106 KB	01/10/2017 14:42
	Compras_Bolsa_Internacional_(kWh)_2017		77 KB	03/10/2017 14:30
	Compras_Bolsa_Nacional_(kWh)_2016		5469 KB	01/10/2017 14:47
	Compras_Bolsa_Nacional_(kWh)_2017		4756 KB	03/10/2017 14:41
	Compras_Bolsa_TIE_(kwh)_2016		76 KB	01/10/2017 14:48
	Compras_Bolsa_TIE_(kwh)_2017		53 KB	03/10/2017 14:50
	Compras_Contrato_(kWh)_2016		7053 KB	01/10/2017 14:52
	Compras_Contrato_(kWh)_2017		5014 KB	03/10/2017 9:46
	Costo_Marginal_Despacho_Programado_($kWh)_2016		123 KB	01/10/2017 14:54
	Costo_Marginal_Despacho_Programado_($kWh)_2017		112 KB	03/10/2017 9:48
	Delta_Internacional_Delta_Nacional_($kWh)_2016		44 KB	01/10/2017 14:57
	Delta_Internacional_Delta_Nacional_($kWh)_2017		39 KB	03/10/2017 9:51
	Desviaciones_(kWh)_2016		333 KB	01/10/2017 15:00

Inteligencia de Negocios

energyone

HOME | ABOUT US | MARKETS SERVED | PRODUCTS | SERVICES | INVESTORS | CONTACT US

EnergyDashboard

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

Inteligencia de Negocios

Login

argus

Crude Oil | Oil Products | Coal | News & Insight | About Arg
LPG/NGL | Bioenergy | Power | Argus Consulting | Methodolo
Emissions | Natural Gas/LNG | Transportation | Services | Reference
Fertilizer | Petrochemicals | Metals | Events | Contact Us

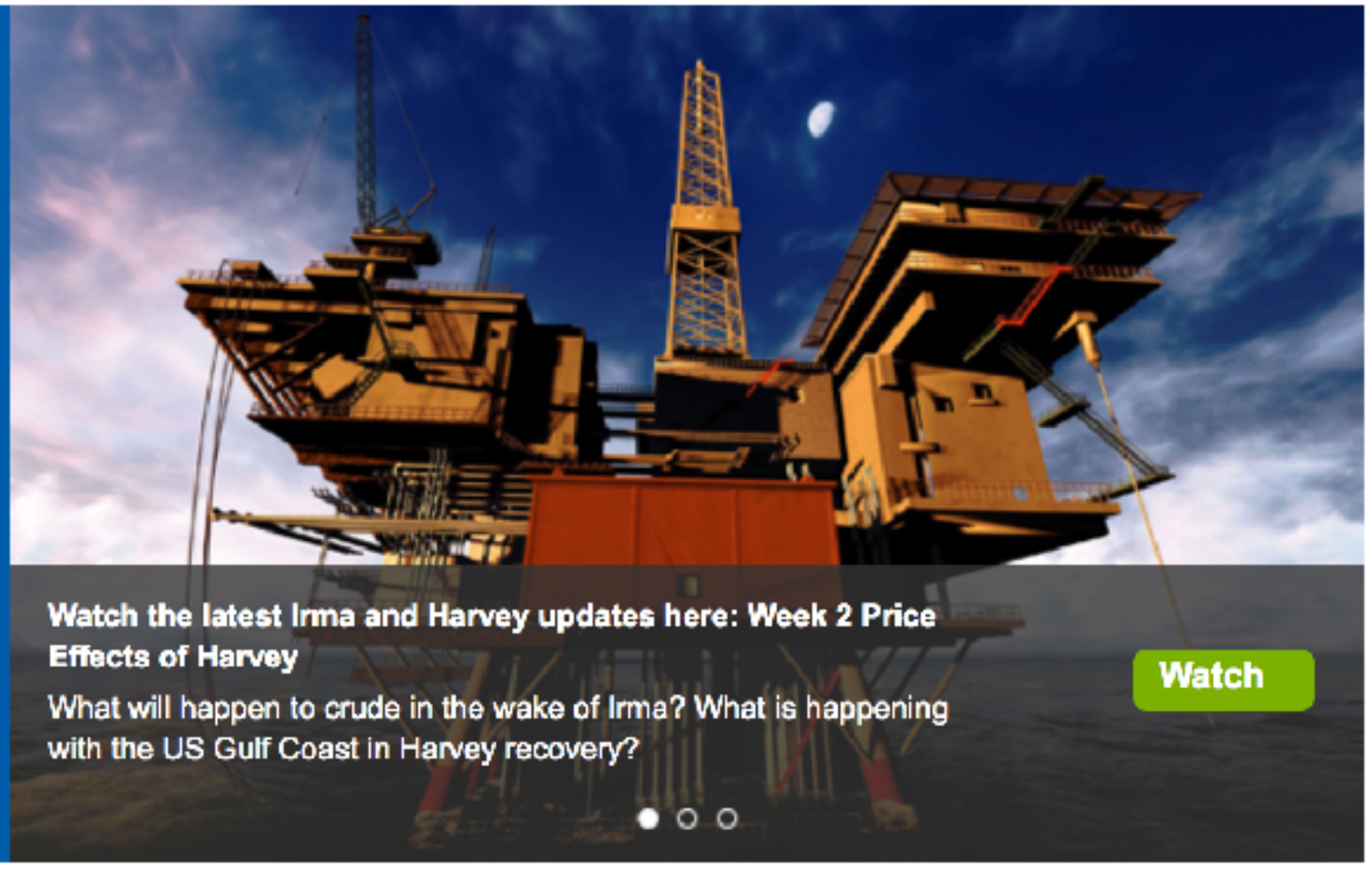
Home / Crude Oil

Русский | Portuguese | 日本 | 中国 | Español

Crude Oil

Argus crude services cover global and specialist local markets, offering exclusive prices and insight.

Our product range includes daily international prices and trading information, intelligent analysis and fundamental data on supply and demand, stock levels as well as logistical updates, events and consulting services.



Almacenamiento de los datos

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}
```

Big Data (Hadoop)

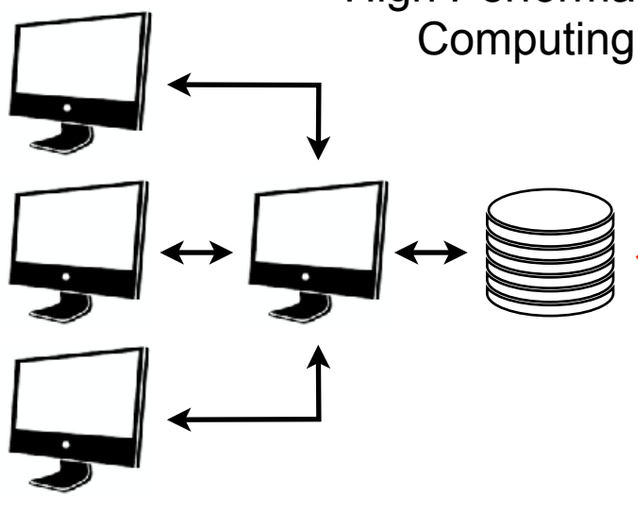
Aproximación Tradicional
(RDBMS y data warehouse)

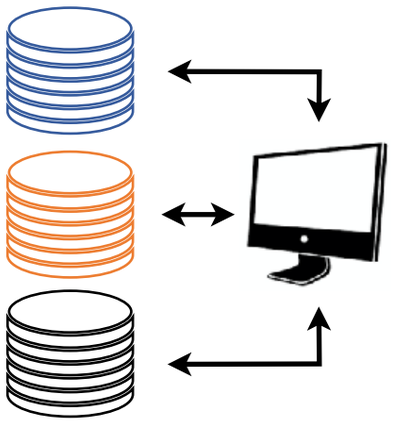
- Alto volumen de datos
- Lectura lenta
- Escritura + lenta
- Búsqueda ++ lenta

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

Procesamiento distribuido

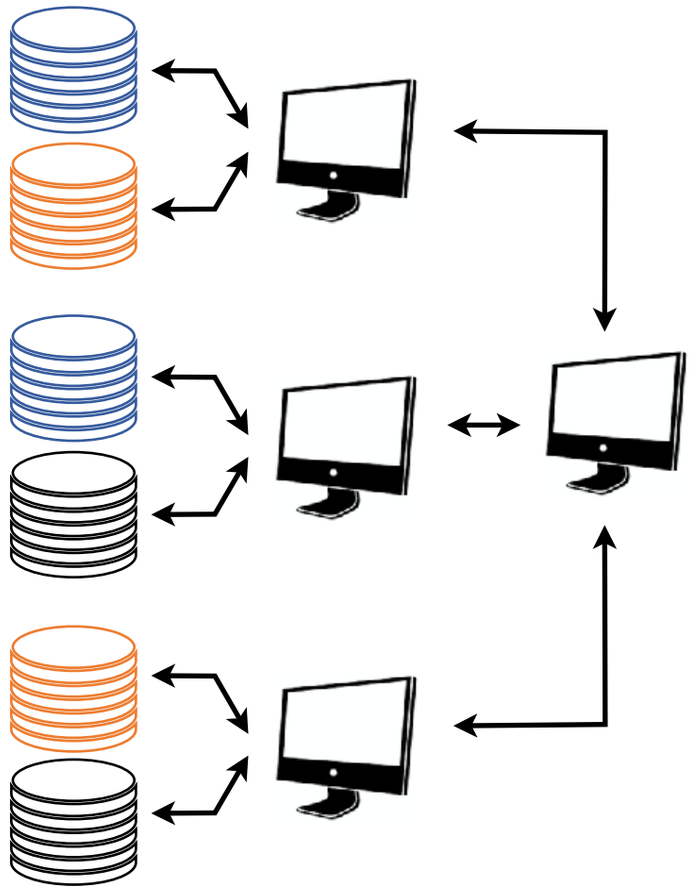
High Performance Computing





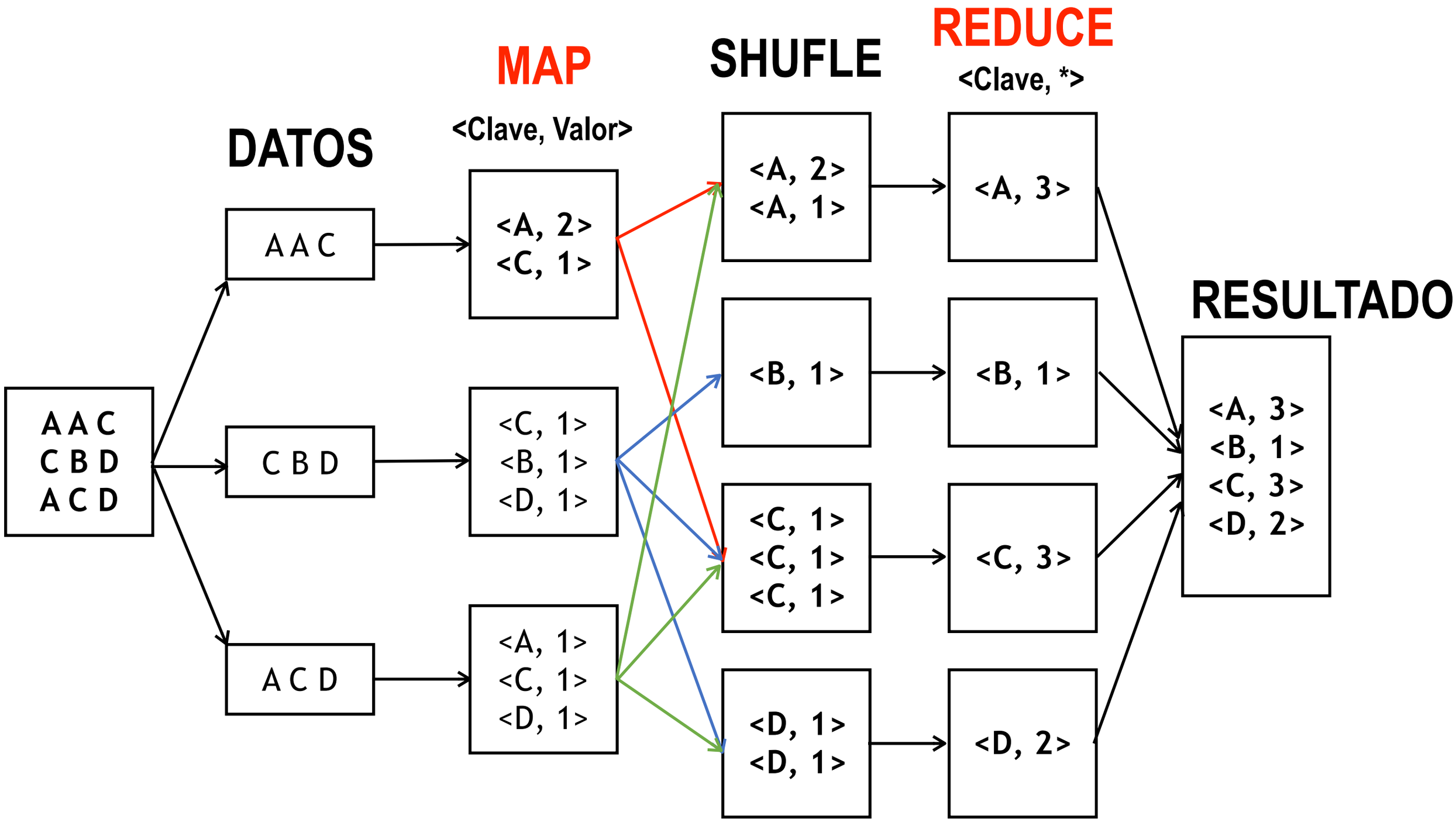
Almacenamiento distribuido

Hadoop Distributed Filesystem (HDFS)



Nodos de Datos + Nodos de Computo

Hadoop / MapReduce



Ecosistema Apache Hadoop

HDFS
Operaciones básicas del sistema de archivos

MapReduce
(Java)

Hive
Lenguaje de consultas similar a SQL.

Mahout
Algoritmos de Machine Learning implementados usando MarReduce (Java)

Pig
Lenguaje para el procesamiento de datos.

HBase
Base de datos orientada a columnas.

Sqoop
RDBMS <—> Hadoop

YARN
Programación de trabajos y manejo de recursos en clusters

AVRO
Sistema de serialización de datos

CASSANDRA
Sistema de Base de Datos

ZooKeeper
Coordinación

Oozie
Flujo y programación de trabajos

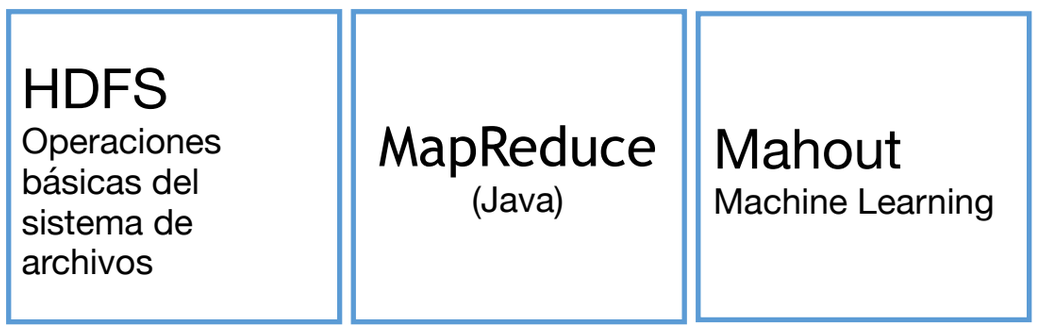
Ejemplo de Pig

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

Ejemplo de Hive

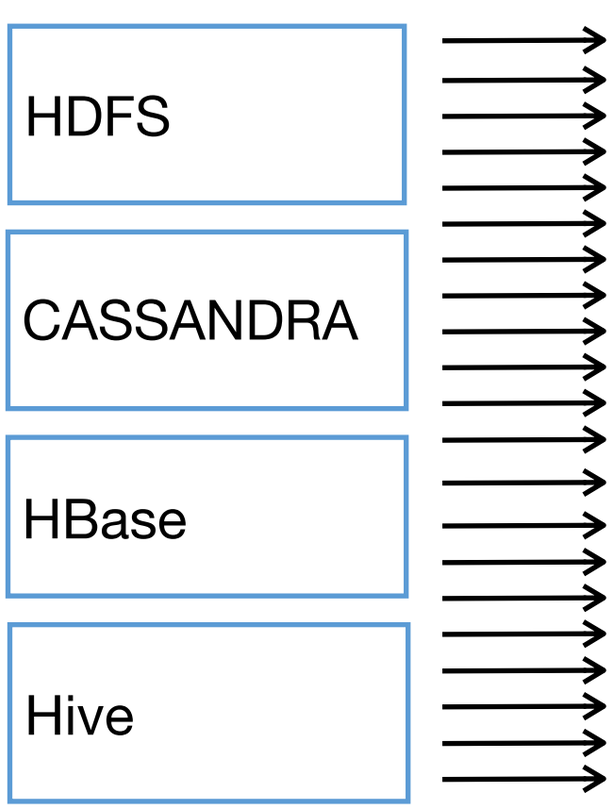
```
CREATE TABLE records (year STRING, temperature INT, quality INT)
  ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';
LOAD DATA LOCAL INPATH 'sample.txt' OVERWRITE INTO TABLE records;
SELECT year, MAX(temperature) FROM records GROUP BY year;
```

Mahout, MLib, R y Python



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

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



Java
Scala
Python
R

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

Mahout vs MLib

Collaborative Filtering *with CLI drivers*

User-Based Collaborative Filtering

Item-Based Collaborative Filtering

Matrix Factorization with ALS

Matrix Factorization with ALS on Implicit Feedback

Weighted Matrix Factorization, SVD++

Classification *with CLI drivers*

Logistic Regression - trained via SGD

Naive Bayes / Complementary Naive Bayes

Hidden Markov Models

Clustering *with CLI drivers*

Canopy Clustering

k-Means Clustering

Fuzzy k-Means

Streaming k-Means

Spectral Clustering

Dimensionality Reduction *note: most scal*
reduction algorithms are available through t
Core Library for all engines
Singular Value Decomposition

Lanczos Algorithm

Stochastic SVD

PCA (via Stochastic SVD)

QR Decomposition

Topic Models

Latent Dirichlet Allocation

Miscellaneous

RowSimilarityJob

Collocations

Sparse TF-IDF Vectors from Text

XML Parsing

Email Archive Parsing

Evolutionary Processes

MLib
Machine Learning

Algorithms

MLlib contains many algorithms and utilities.

ML algorithms include:

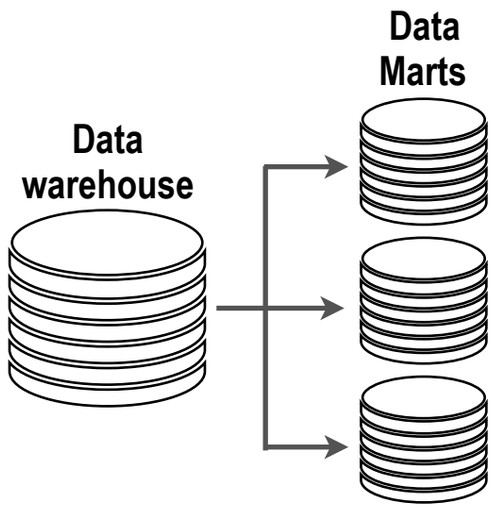
- Classification: logistic regression, naive Bayes,...
- Regression: generalized linear regression, survival regression,...
- Decision trees, random forests, and gradient-boosted trees
- Recommendation: alternating least squares (ALS)
- Clustering: K-means, Gaussian mixtures (GMMs),...
- Topic modeling: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

ML workflow utilities include:

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

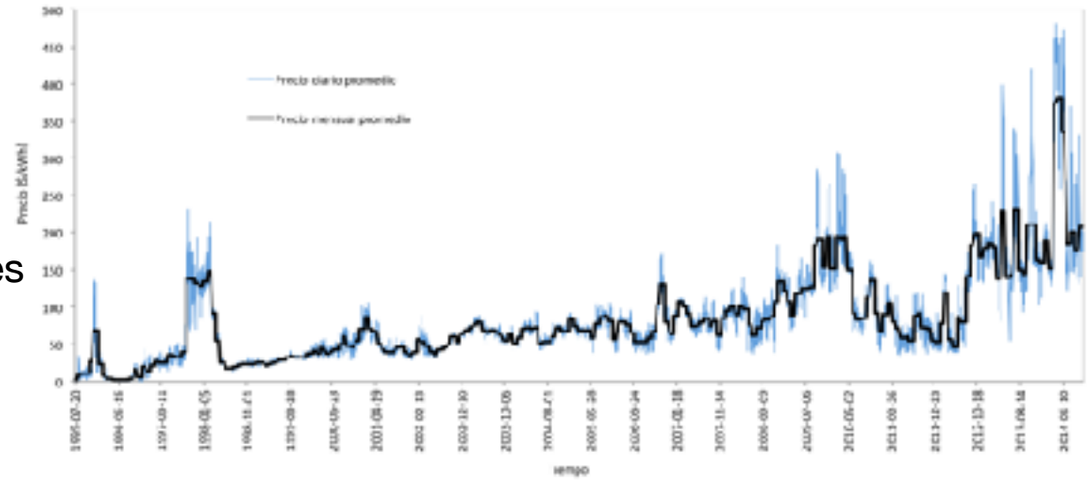
Other utilities include:

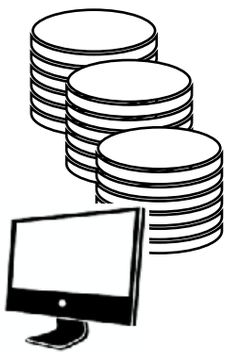
- Distributed linear algebra: SVD, PCA,...
- Statistics: summary statistics, hypothesis testing,...



Data Mining

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





Data warehouse
NoSQL
Datos columnares
Texto
CSV
Datos no estructurados
Web
IoT

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?)

Analytics

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

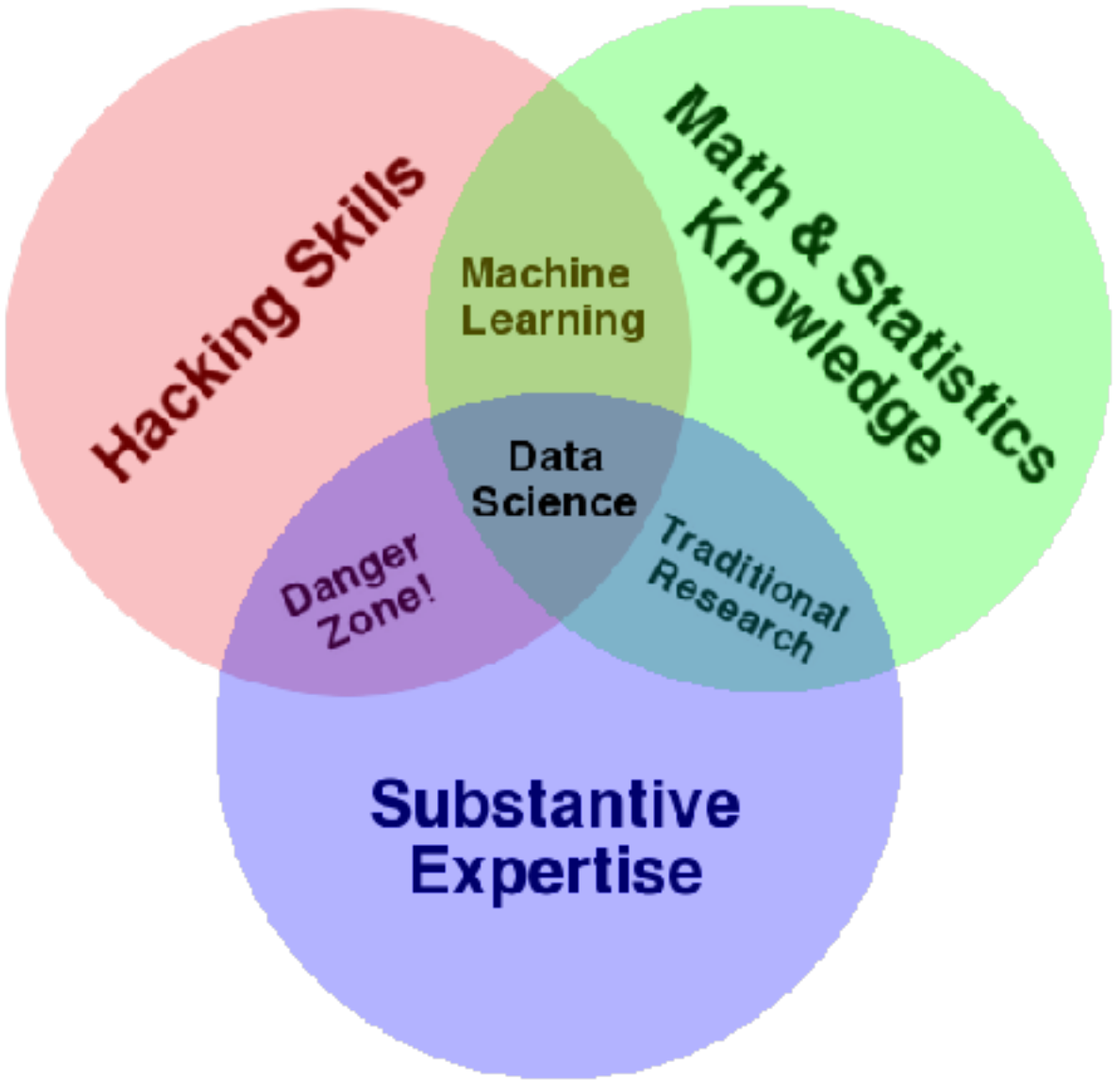
I. Problema organizacional
II. Transformación en un problema de analytics
III. Datos
IV. Selección de la metodología
V. Desarrollo del modelo
VI. Puesta en marcha (deploy)
VII. Gestión del ciclo de vida del modelo

Big Data Analytics

Analytics sobre grande conjuntos de datos

Algoritmos, técnicas y metodologías

Diagrama de Ven explicando qué es Data Science y habilidades requeridas



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

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

Similitudes y Diferencias

DATA SCIENCE

Programación.

Adquisición, limpieza, preprocesamiento y visualización de datos.

Investigación reproducible.

Modelado de Datos (minería de datos)

- Inferencia Estadística.
- Modelos estadísticos
- Aprendizaje de Máquinas

Productos de Datos.

ANALYTICS

Programación

Adquisición, limpieza, preprocesamiento y visualización de datos.

Modelado de Datos (modelado predictivo)

- Inferencia Estadística.
- Modelos estadísticos
- Aprendizaje de Máquinas

Productos de Datos.

Inteligencia de Negocios.

Simulacion.

Optimización.

Métodos prescriptivos: modelos predictivos + optimización.

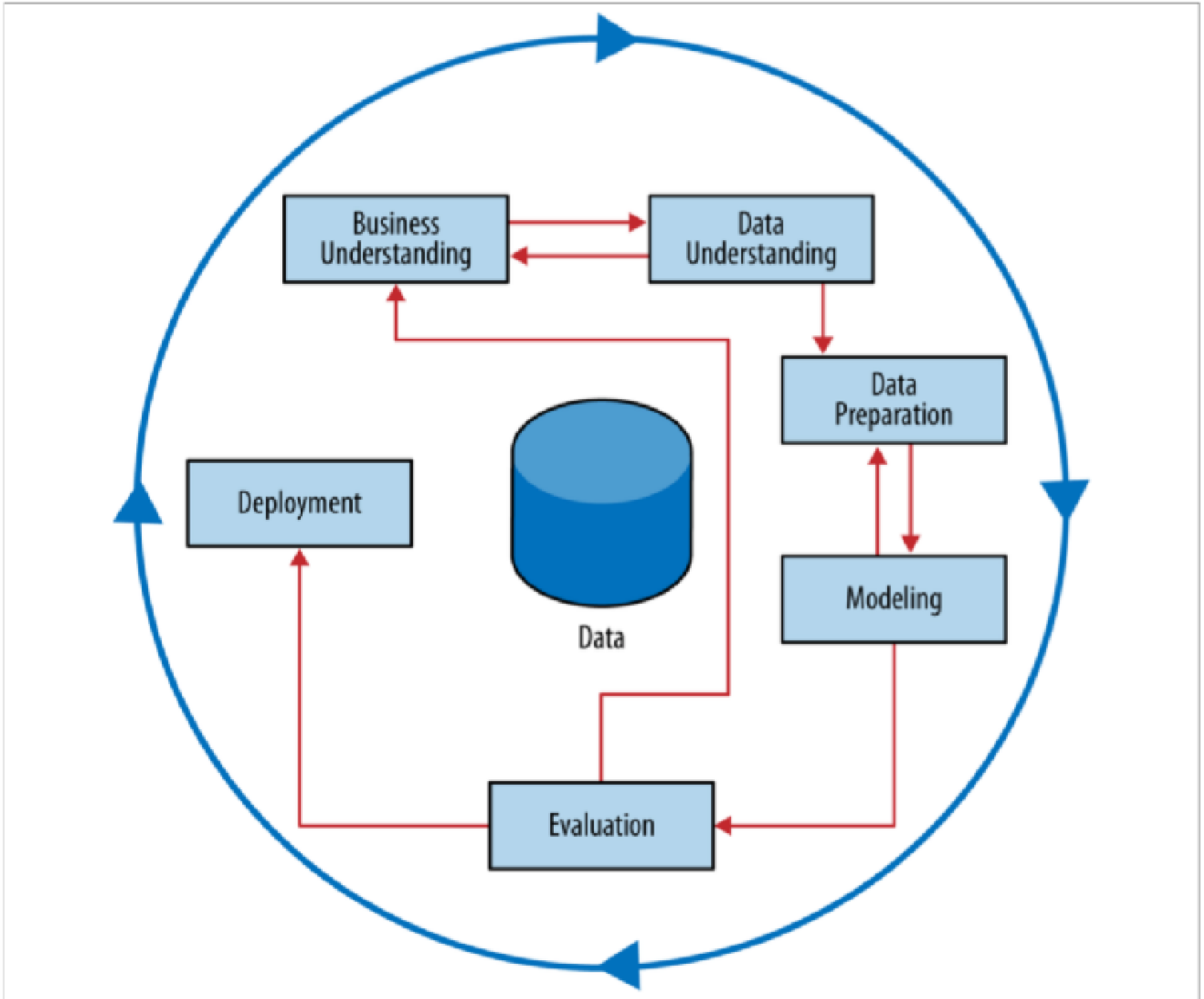
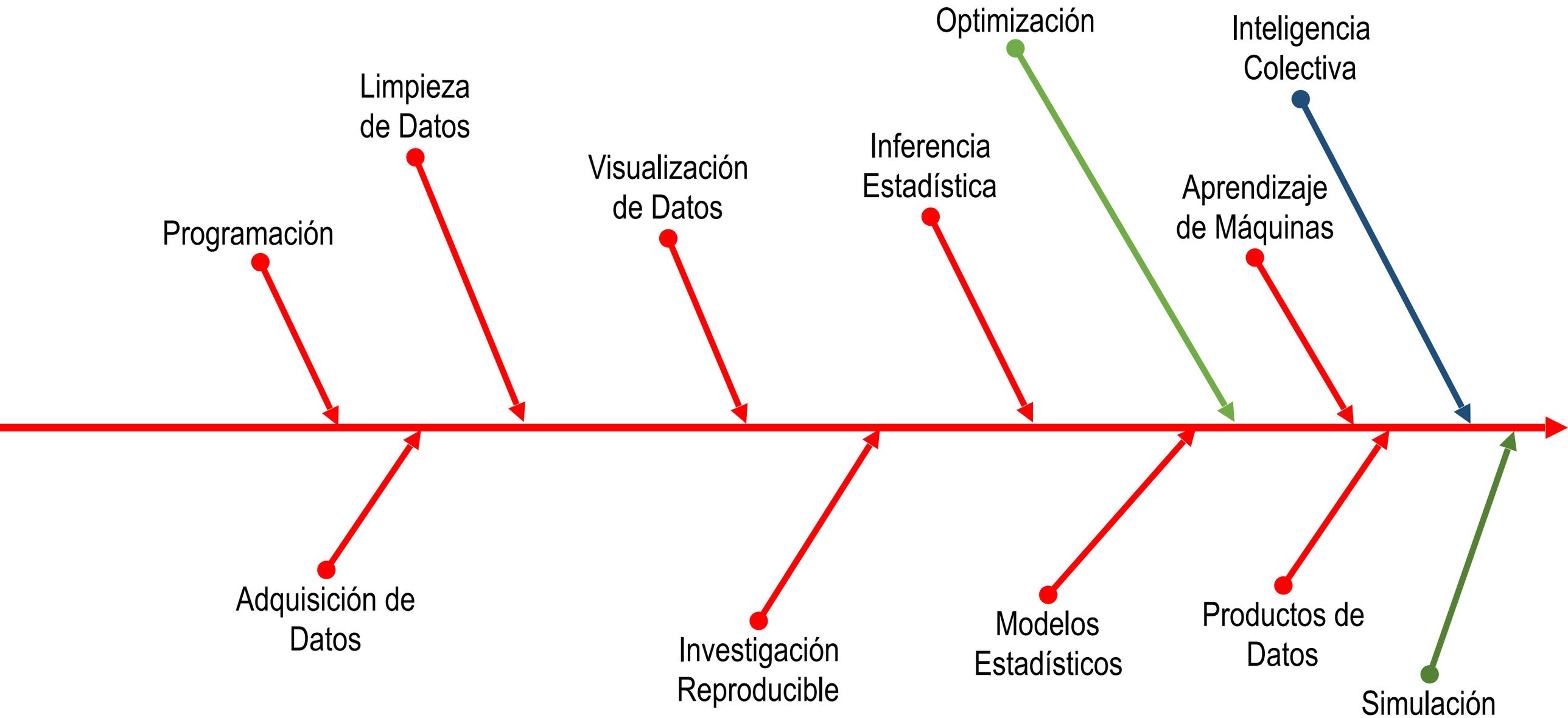


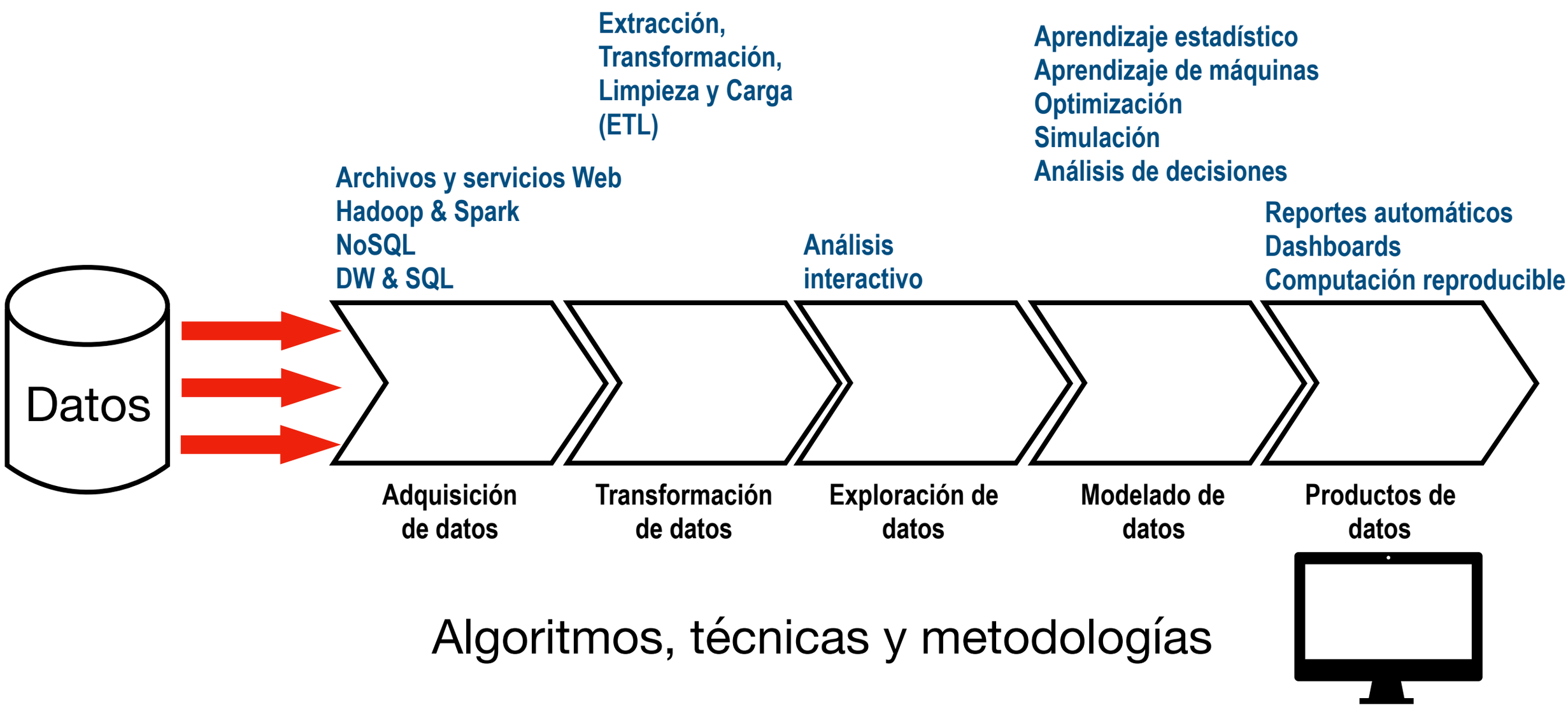
Figure 2-2. The CRISP data mining process.

Componentes de Data Science / Analytics



Data-driven decision making!

Proceso de modelado en Analytics



Algoritmos, técnicas y metodologías

Infraestructura computacional

- Un procesador
- Muchos procesadores

- Computación en máquinas locales
- Computación en la nube

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

¿Ordenar un vector de números?

¿Calcular la suma de los primeros 20 números primos?

¿Computar la inversa de una matriz?

Programación para ingeniería

Cómputo numérico.

Programación para Computer Sciences

Manipulación de texto.

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

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

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

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

## etcetera …

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

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

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

Ejemplo de Pig

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

Ejemplo de Hive

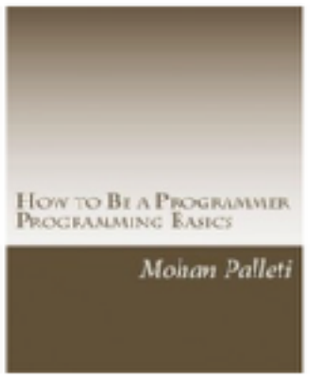
```
CREATE TABLE records (year STRING, temperature INT, quality INT)
  ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';
LOAD DATA LOCAL INPATH 'sample.txt' OVERWRITE INTO TABLE records;
SELECT year, MAX(temperature) FROM records GROUP BY year;
```

Programación -- www.allitebooks.com

Algorithms Unlocked

By: Thomas H. Cormen

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



How to Be a Programmer: Programming Basics

By: Mohan Palleti

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

Programación **vs** Aplicaciones de usuario final

R

Python

Julia

Bash

C

Revolution Analytics (R)
(Microsoft R)

Enthought Canopy
(Python)

Anaconda

Usuario

Analytics Solver
(Frontline Systems)

KNIME

Pentaho

IBM Watson explorer
IBM SPSS Modeler

SAS predictive
Analytics
RapidMiner

Oracle Data Mining

TIBCO Analytics

Salford Systems

Tableau

Qilk

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

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

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

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

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

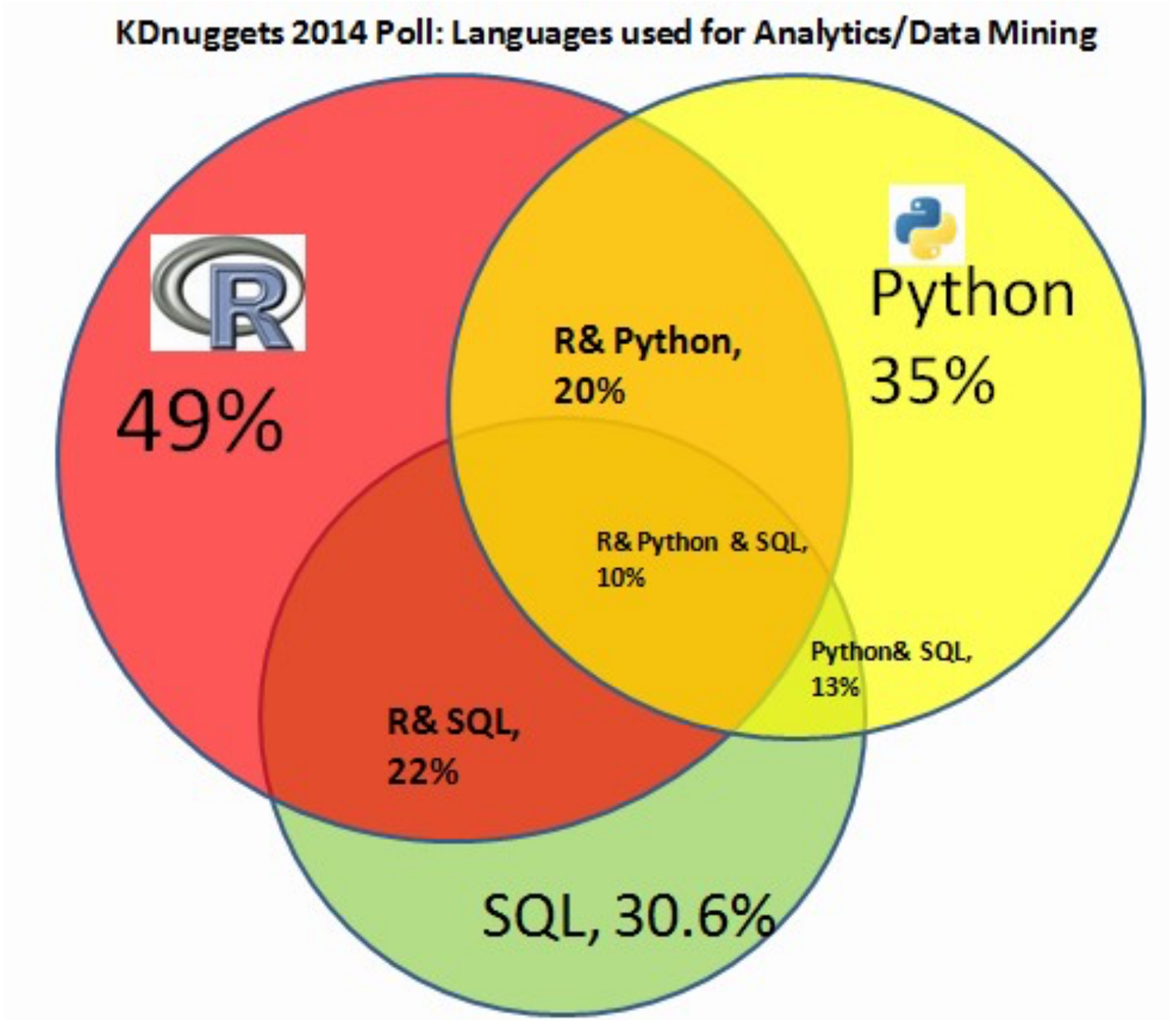
http://spectrum.ieee.org/computing/software/the-2016-top-programming-languages

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

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

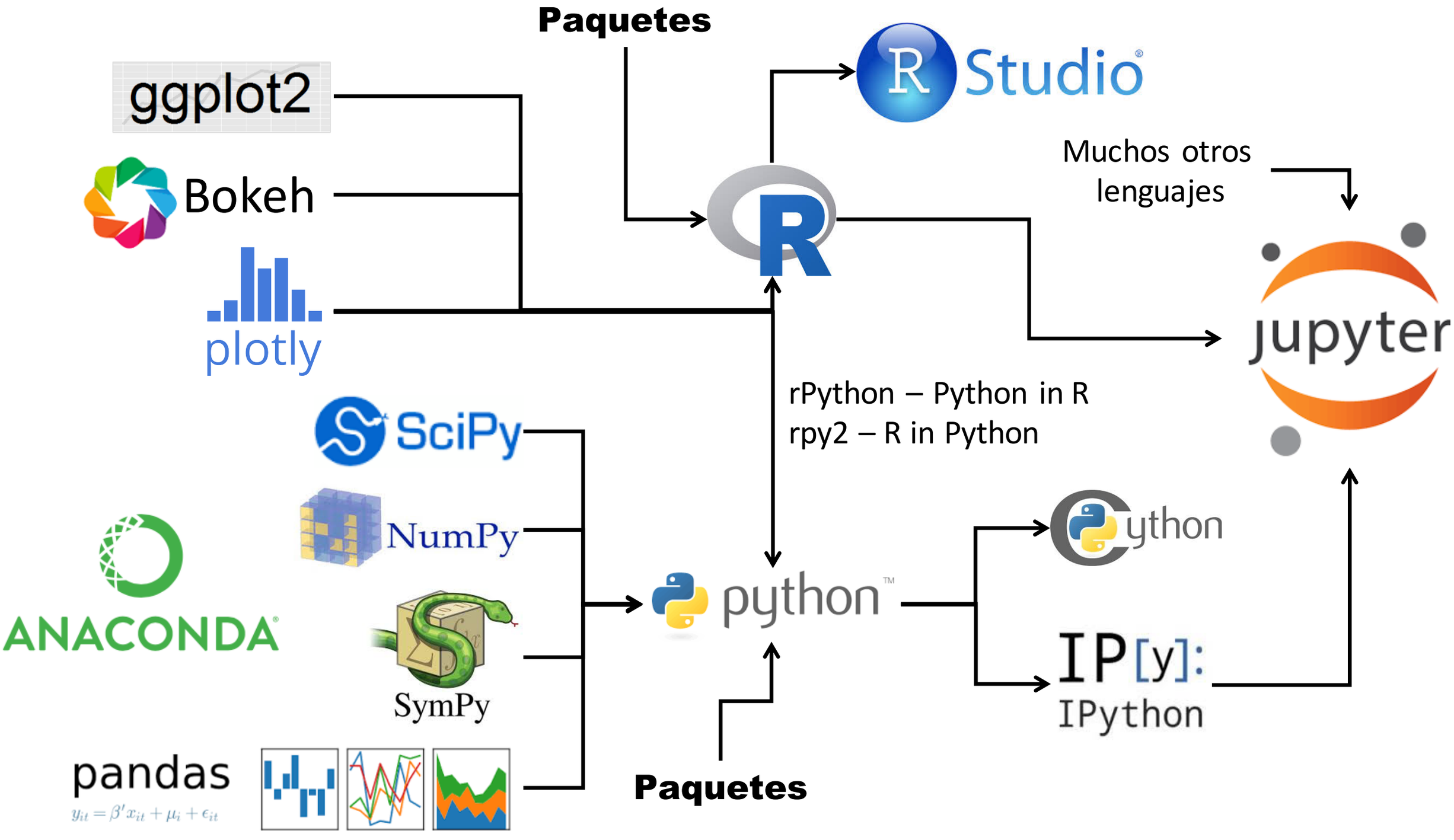
http://spectrum.ieee.org/computing/software/the-2017-top-programming-languages

Popularidad de los lenguajes



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

Ecosistema de herramientas y lenguajes de programación



Adquisición y Limpieza de Datos

TXT, Excel, CSV, PDF, *.docx.

Páginas web (HTML) y Google Groups.

Bases de datos relacionales.

Lenguaje Natural.

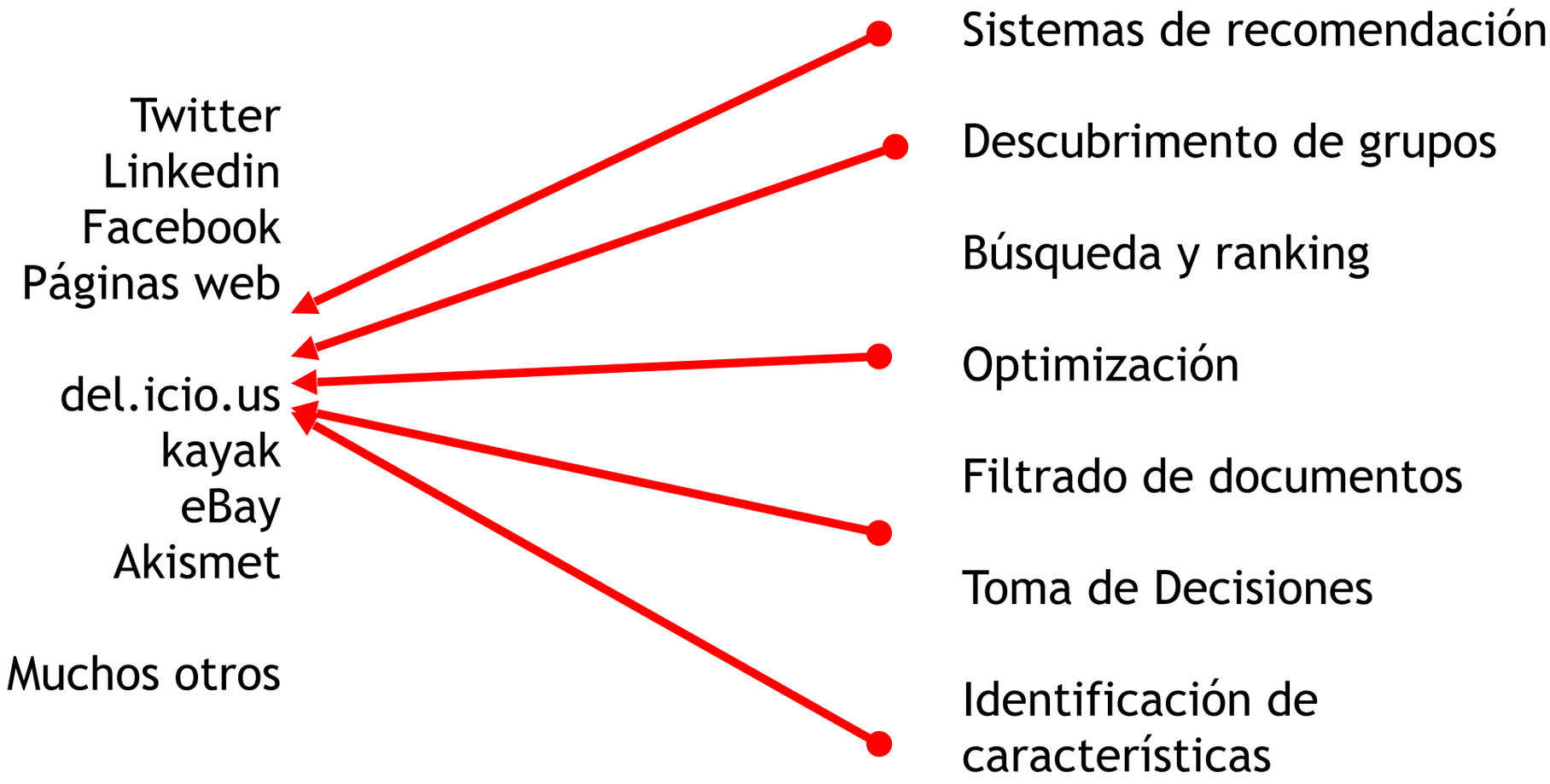
Imágenes (Captcha)

Manipulación de texto

Conversión de un formato a otro

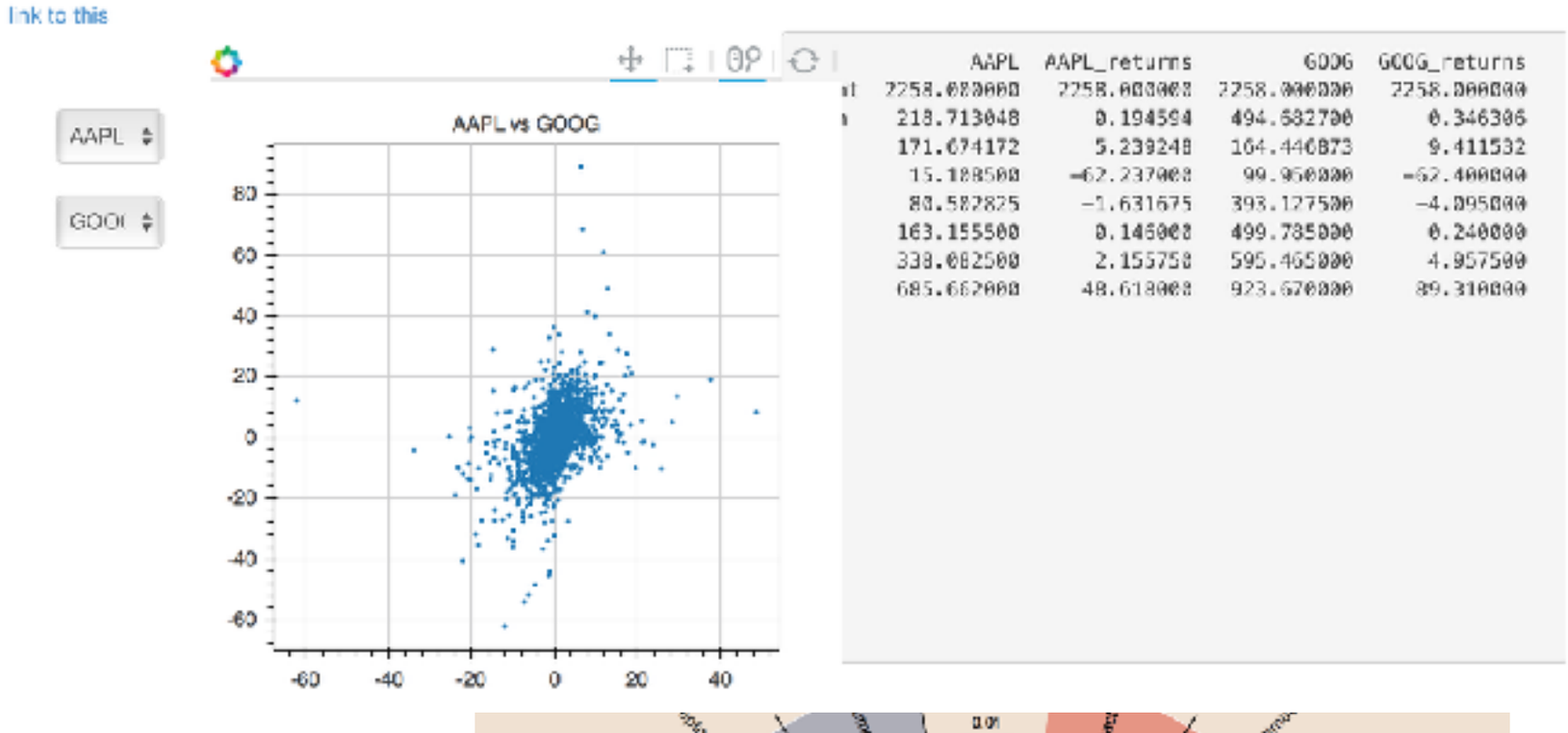
Detección de datos faltantes, datos nulos, datos inconsistentes

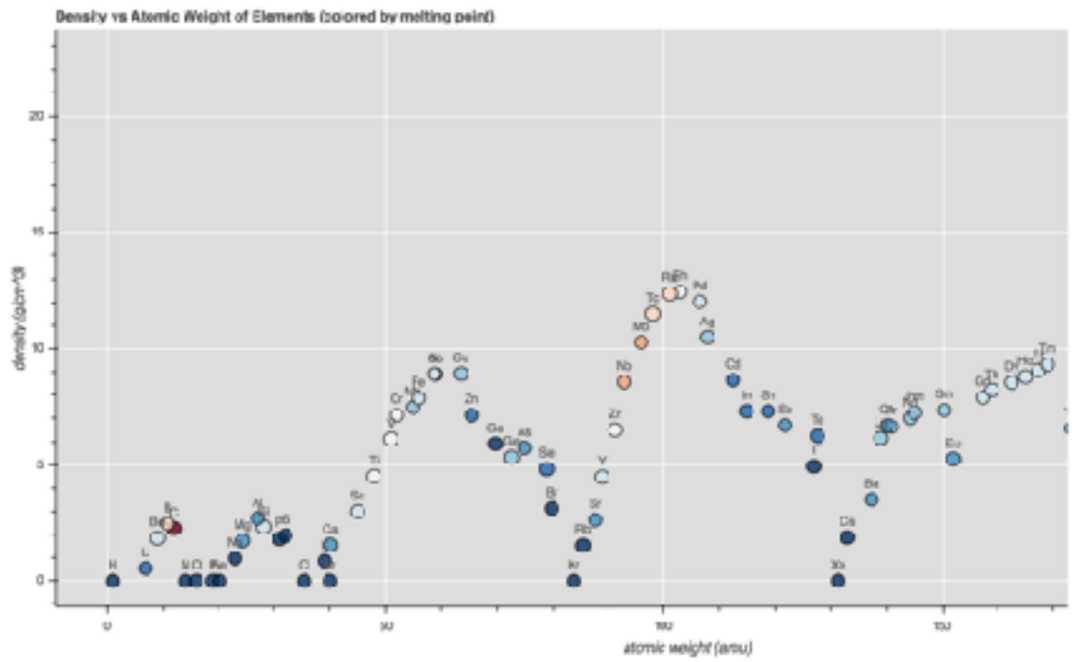
Adquisición de datos -- Inteligencia colectiva

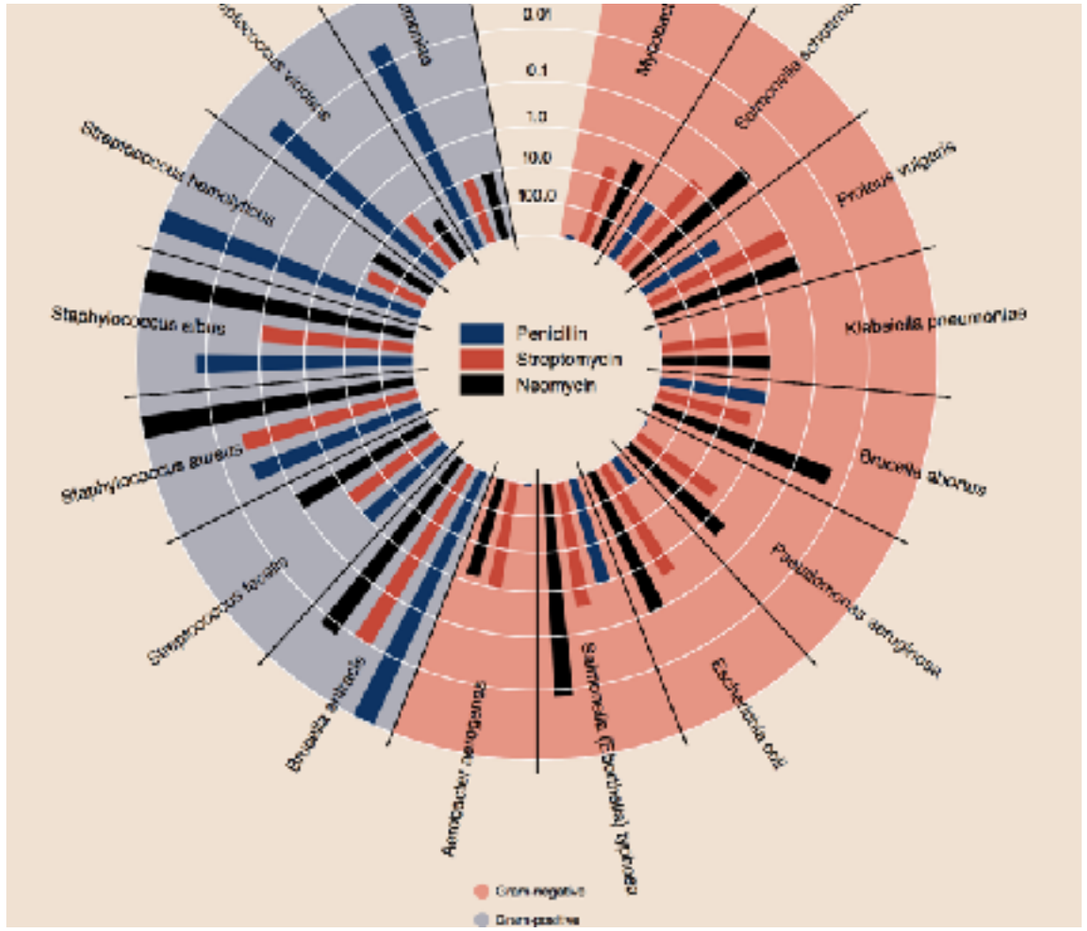


Visualización de Datos

- Bokeh
- Matplotlib
- R
- ....

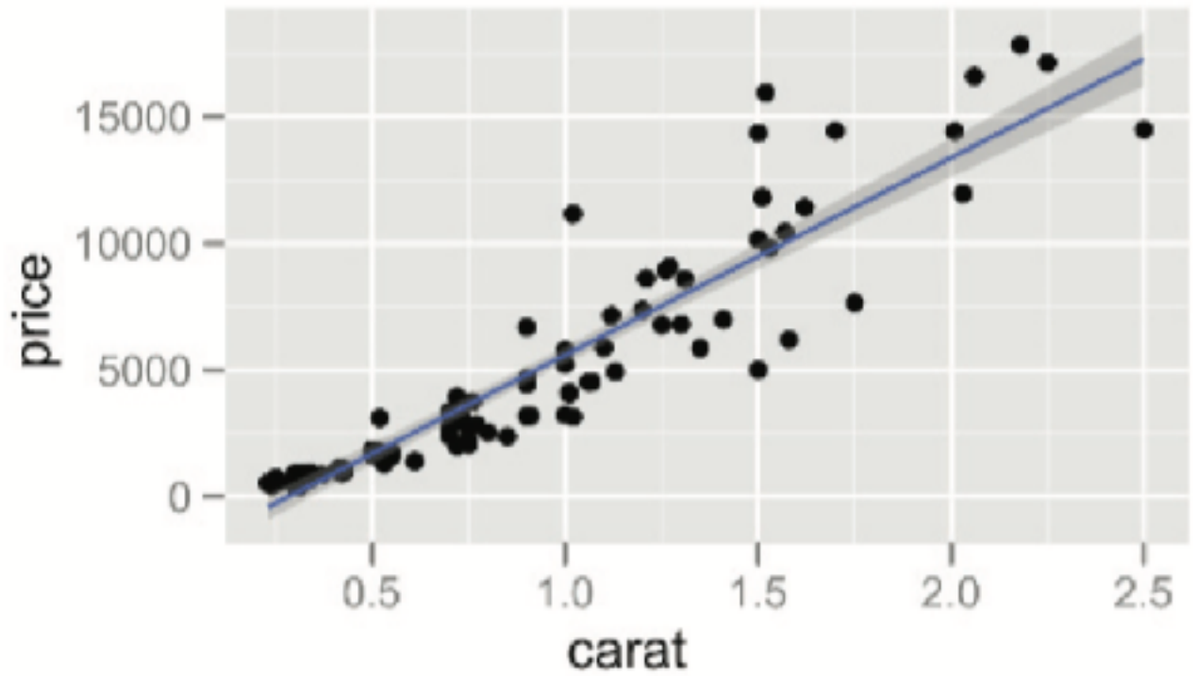


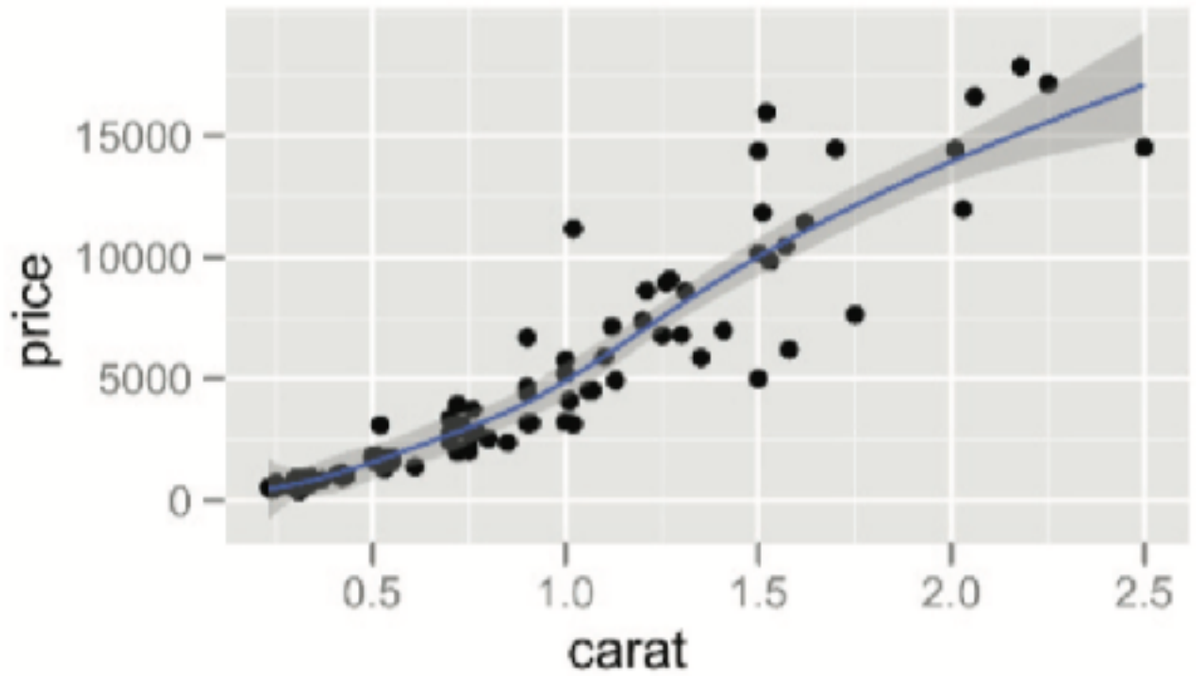




Modelado Estadístico y Aprendizaje de Máquinas

Aplicación clásica





¿Y si hay 10 millones de datos?

Computación Reproducible (Markdown)

Markdown Editor

/

Input

```
Inline link: [destination](<index.html>)

Reference link: [destination][1]

Reference link: [reference link]

[1]: <index.html>
[reference link]: <http://www.infopark.com> "Link title"

Automatic link: <http://daringfireball.net/projects/markdown />

    This is a blockquote
    (pre + code)

---

Heading 1
=========

Heading 2
---------

### Heading 3

#### Heading 4

##### Heading 5

###### Heading 6


* List item 1
* List item 2
  * Subitem 2.1
  * Subitem 2.2
```

Preview

Inline link: destination
Reference link: destination
Reference link: reference link
Automatic link:

```
This is a blockquote
(pre + code)
```

Heading 1

Heading 2

Heading 3

Heading 4

Heading 5

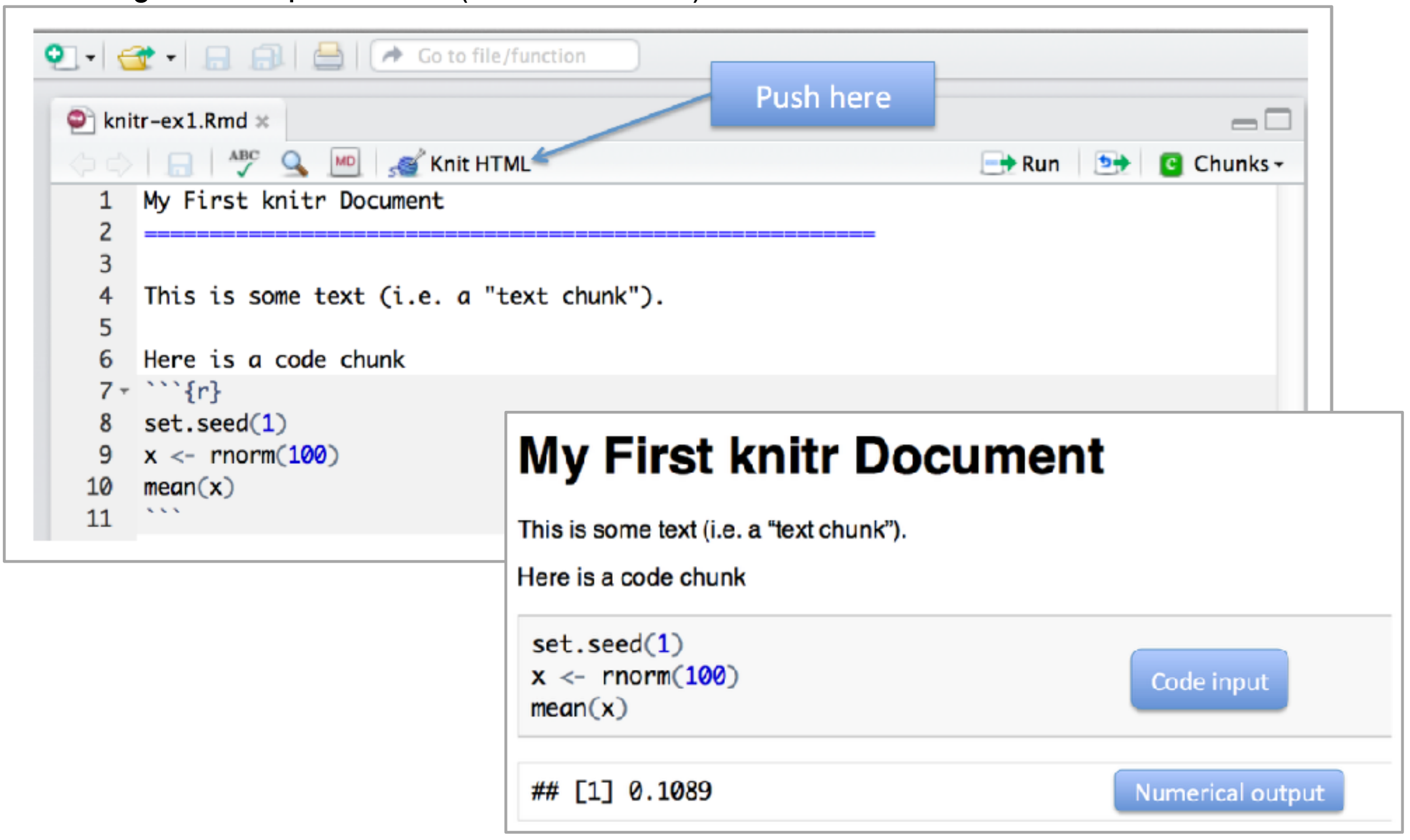
Heading 6

- List item 1
- List item 2
  - Subitem 2.1
  - Subitem 2.2

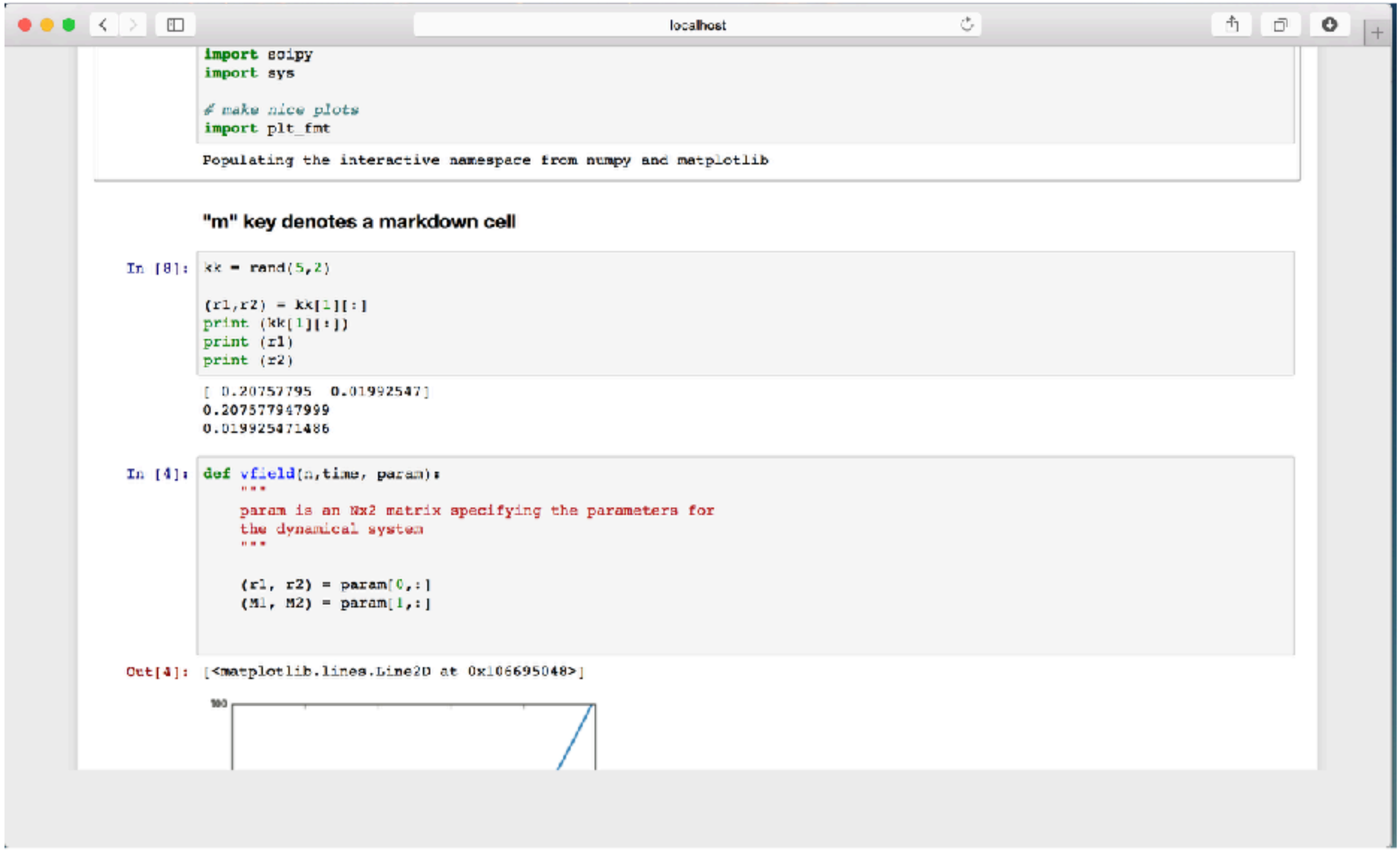
Help on this page

Ok Cancel

Investigación Reproducible (Markdown + R)

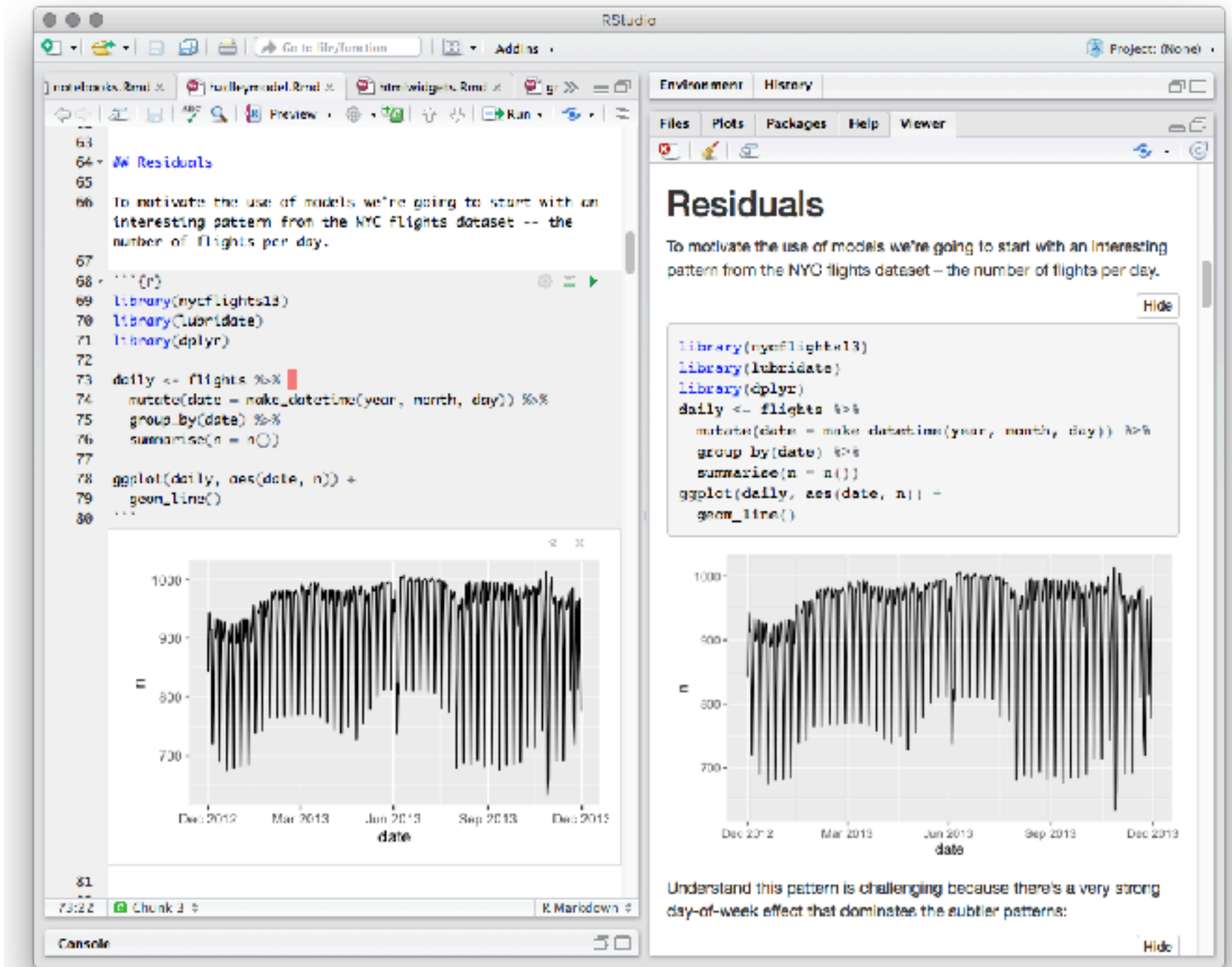


Computación Reproducible (Jupyter Notebook)



Productos de datos

- Informes autocalculables.
- Tableros de control (Dashboards)
- Aplicaciones de datos



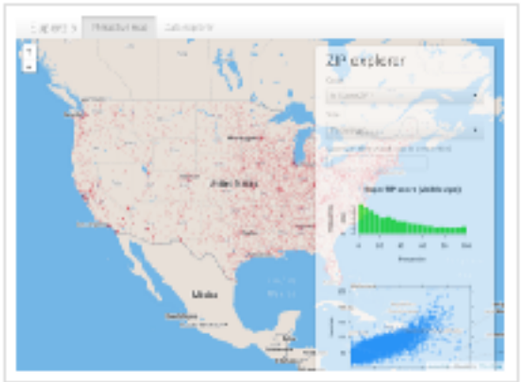
R Shiny

Gallery

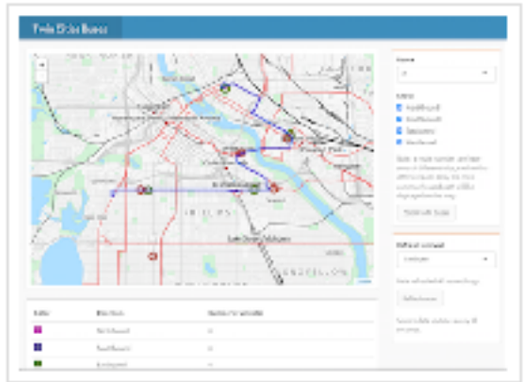
This gallery contains useful examples to learn from. Visit the Shiny User Showcase to see an inspiring set of sophisticated apps.

Interactive visualizations

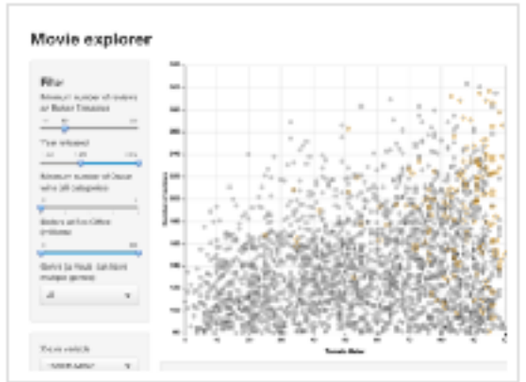
Shiny is designed for fully interactive visualization, using JavaScript libraries like d3, Leaflet, and Google Charts.



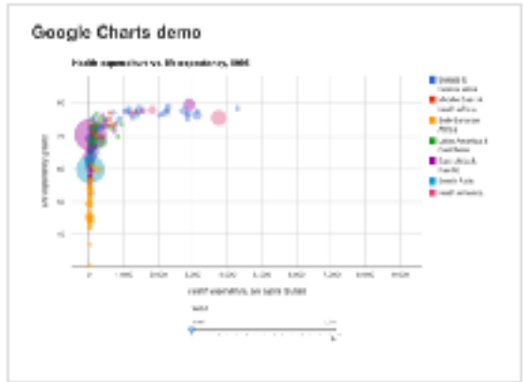
SuperZip example



Bus dashboard



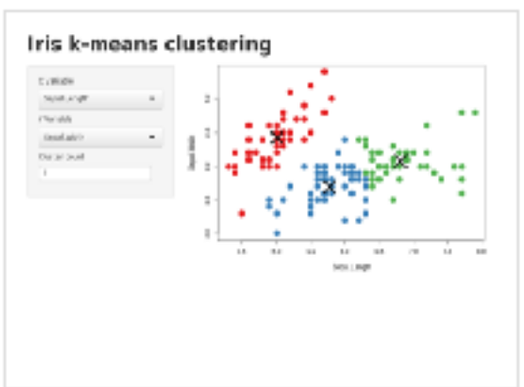
Movie explorer



Google Charts

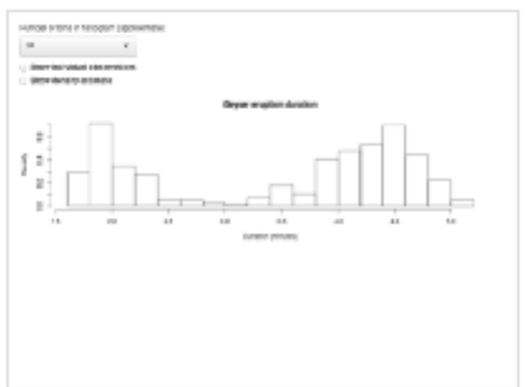
Start simple

If you're new to Shiny, these simple but complete applications are designed for you to study.

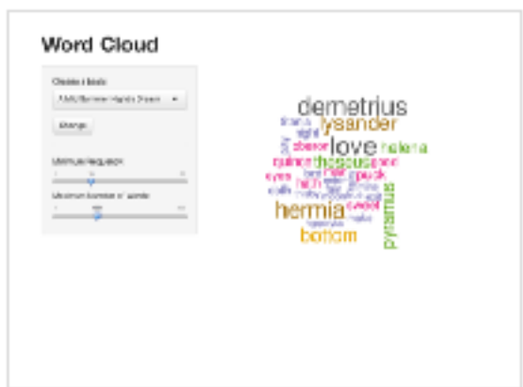


Kmeans example

Telephones by region

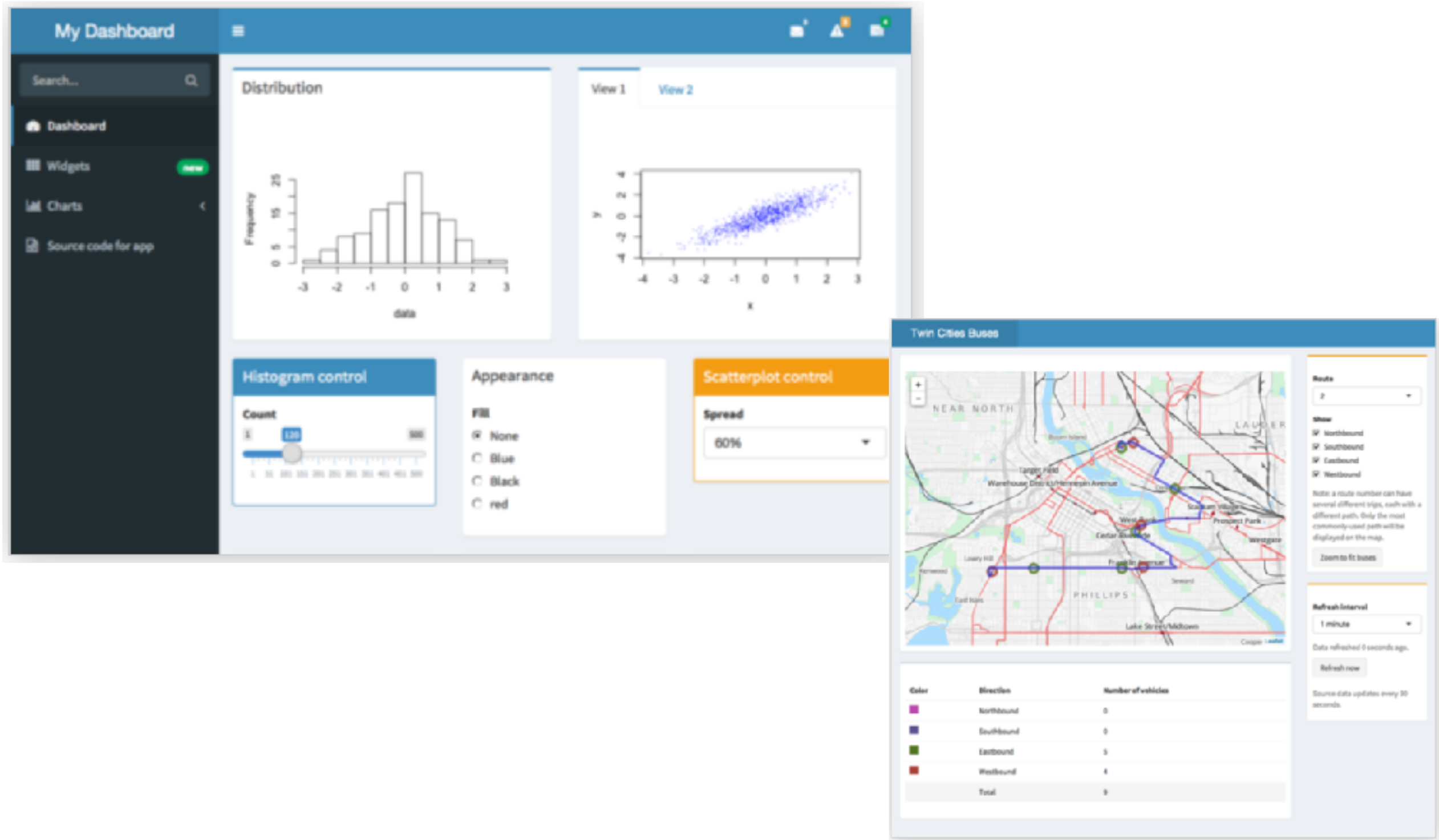


Faithful



Word cloud

R Dashboards



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

(con ejemplos en mercados de energía)

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

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

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

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